CREDIT EDA CASE STUDY

By Sudharma BG and Syed Ahisan

Problem Statement

- 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.

- The bank wants to understand the driving factors behind the loan default. i.e the variables which are strong in loan default or we can put it in a way that the segments that the bank wants to target to provide loans and make a profit in the process

1. Reading the Data

Application Data

- 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.

- The data has 307511 rows and 122 columns.

- The data has 51 columns with missing values greater than 25% which are removed.

Previous Data

- 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.

- The data has 1670214 rows and 37 columns .

- 4 columns have missing values at more than 50% and such columns are removed.

2. Data Quality checks

Data Quality checks on Application data

- Missing values in variable NAME_TYPE_SUITE replaced with the MODE value
- For categorical variable the value which should be imputed with maximum in frequency. The columns which are imputed are OBS_30_CNT_SOCIAL_CIRCLE , DEF_30_CNT_SOCIAL_CIRCLE , OBS_60_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE
- The columns:- AMT_REQ_CREDIT_BUREAU_HOUR,AMT_REQ_CREDIT_BUREAU_DAY,AMT_REQ_CREDIT_BUREAU_WEEK,AMT_REQ_CREDIT_BUREAU_MON,AMT_REQ_CREDIT_BUREAU_QRT,AMT_REQ_CREDIT_BUREAU_YEAR are filled with 0
- The following columns are converted to int data type as they can be only whole numbers :- DAYS_REGISTRATION,CNT_FAM_MEMBERS,OBS_30_CNT_SOCIAL_CIRCLE,DEF_30_CNT_SOCIAL_CIRCLE,DAYS_LAST_PHONE_CHANGE,AMT_REQ_CREDIT_BUREAU_HOUR

- Some columns are converted for Categorical Analysis (We have considered No as 0 and Yes as 1)

- The columns such as Gender(CODE_GENDER) and type of organization(ORGANIZATION_TYPE) where the value is mentioned as 'XNA' which means 'Not Available'.

- So we have find the number of rows and columns and implement suitable techniques on them to fill those missing values or to delete them.

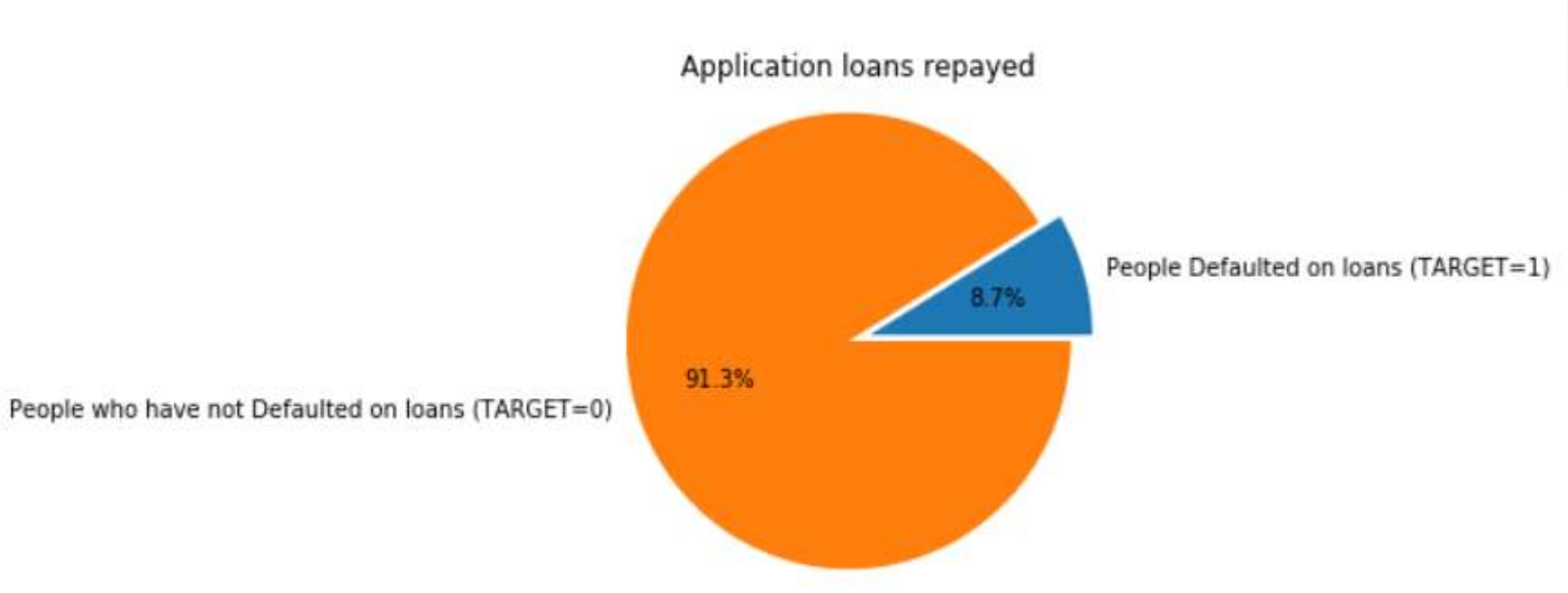
Data Quality checks on Previous data

- For the missing values of loan annuity(AMT_ANNUITY) and accompanying the clients (NAME_TYPE_SUITE) columns we have imputed the columns with mean and median respectively.

