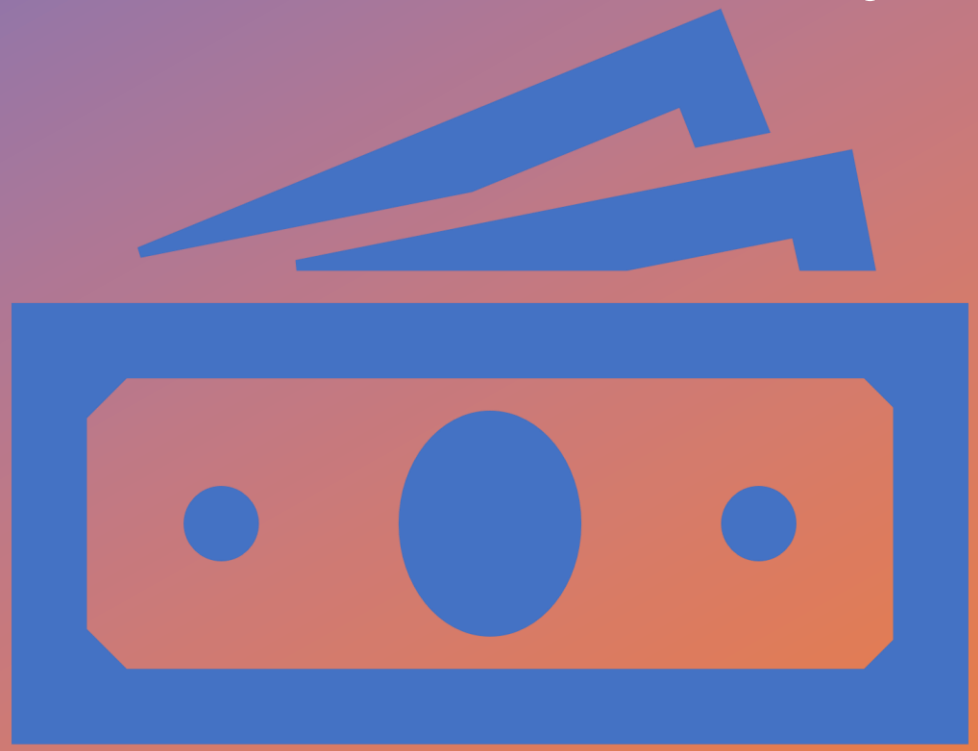
Credit EDA Case Study



-Submitted by Rajat Jain & Abhilasha Ranjan

1. DATA UNDERSTANDING

- There are primarily two data sets given. Namely the Application dataset & the Previous Application dataset.
- The Application dataset has 307511 rows & 122 columns(307511x122).
- The Previous Application dataset has 1670214 rows & 37 columns (1670214x37).
- TARGET column- it is the target variable for application dataset and consists of two values 0 & 1 and define the customers with payment difficulties.
- NAME_CONTRACT_STATUS - it is the target variable for previous application dataset and consists of four values namely, “Approved”, “Cancelled”, “Refused”, “Unused offer” and it defines the status of loan contract as per previous applications.
- The Application dataset and Previous Application dataset have a common key column; i.e, SK_ID_CURR.

2. DATA CLEANING

A. Null Values Handling

- Application dataset has 67 columns with null values and out of which 49 columns have more than 40% null values across the dataset.
- Previous application dataset has 16 columns with null values and out of which 11 columns have more than 40% null values across the dataset.
- One column named as NFLAG_MACRO_CASH is present in column description.csv file but missing in the actual Previous application dataset.

B. Impute and Remove Missing Values

- Dropped the columns in application dataset & previous application dataset having more than 40% null values.
- Column NAME_TYPE_SUITE in application dataset had fewer null values(0.42%), so replaced null values with the mode value of this column, i.e. ‘Unaccompanied’.
- Column OCCUPATION_TYPE in application dataset had 96391 null values(31%), It can be risky to replace null values with the mode or some other values. Hence, filled missing values with the value ‘Missing’.
- There were some numerical columns which were having null values, so replaced these null values with median of column values.
- Fixed the invalid values(365243) in DAYS_EMPLOYED column in application dataset.

C. Fixation of Column Datatypes

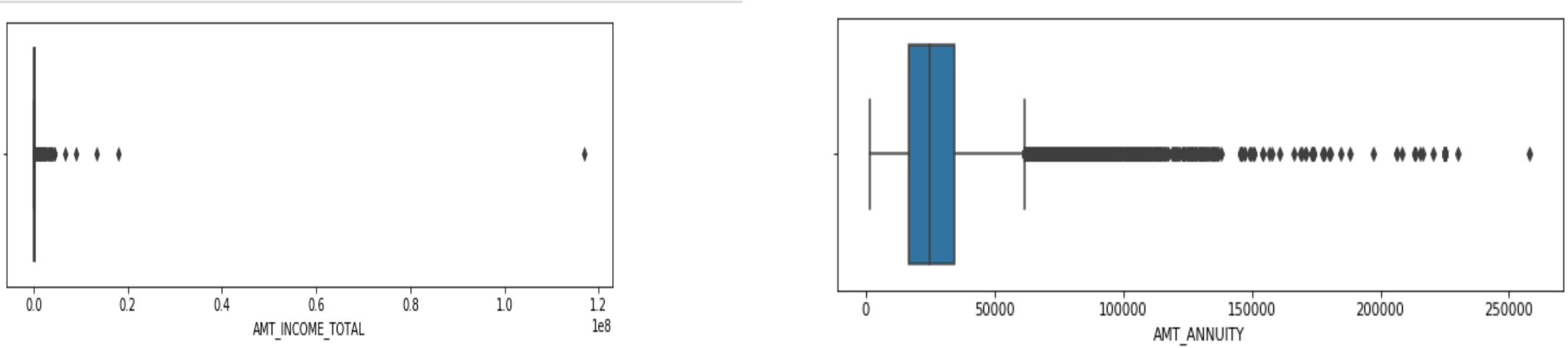
```
In [20]: app_data.select_dtypes('float').columns

Out[20]: Index(['AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE',
                'REGION_POPULATION_RELATIVE', 'DAYS_REGISTRATION', 'OWN_CAR_AGE',
                'CNT_FAM_MEMBERS', 'EXT_SOURCE_1', 'EXT_SOURCE_2', 'EXT_SOURCE_3',
                'APARTMENTS_AVG', 'BASEMENTAREA_AVG', 'YEARS_BEGINEXPLUATATION_AVG',
                'YEARS_BUILD_AVG', 'COMMONAREA_AVG', 'ELEVATORS_AVG', 'ENTRANCES_AVG',
                'FLOORSMAX_AVG', 'FLOORSMIN_AVG', 'LANDAREA_AVG',
                'LIVINGAPARTMENTS_AVG', 'LIVINGAREA_AVG', 'NONLIVINGAPARTMENTS_AVG',
                'NONLIVINGAREA_AVG', 'APARTMENTS_MODE', 'BASEMENTAREA_MODE',
                'YEARS_BEGINEXPLUATATION_MODE', 'YEARS_BUILD_MODE', 'COMMONAREA_MODE',
                'ELEVATORS_MODE', 'ENTRANCES_MODE', 'FLOORSMAX_MODE', 'FLOORSMIN_MODE',
                'LANDAREA_MODE', 'LIVINGAPARTMENTS_MODE', 'LIVINGAREA_MODE',
                'NONLIVINGAPARTMENTS_MODE', 'NONLIVINGAREA_MODE', 'APARTMENTS_MEDI',
                'BASEMENTAREA_MEDI', 'YEARS_BEGINEXPLUATATION_MEDI', 'YEARS_BUILD_MEDI',
                'COMMONAREA_MEDI', 'ELEVATORS_MEDI', 'ENTRANCES_MEDI', 'FLOORSMAX_MEDI',
                'FLOORSMIN_MEDI', 'LANDAREA_MEDI', 'LIVINGAPARTMENTS_MEDI',
                'LIVINGAREA_MEDI', 'NONLIVINGAPARTMENTS_MEDI', 'NONLIVINGAREA_MEDI',
                'TOTALAREA_MODE', 'OBS_30_CNT_SOCIAL_CIRCLE',
                'DEF_30_CNT_SOCIAL_CIRCLE', 'OBS_60_CNT_SOCIAL_CIRCLE',
                'DEF_60_CNT_SOCIAL_CIRCLE', 'DAYS_LAST_PHONE_CHANGE',
                'AMT_REQ_CREDIT_BUREAU_HOUR', 'AMT_REQ_CREDIT_BUREAU_DAY',
                'AMT_REQ_CREDIT_BUREAU_WEEK', 'AMT_REQ_CREDIT_BUREAU_MON',
                'AMT_REQ_CREDIT_BUREAU_QRT', 'AMT_REQ_CREDIT_BUREAU_YEAR'],
               dtype='object')
```

- There were columns in application dataset whose datatype was FLOAT but as per column definition those should be in the integers, so converted it into an INT.

D. Handling Outliers

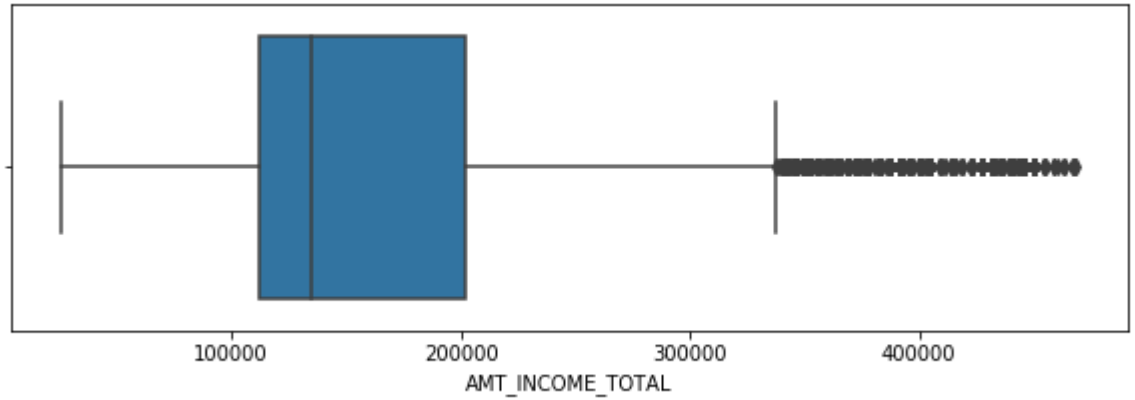
- Considering below four columns for outliers handling as from the data description these four columns shows some outliers.

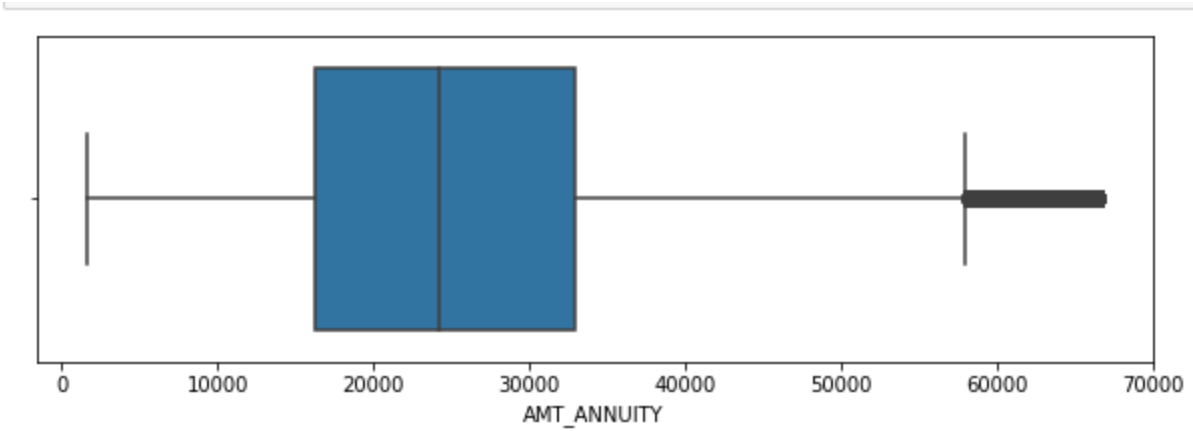
AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE



Box plot of AMT_INCOME_TOTAL & AMT_ANNUITY before handling outliers

Box Plot visualization





Box plot of AMT_INCOME_TOTAL & AMT_ANNUITY after handling outliers

- Considered the top 99%tile of the values and excluded remaining 1% outliers to handle the outliers

E. Binning of Attribute Values

Age variable

```
In [259]: app_data['AGE_GROUP']= pd.cut(app_data['AGE'],bins=np.linspace(20 ,70,11))
          app_data['AGE_GROUP'].value_counts()
```

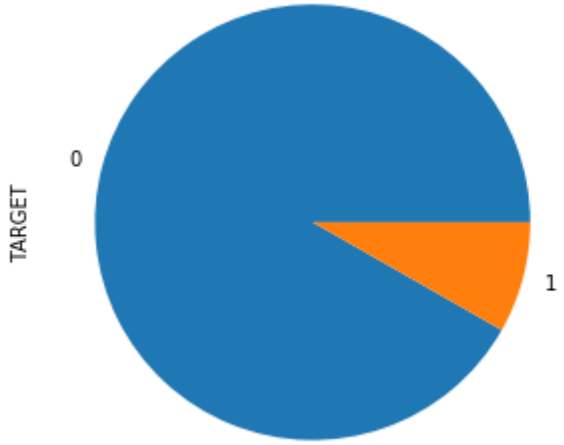
```
Out[259]: (35.0, 40.0]    41484
          (40.0, 45.0]    37826
          (30.0, 35.0]    37724
          (25.0, 30.0]    35354
          (50.0, 55.0]    33406
          (45.0, 50.0]    32639
          (55.0, 60.0]    31345
          (60.0, 65.0]    23637
          (20.0, 25.0]    16047
          (65.0, 70.0]     5006
          Name: AGE_GROUP, dtype: int64
```

```
In [264]: app_data.loc[:,'INCOME_RANGE'].value_counts()
```

```
Out[264]: High         125448
          Low           66984
          Medium        48373
          Very_low      33348
          Very_high     20316
          Name: INCOME_RANGE, dtype: int64
```

- Considered some of the attributes for binning as it will be useful for further analysis of these attributes.
- AGE, INCOME, RATINGS , these are some of the variables chosen for binning of values.

