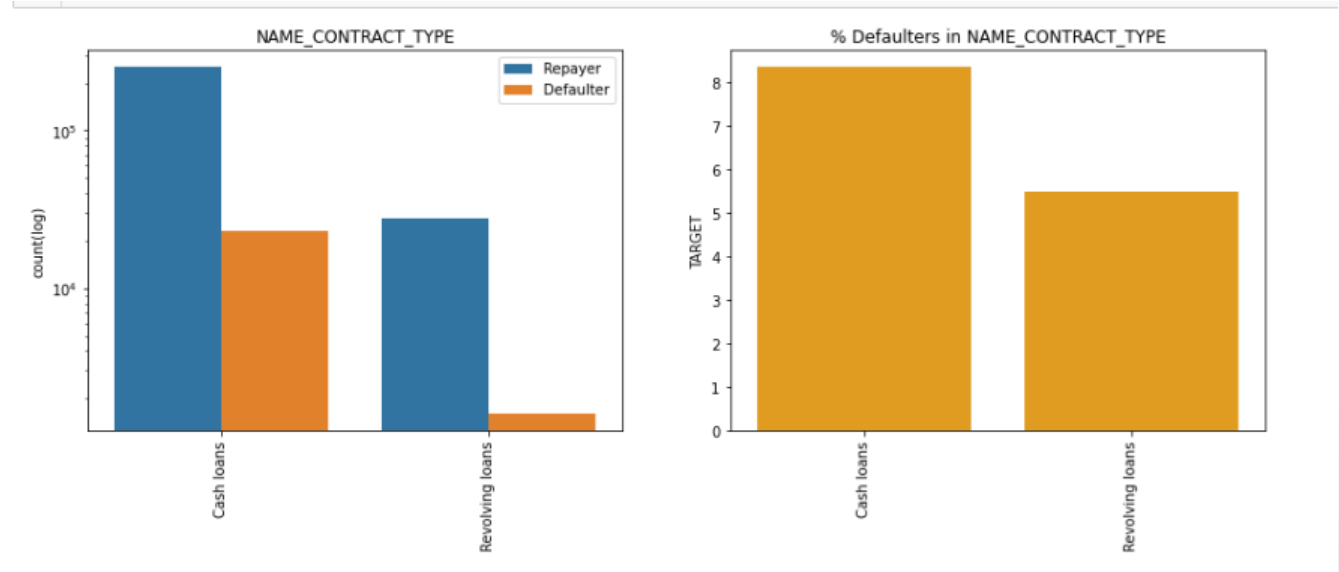
CREDIT EDA CASE STUDY

By Chetan S Daddikar

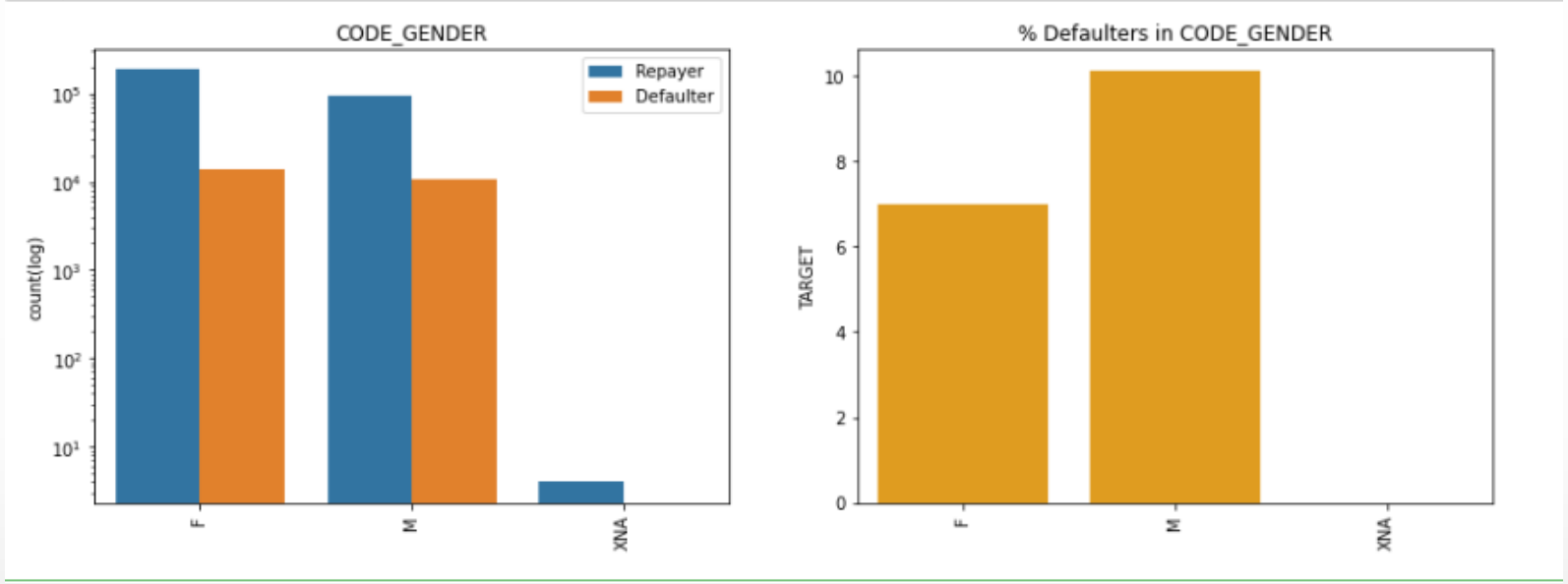
- **Univariate Analysis of Categorical Columns**

1) Count distribution of Repayers and Defaulters in 'NAME_CONTRACT_TYPE' col based on the 'TARGET' col



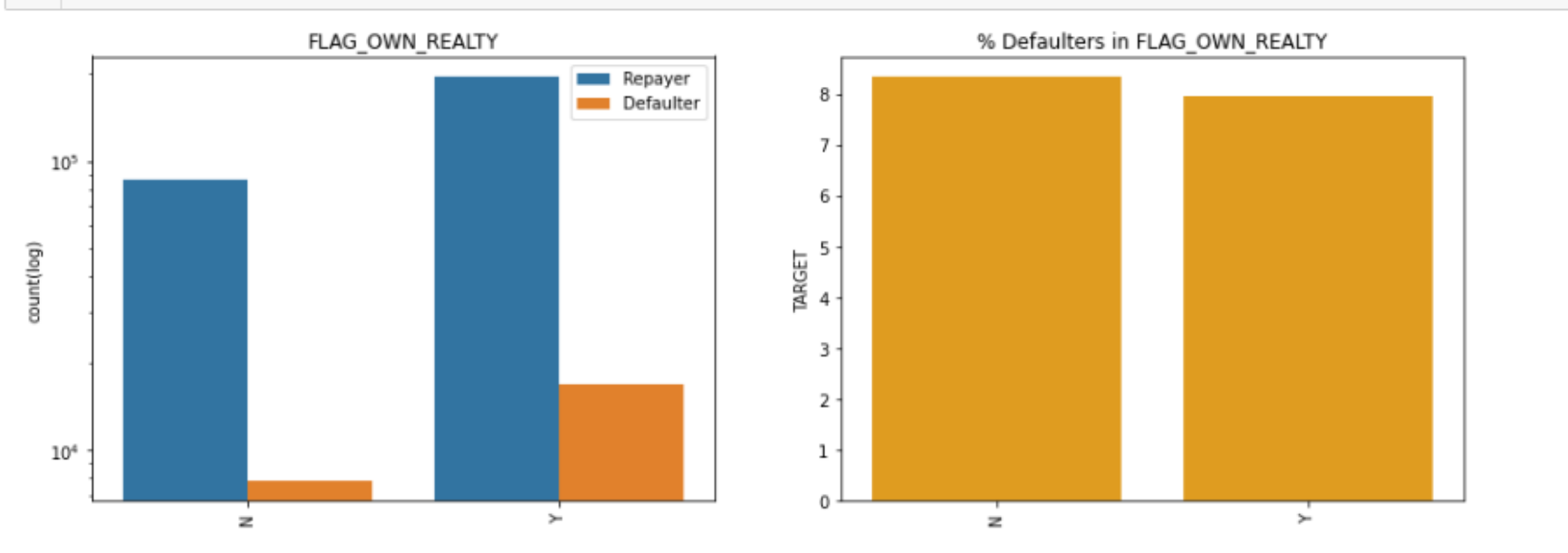
- There are two loan contract types 'Cash loans' & 'Revolving loans' present in our data. The applications for Cash loans are more than the Revolving loans. 8-9% defaulters (Clients with payment difficulty) are present in Cash loan sections and 5-6% defaulters are present in Revolving loan section.
- The clients with payment difficulty are more in Cash loan section than Revolving because of more number of applications.

