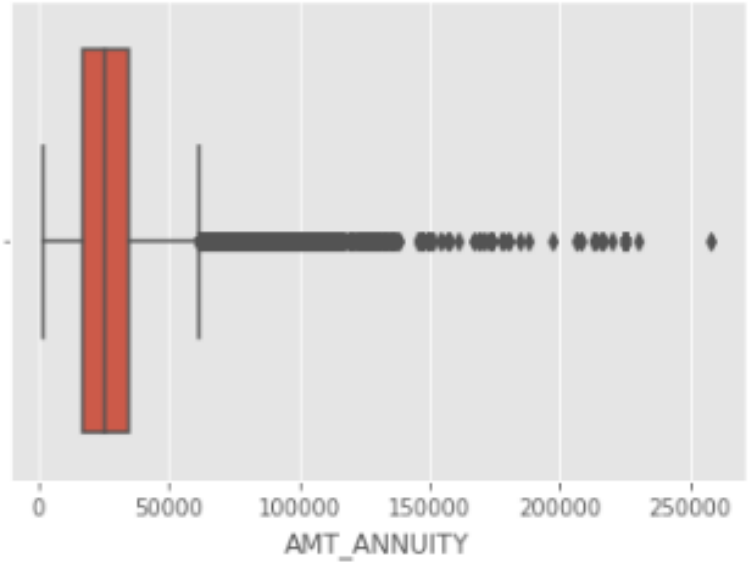
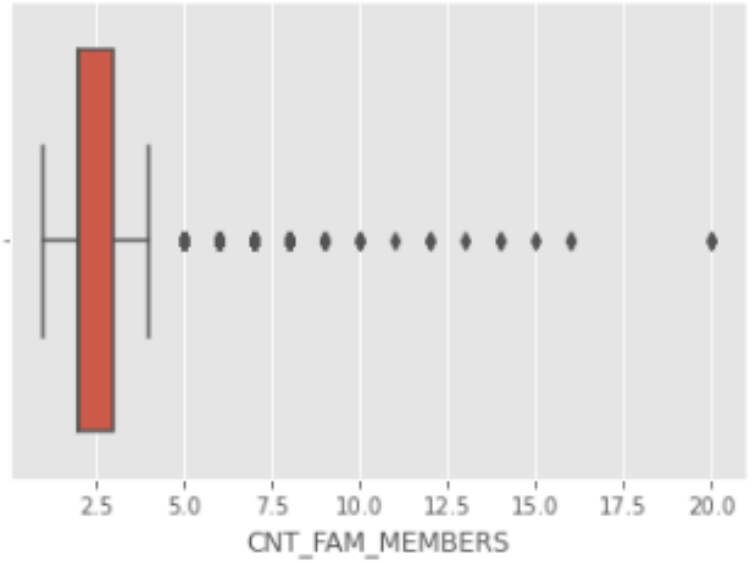
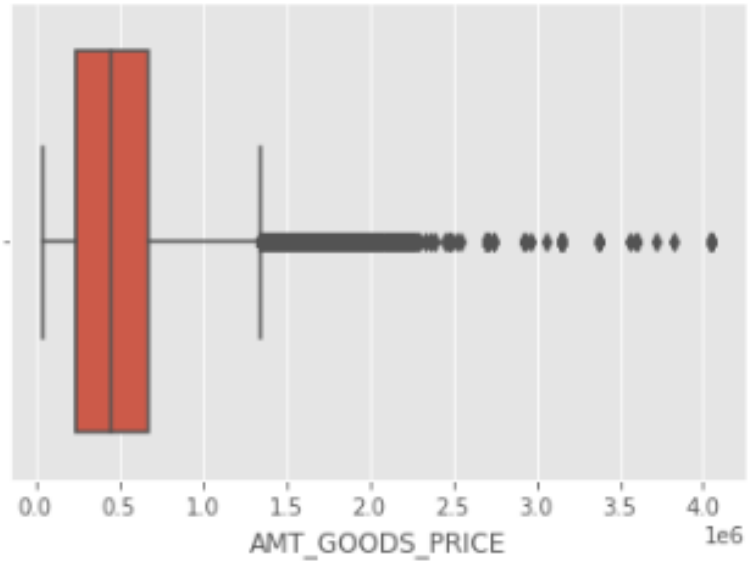
Credit Exploratory Data Analysis Case Study

Kamal Tamang and Siddesh Sawant

Continuous Variable Outliers



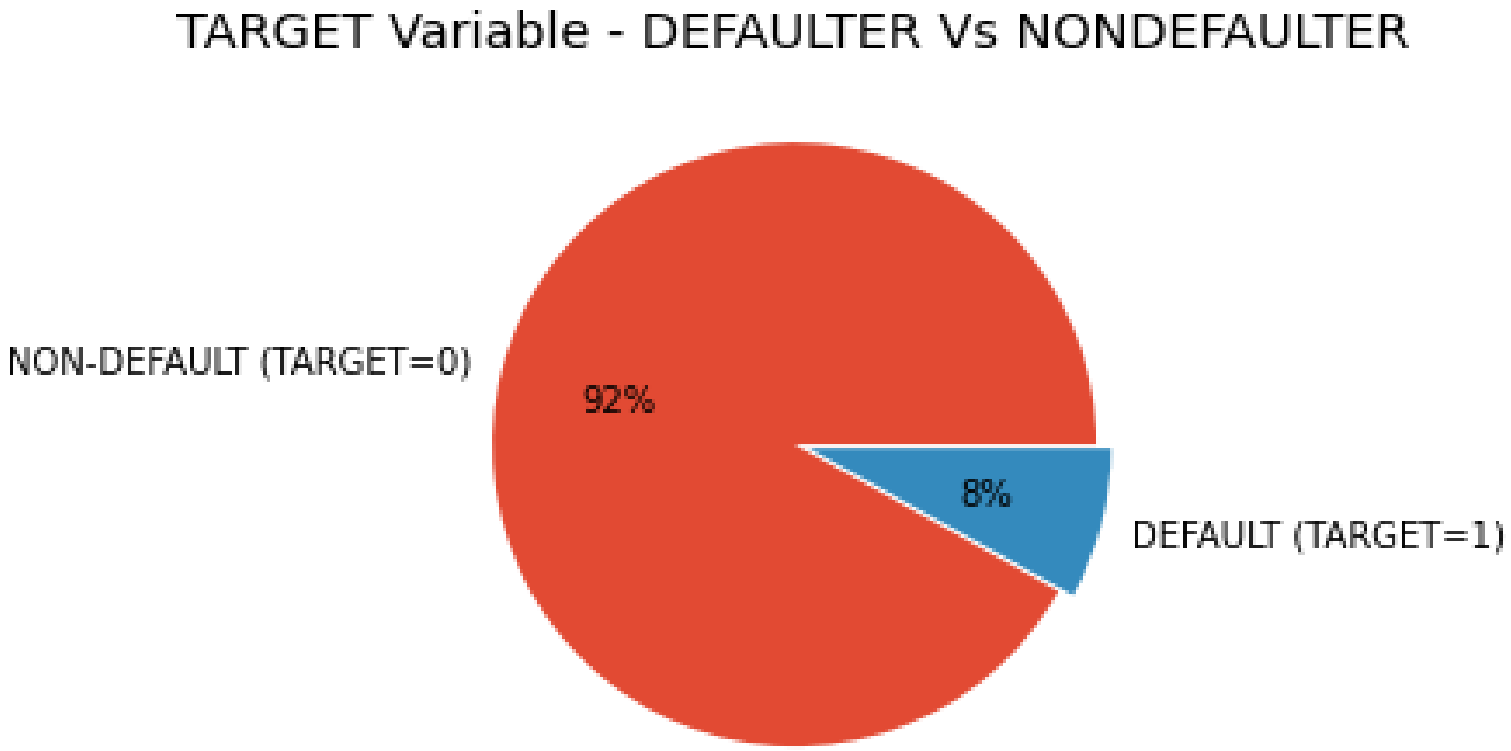




1. AMT_ANNUITY has outliers, the column can be imputed using the median of the column i.e. 24903.0
2. CNT_FAM_MEMBERS has outliers, the column can be imputed using the median of the column i.e. 2.0
3. AMT_GOODS_PRICE has outliers, the column can be imputed using the median of the column i.e. 450000.0

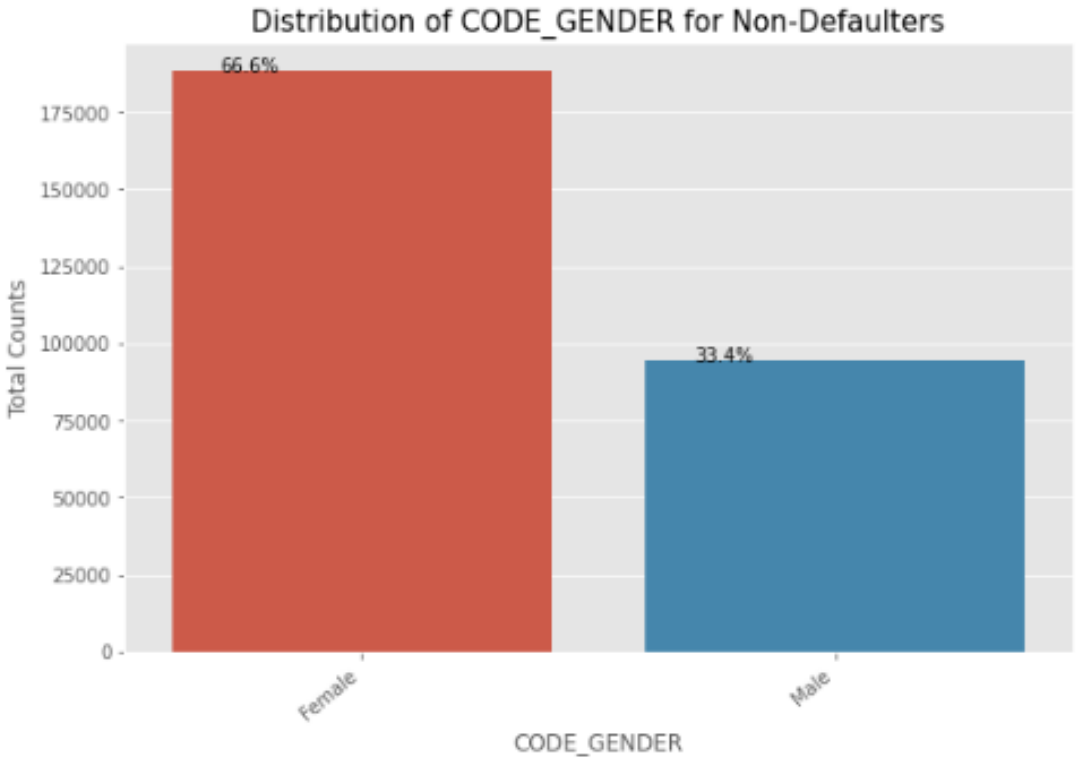
Imbalance in Target

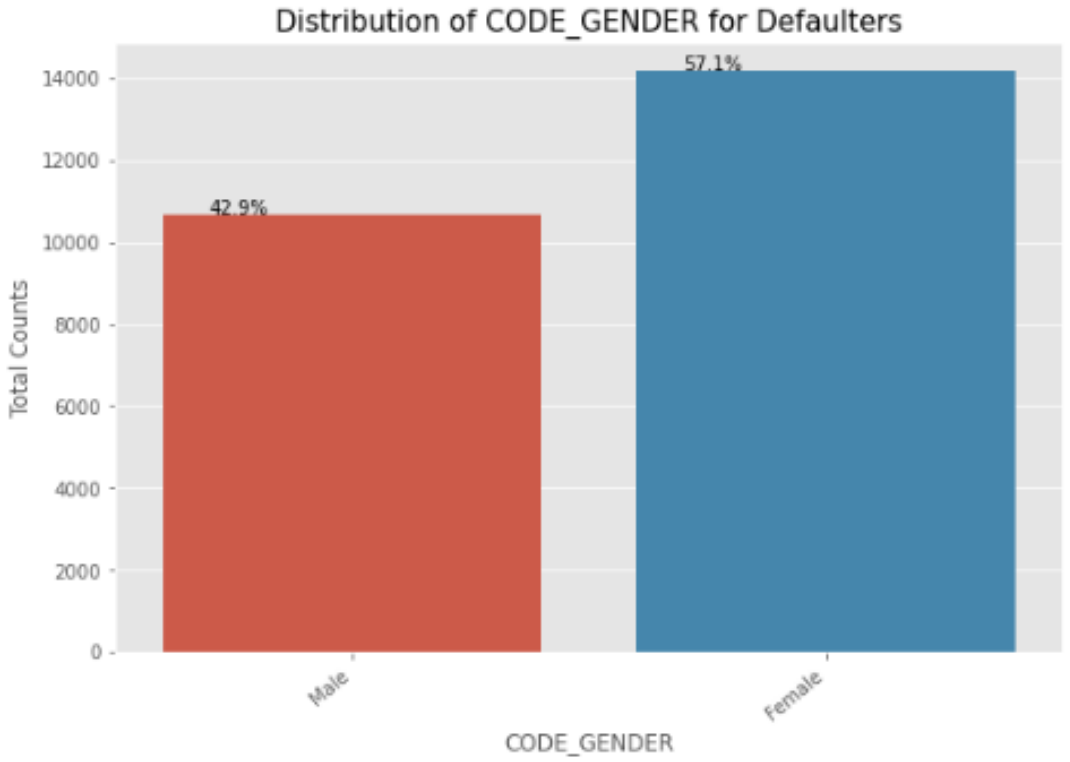
Its clear that there is an imbalance between people who defaulted and who didn't default. More than 92% of people didn't default as opposed to 8% who defaulted.



