Credit EDA case Study

Introduction

This case study aims to give us an idea of applying EDA in a real business scenario. In this case study, we develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

Business Objectives

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

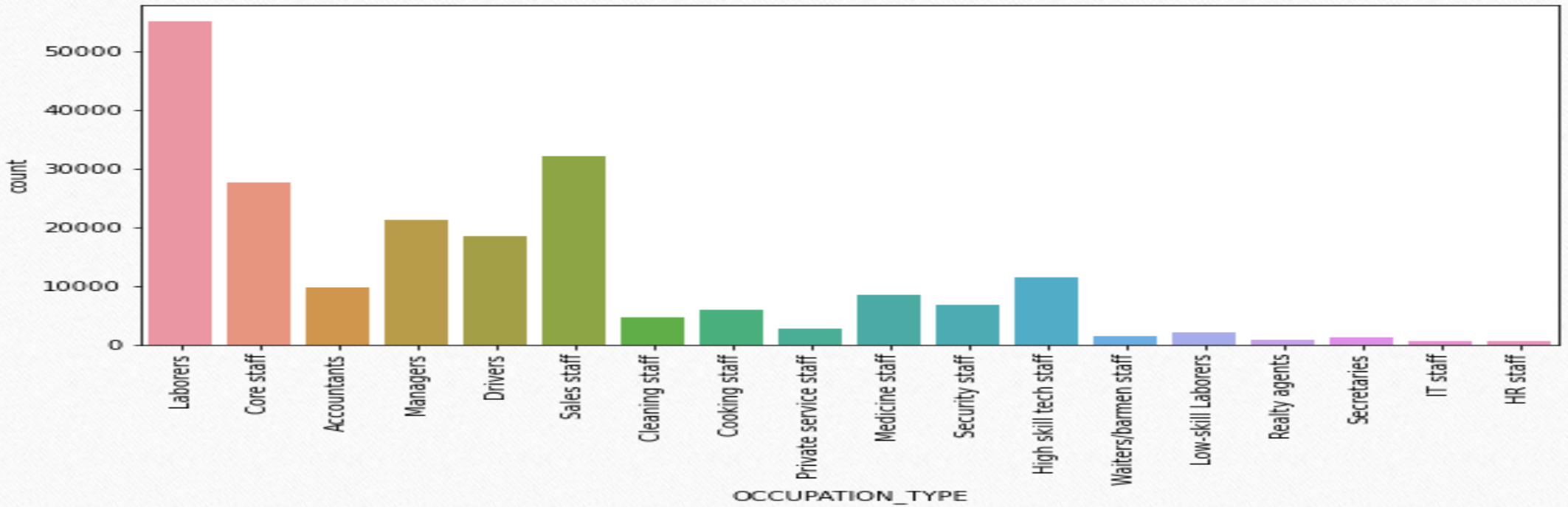
Steps

- **Data Understanding**
- Import Libraries
- Loading The Data
  - Inspecting The Dataframe
  - Data Cleaning
  - Best Metric To Impute Missing Values In Some Columns
  - Errors in Data types and Data
  - Binning of continuous variable

Steps

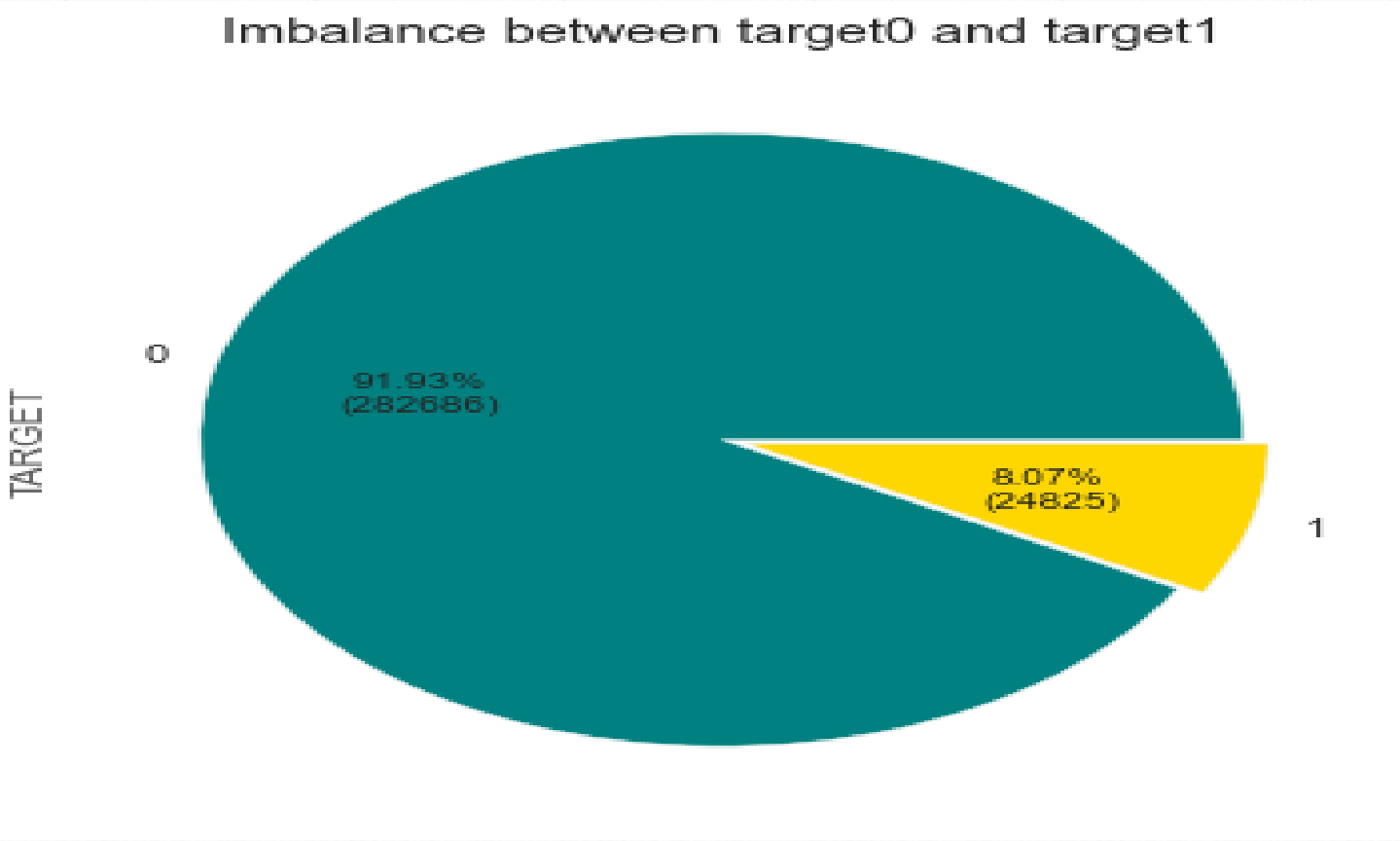
- **Data Analysis**
  - ✓ Univariate Analysis of Categorical Variables
  - ✓ Univariate Analysis of Numerical Variables on the basis of 'Target' Variable
  - ✓ Bivariate Analysis
- Merging the files and analyzing the data

Occupation Type Analysis



Observations

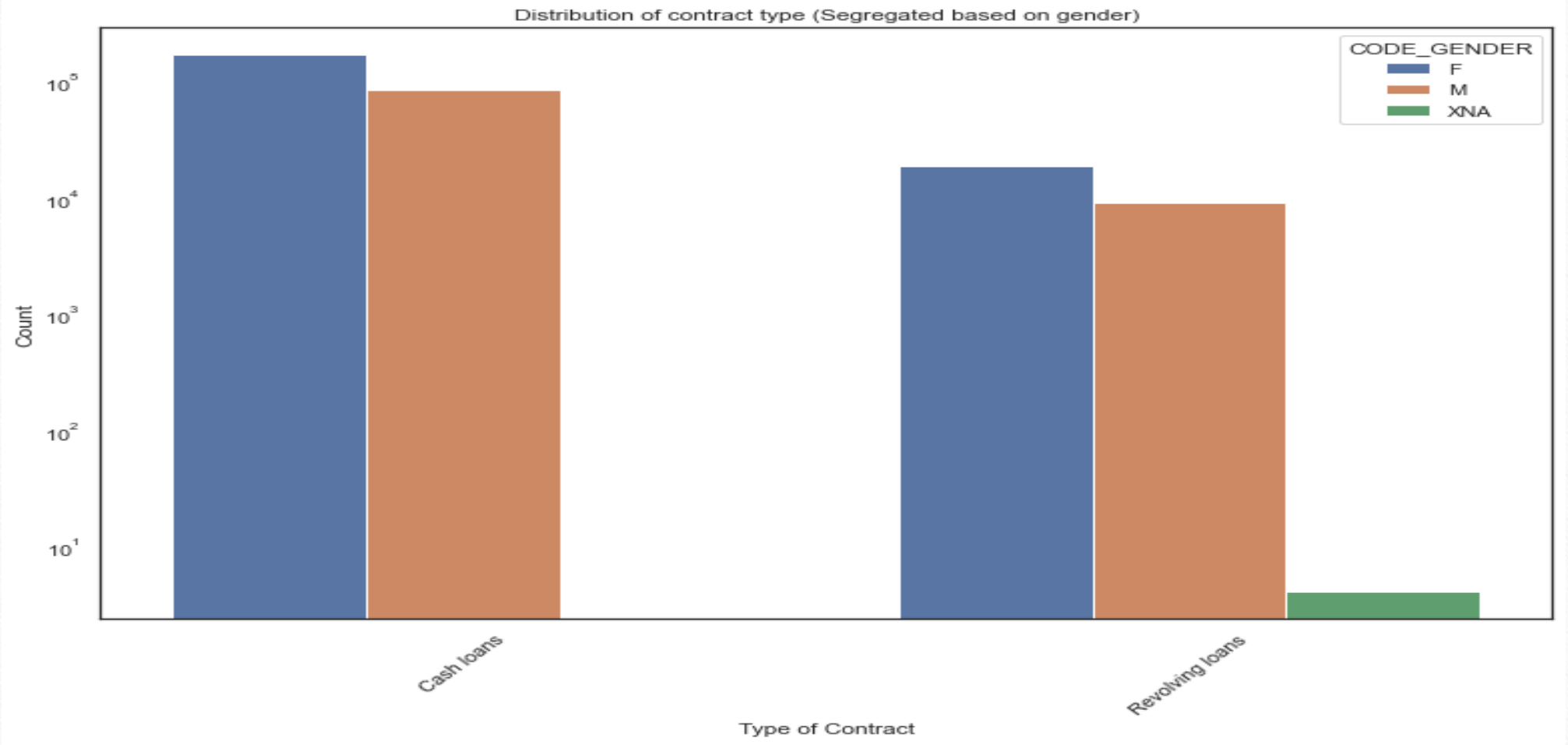
- Looking at the plot, Laborers has the highest number of loan applicants
- For imputation, it would be better to leave the data as is (missing values being 31.35%) and not impute to min/min/mode/median as it may bias the data in later computations



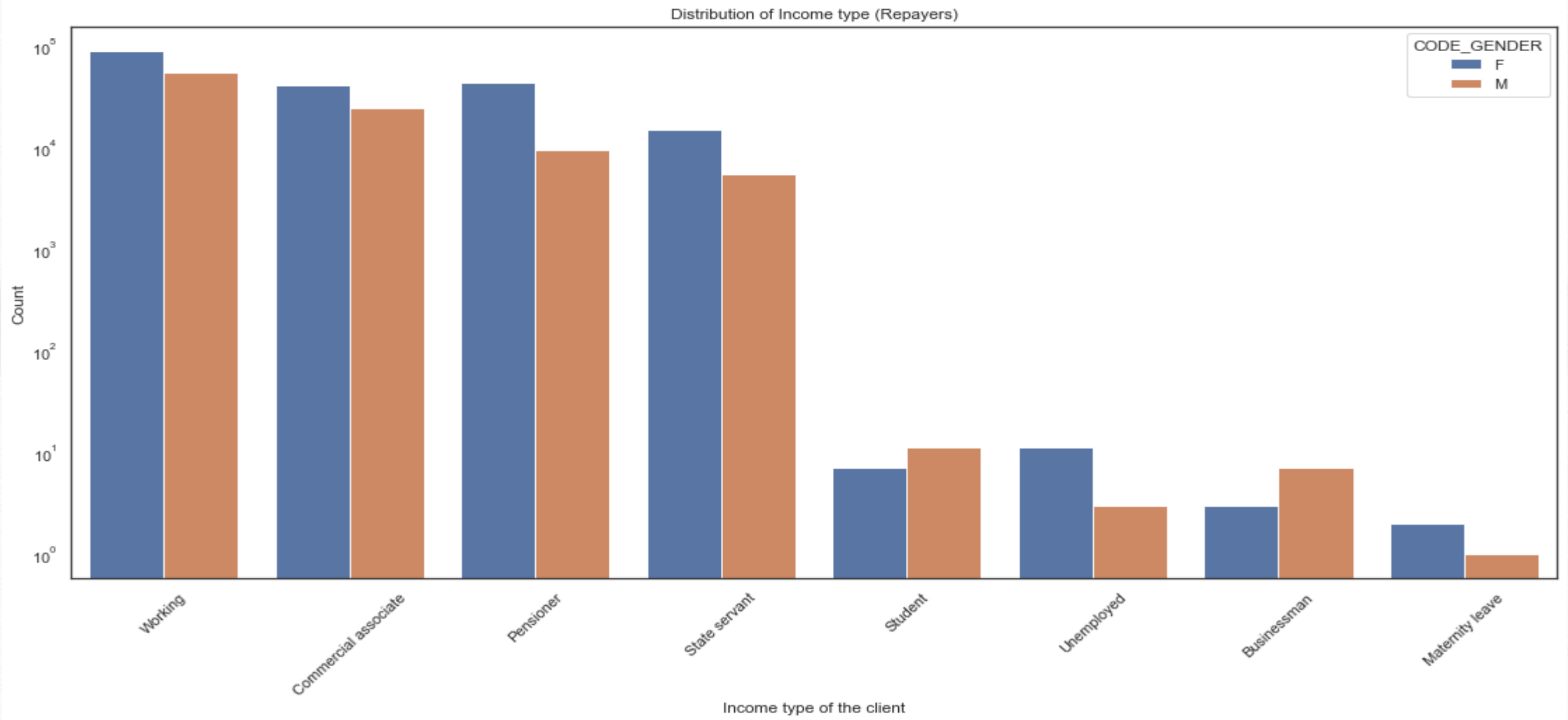
Observations

8.07% of clients are clients with payment difficulties. 91.21% of clients fall under the 'all other cases' category.

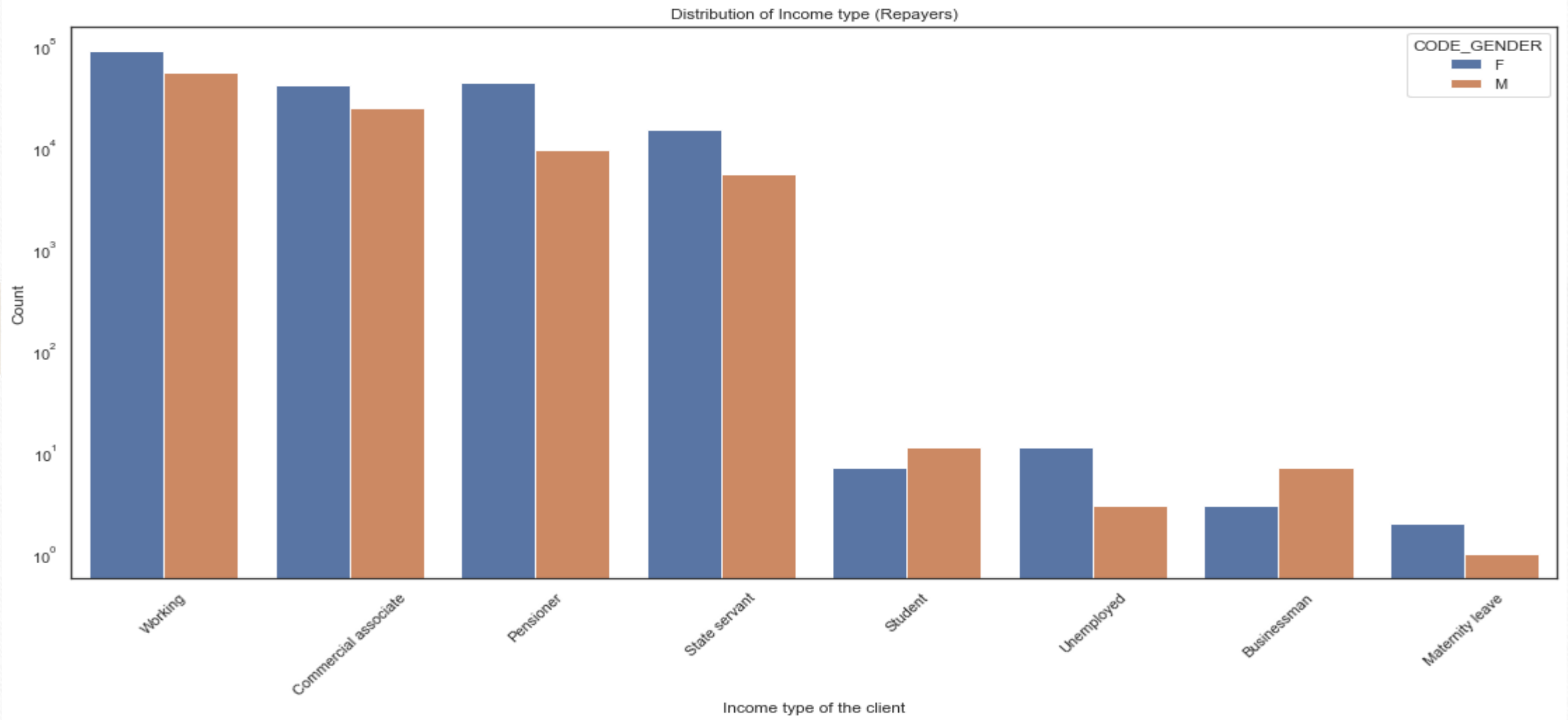
Distribution of contract type (Segregated based on gender)



