CREDIT EDA CASE STUDY

-BY HITIK NEGI & ARUN TYAGI

PROBLEM STATEMENT

- 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 as their advantage by becoming a defaulter.

- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

  1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

  2. 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.

OBJECTIVE

- Aim is 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.

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

APPROACH

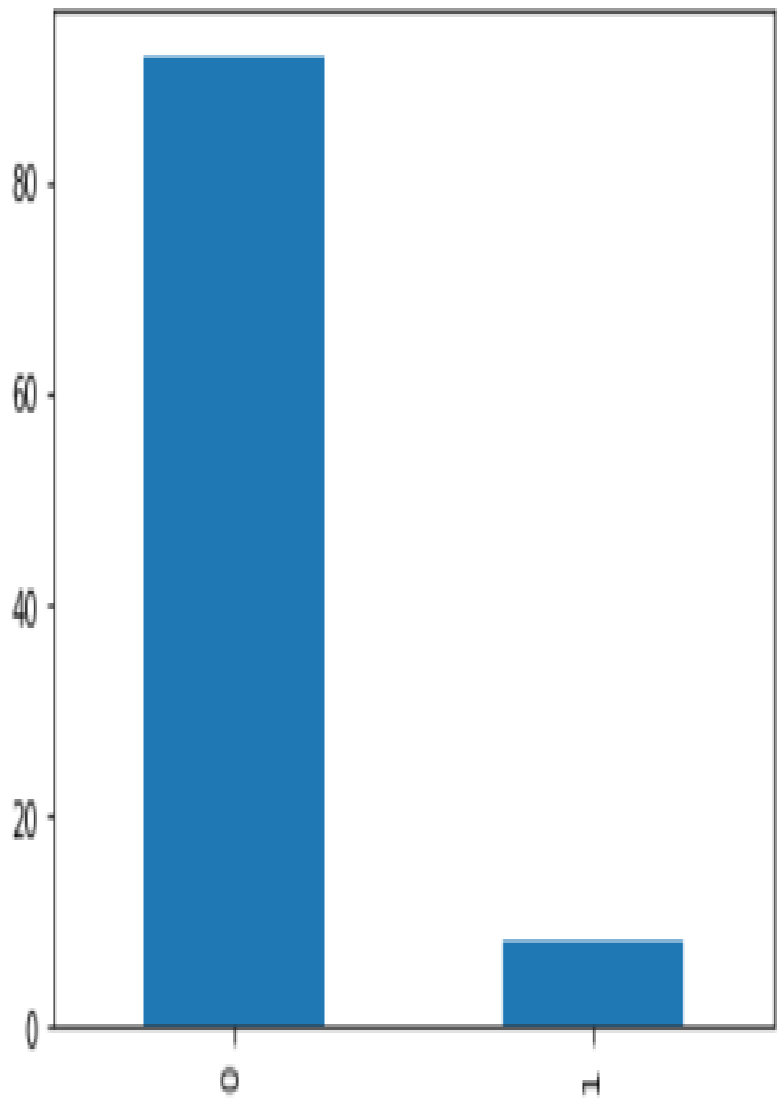
Following is the approach adopted while performing the exploratory data analysis on the given bank datasets:

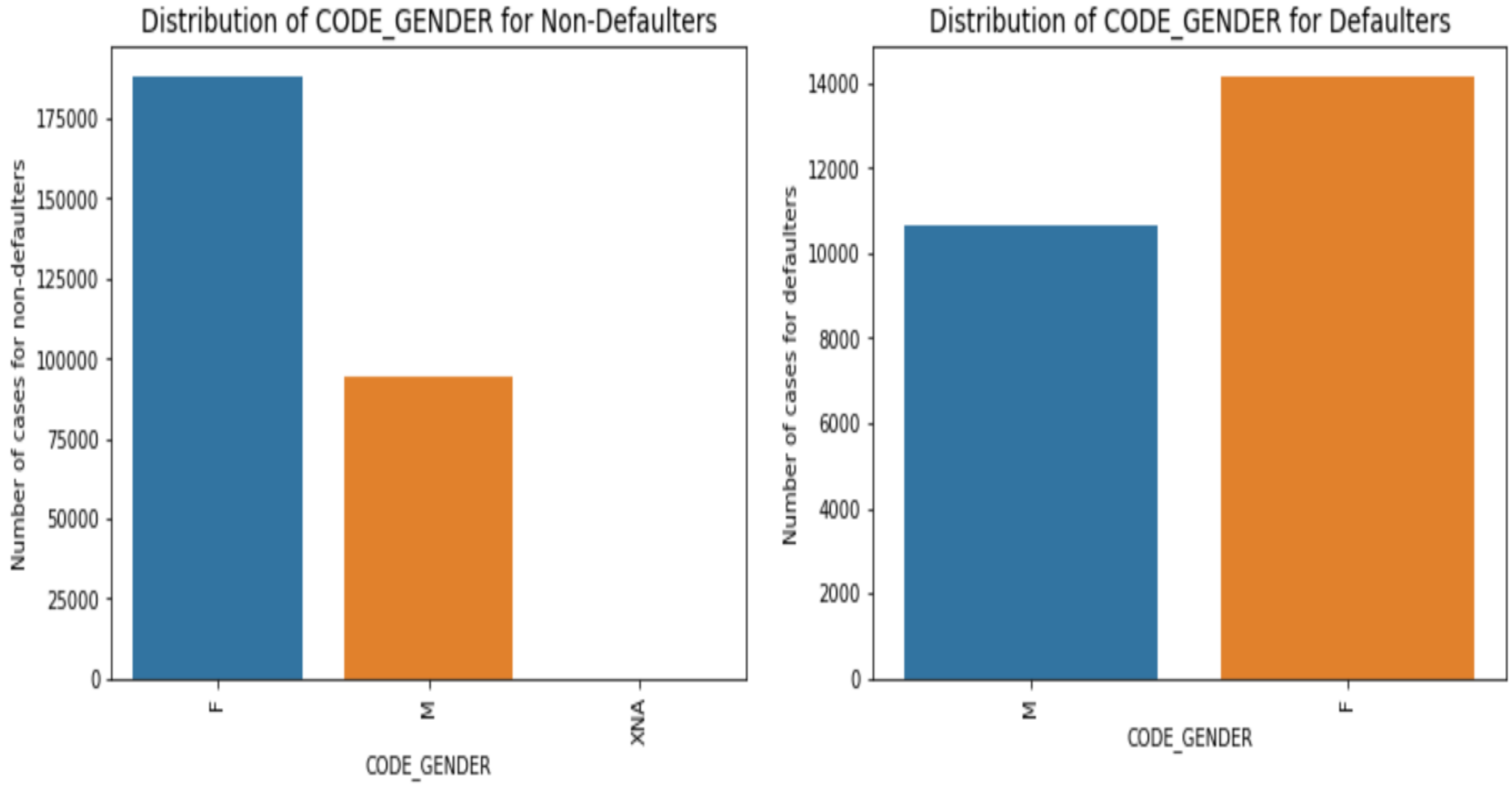
- Importing the libraries and reading the file
- Data Understanding
- Data Cleaning
- Univariate Analysis
- Bivariate Analysis
- Finding the Correlation between columns

TARGET VARIABLE

- Target variable has segmented the data into 2 types - 1 which represents people with payment difficulties (defaulters) and 0 which represents all other cases (non-defaulters).

- As per the given data, there are approximately 92% of non defaulters and 8% of defaulters which reflects the data is imbalance.

