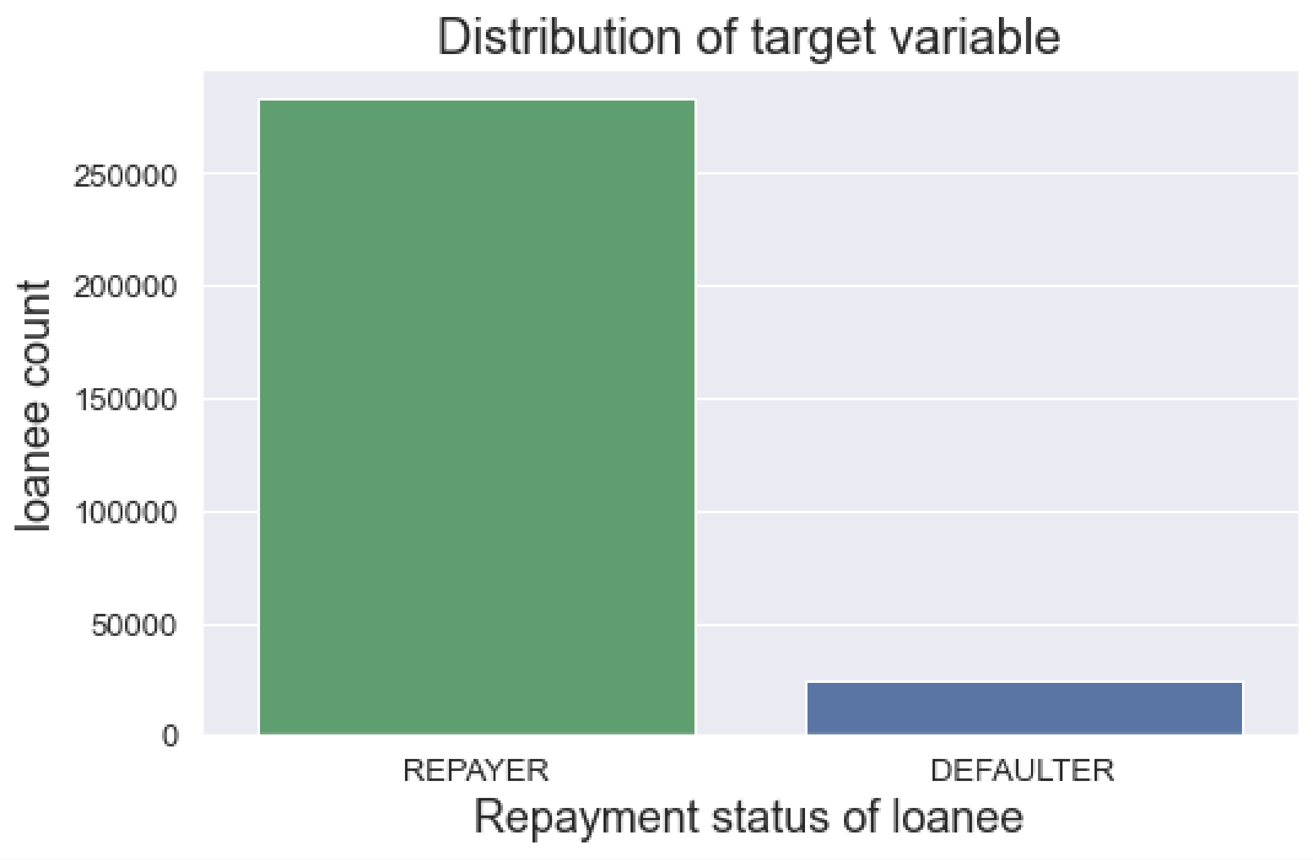
CREDIT EDA CASE STUDY

By

Sindhuja Macharla and Syed Murtuza Ali

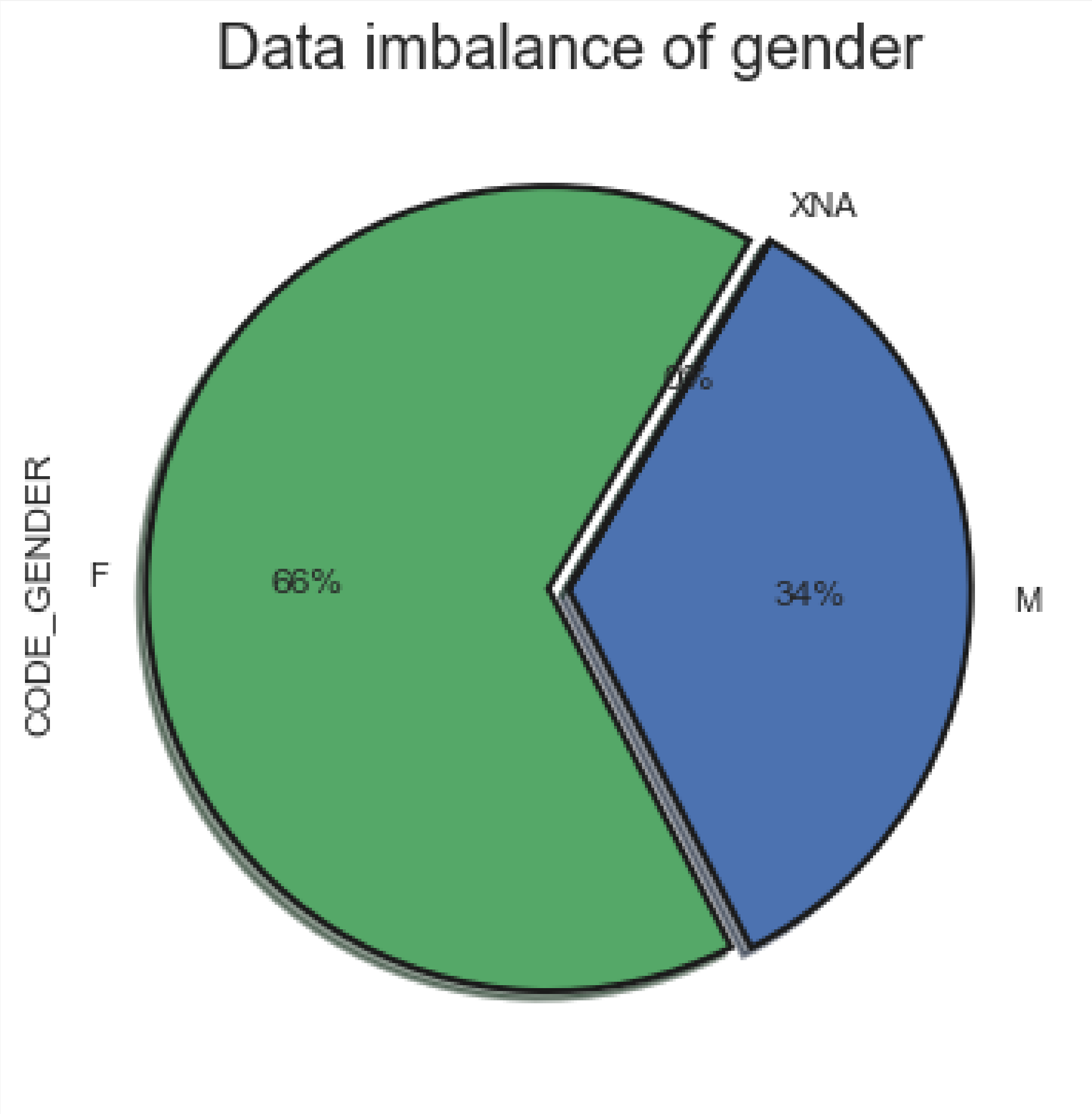
BUSINESS OBJECTIVES

- This case study aims to identify patterns which indicate if a client has difficulty paying their installments 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.



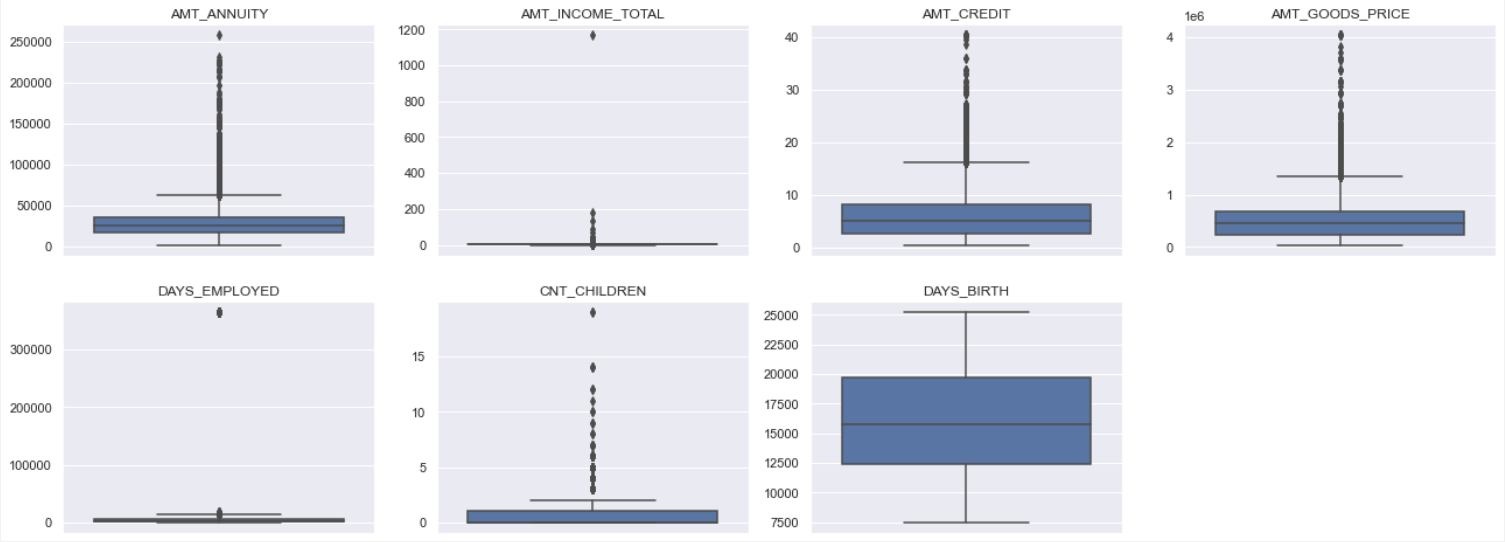
DISTRIBUTION OF TARGET VARIABLE

- The plot shows the distribution of target variable from the application data set.
- The repayors are comparatively higher than the defaulters.



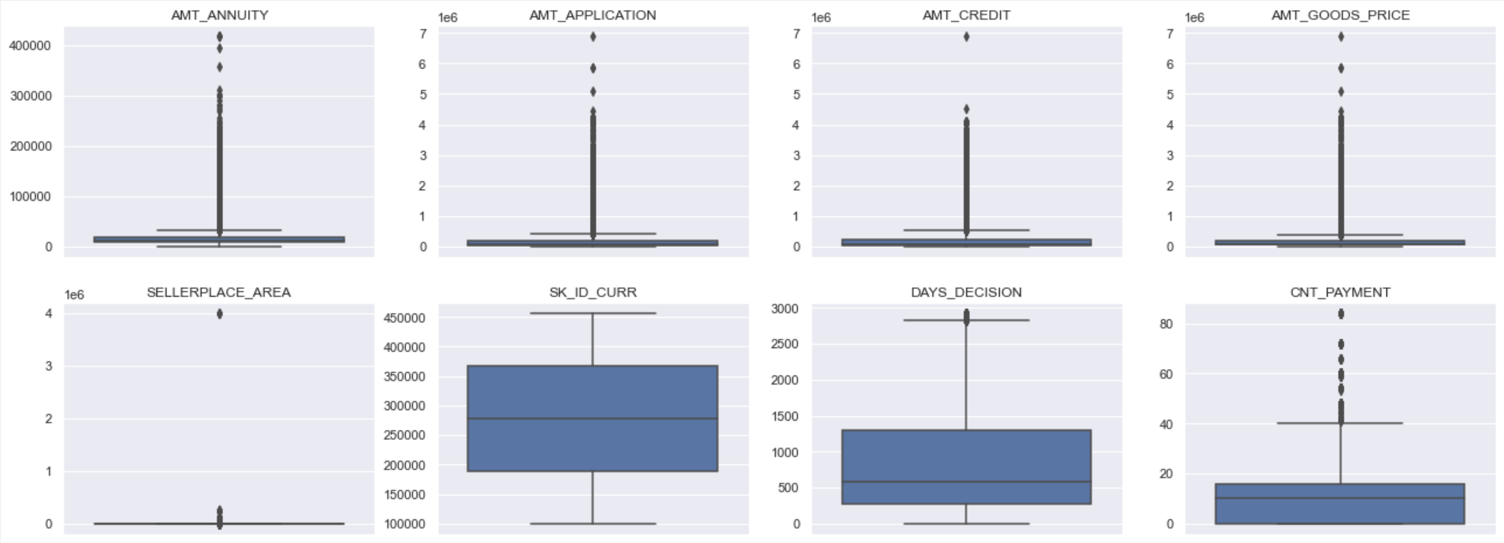
DATA IMBALANCE PERCENTAGE

- The pie chart clearly states that the percentage of female applicants are greater than that of male applicants.
- It clearly shows that there is data imbalance in CODE_GENDER variable.



APPLICATION DATA OUTLIERS

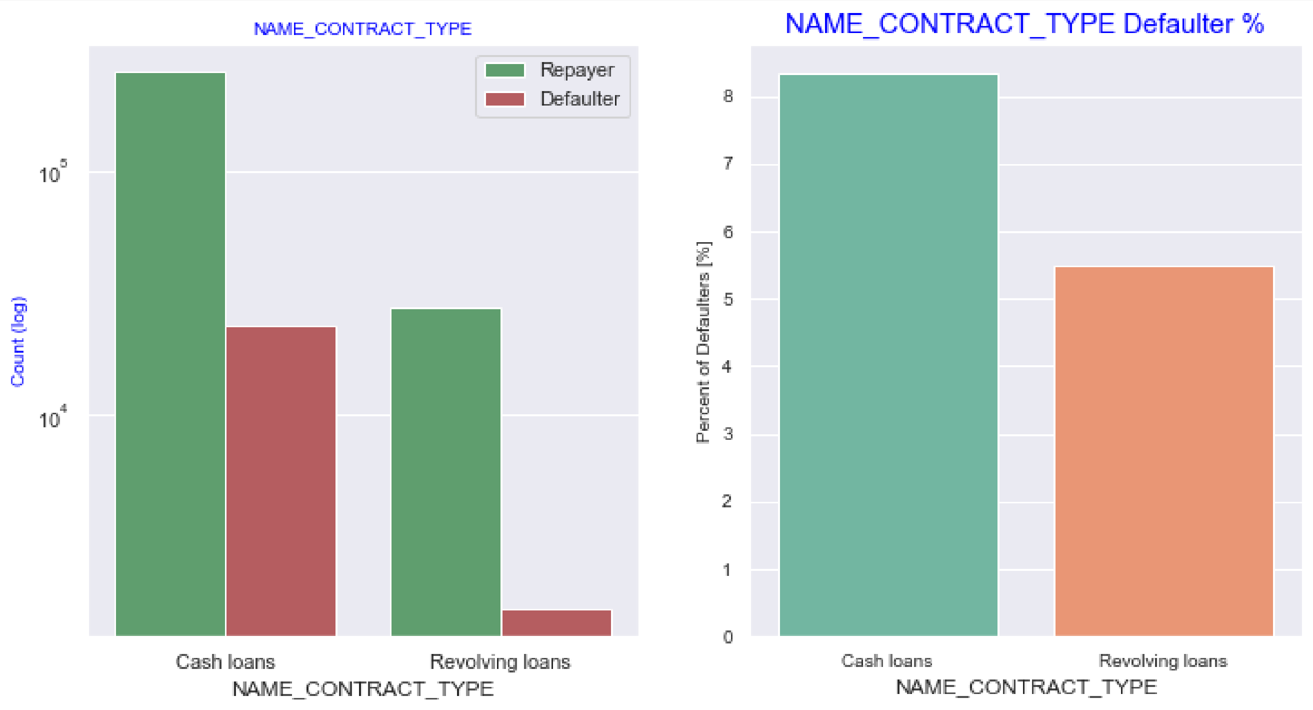
- AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.
- AMT_INCOME_TOTAL has huge number of outliers which indicate that few of the loan applicants have high income when compared to the others.
- DAYS_BIRTH has no outliers which means the data available is reliable. DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and hence this has to be incorrect entry.



