CREDIT EDA CASE STUDY

ASSIGNMENT

BY : Pritha Bera

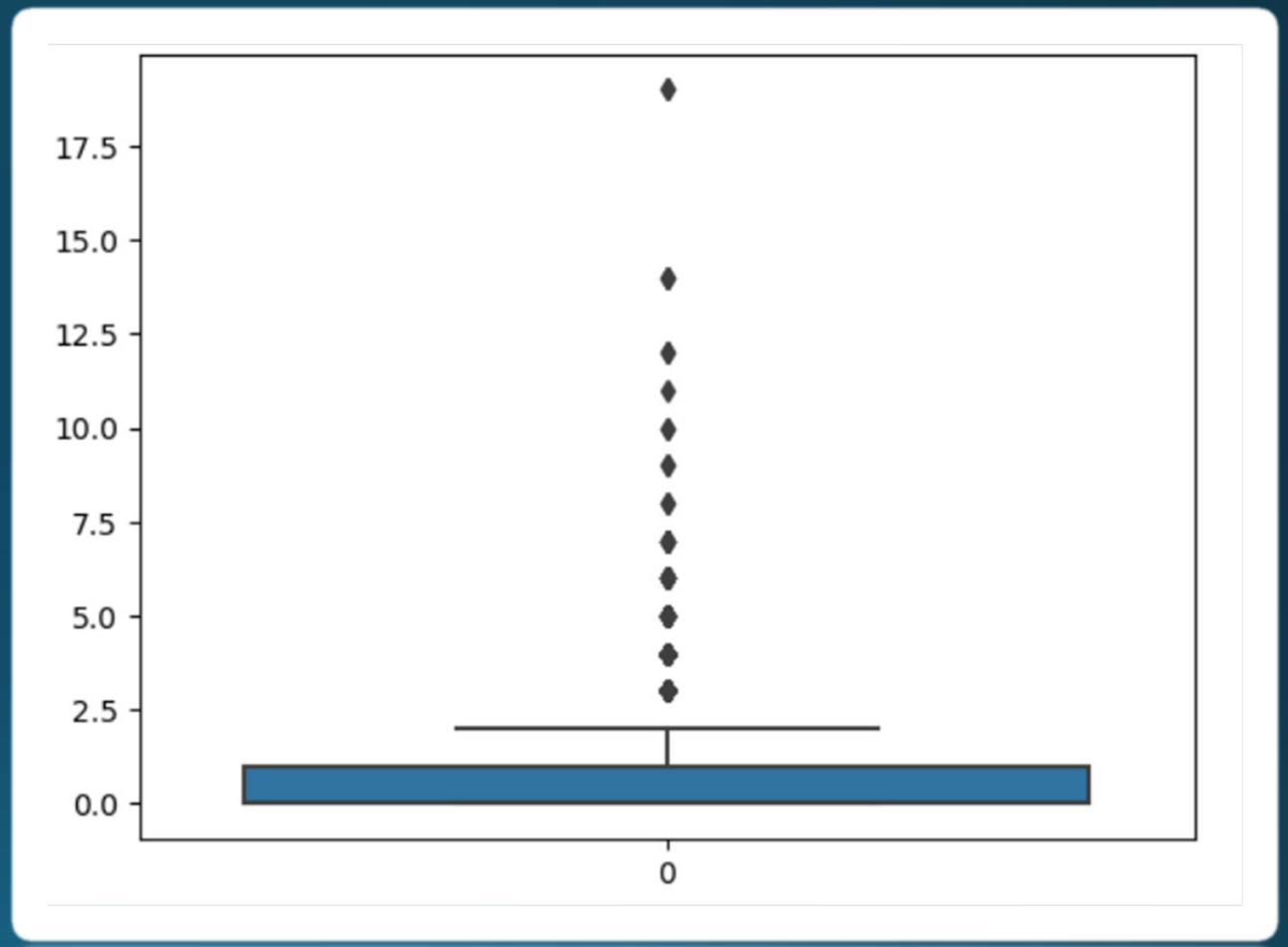
CREDIT EDA ASSIGNMENT TOPICS

- Importing Library's
- Columns Description Data
- Application Data
- Data cleaning
- Data Analysis
- Top correlations
- Bivariate Analysis
- Multivariate Analysis
- Merging Application Data and Previous Data

CATEGORICAL UNIVARIATE ANALYSIS,DROPPING THE ROWS BY FINDING OUTLIERS

Dropping the rows of very high no of children.

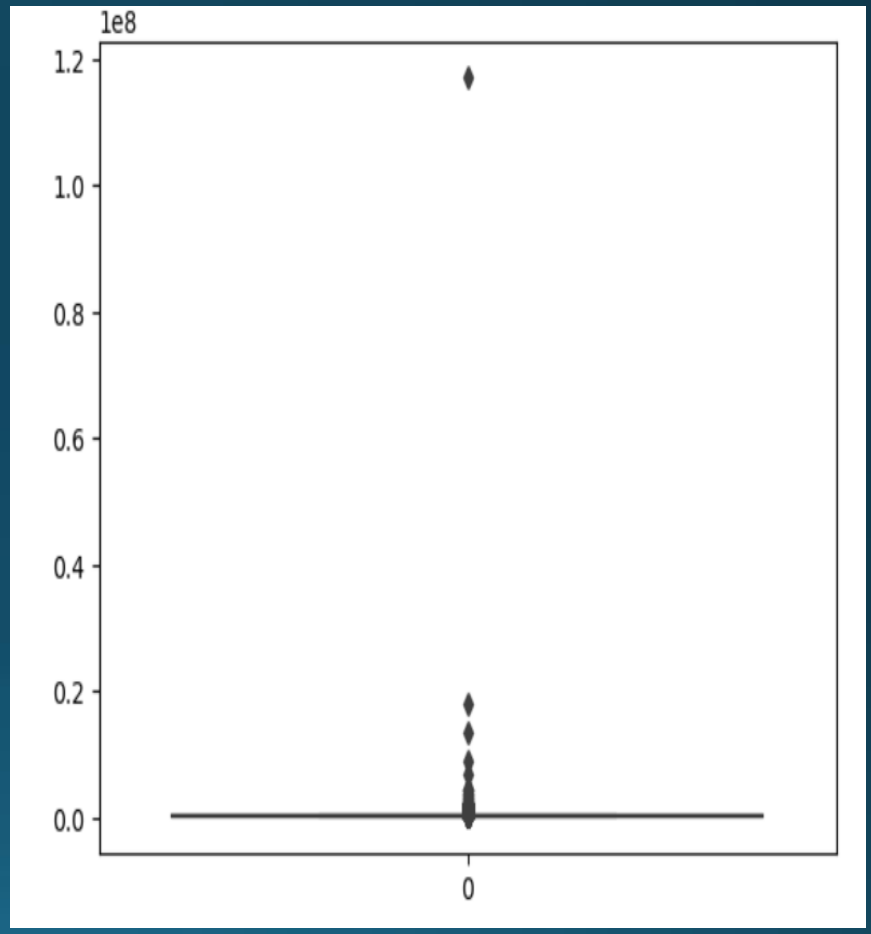
There are one outlier having children 19 which is not possible, so we can simply drop the row .



Dropping the rows of very high income.

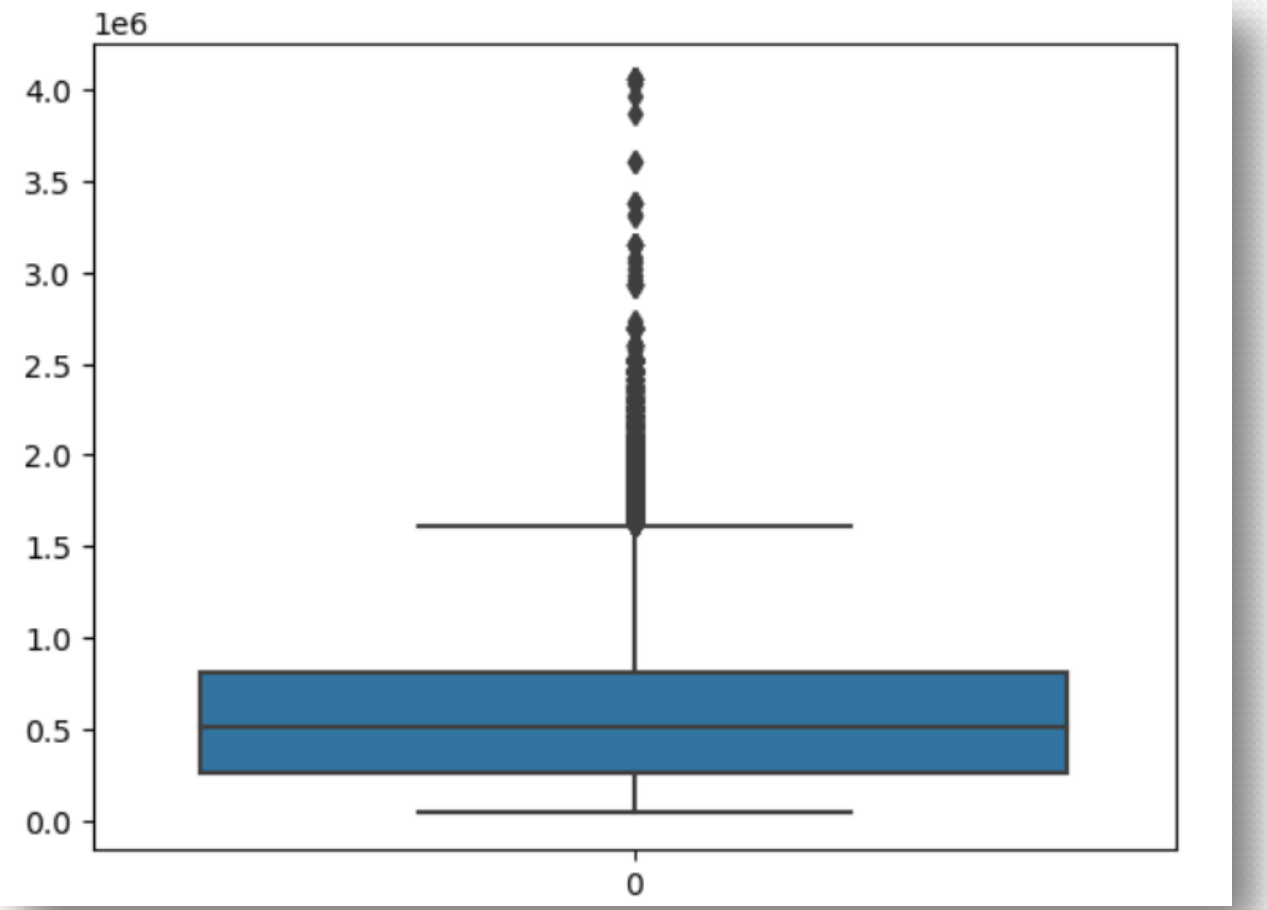
- Points to be considered from the Graph

- One outlier has value 117000000 which has no use, so we will drop the row.



Dropping the rows Of very high credit

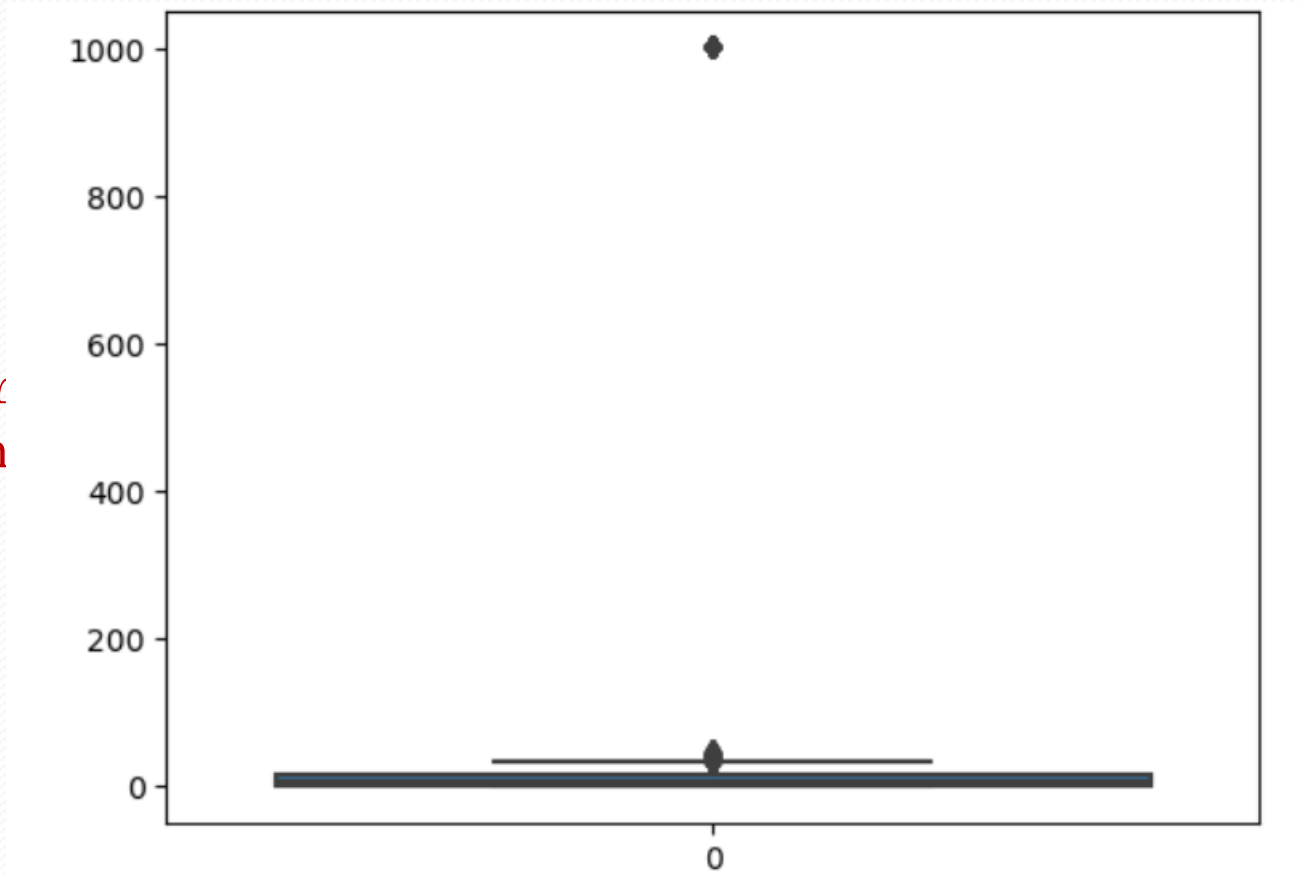
- There are so many outliers in the Amount_credit.
- In many rows the amount of credit is much higher than amount of income.
- We will drop the rows for our calculation.



