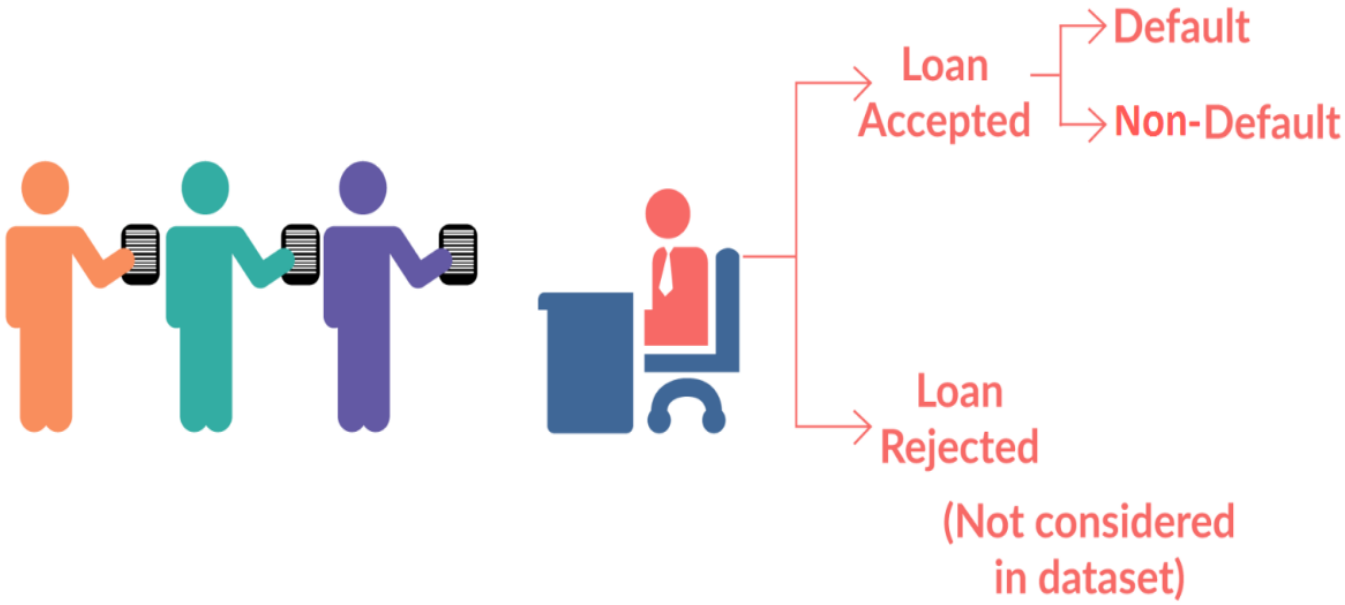
CREDIT EDA CASE STUDY



CONTRIBUTED BY:

Nikita Bhargava
Saurabh Gupta

Contents

- Data Preparation
- Reading The Data
- Data Cleaning
  - ✓ Fix Rows And Columns
  - ✓ Fix Missing Values
  - ✓ Handling Outliers
  - ✓ Filter Data
- Derived Metrics
- Univariate Analysis
- Segmented Analysis
- Bivariate Analysis
- Correlation
- Merging The Dataset
- Final Conclusions

CASE STUDY OBJECTIVE:

- 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.

INSPECTING THE DATASETS:

- We started with inspecting the data frame by checking its head ,tail ,shape , number of entries in the dataset.
- We did this for both the datasets in the similar manner.
- Then we tried to understand the relationship between both the given datasets and the columns of both the dataset.

FILTERING THE GIVEN DATA:

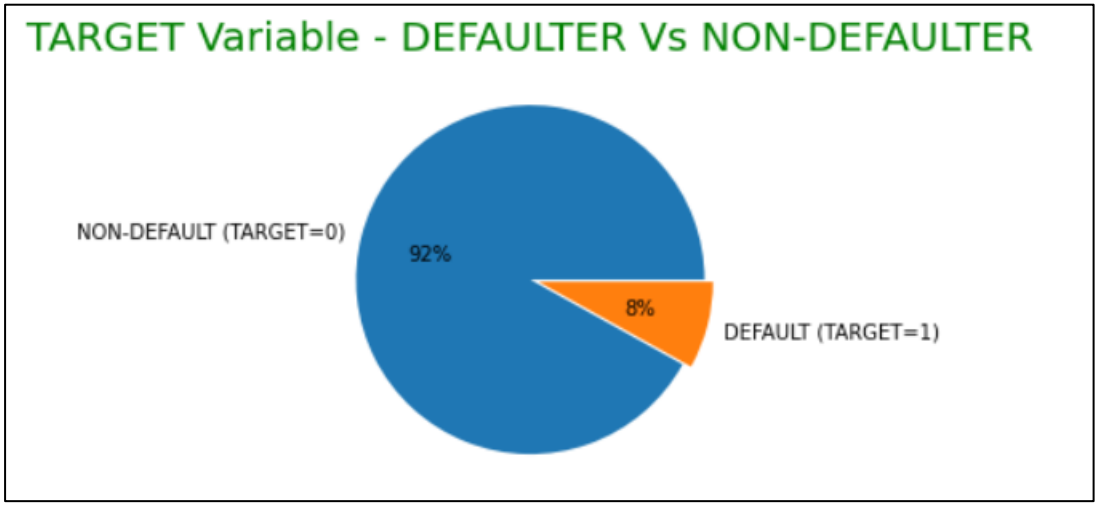
- We checked for all the null values present in the datasets.
- And based on the percentage of the null values , we drop all the columns with more than 40% of the missing data.
- For rest of the columns with missing values we imputed them with mean , median or mode.

DIVIDING THE DATASET INTO TWO HALVES:

- Based on whether the person is defaulter or not we have divided the dataset into two parts.
- In our dataset "TARGET" is the target variable and it has only two values.
- We have divided the data as follows:

1. TARGET=0 as non-defaulters.
2. TARGET=1 as defaulters.

DIVISION OF DATASET :



Here we can see that 92 % are non-defaulters and 8 % are defaulters.

DATA ANALYSIS FOR APPLICATION DATA:

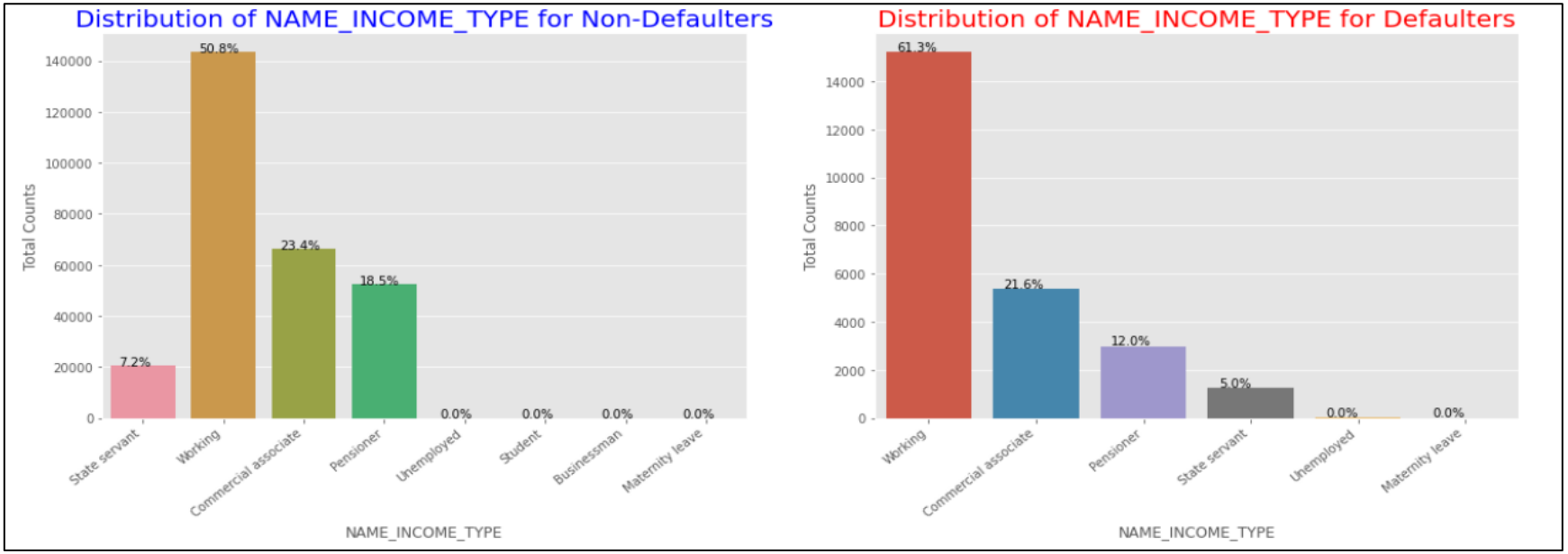
- This dataset contains personal information about customer like whether the client has own car, own house, own reality.
- Apart from this is also contains information related with the family members of the client, number of children , employment type of the client , education type etc.

UNIVARIATE ANALYSIS :

A variable in univariate analysis is just a condition or subset that your data falls into. You can think of it as a "category." For example, the analysis might look at a variable of "age" or it might look at "height" or "weight".

