Credit EDA Assignment

Business Problem

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.

Business Objective : Identify the patterns

- Client with payment difficulties.
- All other cases. When payment is paid on time. Identify the clients capable of repaying the loan.
- The company wants to understand the driving factors/variables behind loan difficulties i.e. variables which are strong indicators of default.
- Clients capable of repaying the loan but applications are rejected.

Session-1 Data Cleaning

Segment-1: Check for missing values

- OCCUPATION_TYPE has over 31% of missing values. We can impute the values either with mode of occupation type or mark those missing values as separate category ‘Not mentioned’.
- CNT_FAM_MEMBERS has 2 missing values. Drop those rows.

Segment-2: Verify if there are any irregularities in the columns

- CODE_GENDER has 4 XNA values. Replace with mode value.

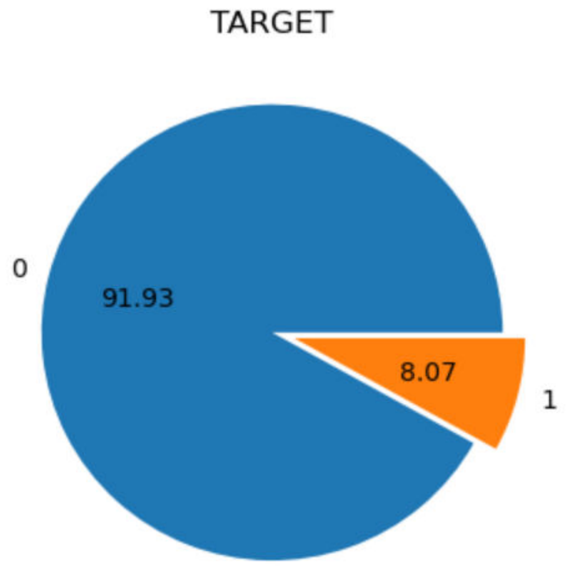
Segement-3: Standardizing values

- DAYS_BIRTH - Client's age in days at the time of application. Convert it to years and drop this column.
- Convert the values Y and N to 1 and 0 for FLAG_OWN_CAR and FLAG_OWN_REALTY.

