CREDIT EDA CASE STUDY

BY CHANDAN KUMAR

INTRODUCTION

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

Business Understanding

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 to 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.

Business Objectives

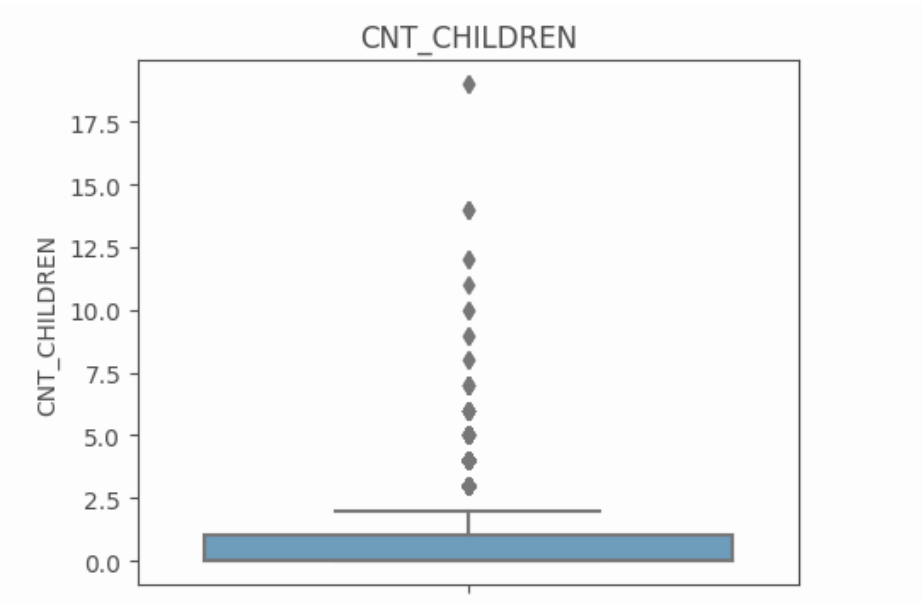
This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such 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.

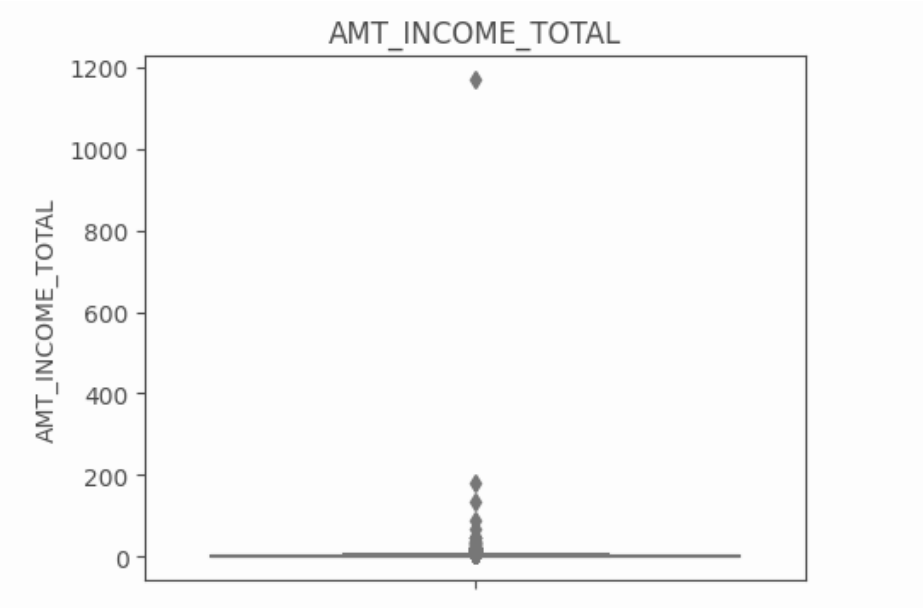
To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

CHECKING FOR OUTLIERS

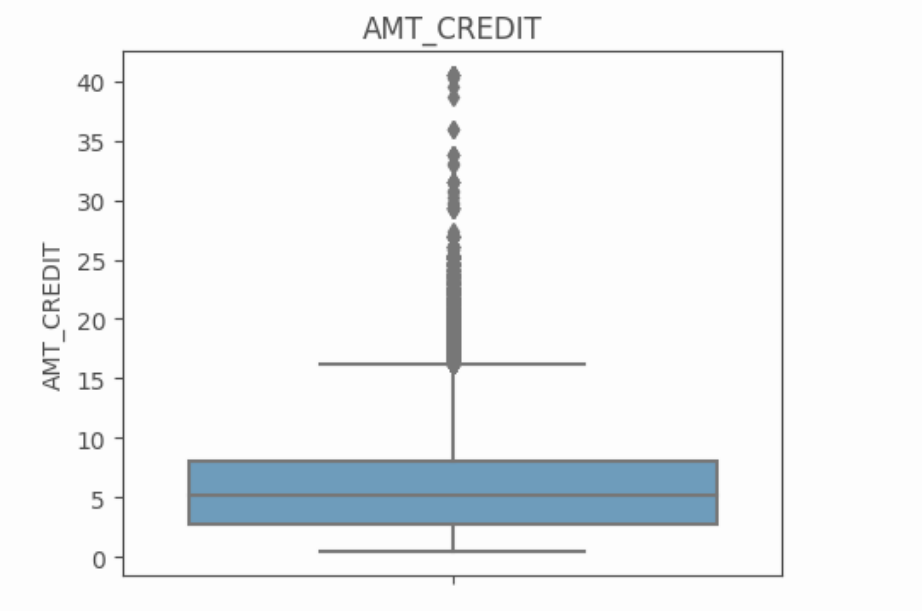
- CNT_CHILDREN have some number of outliers.



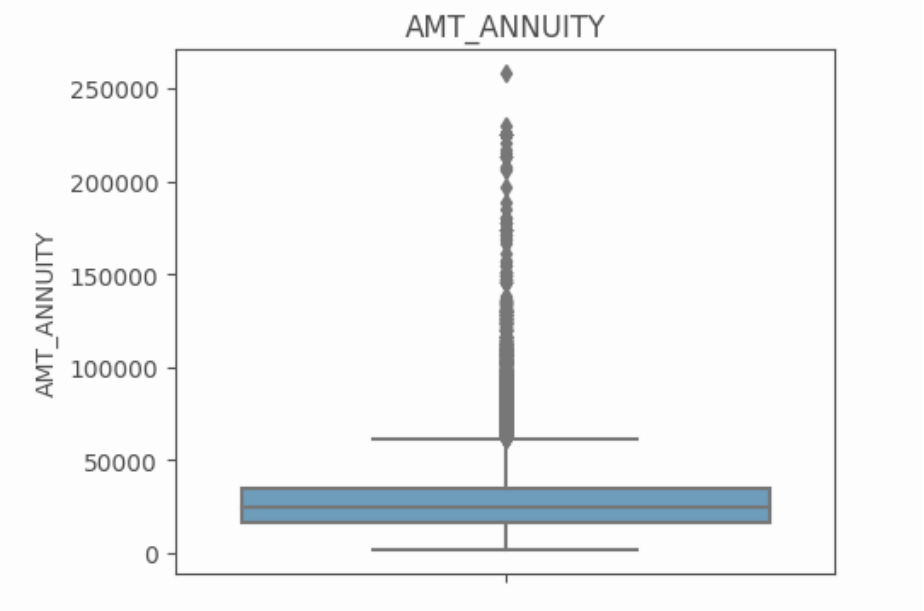
- AMT_INCOME_TOTAL has huge number of outliers which indicate that few of the loan applicants have very high income when compared to the others.



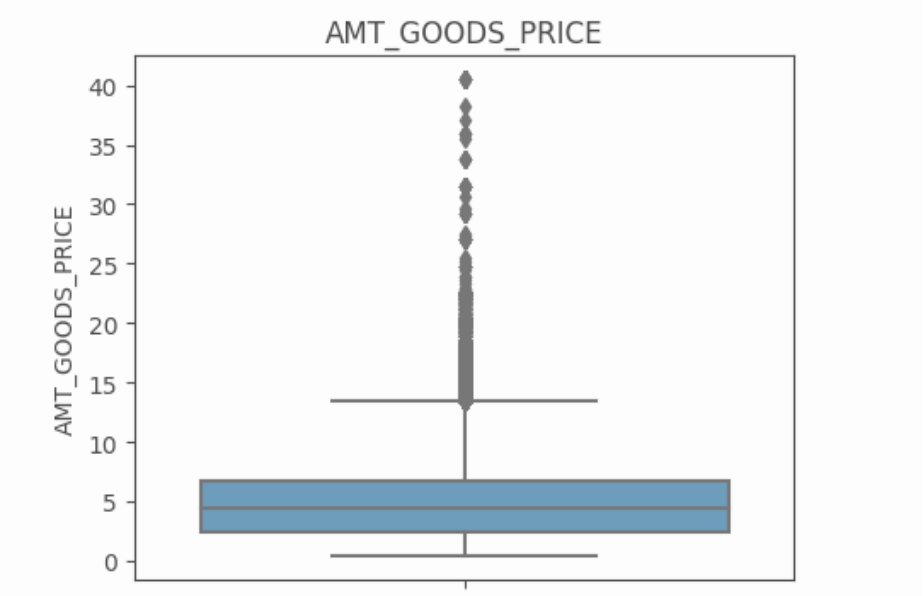
- The outliers in AMT_CREDIT is most likely relevant value. This value can be binned when analyzing.



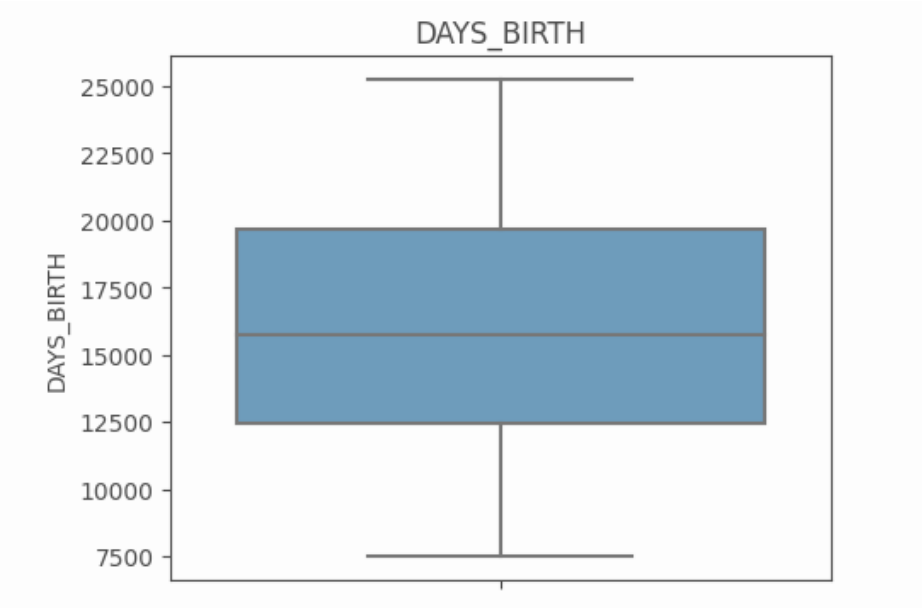
- The outliers in AMT_ANNUITY is most likely relevant value. This value can be binned when analyzing.



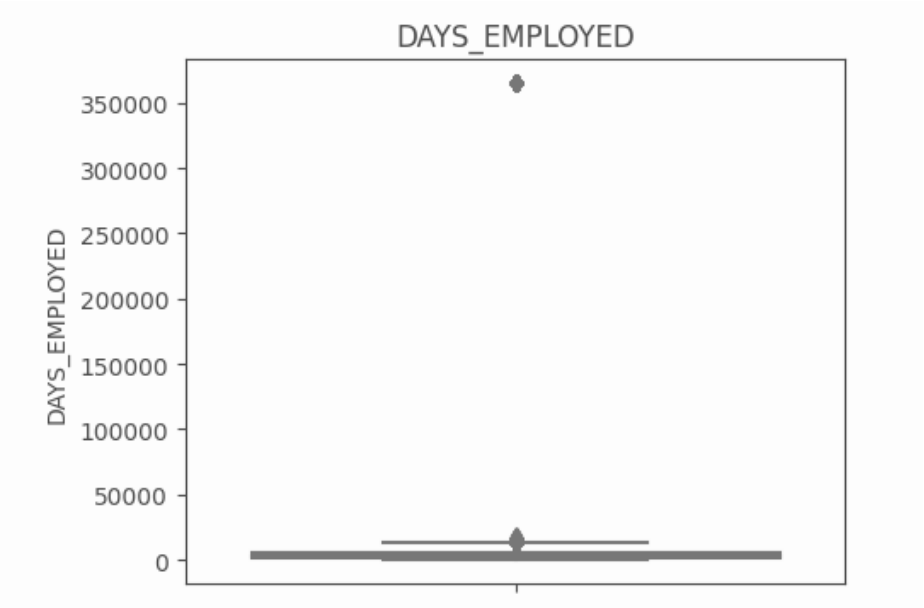
- AMT_GOODS_PRICE have some number of outliers which is relevent.



