Presentation
On

CREDIT EDA CASE STUDY

By: Snehal Kadlag

Business Objectives -

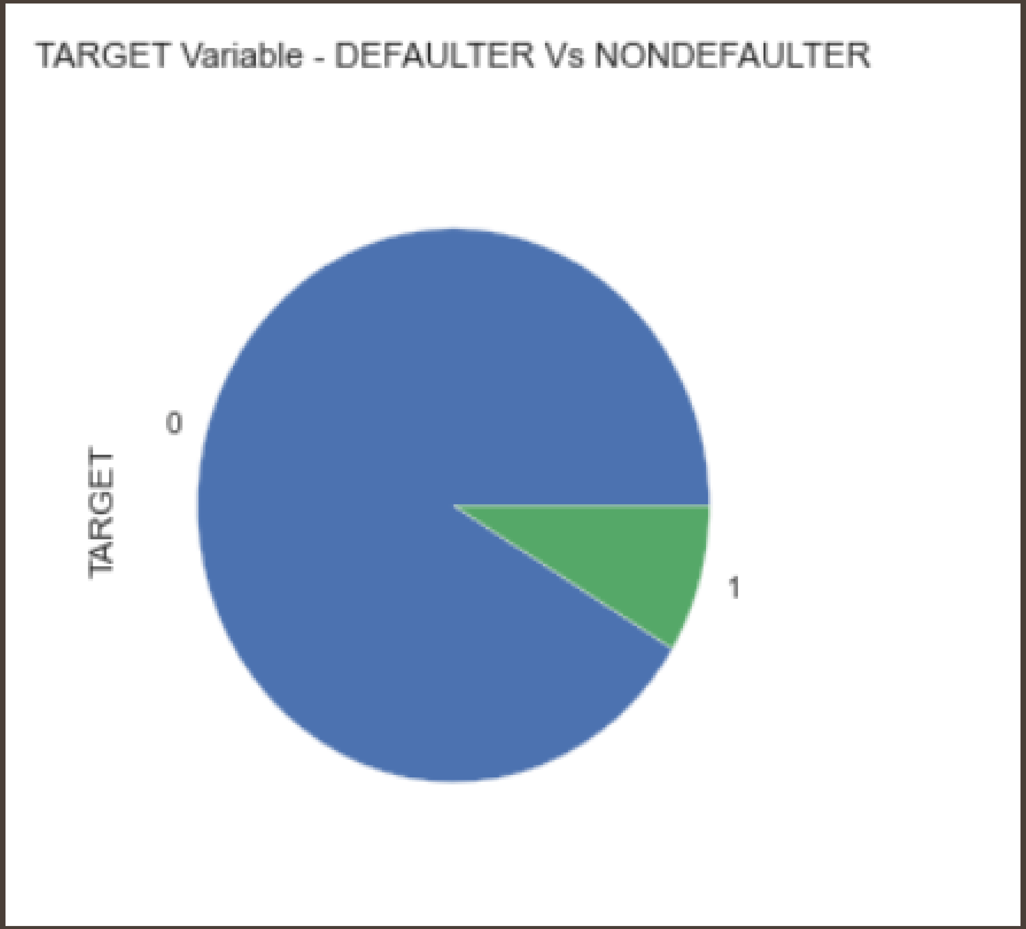
- This case study aims to identify patterns which indicate if a client has difficulty paying their installments. Which will help the company to make decisions for loan approval based on the applicant’s profile. Which will further help control financial and Business loss for the company.

Steps taken:

- 1. Data understanding and sourcing.
- 2. Checking for data Quality issues like missing data, imputing data.
- 3. Checking for data imbalance and segmenting data for target as 0 for non-Defaulters and 1 for Defaulters.
- 4. Doing univariate ,bivariate analysis and finding corelation.
- 5. Merging of application data with previous application data.
- 6.Data analysis by univariate, segmented univariate, bivariate analysis and correlation.
- 7. Recommendation and Risks.

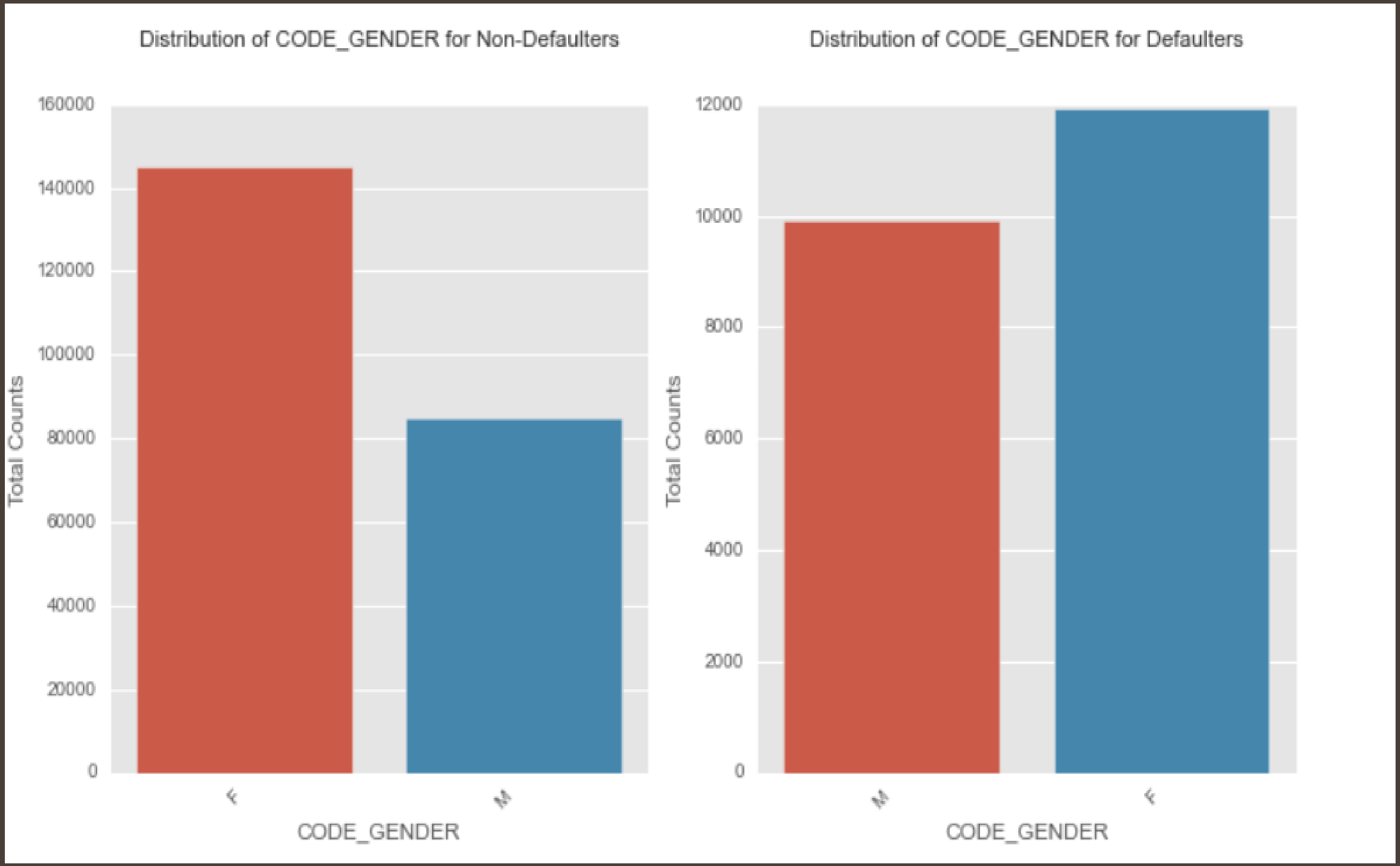
Indicating:
Target imbalance

inference clearly show's imbalance between people who defaulted and who didn't default



