

EDA CASE STUDY SOLUTION

By Deepak A

INTRODUCTION–

This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

BUSINESS UNDERSTANDING – 1

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 as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse 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:
- If the applicant is likely to repay the loan, then not approving the loan results 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 UNDERSTANDING –2

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios: • The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample, • All other cases: All other cases when the payment is paid on time. When a client applies for a loan, there are four types of decisions that could be taken by the client/company): 1. Approved: The Company has approved loan Application 2. 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. 3. Refused: The company had rejected the loan (because the client does not meet their requirements etc.). 4. Unused offer: Loan has been cancelled by the client but on different stages of the process. In this case study, we will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

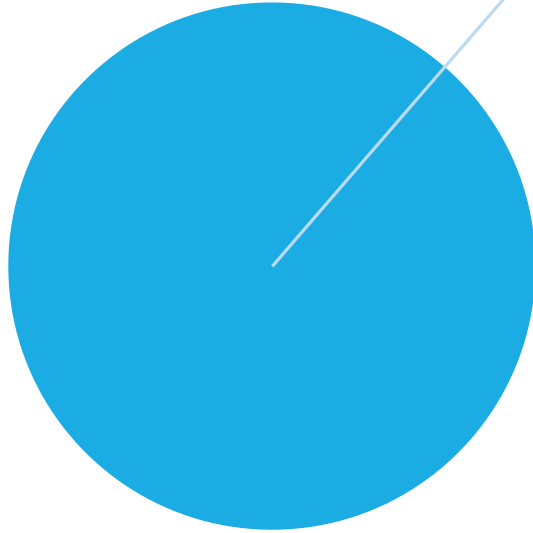
BUSINESS OBJECTIVES

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 applicant's 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 utilize this knowledge for its portfolio and risk assessment. • To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough)

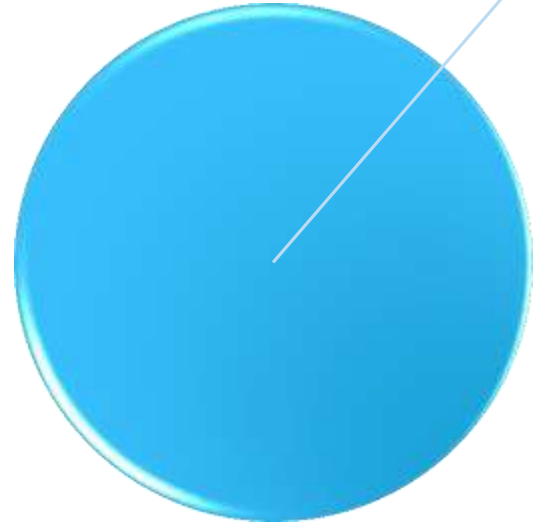
DATA UNDERSTANDING

This dataset has 3 files as explained below:

- 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 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.

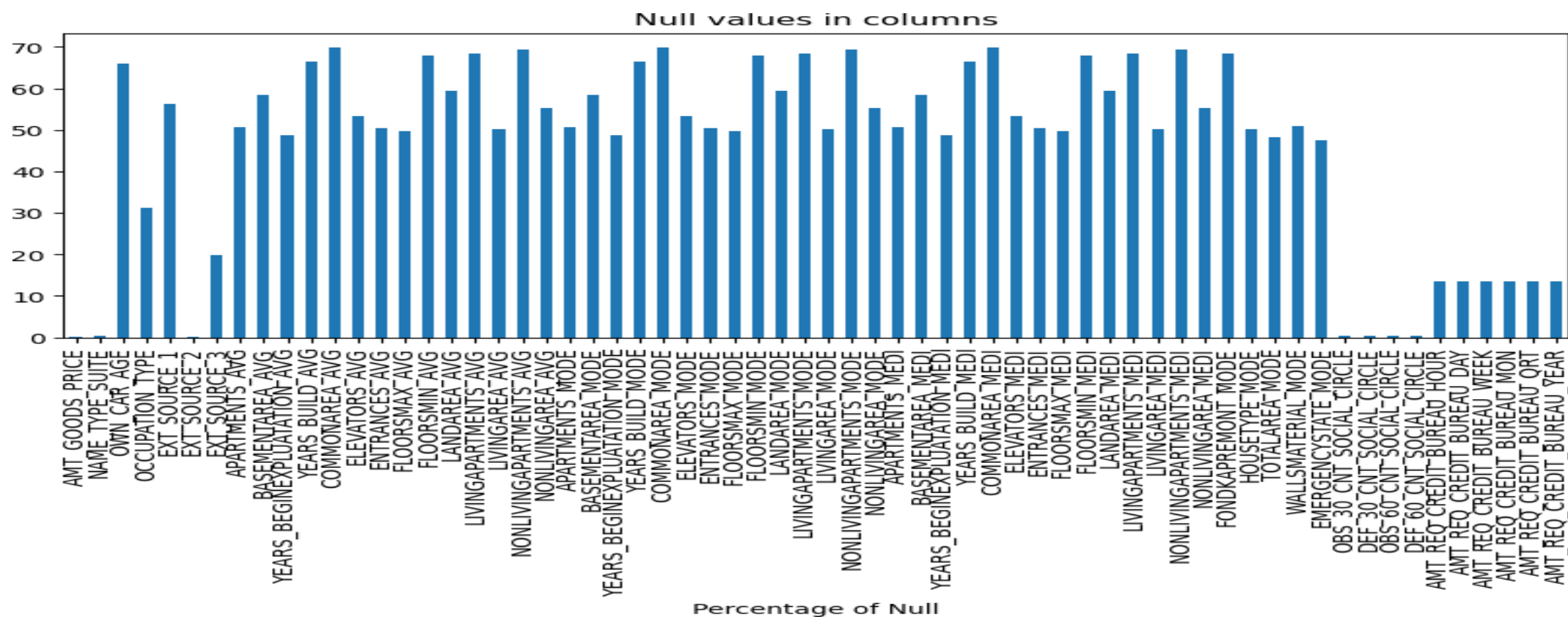


Analysis of
information
of the client
at the time of
application

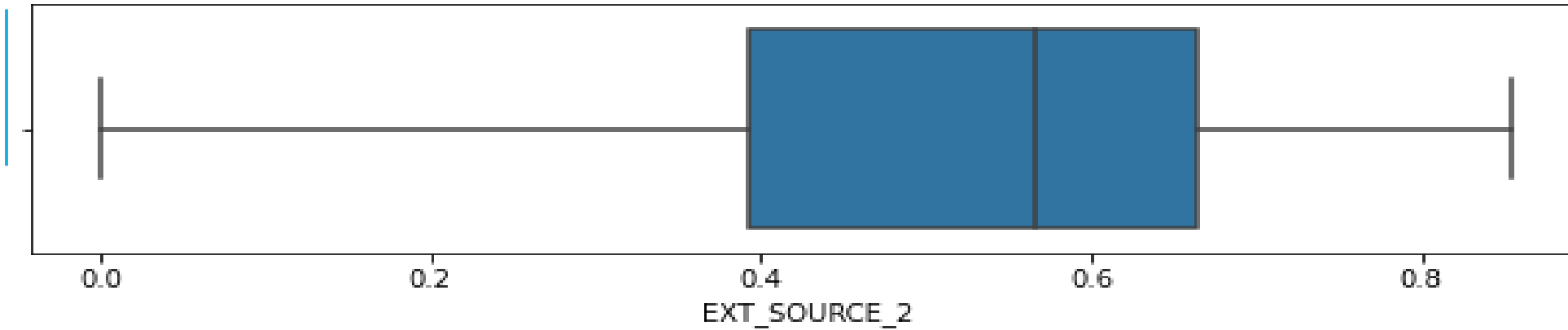


Outlier
analysis

Visualizing Null values of columns in graph

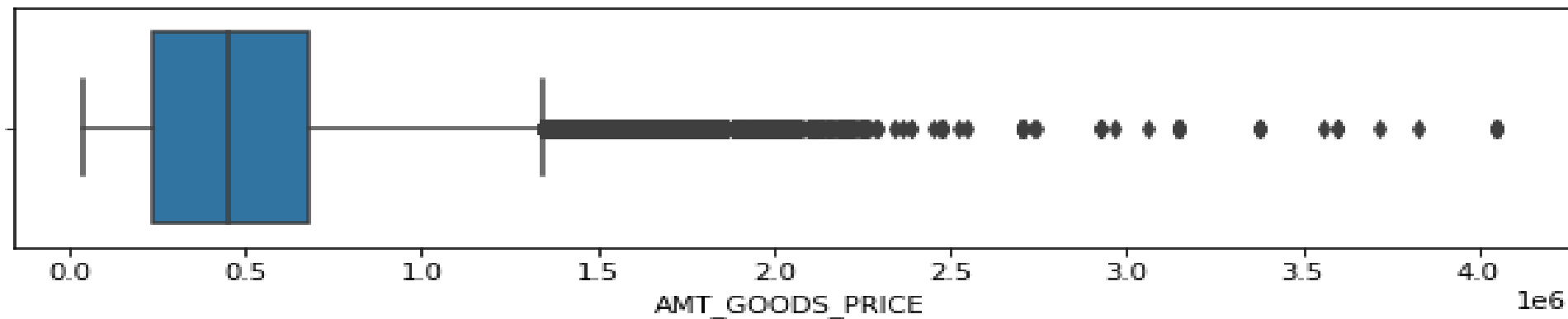


Continuous variable



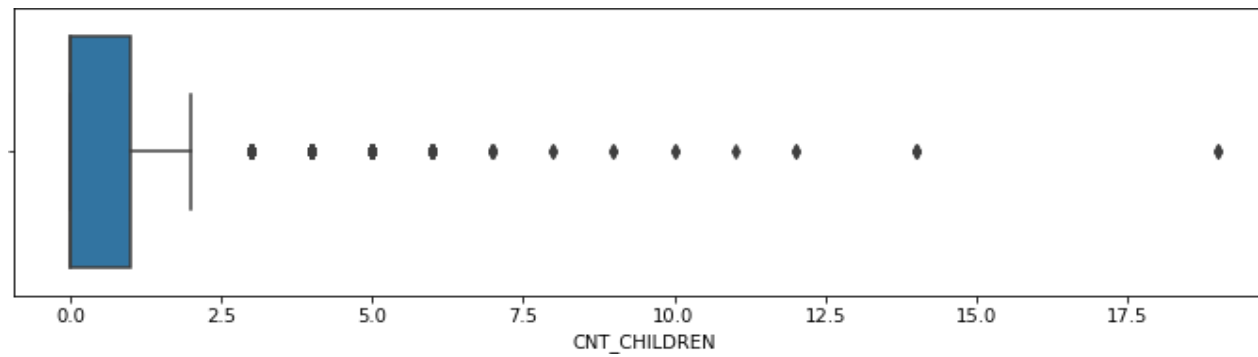
Observation from Boxplots:

For 'EXT_SOURCE_2' no outliers present. So data is rightly present. For 'AMT_GOODS_PRICE' outlier present in the data. so need to impute with median value: 4

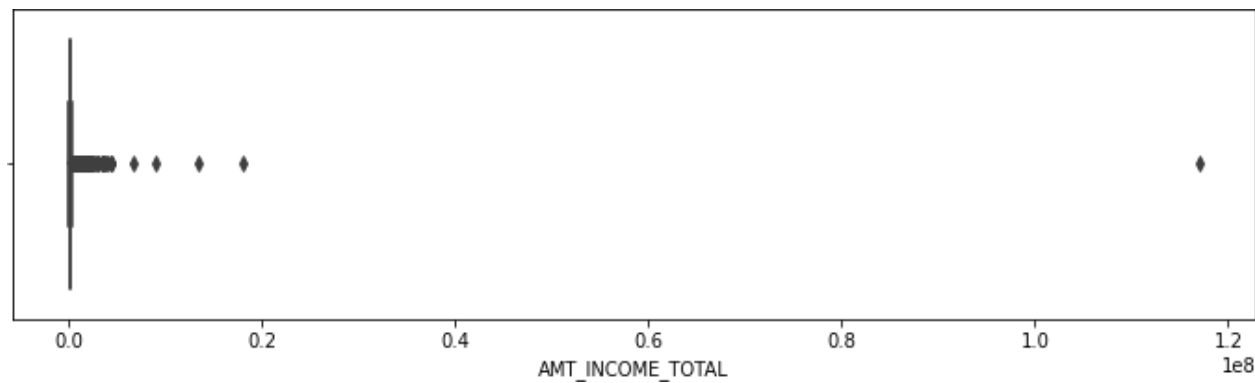


From the above we can see that first two (EXT_SOURCE_2, AMT_GOODS_PRICE) are continuous variables and remaining are categorical variables

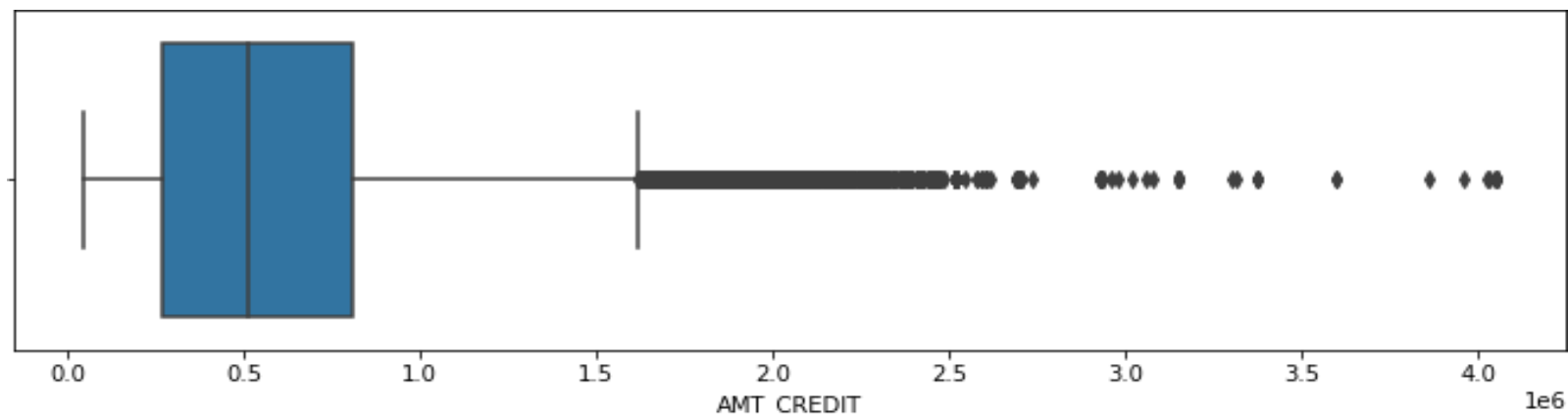
check box plot for 'CNT_CHILDREN',
'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION'



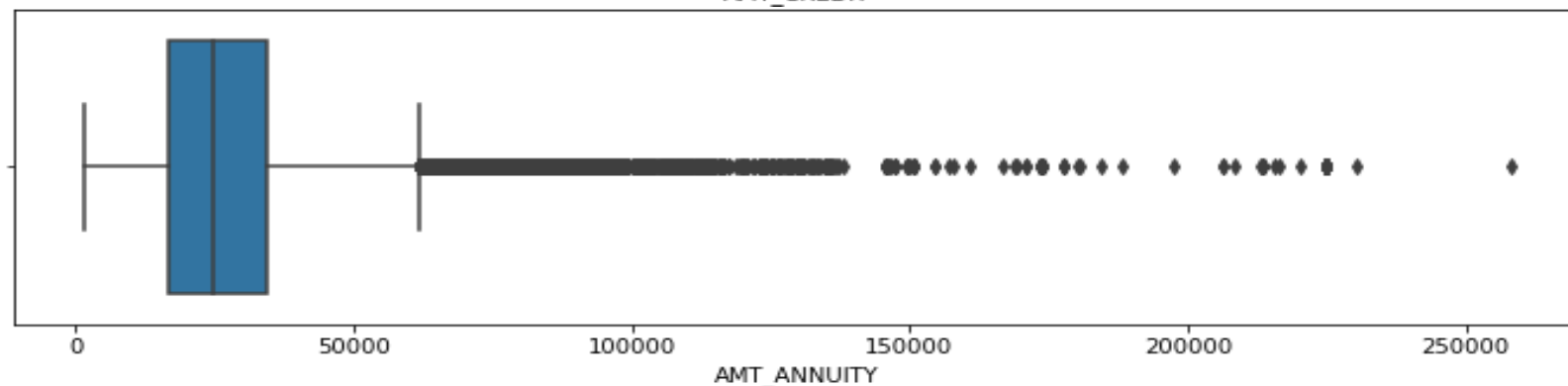
1st quartile is missing for
CNT_CHILDREN which means most of
the data are present in the 1st quartile.



