CREDIT EDA ASSIGNMENT

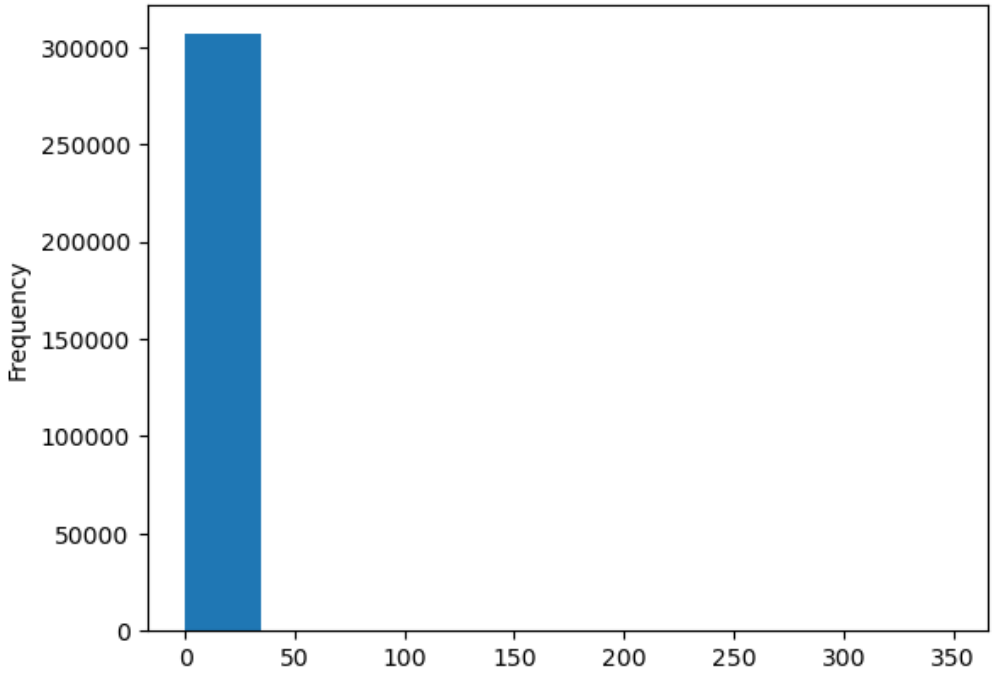
BY

HARSHAVARDHINI J S

Application_data.csv

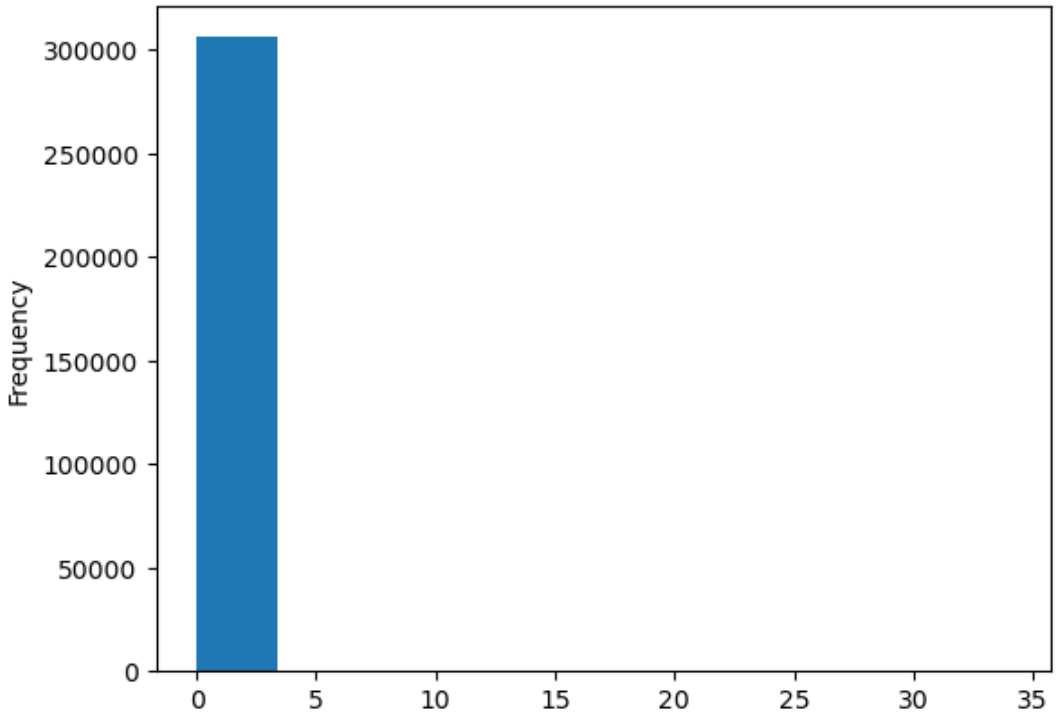
Handling missing values

Column: OBS_30_CNT_SOCIAL_CIRCLE



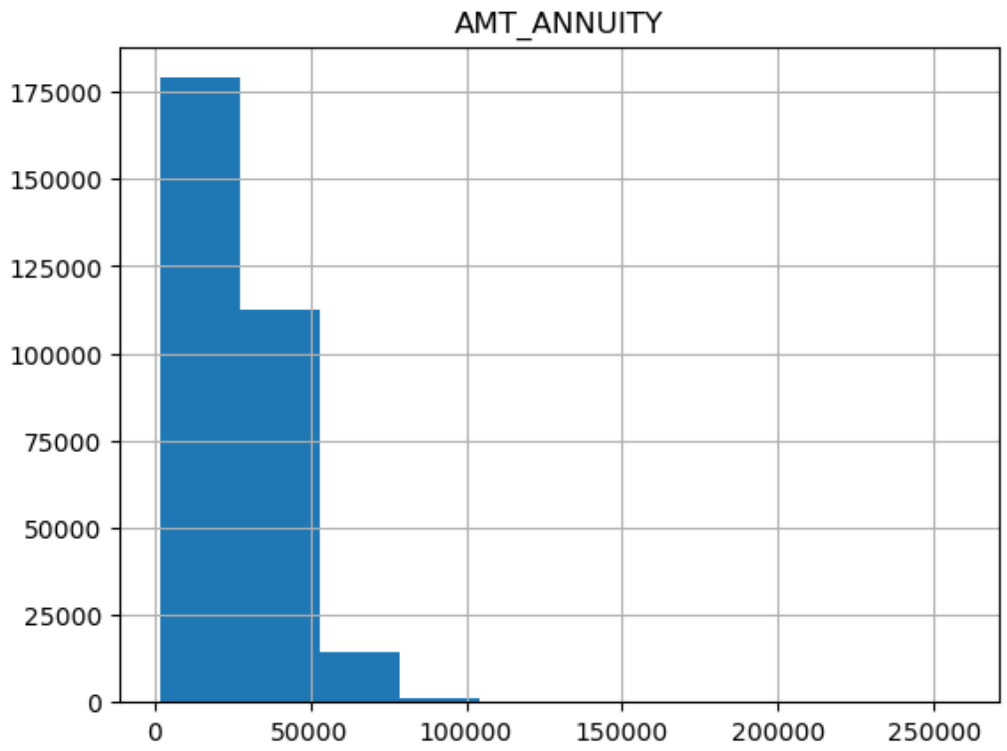
The histogram plot shows that the column has a skewed distribution and hence the suggested imputation method for the null values can be replaced by mode=0 which is the maximum frequency value.

Column:DEF_30_CNT_SOCIAL_CIRCLE



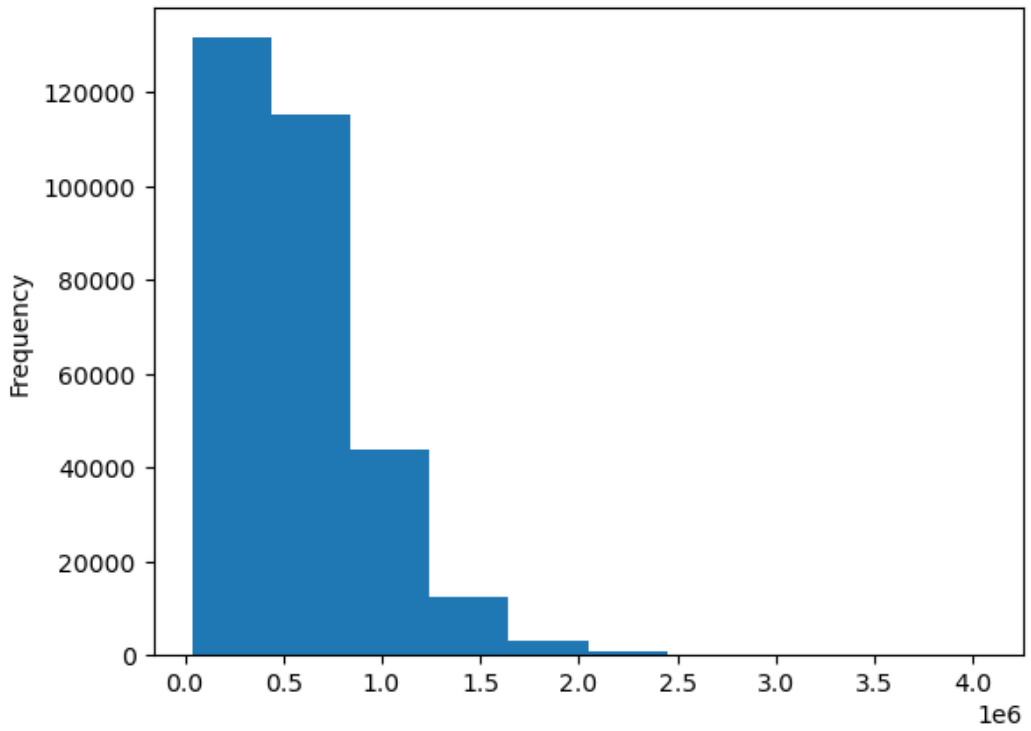
The suggested imputation method for the null values in the column DEF_30_CNT_SOCIAL_CIRCLE can be replaced by mode=0 which is the maximum frequency value.

Column: AMT_ANNUITY



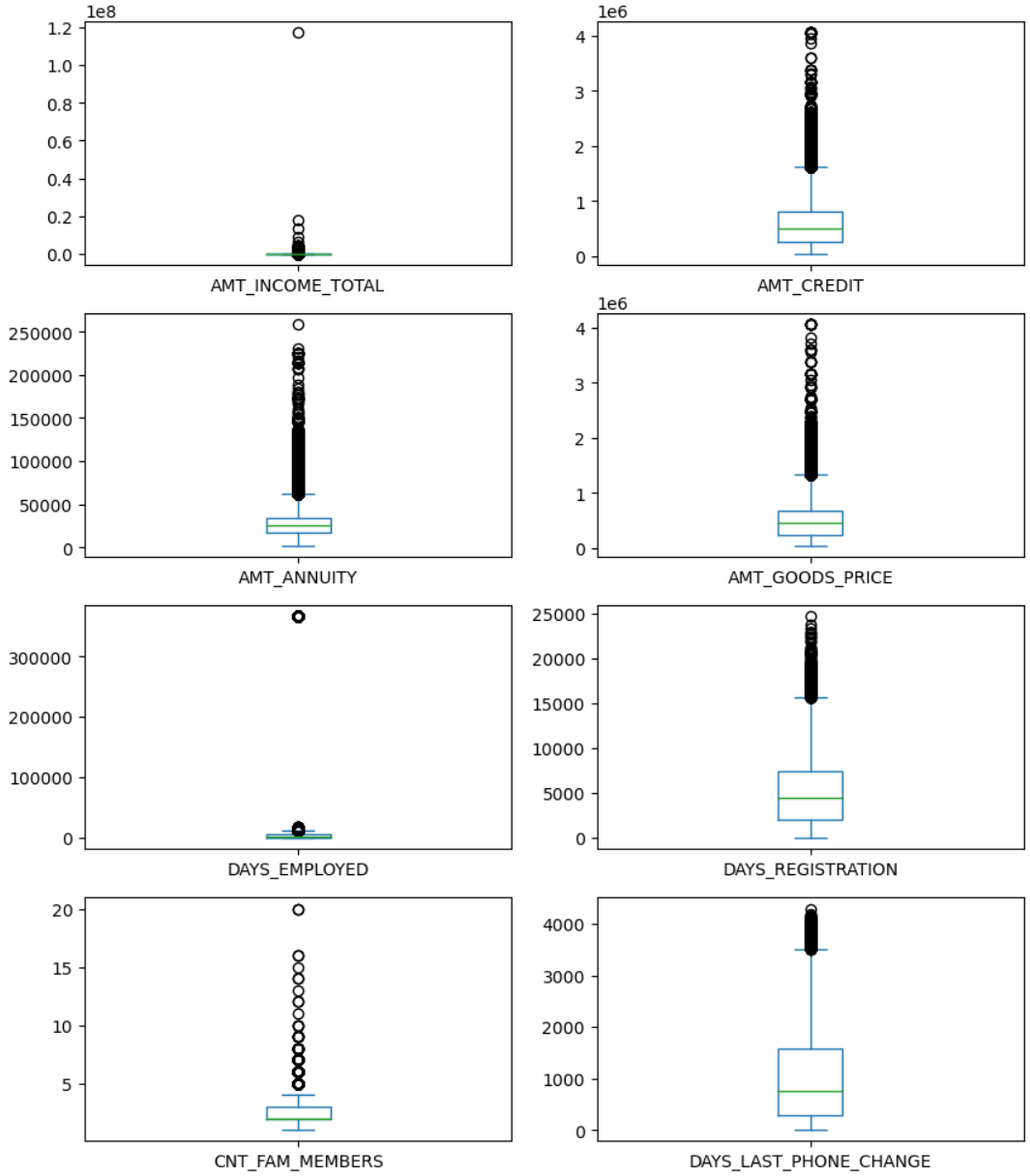
The column has an asymmetrical distribution and hence the mean values is skewed. Since the null values are very less in percentage the columns can be either dropped off or can be imputed with median value = 24903.

Column: AMT_GOODS_PRICE



The column AMT_GOODS_PRICE has an asymmetrical distribution and hence the mean values is skewed. Since the null values are very less in percentage the columns can be either dropped off or can be imputed with median value = 450000

Handling outliers

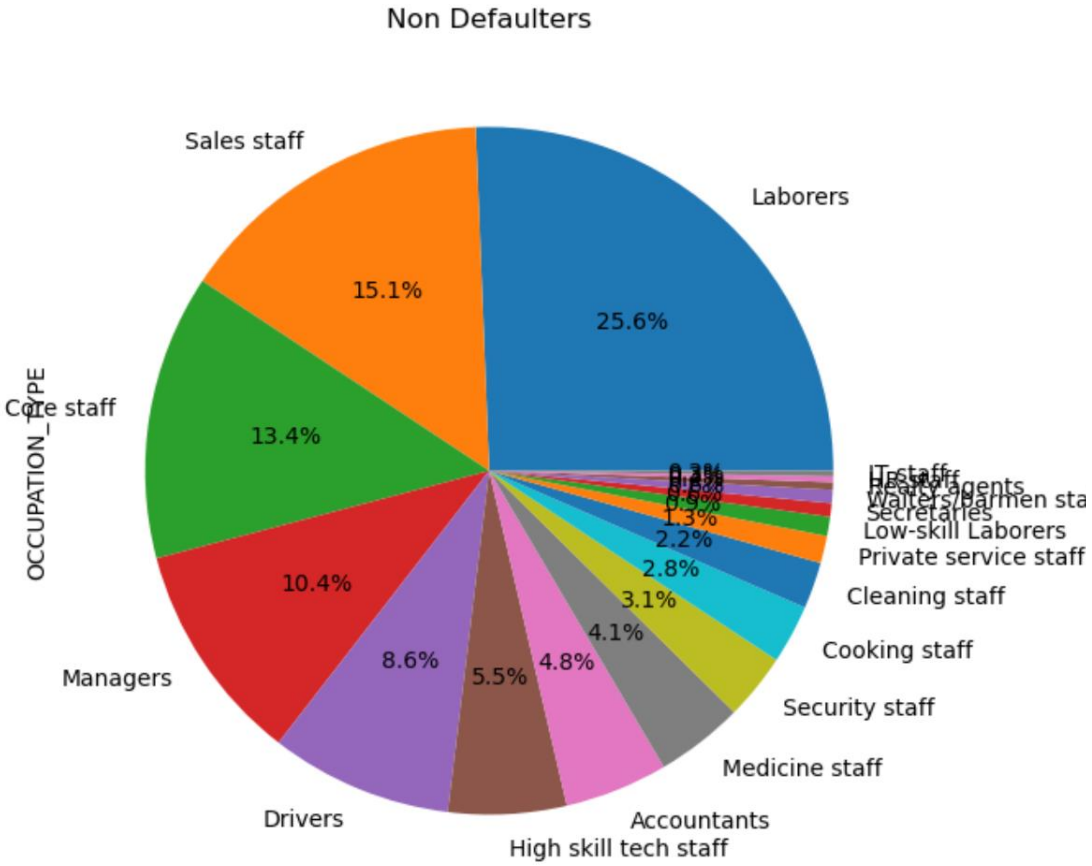


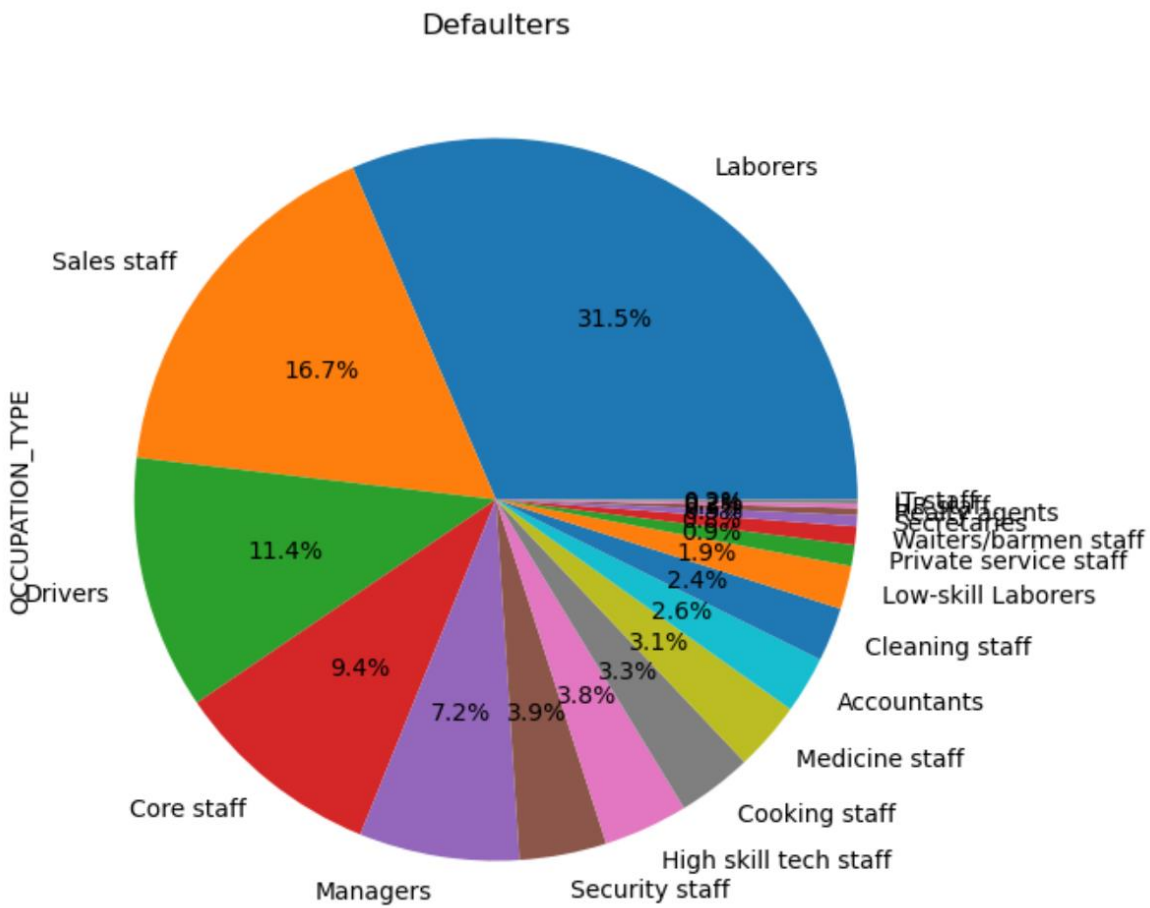
SUGGESTIONS

- All the variables in num_cols exhibit very low percentage of outliers which is less than 5% hence these outliers can be dropped off so that there is not much effect on the prediction of the results.
- The variable DAYS_EMPLOYED shows 18.33% of outliers which can be capped with the upper and lower bound of the interquantile range for further analysis.

