Risk Analytics in Banking and Financial Services: An Exploratory Data Analysis

Executive Summary

This report presents an exploratory data analysis (EDA) to identify patterns and factors that indicate the likelihood of loan default among customers of a consumer finance company. The goal is to minimize financial losses and improve decision-making regarding loan approvals by leveraging data-driven insights.

Introduction

This case study aims to provide an understanding of risk analytics in banking and financial services by applying EDA in a real business scenario. The objective is to analyze how data can be used to minimize the risk of financial loss when lending to customers with varying credit histories.

Business Understanding

Loan providing companies face significant challenges in lending due to insufficient or non-existent credit histories of applicants. This leads to two primary risks:

1. **Loss of business** when creditworthy applicants are rejected.
2. **Financial loss** when loans are approved for applicants likely to default.

Scenarios in Loan Applications

The dataset includes various scenarios of loan applications:

- **Approved**: Loan applications approved by the company.
- **Cancelled**: Applications cancelled by the client during the approval process.
- **Refused**: Applications rejected by the company due to unmet requirements.
- **Unused Offer**: Loans cancelled by the client at different stages of the process.

Business Objectives

The aim is to identify patterns indicating payment difficulties, enabling actions such as denying loans, adjusting loan amounts, or lending at higher interest rates to risky applicants. The key objective is to understand the driving factors behind loan defaults for better portfolio and risk assessment.

Data Description

The dataset includes information about loan applications at the time of applying for a loan. It consists of variables related to the applicant's demographic, financial status, and historical loan performance.

Methodology

Data Cleaning and Preparation

- Handling missing values.
- Outlier detection and treatment.
- Data normalization and transformation.

Exploratory Data Analysis

- Descriptive statistics for all relevant variables.
- Visualization of data distributions and relationships.
- Identification of key patterns and anomalies.

Statistical Analysis

- Correlation analysis to identify relationships between variables.
- Segmentation analysis to categorize customers based on repayment behavior.

Detailed Analysis

Descriptive Statistics

- Summary statistics of applicant demographics and loan characteristics.
- Distribution of loan amounts, interest rates, and repayment durations.

Customer Segmentation

- Categorization based on credit history, income levels, and loan amounts.
- Analysis of default rates across different customer segments.

Transaction Analysis

- Trends in loan application volumes and approval rates.
- Seasonal patterns in loan applications and repayments.

Product Performance

- Analysis of different loan products and their default rates.
- Performance comparison of products across customer segments.

Churn Analysis

- Factors contributing to loan defaults.
- Predictive modeling to identify high-risk applicants.

Recommendations

- Strategies for improving loan approval processes.
- Risk mitigation techniques for high-risk applicants.
- Adjustments to loan products based on performance analysis.

Findings and Discussion

- Key insights from the EDA, highlighting variables strongly associated with loan defaults.
- Interpretation of results and their implications for the company's lending strategy.

Conclusion

After analyzing the datasets, several attributes have been identified that can help predict whether a client will repay the loan or not. The analysis is summarized below:

Factors Indicating a Likely Repayer:

- **NAME_EDUCATION_TYPE**: Applicants with academic degrees have fewer defaults.
- **NAME_INCOME_TYPE**: Students and businessmen show no defaults.
- **REGION_RATING_CLIENT**: Rating 1 regions are safer.
- **ORGANIZATION_TYPE**: Clients in Trade Type 4 and 5, and Industry Type 8 have defaulted less than 3%.
- **DAYS_BIRTH**: Applicants above the age of 50 have a low probability of defaulting.
- **DAYS_EMPLOYED**: Clients with over 40 years of employment have less than a 1% default rate.
- **AMT_INCOME_TOTAL**: Applicants with incomes over 700,000 are less likely to default.
- **NAME_CASH_LOAN_PURPOSE**: Loans for hobbies and buying garages are mostly repaid.
- **CNT_CHILDREN**: Applicants with zero to two children tend to repay their loans.

Factors Indicating a Likely Defaulter:

- **CODE_GENDER**: Men have a relatively higher default rate.
- **NAME_FAMILY_STATUS**: Those who are single or in civil marriages default more.
- **NAME_EDUCATION_TYPE**: Lower secondary and secondary education levels are associated with higher defaults.
- **NAME_INCOME_TYPE**: Those on maternity leave or unemployed default more frequently.
- **REGION_RATING_CLIENT**: Regions with a rating of 3 have the highest defaults.
- **OCCUPATION_TYPE**: Low-skill laborers, drivers, waiters/barmen, security staff, laborers, and cooking staff have high default rates.
- **ORGANIZATION_TYPE**: High default rates are seen in Transport Type 3, Industry Type 13, Industry Type 8, and restaurants. Self-employed individuals also have high default rates.
- **DAYS_BIRTH**: Young people aged 20-40 have higher default probabilities.
- **DAYS_EMPLOYED**: Those with less than 5 years of employment have high default rates.
- **CNT_CHILDREN & CNT_FAM_MEMBERS**: Clients with 9 or more children default 100% of the time.
- **AMT_GOODS_PRICE**: Credits above 3 million show an increase in defaults.

Strategic Recommendations:

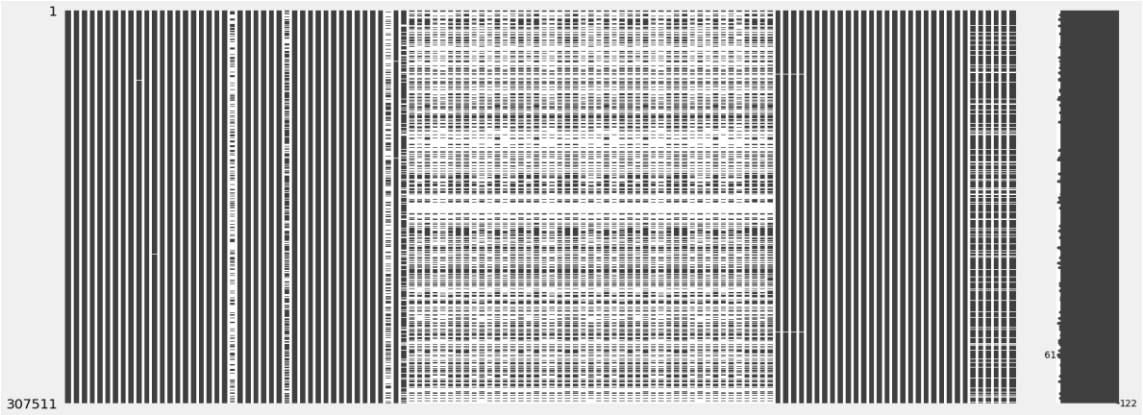
- **NAME_HOUSING_TYPE**: Offer loans with higher interest rates to those in rented apartments or living with parents.
- **AMT_CREDIT**: Apply higher interest rates to loans in the 300-600k range.
- **AMT_INCOME**: Higher interest rates for applicants with incomes less than 300,000.
- **CNT_CHILDREN & CNT_FAM_MEMBERS**: Apply higher interest rates for clients with 4-8 children.
- **NAME_CASH_LOAN_PURPOSE**: Continue the approach of high-interest rates or rejections for loans aimed at repairs.

Other Suggestions:

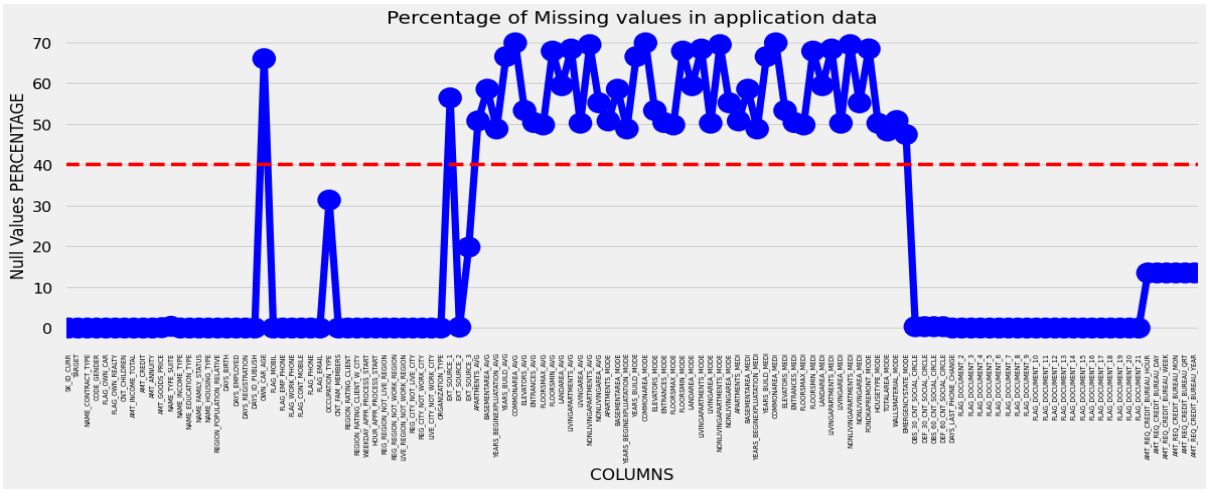
- Record reasons for cancellations to negotiate terms with potential repaying customers.
- Document reasons for rejections to identify and approach potential repaying clients in the future.

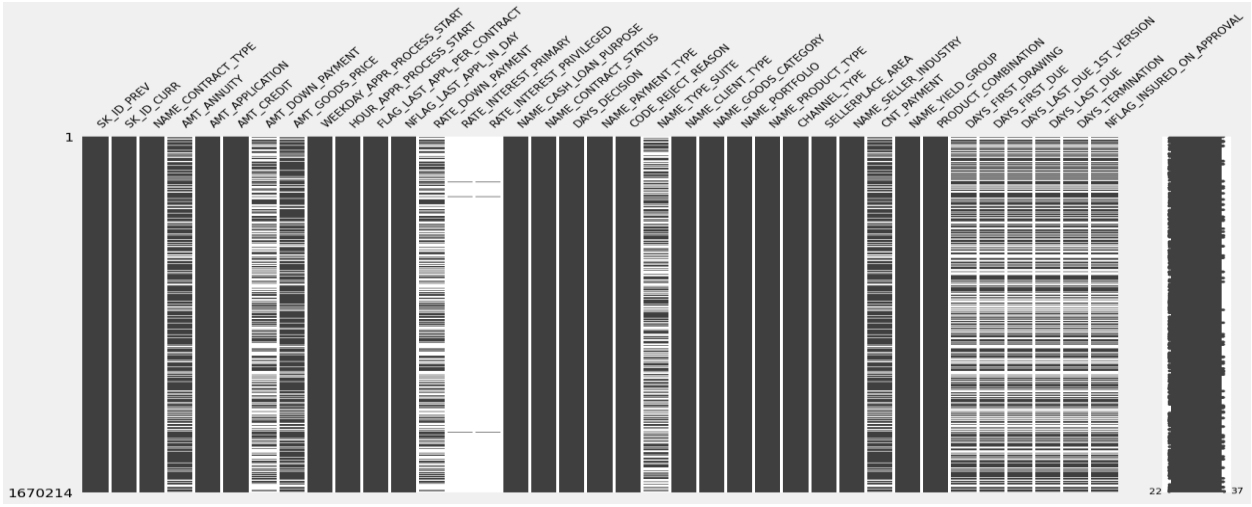
This comprehensive analysis provides actionable insights to minimize financial risks and optimize loan approval processes, enhancing the company's financial stability and customer satisfaction.

Appendices

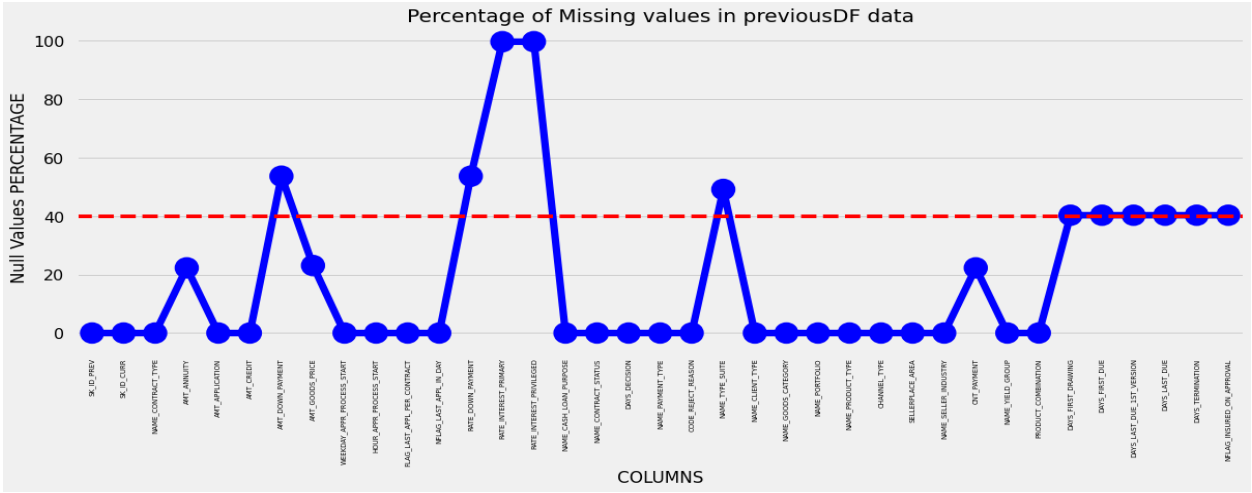


Missing value visualization using matrix in application data.