Dropping the rows of With illogical year of Employment.

- There is an outlier of employment year Column which exceeds 1000 years which is never possible .

- So we will drop the row.



UNIVARIATE ANALYSIS FOR TARGET 0 AND TARGET 1.

- As we know from the instruction that Target =1 means the customers are defaulters and Target = 0 means the customers are paying their loans on time.

- We will assign two variables , def_r for defaulters i.e. Target=1 and non_def for non defaulters i.e. Target =0.

- We will do our rest of univariate analysis of both Target 0 and Target 1 category.

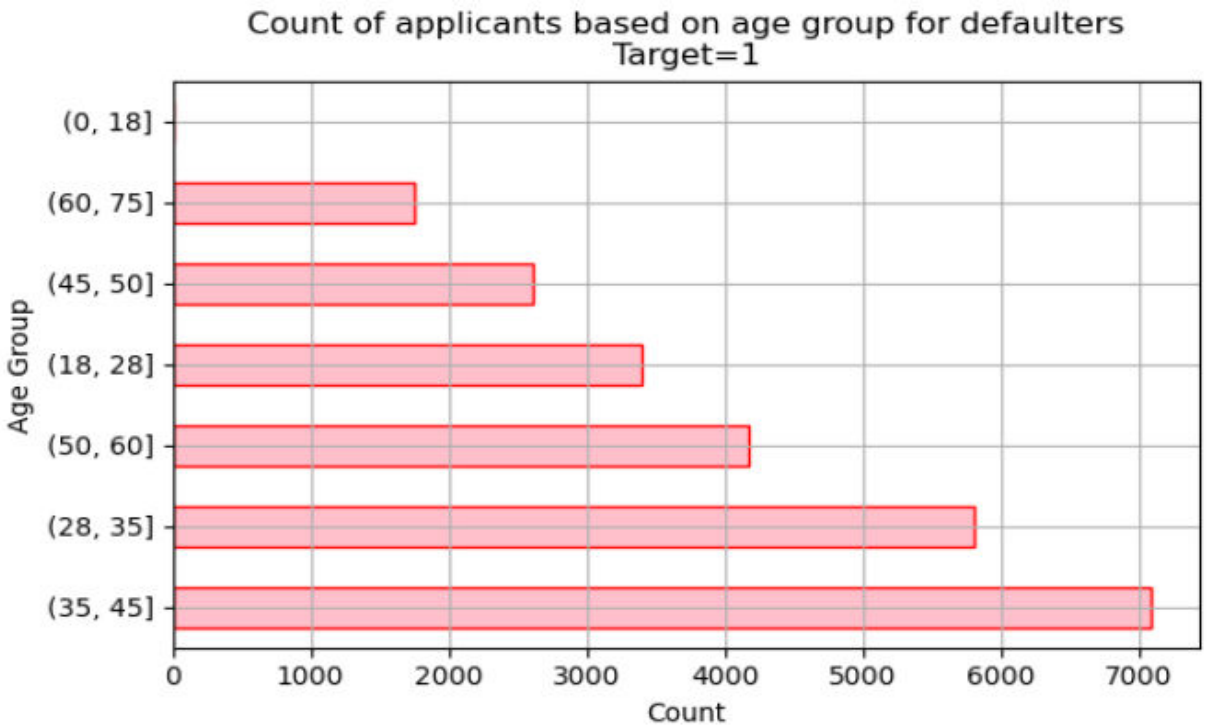
Target variable : Gender

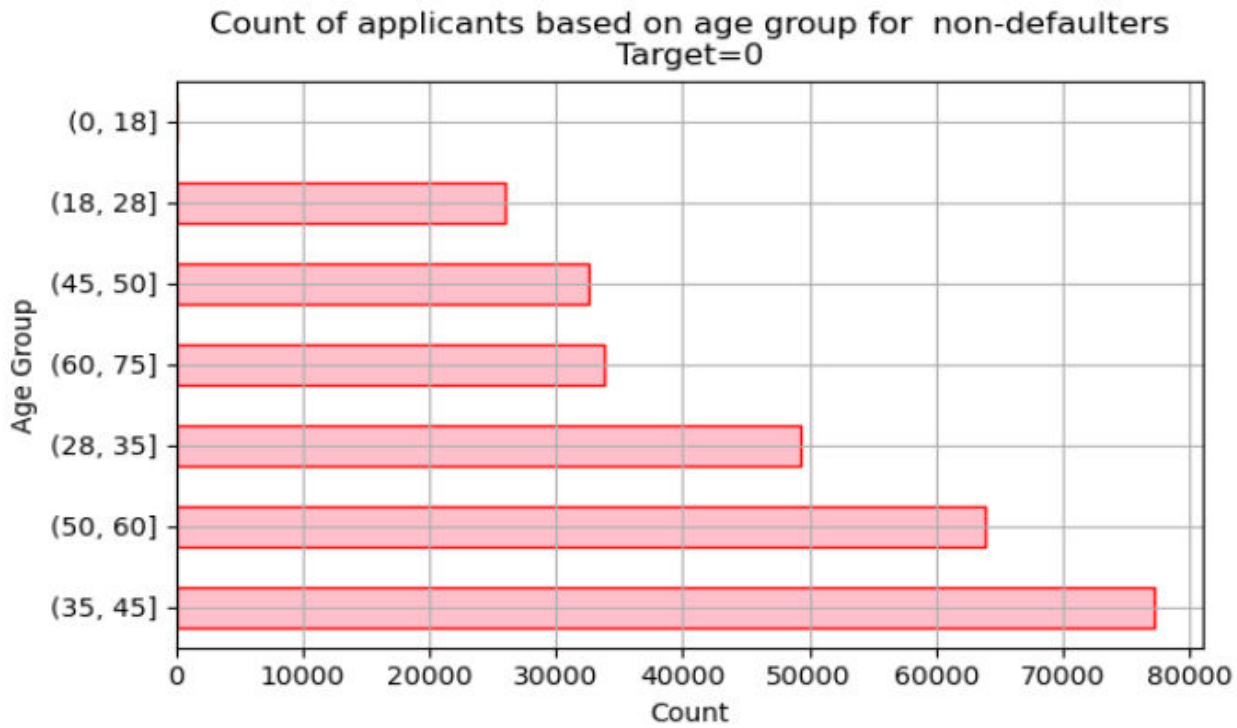
- Close to 60% applicants are female in defaulters and close to 70% applicants are female in non-defaulters.

- It seems that there are more chances of women applicants are being defaulters and not defaulters both.

Target Variable : Age Group

- Most of the applicants are falling under age group of 35-45 years in both defaulters and non-defaulters category.
- There are no applicant less than 18 years of age in both the category.
- Non defaulter category: 2nd highest are in 50-60 age group and 3rd highest in 28-35 age group
- Defaulter category : 2nd highest are in 28-35 age group and 3rd highest in 50-60 age group

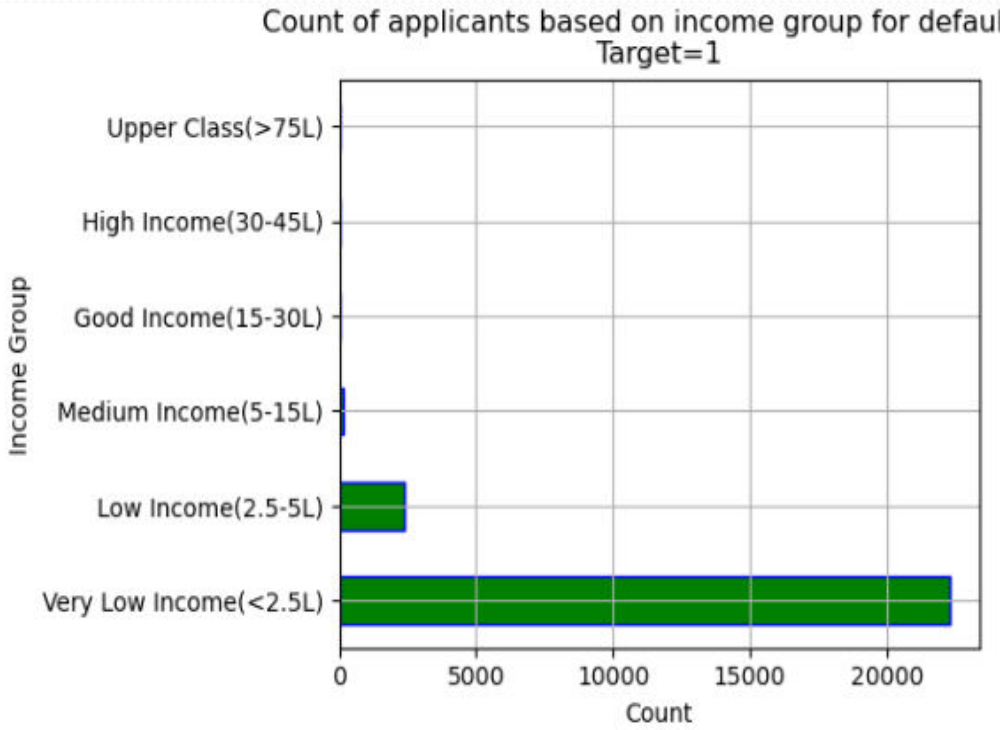


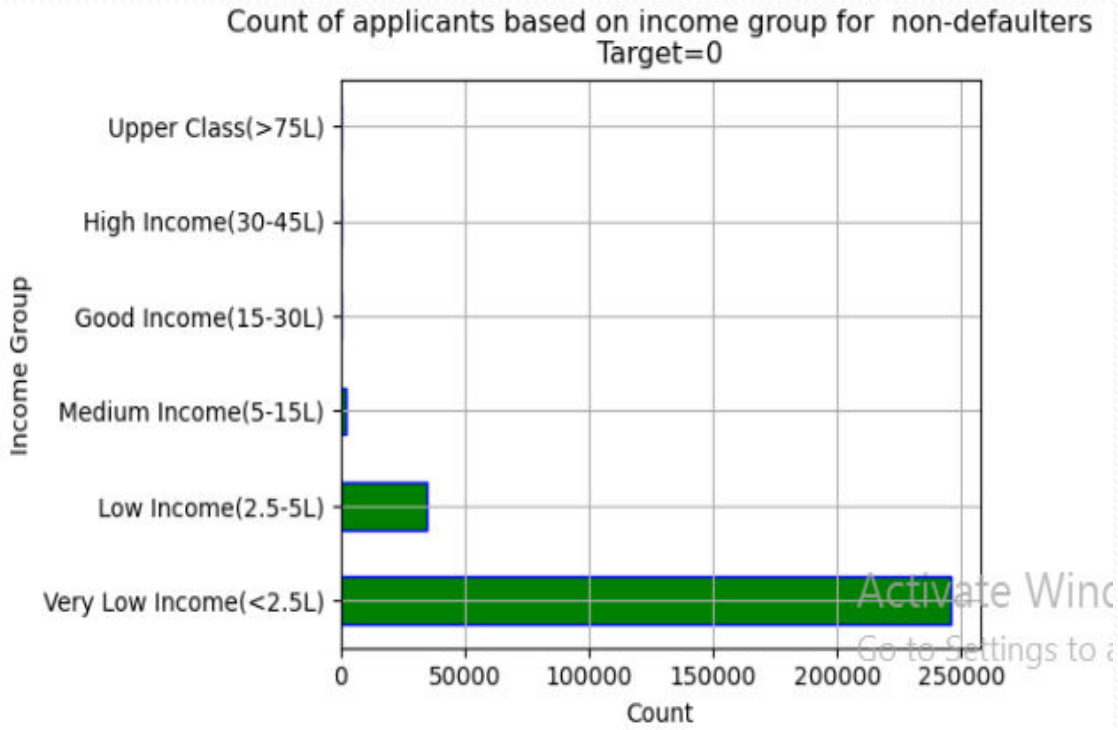


Target Variable : Income Group

❖Most of the applicants are from very low income group i.e. less than 2.5L in both the category.

❖2nd highest is from Low income group and 3rd highest is from Medium income group and there are very few applicants having income more than 15L