2) Count distribution of Repayers and Defaulters in 'CODE_GENDER' col based on the 'TARGET' col



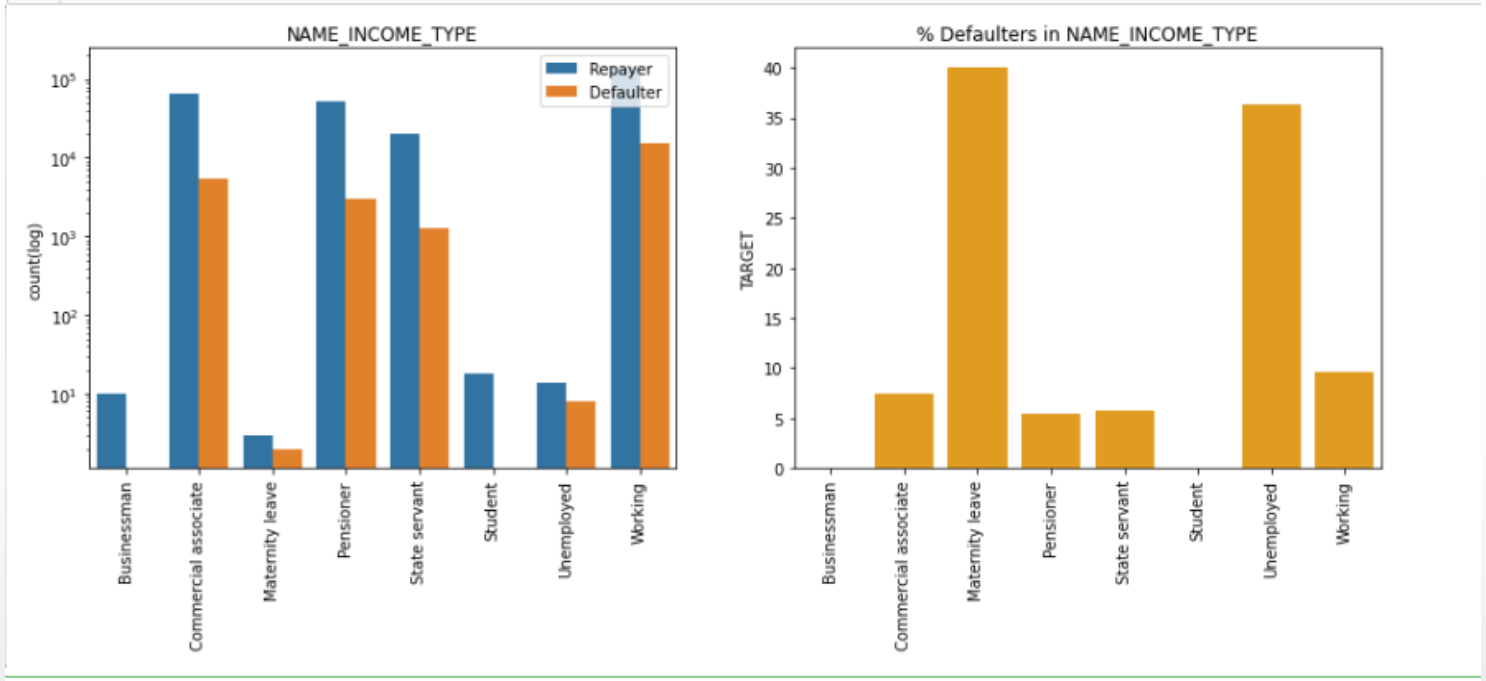
- From the above plots we can see that the male applicants have more probable chances of having payment difficulty than female applicants. The number of female applicants are also more than male.

3) Count distribution of Repayers and Defaulters in 'FLAG_OWN_REALTY' col based on the 'TARGET' col



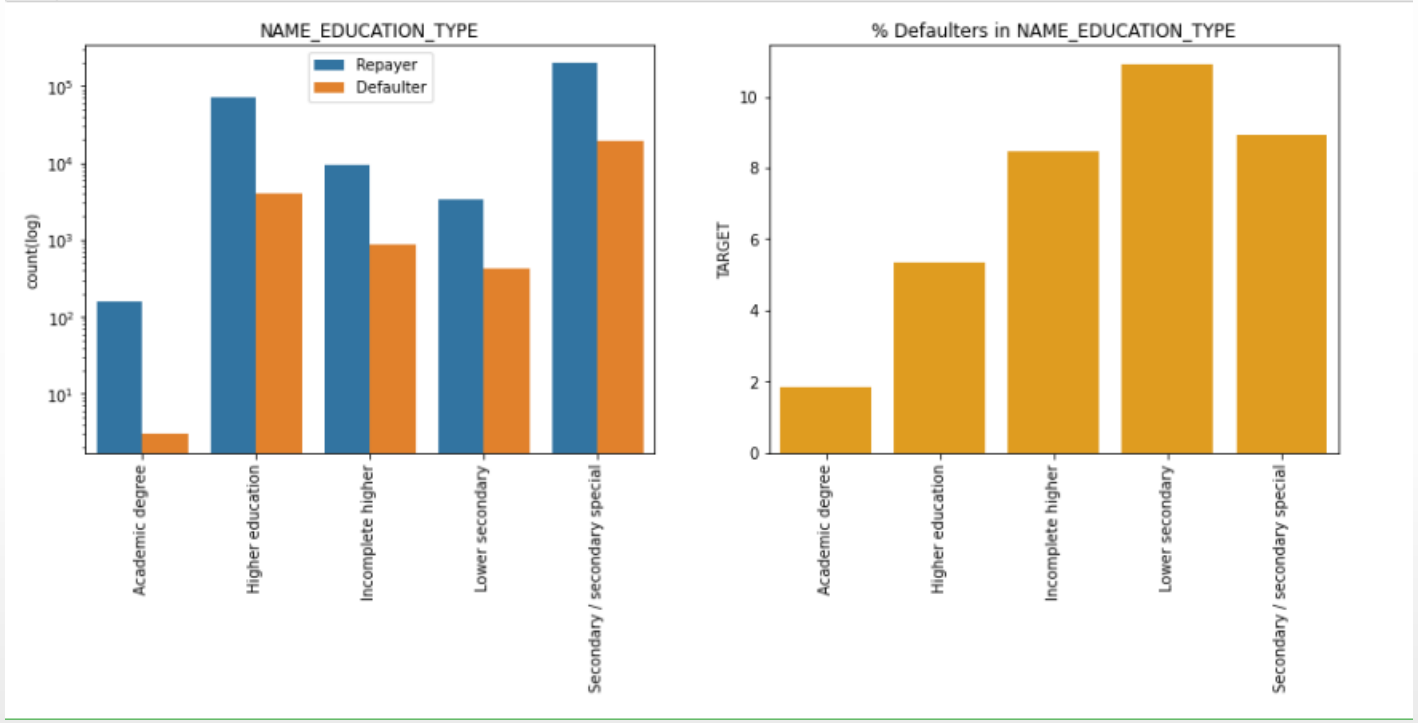
- The clients who purchased real estate property also have the payment difficulty almost equal to the clients who don't purchased the real estate property.
- From that we can say that there is non correlation between owning the real estate property and defaulting the loan.

4) Count distribution of Repayers and Defaulters in 'NAME_INCOME_TYPE' col based on the 'TARGET' col



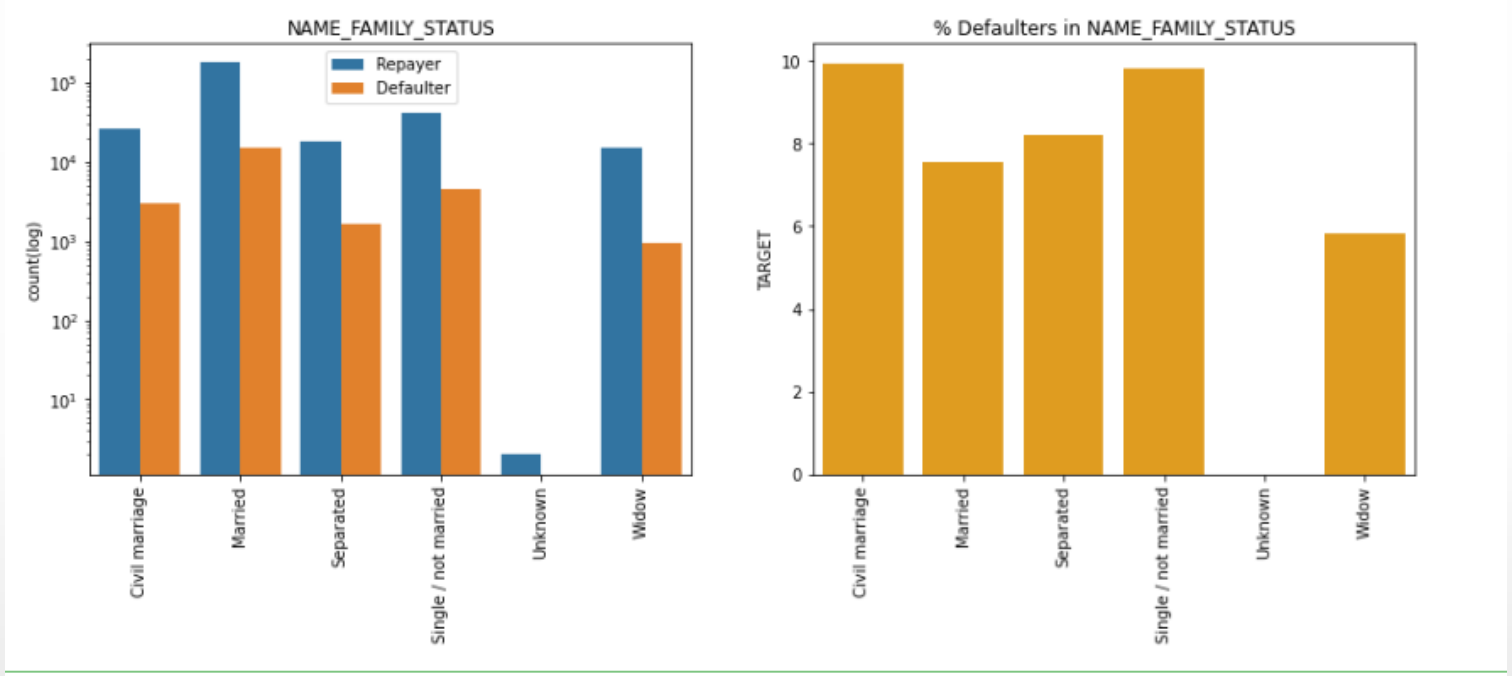
- Most of the clients applied for loans are from working category followed by commercial associate, pensioner and state servant.
- The clients from maternity leave have 40% of chances of payment difficulty followed by unemployed category. Businessman and students are safest categories to provide loan.