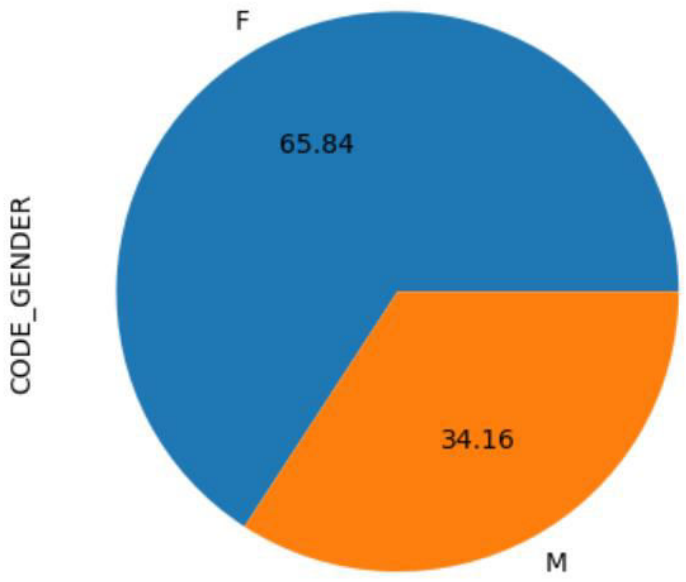
Session- 2 Univariate Analysis

Segment-1 Categorical unordered univariate analysis

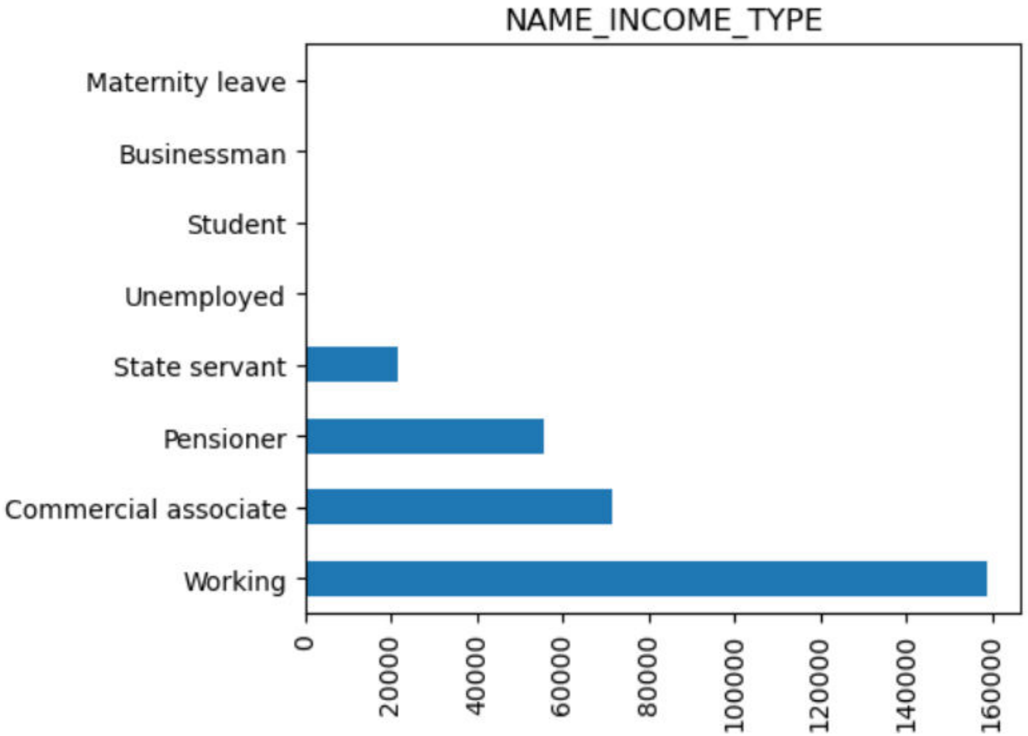
- TARGET : 8.07% of clients are having payment difficulties and majority are 91.93% of clients paid the payment on time.



- CODE_GENDER : Clients who have availed loan, 34.16% are male and 65.84% are female.