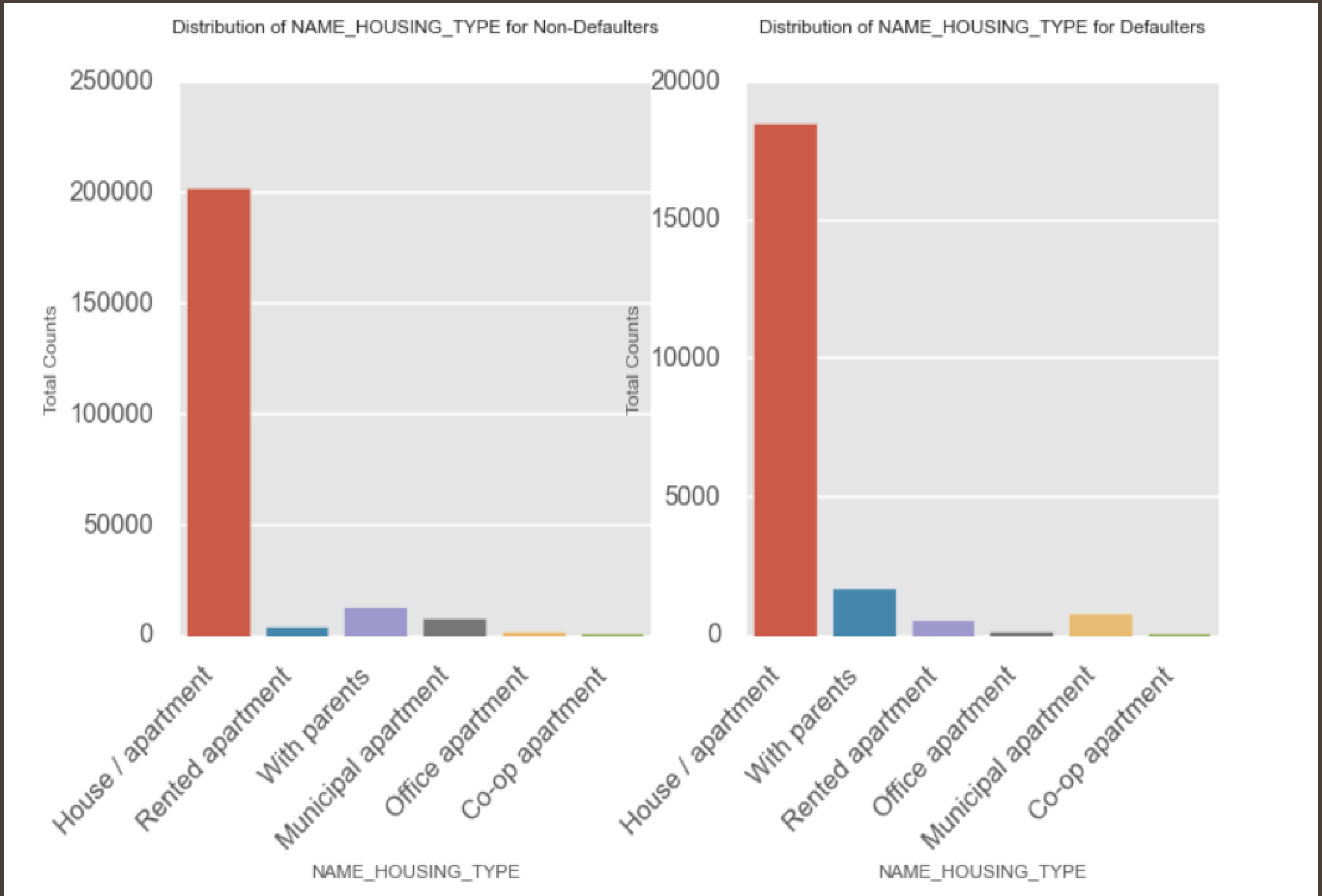
Univariate Categorical Ordered Analysis

We can see that Female's contribute to the non-defaulters higher than male's but at the same time female's also are more in defaulters. We can conclude that We see more female's applying for loans than male's and hence the more number of female's defaulters as well. But the rate of defaulting of FEMALE is much lower compared to their MALE counterparts as per loan application. Hence it is safe to infer female's could be potential customers



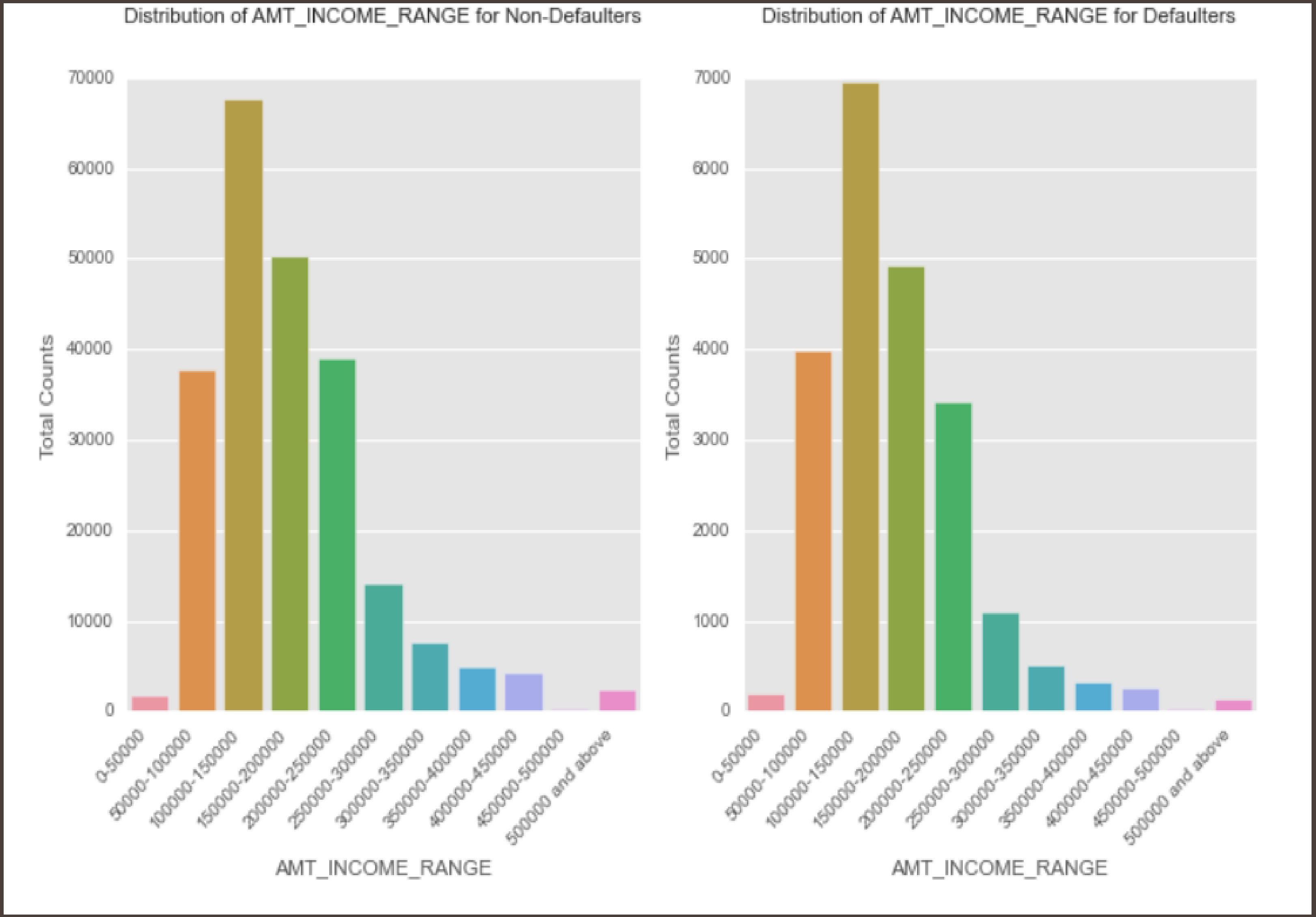
Distribution of Name house type

It is clear from the plot that people who have House/ Appartement, 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.



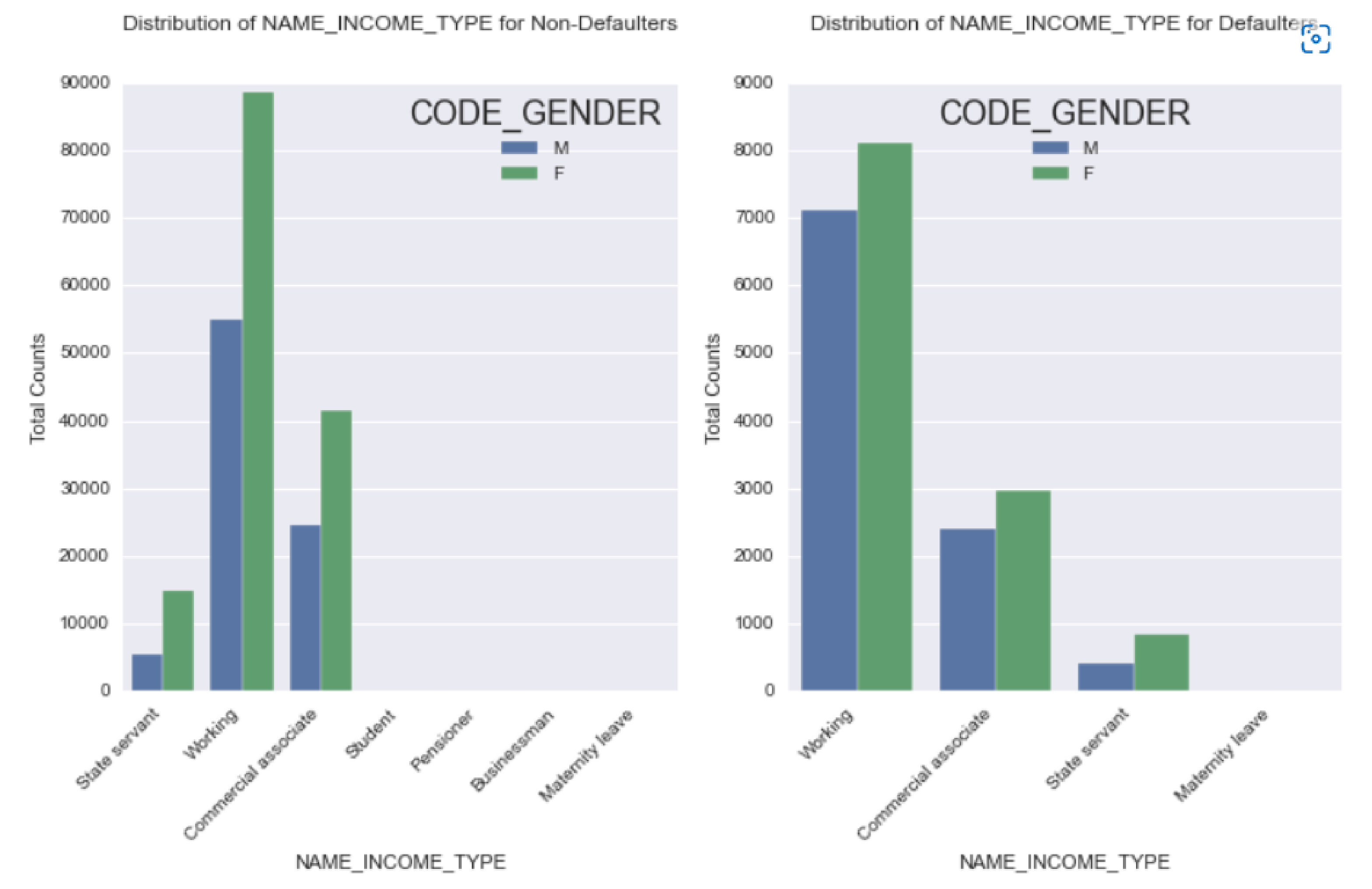
Distribution of amount income range:

From the above graph we can infer that people applying for loan are higher in income range from 100000-150000 , at same time they are higher on defaulters side comparative to non defaulters , similarly people in income range from 50000-100000 are more on defaulters list than compare to non defaulters hence makes them risky customers . In this case people with income range between 200000-450000 can be good potential customers as they are on higher side of income so less likely to default.



