CREDIT EDA CASE STUDY

BY:

DEEPA B

K M SAI GANESH

INTRODUCTION

This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

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. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.
- 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:
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- 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.

The data given contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
- **All other cases:** All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

1. **Approved:** The Company has approved loan Application
2. **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
3. **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
4. **Unused offer:** Loan has been cancelled by the client but on different stages of the process.

Business objectives

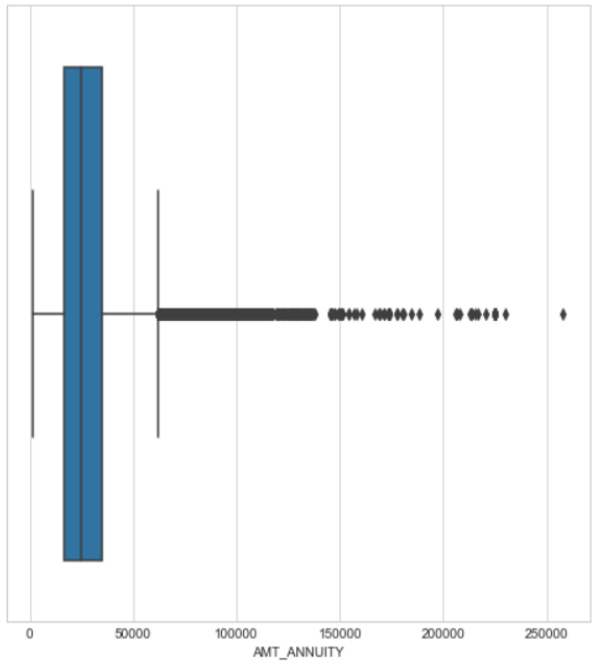
- This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers 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.
- To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

DATA UNDERSTANDING

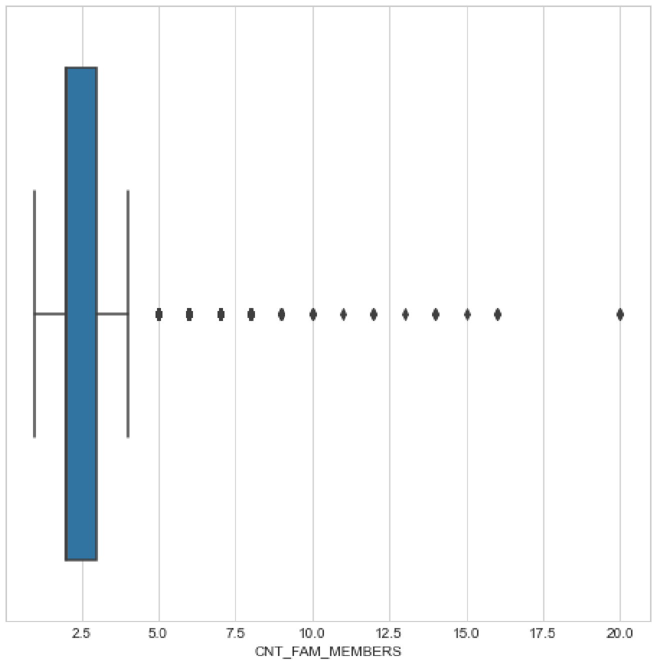
- *1. 'application_data.csv'* contains all the information of the client at the time of application. The data is about whether a **client has payment difficulties.**
- *2. 'previous_application.csv'* contains information about the client's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**
- *3. 'columns_description.csv'* is data dictionary which describes the meaning of the variables