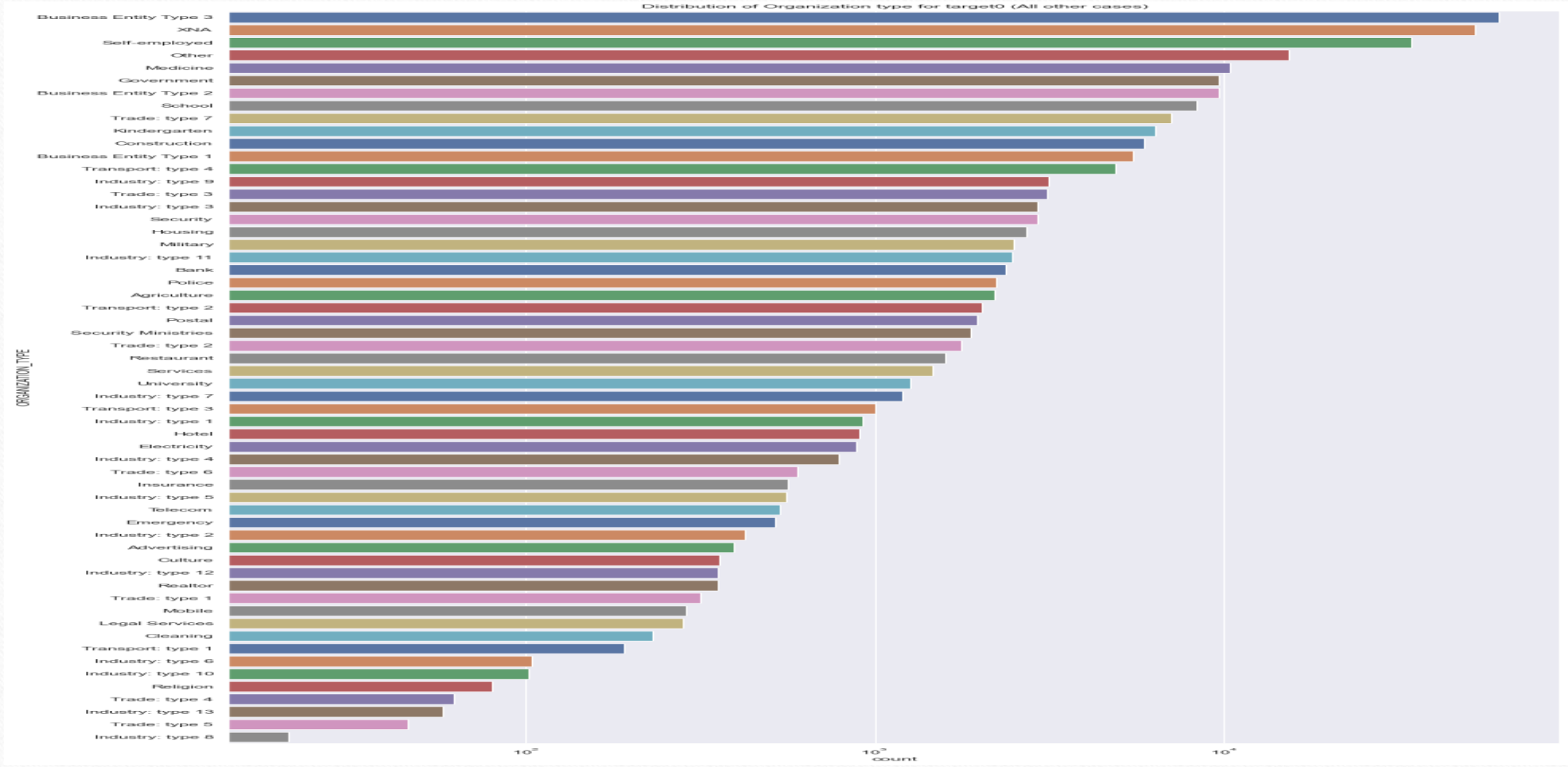
Distribution of Income type (Repayers)



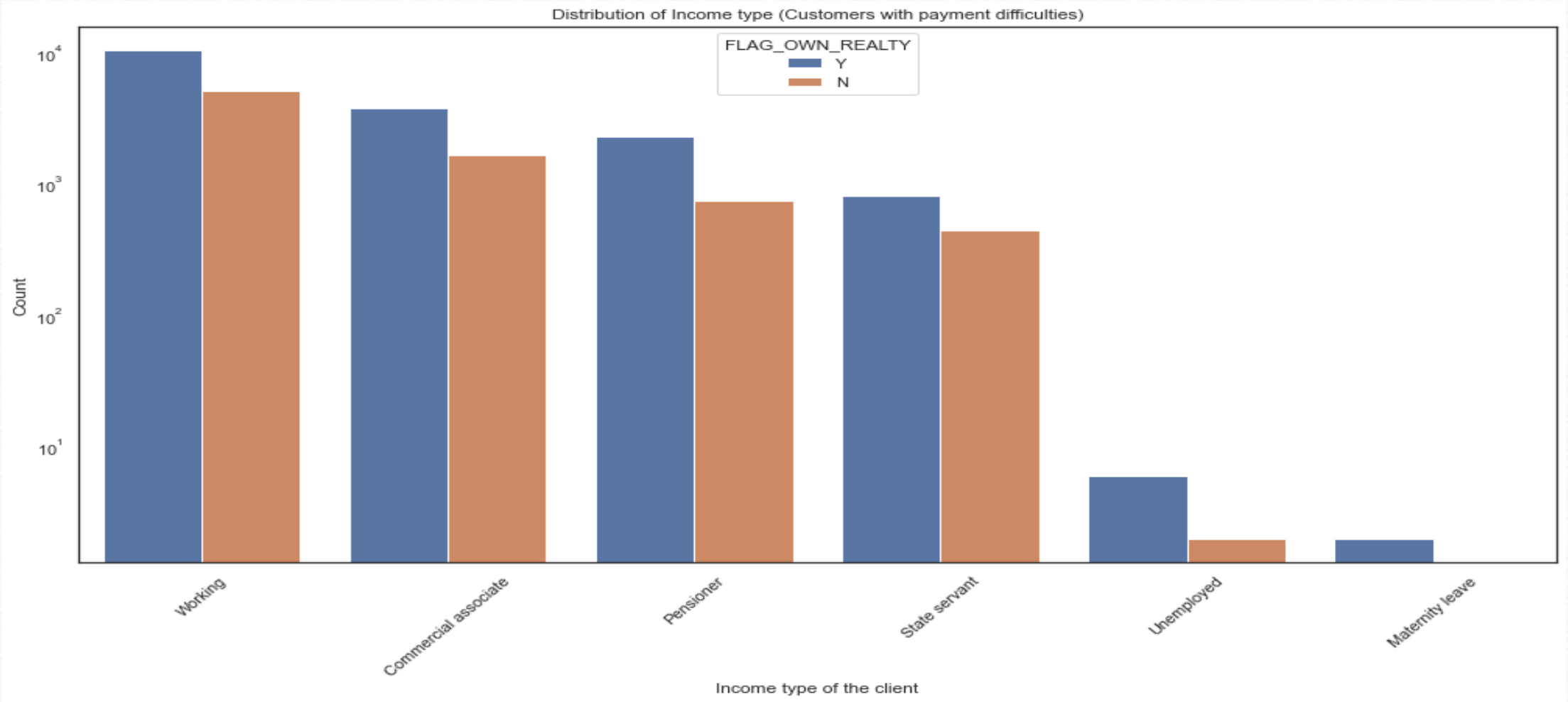
Distribution of Income type (Repayers)



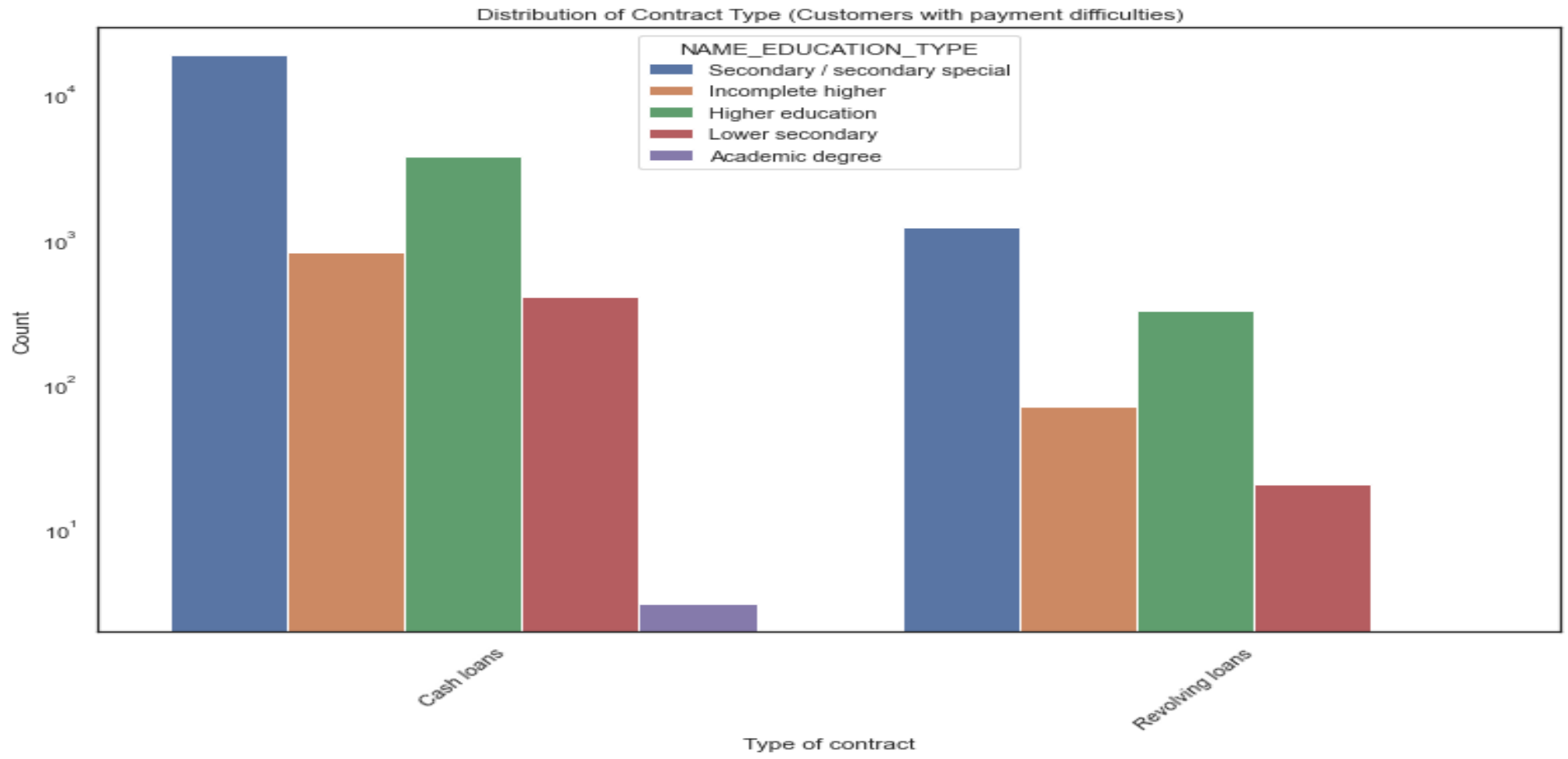
Distribution of Organization type for target0 (All other cases)



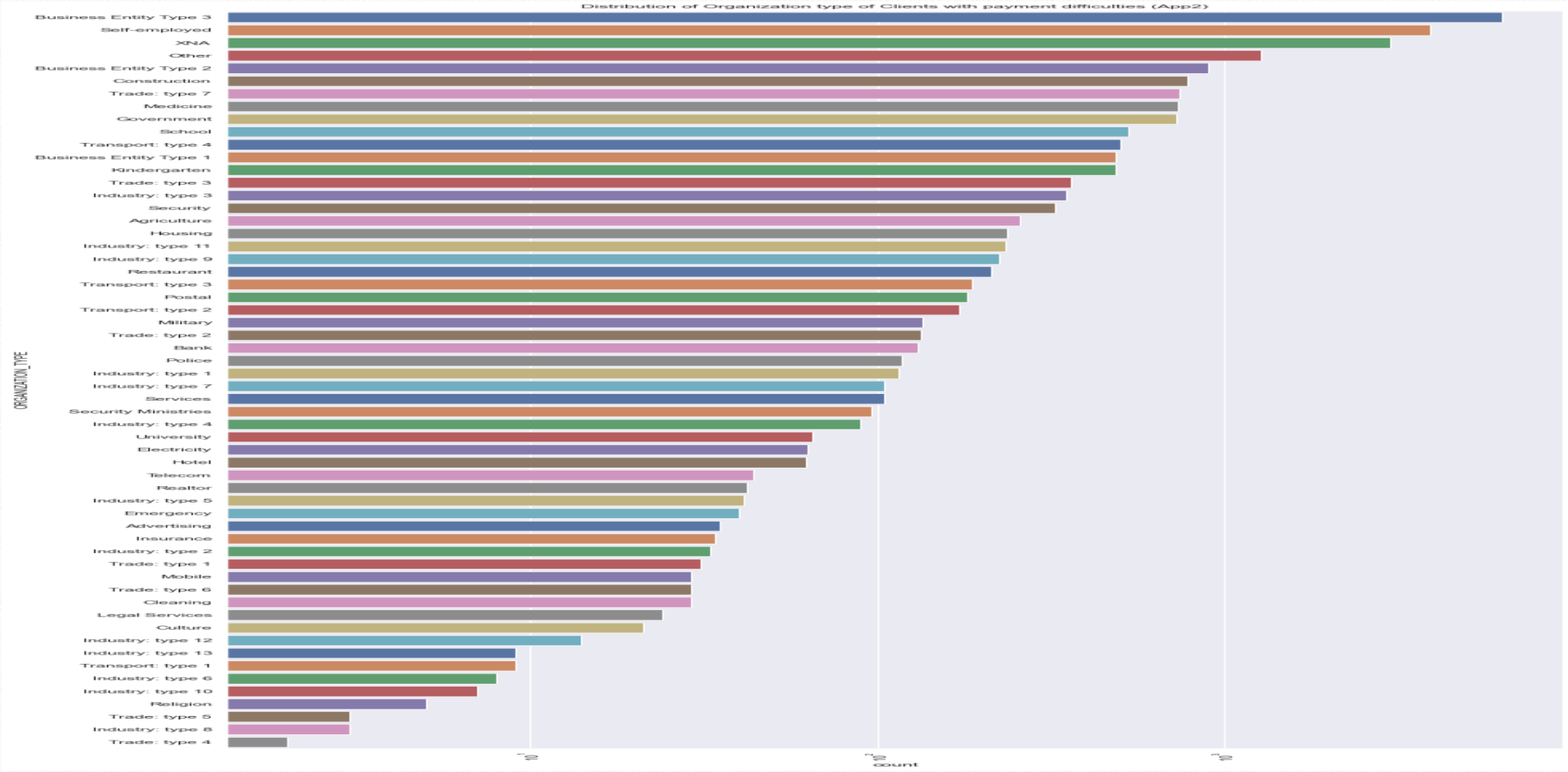
Distribution of Income type (Customers with payment difficulties)



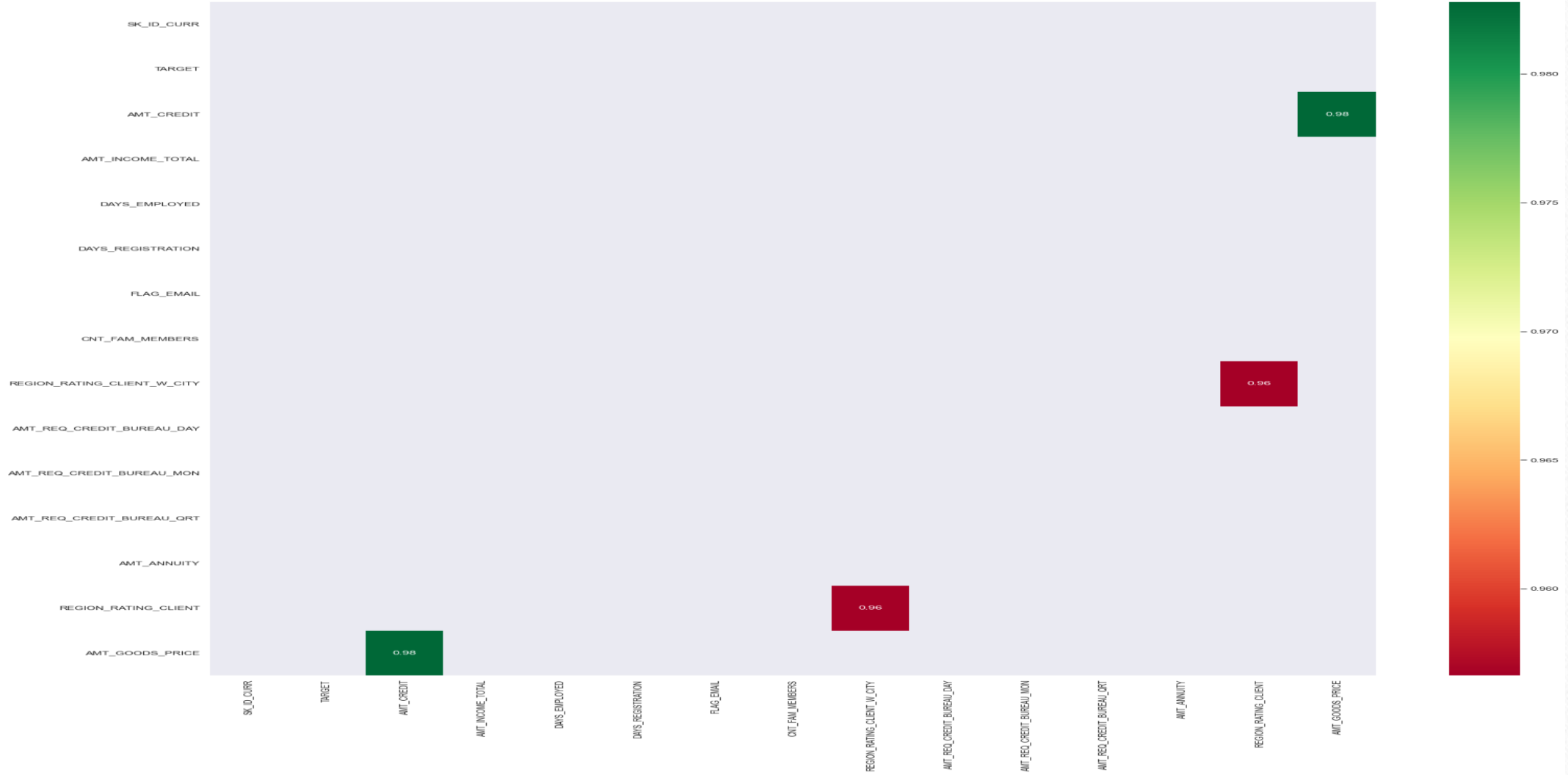
Distribution of Contract Type (Customers with payment difficulties)