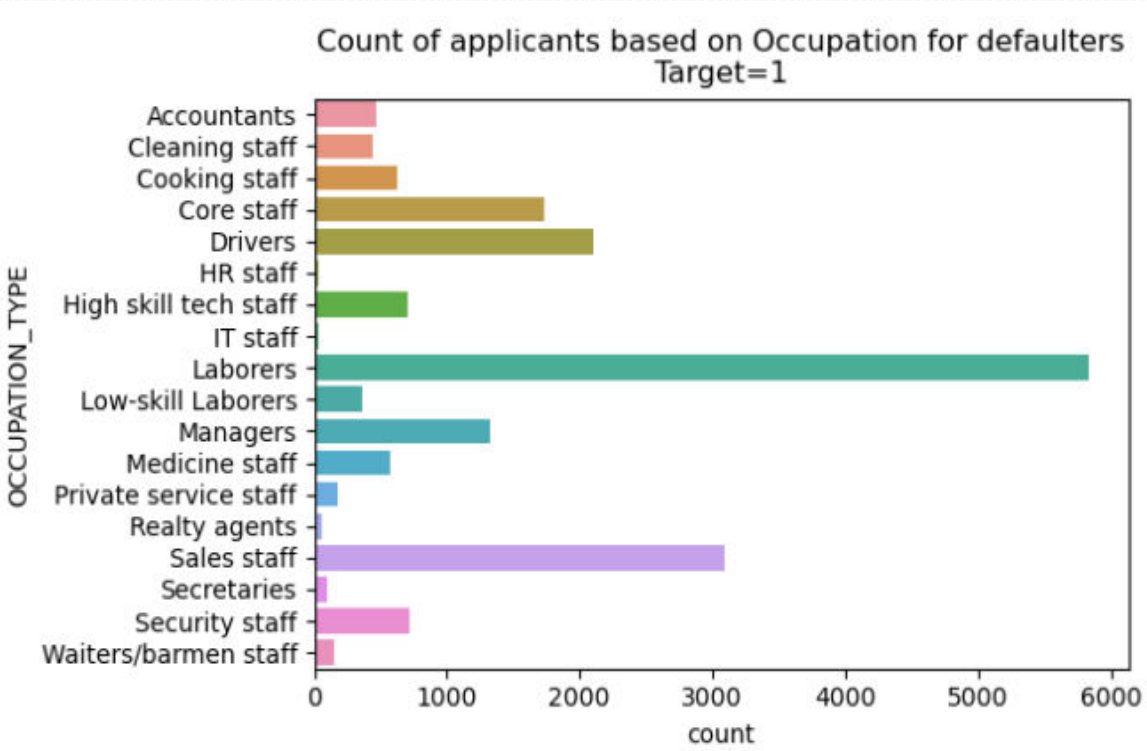


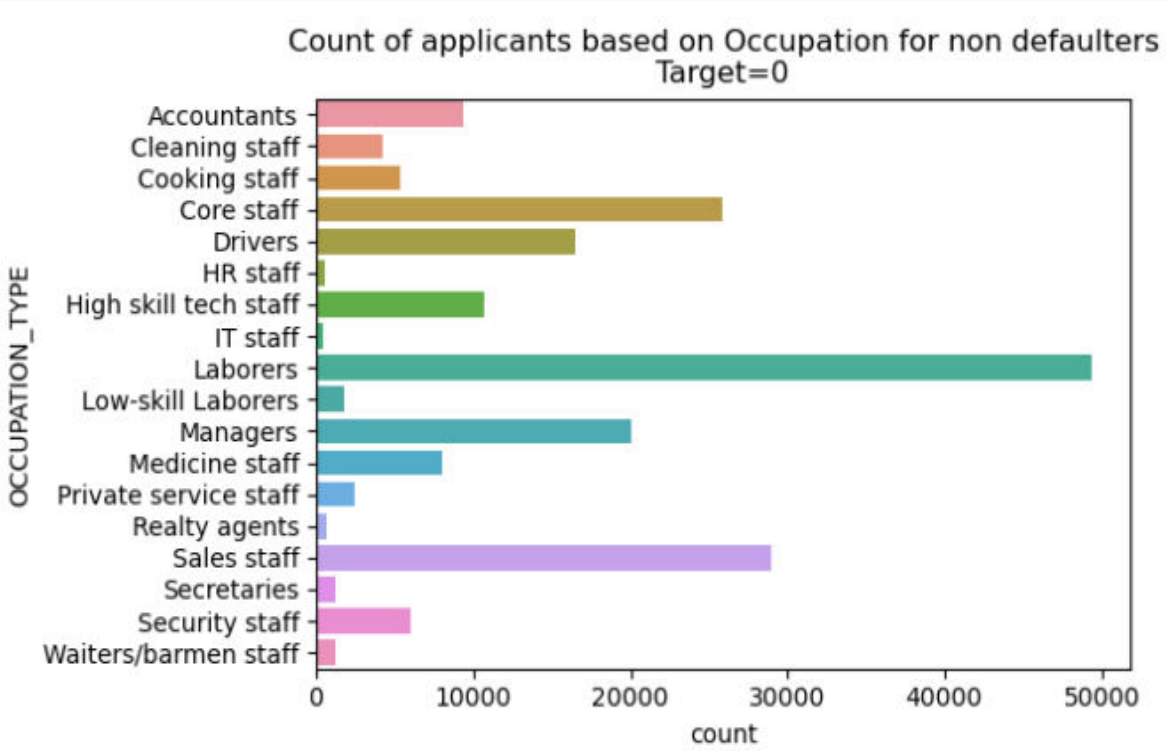


Target Variable : Occupation Type

- Most of the applicants are belong to labourers as occupation , 2nd highest occupation is Sales staff.

- The number of defaulters in every occupation is comparatively less than number of non defaulters.

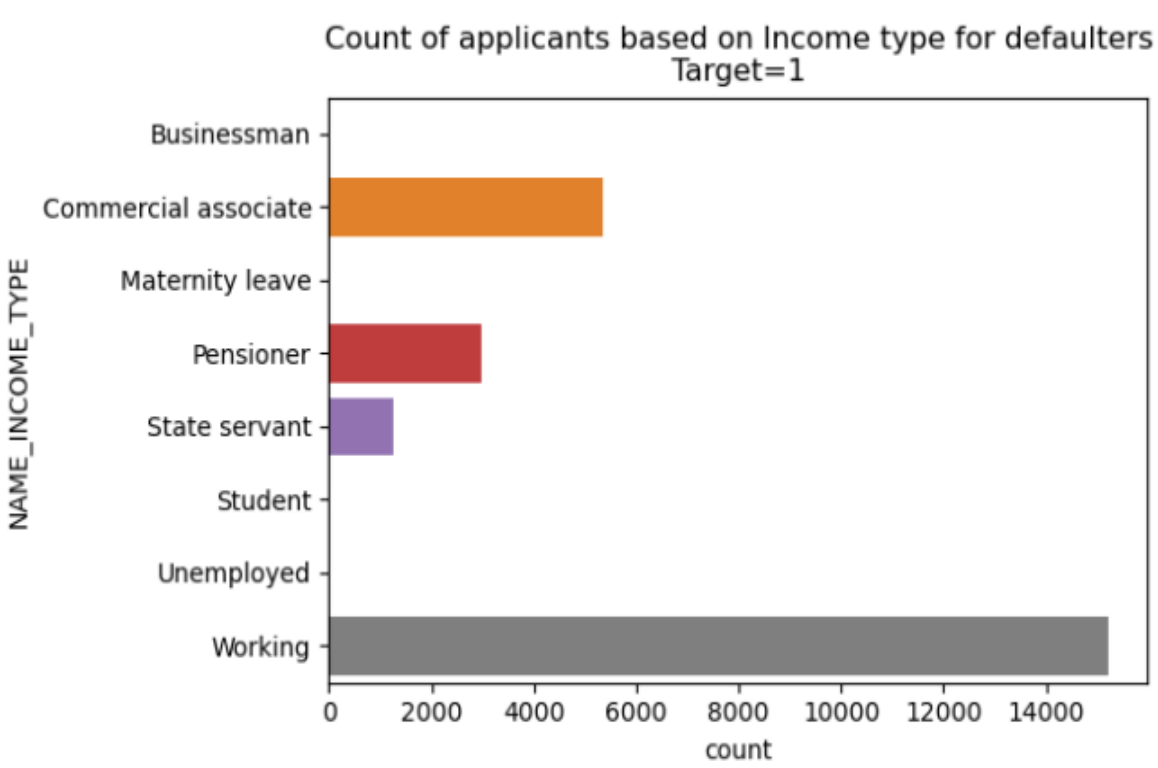


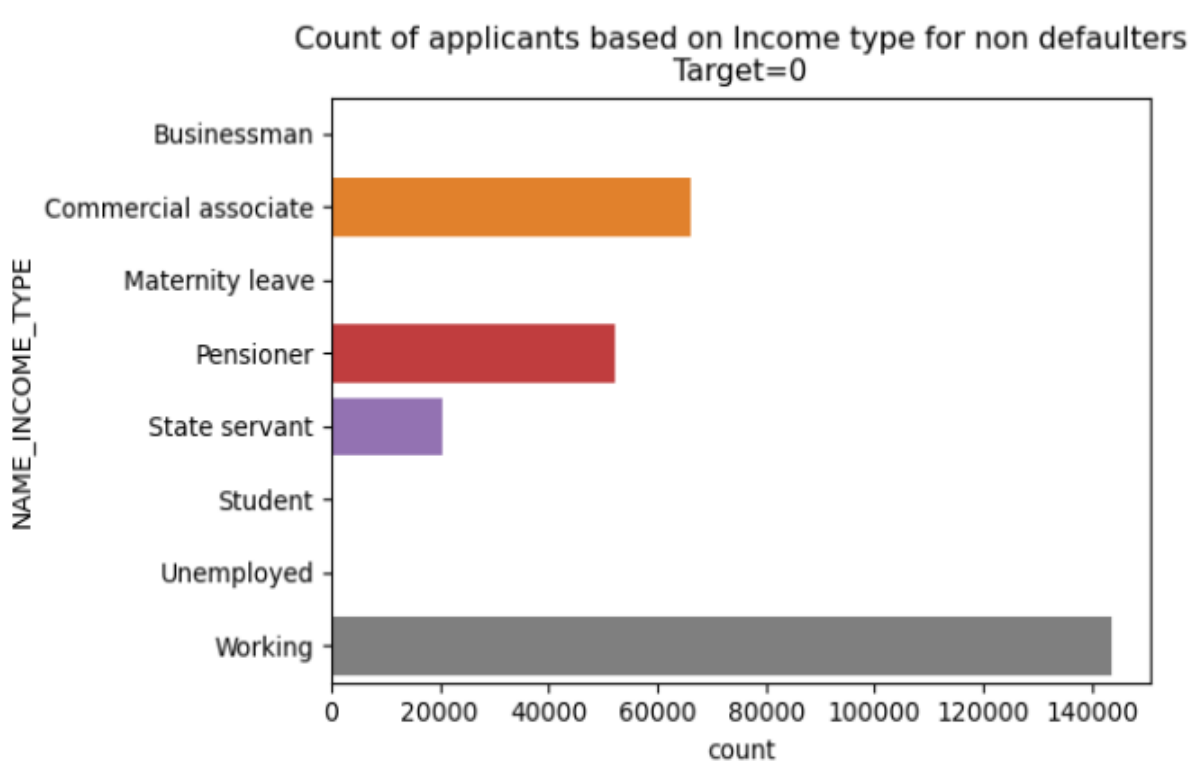


Target Variable : Name Income Type

- Most of the loans are distributed to working class people , Pensioners are also good in number for applying loans and mostly they are non -defaulters.

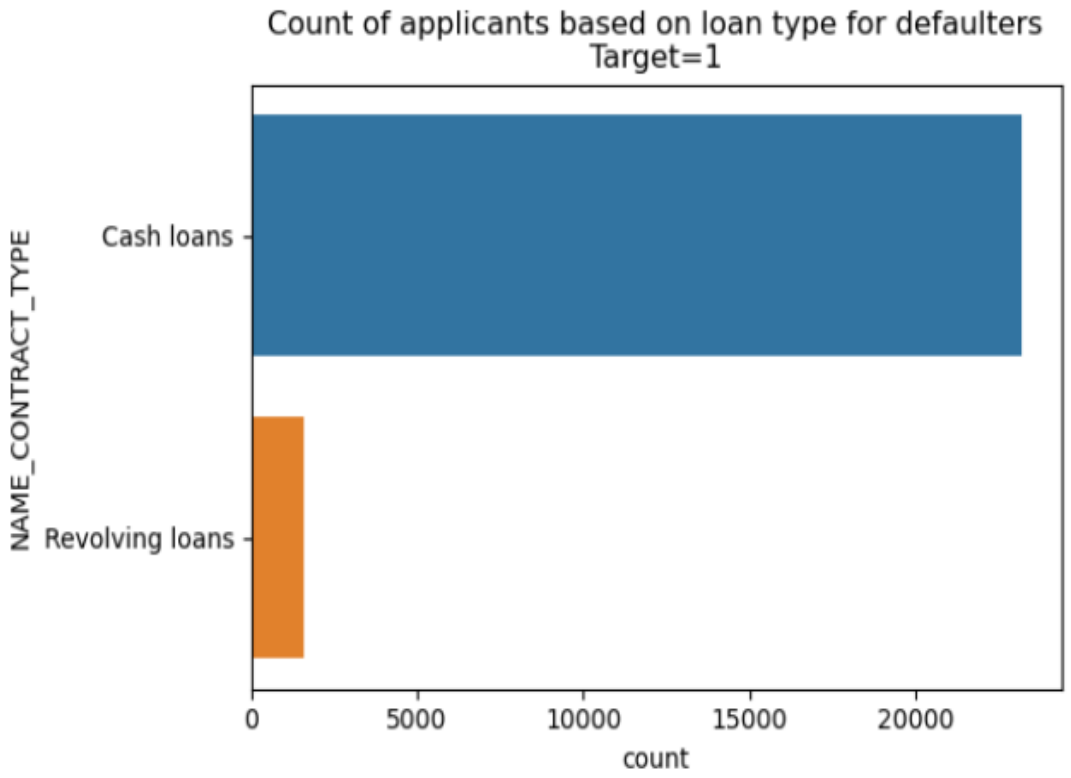
- Businessman, students, unemployed are hardly took a loan.