PREVIOUS DATA OUTLIERS

- AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.
- CNT_PAYMENT has few outlier values. SK_ID_CURR is an ID column and hence no outliers.
- DAYS_DECISION has little number of outliers indicating that these previous applications decisions were taken long back.

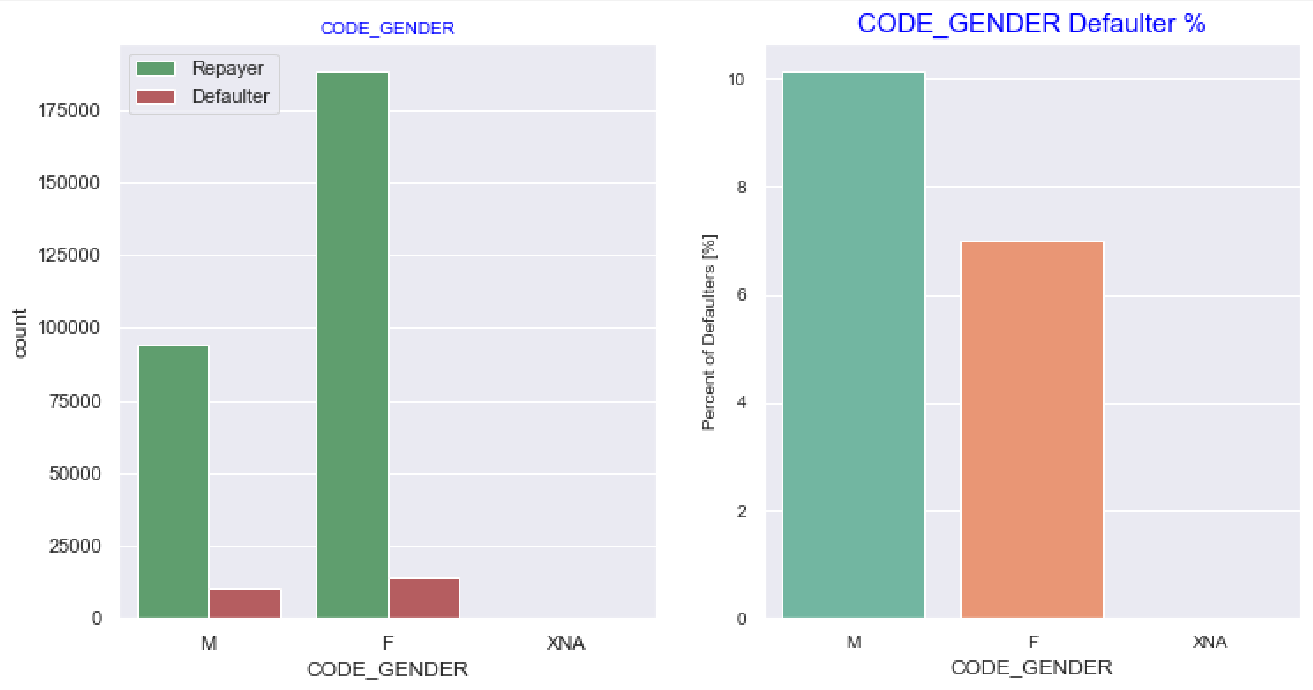
UNIVARIATE CATEGORICAL ANALYSIS



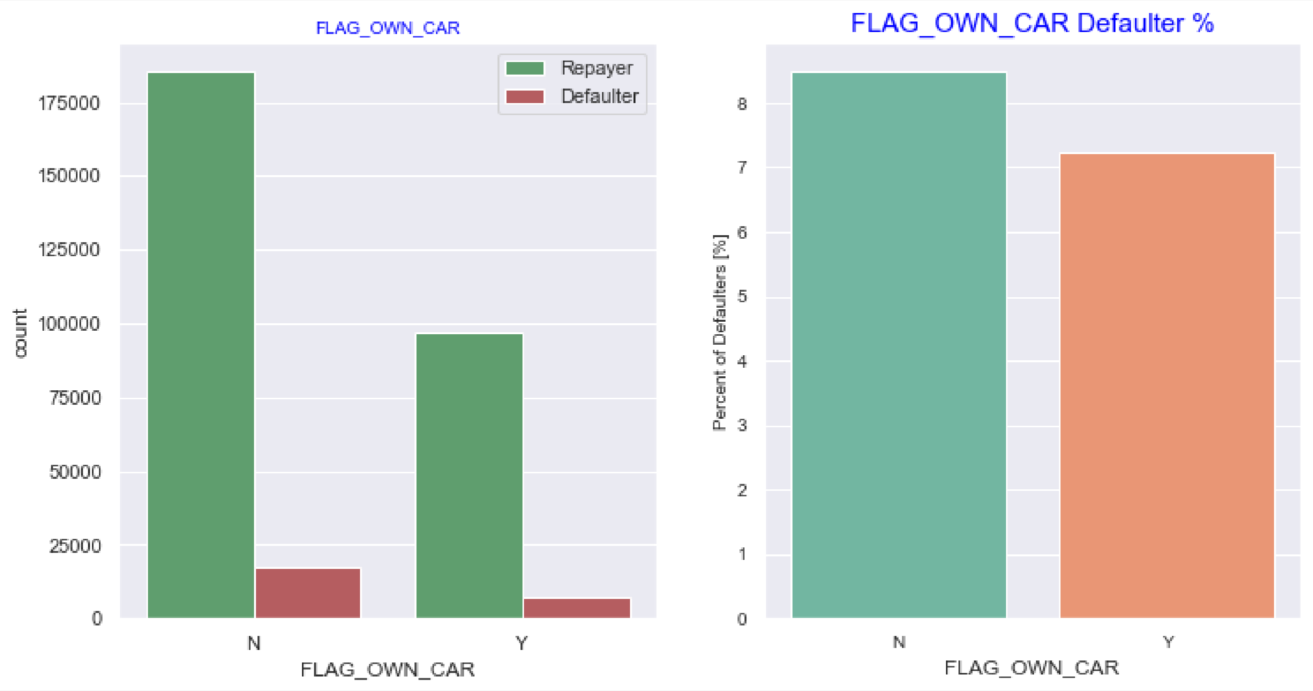
CONTRACT TYPE BASED ON LOAN REPAYMENT STATUS

- Contract type: Revolving loans are just a small fraction (10%) from the total number of loans; in the same time, a larger amount of Revolving loans, comparing with their frequency, are not repaid.

TYPE OF GENDER ON LOAN REPAYMENT STATUS



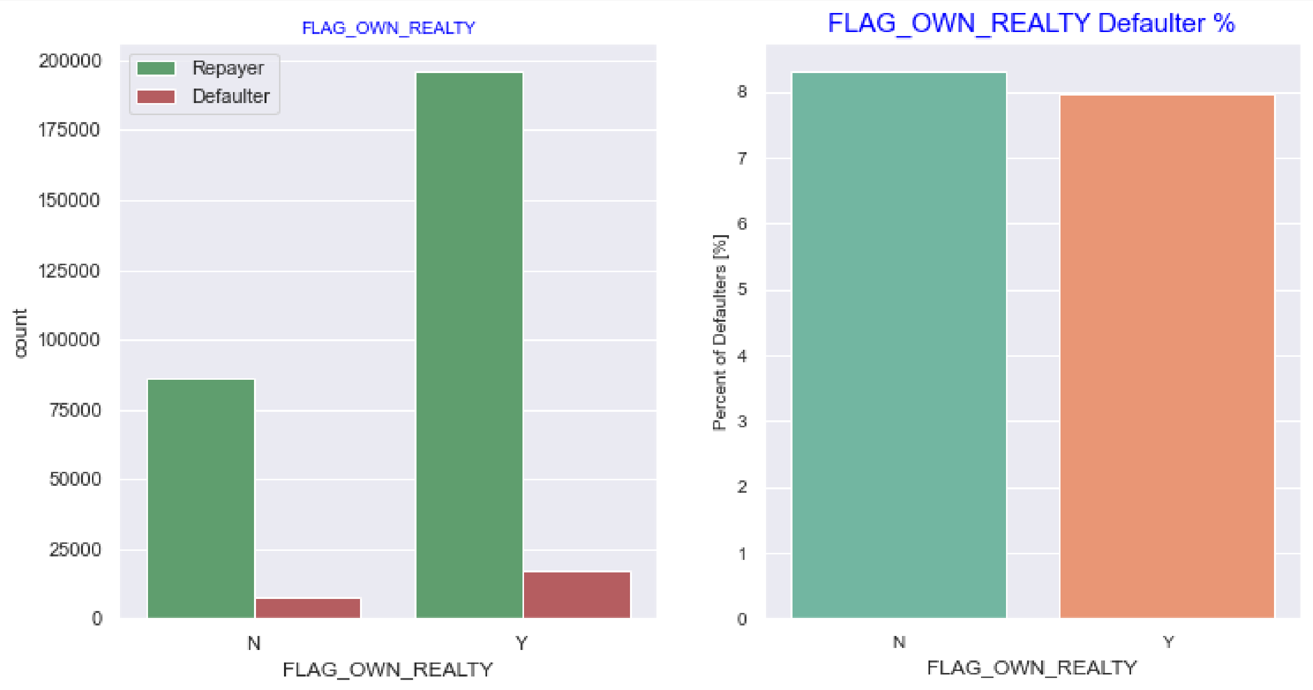
- The number of female clients is almost double the number of male clients. Based on the percentage of defaulted credits, males have a higher chance of not returning their loans (~10%), comparing with women (~7%)



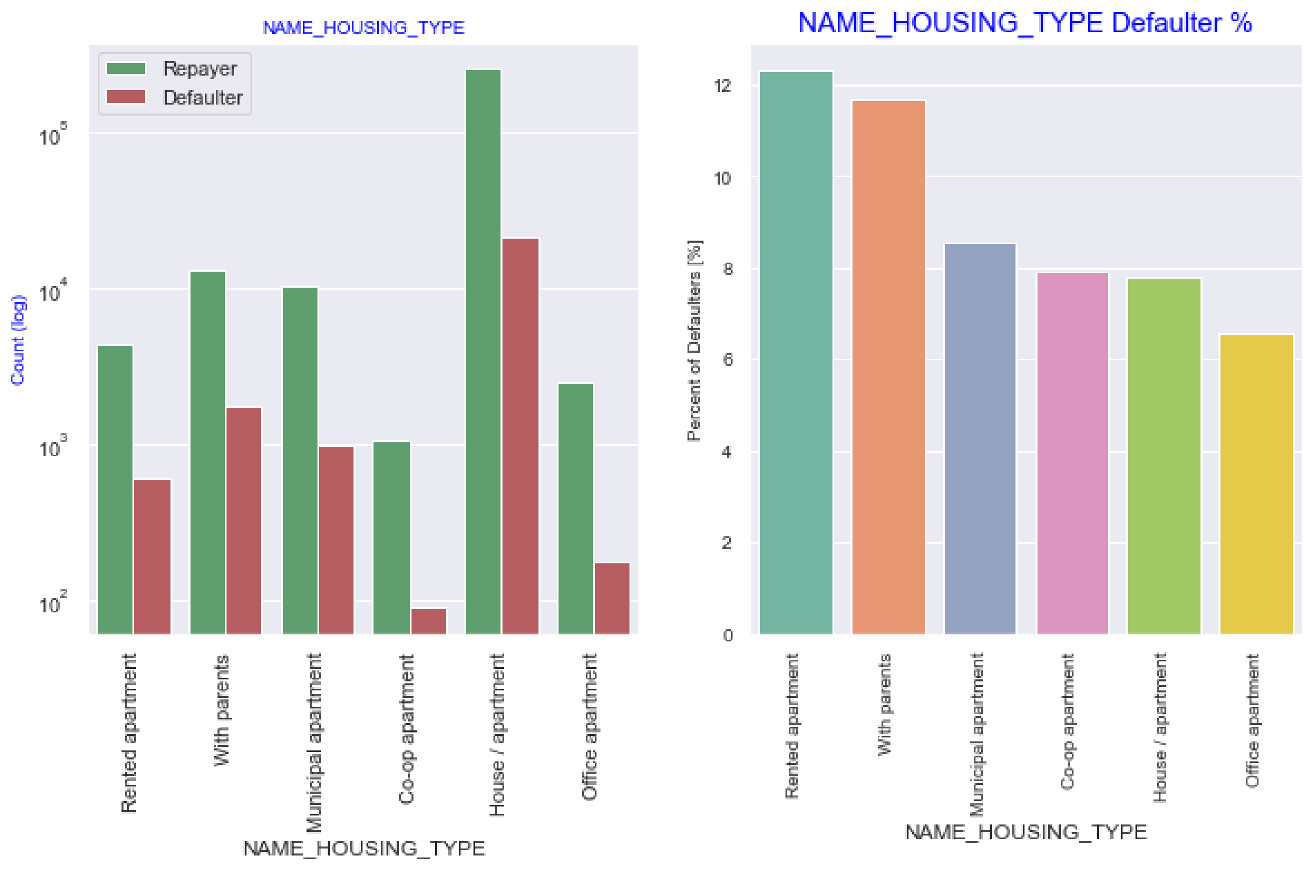
OWNING A CAR AND ITS EFFECTS ON LOAN REPAYMENT STATUS

- Clients who own a car are half in number of the clients who don't own a car. But based on the percentage of default, there is no correlation between owning a car and loan repayment as in both cases the default percentage is almost same.

