CREDIT EXPLORATORY DATA ANALYSIS

By : A S Sushmitha Urs

Problem Statement

- Examine loan characteristics and consumer behavior to determine how they affect loan defaults.
- In order to reduce risk in lending decisions, make sure qualified candidates are not turned down.
- Situations: If creditworthy applicants are turned down, there will be a loss of business.
- monetary damages in the event that high-risk applications are accepted.
- Objective: Determine the "driver variables" that contribute to loan defaults and use them to reduce risk.

Assumptions

- To keep the dataset complete and pertinent, missing data was handled by eliminating columns with more than 19% missing values.
- used the mean, median, or mode to impute the remaining missing values.
- Outliers were found, however they were kept for analysis
- resolved the target variable's data imbalance.

Approach and Methodology

- Data Cleaning: Data cleaning: Columns with a high percentage of missing values were removed, while others were imputed.

- EDA: Performed segmentation analysis to look at patterns unique to defaulters and non-defaulters, univariate analysis to comprehend the distributions of individual variables, and bivariate analysis to investigate correlations between variables.

- Correlation Analysis: To determine the most important variables that are highly linked to loan defaults, pairwise correlations were examined. This analysis shed light on the variables that have a major impact on default risk.

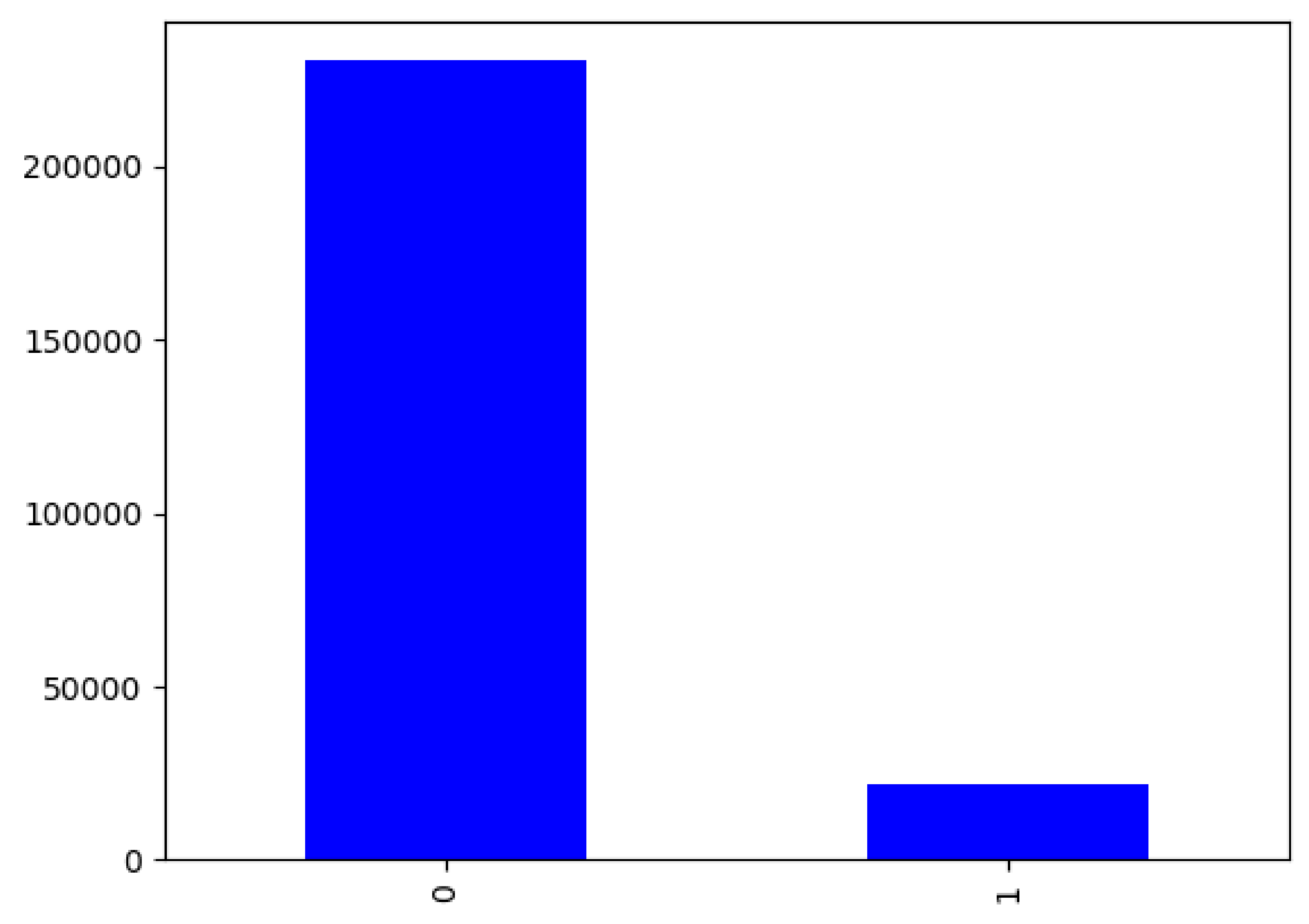
- Visualization: Insightful visualizations were produced using the Python tools Seaborn and Matplotlib.

Application dataset

Data Imbalance

- Dataset Imbalance: The majority of applicants successfully return their loans, indicating a notable imbalance in the dataset.

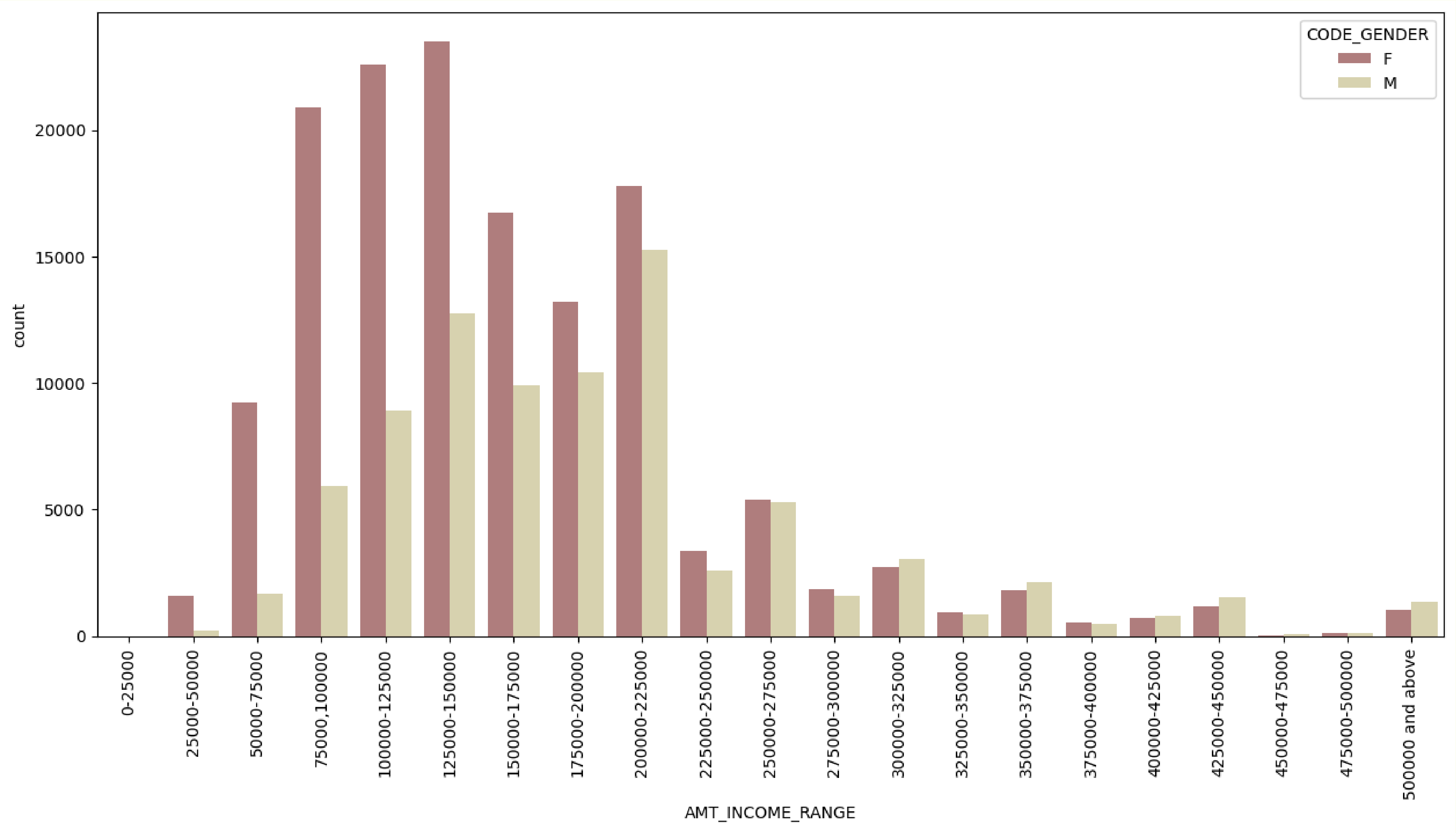
- Defaulter Proportion: There are much fewer applicants who defaulted (target = 1) than those who did not (target = 0), underscoring the need for strategies to address class disparity.



