Presentation on Credit EDA Case Study

Problem Statement:

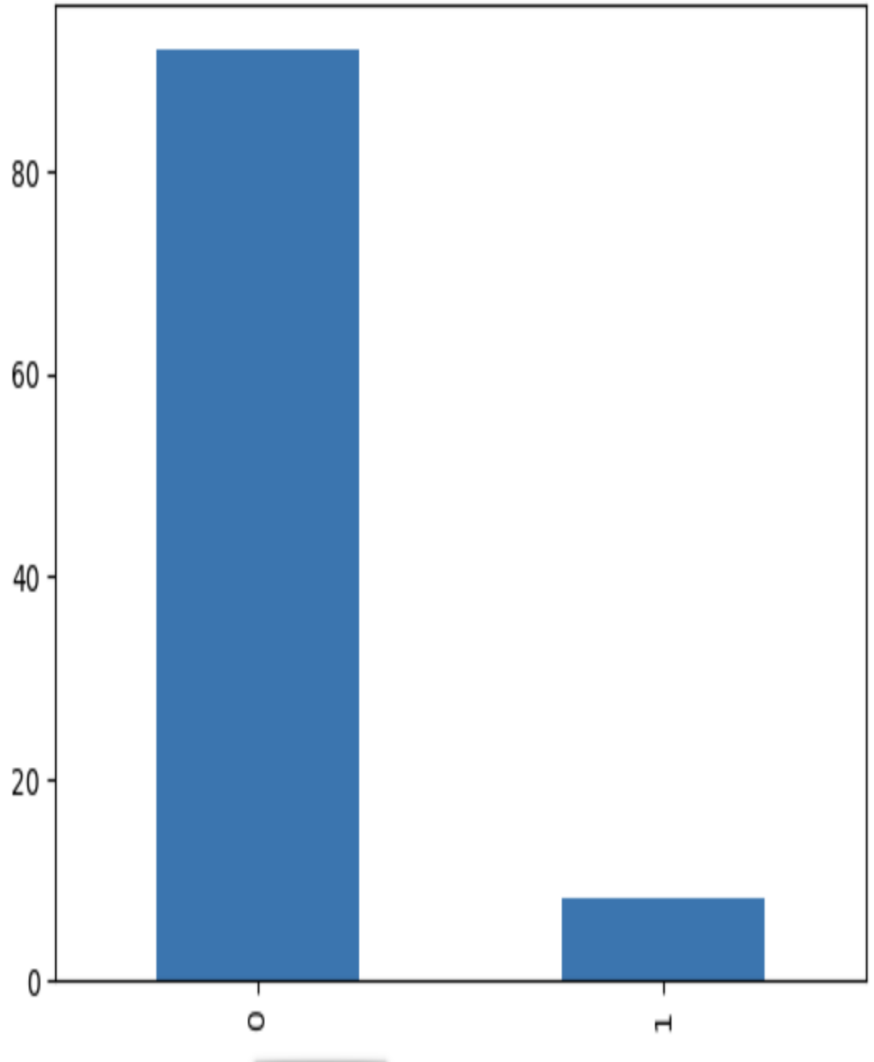
To Identify Patterns: which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate

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.

Analysis Steps:

- Data understanding and sourcing
- Check for Data quality issues and Binning
- Check for data imbalance and univariate, segmented univariate & Bivariate analysis, correlation
- Merging of application data with previous application data
- Data analysis by univariate, segmented univariate, Bivariate analysis and correlation
- Recommendations and Risks

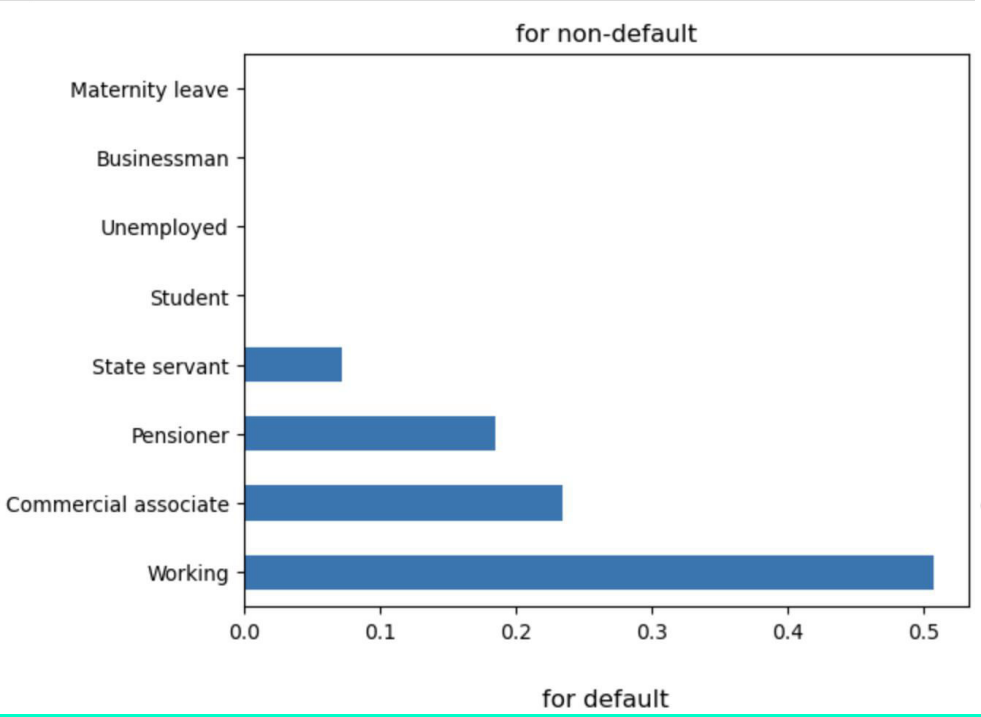


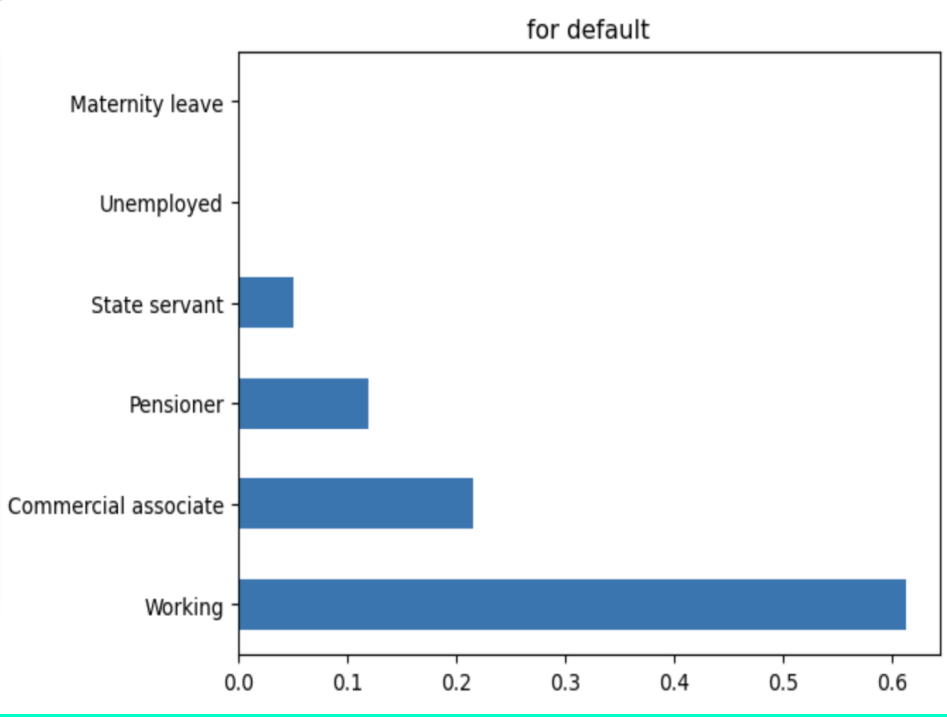
From the Data analysis on Target column we came to know that the data is heavily imbalanced

Dividing Target data in two subsets Target_0 and Target_1.

Perform univariate analysis for categorical variables for both 0 and 1

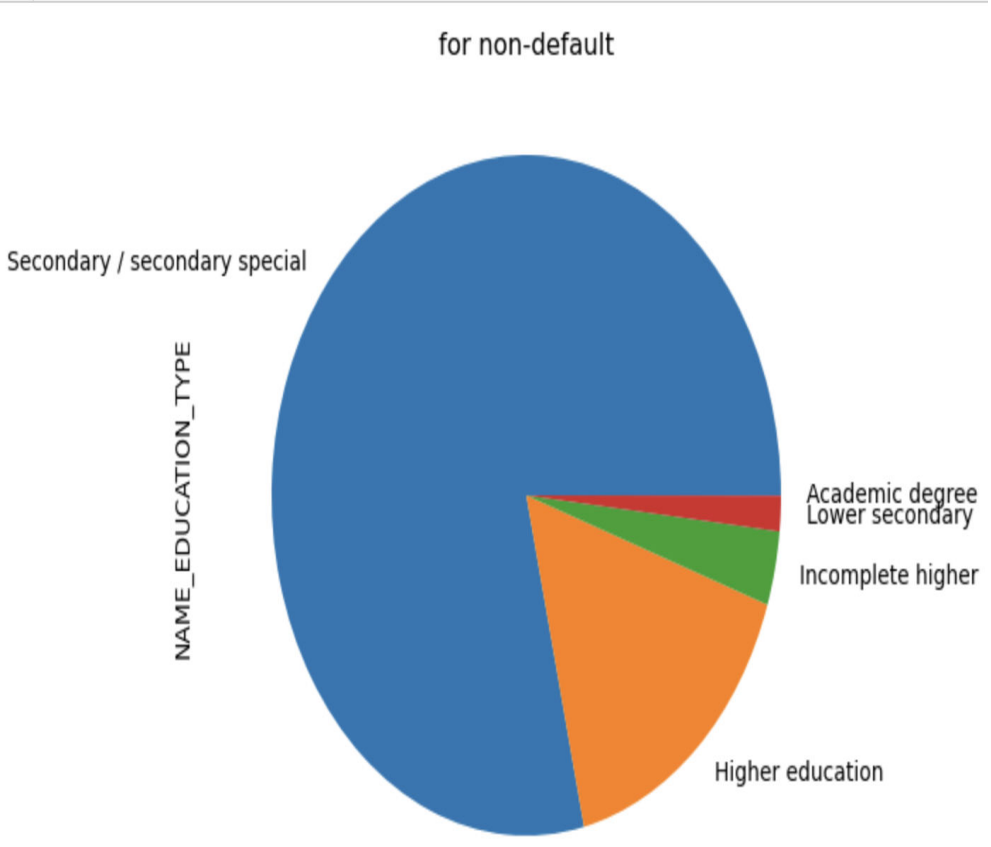
Univariate Analysis of Continuous Data in Application data