- DAYS_BIRTH has no outliers which means the data is reliable.

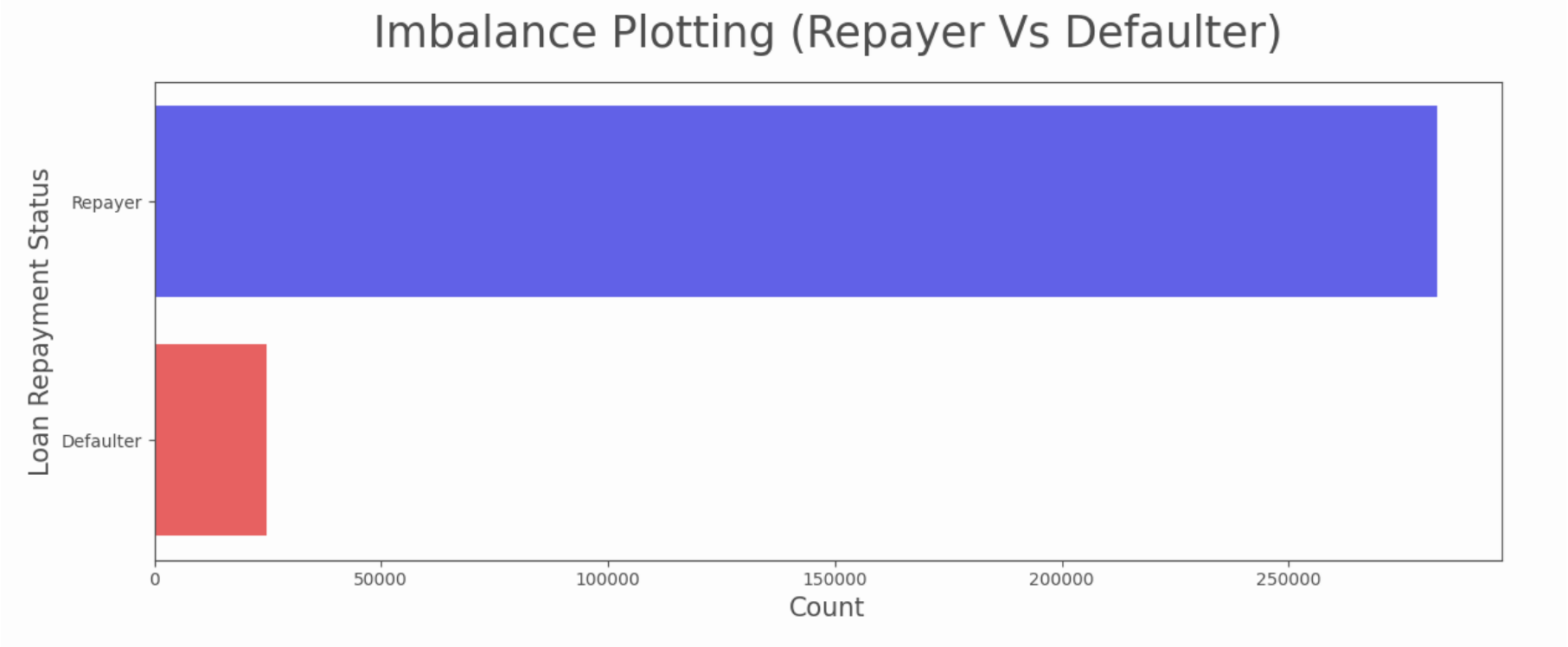


- DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and hence this has to be incorrect entry.



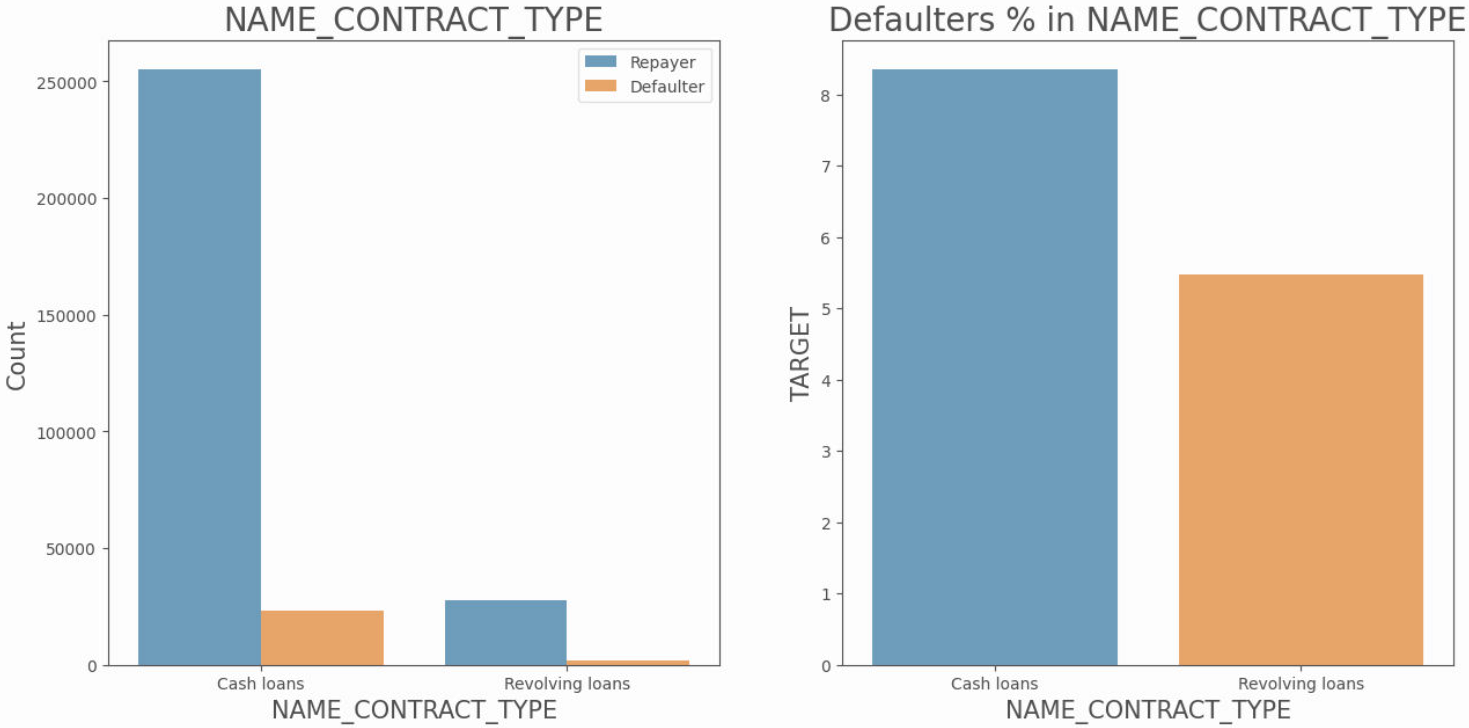
ANALYSIS

Imbalance Data



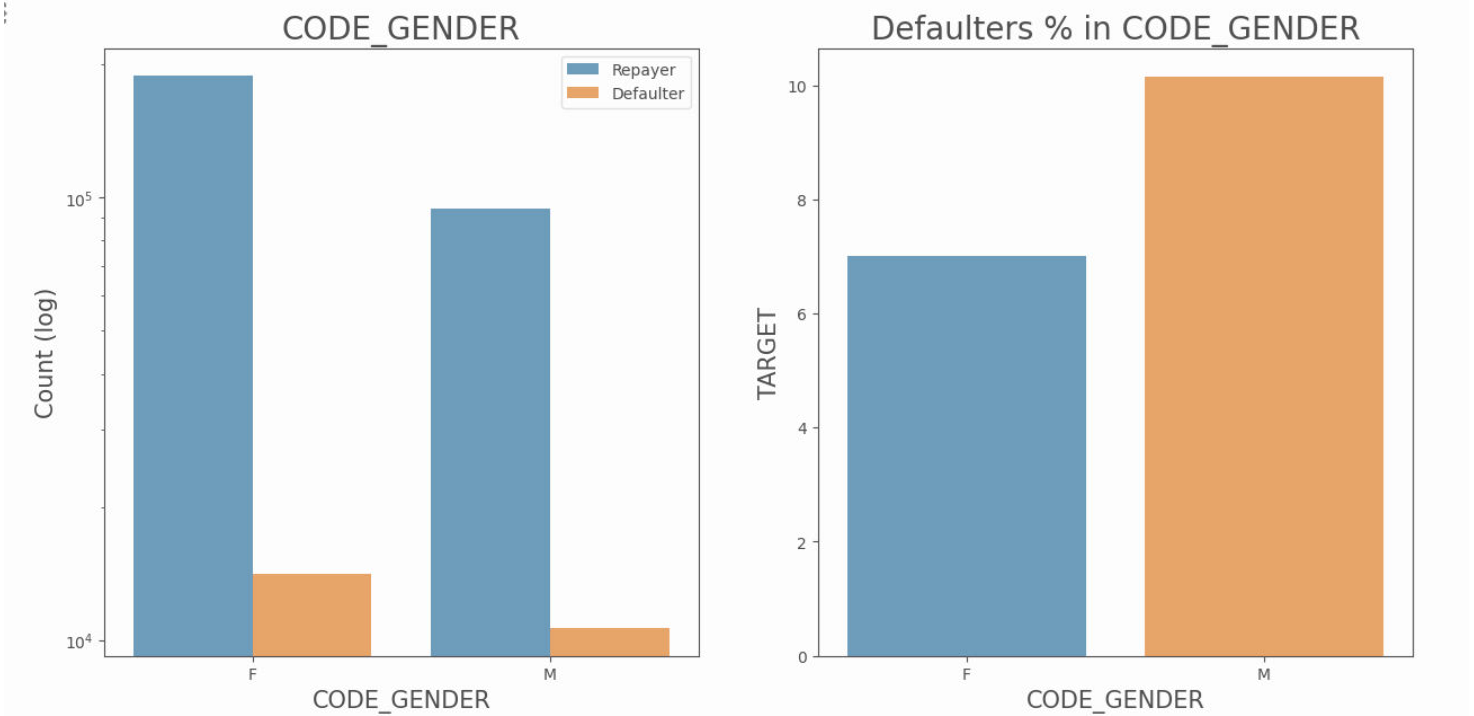
CATEGORICAL VARIATE ANALYSIS

Contract Type



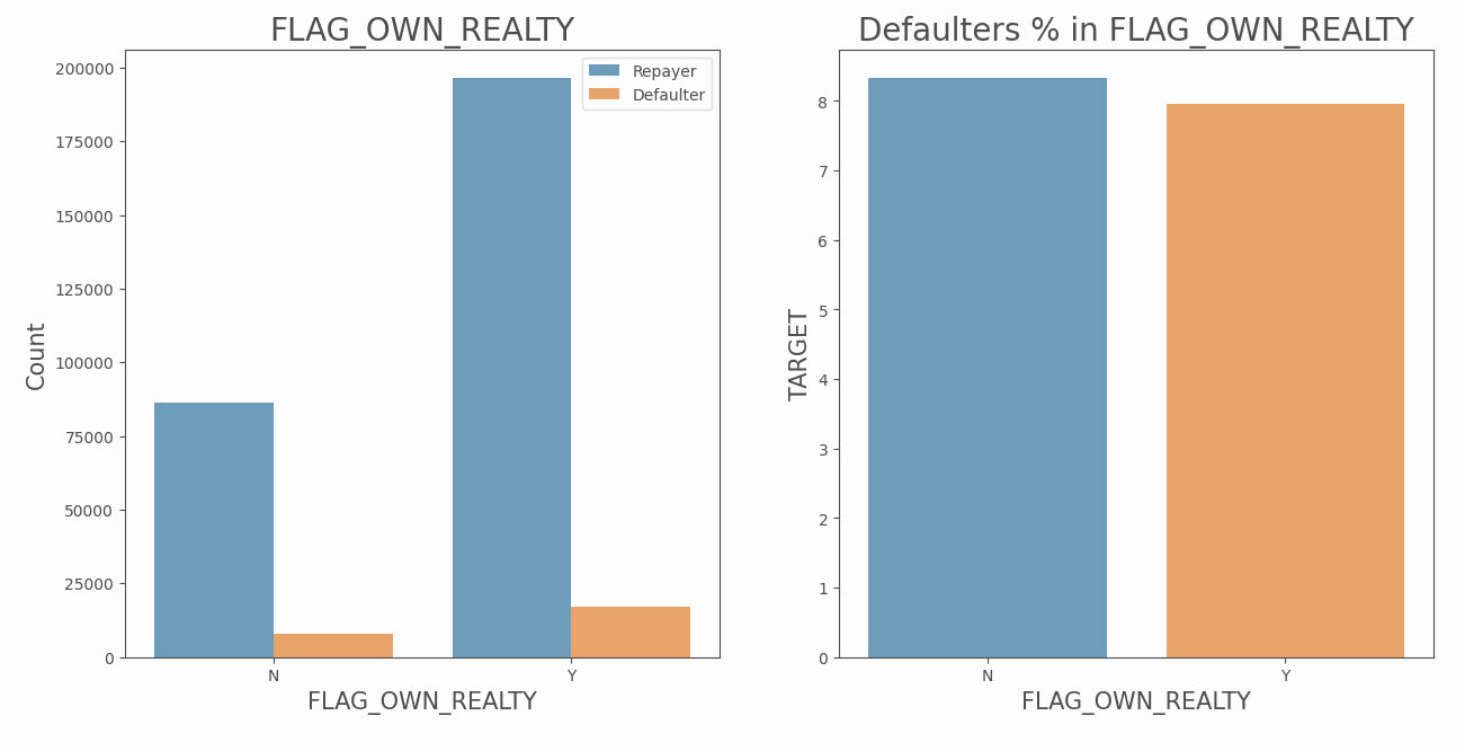
- Revolving loans are just a small fraction (10%) from the total number of loans
- Around 8-9% Cash loan applicants and 5-6% Revolving loan applicant are in defaulters

