EDA Case Study

Group members:
Shraddha Maladkar
Kavita Sardesai

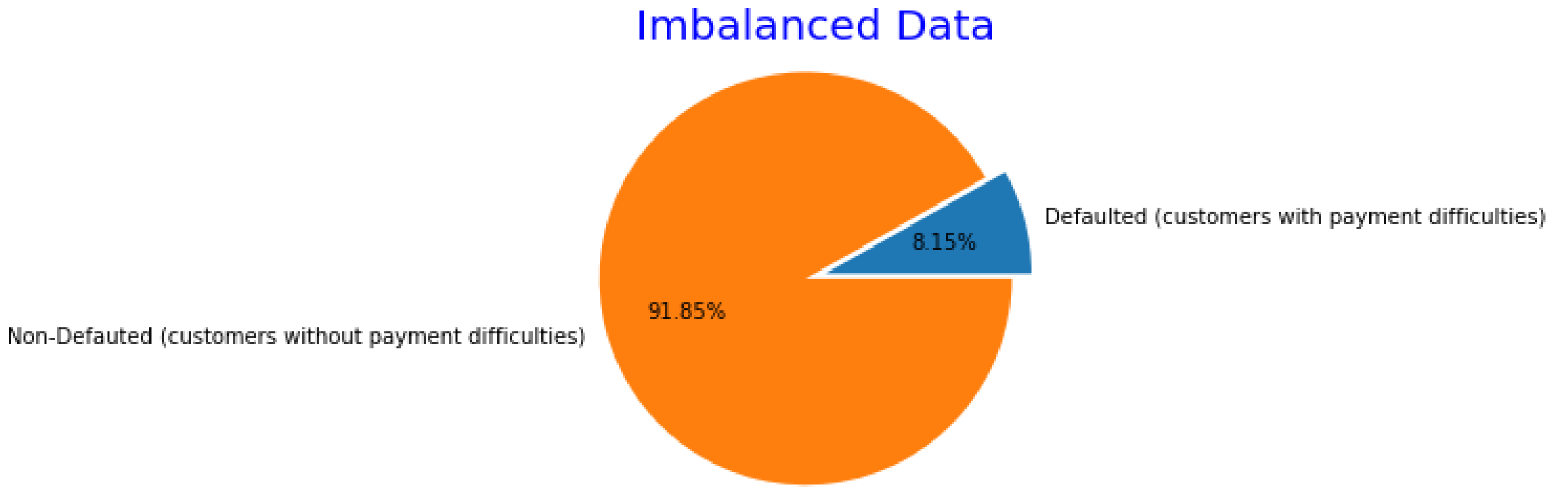
PROBLEM STATEMENT:

- 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:
  - H0 -If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
  - H1-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.

Case study objective

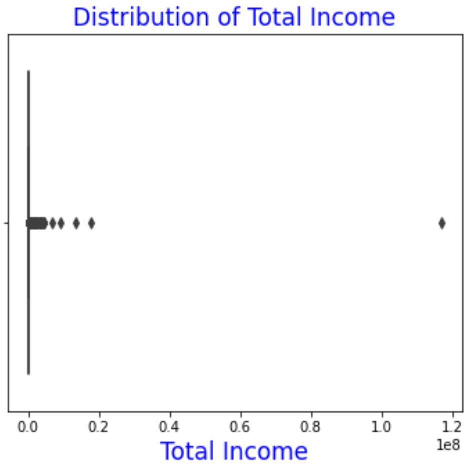
- To clean and standardize the data for analysis
- To understand columns and identify relevance of the columns for analysis
- To extract necessary data from those columns and finding out patterns in that data
- To analyse these patterns which will help to achieve conclusions. Conclusions will help to identify cases that must be 'REJECTED' or 'ACCEPTED' which will be helpful in managing risk for the company
- To ensure that this analysis will help to avoid both risks that were mentioned in the problem statement.

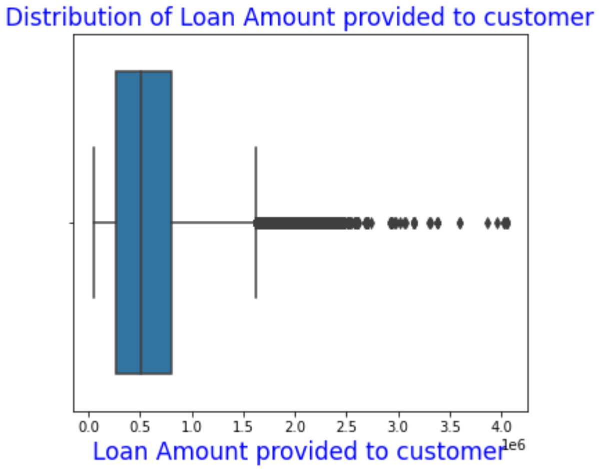
Imbalanced Data Representation

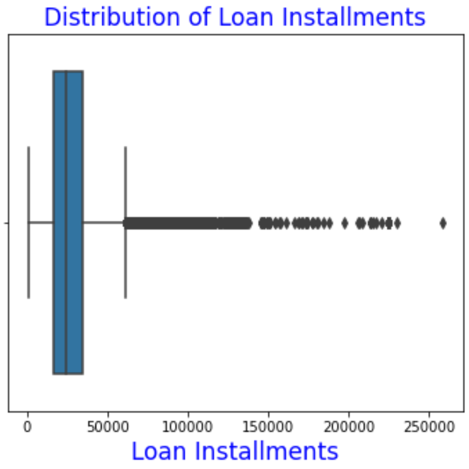


Application data has high imbalance with Defaulted customers at 8.15% as compared to Non-defualted customers at 91.85%.

