BANK LOAN CASE STUDY

KEERTHI NANNEPAMULA

PROJECT DESCRIPTION

This project gives us an idea of applying EDA in a real business scenario. It has also been helpful in understanding risk analytics in banking and financial services and how data is crucial to minimize the risk of losing money while lending money to customers through financial bodies.

BUSINESS UNDERSTANDING

The Loan-providing companies find it hard to give loans to people due to their insufficient or non-existent credit history. When working with consumer finance companies that specialize in lending various types of loans to urban customers, we use EDA to analyze the patterns present in the data.

The types of risks associated with the bank's decision when the company decides on the loan approval on the applicant's profile are:

1. Loss of business occurs when a loan is not approved for an applicant, who is likely to repay the loan.
2. Financial loss to the company happens when a loan is approved for a defaulter.

Types of decisions taken by the client/company when a client applies for a loan are:

1. APPROVED: the loan has been granted by the company

2. CANCELLED: the client had cancelled the application at some point in time
3. REFUSED: the loan has been rejected by the company
4. UNUSED OFFER: loan has been cancelled by the client during the application process

BUSINESS OBJECTIVES

1. Identification of patterns that 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, etc.
2. Identification of applicants defaulters using EDA
3. Understanding of driver variables(driving factors) behind loan default is pivotal for the company's portfolio and risk assessment

APPROACH

PROBLEM STATEMENT

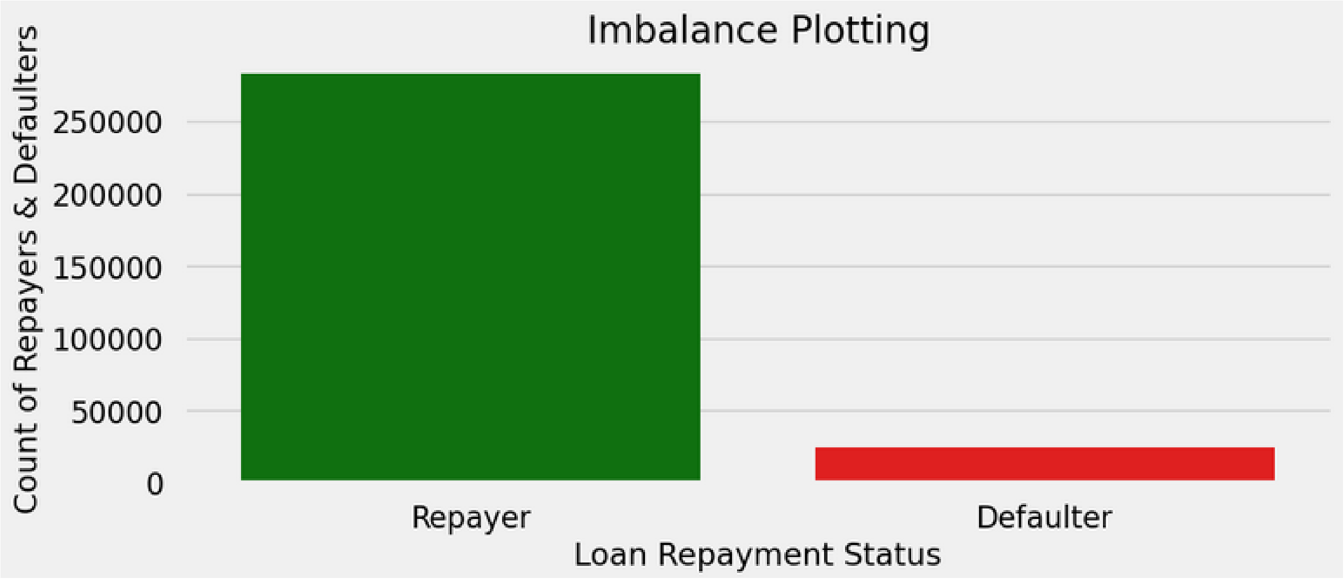
This case study involves the discovery of missing data and the appropriate methods to deal with it. We're also required to identify the outliers in the given dataset and the imbalance with the ratio of imbalance. The results of the univariate, segmented univariate, bivariate, etc, and the top 10 correlation for the client with payment difficulties and all other cases with the necessary visualizations.

ANALYSIS APPROACH

We have applied our basic understanding of risk analysis, EDA, and driving factors to improve the analysis of the case study.

1. DATA CLEANING:
    - Finding and handling missing data
    - Removing unwanted columns
    - Dropping column with high missing value percentage
    - Data manipulation in important columns which cannot be dropped and data with low count of null values
    - Standardizing values

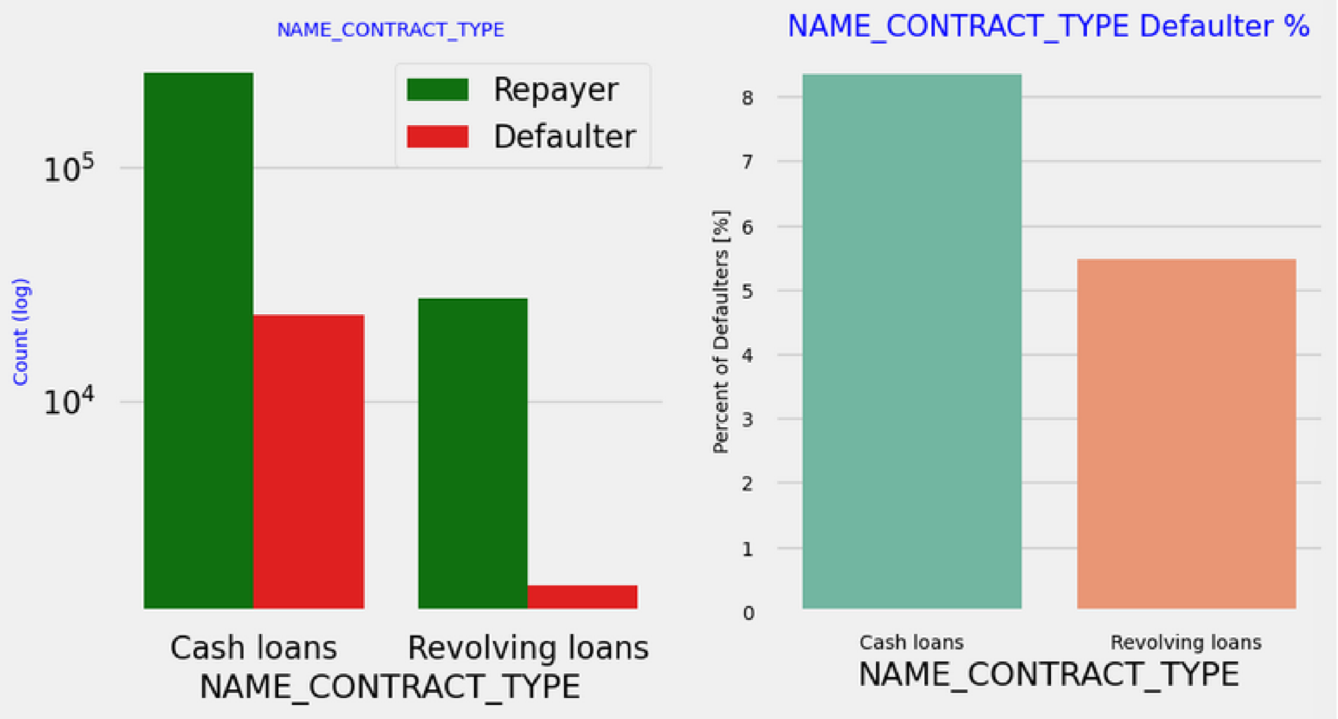
DATA IMBALANCE AND ITS RATIO



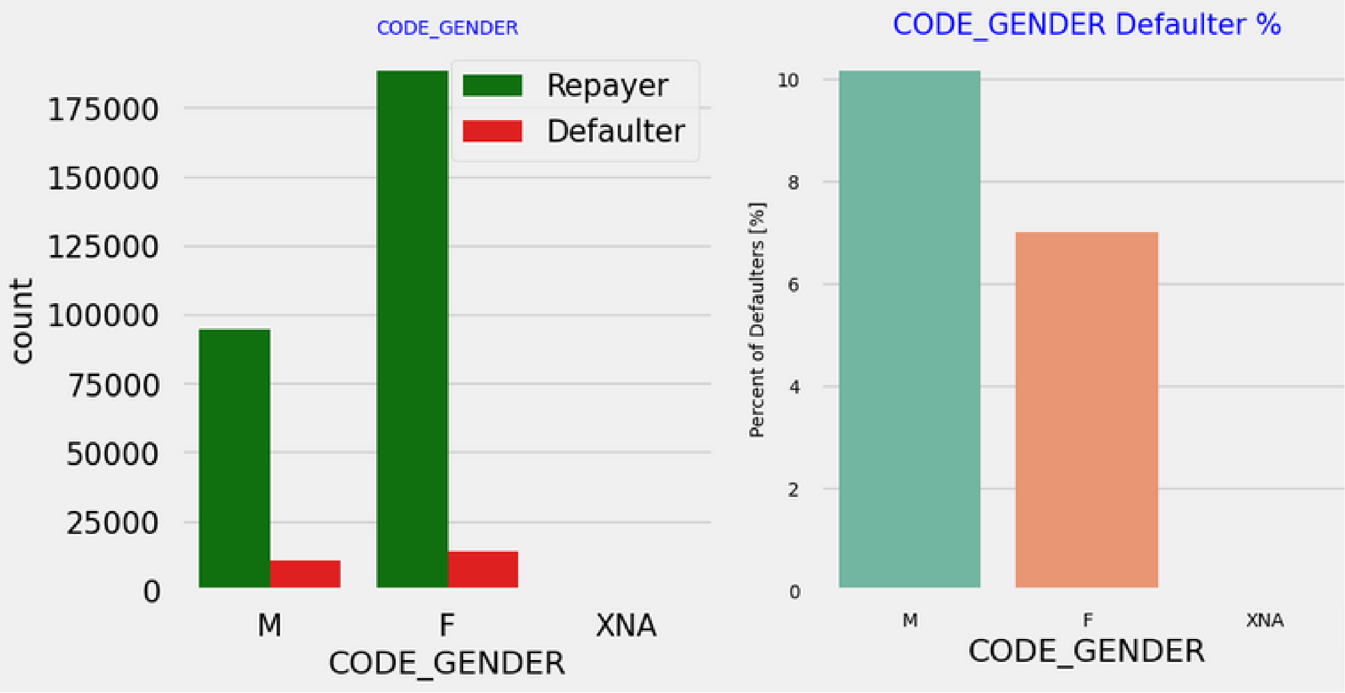
1. **Ratios of imbalance in percentage with respect to Repayer and Defaulter datas are: 91.93 and 8.07**
2. **Ratios of imbalance in relative with respect to Repayer and Defaulter datas is 11.39 : 1 (approx)**

