CREDIT EDA CASE STUDY

BY RAJEEV RANJAN KASHYAP

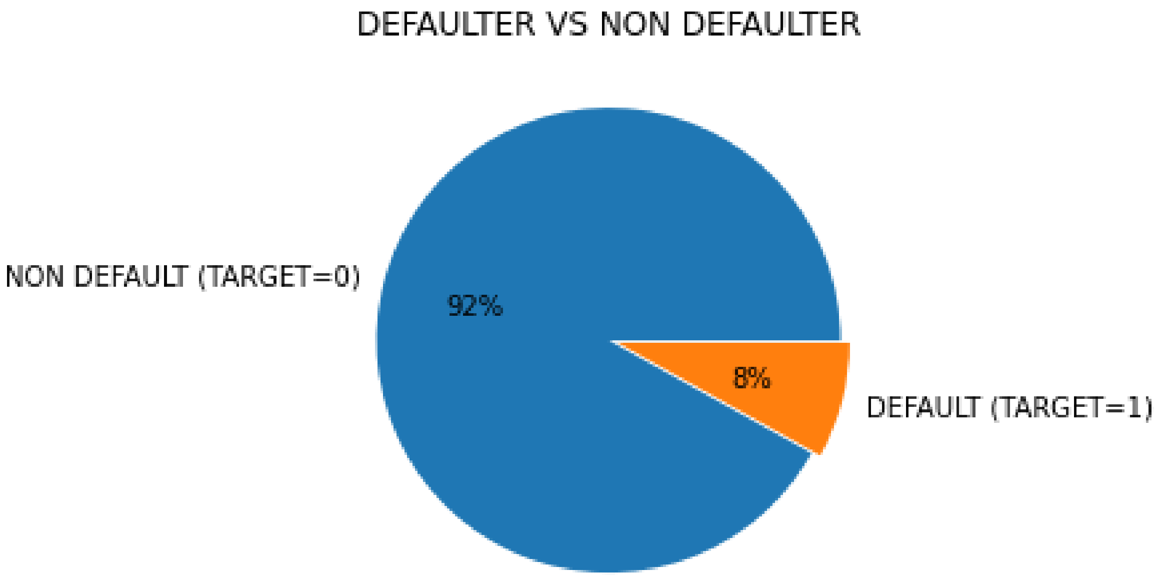
Purpose

Credit risk analysis for financial institution based on
client profile and details in dataset to avoid
risk of default and business loss

Steps

- Understanding of Data and sourcing
- Checking structure of data
- Data quality check
- Check binning
- Check for imbalance in data, Univariate and Bivariate analysis, correlation
- Recommendation

Imbalance check in Target

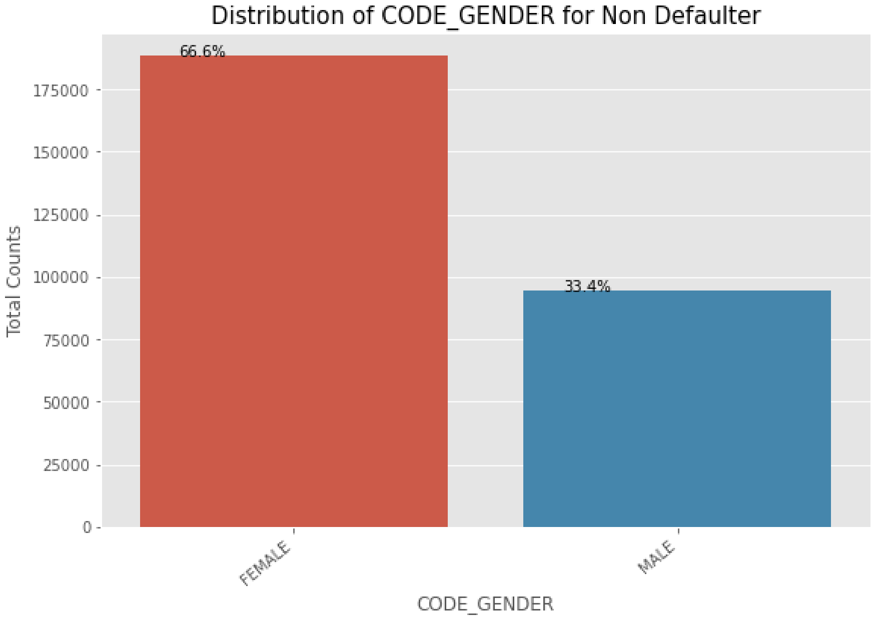


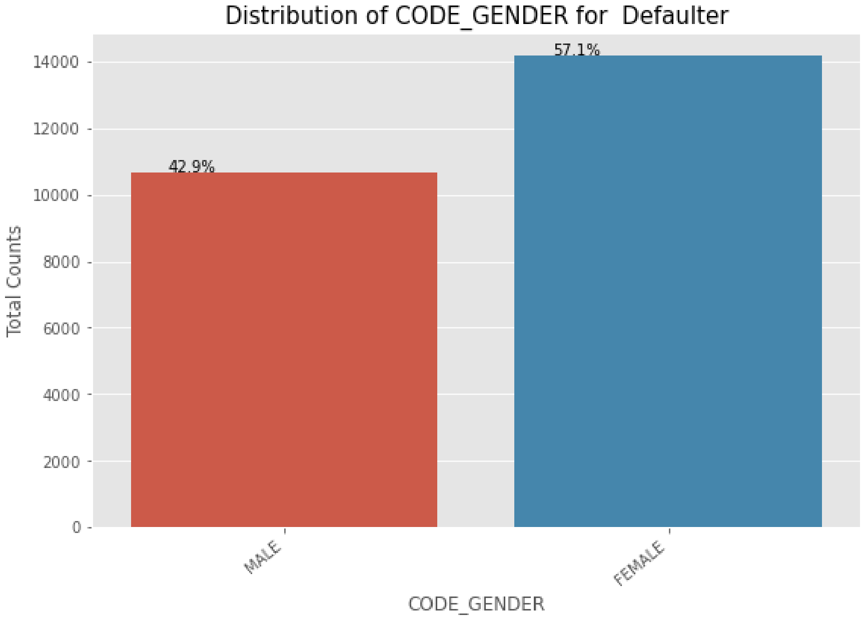
Imbalance check in Target

- From pie chart it is observed that there is following percentage of Non Defaulter and Defaulter

- Non Defaulters are much more than defaulters
- Non Defaulters-92%
- Defaulter-8%

Univariate analysis of categorical variables CODE_GENDER



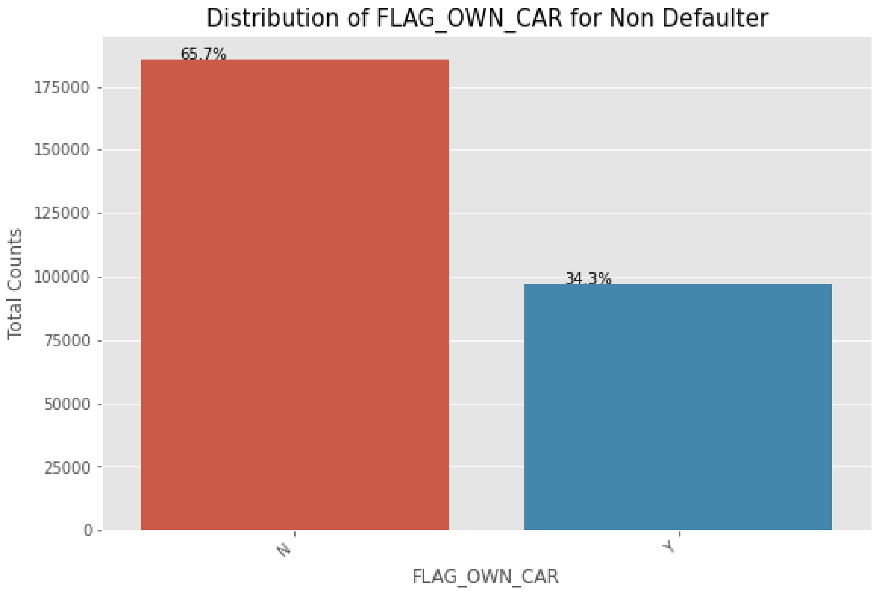


Univariate analysis of categorical variables
CODE_GENDER

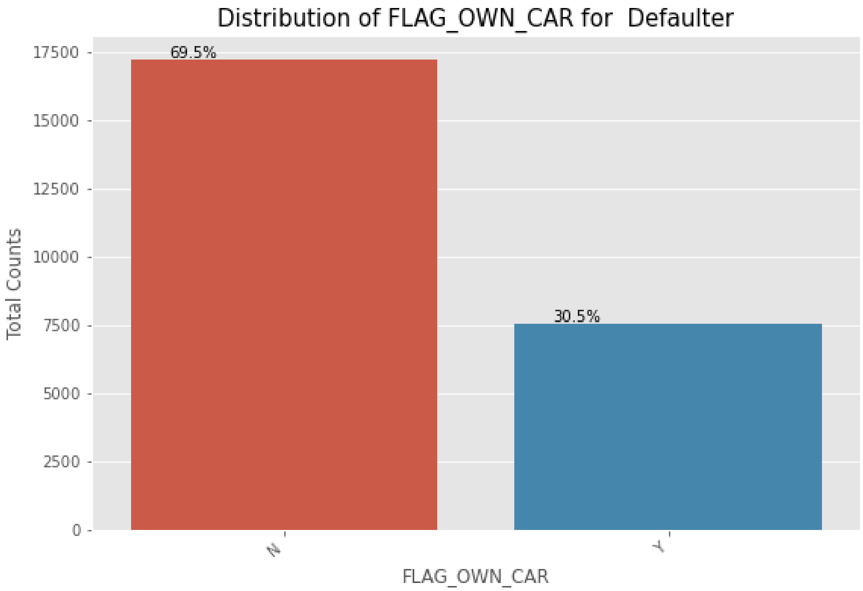
- Observations
- Female contribute 67% in Non defaulter and 57% in defaulter segment
- It implies that female are applying more for loan
- The rate of default is less in female

Univariate analysis of categorical variables

FLAG_OWN_CAR



Distribution of FLAG_OWN_CAR for Non Defaulter



Distribution of FLAG_OWN_CAR for Defaulter

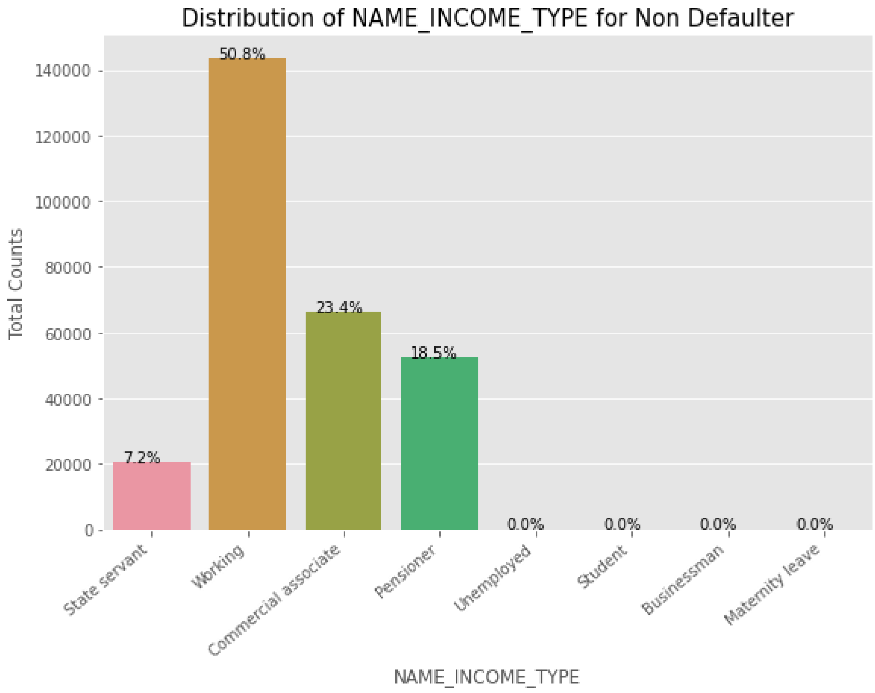
Univariate analysis of categorical variables FLAG_OWN_CAR

