Presentation Report on EDA Case Study

Vishal V Ravi

Problem Statement –

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

The purpose is to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision :

1) If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

2) 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.

to be continued..

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

Approved : The Company has approved loan Application

Cancelled : The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want.

Refused : The company had rejected the loan (because the client does not meet their requirements etc.).

Unused offer : Loan has been cancelled by the client but at different stages of the process.

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency to default.

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.

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

Datasets used in the EDA Case Study :

1. '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.**

2. '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.**

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

Steps involved in the EDA Case Study :

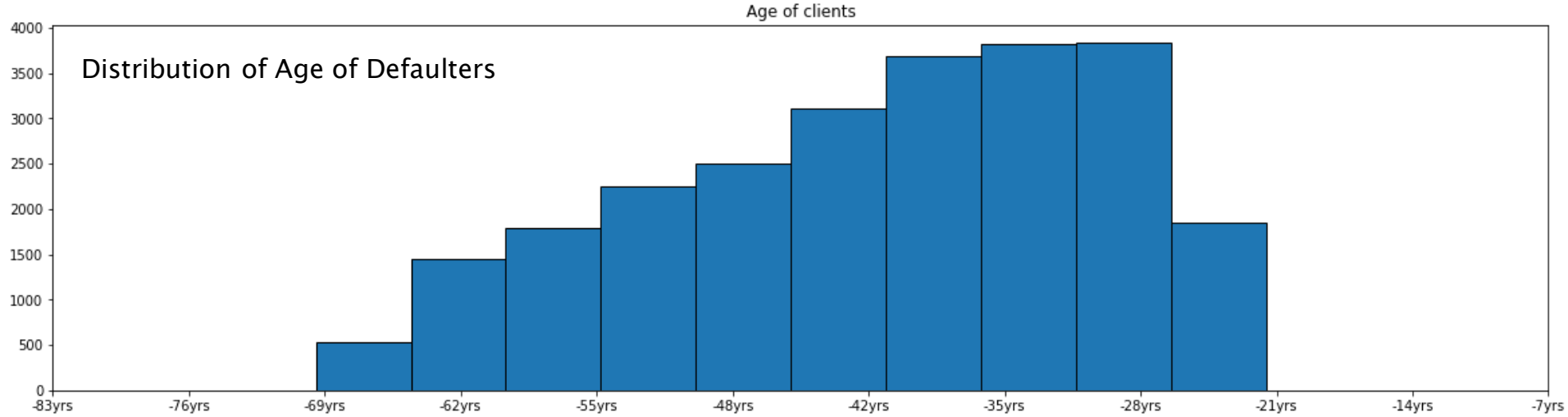
A) Data Cleaning :

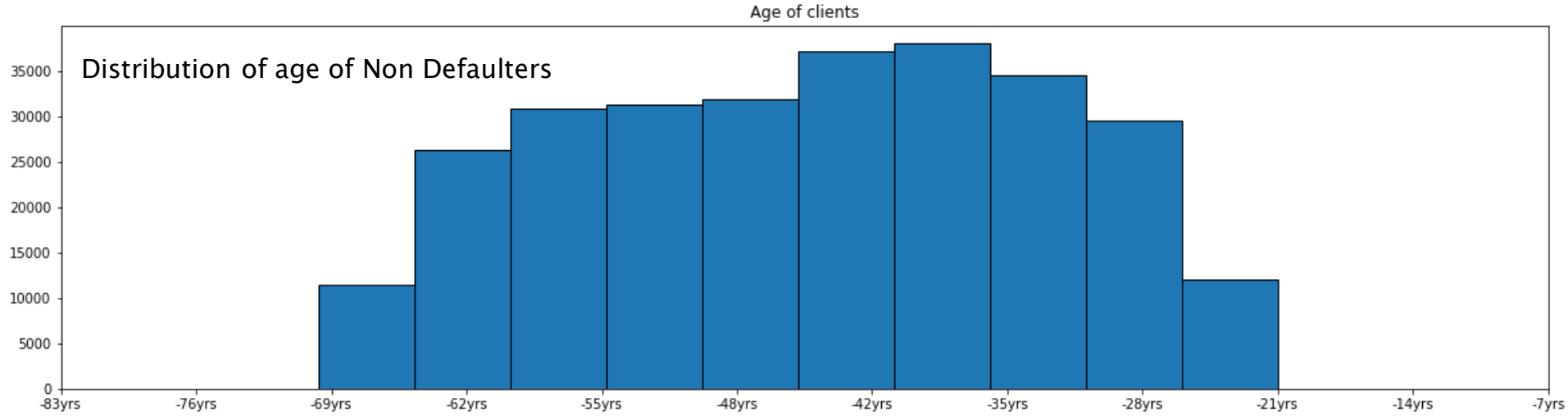
1) Importing the libraries – pandas, numpy, matplotplotlib and seaborn
2) Importing the datasets
3) Calculate null values percentage – remove columns with high null value percentage
4) Fix datatypes if required
5) Impute missing values suitably
6) Check for outliers in numerical variables and suggest how to handle them
7) Perform Binning of numerical variables

B) Data Analysis :

1) Perform univariate and bivariate analysis
2) Make plots accordingly by using data from the variables
3) Draw insights and make report
4) Make a conclusion on the EDA Case Study