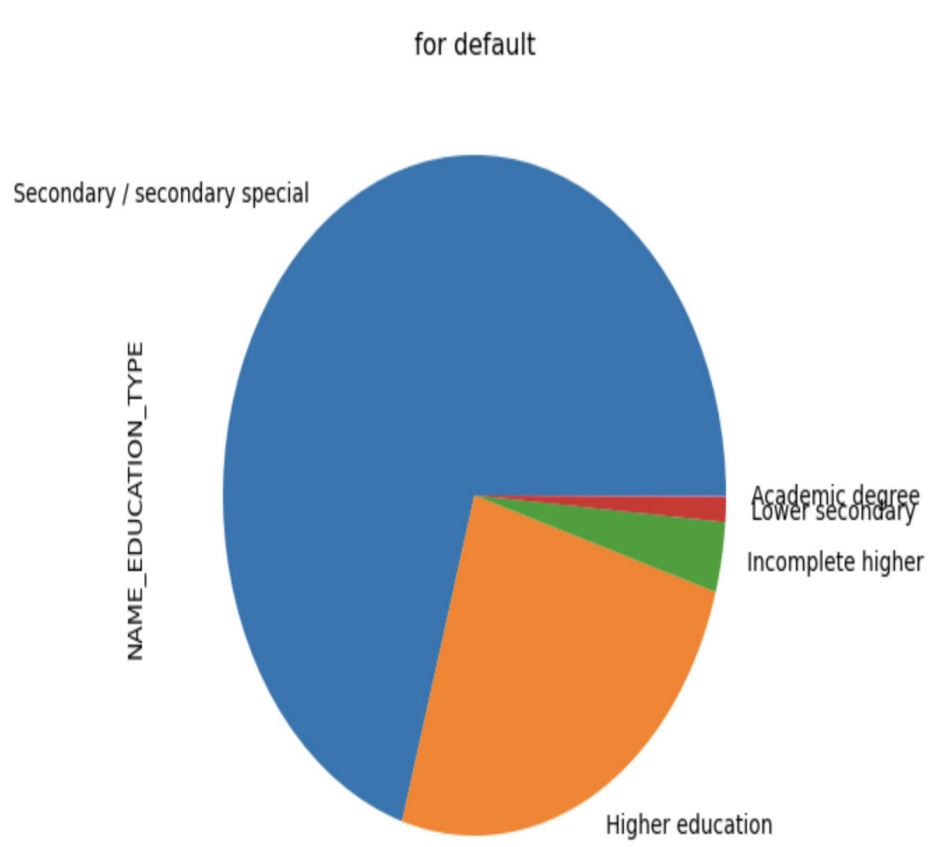




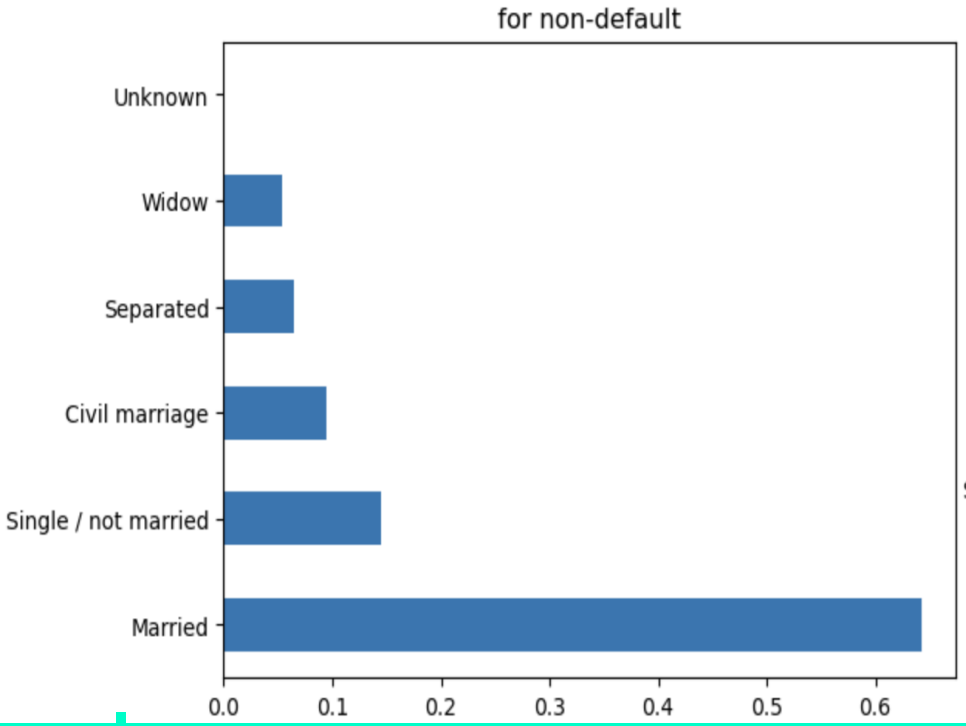
Pensioner of non-default case are high in number compared to Pensioner of default case., it seems there exists both loss and profit due to Pension people to the Bank.

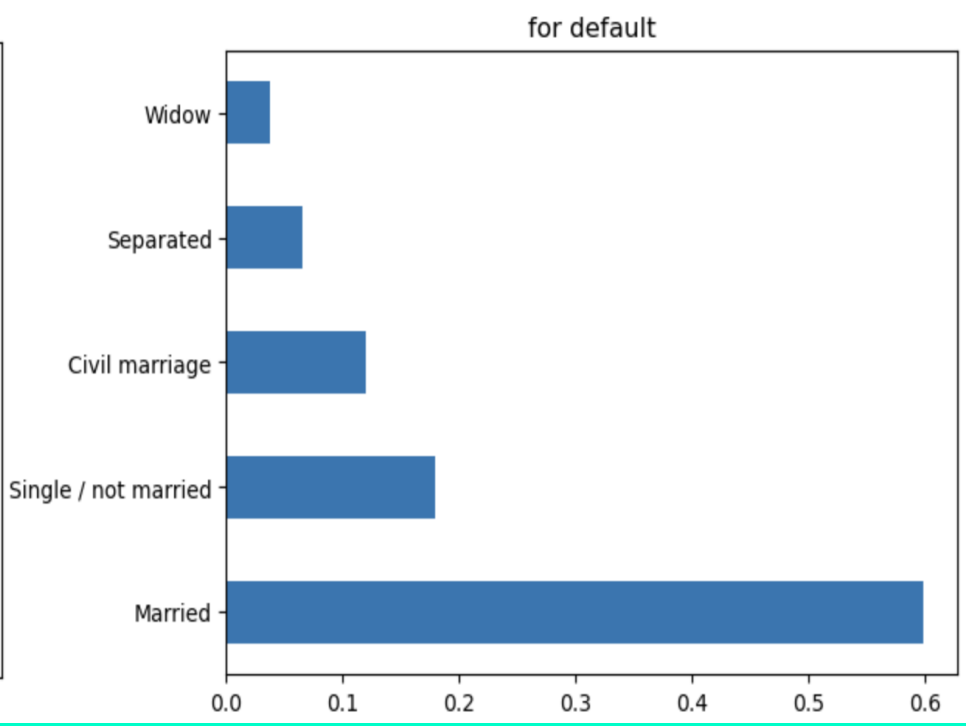
It also shows that majority of defaulters income type is working,and at the same time there is





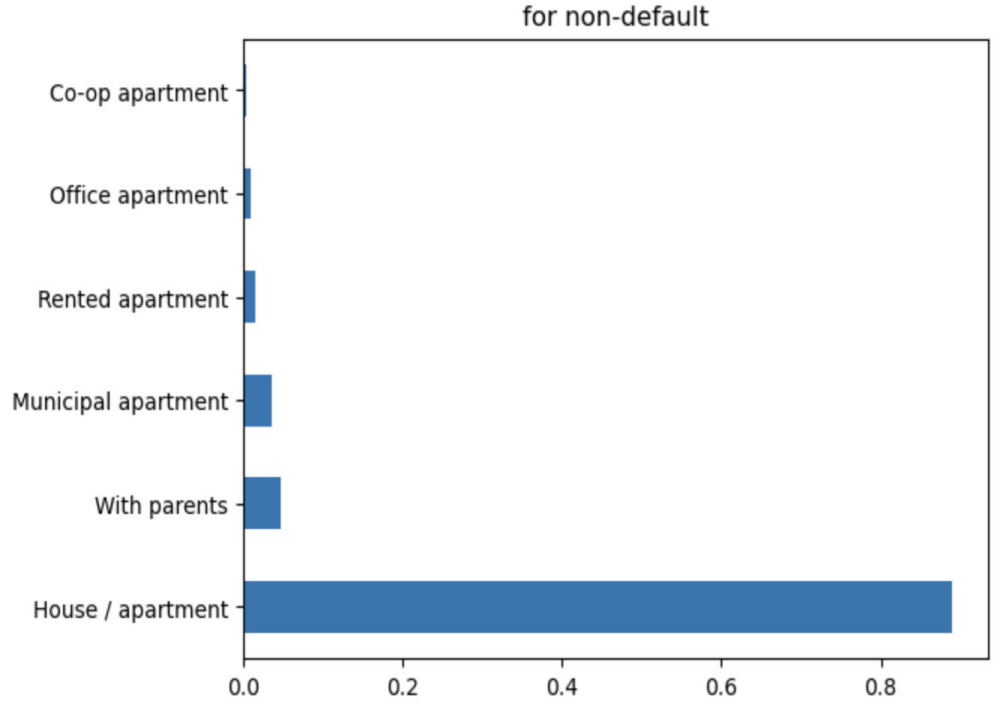
From the plots, we can conclude that secondary/special educated people are applying loans in high in number.
Academic degree educated people are applying loan in least count. for both target= 0 and 1

