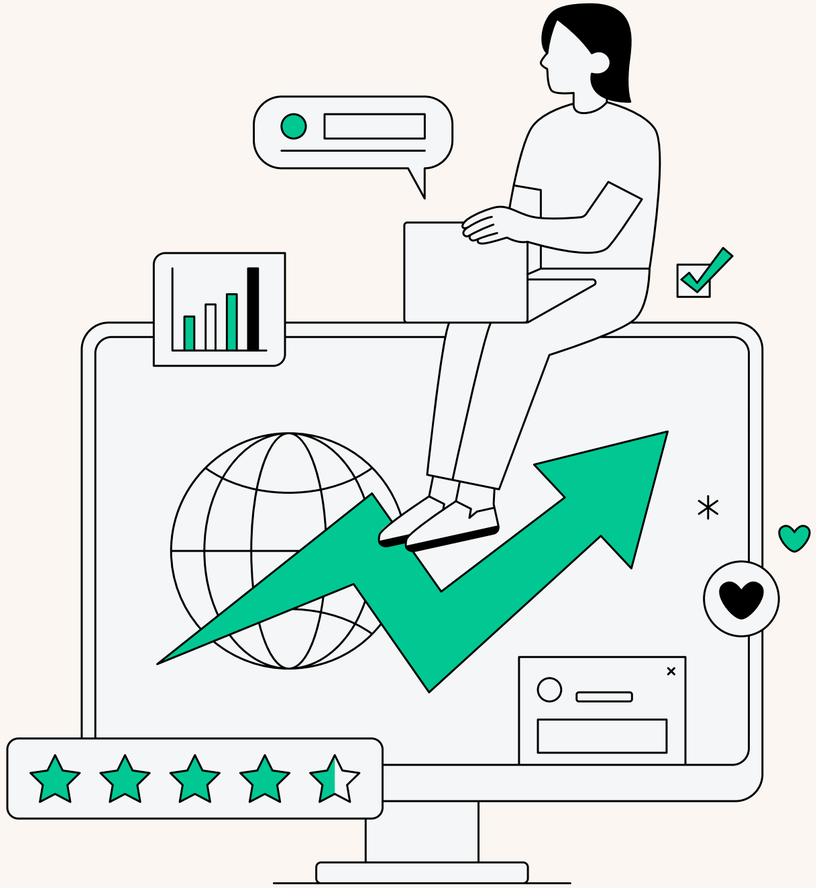
Presented by Aafreen Moyal

Credit Risk Analysis Results



Problem Statement

Objective:

- Analyze the dataset to identify factors contributing to loan defaults.
- Provide insights to assess risks associated with loan lending.

Purpose:

- Understand the reasons behind loan defaults.
- Help in reducing risks for future loan lending.

Approach:

- Perform Exploratory Data Analysis (EDA) to explore and understand the dataset.
- Extract meaningful insights to support better decision-making for future business strategies.

Outcome:

- Gain actionable insights to refine loan approval criteria.
- Minimize default risks and improve overall lending efficiency.

Dataset Description

01.

‘application_data.csv' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.

02.

'previous_application.csv' contains information about the client’s previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.

03.

'columns_description.csv' is data dictionary which describes the meaning of the variables.

Steps Performed

Step 1: Loading the Datasets

Step 2: Merging the Datasets

Step 3: Data Cleaning

Step 4: Insights from Data Imbalance

Step 5: Exploratory Data Analysis (EDA)

- Categorical Data Analysis
- Numerical Data Analysis
- Outlier Analysis
- Significant Columns with Target Variation
- Bivariate Analysis

Step 6: Conclusion

Step 1: Loading the Data & Step 2: Merging the Datasets

1. Imported necessary modules.
2. Loaded and explored the following datasets:
    a. previous_application: 1,670,214 entries, 37 columns.
    b. application_data: 307,511 entries, 122 columns.
3. Merged both datasets on SK_ID_CURR, resulting in:
    a. merged dataset: 1,413,701 entries, 158 columns.
4. Data Preview
    a. Checked dataset structure, data types, and column details.

Step 3: Data Cleaning

1. Columns with >50% Null Values: Removed to improve data quality.
2. Duplicates: Eliminated to ensure data integrity.
3. Observations After Cleaning
    a. Cleaned dataset size: Reduced columns while maintaining rows with valid data.
4. Converted DAYS_BIRTH (age in negative days) into positive years for better understanding.
5. Converted 'DAYS_REGISTRATION' ,'DAYS_TERMINATION','DAYS_FIRST_DRAWING', 'DAYS_ID_PUBLISH' into years.

Step 4: Insights from Data Imbalance

1. Target Variable Analysis:
    a. Target Variable Definition:
        - 1: Loan default (client had difficulty paying).
        - 0: No payment issues.
2. Imbalance Observation:
    a. Ratio of non-defaulters (0) to defaulters (1): 10.551.
3. Implications
    a. **The dataset is highly imbalanced.**

Step 5: Exploratory Data Analysis (EDA)

Analysing Categorical Data

Subset Creation:

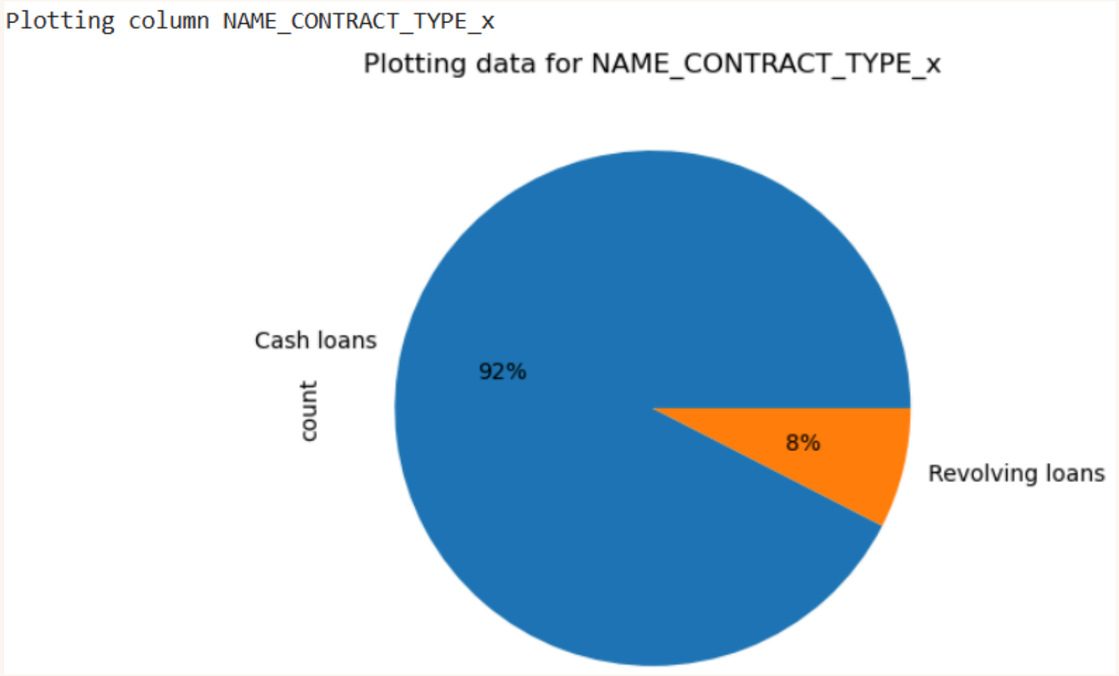
- target_0: Records where TARGET=0 (no payment difficulties).
- target_1: Records where TARGET=1 (payment difficulties).