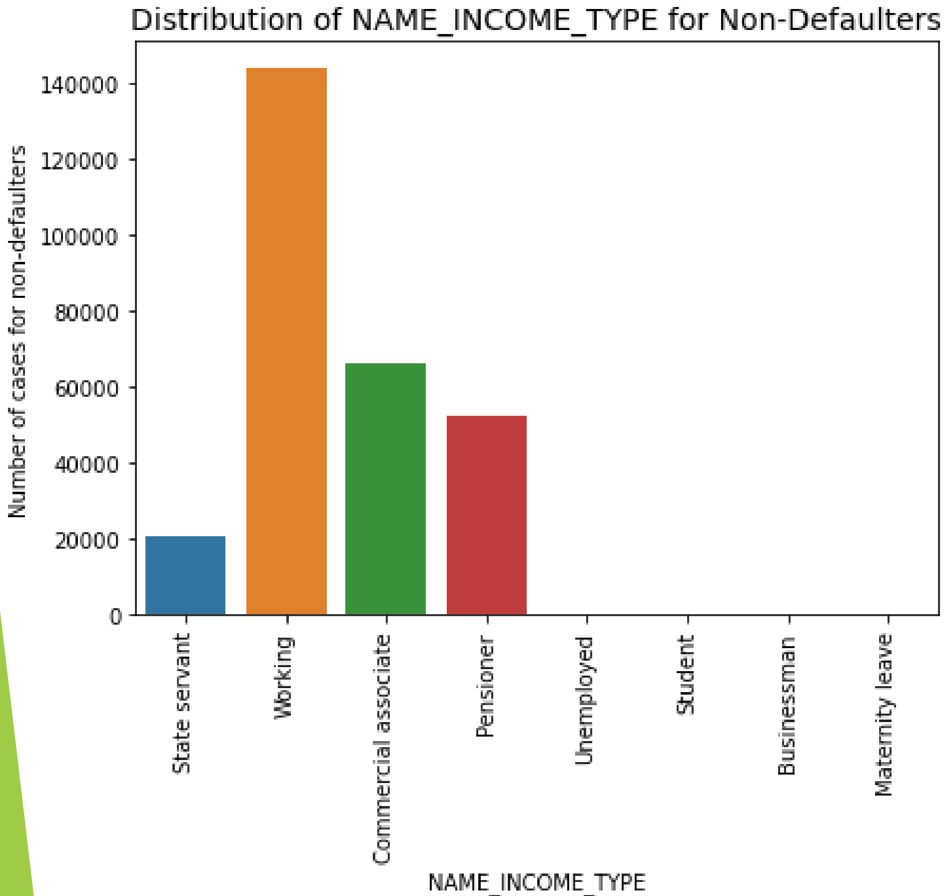


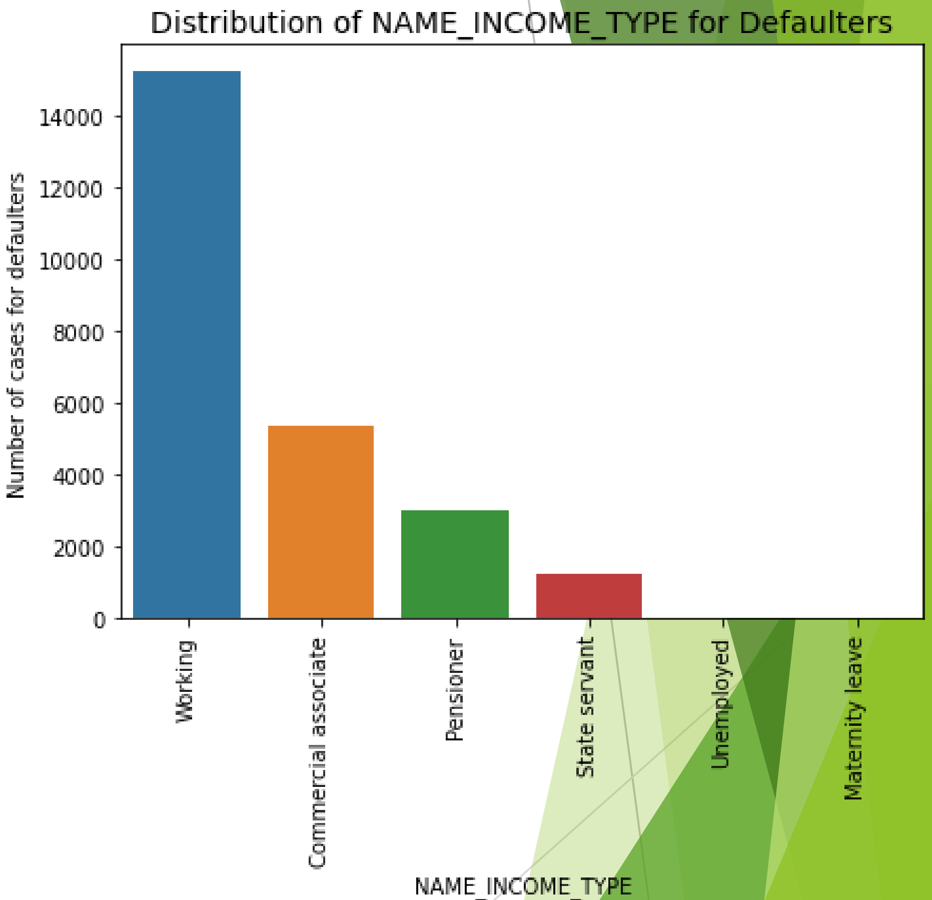


CATEGORICAL ANALYSIS : GENDER

CATEGORICAL ANALYSIS : GENDER

- Females tend to take more loan than Males in general.
- Number of female defaulters are higher than Males.

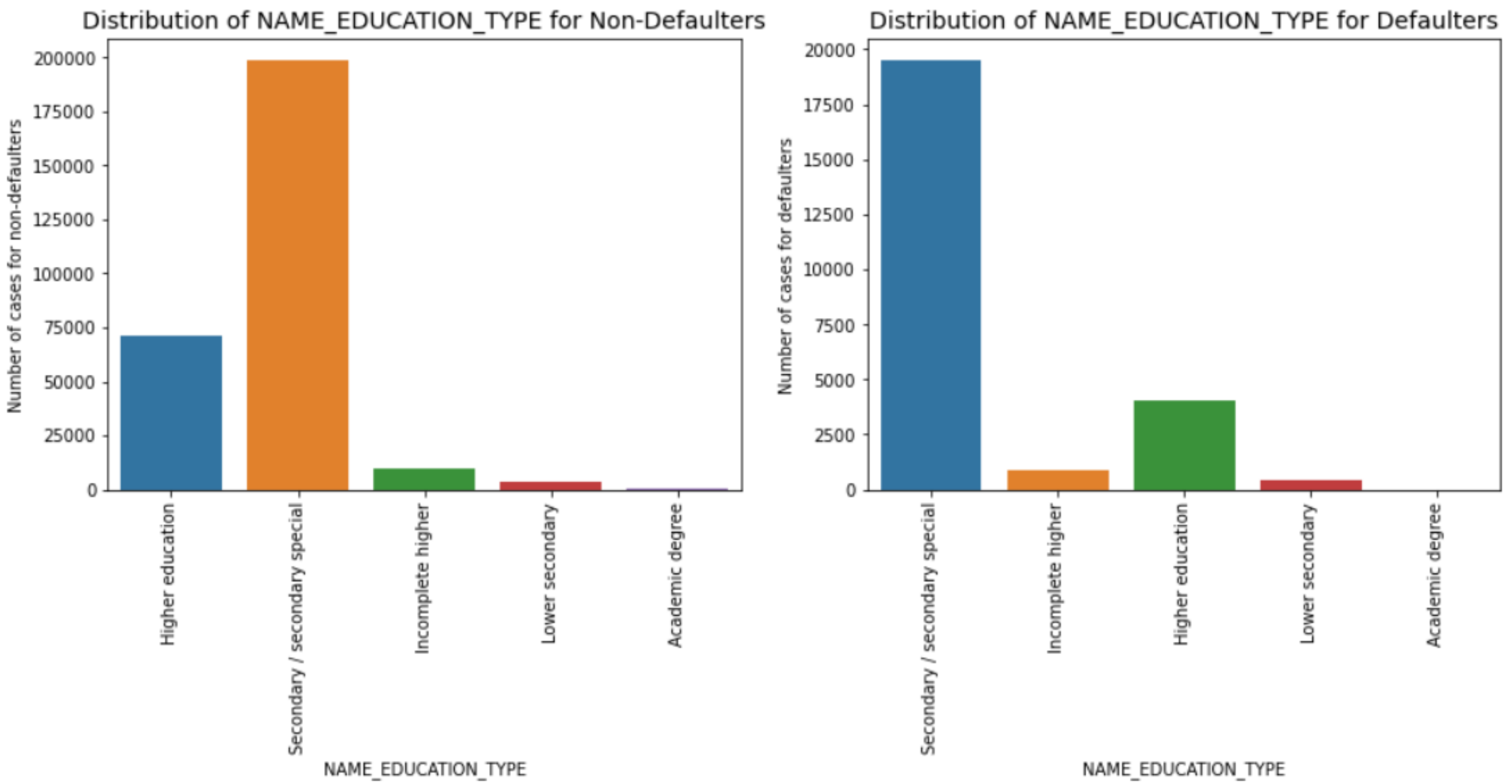

Distribution of NAME_INCOME_TYPE for Non-Defaulters
Number of cases for non-defaulters
0
20000
40000
60000
80000
100000
120000
140000
State servant
Working
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
NAME_INCOME_TYPE


