CREDIT EDA CASE STUDY

ROHAN GHOGARE

JYOTHISH J

PROBLEM STATEMENT

Introduction:

- This study aims at applying EDA principals to understand the risk associated in any financial lending institutions

- This EDA study needs to identify the driving factors which can help to find out potential defaulters and also whether to offer a loan to clients. This way financial institution can cut back financial losses and make some profitable informed business decisions

Data Provided :

- Spread sheet containing all the information of the client at the time of application. The data talks about attributes which has impact on clients payment difficulties.

- Spread sheet contains information about the client's previous loan data. This data shows whether previous application had been Approved, Cancelled, Refused or Unused offer.

SOLUTION APPROACH

Approach for case study:

- Load/ Read all data provided and identify the data structure, information about data like data types, missing values.

- Data Cleaning Techniques :
  - ❖ Investigate the missing values and impute that with suitable method
  - ❖ Analysing if there is any outliers in data
  - ❖ Check data imbalance

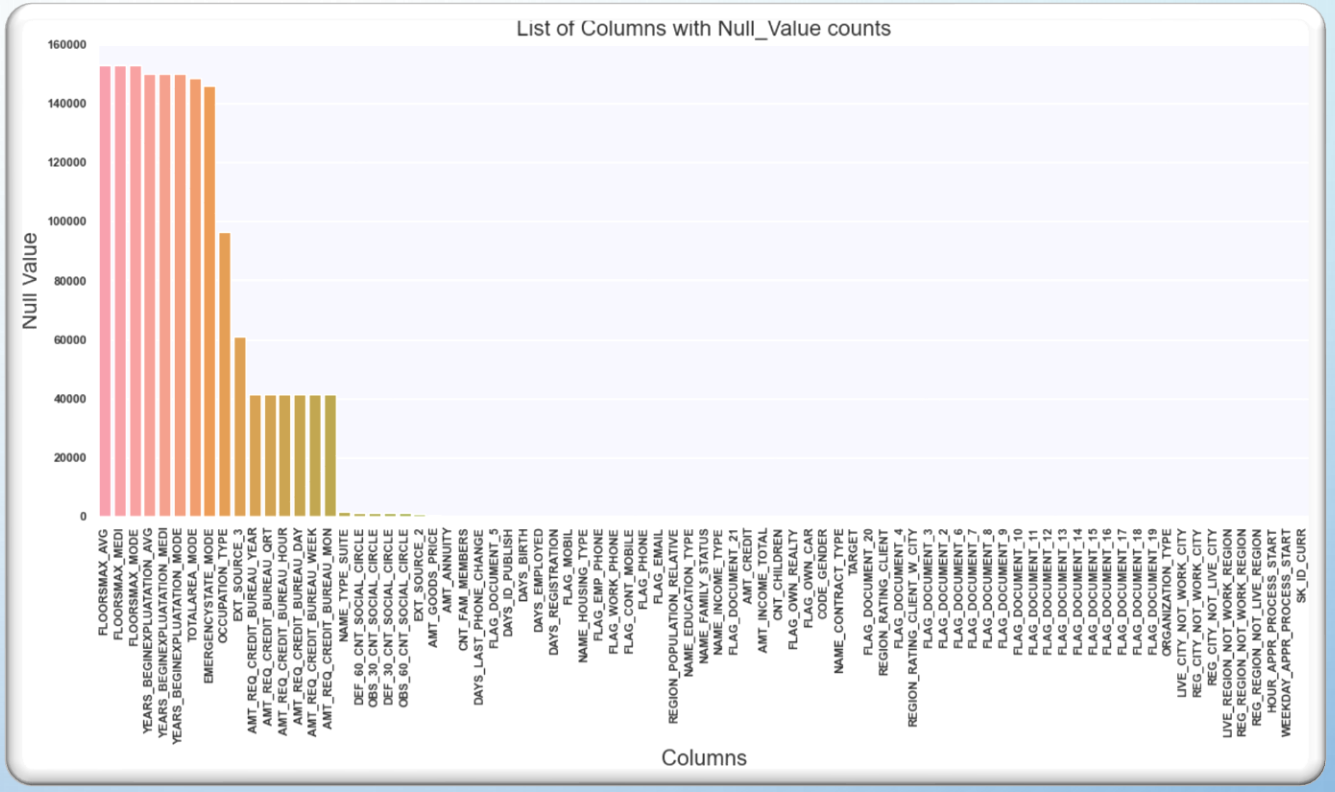
- Data analysis
  - ❖ Cleaned data used to investigate the relationship of variables with that of TARGET variable
  - ❖ Univariate and bivariate analysis to understand the patterns in data and drew conclusions
  - ❖ Data visualization techniques to summarize the results

DATA ANALYSIS

DATA CLEANING

- ***INFERENCES*** **:**

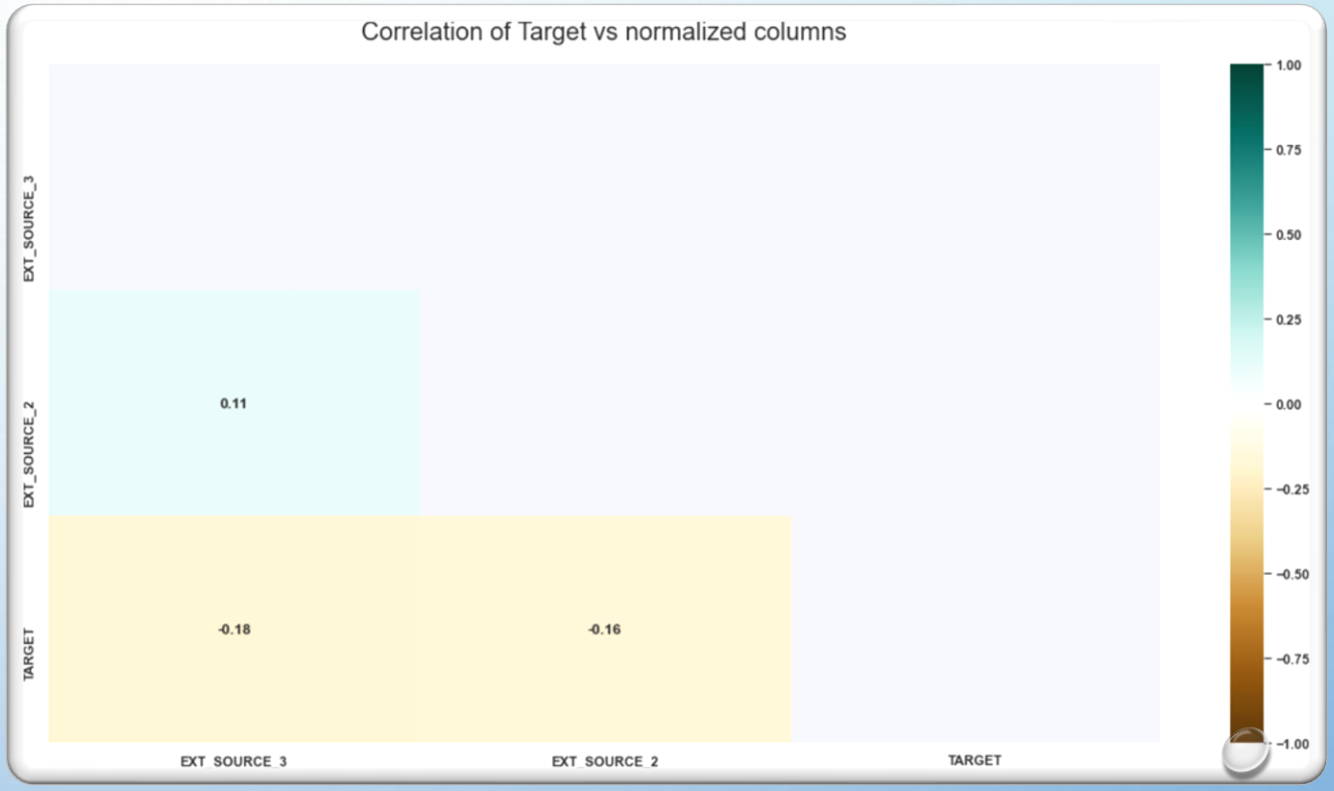
- Initially there are total 122 columns & 307511 rows
- Null values with higher % identified(say >50%) and dropped
- The columns with lesser % of missing values are identified and imputed with suitable technique viz. mean, mode
- Next the columns which have +ve and -ve values require fixing



DATA CLEANING

- ***INFERENCES*** : Correlation is used to find relationship with TARGET variable and if its not can be eliminated

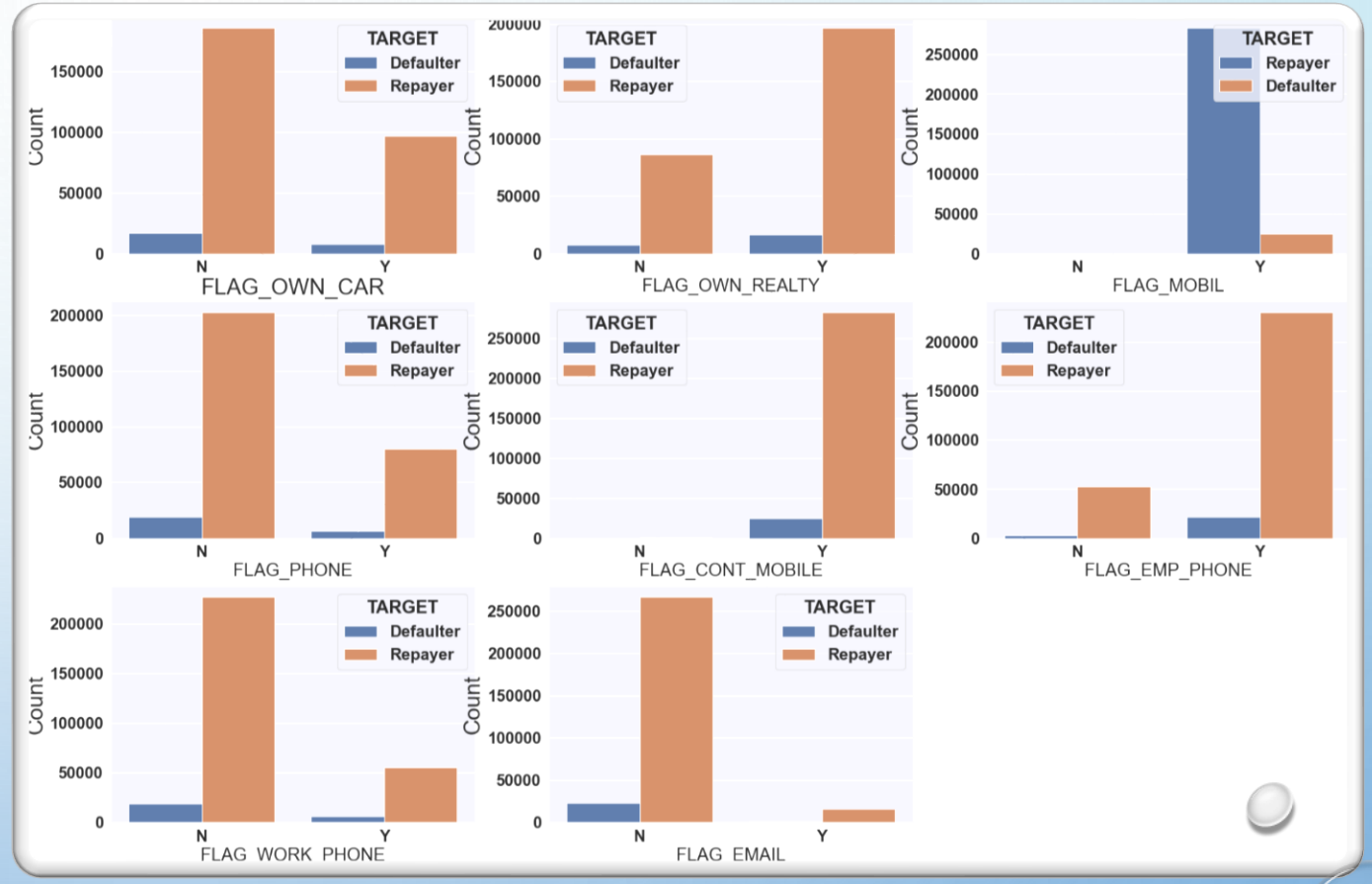
- Correlation used above to represent the statistical measure of linear relationship between two variables and target variable. This can help to measure of dependence between two different variables. The Correlation between Target and other normalized column is Weak negative suggesting no causal relationship. Hence we drop these two columns



DATA CLEANING

- ***INFERENCES*** : Plot all flag columns with respect to Target variable to get insight