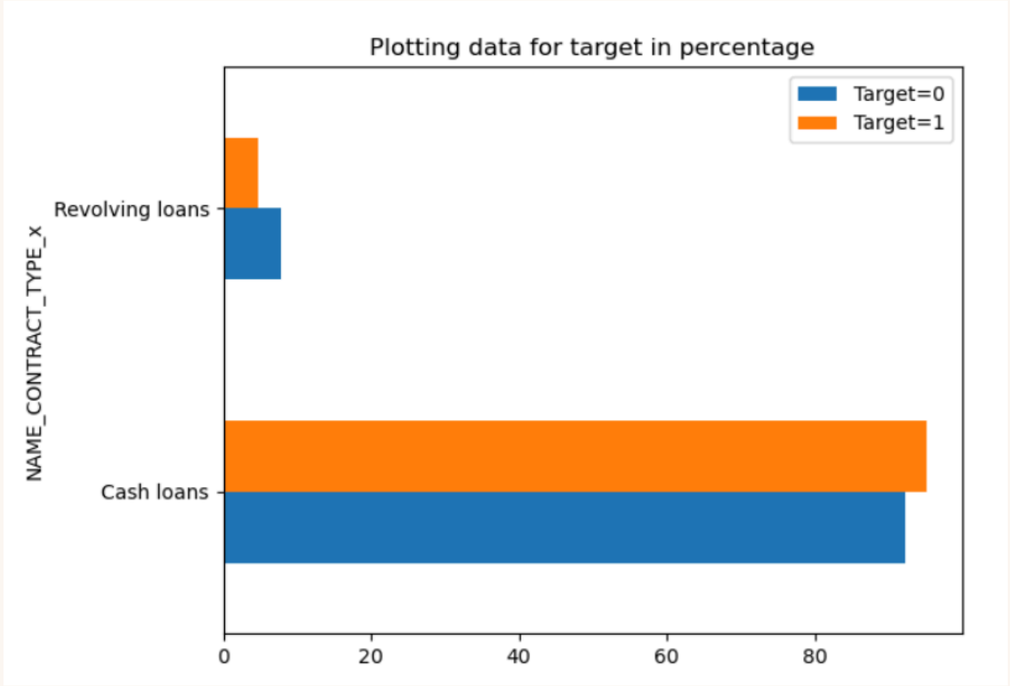
Columns selected for analysis: (as the dataset is quite large we selected columns which has least null values)

'NAME_CONTRACT_TYPE_x', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'NAME_TYPE_SUITE_x', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE', 'OCCUPATION_TYPE', 'WEEKDAY_APPR_PROCESS_START_x', 'ORGANIZATION_TYPE','EMERGENCYSTATE_MODE'

Univariate analysis of the categorical columns

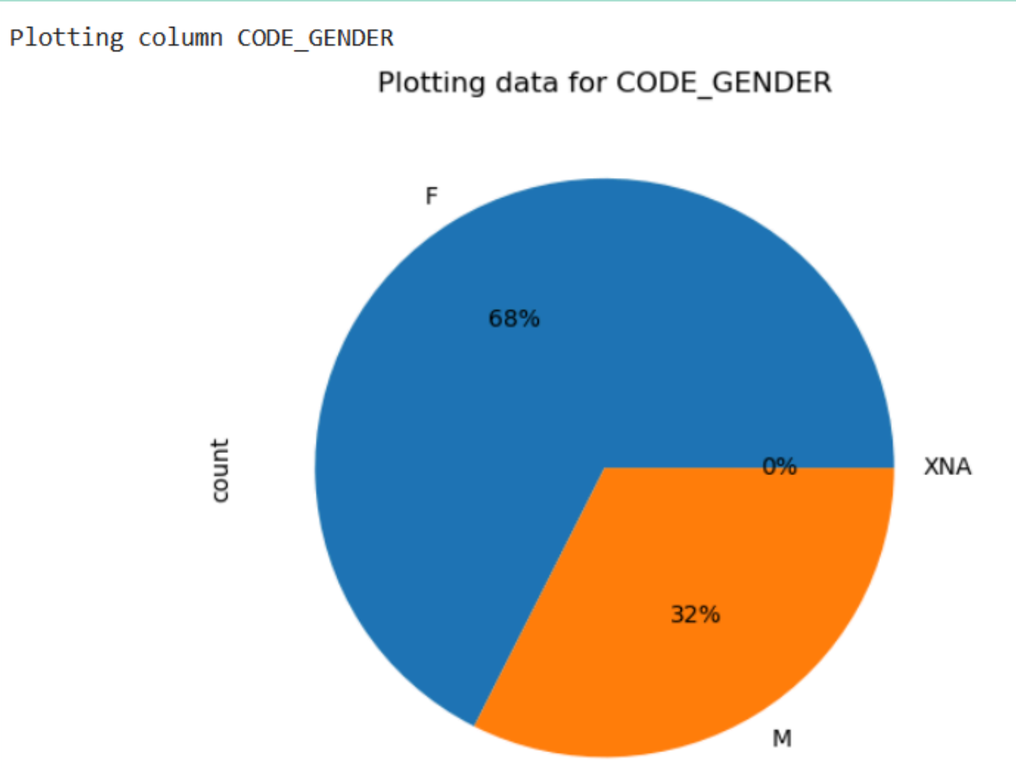
1. Loan Type

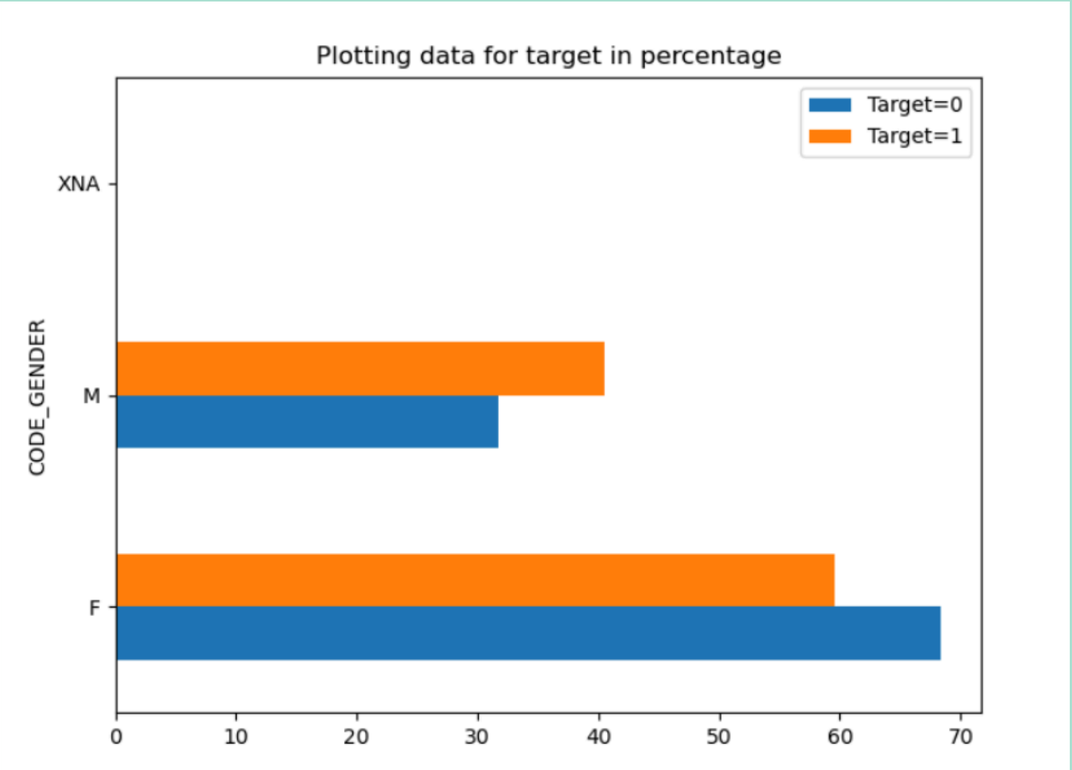




- Majority opt for cash loans over revolving loans due to their simplicity and accessibility.
- People with cash loans often face greater difficulties in repayment compared to those with revolving credit

