CASE STUDY

Submitted by:
Jyoti

PROBLEM STATEMENT

The aim is to identify patterns which can help to understand the clients:

- People who will have difficulties to pay the loan amount should not be given the Loan as there will be business loss.
- People who will make the payments on time should be given the loan as it will help to get the profit.

STEPS

- There are 2 datasets, first will take the application dataset, understand all the dataset and checks the columns description to understand all the parameters involved in the problem.
- After importing the dataset and the libraries will find out the shape, info, description, which will give us the no of rows, columns, null values, mean, median, standard deviation.
- The next step is Data Cleaning, which is the important step, to check the % of null values and how we can impute those values, dropping the columns which are not relevant to the problem at the time.

- So, after checking the missing values, I dropped the columns which were having missing values more than 40% for further positive analysis.
- At the start, dataset had 121 columns, after dropping 40% I was left with 72 columns.
- Then , Took 5 columns which were having null values less than 13% , I have imputed them mean, median or mode after dividing into continuous and categorical variables.
- Categorical columns- OBS_30_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE,NAME_TYPE_SUITE
- Continuous columns- EXT_SOURCE_2,AMT_GOODS_PRICE, AMT_ANNUITY
- I have imputed categorical columns with mode.

- Then converting -ve values into +ve values in “DAYS_LAST_PHONE_CHANGE”, “DAYS_ID_PUBLISH”,"DAYS_REGISTRATION,”DAYS_EMPLOYED”,"DAYS_BIRTH" by using abs() function.
- Converted “DAYS_EMPLOYED" ,”DAYS_BIRTH”,”DAYS_REGISTRATION", “DAYS_ID_PUBLISH”into years.
- Then , from the dataset which columns were not required or did not have any significant impact we can drop them.
- After dropping and filtering those columns we were left with 42 columns.
- After checking the uniques values, “CODE_GENDER” had 3 unique, but there can be only 2 genders Female and Male so I have replaced XNA values with Female and in “ORGANIZATION_TYPE" had XNA values I replaced them with NaN.

CHECKING OUTLIERS

- Checked outliers for 5 continuous columns - “CNT_CHILDREN”,"AMT_CREDIT","AMT_ANNUITY","AMT_GOODS_PRICE","CNT_FAM_MEMBERS" and found out that:-
- CNT_CHILDREN is the count of children a client has so,the values greater than 3 can be considered as outliers.
- AMT_CREDIT is the credit amount of the loan. The values greater than 1.5 can be considered as outliers.
- AMT_ANNUITY is the loan annuity. The values greater than 600000 are consbidered as outliers.
- AMT_GOODS_PRICE is the price of the good for which loan has been given, so above 1.5 we can consider outliers.
- CNT_FAM_MEMBERS is count of family members, so above 4 we can consider outliers.

BIN CREATION

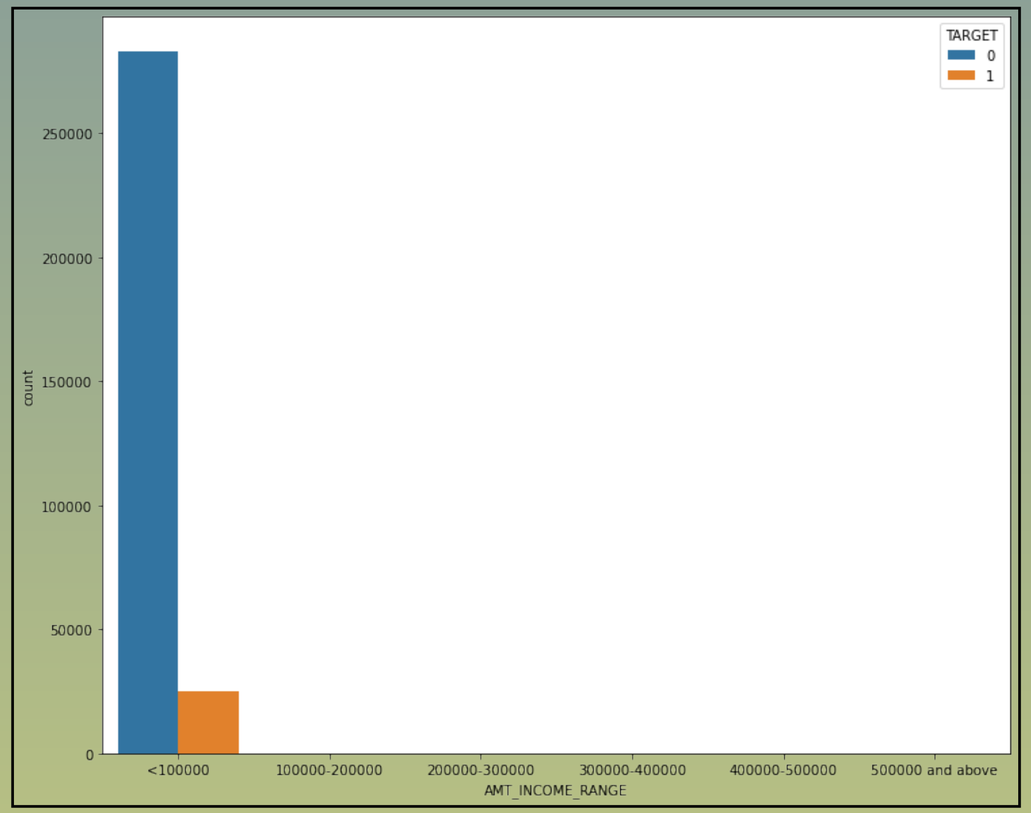
- Need to bin the columns “AMT_INCOME_TOTAL” and “AMT_CREDIT" and created 2 new columns “AMT_INCOME_RANGE” and “AMT_CREDIT_RANGE”.
- This is done to ease out the analysis of the data.