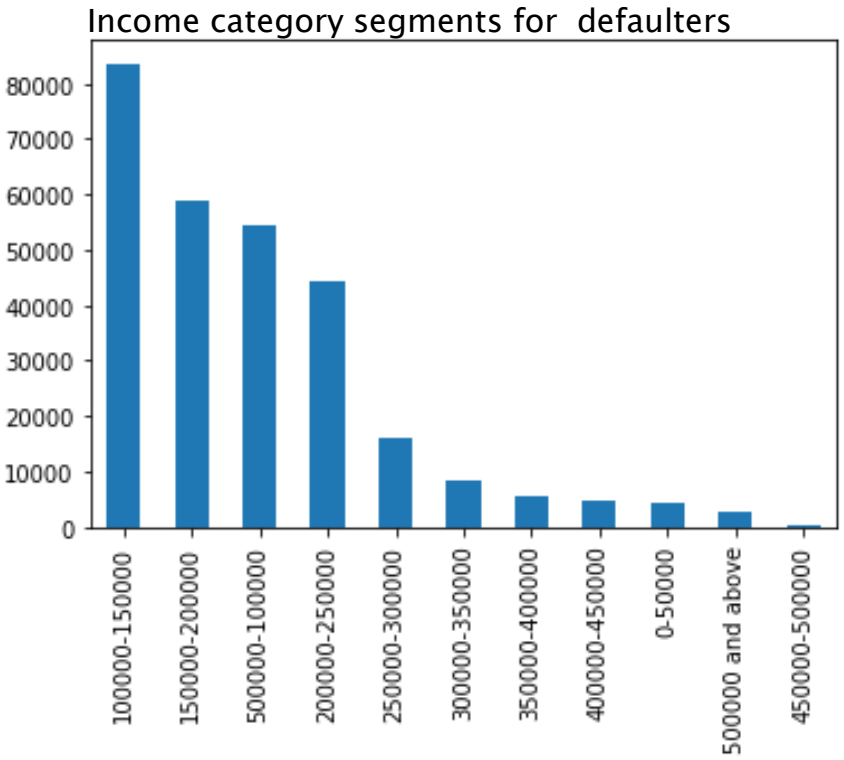
Univariate Analysis in Application Data

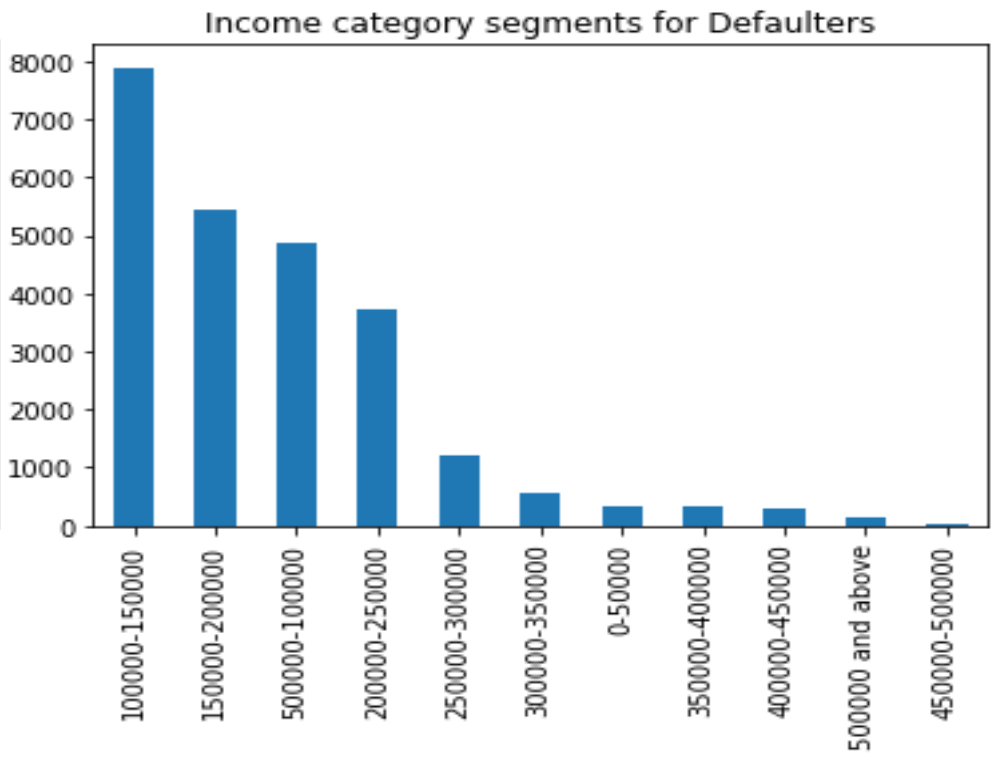




People of age group 26-40yrs are found to be defaulters.
This age group comprises of young people who have high chance to be defaulters.
Old age people are very less likely to be defaulters

Income range of the clients





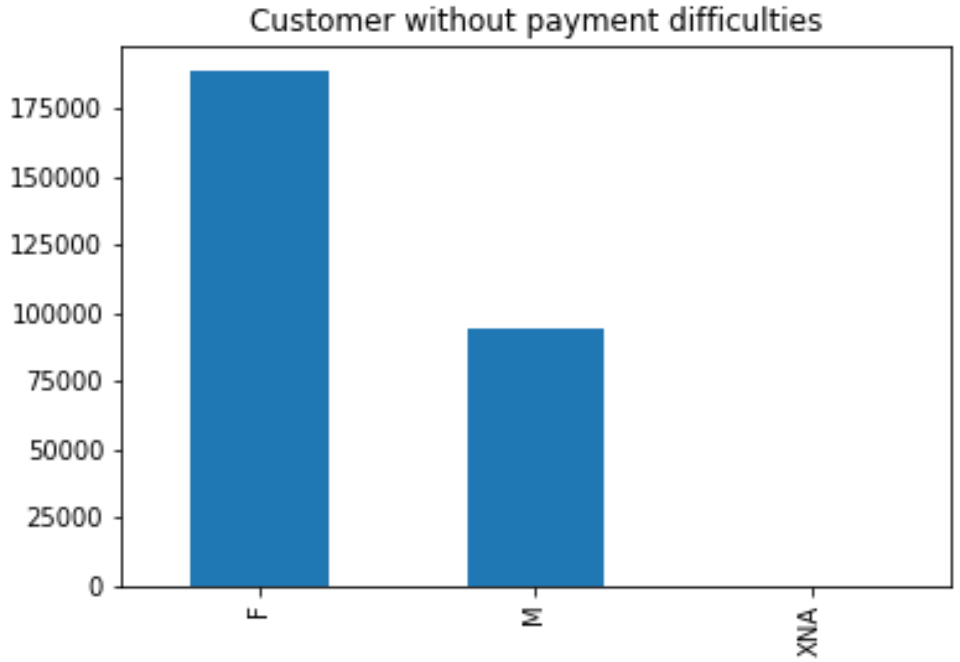
Here, we can see that there are more people earning in the low income range both for the cases of defaulters and non defaulters. Distribution of income is same for both the cases.