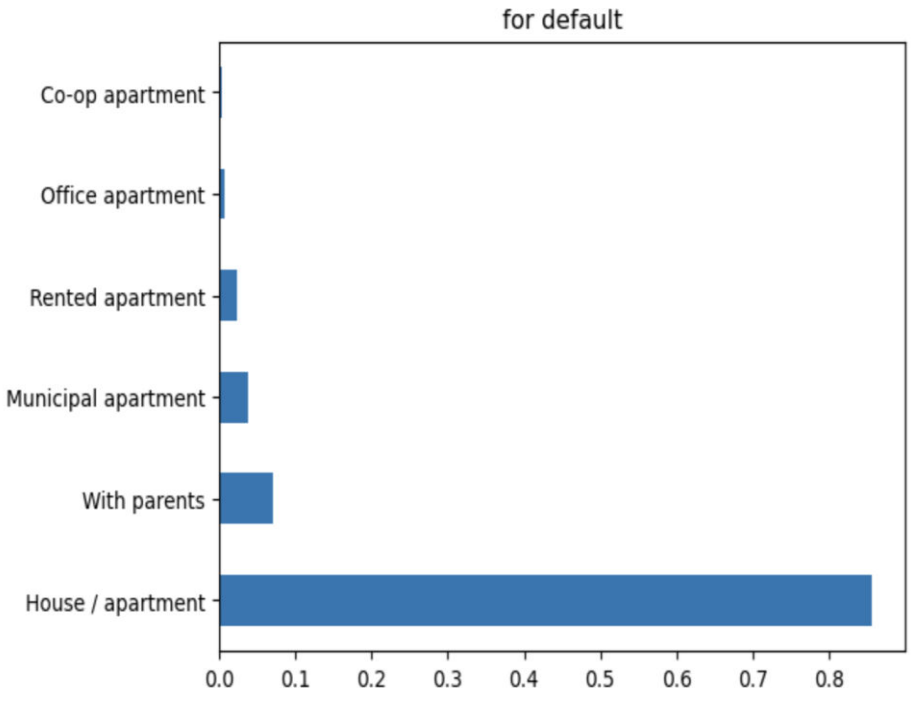




We can say more married people tend to take more Loan as compared to other categories

and being married is not impacting default and not defaulting





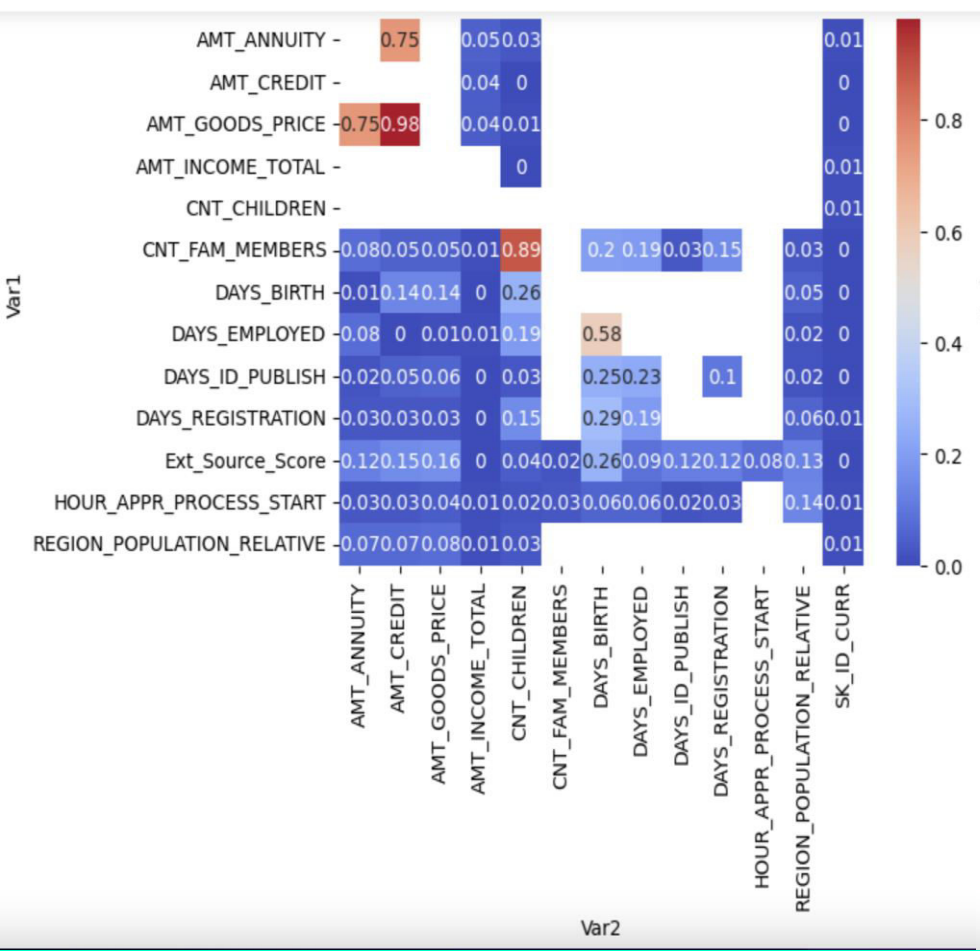
From graph we can conclude that there exists people who have own house, lies in both default and non default.

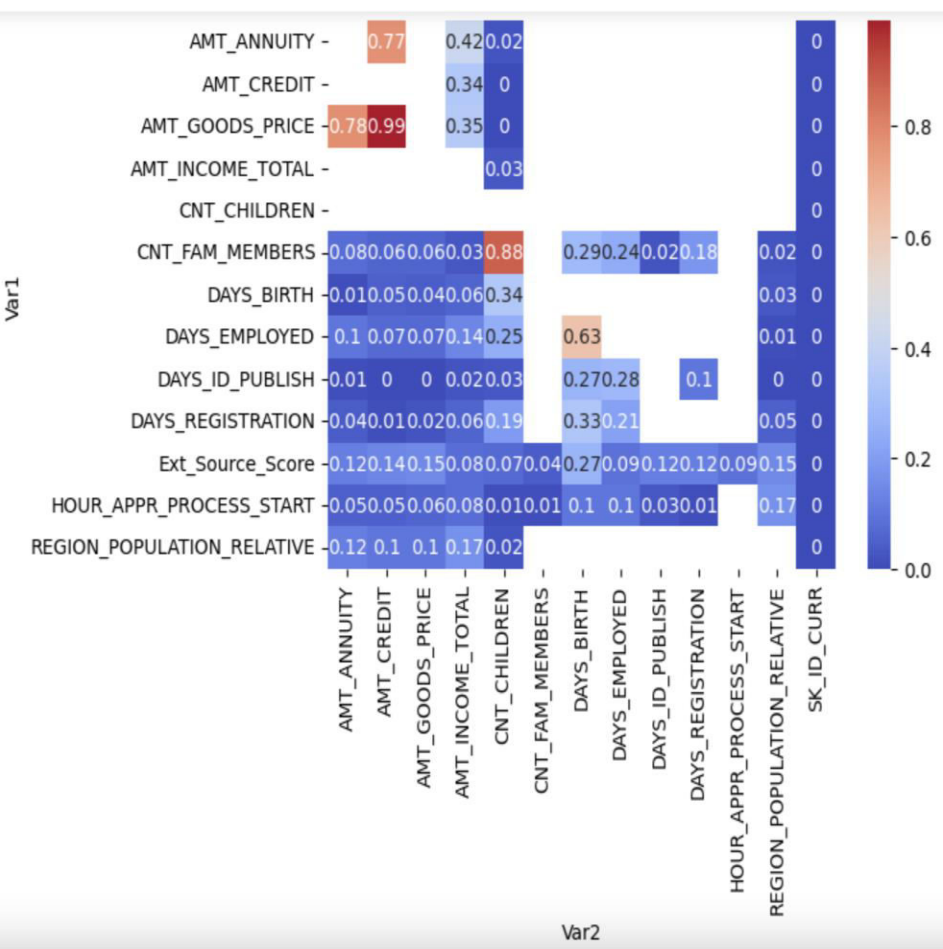
Correlation

Both the Correlation looks same

	Var1	Var2	Coorelation
73	AMT_GOODS_PRICE	AMT_CREDIT	0.98
155	CNT_FAM_MEMBERS	CNT_CHILDREN	0.89
59	AMT_ANNUITY	AMT_CREDIT	0.75
74	AMT_GOODS_PRICE	AMT_ANNUITY	0.75
119	DAYS_EMPLOYED	DAYS_BIRTH	0.58
133	DAYS_REGISTRATION	DAYS_BIRTH	0.29
189	Ext_Source_Score	DAYS_BIRTH	0.26
99	DAYS_BIRTH	CNT_CHILDREN	0.26
147	DAYS_ID_PUBLISH	DAYS_BIRTH	0.25
148	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.23

	Var1	Var2	Coorelation
73	AMT_GOODS_PRICE	AMT_CREDIT	0.99
155	CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
74	AMT_GOODS_PRICE	AMT_ANNUITY	0.78
59	AMT_ANNUITY	AMT_CREDIT	0.77
119	DAYS_EMPLOYED	DAYS_BIRTH	0.63
58	AMT_ANNUITY	AMT_INCOME_TOTAL	0.42
72	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.35
44	AMT_CREDIT	AMT_INCOME_TOTAL	0.34
99	DAYS_BIRTH	CNT_CHILDREN	0.34
133	DAYS_REGISTRATION	DAYS_BIRTH	0.33