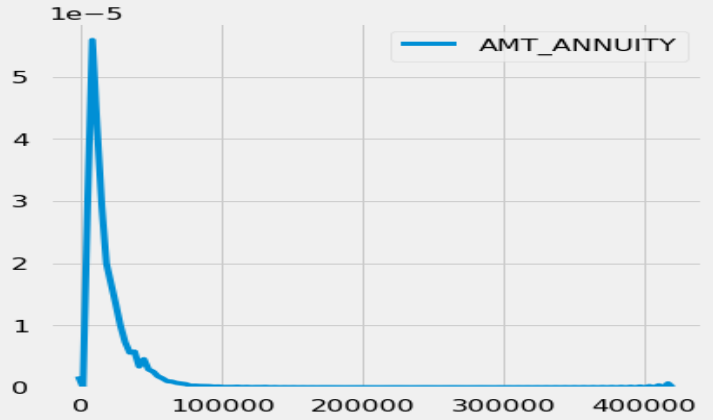




Missing value visualization using matrix in previous data.

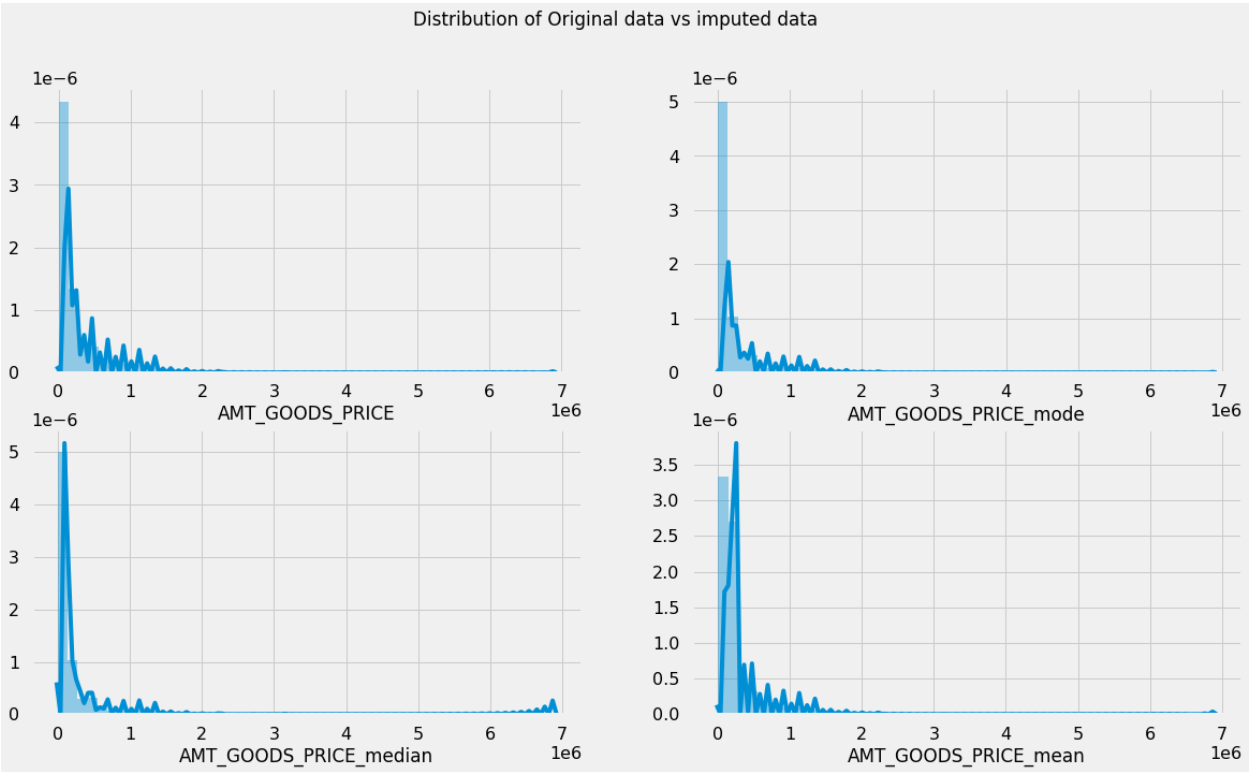


Using correlation property to drop the irrelevant features.

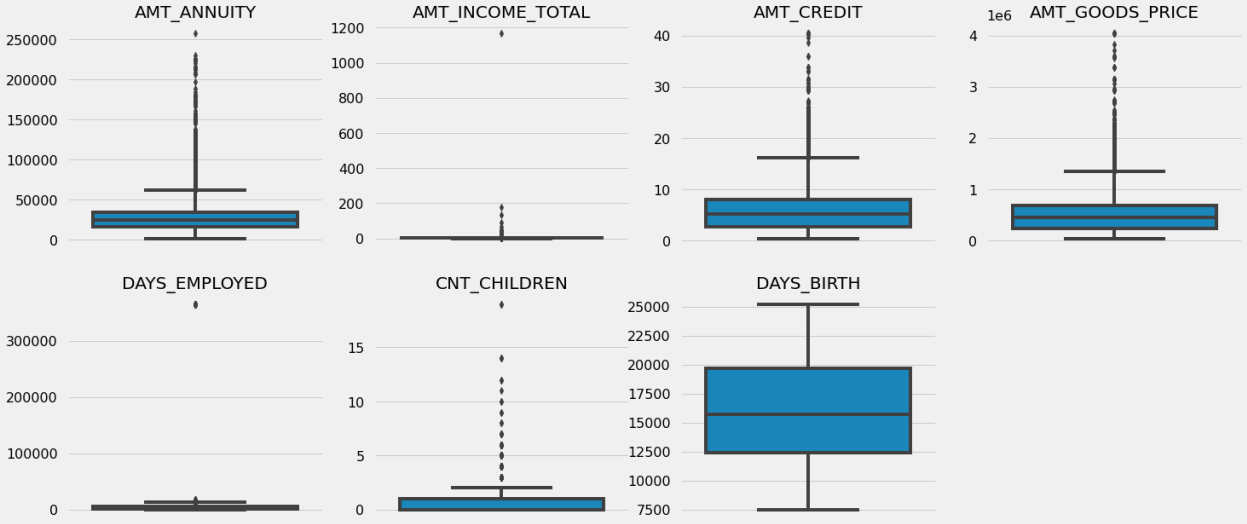


The single peak at left side of graph indicates the presence of outliers and hence imputing mean would not be a right approach and hence imputing with median

ve



There are several peaks along the distribution. So, impute using mean, median, mode to check which one resembles the original distribution



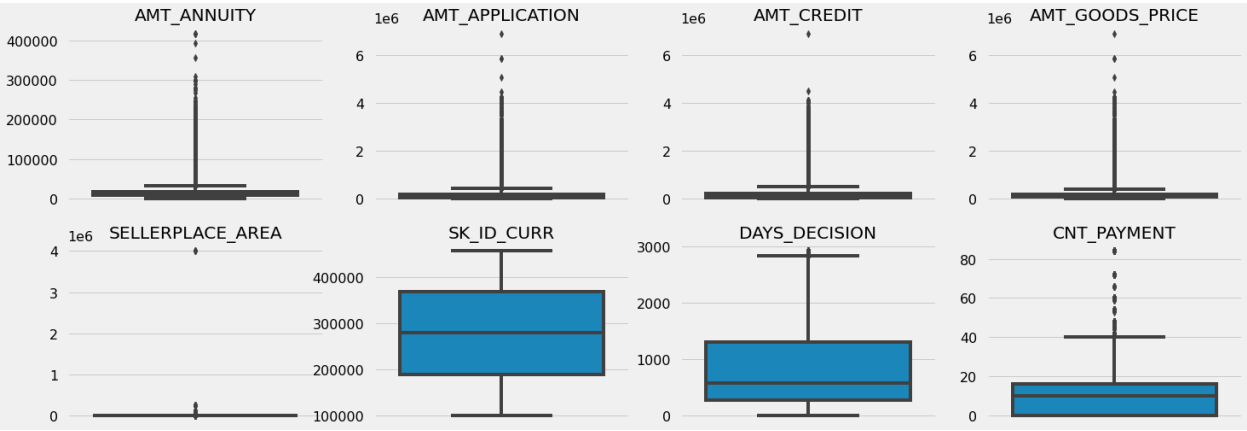
It can be seen that in current application data

AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.

AMT_INCOME_TOTAL has huge number of outliers which indicates 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 must be incorrect entry.

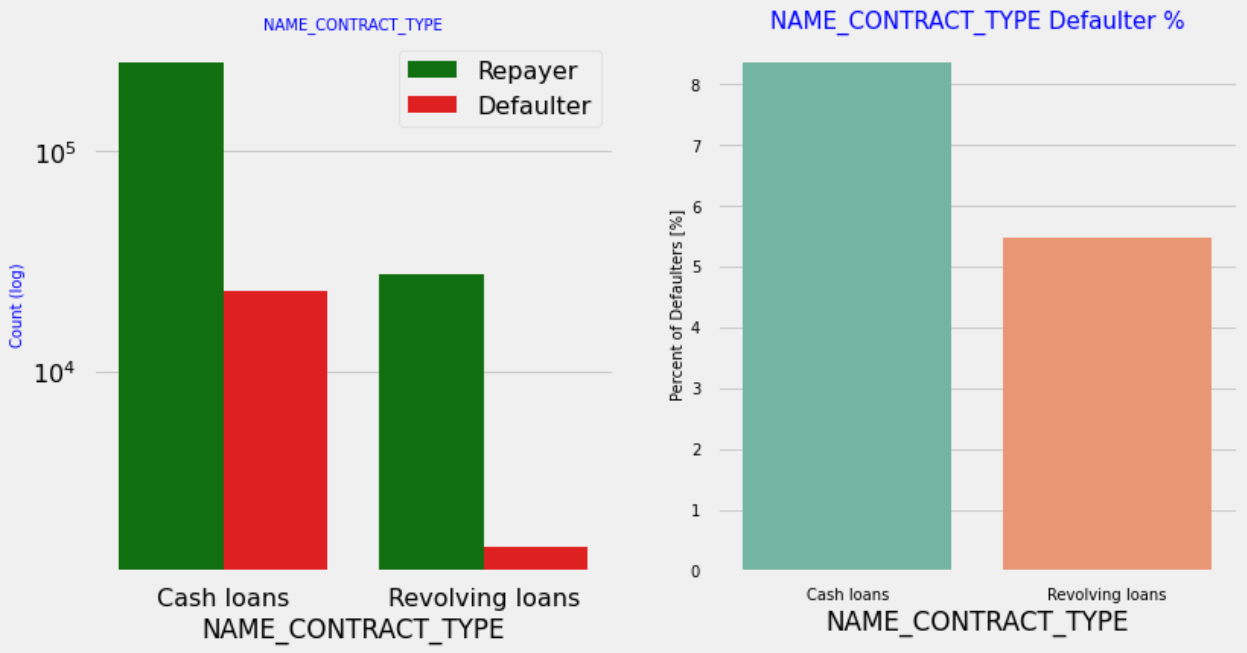


It can be seen that in previous application data

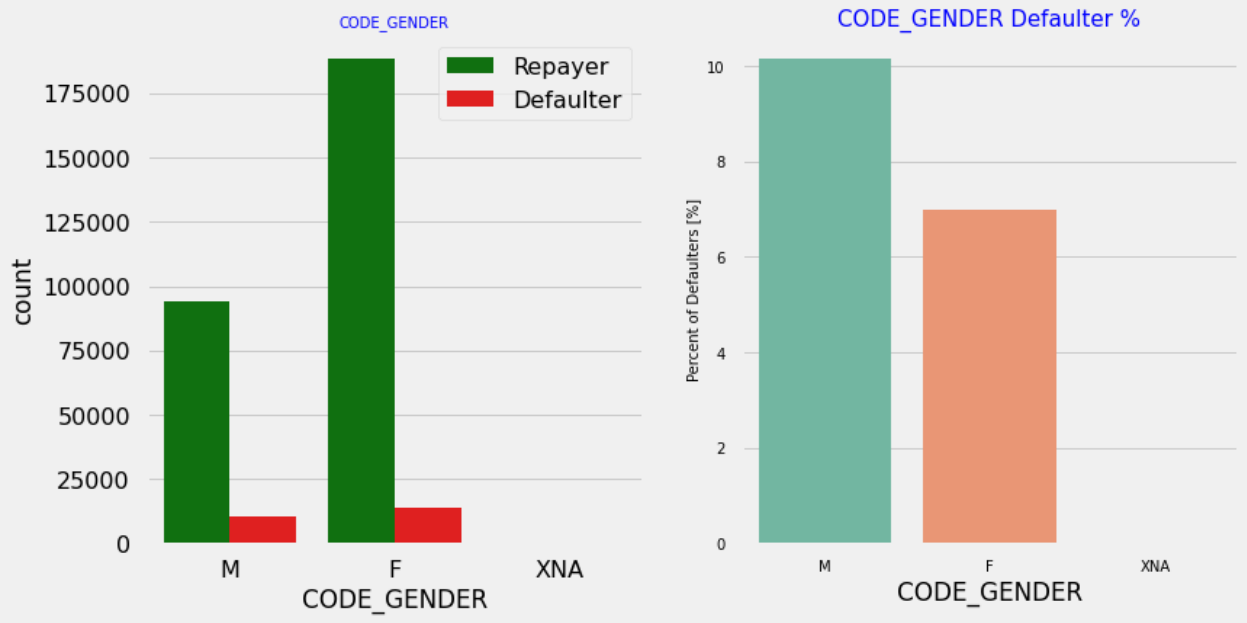
AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.

CNT_PAYMENT has few outlier values.SK_ID_CURR is an ID column and hence no outliers.

DAYS_DECISION has a small number of outliers indicating that these previous applications decisions were taken long back.