Application Data

CHECKING THE OUTLIERS

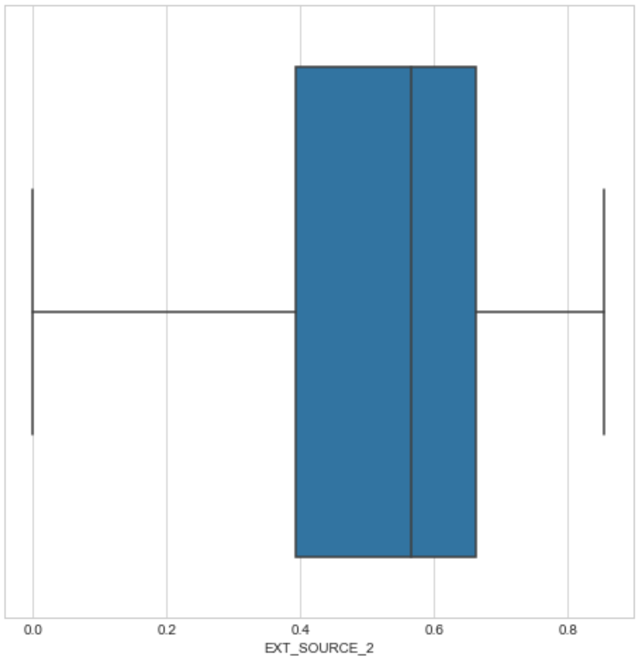


AMT_ANNUITY

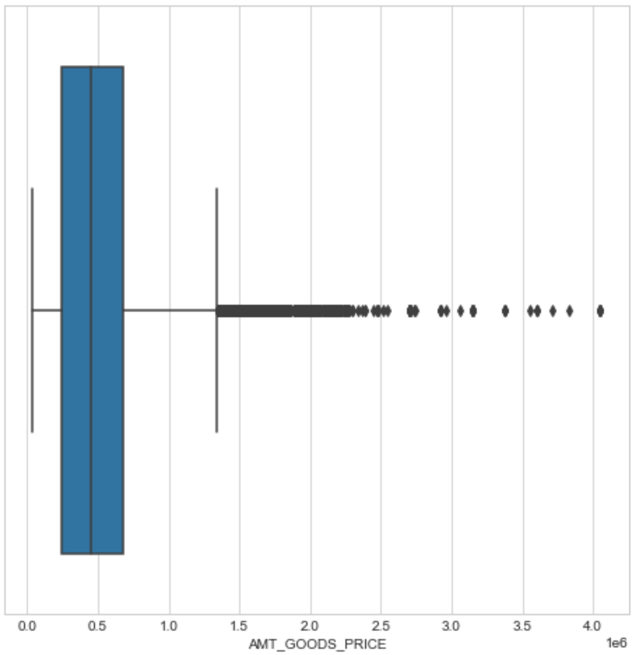


CNT_FAM_MEMBERS

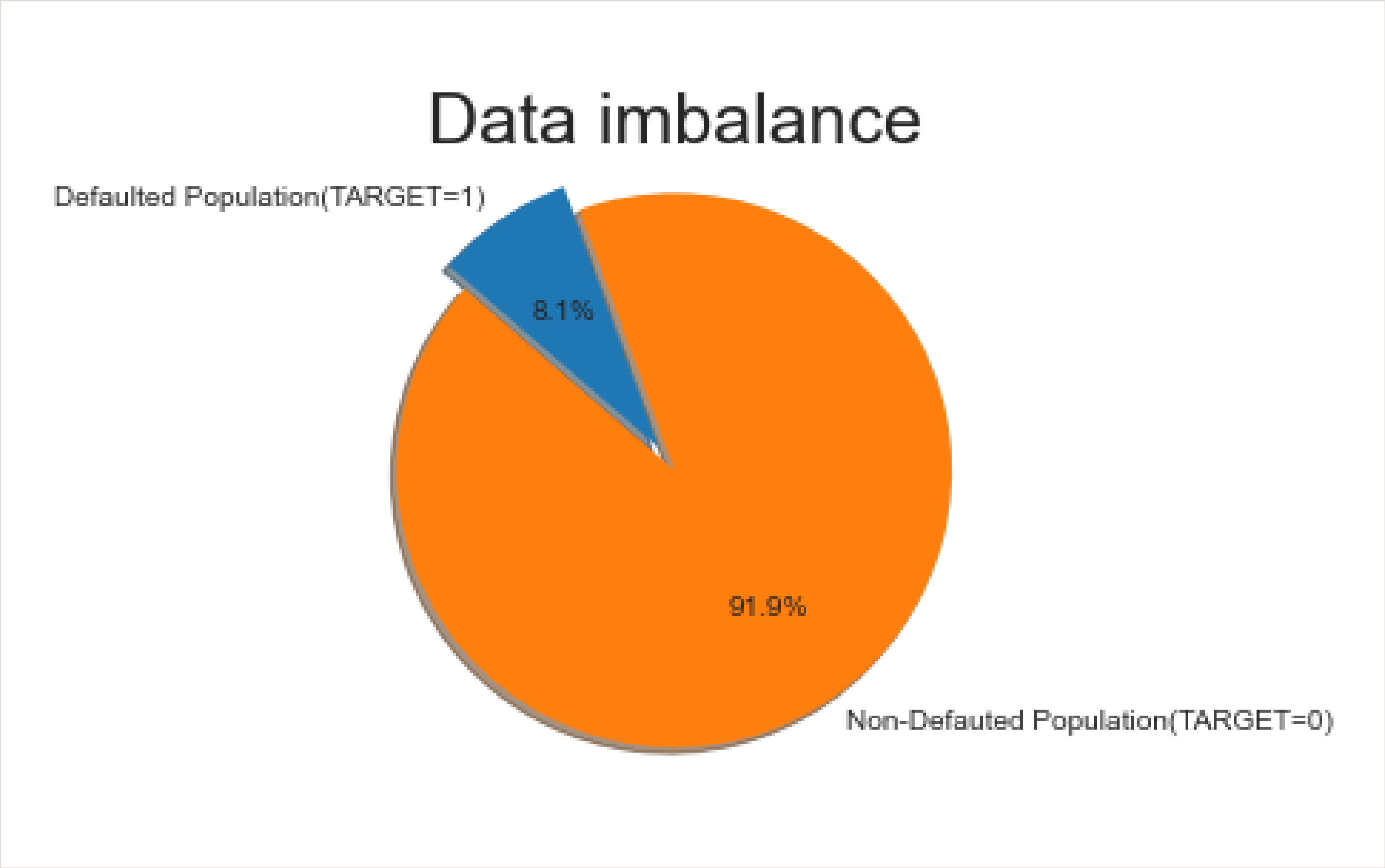
CHECKING THE OUTLIERS



EXT_SOURCE_2

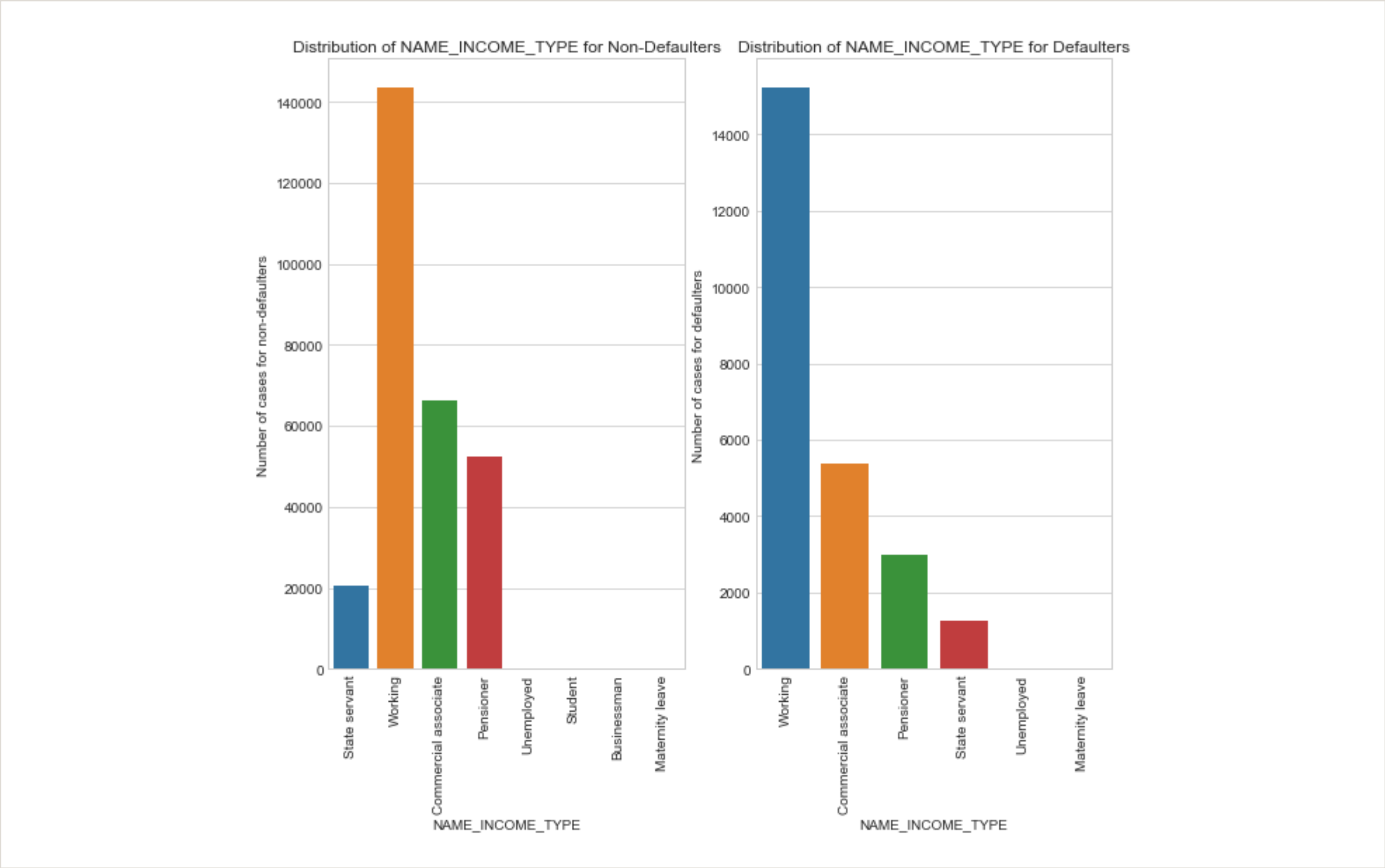


AMT_GOODS_PRICE



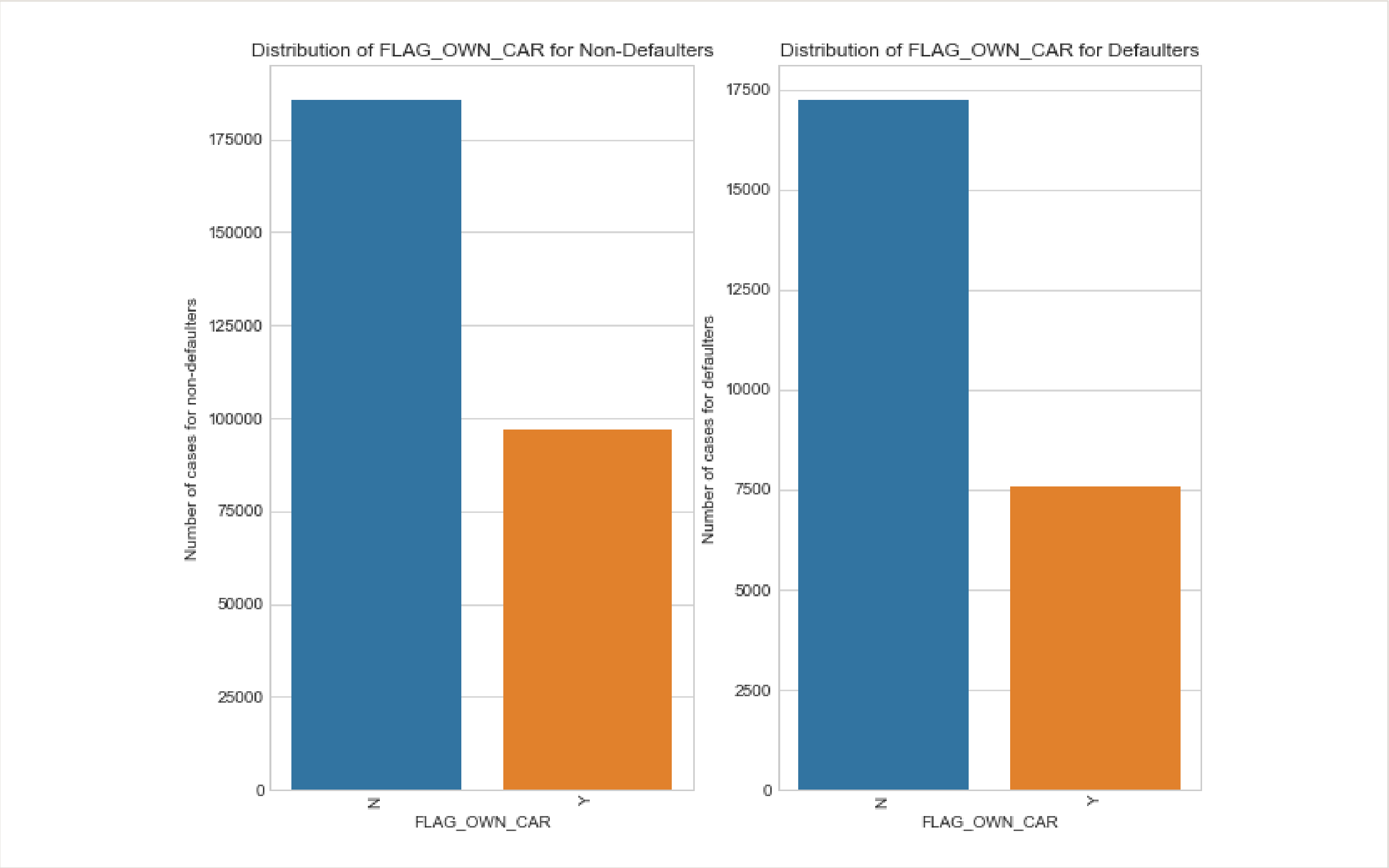
CHECKING IMBALANCE

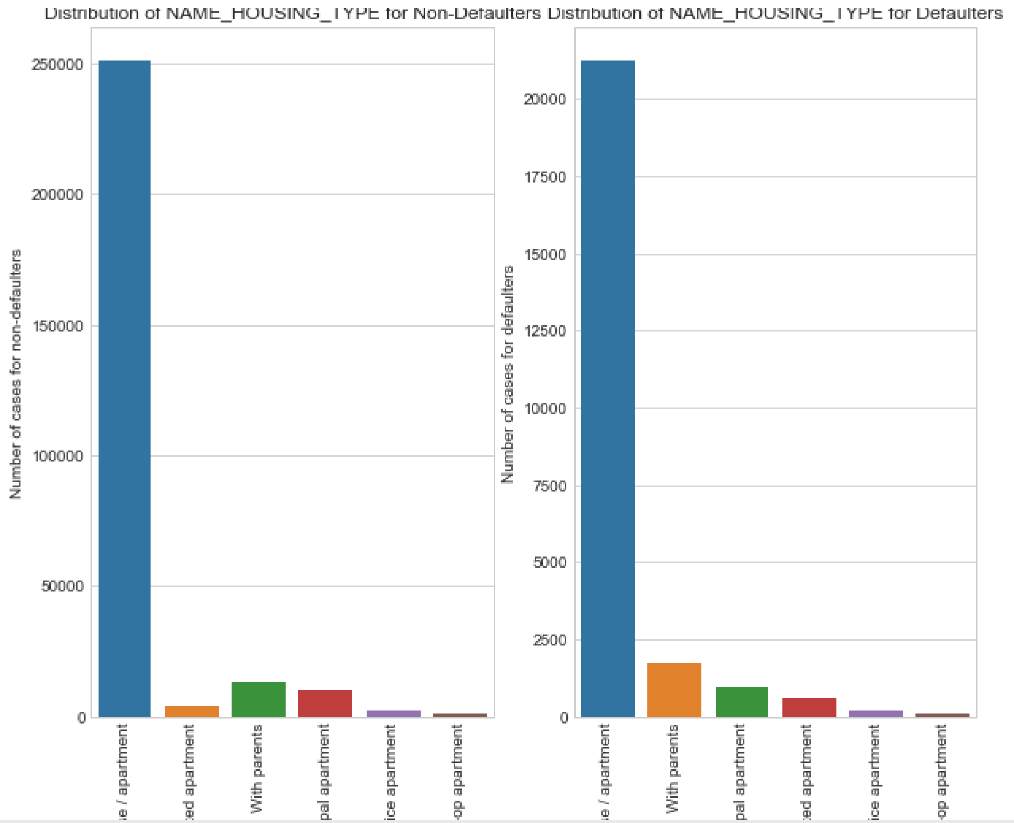
Univariate Categorical Unordered Analysis



From above graph,we observe that people who own business pays the loan amount on time. We also see that students and people who are on maternity leave are less likely to default.This is because they don't need to pay the amount for certain period of time.

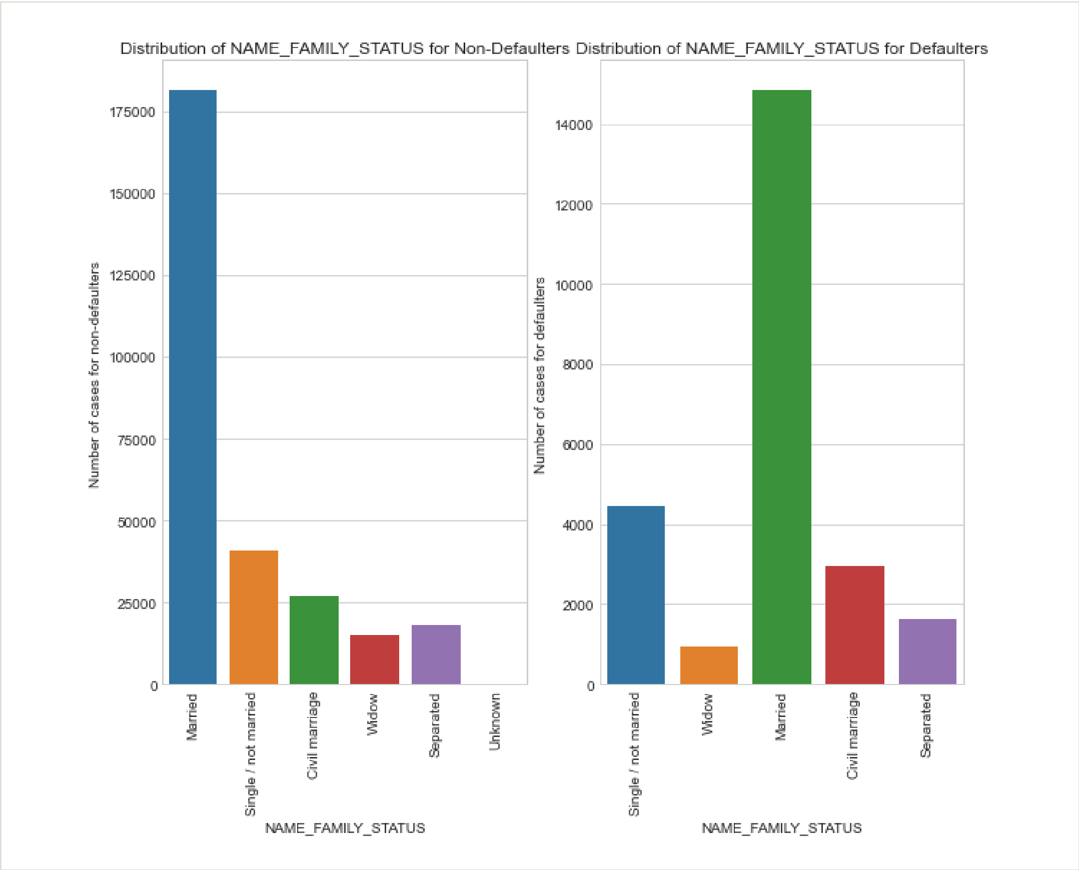
FROM THE ABOVE GRAPH, WE CAN OBSERVE THAT PEOPLE WHO DON'T OWN A CAR ARE MORE LIKELY TO DEFAULT THAN THOSE WHO DO.