OWNING A REALTY AND ITS EFFECTS ON LOAN REPAYMENT STATUS

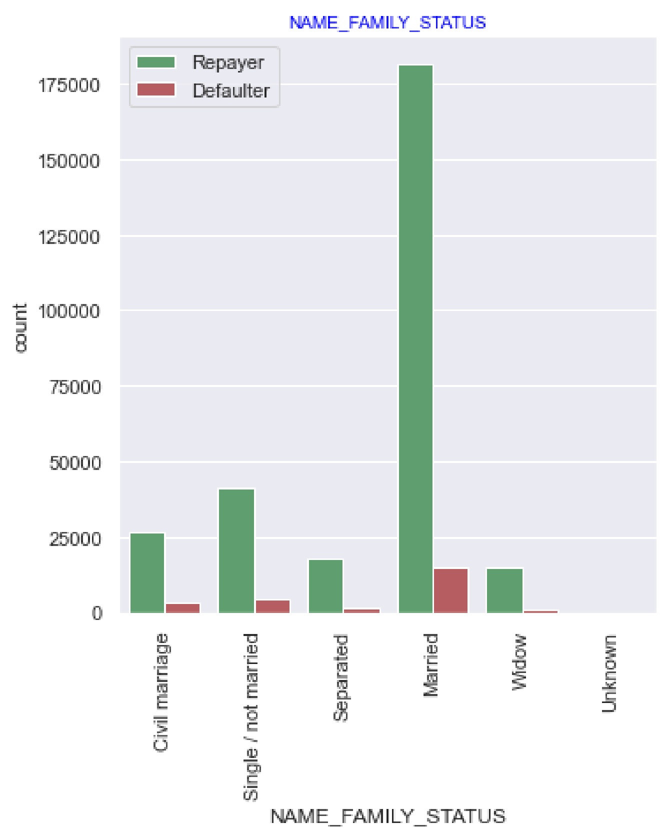


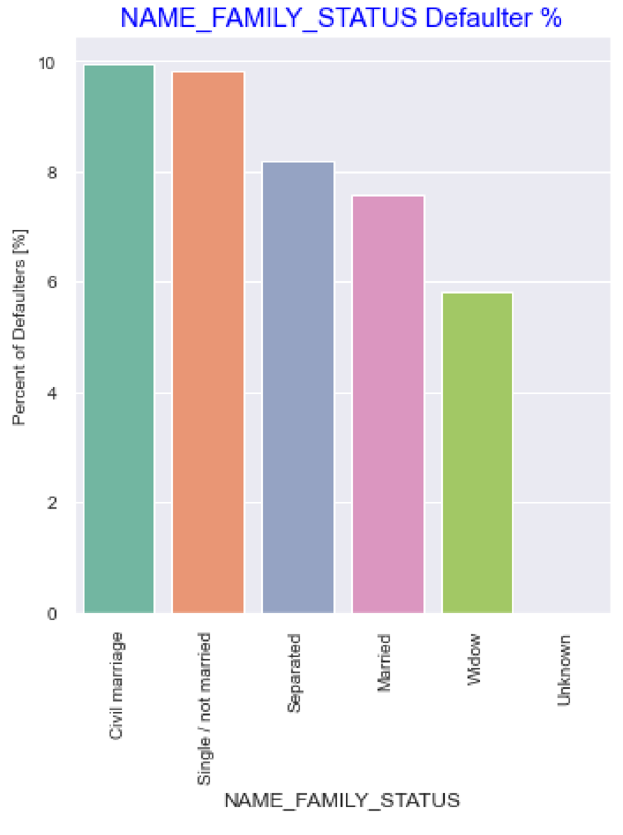
- The clients who own real estate are more than double of the ones that don't own. But the defaulting rate of both categories are around the same (~8%). Thus there is no correlation between owning a reality and defaulting the loan.



HOUSING TYPE BASED ON LOAN REPAYMENT STATUS

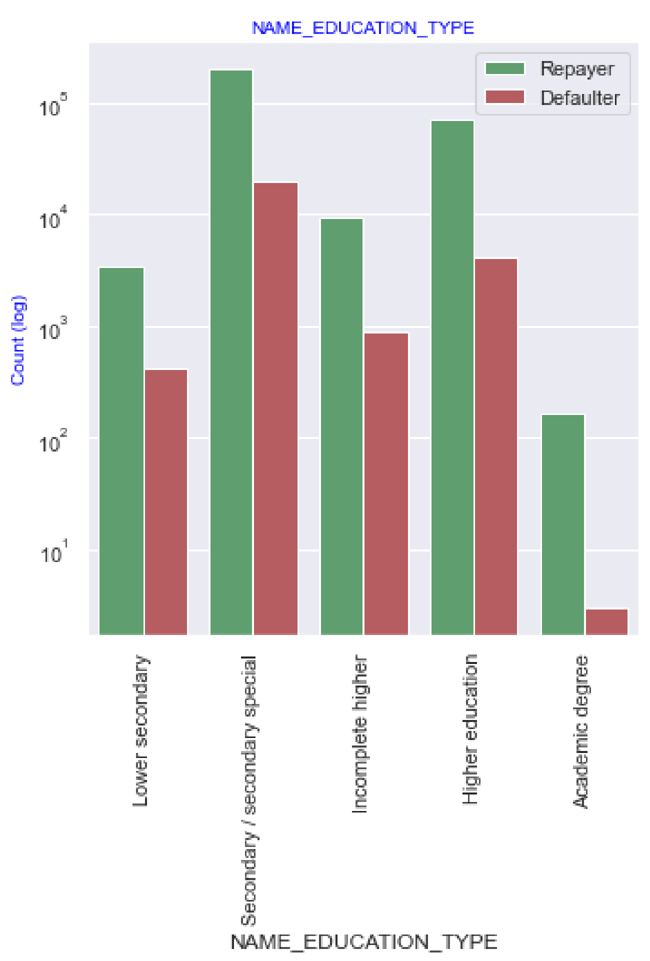
- Majority of people live in House/apartment People living in office apartments have lowest default rate People living with parents (~11.5%) and living in rented apartments(>12%) have higher probability of defaulting

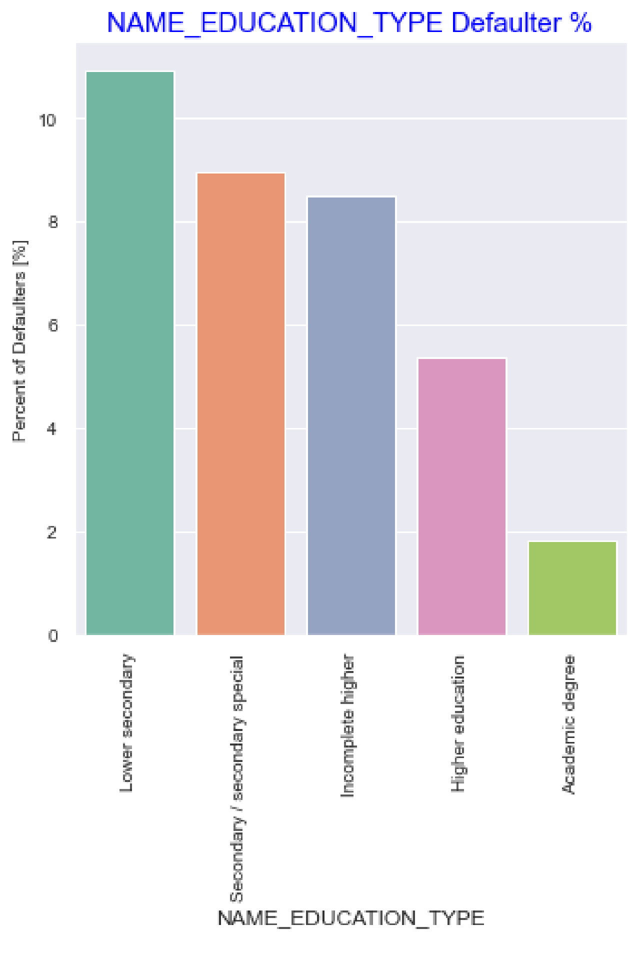




FAMILY STATUS BASED ON LOAN REPAYMENT STATUS

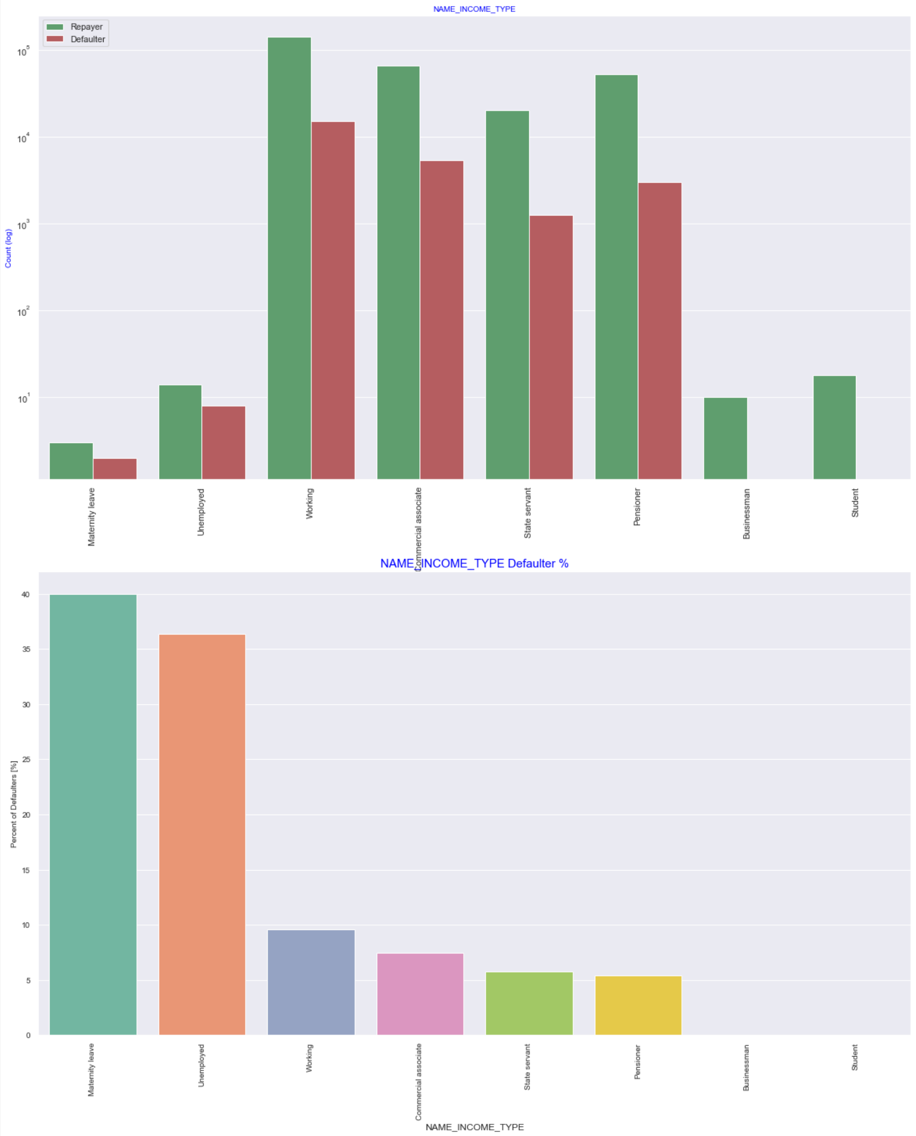
- Most of the people who have taken loan are married, followed by Single/not married and civil marriage In terms of percentage of not repayment of loan, Civil marriage has the highest percent of not repayment (10%), with Widow the lowest (exception being Unknown).





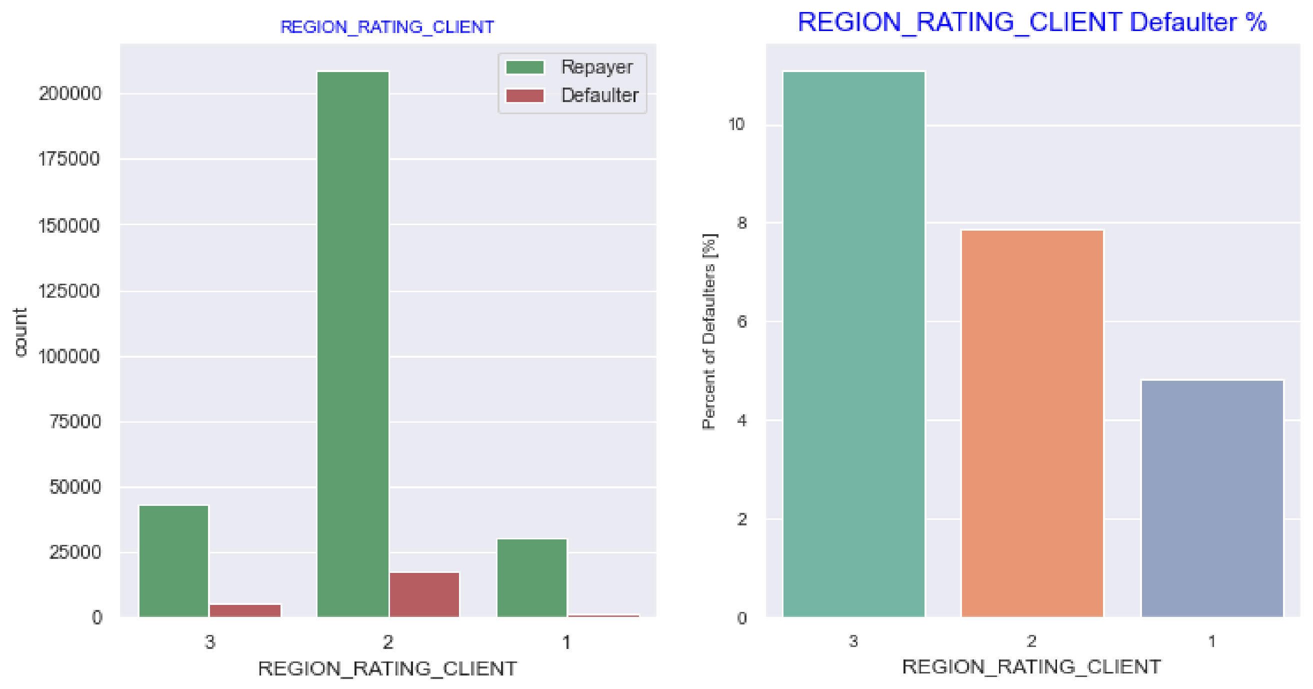
EDUCATION TYPE BASED ON LOAN REPAYMENT STATUS

- Majority of the clients have Secondary / secondary special education, followed by clients with Higher education. Only a very small number having an academic degree The Lower secondary category, although rare, have the largest rate of not returning the loan (11%). The people with Academic degree have less than 2% defaulting rate.



