EDA Credit Assignment

- Ankita Patel

Datasets

- Application Data
- Previous Application

Steps before the Analysis

- Import required libraries
- Import both the datasets
- Understand shape, info and description and get the clarity of the datasets
- Cleaning of Data – Identify the null values
- Drop empty columns with more than 40% of missing values
- To understand certain outliers plot the Boxplot for AMT_Annuity & CNT_FAM_MEMBERS and replace null values with the median for further analysis
- Dropped columns which are of less significance

- Cntd….

Steps before the Analysis

- Convert days columns to no of days and then in the years
- Identify XNA in gender and moving values to Female
- Likewise at all the important columns replace XNA at the identified places to required information i.e. XNA with Retired as an organization against the income type
- Create bins for income amount and credit amount

Analysis stages

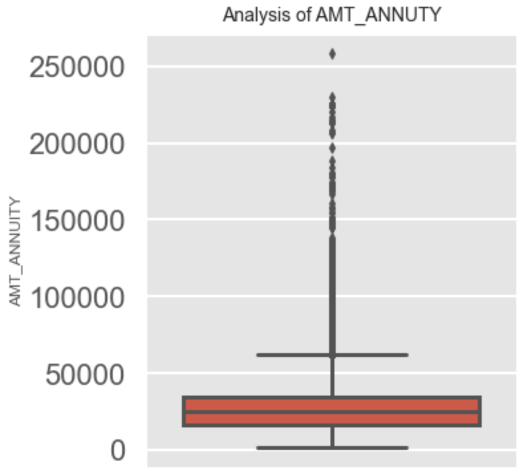
- Divide dataset in to two datasets of target 1 and target 0 respectively representing Defaulters and non Defaulters
- Calculate the Imbalance percentage
- Subsequently carry out the univariate and bivariate analysis
- Identify correlation matrix for both target 1 and target 0
- Once done with Application data pick up the Previous Application and cleanse it and after that merge both Application Data and Previous application
- Analysis of Numerical vs Numerical and Numerical vs categorical data

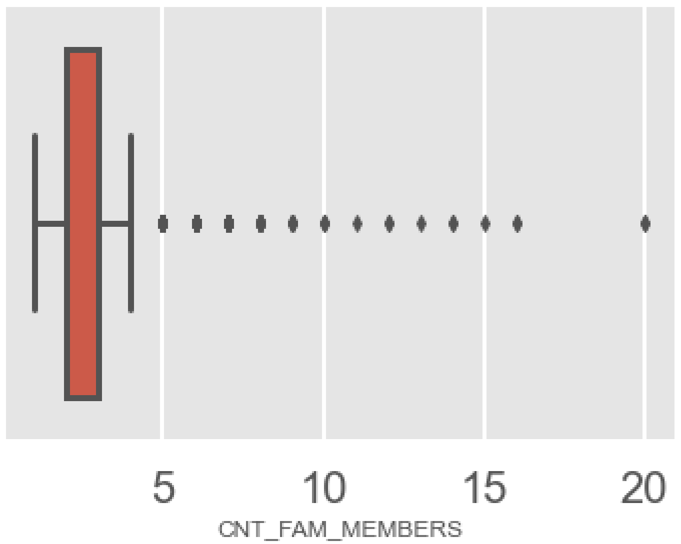
Libraries used

- #import the libraries
- import pandas as pd
- import numpy as np
- import matplotlib.pyplot as plt #Data Visualization Libraries
- %matplotlib inline
- import seaborn as sns #Data Visualization Libraries
- import warnings
- warnings.filterwarnings('ignore')
- pd.set_option("display.max_rows",1000)

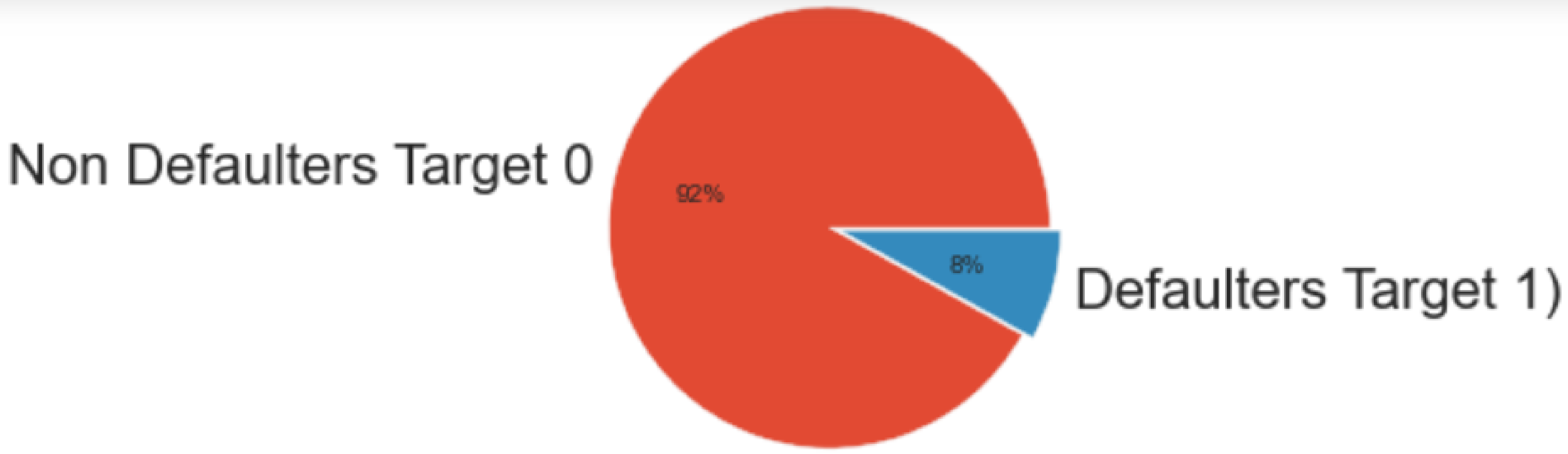
Basic understanding of data

- Through shape, describe, info
- also identified missing values and treated it as needed
- Removed columns with more than 40% missing values
- In remaining columns where ever there were missing values – post preparing a boxplot imputed it with the mean/ median values





Calculation on imbalance for target 0 and 1



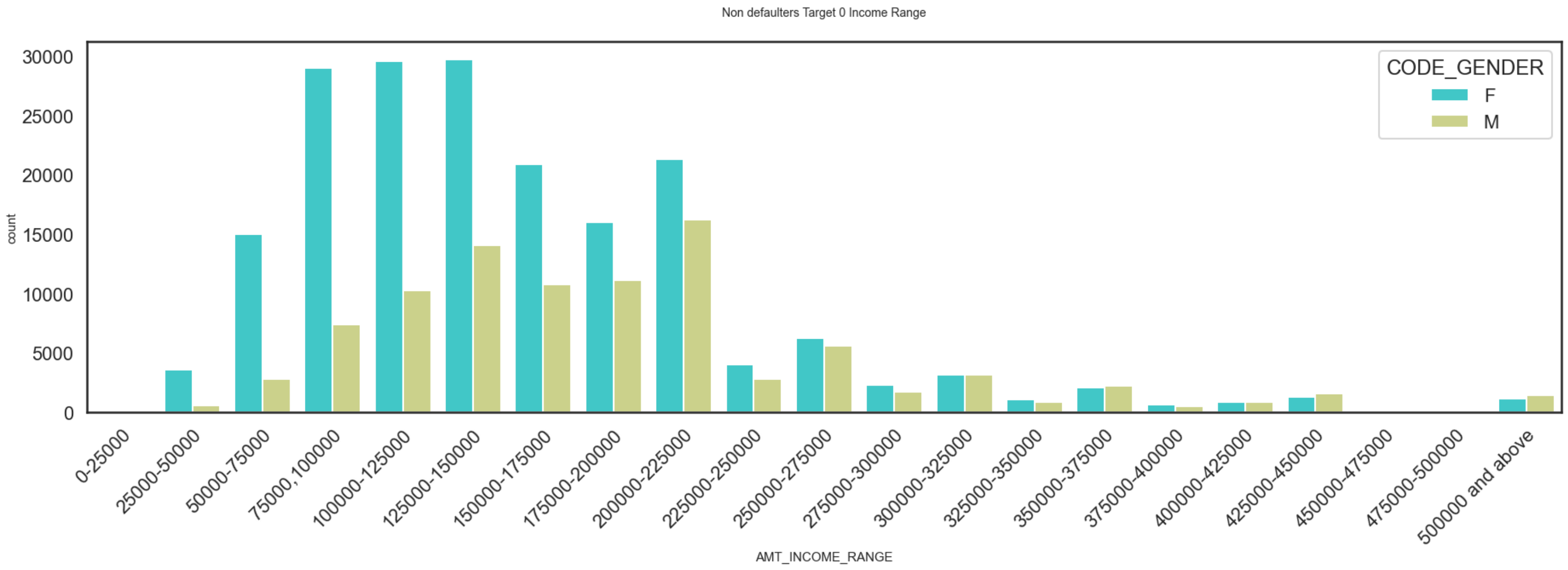

```
# Calculating Imbalance percentage
imbalance=round(len(target0)/len(target1),2)
imbalance
```

```
11.39
```

Age vs Targets

Interpretation :
Defaulters are mostly between the age from 25 to 40 yrs and then it is reducing swiftly

Univariate analysis on Application Data CSV
Categorical Analysis - Income range vs code Gender

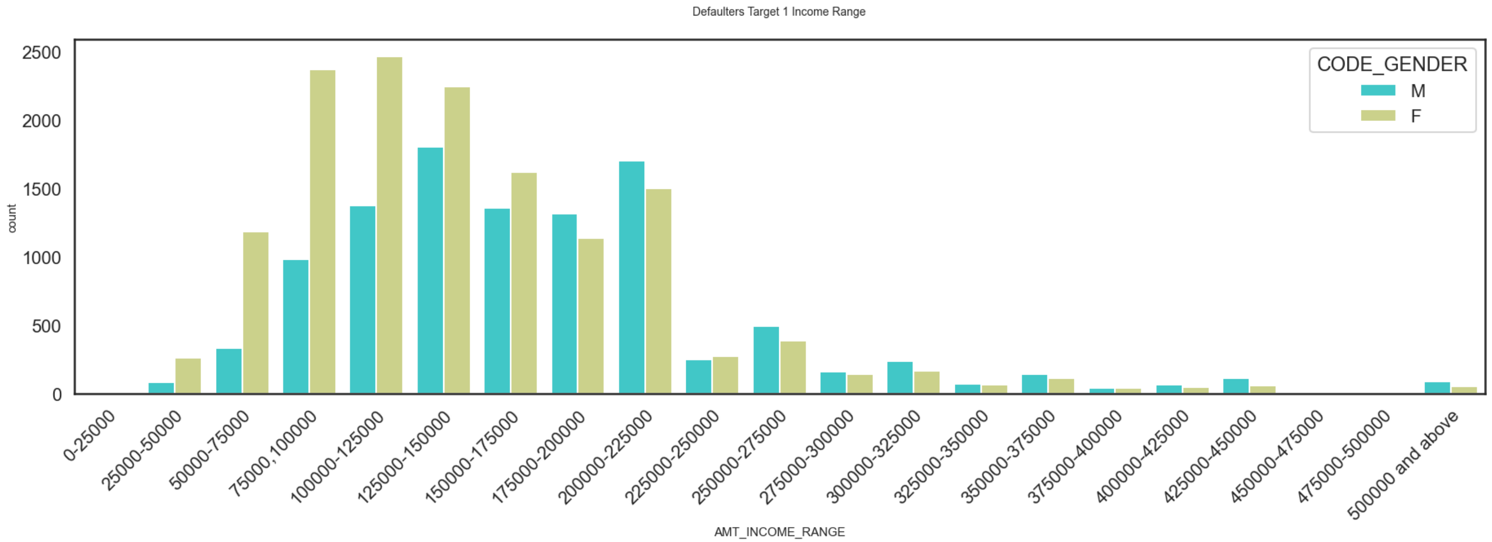


Interpretation :
For gender Non Defaulters
1. Females are higher than males and have more credits for almost all the ranges of income.

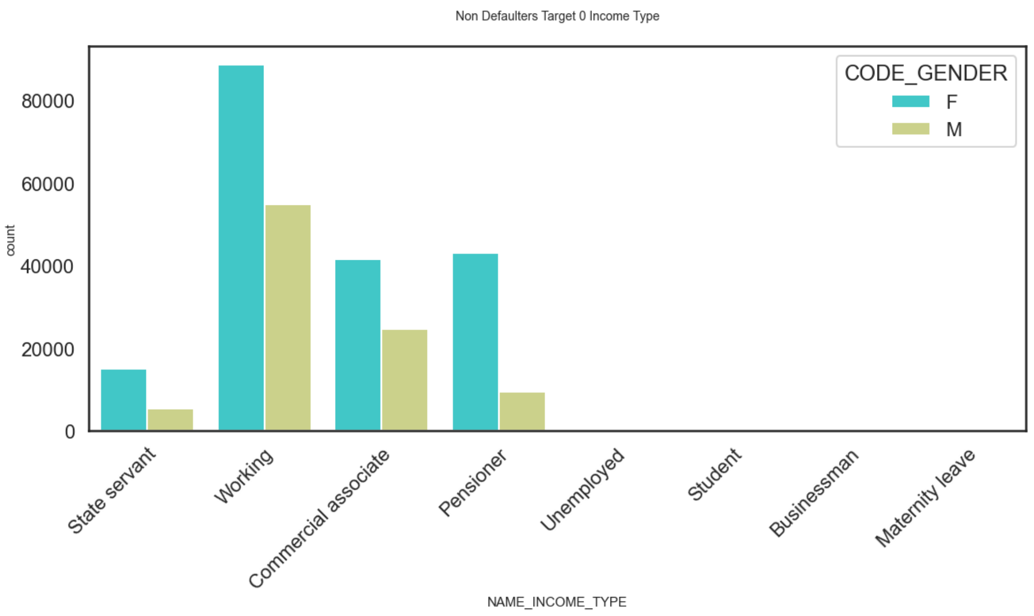
Interpretation:
For defaulters

Males are higher defualters than females for the income of 2 lack and above.