3.Data Analysis

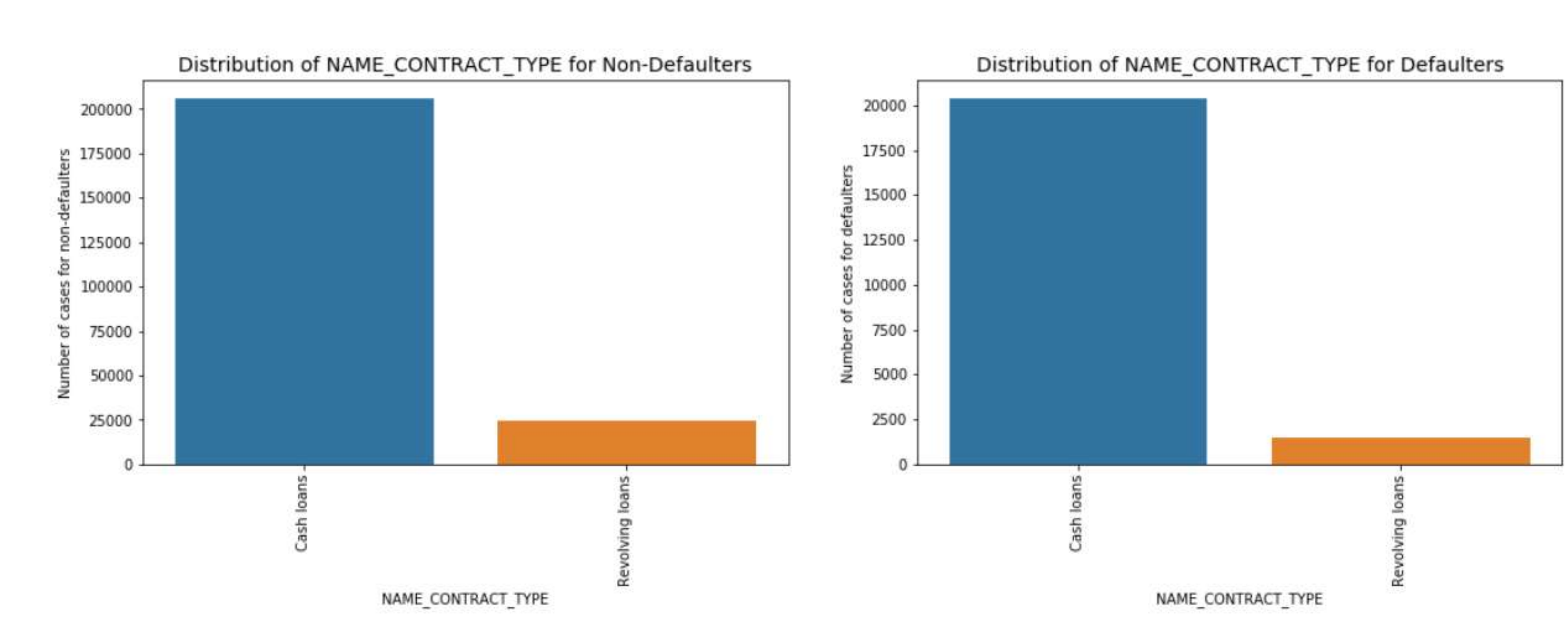
Data Analysis on Application data

Calculation of Loans Paid to Loans Not Paid



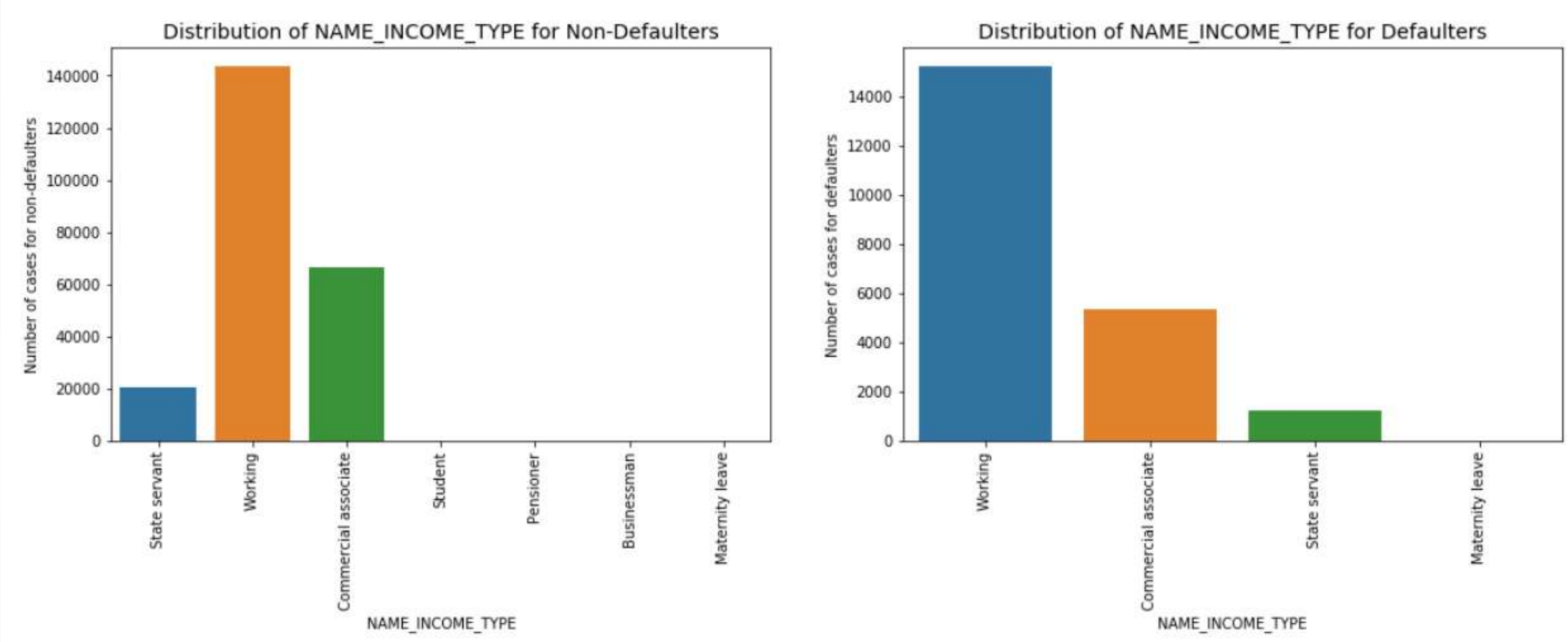
- Points to be concluded from the above graph
- Non-Defaulters are less than Defaulters.
- Hence there is imbalance in the given data.
- The imbalance ratio is 10.55

Distribution of NAME_CONTRACT_TYPE



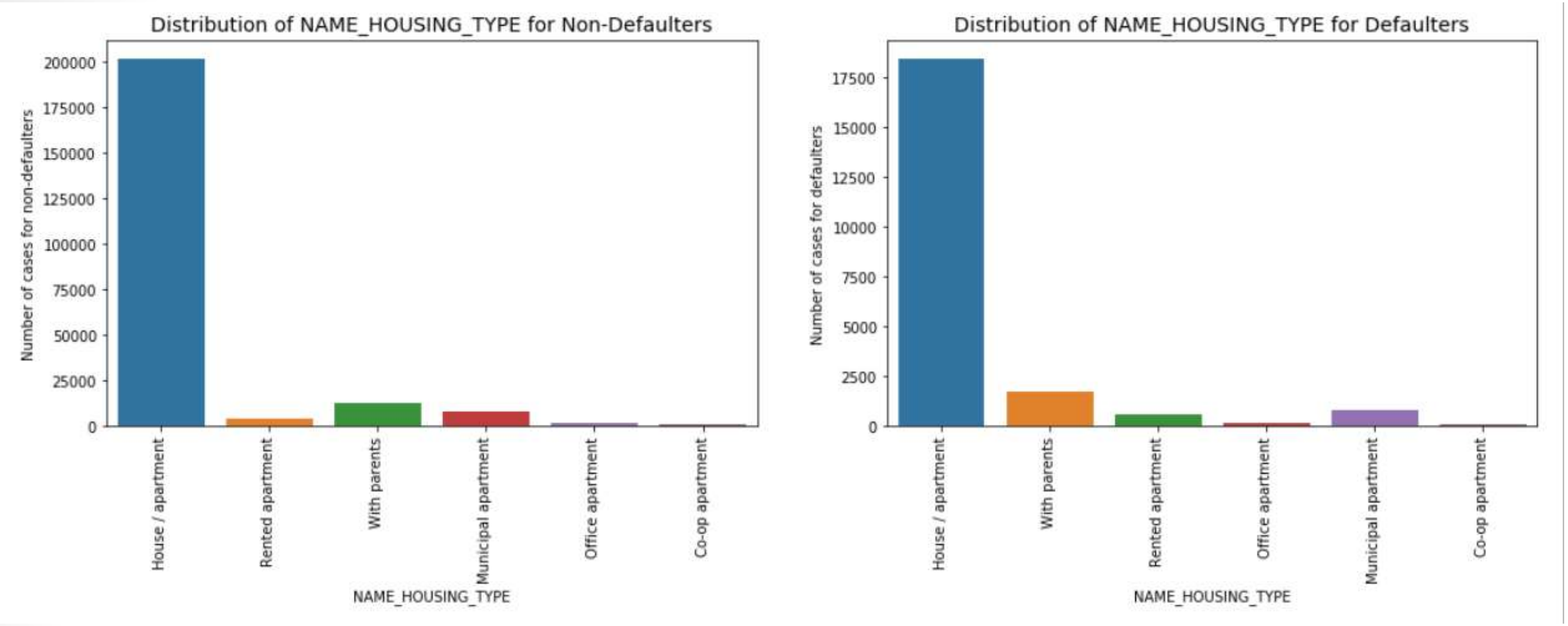
- Points to be concluded from the above graph.
- We can infer that the customers have opted considerably more for cash loans rather the revolving loans (Revolving loans is also called as open-end credit because the length of the loan isn't fixed and is flexible in nature).
- The Revolving loans have less percent of Defaulters when compared to Cash Loans .