Observation-

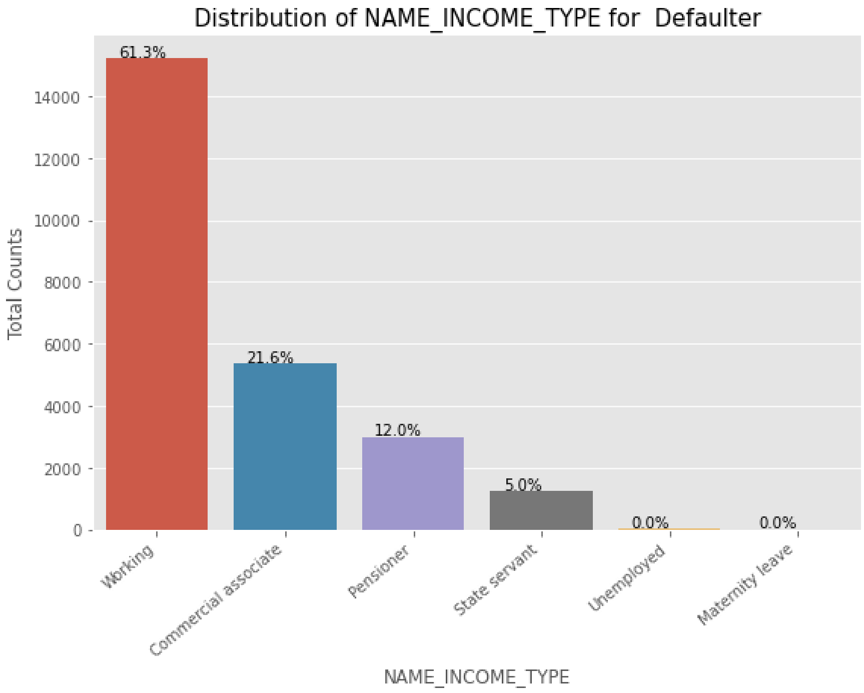
People without car are more non defaulter since there are more people without car

People with car defaults less 30.5% as observed in graph

Univariate analysis of categorical variables
NAME_INCOME_TYPE



Distribution of NAME_INCOME_TYPE for Non Defaulter



Distribution of NAME_INCOME_TYPE for Defaulter

Univariate analysis of categorical variables NAME_INCOME_TYPE

Observations

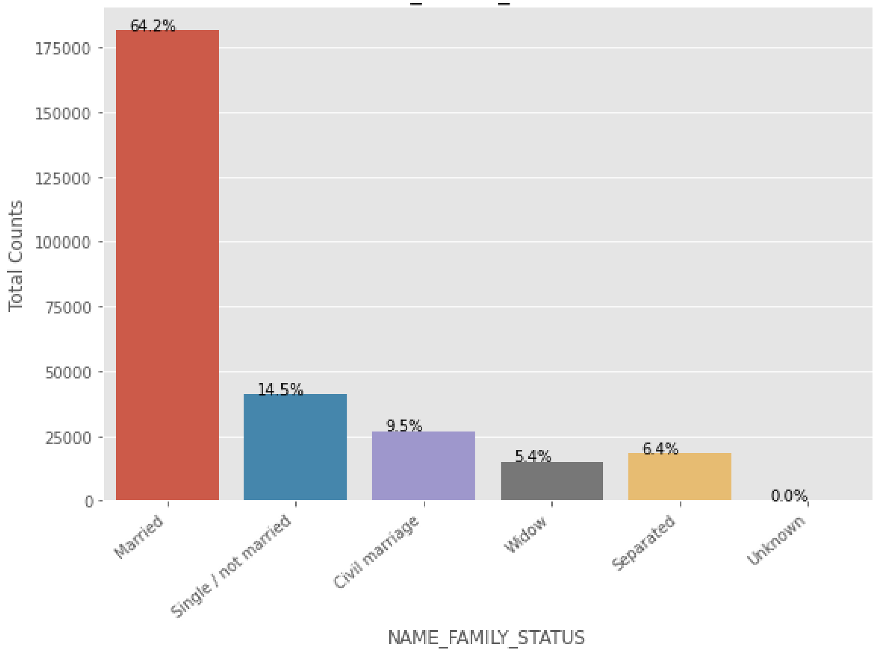
Students do not default since they are not bounded to pay till student life

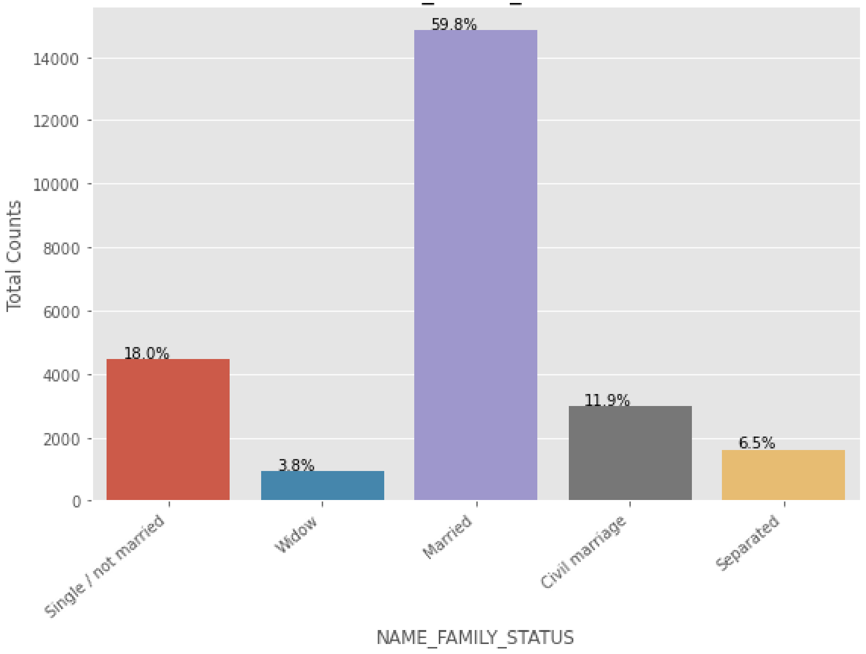
Businessman also don’t default

Working class has more contribution

Working class has more contribution in defaulter category as compared to non defaulter

Univariate analysis of categorical variables NAME_FAMILY_STATUS


Distribution of NAME_FAMILY_STATUS for Non Defaulter


Distribution of NAME_FAMILY_STATUS for Defaulter

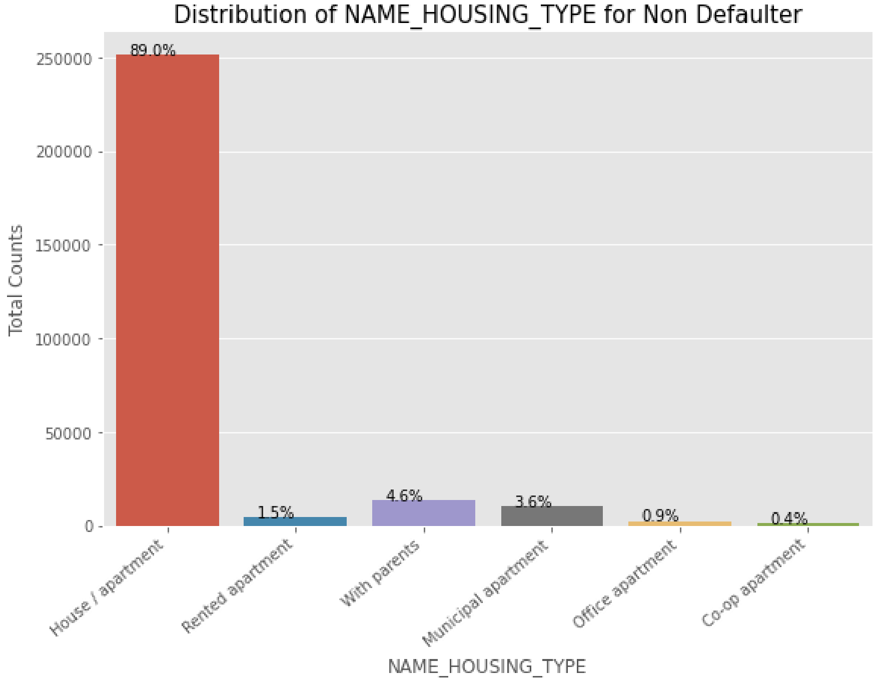
Univariate analysis of categorical variables NAME_FAMILY_STATUS

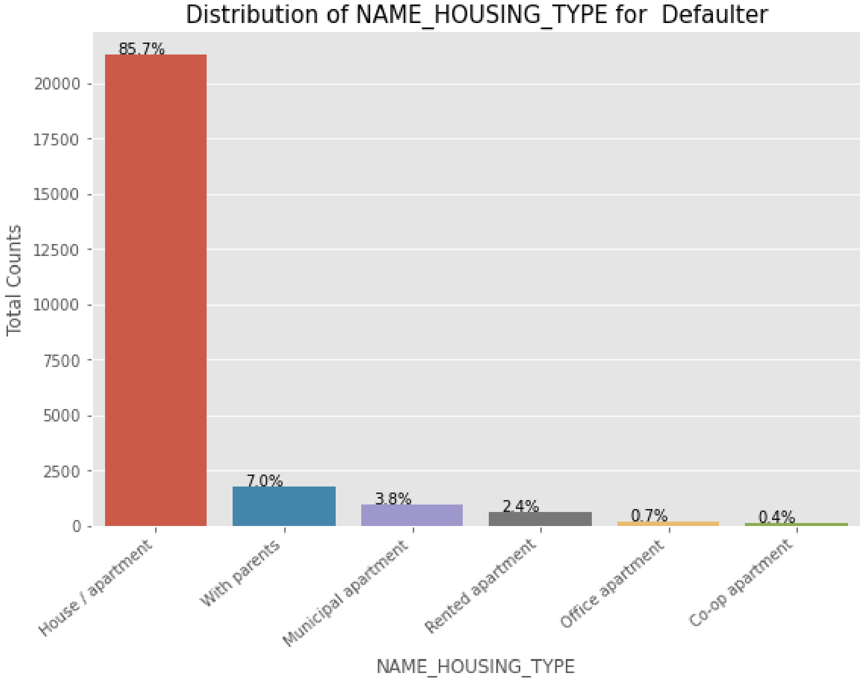
Observation

Married people here has more contribution in non defaulter and defaulter category since they avail more loans

Single people has less liability hence they are second highest contributor in non defaulter category

Univariate analysis of categorical variables NAME_HOUSING_TYPE


Distribution of NAME_HOUSING_TYPE for Non Defaulter


Distribution of NAME_HOUSING_TYPE for Defaulter

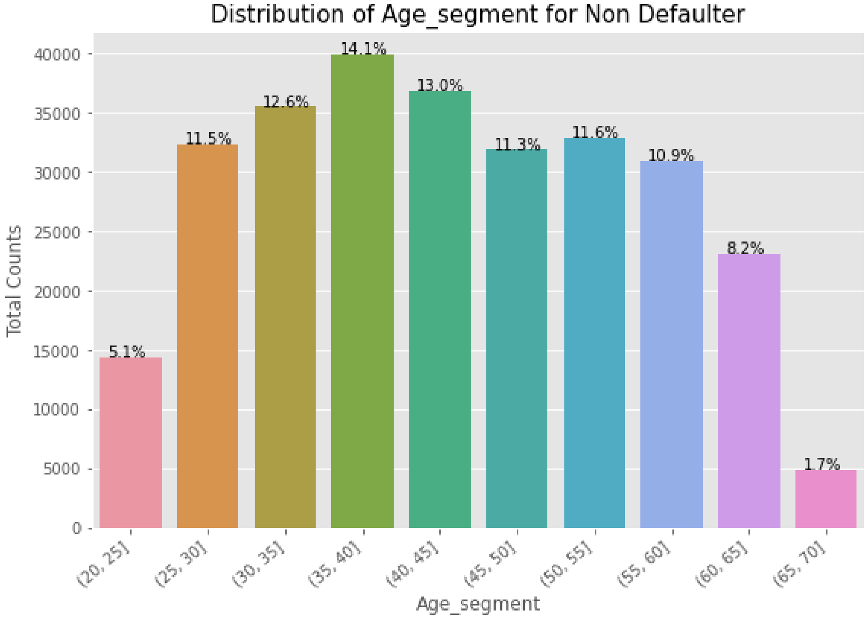
Univariate analysis of categorical variables NAME_HOUSING_TYPE

