Credit EDA Assignment

Radhika Kute

Problem Statement:

- Analyzing Patterns present in the data using EDA to ensure that the applicants capable of repaying the loan are not rejected.
- There are Two types of risks associated with the bank’s decision, When it receives a loan application:
  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
  - If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Problem Statement:

- The Aim is 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.
- To understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

Steps Involved:

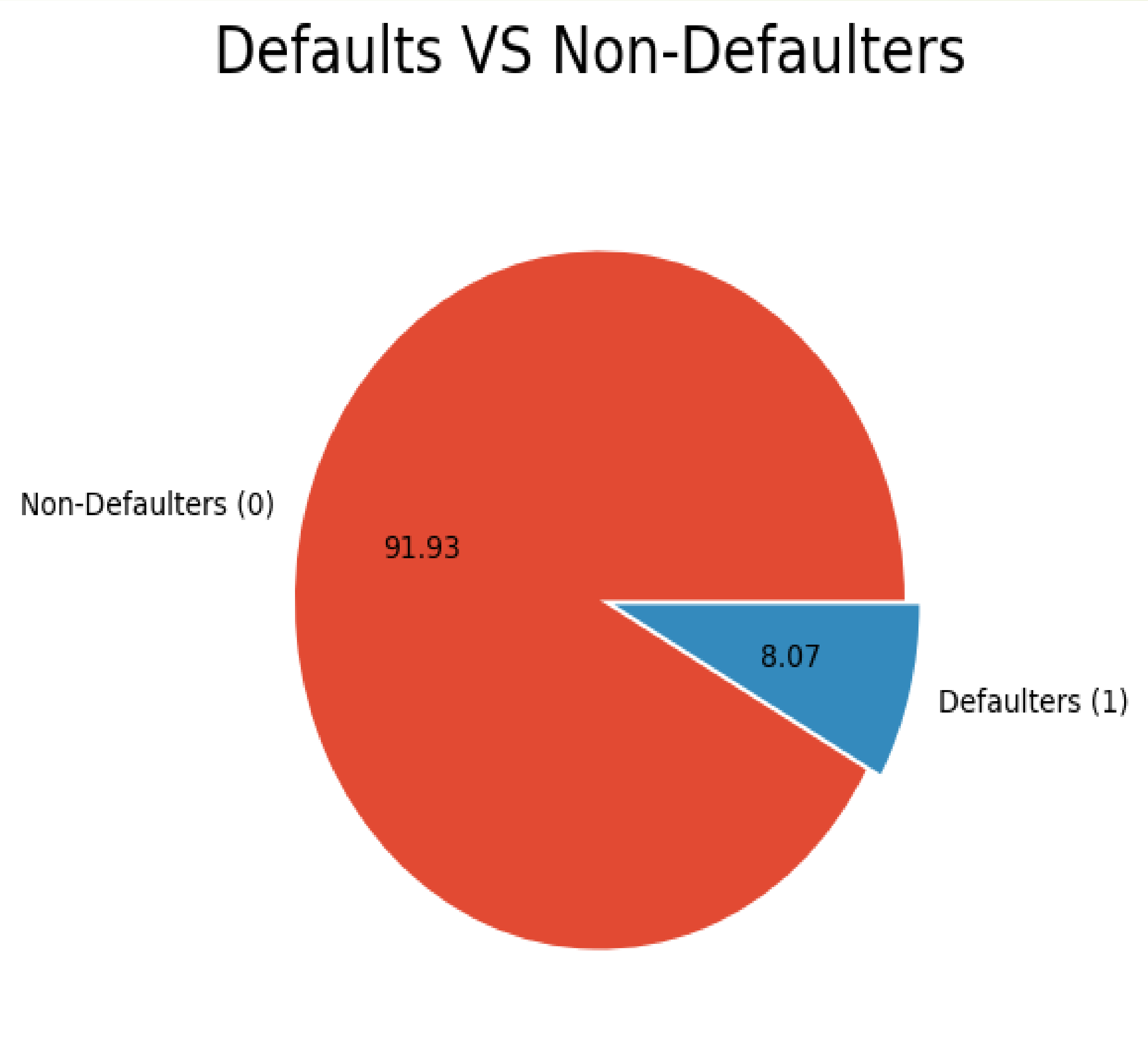
- Import Libraries & input files
- Data Inspection
- Data Cleaning
  1. Data Cleaning for df1 (application_data.csv):
     a) Handling Null Values in df1
     b) Imputing the null values with mean/median/mode values
     c) Standardising Values - Binning Numerical Columns to create a categorical column
     d) Data Type Conversion
  2. Data Cleaning for df2 (previous_application.csv):
     a) Handling Null Values in df2
     b) Imputing the null values with mean/median/mode values
     c) Standardising Values - Binning Numerical Columns to create a categorical column
     d) Data Type Conversion

Steps Involved:

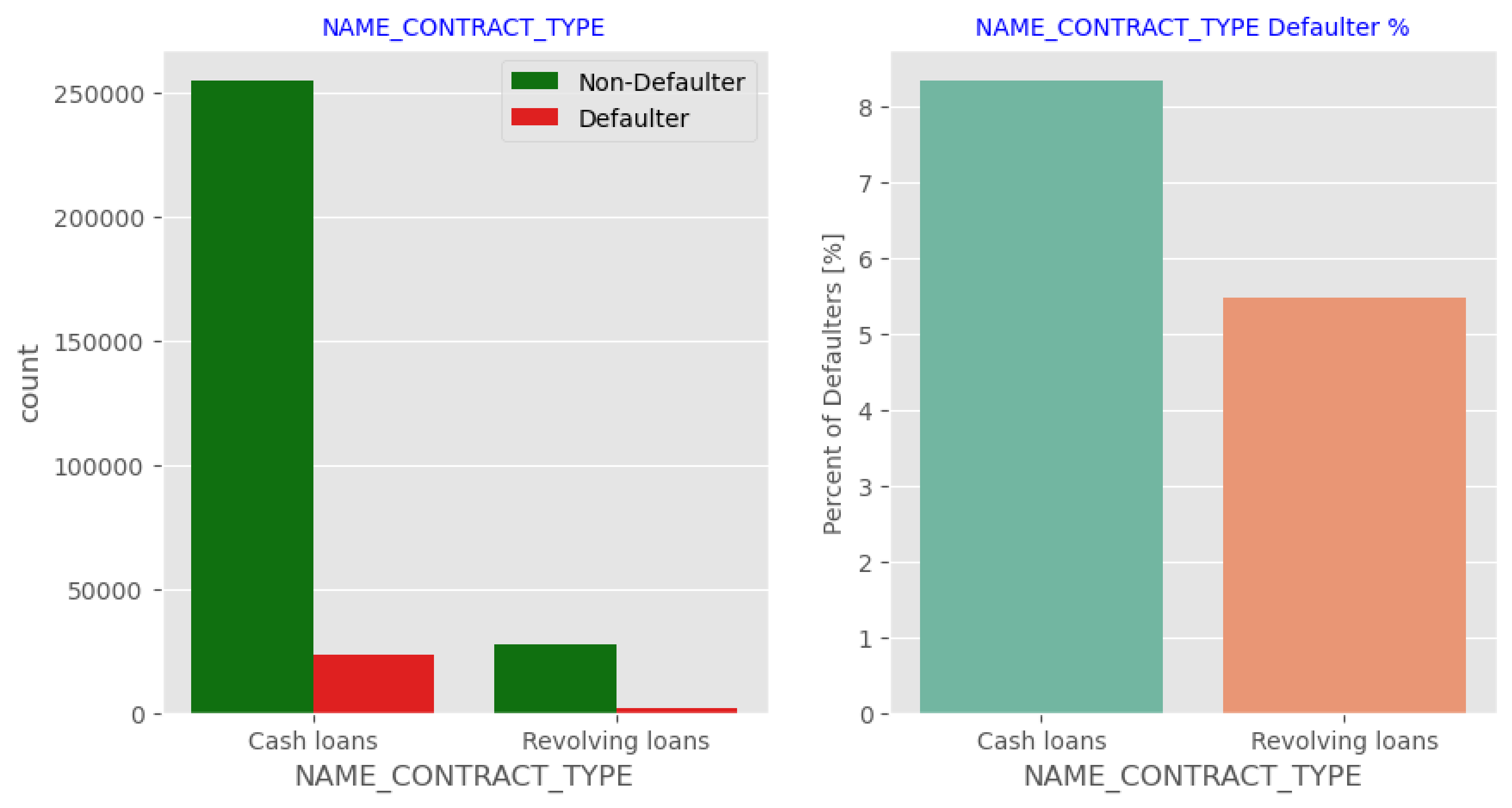
- Data Analysis
  - Imbalance Analysis – Target Variable
  - Categorical Univariate Analysis
  - Categorical Bi/Multivariate Analysis
  - Numerical Bivariate Analysis
  - Merged Dataframe Analysis
- Conclusion

Imbalance Analysis – Target Variable

From this plot we can see that, around 92% people paid their loans on time i.e Non-Defaulters and around 8% people had difficulty in paying their loans on time i.e Defaulters.

