Credit EDA Assignment

Name: Rupam Singh

Batch : DS_C63 Batch

INTRODUCTION:

Often, we come across news where customers after getting the loans are not able to repay the loan. Which leads bank to suffer lose for that.
So, this assignment aims to understand the factors through which bank faces the loss/profit.

When a client applies for the loans four types of decision should be taken by the company or loan providing bank:

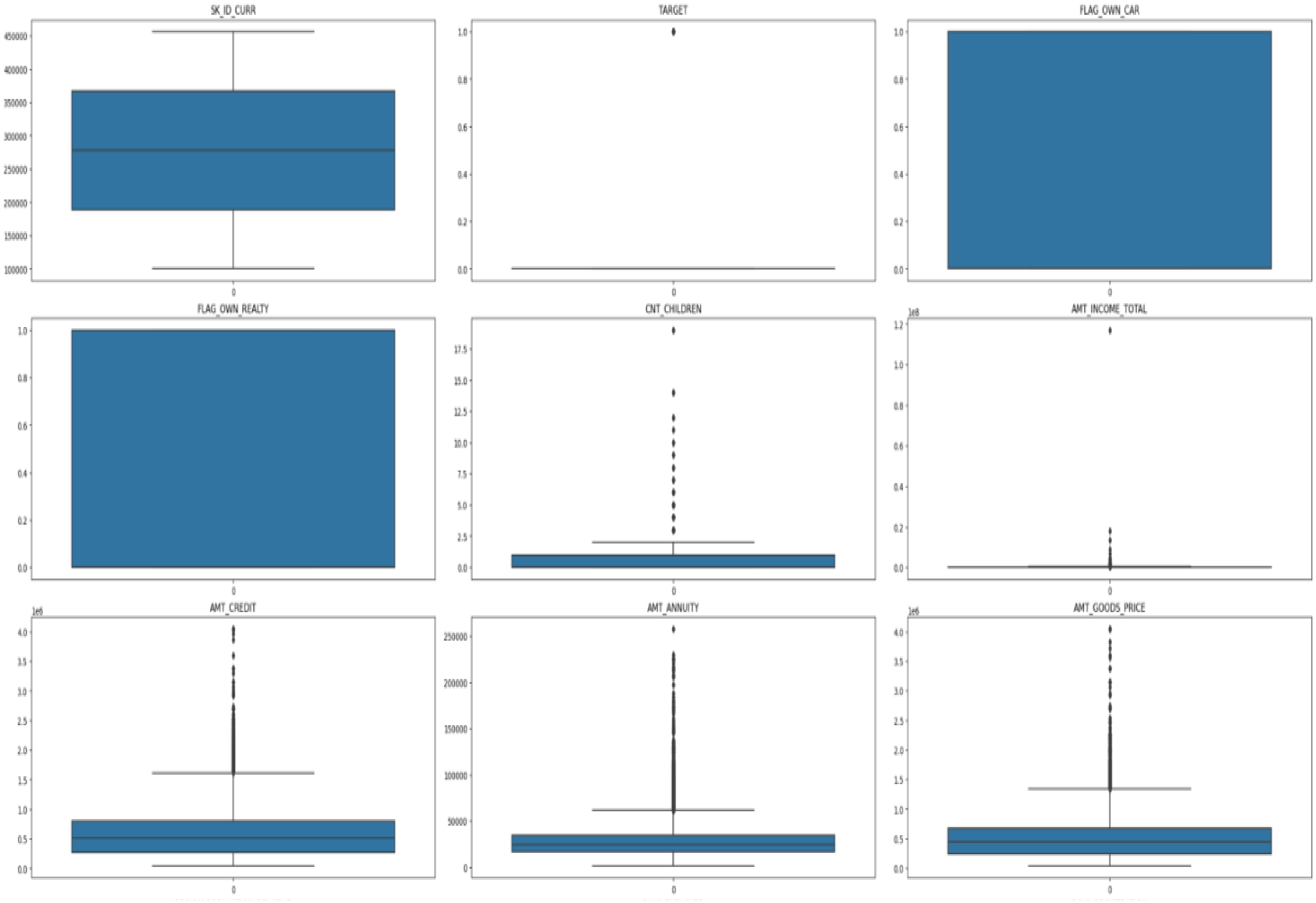
1. Whether to approve the loan.
2. Whether to reject the loan applivation.
3. Whether the client has refused, as they might have got some other better offers.
4. The loan has been cancelled by the client but at the different stages of loan process.

Let us understand the application and previous dataset.

Structure of the Data set:

1. **Application dataset**: It consists of 307511 rows and 122 columns.
2. **Previous dataset** : It consists of 1670214 rows and 37 columns.

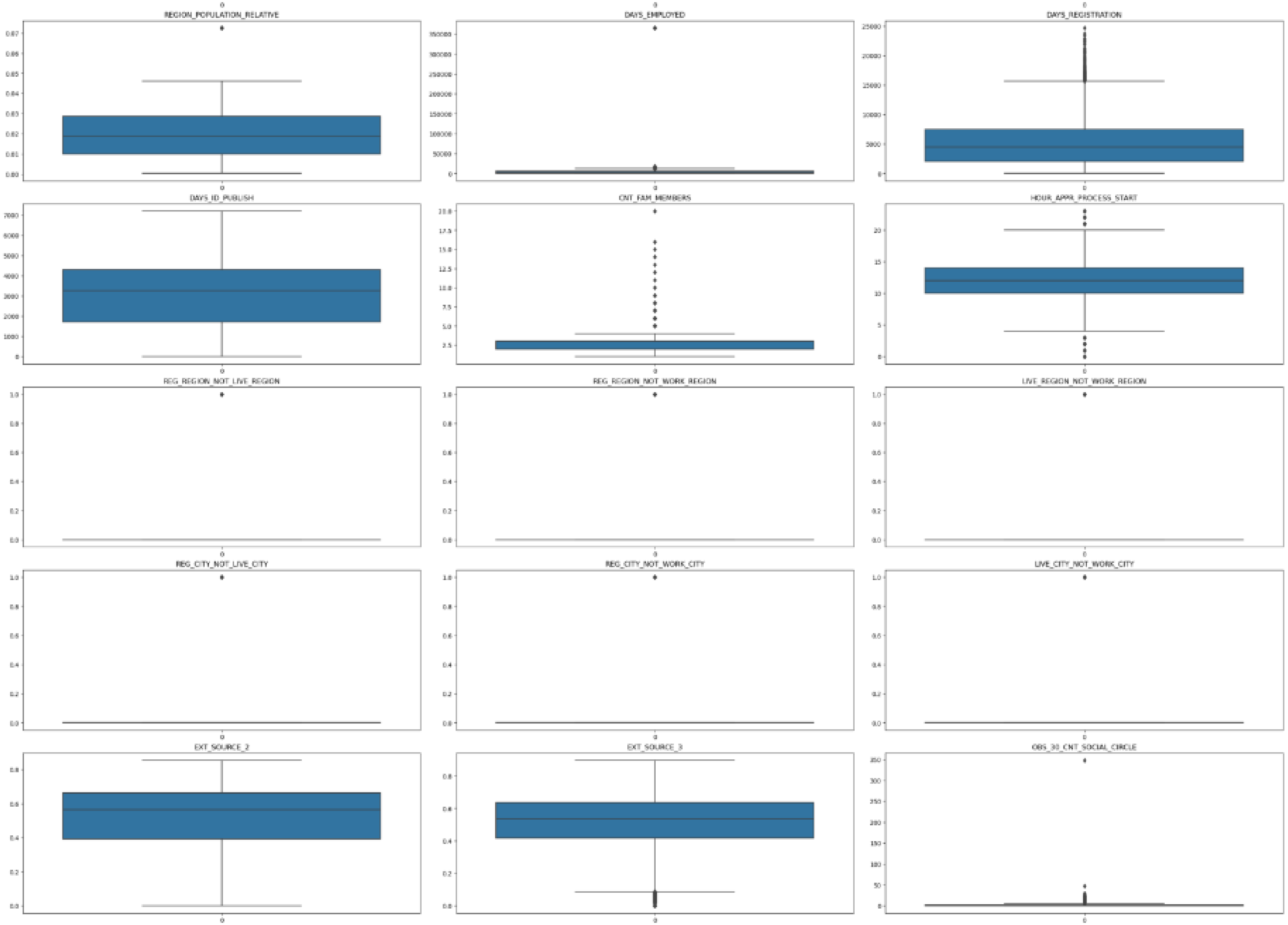
Outlier Analysis of the Application data set :



Insights:

- **FLAG_OWN_CAR** tells whether the client owns car or not. It has been answered in Y and N which has been further replaced by Y=1 and N =0.Therefore the Q1 ana Q3 is from 0 and 1 respectively.
- Similarly, **FLAG_OWN_CAR** tells whether the client owns house or not. It has been answered in Y and N which has been further replaced by Y=1 and N =0.Therefore the Q1 ana Q3 is from 0 and 1 respectively.
- **CNT_CHILDREN:** The above whiskers lies at 2.5. Q1 and Q3 lies somewhere between 0 and 1.5. We can see outliers after the maximum whiskers.The outliers is as high as 17.5.
- **AMT_INCOME_TOTAL:** A thin line can be seen that lies somewhere around 0, outliers can be seen here.
- **AMT_CREDIT:** 3rd quartile is greater than the first quartile which means most of the values lies in the third quartile, having the large numbers of quartiles.
- **AMT_ANNUITY:** 3rd quartile is greater than the first quartile which means most of the values lies in the third quartile.
- **AMT_GOODS_PRICE:** It has too many outliers to notice.

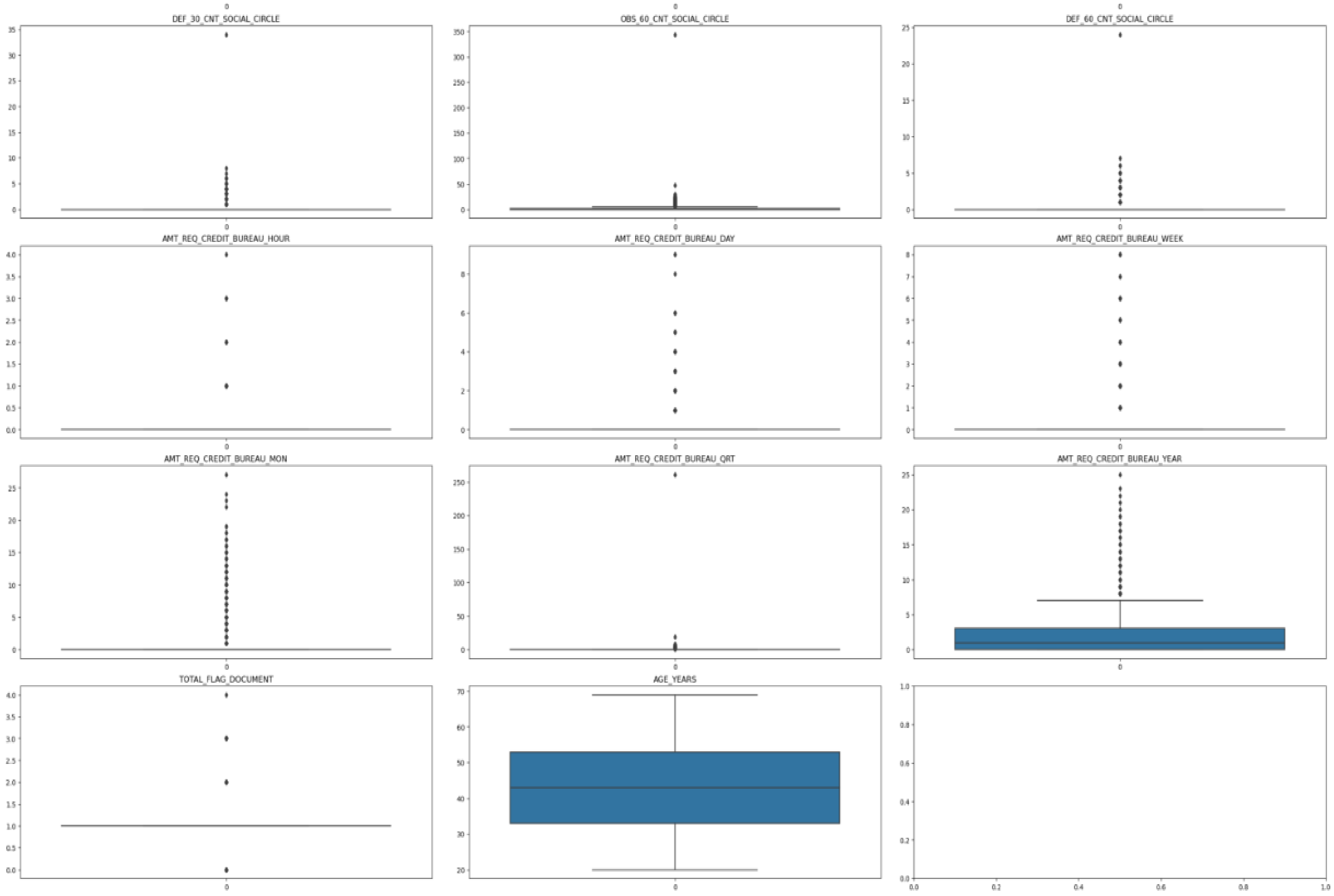
Outlier Analysis of the Application data set :



Insights:

- **REGION_POPULATION_RELATIVE:** Most of the data lies between Q1 and Q2 and very less outliers.
- **DAYS_EMPLOYED:** Only a thin line can be seen.
- **DAYS_REGISTRATION:** Median lies somewhere at 5000. Outliers are above 25000.
- **DAYS_ID_PUBLISH:** 3rd quartile is smaller than the first quartile which means most of the values lies in the first quartile.
- **HOUR_APPR_PROCESS_START:** Outliers are above and below maximum and minimum whiskers respectively.
- **EXT_SOURCE_2:** Q1 is greater than Q3.
- **EXT_SOURCE_3:** Below minimum whiskers outliers are there.
- All the LIVE_REGION_NOT_WORK_REGION, REG_REGION_NOT_WORK_REGION, REG_REGION_NOT_LIVE_REGION, LIVE_REGION_NOT_WORK_CITY etc. these have just a single thin line.