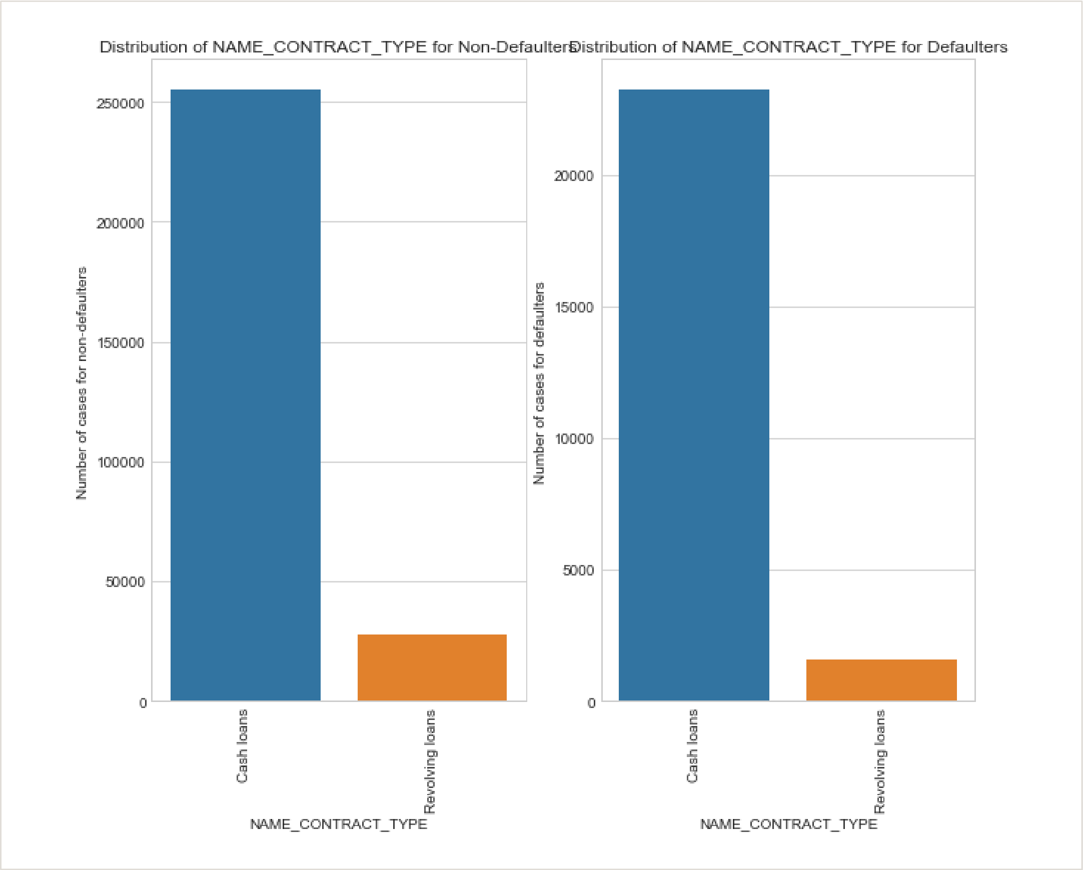


- From the above graph we observe that people who own a house are more likely to default than people who live with parents and people with rented apartment

- From the above graph we can observe that there is not much difference between married people who default and those who don’t default.But we can see that there is high risk involved in providing loans to single/married people.



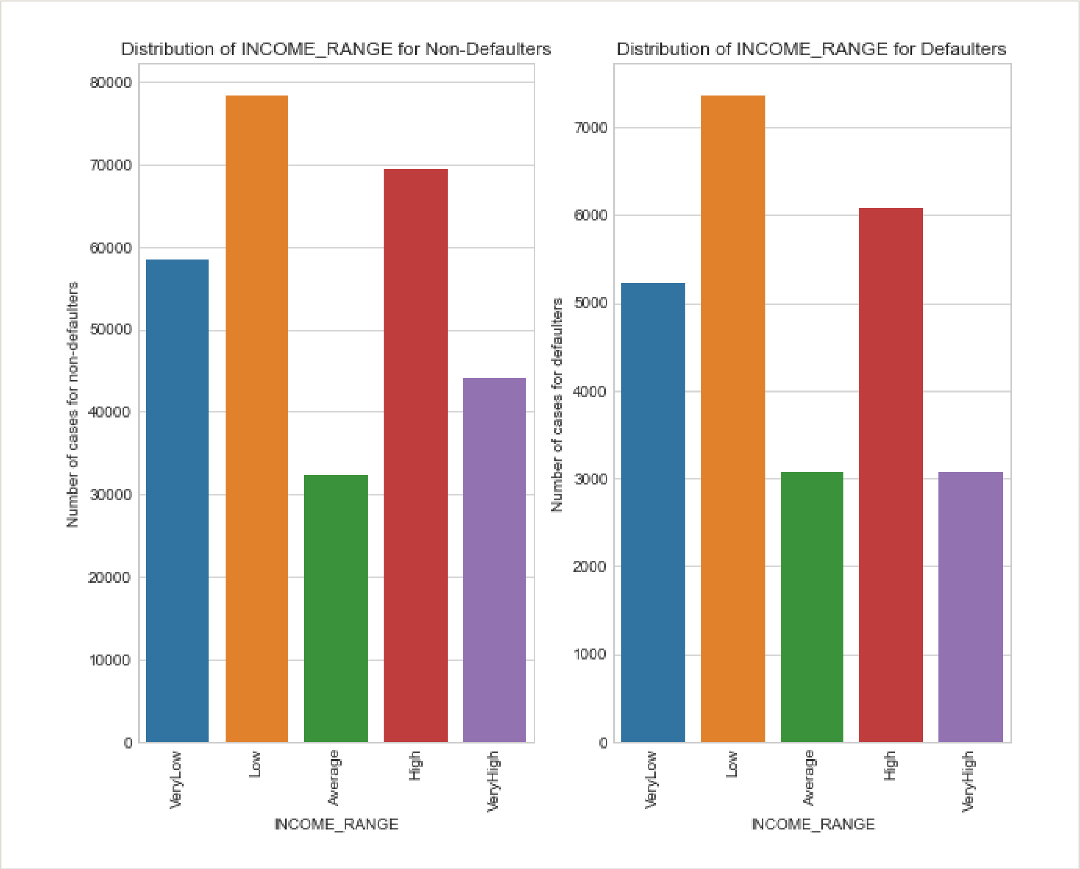
Distribution of NAME_FAMILY_STATUS for Non-Defaulters Distribution of NAME_FAMILY_STATUS for Defaulters

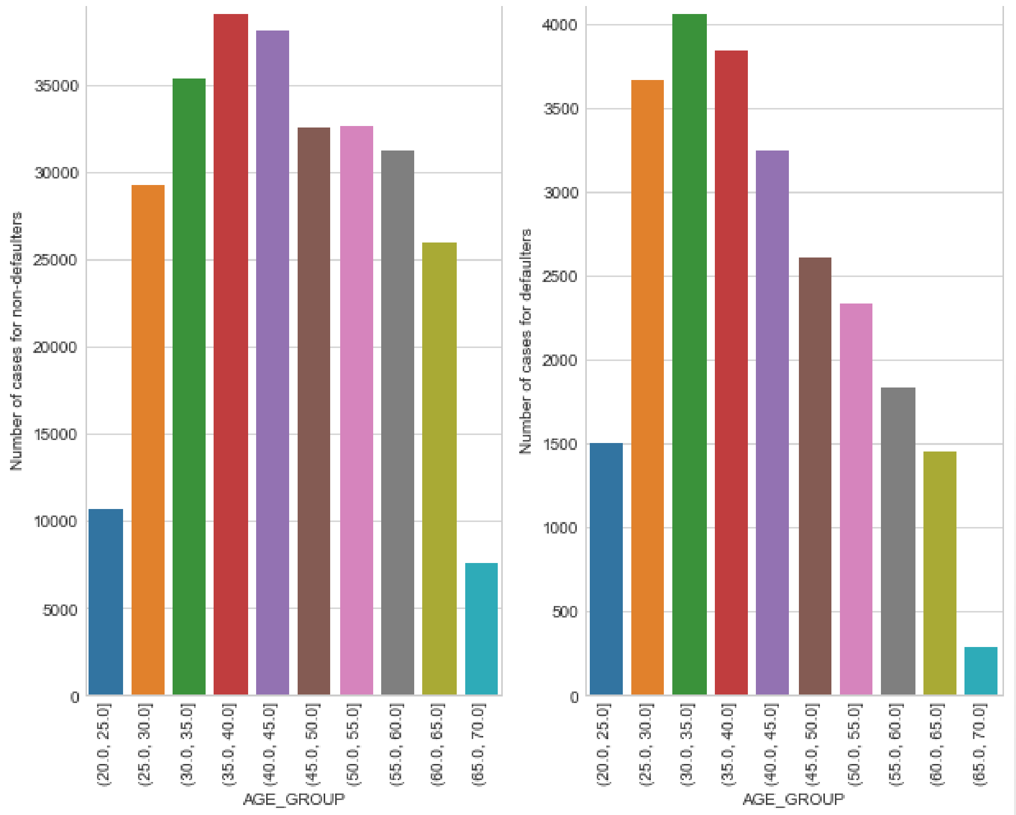


- From the above graph, we can see that there is high risk in providing cash loans compared to revolving loans.

Univariate Categorical Ordered Analysis

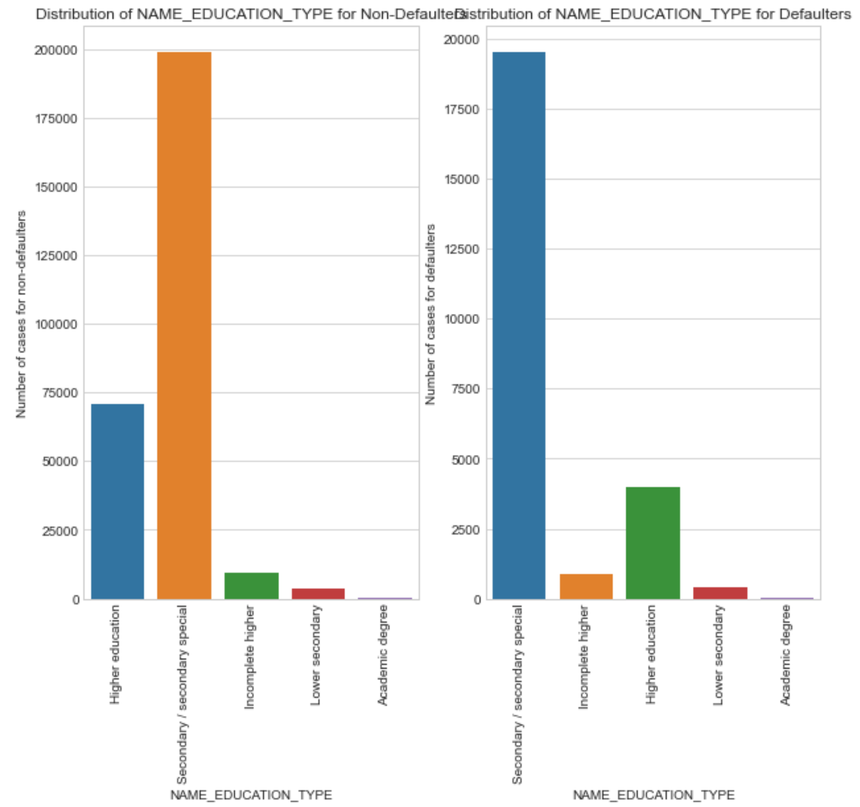
- From the graph, we see that people with low income are more likely to default compared to people with high and very high income. We also see an interesting observation where people with very low income default less compared to people with low income whereas it's assumed vice versa.