RESULTS OF UNIVARIATE, SEGMENTED UNIVARIATE, BIVARIATE ANALYSIS, ETC IN BUSINESS TERMS

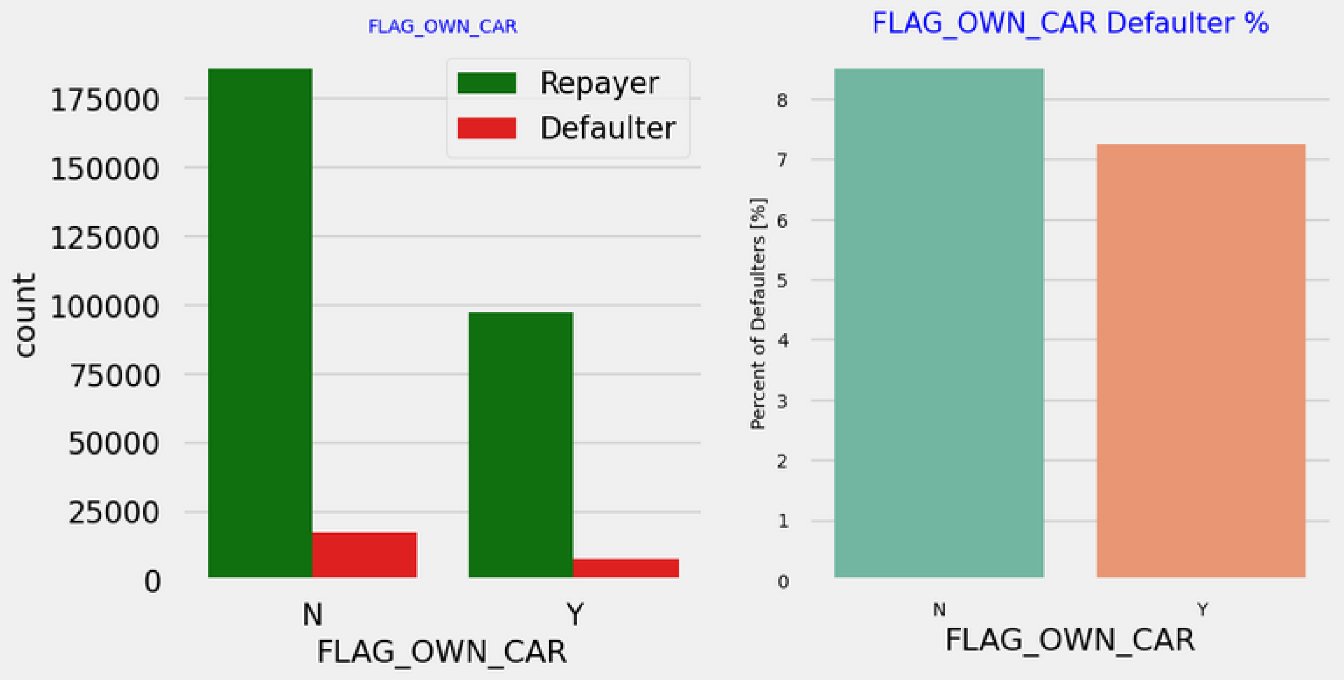
UNIVARIATE ANALYSIS



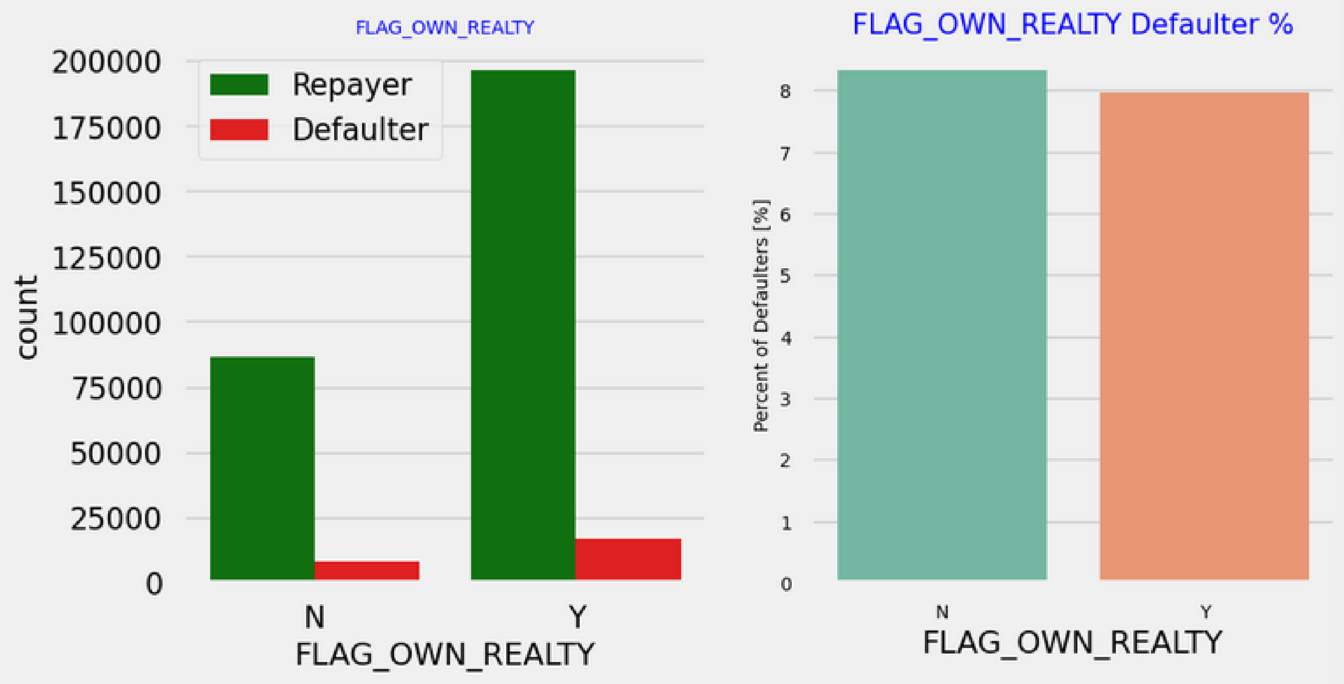
Revolving loans are just a small fraction (10%) from the total number of loans; in the same time, a larger amount of Revolving loans, comparing with their frequency, are not repaid.



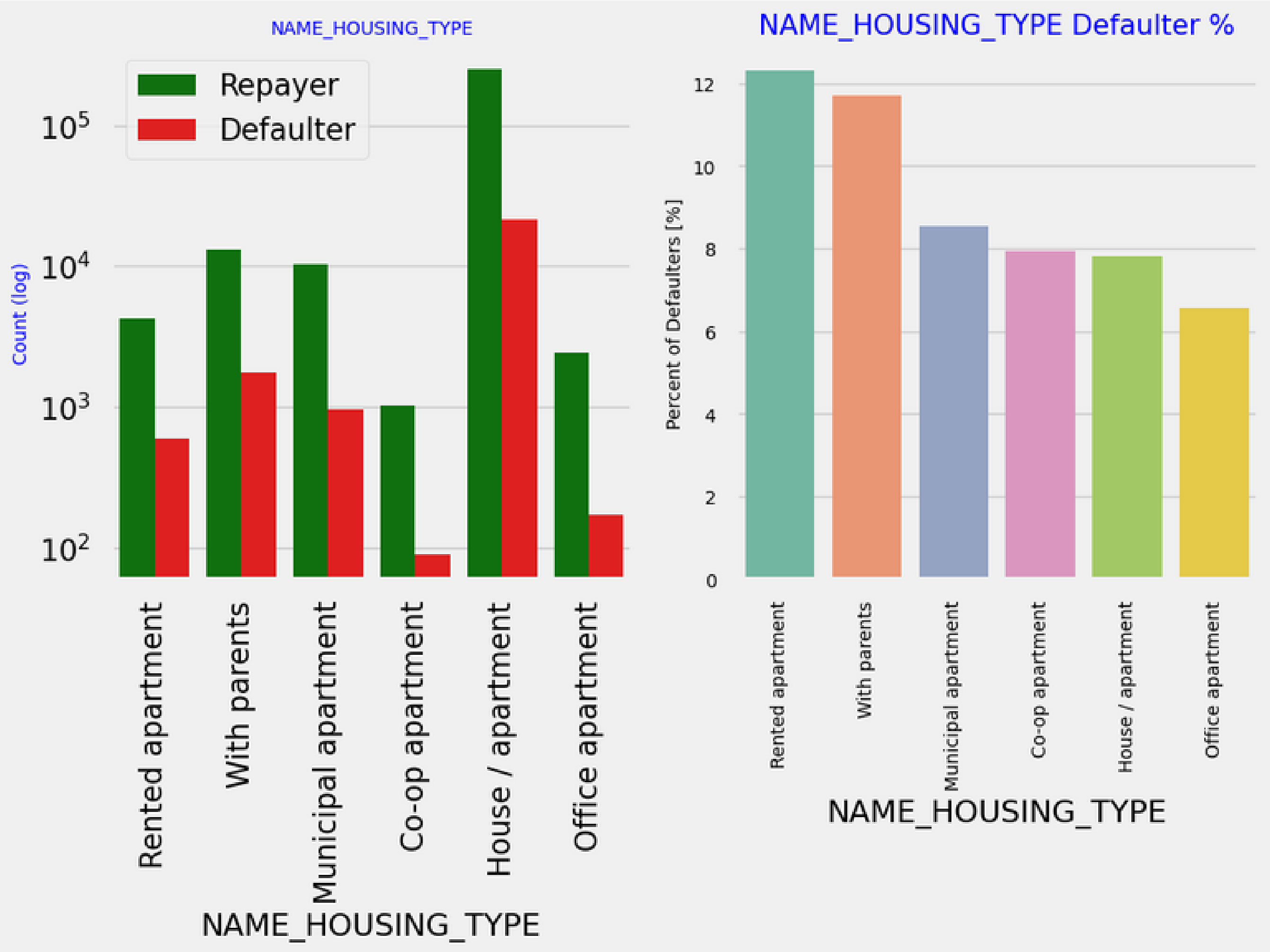
The number of female clients is almost double the number of male clients. Based on the percentage of defaulted credits, males have a higher chance of not returning their loans (~10%), comparing with women (7%)