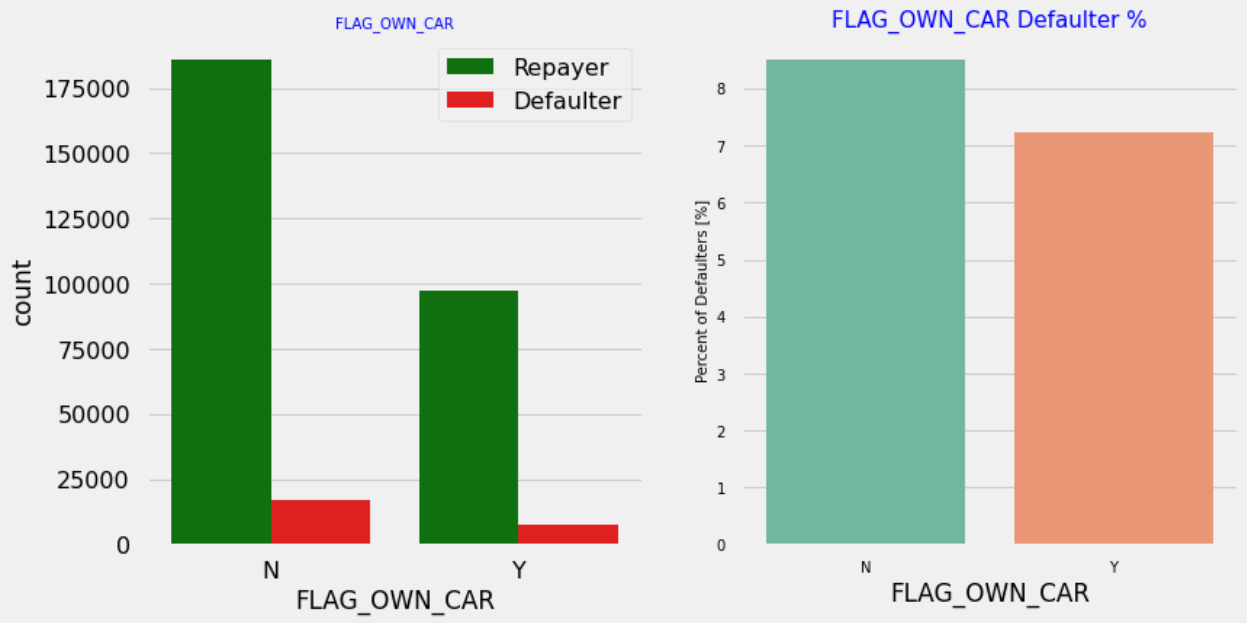


Contract type: Revolving loans are just a small fraction (10%) from the total number of loans; at the same time, a larger amount of Revolving loans, comparing with their frequency, are not repaid.

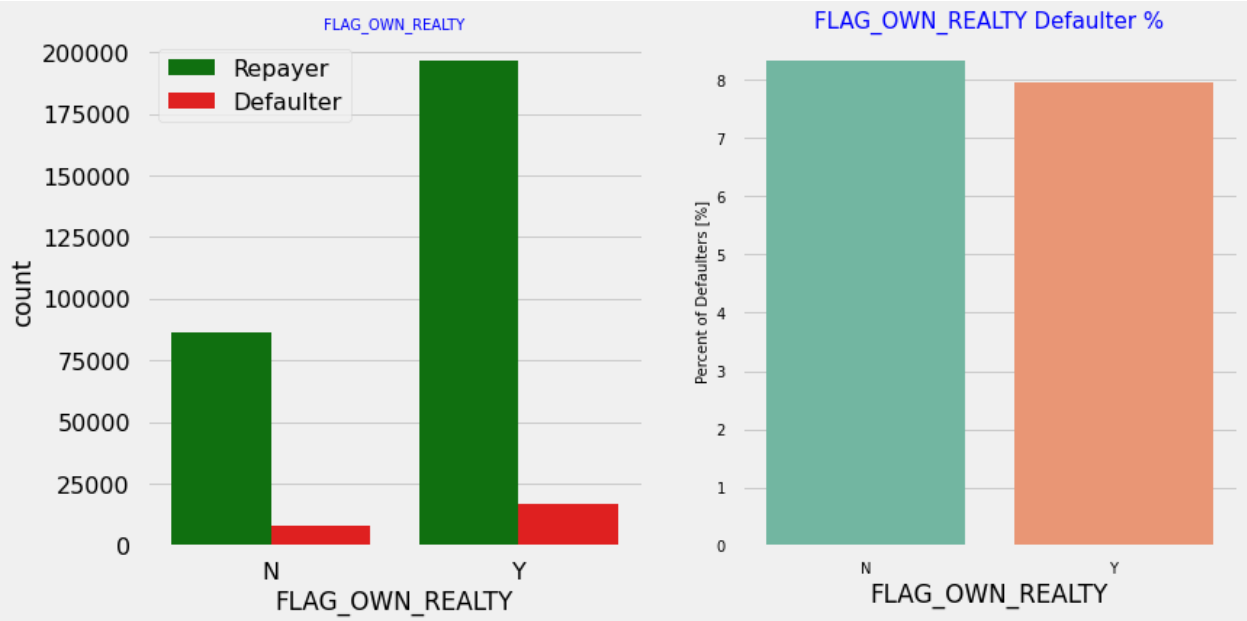


The number of female clients is almost double the number of male clients. Based on the

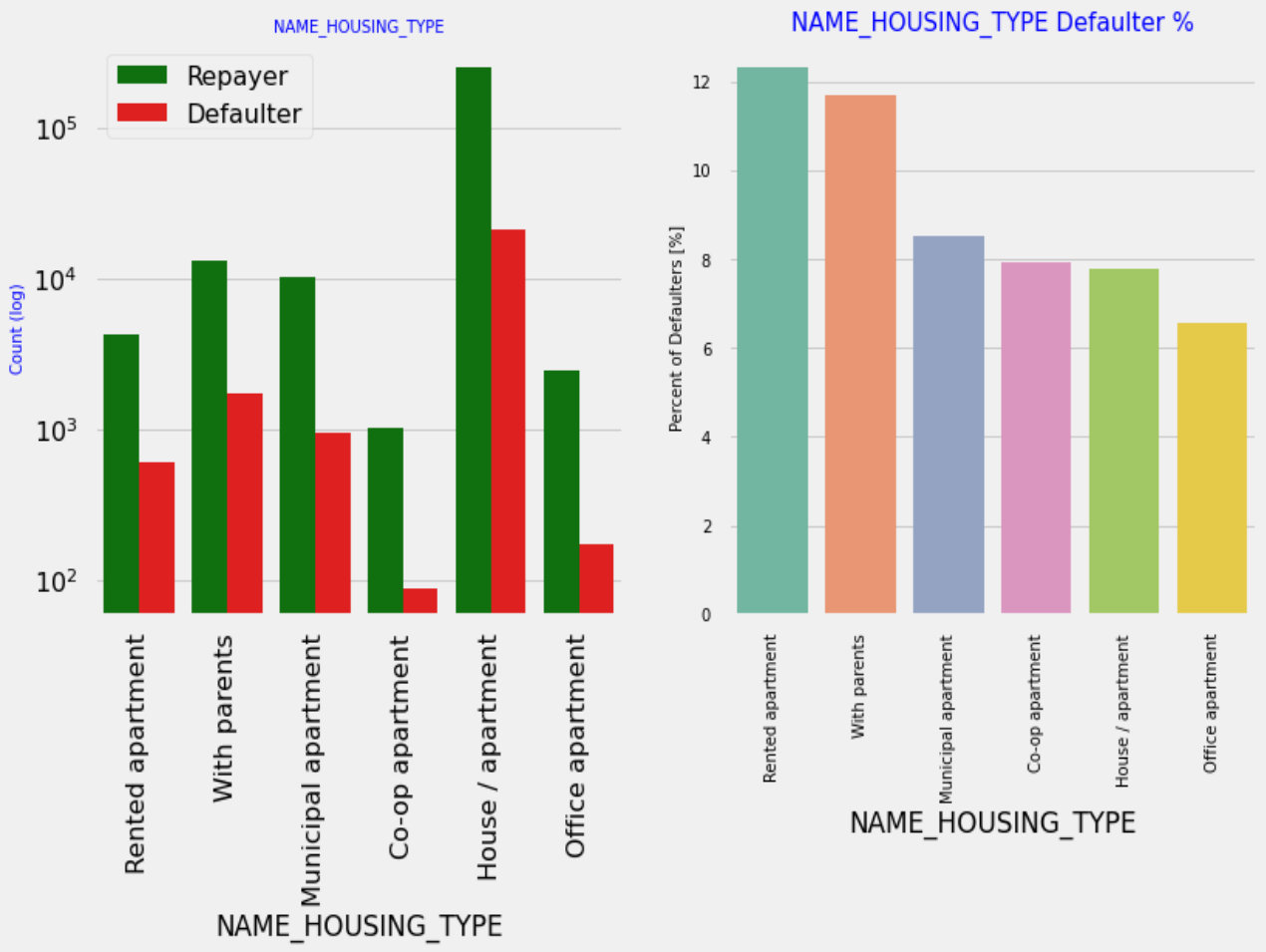
percentage of defaulted credits, males have a higher chance of not returning their loans (~10%), comparing with women (~7%)



Clients who own a car are half 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.



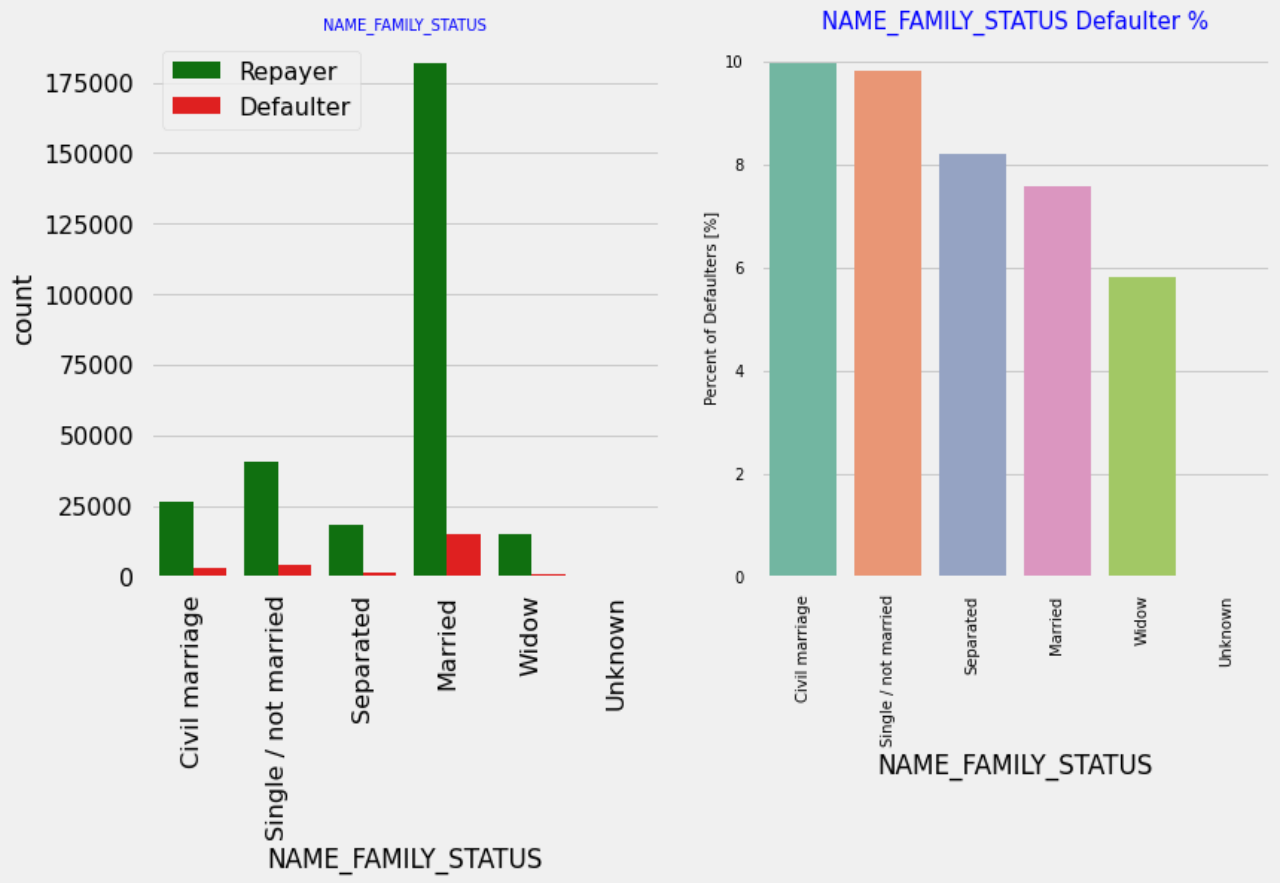
The clients who own real estate are more than double of the ones that don't own. But the defaulting rate of both categories are around the same (~8%). Thus, there is no correlation between owning a reality and defaulting on the loan.