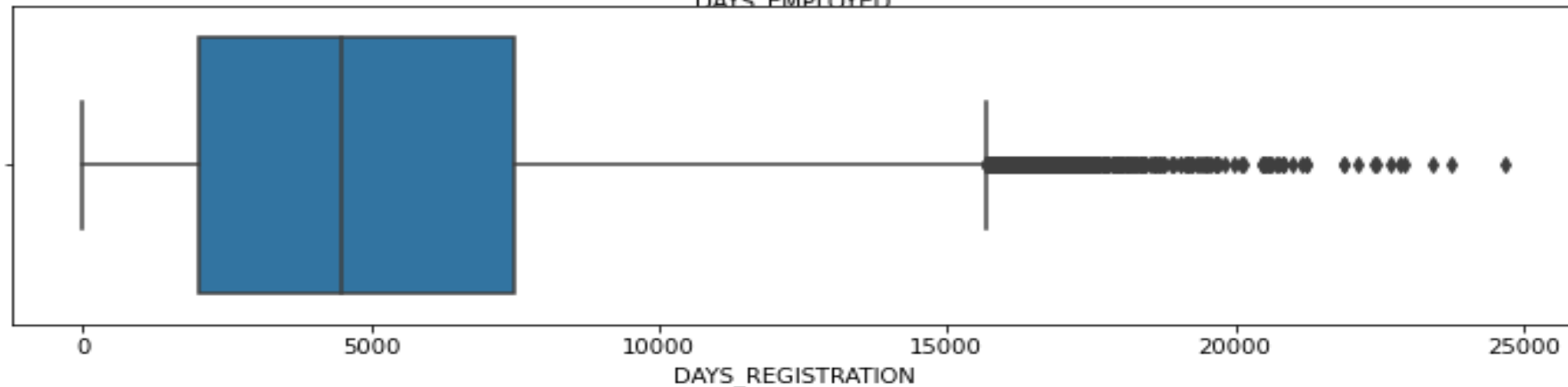
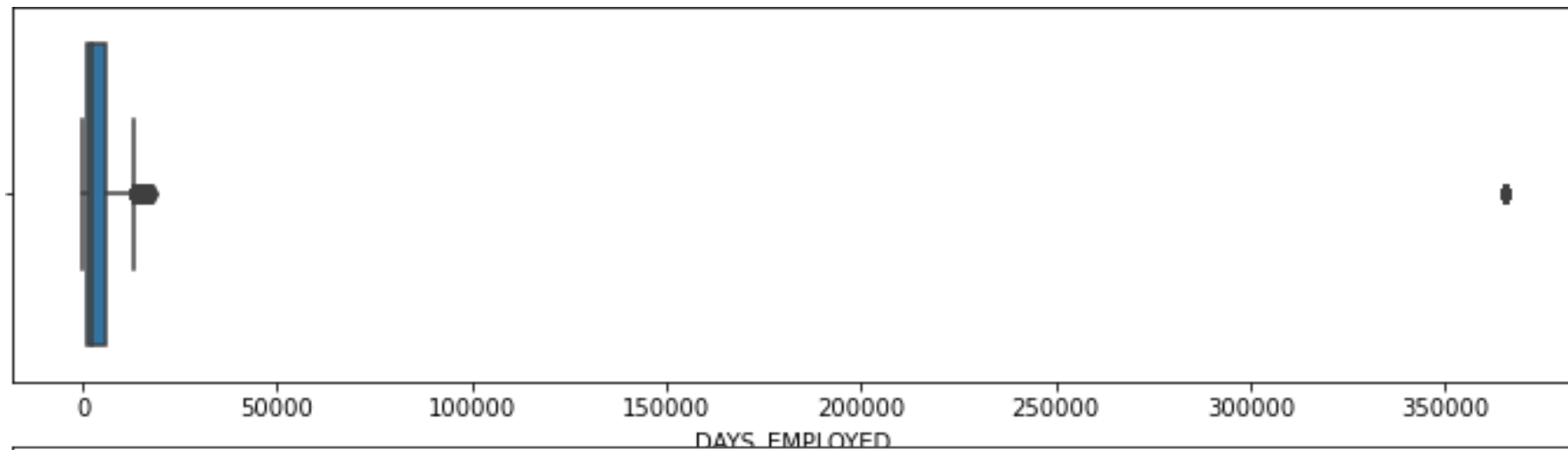
In AMT_INCOME_TOTAL only single high value data point is present as outlier



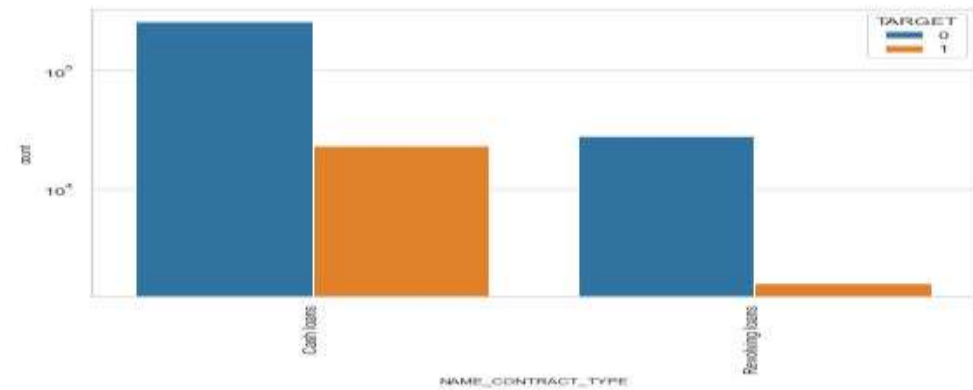
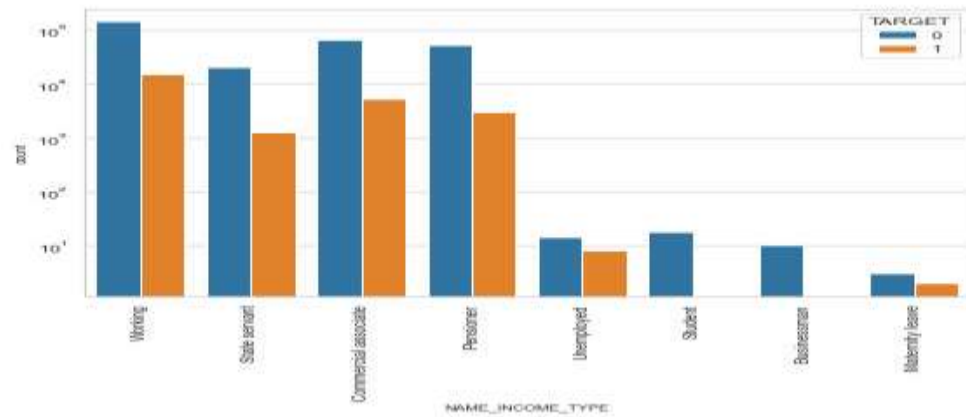
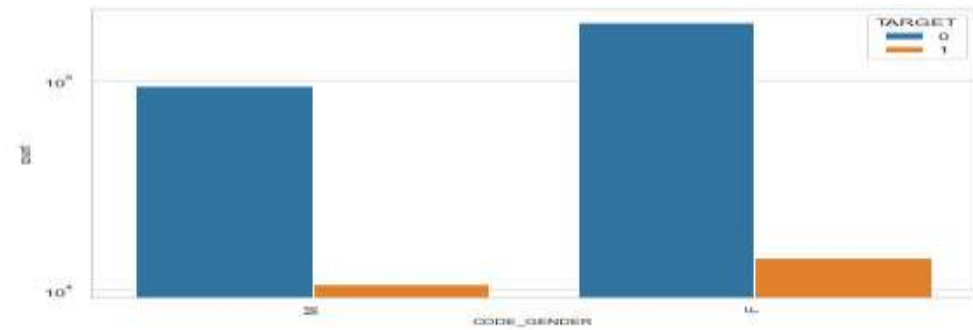
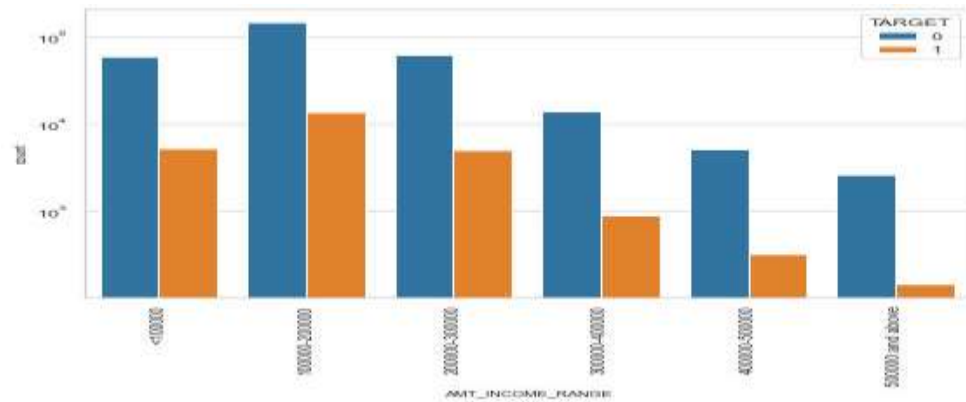
AMT_CREDIT has
little bit more outliers



1st quartiles and 3rd quartile for
AMT_ANNUITY is moved towards first quartile.

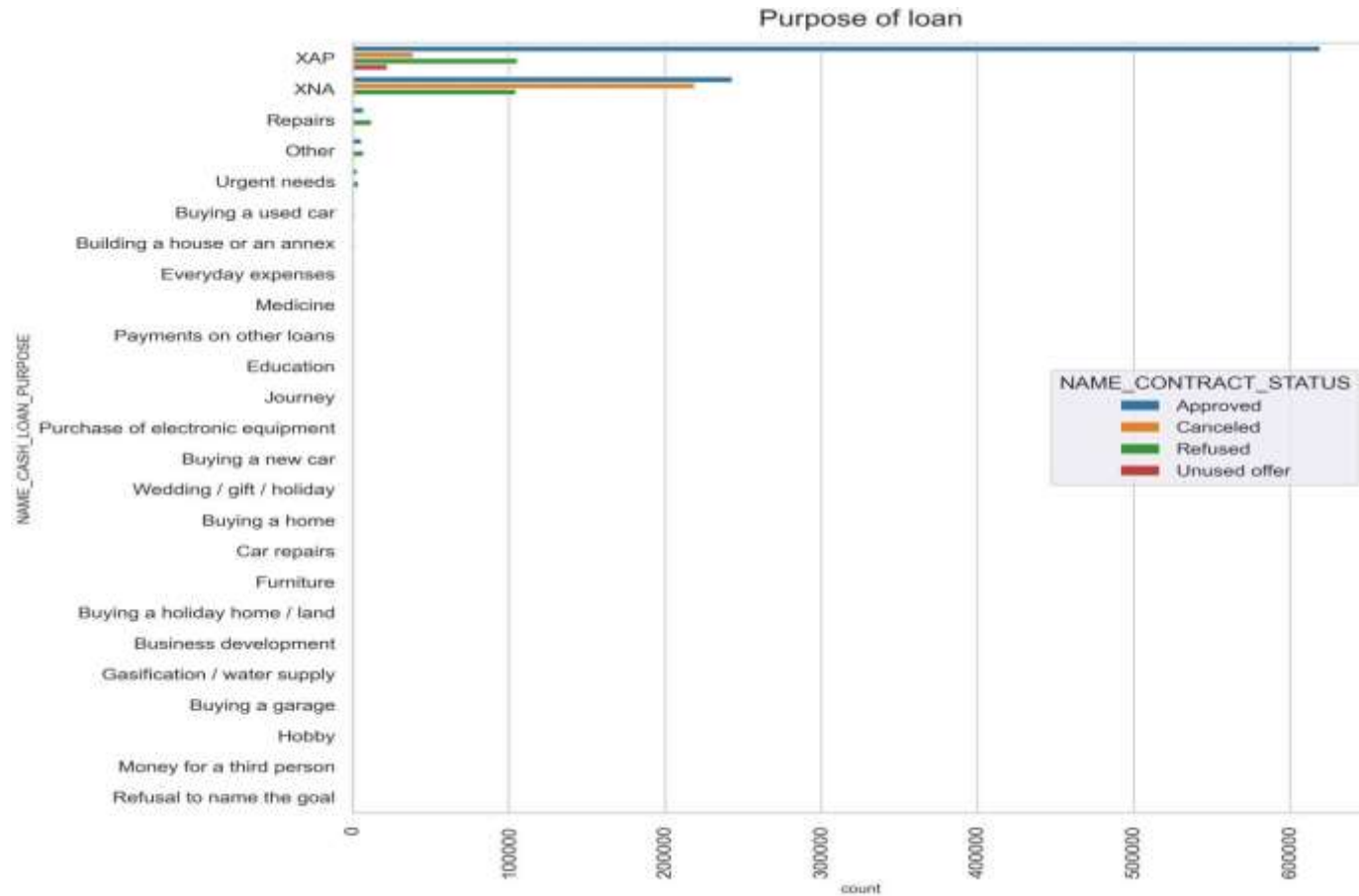


Univariate Analysis



Performing univariate analysis

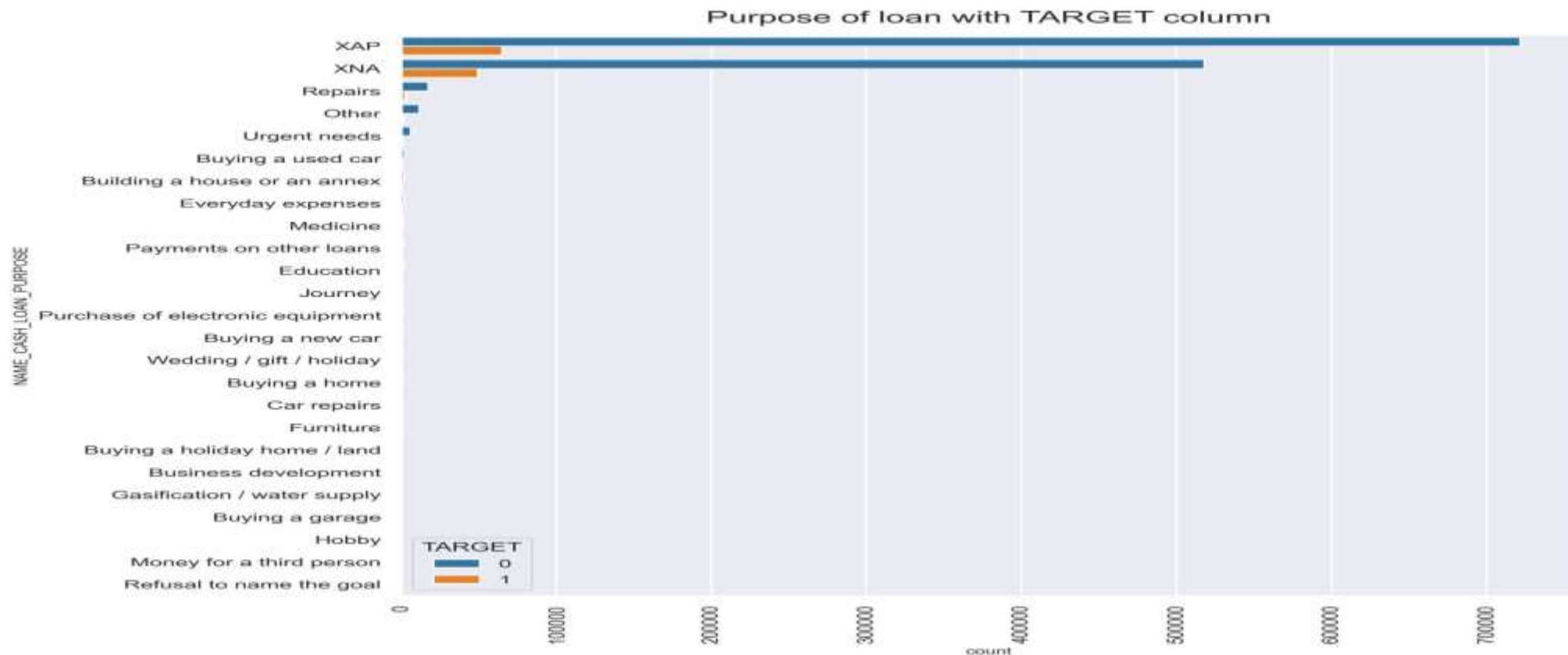
-PURPOSE OF LOAN



Observation

Most of loan rejection was from 'repairs'