Outlier Analysis of the Application data set :

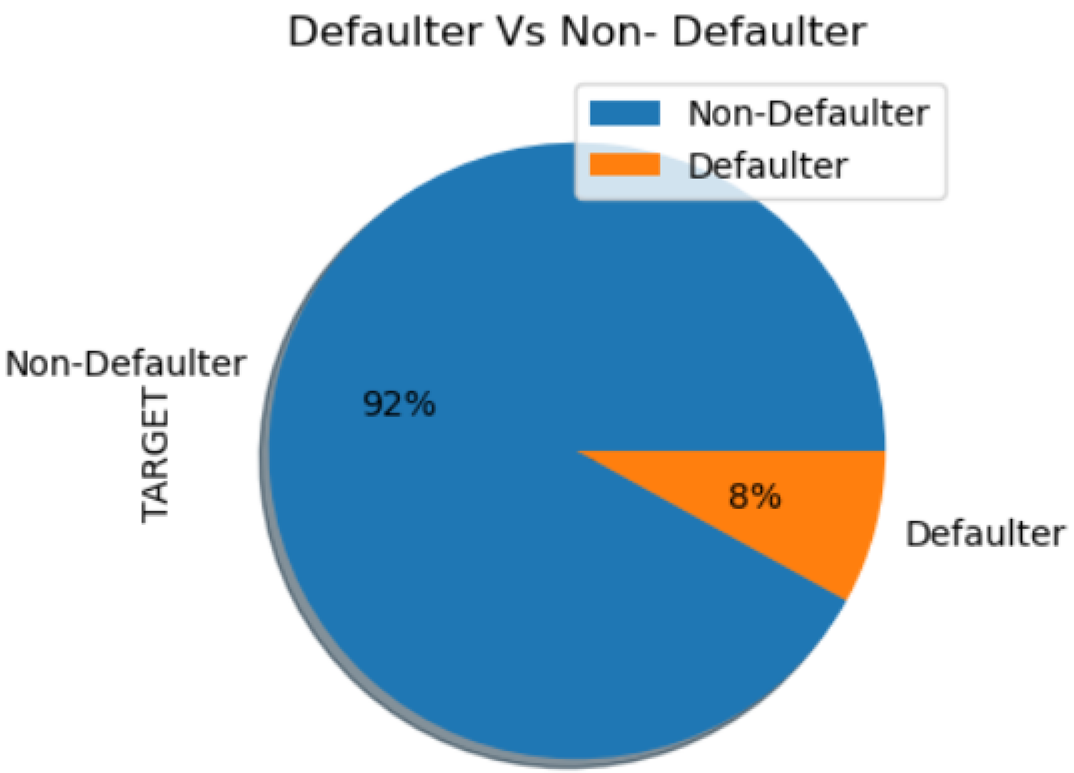


Insights:

- **AGE_YEARS:** It can be seen that the client between age group 35 to 55 has applied for the loan mostly. Average people are mostly 45 years old.
- **TOTAL_FLAG_DOCUMENT:** The maximum document submitted were 4 and the minimum documents were 0 as well.

Data Imbalance:

Here, “TARGET” column helps us to understand the defaulters and non – defaulters.
Target with value 1 : shows client has the payment difficulty/defaulters
Target with value 0 : shows client not having the payment difficulties/non-defaulters.



The numbers of defaulters are 24825, which is 8%.
The number of non-defaulters are 282686, which is 92% of overall data.

We can calculate the percentage ratio of data imbalance:

Data imbalance% = $\frac{\%\, of\ non{-}defaulters}{\%\, of\ defaulters}$

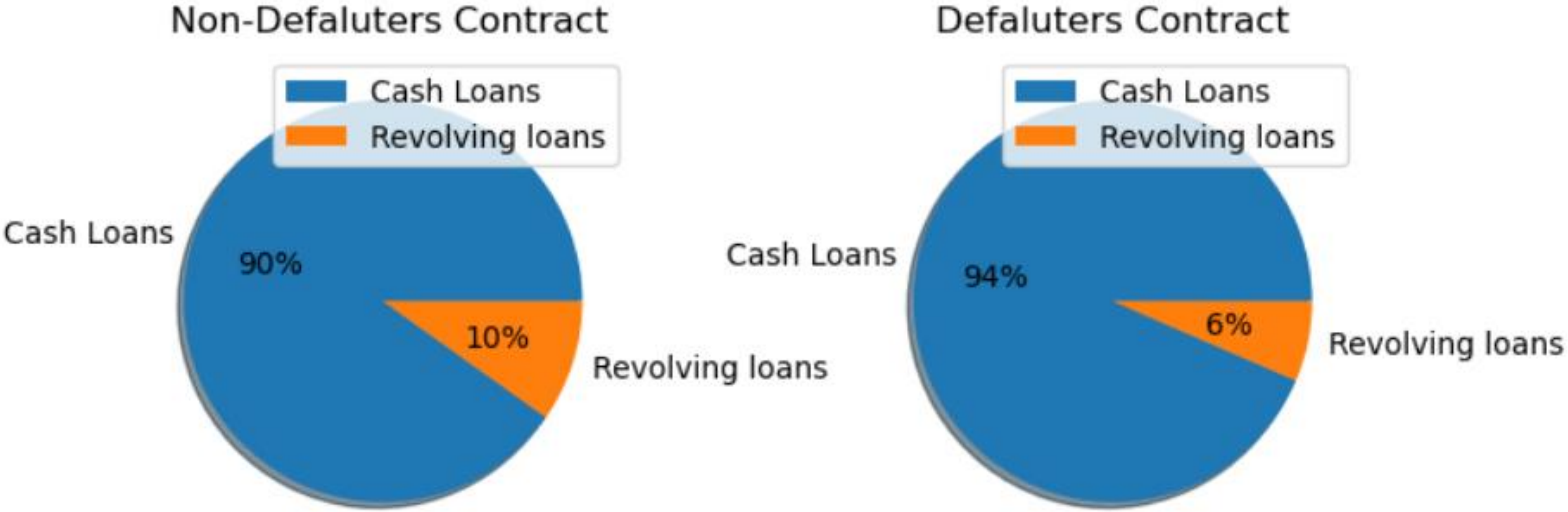
$= \frac{92}{8}$

$= 11.5\%$

Univariate Analysis: Contract type

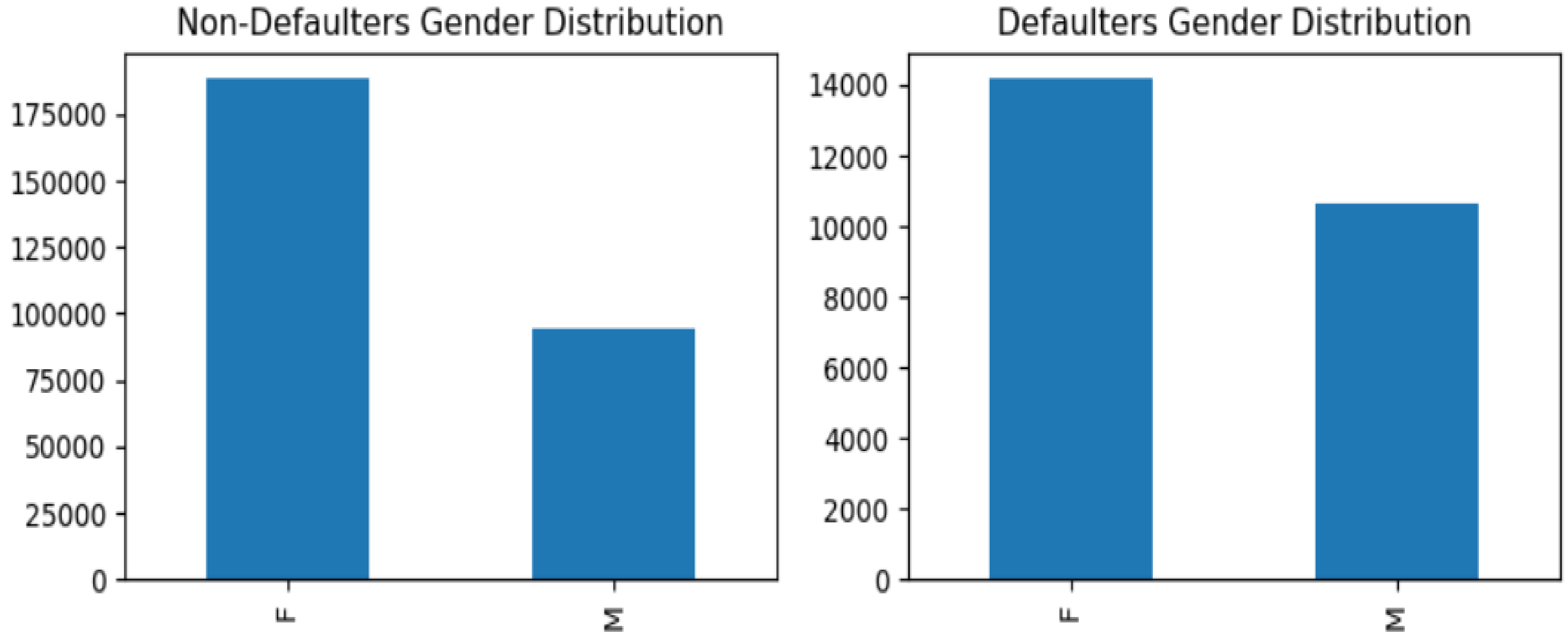
After understanding the data imbalance let us make analysis on the basis of defaulters and non- defaulters.
Here we are dividing the application set data into two:

1. Non-defaulters
2. Defaulters



Insights: In both the cases, the contract type is mostly the cash loans. Only few clients have opted for the revolving loans.

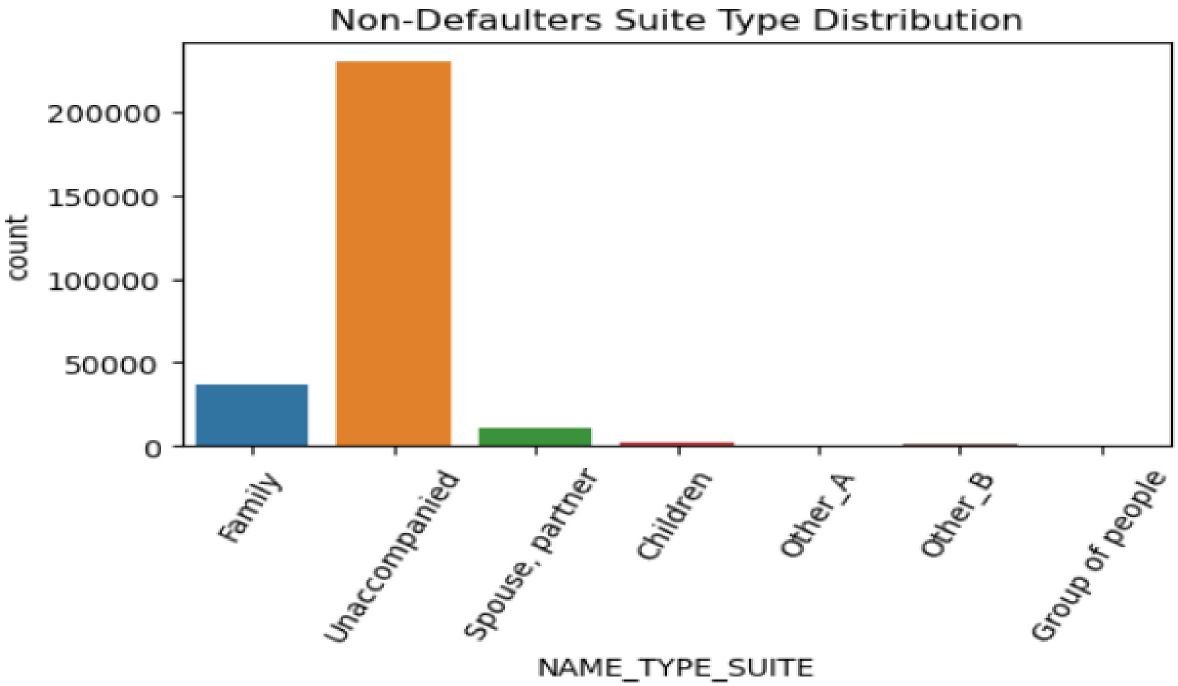
Univariate Analysis: Gender



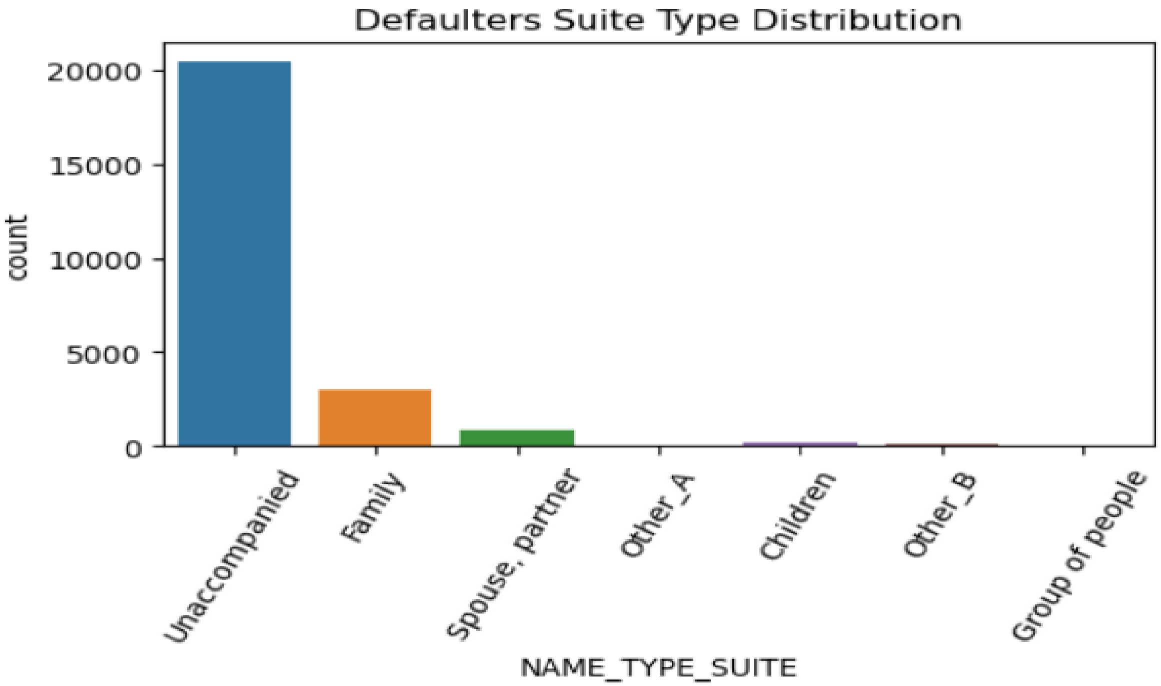
Insights: In both the cases, Female client is more than the male clients.