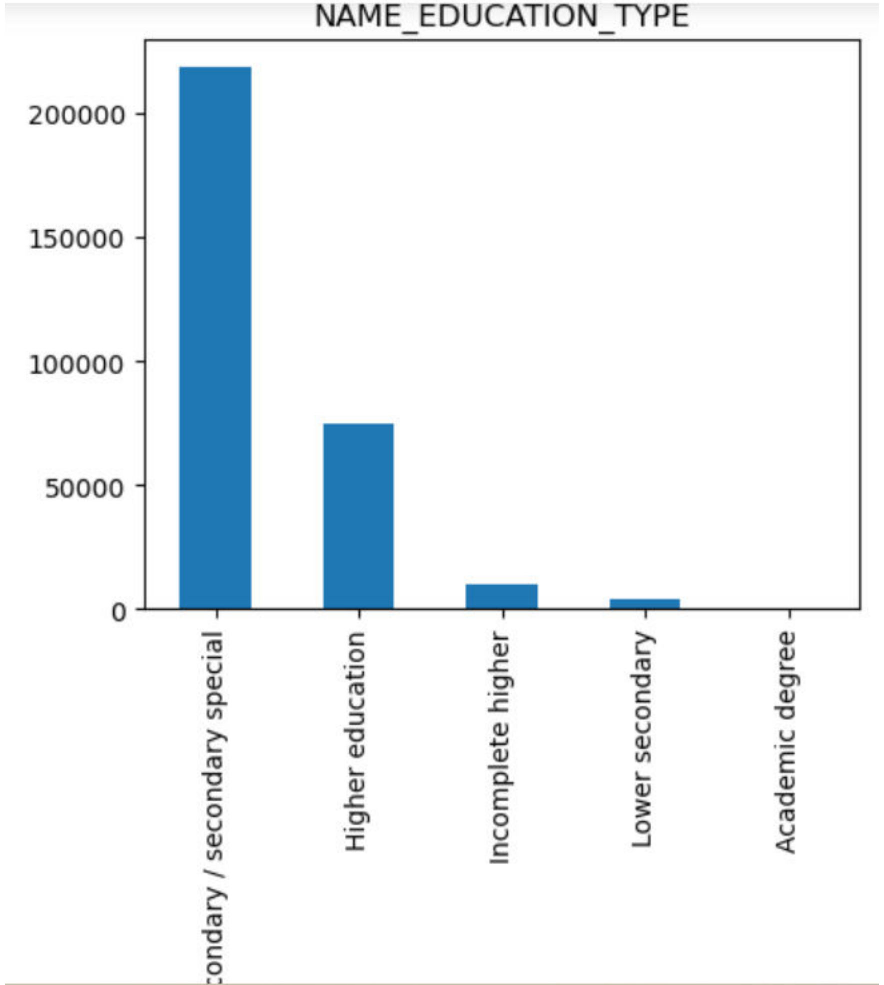


NAME_INCOME_TYPE: Client income type of working category are more among the clients who has taken the loan.