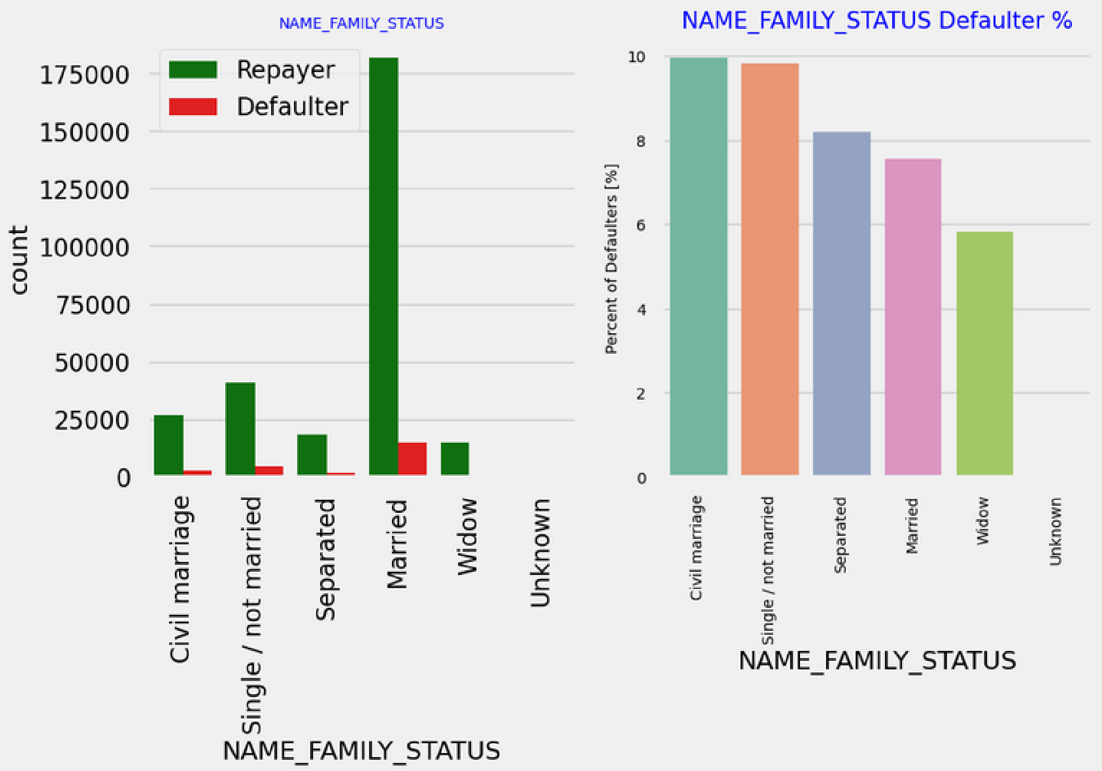
Clients who own a car are half in number of the clients who dont own a car. But based on the percentage of deault, there is no correlation between owning a car and loan repayment as in both cases the default percentage is almost same



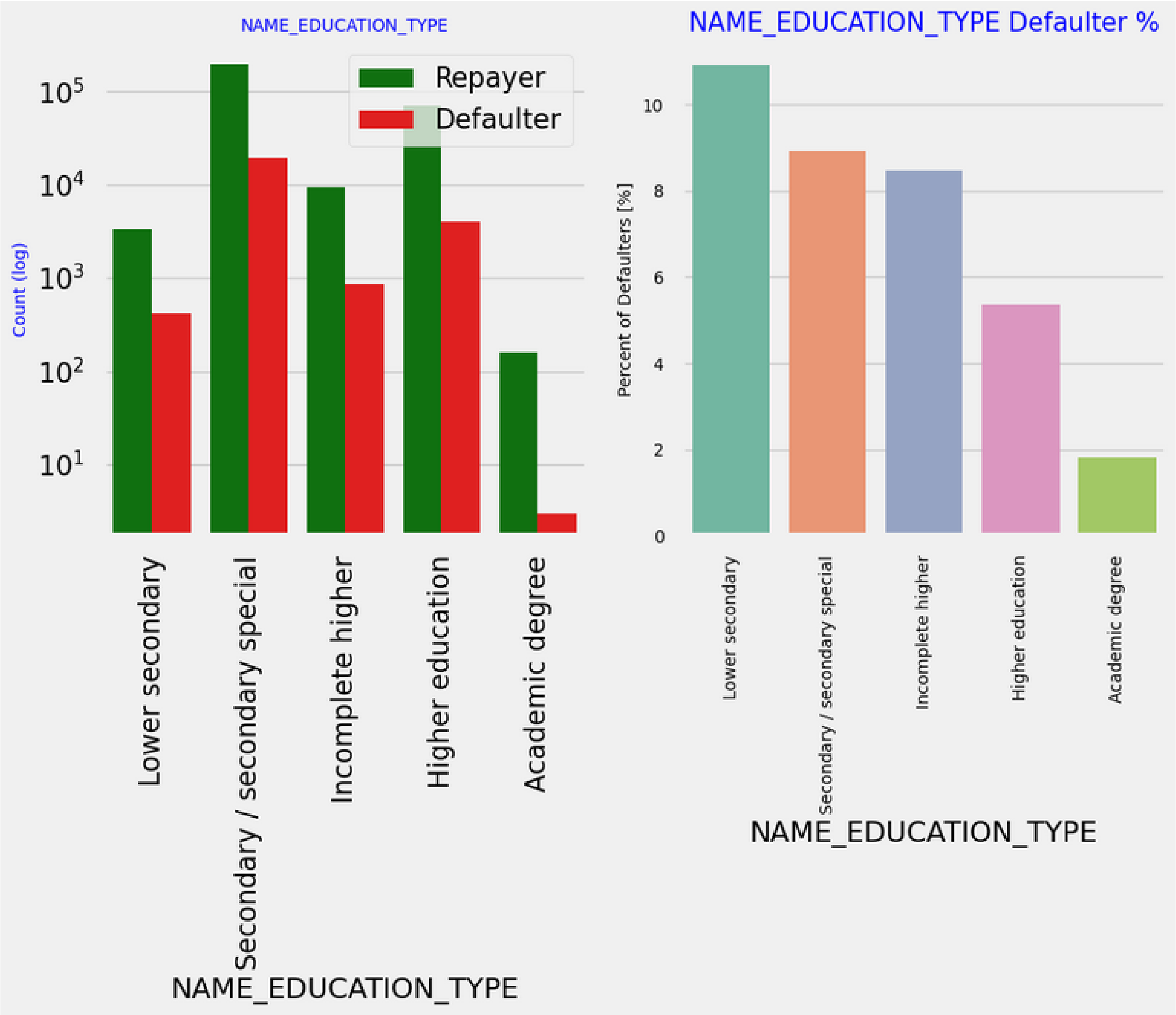
The clients who own real estate are more than double of the ones that don't own. But the defaulting rate of both categories are around the same (~8%). Thus there is no correlation between owning a reality and defaulting the loan.



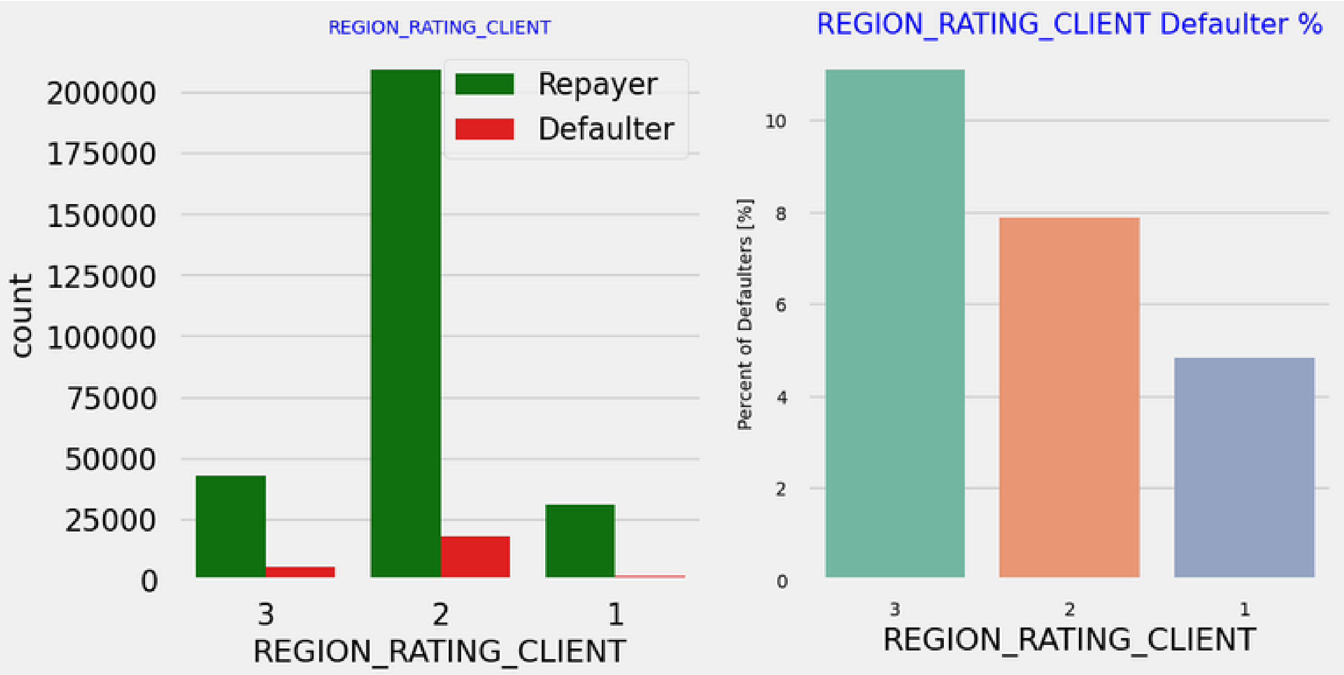
Majority of people live in House/apartment
People living in office apartments have lowest default rate
People living with parents (~11.5%) and living in rented apartments(>12%) have higher probability of defaulting



