

“Credit EDA Case Study”

By SUDHEER.N.POOJARI



Problem Statement:

The loan-providing companies find it hard to give loans to people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming defaulters.

When the bank receives a loan application there are two types of risks associated with the bank's decision:


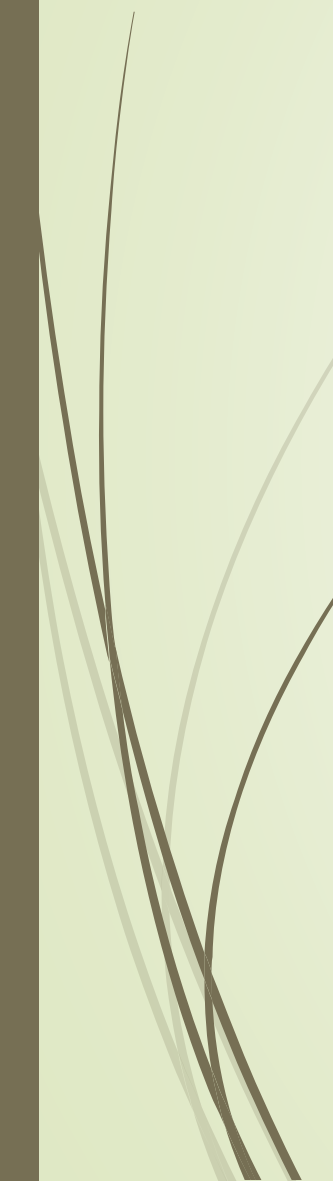
- ❖ If the applicant is likely to repay the loan, then not approving the loan result in a loss of business to the company.
- ❖ If the applicant is not likely to repay the loan, I.e.he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Business Objective:

- ❖ The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.
- ❖ So the company can utilize this knowledge for its portfolio and risk assessment.

OVERALL APPROACH TO EXECUTE CASE STUDY

- ❑ Start by importing the 'application_data' set.
- ❑ Checking the structure of the data(Normal routine check) like 'head','dtypes','info', and 'describe' functions.
- ❑ Data quality check and missing values.
- ✓ Will find the percentage of missing values for all the columns in the application data set.
- ✓ Will remove the column which contains more than 40 % missing value.
- ✓ For the column which has a missing percentage value of around 13%, then we will check the best metric to impute the missing values.
- ✓ If the column is numerical, then we impute the missing value with mean, median or mode value.
- ✓ If the column is categorical, then we impute the missing value mode of the column.
- ✓ Imputation process has checked for five variable by plotting box plot: 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'NAME_TYPE_SUITE', 'CNT_FAM_MEMBERS', 'EXT_SOURCE_2'.
- ✓ Will check the data types of all the columns and change the datatype like negative value(age & date) to positive by using abs() and object to int64,float so on.
- ✓ For numerical columns will check for outliers by plotting box plot and inference the observation. Have checked for ('REGION_POPULATION_RELATIVE','CNT_CHILDREN','AMT_CREDIT','AMT_GOODS_PRICE','AMT_ANNUITY').
- ✓ Binning of a continuous variable into different categories value. Binnig has done for INCOME_VAL and AGE columns.

