Exploratory Data Analysis

Credit EDA Assignment

Submitted by:
MAYUR PADORE

Problem Statement :

- The Loan providing company find it hard to give loans to the people due to their insufficient balance, not paying their EMI on time, non-existent credit history, there may be the people who are new to taking loan from Bank / Company, and people with all different type of Occupation Class.

- We have to perform the EDA to analyze the pattern present in the Data.

- So, we have to ensure that, if the Company receives a loan application, that it check the pattern to whether a application meets the requirement of payment their bills on time.

- There are two scenarios which occur loss for both of them:

➢ If the client can repay the bills, and company rejected it , then there is a loss of business to the company

➢ If the client is not likely to repay the loan, then approving the loan may lead to a financial loss for the company.

Approach / Steps to perform EDA :

- Understand the Data file , which kind of data it is.

- Import the Warnings, Libraries and Plots Libraries for presenting the Graphs.

- Upload the .csv file with the help of pandas library.

- Analyse the Data File, Check the Missing value if possible remove those missing value or Impute it with some value e.g. Mean/Median/Mode. Find the Outliers and Standardize the values.

- After performing cleaning, Analyse the Variables, which looks important for handling the loan application:
- For Analysis, we have three methods :

    1) Univariate Analysis
    2) Bivariate Analysis
    3) Multivariate

- By Analysing the data and plotting the Graphs for Variables depend on their datatypes, we will find the correlation between them.

Dataset :

1. *'application_data.csv'* contains all the information of the client at the time of application. The data is about whether a **applicant has payment difficulties.**

2. *'previous_application.csv'* contains information about the applicant's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

3. *'columns_description.csv'* is data dictionary which elaborates the meaning of the variables.

“ TARGET ” :

```
1 sns.countplot(x = "TARGET", data = AppDf)
2 plt.show()
```

For this Dataset, I create two category ,

“ 0 “ – who can repay their loan time to time and had no difficulties till now. (Non-Defaulters)

“ 1 “ – who has difficulties , or not like to repay. (Defaulters)

From this Countplot, we can see that maximum number of people are ,who can repay their loans

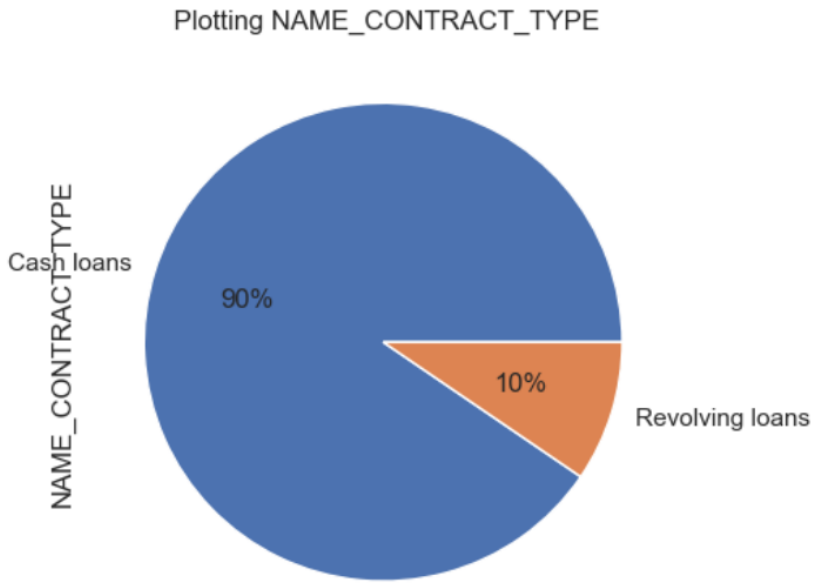
- Here, I decided to choose those only those columns which looks more important.

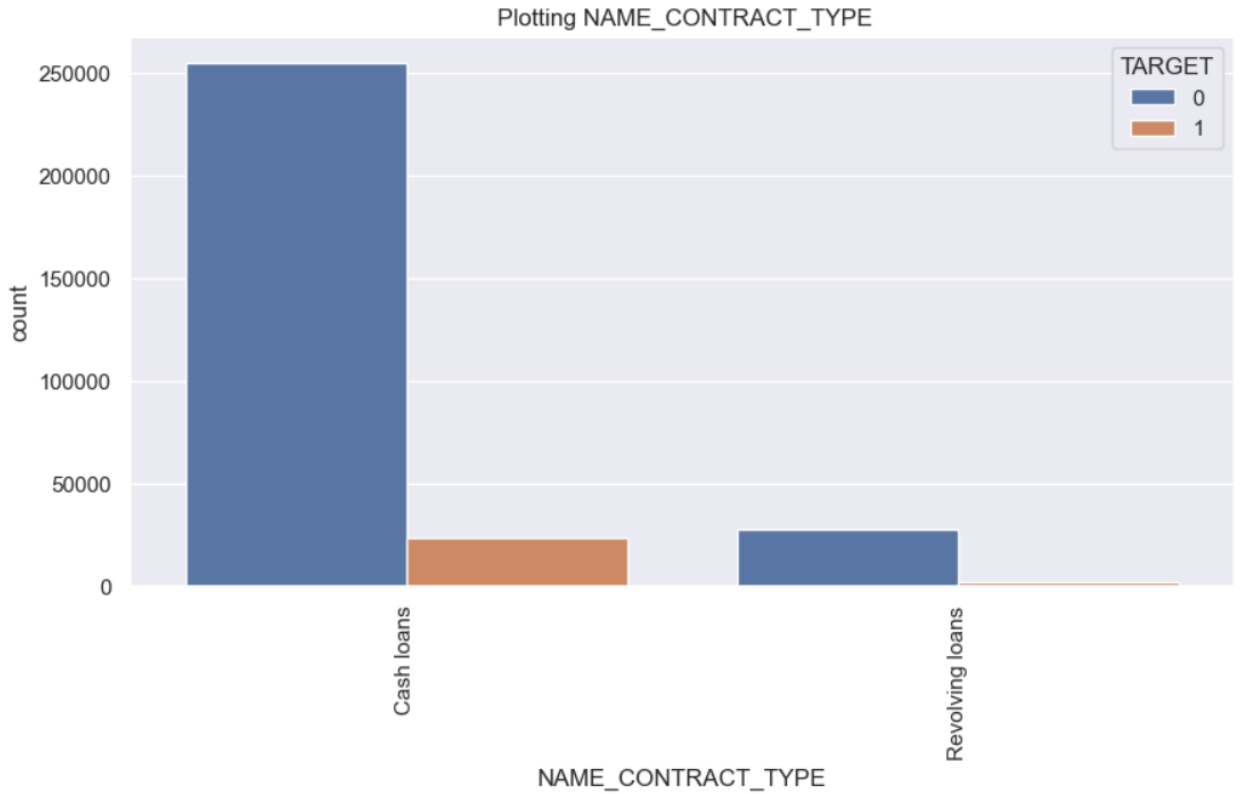
```
In [125]: 1 categorical_column.columns

Out[125]: Index(['NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY',
                 'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE',
                 'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE', 'OCCUPATION_TYPE',
                 'WEEKDAY_APPR_PROCESS_START', 'ORGANIZATION_TYPE',
                 'DAYS_LAST_PHONE_CHANGE'],
                dtype='object')

In [126]: 1 numerical_column.columns

Out[126]: Index(['AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE',
                 'REGION_POPULATION_RELATIVE', 'CLIENTS_AGE', 'DAYS_REGISTRATION',
                 'CNT_FAM_MEMBERS', 'EXT_SOURCE_2', 'EXT_SOURCE_3',
                 'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE',
                 'OBS_60_CNT_SOCIAL_CIRCLE', 'DEF_60_CNT_SOCIAL_CIRCLE'],
                dtype='object')
```