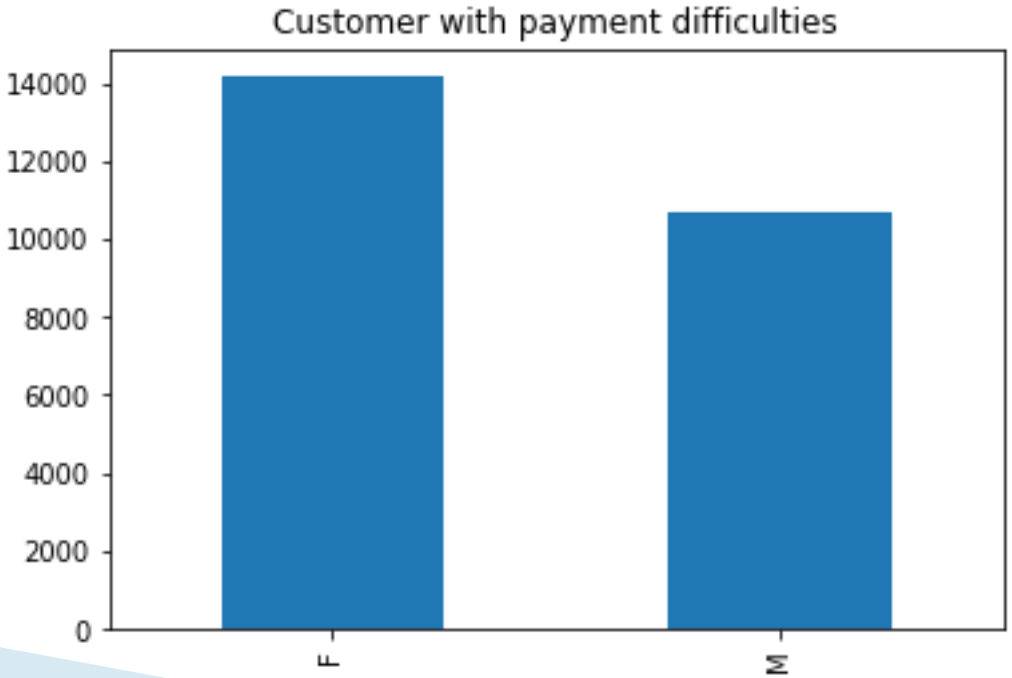
Gender distribution

It is observed that there are more females than males both in the cases of defaulters and non – defaulters.

But in the case of defaulters the difference between the number of females and males is less than in the case of non defaulters.

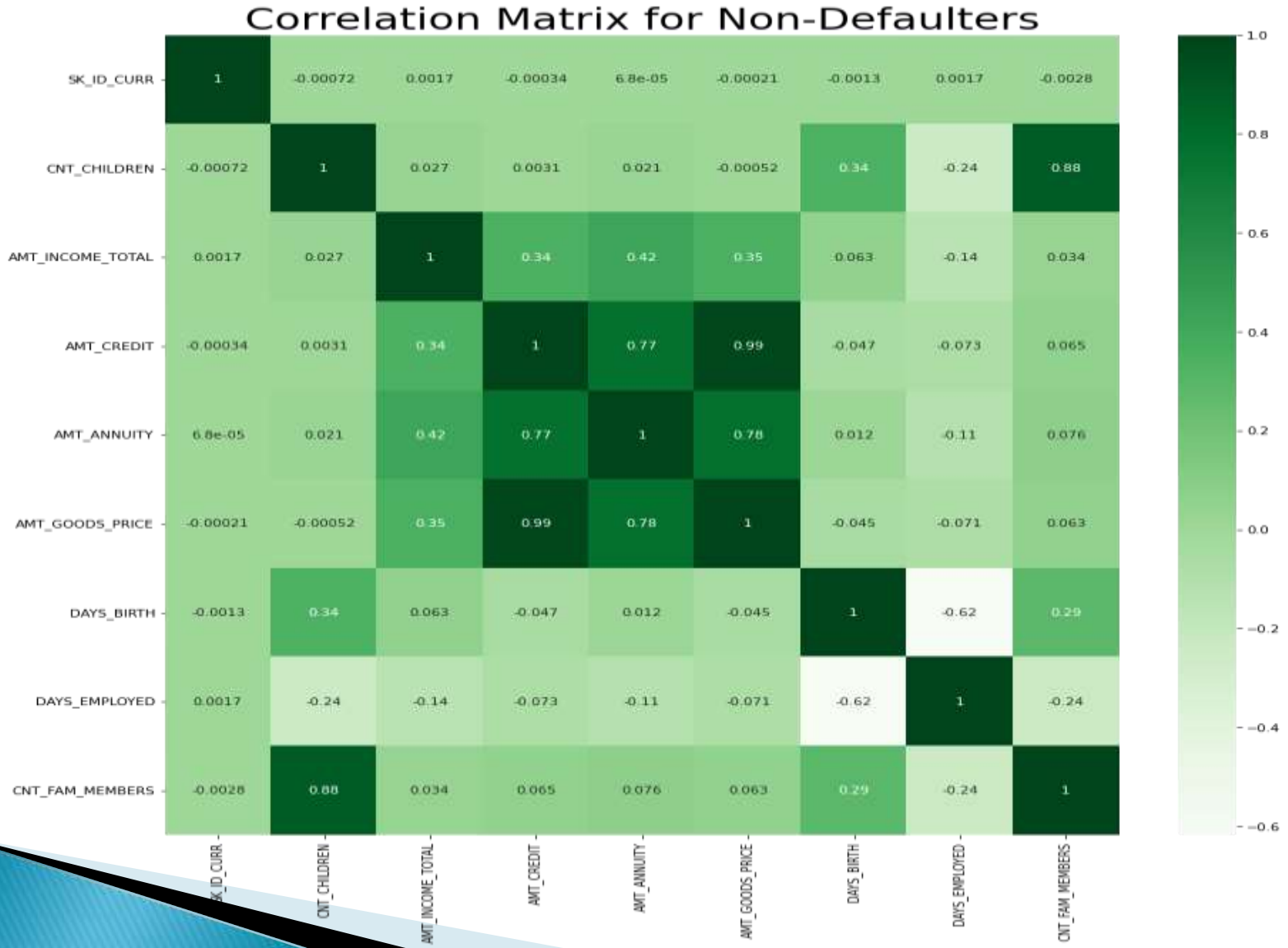
Thus, we can say that females are more likely to be deafulters than the males.