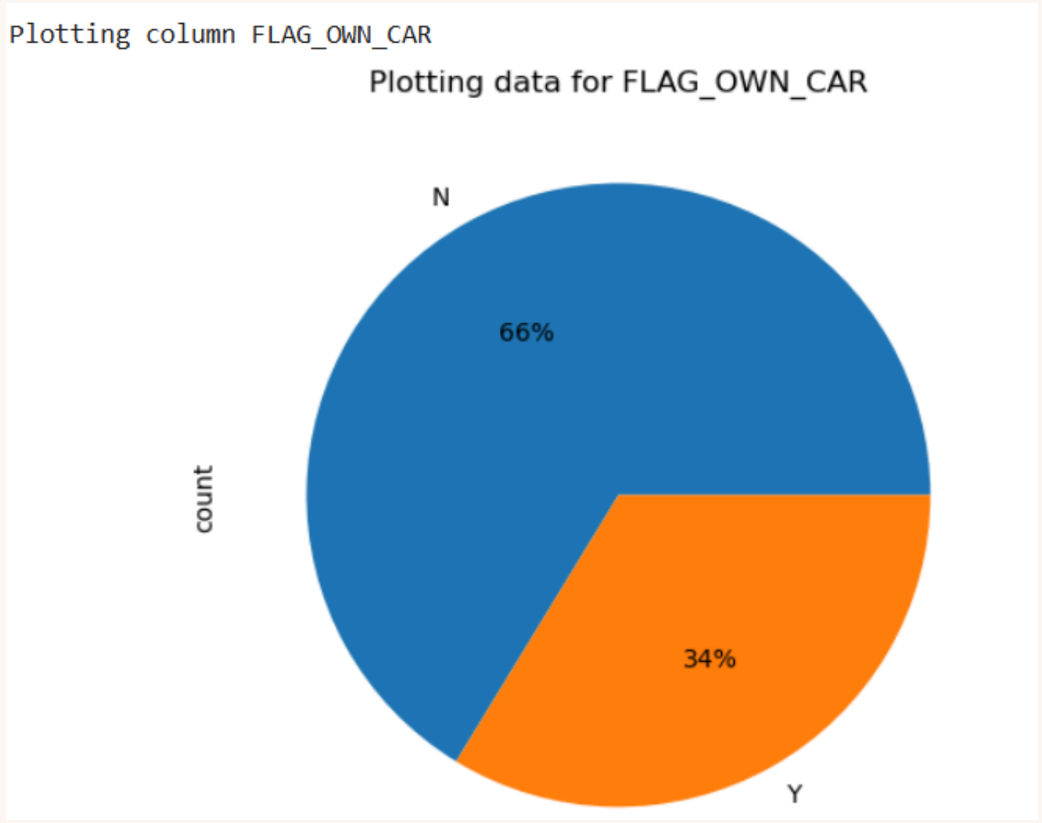
2. Gender

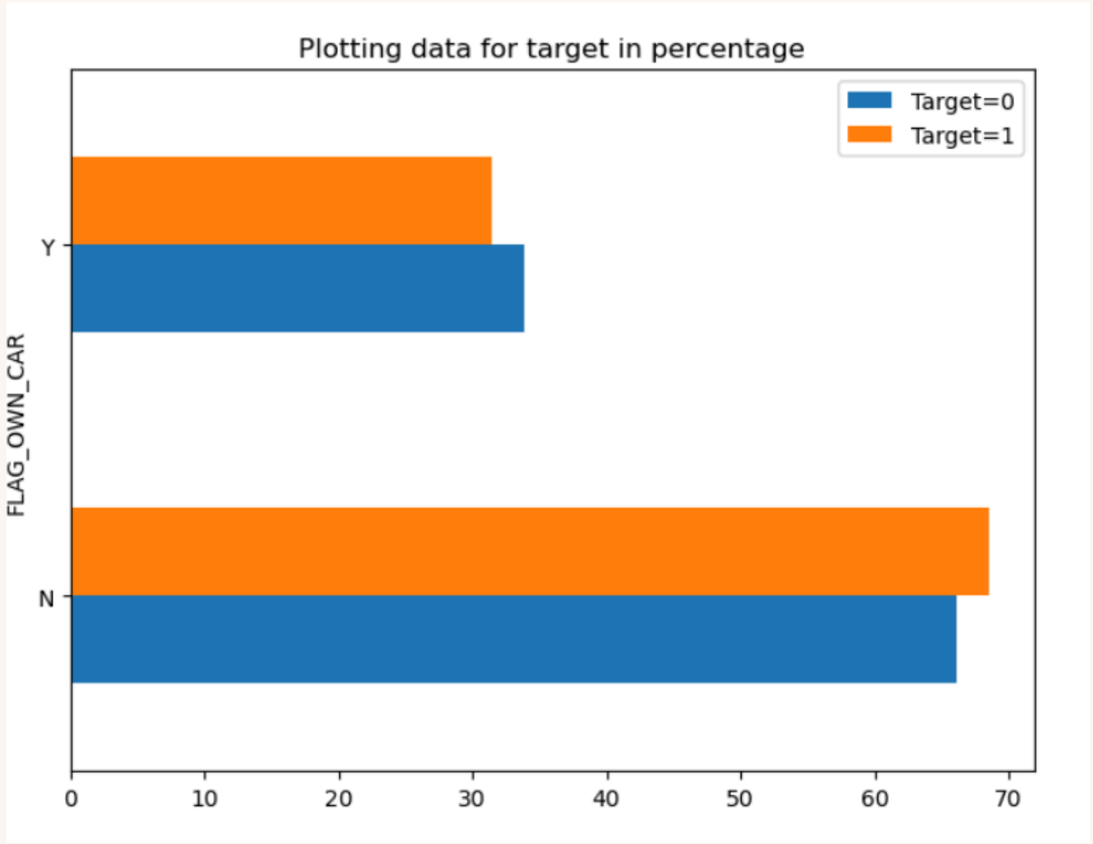




- Women apply for loans more than men.
- Women are more likely to repay loans on time, reflecting financial responsibility. Male defaulters is significantly higher.