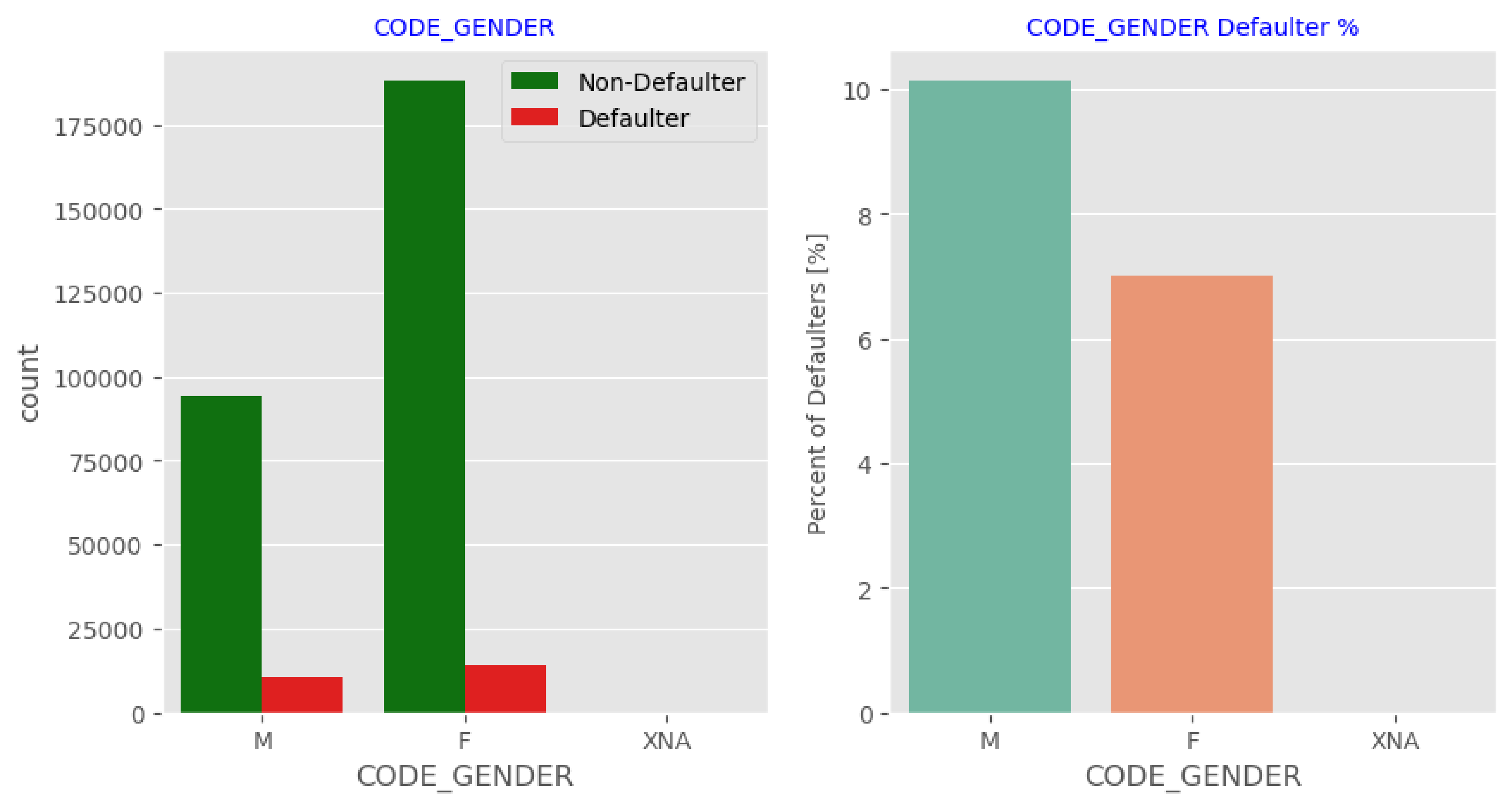


Categorical Univariate Analysis - Contract Type



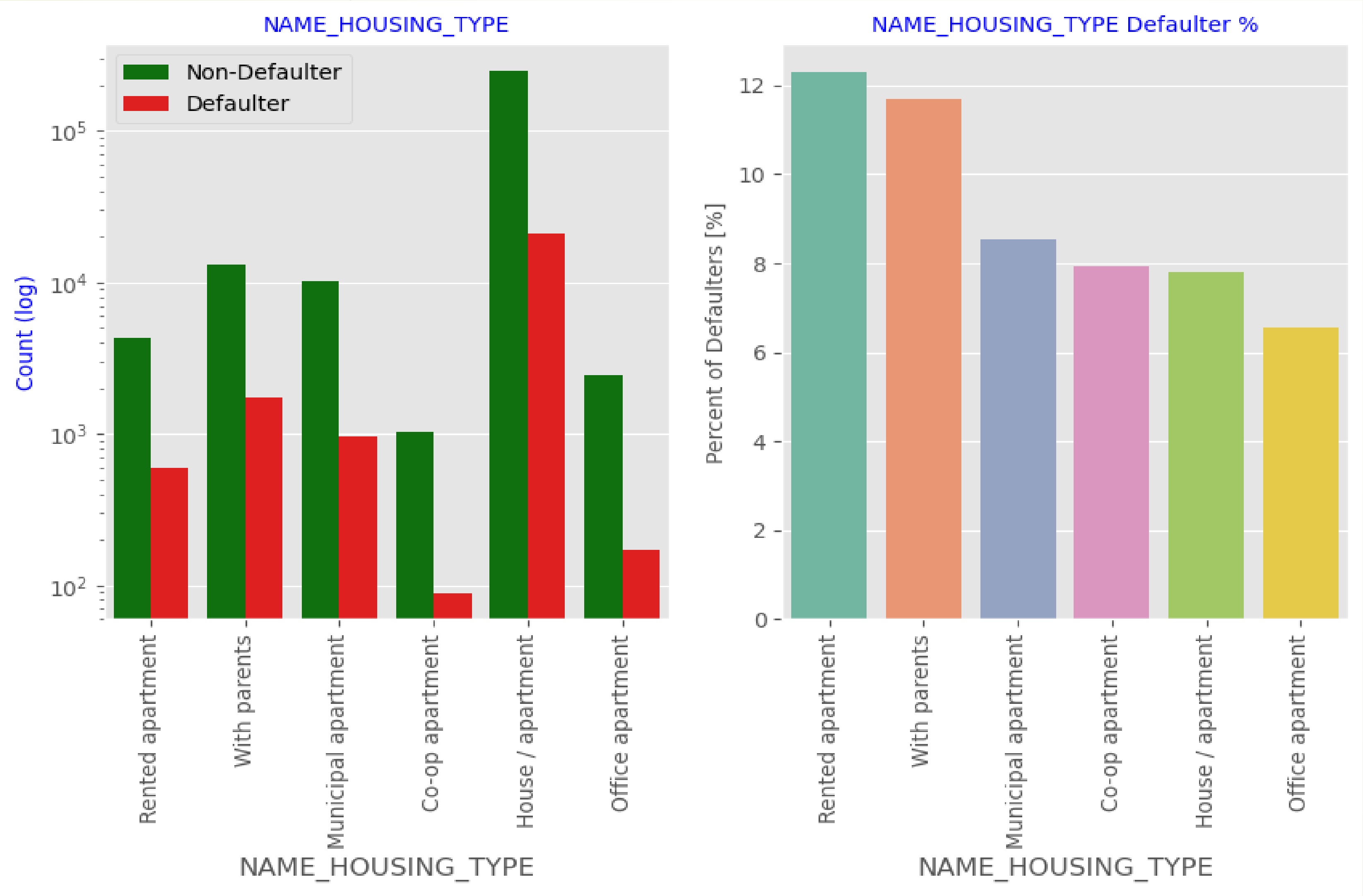
- Revolving loans contribute a small fraction from the total number of loans while Majority are Cash Loans ; but a larger amount of Revolving loans, comparing with their frequency, are not repaid.

Categorical Univariate Analysis - Gender



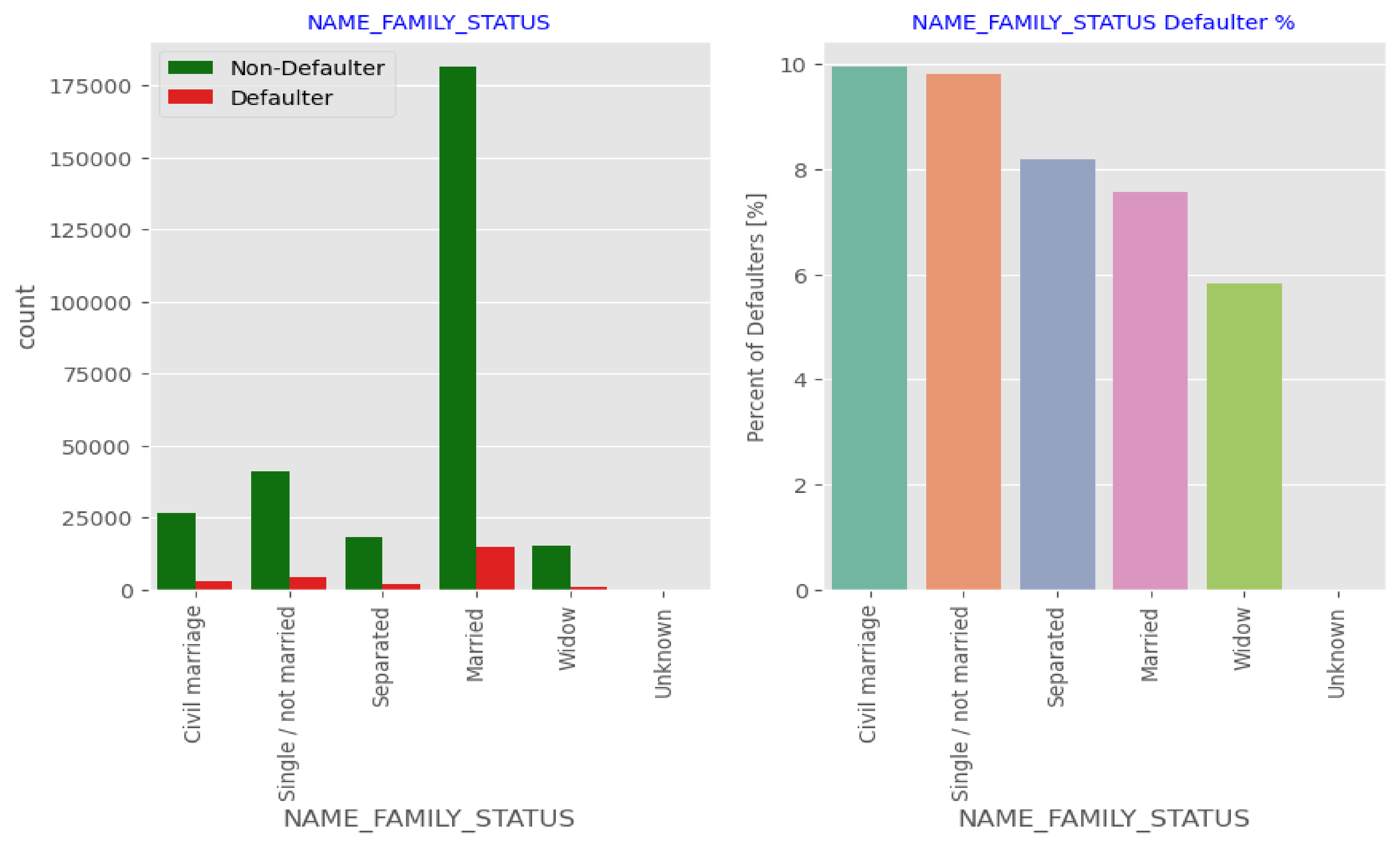
- The number of female clients is almost double the number of male clients.
- Based on the percentage of defaulters, males are more likely to default in repaying their loans (10%), in comparison with women (7%)