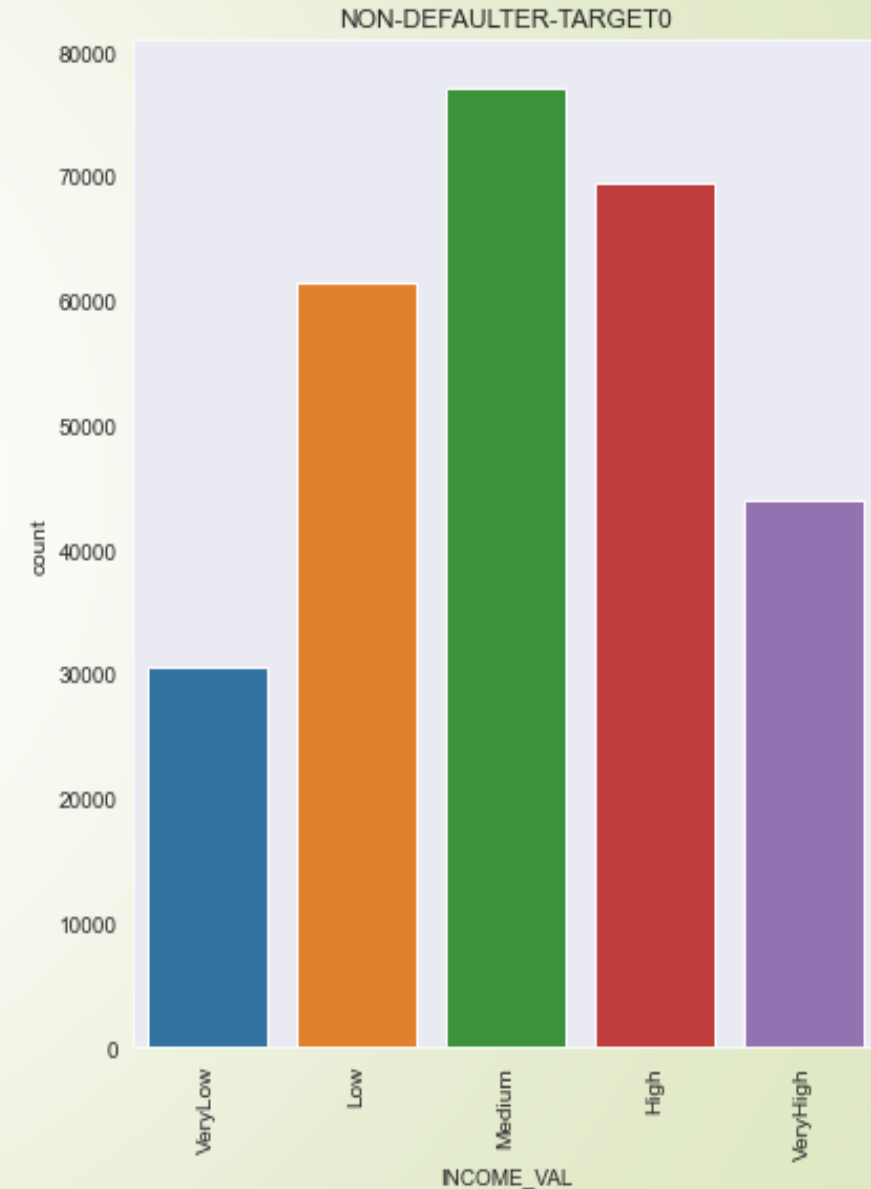
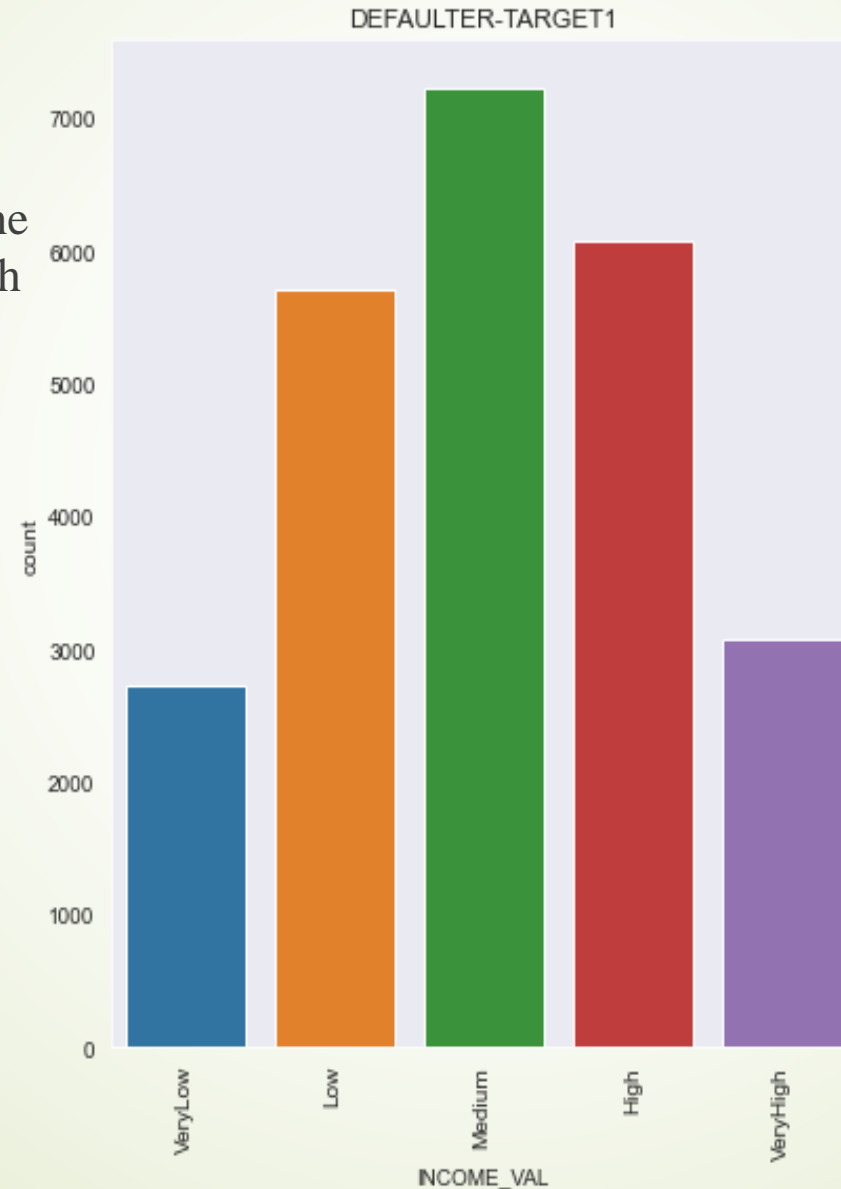
- 
- 
- ☐ Data quality check and missing values.
 - ✓ Checking imbalance percentage between Defaulter vs Non-Defaulter value in application data set.
 - ✓ Divide the application data into two subsets, i.e. Target=1 and Target=0.
 - ✓ Will perform the univariate analysis for the categorical variable for both Target=1 and Target=0 and compare the target variable across categories of a categorical variable using a count plot.
 - ✓ Will check for correlation for a numerical column for both the case, i.e. 0 and 1.
 - ✓ Will find the variable with the highest correlation is the same in both file or not.
 - ✓ Will perform the univariate analysis for the numerical variable for both Target=1 and Target=0 and compare the target variable across categories of a continuous variable using a dist plot.
 - ✓ Performing bivariate analysis for the numerical variable for both Target=1 and Target=0, the analysis included categorical-categorical, numerical-numerical, and categorical-numerical, plot analysis.
 - ☐ Import the previous_application data and will do the normal routine check like data type conversion and finding, replacing missing value.
 - ☐ Will merge the two files application_data and previous_data for final analysis purposes.
 - ☐ We do the same univariate and bivariate analysis for the merge file and will try to get some patterns and insights from the merge data.

UNIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES

FOR TARGET 0 & 1

OBSERVATION:

- **INCOME_VAL (Insights):** From the plot we conclude that the Very High income group tend to default less often. While compare to the total number of Non-defaulter.