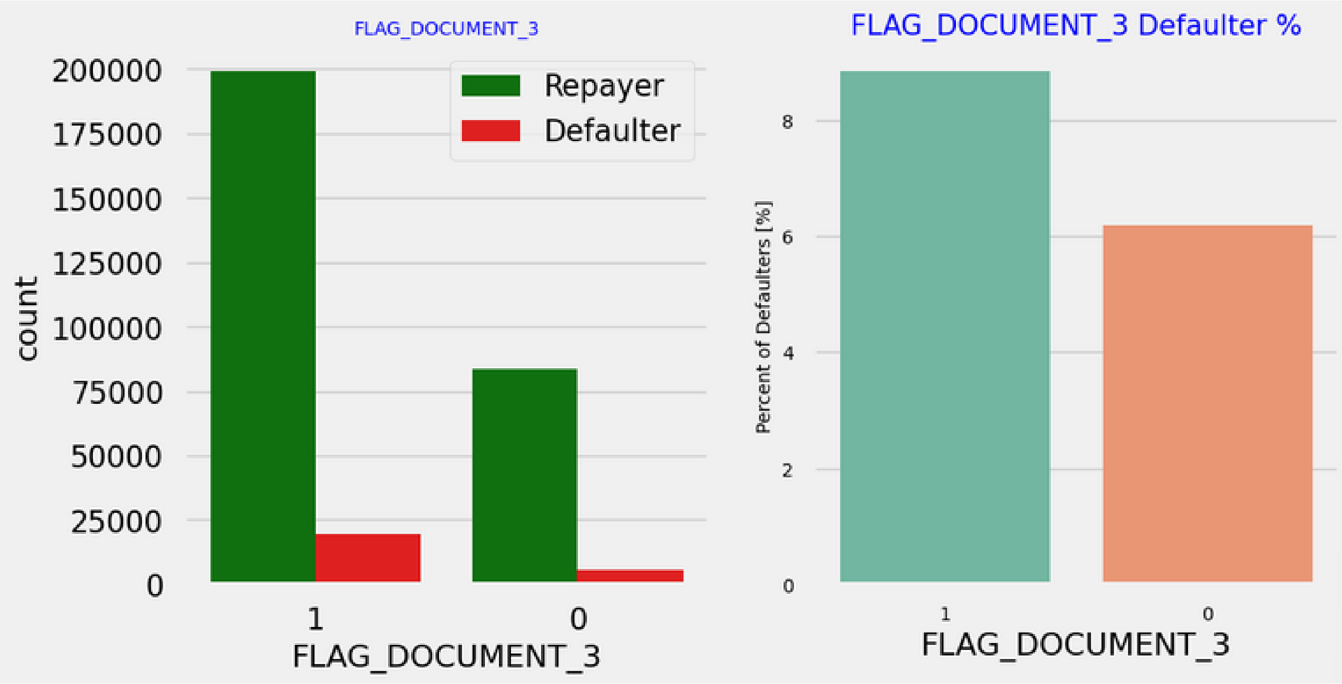
Most of the people who have taken loan are married, followed by Single/not married and civil marriage
In terms of percentage of not repayment of loan, Civil marriage has the highest percent of not repayment (10%), with Widow the lowest (exception being Unknown).



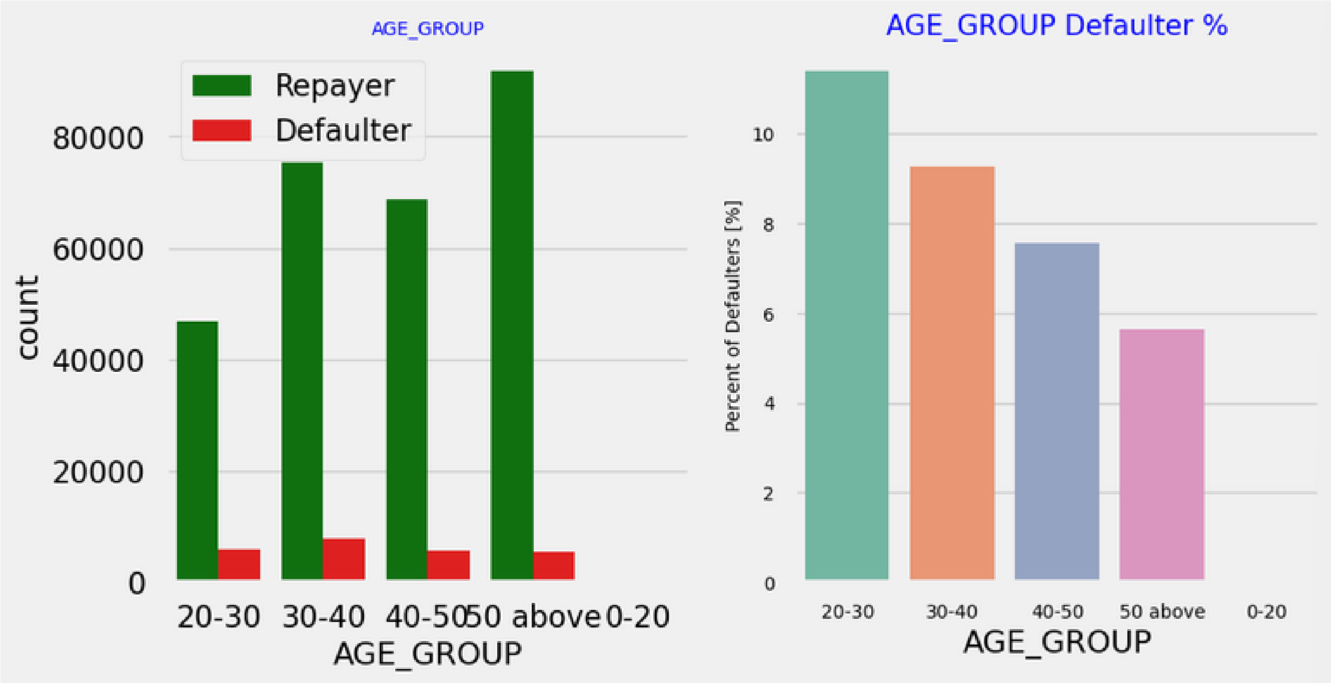
Majority of the clients have Secondary / secondary special education, followed by clients with Higher education. Only a very small number having an academic degree
The Lower secondary category, although rare, have the largest rate of not returning the loan (11%). The people with Academic degree have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



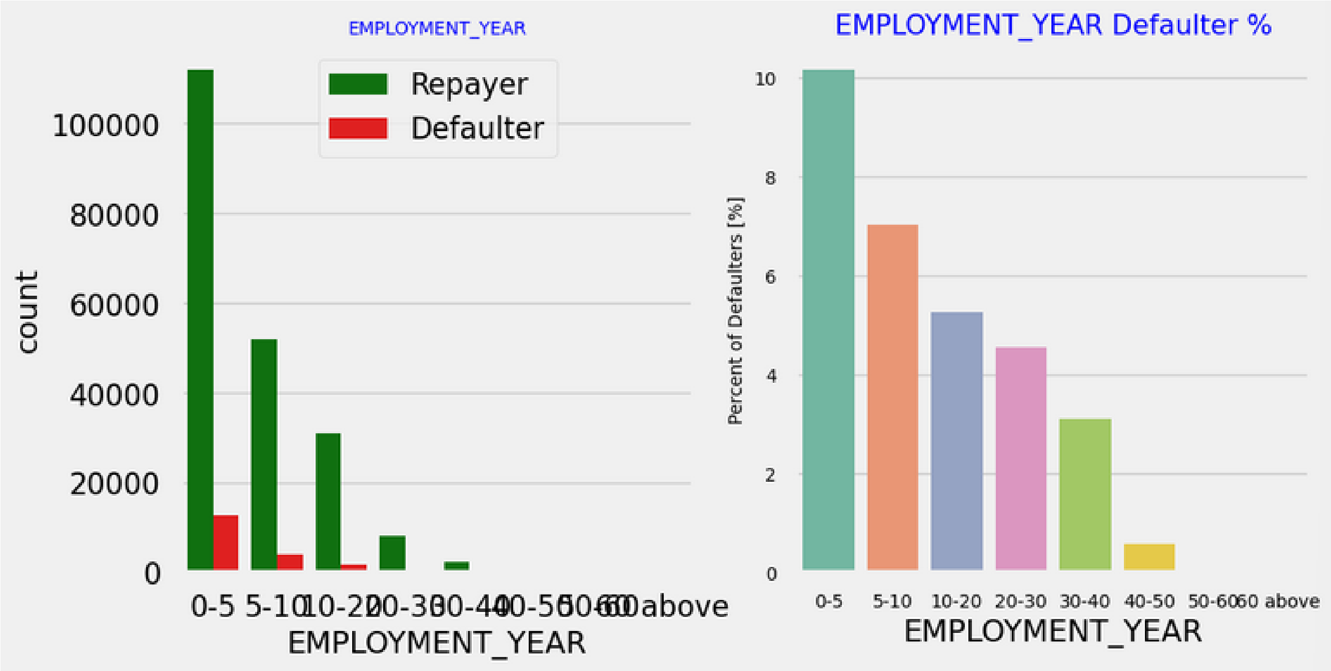
1. Most of the applicants are living in Region_Rating 2 place.
2. Region Rating 3 has the highest default rate (11%)
3. Applicant living in Region_Rating 1 has the lowest probability of defaulting, thus safer for approving loans



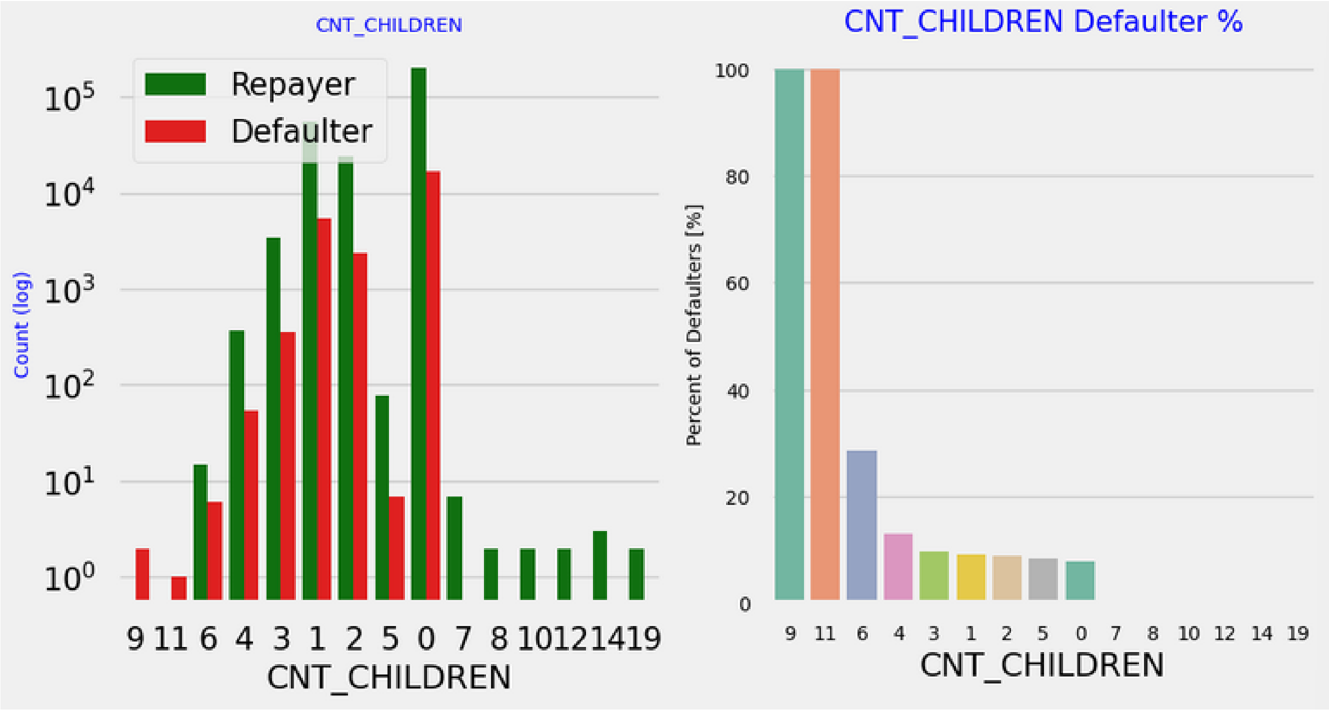
There is no significant correlation between repayers and defaulters in terms of submitting document 3 as we see even if applicants have submitted the document, they have defaulted a slightly more (~9%) than who have not submitted the document (6%)



