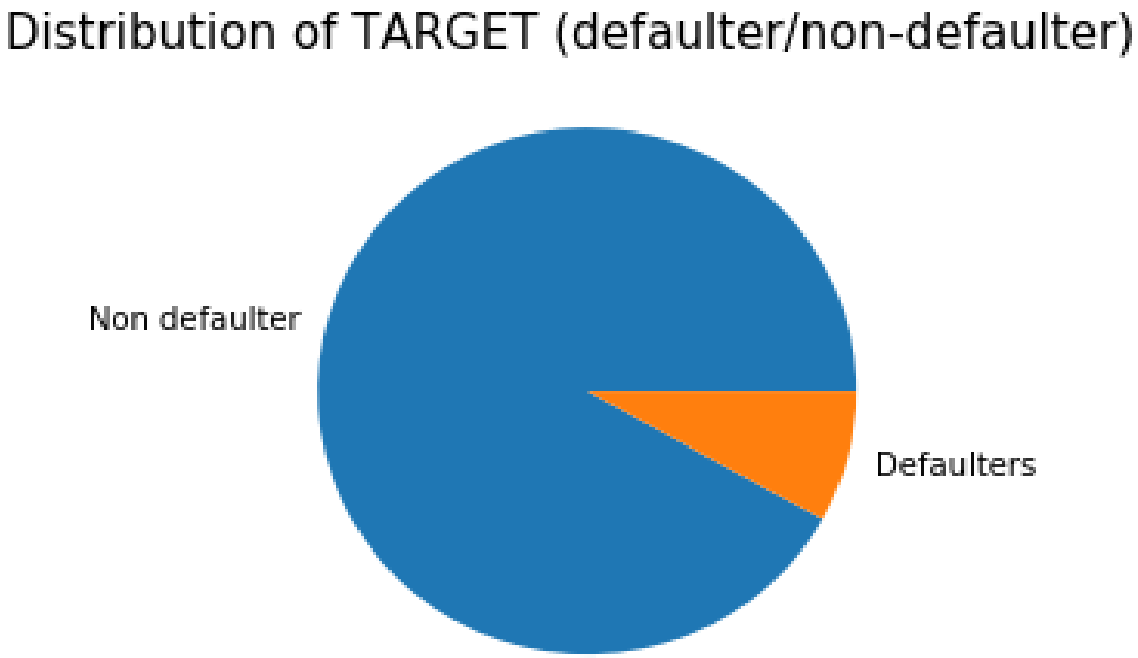
CREDIT EDA CASE STUDY

BY:: Rohit Lal and Monica Fatwani

- Application Data Frame

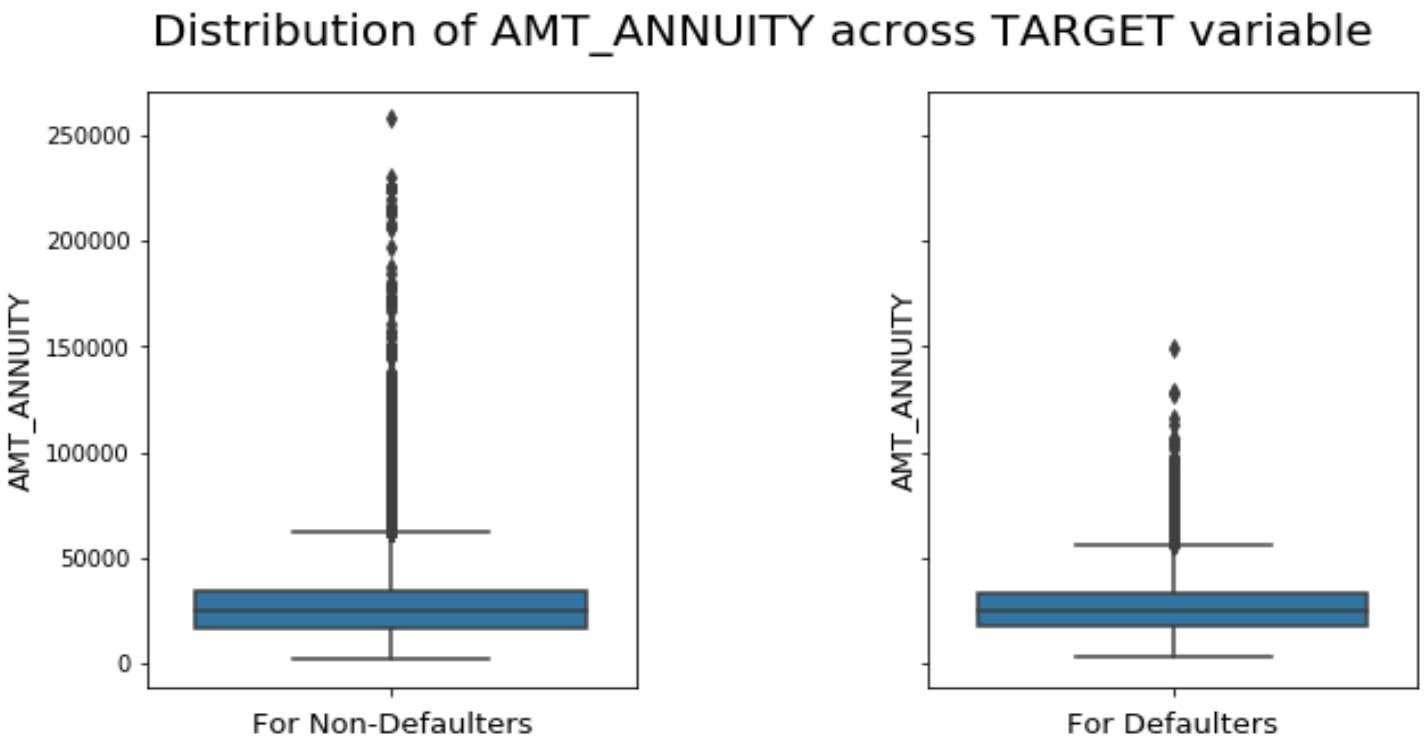
Target


Distribution of TARGET (defaulter/non-defaulter)
Non defaulter
Defaulters

Univariate Analysis –
For Continuous Data Types:

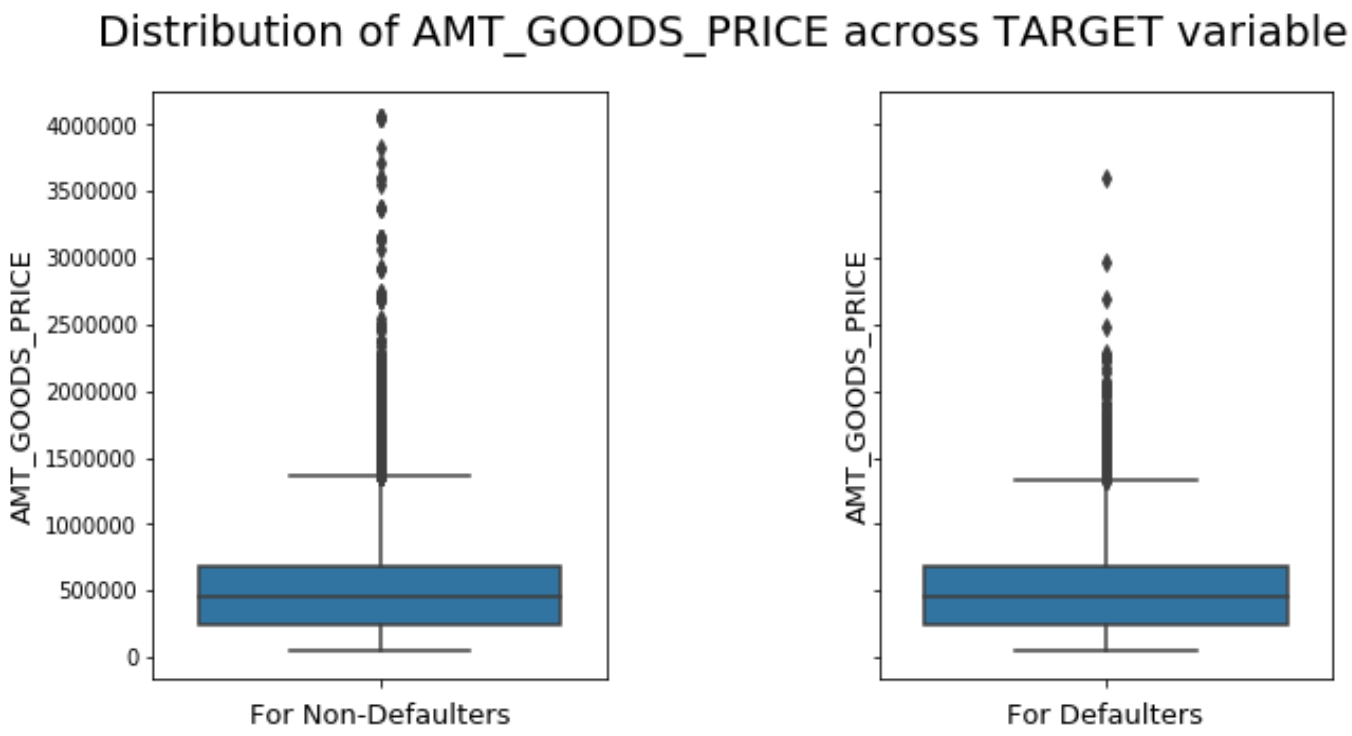
AMT_ANNUITY

1) No strong comparison found in AMT_ANNUITY with the TARGET variable
2) Non-defaulters have in generally loans with very high Annuity amounts (see, outliers in the above plot. Non-defaulters have very high magnitude of Outliers.)
3) This can be due to the High Annuity loans are provided only to trusted people which are Non-defaulters



AMT_GOODS_PRICE

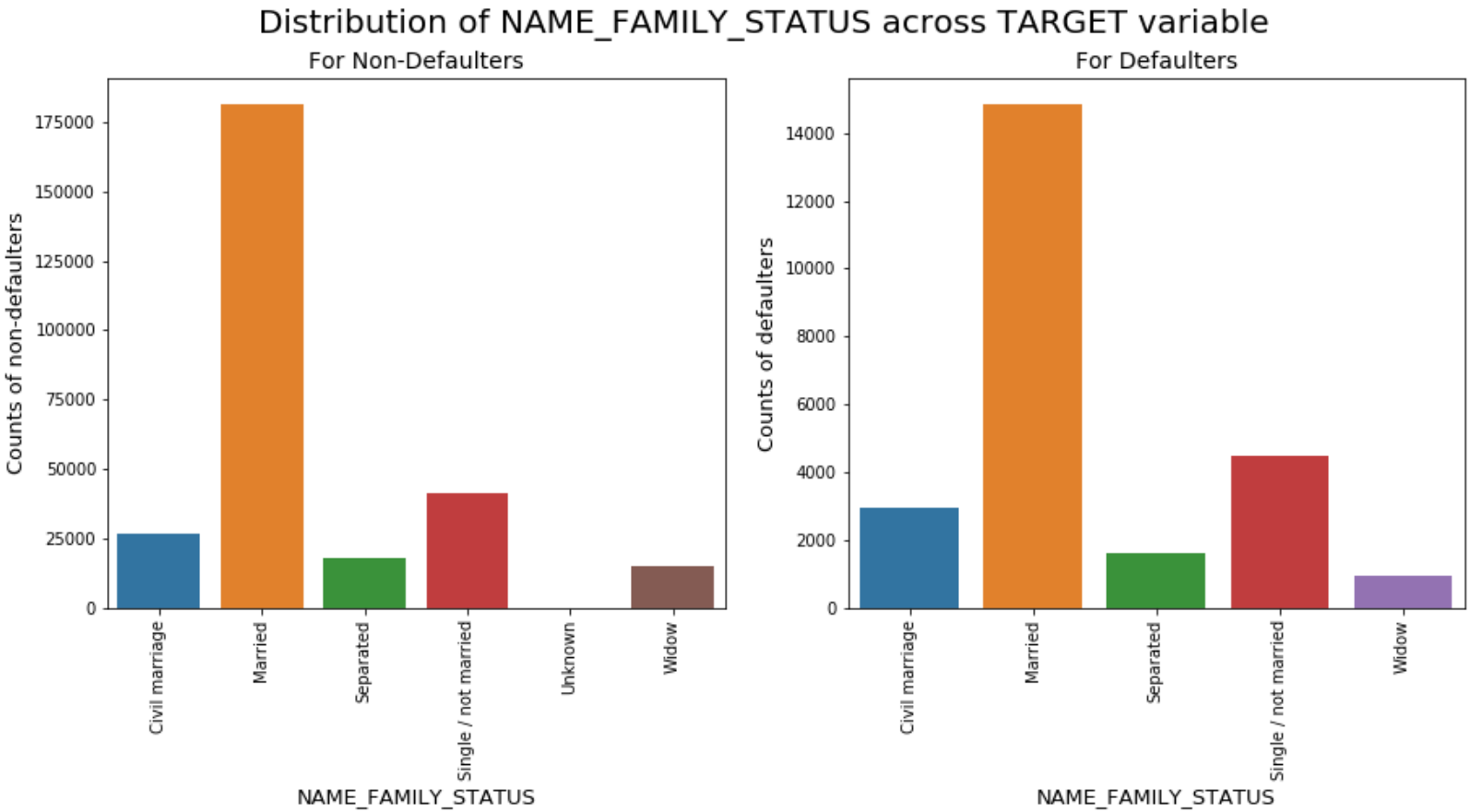
1) AMT_GOODS_PRICE column has quite high numbers of outliers.
2) Again, mean would not be an appropriate measure.
We can use Median to estimate the missing values