Univariate Analysis

Gender





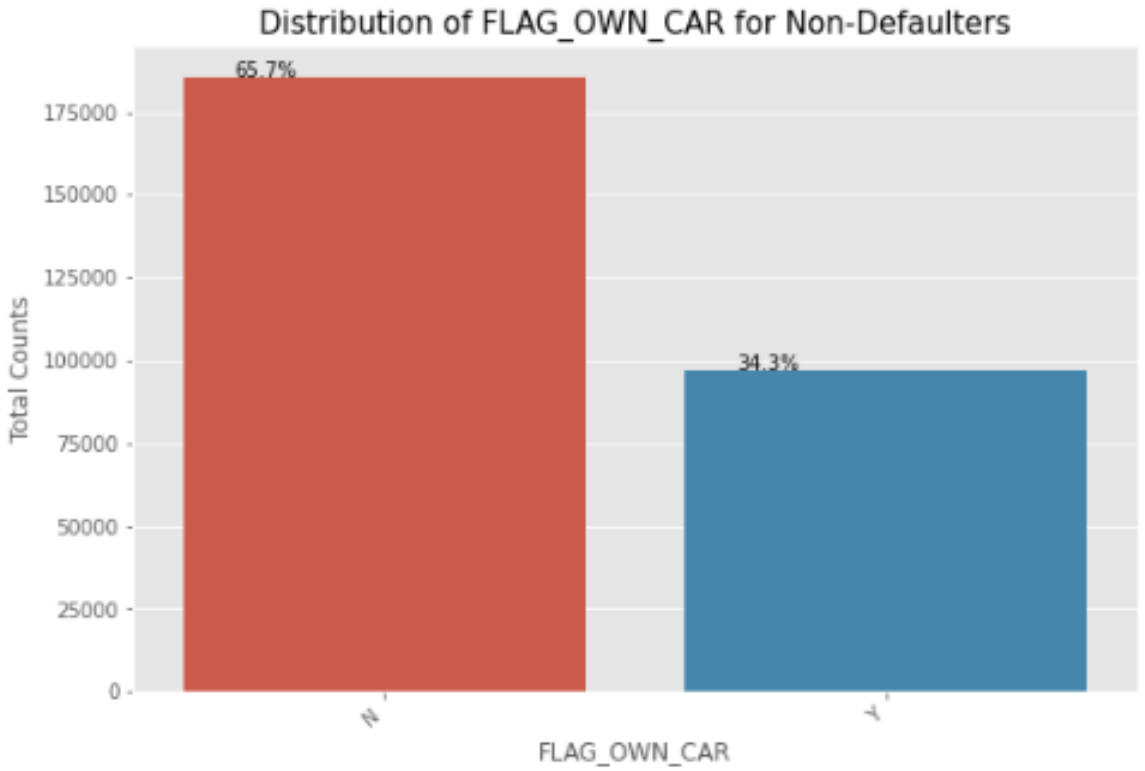
We can see that Female contribute 67% to the non-defaulters while 57% to the defaulters. We can conclude that

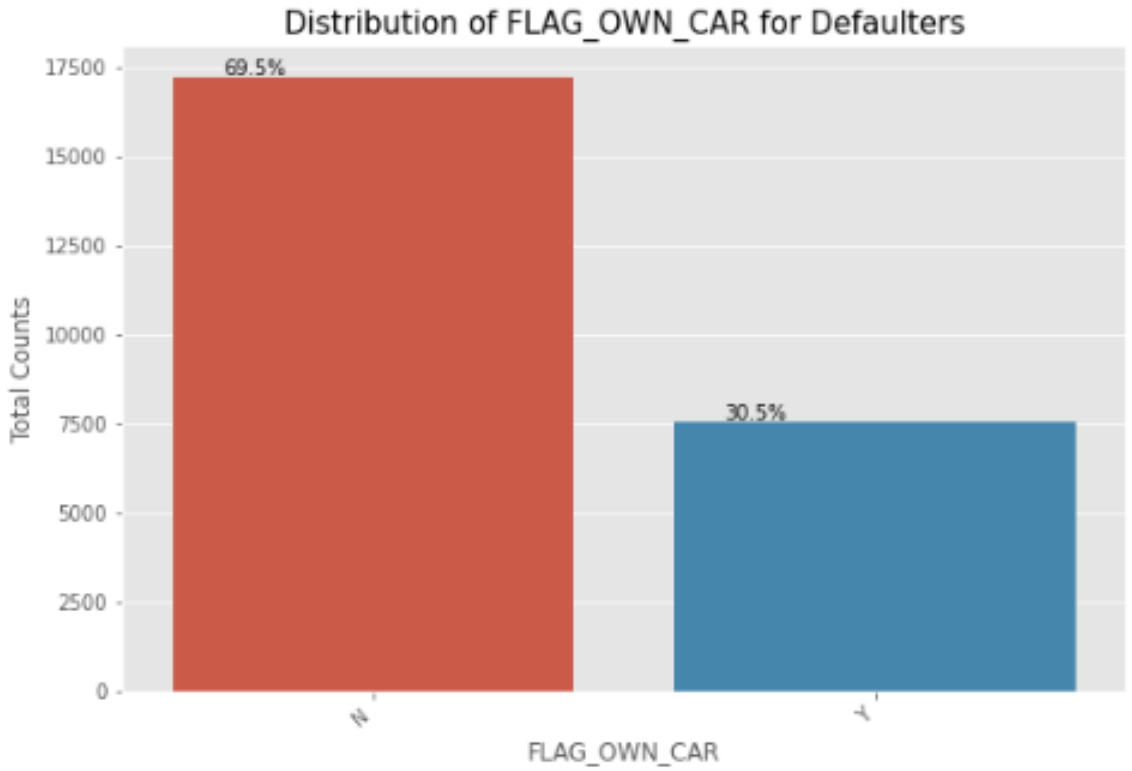
We see more female applying for loans than males and hence the more number of female defaulters as well.

But the rate of defaulting of FEMALE is much lower compared to their MALE counterparts.

Univariate Analysis

Own Car





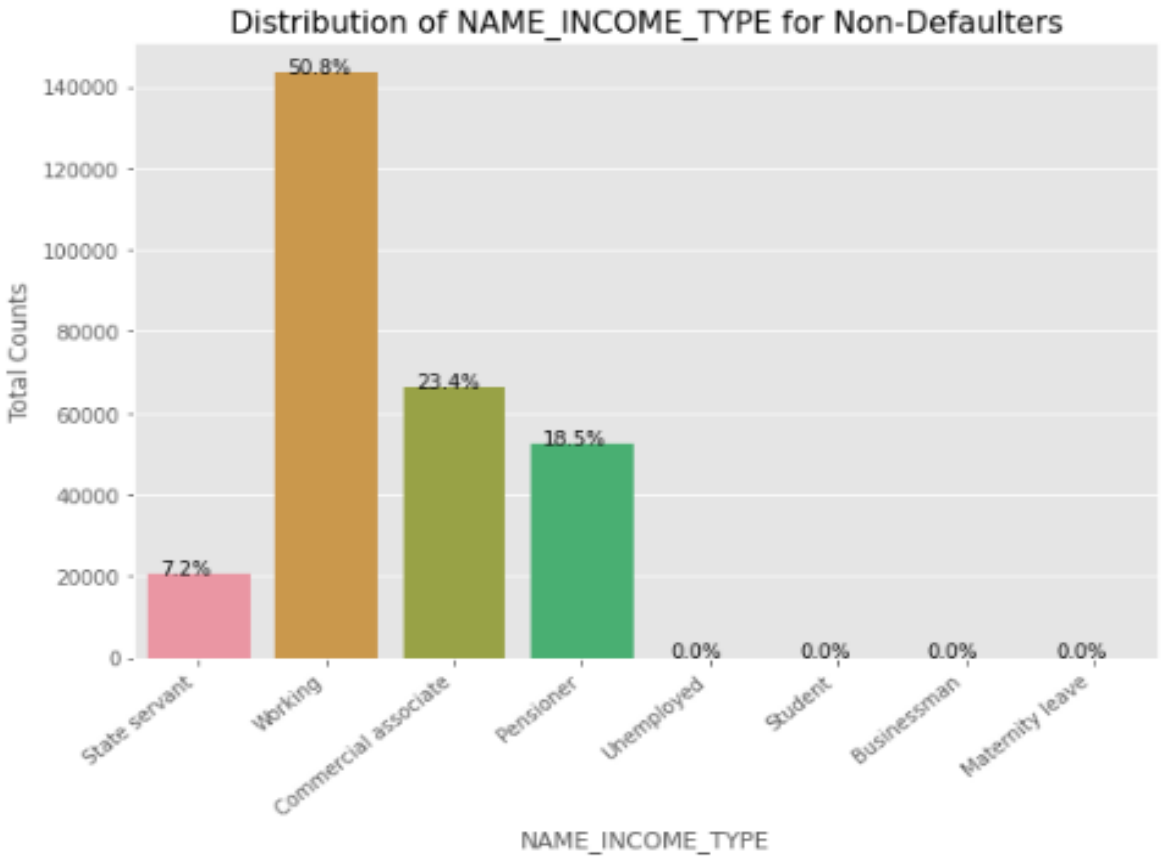
We can see that people with cars contribute 65.7% to the non-defaulters while 69.5% to the defaulters. We can conclude that

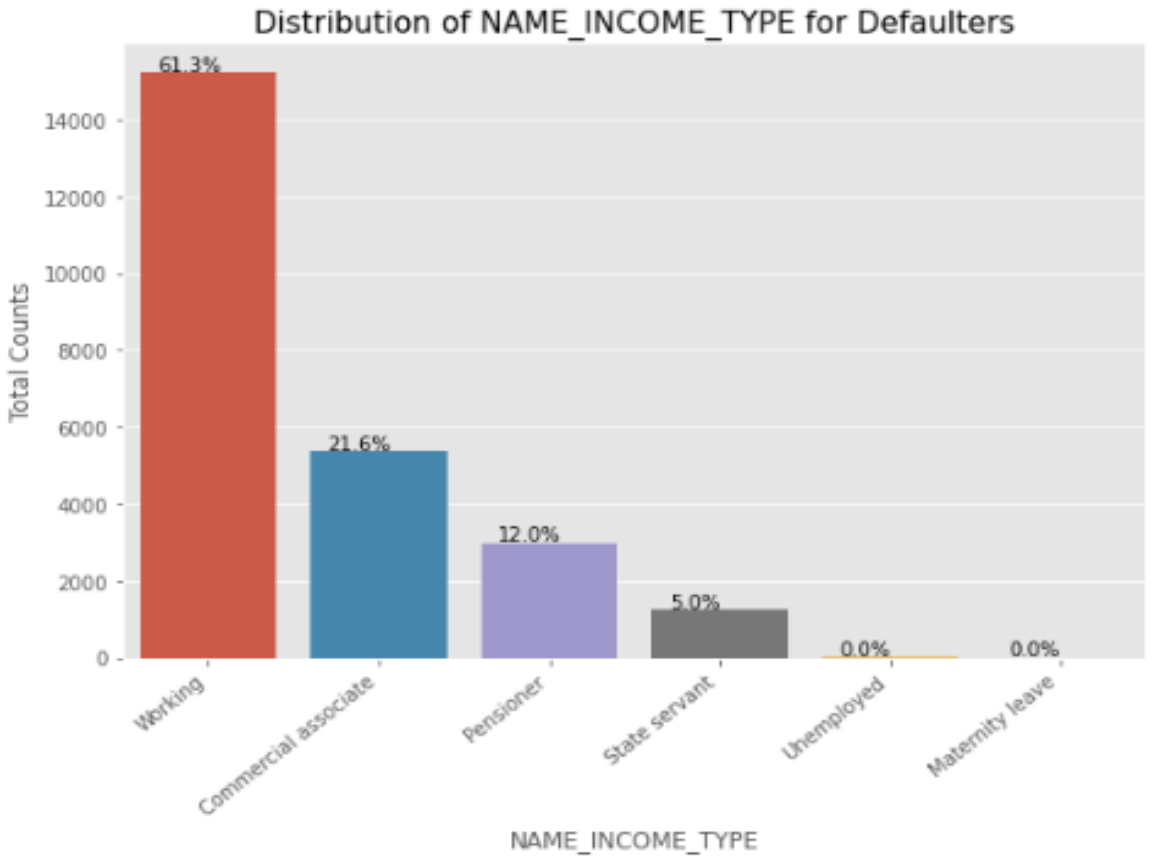
While people who have car default more often, the reason could be there are simply more people without cars

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.