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

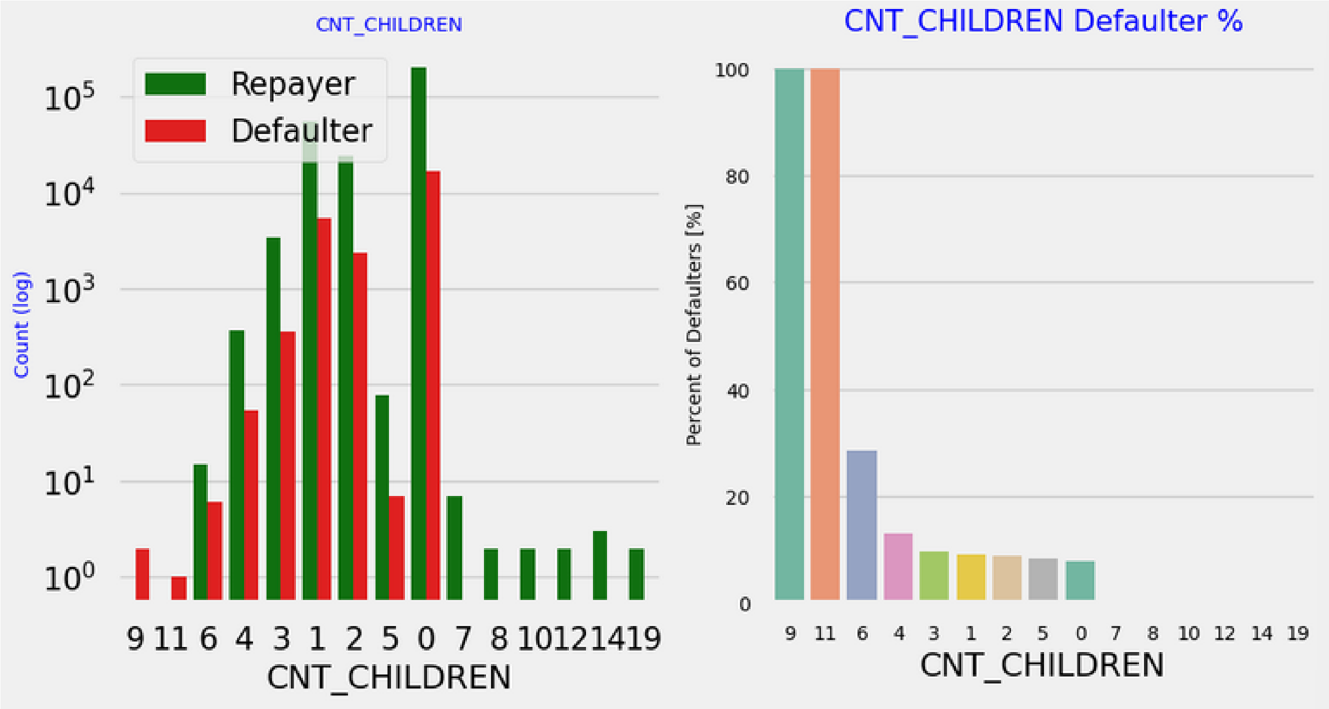


1. Majority of the applicants have been employeed in between 0-5 years. The defaulting rating of this group is also the highest which is 10%
2. With increase of employment year, defaulting rate is gradually decreasing with people having 40+ year experience having less than 1% default rate



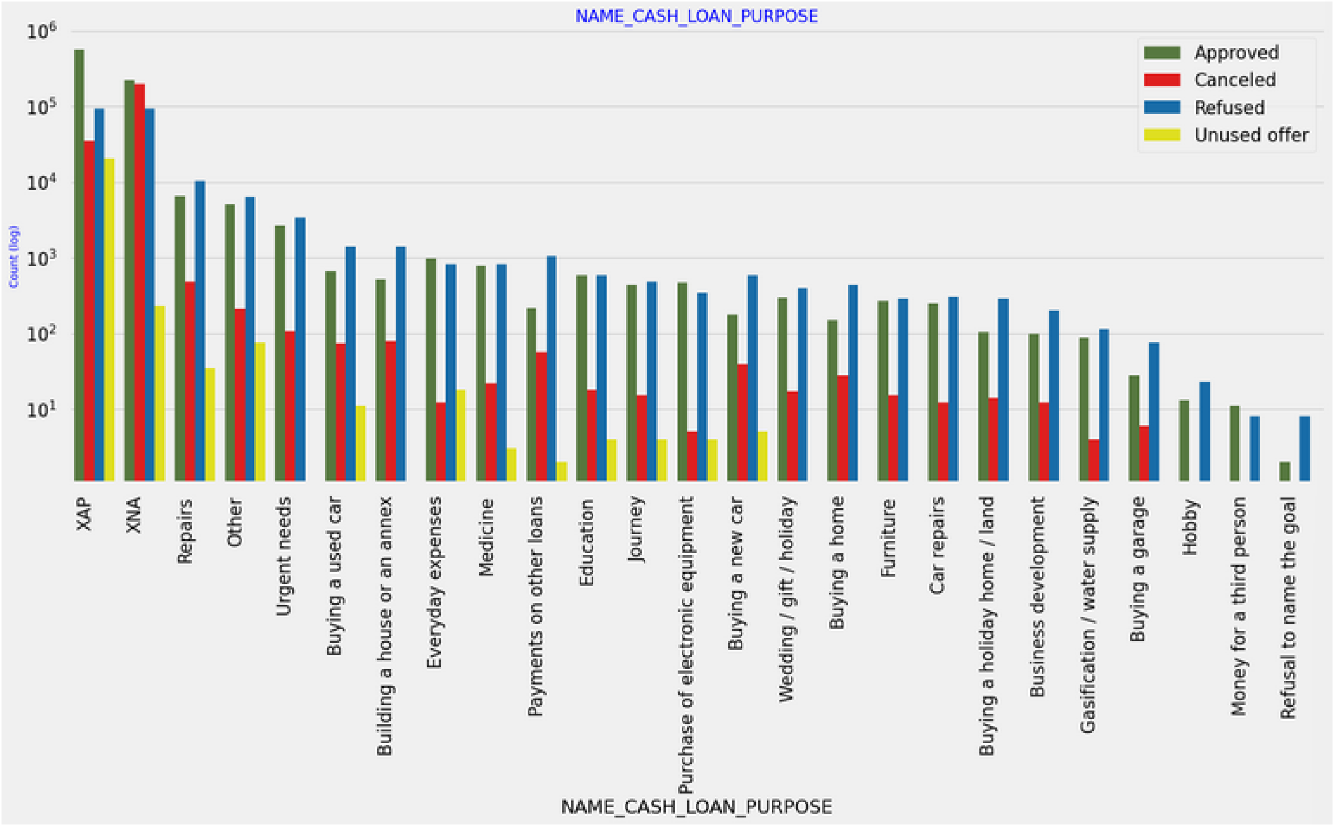
1. **Most of the applicants do not have children**
2. **Very few clients have more than 3 children.**
3. **Client who have more than 4 children has a very high default rate with child count 9 and 11 showing 100% default rate.**

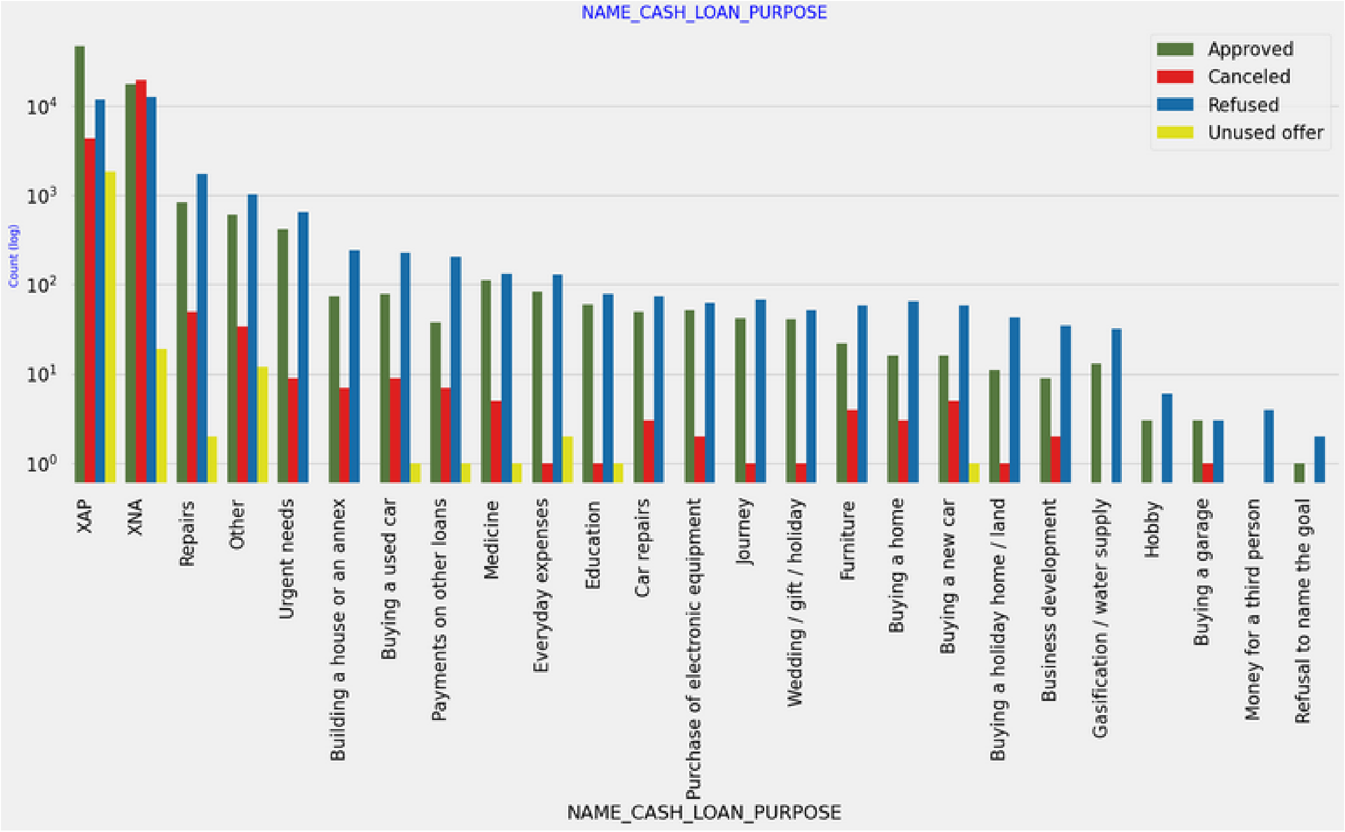
BIVARIATE ANALYSIS



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.