Code Gender



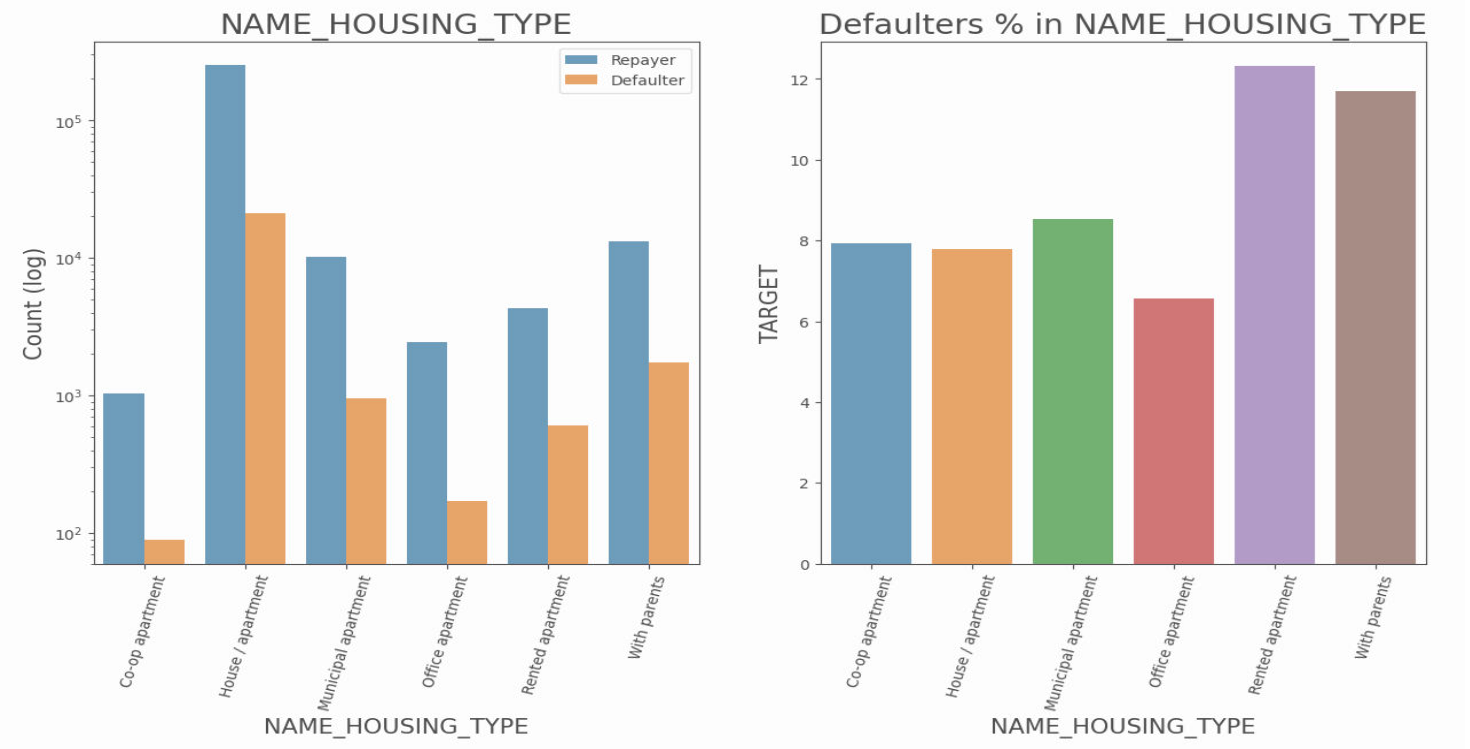
- The number of female clients is almost double the number of male clients.
- Based on the percentage of defaulted credits, males have a higher chance of not returning their loans about 10%, comparing with women about 7%

FLAG OWN REALTY



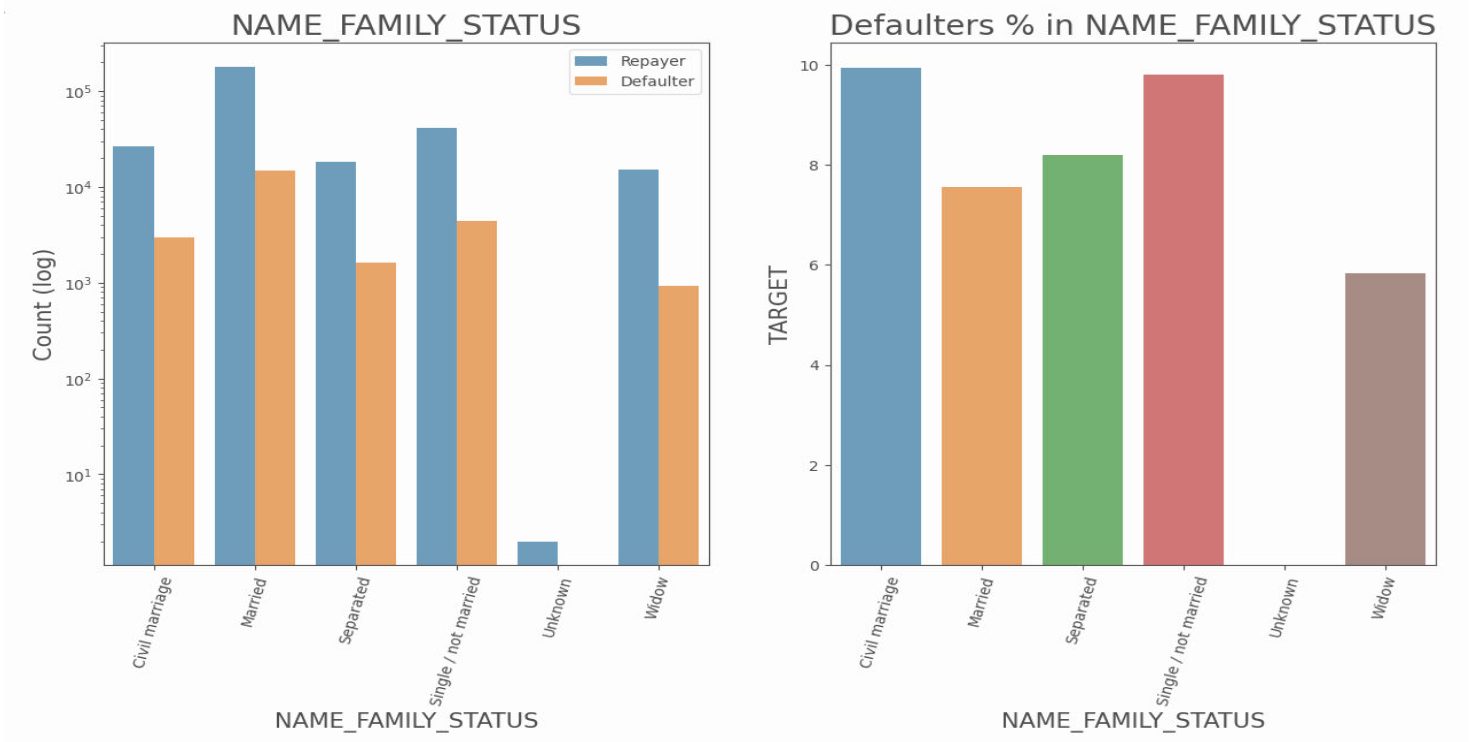
- The clients who own real estate are more than double of the ones that don't own.
- The defaulting rate of both categories are around the same (~8%). Thus we can infer that there is no correlation between owning a reality and defaulting the loan.

HOUSING TYPE



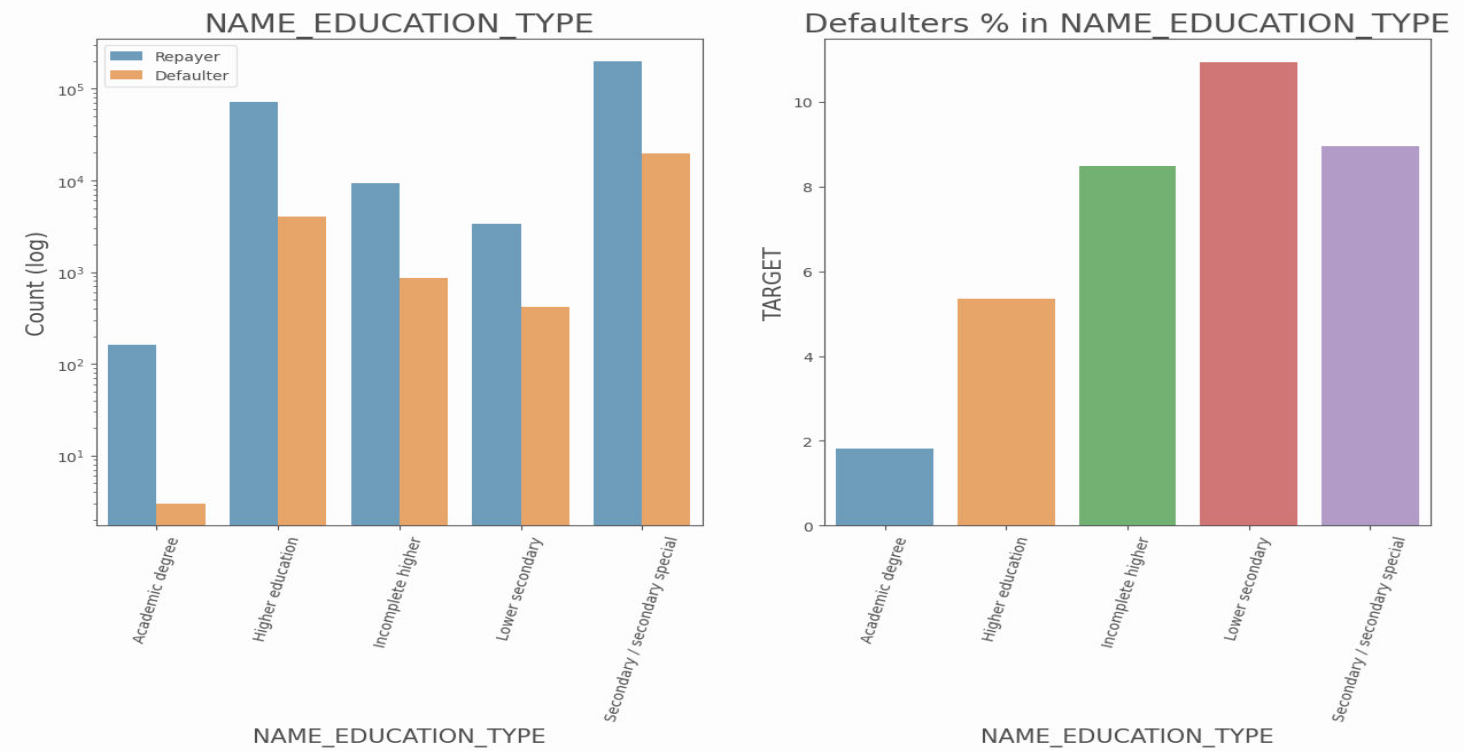
- Majority of people live in House/apartment
- People living in office apartments have lowest default rate
- People living with parents (~11.5%) and living in rented apartments(>12%) have higher probability of