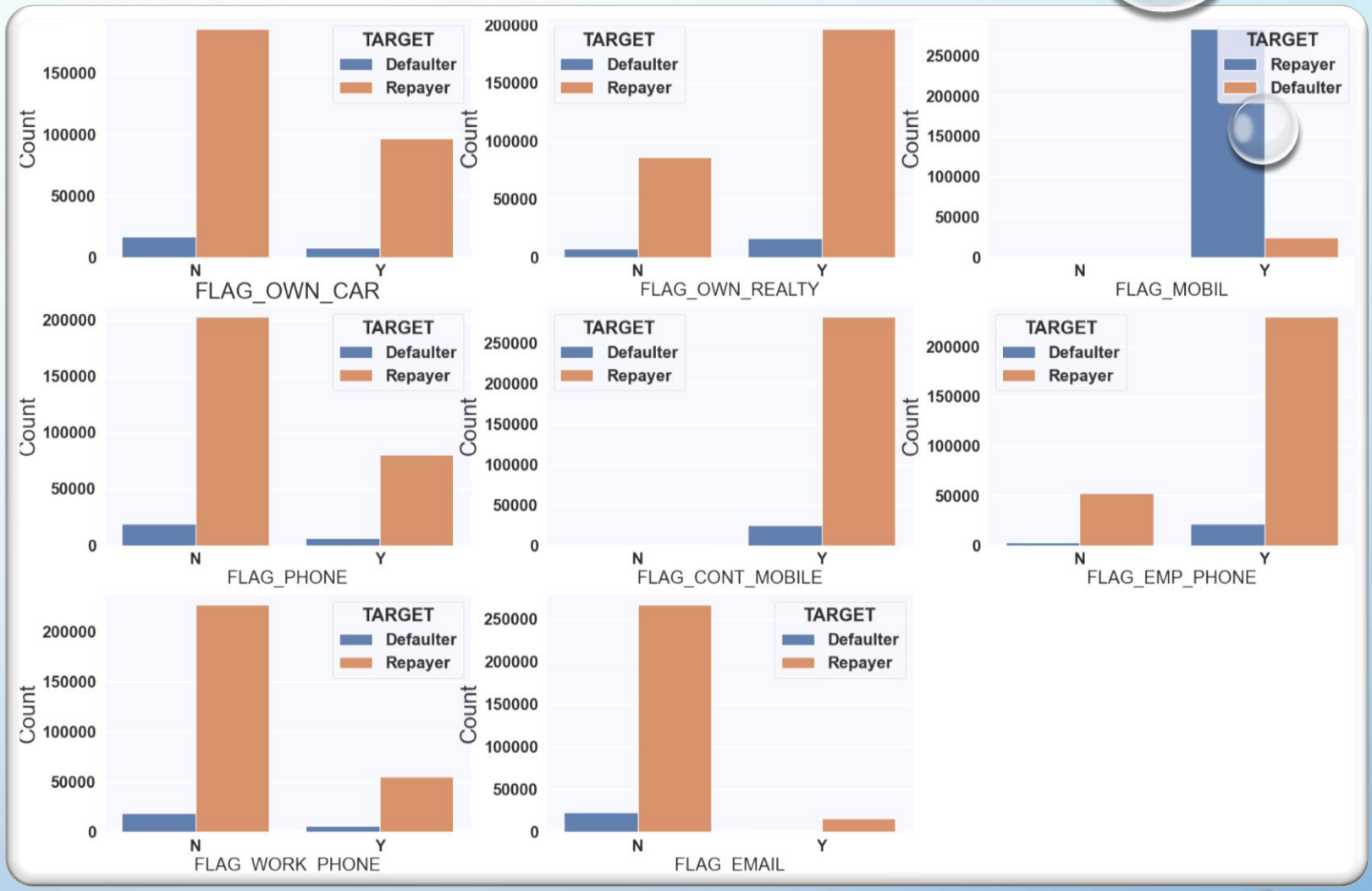
- Columns likes FLAG_OWN_REALTY, FLAG_MOBIL , FLAG_CONT_MOBILE, FLAG_EMP_PHONE, shows there are more repayor's than defaulter and hence keep these columns and remove all other columns for further analysis



DATA CLEANING

- ***INFERENCES*** : Plot all flag columns with respect to Target variable to get insight

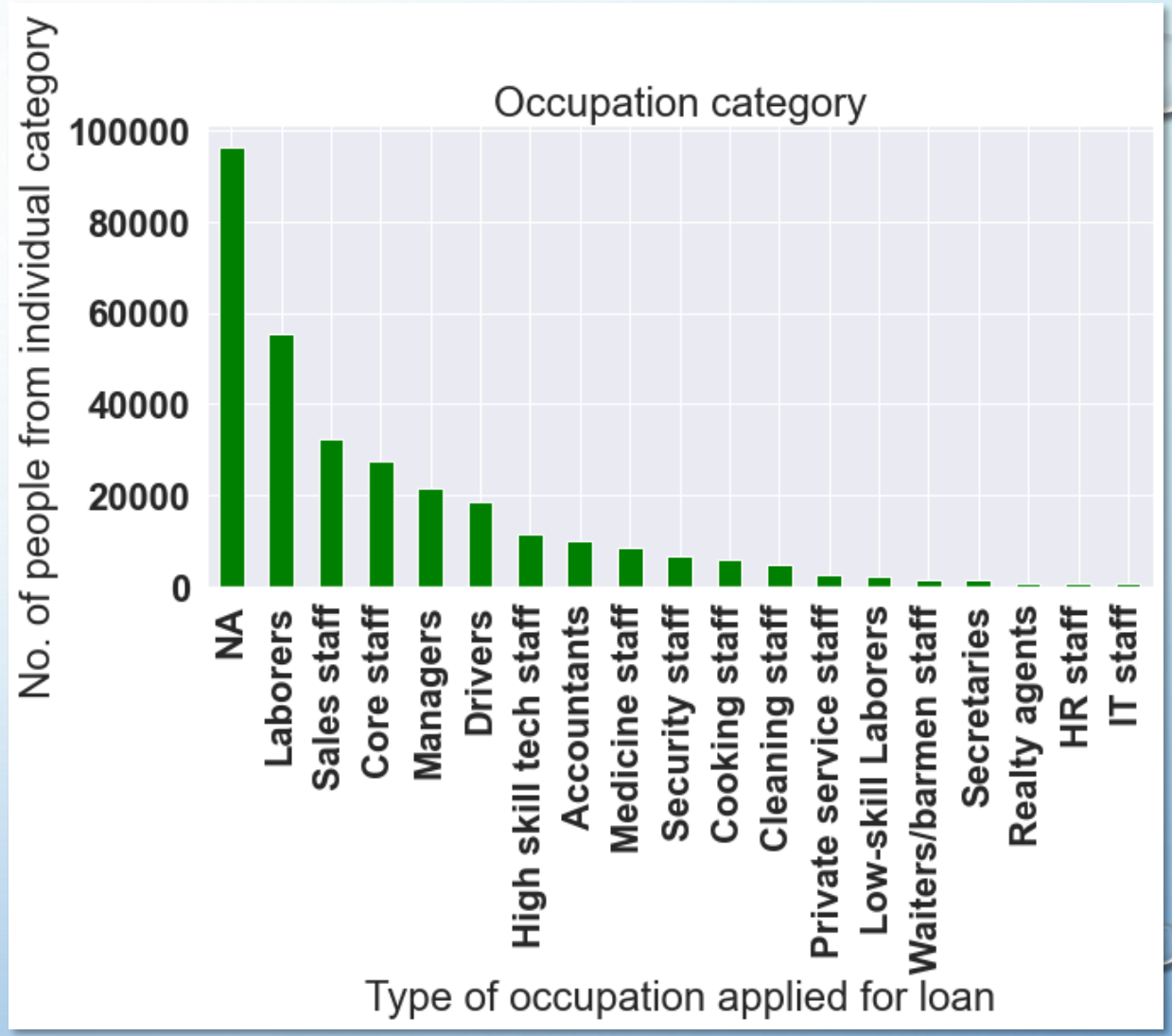
  - Columns likes FLAG_OWN_REALTY, FLAG_MOBIL,FLAG_CONT_MOBILE, FLAG_EMP_PHONE, shows there are more repayor's than defaulter and hence keep these columns and remove all other columns for further analysis



Note : Similar approach has been followed with other variable and shown in python file

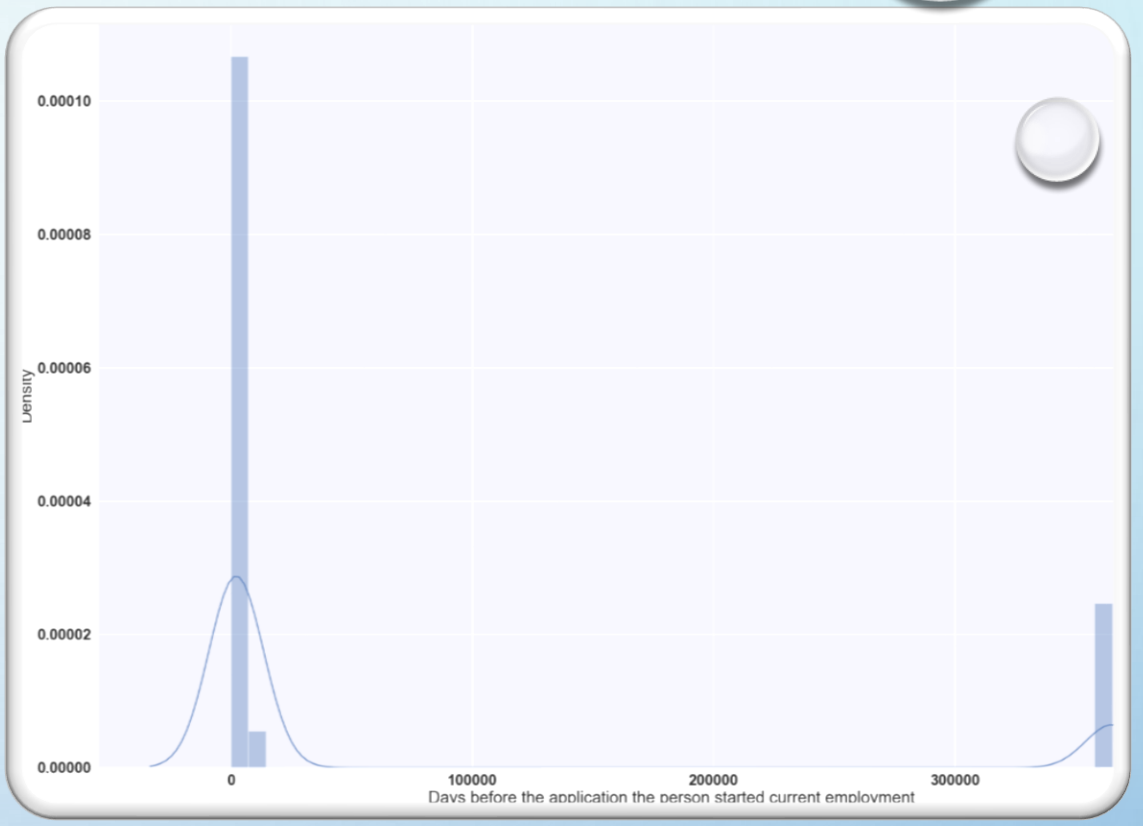
DATA IMPUTATION

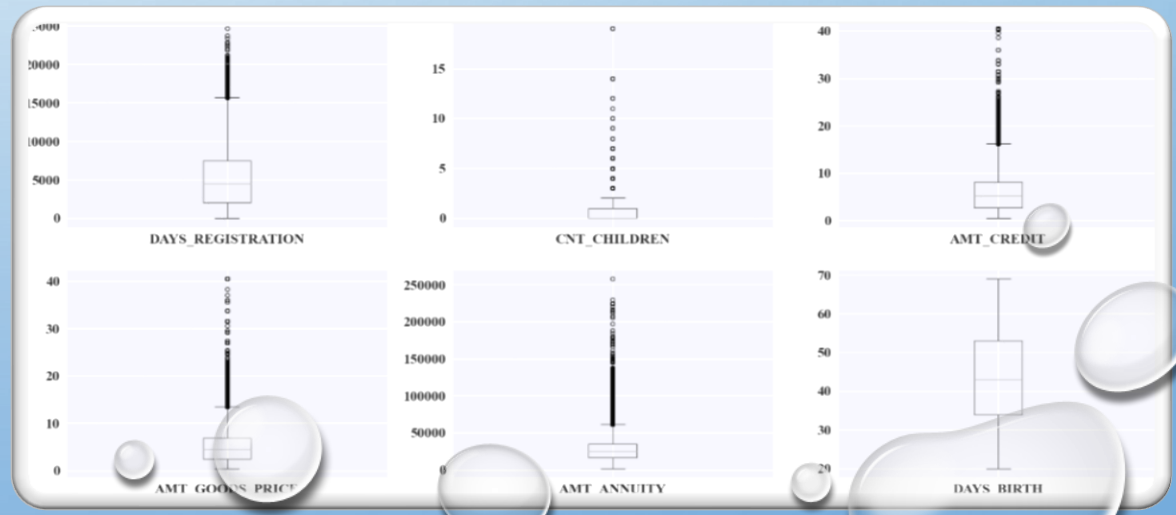
- columns having null values greater than 1% , will impute that lets do it one by one
- ***INFERENCES:***
  - NA are the highest % amongst all people applied for loan
  - The columns imputed with mode and median



INVESTIGATING OUTLIERS

- ***INFERENCES*** :
- The column 'DAYS_EMPLOYED' shows Days before the application the person started current employment. Hence if it is more than 20,000 that can be outlier
- Does not see any outlier in DAYS_BIRTH
- In column 'AMT_ANNUITY' a single value of > than 250000 is an outlier.
- In the column 'DAYS_REGISTRATION' a value greater than 24000 is an outlier.



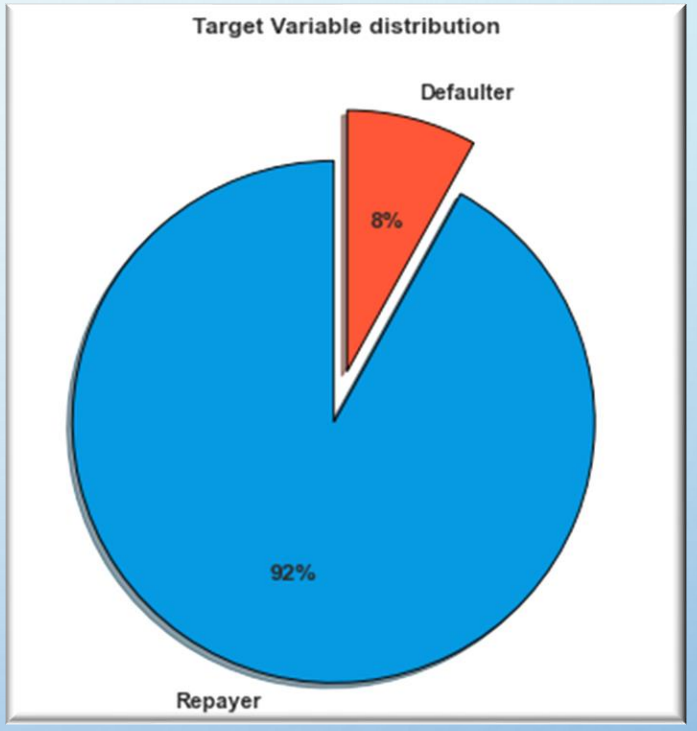


IDENTIFYING DRIVING FACTORS FOR BUSINESS DECISIONS USING DATA ANALYSIS

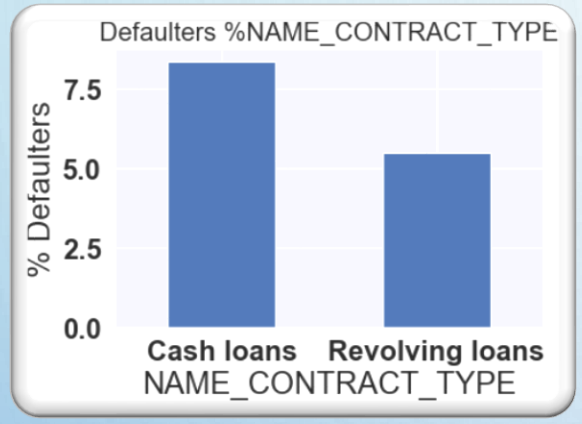
DATA IMBALANCE

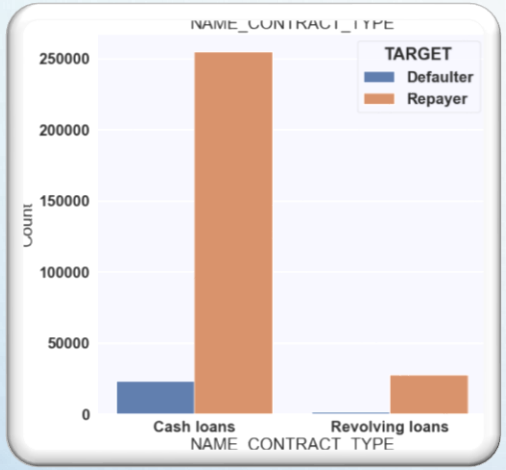
INFERENCES :