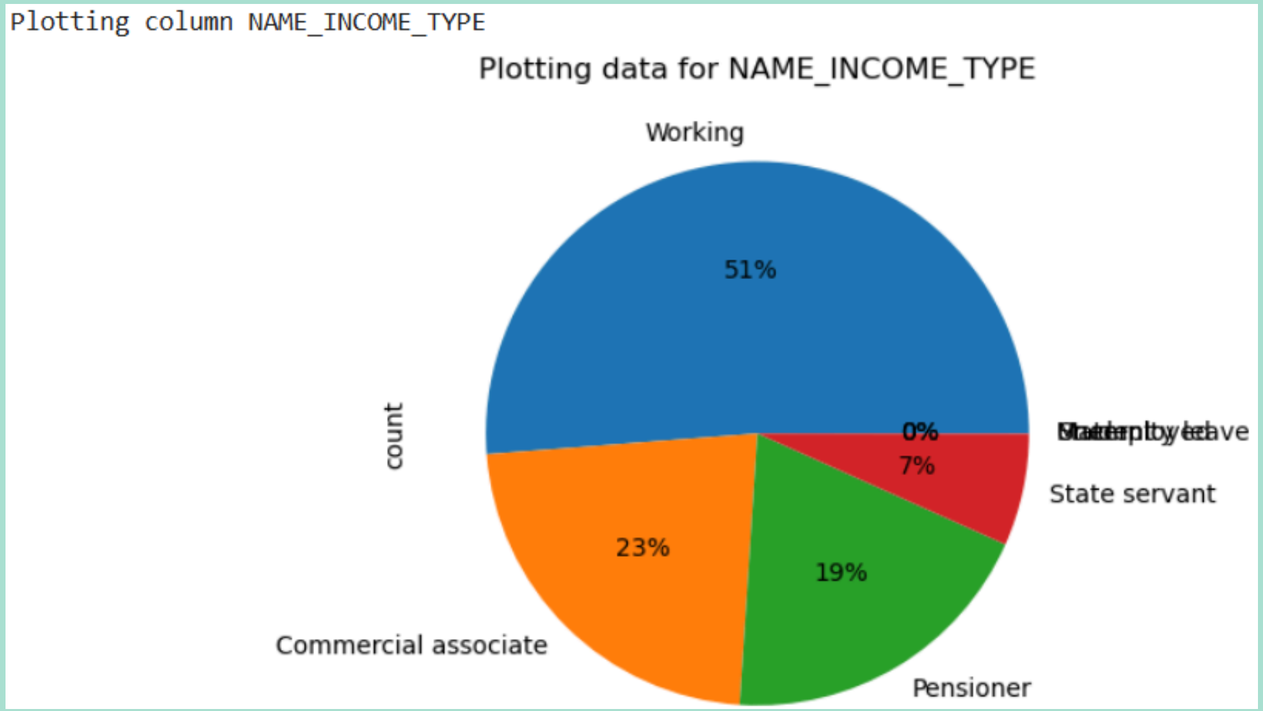
3. Own car or not

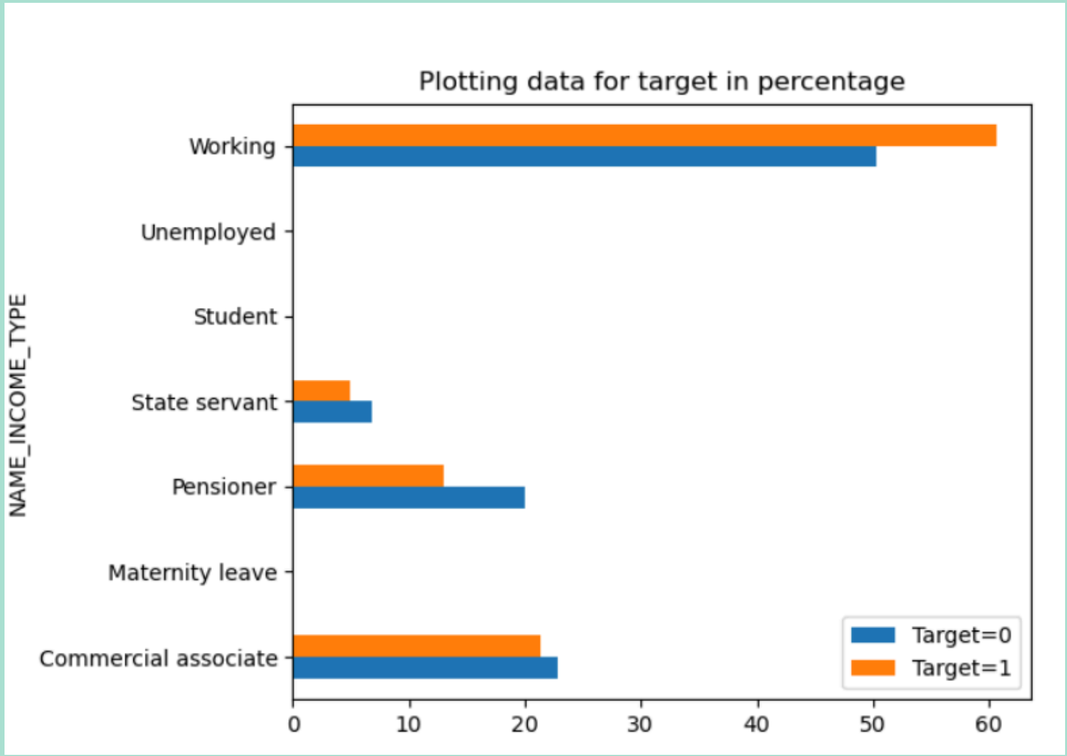




- People without cars apply for loans more often.
- People without car face difficulties while repaying, than people with car.

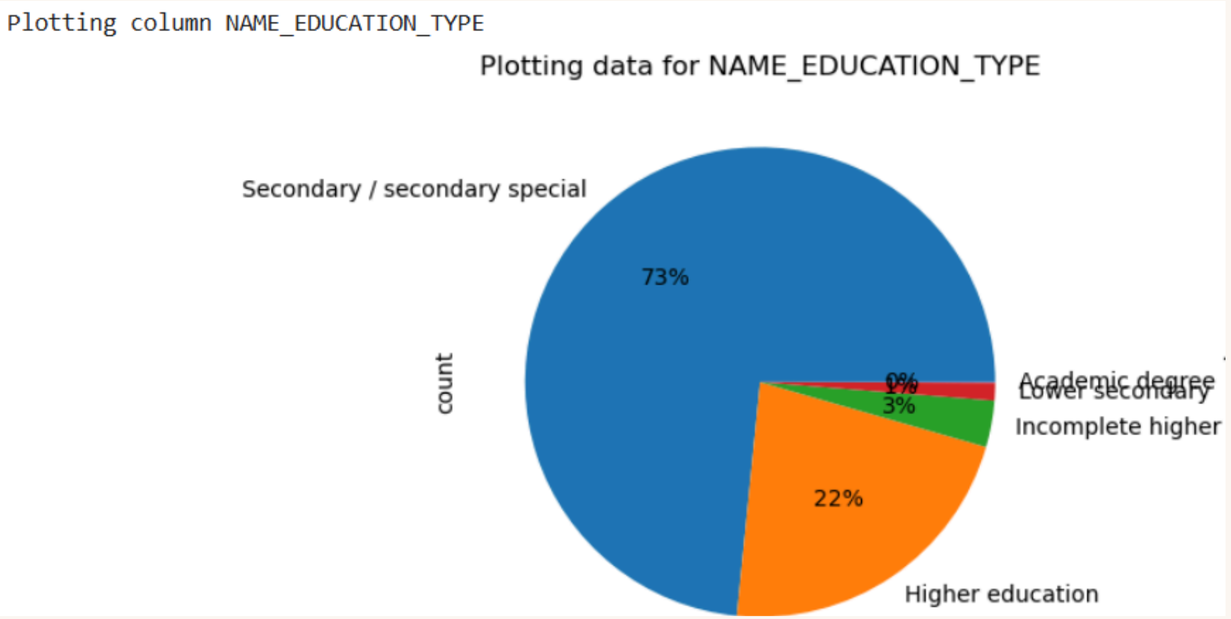
4. Income Type

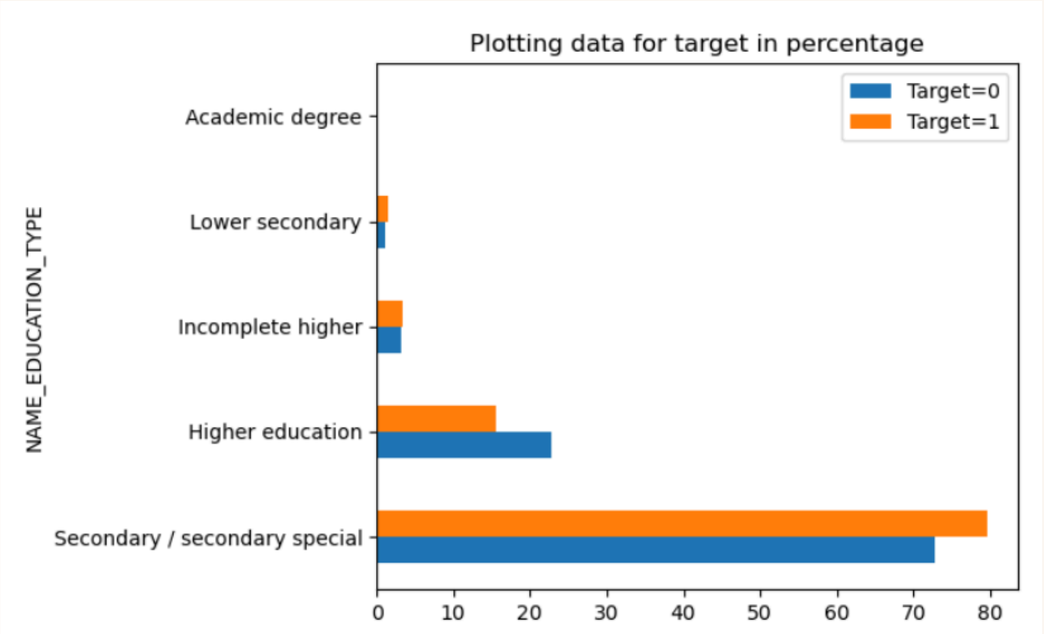




- Working class apply for loan more than any other Income type field.
- Pensioners have a lower default rate, showing that even with a fixed or smaller income, they are more reliable in repaying loans.