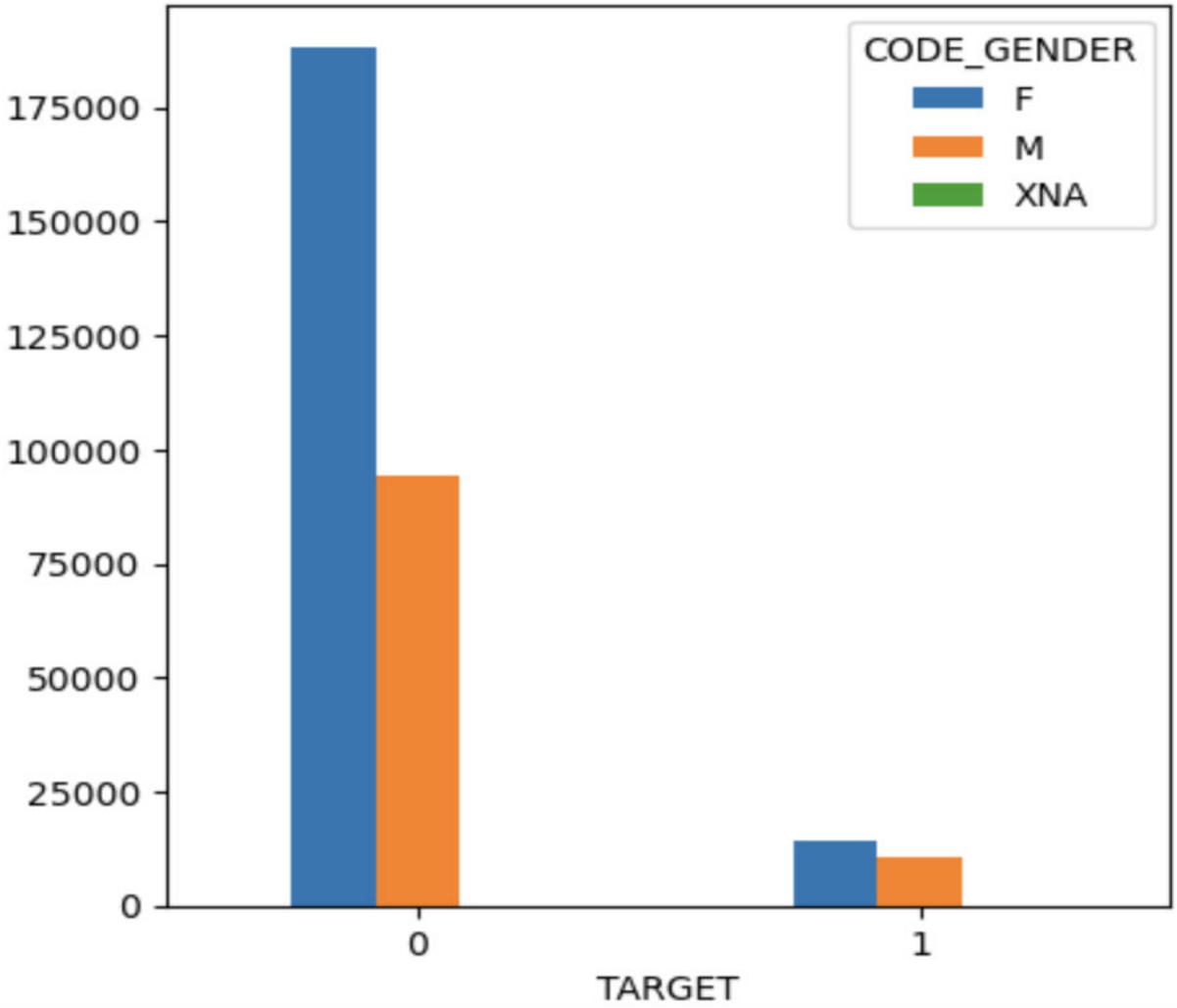
Top Co Relation For Defaulters & Non Defaulters

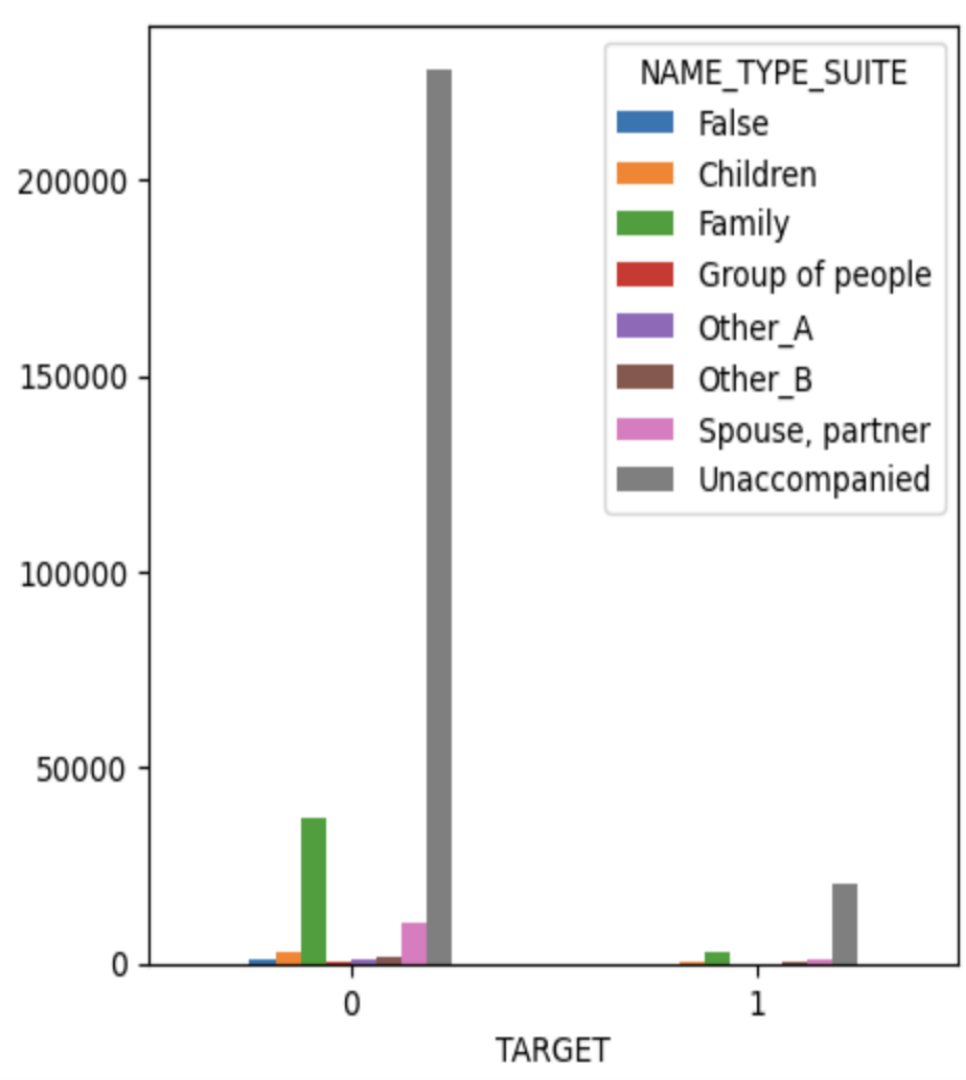
Bivariate Analysis

On Target data Frames (0 & 1)

Females take more loans

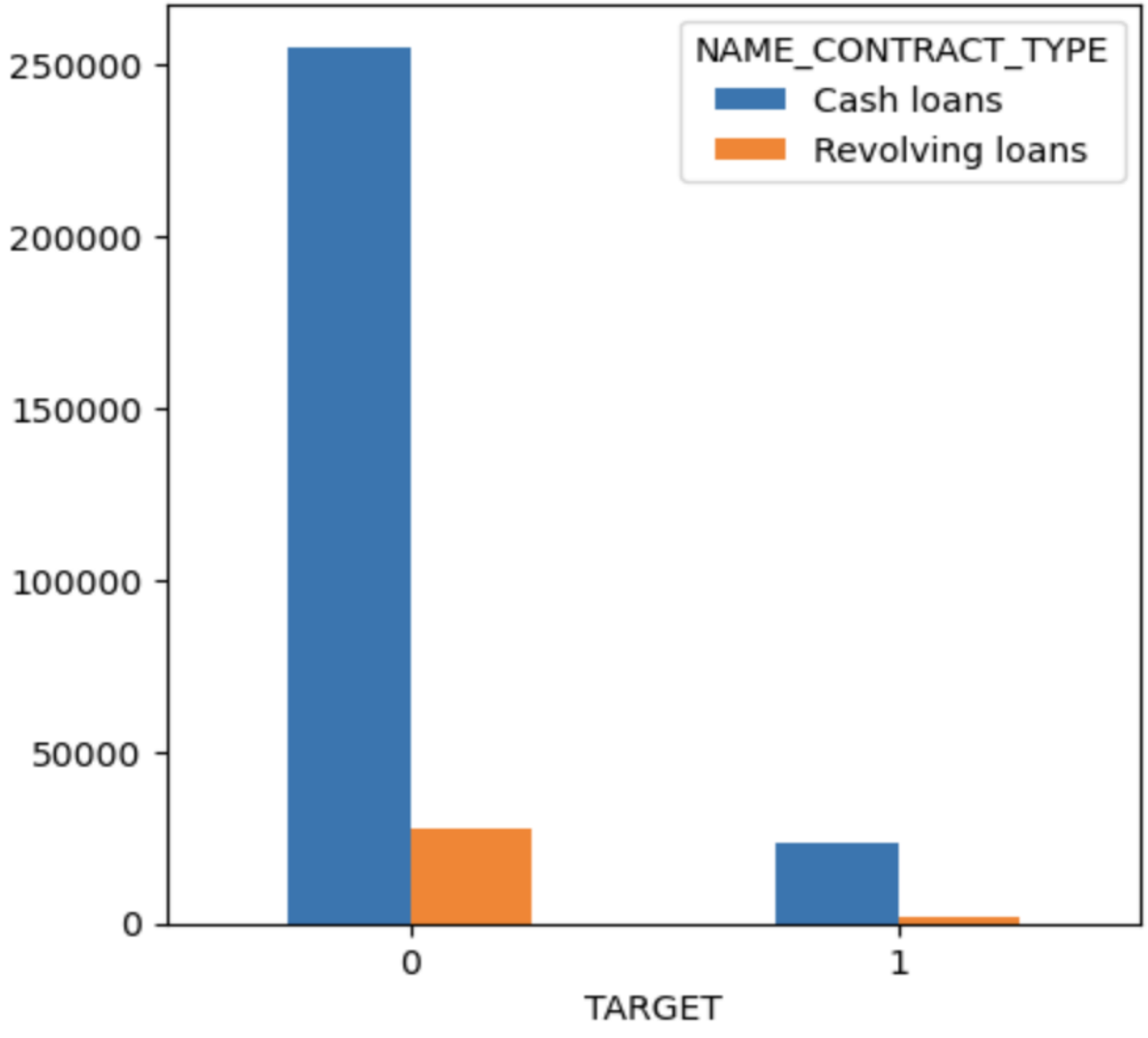
XNA in Code_Gender is consists of 4 rows and is not known, so we can ignore