Distribution of NAME_INCOME_TYPE for Defaulters
Number of cases for defaulters
0
2000
4000
6000
8000
10000
12000
14000
Working
Commercial associate
Pensioner
State servant
Unemployed
Maternity leave
NAME_INCOME_TYPE

INCOME TYPE

INCOME TYPE

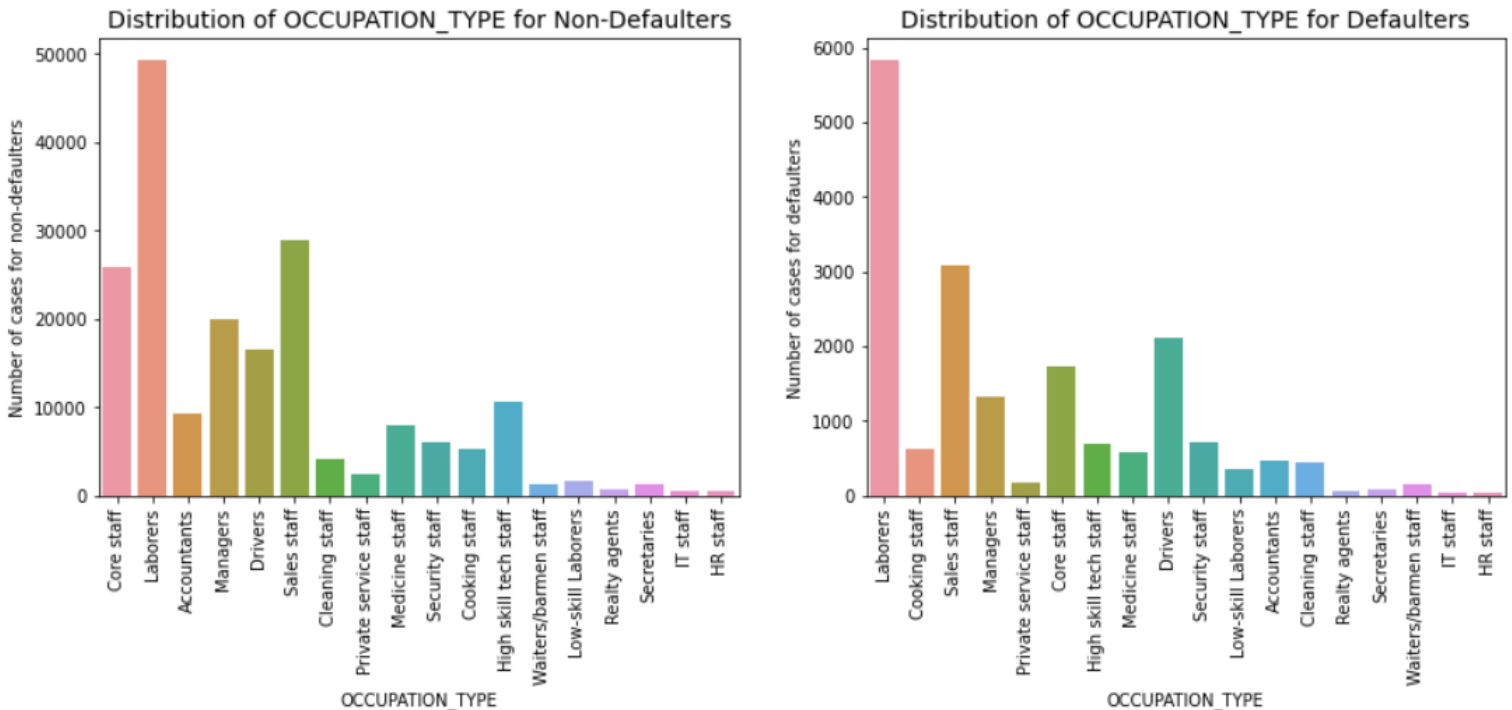
- Majority of defaulters income type is working and at the same time they are good income to bank as well.
- Pensioner of not default case are high in number compared to Pensioner of default case.



EDUCATION TYPE

EDUCATION TYPE:

- From the graph above, we can say that secondary/special educated people are applying loans high in number.
- Higher education count is lesser in defaulted population as compared to non defaulted population.
- People holding Academic degree have applied for loan the least number of times.