3. UNIVARIATE ANALYSIS

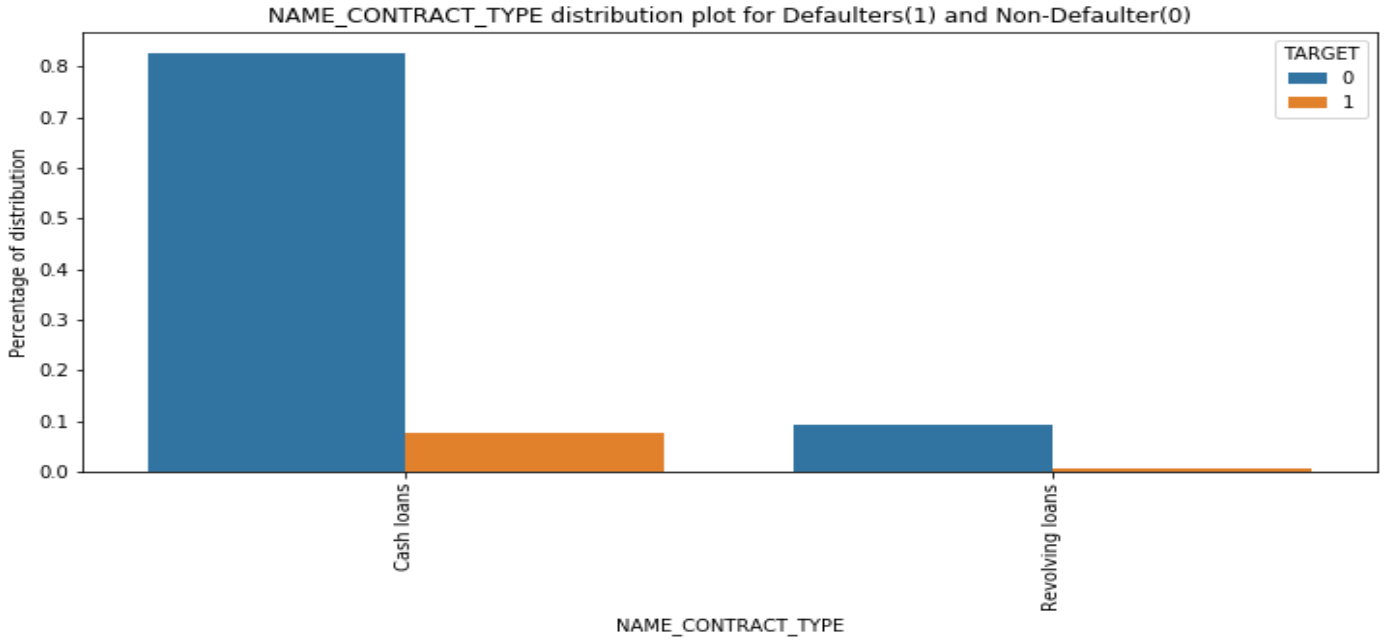


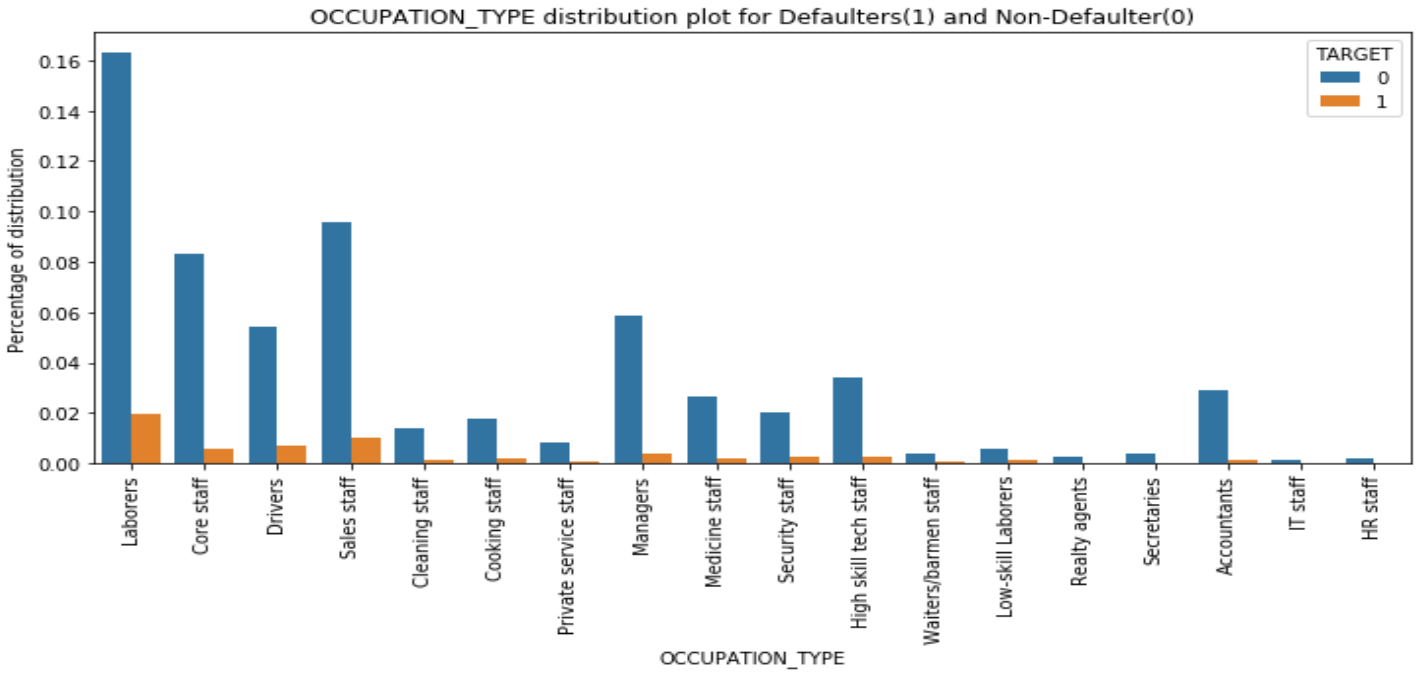
Target varaible of application dataset

```
In [268]: app_data.TARGET.value_counts(normalize=True)

Out[268]: 0    0.91757
          1    0.08243
          Name: TARGET, dtype: float64
```

- From the derived Pie chart we can infer that in application dataset 92 % of data is for the Non-Defaulters and around 8 % data is for Defaulters category.

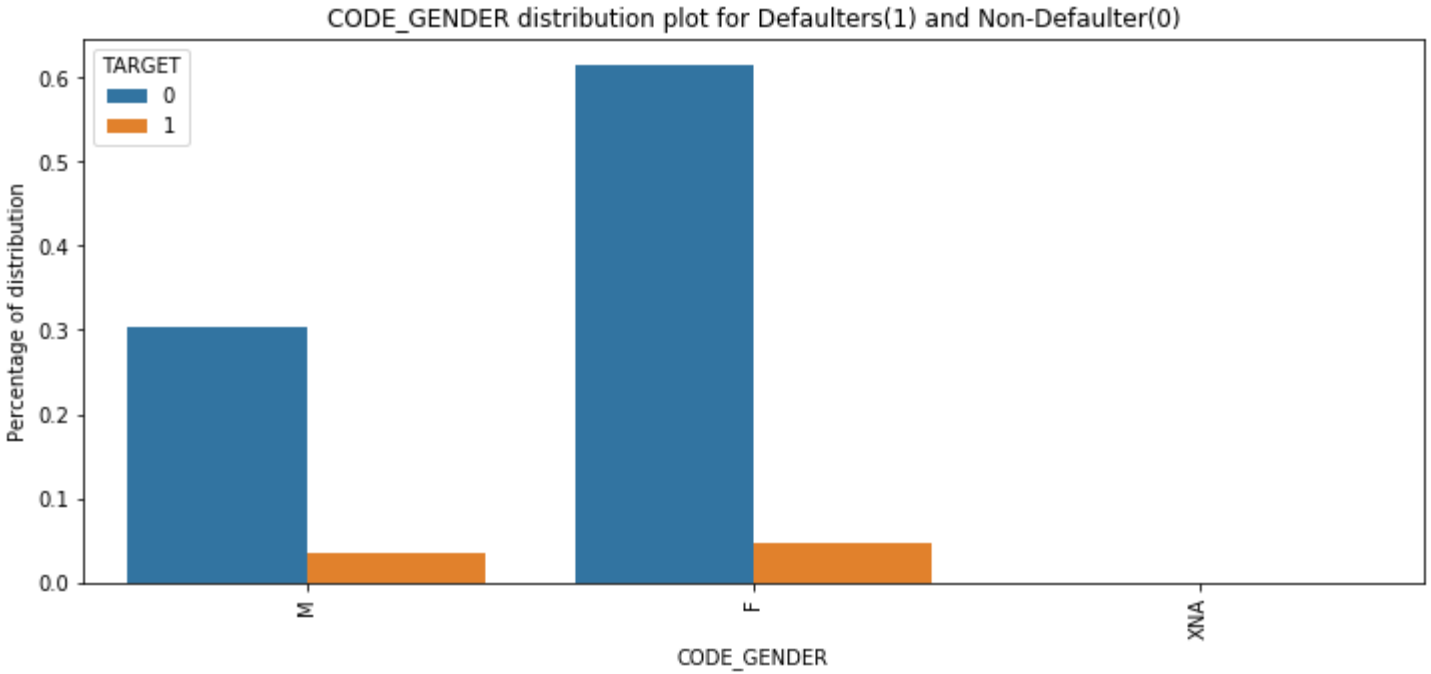
- This makes the data imbalanced in the application dataset but in this real time scenario getting the balanced data is not possible. Hence, performing further analysis based upon the given application dataset.

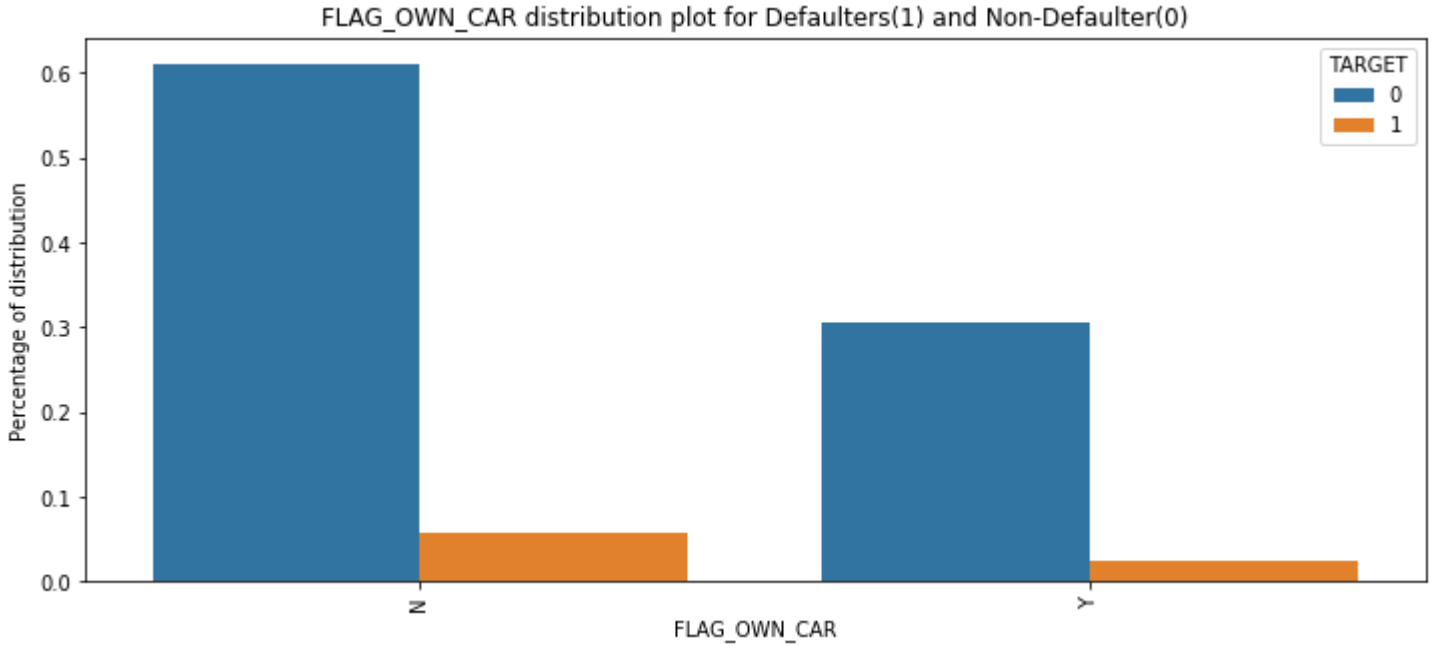




Insights :

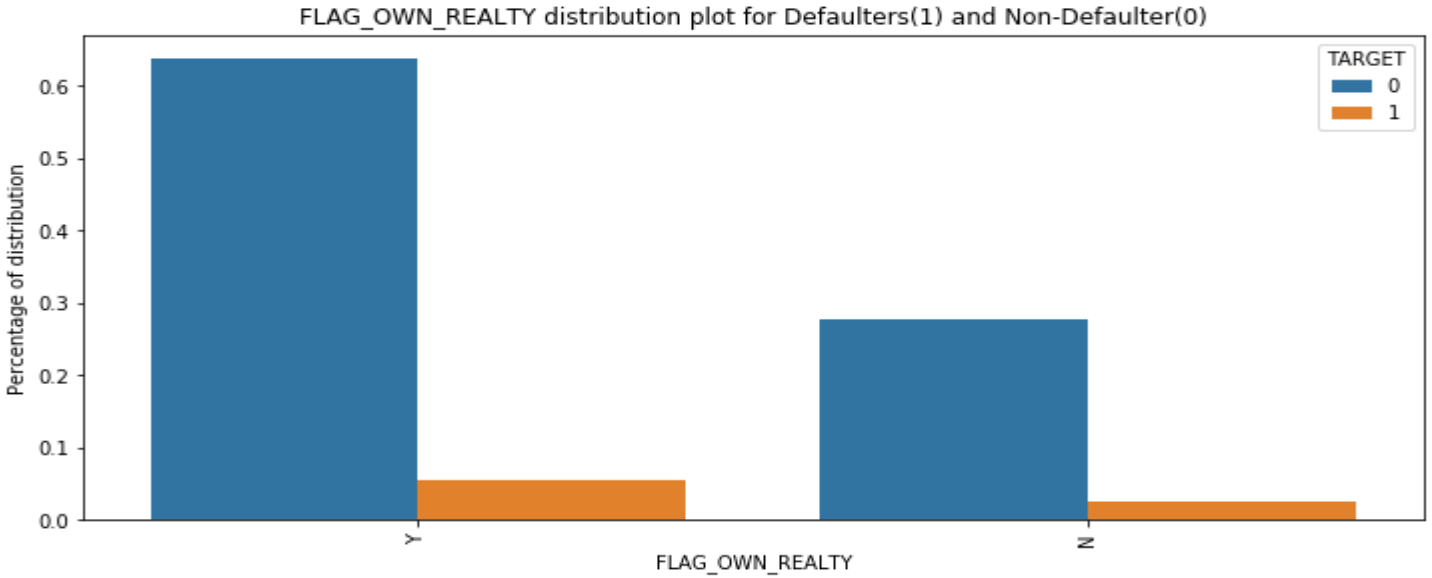
- Providing Revolving loan is safer than Cash loan as there are more proportion of defaulters in cash loan.
- Laborer, Driver and sales staff are likely to be defaulter in the OCCUPATION TYPE category.