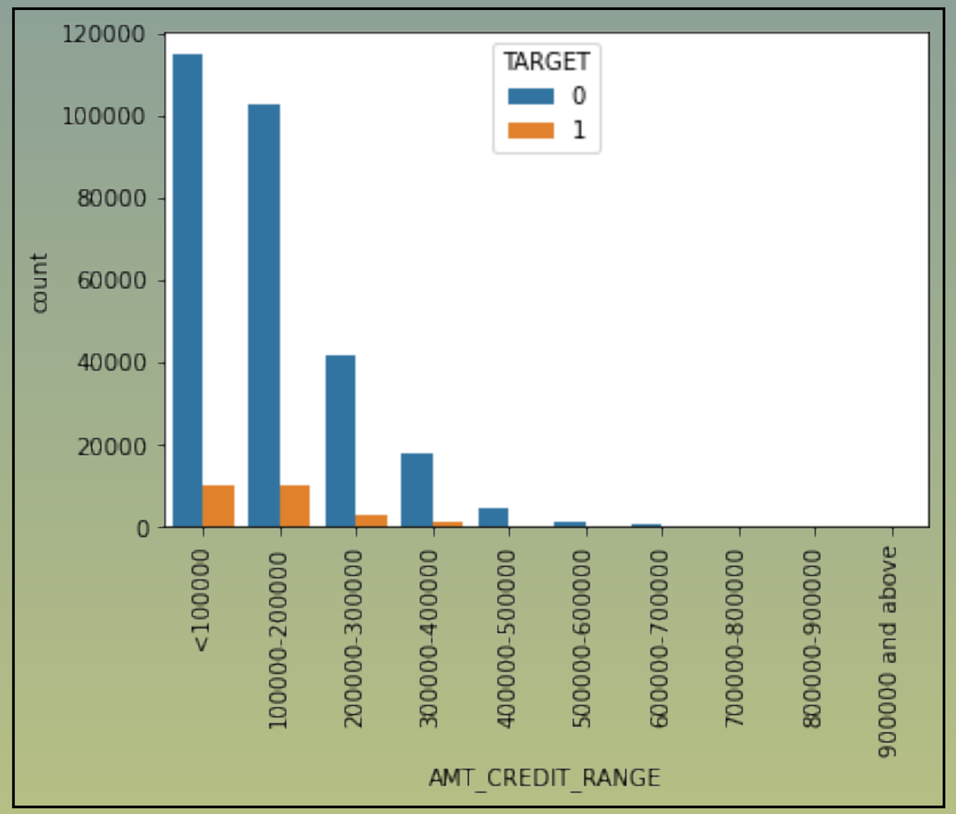
DATA ANALYSIS

- First of , splitted the “TARGET”column into 2 columns as it has 2 values 0, 1.
- df_0 and df_1 is non-defaulters and defaulters respectively.
- After checking out the % it seems Perc_nondefaulters:(91.93 %) is more than Perc_defaulters(8.07 %).
- After checking the Data Imbalance which comes out to be 11.38.

UNIVARAITE ANALYSIS OF CONTINUOUS COLUMNS

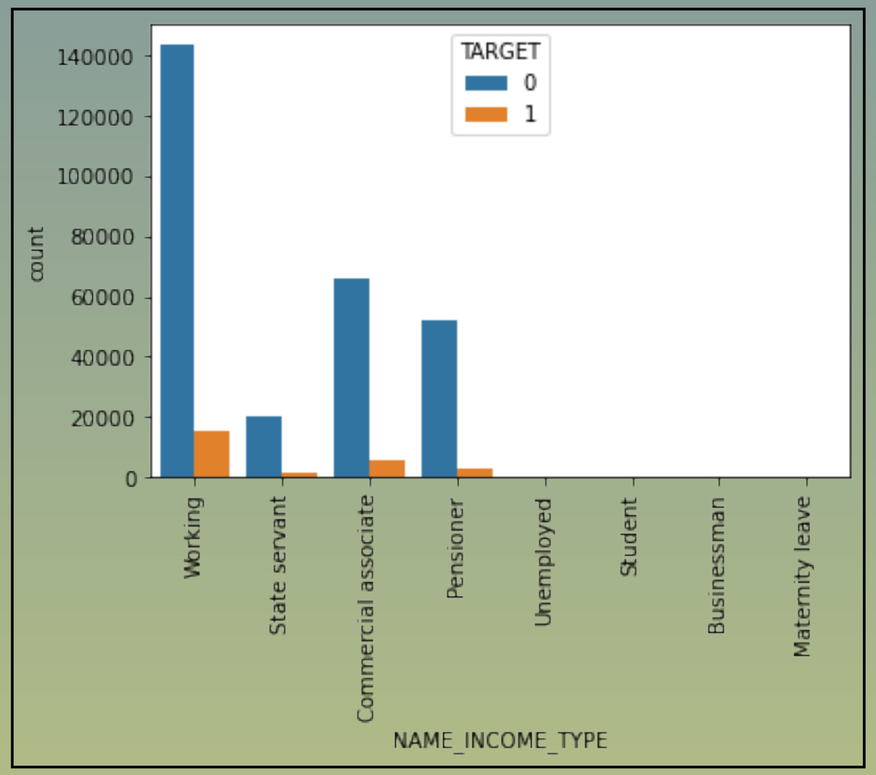
AMT_INCOME_RANGE AND AMT_CREDIT_RANGE

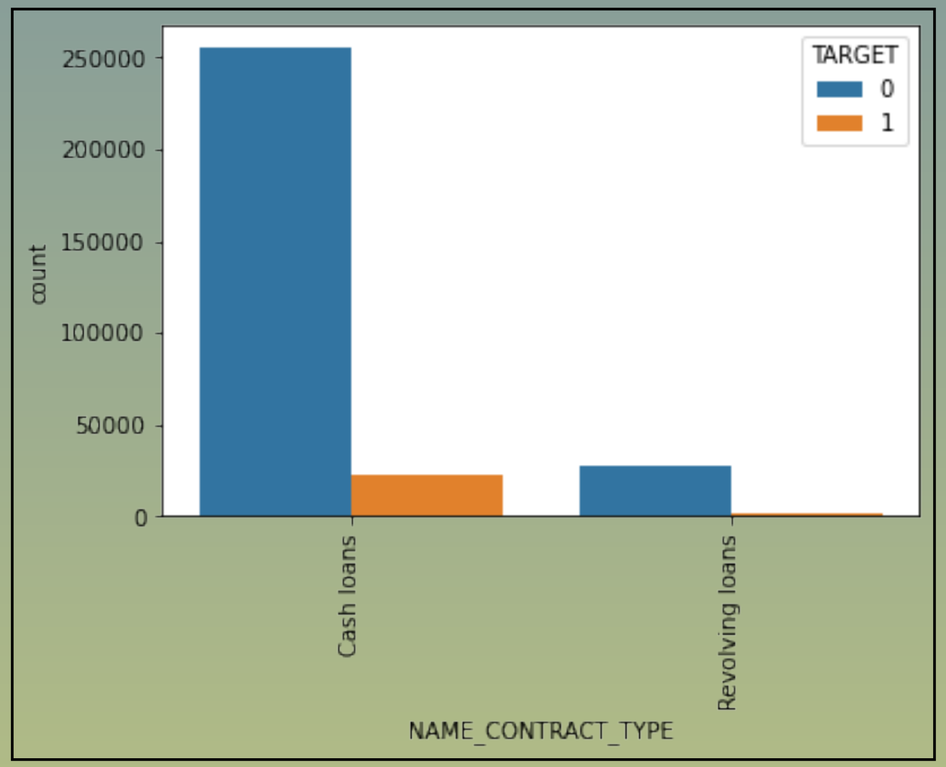




- In amount income range, we can see that low income people <100000 will become defaulters and same time they are the with most no. of loans.
- In amount credit range, defaulters are in low range .