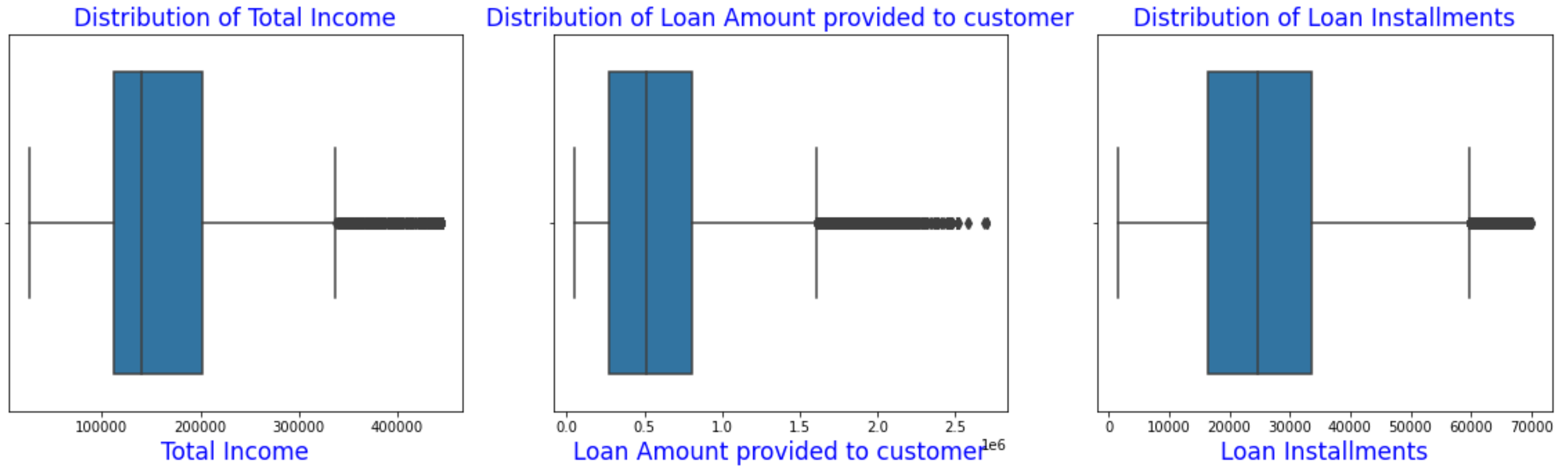
Finding Outliers in data


Distribution of Total Income


Distribution of Loan Amount provided to customer

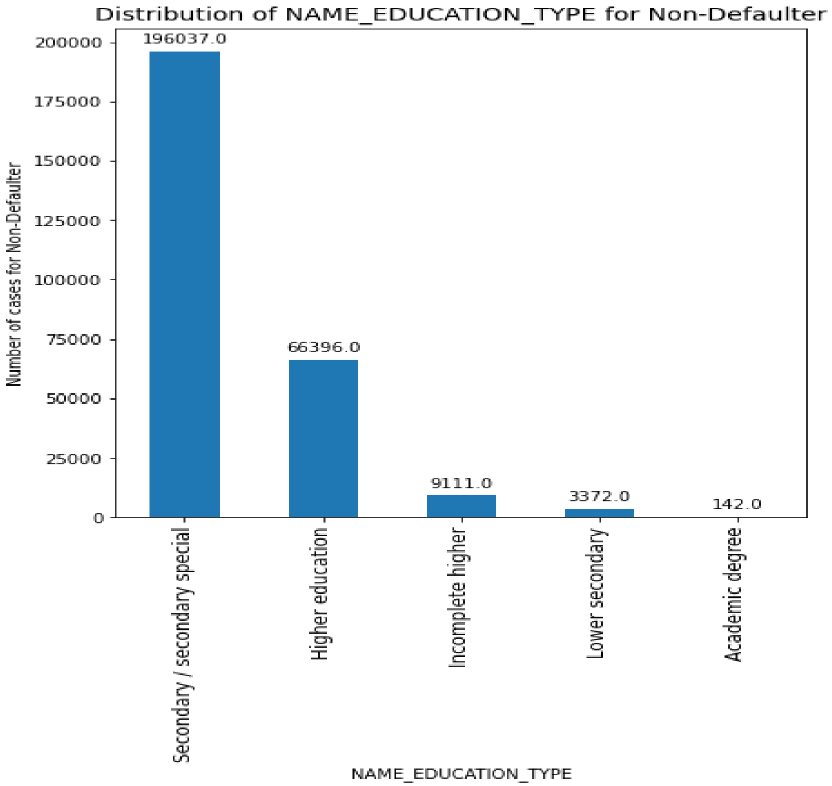

Distribution of Loan Installments

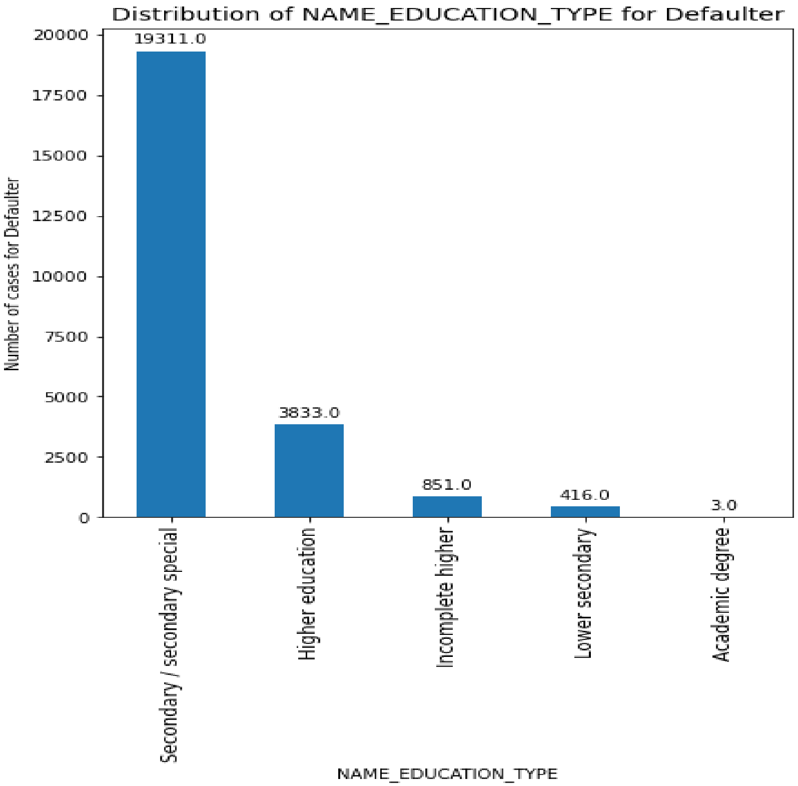
Handling outliers



UNI-VARIATE ANALYSIS:

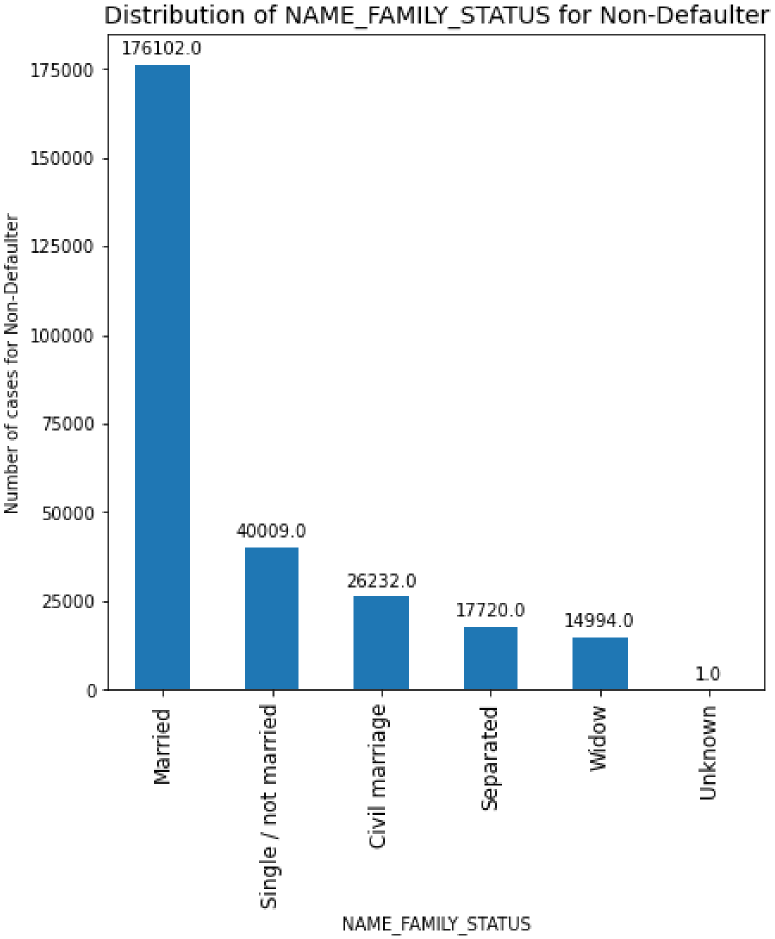
DISTRIBUTION BY EDUCATION



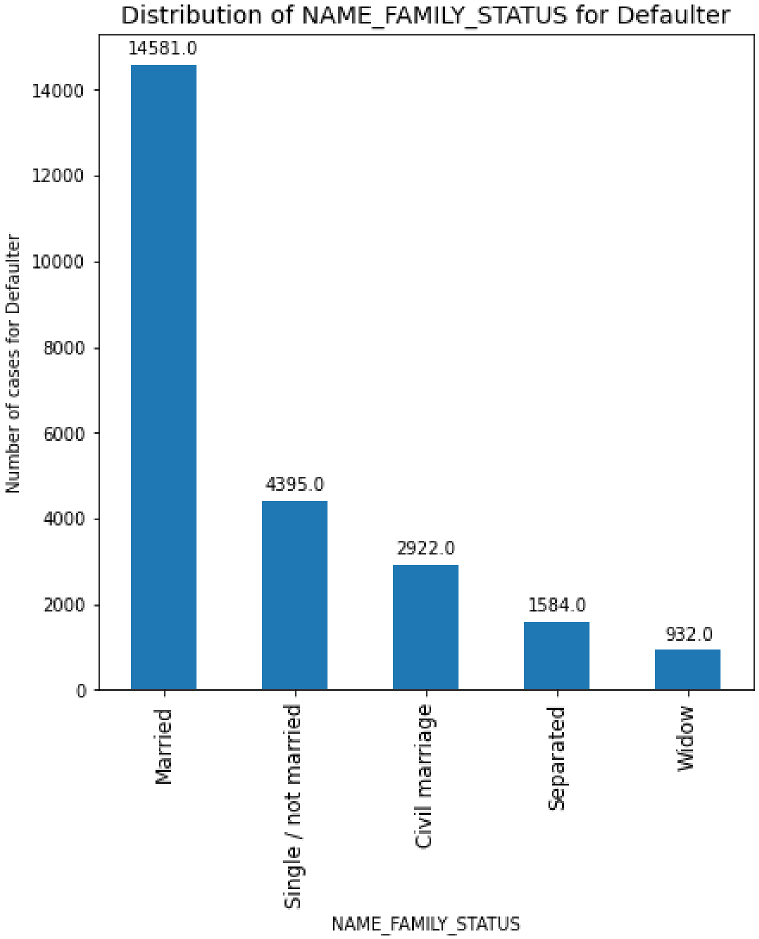


- **Higher education count is proportionally lesser in dafualted population as compared to non defaulted population. Hence higher the education level, lower the default rate. This is logical as higher degree category should be earning more and hence easier to pay off loan installments**