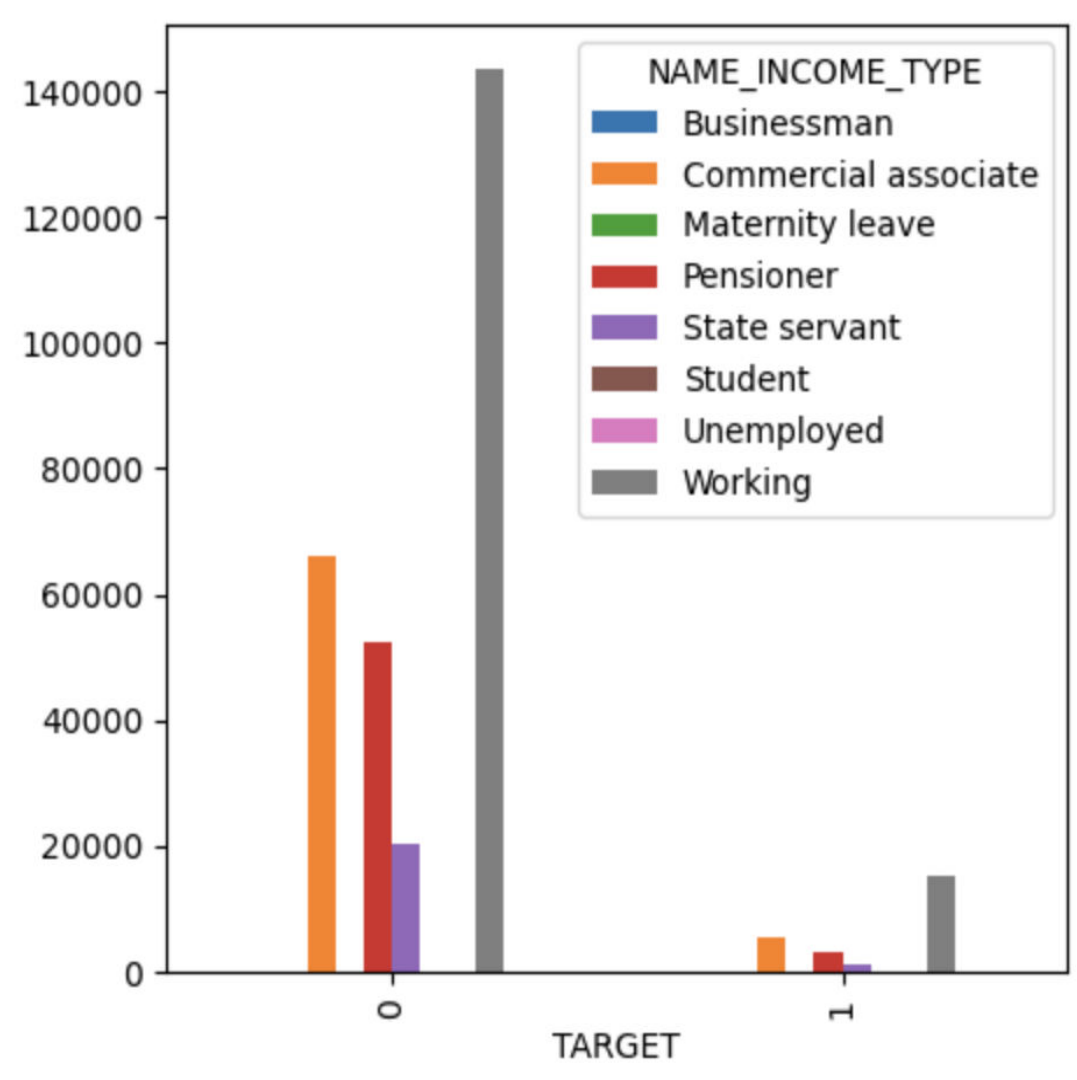




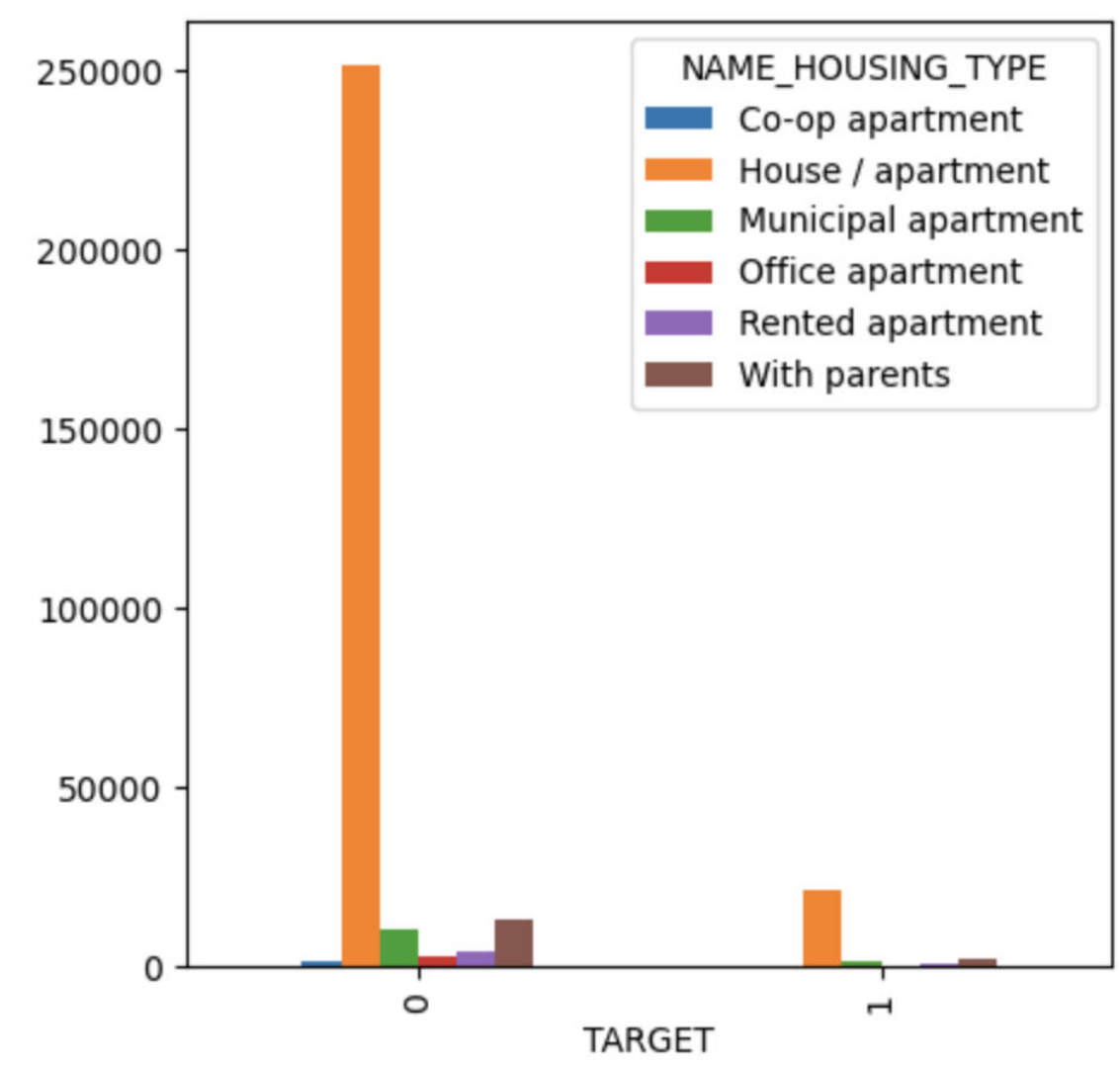
Most of the people come alone when taking a loan

Cash loans have the highest count of Approved loans



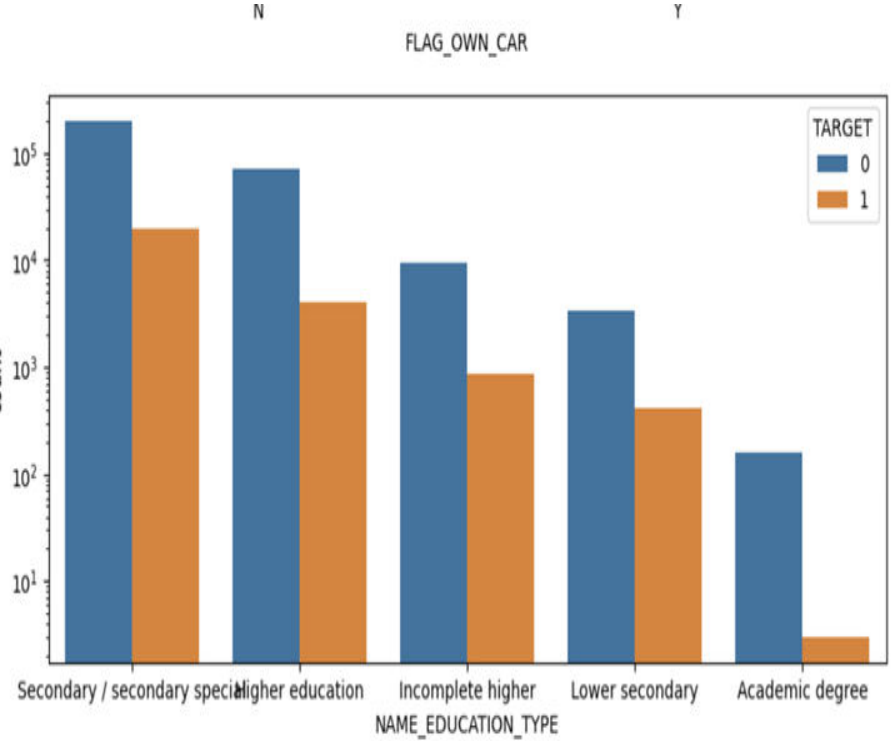


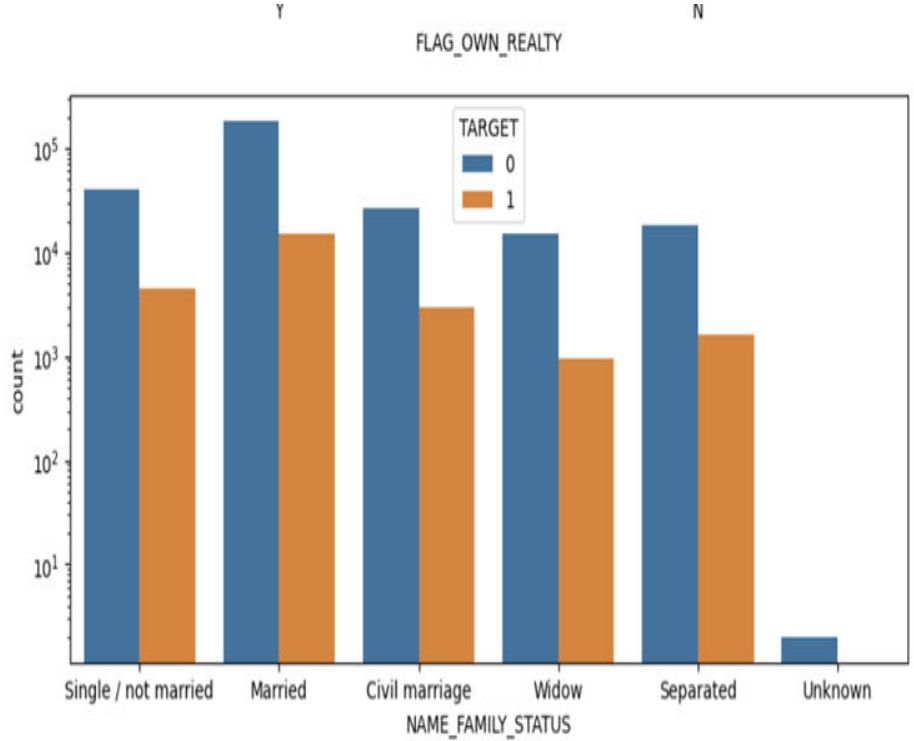
Highest number of approvals for working applicants