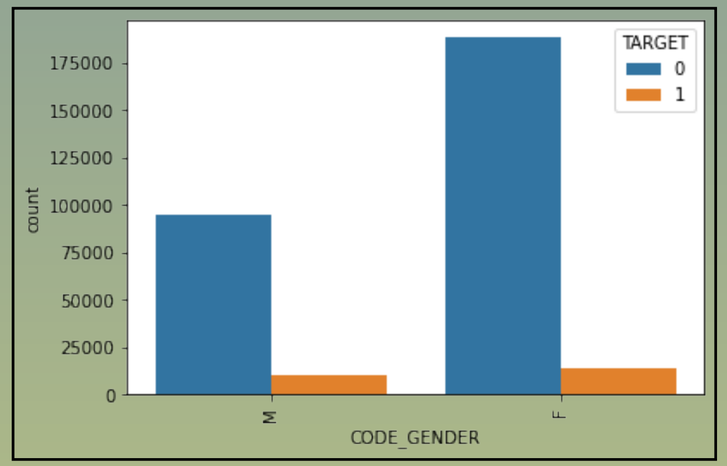
INCOME TYPE AND CONTRACT TYPE

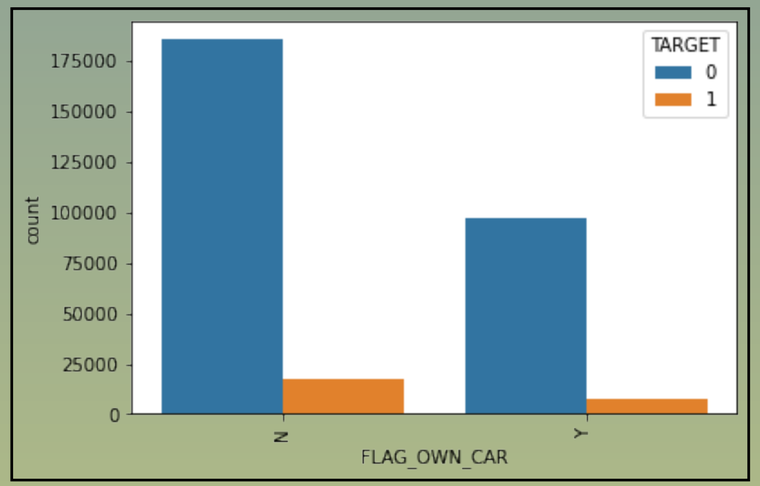


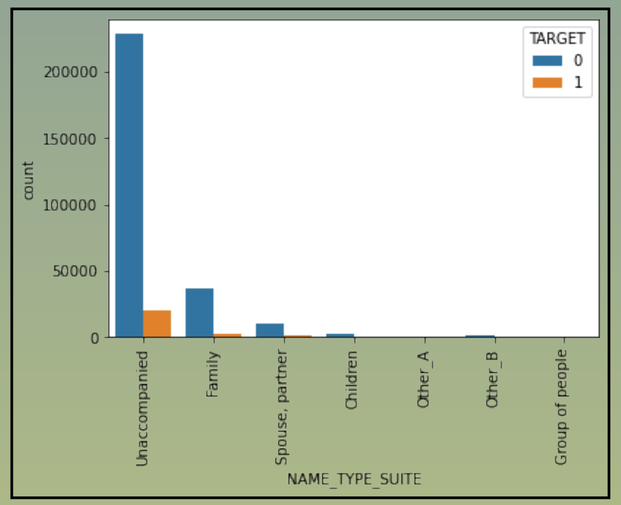


- Working customers seems to become more defaulters as compare to state servant and associate and student, businessman can pay loans easily.
- Cash loans are higher in range and they likely become defaulters than revolving loans.

CODE_GENDER , FLAG_OWN_CAR, NAME_TYPE_SUITE

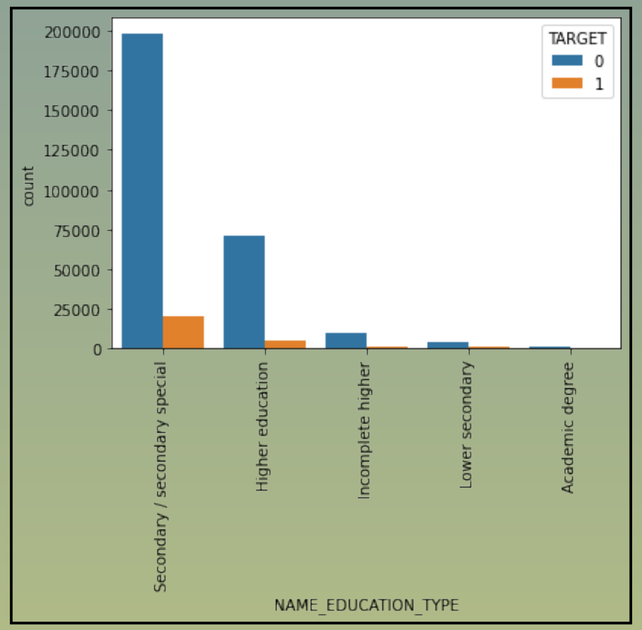


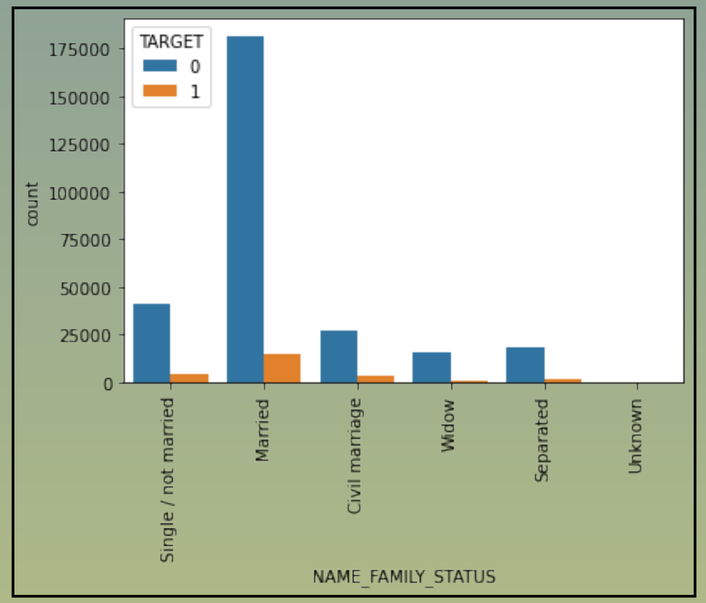


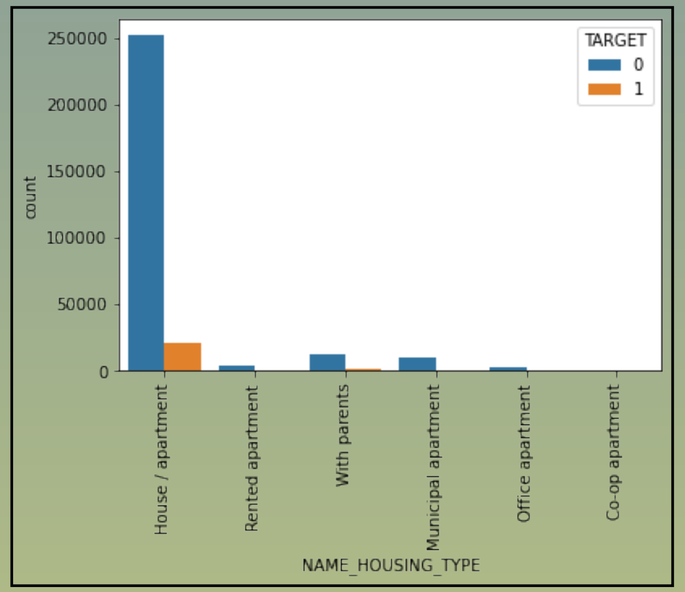


- Female will take more loans and likely to become more defaulters than males.
- People with cars will take loans but they find less difficulty with repayment than people with non-owners who are higher in range of taking loans and becoming defaulters.
- Unaccompained will become defaulters than people with having family and people with spouse partner.