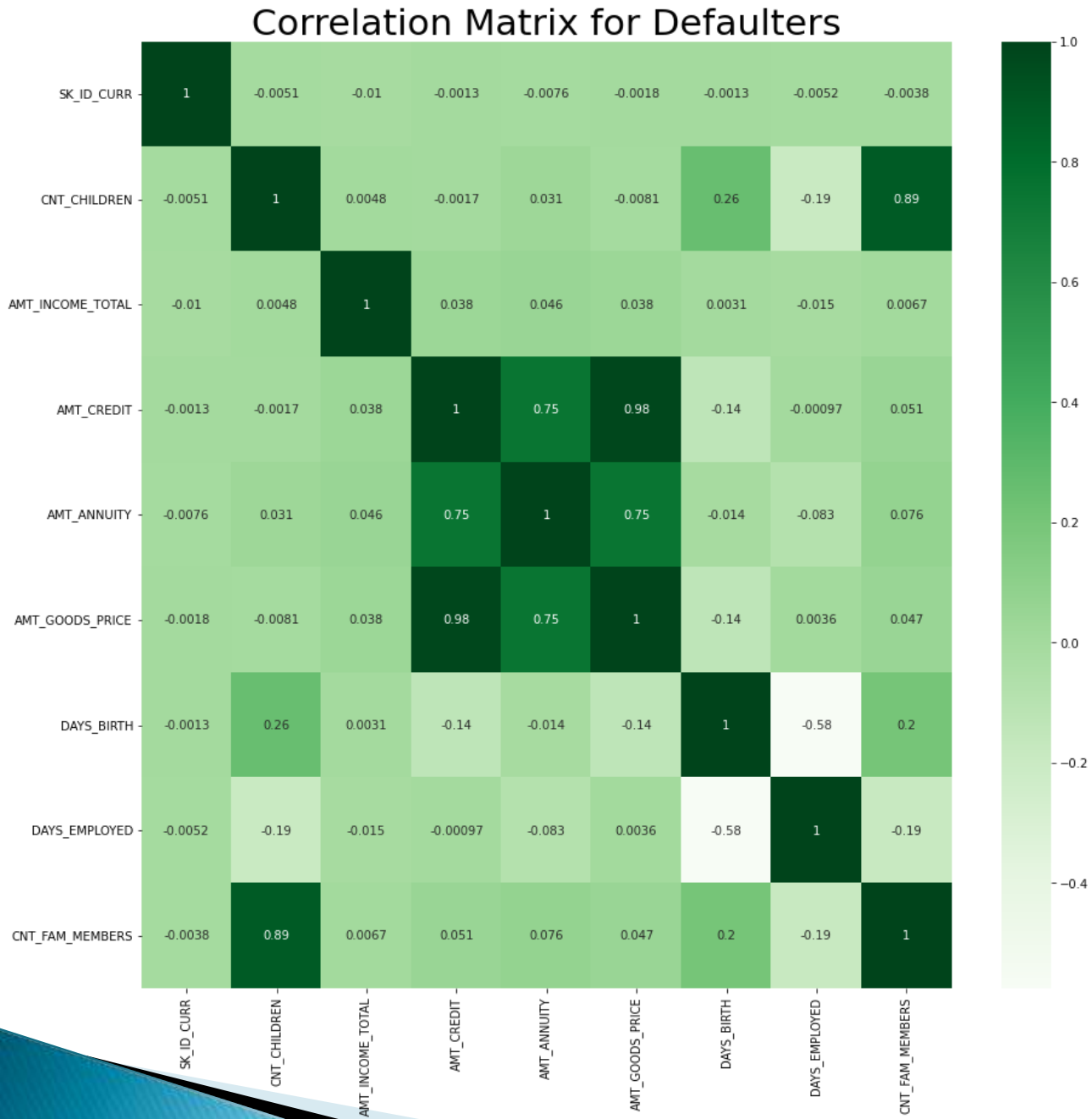


Bivariate analysis of Application data

Correlation matrix for non defaulters and defaulters dataframes



