EDA Credit : Assignment

-Data Tool Kit 1

A Report by-
Lavkesh Sharma

Introduction:

- This assignment aims to give you an idea of applying EDA in a real business scenario.

- In this assignment, 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.
- 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.
- There are 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 installments of the loan in our sample,
- **All other cases:** All other cases when the payment is paid on time.

Business Objetives:

- 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

Case Study Steps performed:

- **Understanding the data and Inspecting**
  - Data Sourcing
  - Data Inspection
  - Inspecting the null values

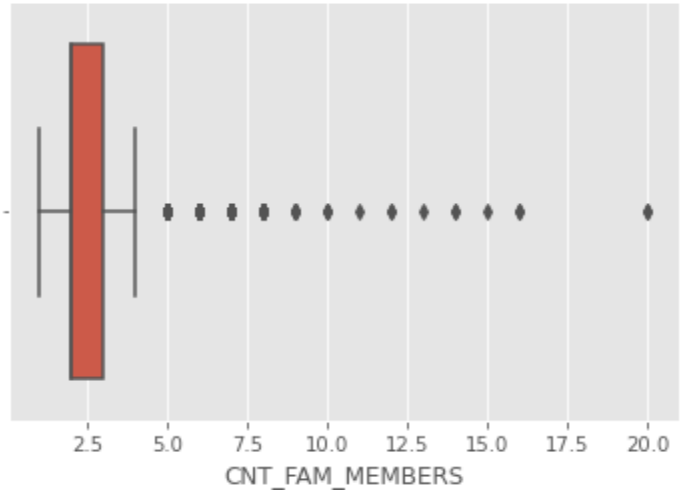
- **Data Manipulation and Data cleaning**
  - Check data types
  - Conversions from negative to positive
  - Outliers

- **Data Analysis**
  - Data imbalance
  - Categories Target 0 and target 1
  - Correlation matrix analysis
  - Univariate Analysis and Bivariate Analysis for numerical variable for target 0 and target 1
  - Load the previous data
  - Join previous and application data

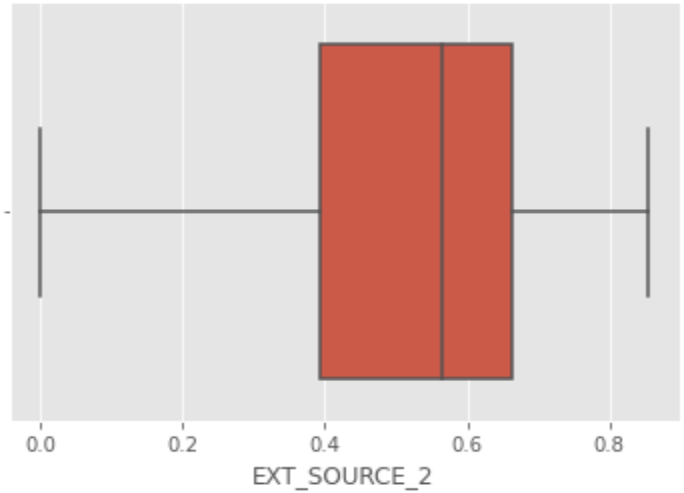
Data type Handling:

- Some of the was changed to numeric data type.
- Some of the columns were having the values as negative, which were taken as absolute value.

Handling of the Missing Values:

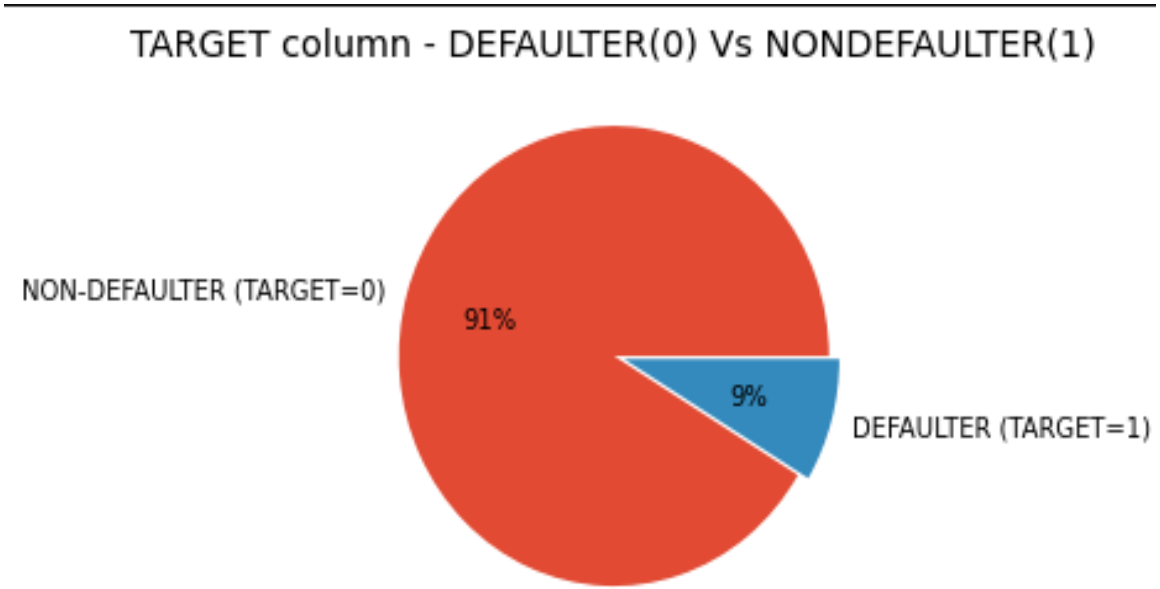


As CNT_FAM_MEMBERS have the outlier , we will have to impute the null values with median