Univariate Analysis: Categorical type

1. **Analysis on Suite Type Distribution** : This column shows who was accompanying client when he/she was applying for the loan.



Non-Defaulters Suite Type Distribution

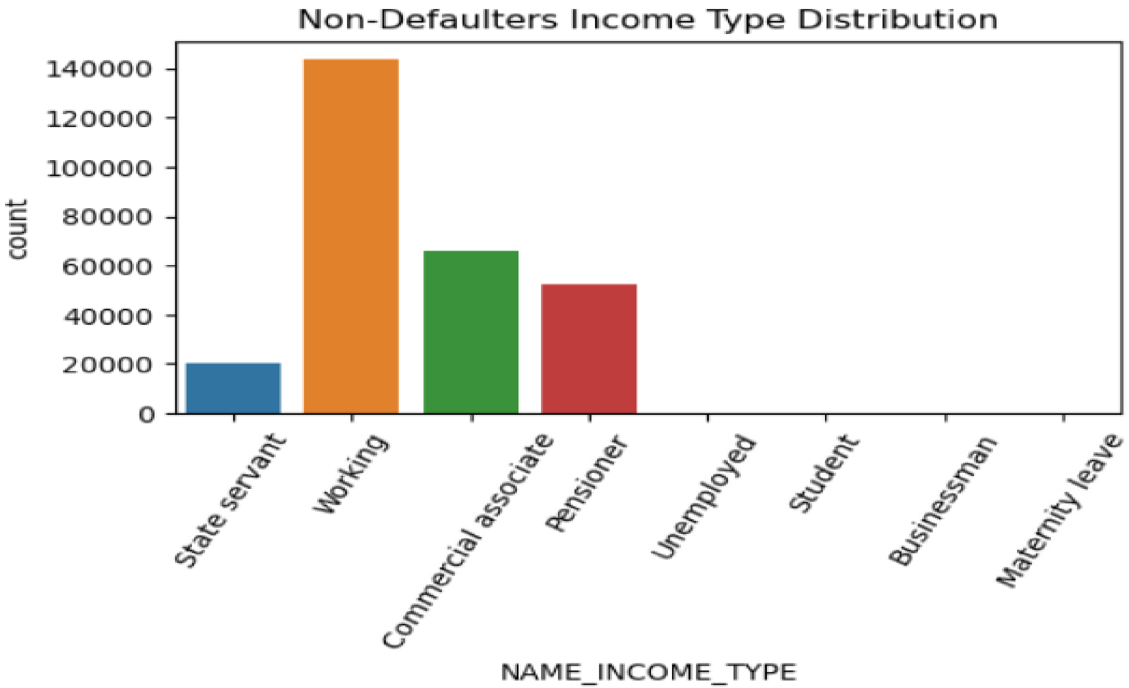


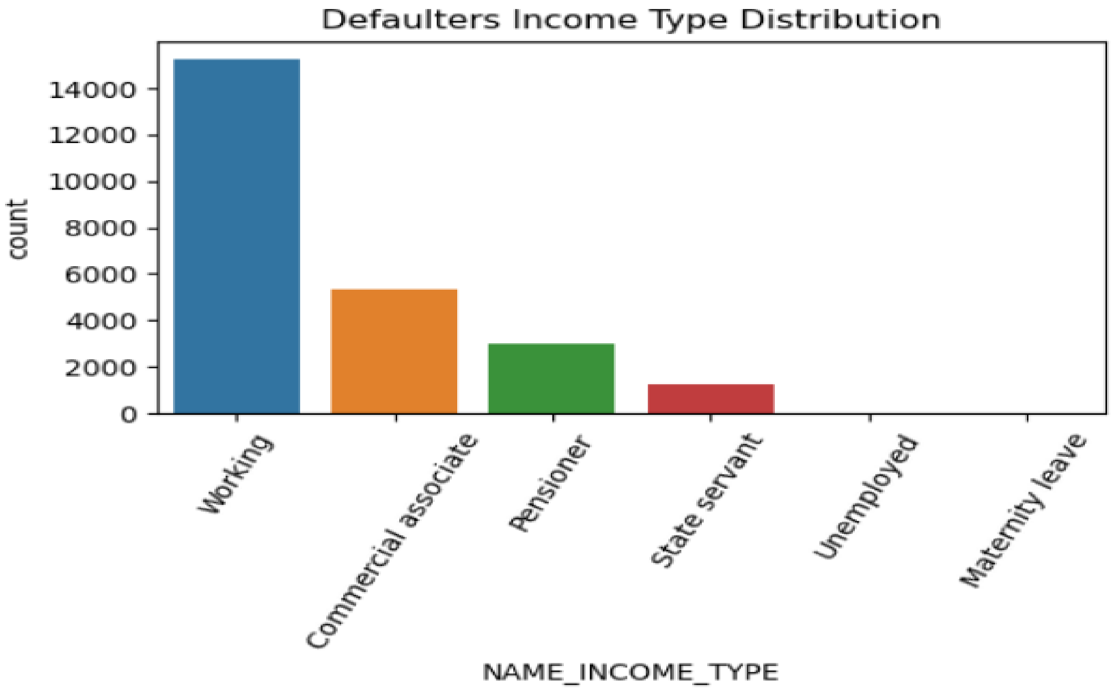
Defaulters Suite Type Distribution

Insights: In both the cases, mostly the client was unaccompanied then followed by family, followed by partner and so on. So we can't actually say whether the client being defaulter or non-defaulter does really depend upon the person accompanying him/her.

Univariate Analysis: Categorical type

2. **Income type Distribution** : This column shows clients income type. Different type of employment are state servant, working, commercial associate, Pensioner, unemployed, student, businessman etc.



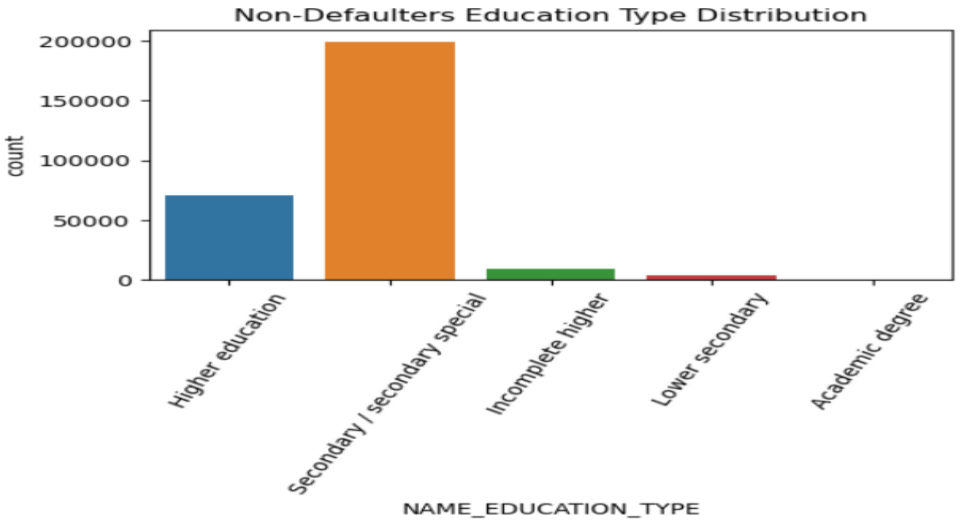


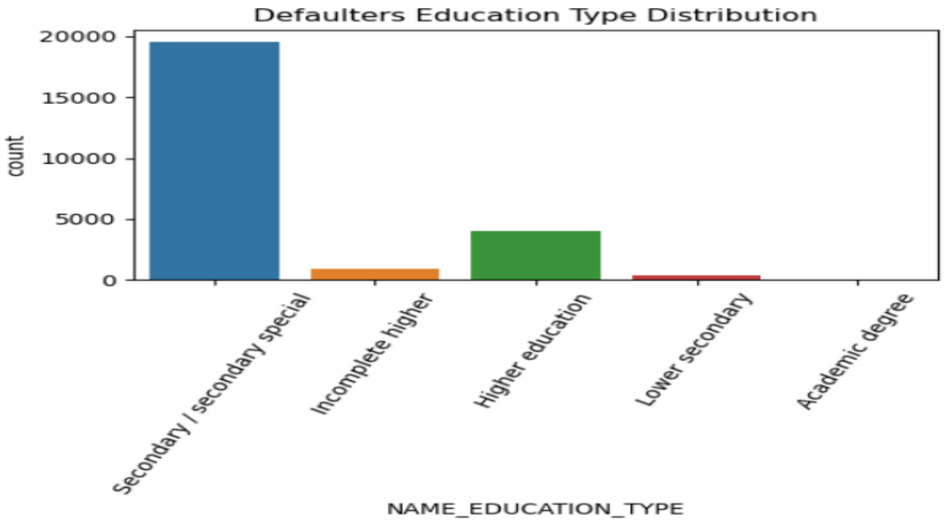
Insights:

- In both the cases, large number of client belongs to working income type.
- The interesting point to note here that is businessman client are very few.
- None of the businessman is in the defaulters list.
- If students have applied for the loan then they are able to pay the loan as they must be supported by their family.(it is just the assumption)

Univariate Analysis: Categorical type

3. **Education type Distributio**n : This column shows level of highest education the client achieved.

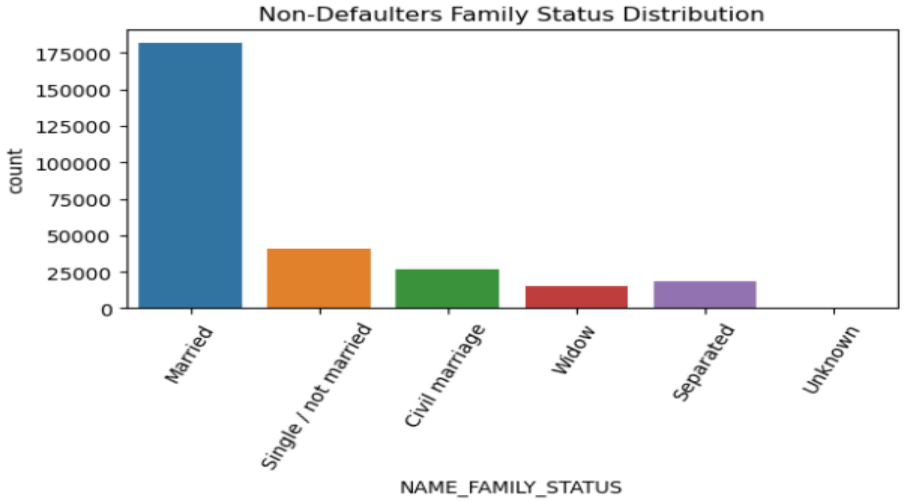


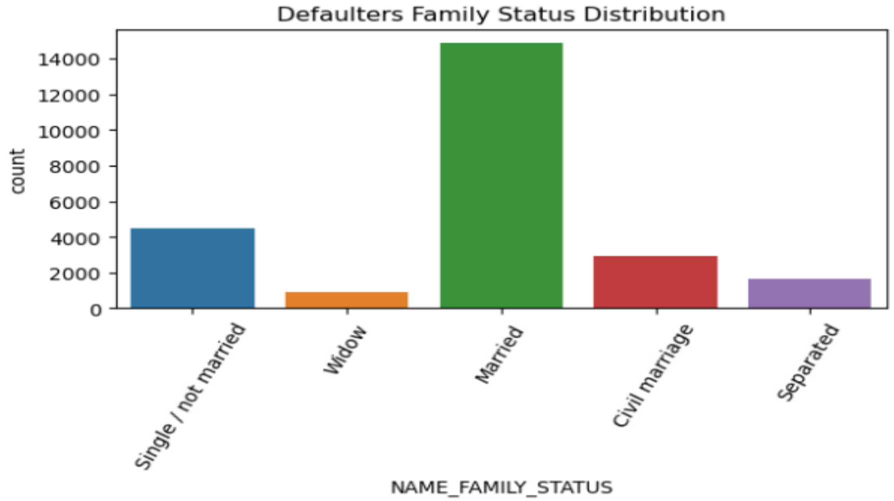


Insights:

- In both the cases, the client has secondary special as highest education level , followed by higher education.