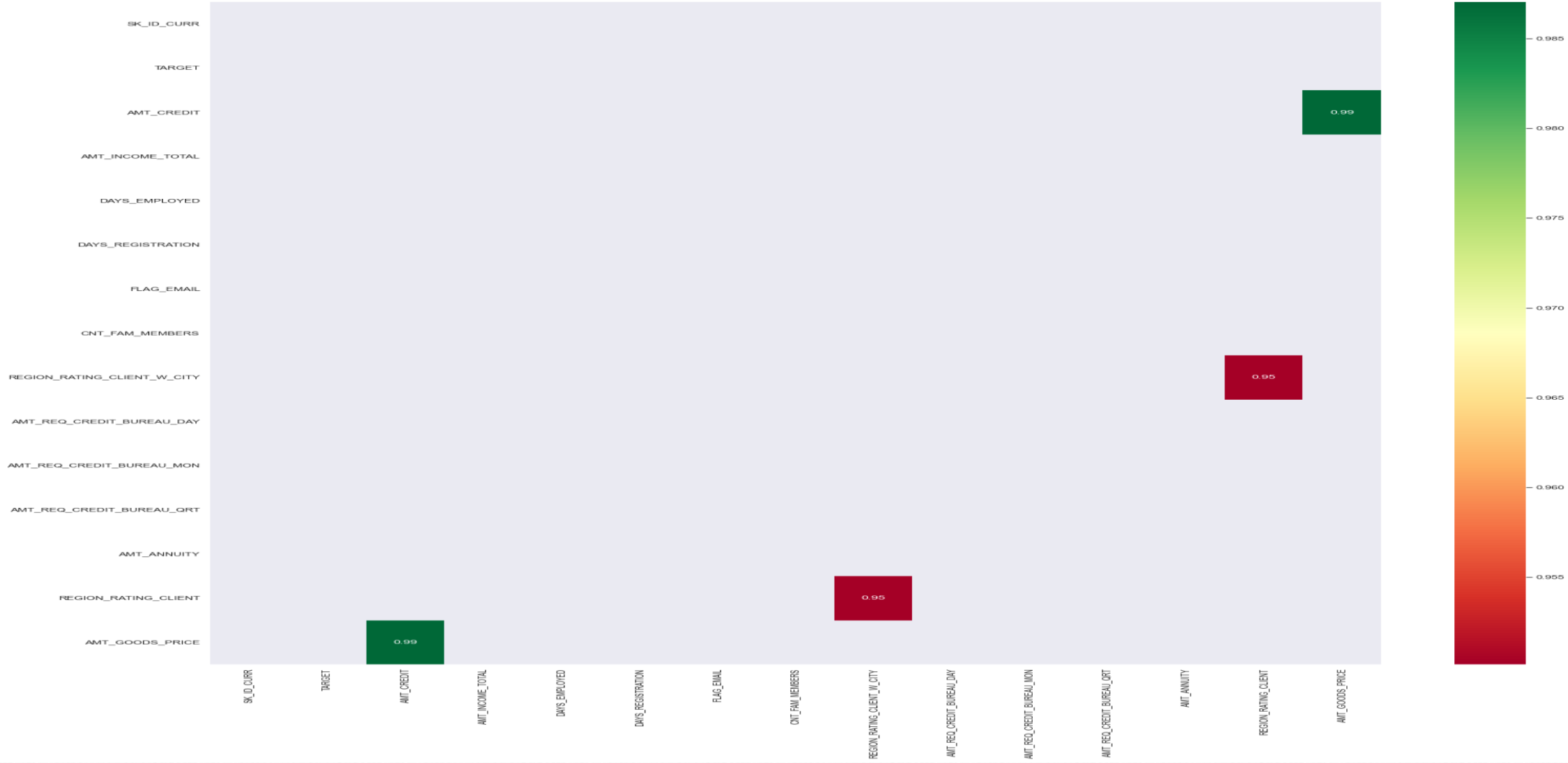
Distribution of Organization type of Clients with payment difficulties



Correlation analysis of numerical variables(Heat Map)

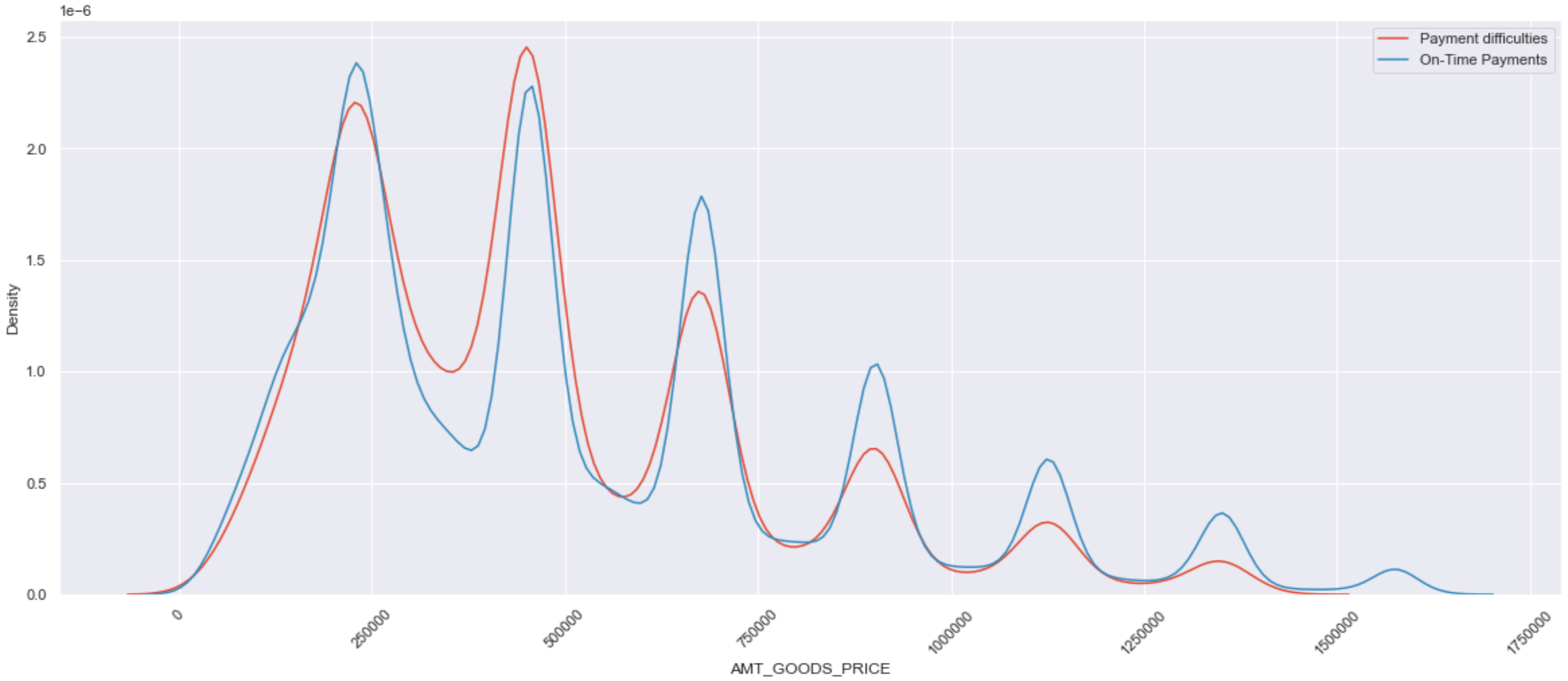


Correlation analysis of numerical variables(Heat Map)

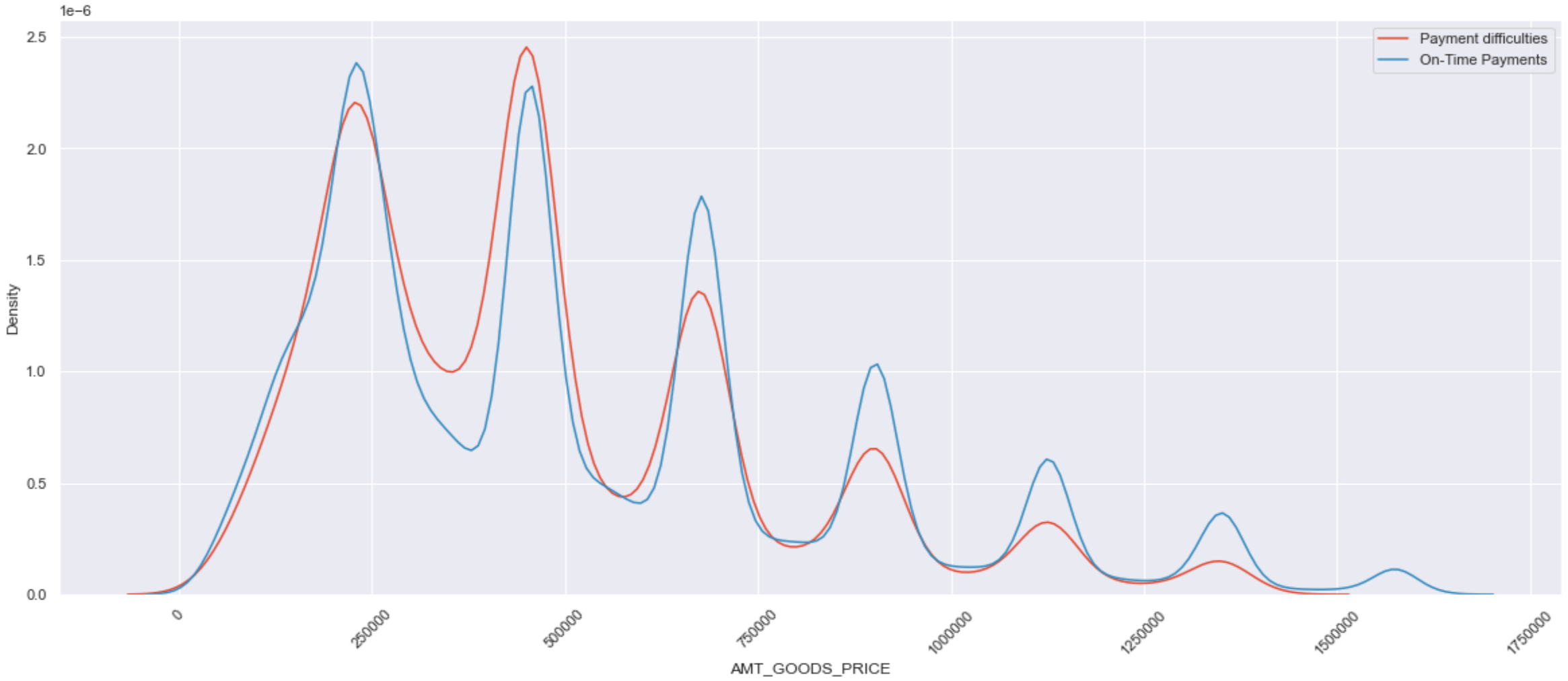


Univariate analysis of numerical variables

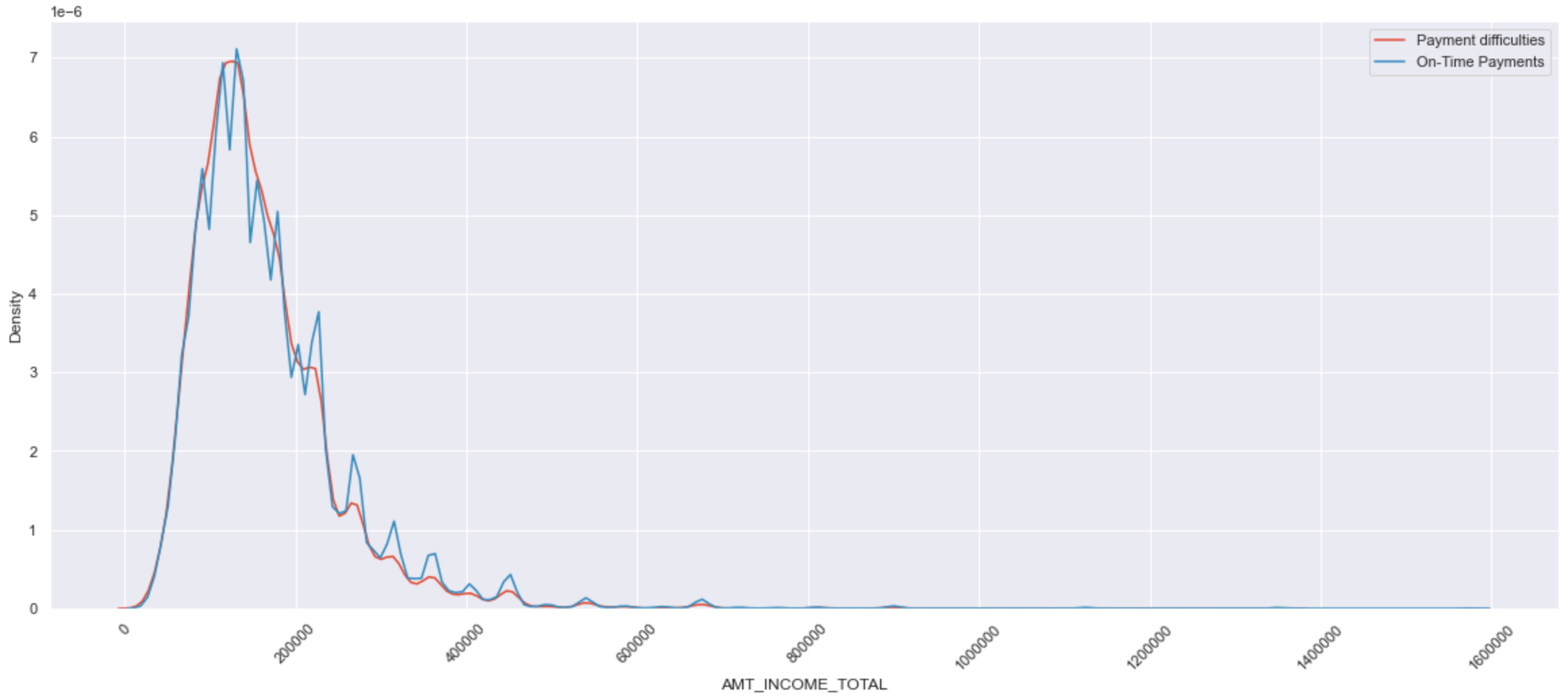
Univariate analysis of numerical variables



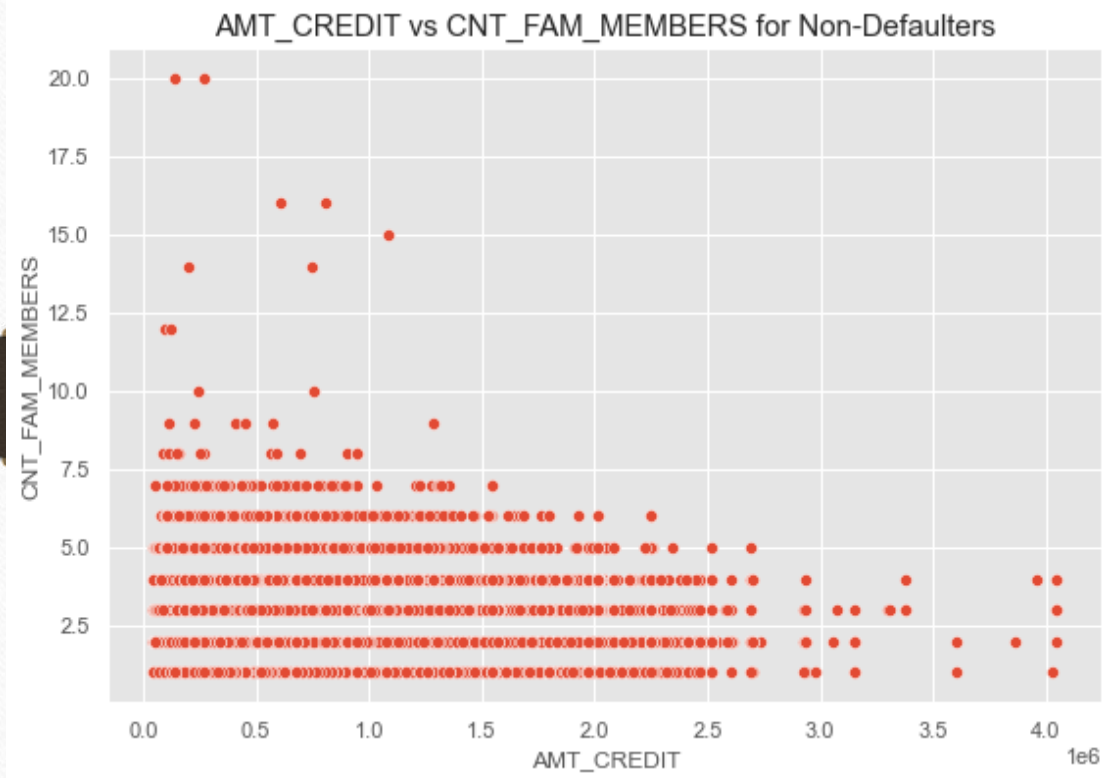
Univariate analysis of numerical variables