Univariate Analysis

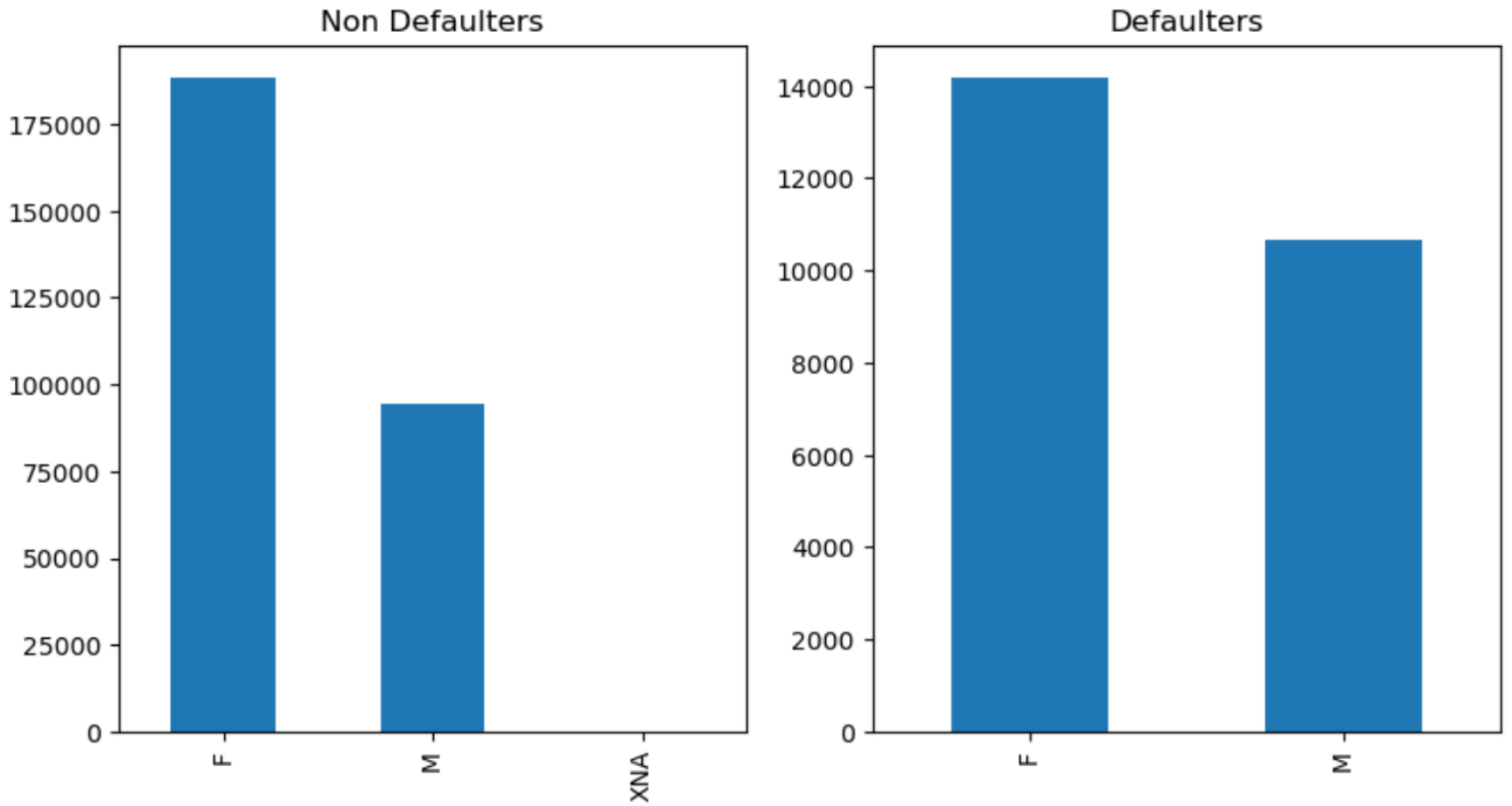
OCCUPATION_TYPE





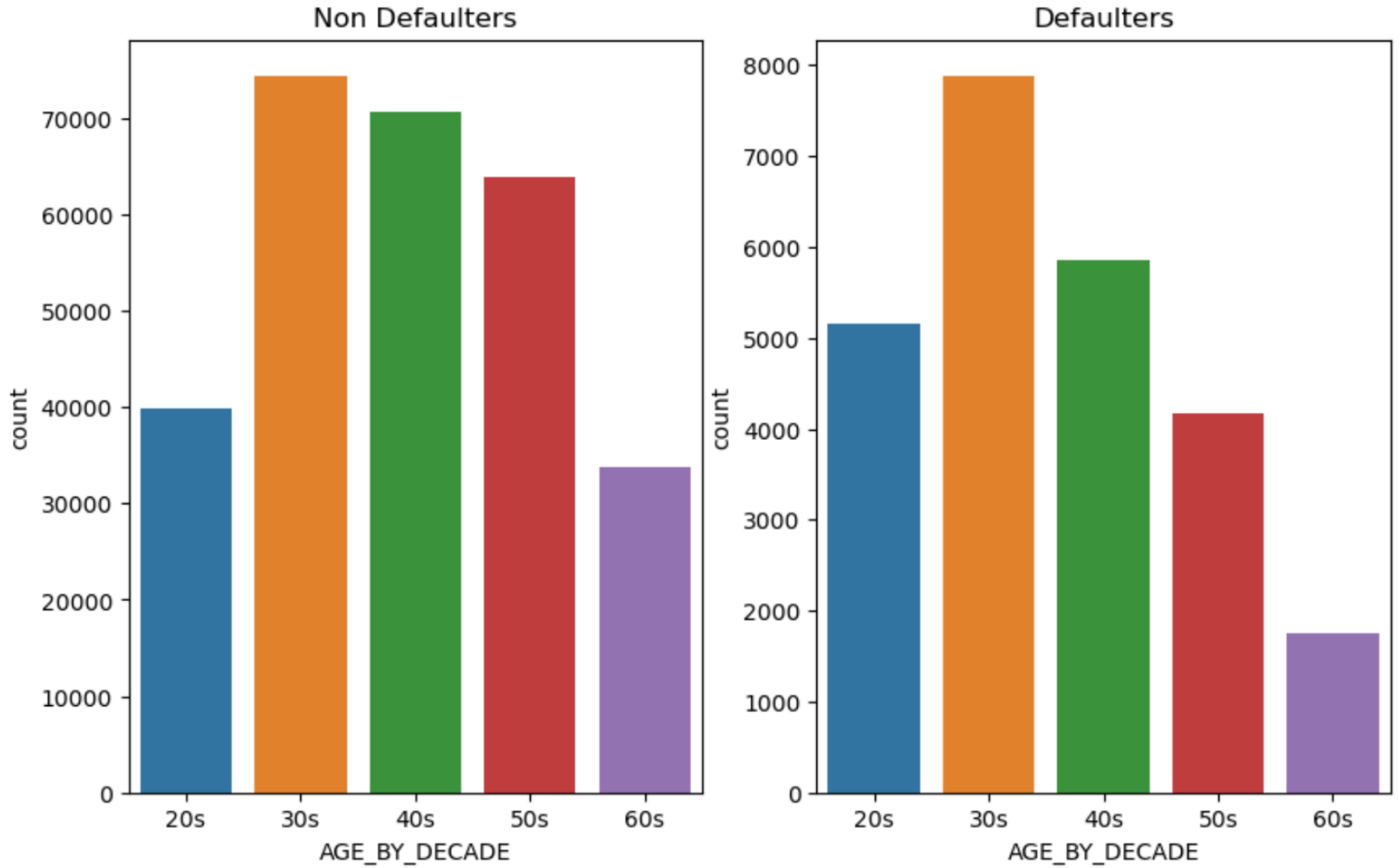
- Laborers, Sales staff and Drivers contribute to more than 50% of defaulters list.
- Laborers, Sales staff, Drivers, Core staff and Managers contribute to more than 75% of defaulters list.
- 31.5% of defaulters are with occupation type as Laborers
- 25.6% of Non Defaulters are with occupation type as Laborers

CODE_GENDER



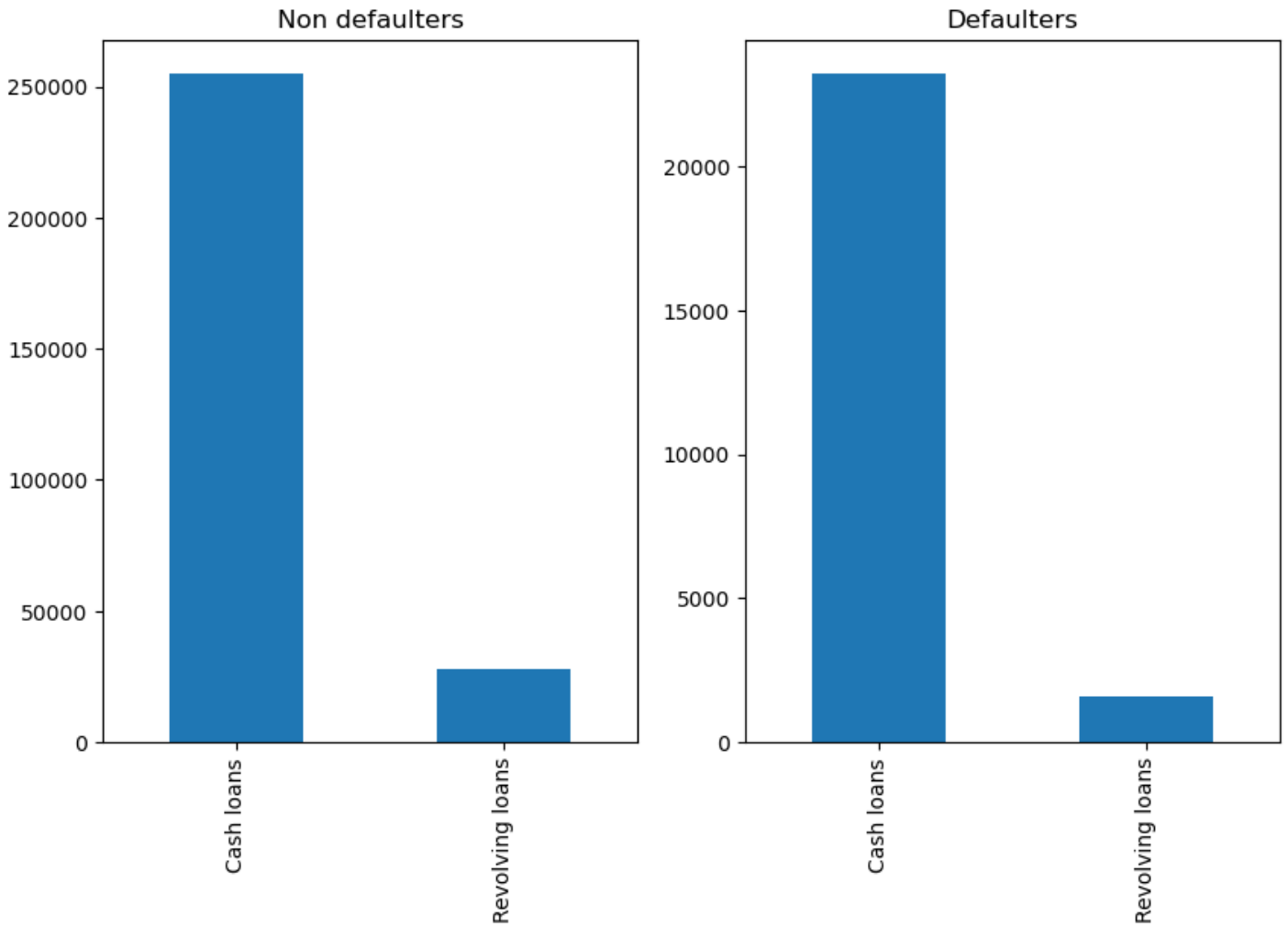
- The bar plot depicts that the female gender were higher as compared to the male in applying credit loans
- The female gender percentage is 66.6% while the male percentage is 33.3% in case of Non defaulters
- The female gender percentage is 57.0% while the male percentage is 42.9% in case of defaulters The female receiving loan tend to repay in higher percentage than male receiving loans

AGE_BY_DEFAULT



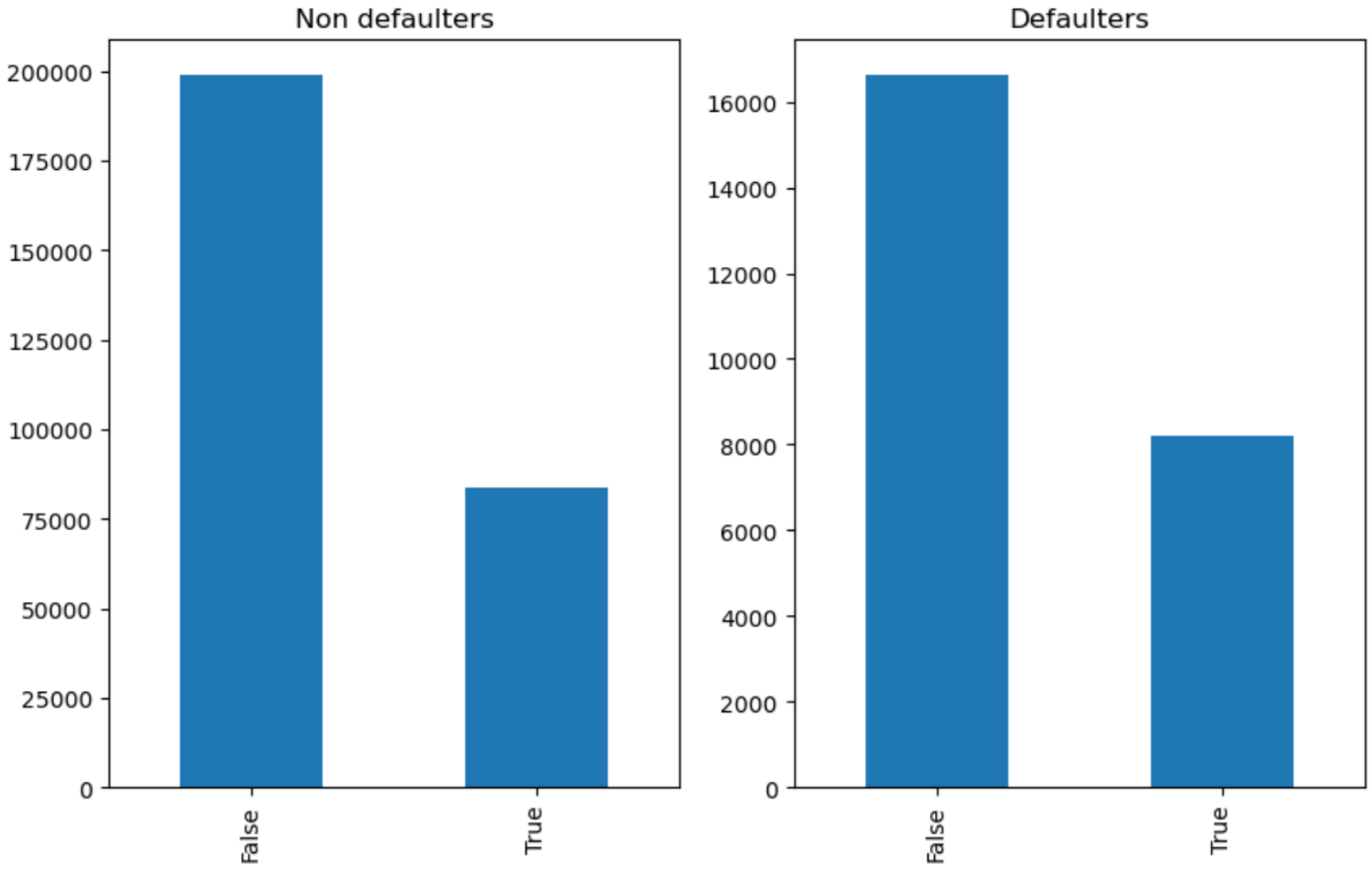
- The people in the age category of mid 30s and 40s contribute nearly 50% of the total people applying loans.
- The plot shows that 30s age category people mostly fail to repay the loan as compared to other age categories.
- The younng and the older aged people are less likely to face difficulties in repaying the loan.