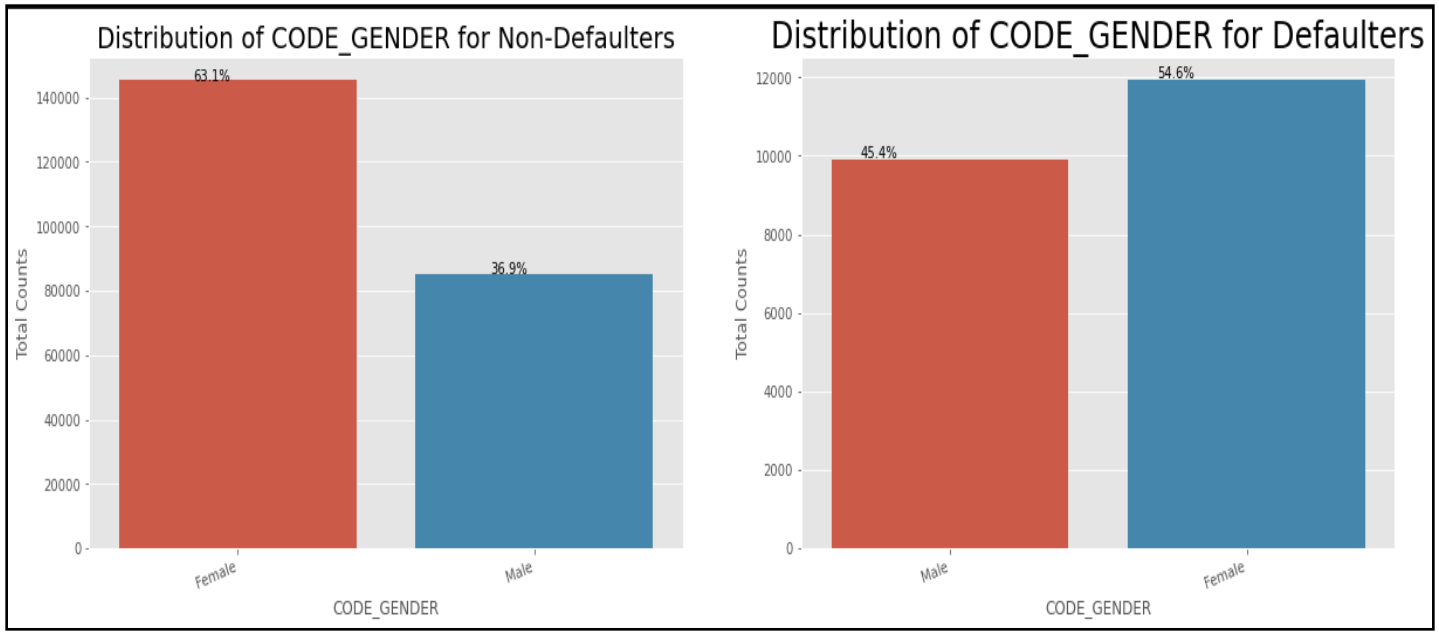
As EXT_SOURCE_2 do not have the outlier , we will have to impute the null values with mean

Checking the Data Imbalance:



Its clear that there is an imbalance between people who defaulted and who didn't default. More than 91% of people didn't default as opposed to 9% who defaulted

Univariate Categorical Ordered Analysis:

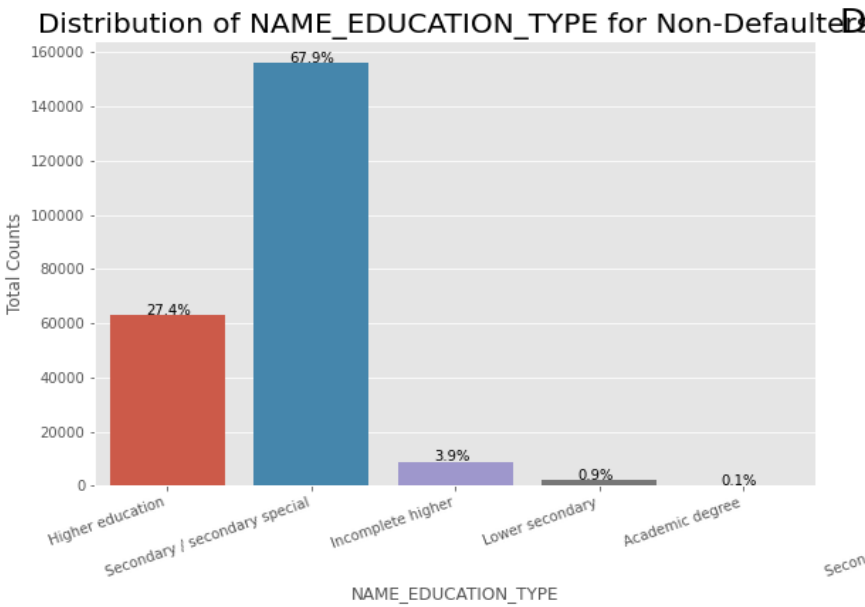


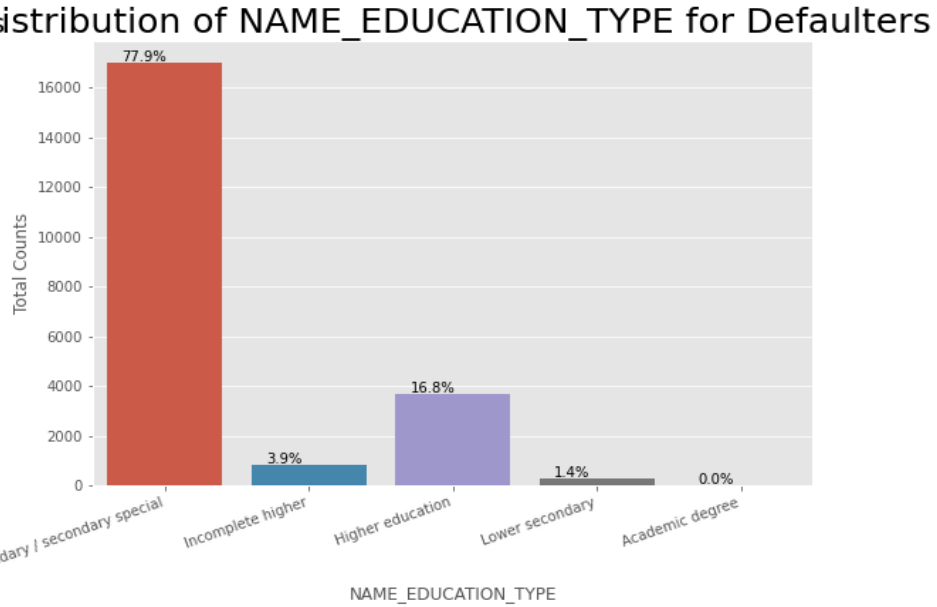
•We can see that Female contribute 63% to the non-defaulters while 55% to the defaulters. We can conclude that

•We see more female applying for loans than males and hence the more number of female defaulters as well.

•But the rate of defaulting of FEMALE is much lower compared to their MALE counterparts.