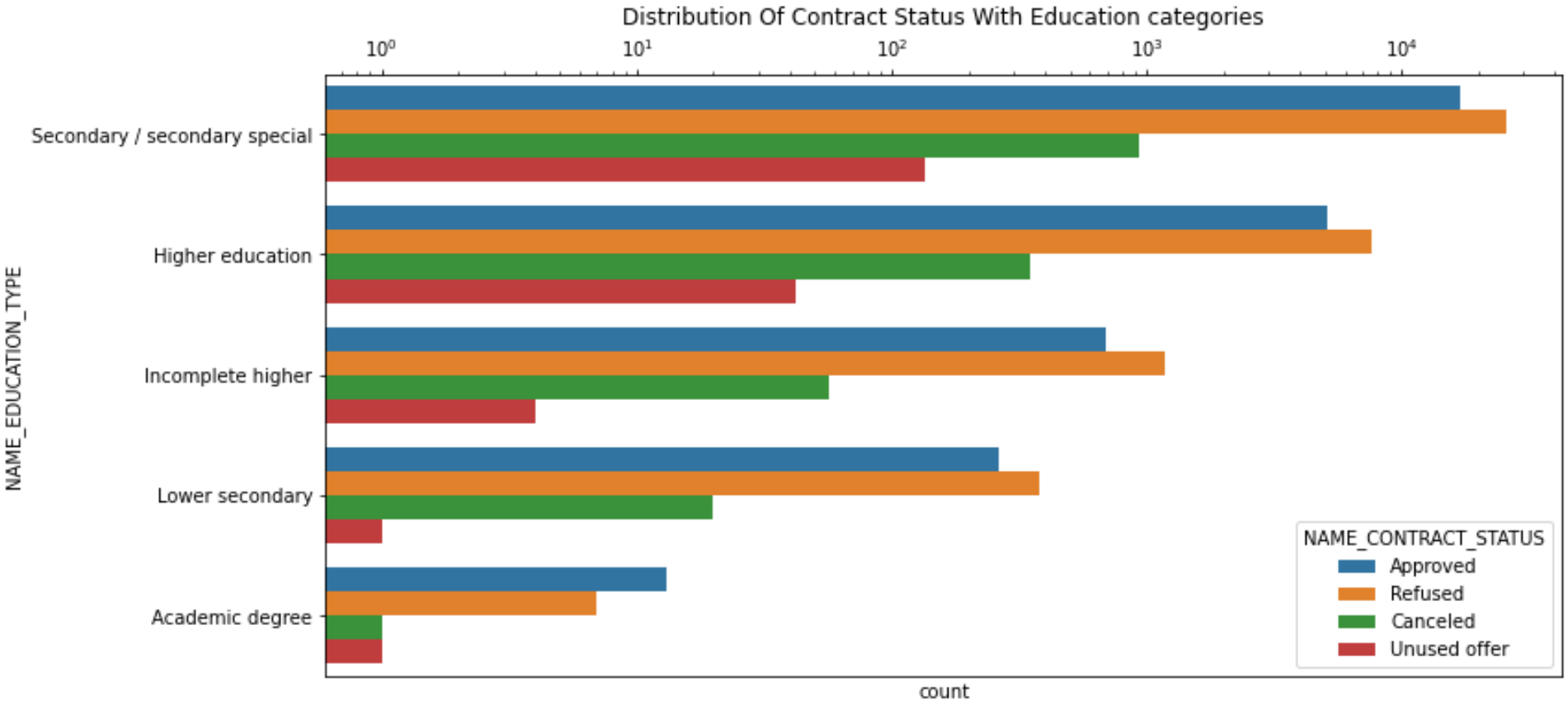
to be continued…



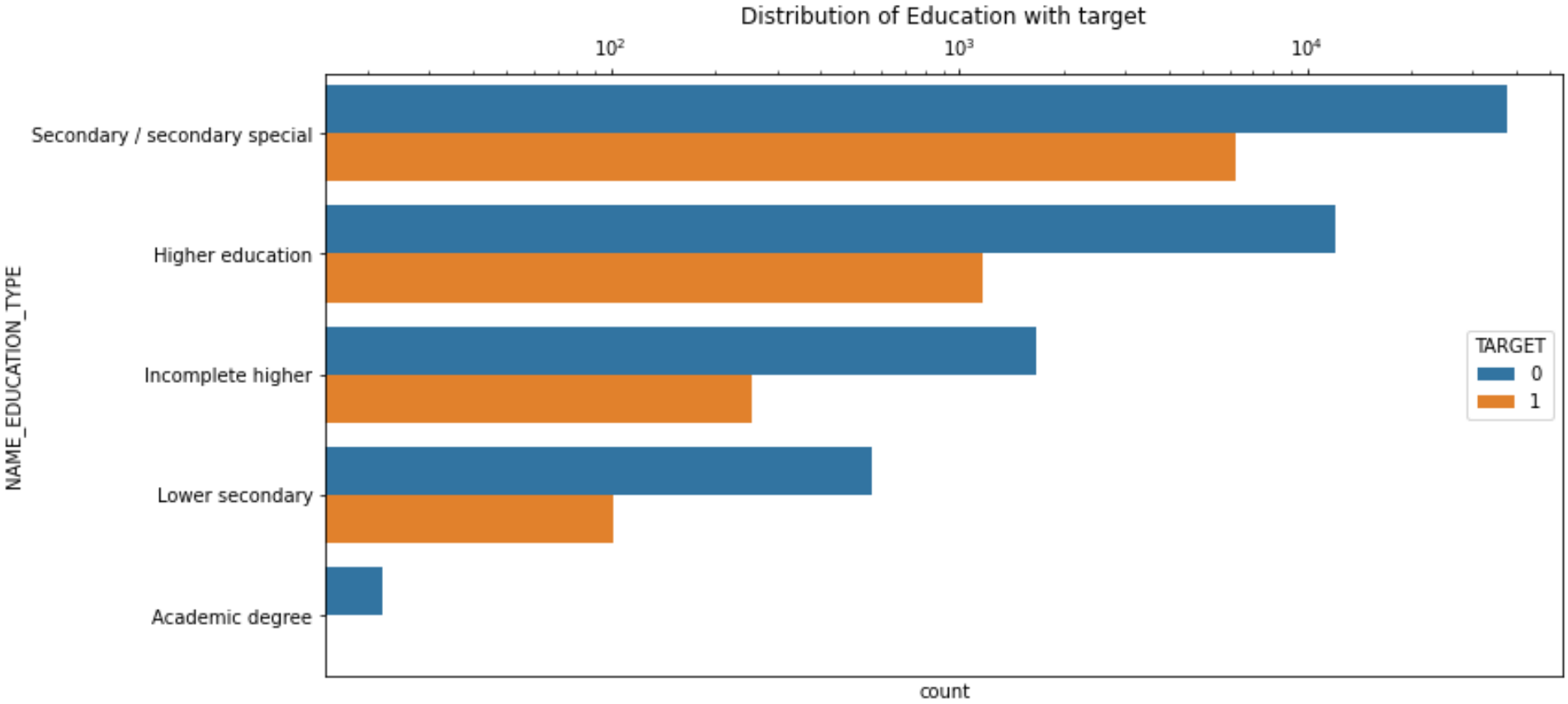
The heatmaps shows correlation values between the mentioned variables.