Univariate Analysis –
For Categorical Data types(Unordered):

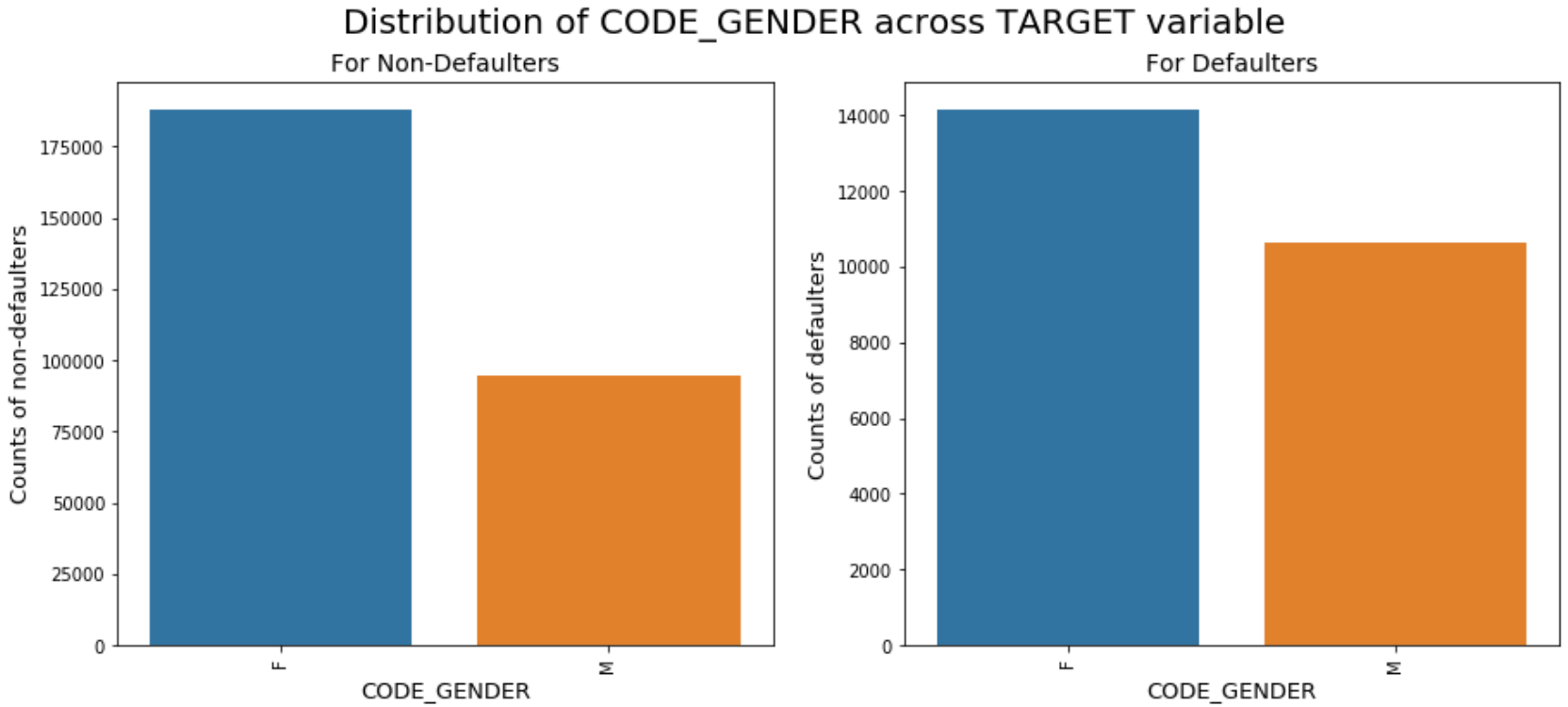
NAME_FAMILY_STATUS

1) Married people are at top in both defaulters and non-defaulters
2) Not-married ratio in defaulters is comparatively higher that that of non defaulted population.



CODE_GENDER

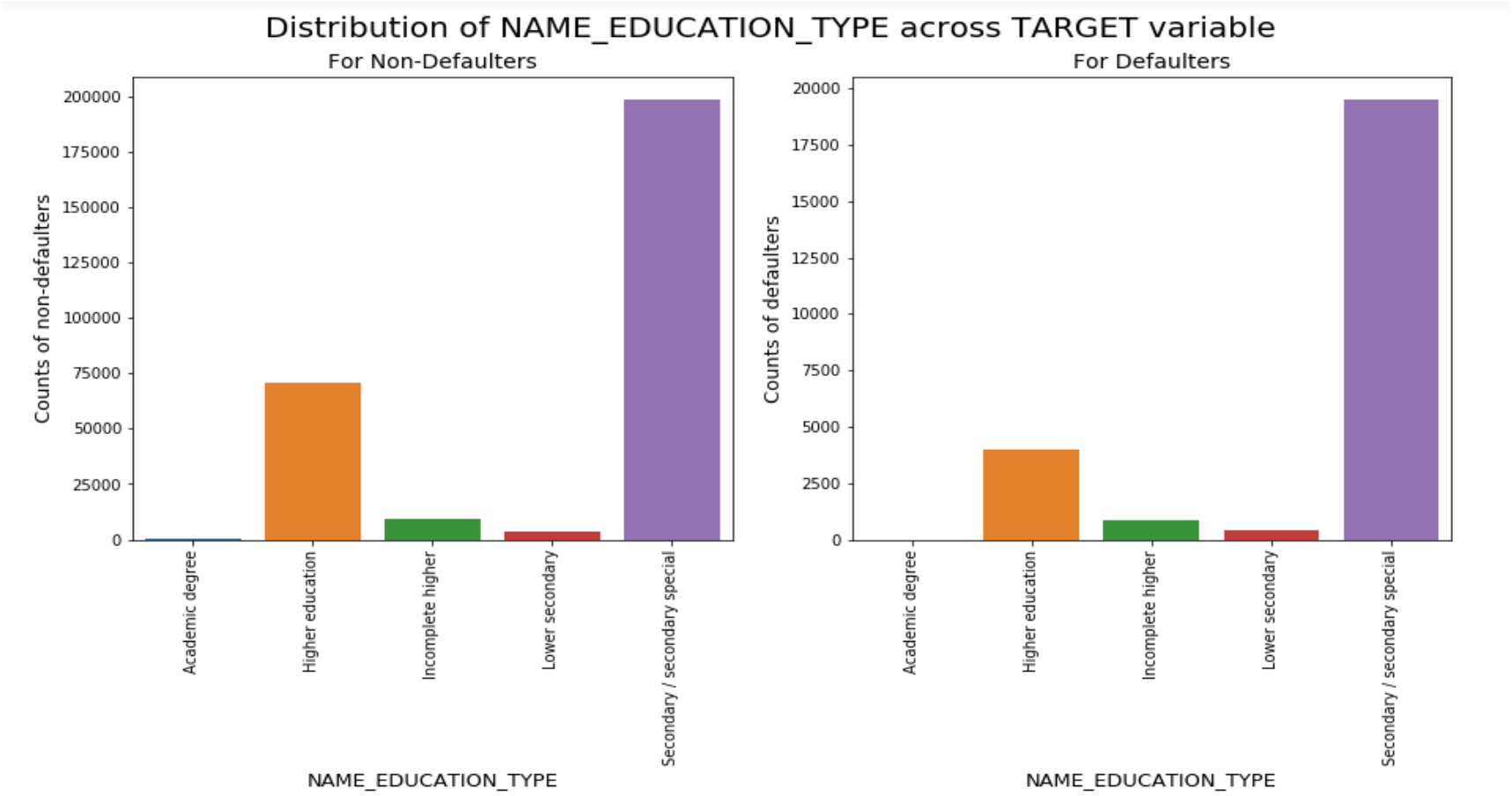
1) Ratio of Female to Male in case of defaulters is comparatively less than that in case of Non-defaulters.



- Univariate Analysis –
  For Categorical Data types(ordered):

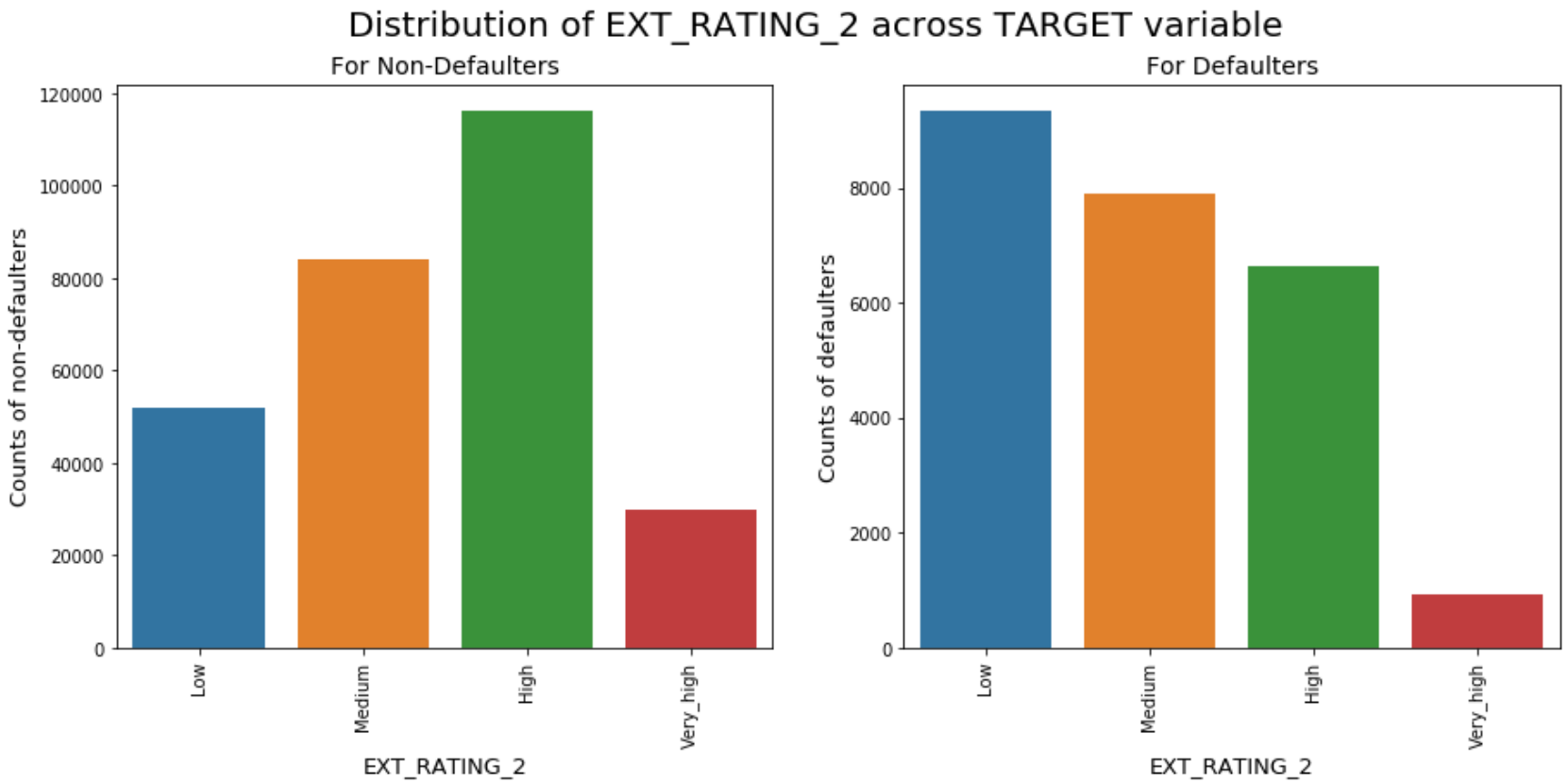
NAME_EDUCATION_TYPE

1) There is no strong relation between NAME_EDUCATION_TYPE & TARGET variable.
2) In general, Education Type 'Secondary/Secondary Special' is highest and Education Type 'Academic degree' is least for both TARGET categories.