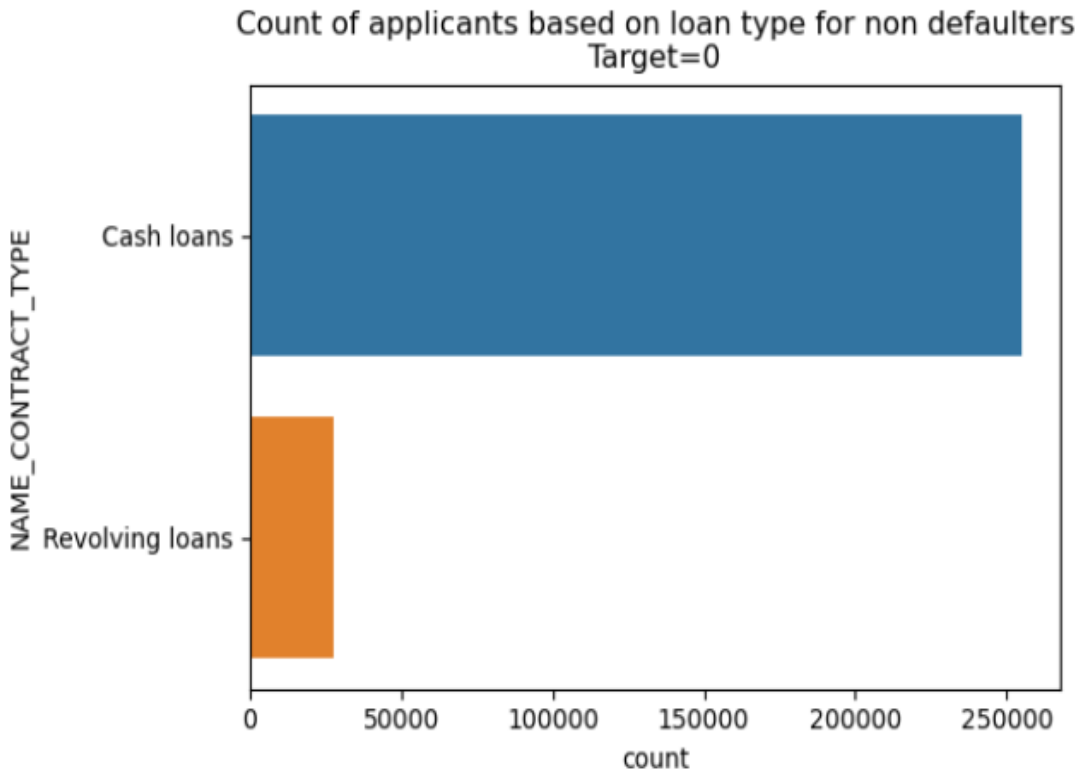




Target Variable : Name Contract Type

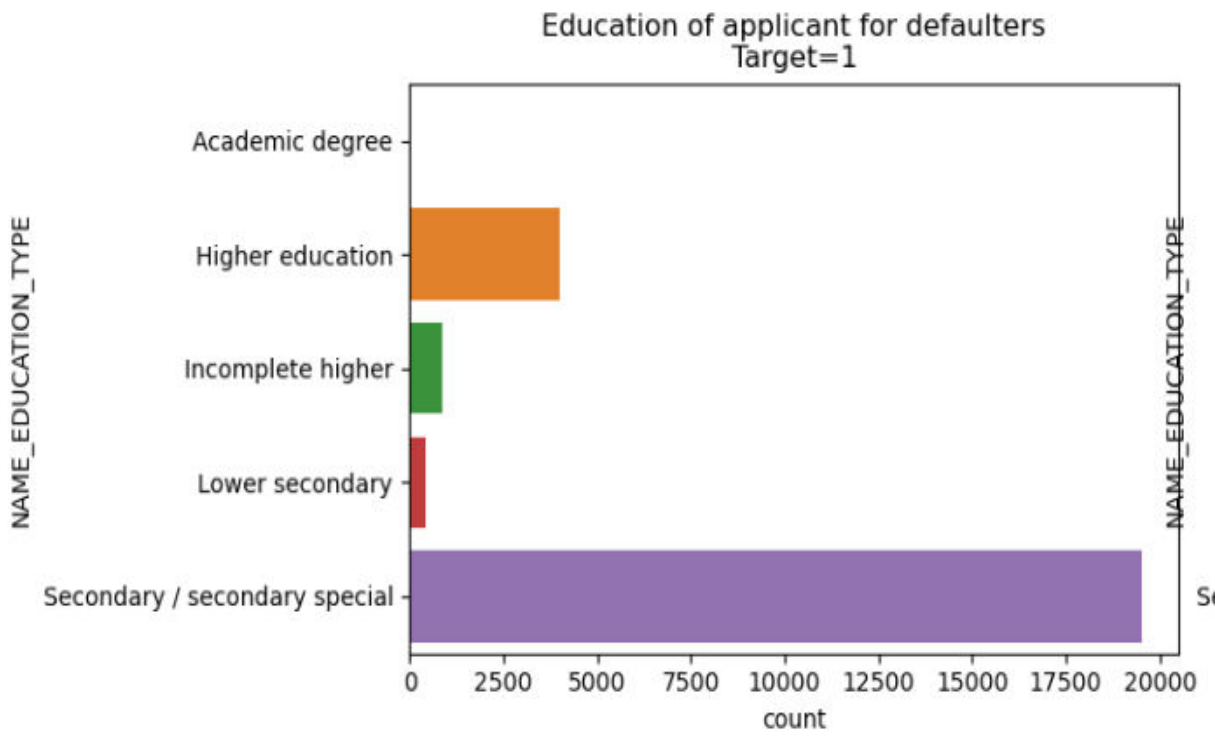
- **Most of the loans types are cash loans in both category .**

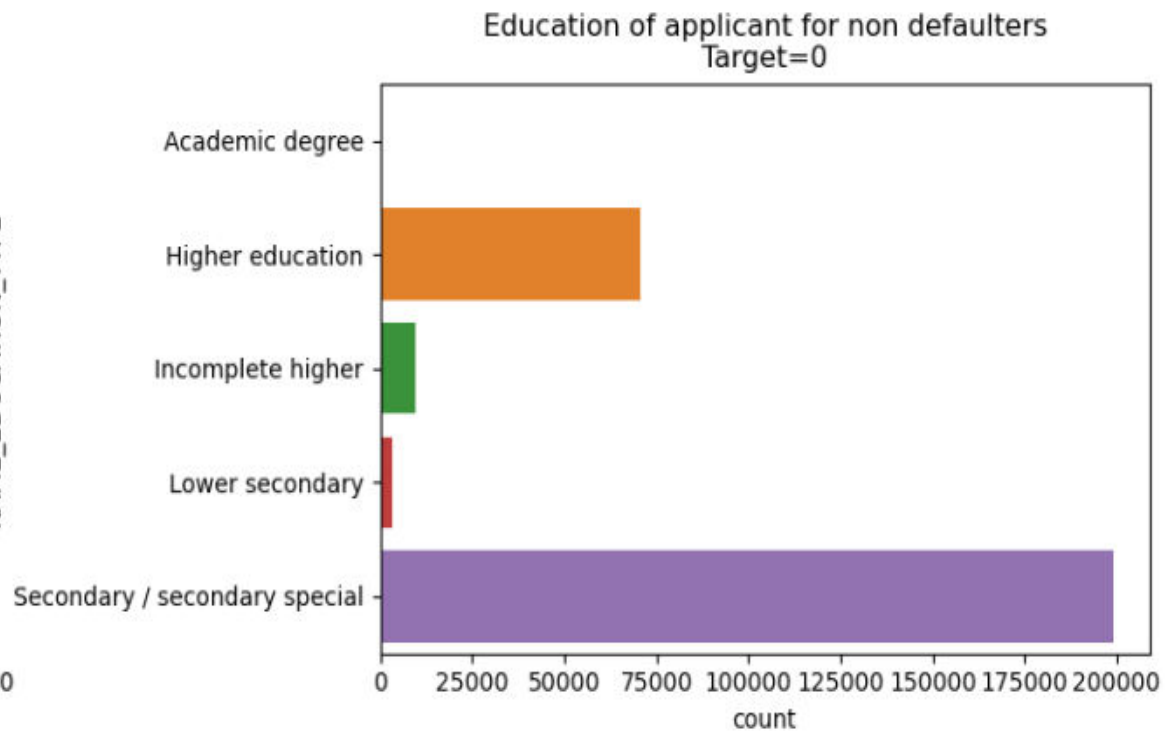




Target Variable : Name Education Type

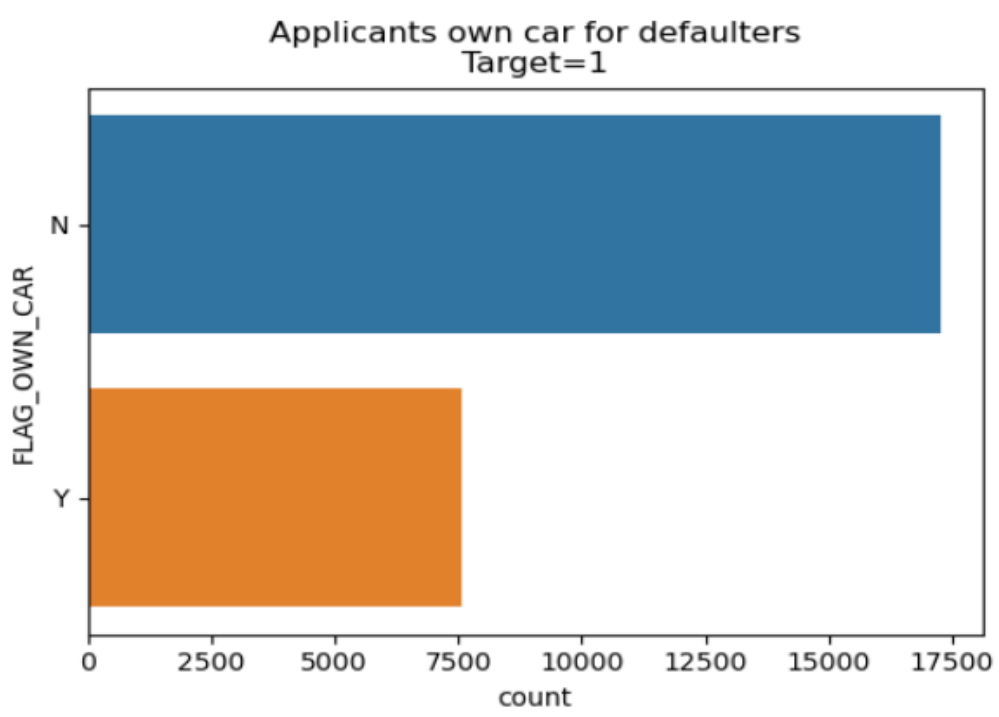
❖Most applicants have completed secondary education in both category, secondly many of them have completed higher education and lastly academic degree holders are almost negligible in both category.

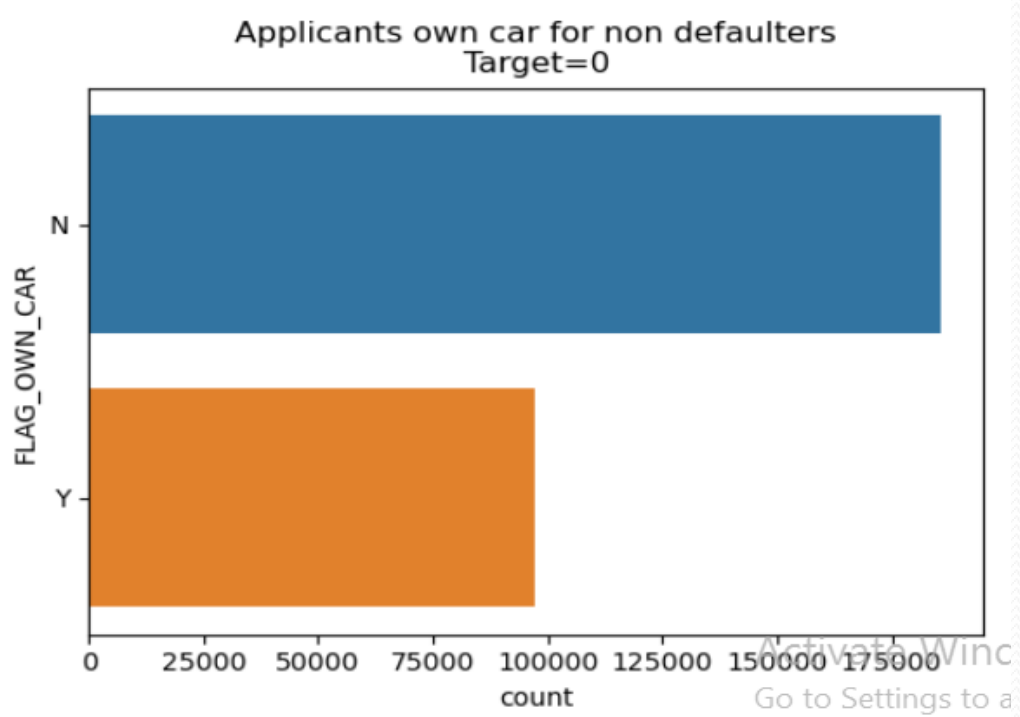




Target Variable : Flag own car

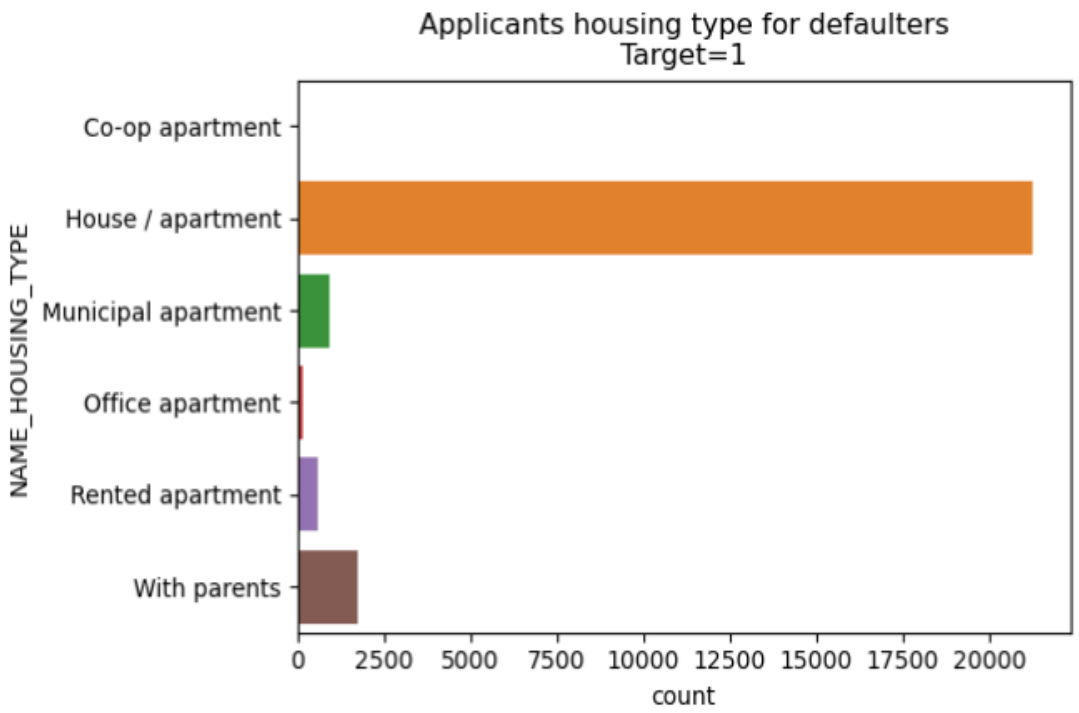
- Most of the applicants don't have a car, we can also say that the number of default of people having car is low compared to people who don't default .

