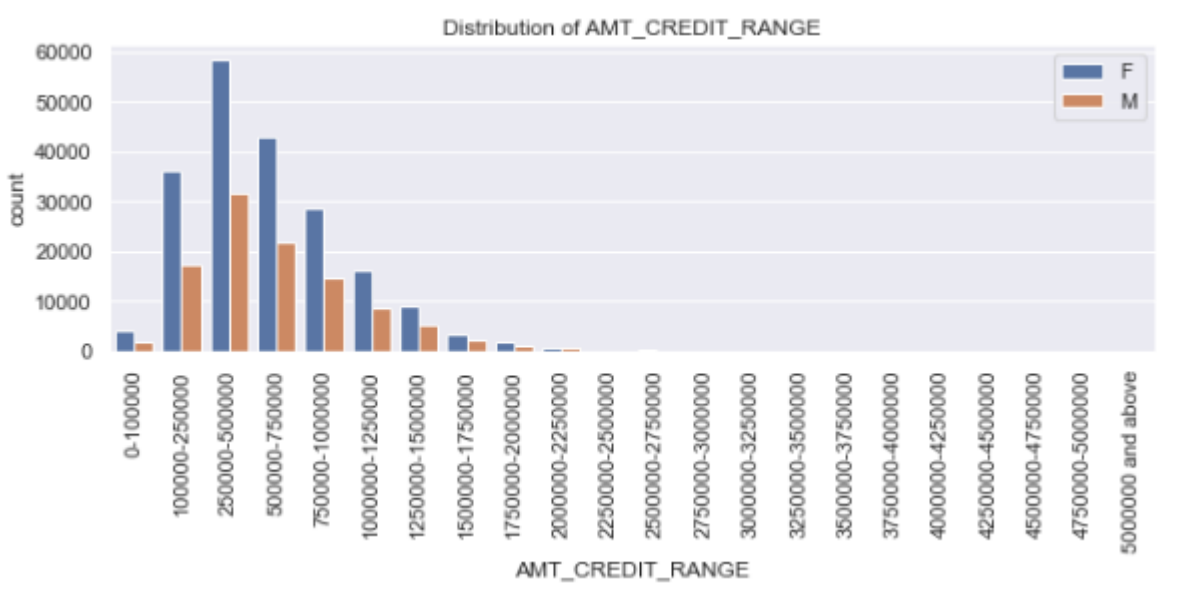
CREDIT EDA CASE STUDY

1. Kiran Sakegaonkar
2. Vinayak Joshi

Distribution of AMT_CREDIT_RANGE

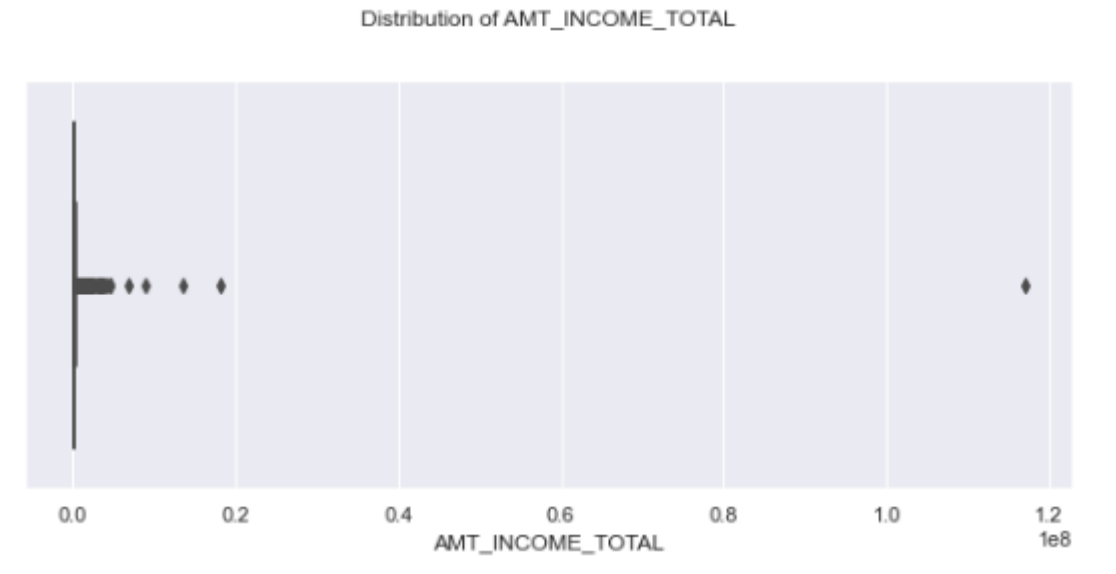


Interfernce :

Points to be concluded from the above graph.

1. Female counts are higher than male.
2. Income range from 100000 to 200000 is having more number of credits.
3. This graph show that females are more than male in having credits for that range.
4. Very less count for income range 400000 and above.

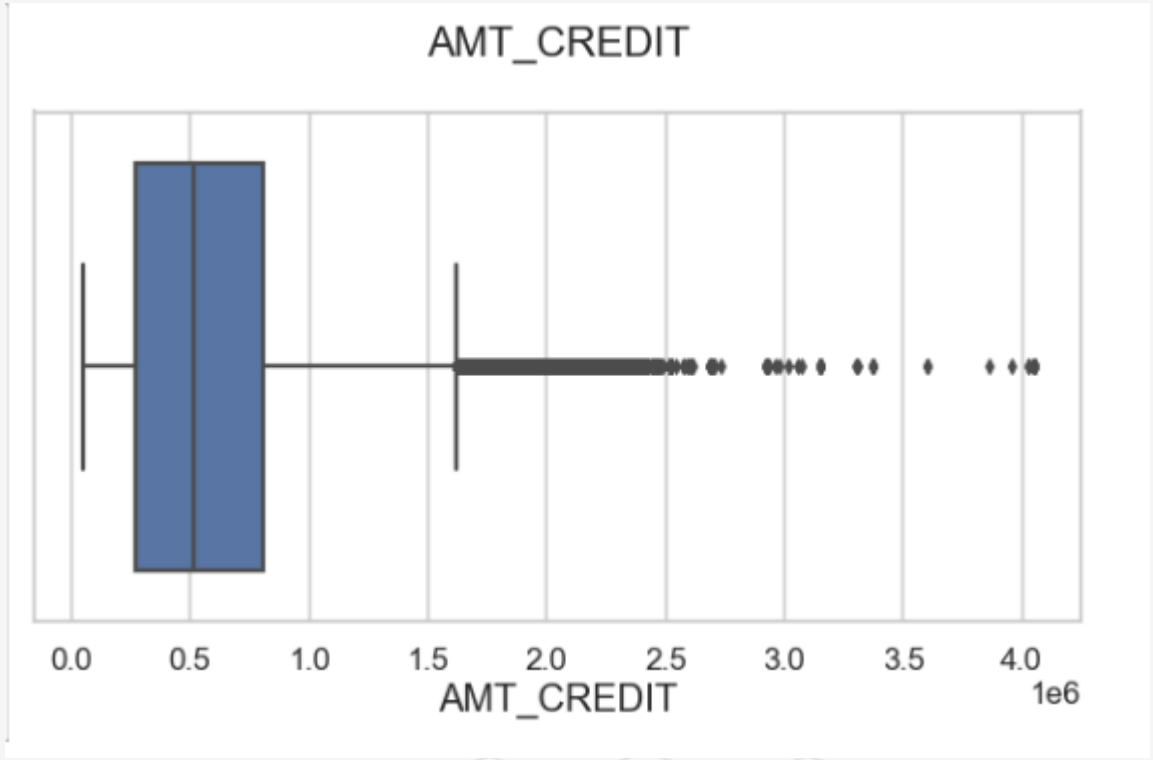
Distribution of AMT_INCOME_TOTAL



Interference: [AMT_INCOME_TOTAL]

1. This varibale indictes the Income of the client.As we can see from the plot there is one value which is too high compared to others.
2. Some outliers are noticed in income amount.
3. The third quartiles is very slim for income amount.

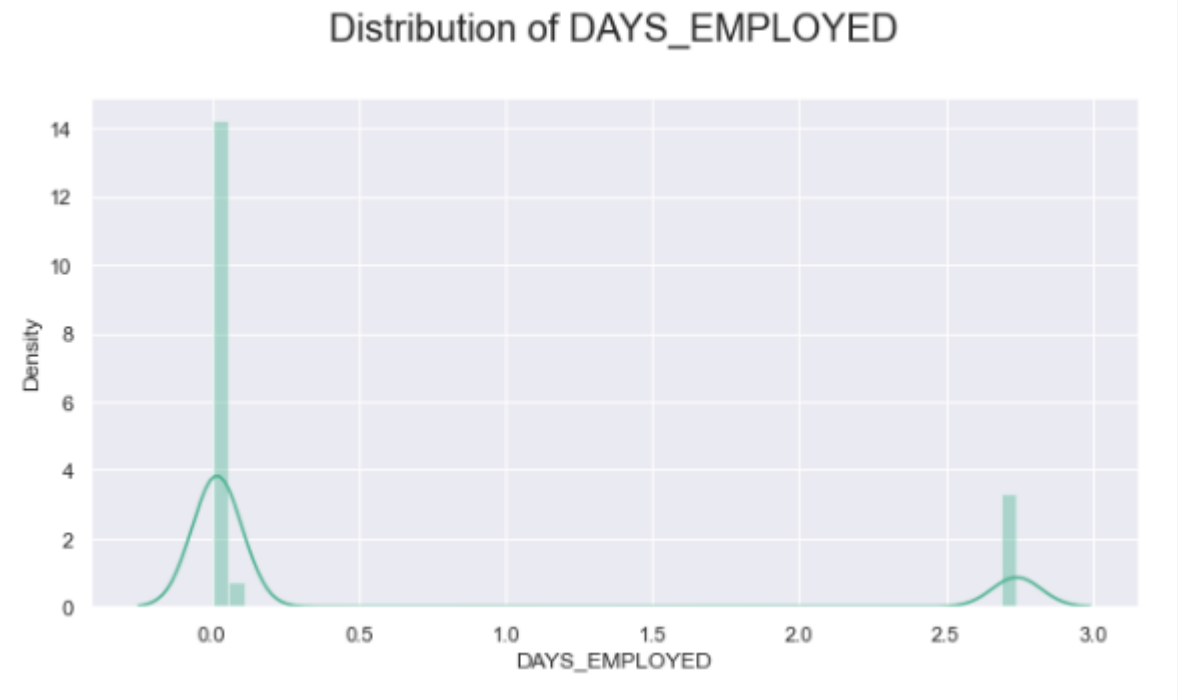
AMT_CREDIT



Interference:

1. Some outliers are noticed in credit amount.
2. The first quartile is bigger than third quartile for credit amount which means most of the credits of clients are present in the first quartile.

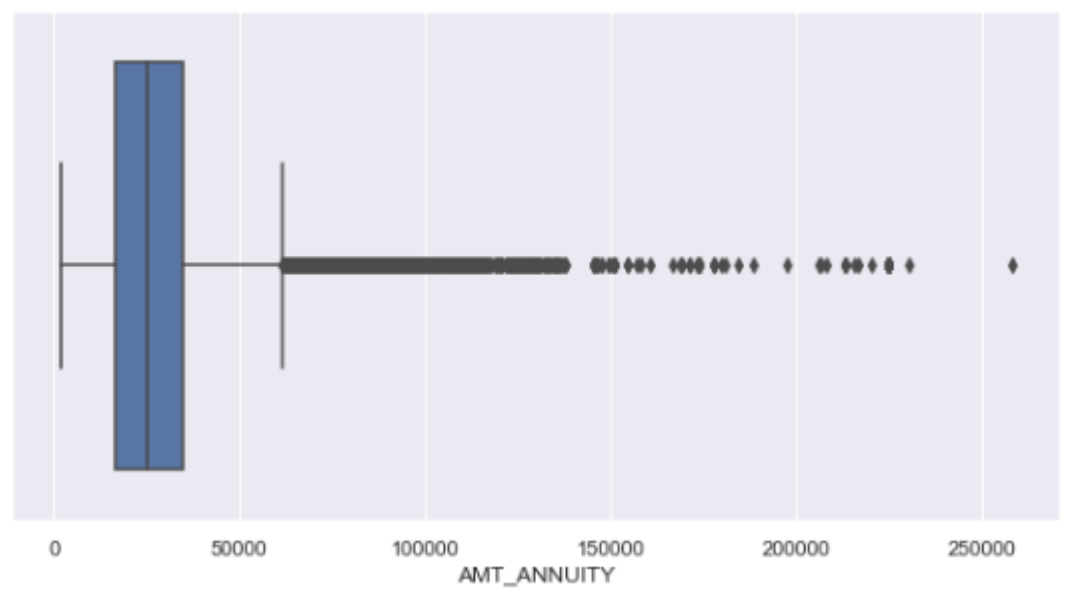
Distribution of DAYS_EMPLOYED



Interfernce:

1. Here, in the 'DAYS_EMPLOYED' which tells how many days before the application the person started current employment.
2. We observe a value which is greater that 20,000 which is surely an outlier because 25,000/365 will be around 54 years.

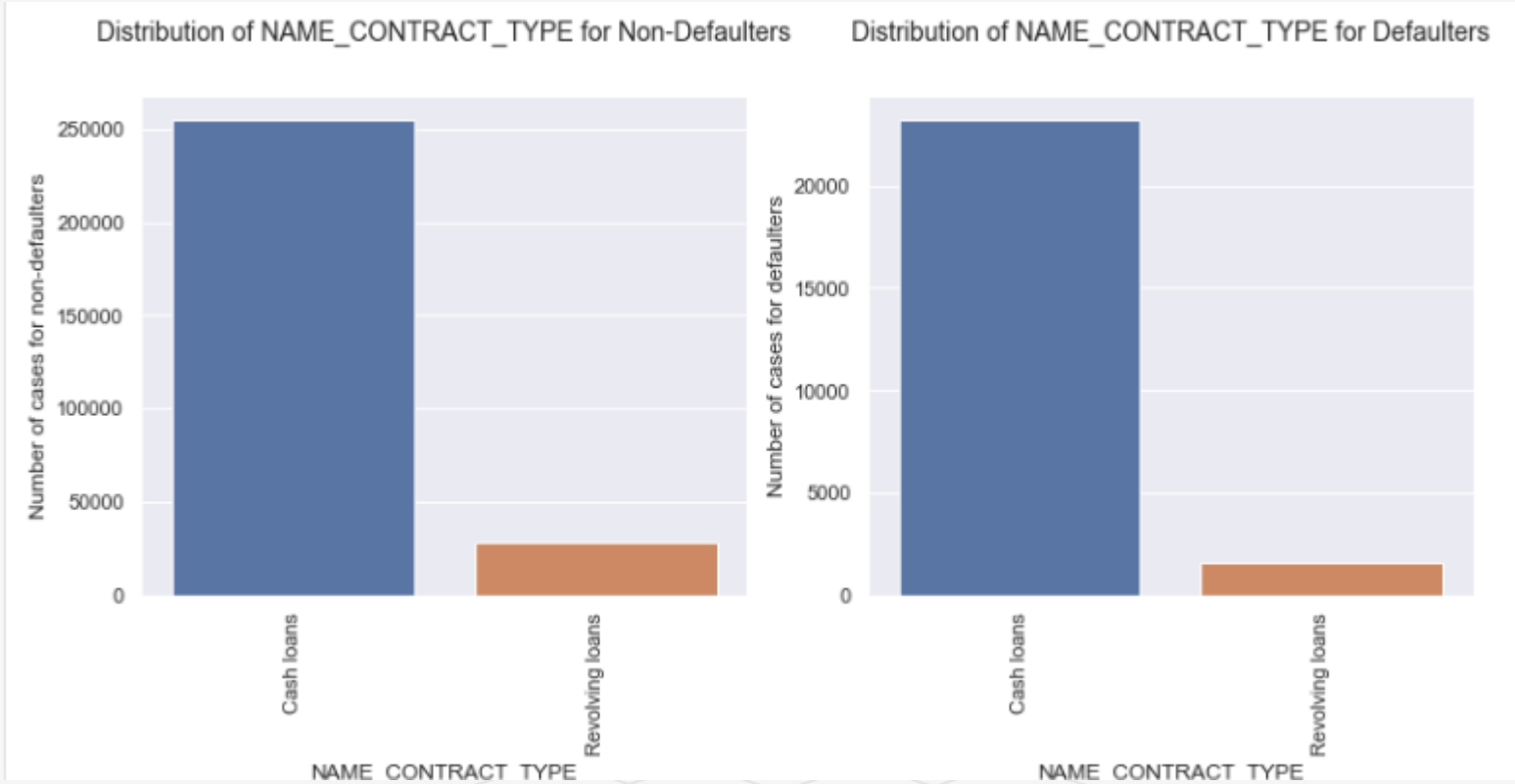
AMT_ANNUITY



Interfernce:

we observe that the credit amount lies between 250000 to around 500000 for defaulters

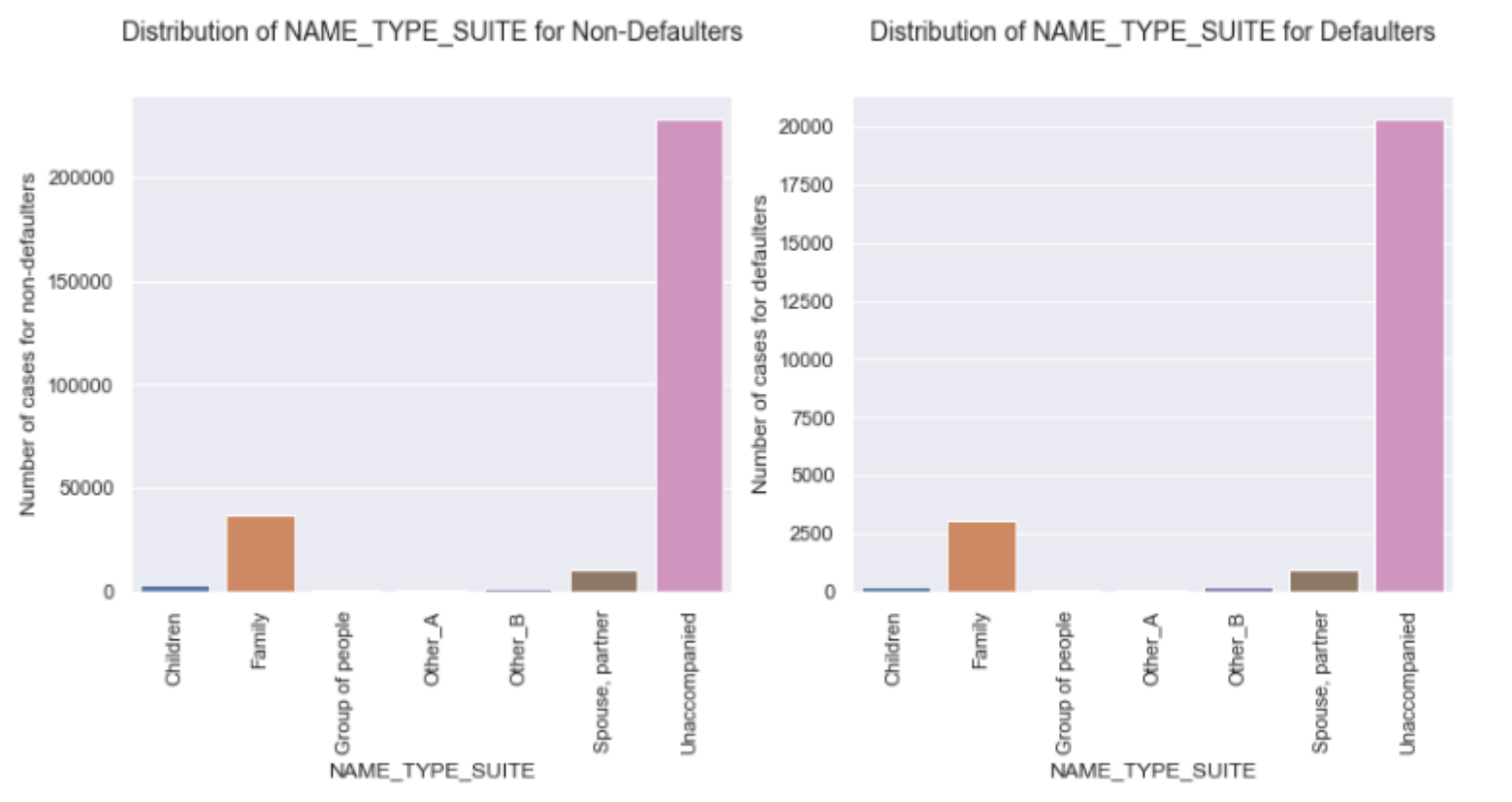
NAME_CONTRACT_TYPE



Interfernce:

We can notice that revolving loans are lesser in the defaulted population. Hence we can infer that revolving loans have comparatively safer. This may be attributed to the Nature of revolving loan as it is considered a flexible financing tool due to its repayment and re-borrowing flexibility

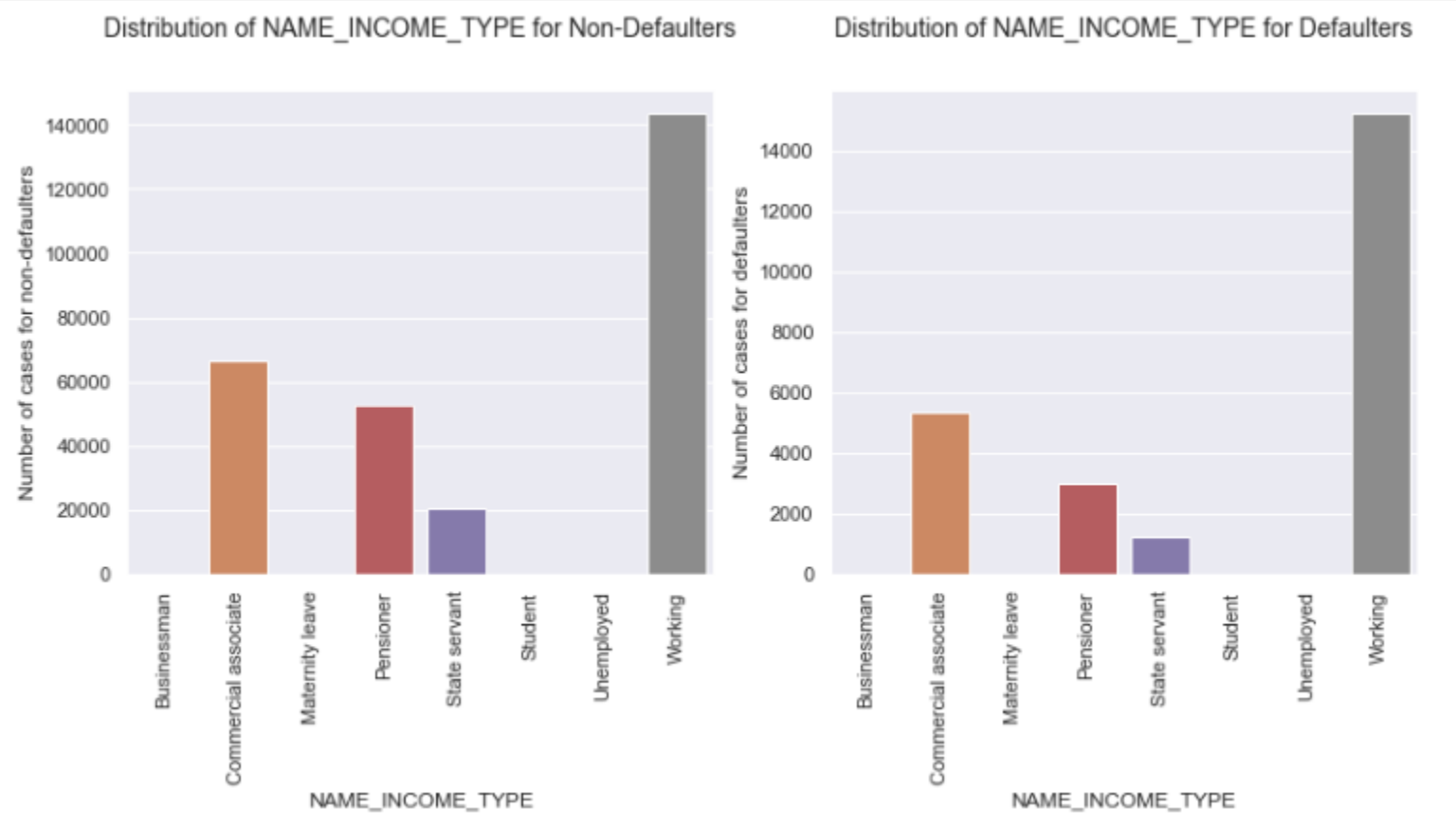
NAME_TYPE_SUITE



Interfernce:

Who was accompanying client when he was applying for the loan does not have any impact on the default. Both populations have same proportions.

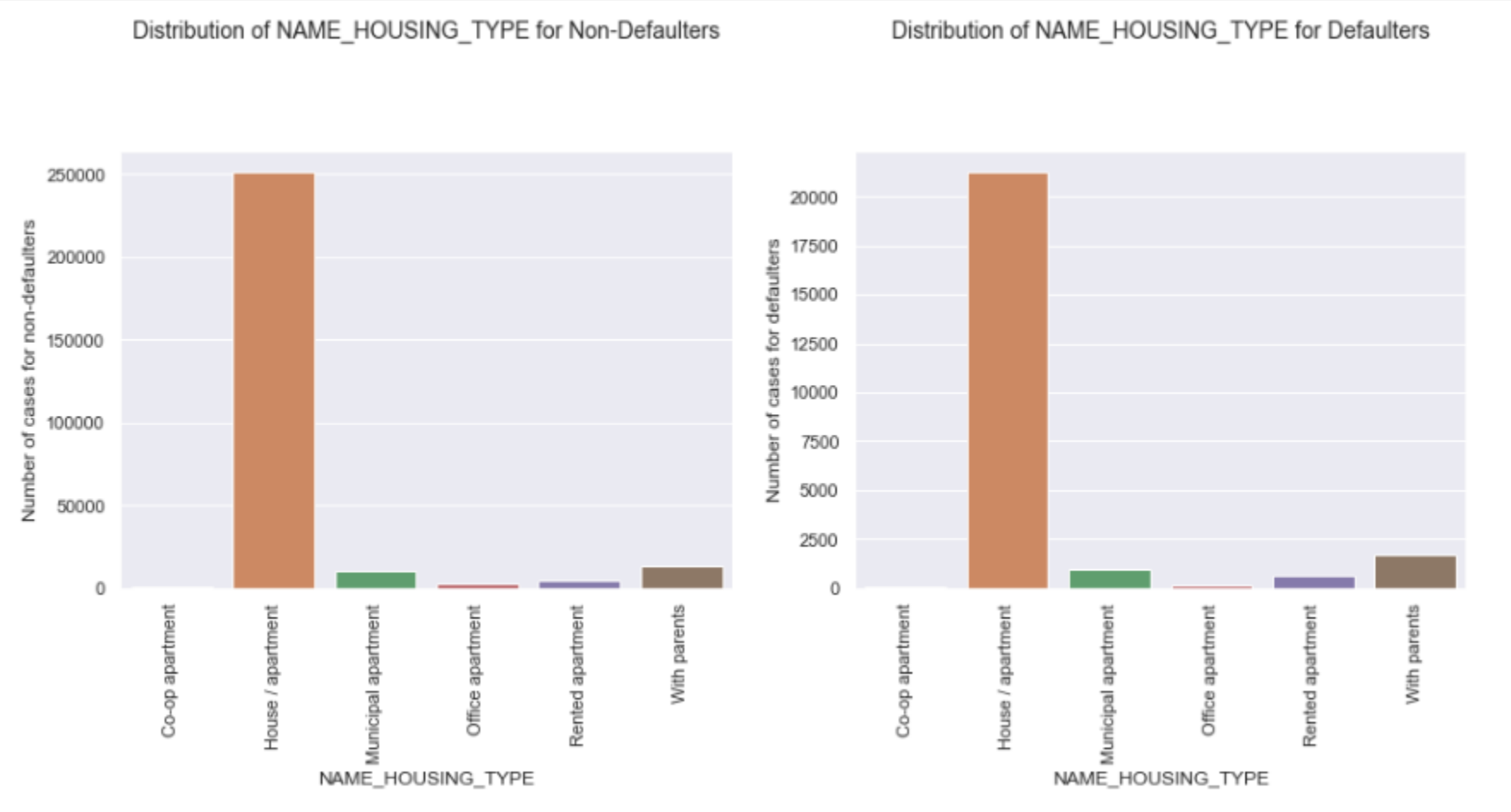
NAME_INCOME_TYPE



Interfernce:

1. Most of the defaults are from Working population

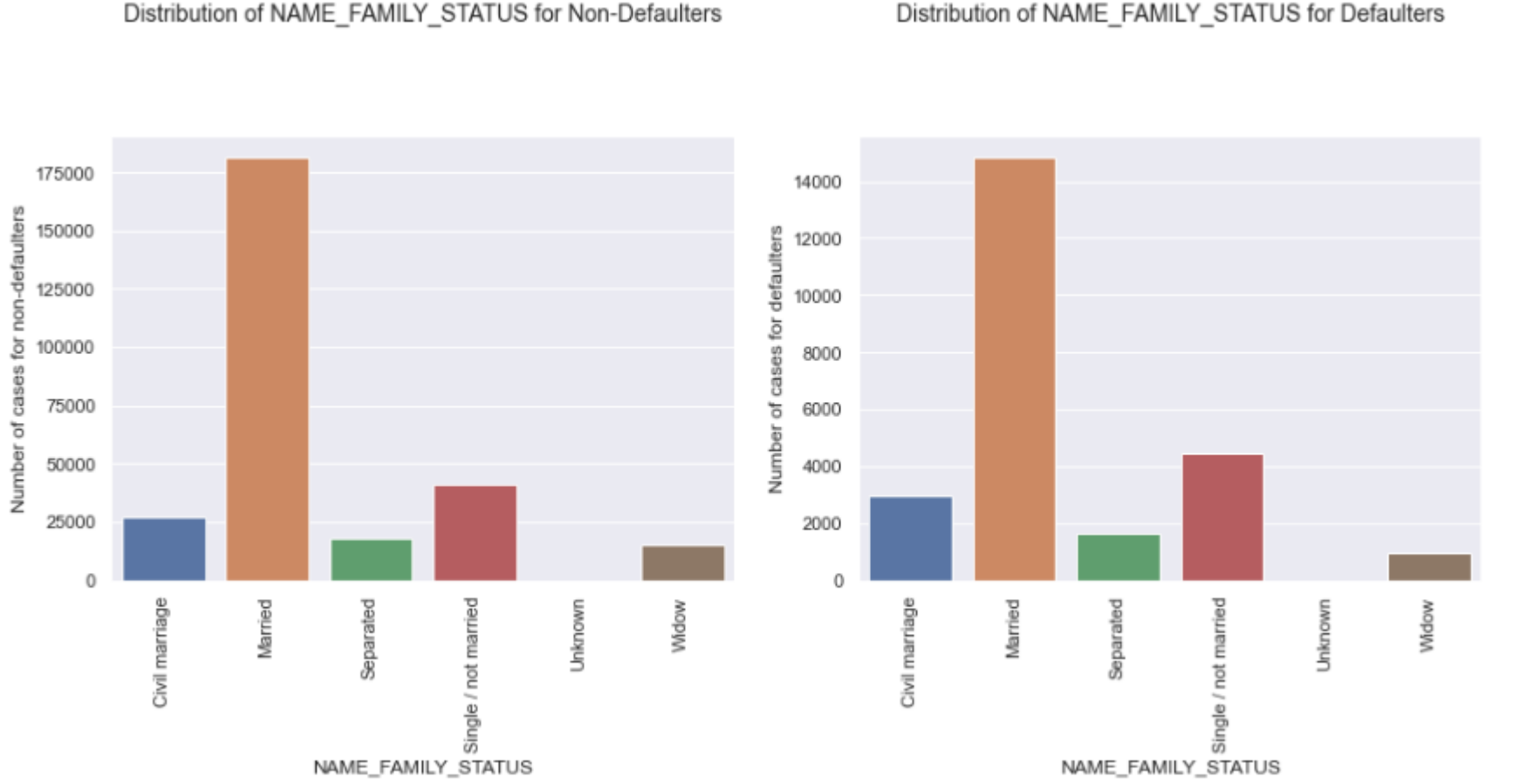
NAME_HOUSING_TYPE



Interference:

1. Population living in Rented apartments and those living with parents have higher default rate as they have higher proportion in the Defaulted population as compared to non defaulted population.
2. Living in rental apartment means a cash outfow towards rent and thus less cash left for repayment of loan.
3. Living with parents may suggest that the incme is not too high anf thus difficulty in repayment of loan.

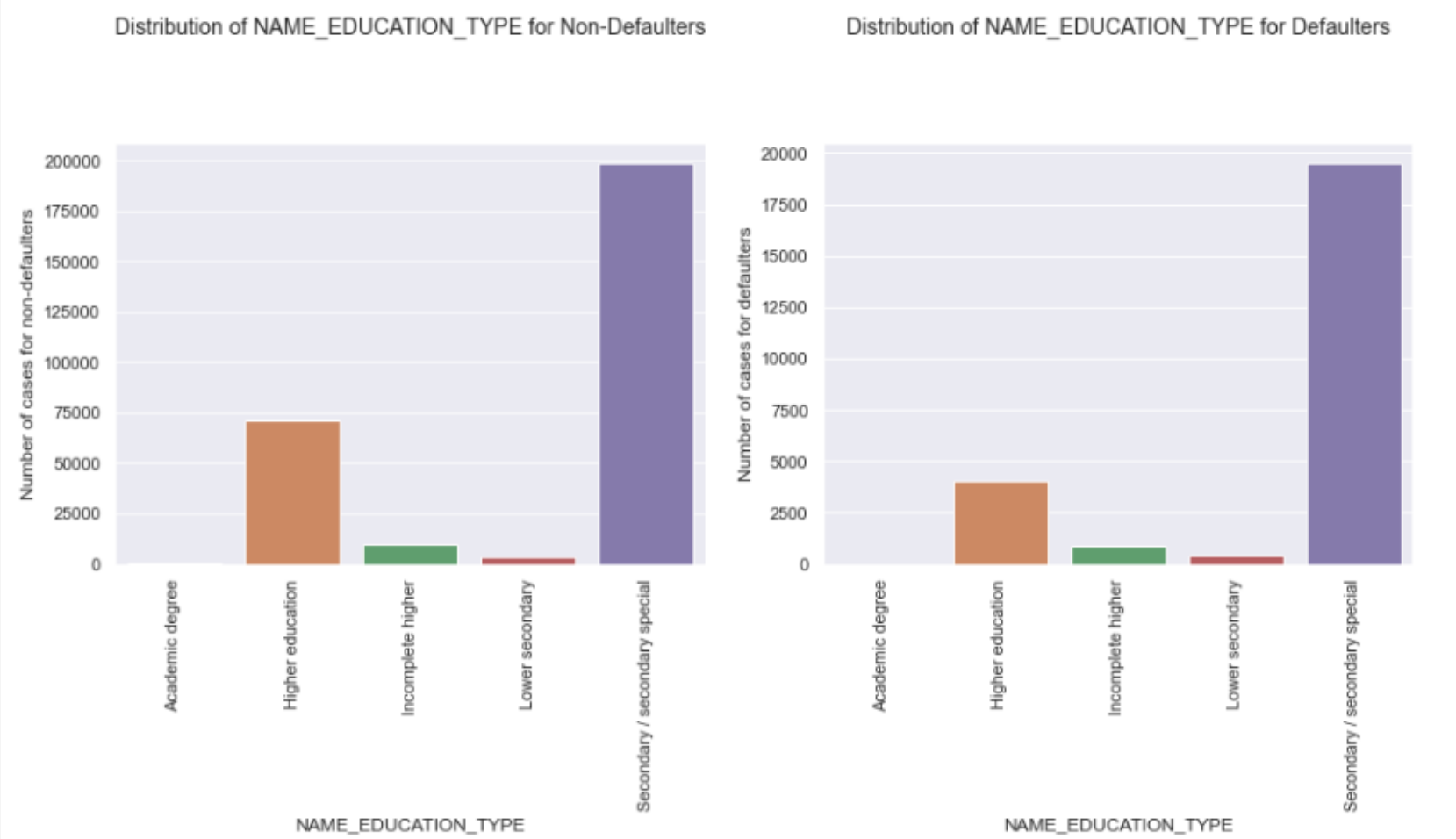
NAME_FAMILY_STATUS



Interfernce:

1. Single/ not married is proportionally higher in dafaulted population as compared to non defaulted population
2. This shows that Single applicants have higher defaults.