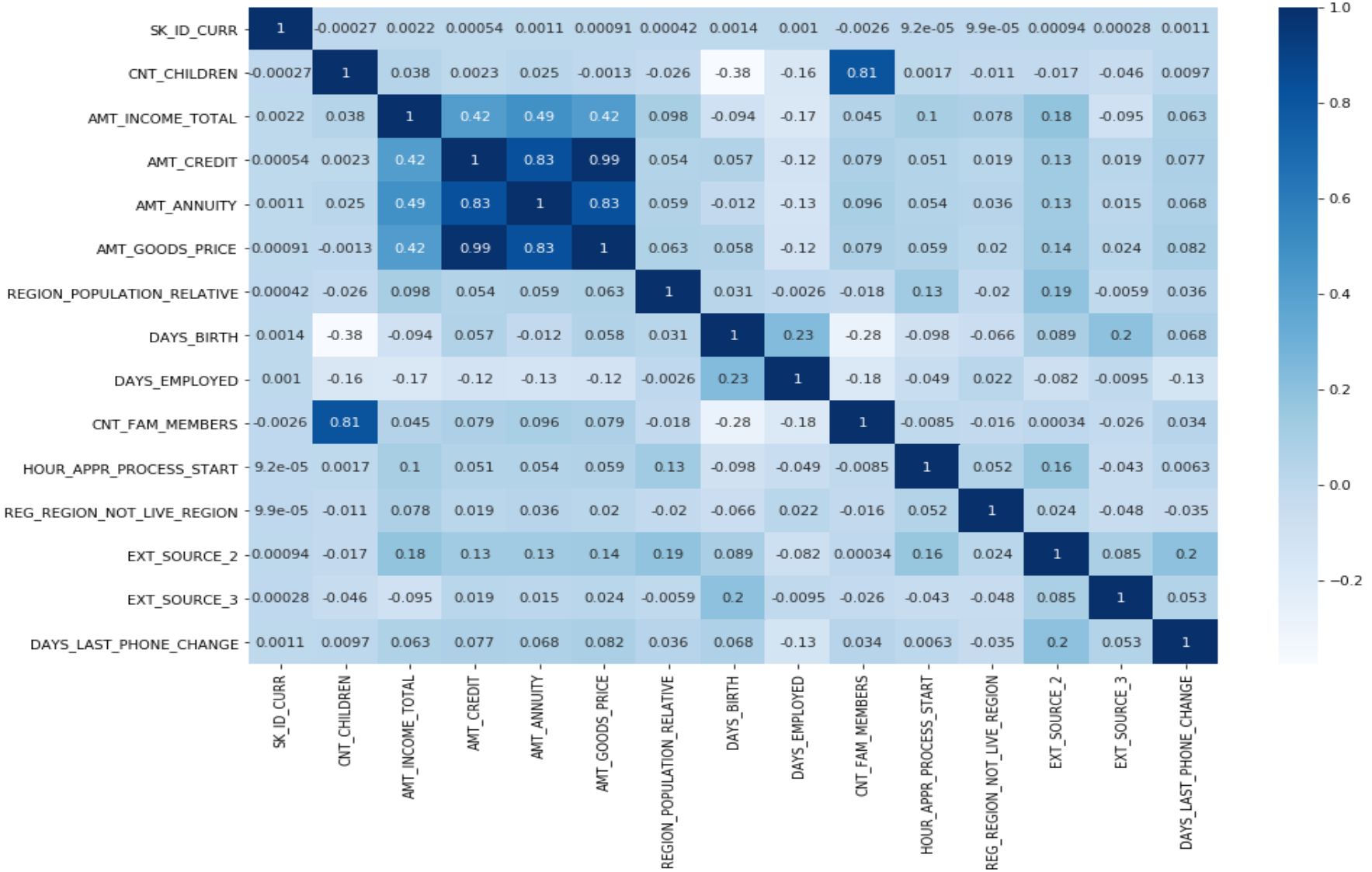
EXT_RATING_2

1) Low Ratings ratio in the Defaulters population is clearly higher as compared to the Non-defaulter population.
2) EXT_RATING_2 shows strong relationship with our target variable (i.e., defaulters/non-defaulters)



- Correlation for Non-Defaulters

Correlation for Non-Defaultors

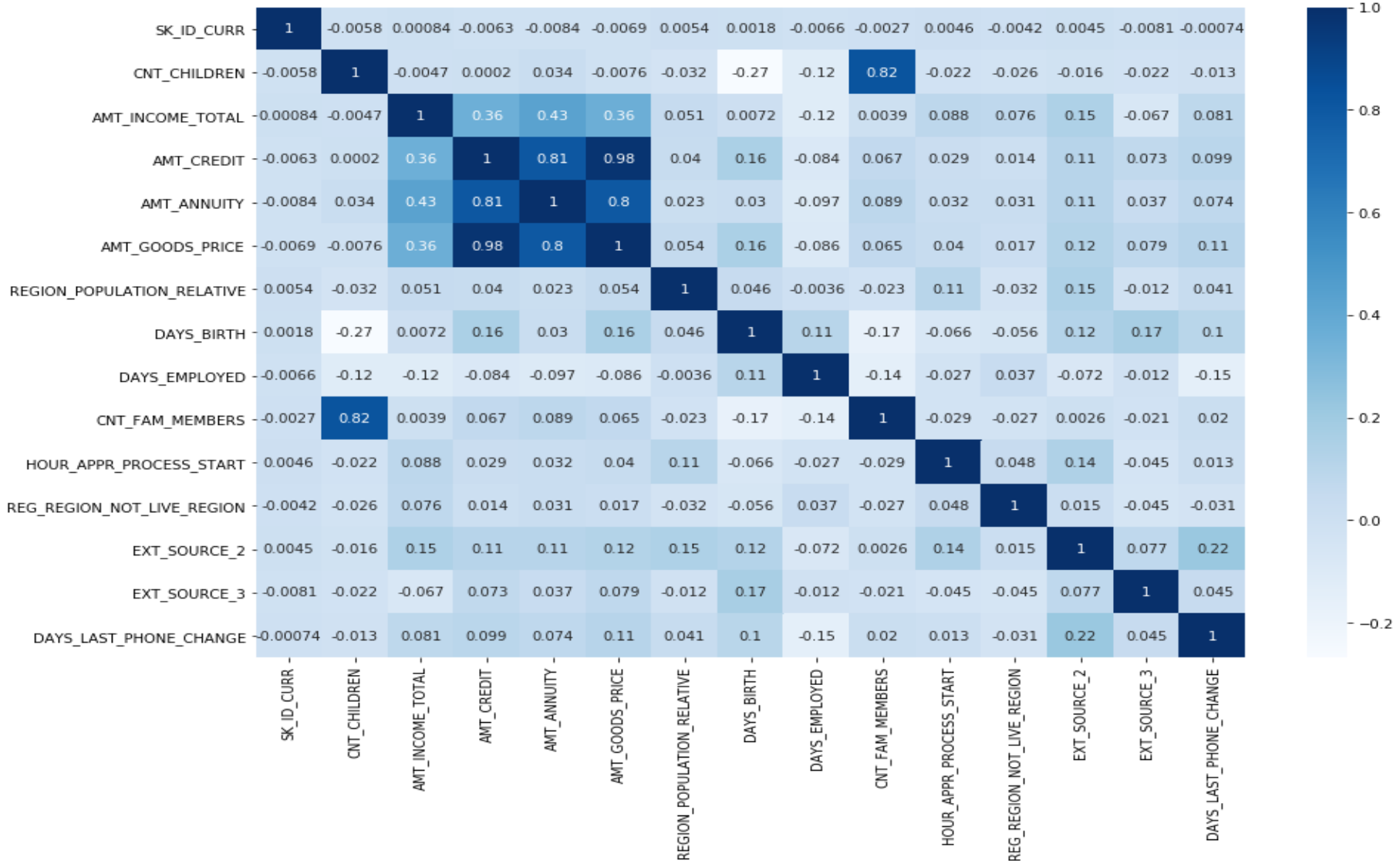


Correlation for Non-Defaulters(Type 0):

- Total income and Credit amount are strongly related. It means wealthy people take generally higher credit amounts.
- Credit Amount and Annuity are strong related. It is due the behaviour of business - For high credit amount loans, annuity will always be higher.
- Children count and Family member count shows strong relation. Again, it is due the behaviour of the variable. More children constitutes to more family members.
- Children and population density are also inversely related, means in densely populated region, clients have less children. This may due to space related constraints or some government body guidelines in dense regions.
- Credit amount and population density are positively related. It means denser areas constitute good credit amount.

- Correlation for Defaulters

Correlation for Defaultors



Correlation for Non-Defaulters(Type 1):

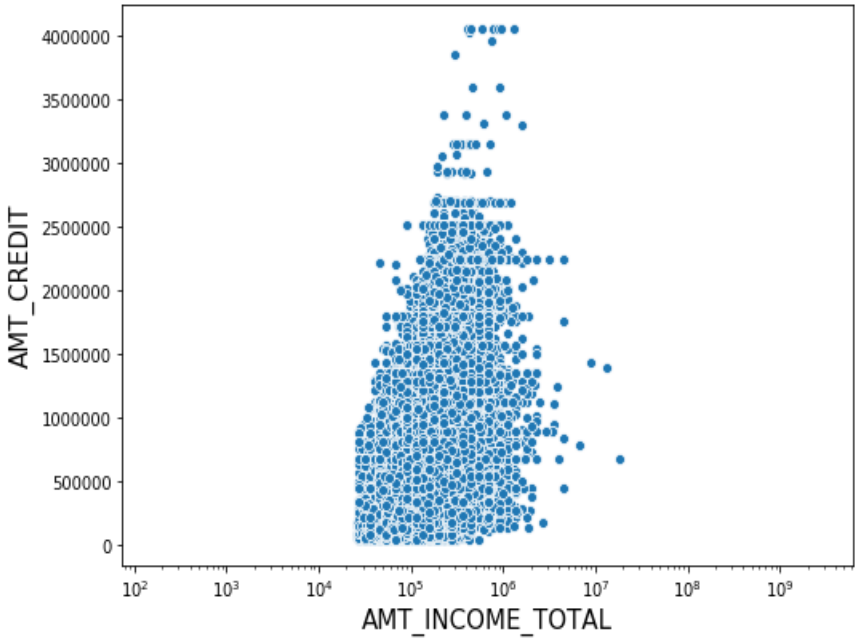
- **For Defaulters also, the prominent correlated variables are same as non-defaulters.**
- Total income and Credit amount are strongly related. It means wealthy people take generally higher credit amounts.
- Credit Amount and Annuity are strong related. It is due the behaviour of business - For high credit amount loans, annuity will always be higher.
- Children count and Family member count shows strong relation. Again, it is due the behaviour of the variable. More children constitutes to more family members.
- Children and population density are also inversely related, means in densely populated region, clients have less children. This may due to space related constraints or some government body guidelines in dense regions.
- Credit amount and population density are positively related. It means denser areas constitute good credit amount.

Bivariate (Continuous-Continuous)

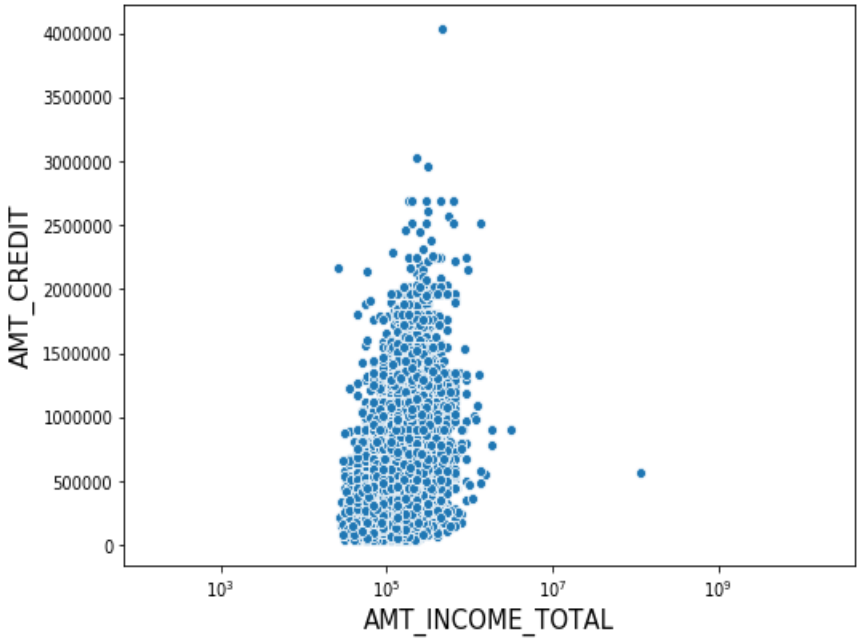
CREDIT v/s INCOME

- Higher income candidates in general present in Non-defaulter population.
- Log Scale for INCOME_TOTAL in above plots is chosen to take care of outlier points.