And, based on their datatypes, Lets Perform Analysis:

- **Univariate Analysis**
  - **On Categorical Data**
  - **On Numerical Data**

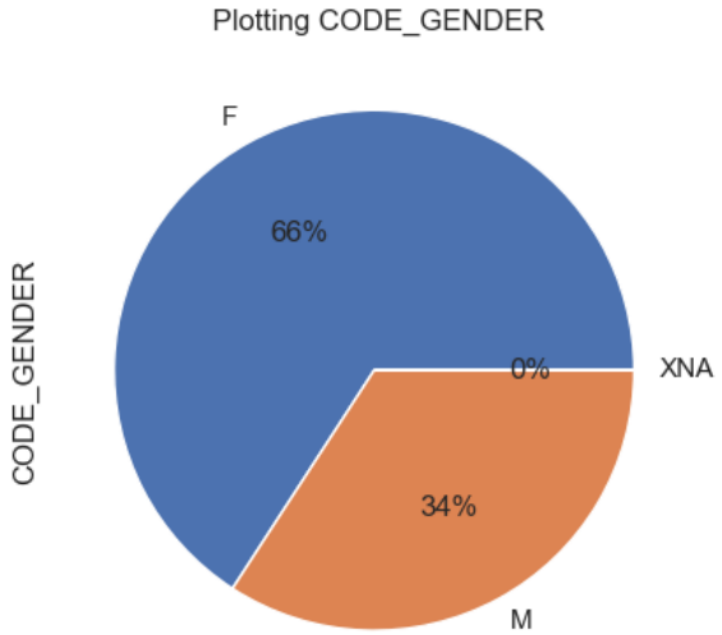
Identification if loan is cash or revolving :



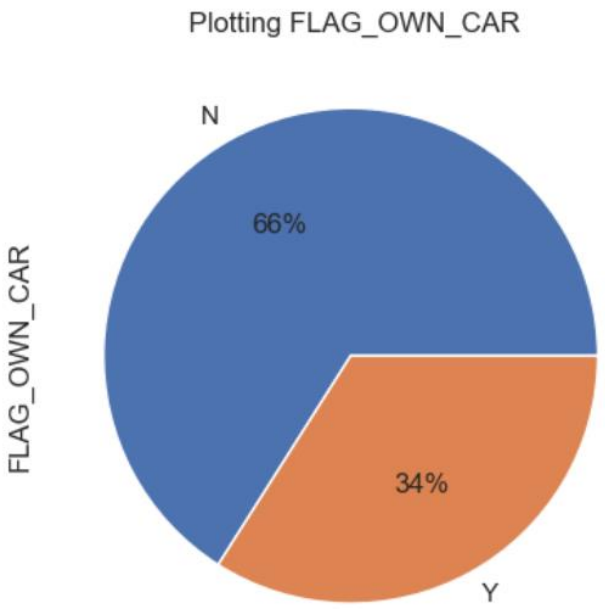


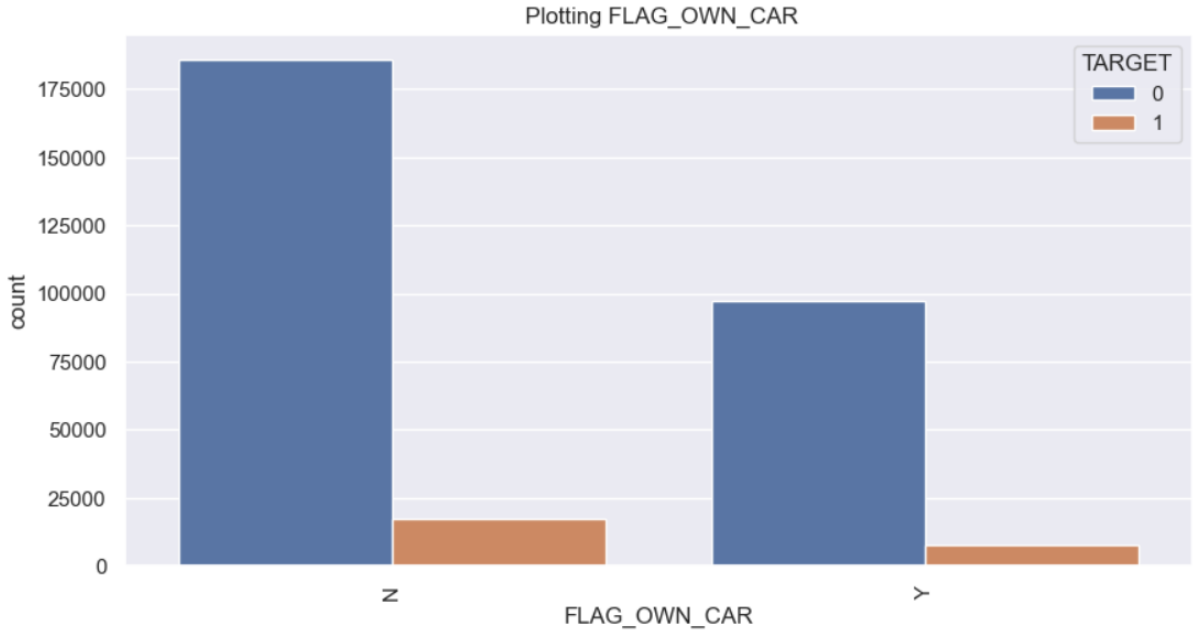
Applicant Gender :

- We can see that, Female Applicants are applying for loan are more than Male Applicants, as well as their are maximum numbers of Non-defaulter.