NAME_CONTRACT_TYPE



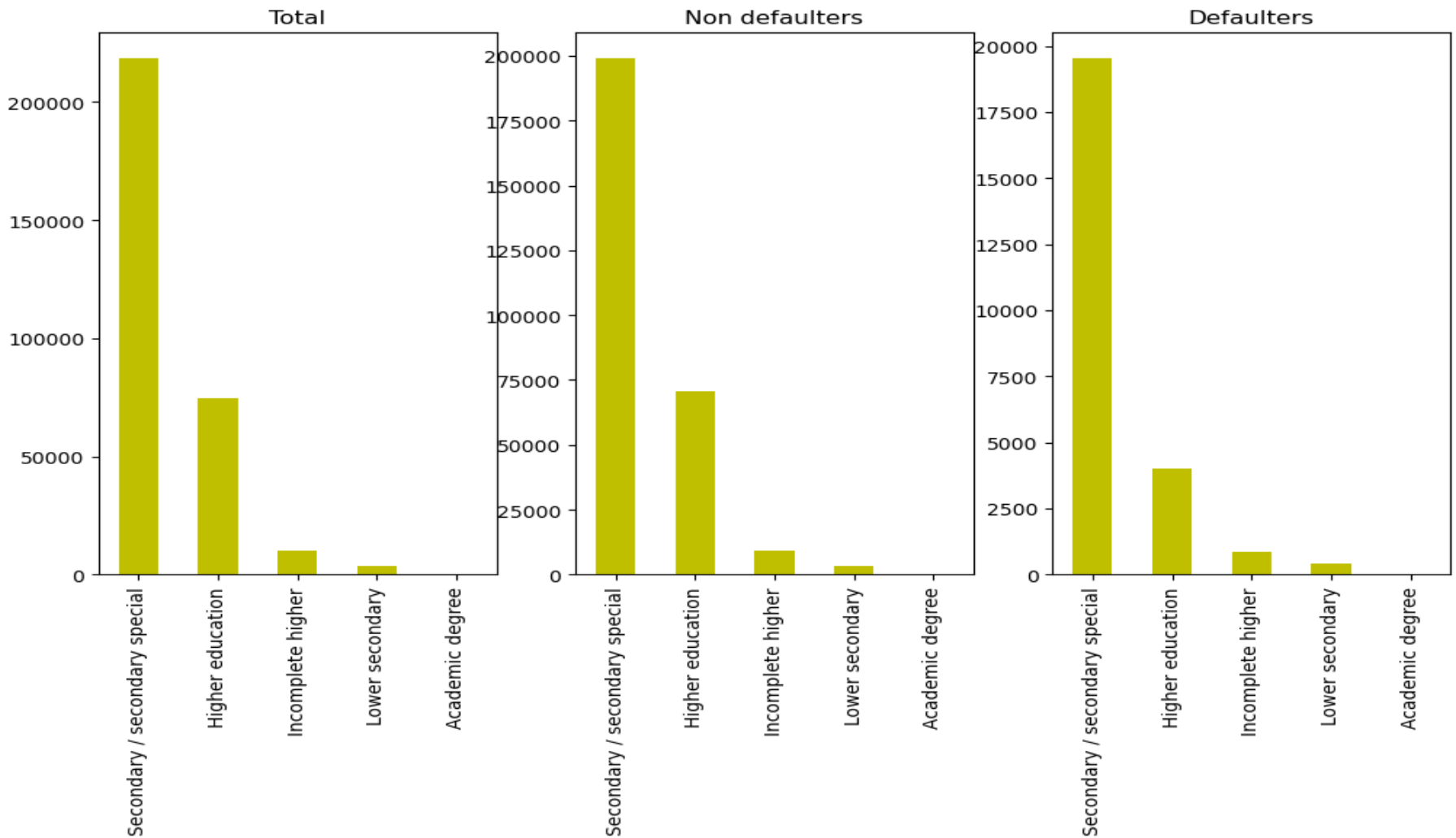
The plot shows that most of the people opted for cash loans. Only a small proportion of people opted revolving loans for both defaulters and non defaulters.

HAS_CHILDREN



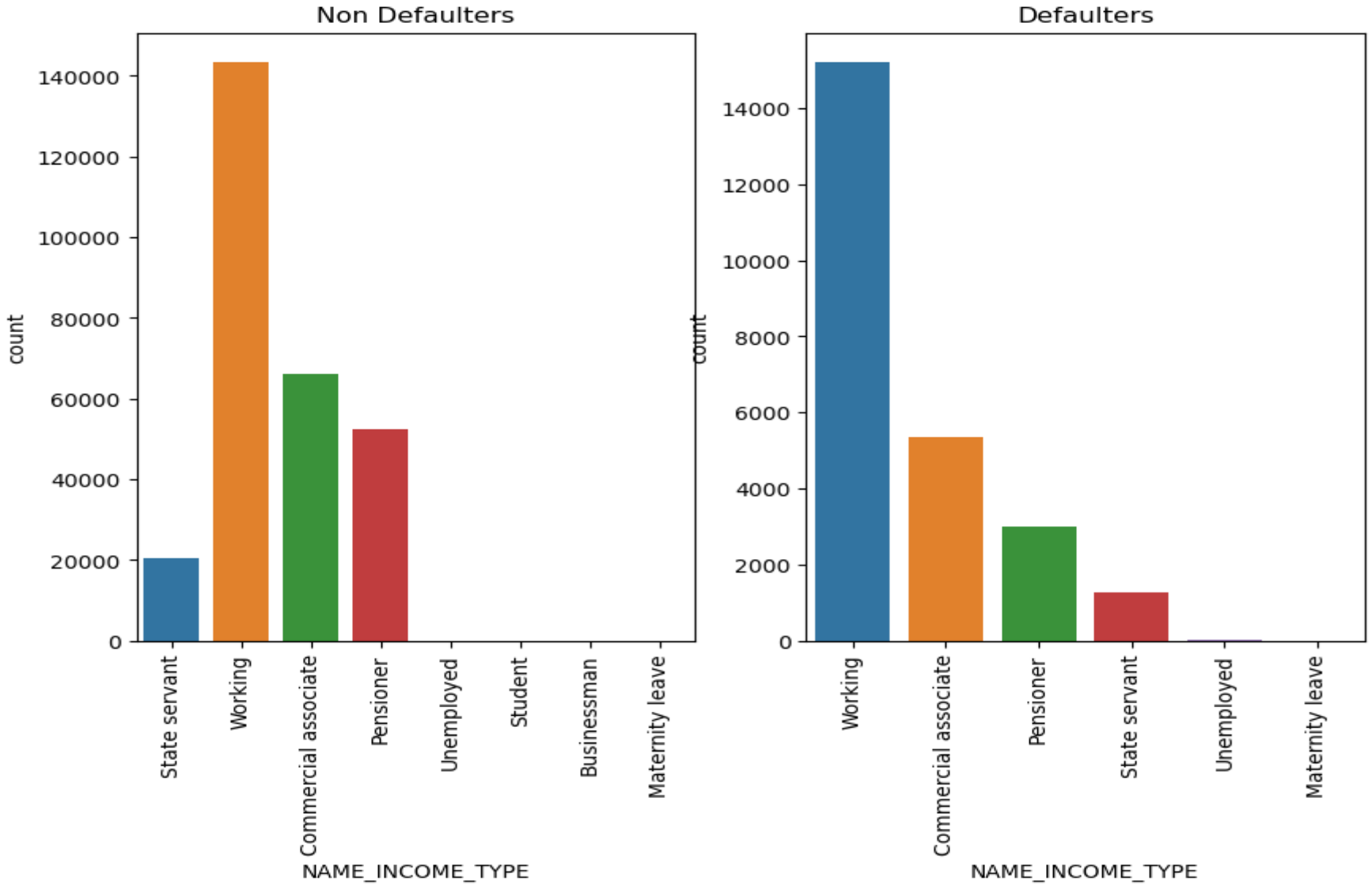
From the plot it is shown that both the defaulters and non defaulters the person who does not have children tends to apply for bank loan than people who has children.

NAME_EDUCATION_TYPE



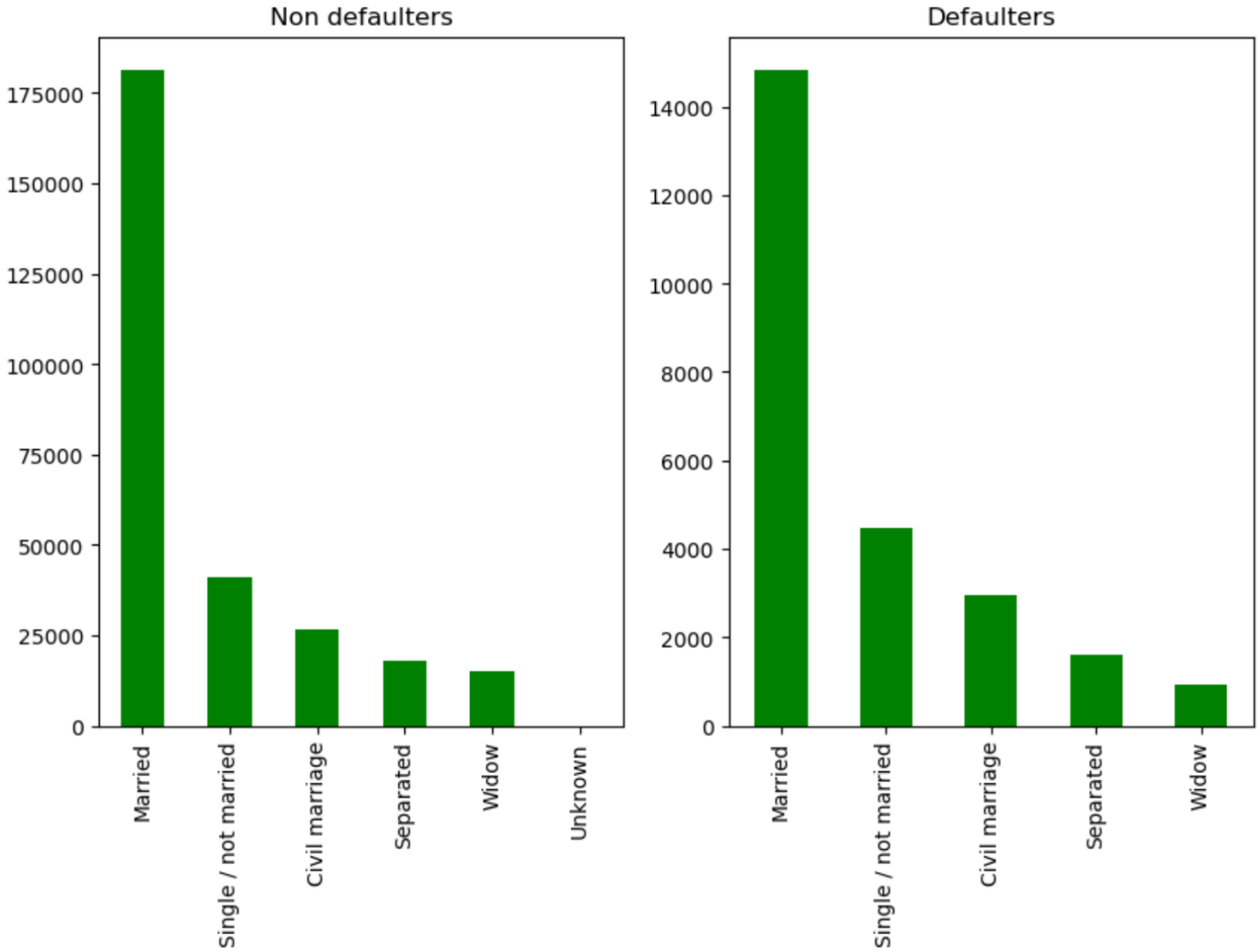
From the bar graph it is observed that people with Secondary education level apply for loans than other qualified people and they also face difficulties in repaying the loan.

NAME_INCOME_TYPE



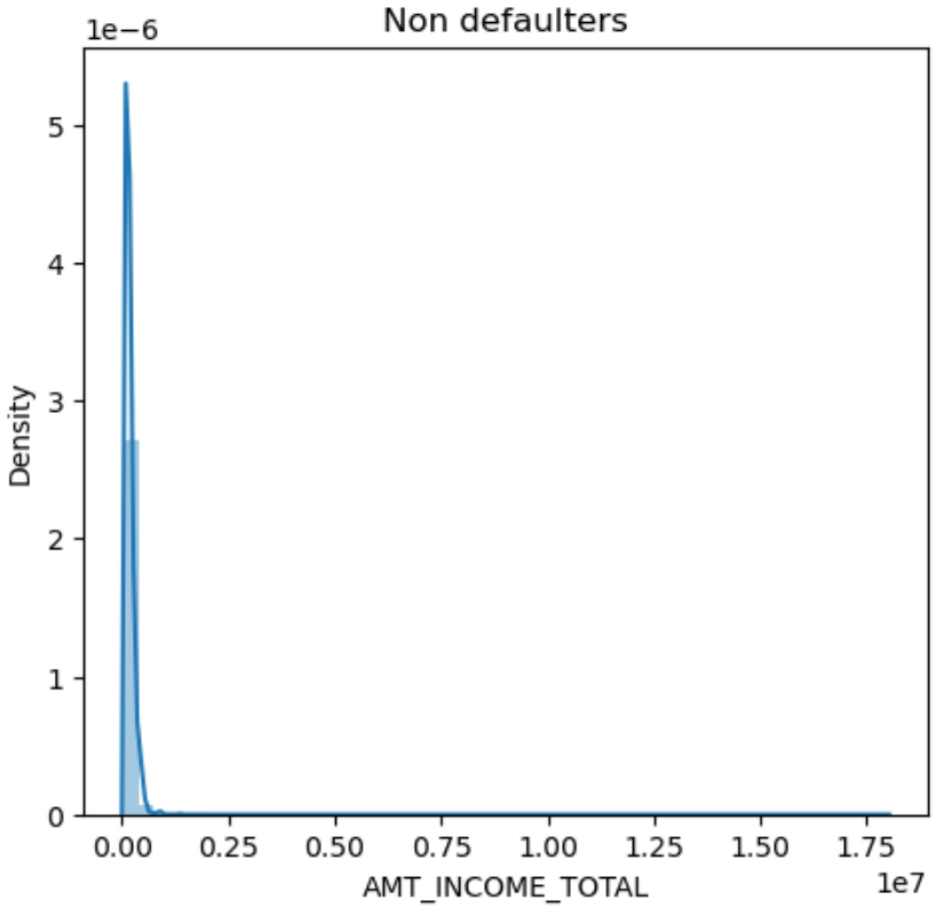
- The plot shows that the Working professionals are in higher proportions for bank loan requirements than any other professsionals and there is a higher risk for loan default.
- The pensioner people and State servant applying for loan has lesser risk to default.

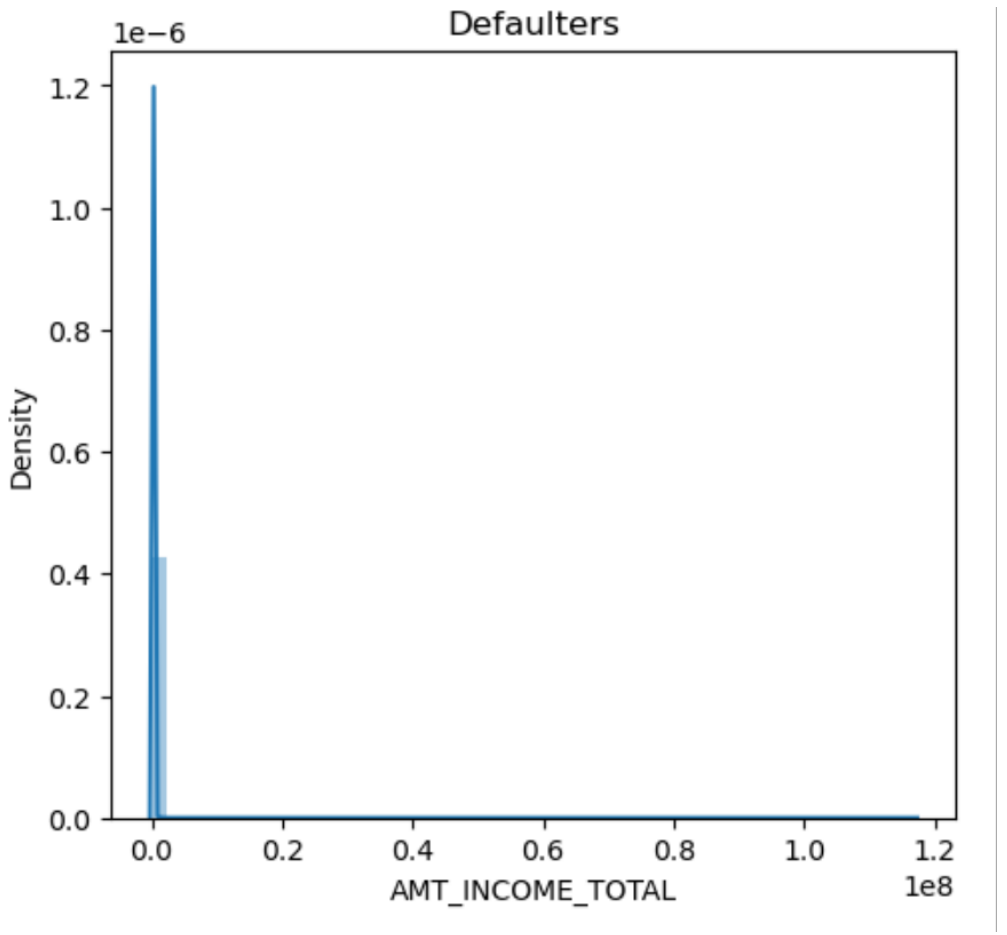
NAME_FAMILY_STATUS



The graph shows that the Married people require bank loan than others in case of both defaulters and non defaulters and Widowed people who applies for loan has lesser risk to loan default.

AMT_INCOME_TOTAL

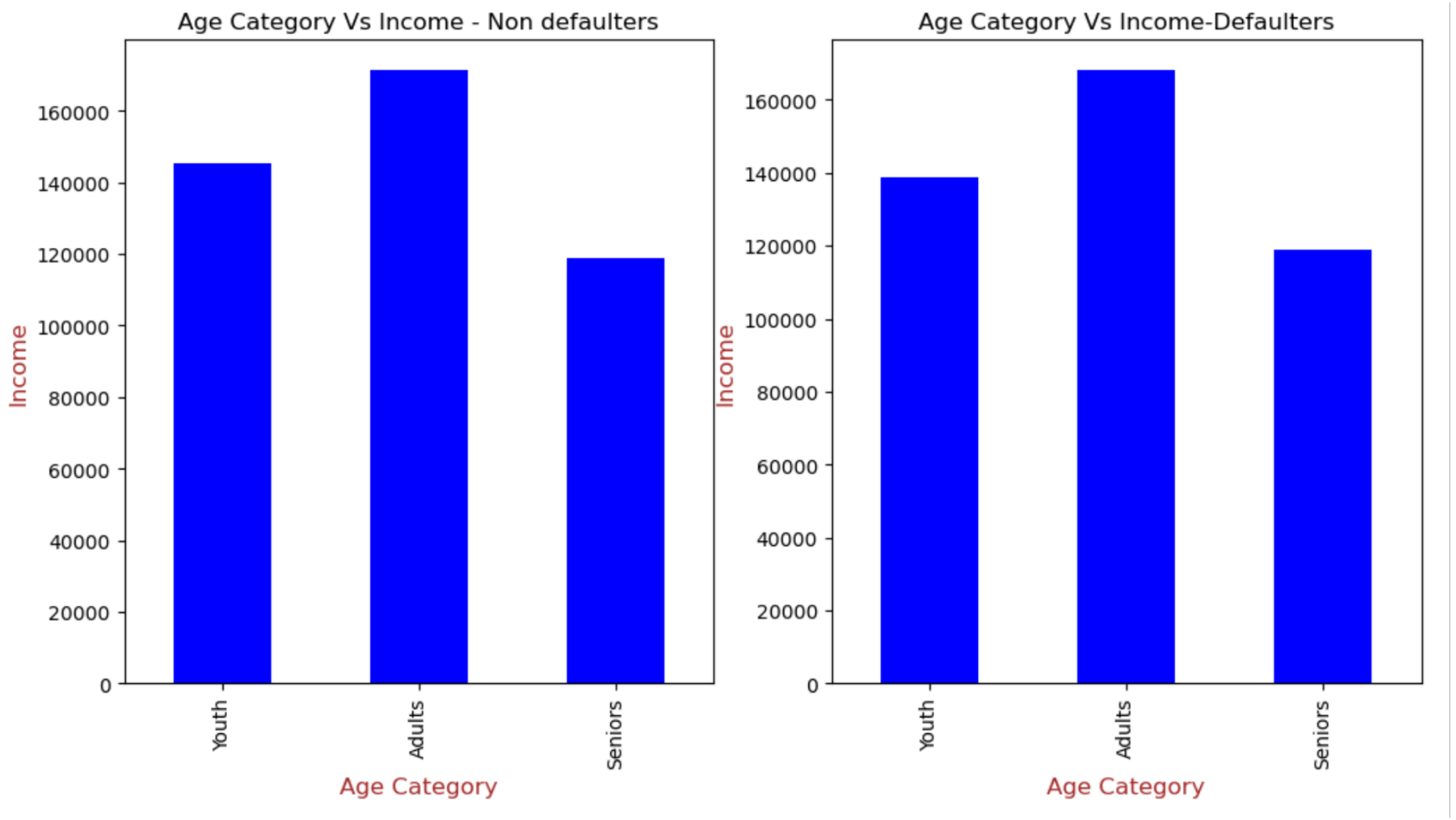




- In summary, the plot suggests that there is a difference in the distribution of income between non-defaulters and defaulters.
- Defaulters have a wider range of income levels, including some with high incomes, while non-defaulters have income levels more concentrated in the lower to middle range.