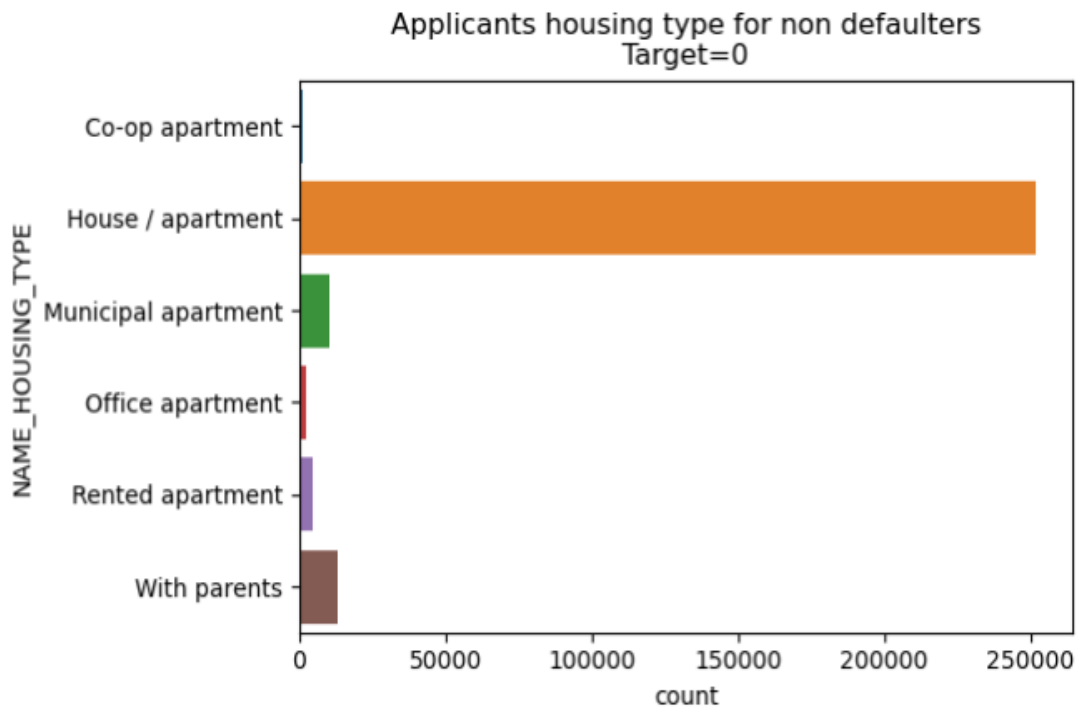




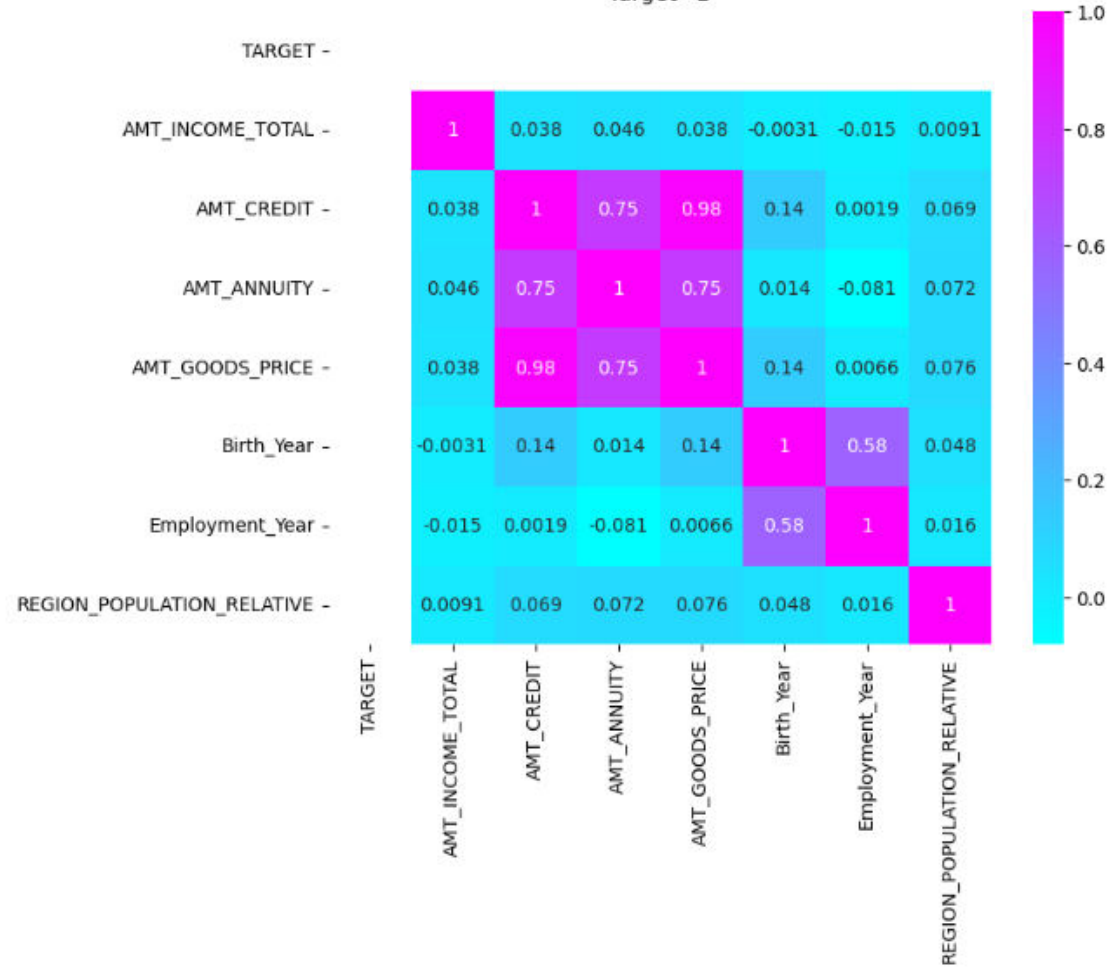
Target Variable : Name housing type

- No of non default applicants who own a house is more than no of default applicant who own a house.





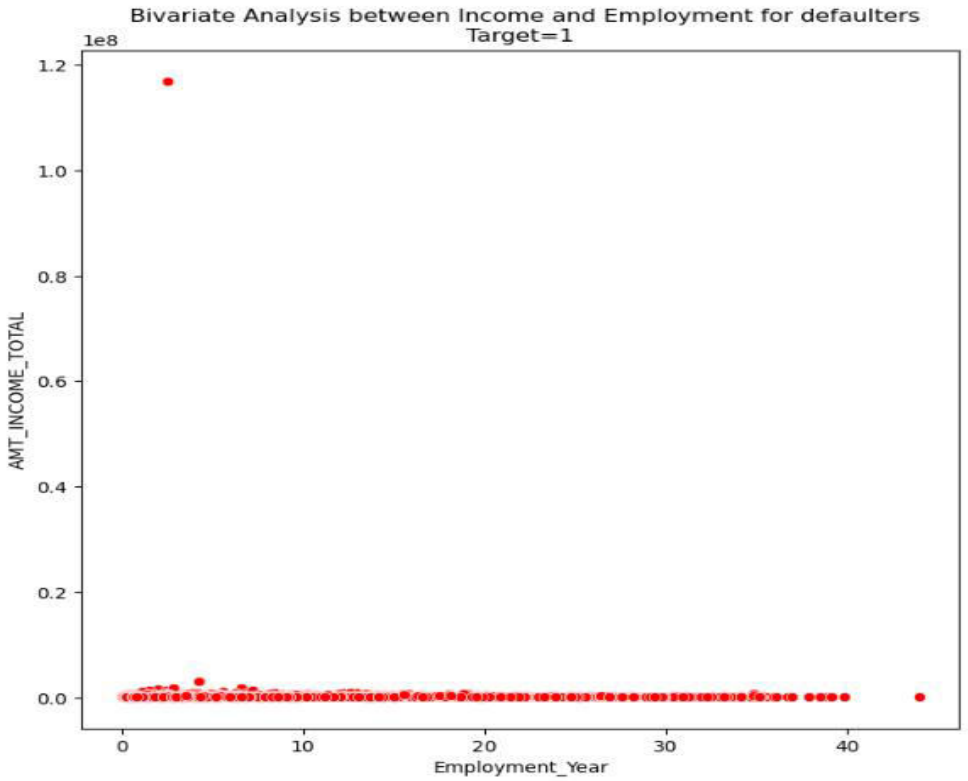
Bivariate analysis in both Target 1 and Target 0 category using Heatmap, Scatterplot and Boxplot .

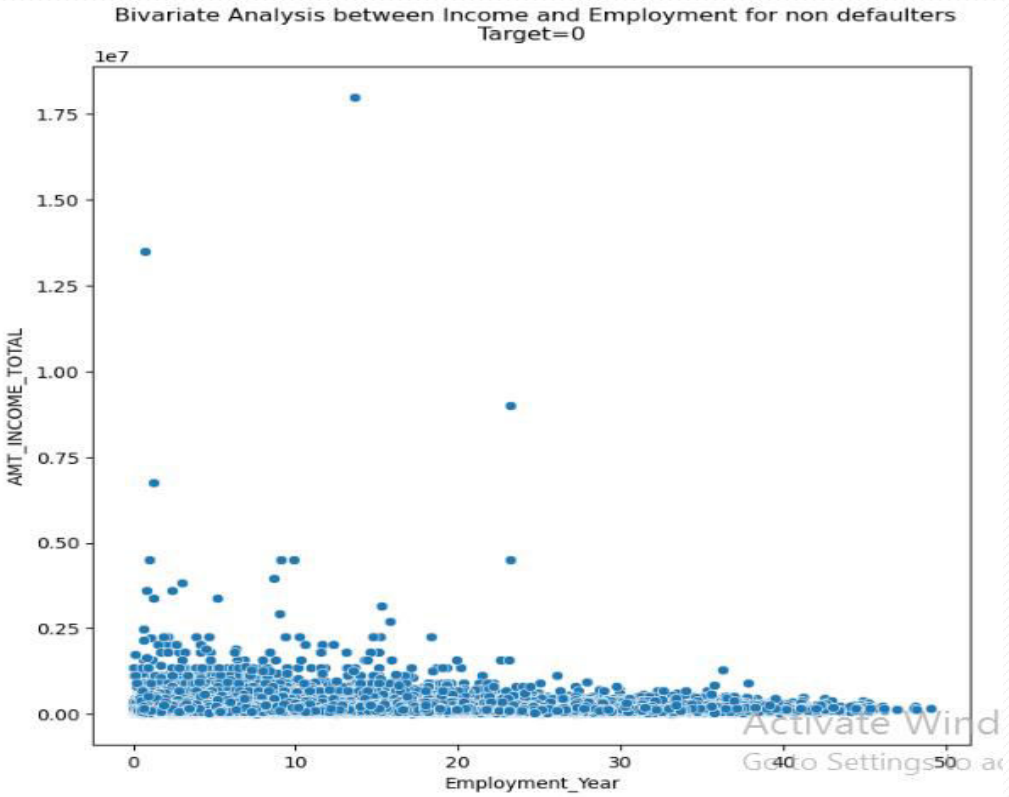


Through the correlation , using heatmap, we can conclude that correlations are almost same while seeing targt 0 & 1 .

Target Variable : Amt_Income_total & Employment_Year

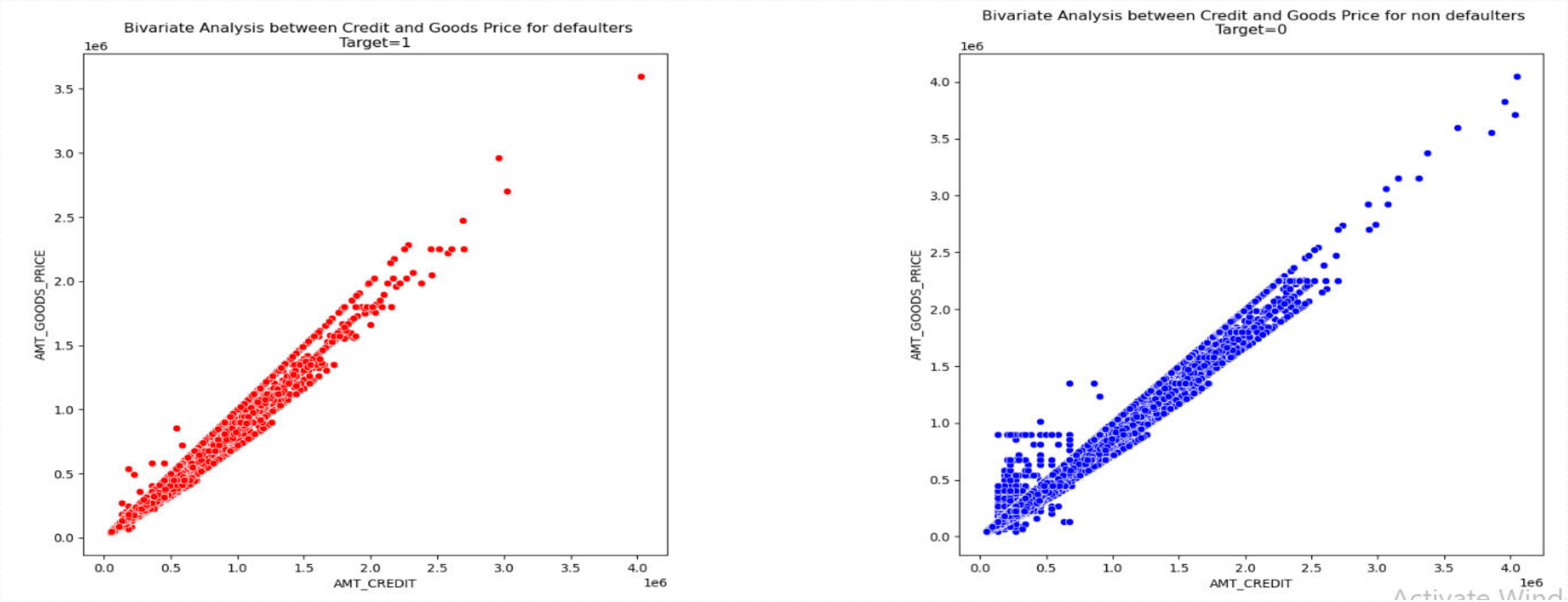
- It is quite evident from the above plots that most of the defaulters have very low income regardless of their duration of employment in years.
- However, we can see that non defaulters have comparatively higher income than defaulters and it can be seen that total income is gradually reducing with increase in years of employment.





