Exploratory Data Analysis Assignment

Problem Statement

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample, **All other cases:** All other cases when the payment is paid on time.

Problem Statement

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

Approved: The Company has approved loan Application

Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want.

Refused: The company had rejected the loan (because the client does not meet their requirements etc.).

Unused offer: Loan has been cancelled by the client but at different stages of the process.

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency to default.

Overall Approach

1. application_data.csv

- Importing data
- Understanding domain variables
- Changing datatypes of some columns
- Checking the structure of data
- Dealing with missing or null values
- Removed columns having more than 20% null values
- Removed less important columns
- Dealing with Outliers(not deleting them)
- Univariate analysis (Pie charts, bar charts, count plots, box plots etc.)
- Bivariate or Multivariate analysis (Scatter plot, pair plot, heatmap etc.)

Overall Approach

2. previous_application.csv

- Importing data
- Understanding domain variables
- Checking the structure of data
- Dealing with missing or null values
- Removed column having more than 20% null values
- Dealing with Outliers
- Univariate analysis (Pie charts, bar charts, count plots, box plots etc.)
- Bivariate or Multivariate analysis (Scatter plot, pair plot, heatmap etc.)

3. Merged file of both csv files

- Merged using inner join based on SK_ID_CURR
- Studying the merged file
- Data visualization on the merged file

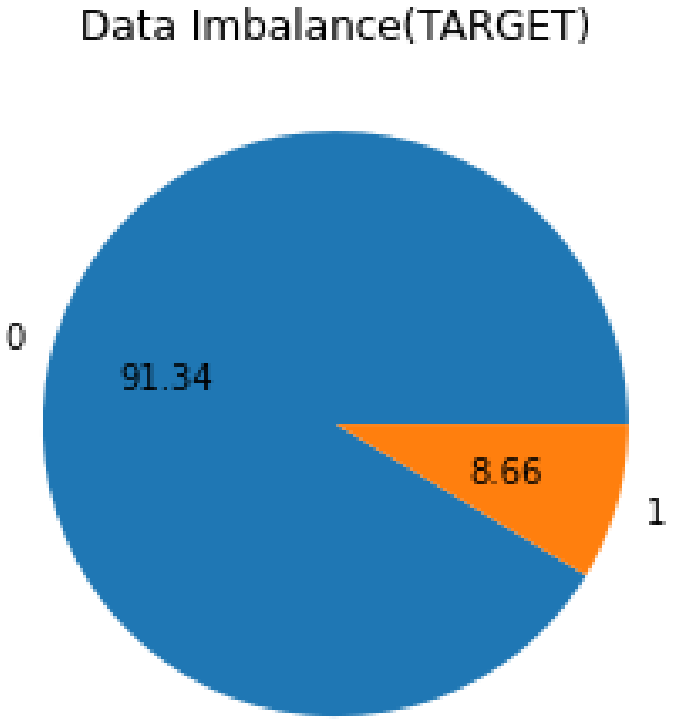
Data Imbalance

Data Imbalance with the TARGET variable

Non defaulters percentage: 91.34%

Defaulter percentage: 8.66%

This means very less percentage of applicants are defaulters.

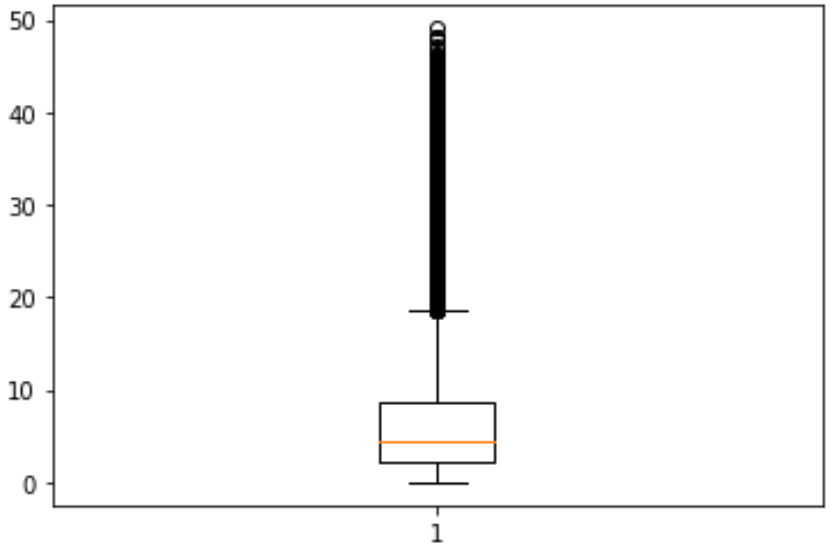


Relevant Results

- Removed XNA values in CODE_GENDER
- There is a DATA ERROR in the outliers of DAYS_EMPLOYED i.e. 55352 values are exactly same, showing 1000 years of employment i.e., 18% of the data.

```
b = final_data.DAYS_EMPLOYED/365
c = b[b>0]
c
```

```
8         1000.665753
11        1000.665753
23        1000.665753
38        1000.665753
43        1000.665753
             ...
307469    1000.665753
307483    1000.665753
307487    1000.665753
307505    1000.665753
307507    1000.665753
Name: DAYS_EMPLOYED, Length: 55352, dtype: float64
```



Relevant Results(Null values Analysis)

- Removed columns having more than 20% null values

```
Cols_to_drop = a[a > 20].to_frame().index.tolist()
Cols_to_drop
```

- Removed columns which are no much important for the analysis

```
final_data.shape
```

```
(307216, 30)
```

- Removed the remaining null values

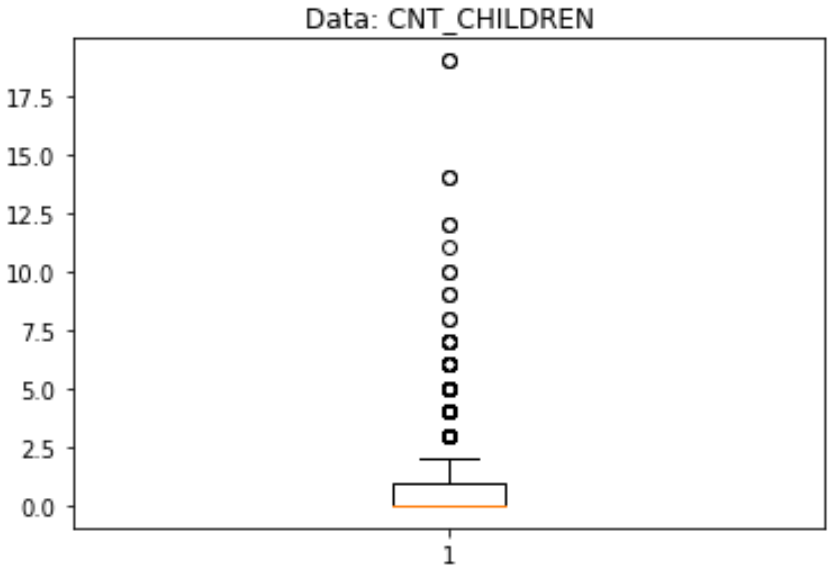
```
final_data.dropna(inplace = True)
```

```
final_data.shape
```

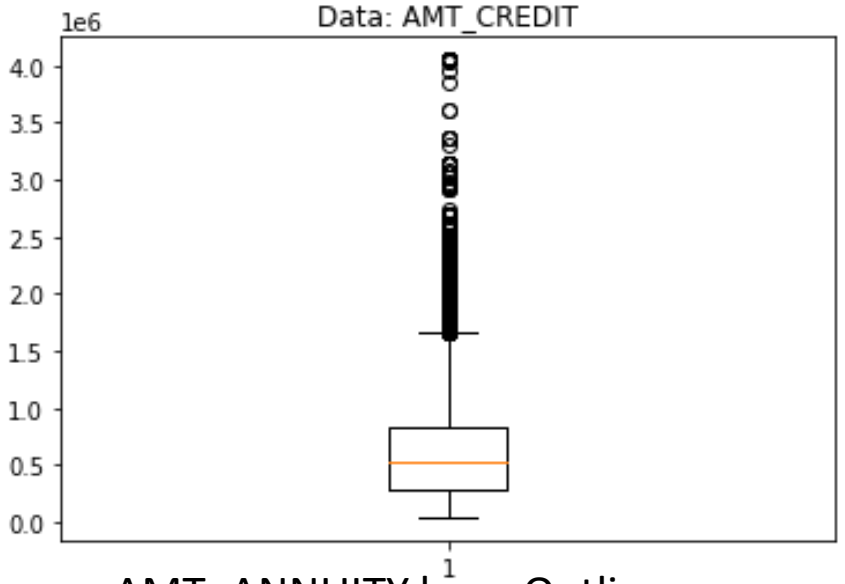
```
(251864, 29)
```

Relevant Results(Outlier Analysis)

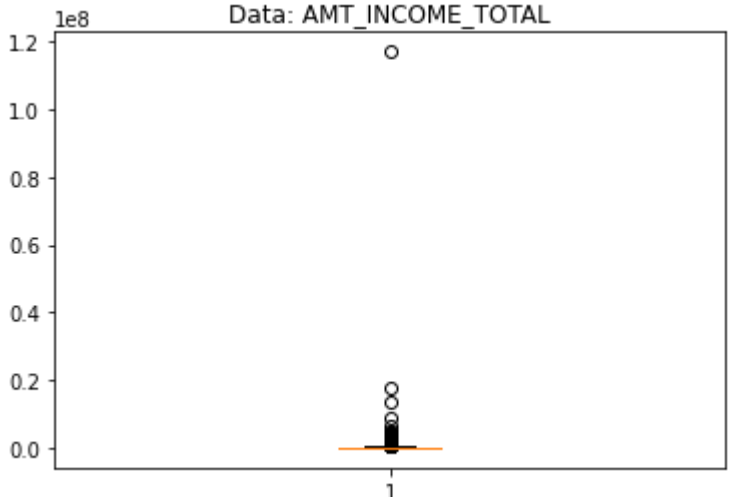
- There are many outliers in CNT_CHILDREN



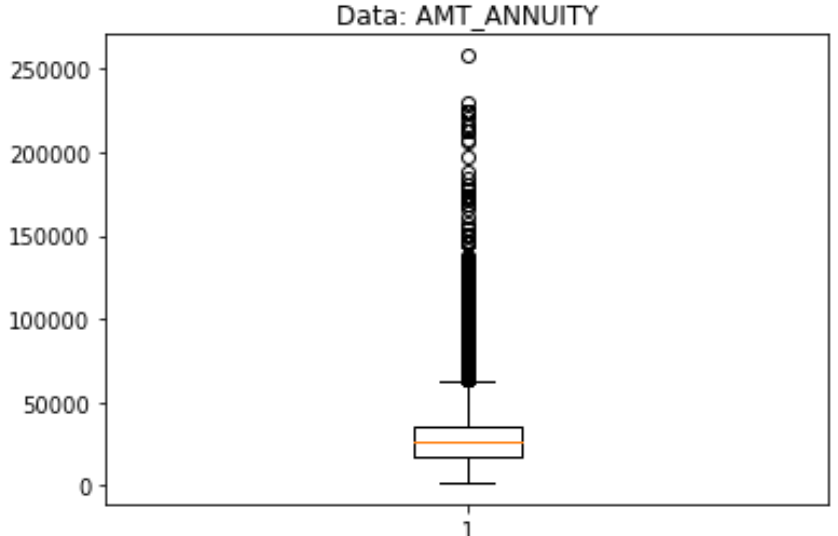
- AMT_CREDIT have Outliers.