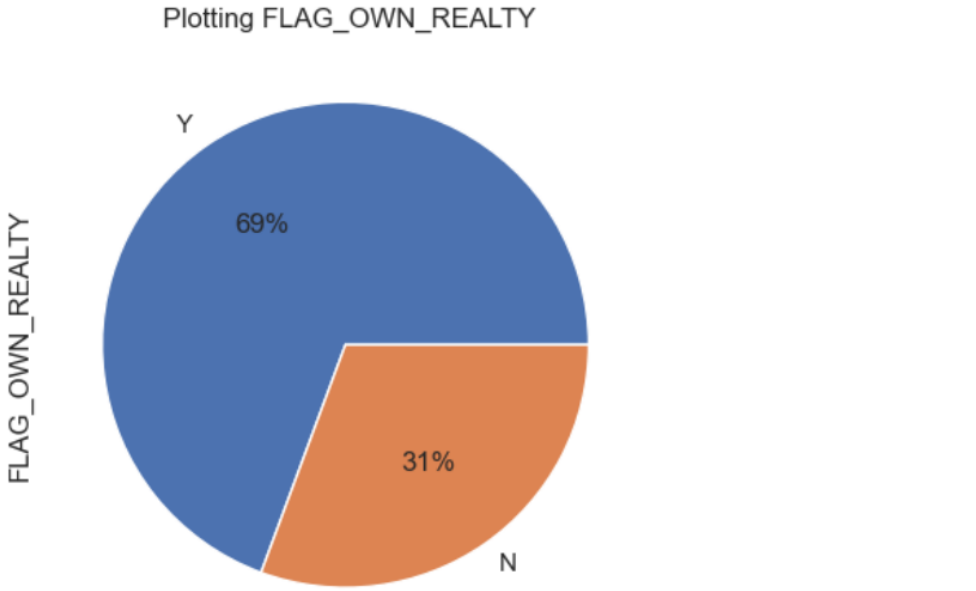
Applicant Own A Car :

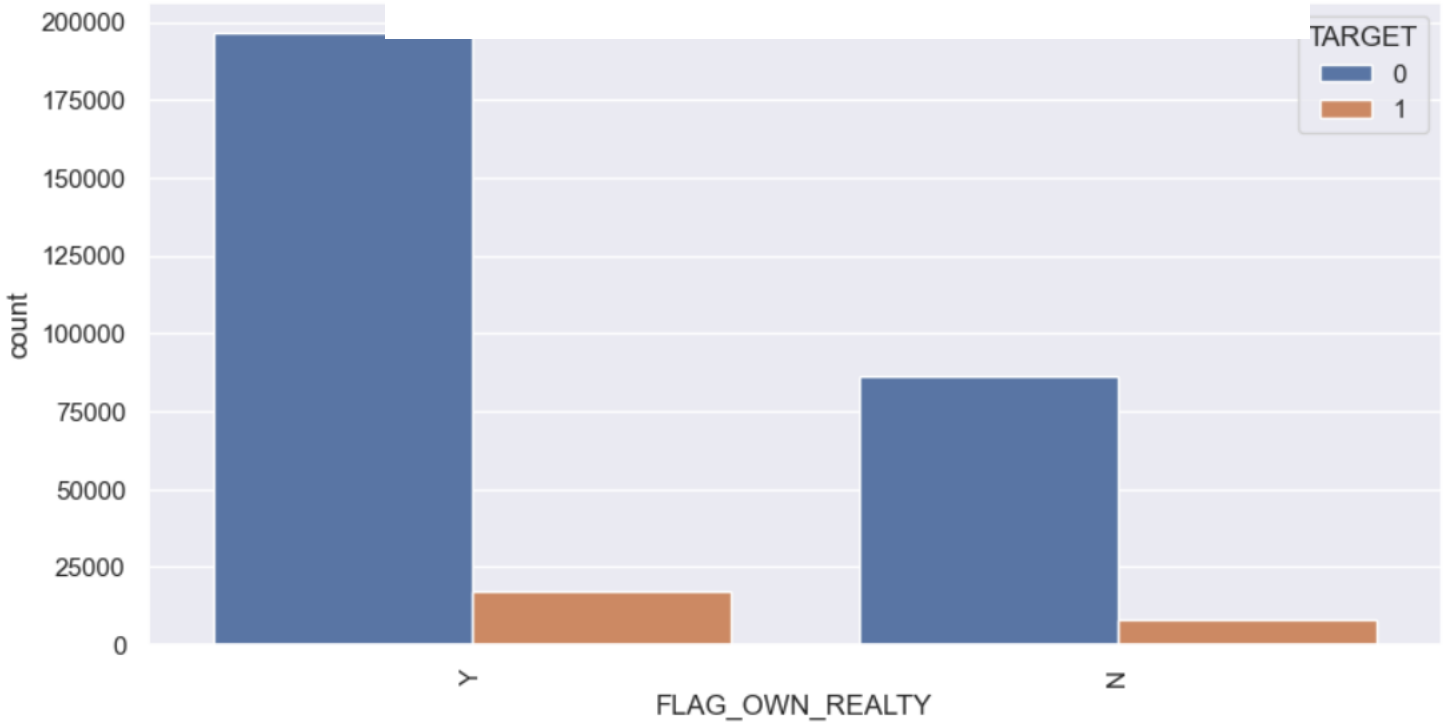




Applicant Own a House :

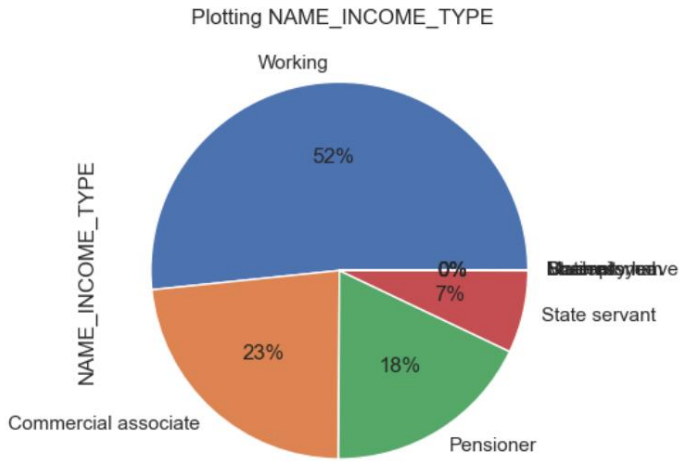
- From two Data , Applicant who don’t own a Car are high in numbers but also we can see that own a House, from this we can say they are middle-class / stable income type.