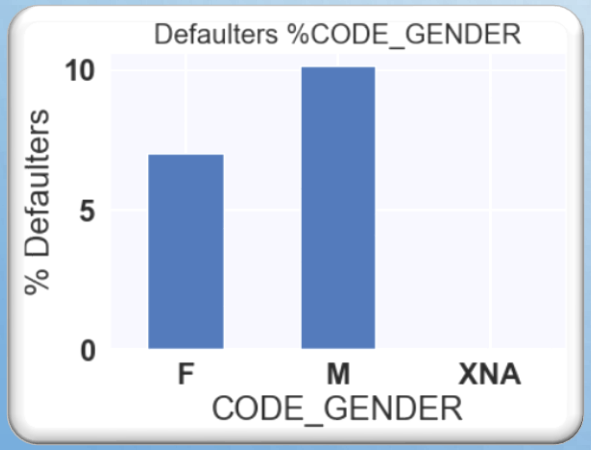
- The data when analysed w.r.t Target variable it clearly indicates imbalance. (Ratio of Data Imbalance 11.4)
- The univariate and bivariate analysis will make things more clear

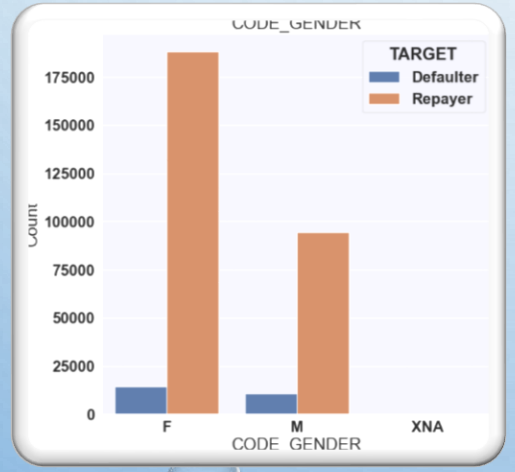


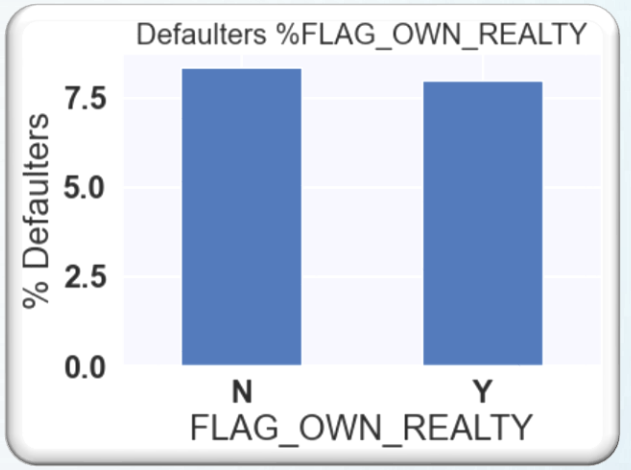
UNIVARIATE ANALYSIS

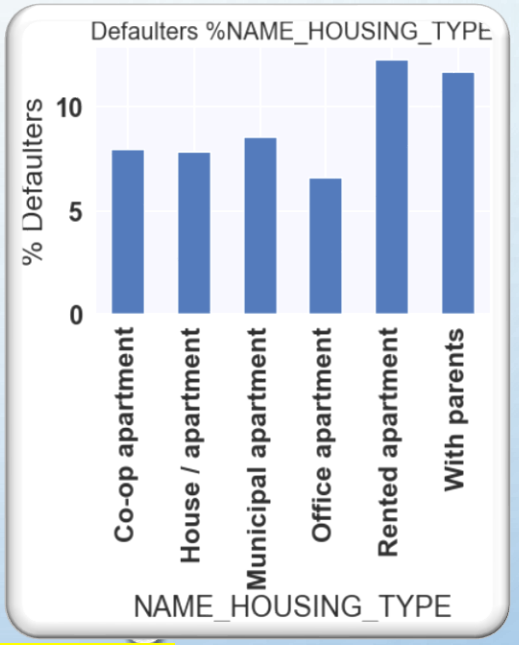


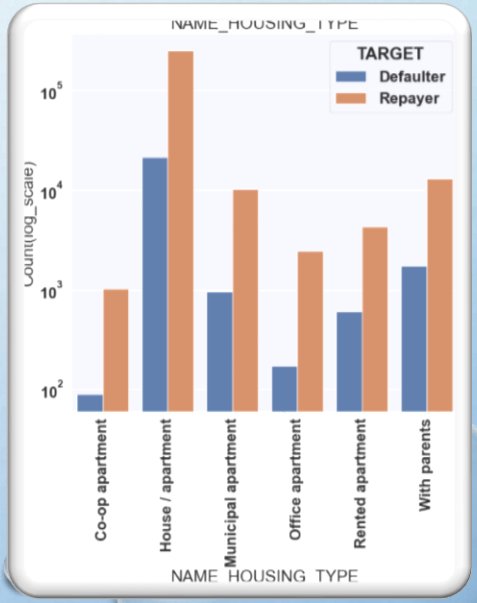






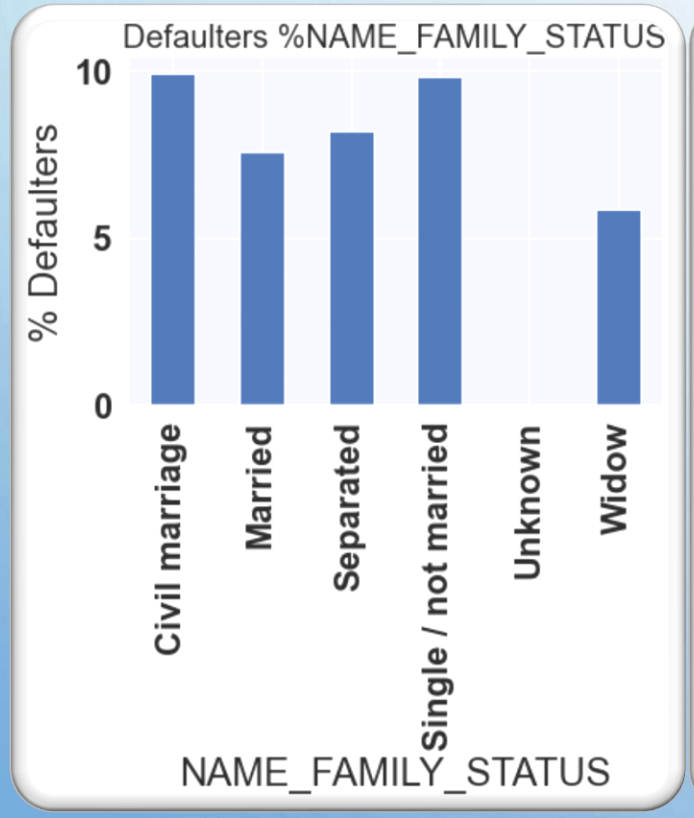


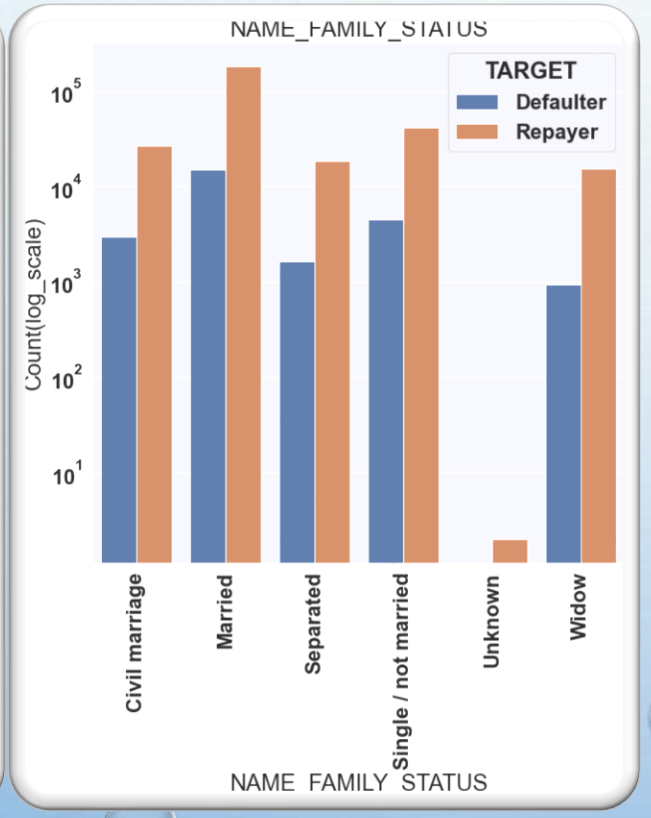


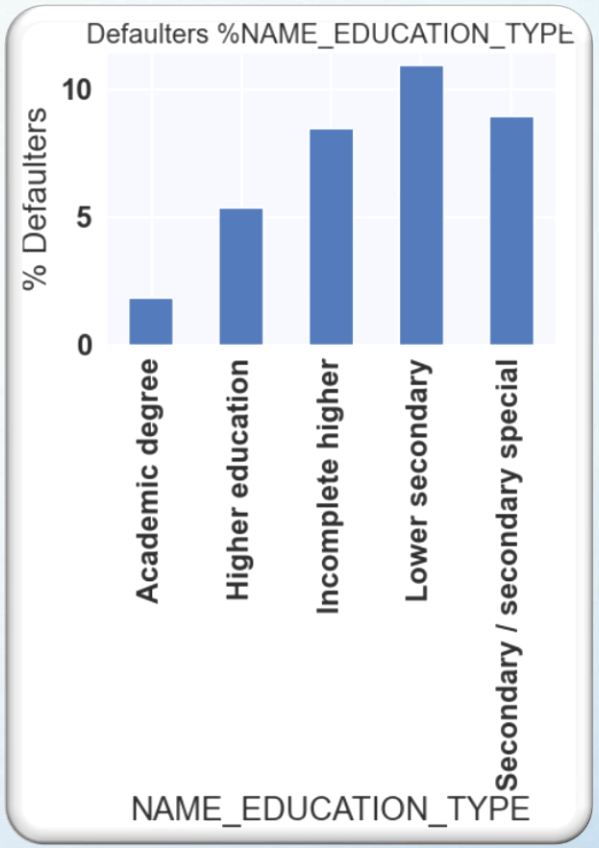


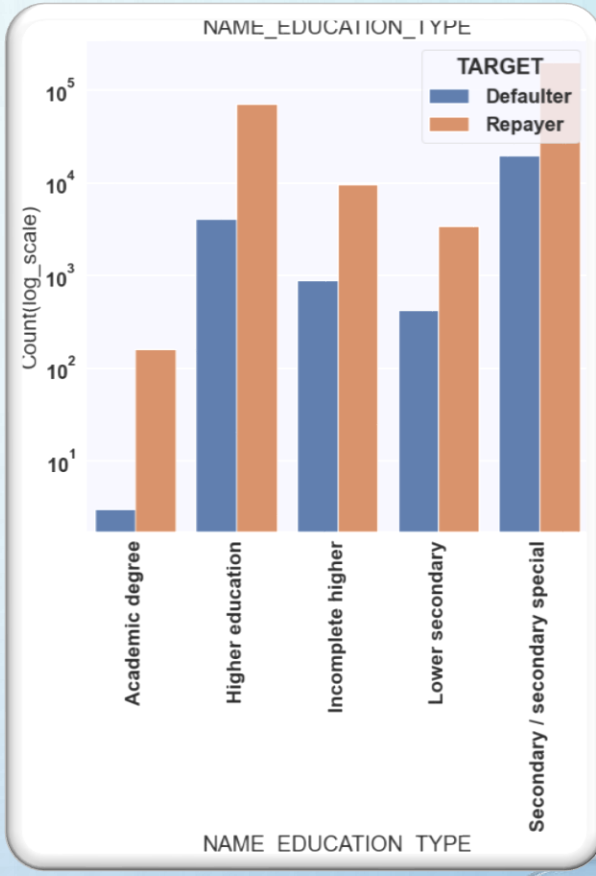
Inferences : Columns likes FLAG_OWN_REALTY, FLAG_MOBIL , FLAG_CONT_MOBILE, FLAG_EMP_PHONE, shows there are more repayers than defaulter and hence keep these columns

UNIVARIATE ANALYSIS





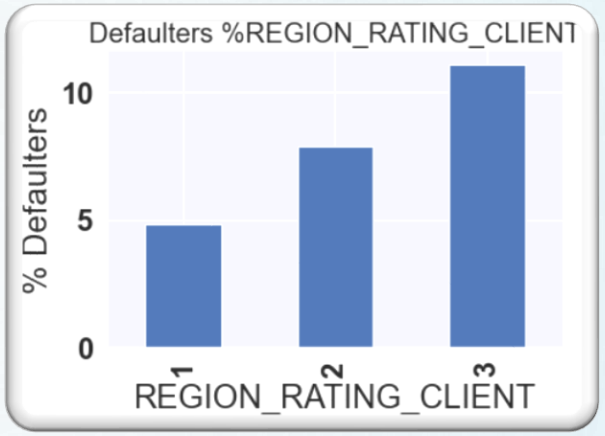


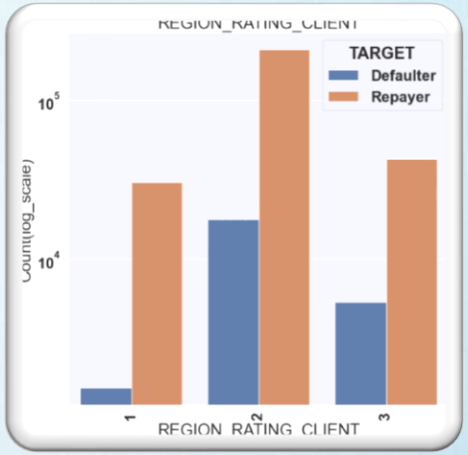


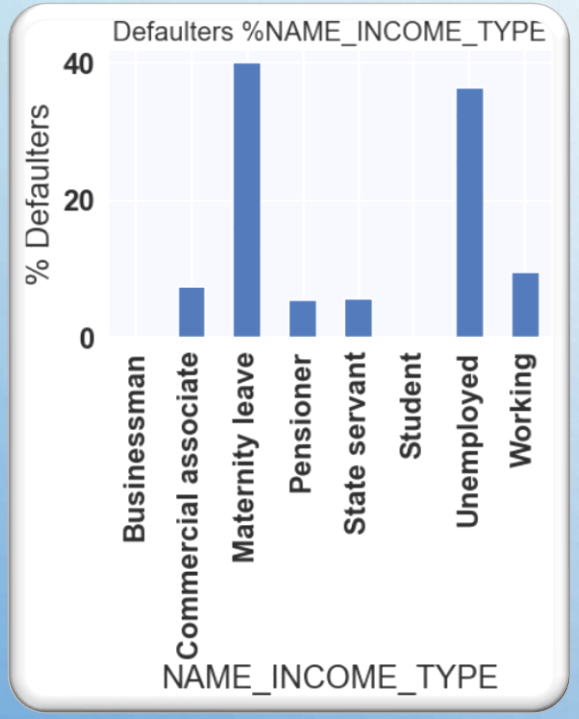
Inferences:

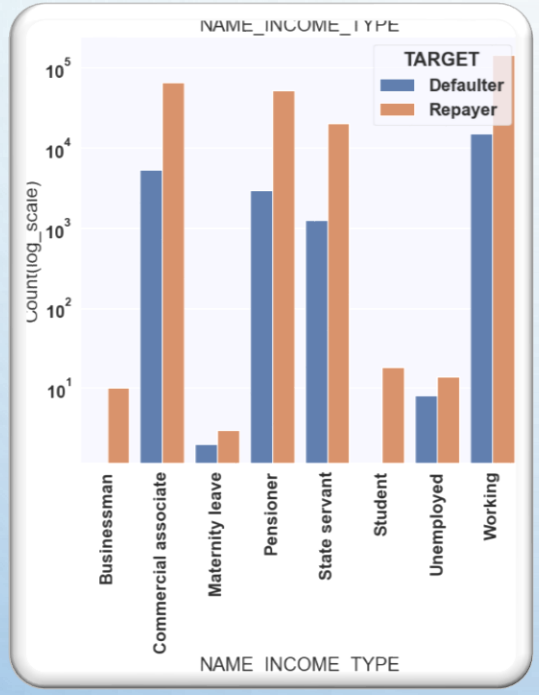
1. Clients with Single/not married and Civil marriage status have highest chances of being defaulter. But Married people has least chances of being defaulter
2. Clients with Lower secondary education have more chances than any client with any other education type and but there are more client with secondary/secondary special education Type

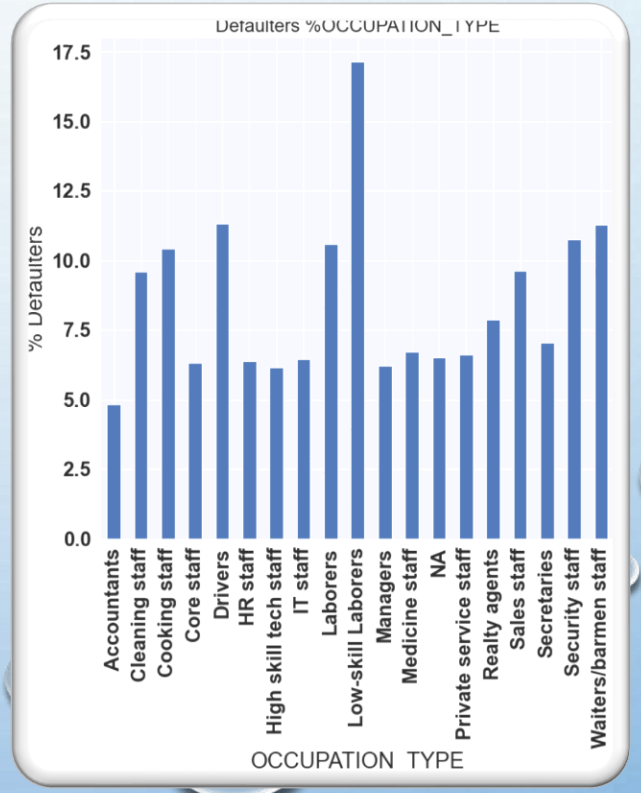
UNIVARIATE ANALYSIS

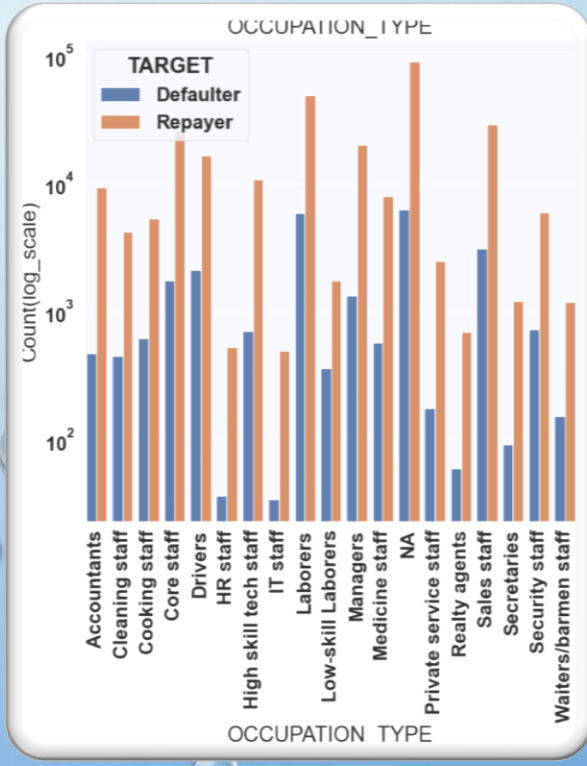






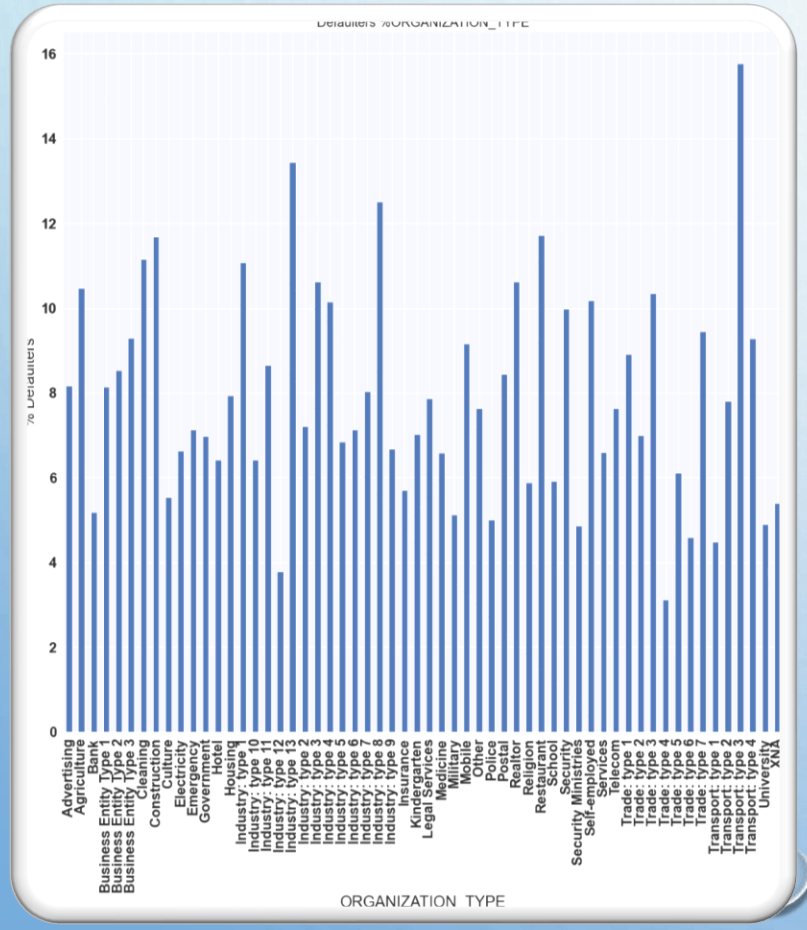


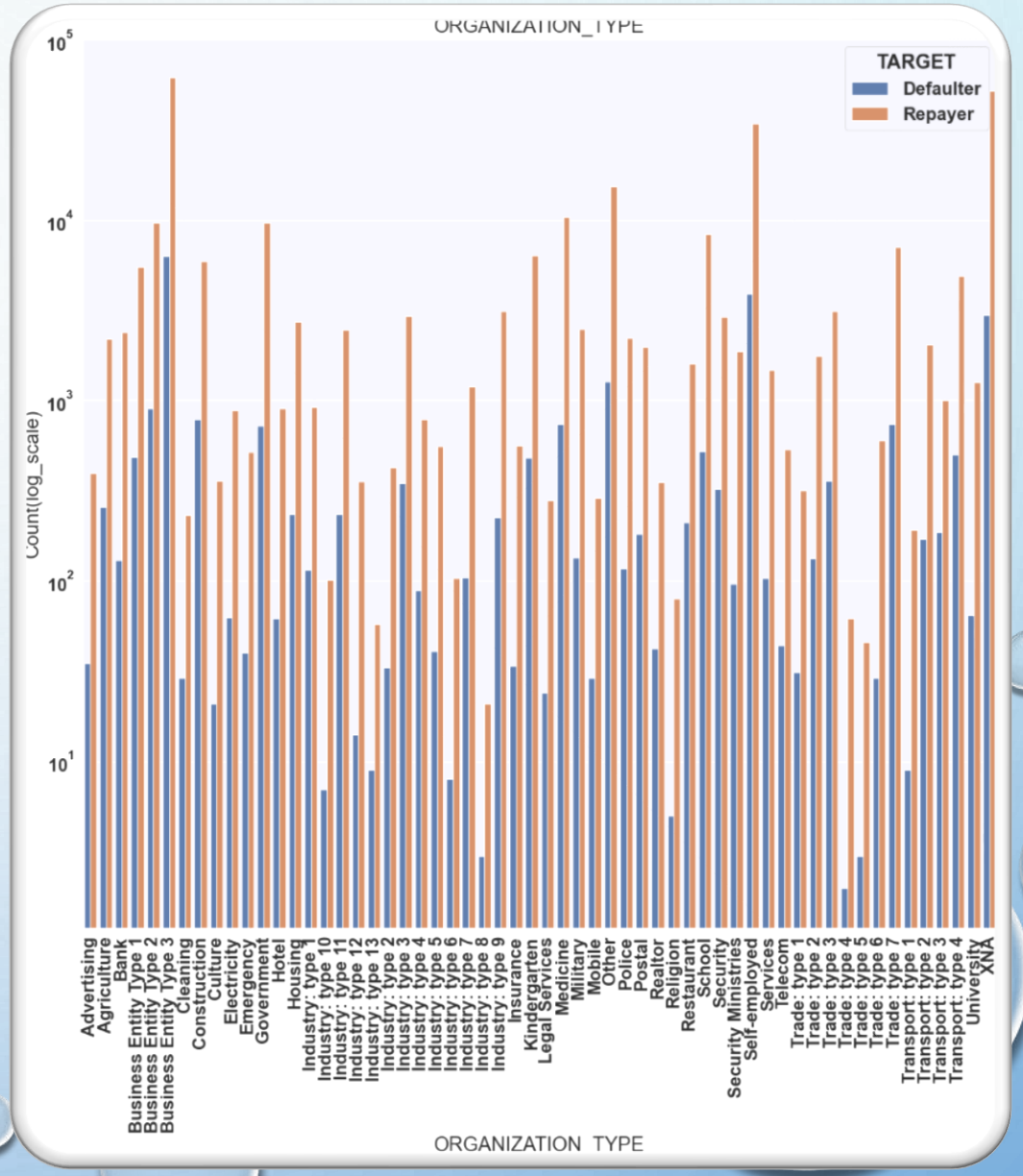




Inferences : Low skills labourer have highest chance of being defaulter and IT staff have less chances of being applied for loan

UNIVARIATE ANALYSIS

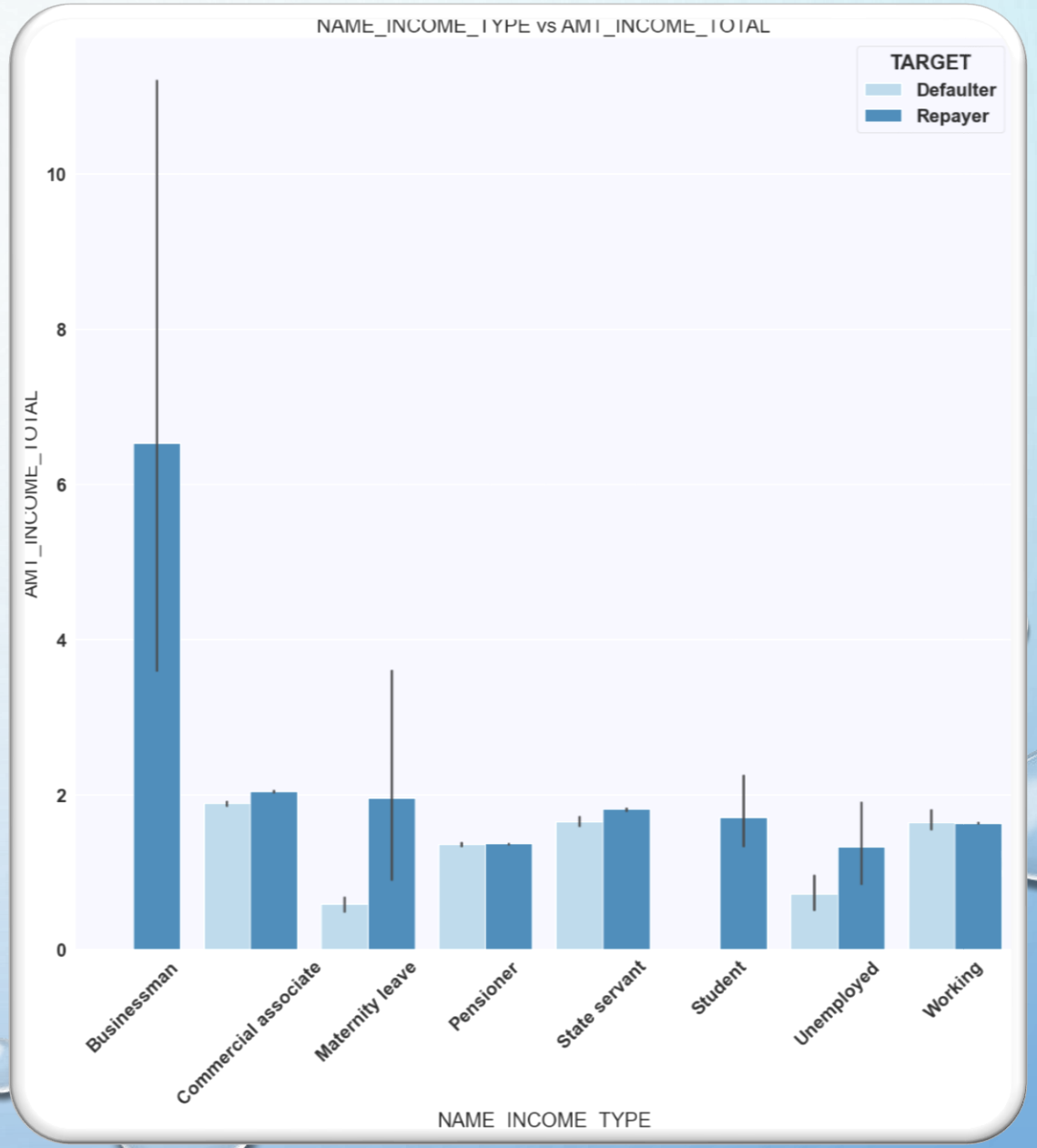




Inferences : Clients with type 3 are highest defaulter and most application are from XNA and Business entity 3