Majority of people live in House/apartment

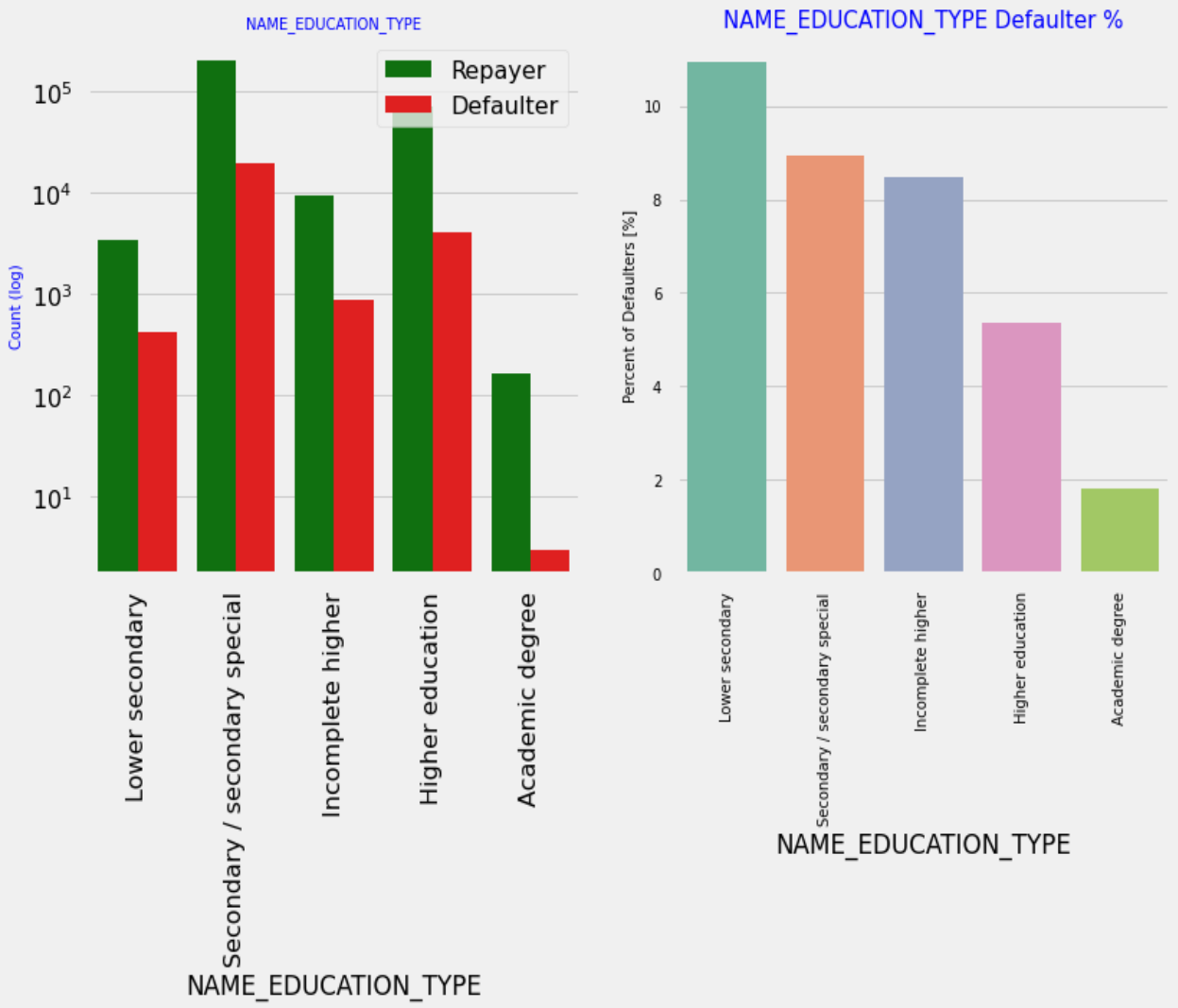
People living in office apartments have lowest default rate

People living with parents (~11.5%) and living in rented apartments (>12%) have higher probability of defaulting



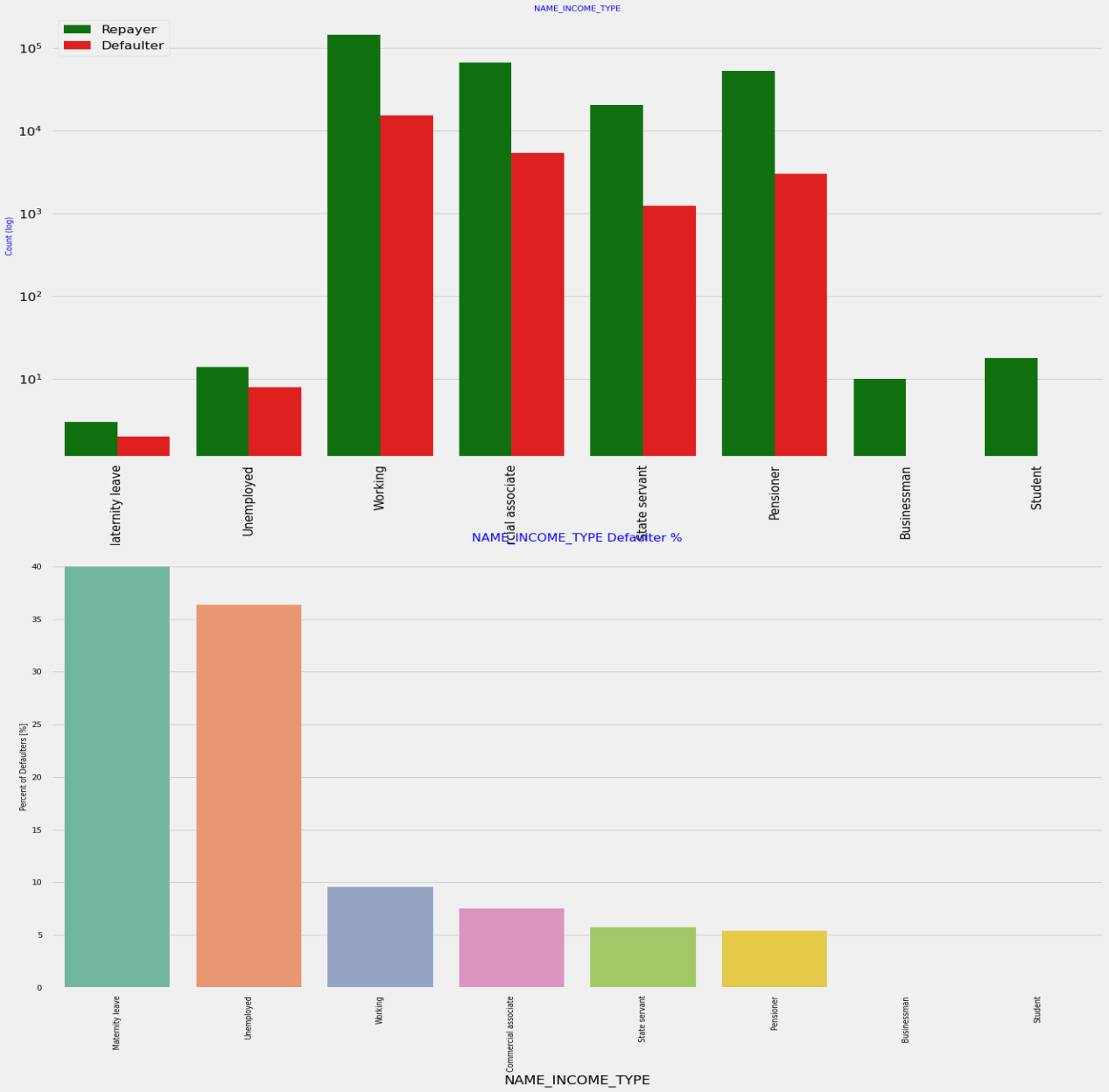
Most of the people who have taken loan are married, followed by Single/not married and civil marriage

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

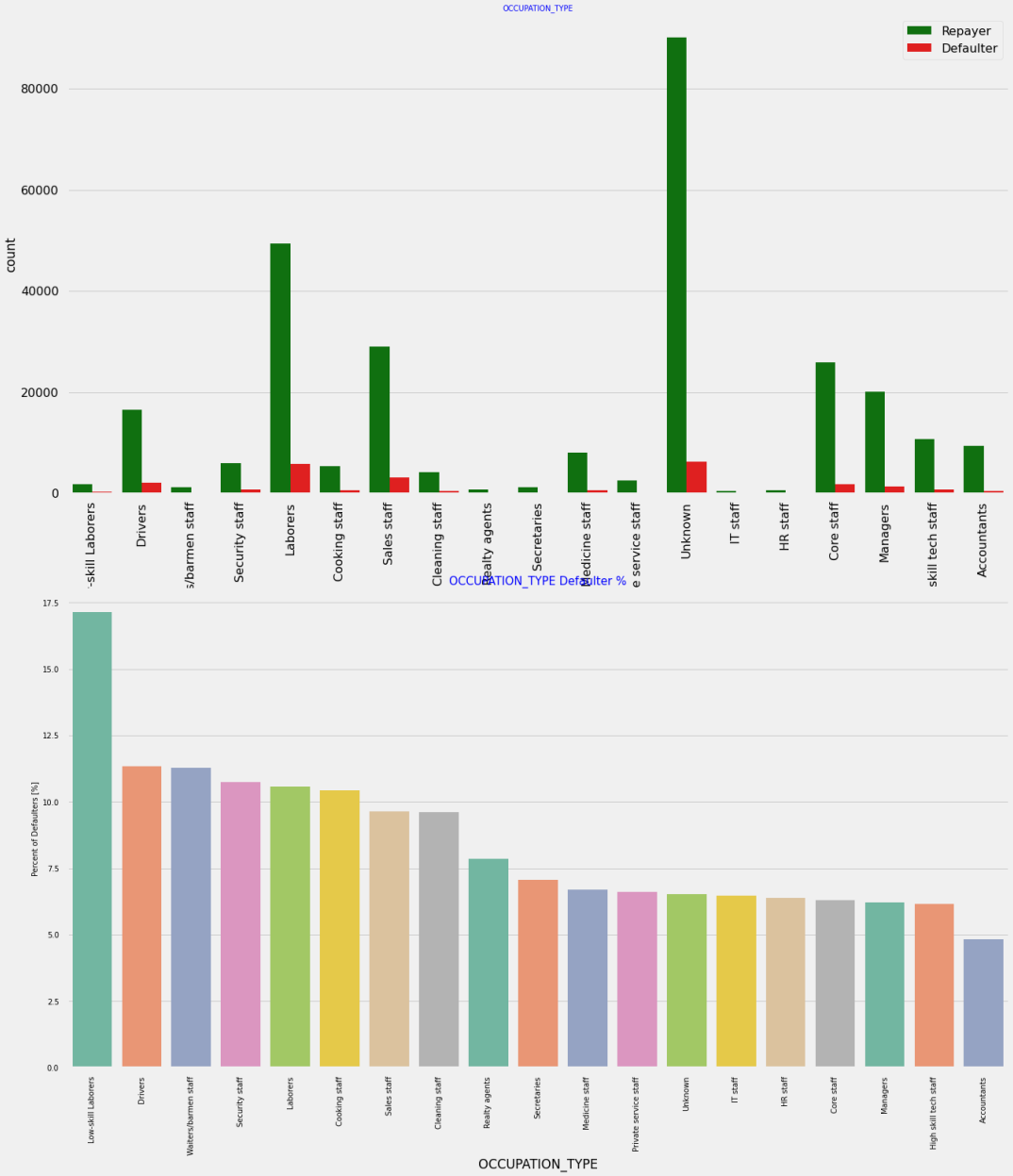
The Lower secondary category, although rare, has the largest rate of not returning the loan (11%). The people with an Academic degree have less than 2% defaulting rate.



Most of applicants for loans have income type as Working, followed by Commercial associate, Pensioner and State servant.

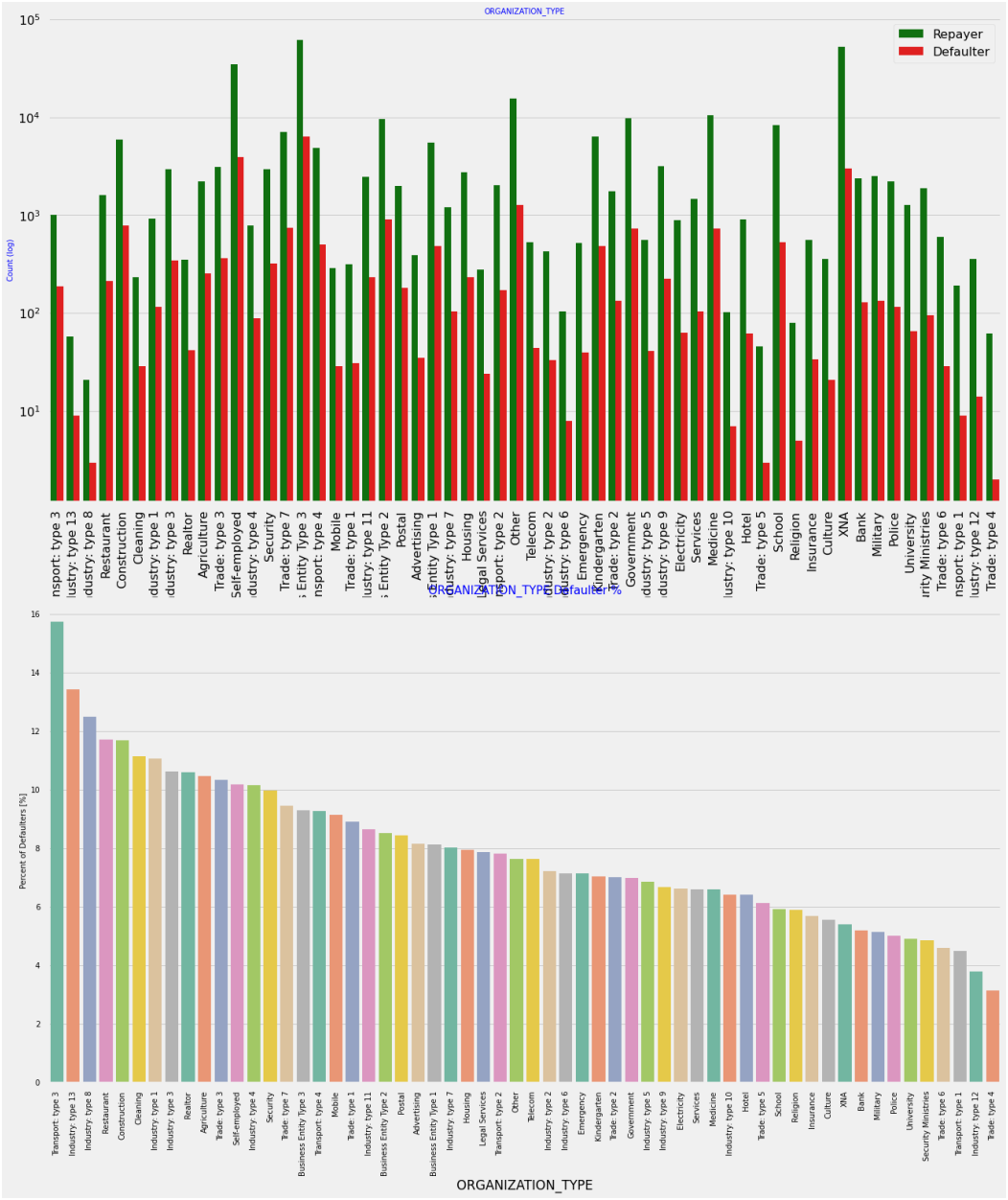
The applicants with the type of income Maternity leave have almost 40% ratio of not returning loans, followed by Unemployed (37%). The rest of types of incomes are under the average of 10% for not returning loans.