Observation-

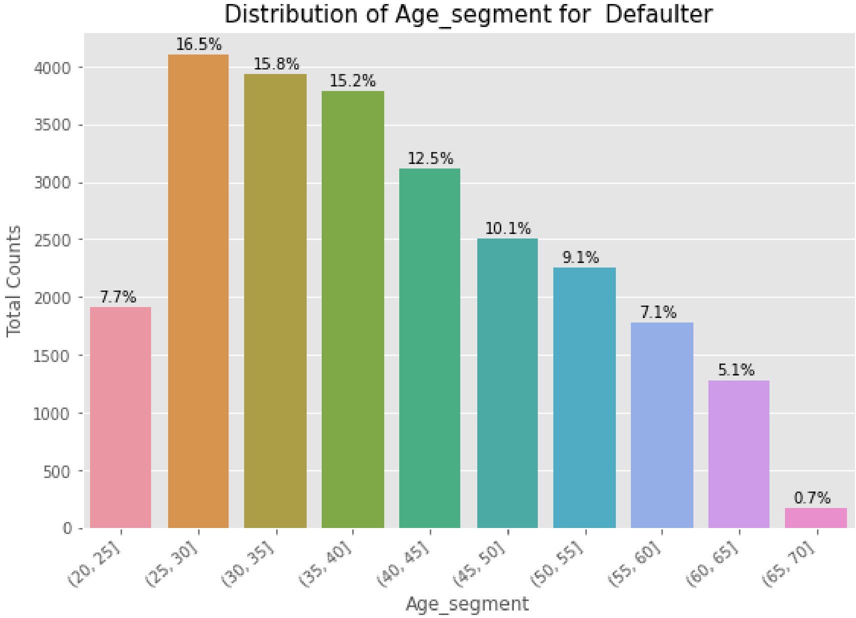
House/apartment segment has highest contribution in Non defaulter and Defaulter segment since they apply and get most of the loan

Person living with person has high level of dafault due high expenses and liabilities

Univariate analysis of categorical variables 'Age_segment'



Distribution of Age_segment for Non Defaulter



Distribution of Age_segment for Defaulter

Univariate analysis of categorical variables 'Age_segment'

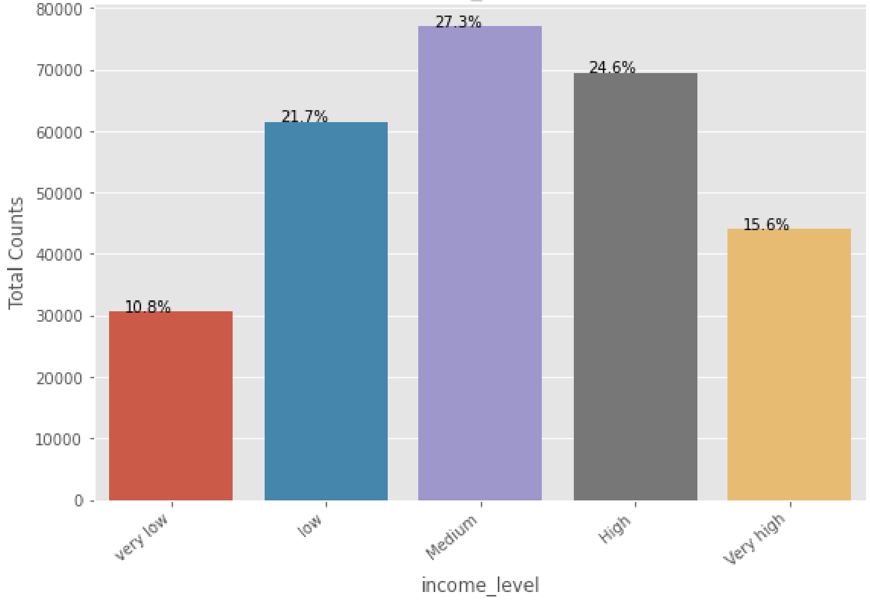
Observation

We observe here that person with age segment (25-30) tends to default more and the most risky segment to provide loan

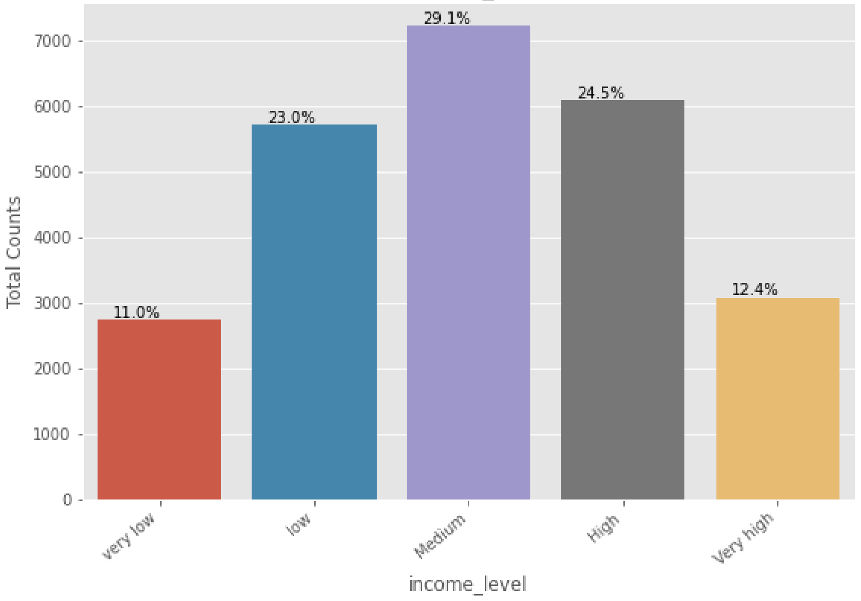
With age stability increases

Univariate analysis of categorical variables ‘income_level'

Distribution of income_level for Non Defaulter



Distribution of income_level for Defaulter



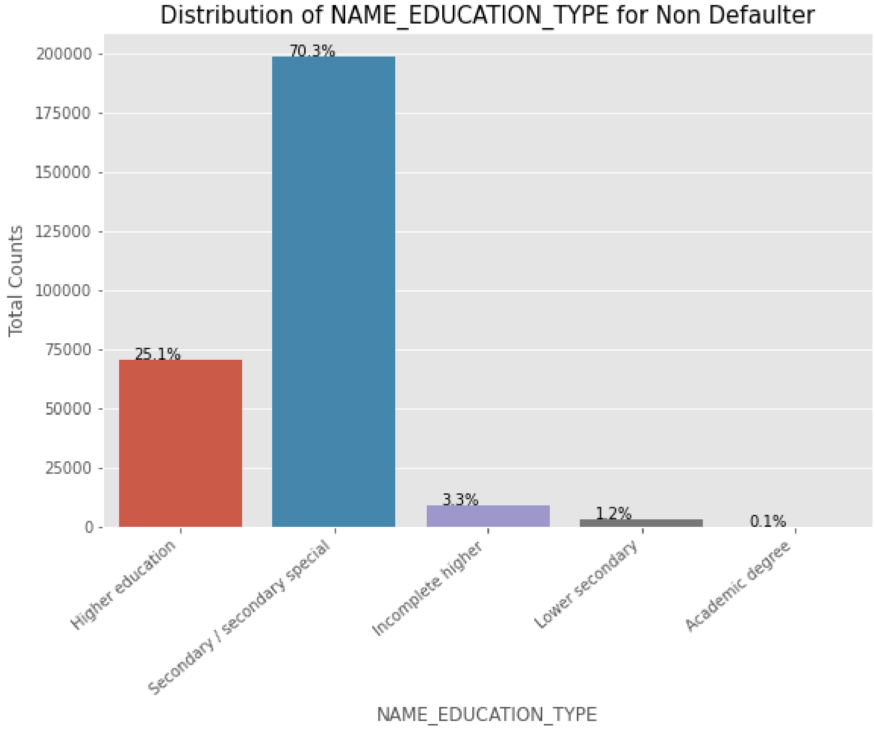
Univariate analysis of categorical variables ‘income_level'

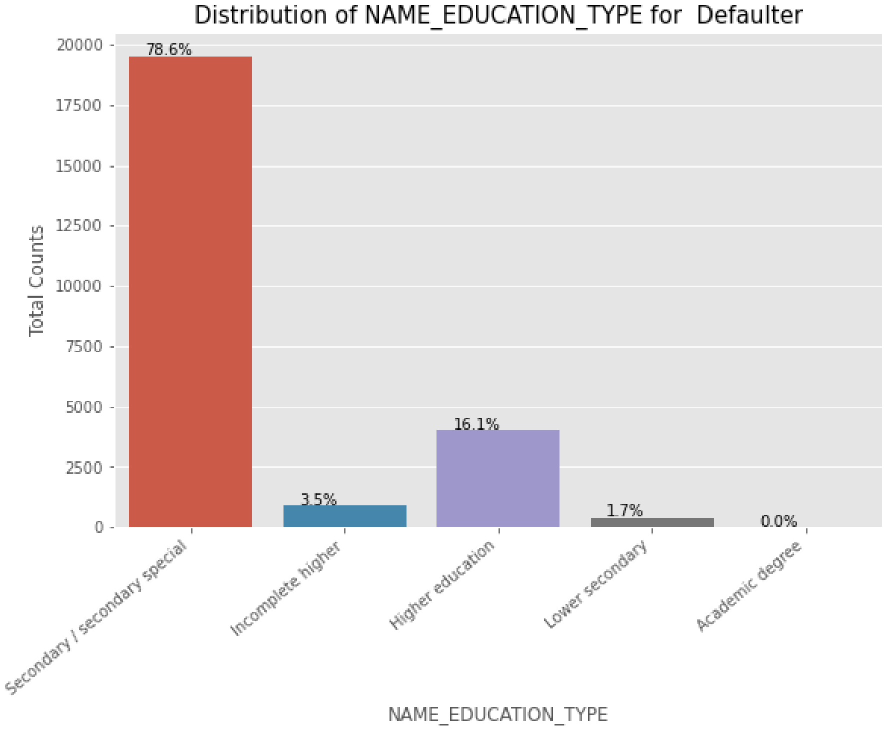
Observation

Very High income level people tends to default less while medium level most

Very low income group are also defaults less because they are given less loans

Univariate analysis of categorical variables
NAME EDUCATION TYPE





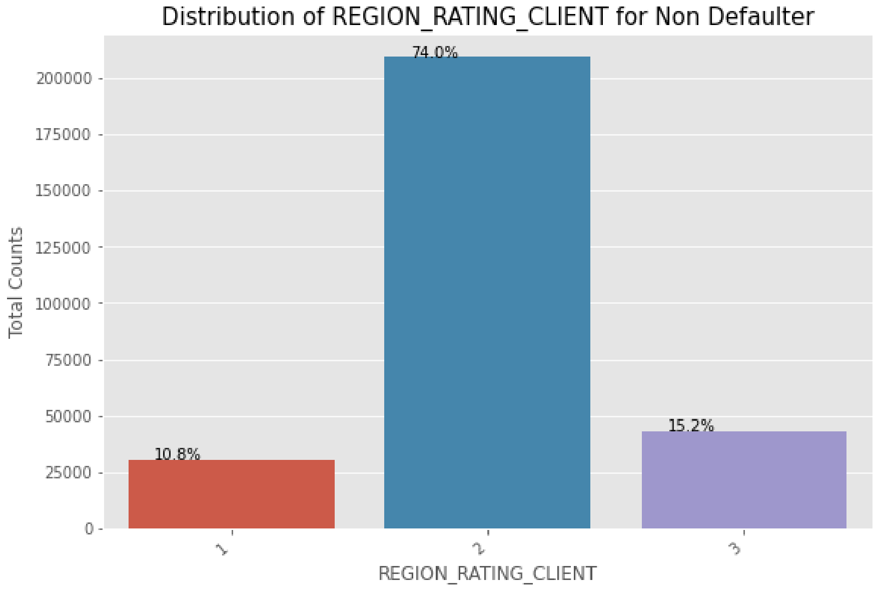
Univariate analysis of categorical variables NAME_EDUCATION_TYPE