Univariate Categorical Ordered Analysis:





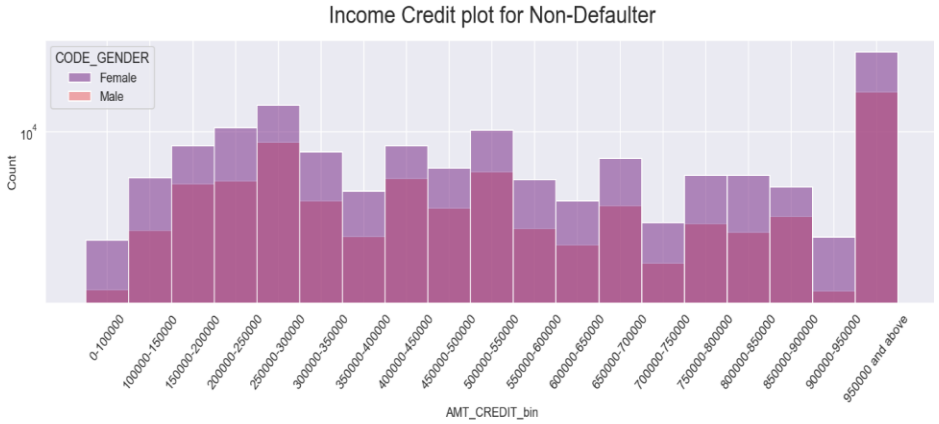
- Almost all of the Education categories are equally likely to default except for the higher educated ones who are less likely to default and secondary educated people are more likely to default
- as above we can see that people with cars are 65.7% to the non-defaulters while 69.5% to the defaulters. We can conclude that While 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.

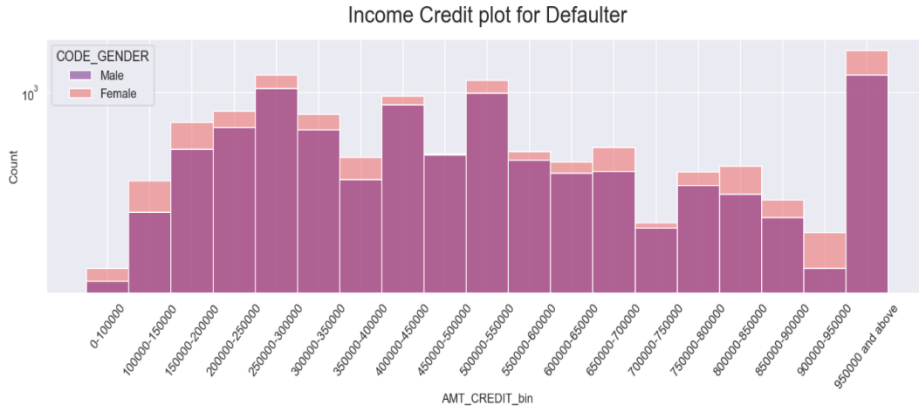
Univariate Categorical Ordered Analysis:

•We can see that Female contribute 63% to the non-defaulters while 55% to the defaulters. We can conclude that

•We see more female applying for loans than males and hence the more number of female defaulters as well.

•But the rate of defaulting of FEMALE is much lower compared to their MALE counterparts.



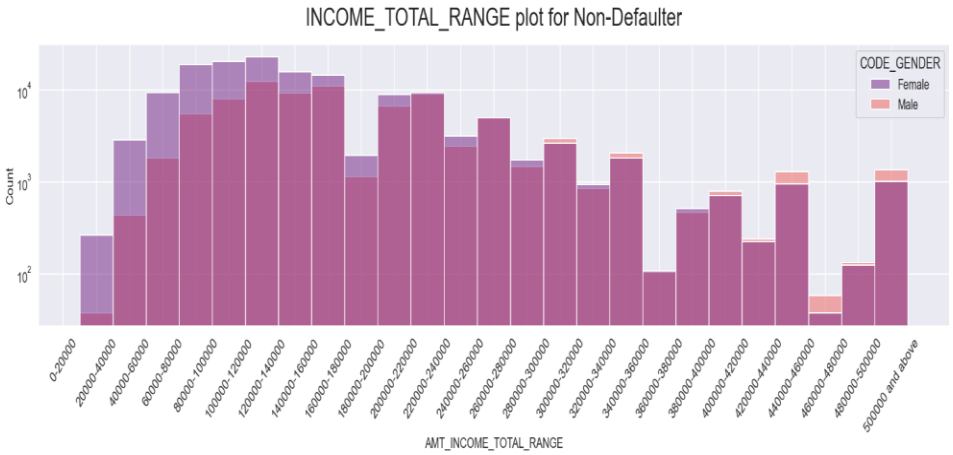


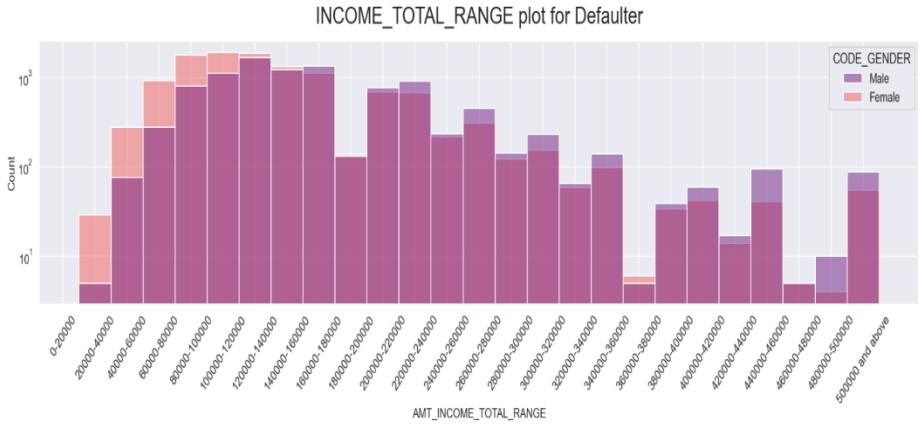
Univariate Categorical Ordered Analysis:

•We can see that Female contribute 63% to the non-defaulters while 55% to the defaulters. We can conclude that

•We see more female applying for loans than males and hence the more number of female defaulters as well.

•But the rate of defaulting of FEMALE is much lower compared to their MALE counterparts.