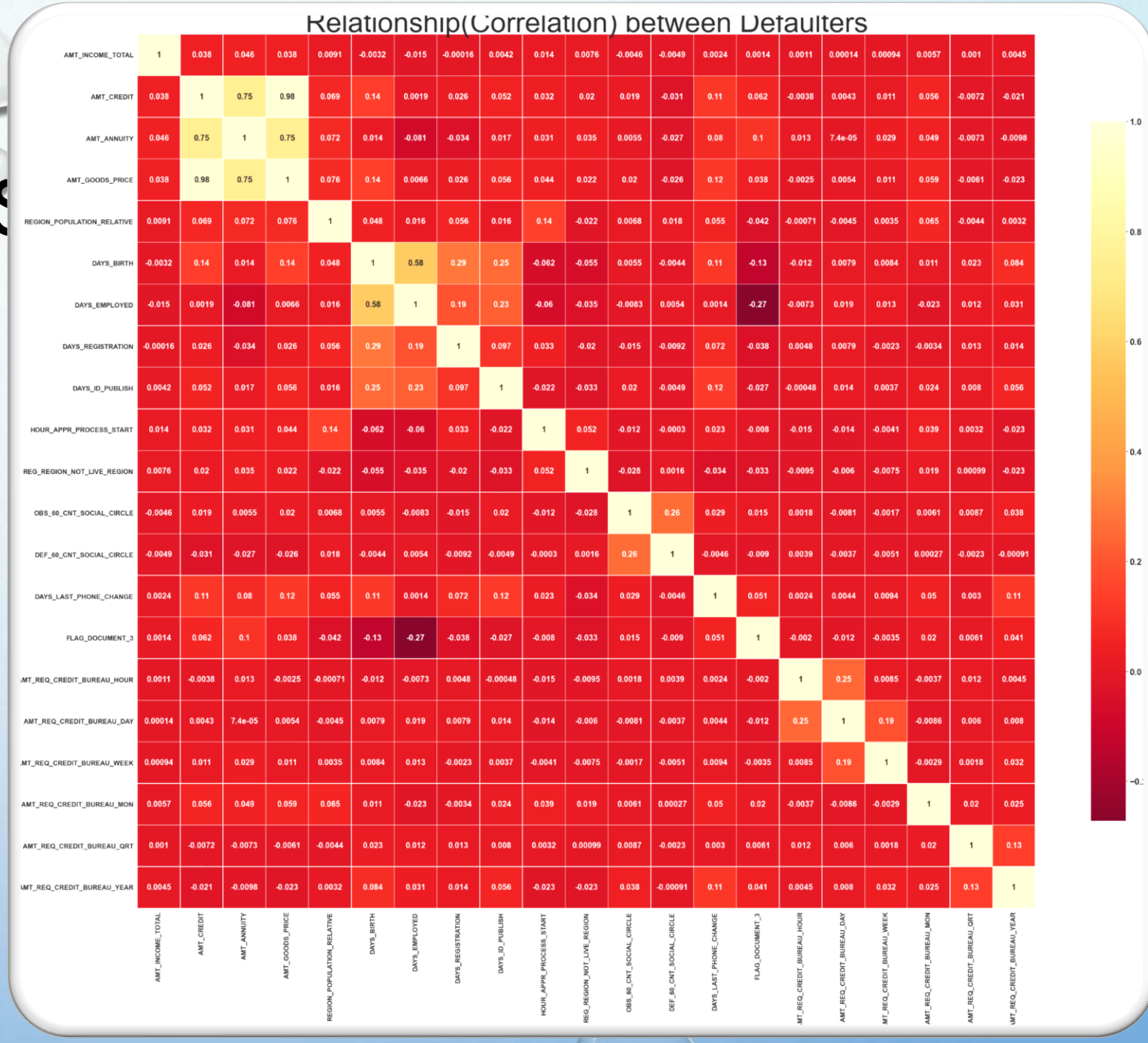
BIVARIATE ANALYSIS

Inference: Businessmen category has highest income with range of ~ 3.5 - 10.5 and more likely to apply for loan



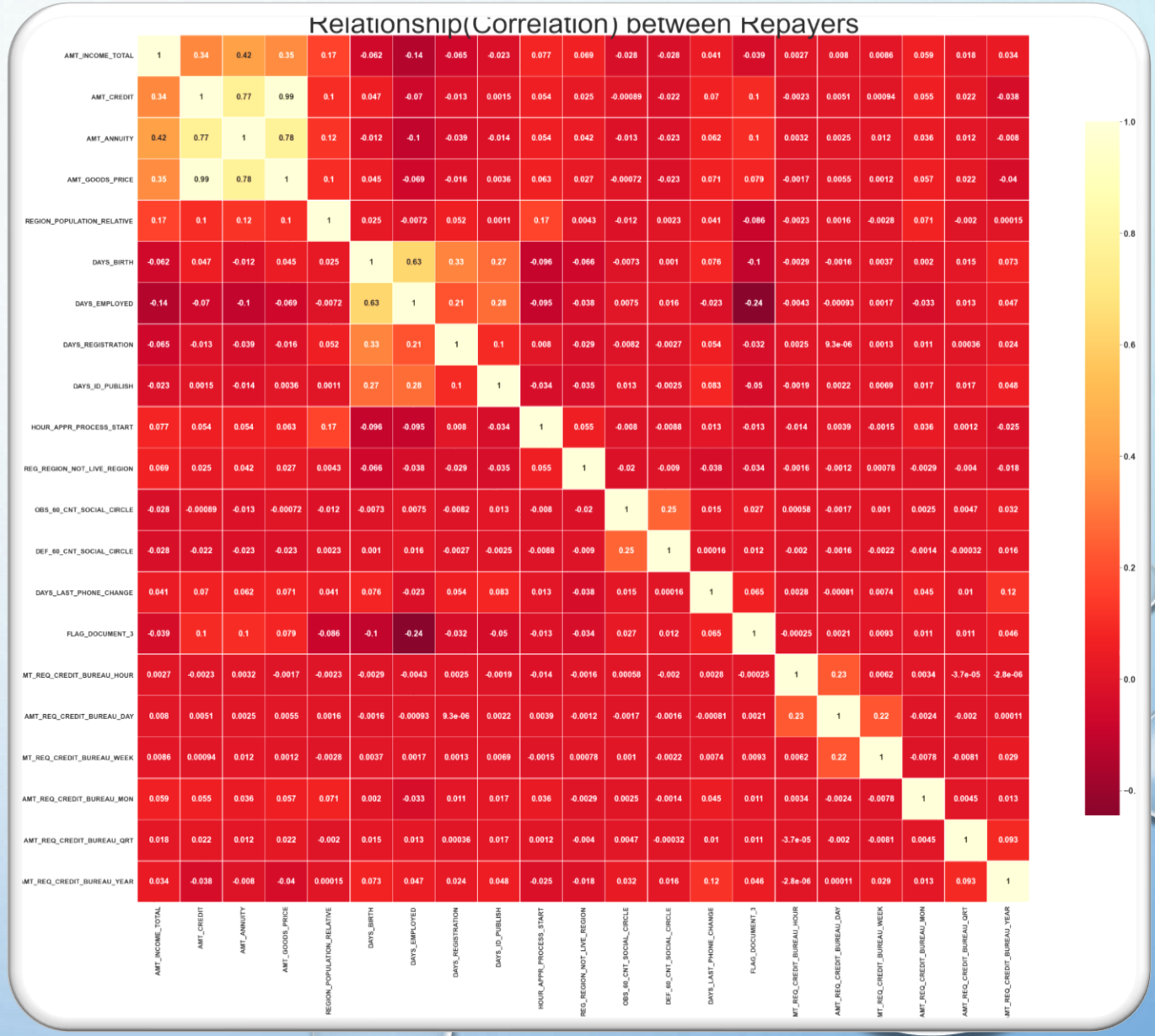
BIVARIATE ANALYSIS DEFAULTERS

Inferences : Credit amount of the loan is correlated well with loan annuity and the price of the goods for which the loan is given. Also number of days employed carry more weightage with loan repayment

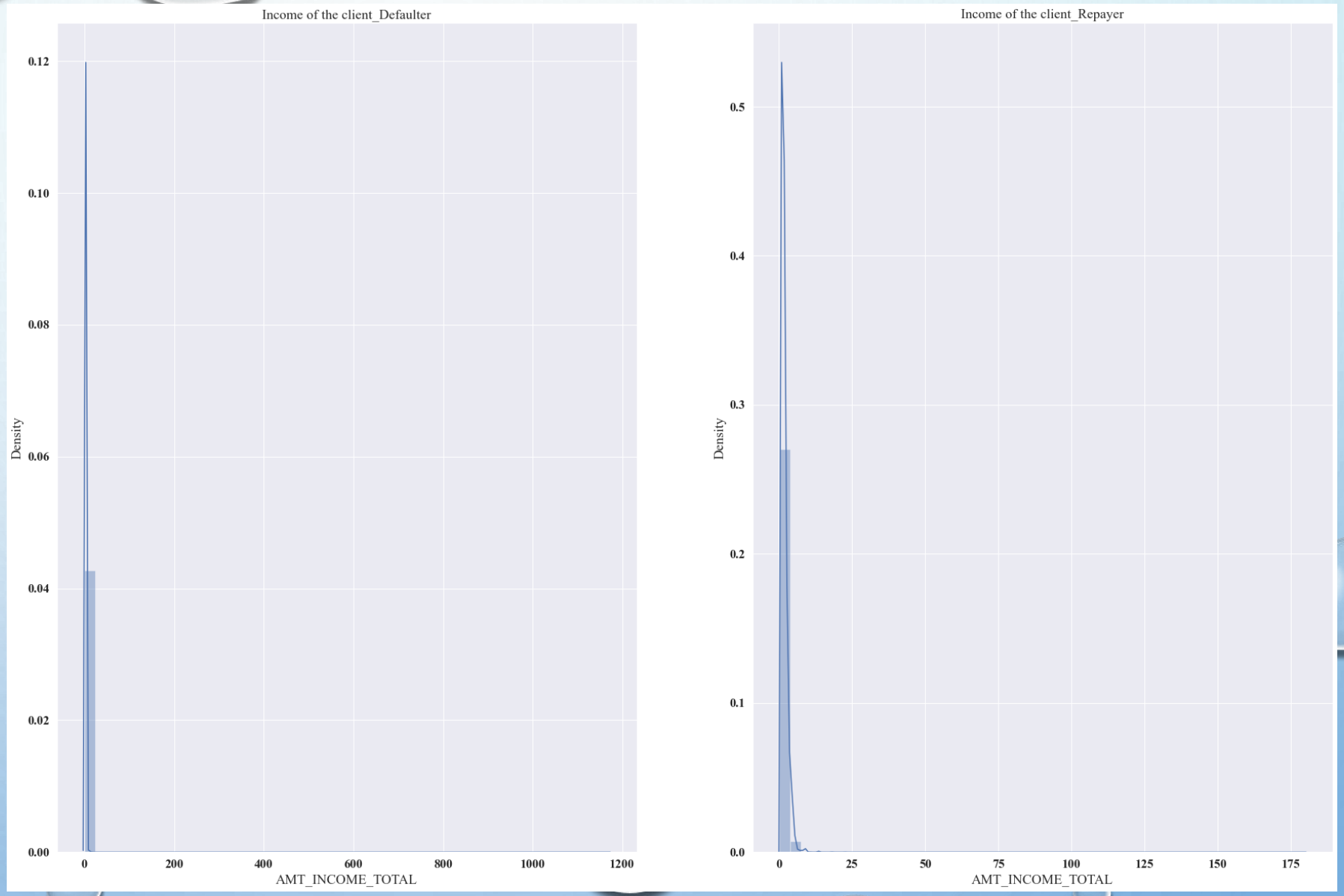


BIVARIATE ANALYSIS REPAYERS

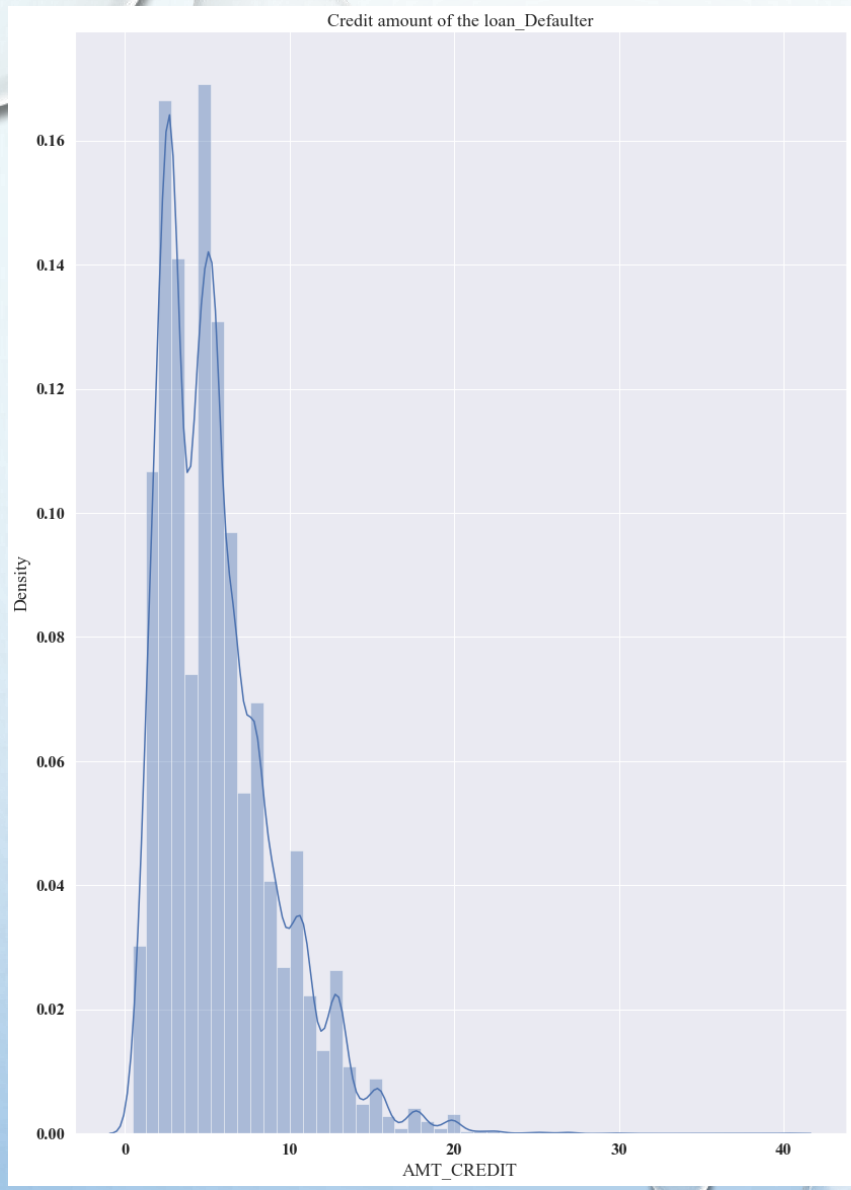
Inferences : Credit amount is strongly correlated with loan annuity and the price of the goods for which the loan is given and Total Income. Also repayers have good correlation in number of days client employed

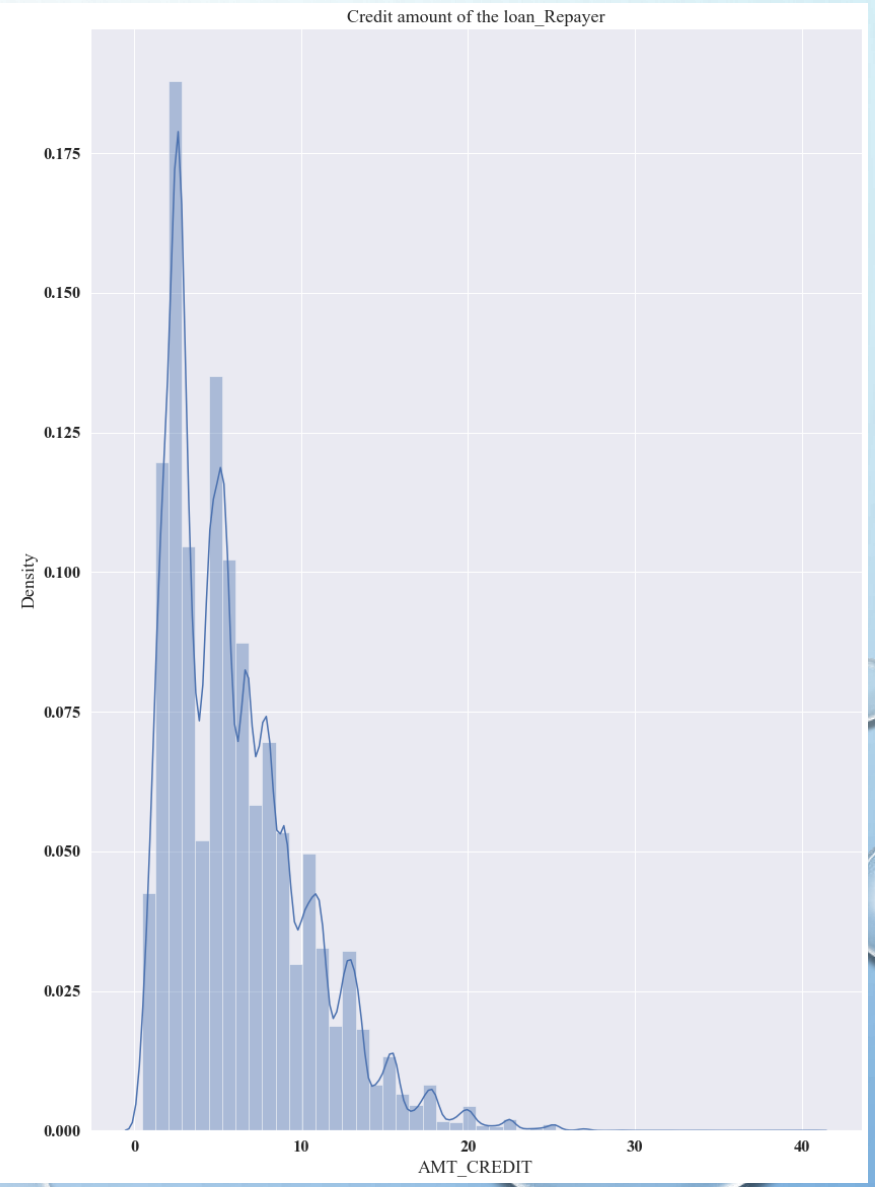


BIVARIATE ANALYSIS

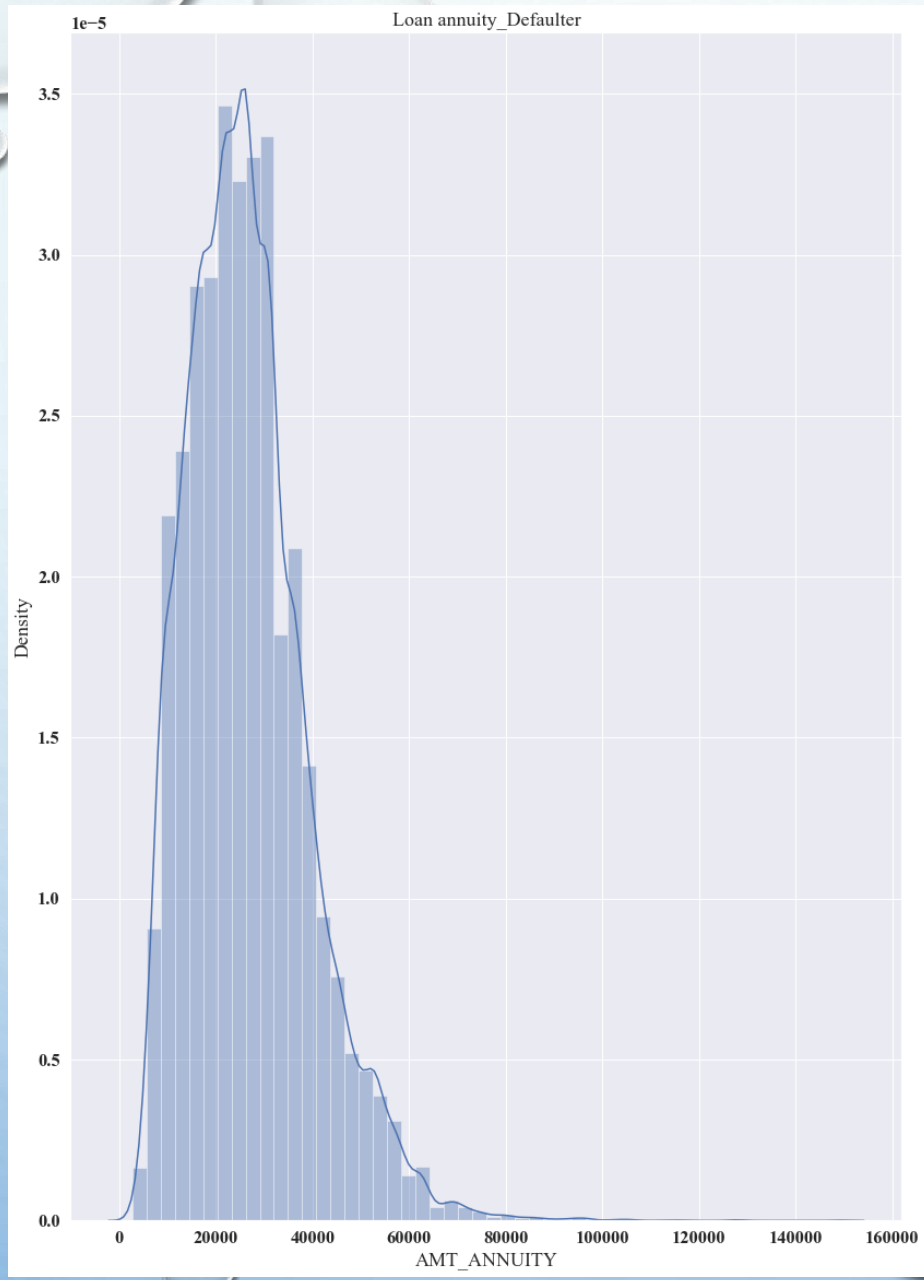


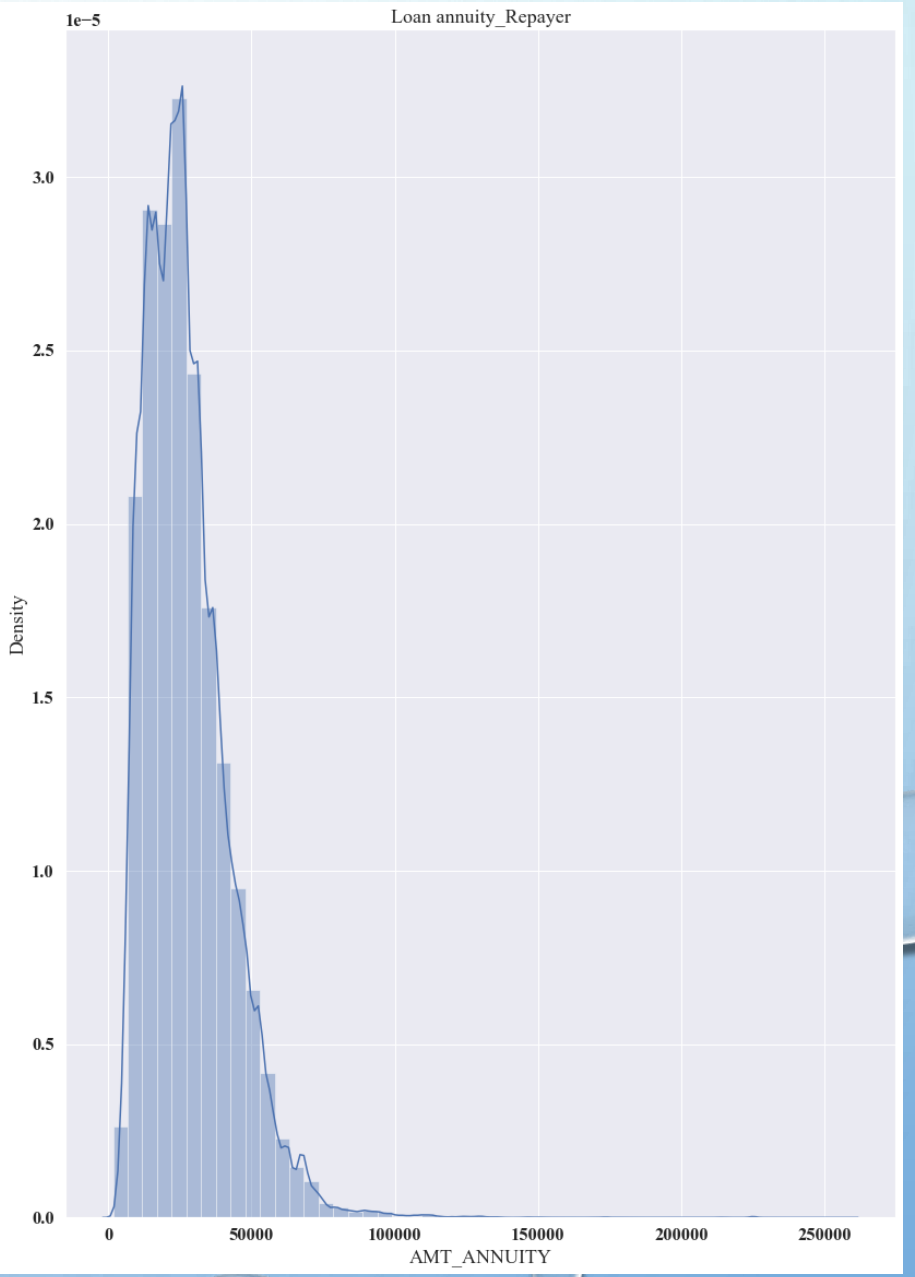
BIVARIATE ANALYSIS





BIVARIATE ANALYSIS REPAYERS

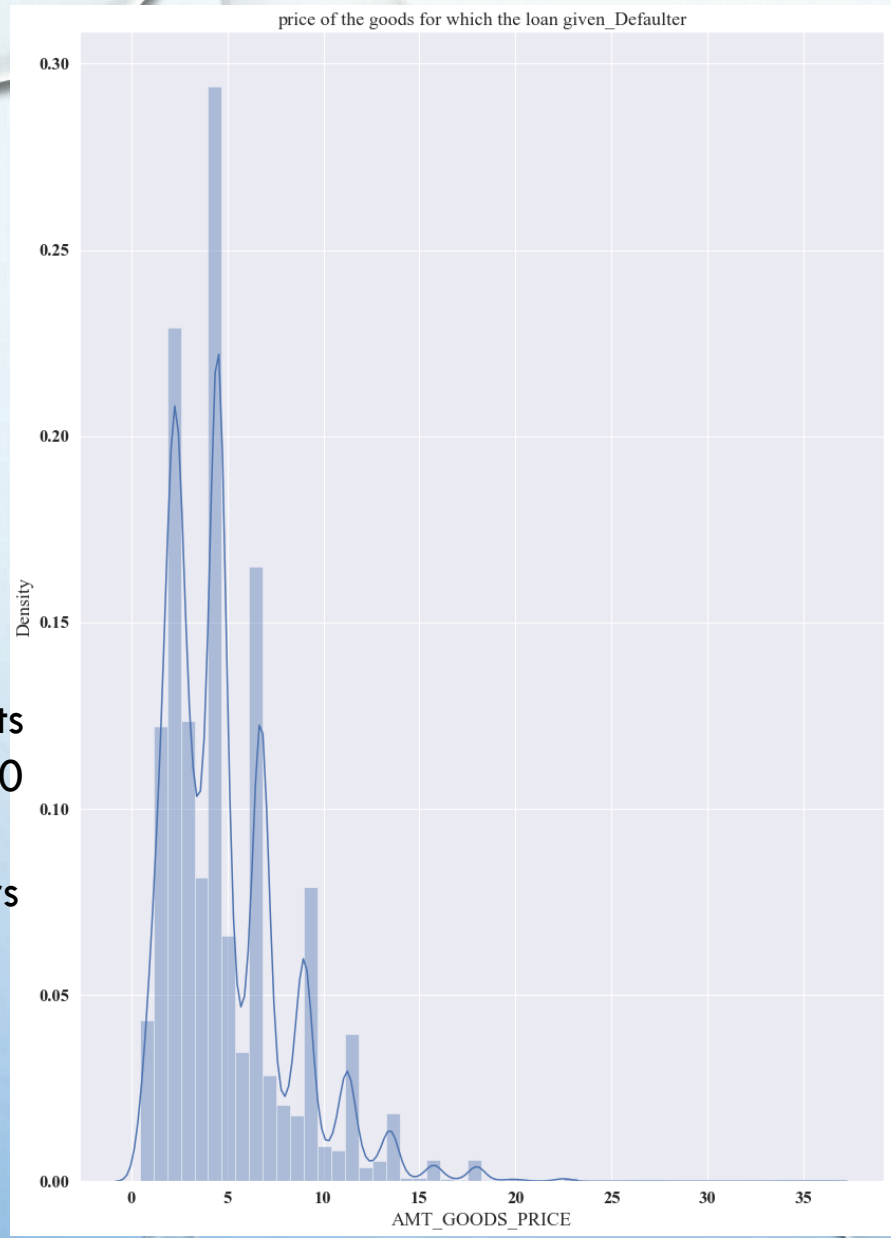


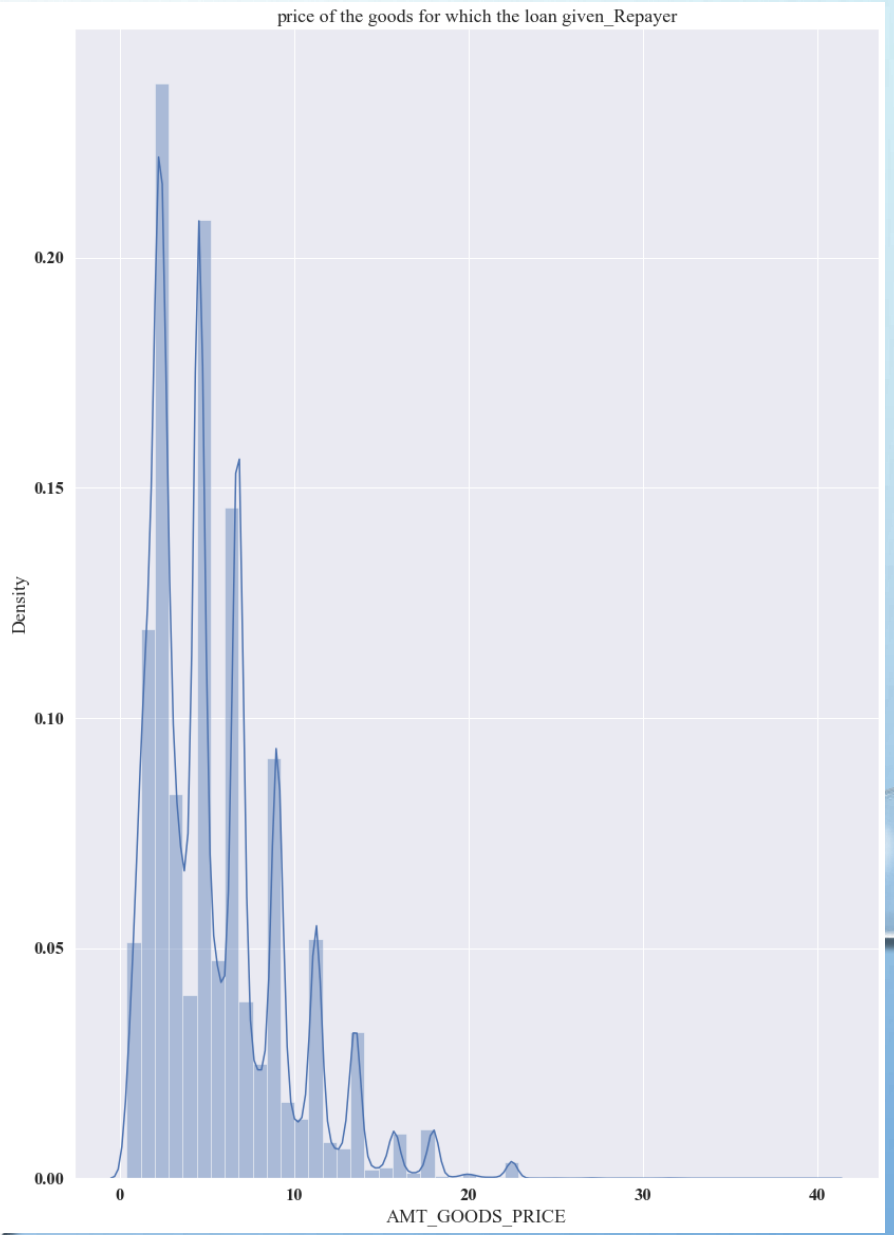


BIVARIATE ANALYSIS

Inferences :

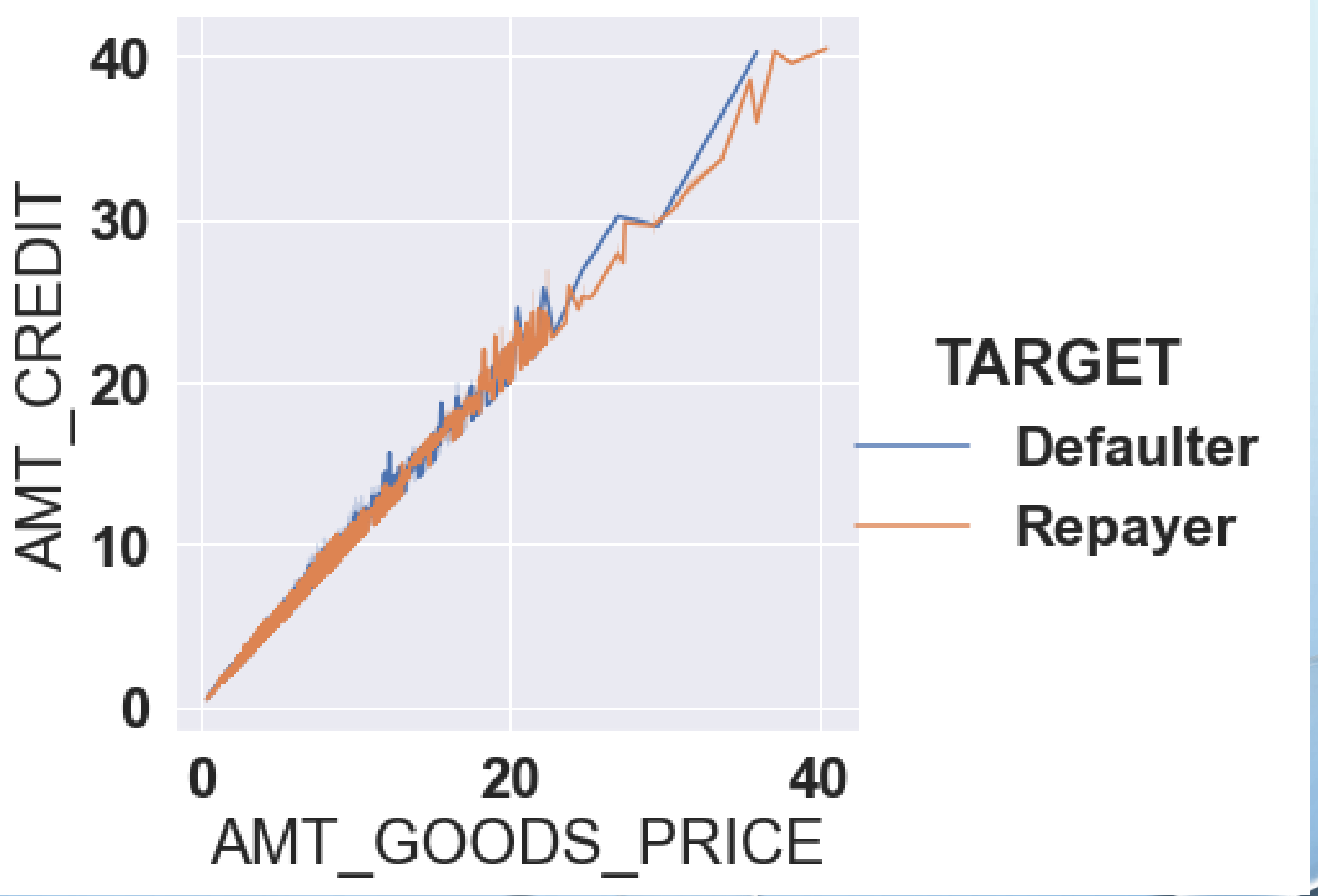
1) Credit amount given to clients is approximately less than 10 lakhs
2) The repayers and defaulters shows similar distribution for all above four graphs and hence it may not suitable to make judgement based on this single varibale





BIVARIATE ANALYSIS

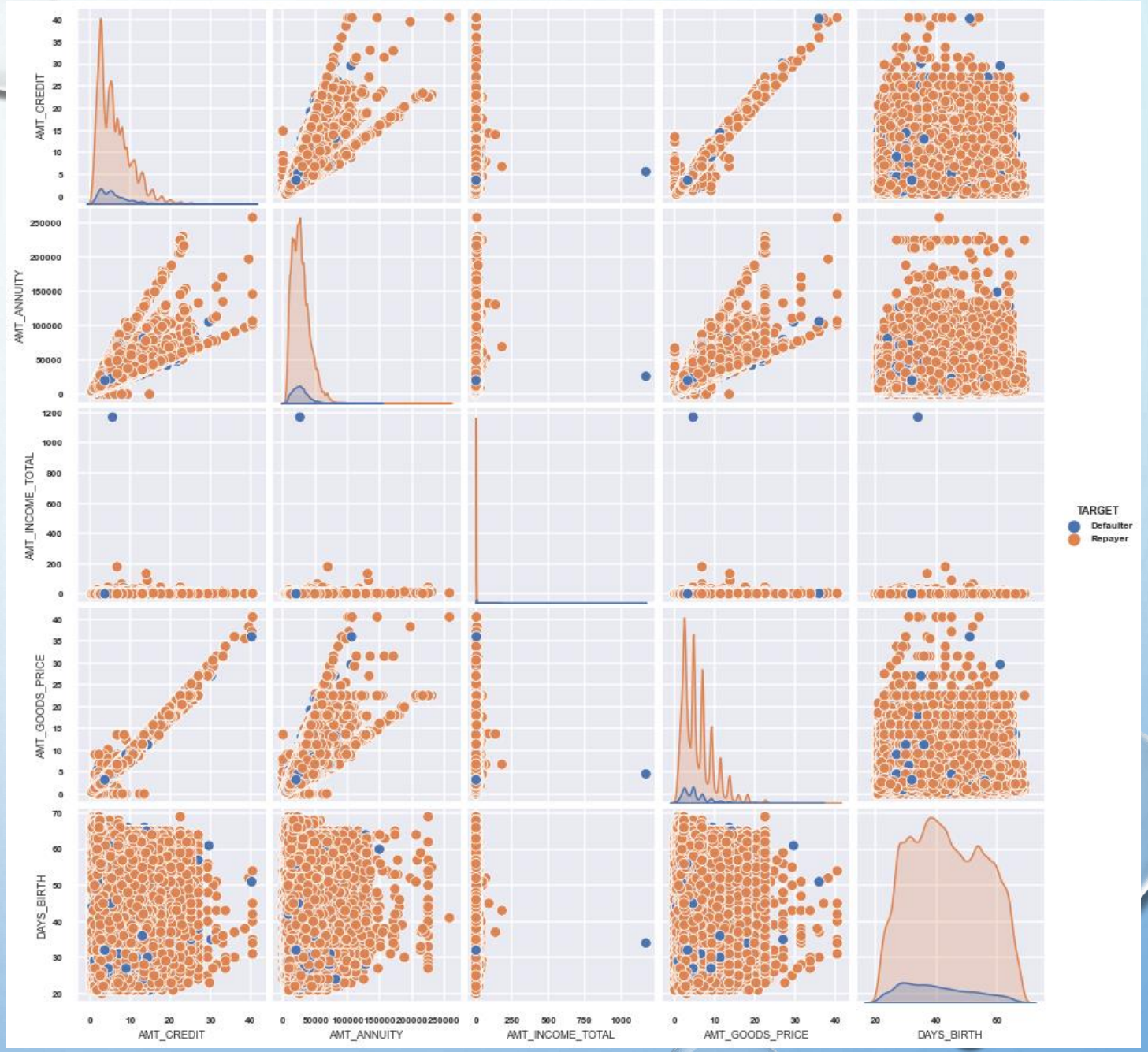
Inferences :
With loan amount more than 30 lakh the defaulter people increases



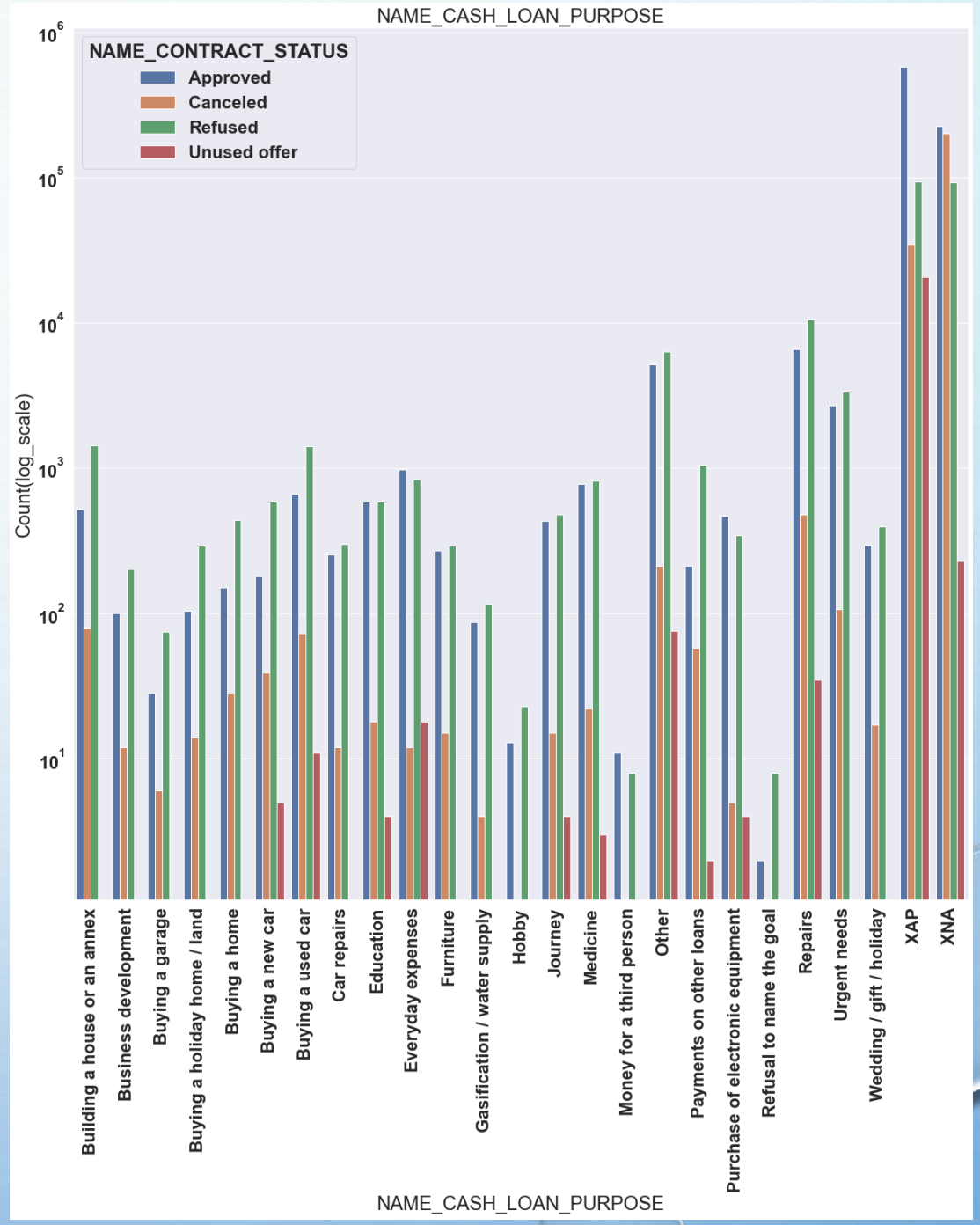
BIVARIATE ANALYSIS

Inferences :

1. Loan Amount and Goods price shows positive correlation
2. AMT_INCOME_TOTAL does not give any clear indication about defaulters and repayers



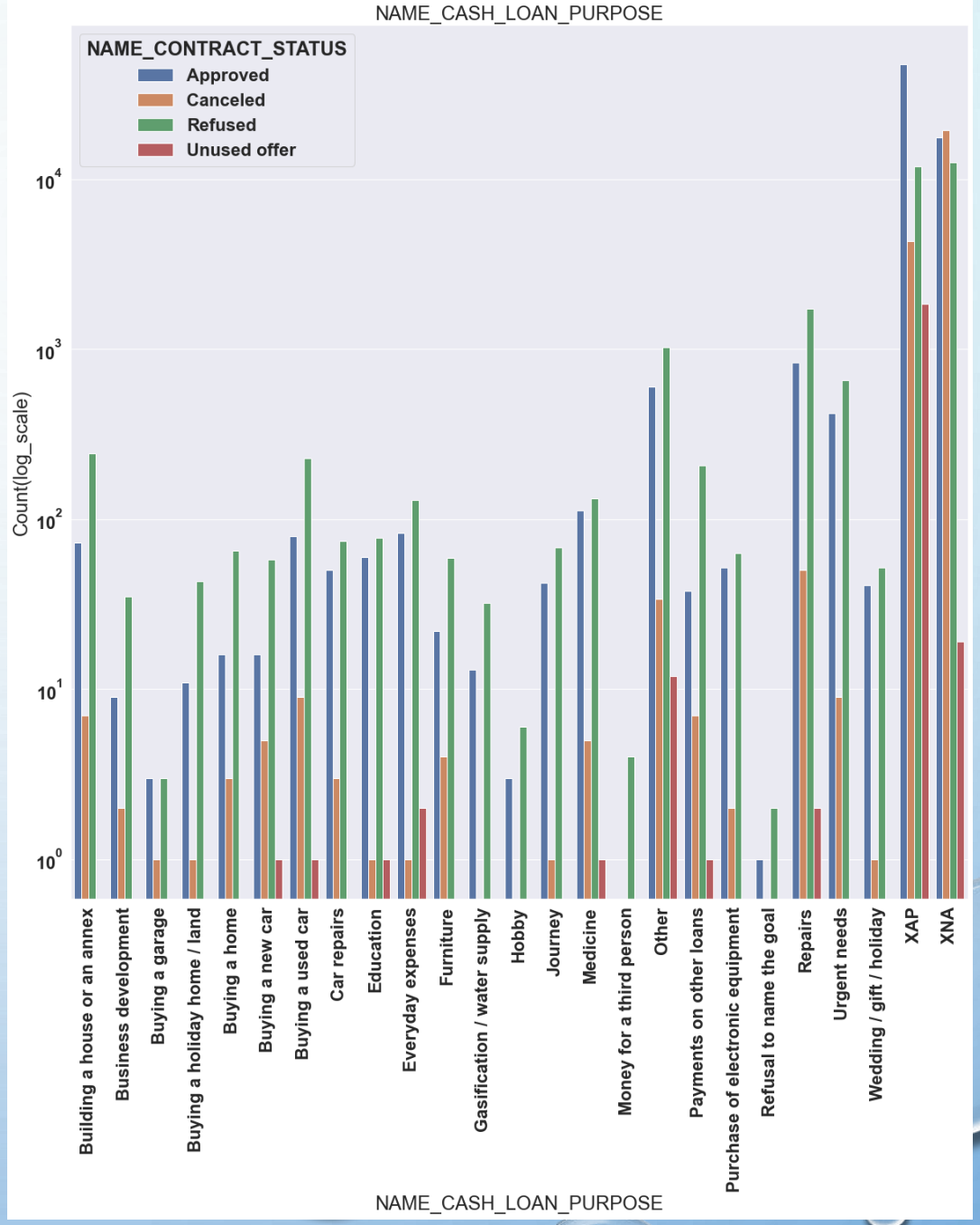
BIVARIATE ANALYSIS



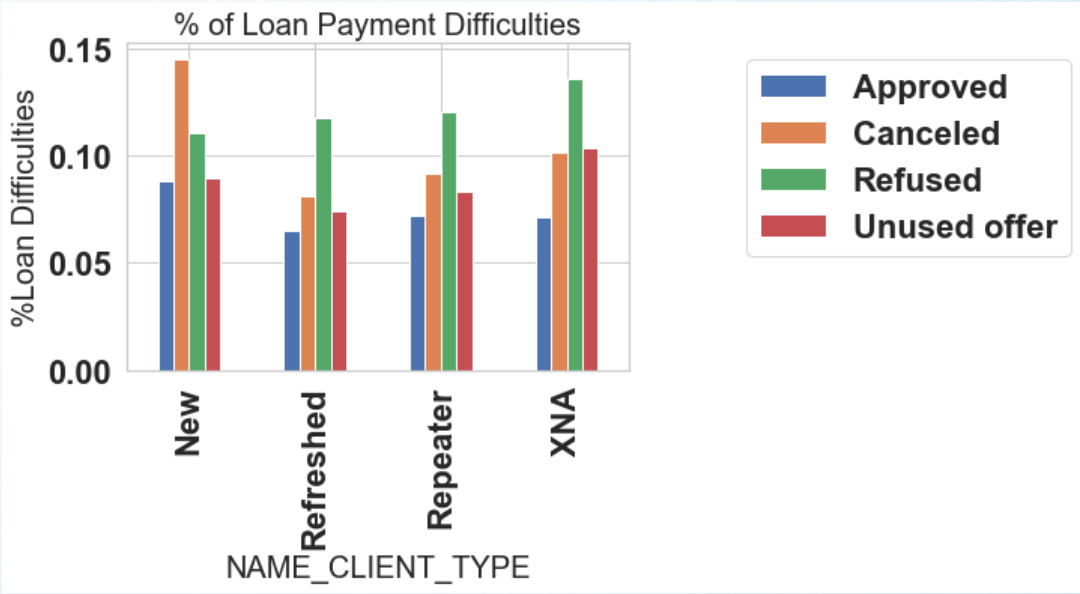
BIVARIATE ANALYSIS

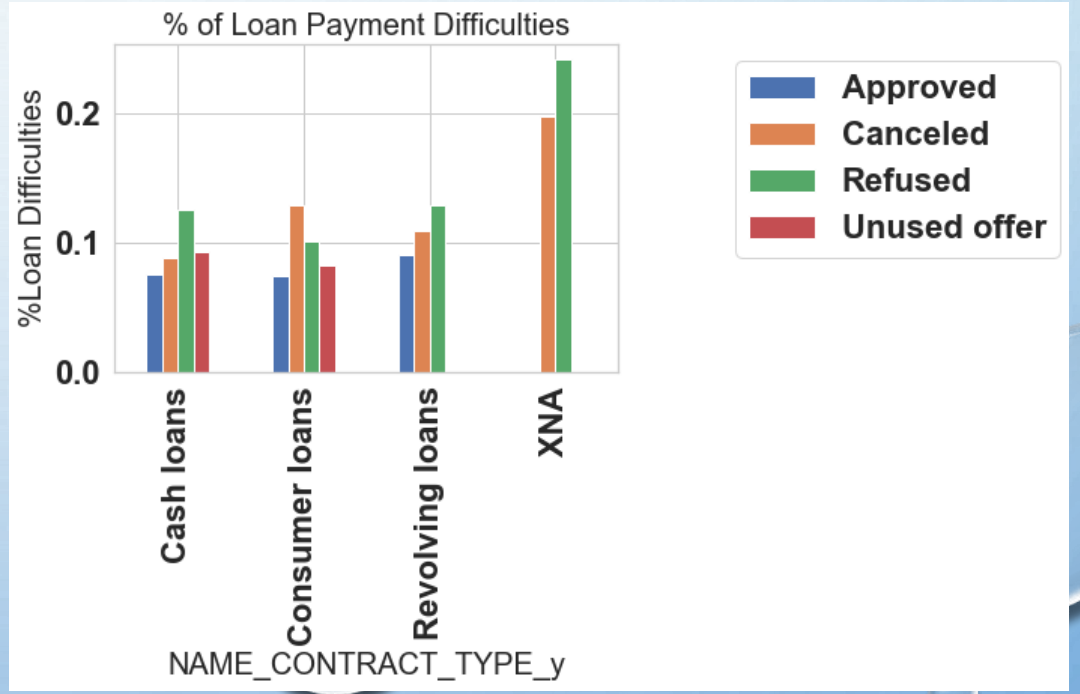
Inferences :

1) Purpose of the cash loan for repair and other has substantial applicant and in this category more applications are rejected
2) Also XNA and XAP which are the purpose of loan have the most application for



BIVARIATE ANALYSIS REPAYERS





Inferences :

1. Client who are "NEW and could secure loan application previously ('Cancelled')tend to be more % of Loan-Payment Difficulties
2. Clients with 'Revolving loans' and also 'Refused' status previously have more difficulties in % of Loan-Payment Difficulties. More over no conclusions can be drawn for XNA loan application which is unknown

CONCLUSION

FACTORS WHICH PREDICT THE CLIENT CAN BE DEFAULTER

1. LOWER SECONDARY' IN 'NAME_EDUCATION_TYPE' HAS MORE DIFFICULTY IN REPAYING LOANS
2. LOW SKILLED LABORERS' IN 'OCCUPATION_TYPE' ALSO HAS MAXIMUM DIFFICULTY IN LOAN REPAYMENT
3. 'MATERNITY LEAVE' IN 'NAME_INCOME_TYPE' ALSO HAS MAXIMUM DIFFICULTY IN LOAN REPAYMENT.
4. 0-5 YEARS OF EMPLOYMENT ARE MORE LIKELY TO DEFAULT LOAN
5. AMT_INCOME_TOTAL:HIGHER INCOME PEOPLE LESS LIKELY TO DEFAULTERS
6. CNT_CHILDREN: MORE NUMBER OF CHILDREN >3 ARE MORE LIKELY TO DEFAULTERS

CONCLUSION

FACTORS WHICH PREDICT THE CLIENT CAN BE A POTENTIAL REPAYER AND CAN BE HELPFUL FOR BUSINESS PROFIT

1. PEOPLE WHO ARE STUDENT AND BUSINESSMEN ARE MOST LIKELY TO REPAY LOANS
2. LOANS WHICH ARE PERSONAL IN NATURE LIKE HOBBY ARE LIKELY TO BE REPAY
3. PEOPLE WHO ARE TOWARDS THEIR RETIREMENT I.E. AGE > 50 HAVE GOOD TRACK RECORD OF REPAYMENT
4. WOMEN HAVE HIGHER CHANCES OF REPAYMENT OF LOAN
5. MARRIAGE ALSO IS DRIVING FACTOR WHILE DECIDING LOAN REPAYMENT I.E. APPLICANT WHO ARE SINGLE AND CIVIL MARRAIGE ARE MORE LIKELY TO REPAY LOAN
6. IF PRICE OF THE GOODS FOR WHICH THE LOAN GIVEN IS LESS THAN ~3 LAKH THERE ARE HIGHER CHANCES GETTING REPAYMENT

CONCLUSION

FACTORS WHICH CAN BE HELPFUL TO OFFER THE LOAN BUT REDUCE ITS AMOUNT & AT HIGHER INTEREST RATES

1. BASED ON NUMBER OF FAMILY MEMBERS (I.E. IF NUMBER IS LARGE) REDUCE AMOUNT OF LOAN THAN ELIGIBLE (SAY SOME % OF TAKE HOME SALARY) CAN BE OFFERED AT HIGHER INTEREST RATES
2. PEOPLE WHO HAVE GOODS PRICE AND LESS INCOME CAN BE POTENTIAL CUSTOMER FOR REDUCE LOAN AND HIGHER INTEREST RATES
3. IF CLIENT REFUSED TO MENTION THE PURPOSE OF LOAN CAN BE TARGETED WITH REDUCED LOAN AMOUNT AND HIGHER INTEREST RATE
4. CLIENT WHO HAVE REFUSED LOAN PREVIOUSLY AND REFUSED PURPOSE CAN BE TARGETED WITH EXTRA SECURITY ASSURANCE TO INCREASE THE BUSINESS

THANK YOU