Univariate Analysis

Income Type

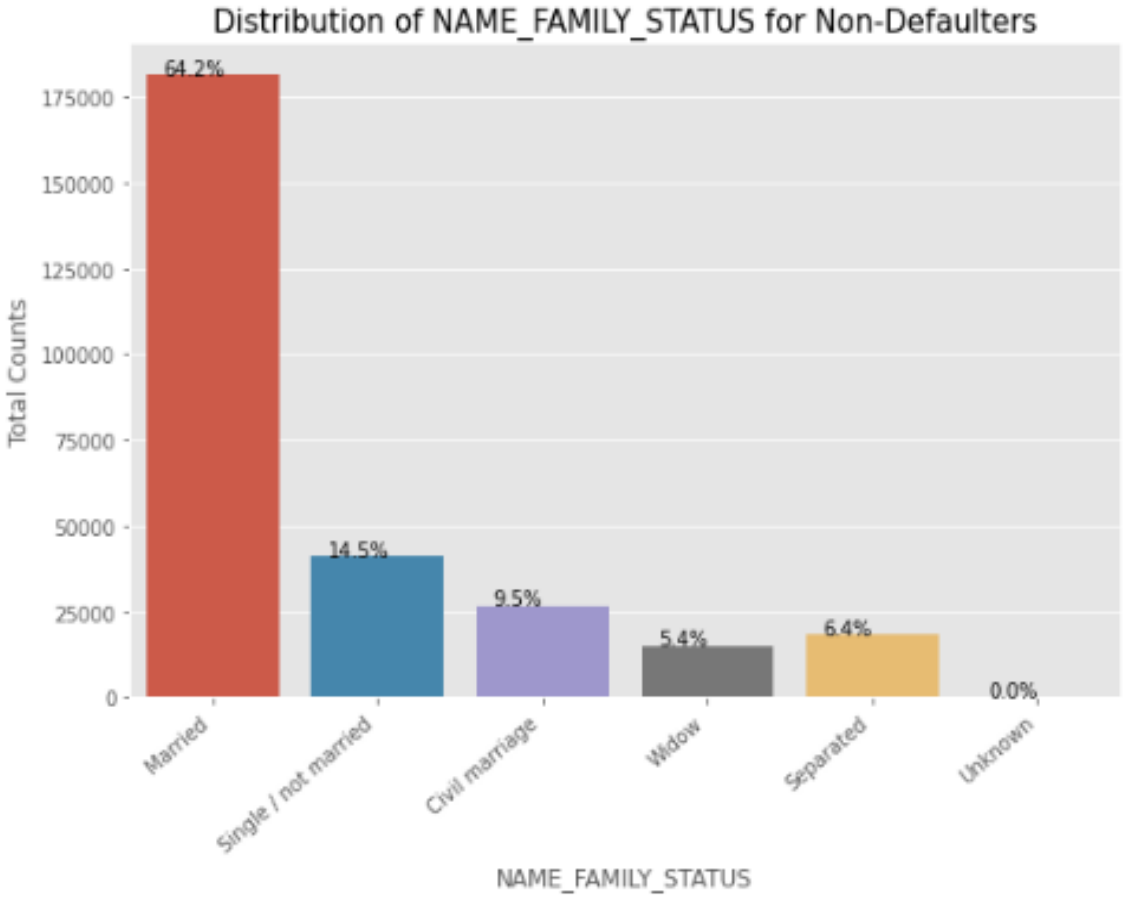


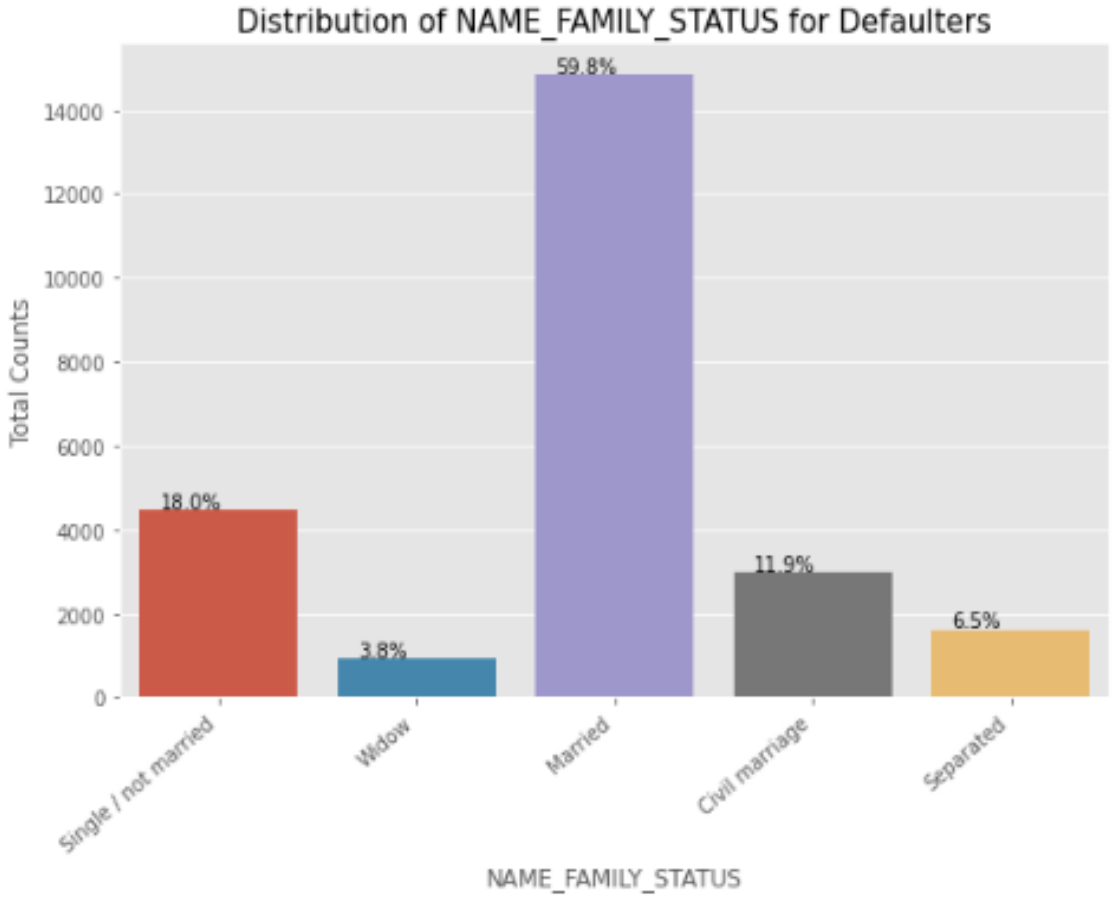


We can notice that the students don't default. The reason could be they are not required to pay during the time they are students.
We can also see that the Business Men never default.
Most of the loans are distributed to working class people
We also see that working class people contribute 51% to non defaulters while they contribute to 61% of the defaulters. Clearly, the chances of defaulting are more in their case.

Univariate Analysis

Family Status

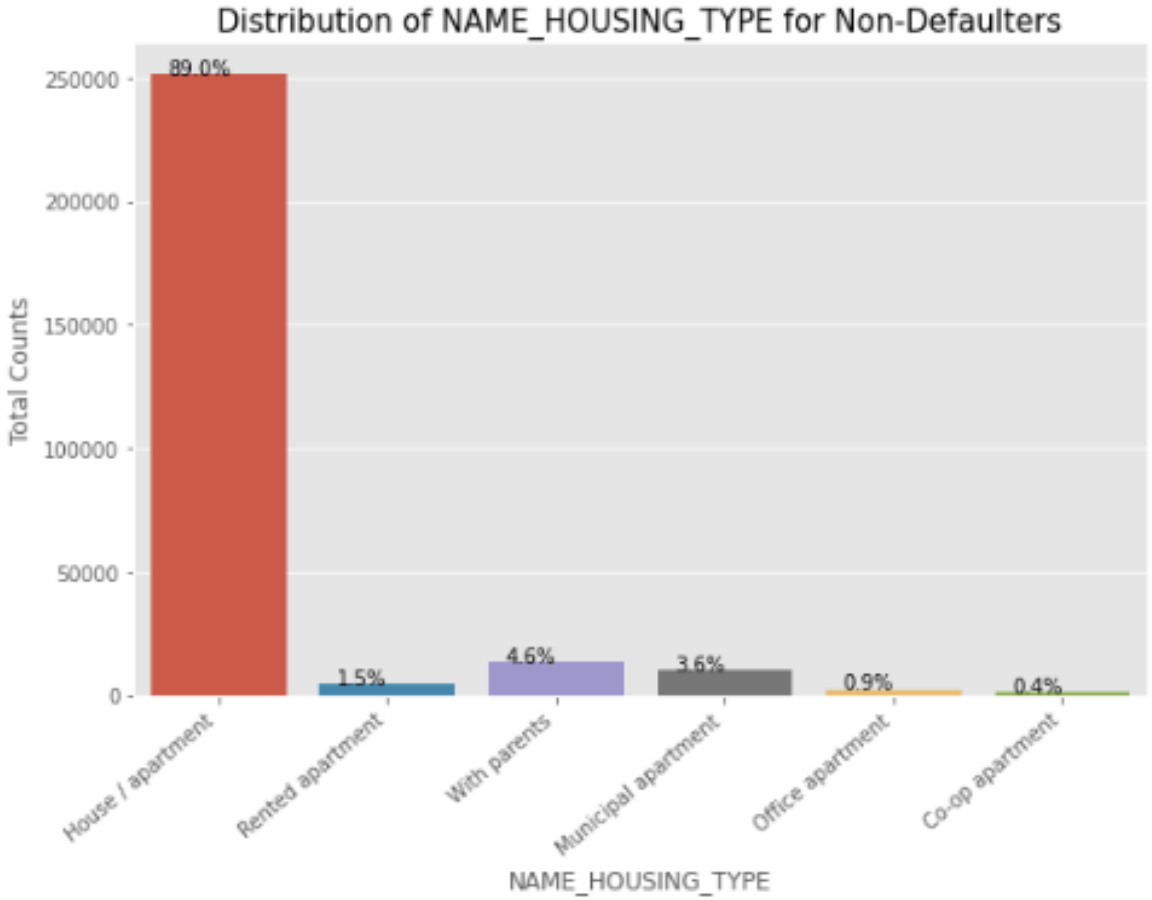


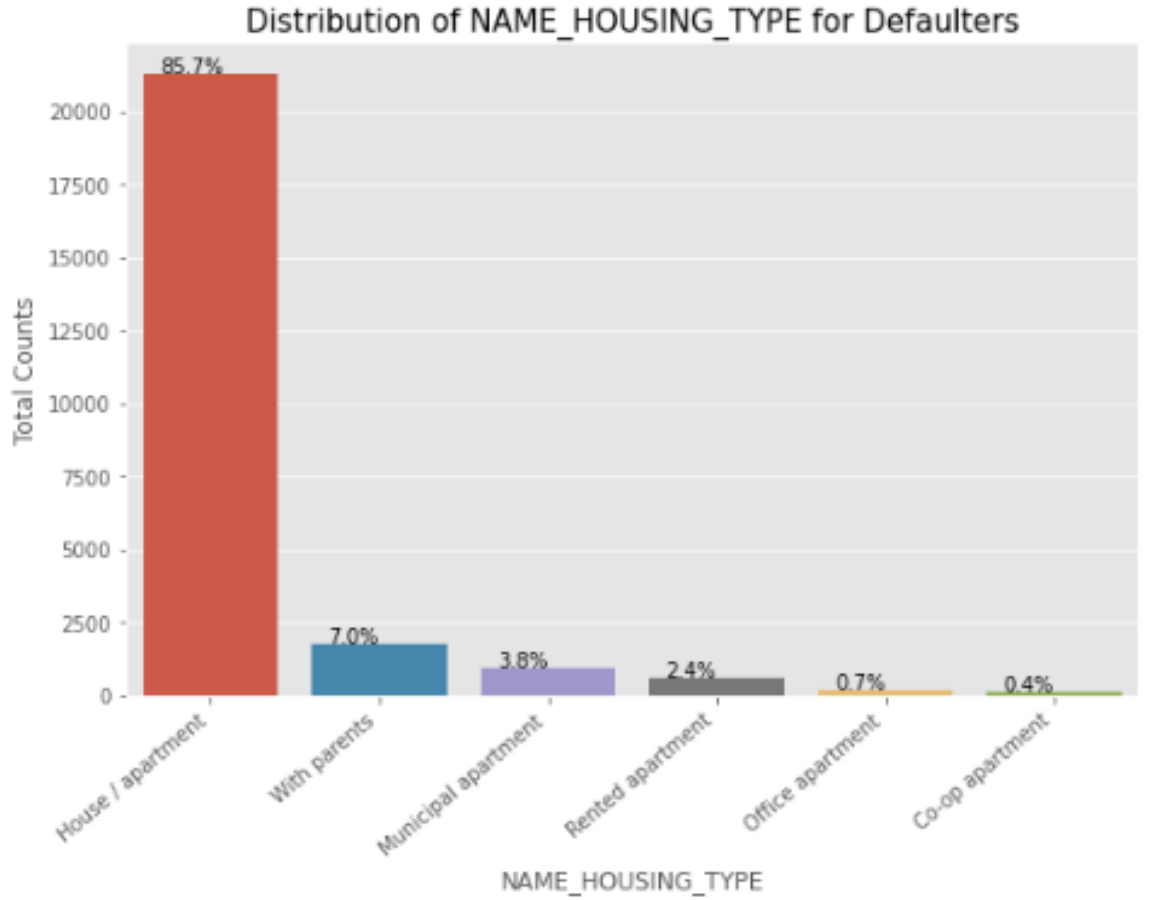


Married people tend to apply for more loans comparatively.
But from the graph we see that Single/non Married people contribute 14.5% to Non Defaulters and 18% to the defaulters. So there is more risk associated with them.

Univariate Analysis

Housing Type

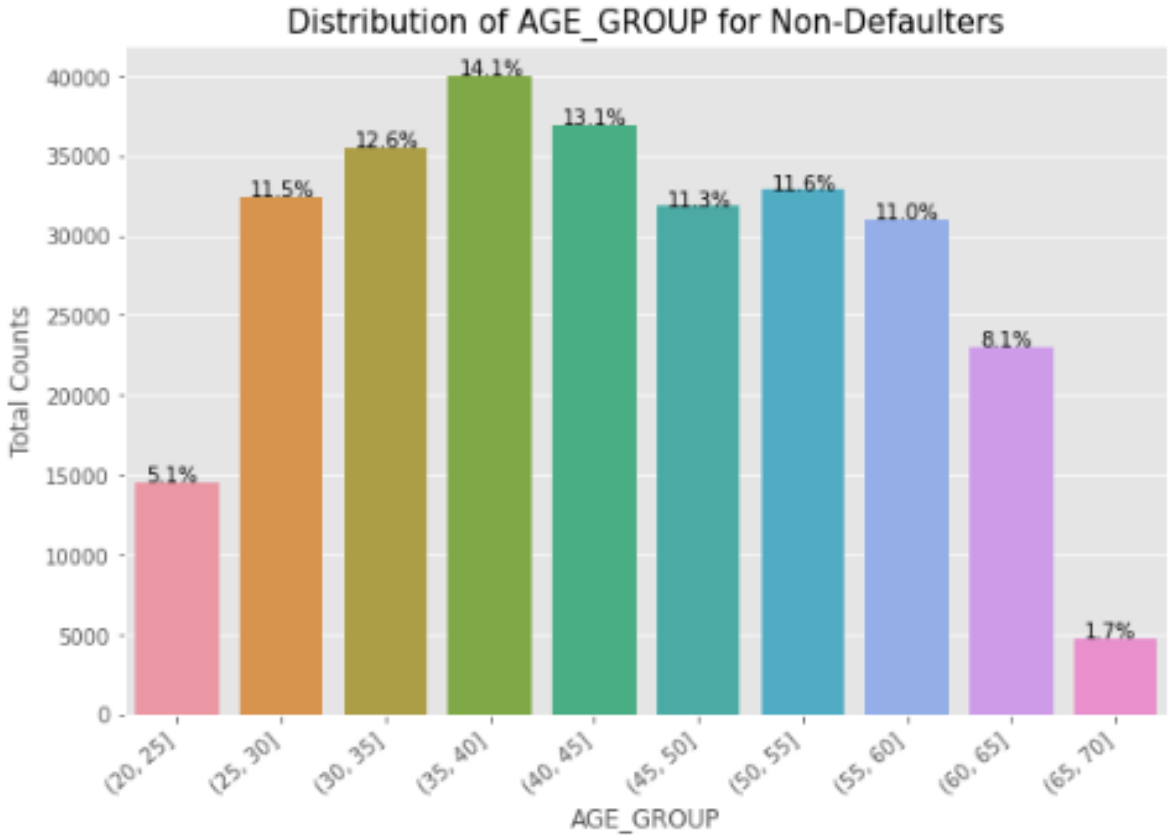


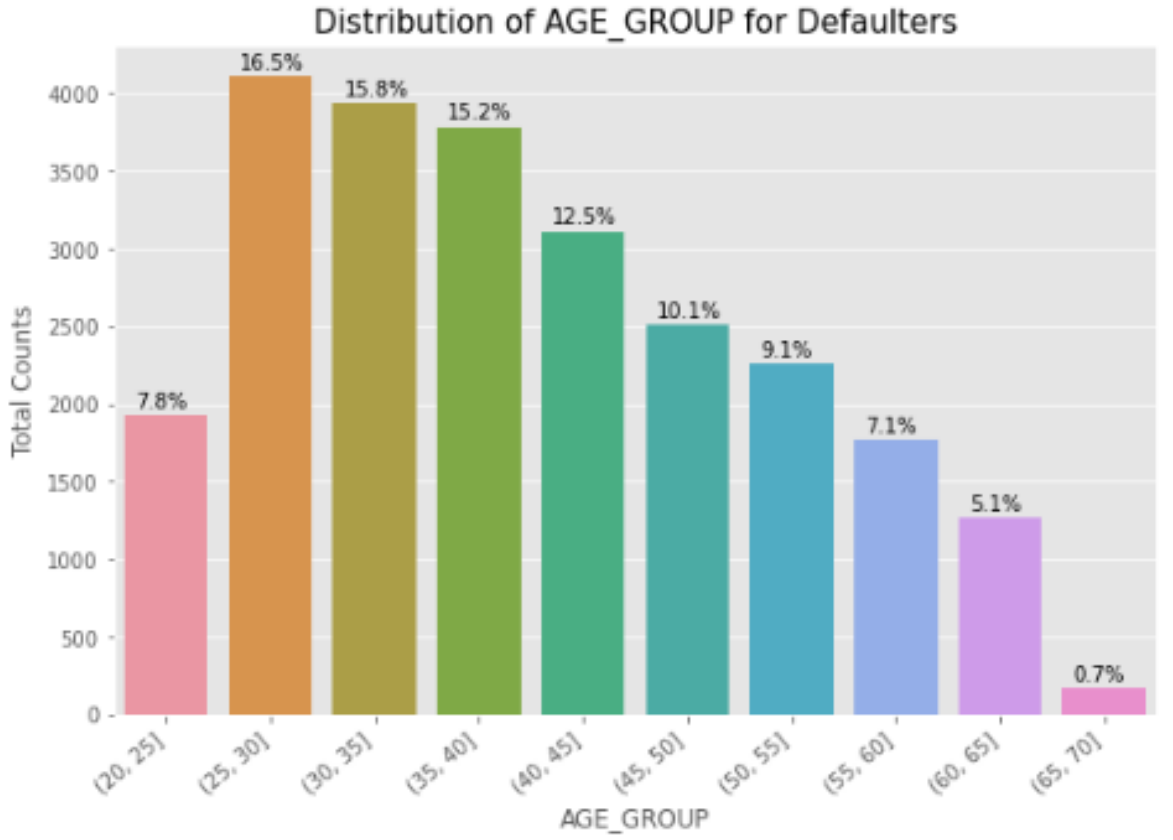


It is clear from the graph that people who have House / Apartment, tend to apply for more loans. People living with parents tend to default more often when compared with others. The reason could be their living expenses are more due to their parents living with them.