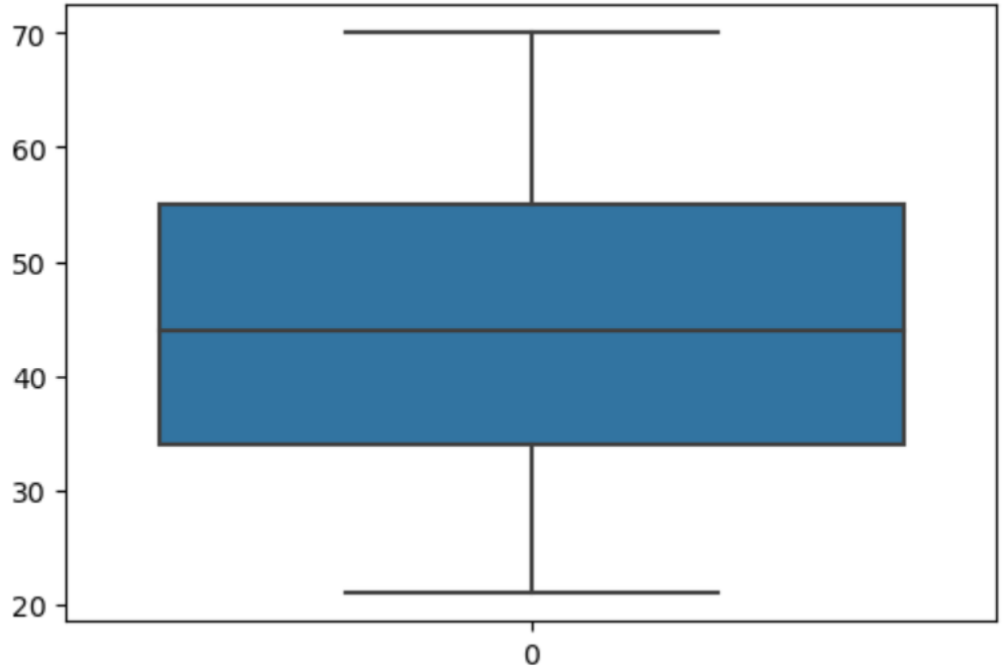


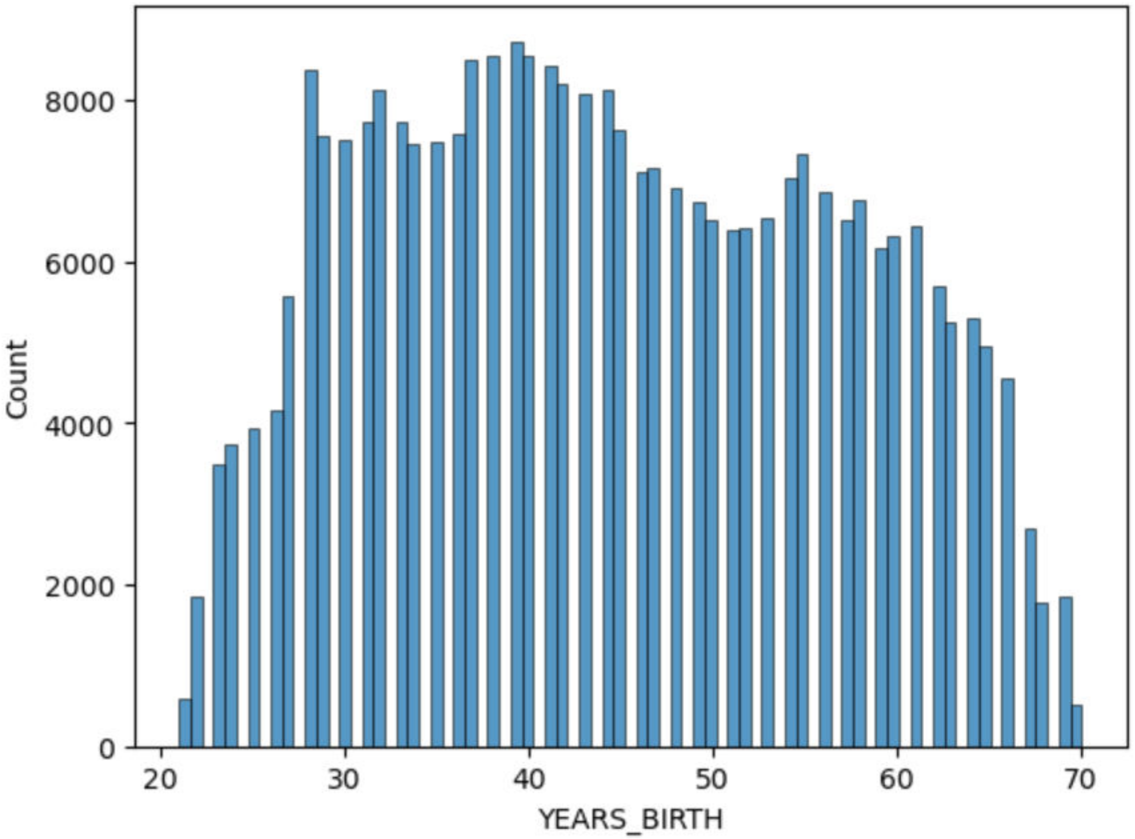
Segment-2 Categorical ordered univariate analysis.

NAME_EDUCATION_TYPE : The maximum of clients who has availed loan are from Secondary/secondary special.



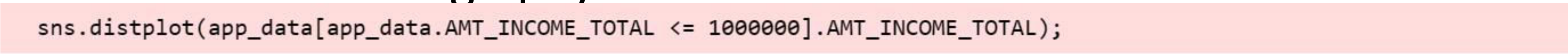
- YEARS_BIRTH : Min and Max age of clients are 21 and 70 resp and clients who were in 40’s are the majority to avail loan.