Categorical Univariate Analysis - Housing Type



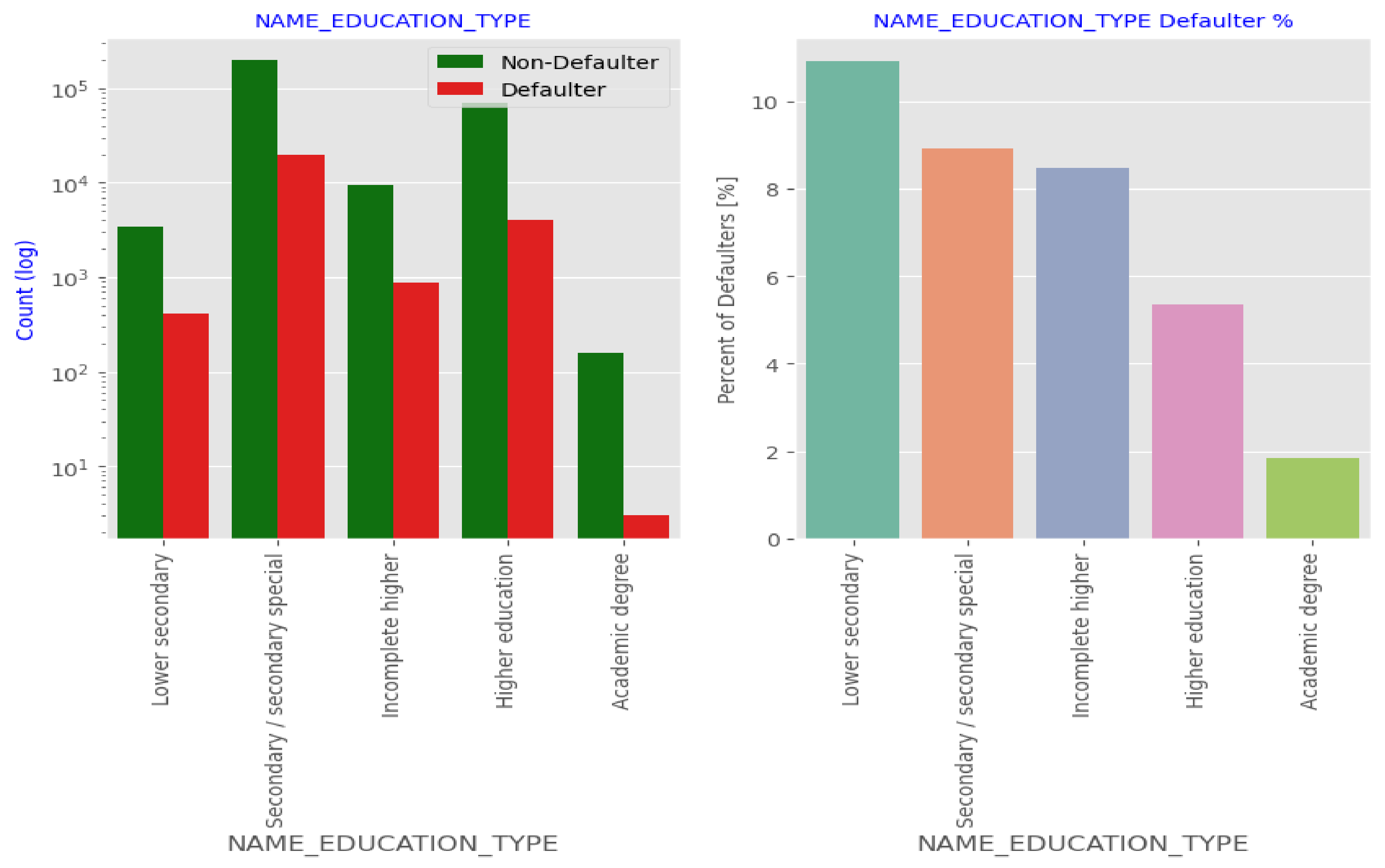
- Majority of people live in House/apartment,
- People living in office apartments have lowest default rate,
- People living in rented apartments and with parents have higher probability of defaulting

Categorical Univariate Analysis - Family Status



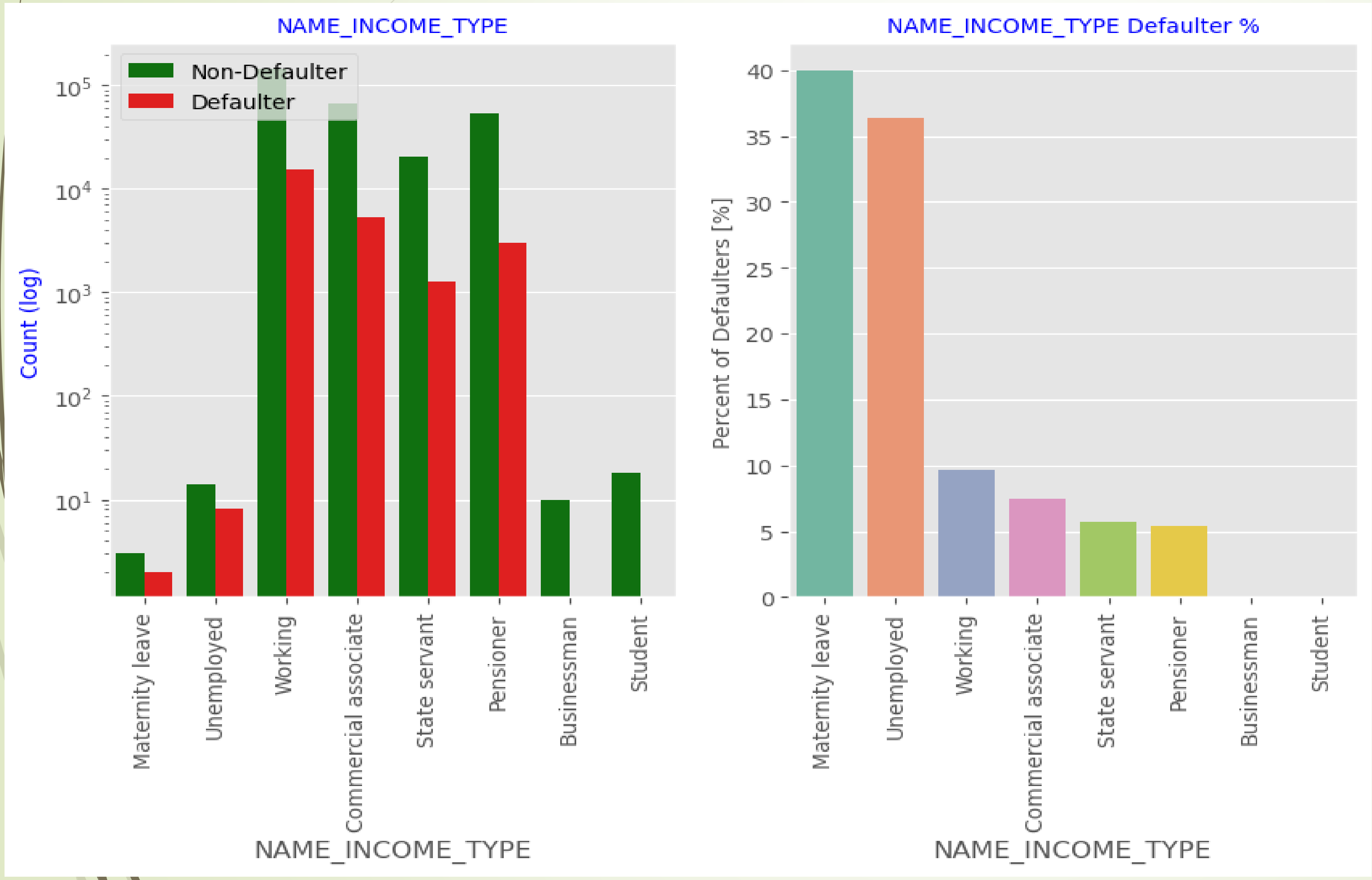
- Most of the people who have taken loan are married, followed by Single/not married and civil marriage
- In terms of defaulter percentage, Civil marriage and Single/not married has the highest percent of not repayment, with Widow the lowest.

Categorical Univariate Analysis - Education Type



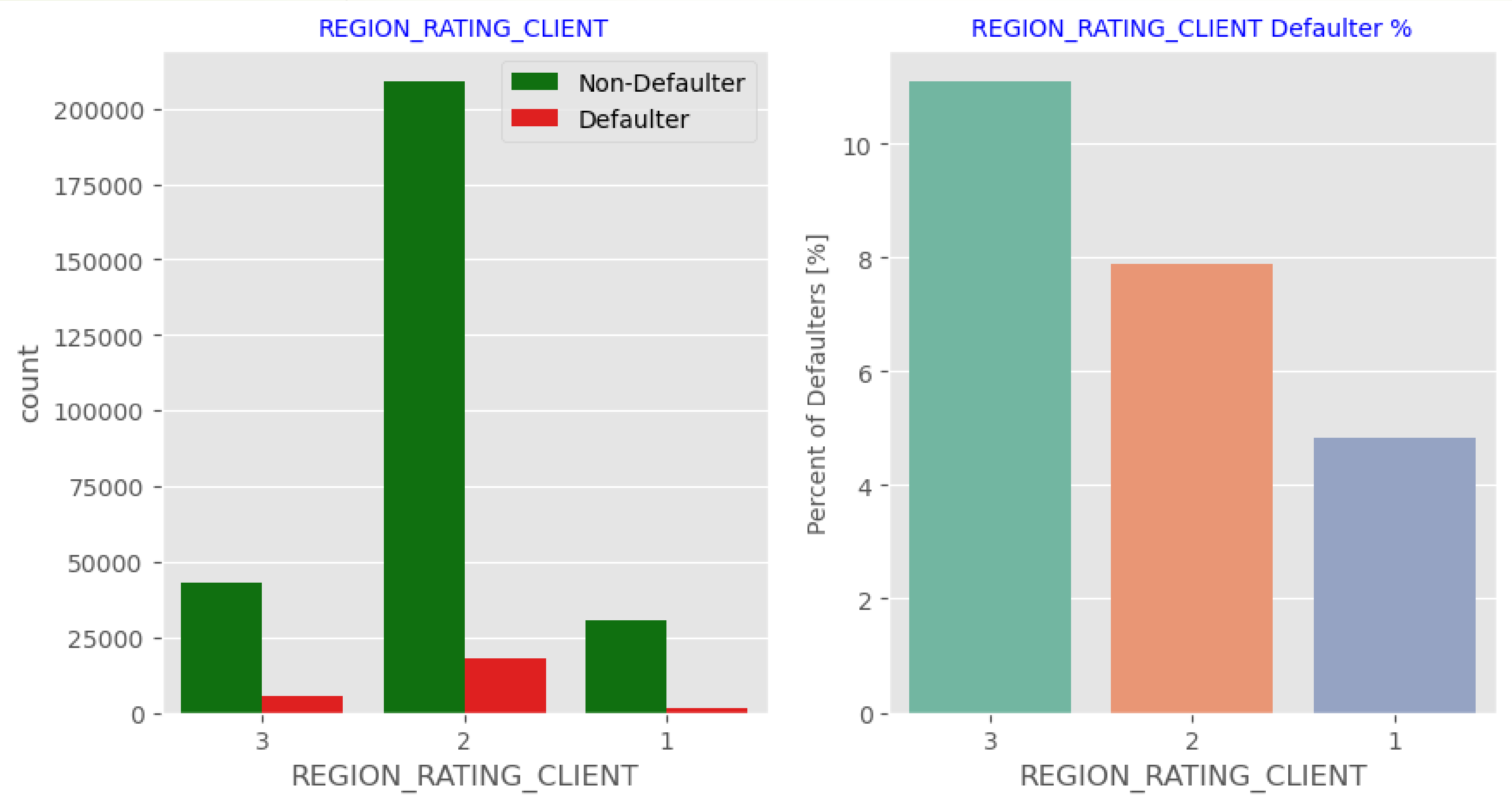
- Majority of the clients have Secondary / secondary special education, followed by clients with Higher education. Only a very small number having an academic degree
- The Lower secondary category, have the highest defaulter rate (11%) while the ones with Academic degree have the lowest defaulter rate.