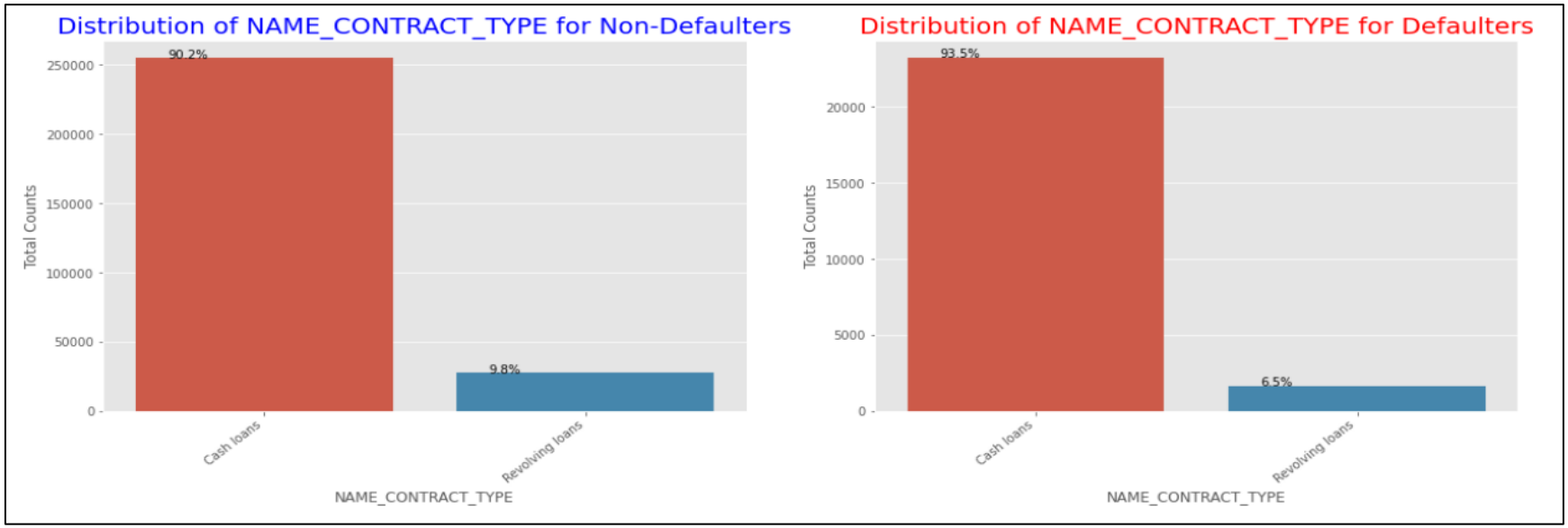
Univariate analysis is the simplest form of analyzing data. "Uni" means "one", so in other words your data has only one variable.

DISTRIBUTION FOR INCOME TYPE:



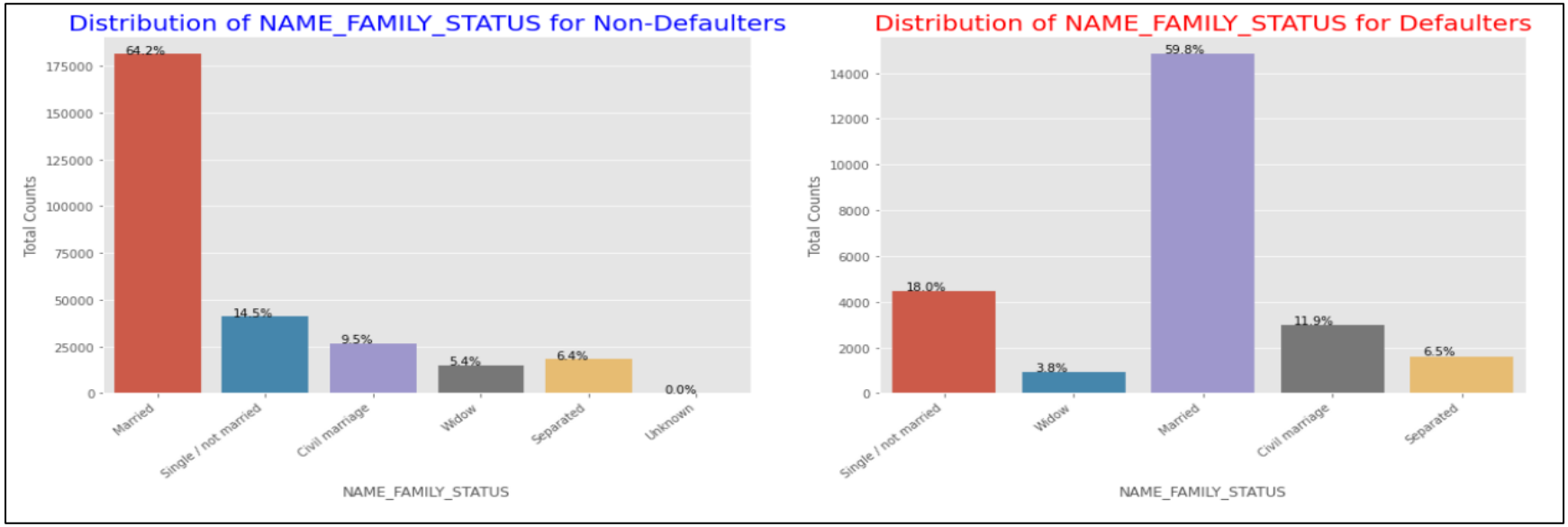
- It is clearly visible that students don't default. The reason for this might be that the students don't pay the loan while they are students.
- Also in both the cases working people are higher with almost 50% in the case of Non-defaulters and 61% in case of defaulters.

DISTRIBUTION FOR CONTRACT TYPE:



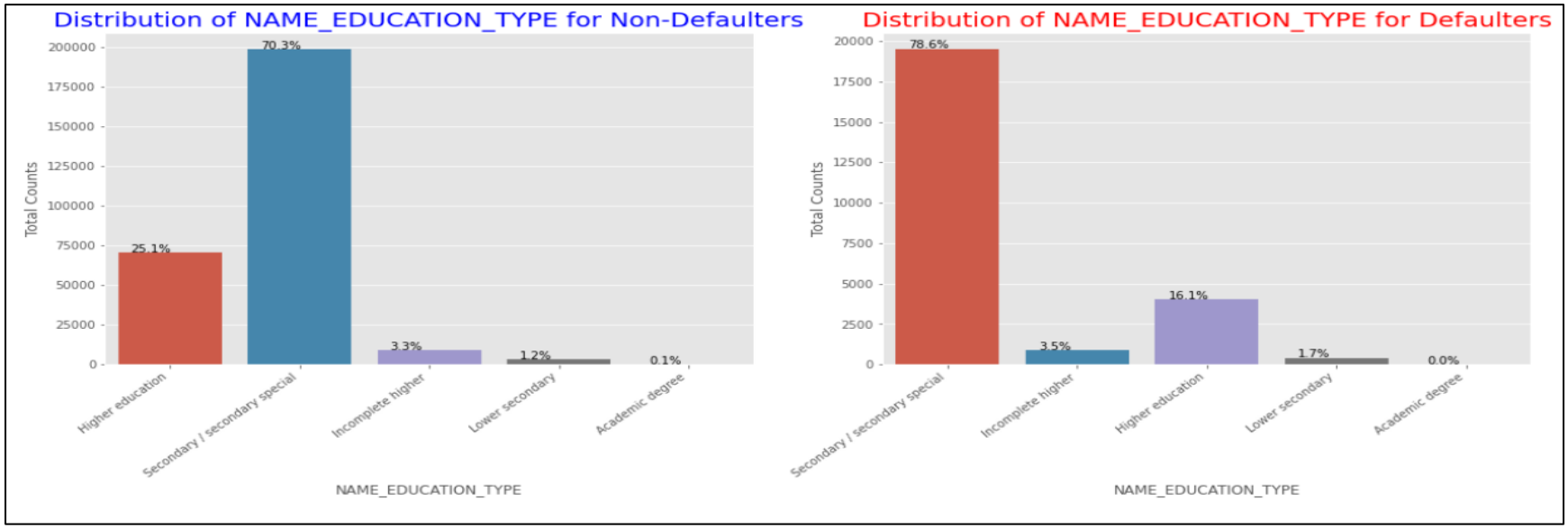
- Cash loans are very popular in the case of defaulters and non-defaulters.