For less than 2 lac, female defaulters are maximum

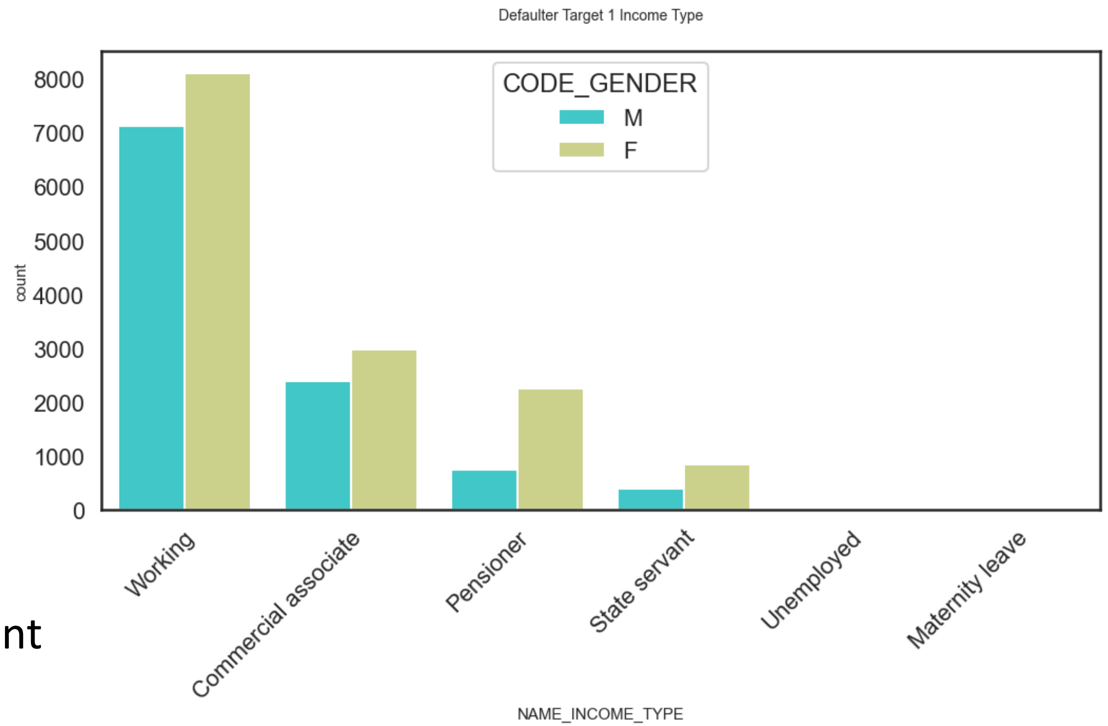


Income type analysis vs code gender for Both the targets



Interpretation:
For Non defaulters

1. In all income type females are higher loan takers and non defaulters

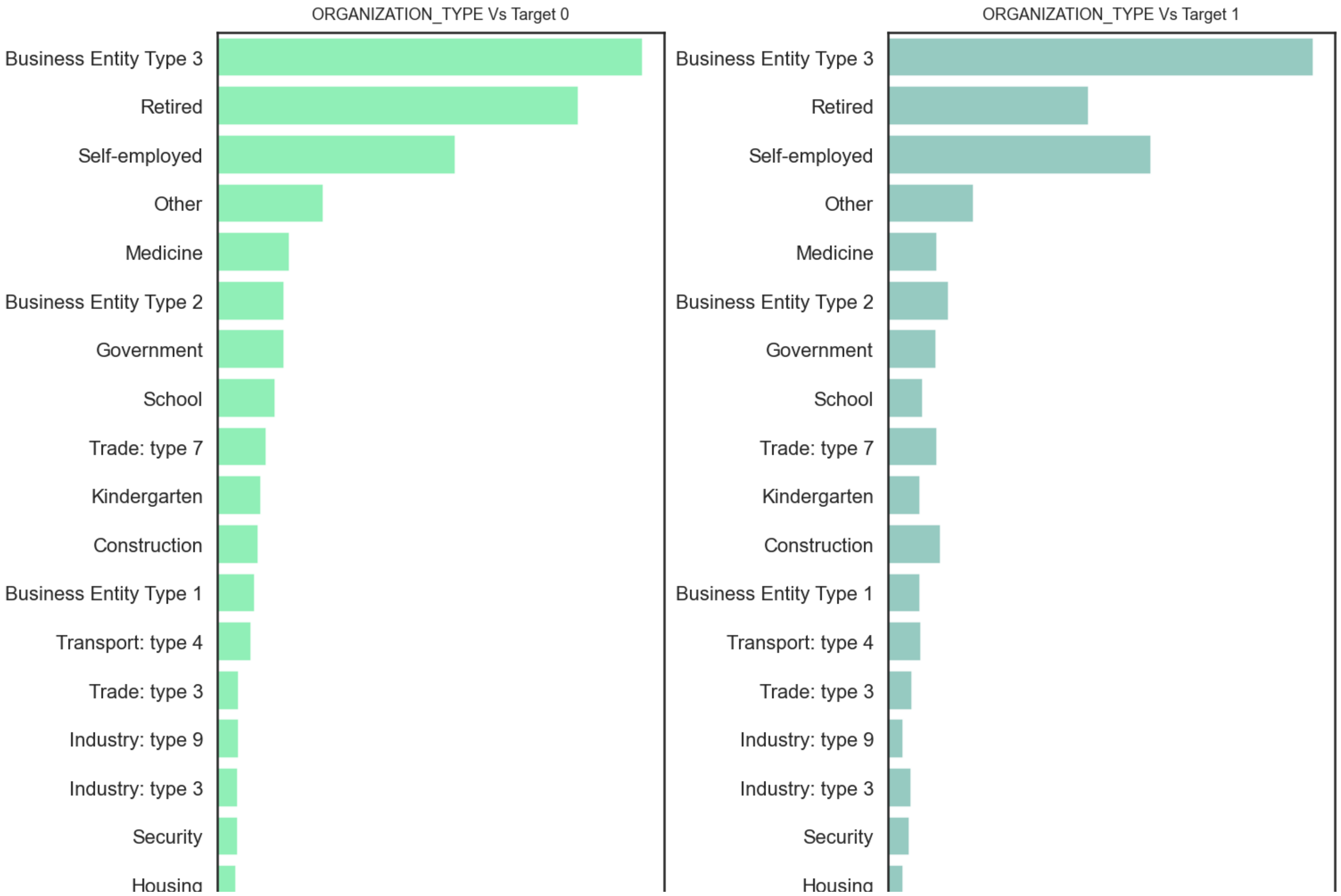


Interpretation:
For defaulters

1. Both Males and females on working class are defaulters Mostly

2. Females as commercial associate, pensioner and state servant also defaults more that Men

Organization Type analysis VS both the targets individually

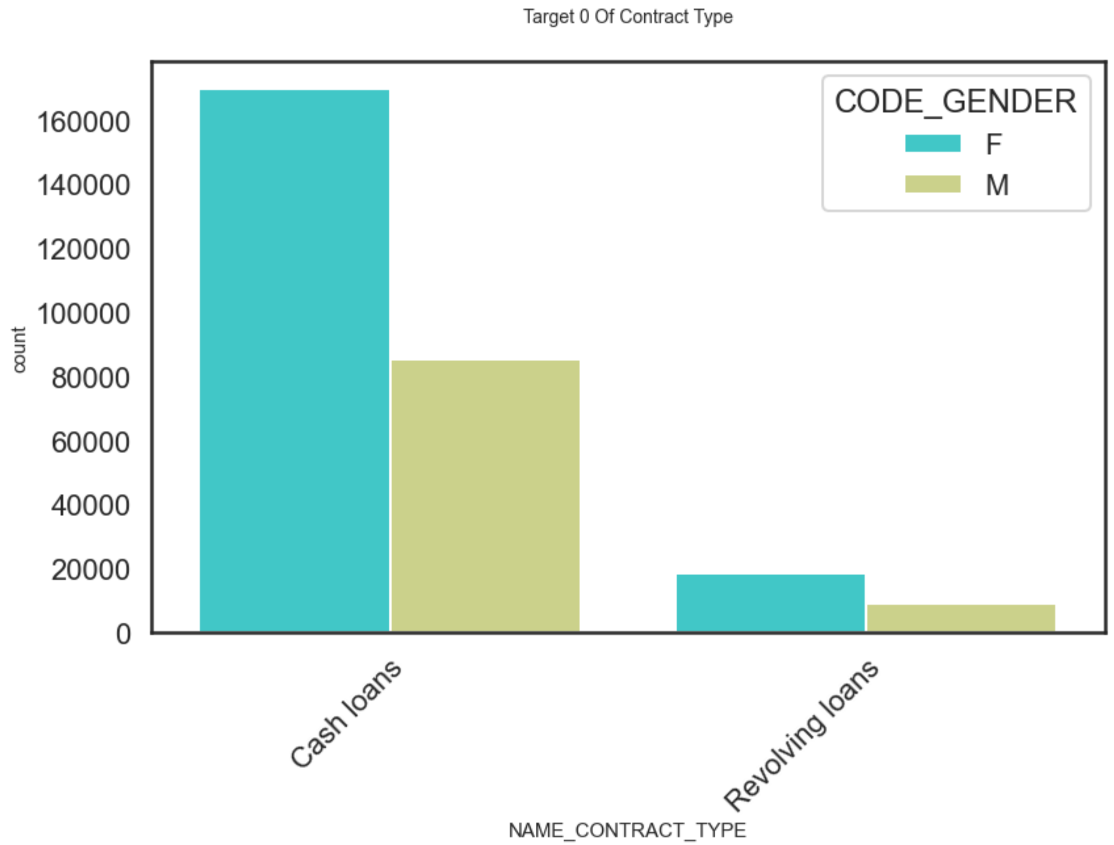


Interpretation:

1. Business Entity type 3, self employed and construction type tend to default more

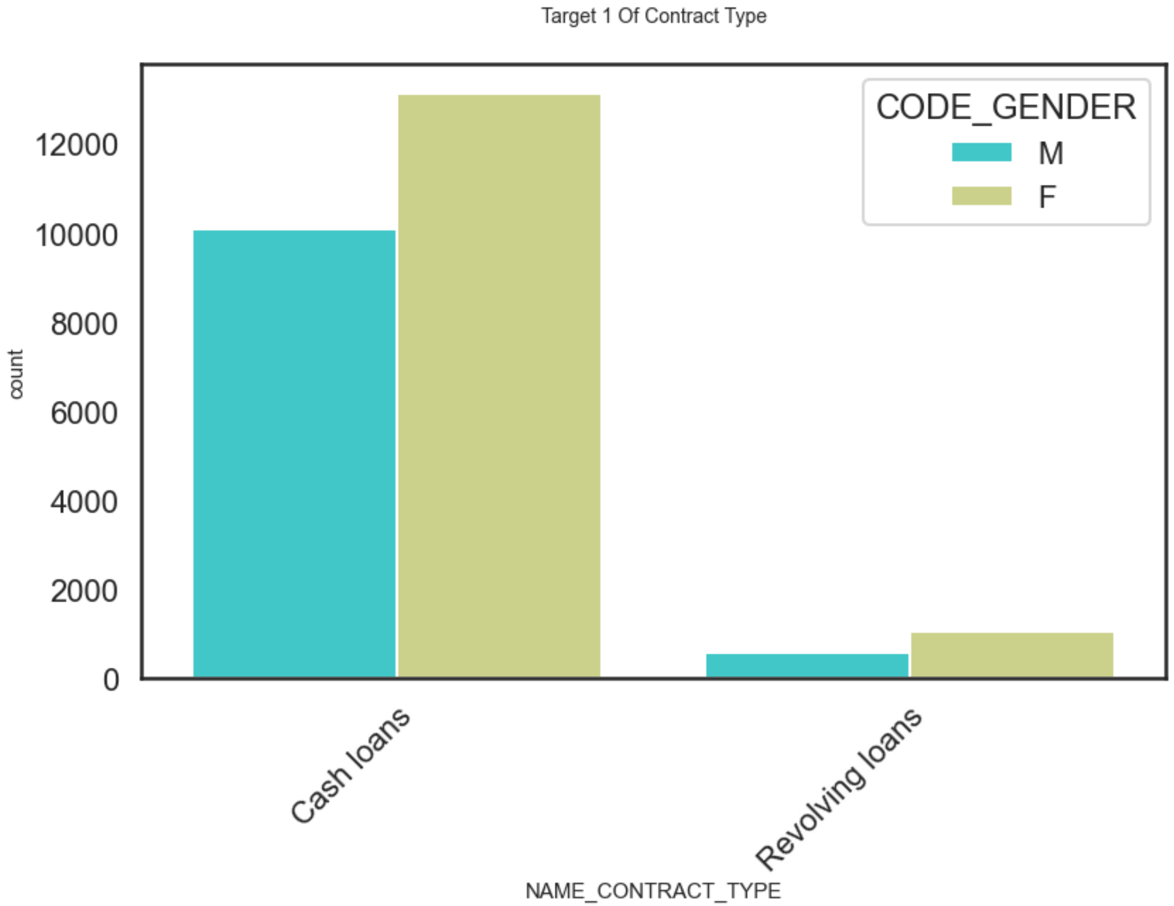
Retired person in comparison are less defaulters then top 3 mentioned above