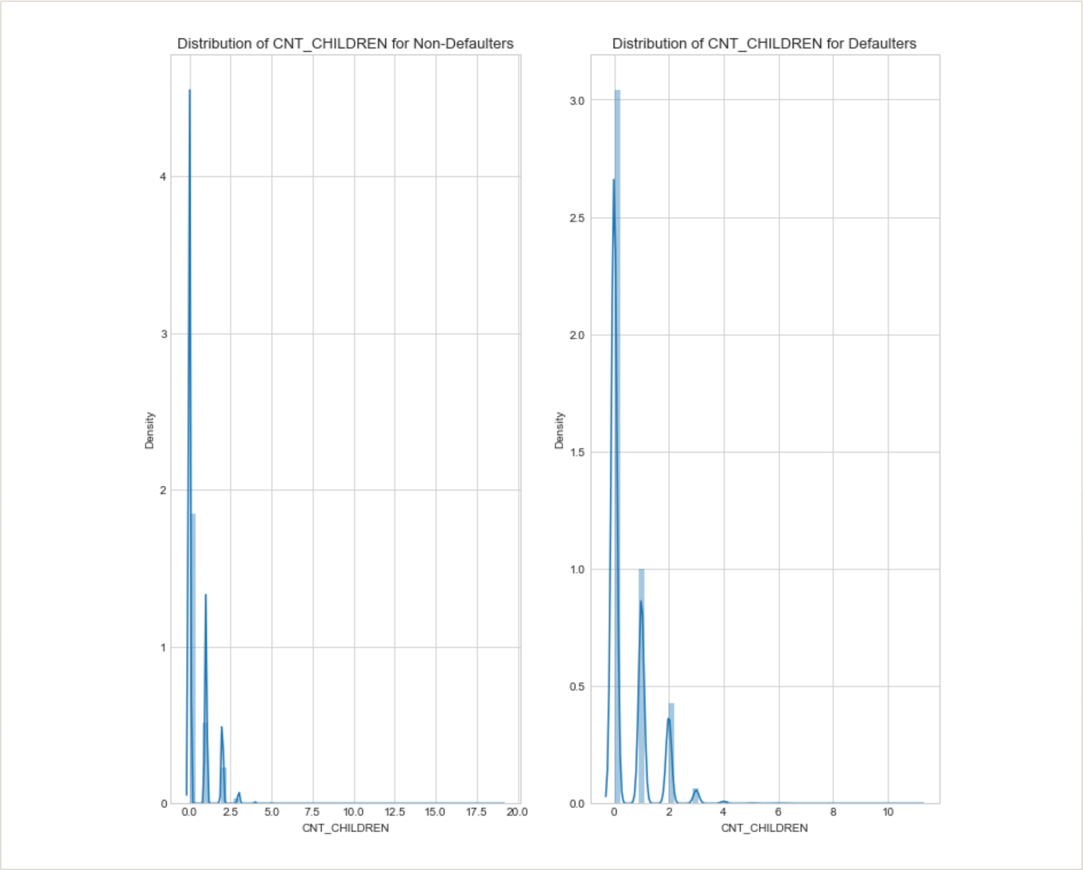


- From the above graph, we observe that most people who apply for loans are in age group between 30-40 yrs.

- 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 for Numerical Variables

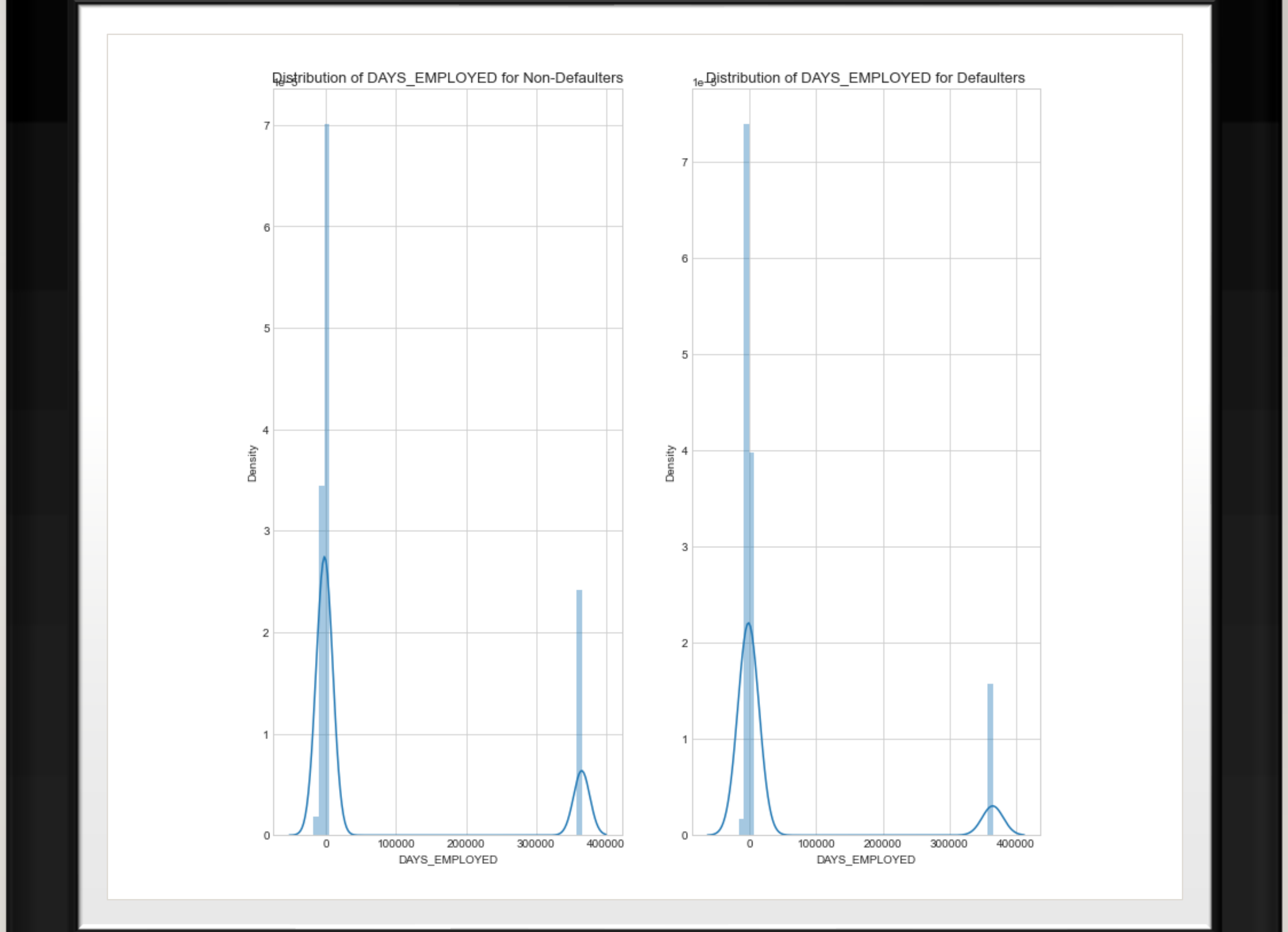


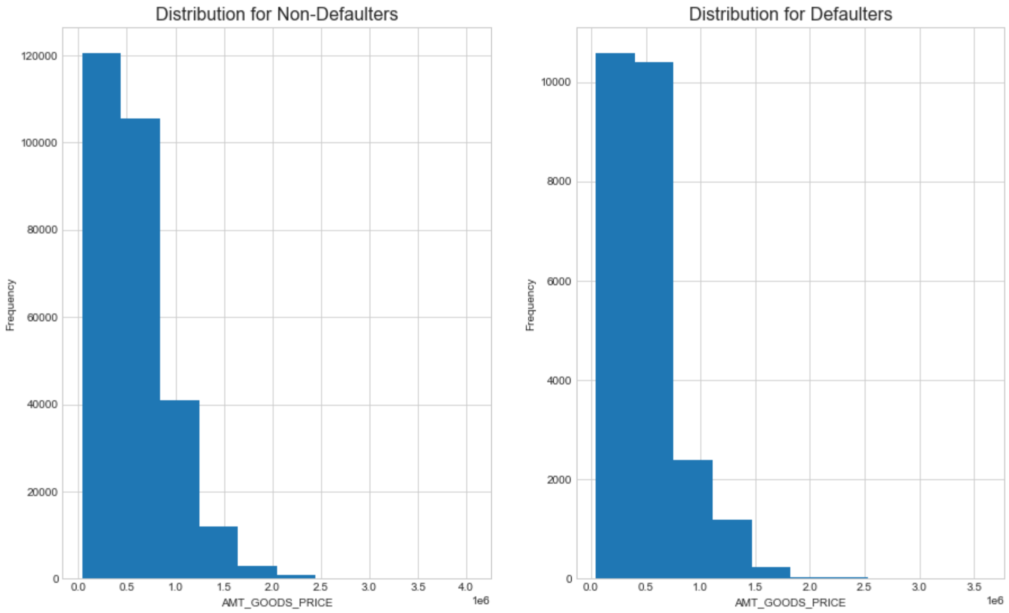
Distribution of CNT_CHILDREN for Non-Defaulters

Distribution of CNT_CHILDREN for Defaulters

- From the above graph we observe that people with 2 or more children are less likely to default.

- Here, we see that people who have been employed for a long time are less likely to default.





