Credit EDA Assignment Presentation of Case Study

Divya Shah.
Advanced Certification Course in Data Science, IIITB.
Upgrad Batch – C44

Problem Statement

This case study aims 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. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Overall Approach

- Data Cleaning for Application Dataset.
  1. Identifying and Binning variables for better analysis.
  2. Checking Target Imbalance.
  3. Segmenting data frames based on target
  4. Segmented Univariate Analysis
  5. Segmented Bivariate Analysis

- Data Cleaning for Previous Application Dataset
  1. Univariate Analysis
  2. Top 10 correlations with important variables
  3. Bivariate Analysis

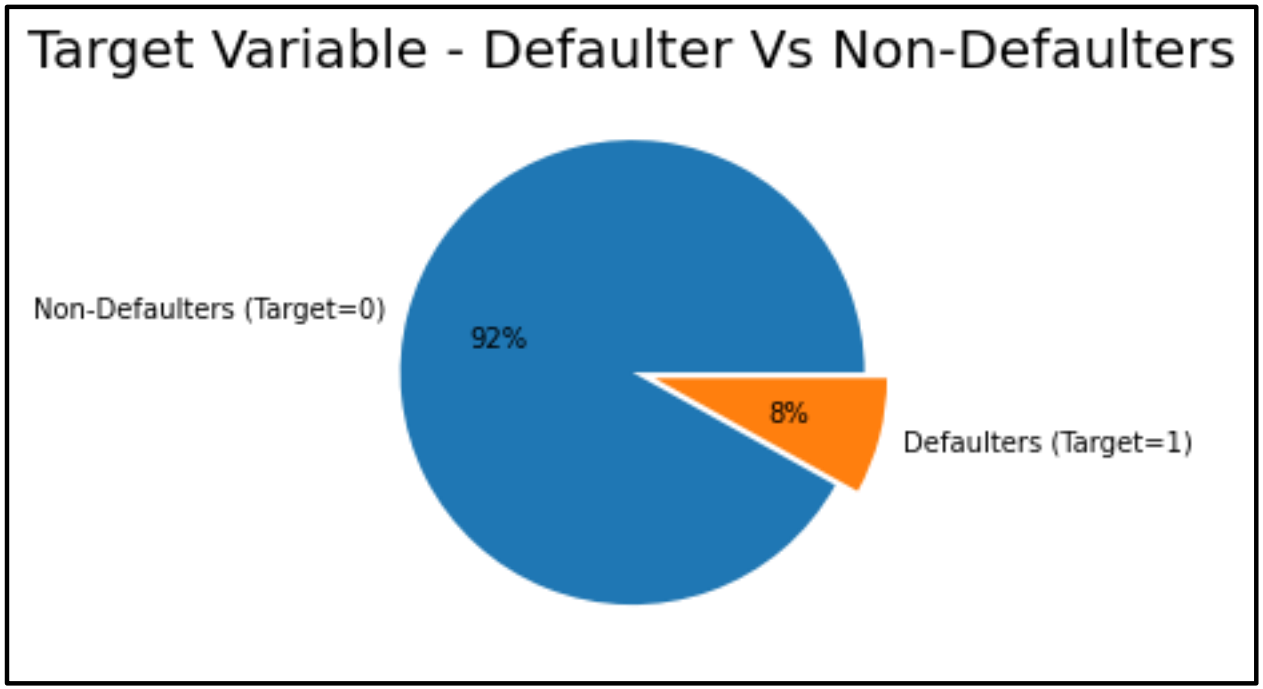
- Merging Application and Previous Application data sets.
  1. Finding relation of potential driver variables
  2. Performing Univariate and Bivariate analysis.

Assumptions

- Multiple choices of annuity are given to clients.
- Loans taken by unemployed clients are for starting new business to take career paths.
- Clients of age group more than 25 tend to less default because they get employed or start earning.

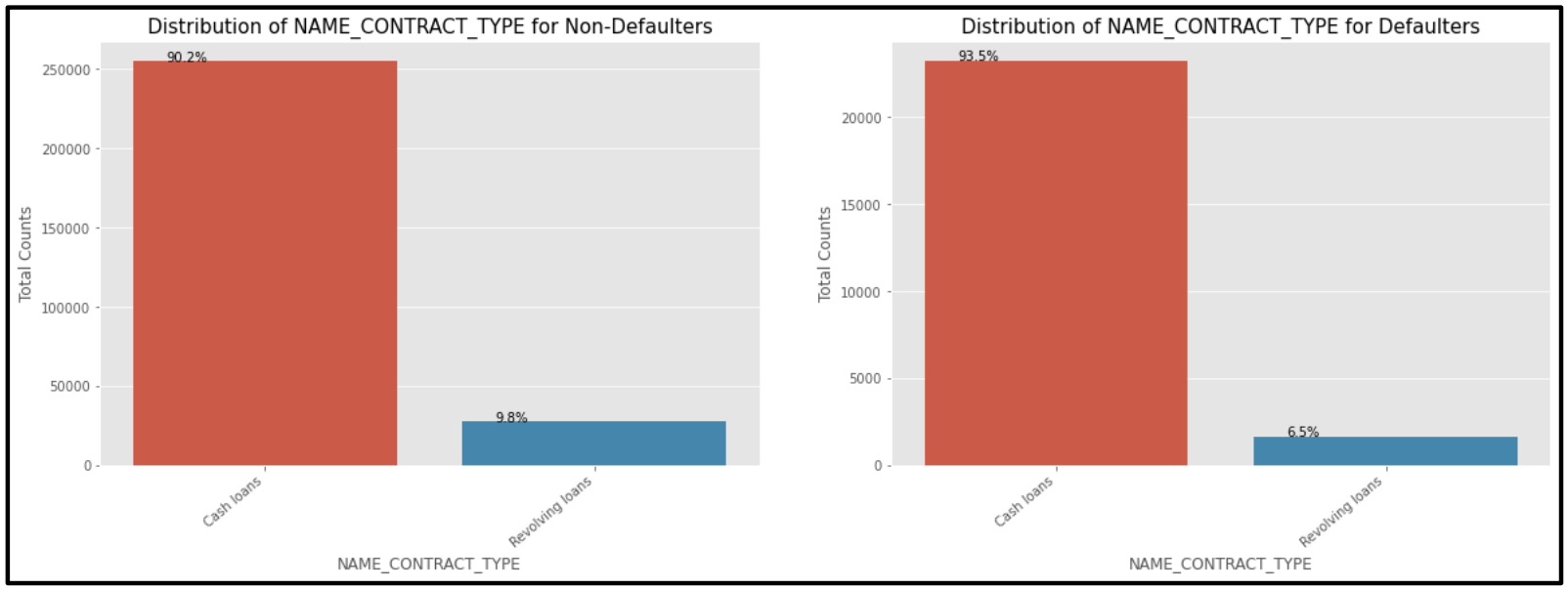
Target Imbalance through Pie Plot

There is an imbalance in Target variable; 92% of the clients are not defaulters but 8% are defaulter; which is quiet significant in order to analyze the risk associated with defaulters.

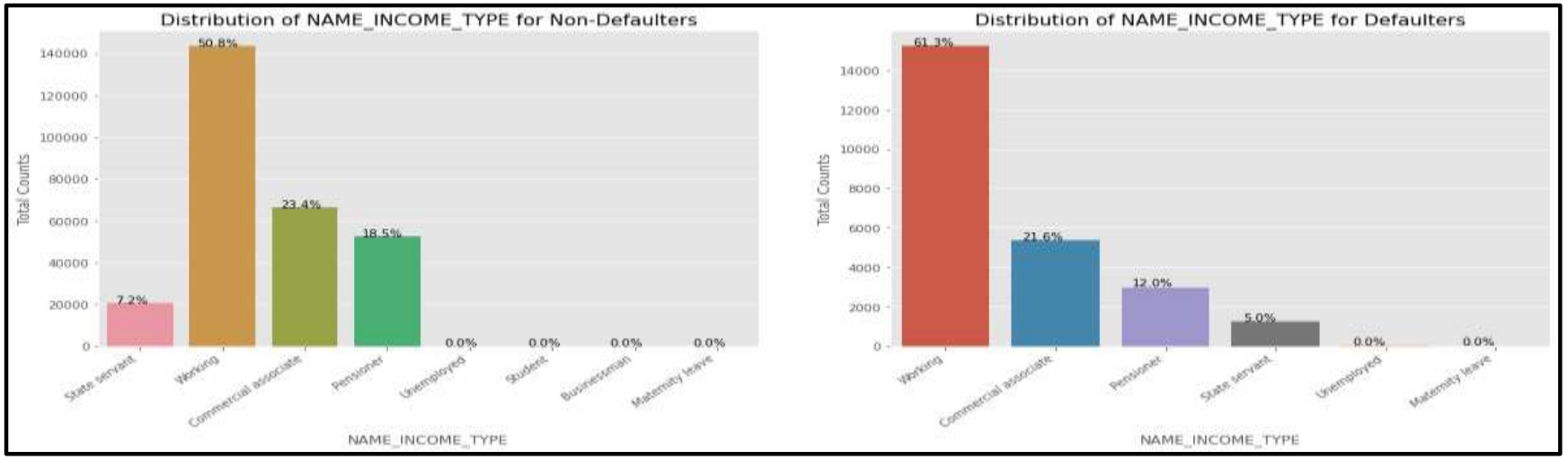


Univariate Ordered Categorical Analysis

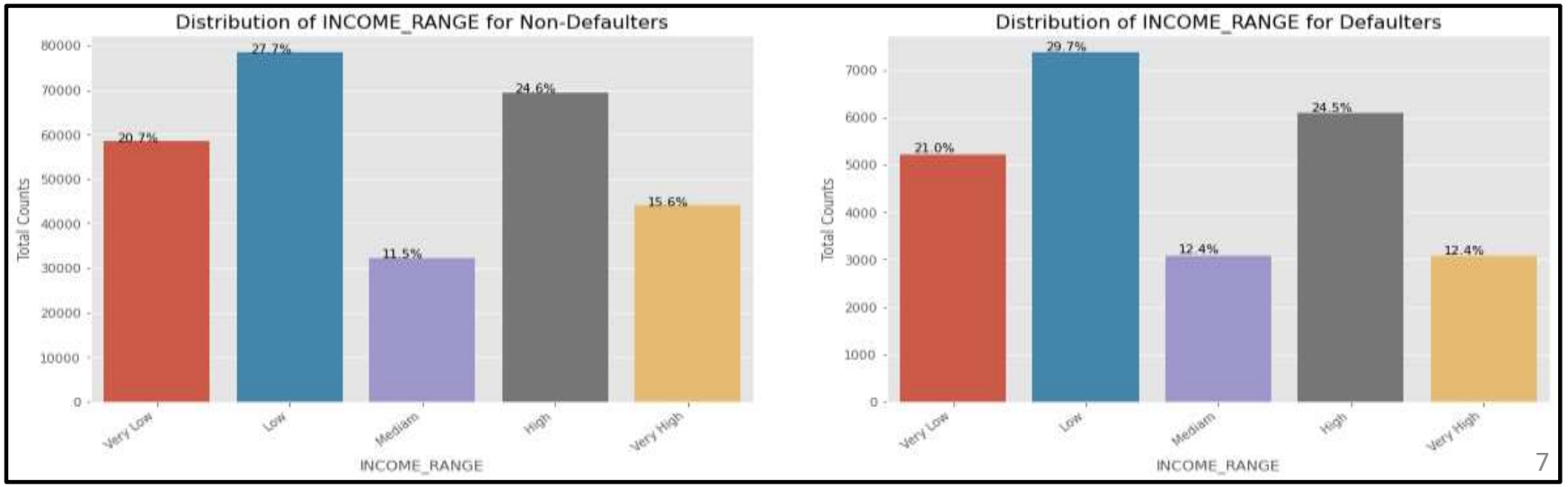
From below chart, we can see that in both cases, applications for Cash loans are higher than revolving loans. **93.5%** of defaulters had applied for cash loans and 6.5% for revolving loans. Cash loans is a very popular loan type for this company yet quiet risky as well.



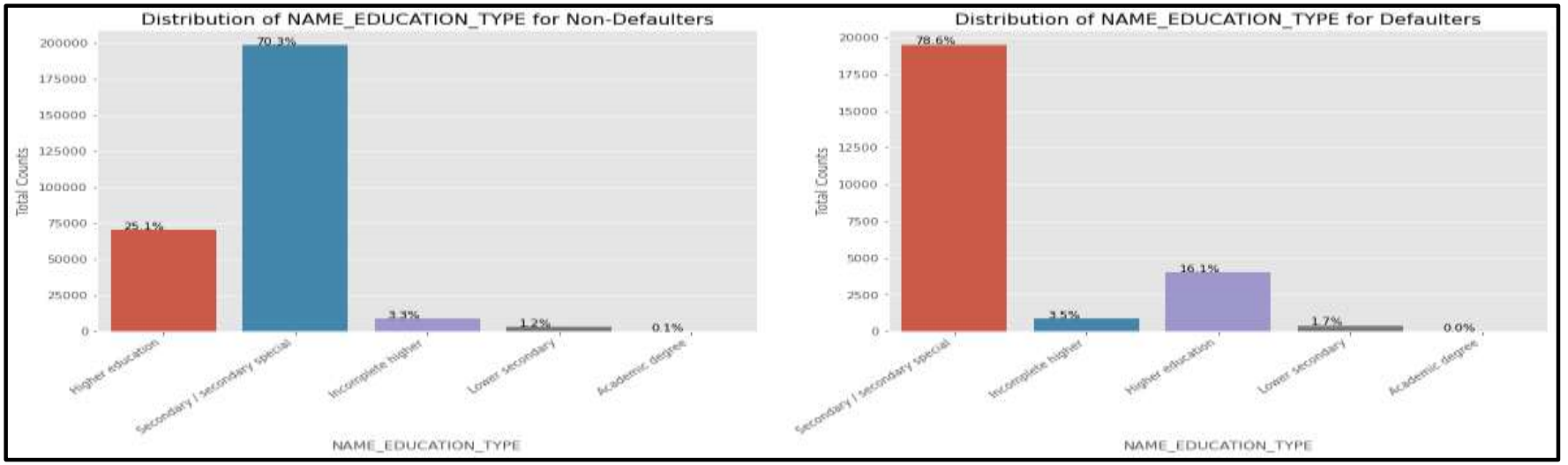
In both the cases, most of loans are given to working class clients and they are the highest contributors. 61% of the working class people are defaulters. Students and businessmen are less likely to default.



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.

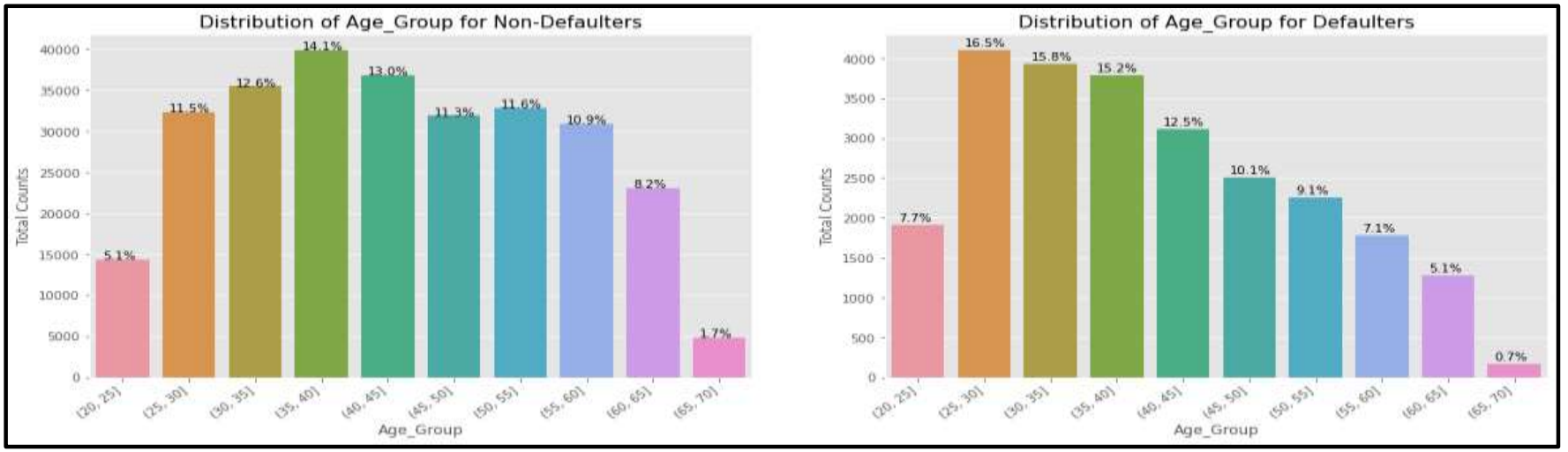


In both cases, Clients who have secondary/secondary special education are highest contributors. Which shows that, most of the loans are taken by these type of clients; 78% of the defaulters are from secondary education background.

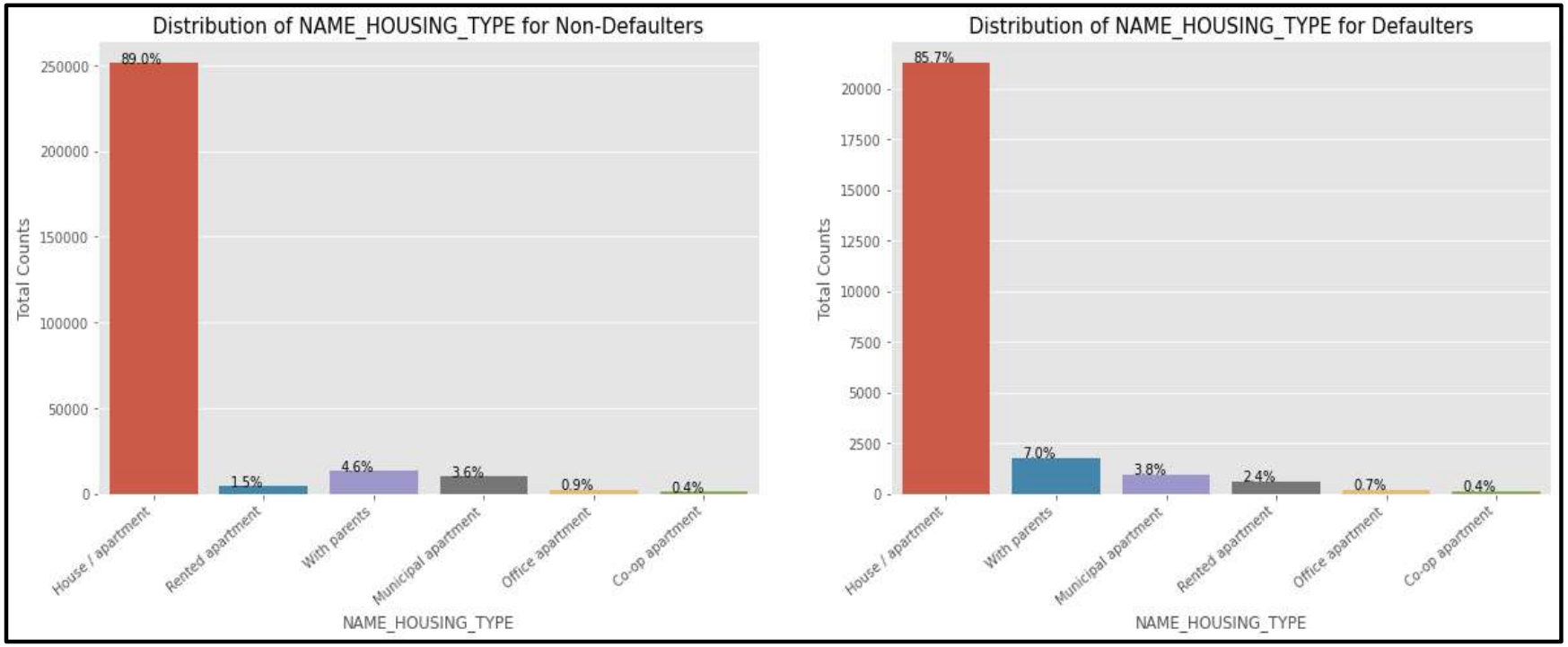


We see that (25,30] age group tend to default more often. With increasing age group, people tend to default less starting from the age 25.

One of the reasons could be they get employed around that age and with increasing age, their salary also increases.



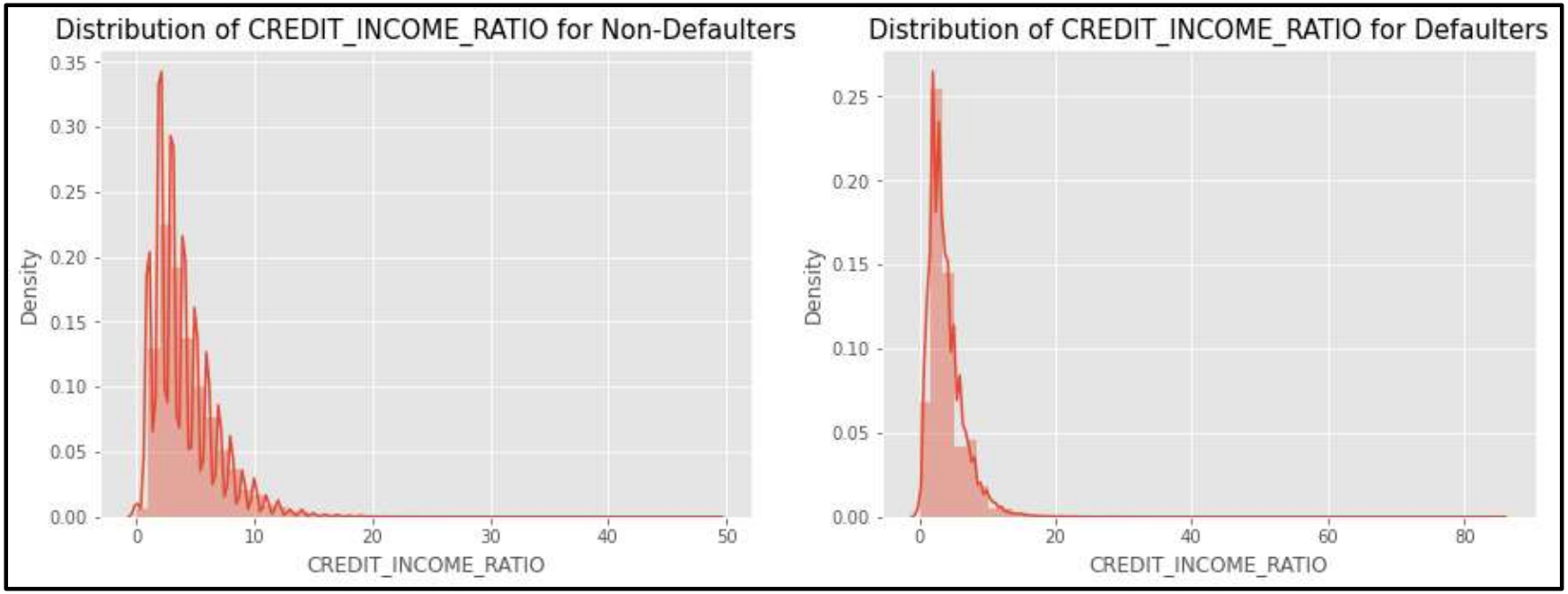
From below chart it is clear that clients staying in or owning house/apartment tend to apply for loans and also 85.7% of the clients staying in house/apartment are defaulters.



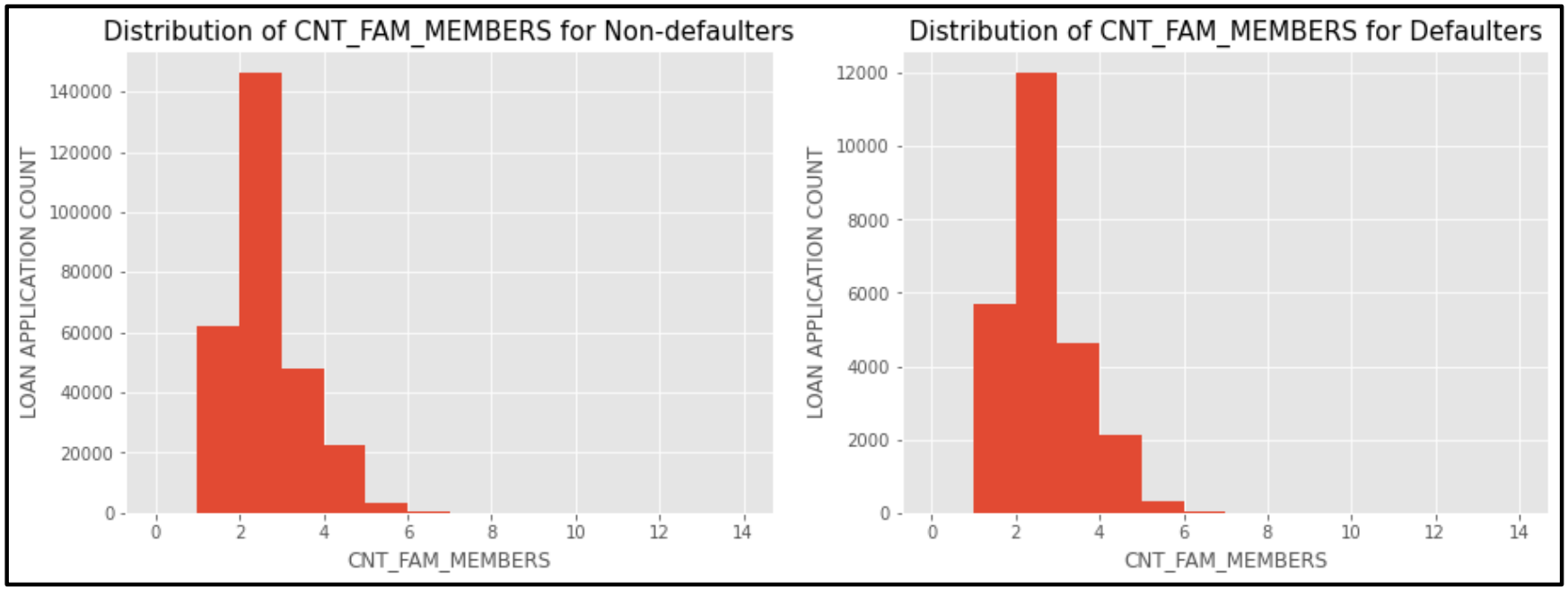
Univariate Continuous Variable Analysis

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 tend to default.

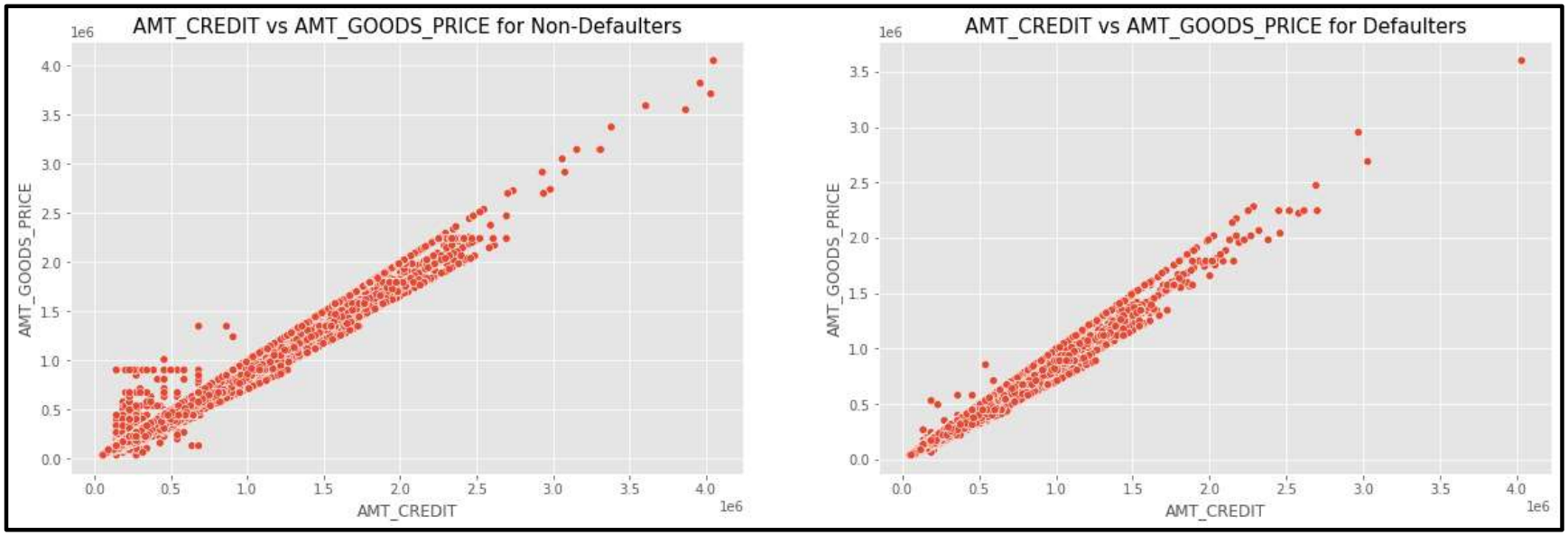


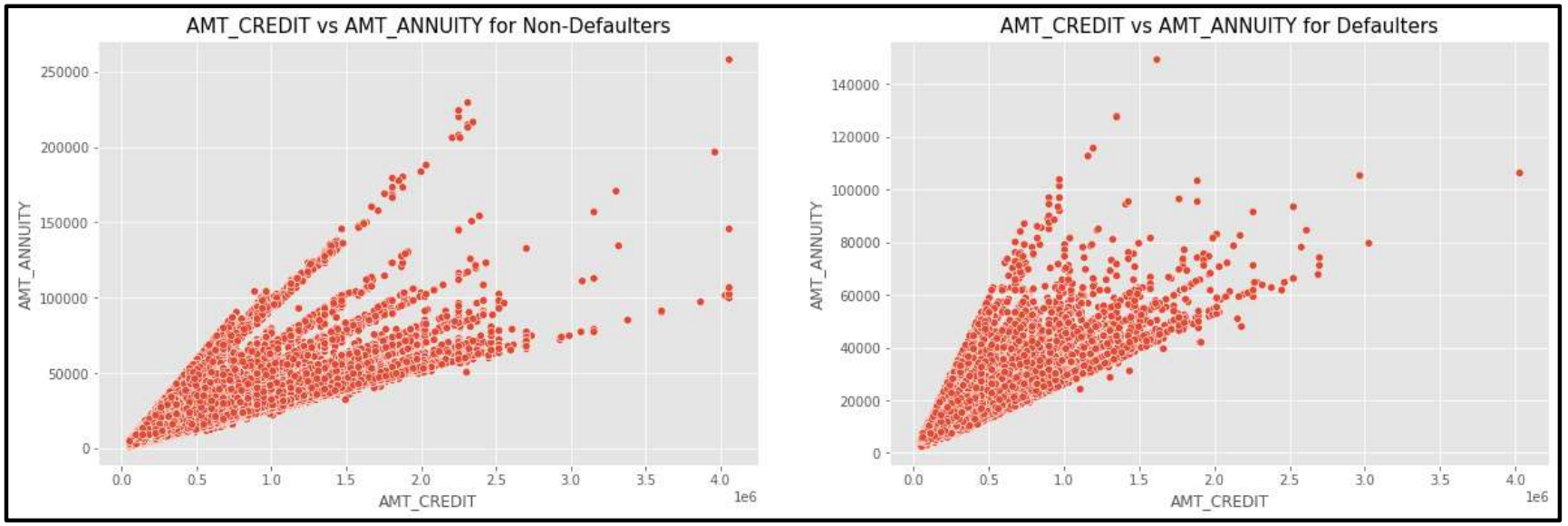
We can see that a family of 3 applies loan more often than the other families.



Bivariate Analysis on Numerical Variables

The plot below doesn't show much difference; amount of credit is directly proportional to amount of goods price.





There is no stark difference between two of the above plotted graphs.

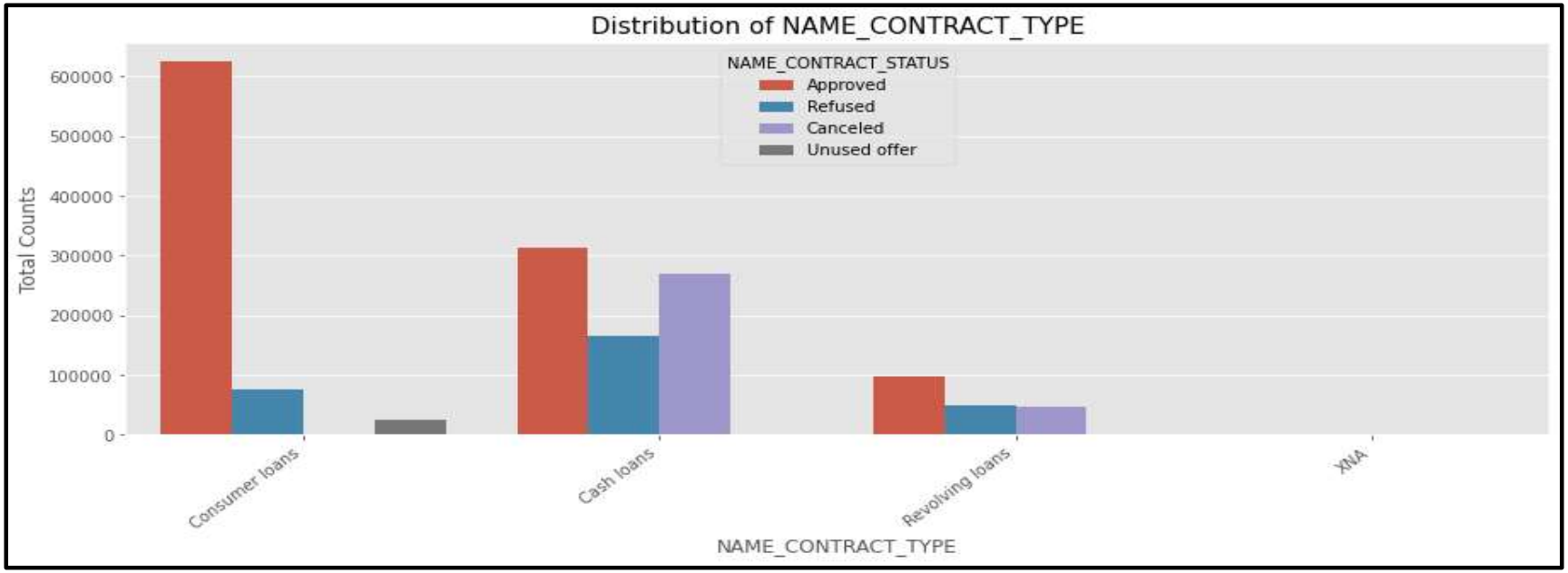
For same amount of credit the annuity varies. This must be because there can be multiple choices to select annuity based on monthly income and other related factors.

The defaulters tend to choose comparatively less amount for annuity than non-defaulters.

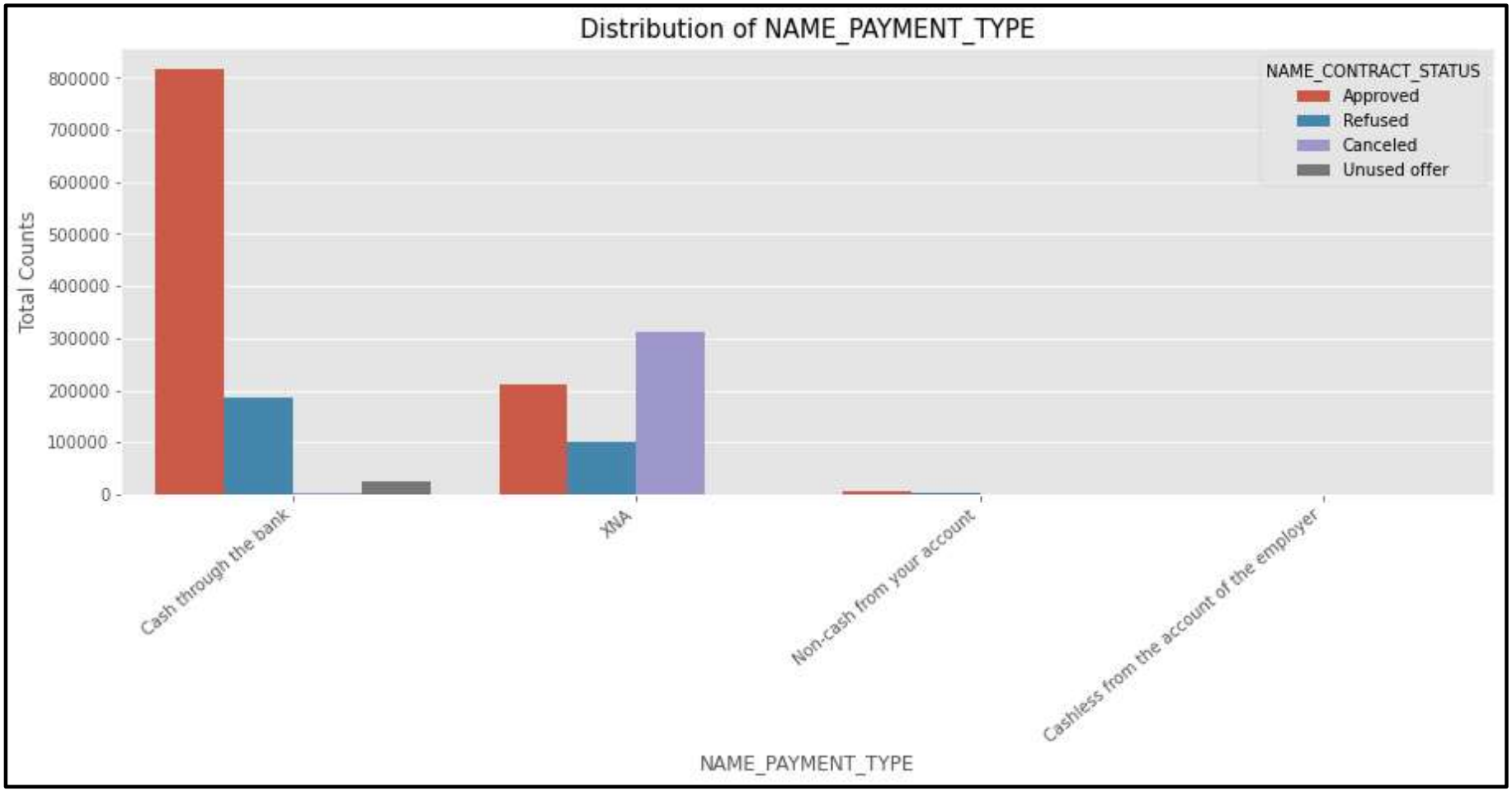
Data Analysis on Previous Application DataSet

From below chart we can infer that, majority of the applications are for Consumer Loans and Cash Loans.

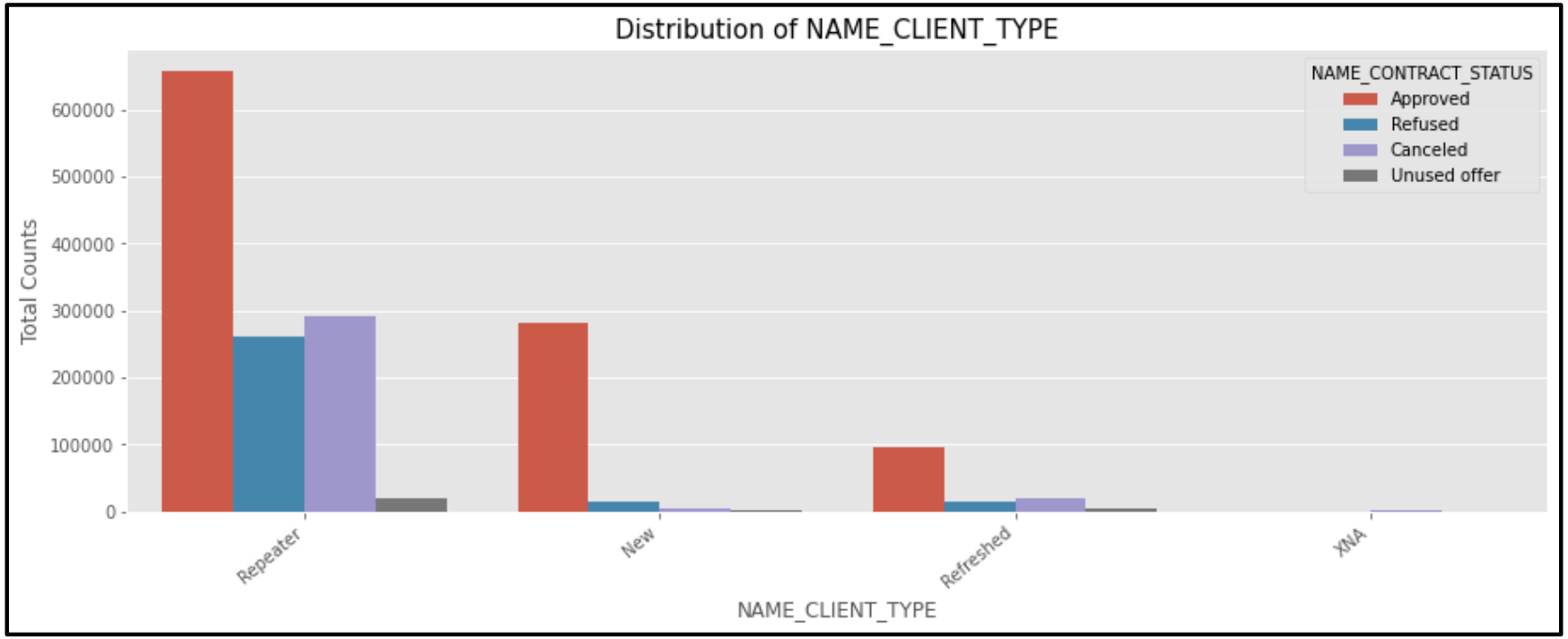
However, majority of the applications have been rejected for cash loans among others.



From below plot, we can infer that, most of the clients chose to replay via Cash through the bank. Non-cash and Cashless from account of client seems unpopular.



Most of the loan applications are from repeaters; some of the application have also been refused.



Bivariate Analysis Using Pair Plot

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

(1) The amount of credit did client asked on the previous application

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

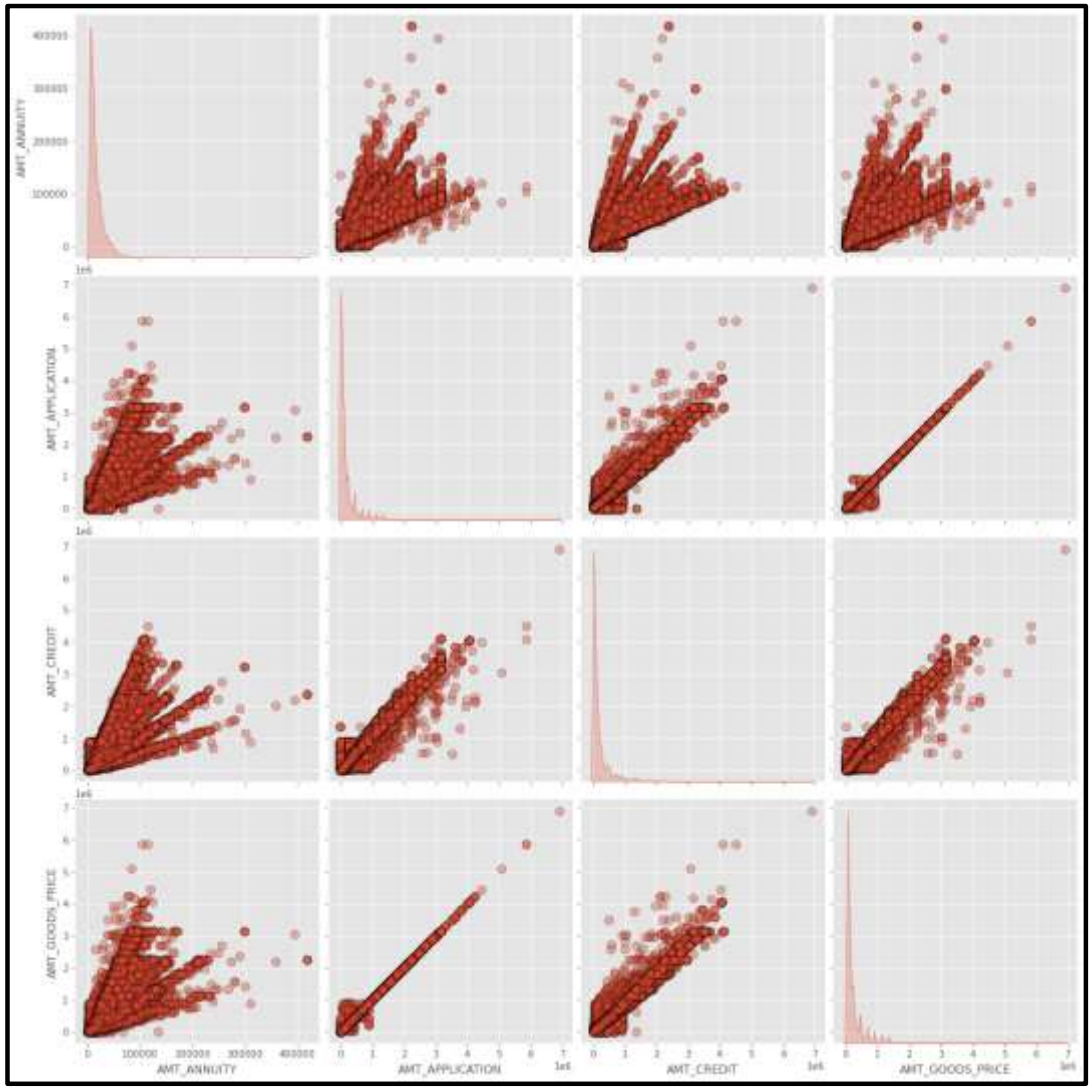
(3) Goods price of good that client asked for on the previous application.

For the amount of credit that the client asked on the previous application is highly influenced by the amount of goods price.

Final credit amount disbursed to the customer after approval is same as amount of application and amount of goods price; from this we can infer that when loan was approved the full price of good has been offered by the company.

Amount of Goods price shows similar relation with other three columns, it is almost equal to amount of application and amount that has been credited while taking loan in the client's account.

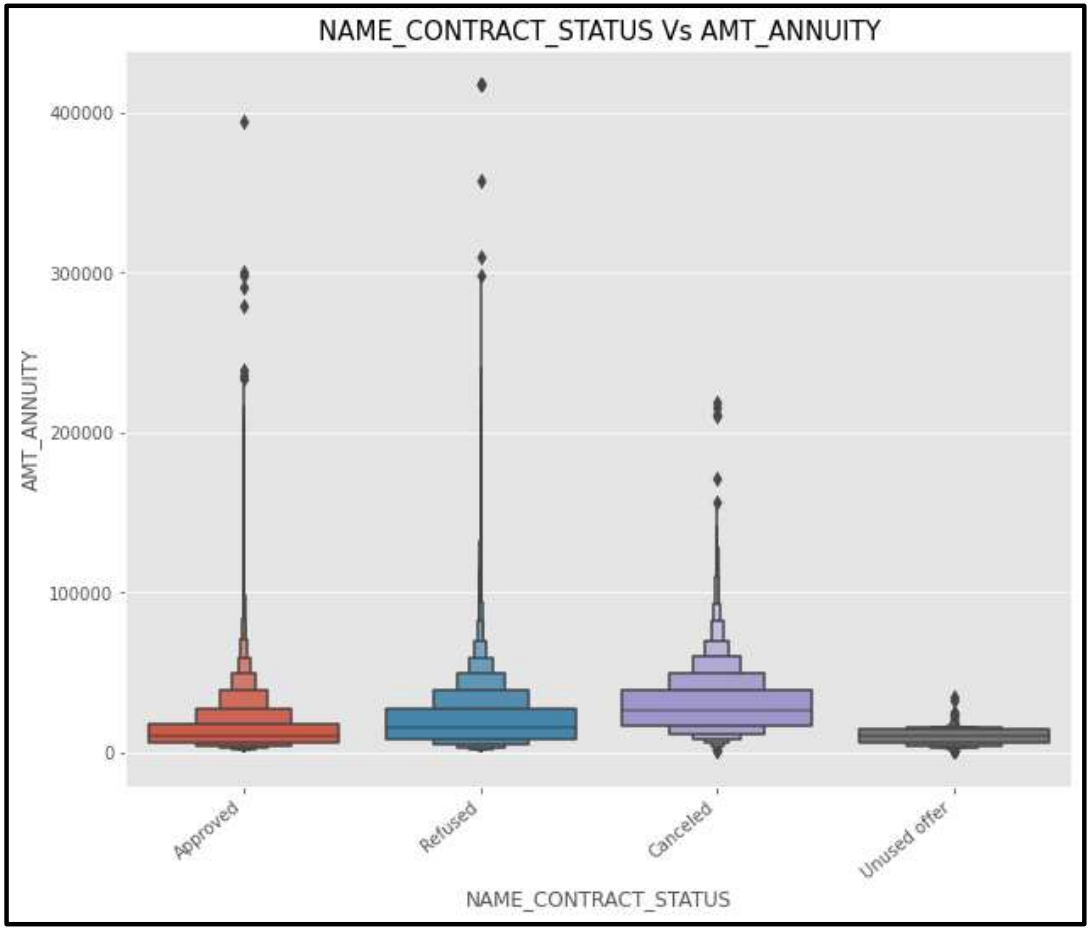
Amount of goods price vs amount of annuity is somewhat directly proportional; client might have choice among certain options to decide the annuity according to financial capacity.



Bivariate Analysis using boxplots

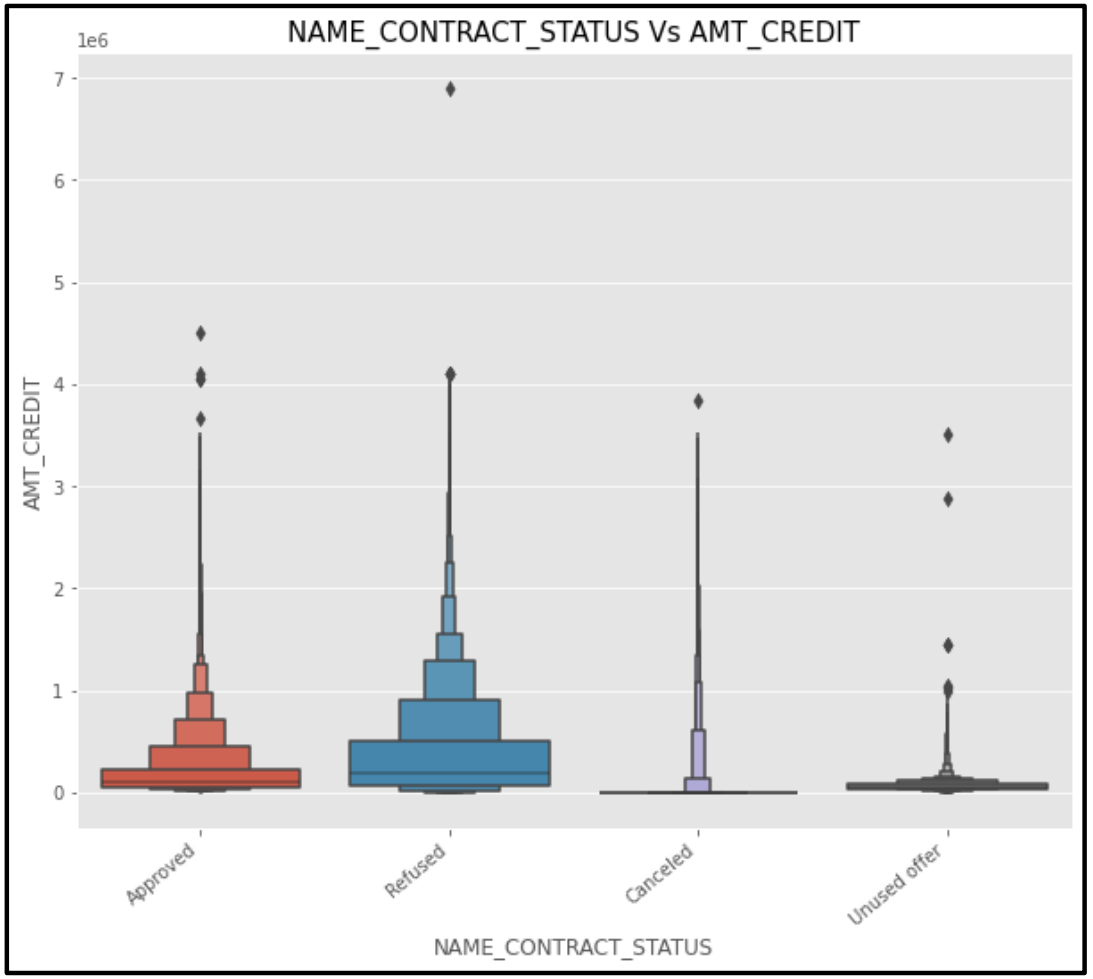
• From the plot we can infer that application with lower annuity fall in category of cancelled or unused offer. On the contrary applications with higher amount of annuity got refused.

• Approved applications have lesser annuity compared to refused applications.



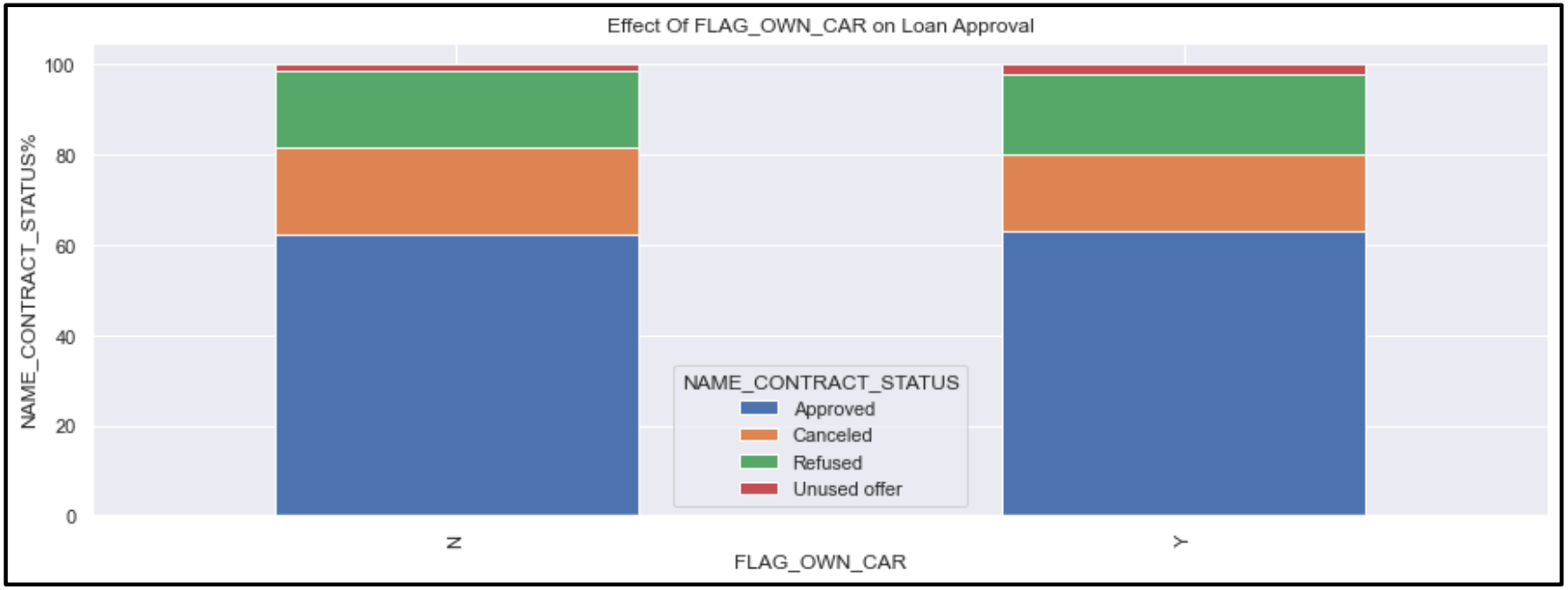
• We can infer that applications with very low credit amount either gets canceled or remain unused by clients.

• On the other hand applications with very high credit amount gets refused by company.

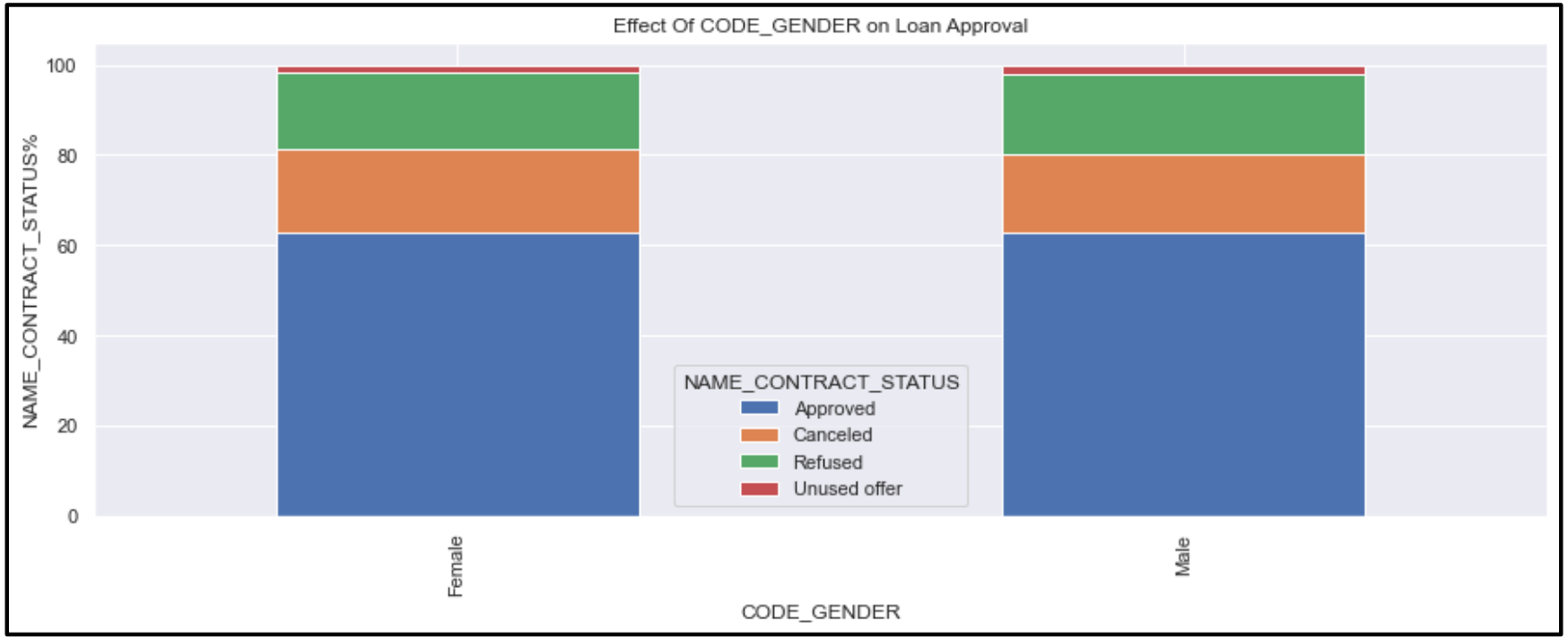


Data Analysis of Merged Dataset

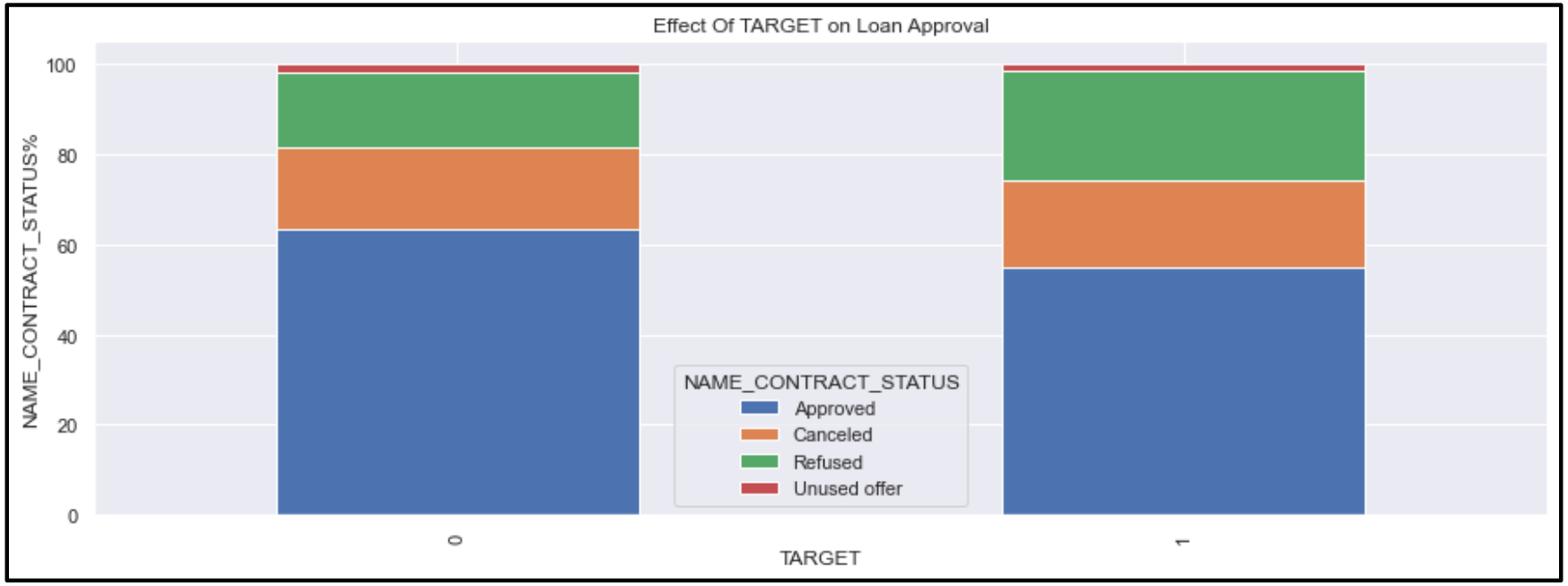
We see that car ownership doesn't have any effect on application approval or rejection.
The bank can add more weightage to car ownership while approving a loan amount.



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.

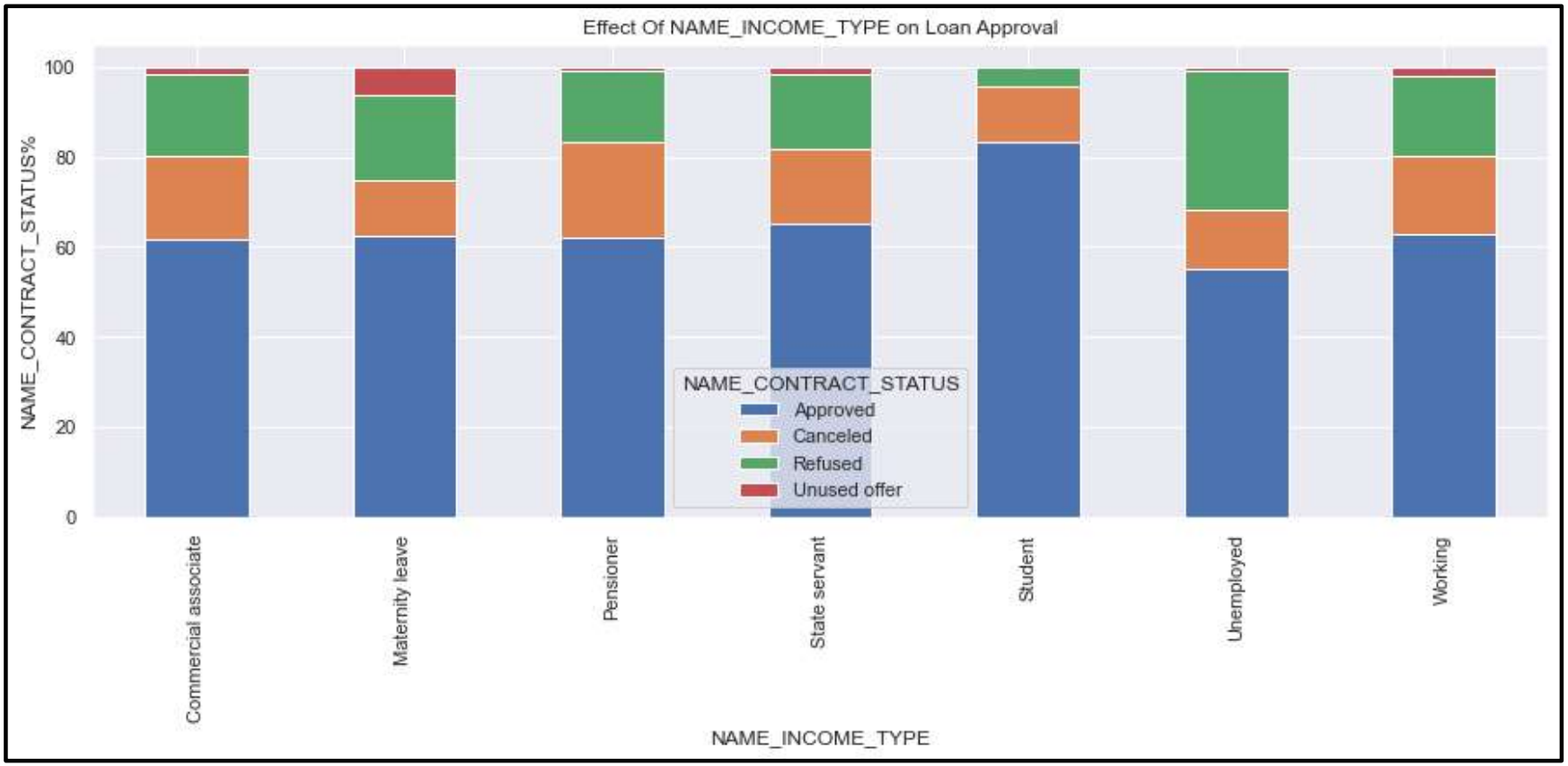


We can see that the people who were approved for a loan earlier, defaulted less often where as people who were refused a loan earlier have higher chances of defaulting.

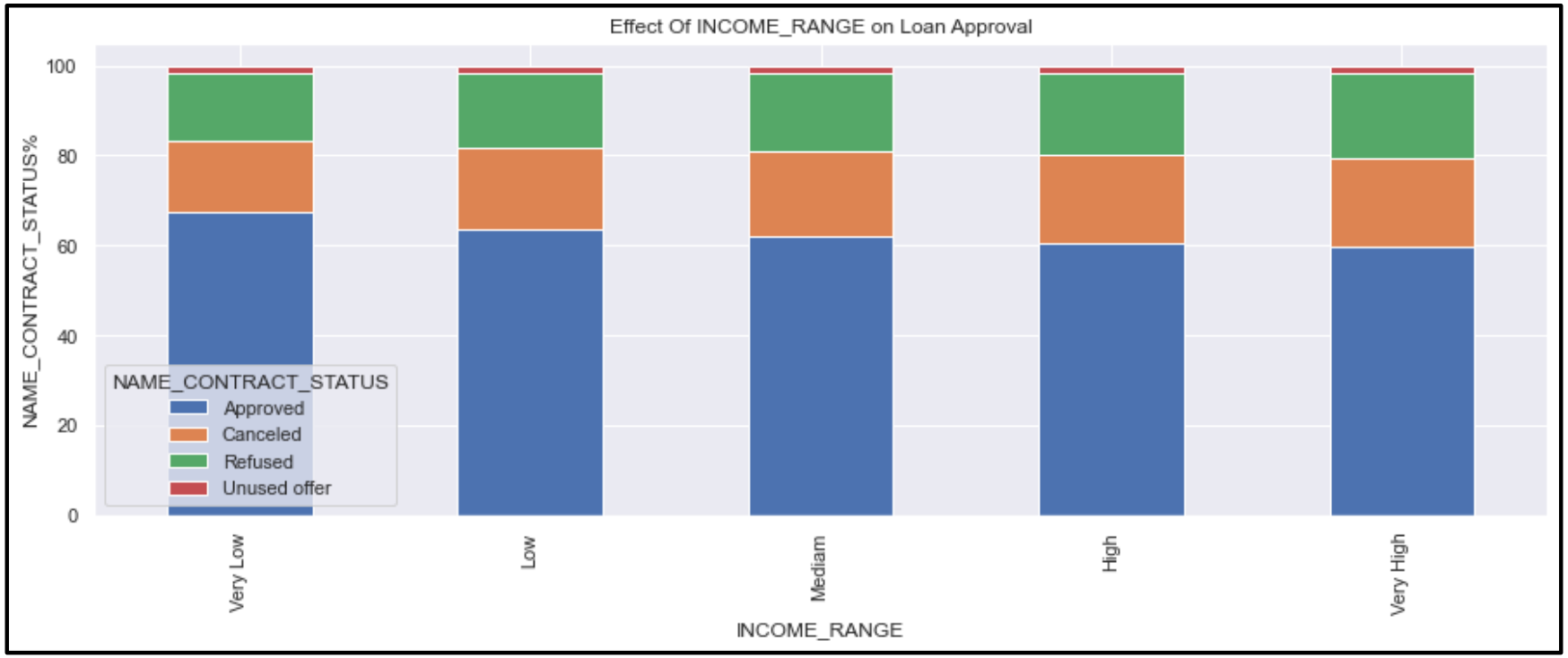


From above plot we can see that majority of the loans are approved for students. Earlier in the analysis we have also noted that students are less likely to default. Hence, bank/company can introduce more lucrative offers for students.

Majority of the loans have been refused for clients who are unemployed; however, the loan has been approved for 50% of the unemployed clients and earlier we have seen that they didn't default. This can be for the purpose of small business. For this class as well by identifying the needs; offers can be made lucrative.



This chart doesn't show any stark difference over the income range. Percentages for all contract types are almost same over the income range.



Conclusion & Recommendations

- Cash loans are popular and the risk associated with it is higher than other types.
- Students and businessmen are less likely to default.
- Even clients from unemployed category show less percentage of default.
- Lucrative loan offers can be made for students, businessmen and unemployed to lower the risk.
- Giving annuity choices in a certain range by considering the income range of customer can be a good practice to lower the risk.
- From previous observations it can be seen that majority of applicants are females and majority of non-defaulters are also females; however gender should not be the driving variable while analysing risk as loan can be taken on the name of any of the family member.