EDUCATION TYPE, FAMILY STATUS AND HOUSING TYPE

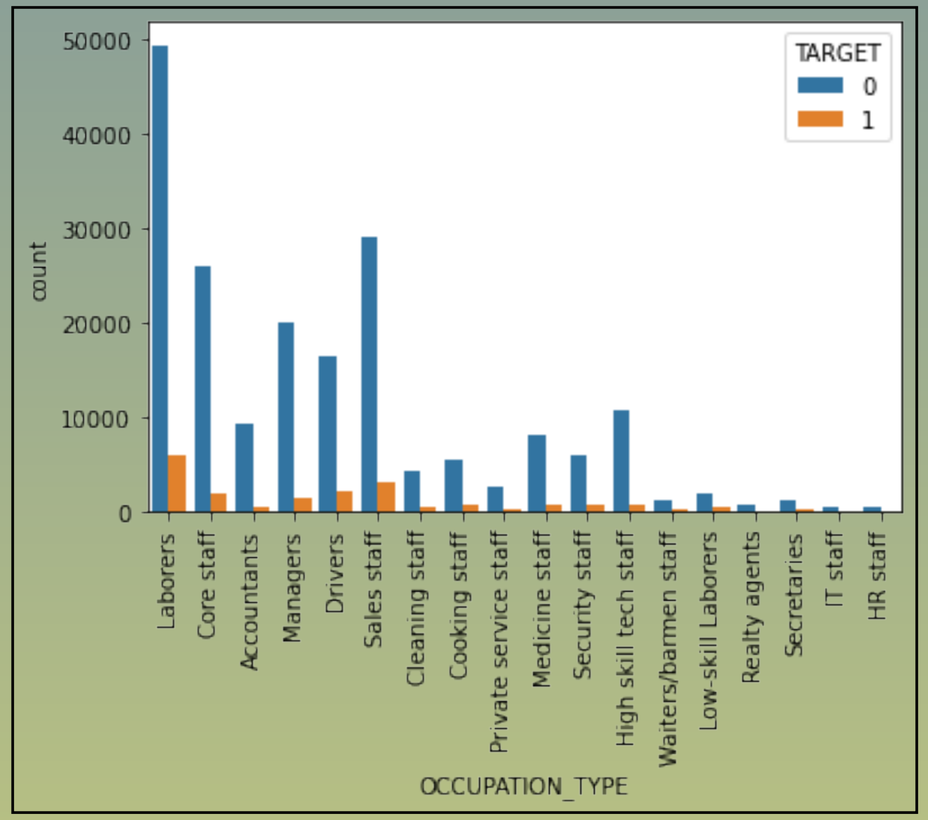


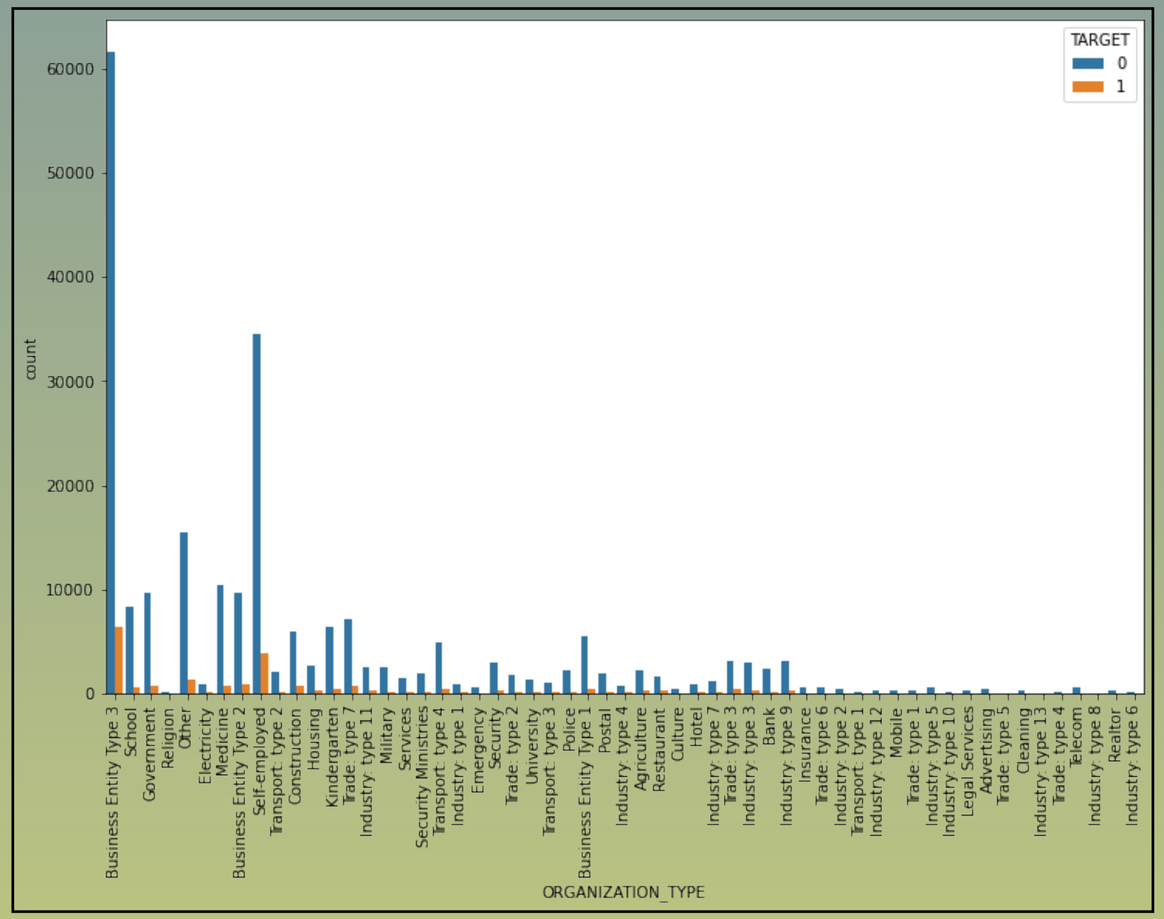




- Customers with having secondary education will be defaulters as compare to higher education and low secondary.
- Married people will take more loans and become defaulters as compare to single, widow,separated and civil marriage.