Income vs. Loan

- Status Income and Default Tendency: Applicants who earn more money are less likely to experience loan default.

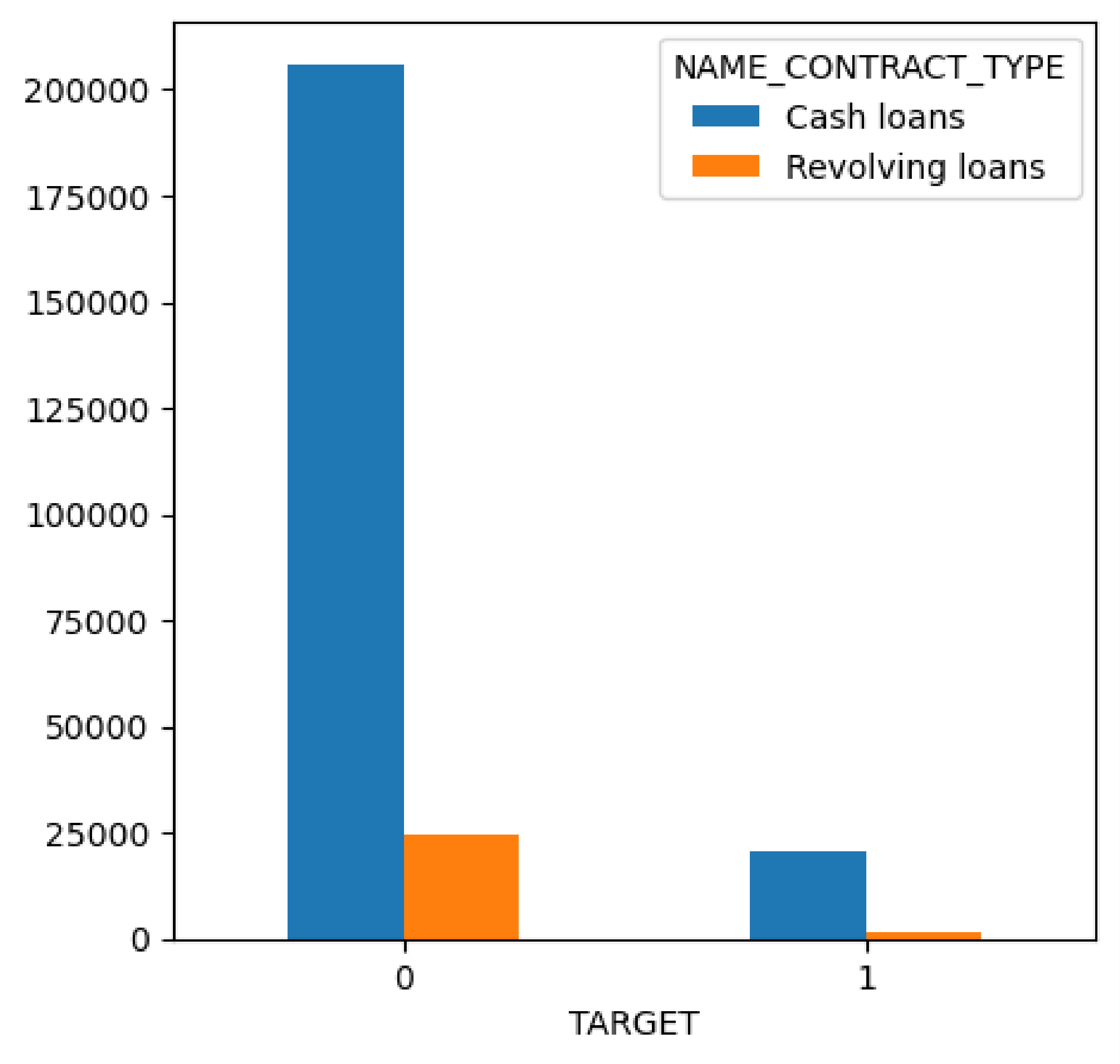
- Threshold Insight: Income is a key predictor of repayment capabilities, with those making above $500,000 having a notably lower default rate.



Analysis of Loan Types

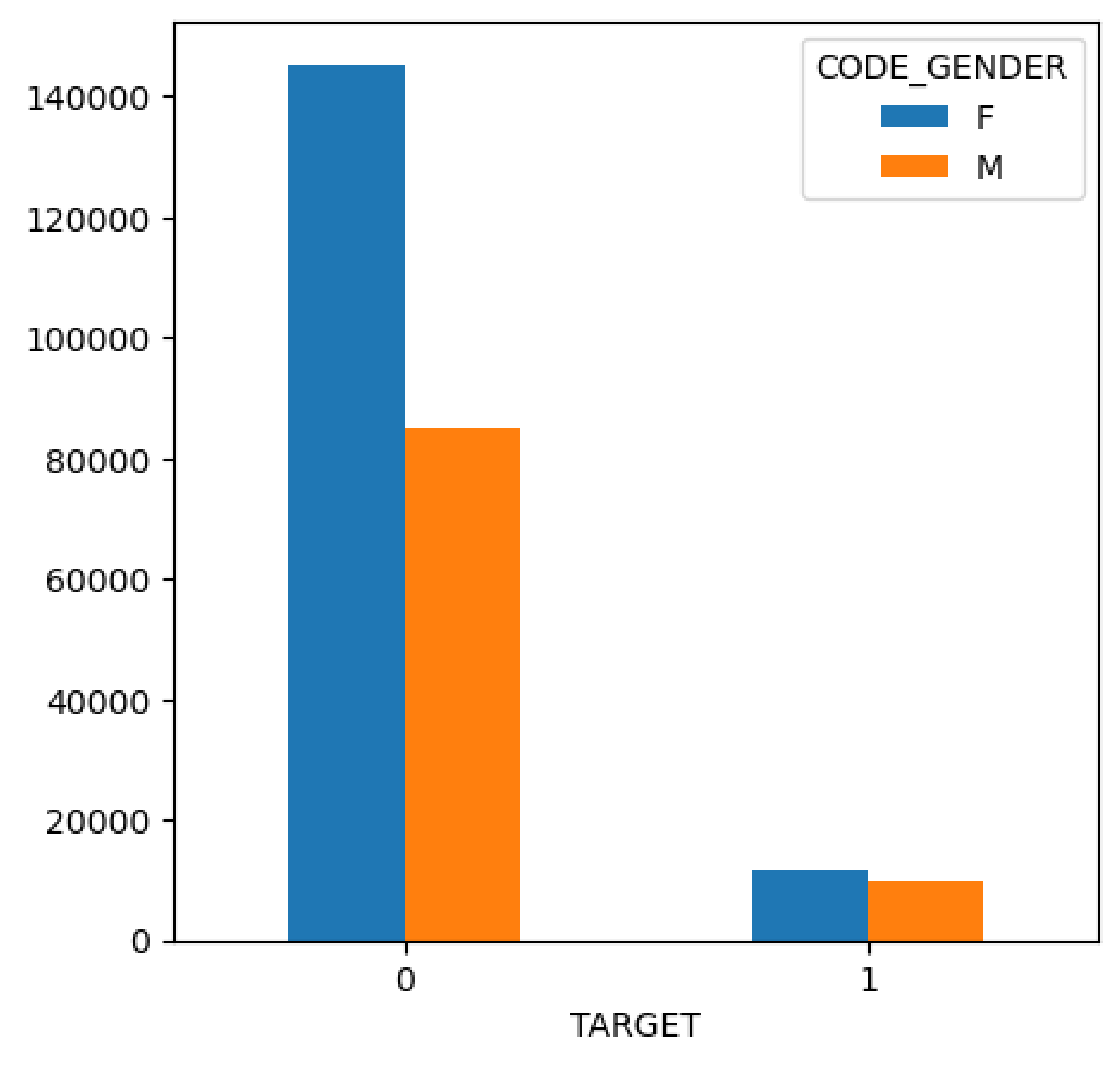
- Distribution of Loan Types: The bulk of loan applications in the dataset are for cash loans, which are the most prevalent type of loan.