Student and Businessmen, though less in numbers do not have any default record. Thus, these two categories are safest for providing loans.



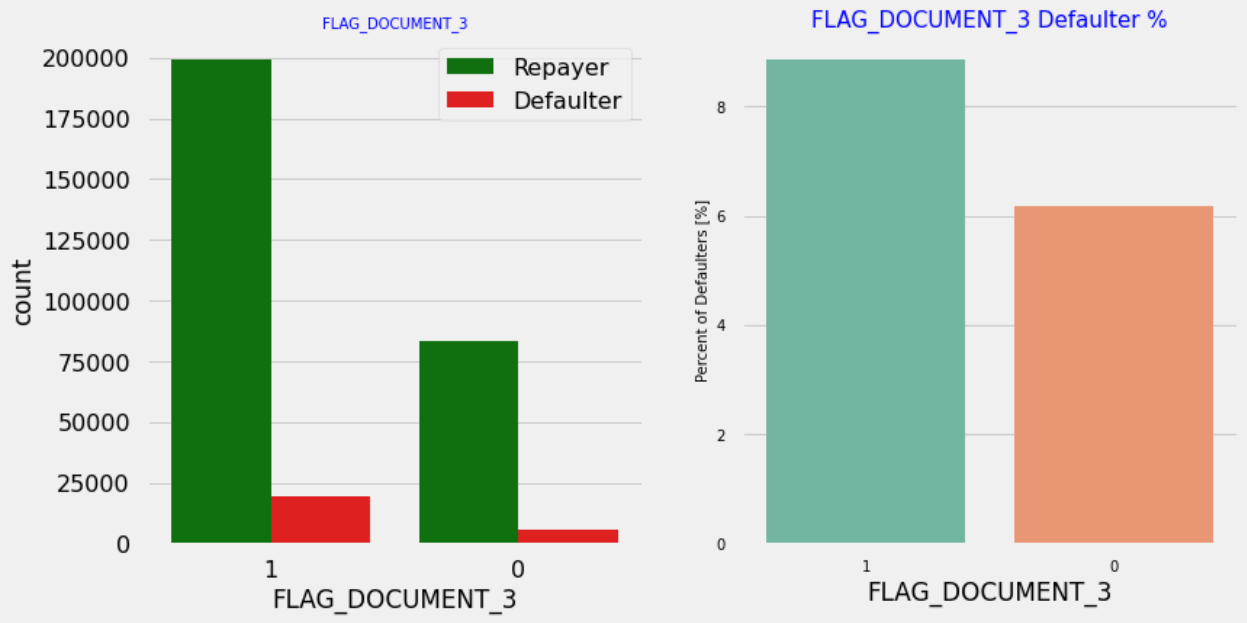
Most of the loans are taken by Laborers, followed by Sales staff. IT staff take the lowest amount of loans.

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.

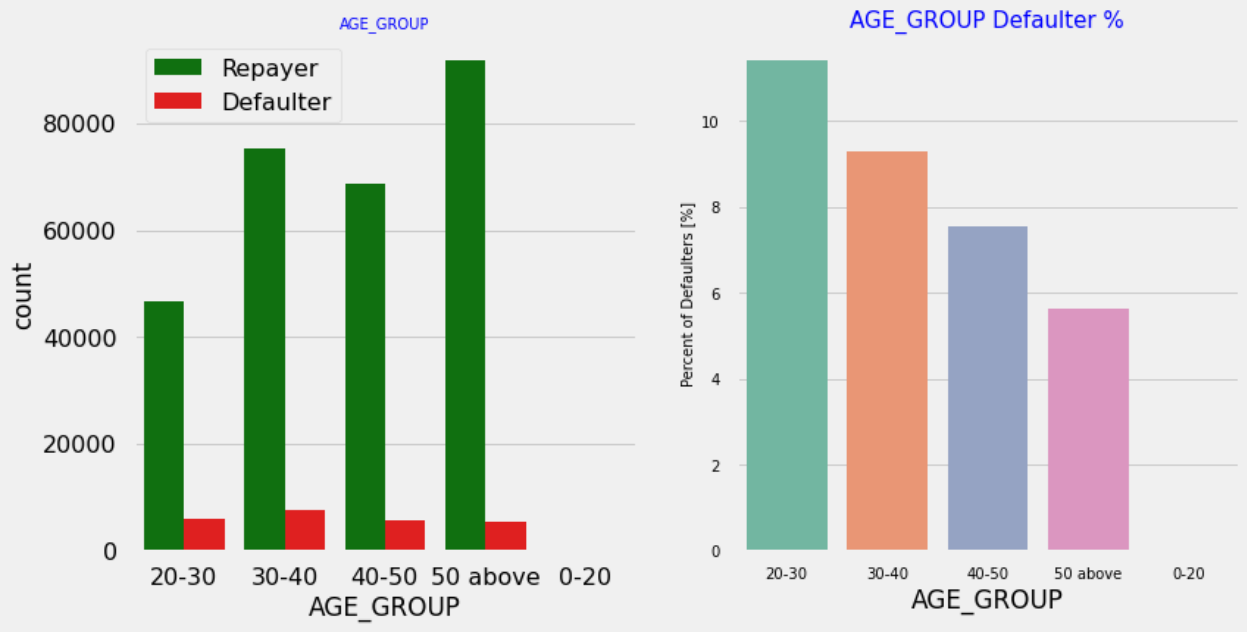


Organizations with the highest percentage of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self-employed people have relatively high defaulting rate, and thus should be avoided to be approved for loan or provide loan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

Following category of organization type has lesser defaulters thus safer for providing loans: Trade Type 4 and 5, Industry type 8

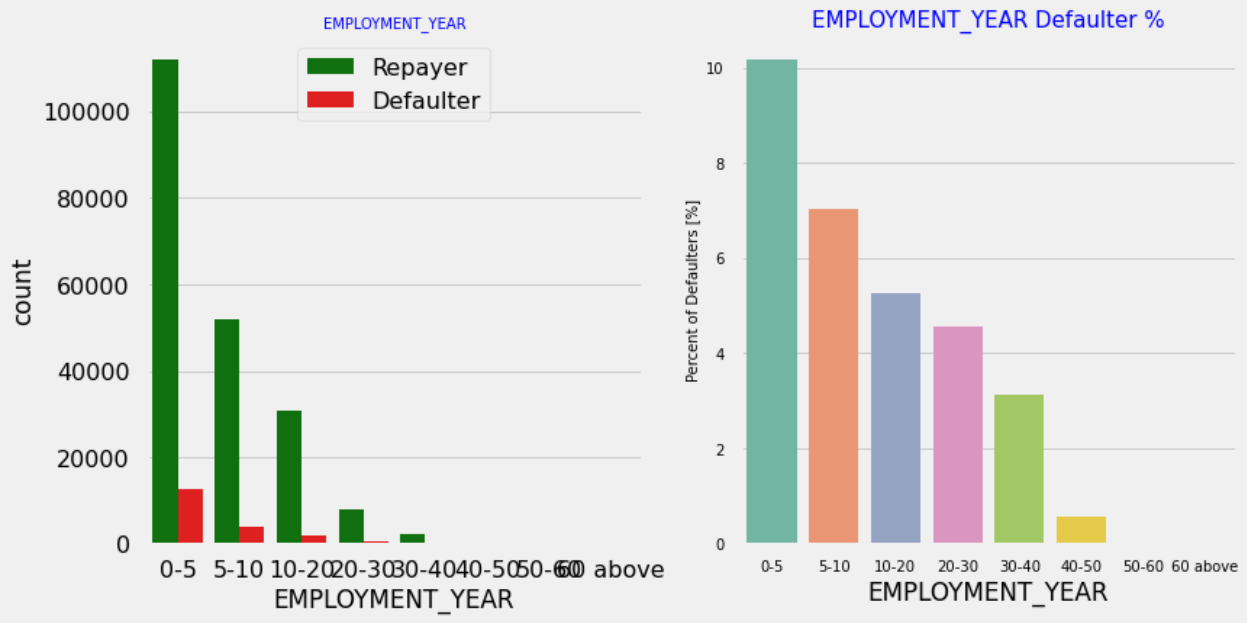


There is no significant correlation between repayers and defaulters in terms of submitting document 3 as we see even if applicants have submitted the document, they have defaulted a slightly more (~9%) than who have not submitted the document (6%)



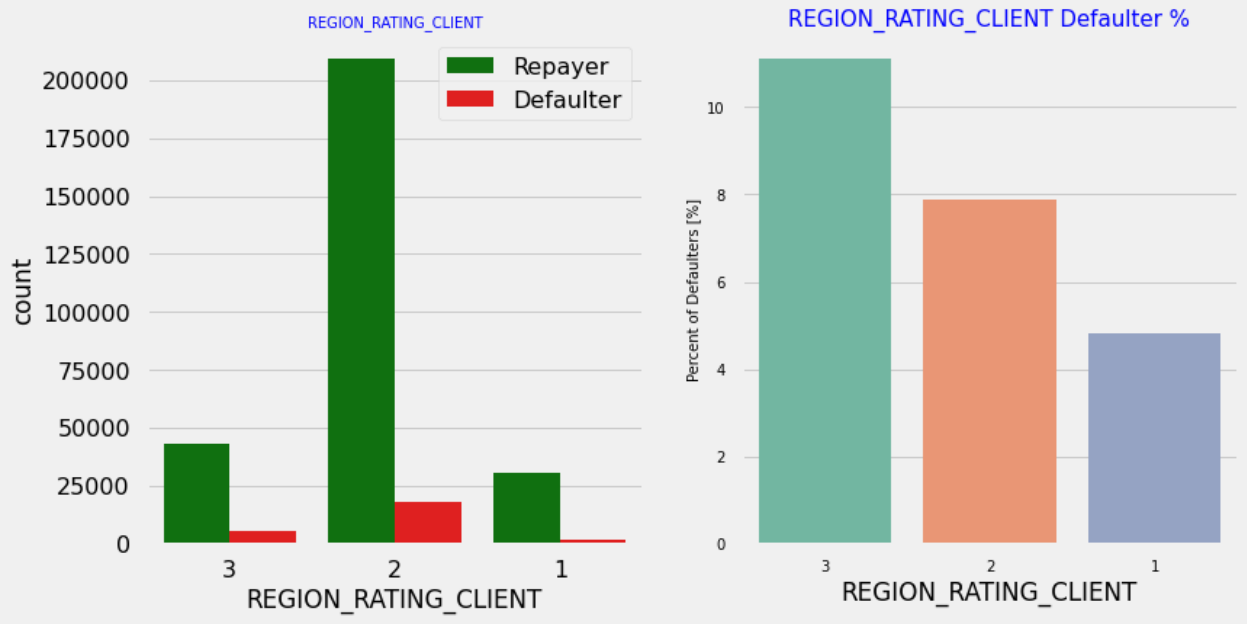
People in the age group range 20-40 have higher probability of defaulting

People above the age of 50 have a low probability of defaulting.



Majority of the applicants have been employed for between 0-5 years. The defaulting rating of this group is also the highest which is 10%

With the increase of employment year, defaulting rate is gradually decreasing with people having 40+ year experience having less than 1% default rate.



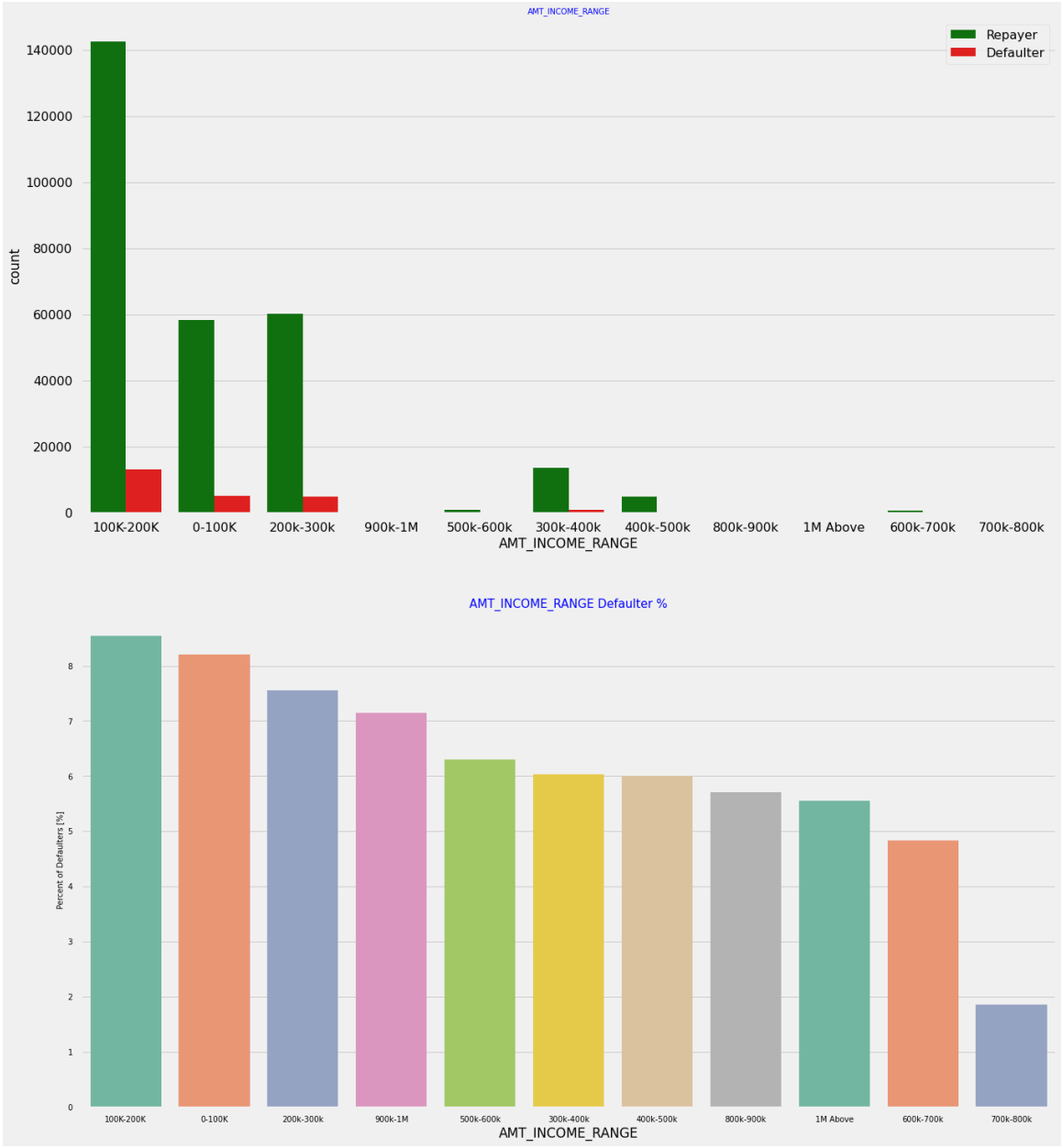
Most of the applicants are living in Region_Rating 2 place.