CREDIT vs INCOME for Non-Defaulters



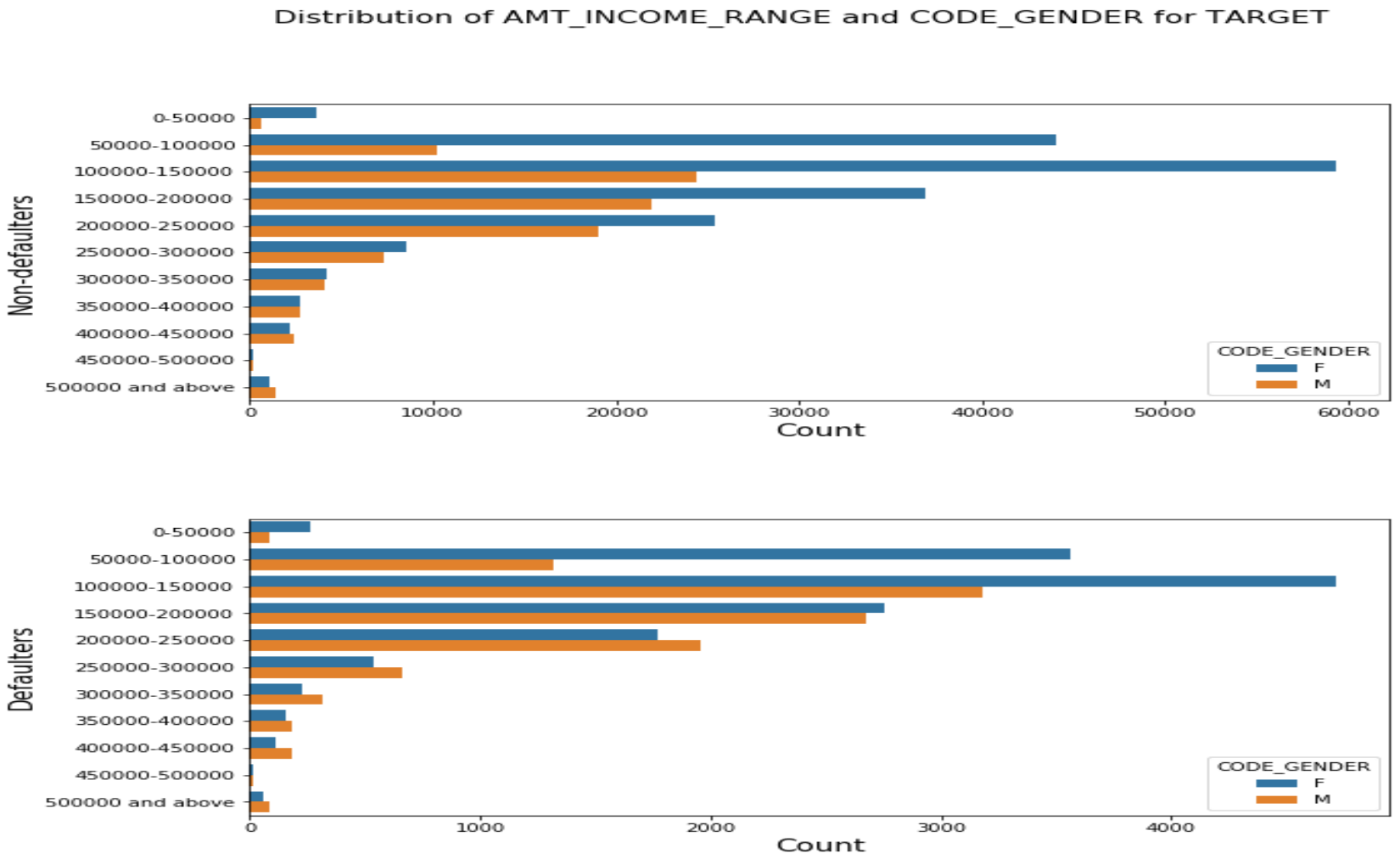
CREDIT vs INCOME for Defaulters



- Bivariate(Continuous-Categorical)

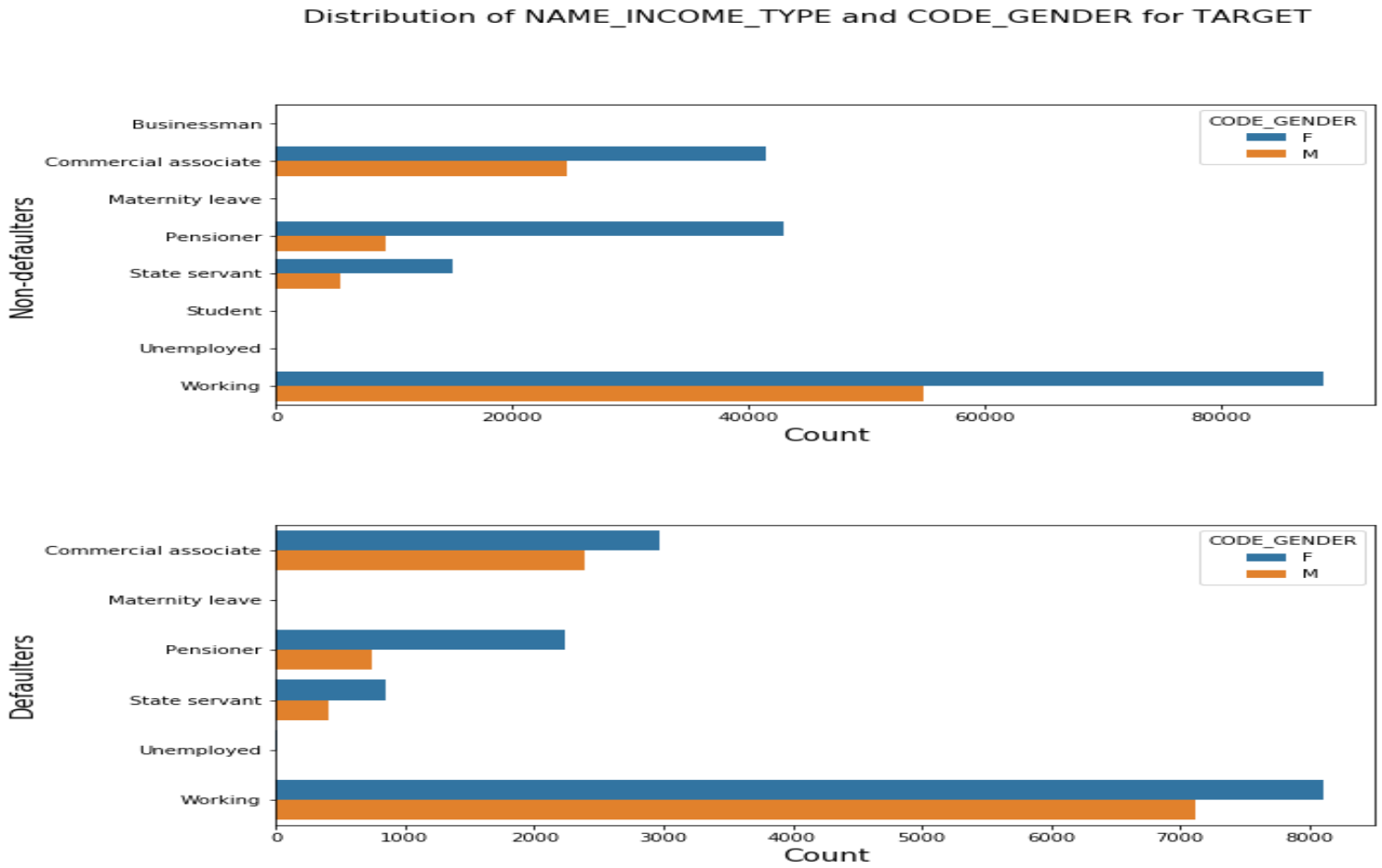
AMT_INCOME_RANGE v/s CODE_GENDER

1) Female counts are higher than male.
2) Income range from 50000 to 250000 is having more number of credits.



NAME_INCOME_TYPE v/s CODE_GENDER

1) For income type working, commercial associate, and State Servant, pensioner the number of credits are higher than others.
2) For this Females are having more number of credits than male.
Less number of credits for income type student, Businessman and Unemployed.

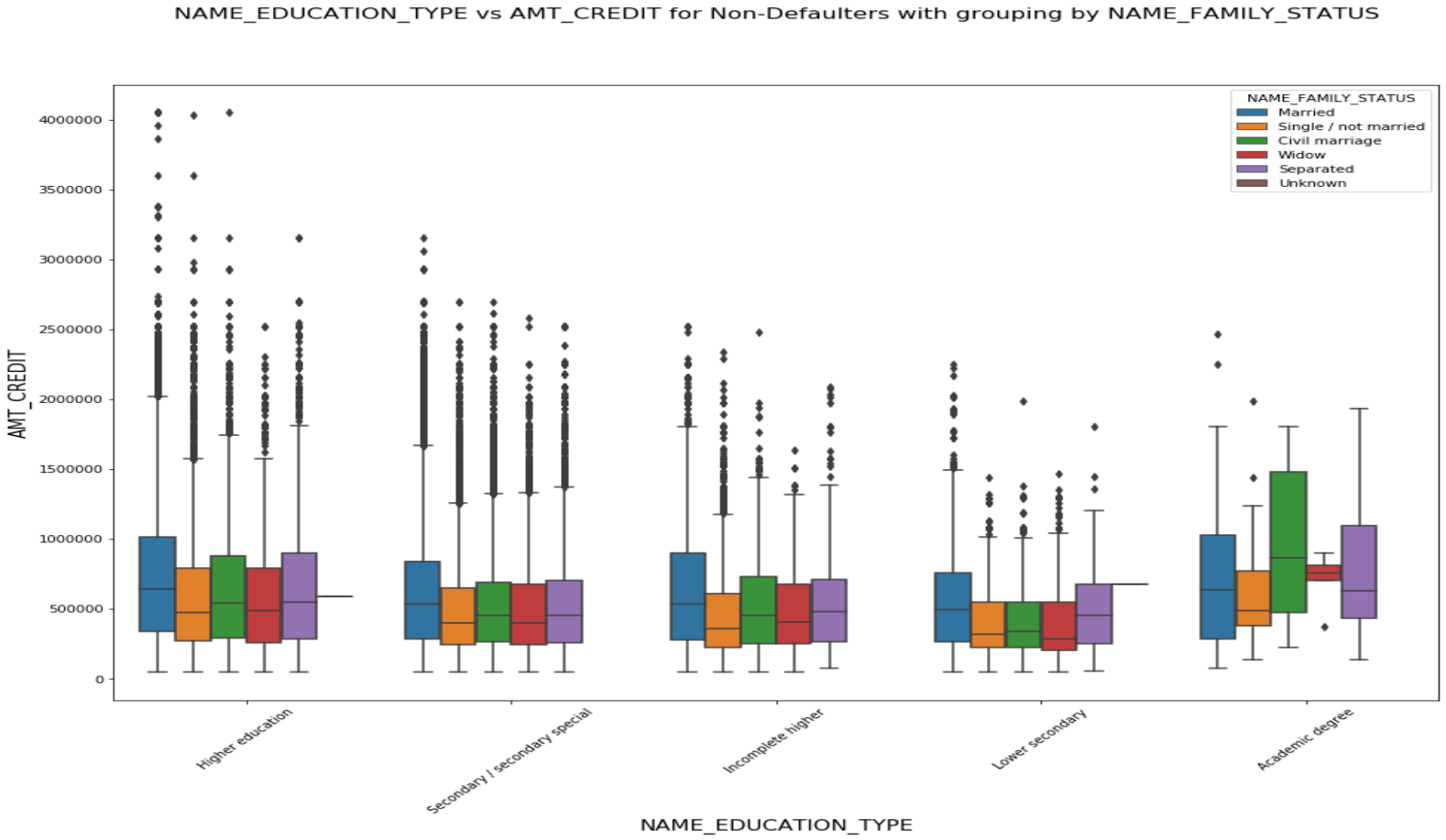


- Multivariate Analysis:

NAME_EDUCATION_TYPE v/s AMT_CREDIT v/s NAME_FAMILY_STATUS - Non Defaulters

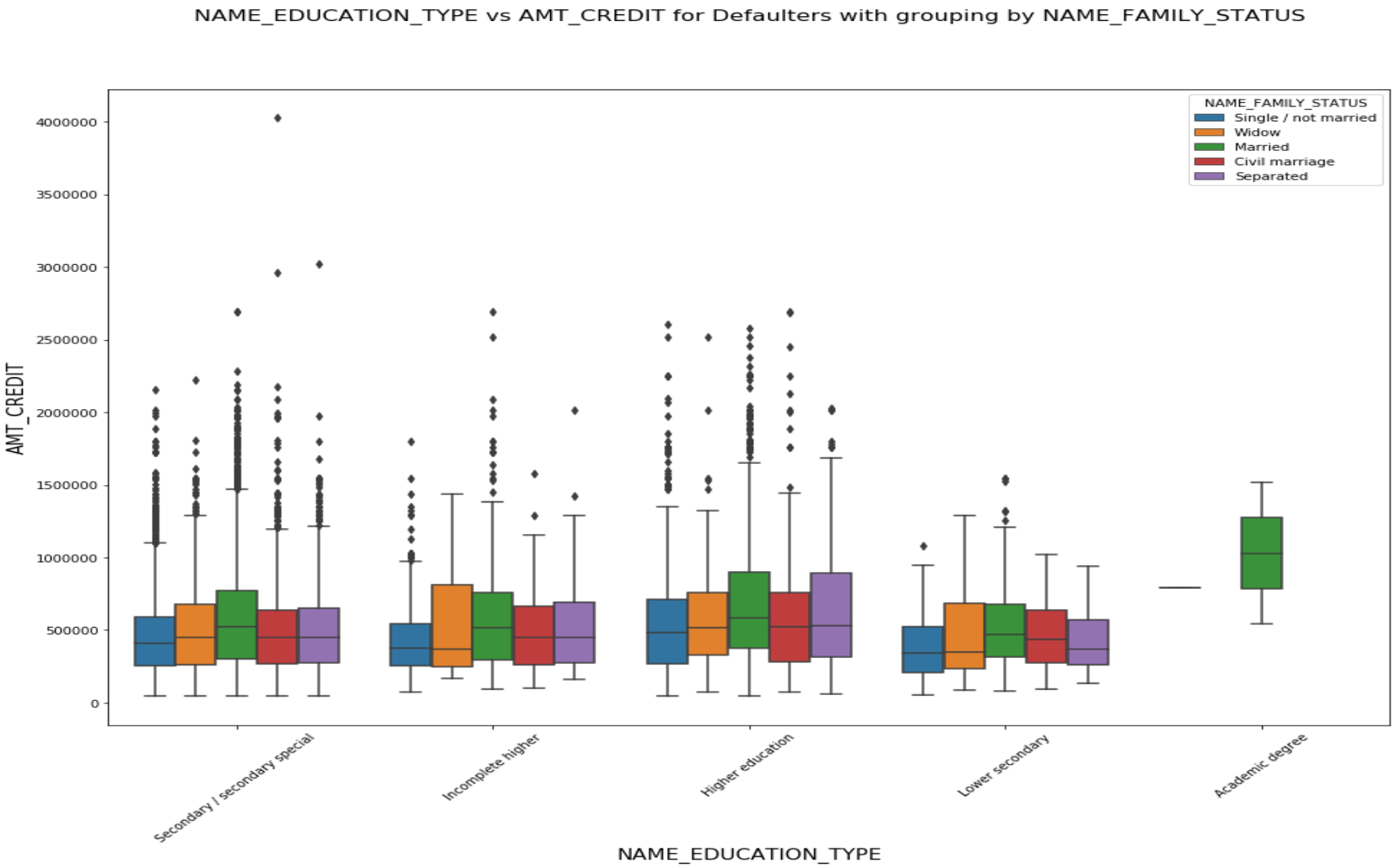
1)In Academic degree category, AMT_CREDIT counts are much higher for Civil Marriage group.

2)In Academic degree category, AMT_CREDIT counts are minimum for Widow group.



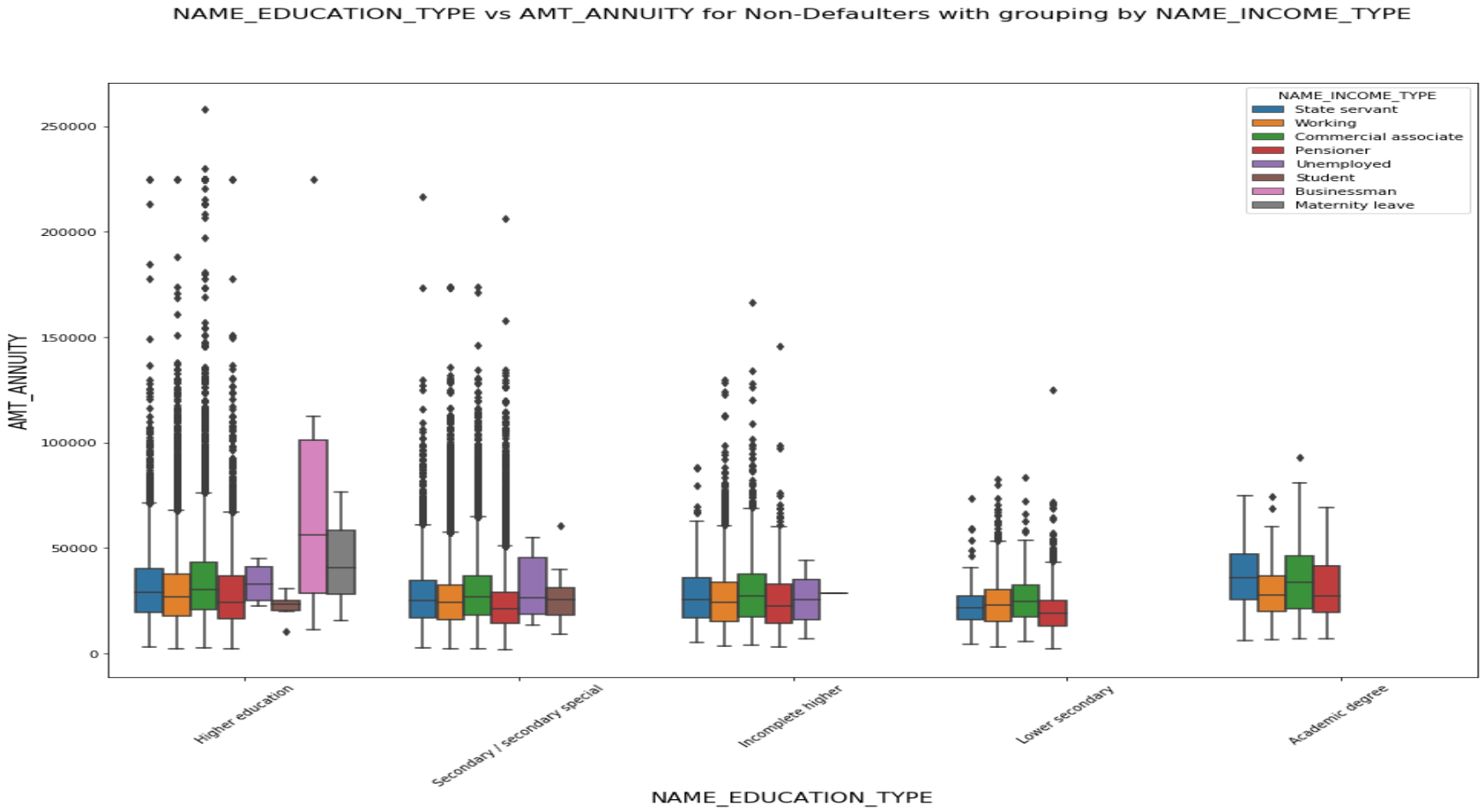
NAME_EDUCATION_TYPE v/s AMT_CREDIT v/s NAME_FAMILY_STATUS - Defaulters

1) In Academics category, only Married group has defaulters.
2) So, bank must prefer applicants with Academics degree and Non-married.



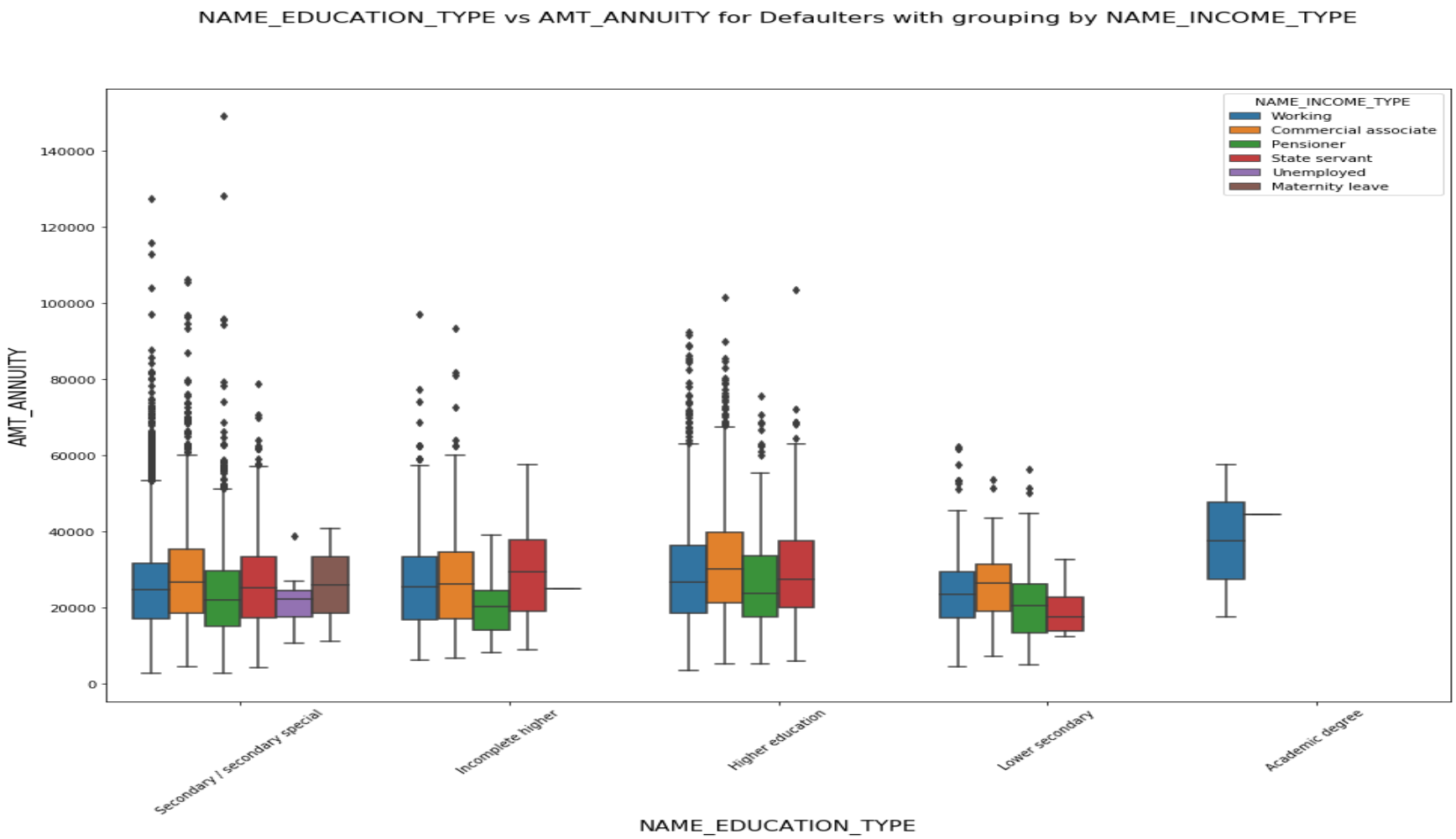
NAME_EDUCATION_TYPE v/s AMT_ANNUITY v/s NAME_INCOME_TYPE - Non-Defaulters

1) There are many applicants from Secondary/secondary special education type whose loan gets approved.
2) Higher Education:Businessman , Higher Education:Maternity leave and Higher Education:students should be targeted as per the data



NAME_EDUCATION_TYPE v/s AMT_ANNUITY v/s NAME_INCOME_TYPE- Defaulters

- Bank must focus on Academic degree Education Type non-working type as they are having higher number of successful payments on time.