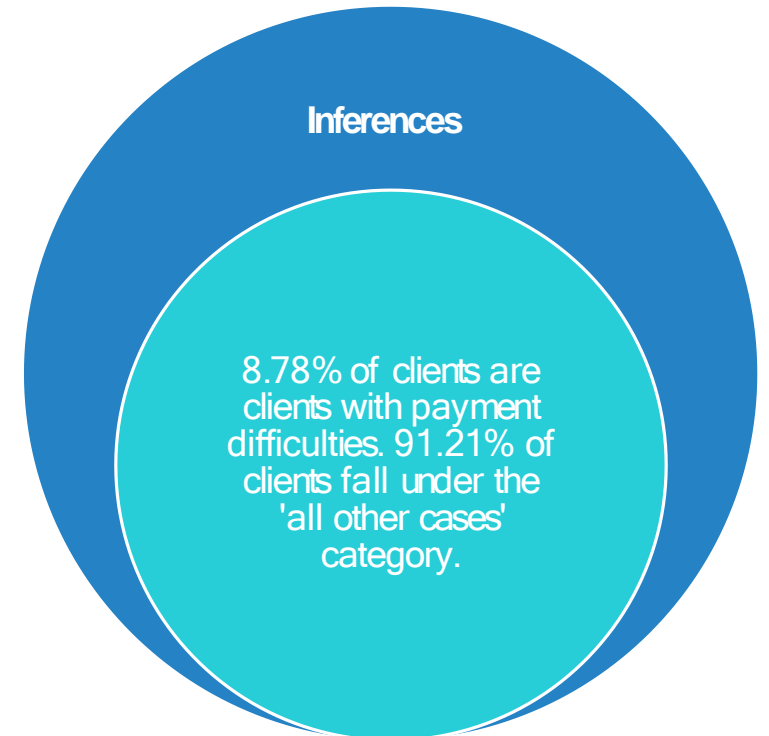
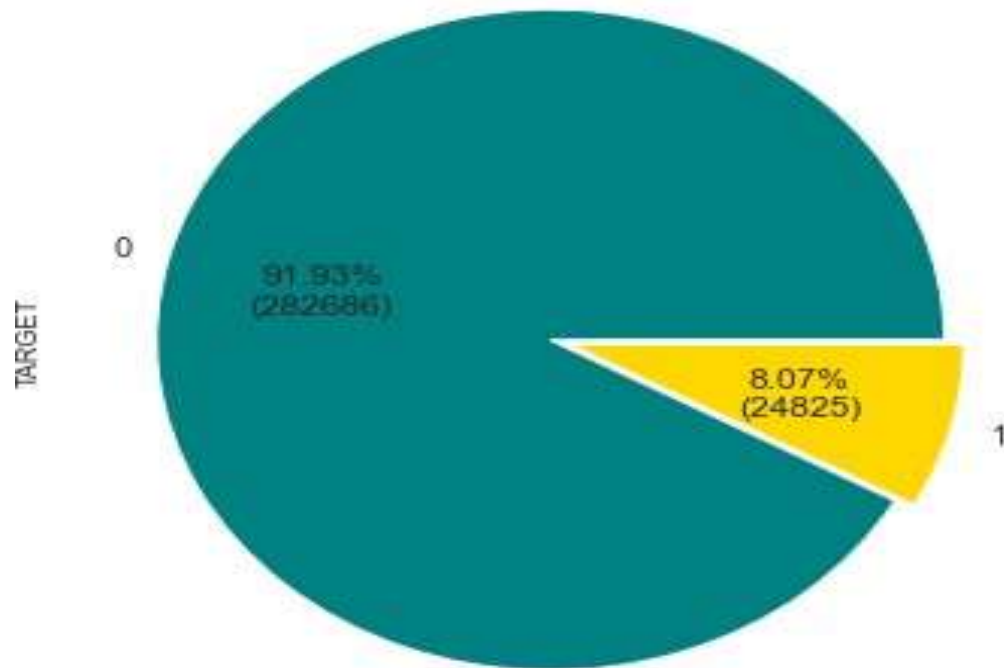
Purpose of loan with TARGET column



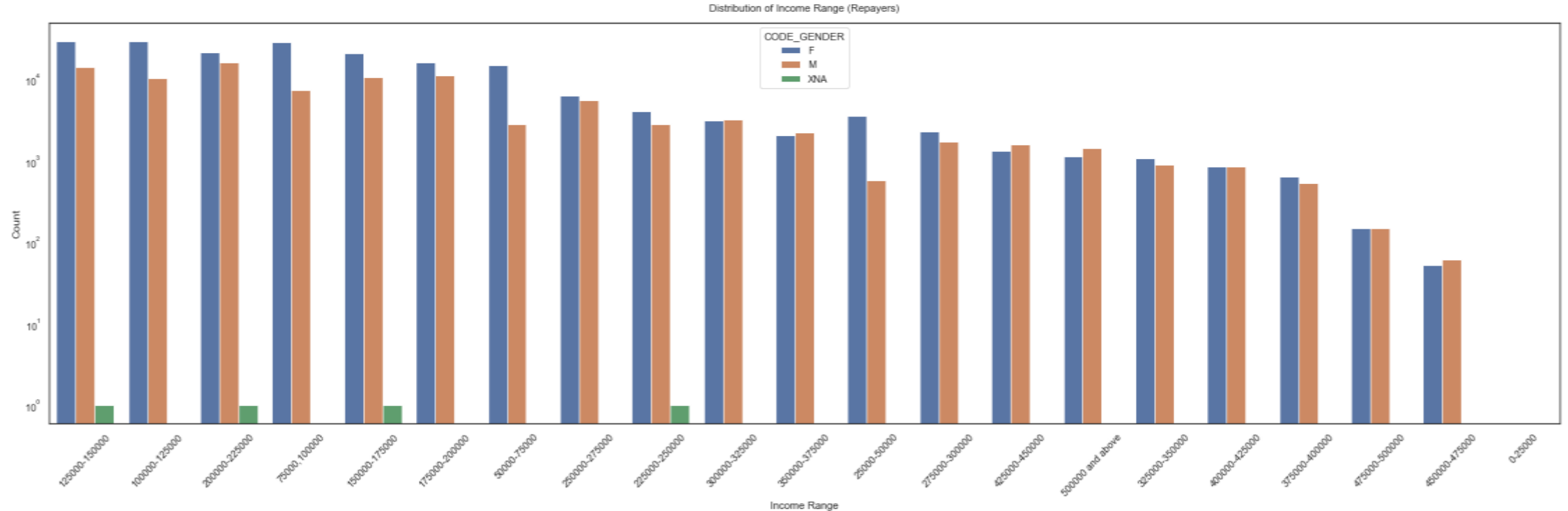
Calculating Imbalance percentage for target 0 and target 1.

Visualizing the above result in a pie plot

Imbalance between target0 and target1



Plotting graphs for Target0 (Customers with no payment difficulties)



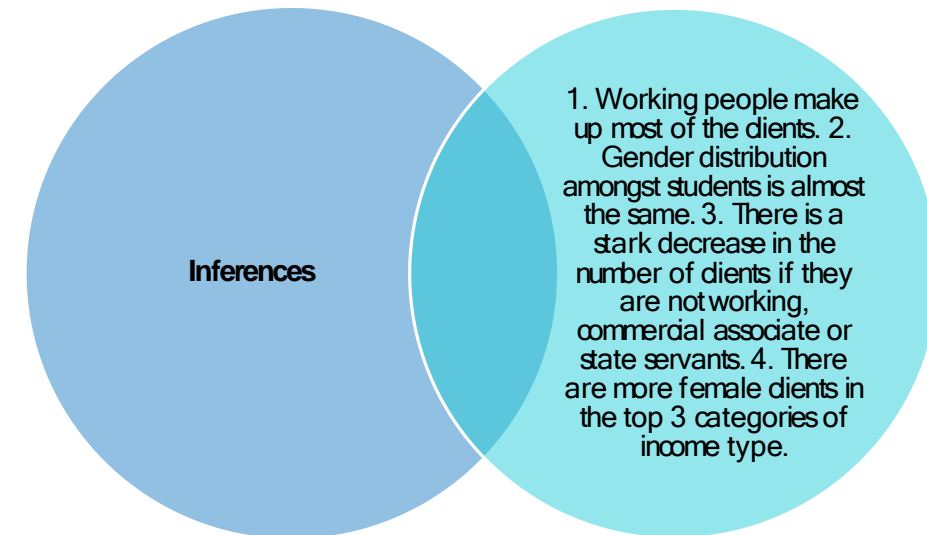
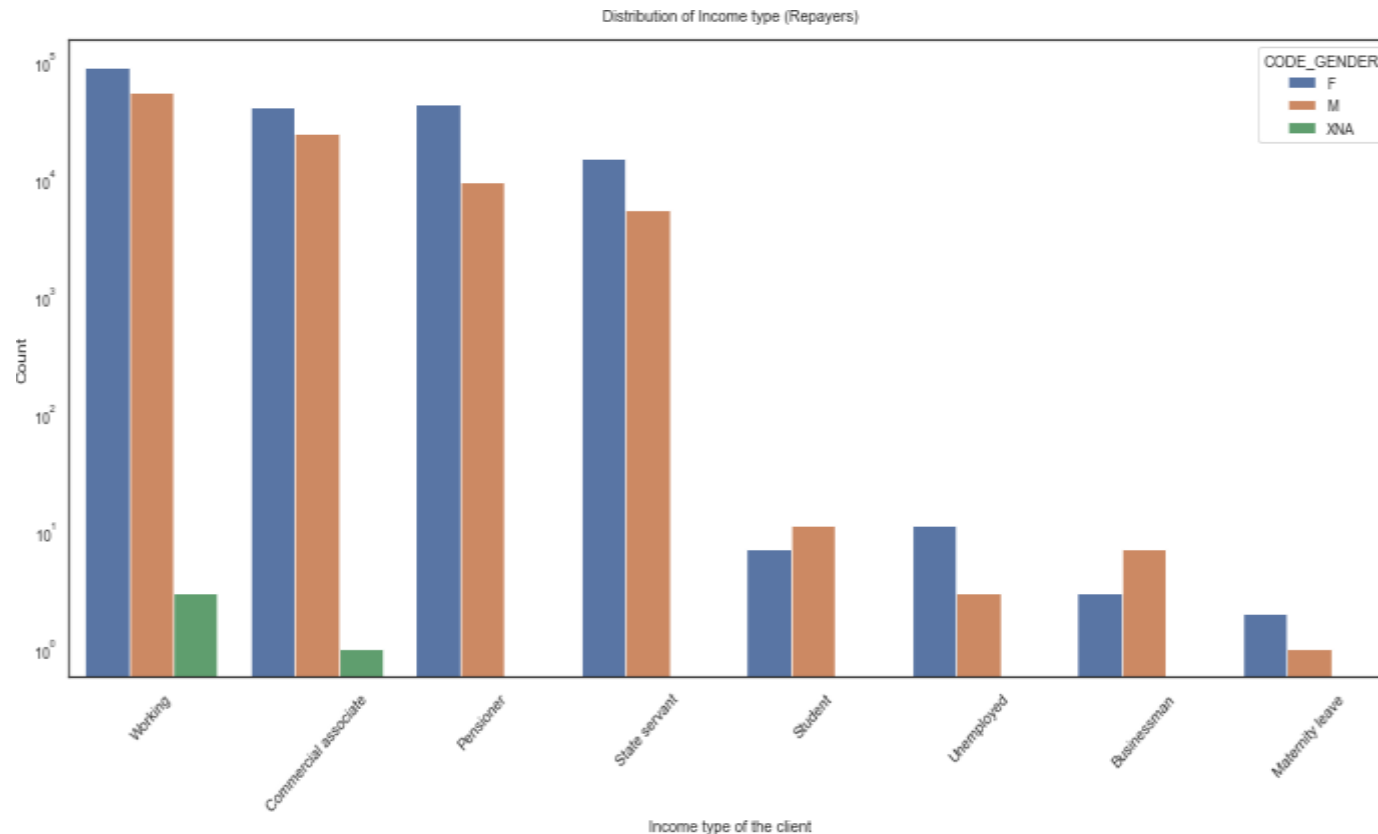
Inferences

In majority of the cases, female counts are higher than male.

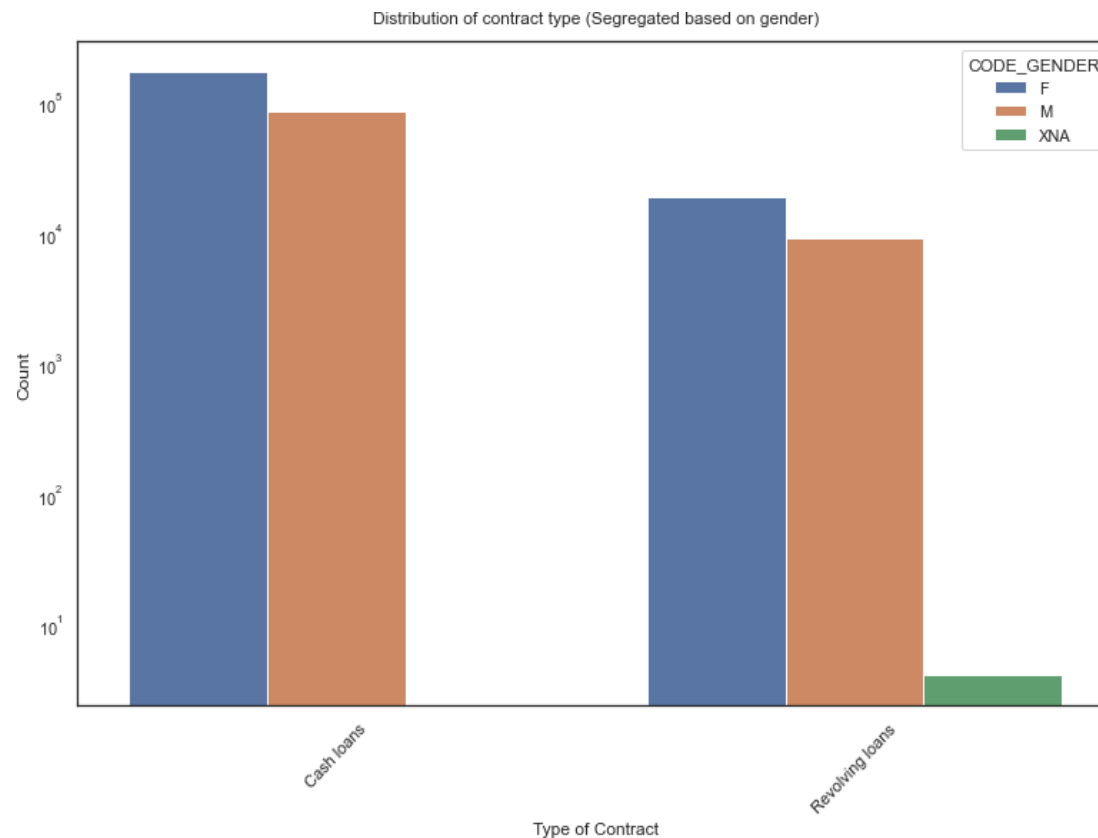
Income range from 100000 to 200000 is having more number of credits.