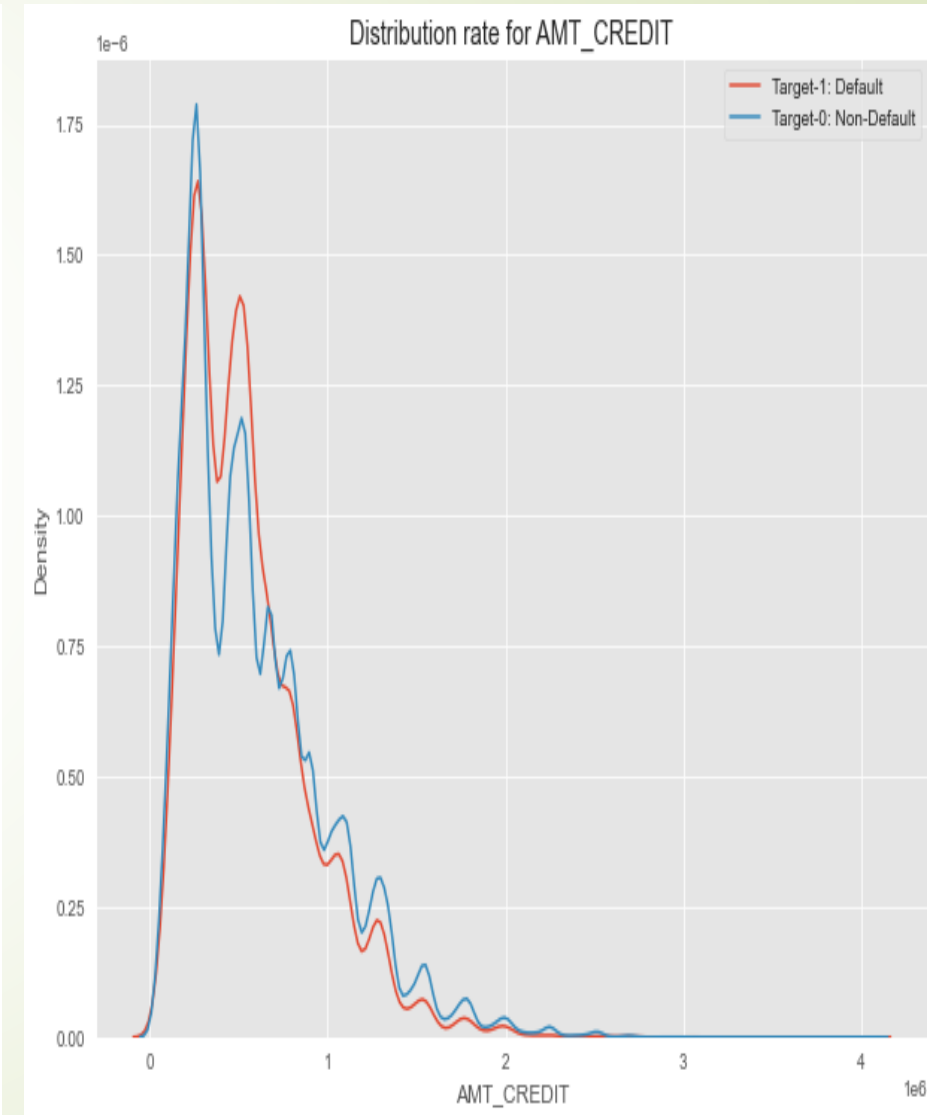
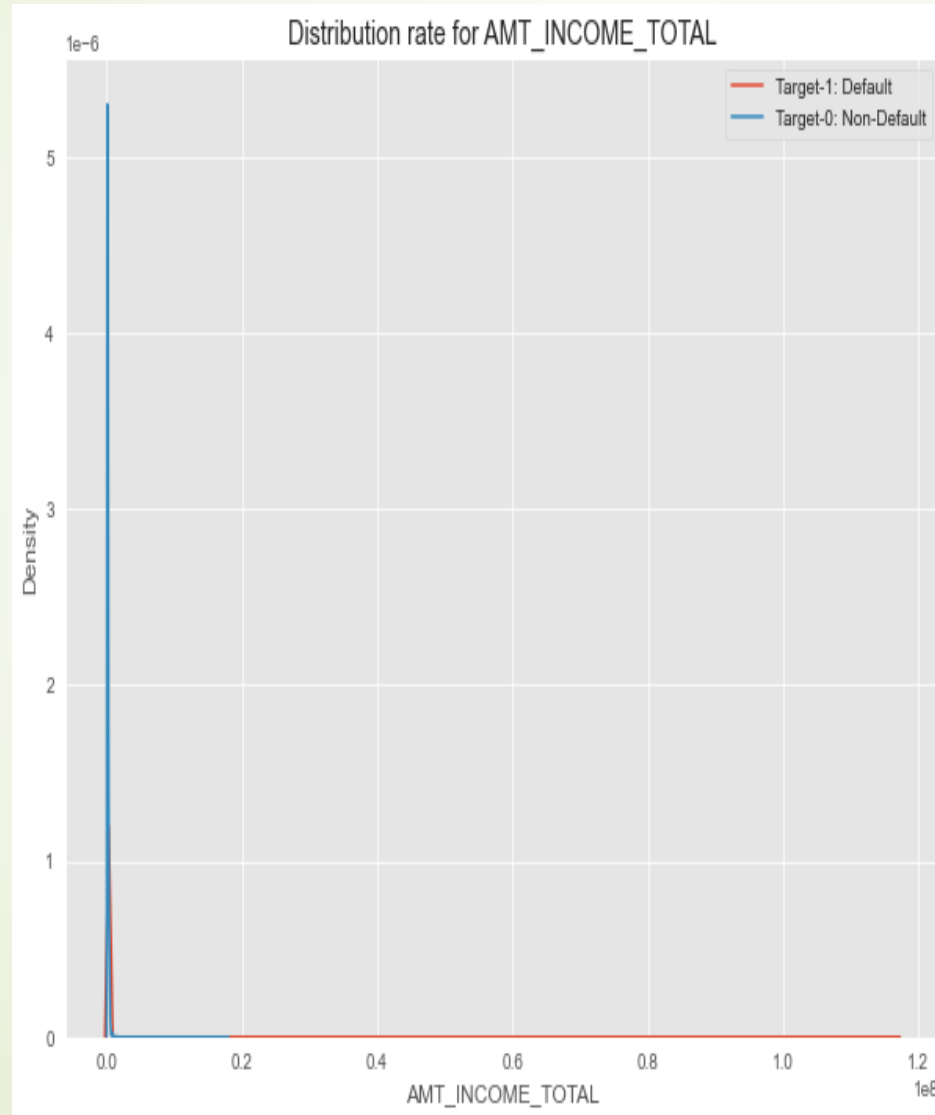