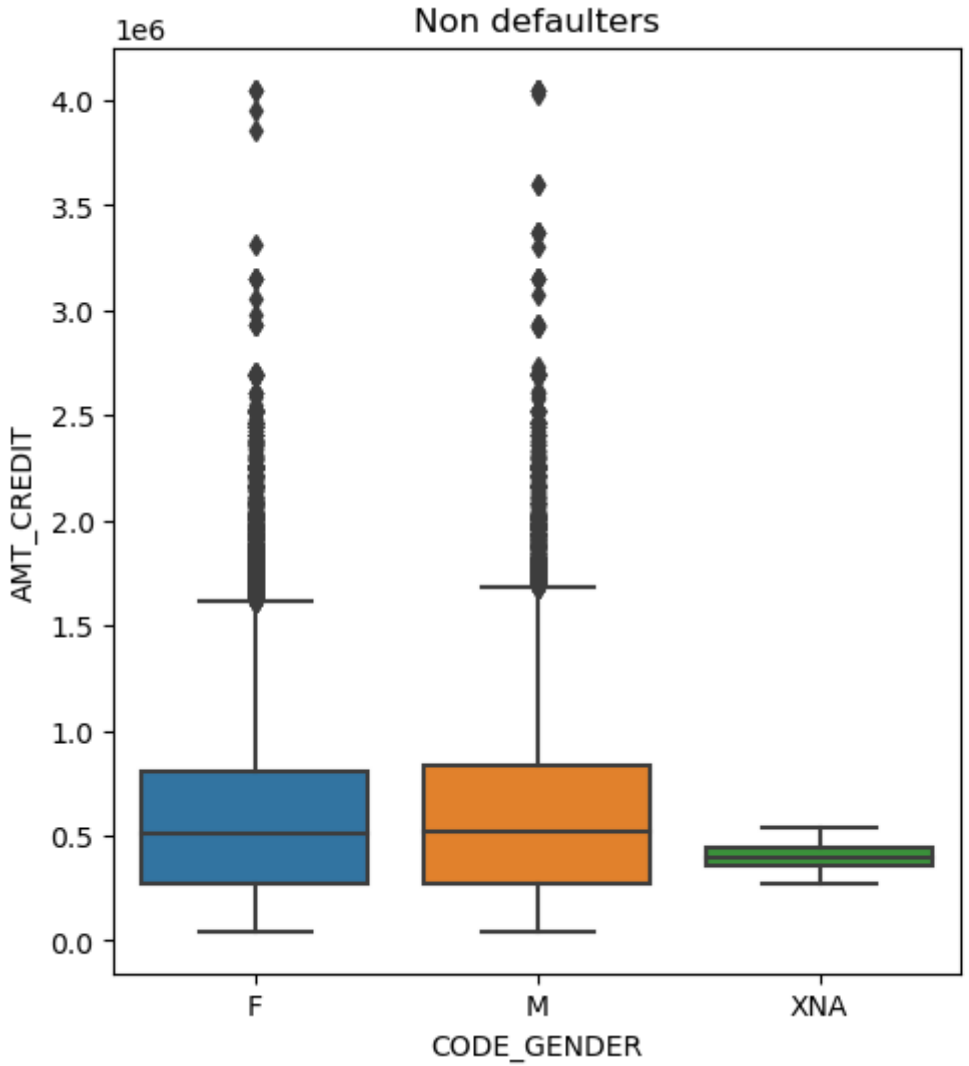
Bivariate and Multivariate analysis

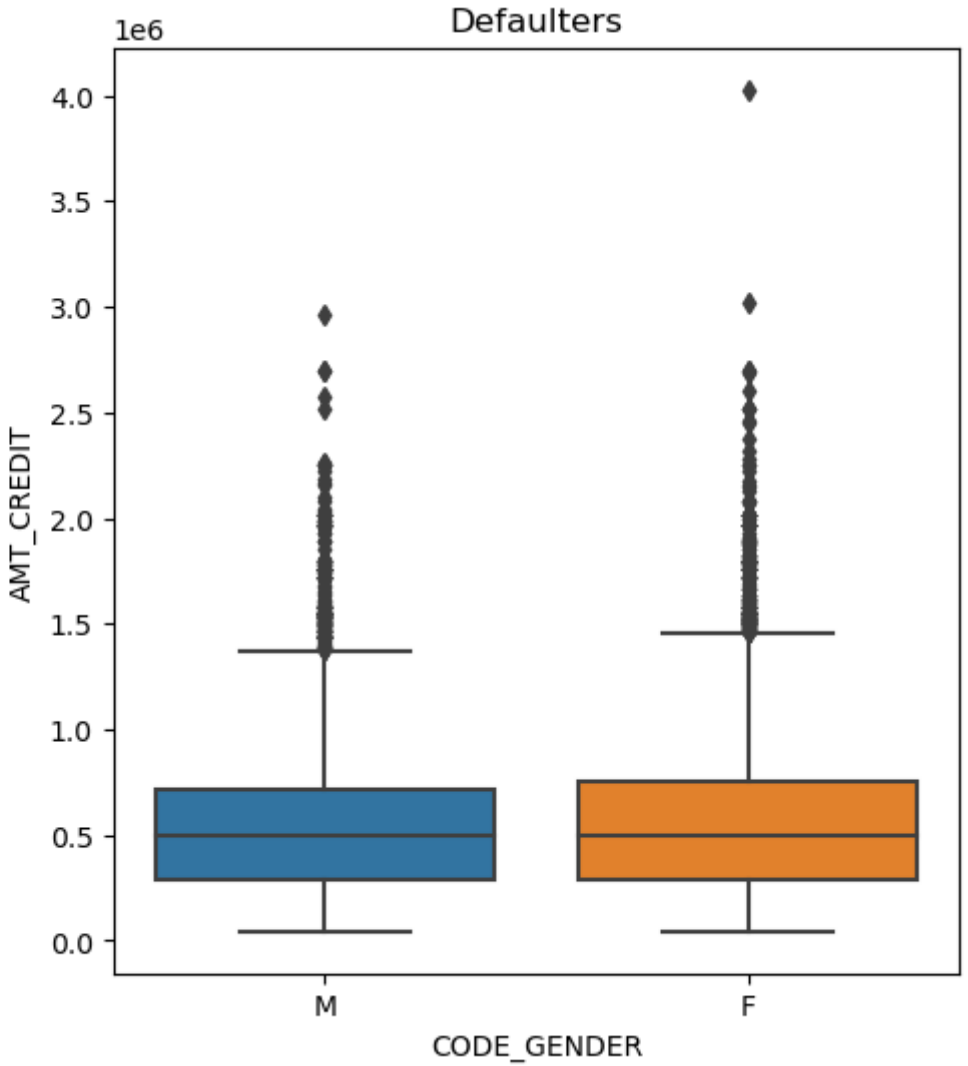
Categorizing the Income of the persons with their Age



- In both plots, there is a clear trend where the average income tends to increase with age, up to a certain point, and then it starts to stabilize or slightly decrease in the older age categories.
- Younger age categories have lower average incomes, while older age categories generally have higher average incomes.
- For non-defaulters, the average income generally increases with age. For defaulters, the trend is similar but less pronounced.

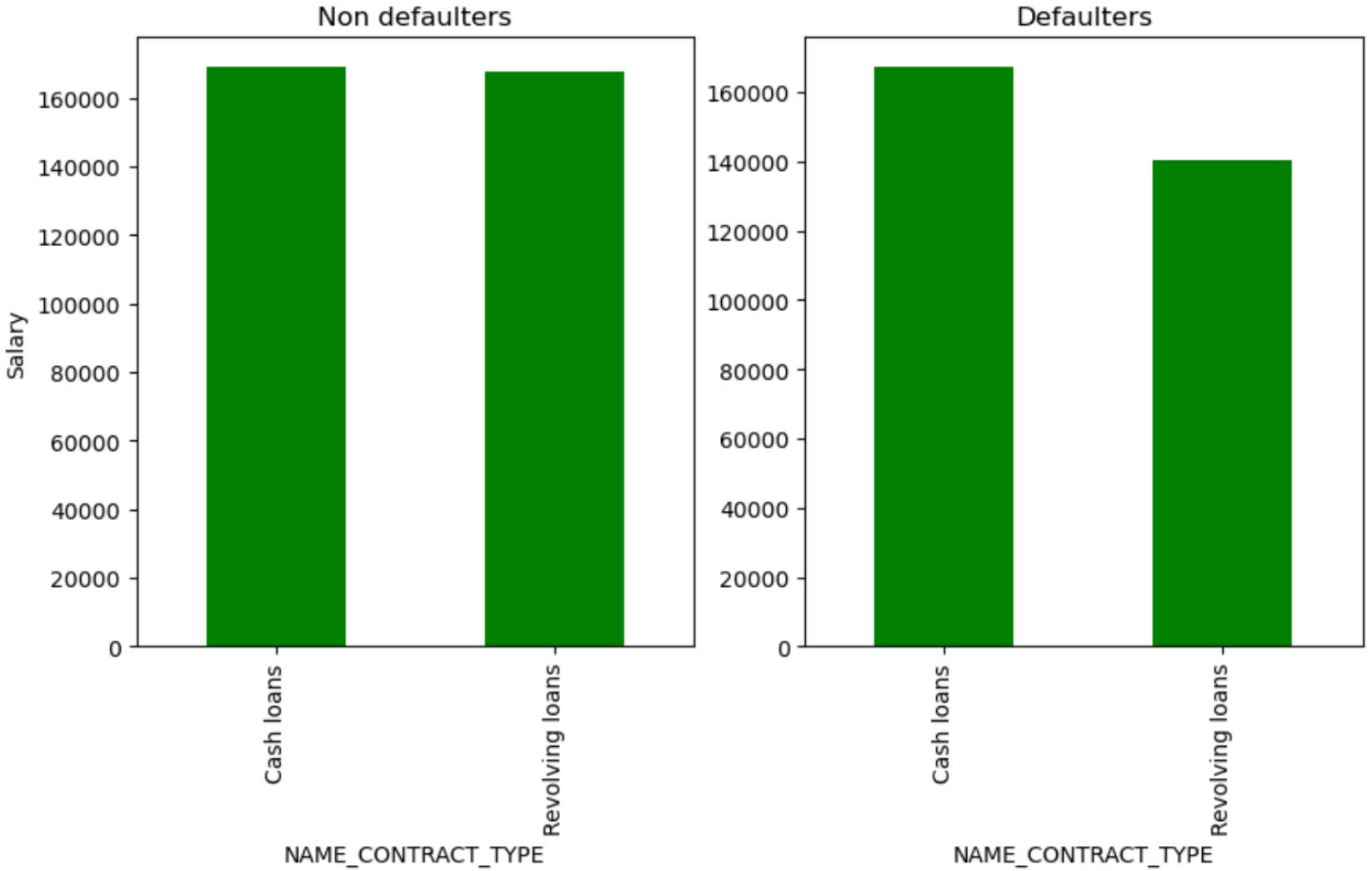
CODE_GENDER Vs AMT_CREDIT





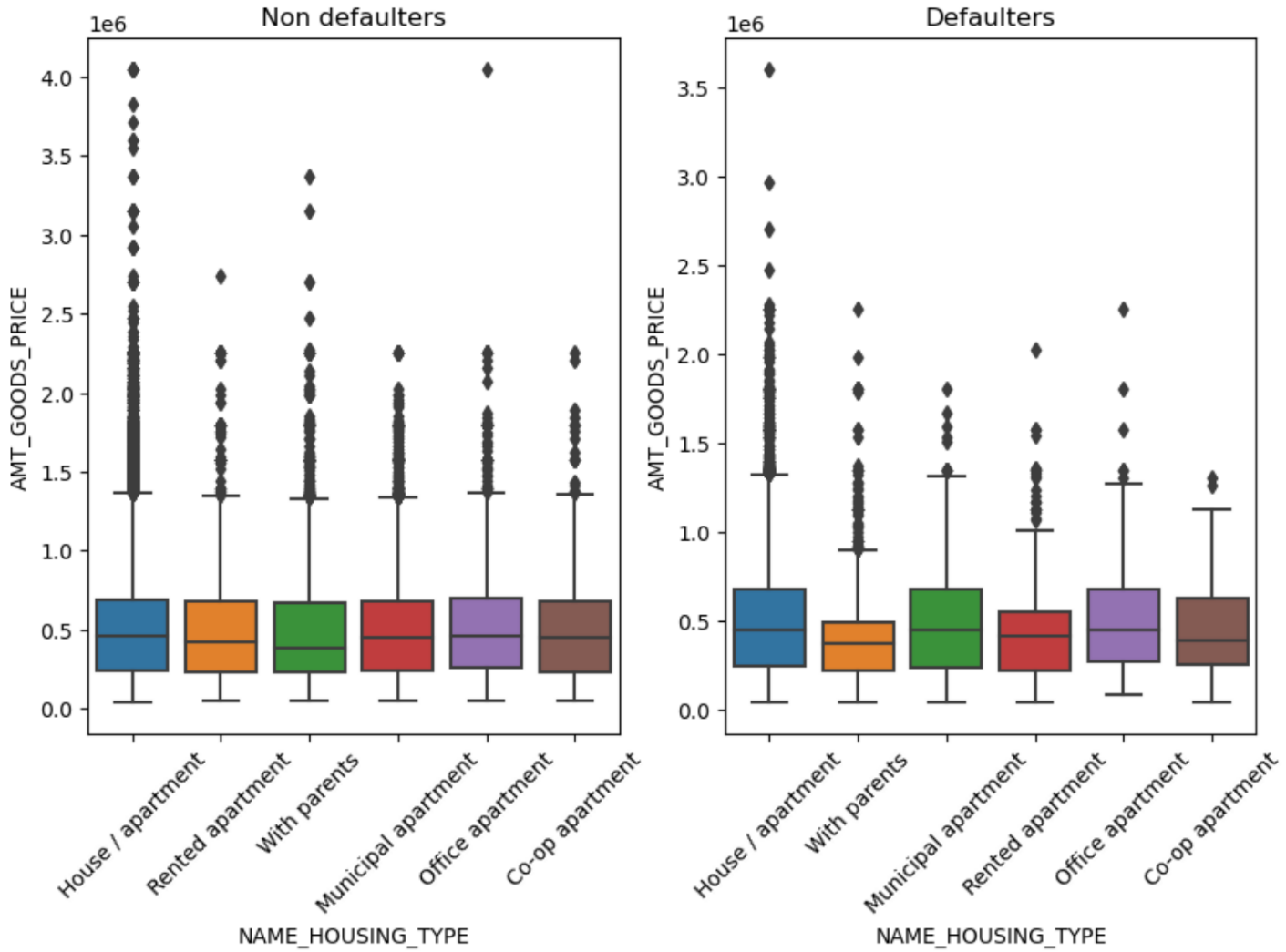
In the Non defaulters plot the credit distribution for both the genders are relatively similar. In the Defaulters plot the credit distribution for males appears to have wider spread than females. Thus there is more variation in the credit amount for male compared to female defaulters.

NAME_CONTRACT_TYPE Vs AMT_INCOME_TOTAL



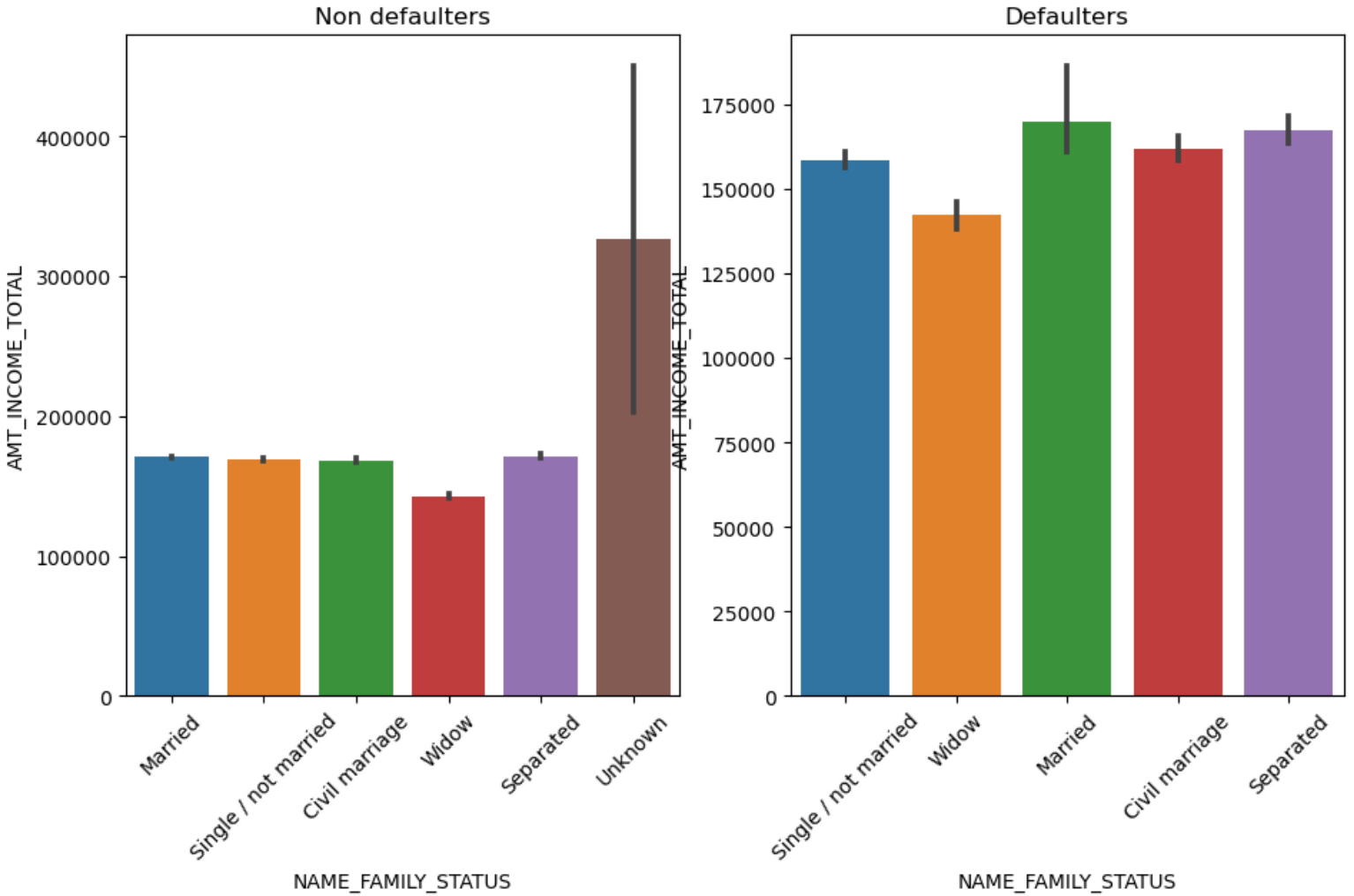
- In summary, the plot suggests that the relationship between contract type and mean income differs between non-defaulters and defaulters.
- Among non-defaulters, those with "Revolving loans" tend to have a higher mean income, while among defaulters, those with "Cash loans" tend to have a higher mean income.