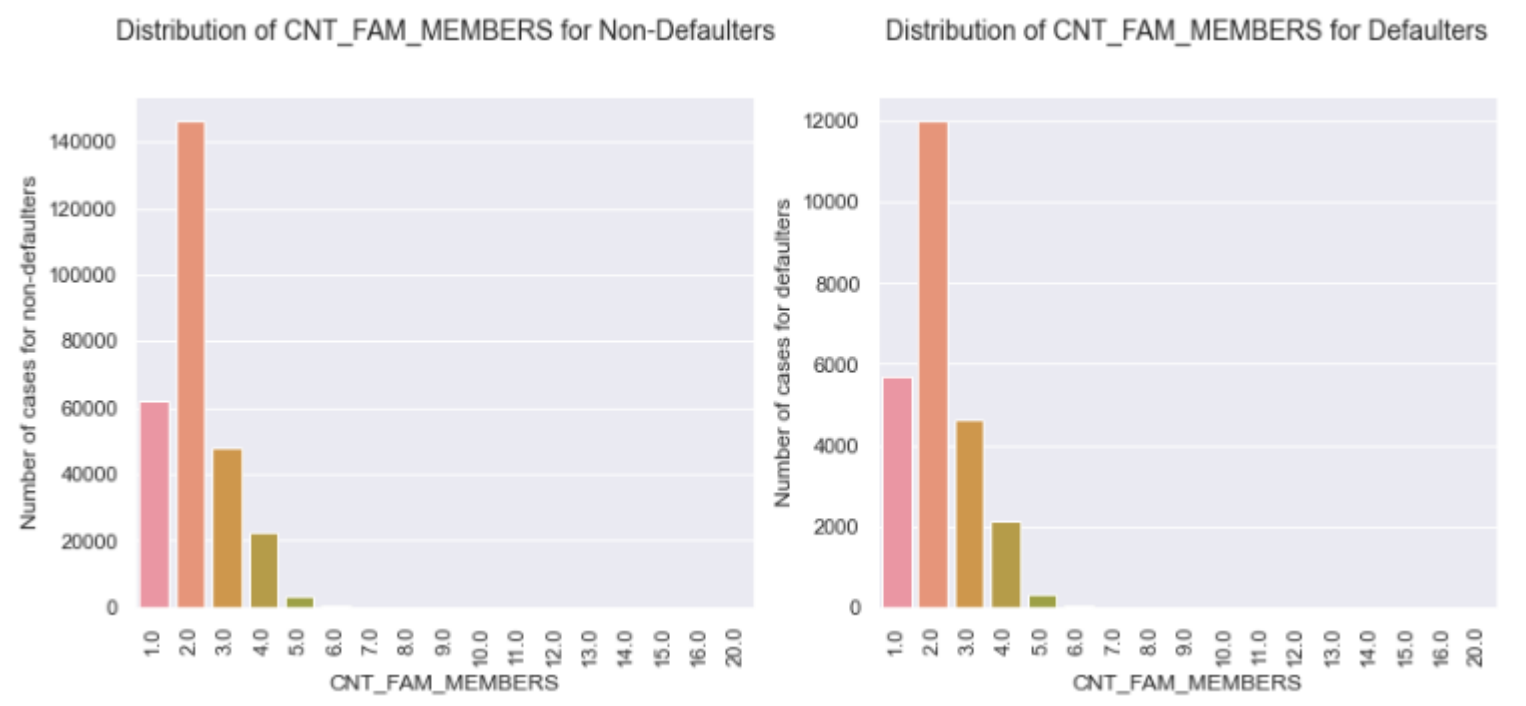
NAME_EDUCATION_TYPE



Interfernce:

1. Higher education count is proportionally lesser in dafualted population as compared to non defaulted population. Hence Higher the education level, lower the default rate. This is logical as higher degree category should be earning more and hence easier to pay off loan installments.
2. customer having payment difficulties in secondary/ secondary special in both the cases.

CNT_FAM_MEMBERS



Interfernces:

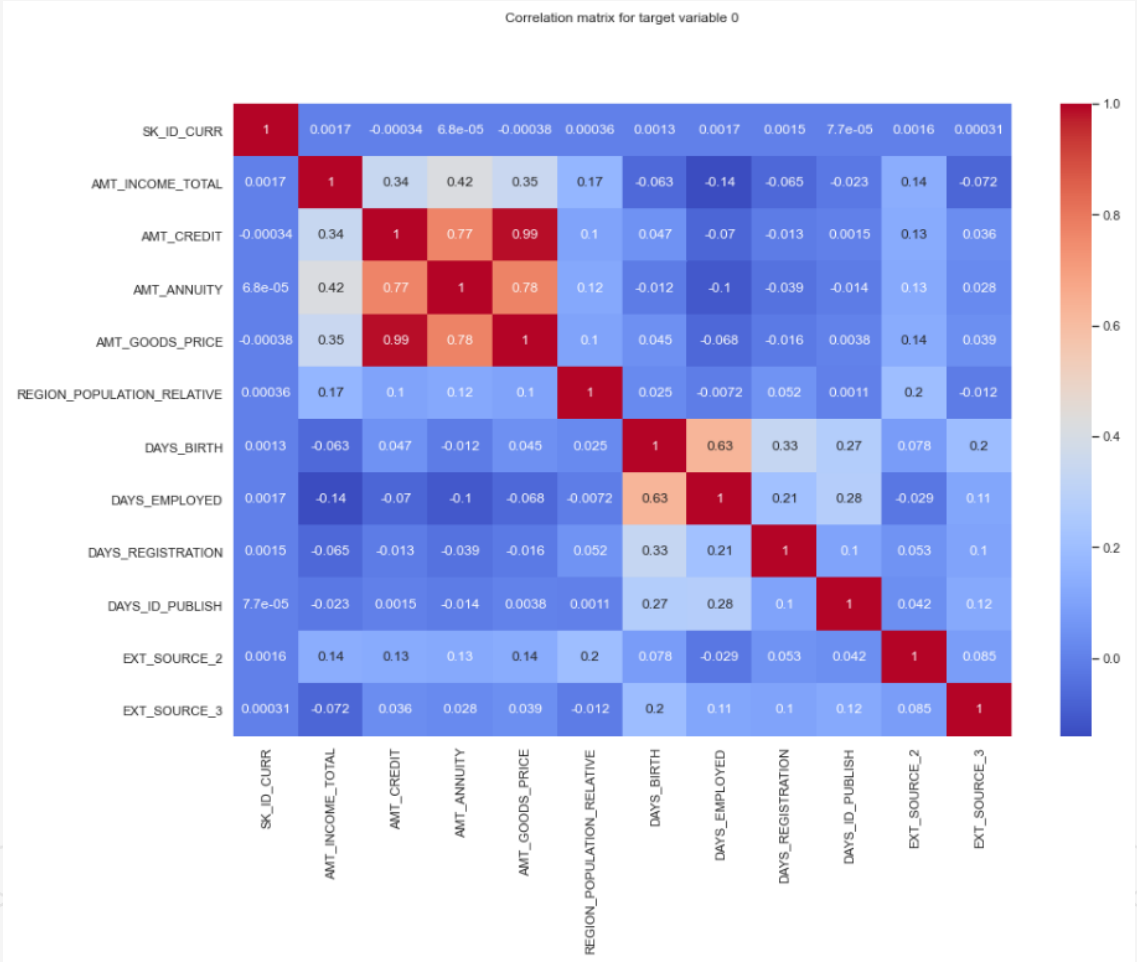
1. Applicants having 2 members in family i.e have no childrens, have higher number of loan applications.
2. Applicants with more than 8 members in family have higher chances of not returning their loans.

Rating2

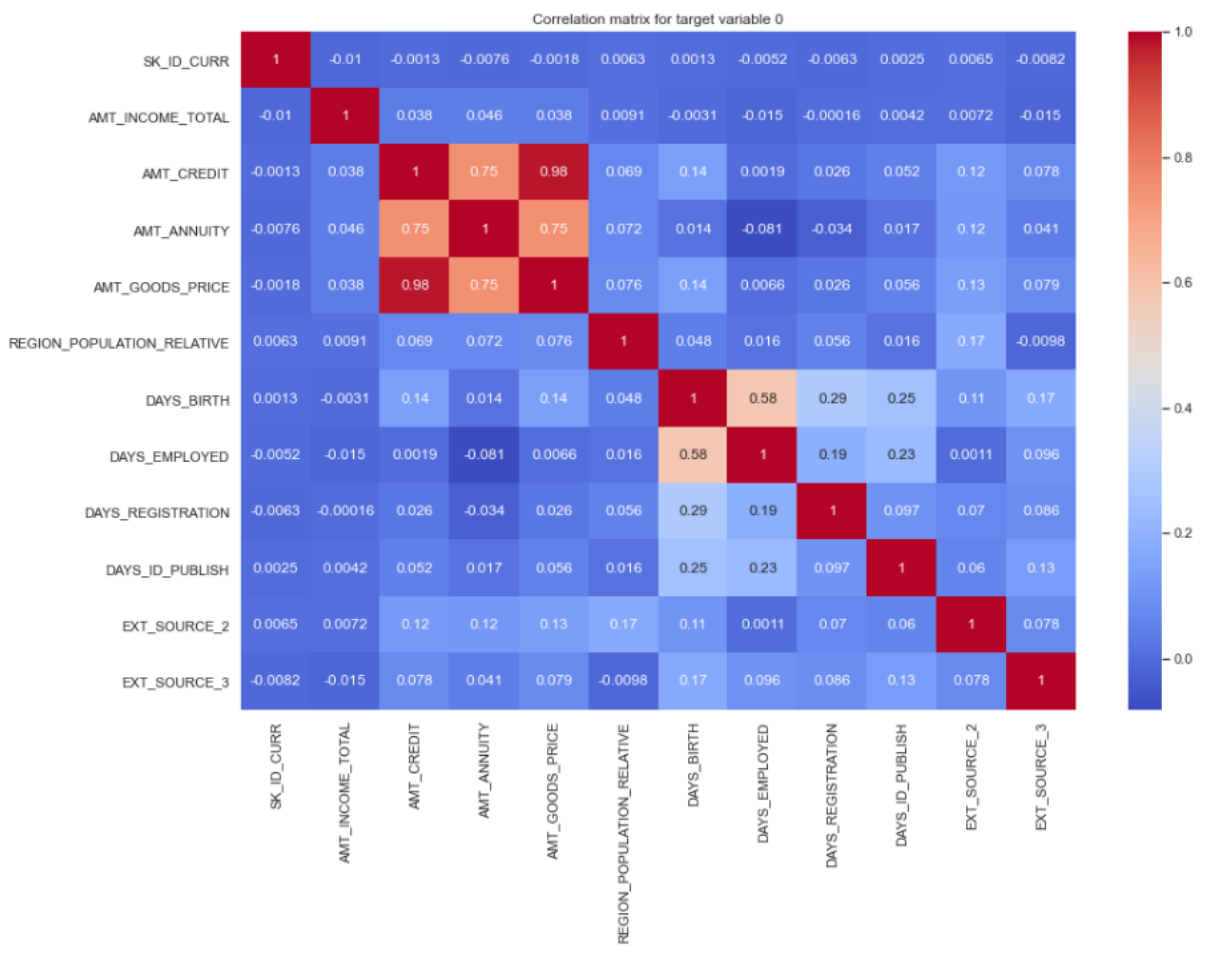
Interfernce:

1. Low income range has higher defaults as their proportion in defaulted population is higher than in the non defaulted population

Correlation matrix for target variable 0



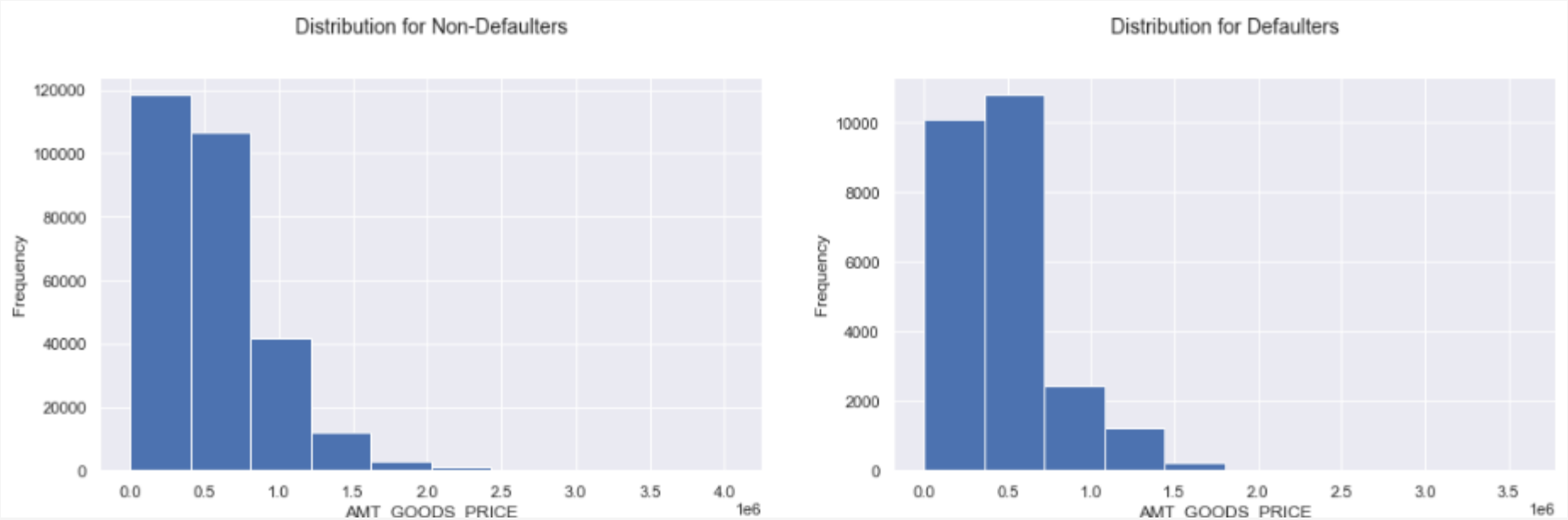
Correlation matrix for target variable 1



Interfernces:

1. If AMT_ANNUITY is higher, Credit is also higher
2. If AMT_ANNUITY is higher, goods wth higher price is purchased
3. If AMT_CREDIT is higher, goods wth higher price is purchased
4. CNT_FAM_MEMBERS is directly related on CNT_CHILDREN
5. REGION_RATING_CLIENT is directly propotional to REGION_RATING_CLIENT_W_CITY
6. If the population of the region is higher, the REGION_RATING_CLIENT is lower
7. If DAYS_BIRTH increases, days_employed increases

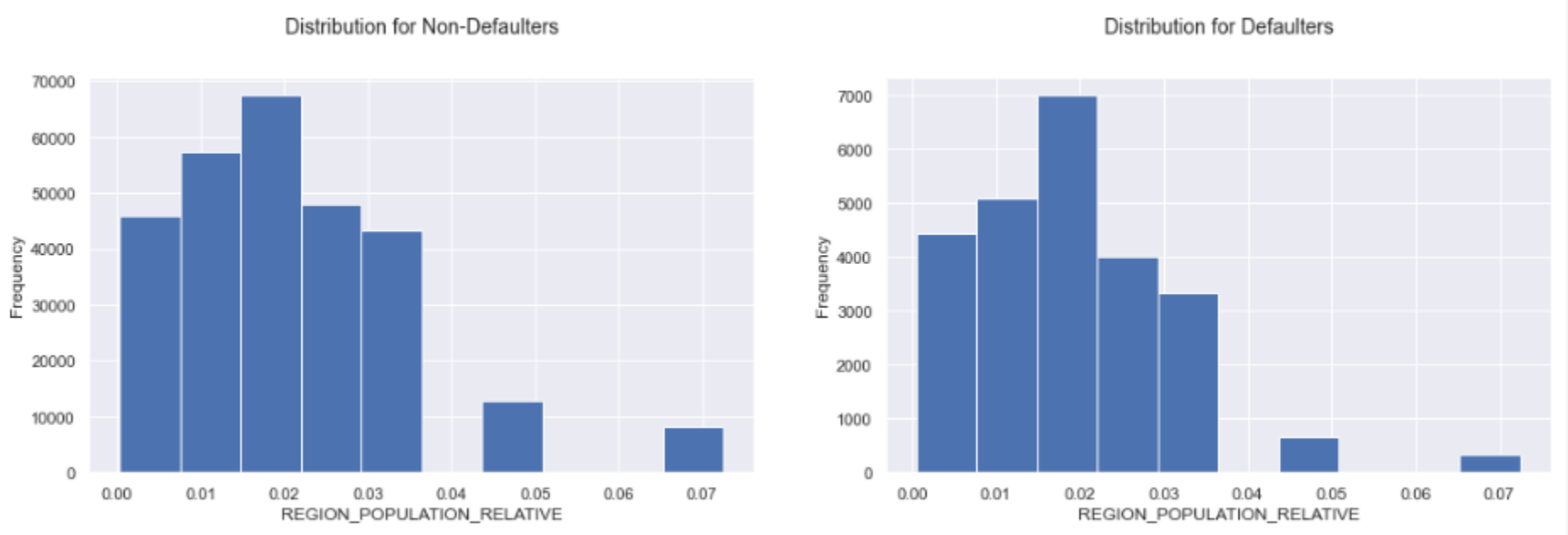
AMT_GOODS_PRICE



Interfernces:

1. Range in Non-defaulters and defaulter in minimum range is high

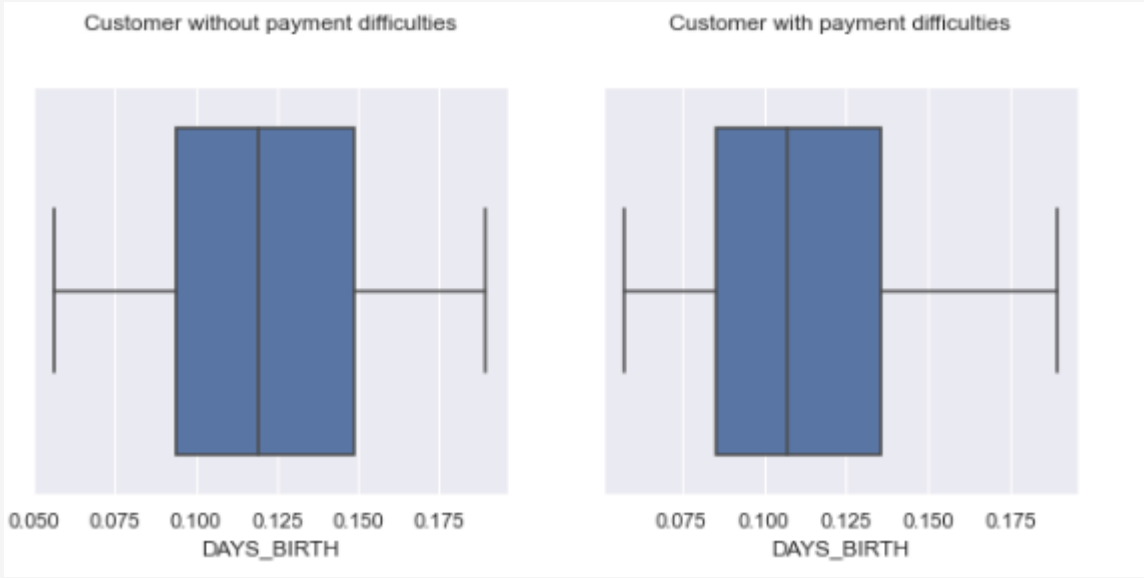
REGION_POPULATION_RELATIVE



Interfernce:

1. A. Range in Non-defaulters and defaulter in population is almost same.

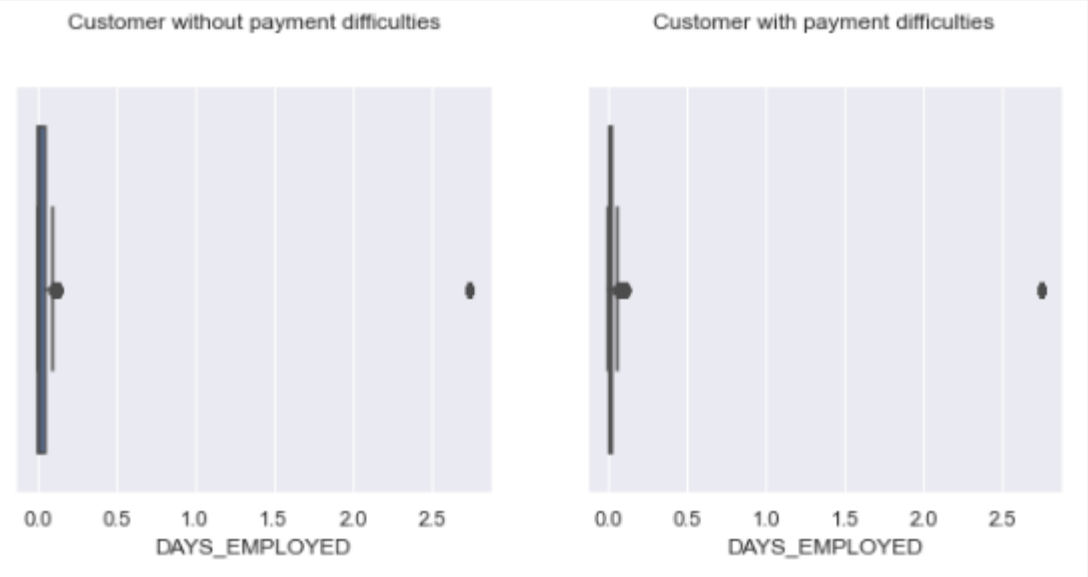
Univariate Analysis for numerical variables [DAY_BIRTH]



Interfernce:

1. customer without payment difficulties having year in between 34 to 54 years.
2. And coustomer with payment difficulties having in between 31 to 50 years

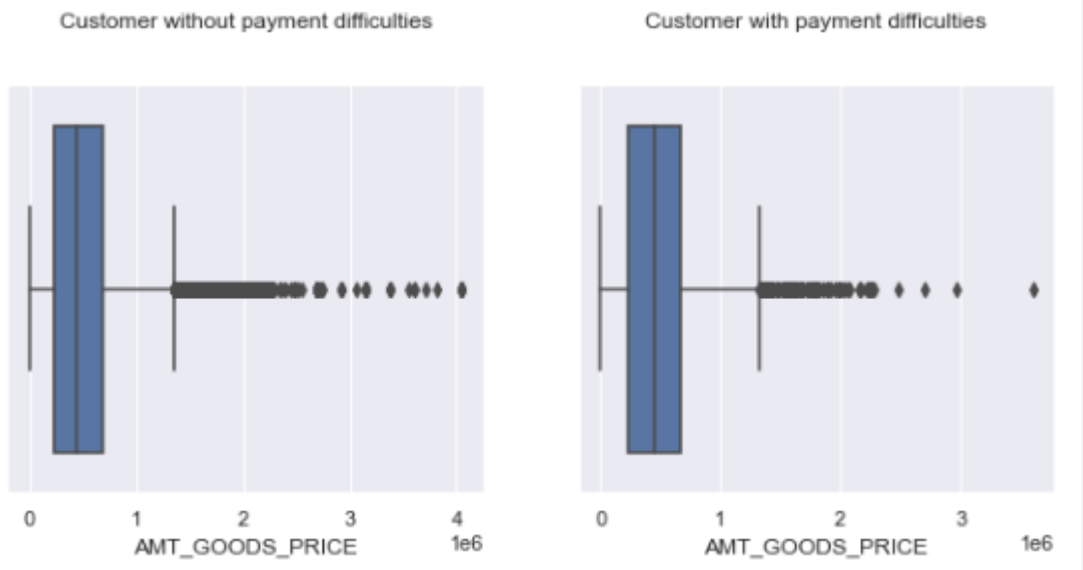
DAYS_EMPLOYED



Interfernce:

1. customer without payment difficulties having year in between 34 to 54 years
2. And coustomer with payment difficulties having in between 31 to 50 years.

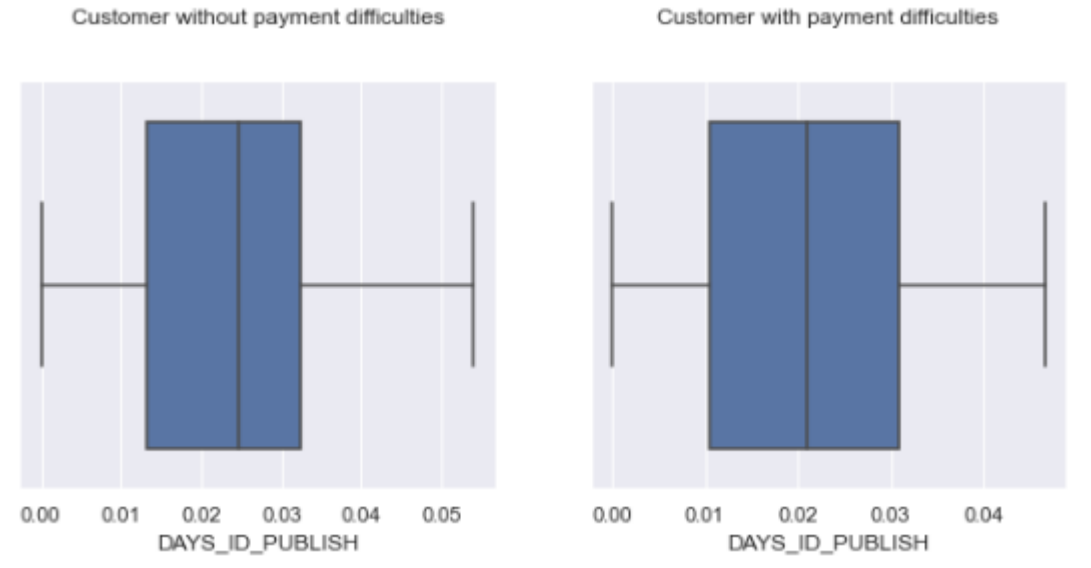
AMT_GOODS_PRICE



Interfernce:

1. customer without payment difficulties lies in between 0.3 to 0.7
2. the customer with payment difficulties lies in between the same as of the without payment 0.3 to 0.7.
3. both are having the mid value about 0.5.

DAYS_ID_PUBLISH



Interfernce:

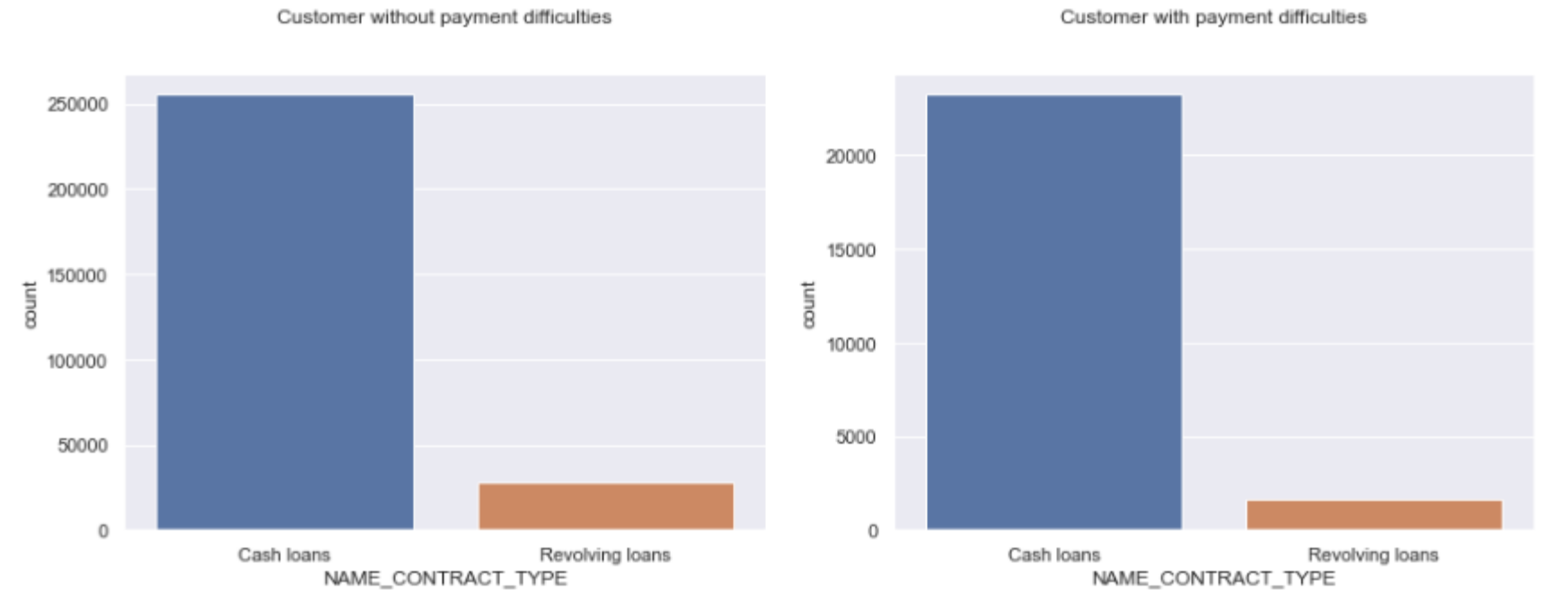
1. customer without payment difficulties lies in between 5 to 11.
2. customer with payment difficulties lies in between 3 to 11

AMT_ANNUITY

Interfernce:

1. maxinum number of defaulters have Low_annuity Values, while maximum number of non-defaulters have high annuity

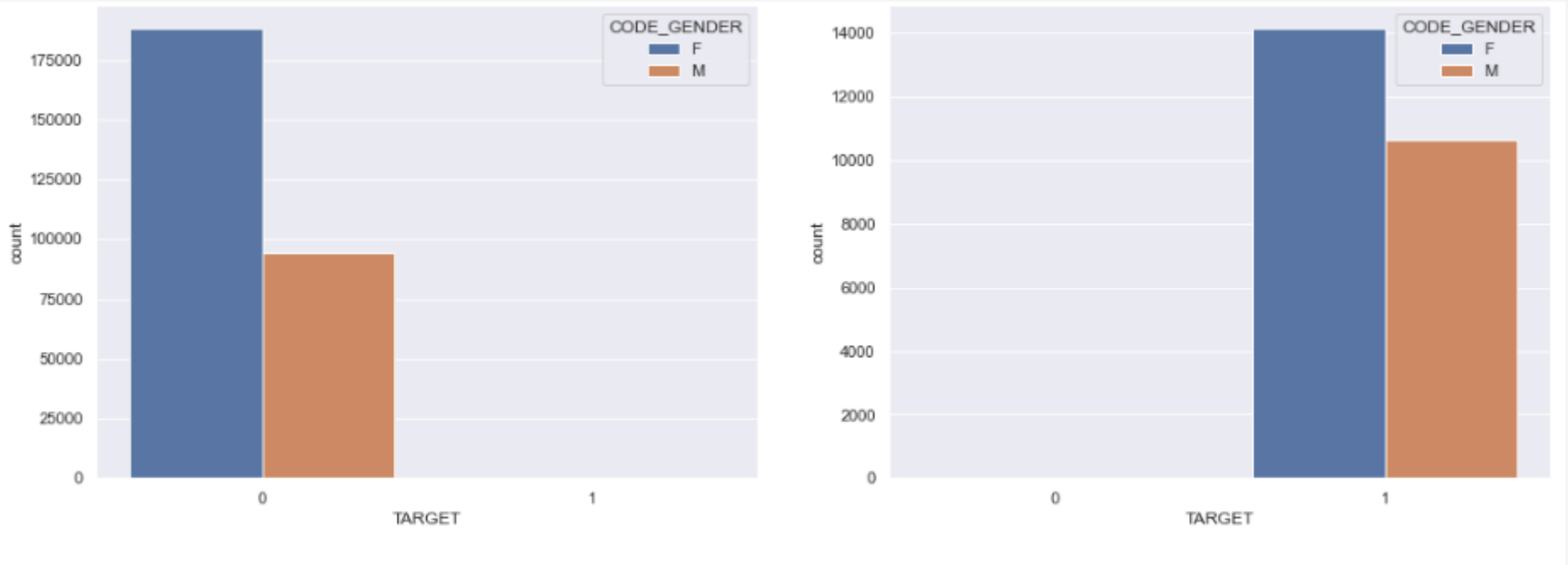
For Categorical Variables [NAME_CONTRACT_TYPE]



Interfernce:

1. customer without payment and customer with payment difficulties both are taking cash loans

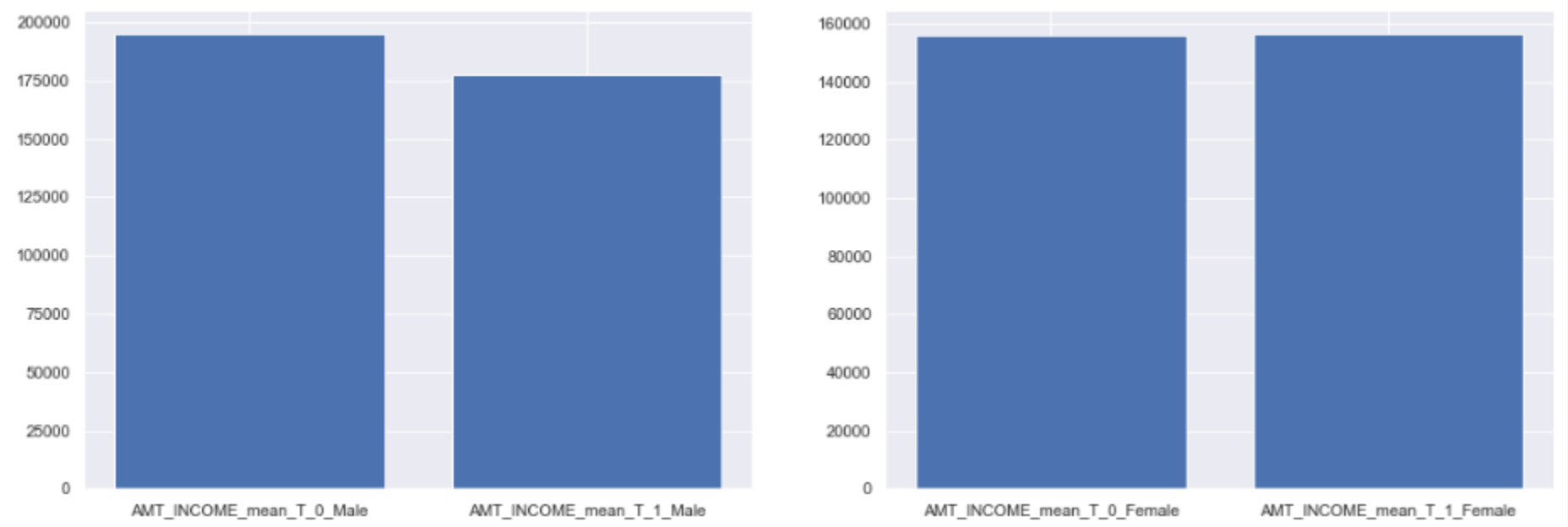
Segmented Analysis Male vs Female [Target]



Interfernce:

1. Number of female applicants is almost double the number of male applicants.
2. Males have a higher chance of not returning their loans (~~10%), comparing with women~~ (7%)

AMT_INCOME_mean_T_1_Male and Female



Interfernce:

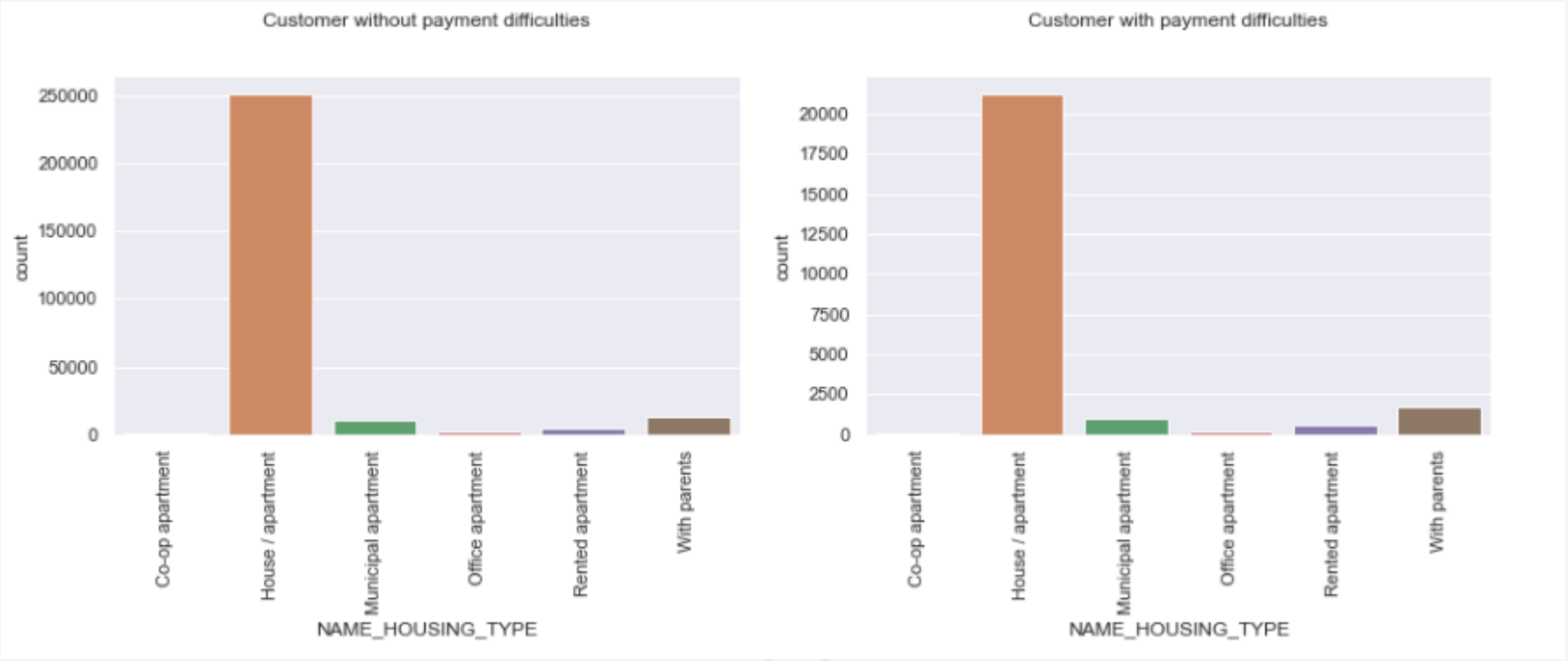
1. Mean among male defaulters do have less income compared to non-defaulters.

NAME_EDUCATION_TYPE

Interfernce:

1. customer having payment difficulties in secondary/ secondary special in both the cases.
2. Applicants having Lower secoondary education have higher chances of not returning their loans.

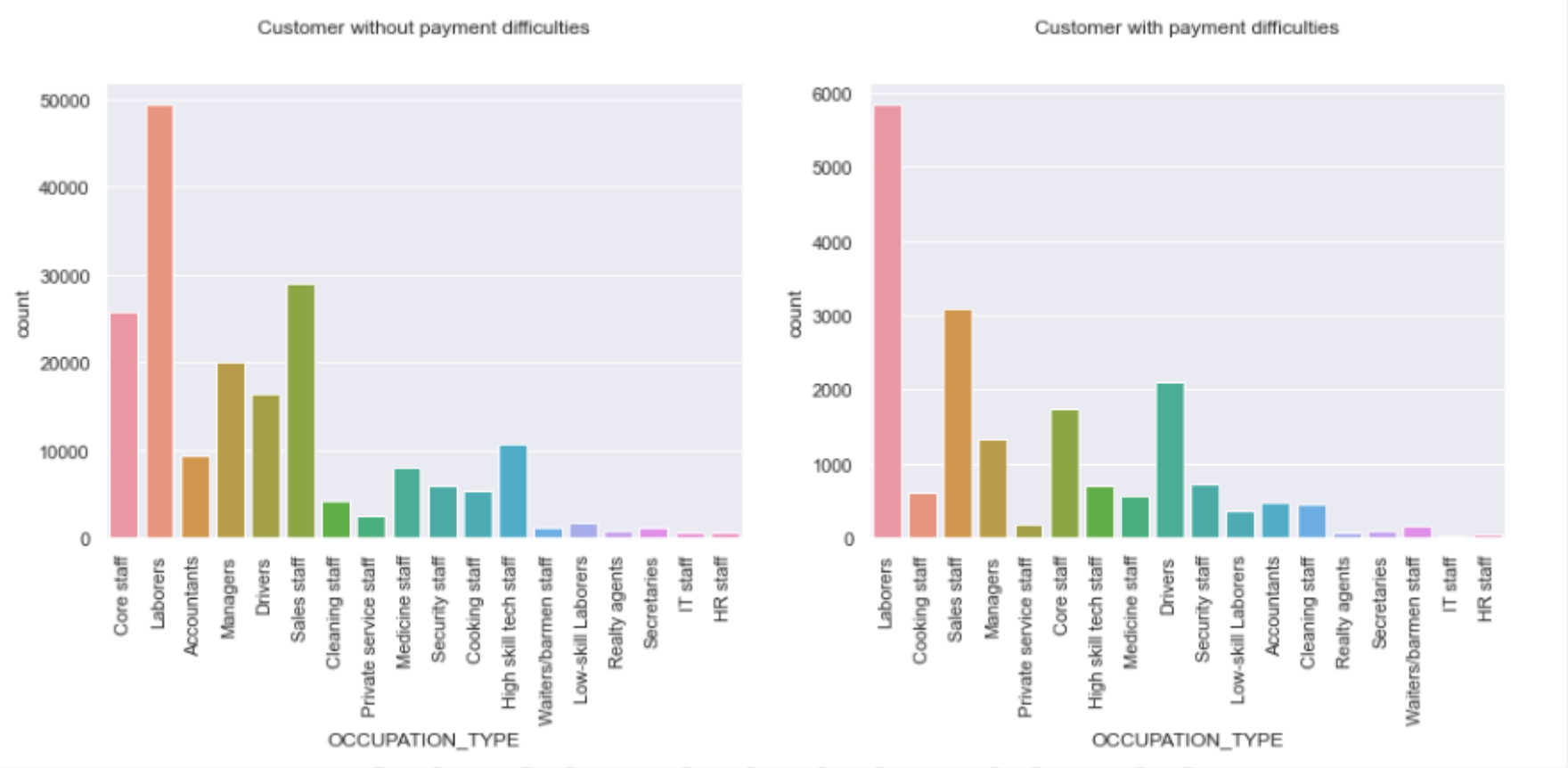
NAME_HOUSING_TYPE



Interfernce:

1. Applicants living in 'House/apartment' have higher number of loan applications.
2. Applicants living in 'Rented apartment'/'With parents' have higher chances of not returning their loans.

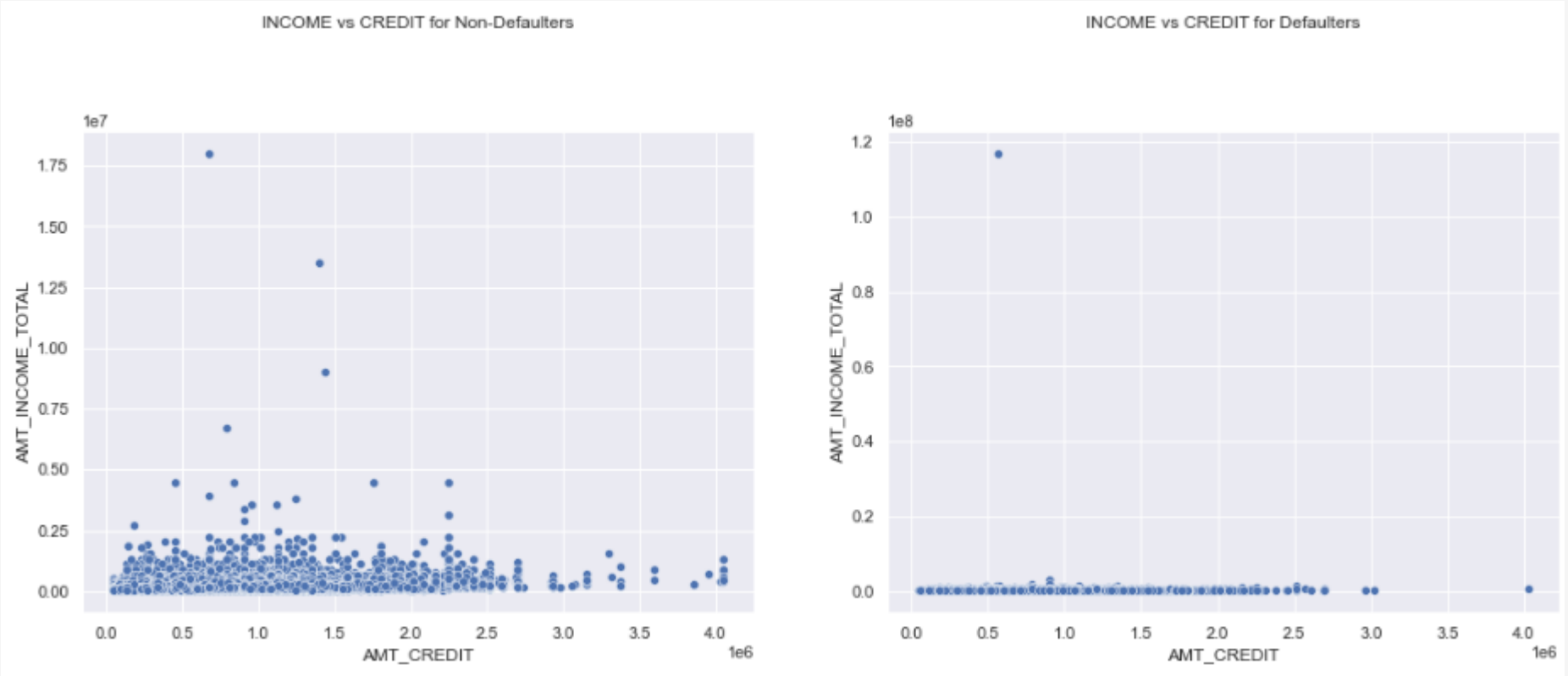
OCCUPATION_TYPE



Interfernce:

1. laborers are having more difficulties in repaying the loan and also the core staff and the sales staff.
2. laborers those who have without payment is way more then with having the payment.

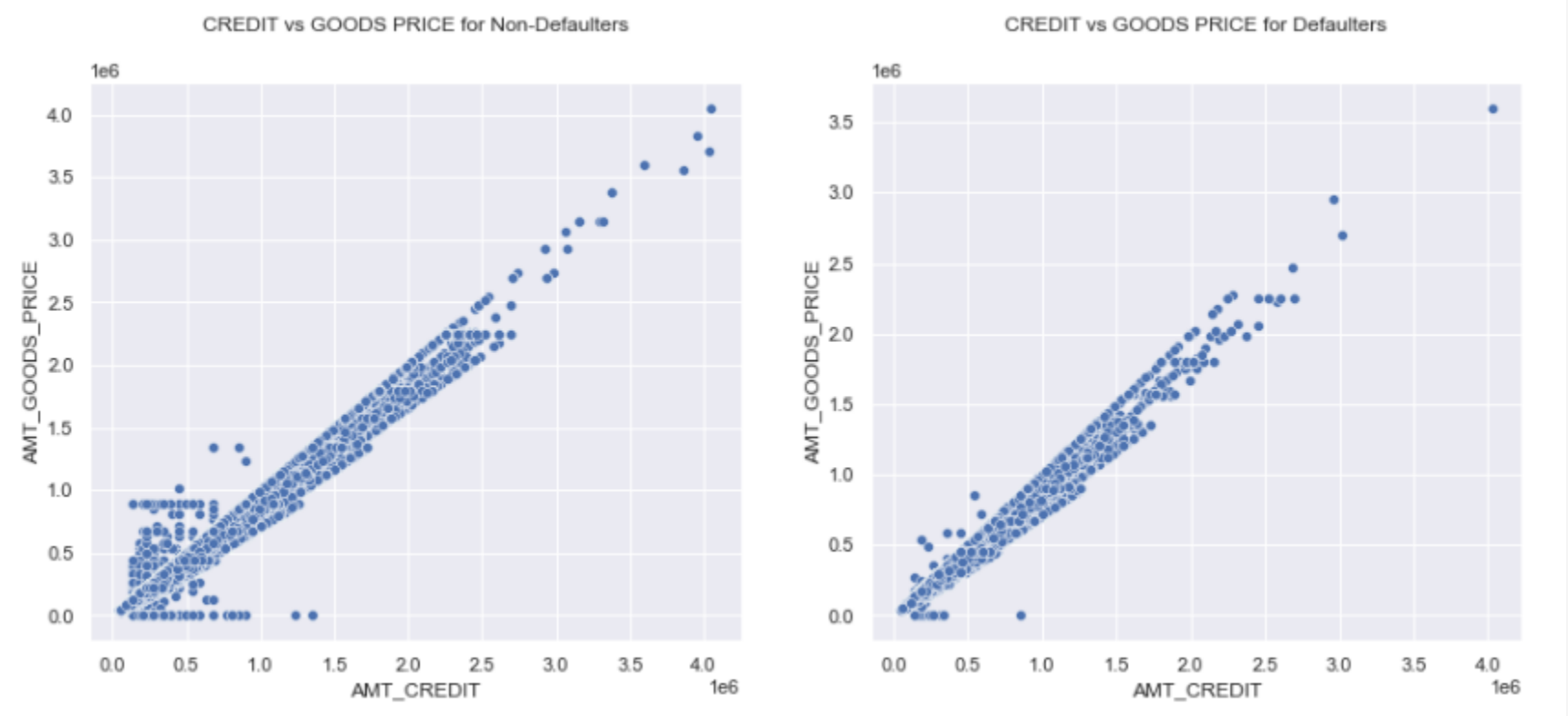
Bivariate Analysis (Numerical-Numerical bivariate analysis) INCOME vs CREDIT for Defaulters



Interfdernce:

Lower desity of defaults where income is higher than 300k or credit is lower than 200k.

CREDIT vs GOODS PRICE for Defaulters



Interfernce:

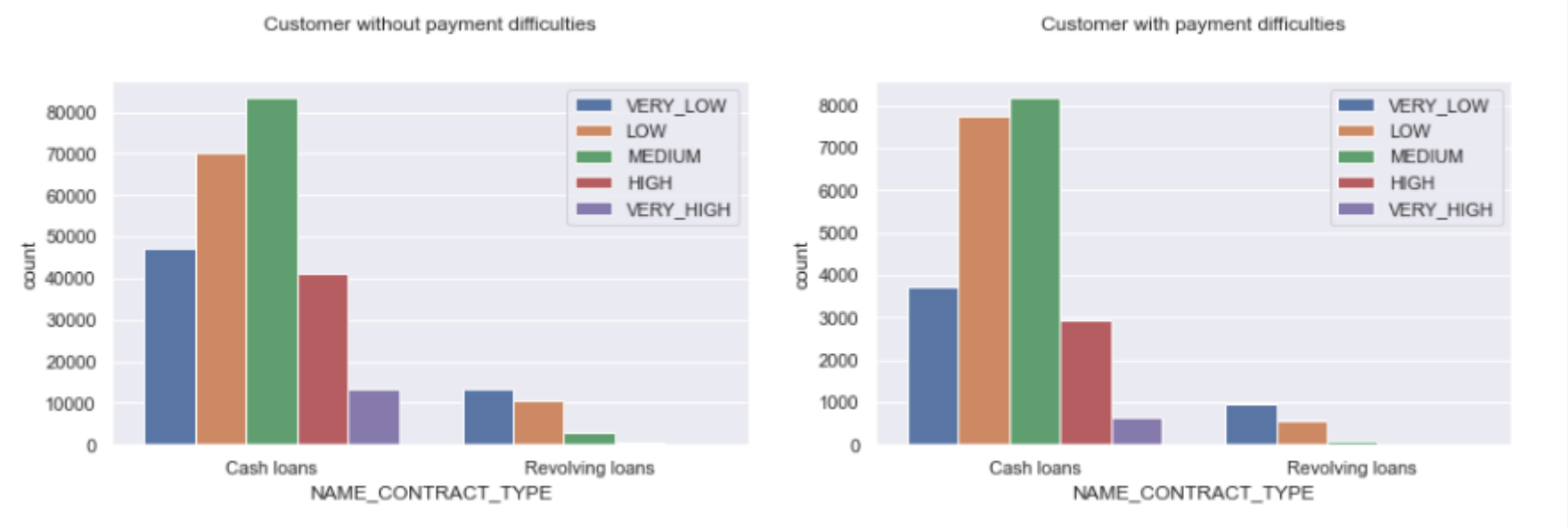
1. Defaulters are less if price of good is upto 500k and amount credit is also less than 500k
2. we can see that goods price is positively correlated with credit amount.

AMT_CREDIT vs AMT_ANNUITY

Interfernce:

People without payment difficuties take more credit for the annuity that they have

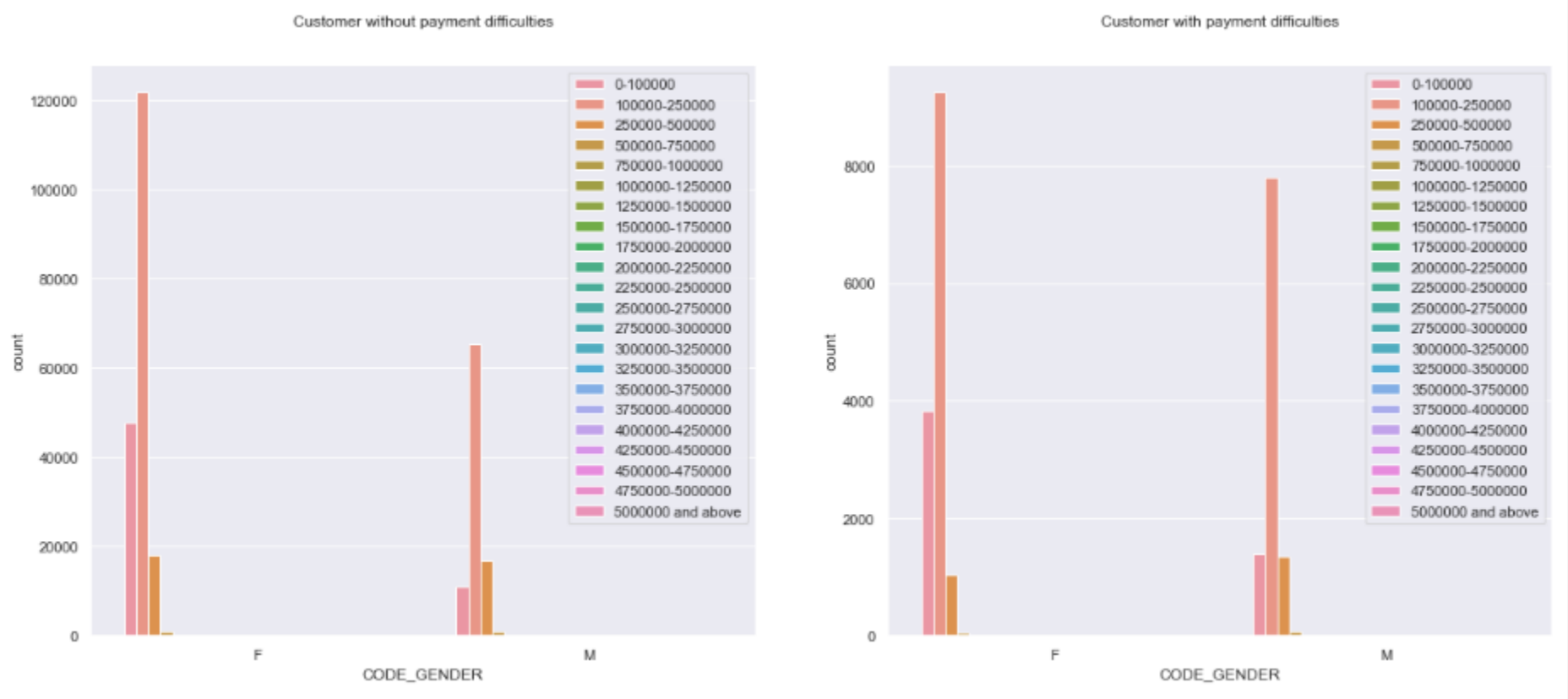
categorical - categorical bivariate analysis NAME_CONTRACT_TYPE



Interfernce:

1. Applicants apply for 'Cash loans' around approximately 11 times greater then 'Revolving loans'.
2. 'Revolving loans' count is very negligible compare to 'Cash loans', but they have comparatively higher chances of not returning loans

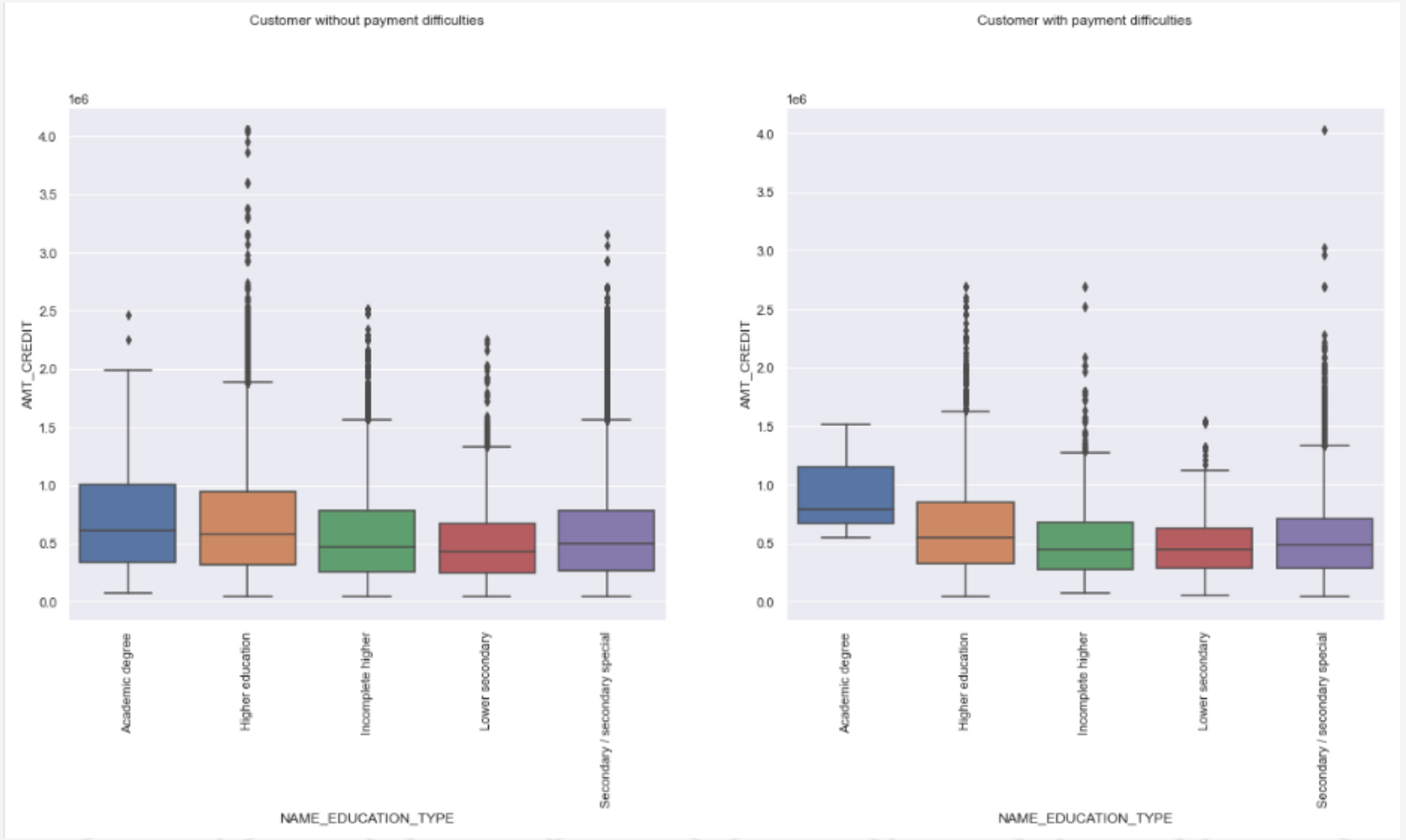
CODE_GENDER



Interfernce:

1. in without difficulties and with payment difficulties female range is higher than male

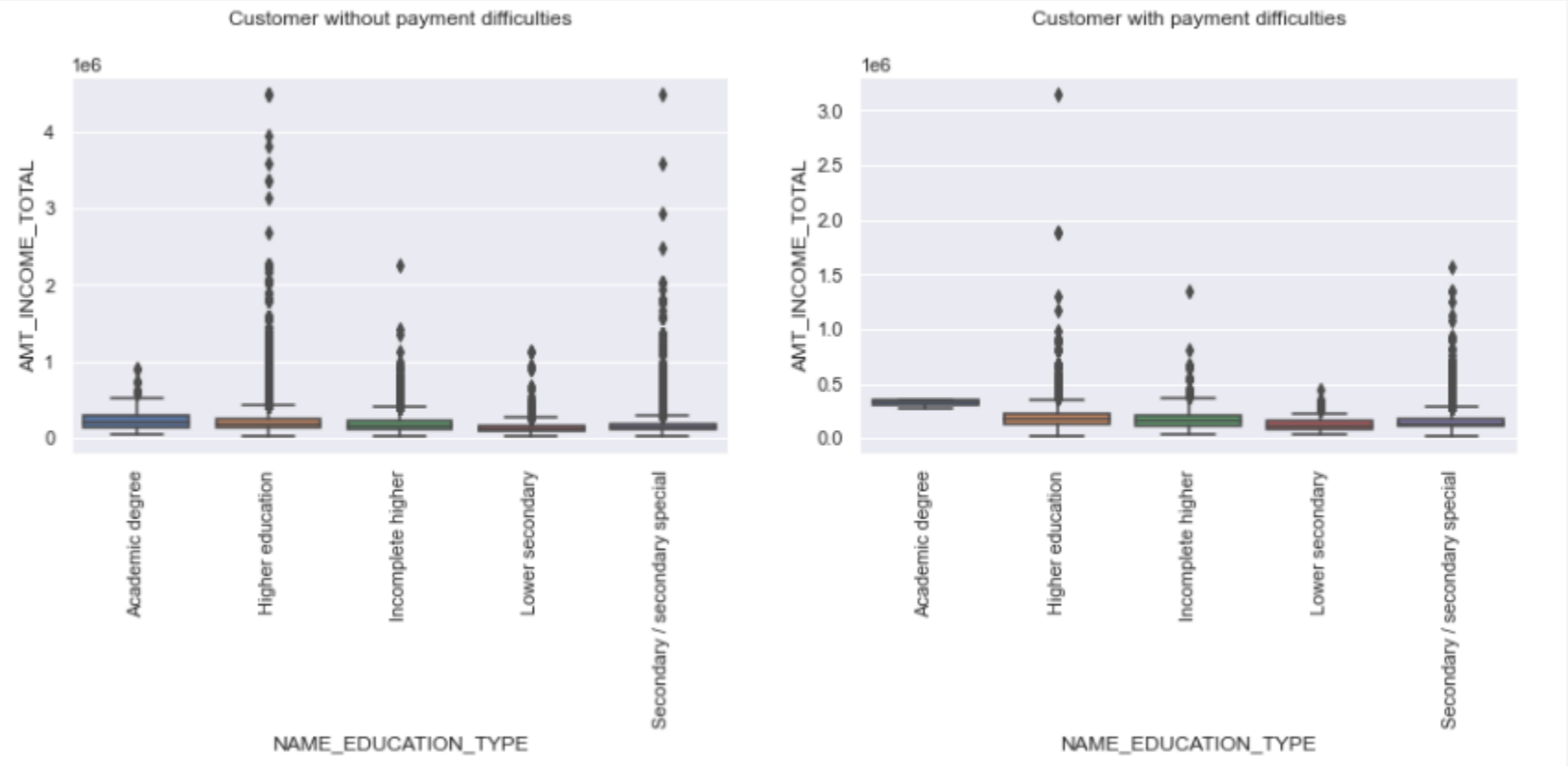
Numerical- Categorical bivariate analysis
AMT_CREDIT vs NAME_EDUCATION_TYPE



Interfernce:

Range of customers without payment of Academic degree is higher than the customer of with payment. And the rest of the Education type is almost same for both the cases.

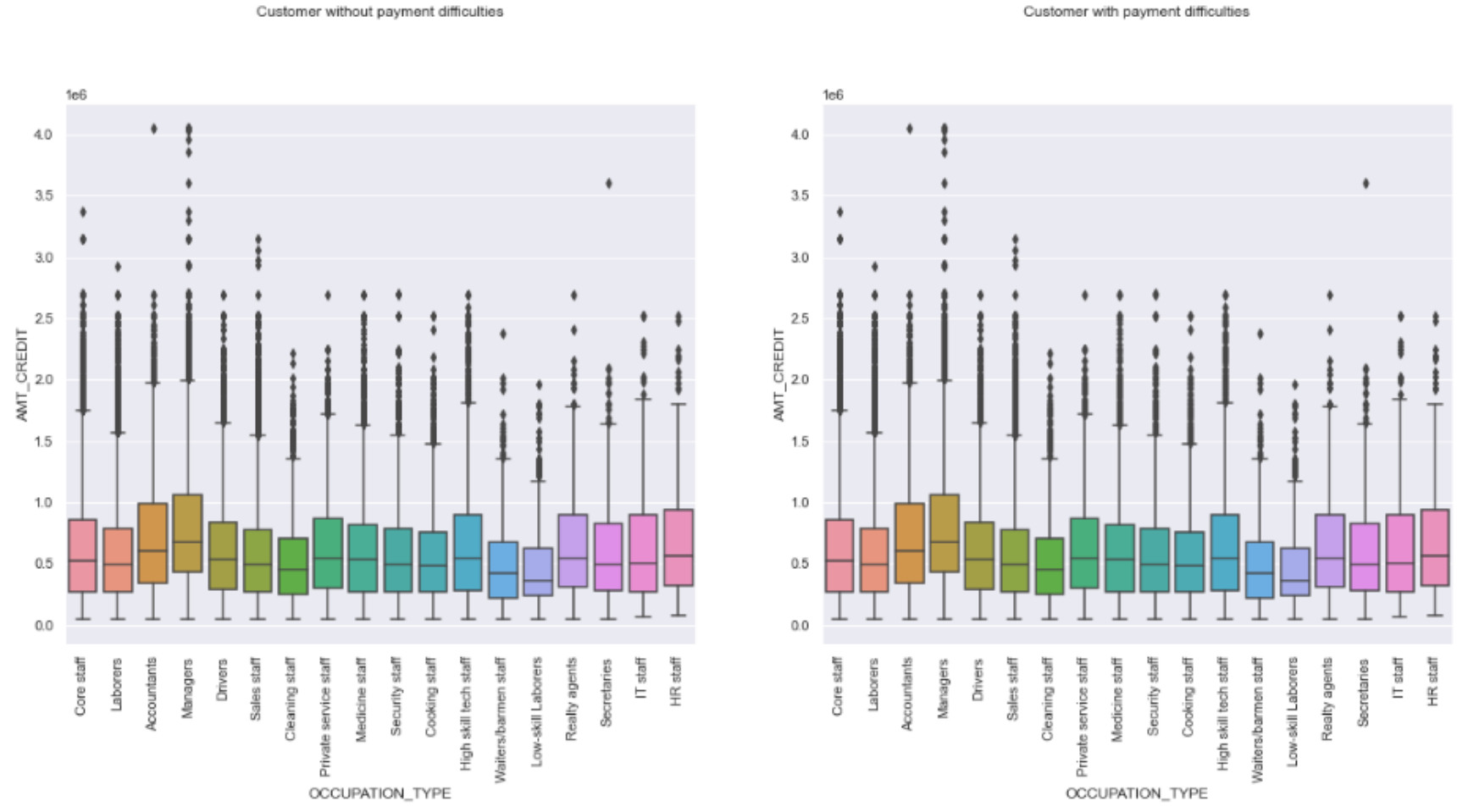
AMT_INCOME_TOTAL vs NAME_EDUCATION_TYPE



Interference:

Customers without payment is having more outliers as compare to the customer with payment.

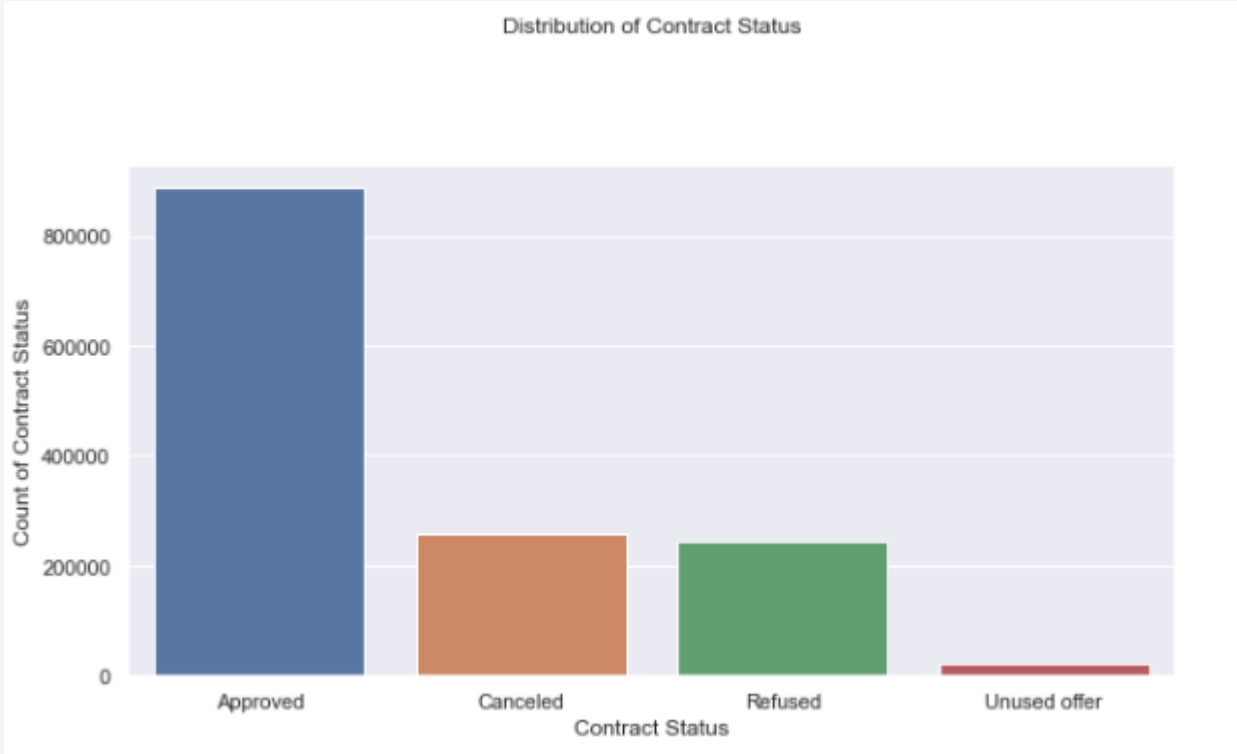
AMT_CREDIT vs OCCUPATION_TYPE



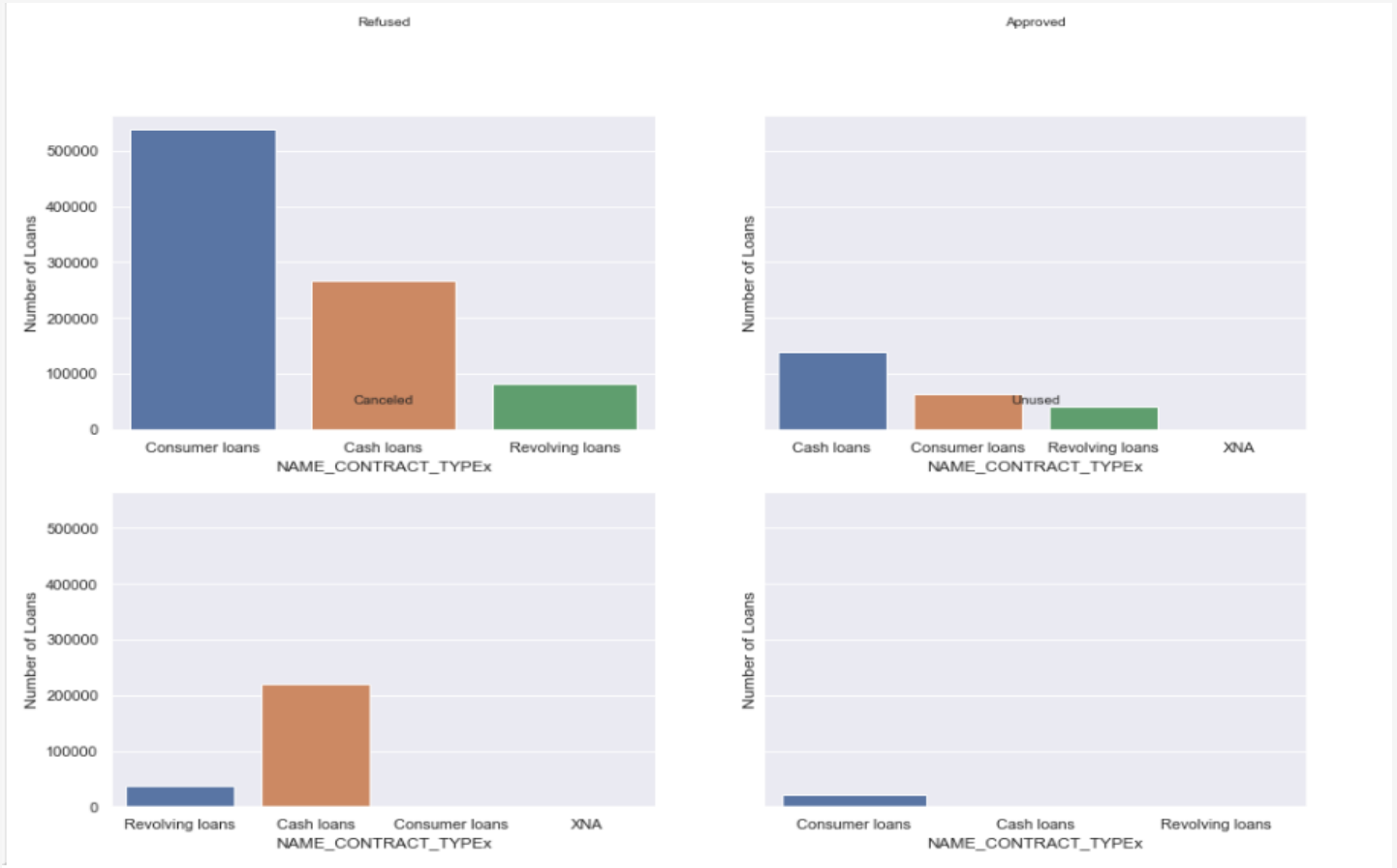
Interfernce:

Range of the customers without payment more as compare to the customers with payment.

previous_application dataframe (Distribution of Contract Status)



Refused,Approved,Cancelled,Unused Categories lone

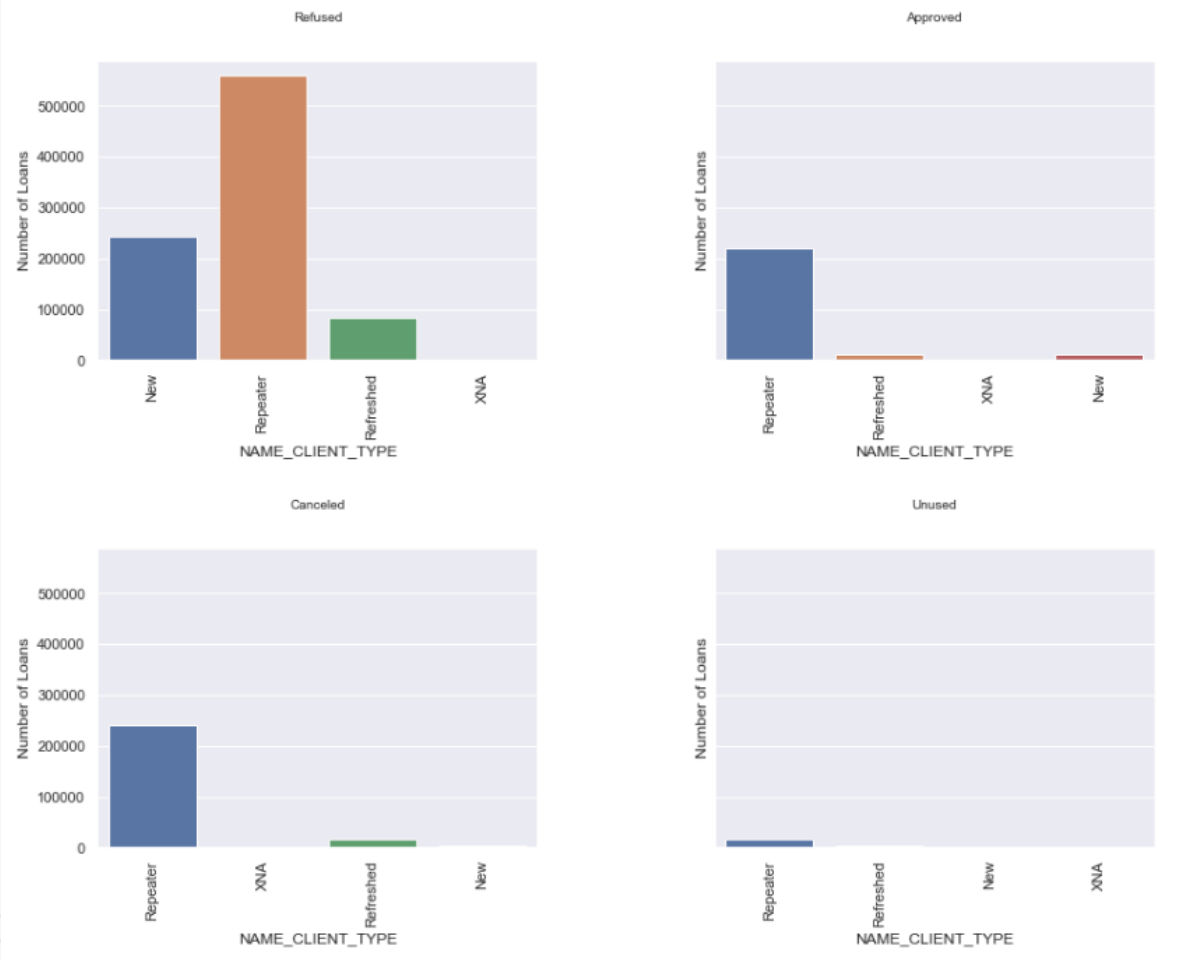


Interfernce:

Revolving loan is much more acceptable as compare to the cash and consumer loans.

As we can see that to visualize 4 plots we wrote same code multiple times. so to avoid redundancy, and to save our time, we will put the above code in a function and generalize it for our following plots, so that its easy to visualize and saves time

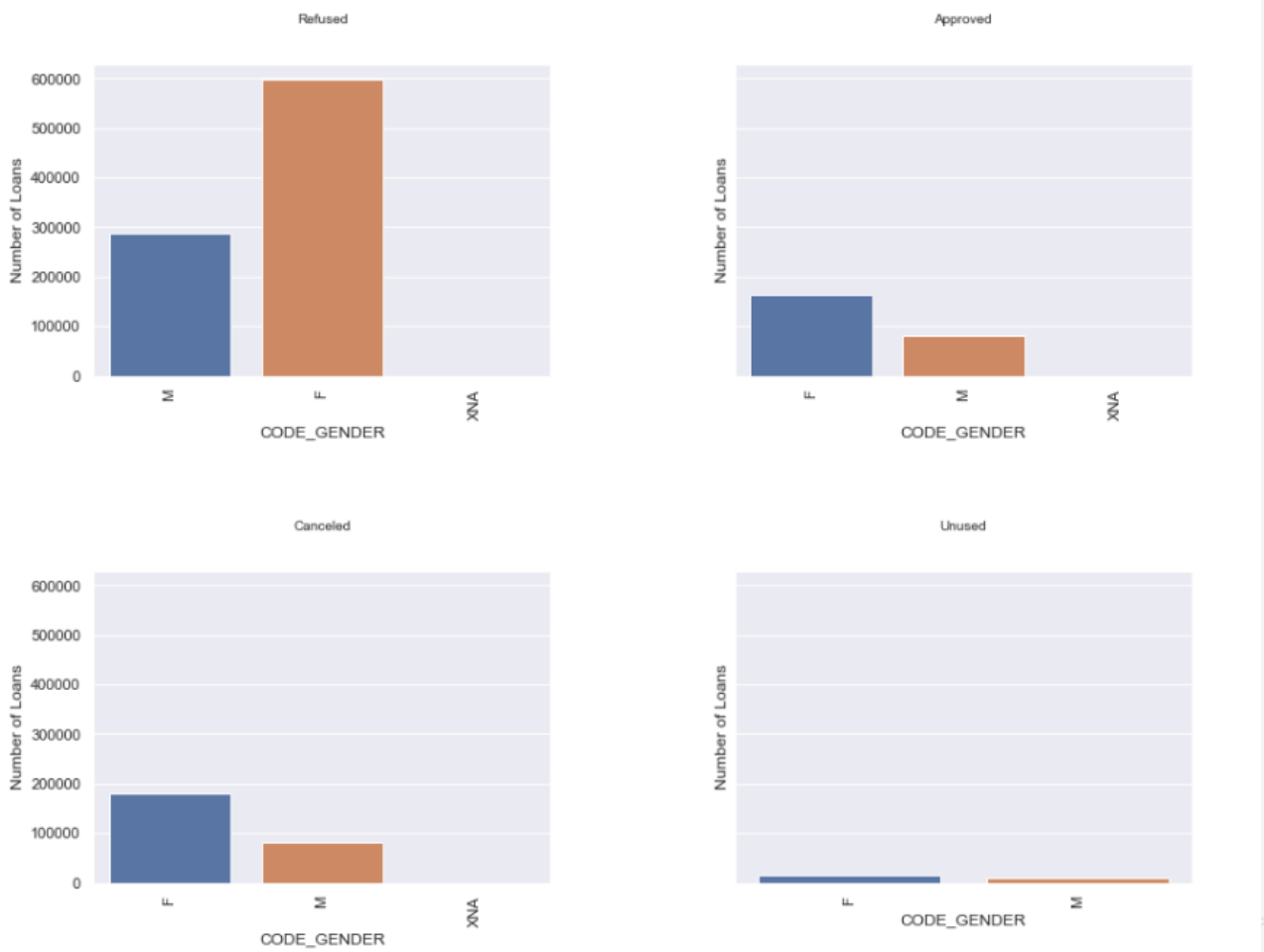
NAME_CLIENT_TYPE



Interfernce:

Repeater is getting more Refused but also we can see that the it also getting more apporved and even that it is getting more canceled and more ususued.

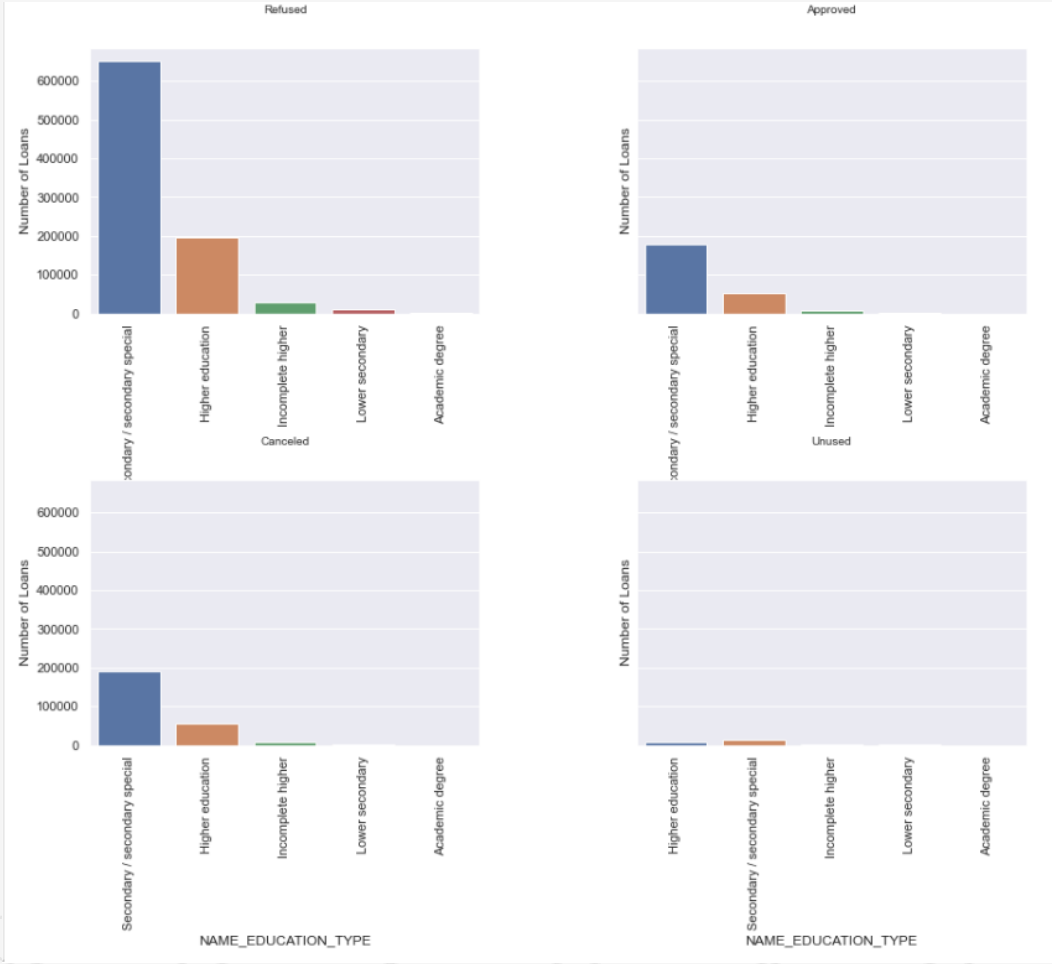
CODE_GENDER



Interfernce:

Female is getting more Refused more approved more canceled more unused but in case of male it is having average in every category.

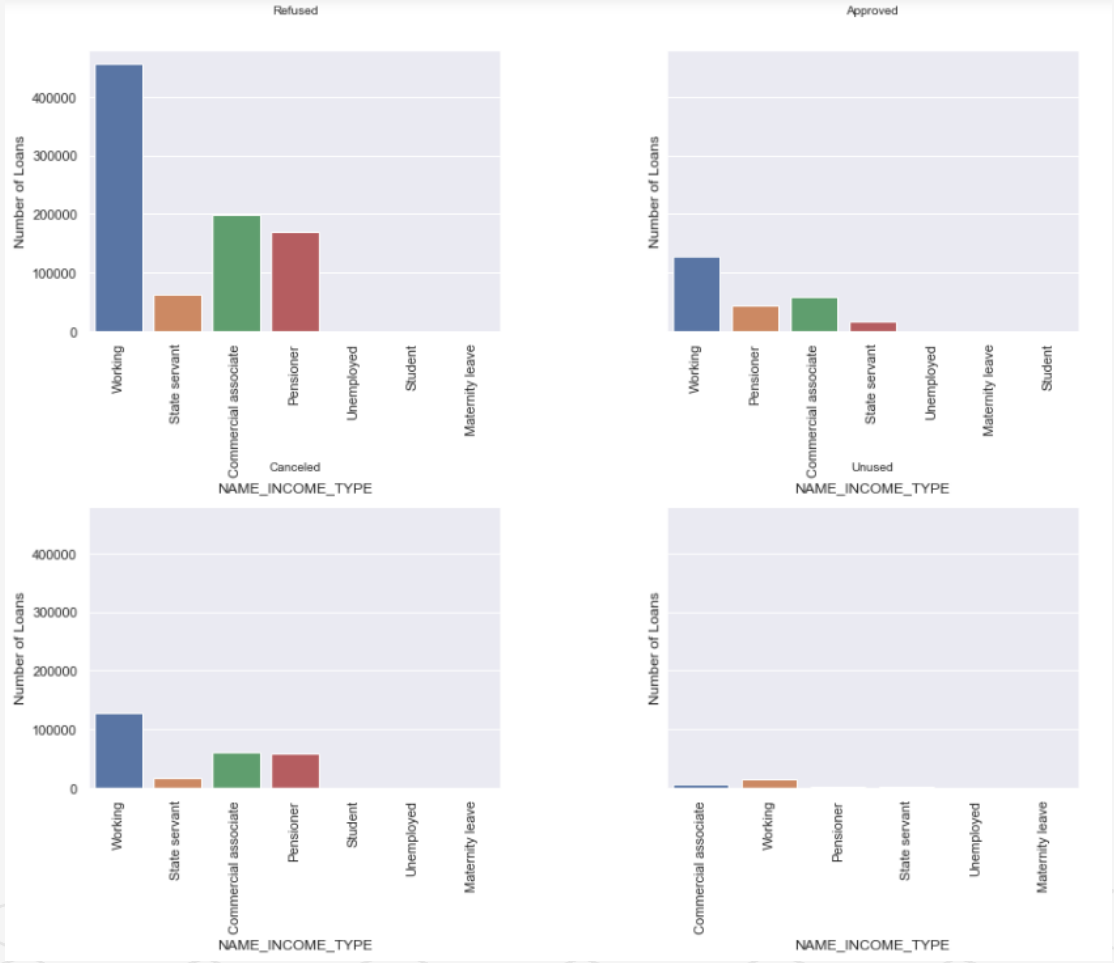
NAME_EDUCATION_TYPE



Interfernce:

Secondary/ Secondary special is more effective in every case

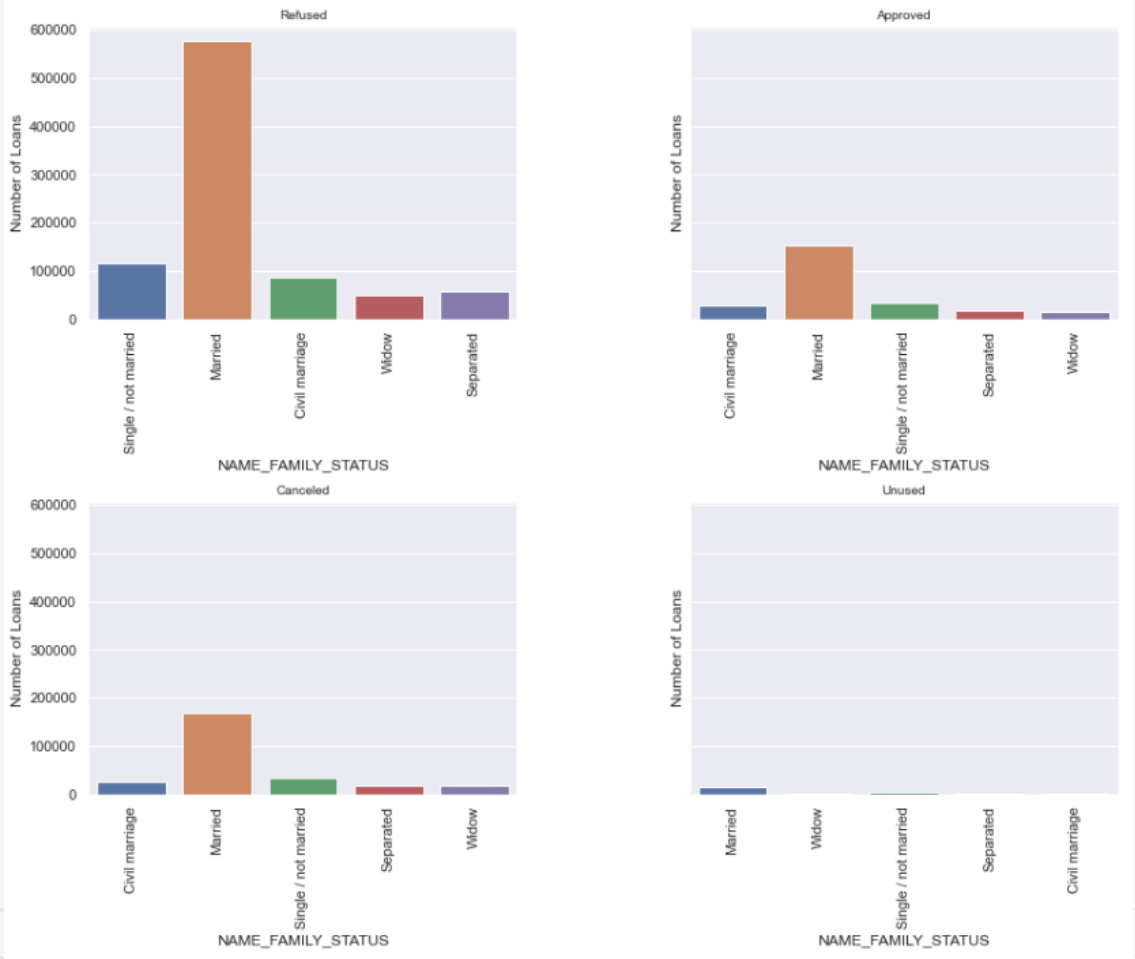
NAME_INCOME_TYPE



Interfernce:

The working type people are applying more loans as compare to others and also Commercial associates people are taking more loans.

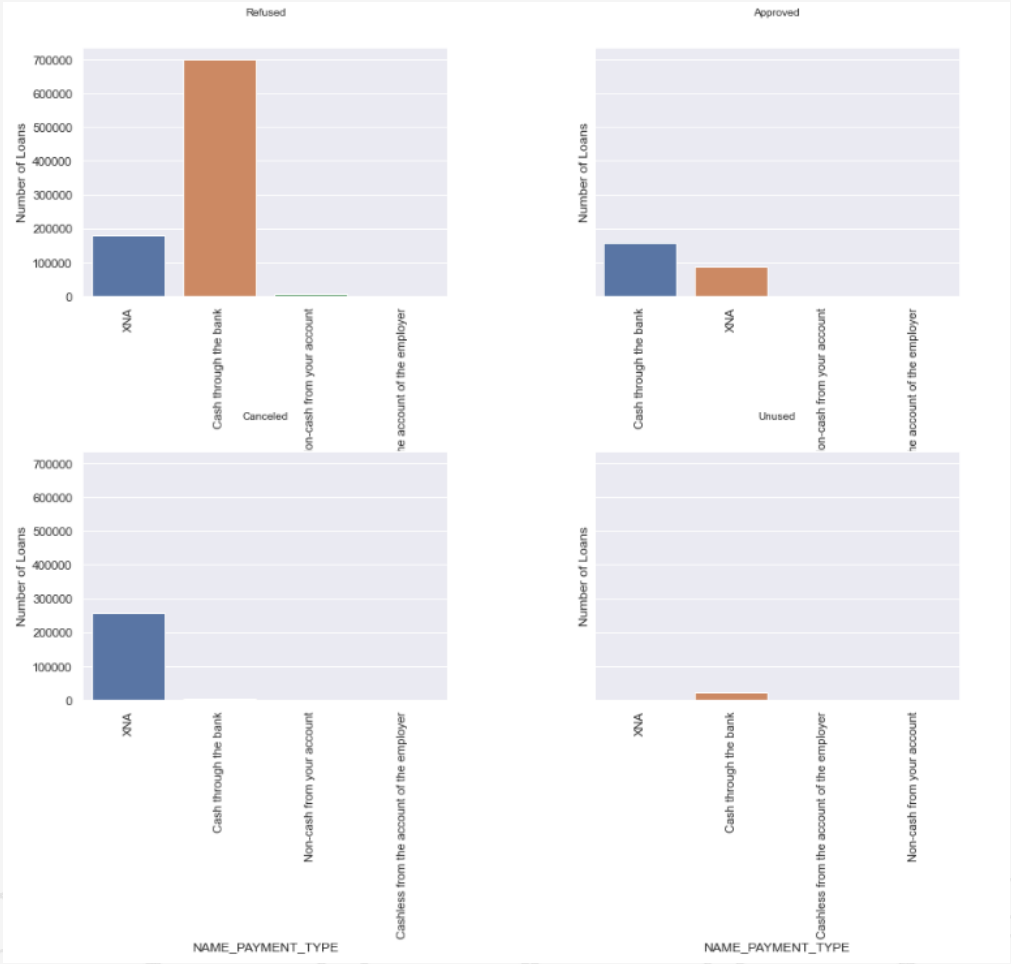
NAME_FAMILY_STATUS



Interfernce:

Married people are applying and taking loans more than the others

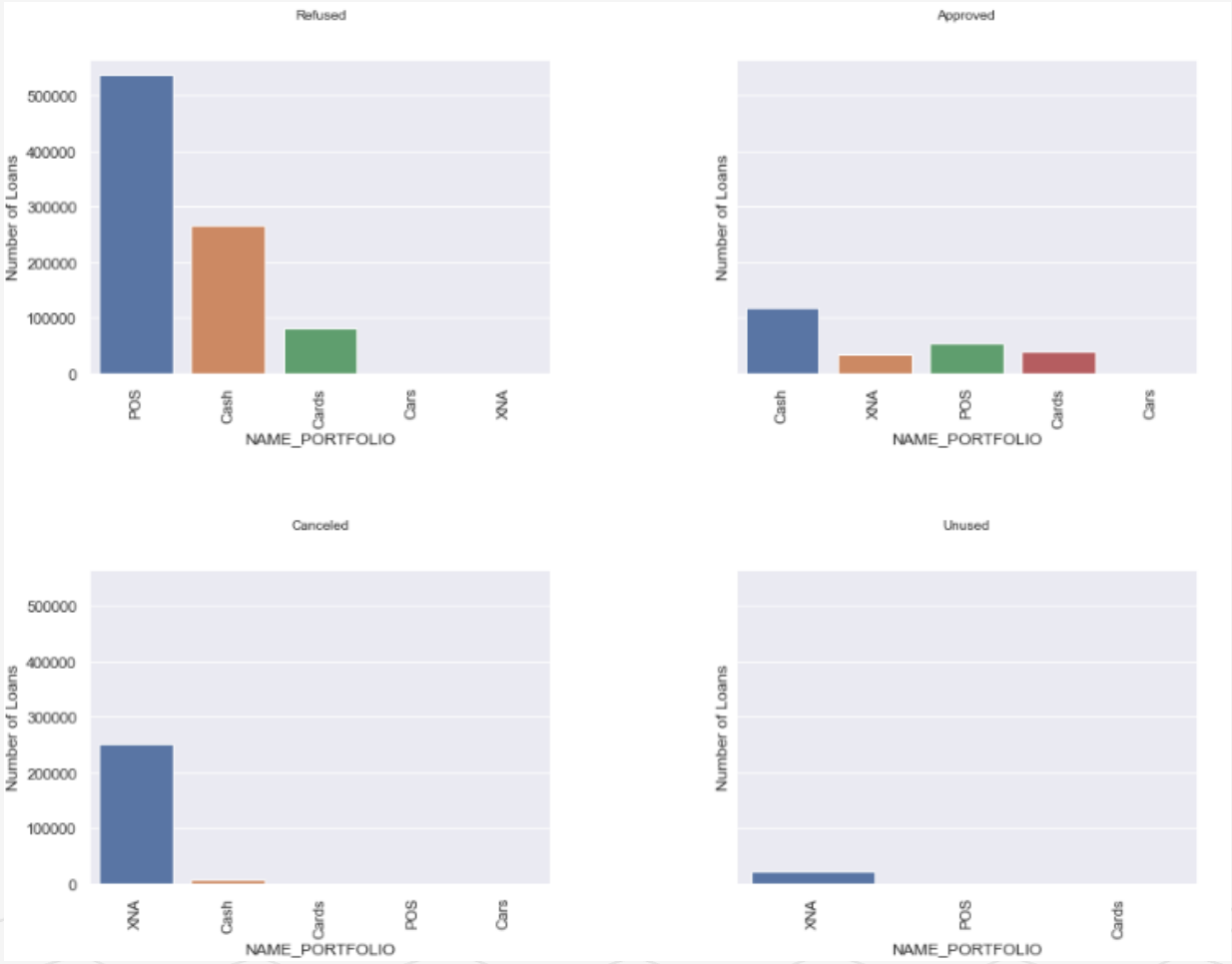
NAME_PAYMENT_TYPE



Interfernce:

The people are taking more loan in format of cash through the bank.

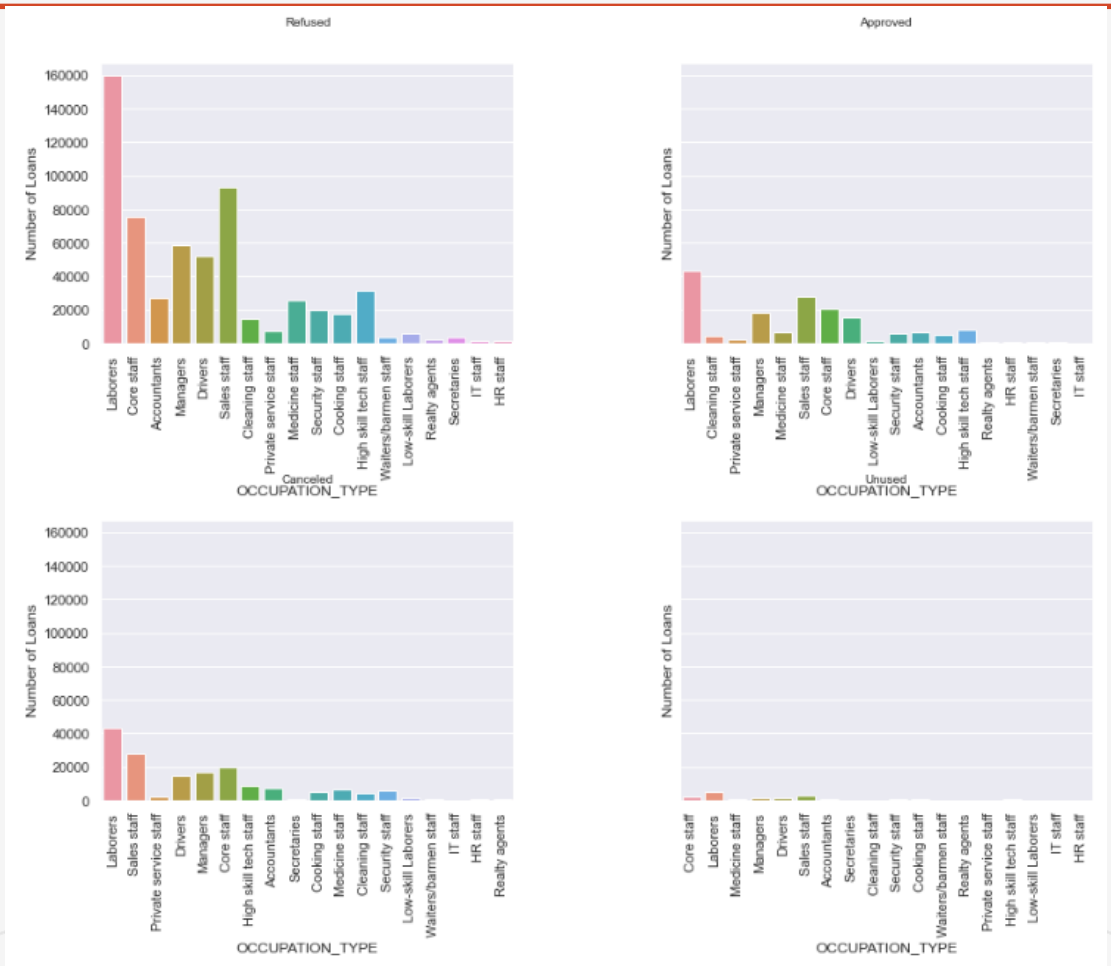
NAME_PORTFOLIO



Interfernce:

Here most approved loan were through POS and Most refused loans were in cash.

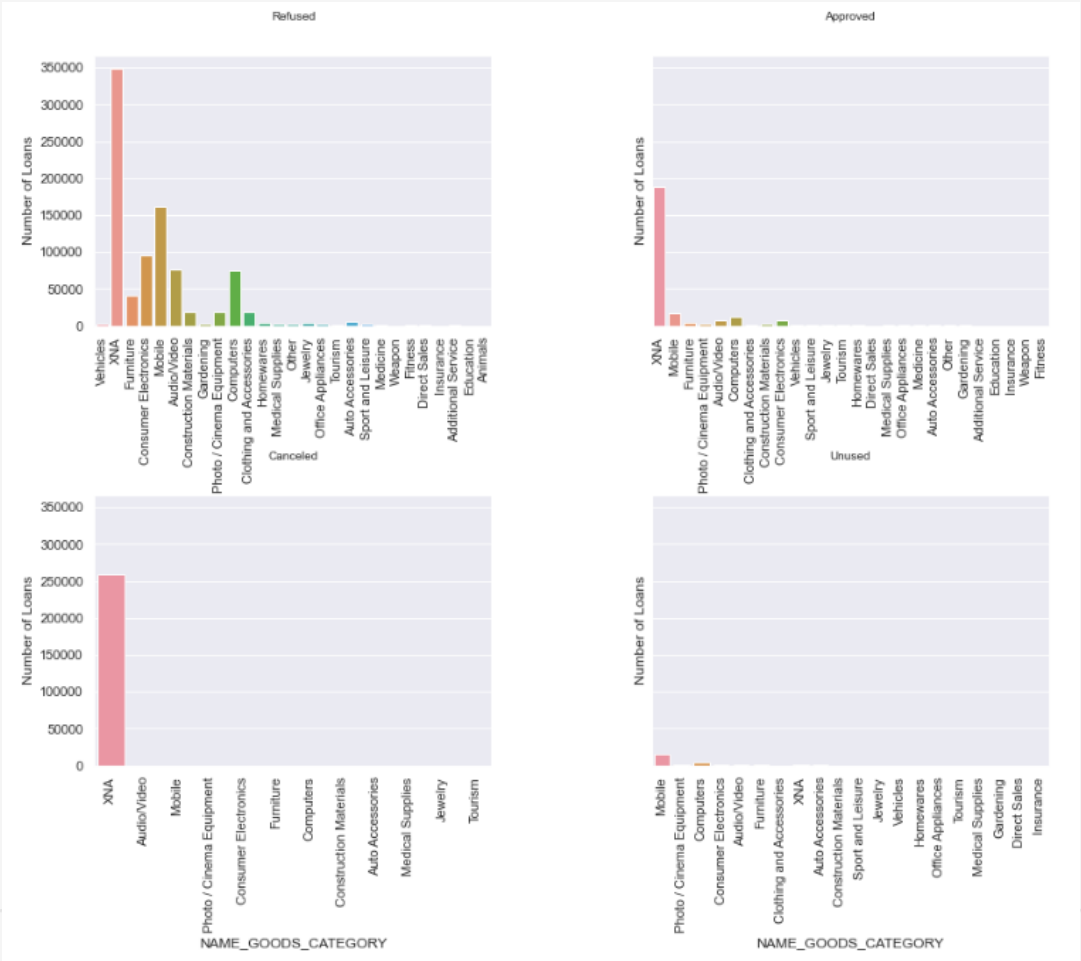
OCCUPATION_TYPE



Interfernce:

1. laborers are getting most refused and most approved loans.
2. Sales staff is also getting the second most refused and approved loans.

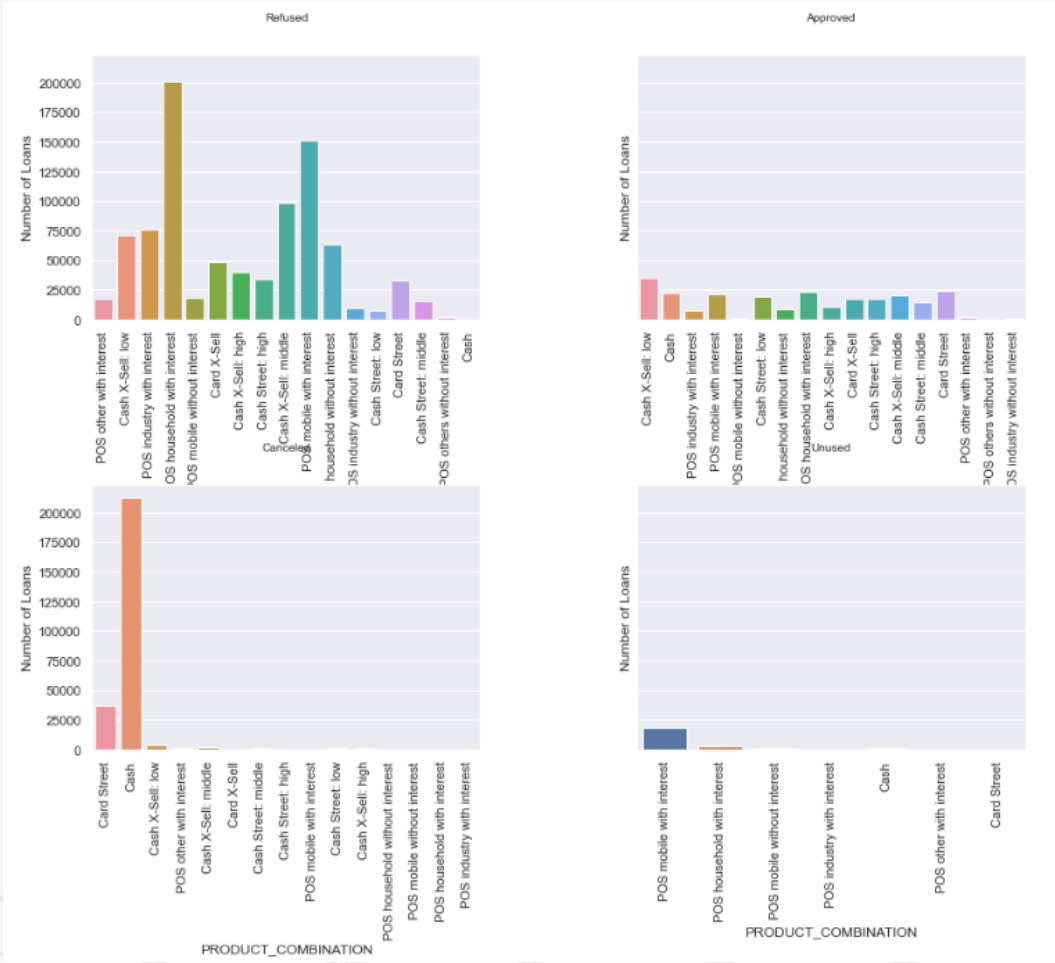
NAME_GOODS_CATEGORY



Interfernce:

Most Refused loan is of Mobile and most approved loan is Mobile

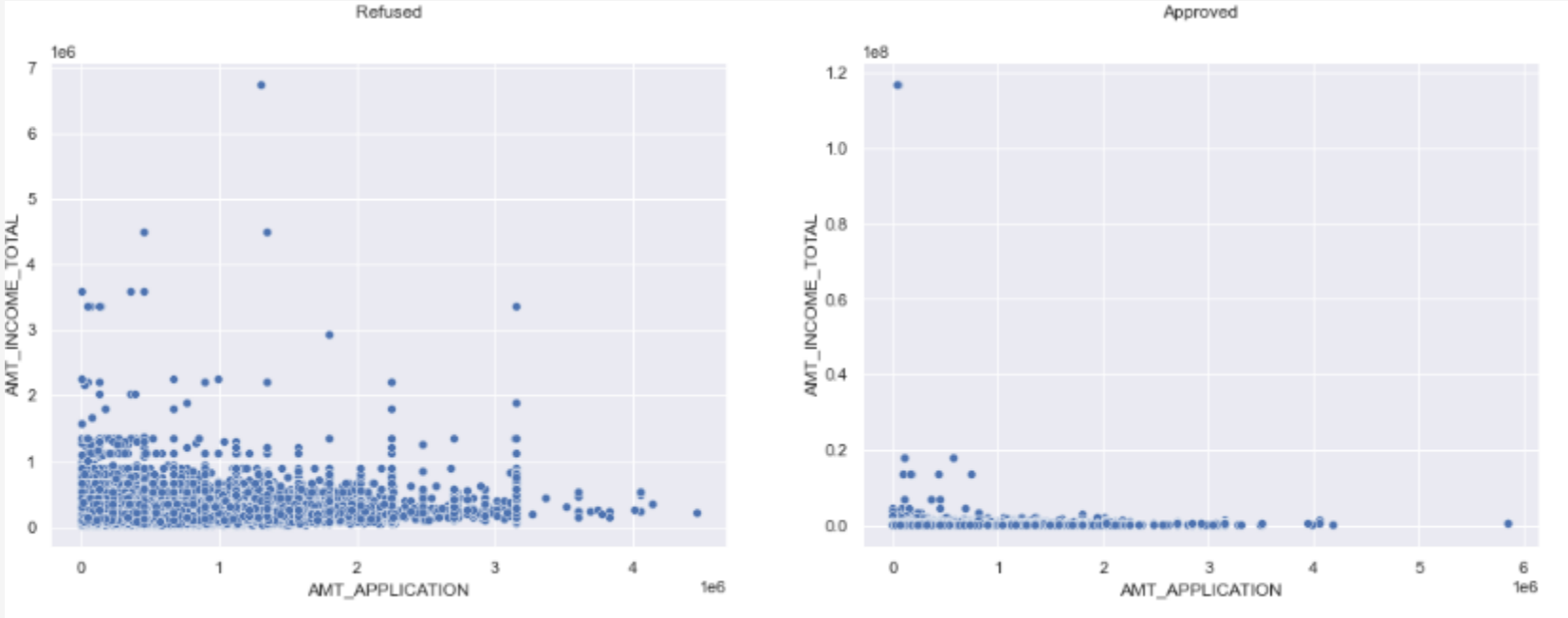
PRODUCT_COMBINATION



Interfernce:

The most accepting loan is Cash X-sell: low And most canceled loan is Cash and Most Unused loan is POS mobile with interest

AMT_APPLICATION vs AMT_INCOME_TOTAL



Interfernce:

Loan request higher than 200k had a higher rejection rate. Also loan rejection rate was much lower if the income was higher than 500k

Thank You