Univariate Analysis

Age Group

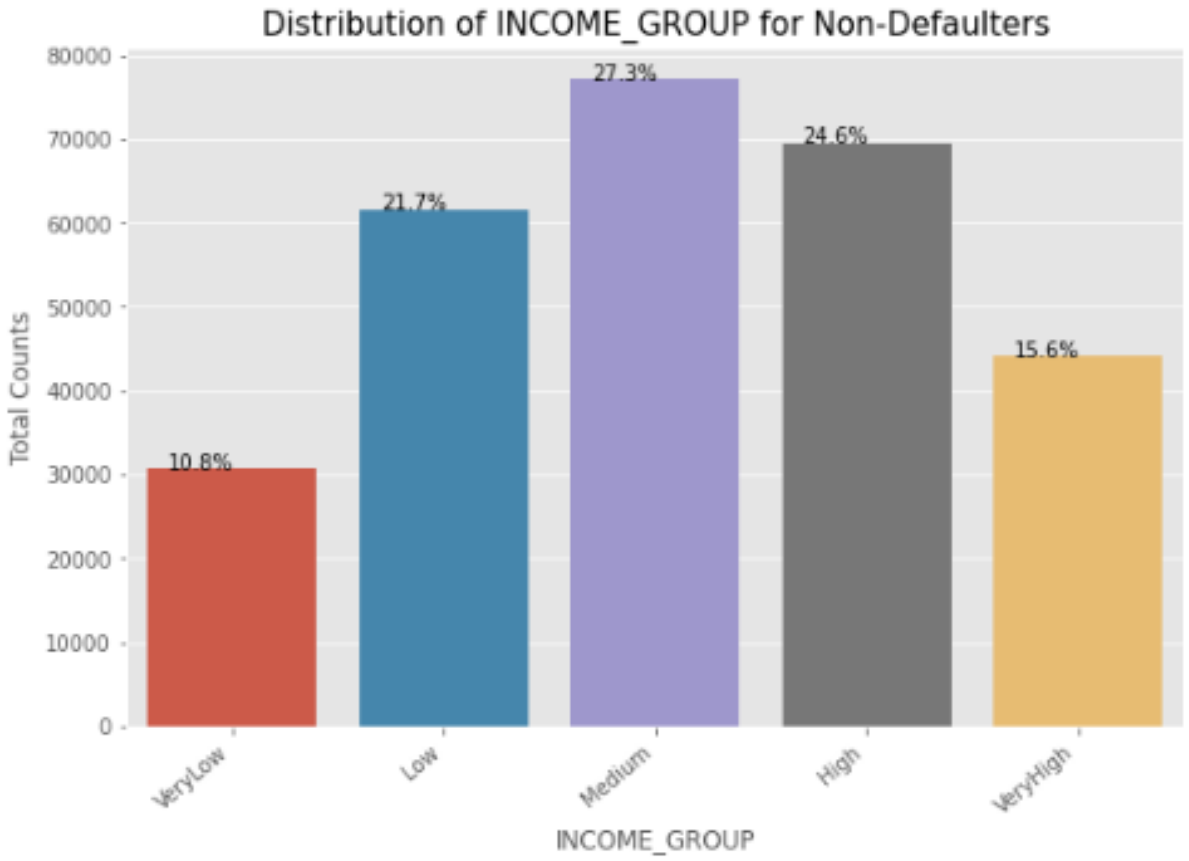


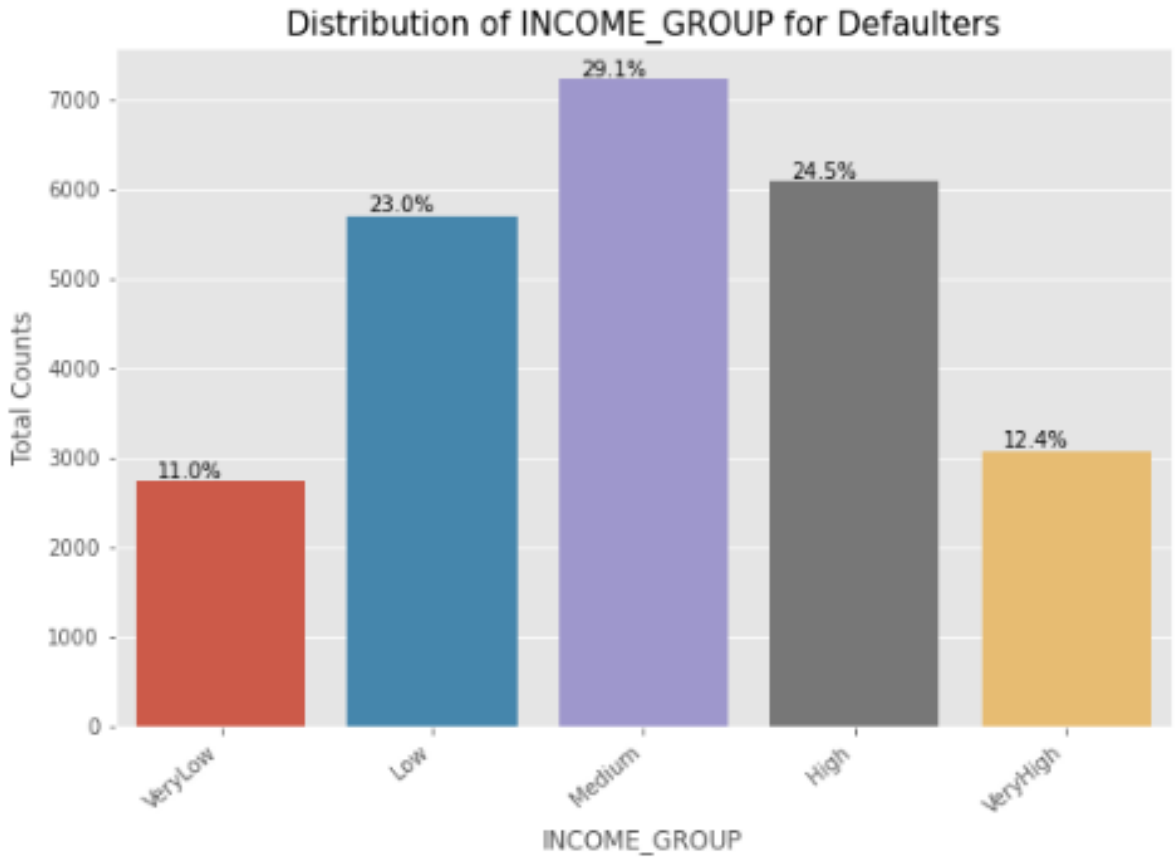


We see that (25,30] age group tend to default more often. So they are the riskiest people to loan to. With increasing age group, people tend to default less starting from the age 25. One of the reasons could be they get employed around that age and with increasing age, their salary also increases.

Univariate Analysis

Income Group

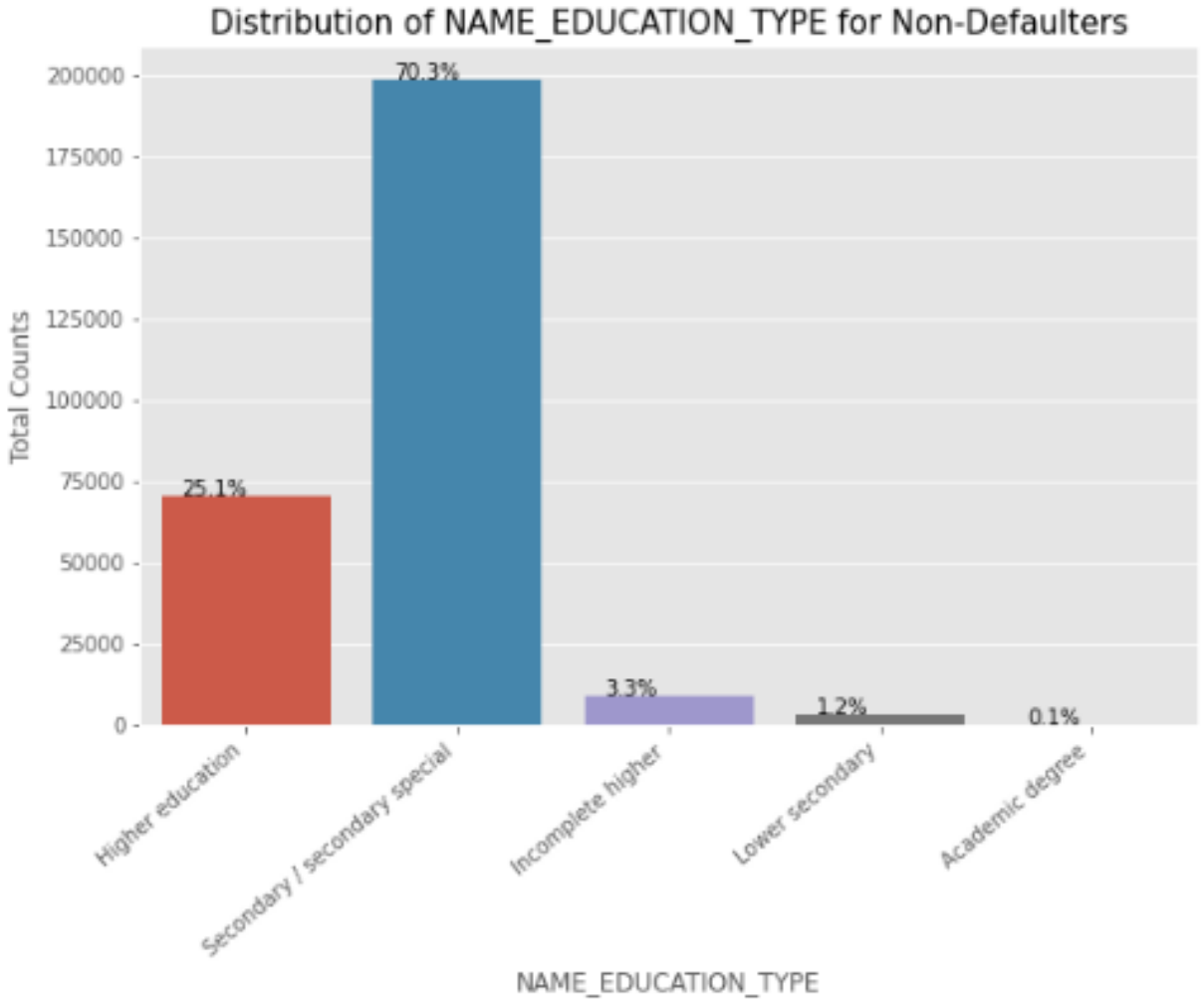


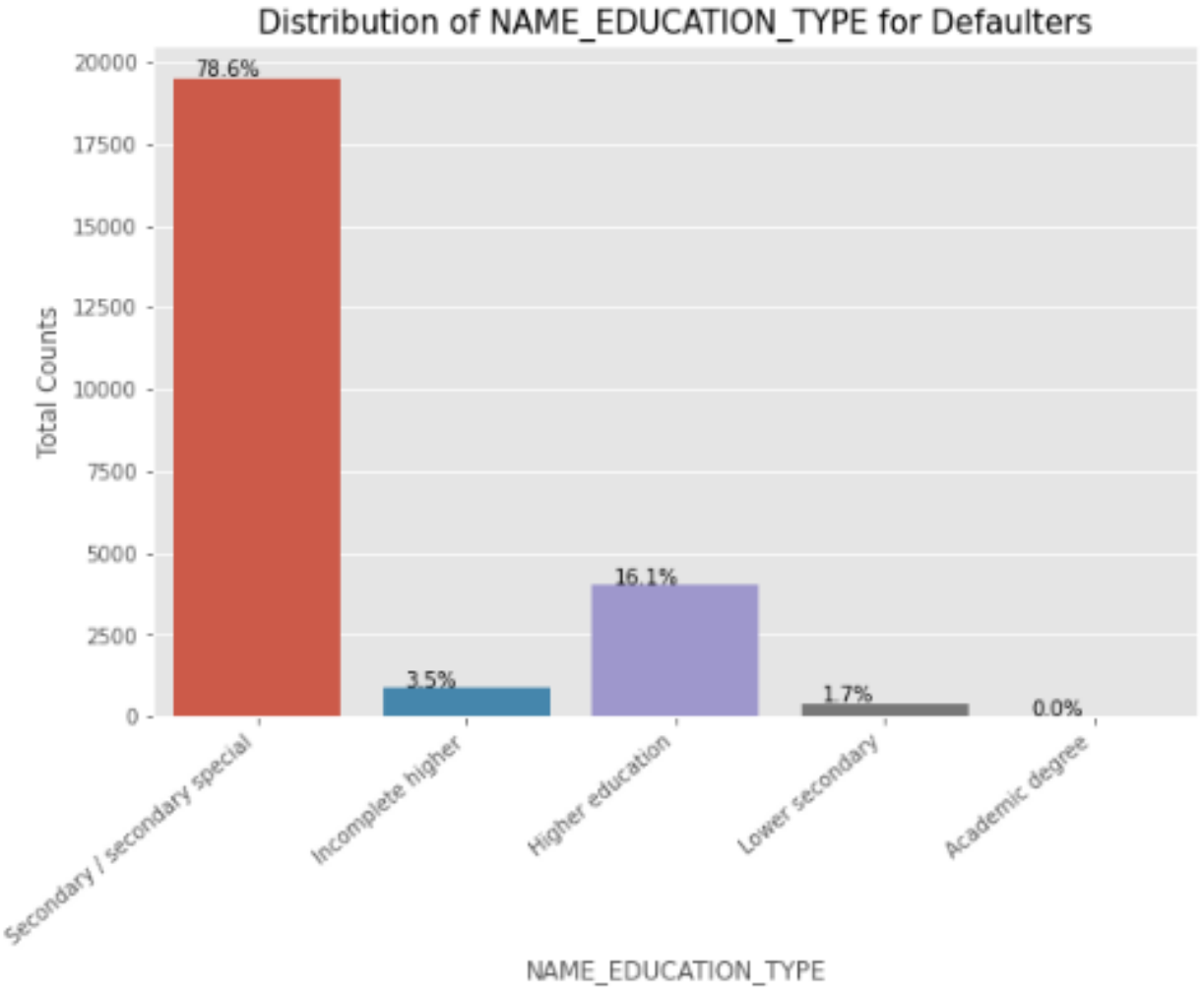


The Very High income group tend to default less often. They contribute 12.4% to the total number of defaulters, while they contribute 15.6% to the Non-Defaulters.