We can see that there is a high correlation between AMT_CREDIT and AMT_GOODS_PRICE and such high values can also observed between other pairs of variables as well. Moderate correlation values can be found Between CNT_CHILDREN and CNT_FAM_MEMBERS.

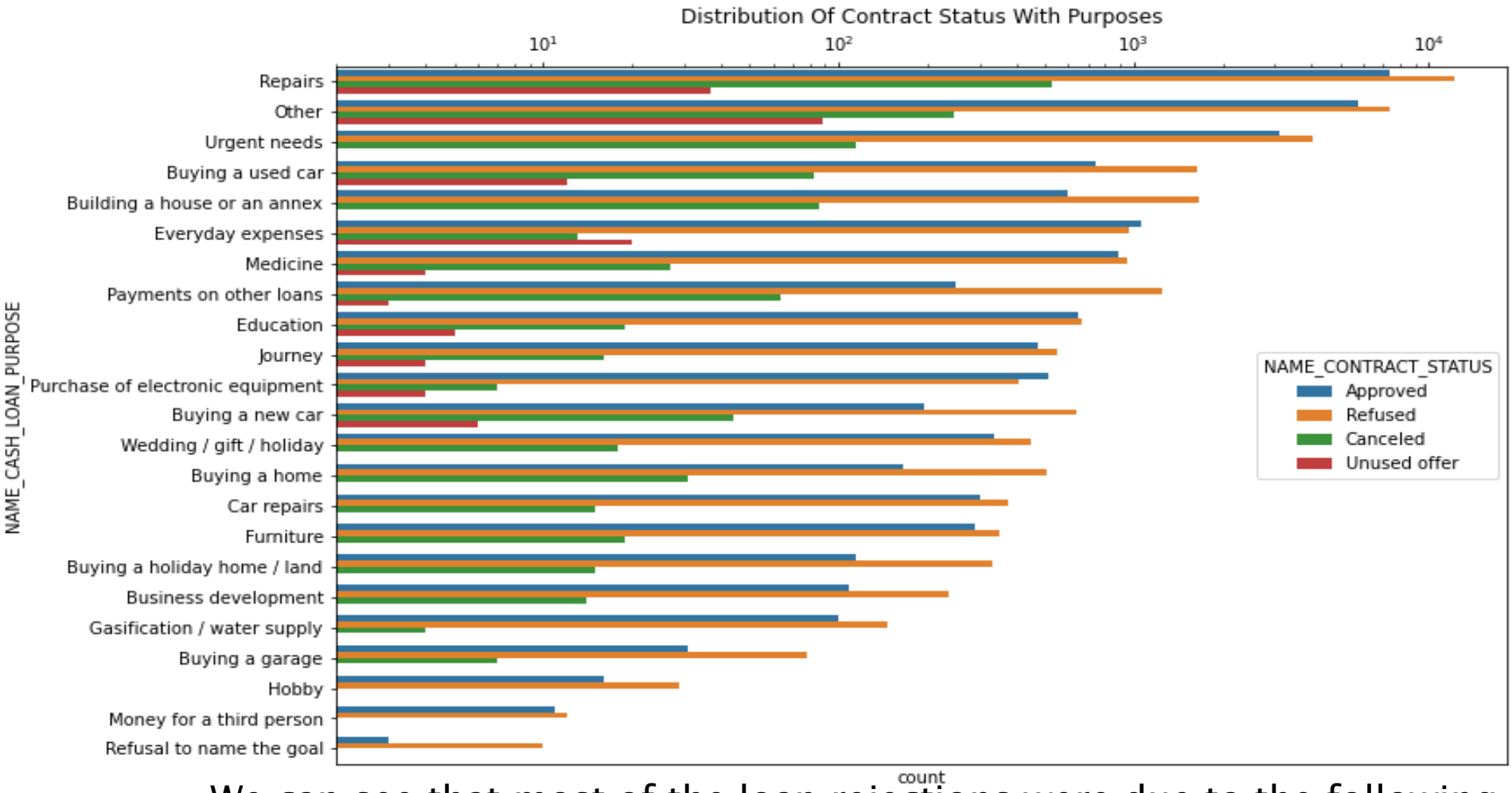
Analysis of Merged dataset



Here it is observed that people with secondary and higher education have their loan applications approved and rejected the most than those with other education fields.