DISTRIBUTION FOR FAMILY STATUS:



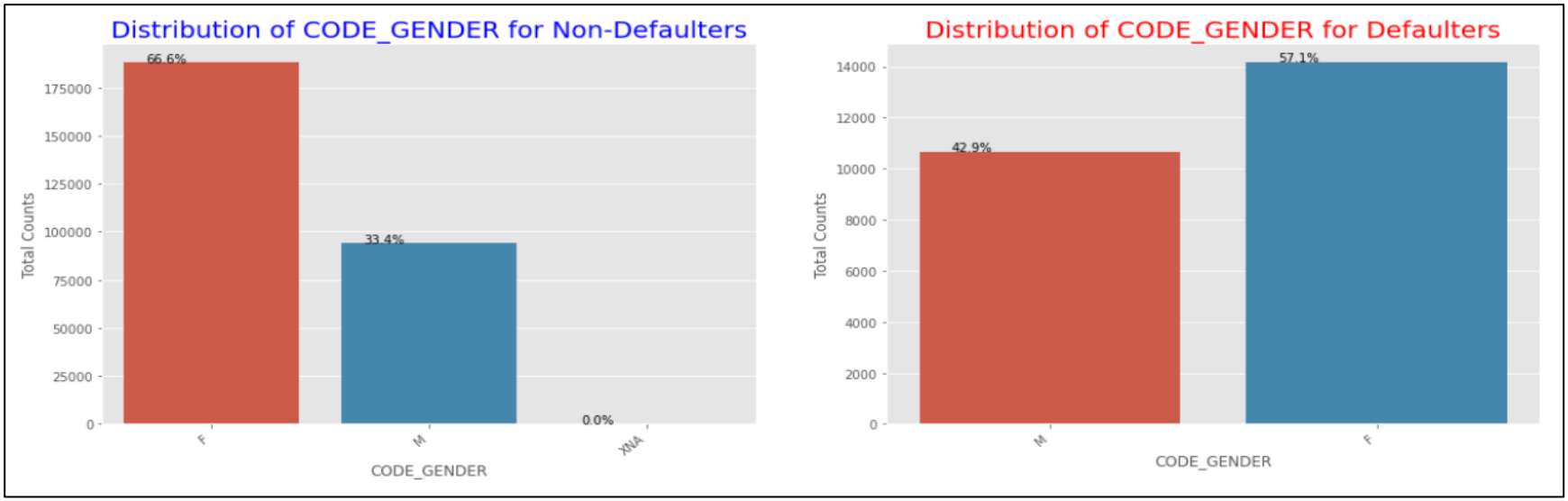
- Married people tend to take more loans and the reason for this is they have two helping hands and so taking loan become easy for them.
- But married people are ahead of everyone in case of defaulters.
- Separated people take less loan and the reason for this might be that they already have problems in their life and so don't want to spend much money.

DISTRIBUTION FOR EDUCATION TYPE:



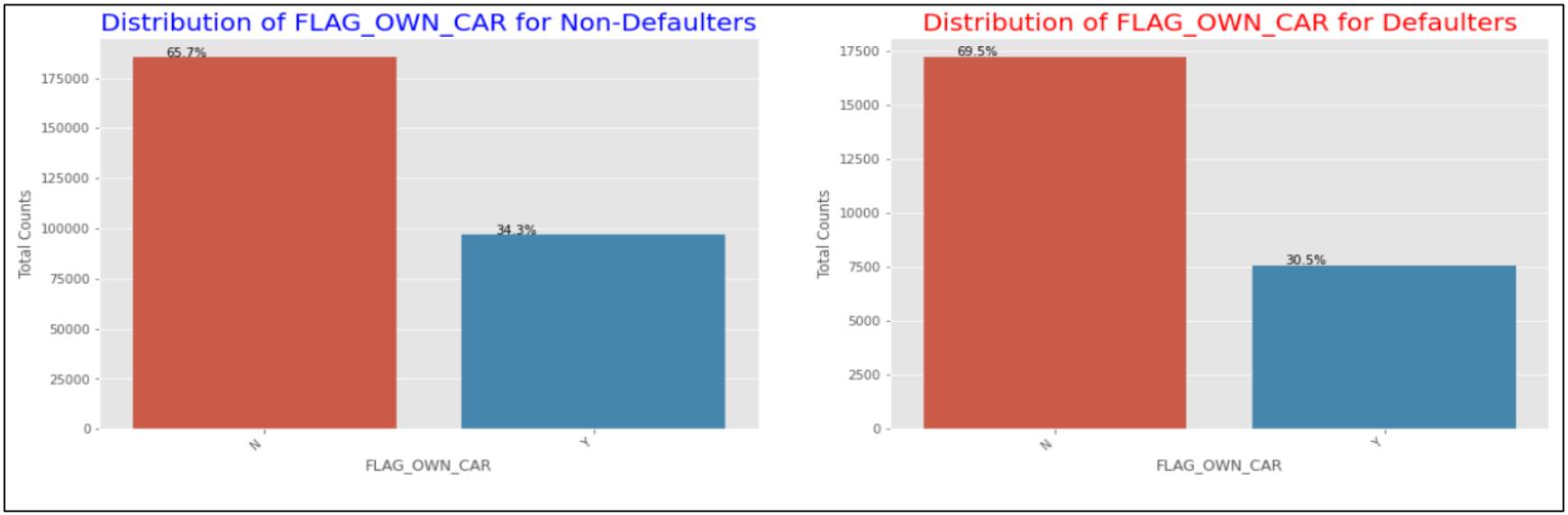
- In case of defaulters and non-defaulters , loans for secondary or secondary special educations are taken much because they are pretty much higher as compared to other type of education and many people preferred to take secondary education.
- But in case of defaulters , loans taken for secondary education suffers huge loss, due to many people not able to repay the loan.

DISTRIBUTION FOR GENDER:



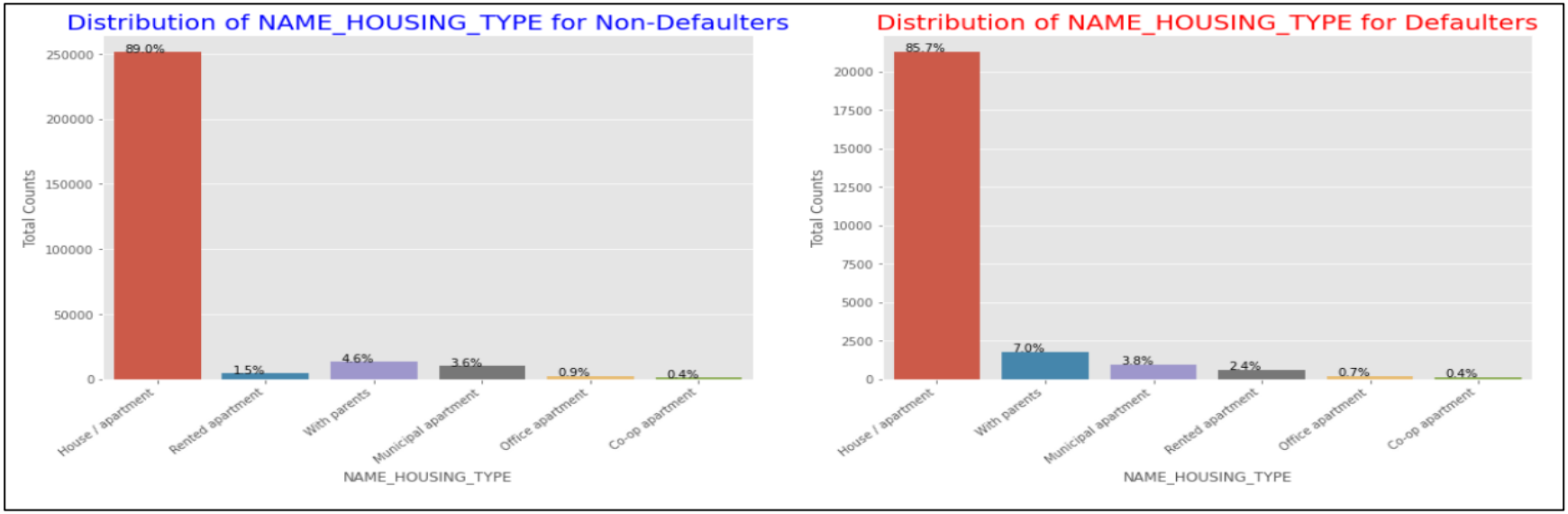
- We can see that Female contribute almost 67% to the non-defaulters while 57% to the defaulters.
- We can conclude that more female applying for loans than males and hence the more number of female defaulters as well.
- But the rate of defaulting of female is much lower compared to their male counterparts.

DISTRIBUTION FOR OWN CAR COLUMN:



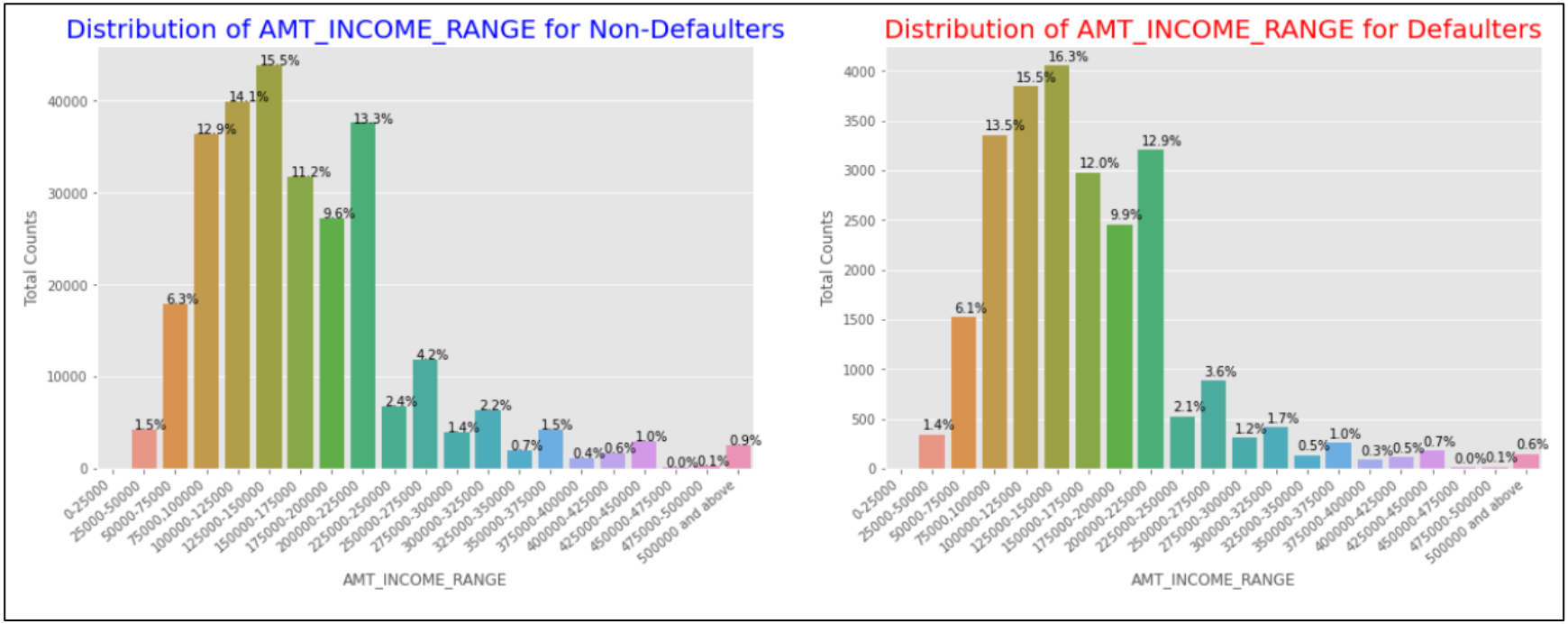
- We can see that people with cars contribute 65.7% to the non-defaulters while 69.5% to the defaulters.
- We can conclude that while people who have car default more often, the reason could be there are simply more people without cars looking at the percentages in both the charts, we can conclude that the rate of default of people having car is low compared to people who don't.

DISTRIBUTION FOR HOUSING TYPE:



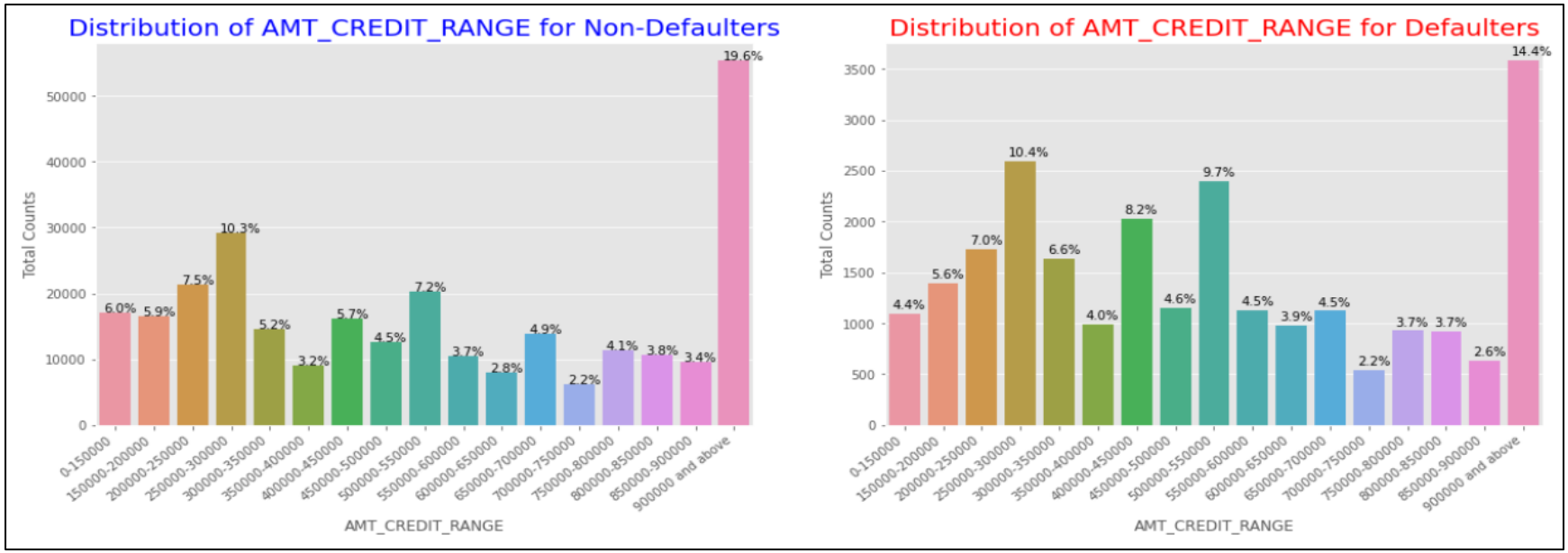
- It is clear from the graph that people who have House/ Apartment , tend to apply for more loans.
- People living with parents tend to default more often when compared with others . The reason could be their living expenses are more due to their parents living with them.

DISTRIBUTION FOR INCOME RANGE:



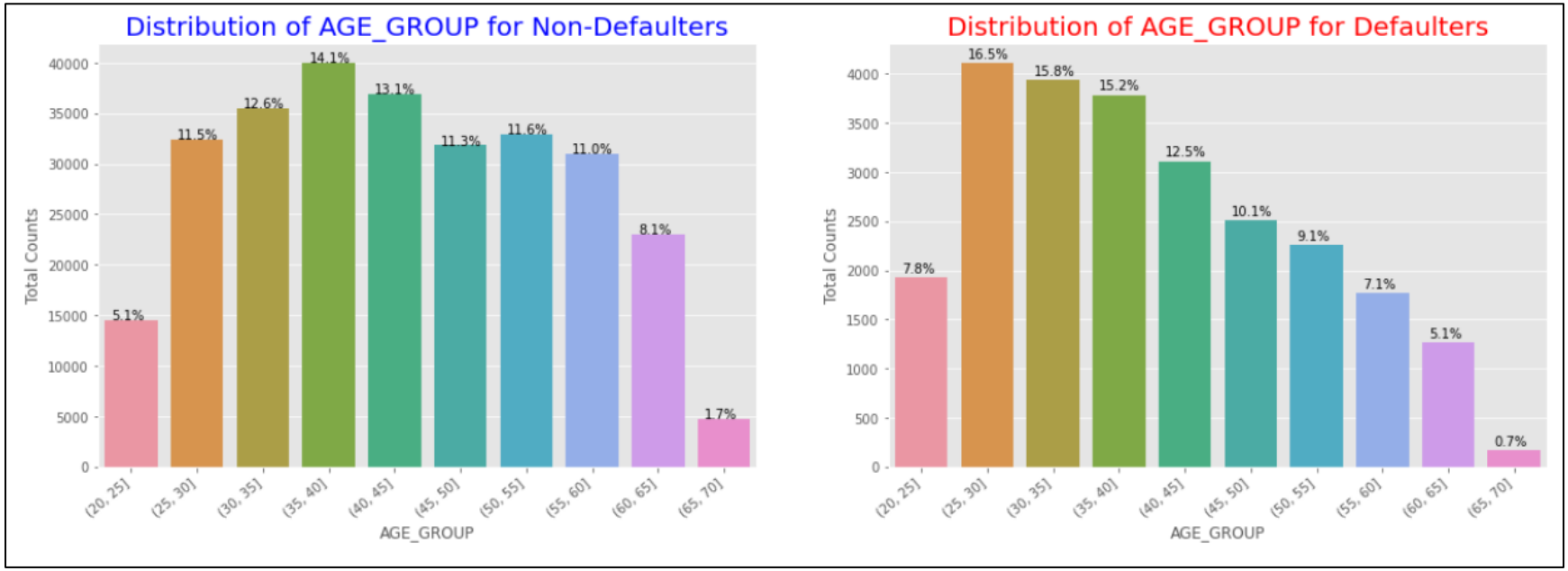
- People with 125000-200000 salary take much loan.
- Most of the defaulters have 125000-150000 as their salary.

DISTRIBUTION FOR CREDIT RANGE:



- For defaulters - maximum credit amount is 14.4% for 9 Lakhs and above.
- For non-defaulters - maximum credit amount is 19.6% for 9 Lakhs and above.