5) Count distribution of Repayers and Defaulters in 'NAME_EDUCATION_TYPE' col based on the 'TARGET' col



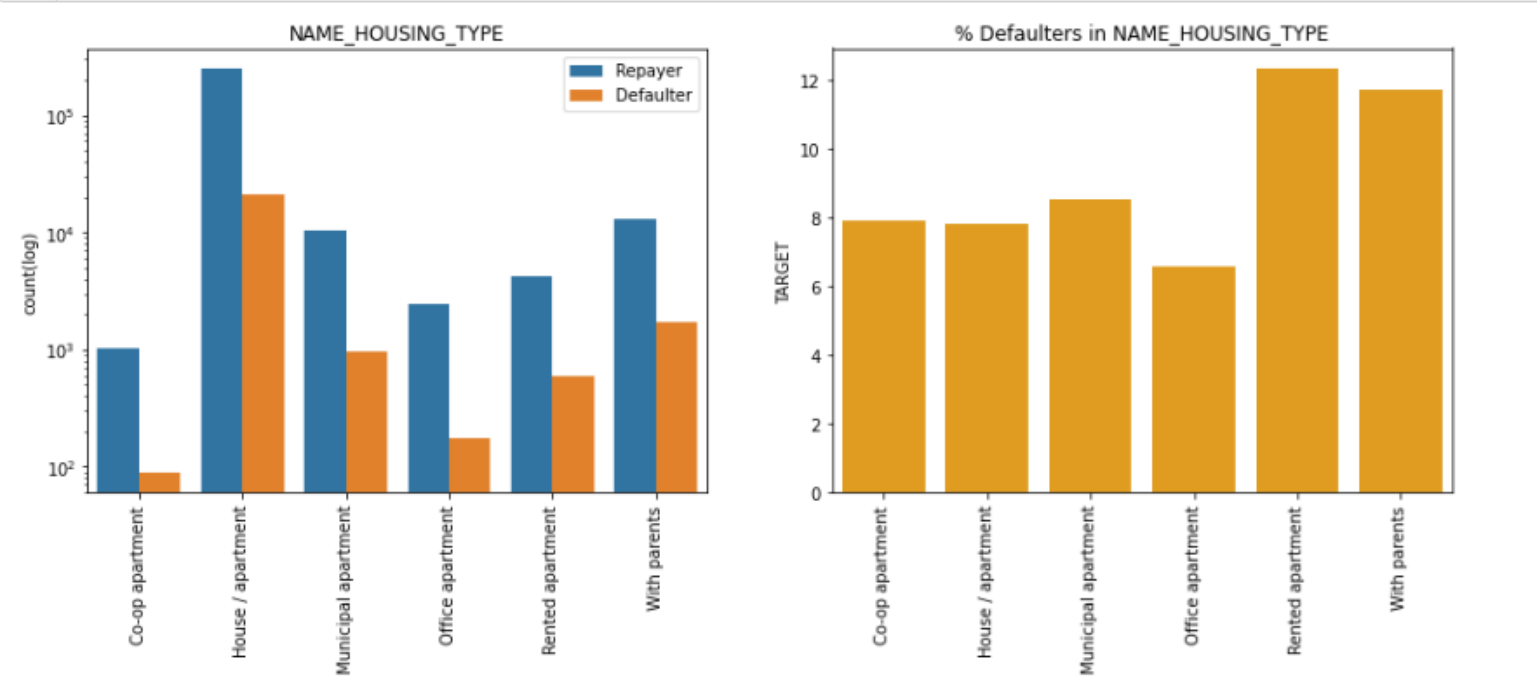
- Most of the loan applicants have completed their secondary education and higher education. The clients who completed their lower secondary education have highest rate of payment difficulty.
- There are less applicant who have academic degree and also have less chances of defaulting the loan.

6) Count distribution of Repayers and Defaulters in 'NAME_FAMILY_STATUS' col based on the 'TARGET' col



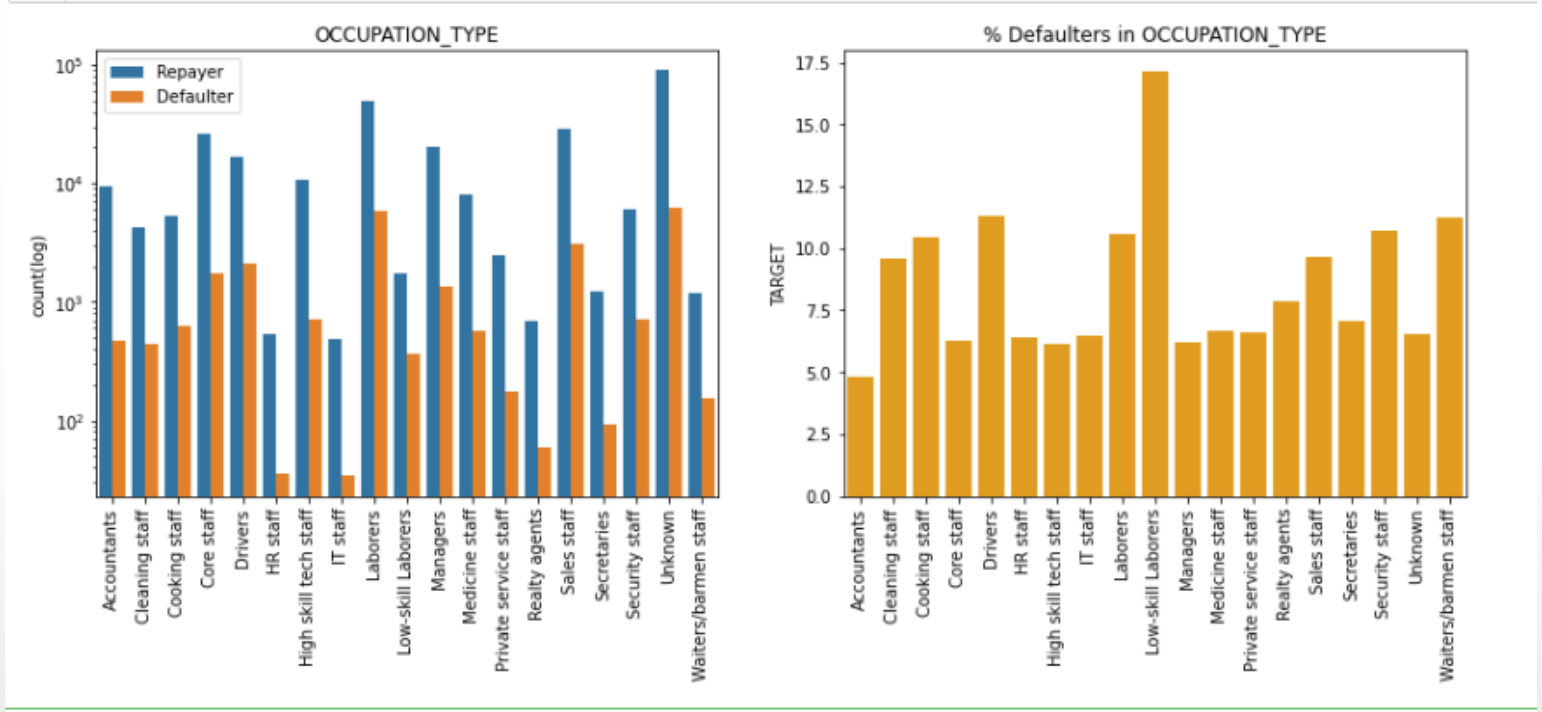
- There are more number of applicants from married category followed by single. The applicants from civil marriage category and single (not married) category have high chances of defaulting the loan.

7) Count distribution of Repayers and Defaulters in 'NAME_HOUSING_TYPE' col based on the 'TARGET' col



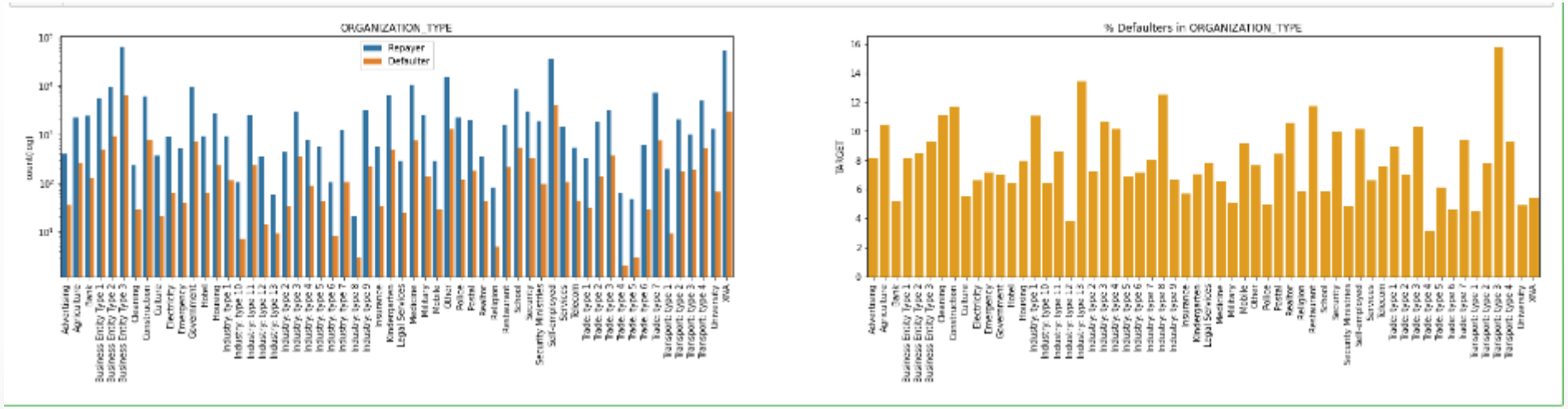
- Most of the clients applied for loan live in house/apartment. The people who rented apartment or live with parents have high chances of having payment difficulty.
- The people who live in office apartment is the best option to provide the loan as they contain less defaulters.