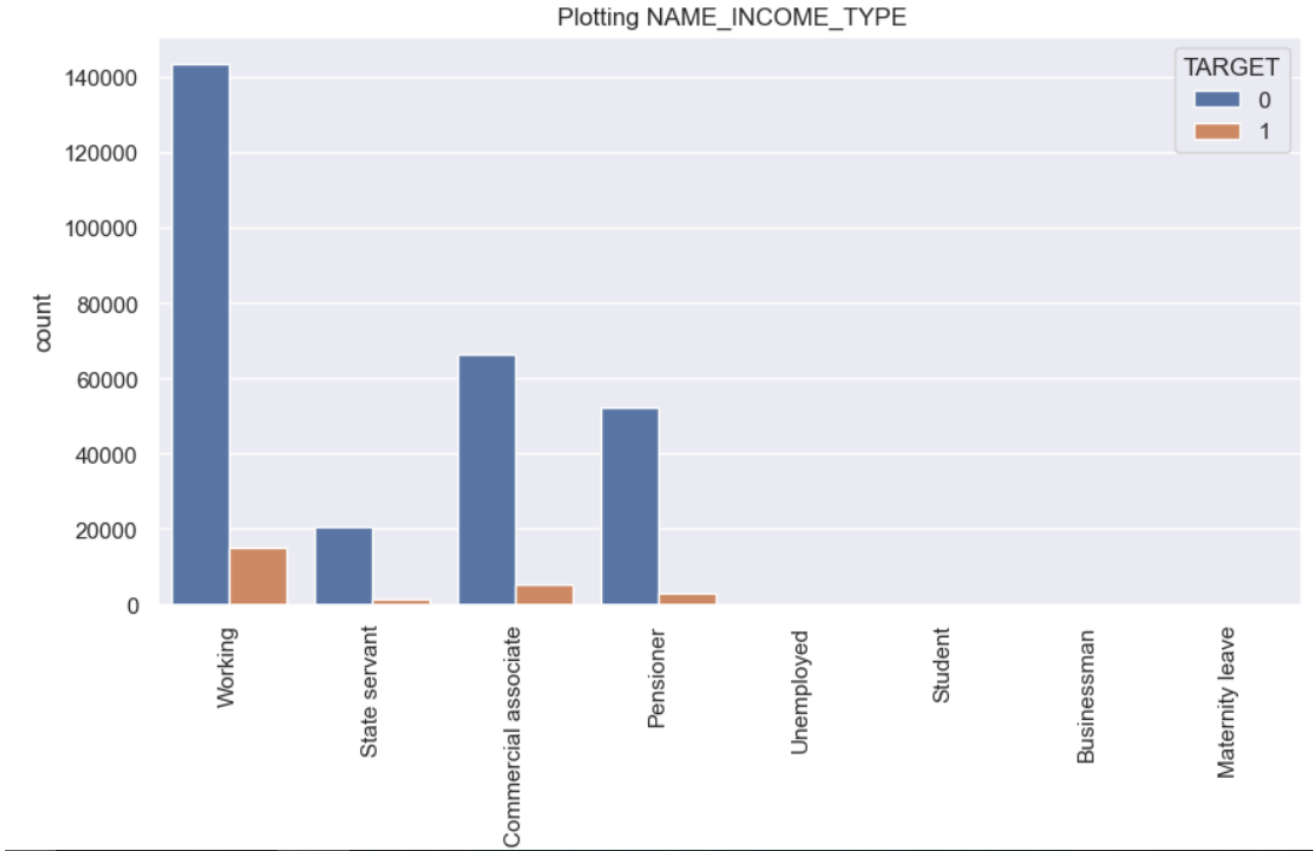




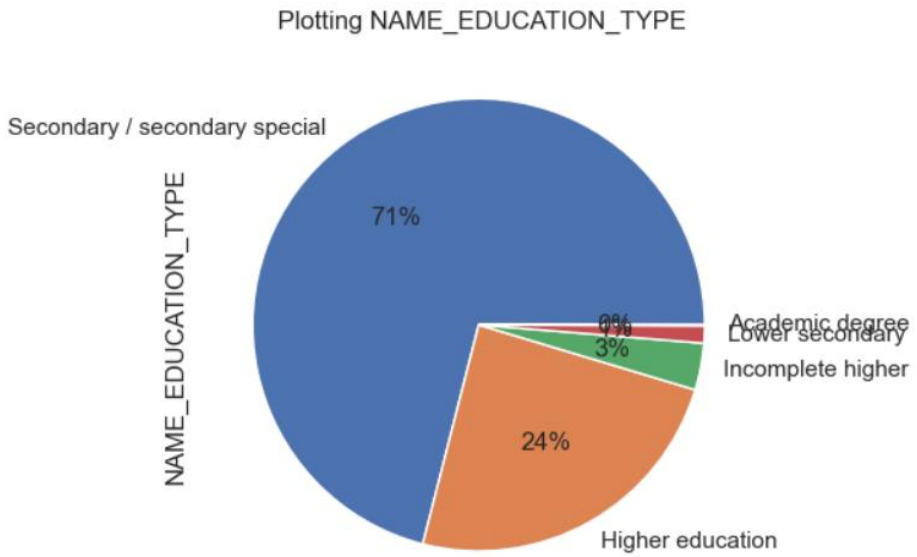
Applicant Income Type :

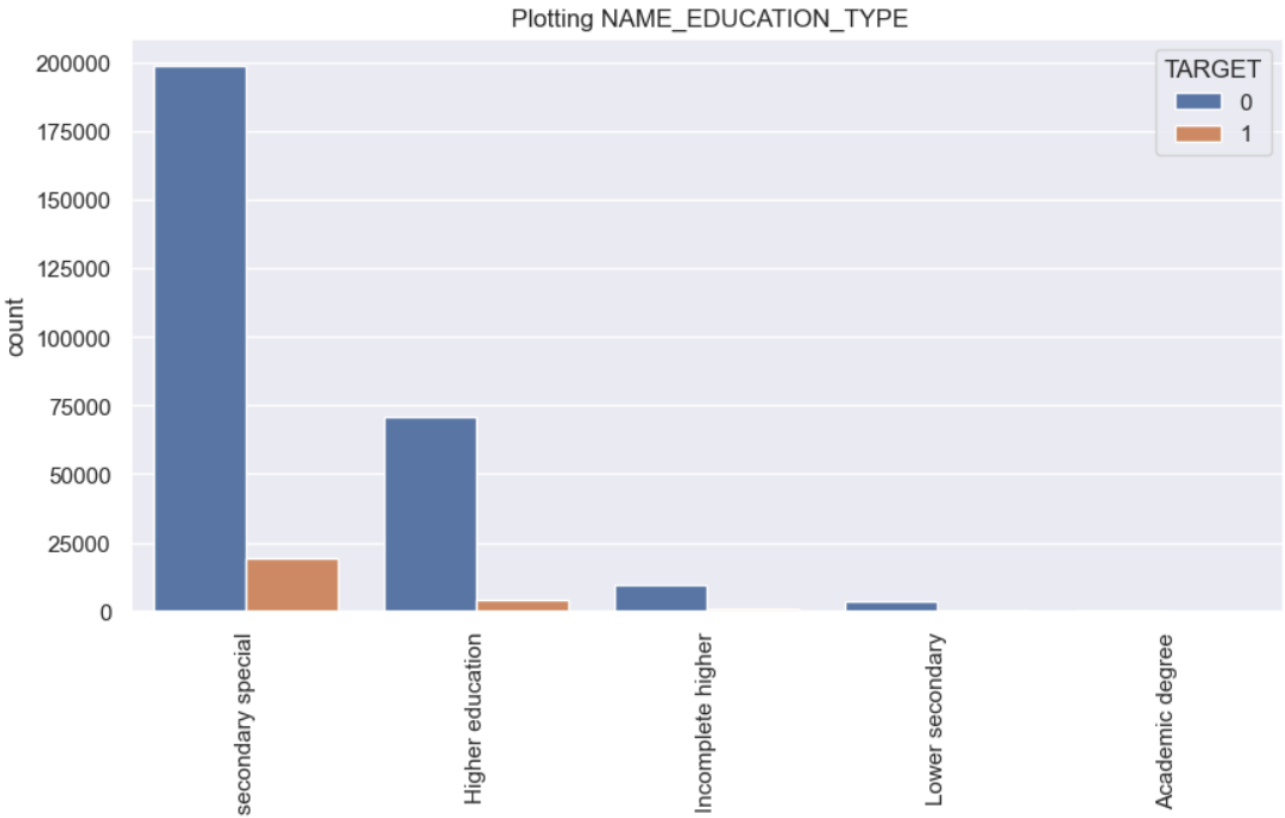
- Half of the Majority Applicants are working Class, they has no non-repaying issue.