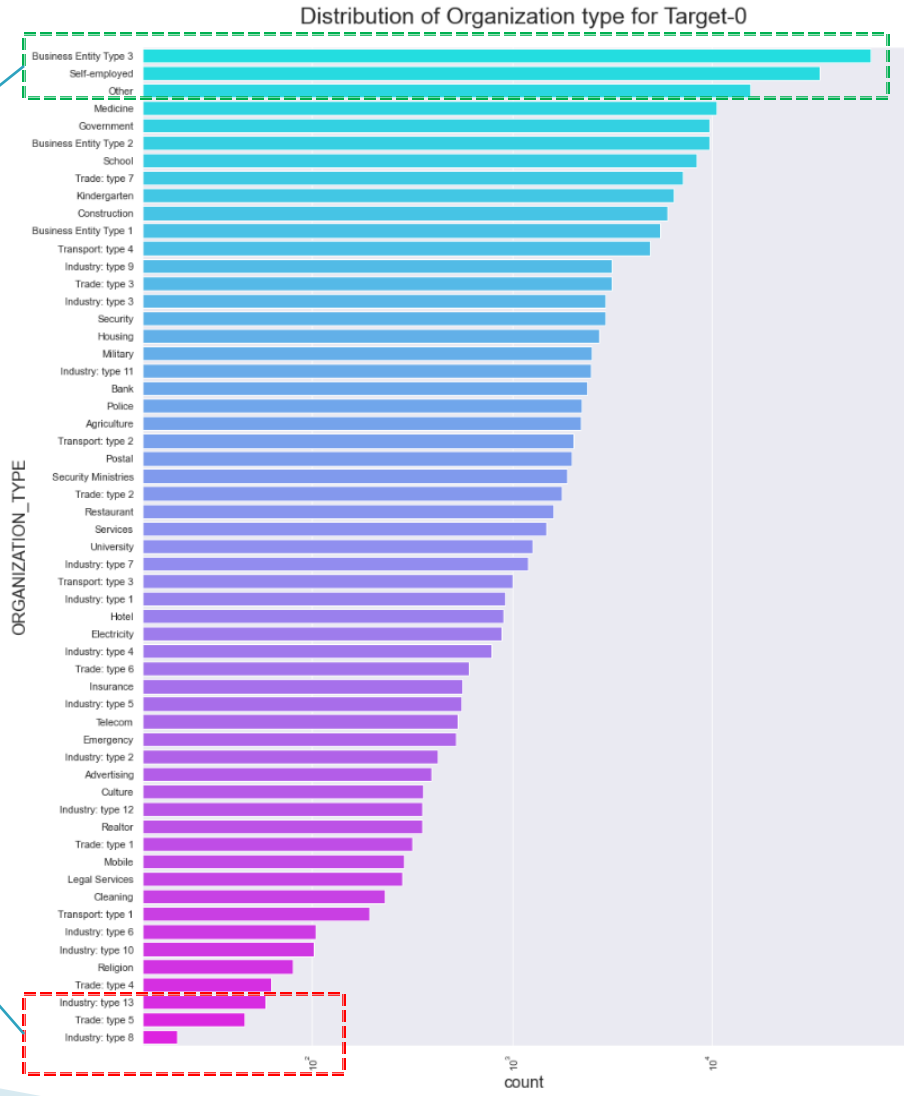


Univariate Categorical Ordered Analysis:

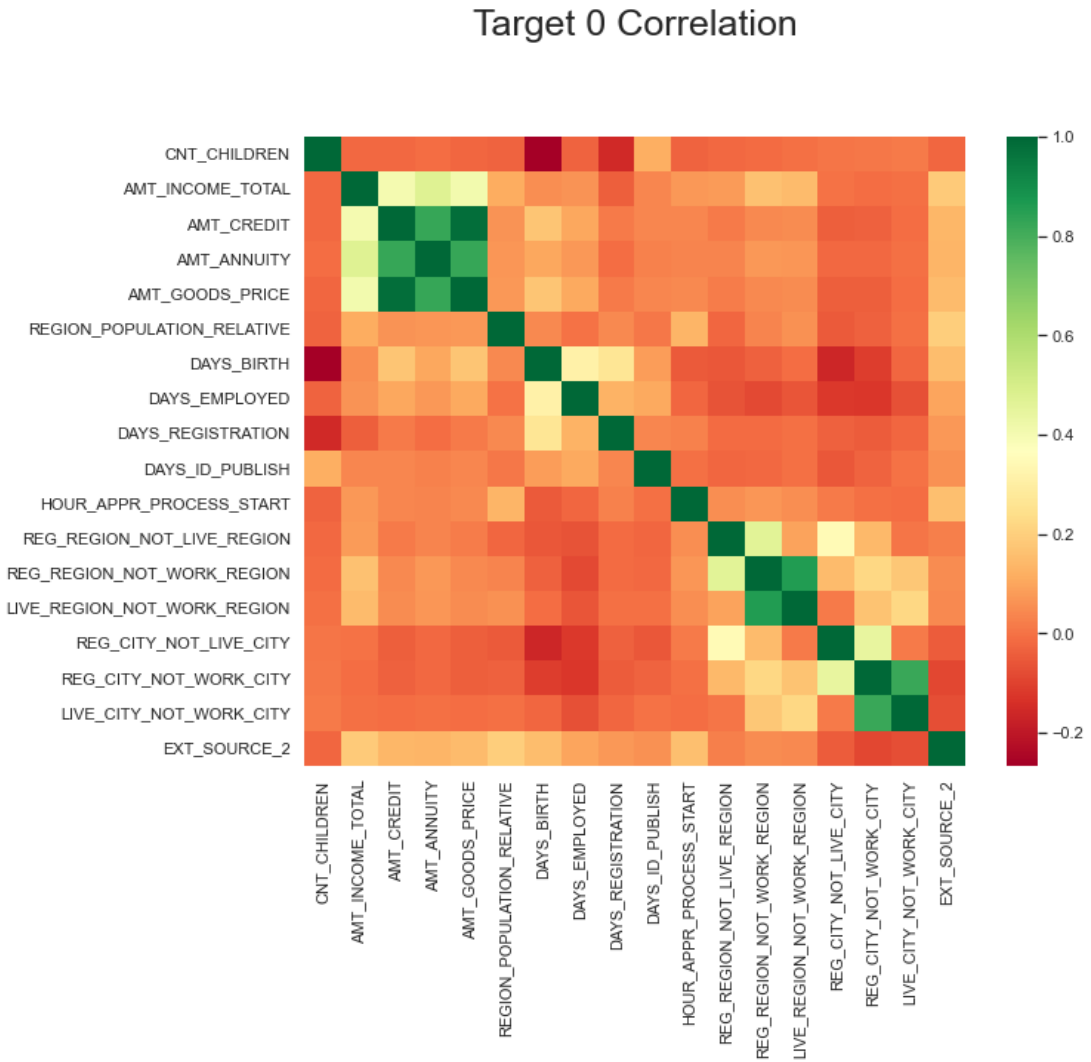
Findings from the Organization type plot:

- we can see that the top 5 categories are 'Business entity Type 3' 'Self employed' , 'Other' , Medicine' and 'Government' which all applied for the loan.

- Industry type 8, type 13 and Trade type 5, type4 and Religion are categories which applied for the loan least times.