Plotting for NAME_CONTRACT_TYPE for both the targets



Interpretation:
For Non defaulters

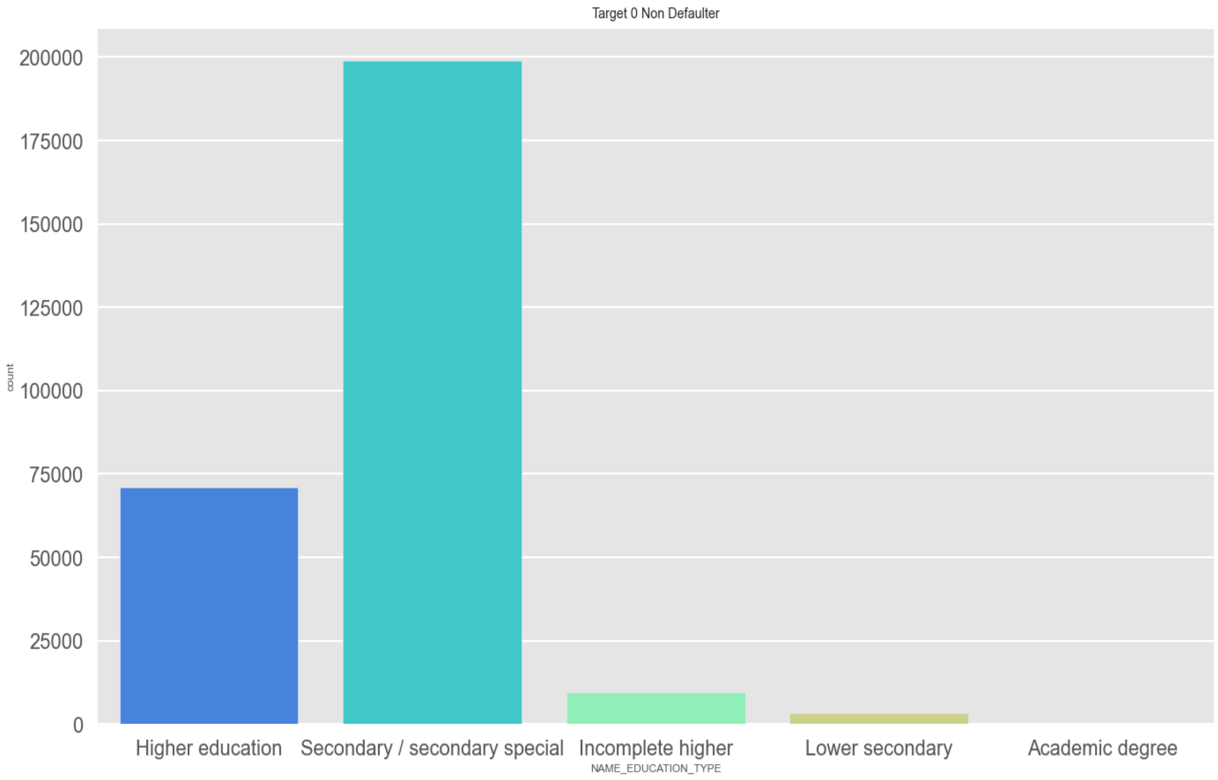
Cash Loan contracts have a higher number female loan takers and than of revolving loan contracts

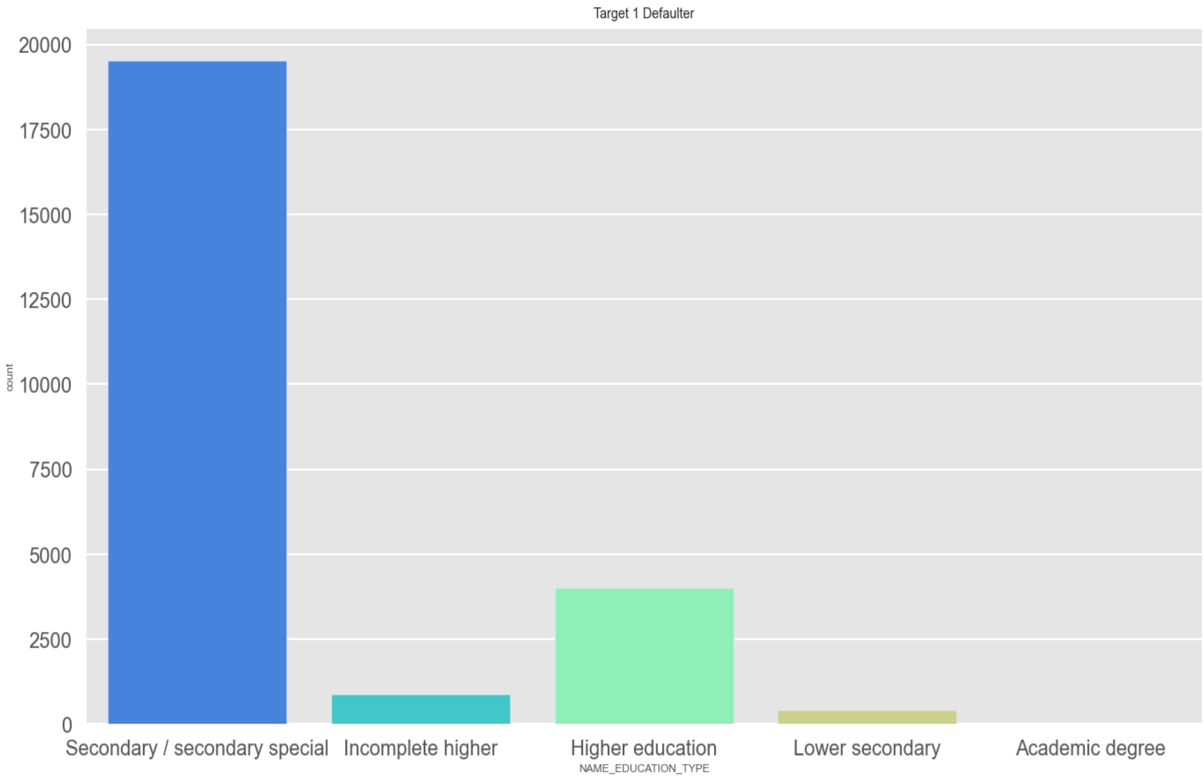


Interpretation:
For defaulters

Cash Loan contracts have the highest female defaulters and also in revolving loan contracts females are higher defaulters

Plotting for NAME_EDUCATION_TYPE for both the targets

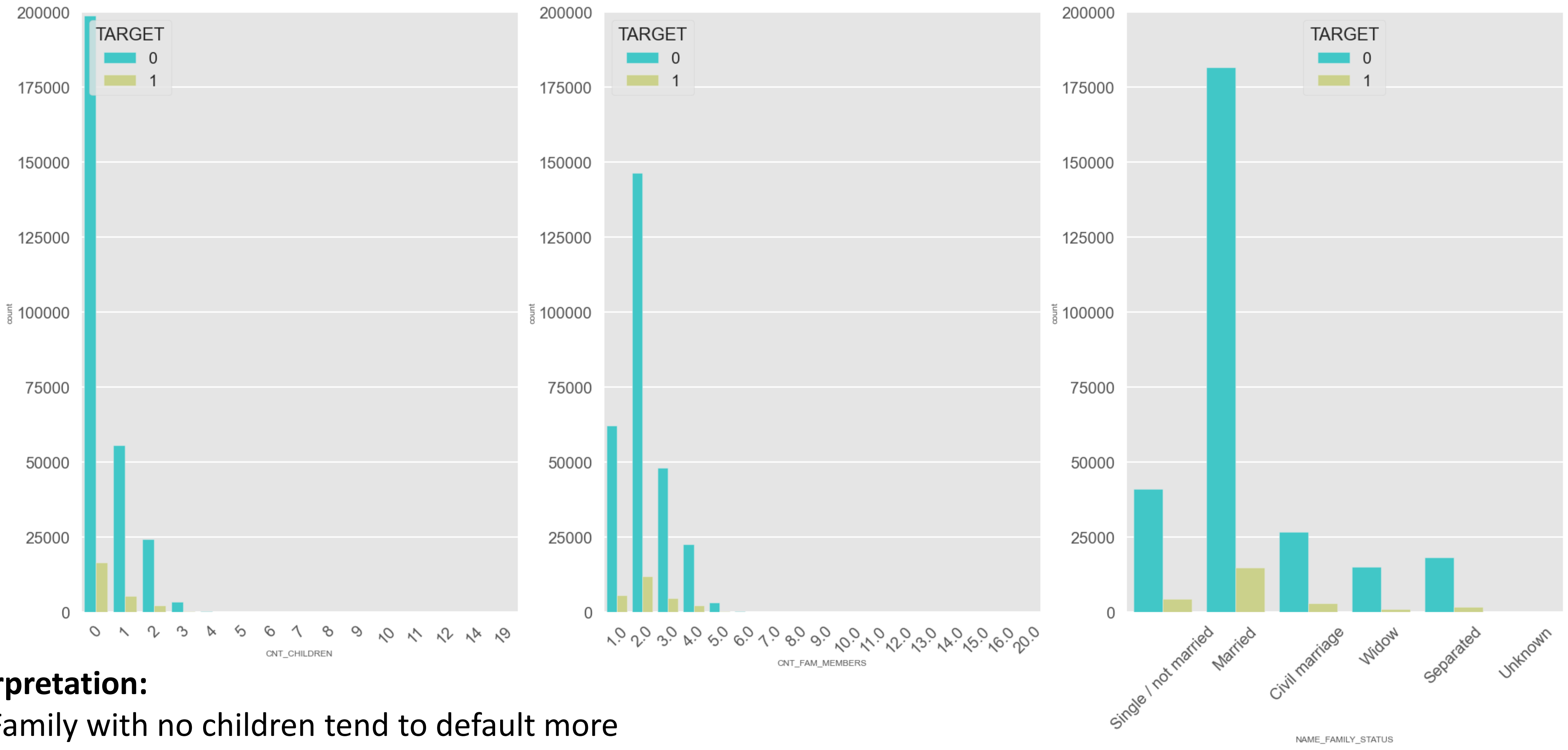




Interpretation:

Secondary educated loan takers are the highest defaulters, followed by higher educated people and them incomplete higher

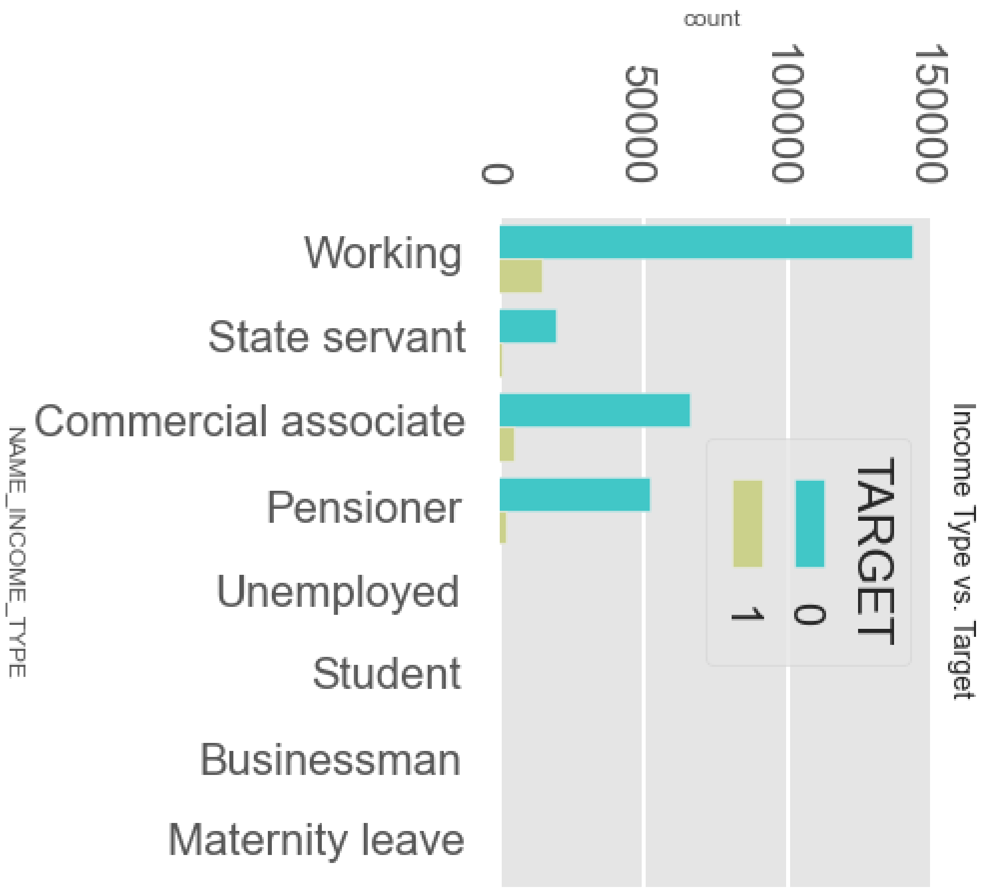
Count of Children, family members and family status vs Targets



Interpretation:

1. Family with no children tend to default more
2. Family with 2 members followed by 3 and 1 member are defaulting
3. Loan takers with married status are tend to default more than singe and civil marriages

Income type vs Targets



Interpretation:
For Non defaulters

Working people are tend to default more followed by commercial associates and then pensioners.