- Previous Application dataframe

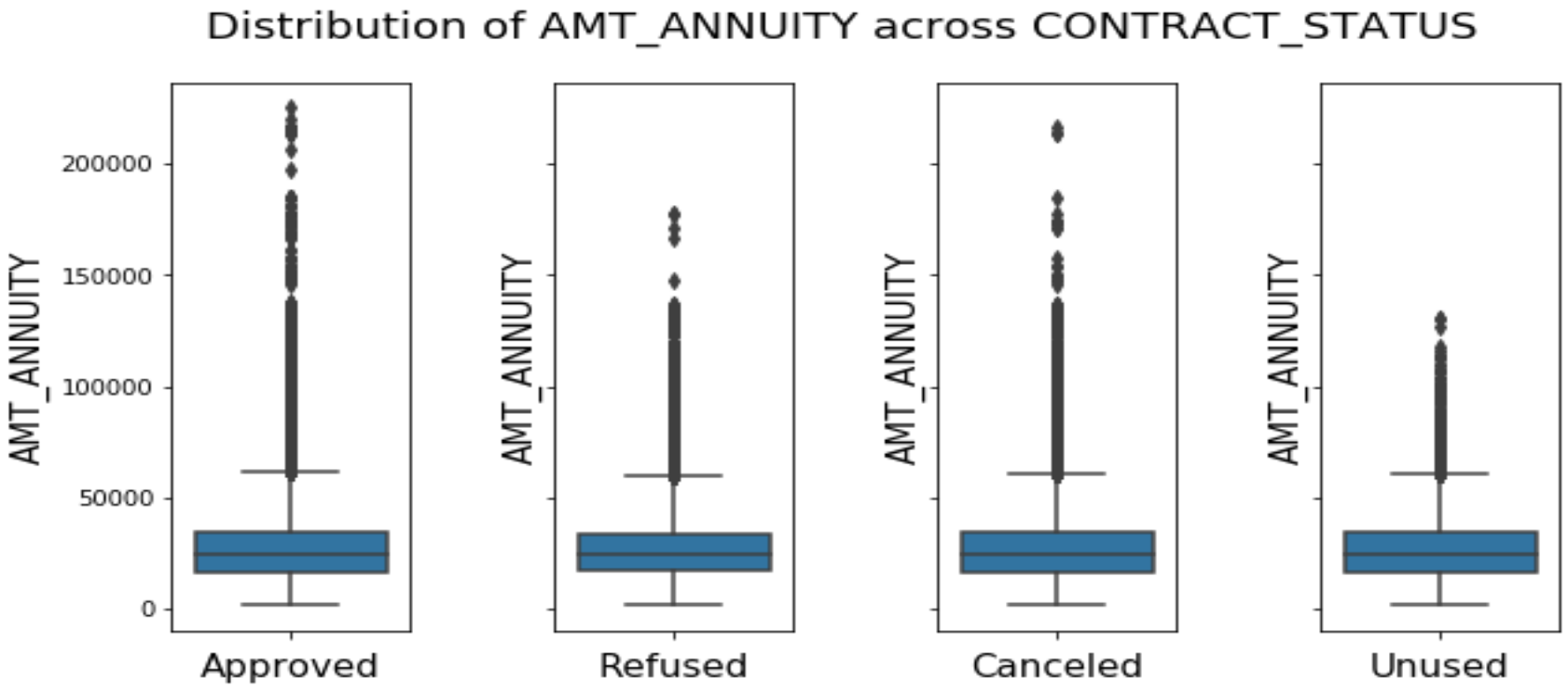
NAME_CONTRACT_STATUS

Univariate Analysis-

(Continuous Variables across Contract Status)

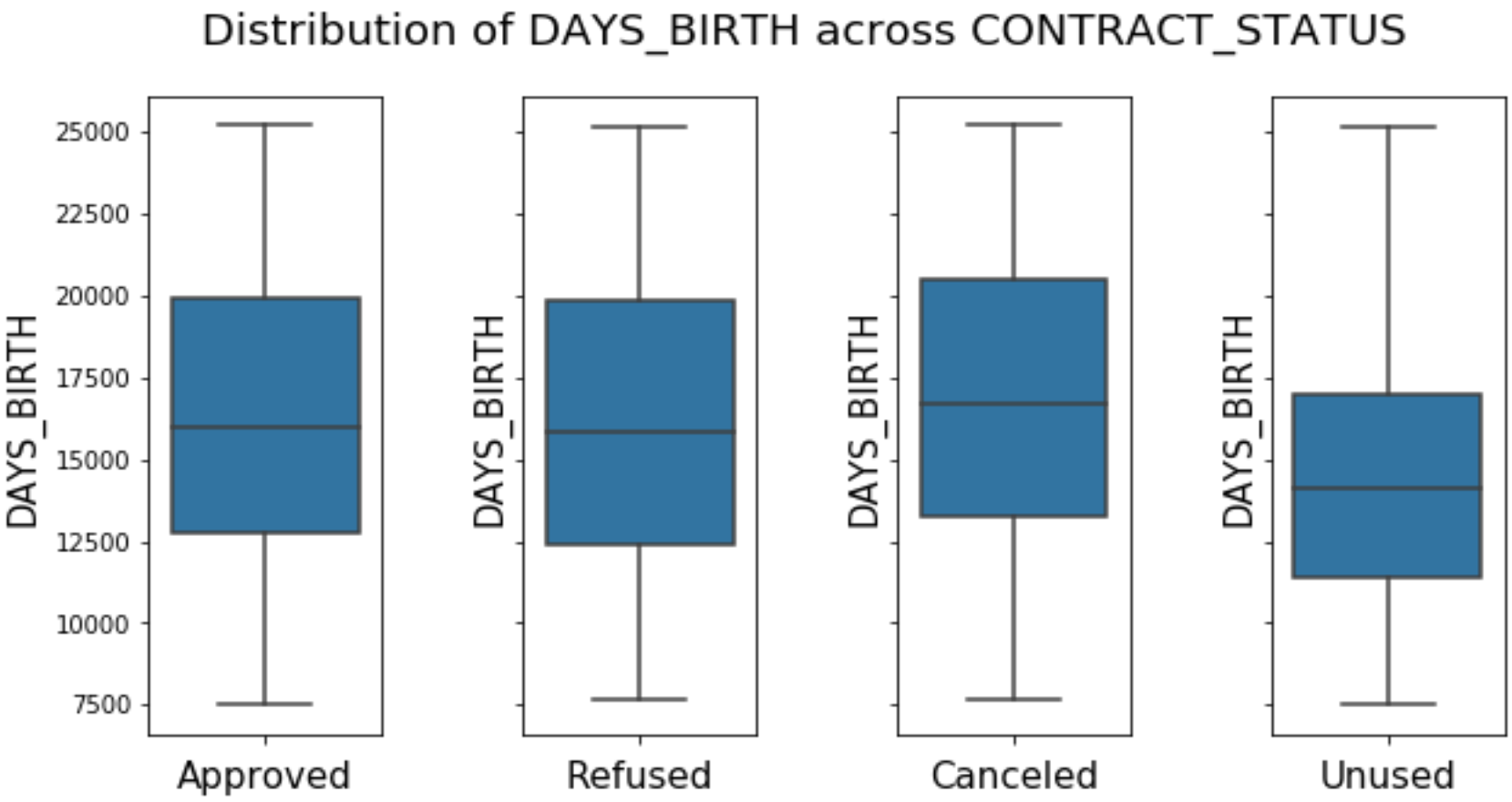
AMT_ANNUITY

- No strong comparison visible in AMT_ANNUITY with the CONTRACT_STATUS variable
- Refused & Unused have comparatively lesser Outliers count than the Approved & Canceled.



DAYS_BIRTH

- Unused loans have in general lesser DAYS_BIRTH, i.e., younger clients are often constituting Unused loans.

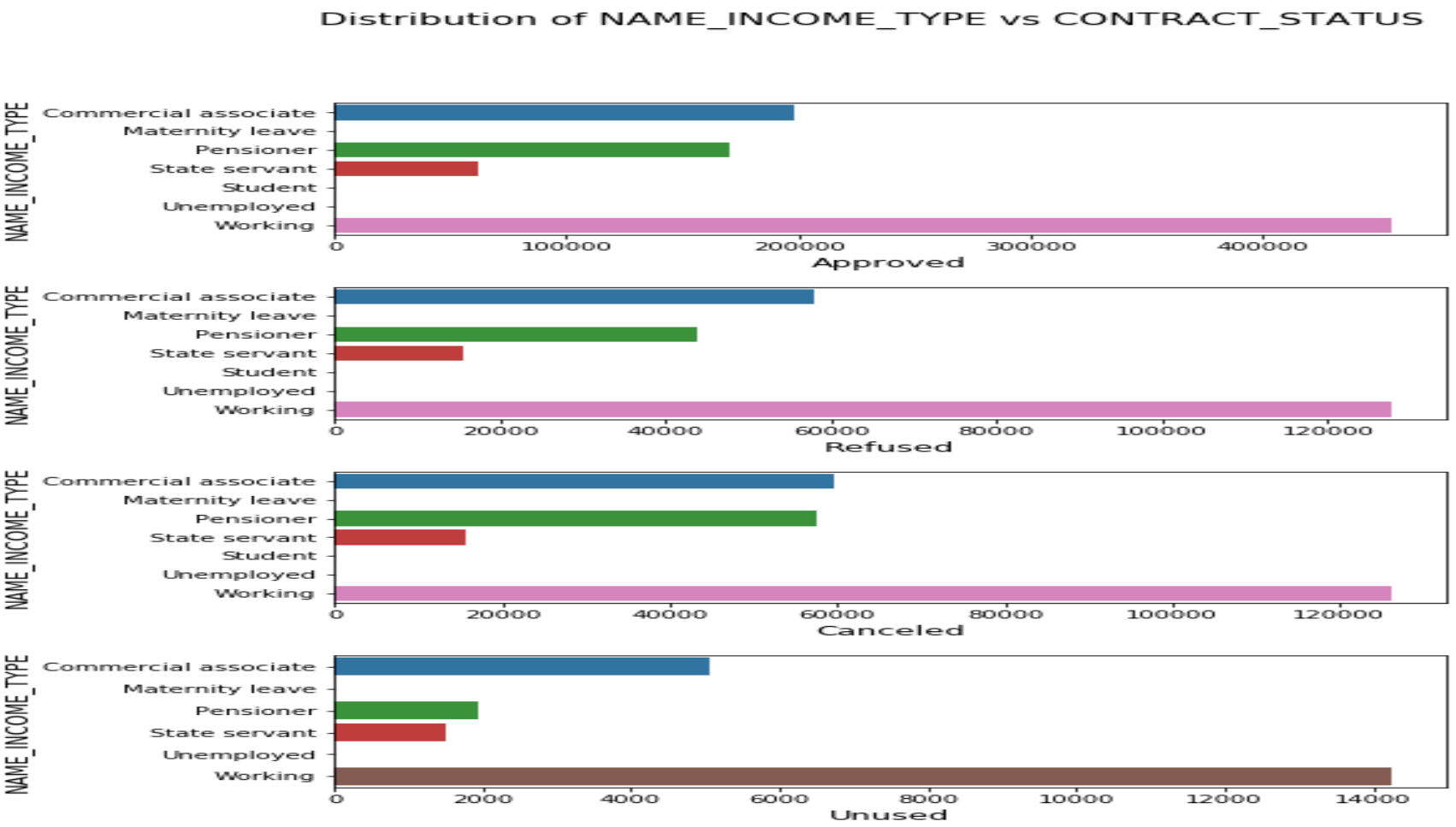


Univariate Analysis-

(Categorical Variables across Contract Status)