In the slots 250000-275000 and 375000-400000, the count for both males and females are almost the same

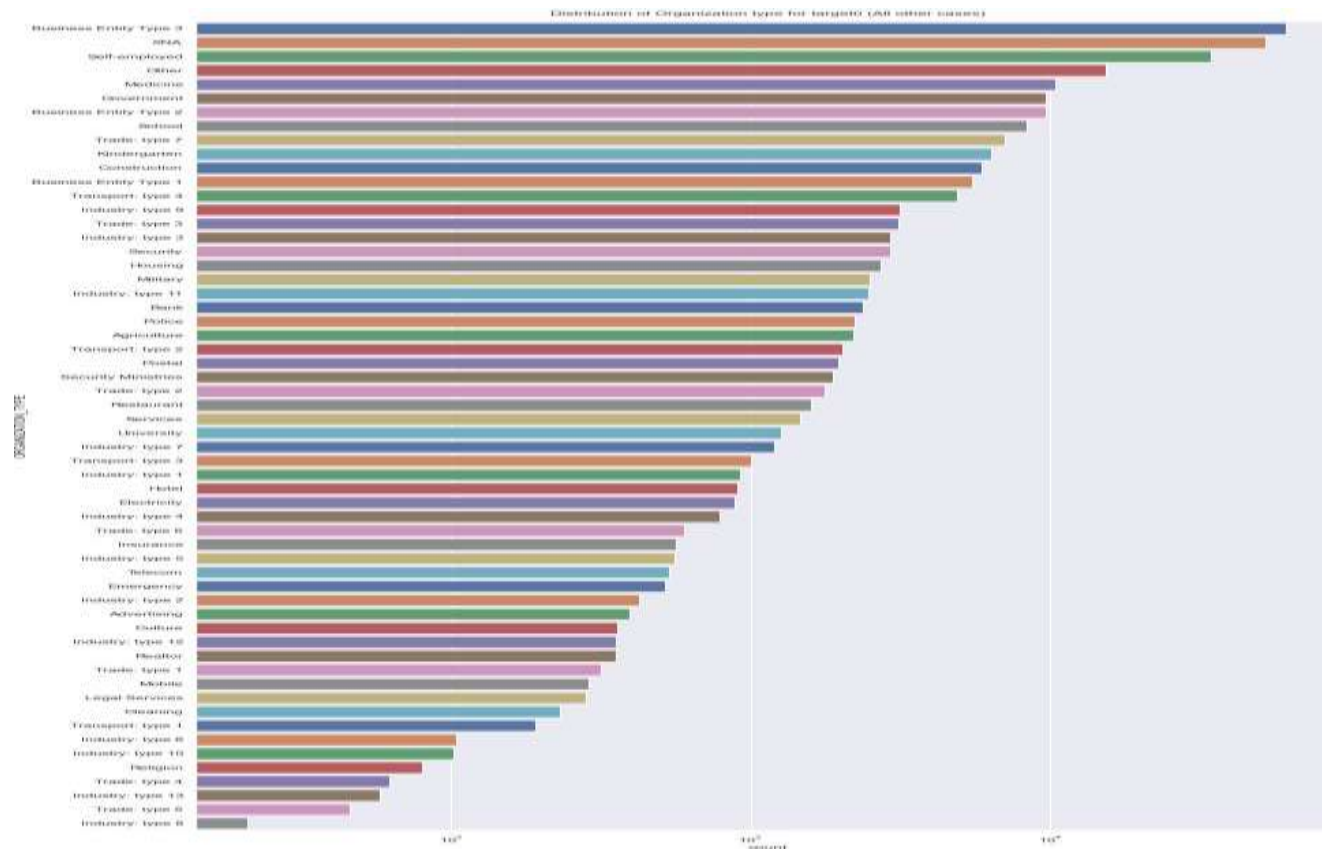
Plotting for Income type (NAME_INCOME_TYPE) (Segregated based on gender)



Plotting for Contract type (NAME_CONTRACT_TYPE) (Segregated based on gender)



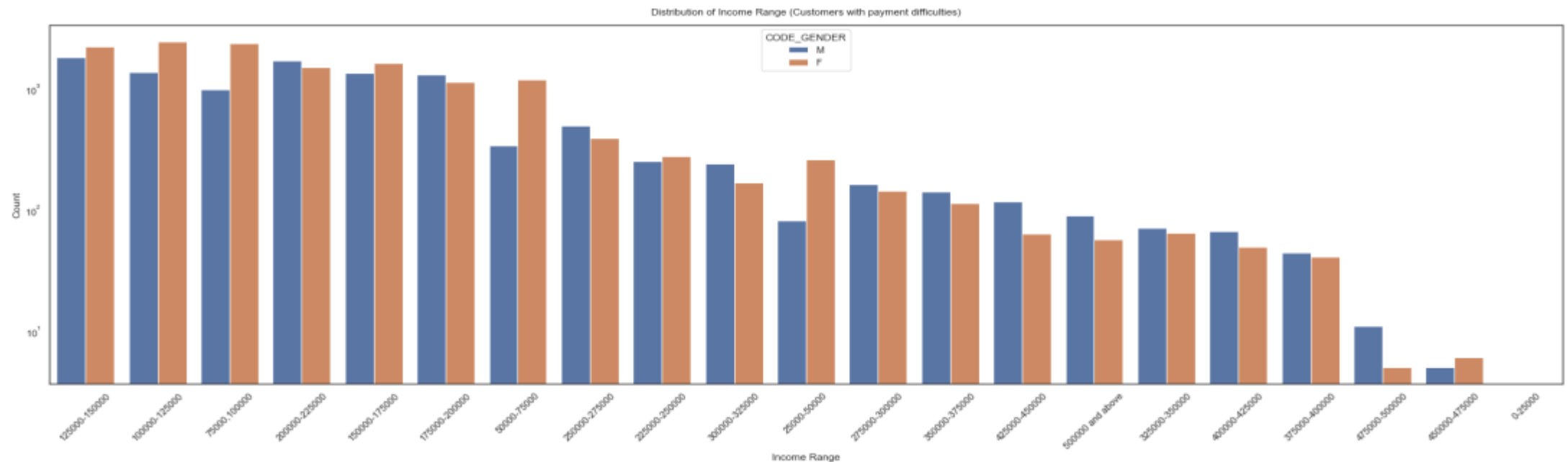
Plotting for Organization type in logarithmic scale



Inferences

Clients which have applied for credits are from most of the organization type 'Business entity Type 3', 'Self employed', 'Other', 'Medicine', 'Government' and 'Business entity type2'.

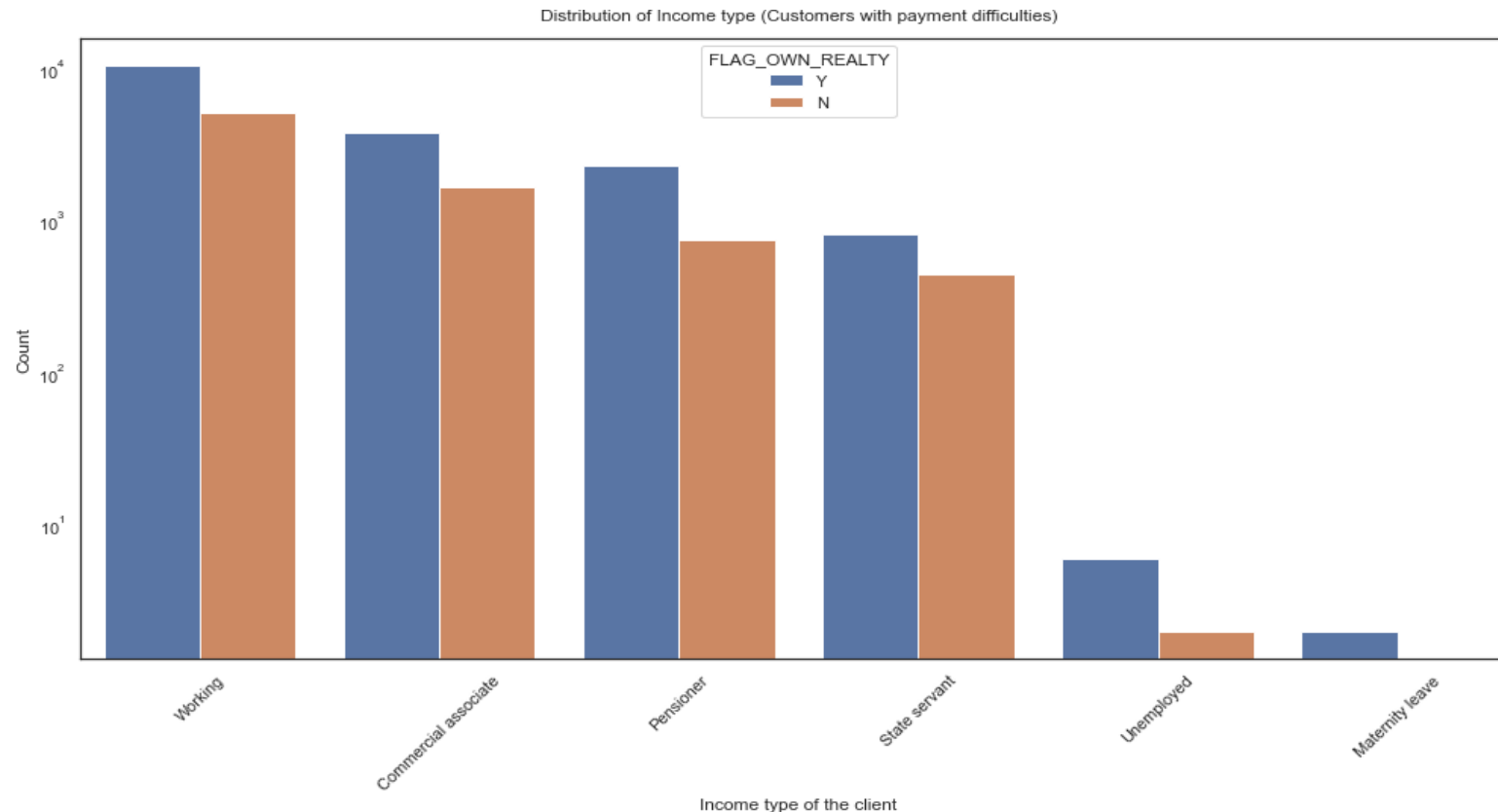
Plotting graphs for Target1 (Customers with payment difficulties)



Inferences

1. Income range from 100000 to 200000 is having the highest number of credits.
2. Very less count for income range 400000 and above.
3. On average, there are more number of male clients where the number of credits are less.

Plotting for Income type (Segregated based on house

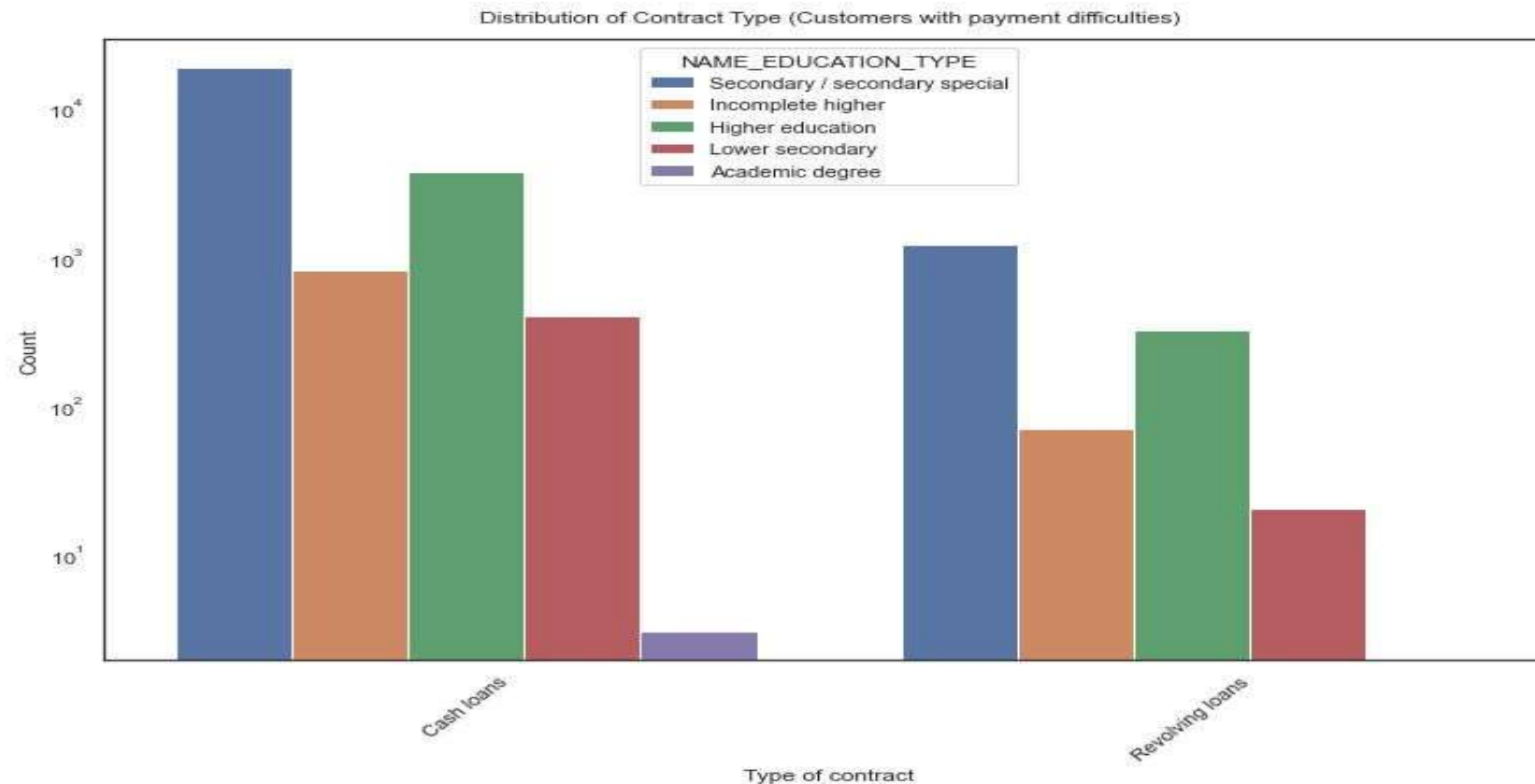


Inferences

1. Working customers, obviously, have a higher count.

2. As we can see, most customers do have their own property (house or a flat) but a large number of customers can be stated as otherwise.

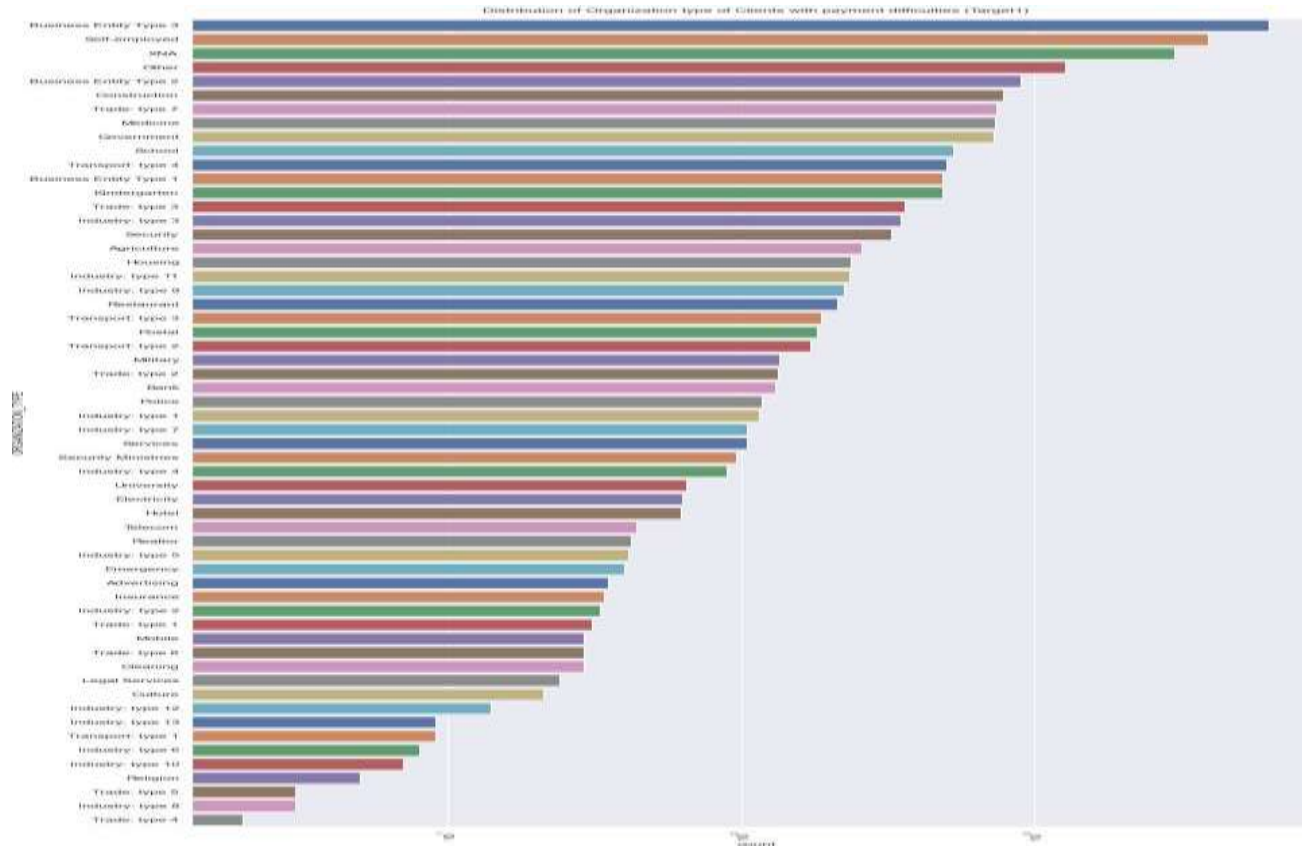
Plotting for Contract type (Segregated based on education level)



Inferences

1. Cash loans, as we can see, are preferred by clients of all education backgrounds with an overwhelming majority.
2. People with only an academic degree do not prefer revolving loans at all.