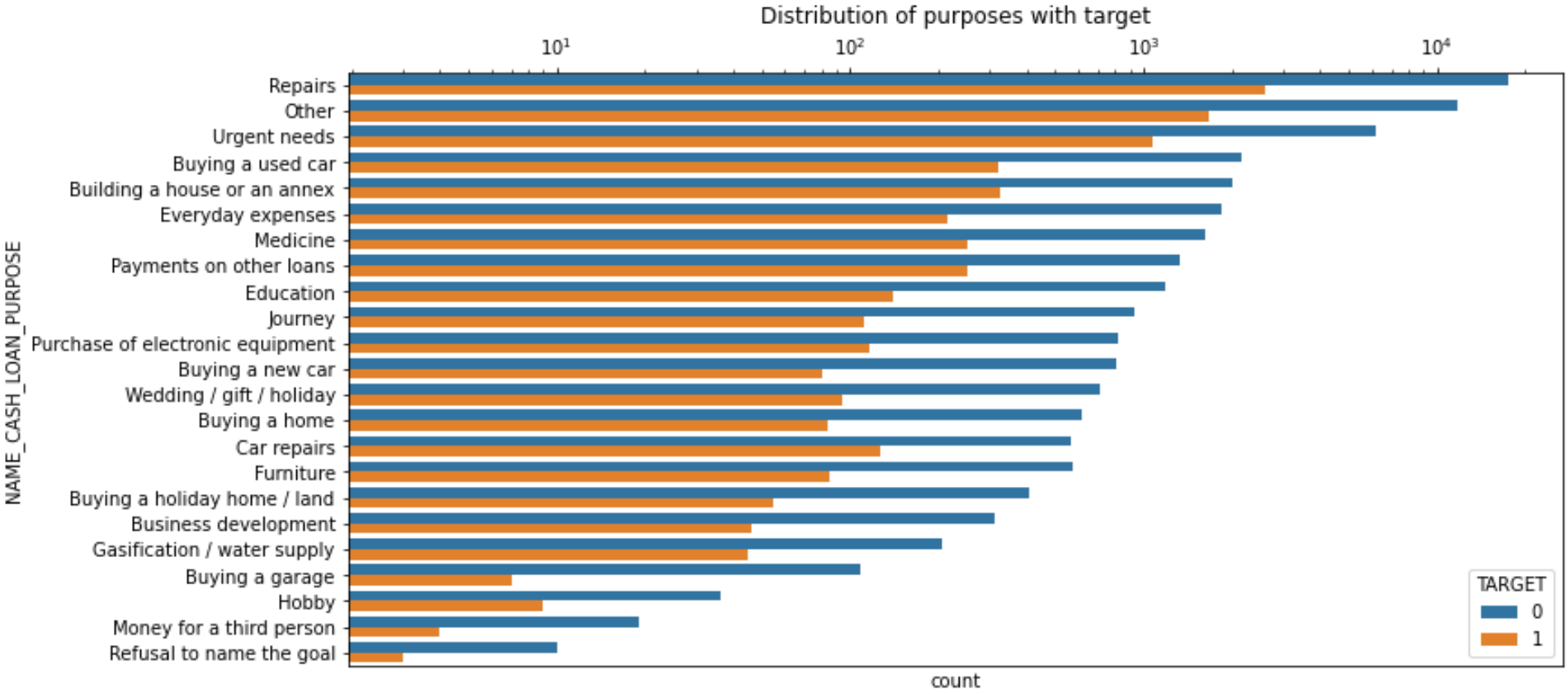


Clients with secondary education are more likely to not repay the loan in time and also more likely to repay the loan in time.