AMT_GOODS_PRICE Vs NAME_HOUSING_TYPE



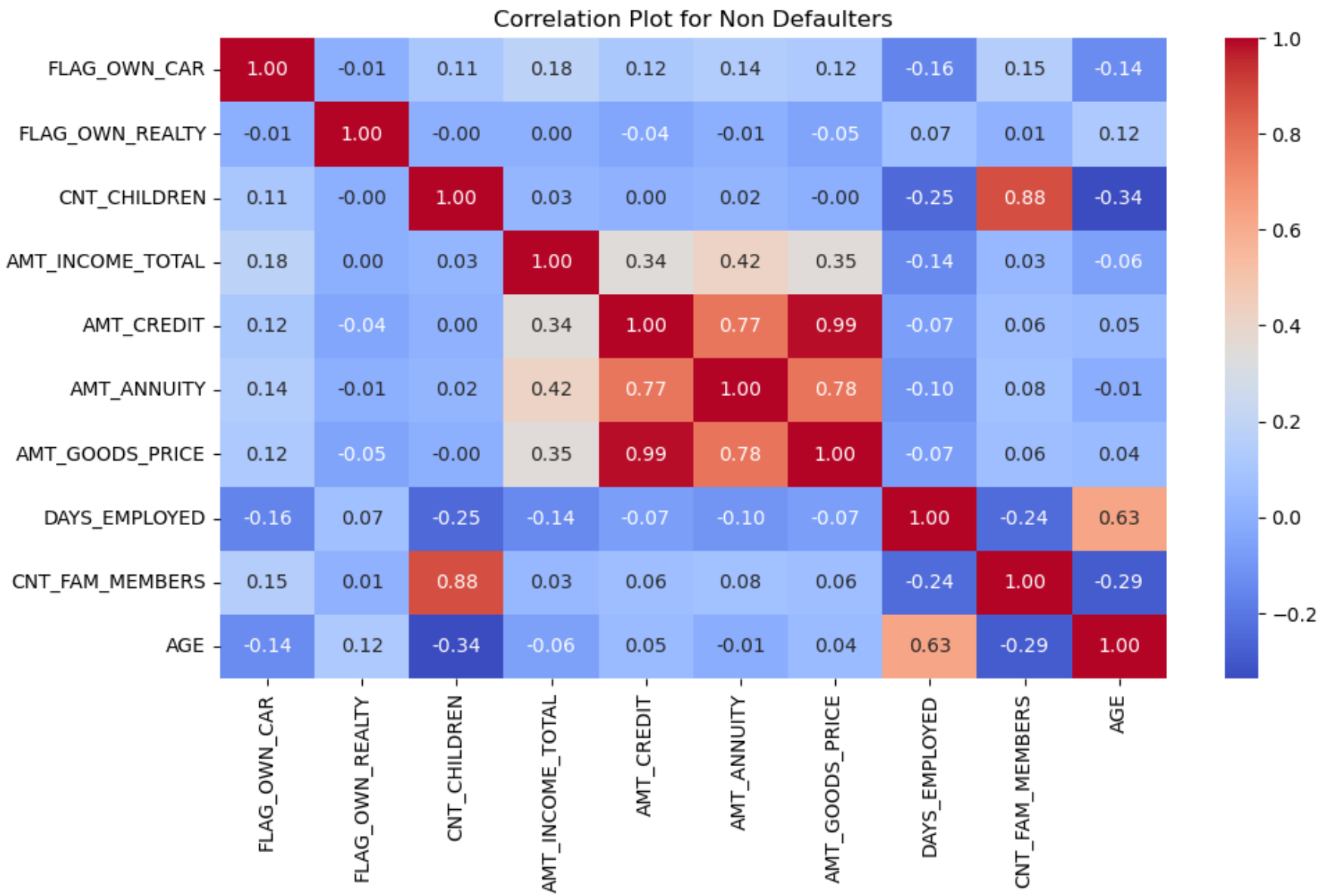
- The persons with House/apartment, Municipal apartment, Office apartment tends to have higher variability than other housing types among defaulters.
- While the non defaulters almost has similar variability of the amount of good owned for all the housing types and does not exhibit much variability in their distribution.

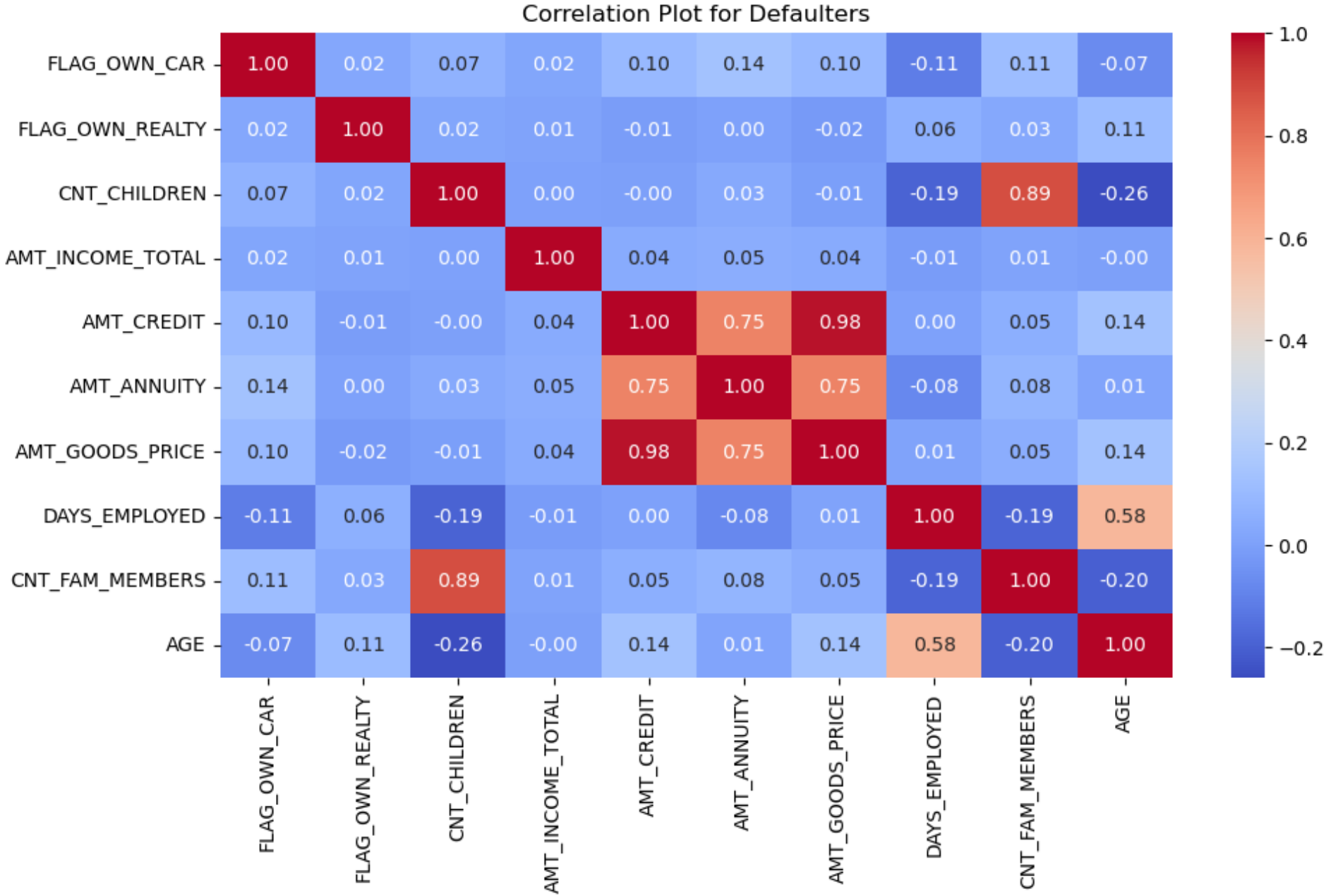
NAME_FAMILY_STATUS Vs AMT_INCOME_TOTAL



1. Civil marriage and Married tend to have higher mean incomes compared to other family statuses.
2. Single/not married and Separated have lower mean incomes than the others for both defaulters and non defaulters.

Correlation Plots

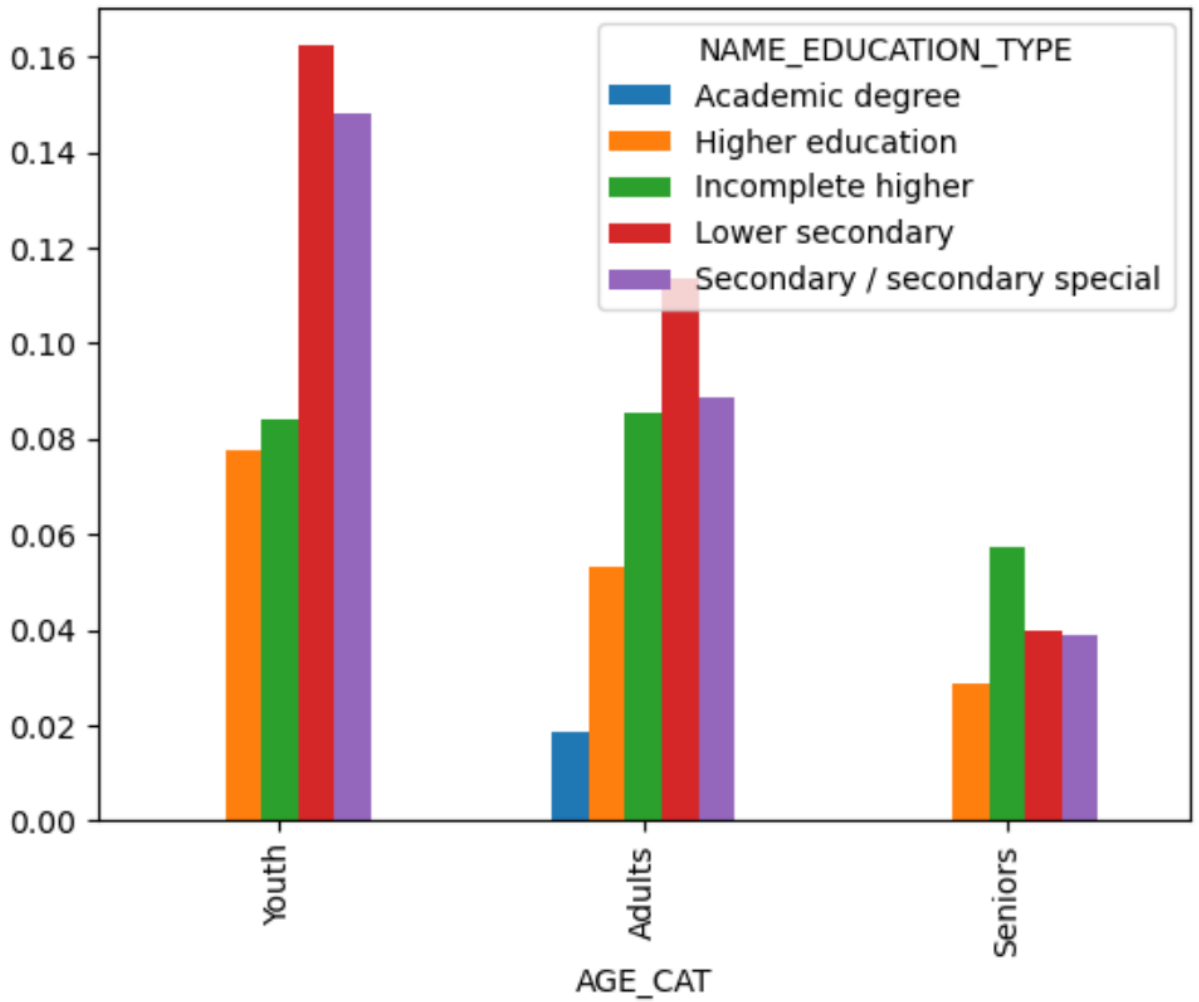



Correlation Plot for Defaulters
FLAG_OWN_CAR 1.00 0.02 0.07 0.02 0.10 0.14 0.10 -0.11 0.11 -0.07
FLAG_OWN_REALTY 0.02 1.00 0.02 0.01 -0.01 0.00 -0.02 0.06 0.03 0.11
CNT_CHILDREN 0.07 0.02 1.00 0.00 -0.00 0.03 -0.01 -0.19 0.89 -0.26
AMT_INCOME_TOTAL 0.02 0.01 0.00 1.00 0.04 0.05 0.04 -0.01 0.01 -0.00
AMT_CREDIT 0.10 -0.01 -0.00 0.04 1.00 0.75 0.98 0.00 0.05 0.14
AMT_ANNUITY 0.14 0.00 0.03 0.05 0.75 1.00 0.75 -0.08 0.08 0.01
AMT_GOODS_PRICE 0.10 -0.02 -0.01 0.04 0.98 0.75 1.00 0.01 0.05 0.14
DAYS_EMPLOYED -0.11 0.06 -0.19 -0.01 0.00 -0.08 0.01 1.00 -0.19 0.58
CNT_FAM_MEMBERS 0.11 0.03 0.89 0.01 0.05 0.08 0.05 -0.19 1.00 -0.20
AGE -0.07 0.11 -0.26 -0.00 0.14 0.01 0.14 0.58 -0.20 1.00
FLAG_OWN_CAR
FLAG_OWN_REALTY
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
DAYS_EMPLOYED
CNT_FAM_MEMBERS
AGE
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Insights from heat maps Defaulters and Non defaulters

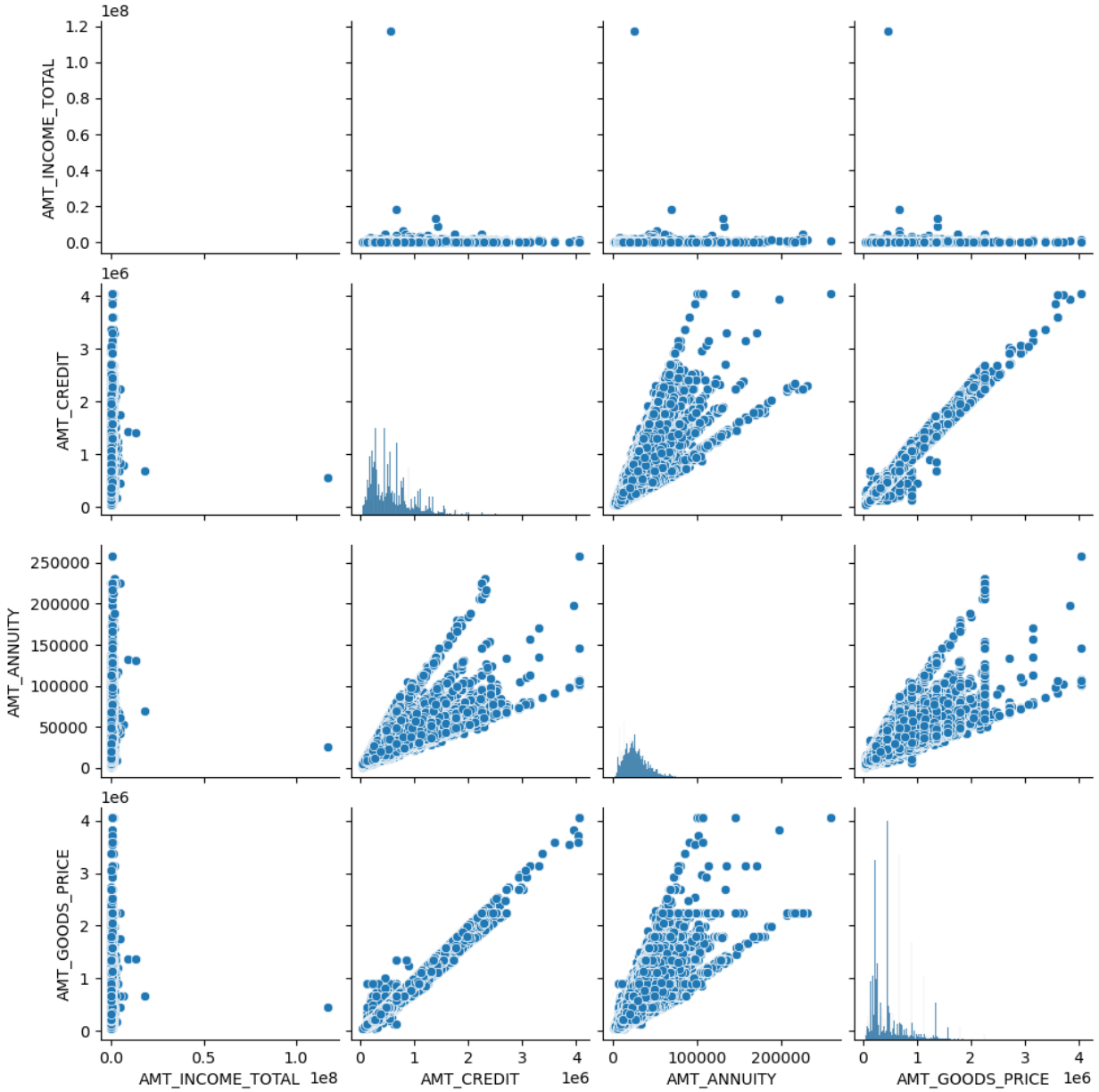
1. AMT_CREDIT and AMT_GOODS_PRICE has strong positive correlation which implies that the loan amount is increased when the price of goods/asset increases for both Defaulters and Non defaulters
2. CNT_CHILDREN and CNT_FAM_MEMBERS are positively correlated with each other. When the count of children in the family increases , the family members count also gets increased for both Defaulters and Non defaulters.
3. AMT_CREDIT and AMT_ANNUITY are directly proportional to each other. The total annual loan payment increases when the credit amount is high.
4. AMT_GOODS_PRICE, AMT_CREDIT are negatively correlated with CNT_CHILDREN which shows that the loan amount is higher for those who has fewer children.
5. AMT_ANNUITY are inversely proportional to DAYS_EMPLOYED which shows that the annual loan amount to be paid is higher for people employed for less number of days.