Distribution by family status



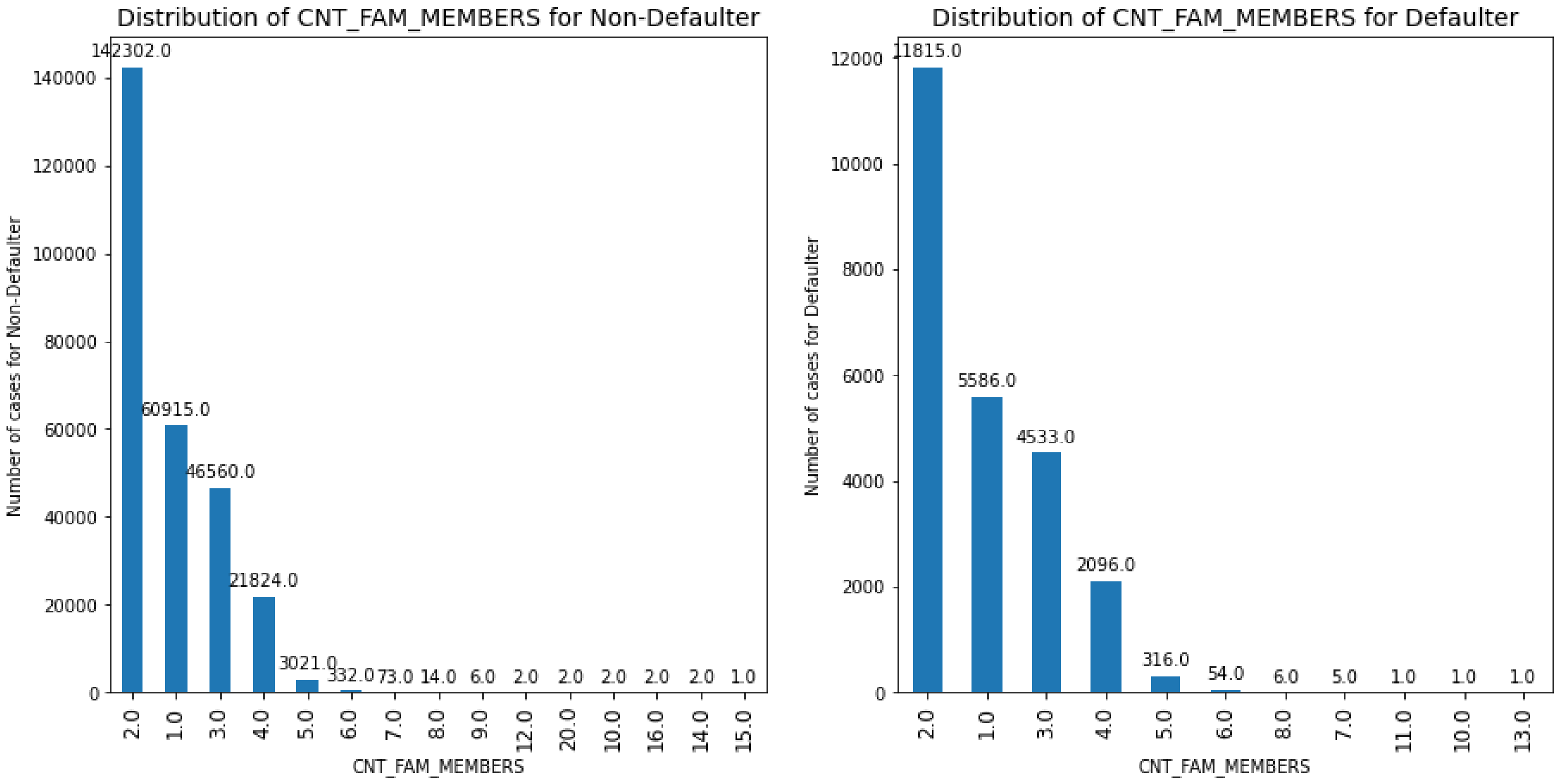
Distribution of NAME_FAMILY_STATUS for Non-Defaulter



Distribution of NAME_FAMILY_STATUS for Defaulter

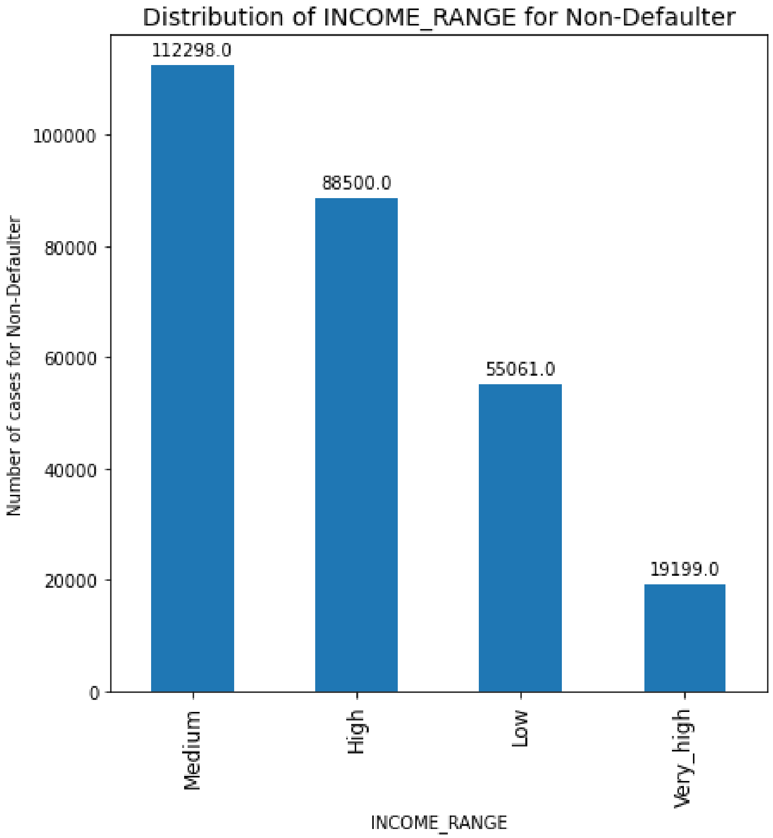
- Single people are having more payment difficulties

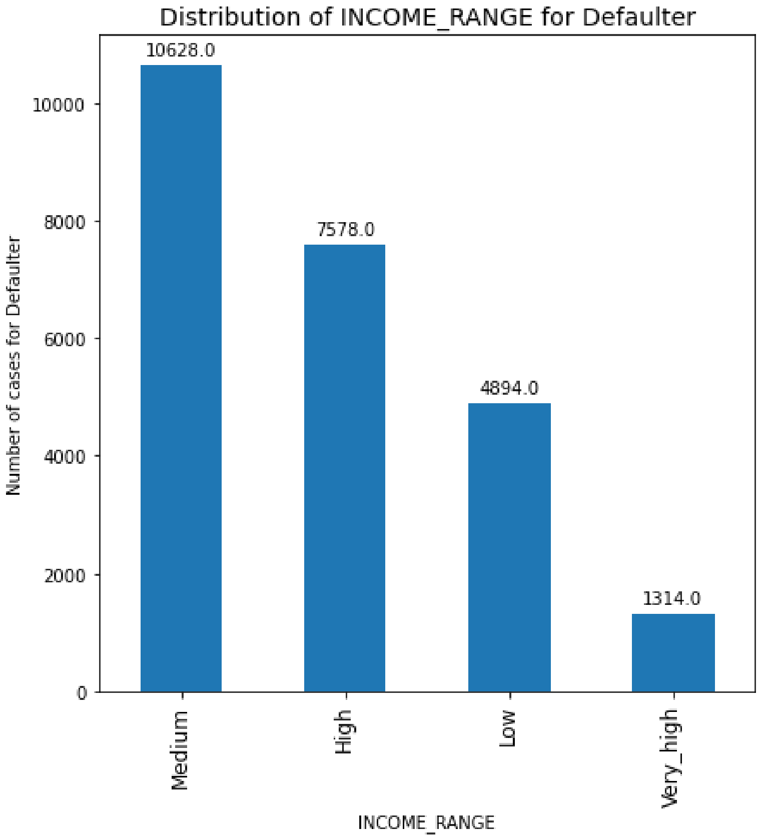
DISTRIBUTION BY CHILD COUNT



- Children count seem to have a little impact on default rate. Although the proportion for higher count of children is a bit more in defaulters as compared to non defaulted

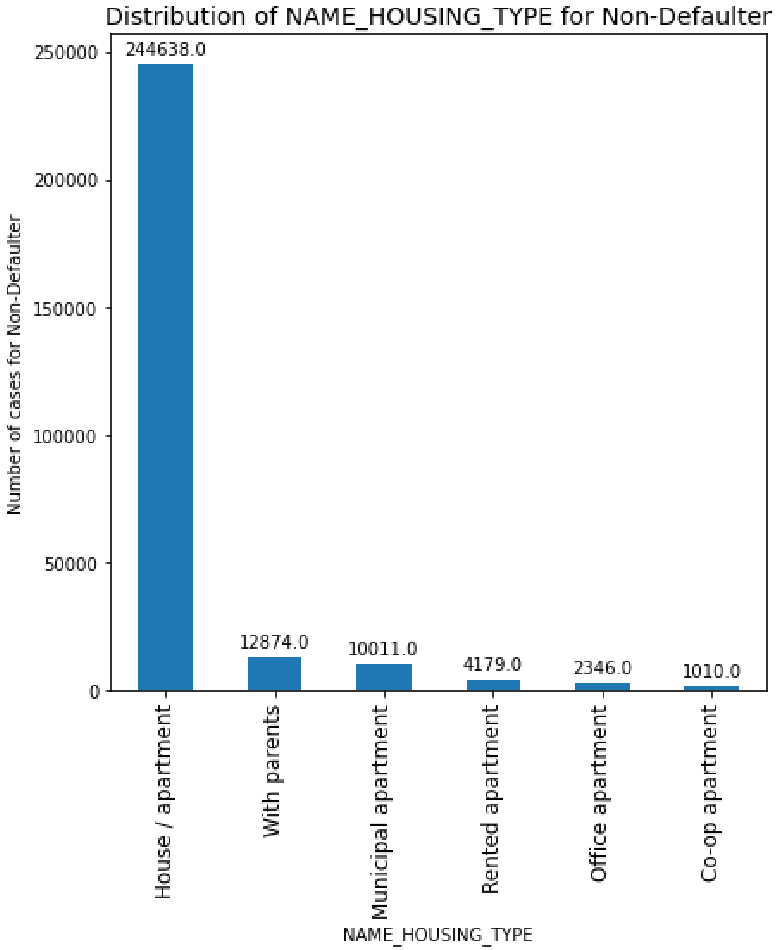
Distribution by income range

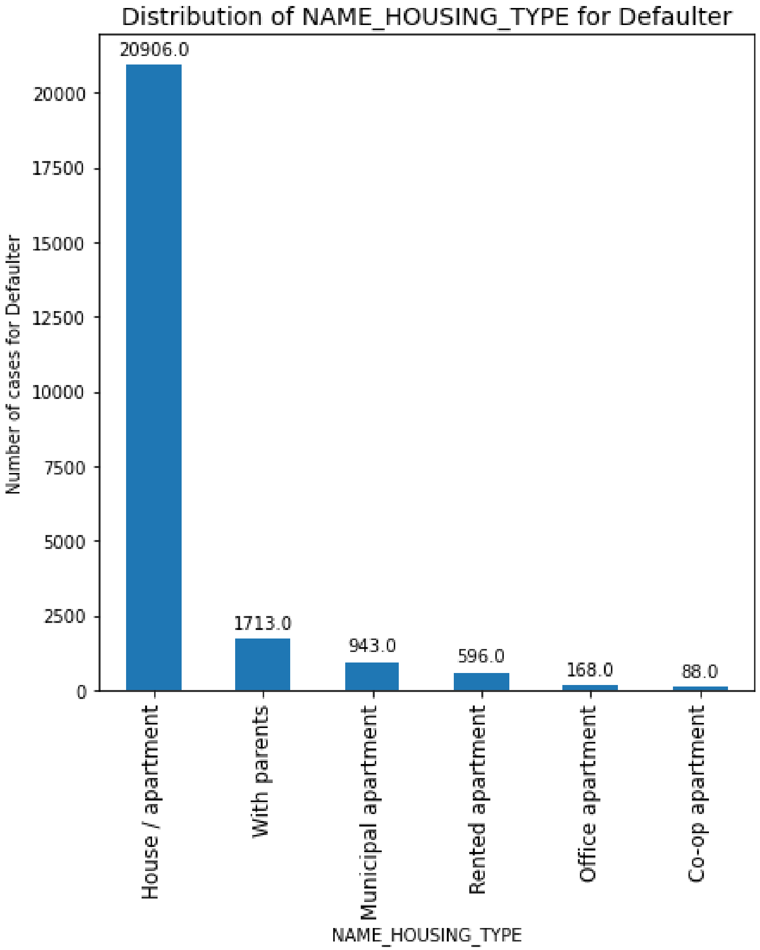




- This plot shows that customers of very high and high income range are non-defaulters which confirms that these customers can afford regular payment of loan and likelihood to default is very less

Distribution by Housing Type

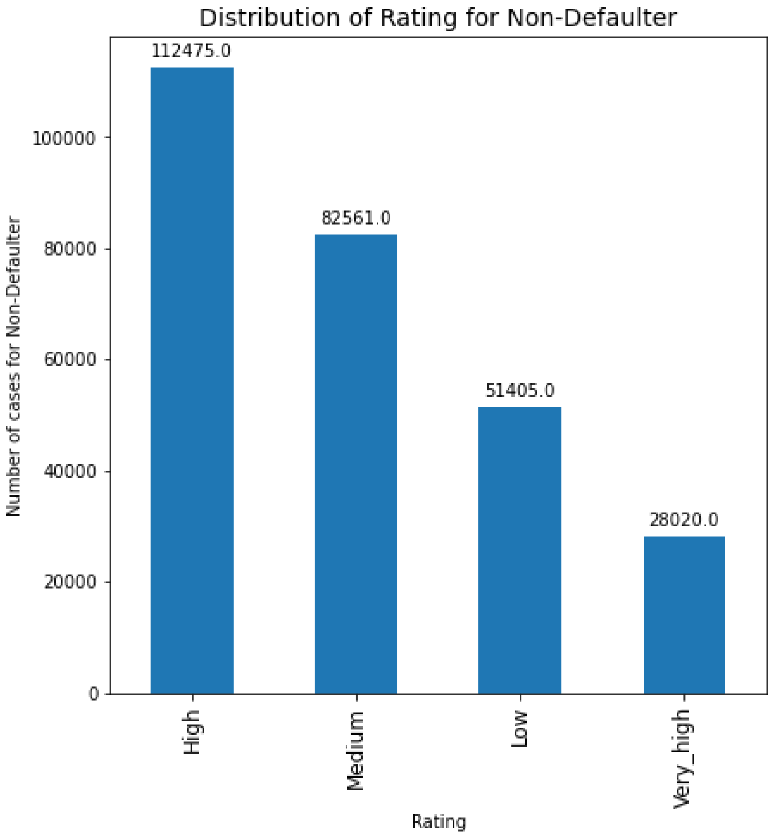


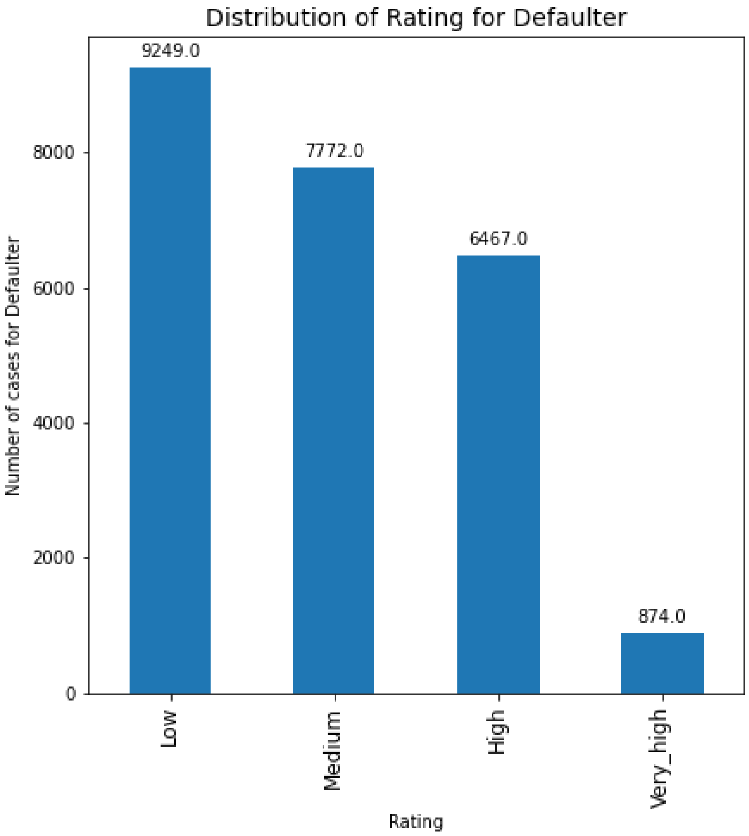


INSIGHTS FROM HOUSING TYPE GRAPHS

- People living in Rented apartments and those living with parents have higher proportion in the Defaulter as compared to non defaulter. There can be difficulty in payment of loan installments for this population as it might suggest that people living with parent may have no or very less income. Also, people with rented apartments must be spending a portion of their income on rent, therefore facing difficulty during repayment of loan.

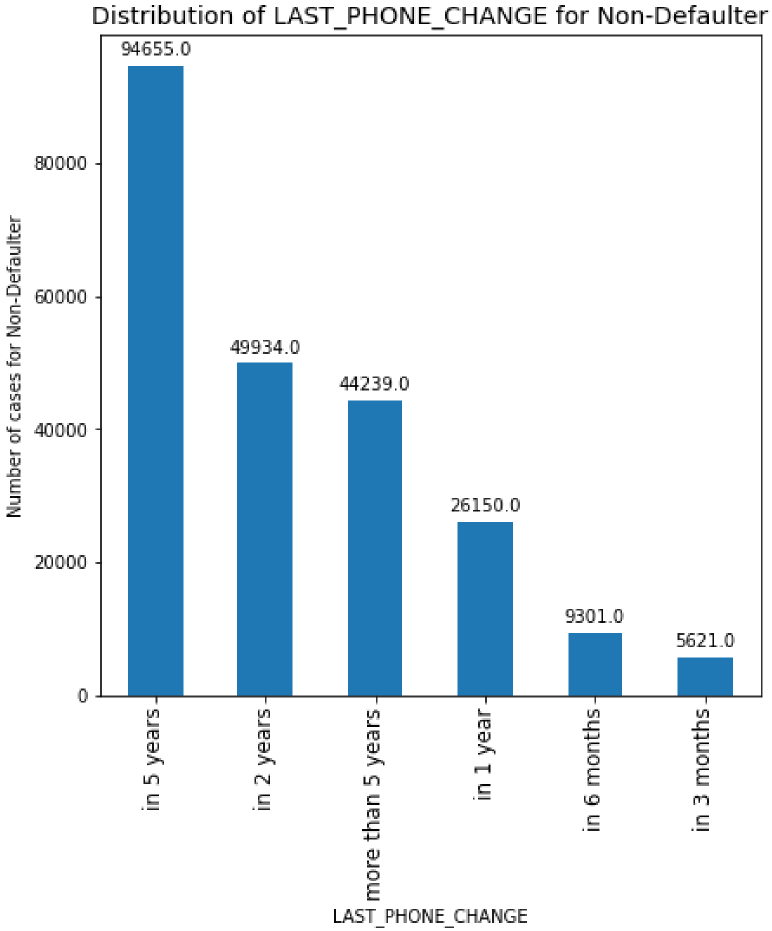
Distribution by rating

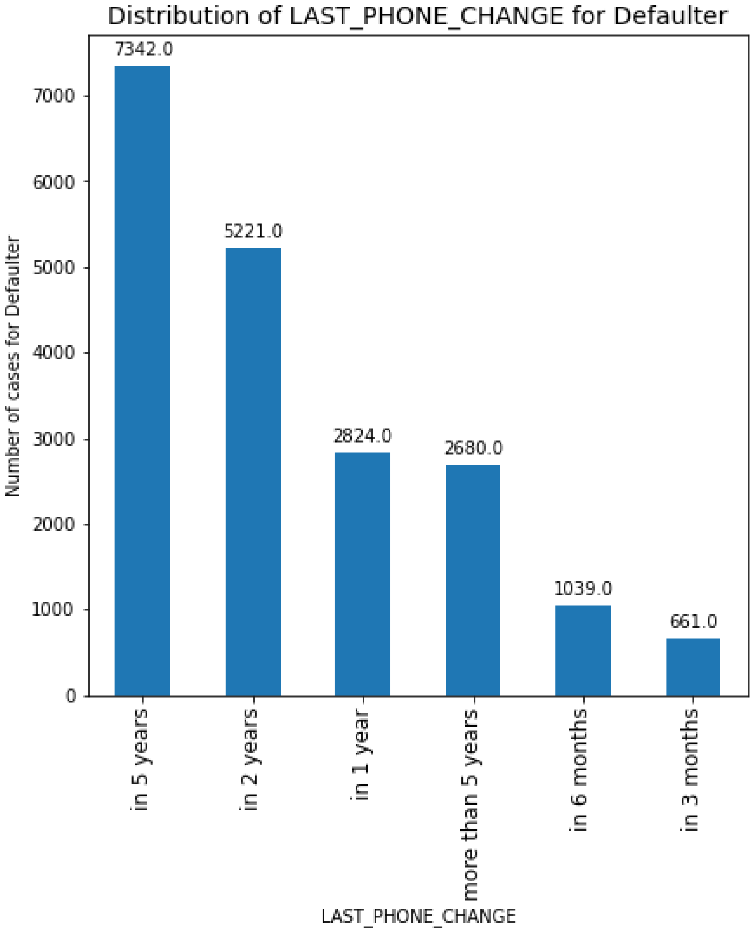




People with 'low' Ratings tends to Default more as compared to non-defaulters. Also they have less 'high' and 'very high' Ratings.

DISTRIBUTION BY LAST PHONE CHANGE

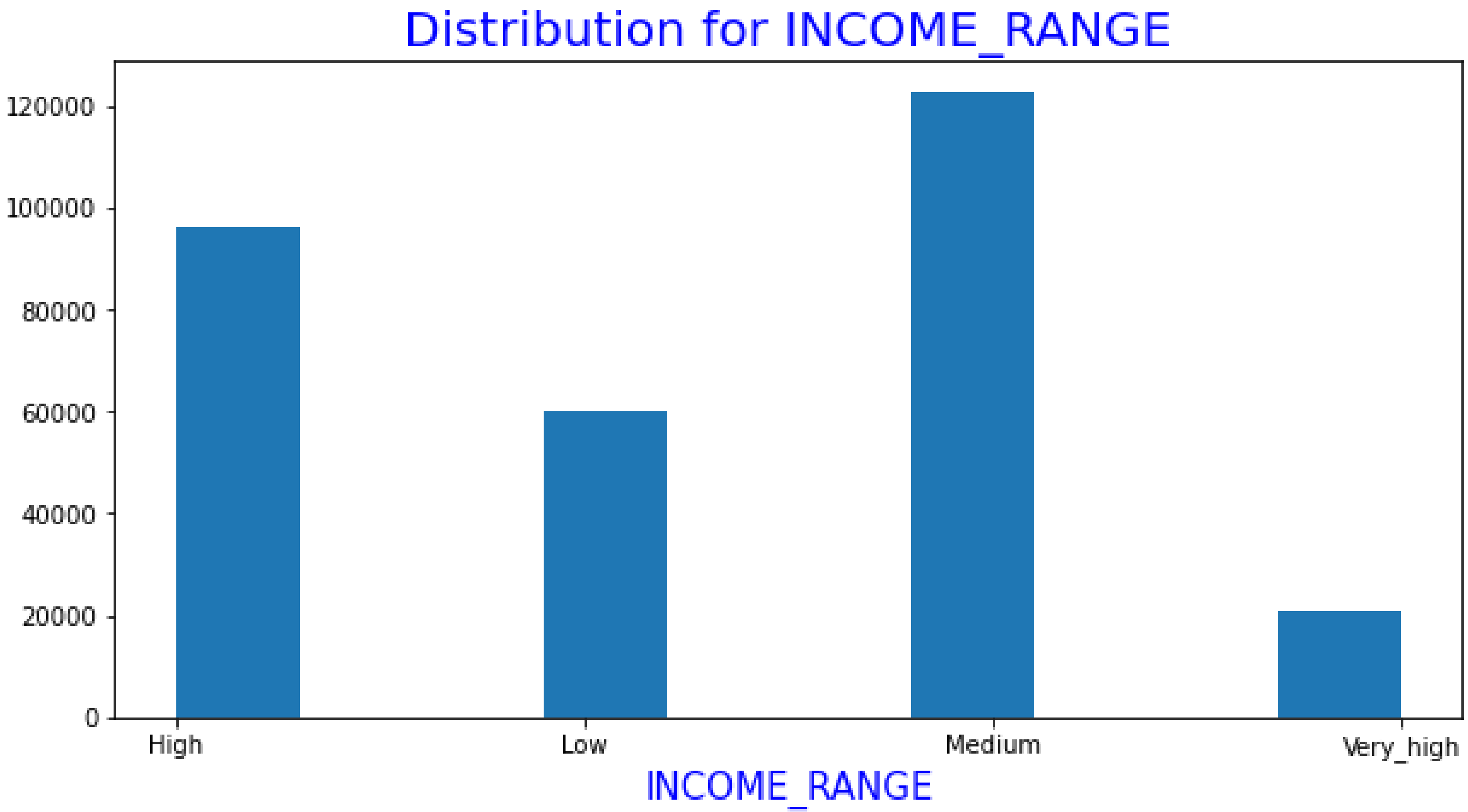




INSIGTHS BY LAST PHONE CHANGE

- This plot clearly displays that customers who have changed their phone in last 2 years in defaulters population are more as compared to customers in non-defaulter population. This demonstrates that customers who are most likely to default, change their phone to avoid repayment of loan. Such customers become difficult to get in contact with for reminders of payments.

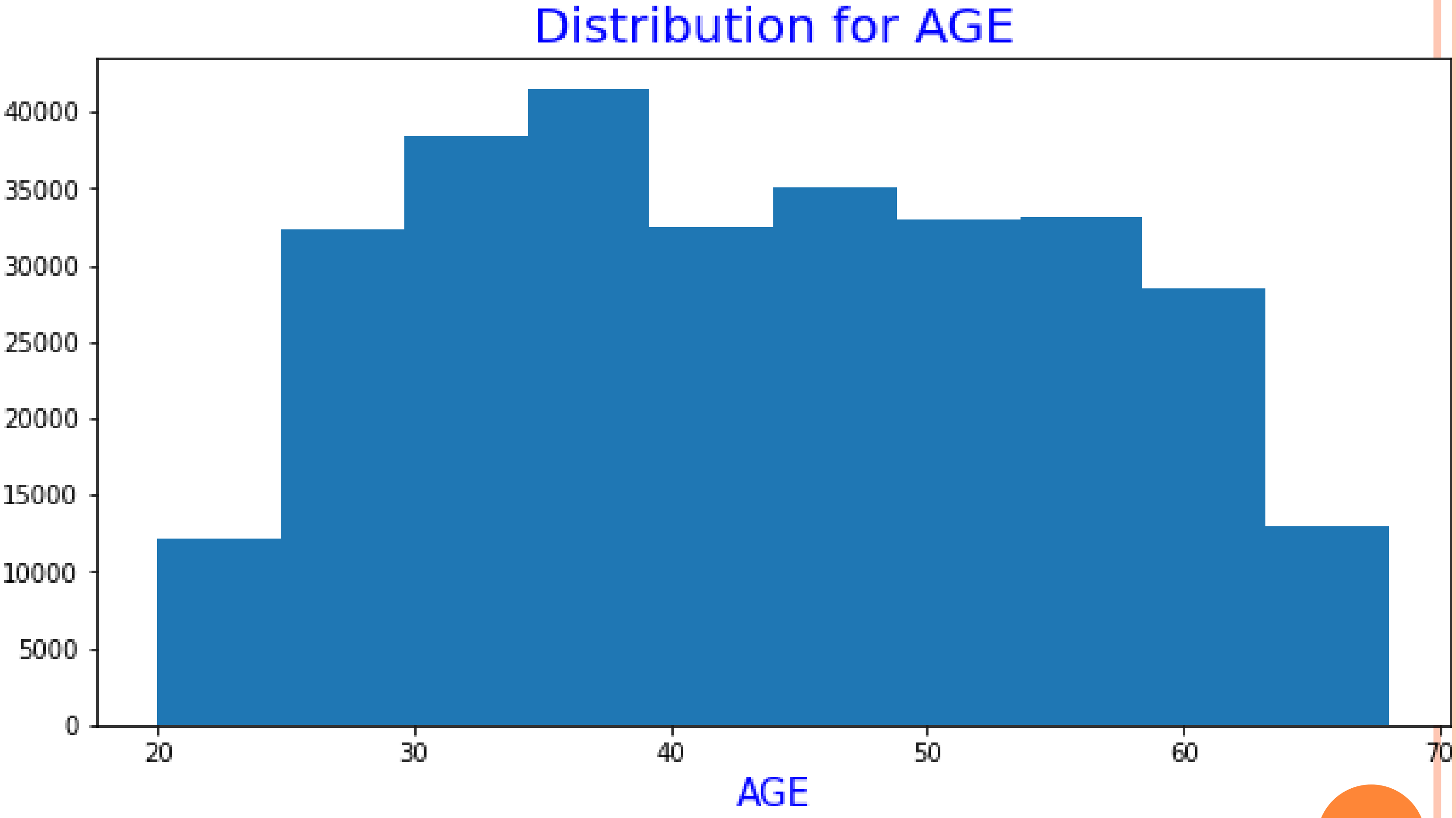
Distribution of income range



Insights from income range plot

- Above histogram insinuates that people with medium income range are more likely to apply for loan. Proportion for 'very high' income is less which can be due to the fact that these people make enough money, hence donor require loan more often. Proportion of people with low income is comparatively less. This may be because they don't earn enough to find the money for loan repayments.

Distribution for age group

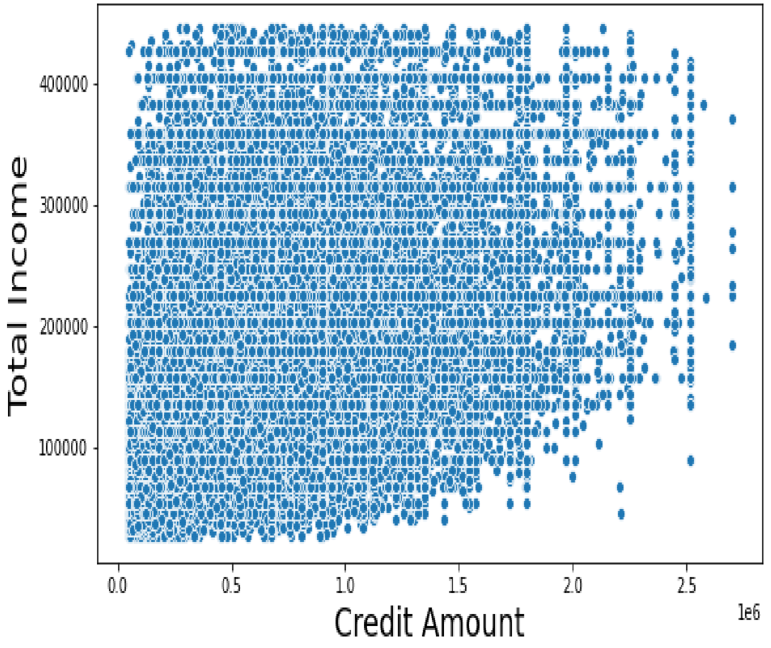


Observation of Age Distribution graph

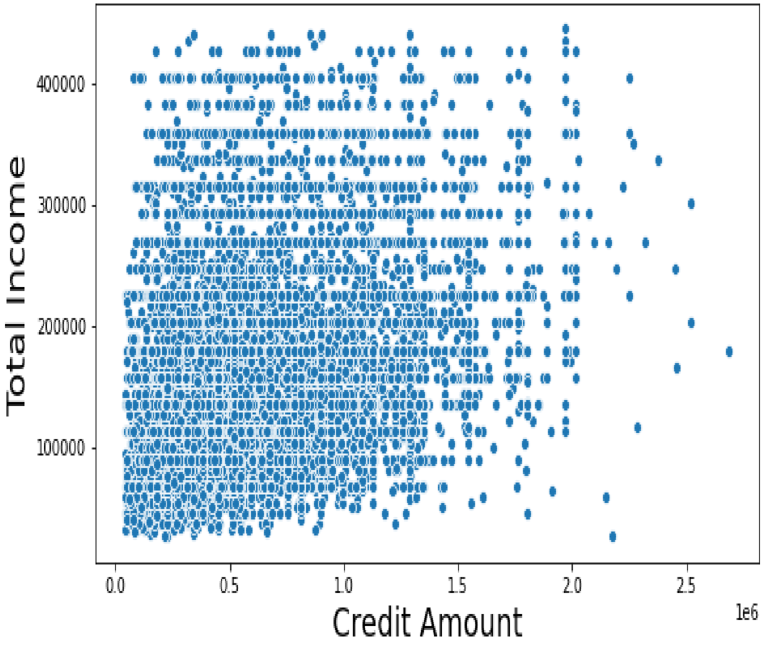
- Majority of people that have applied for loan, belong to age group between 25 and 60. Even in this range, people belonging to age group 30 to 40 are more likely to apply for loan.

Bivariate analysis (Total income vs credit amount)

Credit given against Income for Non-Defaulters



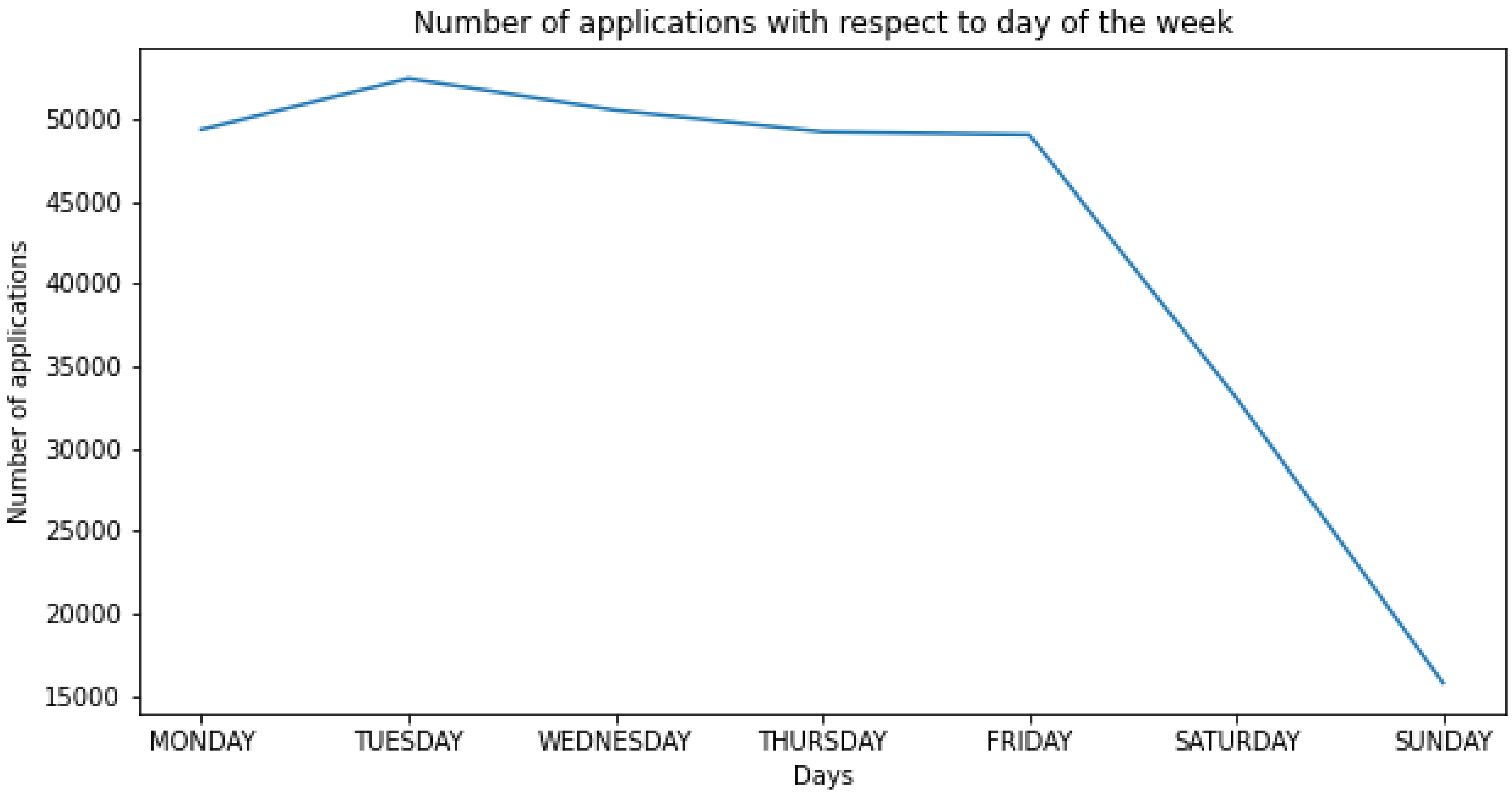
Credit given against Income for Defaulters



Observations of bivariate analysis :

- Cases where loan amount more than 20 lacs provided to defaulters are very less. Whereas in case of non-defaulters, high credit amount is provided to customers with higher total income. This demonstrates that in case of defaulters, higher the requested loan amount, more are the chances of that application to be rejected irrespective of their income.

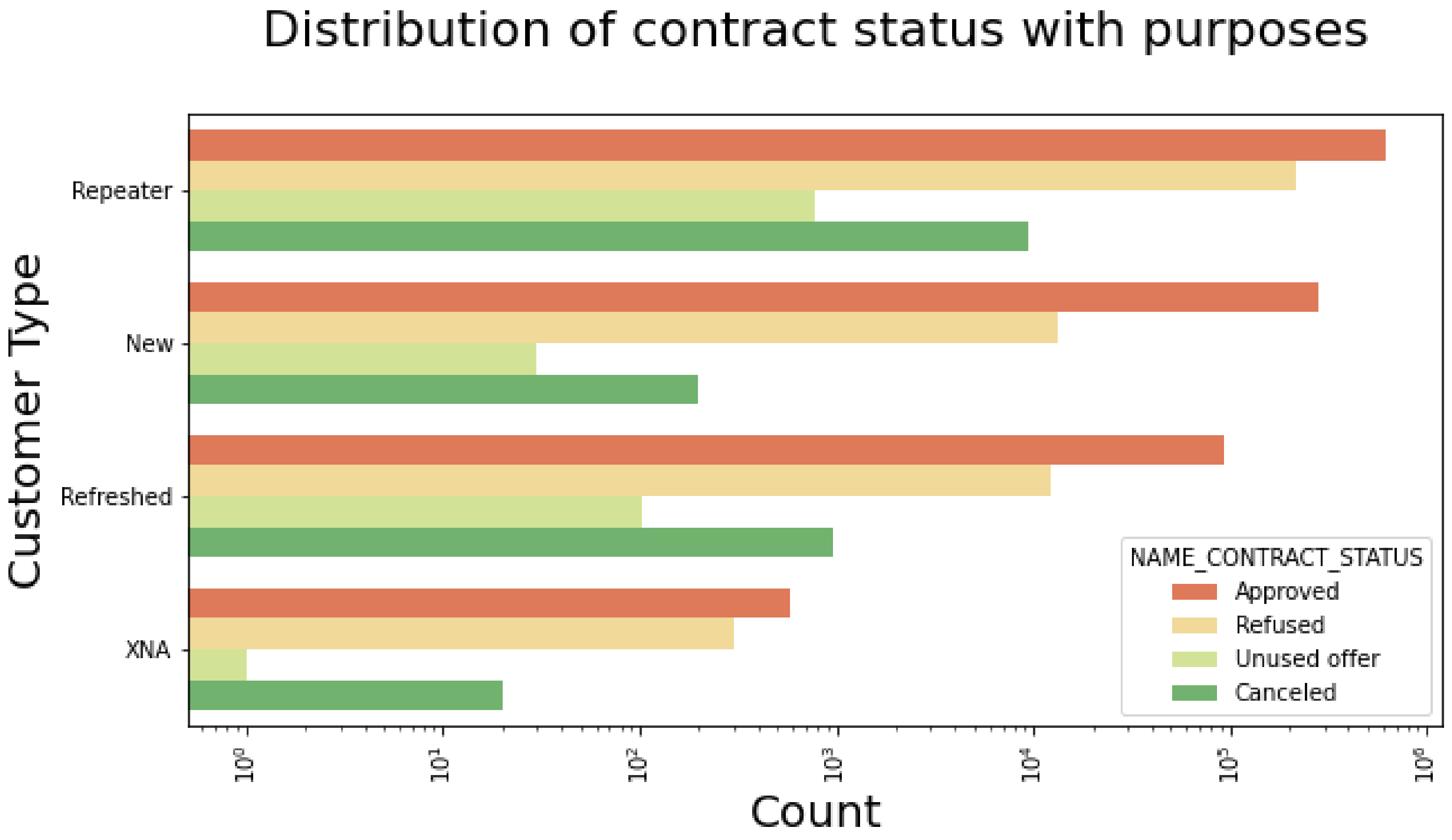
Number of applications each day



Insights from number of applications each day graph

- Majority of applications applied by customers are at the start of the week. Number of applications decreases at the end of the week. People are less likely to apply for loan on Friday, Saturday and Sunday. Hence, executives can approach customers at the start of the week.

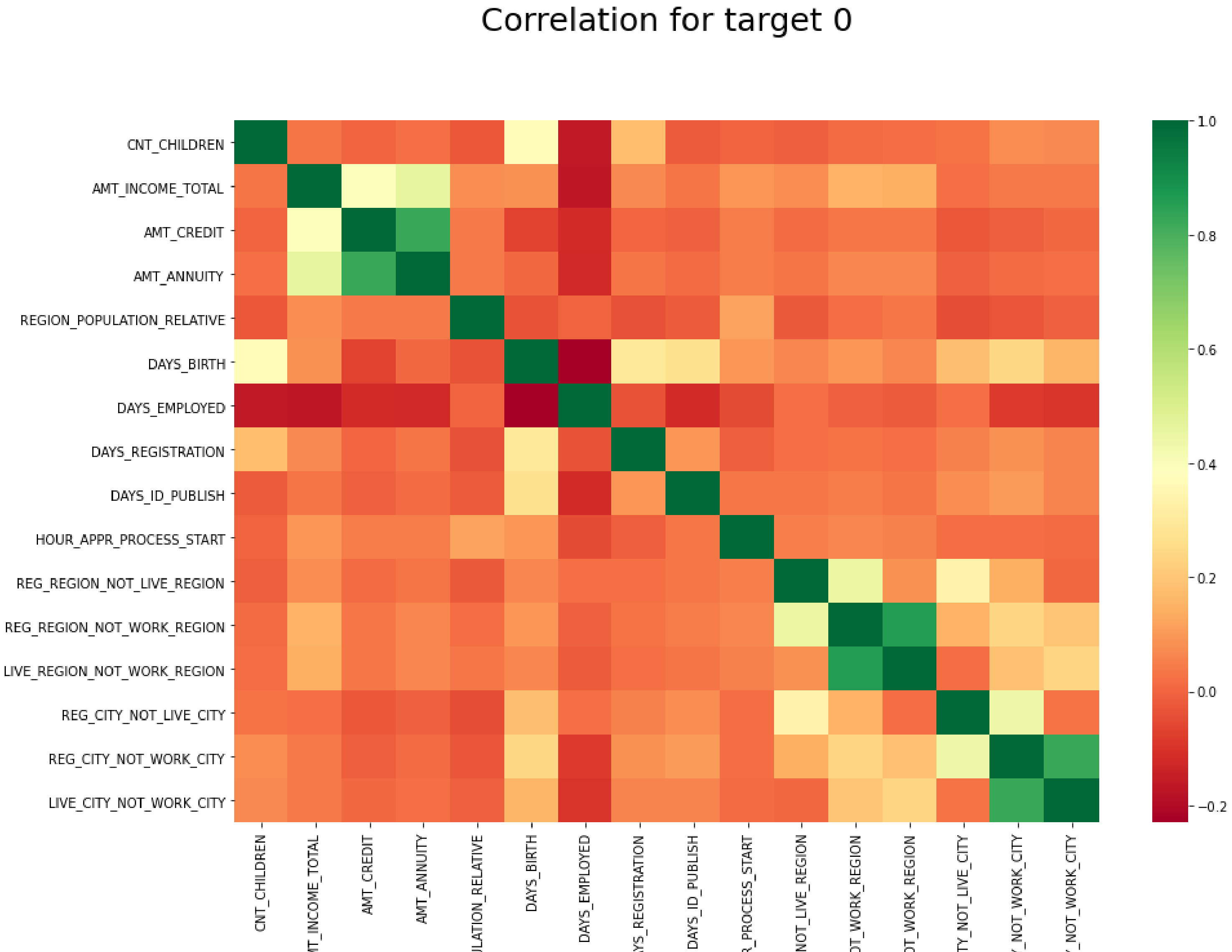
Distribution of Contract Status



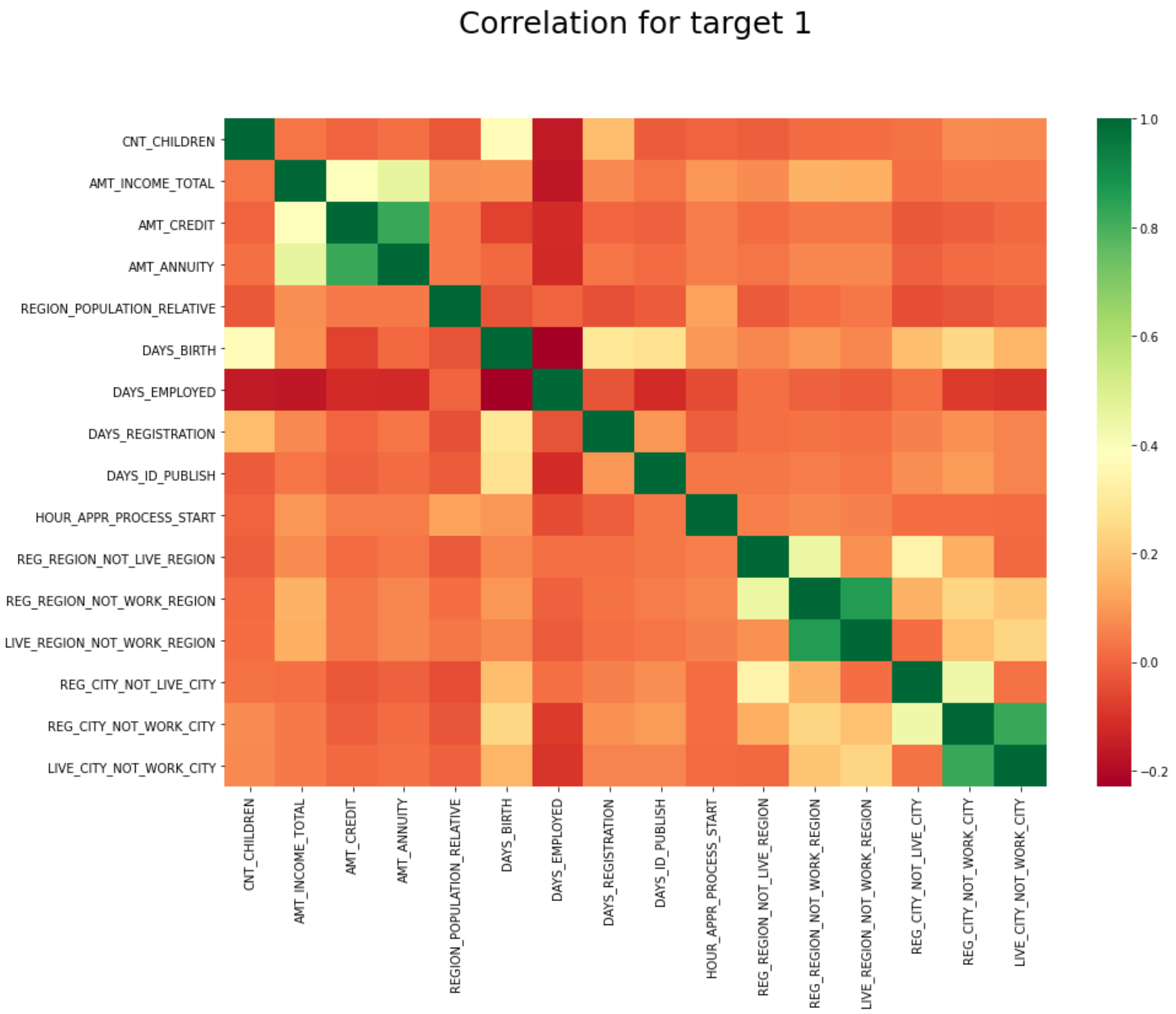
Observations from contract Status

- It seems that Rejection count is greater in case of Repeaters as compared to New customers. It may be because customers who have history of previous application rejections are more likely to get their new application rejected based on reasons for previous rejections. New customers may not have any reasons for rejection as their is no previous history.