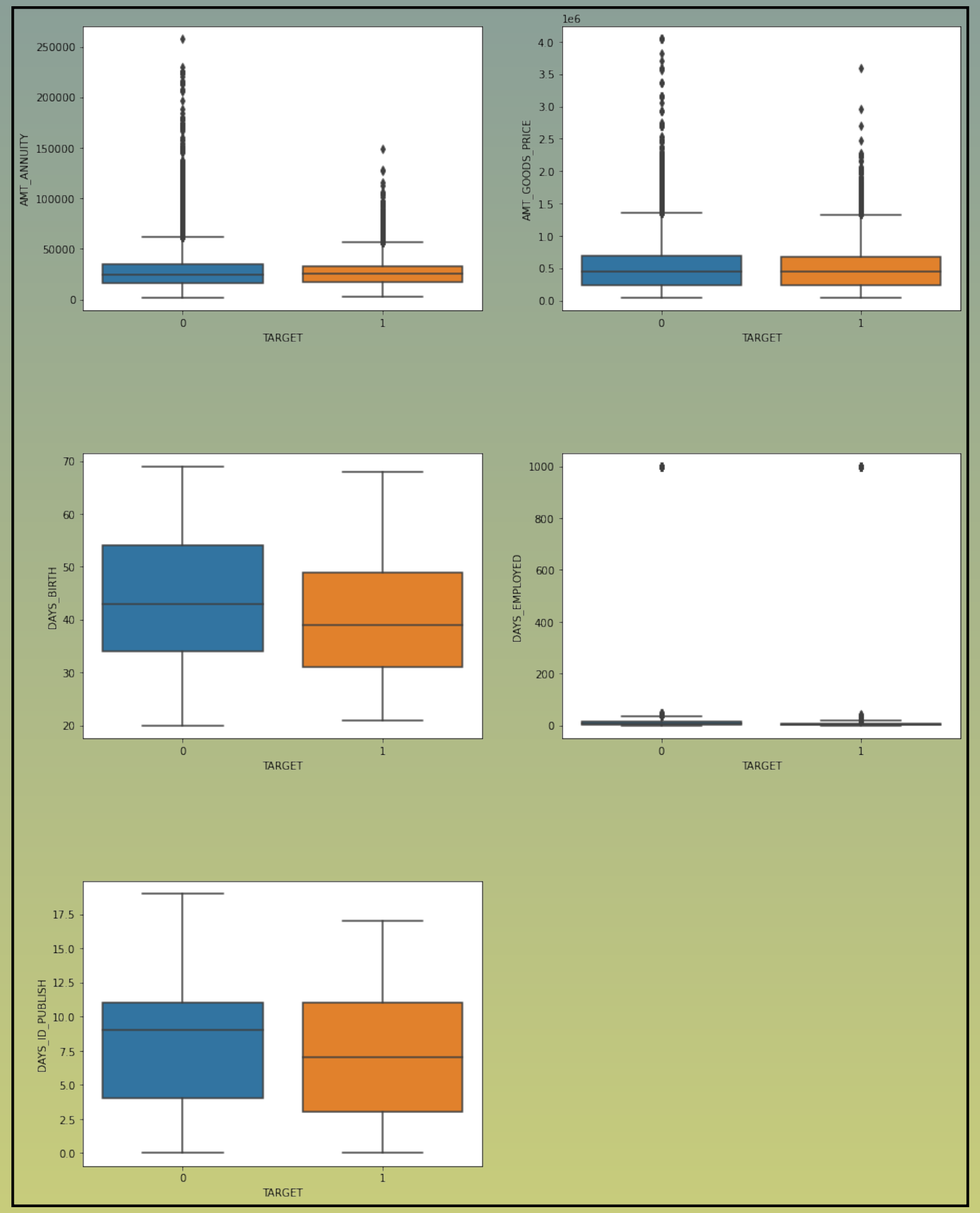
OCCUPATION TYPE AND ORGANISATION TYPE





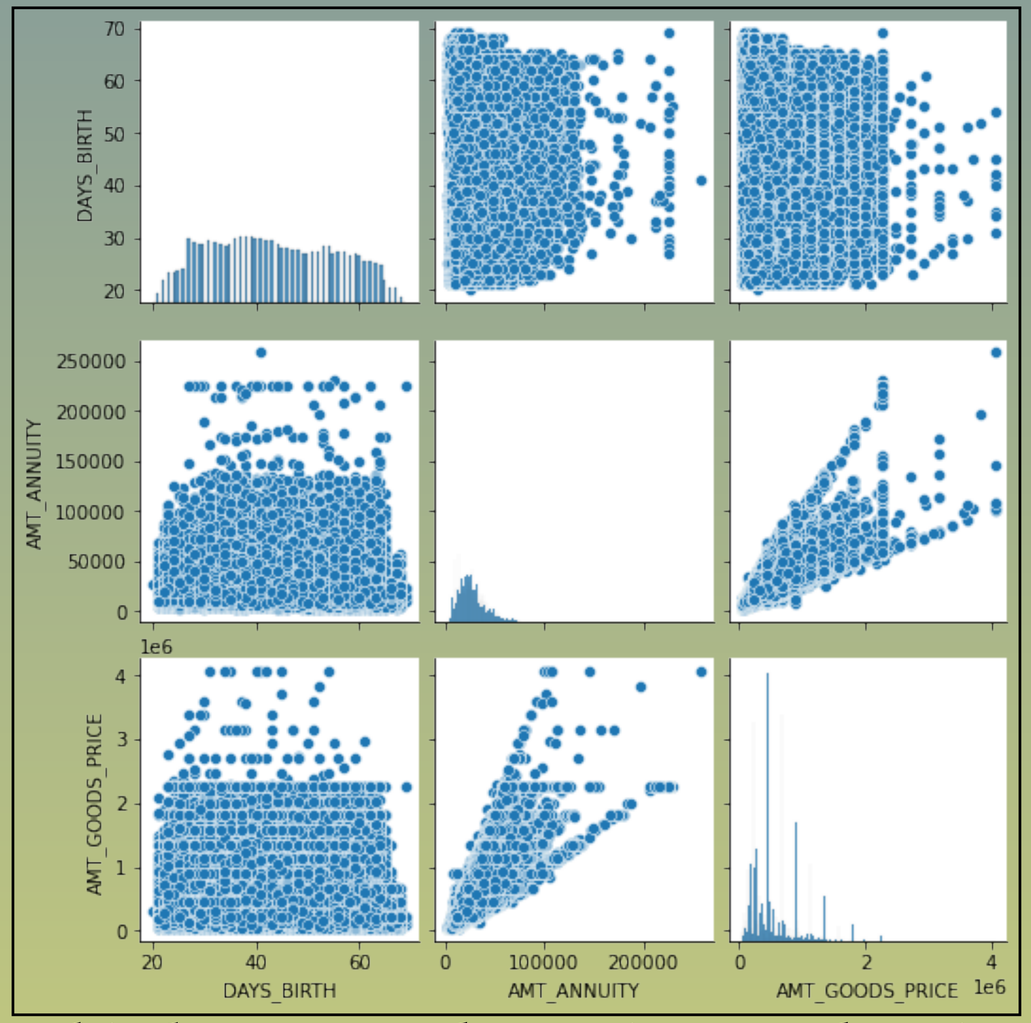
- Laborers are higher in number of the defaulters list.
- Business entity type-3 are more defaulters than other organisation people.