Analysing loan defaulters by Education Type and Age Category



1. The plot shows that the youth who completed lower secondary education exposes high risk of loan default.
2. The Senior age group people who are highly educated shows minimal risk of repayment.

Pairplots of continuous variables



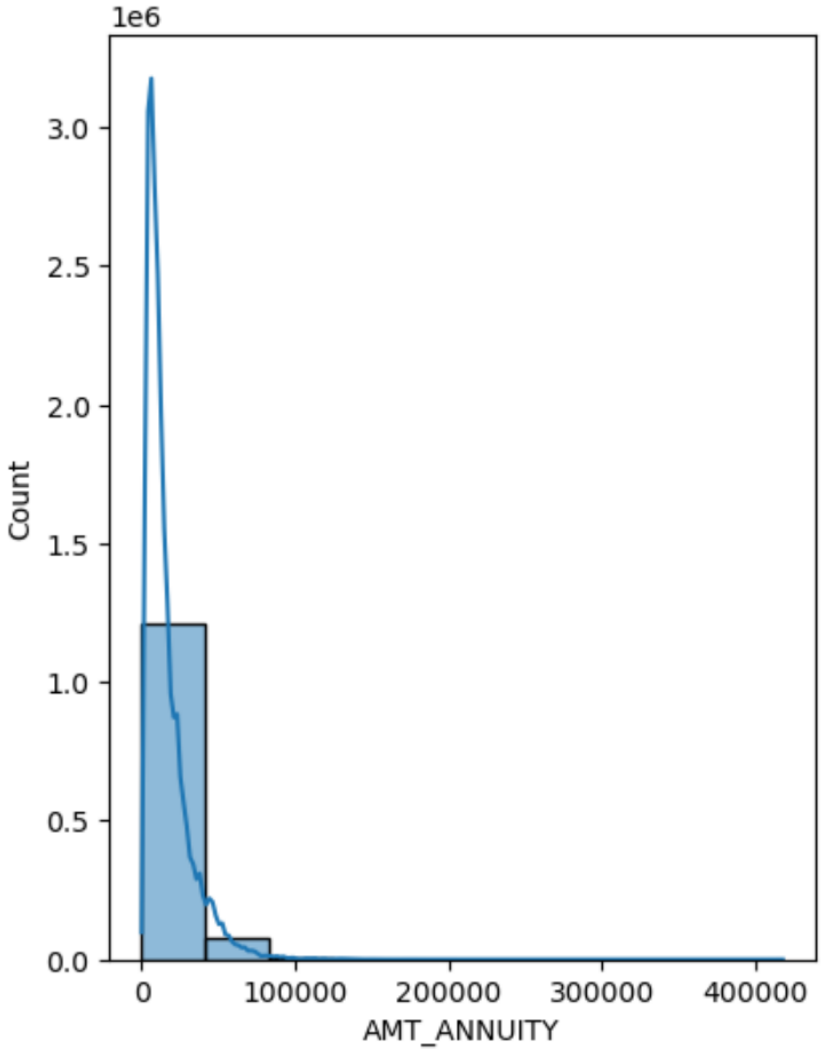
Inference

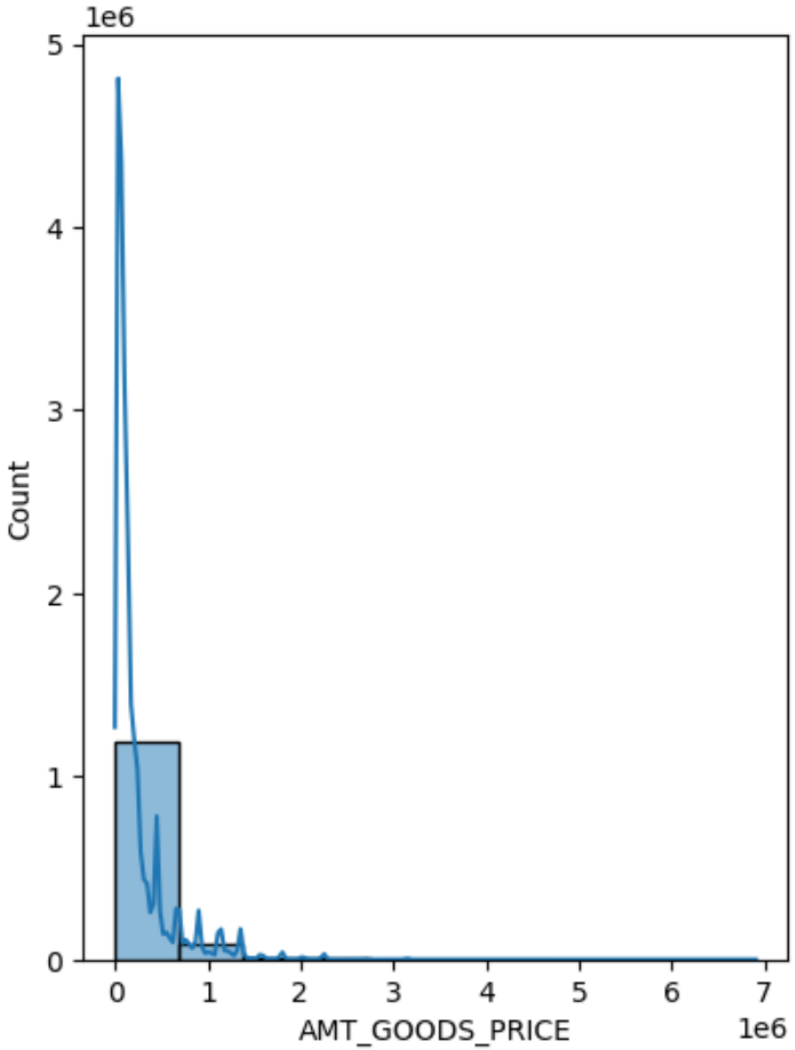
1.AMT_CREDIT and AMT_GOODS_PRICE shows strong correlated with each other as an increase in the credit amount is typically associated with a higher goods price.

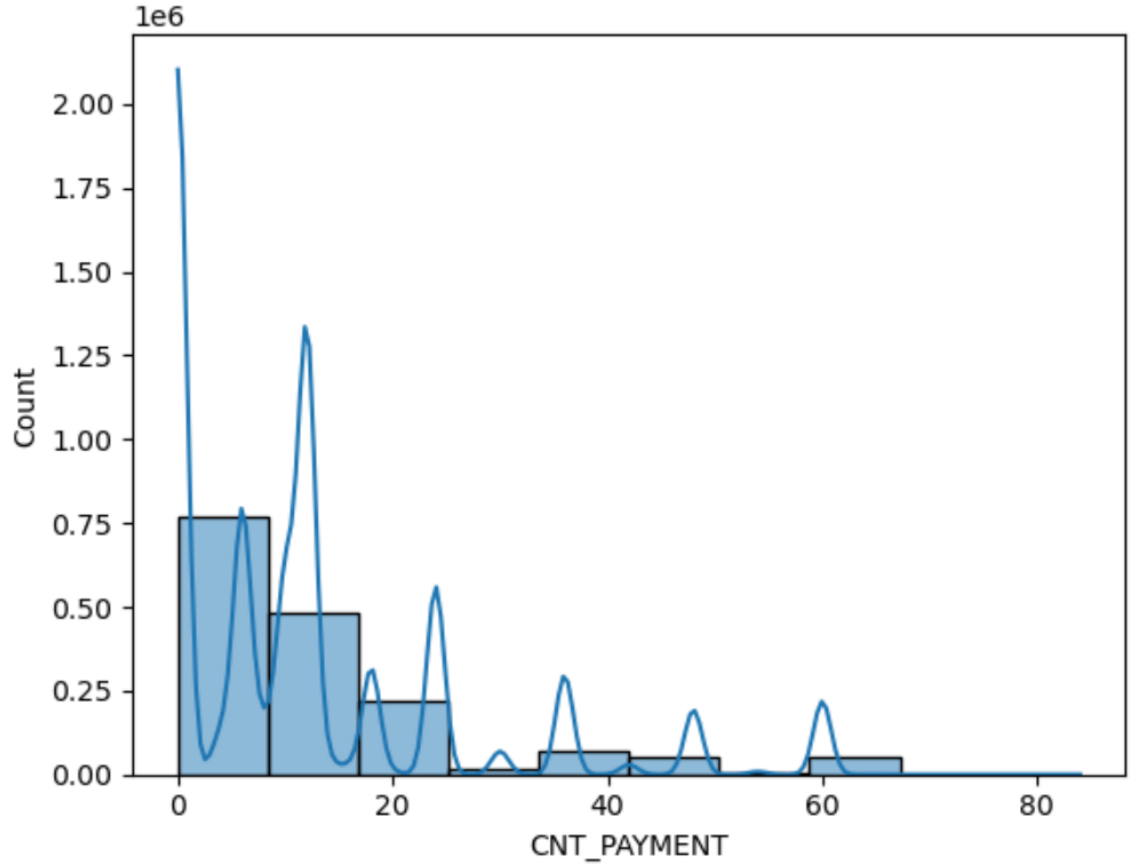
2.AMT_ANNUITY and AMT_CREDIT may reveal a positive linear relationship, where higher credit amounts are associated with higher annuities.

Previous_application

Handling missing values columns whose missing value is < 13% are imputed

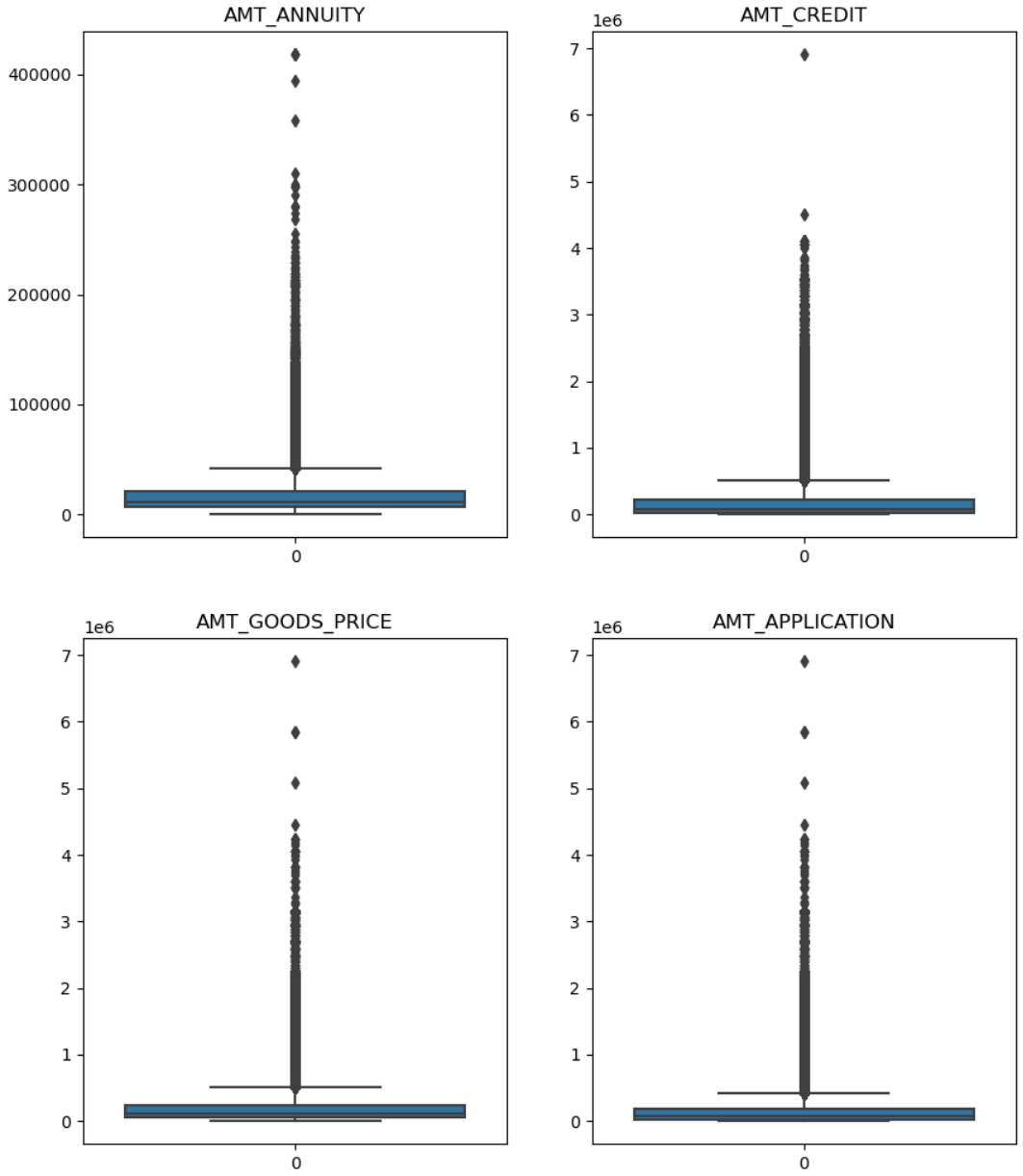






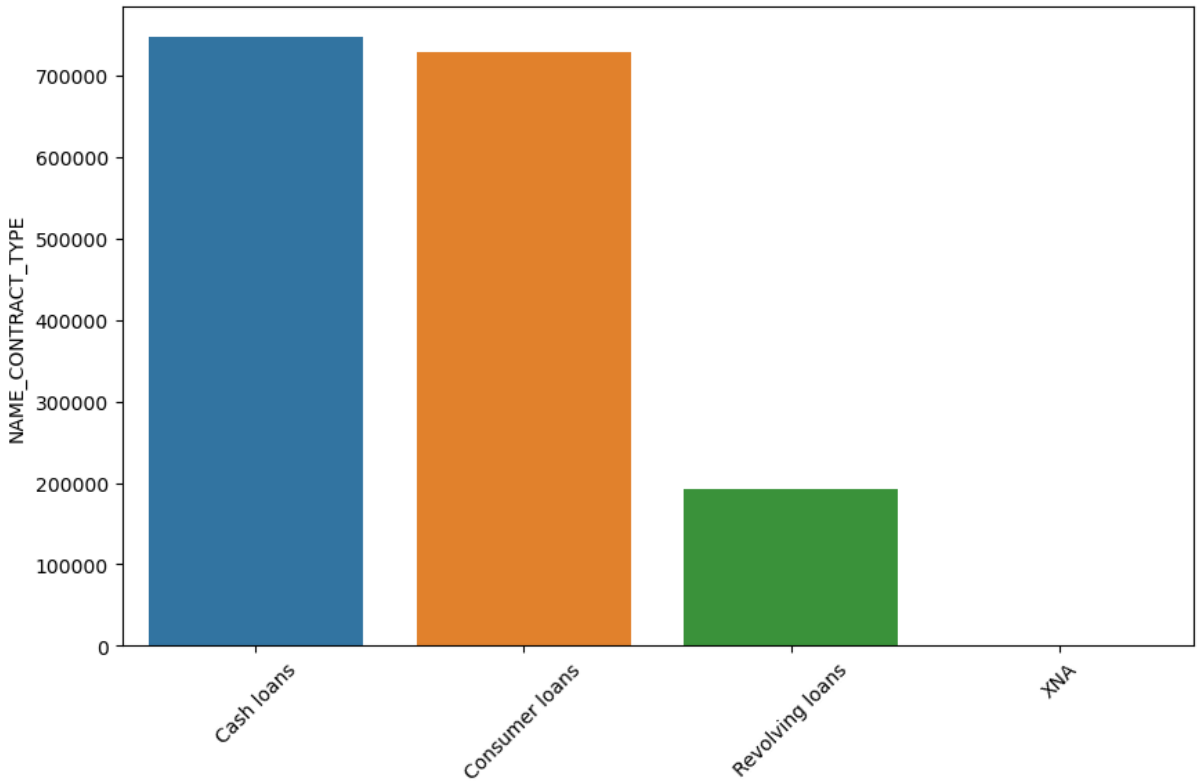
The distribution plot of the variable AMT_ANNUITY, AMT_GOODS_PRICE and CNT_PAYMENT is rightly skewed and hence the null values can be replaced with median value.

Handling outliers

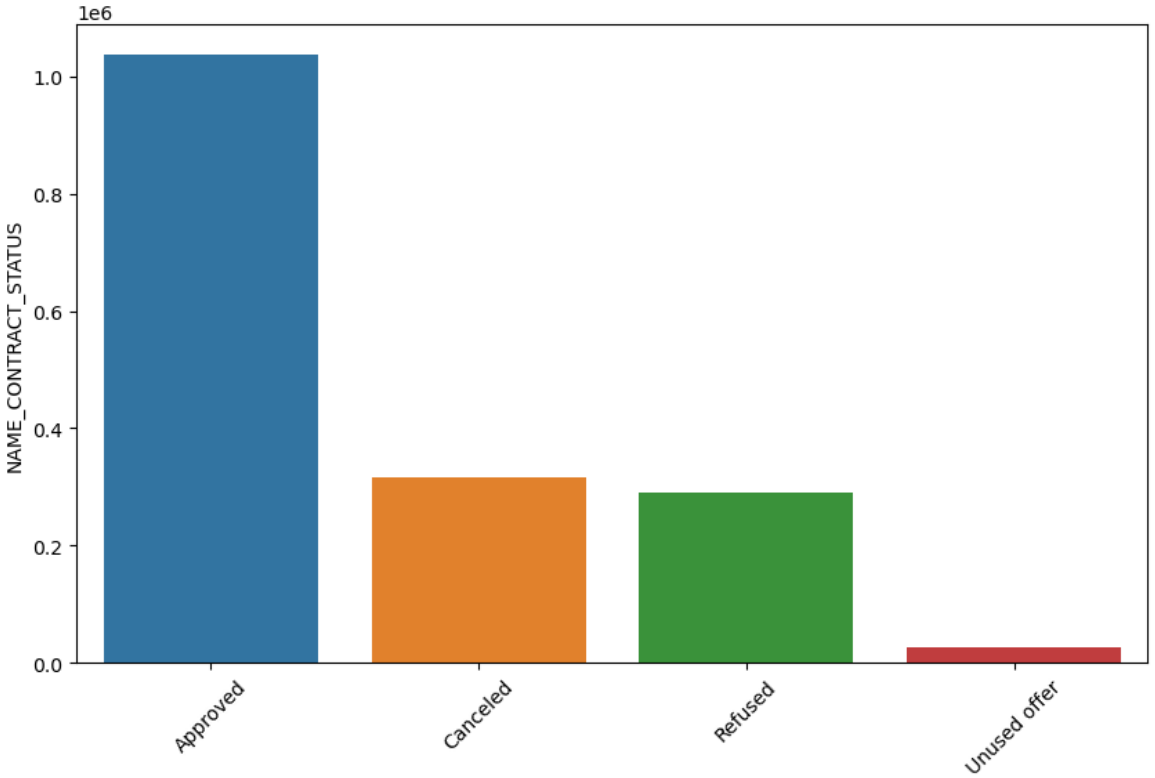


Since the percentage of outliers are very less compared to the total data, the outliers in all the variables can be dropped off, since it does not have an effect on the overall analysis.