UNIVARIATE ANALYSIS FOR CONTINUOUS VARIABLES FOR TARGET 0 & 1

OBSERVATION:

As from the plotted graph following inferences we can conclude:

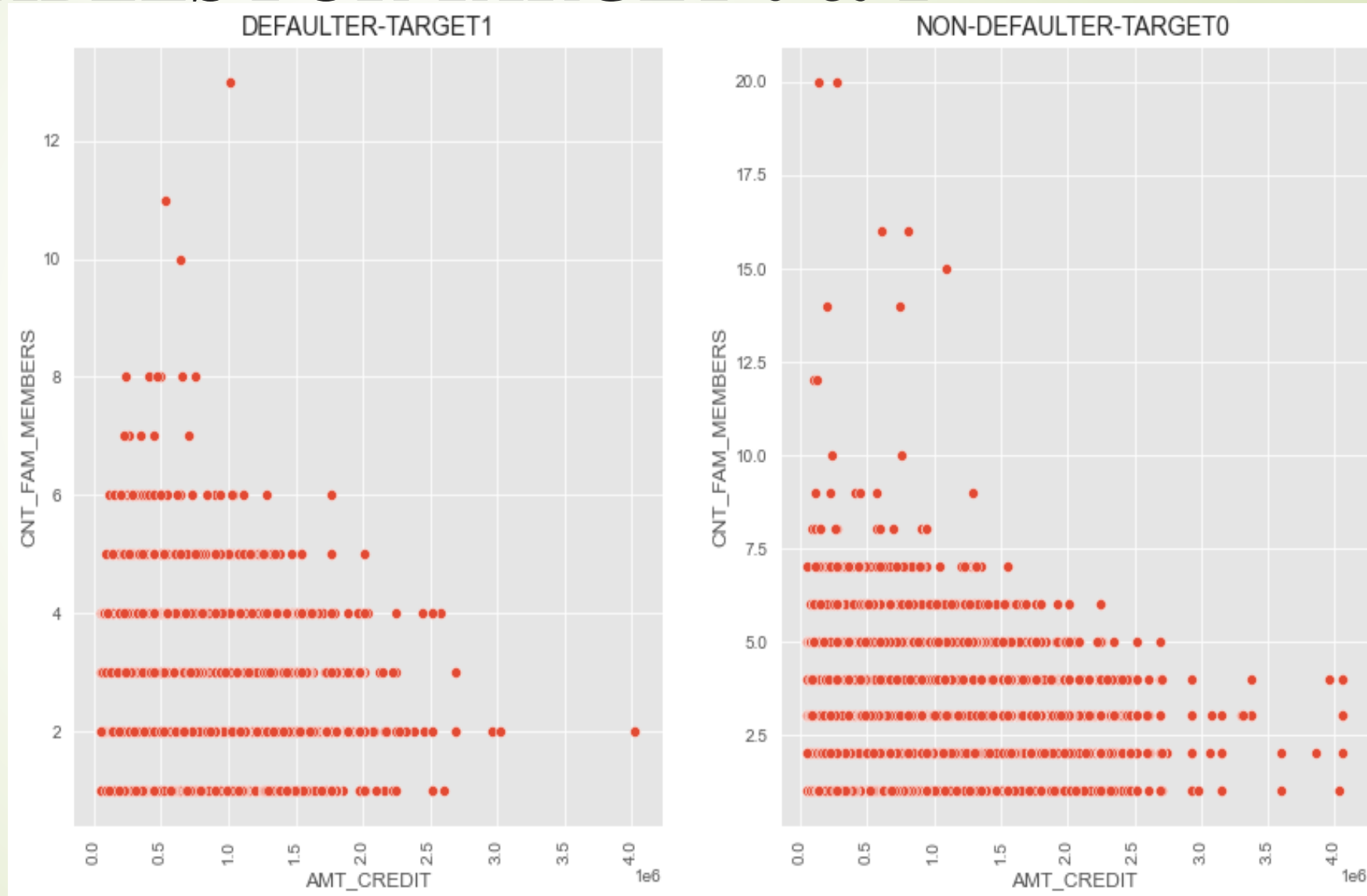
1. People with lower total income are more likely to take loan or default.
2. People with lower credit amount are more likely to default.