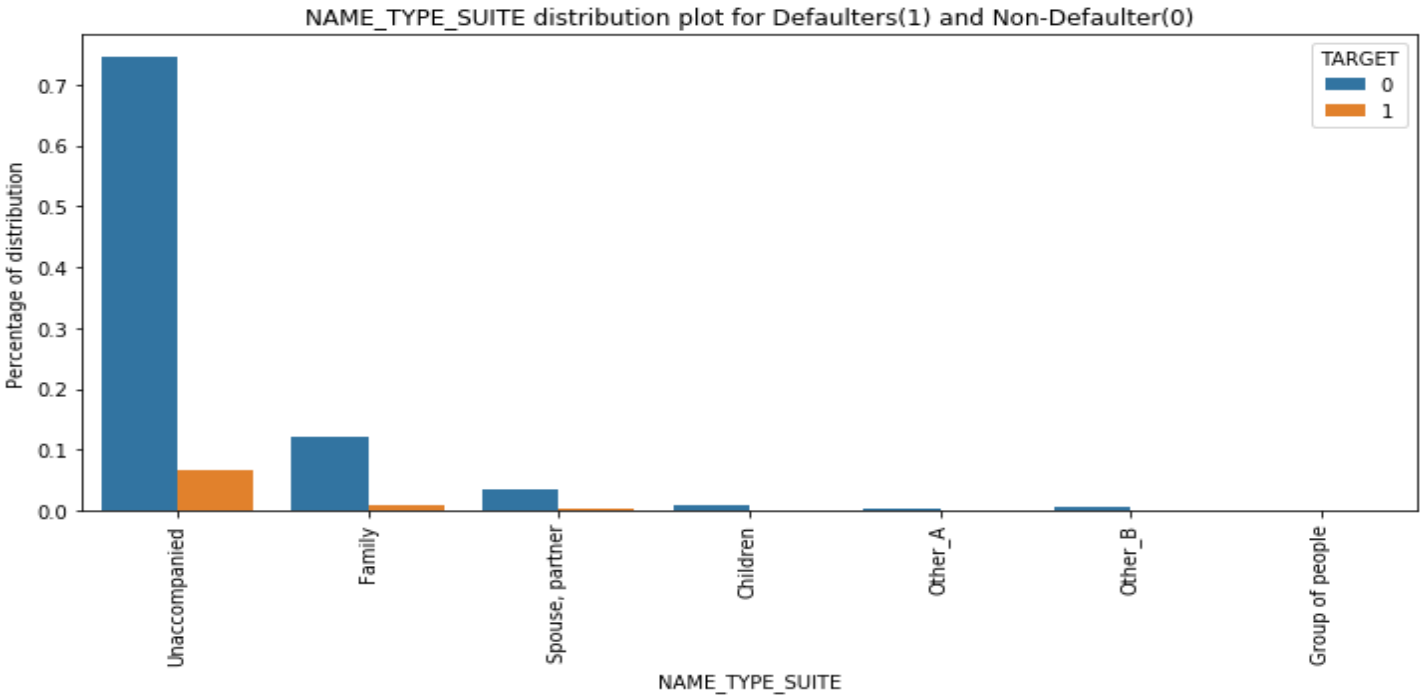




Insights :

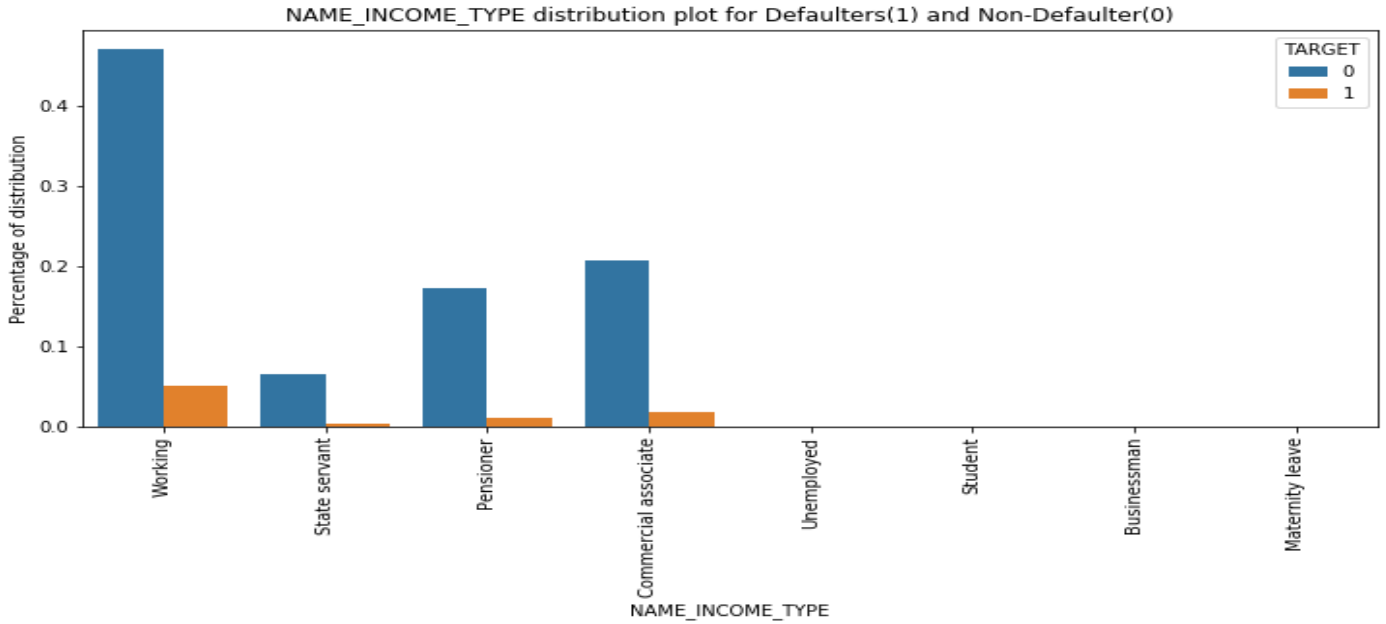
- Male proportion for being defaulter is higher than the Female one. It's safe to give loan to Female.
- Customer not owing the car are more likely to be default.

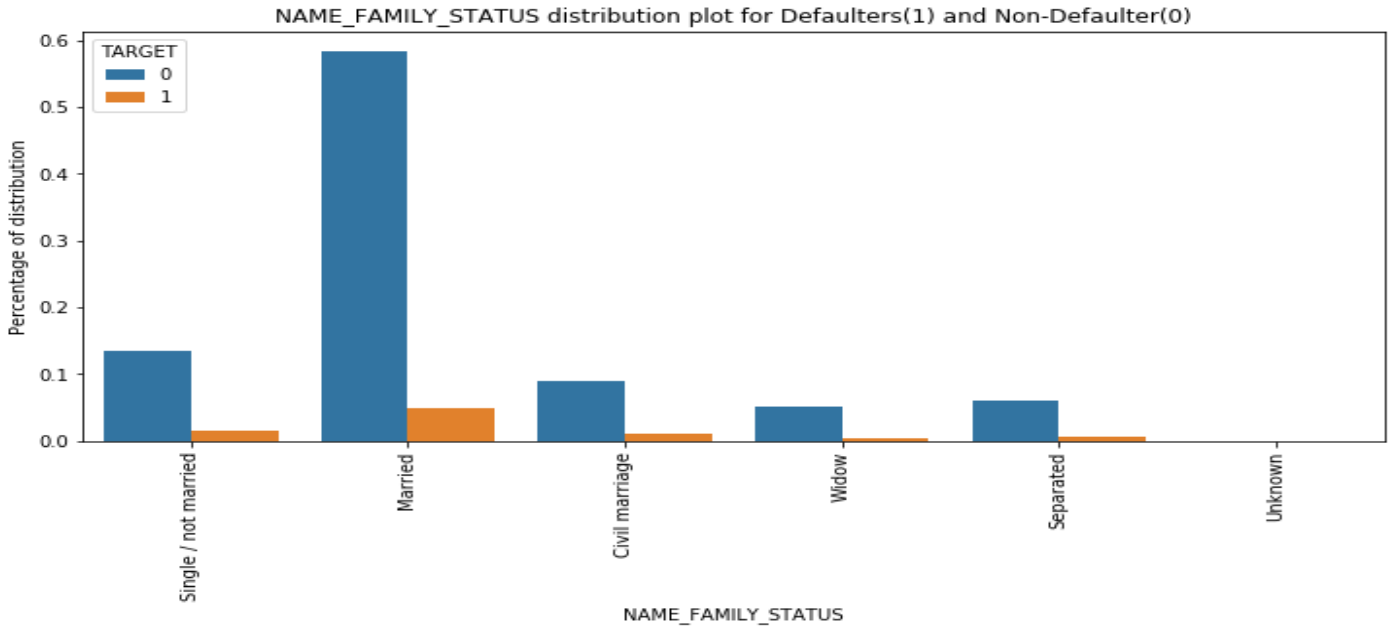




Insights :

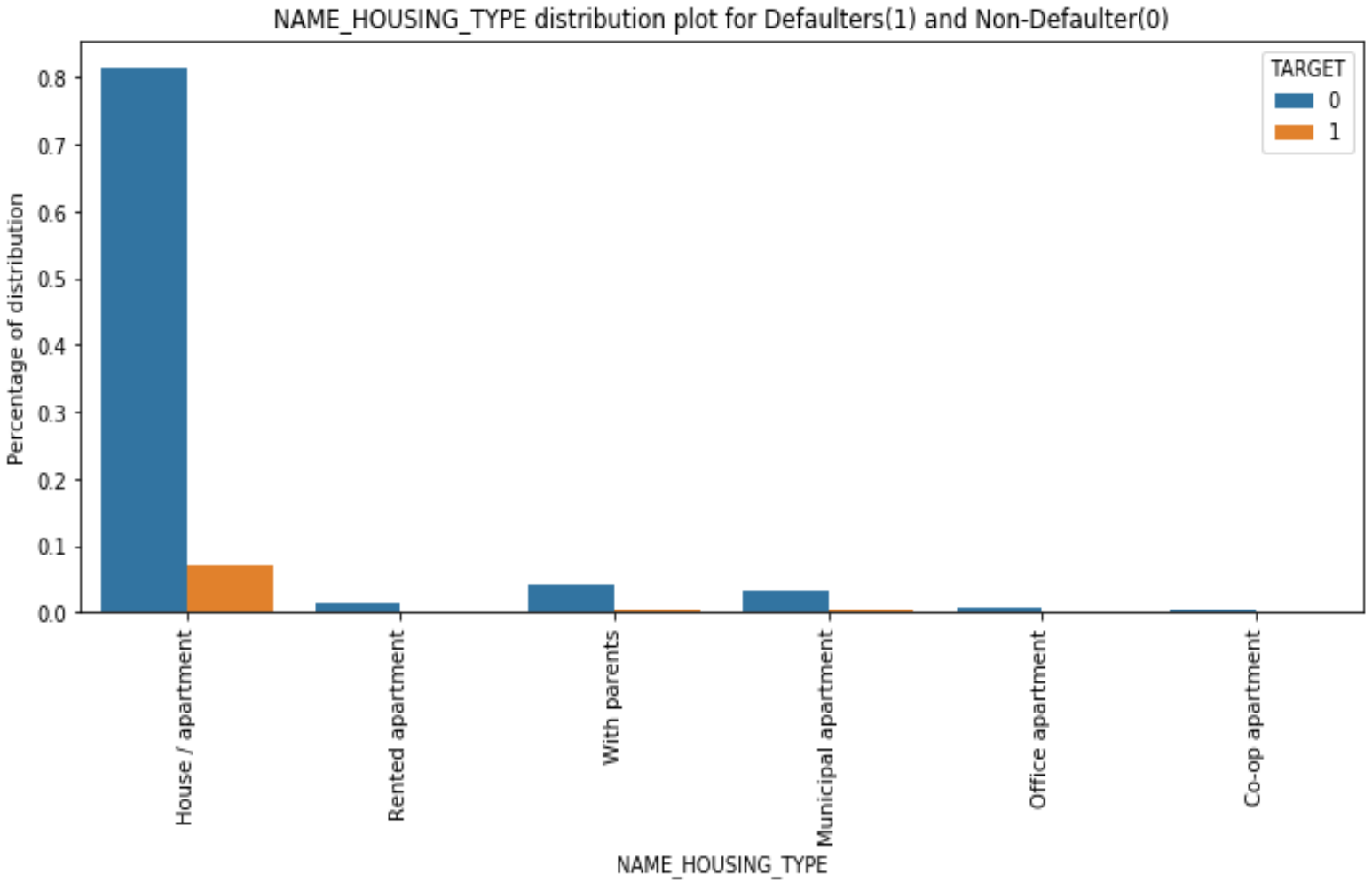
- Customer owing Real state are safer to provide the loan as customer not owing real state are likely to be defaulter.
- Unaccompanied category of the NAME_TYPE_SUIT attribute is highly likely to be defaulter than the other.





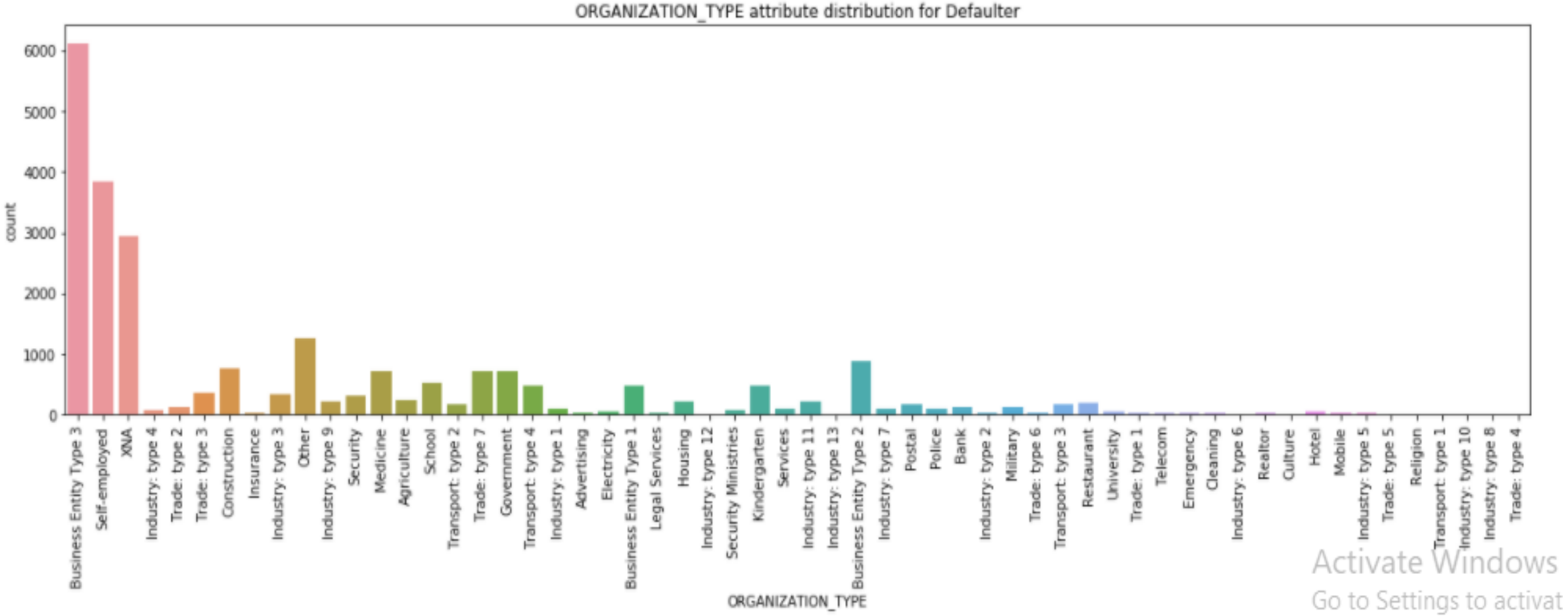
Insights :

- Working professional are likely to be defaulter than others but their proportion in dataset is also high.
- Married person are more likely to be defaulter and single/not married are less likely to be defaulter.