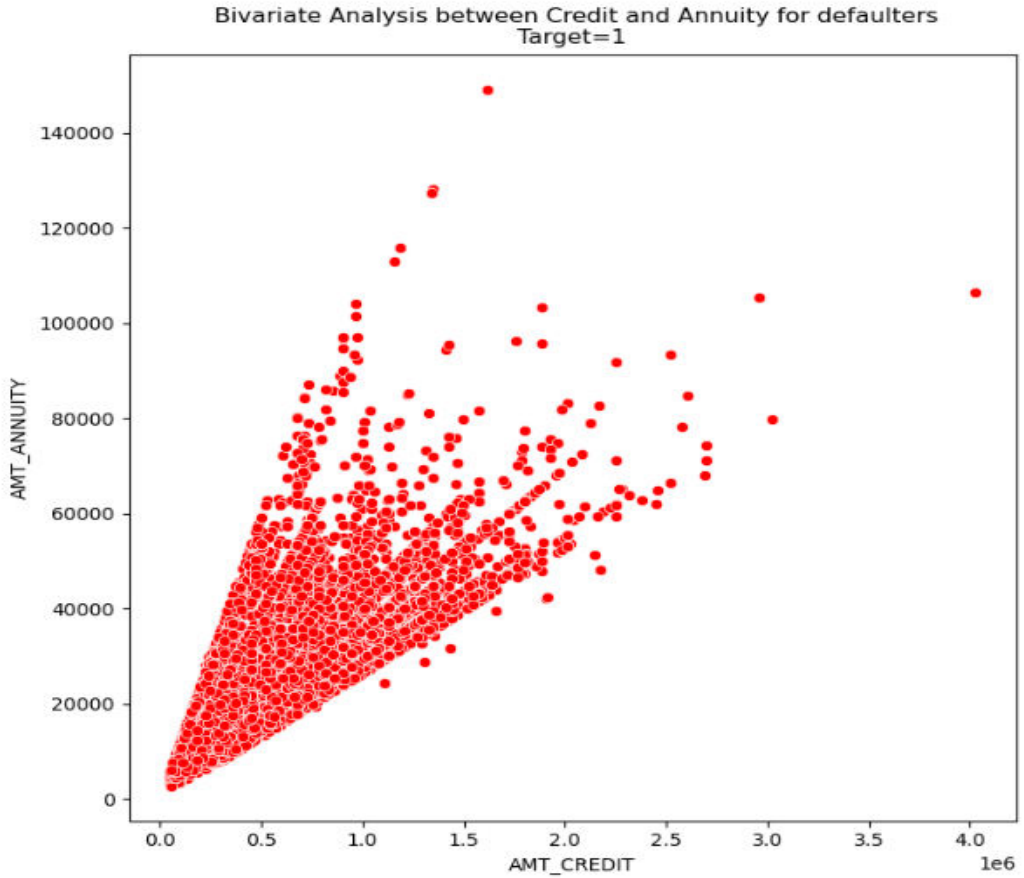
Target Variable : AMT_CREDIT and AMT_GOODS_PRICE

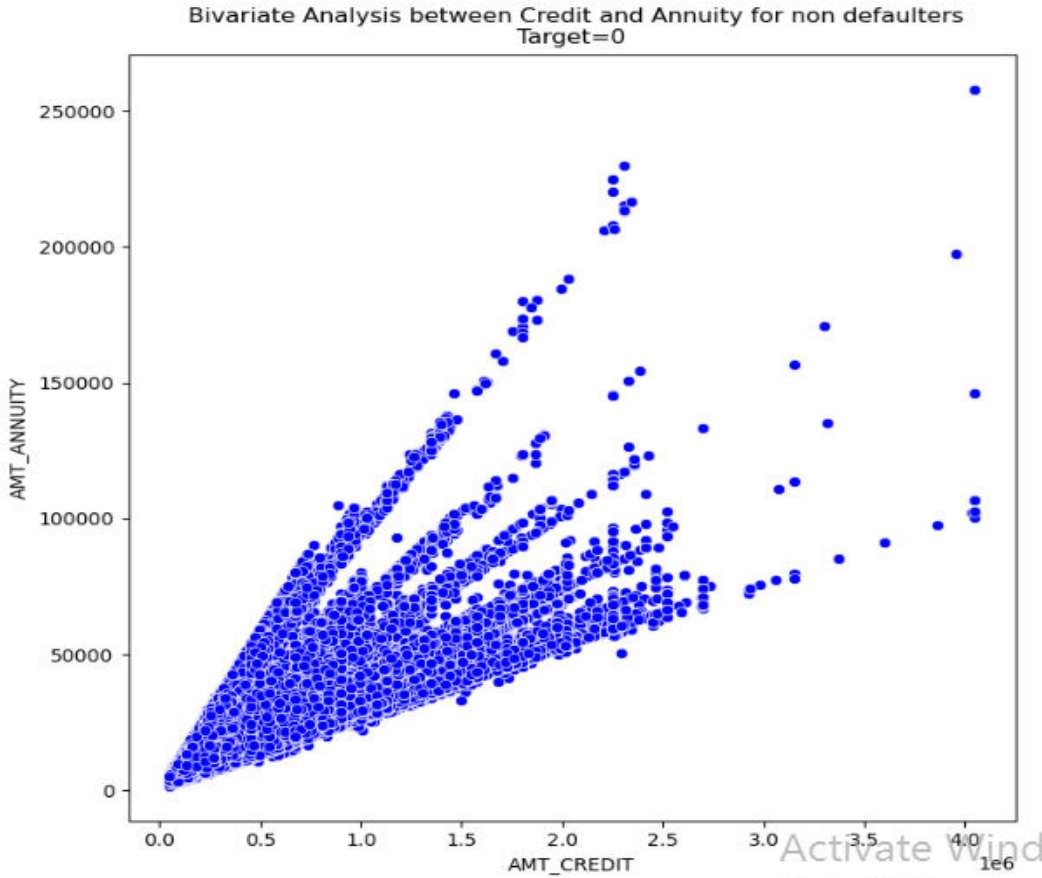
Amount Credit and Amount of goods price are showing same trend in both the cases .



Target variable : AMT_CREDIT and AMT_ANNUITY

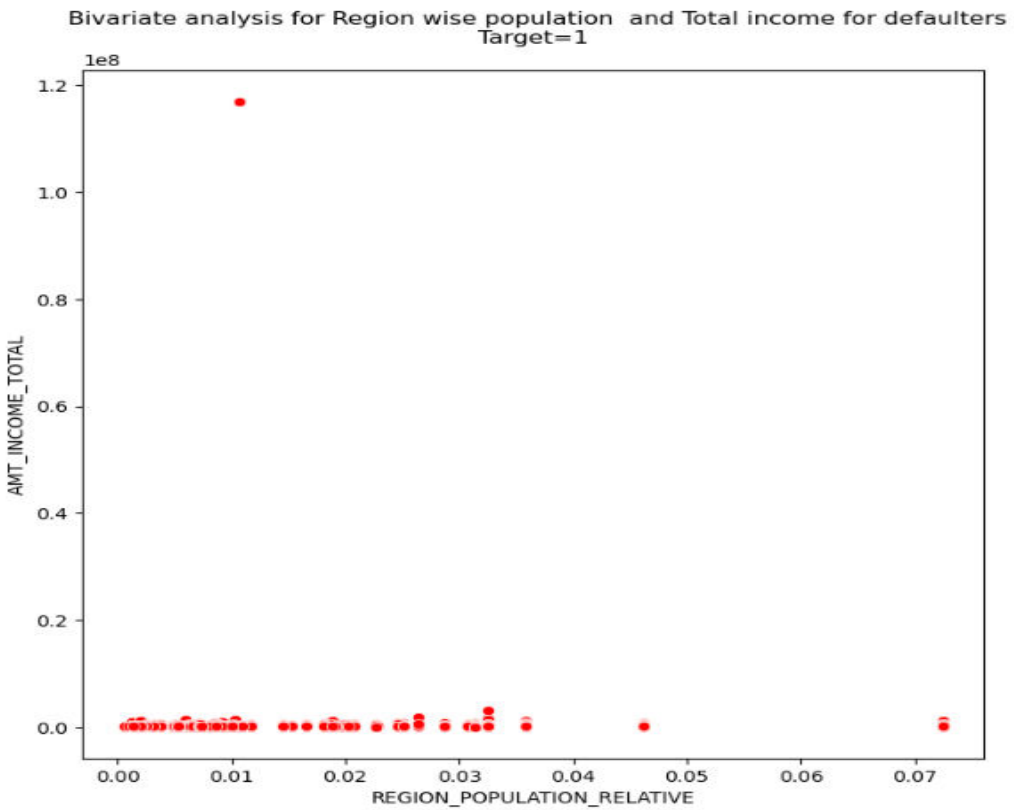
Amount Credit and Amount Annuity are showing same trend .

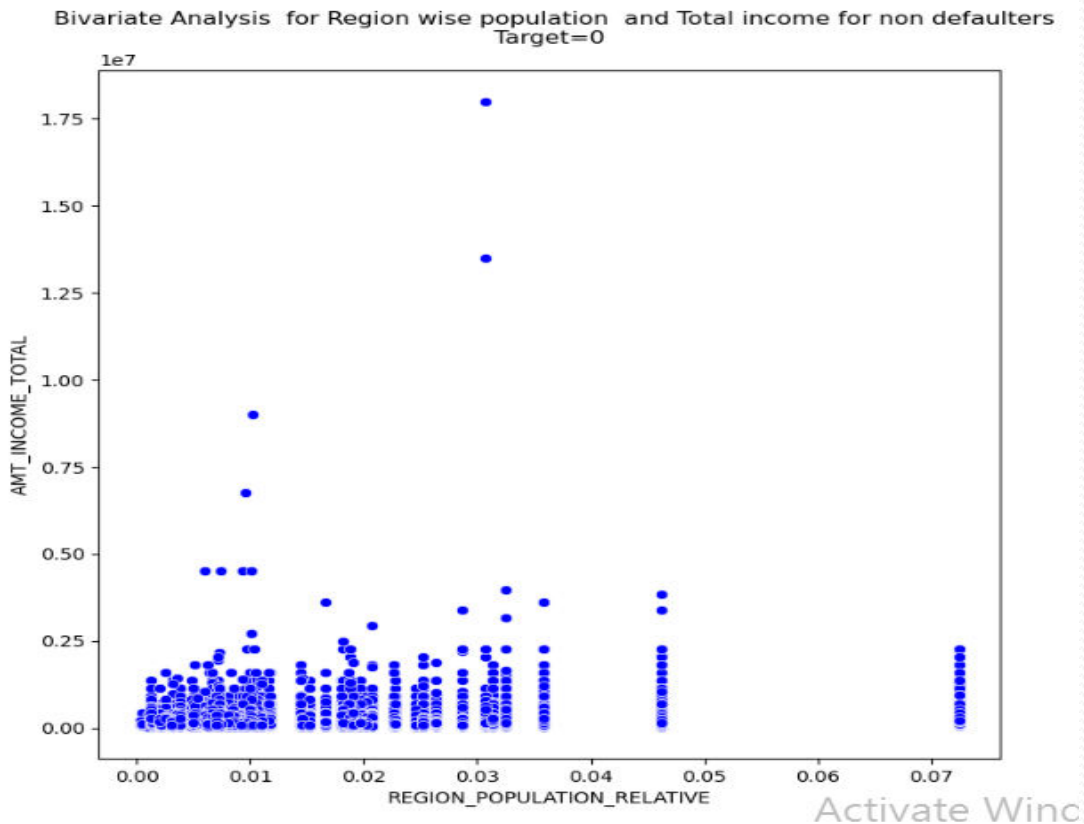




Target variable : REGION_POPULATION_RELATIVE & AMT_INCOME_TOTAL

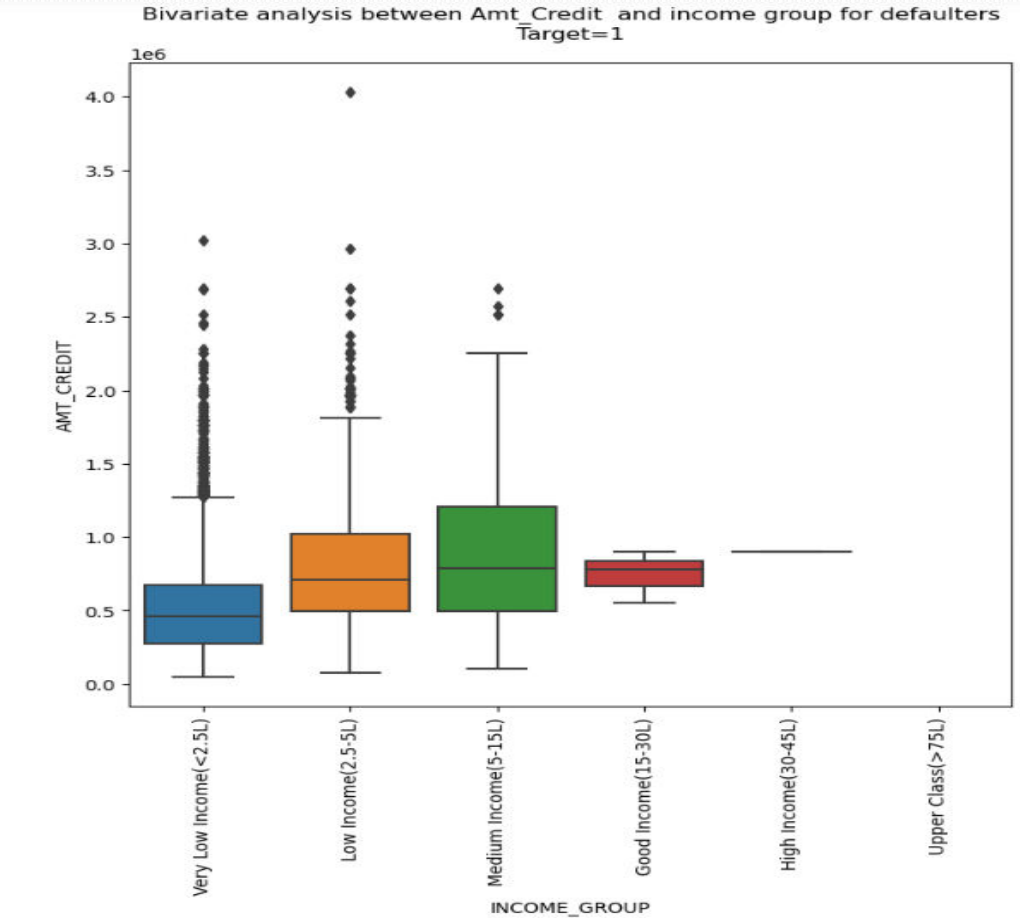
- It can be seen from the above plot that most of the defaulters have very low income where region population is less dense. We can also see that non defaulters have comparatively higher incomes than defaulters in the same polpulation regions.