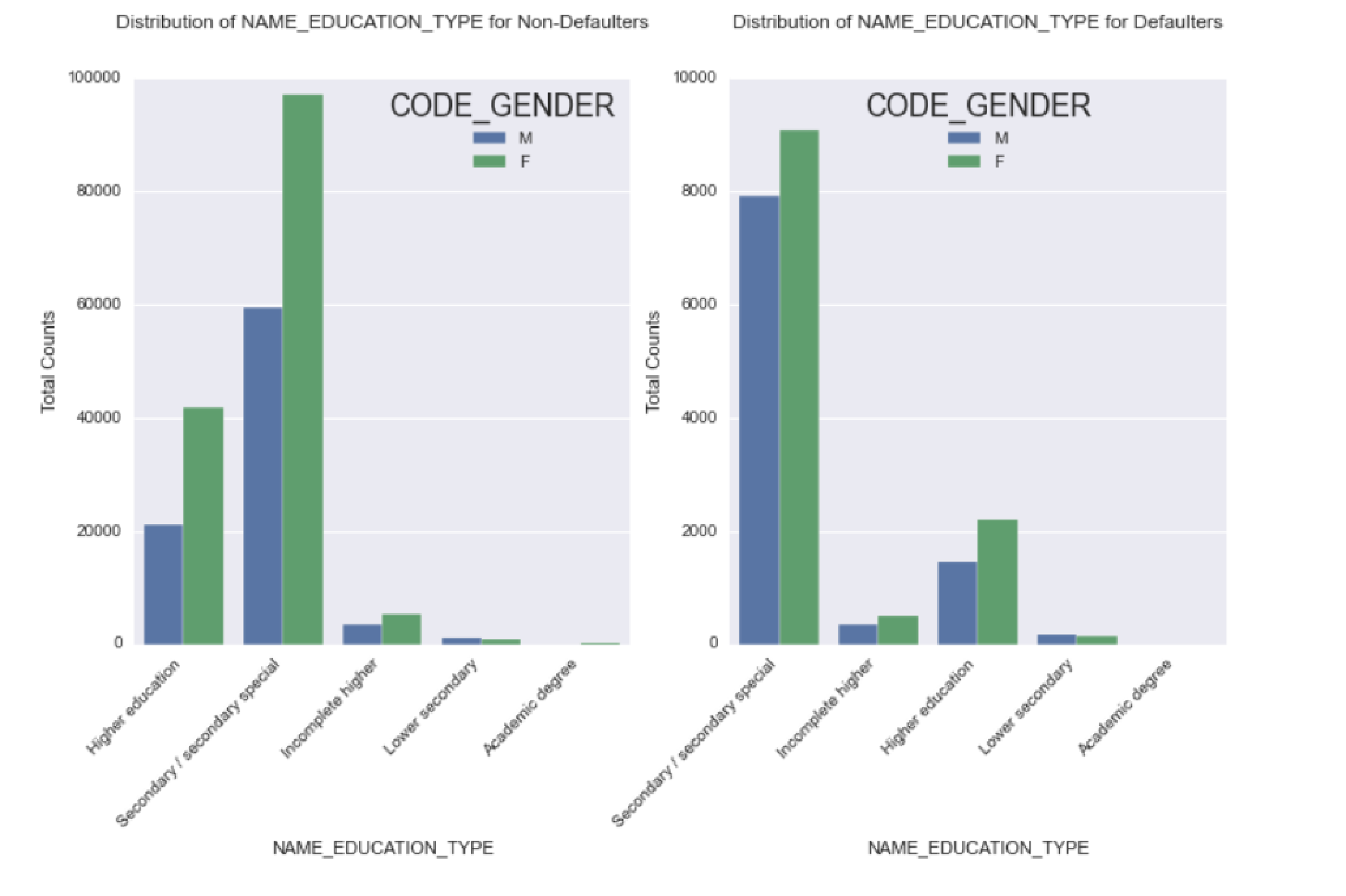
Segmented univariate analysis:

Most of the loans are distributed to working class people and in that also female have high count but at same time We can see that working class people contribute maximum to defaulters, while the males are comparatively high on defaulters side then those who applied for loan. Clearly, It state's that chances of defaulting are more in their case. Where as in case of commercial associate are also high chances of Default but in this case female can be the most reliable customers.



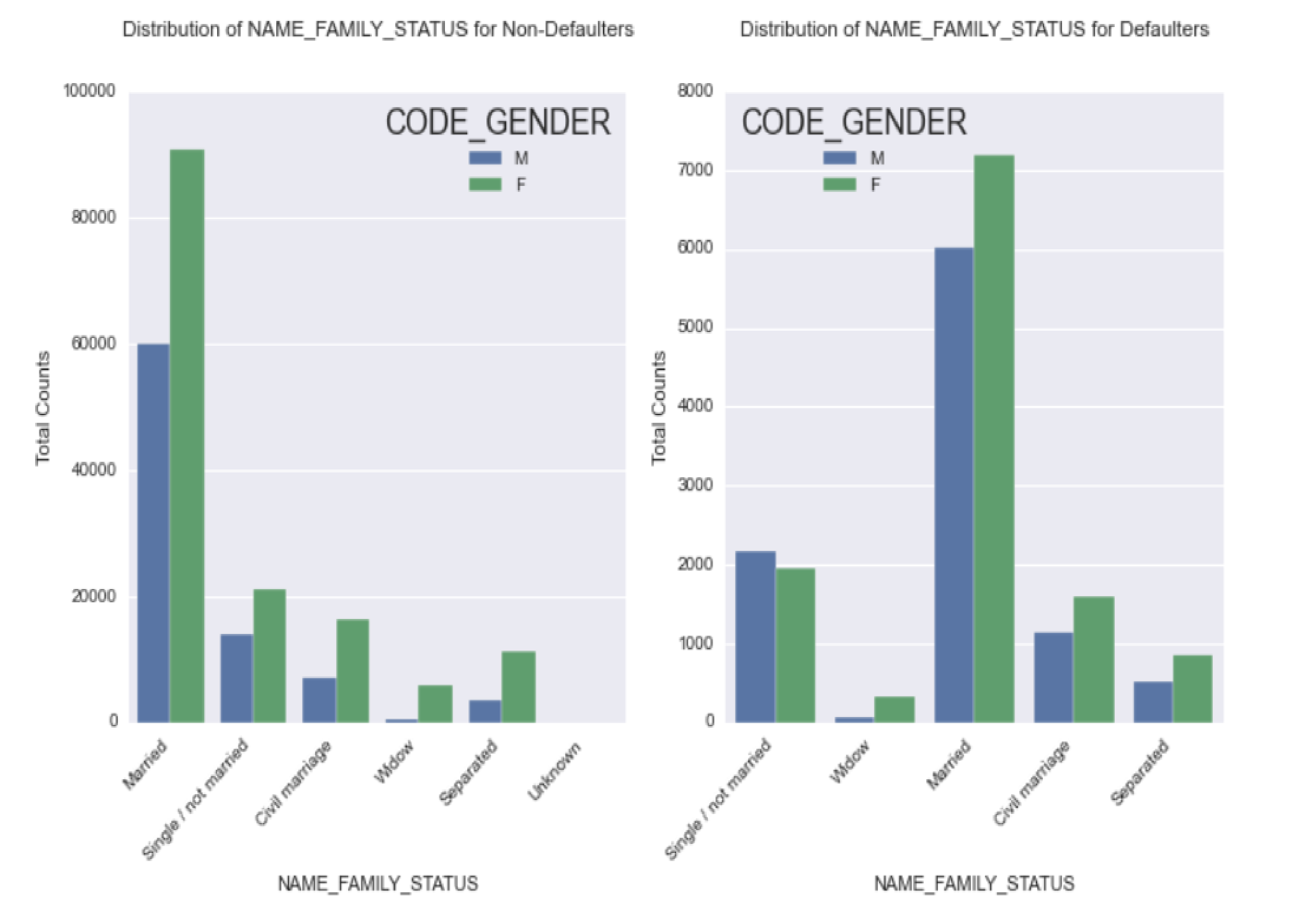
Distribution on Education between male and female:

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 especially the males



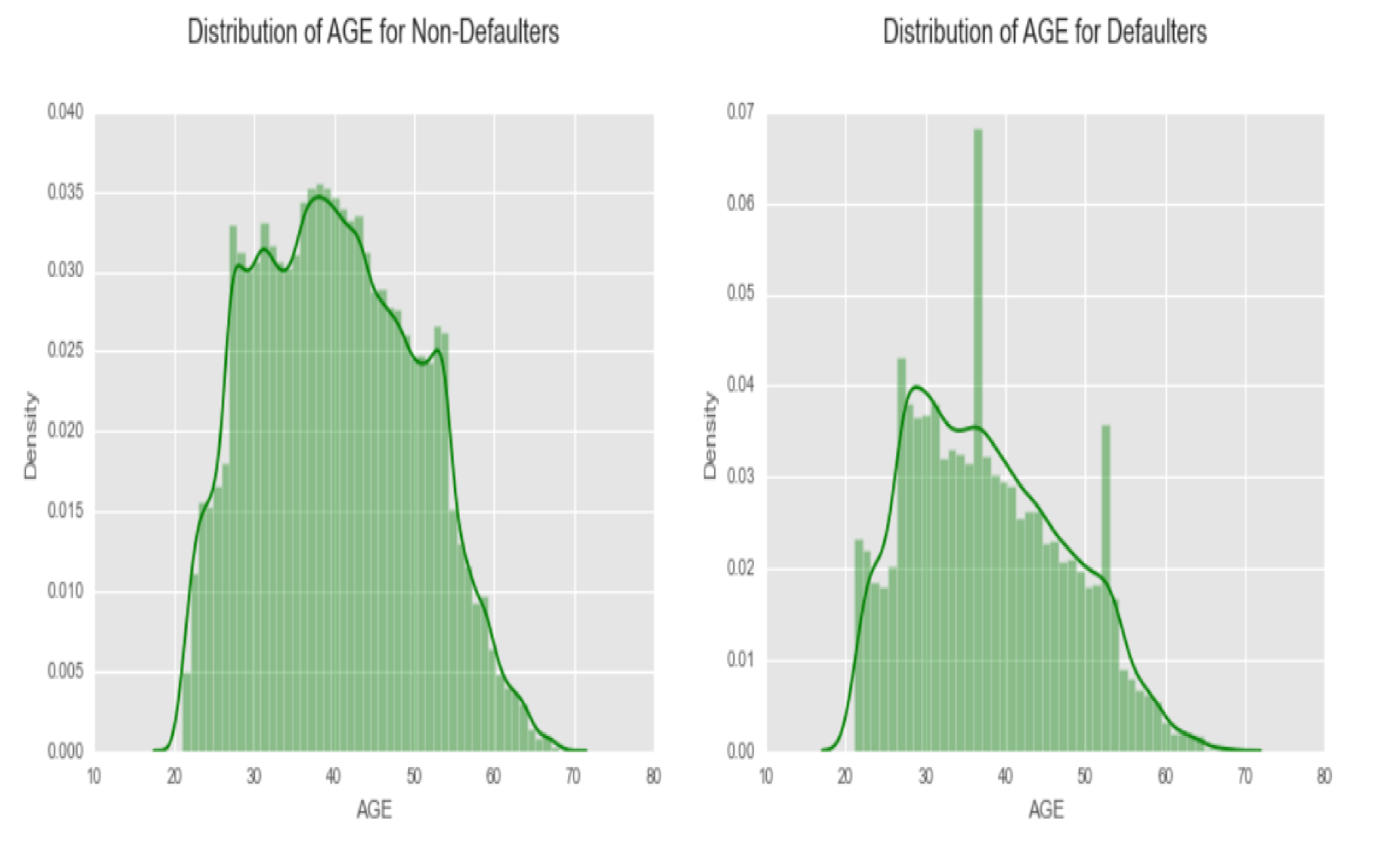
Distribution based on family status:

The plot shows Married people tends to apply for more loans comparatively but at the same time are high on defaulter list . Married female can be our potential customer as her ratio of defaulter against non defaulter is less compare to married males , could be because of more financial responsibility might be shared by male's. we also see that Single/non Married people take less loan but single males are high on default side then females.



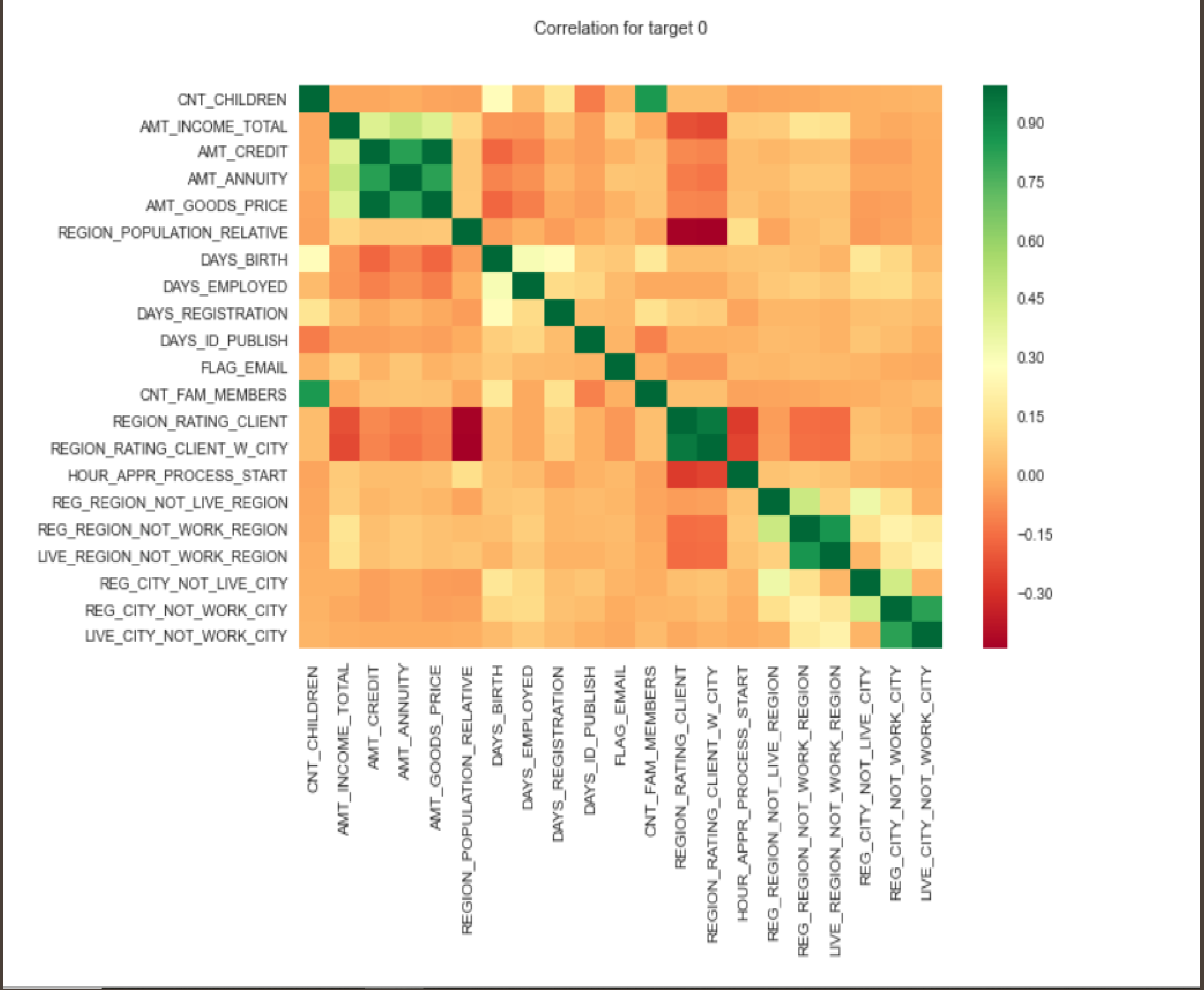
Distribution on basis of age:

In age graph we can observe that age group between 35-45 yrs are high taking loans and are less on non defaulter side except age group of 36-38 which show high peak on defaulter side, again age group between 25-35 that's is more younger age group are also on high defaulter side . Company can rely on people after 55 age group.



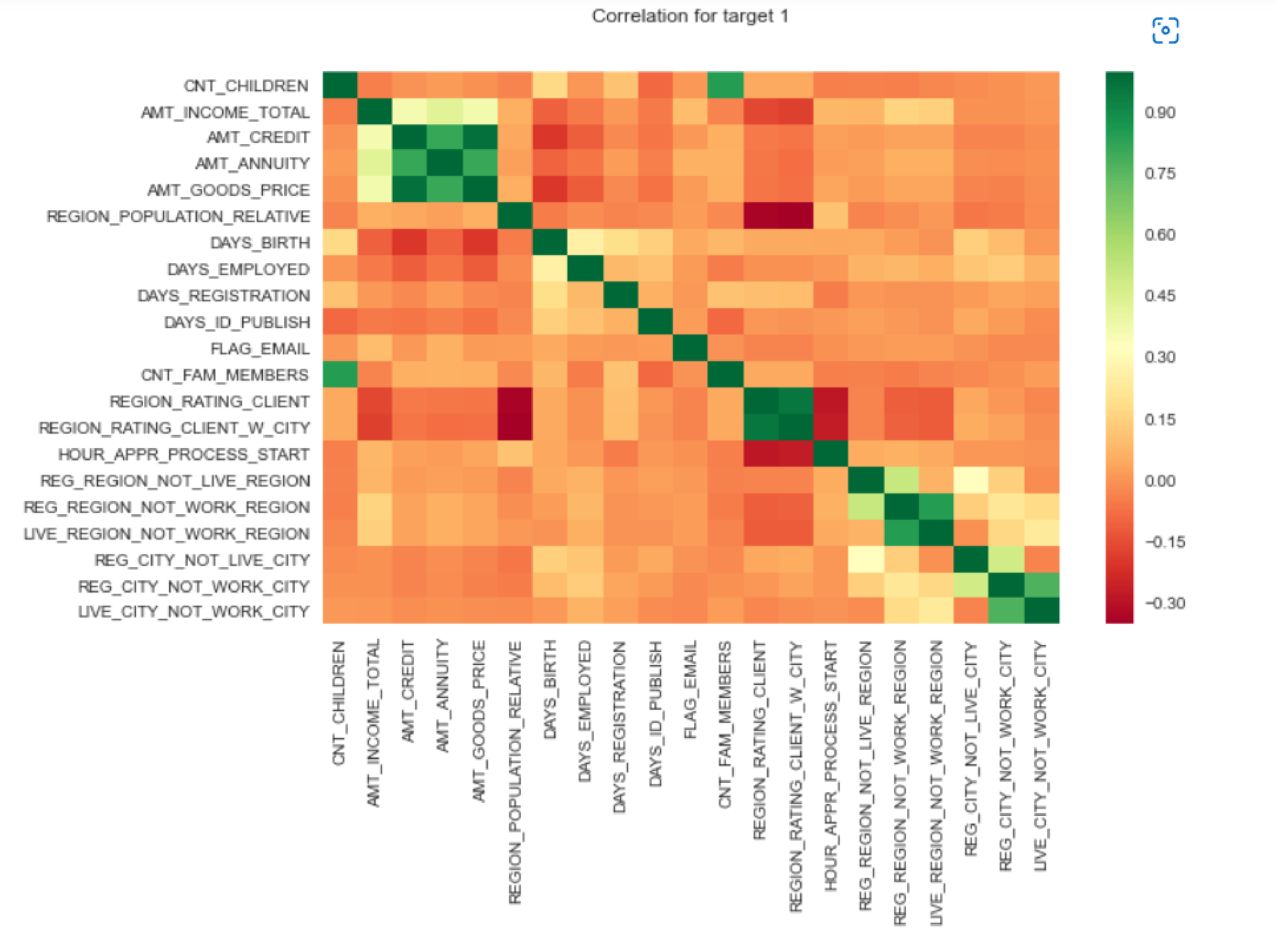
Correlation of TARGET non defaulter's(o):

:Credit amount is higher for low age and vice-versa.
:Credit amount is higher for less children count client have and vice-versa.
: More income for less family member client have and vice-versa.
: Income increases with high population density region.
: High Credit amount is provided for Goods price.