Insights :

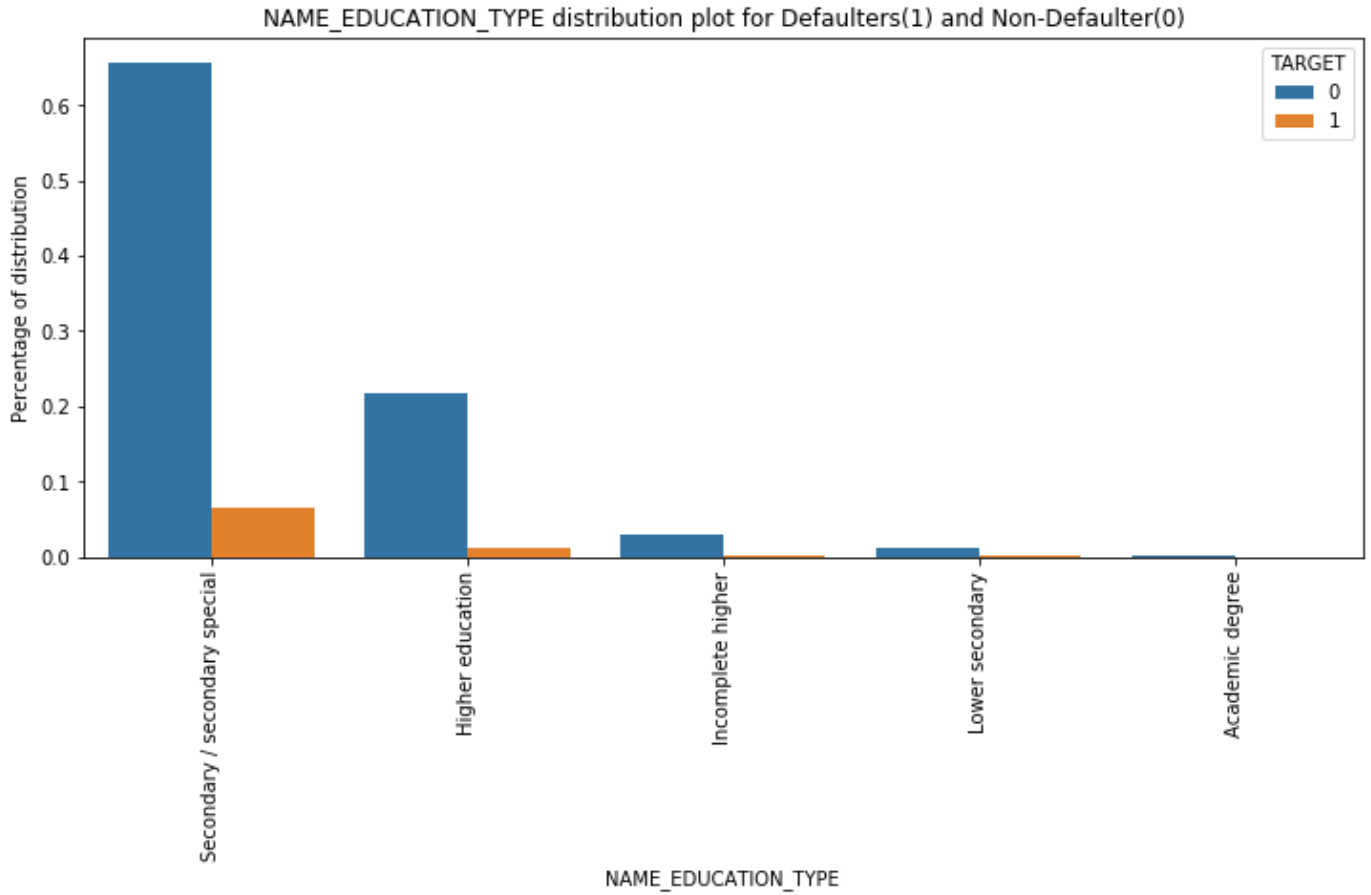
- There are around 80% of loan customer, having House/apartment as their housing type.
- And customer having house type as House/apartment are likely to be defaulter.



Insights :

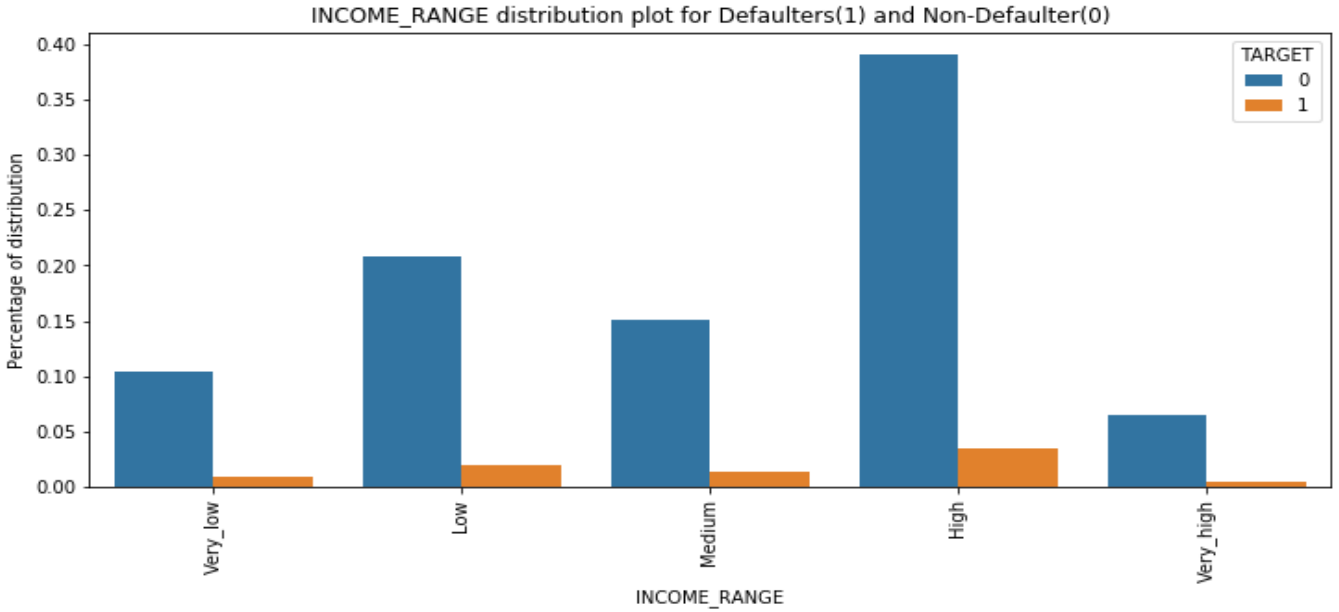
- The above plot showing the distribution of the defaulter with ORGANIZATION_TYPE.
- Business severity type 3 & Self-employed are likely to be defaulter.

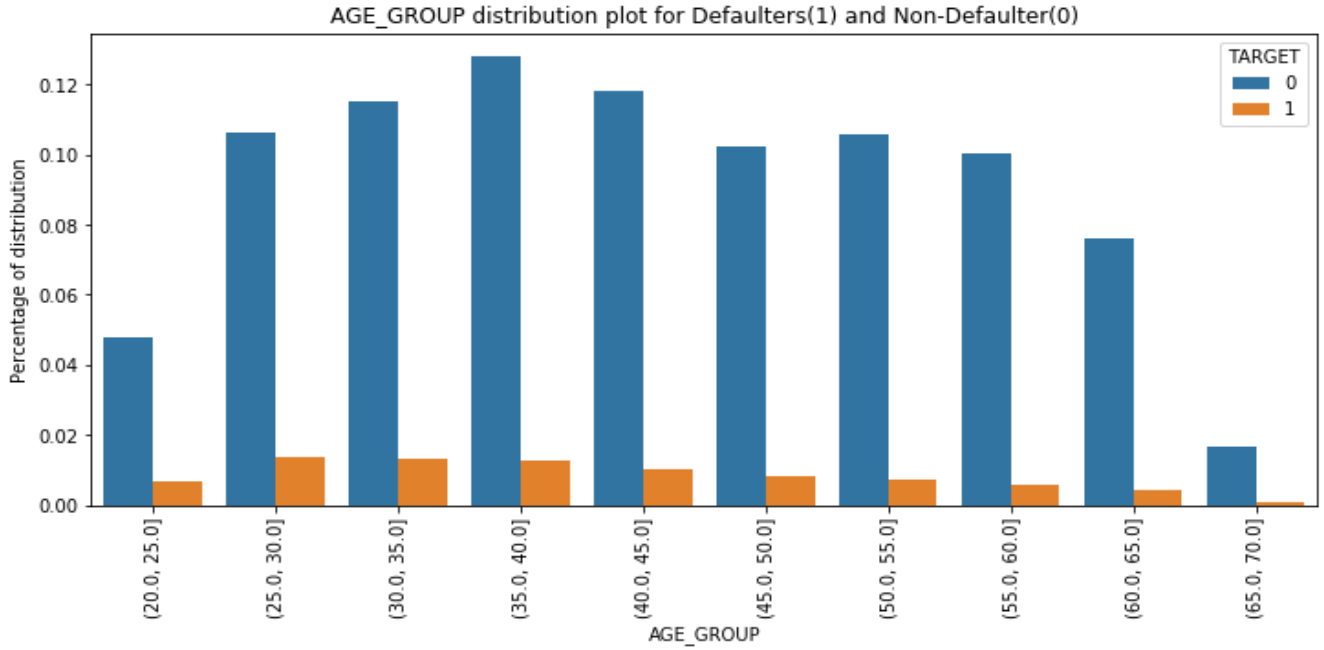
Ordinal categorical variable distribution around TARGET variable



Insights :

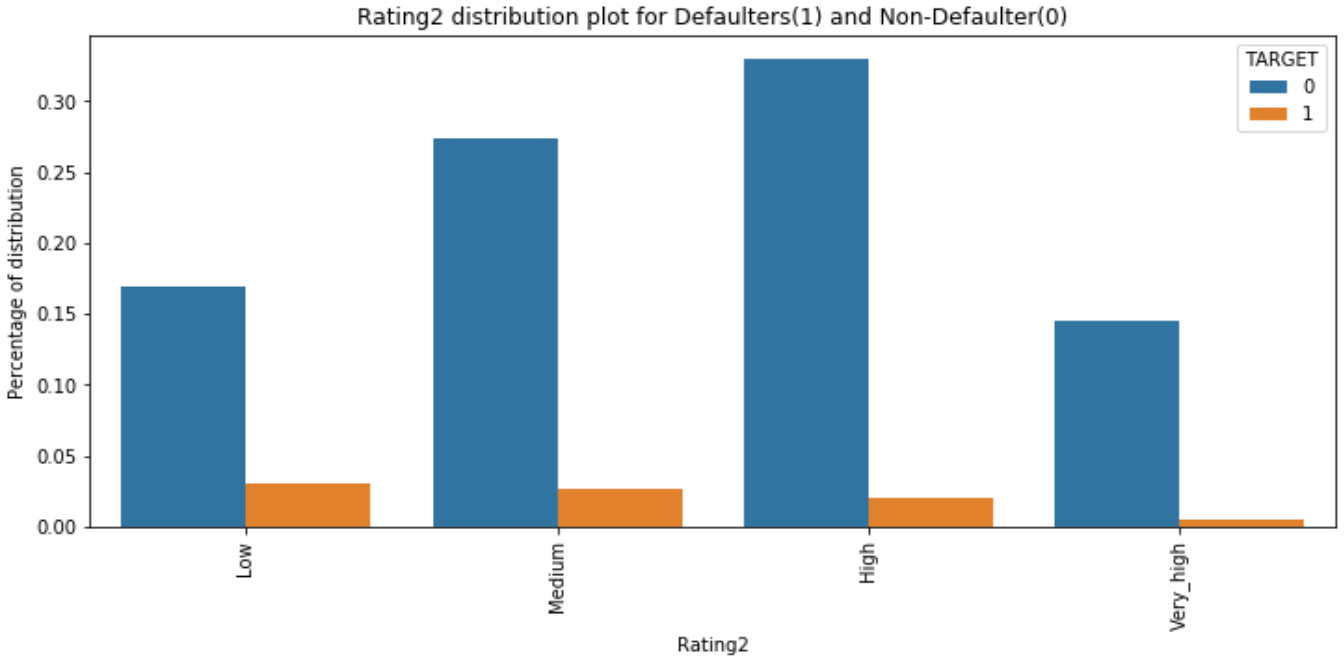
- Secondary/Secondary special are more likely to be defaulter as they are less educated and might not be earning well.

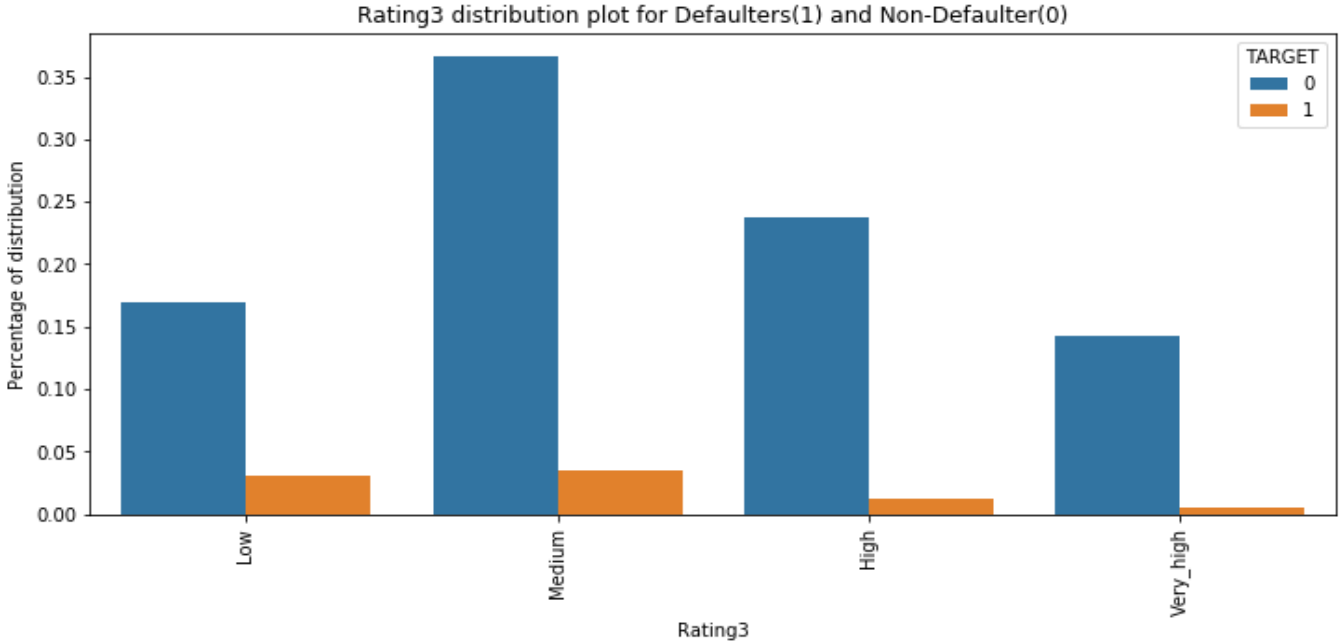




Insights :

- People with age more than 45 are lee likely to be defaulter and are safe for providing the loans.
- Very high and very low customer income range are less likely to be defaulter.