- Default Rate by Loan Type: Compared to revolving loans, cash loans have a higher default rate, indicating a higher level of risk involved.



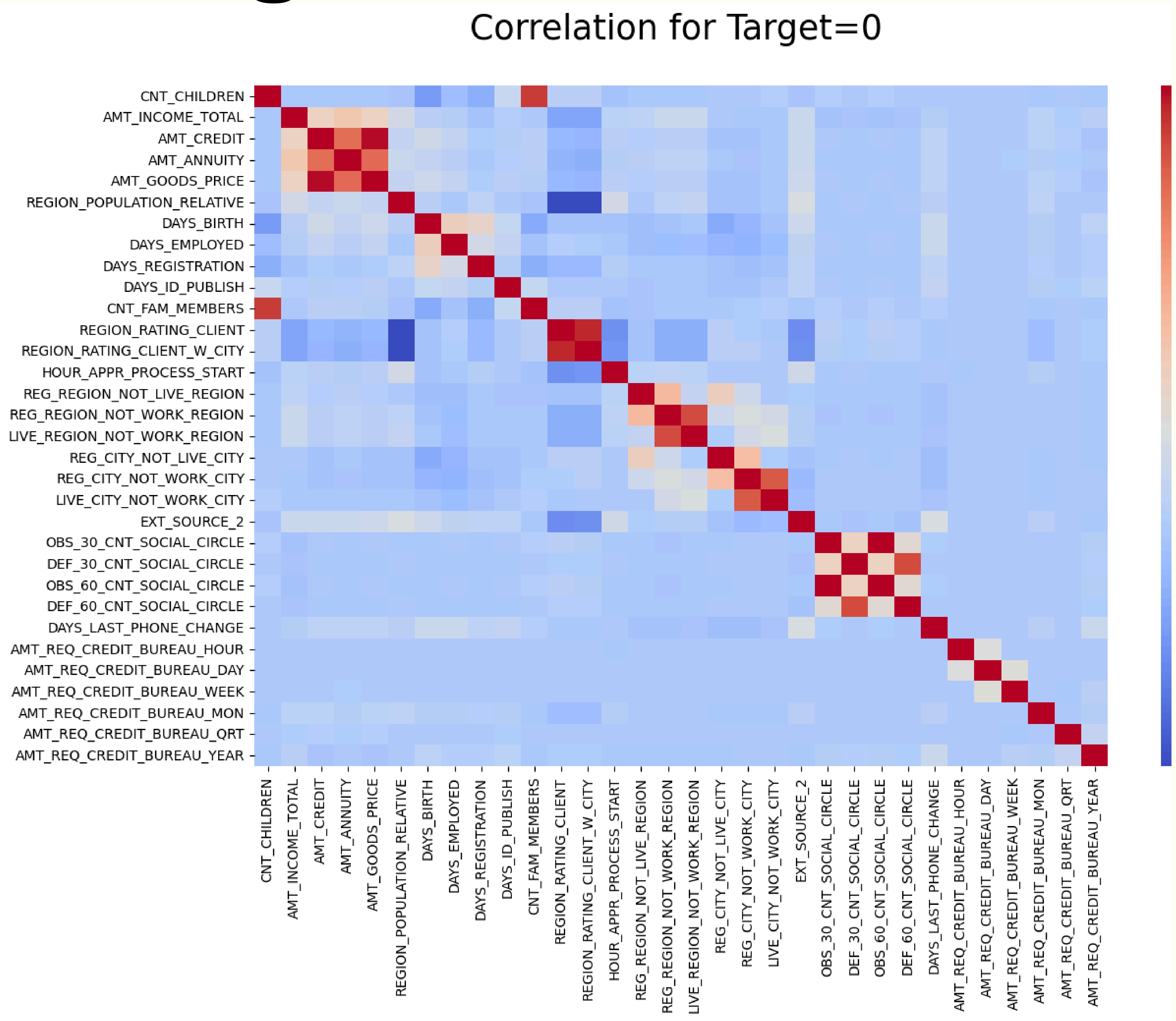
Loan Status and Gender

- Gender and Loan Application: Compared to their male counterparts, female applicants are more likely to obtain loans.
- Gender and Default Rates: The default rates for male and female applicants do not differ much, suggesting that gender has little bearing on loan repayment patterns.



Correlation Insights

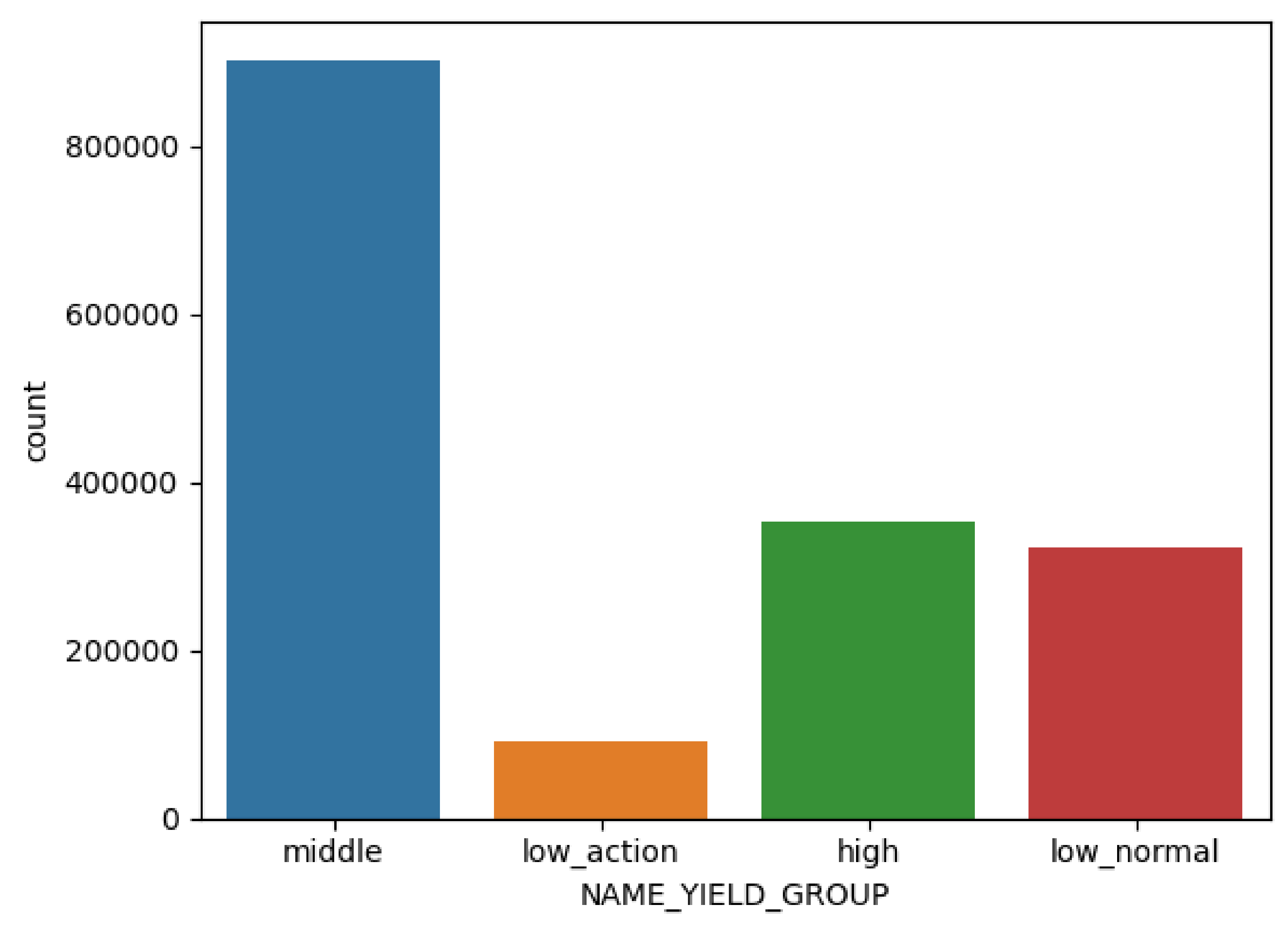
- The connection between CNT_CHILDREN and CNT_FAM_MEMBERS is strong (0.8).
- AMT_ANNUITY and AMT_CREDIT_Current: Strong association (0.7).