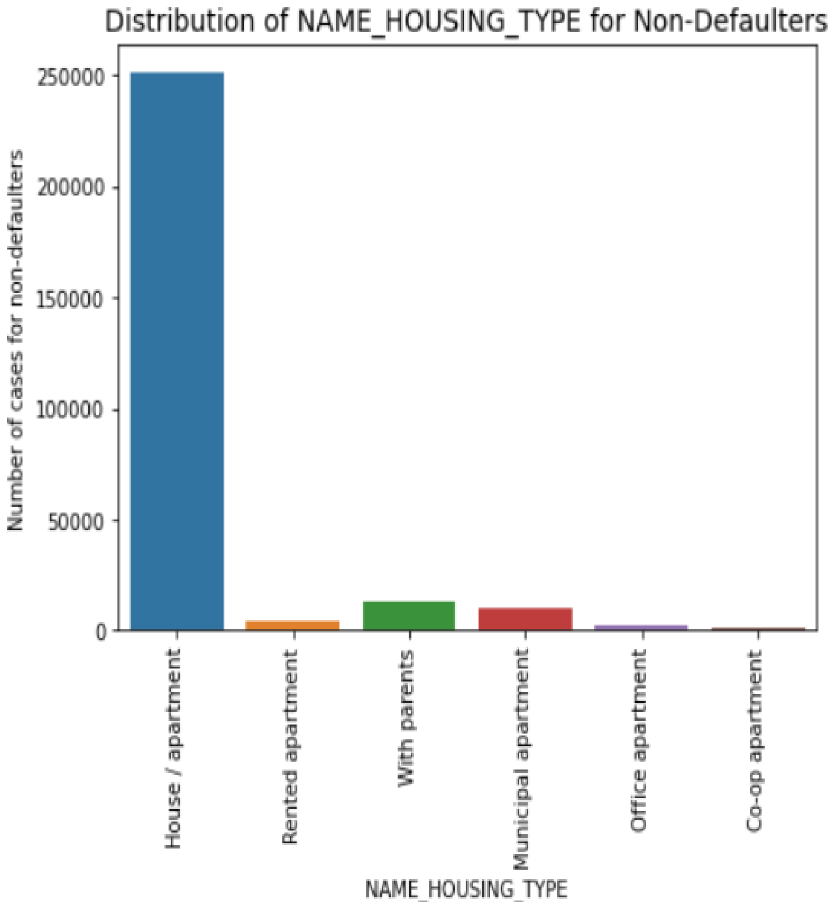
OCCUPATION TYPE

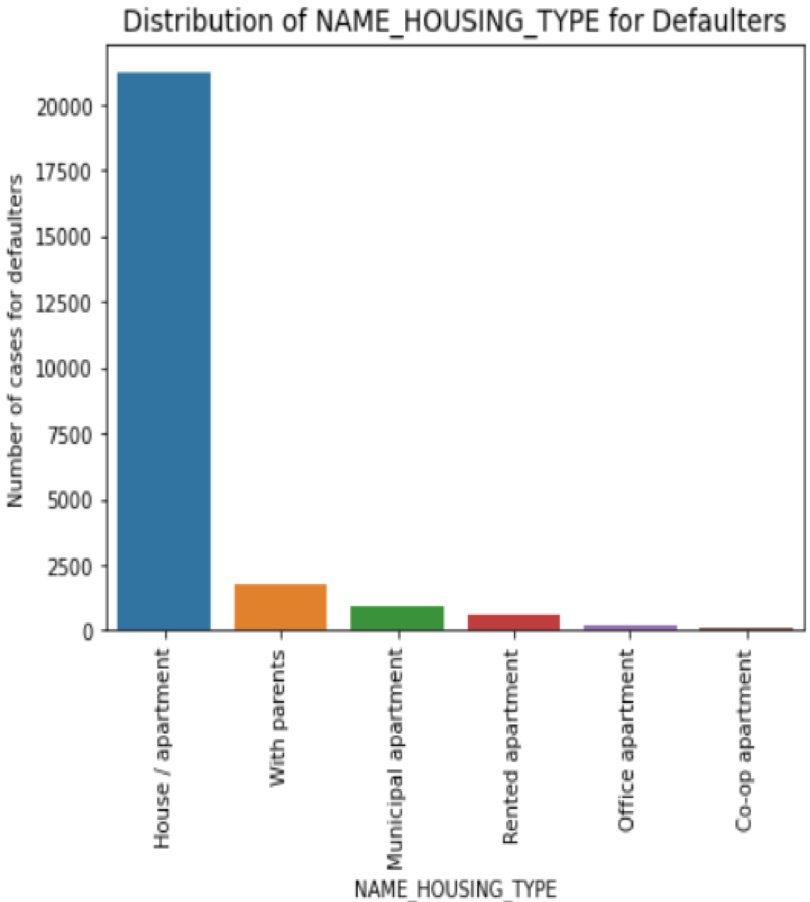
OCCUPATION TYPE

Population with Occupation type as 'Laborers' have the highest number of loan count

Popualtion with Occupation type as 'Core staff' have significantly high proportion of non-defaulters when compared to defaulters.

Population with "HR staff" as Occupation type has lowest loan count.



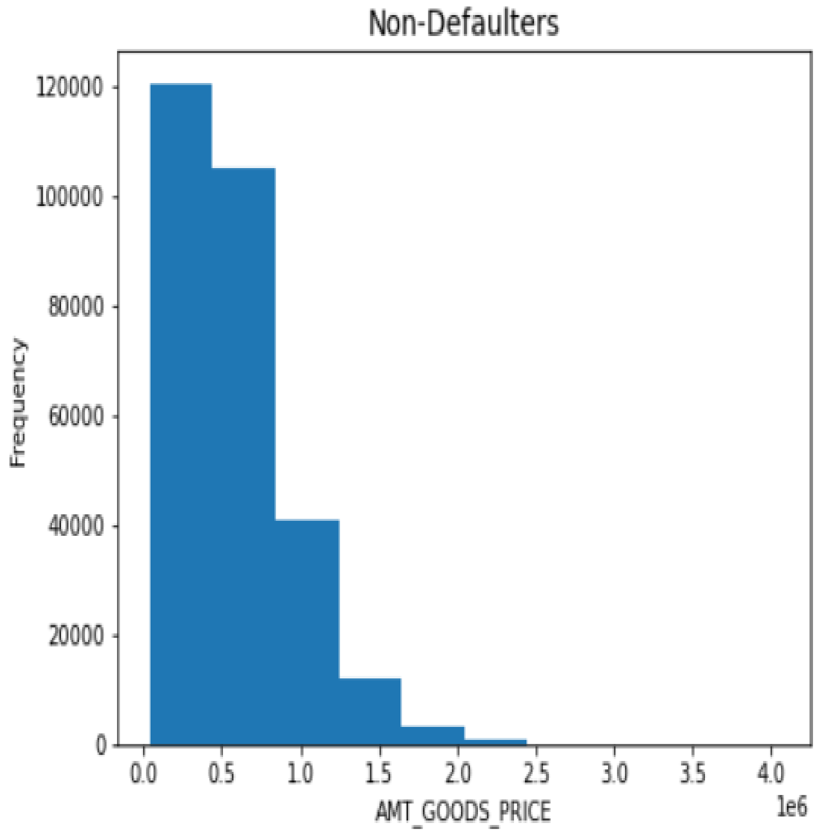


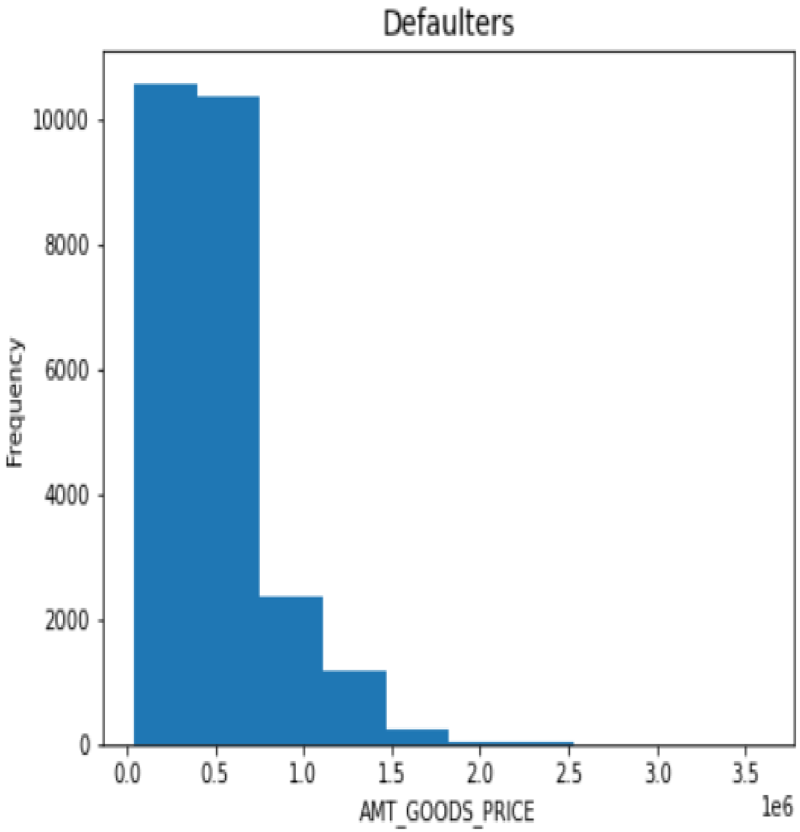
HOUSING TYPE

HOUSING TYPE

- Population living in Rented apartments and with parents have higher default rate as their proportion in defaulted population is higher than non-defaulted population.

- Population living in House/apartment tend to take more loans than any other category