Insights :

- Customer with higher rating from external source 2 are less likely to be defaulter.
- Customer with higher rating from external source 3 are less likely to be defaulter and with rating as low are highly likely to be defaulter.

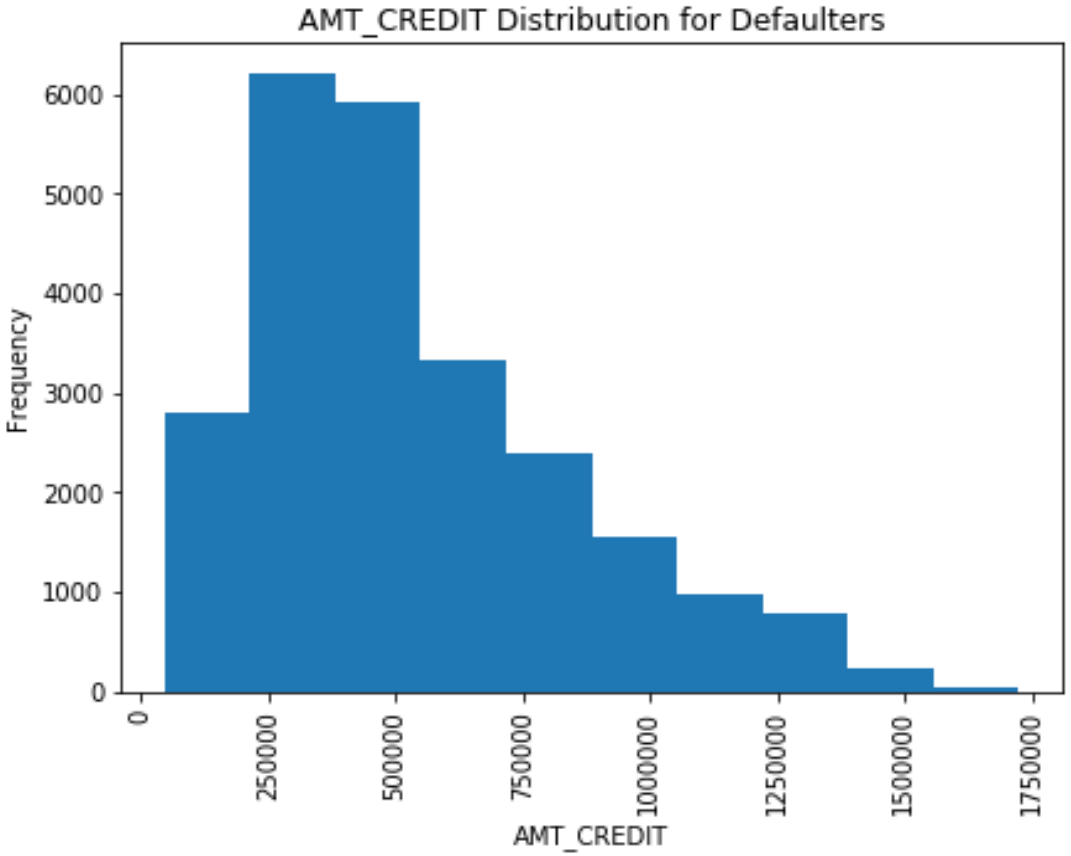
Top 10 Highly Correlated Attributes

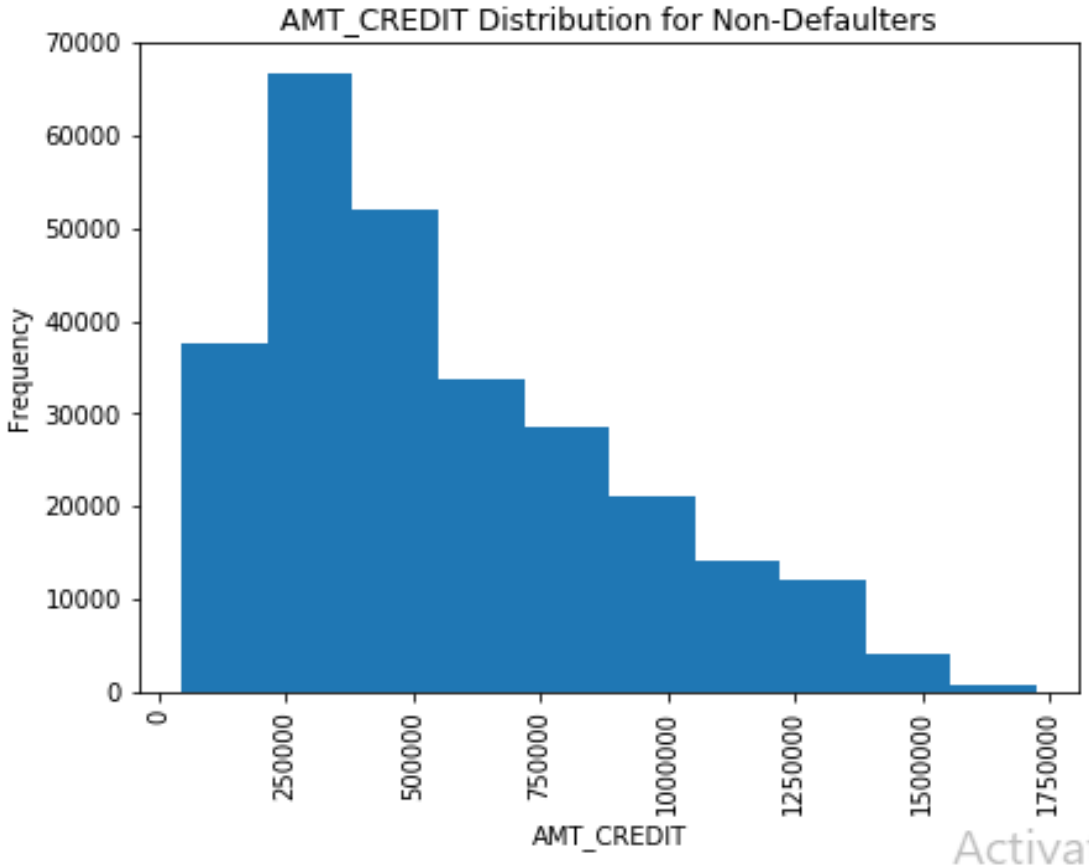
- **Positively Correlated Attributes**

AMT_REQ_CREDIT_BUREAU_YEAR	AMT_REQ_CREDIT_BUREAU_YEAR	1.00
AMT_GOODS_PRICE	AMT_CREDIT	0.98
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.96
CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.85
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.78
AMT_ANNUITY	AMT_GOODS_PRICE	0.74
DAYS_BIRTH	FLAG_EMP_PHONE	0.58
REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.50

- **Negatively Correlated Attributes**

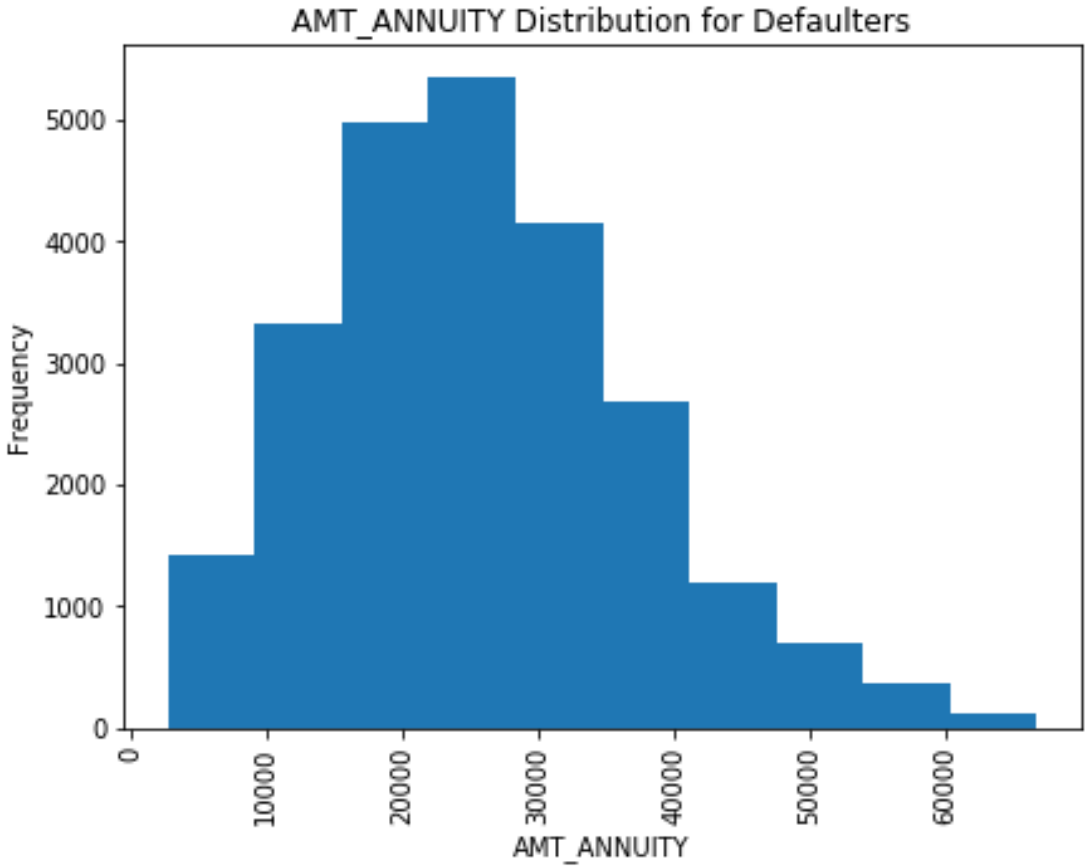
HOUR_APPR_PROCESS_START	REGION_RATING_CLIENT_W_CITY	-0.28
AGE	DAYS_REGISTRATION	-0.29
DAYS_EMPLOYED	FLAG_EMP_PHONE	-0.32
DAYS_BIRTH	FLAG_DOCUMENT_6	-0.39
REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT	-0.44
FLAG_DOCUMENT_3	FLAG_DOCUMENT_6	-0.48
FLAG_DOCUMENT_8	FLAG_DOCUMENT_3	-0.52
AGE	FLAG_EMP_PHONE	-0.58
FLAG_DOCUMENT_6	FLAG_EMP_PHONE	-0.62
Total_EXP	DAYS_EMPLOYED	-1.00

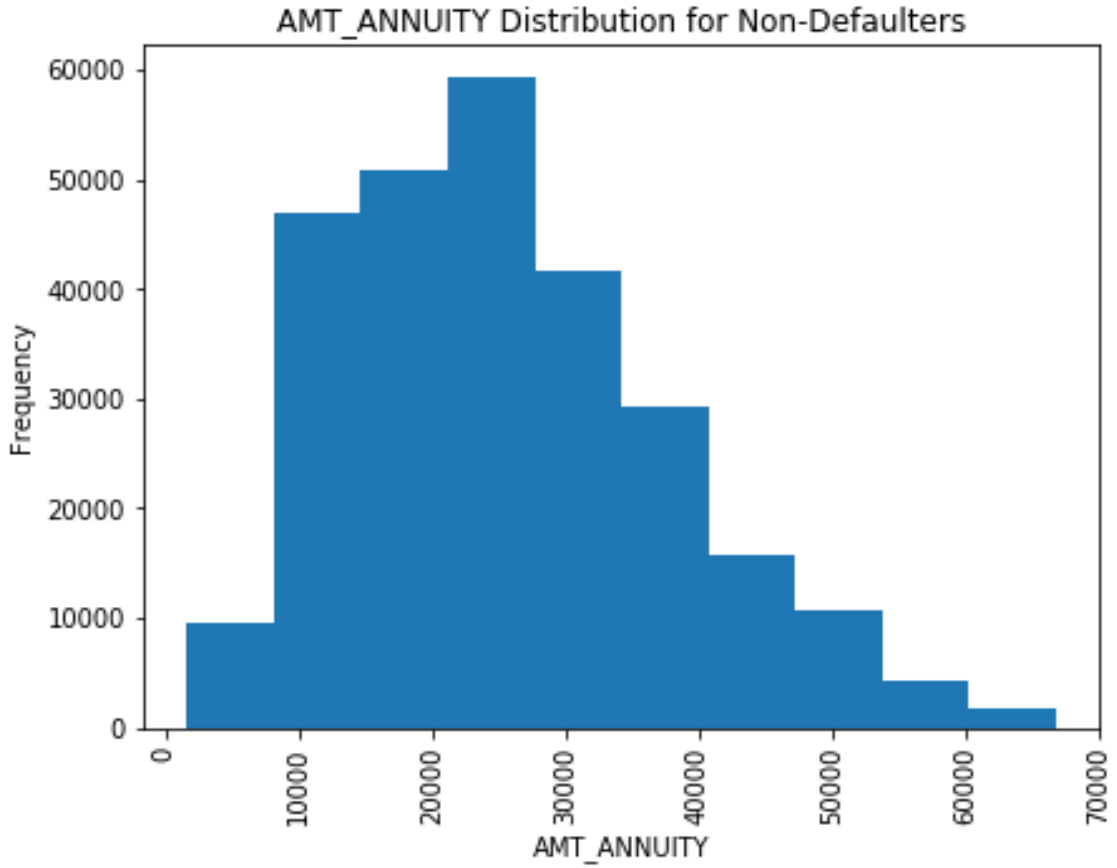




Insights :