Region Rating 3 has the highest default rate (11%)

Applicants living in Region_Rating 1 have the lowest probability of defaulting, thus safer for approving loans.

More than 80% of the loan provided are for amount less than 900,000

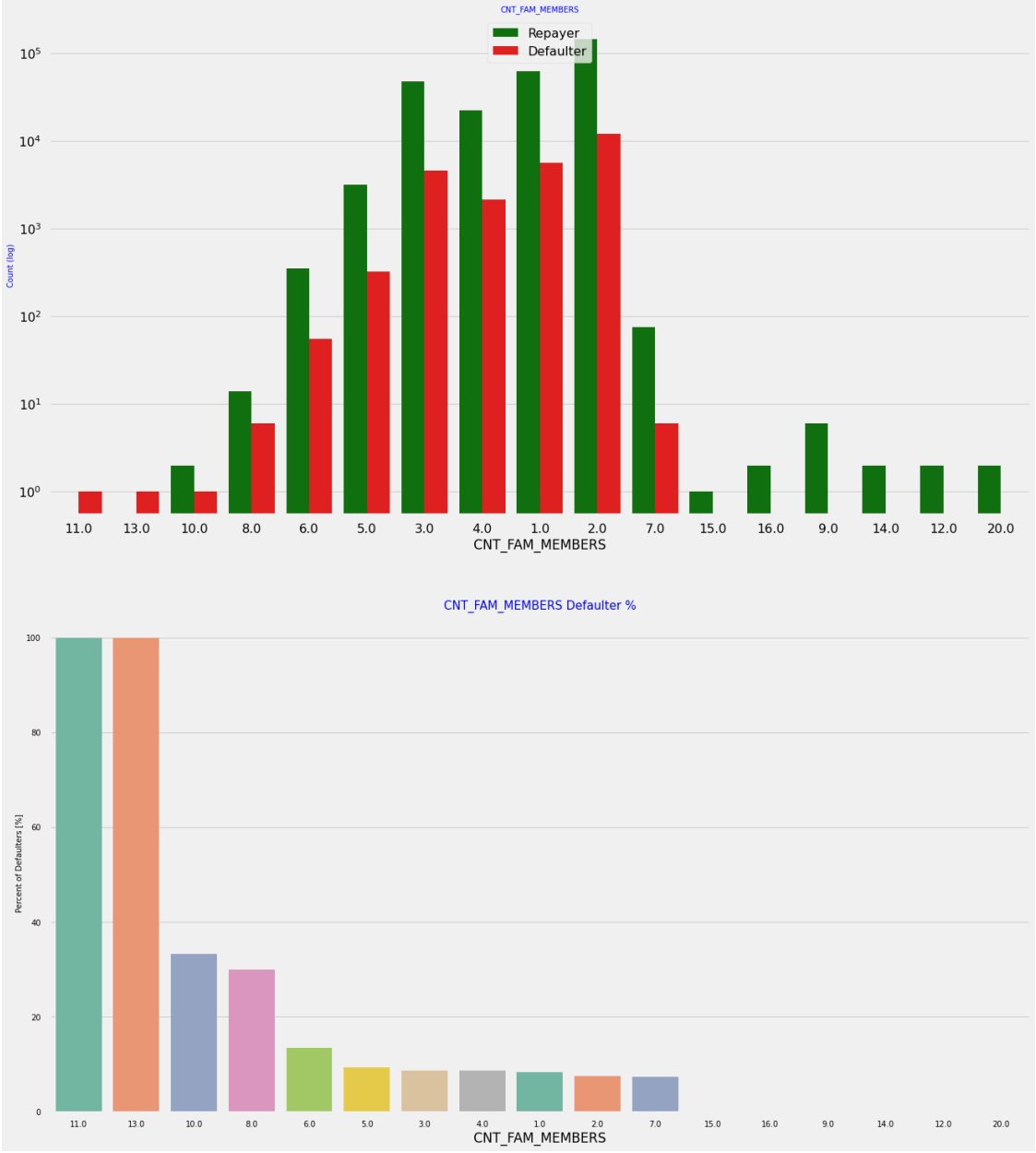
People who get a loan for 300-600k tend to default more than others.



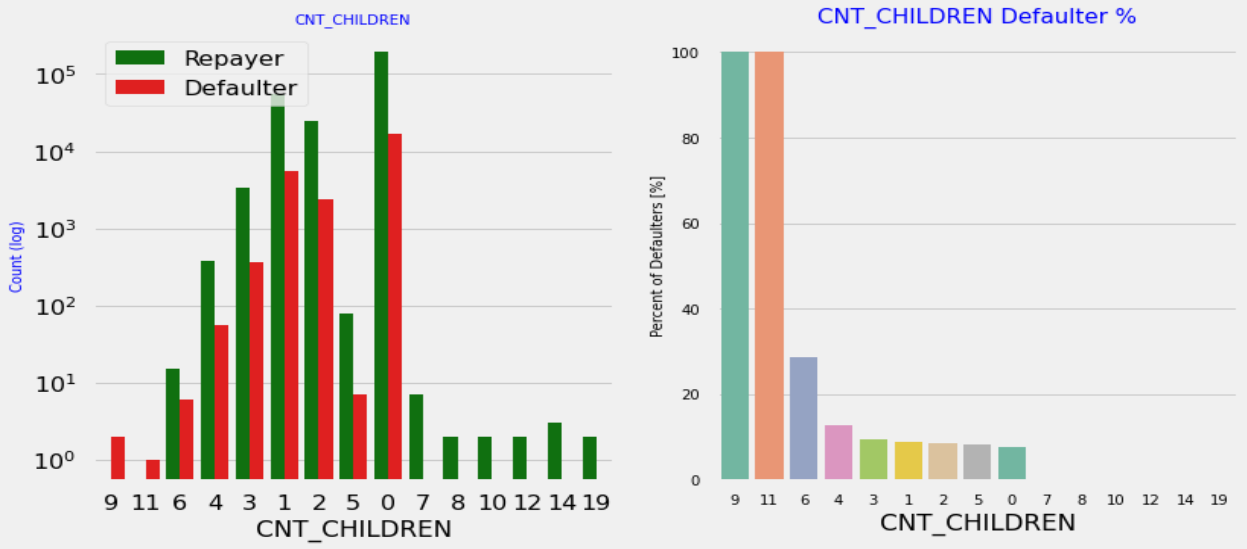
90% of the applications have Income total less than 300,000

Application with Income less than 300,000 has high probability of defaulting

Applicant with Income more than 700,000 are less likely to default.



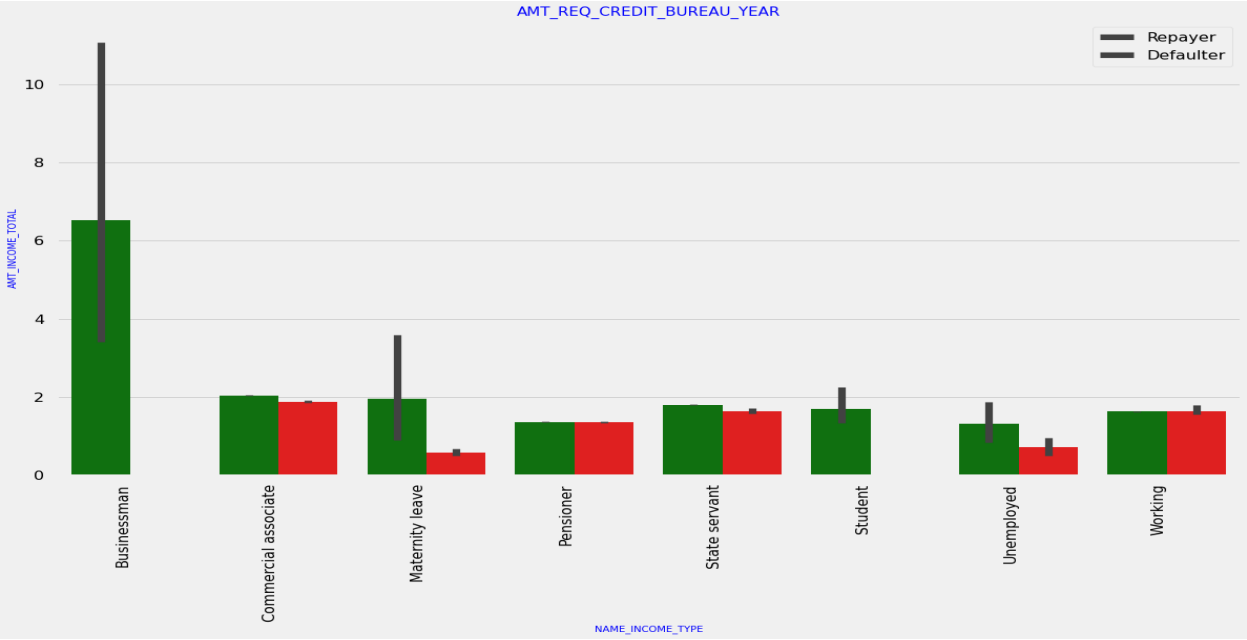
Family member follows the same trend as children were having more family members increases the risk of defaulting



Most of the applicants do not have children

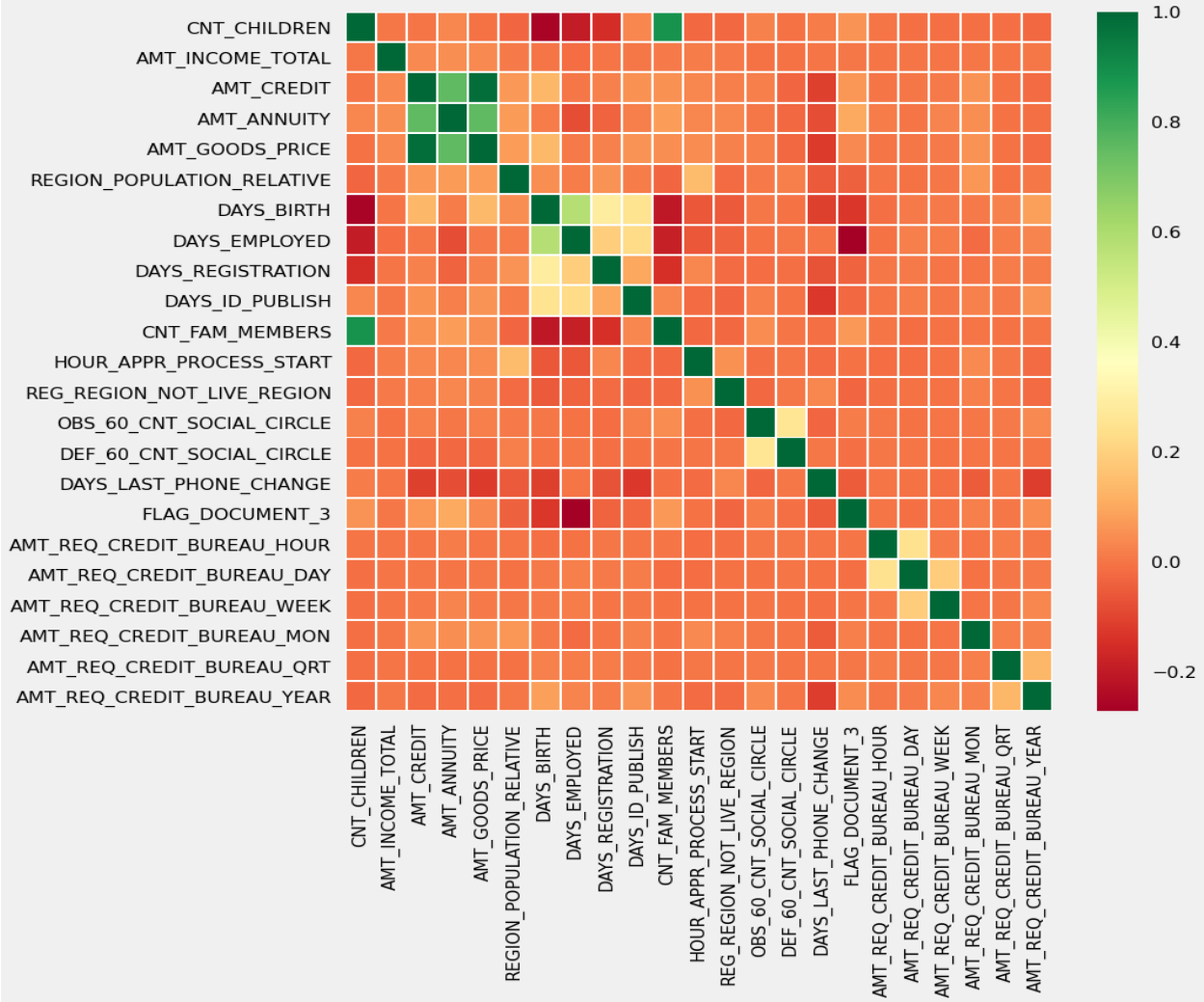
Very few clients have more than 3 children.