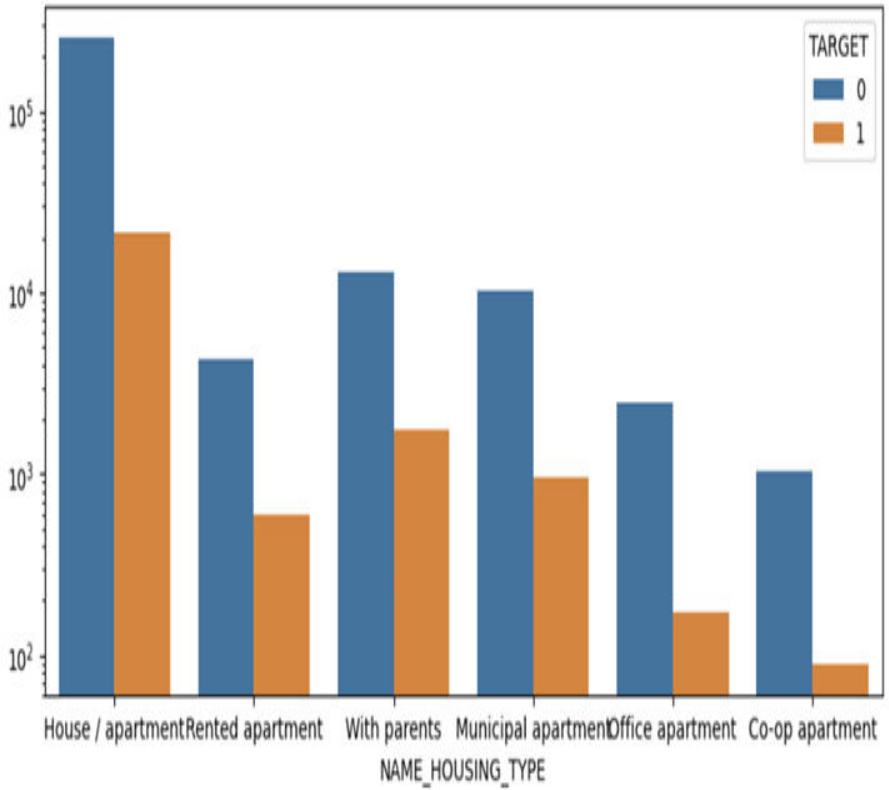
People having house/appartment tend to take more loans

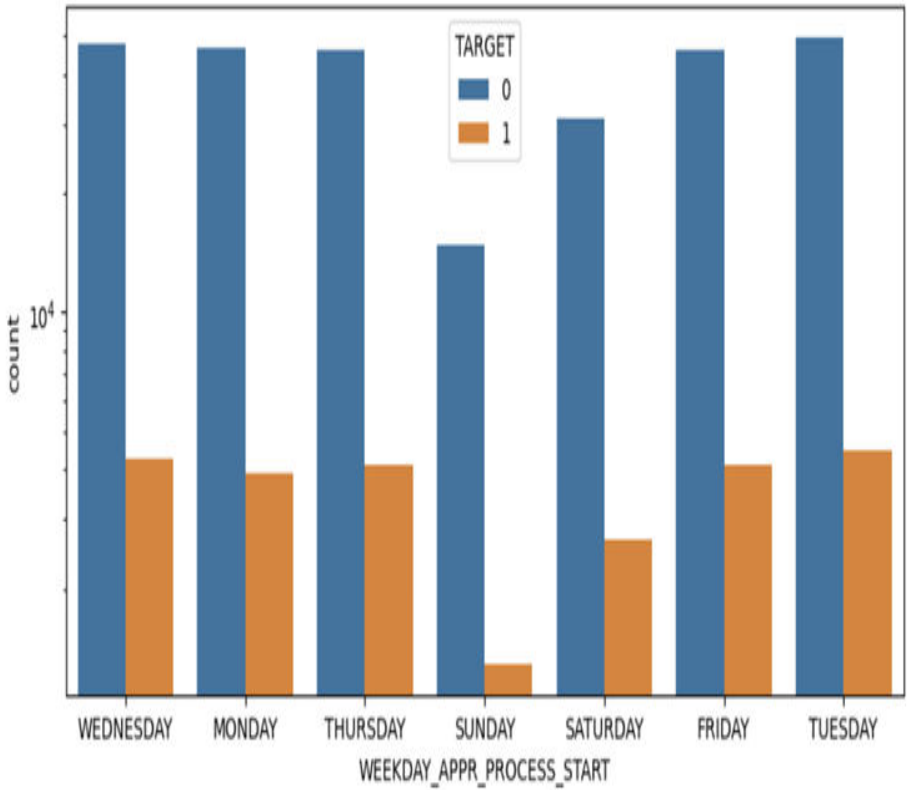
Categorical Analysis across Traget 0 & 1





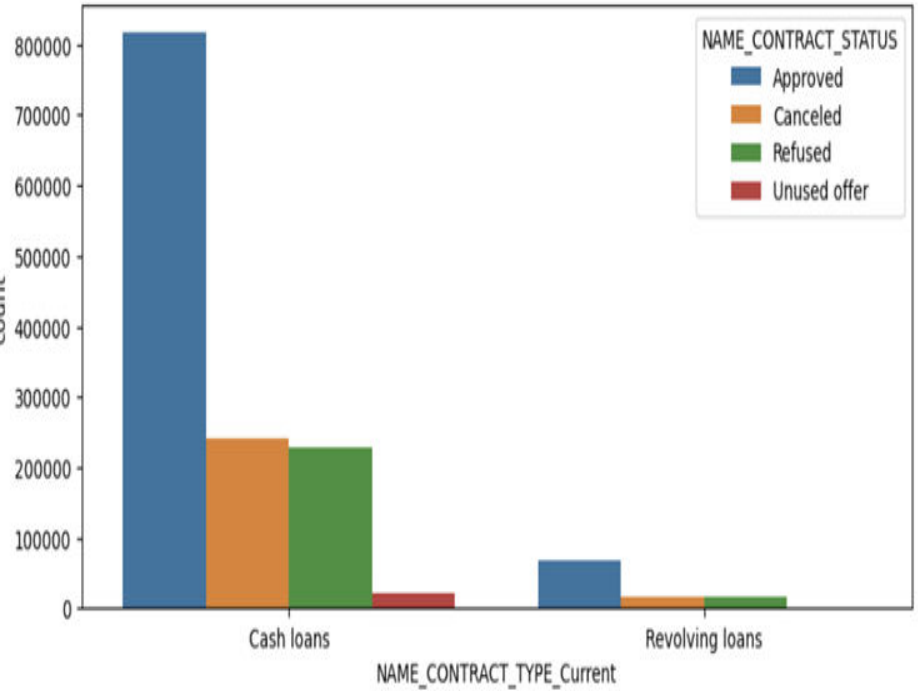
Seconday/Special People are applying for loan in high number and Maried people use to take more loans