Univariate Analysis

Education

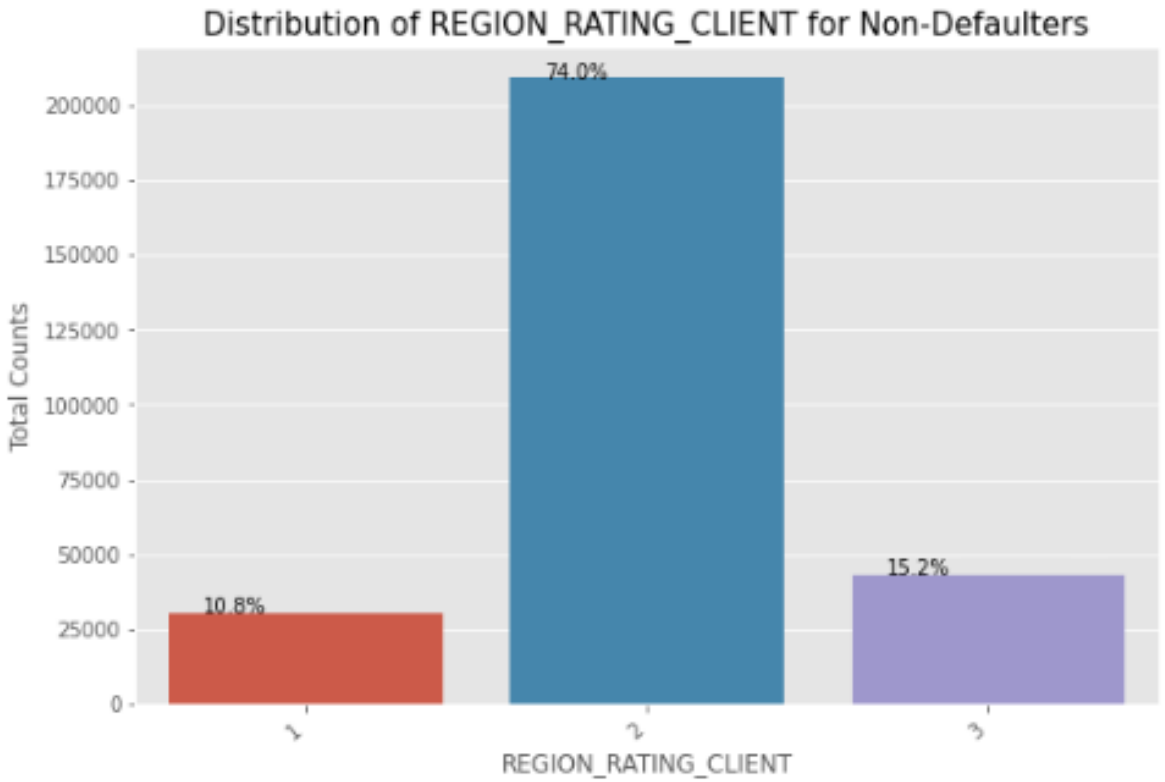


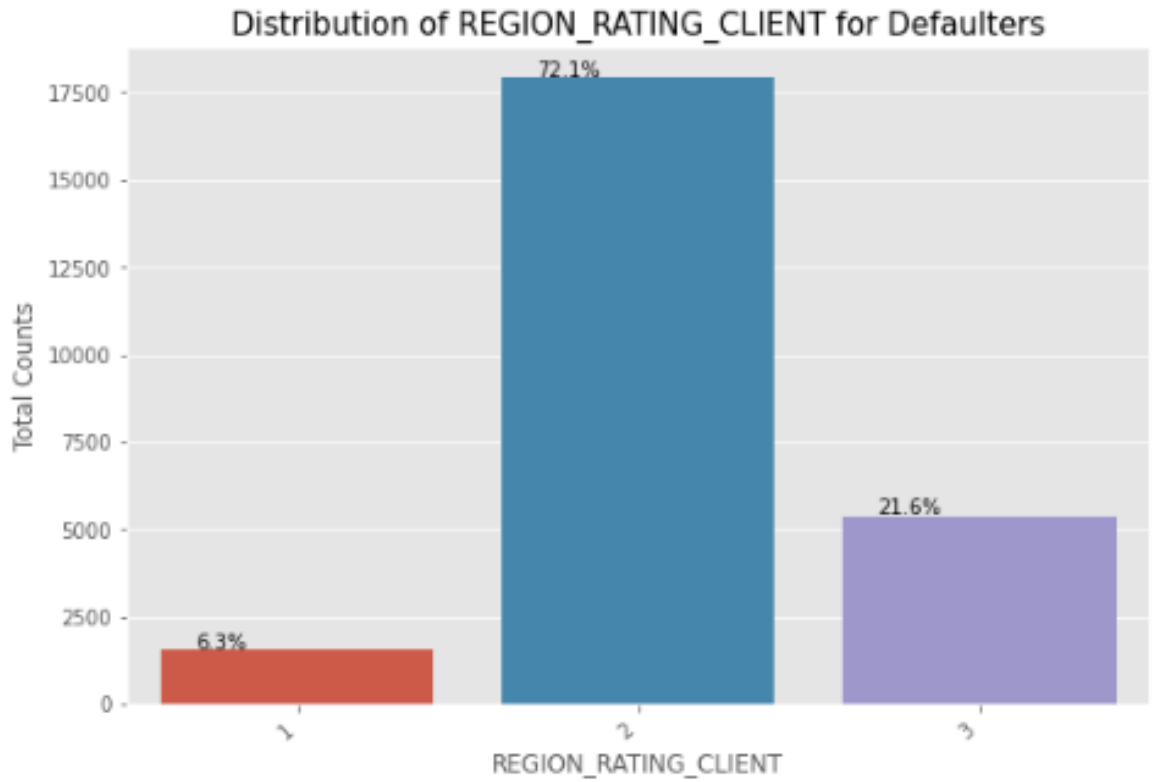


Almost all of the Education categories are equally likely to default except for the higher educated ones who are less likely to default and secondary educated people are more likely to default

Univariate Analysis

Rating Client

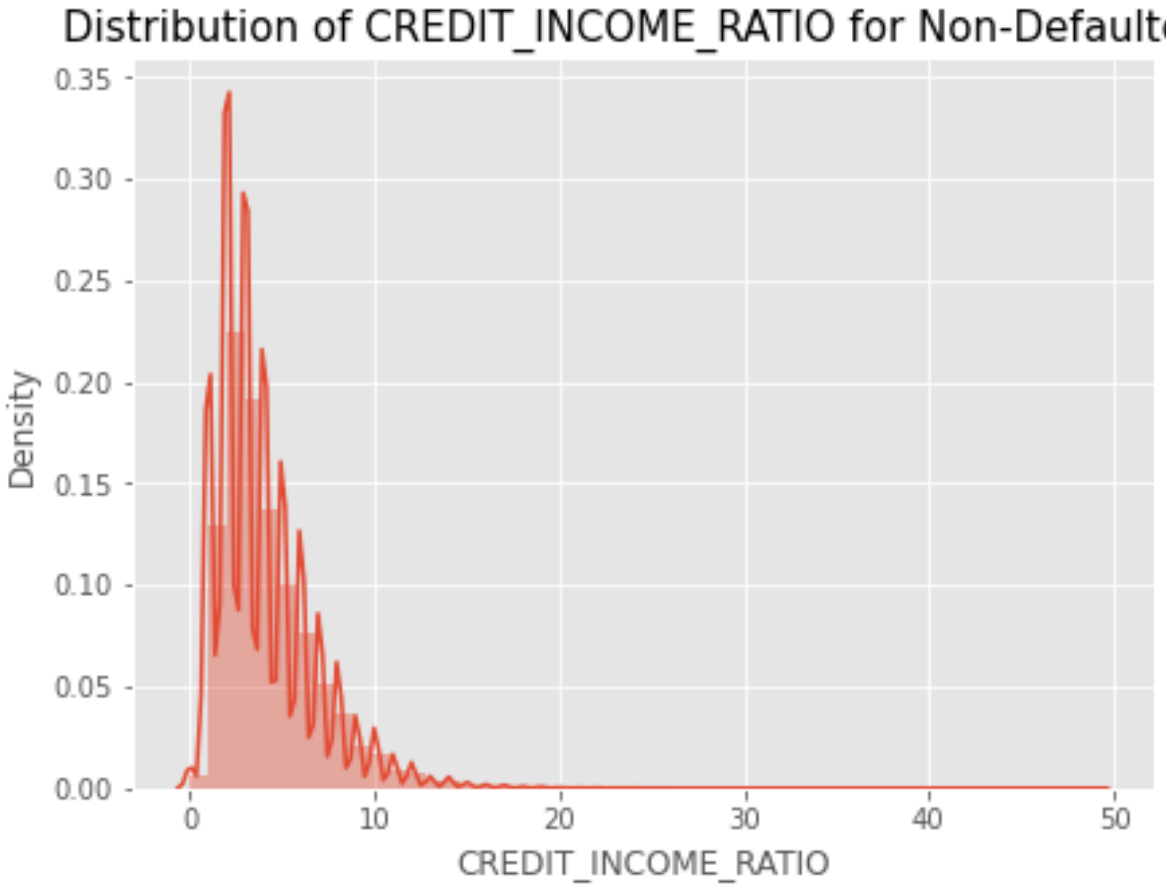


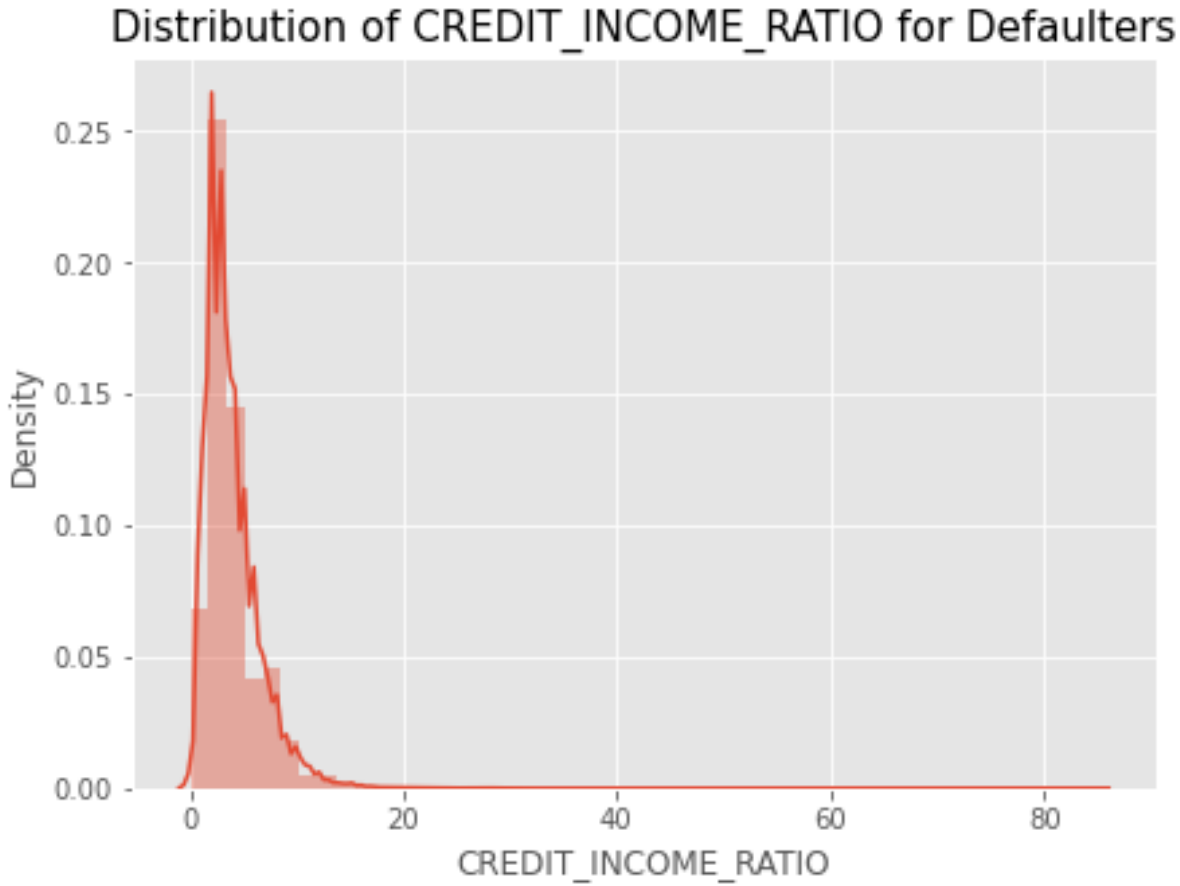


More people from second tier regions tend to apply for loans.
We can infer that people living in better areas(Rating 3) tend contribute more to the defaulters by their weightage.
People living in 1 rated areas