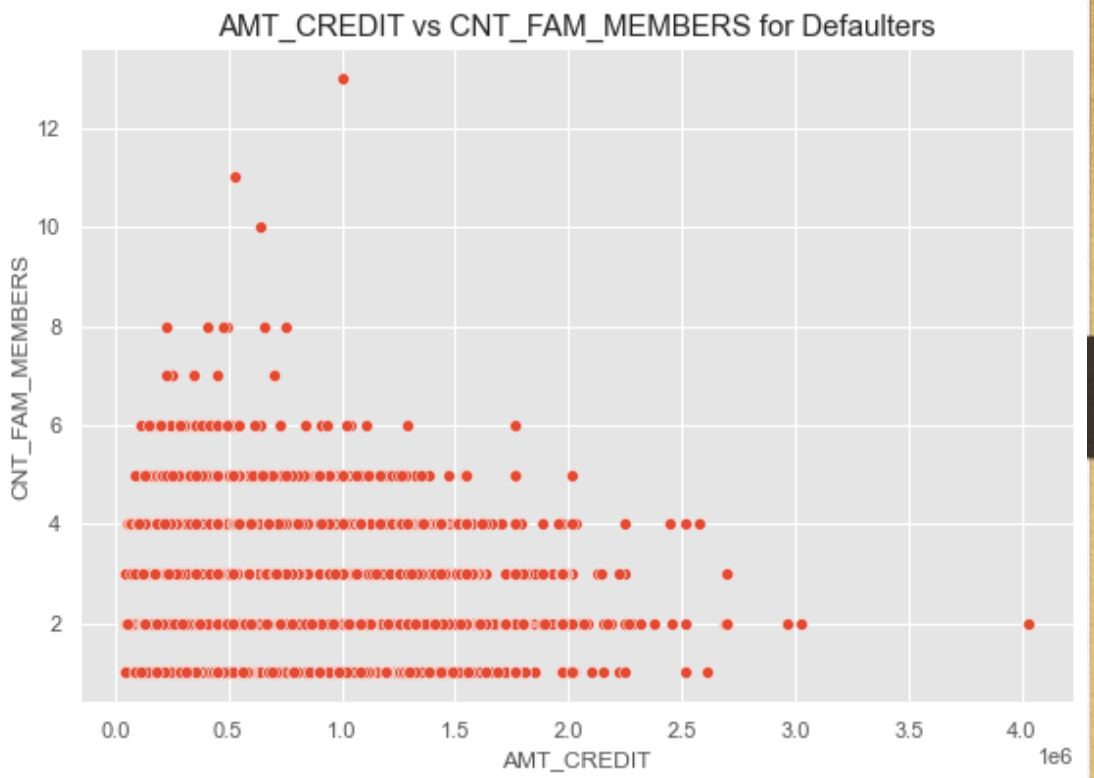


Univariate analysis of numerical variables

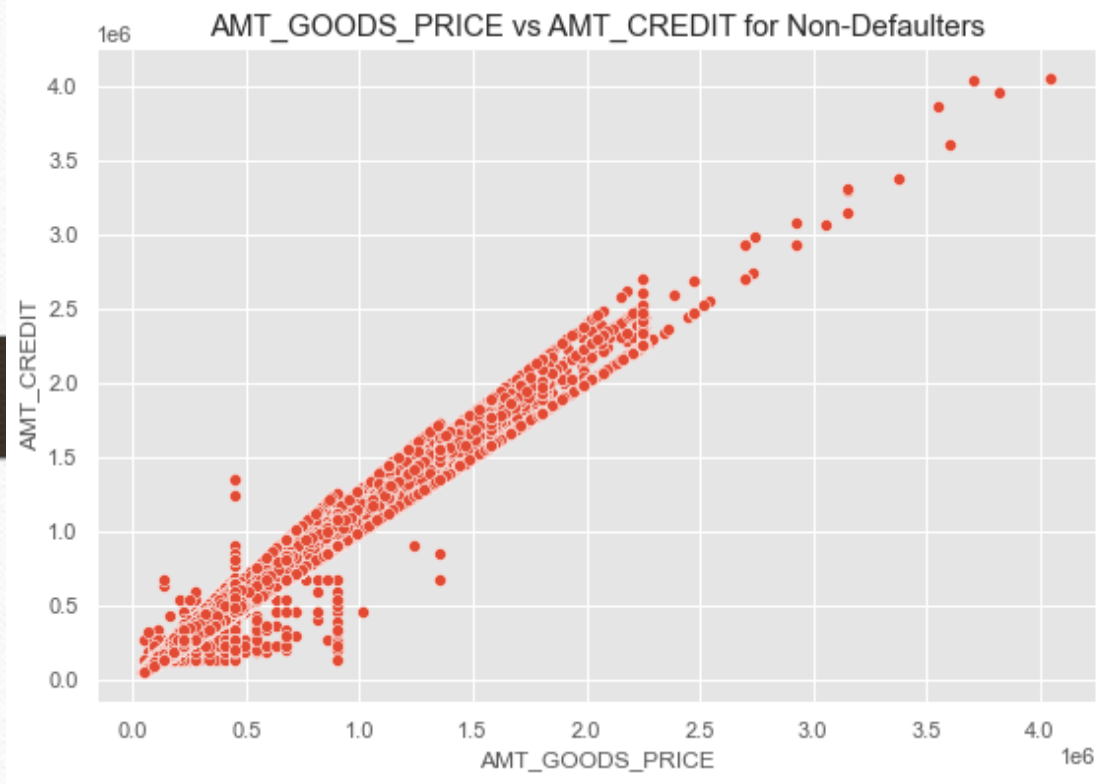


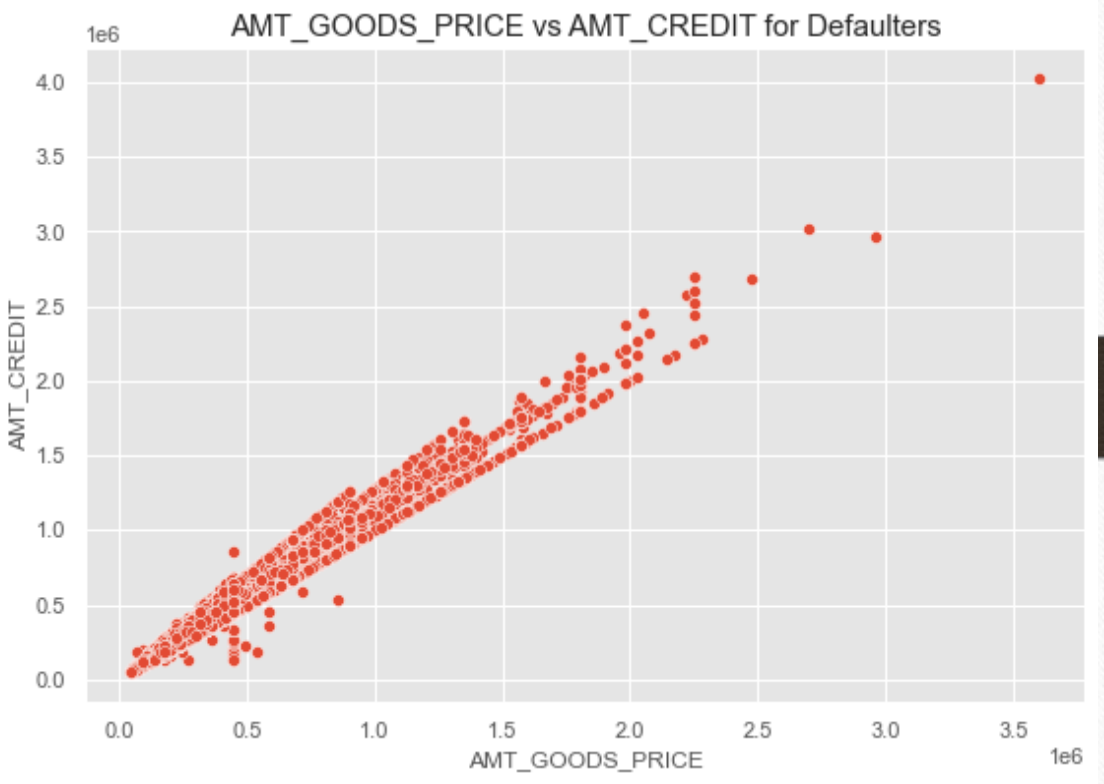
Univariate analysis of numerical variables