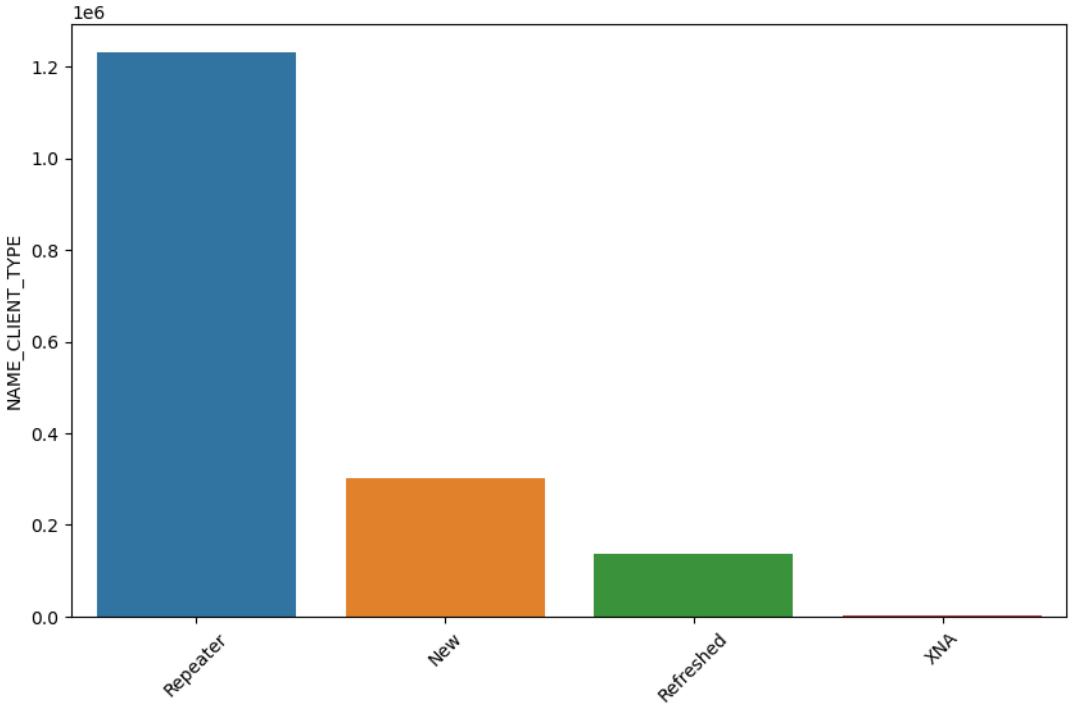
Categorical variables



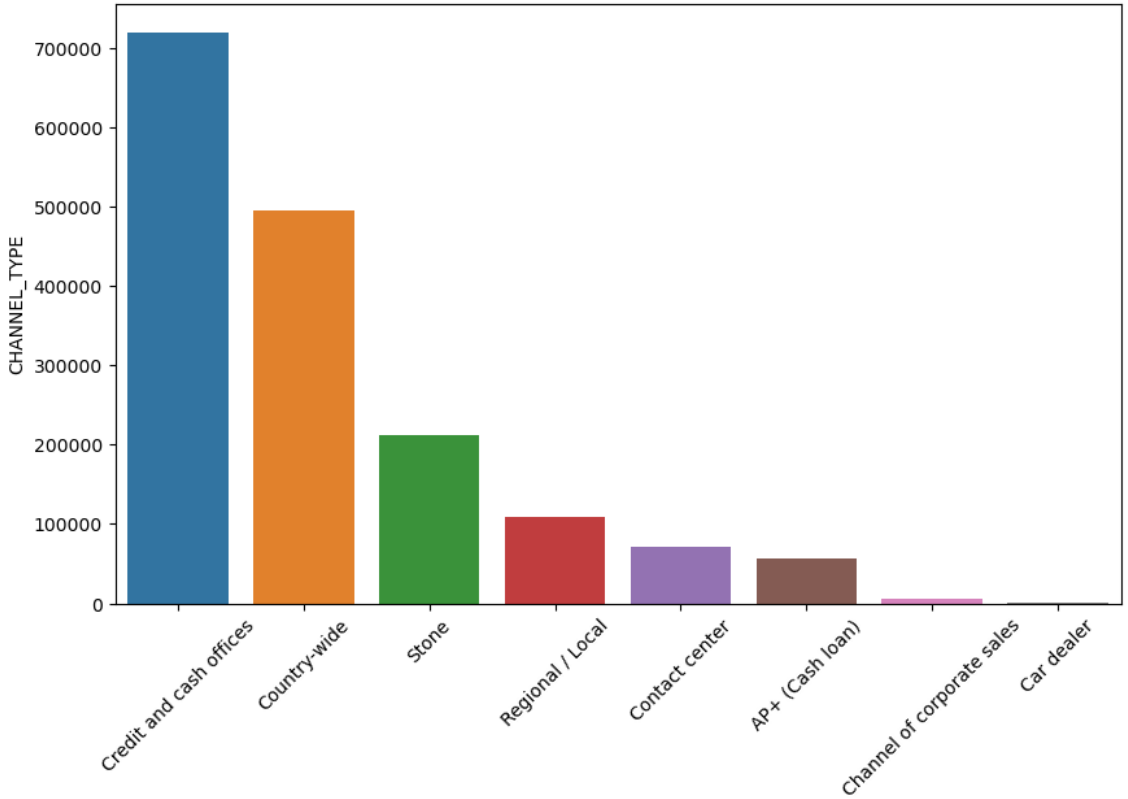
NAME_CONTRACT_TYPE



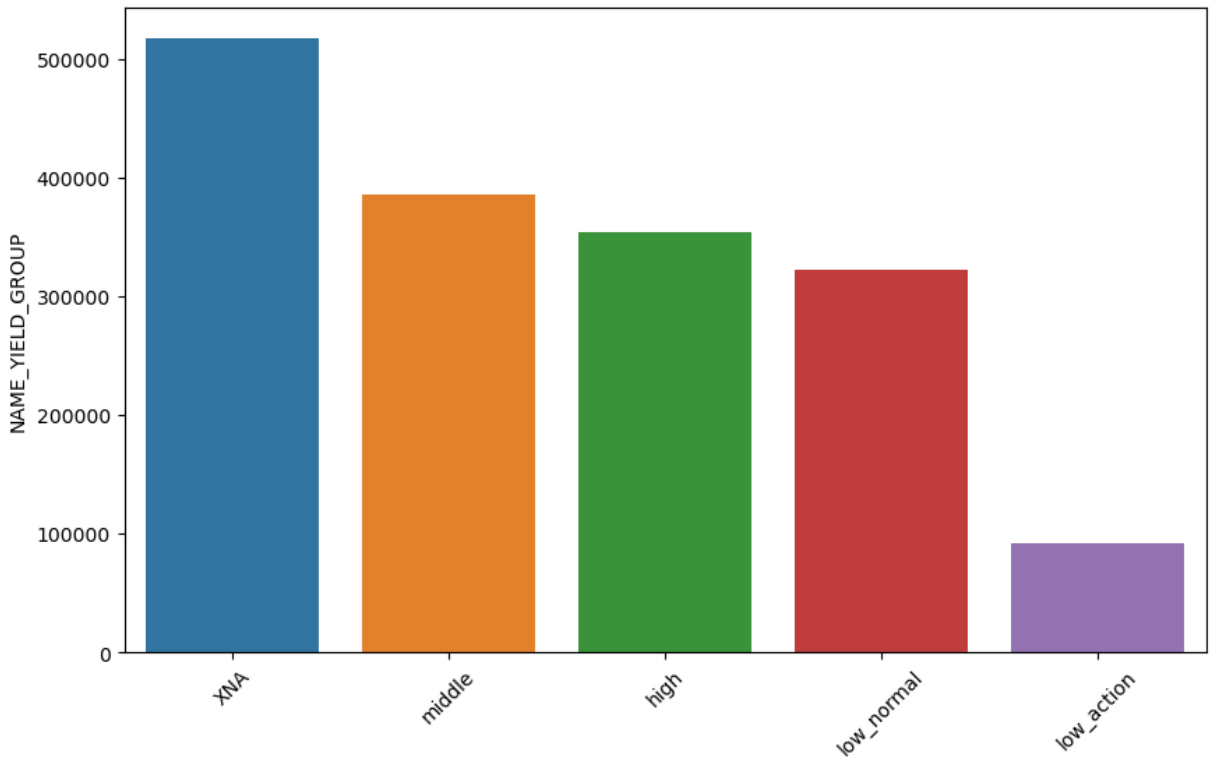
NAME_CONTRACT_STATUS



NAME_CLIENT_TYPE



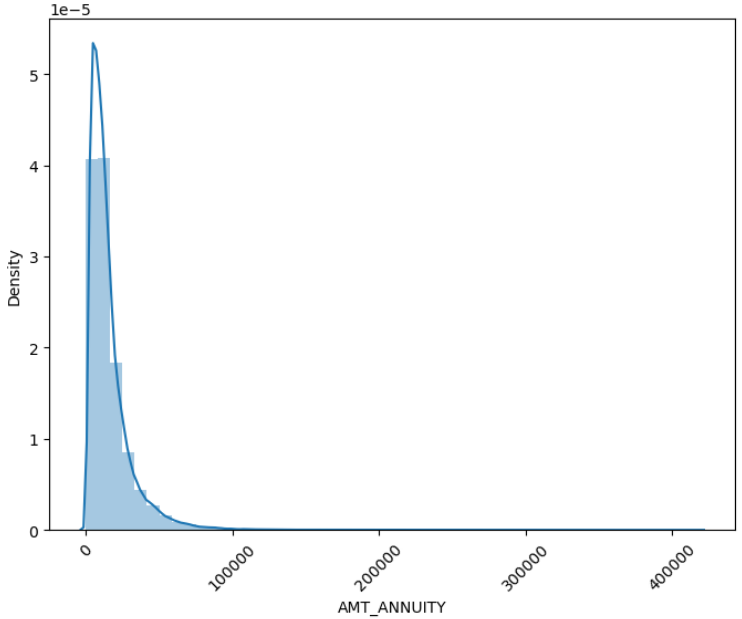
CHANNEL_TYPE

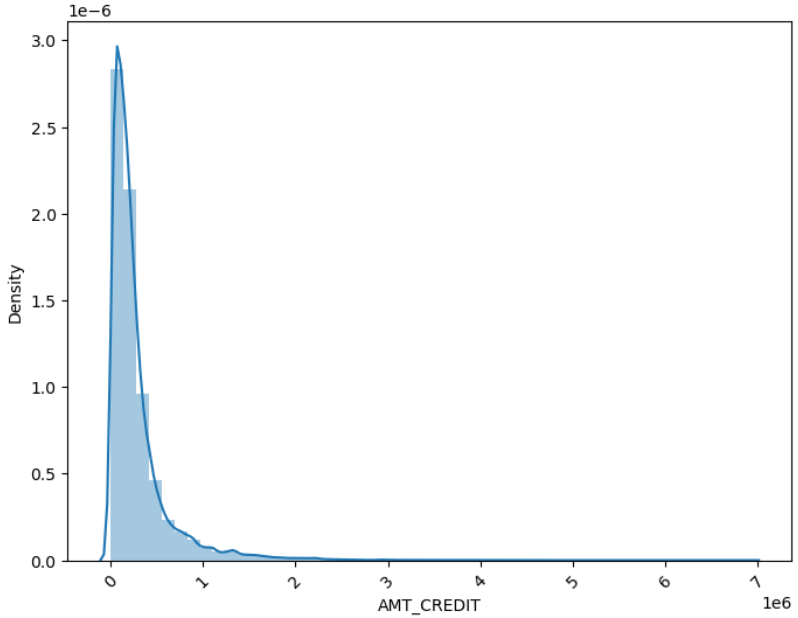


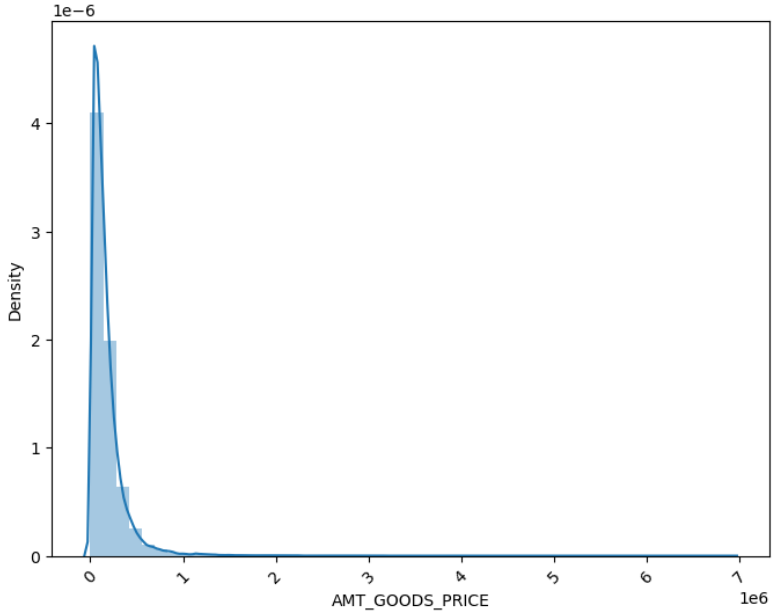
Insights from the Categorical variables

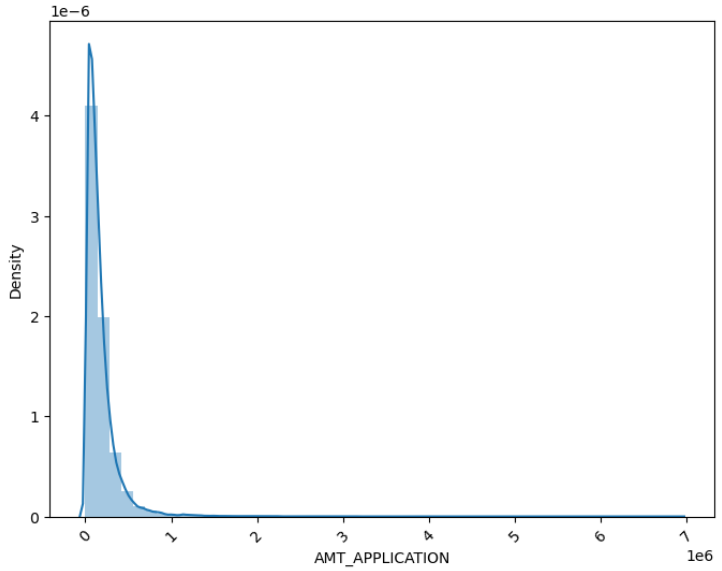
1. NAME_CONTRACT_TYPE: Cash loans are higher followed by Consumer loans
2. NAME_CONTRACT_STATUS: Most of the previous loan application status are approved and only a meagre amount of applicant did not use the offer.
3. NAME_CLIENT_TYPE: Most of the clients are already a customer of the bank
4. CHANNEL_TYPE: The clients approached by the Credit and cash offices are higher in proportions.
5. NAME_YIELD_GROUP: The interest were in the middle category for most of the loan applicants.