8) Count distribution of Repayers and Defaulters in 'OCCUPATION_TYPE' col based on the 'TARGET' col



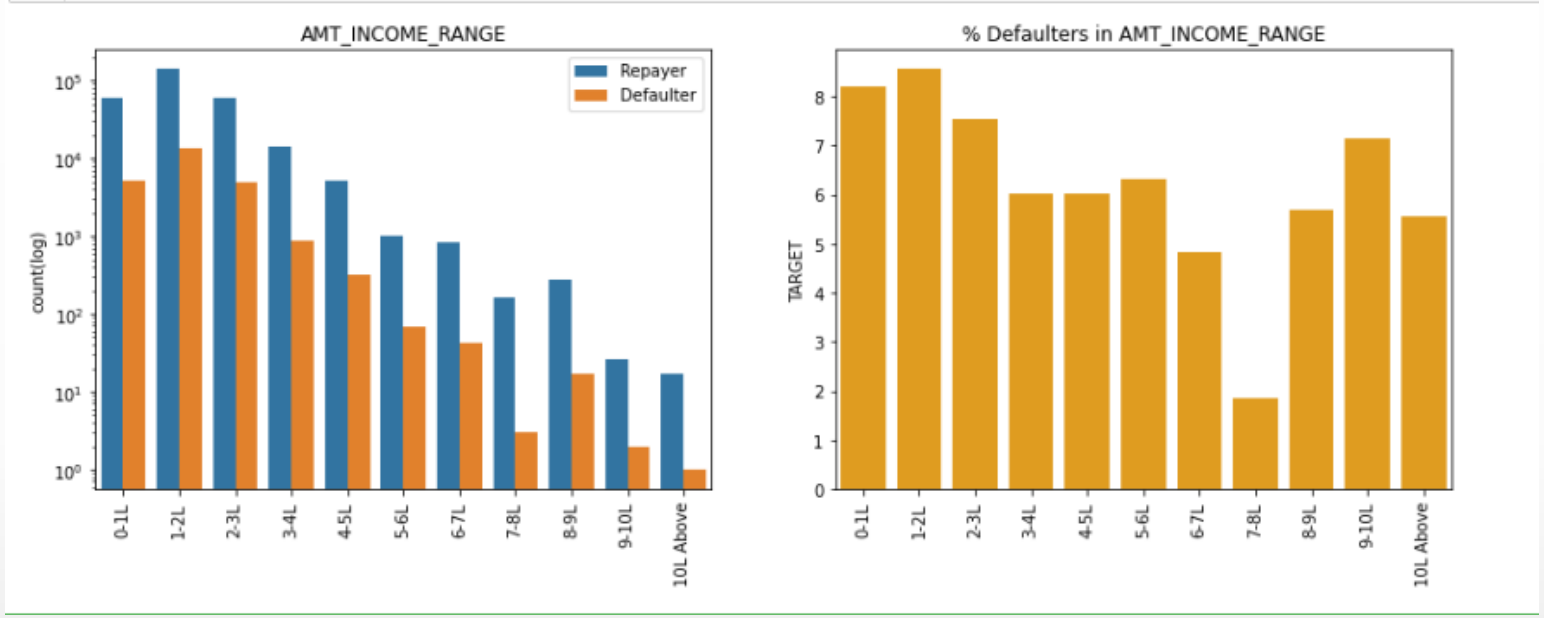
- The occupation type of most of the clients is unknown. Low skill laborers and drivers have more chances of defaulting the loan. The people who work as a accountant are likely to pay the loan.

9) Count distribution of Repayers and Defaulters in 'ORGANIZATION_TYPE' col based on the 'TARGET' col



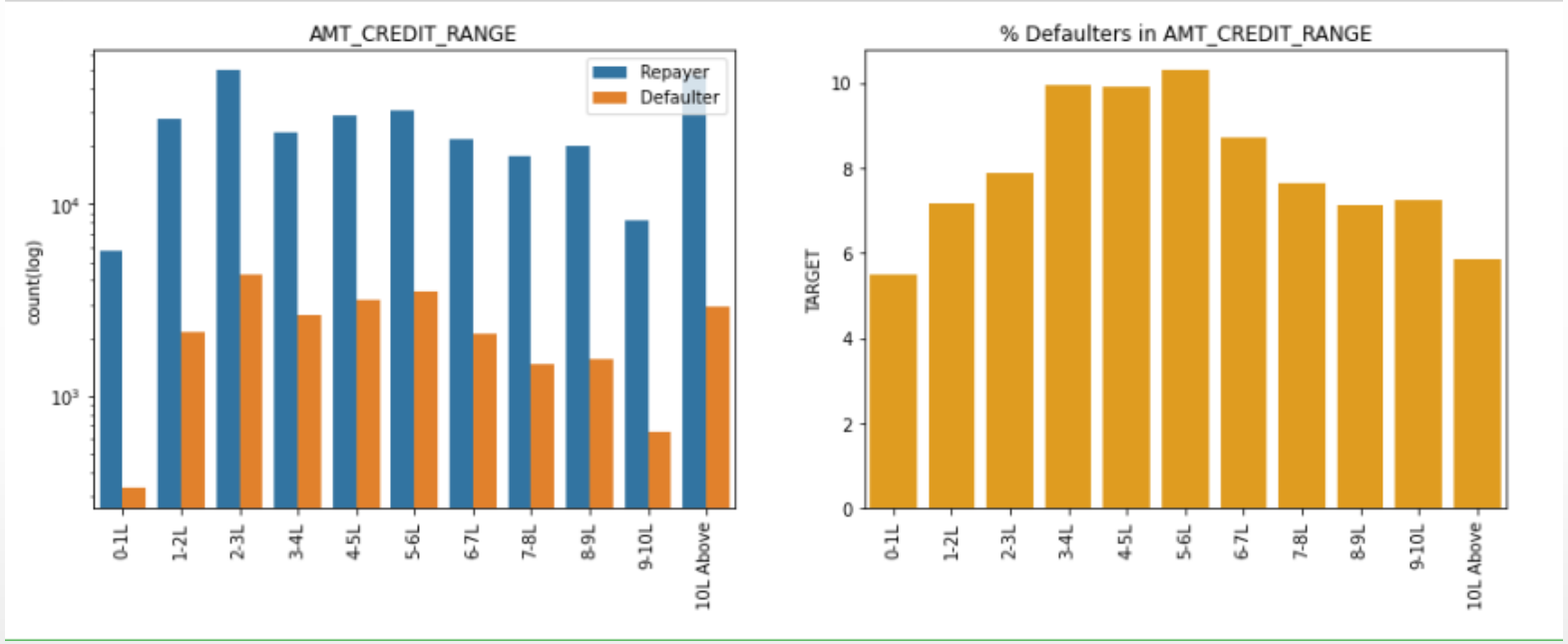
- The Organization type of most of the applicants is of business entity type 3. Followed by that most of the applicant are self employed.
- The applicants whose organization type is transport type 3 have more chances of having payment difficulties followed by that the applicants from industry type 13.
- The applicants working in industry type 12, trade type 4 and transport type 1 have less % of defaulting rate. For very high number of applications the information of organization type is not provided.

10) Count distribution of Repayers and Defaulters in 'AMT_INCOME_RANGE' col based on the 'TARGET' col



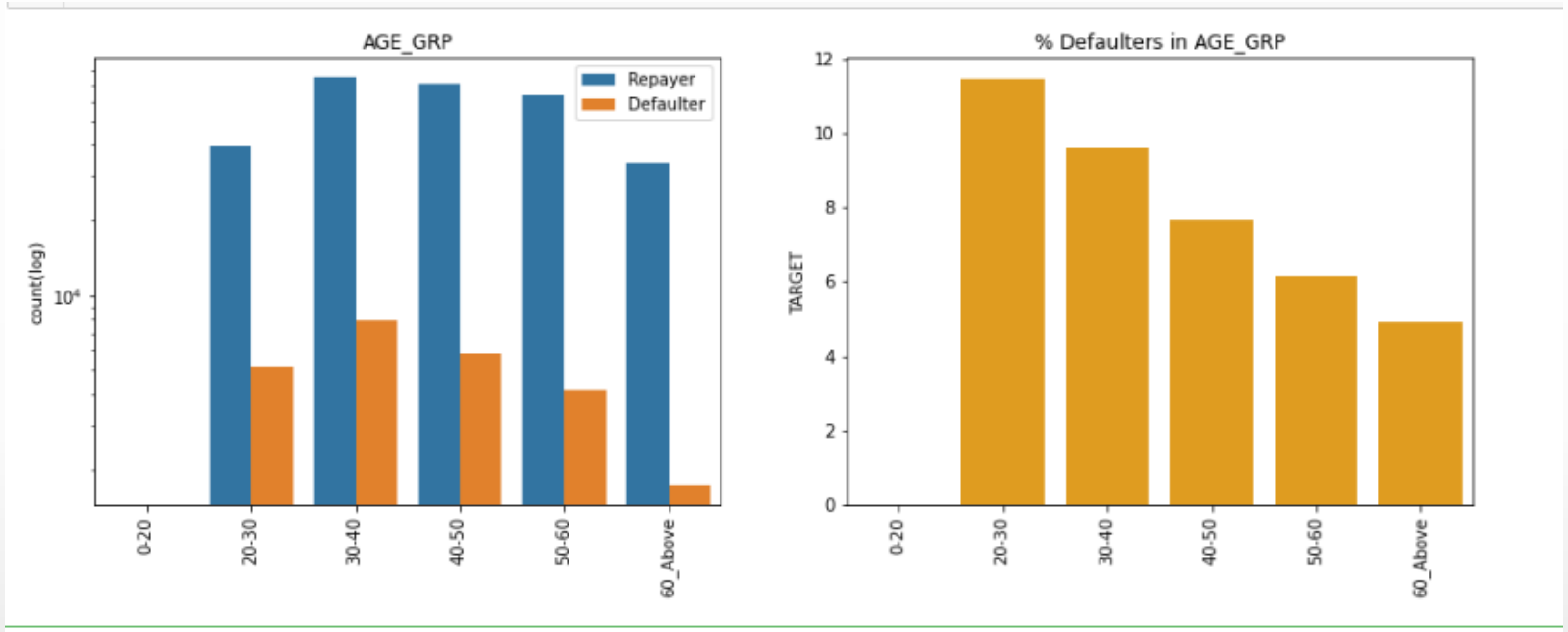
- The income range of most of the applicants is in 1-2L also the rate of having payment difficulty is more than 8% which is max. The people whose income range is 7-8L are more likely to pay the loan.