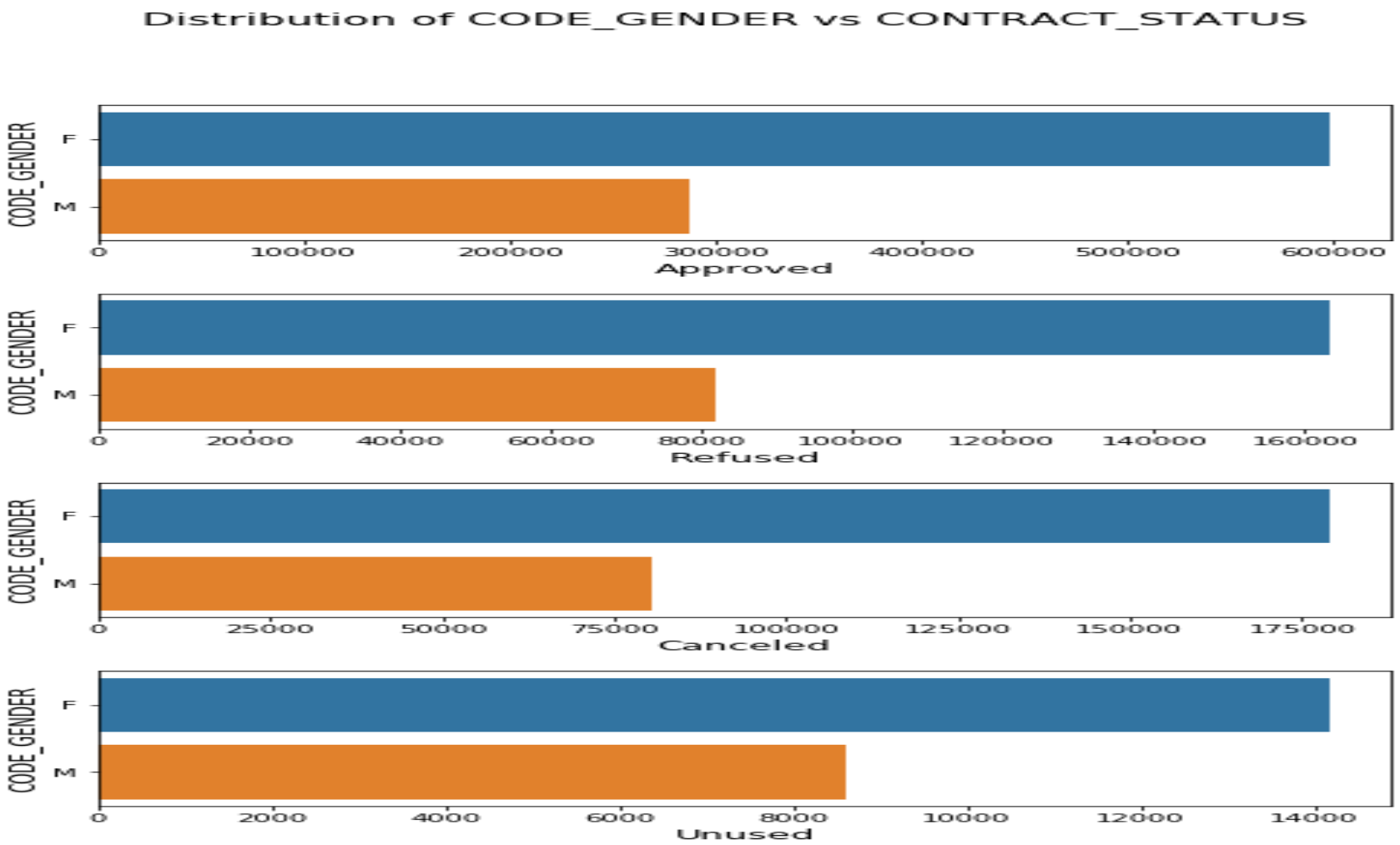
NAME_INCOME_TYPE

In Canceled category, there is a huge difference in income type Pensioner & State servant



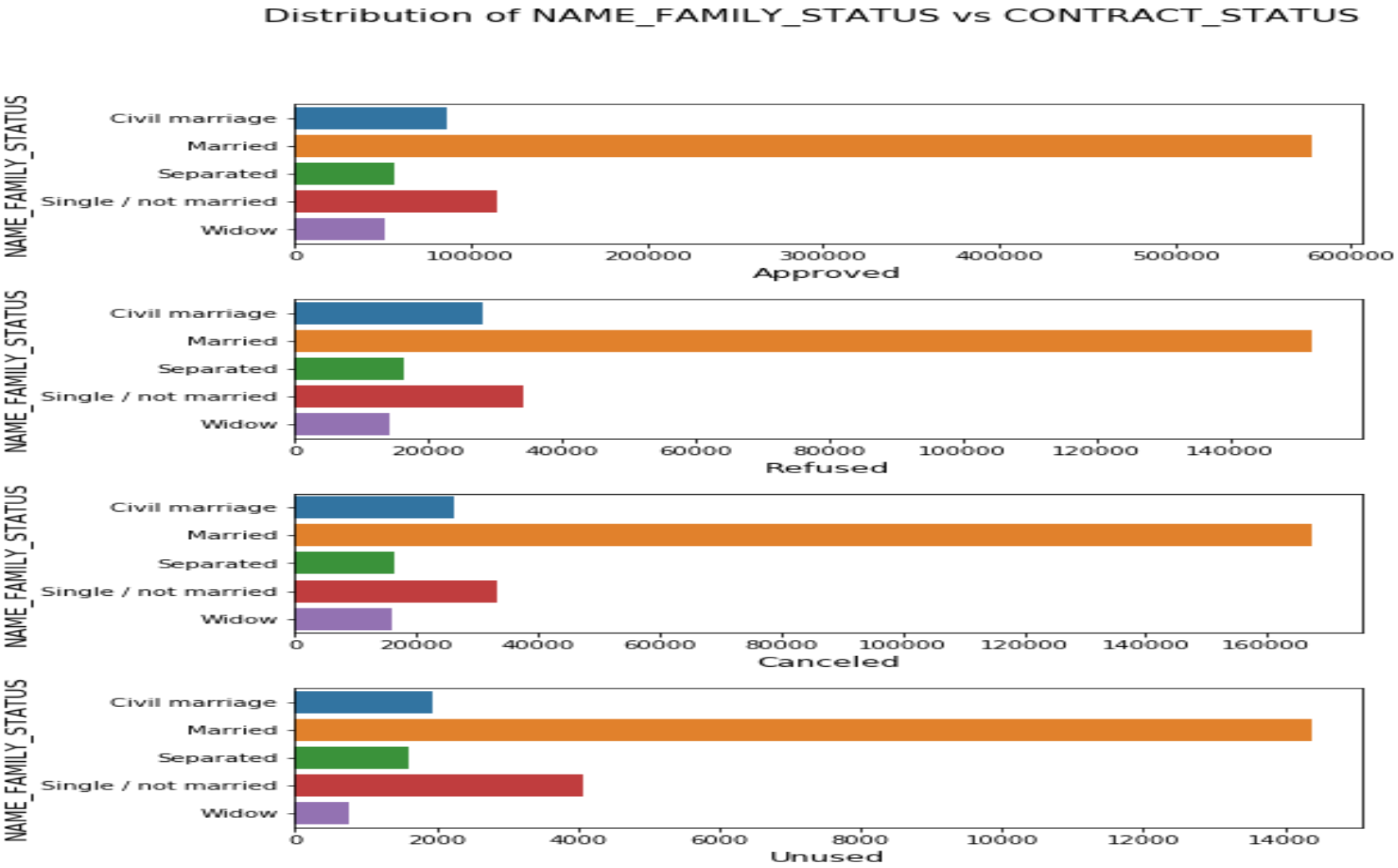
CODE_GENDER

Males have considerably lesser counts in Canceled category as compared to Females.



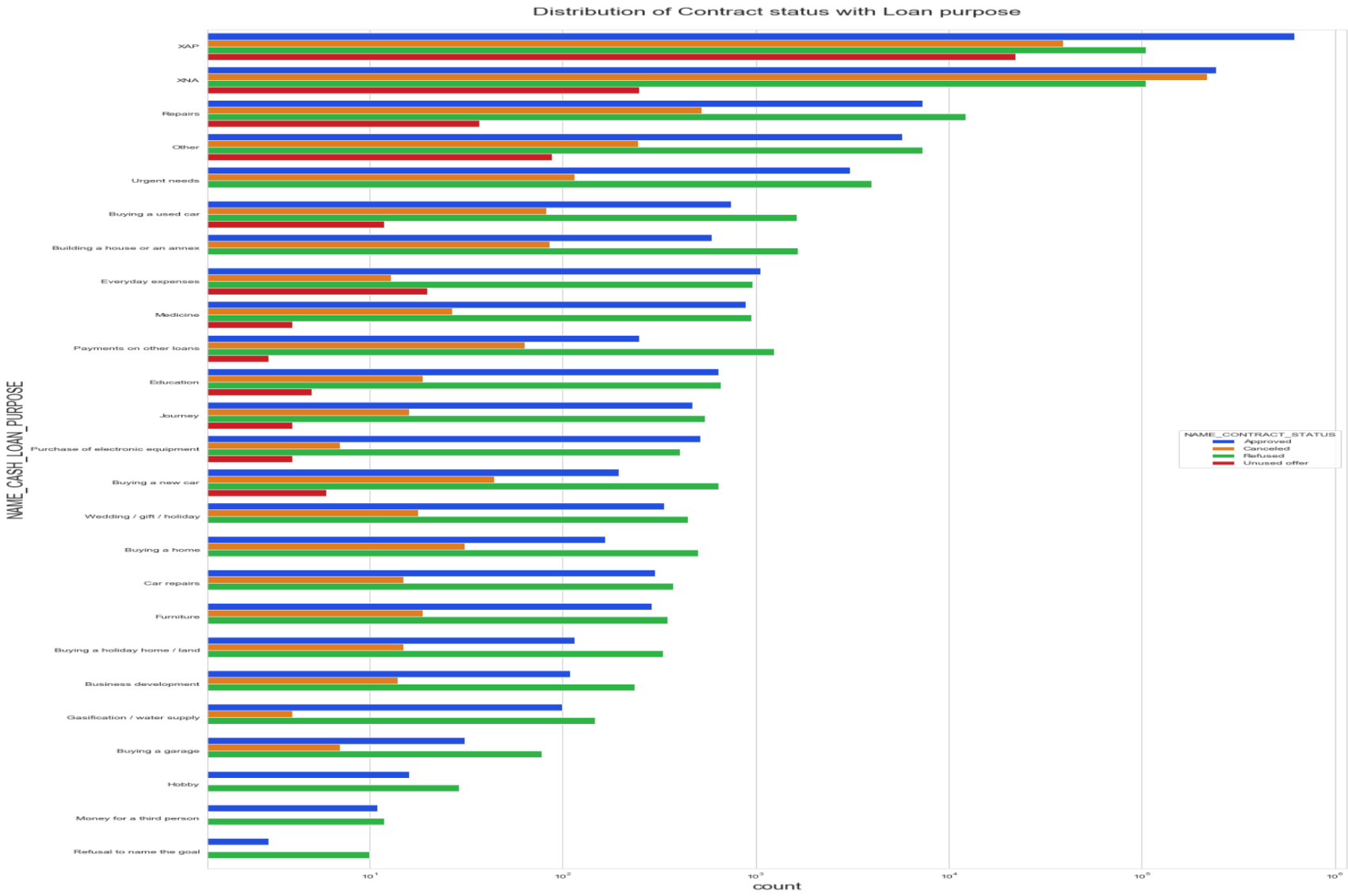
NAME_FAMILY_STATUS

1) Married people are at top in all CONTRACT_STATUS categories.
2) Single/Not-married ratio is comparatively higher in Unused category.