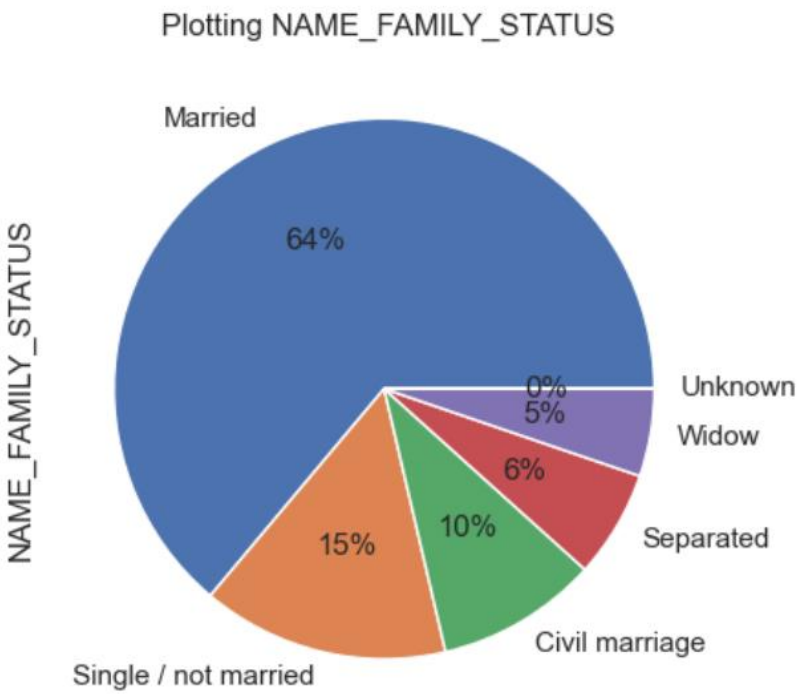
Applicant Education Type :

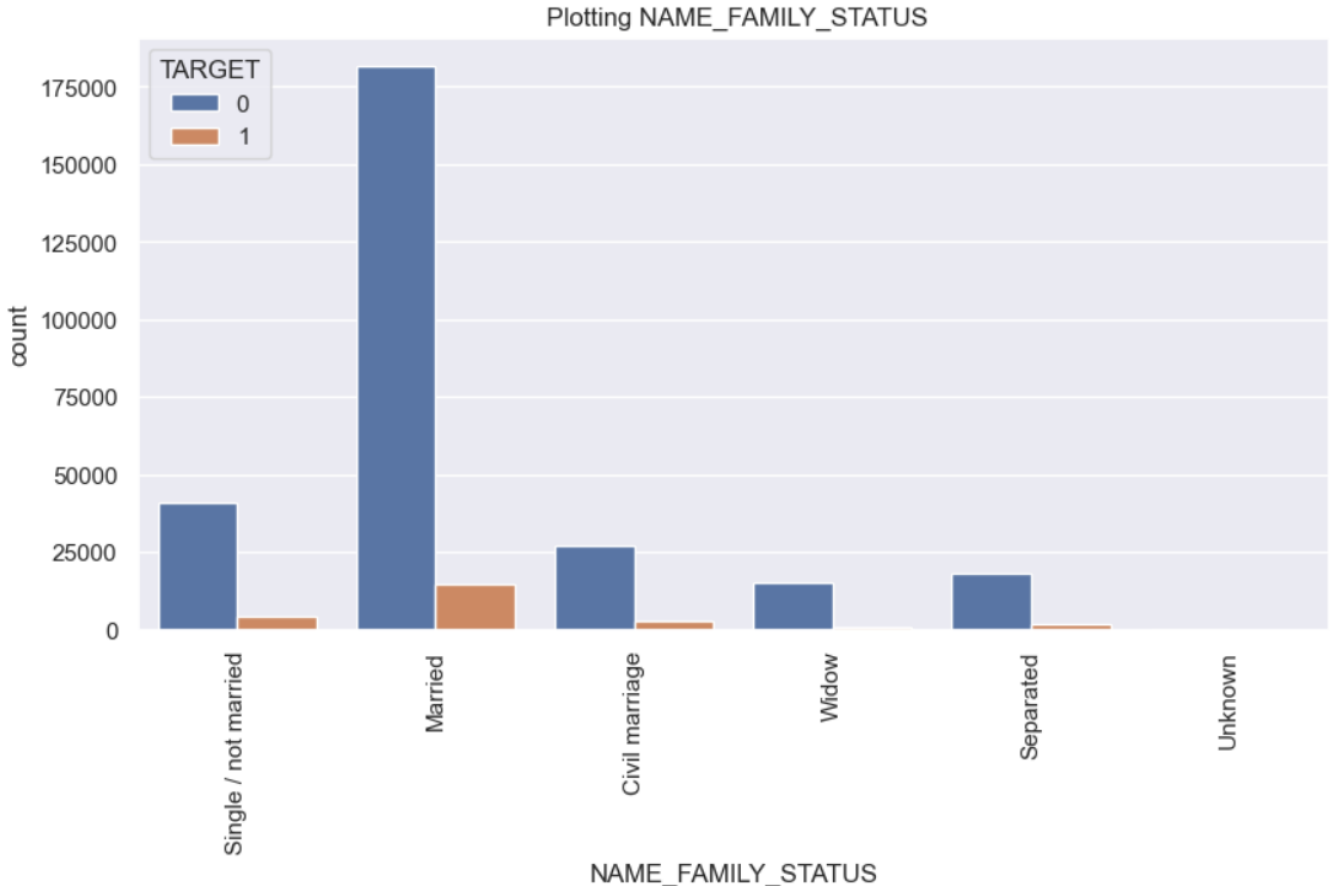




- Most of defaulters and non-defaulters have education level up to Secondary / Secondary special.

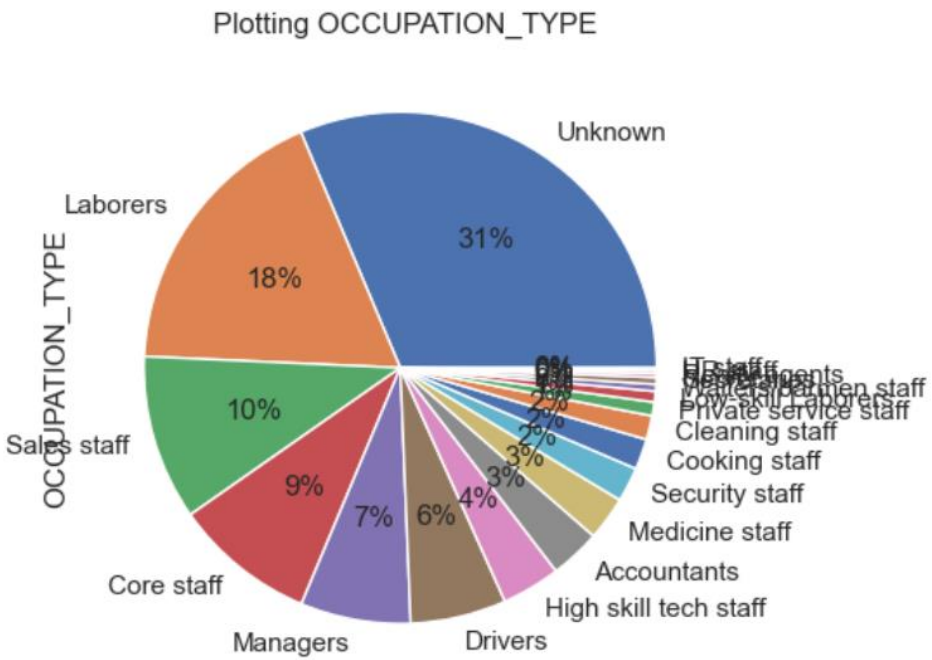
Applicant Family Status :