FAMILY STATUS



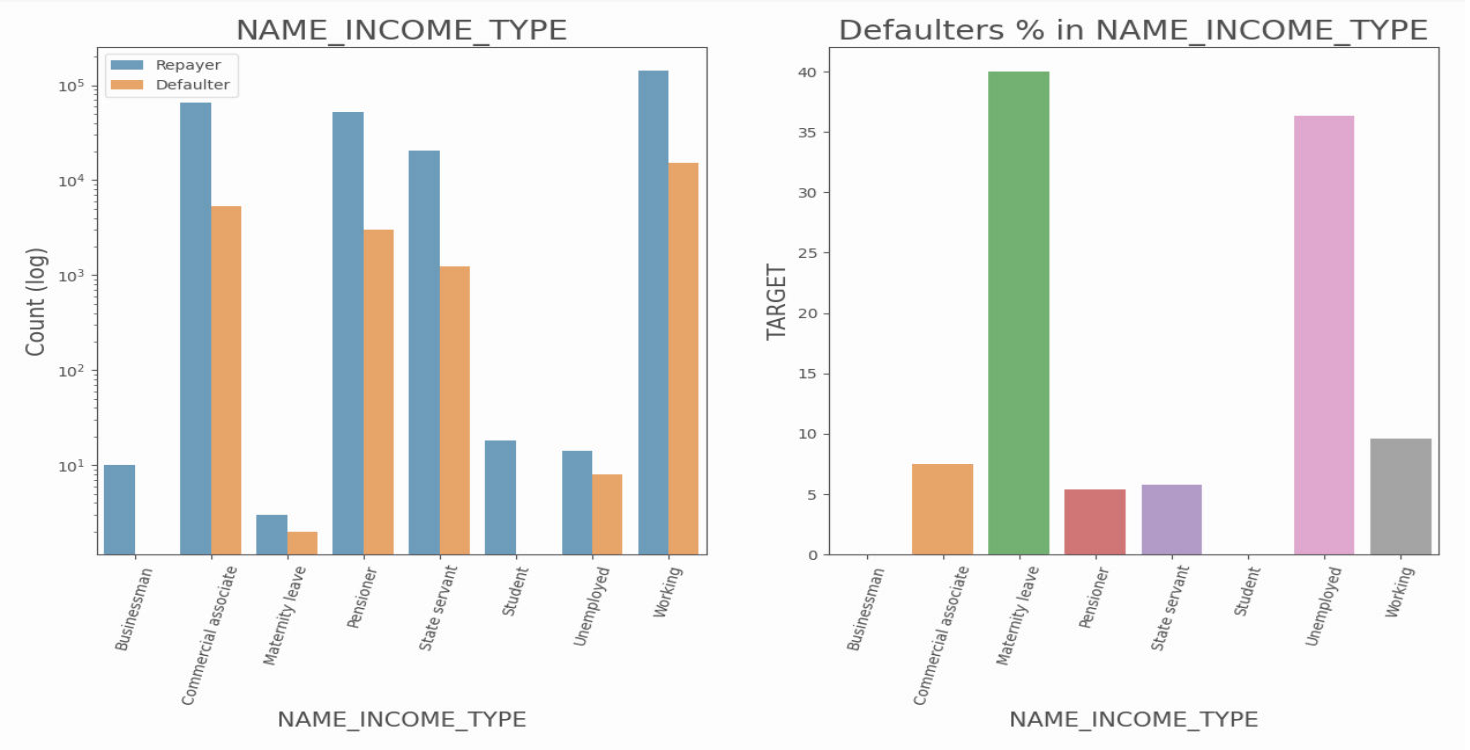
- Most of the people who have taken loan are married, followed by Single/not married and civil marriage
- In Percentage of defaulters, Civil marriage has the highest percent around (10%) and widow has the

EDUCATION TYPE



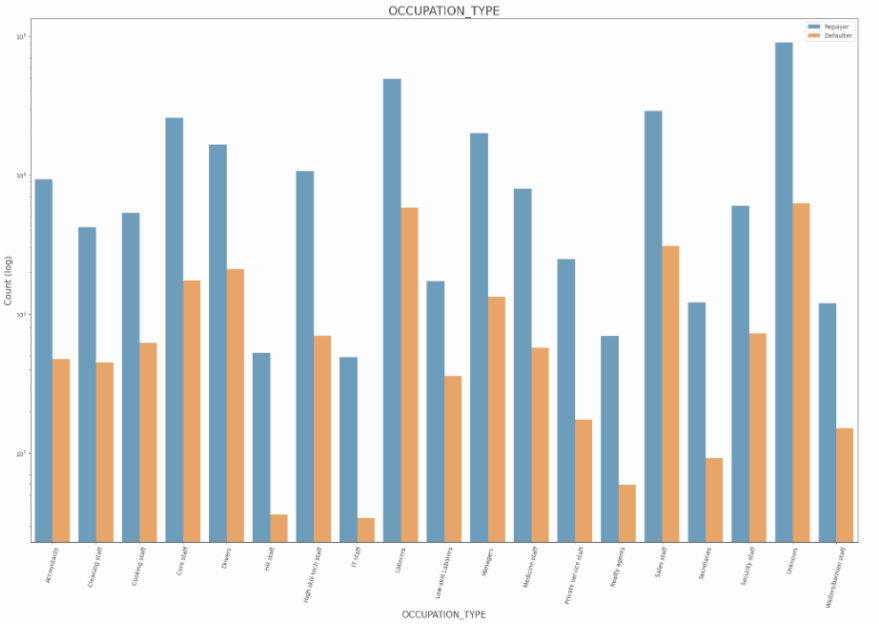
- Majority of clients have Secondary/secondary special education, followed by clients with Higher education.
- Very few clients have an academic degree
- Lower secondary category have highest rate of defaulting around 11%.

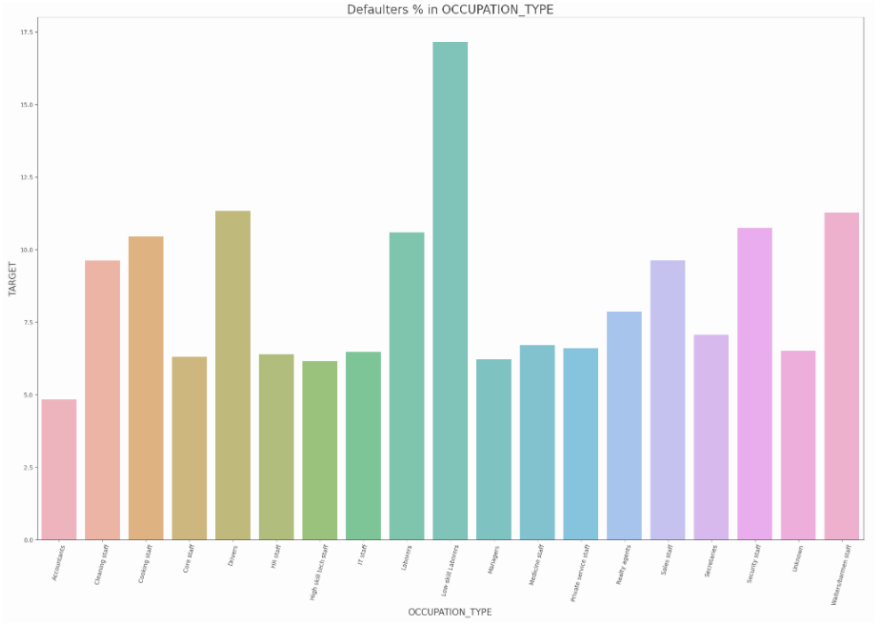
INCOME TYPE



- Most of applicants for loans income type is Working, followed by Commercial associate, Pensioner and State servant.
- The applicants who are on Maternity leave have defaulting percentage of 40% which is the highest, followed by Unemployed (37%). The rest under average around 10% defaultees.
- Student and Businessmen though less in numbers, do not have default record. Safest two categories for providing loan.

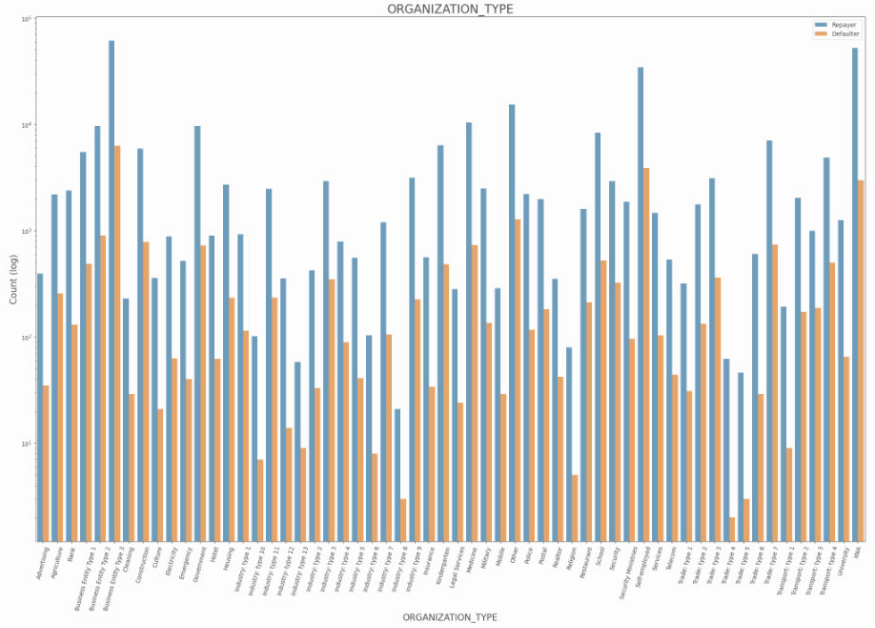
OCCUPATION TYPE

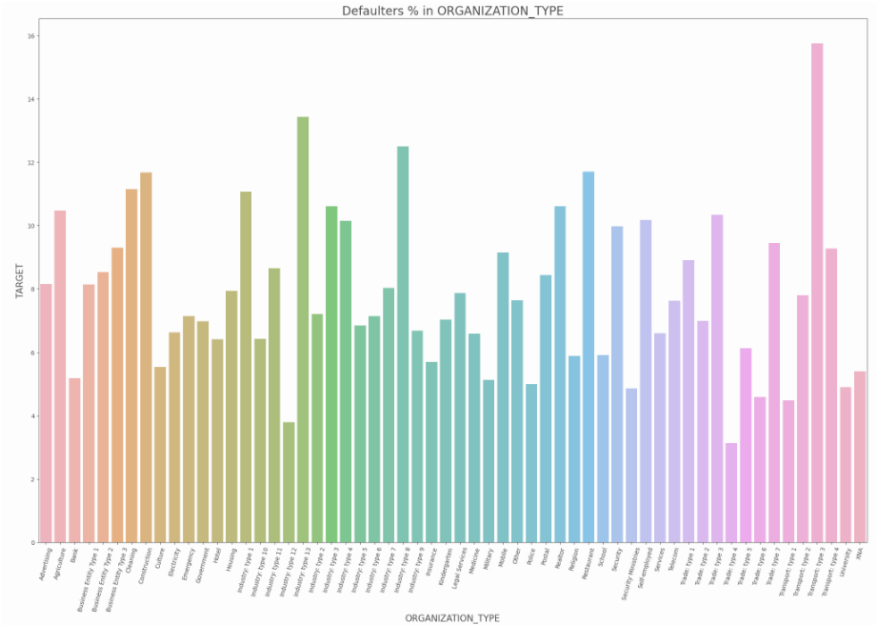




- Most of the loans are taken by Laborers, followed by Sales staff.
- IT staff and HR Staff are less likely to apply for Loan.
- Category with highest percent of defauters are Low-skill Laborers (above 17%), followed by Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff

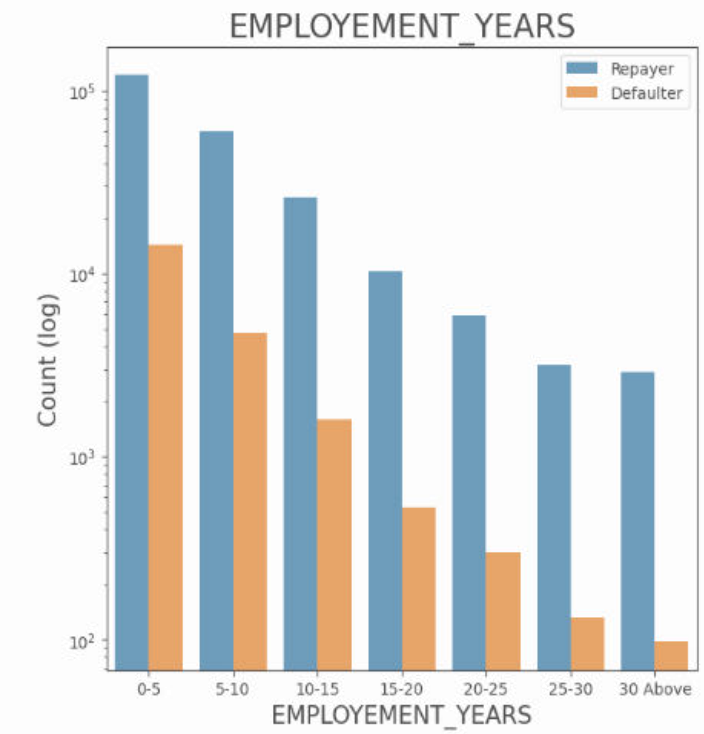
ORGANIZATION TYPE

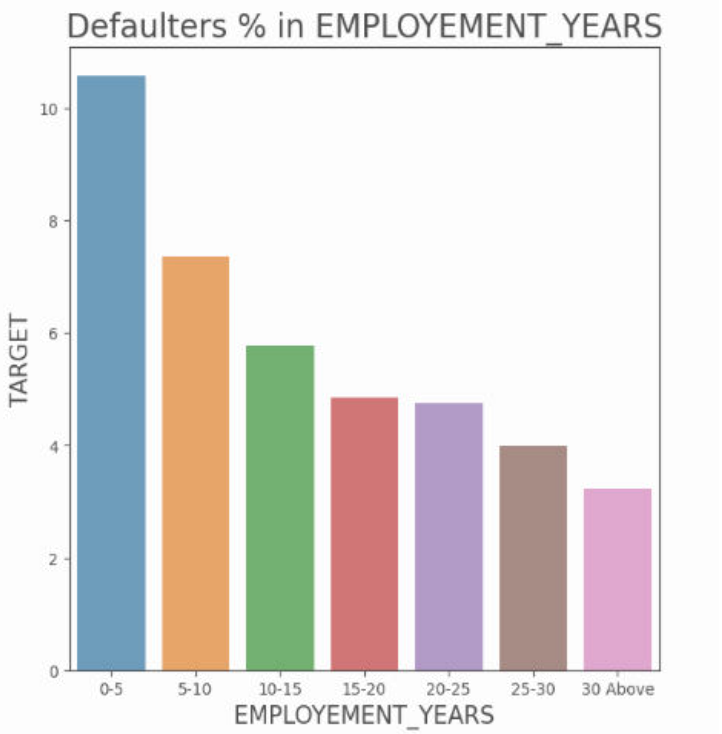




- Organizations with highest percent of defaultess are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%).
- Self employed people have relative high defaulting rate,to be safer side loan disbursement should be avoided or provide loan with higher 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 unavailable(XNA)