4. **Family Status Distributio**n : This column shows family status of the client.



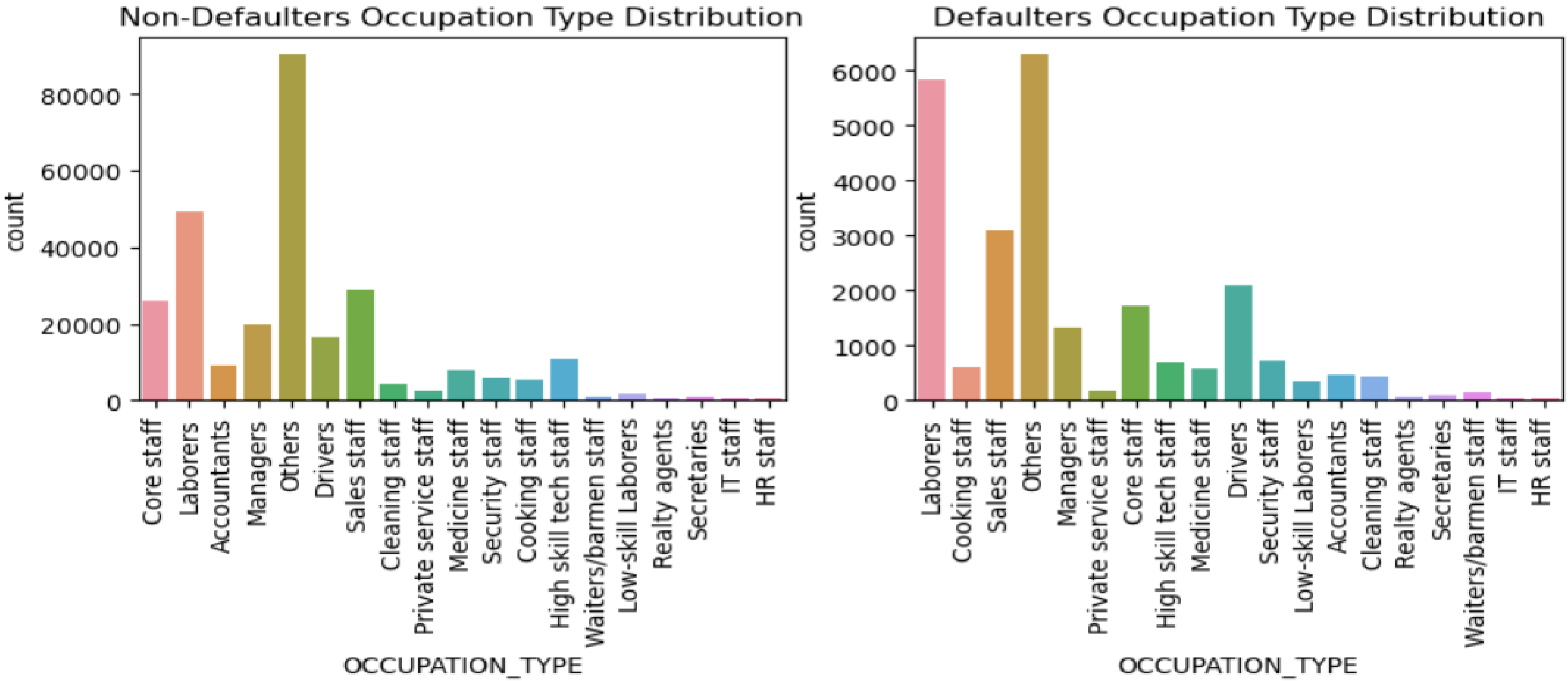


Insights:

- In both the cases, the client who have taken loan are mostly married. Then followed by singles. There is a group of people who have not shared their family status.

Univariate Analysis: Categorical type

5. **Occupation type Distributio**n : This column shows what kind of occupation does the client have.

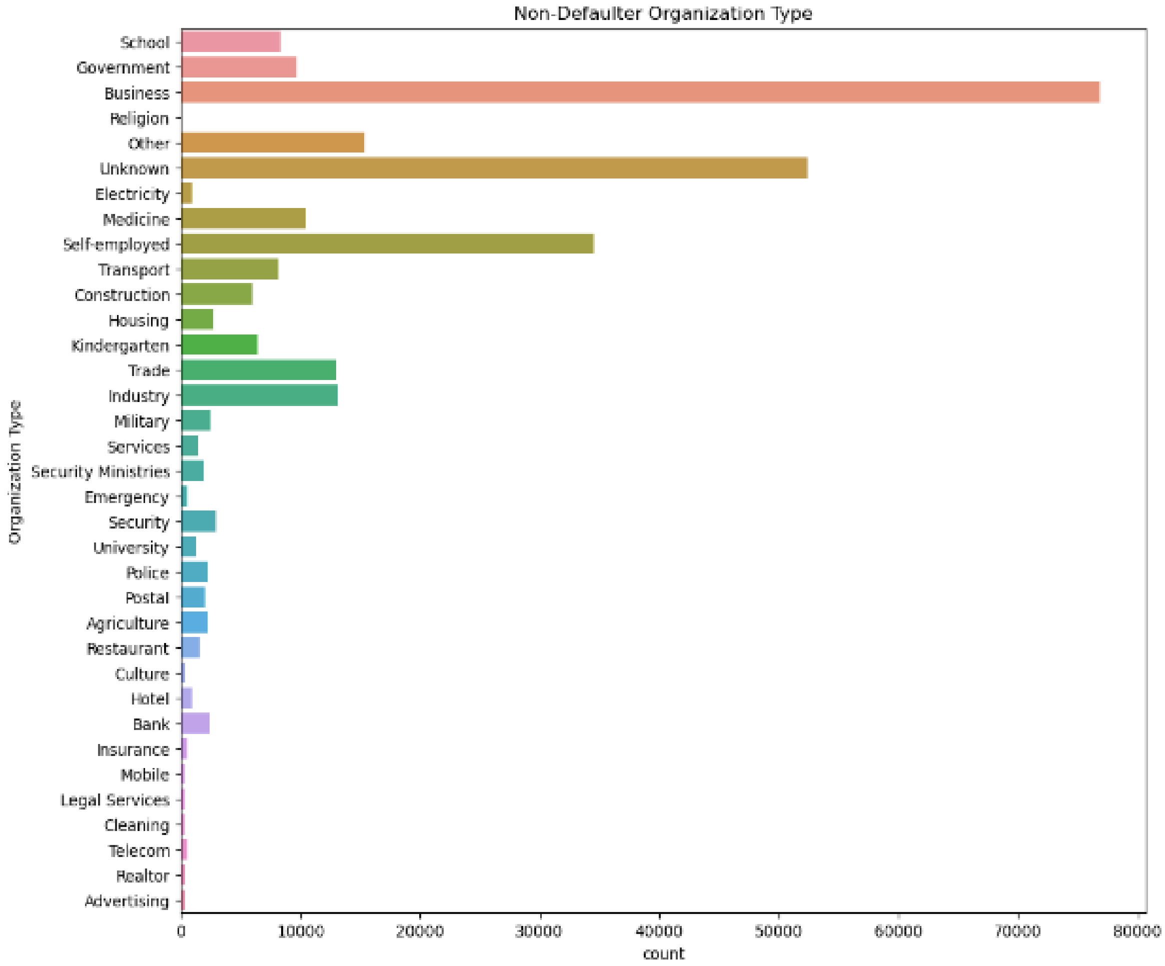


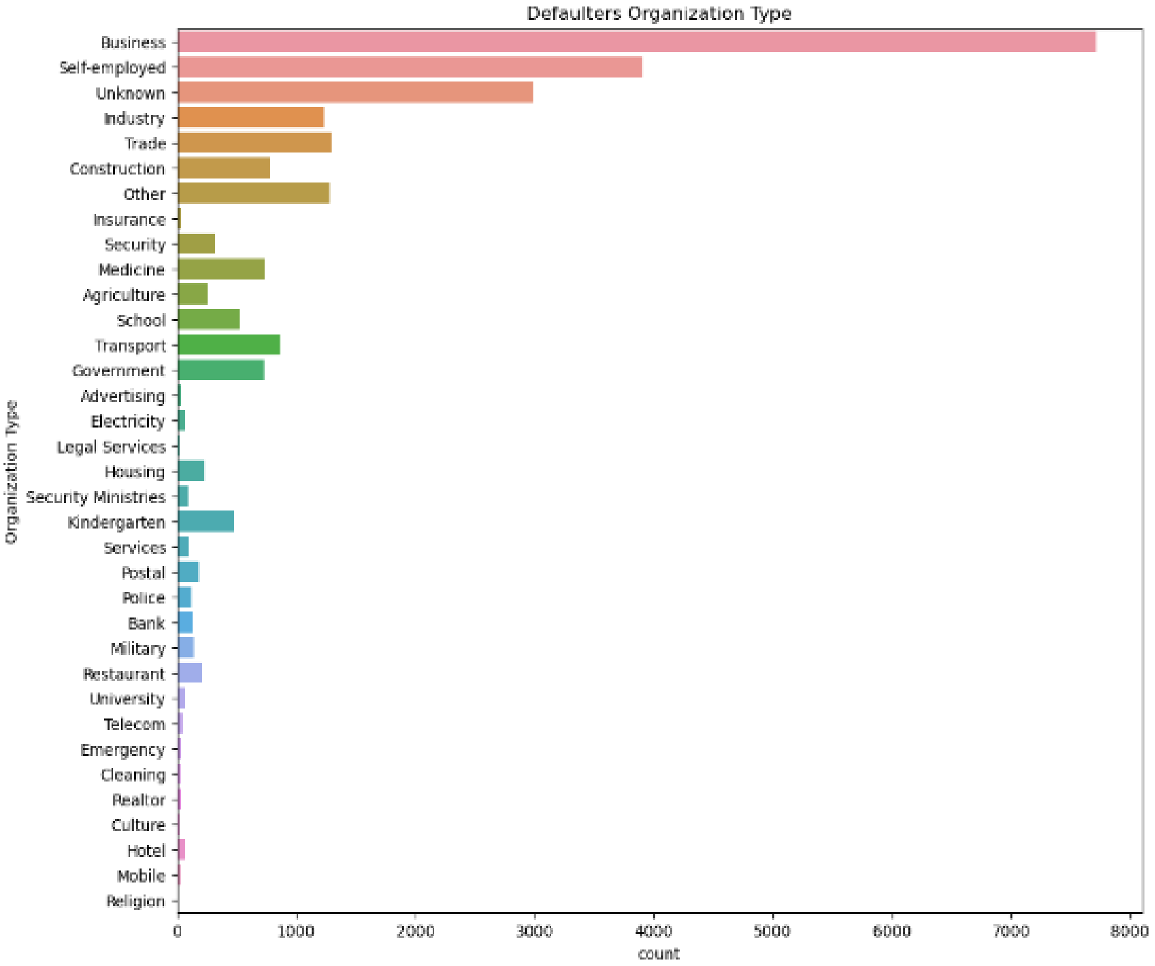
Insights:

- In both the cases, the client has mostly not shared their occupation type. So we can say that laborers are mostly high in both the cases.

Univariate Analysis: Categorical type

6. **Organization type Distributio**n : This column shows type of organization where client works.



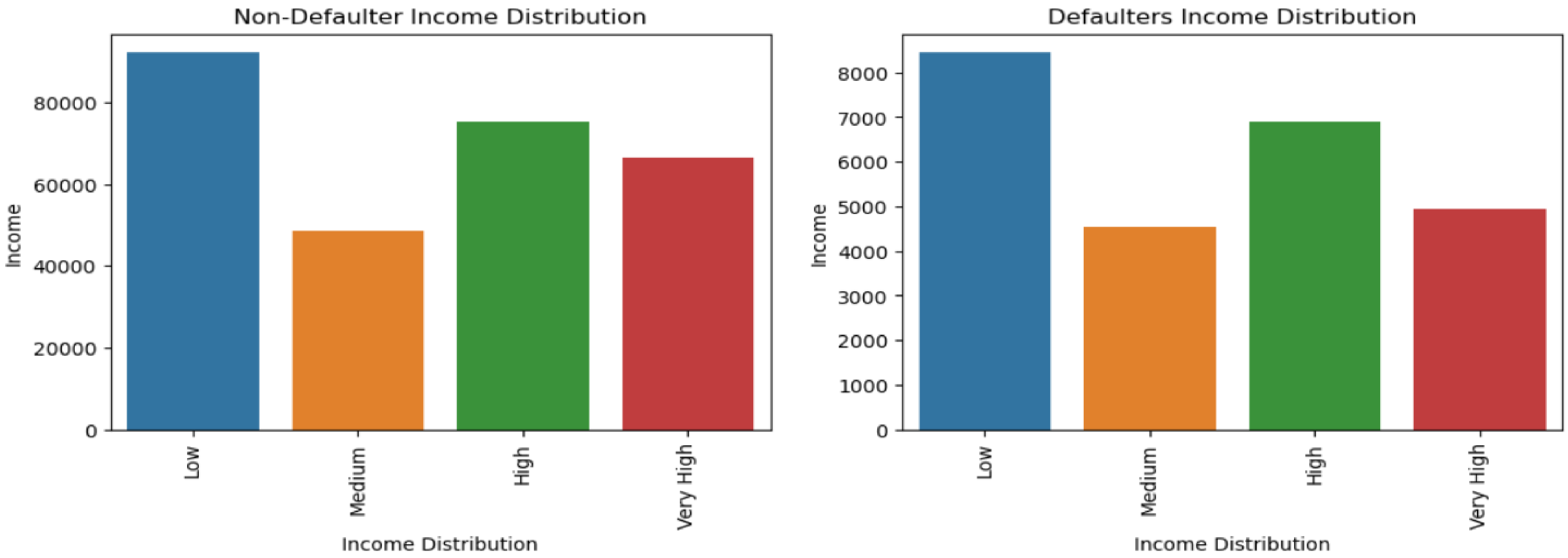


Insights: For Non-Defaulters: Business, self employed, trade, industry and others have applied most for the loans.