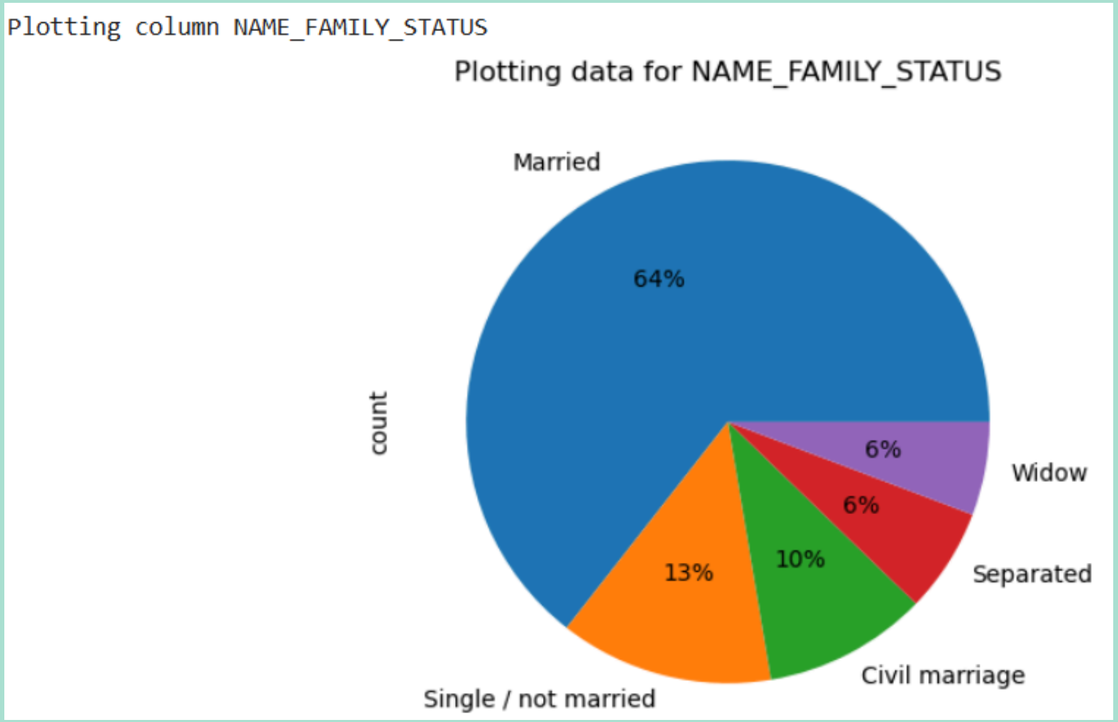
5. Education Type

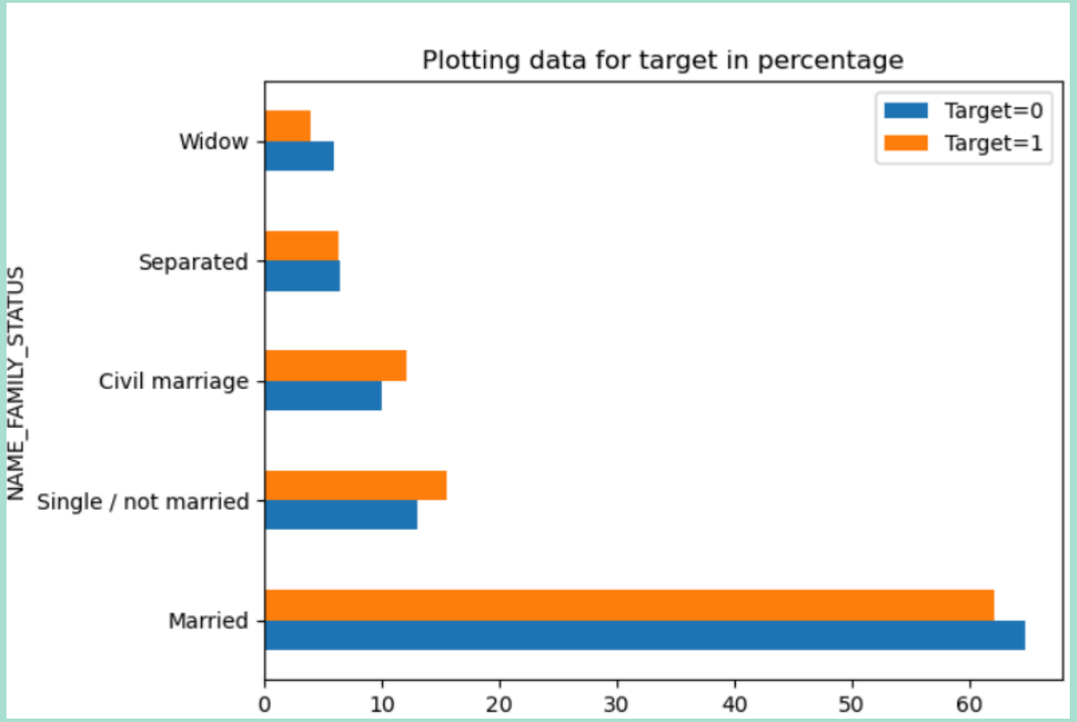




- Most loans are taken by people with secondary education, but their default rate is much higher compared to those with higher education. This indicates that education level plays a key role in loan repayment ability.

6. Family Status

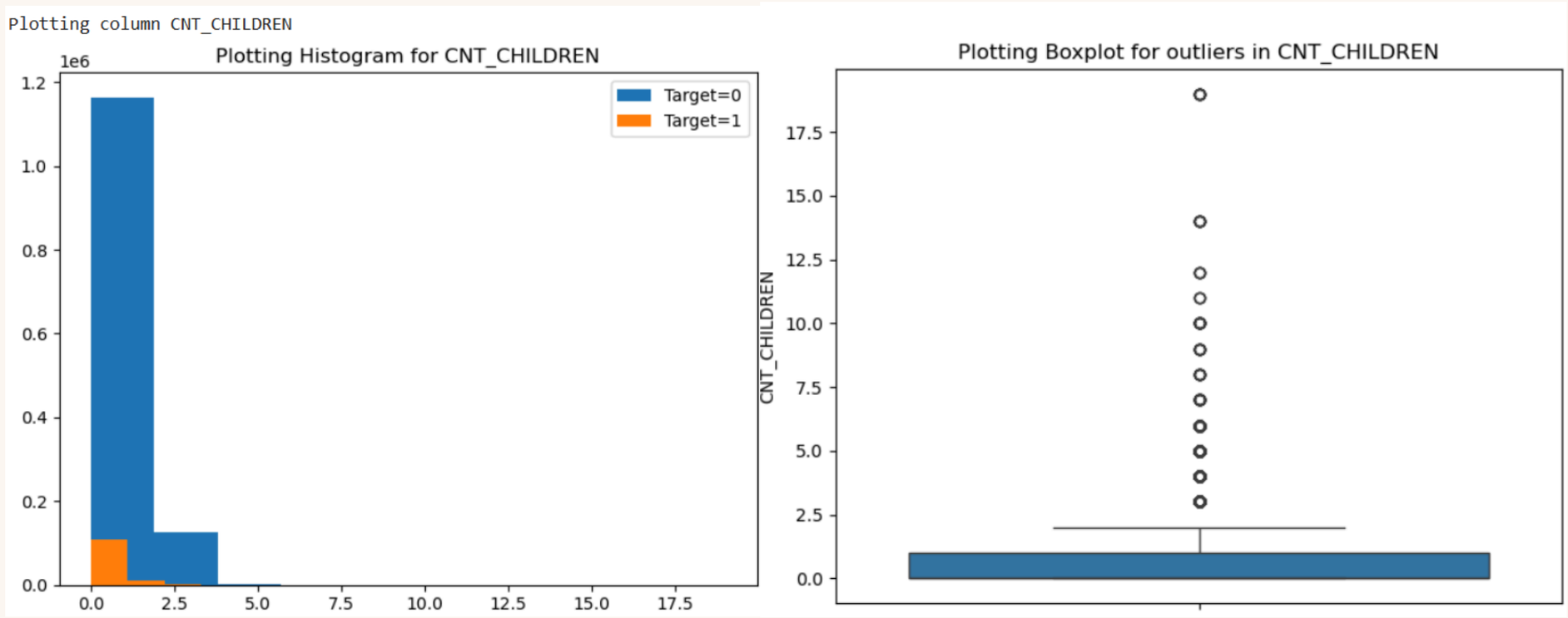




- Married people apply for loans the most and are less likely to default. On the other hand, singles and those in civil marriages have higher default rates, suggesting marital status affects repayment behavior.

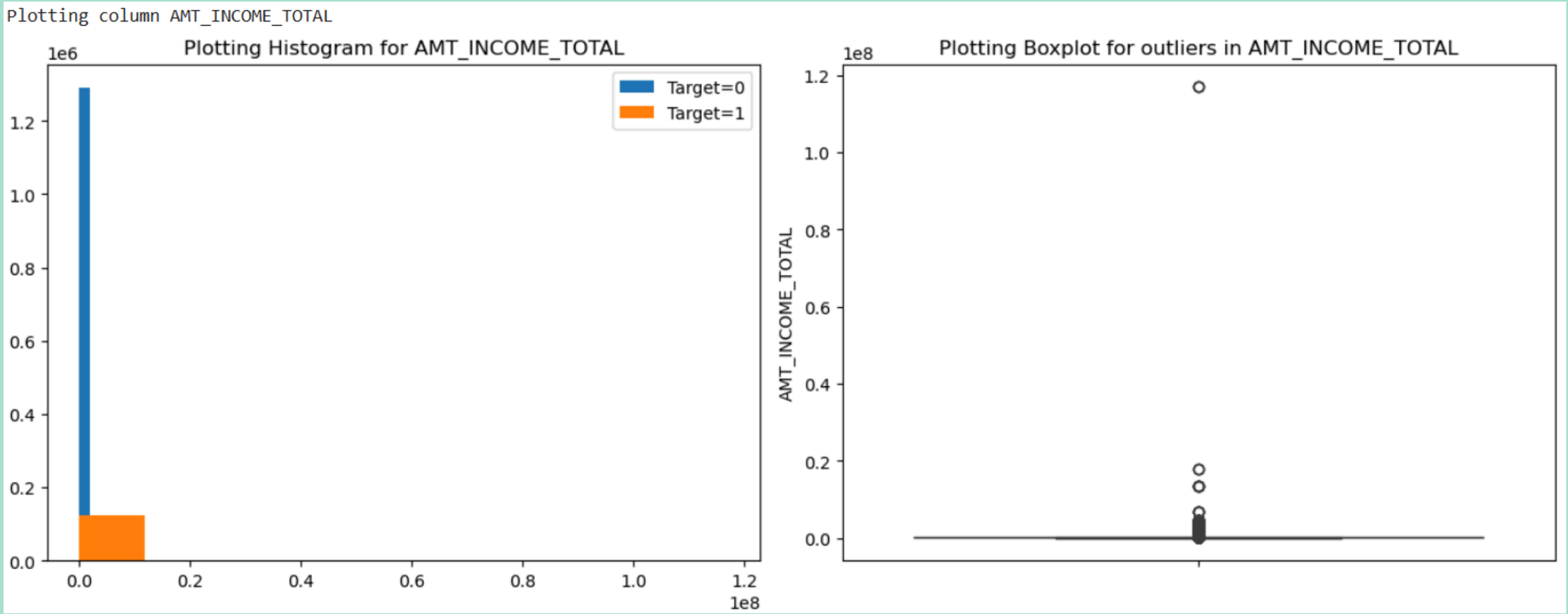
Univariate analysis of the numerical columns