Correlation in the application Data dataset overall

	Column1	Column2	Correlation	Abs_Correlation
211	CNT_FAM_MEMBERS	CNT_CHILDREN	0.879160	0.879160
299	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.860627	0.860627
359	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.825575	0.825575
99	AMT_ANNUITY	AMT_CREDIT	0.770127	0.770127
159	DAYS_EMPLOYED	DAYS_BIRTH	0.623879	0.623879
279	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.450804	0.450804
339	REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.440409	0.440409
317	REG_CITY_NOT_LIVE_CITY	REG_REGION_NOT_LIVE_REGION	0.339232	0.339232
178	DAYS_REGISTRATION	DAYS_BIRTH	0.331796	0.331796
135	DAYS_BIRTH	CNT_CHILDREN	-0.330893	0.330893

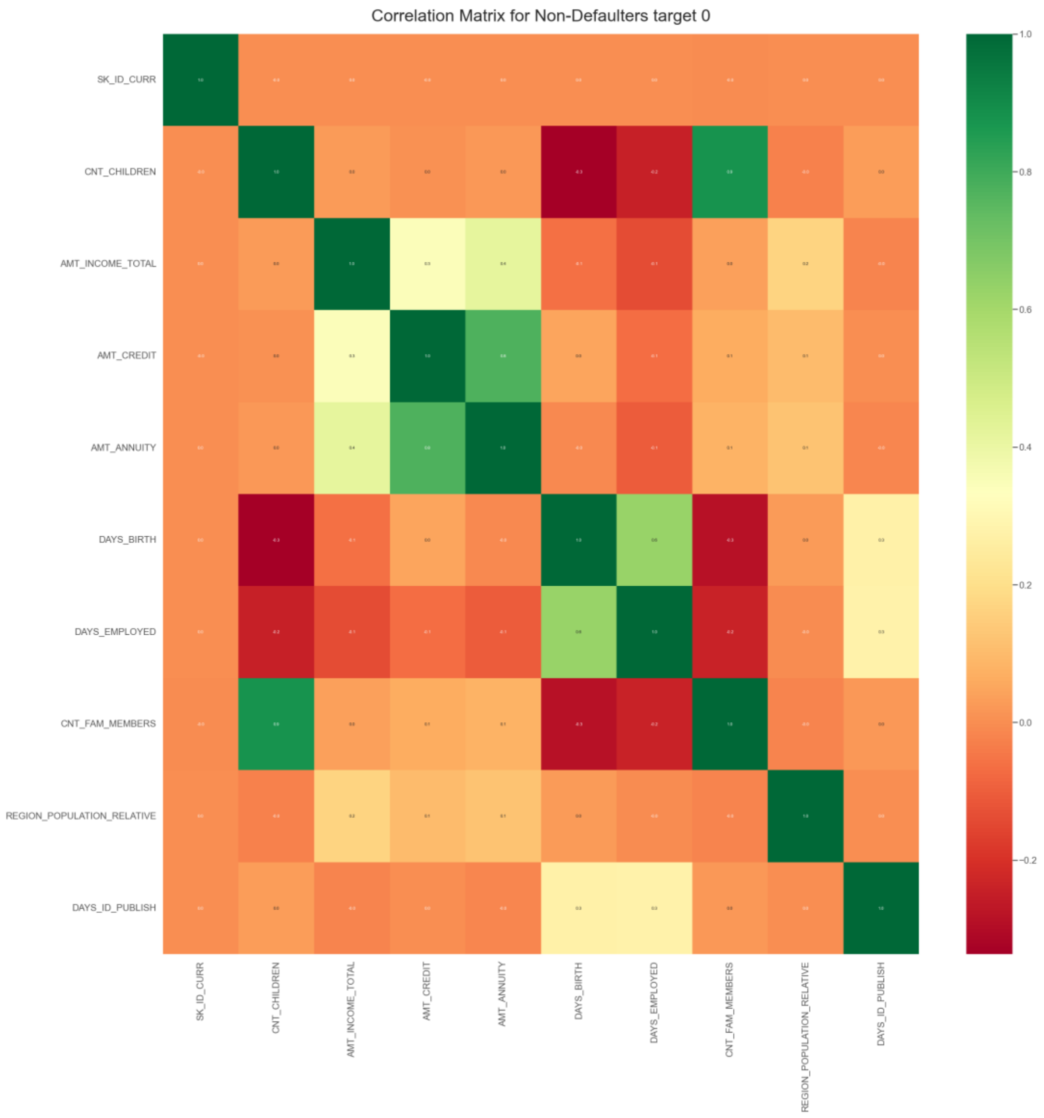
Top 10 correlation target 0

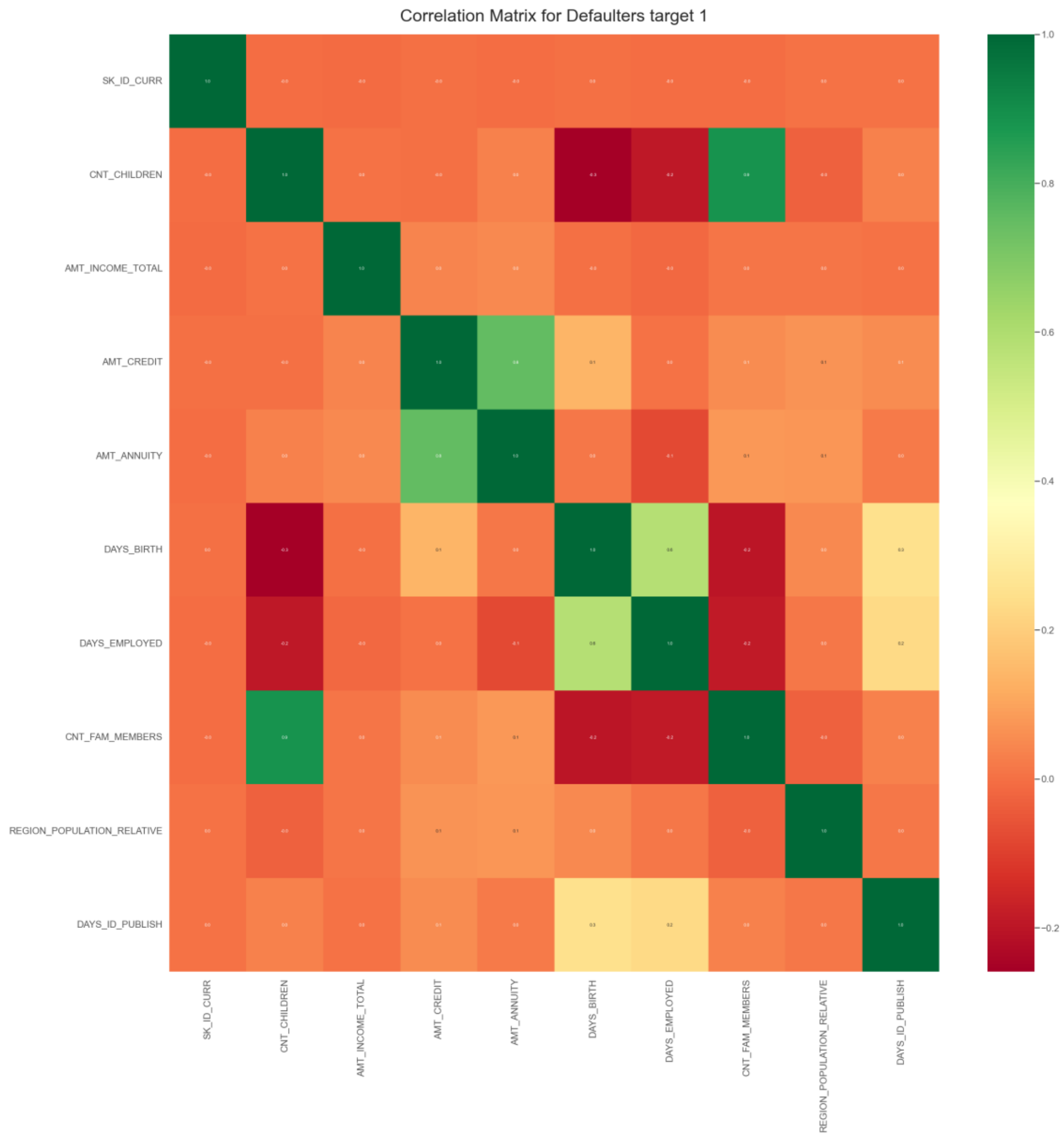
```
SK_ID_CURR                   SK_ID_CURR                    1.000000
CNT_CHILDREN                 CNT_FAM_MEMBERS               0.878571
LIVE_REGION_NOT_WORK_REGION  REG_REGION_NOT_WORK_REGION    0.861861
LIVE_CITY_NOT_WORK_CITY      REG_CITY_NOT_WORK_CITY        0.830381
AMT_CREDIT                   AMT_ANNUITY                   0.771297
DAYS_EMPLOYED                DAYS_BIRTH                    0.626028
REG_REGION_NOT_LIVE_REGION   REG_REGION_NOT_WORK_REGION    0.446101
REG_CITY_NOT_WORK_CITY       REG_CITY_NOT_LIVE_CITY        0.435514
AMT_INCOME_TOTAL             AMT_ANNUITY                   0.418948
AMT_CREDIT                   AMT_INCOME_TOTAL              0.342799
REG_CITY_NOT_LIVE_CITY       REG_REGION_NOT_LIVE_REGION    0.341571
dtype: float64
```

Top 10 correlation target 1

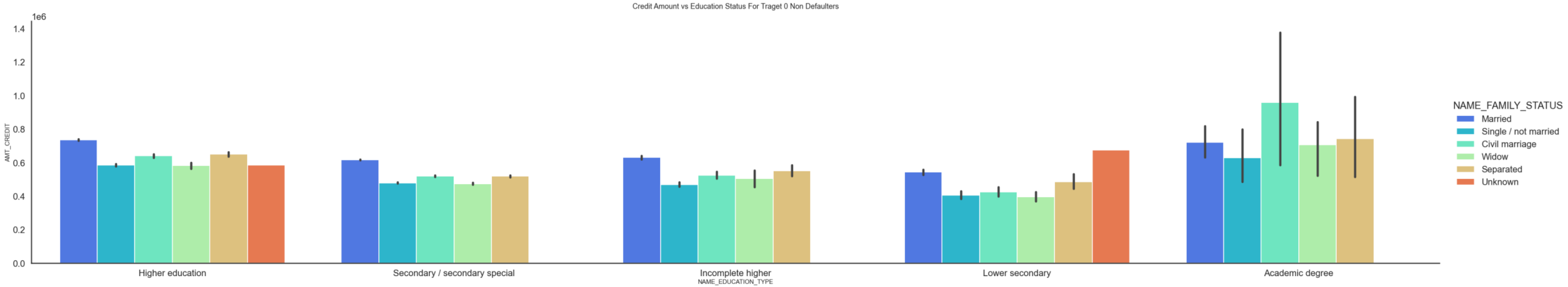
```
SK_ID_CURR                   SK_ID_CURR                    1.000000
CNT_FAM_MEMBERS              CNT_CHILDREN                  0.885484
LIVE_REGION_NOT_WORK_REGION  REG_REGION_NOT_WORK_REGION    0.847885
REG_CITY_NOT_WORK_CITY       LIVE_CITY_NOT_WORK_CITY       0.778540
AMT_ANNUITY                  AMT_CREDIT                    0.752195
DAYS_EMPLOYED                DAYS_BIRTH                    0.582441
REG_REGION_NOT_WORK_REGION   REG_REGION_NOT_LIVE_REGION    0.497937
REG_CITY_NOT_WORK_CITY       REG_CITY_NOT_LIVE_CITY        0.472052
REG_CITY_NOT_LIVE_CITY       REG_REGION_NOT_LIVE_REGION    0.322628
DAYS_REGISTRATION            DAYS_BIRTH                    0.289116
DAYS_BIRTH                   DAYS_ID_PUBLISH               0.252256
dtype: float64
```

Correlation matrix Heatmap representation





NAME_EDUCATION_TYPE vs AMT_CREDIT for each family status

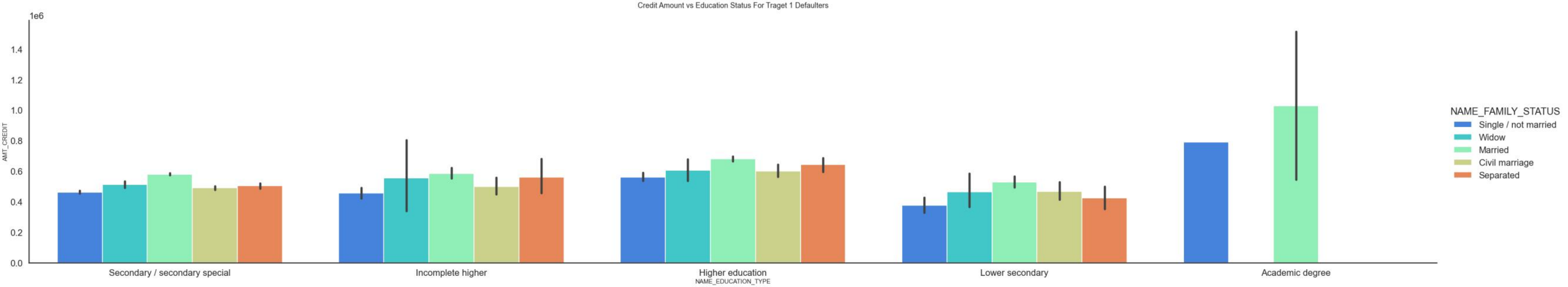


Interpretation:

1. Married people with academic degree has more defaulters followed by Higher education and secondary educated people
Are tend to default more