BIVARIATE ANALYSIS FOR CONTINUOUS VARIABLES FOR TARGET 0 & 1

OBSERVATION:

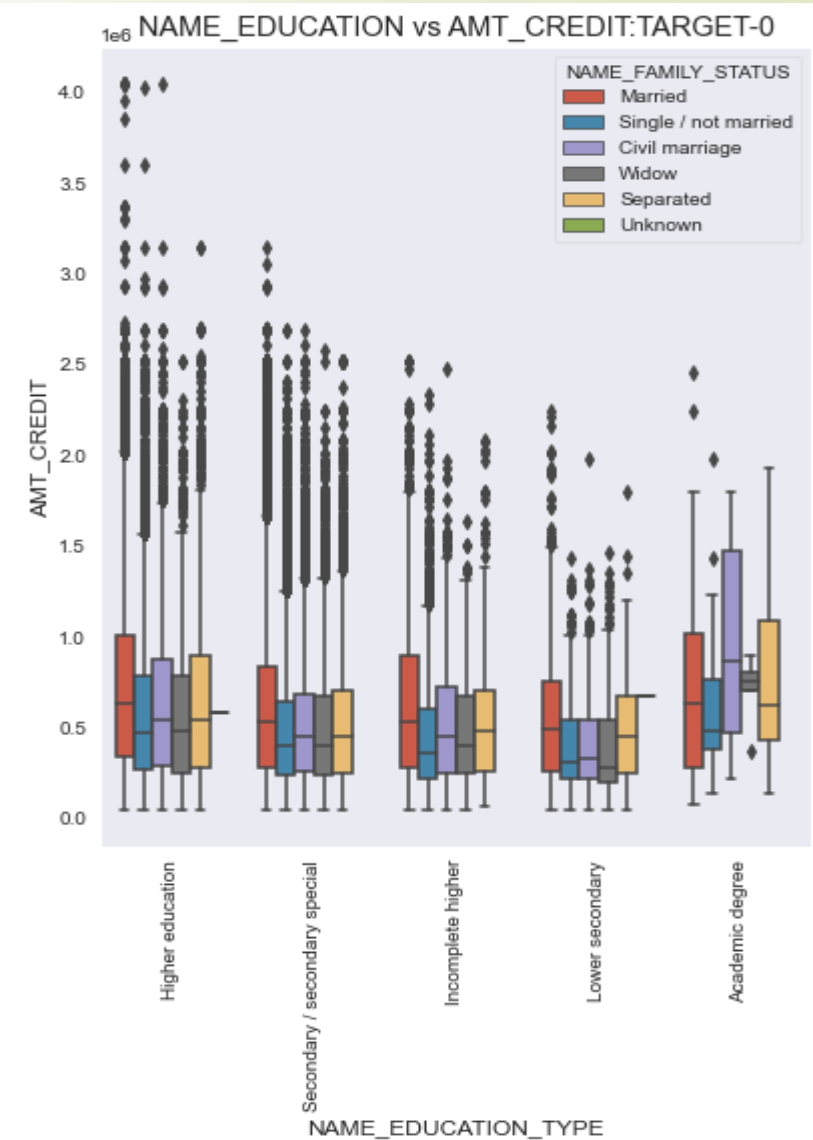
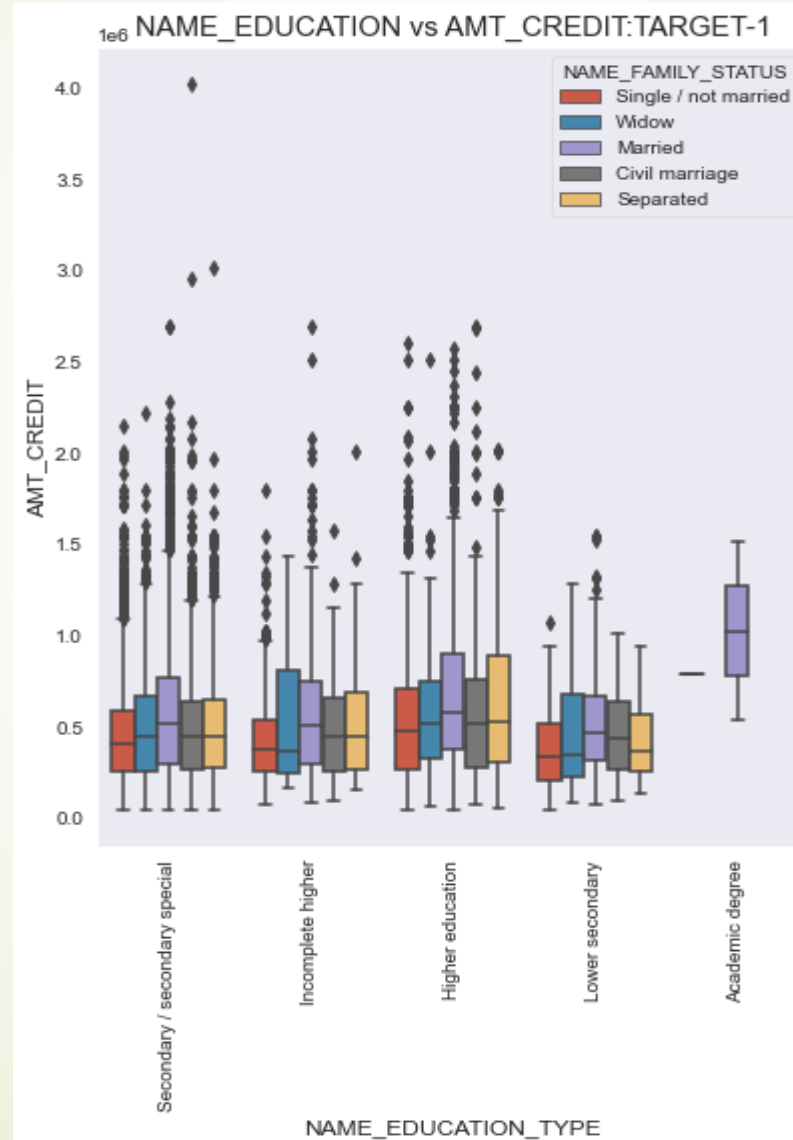
- From the graph, we can see in both plot, the density in the lower left corner is similar. So we can conclude that people are equally likely to default if the amount of credit and family is low also larger families and people with larger AMT_CREDIT default less often.