1. Number of children



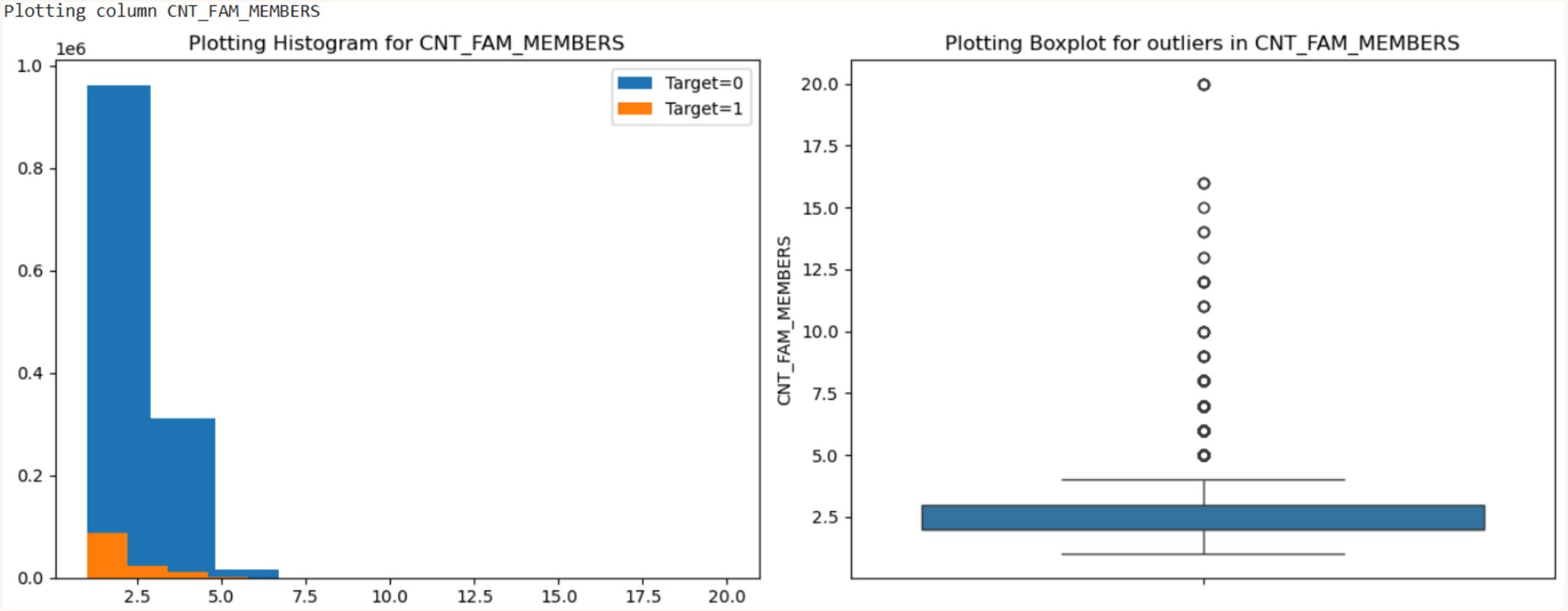
- This tells that clients with fewer children are less likely to loan default than those who have no children
- Clients with more than 10 children are considered unusual and need closer inspection.

2. Income of the client



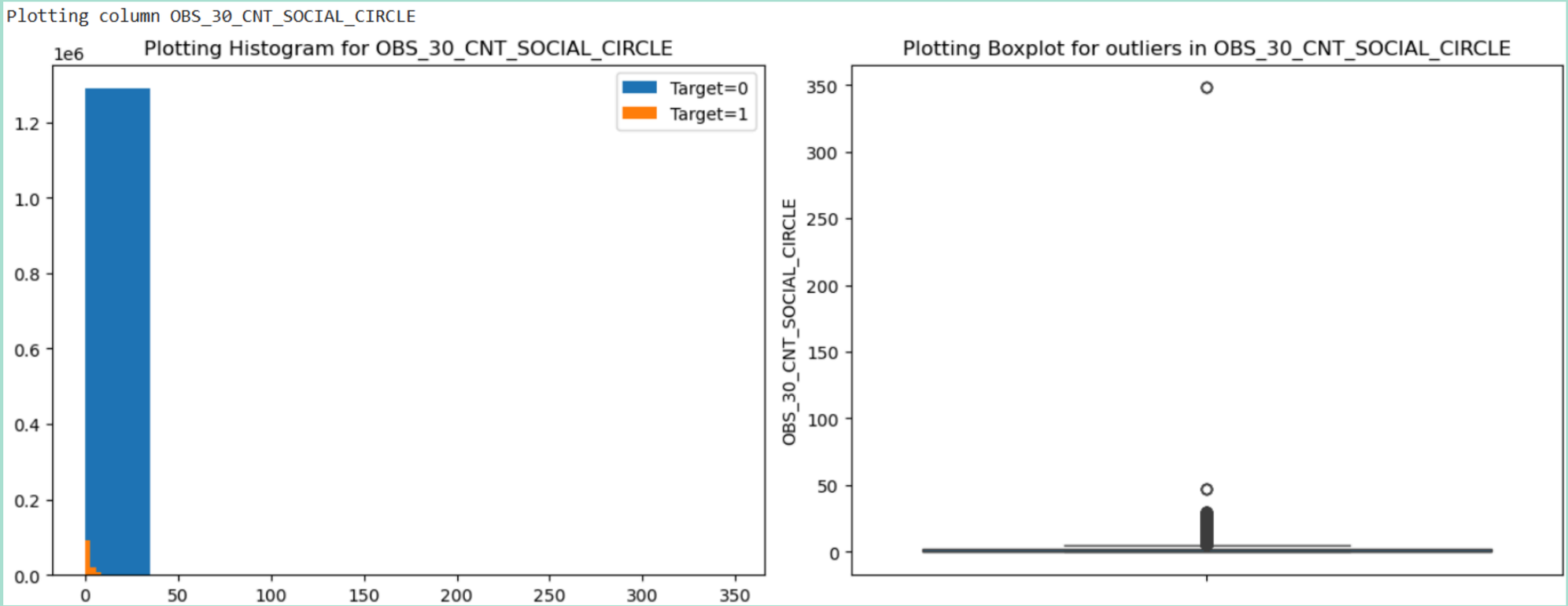
- A specific income value of around 120 million (1.2e8) seems very far from the others, making it an outlier that might need to be removed.
- Clients with higher income repay the loan in time.