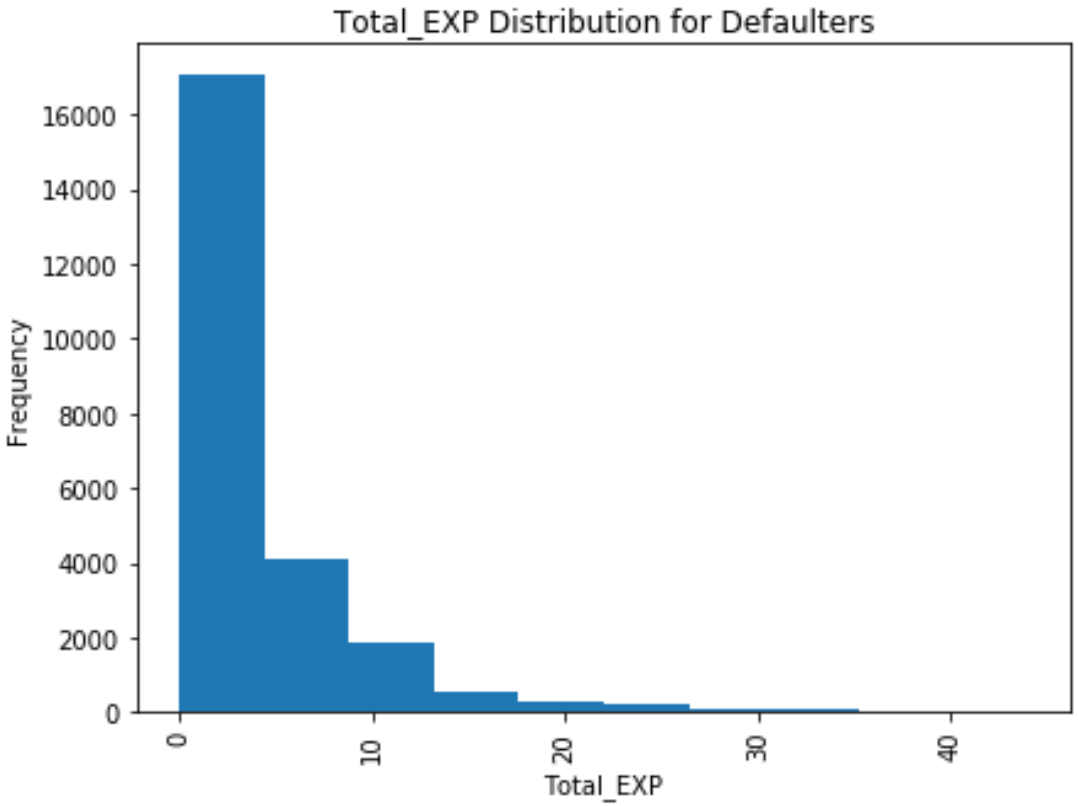
- Customer with amount credit in range from 3 lac to 5 lac are more likely to be defaulter as amount is in moderate range.

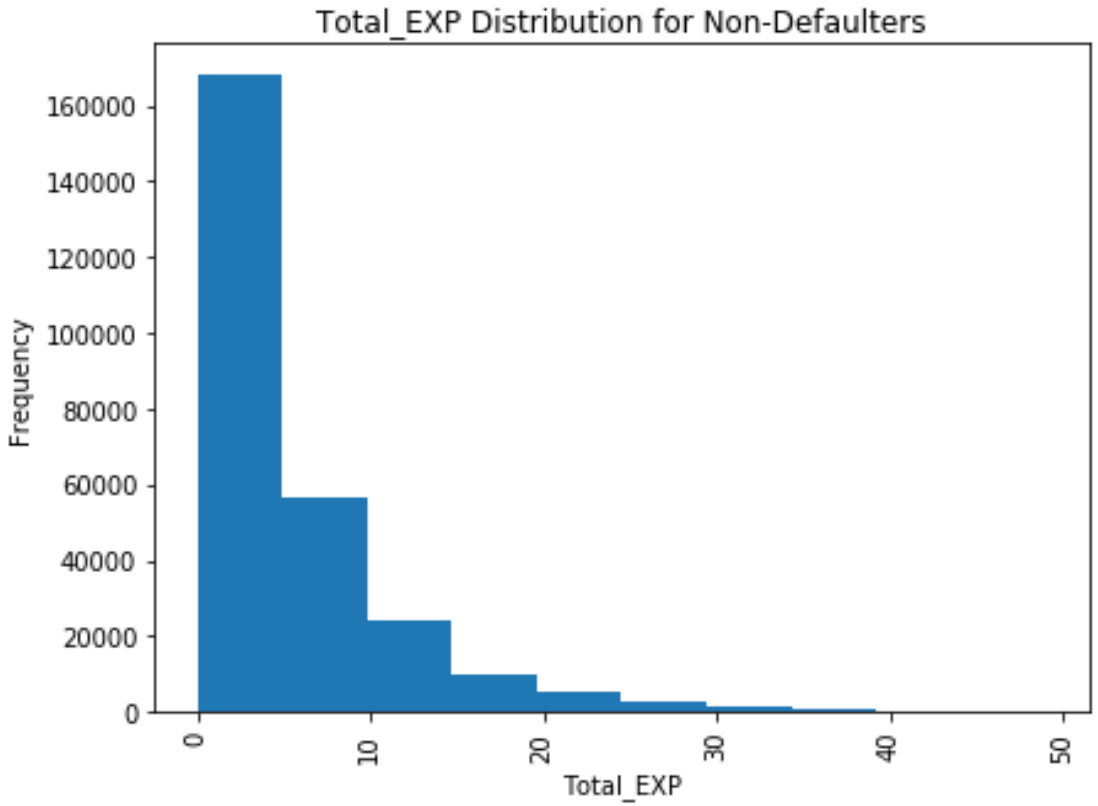




Insights :

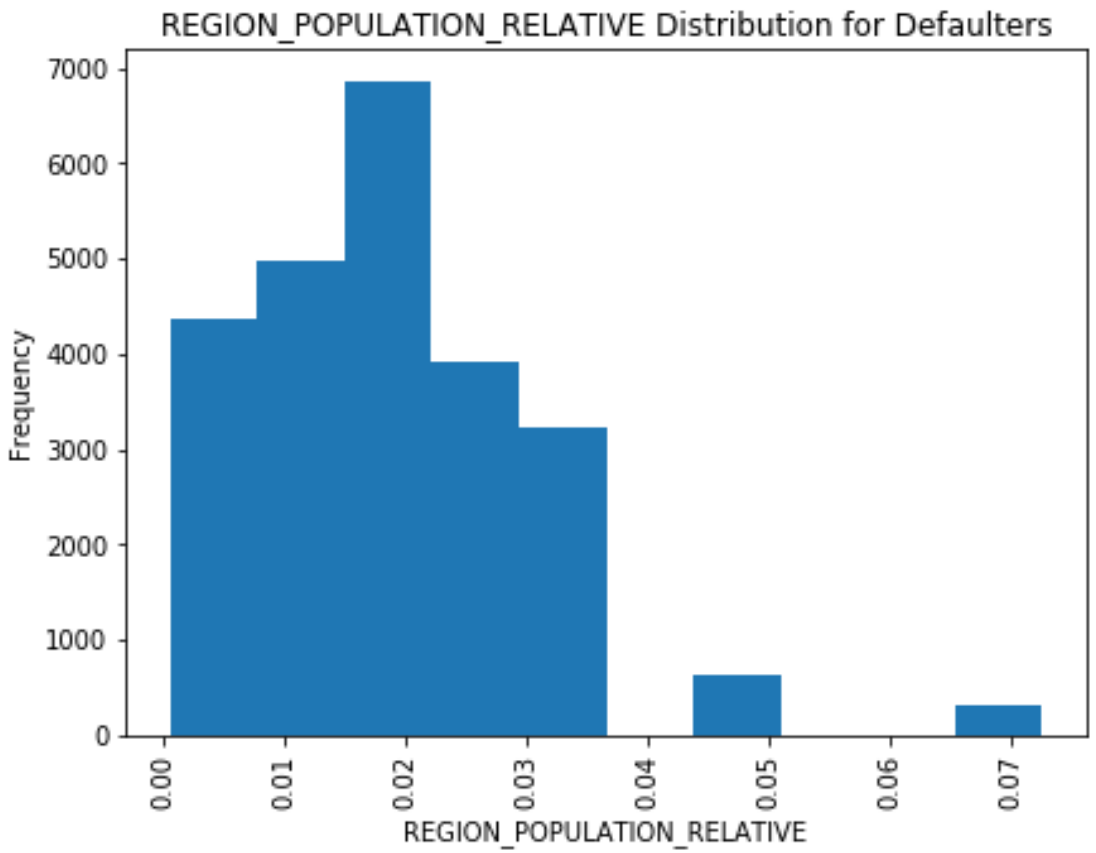
- Customer with amount annuity in range 1 lac to 2 lac are less likely to be defaulter.

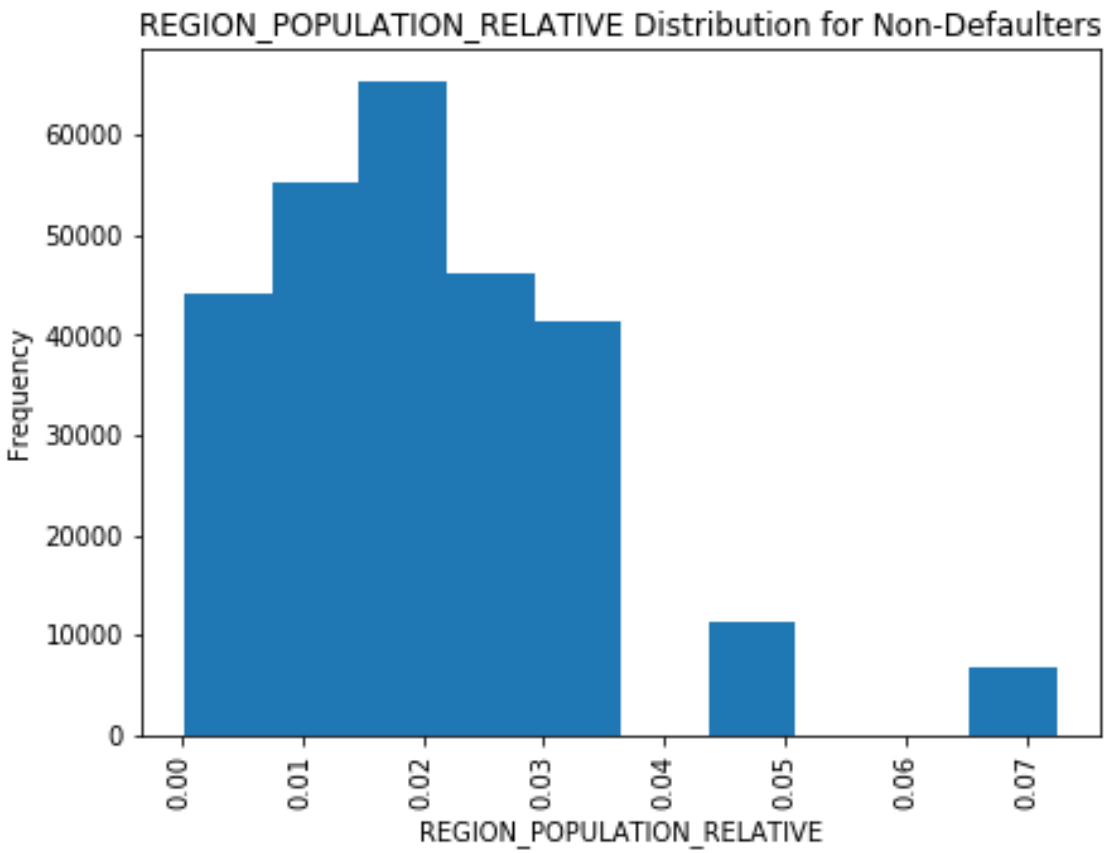




Insights :

- People with 5 or more years of experience are less likely to be defaulter than other.

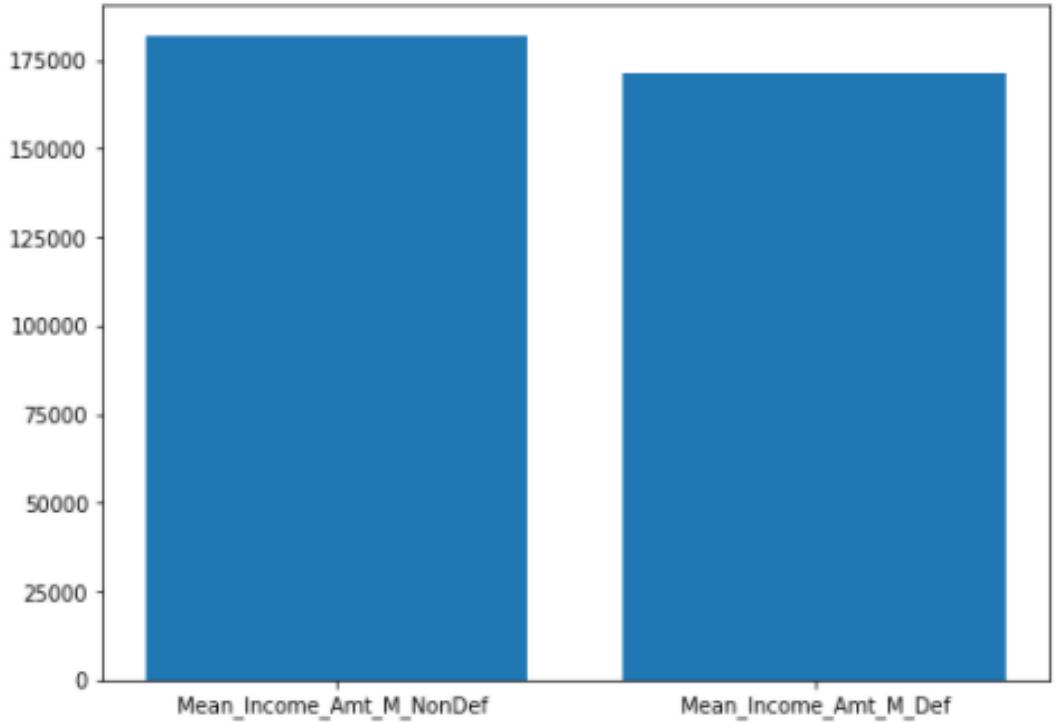


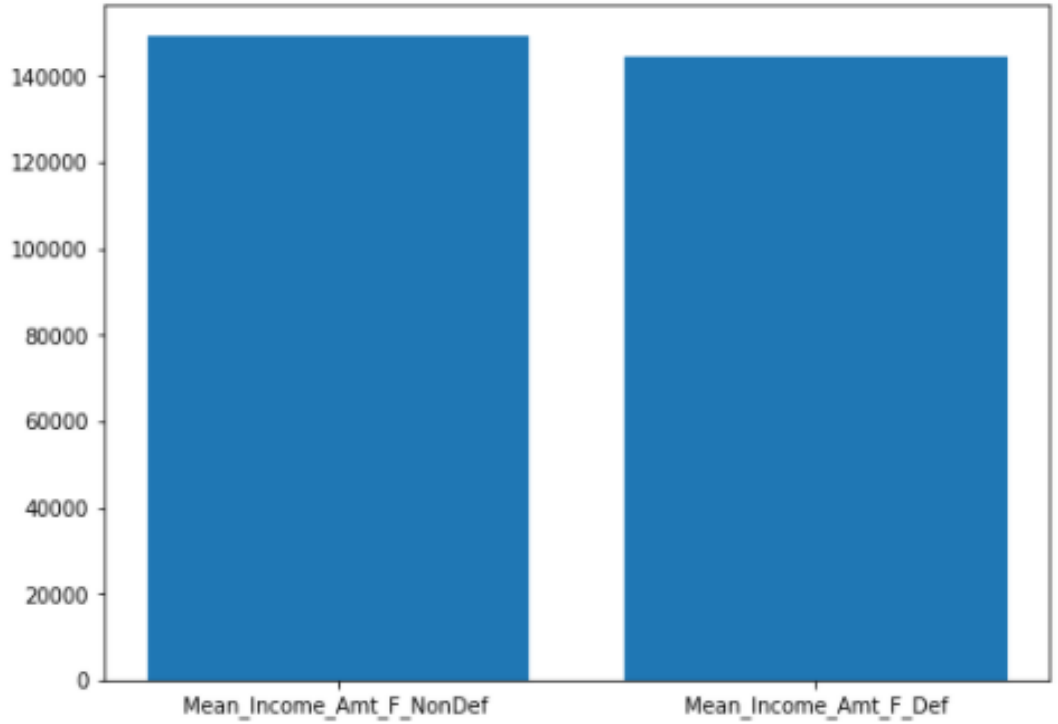


Insights :

- People with region population relative greater than 0.02 are less likely to be defaulter.