Correlation on TARGET defaulter’s(1):

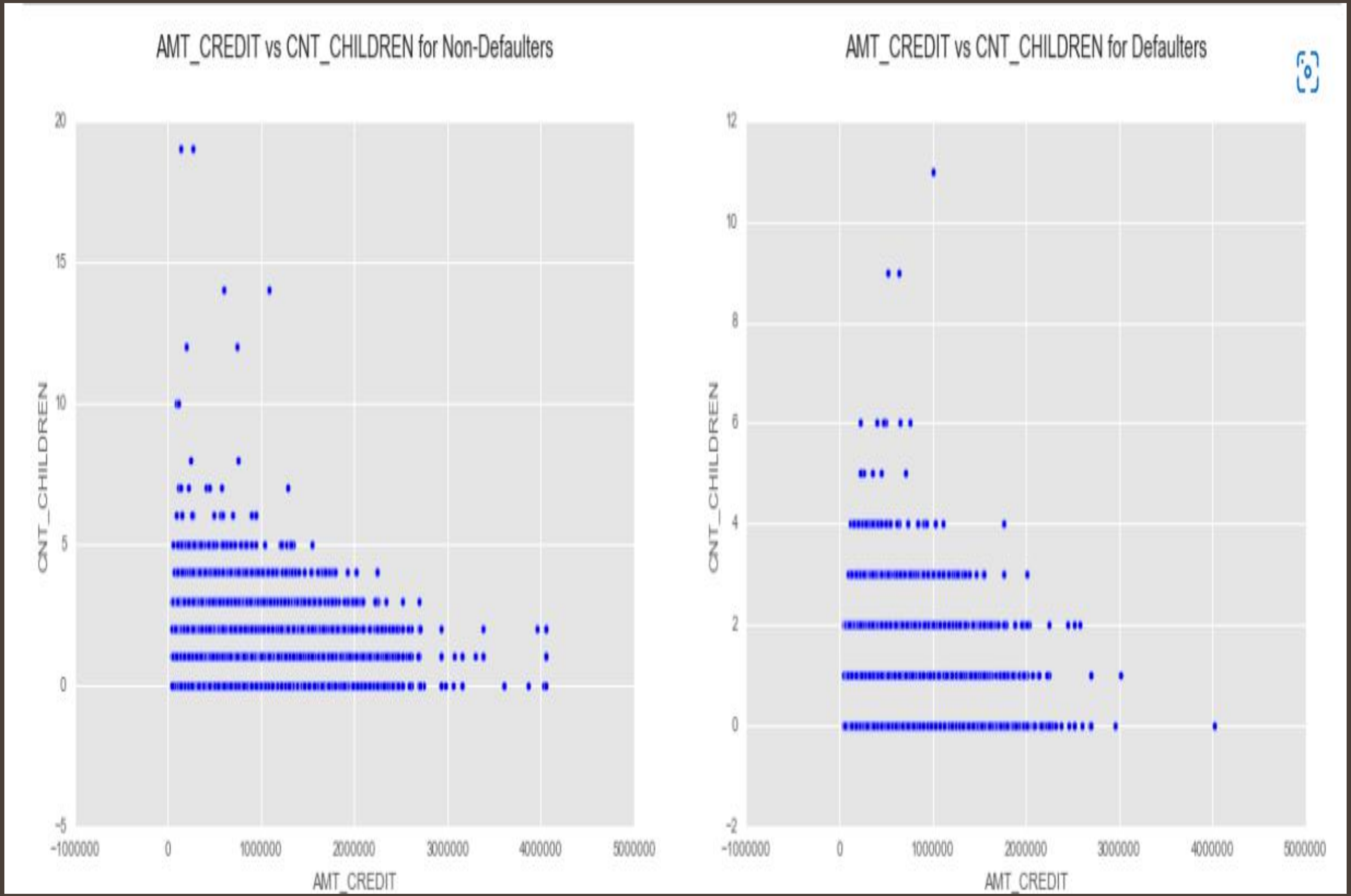
This heat map for Target 1 is also having quite a same observation just like that Target 0.
: The client's permanent address does not match contact address that are having less children and vice-versa.
: The client's permanent address does not match work address are having less children and vice-versa.



Bivariate Analysis of numerical variables:

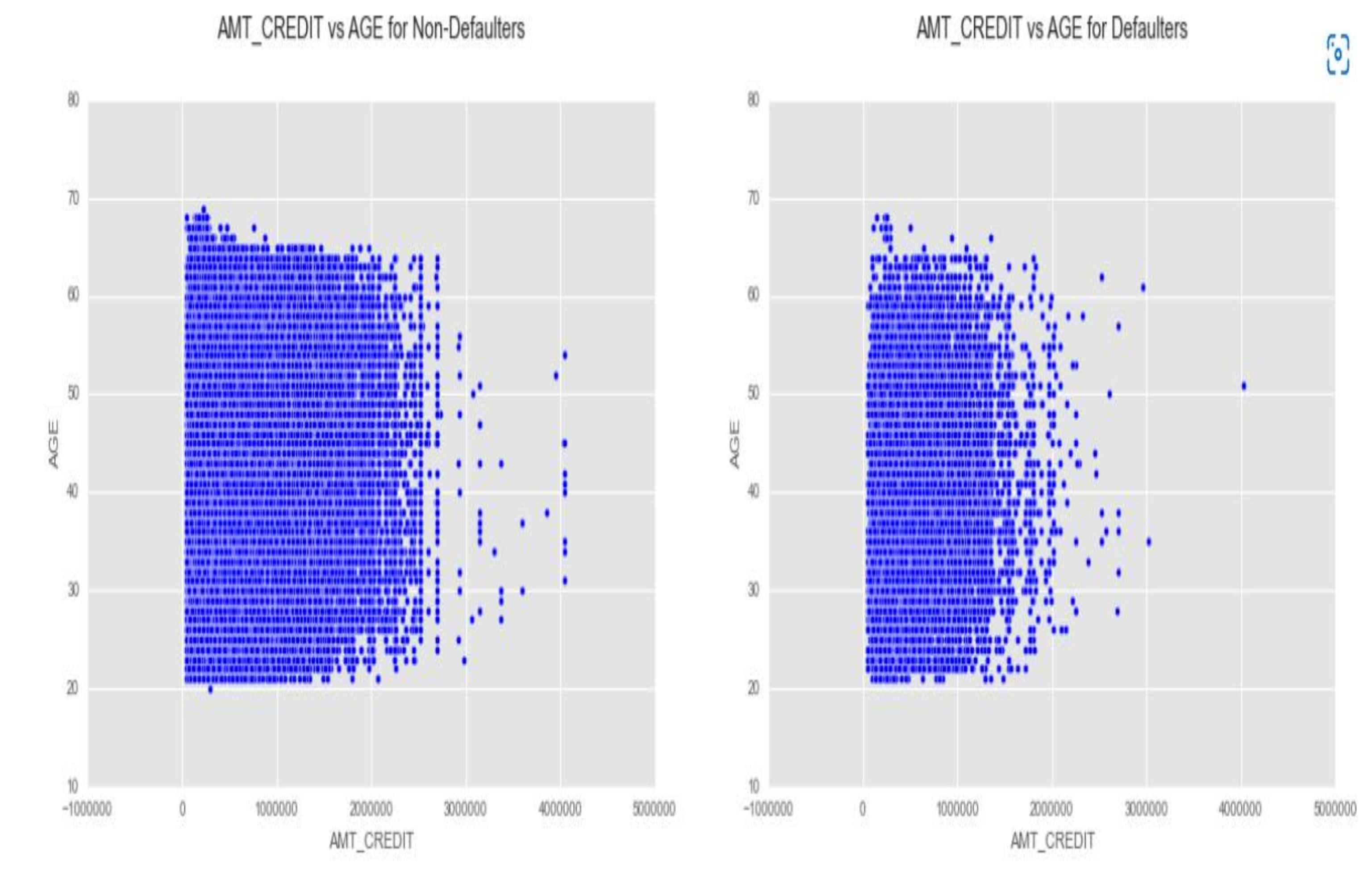
Credit amount and on basis of number of children

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 they have less children and the AMT_CREDIT is low.One more inference can be taken people with high number of children even having low amount of loan are more likely to default where as higher loan amount with less children are low on defaulter side



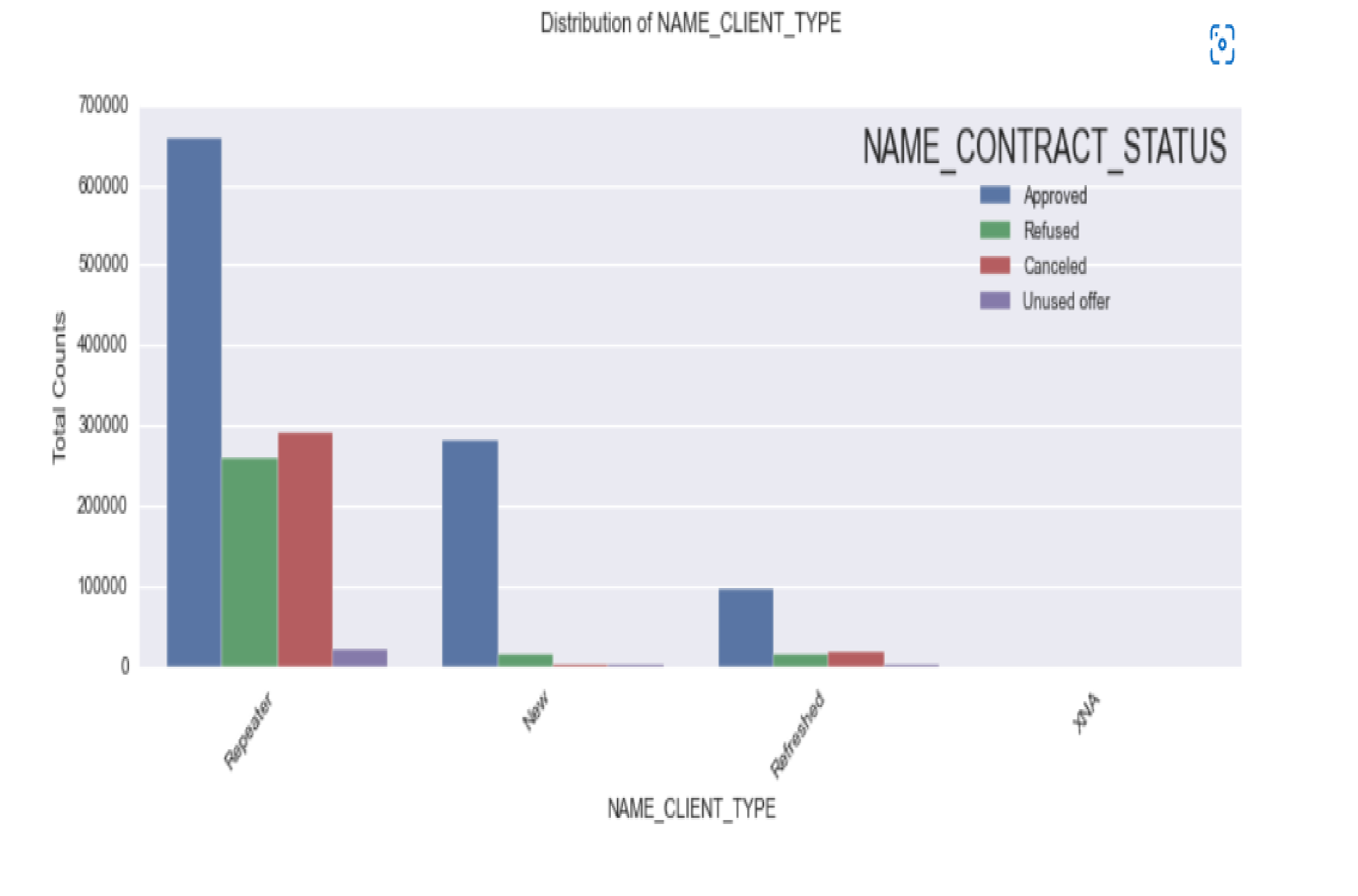
Bivariate Analysis on credit amount and age:

In this graph we can see people of all age group(20-69) are more on defaulter side who have taken loan amount between 0-1200000. however people with higher loan amount and of higher age are less on defaulter side



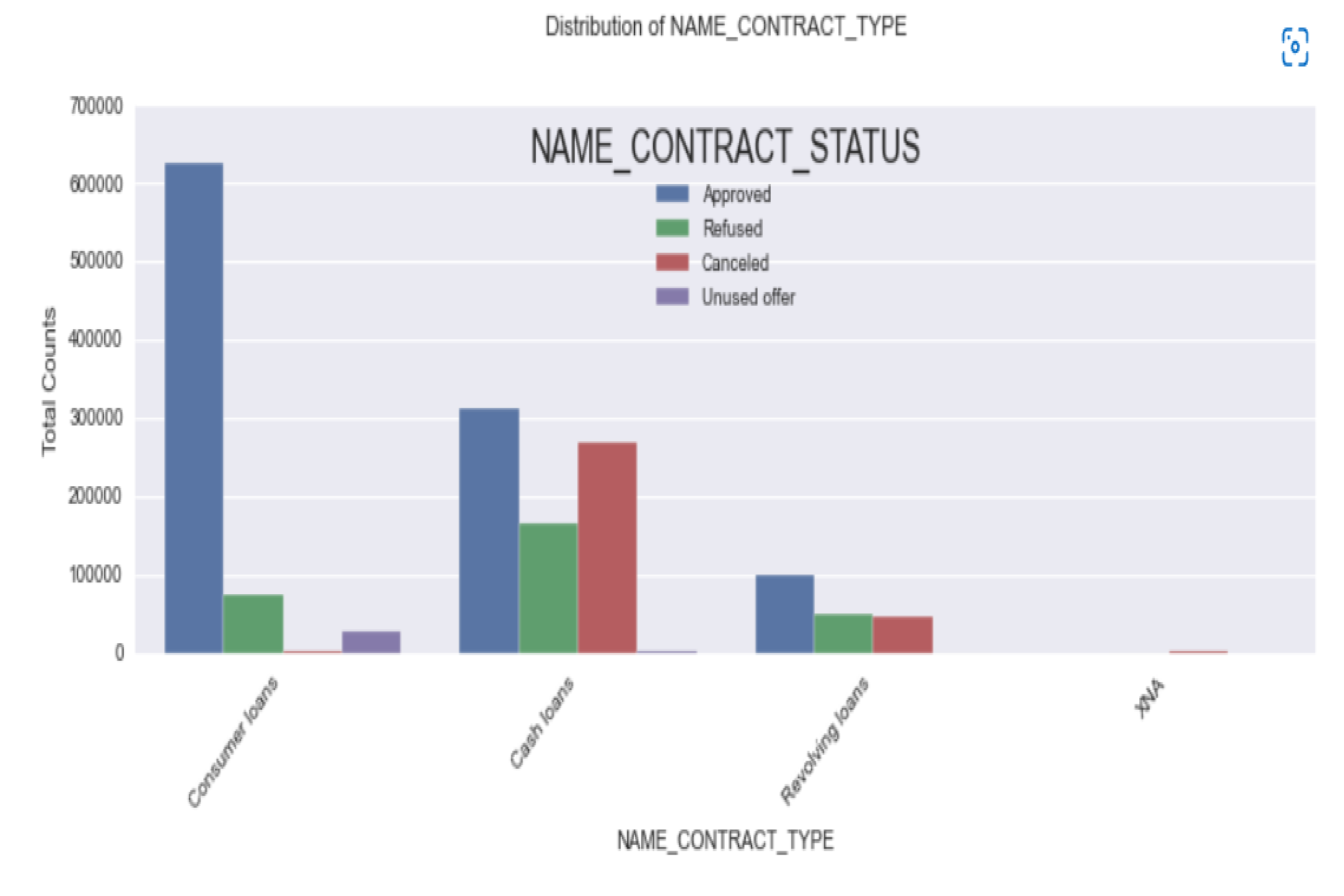
Univariate analysis : For Previous data. name client type on name contract status

Most of the loan applications are from repeater customers, out of the total applications 70% of customers are repeaters. They also get refused most often and there are high numbers of cancel as well ,where as it is interesting to note in new applications maximum are approved



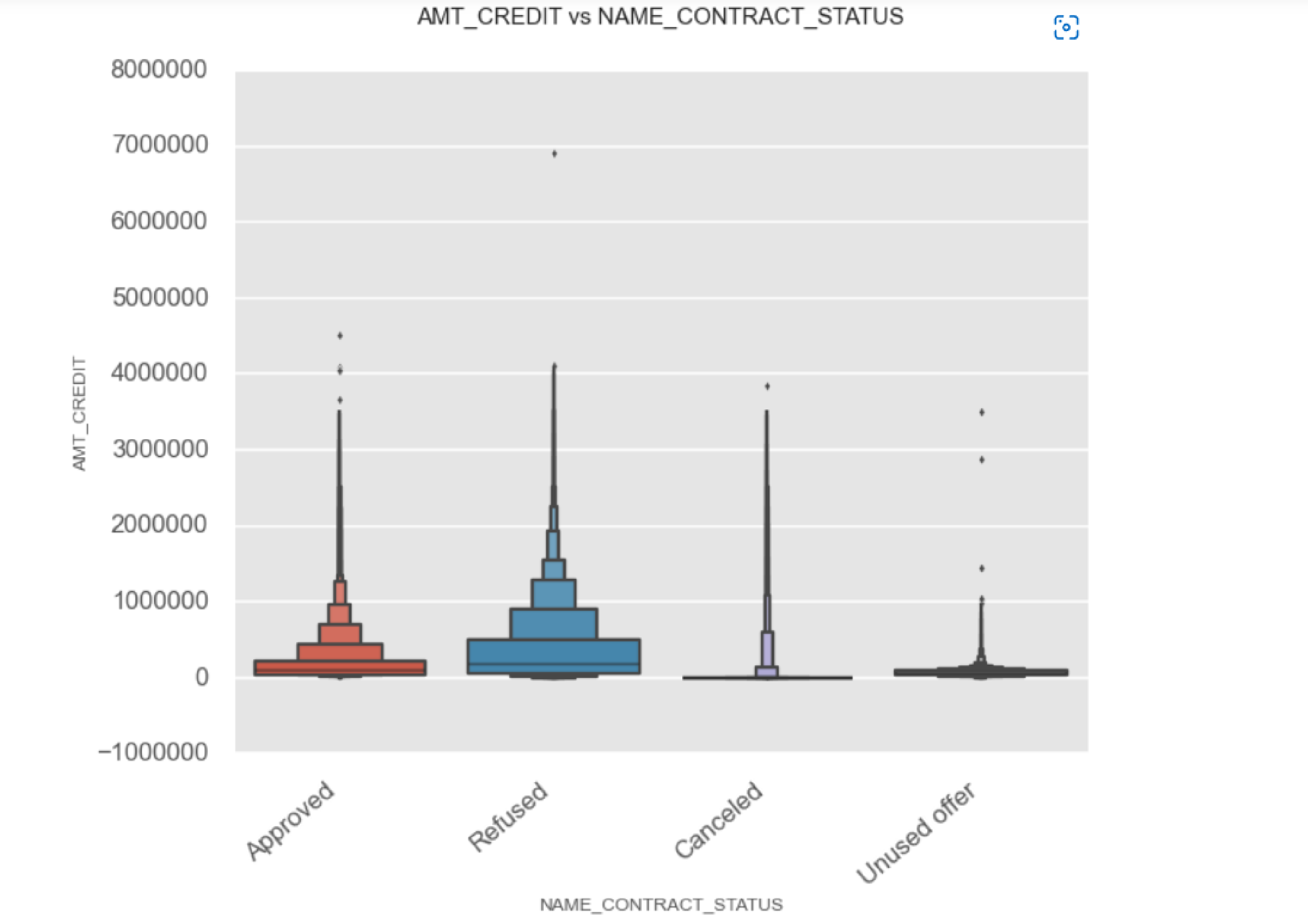
Univariate analysis: on name contract type and name contact status:

From the graph, we can infer that, most of the applications are for 'Consumer loan' and 'cash loan’. Moreover the cash loans are canceled and refuse more often than others where as consumer loans are more readily excepted.



Bivariate analysis on categorical vs numeric columns on Previous data:

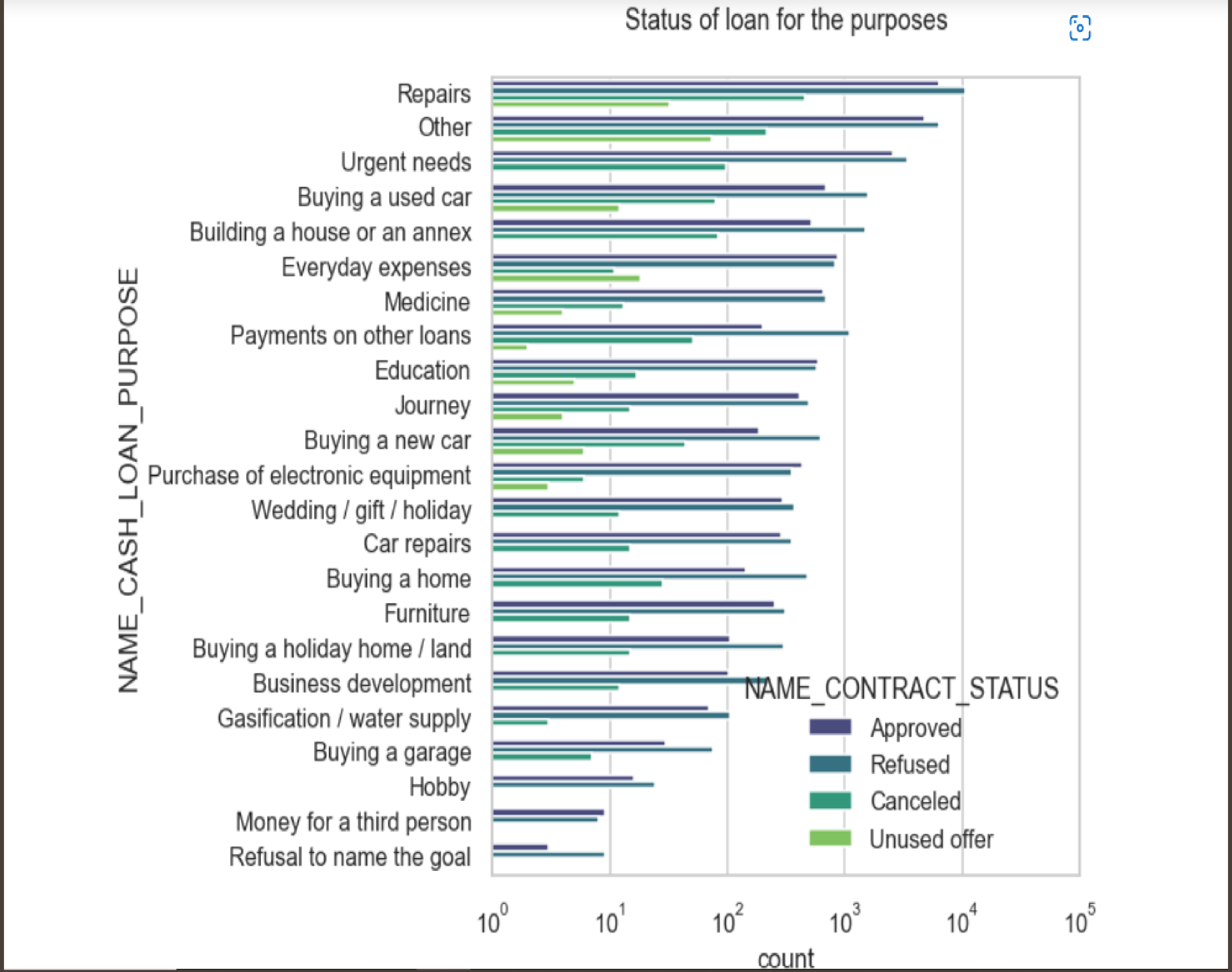
From the above plot we can see that loan application for people with lower amount credit gets canceled or Unused most of the time. We also see that applications with too high amount credit are also getting a time's refused than others. Potential customer's for the company would be people applying for credit amount between 0-2000000.



UNIVARIATE ANALYSIS AFTER MERGING PREVIOUS DATA

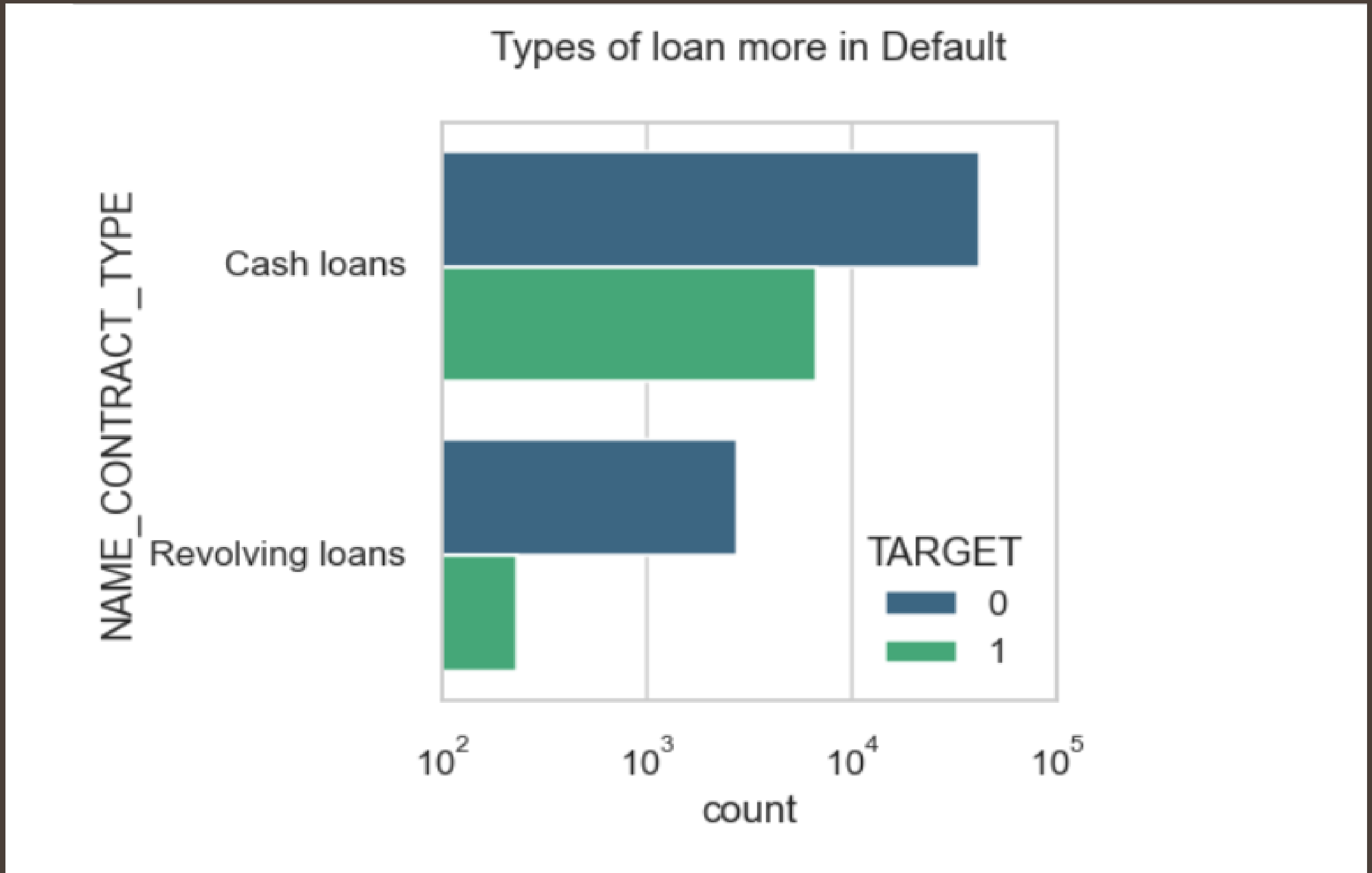
Distribution between Status of loan for the purposes and name contract status:

conclusion from above plot: Most rejection of loans are from 'repairs' purpose. For education purposes we have equal number of approves and rejection Payment on other loans and buying a new car is having significant higher rejection than approves.