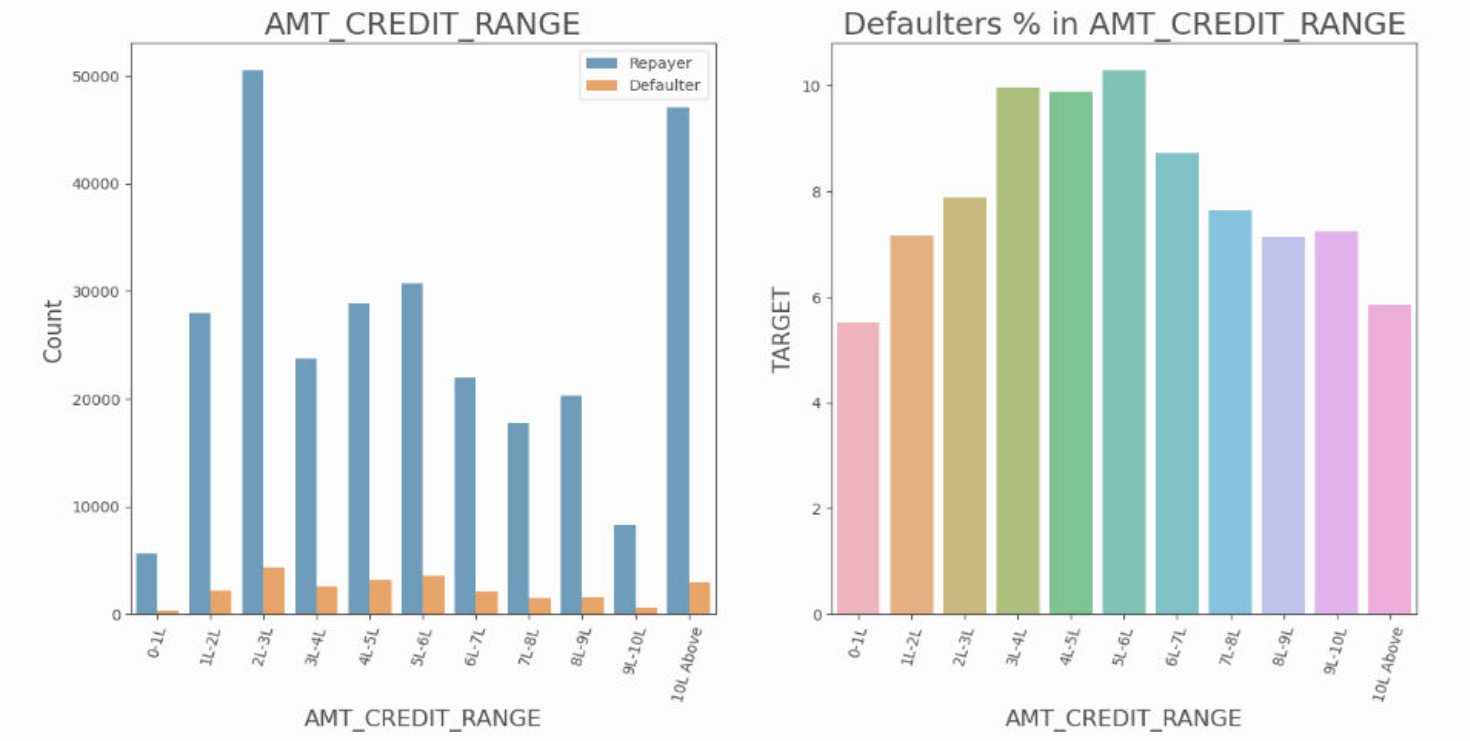
EMPLOYMENT YEARS





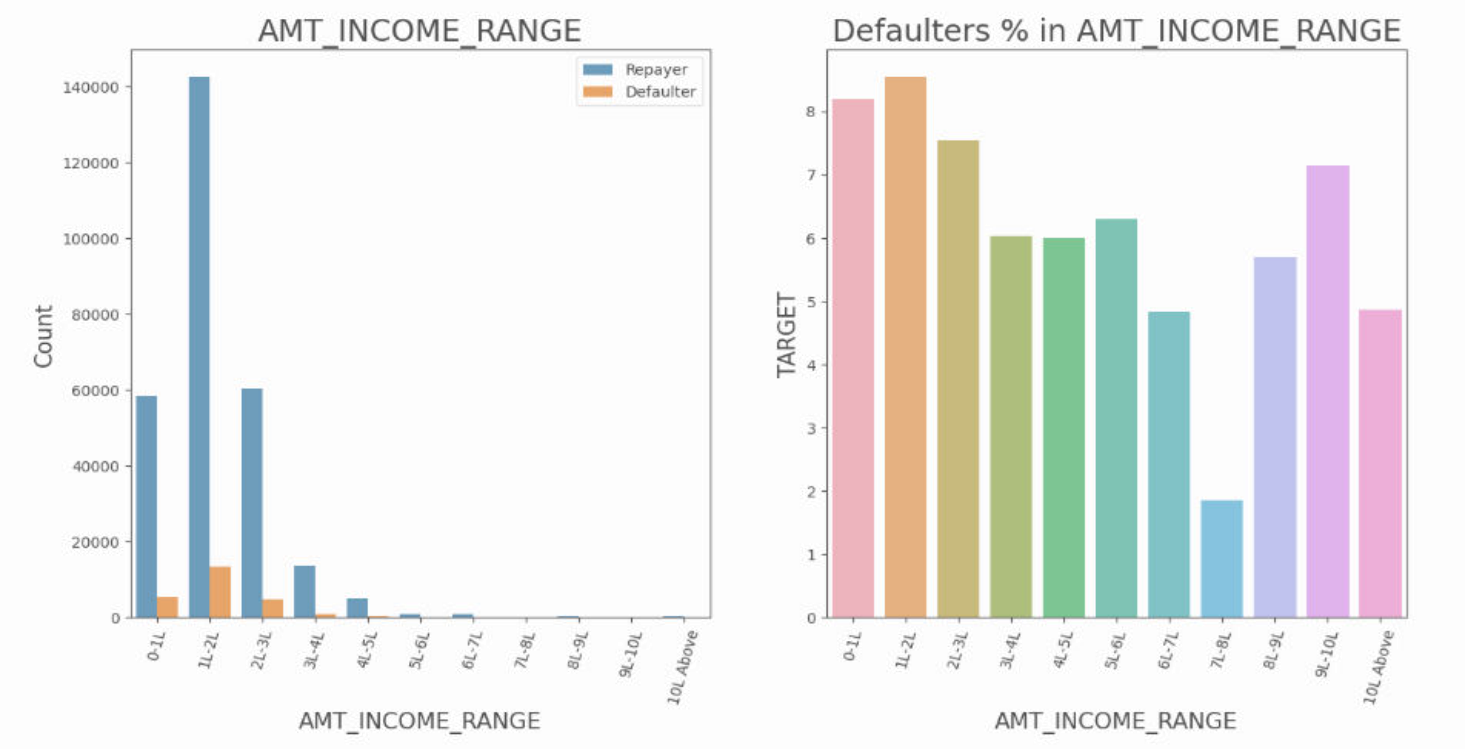
- Majority of the applicants having working experience between 0-5 years are defaultees. The defaulting rating of this group is also the highest which is around 10%
- With increase of employment year, defaulting rate is radually decreasing.
- With people having 40+ year experience have less than 1% default rate

AMOUNT CREDIT



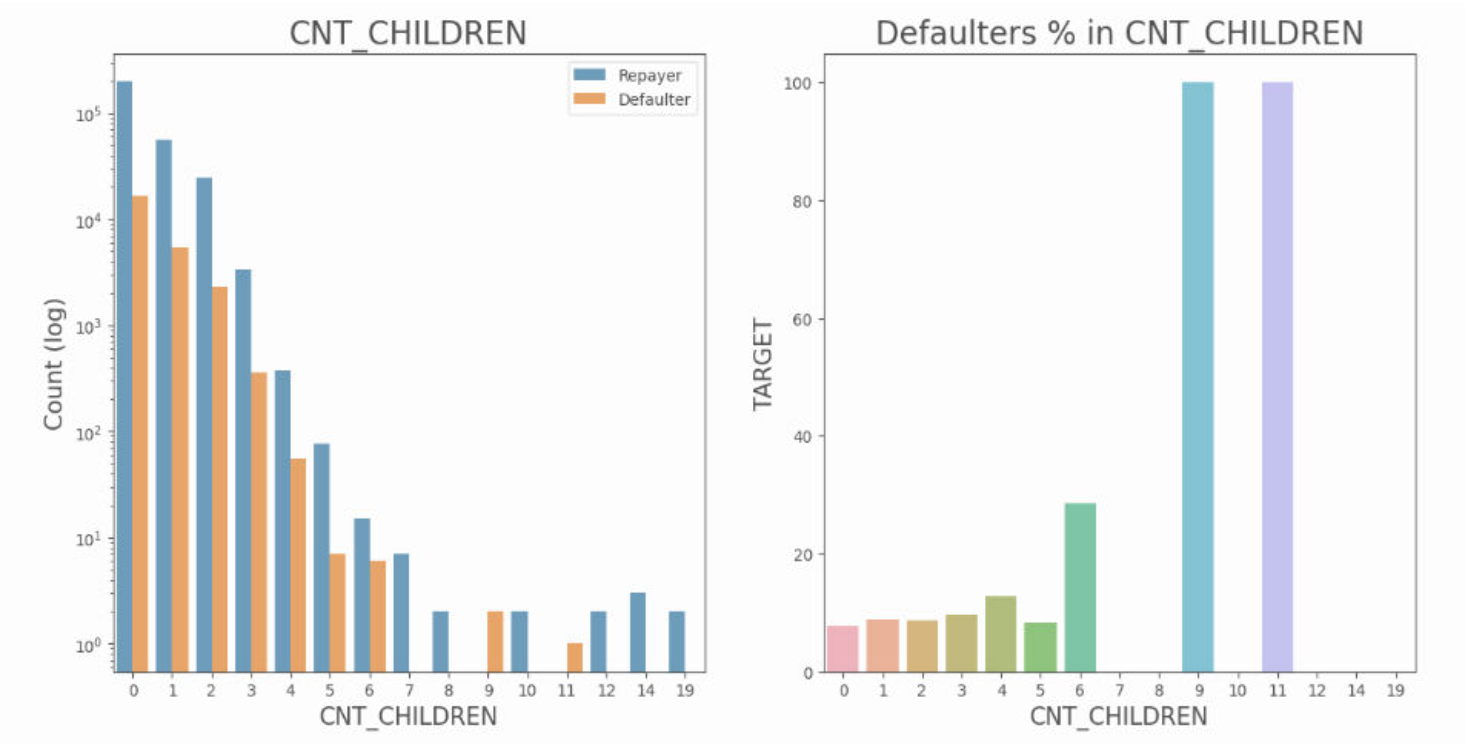
- There are high number of applicants have loan in range of 2-3 Lakhs followed by 10 Lakh above range
- People who get loan for 3-6 Lakhs have most number of defaulters than other loan range.

AMOUNT INCOME



- Majority of the applications have Income total less than 3 Lakhs.
- Application with Income less than 3 Lakhs has high probability of defaulting
- Applicant with Income 7-8 Lakhas are less likely to default.