MERGED DATA FRAMES ANALYSIS



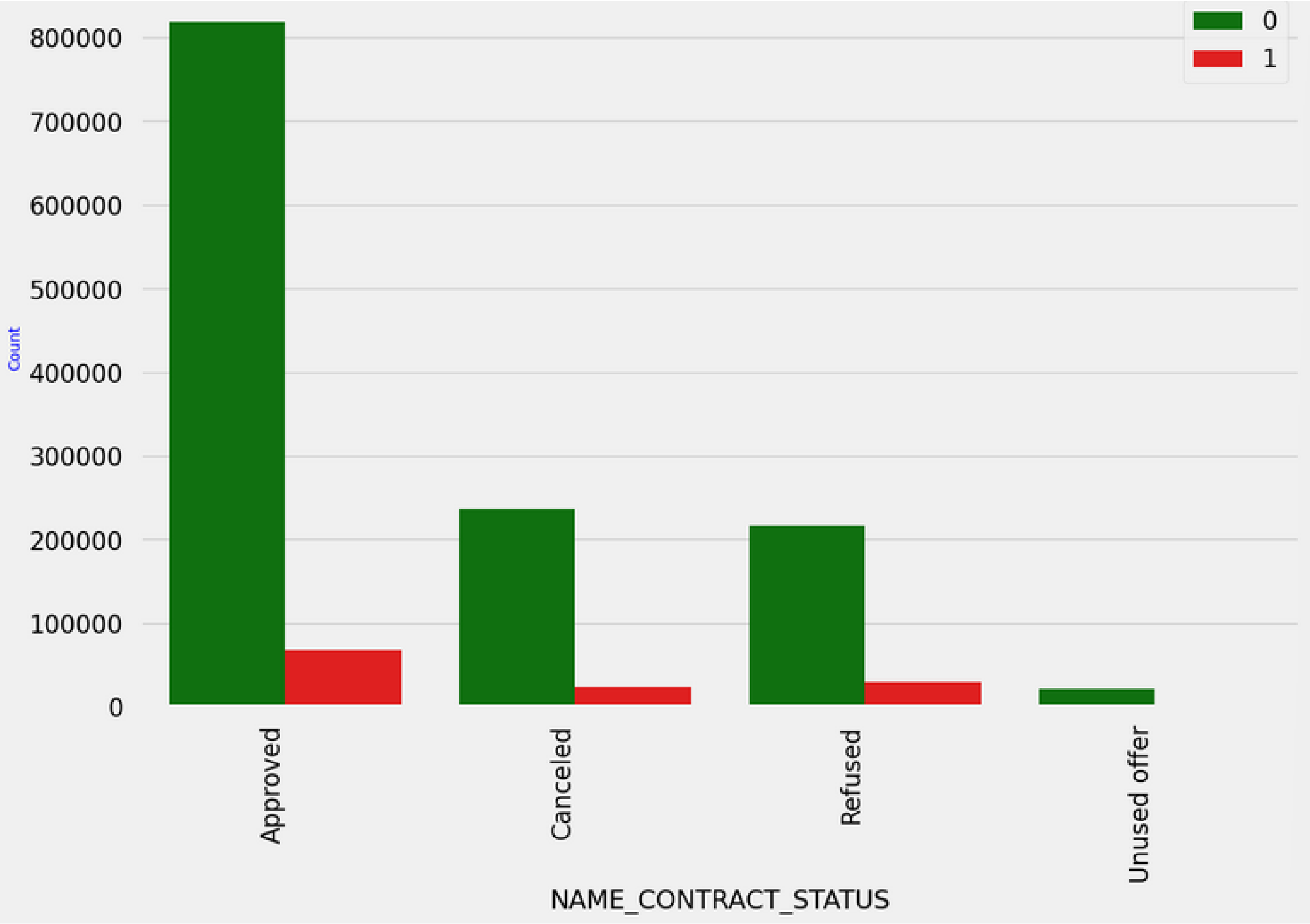


1. **Loan purpose has high number of unknown values (XAP, XNA)**

2. **Loan taken for the purpose of Repairs seems to have highest default rate**

3. **A very high number application have been rejected by bank or refused by client which has purpose as repair or other.**

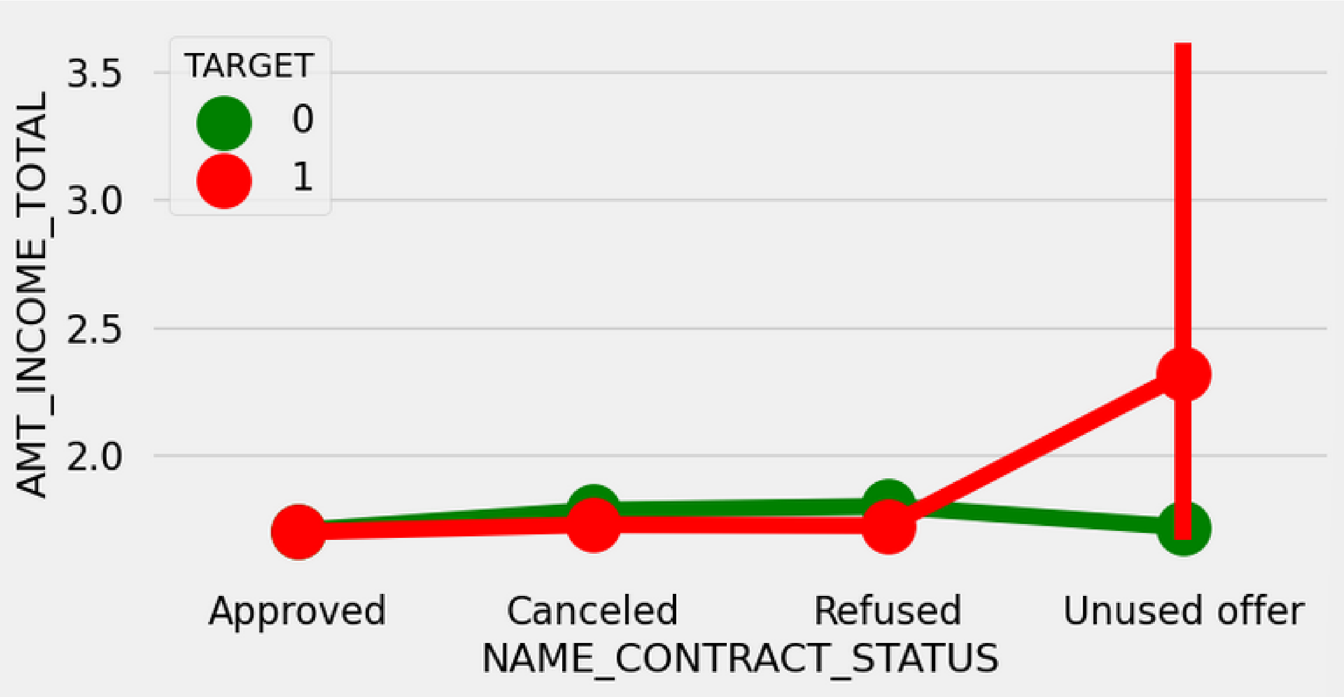
4. **This shows that purpose repair is taken as high risk by bank and either they are rejected or bank offers very high loan interest rate which is not feasible by the clients, thus they refuse the loan.**



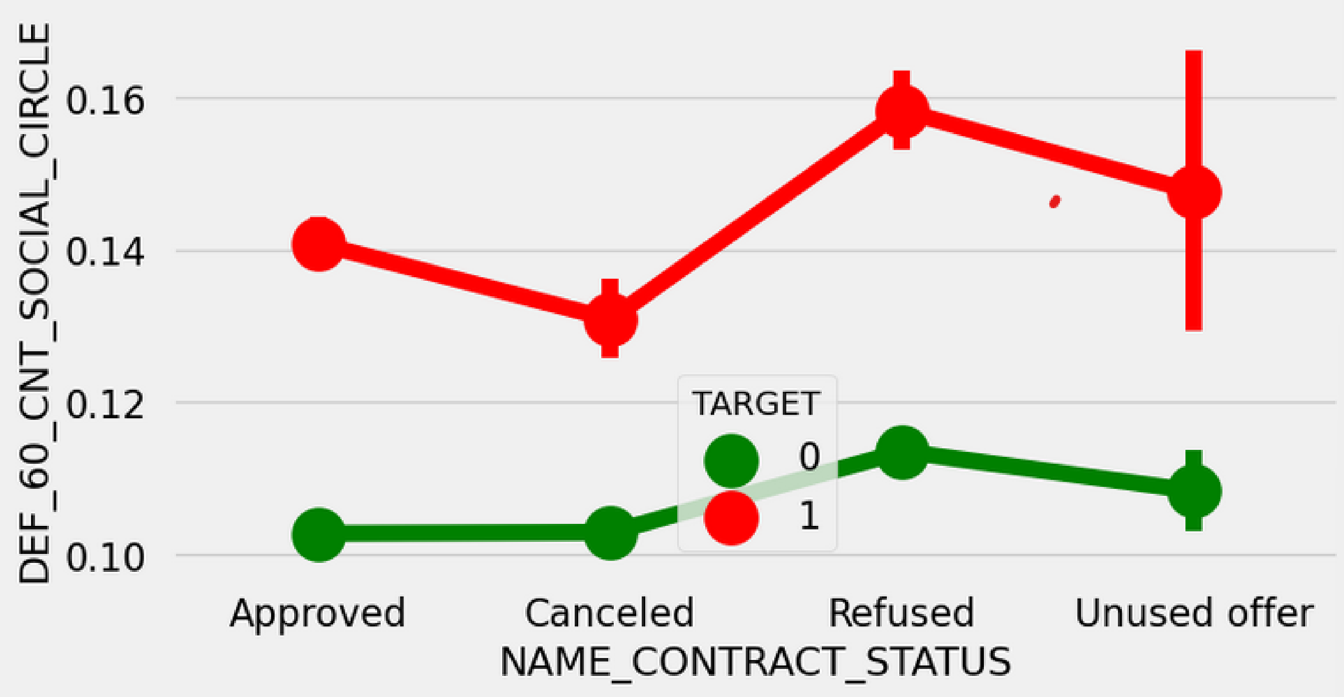
1. **90% of the previously cancelled client have actually repayed the loan. Revisiting the interest rates would increase business opoortunity for these clients**