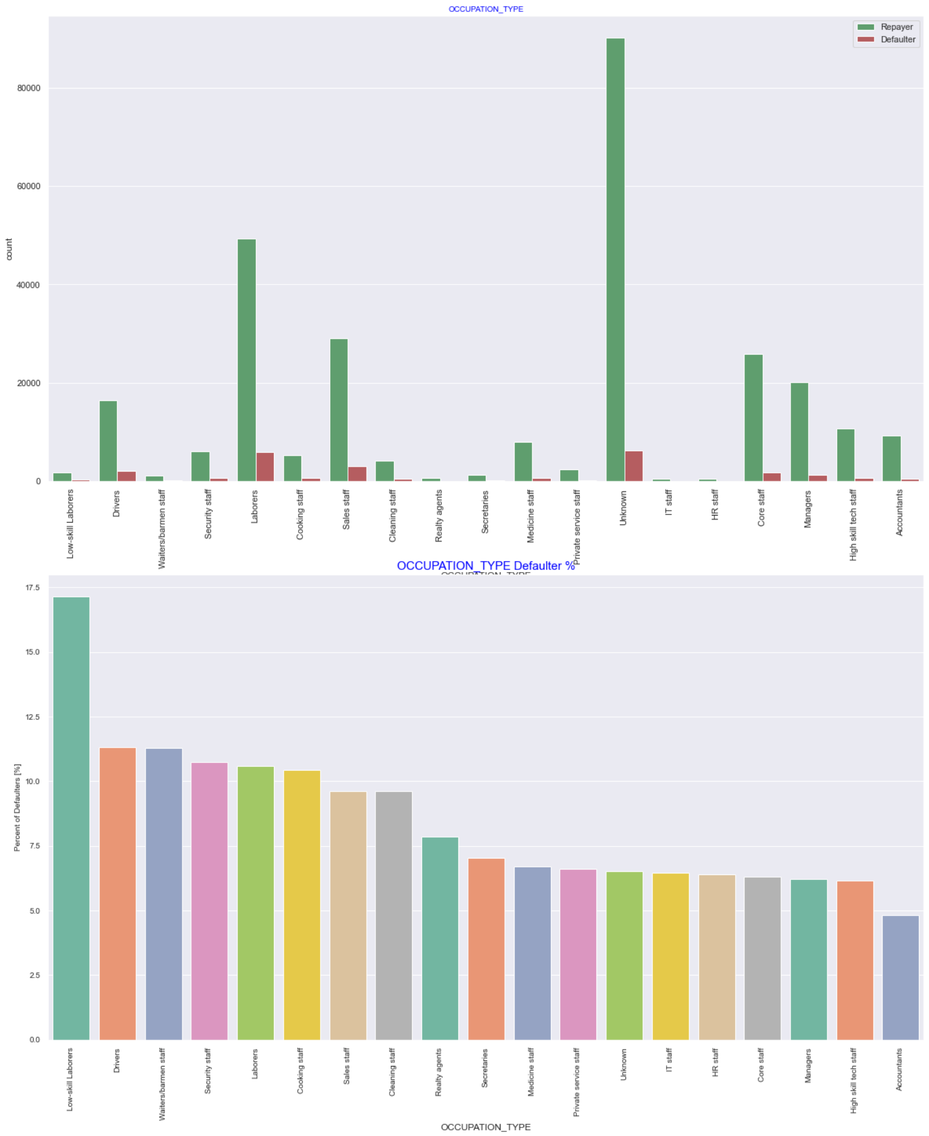
INCOME TYPE BASED ON LOAN REPAYMENT STATUS

- Most of applicants for loans have income type as Working, followed by Commercial associate, Pensioner and State servant. The applicants with the type of income Maternity leave have almost 40% ratio of not returning loans, followed by Unemployed (37%). The rest of types of incomes are under the average of 10% for not returning loans. Student and Businessmen, though less in numbers do not have any default record. Thus these two category are safest for providing loan.



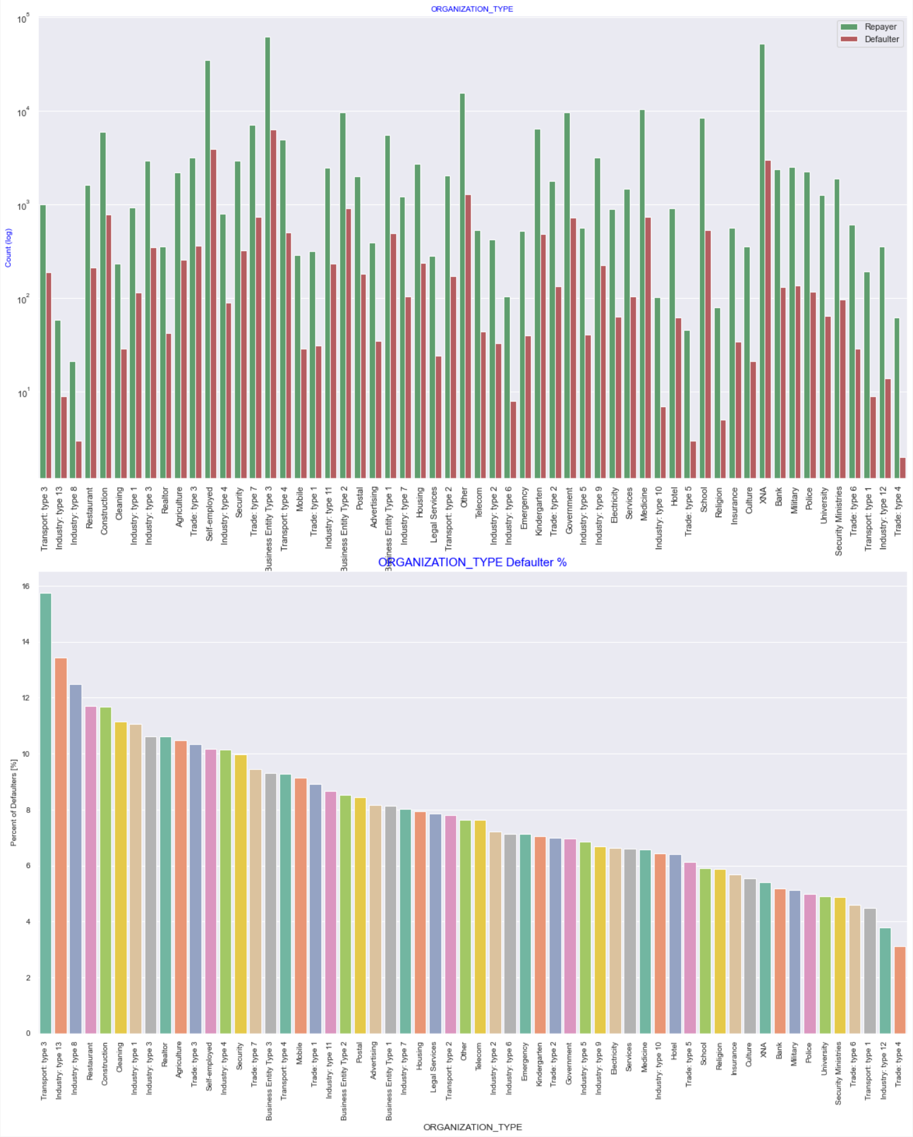
REGION RATING WHERE APPLICANT LIVES BASED ON LOAN REPAYMENT STATUS

- Most of the applicants are living in Region_Rating 2 place. Region Rating 3 has the highest default rate (11%) Applicant living in Region_Rating 1 has the lowest probability of defaulting, thus safer for approving loans



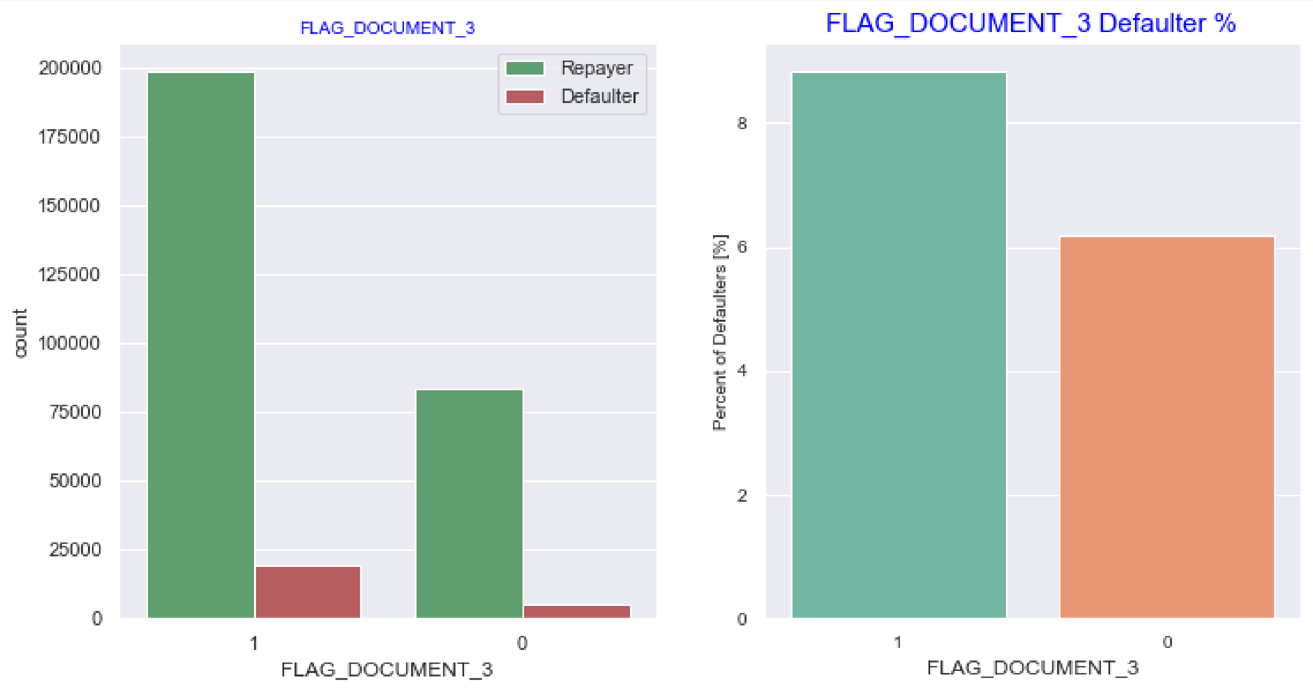
OCCUPATION TYPE WHERE APPLICANT LIVES BASED ON LOAN REPAYMENT STATUS

- Most of the loans are taken by Laborers, followed by Sales staff. IT staff take the lowest amount of loans. The category with highest percent of not repaid loans are Low-skill Laborers (above 17%), followed by Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff.



LOAN REPAYMENT STATUS BASED ON ORGANIZATION TYPE

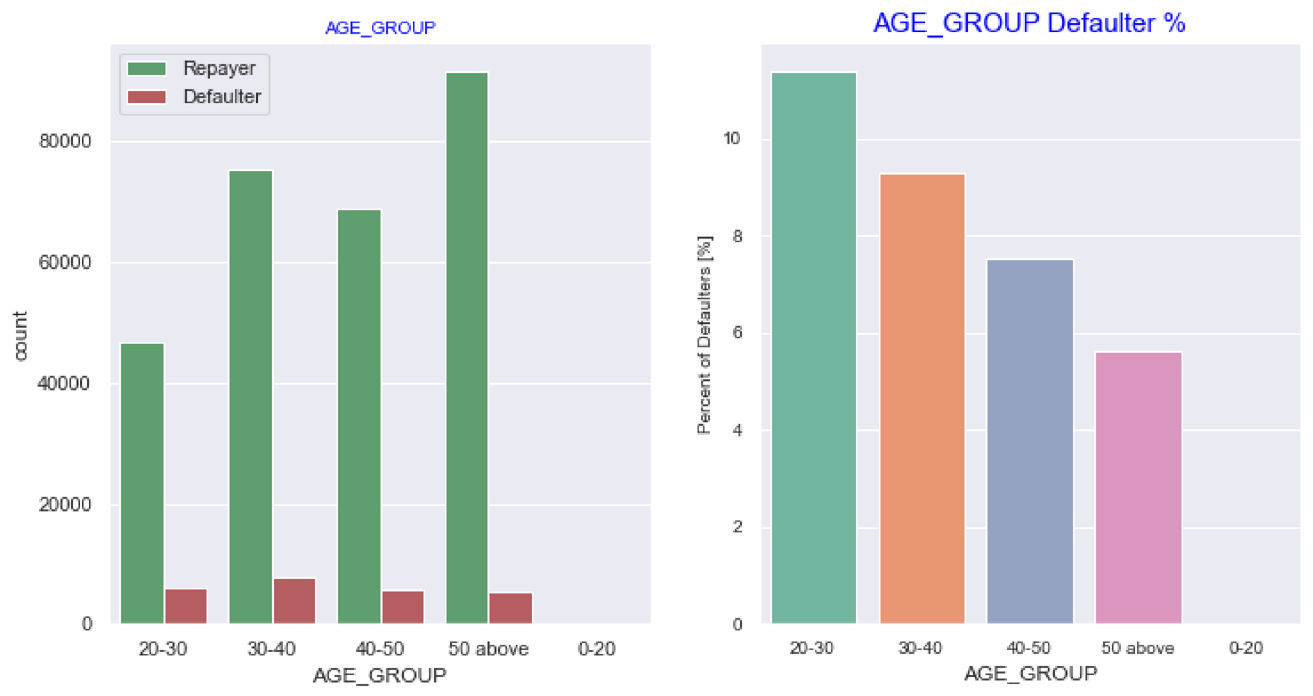
- Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self employed people have relative high defaulting rate, and thus should be avoided to be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting. Most of the people application for loan are from Business Entity Type 3 For a very high number of applications, Organization type information is u