Univariate continuous variable Analysis

Credit Income Ratio

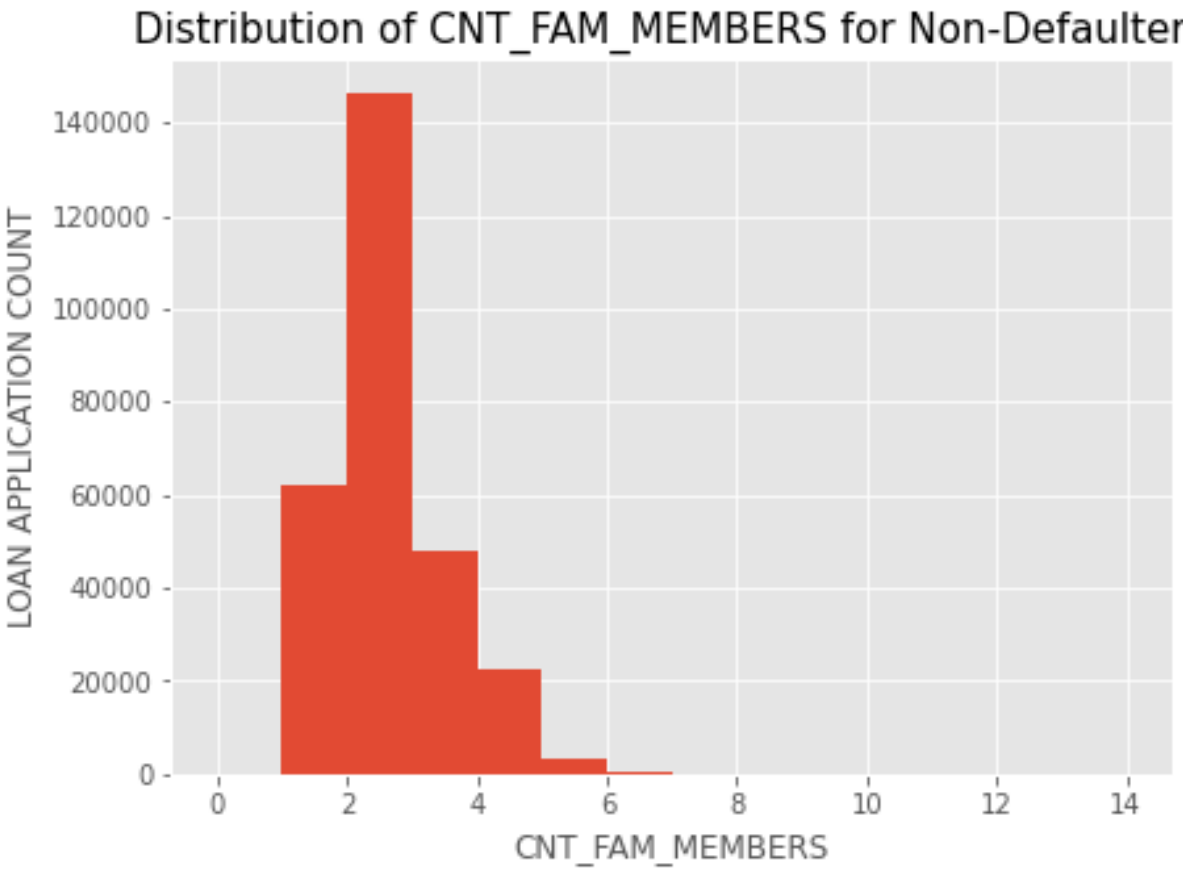


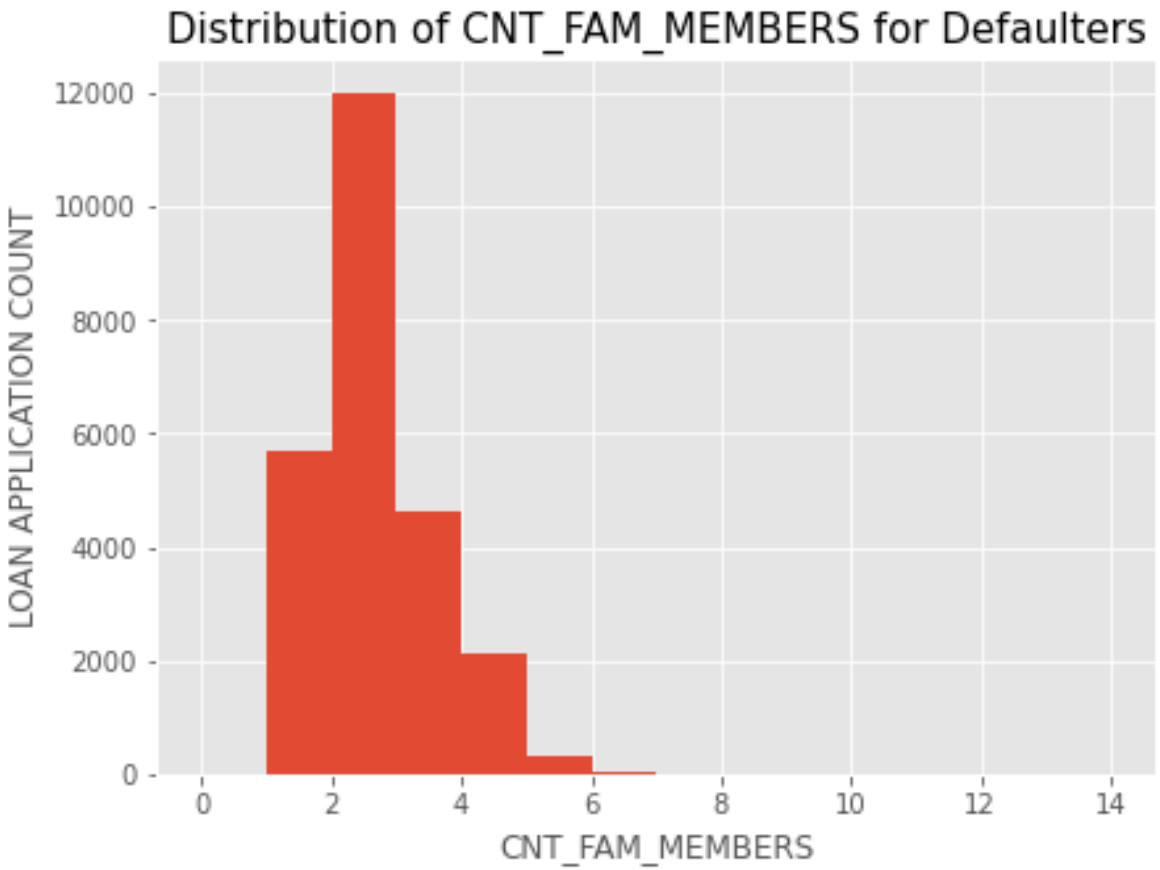


There doesn't seem to be a clear distinguish between the group which defaulted vs the group which didn't when compared using the ratio, we can see that when the CREDIT_INCOME_RATIO is more than 50, people default.

Univariate continuous

Family Members

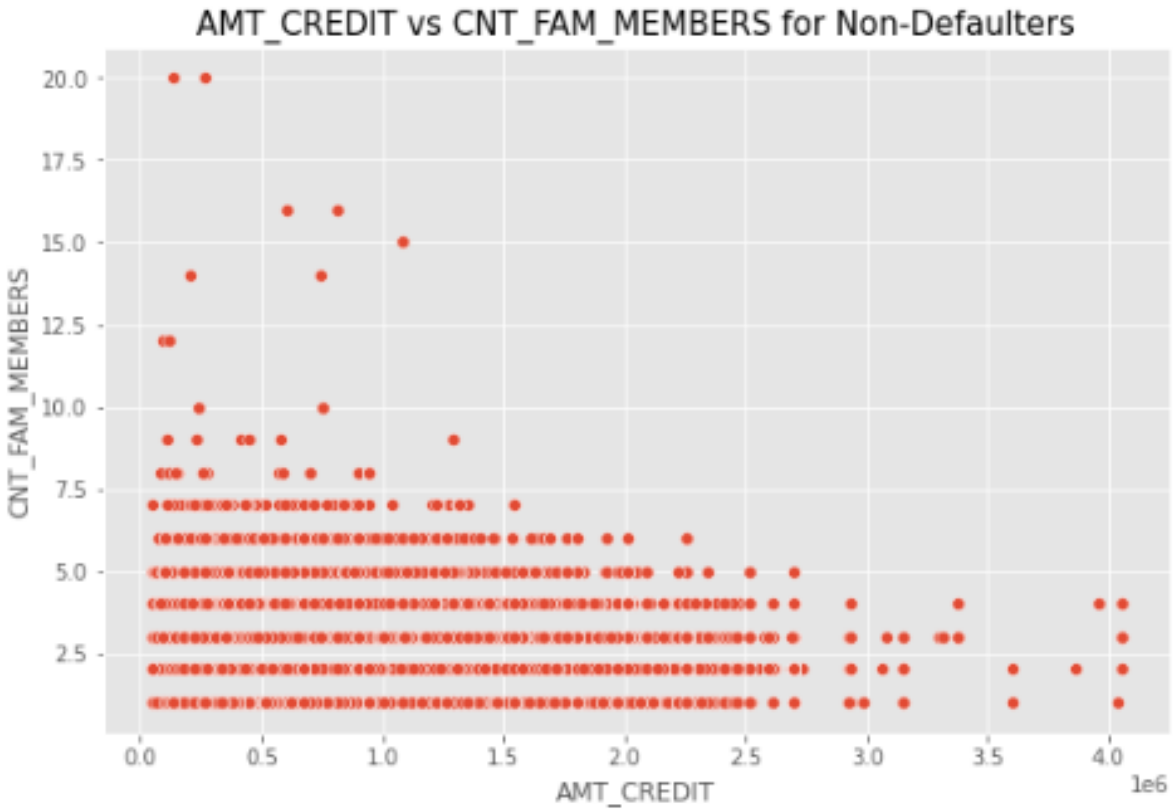


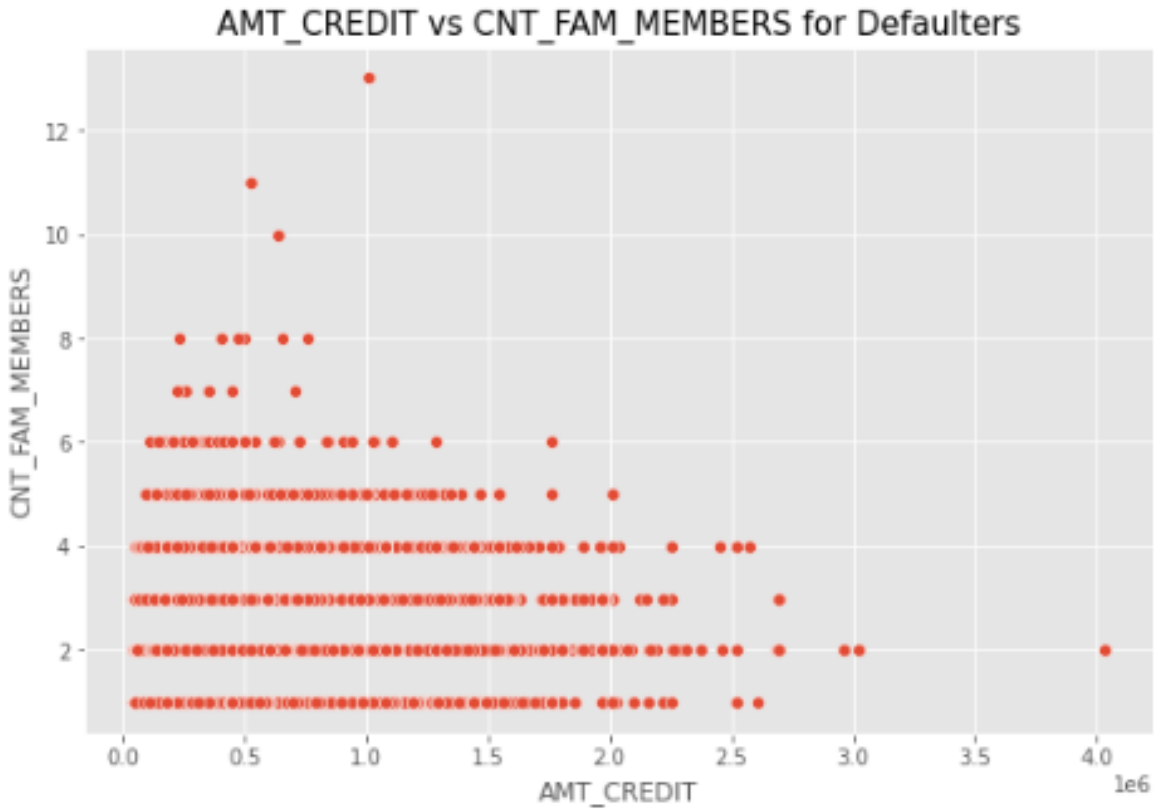


From the graph we can clearly see that a family of 3 applies loan more often than the other families

Bivariate Analysis

AMT Credit vs CNT Family Members

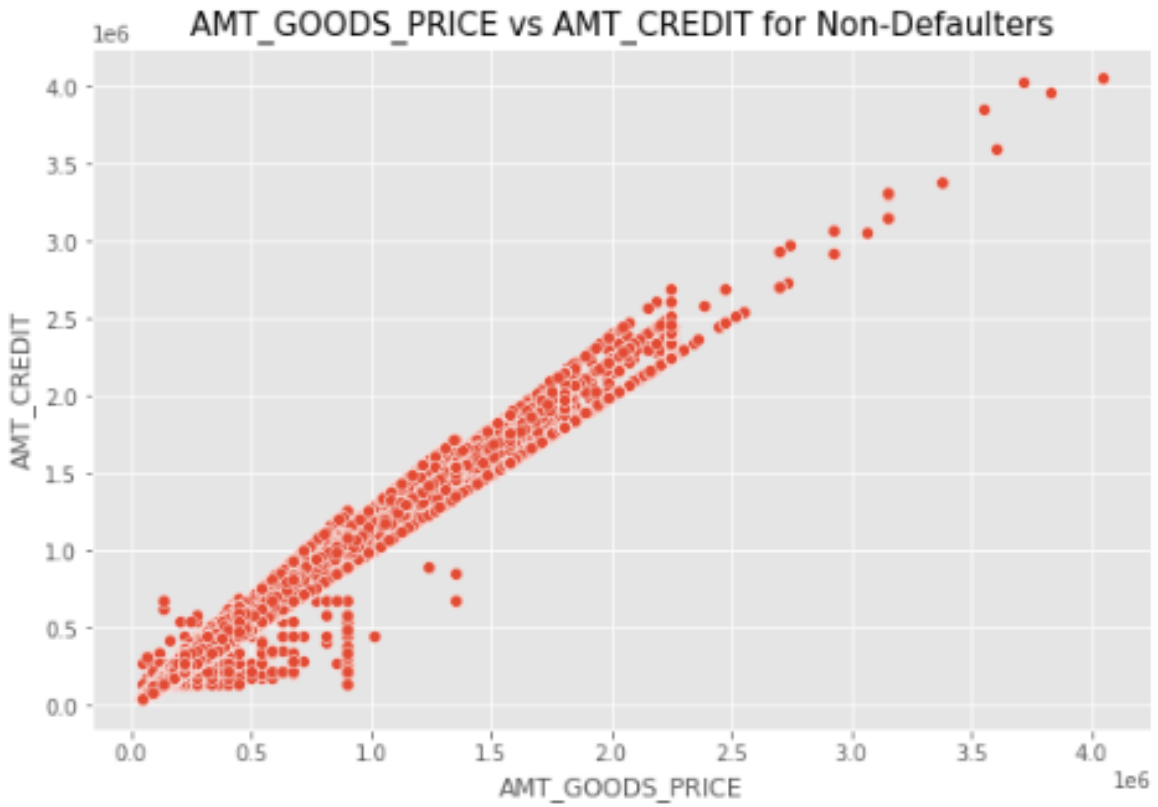


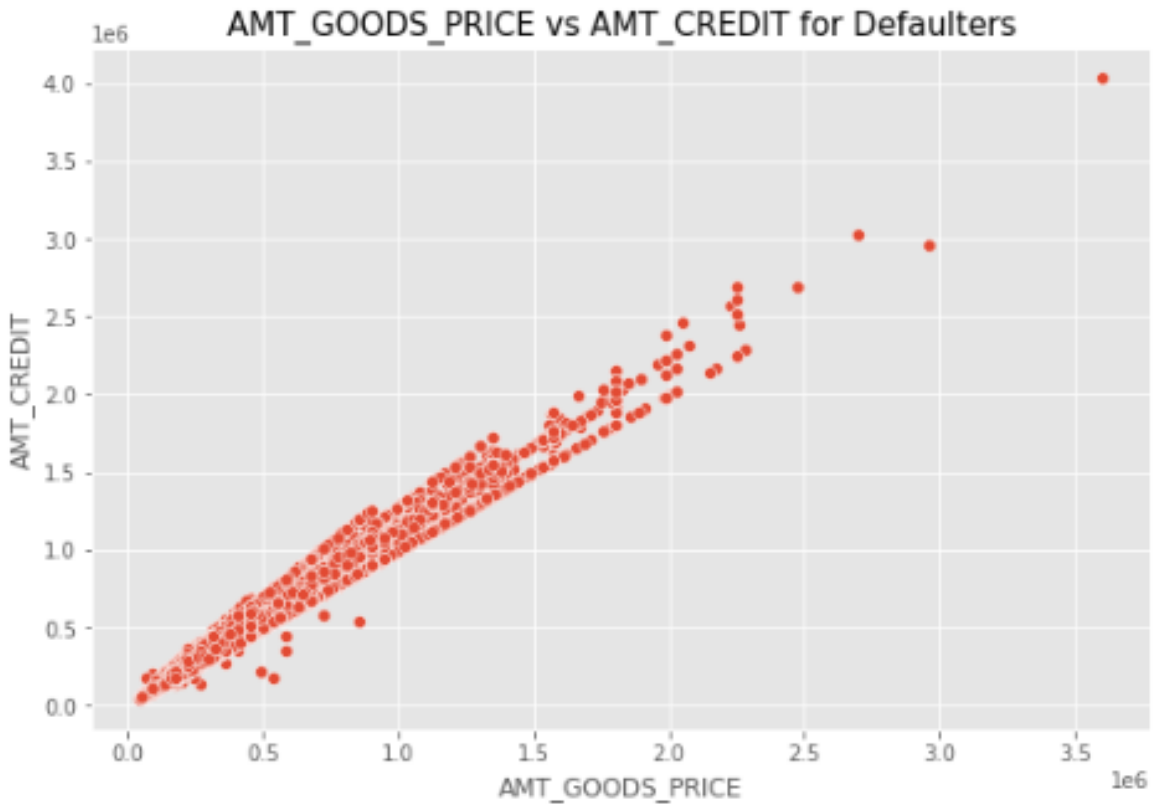


We can see that the density in the lower left corner is similar in both the case, so the people are equally likely to default if the family is small and the AMT_CREDIT is low. We can observe that larger families and people with larger AMT_CREDIT default less often