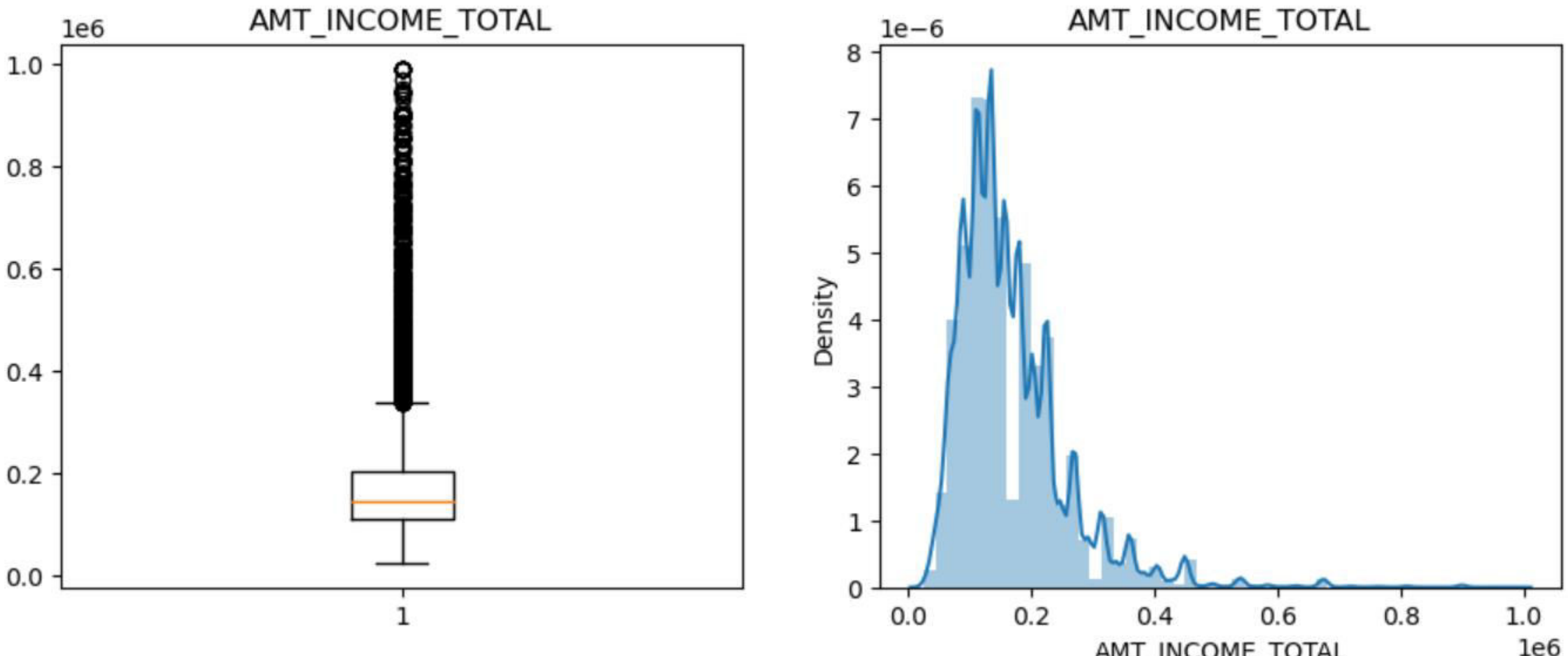


Segment-3 Continuous variable

AMT_INCOME_TOTAL: There are outliers but we will not consider them as outliers as there could be clients of high pay who also avail for loan.