Distribution of NAME_INCOME_TYPE



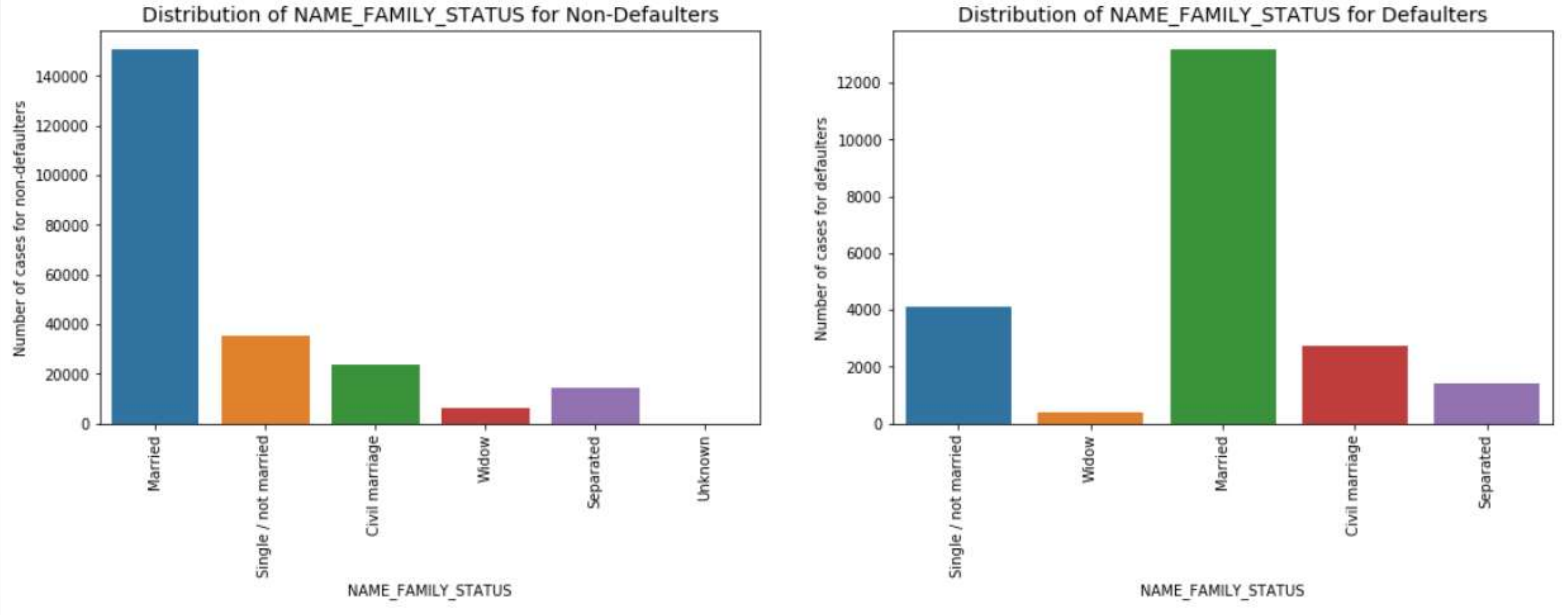
- Points to be concluded from the above graph
- By comparing the above graphs we can infer the state servant (Government employees) have repaid their loans on time and have not defaulted on their loans comparatively. This could be because of the fact that Government employees have a stable income and risk free career.
- Working class people have taken the most of the loans and have defaulted the most on loans and hence the banks must not focus much on them.

Distribution of NAME_HOUSING_TYPE



- Points to be concluded from the above graph
- People who own house/apartment have taken the most loans.
- This could be due to the fact that the they can provide security required for the loans

Distribution of NAME_FAMILY_STATUS



- Points to be concluded from the above graph
- Single/not Married people have defaulted the most on the loans comparatively.

Univariate Analysis

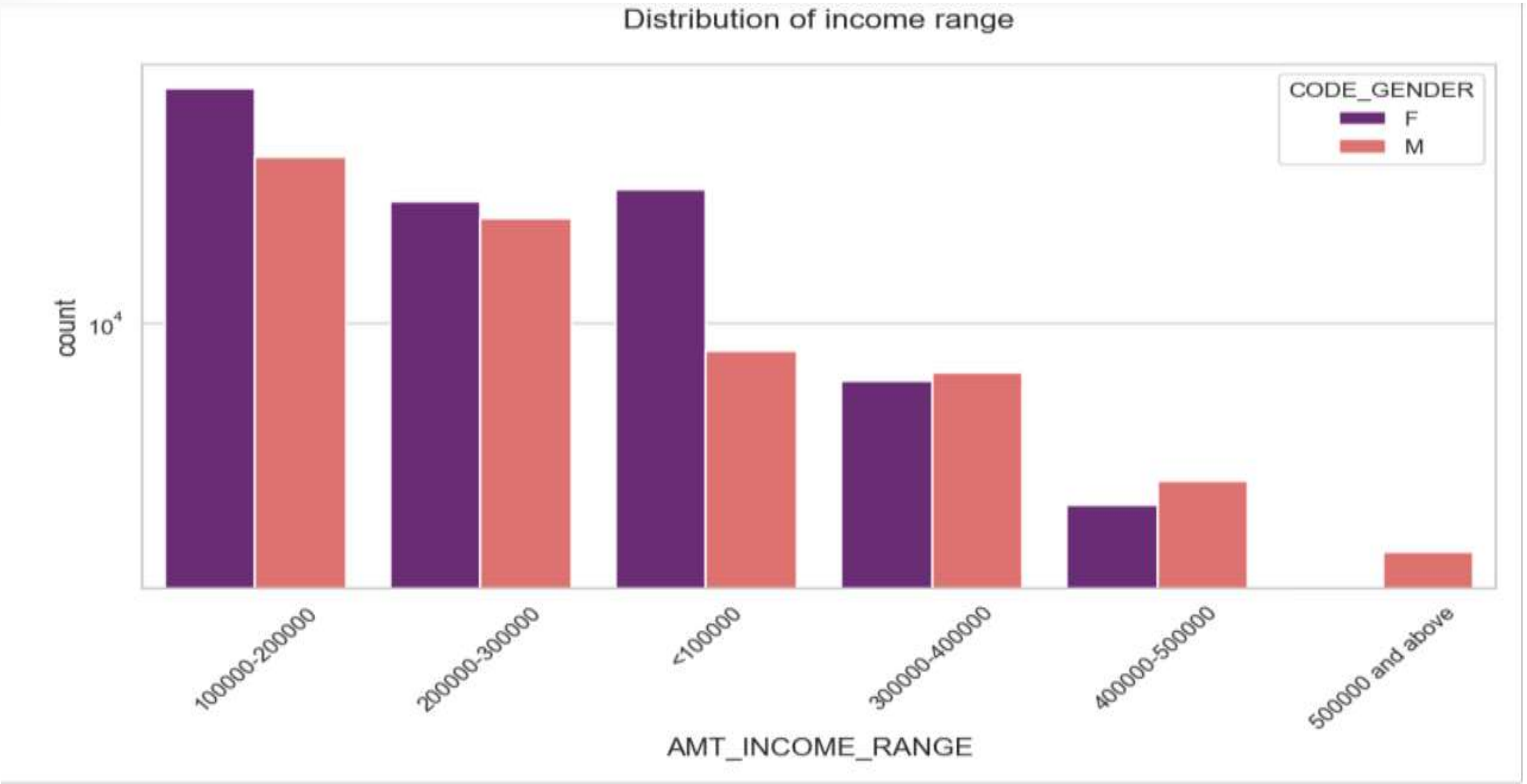
- We have binned the continues data into multiple segments for better analysis

- We have splitted the dataframe into 2 new data frames:-

  0.Non-Defaulters

  1.Defaulters

Distribution of AMT_INCOME_RANGE



- Points to be concluded from the above graph
- We can observe towards the higher income range the number females are more compared to males.
- Towards the lower income range we can observe that the number of females are less compared to males