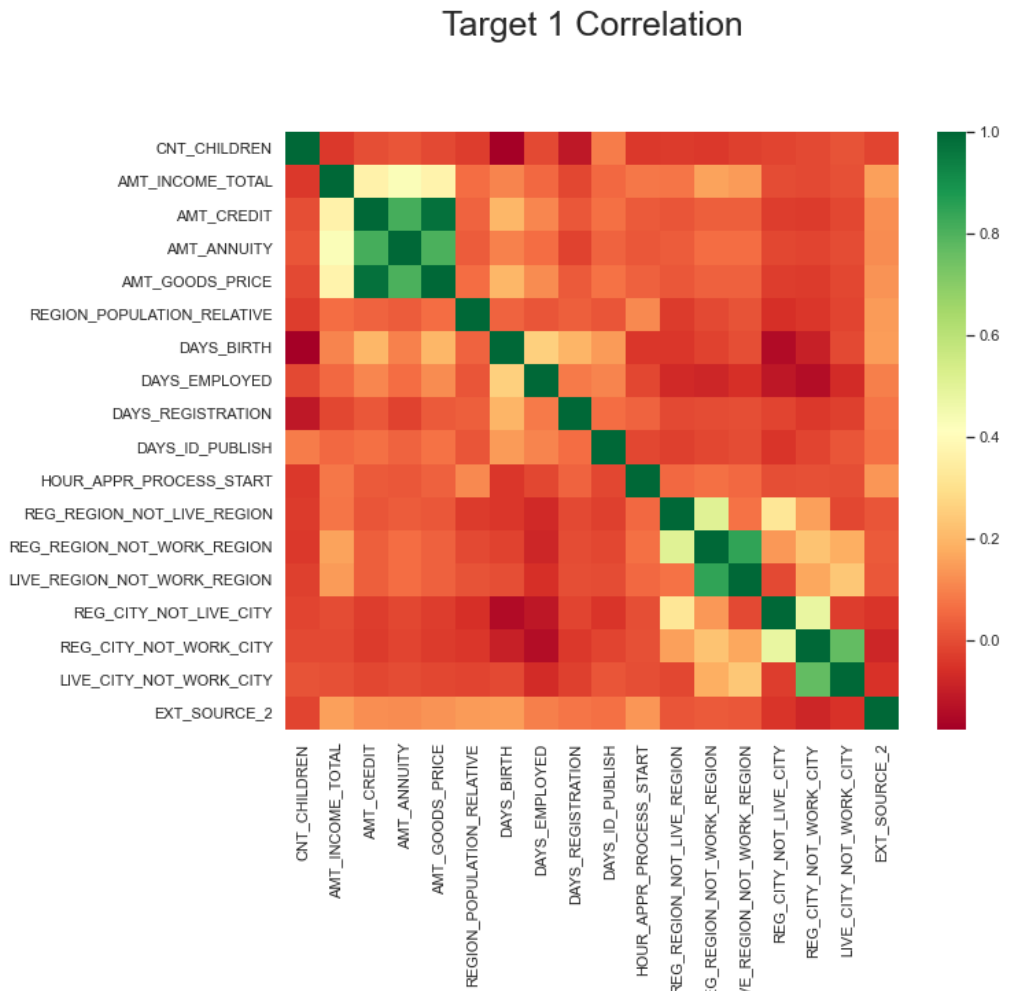
Co-relation for Target_0:



Findings from the Heat Map above for Target0 correlation :

- Credit amount is inversely proportional to the number of children client have, that means Credit amount is higher for less children count client have and vice-versa.
- Income amount is inversely proportional to the number of children customers have, that means more income for less children - customers have and vice-versa.
- Customers with less children have in densely populated area.
- Credit amount is higher to densely populated area.
- The income is also higher in densely populated area.

Co-relation for Target_1::



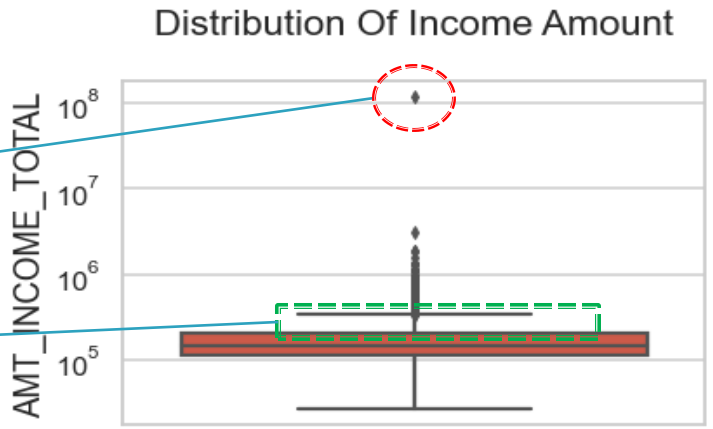
Findings from above Target1 Heat map:

- Customers which have few children, their permanent address not likely to match contact address and vice-versa.

Handling Outliers:

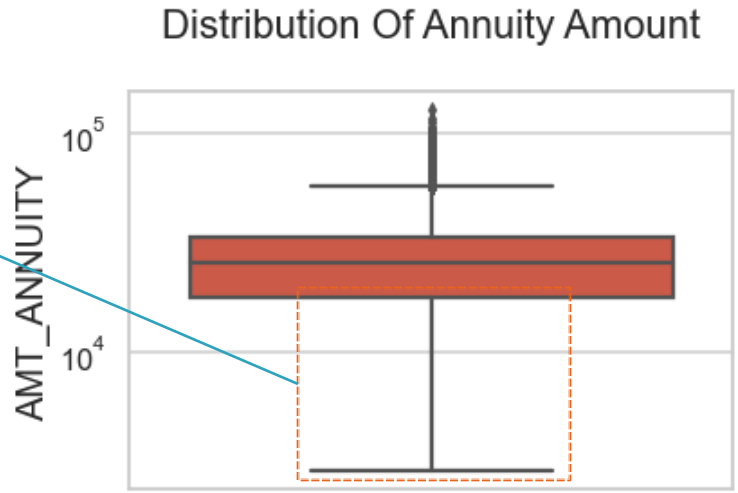
Findings from Income Box Plot

- There is one outliner which is very far from other outliners , it is either by mistake or a exception in data.
- Third quartile is very narrow compared to other and has very less values compare to other quartiles.



Findings from Annuity Amount Box Plot:

- Most of the customer fall in the first quartile.