BIVARIATE ANALYSIS FOR CONTINUOUS & CATEGORICAL VARIABLES FOR TARGET 0 & 1

OBSERVATION:

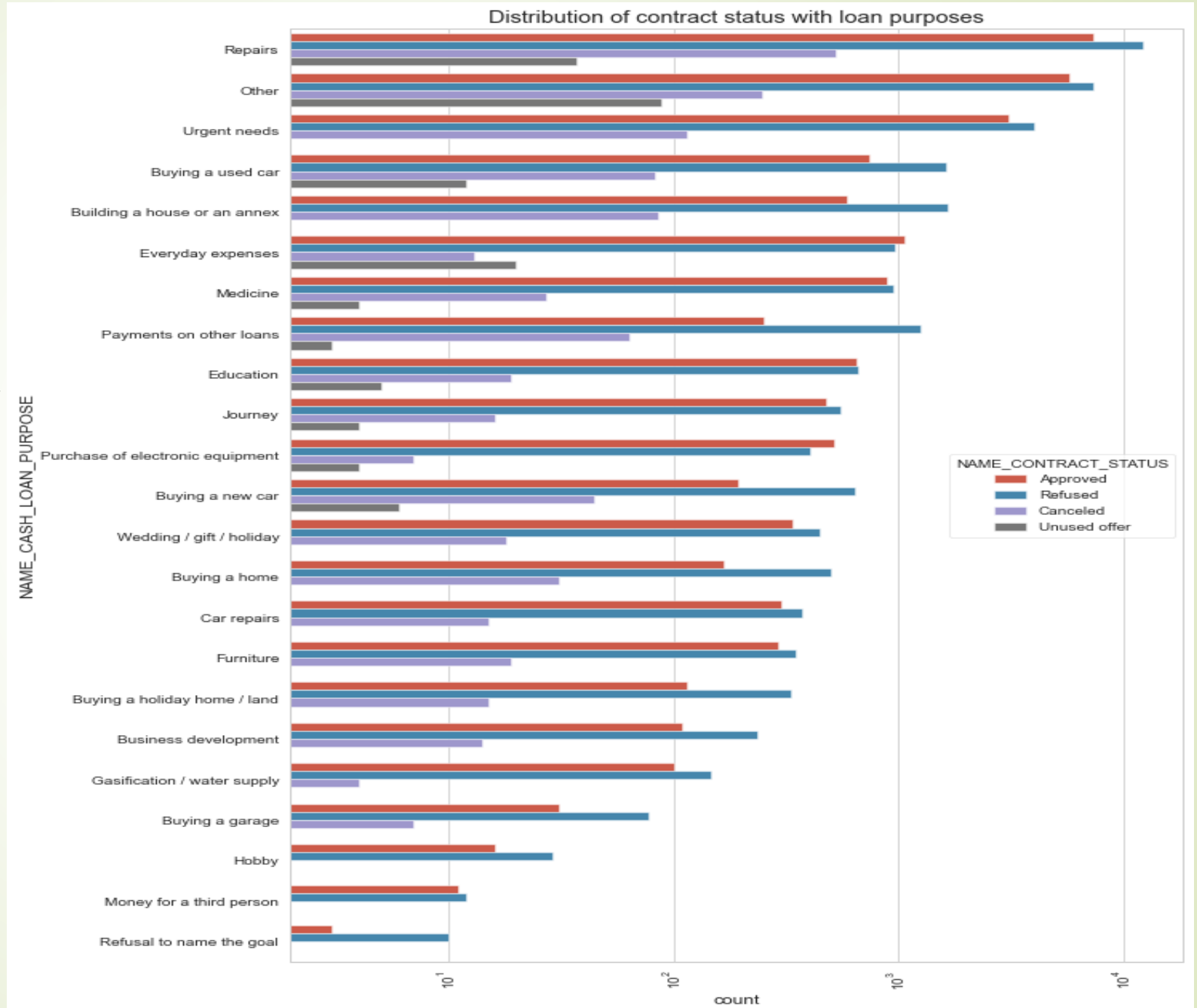
- From the box plot, we can conclude that the Family status of 'civil marriage', 'marriage', and 'separated' of Academic degree education are having a higher number of credits than others. Also, higher education of family status of 'marriage', 'single', and 'civil marriage' are having more outliers. Civil marriage for an Academic degree is having most of the credits in the third quartile.