Numerical variables



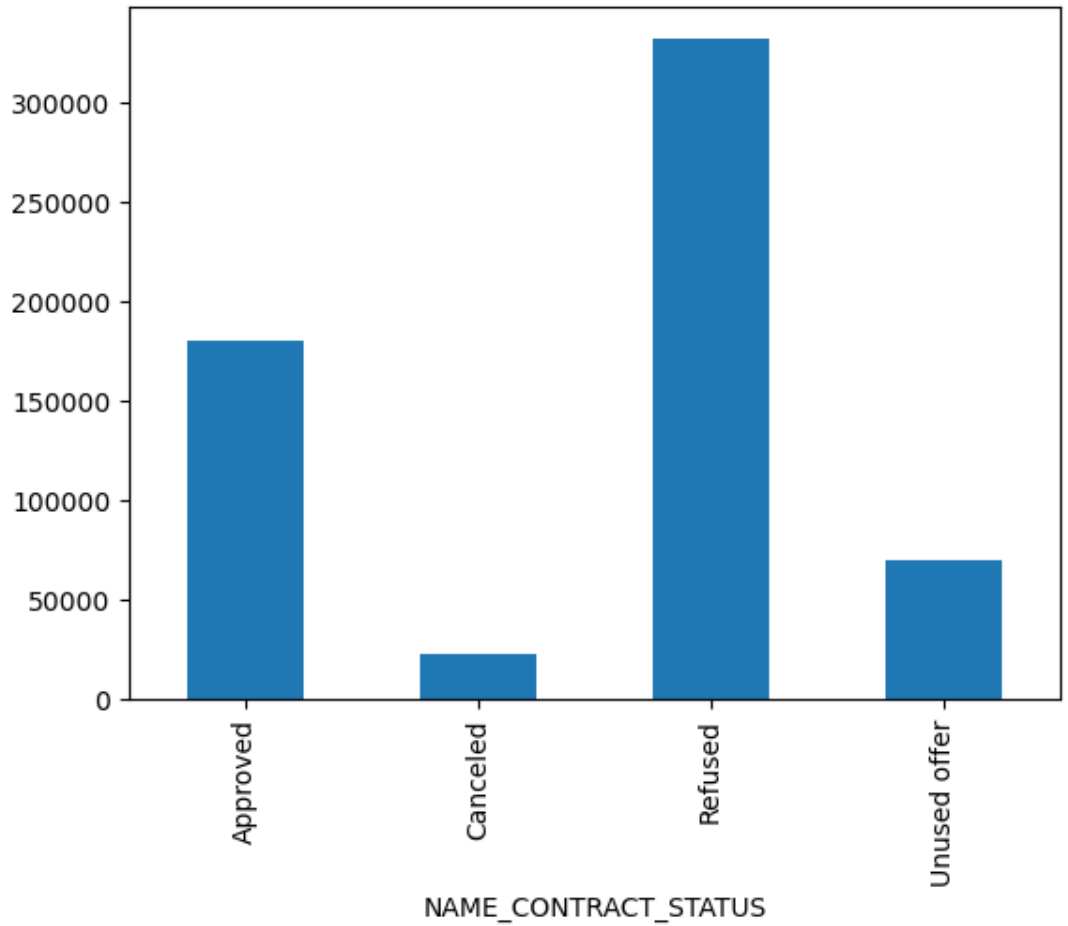






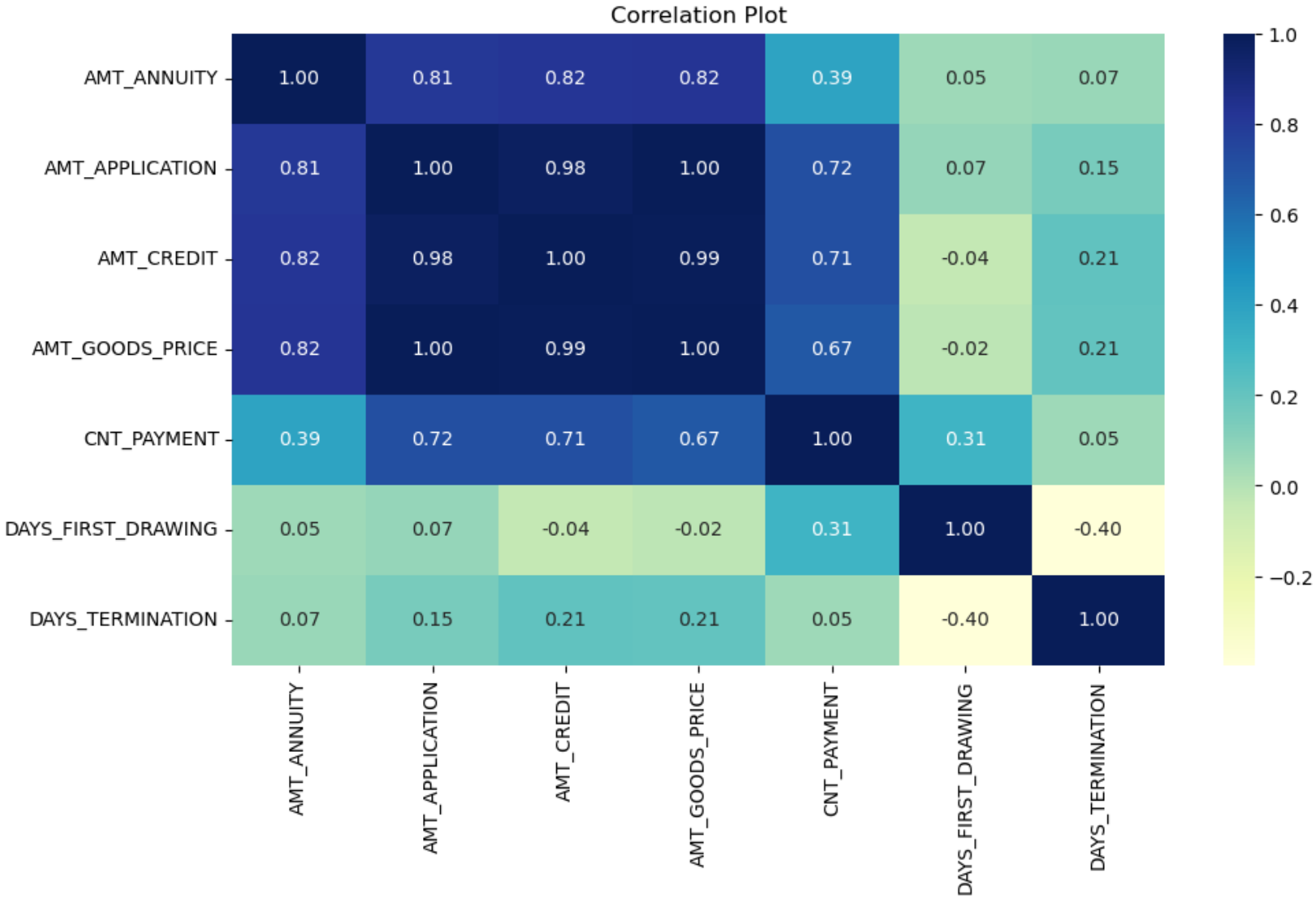
Bivariate analysis

Average Application amount by Contract Status



1. The plot shows that most of the clients were refused who had a very high application amount.
2. The clients who opted for average amount of application at a moderate range were accepted and those with lower amount were cancelled.

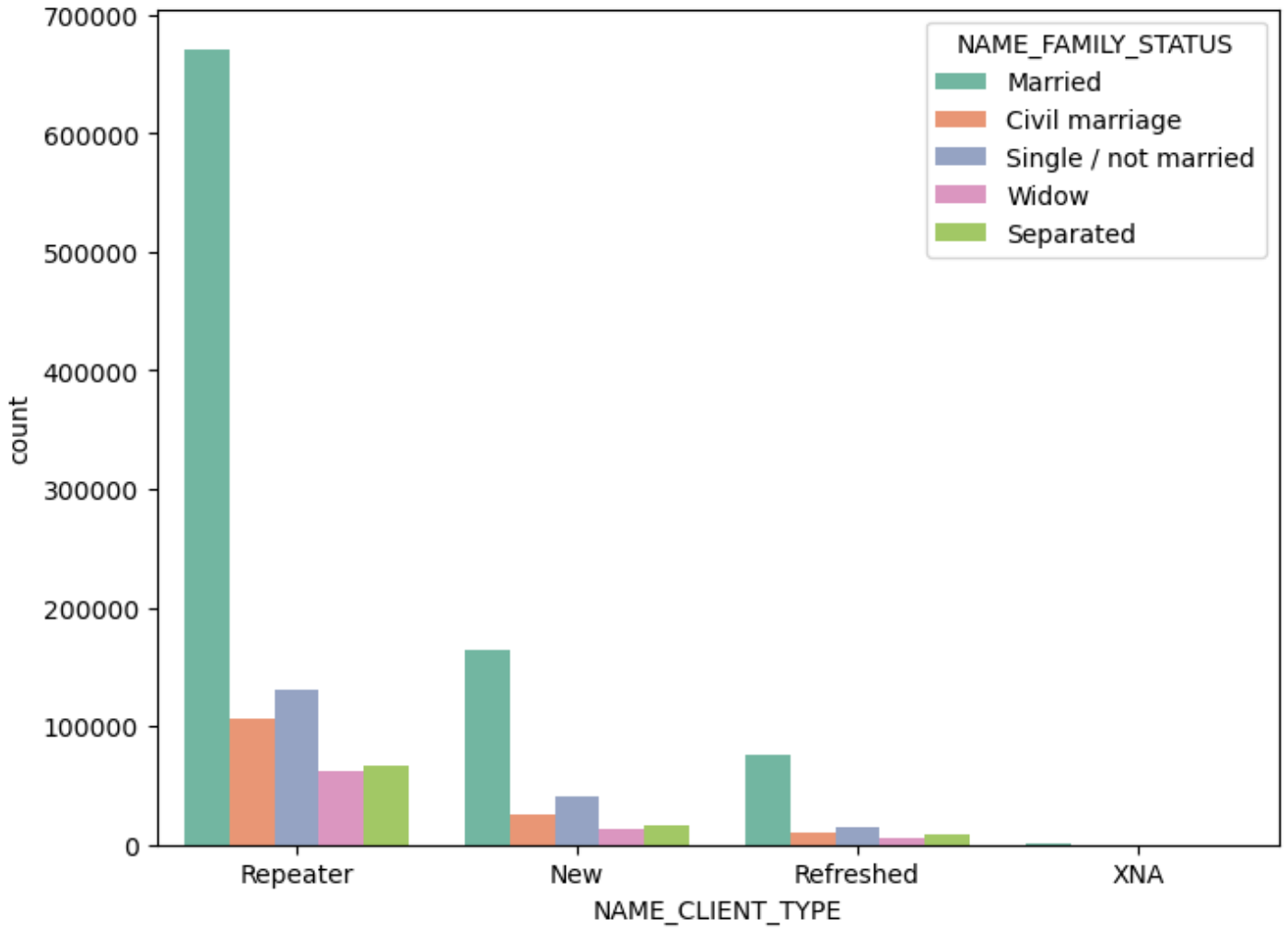
Correlation plots for numerical variables



- The variable AMT_ANNUITY,AMT_APPLICATION,AMT_CREDIT and AMT_GOODS_PRICE shows strong positive correlation with each other.
- While DAYS_FIRST_DRAWING and DAYS_TERMINATION exhibit moderate correlation with other variables.

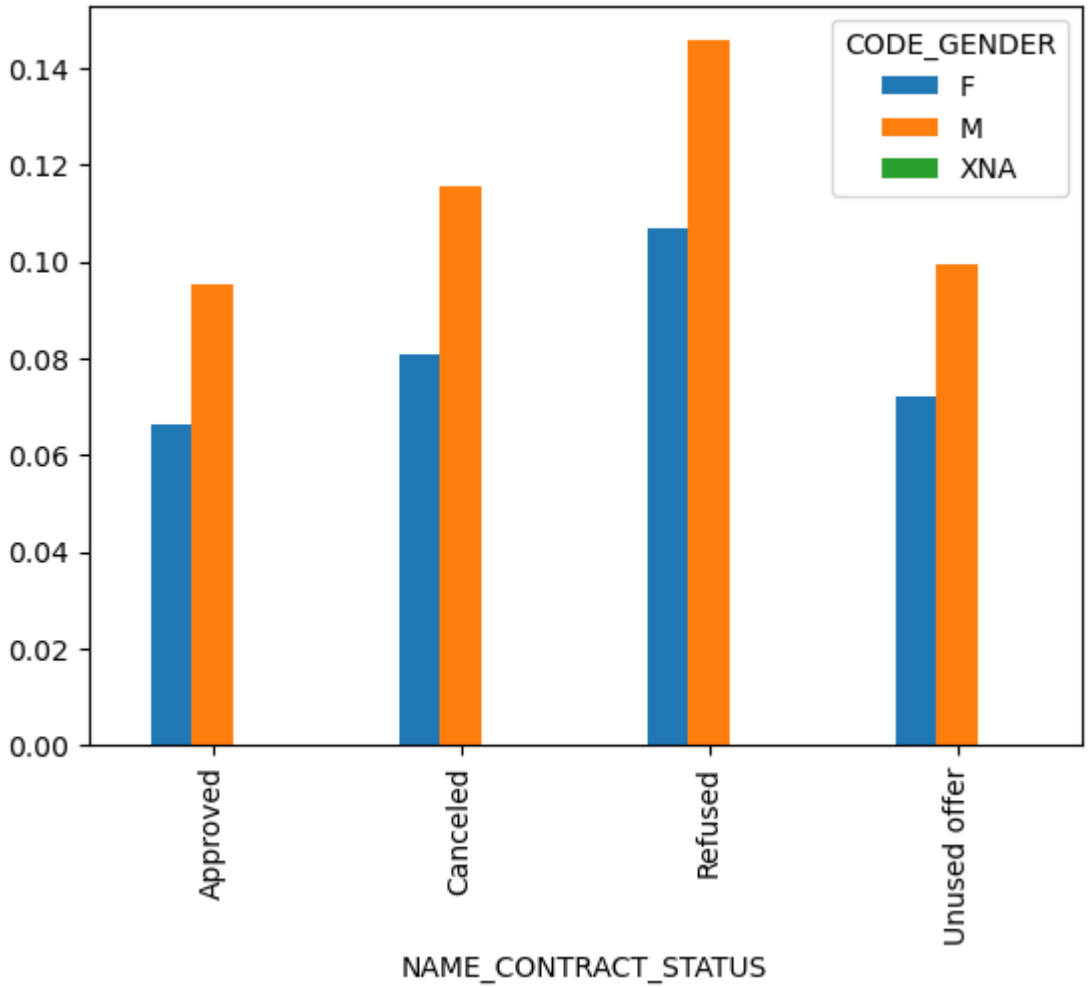
Merged Data analysis

Types of Client based on Family status



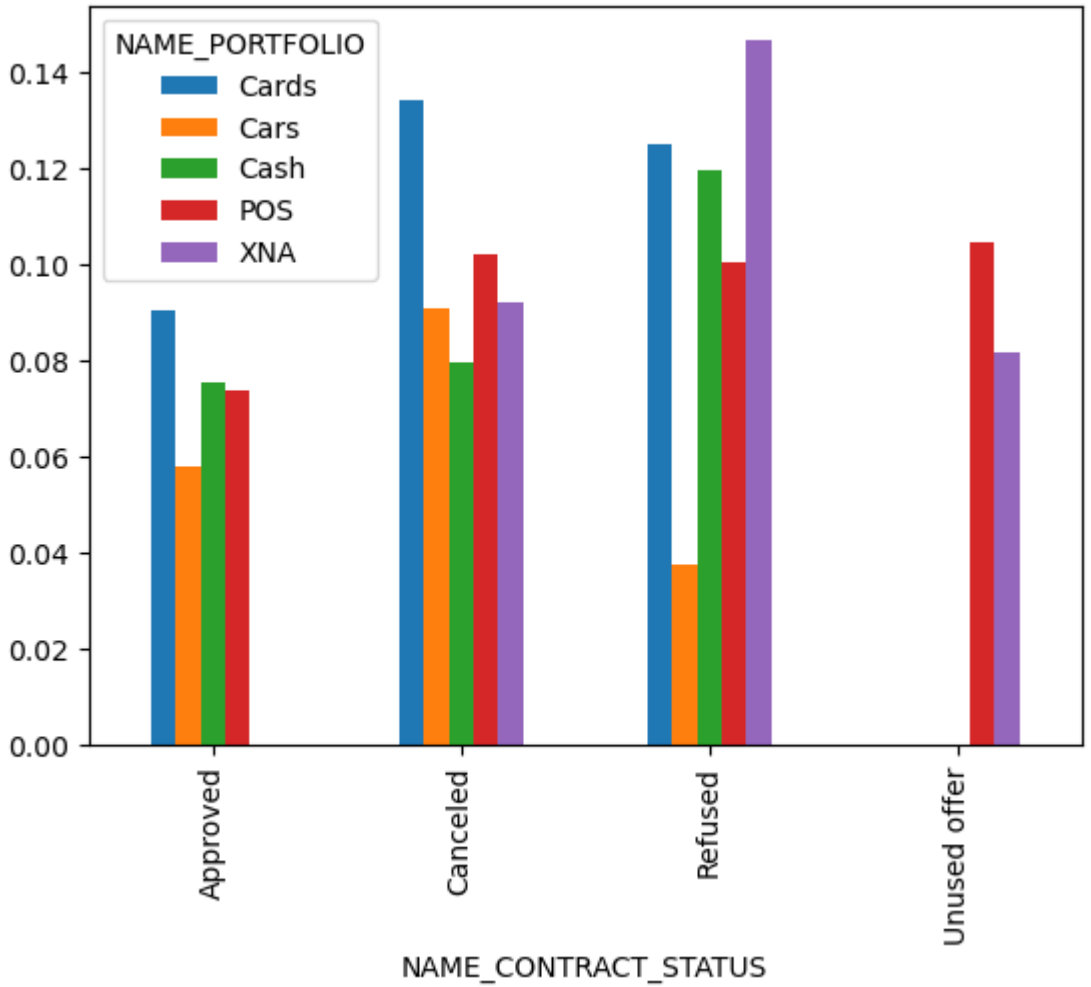
The count plot of the type of clients based on their family status shows that most of the clients approved for credit were repeated clients and married.

Current loan defaulters based on Previous application status wrt gender



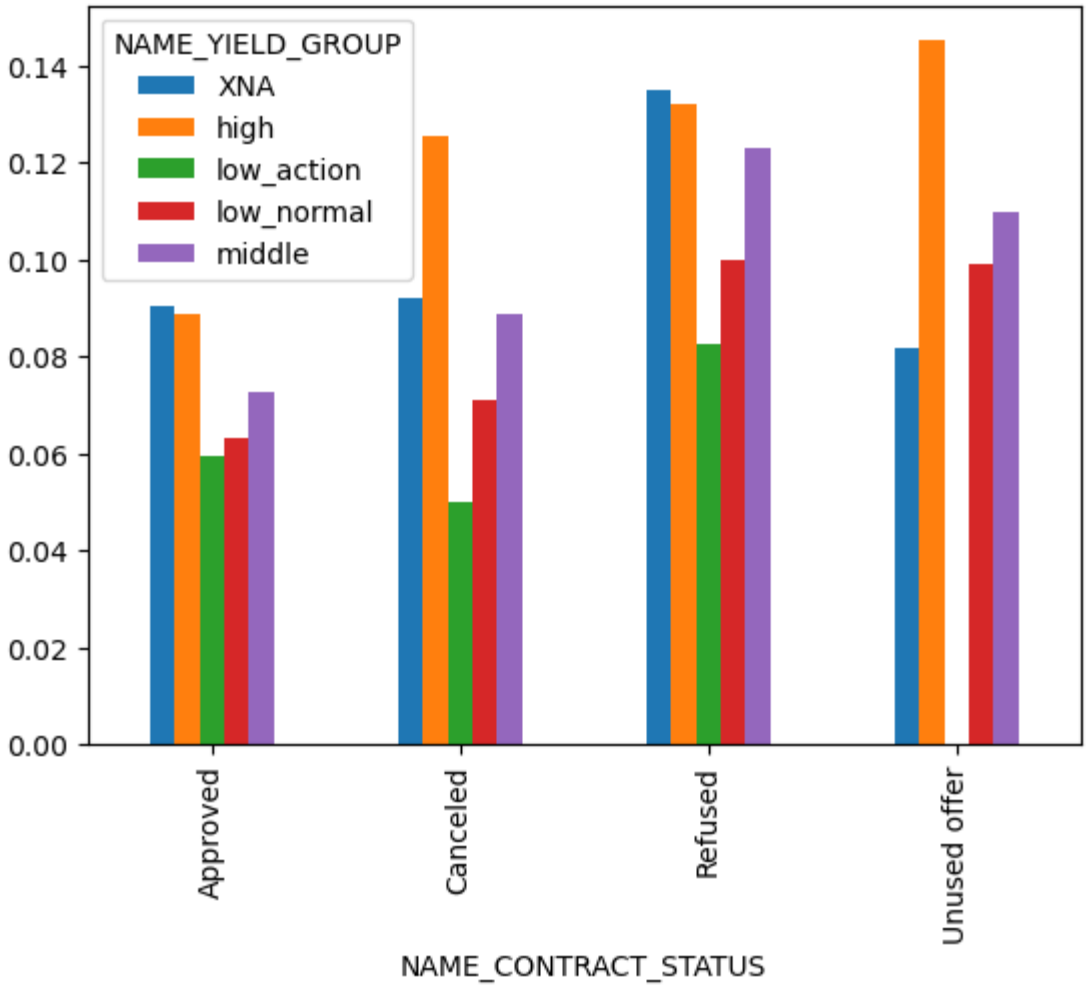
The plot shows that males are higher in proportion than female defaulters and loan to Previously refused clients has high risk of loan default.

Current loan defaulters based on Previous application status wrt Name portfolio



1.The loan defaulters are less for those approved for cars from previous application and exhibits high risk for cards portfolio.
2.For those previously refused applicants, the default loan status is less for cars portfolio.
3.For Previously cancelled applicants, cash exhibits low risk of loan default.

Current loan defaulters based on Previous application status wrt Interest category



- There is a risk of loan default for those having high grouped interest amount.
- In previously approved application, low-action interest group has lesser loan default risk. The same pattern goes for both cancelled and refused loan applicants.

Key Insights from the analysis

Targets based on Categories

- **OCCUPATION_TYPE** - Labourers
- **CODE_GENDER** - Females
- **AGE_BY_DECADE** – 30s age category
- **NAME_CONTRACT_TYPE** – Cash loans
- **HAS_CHILDREN** – Does not have children
- **NAME_EDUCATION_TYPE** – Secondary/ Secondary special
- **NAME_INCOME_TYPE** – Working professionals
- **NAME_FAMILY_STATUS** – Married
- **AMT_INCOME_TOTAL** – lower to middle range income

- Most of the clients approved for credit were repeated clients and married.
- Loan to Previously refused clients has high risk of loan default.
- The loan defaulters are less for those approved for cars from previous application
- Low-action interest group has lesser loan default risk

THANK YOU