Bivariate Analysis

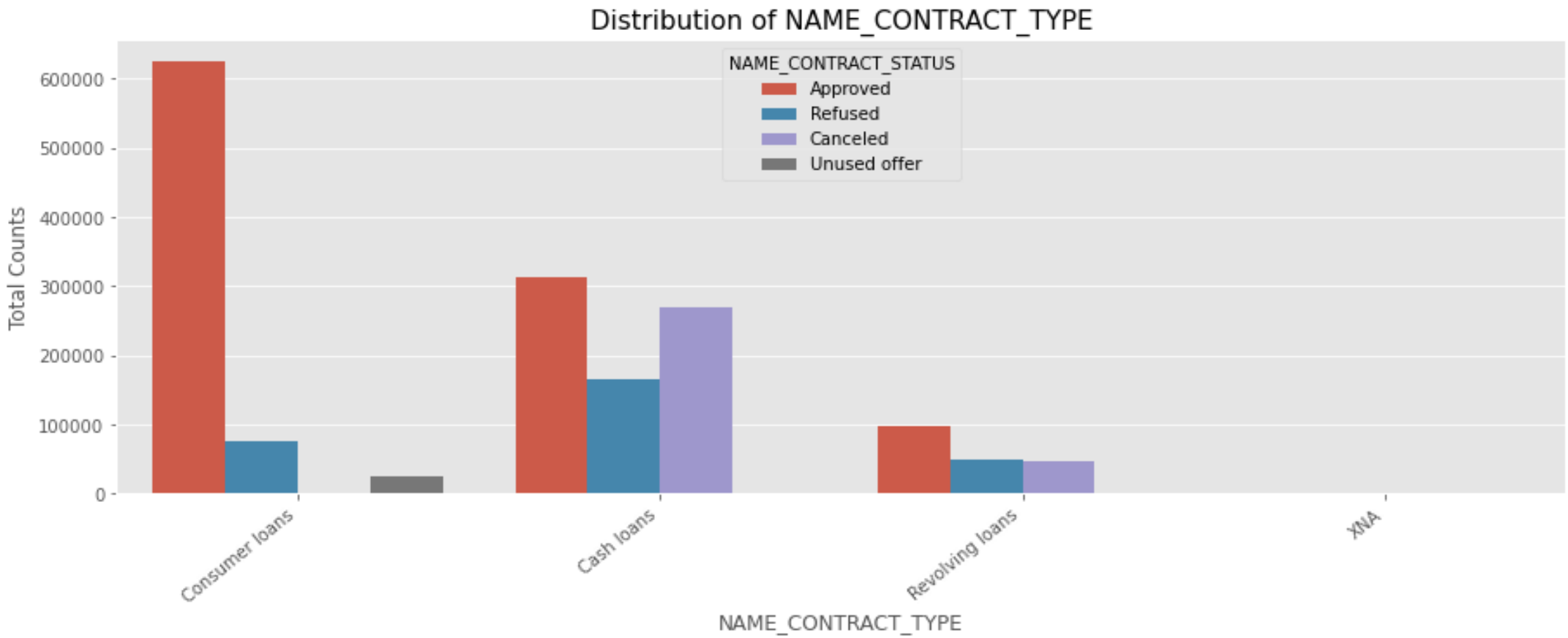
AMT Goods Price vs AMT Credit





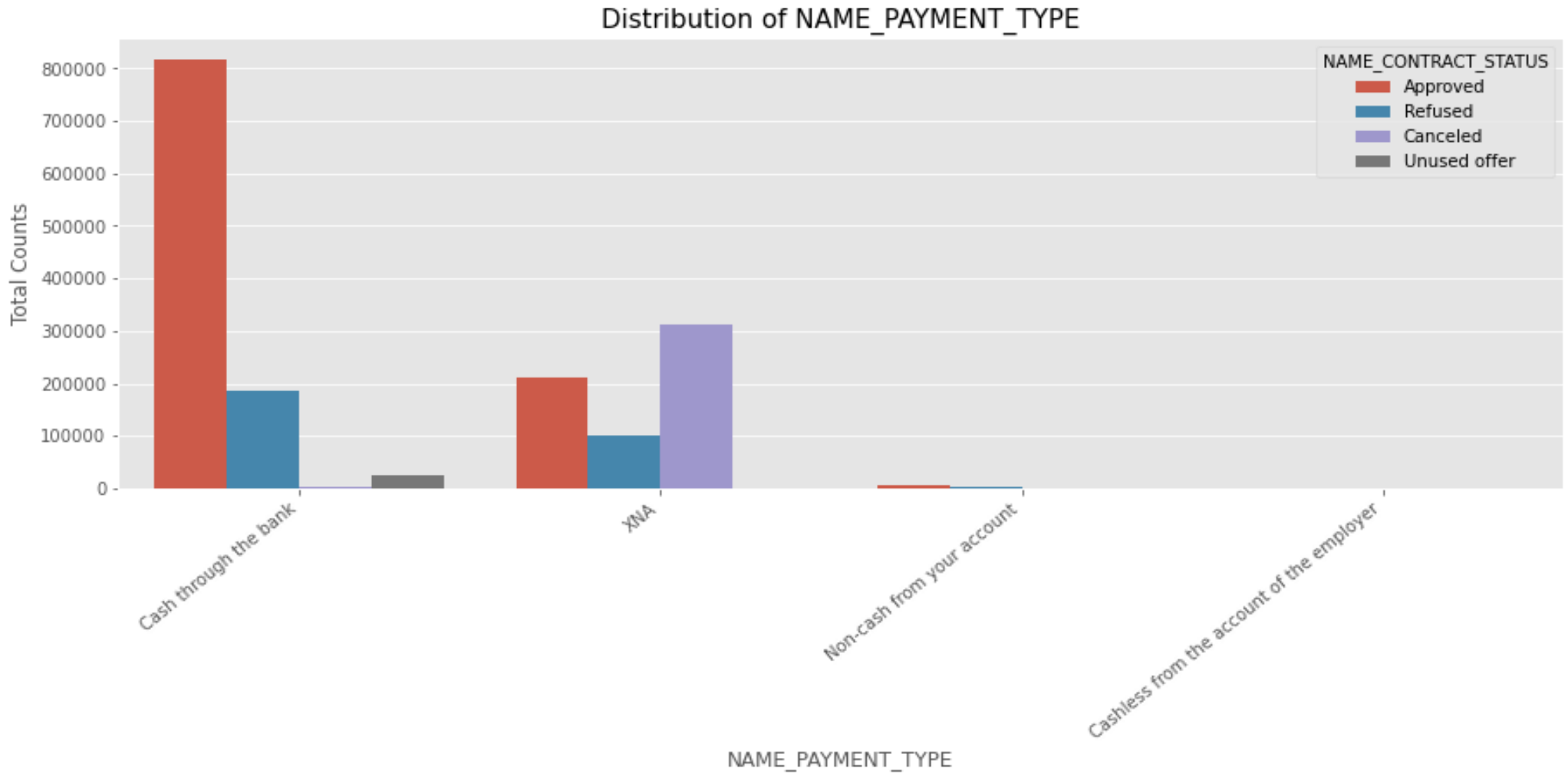
We can see that the density in the lower left corner is more in non- defaulters as compaired with defaulters.

Univariate Analysis - Previous Application Data



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.

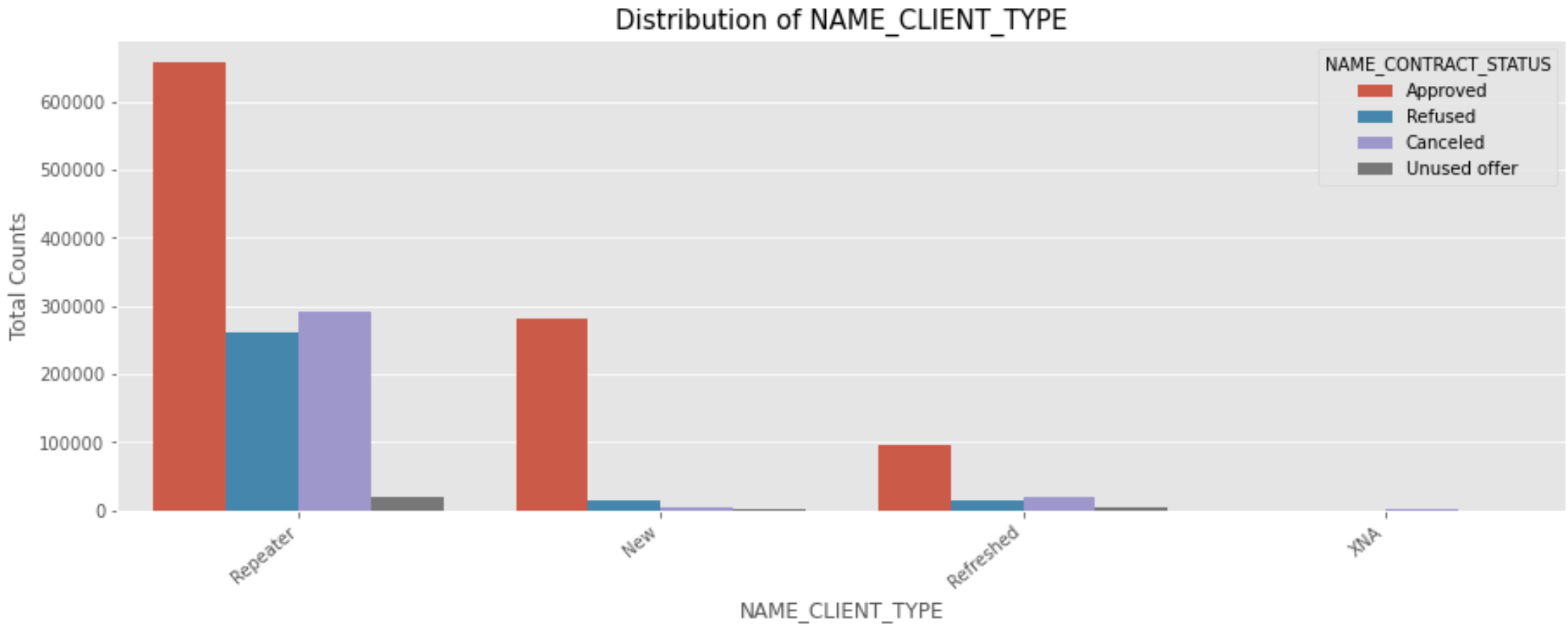
Univariate Analysis - Previous Application Data



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

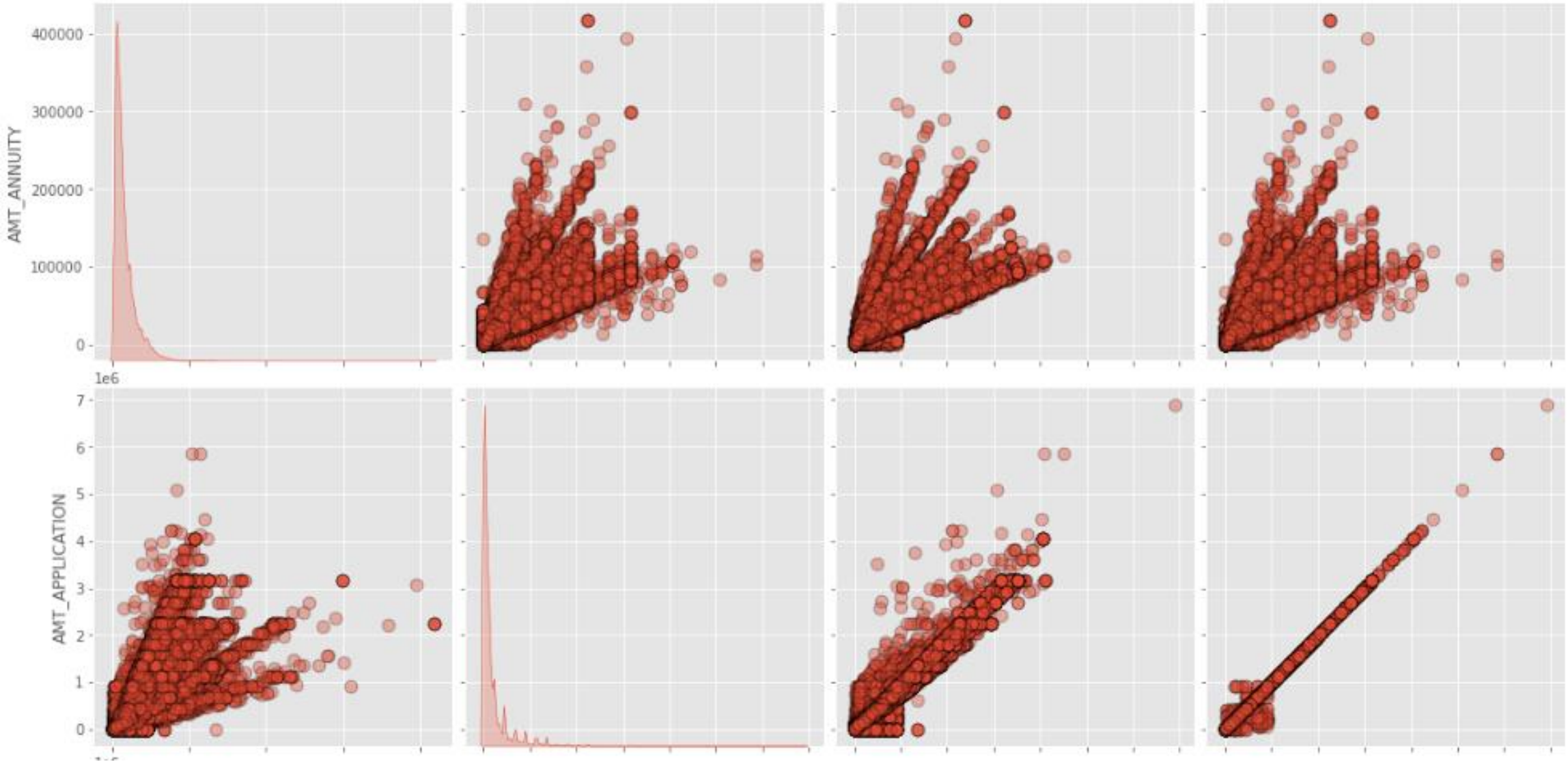
We can also see that 'Non-Cash from your account' & 'Cashless from the account of the employee' options are not at all popular in terms of loan repayment amongst the customers.