- AMT_INCOME_TOTAL have Outliers.

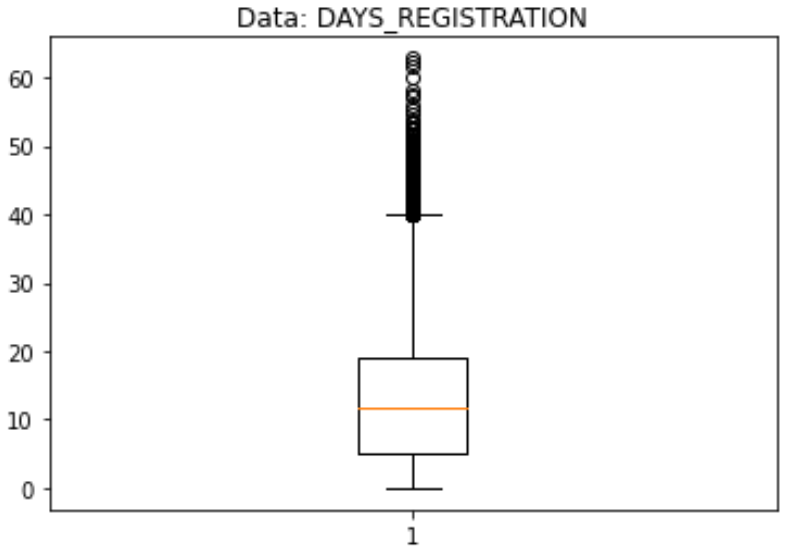


- AMT_ANNUITY have Outliers.

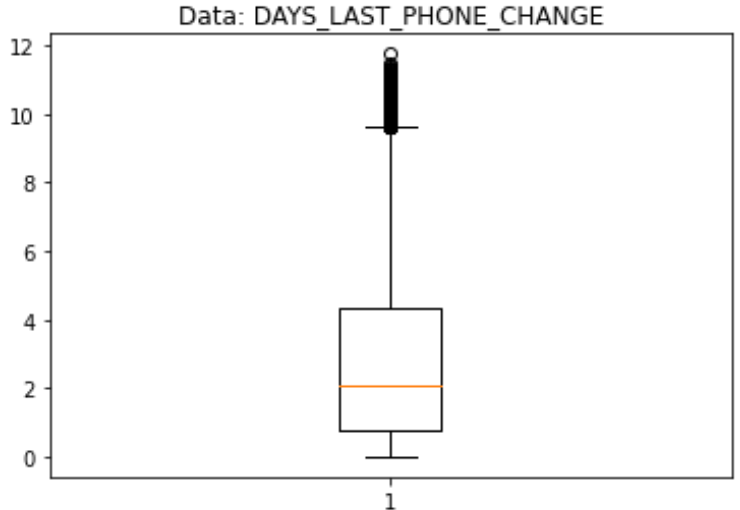


Relevant Results(Outlier Analysis)

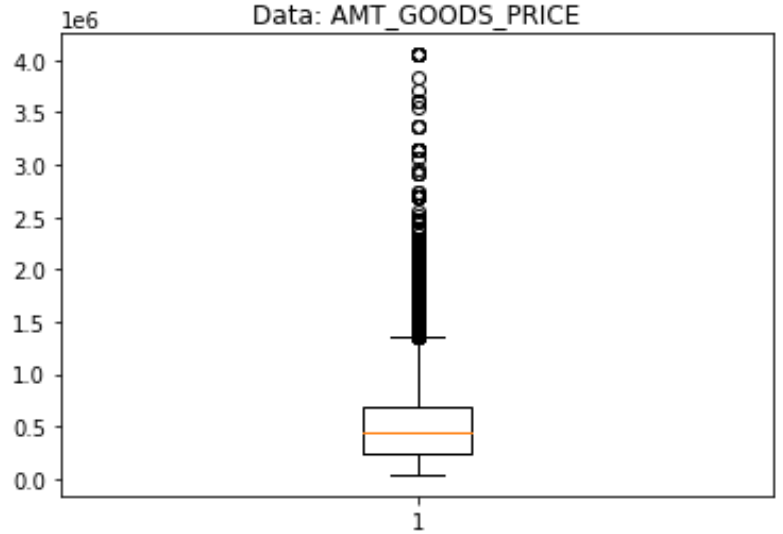
- DAYS_REGISTRATION have outliers.



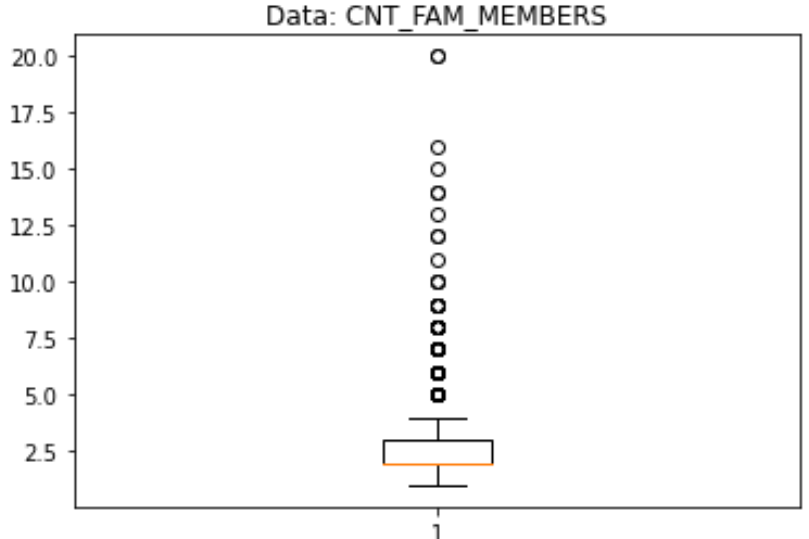
- DAYS_LAST_PHONE_EXCHANGE have Outliers



- AMT_GOODS_PRICE have outliers.

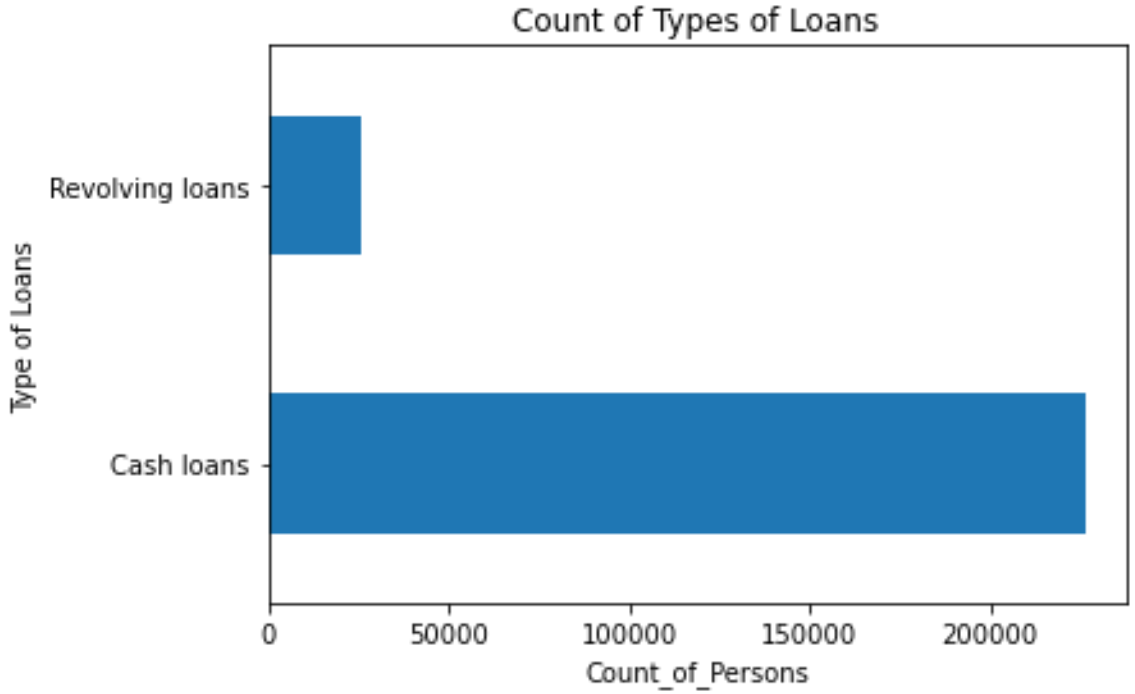


- CNT_FAMILY_MEMBERS have Outliers

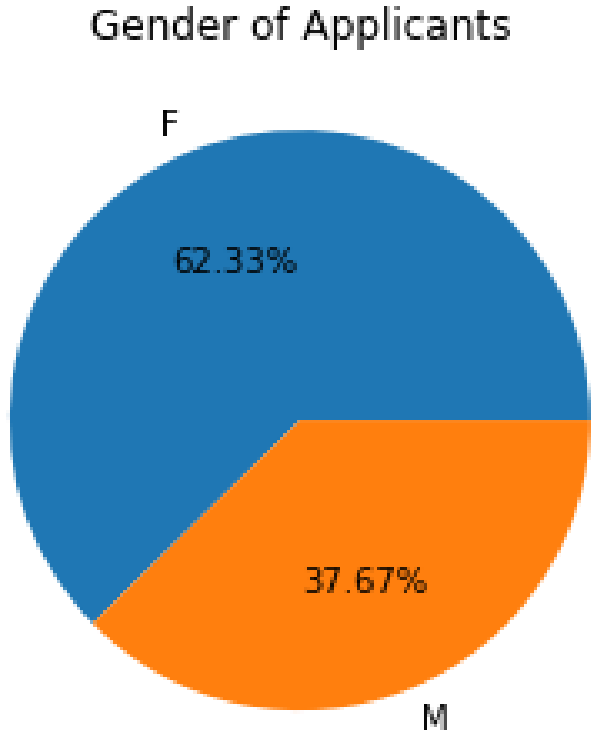


Relevant Results(**Univariate Analysis**)

- There are more cash loans comparative to Revolving Loans.