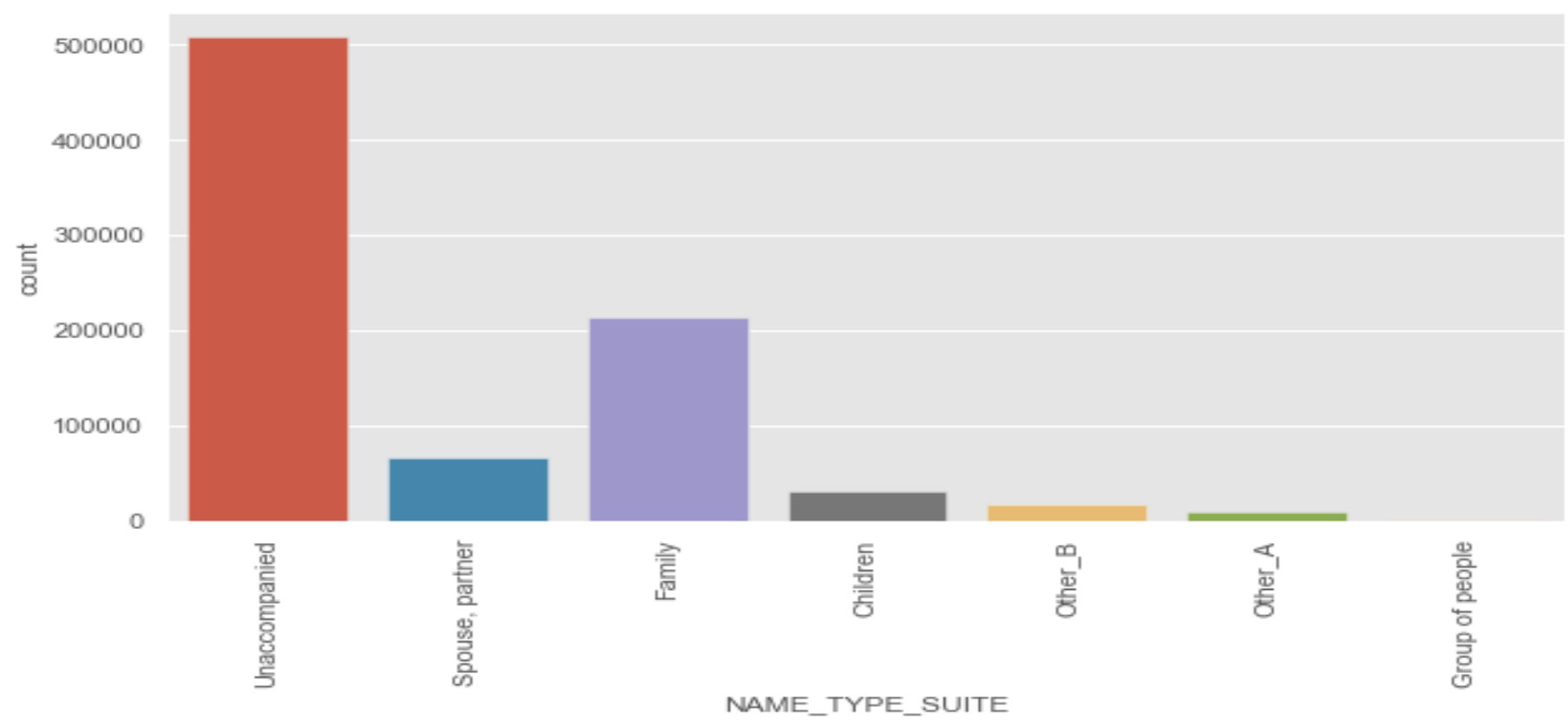


Univariate analysis of numerical variables

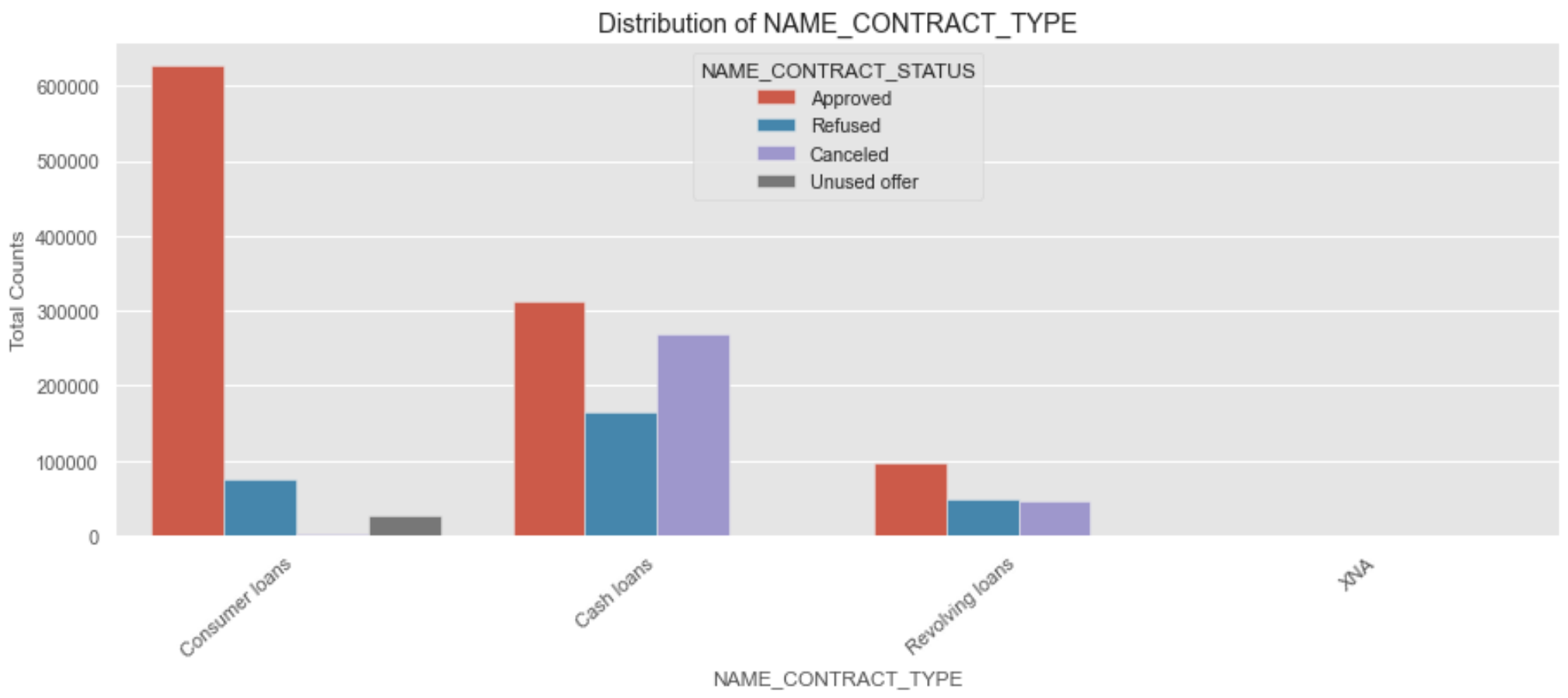




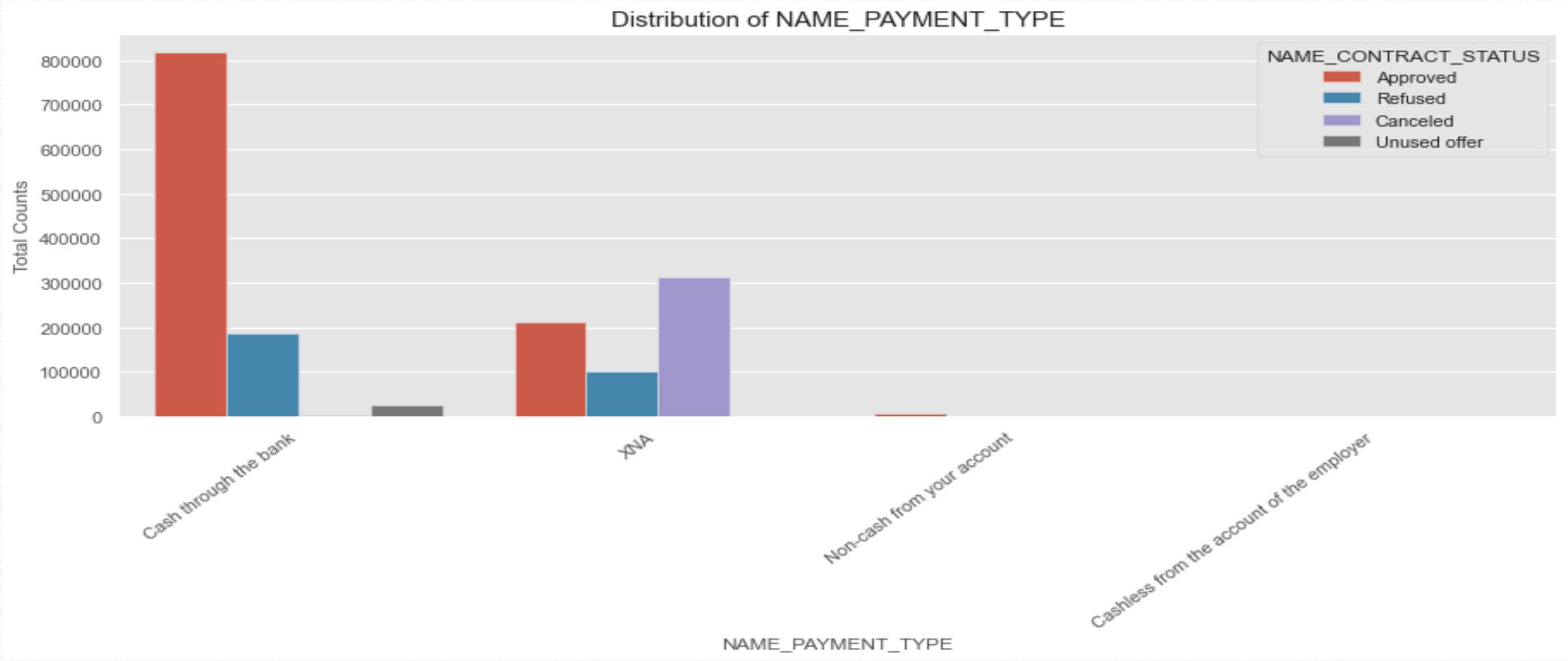
Previous Application Data Analysis



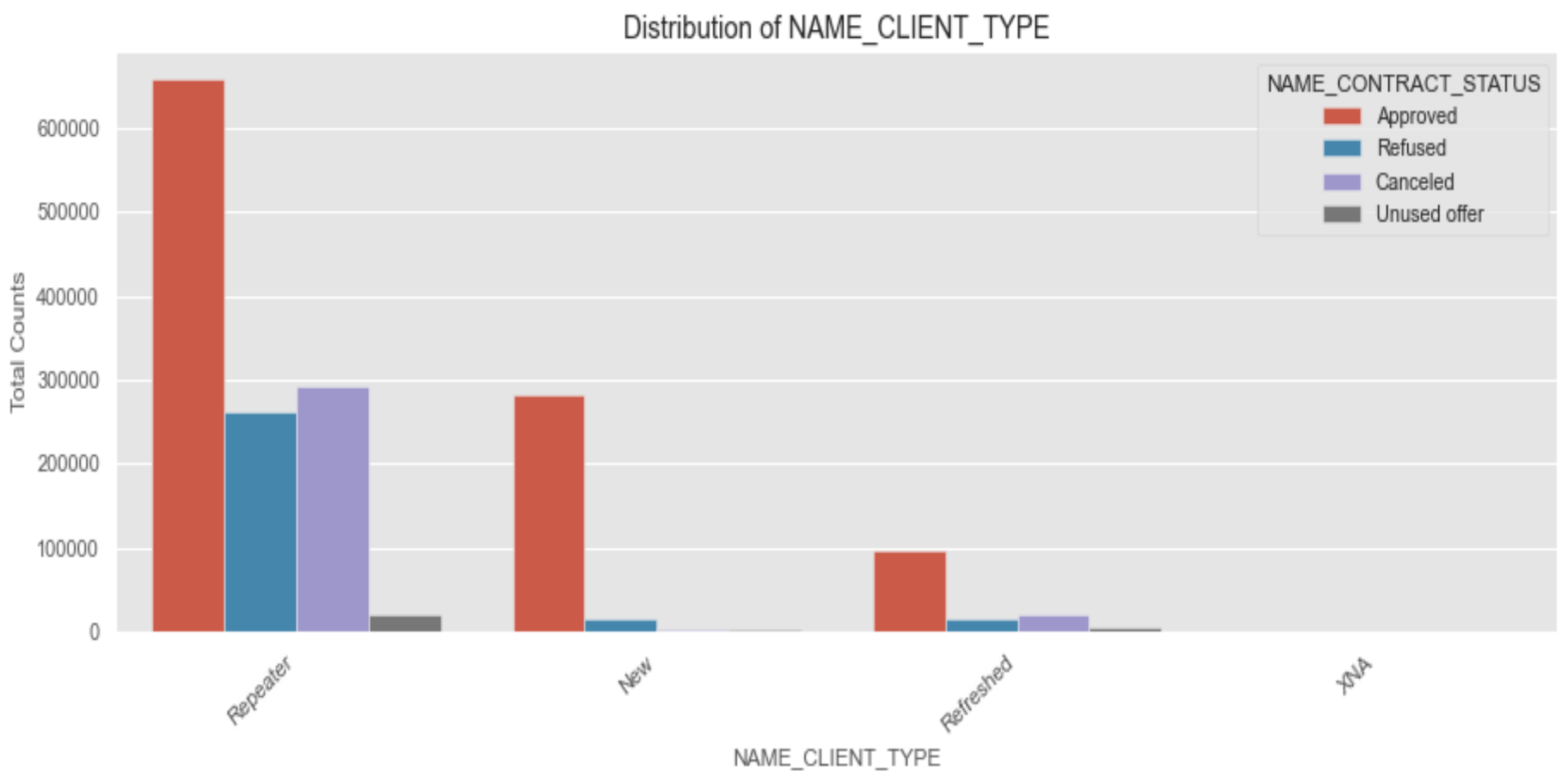
Previous Application Data Analysis



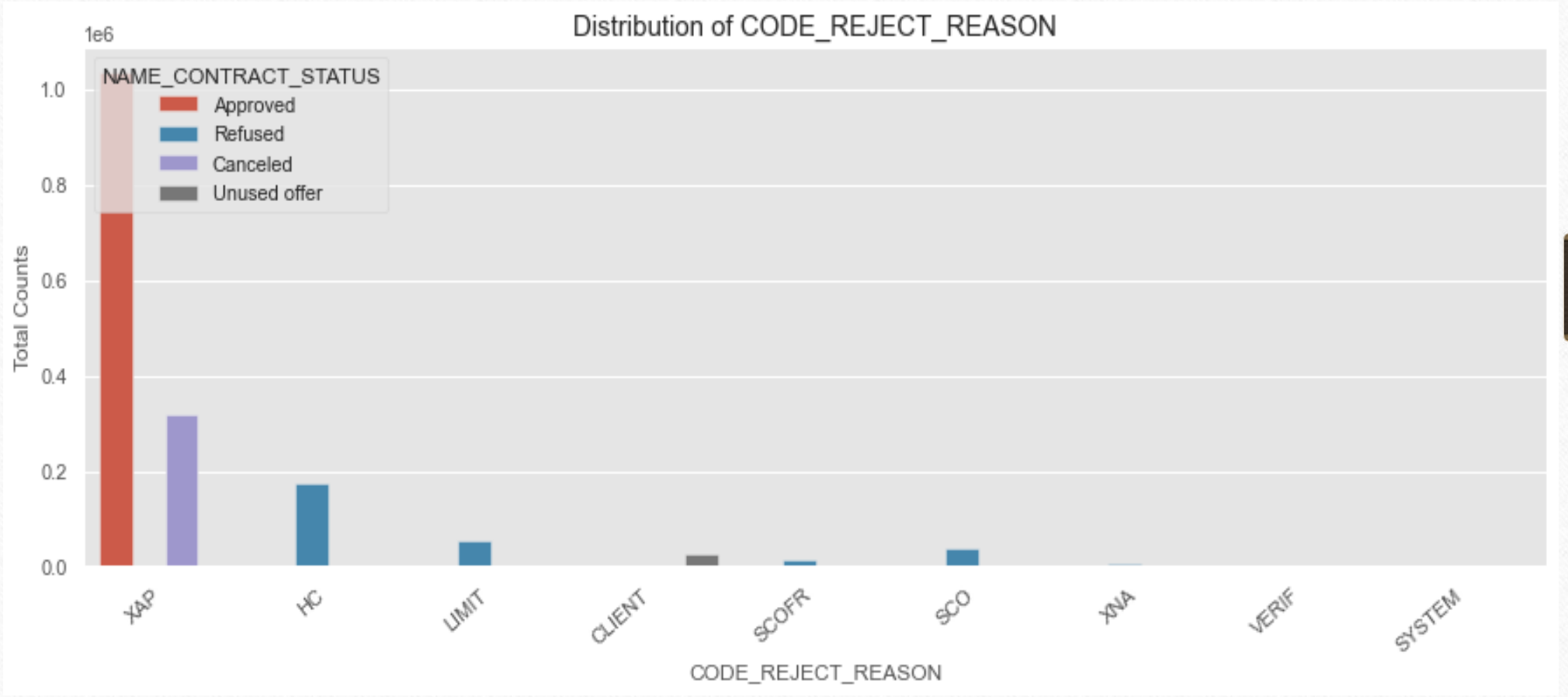
Previous Application Data Analysis

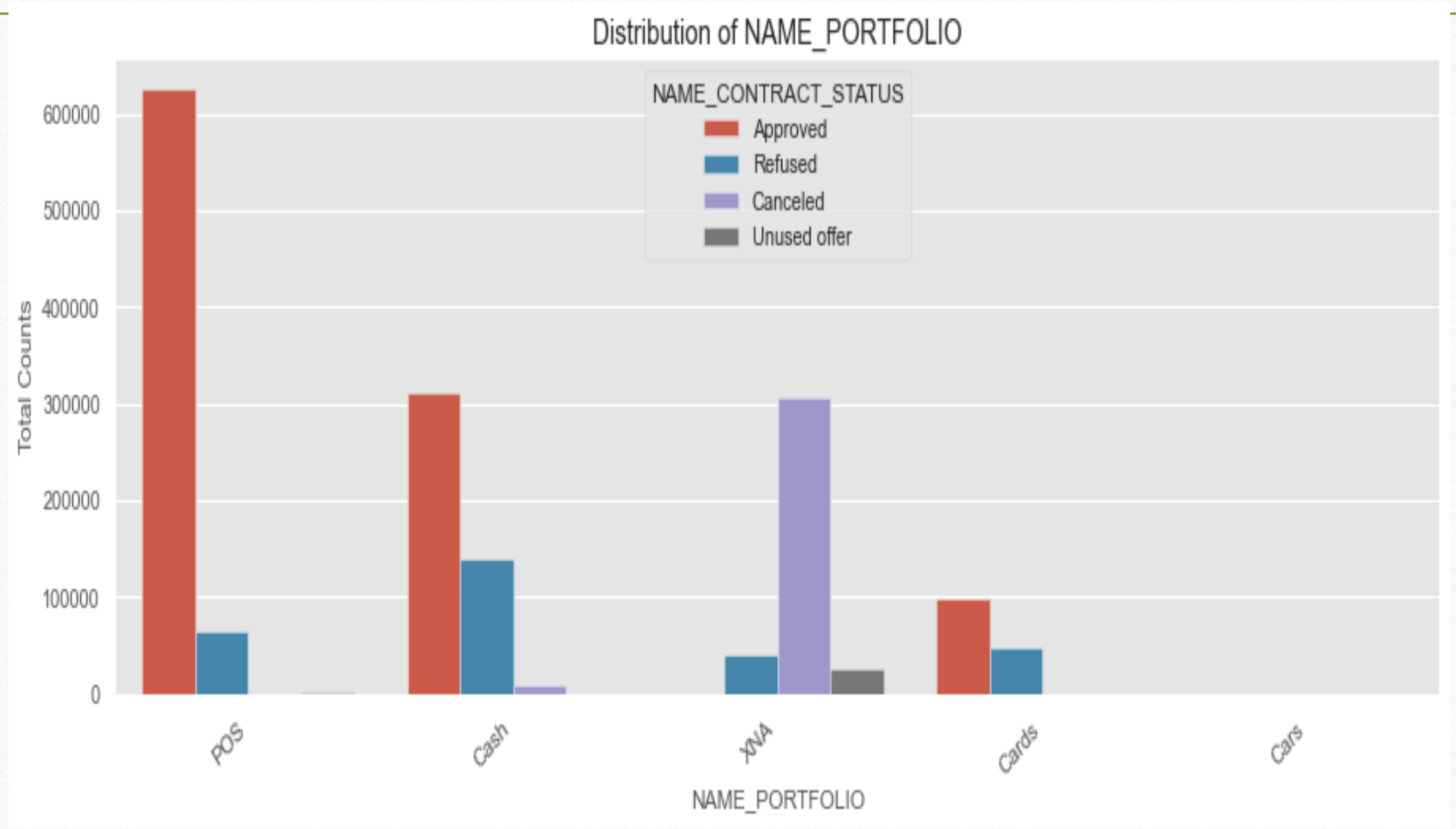


Previous Application Data Analysis

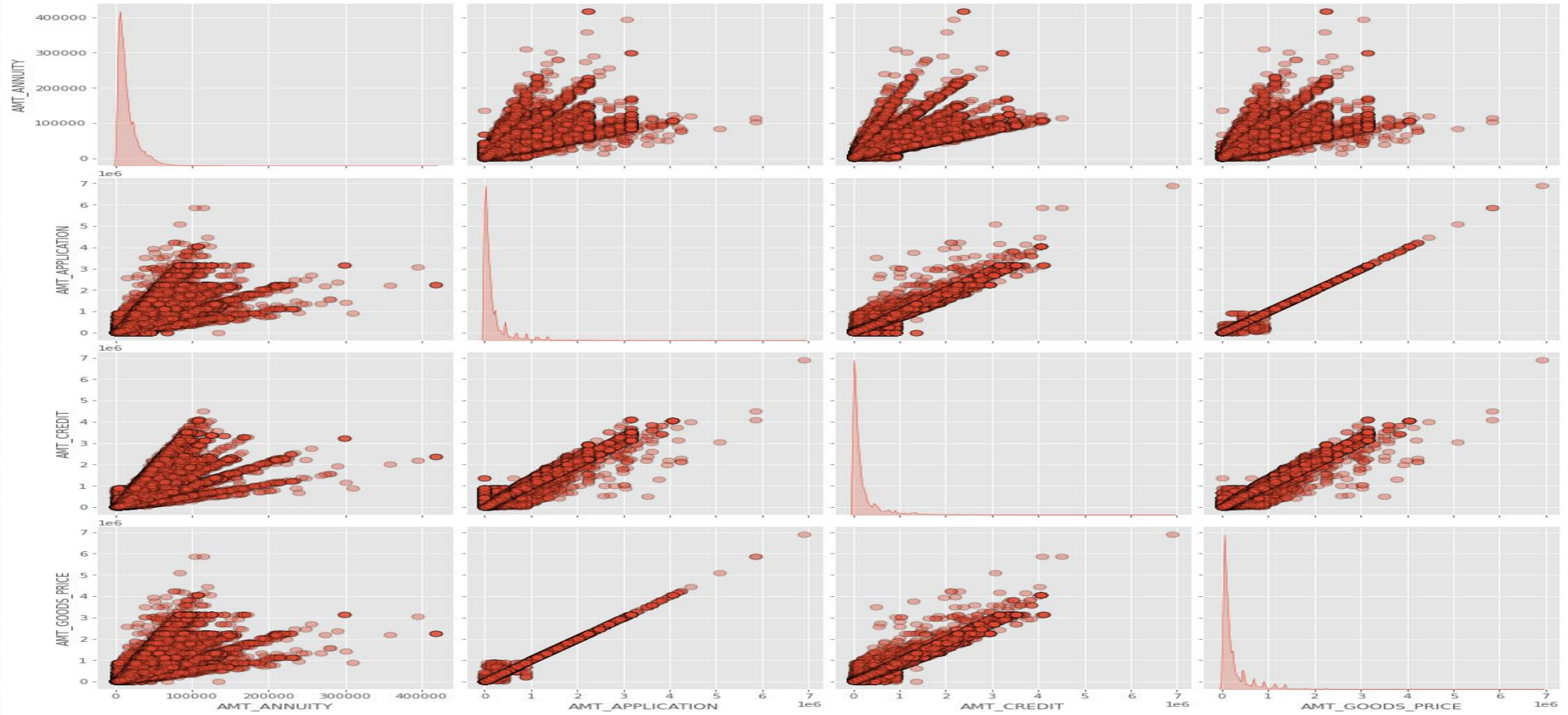


Previous Application Data Analysis

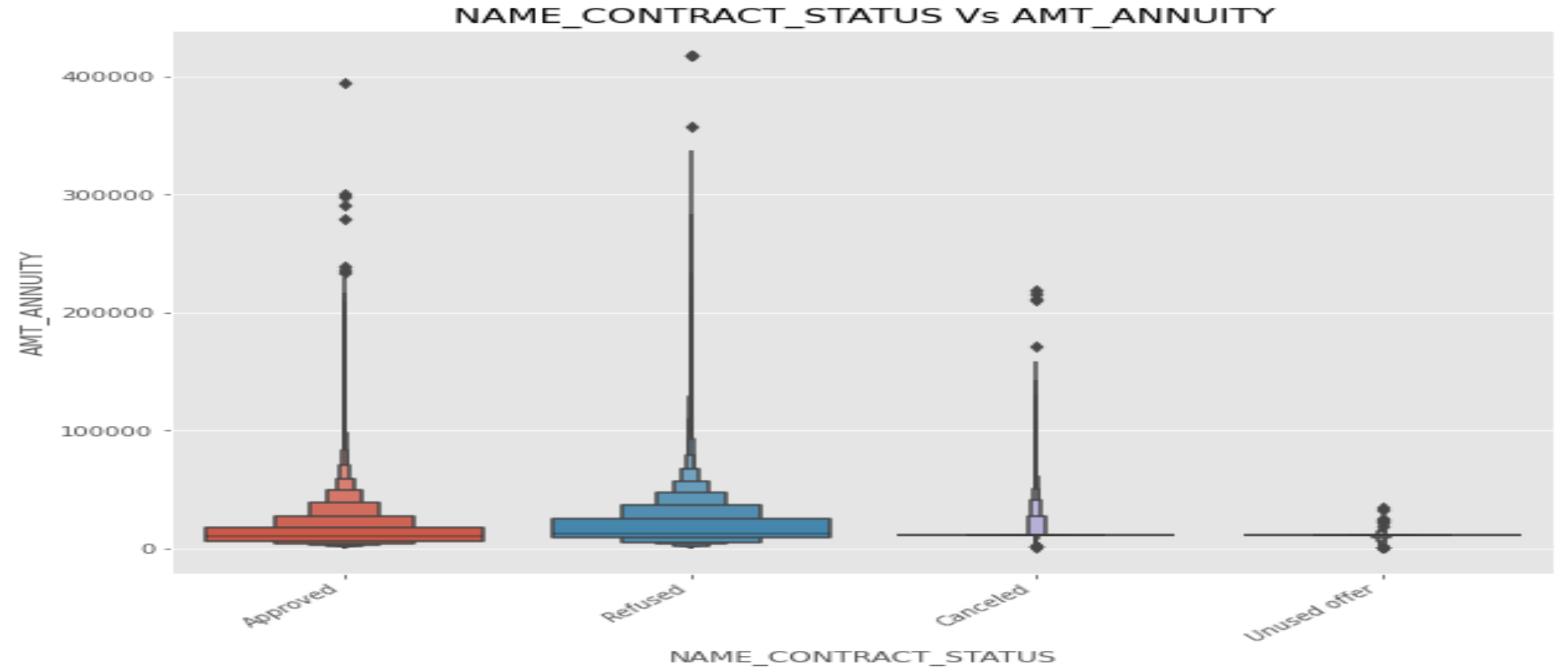



Distribution of NAME_PORTFOLIO
NAME_CONTRACT_STATUS
Approved
Refused
Canceled
Unused offer
Total Counts
600000
500000
400000
300000
200000
100000
0
POS
Cash
XNA
Cards
Cars
NAME_PORTFOLIO