Univaraite analysis of continuous columns¶



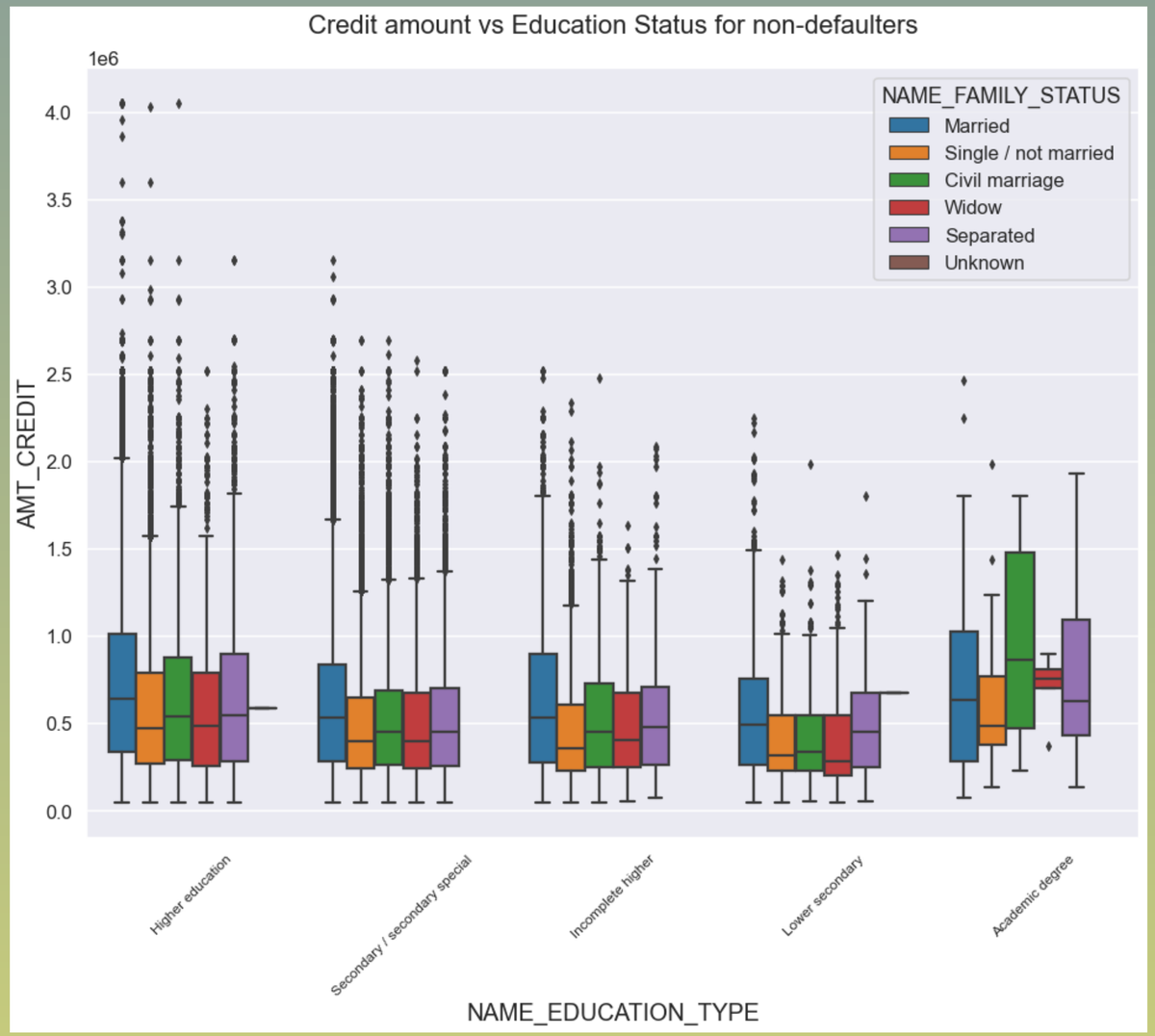
- Boxplot shows that are the outliers present in Amt_annunity plot and amt_goods_price plot.
- Customers with age in the range of 60-70 likely to become defaulters and higher in range of taking loans.
- Customers who have changed their Id will become defaulters.

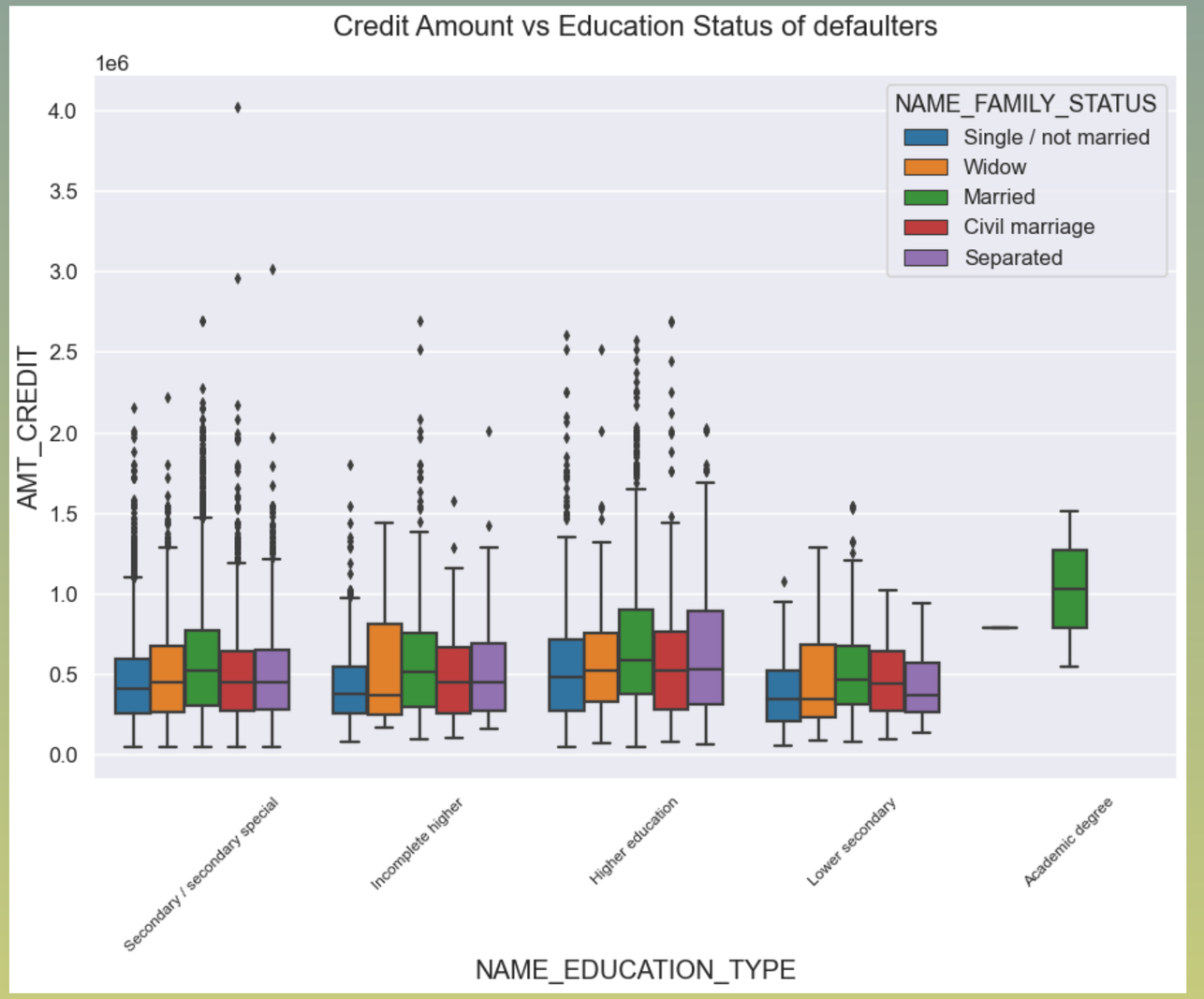
Bivariate analysis of numerical- numerical columns



- Pair plot shows that birth years and annuity, goods price has high correlation.