Plotting for Organization type



Inferences

1. As compared to the clients with NO payment difficulties, clients WITH payment difficulties have the 'construction' business type in the top 5 count replacing the 'medicine' business type.
2. Most of the business types are the same as clients with NO payment difficulties, except we have the business type 'Transport: type1' in the case of clients WITH payment difficulties which wasn't present before.

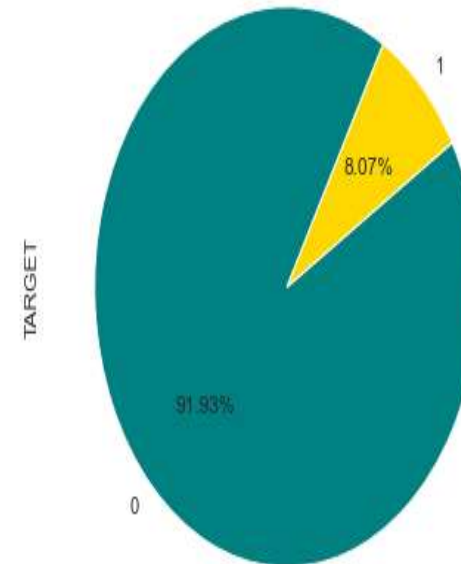
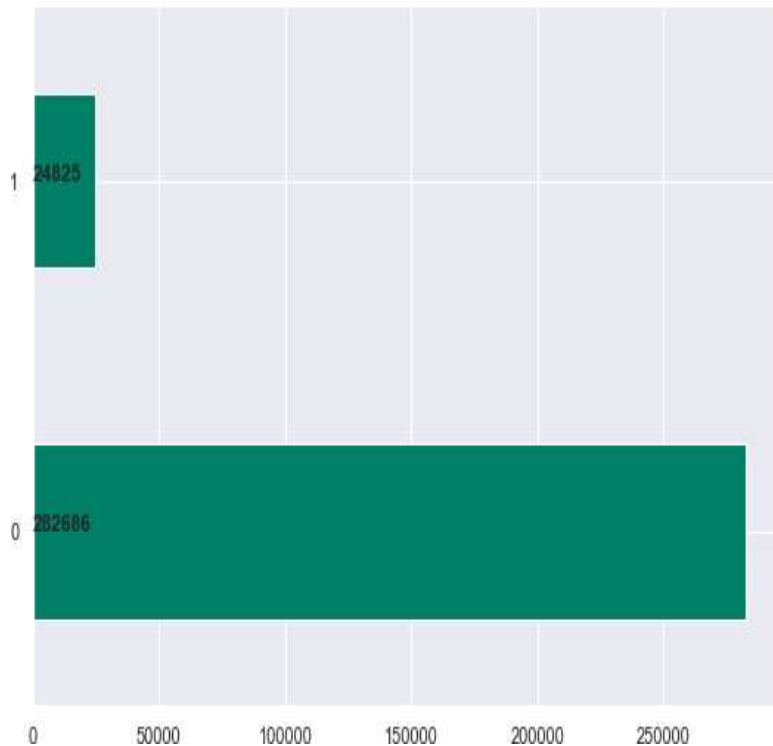
More Analysis to find patterns:

Distribution of Target variable¶

Target variable: 1 -client with payment difficulties

0 -all other cases

Distribution of clients with difficulties and all other cases

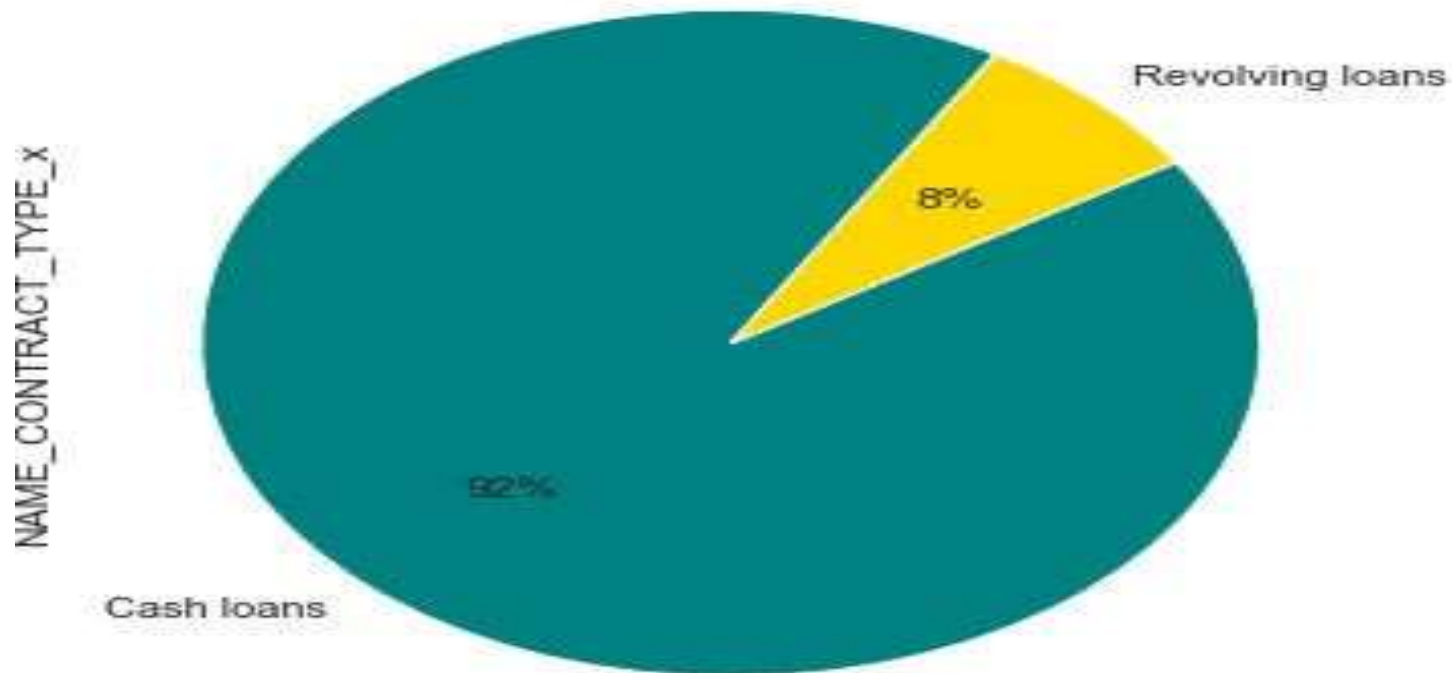


Inferences

8.79% (18547) out of total client population (192573) have difficulties in repaying loans.

Distribution in Contract types in data (Combined dataset)

distribution of contract types in data (combined dataset)



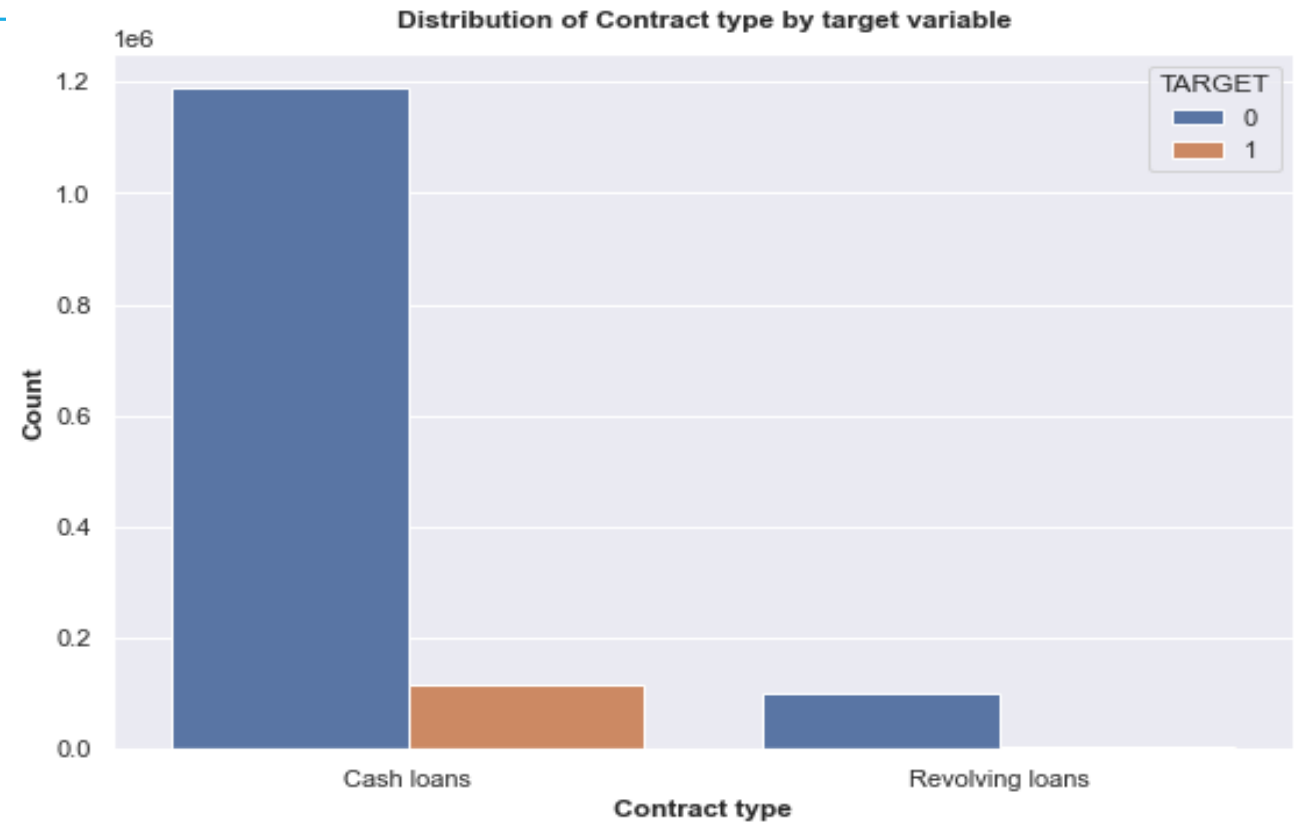
Inferences

The percentage of revolving loans and cash loans are 8% & 92%.

Point to infer from the graph

In the applicationData file, we saw females had 61% and males had 39% but now in the combined dataset we see:- Females: 62% Males: 38%

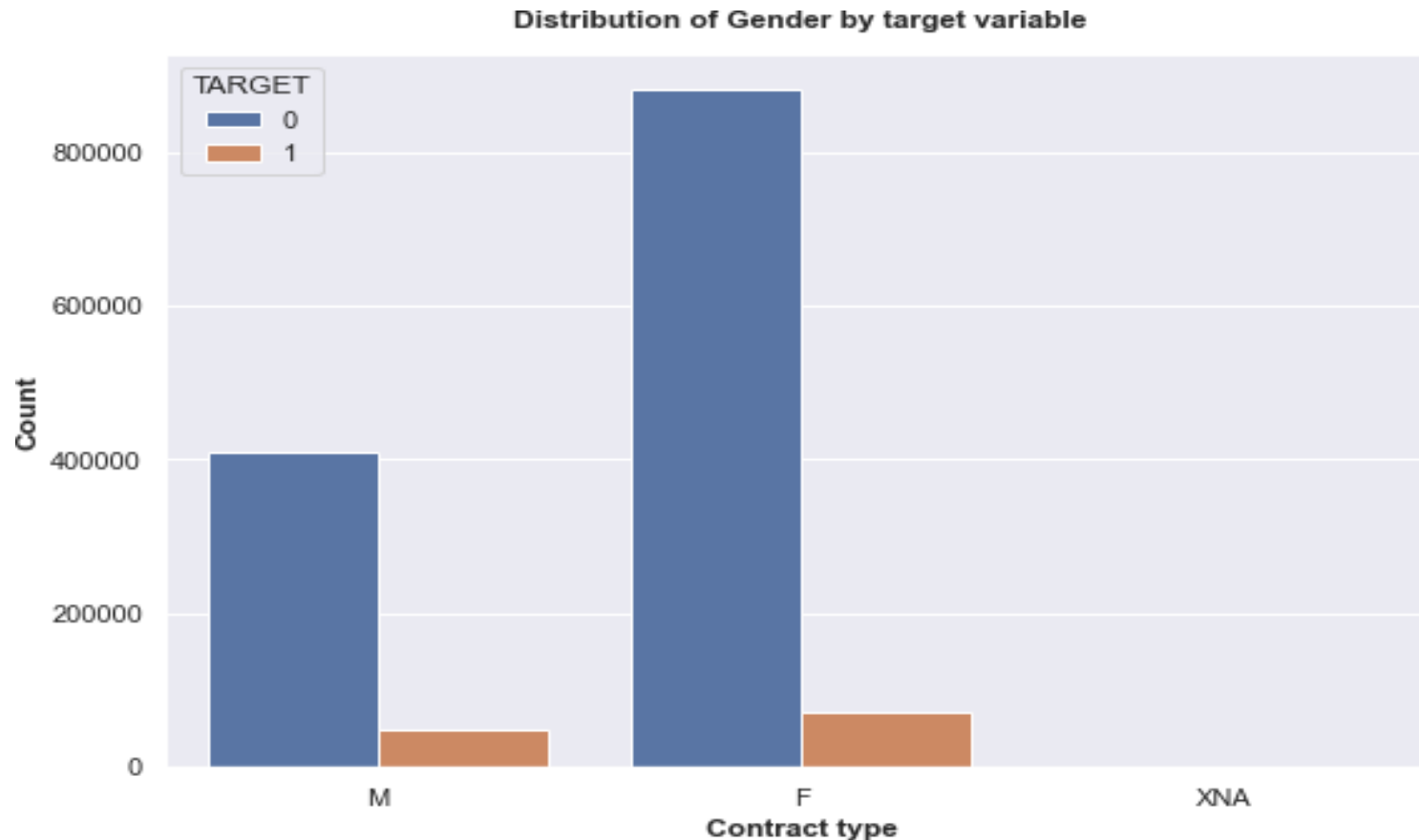
Distribution of Contract type by target (repayment status)



Inferences

Both set of clients (Target 0 and target 1) prefer cash loans over revolving loans with overwhelming numbers

Distribution of Gender by target (repayment status)

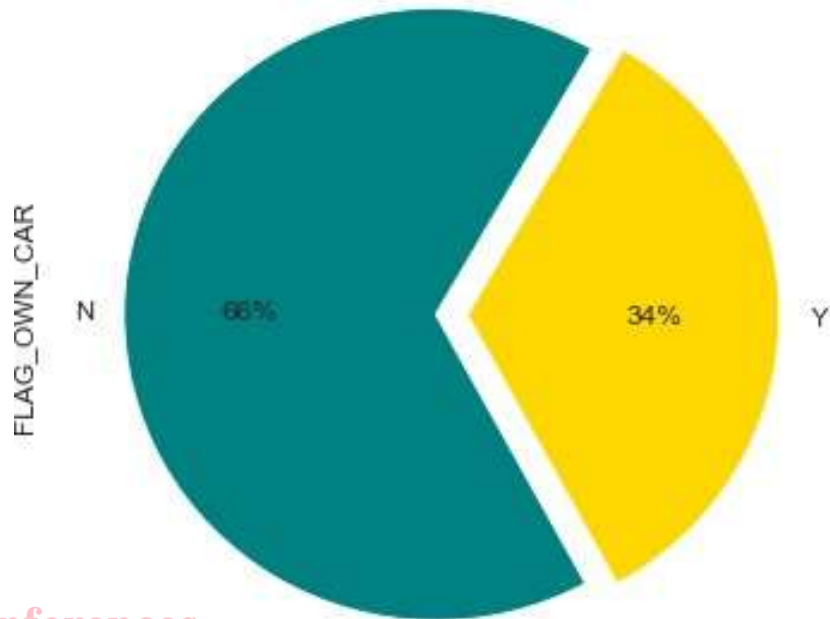


Inferences

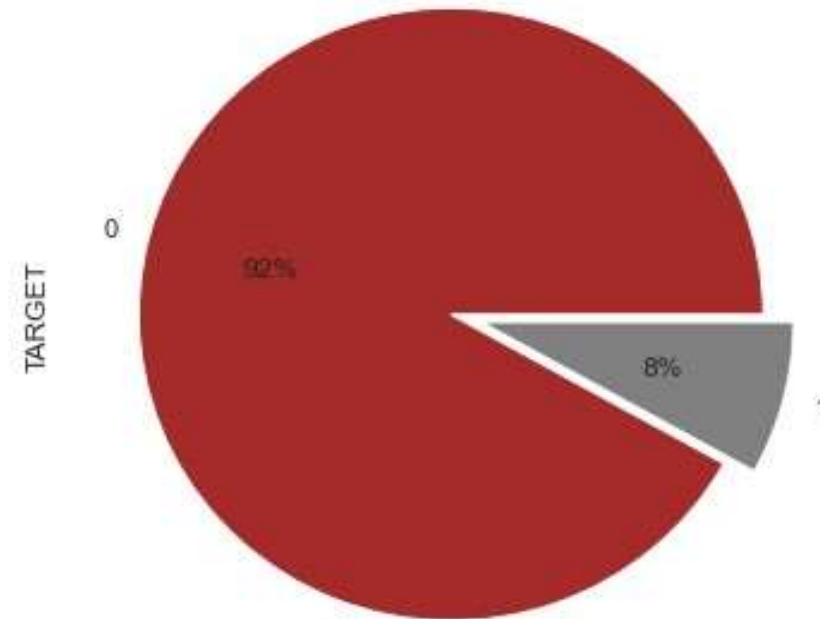
1. Clearly, female clients are the best repayers of their loan (almost double the amount of males).
2. Amount of defaulters in both genders are almost equally distributed.