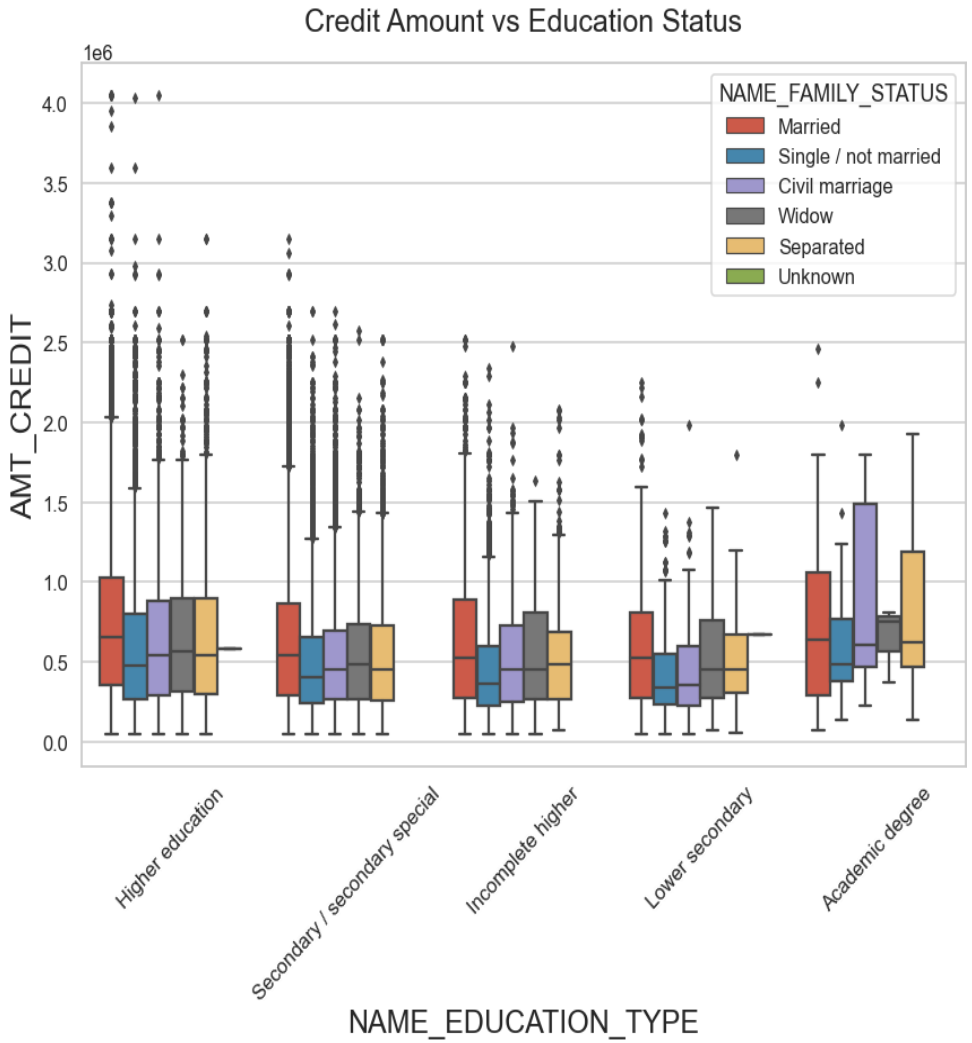


BIVARIATE ANALYSIS :

Findings from the Credit Amount vs Education Status:

➢Customer which all have Academic degree along with Family Status of 'Civil marriage' , 'Married', 'Separated' are having higher credits than others.

➢Higher educated clients along with Family Status either of 'Married' ,'Single', 'Civil Marriage' have more outliers.

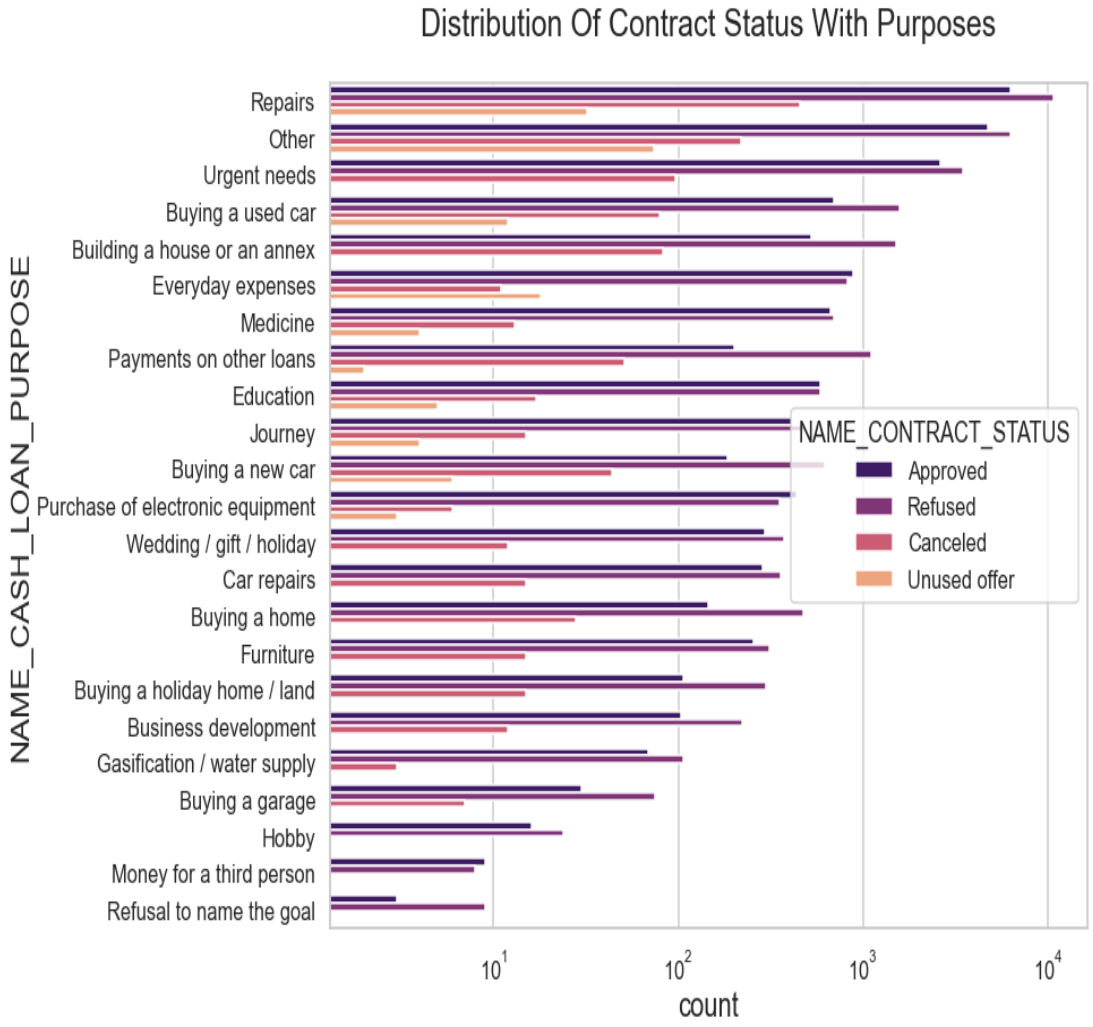


BIVARIATE ANALYSIS :

Findings from Above Count Plot:

➢'Payments on other loans' has most refused rate and 'Buying a new car' also has most refused rate than approved.