UNIVARIATE ANALYSIS AFTER MERGING PREVIOUS DATA

OBSERVATION:

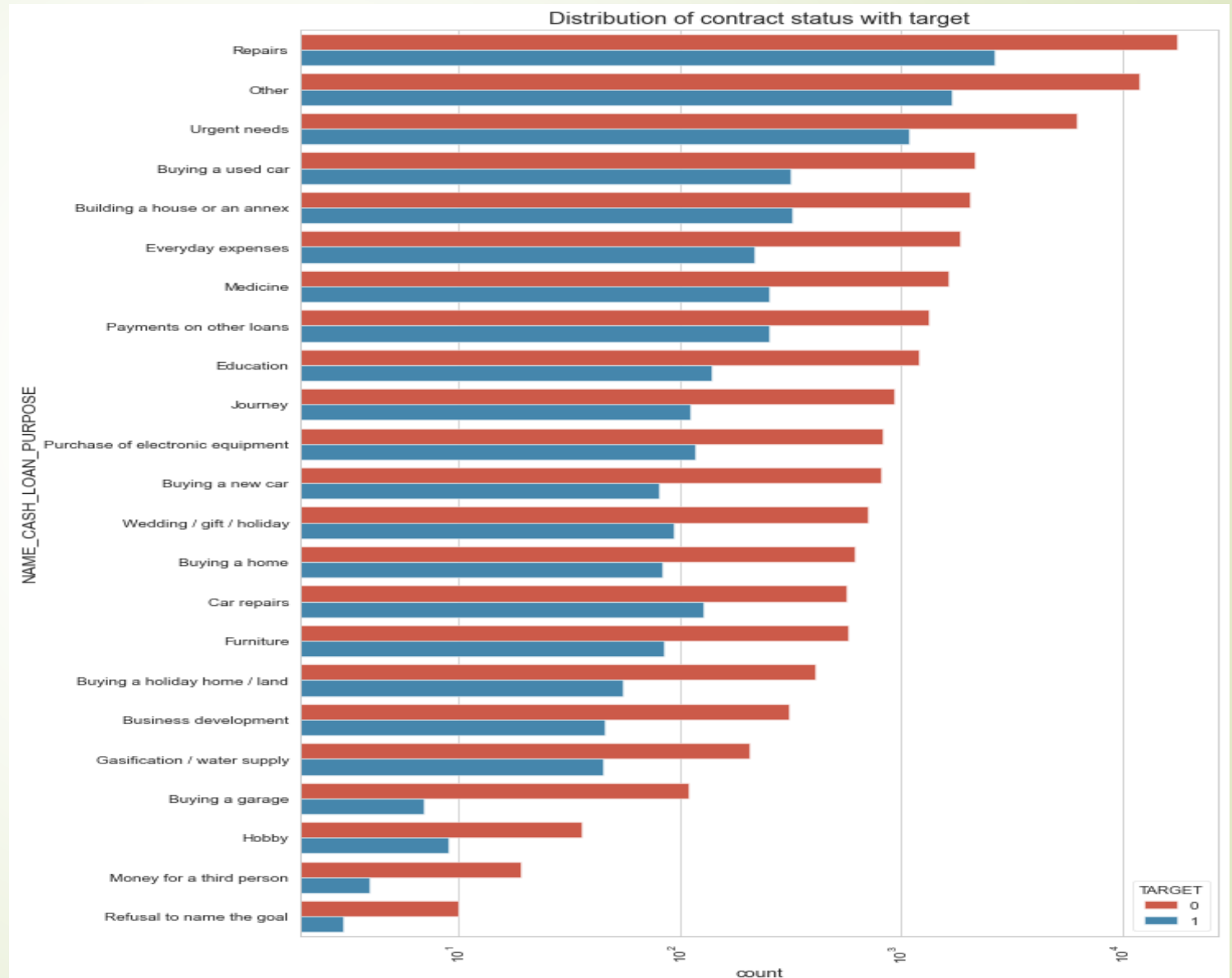
1. From the plot we can say that education purpose loans have equal numbers of approval and refusal.
2. Most loan applications are rejected by the people of purpose 'repairs'.
3. Paying in other loans and buying a new car is having significantly higher rejections than approval.



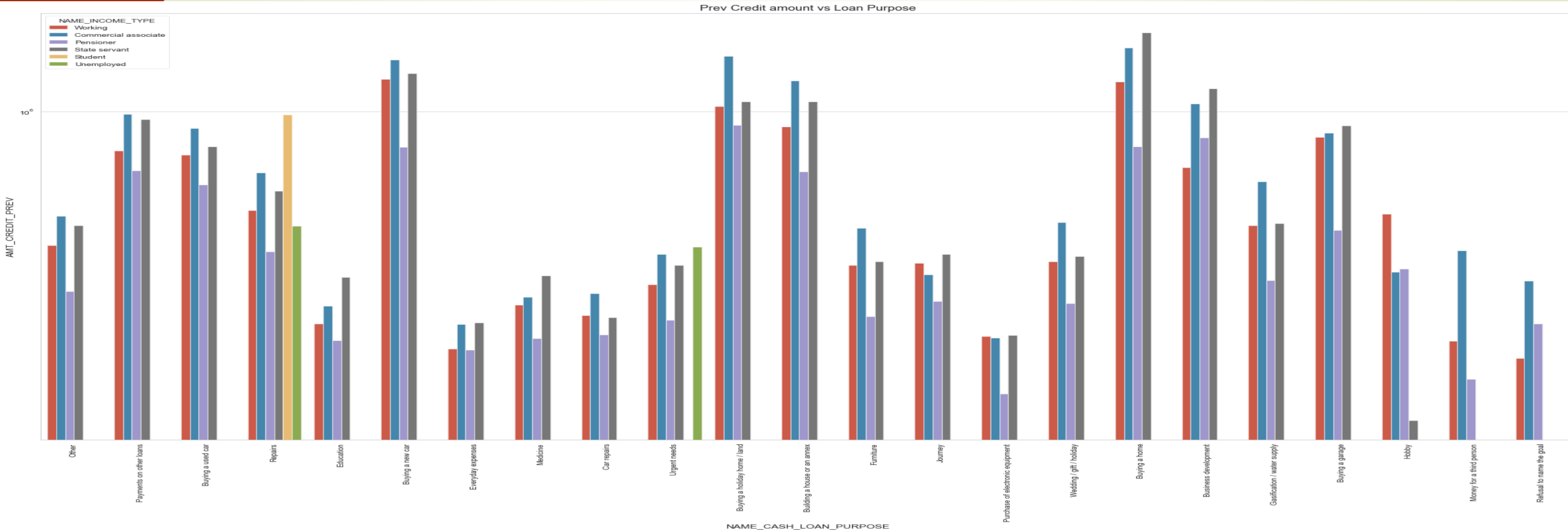
DISTRIBUTION OF TARGET VARIABLE WITH COLUMN LOAN PURPOSE