Distribution of client owning a car and by target.

Distribution of Client by car ownership



Distribution of Client by car ownership based on repayment status

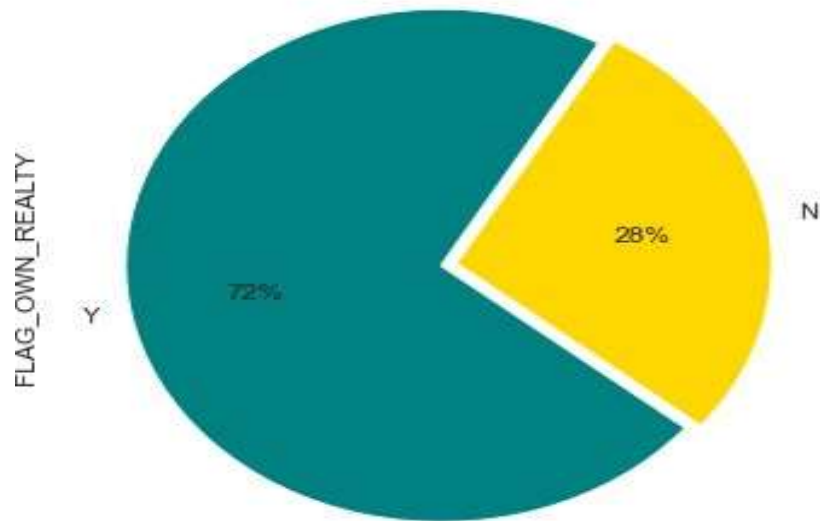


Inferences

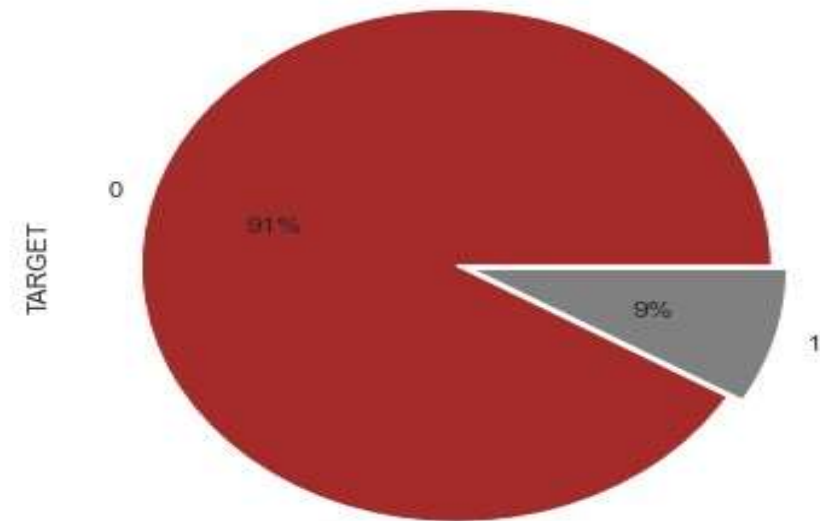
1st pie plot : Only 38% of clients own a car .2nd pie plot : Only 8% of clients who own a car have difficulty in payments

Distribution of client owning a house or flat and by target

Distribution of Client by house ownership



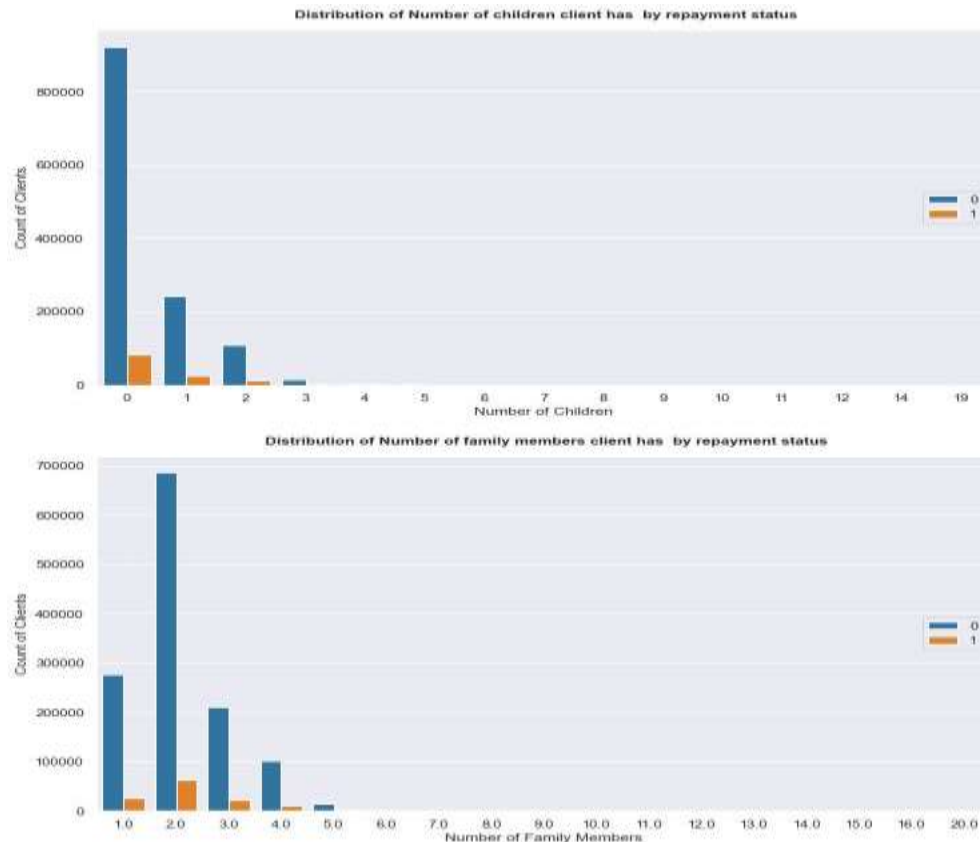
Distribution of client by house ownership based on repayment status



Inferences

SUBPLOT 1 : 71% of clients own a house or a flat. SUBPLOT 2 : Out of all the clients who own a house, 9% of clients have difficulty in making payments.

Distribution of Number of children and family members of client by repayment status (Based on target).



Inferences

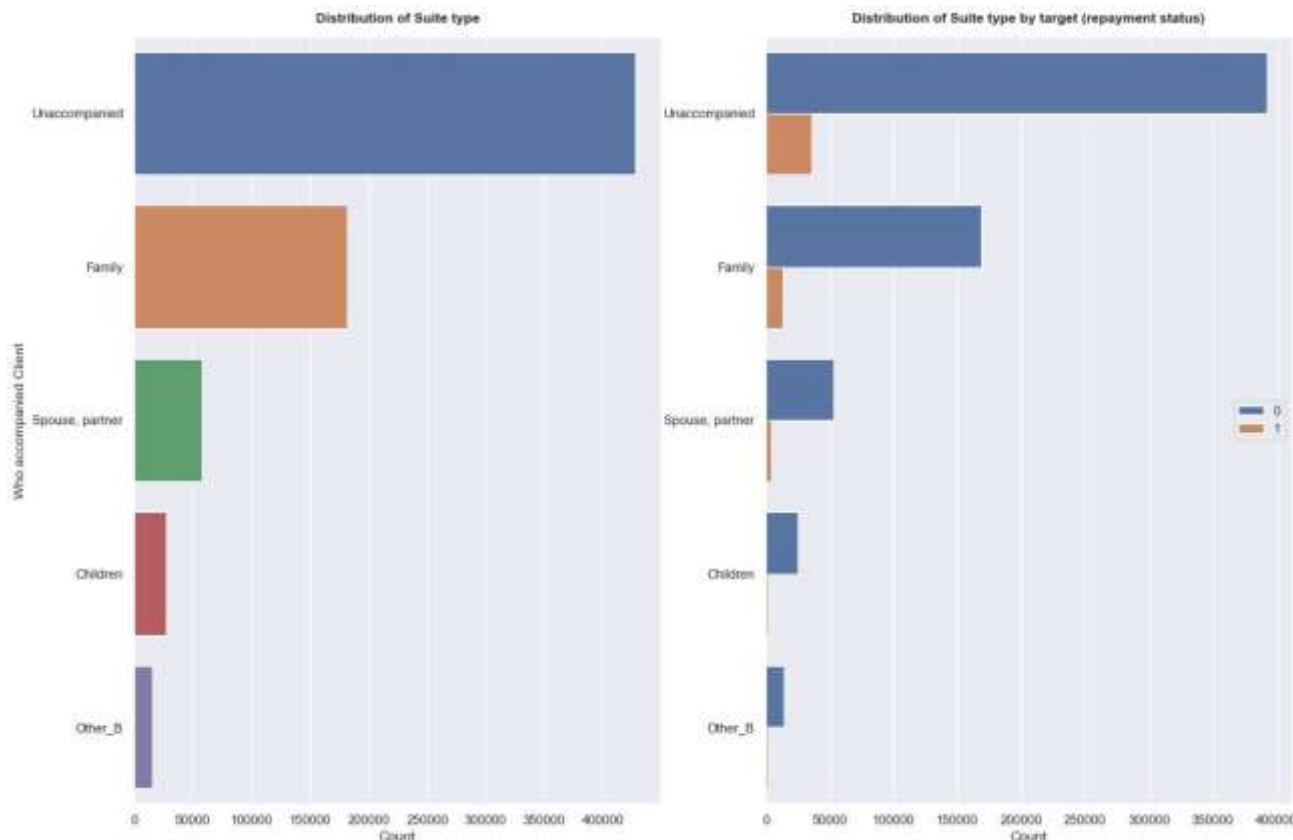
Subplot1:

1. The majority as per both cases of repayment status, have zero children.
2. Clients with more than 2 children do not have difficulty in making payments.
3. Clients with 0 children have the majority in terms of having difficulty in making payments.

Subplot2:

1. Clients with 2 family members living together are in high numbers as per both cases of repayment status
2. Also, from point 1, the majority of clients having difficulty in payments have 2 family members

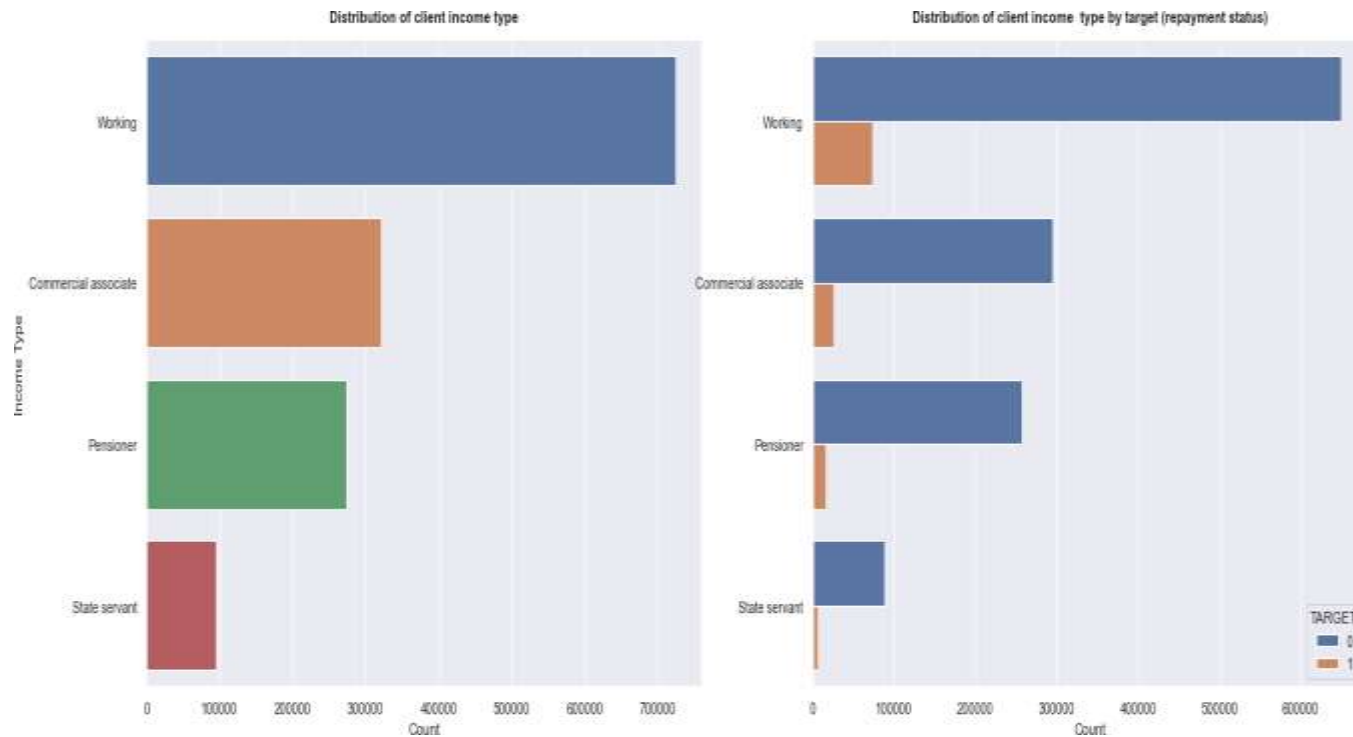
Distribution of Suite type



Inferences

1. Note: Missing data was labelled as 'missing' during the data cleaning process so we can ignore it. of the clients are (in both cases of repayment status) unaccompanied (without anyone to help/guide them)
2. Least amount of clients are in the company of their children.