2. Separated people with Higher education and Incomplete higher also tend to default

3. Singles with an academic degree also has defaulted a lot

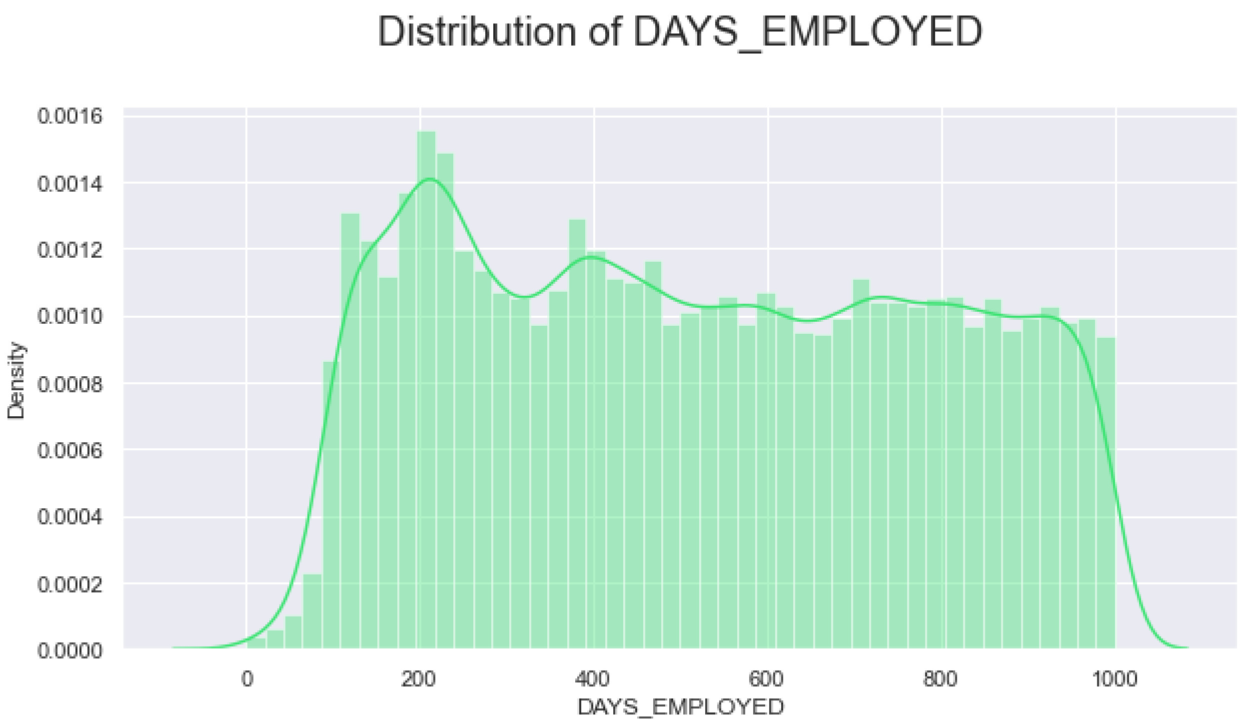


Analysis on distribution of 'DAYS_EMPLOYED

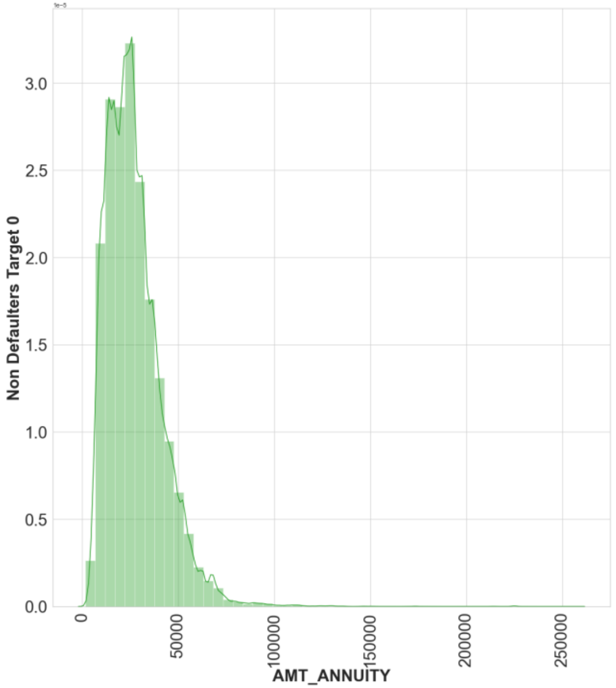
Interpretation:

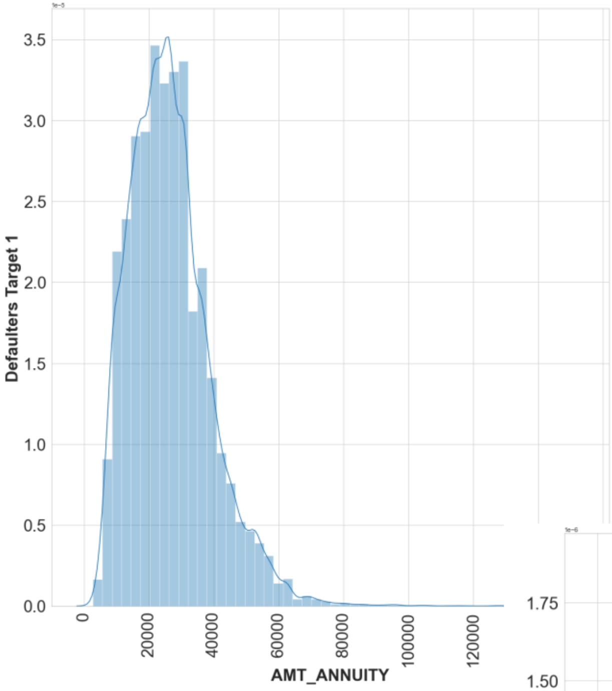
1. People between 0.5 to 1.5 years of employment has more tendency towards defaulting

2. People with about half a year of employment has the maximum defaulters



Analysis on AMT_Annuity and AMT_Credit vs Targets



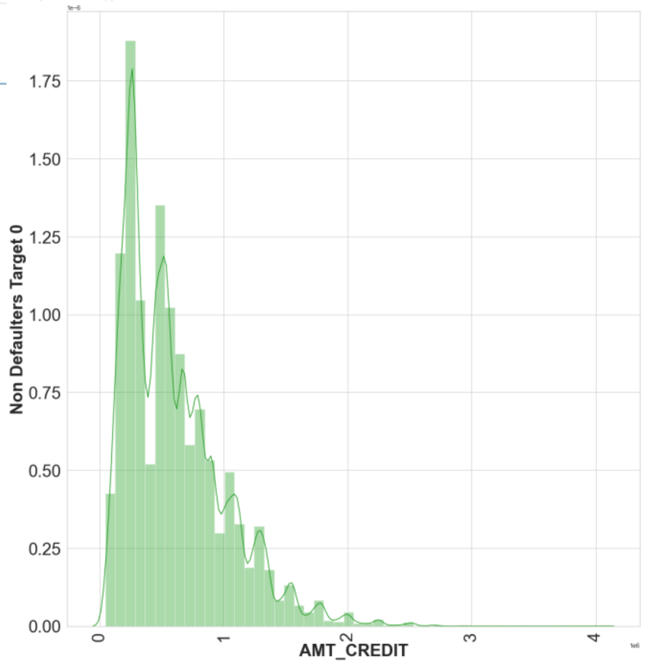


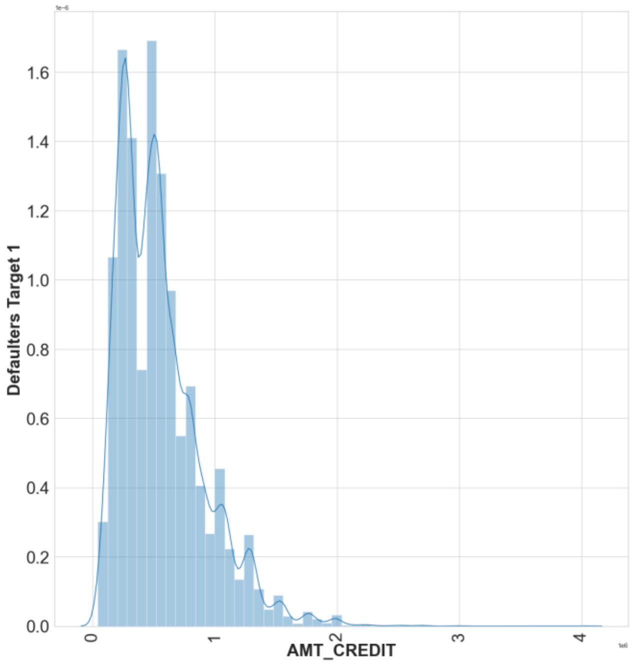
Interpretation:

1. People with amount annuity between 2,000 to 4,000 are defaulting more

Interpretation:

1. People with amount credit between 0.5 to 1 are defaulting more



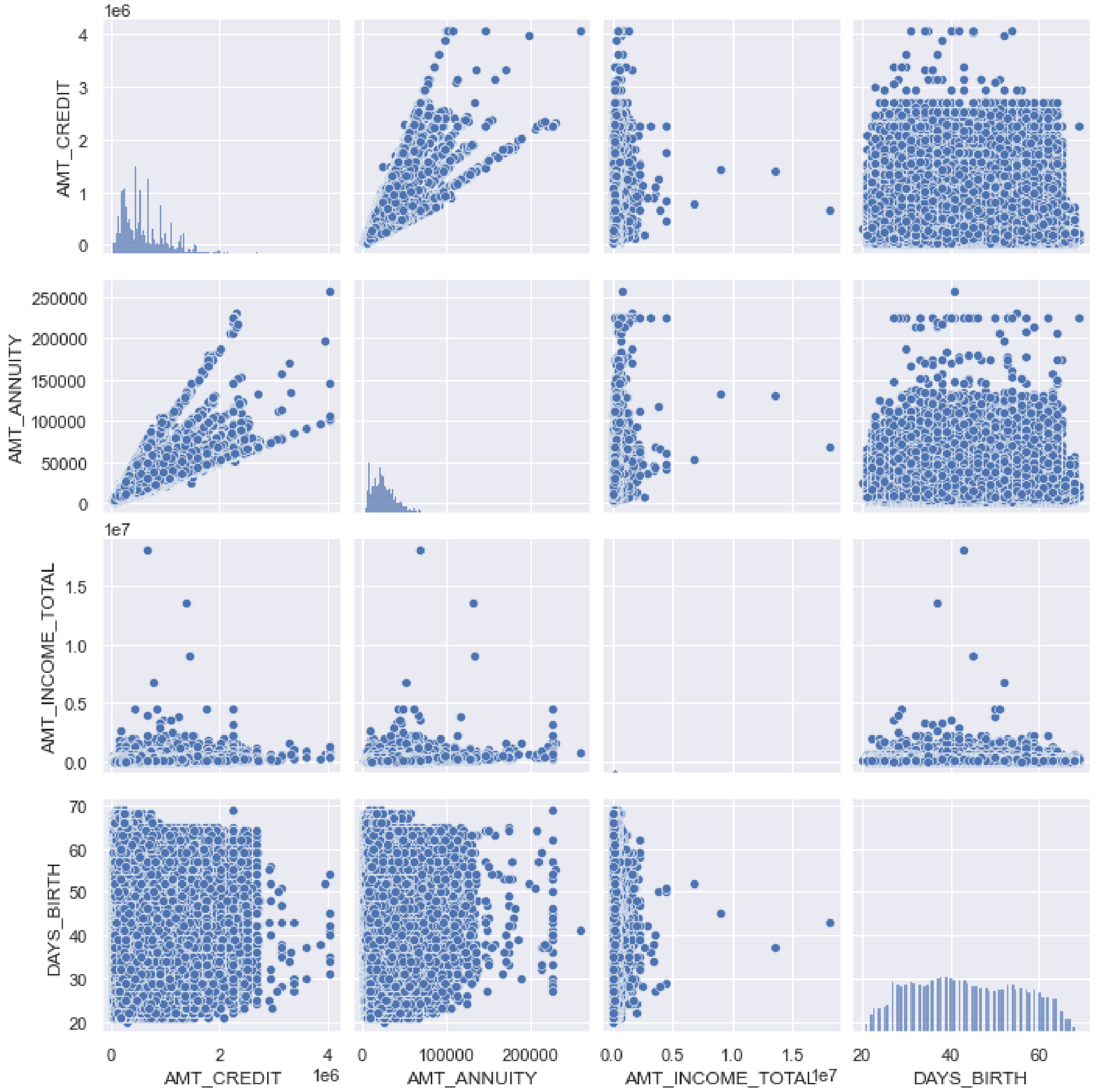


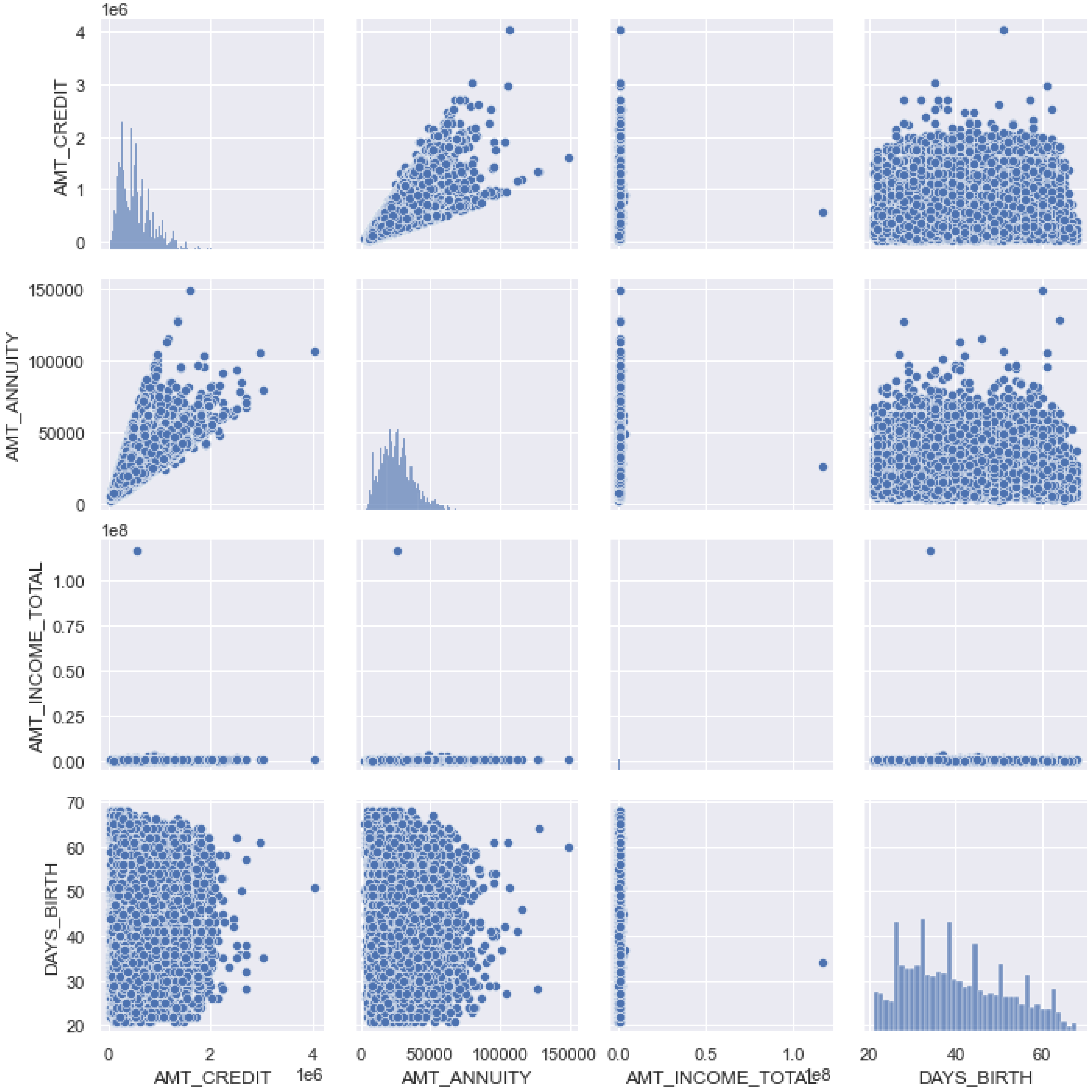
Bivariate Analysis
Numerical vs Numerical analysis

Target 0 Non Defaulters

Interpretation:

1. It is very clear that AMT_ANNUITY and AMT_CREDIT has positive correlation





Target 1 Defaulters

Interpretation:

1. It is very clear that AMT_ANNUITY and AMT_CREDIT has positive correlation

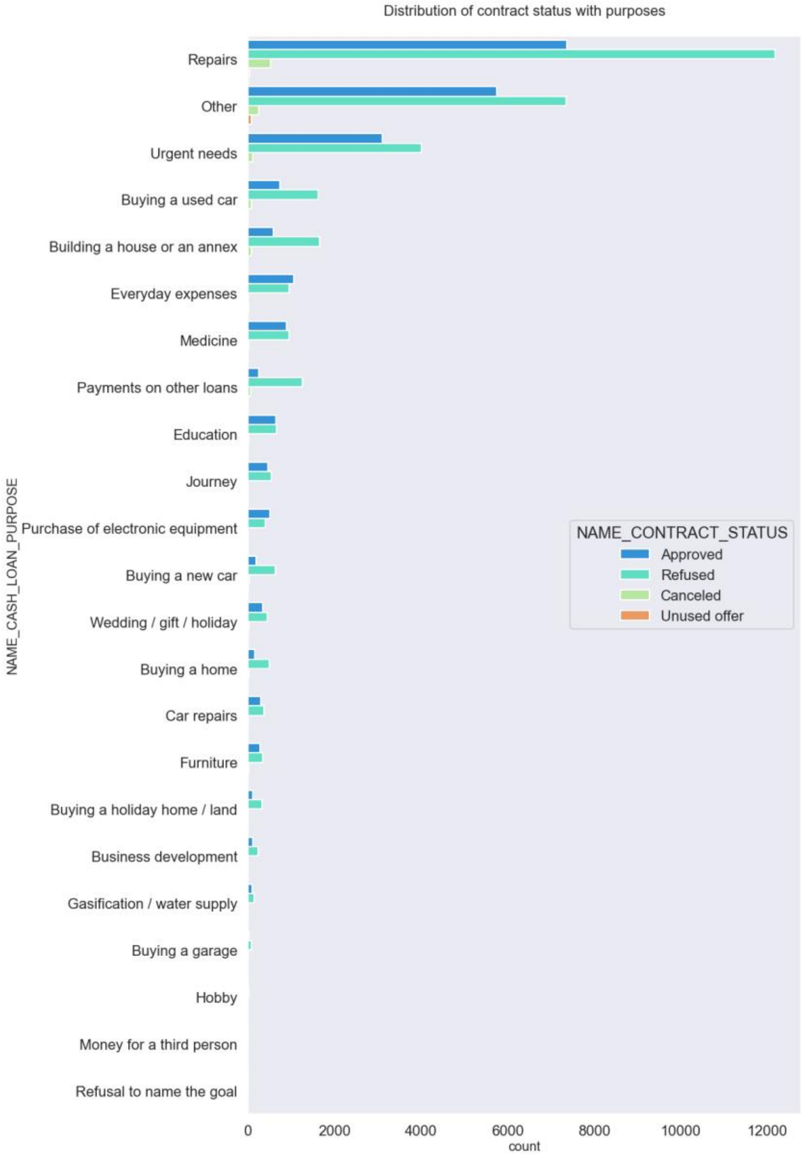
Data analysis of Previous Application

- Merging both Application dataset with previous application dataset with an inner merge

Merge Data set Univariate analysis

Interpretation:
Those who had approved loan earlier has lesser defaulted

Univariate Analysis
Distribution of contract status vs purpose



Interpretation:

Repairs have the highest refusals followed by the other category of application and then urgent needs

Also, loan taken for others show an unused offer

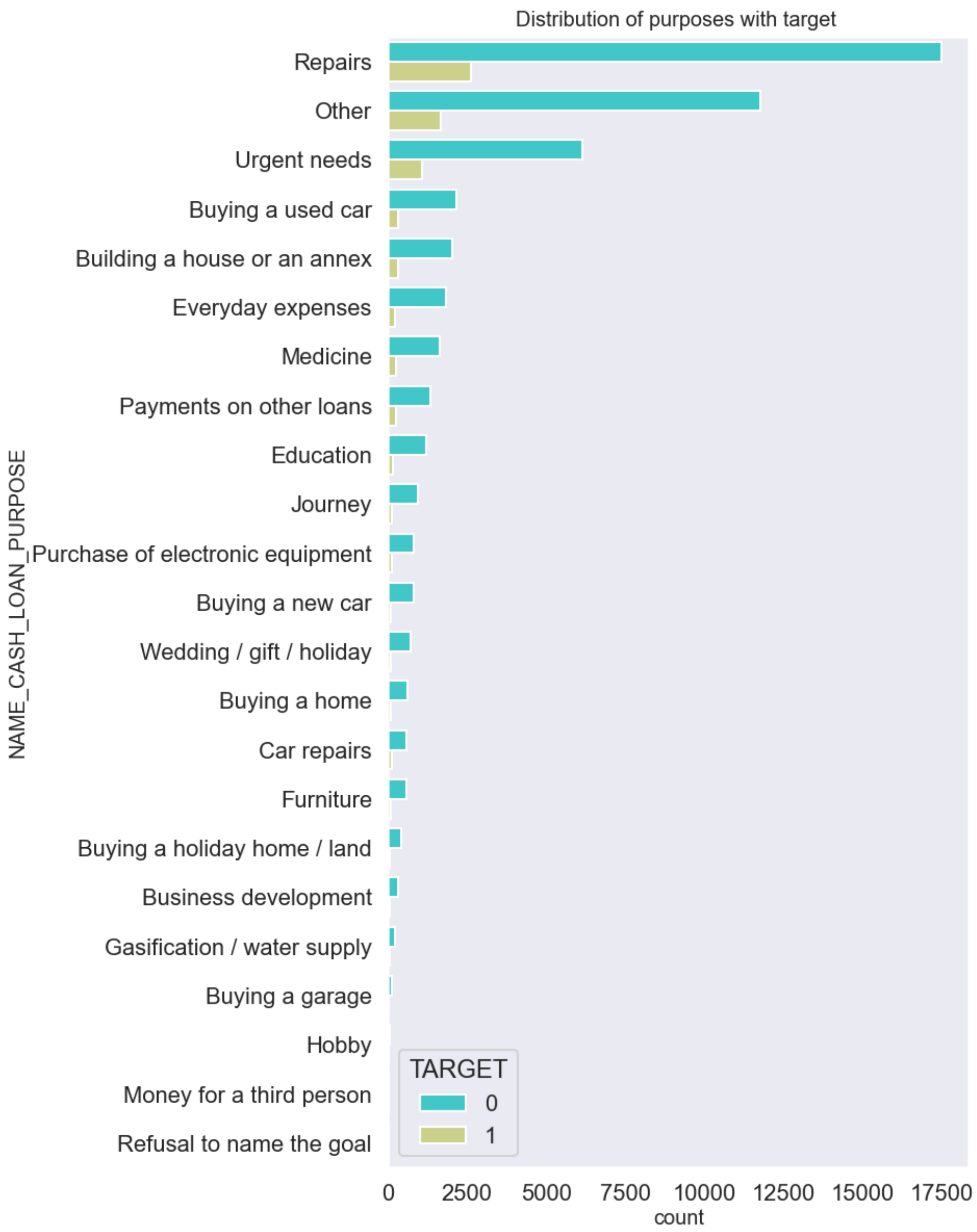
education purposes there is an similar number of approves and rejection

For buying cars, home and buying holiday home also has more rejections

Distribution of purpose vs target

Interpretation:
Repairs have the highest defaulters followed by others and then urgent needs

Majority people taking loan for Education has returned it but still few defaulters are visible there as well



Bivariate analysis on merged data

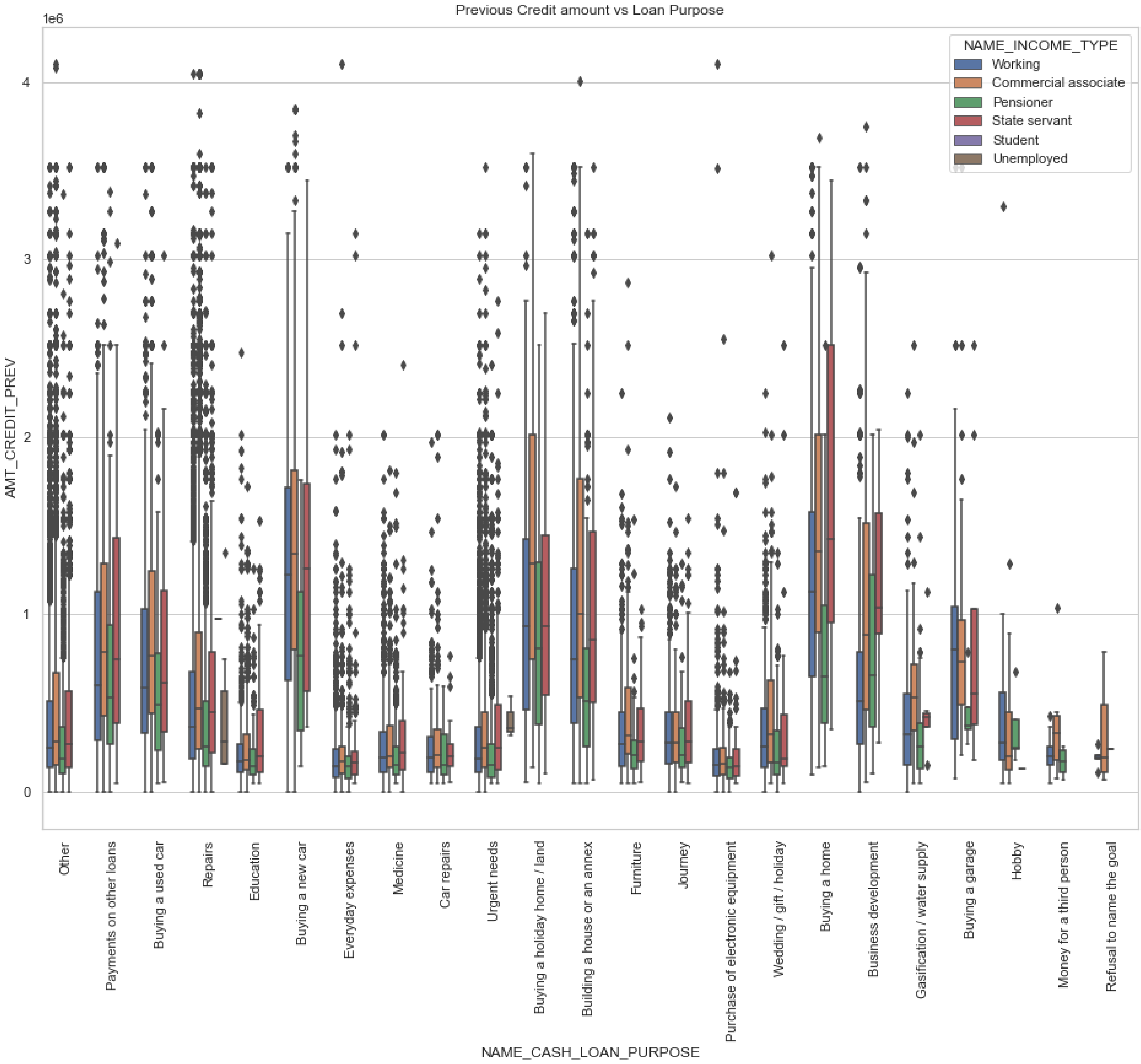
Box plotting for previous Credit amount vs Housing type in logarithmic scale