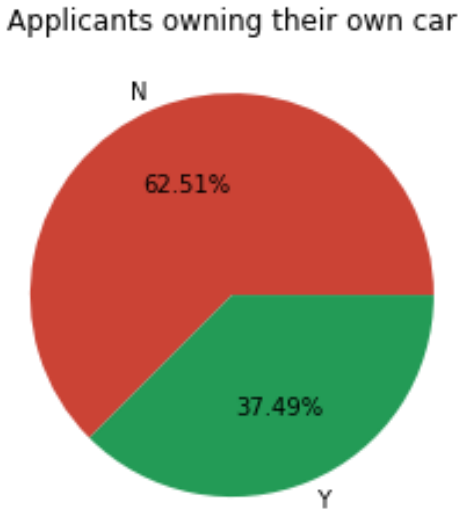


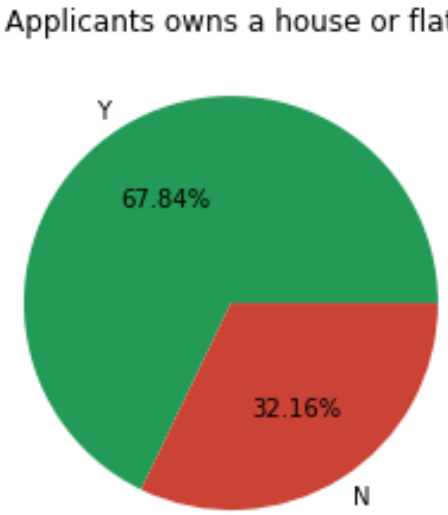
- There are more female applicants than male applicants.



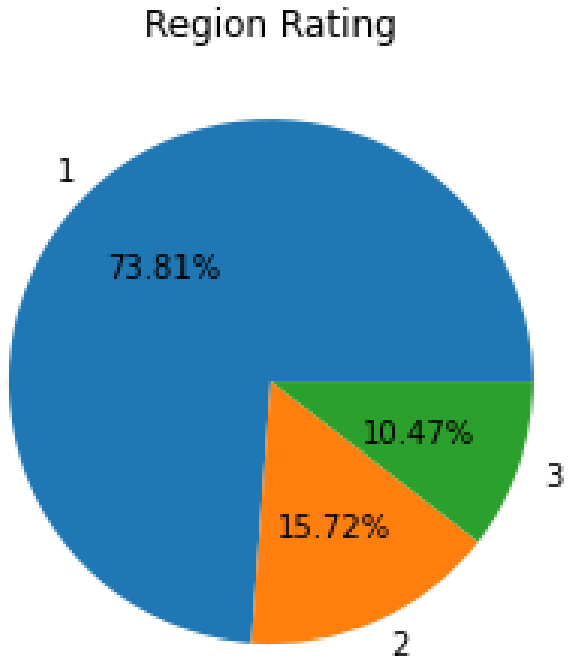
Relevant Results(**Univariate Analysis**)

- There are less applicants who owns their car.
- There are more applicants who own their house/flat.



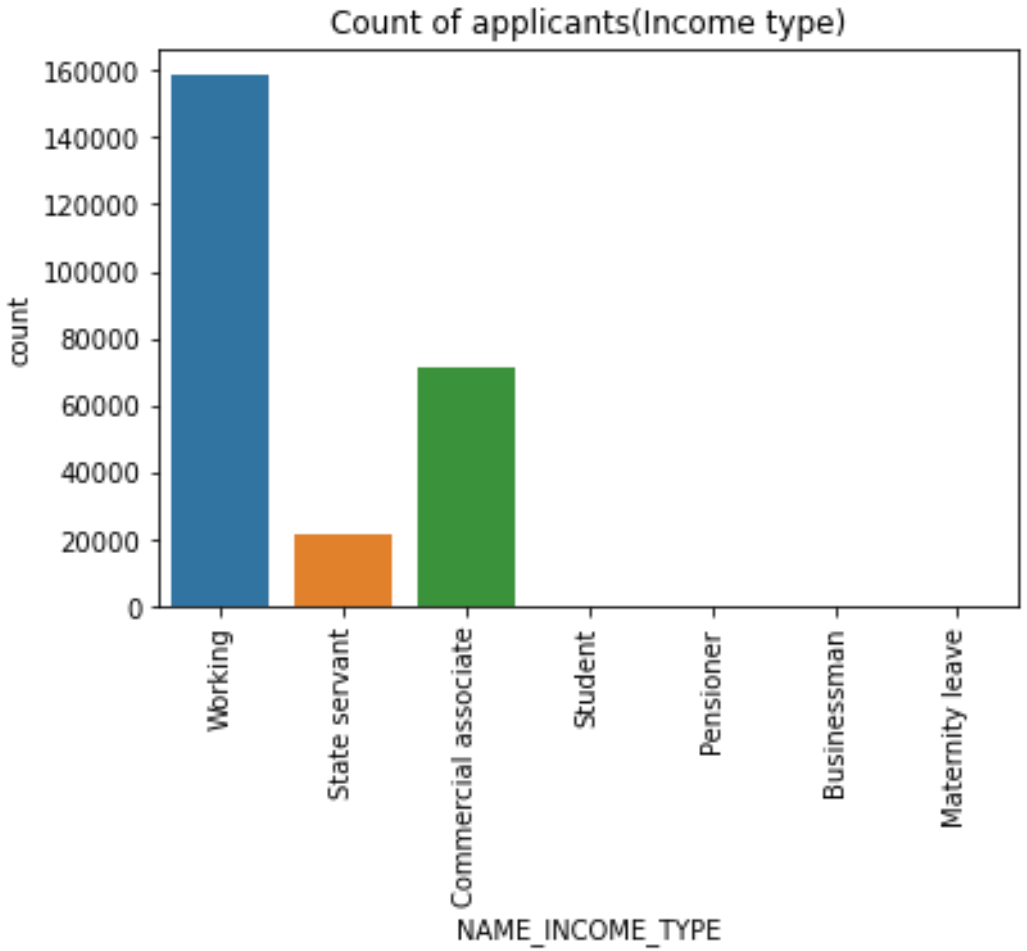


- Region rating 1 has maximum applicants.

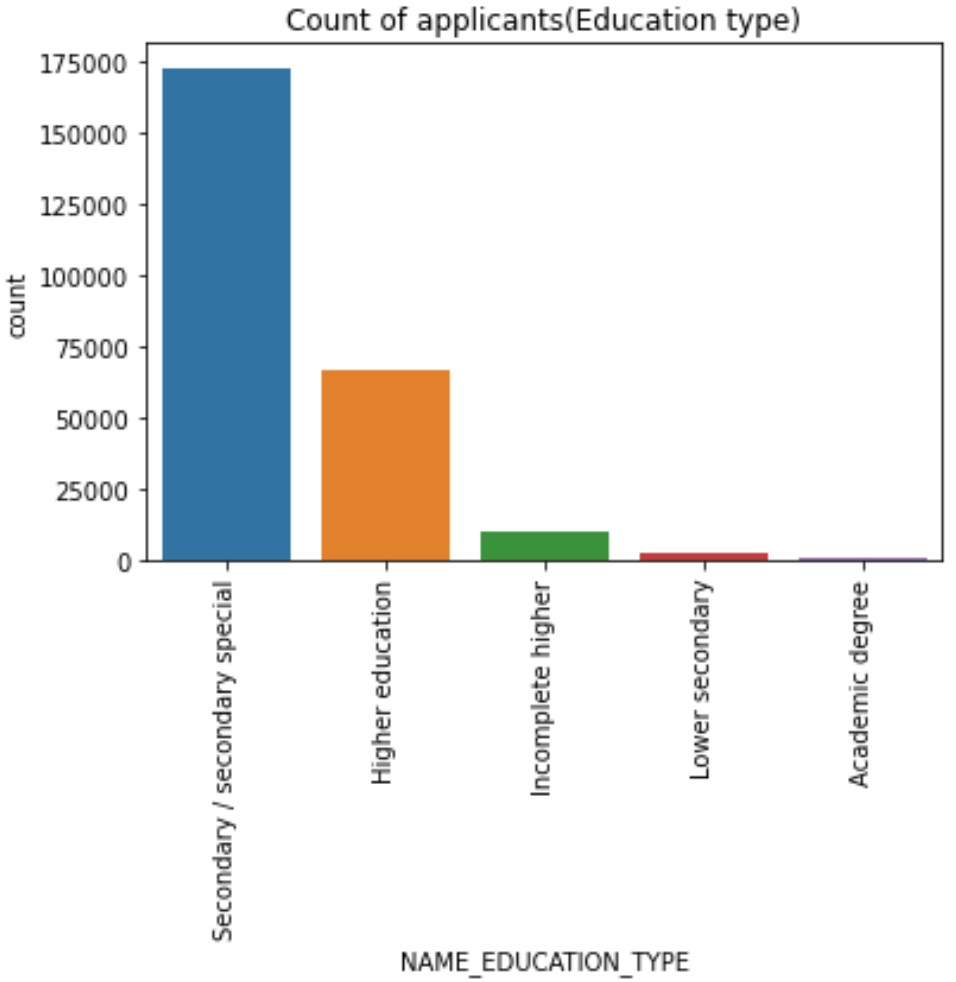


Relevant Results(**Univariate Analysis**)

- Most of the applicants are working professionals.

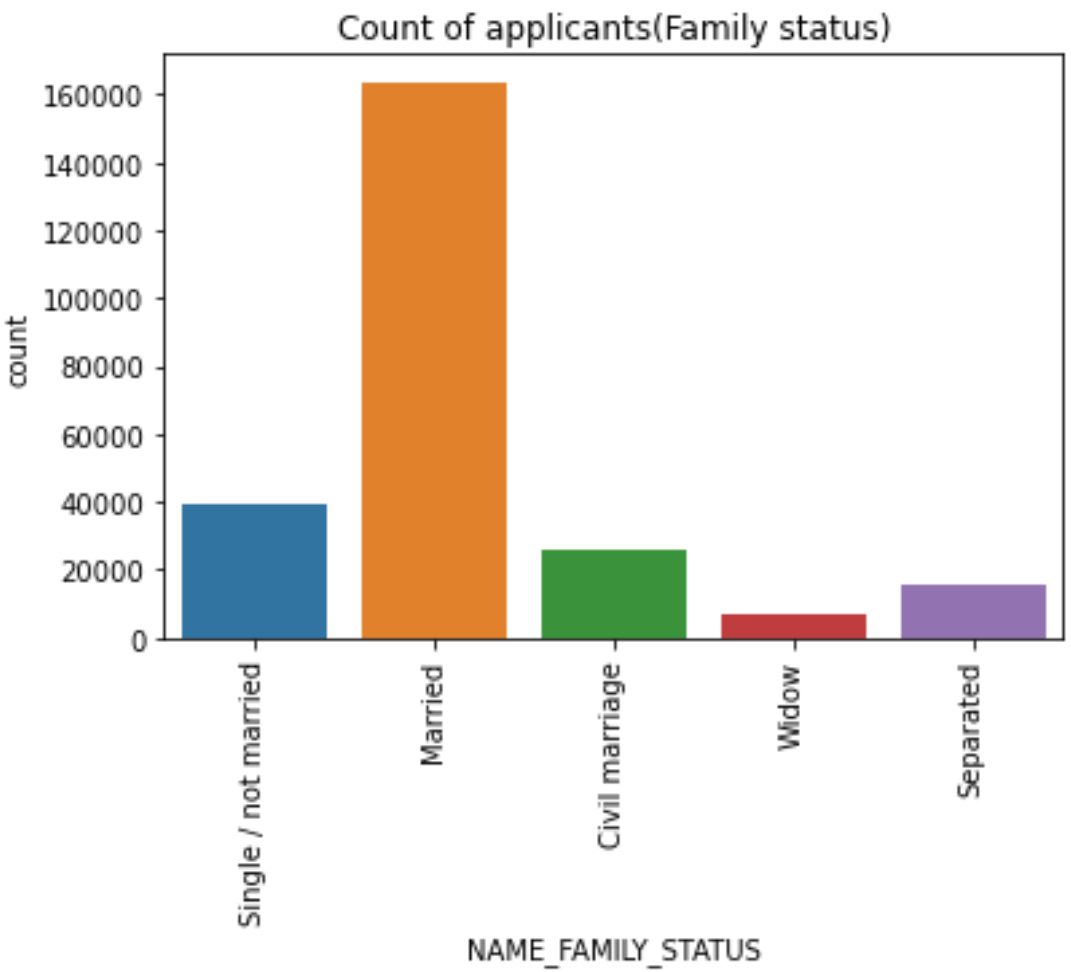


- Most of the applicants have secondary education as highest education.

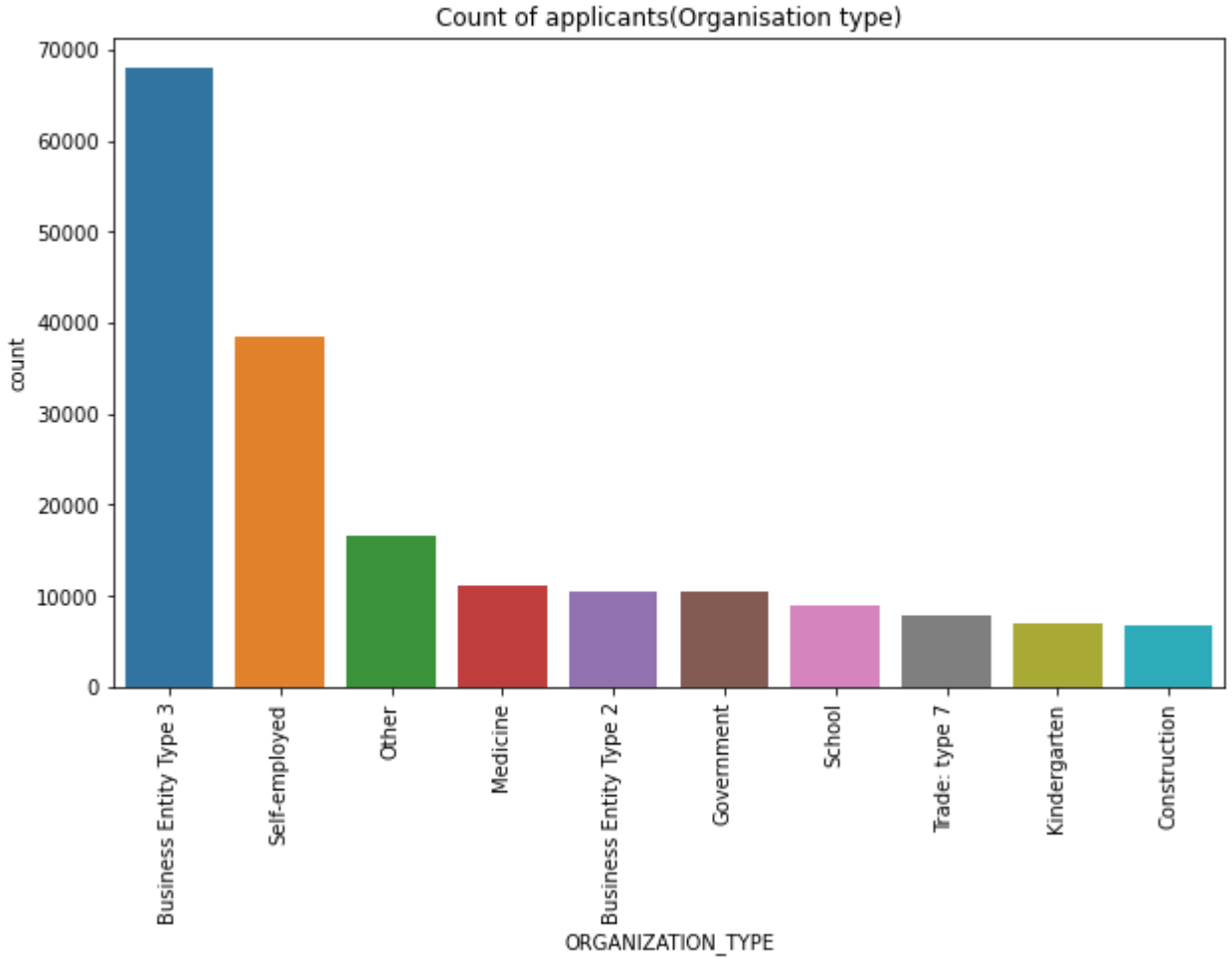


Relevant Results(**Univariate Analysis**)

- Most of the applicants who have taken loans are married

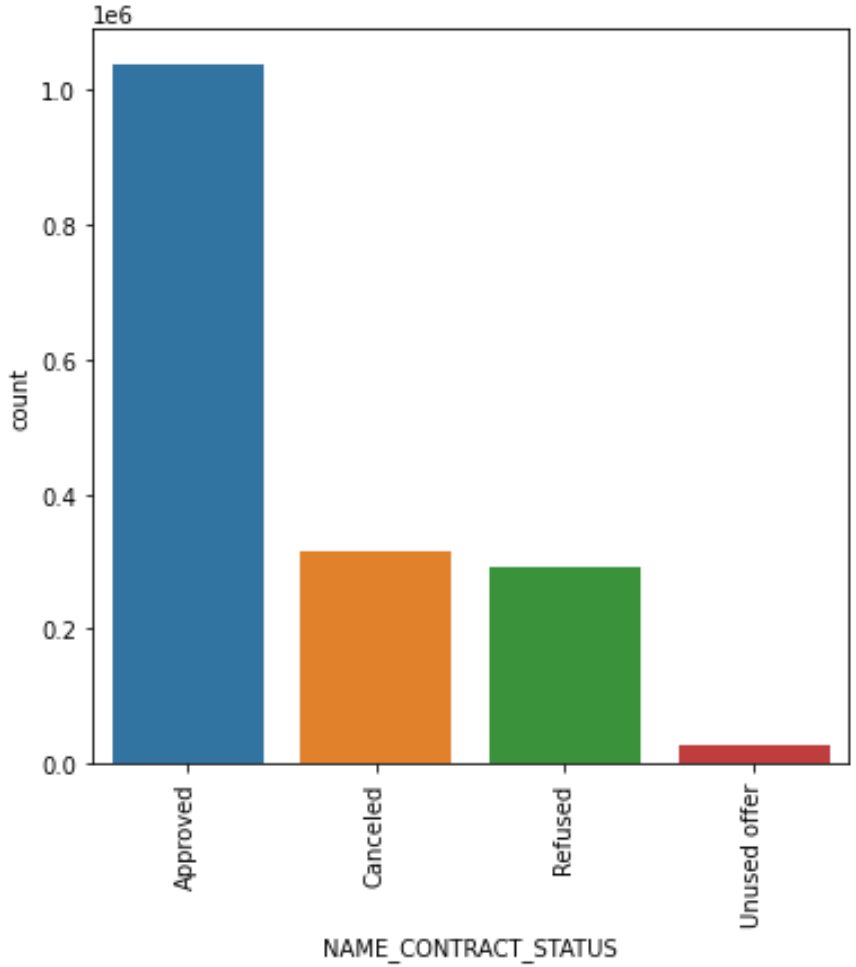


- Most of the applicants have business Entity type 3 as their organization.

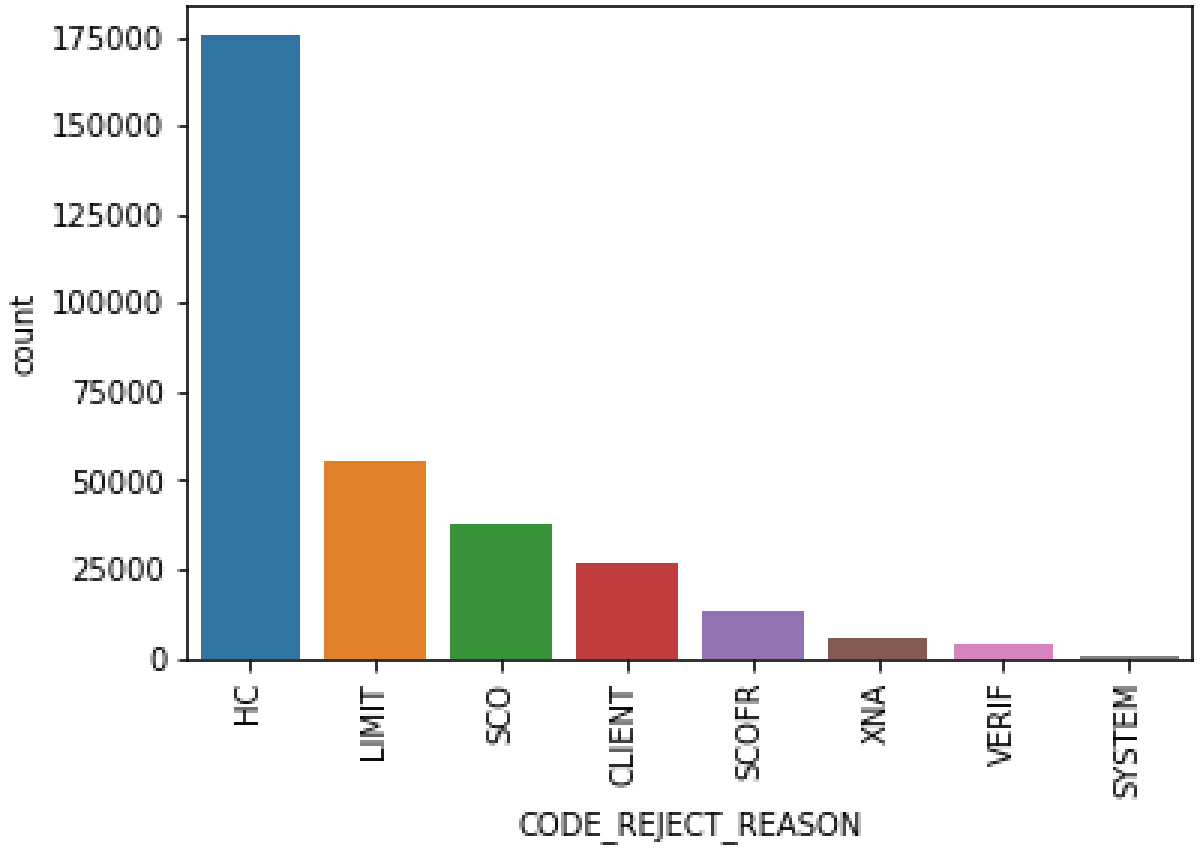


Relevant Results(**Univariate Analysis- prev_data**)

- Most of the applicants have approved loans

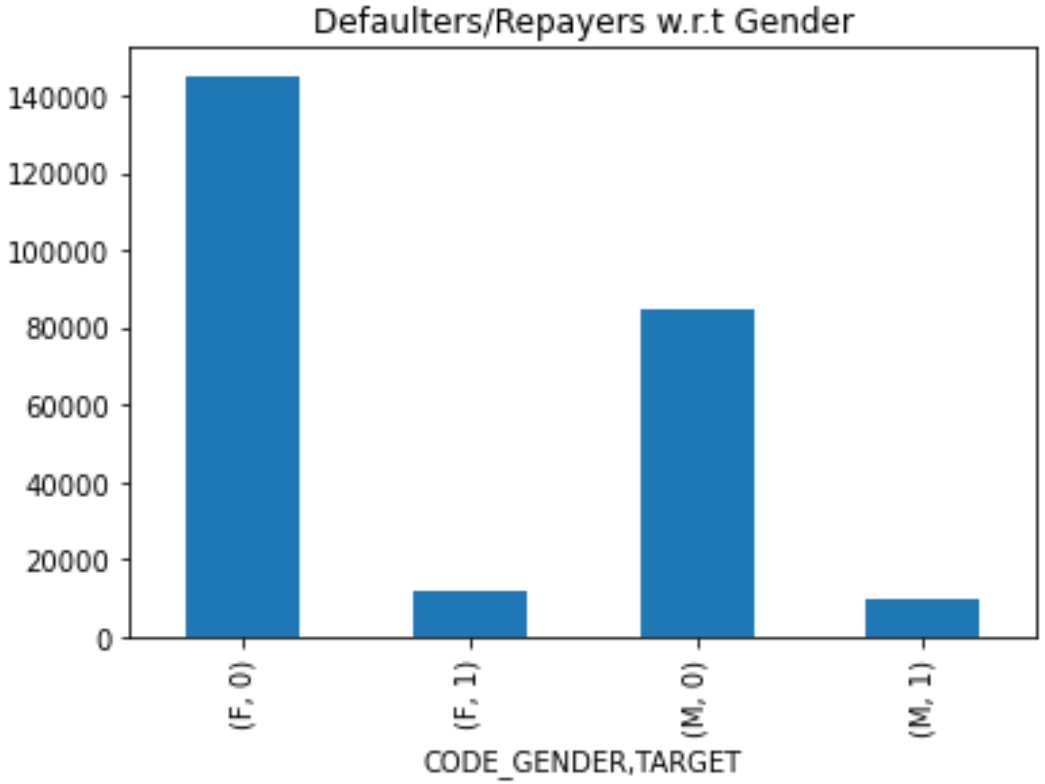


- Most of the applicants have reason of rejection as HC.

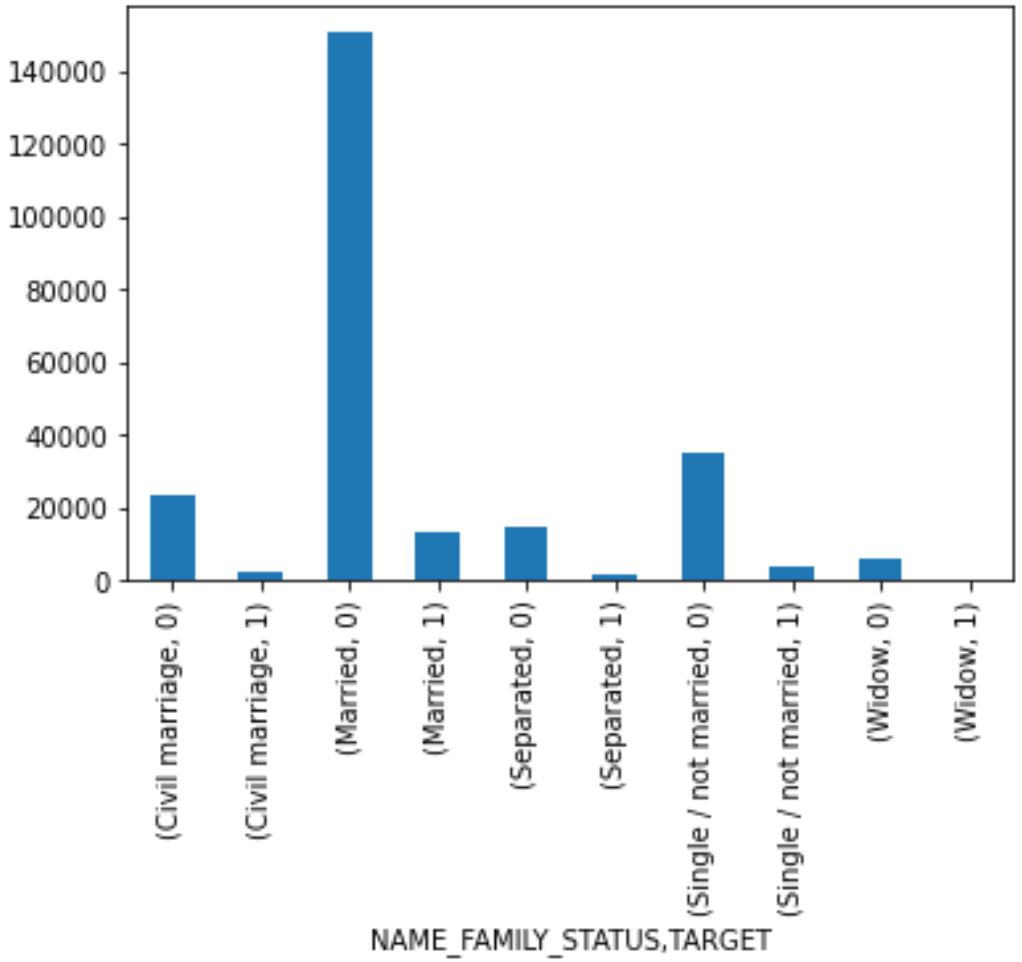


Relevant Results(**Bivariate/Multivariate Analysis**)

- Females are more defaulters as compared to males.

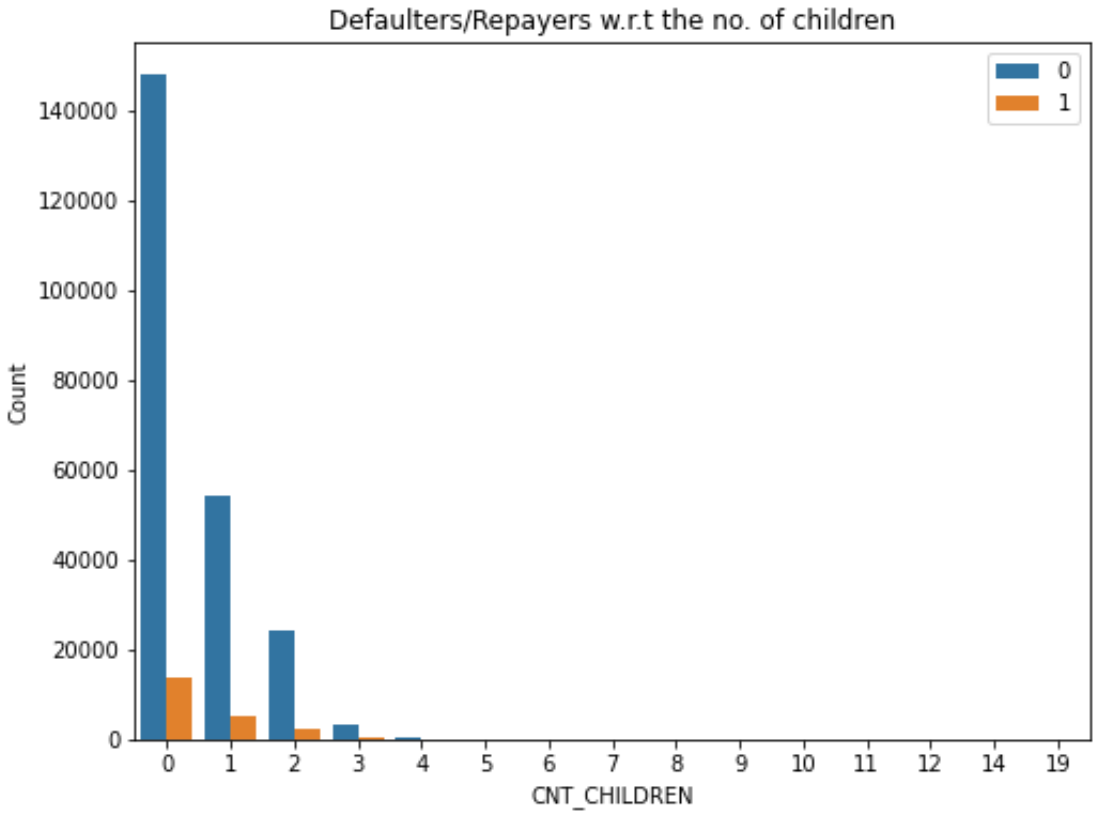


- Civil marriages and singles are more likely to be defaulters.

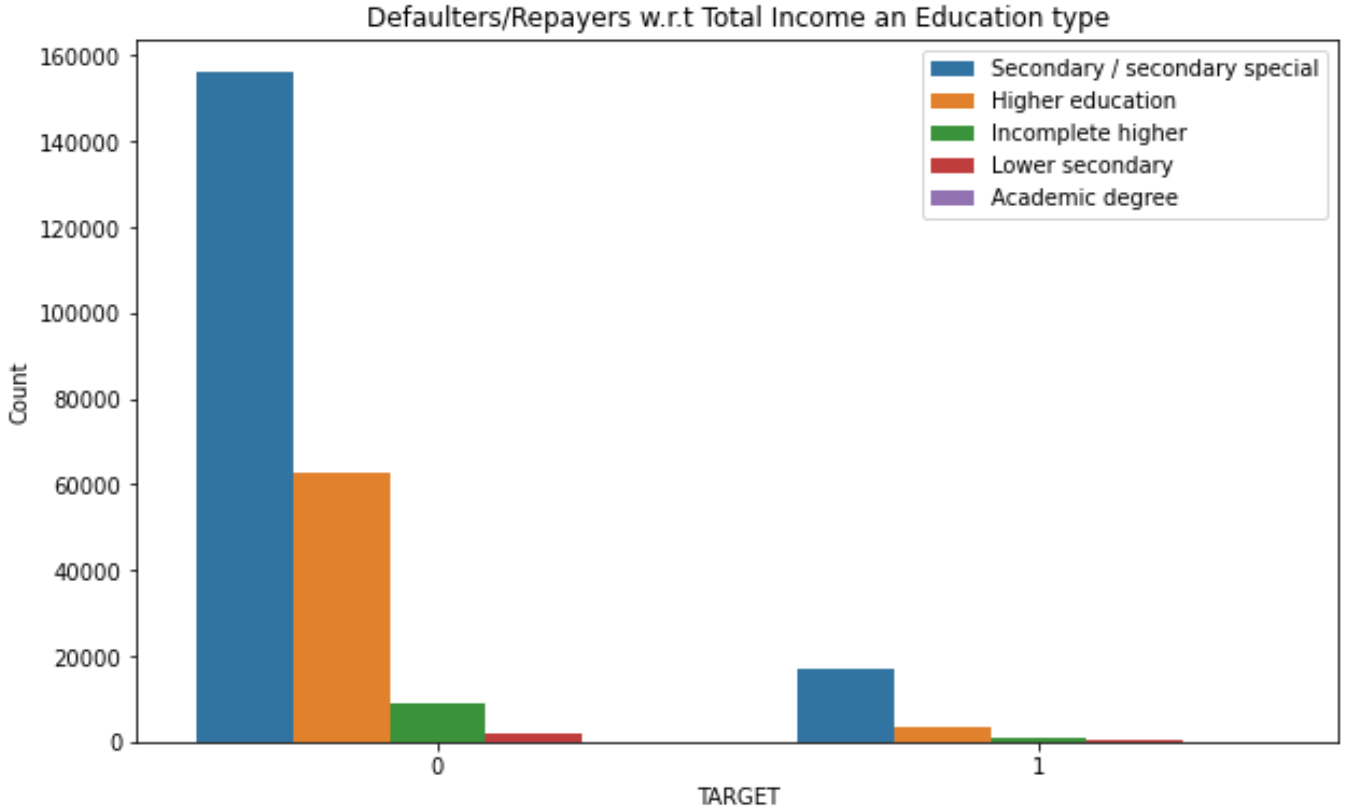


Relevant Results(**Bivariate/Multivariate Analysis**)

- Applicants having 0-2 children are more likely to be repayers.

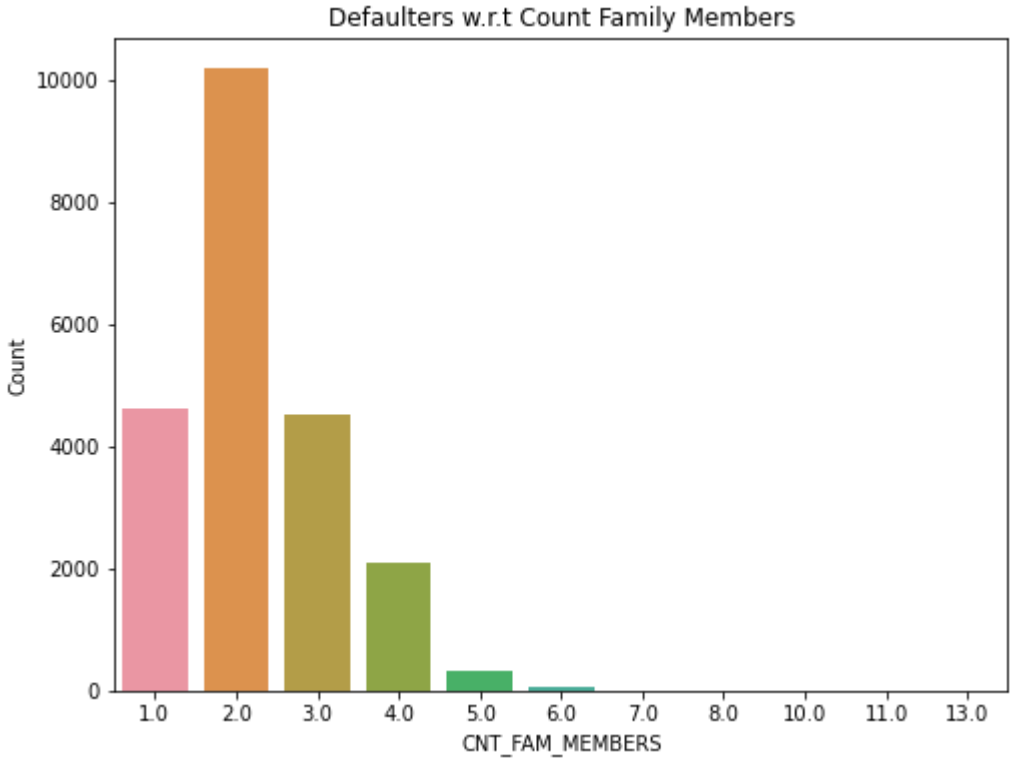


- Most of the defaulters have secondary degree and in repayers too, most of the people are having secondary degree.

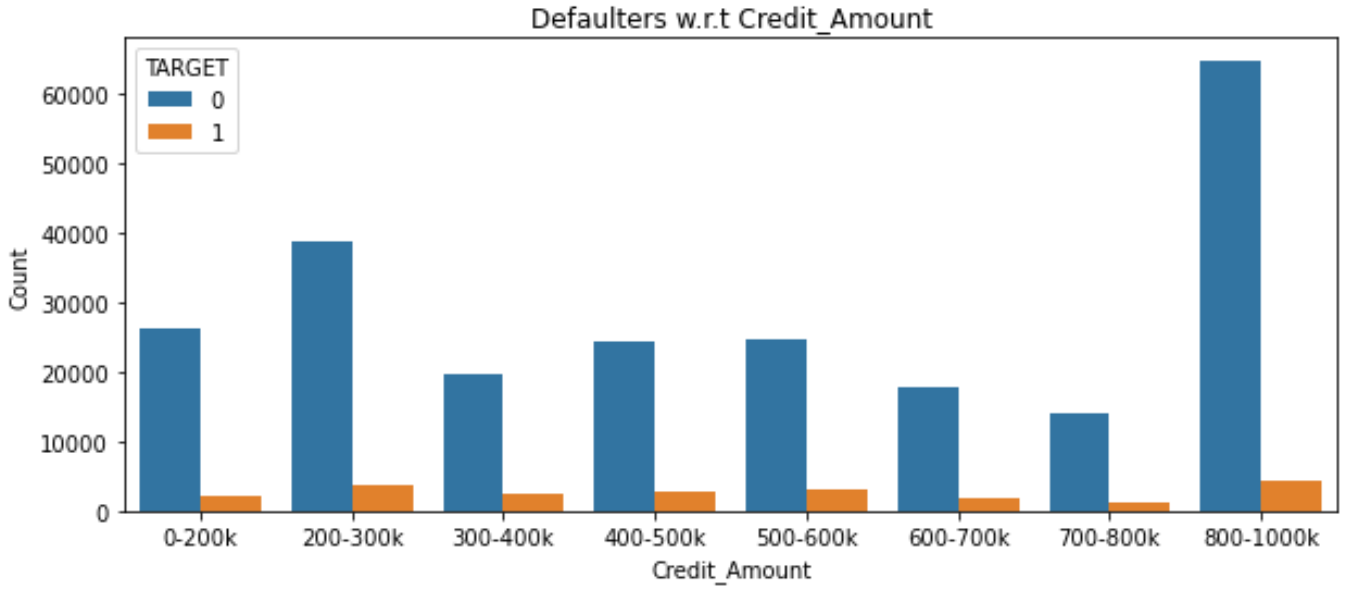


Relevant Results(**Bivariate/Multivariate Analysis**)

- Applicants having 2 family members are more likely to be defaulters.



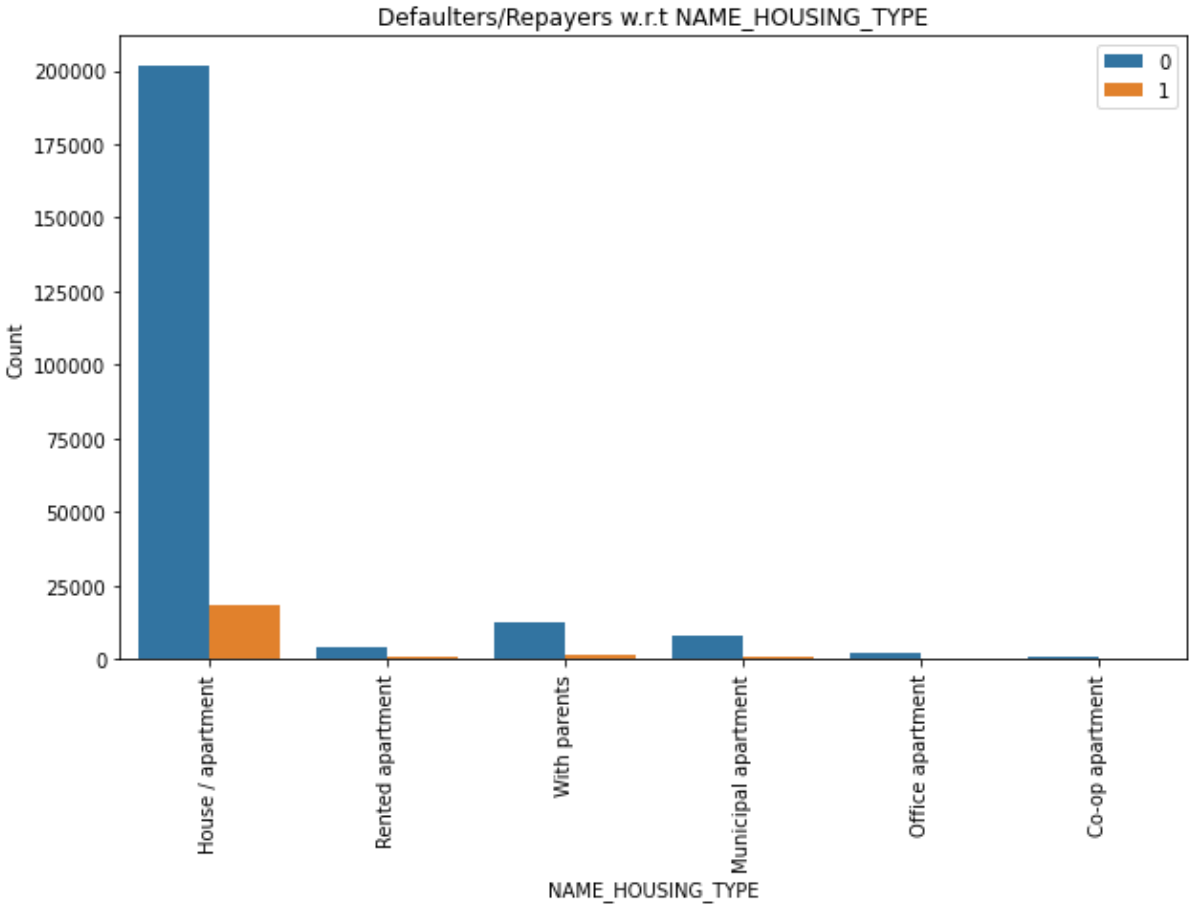
- Most of the defaulters lie in the range of Amount Credit 200k – 500k



Relevant Results(**Bivariate/Multivariate Analysis**)

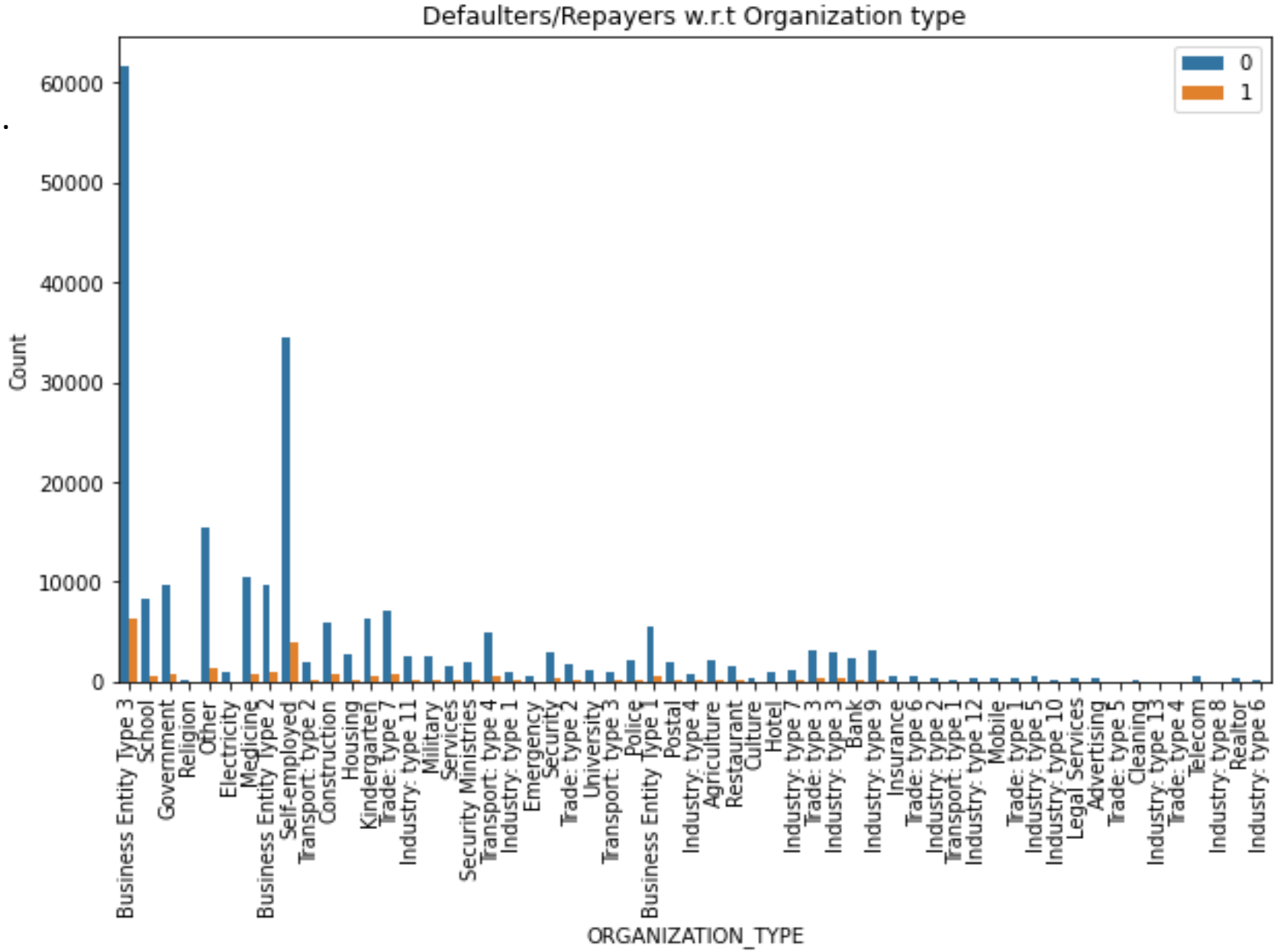
- We have more defaulters in REGION_RATING_CLIENT type 3.

- We have more defaulter count present in Rented apartments and living with their parents than own house.



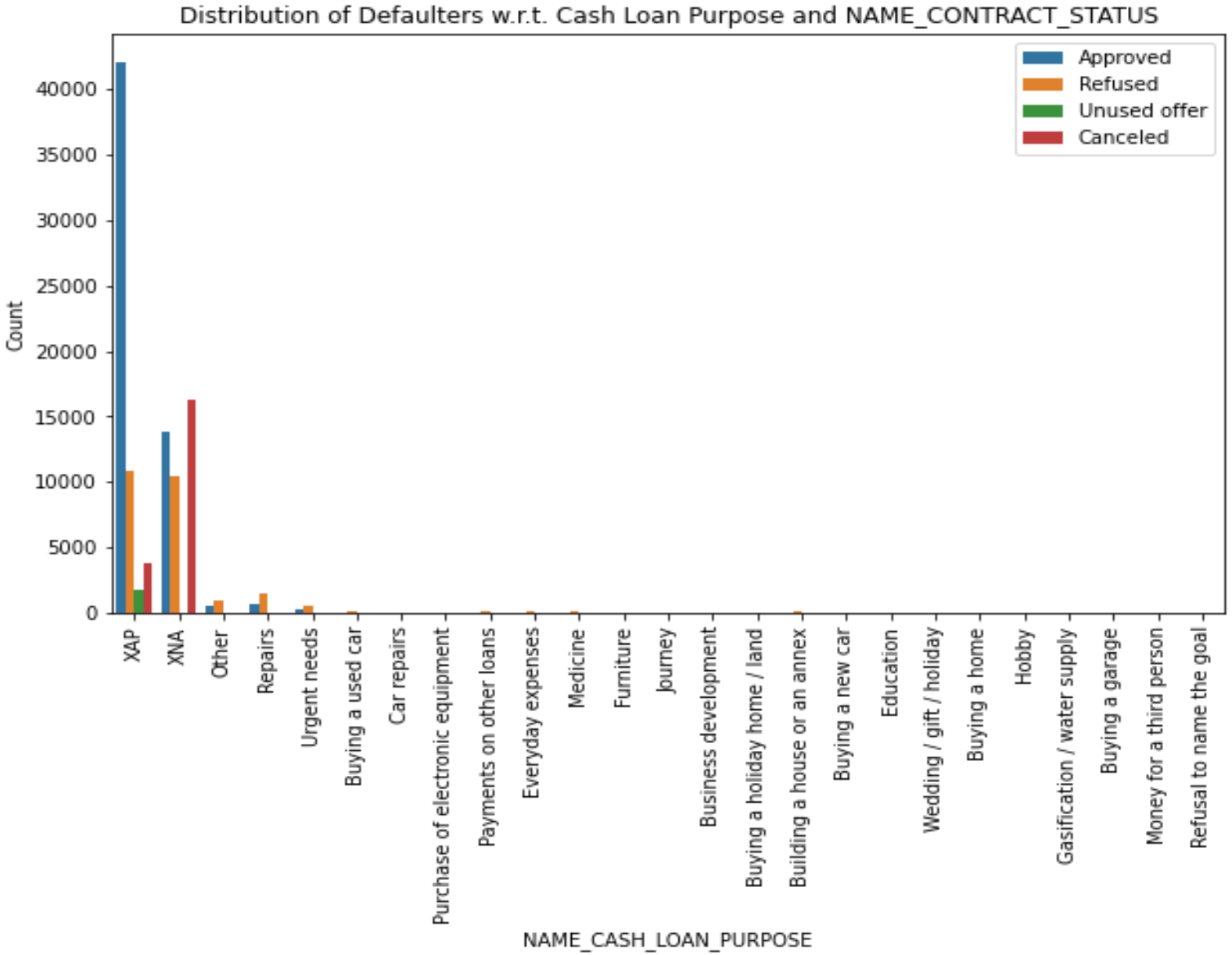
Relevant Results(**Bivariate/Multivariate Analysis**)