Distribution for Non-Defaulters

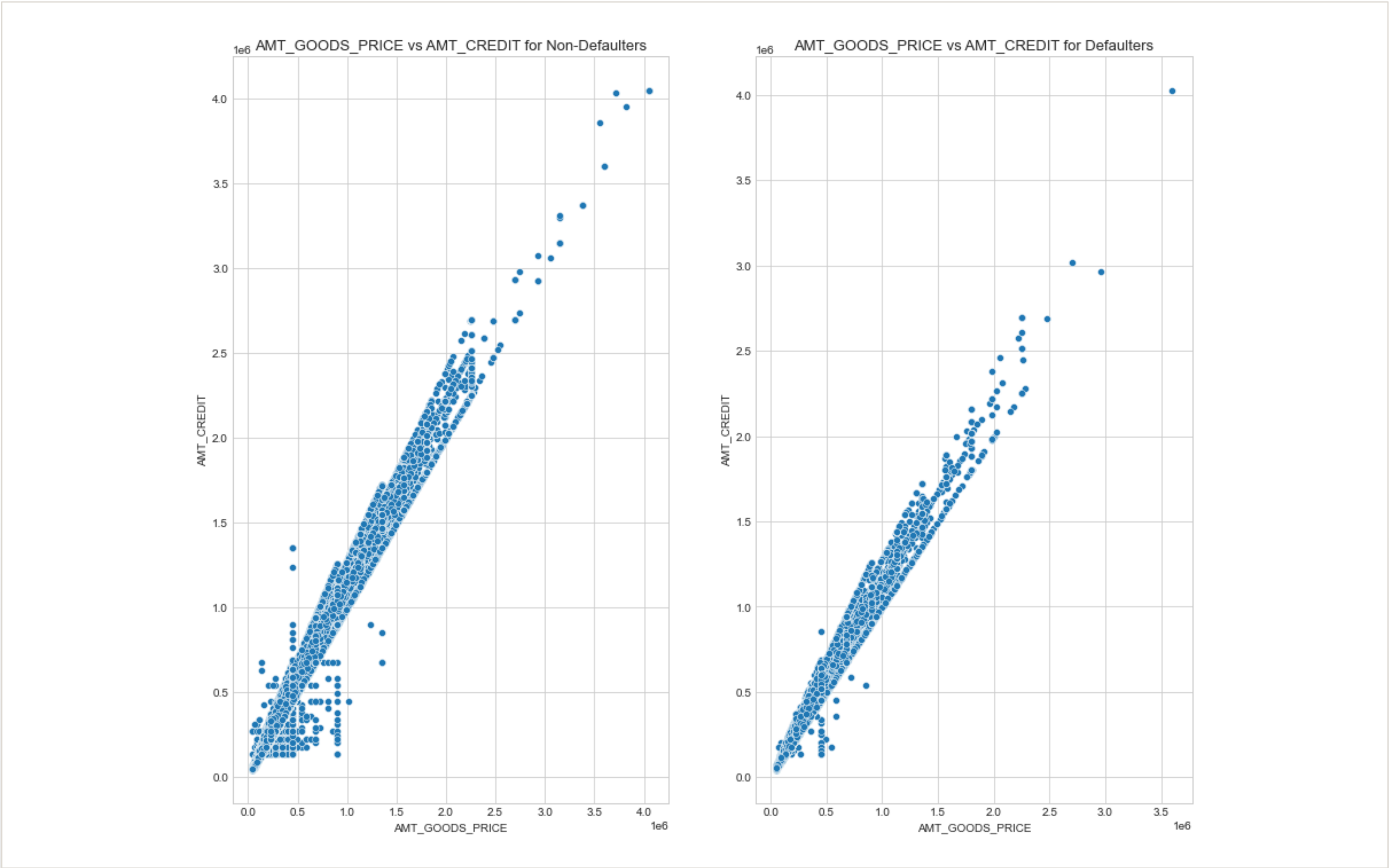
Distribution for Defaulters

- From the above graph, we see that defaults are higher for price of the goods for which the loan is given is lesser (between 0 to 500,000).

Bivariate Analysis

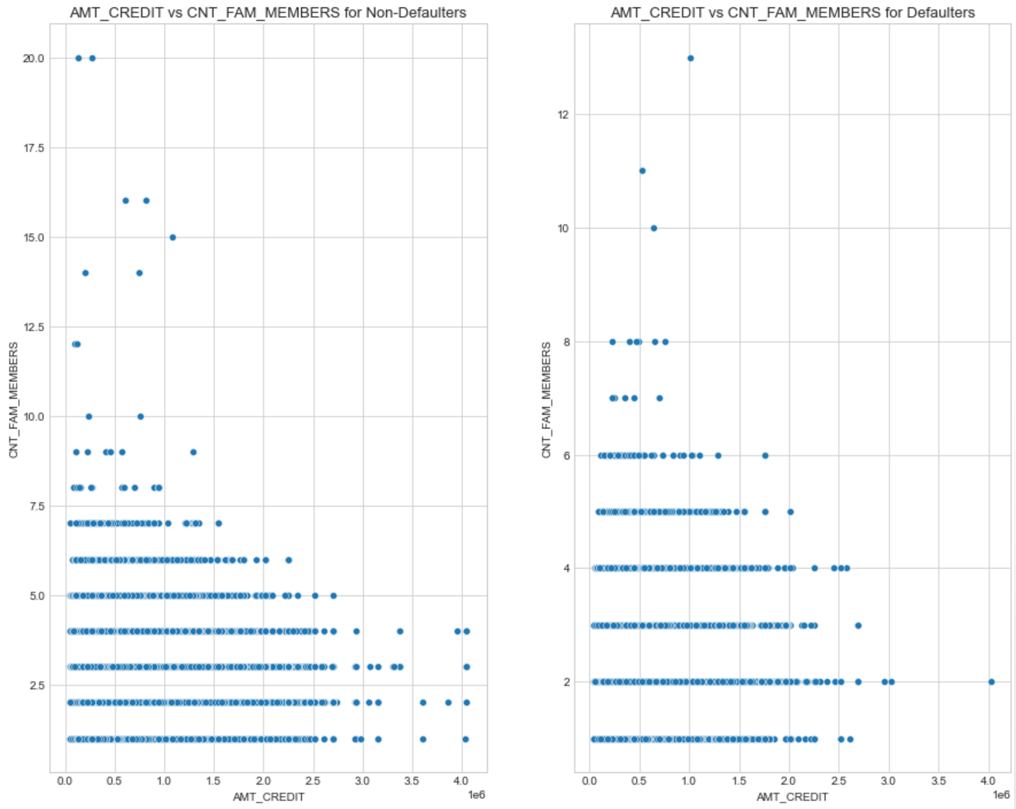
Bivariate Analysis of numerical variables

From the above graph, we see that defaulters are less if price of good is upto 500k and amount credit is also less than 500k.



AMT_GOODS_PRICE vs AMT_CREDIT for Non-Defaulters

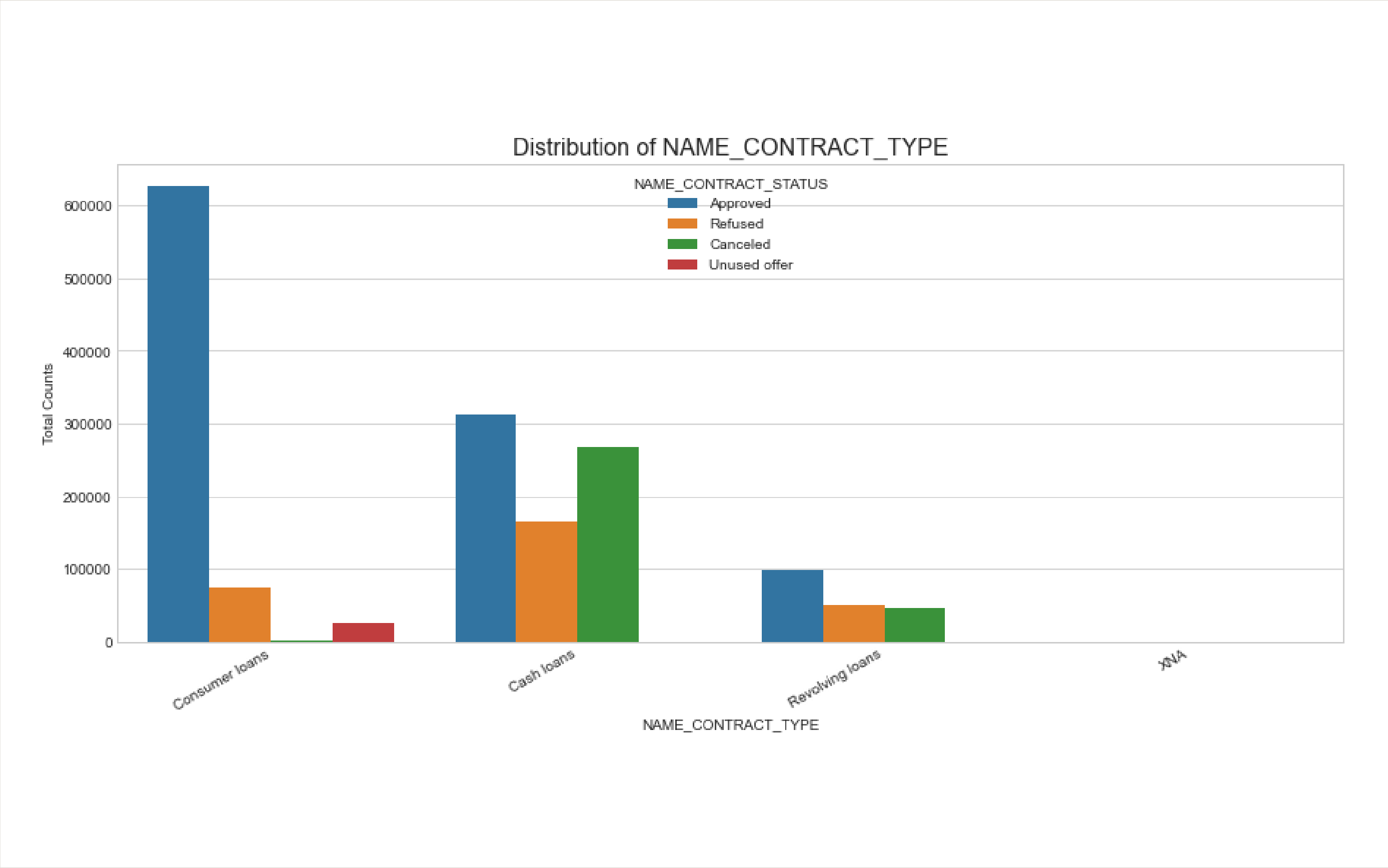
AMT_GOODS_PRICE vs AMT_CREDIT for Defaulters



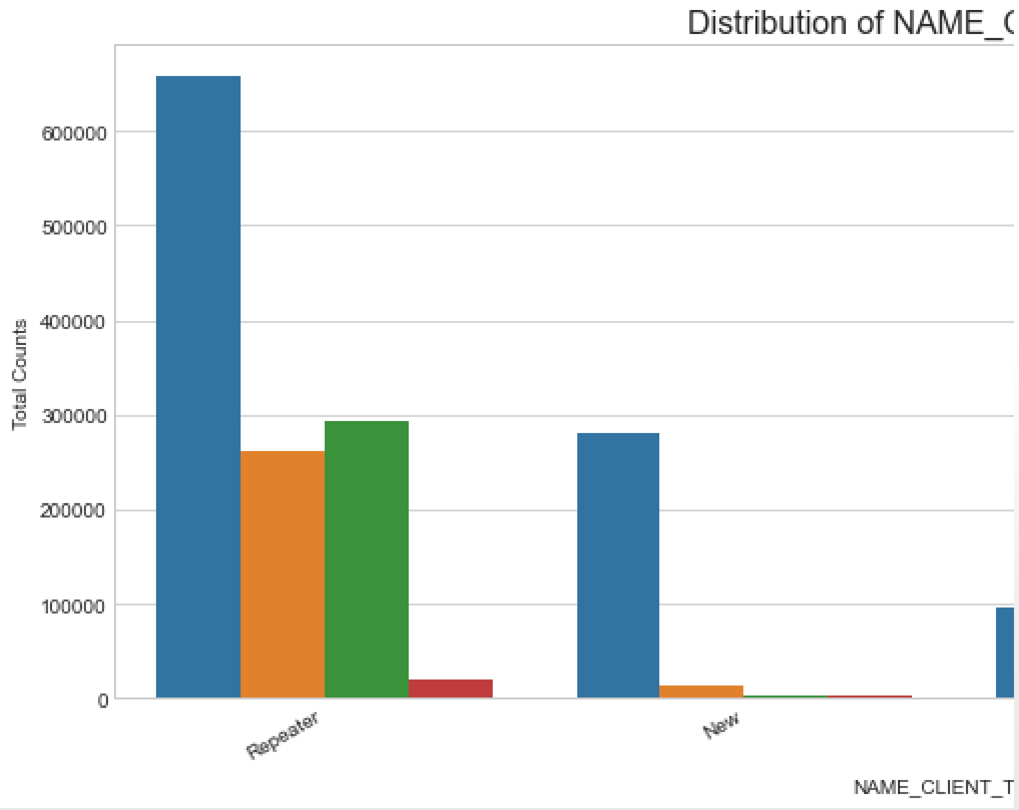
- From the above graph,we can see that the density in the lower left corner is similar in both the case, so the people are equally likely to default if the family is small and the AMT_CREDIT is low. We can observe that larger families and people with larger AMT_CREDIT default less often.