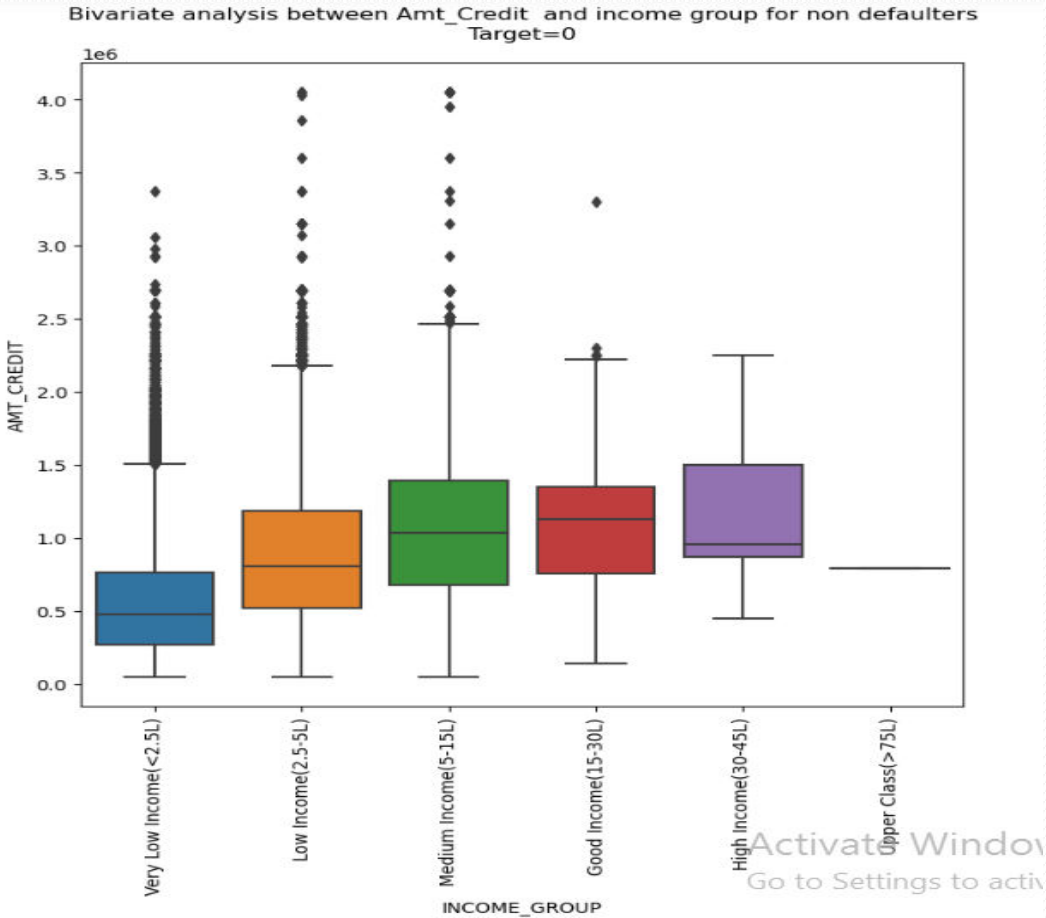




Target variable : AMT_CREDIT and INCOME_GROUP

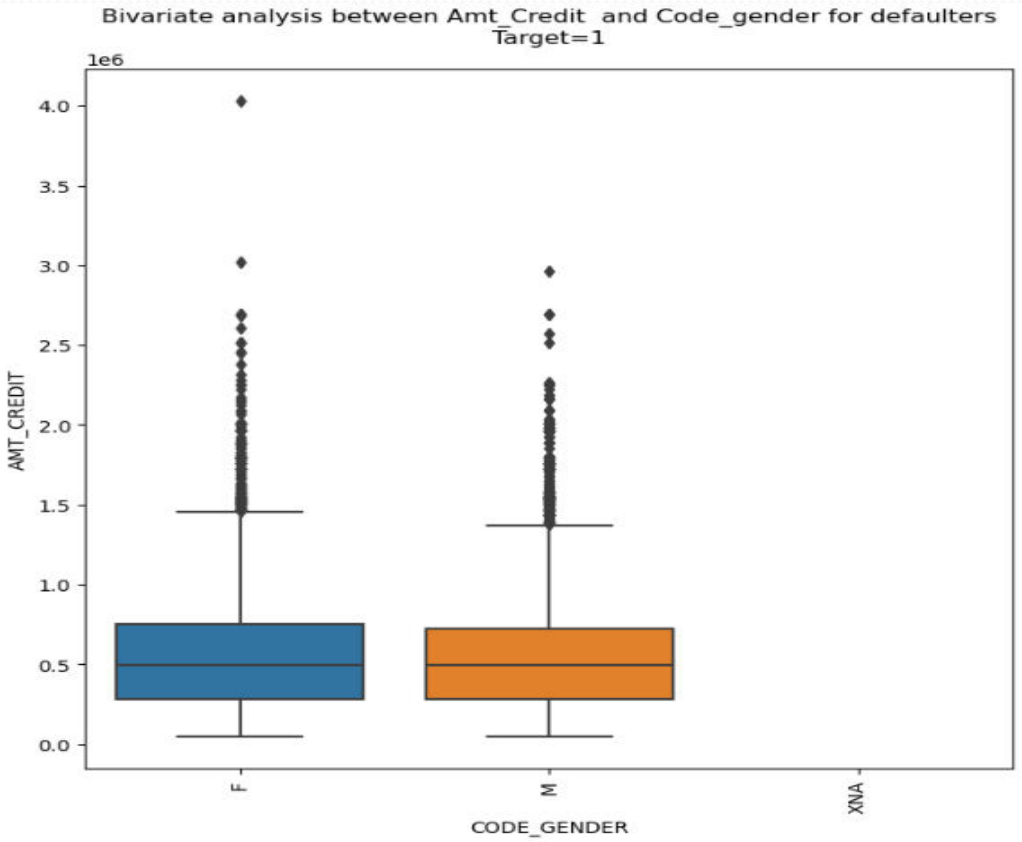
❖We can infer that although the maximum no of loans is given to the Medium Income group, but then default value per loans highest in higher income group. The loan book of the financial institution can get affected duo to higher amount not being paid back.

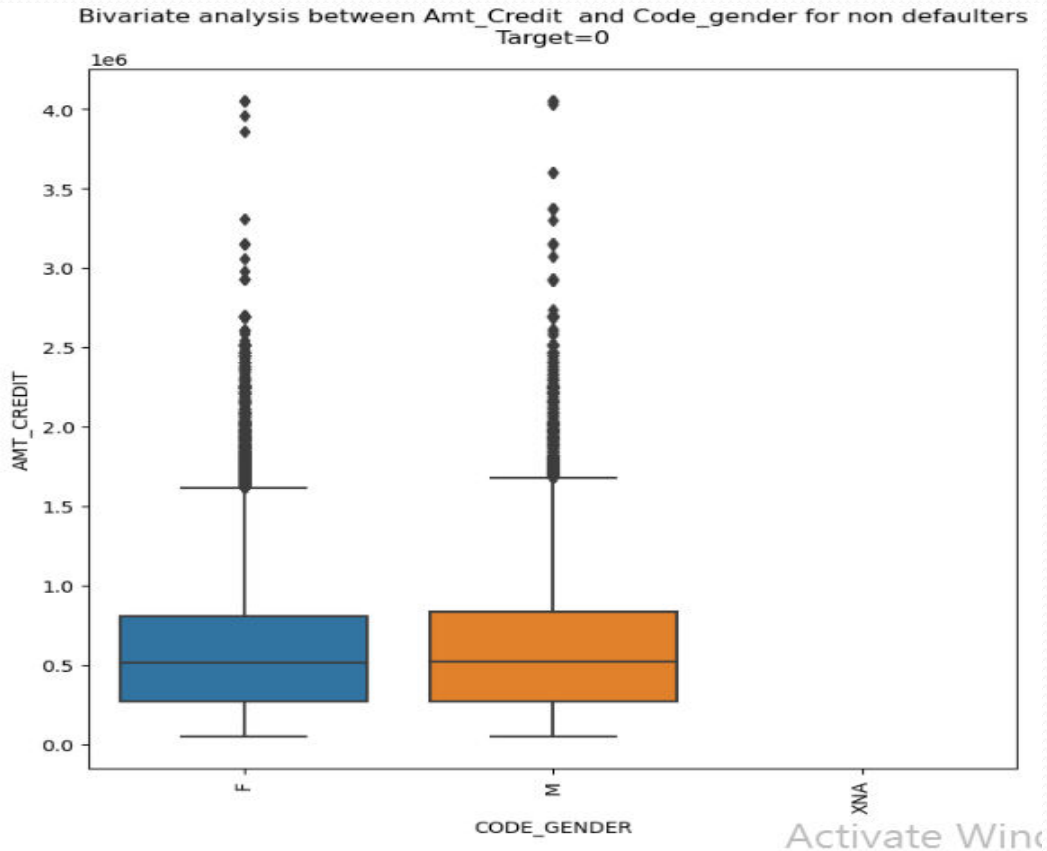




Target variable : Amt_credit & Gender

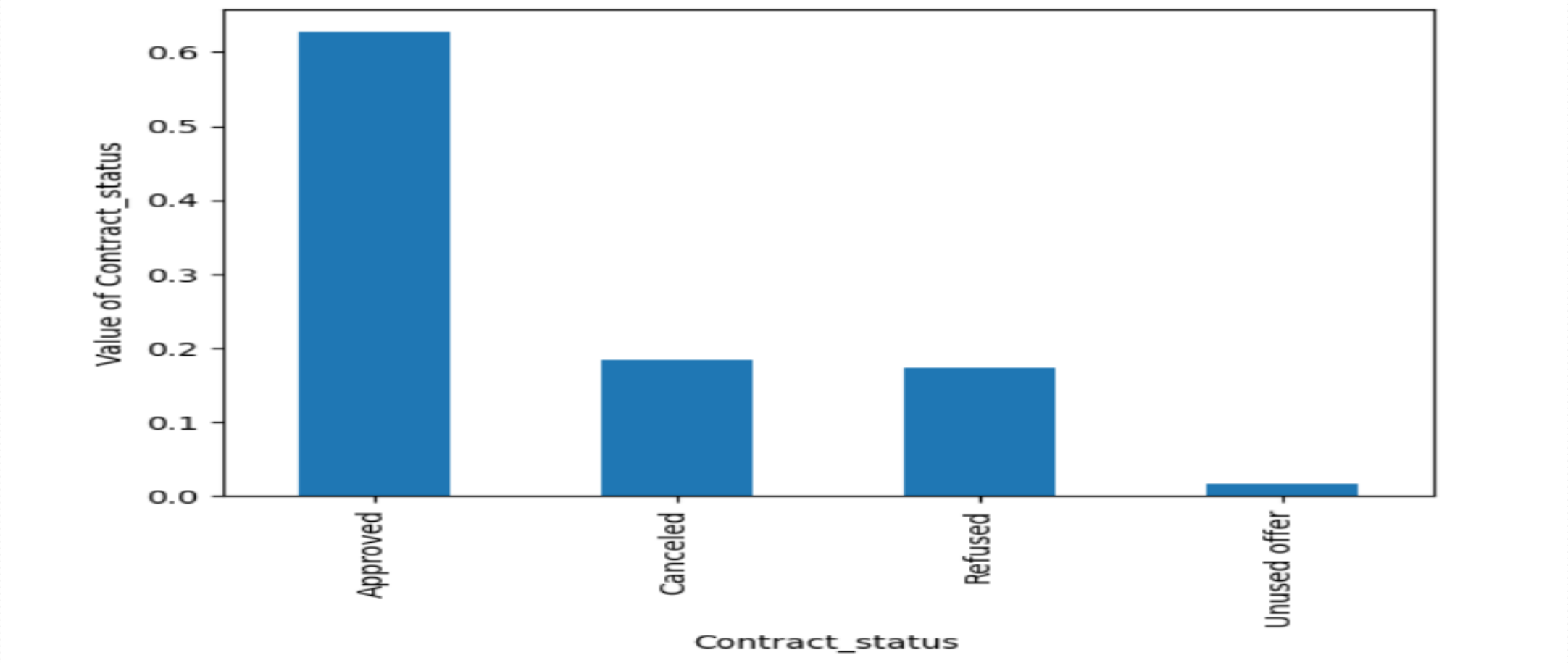
- However there are no big difference between male and female group but females have more amount credit range in defaulters. No of non defaulters in both male and female category are higher than defaulters.



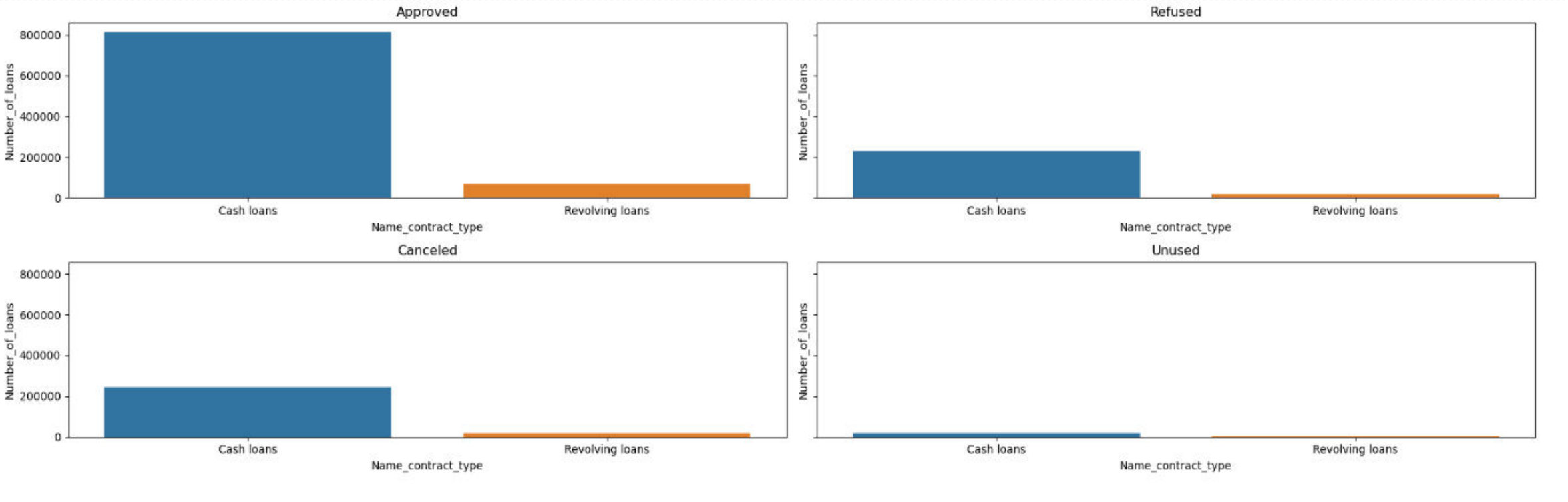


Analysis After Merging Two Dataframes .

1. Reading the previous data frame.
2. Merging the two data frame on the basis of customer id ('SK_ID_CURR').
3. Creating a new data base named ‘loandata’ .
4. Creating 4 variable based on Contract_status and analyse the further database in 4 categories.