Observations

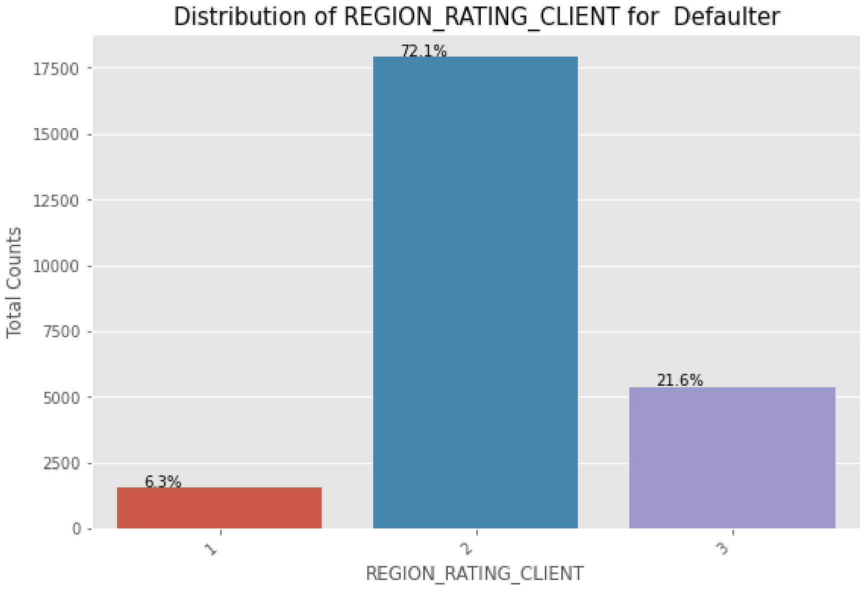
Secondary educated people has most non defaulter contribution and more prone to default

Less educated people has low share

Univariate analysis of categorical variables REGION_RATING_CLIENT



Distribution of REGION_RATING_CLIENT for Non Defaulter



Distribution of REGION_RATING_CLIENT for Defaulter

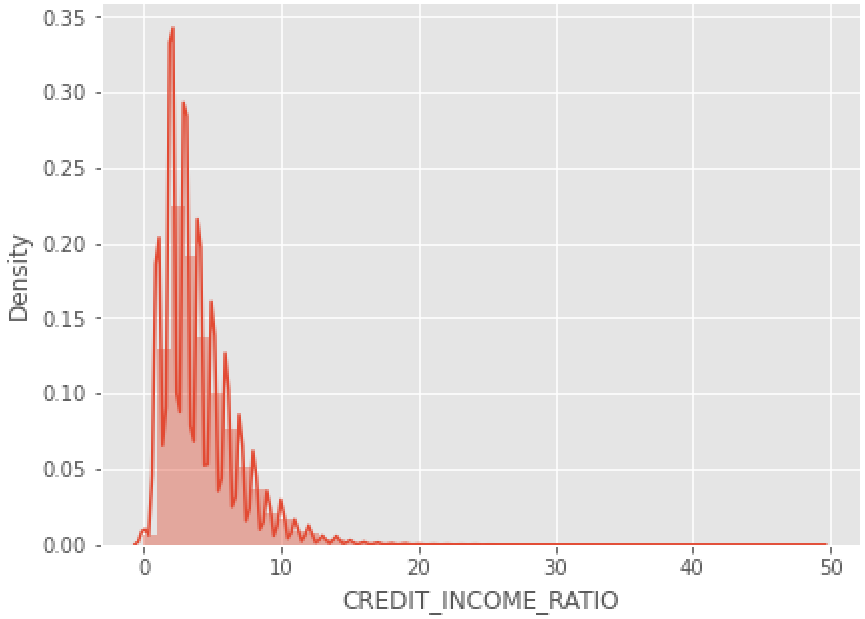
Univariate analysis of categorical variables REGION_RATING_CLIENT

Observation

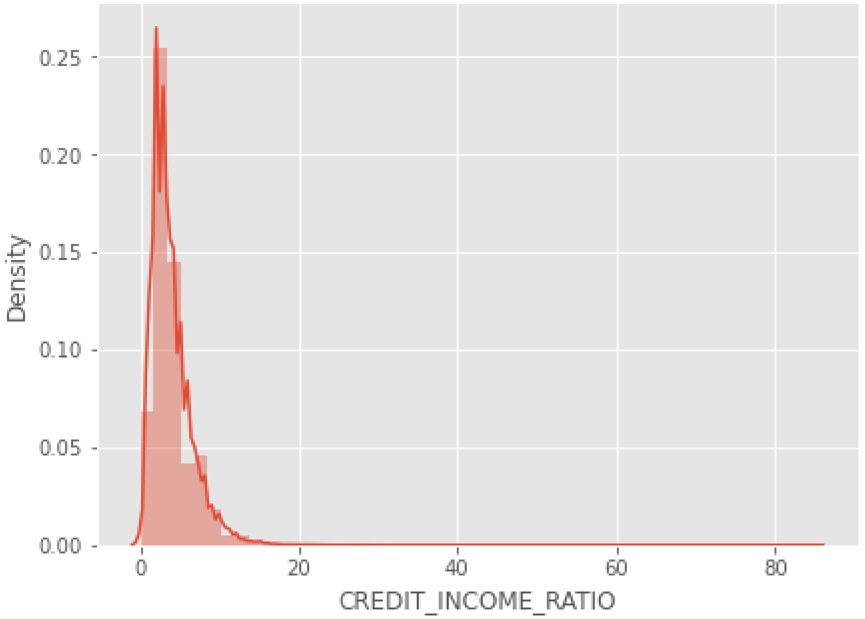
There is more default in 2 rated region and they apply more

1 rated region has least availed the loan

Univariate analysis of continuous variables
CREDIT_INCOME_RATIO



Distribution of CREDIT_INCOME_RATIO for Non Defaulter



Distribution of CREDIT_INCOME_RATIO for Defaulter

Univariate analysis of continuous variables CREDIT_INCOME_RATIO

Observation

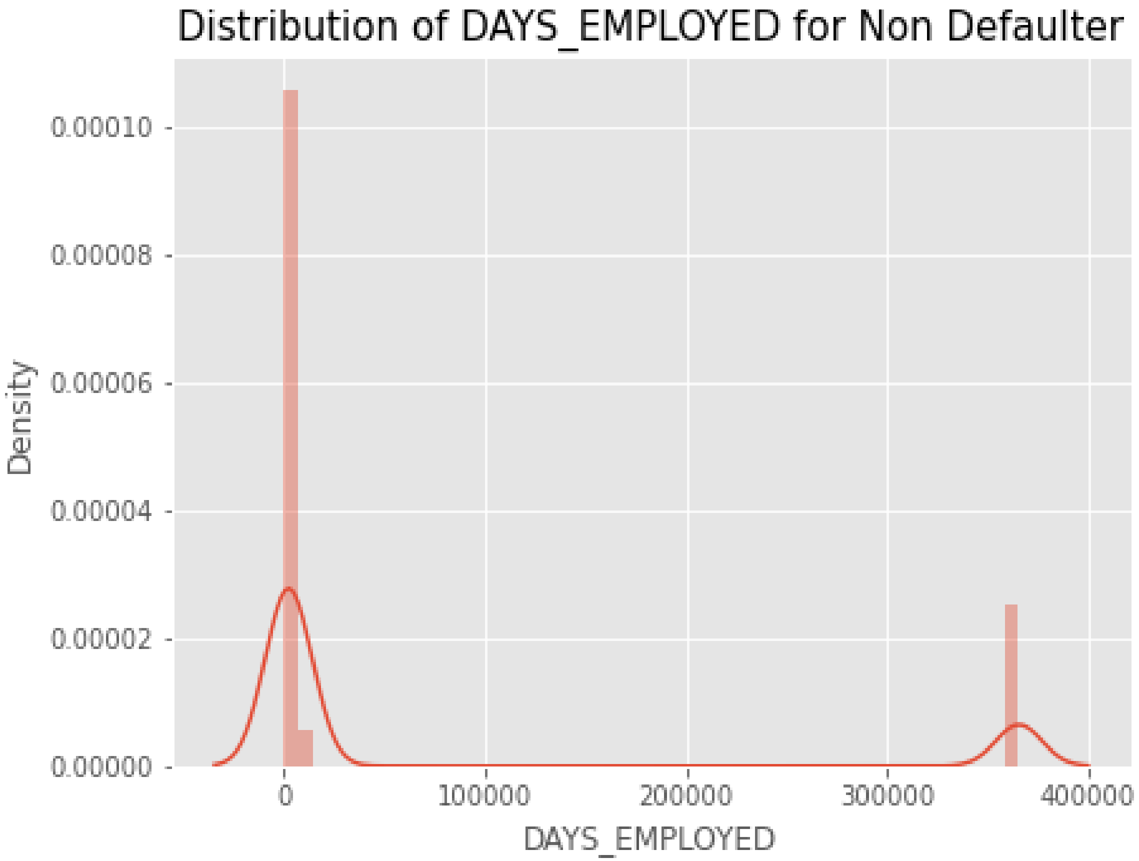
There is no much difference between people group who defaulted and who were non defaulter

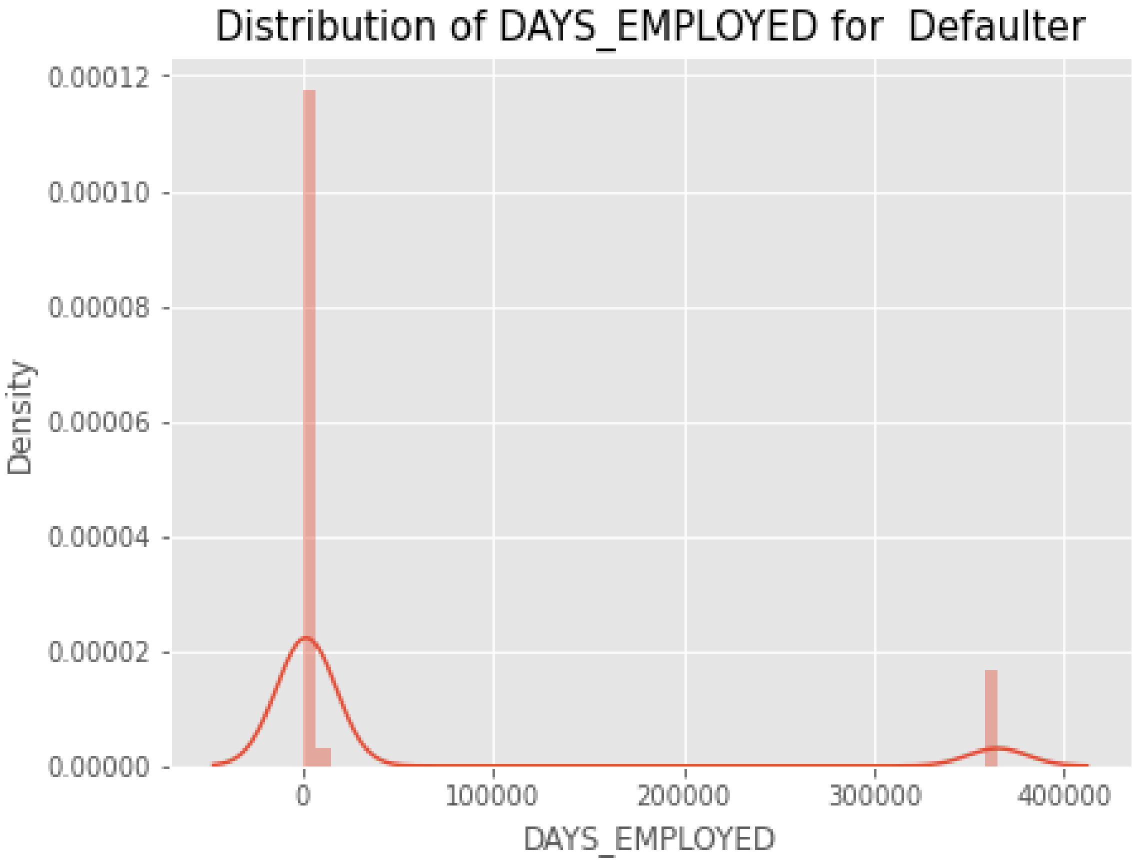
It shows that when CREDIT_INCOME_RATIO is more than 50, people default