DISTRIBUTION FOR AGE GROUP:



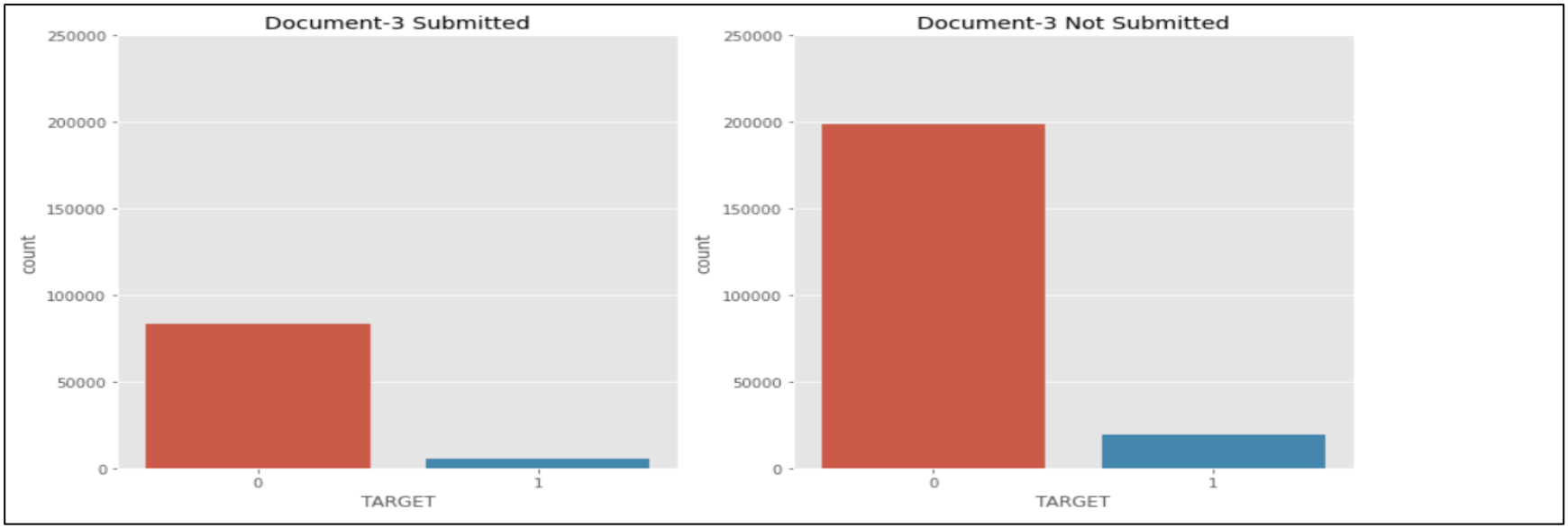
- People of age 25 to 30 have higher default rate.
- Default cases are less for applicants more than 40 years old.

BIVARIATE ANALYSIS:

Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as *X*, *Y*), for the purpose of determining the empirical relationship between them.

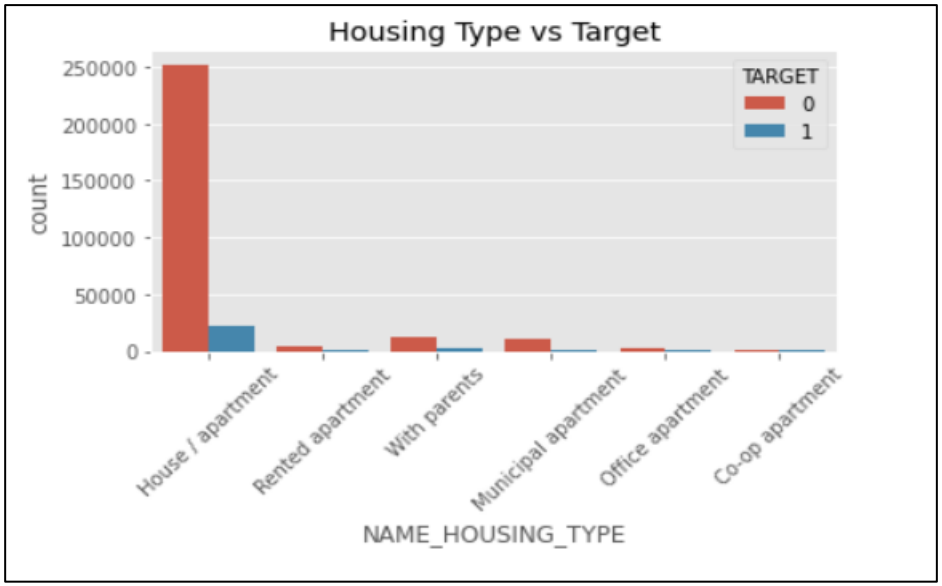
Bivariate analysis can be contrasted with univariate analysis in which only one variable is analysed. Like univariate analysis, bivariate analysis can be descriptive or inferential. It is the analysis of the relationship between the two variables

DISTRIBUTION FOR THOSE WHO HAVE SUBMITTED DOCUMENT-3 AND THOSE WHO DIDN'T:



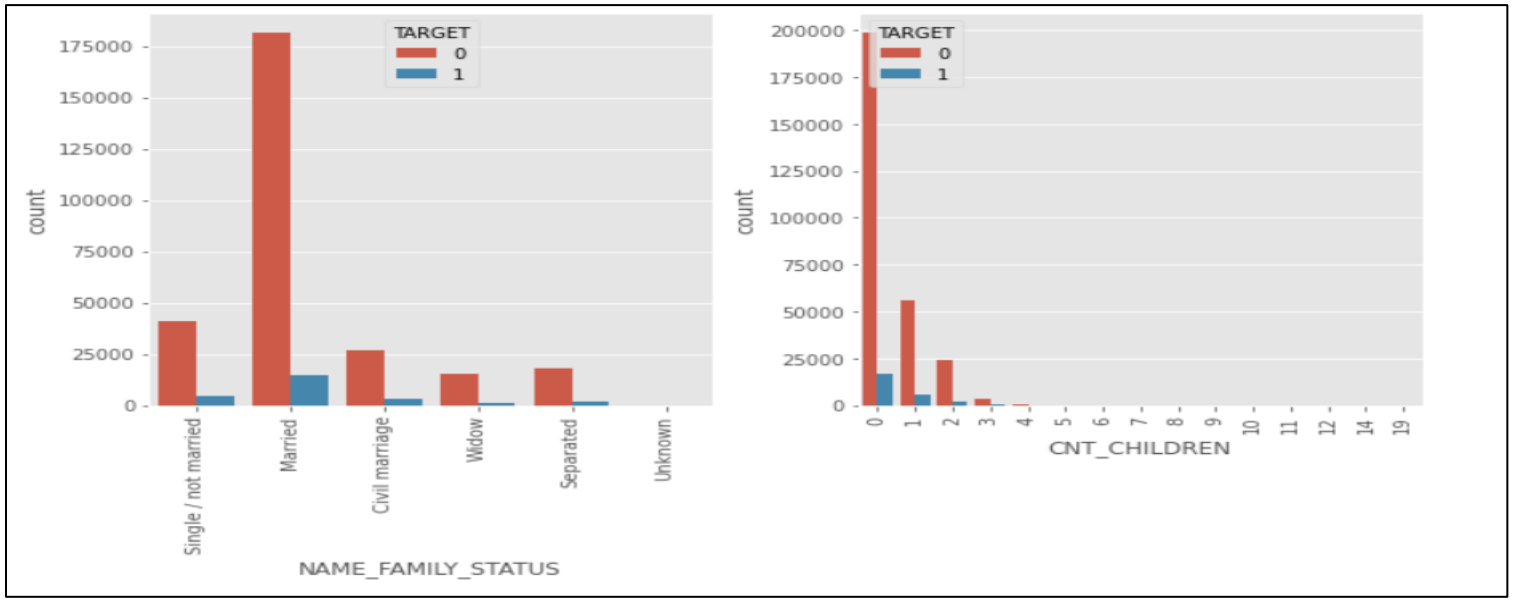
- FLAG_DOCUMENT_3 is showing similar trend for both non-defaulters and defaulters.
- Hence, this column can be dropped also.

DISTRIBUTION FOR HOUSING TYPE:



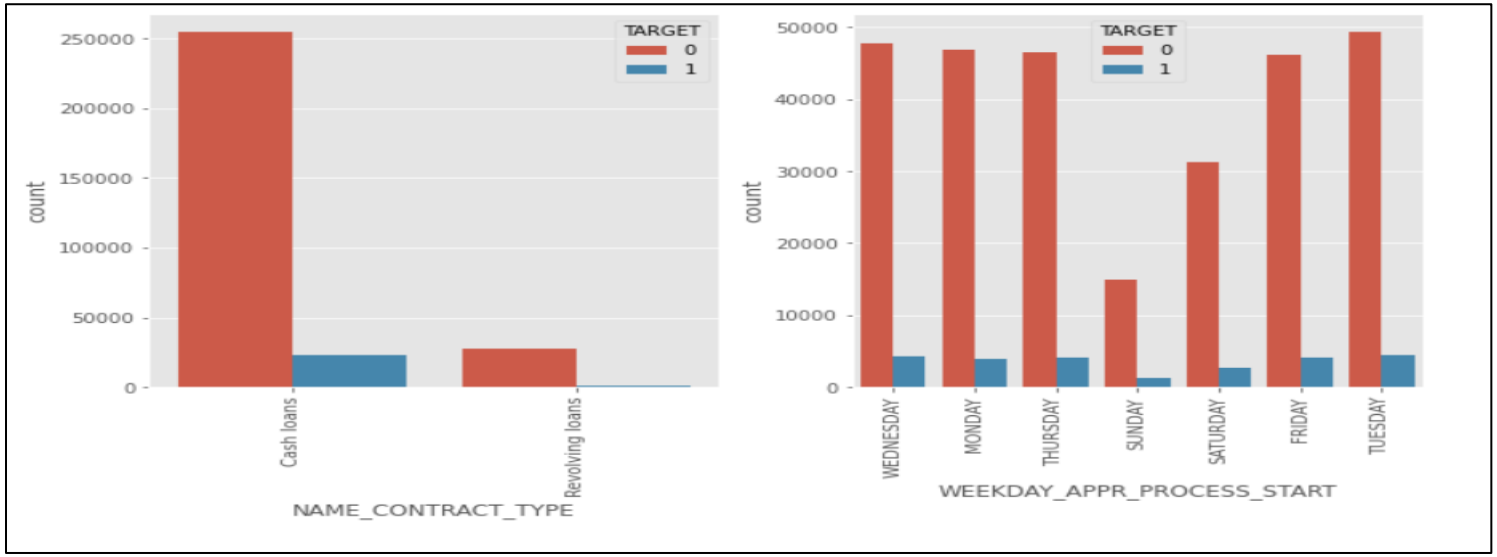
- So here we can see most of the applicants live in House / Apartment for both defaulters and non-defaulters.