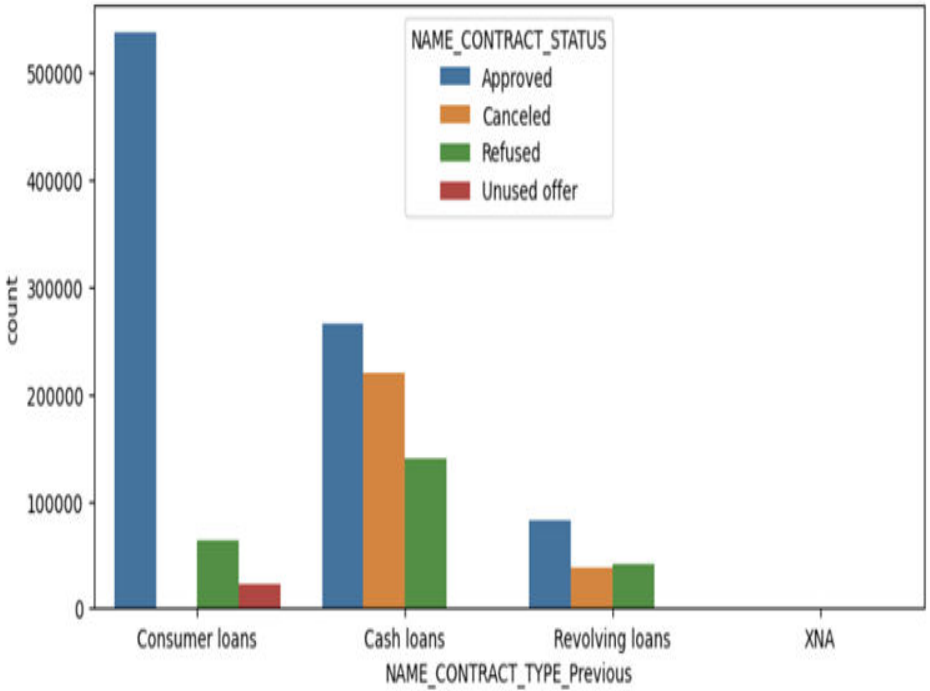




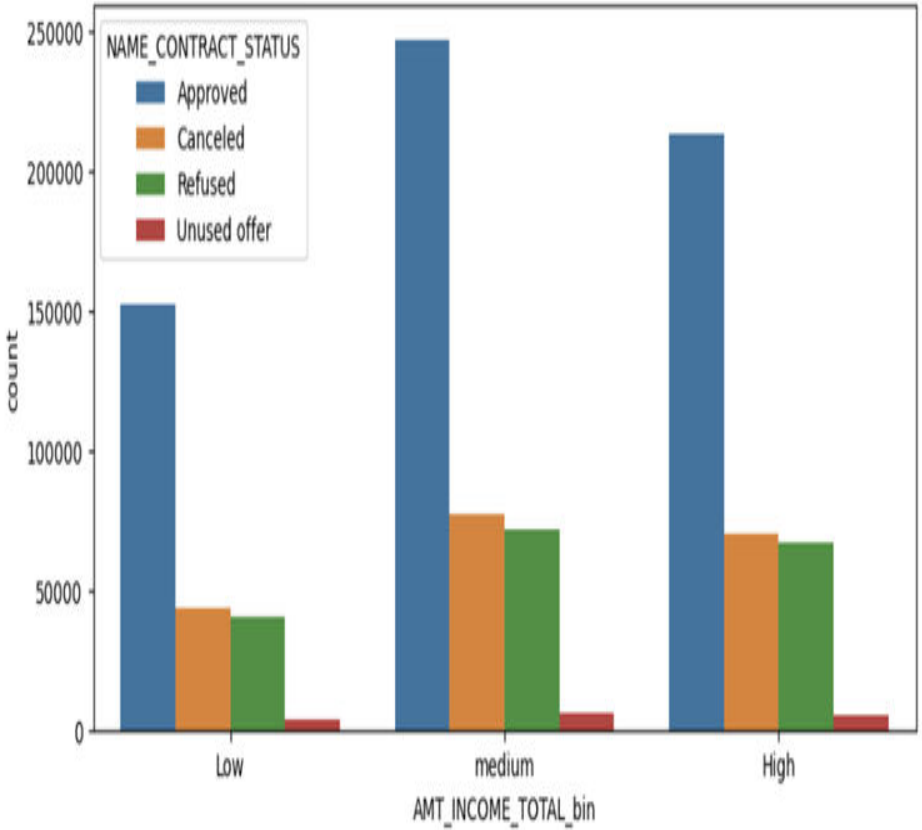
People with House/apartment tends to take more loans, Saturday and sunday are less busy for bank in terms of loan applications

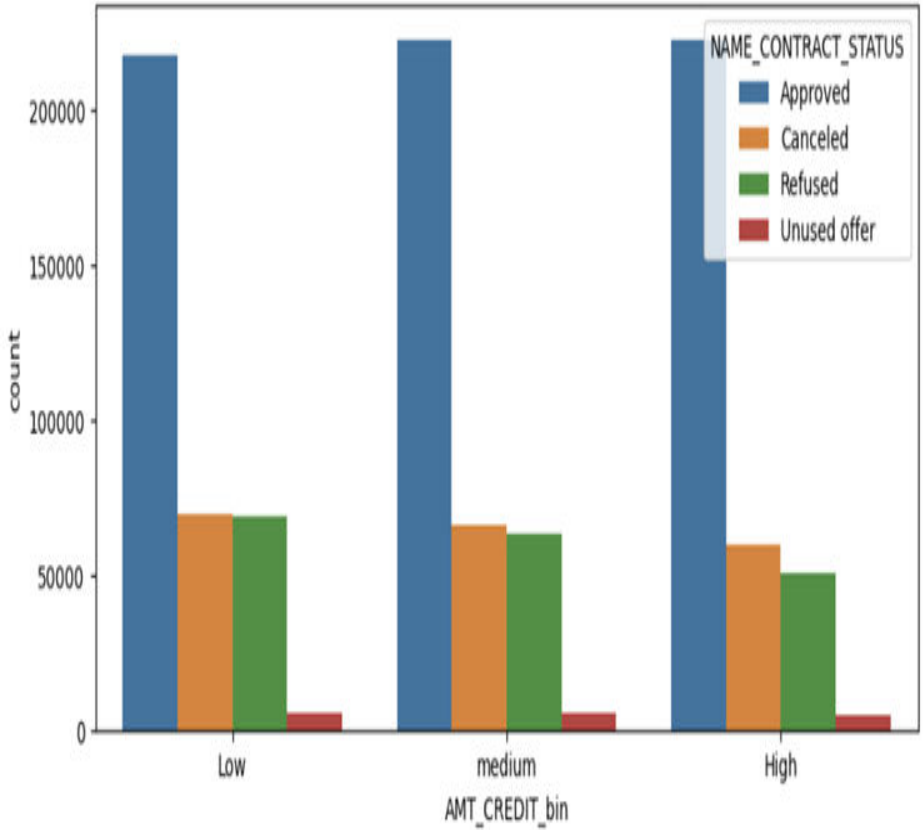
Categorical Analysis on application data(_Current) & Previous application data(_Previous)





Number of consumer loans were highest previously and now highest number is Cash loans.