4. BIVARIATE ANALYSIS

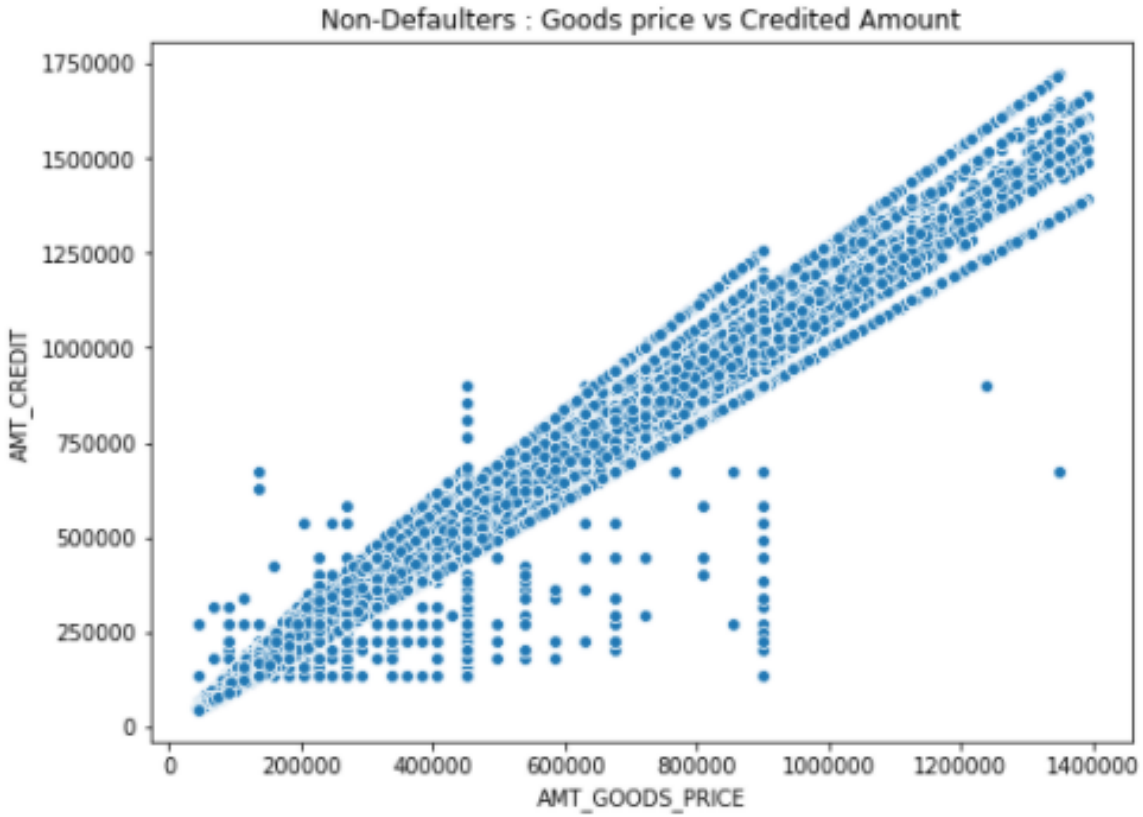




Insights :

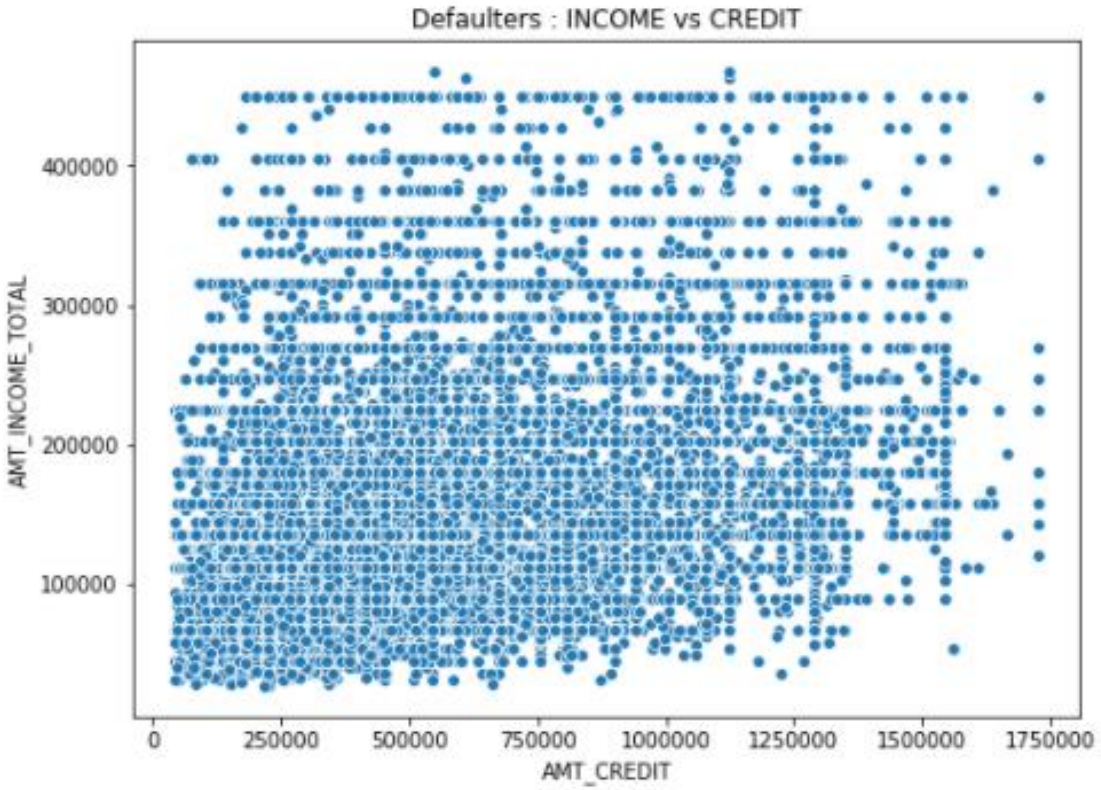
- Mean income of Male defaulter is lesser than the non-defaulter. That is the expected one as lesser income are higher chance of missing installment.

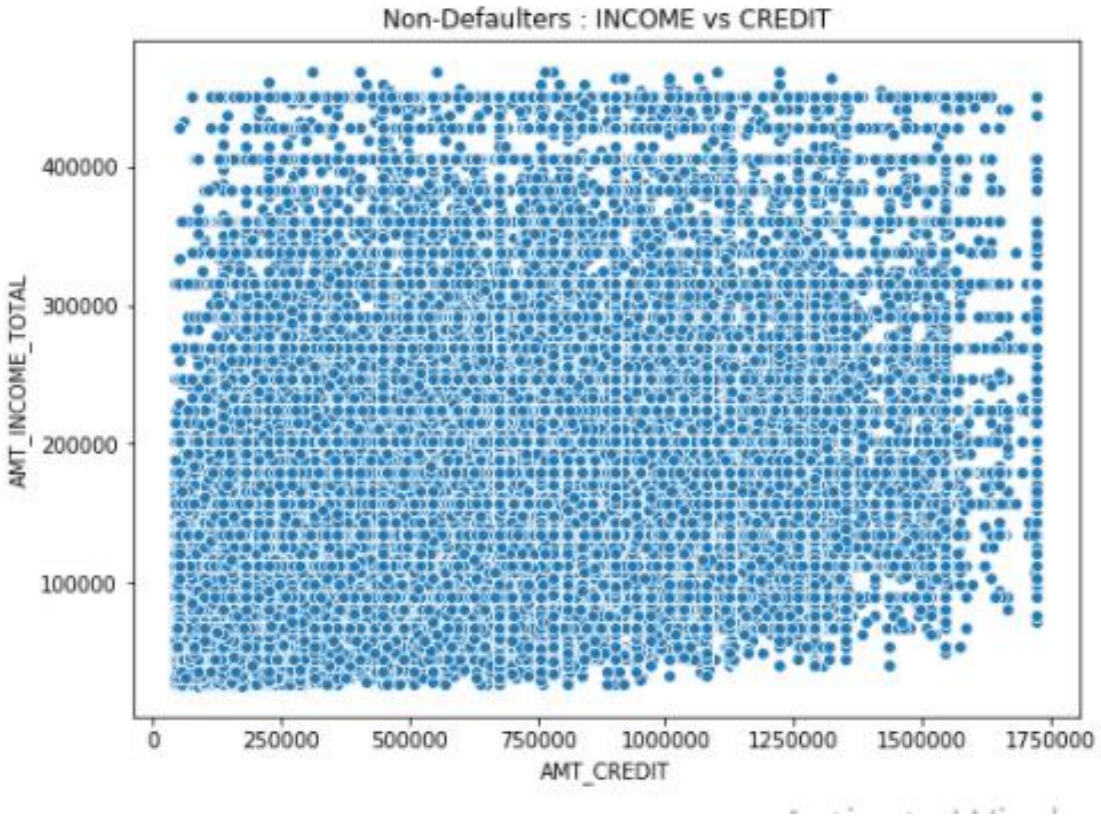
- Mean income of Male defaulter is lesser than the non-defaulter. That is the expected one as lesser income are higher chance of missing installment.



Insights :

- Customers with Amount goods price in between 10 thousand to 4 lac and amount credit upto 3 lac are less likely to be defaulter.





Insights :

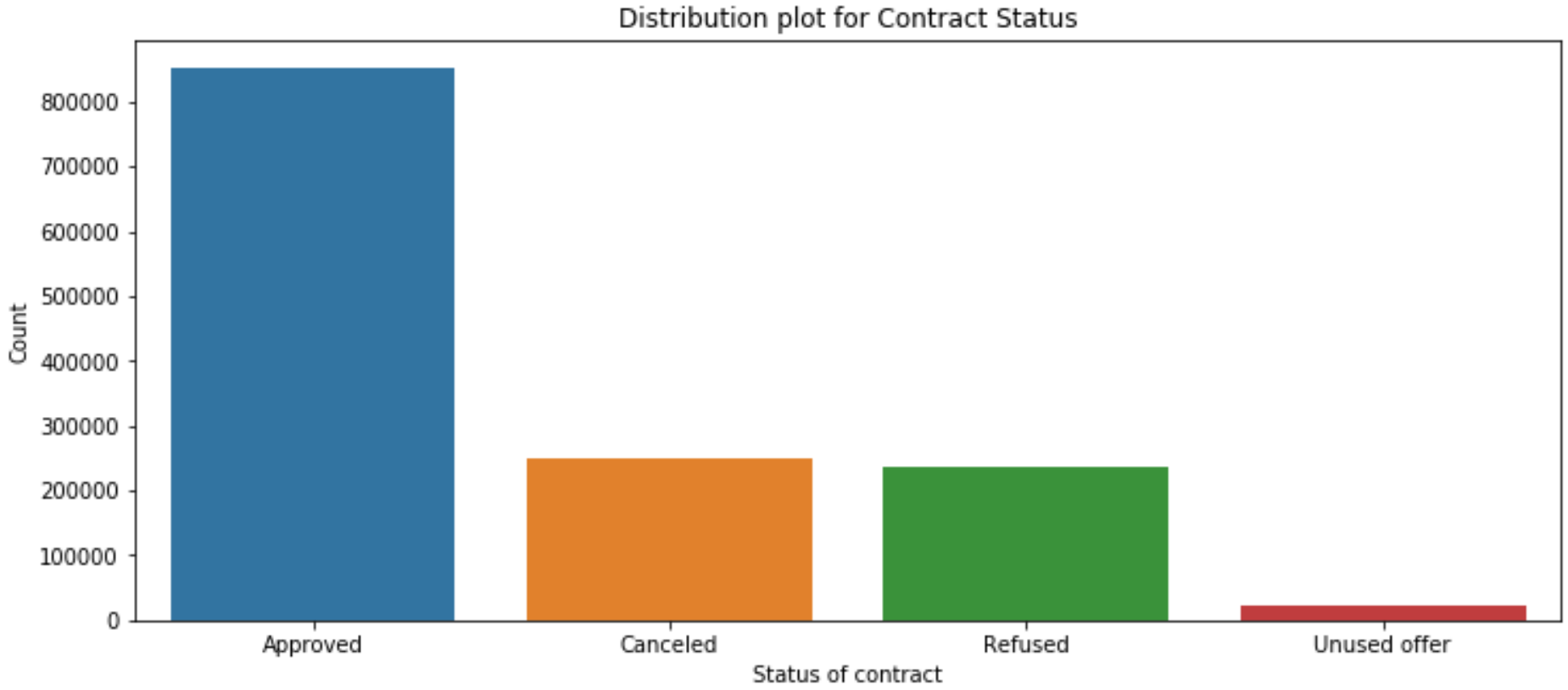
- Customer with amount credit up to 6 lac and income total up to 2.5 lac are less likely to be defaulters.

5. MERGING DATASET

- Merged application and previous application dataset on key attribute SK_ID_CURR.
- After merging both the datasets, the total number of rows were found to be 1361632 & columns were found to be 105 (1361632x105).
- Further analyzed the merged dataset on below four categories.