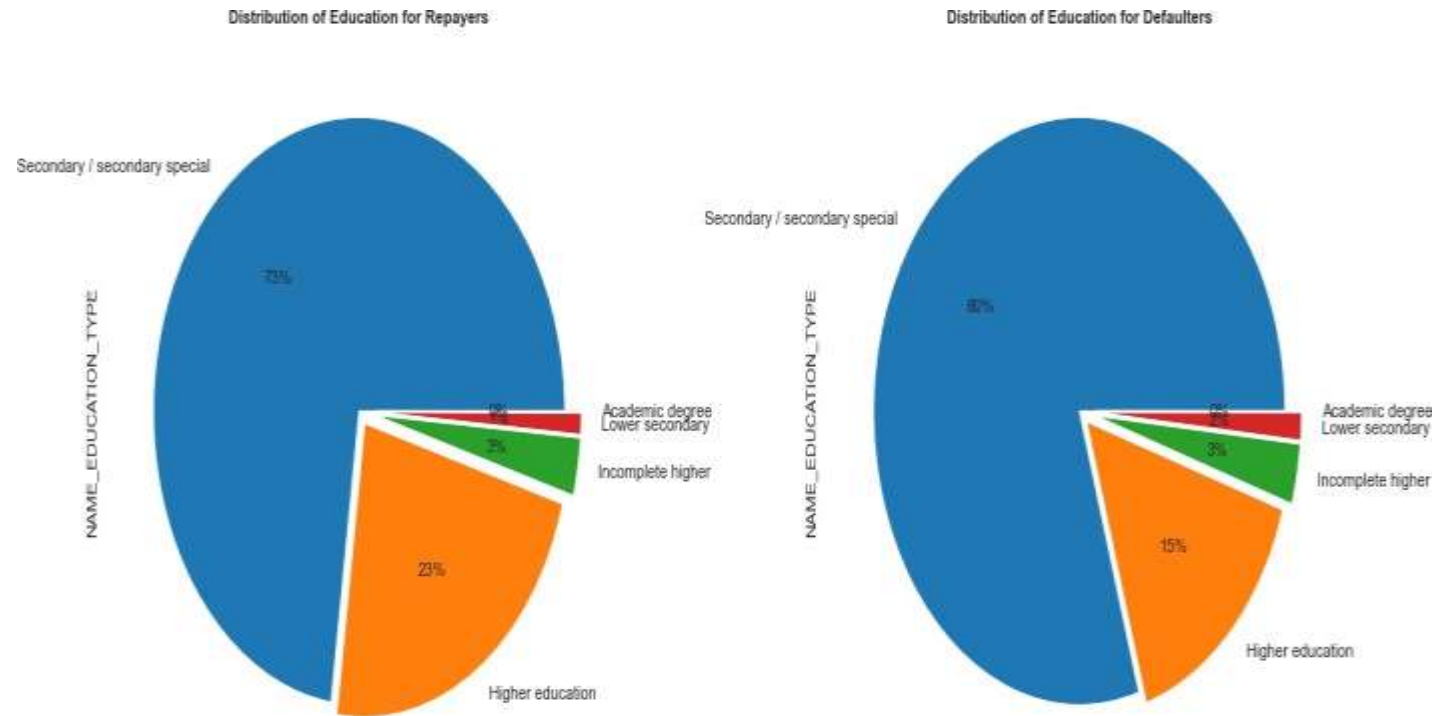
Distribution of client income type



Inferences

1. Most clients as per both cases of repayment status, are working.
2. Conversely, the least amount of clients are pensioners (retired clients)

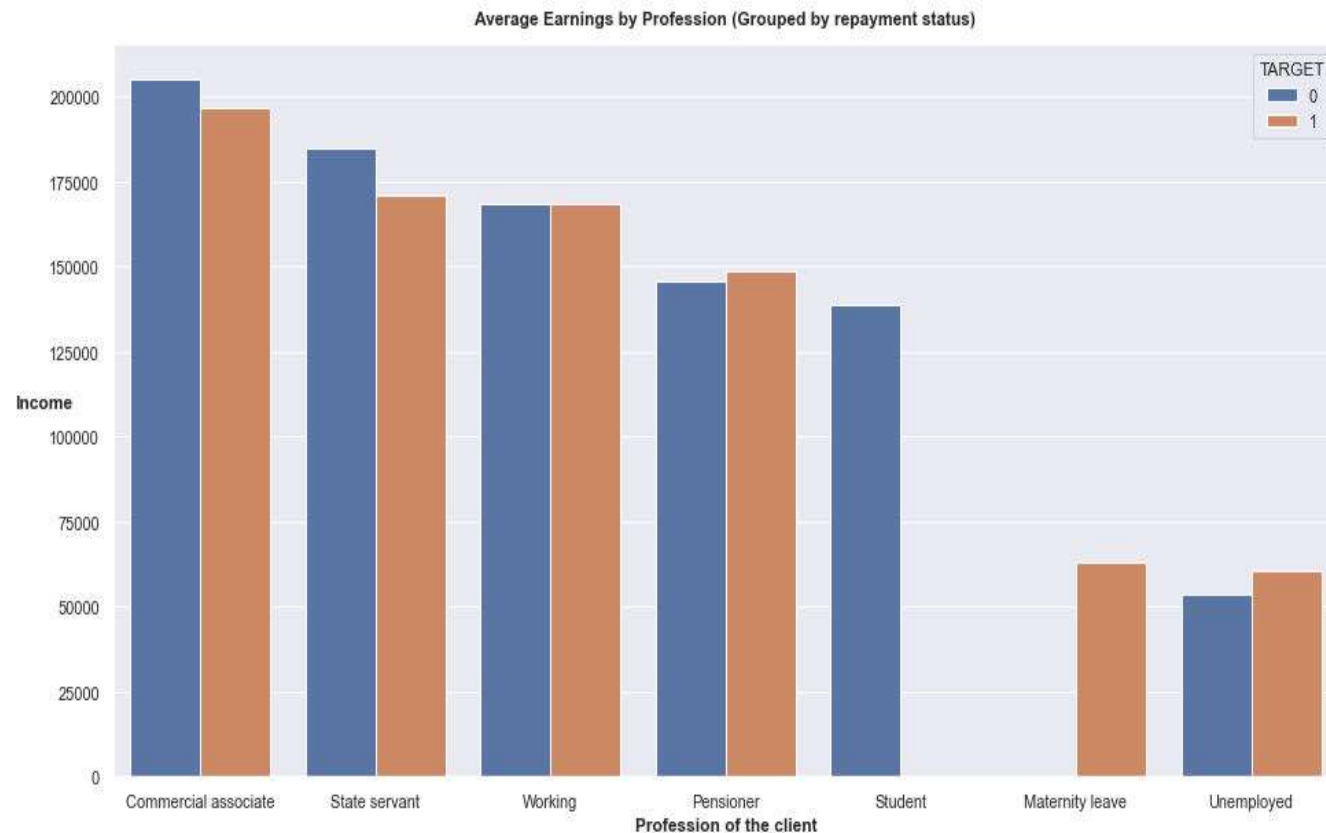
Distribution of Education type by repayment status



Point to infer from the graph

1. Clients who default are proportionally 9% higher compared to clients who do not default (for clients with education as secondary).
2. In the higher education category, clients who default are 8% fewer.
3. In both cases of repayment status, lower secondary and academic degree categories are the minority.

Average Earnings by different professions based on target (repayment status)

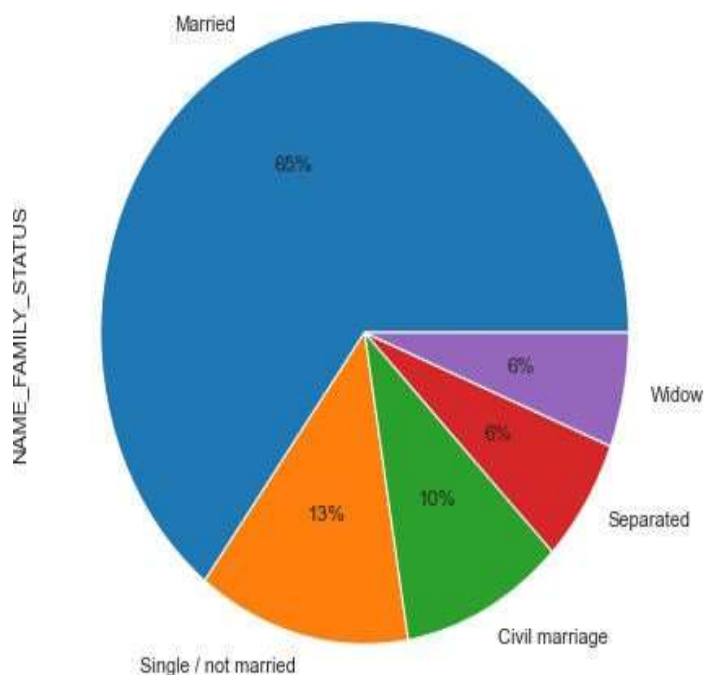


Inferences

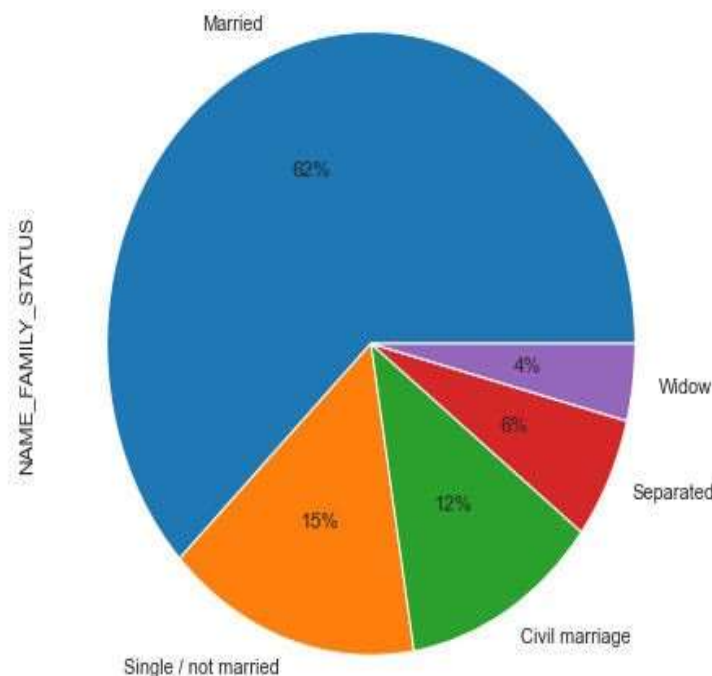
1. In both cases of repayment status, commercial associate clients are the highest earners.
2. Clients who are on maternity leave (therefore, female clients) have difficulty in making payments.
3. Pensioners and students do not have any difficulties in repayments.
4. There are almost an equal number of clients under the working category who repay and default.

Distribution of Education type by loan repayment status

Distribution of Family status for Repayers (Target0)



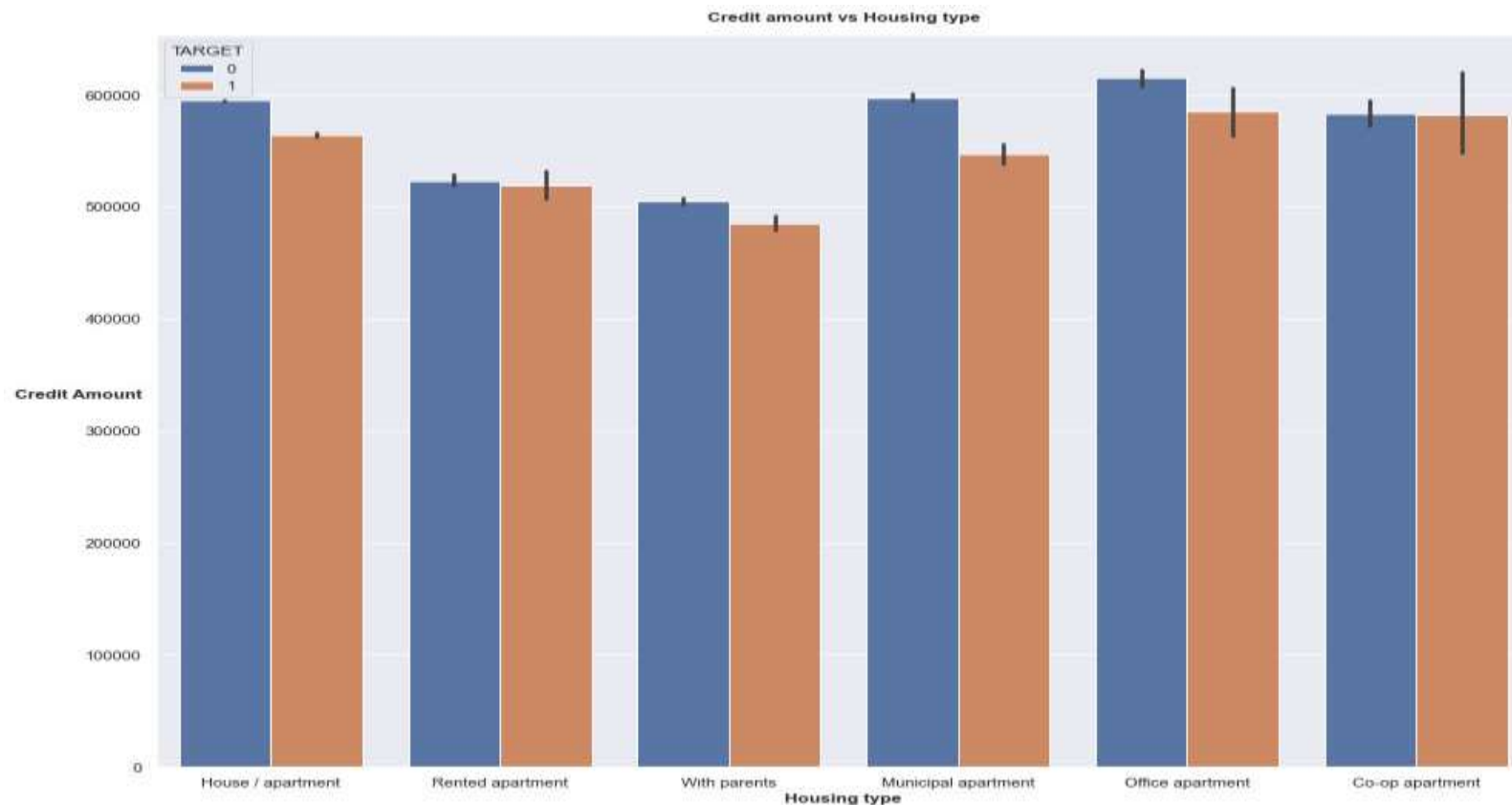
Distribution of Family status for Defaulters (Target1)



Inferences

1. There's a difference of -4% in married clients who have difficulty in making payments.
2. Family status for both cases of repayment status have an almost evenly distributed family status (family members living with the client)

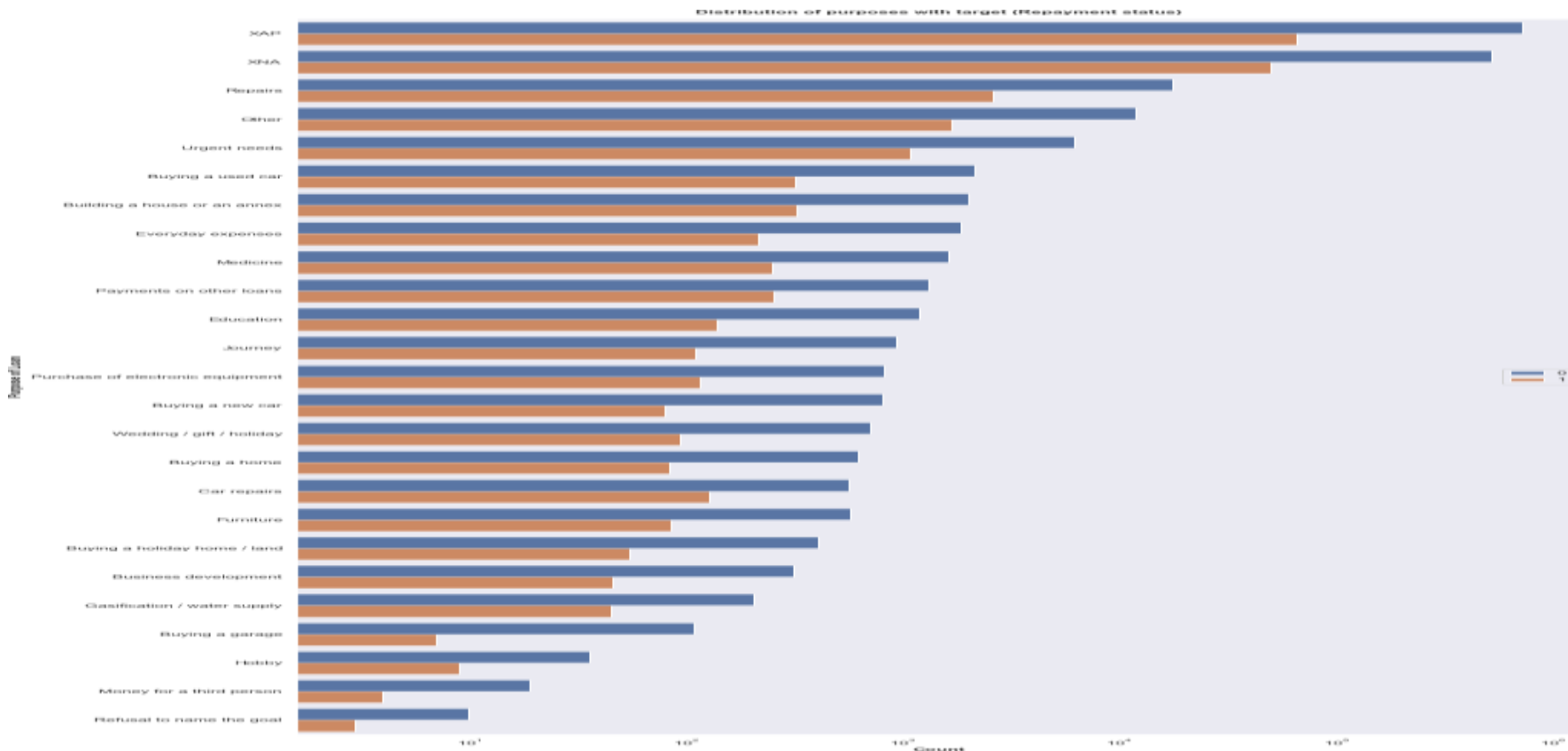
Distribution of credit amount and housing type



Inferences

1. Clients with office, co-op, municipal apartments have the highest repayers.
2. Clients living with parents or in a parents' apartment have the least amount of repayers and defaulters.