A client who has more than 4 children has a very high default rate with a child count of 9 and 11 showing 100% default rate.

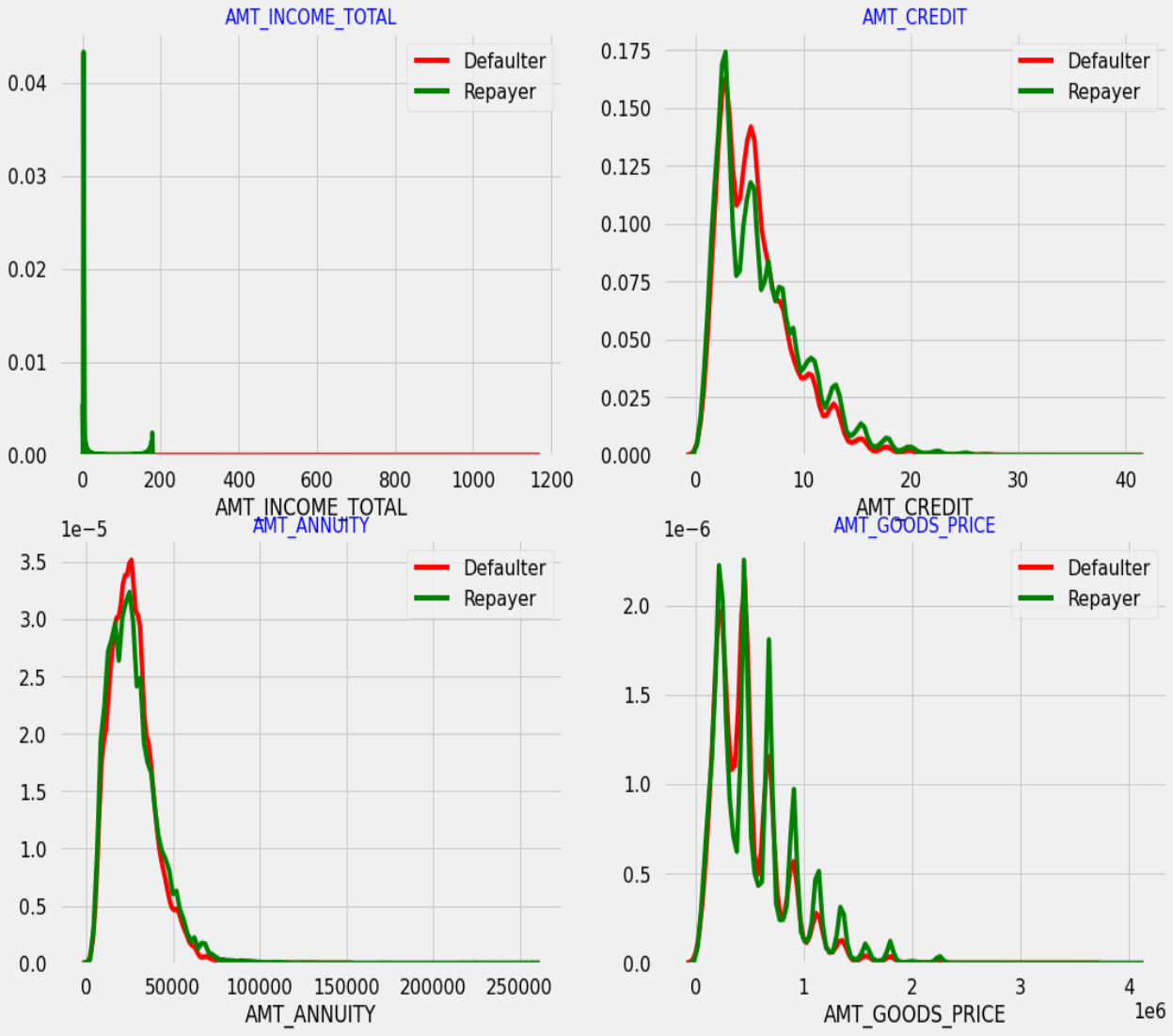


It can be seen that businessman's income is the highest and the estimated range with default 95% confidence level seem to indicate that the income of a businessman could be in the range of slightly close to 4 lakhs and slightly above 10 lakhs.

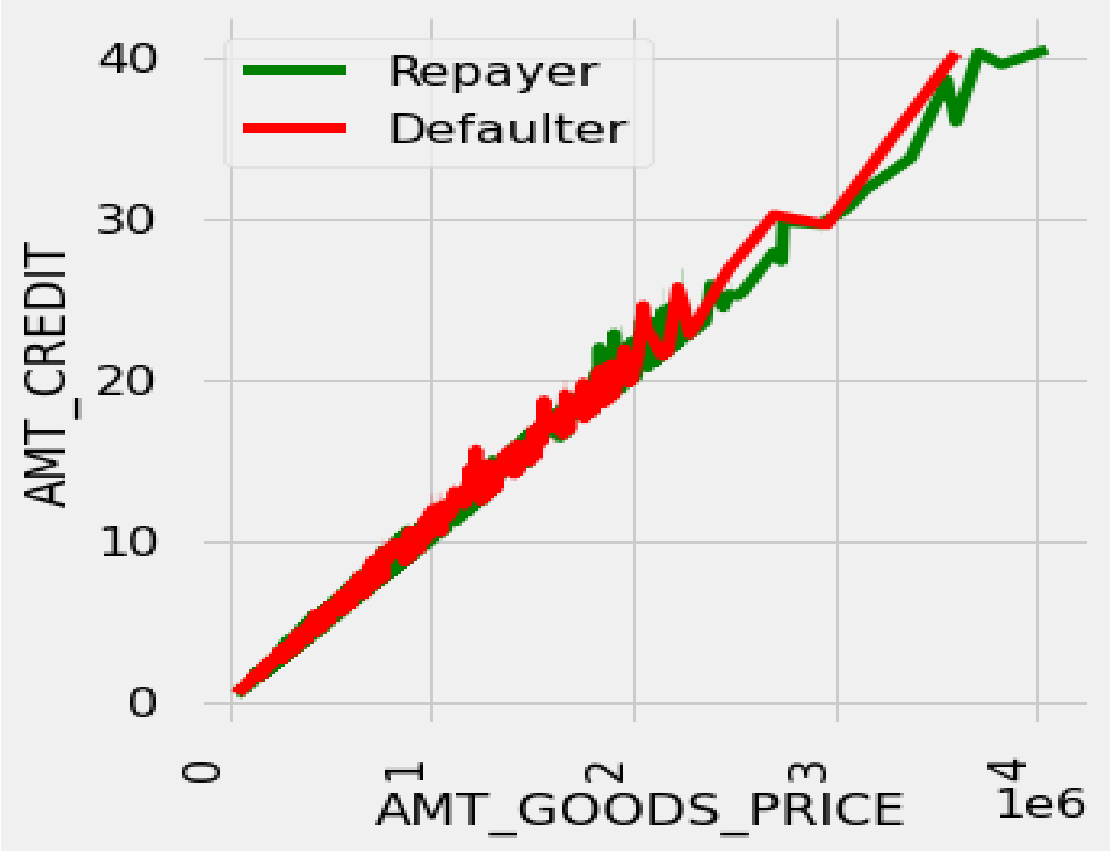
- Correlating factors amongst repayers:
- Credit amount is highly correlated with
- amount of goods price
- loan annuity
- total income
- We can also see that repayers have a high correlation in the number of days employed.



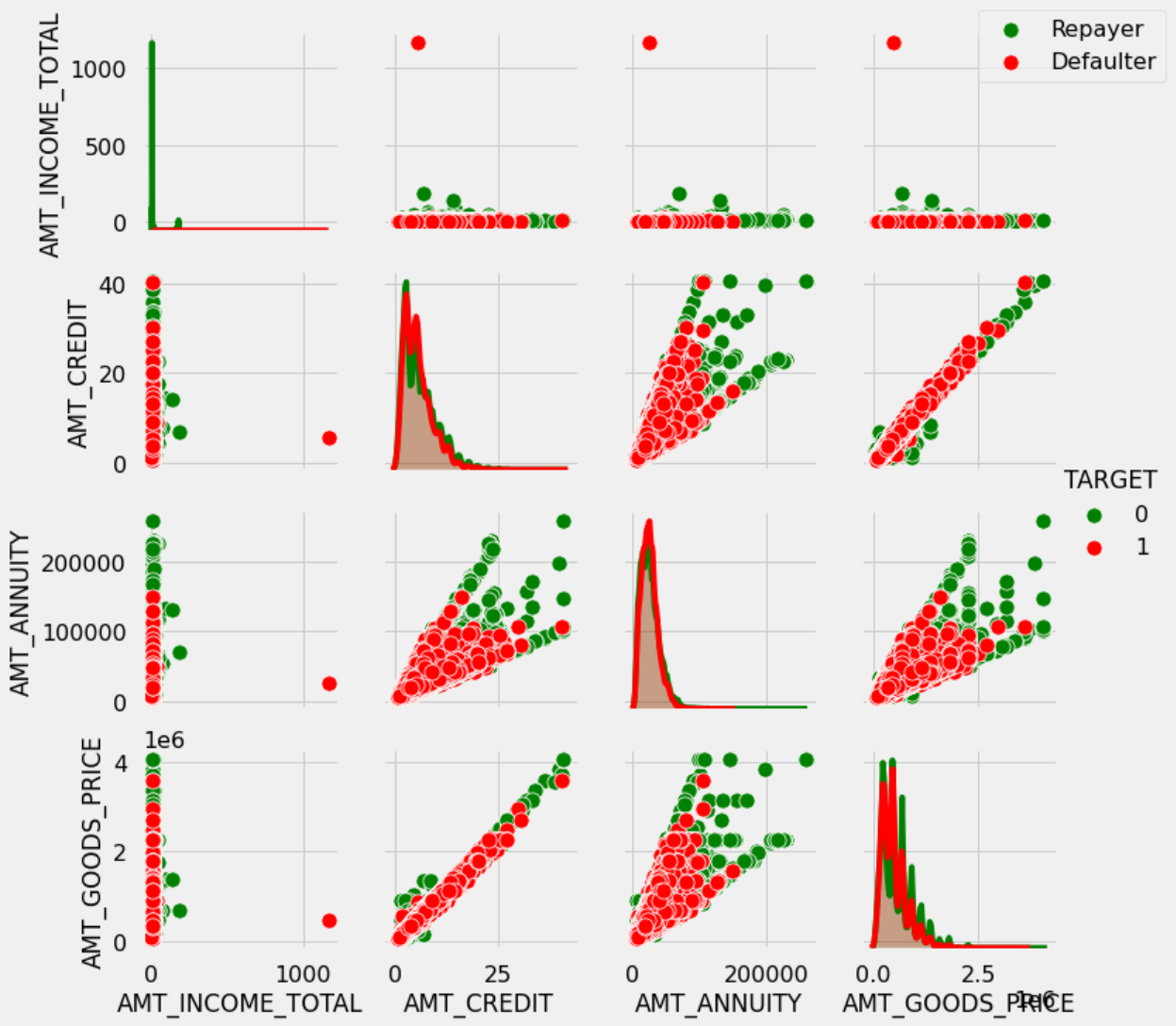
- Credit amount is highly correlated with amount of goods price which is same as repayers.
- But the loan annuity correlation with credit amount has slightly reduced in defaulters (0.75) when compared to repayers(0.77)
- We can also see that repayers have a high correlation in number of days employed (0.62) when compared to defaulters (0.58).
- There is a severe drop in the correlation between the 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)



- 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 repayers and defaulters' distribution overlap in all the plots and hence we cannot use any of these variables in isolation to decide.