PREVIOUS APPLICATION

Univariate Analysis of Categorical Variables

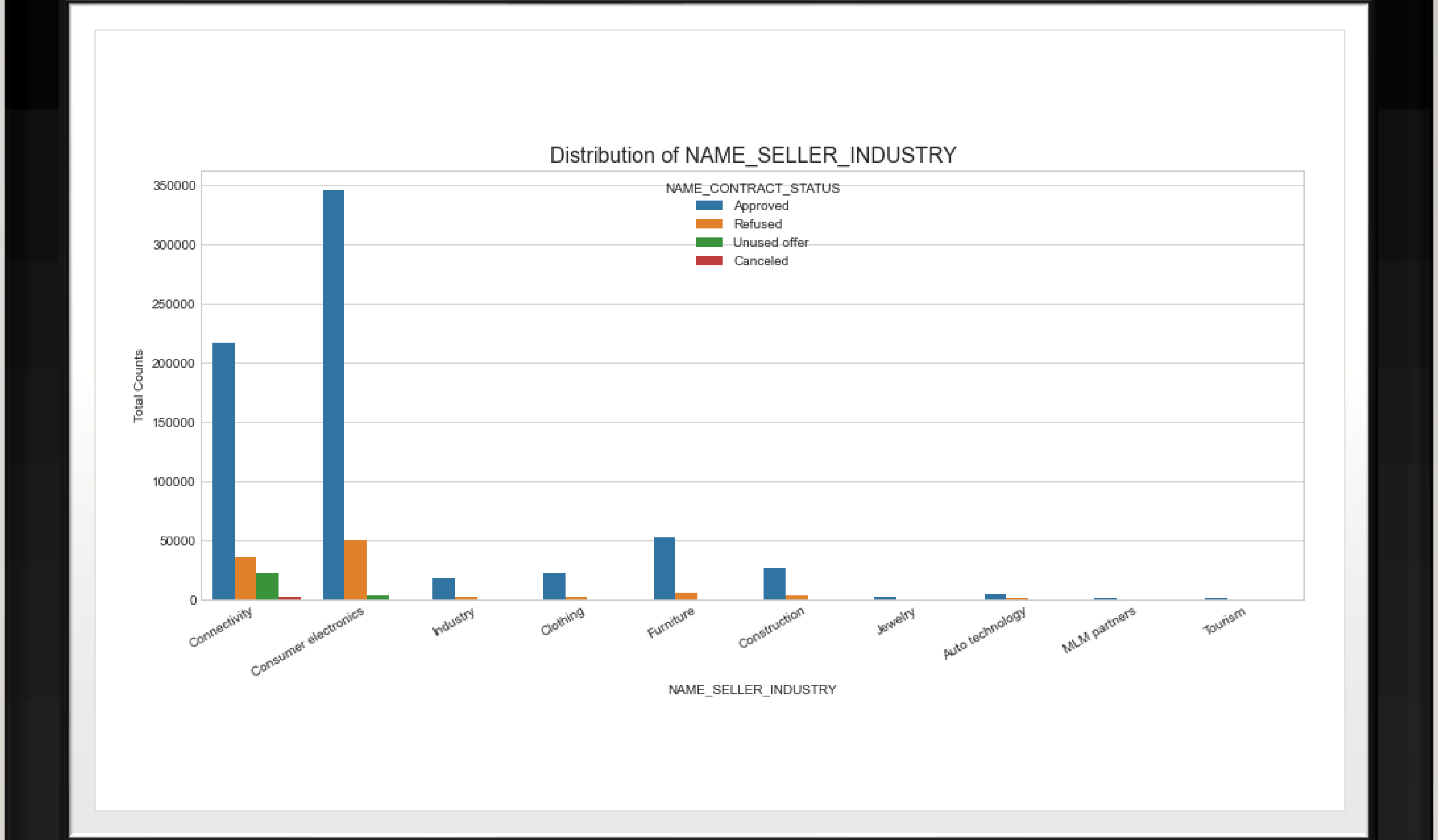


As we can see above that most of the applications are for 'cash loan' and 'consumer loan' even they are canceled.

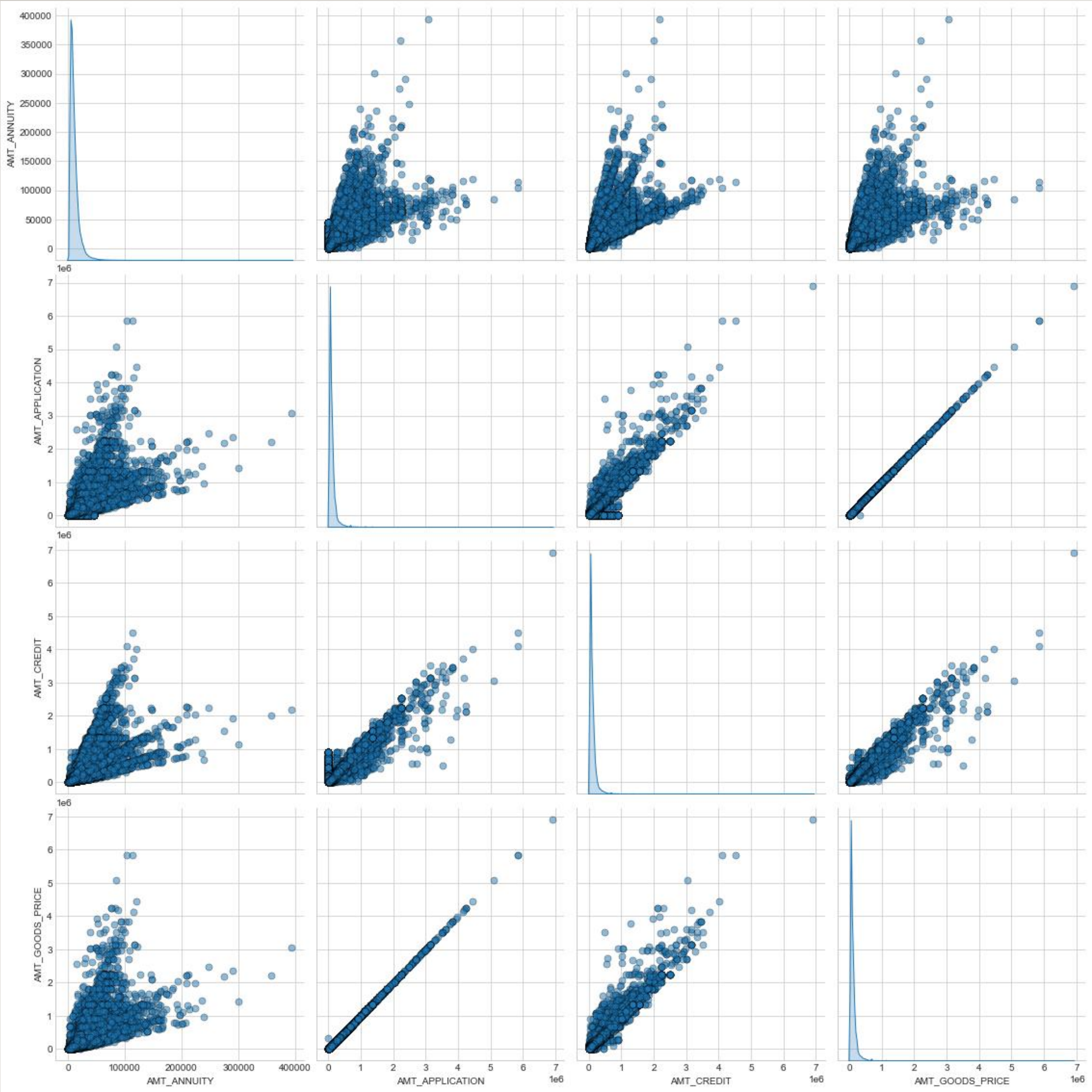


- As we can see that most of the approved loans are for the repeaters category.

We can see that loans are most taken by consumer electronics and most approved are also the same category.



Multivariate Analysis



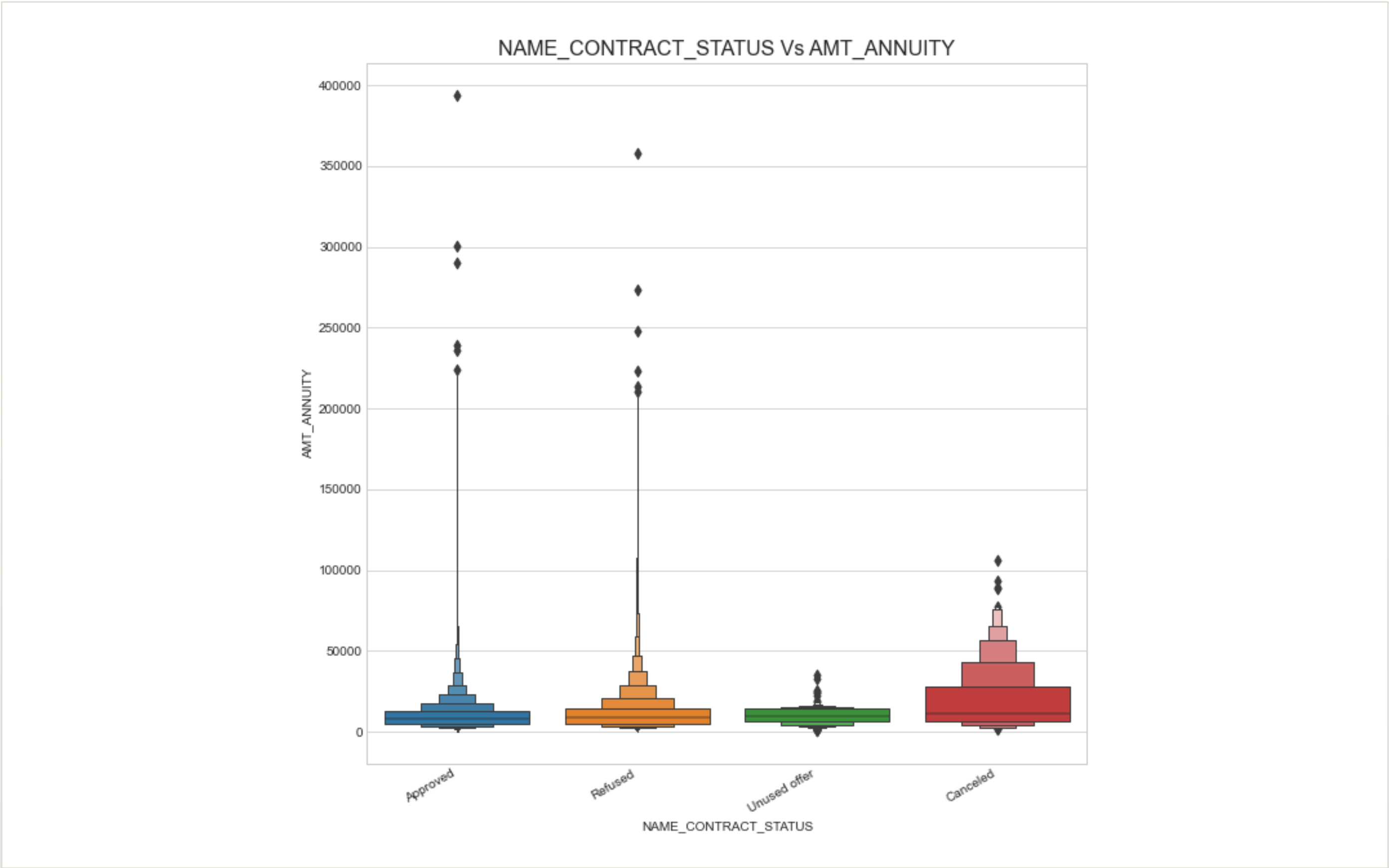
- Annuity of previous application has a very high and positive influence over: (Increase of annuity increases below factors) -How much credit did client asked on the previous application
- -Final credit amount on the previous application that was approved by the bank - Goods price of good that client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