11) Count distribution of Repayers and Defaulters in 'AMT_CREDIT_RANGE' col based on the 'TARGET' col



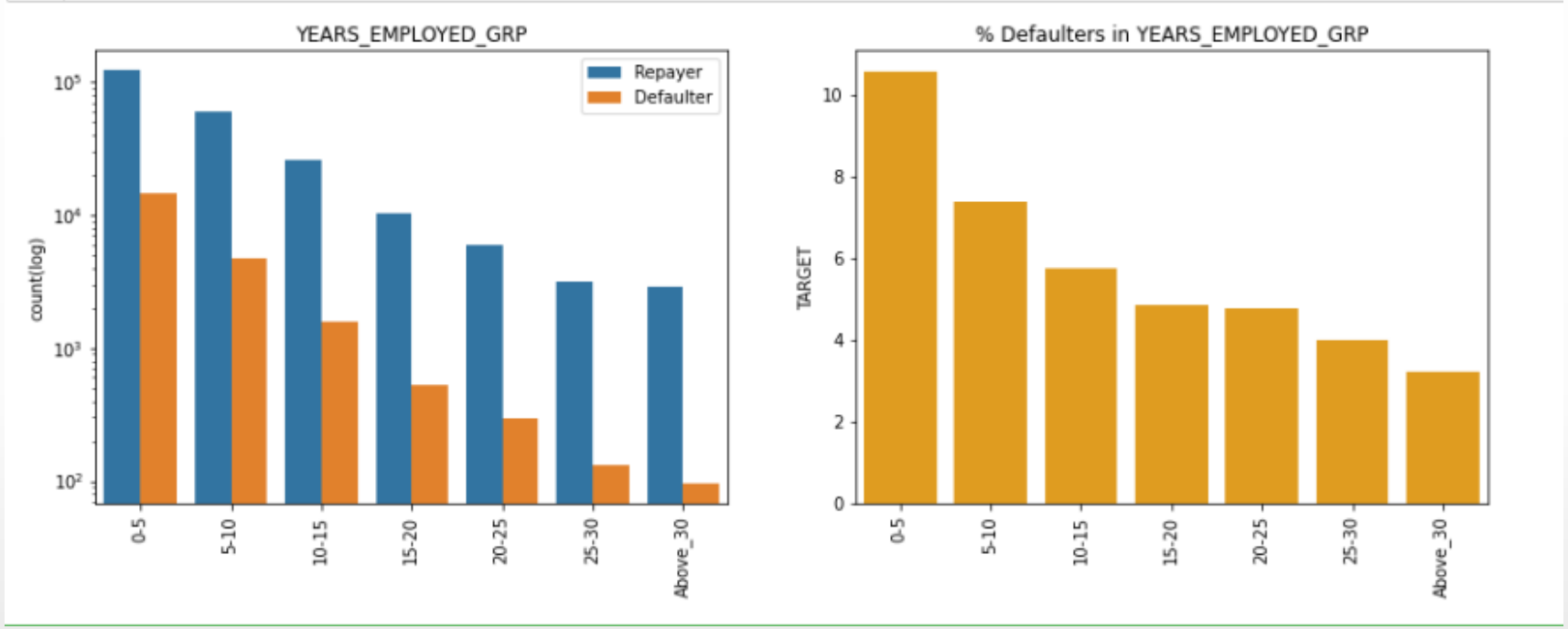
- Most of the applicants have loan in range of 2-3L and above 10L. The applicants having loan ranges from 3-6L have more defaulters.

12) Count distribution of Repayers and Defaulters in 'AGE_GRP' col based on the 'TARGET' col



- We can see here, most of the loan applicants are from 30-60 years old. The applicants whose age is 20-30 years have defaulters more than 11%.
- The applicants above 60 years old are more likely to pay the loan.

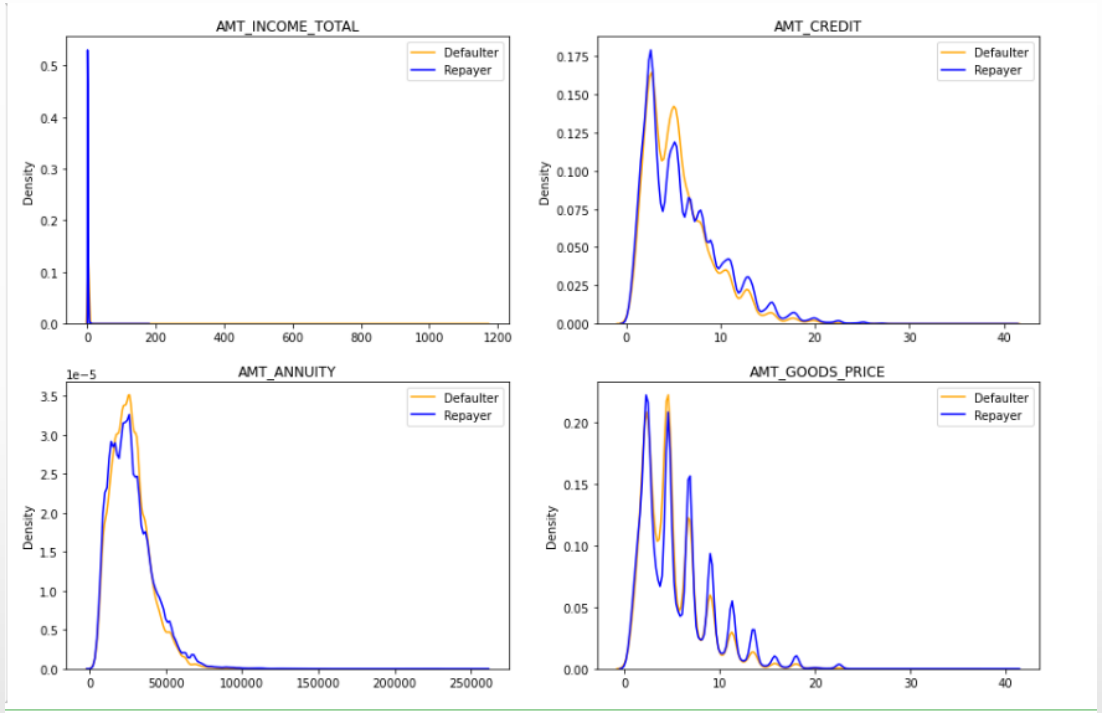
13) Count distribution of Repayers and Defaulters in 'YEARS_EMPLOYED_GRP' col based on the 'TARGET' col



- The applicants having more than 30 years of experience or employed more than 30 years have less than 4% defaulters and the applicants having 5-10 years of experience have more number of defaulters.
- With the increase in number of years of experience the defaulting rate decreases.

- **Univariate Analysis of Numerical Columns**

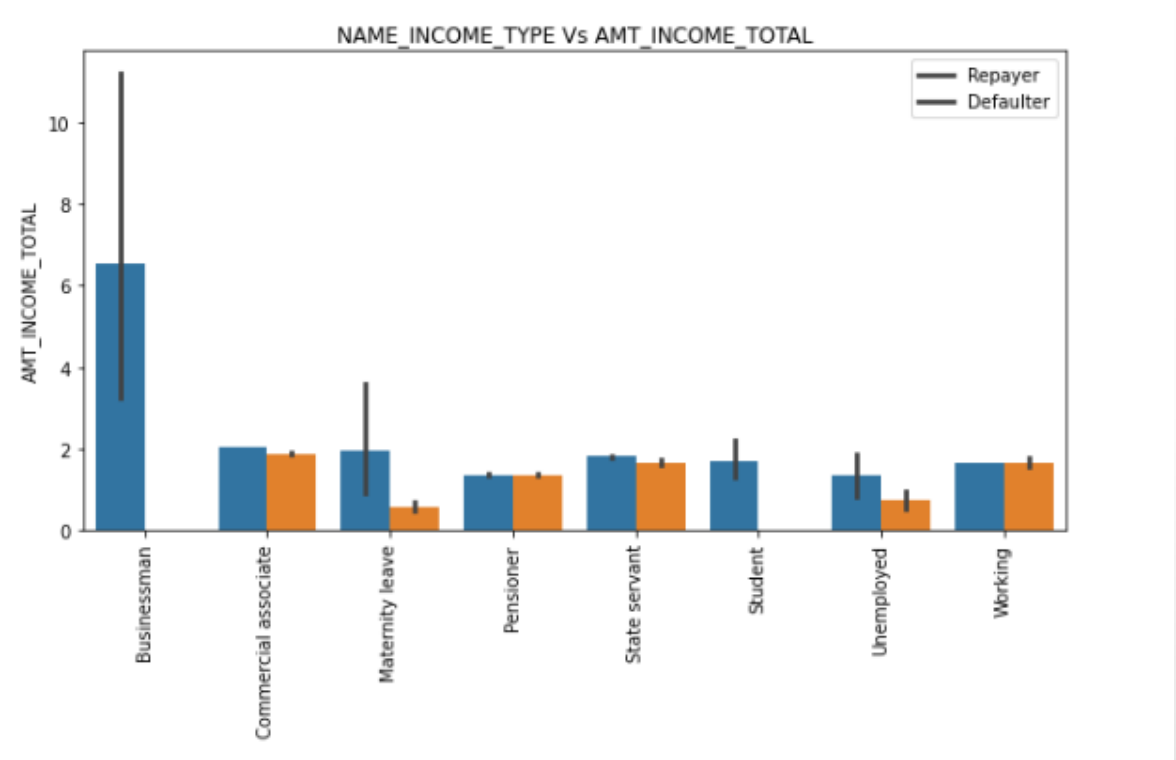
1) Subplot of distributions of each of 'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE‘ columns w.r.t TARGET column



- As we can see, the graphs for Defaulter, and for Repayers are overlapping means that we cannot use either of these variables to make a decision. The loan amount credited mostly is below 10L.
- For the loan credited the annuity paid by most of the people is below 50K. The maximum number of loans given to Goods price are below 10L.