Bivariate analysis of Numerical and Categorical variables





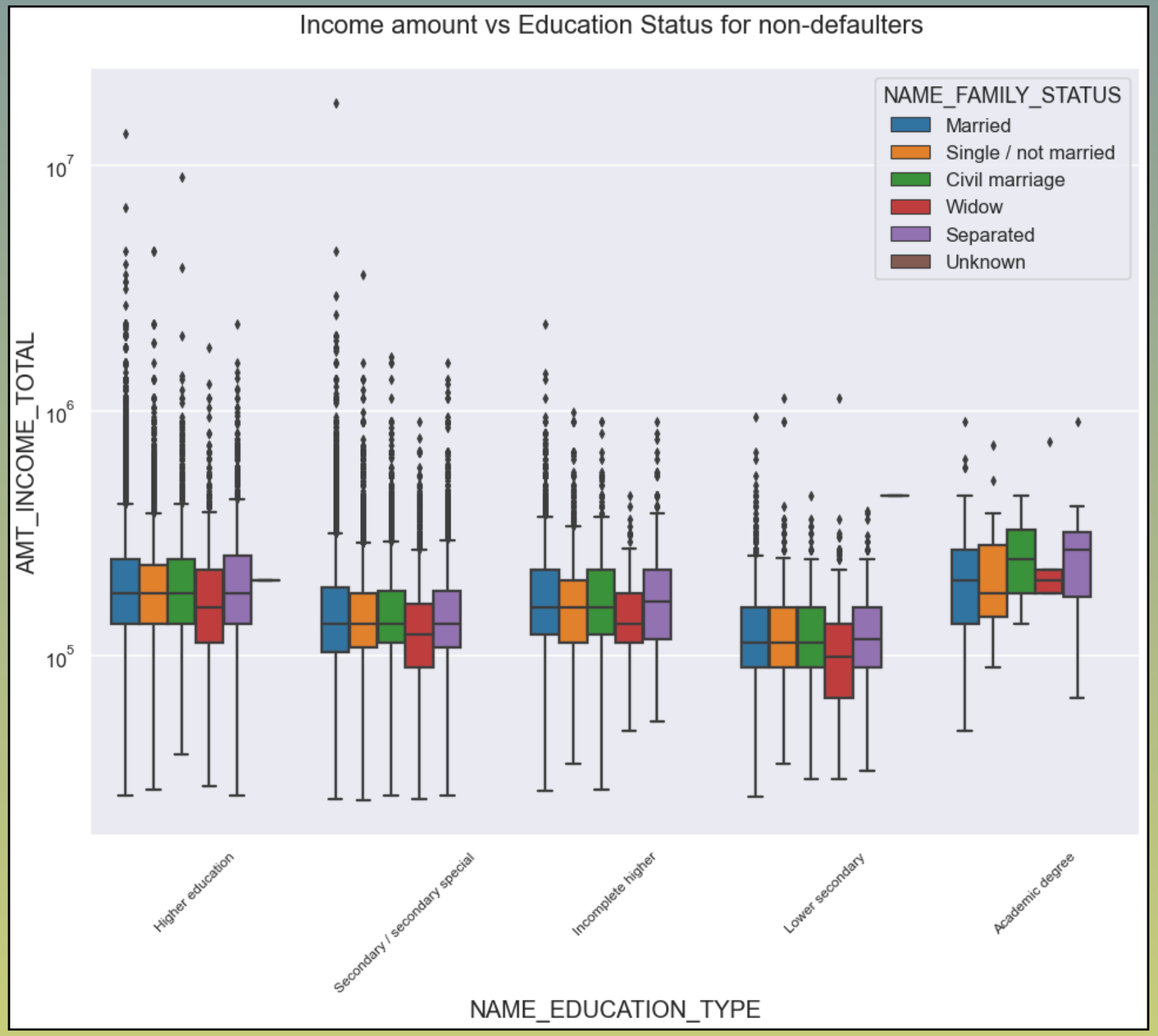
Insights from Credit amount vs education type non-defaulter plot:

- Clients with different Education types except Academic degrees have a large number of outliers.
- Most of the population i.e. clients' credit amounts lie below 25%.
- Clients with an Academic degree and who is a widow tend to take higher credit loan.
- Some of the clients with Higher Education, Incomplete Higher Education, Lower Secondary Education and Secondary/Secondary Special Education are more likely to take a high amount of credit loans.

Insights from Credit amount vs education type defaulter plot:

- The income amount for Married clients with an academic degree is much lesser as compared to others.
- (Defaulter) Clients have relatively less income as compared to Non-defaulters.

INCOME AMOUNT VS EDUCATION STATUS FOR NON-DEFAULTERS



Insights from non-defaulter plot:

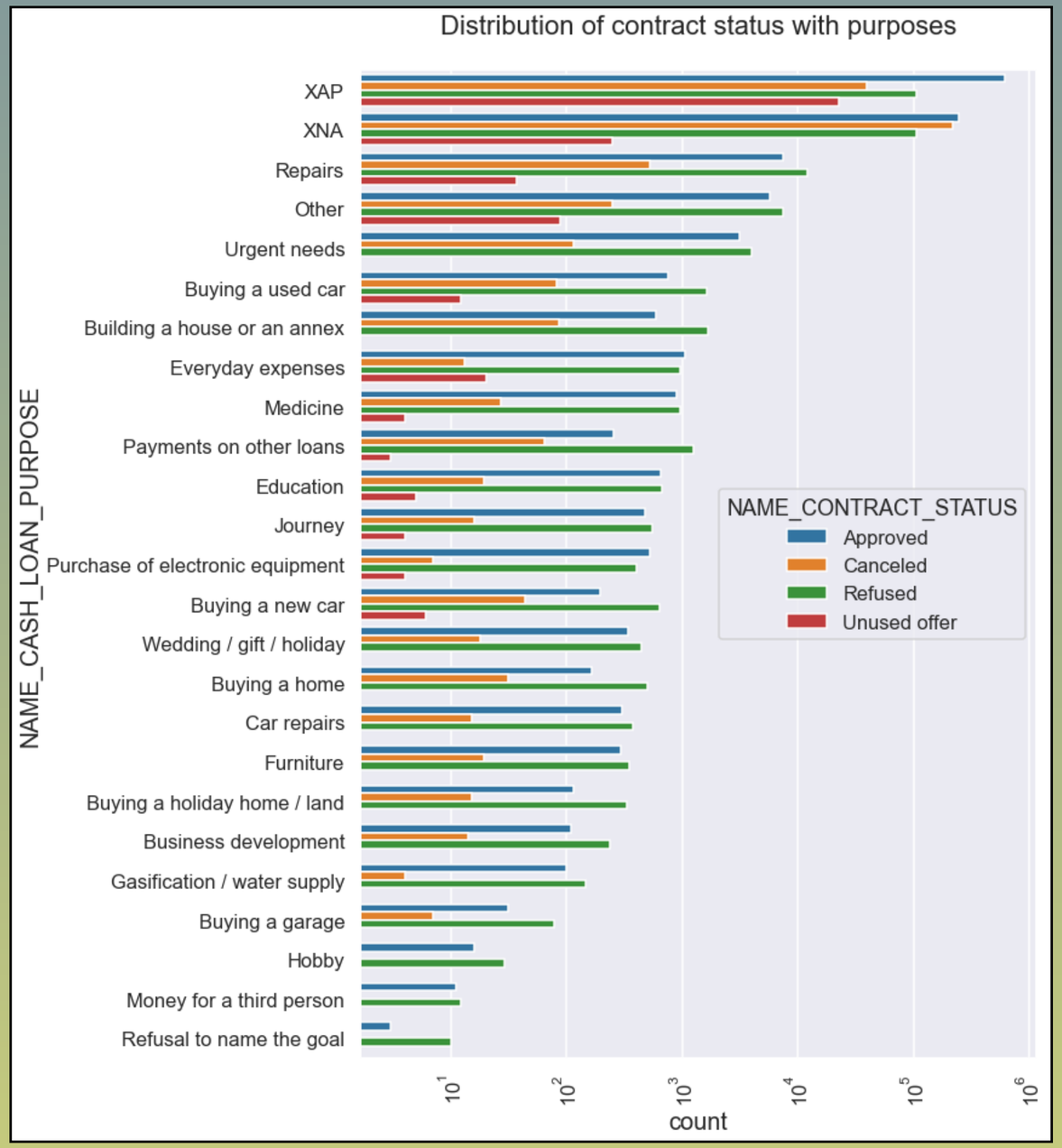
- Widow Client having Academic degree have very few outliers and doesn't have 1st and 3rd quartile. Also, Clients with all types of family statuses having academic degrees have very less outliers as compared to other types of education.
- Income of the clients with all types of family status having rest of the education type lie Below the 1st quartile i.e. 25%
- Clients having Higher Education, Incomplete Higher Education, Lower Secondary Education and Secondary/ Secondary Special have a higher number of outliers.
- From the above figure, we can say that some of the clients having Higher Education tend to have the highest income compared to others.
- Though some of the clients who haven't completed their Higher Education tend to have higher incomes.
- Some of the clients having Secondary/Secondary Special Education tend to have higher incomes.