Previous application dataset

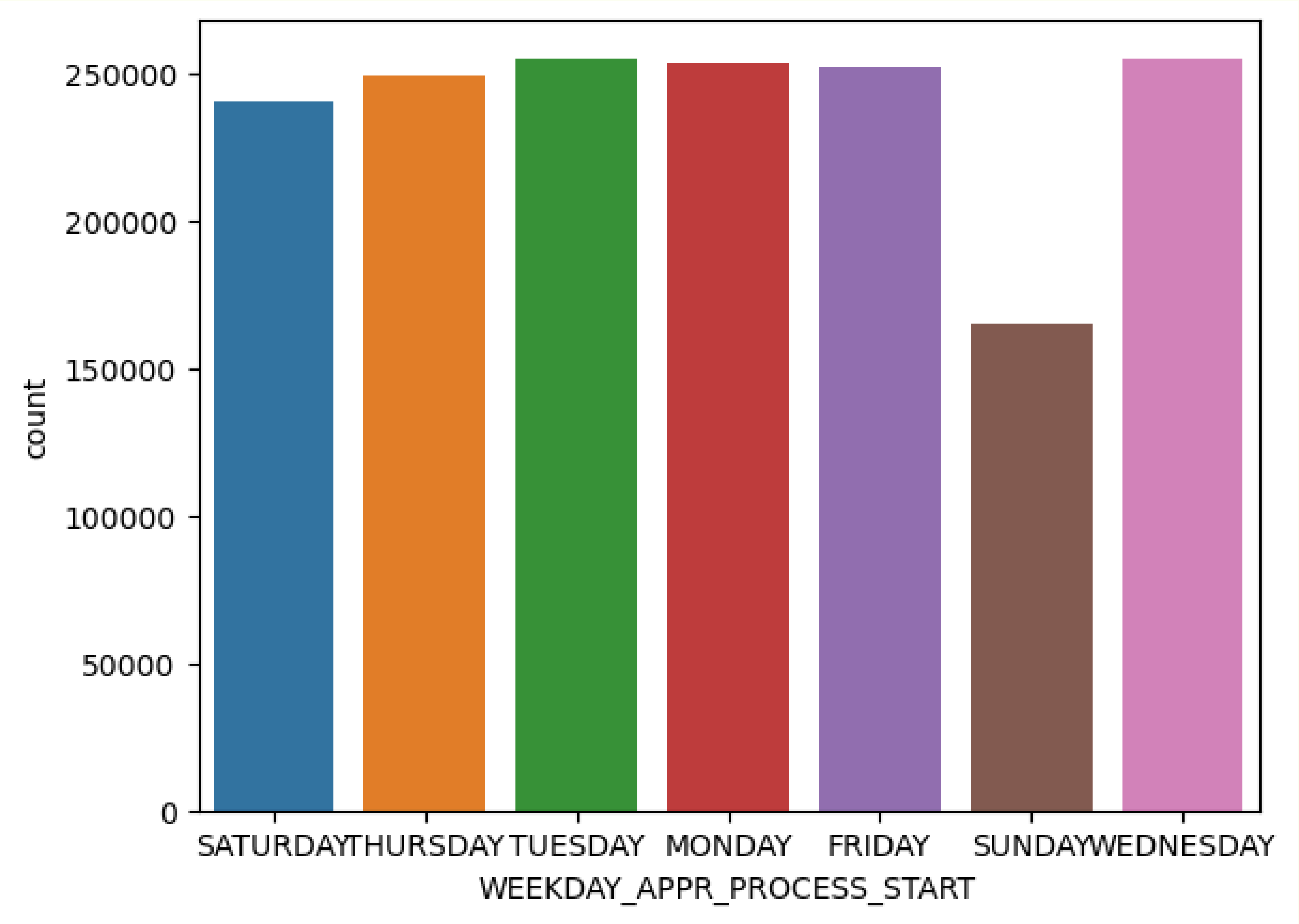
Yield group analysis

- The "middle" yield group includes the majority of loans, suggesting a preference for loan products with a moderate level of risk. The prevalence of the "high" and "low_normal" yield groups is balanced, however the "low_action" group is the least prevalent.