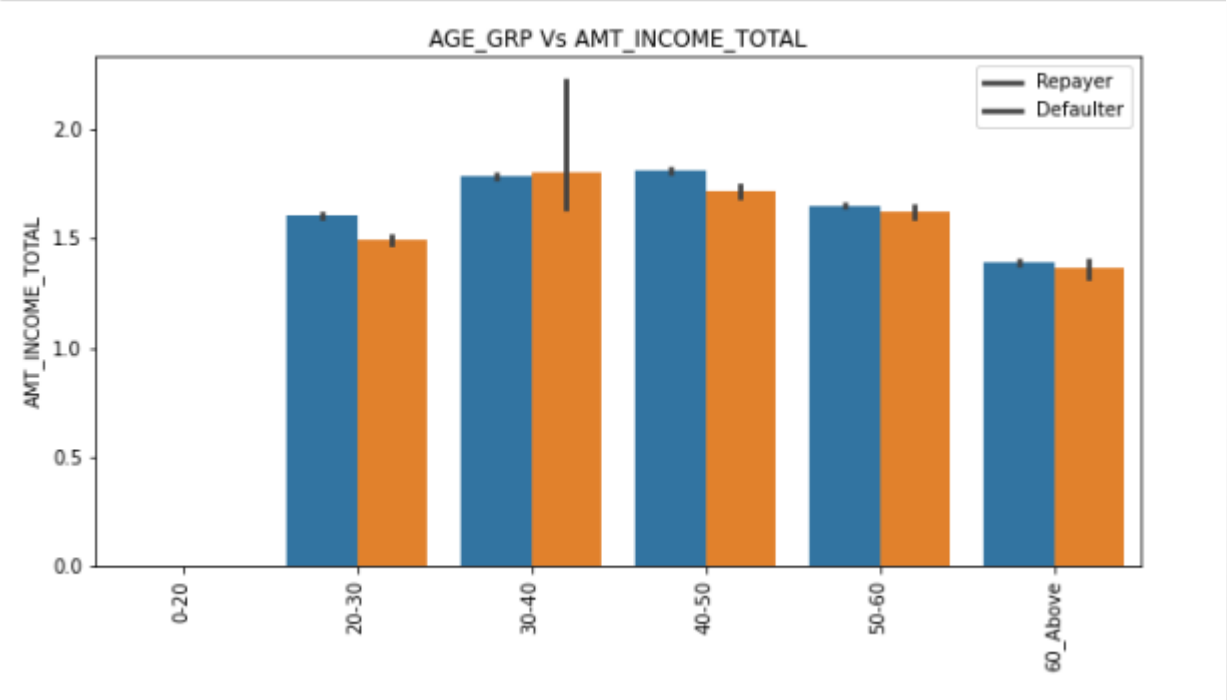
- **Bivariate Analysis of Categorical Columns**

1) Bar plot for 'NAME_INCOME_TYPE‘, 'AMT_INCOME_TOTAL‘ columns



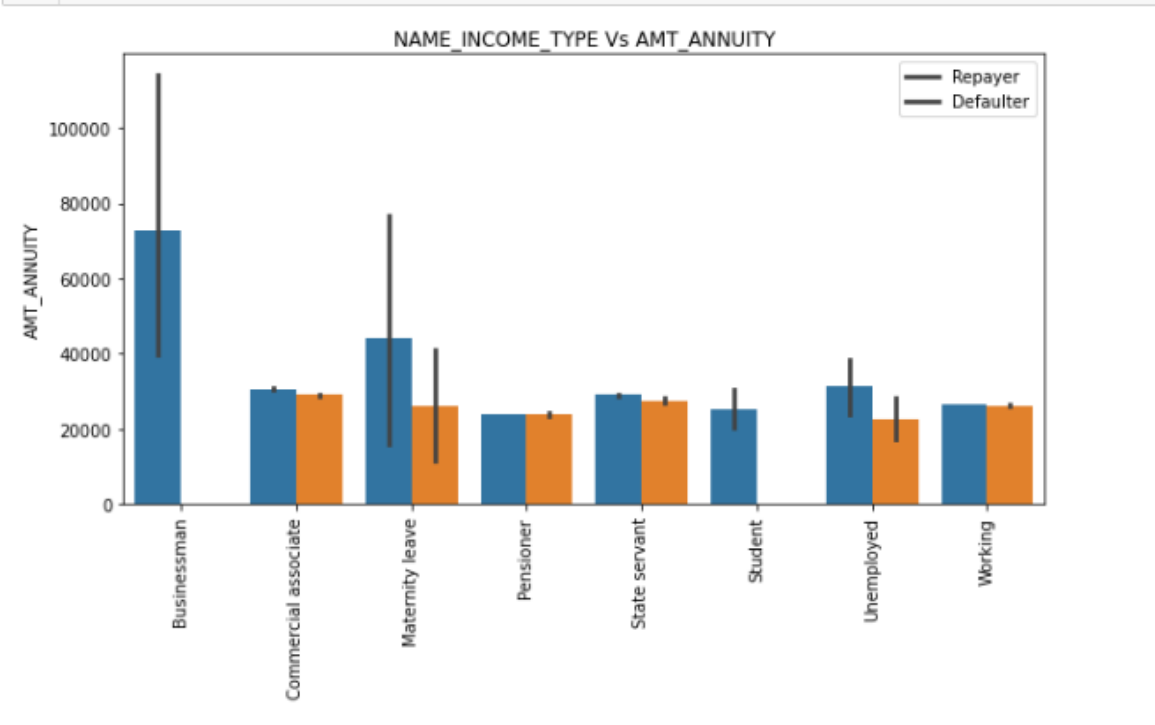
- From above statistical data and graph we can see that, the average total income of businessman income type category is 6.5L which is maximum.
- There are no defaulters in businessman and student income type category. These categories are the safest to provide the loan. The average income of student type category is 1.7L.

2) Bar plot for 'AGE_GRP‘, 'AMT_INCOME_TOTAL‘ columns



- The average total income of people having age group between 30-50 is approx. 1.8L.

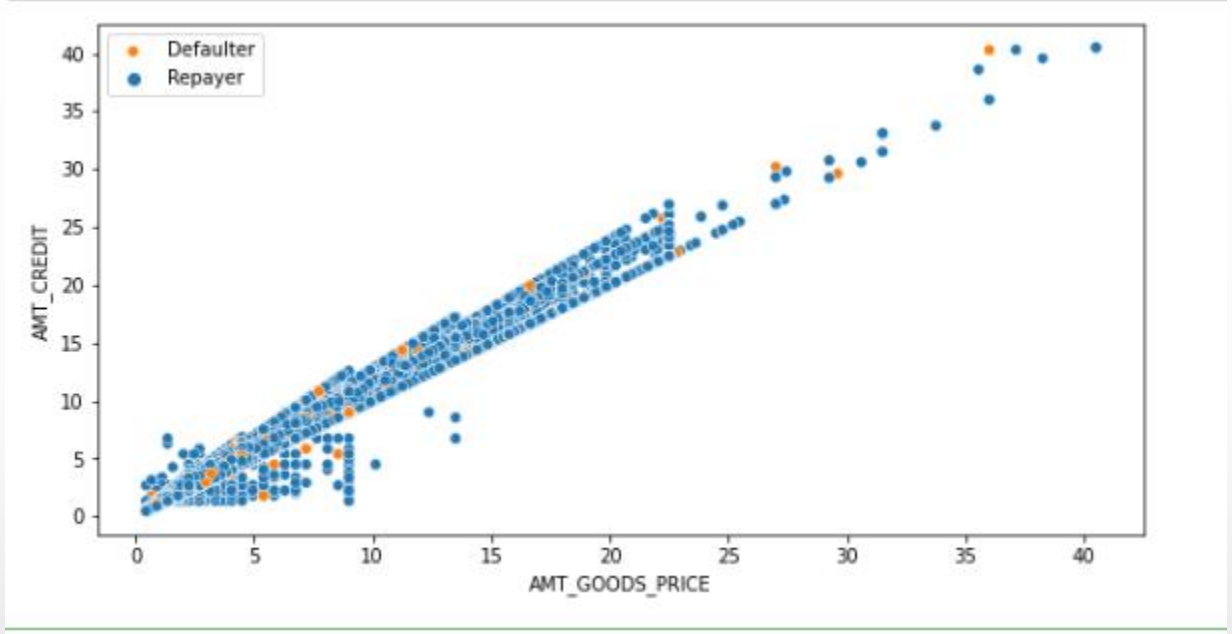
3) Bar plot for 'NAME_INCOME_TYPE', 'AMT_ANNUITY' columns



- As we have seen the average total income of businessman category is maximum. Therefore the annuity paid by the businessman category is also maximum ranges from 40K to more than 1L.
- The average annuity paid by the businessman is approx. 72K. The annuity paid by the maternity leave category is second maximum ranges from 20K to 80K.