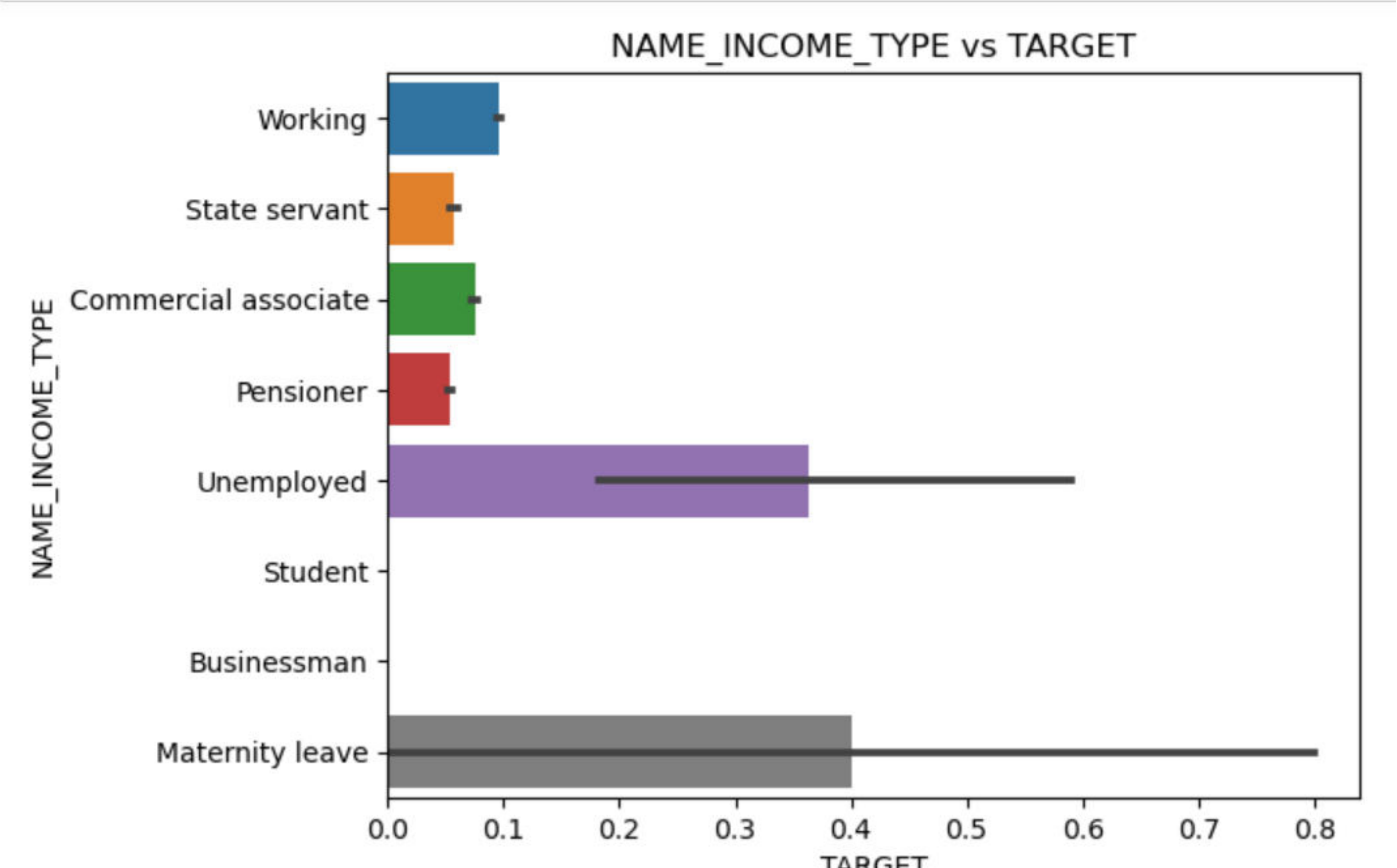


```
sns.distplot(app_data[app_data.AMT_INCOME_TOTAL <= 1000000].AMT_INCOME_TOTAL);
```