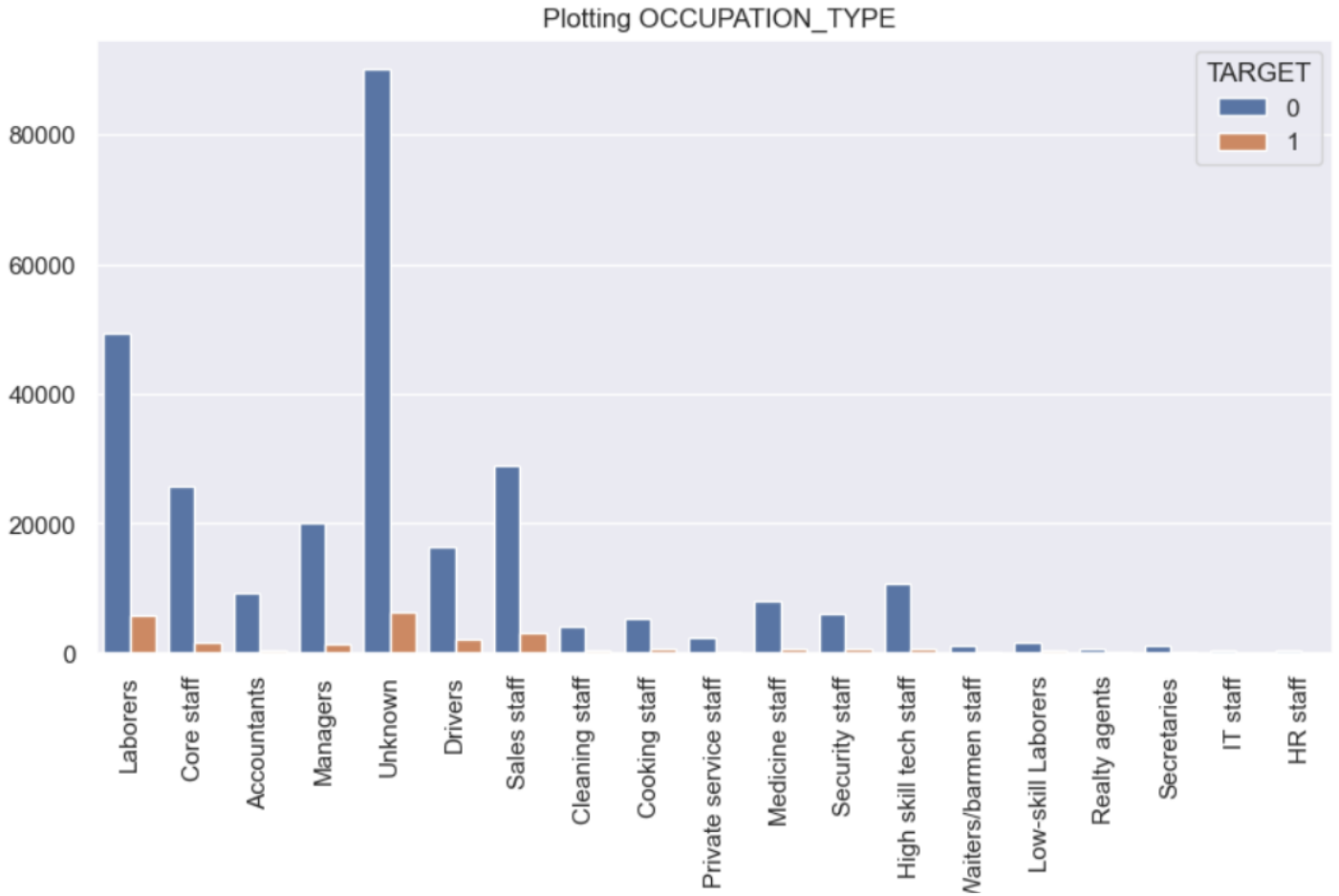




- Most of the applicants are married in both the segments.

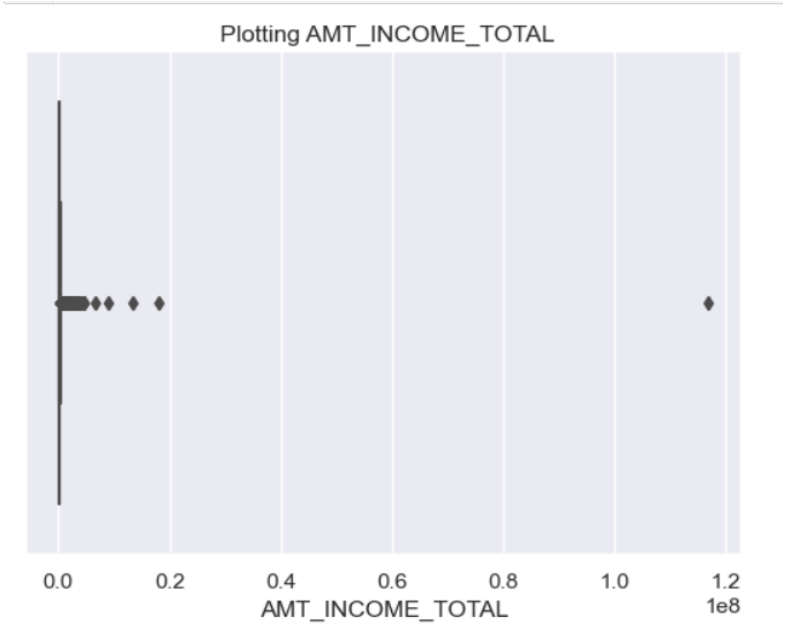
Applicant Occupation Type :