2. **88% of the clients who have been previously refused a loan has payed back the loan in current case.**

3. **Refusal reason should be recorded for further analysis as these clients would turn into potential repaying customer**

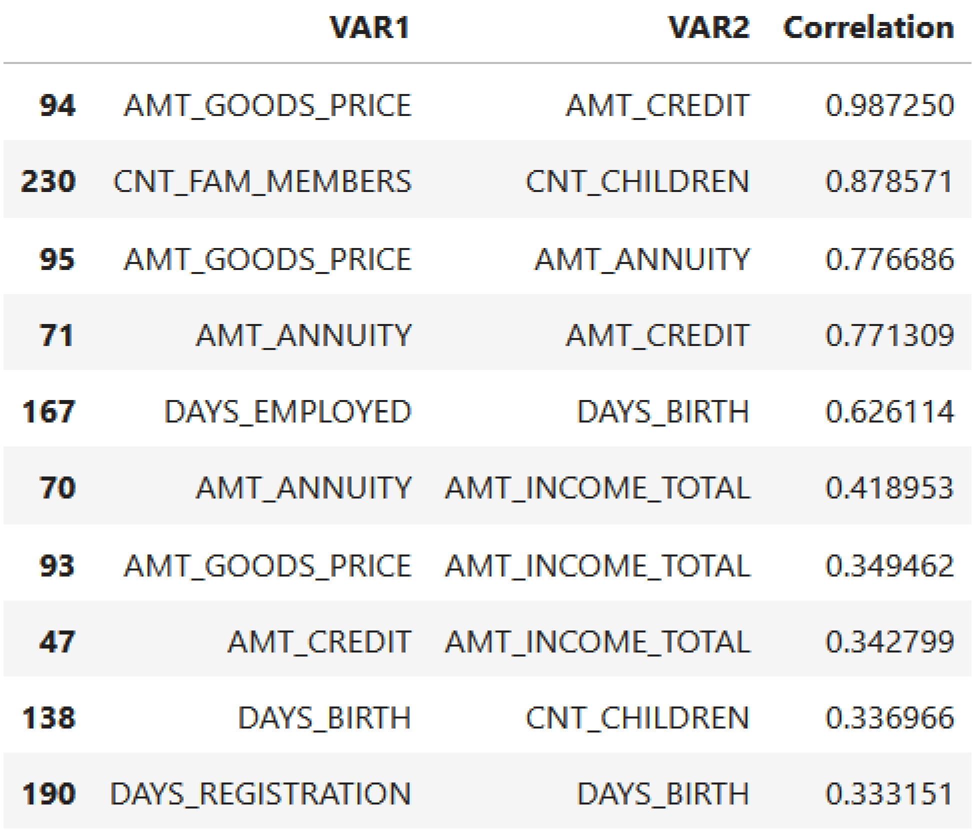


The point plot show that the people who have not used offer earlier have defaulted even when there average income is higher than others

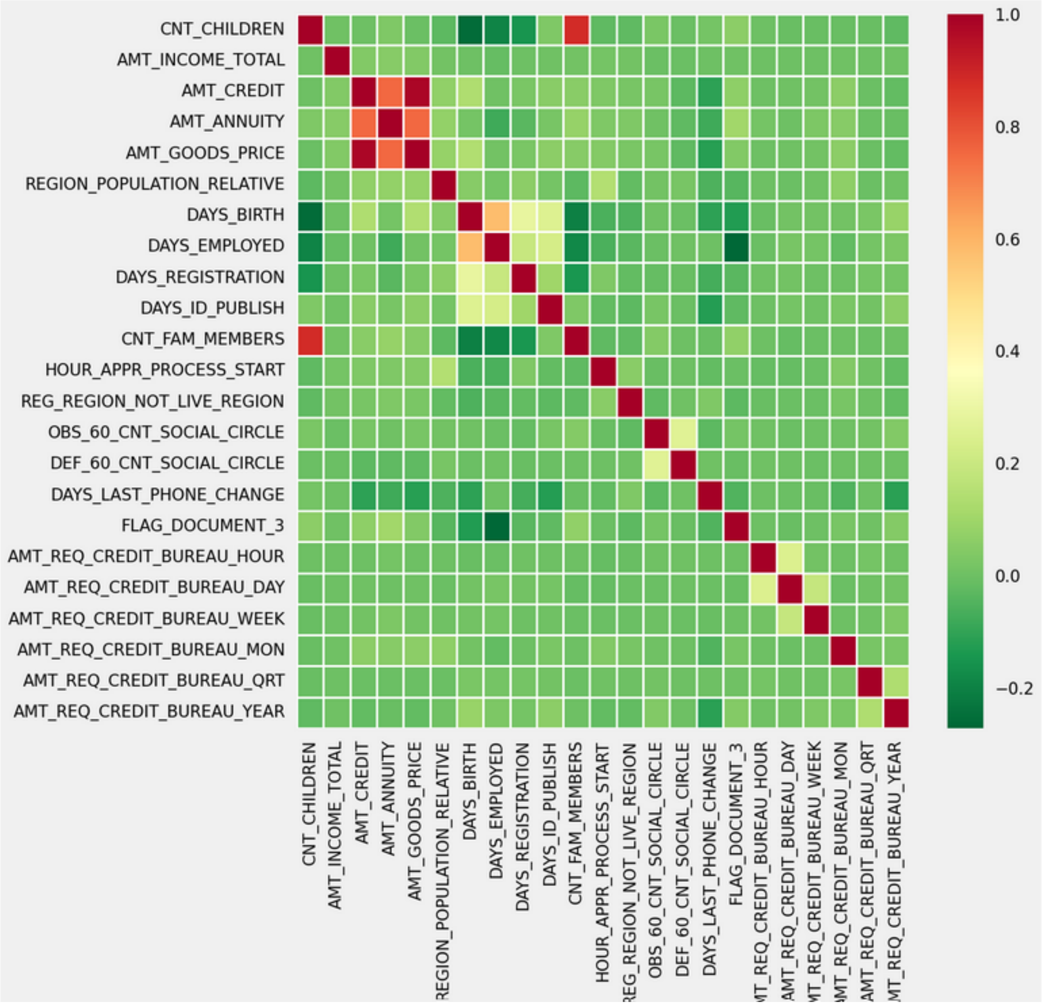


Clients who have average of 0.13 or higher DEF_60_CNT_SOCIAL_CIRCLE score tend to default more and hence client's social circle has to be analysed before providing the loan.

TOP 10 CORRELATION FOR THE CLIENT WITH PAYMENT DIFFICULTIES AND ALL OTHER CASES(TARGET VARIABLE)



	VAR1	VAR2	Correlation
94	AMT_GOODS_PRICE	AMT_CREDIT	0.987250
230	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
95	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686
71	AMT_ANNUITY	AMT_CREDIT	0.771309
167	DAYS_EMPLOYED	DAYS_BIRTH	0.626114
70	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418953
93	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349462
47	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799
138	DAYS_BIRTH	CNT_CHILDREN	0.336966
190	DAYS_REGISTRATION	DAYS_BIRTH	0.333151



Credit amount is highly correlated with amount of goods price which is same as repayers.

But the loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to repayers(0.77)

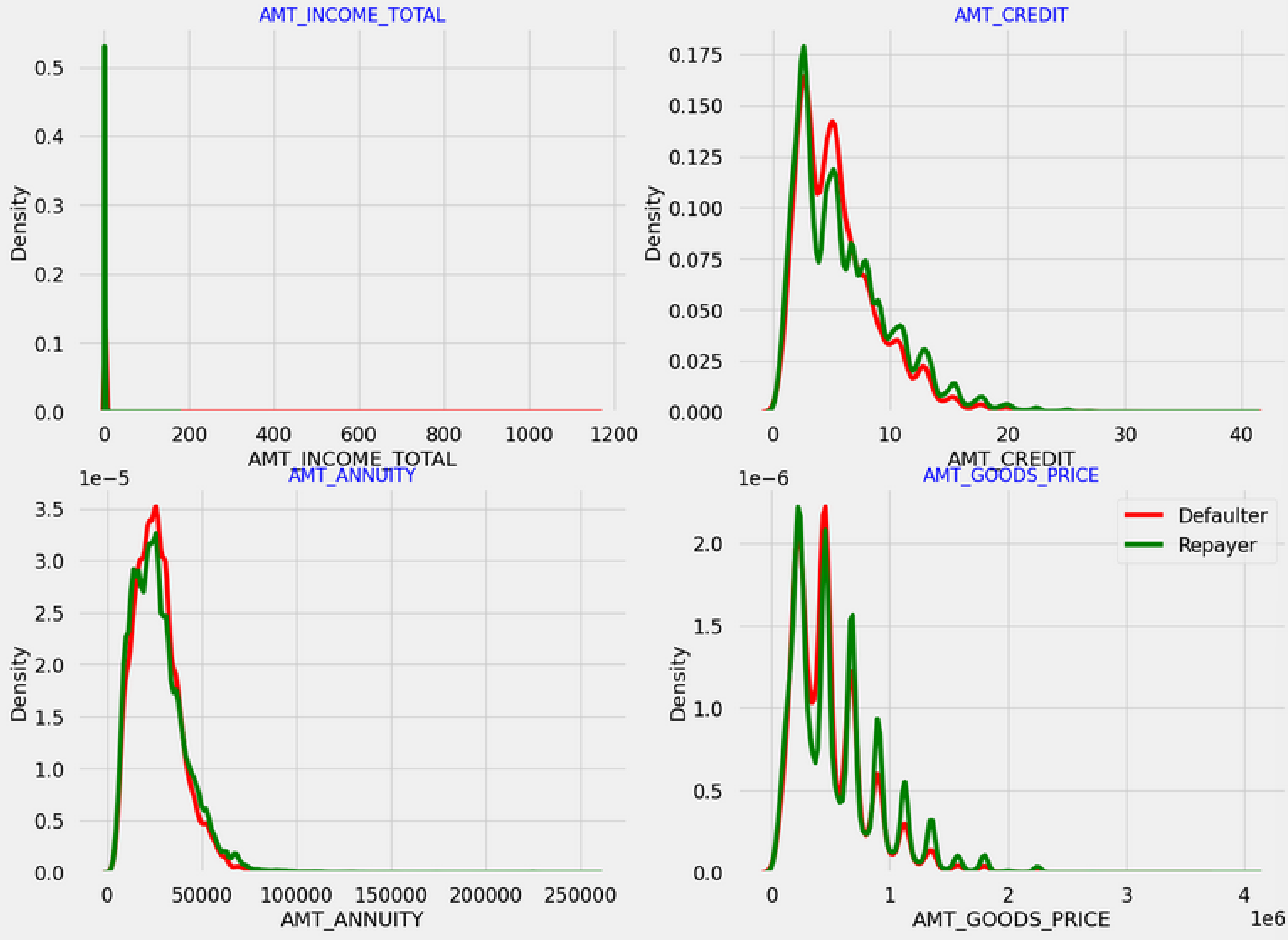
We can also see that repayers have high correlation in number of days employed(0.62) when compared to defaulters(0.58).