Univariate analysis of continuous variables

DAYS_EMPLOYED

Less days employed people defaults more

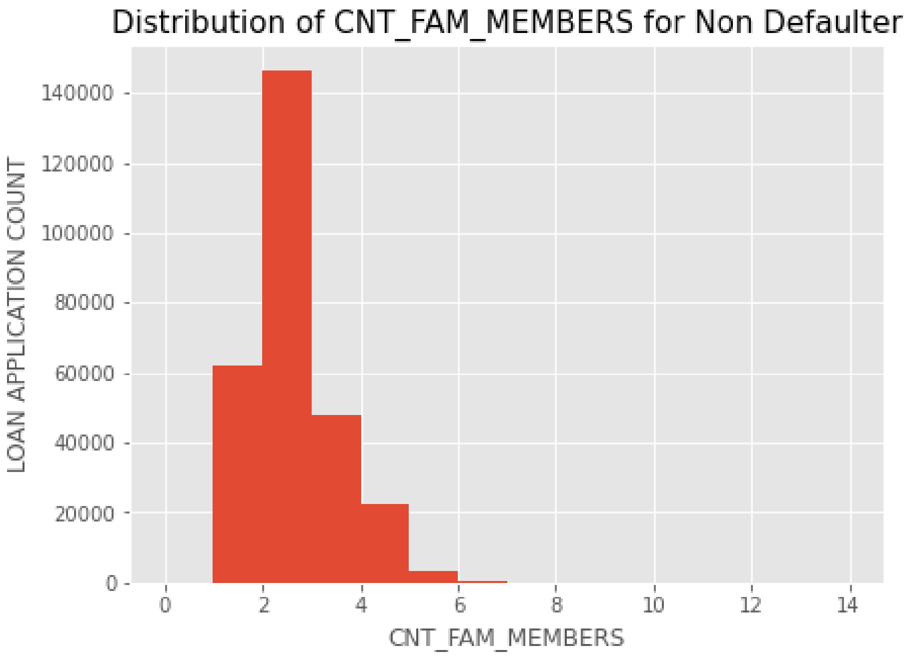


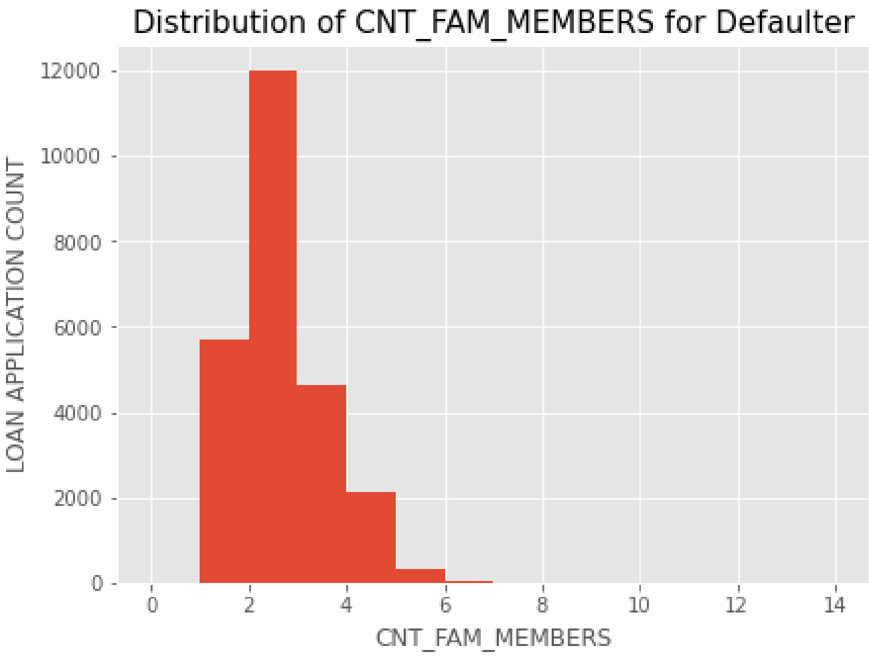


Univariate analysis of CNT_FAM_MEMBERS

Observation

Family of 3 applies for loan more frequently





Variables with high correlation in Non defaulter category

	Column1	Column2	Correlation	Abs_Correlation
919	FLAG_EMP_PHONE	DAYS_EMPLOYED	-0.999705	0.999705
4768	Age	DAYS_BIRTH	0.999691	0.999691
2767	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998269	0.998269
2481	FLOORSMAX_MEDI	FLOORSMAX_AVG	0.997187	0.997187
2410	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.996124	0.996124
2483	FLOORSMAX_MEDI	FLOORSMAX_MODE	0.989195	0.989195
2341	FLOORSMAX_MODE	FLOORSMAX_AVG	0.986594	0.986594
424	AMT_GOODS_PRICE	AMT_CREDIT	0.982783	0.982783
2270	YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.980466	0.980466
2412	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.978073	0.978073

Variables with high correlation in Non defaulter category

- Highly correlated Variables are Age Days_Birth(date of birth) which is obvious
- As soon as goods prices increases ,credit amount also increases or decreases accordingly

Variables with high correlation in Defaulter category

	Column1	Column2	Correlation	Abs_Correlation
308	AMT_GOODS_PRICE	AMT_CREDIT	0.983103	0.983103
297	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.956637	0.956637
208	SOCIAL_CIRCLE_60_DAYS_DEF_PERC	SOCIAL_CIRCLE_30_DAYS_DEF_PERC	0.874562	0.874562
321	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699	0.752699
272	AMT_ANNUITY	AMT_CREDIT	0.752195	0.752195
74	CREDIT_INCOME_RATIO	AMT_CREDIT	0.639744	0.639744
310	AMT_GOODS_PRICE	CREDIT_INCOME_RATIO	0.623163	0.623163
274	AMT_ANNUITY	CREDIT_INCOME_RATIO	0.381298	0.381298
113	DAYS_REGISTRATION	DAYS_EMPLOYED	-0.188929	0.188929
149	CNT_FAM_MEMBERS	DAYS_EMPLOYED	-0.186561	0.186561

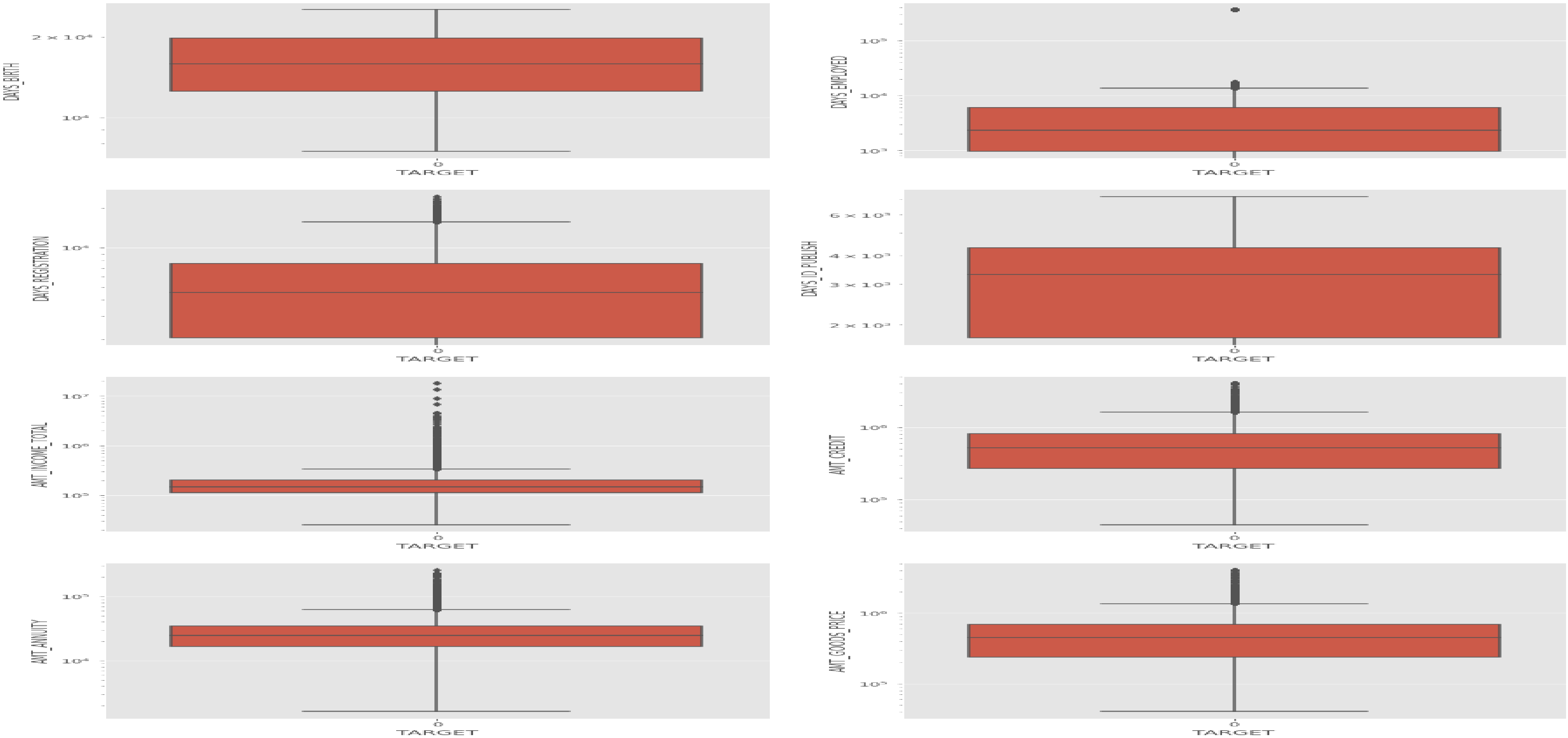
Variables with high correlation in Defaulter category

Observation

Credit amount varies with price of goods

Annuity amount varies with price of goods

Bivariate continuous plots



Bivariate continuous plots

Observations

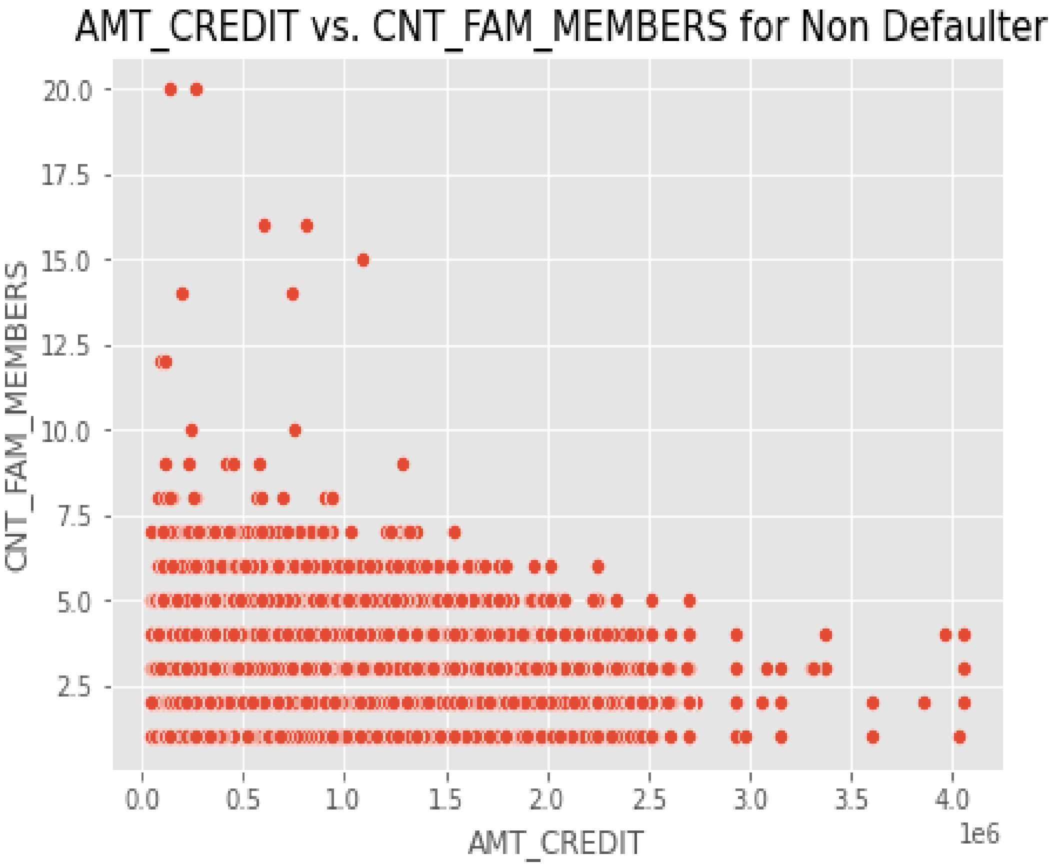
In non default case,AMT_GOOD_PRICE contains more outlier than default cases