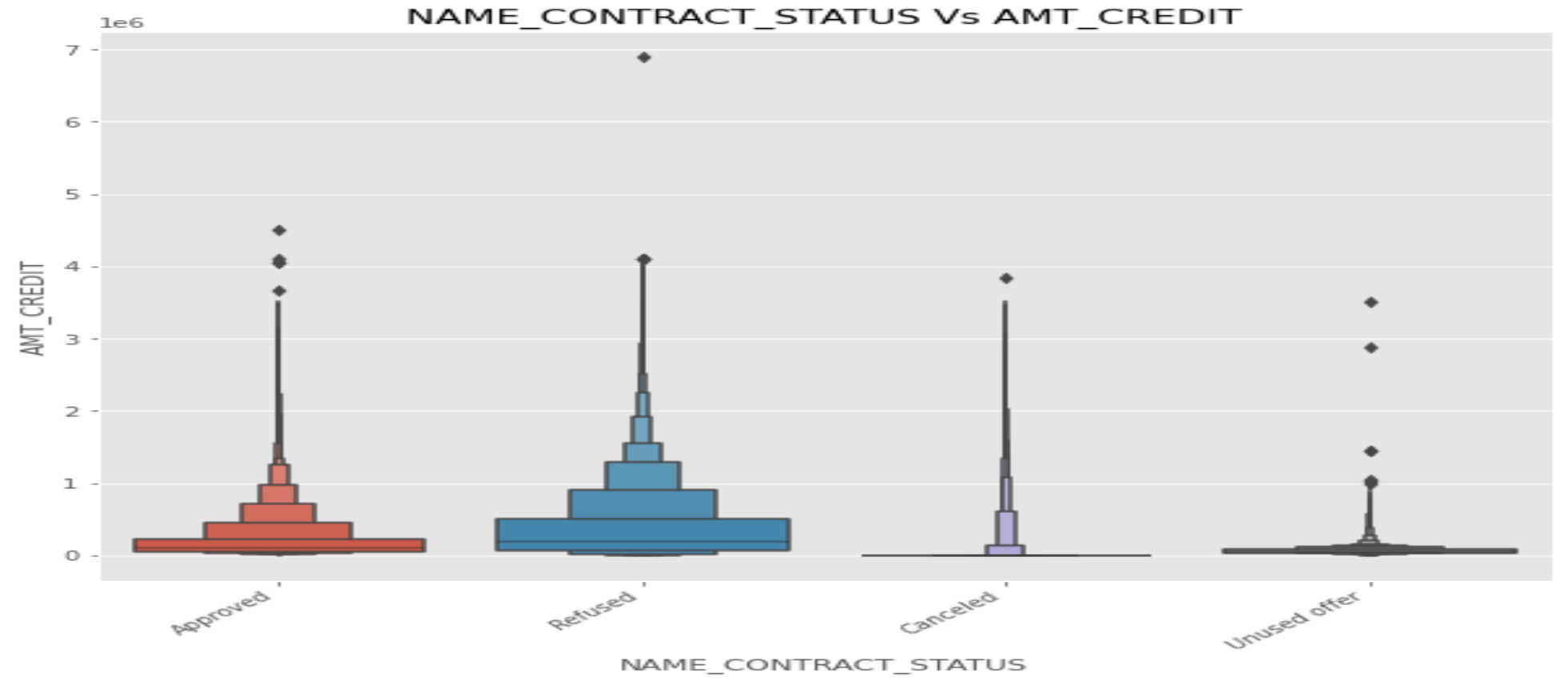
Previous Application Data Analysis



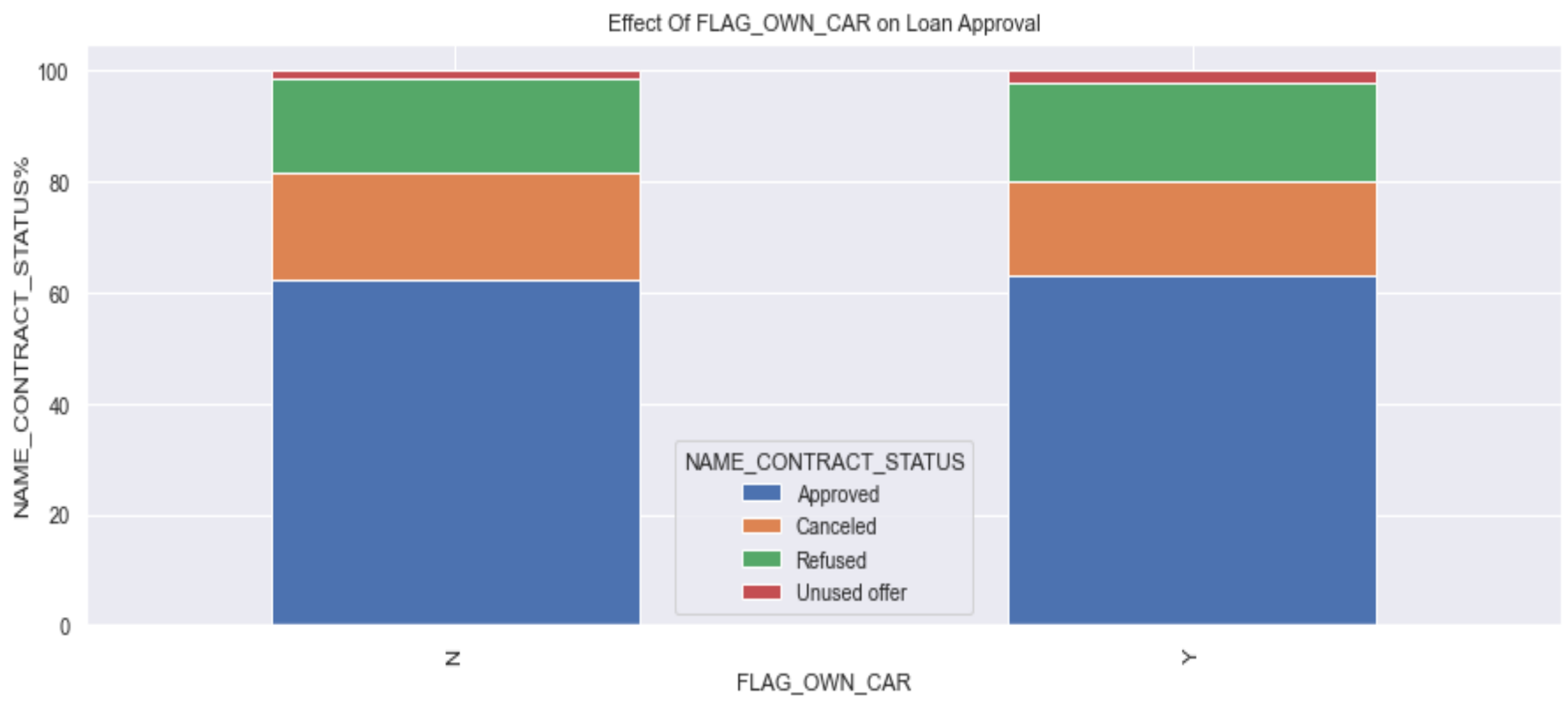
NAME_CONTRACT_STATUS' Vs 'AMT_ANNUITY



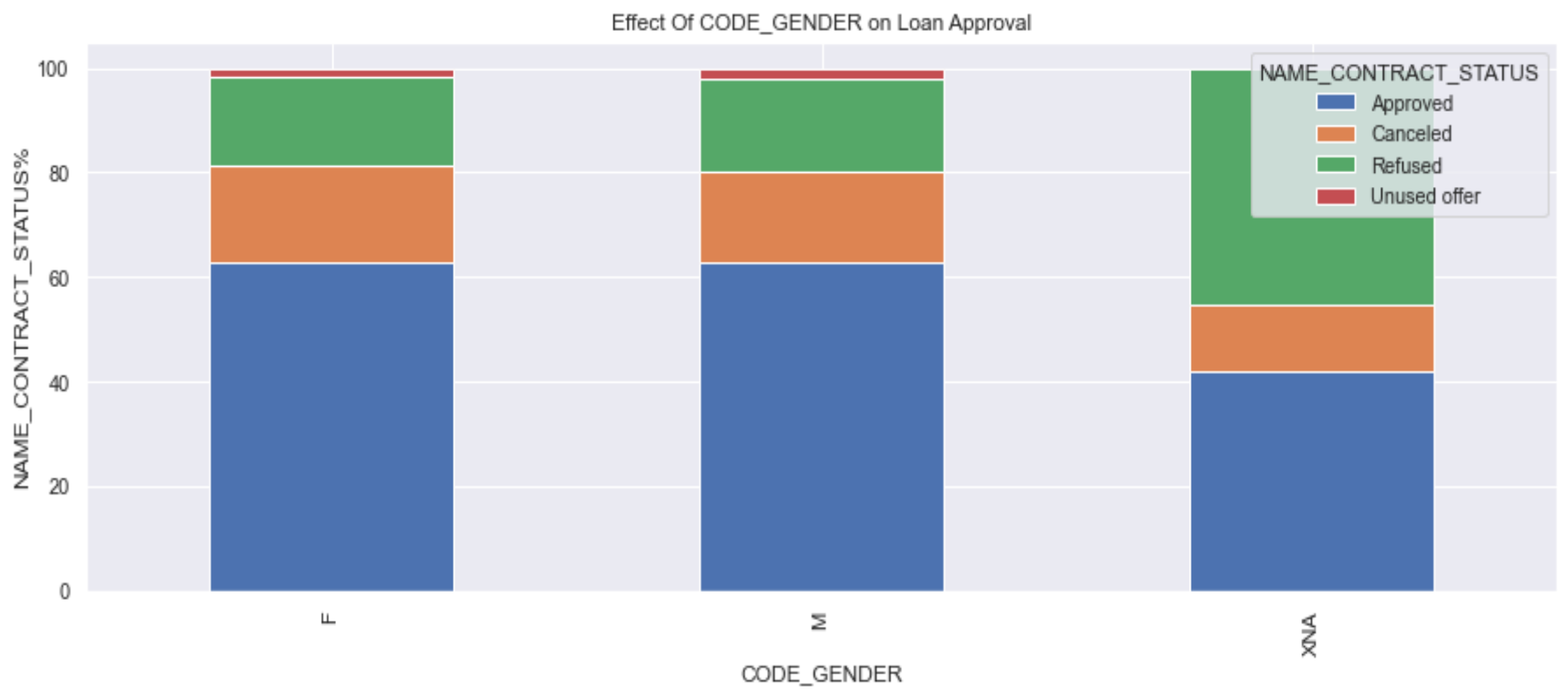
NAME_CONTRACT_STATUS' Vs AMT_CREDIT



Merging of the two data sets by a common column (SK_ID_CURR) for further comparison and analysis



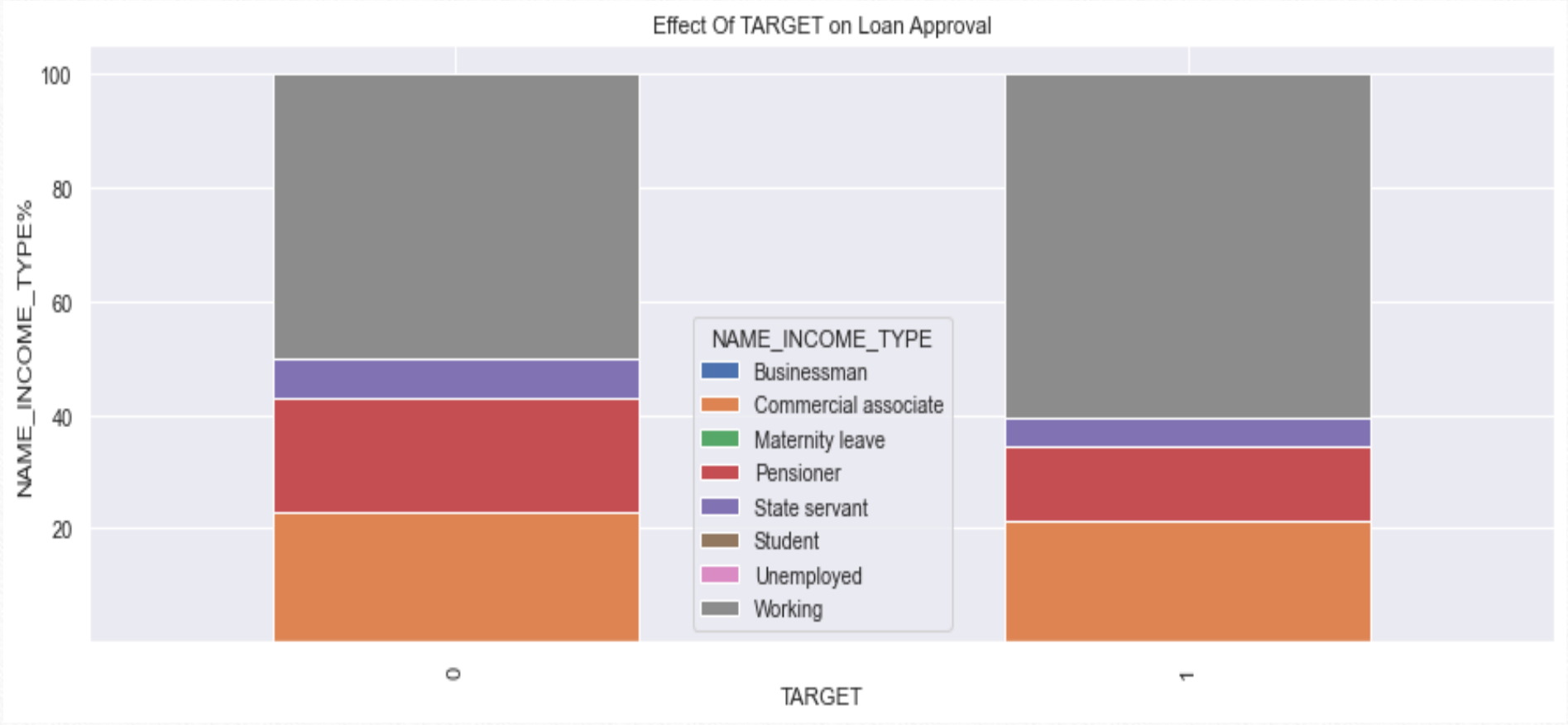
Merging of the two data sets by a common column (SK_ID_CURR) for further comparison and analysis



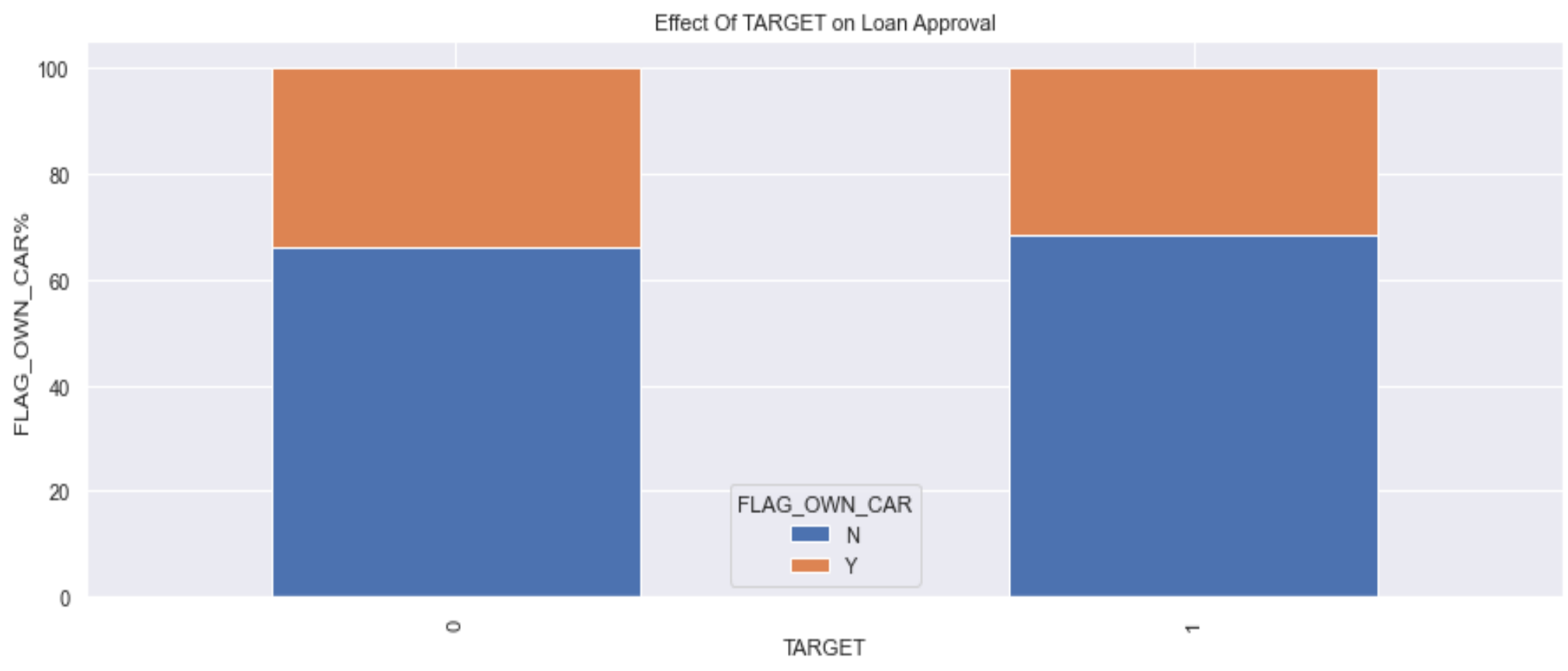
Merging of the two data sets by a common column (SK_ID_CURR) for further comparison and analysis