FLAG_DOC_3 SUBMISSION STATUS BASED ON LOAN REPAYMENT STATUS

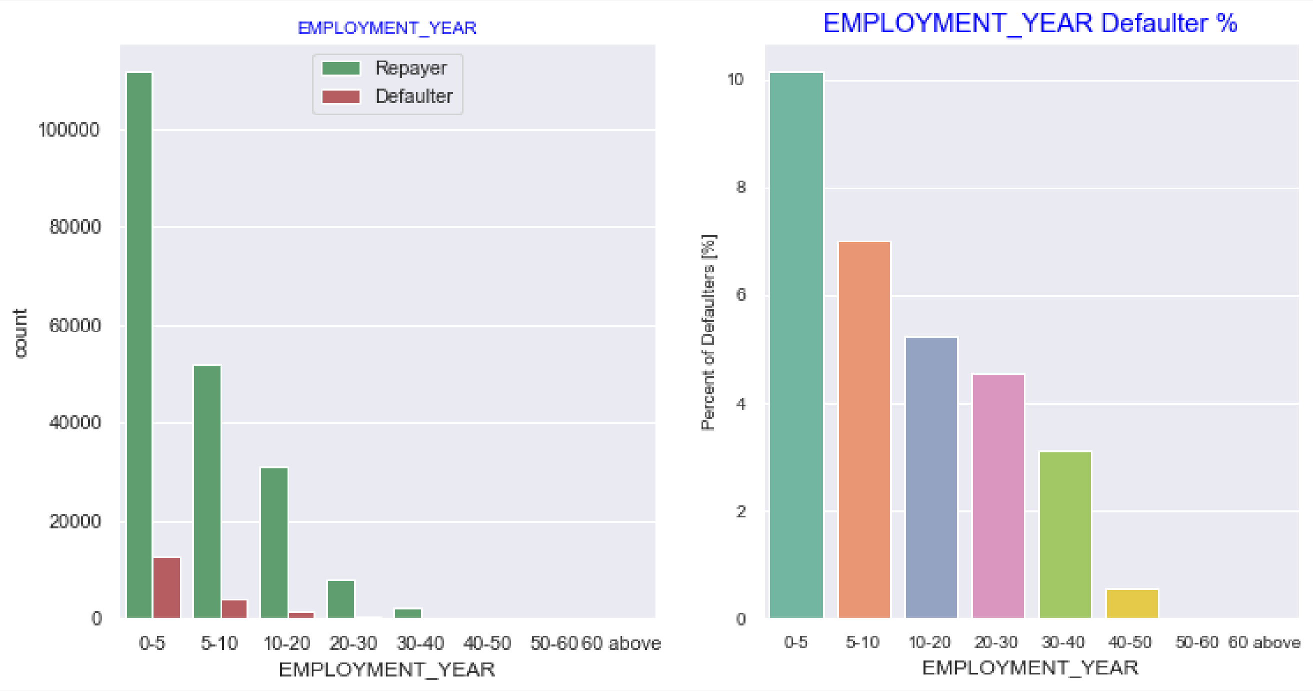
- There is no significant correlation between repayors and defaulters in terms of submitting document 3 as we see even if applicants have submitted the document, they have defaulted a slightly more (~9%) than who have not submitted the document (6%)

AGE GROUP BASED ON LOAN REPAYMENT STATUS



- People in the age group range 20-40 have higher probability of defaulting People above age of 50 have low probability of defaulting

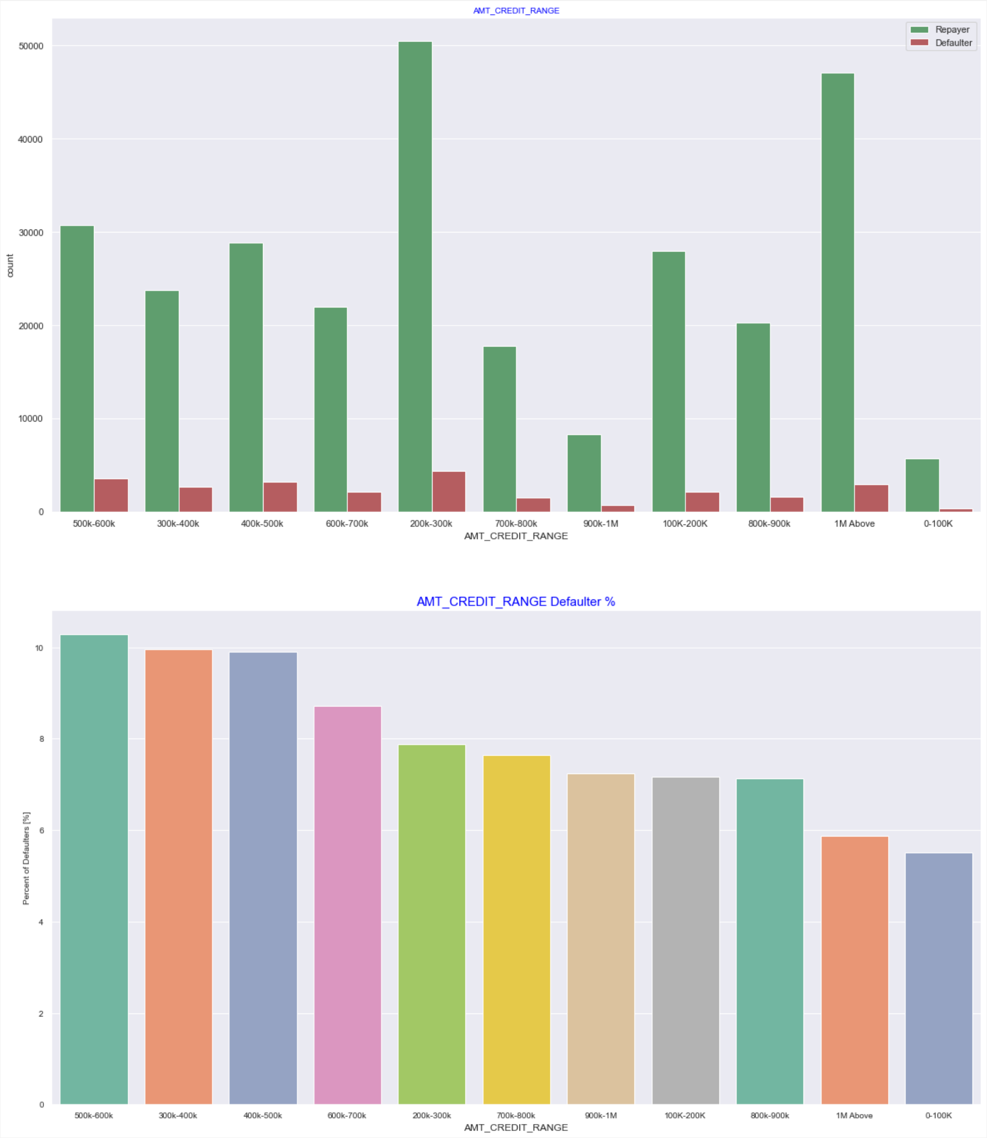
EMPLOYMENT_YEAR BASED ON LOAN REPAYMENT STATUS



EMPLOYMENT_YEAR

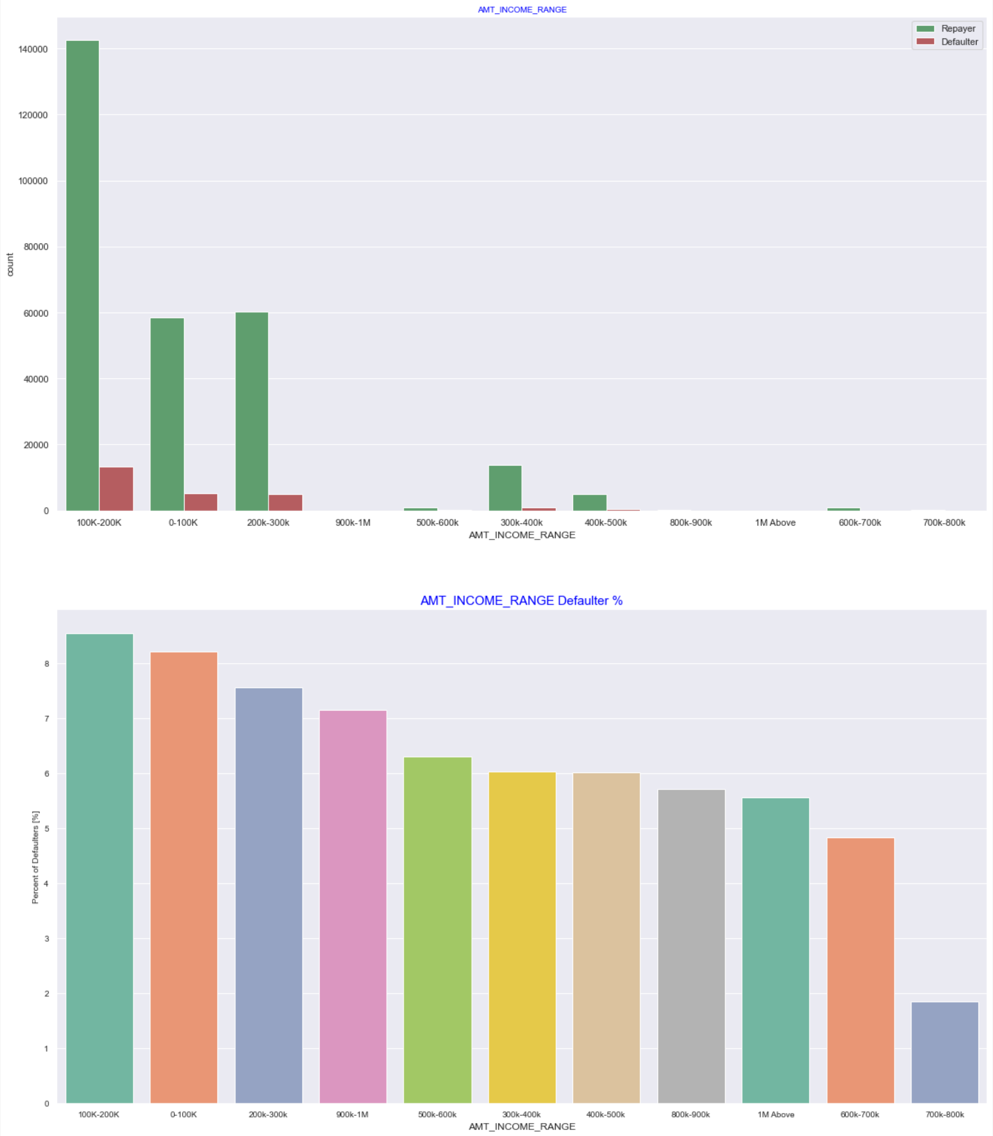
EMPLOYMENT_YEAR Defaulter %

- Majority of the applicants have been employed in between 0-5 years. The defaulting rating of this group is also the highest which is 10% With increase of employment year, defaulting rate is gradually decreasing with people having 40+ year experience having less than 1% default rate



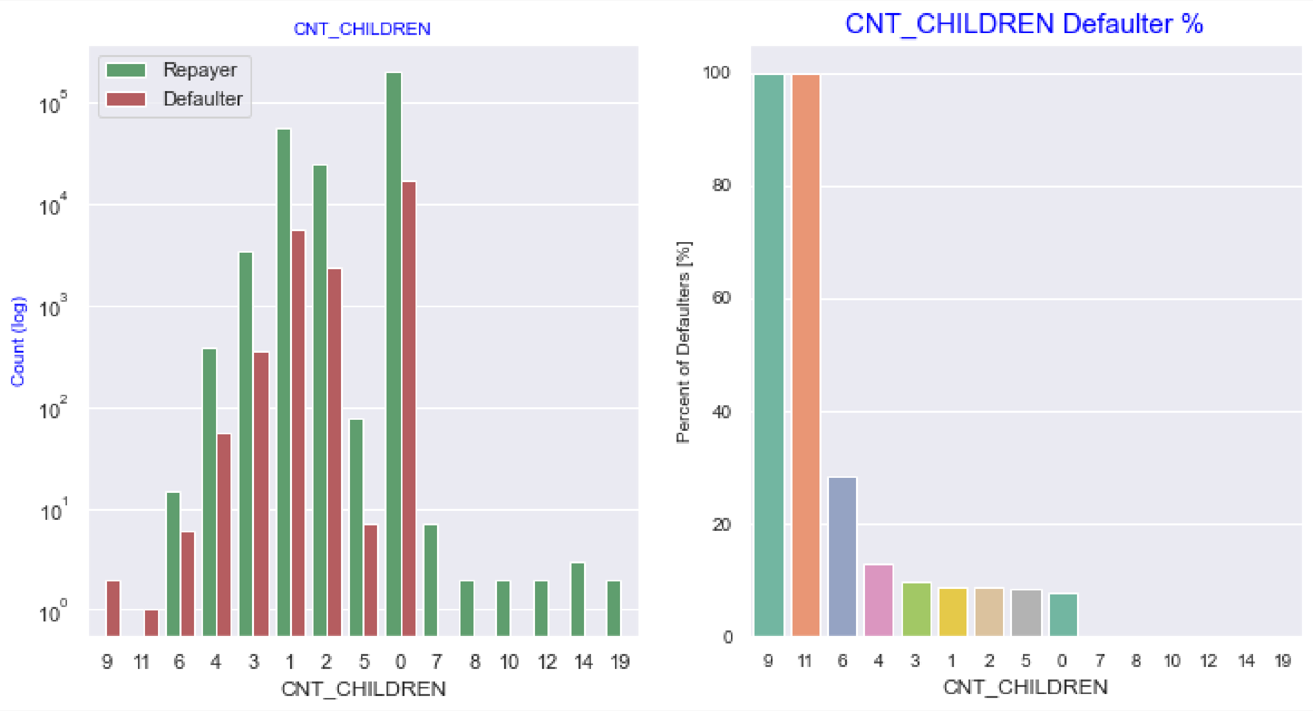
AMOUNT_CREDIT BASED ON LOAN REPAYMENT STATUS

- More than 80% of the loan provided are for amount less than 900,000 People who get loan for 300-600k tend to default more than others.