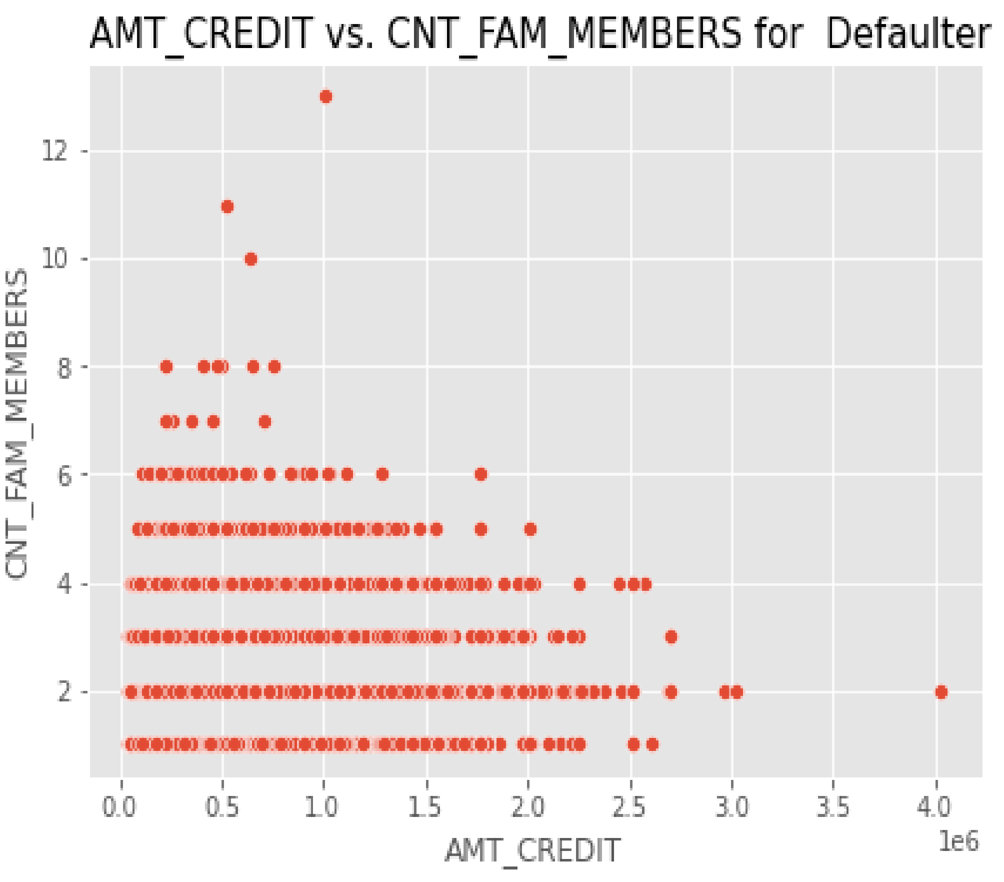
People with higher no of employment days tends to default less

In default case,most of the client amount annuity is greater than median value i.e. 25000

Mostly defaulters have less income

Bivariate analysis of numerical variable AMT_CREDIT,CNT_FAM_MEMBERS





Bivariate analysis of numerical variable AMT_CREDIT,CNT_FAM_MEMBERS

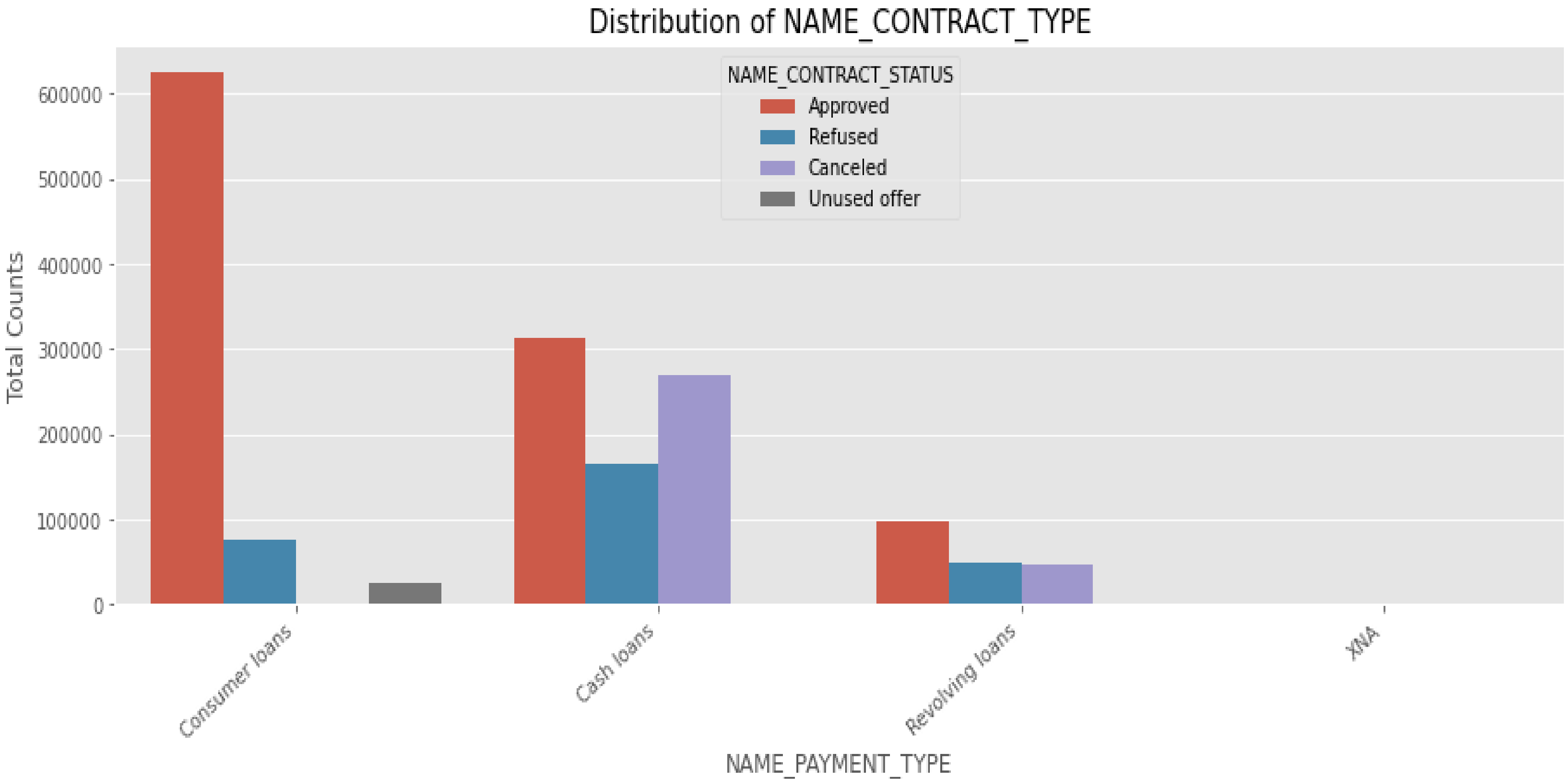
Observations

Small family and low credit amount tends to default less

Large family with high credit amount default less often

Univariate analysis on previous dataset NAME_CONTRACT_TYPE

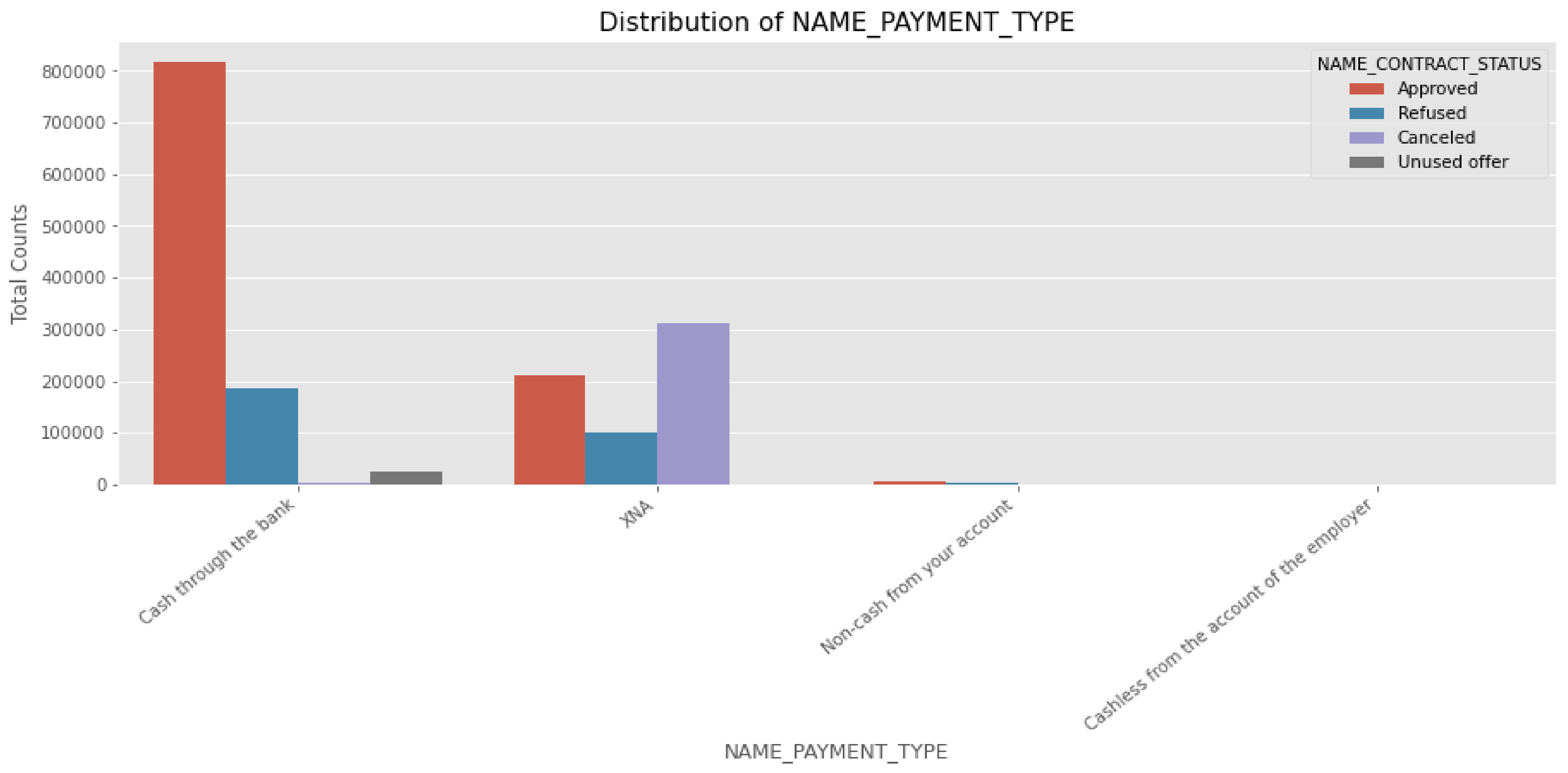
Observation:Most loans are consumer and cash ,cash loans are most rejected



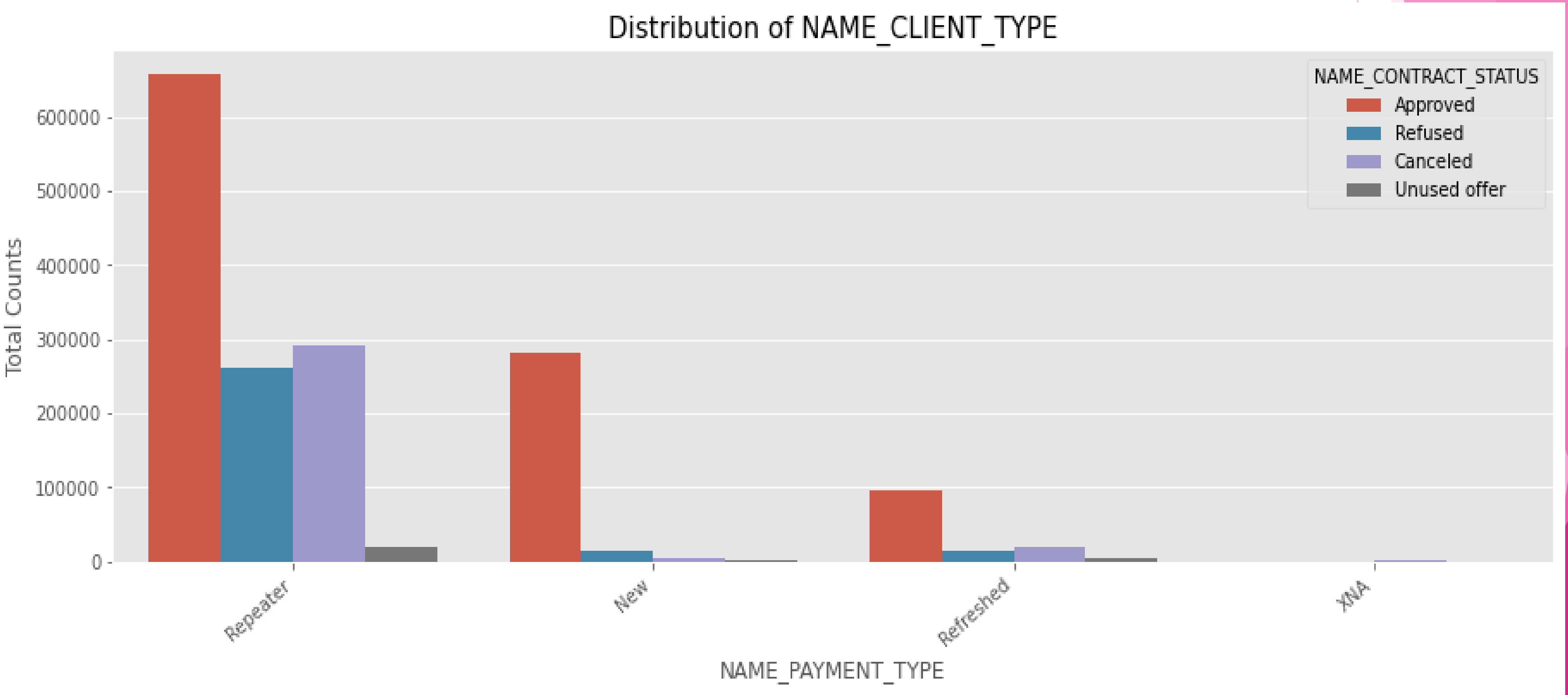
Univariate analysis on previous dataset

NAME_PAYMENT_TYPE

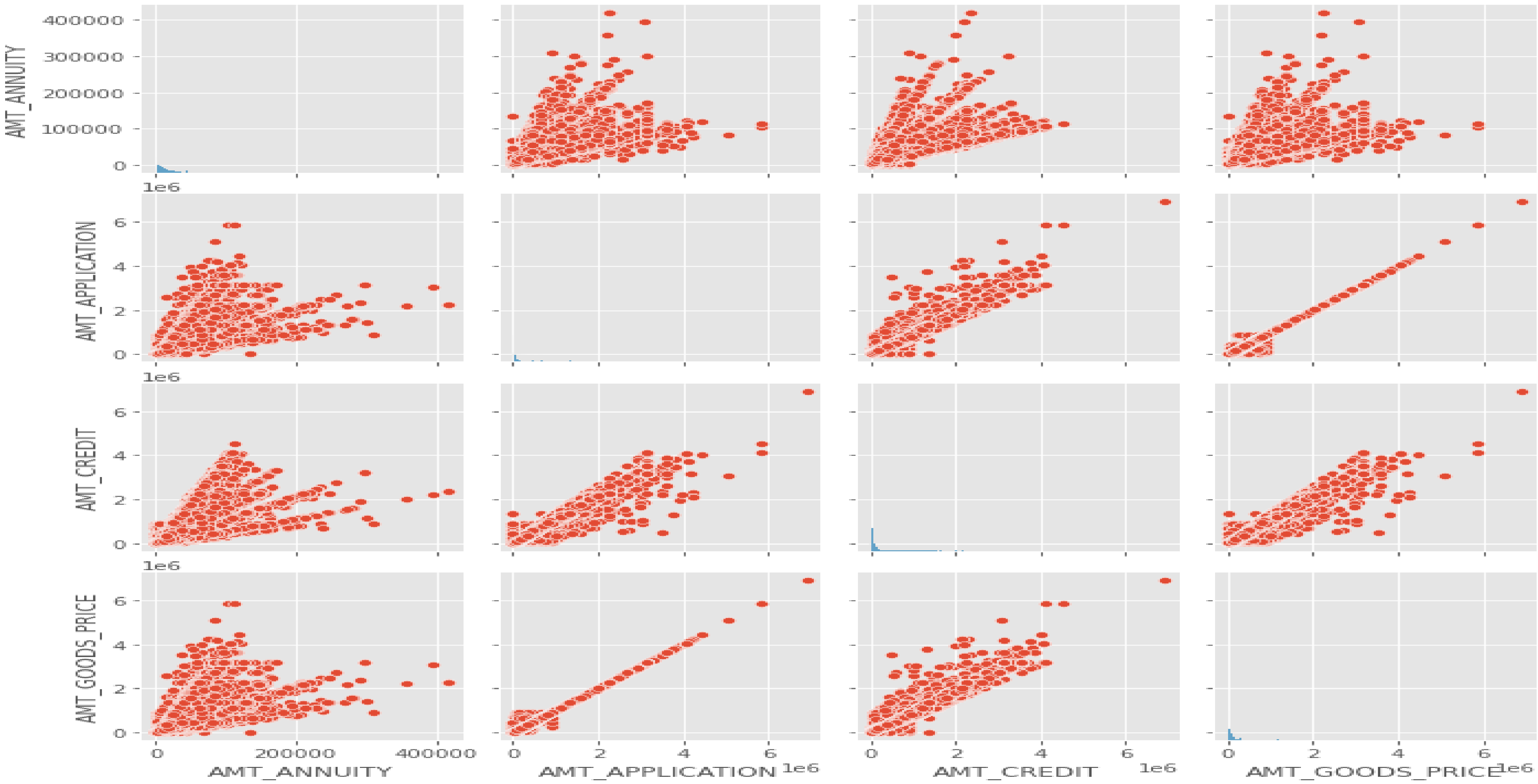
Observation: Loan repayment is mostly through the bank



Univariate analysis on previous dataset NAME_CLIENT_TYPE Observation:Most of loan requests are from repeated customers



Previous data Bivariate pairplot



Previous data Bivariate analysis

Previous data has high impact of annuity on credit,final amount and goods price

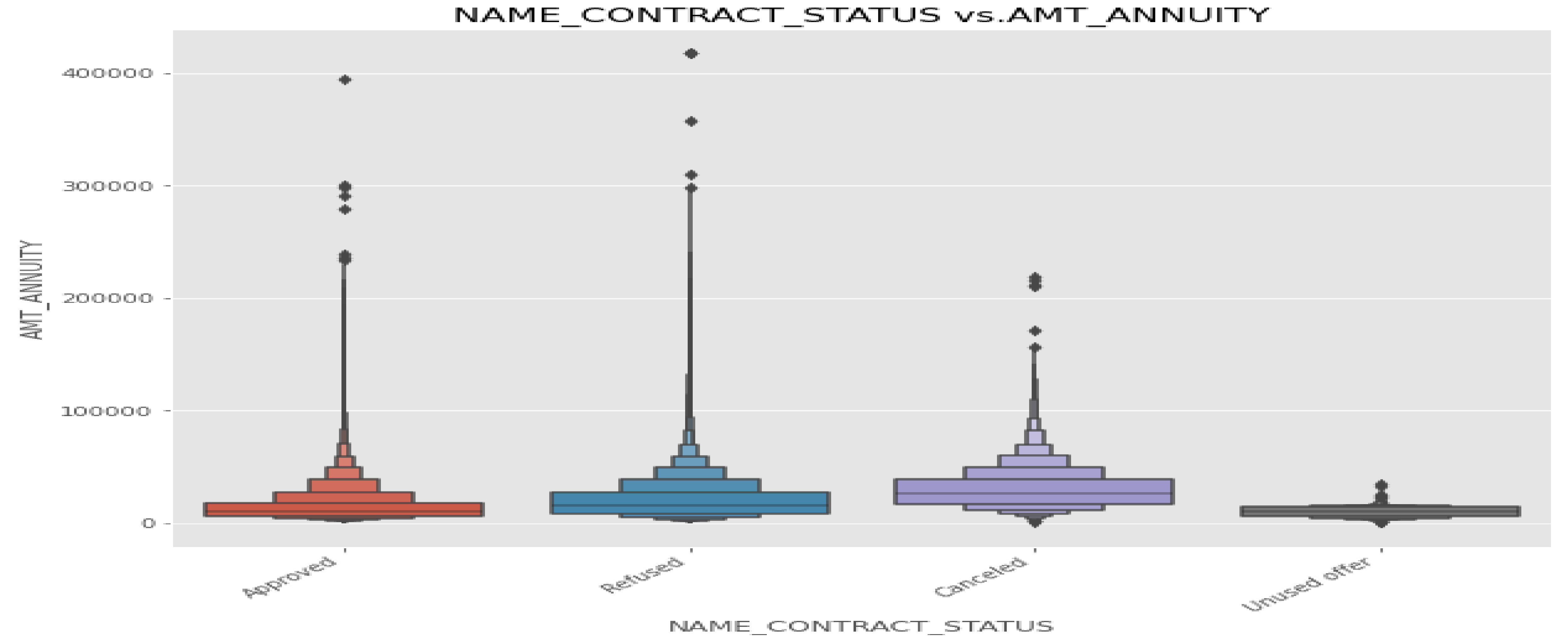
Credit amount asked by client has been highly related to goods price

The amount released is highly related to amount asked and goods price

Previous data Bivariate analysis
NAME_CONTRACT_STATUS','AMT_ANNUITY

Loan application with low Annuity gets canceled or unused

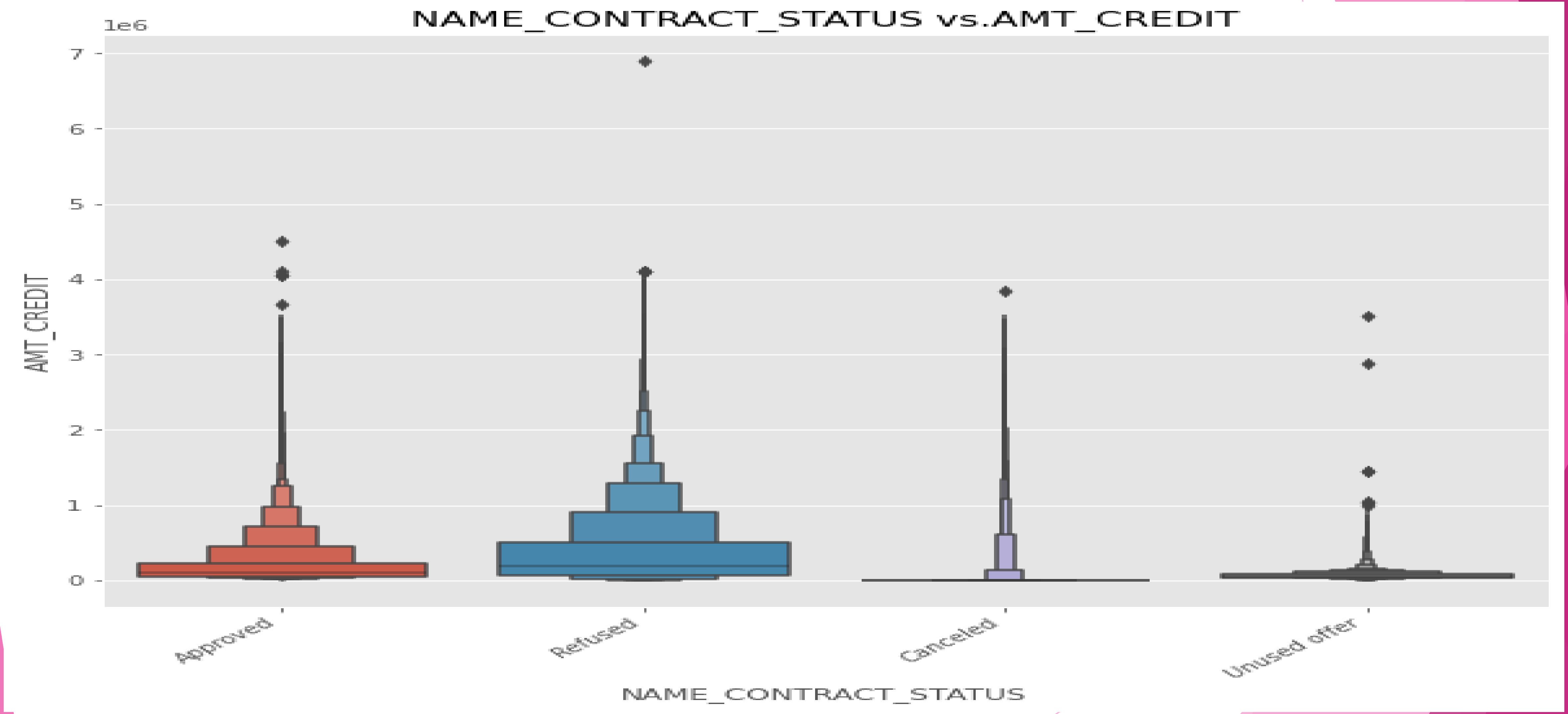
High annuity also gets rejected



Previous data Bivariate analysis

NAME_CONTRACT_STATUS','AMT_CREDIT

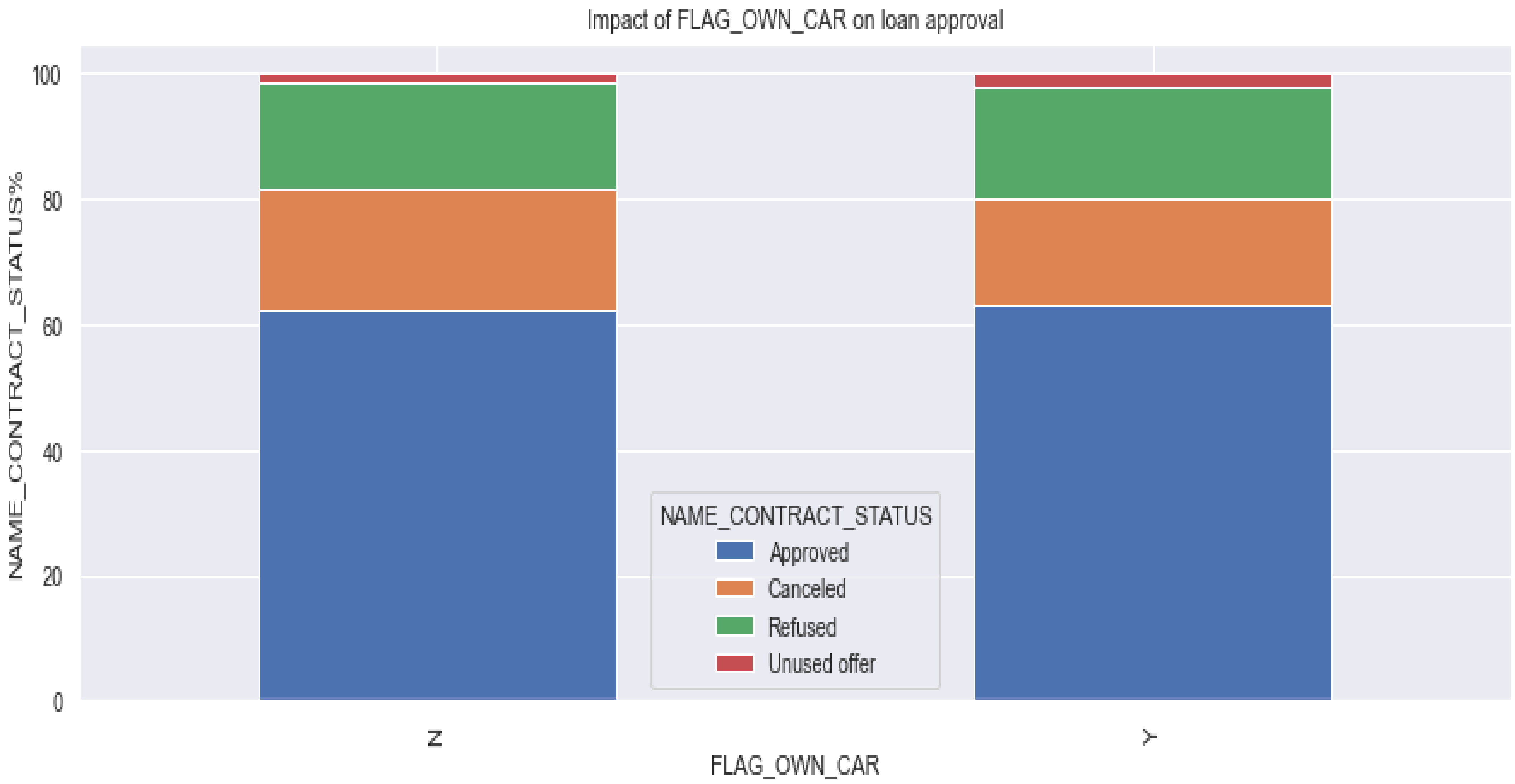
If amount credit is low then it gets cancelled or unused



Plots after merging data

FLAG_OWN_CAR','NAME_CONTRACT_STATUS

It shows that car owner ship does not impact on loan but earlier it was there hence more weightage should be given to it

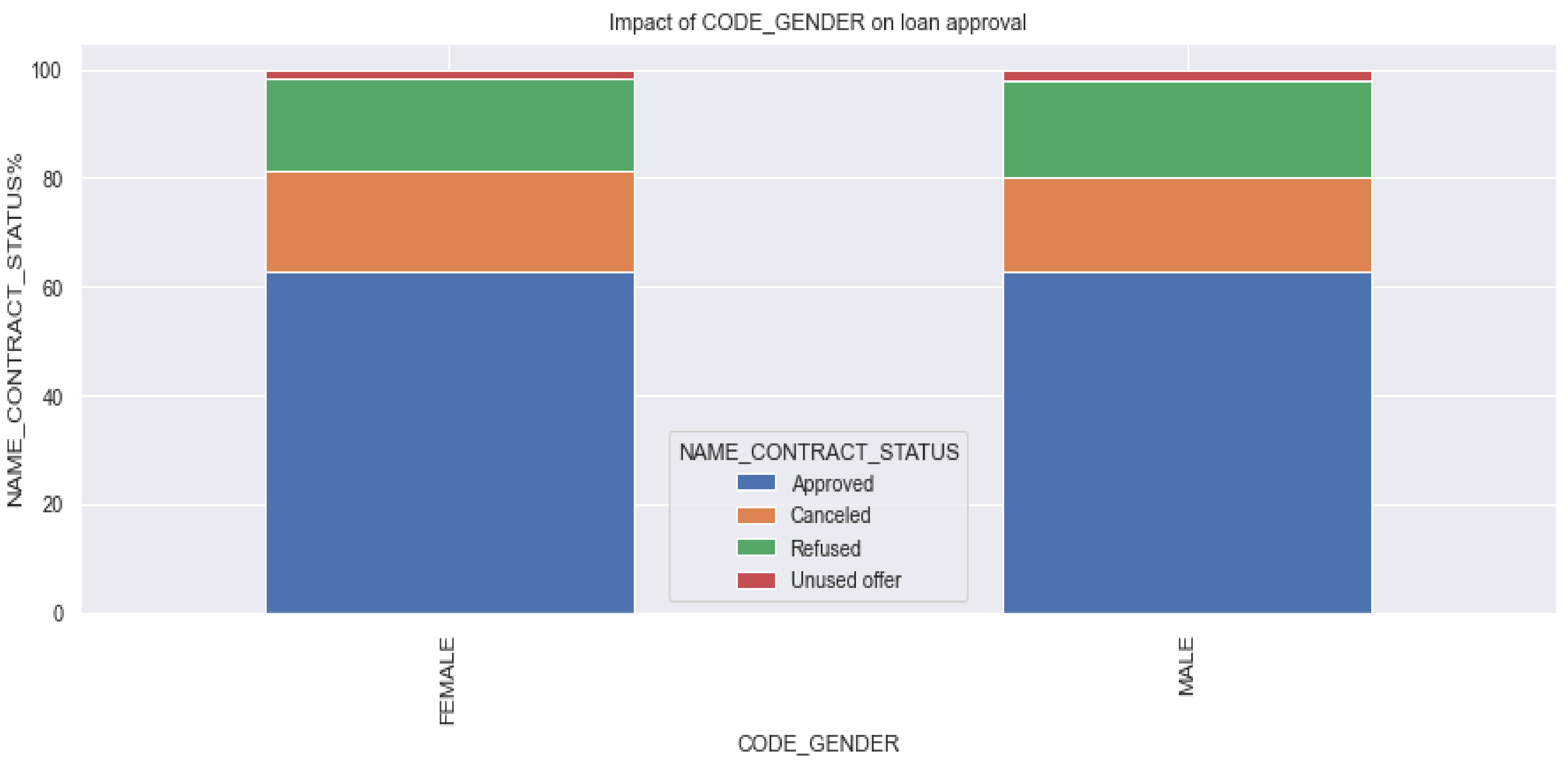


Impact of FLAG_OWN_CAR on loan approval

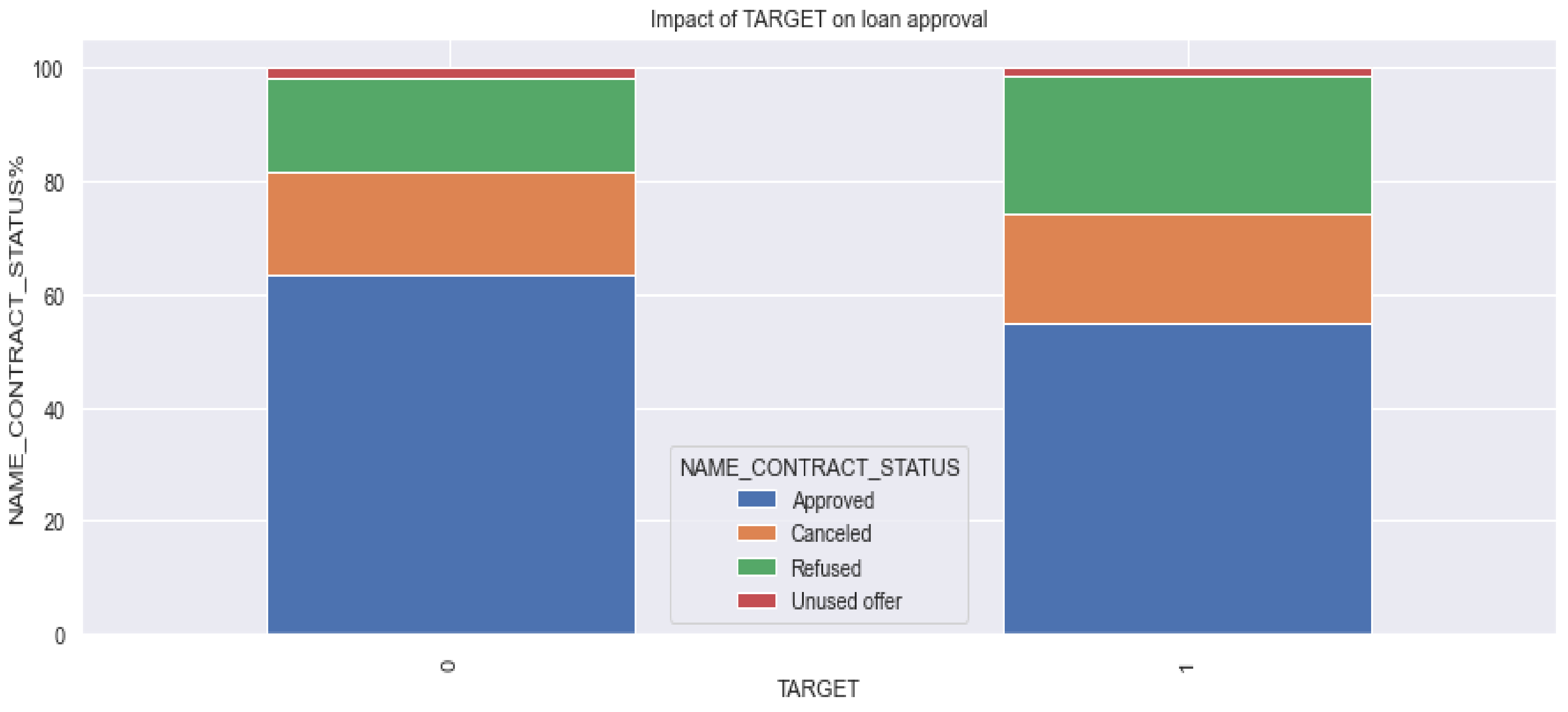
Plots after merging data

CODE_GENDER','NAME_CONTRACT_STATUS

There is no impact of gender however earlier female were less defaulter hence more weightage



Plots after merging data
TARGET','NAME_CONTRACT_STATUS
People who has already availed loan are less dafaulter



Recommendation

Following groups are less likely to default

Client with high income group

Old people of any income group

Old female client

Client with high education

Client who has availed loan earlier

High Risk Group

- Low educated clients who whose previous loans were rejected
- Male client with civil marriage
- Group who has been denied loan earlier

Thank You