For Defaulters: It is from Business, selfemployed, industry, trade, other and transport are the ones have mostly applied for the loans.

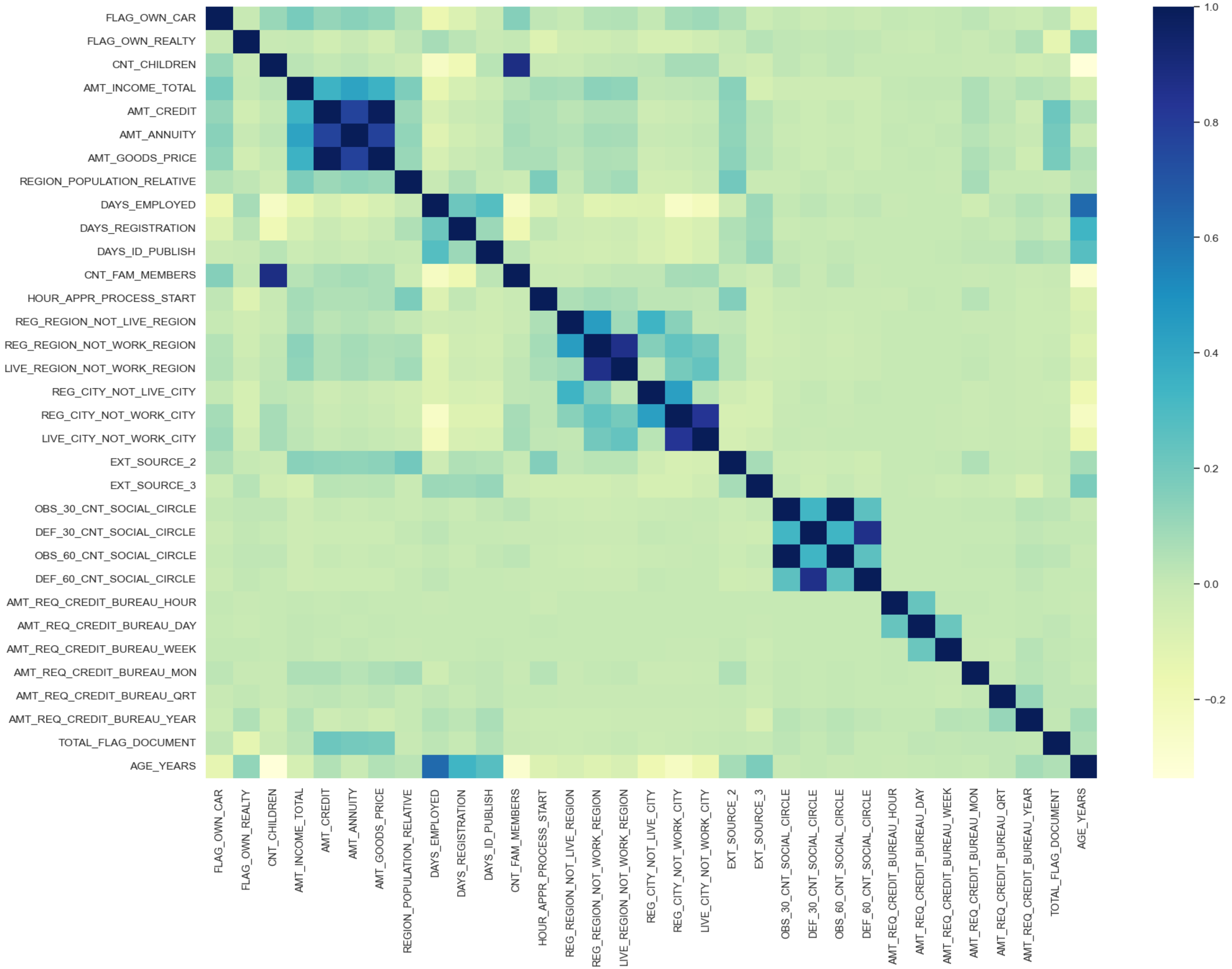
Univariate Analysis: Categorical type

7. **Income type Distributio**n : This column shows income of the client



Insights: For both non-defaulters and defaulters, the income of the client is mostly low. A good number of high income people have also opted for loan.

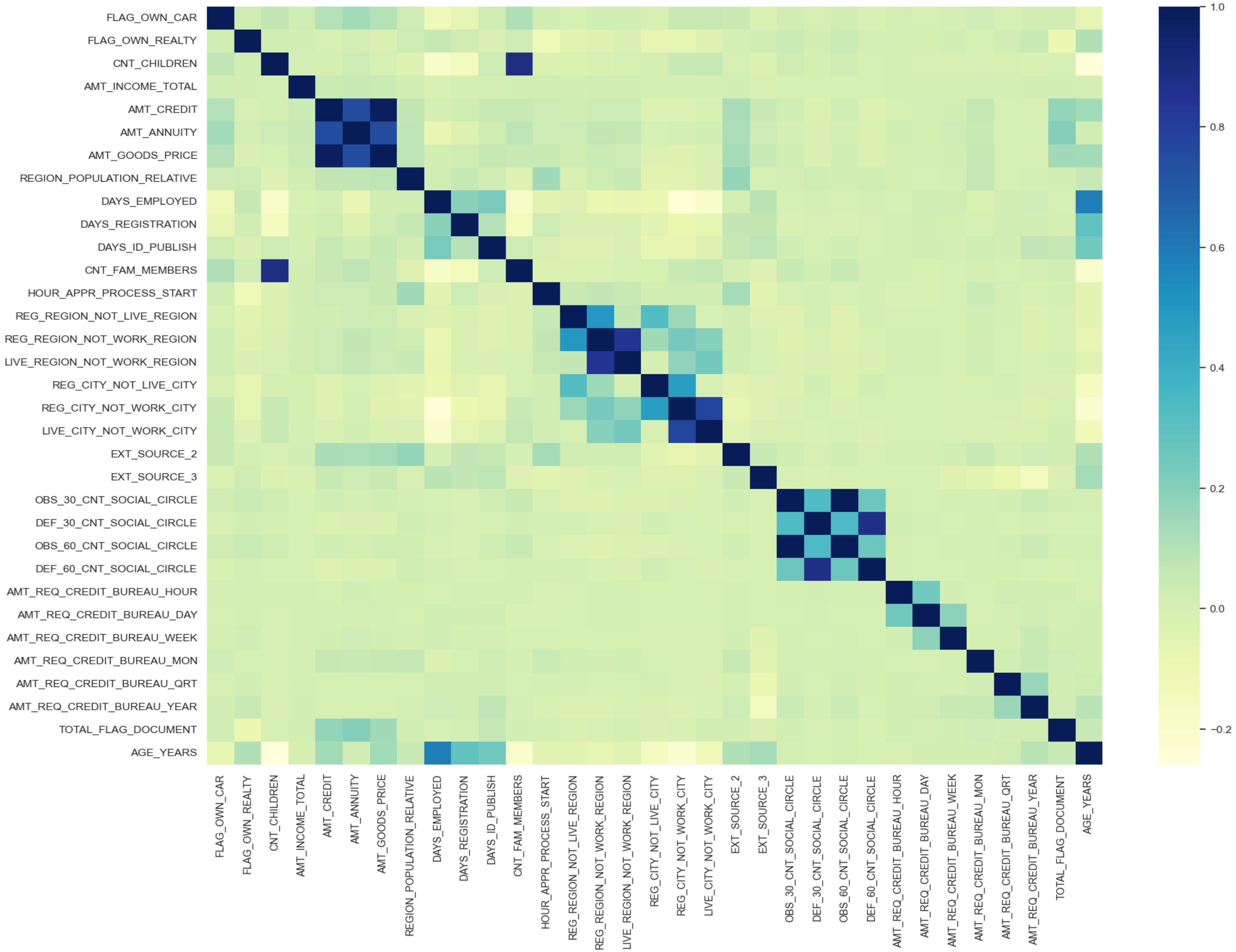
Correlation between numerical variables: Non-Defaulters



Insights:

- Clients belonging to high age group have low amount credit.
- Children count is directly proportional to family members count.
- AMT_ANNUITY is highly proportional to AMT_INCOME_TOTAL
- DAYS_EMPLOYED is less common with AMT_INCOME_TOTAL.
- AMT_INCOME_TOTAL is proportional to REGION_POPULATION_RELATIVE.
- TOTAL_FLAG_DOCUMENT is proportional with high credit amount and annuity amount.

Correlation between numerical variables: Defaulters

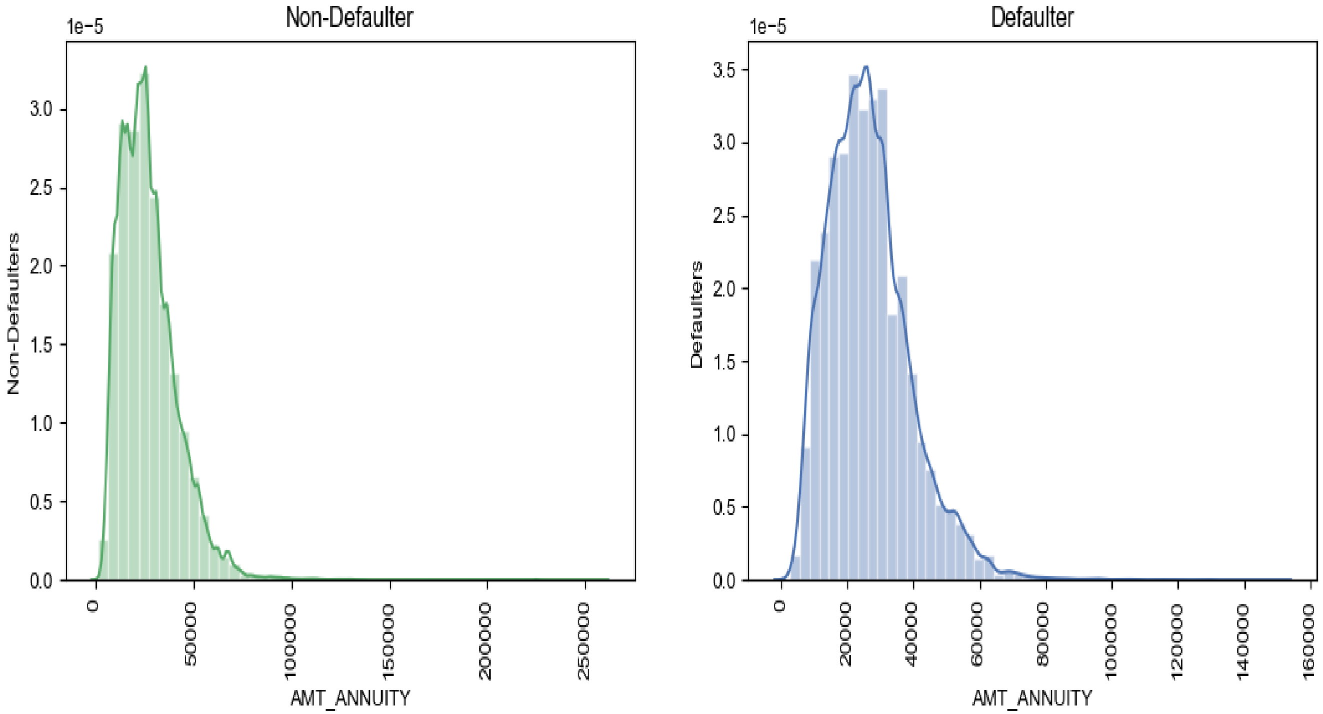


Insights:

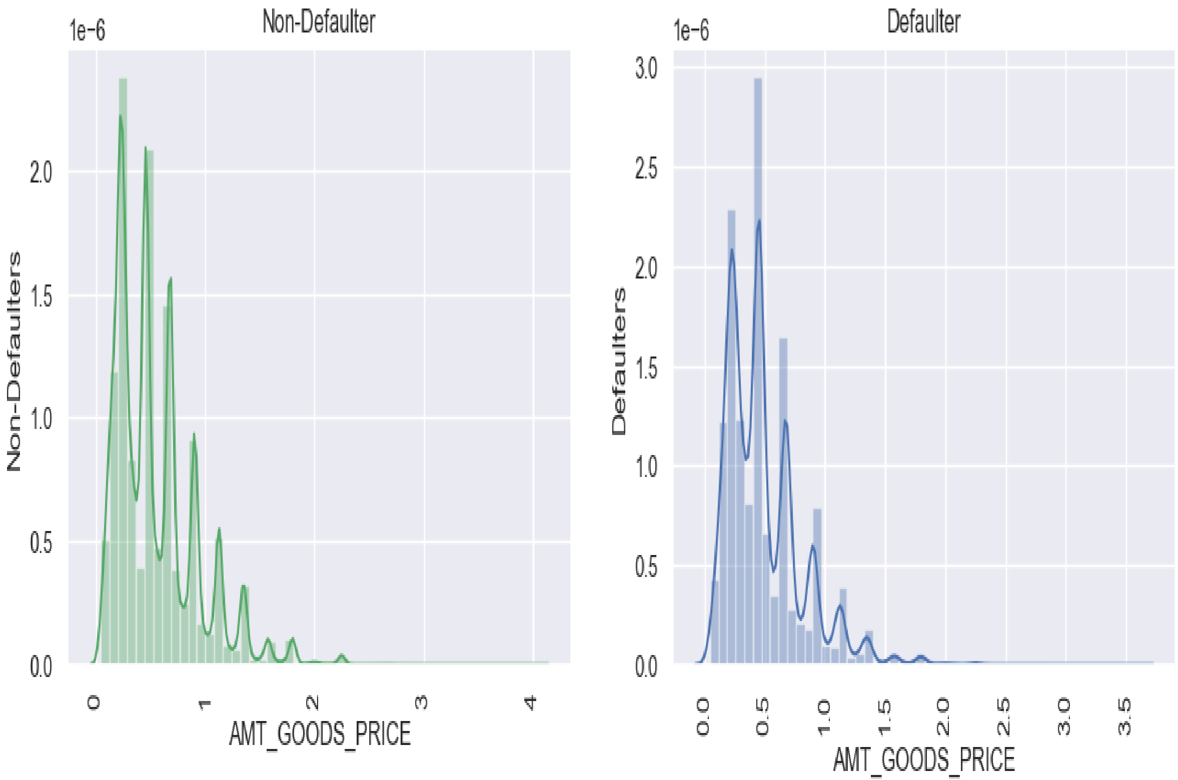
- Clients belonging to high age group have low amount credit.
- Children count is directly proportional to family members count.
- AMT_ANNUITY is highly proportional to AMT_INCOME_TOTAL
- DAYS_EMPLOYED is less common with AMT_INCOME_TOTAL.
- AMT_INCOME_TOTAL is proportional to REGION_POPULATION_RELATIVE.
- TOTAL_FLAG_DOCUMENT is proportional with high credit amount and annuity amount.
-

Univariate Analysis: Numerical types

1. **AMT_ANNUITY**: This column shows Loan annuity.



2. **AMT_GOODS_PRICE**: This column shows for consumer loans it is the price of the goods for which the loan is given



Insights:

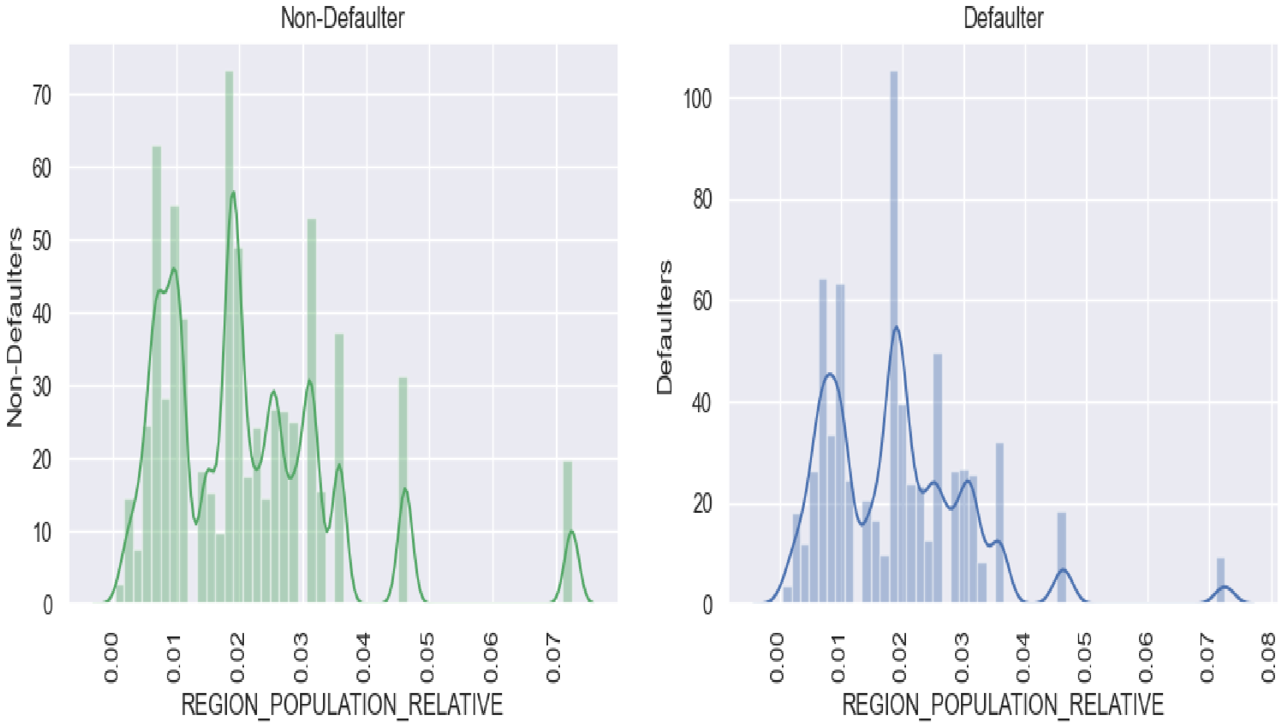
- For non-defaulters, most loan annuity ranges between 0 to 75000.
- For defaulters, its ranges between 0 to 60000.

Insights:

- Both non-defaulters and defaulters have similar shape.

Univariate Analysis: Numerical types

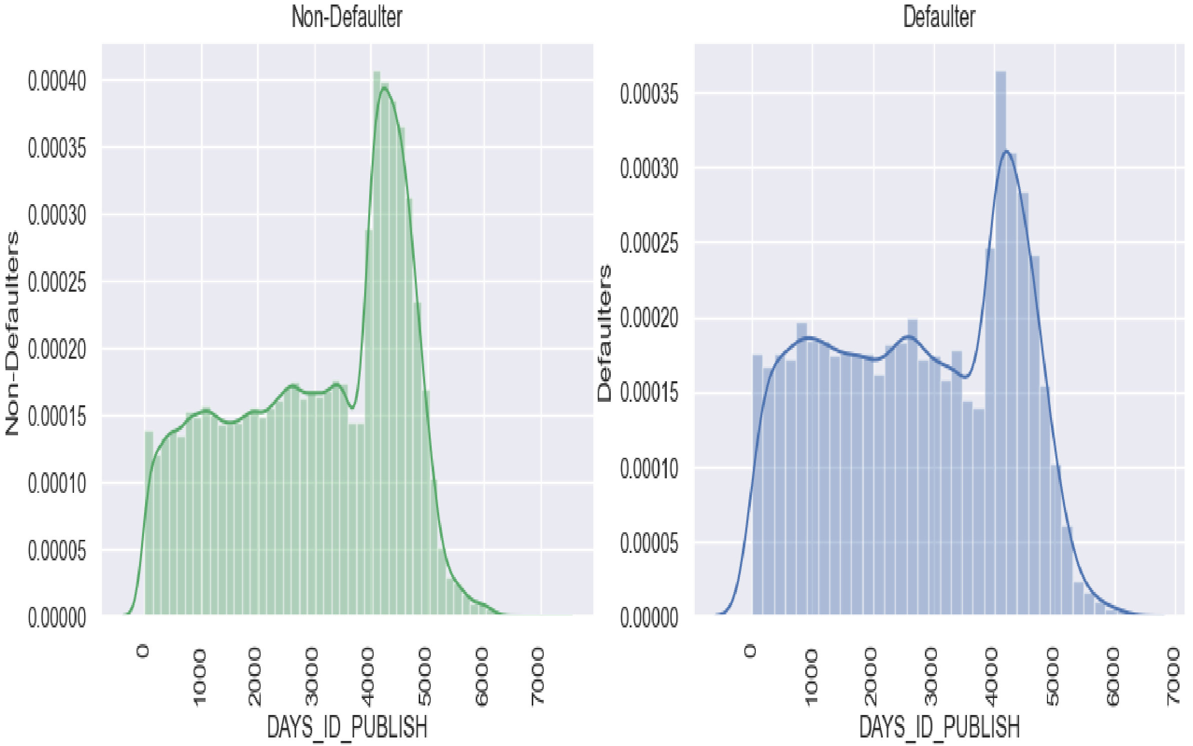
3. **REGION_POPULATION_RELATIVE** : This column shows normalized population of region where client lives



Insights:

Both non-defaulters and defaulters distribution is not normal.

4. **DAYS_ID_PUBLISH**: This column shows How many days before the application did client change the identity document with which he applied for the loan

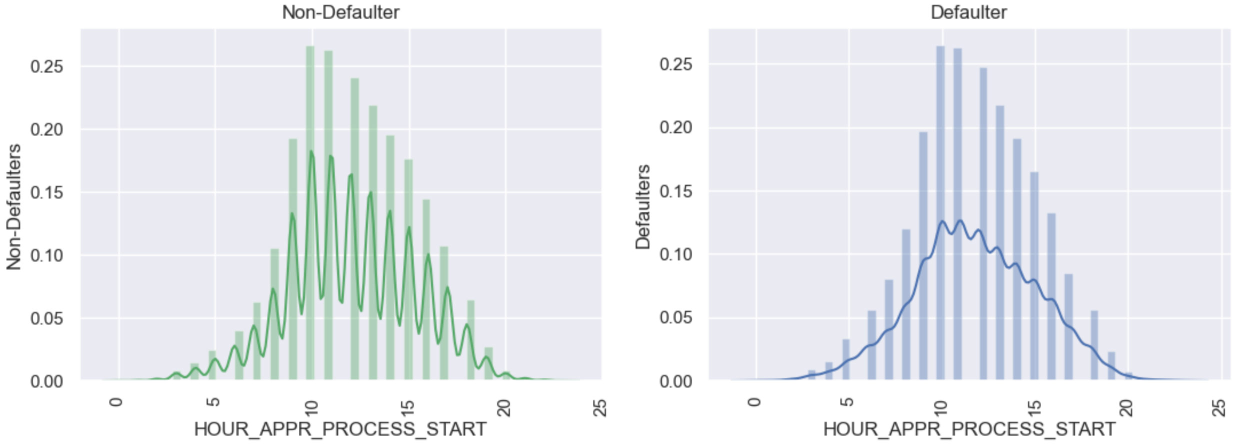


Insights:

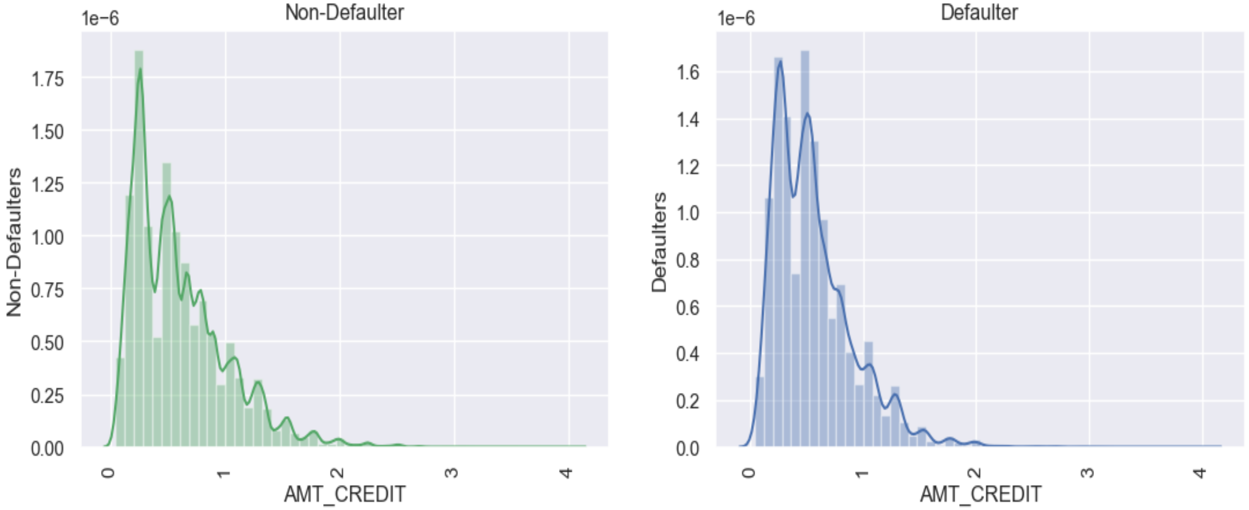
- Both non-defaulters and defaulters have similar shape.