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

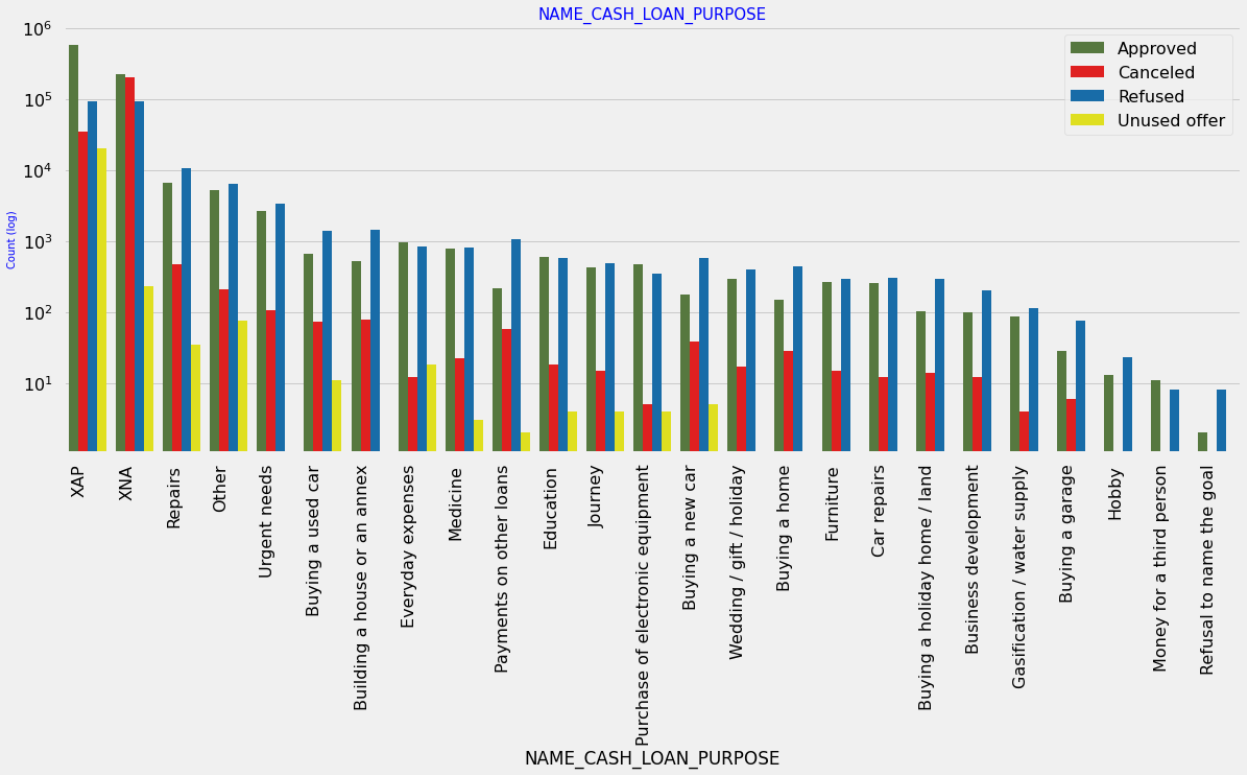


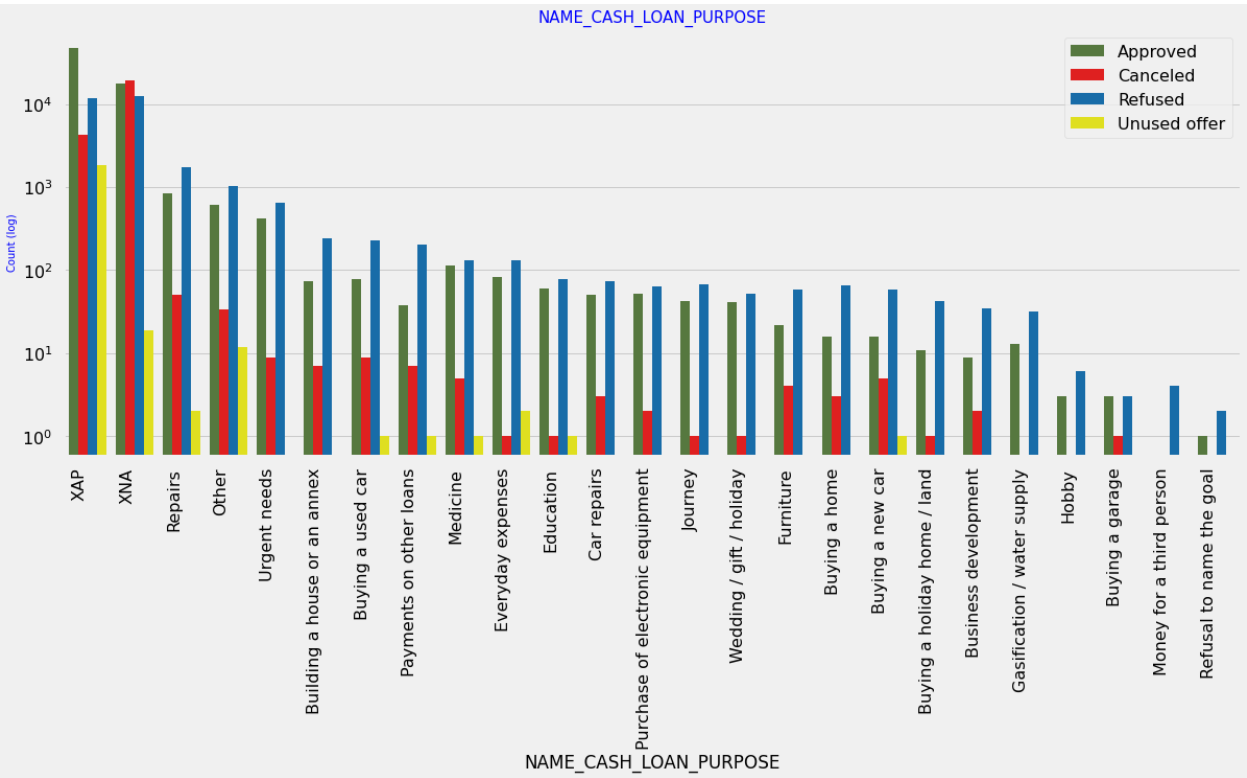
When amt_annuity >15000 & amt_goods_price> 3M, there is a lesser chance of defaulters

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

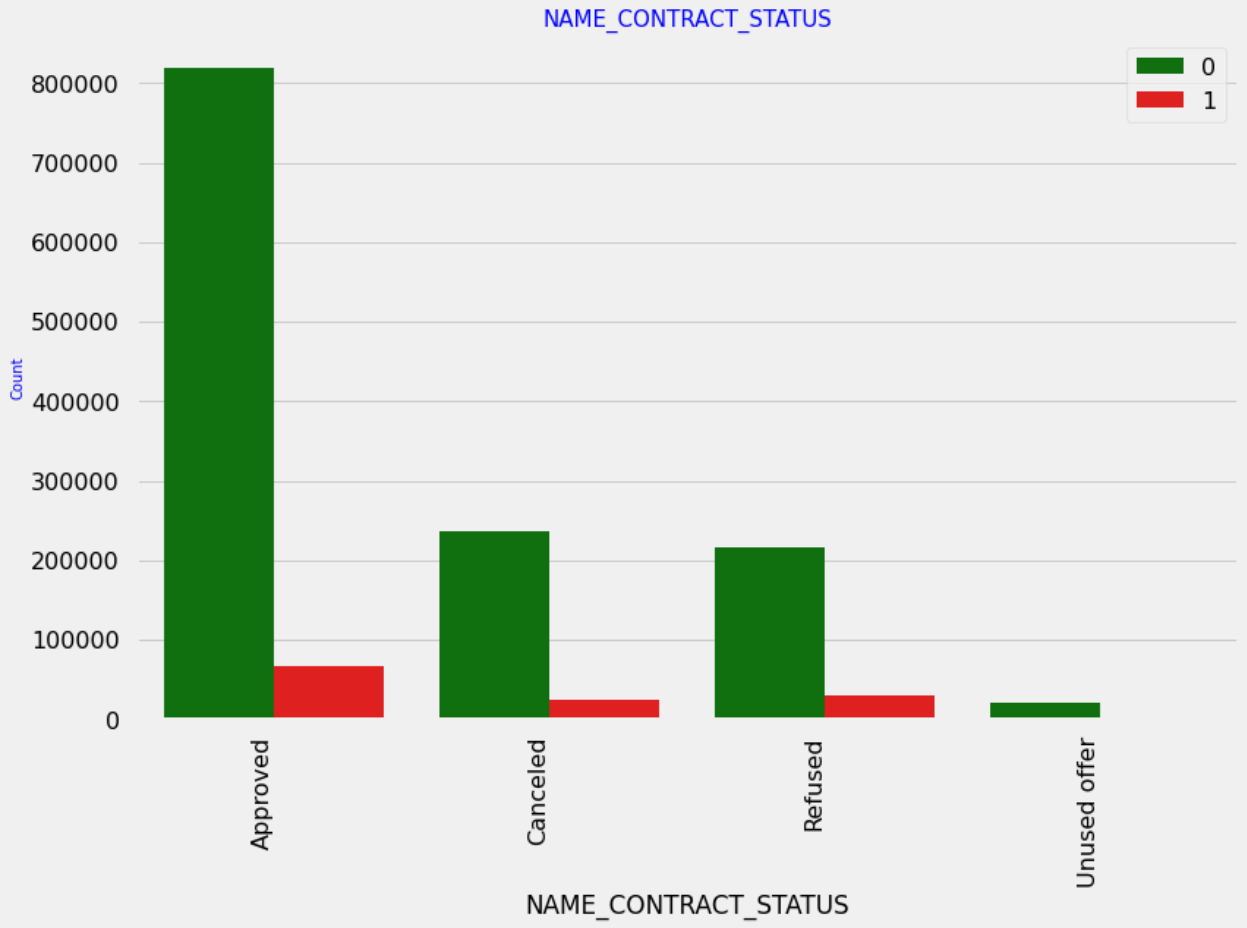
Inferences related to distribution plot has been already mentioned in previous distplot graphs inferences section.





- Loan purpose has high number of unknown values (XAP, XNA)

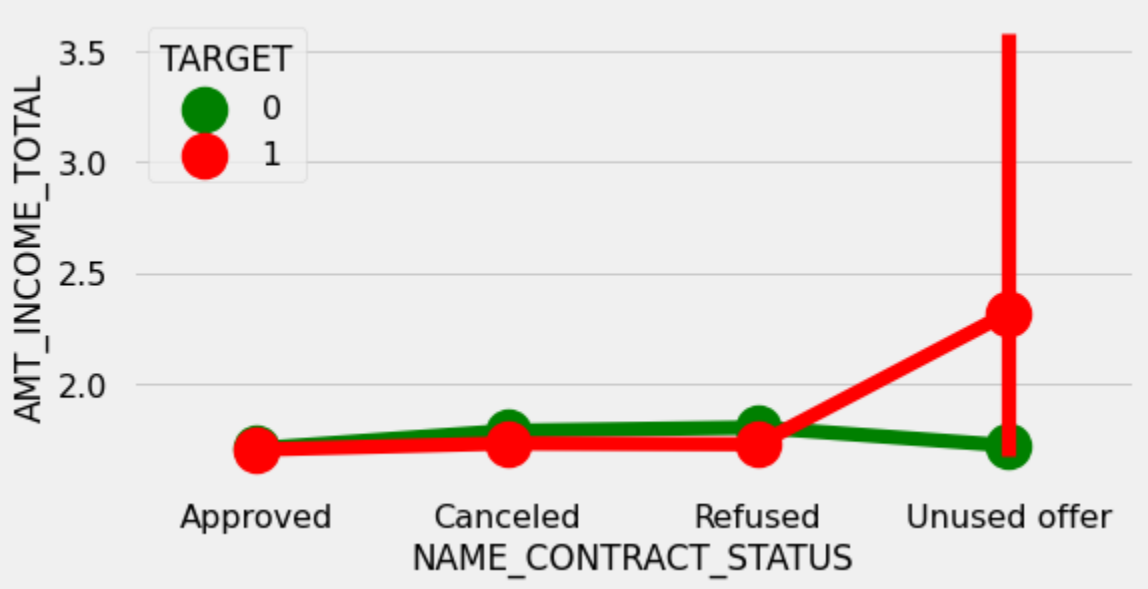
- Loan taken for the purpose of Repairs seems to have highest default rate
- A very high number of applications have been rejected by banks or refused by clients which have purpose as repair or other. This shows that the purpose of repair is taken as high risk by the bank and either they are rejected,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



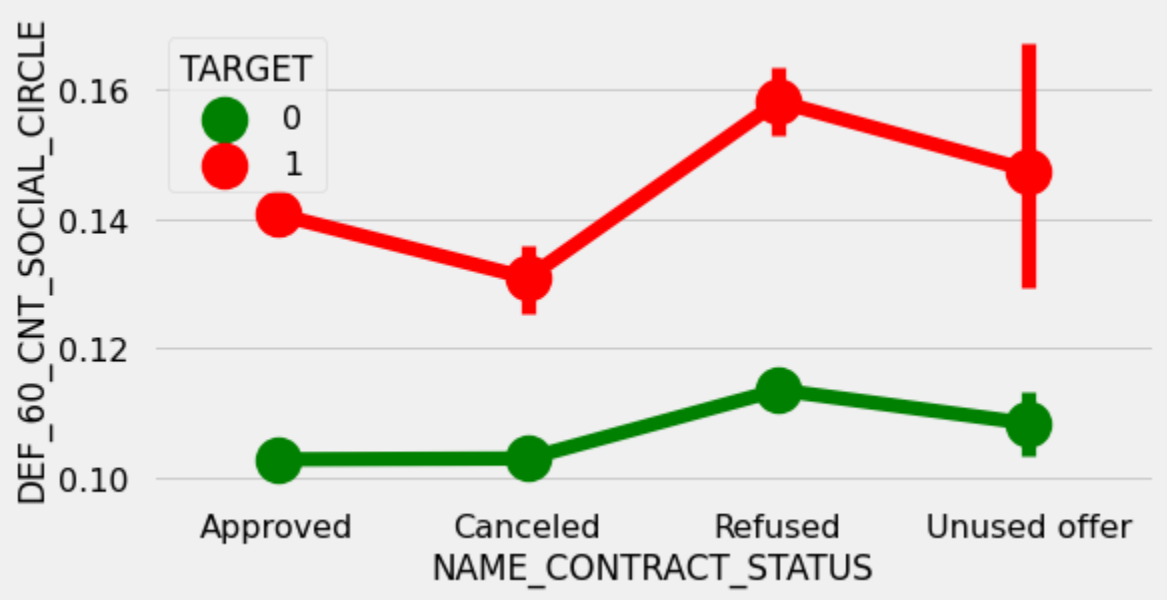
90% of the previously cancelled clients have repaid the loan. Revisiting the interest rates would increase business opportunity for these clients

88% of the clients who have been previously refused a loan have paid back the loan in the current case.

Refusal reasons should be recorded for further analysis as these clients would turn into potential repaying customers.



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



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