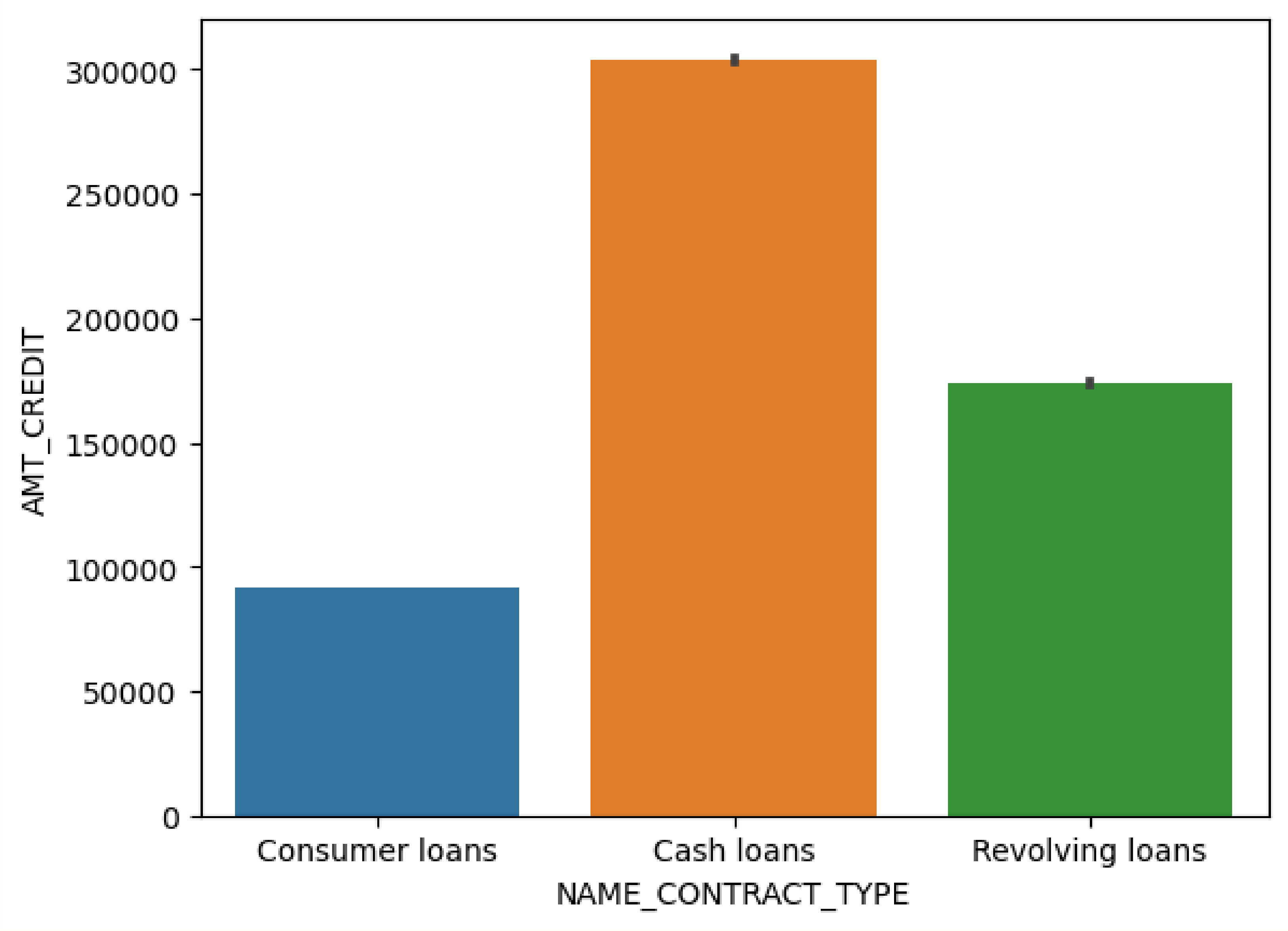
Loan application for the days of the Week

- Weekdays showed an increase in loan applications, especially in the middle of the week.
- Weekends witnessed a steep drop in applications, suggesting less activity.



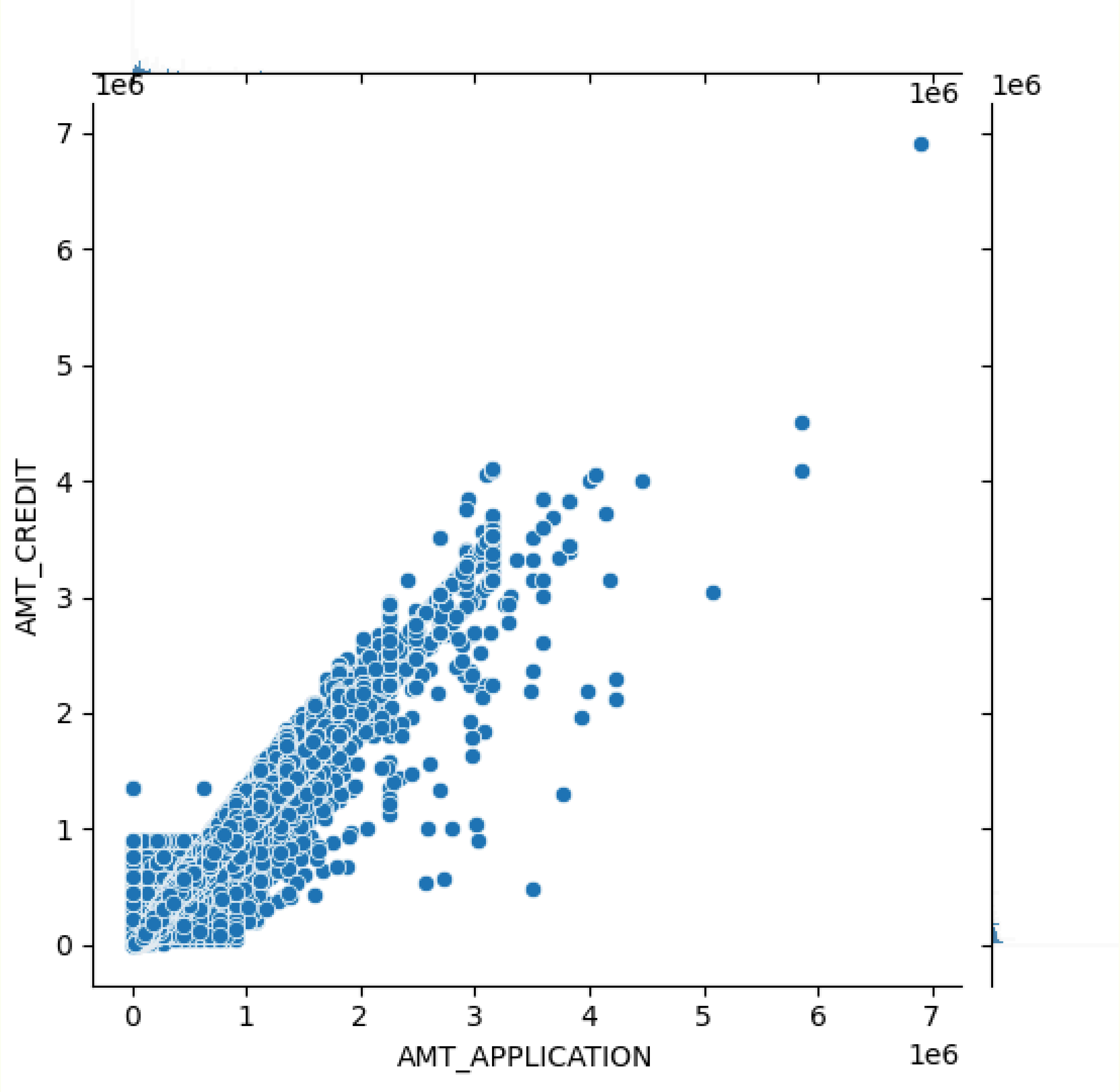
Details of the Contract Type

- Preference for Loans with Cash: Since cash loans give the biggest credit amount of the three categories, borrowers who require greater loan amounts are more inclined to choose them.
- Smaller-Needs Consumer Loans: Because consumer loans often have much lower loan amounts, they are intended for borrowers with modest financial needs.