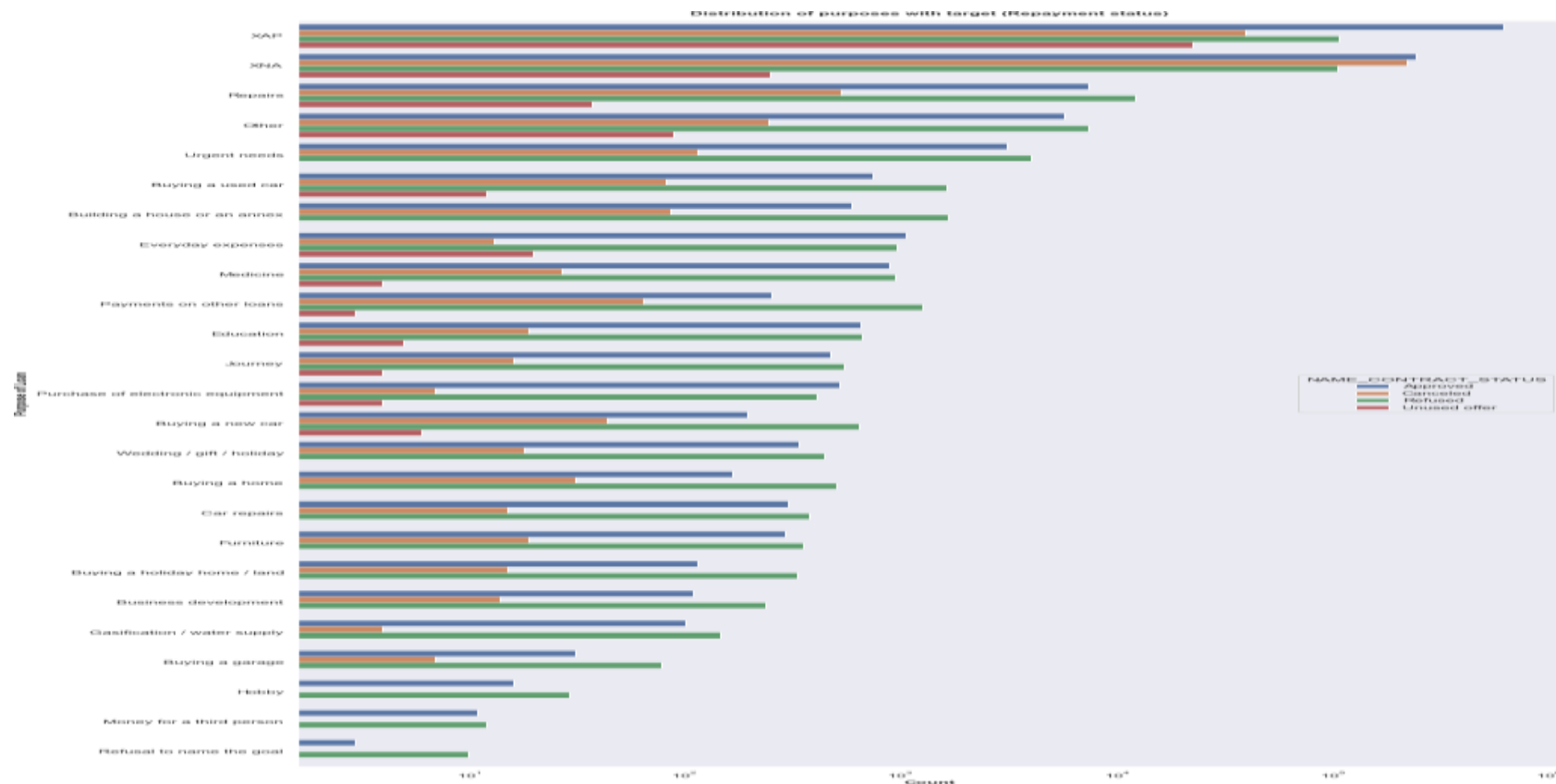
Distribution of Loan purpose (Segregated by repayment status (Target))



Inferences

1. Repair purposes are on top with most defaulters and repayers.
2. Proportion wise, there are high amount of repayers when the client refuses to name the purpose of the loan. Although such clients are rare.

Distribution of contract status



Inferences

1. Most rejection of loans is when the purpose of the client is based on Repairs. 2. For education purposes we have equal number of approvals and refusals.

TOP10 Correlation variables

From the above output, the top10 correlated columns are:
(double click to view in proper format)

OBS_30_CNT_SOCIAL_CIRCLE OBS_60_CNT_SOCIAL_CIRCLE
1.00

AMT_CREDIT_y AMT_APPLICATION 0.97

DAYS_TERMINATION DAYS_LAST_DUE 0.93

CNT_FAM_MEMBERS CNT_CHILDREN 0.90

REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION 0.88

DEF_30_CNT_SOCIAL_CIRCLE DEF_60_CNT_SOCIAL_CIRCLE
0.87

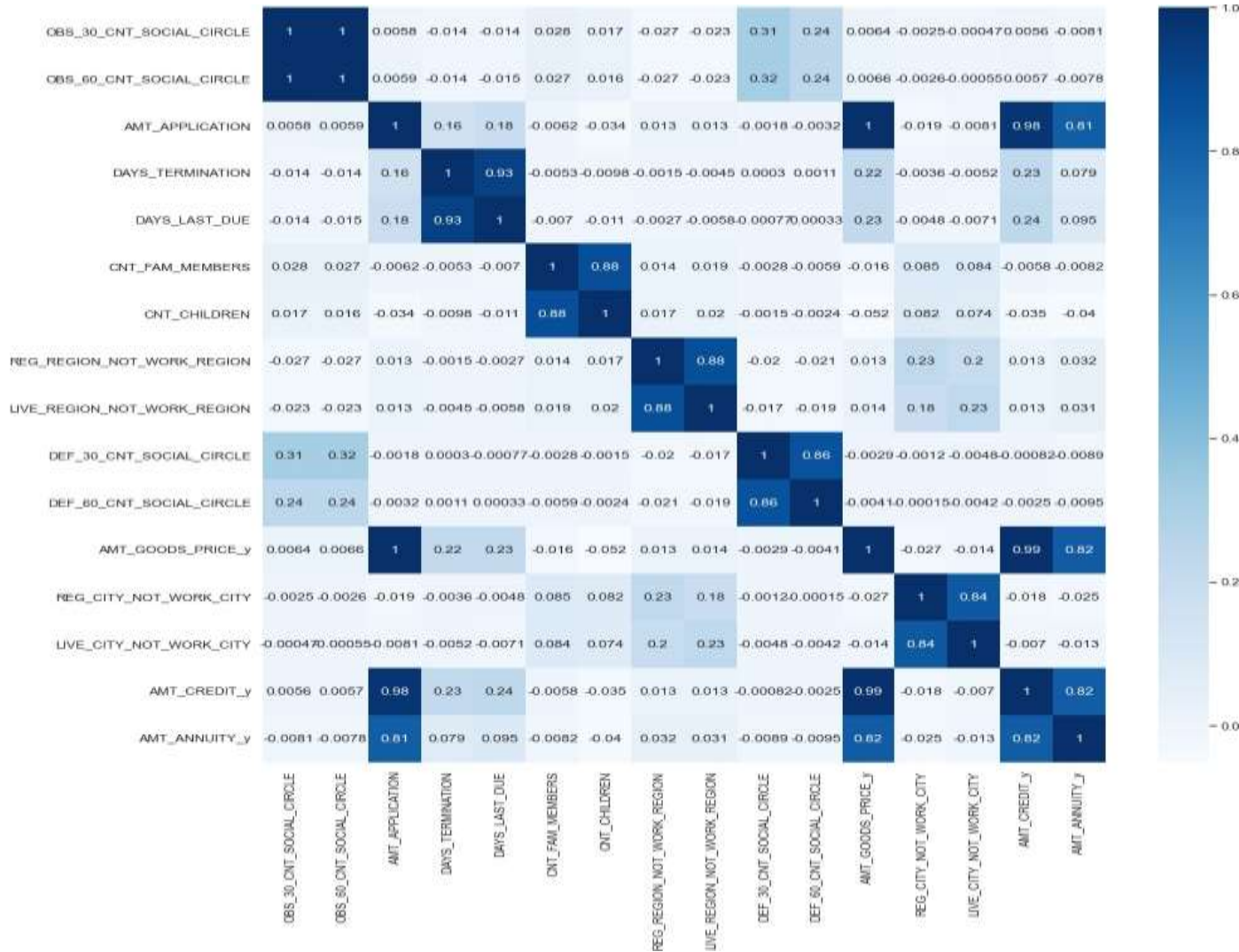
AMT_GOODS_PRICE_y AMT_CREDIT_y 0.86

AMT_APPLICATION AMT_GOODS_PRICE_y 0.85

REG_CITY_NOT_WORK_CITY LIVE_CITY_NOT_WORK_CITY
0.83


AMT_CREDIT_y AMT_ANNUITY_y 0.81

Visually showcasing the top10 correlated columns through a heatmap



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_ANNUIITY 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. If the client's contact address does not match the work address, then there's a high chance that the client's permanent address also does not match the work address. 4. First due of the previous application is highly correlated with Relative to the expected termination of the previous application 5. CNT_CHILDREN and CNT_FAM_MEMBERS are highly correlated which means a client with children is highly likely to have family members as well.



More Analysis to find patterns:
Distribution of Target variable
Target variable:
1 - client with payment difficulties
0 - all other cases

OBS_60_CNT_SOCIAL_CIRCLE OBS_30_CNT_SOCIAL_CIRCLE 1.00

AMT_APPLICATION AMT_CREDIT_y 0.97

DAYS_TERMINATION DAYS_LAST_DUE 0.95

CNT_FAM_MEMBERS CNT_CHILDREN 0.90

LIVE_REGION_NOT_WORK_REGION REG_REGION_NOT_WORK_REGION 0.87

DEF_30_CNT_SOCIAL_CIRCLE DEF_60_CNT_SOCIAL_CIRCLE 0.86

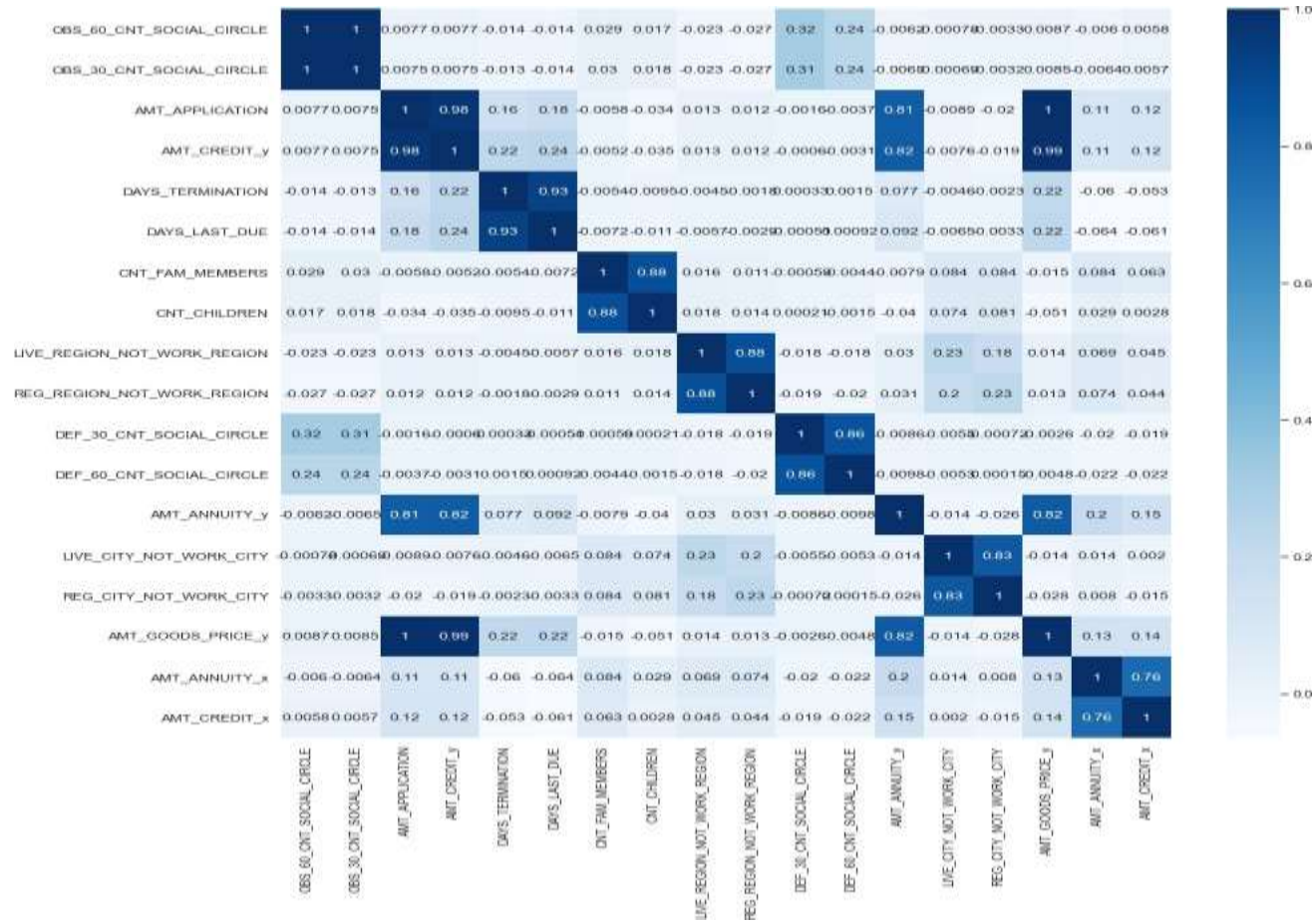
AMT_CREDIT_y AMT_ANNUIITY_y 0.83

LIVE_CITY_NOT_WORK_CITY REG_CITY_NOT_WORK_CITY 0.78

AMT_ANNUIITY_y AMT_GOODS_PRICE_y 0.76

AMT_ANNUIITY_x AMT_CREDIT_x 0.74

Adding the top10 correlated columns into a new dataframe:



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. In comparison to the repayer heatmap, If the client's contact address does not match the work address, then there's a high chance that the client's permanent address also does not match the work address. 4. Higher the goods price, higher the credit by the client 5. 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) 6. CNT_CHILDREN and CNT_FAM_MEMBERS are highly correlated which means a client with children is highly likely to have family members as well (same as with the repayer heatmap)

Conclusion from the Analysis

Banks must target more on contract type 'Student', 'Pensioner' and 'Businessman' for profitable business

Banks must focus less on income type 'Working' as it has most number of unsuccessful payments in order to get rid of financial loss for the organization

1. 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.

2. Female clients on maternity leave should **NOT** be targeted as they have no record of repayments (therefore they are highly likely to default and targeting them would lead to a loss)

3. While clients living with parents have the least amount of repayers, they also have the least amount of defaulters. So, in cases where the risk is less, such clients can be **TARGETED**.

4. Clients who are working need to be targeted **LESS** by the bank as they have the highest amount of defaulters.

5. Clients should **NOT** be targeted based on their education type alone as the data is very inconclusive.

6. Banks **SHOULD** target clients who own a car.

7. There are **NO** repayers/negligible repayers when the contract type is of revolving loan.

8. Banks **SHOULD** target more people with no children.

9. '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.

10. Banks **SHOULD** also target female clients as they are the highest repayers (almost as double as males) amongst both the genders.