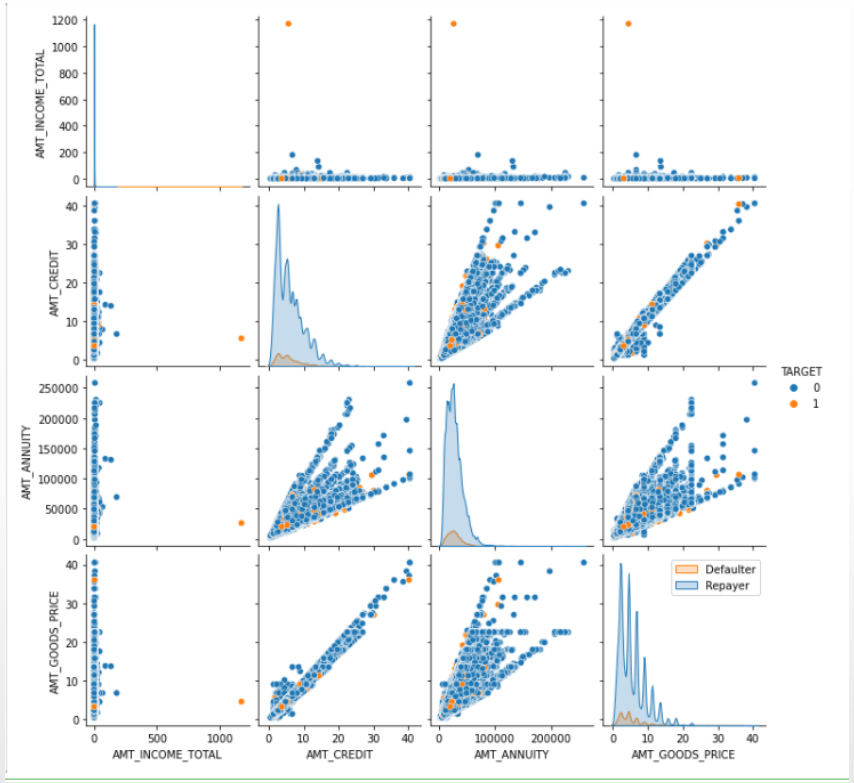
- **Bivariate Analysis of Numerical Columns**

1) Relation between two numeric columns 'AMT_GOODS_PRICE' and 'AMT_CREDIT'



- We can see here, as the credit amount increases above 25-30L the number of defaulters decreases.

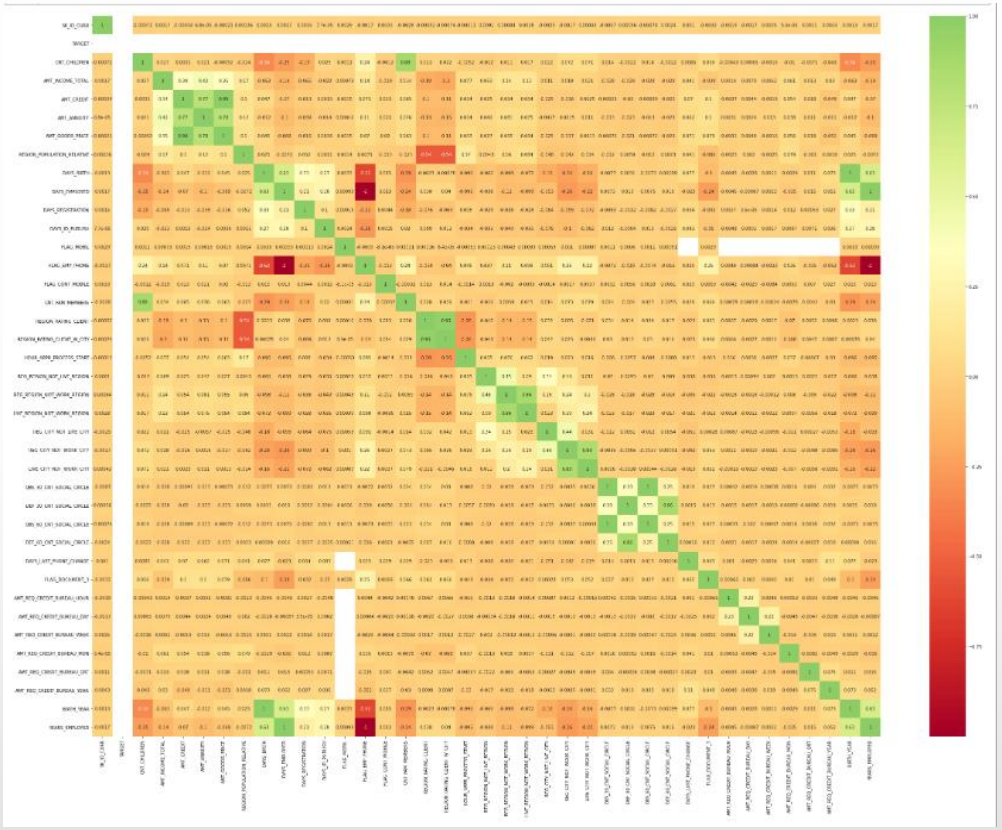
2) Pair plot for AMT variables to check the relation between each AMT column with respect to TARGET column.



- There is a high correlation between 'AMT_CREDIT' and 'AMT_GOODS_PRICE'. 'AMT_CREDIT' linearly increases with 'AMT_GOODS_PRICE'. As the credit amount increases above 20L the number of defaulters decreases. When the credited amount is above 20-25L the annuity paid is above 1.5L, there are less defaulters.

- **Correlation Analysis**

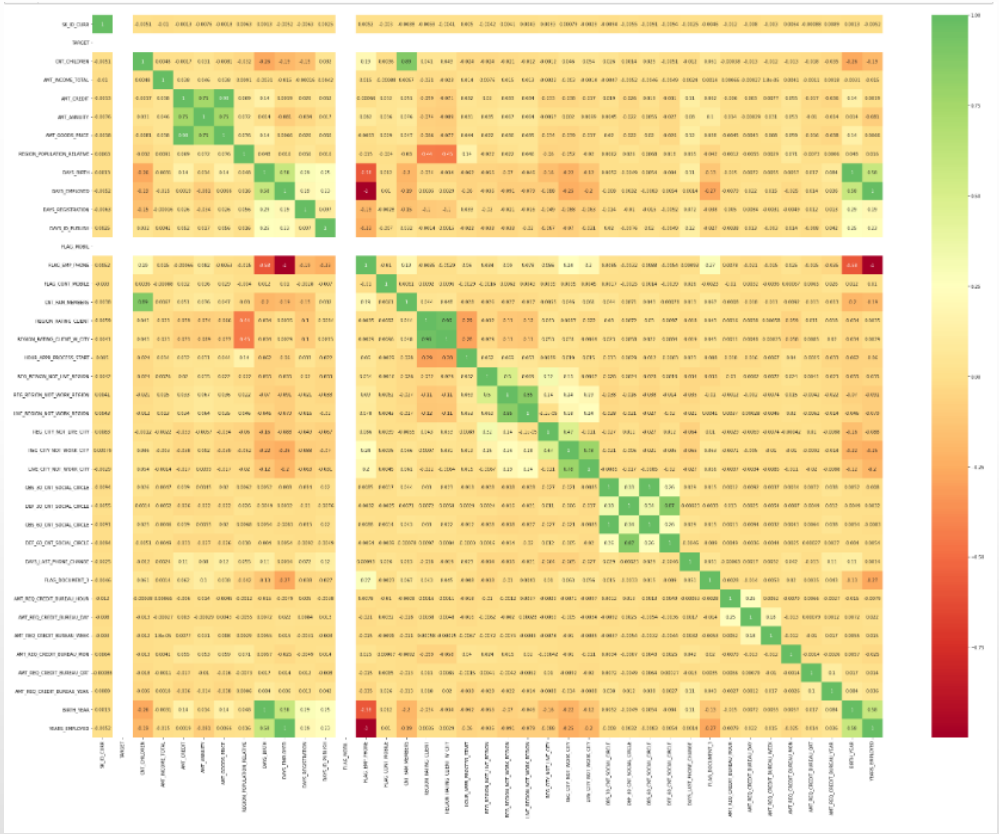
1) Correlation between Repayer_df



- There is high +ve linear correlation between 'AMT_CREDIT' 'AMT_GOODS_PRICE' and 'AMT_ANNUITY'. 'AMT_INCOME _TOTAL' also have +ve correlation with 'AMT_CREDIT' 'AMT_GOODS_PRICE' and 'AMT_ANNUITY'.

- **Correlation Analysis**

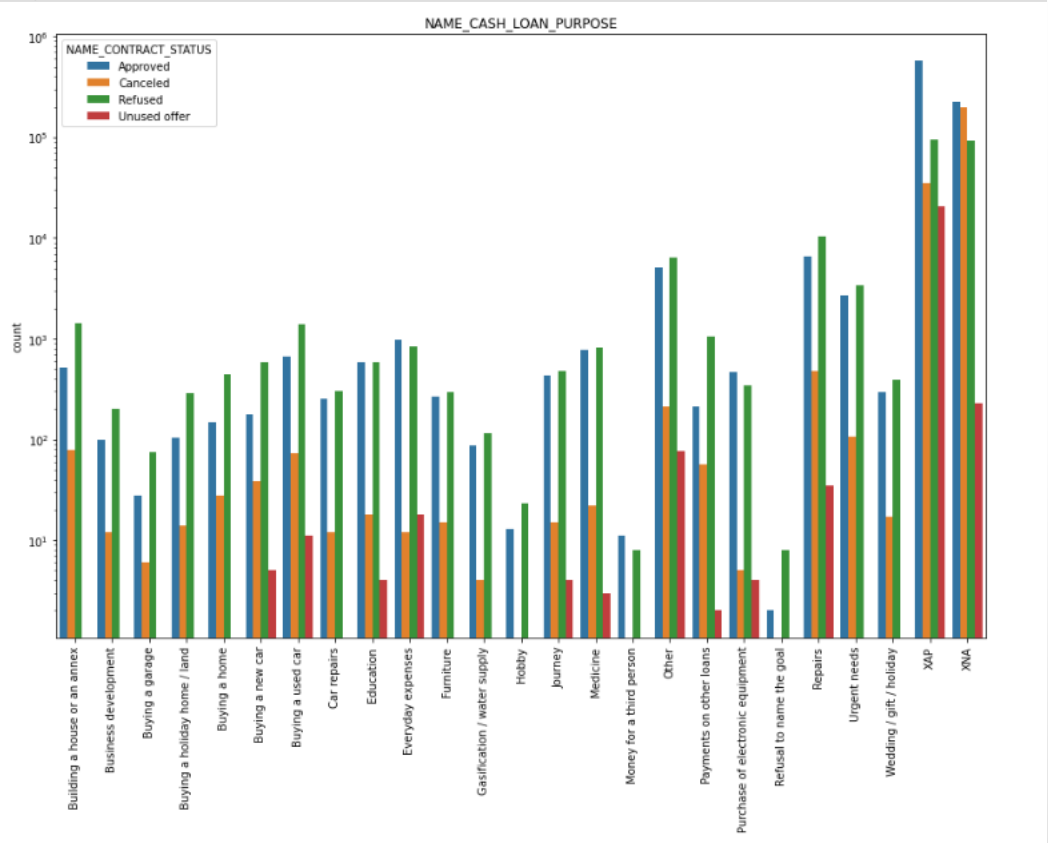
2) Correlation between Defaulter_df



- As in the Repayer_df 'AMT_CREDIT' have strong +ve correlation with 'AMT_GOODS_PRICE' and 'AMT_ANNUITY' with slightly less Pearson coefficient. There is a Sevier drop in correlation between 'AMT_INCOME _TOTAL' and 'AMT_CREDIT'. As compared to Repayer_df the Defaulter_df has lowered correlation between 'DAYS_EMPLOYED'.