3. Number of family members



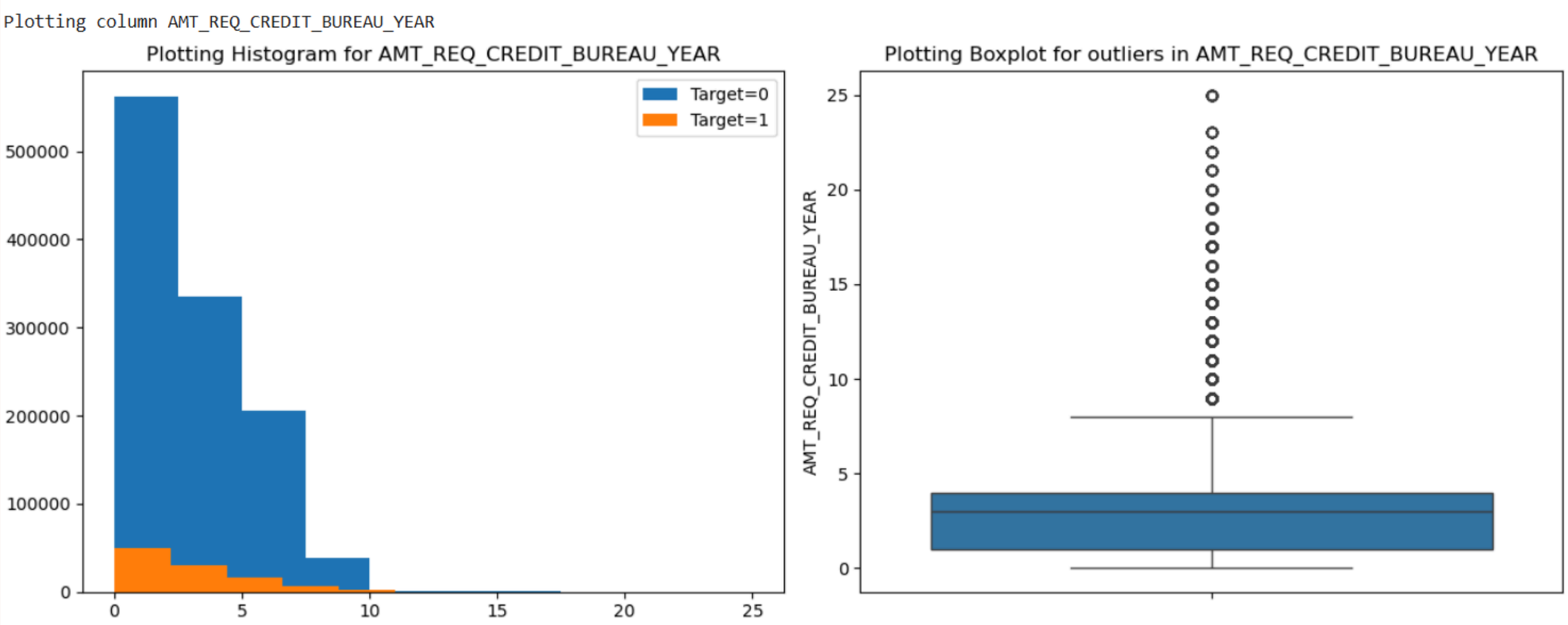
- The greater the number of family members, the lesser will be the loan default.
- We can observe outliers where some client have up to 20 family members.

4. Social Behaviors



- These columns track social behaviors and may also need further attention to understand if any values are unusual.

5. Enquiries to Credit Bureau

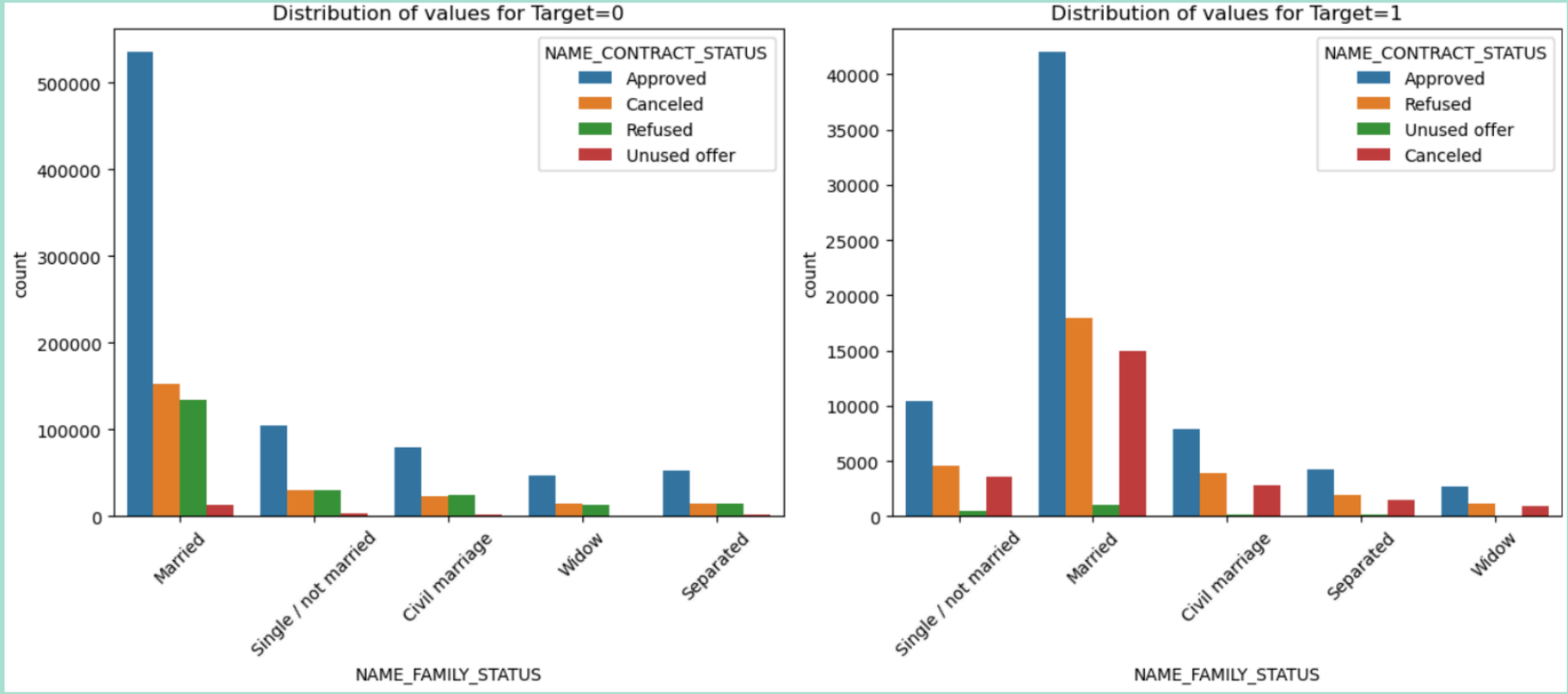


- We can observe outliers having more enquiries

Outlier analysis depends on the business context. If the data is accurate, whether to remove extreme values is based on the business goals. After reviewing different charts, it seems the data is correctly reported. Therefore, instead of removing these extreme values, we will mark the columns that have them for further review and analysis.