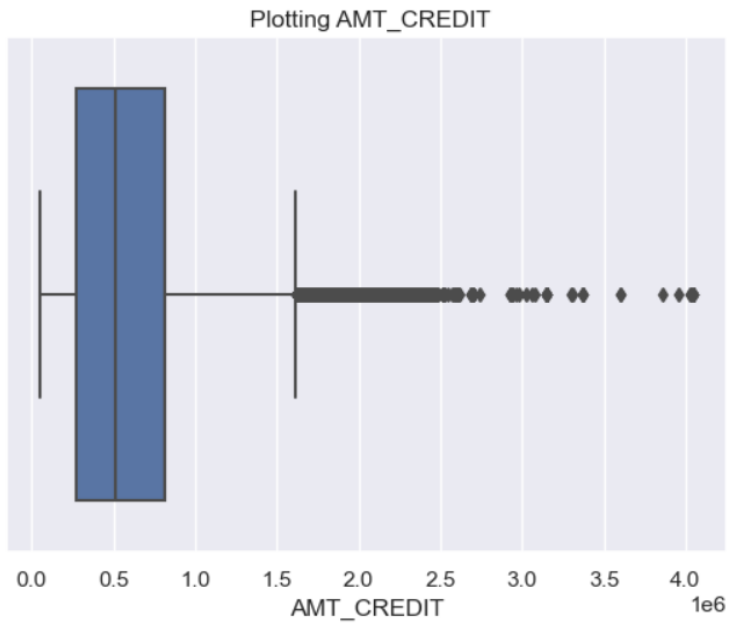


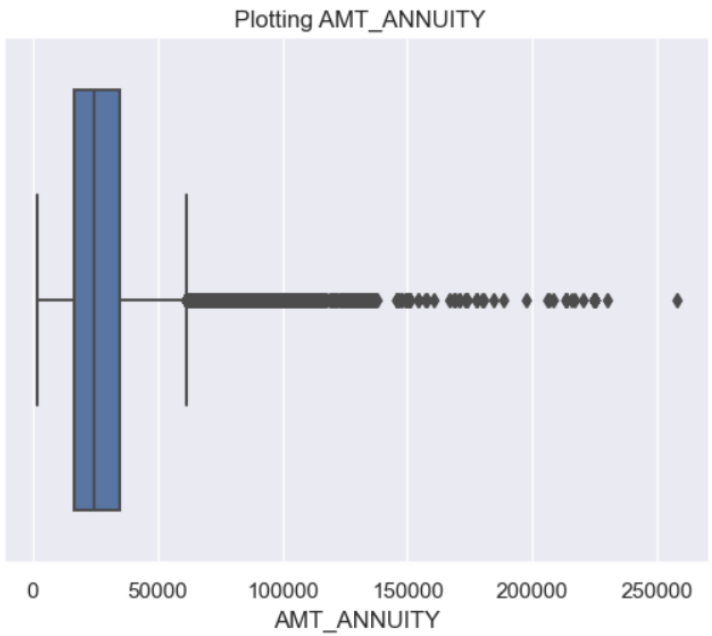


- As we can see 18% are from Labor Class and 31% of the data Occupation was missing hence imputed with Unknown to take care of the missing values

BARPLOT of Income of Applicant, Credit Amount of Loan, and Loan Annuity :







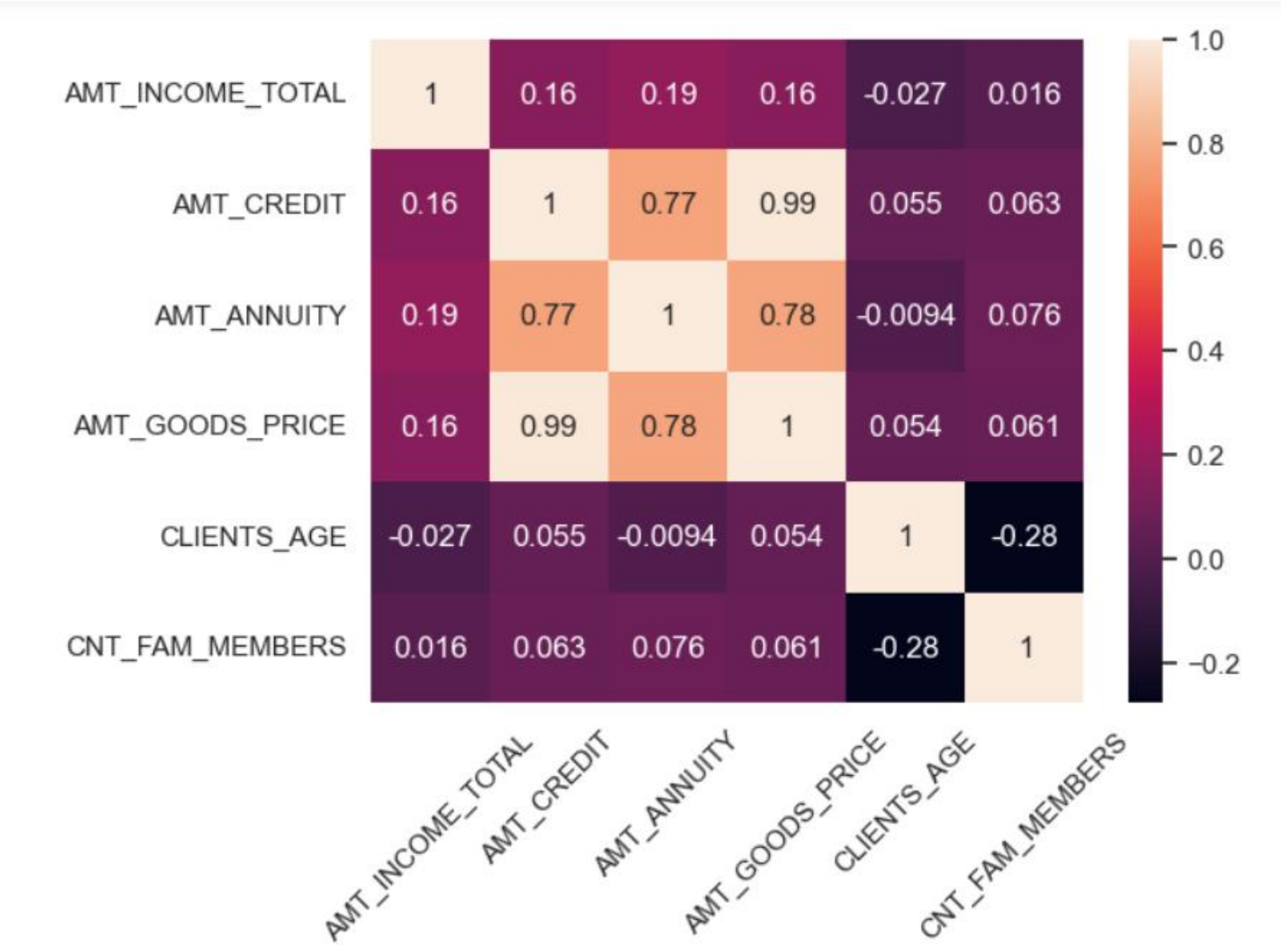
- As we can see that, in Income Data 3 People whose annual are more than 1 Cr are also available .

```
In [184]:  1  AppDf[AppDf['AMT_INCOME_TOTAL']>10000000]
```