DISTRIBUTION FOR FAMILY STATUS AND NUMBER OF CHILDREN:



- Mostly married people take loans and the default rate is also maximum for married the reason for this might be that married people have income for two people.
- Generally loans are taken after marriage and before having kids, because after having kids client has more responsibilities.

DISTRIBUTION FOR CONTRACT TYPE AND WEEK DAY AT WHICH THE LOAN APPLICATION STARTED:

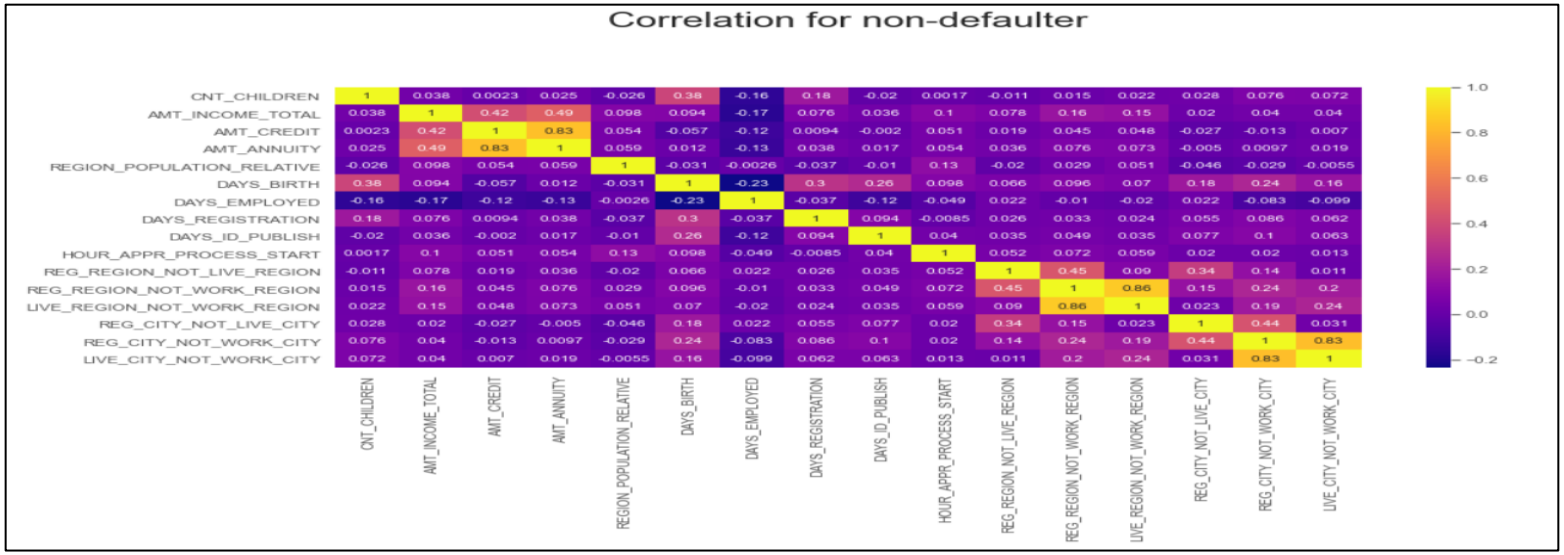


- Number of Cash Loans is quite higher than Revolving Loans.
- All weekdays have similar number of applicants except weekend(Saturday and Sunday).

DATA ANALYSIS FOR PREVIOUS DATA:

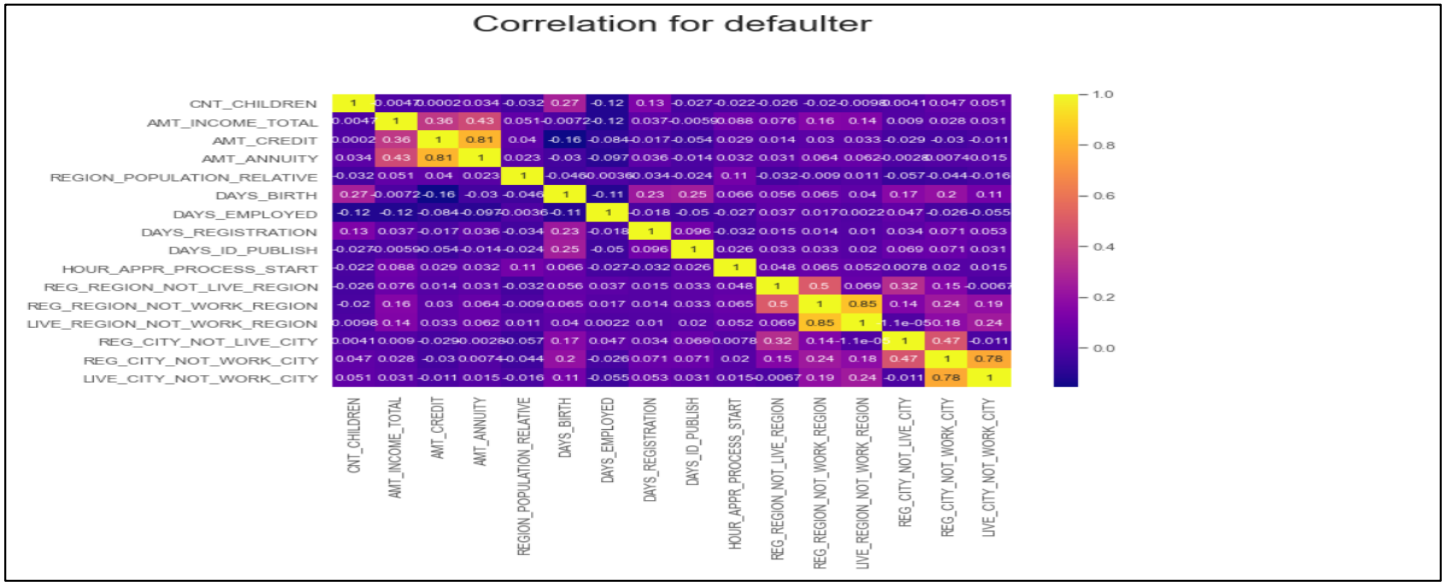
- This dataset contains columns like amount credit, amount annuity, down payment, contract type, payment type, time taken to approve the loan, day of week for which loan application was sent.

CORRELATION FOR NON-DEFAULTERS:



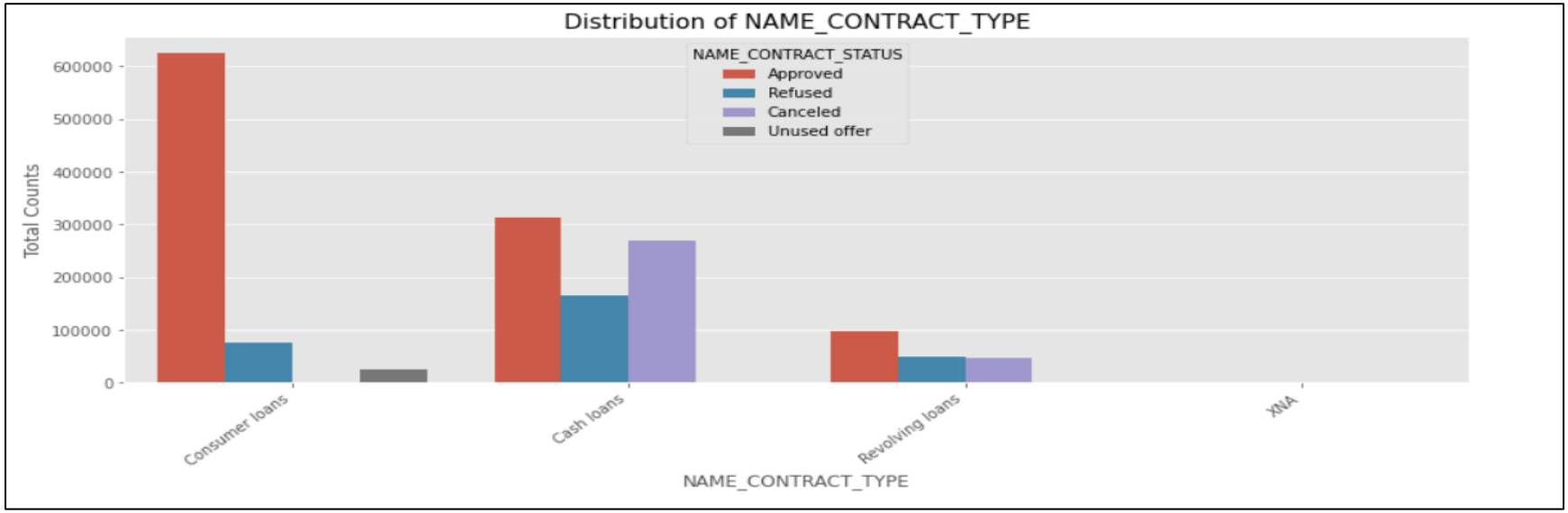
- Credit amount is inversely proportional to the date of birth, which means Credit amount is higher for low age and vice-versa.
- Credit amount is inversely proportional to the number of children client have, means Credit amount is higher for less children count client have and vice-versa.
- Income amount is inversely proportional to the number of children client have, means more income for less children client have and vice-versa.