Distribution between types of loan and name contact type:

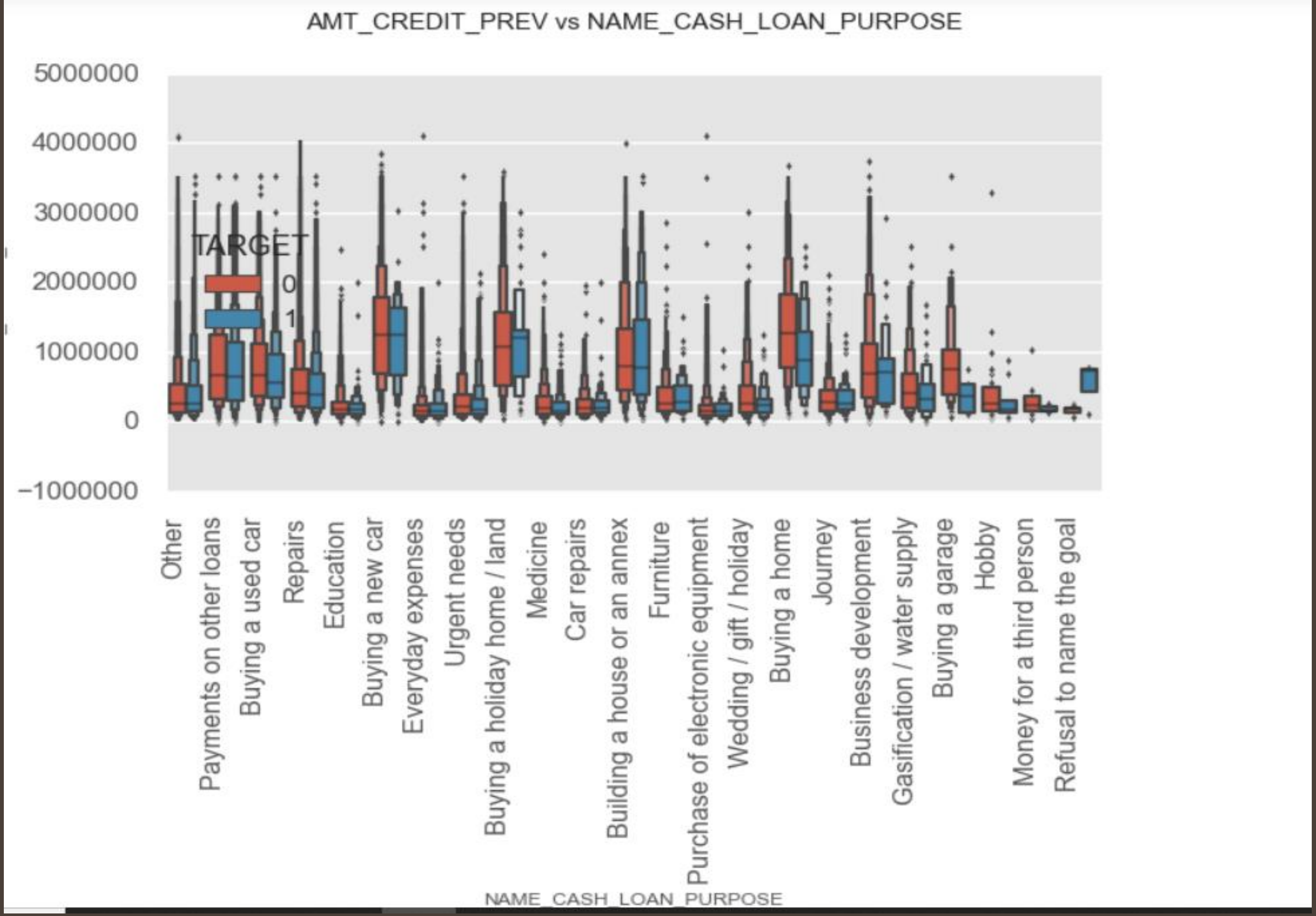
We can infer that cash loan have more defaulters than the revolving one's



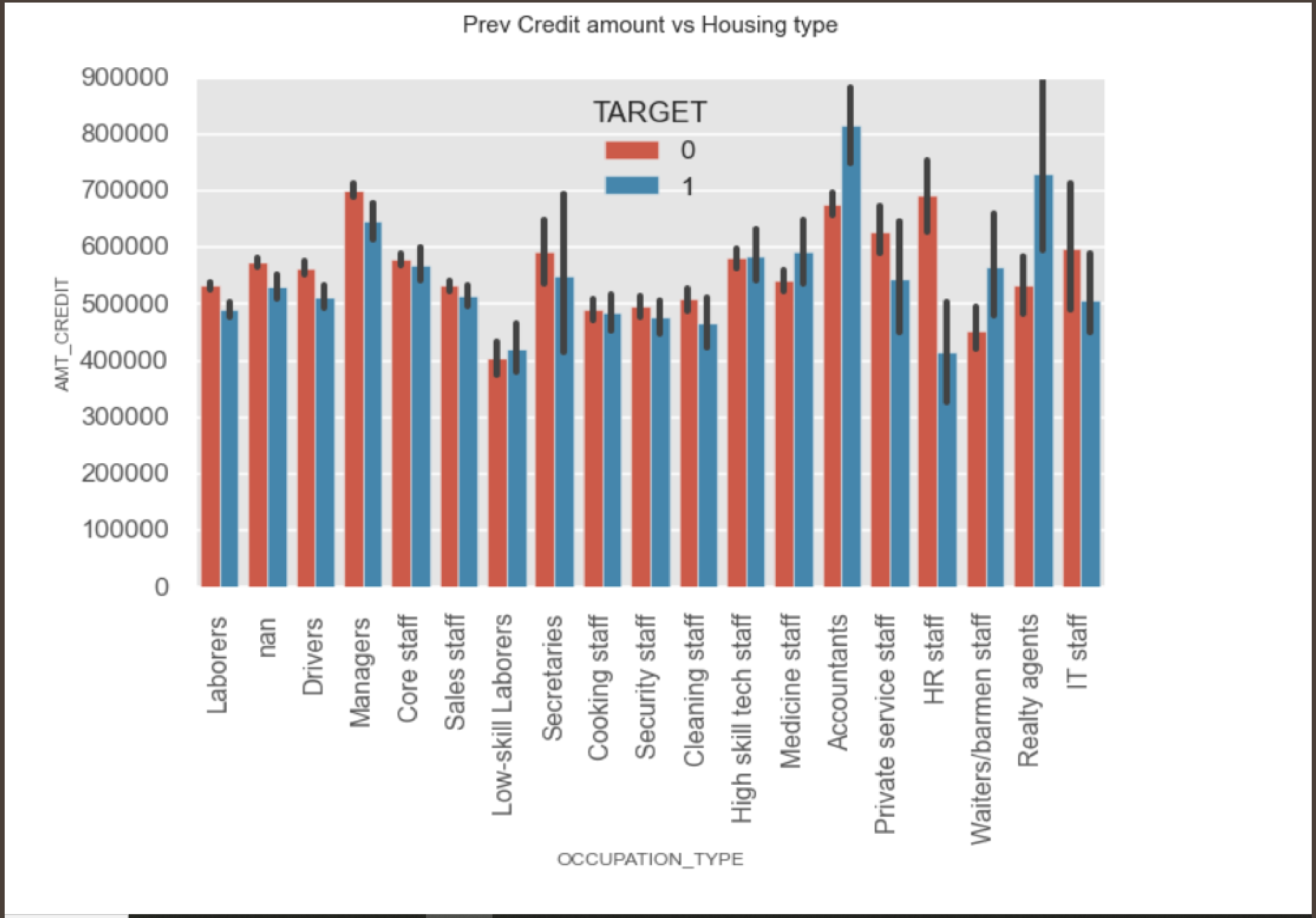
Target variable (0 - Non Defaulter 1 - Defaulter)

Bivariate analysis on Merged data : Cash loan for the purpose and amount credit prev on target variable:

From the above we can conclude some points-
The credit amount of Loan purposes like 'Buying a home' ,' Buying a land' ,' Buying a new car' and 'Building a house' is higher with high number of defaulter's in buying a new car, Payments of other loans and Building a house. best Target people will those who taking small loans and are less on defaulter sides like Education loan(as they have not to pay immediately) , Buying Garage(might gives good earning instantly), Journey loans.



Distribution between Prev credit amount and Housing type.



From the above plot we can observe people with occupation in Realty/agents , Accountants, Waiters are more on defaulter side with very high credit amount . Bank should go with customers like HR staff , IT staff, Mangers, Secretaries

Recommended group:

1. We saw earlier that people who are car owner and house owner are on less side of being defaulters .The company can add weightage to house and car owner along with people who are having small family and less children.
2. People from occupation belonging to HR staff , IT staff , Mangers have been seen on less default side.
3. Company can give more weightage to females specially married one's and belonging from state servant , commercial associate and even for that matter working women.
4. Company should focus more on consumer loans and repeaters..
5. We have also observed people beyond age 40 and above with applying for high credit are low on defaulters sides.
6 For company best potential loans will education loan, every day expense loan , medicine loan , purchase of electronic equipments .

Risky Group:

1. Married and single males are more on defaulter side, company should be more caution before loaning.
2. People belonging to age group below 30 and specially between 35-38 range and 53-55 range are found to be potential defaulters.
3. Male who are Working and are commercial associated should dealt with more care.
4. People applying for credit between 50000-200000 can be consider to be risky customers.
5. Loans for repair purpose , payment for other loans , Buying a new car , Buying a house or home loans are to be given with stringent verification.
6.It has been observed that people from occupation belonging to Realty agent , Accountant , Waiters are more on defaulter side.