Merging of the two data sets by a common column (SK_ID_CURR) for further comparison and analysis

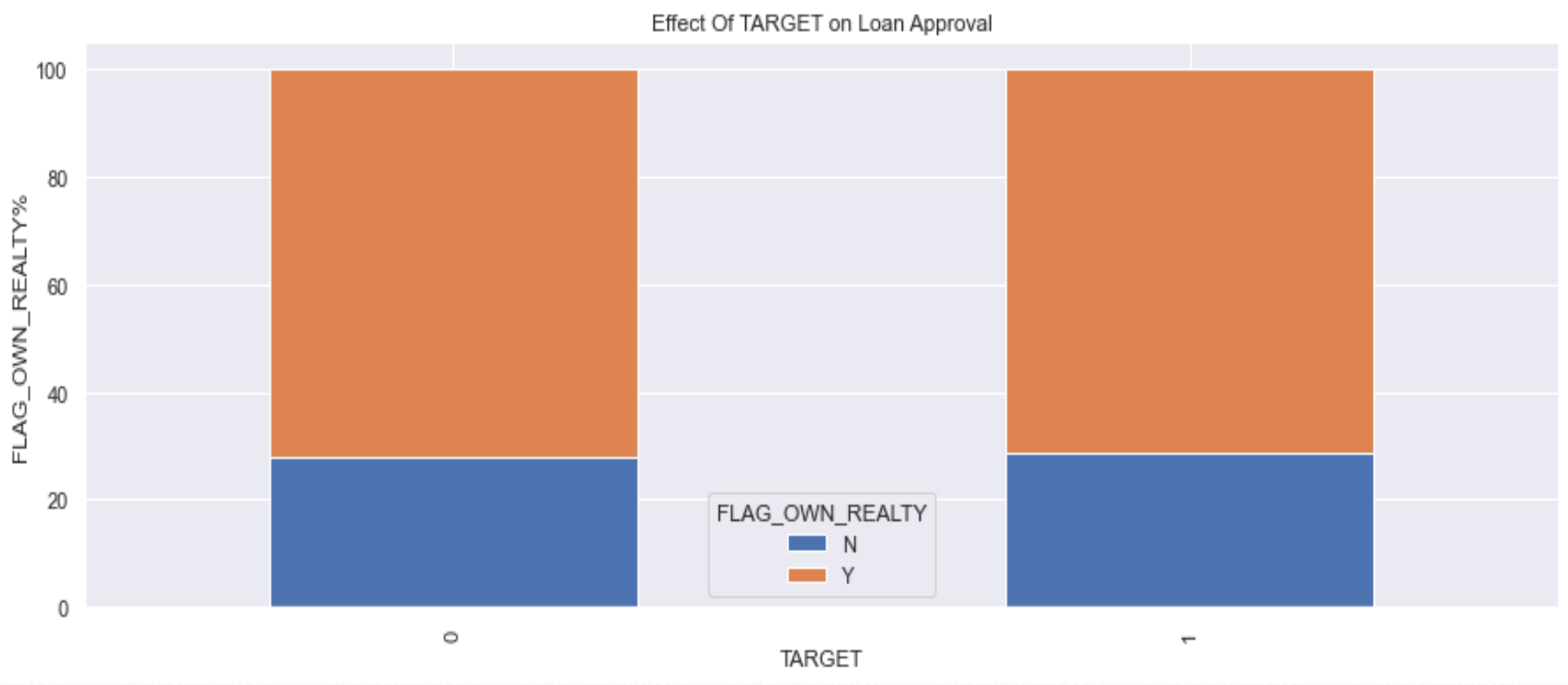
Merging of the two data sets by a common column (SK_ID_CURR) for further comparison and analysis



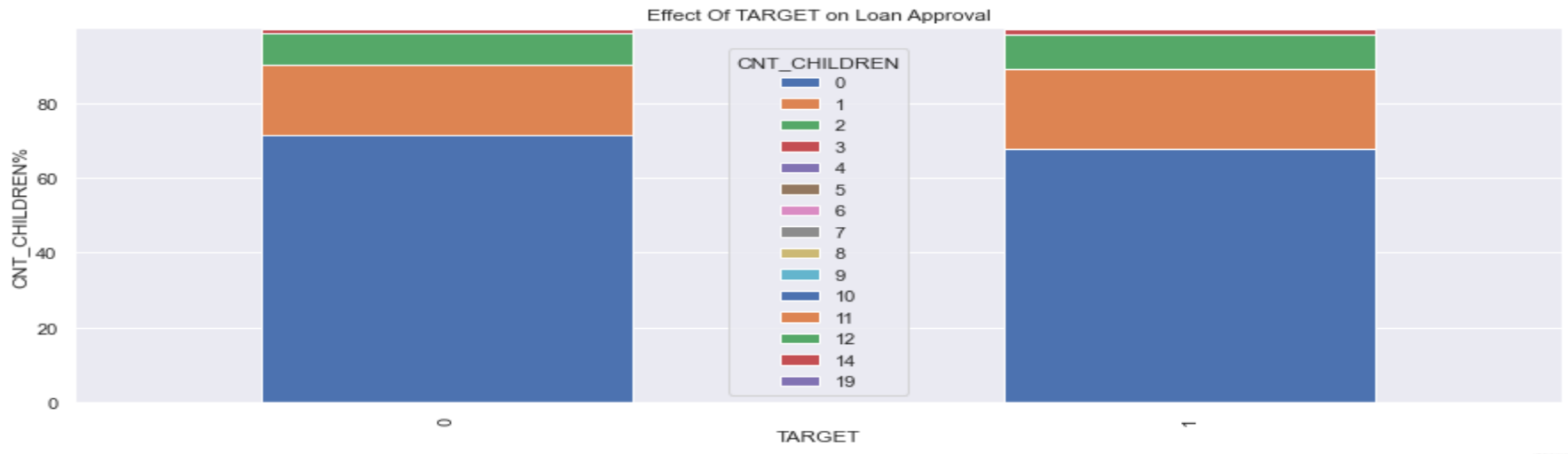
Merging of the two data sets by a common column (SK_ID_CURR) for further comparison and analysis

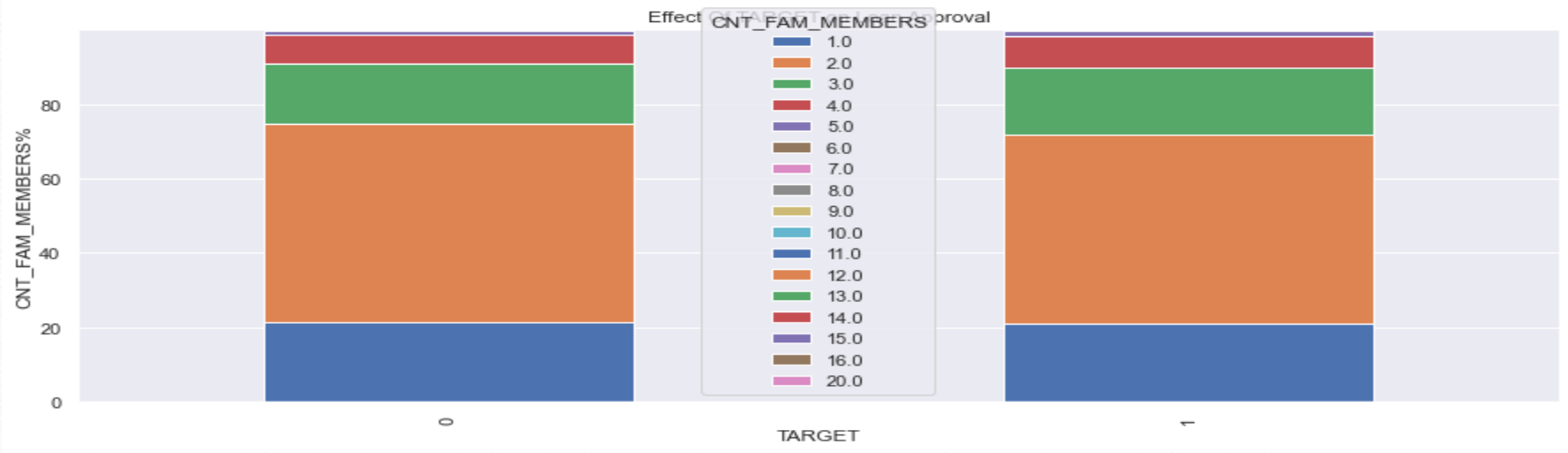


Merging of the two data sets by a common column (SK_ID_CURR) for further comparison and analysis

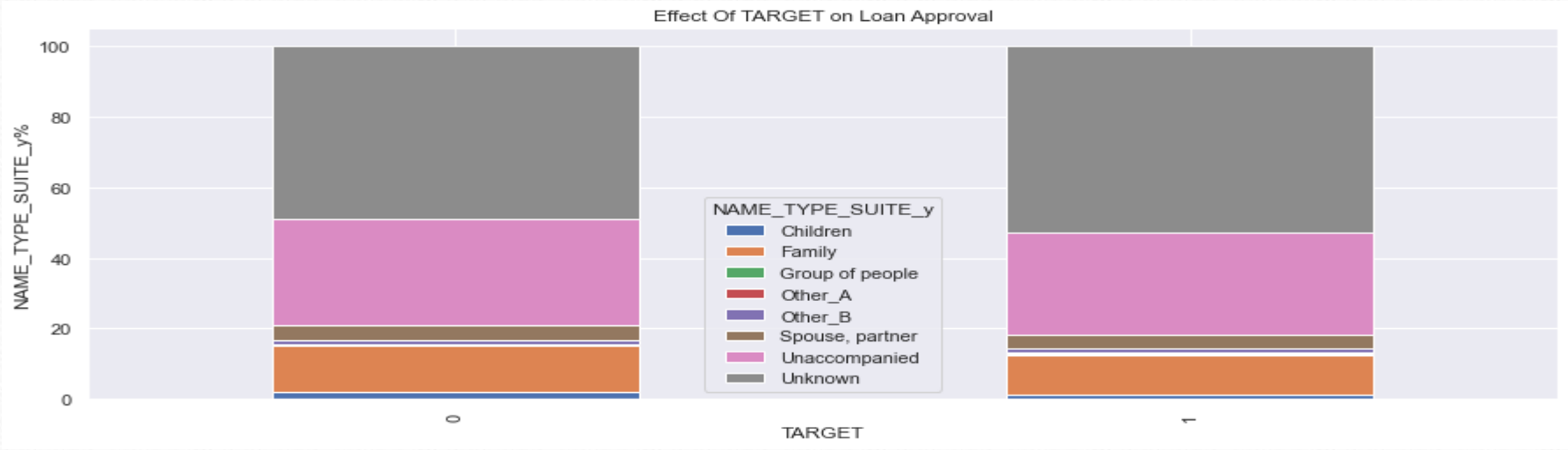


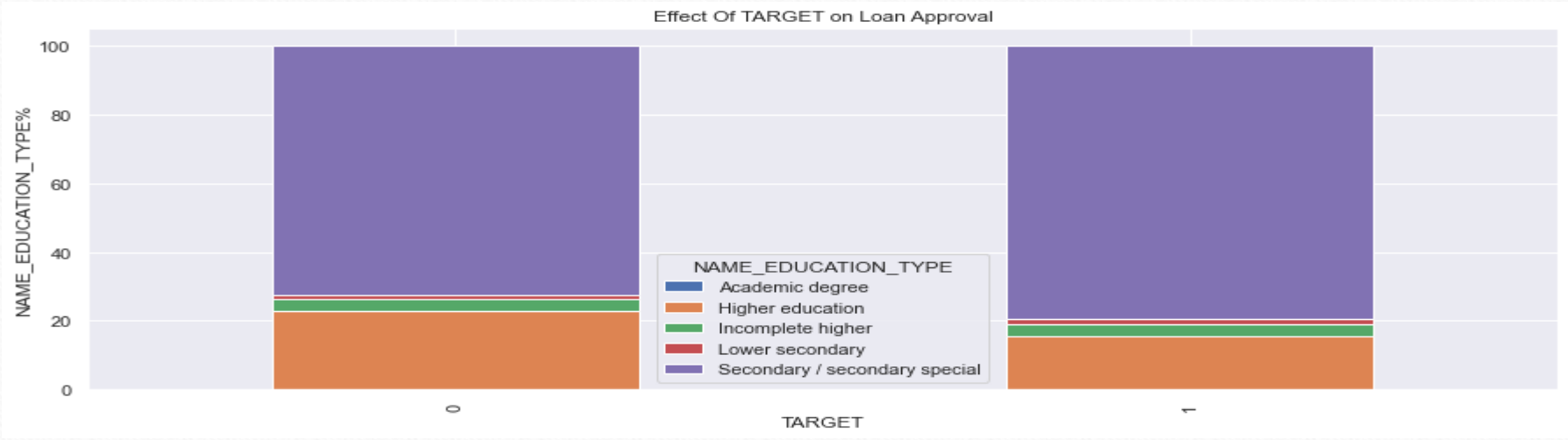
Merging of the two data sets by a common column (SK_ID_CURR) for further comparison and analysis



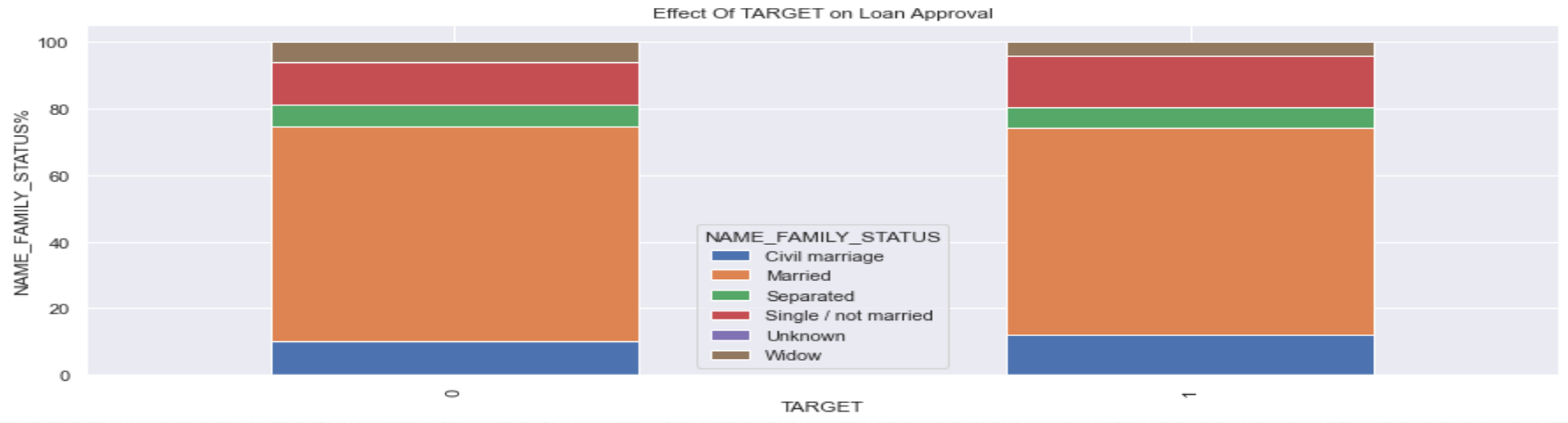


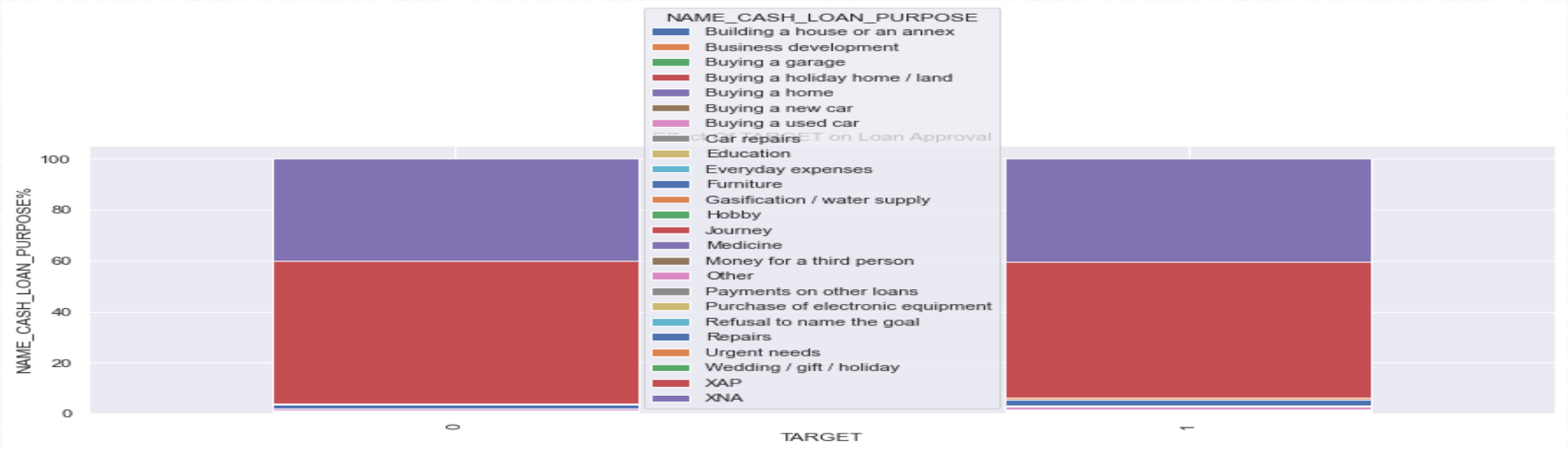
Merging of the two data sets by a common column (SK_ID_CURR) for further comparison and analysis