Interpretation:

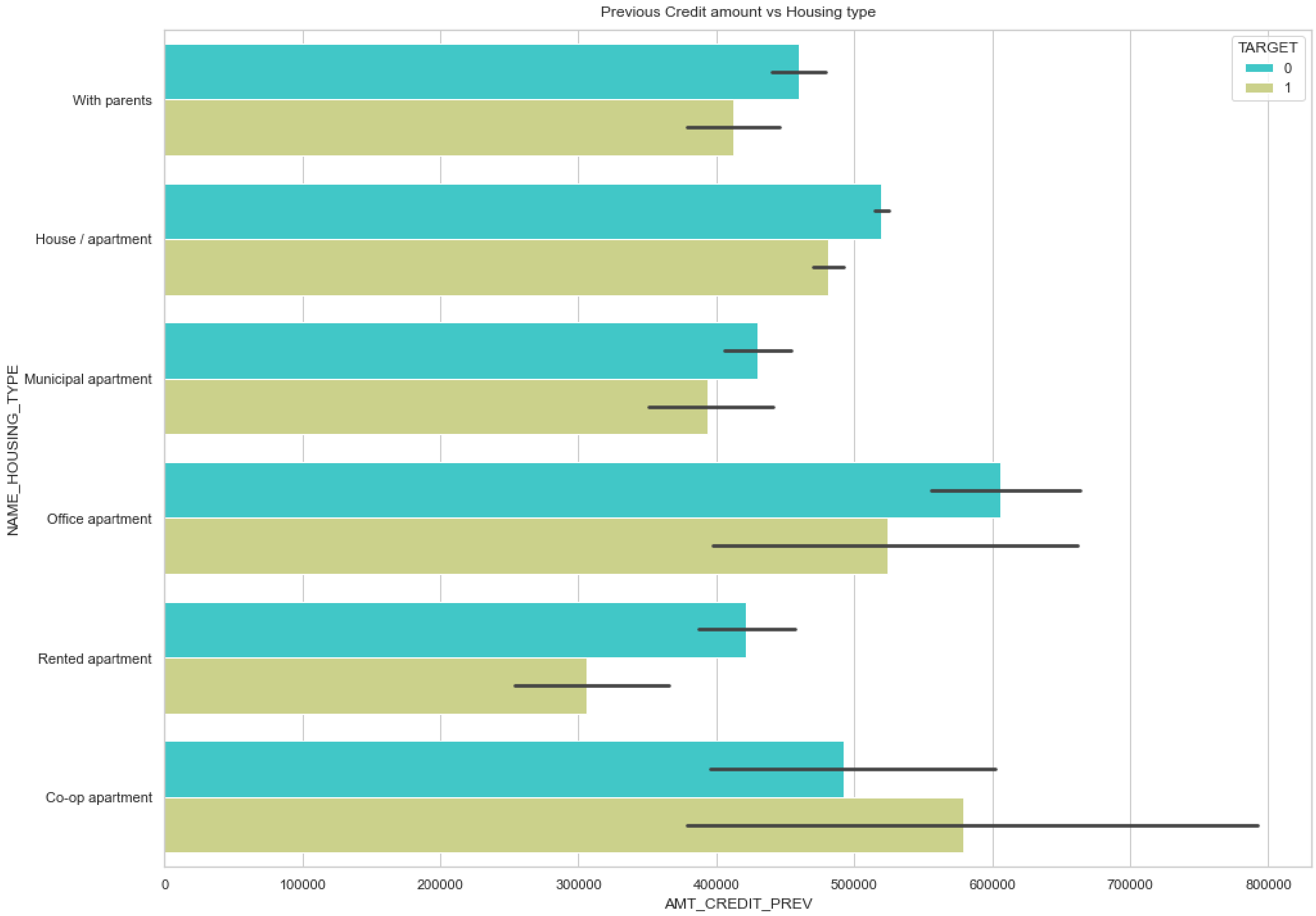
1. Loan purposes of 'Buying a home','Buying a land','Buying a new car' and'Building a house' has a higher credit amount

2. State servants have applied the highest for buying new car and home

3. Commercial associates also have more credit amount for buying new car, home, house, garage and money for third person or a Hobby



Previous Credit amount vs Housing type

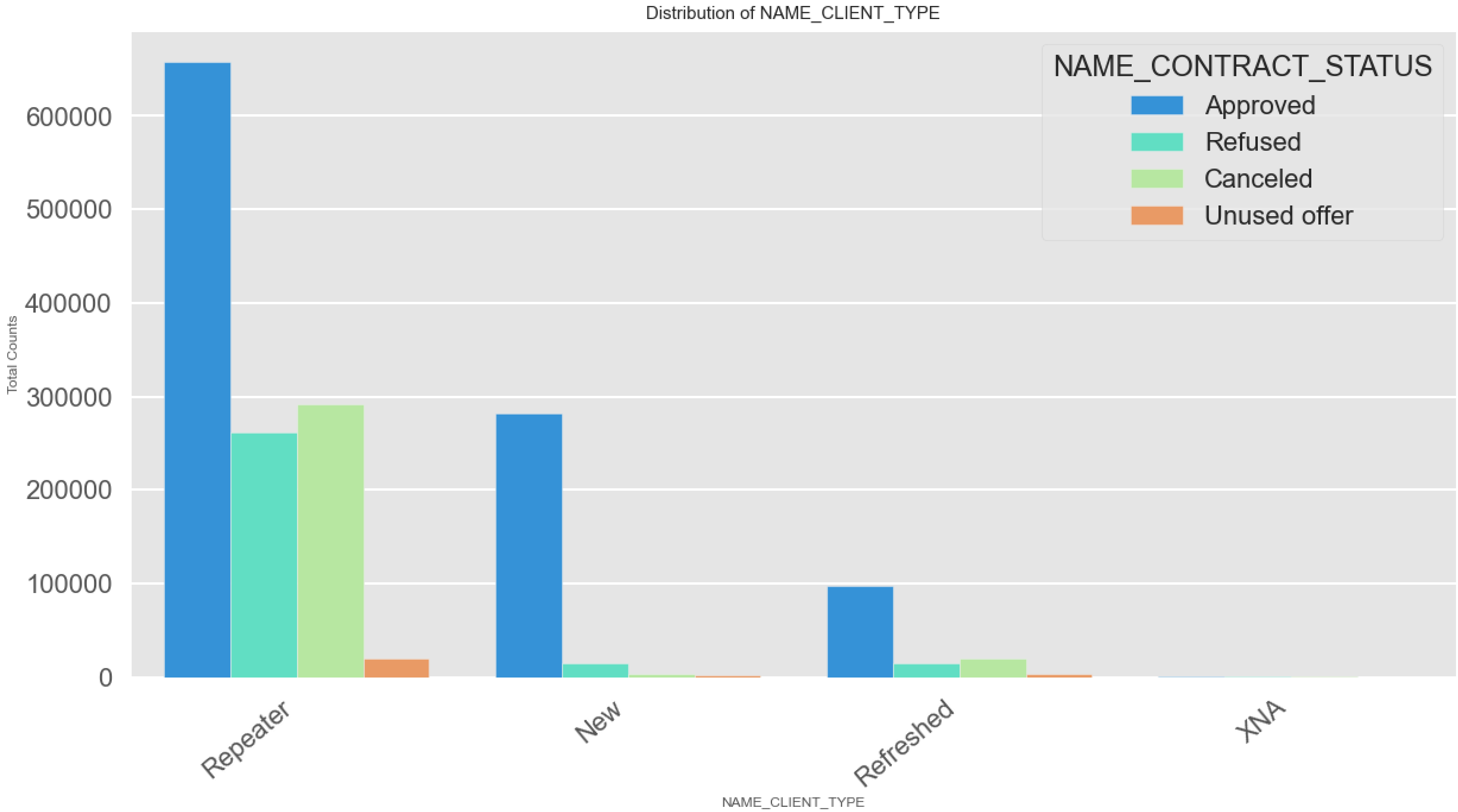


Interpretation:

1. Co-op apartment, office apartment and apartment has higher defaulters

2. Office apartment also has higher defaulters however it also has people with non defaulting background

categorical variables - function to countplot Name client type with contract status



Interpretation:

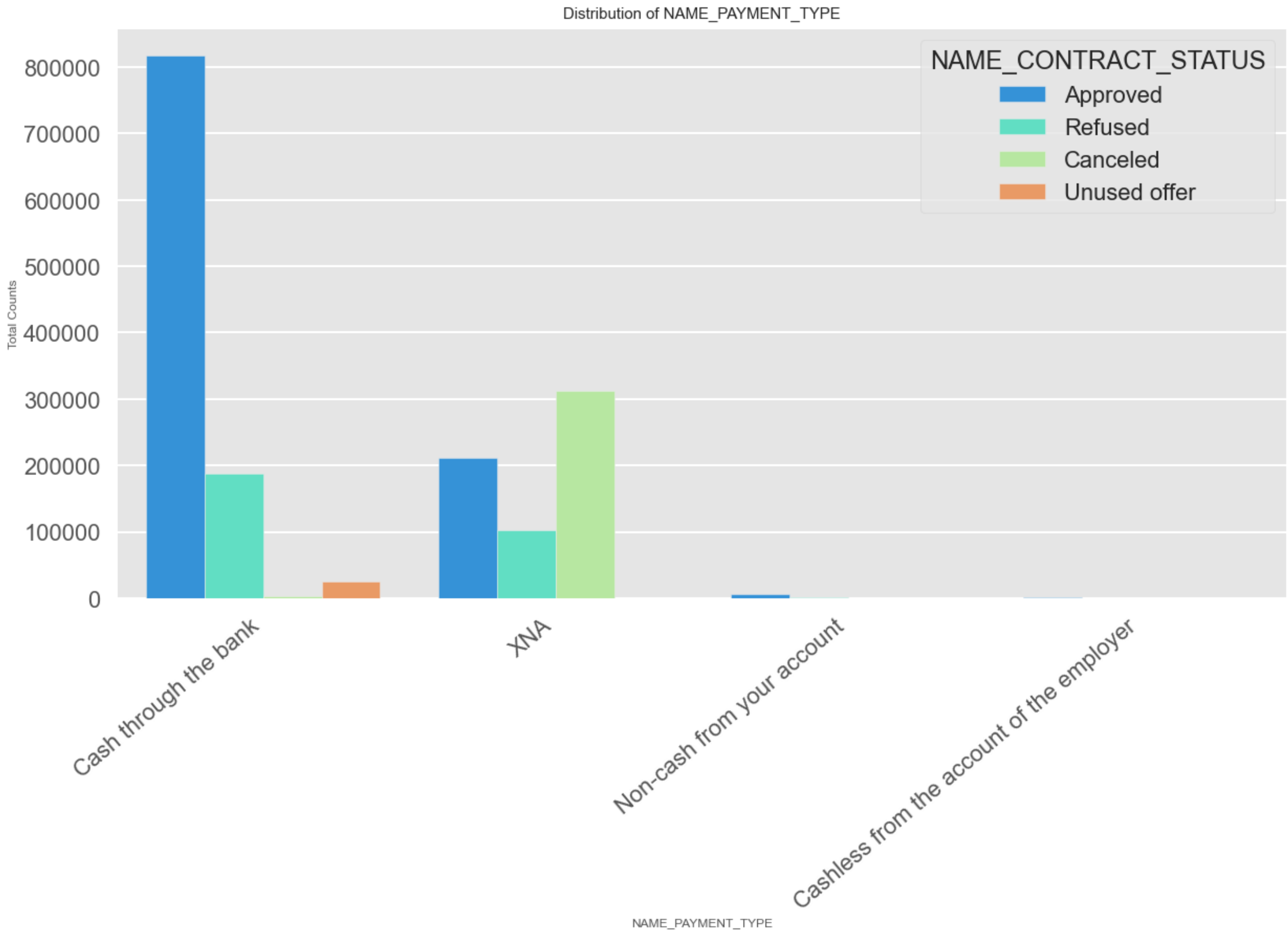
1. Most repeaters have got loan refusal
2. Also few new and refreshed has numbers of refusal
3. Majority repeaters have got their loans approved followed by new and then refreshed

Name payment type with contract status

Interpretation:

1. Majority cahs through the bank loans had been approved

2. Where as the refusal is also significant here

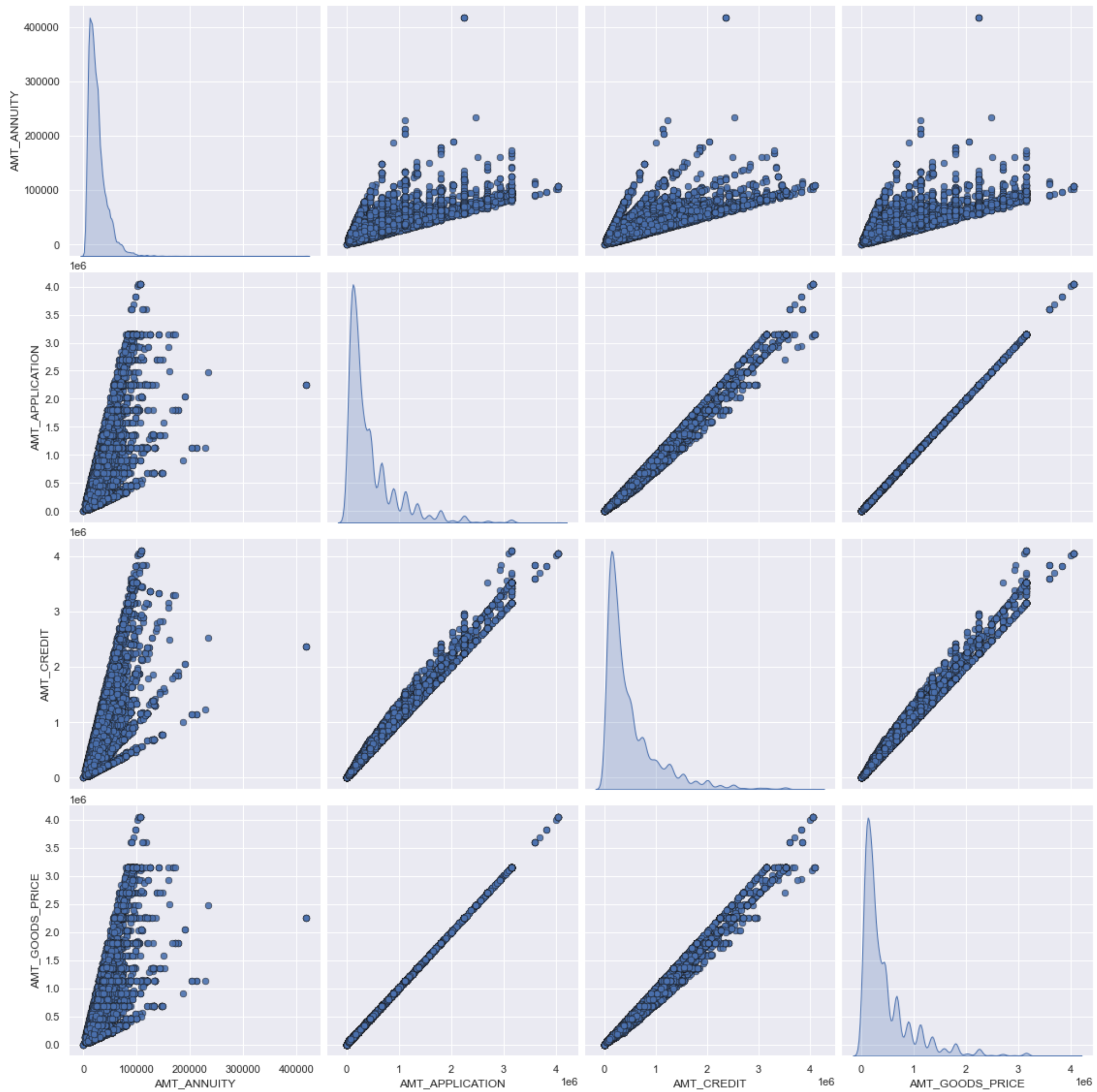


Correlation in the Previous application dataset

	Column1	Column2	Correlation	Abs_Correlation
129	AMT_GOODS_PRICE	AMT_APPLICATION	1.000000	1.000000
87	AMT_CREDIT	AMT_APPLICATION	0.994941	0.994941
130	AMT_GOODS_PRICE	AMT_CREDIT	0.994941	0.994941
417	DAYS_TERMINATION	DAYS_LAST_DUE	0.987981	0.987981
369	DAYS_LAST_DUE_1ST_VERSION	DAYS_DECISION	0.823877	0.823877
65	AMT_APPLICATION	AMT_ANNUITY	0.784131	0.784131
128	AMT_GOODS_PRICE	AMT_ANNUITY	0.784131	0.784131
86	AMT_CREDIT	AMT_ANNUITY	0.780327	0.780327
298	CNT_PAYMENT	AMT_CREDIT	0.677761	0.677761
371	DAYS_LAST_DUE_1ST_VERSION	CNT_PAYMENT	0.661561	0.661561

pairplot for bivariate analysis on numerical columns

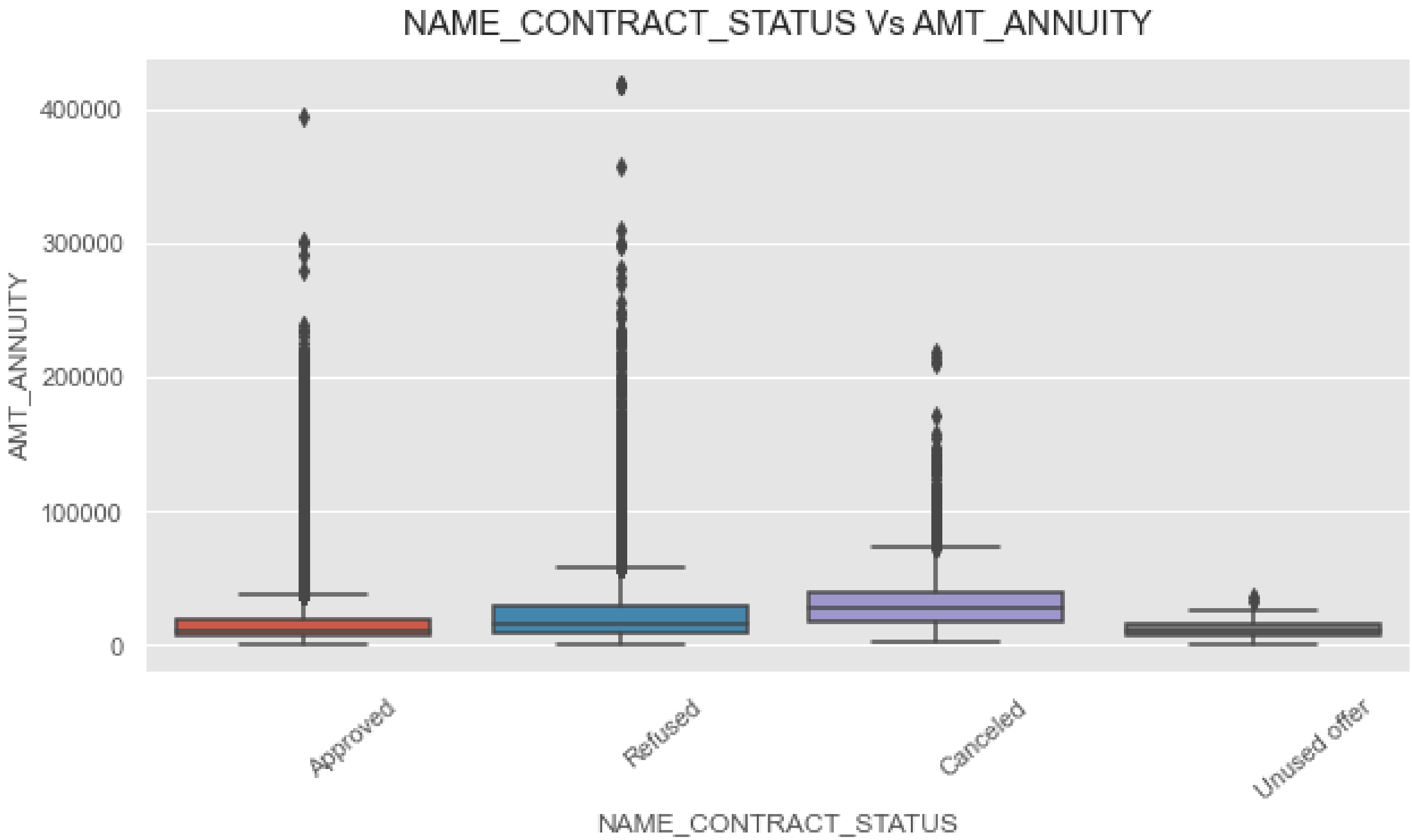
plotting the relation between correlated highly corelated numeric variables



Interpretation:

1. It is significant positive correlation we can observe for AMT_ANNUITY AND AMT_APPLICATION AND AMT_CREDIT AND AMT_GOODS
2. AMT_APPLICAITON has shown positive correlation with AMT_GOODS_PRICE and AMT_Credit

bivariate analysis of Contract status and Annuity of previous application



Interpretation:

1. Cancelled application predominantly is higher for the AMT_ANNUITY less than 1 Lac

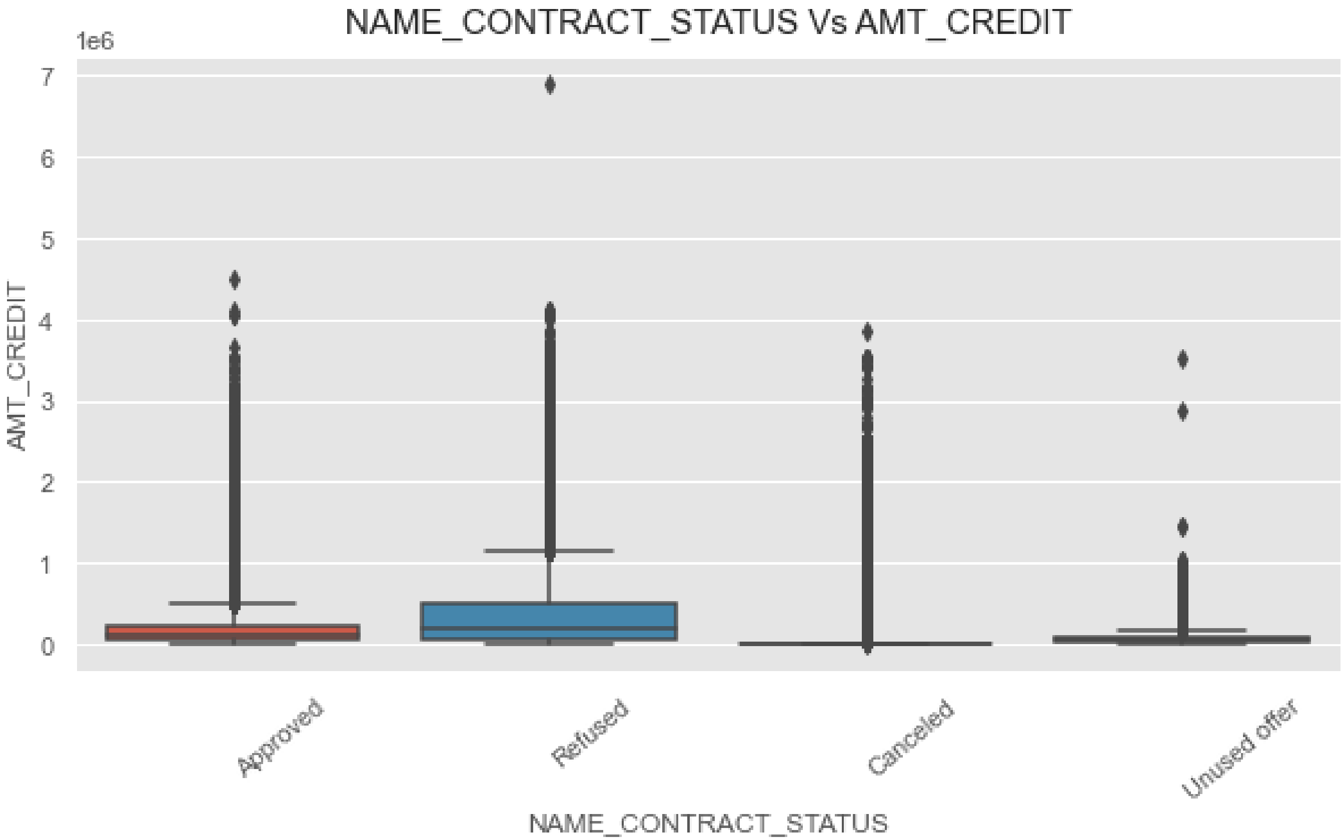
2. Refusal of loan increases eventually where the AMT_ANNUITY is increasing

3. Lesser amount gets approval the quickest

bivarient analysis of Contract status and Final credit amount disbursed to the customer previously, after approval

Interpretation:

when the AMT_CREDIT is too low, it get's cancelled/unused most of the time.



Conclusion

- Data of Application Data is extremely imbalance with the percentage of 11.39% as defaulters are extremely less
- Women have applied more for loans than Men
- People applying for a loan for repairs have the highest defaults
- There are more defaulters between the age of 25 to 40 yrs
- Business Entity Type 3, Self-employed, Other ,Medicine, Government, Business Entity Type 2 applied the most for the loan as compared to others
- Cash loans applications are higher than Revolving loans for both defaulters and non defaulters
- Clients who applied for loans were getting income by Working, Commercial associate and Pensioner are more likely to apply for the loan, highest being the Working class category
- Businessman, students and Unemployed less likely to apply for loan
- Working category have high risk to default

Conclusion. cntd

- State Servant is at Minimal risk to default
- Pensioner being highest followed by laborers have high risk to default
- Clients applying for high and low credit are at high risk of default
- Clients having low and medium income are at high risk to default
- Clients with secondary education are at high risk to default
- Female clients with an Academic degree and high-income type have a higher risk of default
- Male clients with Secondary/Secondary Special Education having all types of salaries have a higher risk of default.
- Male clients with Incomplete Education having very low salaries have a high risk of default.
- Male Clients with Lower Secondary Education having very low or medium have a high risk to default
- Bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment.