CREDIT EDA CASE STUDY

BY
POONAM BHONGE

BUSINESS UNDERSTANDING

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.

We have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

BUSINESS OBJECTIVE

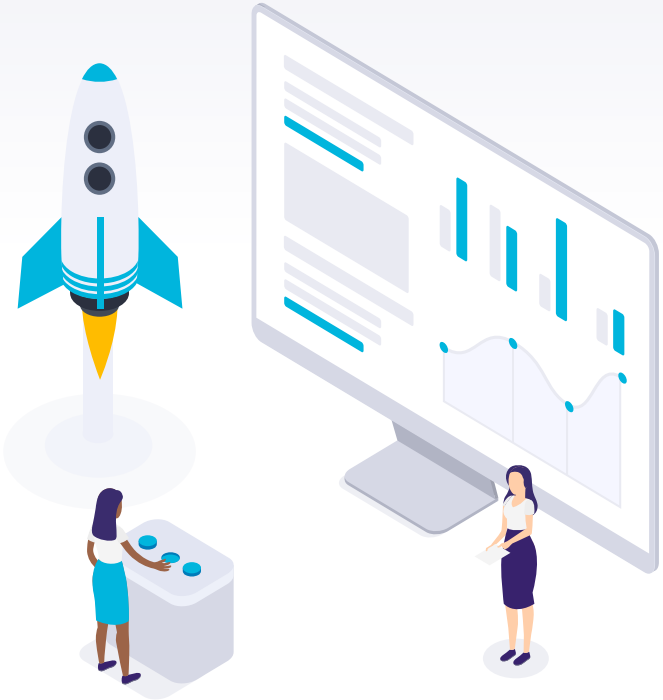
1.Develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

2.To identify patterns which indicate if a client has difficulty paying their instalments 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.

3.Identification of such applicants

4.Company wants to understand the driving factors (or driver variables) behind loan default, i.e., the variables which are strong indicators of default.

AVAILABLE RESOURCES



This dataset has 3 files as explained below:

1.'application_data.csv' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.

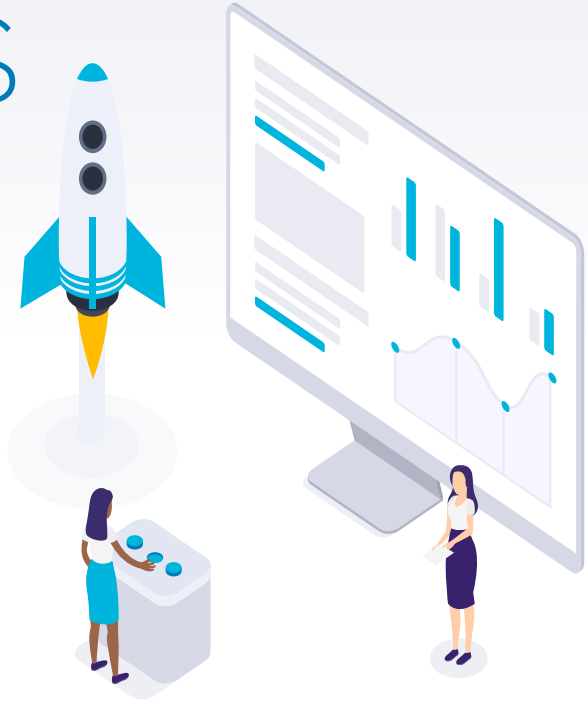
2.'previous_application.csv' contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

3.'columns_description.csv' is data dictionary which describes the meaning of the variables.

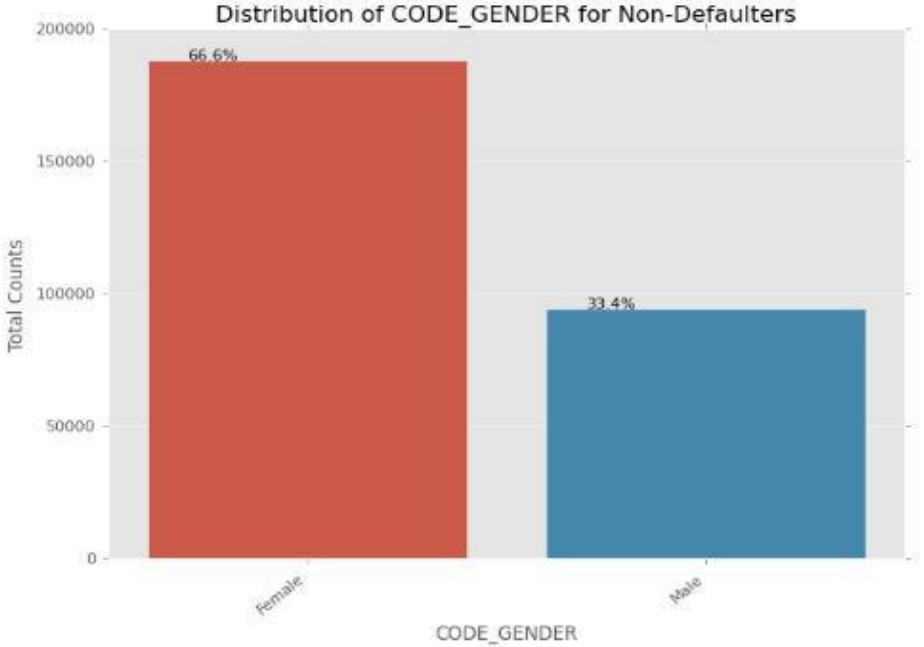
UNIVARIATE ANALYSIS

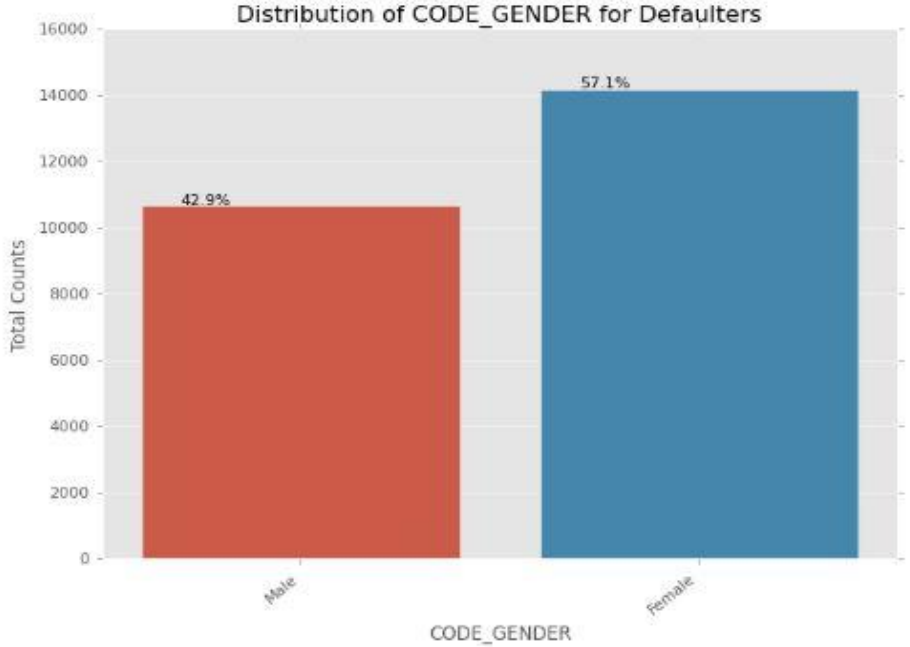
1

UNIVARIATE CATEGORICAL VARIABLES



Distribution of CODE_GENDER for Non-Defaulters & Defaulters

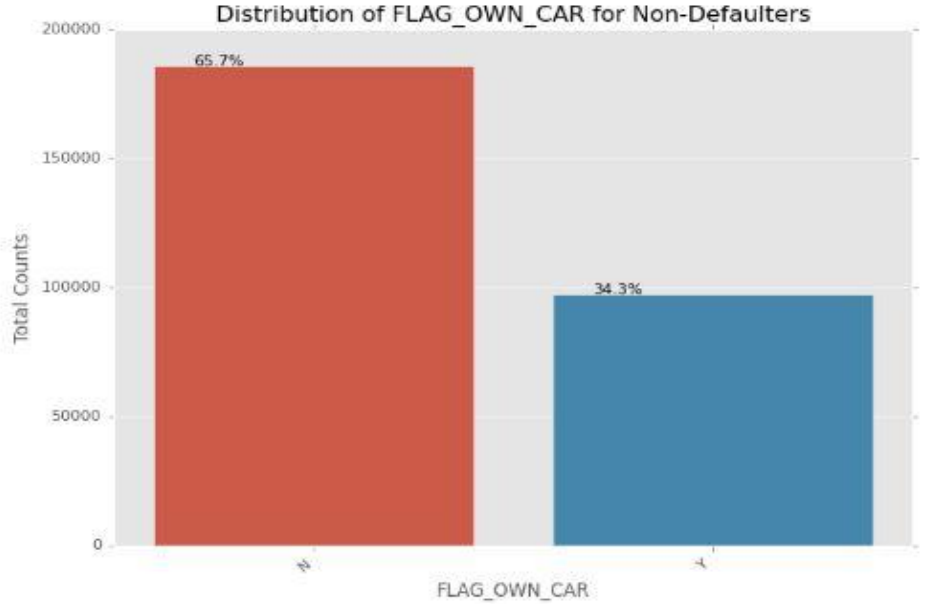


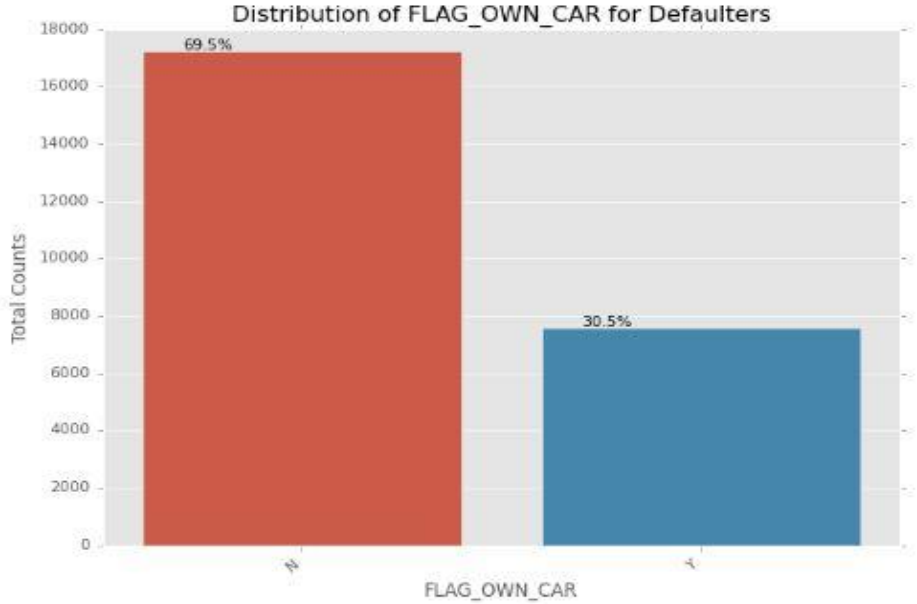


OBSERVATION

- Female contribute 67% to the non-defaulters, while 57% to the defaulters.
- We see more female applying for loans than males and hence large number of female defaulters as well.
- The rate of defaulting of FEMALE is much lower compared to their MALE counterparts.

Distribution of FLAG_OWN_CAR for Non-Defaulters & Defaulters

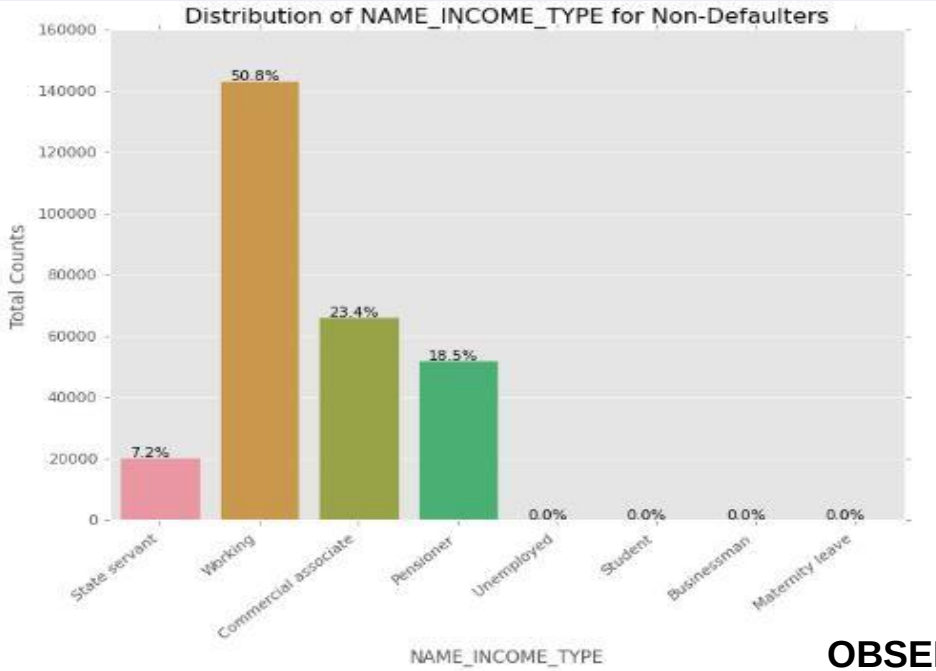


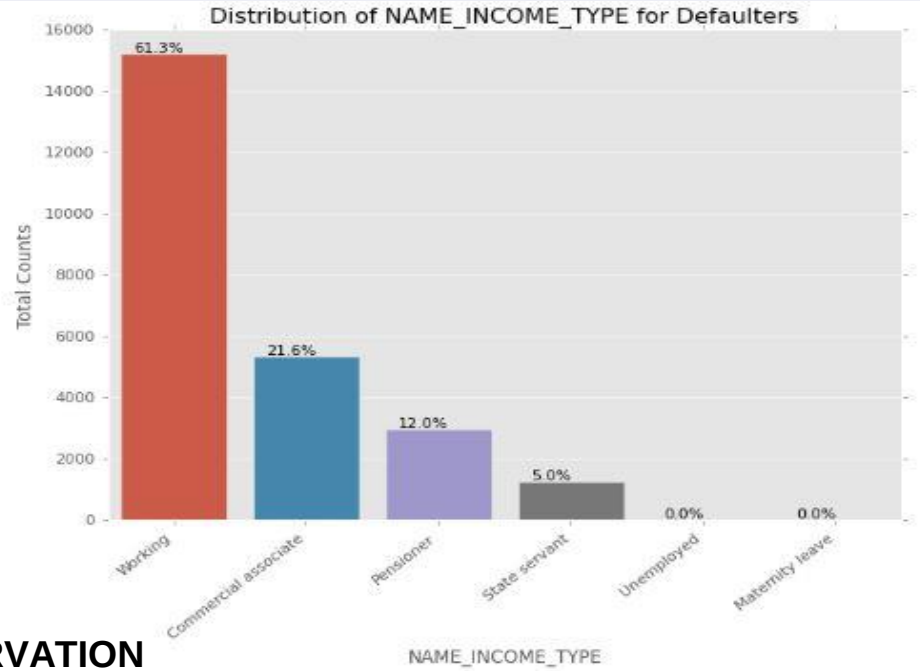


OBSERVATION

- People with cars contribute 65.7% to the non-defaulters while 69.5% to the defaulters.
- We can conclude that, 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 of NAME_INCOME_TYPE for Non-Defaulters & Defaulters

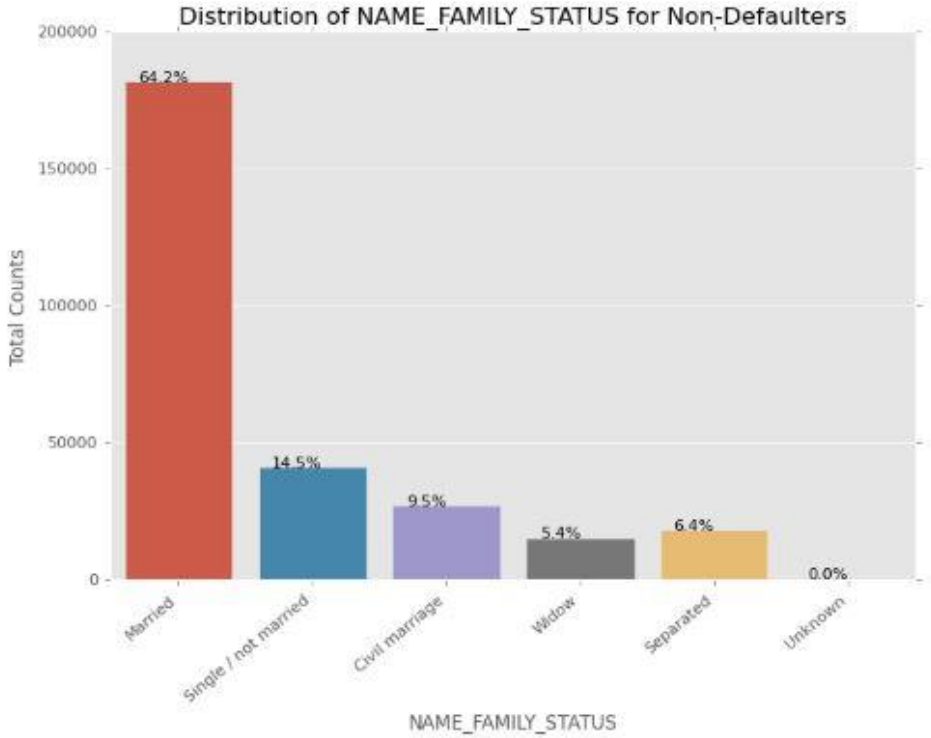


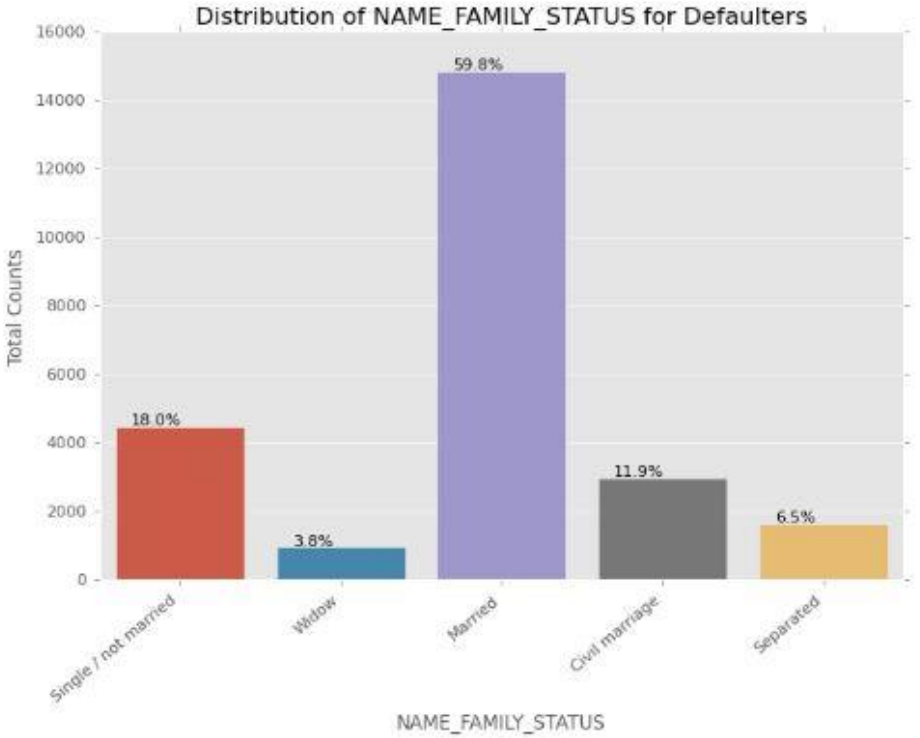


OBSERVATION

- 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.

Distribution of NAME_FAMILY_STATUS for Non-Defaulters & Defaulters

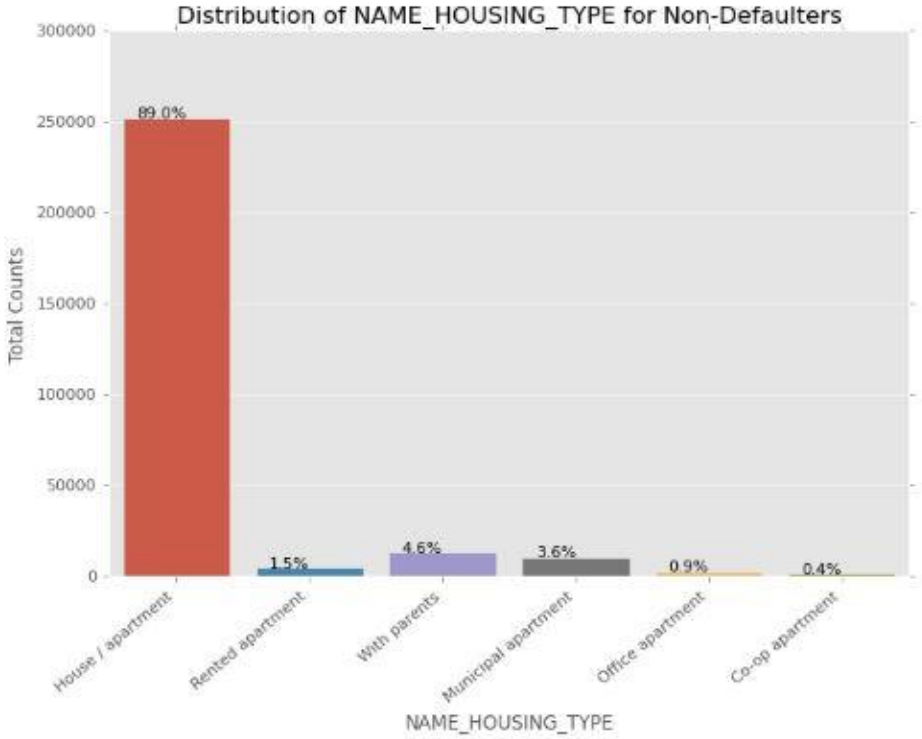


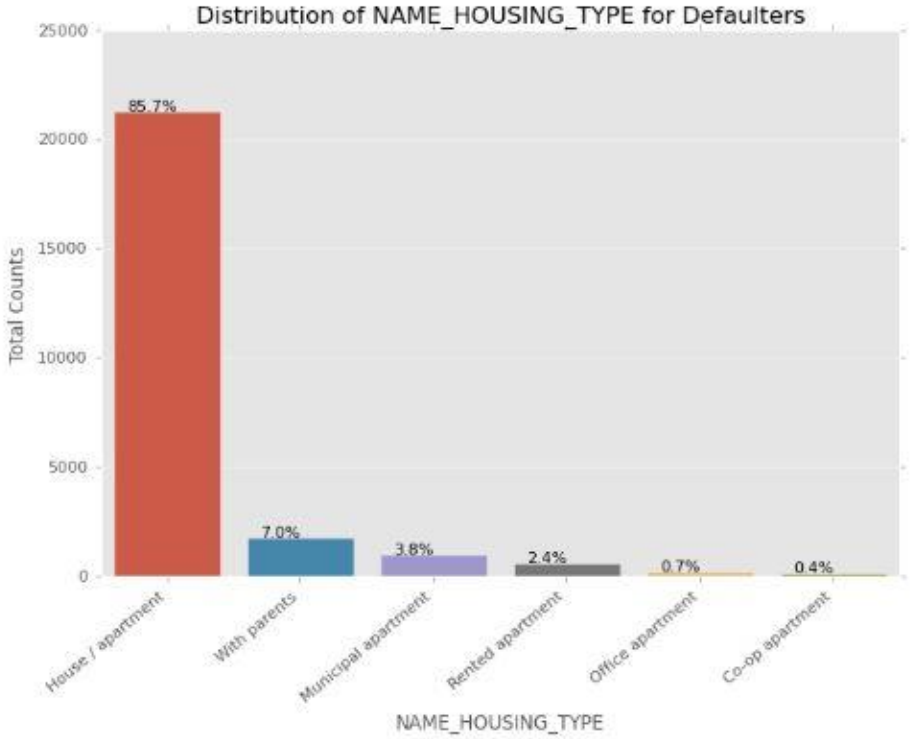


OBSERVATION

- 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.

Distribution of NAME_HOUSING_TYPE for Non-Defaulters & Defaulters

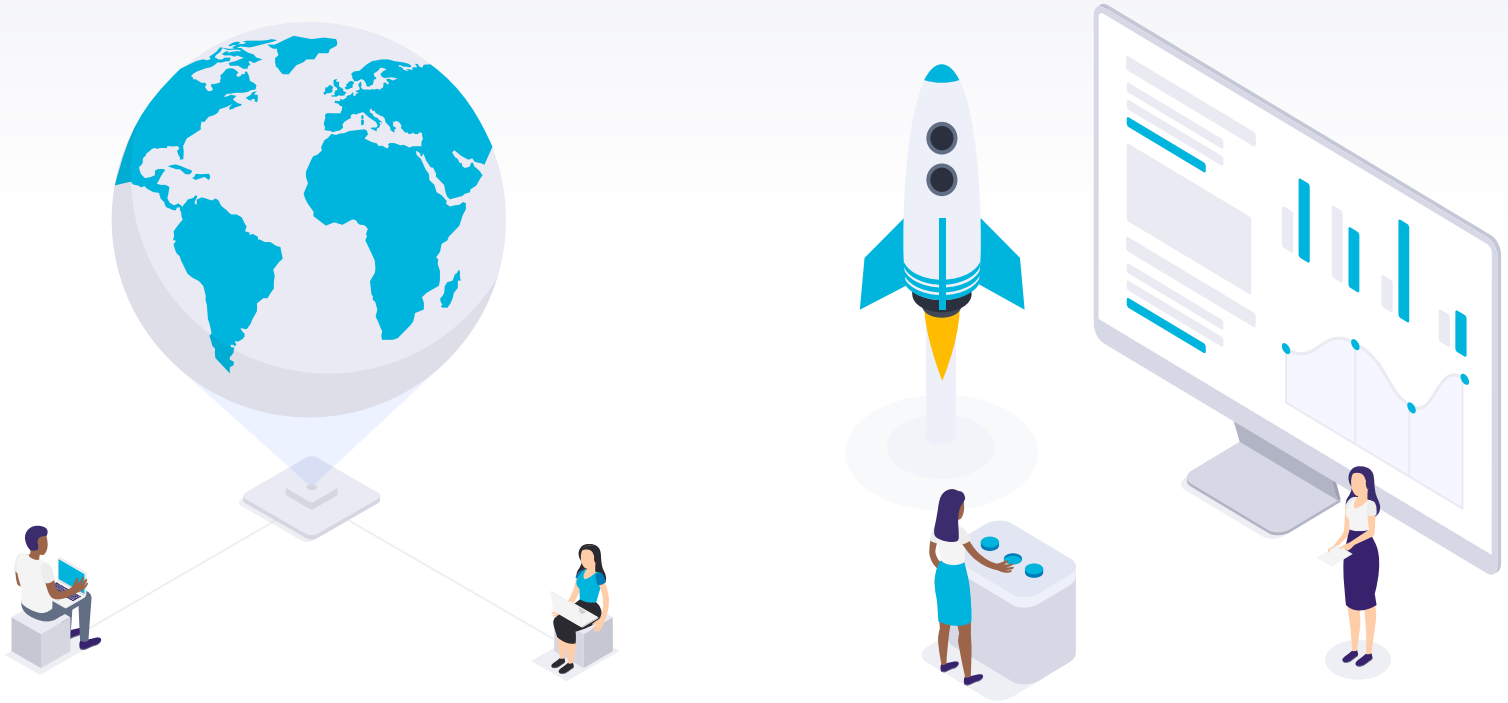




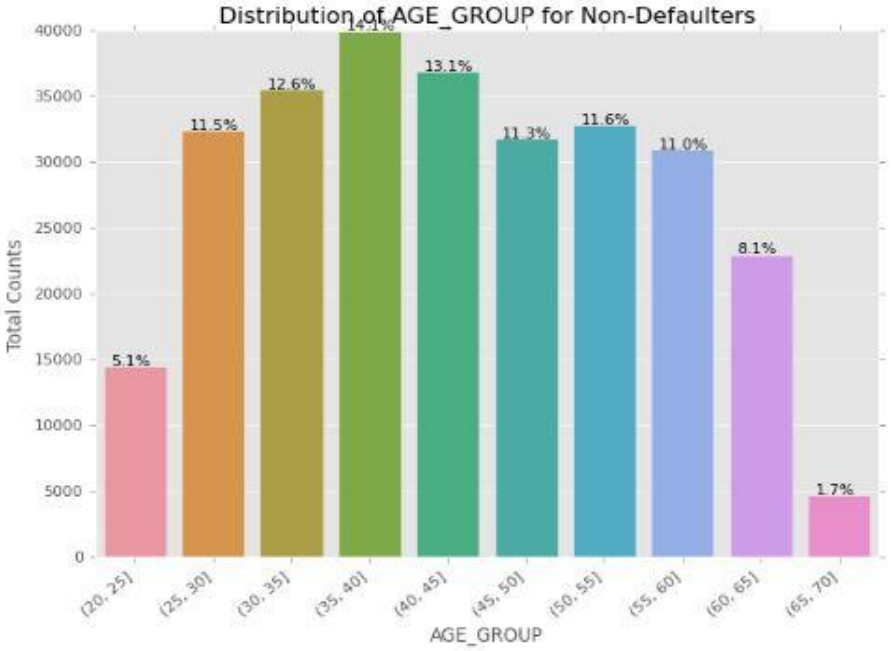
OBSERVATION

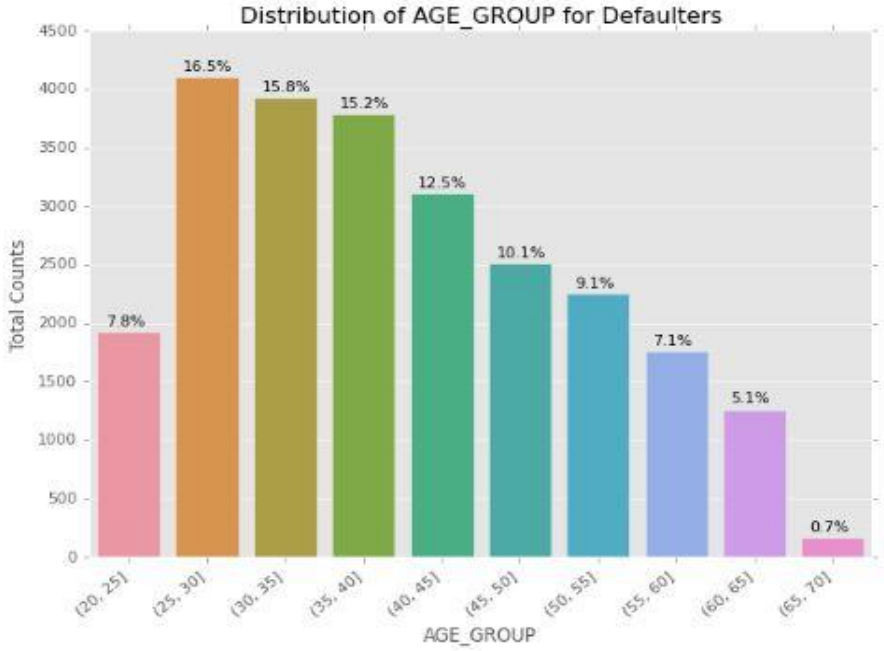
- It is clear from the graph that people who have House/Apartment, tend to apply for more loans.
- People living with parents tend to default more often when compared with others. The reason could be their living expenses are more due to their parents living with them.

Univariate Categorical Ordered Analysis



Distribution of AGE_GROUP for Non-Defaulters & Defaulters

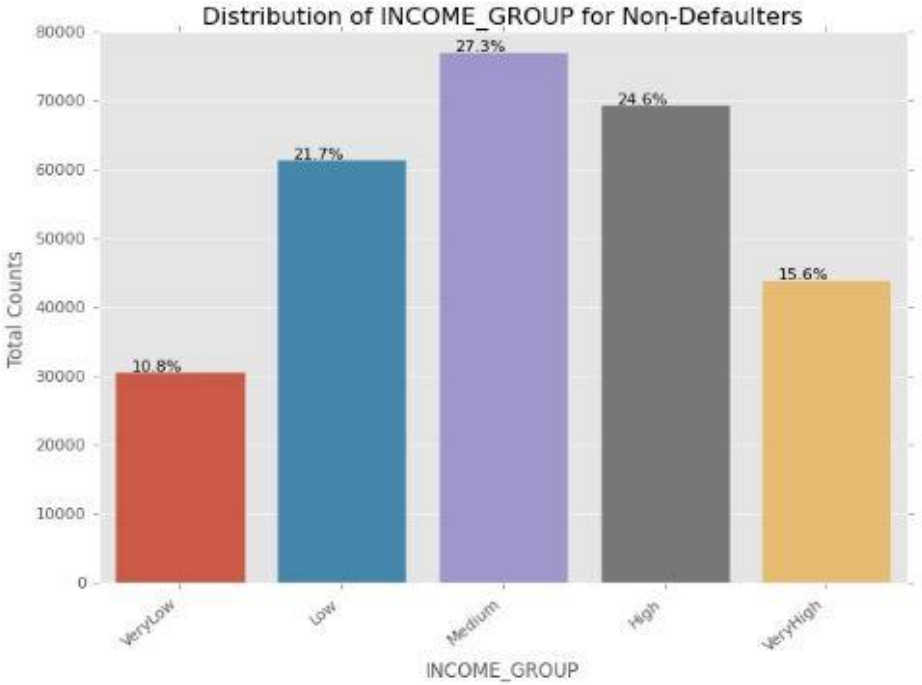


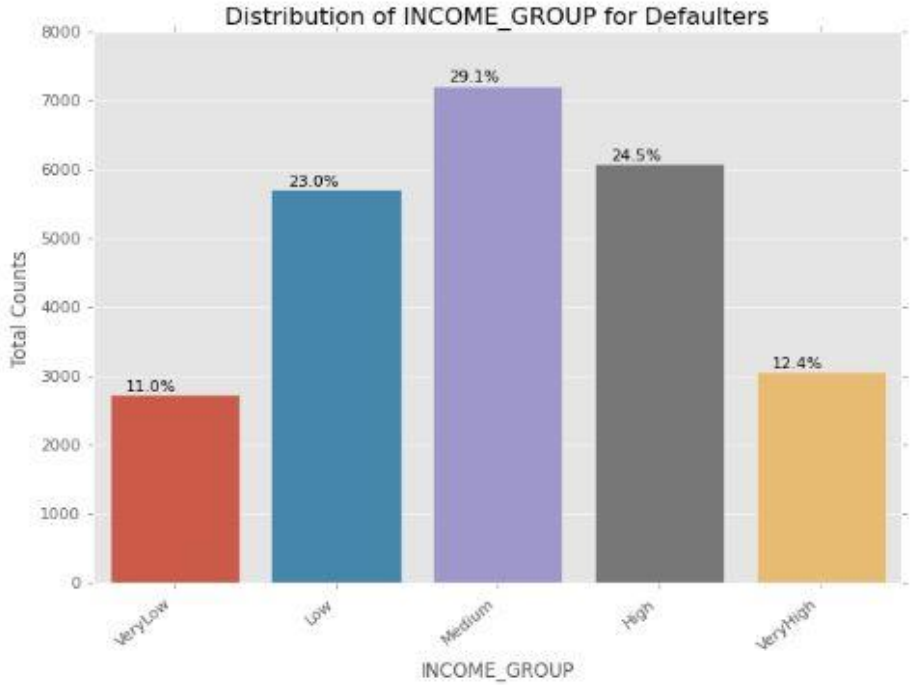


OBSERVATION

- We see that (25,30) age group tend to default more often. So, they are the riskiest people to loan.
- 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.

Distribution of INCOME_GROUP for Non-Defaulters & Defaulters

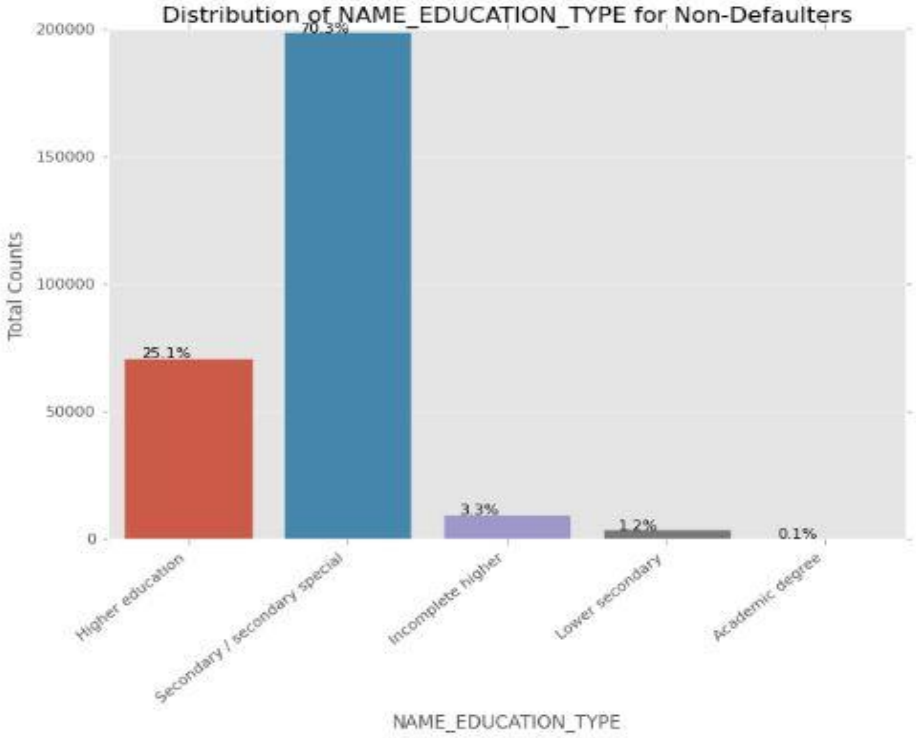


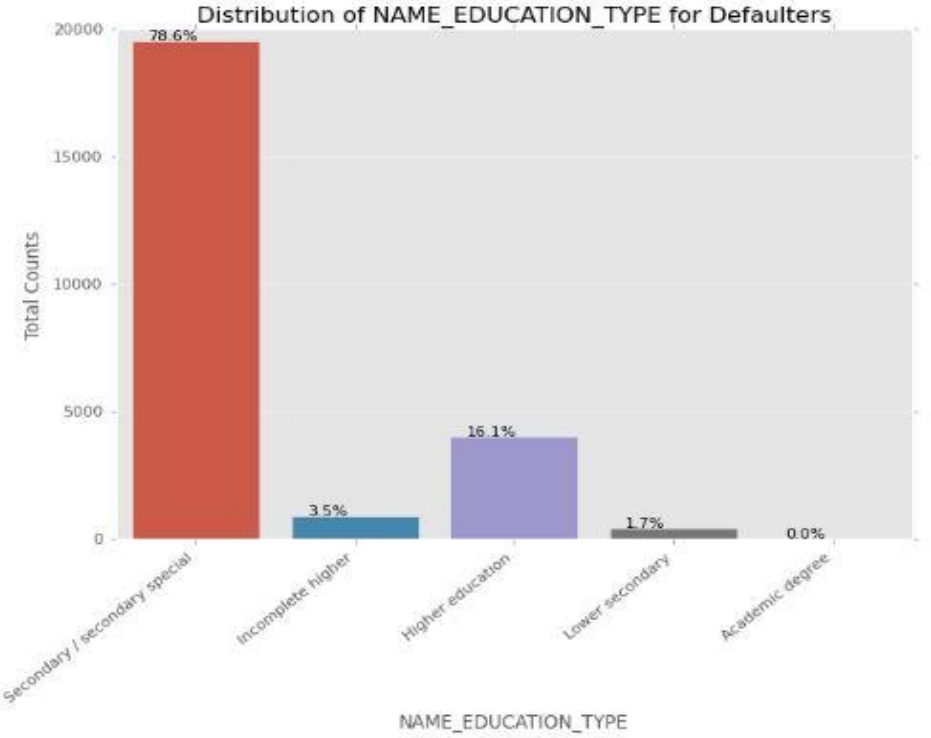


OBSERVATION

- 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.

Distribution of NAME_EDUCATION_TYPE for Non-Defaulters & Defaulters

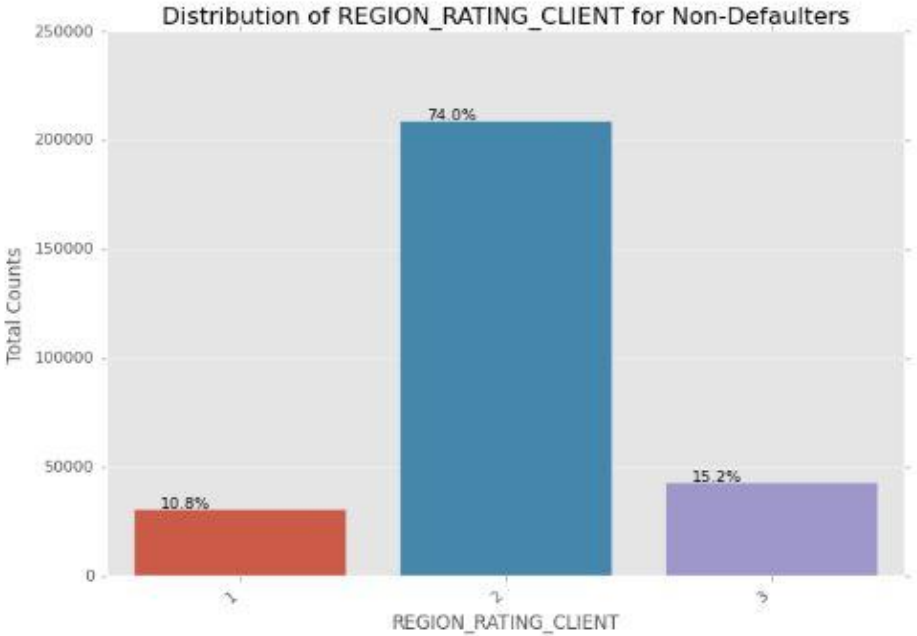


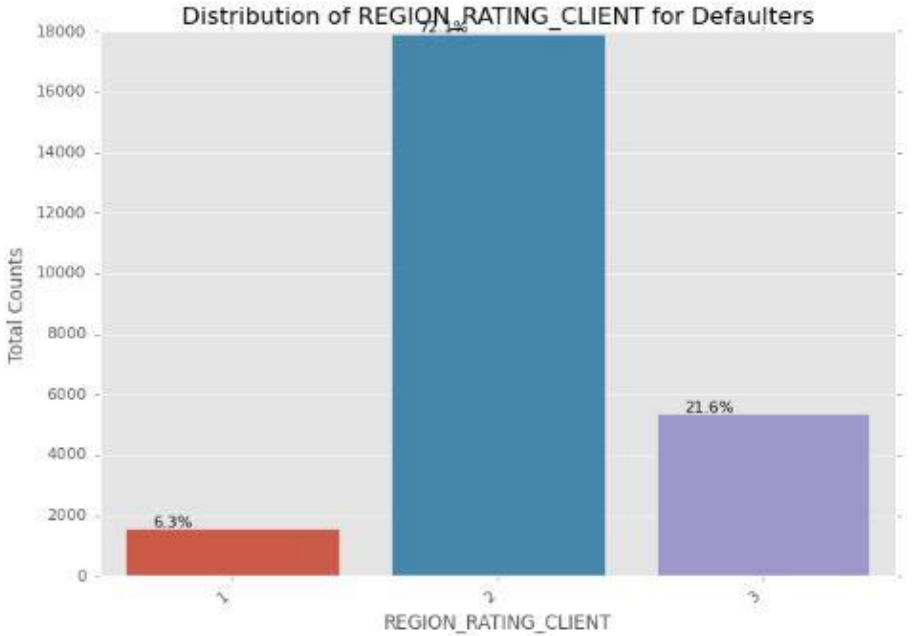


OBSERVATION

- 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

Distribution of REGION_RATING_CLIENT for Non-Defaulters & Defaulters





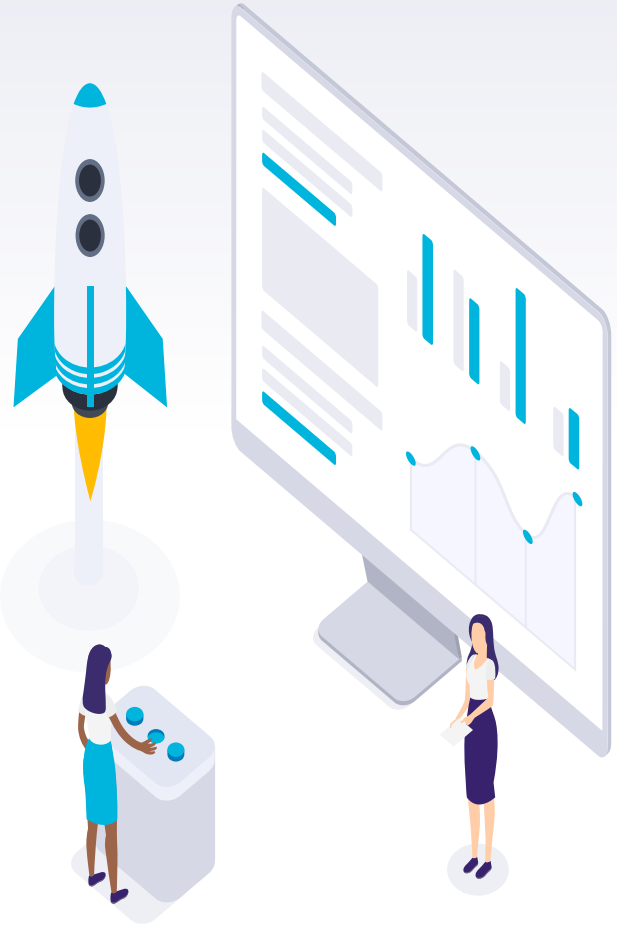
OBSERVATION

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 likely are less defaulters.

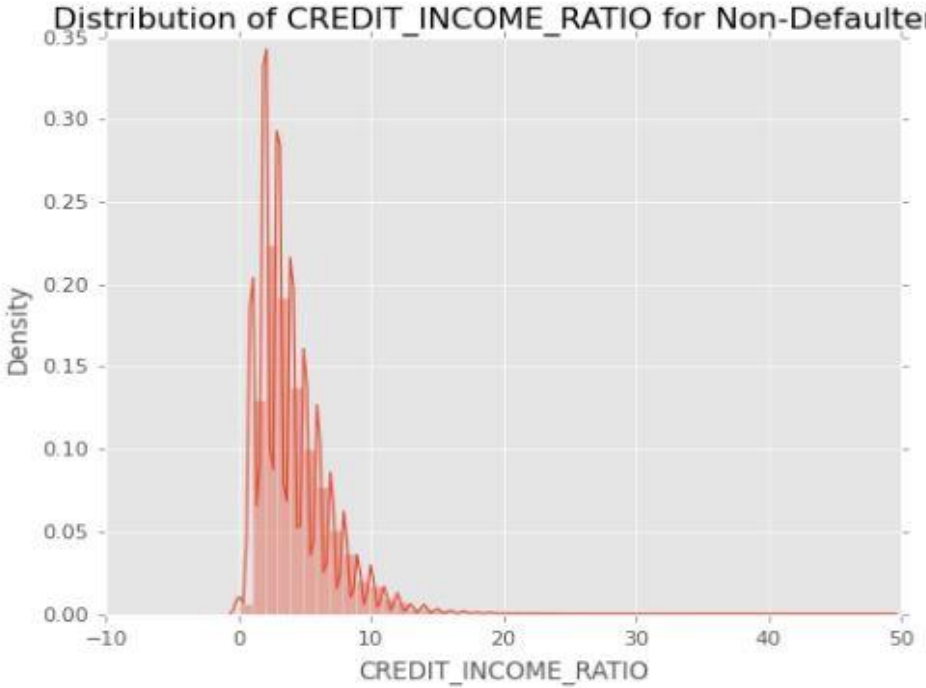
UNIVARIATE ANALYSIS

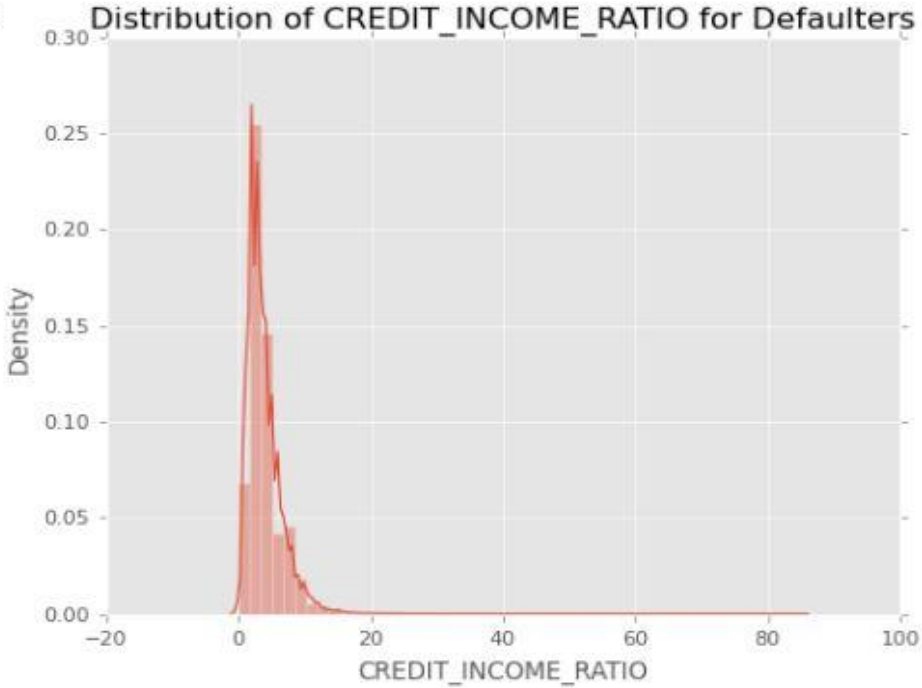
2

Univariate continuous variable analysis



Distribution of CREDIT_INCOME_RATIO for Non-Defaulters & Defaulters

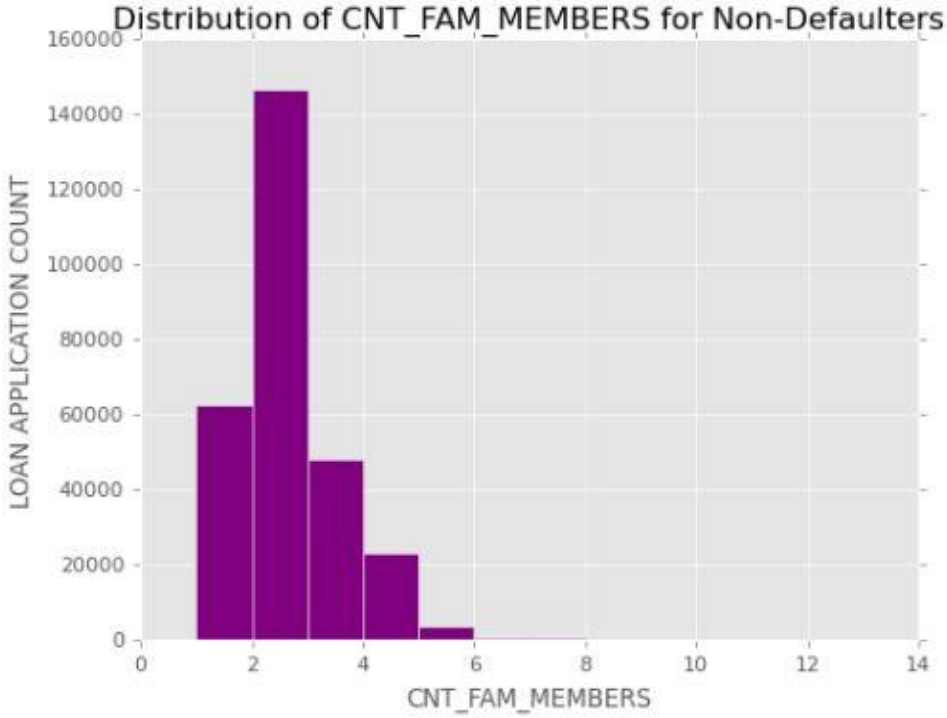


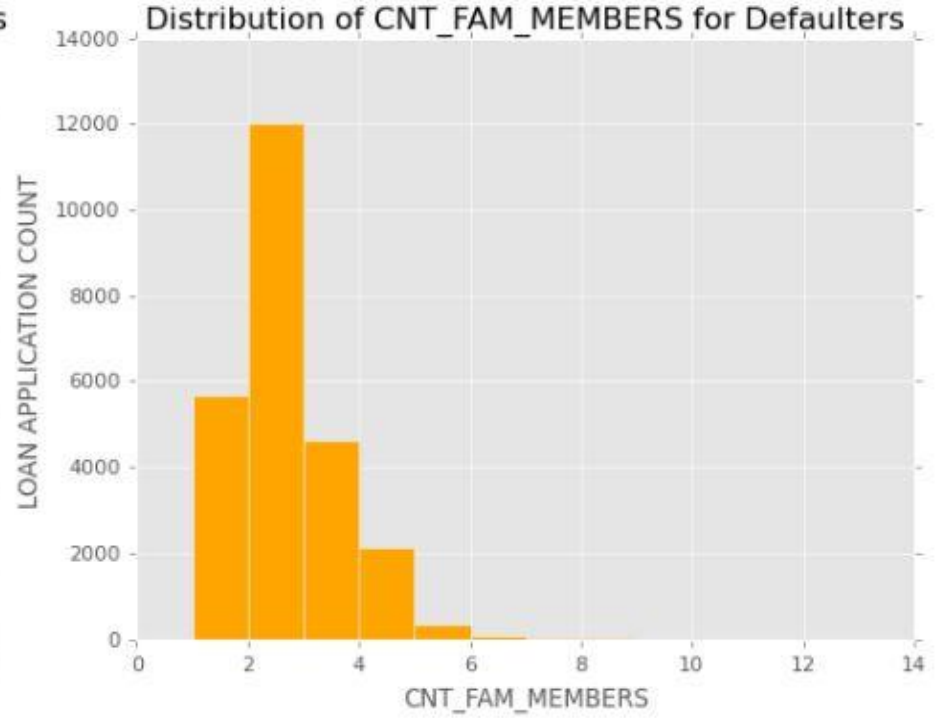


OBSERVATION

- Credit income ratio is the ratio of AMT_CREDIT/AMT_INCOME_TOTAL. Although 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.

Distribution of CNT_FAM_MEMBERS for Non-Defaulters & Defaulters





OBSERVATION

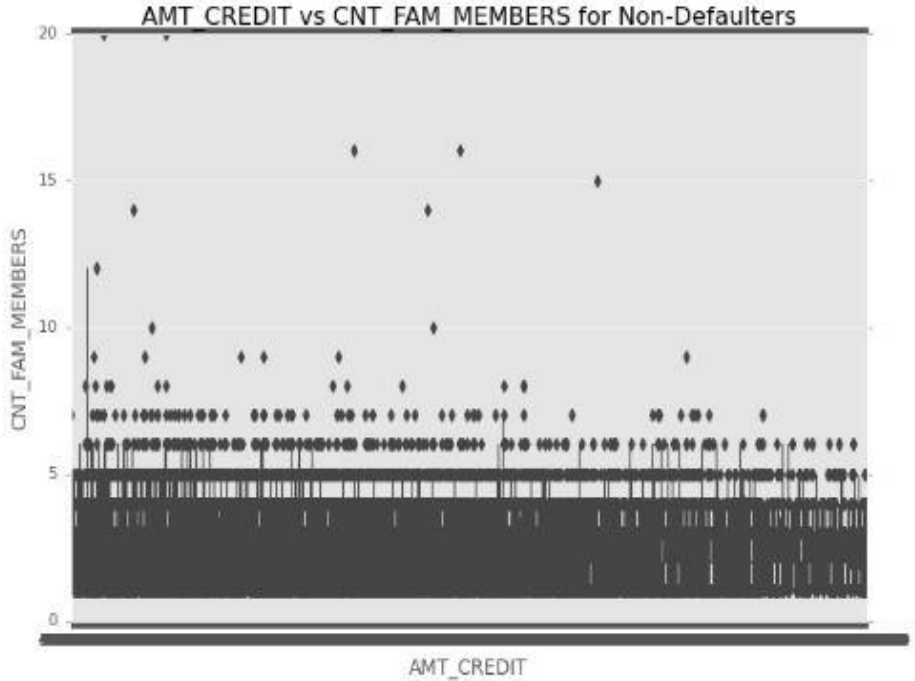
- We can see that a family of 3 applies loan more often than the other families

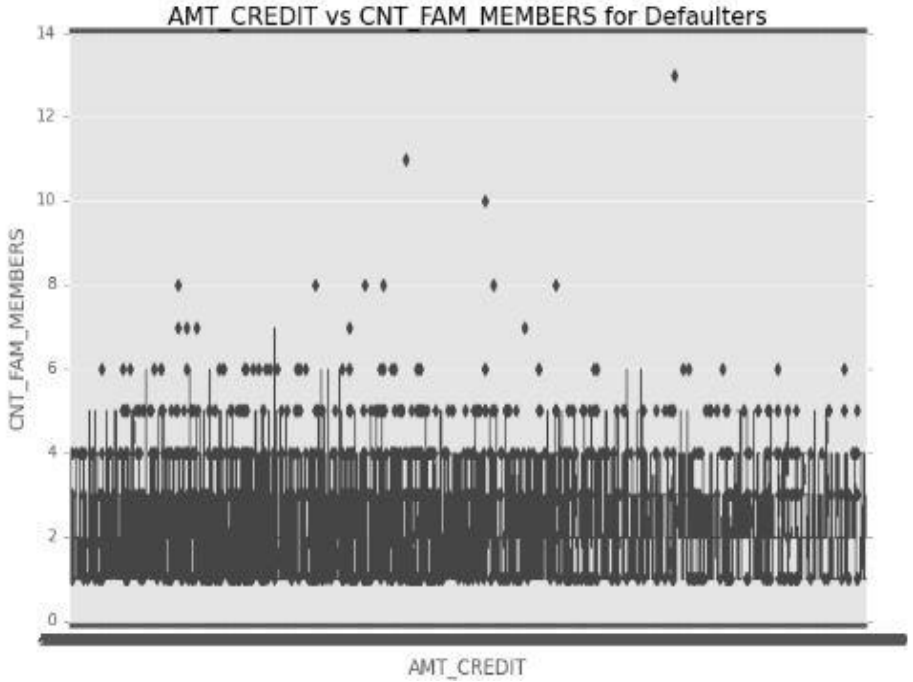
BIVARIATE ANALYSIS

2

Bivariate Analysis of numerical variables

AMT CREDIT Vs FAM MEMBERS for Non-Defaulters & Defaulters





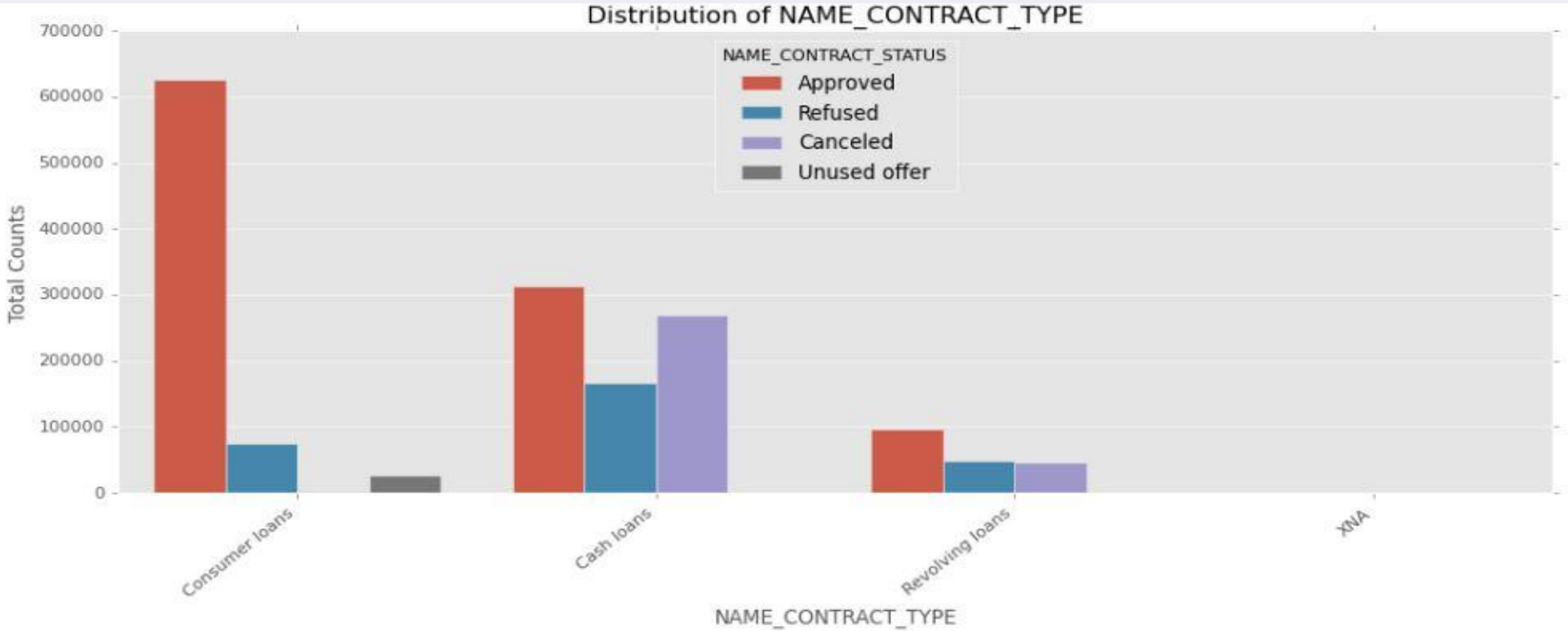
OBSERVATION

- 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

Data Analysis For Previous Application Data

Univariate Analysis

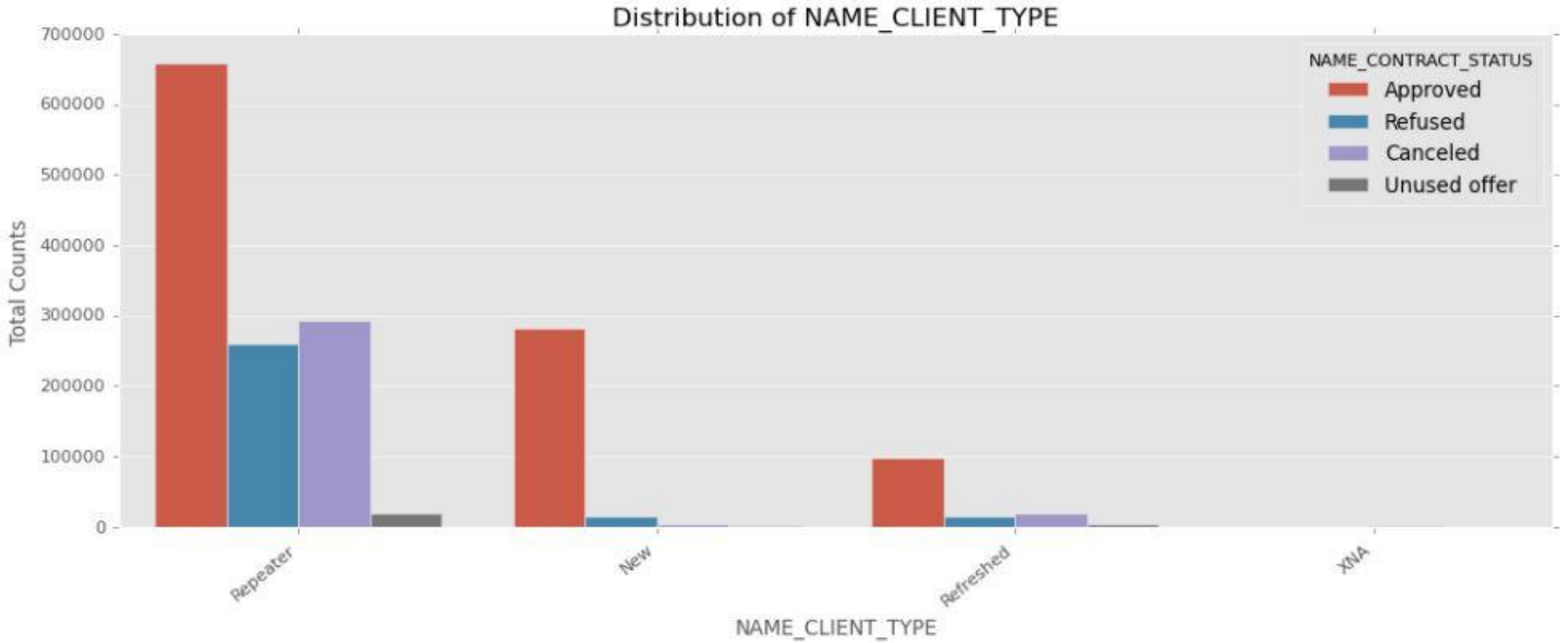
Distribution of NAME_CONTRCT_TYPE



OBSERVATION

- 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.

Distribution of NAME_CONTRCT_TYPE

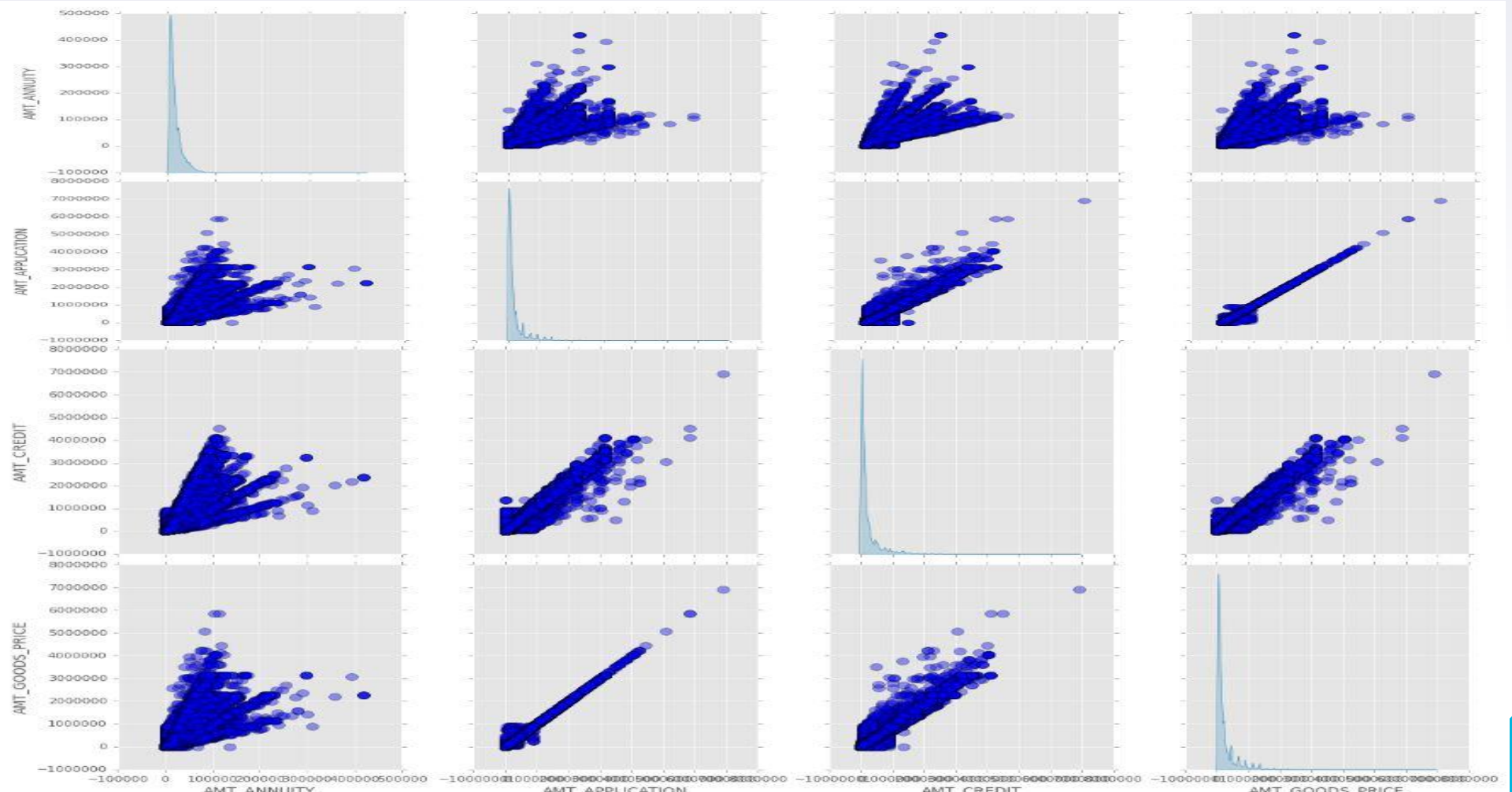


OBSERVATION

- 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

Bivariate Analysis

Plotting the relation between correlated highly corelated numeric variables



OBSERVATION

1. 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.

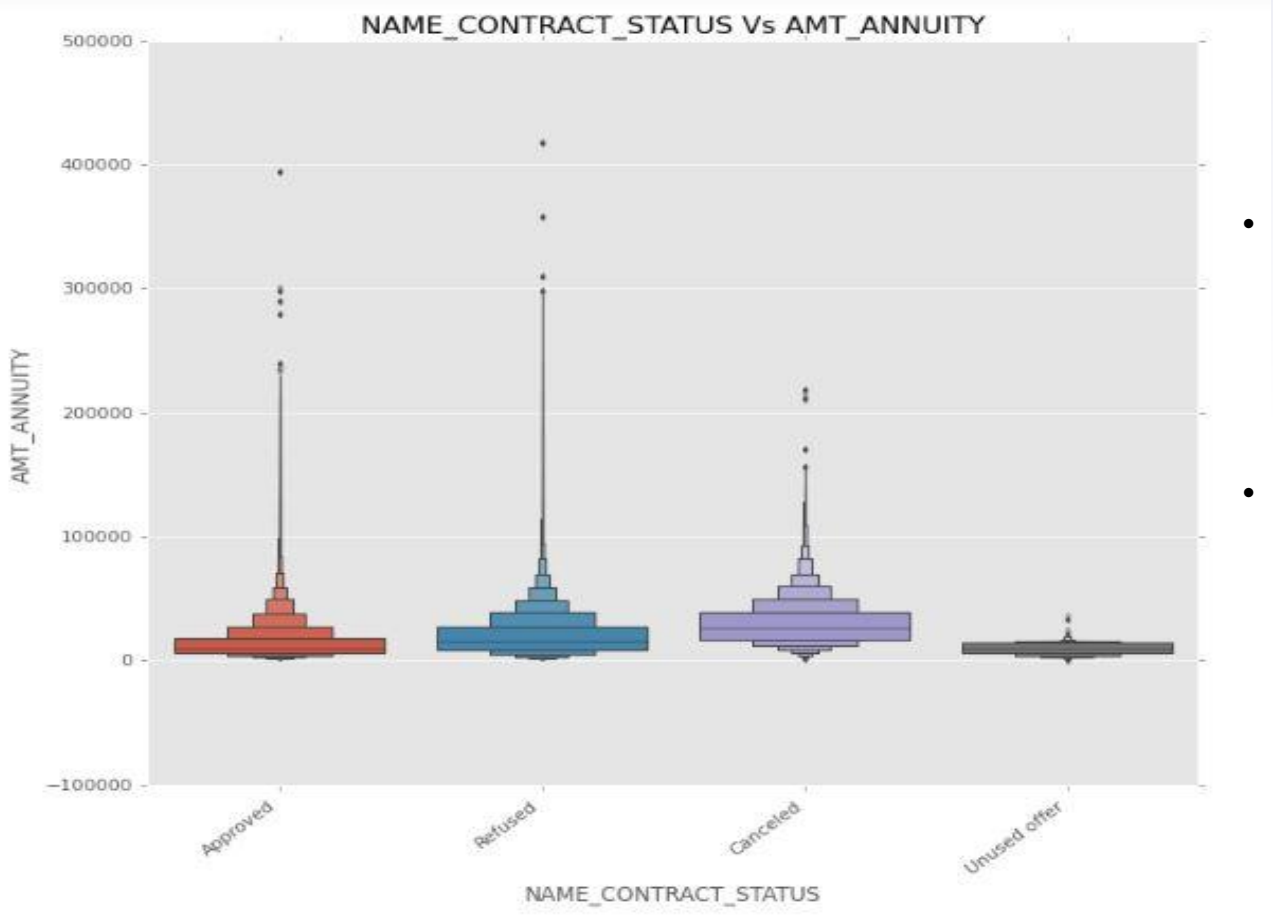
(2)Final credit amount on the previous application that was approved by the bank.

(3)Goods price of good that client asked for on the previous application.

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.

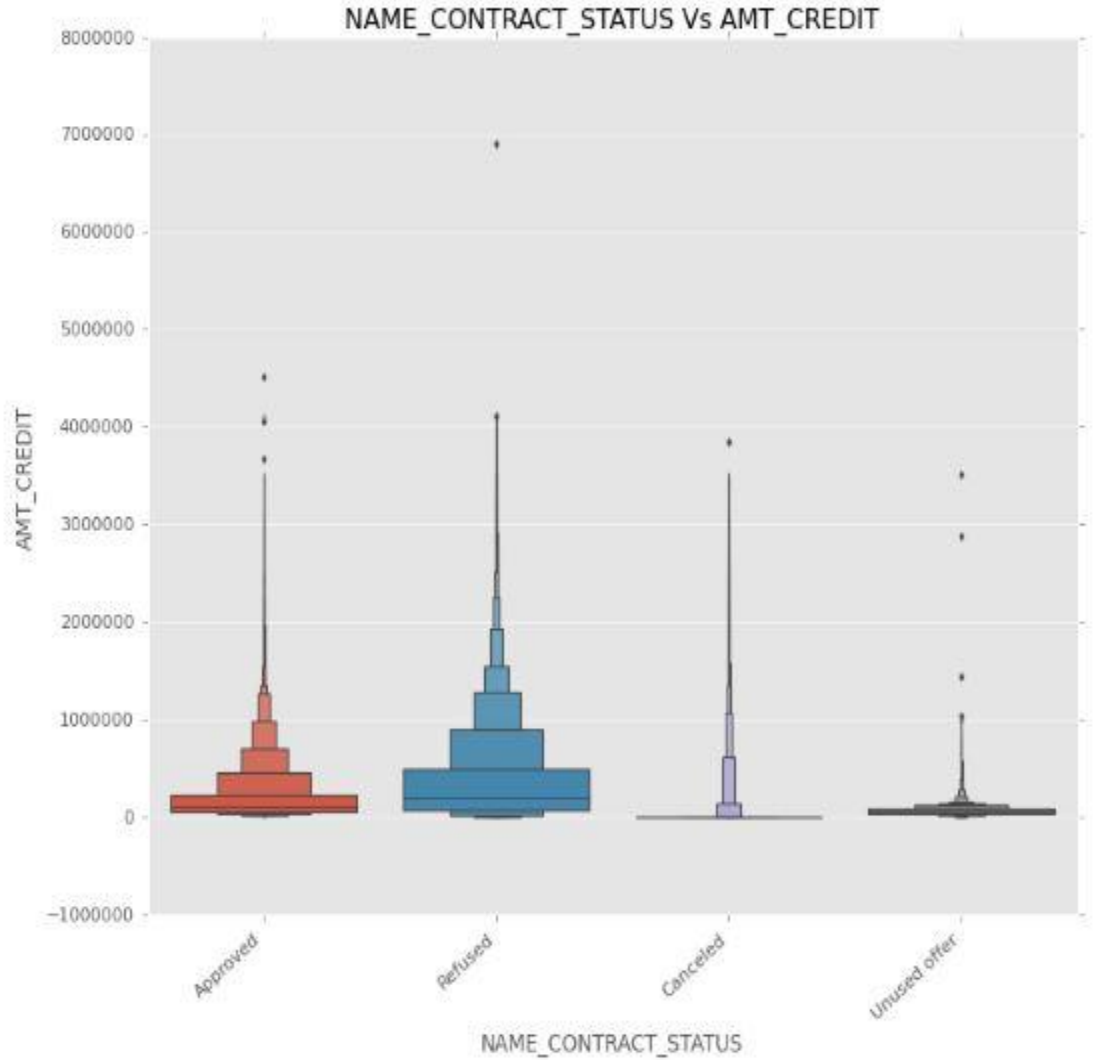
3. Final credit amount disbursed to the customer previously, after approval is highly influence by the application amount and also the goods price of good that client asked for on the previous application.

Using box plot to do some more bivariate analysis on categorical vs numeric columns



OBSERVATION

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



OBSERVATION

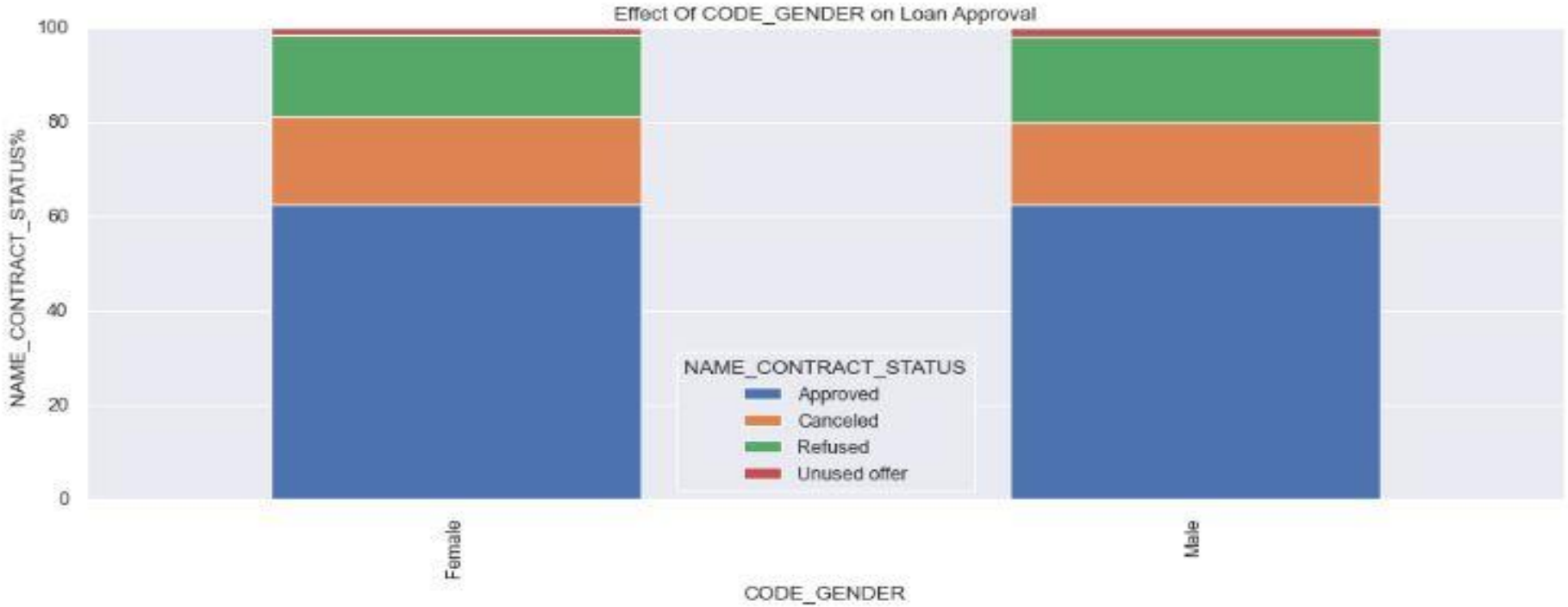
- We can infer that when the AMT_CREDIT is too low, its get's cancelled/unused most of the time.

Effect of FLAG_OWN_CAR on loan Approval

OBSERVATION

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

Effect of CODE_GENDER on loan Approval



OBSERVATION

- 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.

Effect of TARGET on loan Approval

Target variable (0 - Non Defaulter 1 - Defaulter)

OBSERVATION

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

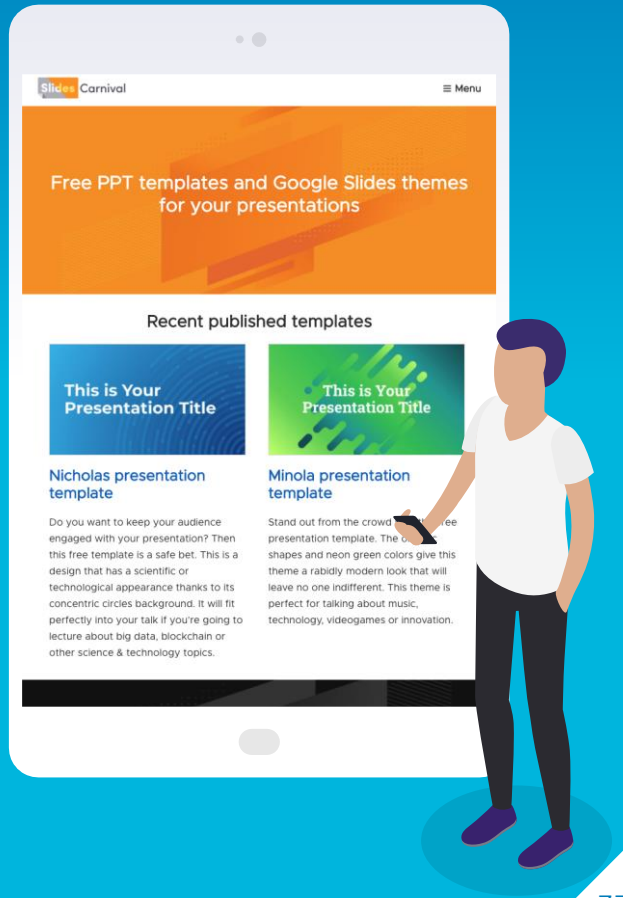
CONCLUSION

Banks should focus more on contract type ‘Student’ ,’pensioner’ and ‘Businessman’ with housing ‘type other than ‘Co-op apartment’ for successful payments.

Banks should focus less on income type ‘Working’ as they are having most number of unsuccessful payments.

Also, with loan purpose ‘Repair’ is having higher number of unsuccessful payments on time.

Get as much as clients from housing type ‘With parents’ as they are having least number of unsuccessful payments.



THANK YOU!