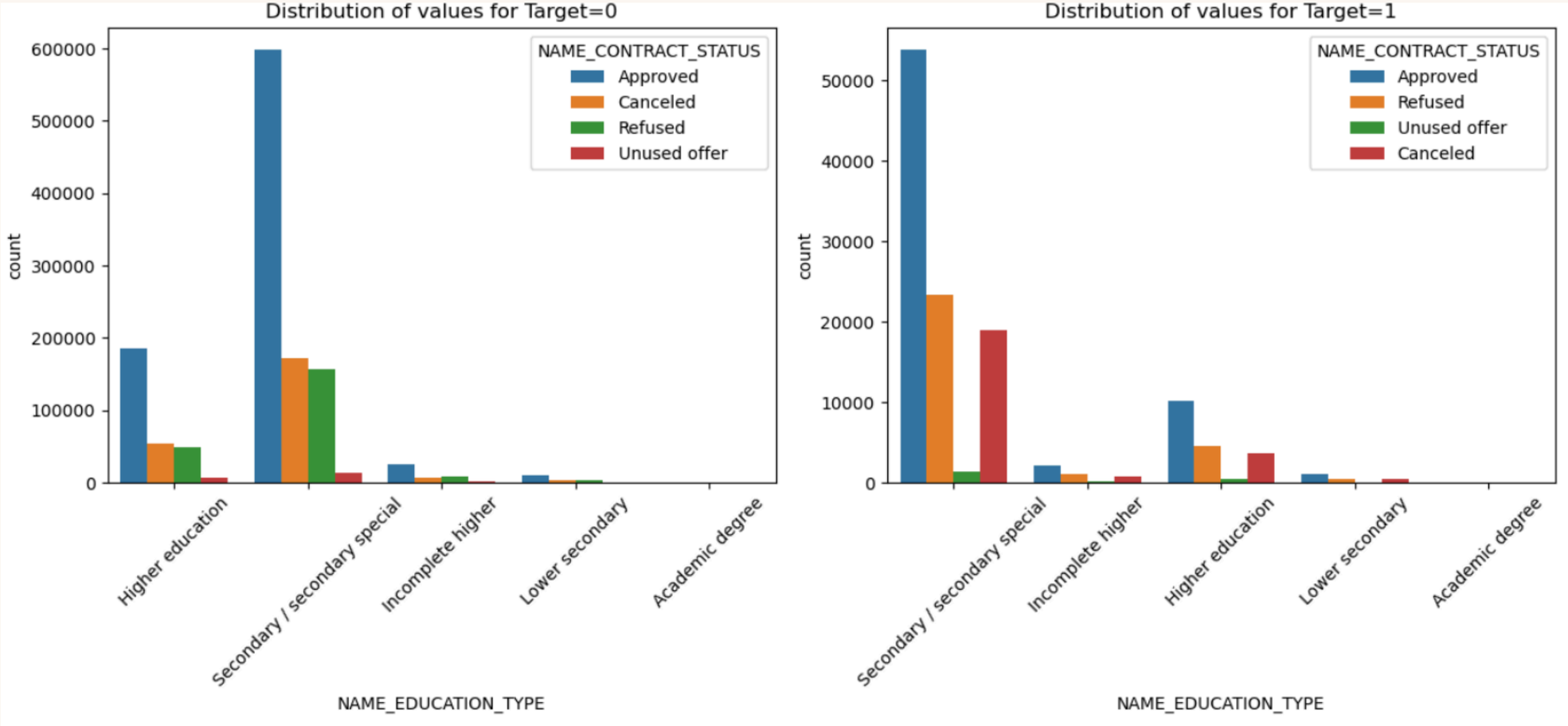
Bivariate analysis

1. NAME_FAMILY_STATUS & NAME_CONTRACT_STATUS



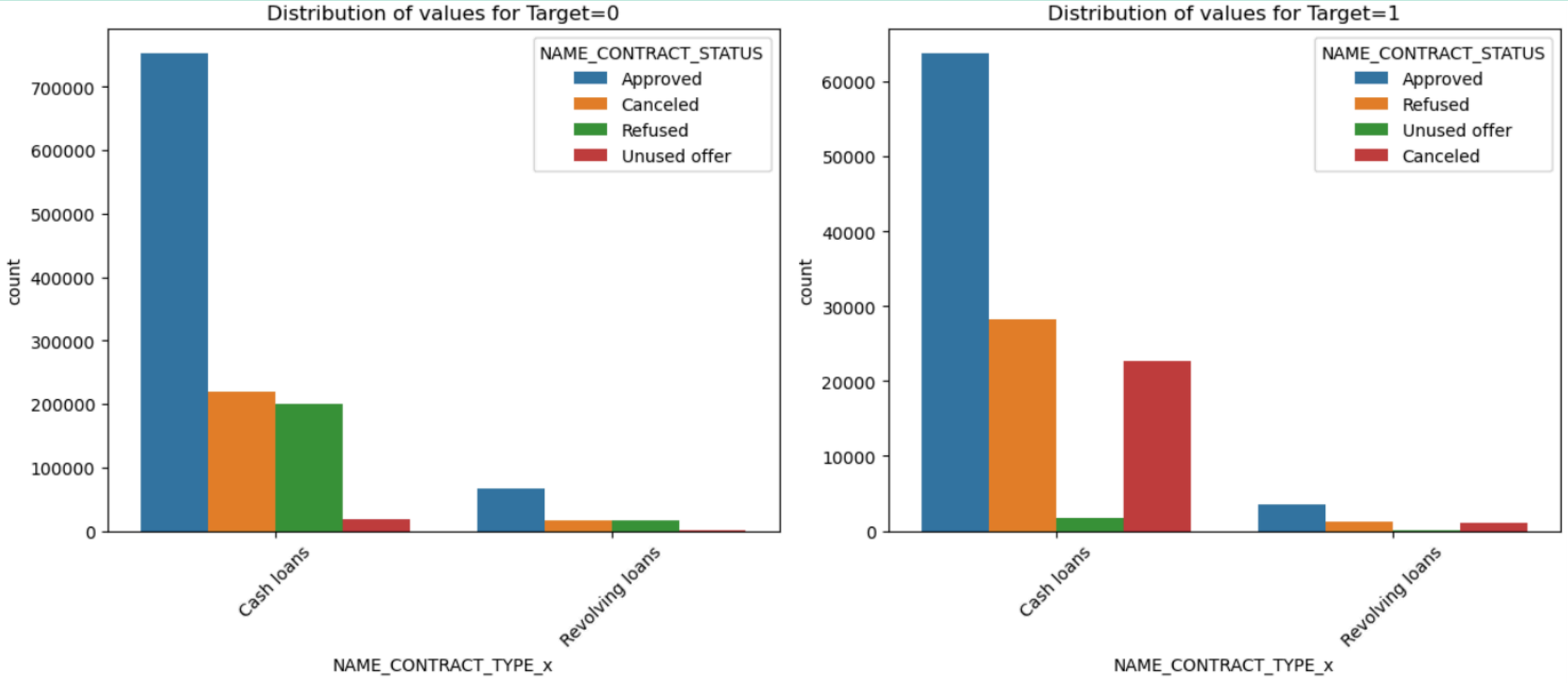
- Married people are better at repaying their loans on time than single people. This is evident from the data showing that married individuals have a higher rate of loan approval.

2. NAME_EDUCATION_TYPE & NAME_CONTRACT_STATUS



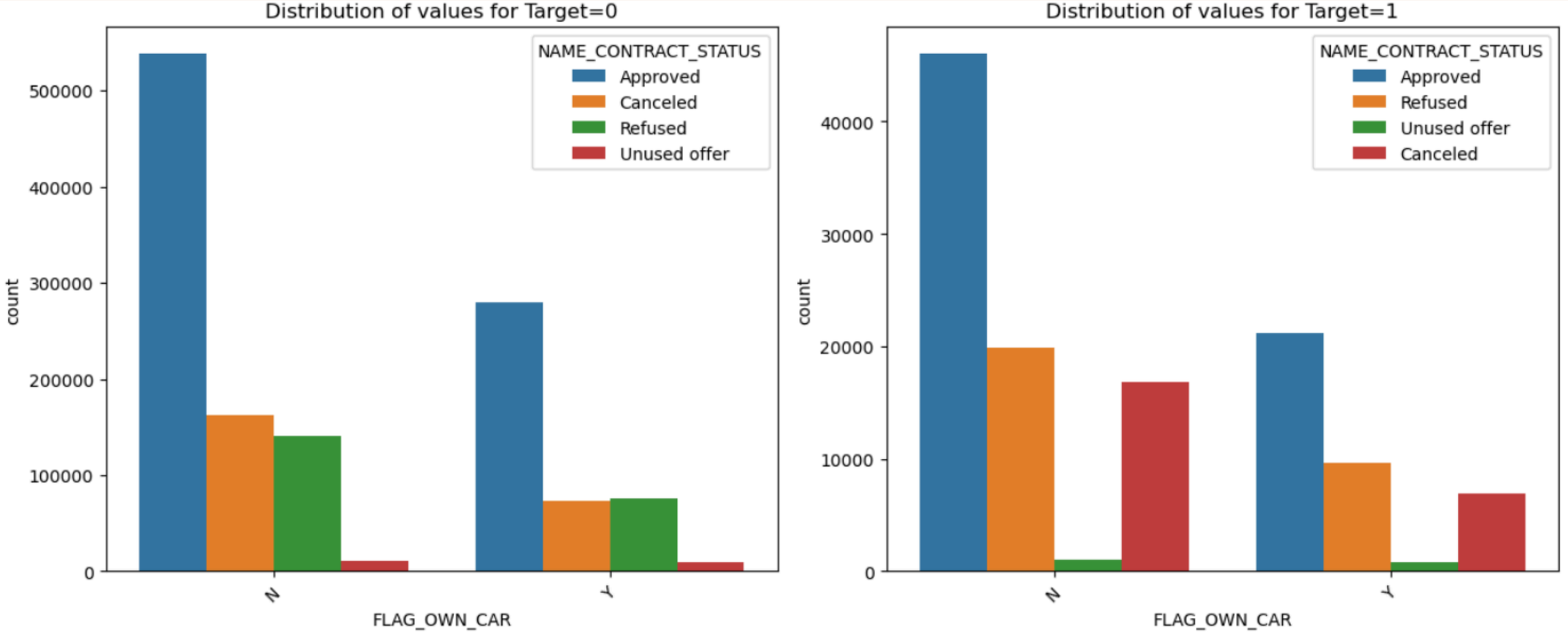
- Most people who apply for loans have an education level of "Secondary special." These individuals not only apply for more loans but also tend to get their loans approved.

3. NAME_CONTRACT_TYPE_x & NAME_CONTRACT_STATUS



- People mostly apply for cash loans and are highly approved

4. FLAG_OWN_CAR & NAME_CONTRACT_STATUS



- People without car are more likely to repay loans

Step 6: Conclusion

- Loan default is influenced by several key factors, including **loan type, gender, income, education, marital status, and car ownership.**

- Cash loans, while simple and accessible, often lead to higher repayment difficulties compared to revolving credit.

- Women are more likely to repay loans on time, whereas men exhibit significantly higher default rates.

- Clients with higher incomes are generally more reliable in repaying loans, while those with lower education levels, such as secondary education, tend to have higher default rates compared to individuals with advanced education.

- Additionally, car ownership impacts repayment behavior, as people without cars apply for loans more frequently but face greater repayment challenges.

- These factors collectively underscore the importance of understanding borrower demographics and financial behaviors to better assess credit risk and tailor loan offerings.

- Marital status also plays a crucial role, with married individuals being less likely to default than singles or those in civil marriages.