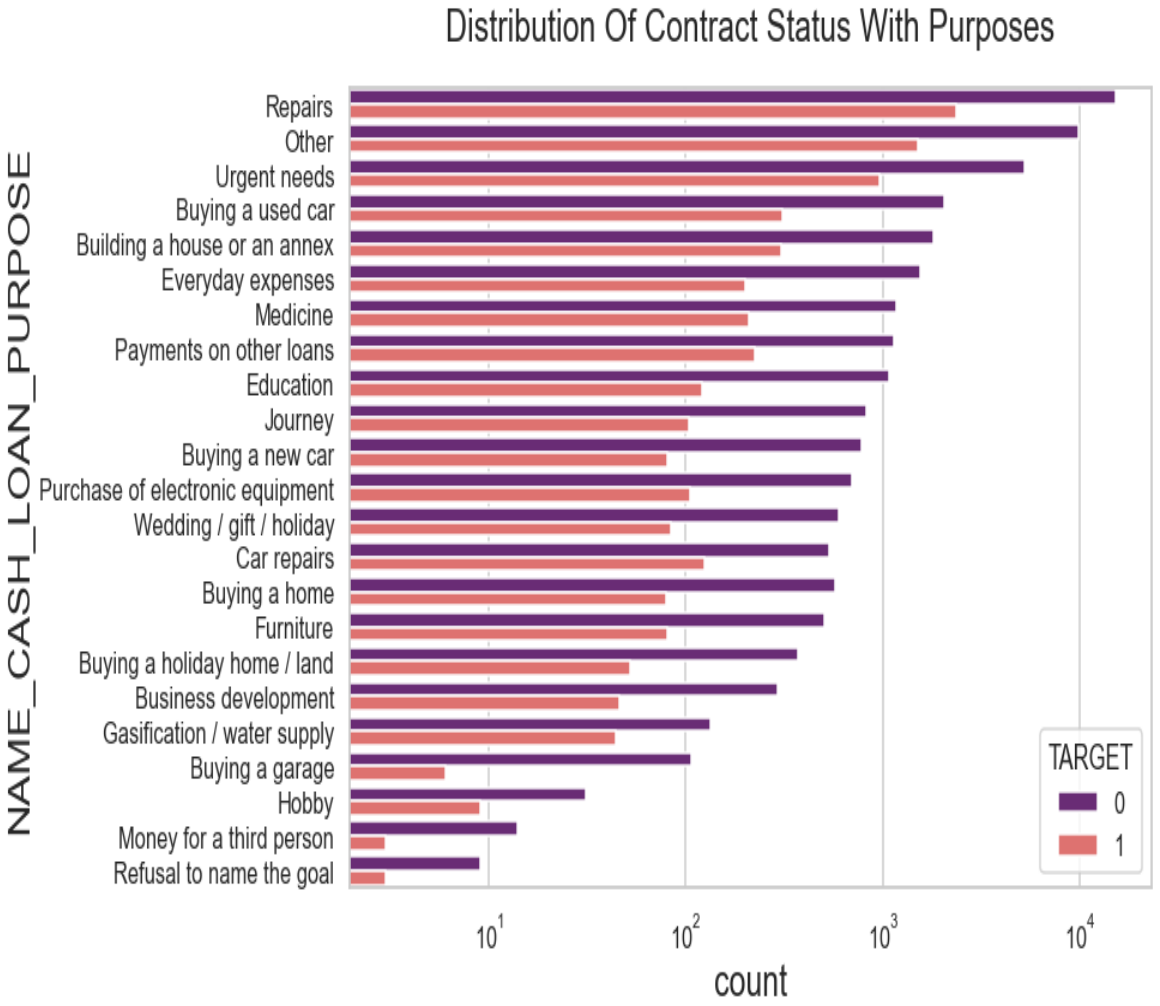
➢'Education' purposes has the same approved and refused rate.



BIVARIATE ANALYSIS :

Findings from above count plot:

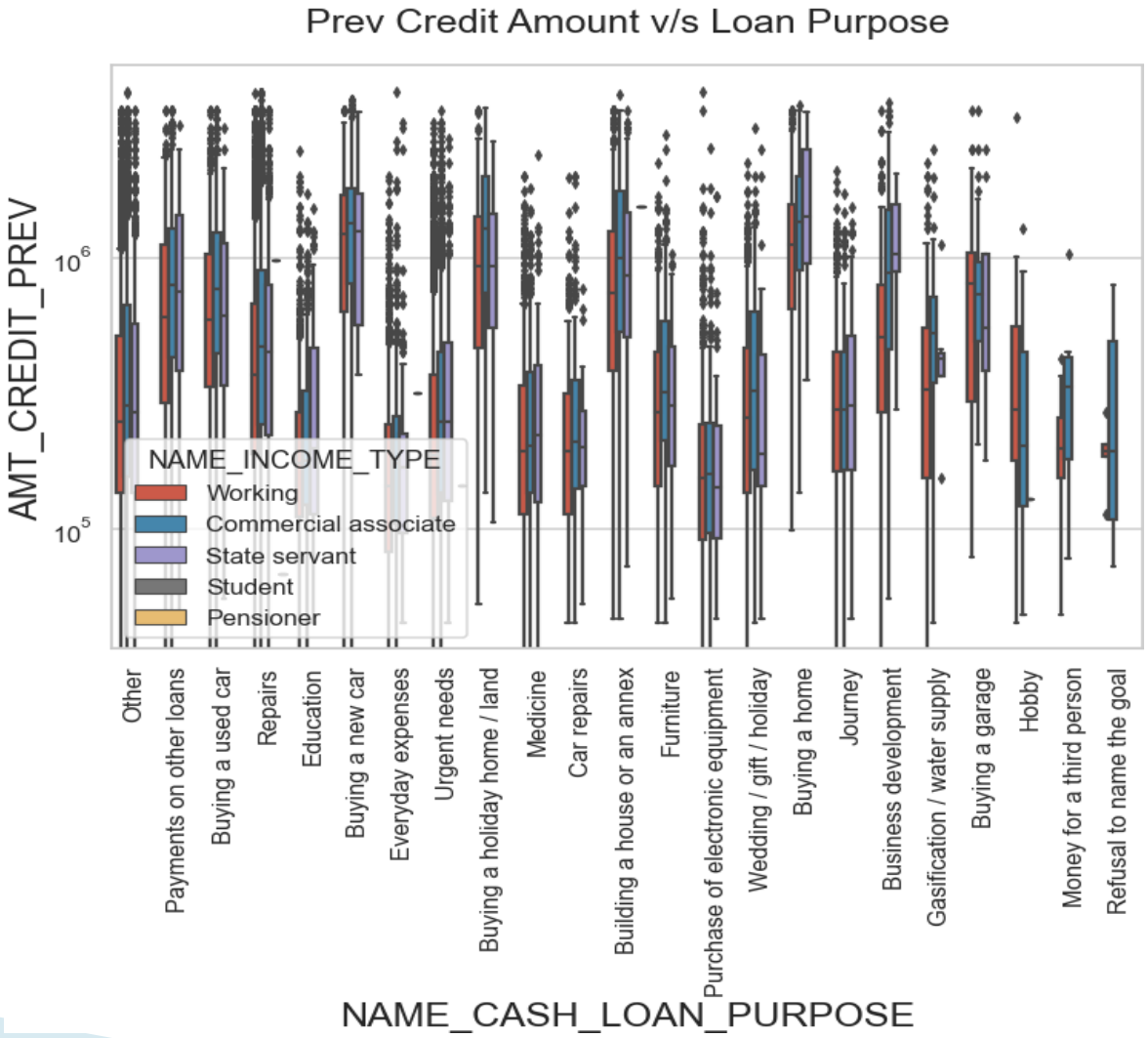
➢Loan purposes with 'Repairs' are facing more difficulties in payment on time. 'Buying a garage', 'Business Development', 'Buying Land', ''Buying a new car', 'Education', are the loan purposes which are having minimal difficulties in payment.



BIVARIATE ANALYSIS :

Findings from above Box Plot:

- 'Servent' income type category seems to have applied in significant amount.
- 'Money for a third person' has least credit applied for.
- 'Buying a land', 'Buying a home', 'Building a home', 'Buying a car', are the one having higher credits..

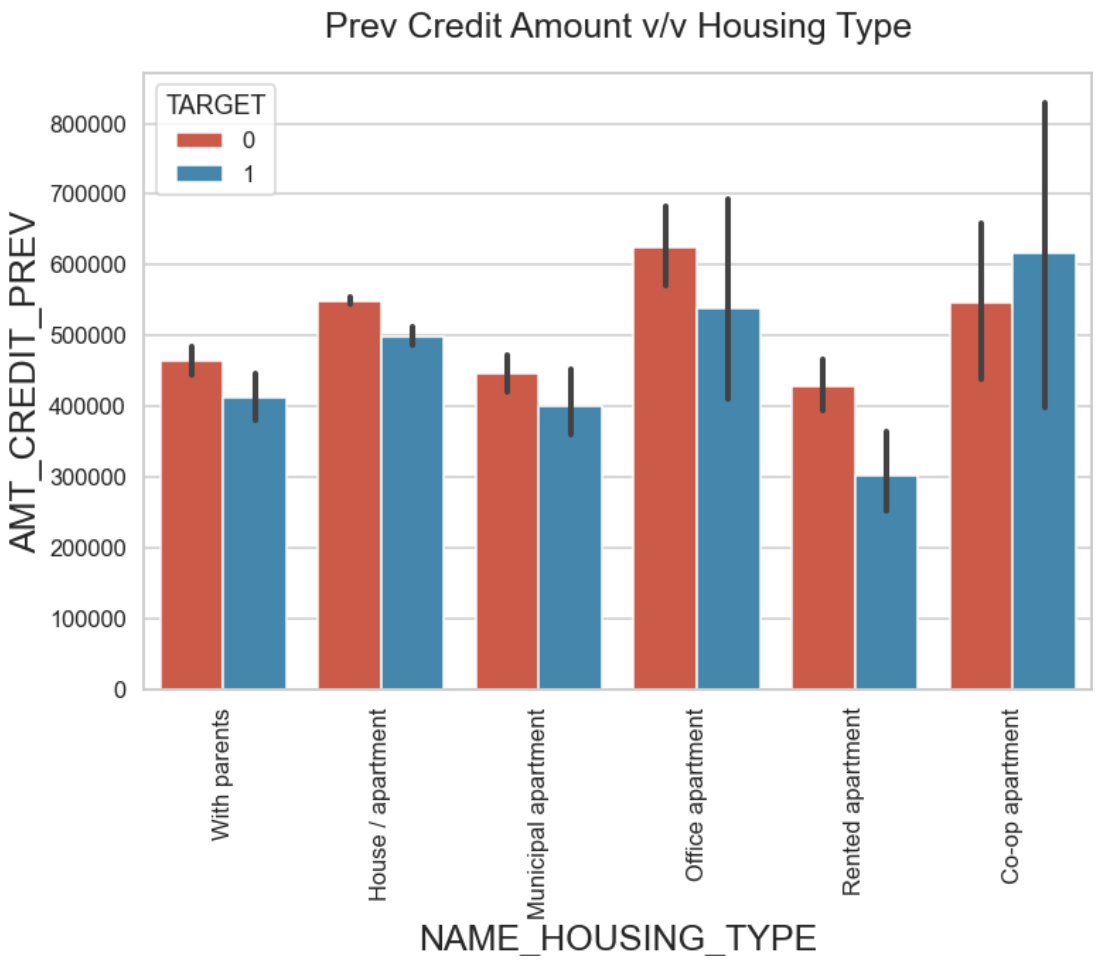


BIVARIATE ANALYSIS :

Finding from above plot:

➢'Co-op apartment' shows difficulties in repay the loan amount as the bar of Target 1 is high, so bank needs to be careful about this category.

➢'Office aprtment' shows good loan repayment response as bar of Target0 is high , so bank should focus more on this category.



Conclusion:

After going through various analysis we can observer following findings

- In Occupation category we can say , bank should focus on the "student", "Pensioner", "Businessman".
- Loan taken for repair are more liking to be a defaulter.
- Customer who live with parents are also likely to be a defaulter.
- "Working" type income category are also fall into high risk category and to be a defaulter.

THANK YOU!