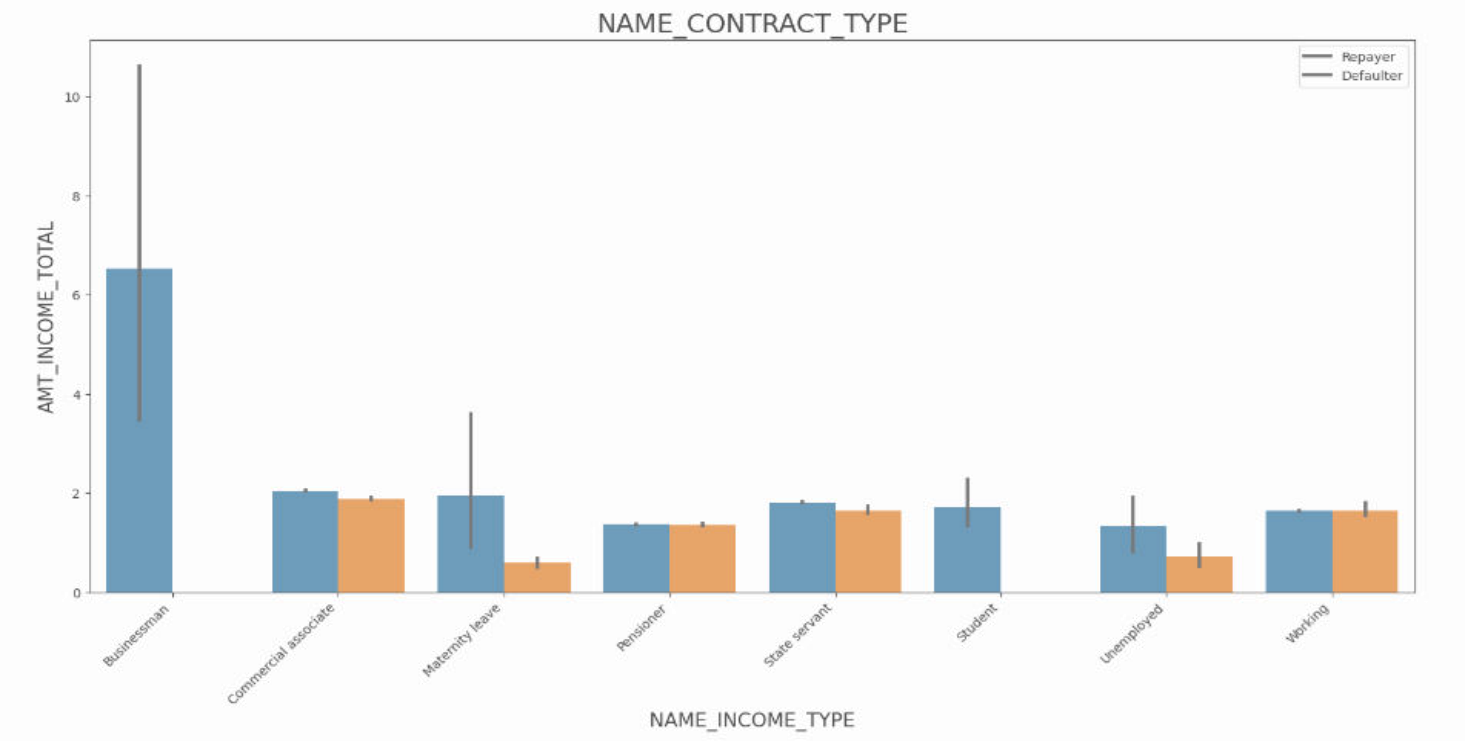
AMOUNT INCOME



- 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

CATEGORICAL BI-VARIATE ANALYSIS

AMOUNT INCOME

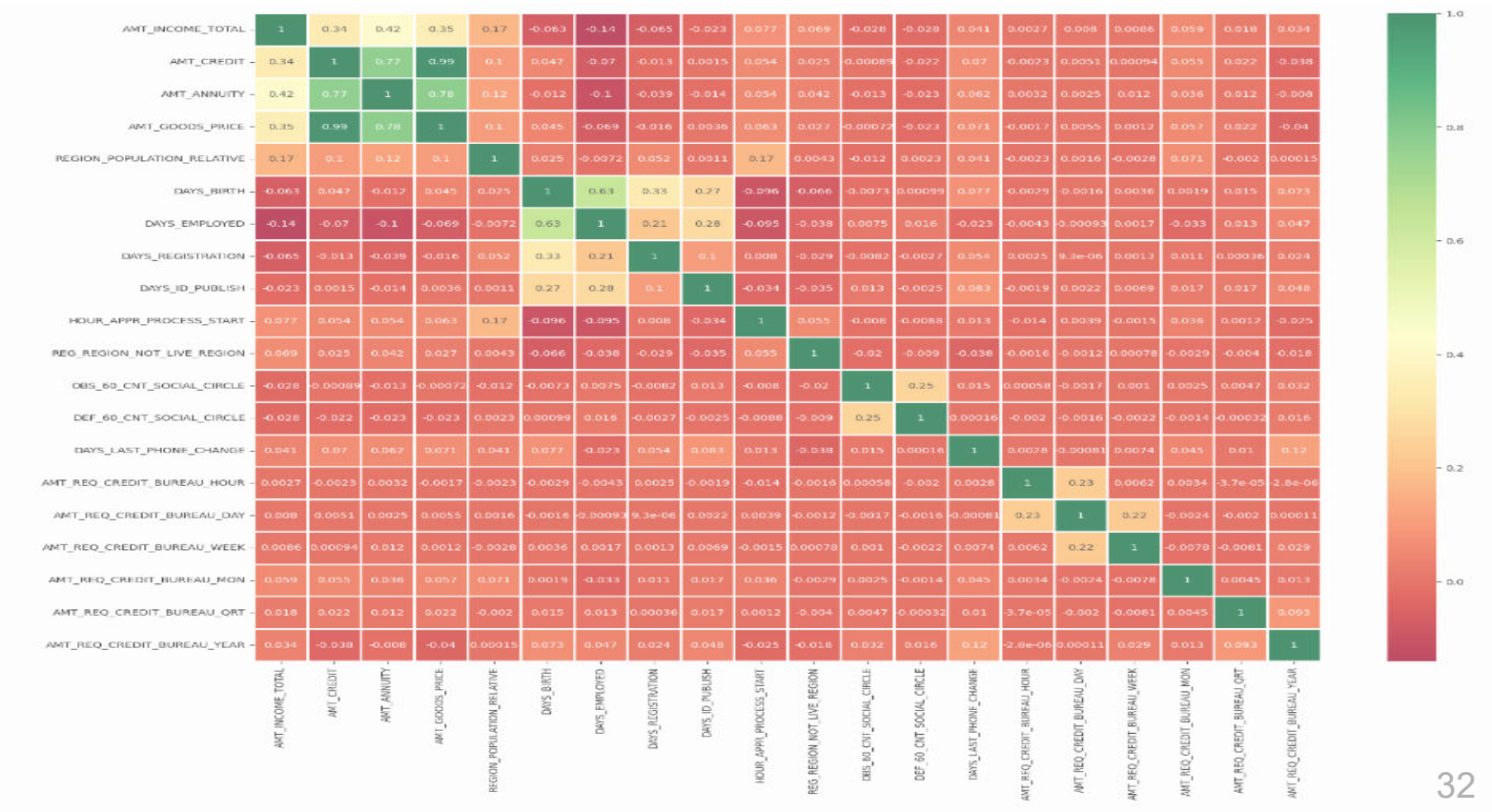


- It can be seen that “Businessman” income is the highest and the estimated range with default 95% confidence level seem to indicate that the income of a Businessman could be in the range of slightly close to 4 lakhs and slightly above 10 lakhs

NUMERIC VARIABLE ANALYSIS

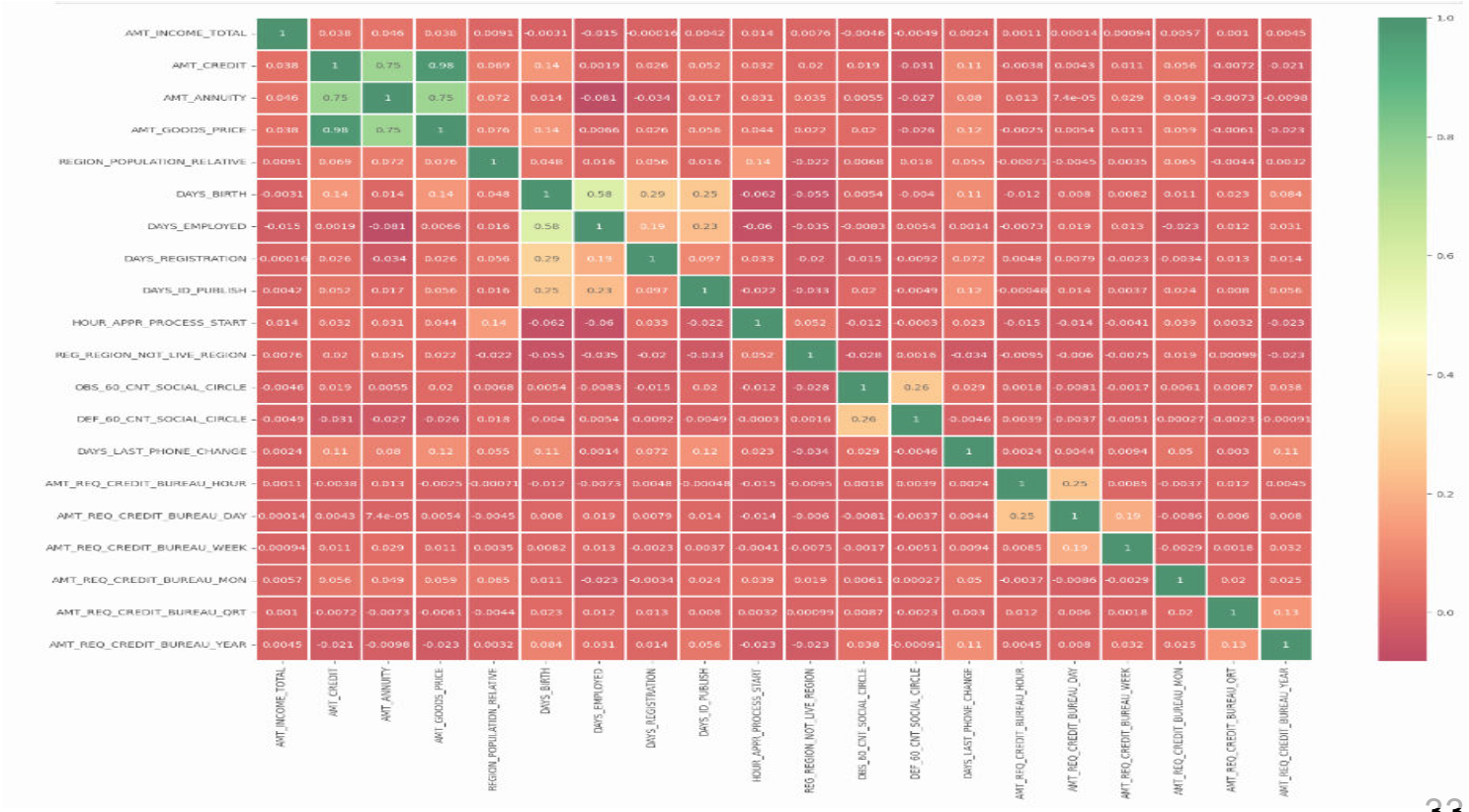
Correlating factors amongst repayers

- 1. Credit amount is highly correlated with:
  - Goods Price Amount
  - Loan Annuity
  - Total Income
- 2. We can also see that repayers have high correlation in number of days employed.



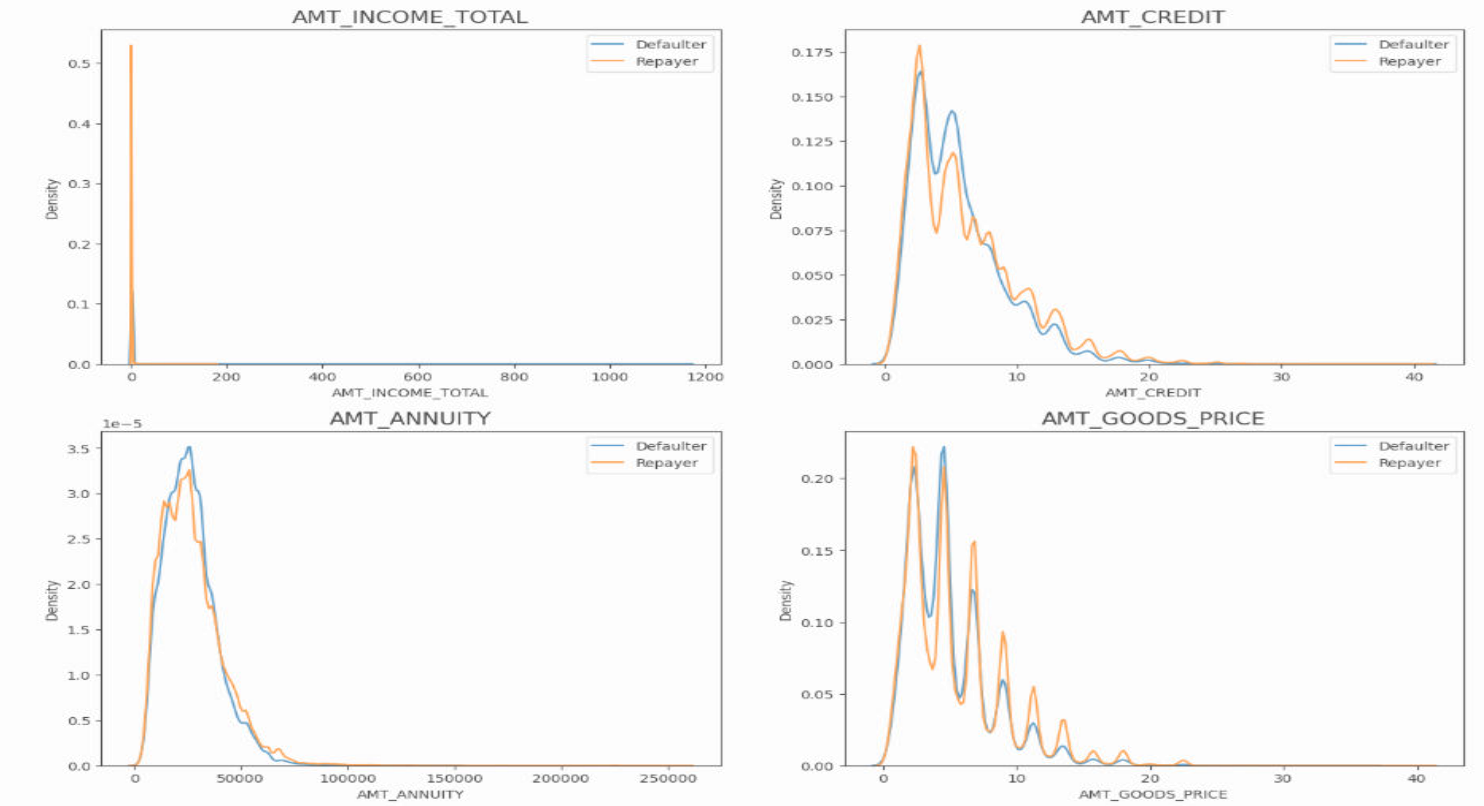
Correlating factors amongst defaulters

- Credit amount is highly correlated with good price amount which is same as repayers.
- Loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to repayers(0.77)
- We can also see that repayers 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 repayers.
- Days_birth and number of children correlation has reduced to 0.259 in defaulters when compared to 0.337 in repayers.
- There is a slight increase in