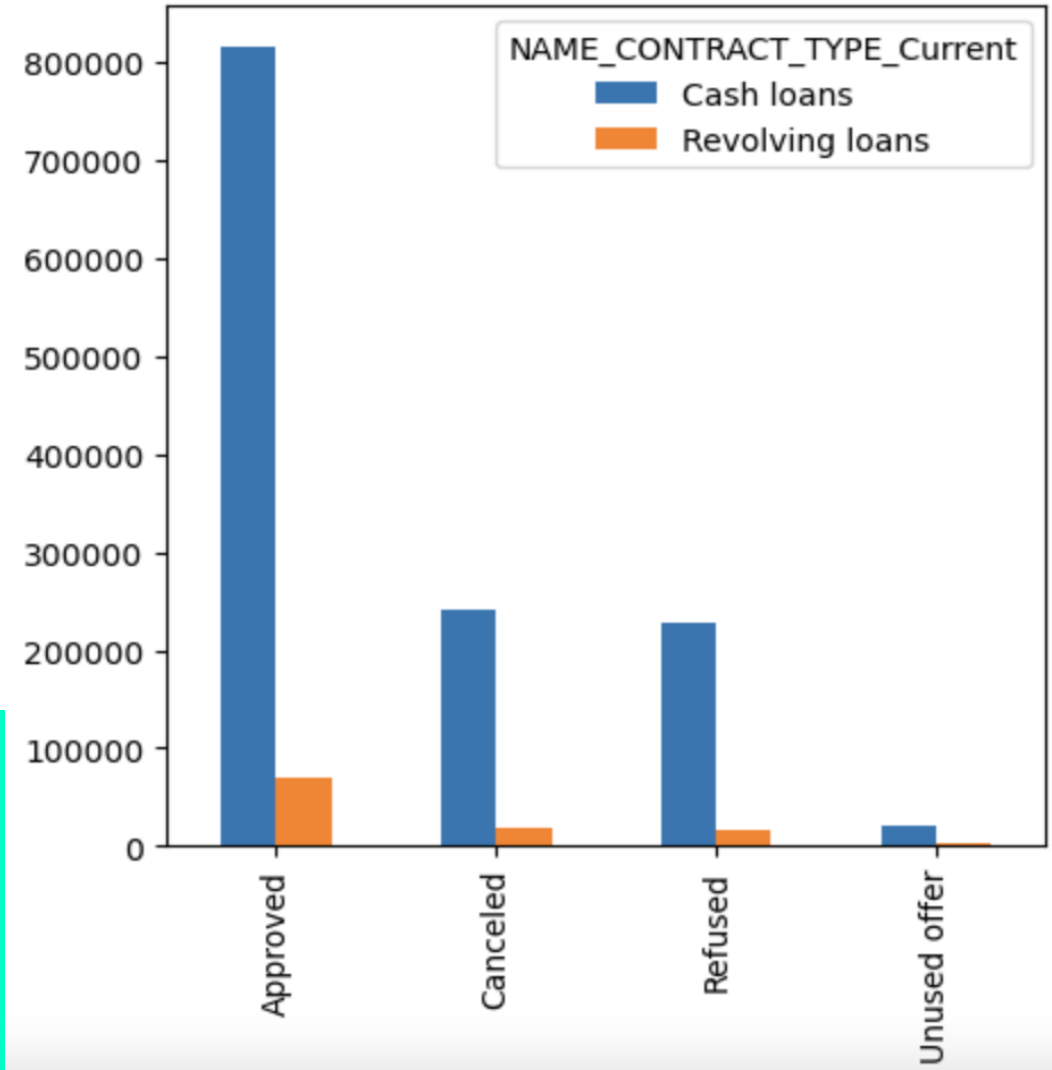




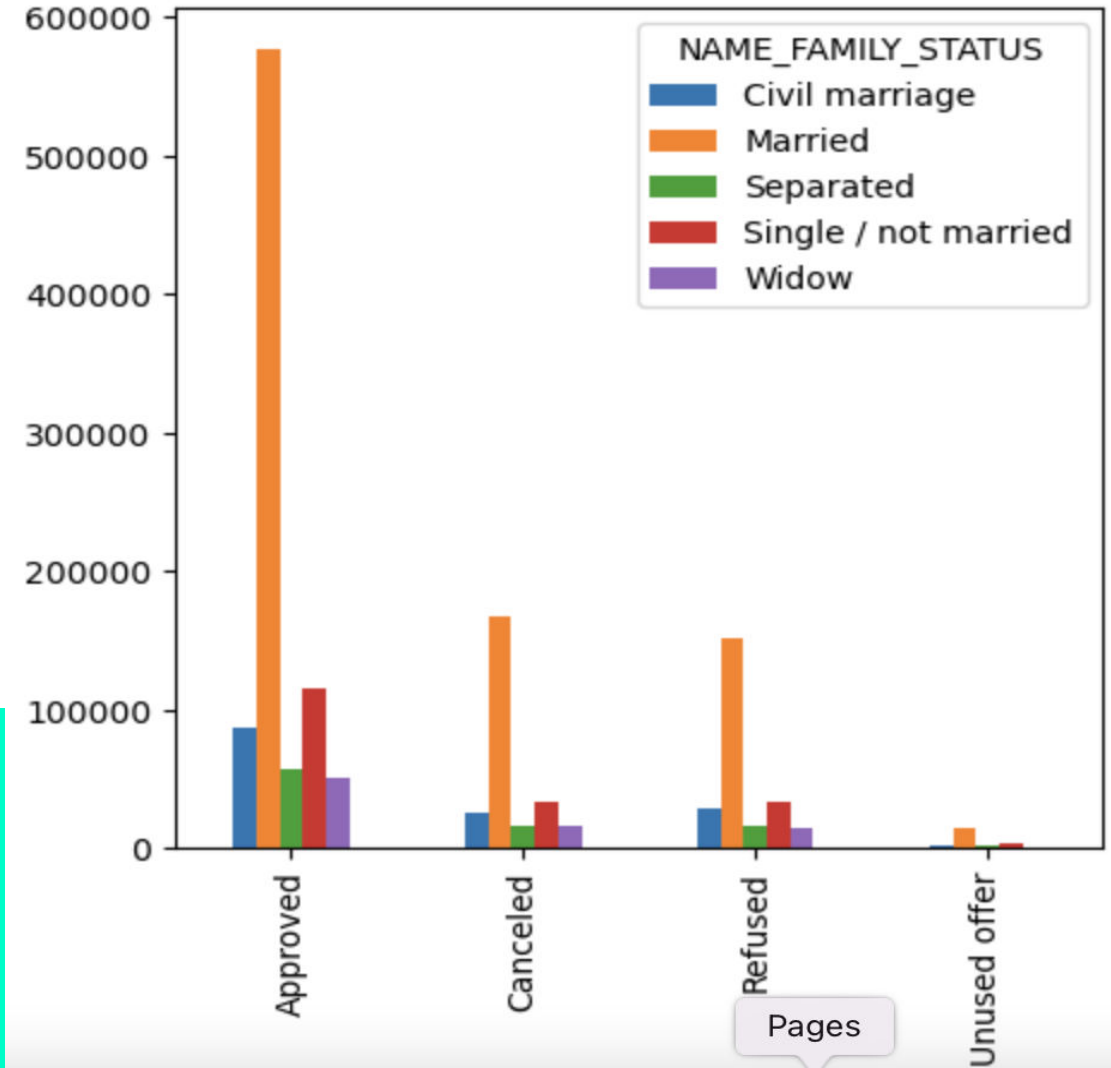
For Medium AMT_INCOME_TOTAL_bin the approval is highest

Bivariate Analysis

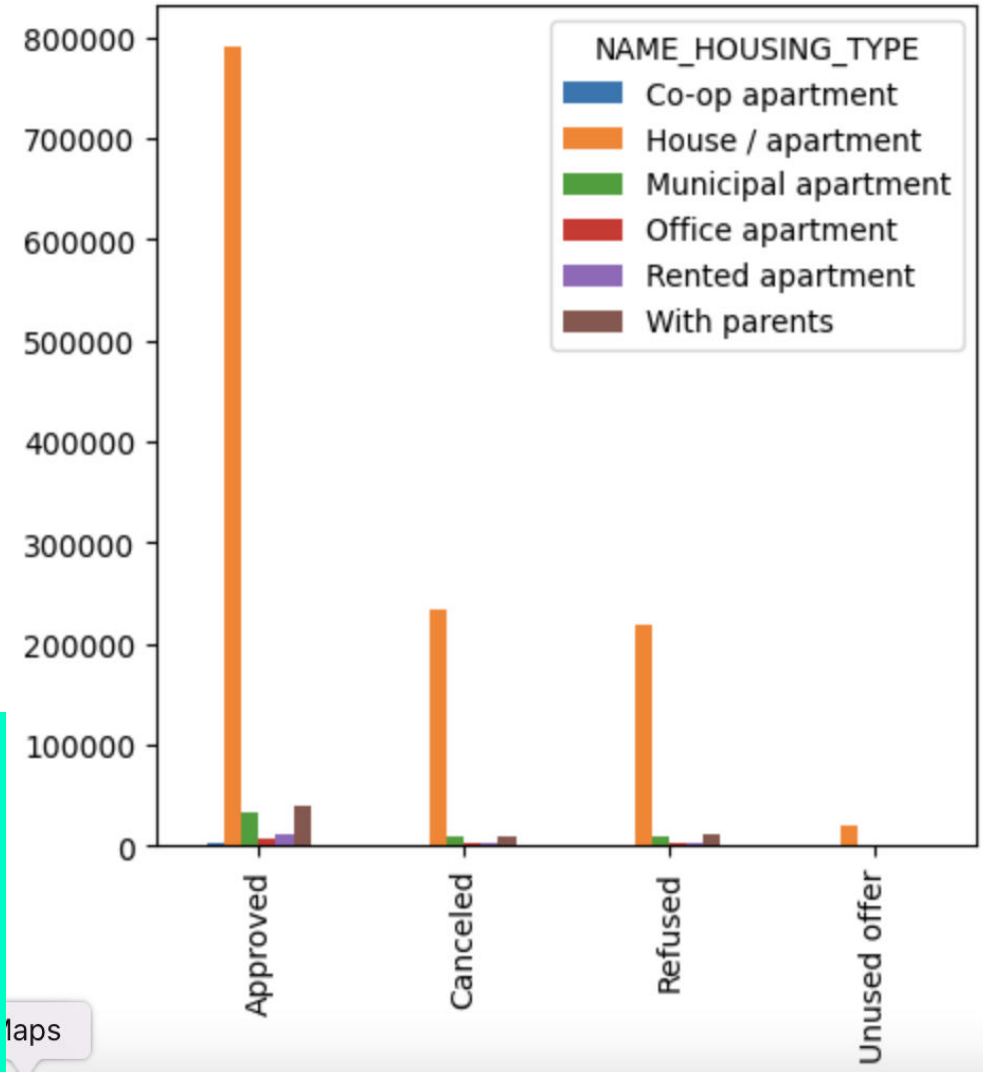
Cash loans have the highest count of Approved loans



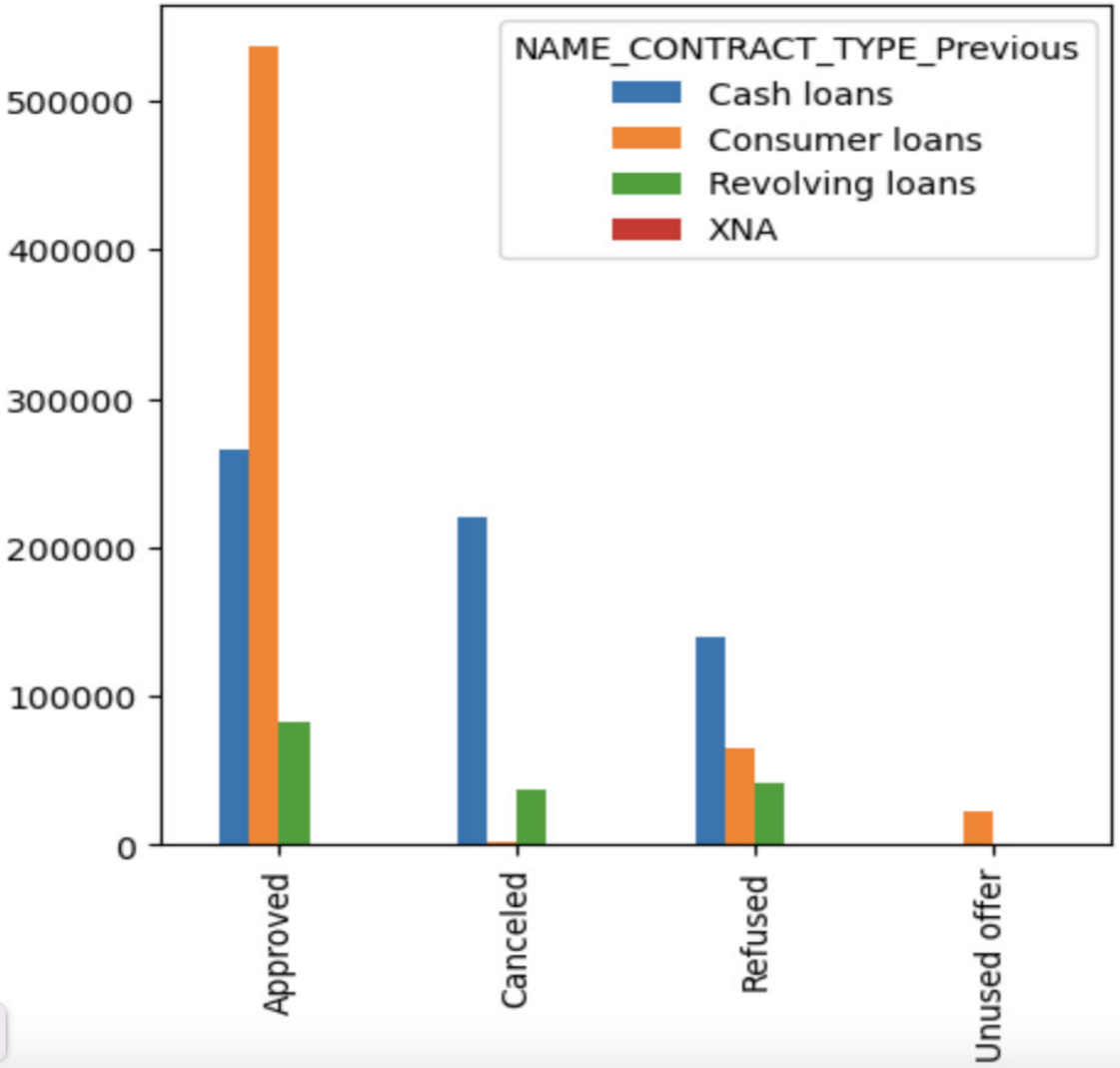
Highest number of approvals for Married applicant



Highest number of approvals for House/apartment owner



Highest number of approvals for Consumer Loans



Final Words

Target/focused variable for Application dataset - **TARGET**

Target/focused variable for Previous dataset - **NAME_CONTRACT_STATUS**

Top Major variables to consider for loan prediction:

1. NAME_EDUCATION_TYPE
2. AMT_INCOME_TOTAL
3. DAYS_BIRTH
4. AMT_CREDIT
5. DAYS_EMPLOYED
6. AMT_ANNUITY
7. NAME_INCOME_TYPE
8. CODE_GENDER
9. NAME_HOUSING_TYPE

The above mentioned variables are to be considered before approving application to minimize risk of loss.