- **Final Analysis of Merged DF**

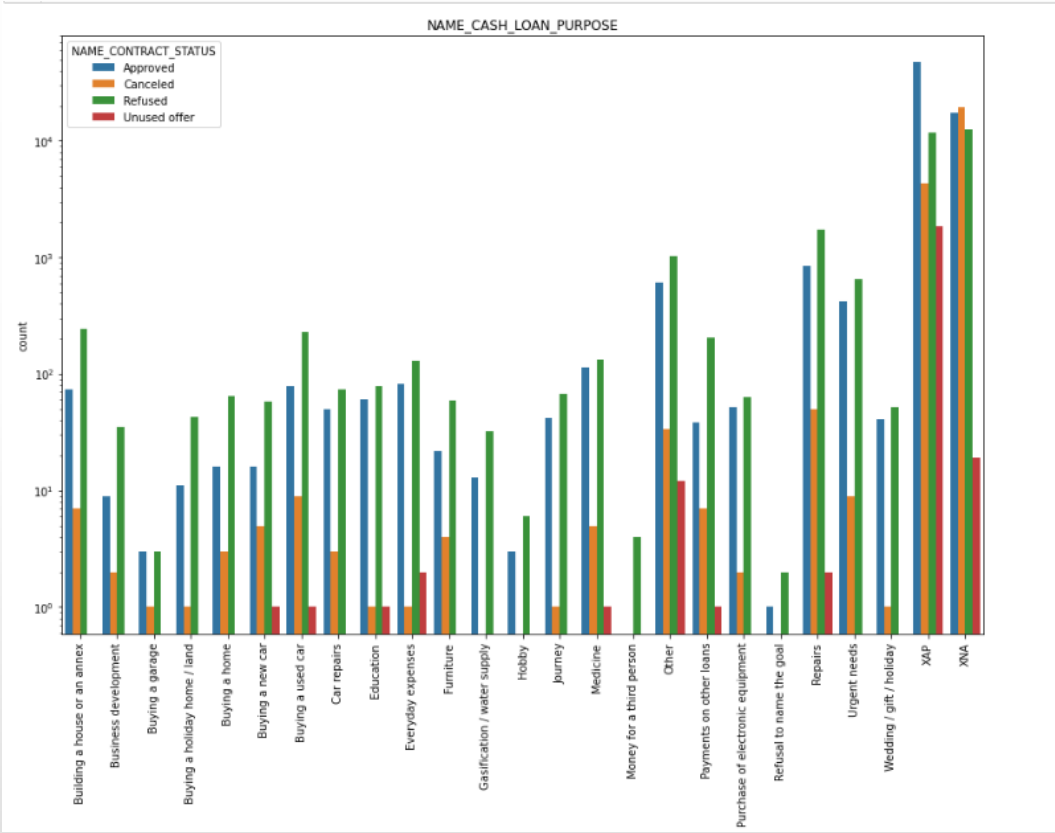
1) Count plot for the purpose of the loan based on the contract status for R0 df (Repayers)



- There are high number of missing values in loan purpose section. Maximum applicants have refused the loan or the loan is cancelled by the bank for loan purpose section 'Repairs' and 'Others'.

- **Final Analysis of Merged DF**

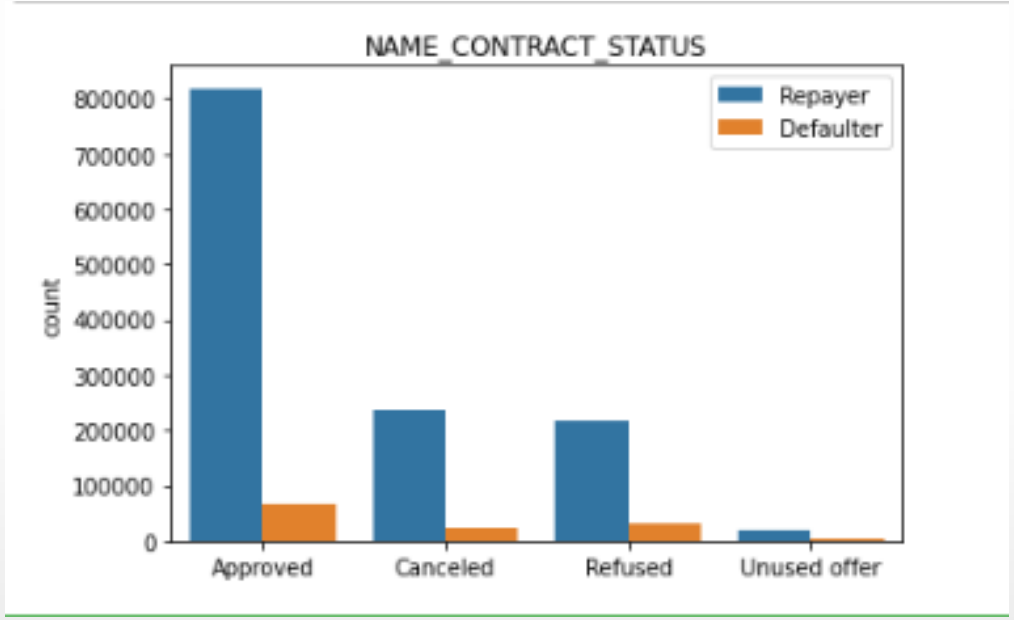
2) Count plot for the purpose of the loan based on the contract status for D1 df (Defaulter)



- The loan purpose section 'Repairs' and 'Others' are considered as high risk by the bank.

- **Final Analysis of Merged DF**

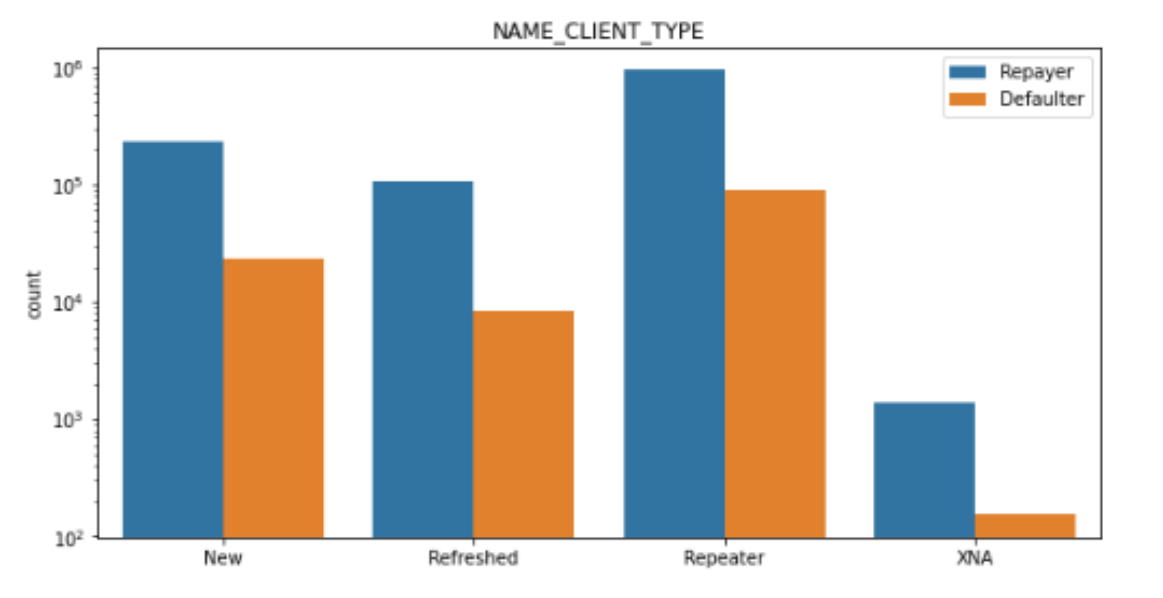
3) Count plot for 'NAME_CONTRACT_STATUS' based on 'TARGET' column



- 90% loan applications of the clients are canceled which are Repayers. There are 88% loan applications of the clients that are refused. These clients are likely to pay the loan.
- The reason for cancellation or refusal of the loan by the client should be noted for further analysis. Revising the interest rate of the loan might help these clients.

- **Final Analysis of Merged DF**

4) Count plot for 'NAME_CLIENT_TYPE' and 'TARGET'



- Most of the clients are from repeater client type. 90% of the loan applications from repeater client type are canceled in earlier case. These clients might be Repayers.
- 88% of the loan applications from new type clients are refused in previous case. These clients are likely to pay the loan.

Conclusion

- 90% of the previously cancelled clients are not defaulters. Recording the reason for cancellation or rejection of the loan and revising the interest rate of the loan might get help the bank to increase the business opportunity in future.
- The loan taken for the purpose of Repairs have high default rate. Maximum number of loan applications are rejected by the bank or canceled by the client in previous loan application for the purpose Repairs and Others. Therefore the purpose Repairs should be taken as a risk by the bank. Either they are rejected or offer very high interest rate is the best option
- The bank must target the clients with income type Student, Businessman, Pensioner for profitable business as they have lower default rate.
- Get as much as clients from housing type ‘Office Apartment’ as they are having less number of unsuccessful payments.