Categorical Univariate Analysis - Income Type



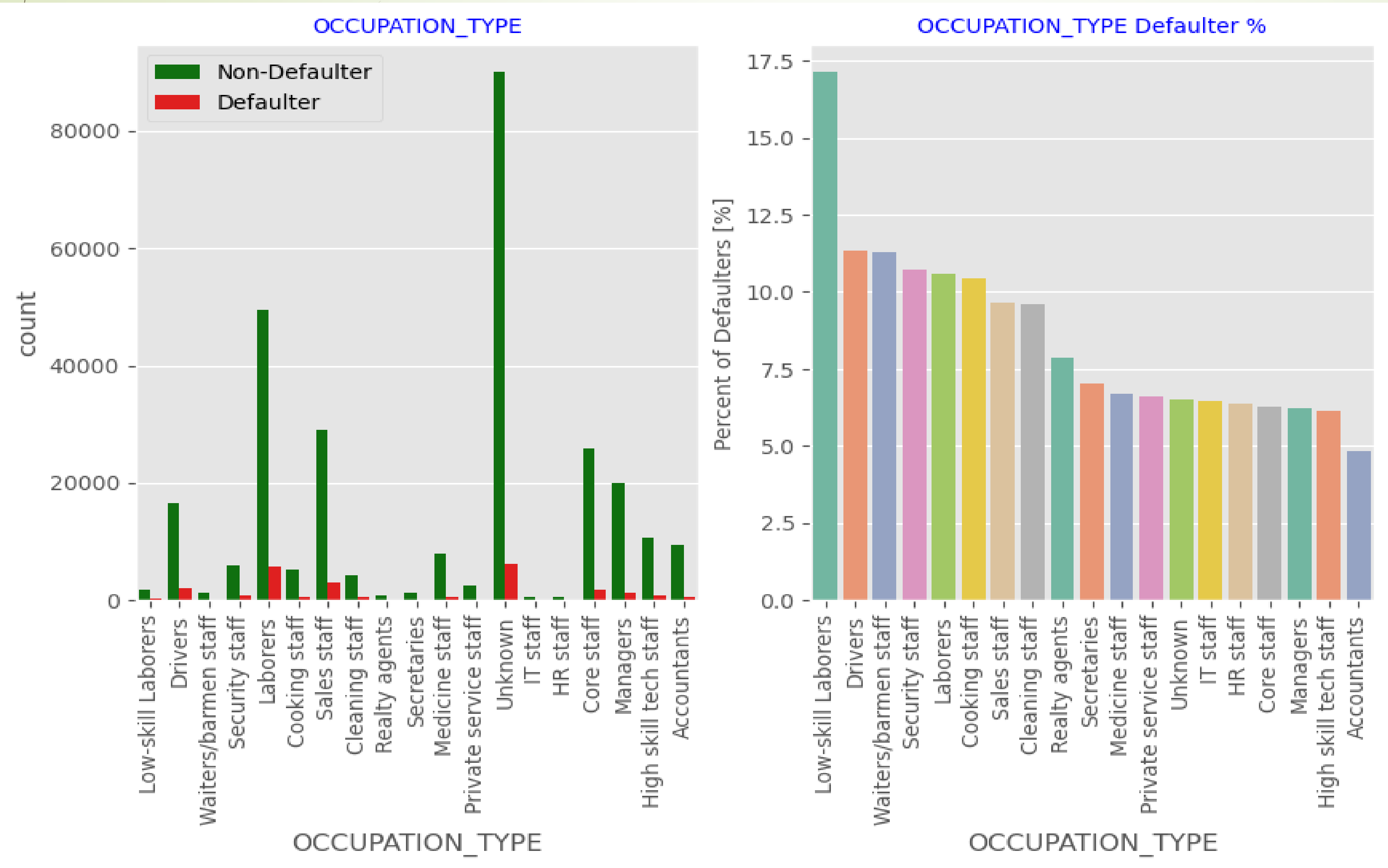
- Most of applicants for loans have income type as Working, followed by Commercial associate, Pensioner and State servant.
- The applicants with income type as Maternity leave have highest defaulter rate (40%), followed by Unemployed (37%).
- The rest of the income types are under the average defaulter rate of 10%. Student and Businessmen, though less in numbers do not have any default record.

Categorical Univariate Analysis - Region Rating



- Most of the applicants are living in Region_Rating 2.
- Region Rating 3 has the highest default rate.

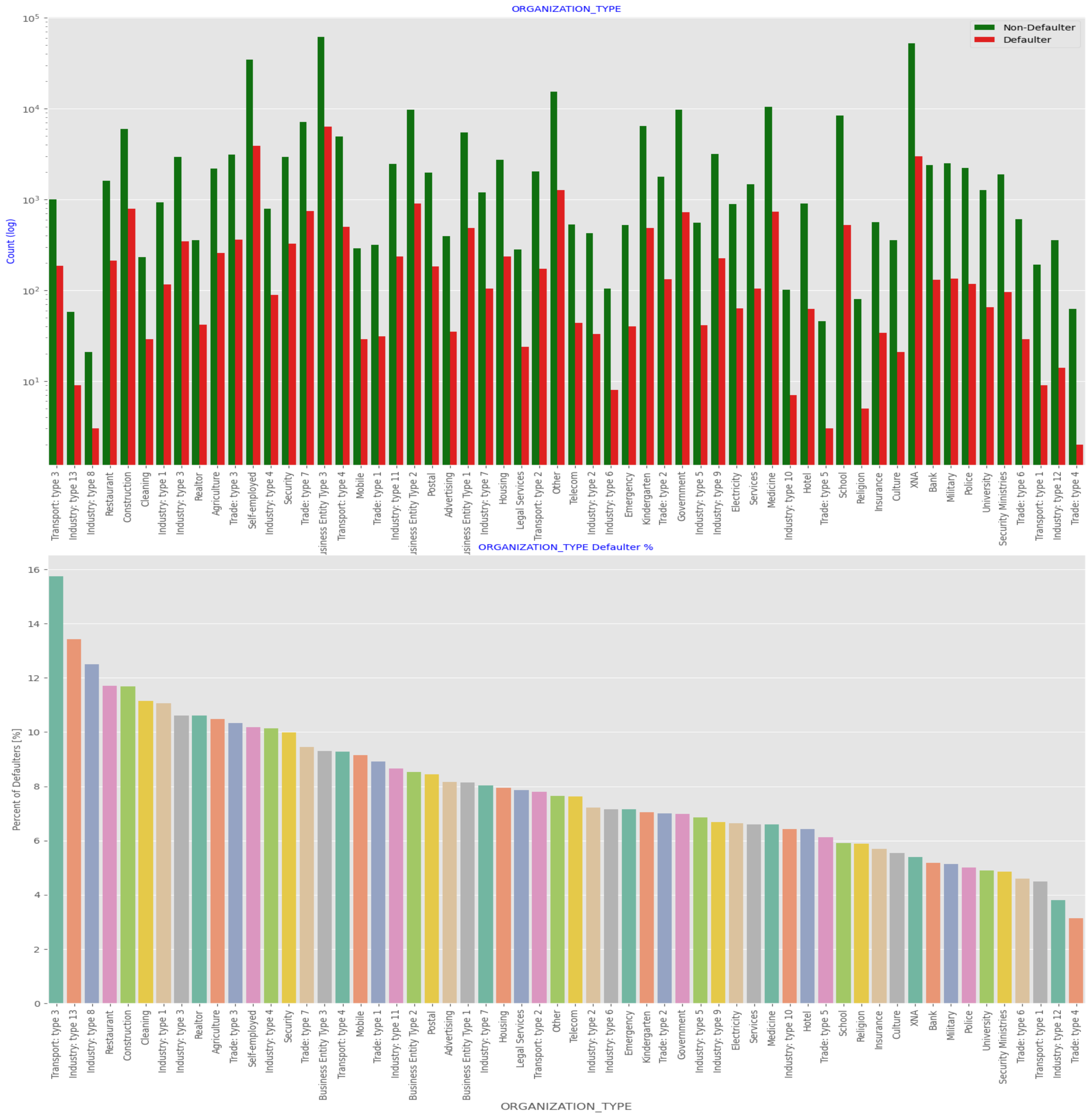
Categorical Univariate Analysis - Occupation Type



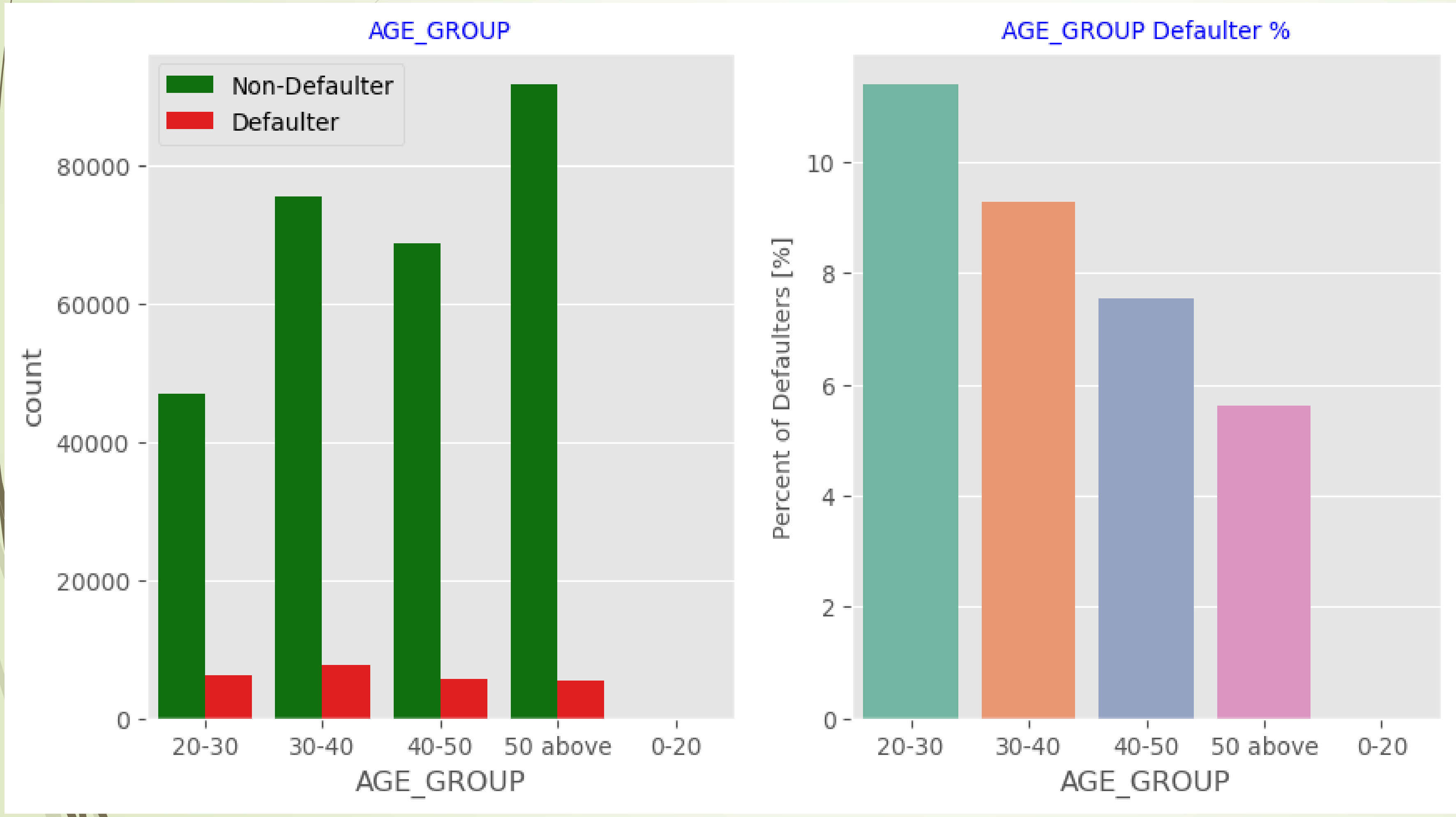
- Most of the loans are taken by Laborers, followed by Sales staff.
- IT staff & HR Staff take the lowest amount of loans.
- Low-skill Laborers have the highest defaulter rate.

Categorical Univariate Analysis – Organization Type

- Business Entity Type 3 has the highest number of loan applications.
- Organizations having highest default rate are Transport: type 3, Industry: type 13, Industry: type 8, Restaurant and Construction.

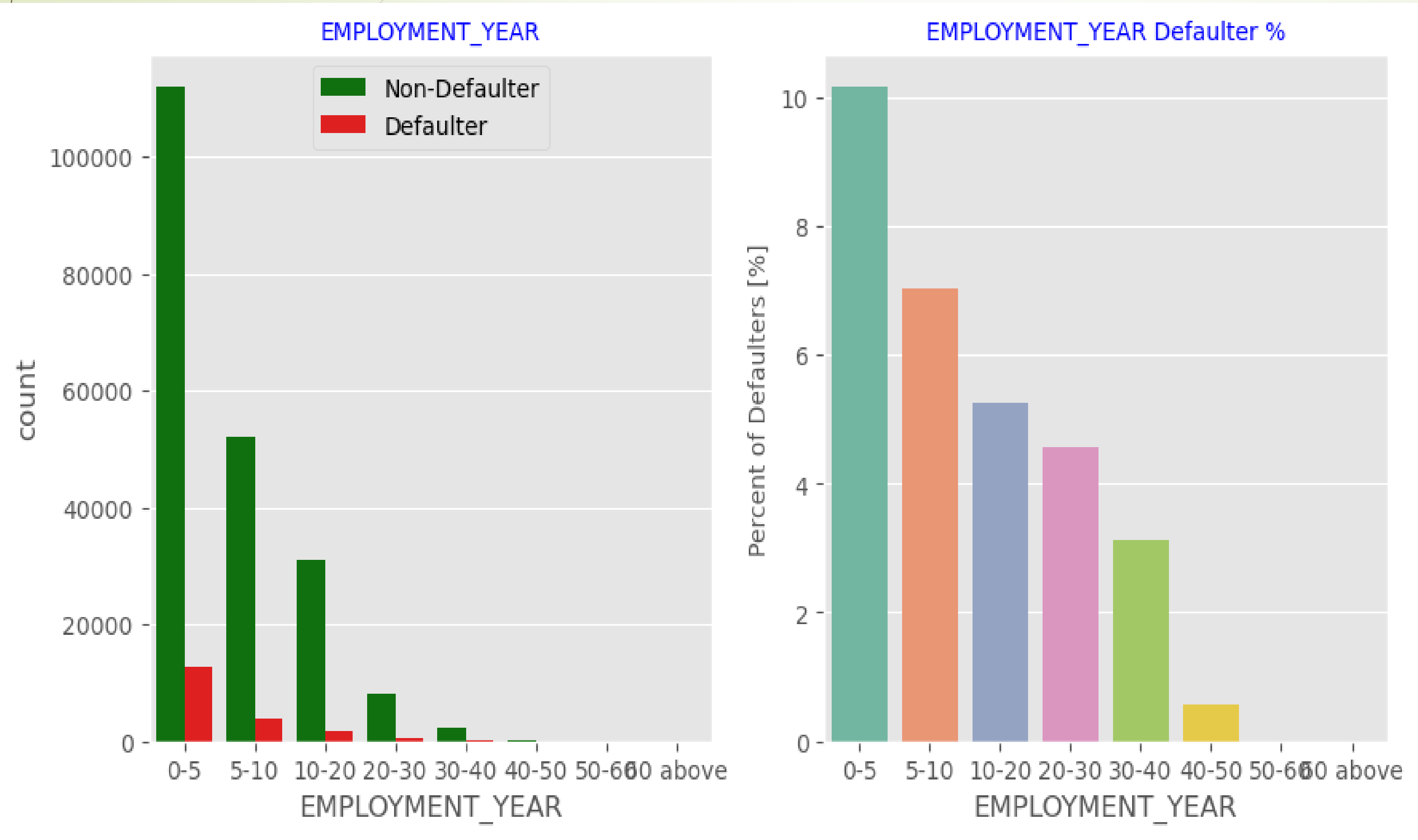


Categorical Univariate Analysis - Age Group



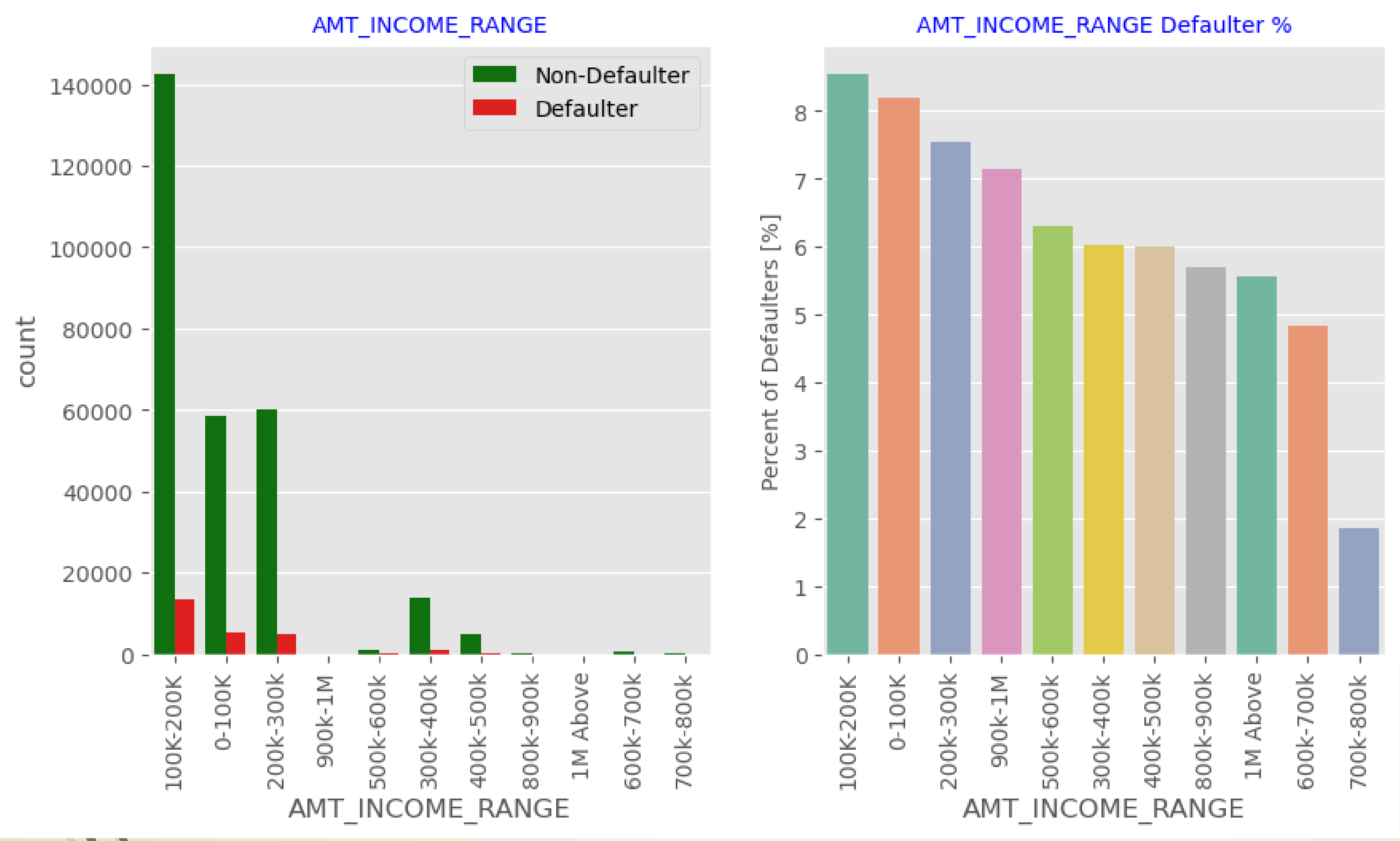
- People above the age of 50 have the highest number of loan applications and low defaulting rate.
- People in the age group range 20-40 have highest defaulting rate.

Categorical Univariate Analysis - Employment Year



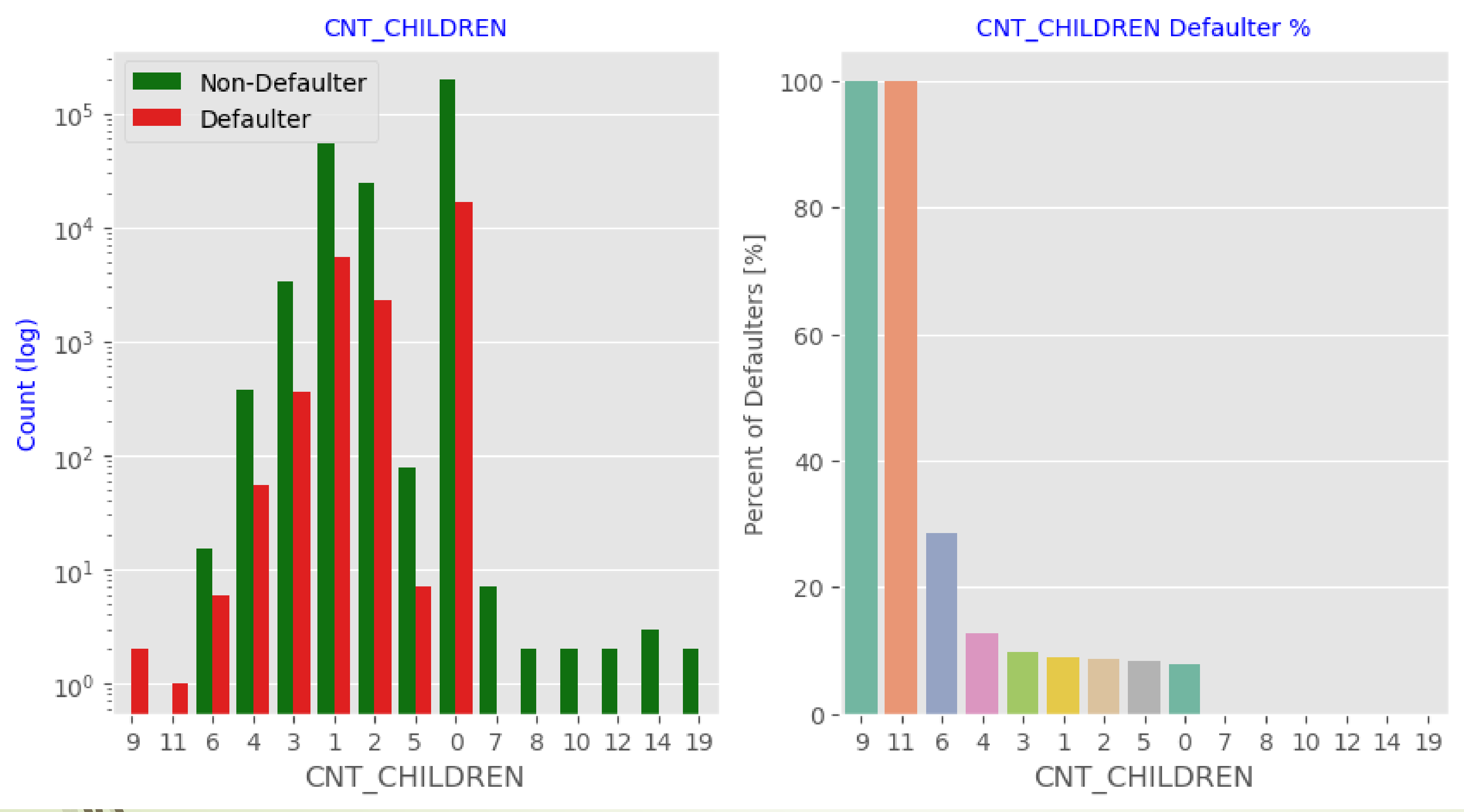
- Majority of the applicants have been employed in between 0-5 years. The defaulter rate of this group is also the highest.
- With increase of employment year, defaulting rate is gradually decreasing with people having 40+ year experience having lowest defaulter rate.

Categorical Univariate Analysis - Income Range



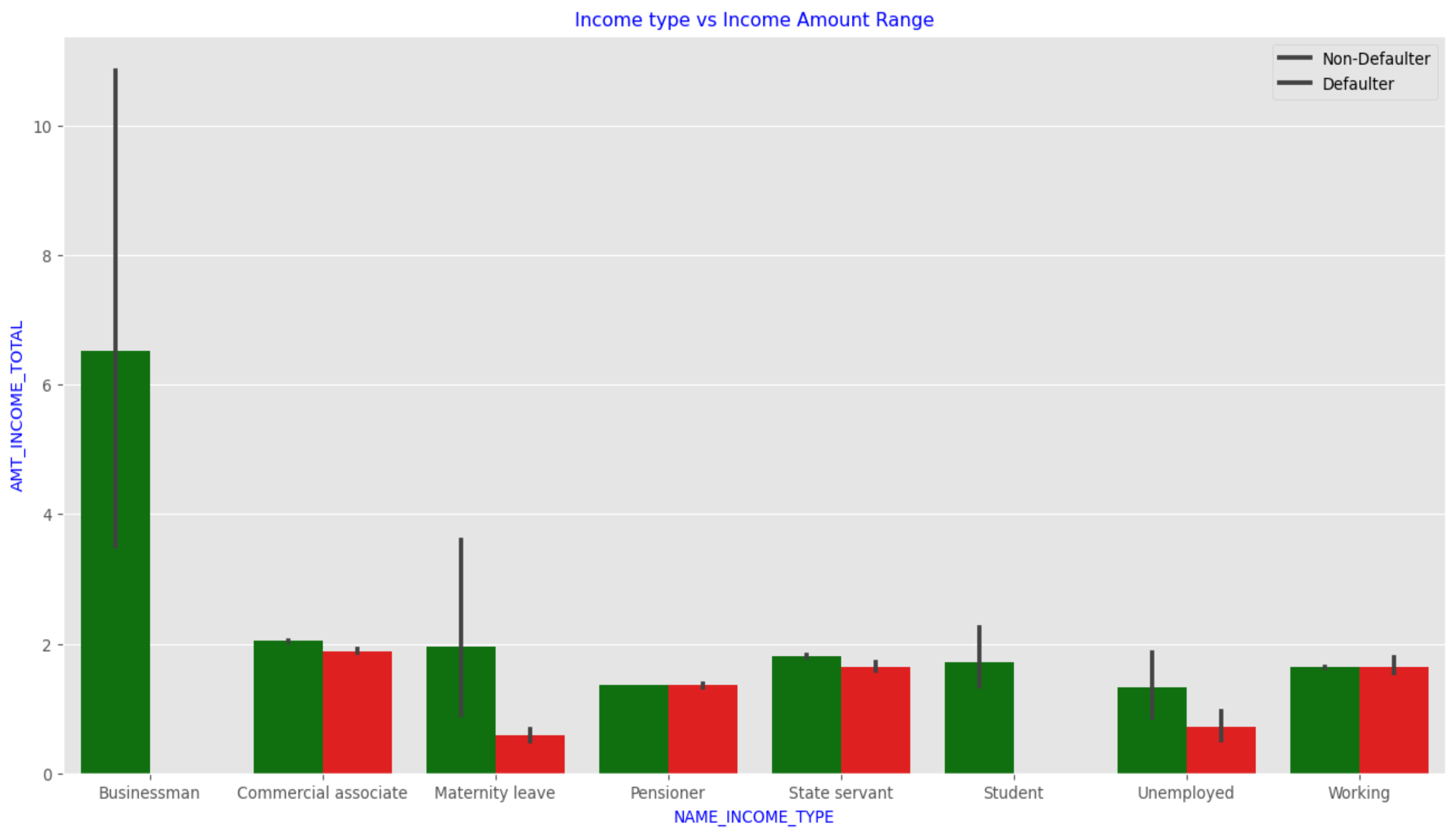
- Majority of the applicants have Income less than 300k, which also has the highest defaulter rate.
- Applicants with Income more than 700k have lowest defaulter rate.

Categorical Univariate Analysis - Children Count



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

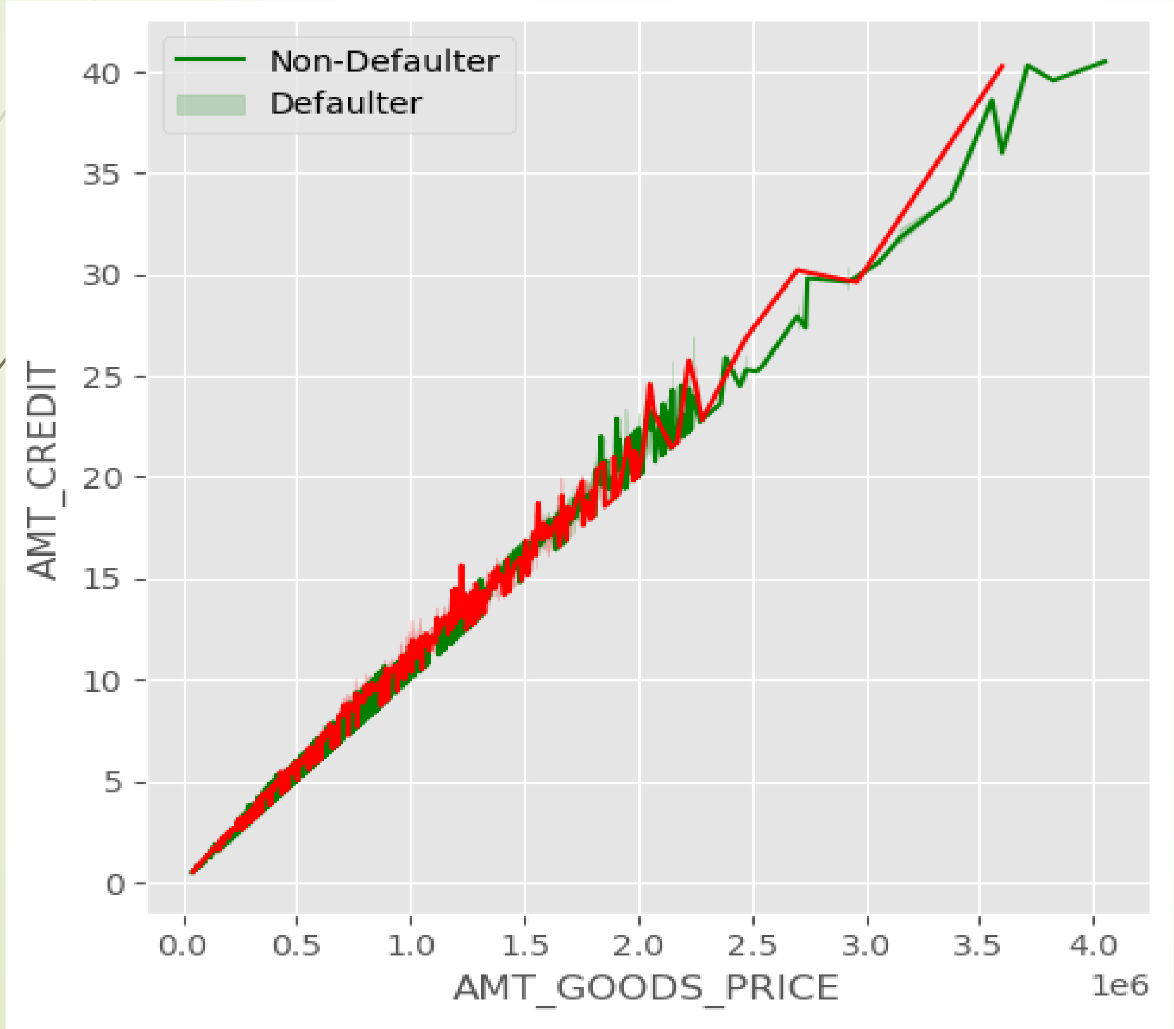
Categorical Bi/Multivariate Analysis



- Business man's income is the highest
- The estimated range 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

Numerical Bivariate Analysis

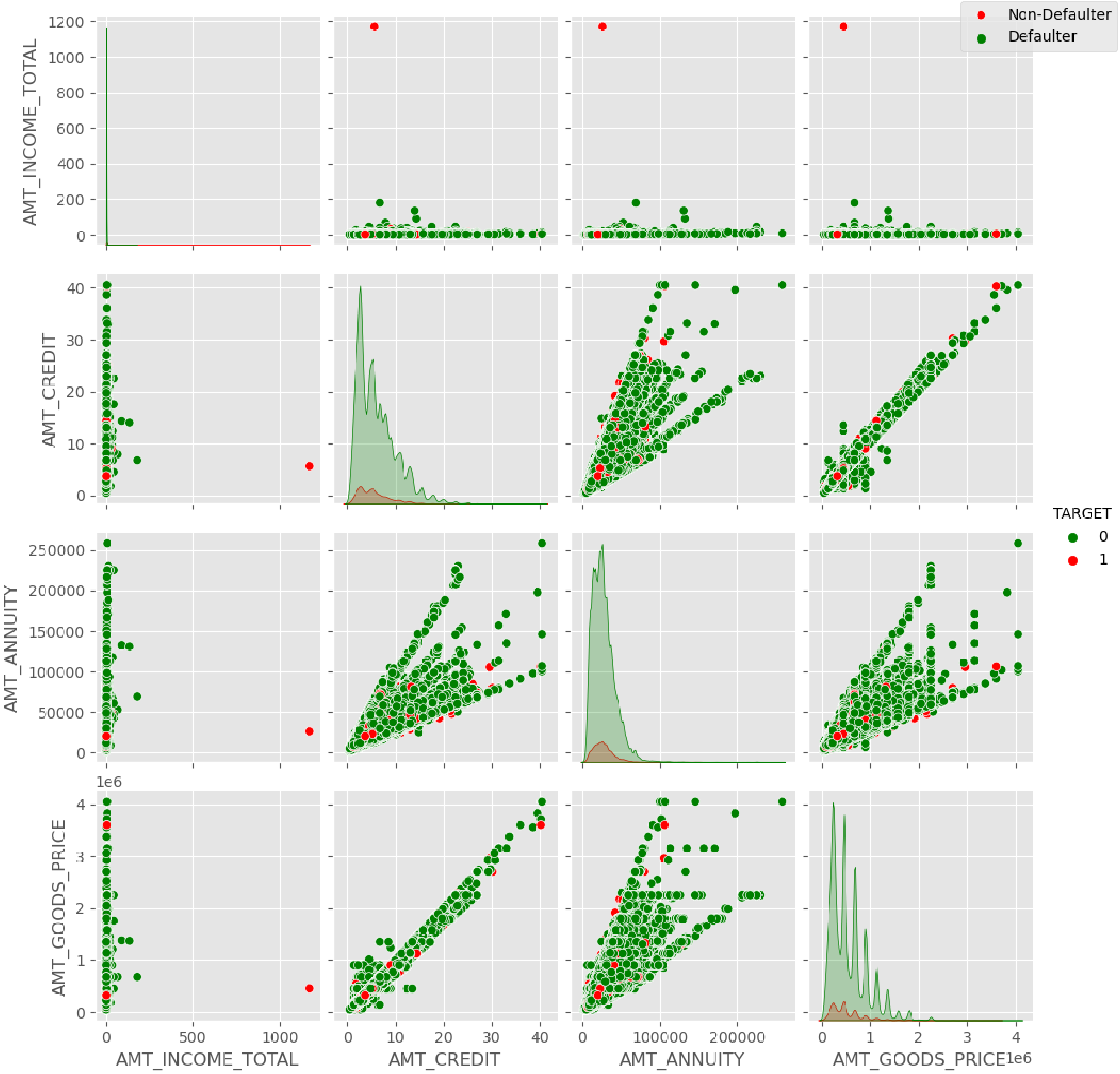
Relationship between Goods price and credit and comparing with loan repayment status



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

Amount Variables vs Target Variable

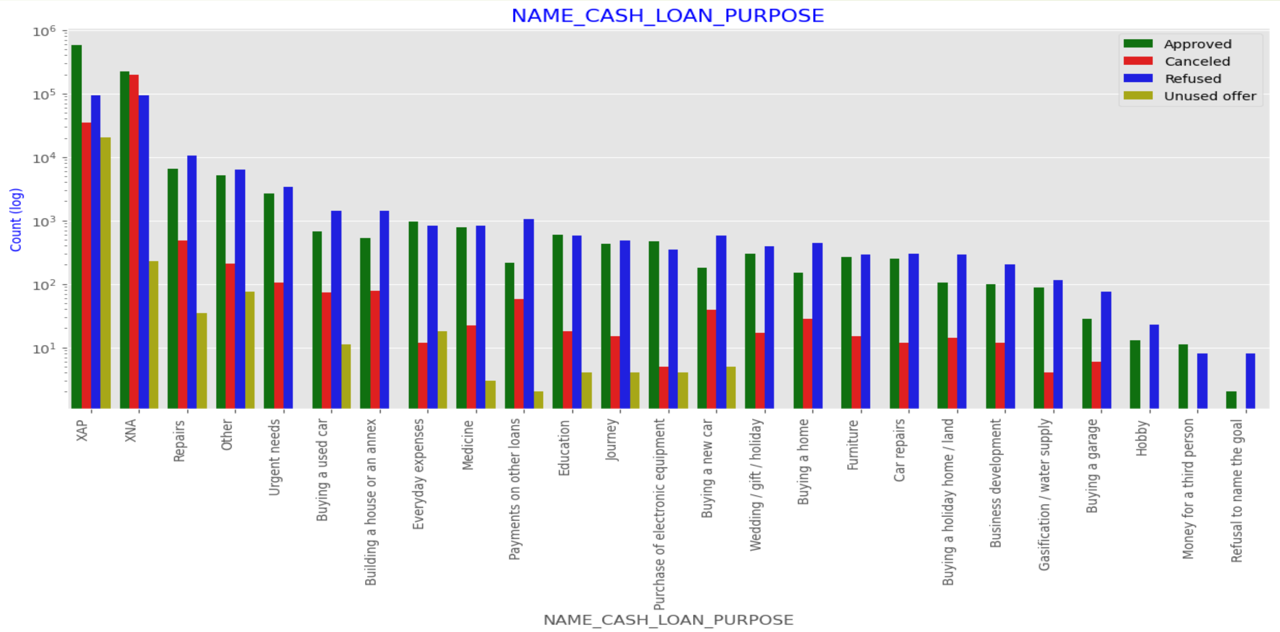
- When amt_annuity >15000 amt_goods_price> 3M, there is a lesser chance of defaulters
- AMT_CREDIT and AMT_GOODS_PRICE are highly correlated as based on the scatterplot where most of the data are consolidated in form of a line
- There are very less defaulters for AMT_CREDIT >3M



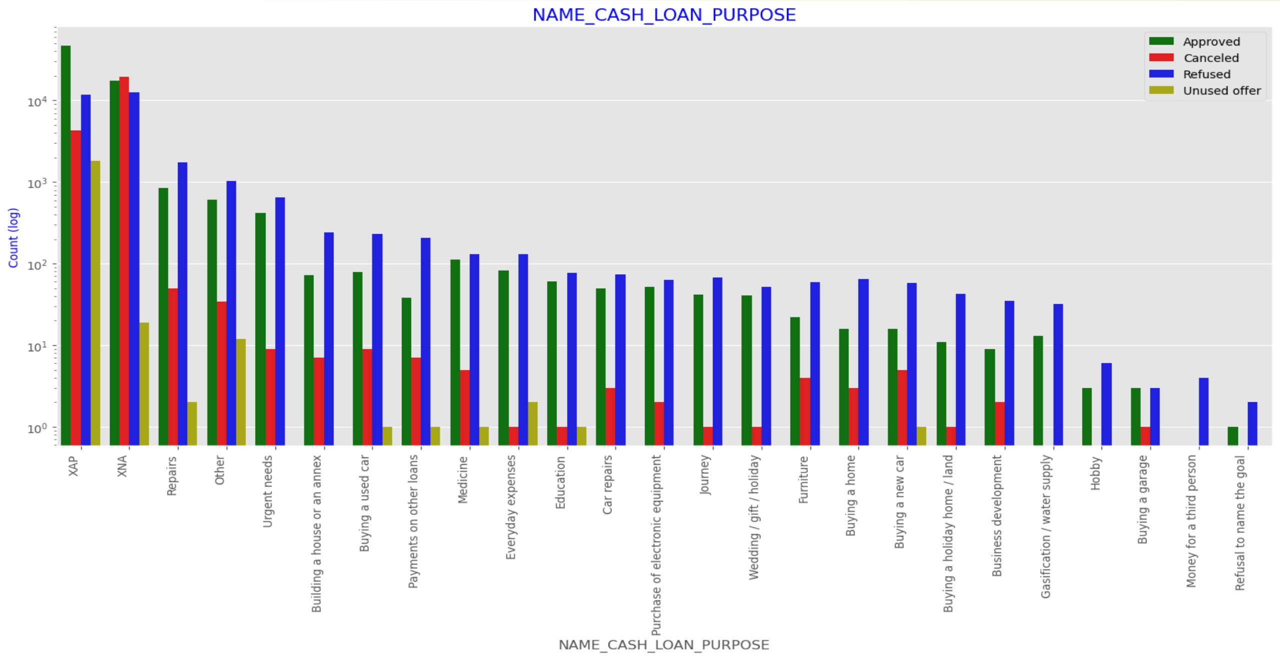
Merged Dataframes Analysis

- 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.

Contract Status vs purpose of the loan for Target 0

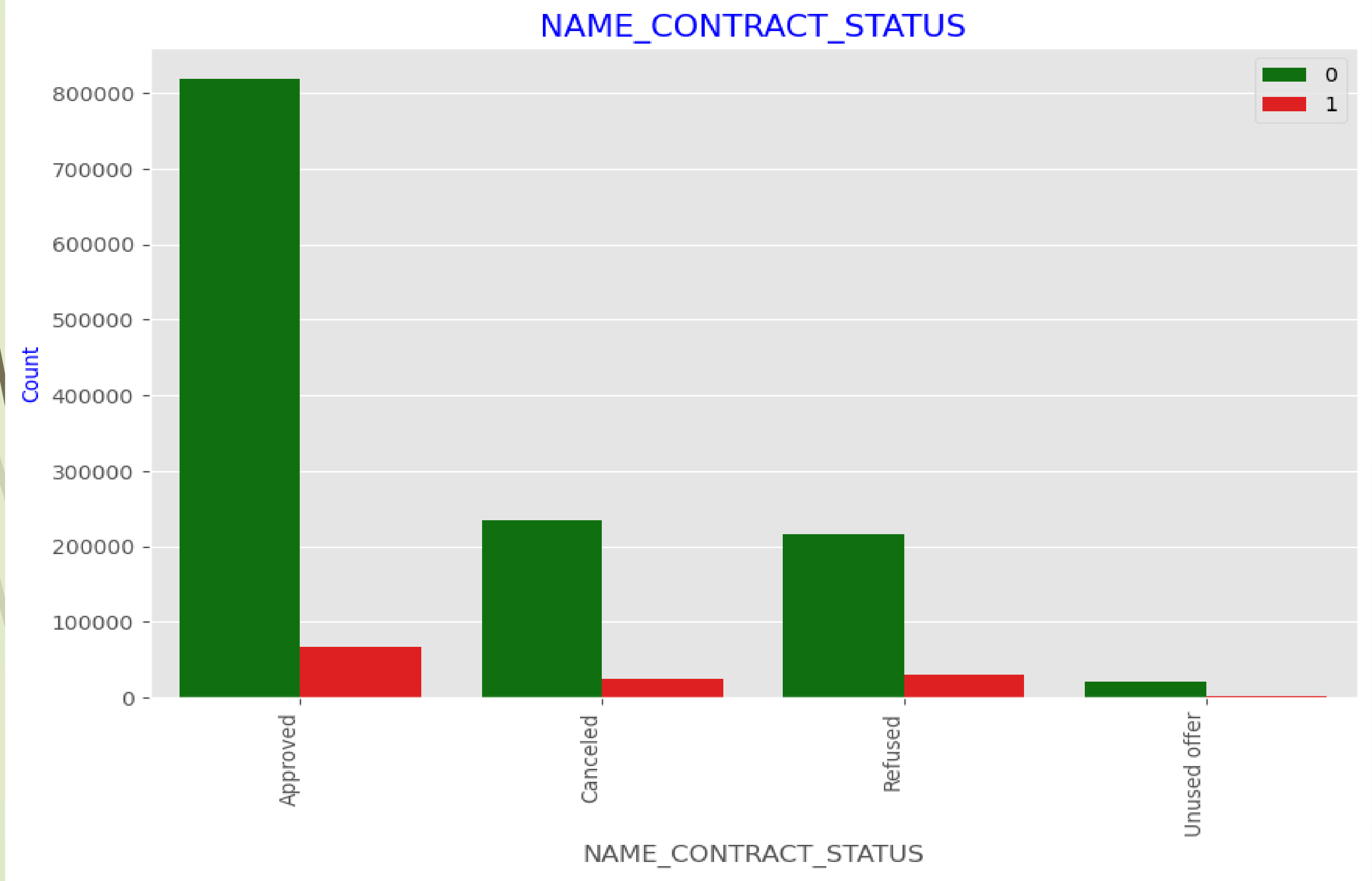


Contract Status vs purpose of the loan for Target 1



Merged Dataframes Analysis

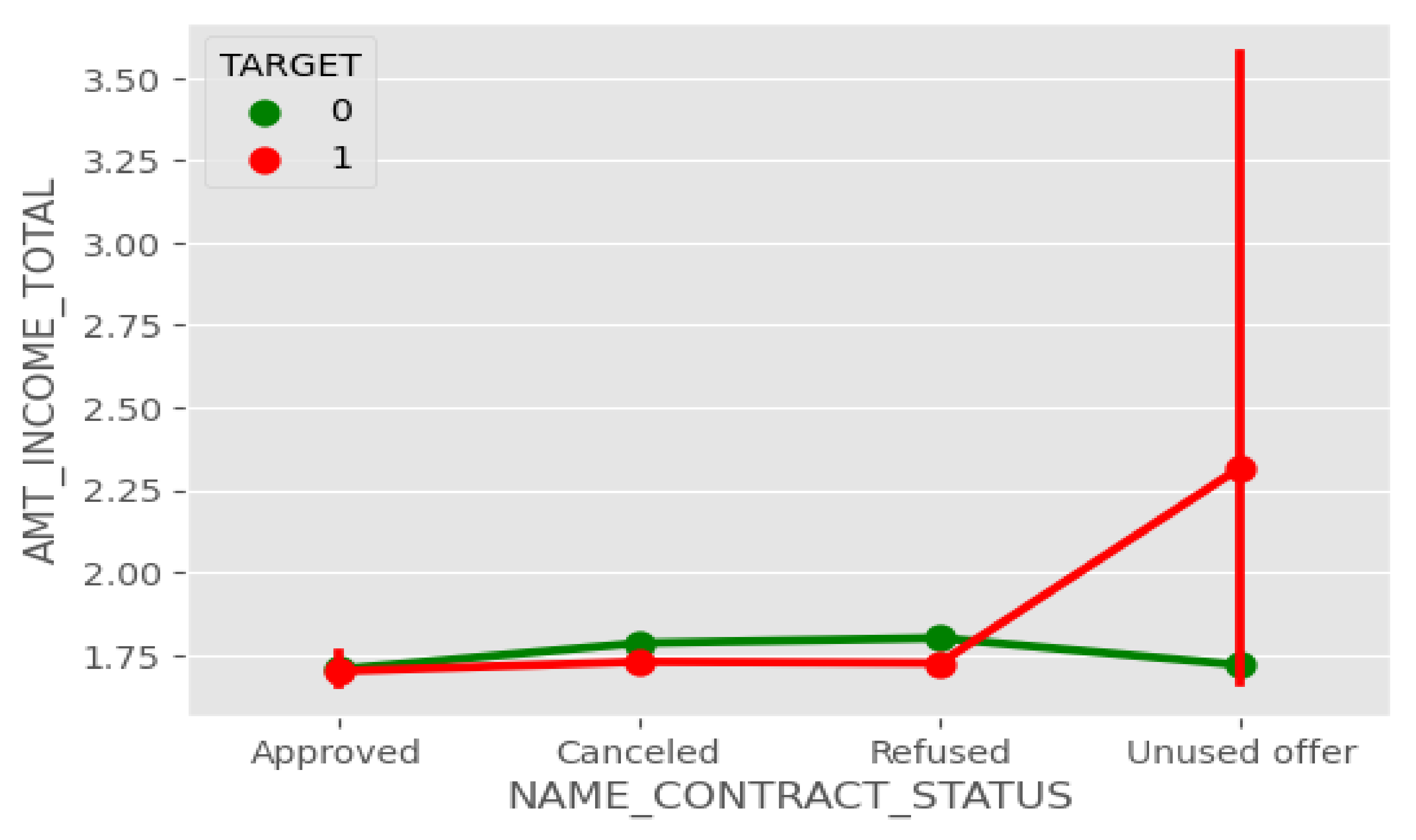
Contract Status based on loan repayment status



- 90% of the previously cancelled client have repayed the loan. And 88% of the clients who have been previously refused a loan have payed back the loan.
- Refual reason should be recorded for further analysis as these clients would turn into potential repaying customer.

Merged Dataframes Analysis

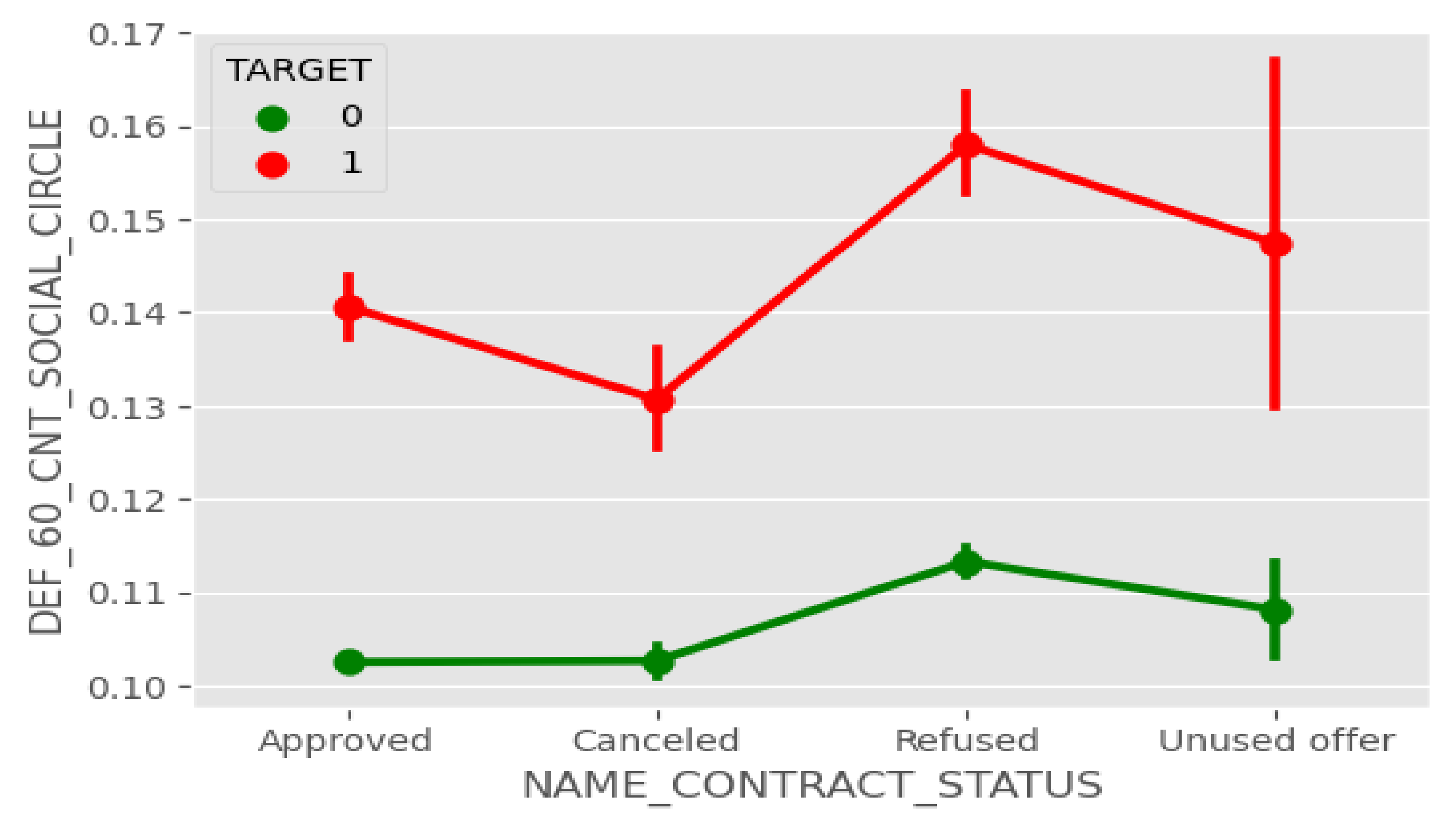
Relationship between Income and Contract status



- Applicants who have not used offer earlier have defaulted even when their average income is higher than others

Merged Dataframes Analysis

Relationship between people who defaulted in last 60 days being in client's social circle and Contract status



- Applicants 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.

Conclusion:

- There are 8.07% of defaulters.
- Larger amount of Revolving loans, in comparison with their frequency, are not repaid.
- There are more loan applications from female clients and they are less likely to default.
- People living in rented apartments and with parents have higher probability of defaulting.
- Single and not married clients are more likely to default.
- People with Lower secondary education, have the highest defaulter rate while the ones with Academic degree have the lowest defaulter rate.
- People above the age of 50 are more likely to repay the loan on time, While People in the age group range 20-40 are more likely to default.
- Applicants who have more than 4 children are more likely to default
- A lot of the previously cancelled and refused clients have turned into repaying clients. Record the reason for cancellation or rejection which might help the bank to determine and negotiate terms with these repaying customers in future for increase business opportunity.
- Applicants who have higher DEF_60_CNT_SOCIAL_CIRCLE score tend to default more