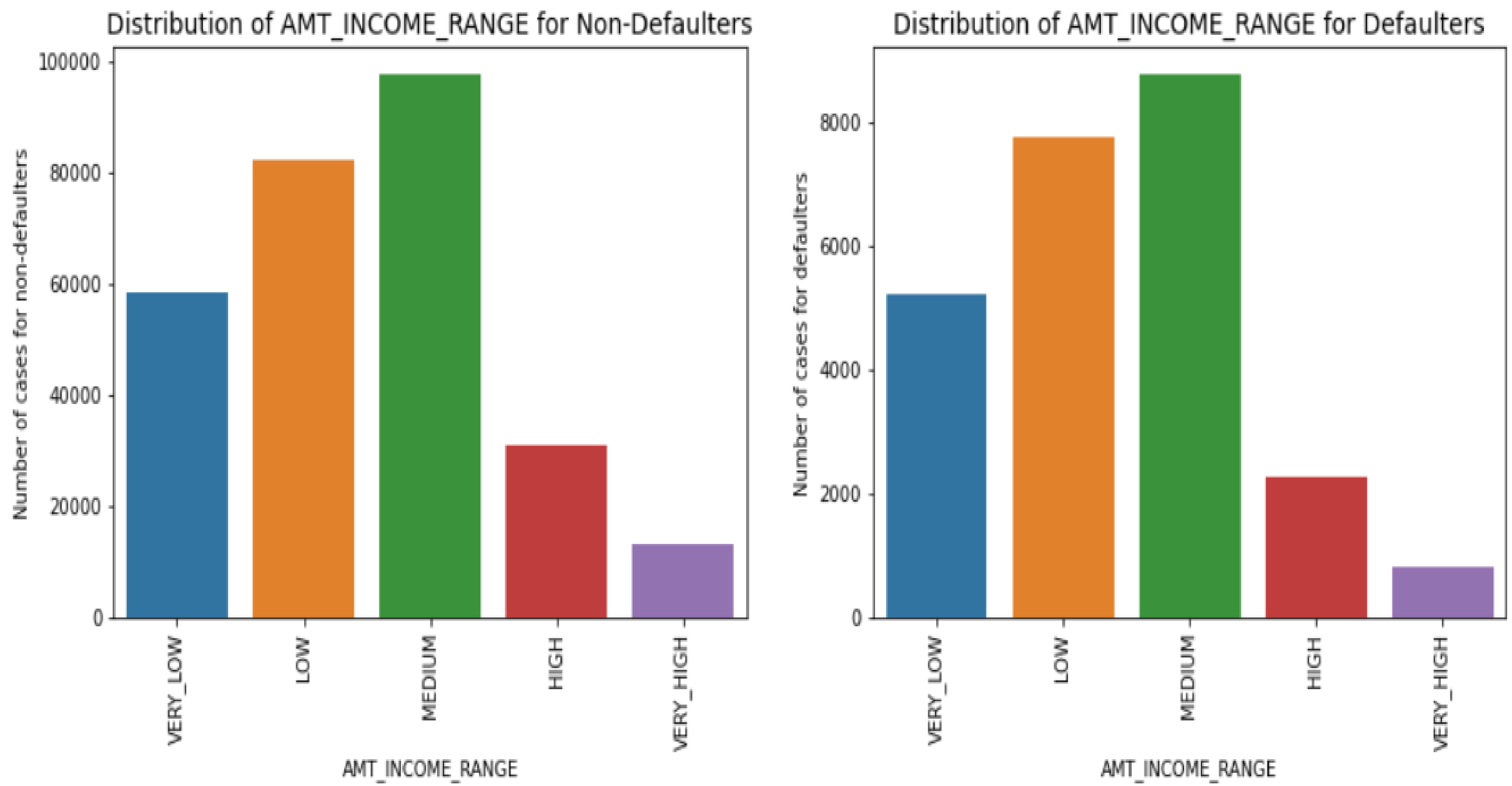

Non-Defaulters
120000
100000
80000
60000
40000
20000
0
Frequency
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
AMT_GOODS_PRICE
1e6


Defaulters
10000
8000
6000
4000
2000
0
Frequency
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
AMT_GOODS_PRICE
1e6

GOODS PRICE

GOODS PRICE

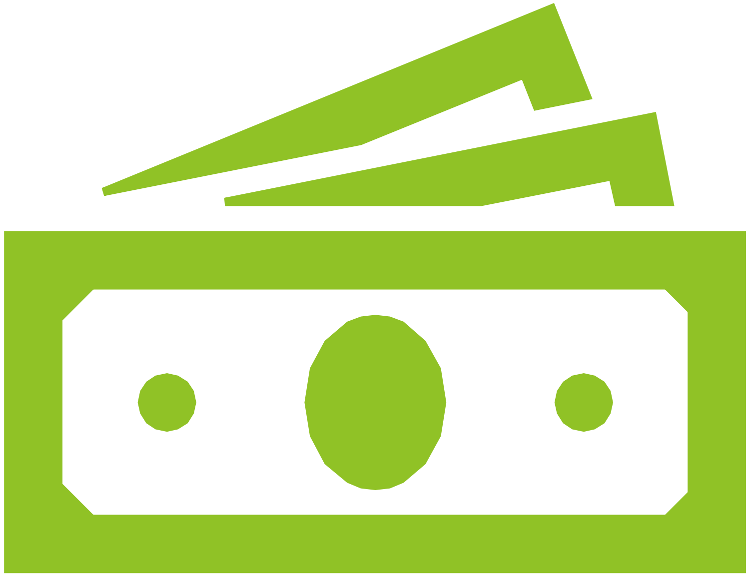
- There's high frequency for loans whose price of the goods for which the loan given is low.
- Defaults are higher for price of the goods for which the loan given is lesser (between 0 to 7,00,000).

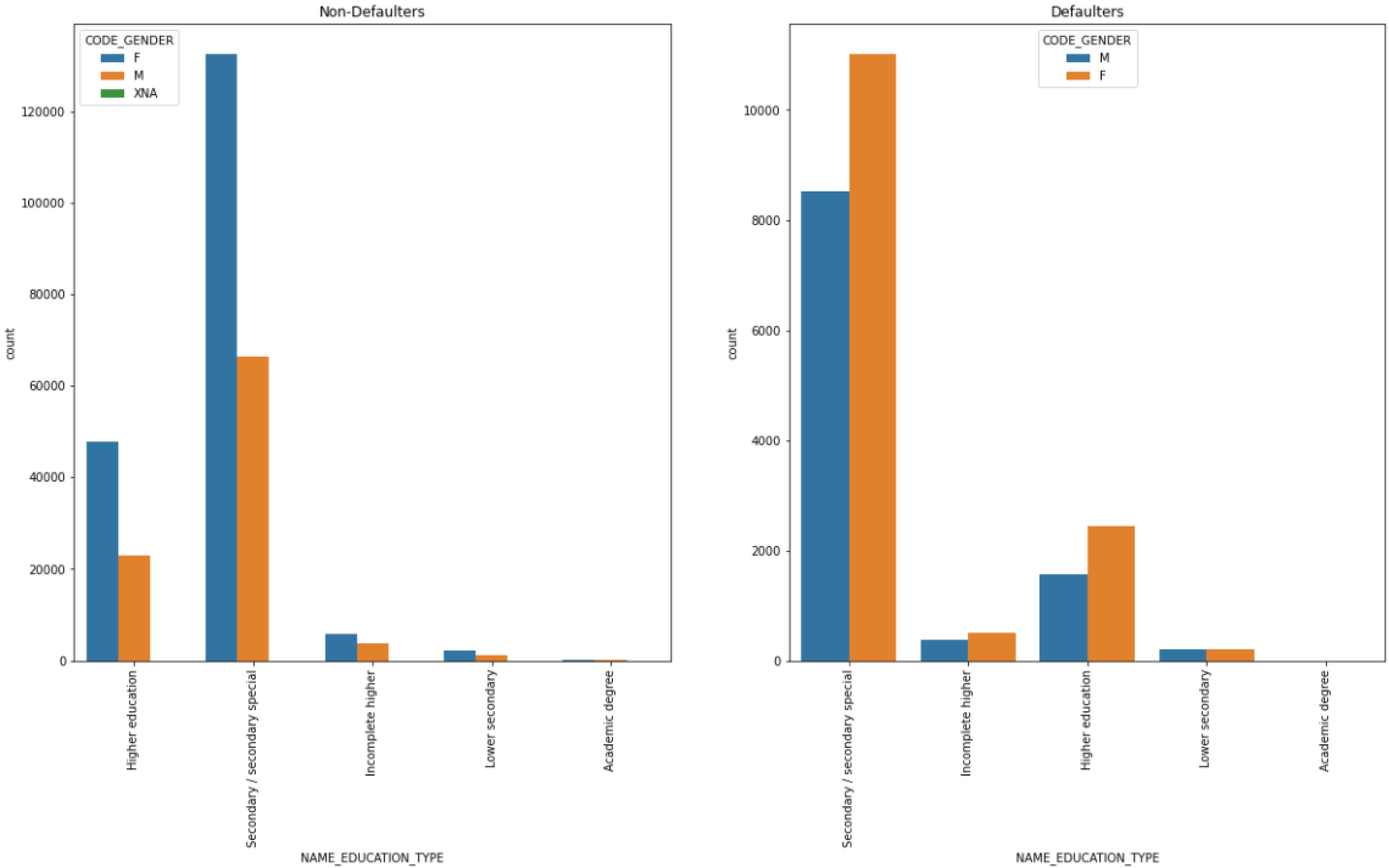


INCOME RANGE

INCOME RANGE

- Population with medium-income range has the highest count of loan.
- Population with low and very low-income range has higher defaults as their proportion in defaulted population is higher than in the non defaulted population.
- Population with very high-income range has the lowest loan count.