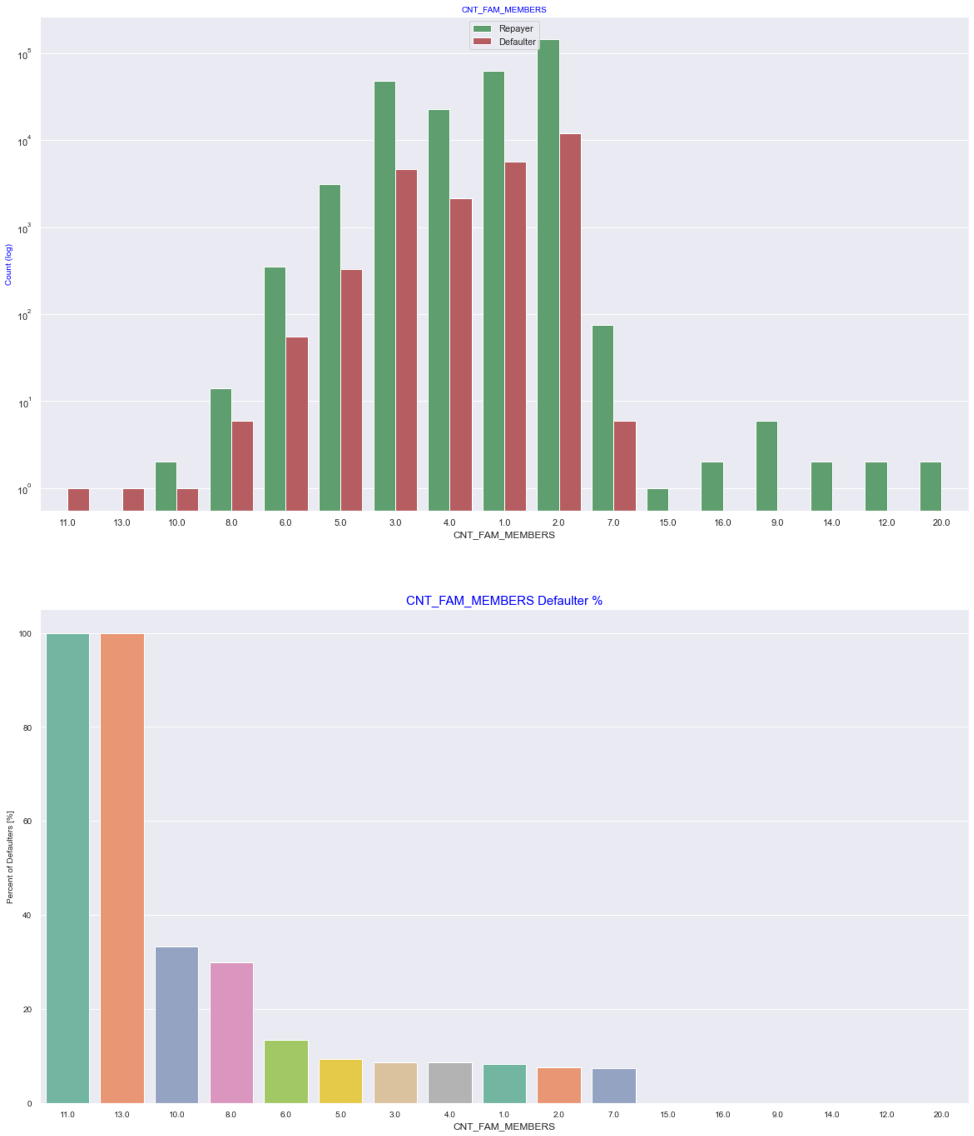
AMOUNT_INCOME RANGE BASED ON LOAN REPAYMENT STATUS

- 90% of the applications have Income total less than 300,000 Application with Income less than 300,000 has high probability of defaulting Applicant with Income more than 700,000 are less likely to default



NUMBER OF CHILDREN BASED ON LOAN REPAYMENT STATUS

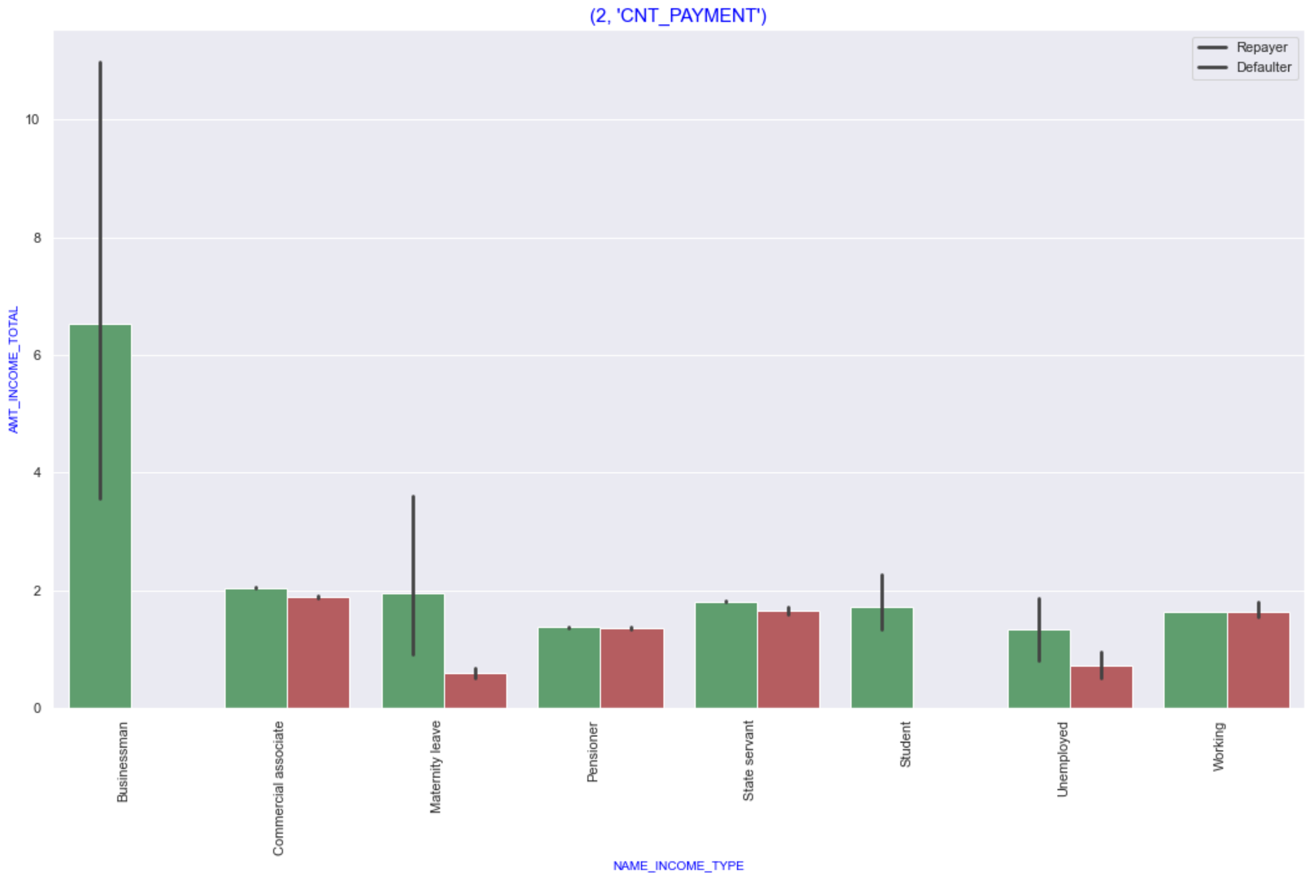
- Most of the applicants do not have children Very few clients have more than 3 children. Client who have more than 4 children has a very high default rate with child count 9 and 11 showing 100% default rate



NUMBER OF FAMILY MEMBERS BASED ON LOAN REPAYMENT STATUS

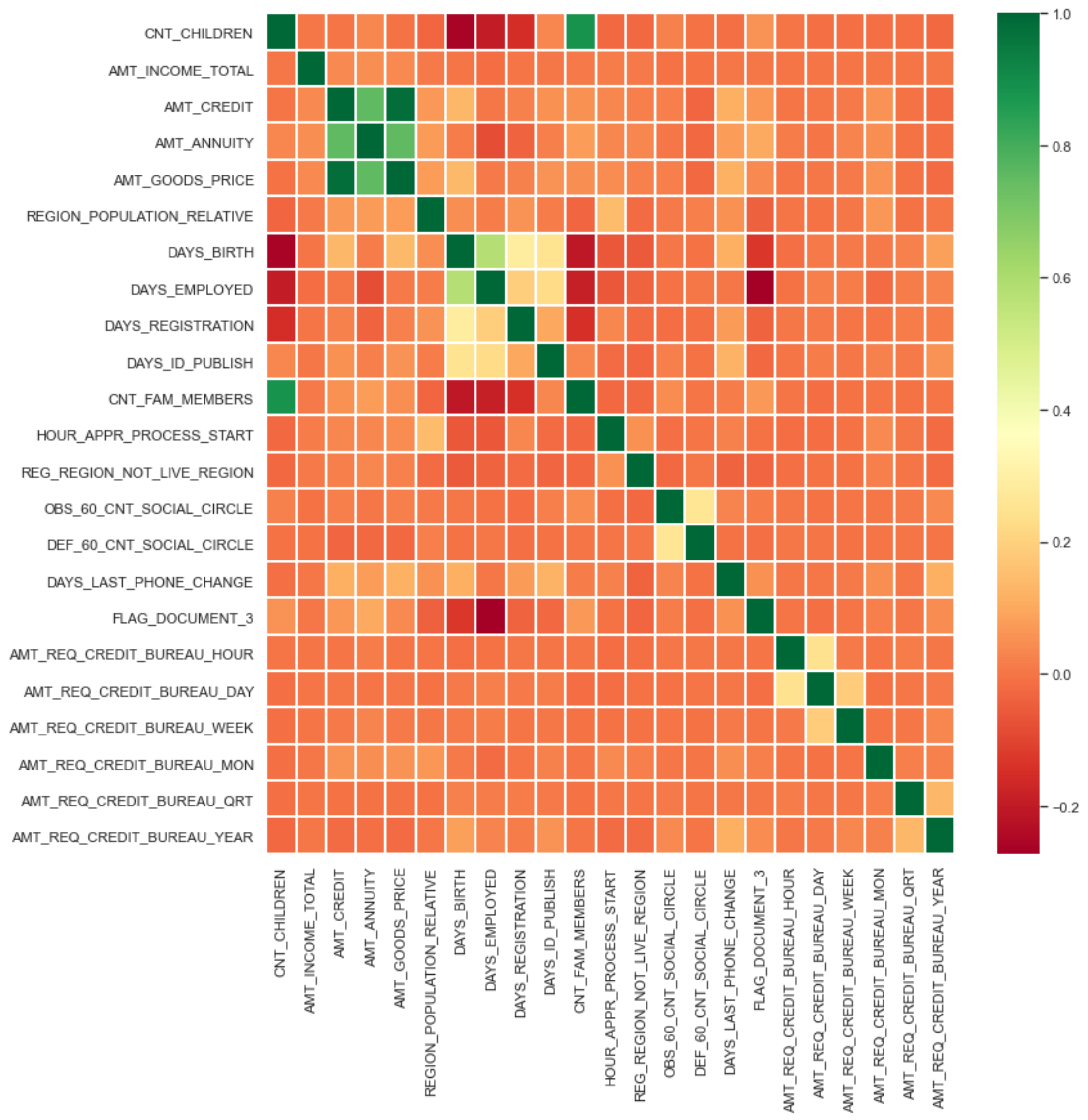
- Family member follows the same trend as children where having more family members increases the risk of defaulting

CATEGORICAL BIVARIATE ANALYSIS



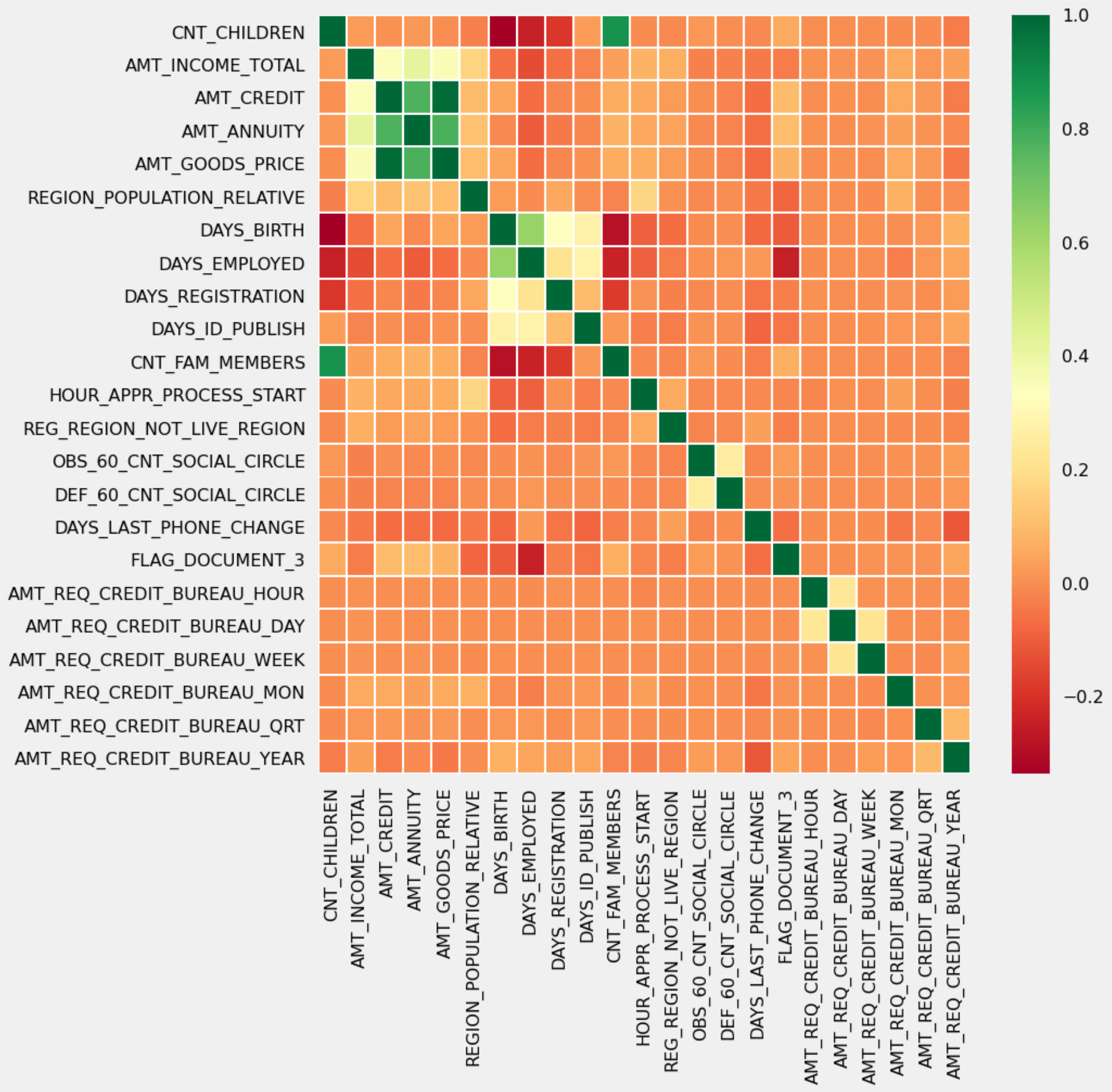
INCOME TYPE VS INCOME AMOUNT RANGE

- It can be seen that business man's income is the highest and the estimated range with default 95% confidence level seem to indicate that the income of a business man could be in the range of slightly close to 4 lakhs and slightly above 10 lakhs