Univariate Analysis: Numerical types

5. **REGION_POPULATION_RELATIVE** : This column shows normalized population of region where client lives



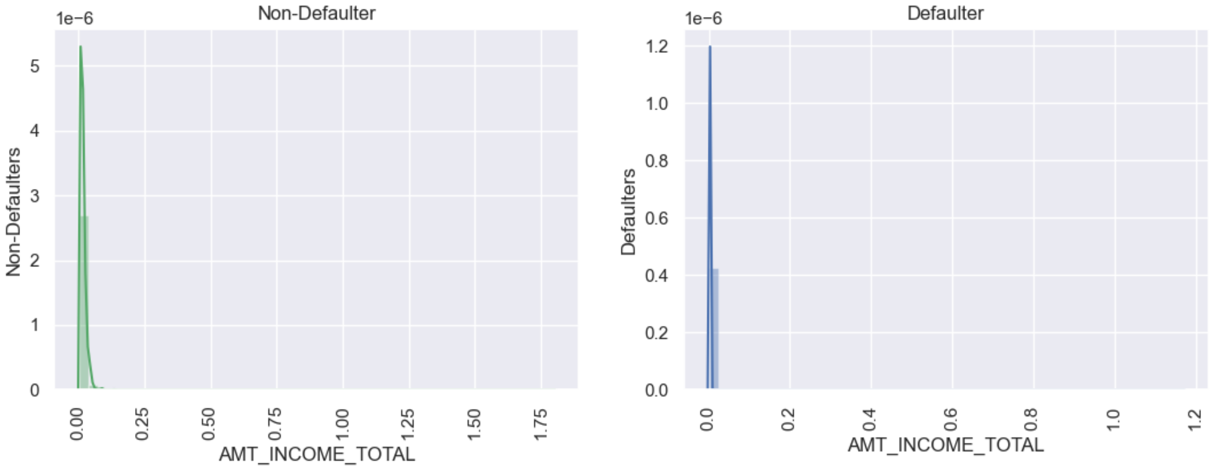
6. **AMT_CREDIT**: It shows credit amount of the loan.



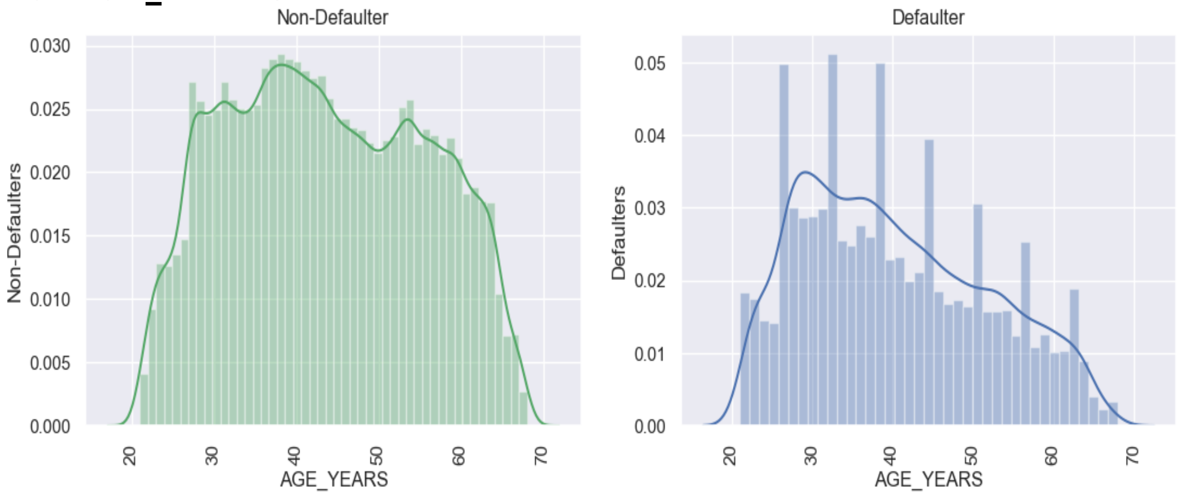
Insights:

Both non-defaulters and defaulters distribution is not normal.

6. **AMT_INCOME_TOTAL**: It shows income of client.



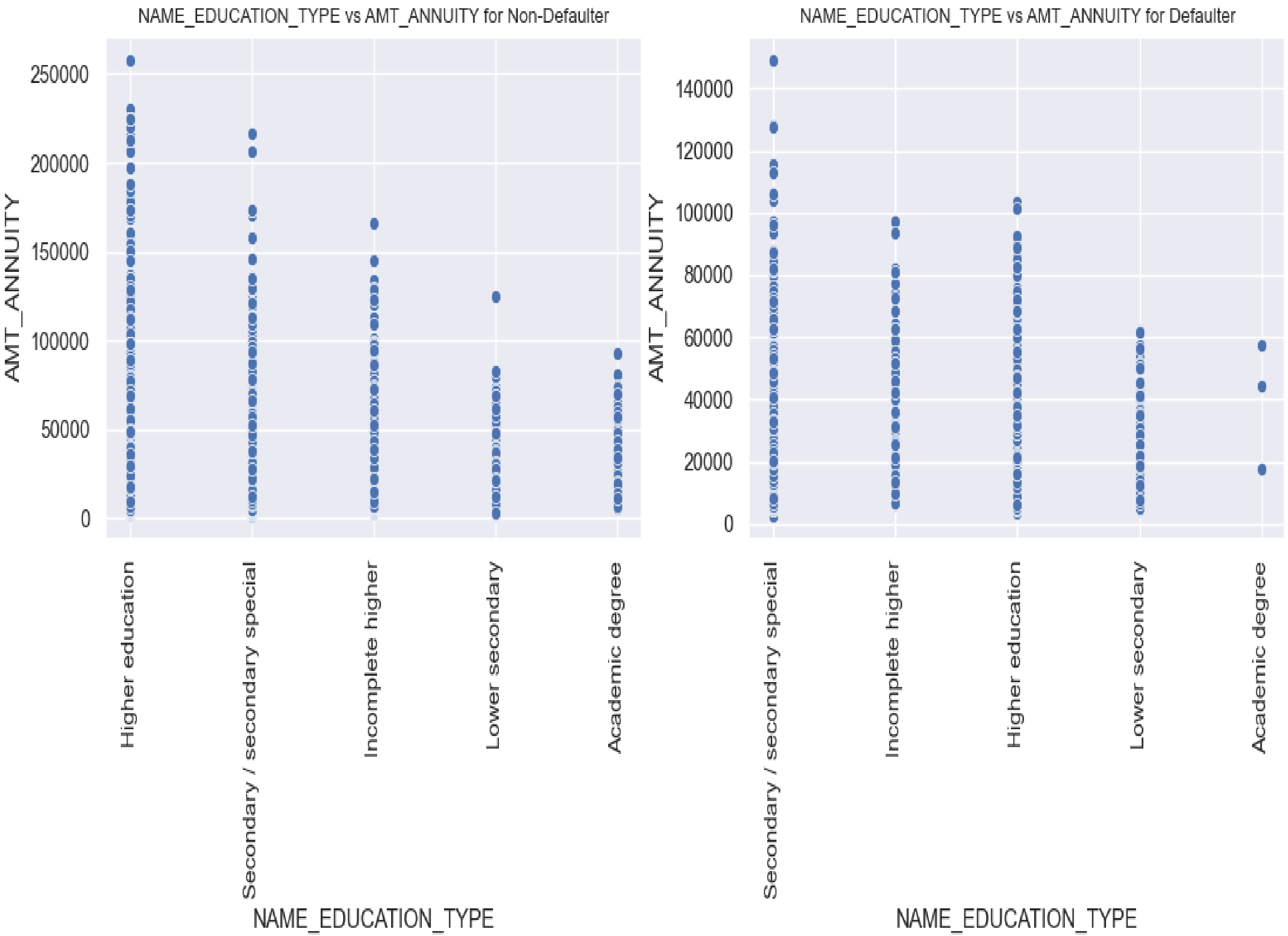
6. **AGE_YEARS**



Insights: Both non-defaulters and defaulters have similar shape. But in age in years is wide in non-defaulters but irt is not in defaulters chart.

Bivariate Analysis: Numerical and Categorical Columns

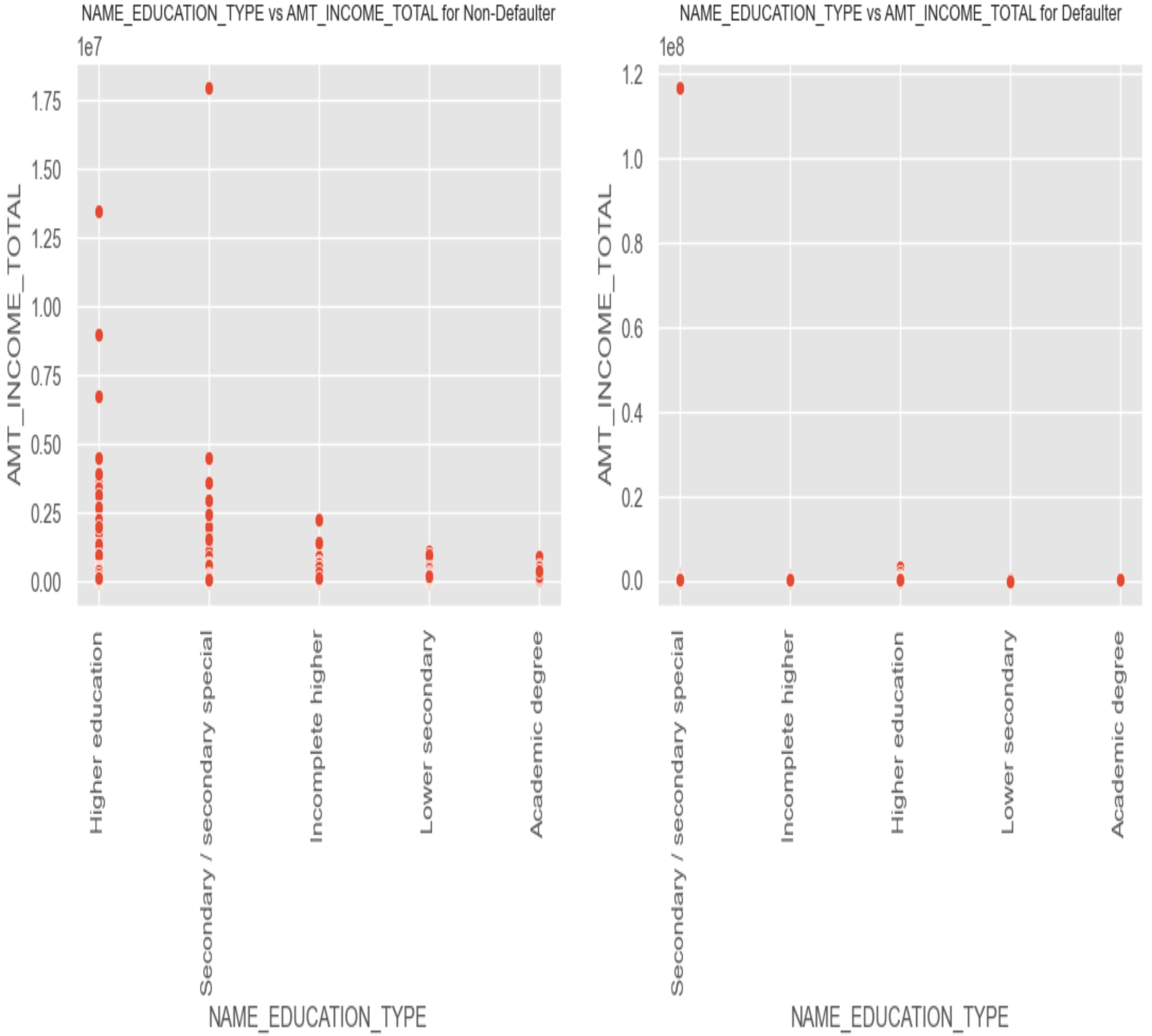
1. AMT_ANNUITY VS NAME_EDUCATION_TYPE



Insights:

Both non-defaulters and defaulters distribution is not normal.

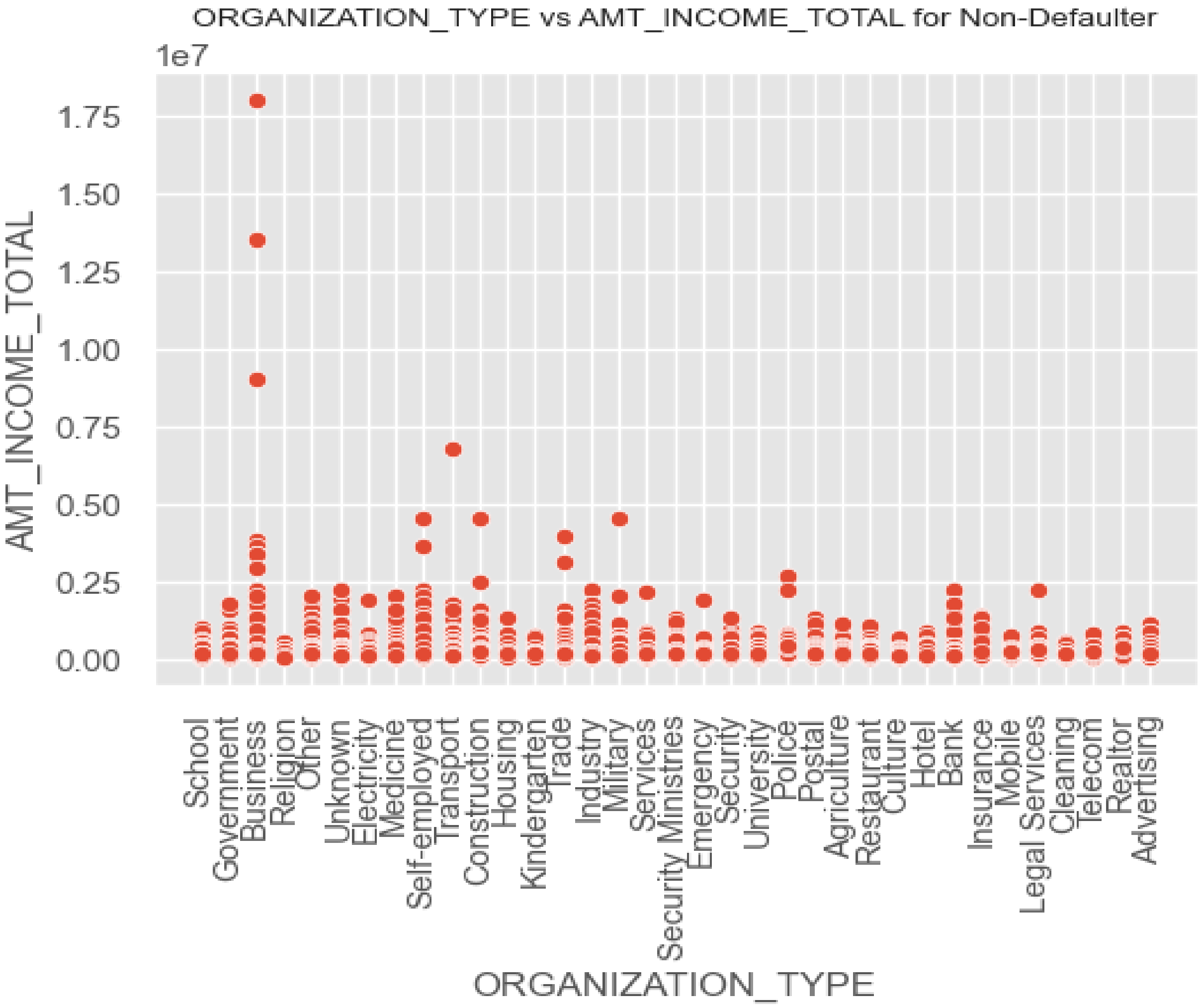
2. NAME_EDUCATION_TYPE VS AMT_INCOME_TOTAL