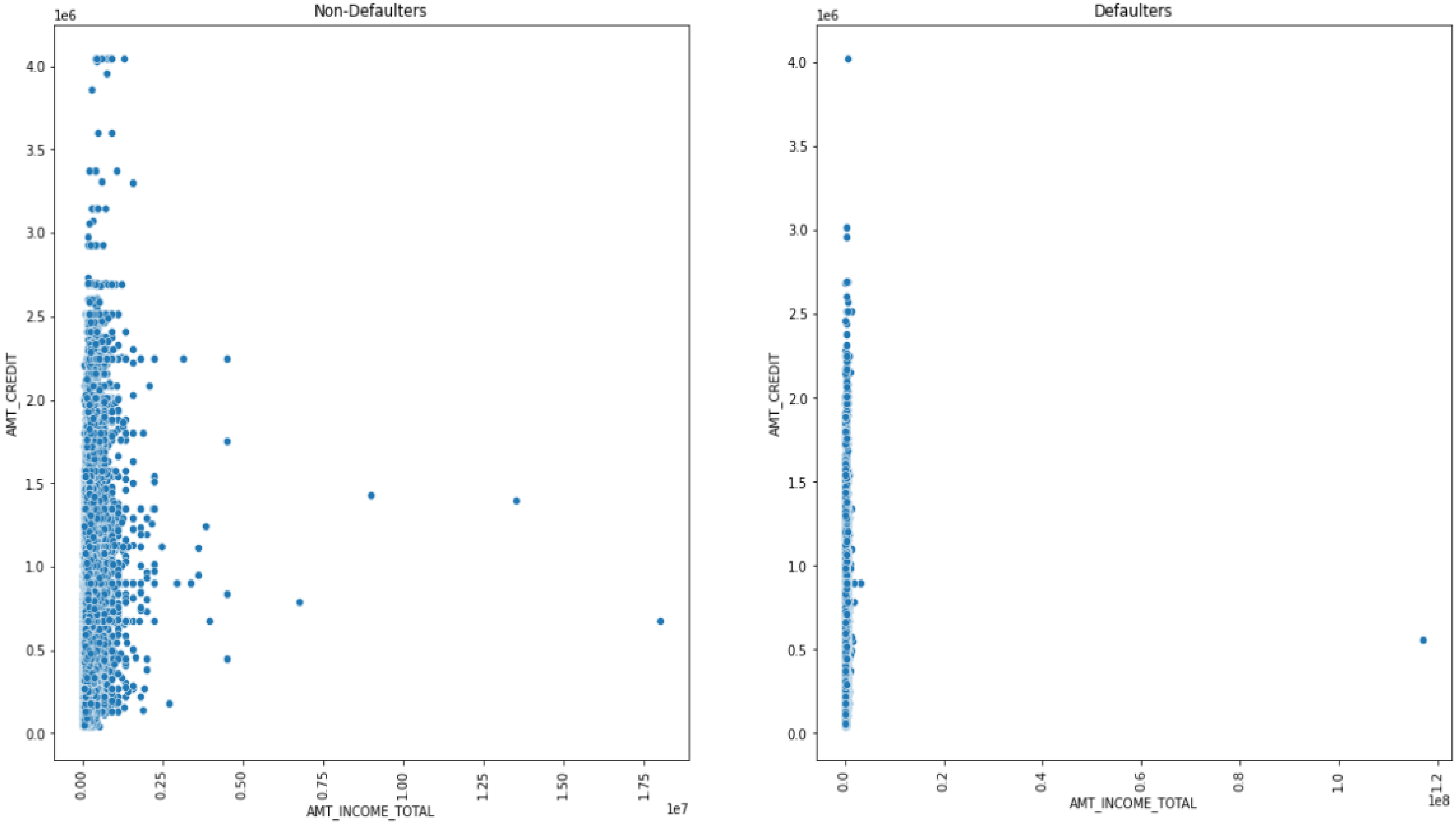




EDUCATION TYPE vs. GENDER

EDUCATION TYPE vs. GENDER

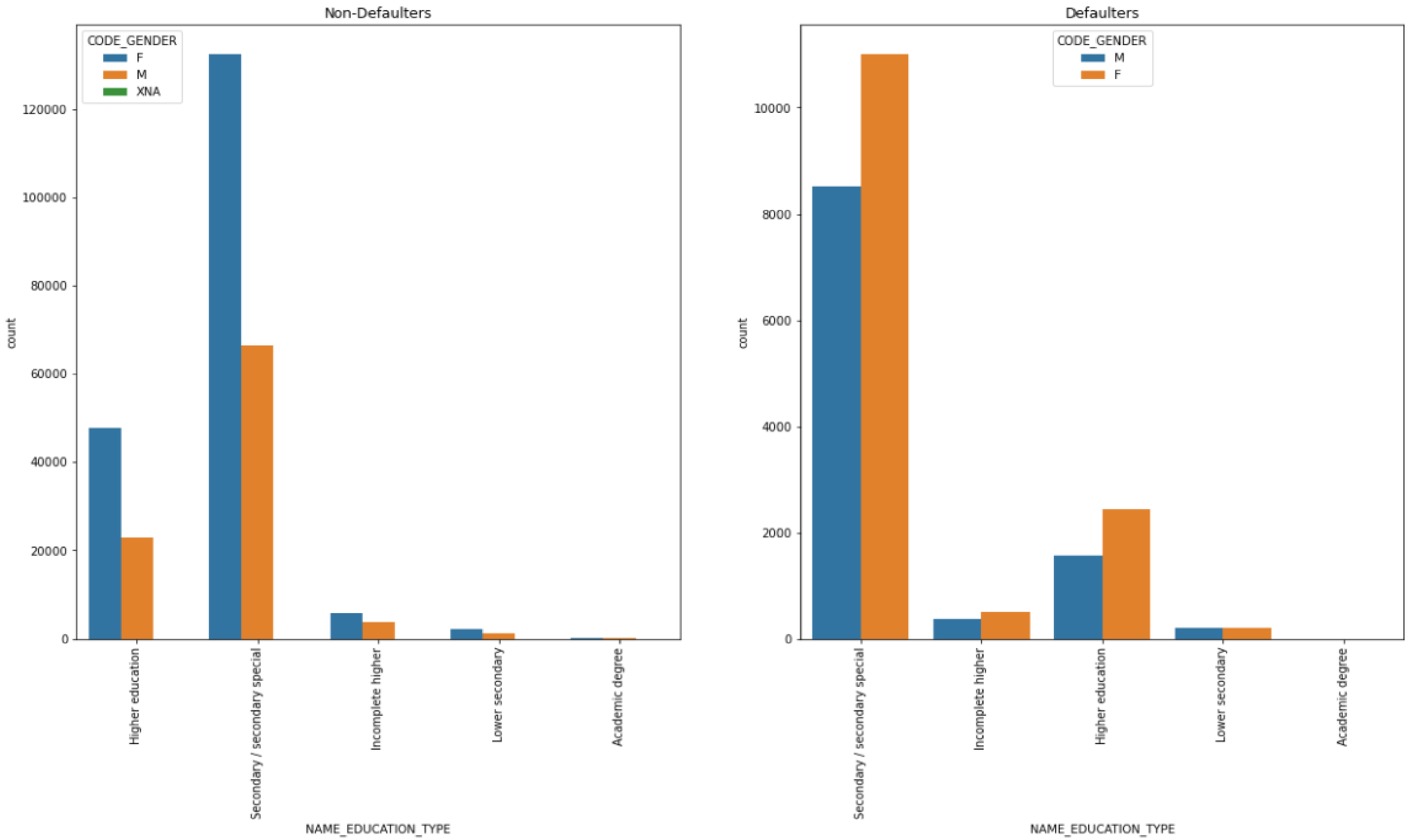
- Both Males and Females with secondary/secondary special education type tend to take more loan than any other category.
- Population with Academic degree has lowest count of loan.



INCOME TYPE vs. CREDIT AMOUNT

INCOME TYPE vs. CREDIT AMOUNT

- Population with lower income are tend to find it difficult to repay the loan .