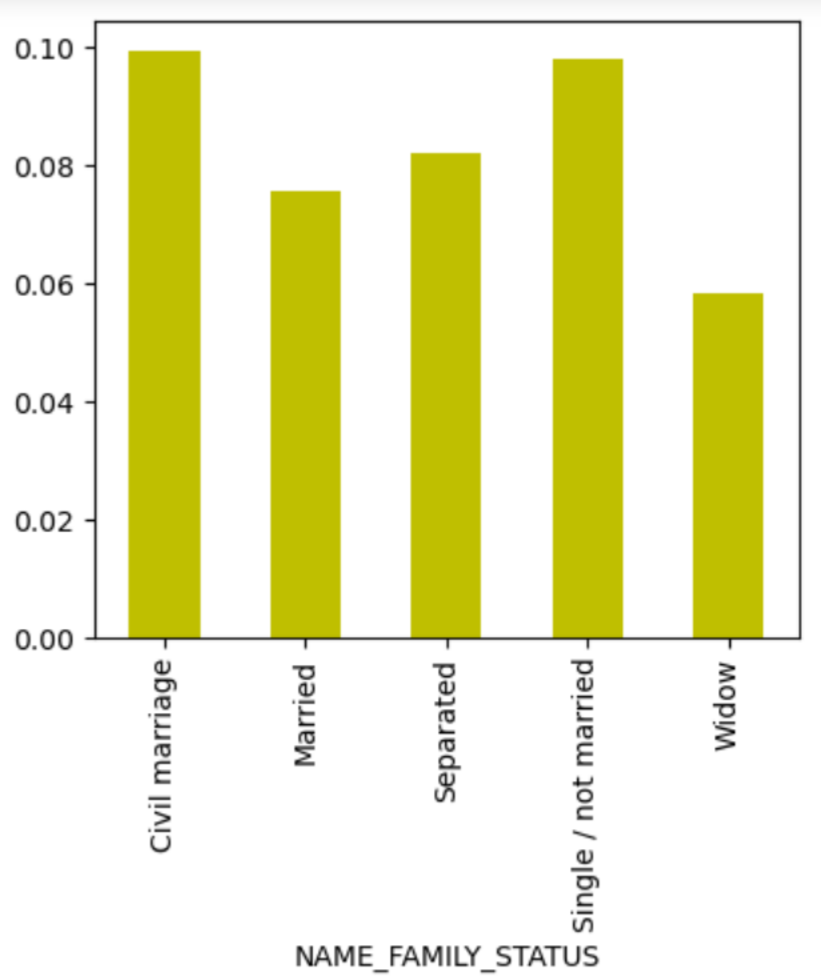


Session-3 Bivariate Analysis

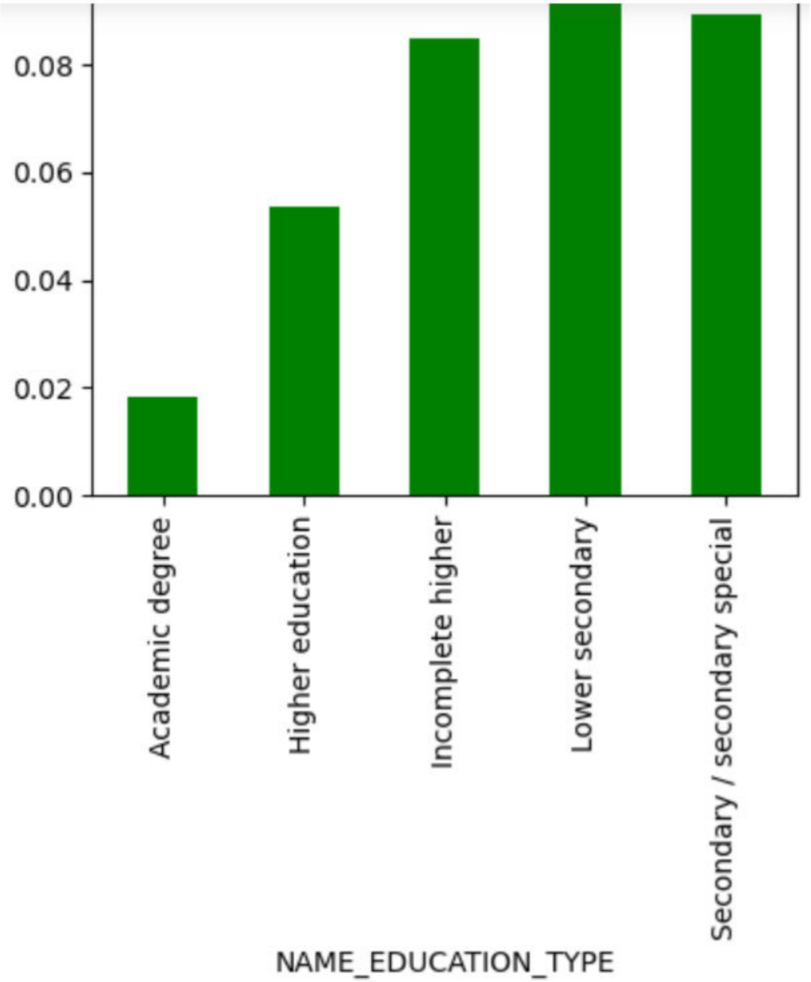
- NAME_INCOME_TYPE vs TARGET : Clients who are unemployed or on maternity leave tends to have payment difficulties.