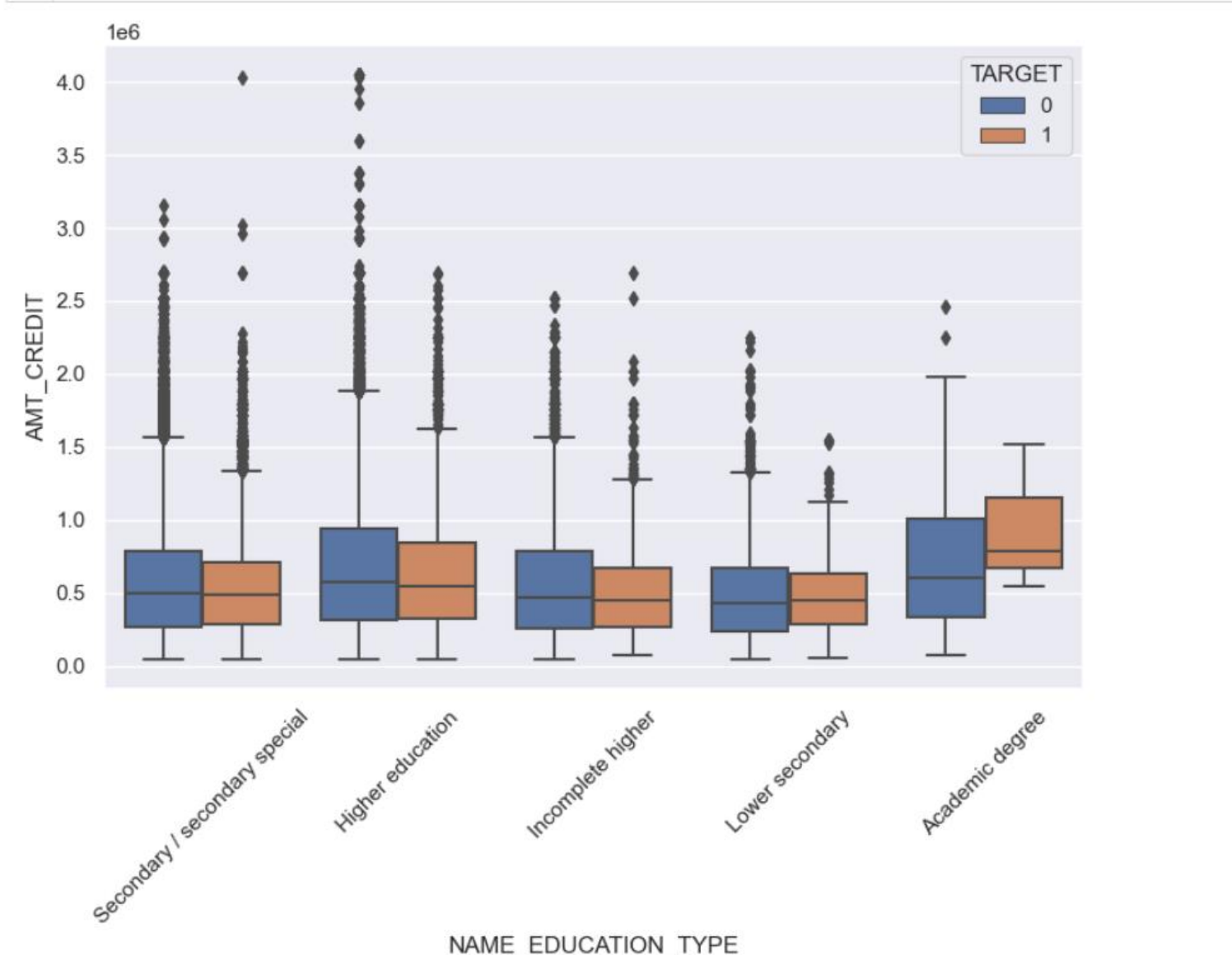
Out[184]:

\G_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY
Y	1	117000000.0	562491.0	26194.5
Y	2	18000090.0	675000.0	69295.5
Y	0	13500000.0	1400503.5	130945.5

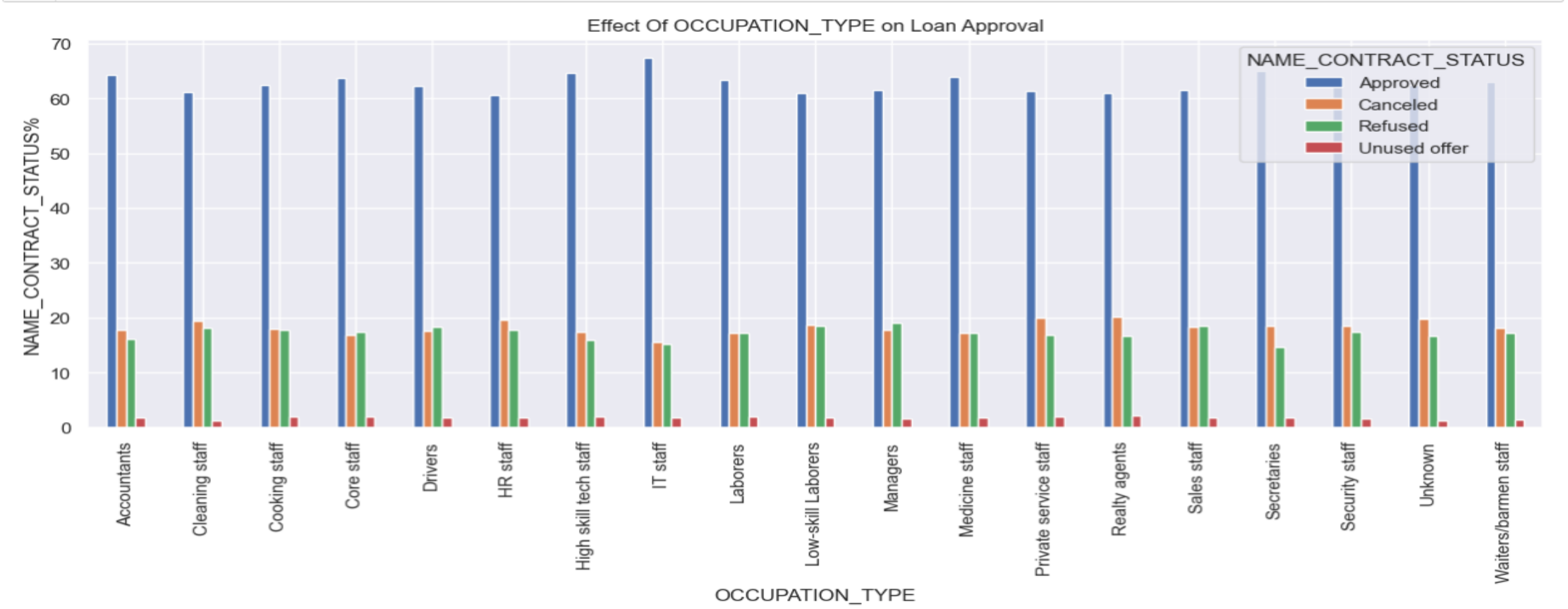
Correlation Plot :



Bivariate Analysis on Categorical against Numerical :



Merging Data :



CONCLUSION AND RECOMMENDATIONS :

- Most of the Applicant are Female, and the Non-Defaulter Applicants greater in numbers as compared to Defaulter.
- Applicants whose Loan Application are “Approved” are Secondary or Higher Class and people who are working professionals.
- There are also a Defaulter case, so not-approving their loan can be another good decision for Company.
- To Reduce the Loan Application, we can drop those Defaulter case, as from our Analysis Clients with repaying issue, are low in income, not married, not a stable working profile and their age lies between 25-35.