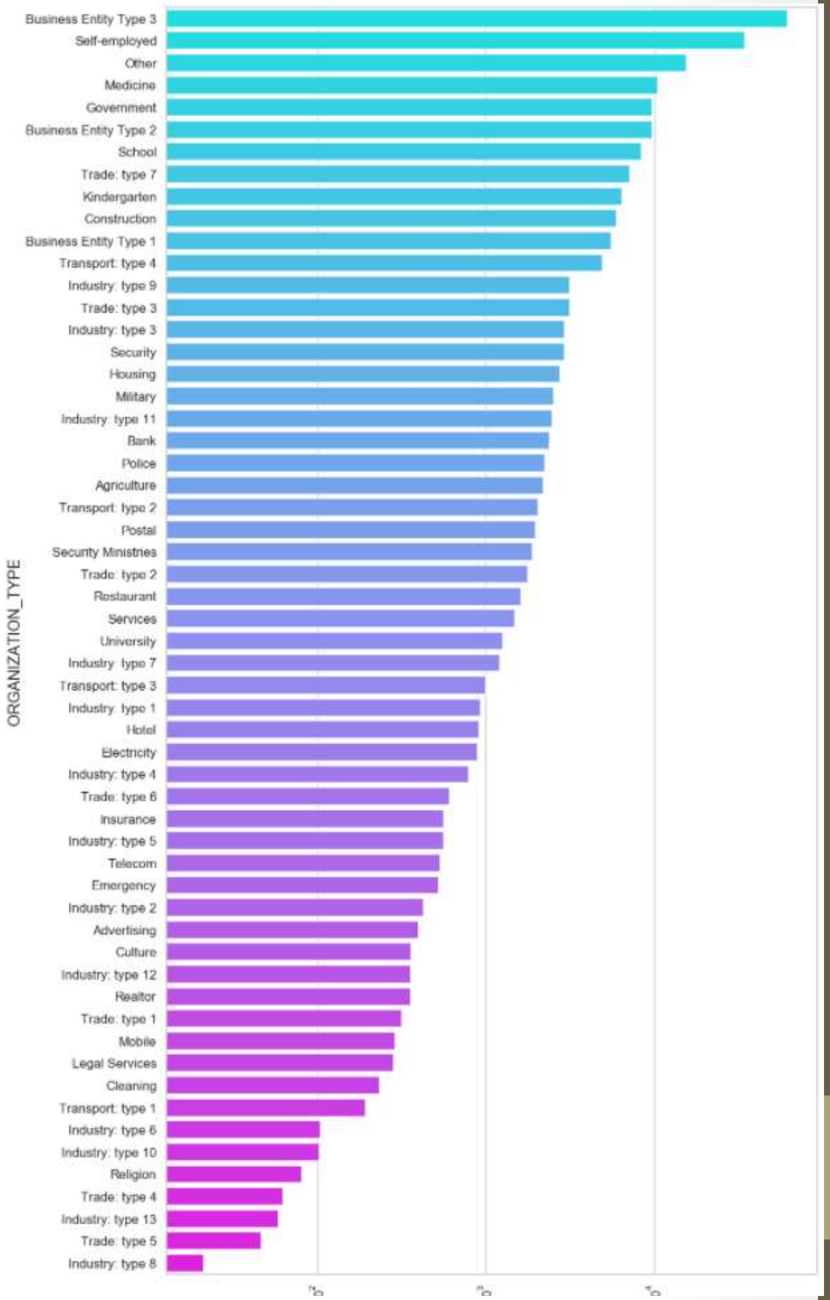
Distribution of ORGANIZATION_TYPE

- The graph is plotted for the target variable 0 (People who have not defaulted).

- Points to be concluded from the graph on the right.

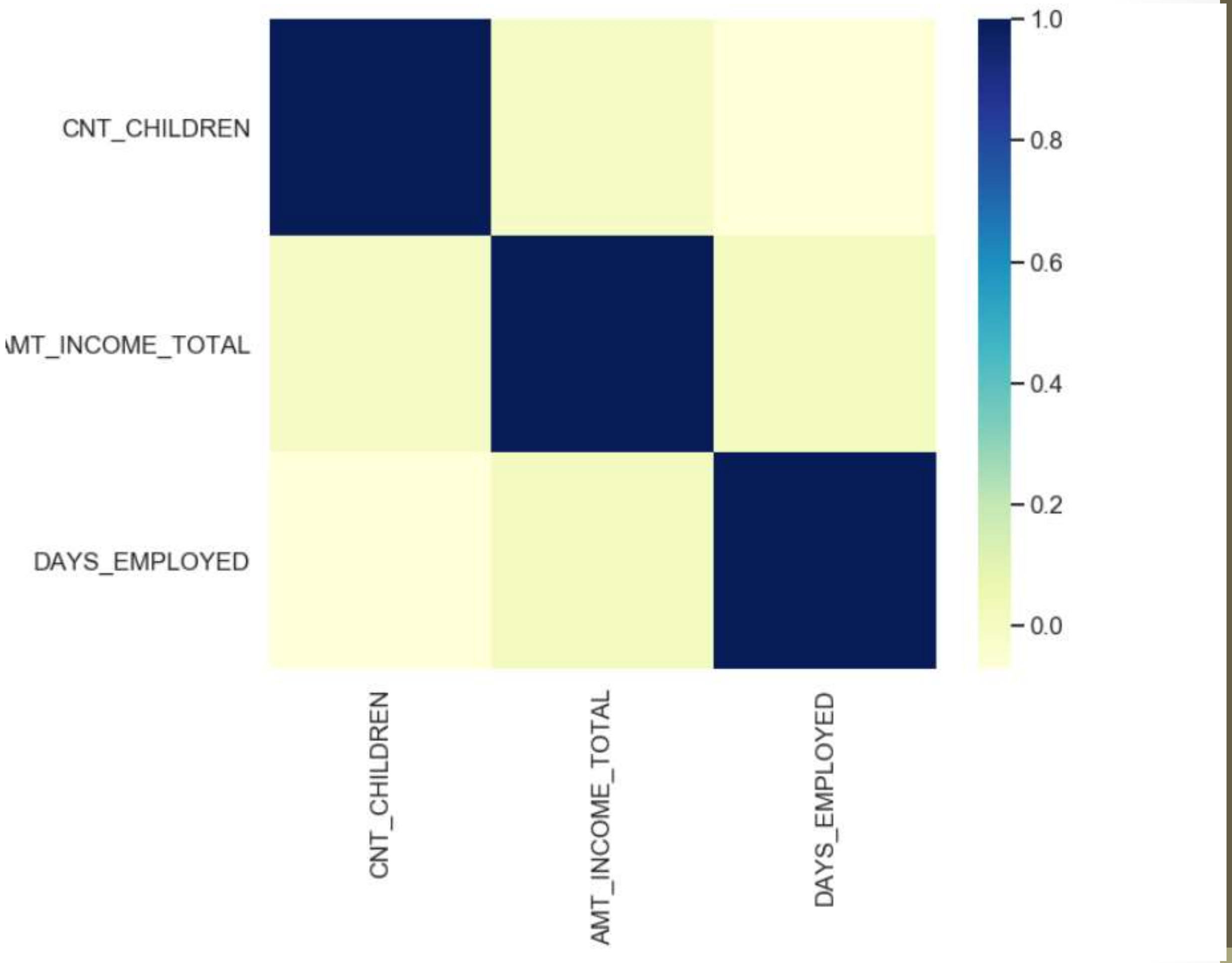
- Clients which have applied for credits are from most of the organization type Business entity Type 3 , Self employed , Other , Medicine and Government.

- Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4.



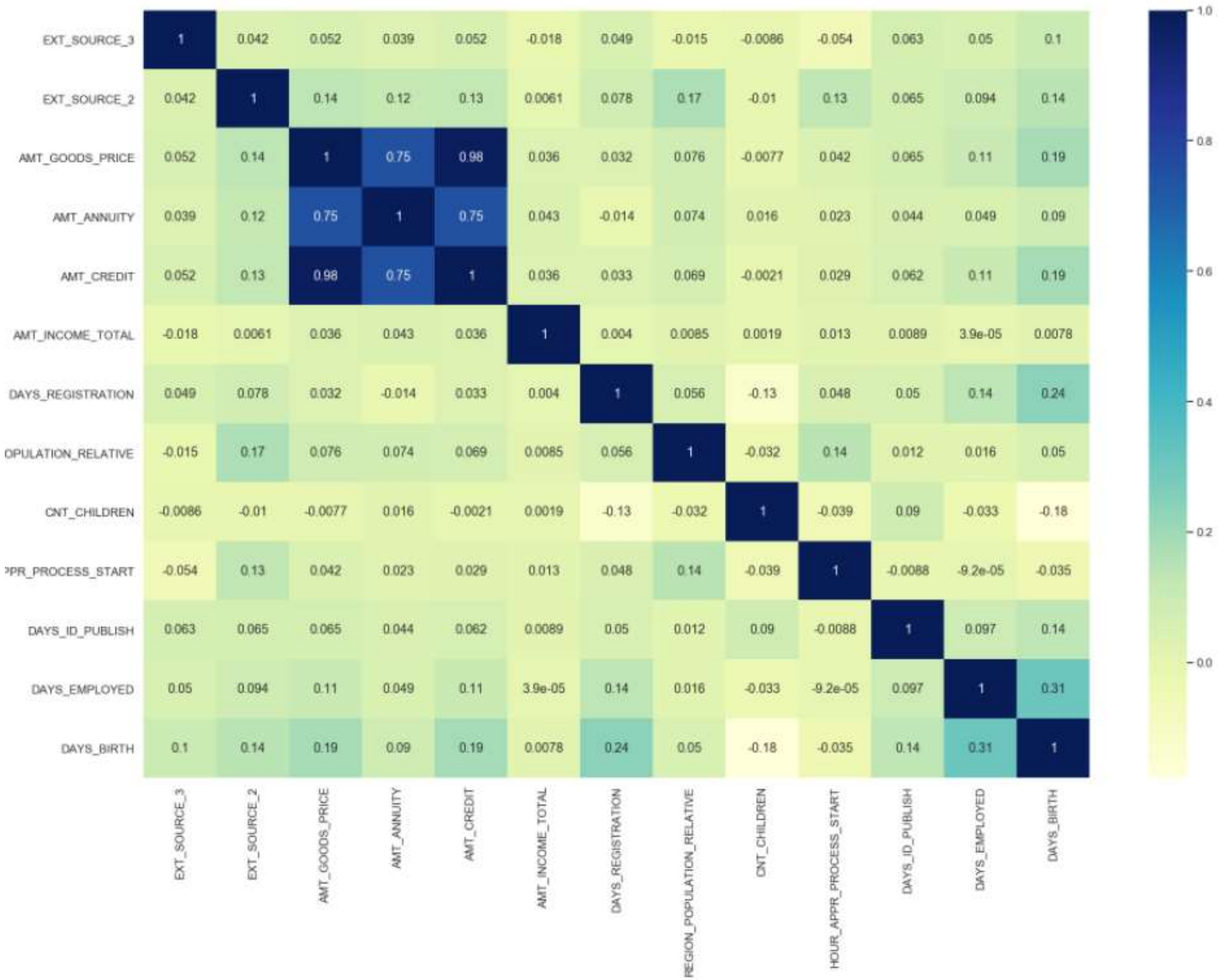
HEATMAP FOR MERGED DATAFRAME COLUMNS

- Among the columns no of children (CNT_CHILDREN) , no of days employees(DAYS_EMPLOYED) and total annual income (AMT_INCOME_TOTAL) there seems to be less co-relation.



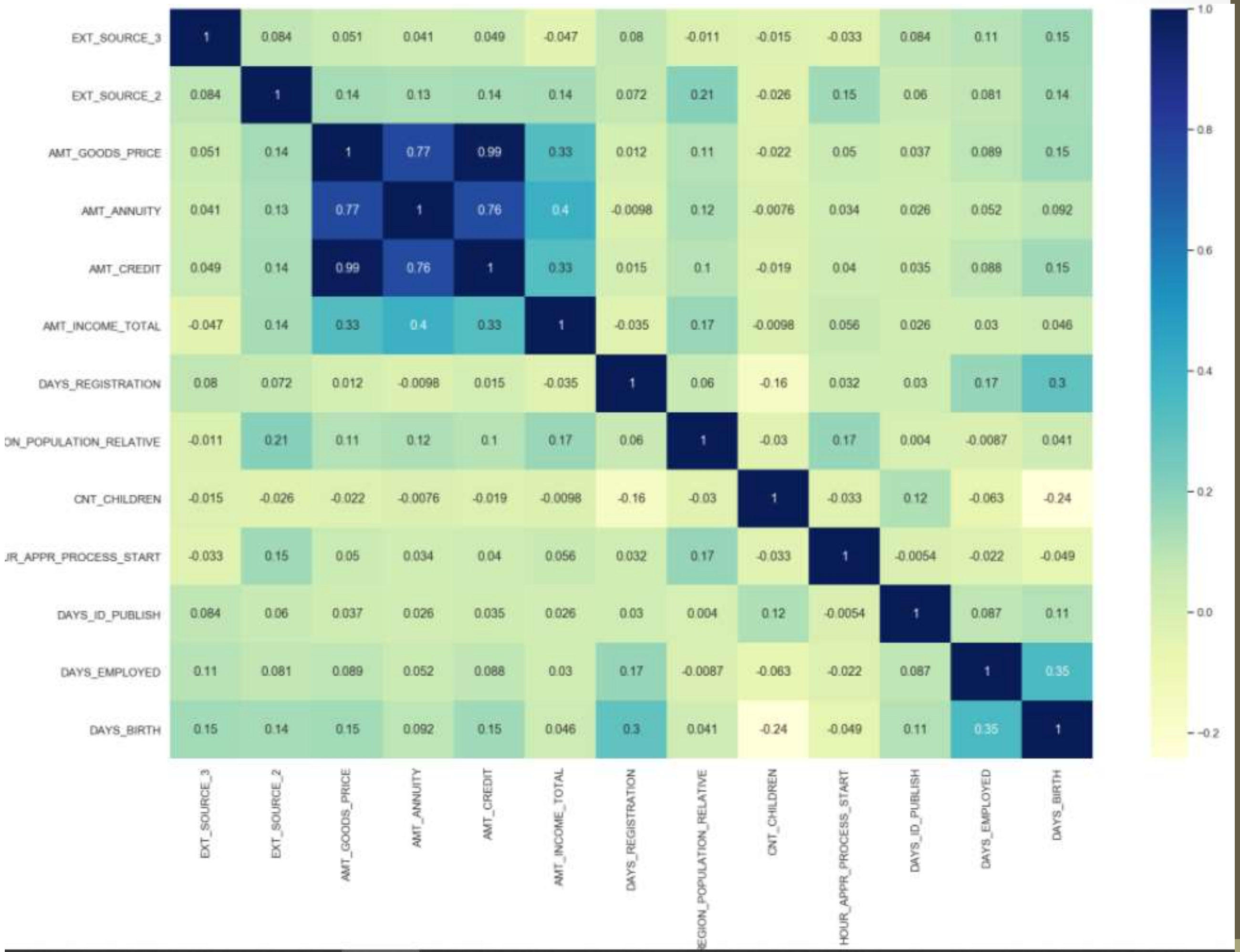
Heatmap for dataframe with target variable 1 (People who have defaulted)

- Highest co-relation:-
- AMT_CREDIT and AMT_GOODS_PRICE
- AMT_CREDIT and AMT_ANNUITY
- Some Columns even show negative correlation such as :-
- EXT_SOURC_3 and REGION_POPULATION_RELATIVE
- DAYS_REGISTRATION and CNT_CHILDREN



Heatmap for dataframe with target variable 0(People who have not defaulted)

Here there seems to be no great co-relation between the columns because the co-relation is >.50

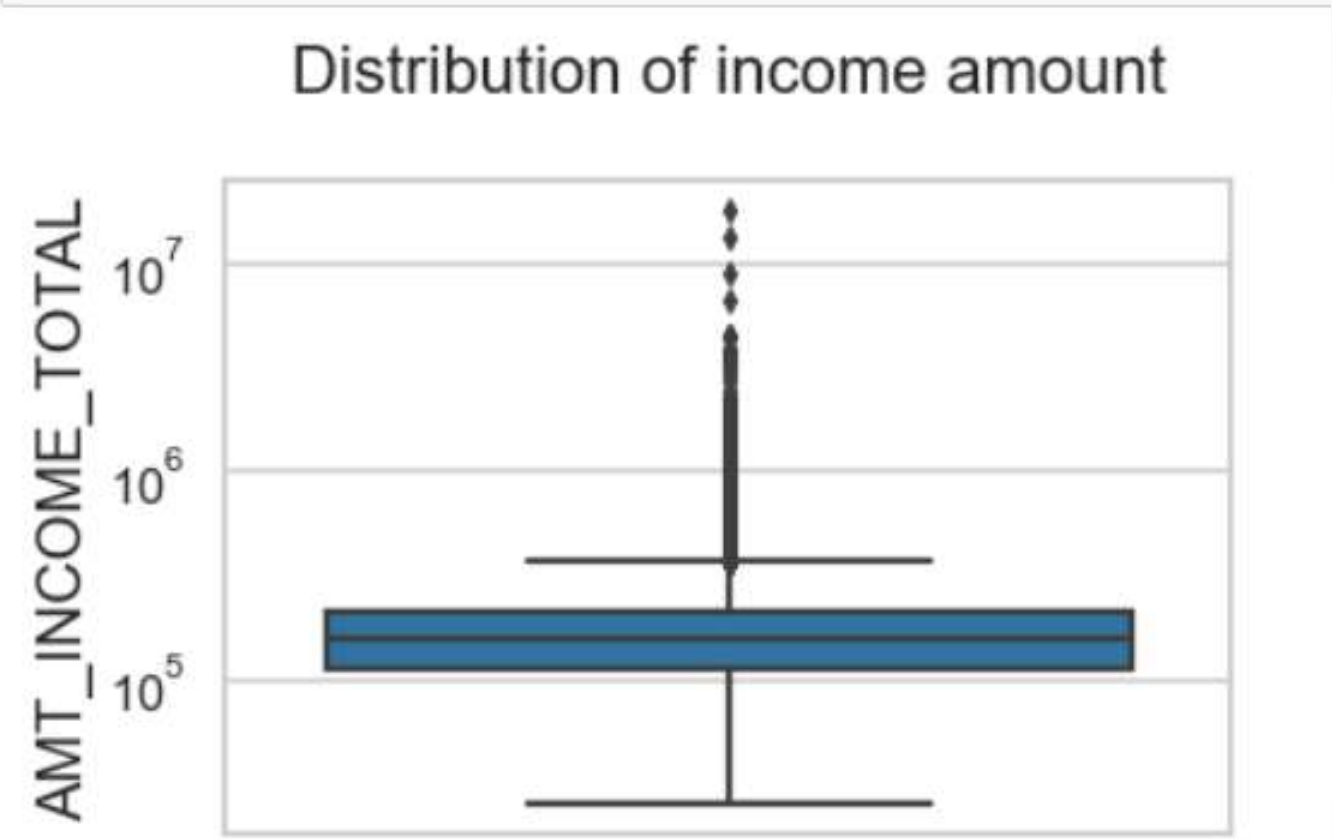


	EXT_SOURCE_3	EXT_SOURCE_2	AMT_GOODS_PRICE	AMT_ANNUITY	AMT_CREDIT	AMT_INCOME_TOTAL	DAYS_REGISTRATION	EGION_POPULATION_RELATIVE	CNT_CHILDREN	HOUR_APPR_PROCESS_START	DAYS_ID_PUBLISH	DAYS_EMPLOYED	DAYS_BIRTH
EXT_SOURCE_3	1	0.084	0.051	0.041	0.049	-0.047	0.08	-0.011	-0.015	-0.033	0.084	0.11	0.15
EXT_SOURCE_2	0.084	1	0.14	0.13	0.14	0.14	0.072	0.21	-0.026	0.15	0.06	0.081	0.14
AMT_GOODS_PRICE	0.051	0.14	1	0.77	0.99	0.33	0.012	0.11	-0.022	0.05	0.037	0.089	0.15
AMT_ANNUITY	0.041	0.13	0.77	1	0.76	0.4	-0.0098	0.12	-0.0076	0.034	0.026	0.052	0.092
AMT_CREDIT	0.049	0.14	0.99	0.76	1	0.33	0.015	0.1	-0.019	0.04	0.035	0.088	0.15
AMT_INCOME_TOTAL	-0.047	0.14	0.33	0.4	0.33	1	-0.035	0.17	-0.0098	0.056	0.026	0.03	0.046
DAYS_REGISTRATION	0.08	0.072	0.012	-0.0098	0.015	-0.035	1	0.06	-0.16	0.032	0.03	0.17	0.3
ON_POPULATION_RELATIVE	-0.011	0.21	0.11	0.12	0.1	0.17	0.06	1	-0.03	0.17	0.004	-0.0087	0.041
CNT_CHILDREN	-0.015	-0.026	-0.022	-0.0076	-0.019	-0.0098	-0.16	-0.03	1	-0.033	0.12	-0.063	-0.24
JR_APPR_PROCESS_START	-0.033	0.15	0.05	0.034	0.04	0.056	0.032	0.17	-0.033	1	-0.0054	-0.022	-0.049
DAYS_ID_PUBLISH	0.084	0.06	0.037	0.026	0.035	0.026	0.03	0.004	0.12	-0.0054	1	0.087	0.11
DAYS_EMPLOYED	0.11	0.081	0.089	0.052	0.088	0.03	0.17	-0.0087	-0.063	-0.022	0.087	1	0.35
DAYS_BIRTH	0.15	0.14	0.15	0.092	0.15	0.046	0.3	0.041	-0.24	-0.049	0.11	0.35	1

Box plotting for univariate variables analysis in logarithmic scale

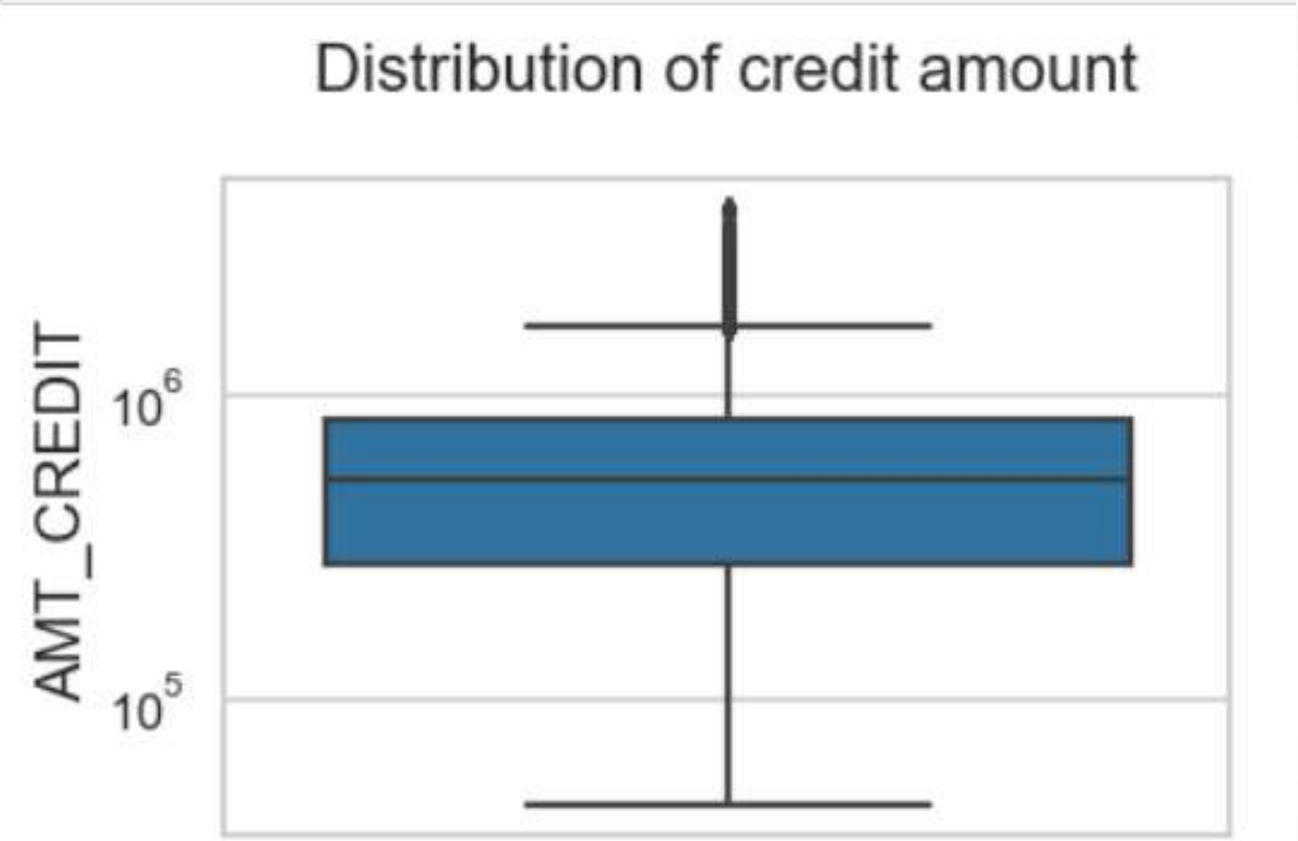
Distribution of income amount

The third quartiles is very slim for income amount.
Distribution of credit amount



Distribution of Credit amount

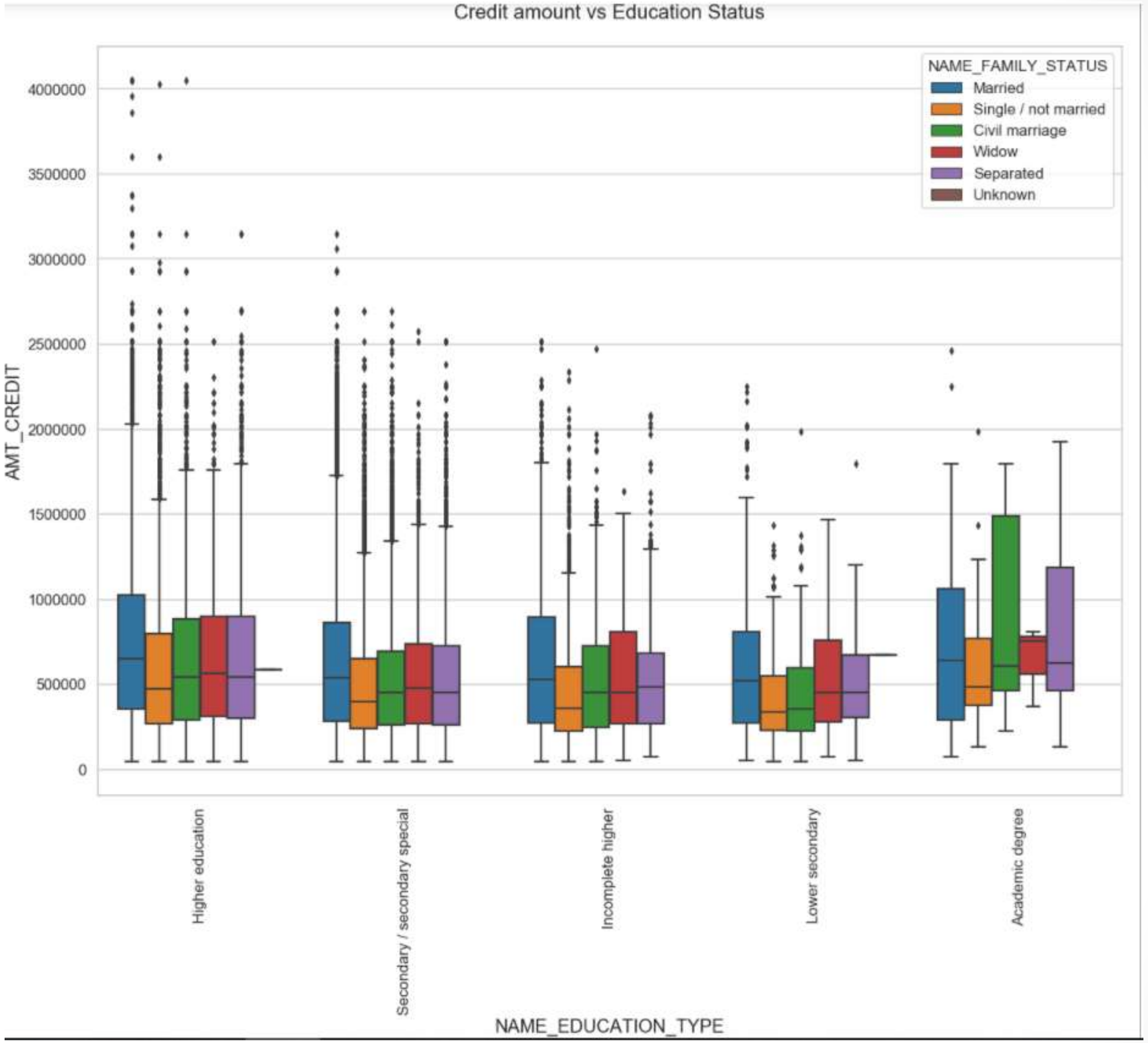
The first quartile is bigger than third quartile for credit amount which means most of the credits of clients are present in the first quartile.



Bivariate analysis

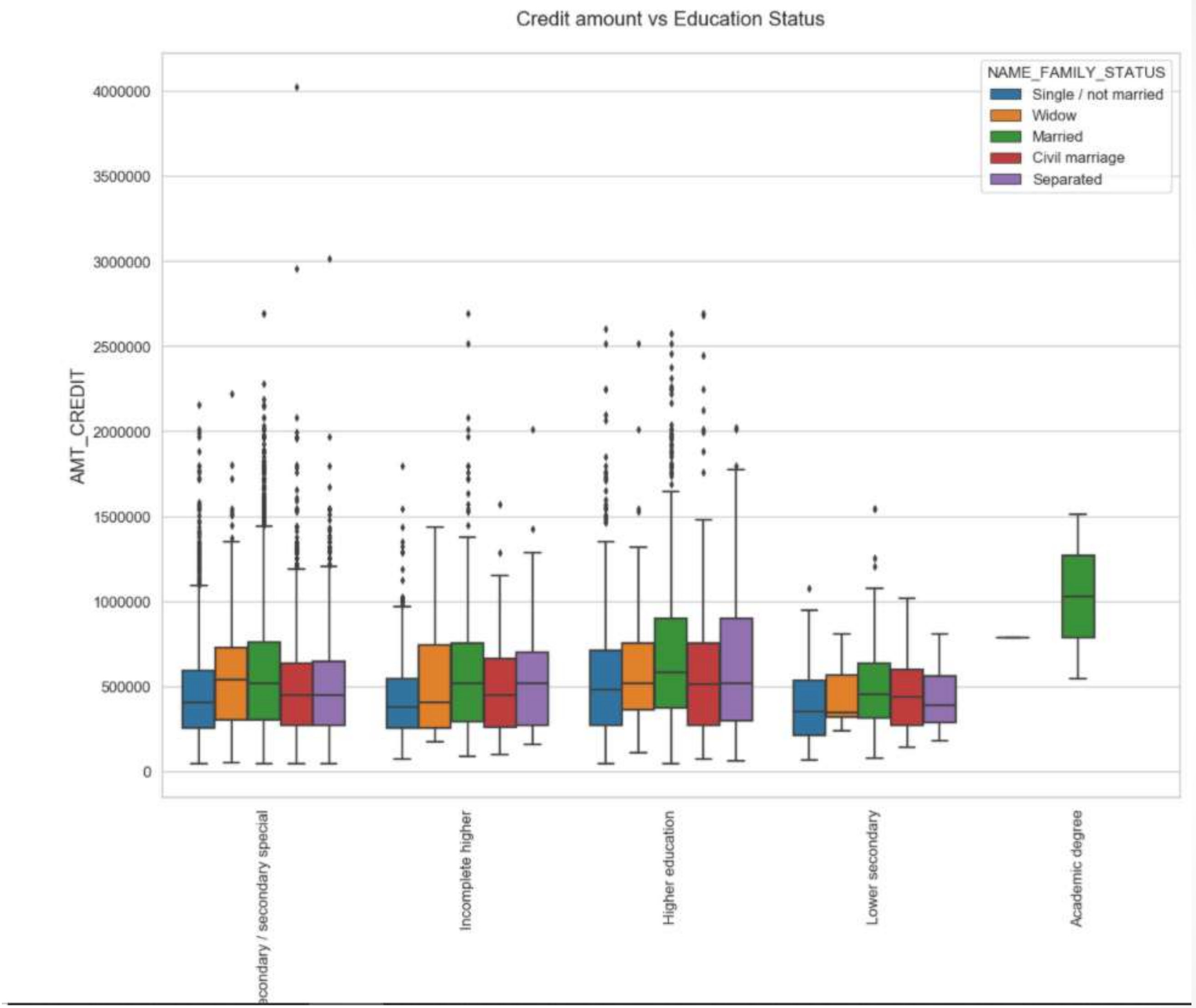
For Target 0 [Non Defaulters](People who have repaid loan on time)

- Observing the above box plot we can infer that the Family status with civil marriage, marriage and separated of Academic degree education are having higher number of credits than others.
- Higher education with family status of marriage, single and civil marriage are having more outliers.
- Civil marriage with Academic degree is having most of the credits in the third quartile.



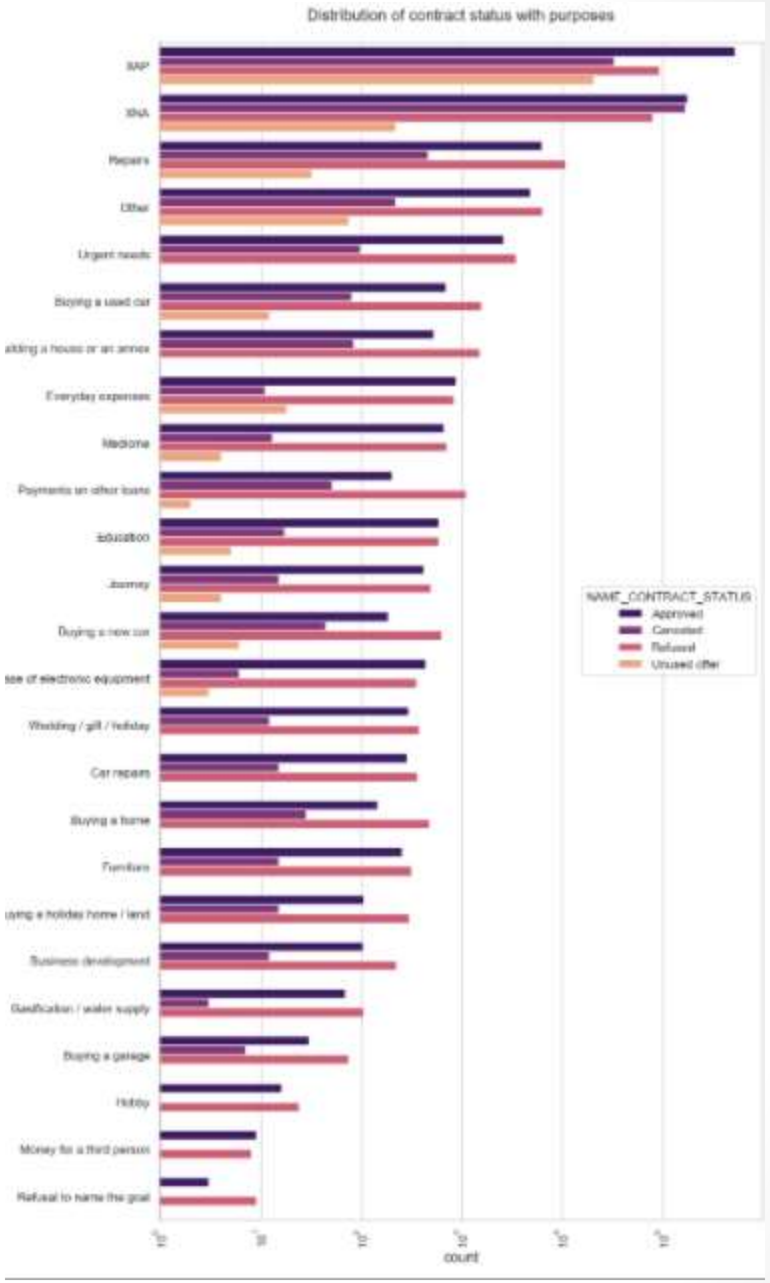
Target 1 [Defaulters (People who have not repaid the loan on time)]

- By observing the box plot we can say that Family status with civil marriage, marriage and separated of Academic degree education are having least number of defaulters than others.Hence banks can target this segment of people
- Most of the outliers are from Education type of Higher education and Secondary.



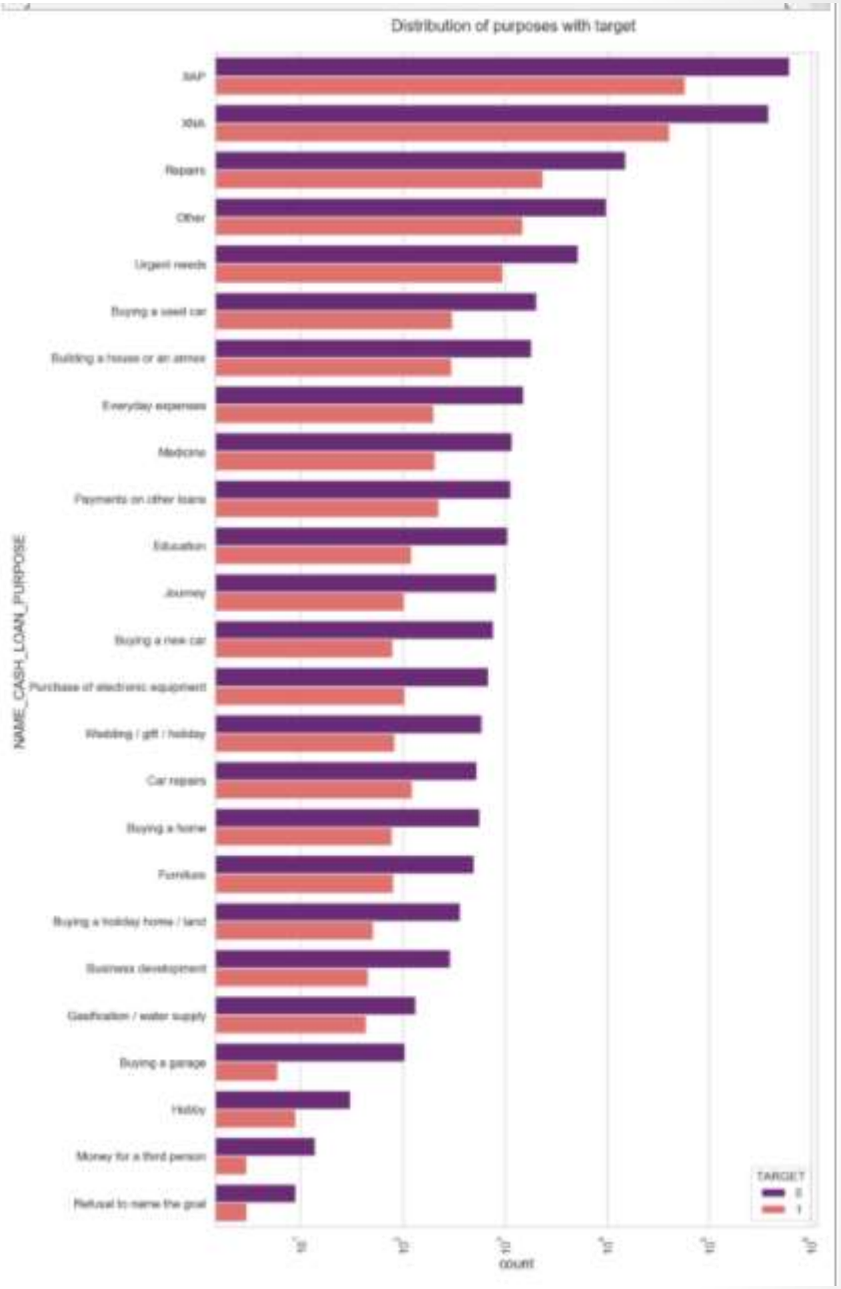
Distribution of contract status with purposes

- Few points can be concluded from the graph.
- Most rejection of loans came from purpose 'repairs'.
- For education purposes we have equal number of approves and rejection
- Paying other loans and buying a new car is having significant higher rejection than approves.



Distribution of purposes with target

- Few points can be concluded from the graph.
- Loan purposes with 'Repairs' are facing more difficulties in payment on time.
- There are few places where loan payment is significant higher than facing difficulties. They are '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 for minimal payment difficulties.



CONCLUSIONS

1. Most of the Loans have been given out is of the type Cash Loans but Revolving type of Loans seems to have comparatively fewer defaulters due to the fact that they are flexible in nature

2. Most of the loans are provided to the segments of people

- who own a house/apartment,
- working class people
- organization type business entity type 3

3.The segments of people who are defaulters(facing difficulties in repaying of loan amount on time) belong to the category

- income type working
- loan purpose repair
- family status not married/single

4.The segments of people who are non defaulters(facing no difficulties in repaying of loan amount on time) belong to the category

- loan purpose for Buying a garage, Business developemt, Buying land,Buying a new car and Education
- education type with people having academic degree
- income type state servant

THANK YOU