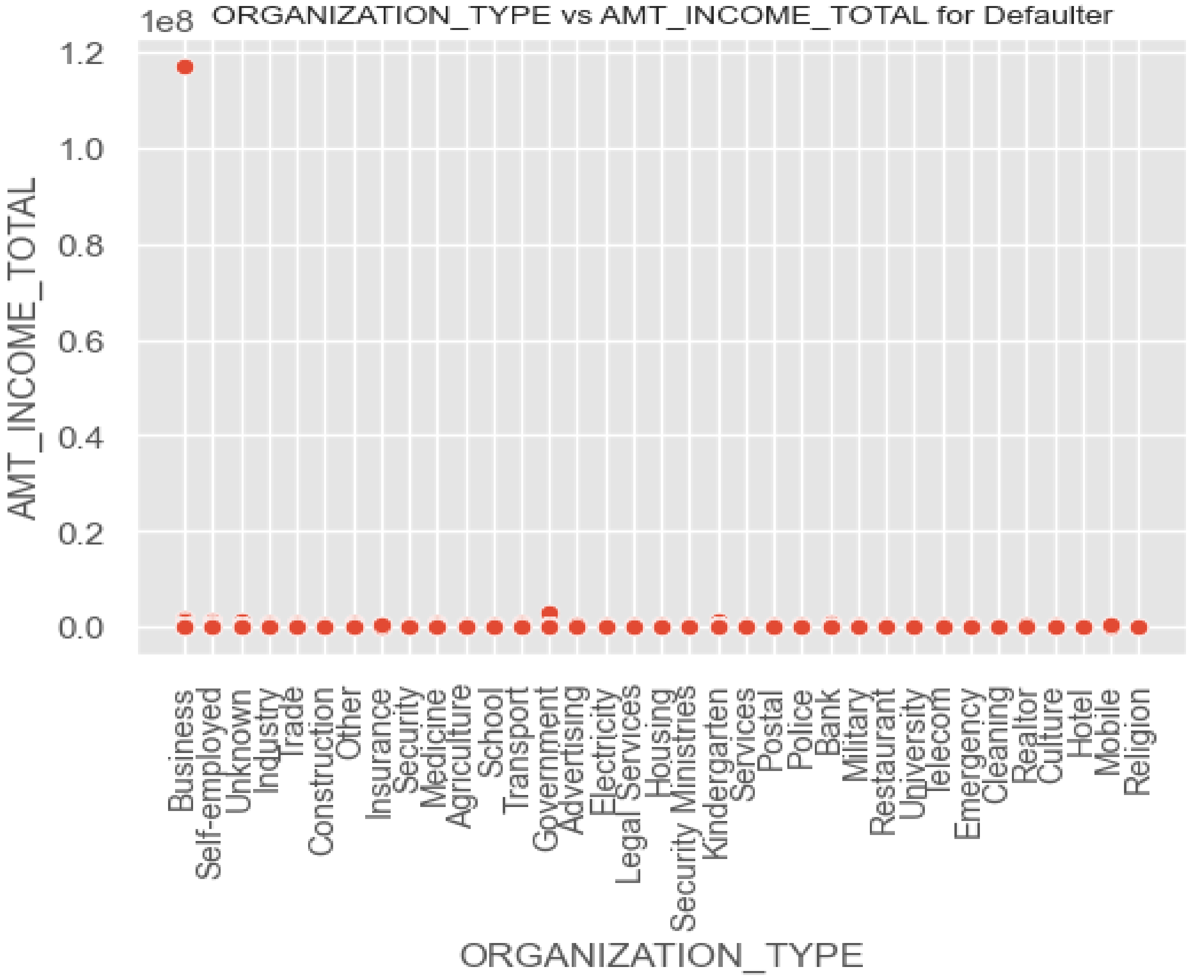


Insights: Both non-defaulters and defaulters have similar shape. But in age in years is wide in non-defaulters but irt i
not in defaulters chart.

Bivariate Analysis: Numerical and Categorical Columns

3. **AMT_INCOME_TOTAL VS ORGANIZATION_TYPE**

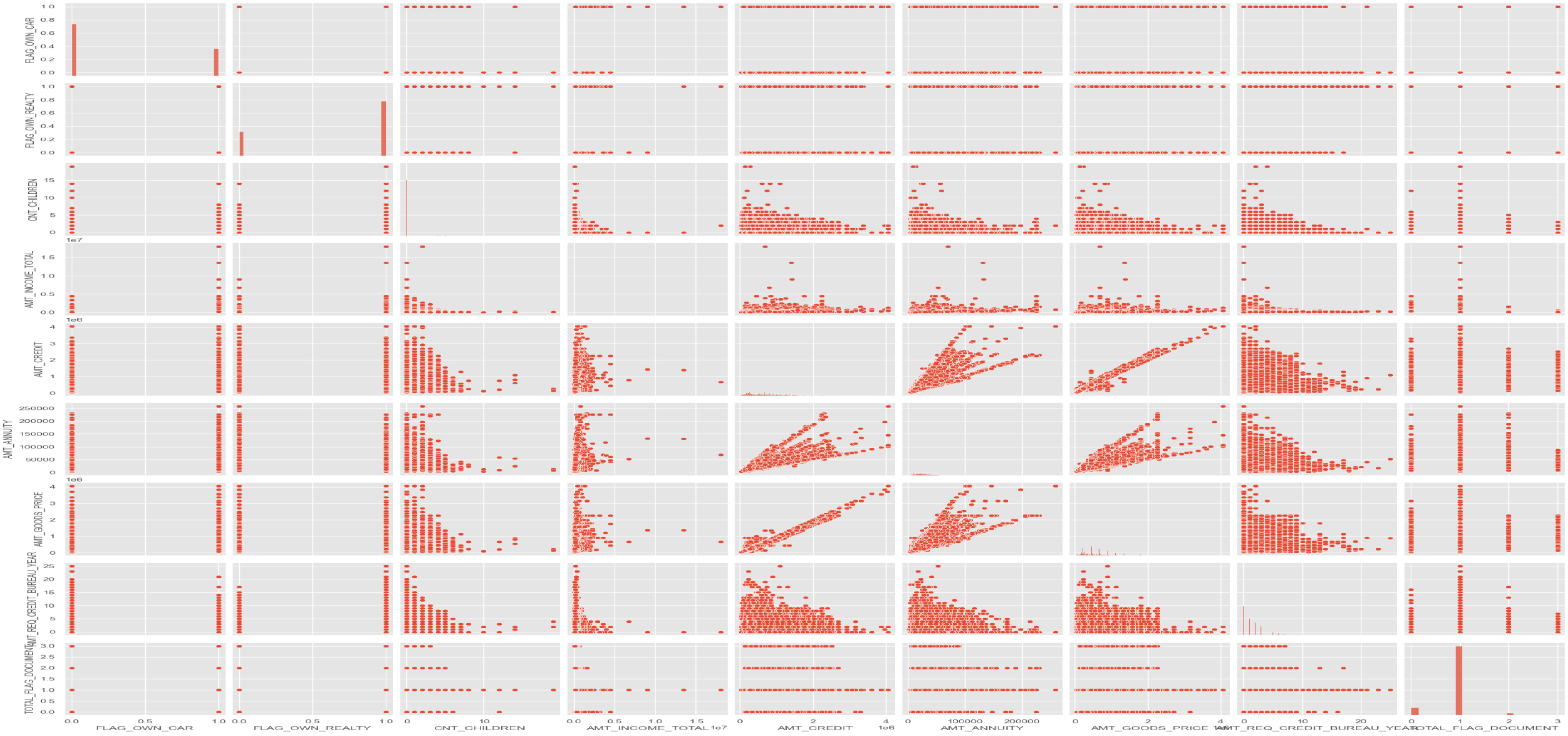

ORGANIZATION_TYPE vs AMT_INCOME_TOTAL for Non-Defaulter


ORGANIZATION_TYPE vs AMT_INCOME_TOTAL for Defaulter

Insights:

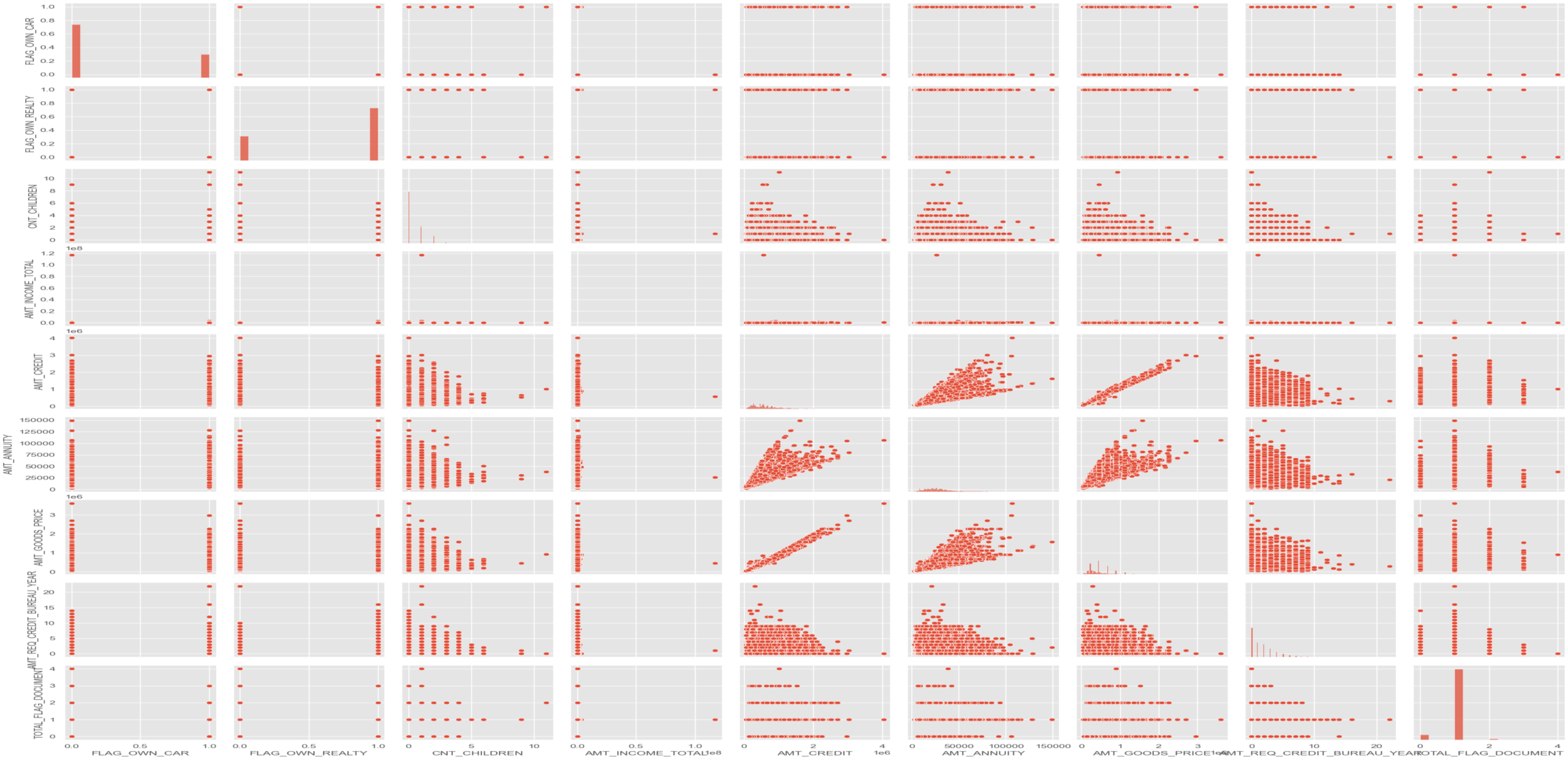
Both non-defaulters and defaulters distribution is not normal.

Pair-plot for non-Defaulters



Insights: AMT_CREDIT, AMT_ANNUITY and GOODS_PRICE are correlated to each other.

Pair-plot for Defaulters



Insights: AMT_CREDIT, AMT_ANNUITY and GOODS_PRICE are correlated to each other.