CORRELATION FOR DEFAULTER:



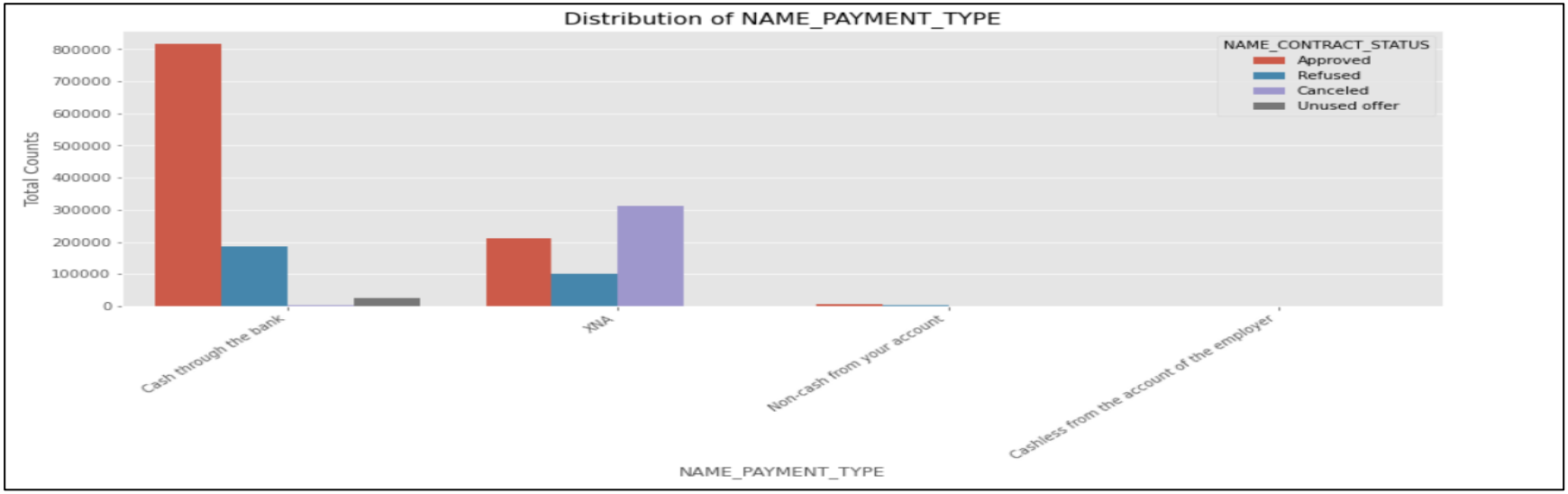
- People with permanent address not matching contact address are having less children (inverse proportional).
- People with permanent address not matching work address are having less children (inverse proportional).

DISTRIBUTION FOR CONTRACT TYPE:



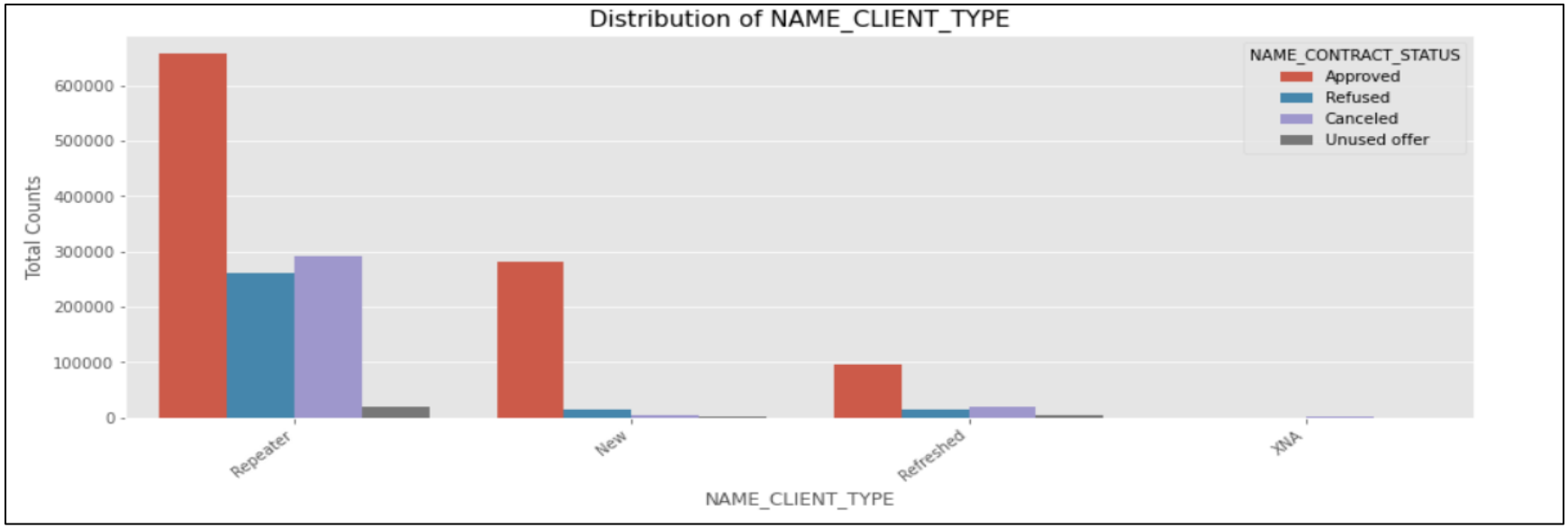
- Consumer and cash loans contract type are very popular among all the clients.
- Cash loans are refused the most , the reason for this might be the client was not able to pay the cash on time.

DISTRIBUTION FOR PAYMENT TYPE:



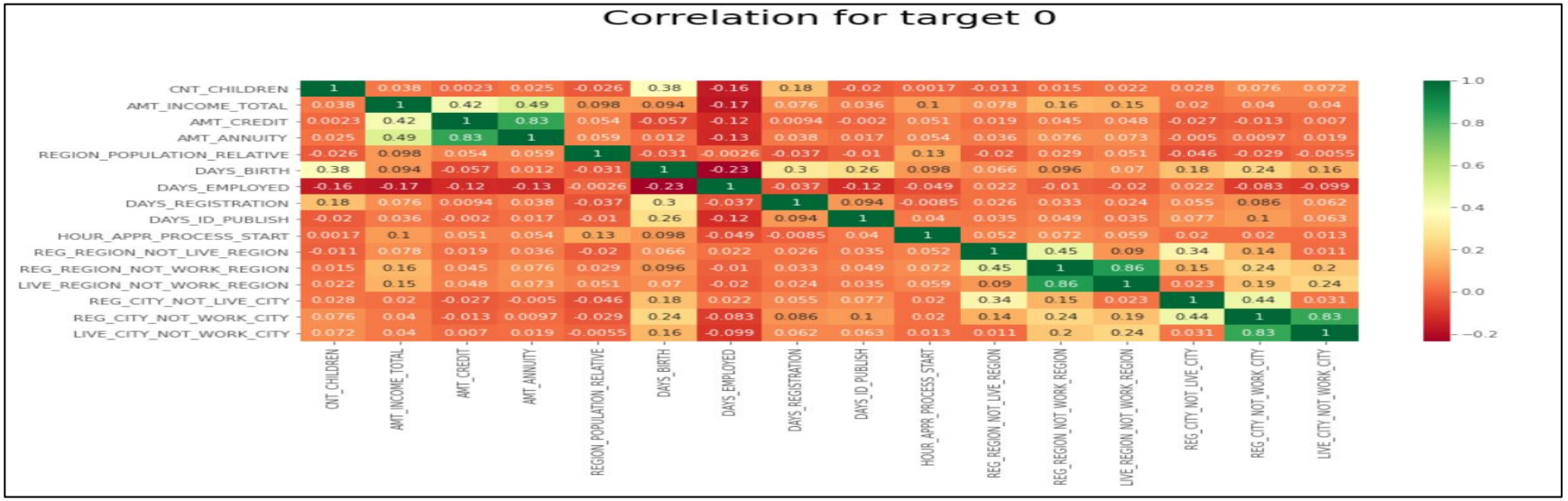
- From the above chart, we can infer that most of the clients choose to repay the loan using the 'Cash through the bank' option .
- We can also see that 'Non-Cash from your account' & 'Cashless from the account of the employee' options are not at all popular in terms of loan repayment amongst the customers.