- We have more defaulters in self employed and business entity type 3.



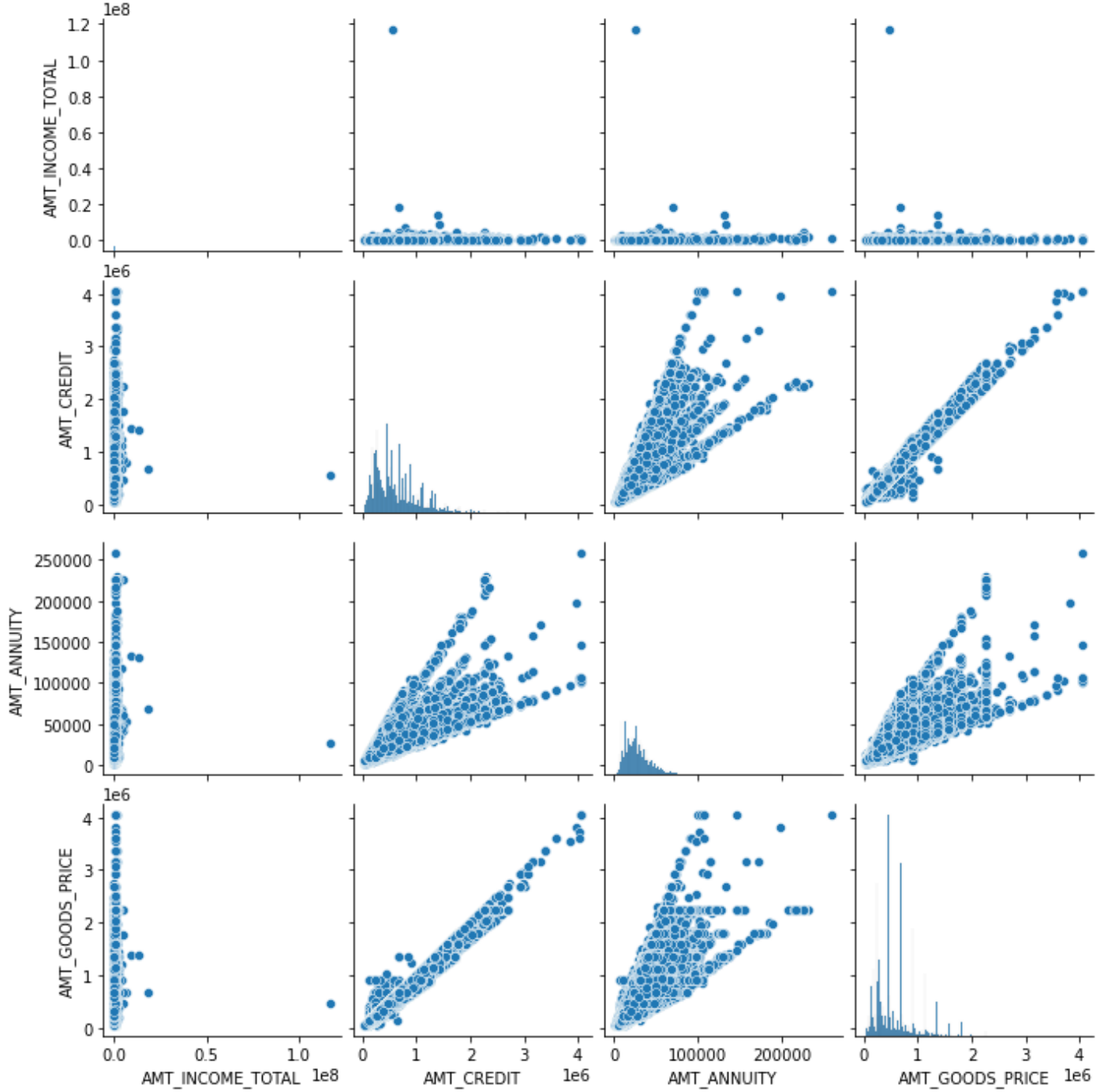
Relevant Results(**Bivariate/Multivariate Analysis**)

- Distribution of defaulters w.r.t Cash Loan purpose and Contract Status.



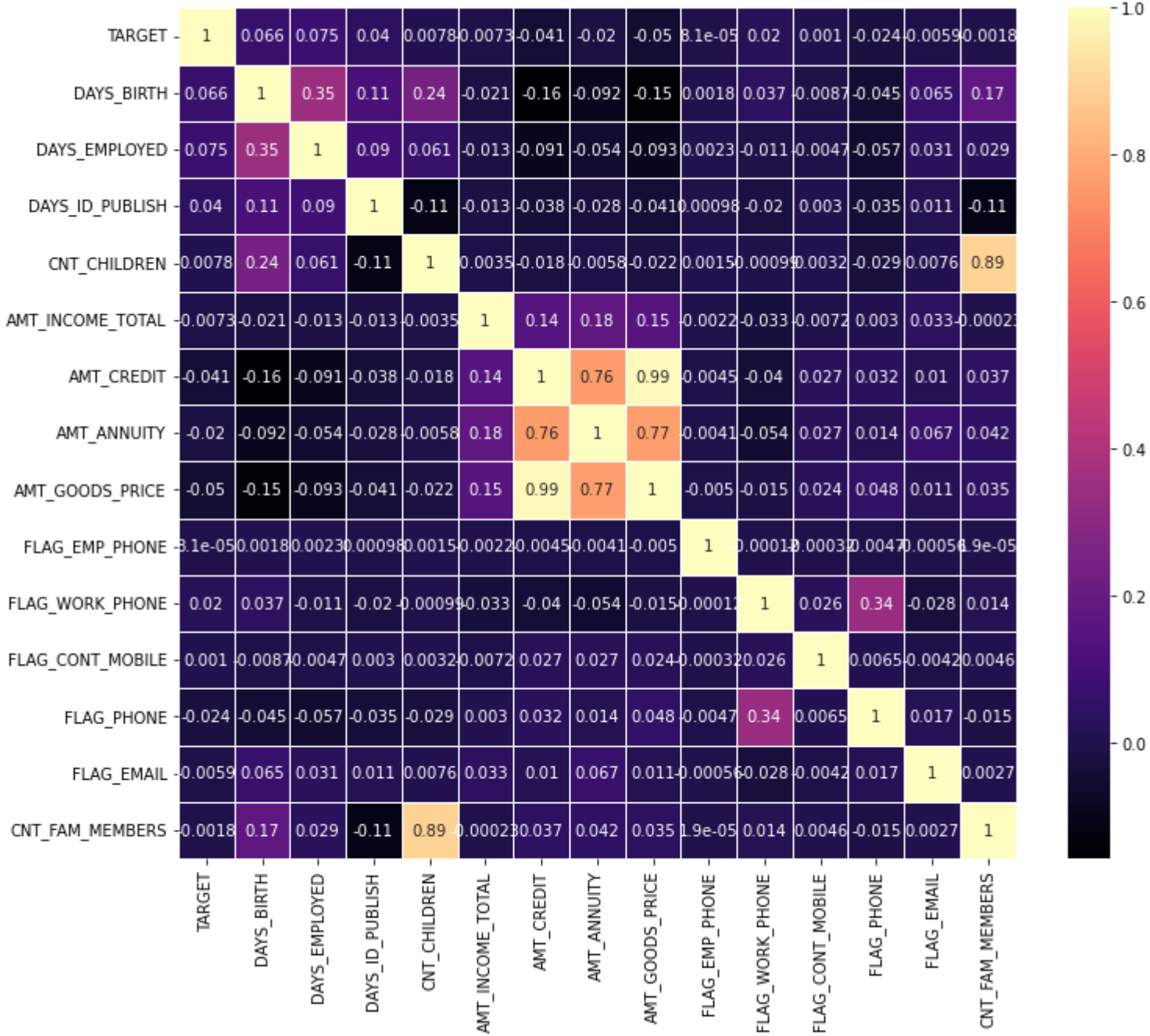
Relevant Results (Bivariate/Multivariate Analysis)

- AMT_CREDIT and AMT_GOODS_PRICE have highly positive correlation.



Relevant Results (Bivariate/Multivariate Analysis)

- AMT_CREDIT and AMT_GOODS_PRICE have highly positive correlation.
- CNT_FAM_MEMBERS and CNT_CHILDREN have highly positive correlation



Relevant Results (Bivariate/Multivariate Analysis – prev_data)

- AMT_CREDIT and AMT_APPLICATION have highly positive correlation.