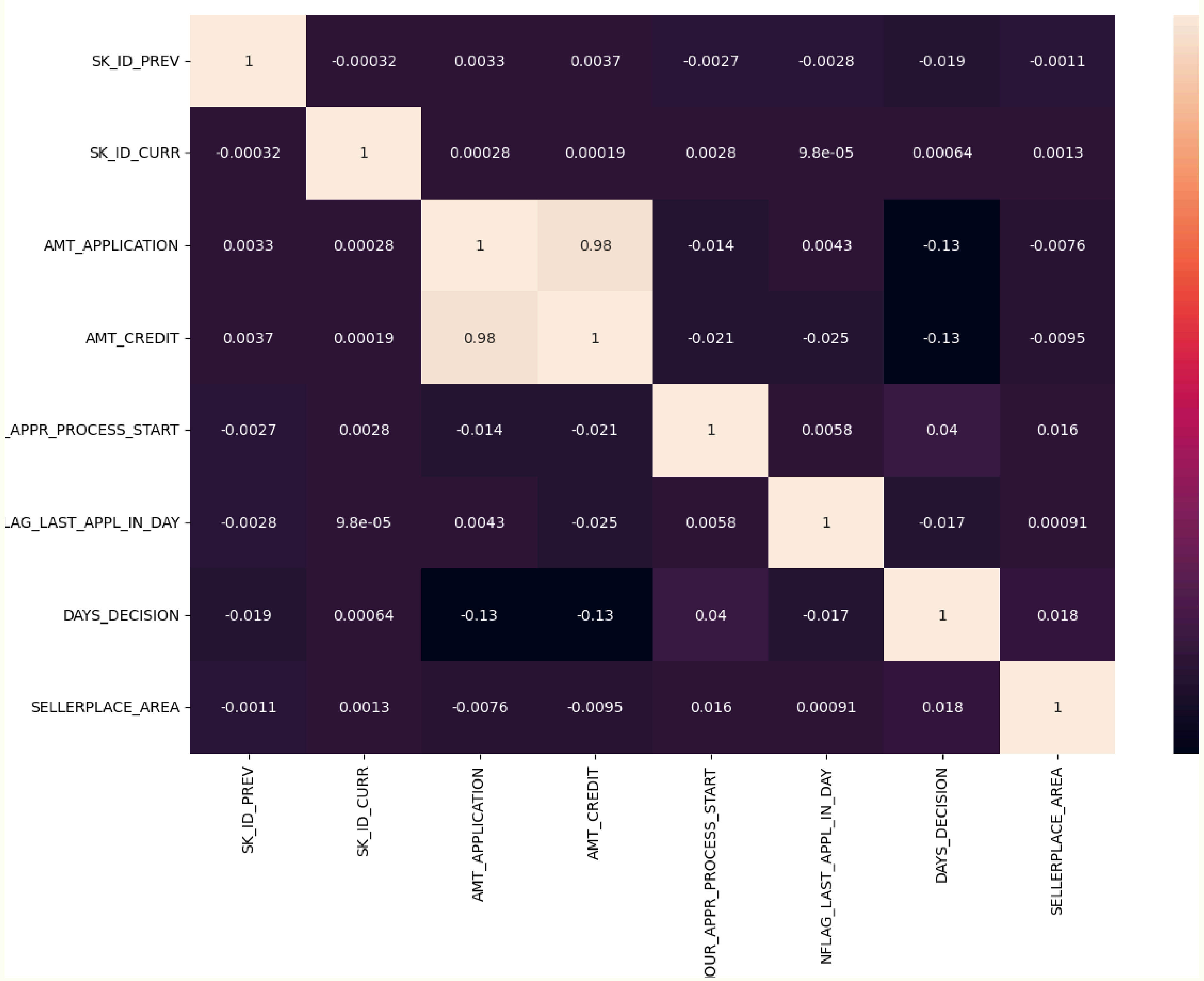
Analysis of Credit Amount

- A higher credit limit is associated with a higher chance of default.



Correlation insights

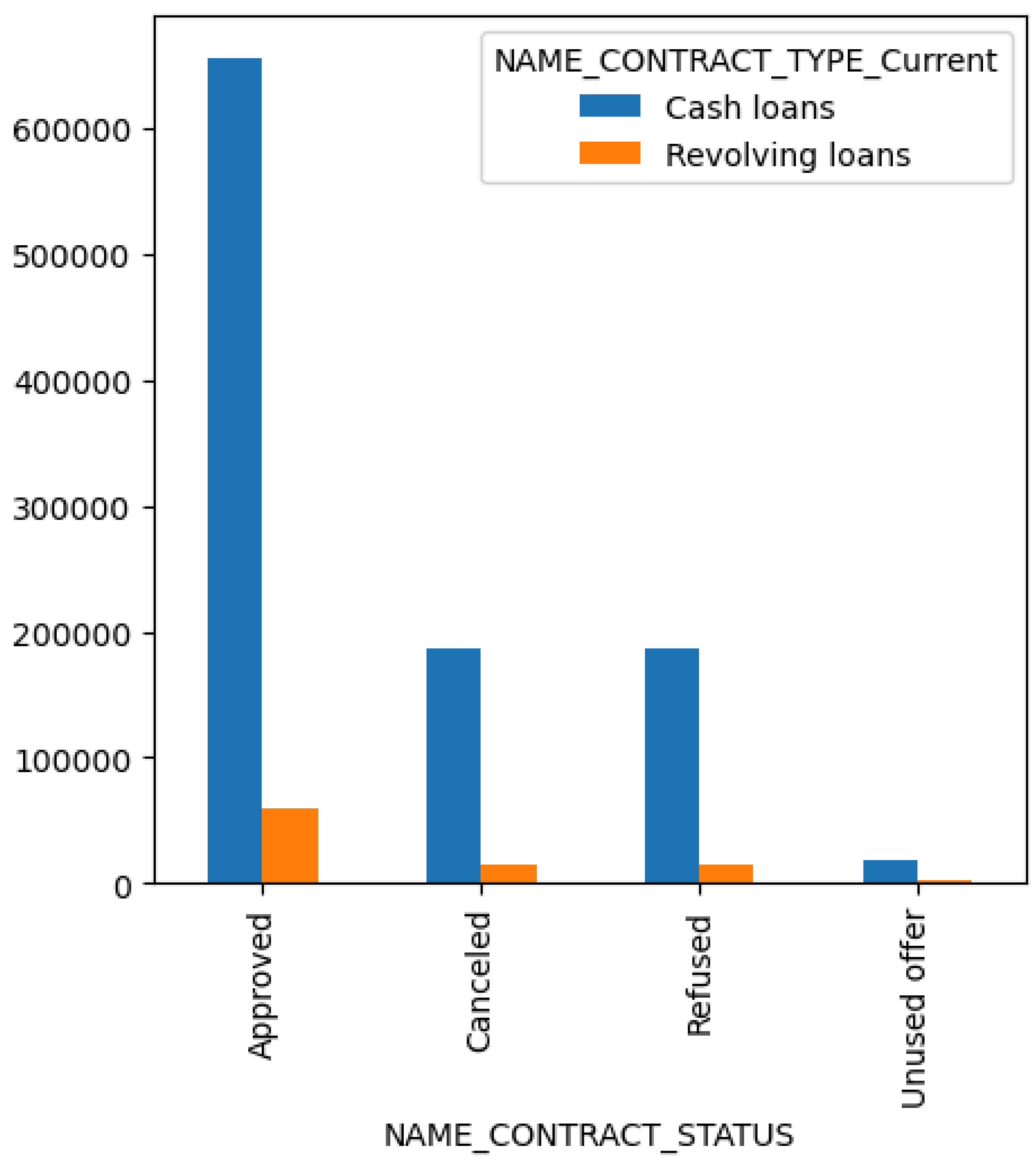
- AMT_APPLICATION and AMT_CREDIT, which have a very high positive correlation of 0.98.