DISTRIBUTION FOR CLIENT TYPE:



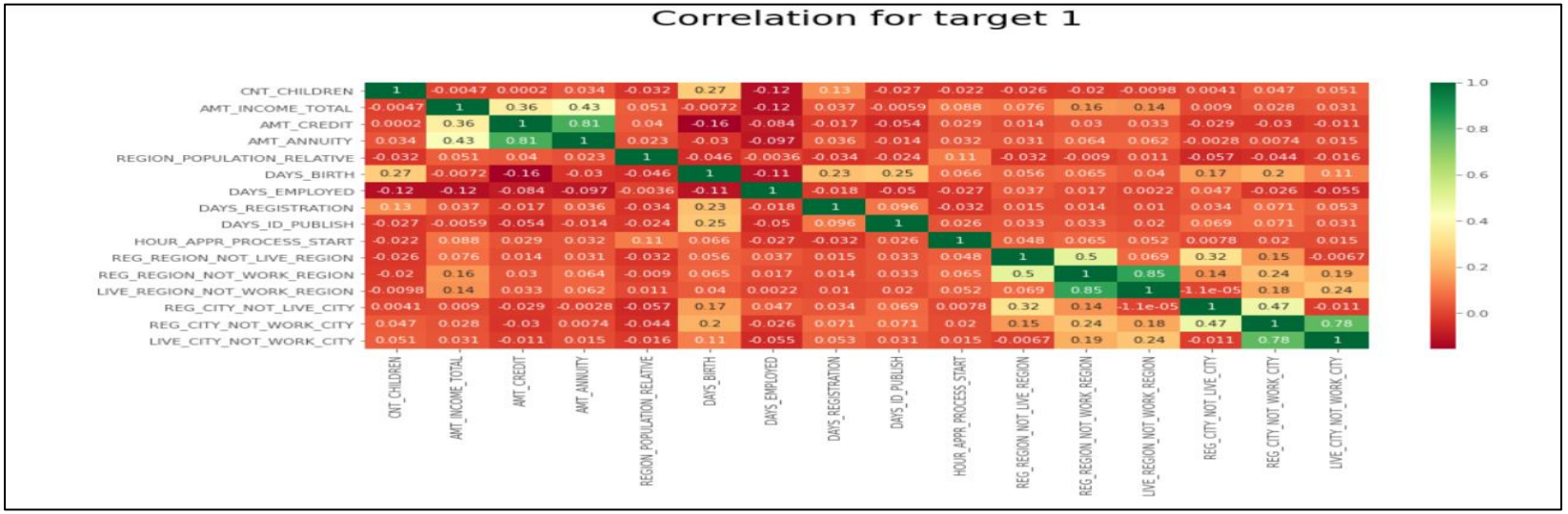
- Most of the loan applications are from repeat customers, out of the total applications 70% of customers are repeaters.
- Also repeaters get refused most often.
- Cancelled loans are also more for repeaters than any other type of client.

CORRELATION FOR NON-DEFAULTERS:



- Credit amount is inversely proportional to the date of birth, which means Credit amount is higher for low age and vice-versa.
- Credit amount is inversely proportional to the number of children client have, means Credit amount is higher for less children count client have and vice-versa.
- Income amount is inversely proportional to the number of children client have, means more income for less children client have and vice-versa.

CORRELATION FOR DEFAULTERS:

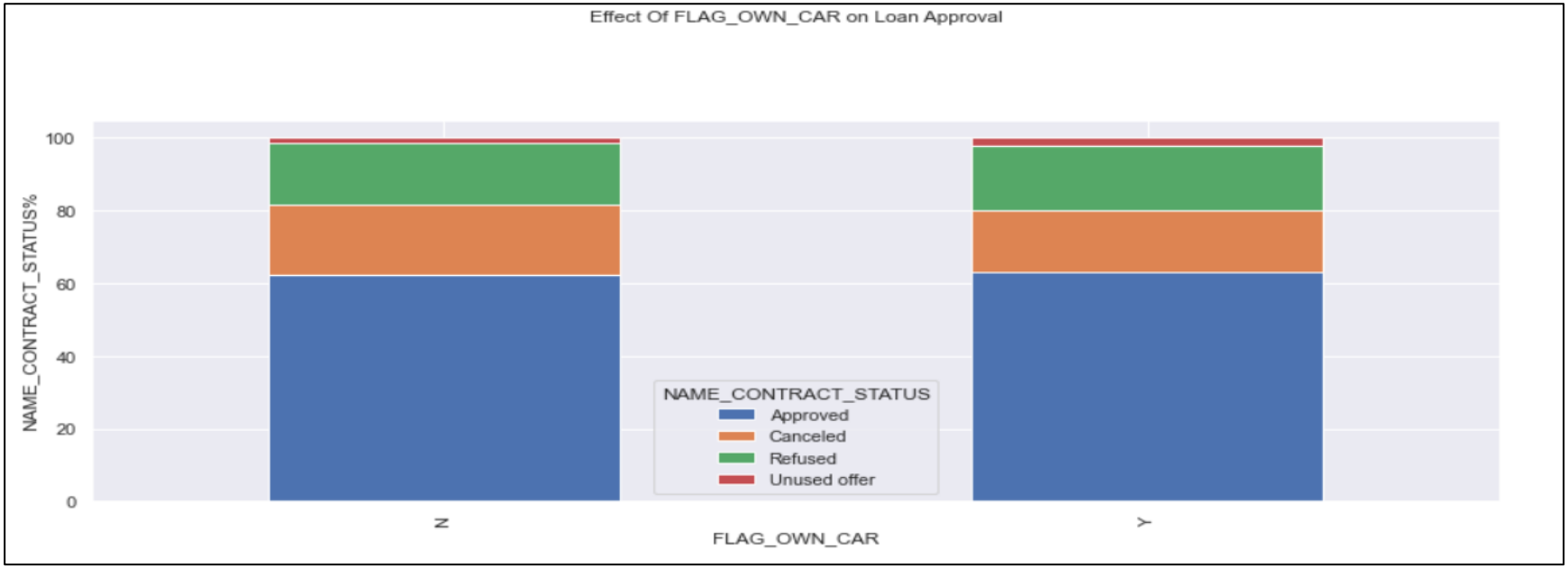


- People with permanent address not matching contact address are having less children (inverse proportional).
- People with permanent address not matching work address are having less children (inverse proportional).

NOW WE HAVE COMBINED BOTH THE DATASETS TO SEE IMPORTANT INSIGHTS

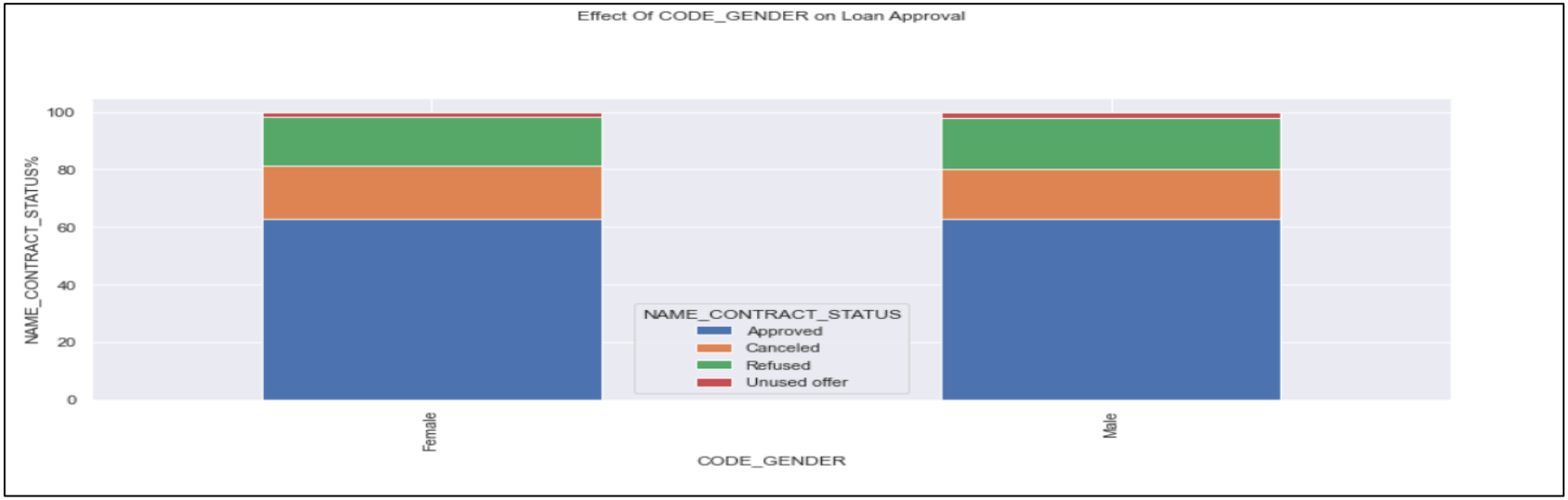
- Application dataset
- Previous dataset

DISTRIBUTION FOR OWN CAR AND CONTRACT TYPE:



- We see that car ownership doesn't have any effect on application approval or rejection.
- But we saw earlier that the people who has a car has lesser chances of default.
- The bank can add more weightage to car ownership while approving a loan amount.

DISTRIBUTION FOR GENDER AND CONTRACT TYPE:



- We see that code gender doesn't have any effect on application approval or rejection.
- But we saw earlier that female have lesser chances of default compared to males.
- The bank can add more weightage to female while approving a loan amount.

FINAL CONCLUSION:

- Focus should be on people with income type "Student", "Pensioner", and "Businessman".
- Housing type should be other than "Co-op apartment".
- People with housing type "With parents" can be targeted as they are having least number of unsuccessful payments.

- Banks should focus less on income type as "Working" as they have the greatest number of unsuccessful payments.
- Also, with loan purpose "Repair" is having higher number of unsuccessful payments on time and so less number of loans should be provided to them for better profits.

THANK YOU!