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.

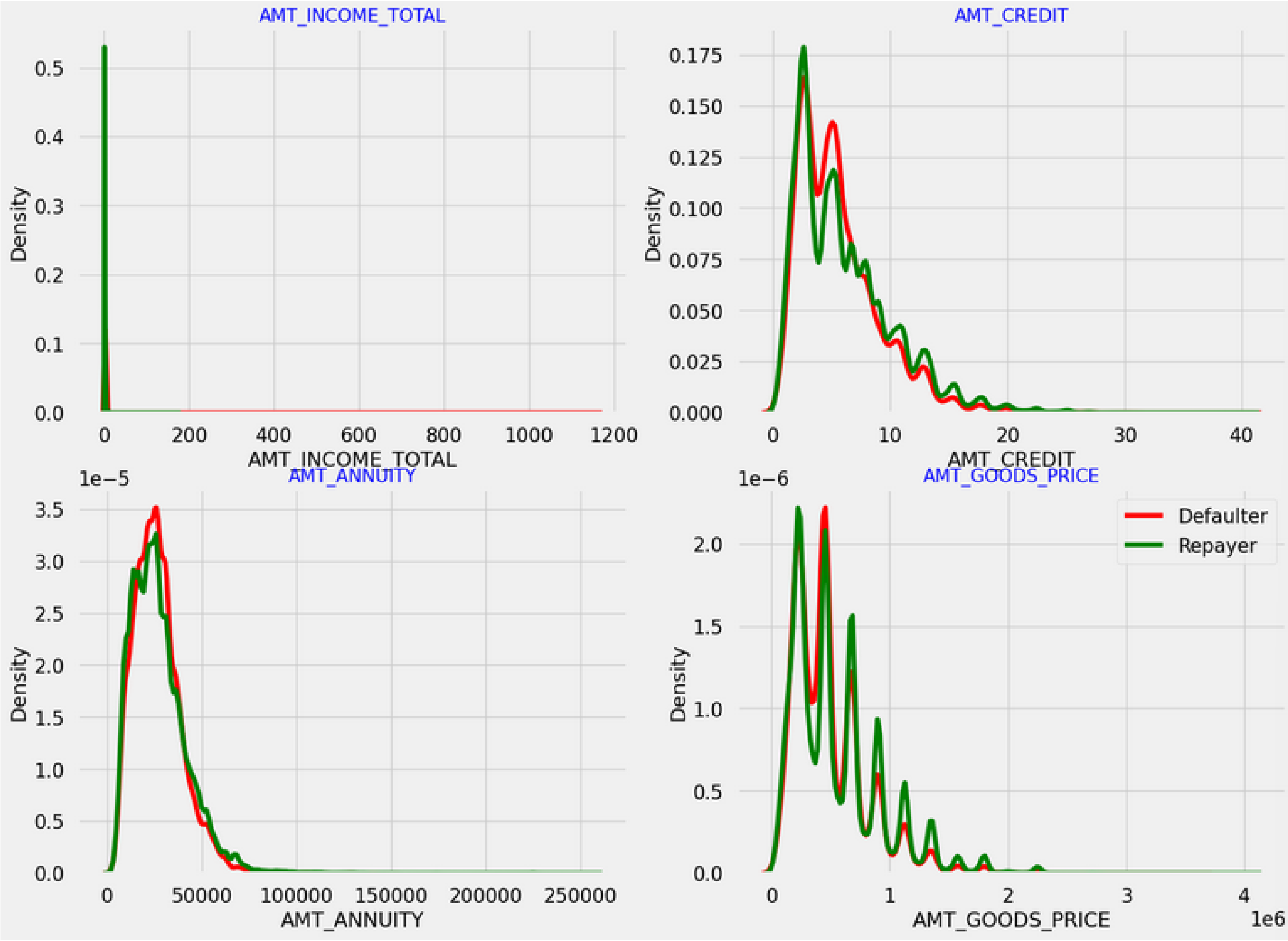
There is a slight increase in defaulted to observed count in social circle among defaulters(0.264) when compared to repayers(0.254

VISUALIZATIONS

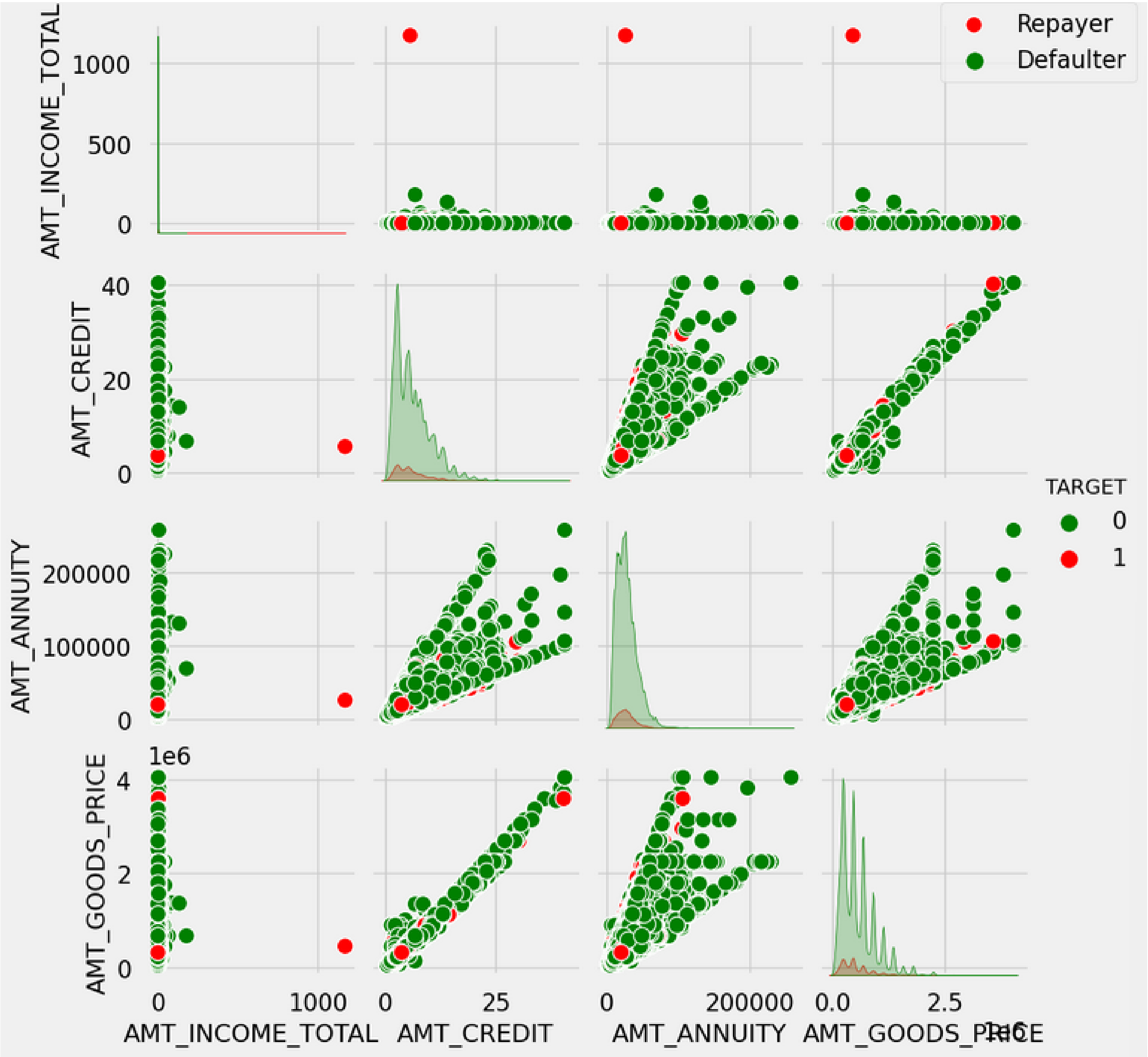


1. Most no of loans are given for goods price below 10 lakhs
2. Most people pay annuity below 50000 for the credit loan
3. Credit amount of the loan is mostly less then 10 lakhs

4. 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



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



1. When amt_annuity >15000 amt_goods_price> 3M, there is a lesser chance of defaulters

2. 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

3. There are very less defaulters for AMT_CREDIT >3M

4. Inferences related to distribution plot has been already mentioned in previous distplot graphs inferences section

TECH STACK

1. Jupyter Notebook
2. Google sheets

INSIGHTS

1. Most of the columns with high missing values are related to different area sizes on apartment owned/rented by the loan applicant.
2. in most of the loan application cases, clients who applied for loans has not submitted FLAG_DOCUMENT_X except FLAG_DOCUMENT_3. Thus, Except for FLAG_DOCUMENT_3, we can delete rest of the columns. Data shows if borrower has submitted FLAG_DOCUMENT_3 then there is a less chance of defaulting the loan.
3. There is almost no correlation between flags of mobile phone, email etc with loan repayment
4. More than 50% loan applicants have income amount in the range of 100K-200K. Almost 92% loan applicants have income less than 300K.
5. More Than 16% loan applicants have taken loan above 1M.
6. 31% loan applicants have 50+ years. More than 55% of loan applicants have 40+ years.
7. More than 55% of the loan applicants have work experience within 0-5 years and almost 80% of them have less than 10 years of work experience.
8. Almost 37% loan applicatants have applied for a new loan within 0-400 days of previous loan decision.
9. AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.
10. AMT_INCOME_TOTAL has huge number of outliers which indicate that few of the loan applicants have high income compared to the others.
11. DAYS_BIRTH has no outliers which means the data available is reliable.
12. DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and hence this is an incorrect entry.
13. AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.
14. CNT_PAYMENT has few outlier values.
15. SK_ID_CURR is an ID column and hence no outliers.
16. DAYS_DECISION has few number of outliers indicating that these previous applications decisions were taken long back.