Merging both the dataset

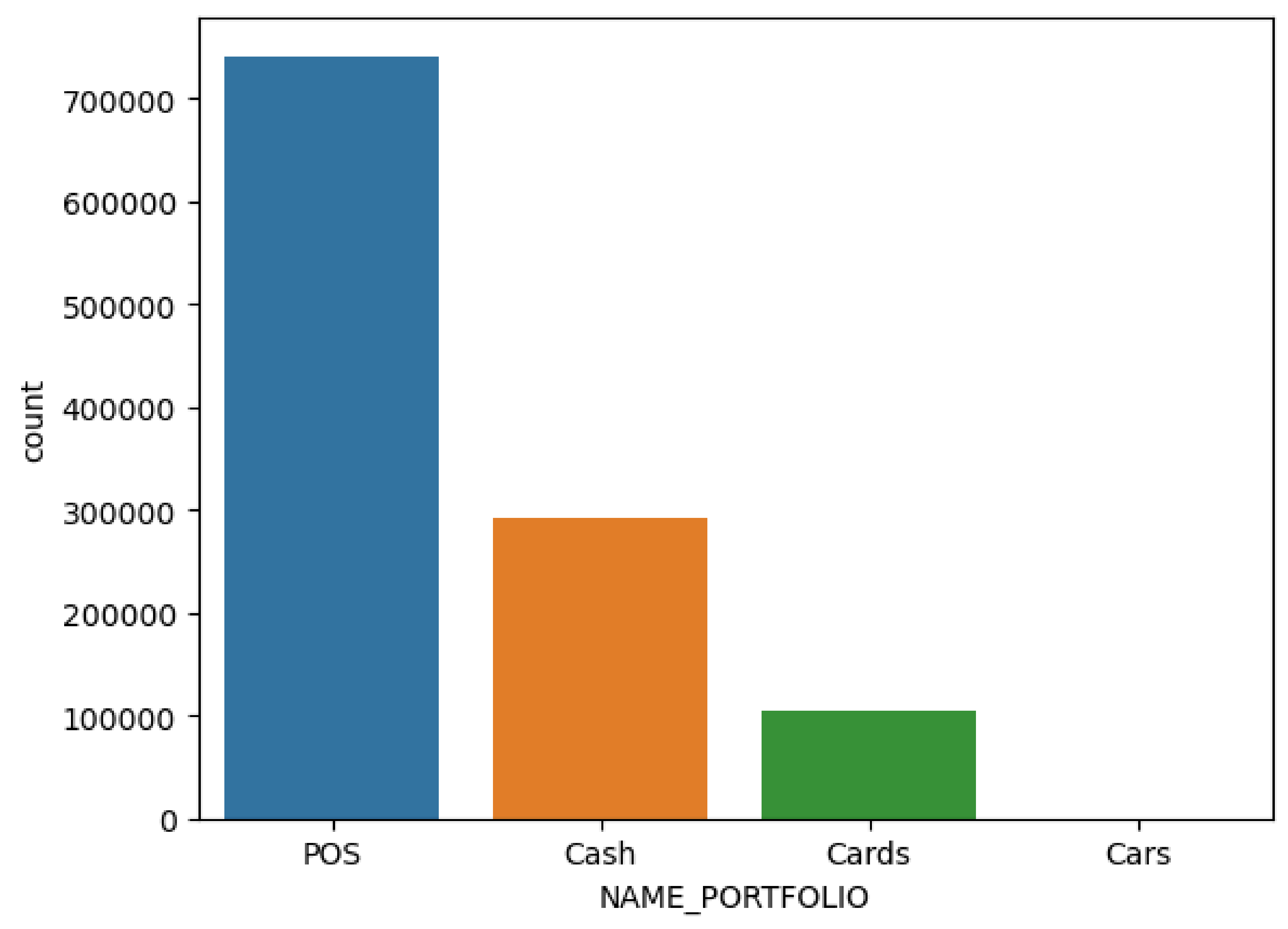
Analysis of Contract Types:

- Cash loans account for the majority of loans granted in both past and present applications.



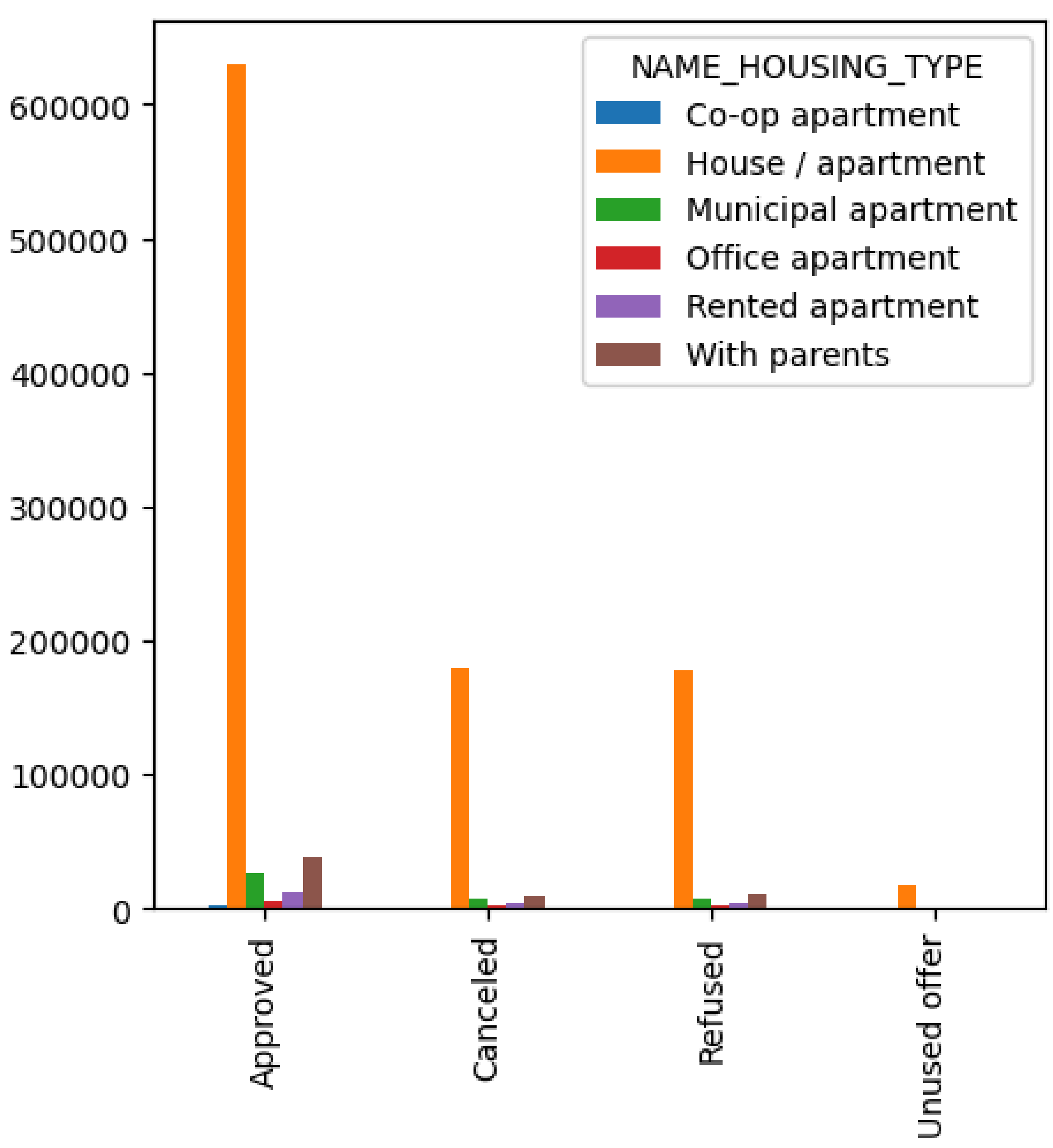
Portfolio Analysis

- The risk profile of the POS portfolio type differs from those of other portfolio types.