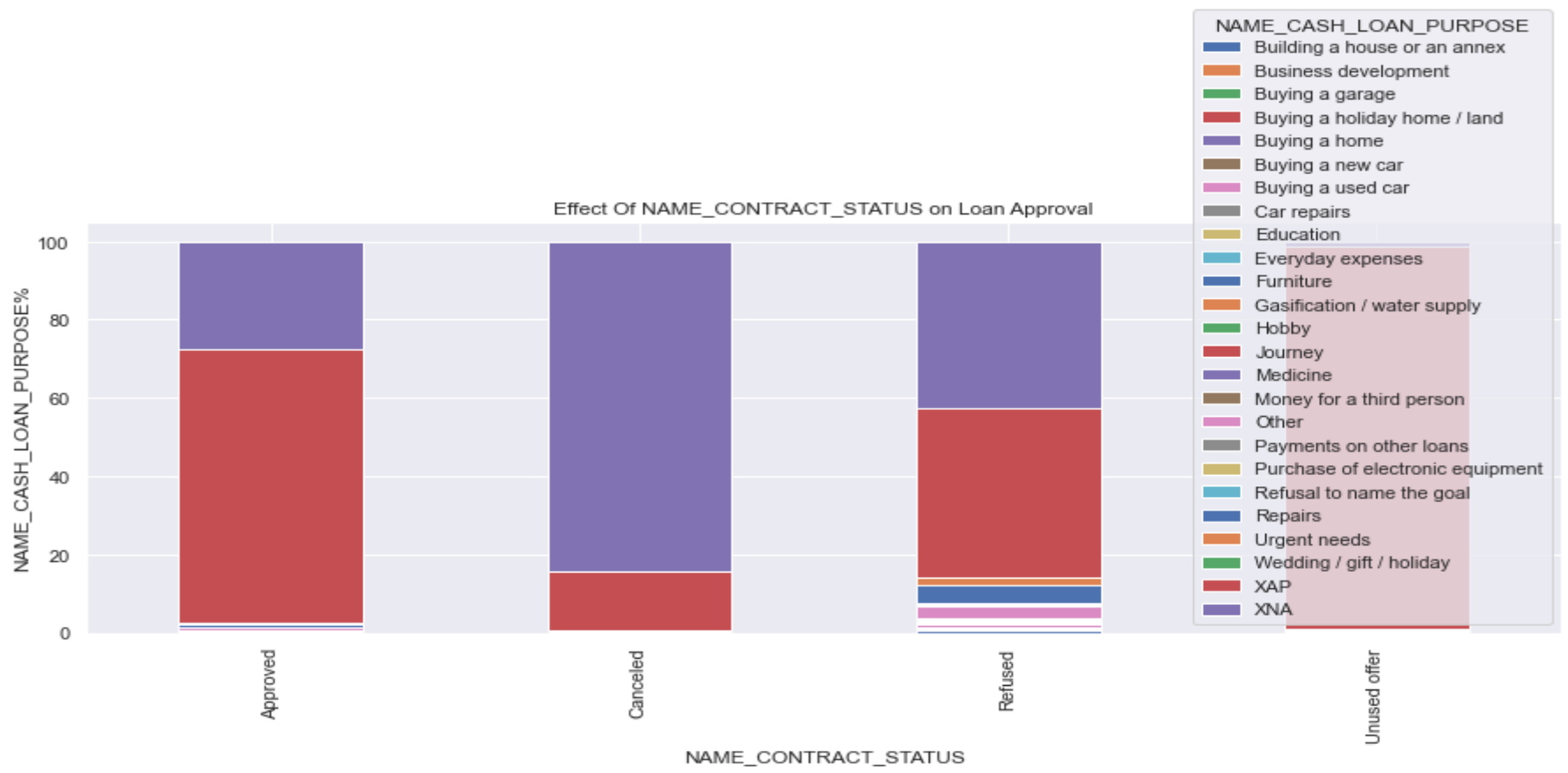


Merging of the two data sets by a common column (SK_ID_CURR) for further comparison and analysis

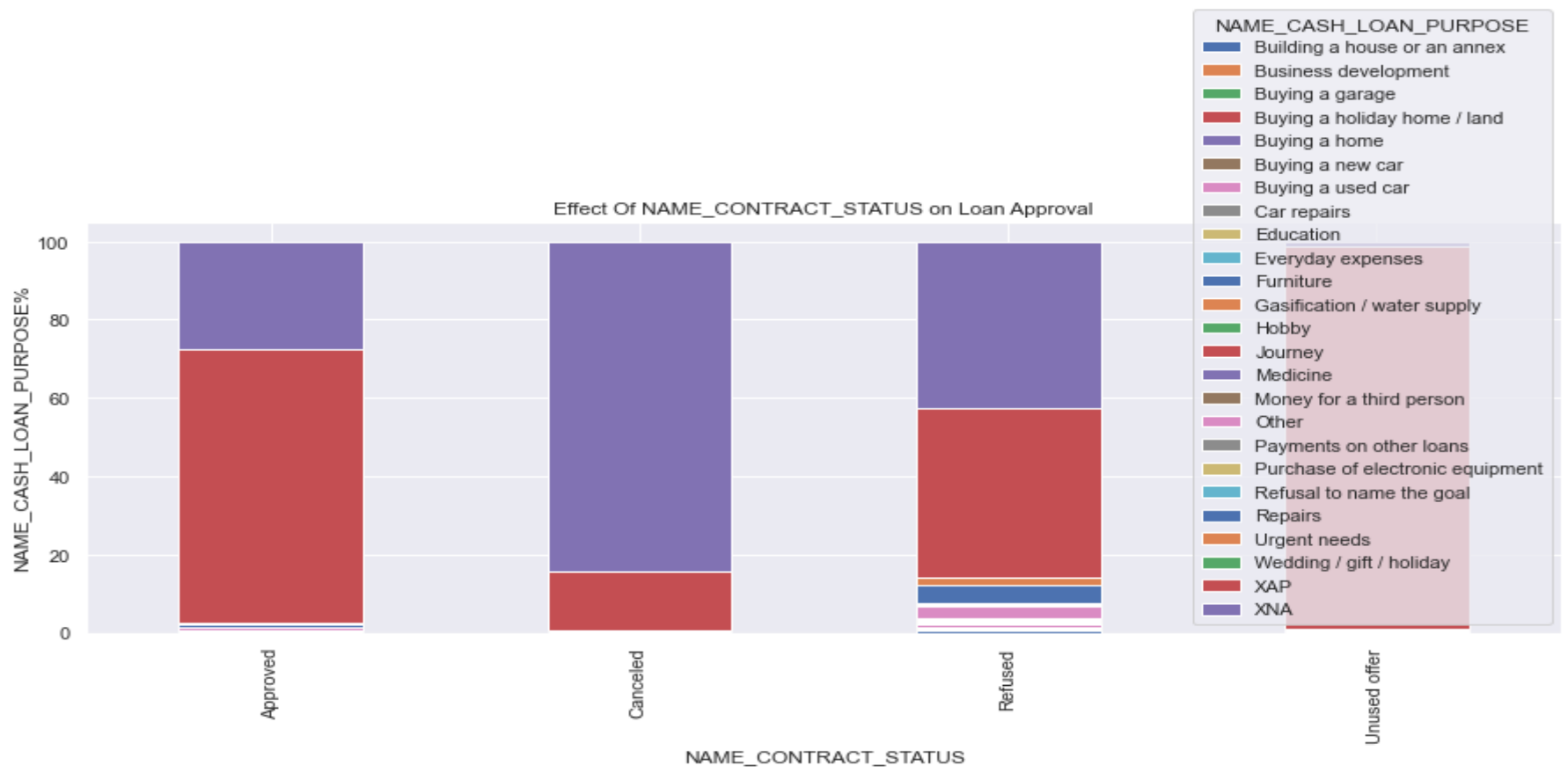




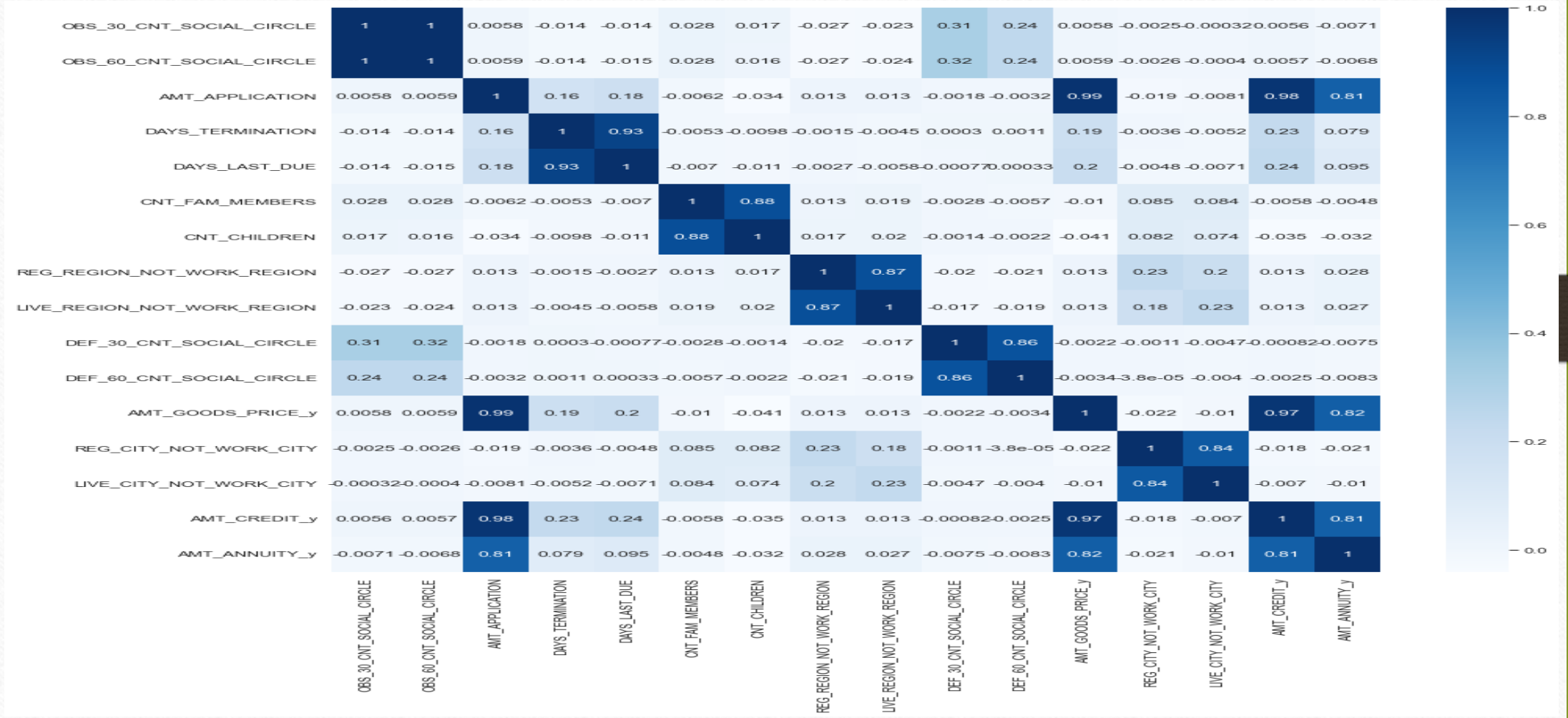
Merging of the two data sets by a common column (SK_ID_CURR) for further comparison and analysis



Merging of the two data sets by a common column (SK_ID_CURR) for further comparison and analysis



TOP Correlation variables



TOP Correlation variables

Inferences:

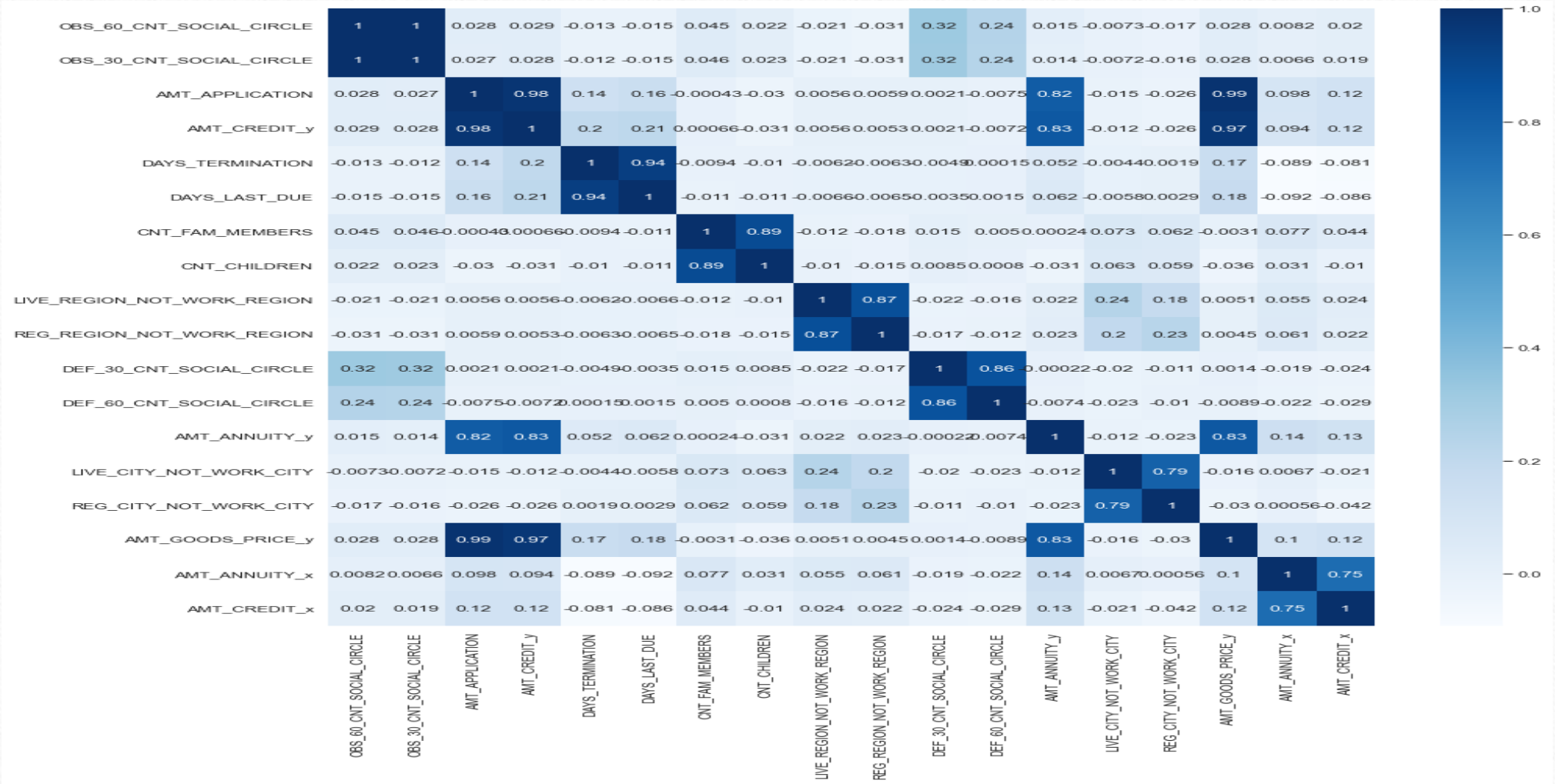
1.AMT_GOODS_PRICE and AMT_APPLICATION have a high correlation, which means the more credit the client asked for previously is proportional to the goods price that the client asked for previously.

2.AMT_ANNUITY and AMT_APPLICATION also have a high correlation, which means the higher the loan annuity issued, the higher the goods price that the client asked for previously.

3.First due of the previous application is highly correlated with Relative to the expected termination of the previous application

4.CNT_CHILDREN and CNT_FAM_MEMBERS are highly correlated which means a client with children is highly likely to have family members as well.

TOP Correlation variables



TOP Correlation variables

Inferences:

1.In comparison to the repayer heatmap, AMT_GOODS_PRICE and AMT_APPLICATION have a high correlation here as well, which means the more credit the client asked for previously is proportional to the goods price that the client asked for previously.

2.In comparison to the repayer heatmap, AMT_ANNUITY and AMT_APPLICATION also have a high correlation, which means the higher the loan annuity issued, the higher the goods price that the client asked for previously.

3.Higher the goods price, higher the credit by the client

4.First due of the previous application is highly correlated with Relative to the expected termination of the previous application (same as with the repayer heatmap)

5.CNT_CHILDREN and CNT_FAM_MEMBERS are highly correlated which means a client with children is higly likely to have family members as well (same as with the repayer heatmap)

Conclusion

- Clients who are Students, Pensioners and Commercial Associates with a housing type such as office/co-op/municipal apartments NEED TO BE TARGETED by the bank for successful repayments. These clients have the highest amount of repayment history.
- Clients who are working need to be targeted LESS by the bank as they have the highest amount of defaulters.
- Clients should NOT be targeted based on their education type alone as the data is very inconclusive.
- Banks SHOULD target clients who own a car.
- There are NO repayers/negligible repayers when the contract type is of revolving loan.
- 'Repairs' purpose of loan is the one with the most defaulters and repayers. Therefore, clients with very low risk SHOULD be given loans for such purpose to yield high profits.
- Banks SHOULD also target female clients as they are the highest repayers (almost as double as males) amongst both the genders.