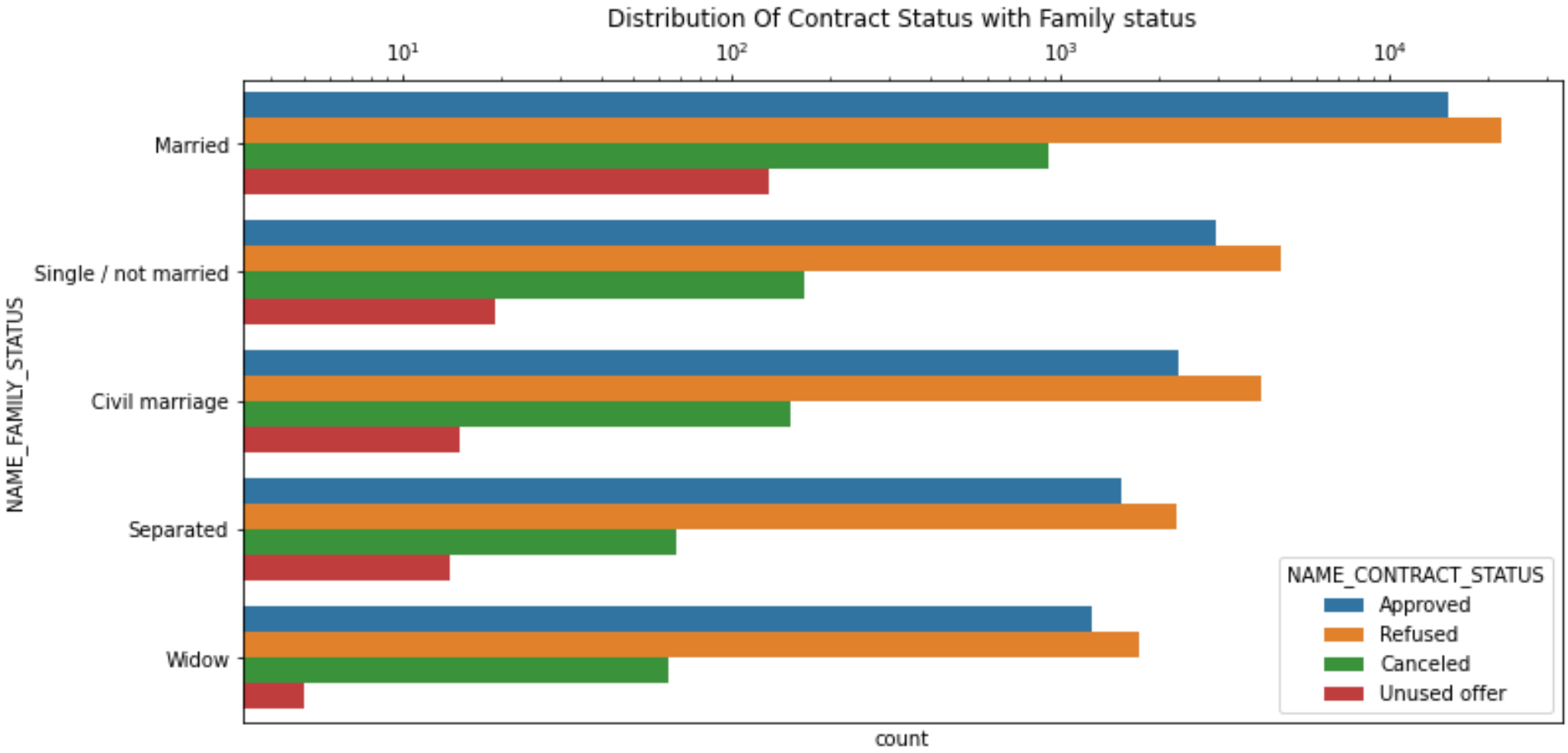


We can see that most of the loan rejections were due to the following causes – Repairs, Urgent needs and Other purposes.

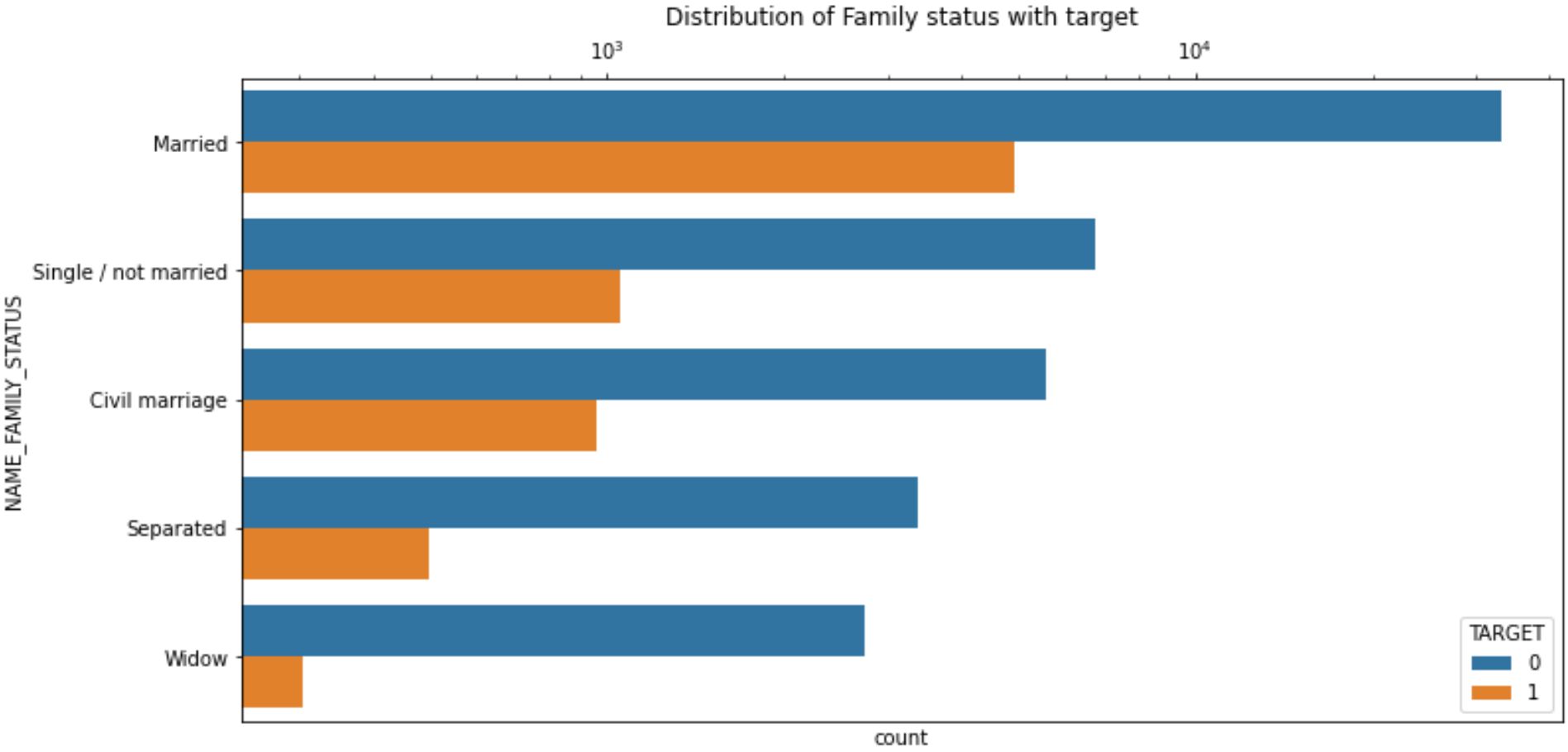
There are some some causes where approvals and rejections look similar like Medicines, Education, Everyday expenses, Journey etc



We can see that the clients with loan purposes – Repairs, Urgent needs and others are more likely to be defaulters.



Married clients have seen their applications approved and rejected the most.



Married clients are more likely to be defaulters whereas widows are more likely to be non defaulters.

Conclusions :

1) Clients who are less likely to be a defaulter :
   a) Older people
   b) People with academic degrees
   c) People who earn in the high income range
   d) Widows who have unused previous loan status
   e) Those who have refused to mention their purpose for loan application

2) Clients who are more likely to be a defaulter :
   a) Those who have their previous application refused or cancelled
   b) People with secondary education
   c) Married clients
   d) People with loan purpose – payments on other loans

3) Imbalance percentage – 11.39%