Univariate Analysis - Previous Application Data



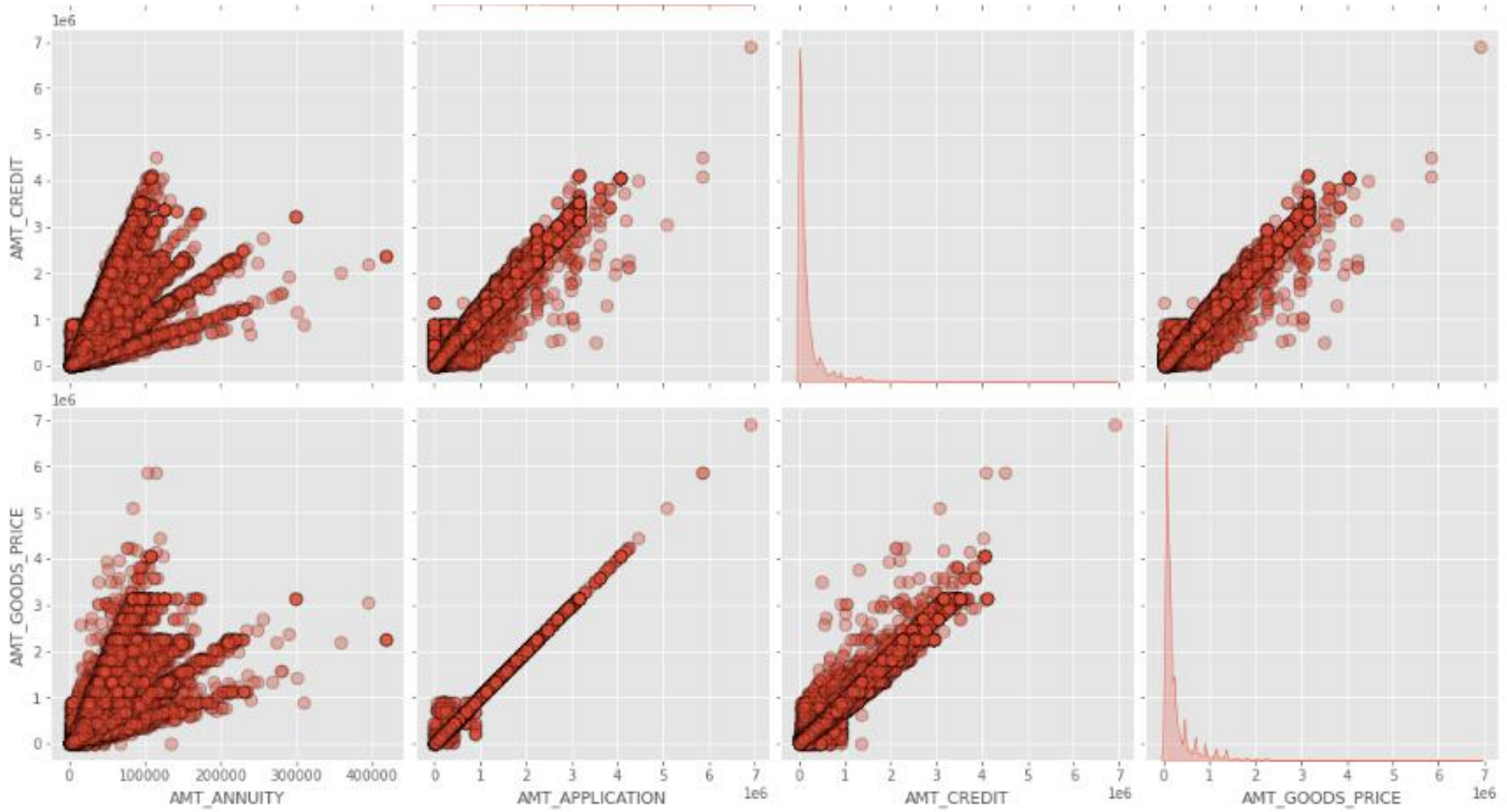
Most of the loan applications are from repeat customers, out of the total applications 70% of customers are repeaters. They also get refused most often.

Pairplot - Bivariate analysis on numerical columns



Continue……

Pairplot - Bivariate analysis on numerical columns



Continue……

Pairplot - Bivariate analysis on numerical columns

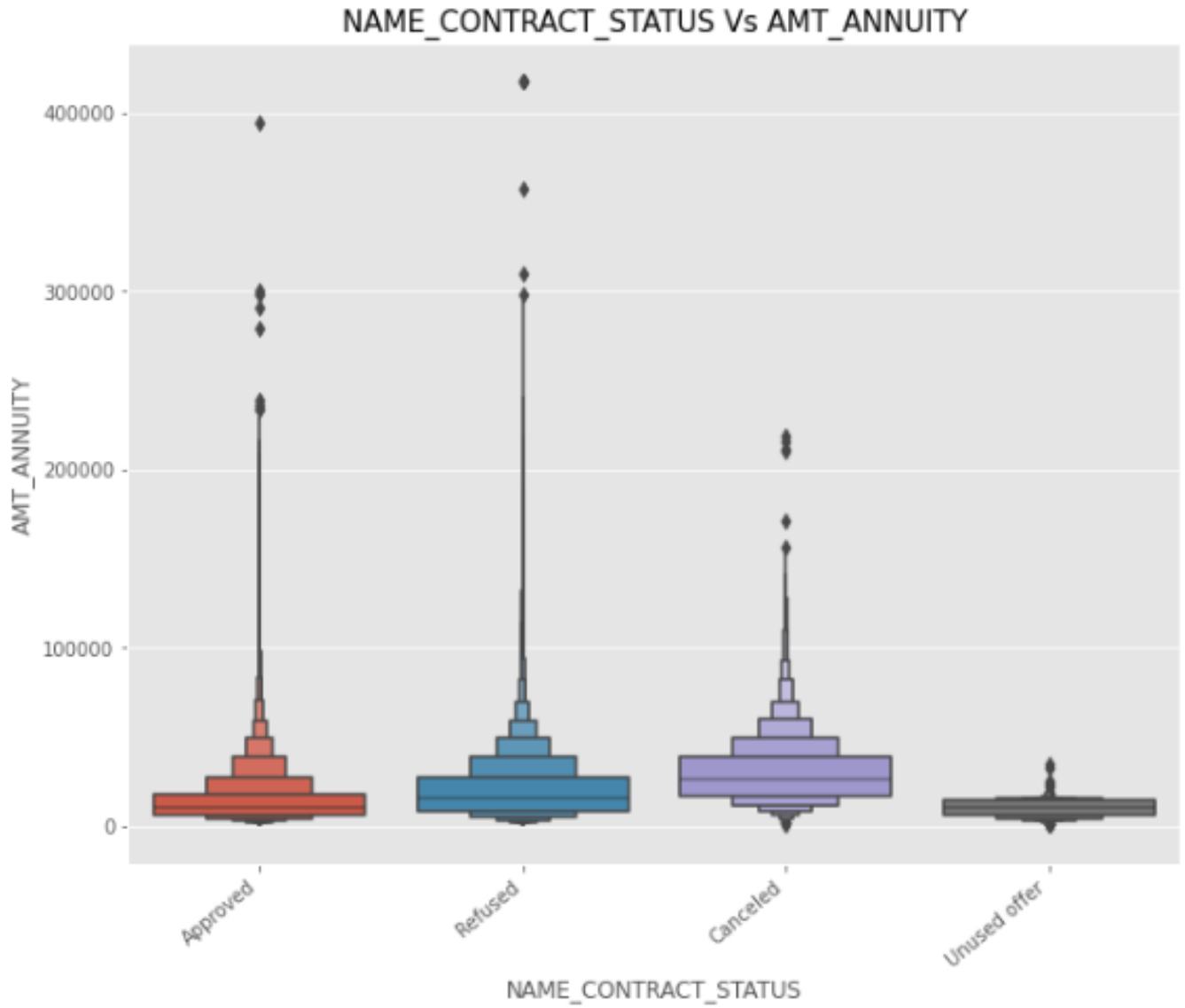
Annuity of previous application has a very high and positive influence over: (Increase of annuity increases below factors)
(1) How much credit did client asked on the previous application
(2)Final credit amount on the previous application that was approved by the bank
(3) Goods price of good that client asked for on the previous application.

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

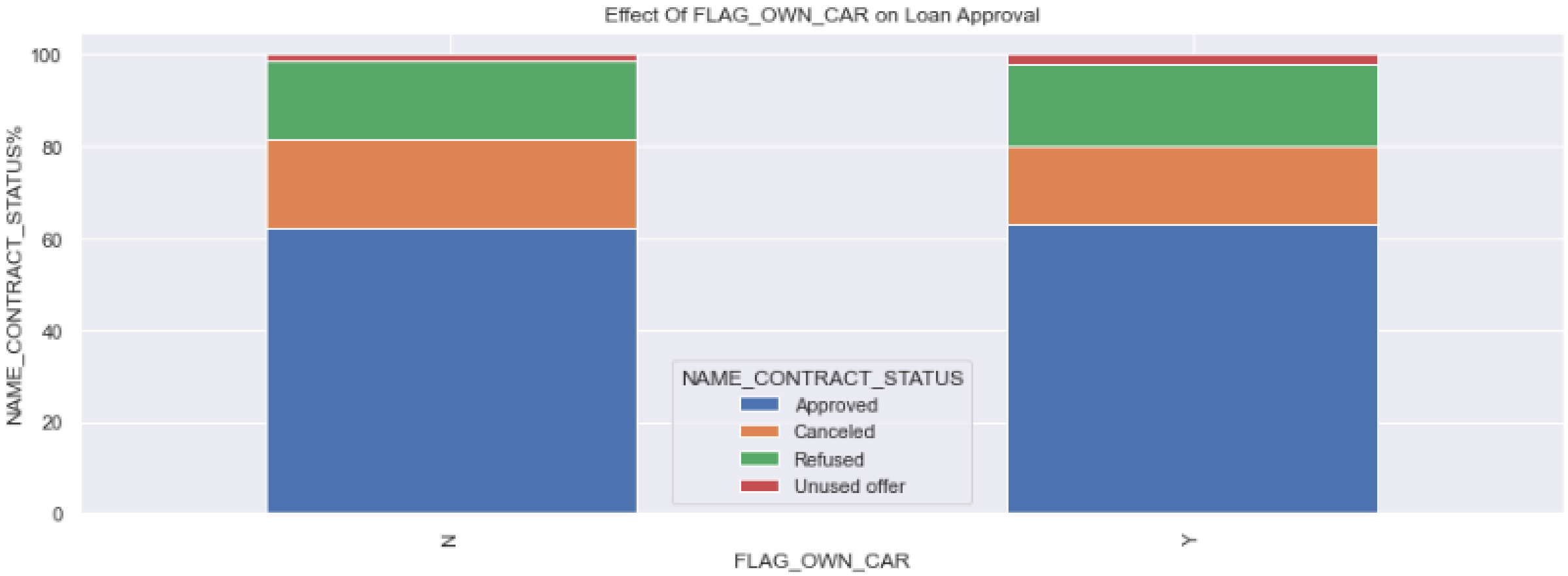
Final credit amount disbursed to the customer previously, after approval is highly influence by the application amount and also the goods price of good that client asked for on the previous application.

Box Plot - Bivariate Analysis on categorical vs numeric columns

From the above plot we can see that loan application for people with lower AMT_ANNUITY gets canceled or Unused most of the time.
We also see that applications with too high AMT ANNUITY also got refused more often than others.

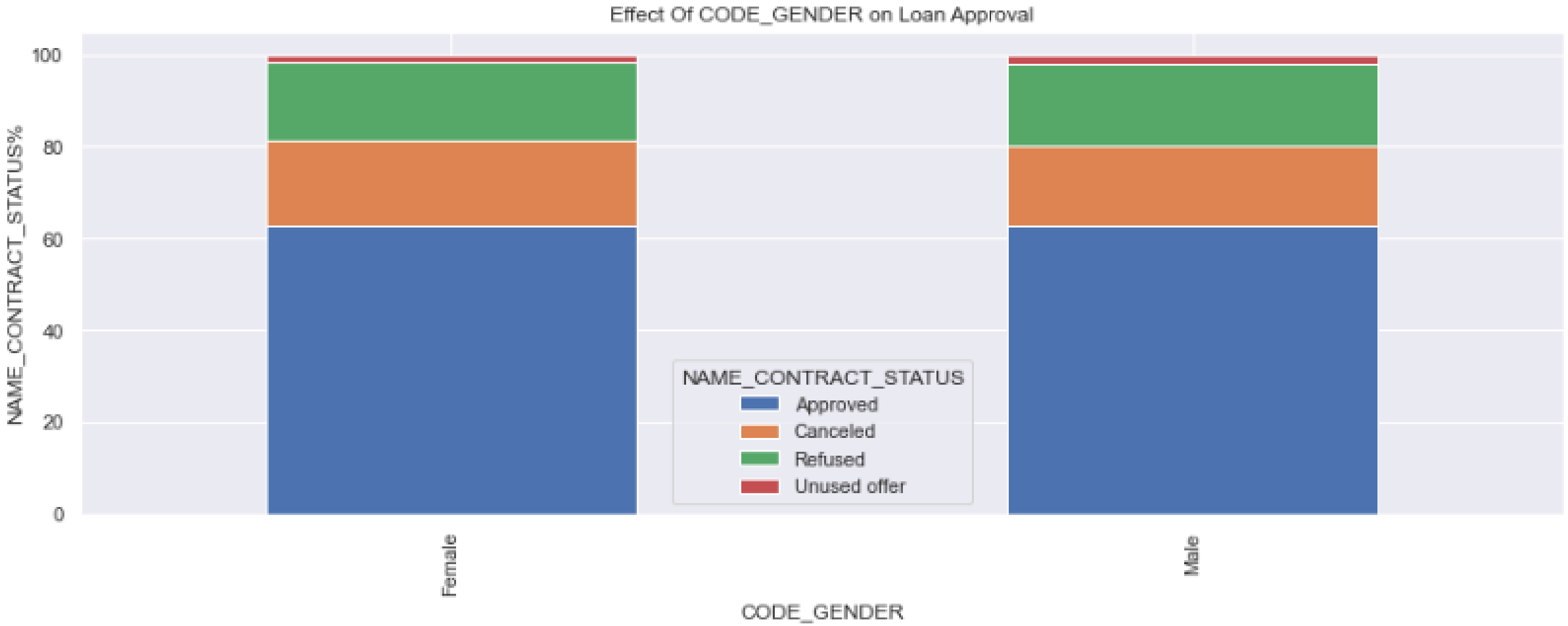


Merging the files and analyzing the data



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

Merging the files and analyzing the data



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.

Merging the files and analyzing the data

Target variable (0 - Non Defaulter 1 - Defaulter)

We can see that the people who were approved for a loan earlier, defaulted less often where as people who were refused a loan earlier have higher chances of defaulting.

Conclusion

- Banks should focus more on ‘Student’ and ‘Businessman’ for successful payments.
- Banks should focus less on income type ‘Working’ as they are having most number of unsuccessful payments.
- Banks should focus more on the Very High income group as it tend to default less often. They contribute 12.4% to the total number of defaulters, while they contribute 15.6% to the Non-Defaulters.
- Age group 25, 30 tend to default more often. So they are the riskiest people to loan.

Thank You