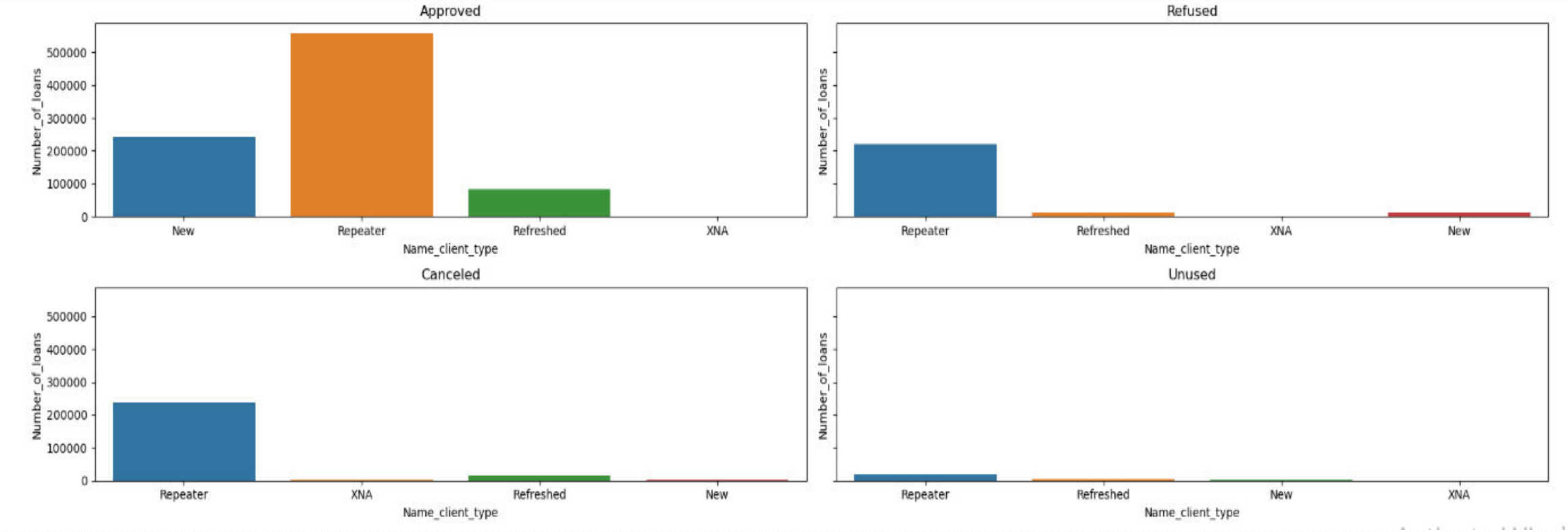


Analysis on variable NAME_CONTRACT_TYPE_X



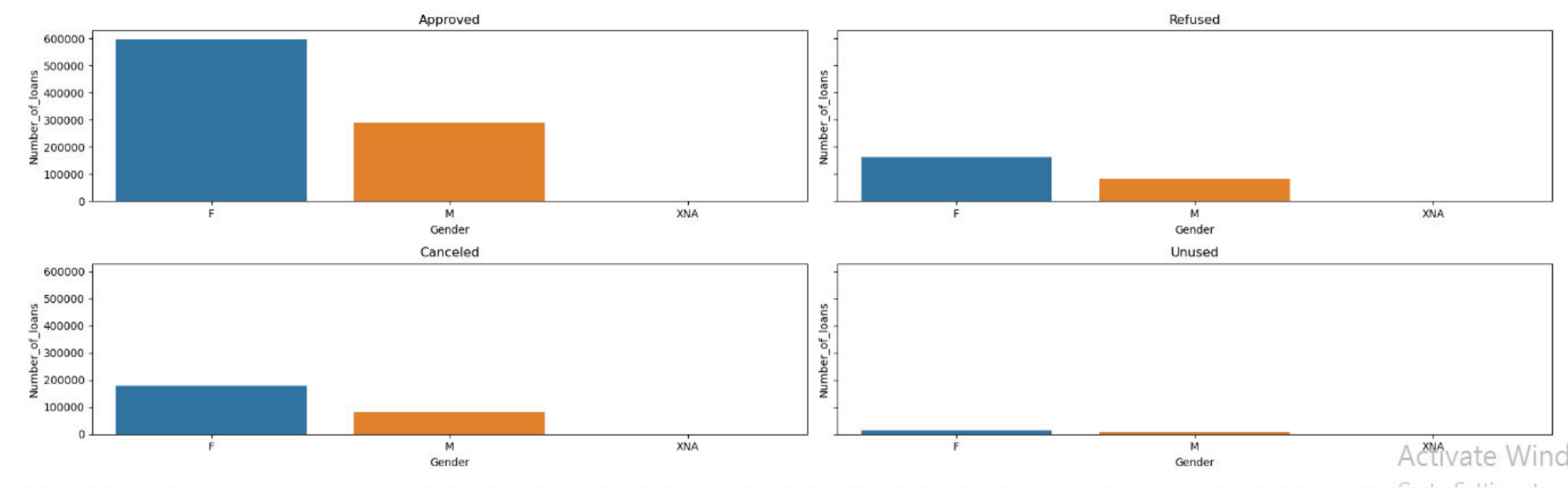
❖It is seen that the amount of cash loans is higher than revolving loans in all 4 categories.

Analysis on variable Name_client_type



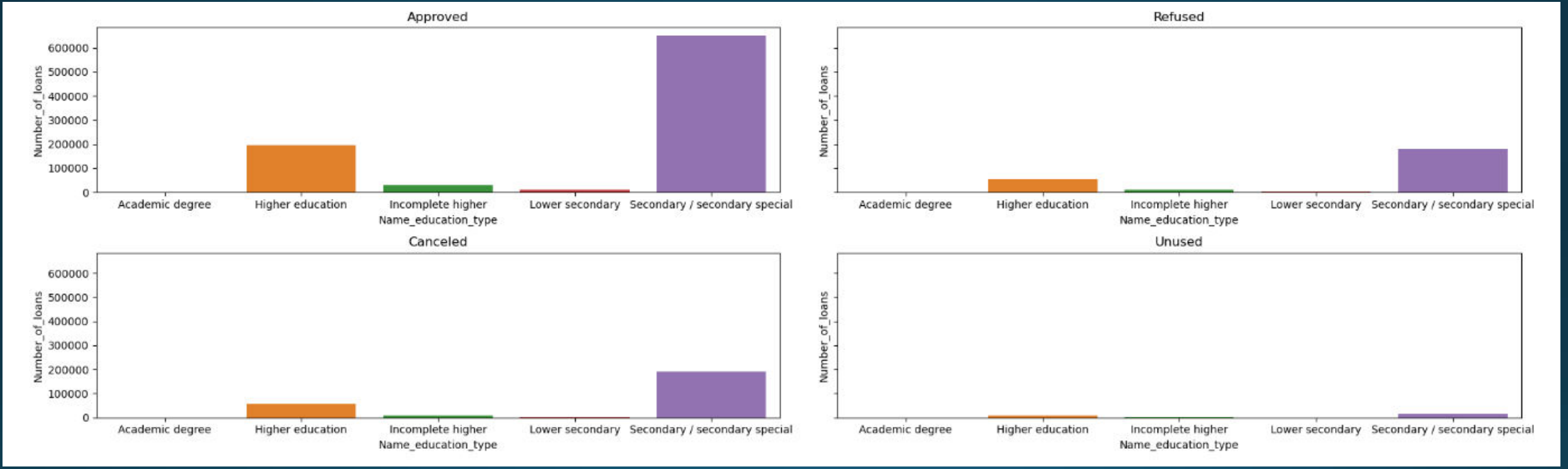
•Here we can see that the repeater is getting more refused and it is also getting more approved even that it is getting more canceled and unused.

Analysis on variable CODE_GENDER



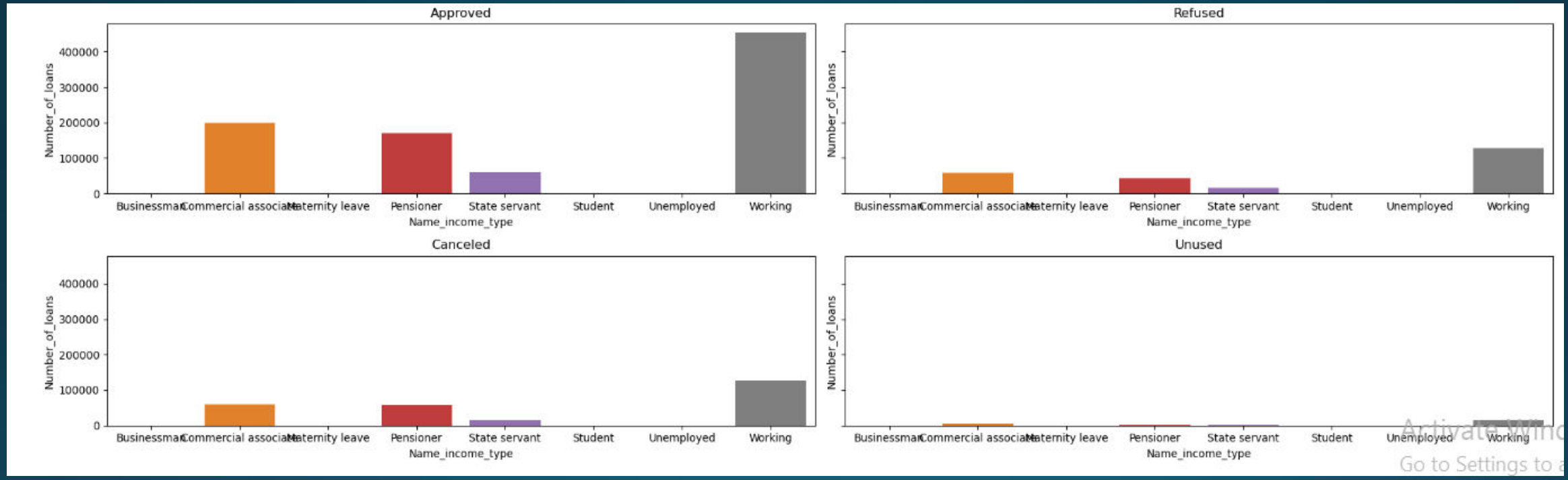
➢Here we can see that female are getting more approved, more refused, more canceled and more unused than male.

Analysis on variable Name_education_type



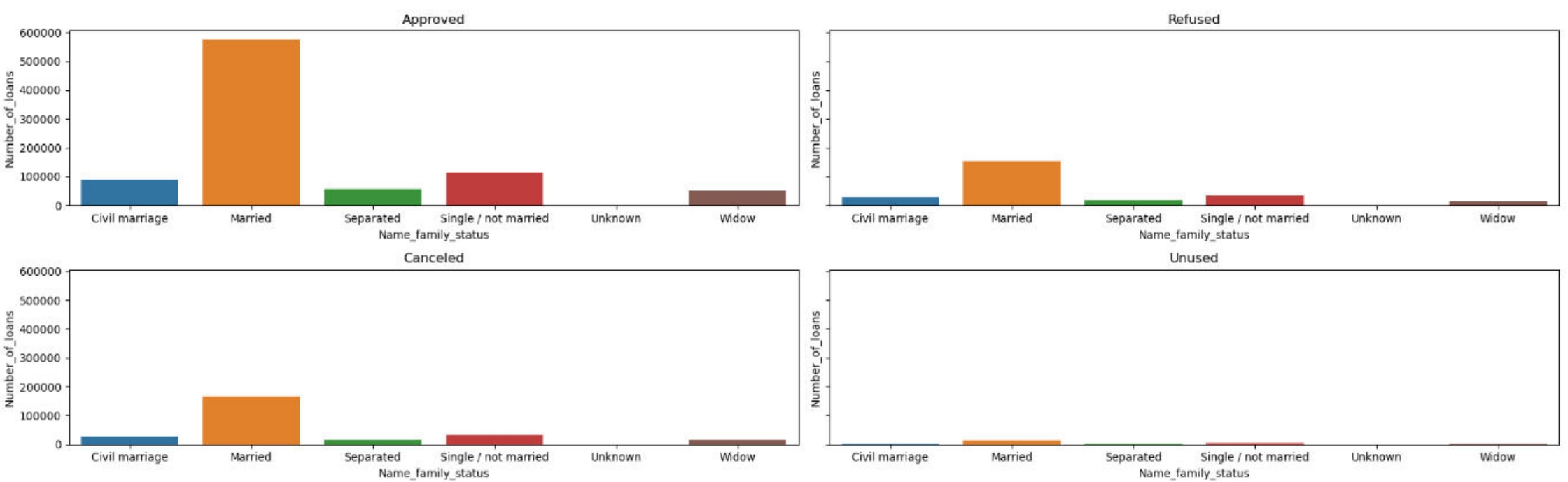
❖Here we can see that the secondary or secondary special is more effective in every cases .

Analysis on Name_Income_type



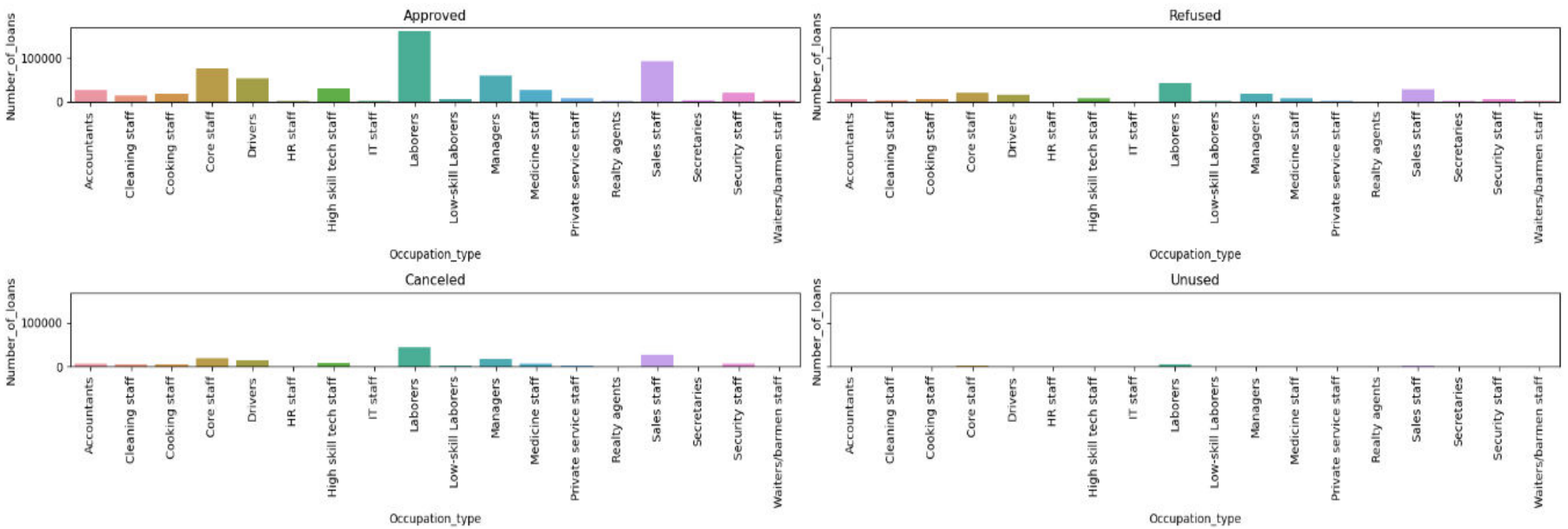
Here we can see that working people are applying more loans as compared to others and also commercial associates are taking more loans.

Analysis on Name_family_status



- Here we can see that married people are applying and taking more loans than others.

Analysis on Occupation_type



❑ **Here labourers are getting most refused and most approved loans. And also sales staff are also getting the second most refused and approved loans, so banks should focus on labourers.**

CONCLUSION

- Banks should focus more on contract type “students” “pensioners” and “Business mans “ with housing “type other than “Co-op apartments for successful payments.

- Banks Should focus less on income “working “ as they are having Most number of unsuccessful payments.

- Get an much as clients from having least housing type “with parents “ as they are having least number of unsuccessful payments.

THANK YOU