OBSERVATION:

1. There are few places where loan payment is significantly higher than facing difficulties. They are 'Furniture', 'Buying a garage', 'Business development', 'Buying land', 'Buying a new car' and 'Education'. Hence we can focus on these purposes for which the client is having minimal payment difficulties.
2. Loan purposes with 'Repairs' are facing more difficulties on-time payment.



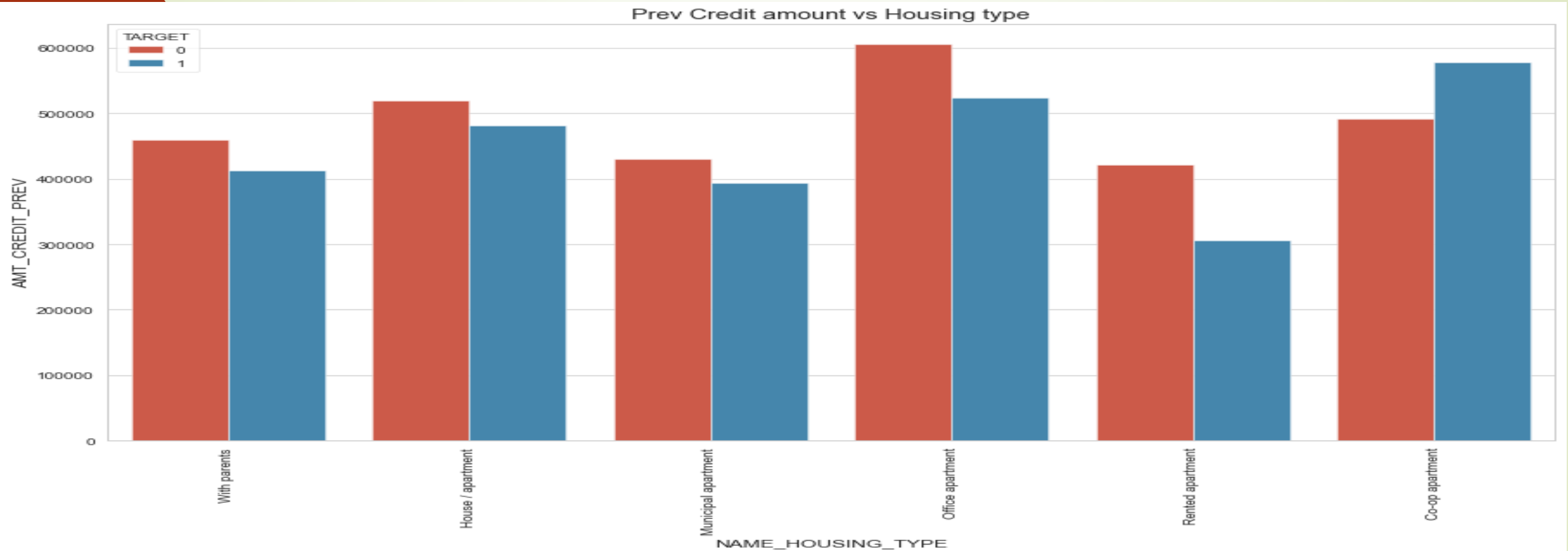
BIVARIATE ANALYSIS AFTER MERGING PREVIOUS DATA



OBSERVATION:

1. The credit amount of Loan purposes like 'Buying a home', 'Buying a land', 'Buying a new car' and 'Building a house' is higher.
2. Income type of state servants have a significant amount of credit applied
3. Money for third person or a Hobby is having less credits applied for.

BIVARIATE ANALYSIS FOR CONTINUOUS & CATEGORICAL VARIABLES AFTER MERGING PREVIOUS DATA



OBSERVATION:

1. Housing type, office apartments is having higher credit of target 0, and co-op apartments is having higher credit of target 1.
2. So, we can conclude that banks should avoid giving loans to the housing type of co-op apartments as they are having difficulties in payment. Bank can focus mostly on housing type with parents or House\apartment or municipal apartment for successful payments.

FINAL OBSERVATION AND CONCLUSION OF EDA CASE STUDY

- ❖ Target/focused variable for Application dataset – TARGET.
- ❖ Target/focused variable for Previous dataset - NAME_CONTRACT_STATUS.
- ❖ Banks should focus less on income type 'Working' and loan purposes 'Repair' as they are having the most number of unsuccessful payments
- ❖ Banks should focus more on contract types 'Student', 'pensioner', and 'Businessman' with housing 'type other than 'Co-op apartment' for successful payments.
- ❖ Banks should focus to get as much as clients from housing type 'With parents' as they are having the least number of unsuccessful payments.
- ❖ Banks must consider the following variables such as, 'NAME_CONTRACT_STATUS', 'NAME_CASH_LOAN_PURPOSE', 'NAME_INCOME_TYPE', 'AMT_CREDIT', 'NAME_HOUSING_TYPE', 'AMT_ANNUITY', 'AMT_INCOME_TOTAL', before approving loan application to minimize the risk of loss.



THANK YOU