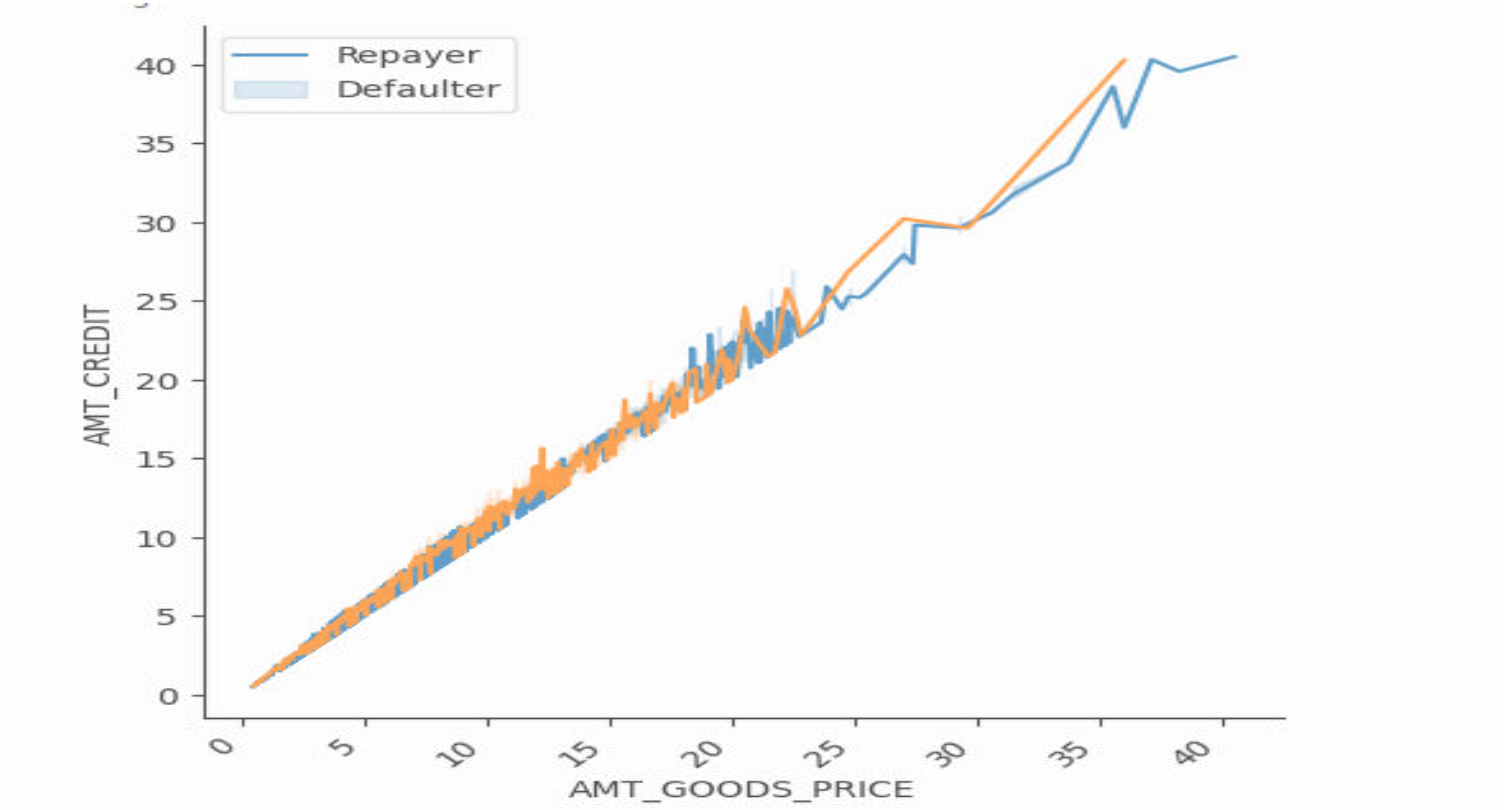
NUMERIC UNI-VARIABLE ANALYSIS

- Most no of loans are given for goods price below 10 lakhs
- Most people pay annuity below 50K for the credit loan
- Credit amount of the loan is mostly less then 10 lakhs
- The repayers and defaulters distribution overlap in all the plots and hence we cannot use any of these variables in isolation to make a decision



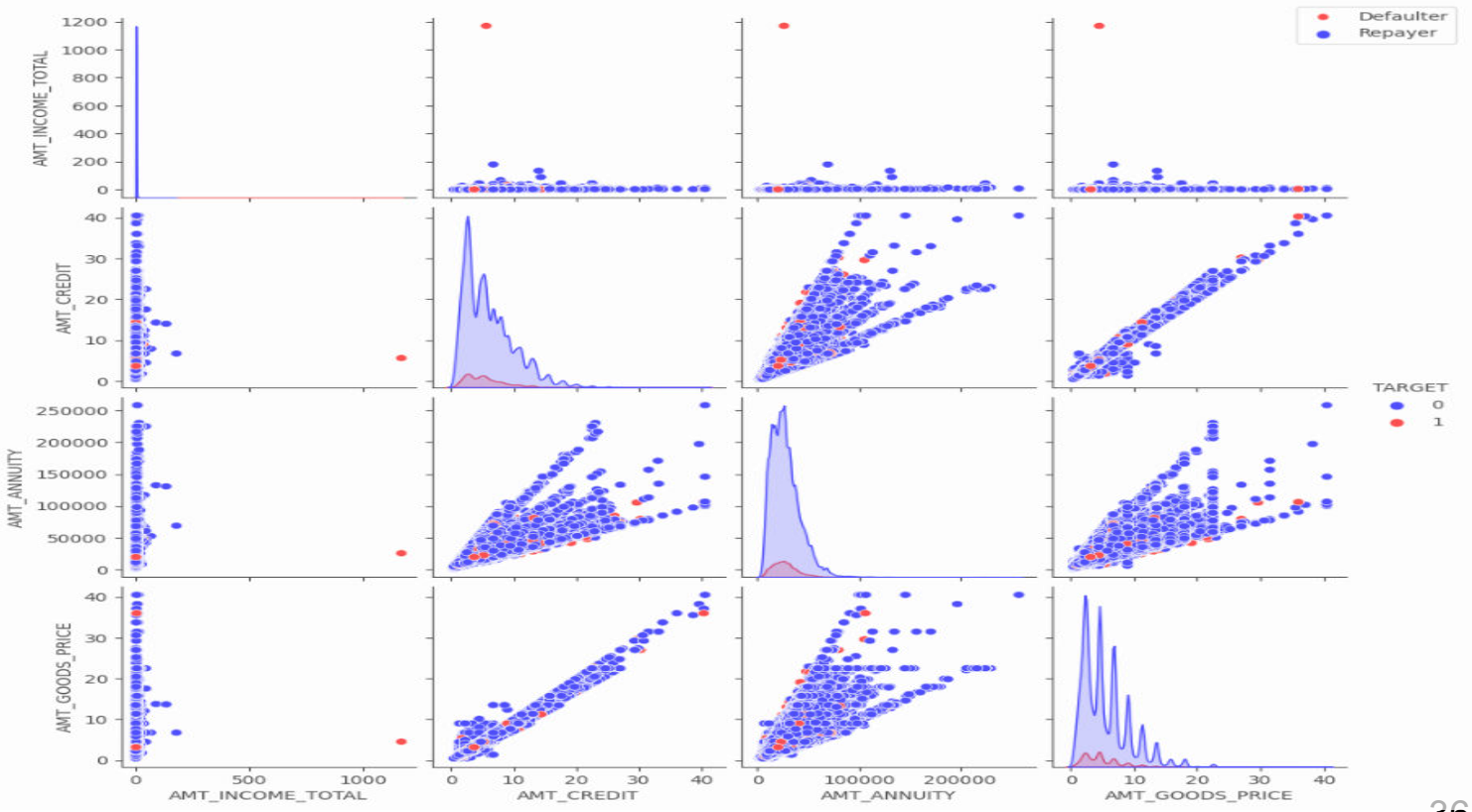
NUMERIC BI-VARIABLE ANALYSIS

- When the credit amount goes beyond 30 Lakhs, there is an increase in defaulters.

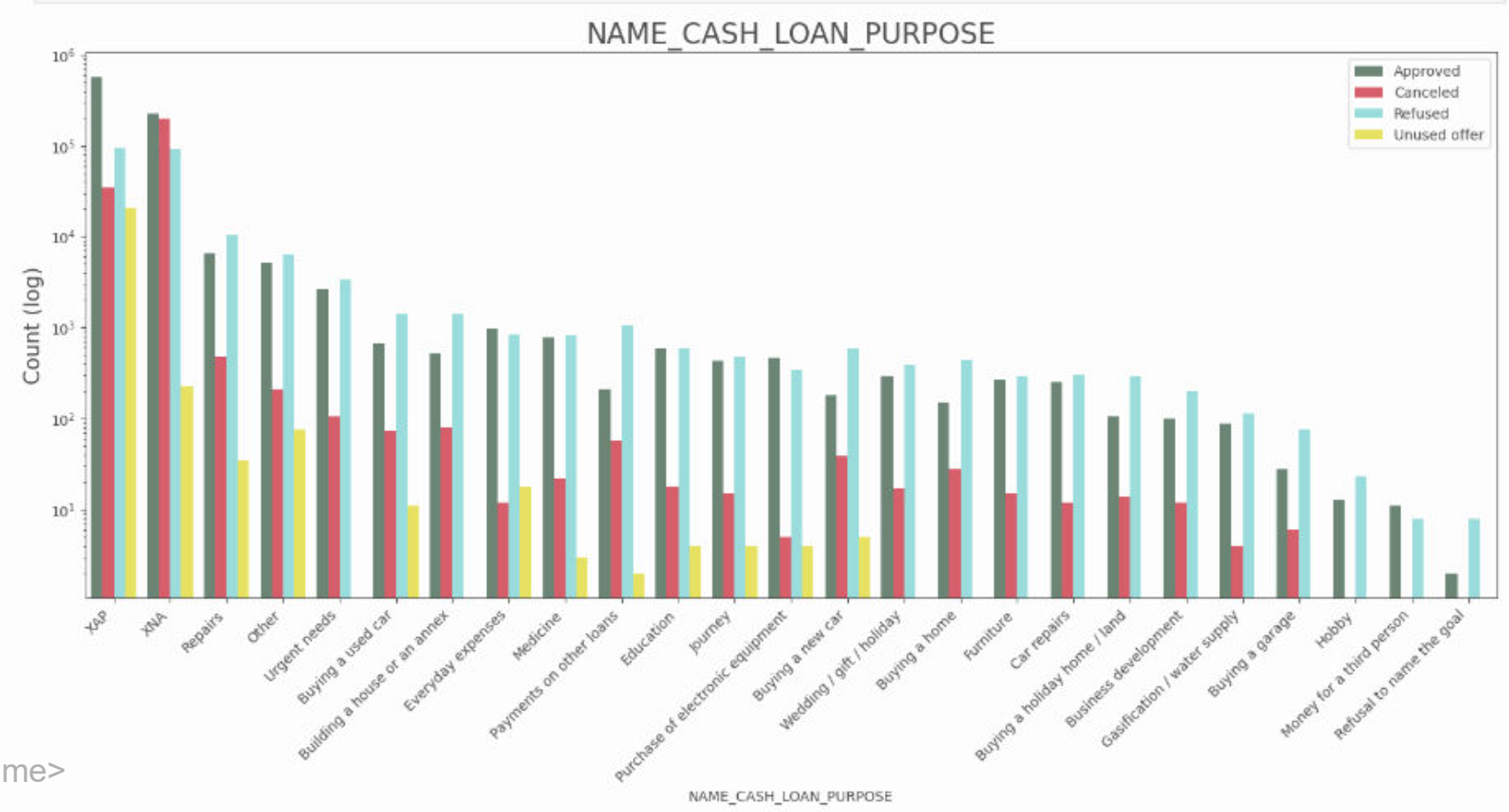


NUMERIC BI-VARIABLE ANALYSIS

- When Annuity Amount > 15K and Good Price Amount > 20 Lakhs, there is a lesser chance of defaulters
- Loan Amount(AMT_CREDIT) and Goods price(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 >20