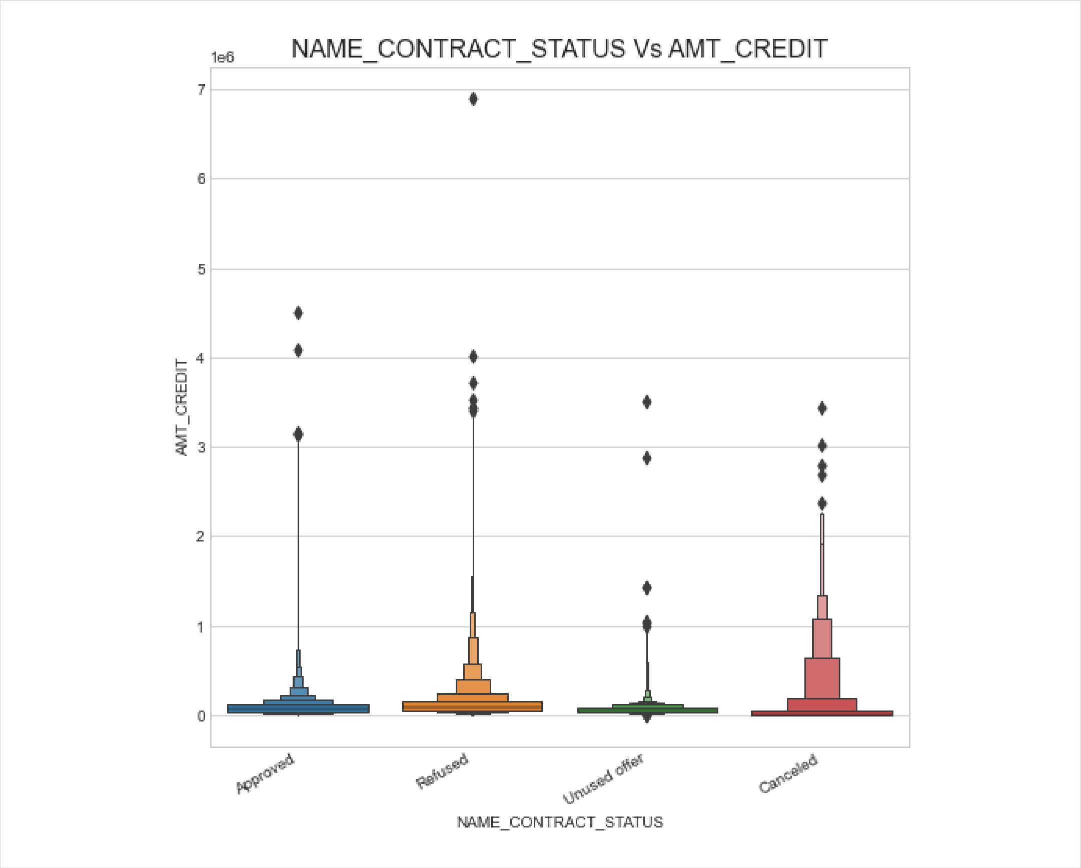
Bivariate Analysis Categorical

Vs

Numerical Columns

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.

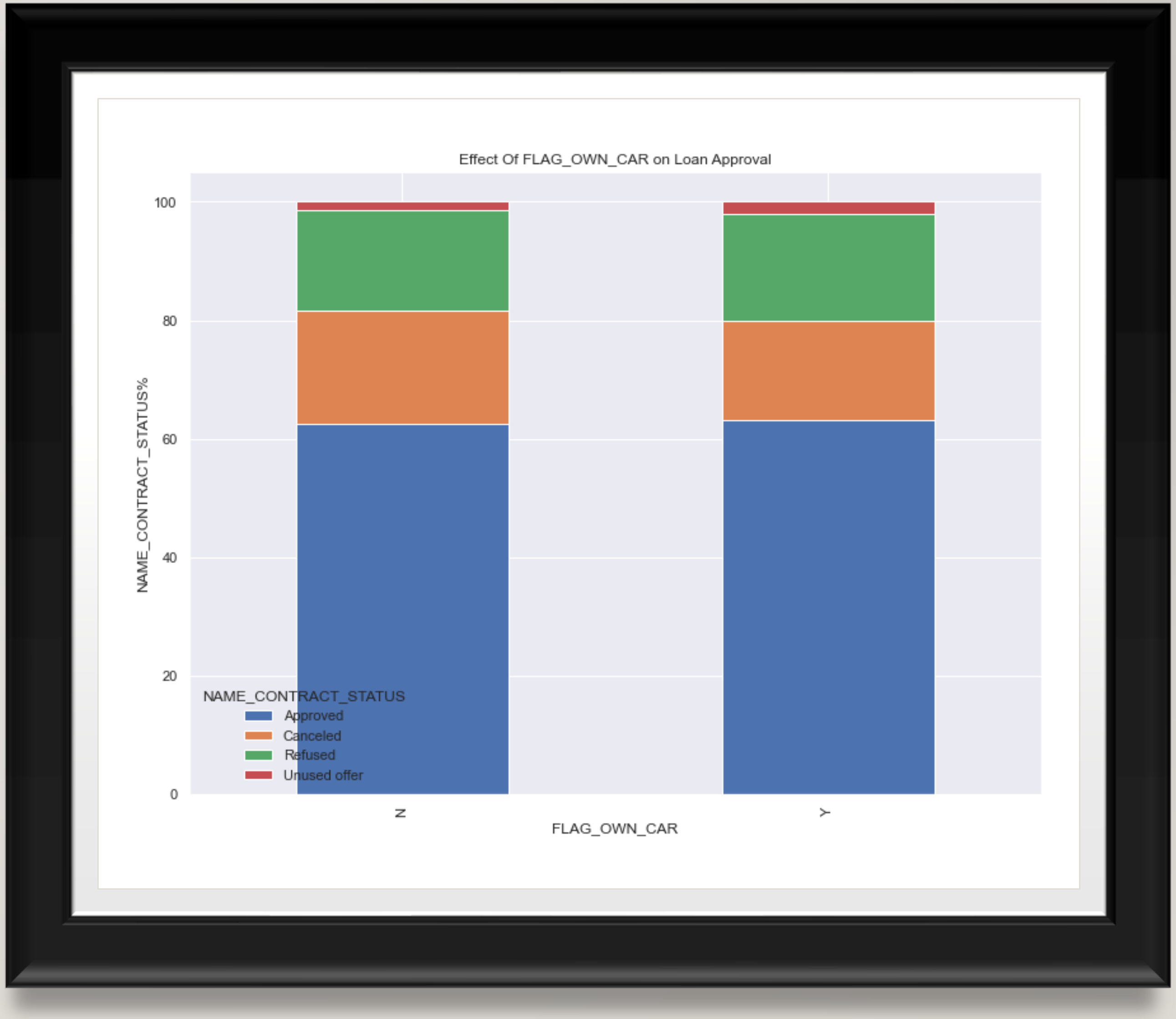


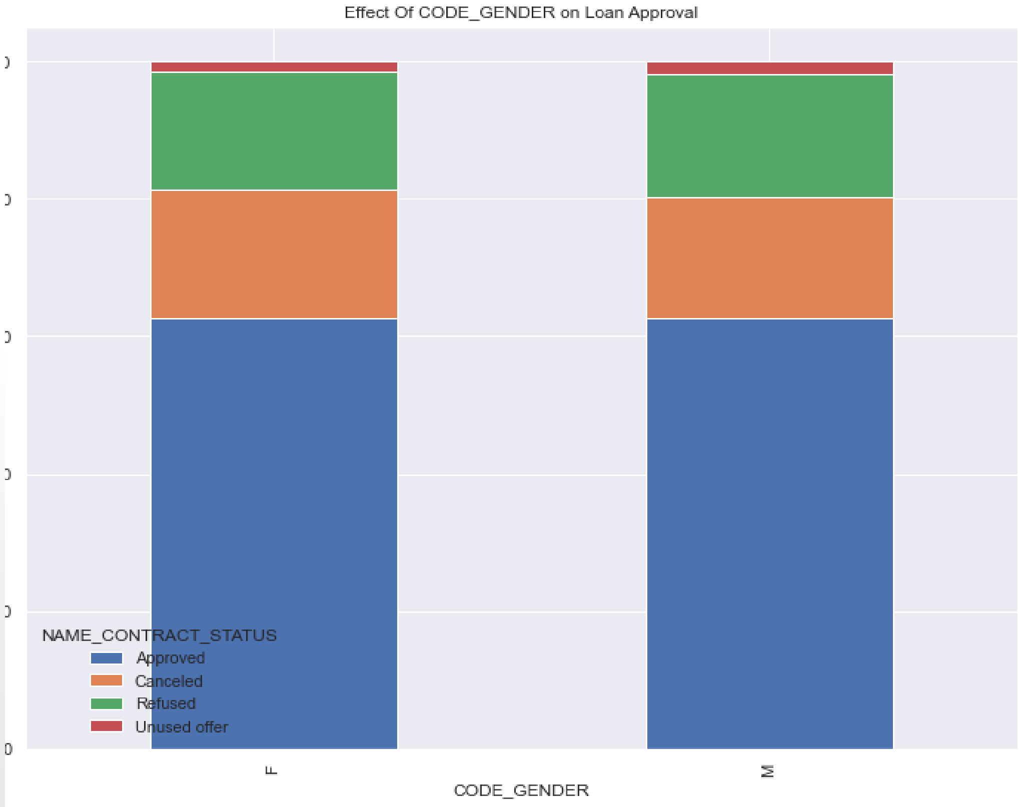


- We can infer that when the AMT_CREDIT is too low, it get's cancelled/unused most of the time.

MERGING APPLICATION DATA & PREVIOUS APPLICATION DATA

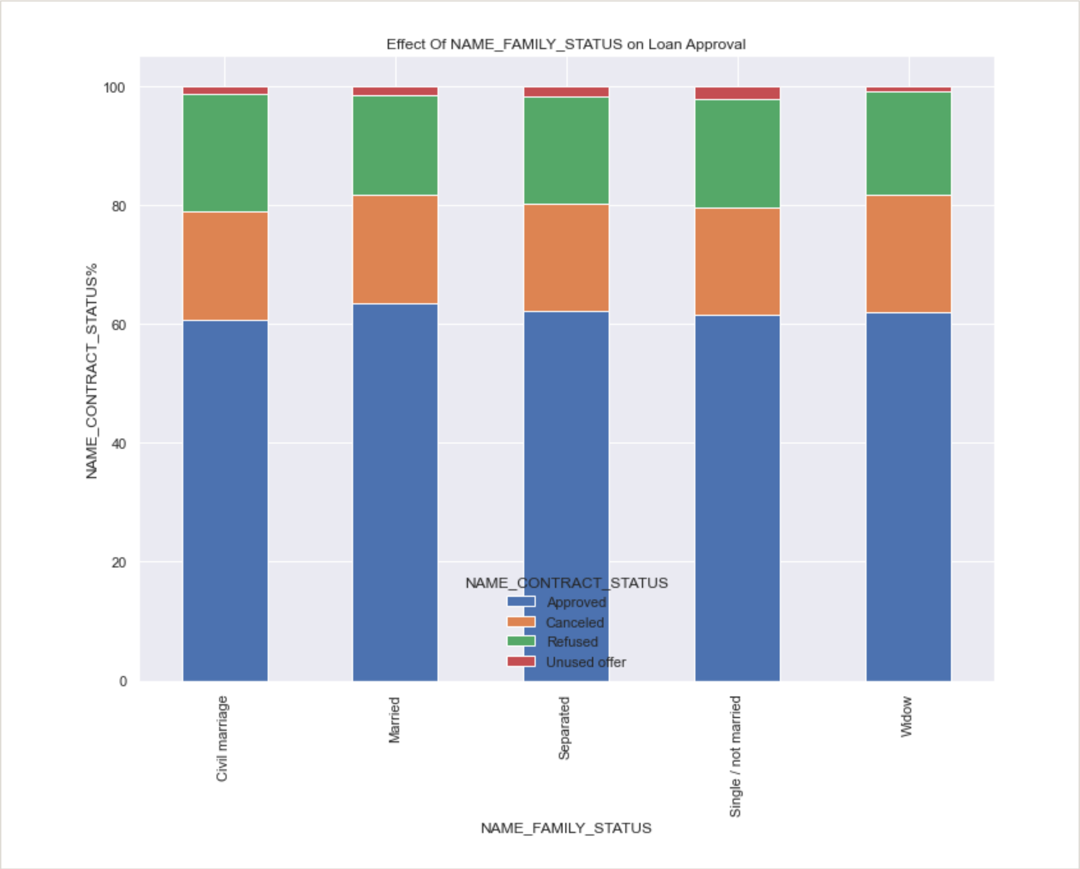
- From the above graph,we see that car ownership doesn't have any effect on application approval or rejection.