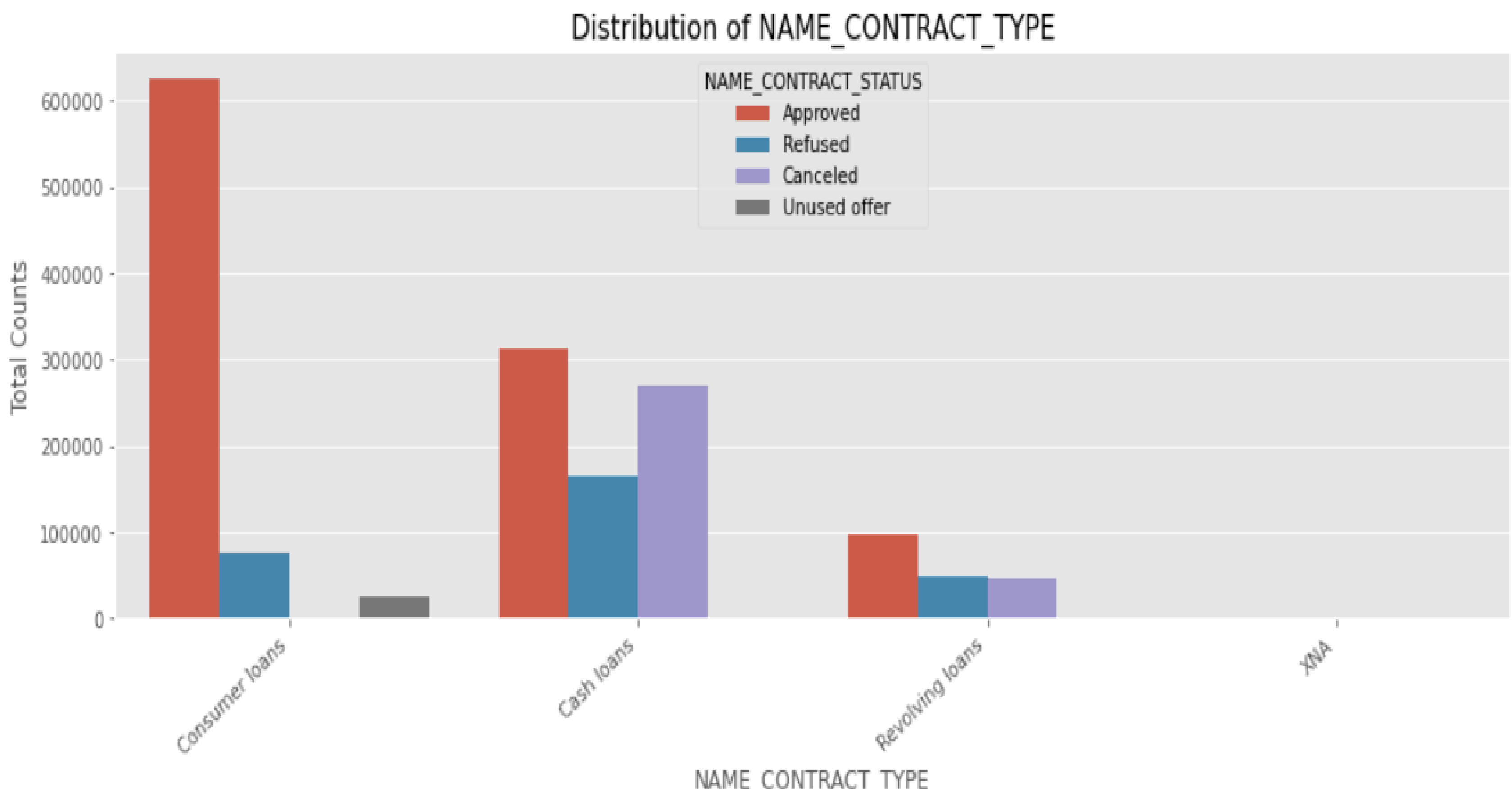


EDUCATION TYPE vs. GENDER

EDUCATION TYPE vs. GENDER

- Both Males and Females with secondary/secondary special education type tend to take more loan than any other category.

- Population with Academic degree has lowest count of loan.

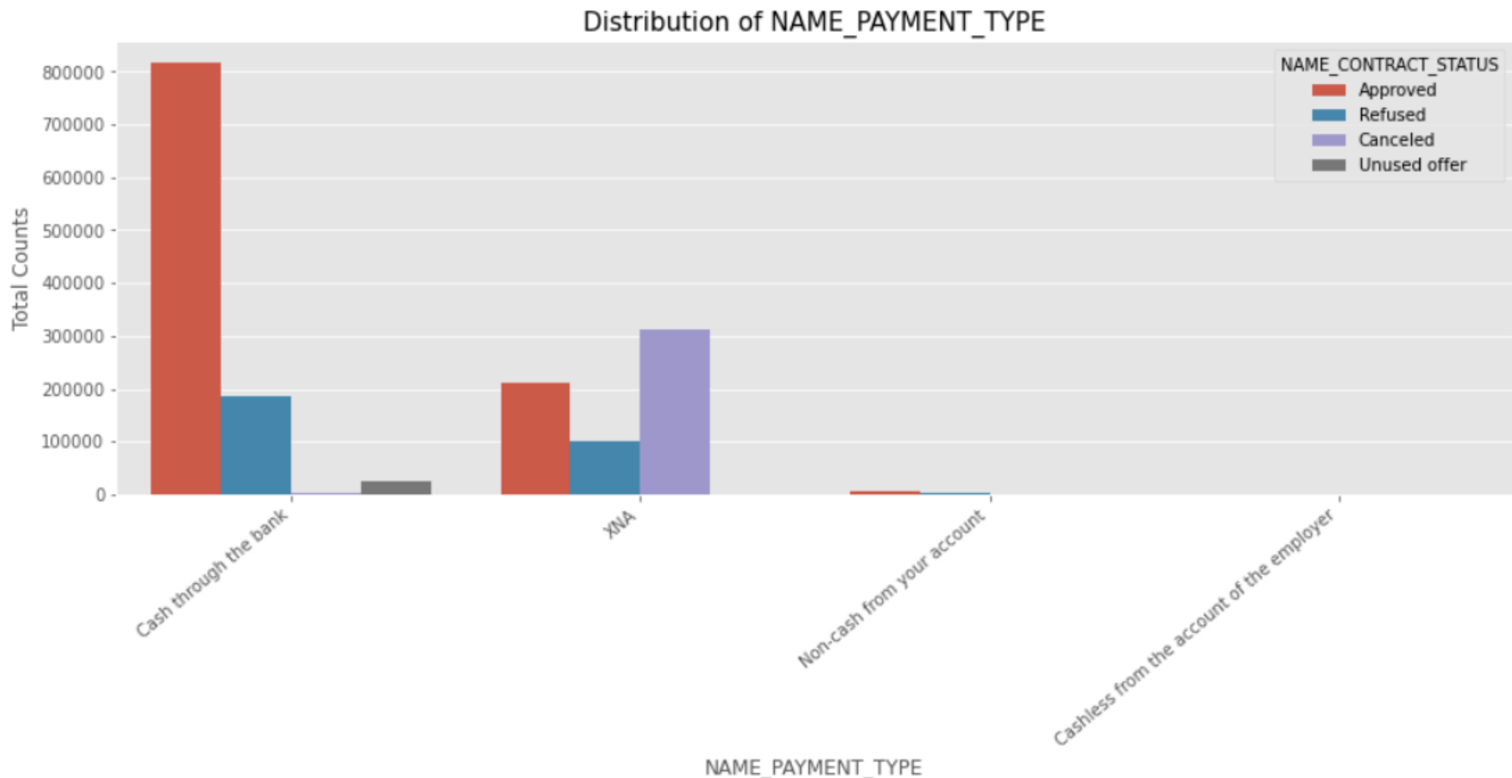


LOAN TYPE vs. APPLICATION STATUS

LOAN TYPE vs. APPLICATION STATUS

From the above chart, we can infer that, most of the applications are for 'Cash loan' and 'Consumer loan’.

Although the cash loans are refused more often than others.