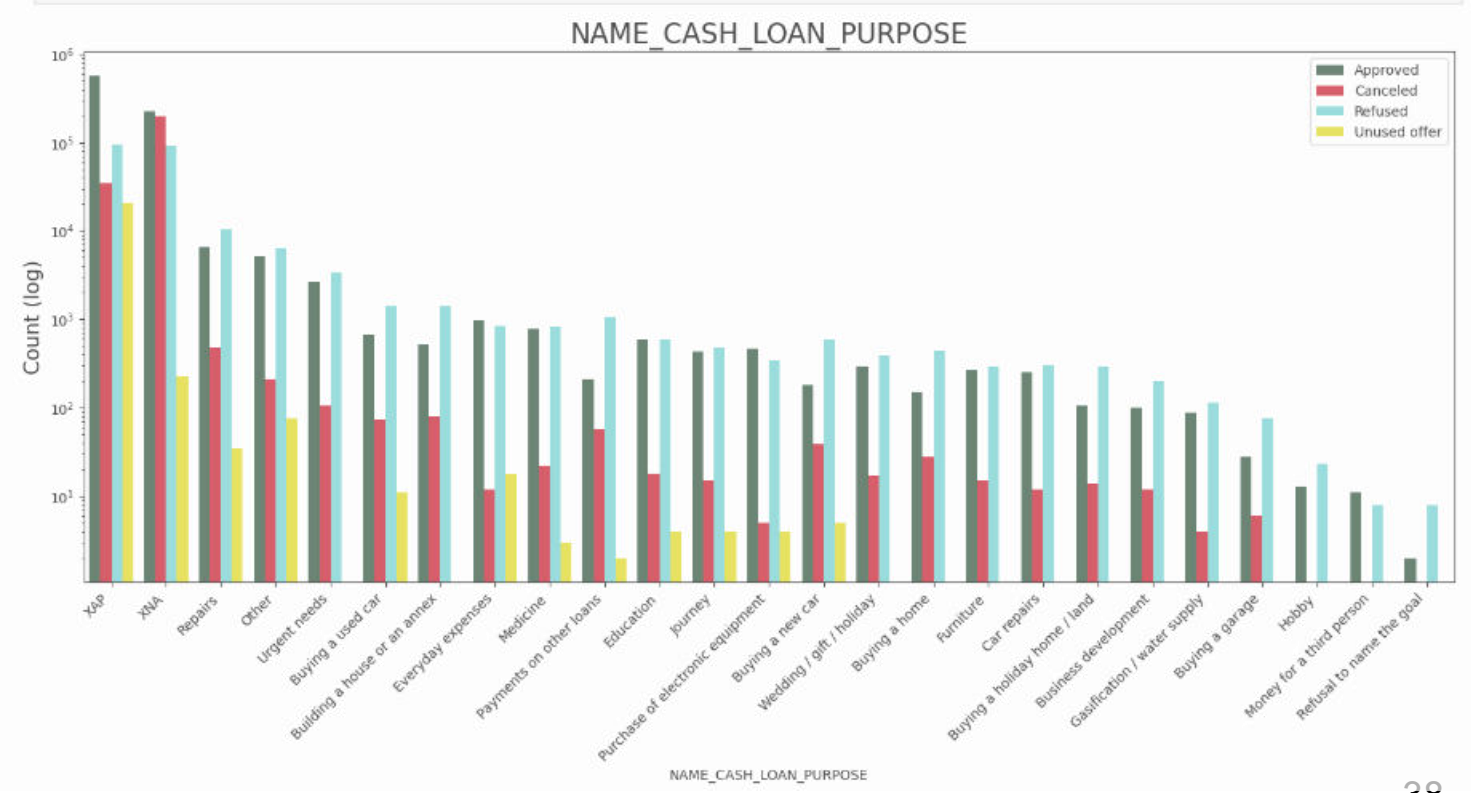


Distribution of Contract Status vs purpose of the loan for Repayer



Distribution of Contract Status vs purpose of the loan for Defaulters

- Loan purpose has high number of unknown values (XAP, XNA)
- Loan taken for the purpose of Repairs looks to have highest default rate
- Huge number application have been rejected by bank or refused by client which are applied for Repair or Other. from this we can infer that repair is considered high risk by bank. Also, either they are rejected or bank offers loan on high interest rate which is not feasible by the clients and



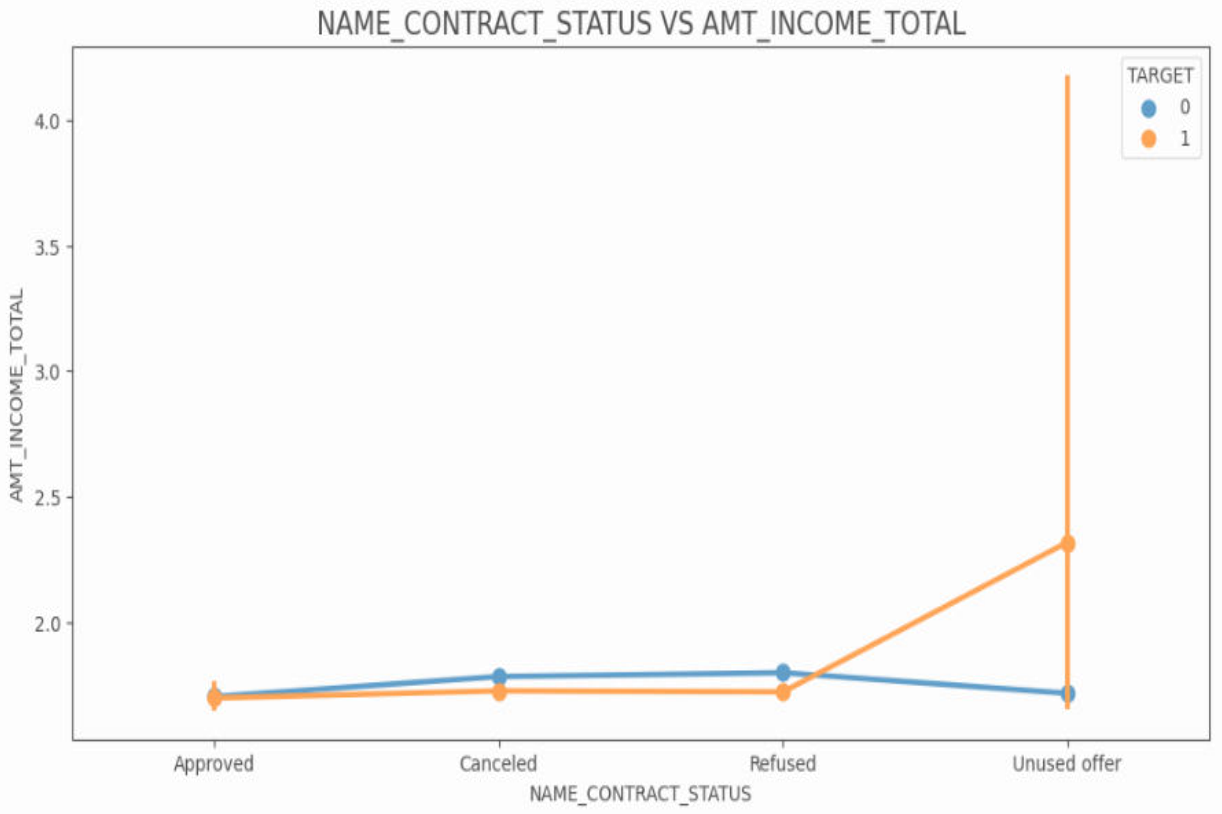
Contract Status based on loan repayment status

- 90% of the previously cancelled client have actually repayed the loan. Revising 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 could turn into potential repaying customer.

NAME_CONTRACT_STATUS	TARGET	Counts	Percentage
Approved	0	818856	92.41%
	1	67243	7.59%
Canceled	0	235641	90.83%
	1	23800	9.17%
Refused	0	215952	88.0%
	1	29438	12.0%
Unused offer	0	20892	91.75%
	1	1879	8.25%

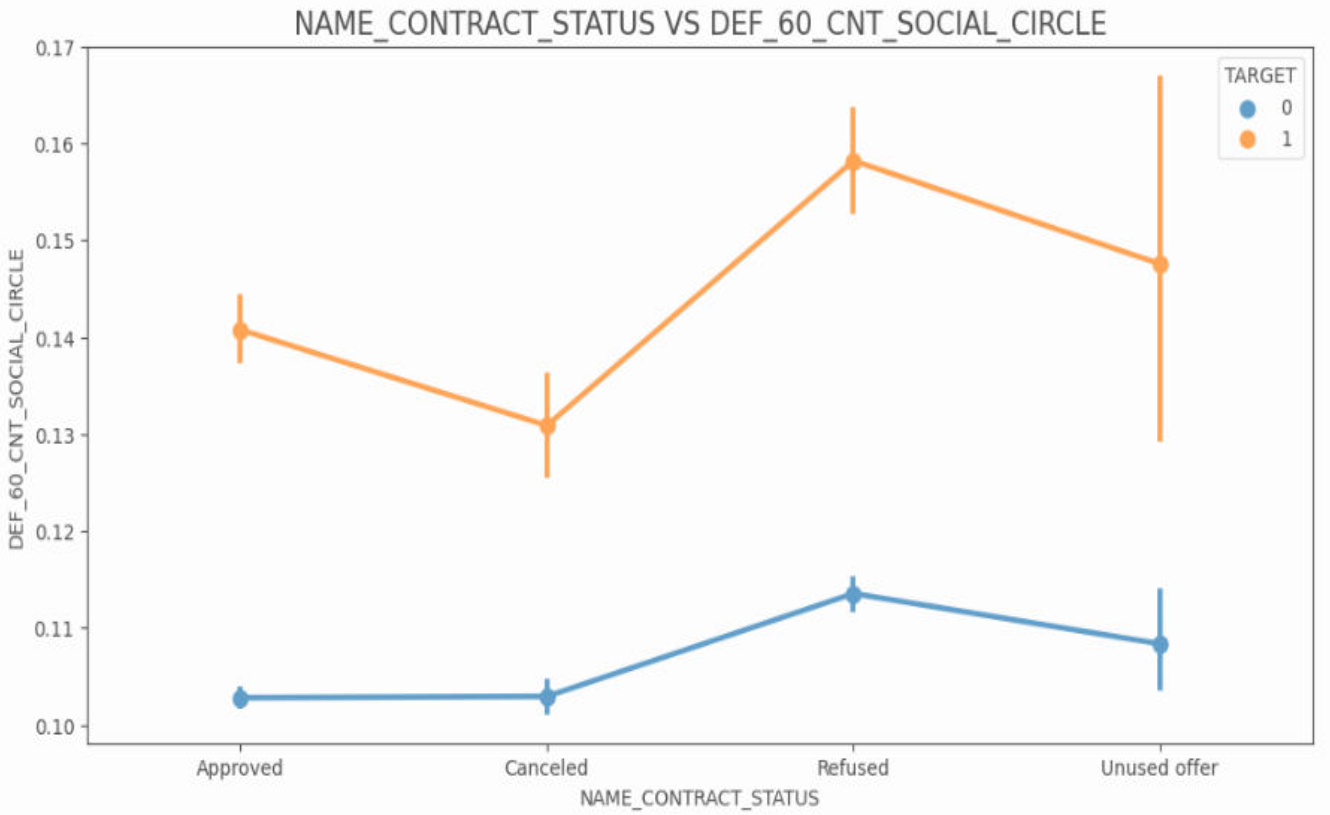
Plotting the relationship between income total and contact status

- 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

- Clients who have average of 0.13 or higher their DEF_60_CNT_SOCIAL_CIRCLE score tend to default more and thus analysing client's social circle could help in disbursment of the loan.



CONCLUSIONS

- **After analysing the datasets, there are few attributes of a client with which the bank would be able to identify if they will repay the loan or not. The analysis is consised as below with the contributing factors and categorization:**
  - 90% of the previously cancelled client have actually repayed the loan. Record the reason for cancellation which might help the bank to determine and negotiate terms with these repaying customers in future for increase business opportunity.
  - 88% of the clients who were refused by bank for loan earlier have now turned into a repaying client. Hence documenting the reason for rejection could mitigate the business loss and these clients could be contacted for further loans.