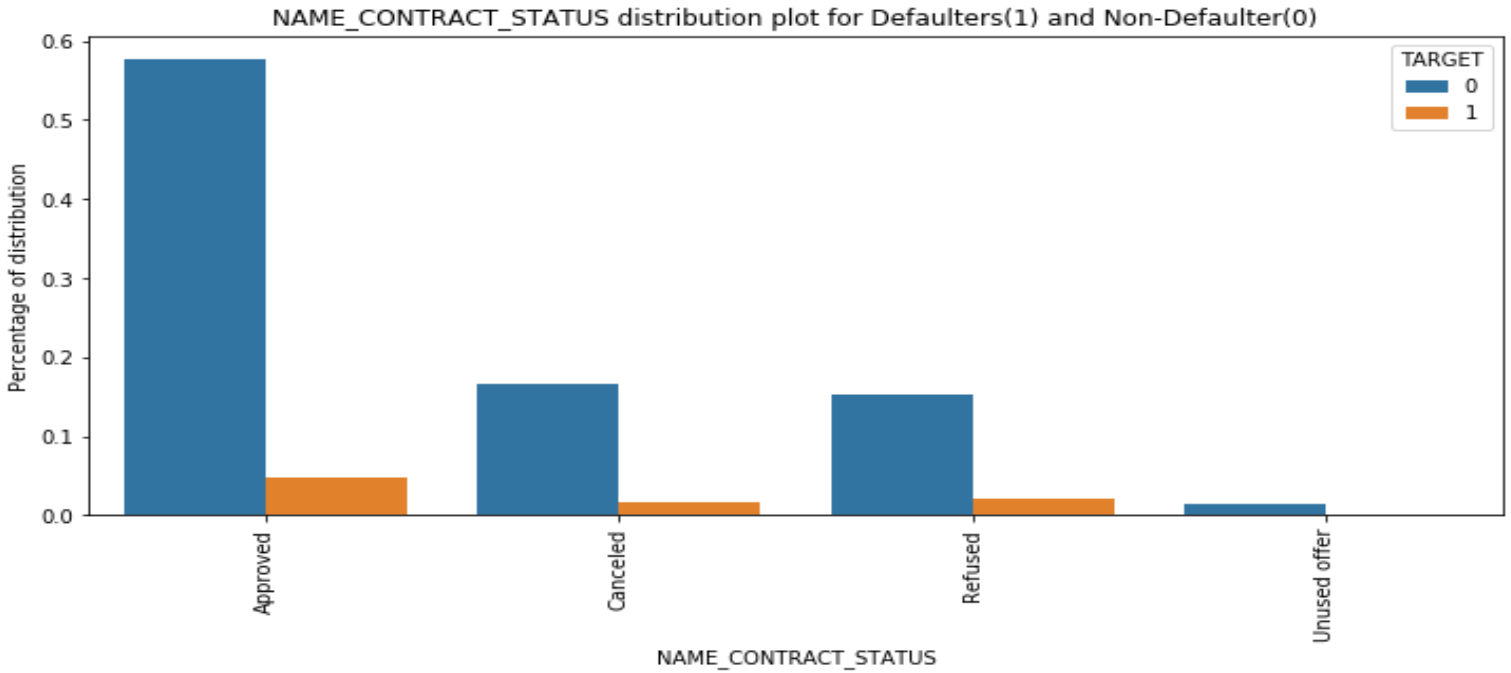
Categorizing the merged dataset into 4 categories:

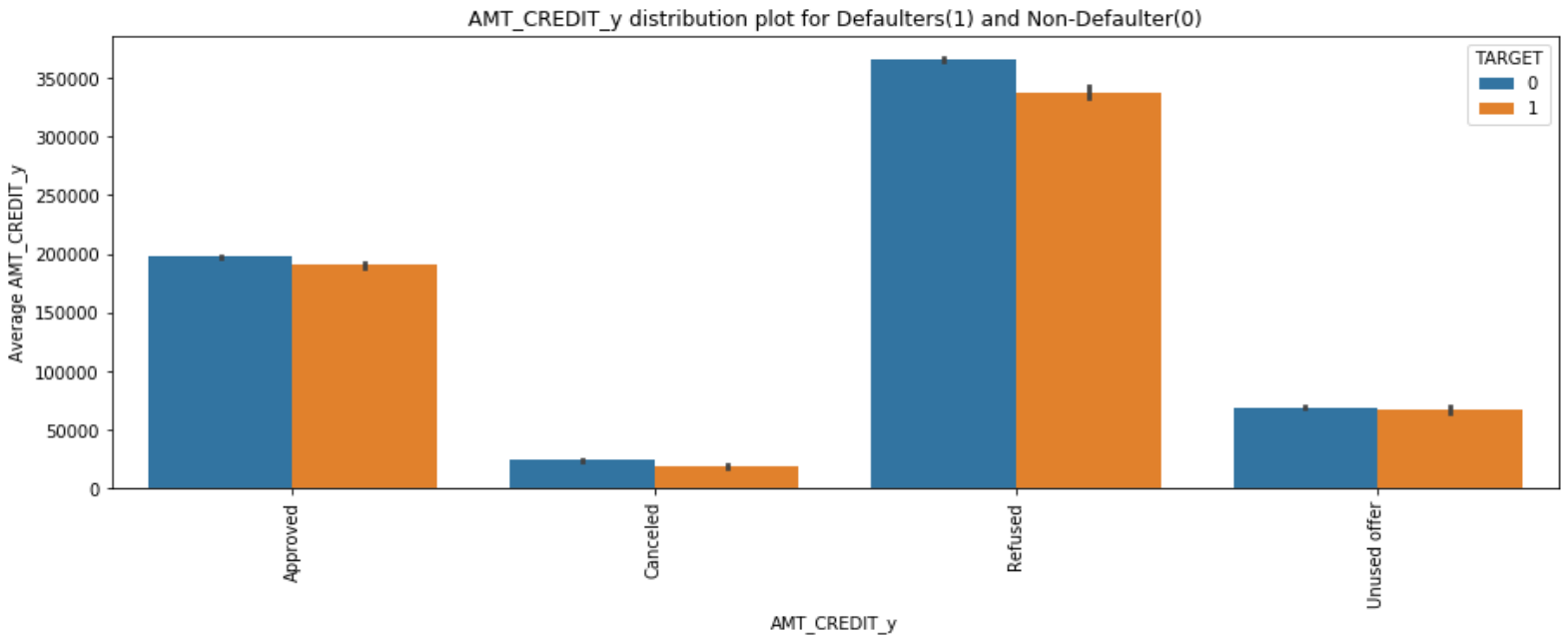
- Approved
- Cancelled
- Refused
- Unused offer



Insights :

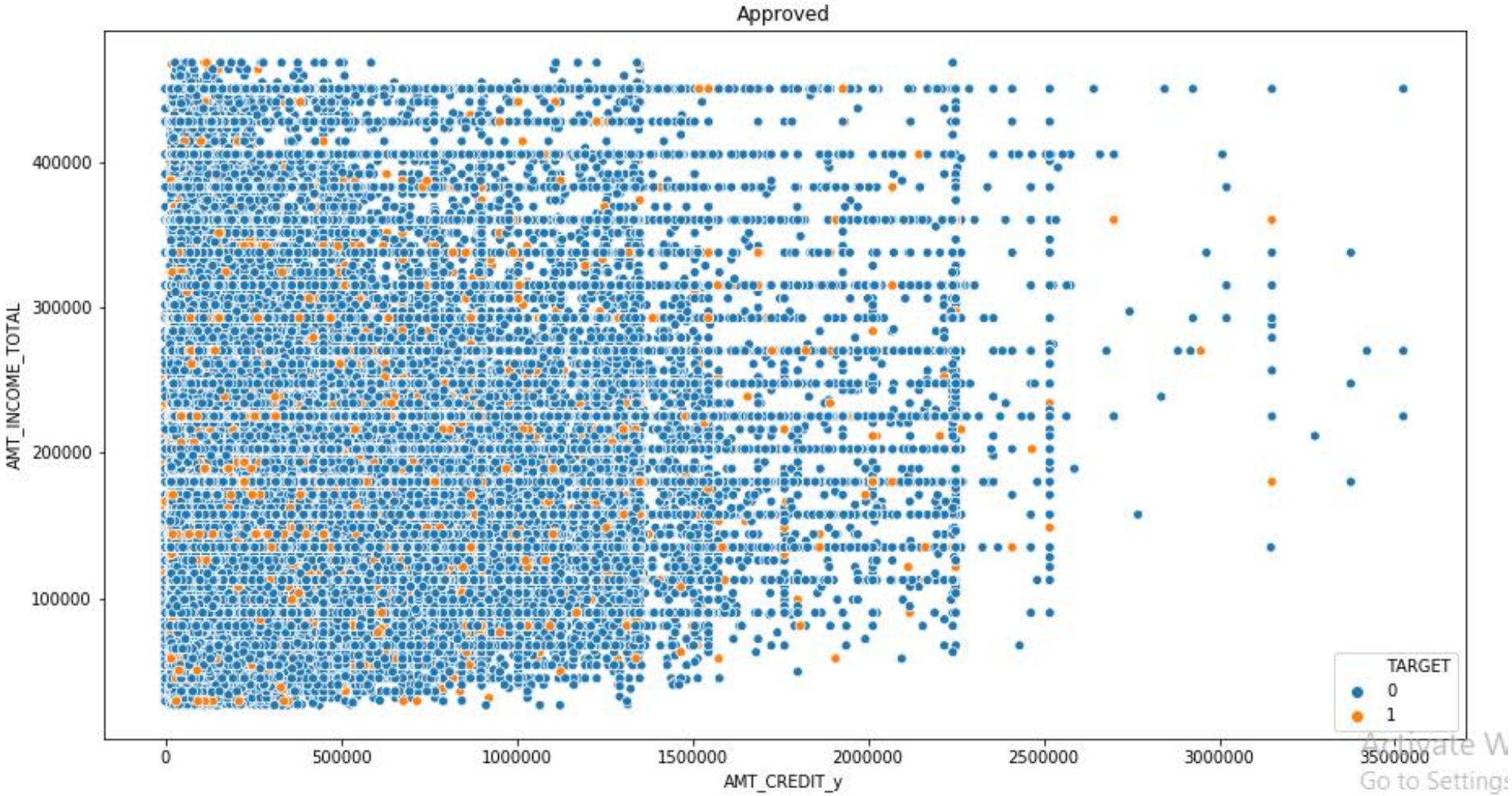
- For the customer in application data set have their previous loan as approved mostly as we can see from the distribution.





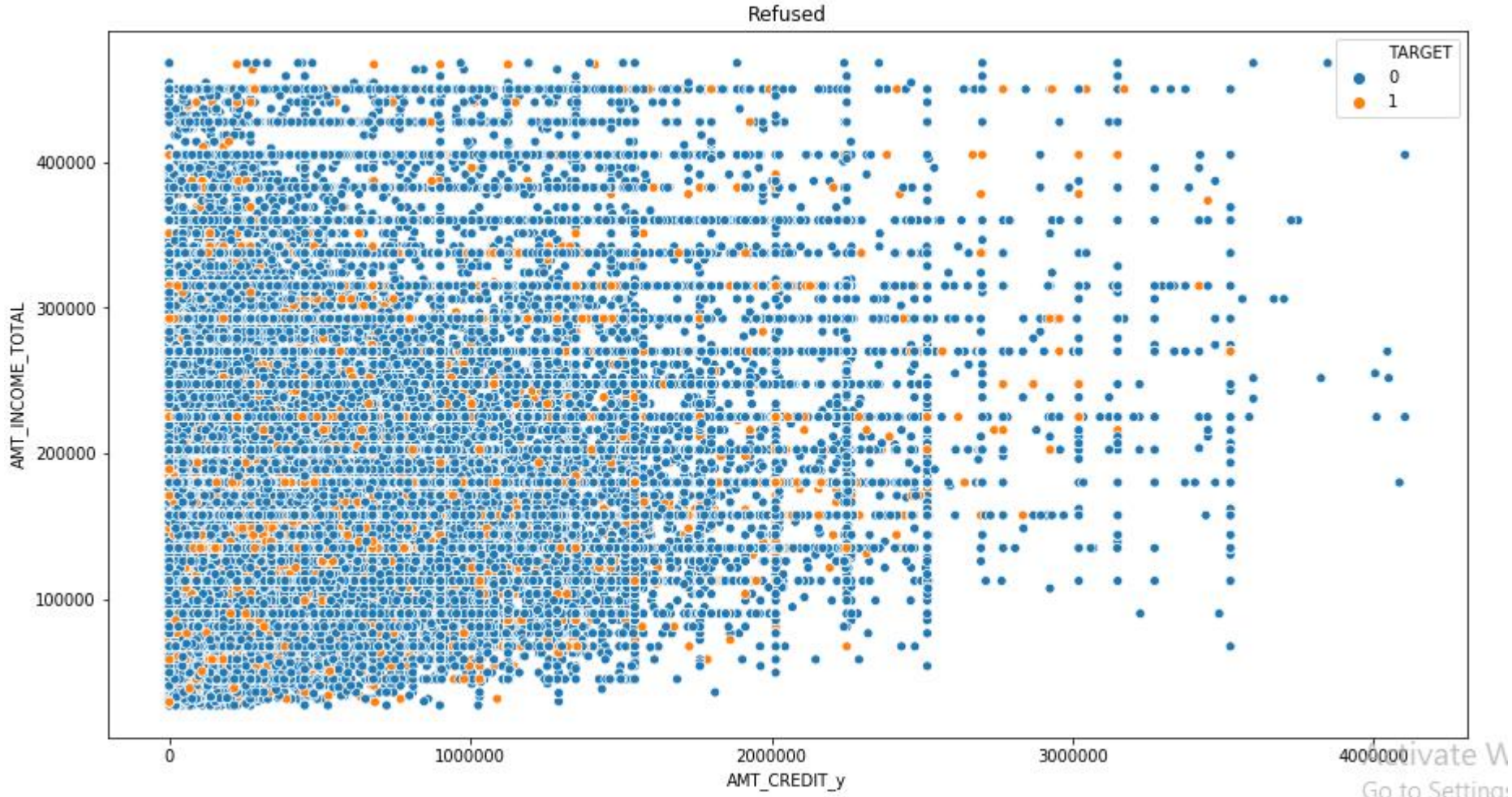
Insights :

- Customer with the previous loan application as approved are more likely to be defaulter.
- Customer with previous loan application status as Refused are less likely to be defaulter.



Insights :

- Customers whose loan application status is approved, having income total upto 3.5 lac and amount credited upto 10 lac are more likely to be defaulters.



Insights :

- Customers whose application has been refused, having income upto 2.5 lacs and amount credited upto 8 lacs would have been a defaulter.

6. CONCLUSIONS

- The Bank should focus majorly on the revolving loan type with that should prioritize female customers more as they are less likely to be defaulters.
- The bank should strictly avoid customers who are unaccompanied at the time of loan application. A guarantor can be a good option in such a case.
- There are large number of customers from working class but their chances of being a defaulter is also high.
- Banks can focus to gain the customers from the working class but from single or unmarried category.
- Banks should prioritize the educated people in the age range 45-65 while granting loan.
- Banks should consider the ratings from external sources.
- The proportion of customers owning a real estate property is more but at the same time their chances of being a defaulter is high. Banks can provide collateral loans on their property to recover the amount later.