Bivariate Analysis

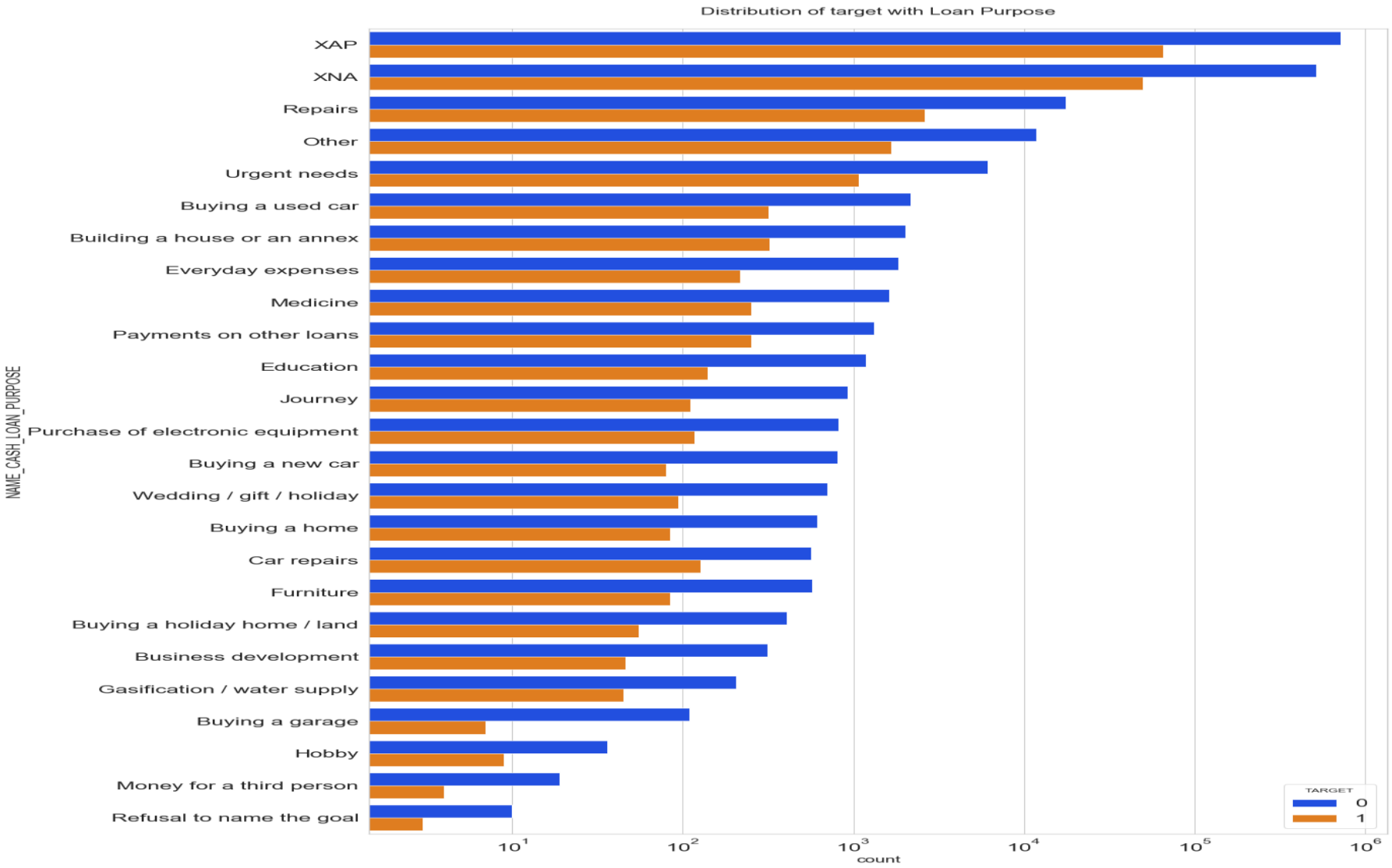
NAME_CASH_LOAN_PURPOSE v/s NAME_CONTRACT_STATUS



NAME_CASH_LOAN_PURPOSE v/s NAME_CONTRACT_STATUS

- Repairs have most reject loans.
- In Medicines we have almost equal number of approved & rejected loans.
- In Education we have almost equal number of approved & rejected loans.
- Paying other loans and buying a new car is having significant higher rejection than approves.

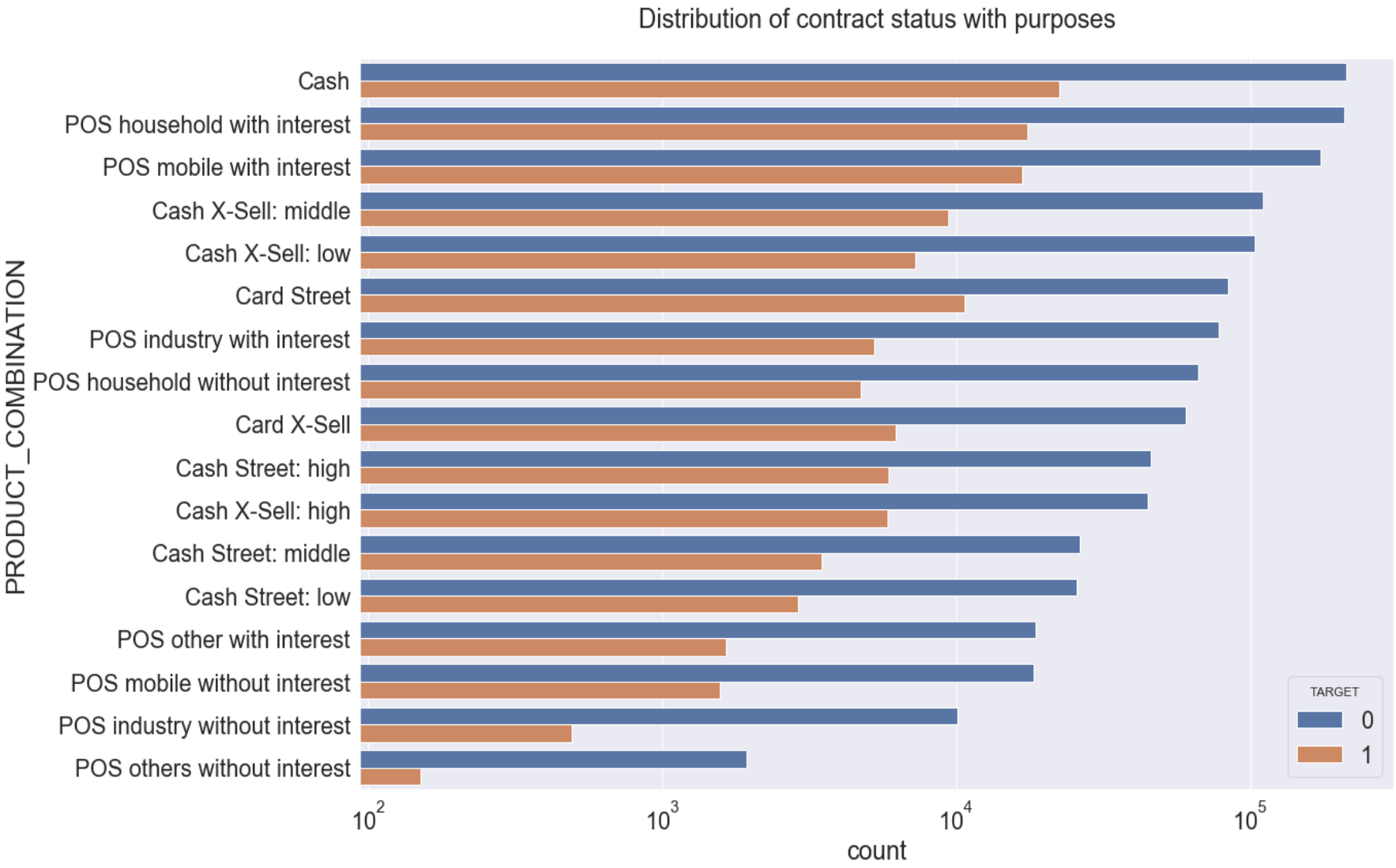
NAME_CASH_LOAN_PURPOSE v/s TARGET



NAME_CASH_LOAN_PURPOSE v/s TARGET

- Repairs have considerably high count of defaulters.
- Buying a garage has comparatively lesser defaulters and more non-defaulters.
- Other loan purposes where non-defaulters count are significantly greater than the defaulter count are - Business Development, Buying a home, Buying a land, Buying a new car.

PRODUCT_COMBINATION v/s TARGET

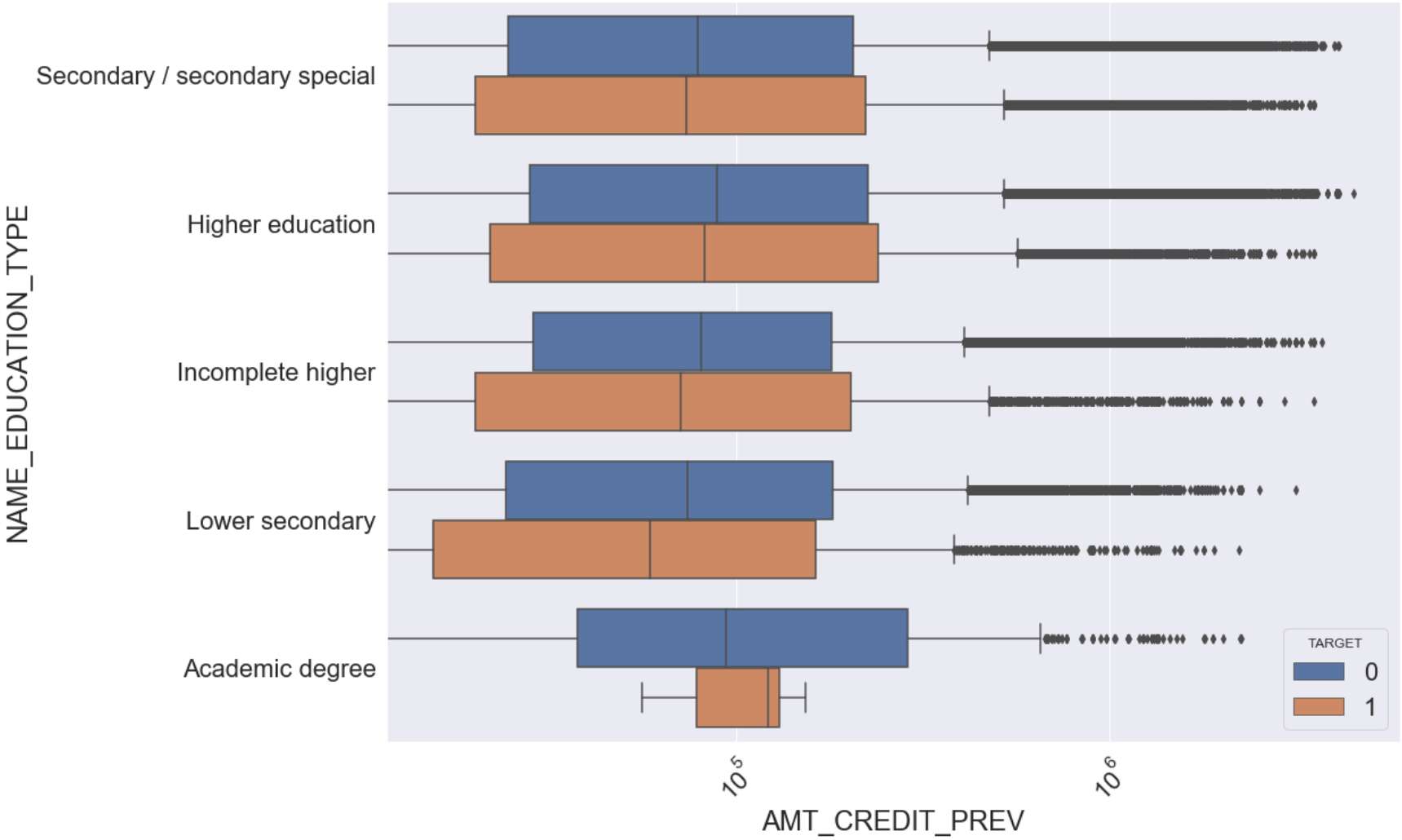


PRODUCT_COMBINATION v/s TARGET

- POS others without interest, POS industry without interest Product combinations should be targeted as their defaulter percentage is low compared to others.

- Highest defaulter rate is for Product combination cash and highest non-defauters are also from Cash product combination. This may be because Cash combination count is high in our data set.

AMT_CREDIT_PREV v/s NAME_HOUSING_TYPE v/s TARGET



AMT_CREDIT_PREV v/s NAME_HOUSING_TYPE v/s TARGET

- Office apartment, Co-op apartment have higher AMT_CREDIT for defaulters.
- Municipal apartment have higher AMT_CREDIT for non-defaulters.
- So, banks can focus on giving loans to House/apartment, Municipal Apartments, as they show positive results being non-defaulters.

CONCLUSION

- Higher Education:Businessman, Higher Education:Maternity leave and Higher Education:students should be targeted as they pay their loan on time as per the data.
- Bank must focus on Academic degree Education Type non-working type as they are having higher number of successful payments on time.
- Loan purpose Repair is having higher number of unsuccessful payments on time.
- Buying a garage, Business development, Buying land, Buying a new car and Education should be targeted as they are having minimum payment difficulties.
- POS others without interest, POS industry without interest Product combinations should be targeted as their defaulter percentage is low compared to others.
- Banks can focus on giving loans to House/apartment, Municipal Apartments, as they show positive results being non-defaulters.

Thanks for reading !