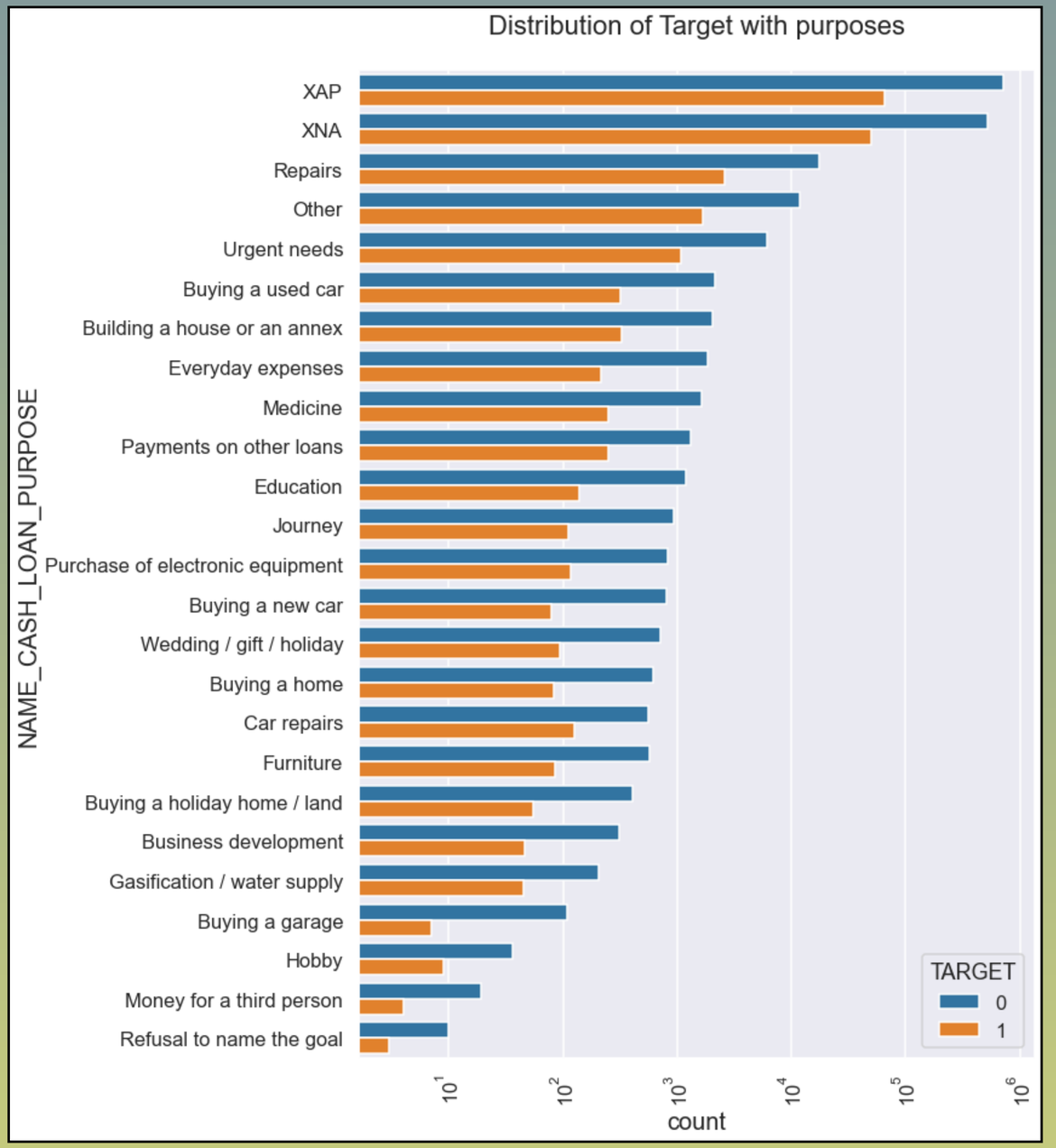
CORRELATION BETWEEN VARIABLES OF DF_0 AND DF-1

- Highest correlation is 1 and is between ‘OBS_60_CNT_SOCIAL_CIRCLE’ and ‘OBS_30_CNT_SOCIAL_CIRCLE’ for both the variables.
- After that we can see, AMT_GOODS_PRICE and AMT_CREDIT are also highly correlated for both the variables.
- AMT_ANNUITY and AMT_CREDIT are co-related with 0.75.

COMBINED DATA

- After that, I have combined Both the datas:- application_data and previous_data.
- Figure out the shape, info, description of the data.
- Checked for the missing values,
- Dropped the null values which are greater than 40%.
- Performed Univariate analysis on combined data.


Distribution of contract status with purposes
NAME_CASH_LOAN_PURPOSE
XAP
XNA
Repairs
Other
Urgent needs
Buying a used car
Building a house or an annex
Everyday expenses
Medicine
Payments on other loans
Education
Journey
Purchase of electronic equipment
Buying a new car
Wedding / gift / holiday
Buying a home
Car repairs
Furniture
Buying a holiday home / land
Business development
Gasification / water supply
Buying a garage
Hobby
Money for a third person
Refusal to name the goal
NAME_CONTRACT_STATUS
Approved
Canceled
Refused
Unused offer
10^1 10^2 10^3 10^4 10^5 10^6
count


Distribution of Target with purposes
NAME_CASH_LOAN_PURPOSE
XAP
XNA
Repairs
Other
Urgent needs
Buying a used car
Building a house or an annex
Everyday expenses
Medicine
Payments on other loans
Education
Journey
Purchase of electronic equipment
Buying a new car
Wedding / gift / holiday
Buying a home
Car repairs
Furniture
Buying a holiday home / land
Business development
Gasification / water supply
Buying a garage
Hobby
Money for a third person
Refusal to name the goal
TARGET
0
1
10^1 10^2 10^3 10^4 10^5 10^6
count

- We can see from the plot (“Distribution of Contract status with Purpose) that “Repairs” have refused for the loan.
- Because they find the difficult to repay the loan hence comes under defaulters.

CONCLUSION

- The proportion of defaulters is 8.07%.
- Low salary income people should not get loan.
- Bank lends more loan to females.
- Working people seems to become defaulters.
- Cash loans are not recommended , it should focus on revolving loans.
- People who don’t own cars are likely to be defaulters.
- Business entity type -3 are defaulters.
- Bank should focus on giving loans to office apartments, municipal appartment as will return the loans.
- Safer to give loan to married as compare to widow, separated or single.
- Highly skilled staff should get more loan as they likely to default less.
- Higher income, higher loan amount likely to default less.

THANK YOU.

THATS ALL FROM MY END. MORE ANALYSIS CAN BE DONE…