Correlation representation of Non Defaulters data using heat map



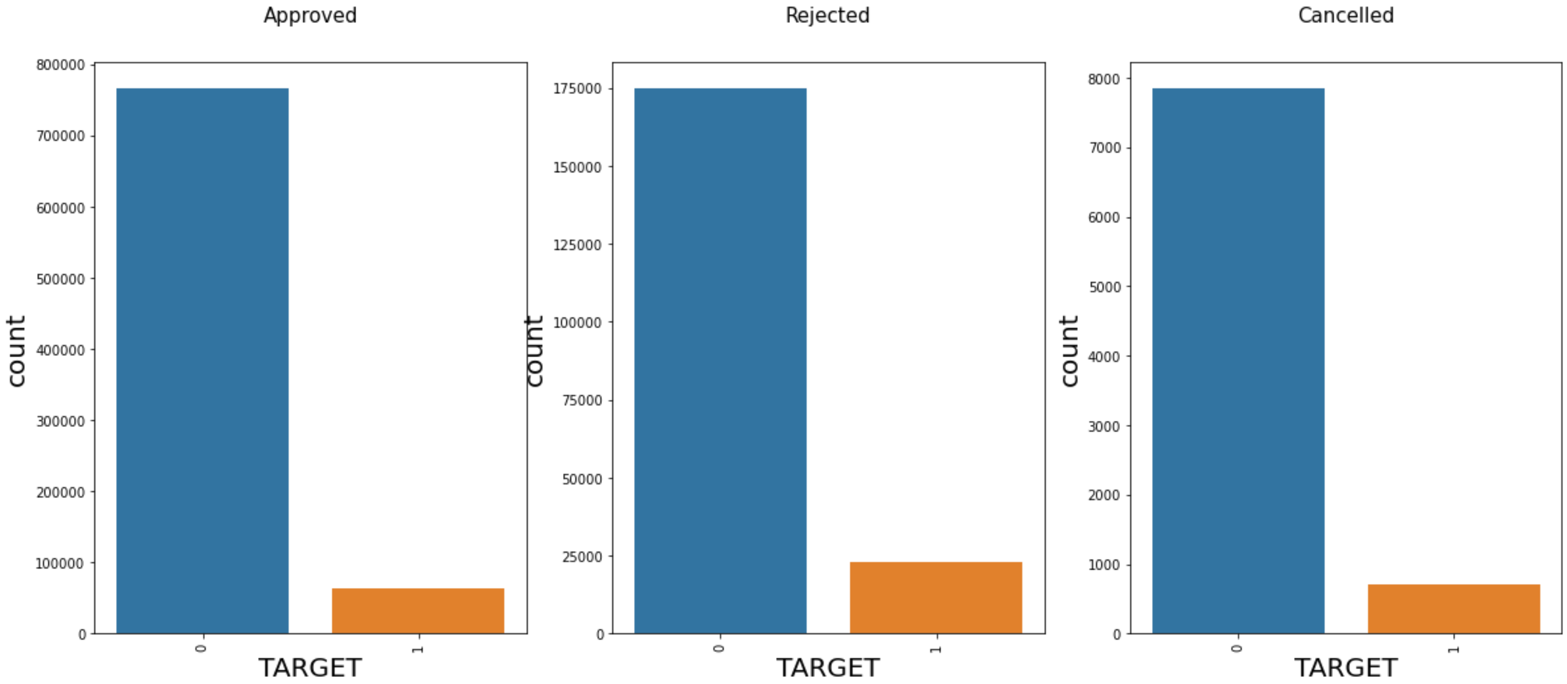
Correlation representation of Defaulters data using heat map



OBSERVATIONS

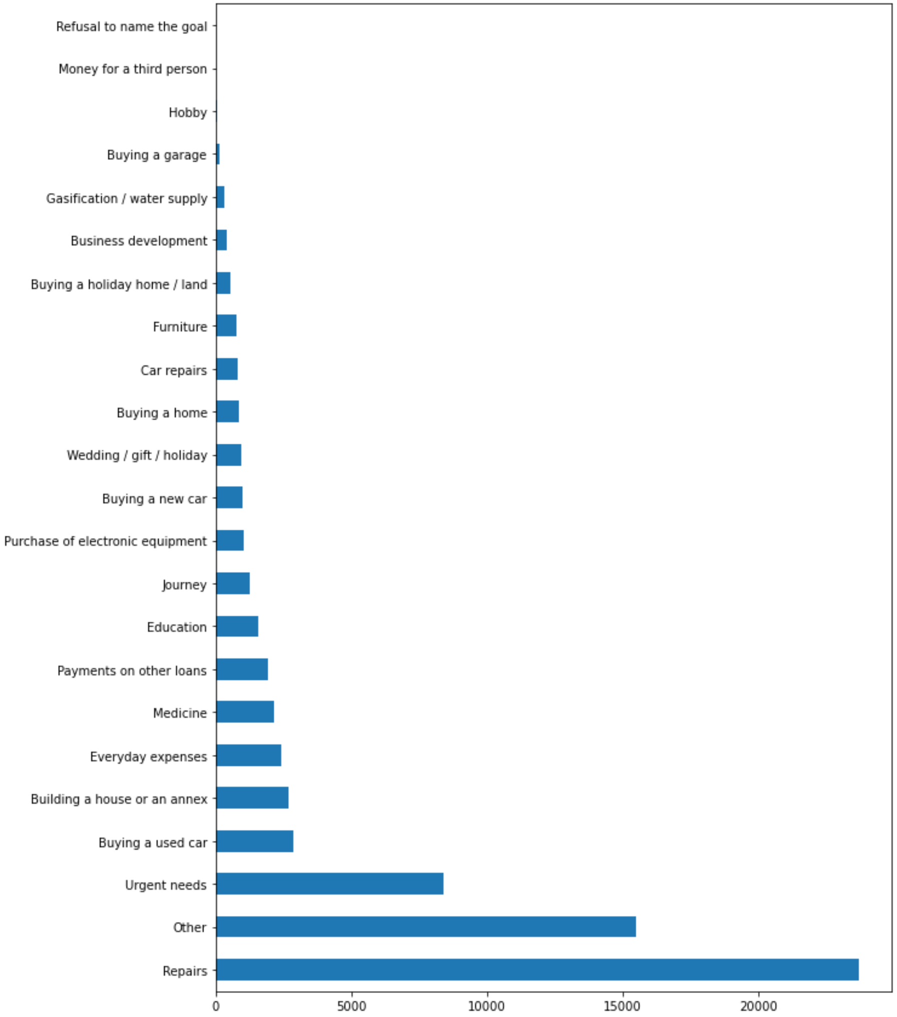
- Credit amount is low for older people, showing inversely proportional relationship. It may be as older people are less likely to be able to make repayment of the loan.
- Credit amount is more for customers with less children count, assuming that such customers can manage to repay the loan easily as they have expenses subjected to one or two children as opposed to those with more children

PREVIOUS APPLICATION DATA



Rejected count is more as compared to Approved count for defaulter (Target 1) customers. It is reasonable considering that these customers are having payment difficulties and are more likely to default.

Distribution By cash loan purpose



Observations of cash loan purpose

- It seems that majority of customers who apply for loan (apart from 'XNA', 'XAP'), have applied for repairs. Top 5 categories for purpose of loan application include buying a car and buying a house and Urgent needs. So executives can approach customers who are looking for buying a car or house.

Conclusions:

- Company should do an excessive background check and history check with credit bureau for customers who are Repeaters for loan application and may have been Rejected earlier by some other company. If such customers are approved for the loan without these checks, then they are more likely to default.
- Customers whose housing type is 'Rented', the company should take their rent into consideration as expense and then calculate income vs expenses analysis before approving their loans.
- Company should target customers with age group between 30 to 50, alse those whose Income falls into 'Medium' or 'High' Income category. Education factor can also be taken into consideration in this. Customers with Higher Education are more likely to repay the loan assuming they earn well.
- Company should look for any previous loans that are open, before approving the loan request. Pervious loan annuity may affect the current loan annuity if current loan is approved. Such customers are high risk customers.
- Company should take Rating from external sources into consideration while making decision. Customers with low rating given by 2 or more sources can prove to be risk to the company as they might default.
- Credit amount should be decided after considering customer's Income, open loans (if any), expenses, age and count of dependent family members.