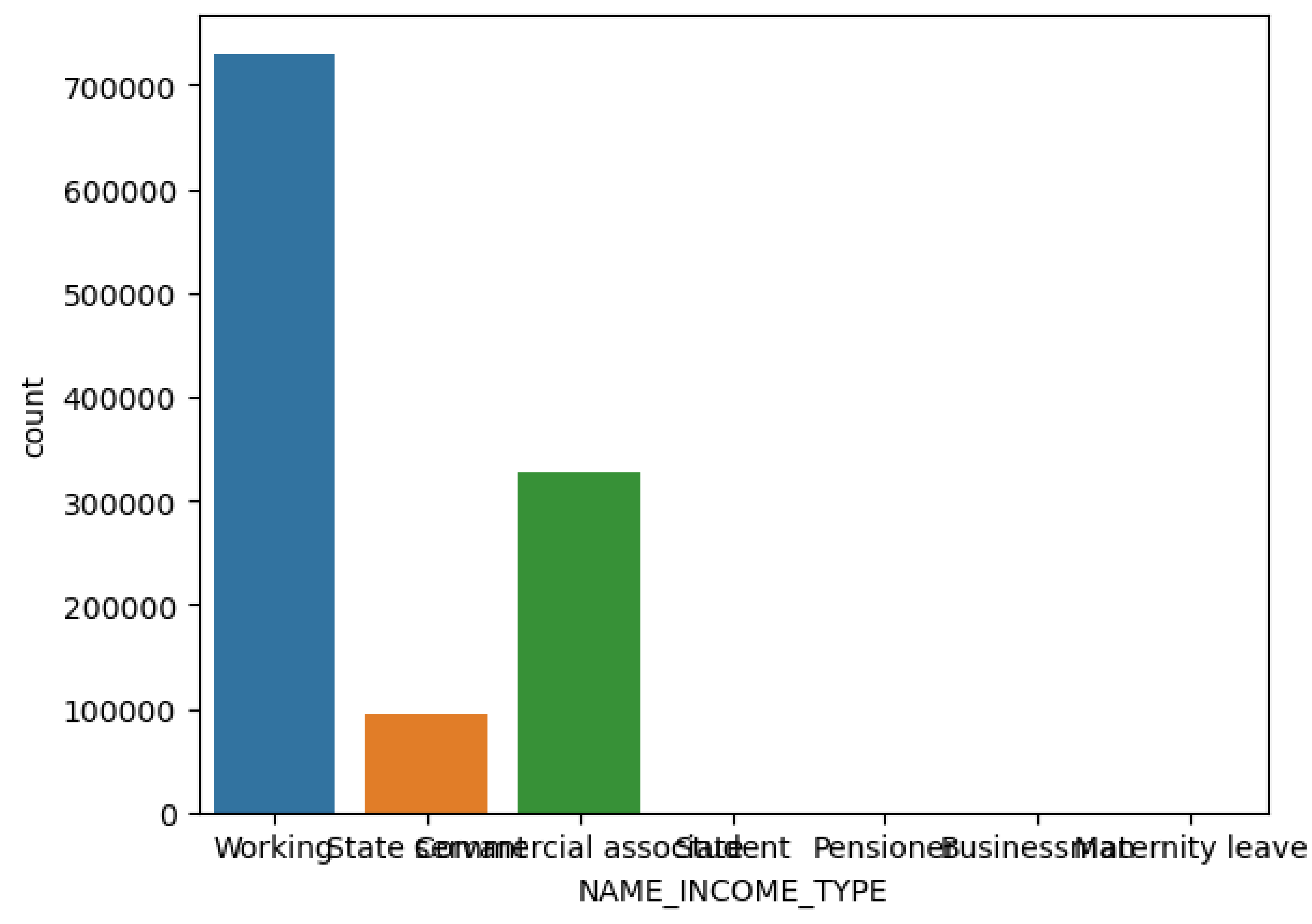
Housing and loan status

- Loan approval rates are higher and default rates are lower for applicants who have solid housing, such as home or flat owners.



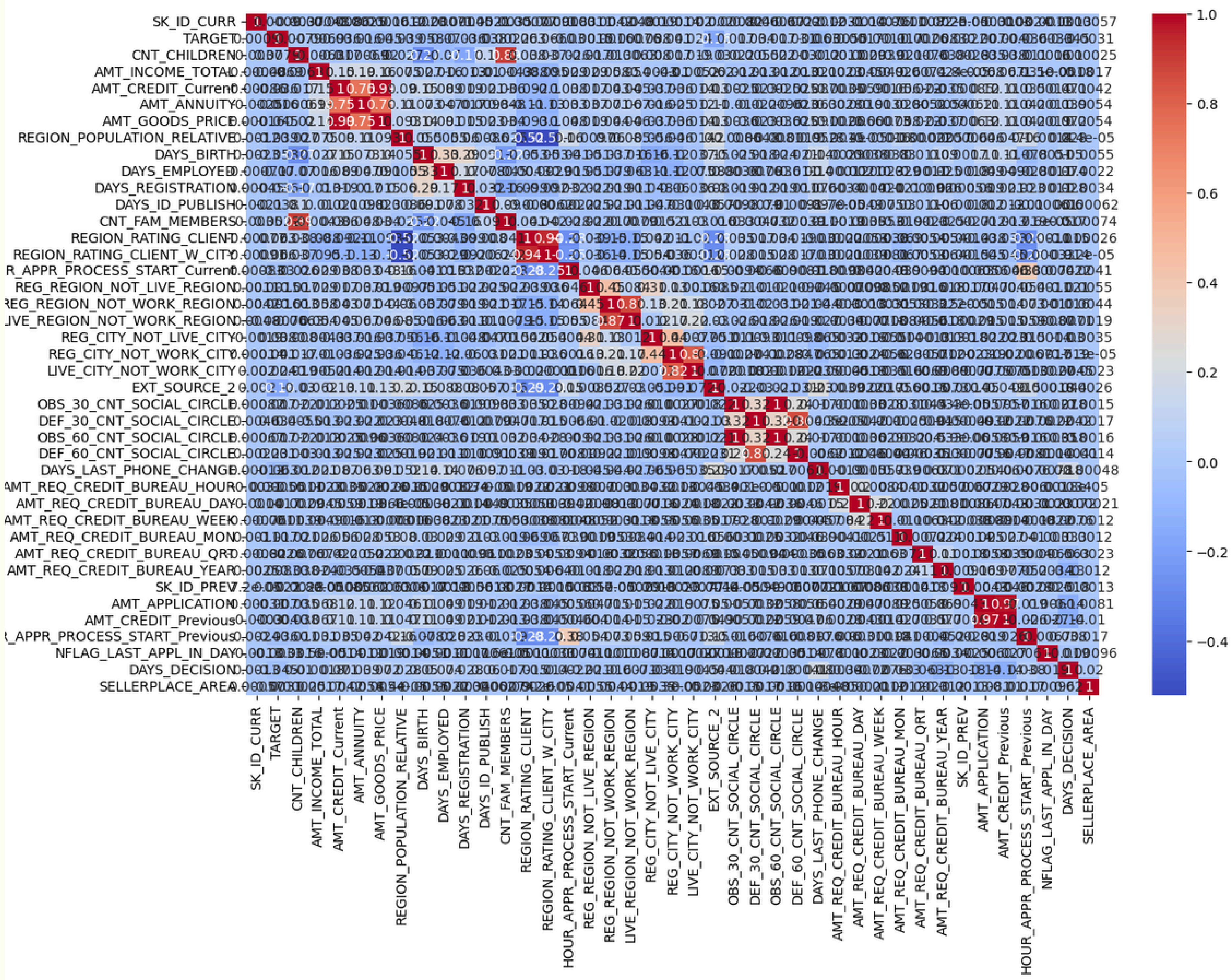
Housing and loan status

- Compared to applicants with other income sources, employed applicants have a balanced risk profile while taking out more loans.



Correlation for mergerd data

- There is a very good association (0.9) between REGION_RATING_CLIENT and REGION_RATING_CLIENT_W_CITY.
- There is a strong correlation (0.8) between DEF_30_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE.
- There is an exceptionally high correlation (0.9) between OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE.



Conclusion

- One important predictor is income: Income level is a crucial consideration when evaluating loan risk because applicants with higher incomes have far lower default rates.

- Loan Types and Risks: Compared to revolving loans, cash loans have a higher default rate despite being the most popular. To reduce risks, cash borrowing tactics must be customized.

- Impact of Demographics: While family status and housing type (e.g., married applicants or house/apartment owners) positively correlate with better repayment behavior, gender has no discernible effect on default risk.

- Outliers in Credit and Annuity Amounts: High-value loan amounts indicate heightened risk, underscoring the need for more investigation in these situations.

Conclusion

- Risk is Informed by Historical Trends: Examining prior loan applications can show recurrent default trends in particular applicant groups, which can inform lending decisions today.

- Age and Employment Tenure Correlation: Older applicants with longer work histories are more likely to make repayments, highlighting the importance of giving these aspects careful consideration when assessing risk.

- Portfolio Insights: POS portfolios show unique patterns that can necessitate tailored assessment procedures to take particular risk profiles into consideration.

- Actionable Suggestions: To maximize loan approvals and reduce defaults, apply EDA findings, tighten screening, divide applicants into risk categories, and balance data imbalances.

THANK YOU