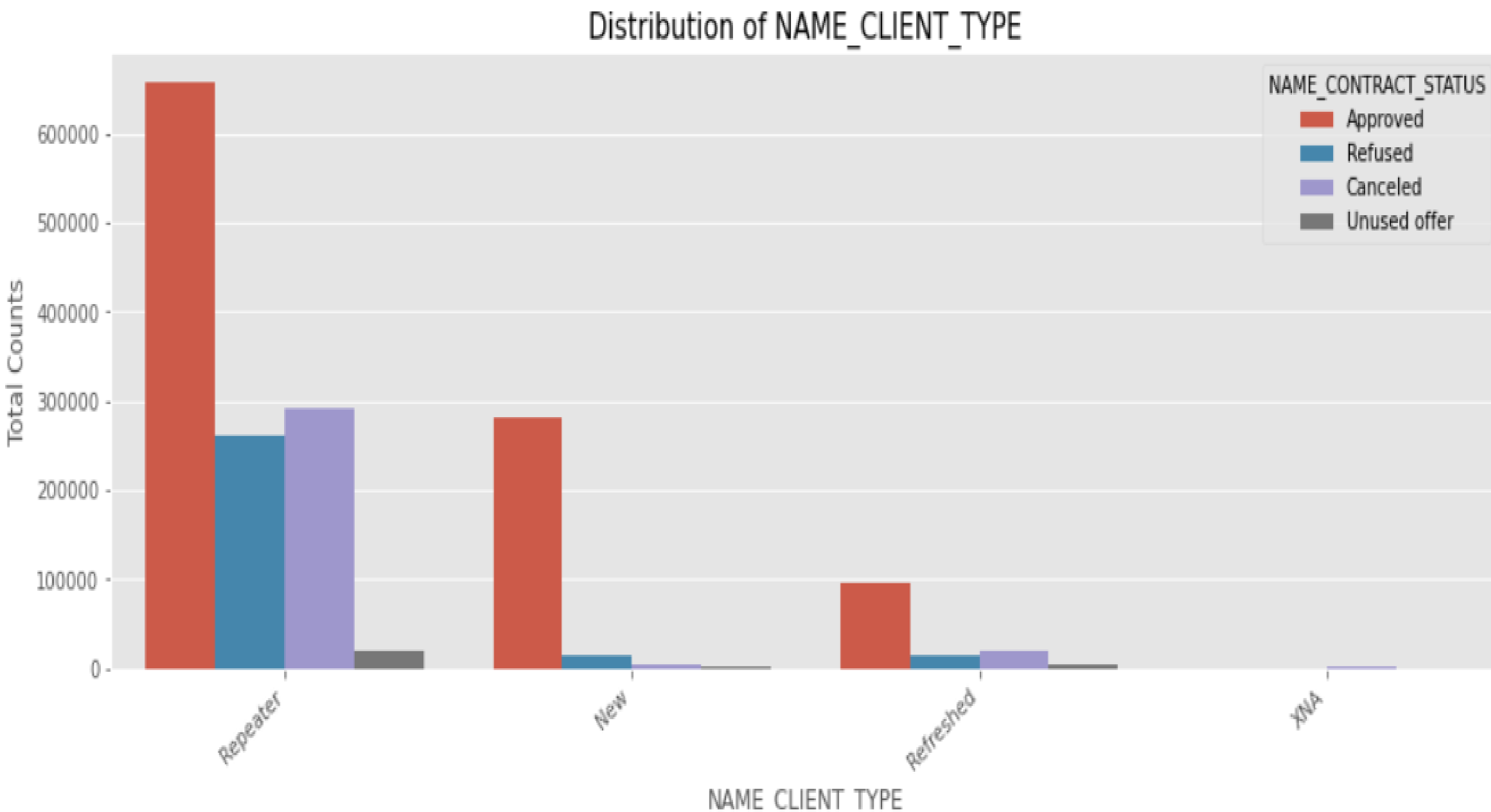


PAYMENT TYPE vs. APPLICATION STATUS

PAYMENT TYPE vs. APPLICATION STATUS

From the above chart, we can infer that most of the clients chose to repay the loan using the 'Cash through the bank' option.

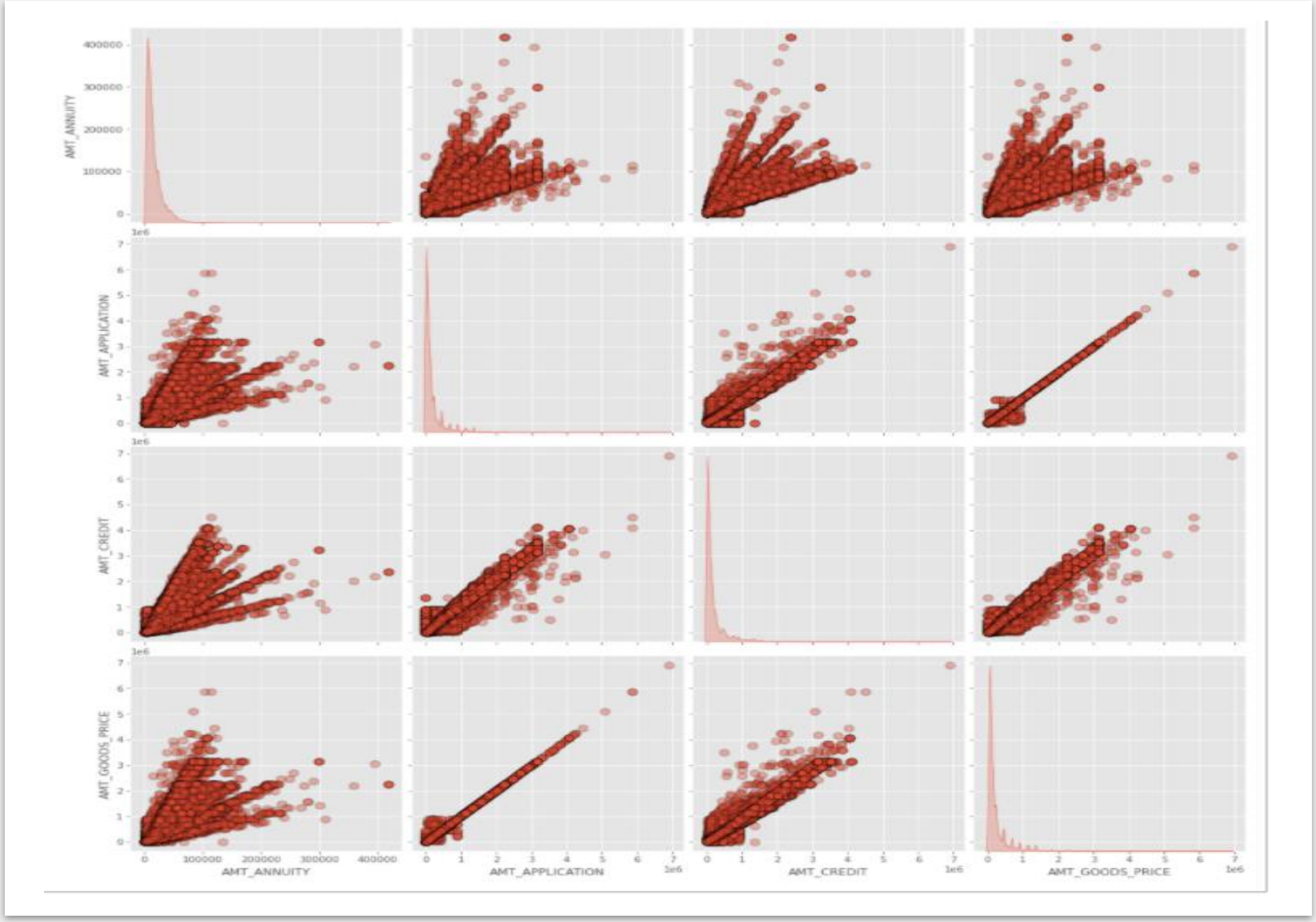
We can also see that 'Non-Cash from your account' & 'Cashless from the account of the employee' options are not at all popular in terms of loan repayment amongst the customers.



CLIENT TYPE vs. APPLICATION STATUS

CLIENT TYPE vs. APPLICATION STATUS

- Most of the loan applications are from repeat customers, out of the total applications 70% of customers are repeaters. They also get refused most often.



CORRELATION BETWEEN DIFFERENT VARIABLES

CORRELATION BETWEEN DIFFERENT VARIABLES

- 1. Annuity of previous application has a very high and positive influence over: (Increase of annuity increases below factors):-

- How much credit did client asked on the previous application
- Final credit amount on the previous application that was approved by the bank
- Goods price of good that client asked for on the previous application.

- 2. For how much credit did client ask on the previous application is highly influenced by the Goods price of good that client has asked for on the previous application

- 3. Final credit amount disbursed to the customer previously, after approval is highly influence by the application amount and also the goods price of good that client asked for on the previous application.

CONCLUSION

- As per our EDA top major variables to consider for loan prediction are as follows:
  - NAME_EDUCATION_TYPE
  - AMT_INCOME_TOTAL
  - DAYS_BIRTH
  - AMT_CREDIT
  - NAME_INCOME_TYPE
  - CODE_GENDER
  - NAME_HOUSING_TYPE
  - AMOUNT_GOODS_PRICE
  - OCCUPATION_TYPE
  - CODE_GENDER
- The above-mentioned variables are to be considered before approving or rejecting the application to minimize the risk of loss or to maximize the profit from potential customers however these variables are not limited.