CORRELATION FOR THE DEFAULTERS DATA

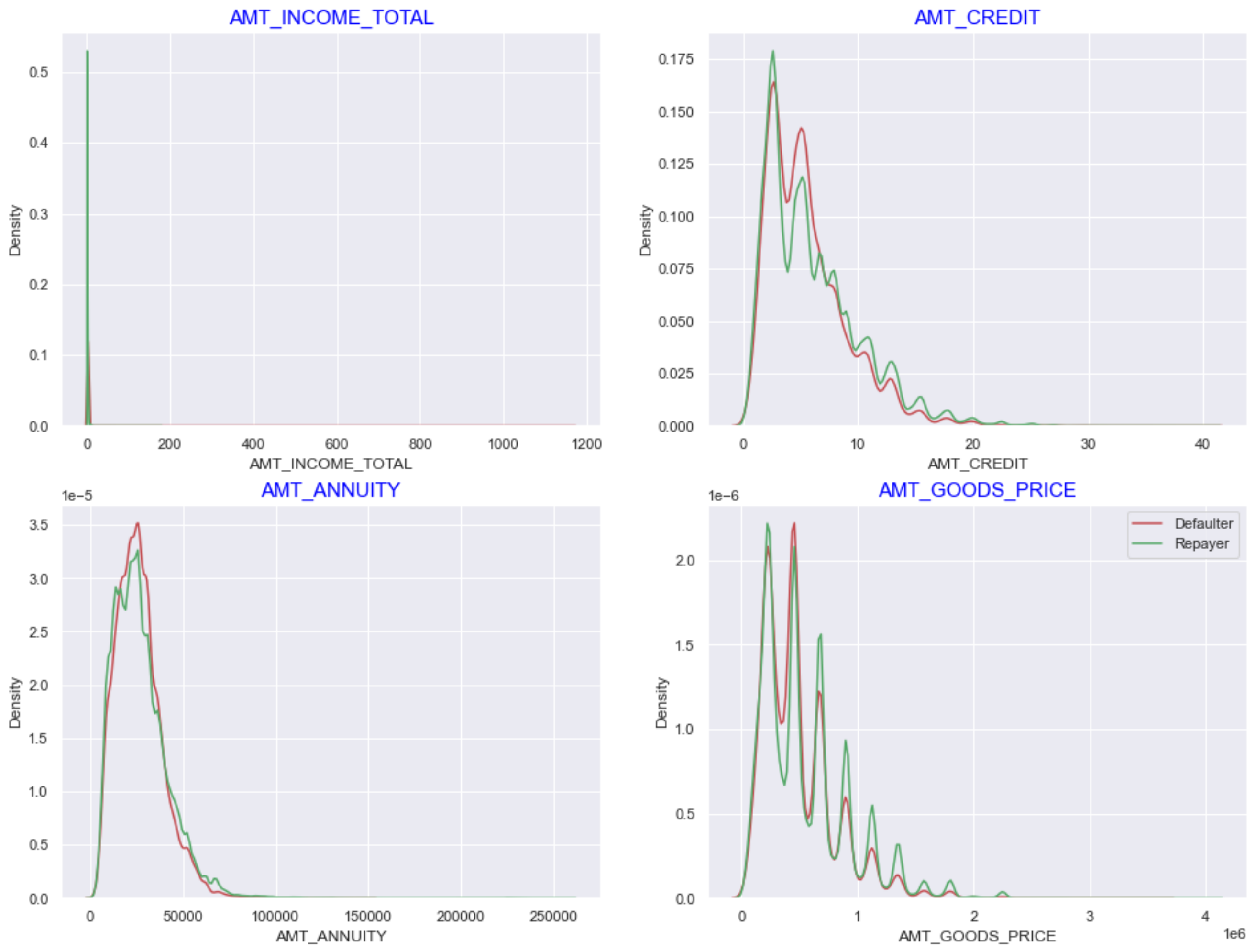
- Credit amount is highly correlated with amount of goods price which is same as repayors. But the loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to repayors(0.77) We can also see that repayors have high correlation in number of days employed(0.62) when compared to defaulters(0.58). There is a severe drop in the correlation between total income of the client and the credit amount(0.038) amongst defaulters whereas it is 0.342 among repayors. Days_



CORRELATION FOR THE REPAYORS DATA

- Correlating factors amongst repayors: Credit amount is highly correlated with amount of goods price loan annuity total income We can also see that repayors have high correlation in number of days employed.

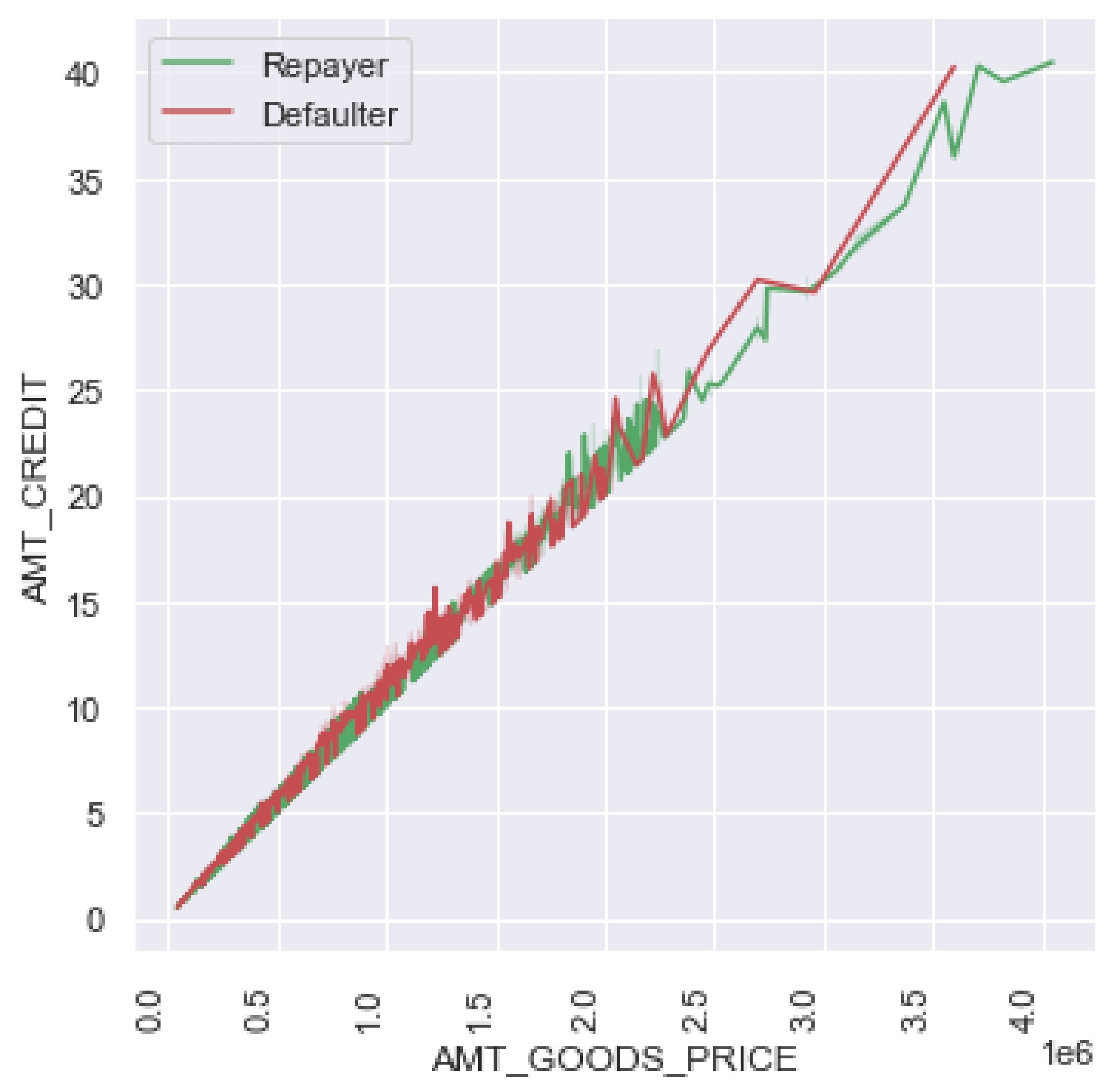
NUMERICAL UNIVARIATE ANALYSIS



PLOTTING THE NUMERICAL COLUMNS RELATED TO AMOUNT AS DISTRIBUTION PLOT TO SEE DENSITY

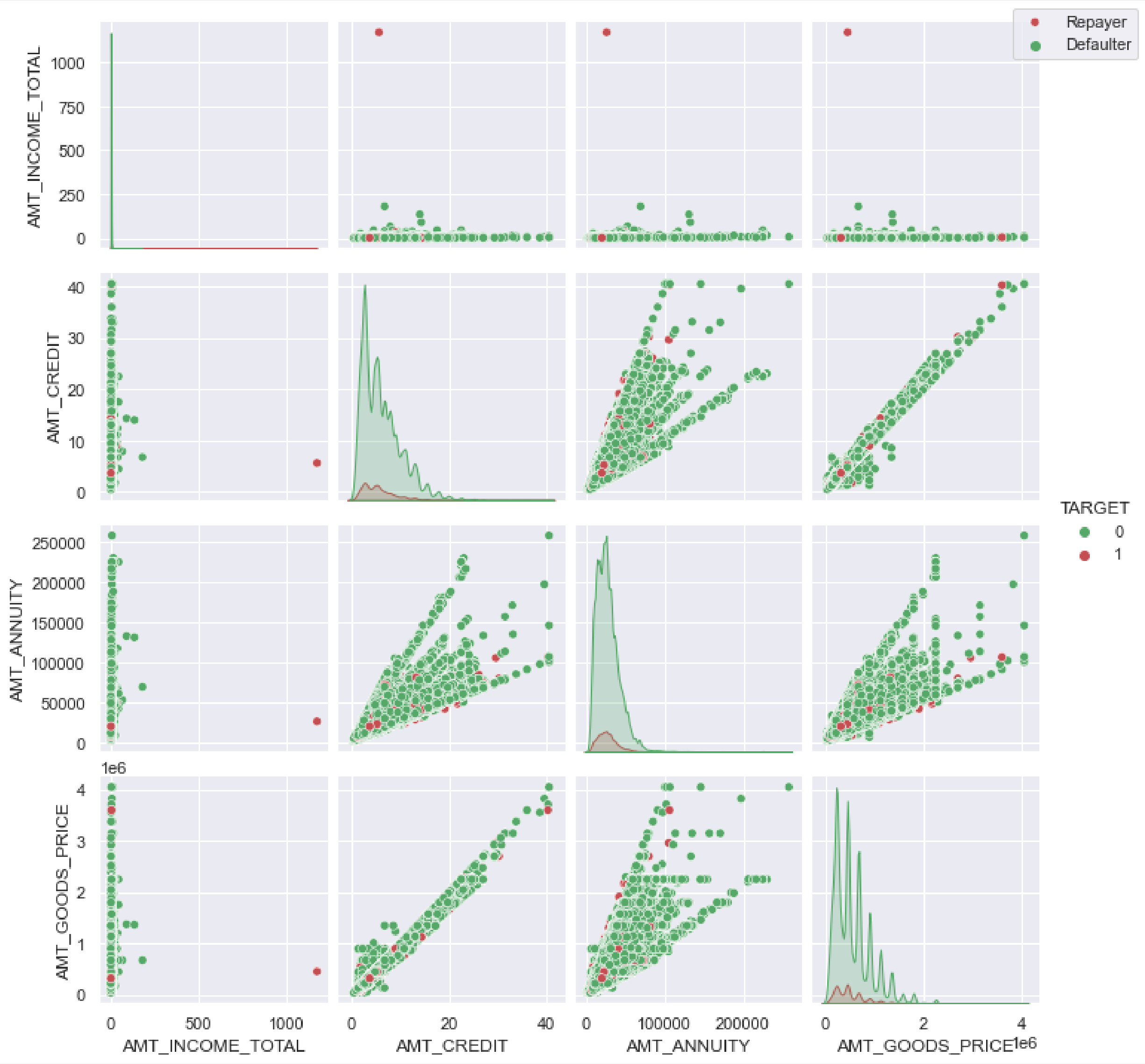
- Most no of loans are given for goods price below 10 lakhs Most people pay annuity below 50000 for the credit loan Credit amount of the loan is mostly less then 10 lakhs The repayors and defaulters distribution overlap in all the plots and hence we cannot use any of these variables in isolation to make a decision

NUMERICAL BIVARIATE ANALYSIS



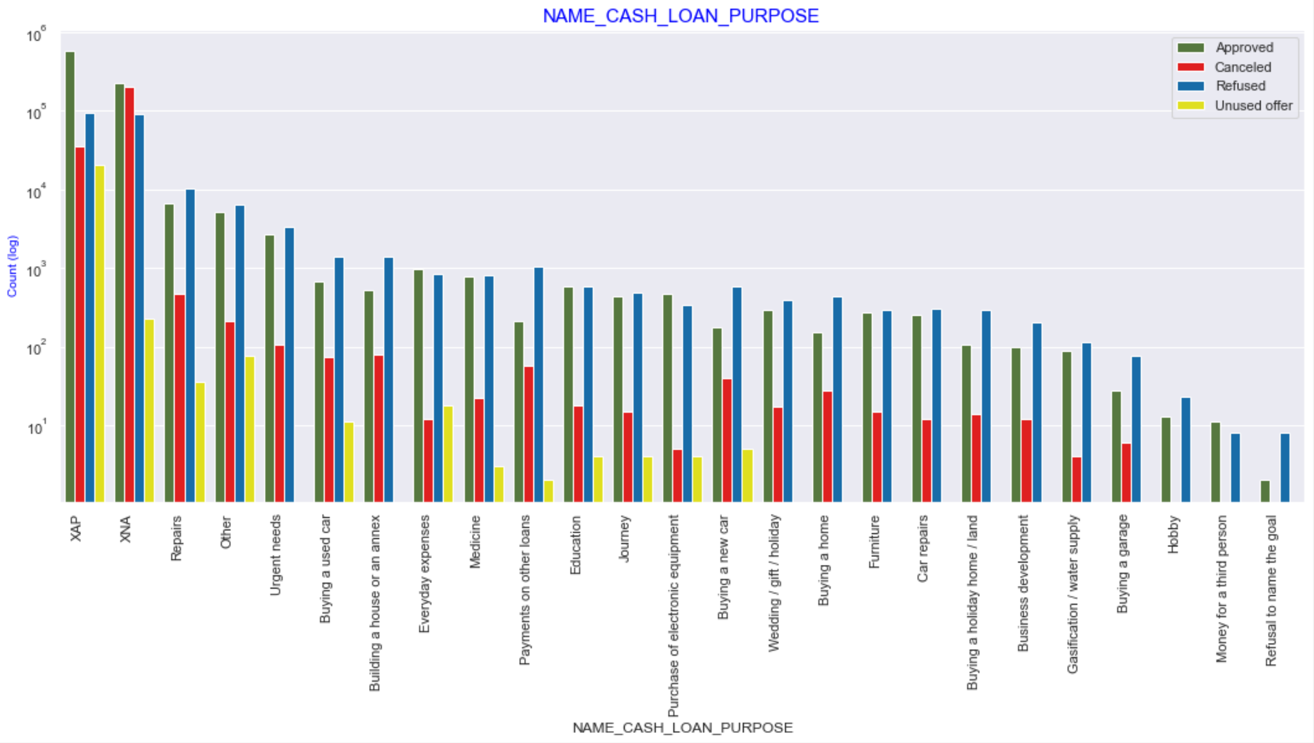
CHECKING THE RELATIONSHIP BETWEEN GOODS PRICE AND CREDIT AND COMPARING WITH LOAN REPAYMENT STATUS

- When the credit amount goes beyond 3M, there is an increase in defaulters.