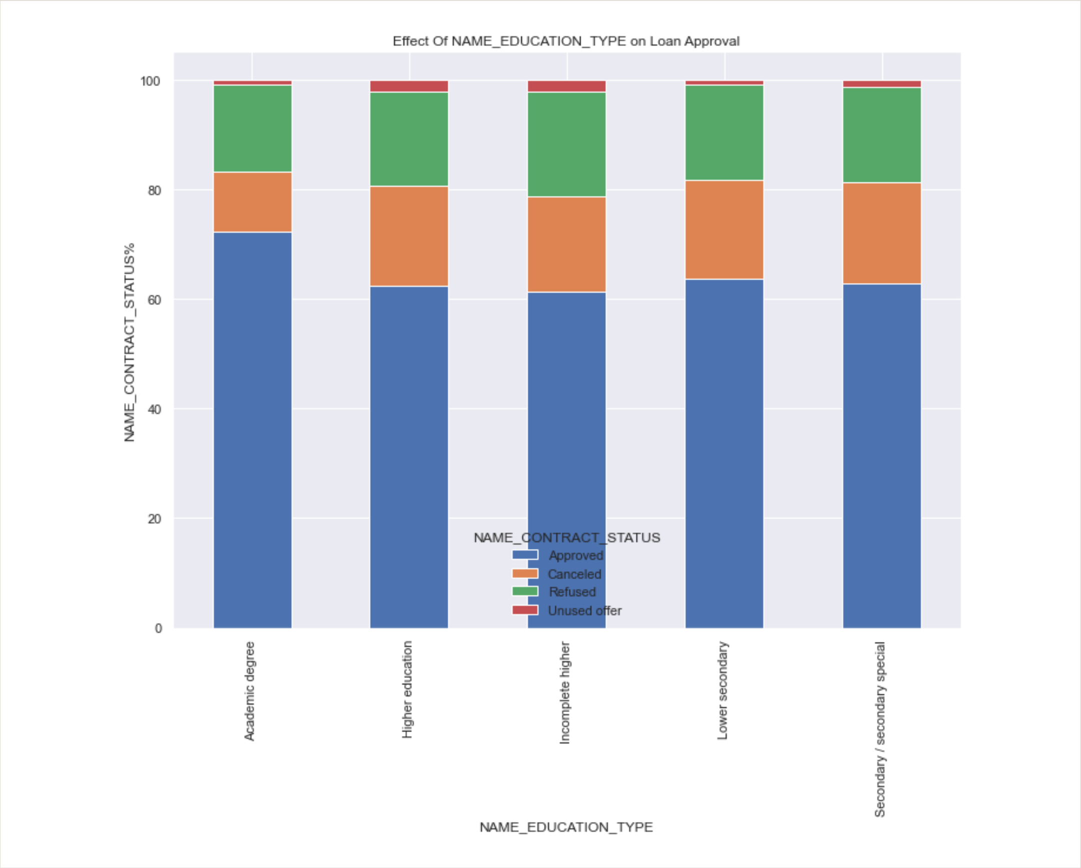




- From the above graph, we can observe that gender of a person has no effect on the loan approval.

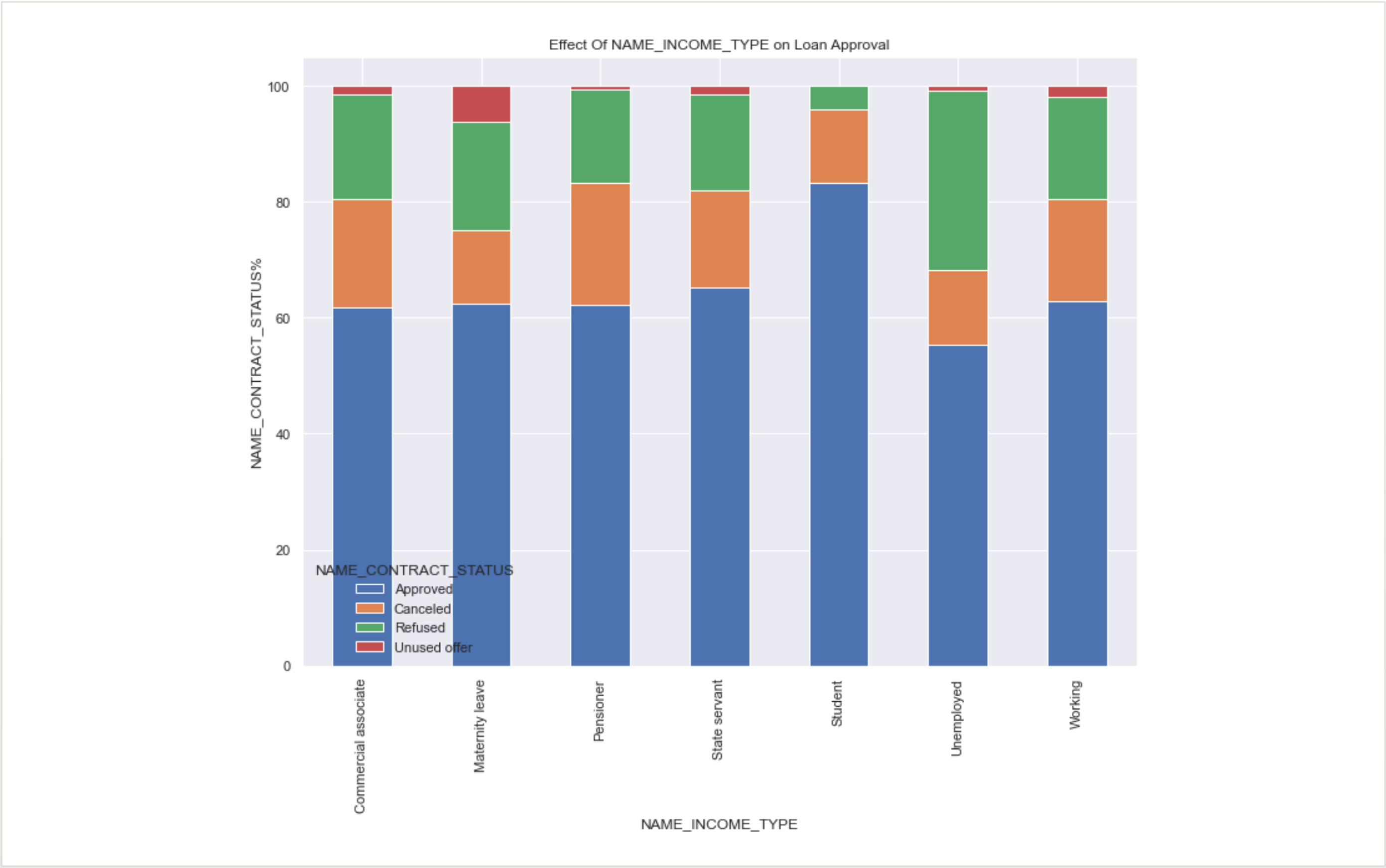
- Here, although the approval rate is slightly higher , not much difference can be seen between family status and approval rate.

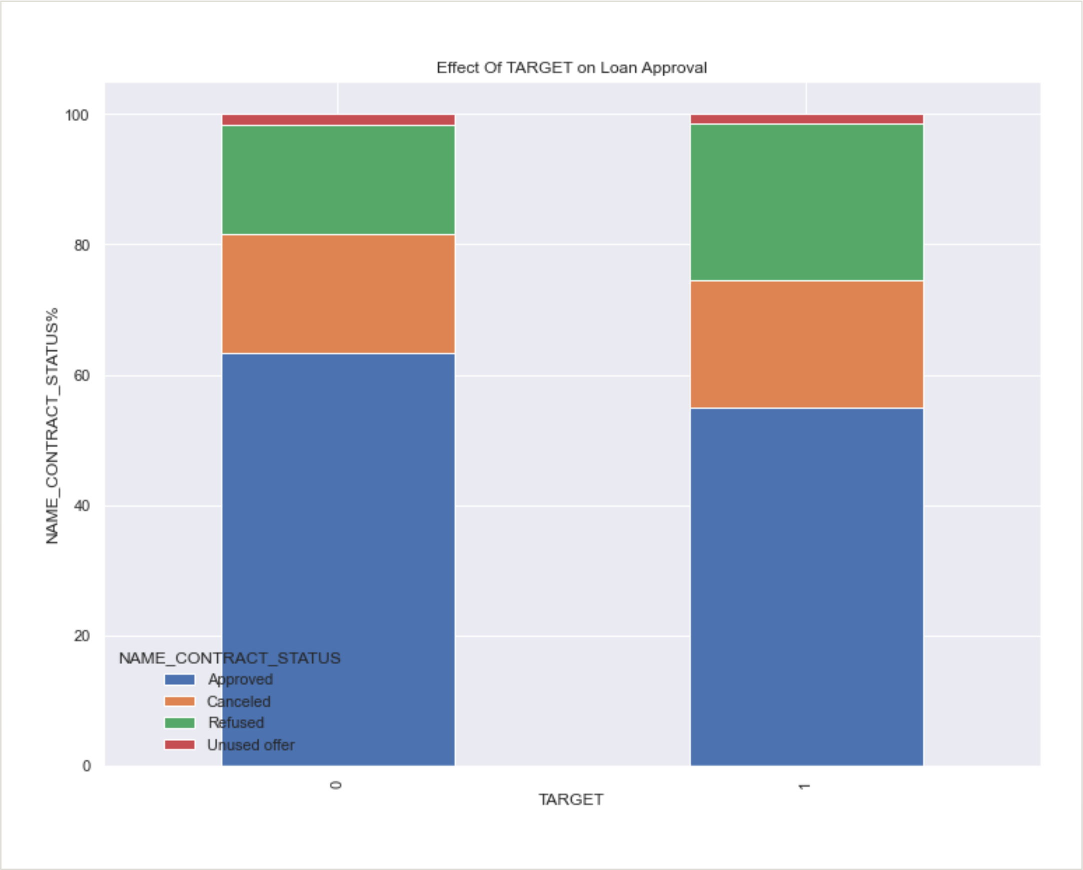




- Here, we can see that the approval rate is higher for the people with an academic degree.

Here,we see that approval rate is higher for the students.This might be because company assume that students will pay the loan once they graduate and start earning.We can also observe the refusal rate is higher for the unemployed.This is due to fact that company is less likely to give a chance to people who aren't employed.





- 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

From this case study, we conclude that loan approval must be provided to those who:

- Had applied for a loan in the past and got an approval by the bank.
- Are either students , businessmen or have high paying jobs.
- Are in the age group between 30 to 40 yrs.
- Have an academic degree.
- Are married with 2 or more children.

Thank
you