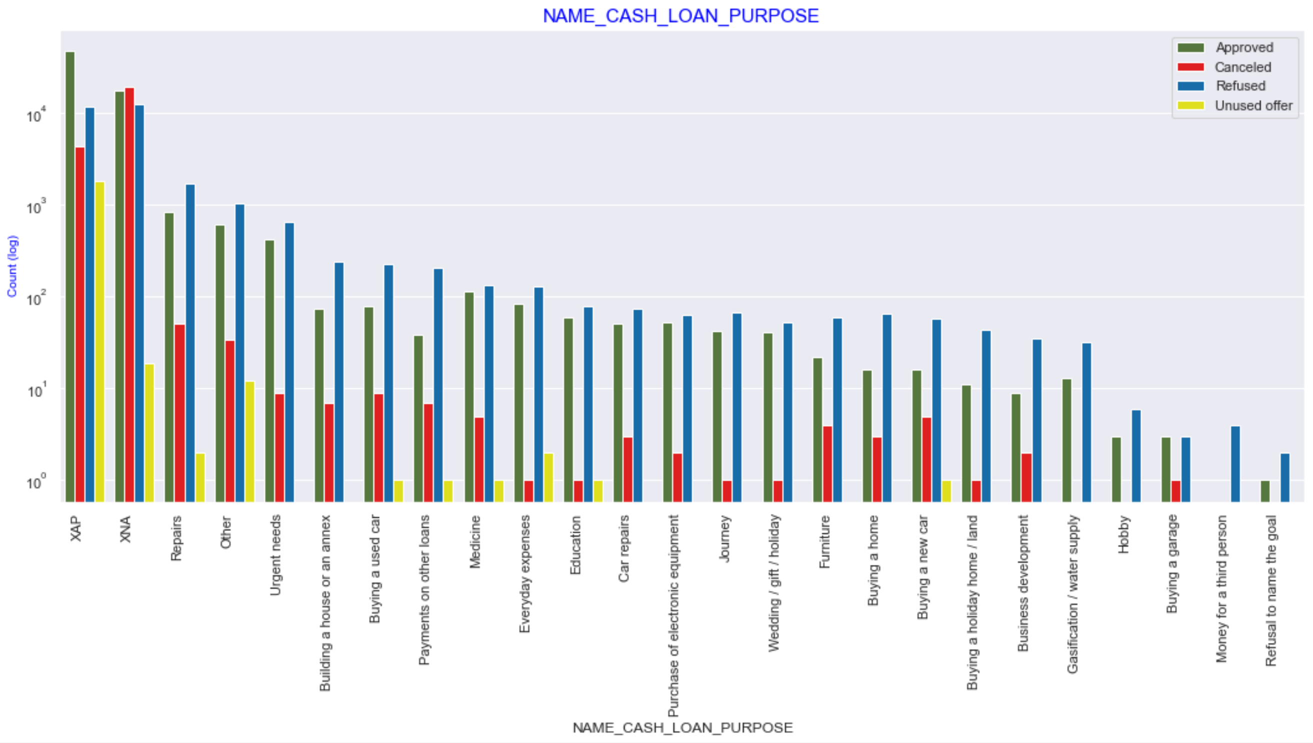
PLOTTING PAIRPLOT BETWEEN AMOUNT VARIABLE TO DRAW REFERENCE AGAINST LOAN REPAYMENT STATUS

- When amt_annuity >15000 amt_goods_price> 3M, there is a lesser chance of defaulters AMT_CREDIT and AMT_GOODS_PRICE are highly correlated as based on the scatterplot where most of the data are consolidated in form of a line There are very less defaulters for AMT_CREDIT >3M Inferences related to distribution plot has been already mentioned in previous distplot graphs inferences section



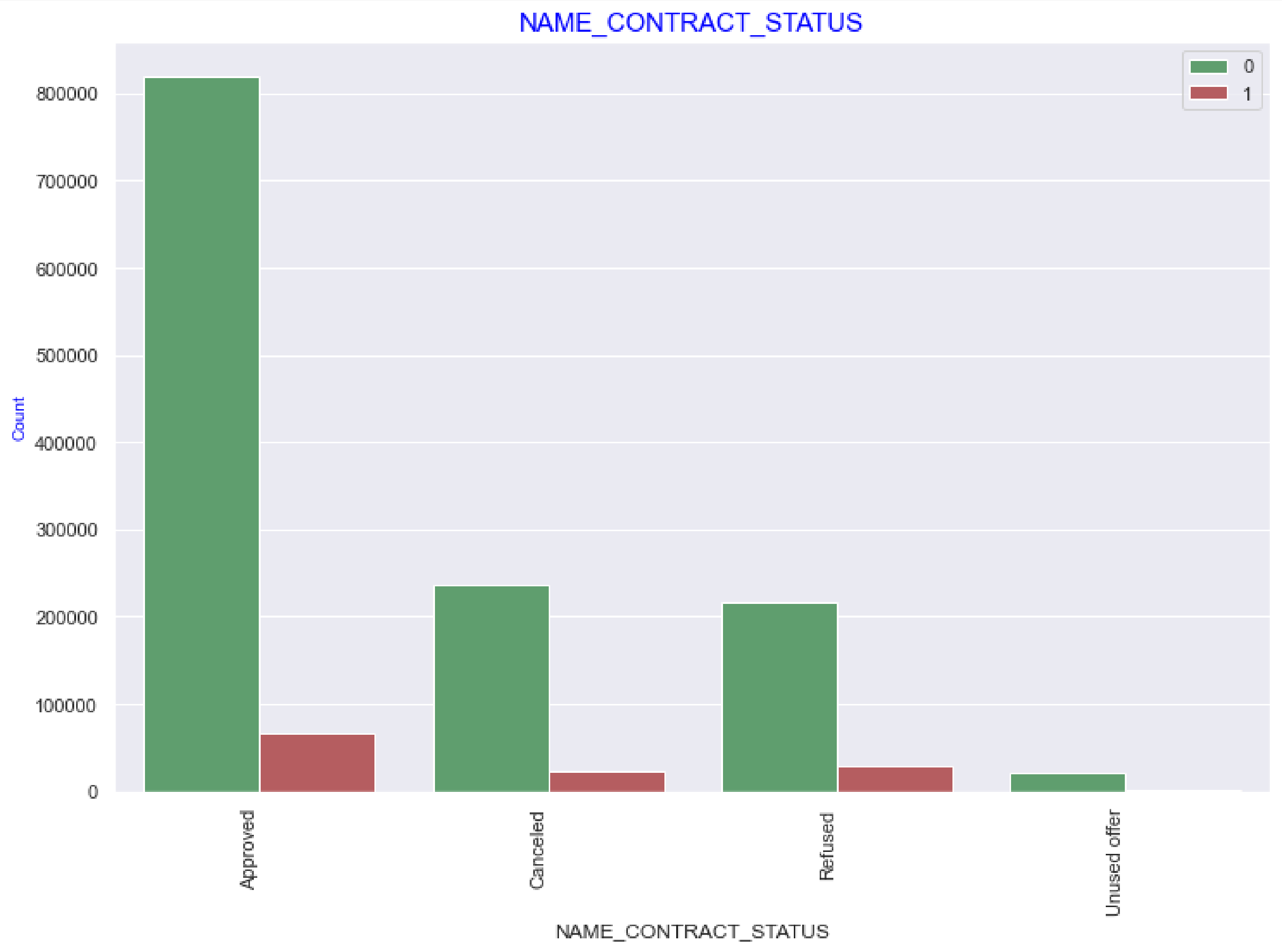
BIFURCATING THE APPLICATIONDF DATAFRAME BASED ON TARGET VALUE 0 FOR CORRELATION AND OTHER ANALYSIS

- Loan purpose has high number of unknown values (XAP, XNA) Loan taken for the purpose of Repairs seems to have highest default rate A very high number application have been rejected by bank or refused by client which has purpose as repair or other. This shows that purpose repair is taken as high risk by bank and either they are rejected or bank offers very high loan interest rate which is not feasible by the clients, thus they refuse the loan.



BIFURCATING THE APPLICATIONDF DATAFRAME BASED ON TARGET VALUE 1 FOR CORRELATION AND OTHER ANALYSIS

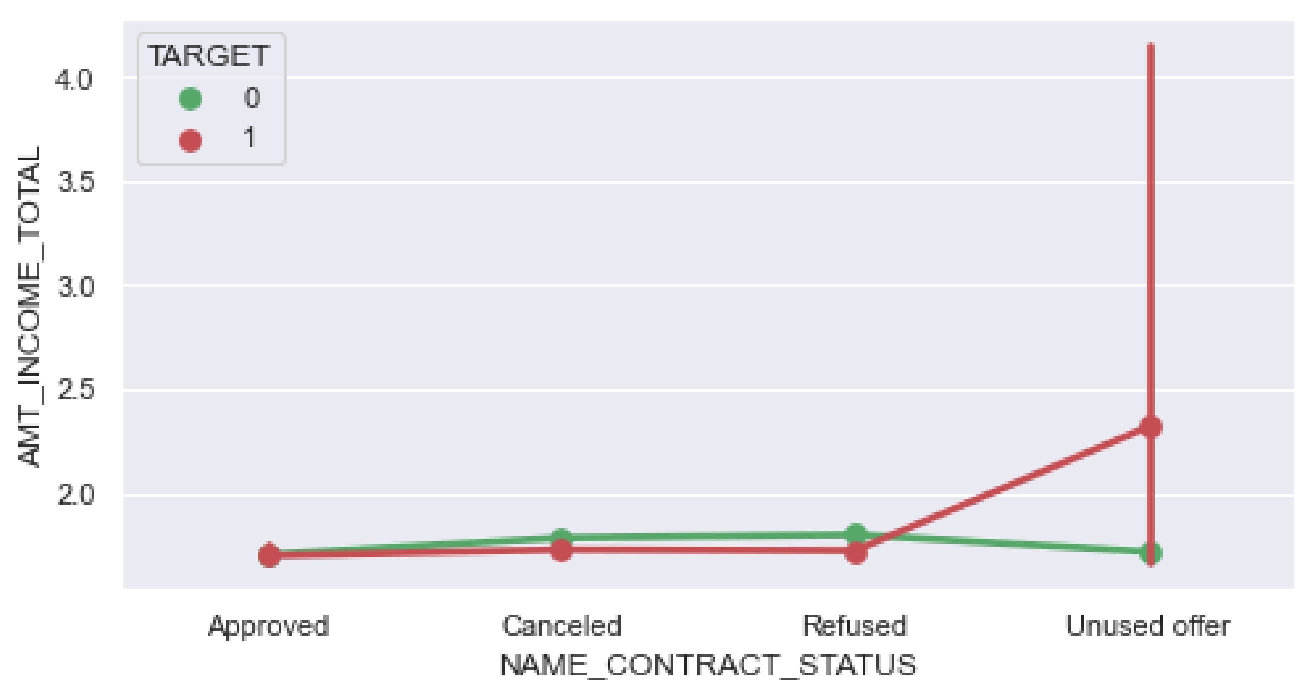
- Loan purpose has high number of unknown values (XAP, XNA) Loan taken for the purpose of Repairs seems to have highest default rate A very high number application have been rejected by bank or refused by client which has purpose as repair or other. This shows that purpose repair is taken as high risk by bank and either they are rejected or bank offers very high loan interest rate which is not feasible by the clients, thus they refuse the loan.



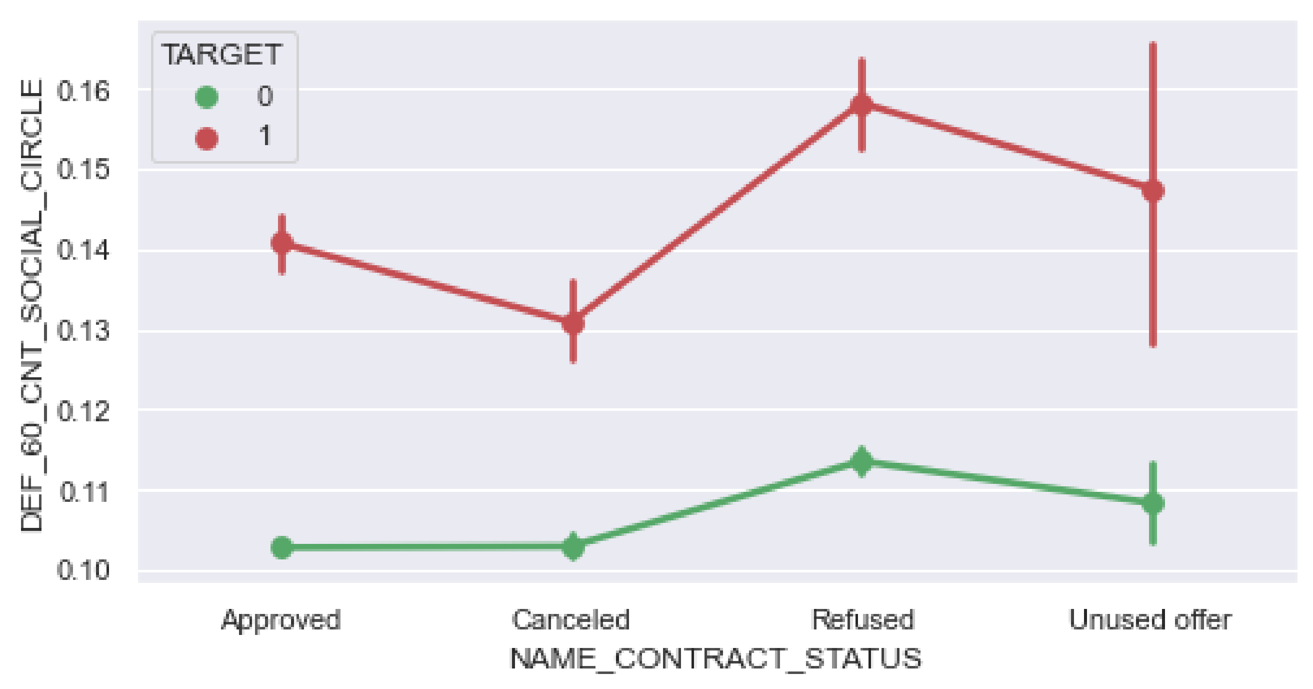
CHECKING THE CONTRACT STATUS BASED ON LOAN REPAYMENT STATUS AND WHETHER THERE IS ANY BUSINESS LOSS OR FINANCIAL LOSS

- 90% of the previously cancelled client have actually repayed the loan. Revisiting the interest rates would increase business opportunity for these clients 88% of the clients who have been previously refused a loan has payed back the loan in current case. Refusal reason should be recorded for further analysis as these clients would turn into potential repaying customer.

PLOTTING THE RELATIONSHIP BETWEEN INCOME TOTAL AND CONTACT STATUS



- The point plot show that the people who have not used offer earlier have defaulted even when there average income is higher than others



PLOTTING THE RELATIONSHIP BETWEEN PEOPLE WHO DEFAULTED IN LAST 60 DAYS BEING IN CLIENT'S SOCIAL CIRCLE AND CONTACT STATUS

- Clients who have average of 0.13 or higher DEF_60_CNT_SOCIAL_CIRCLE score tend to default more and hence client's social circle has to be analyzed before providing the loan.