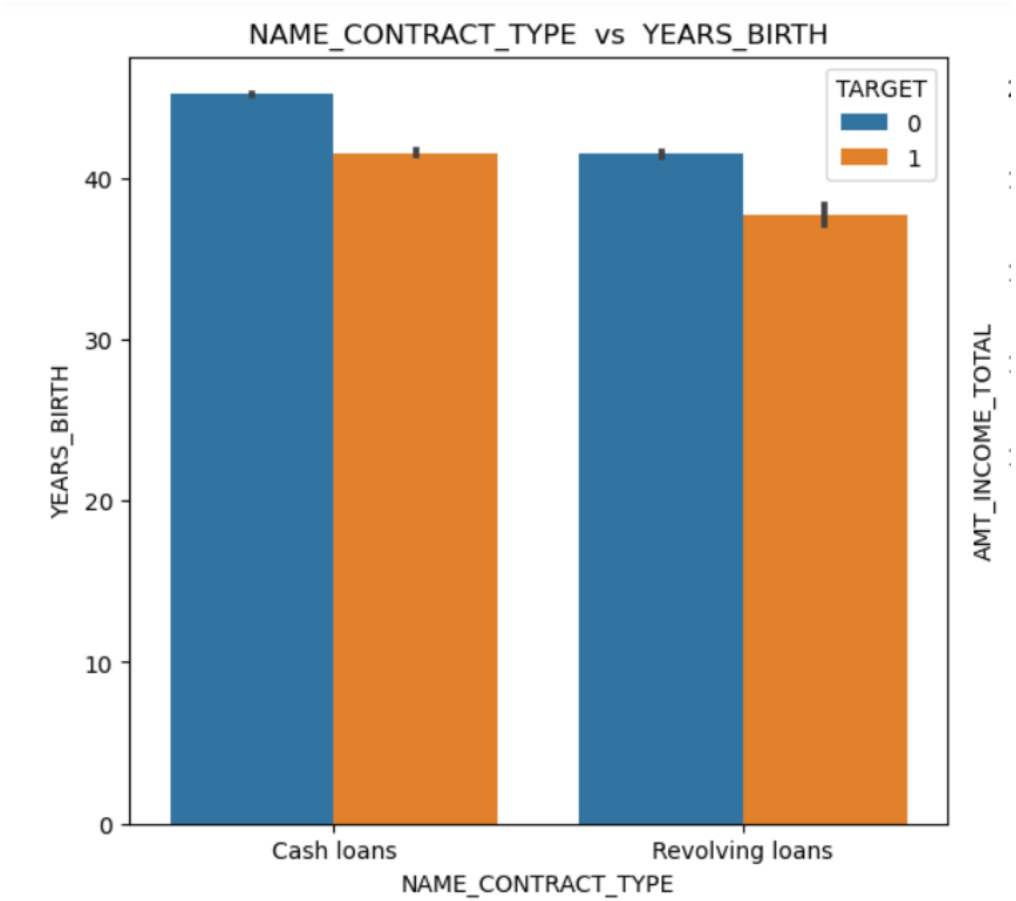
- NAME_FAMILY_STATUS vs TARGET : Clients whose marital status is either Single/not married or whose marriage is civil marriage, tends to have difficulties in repaying the loan.



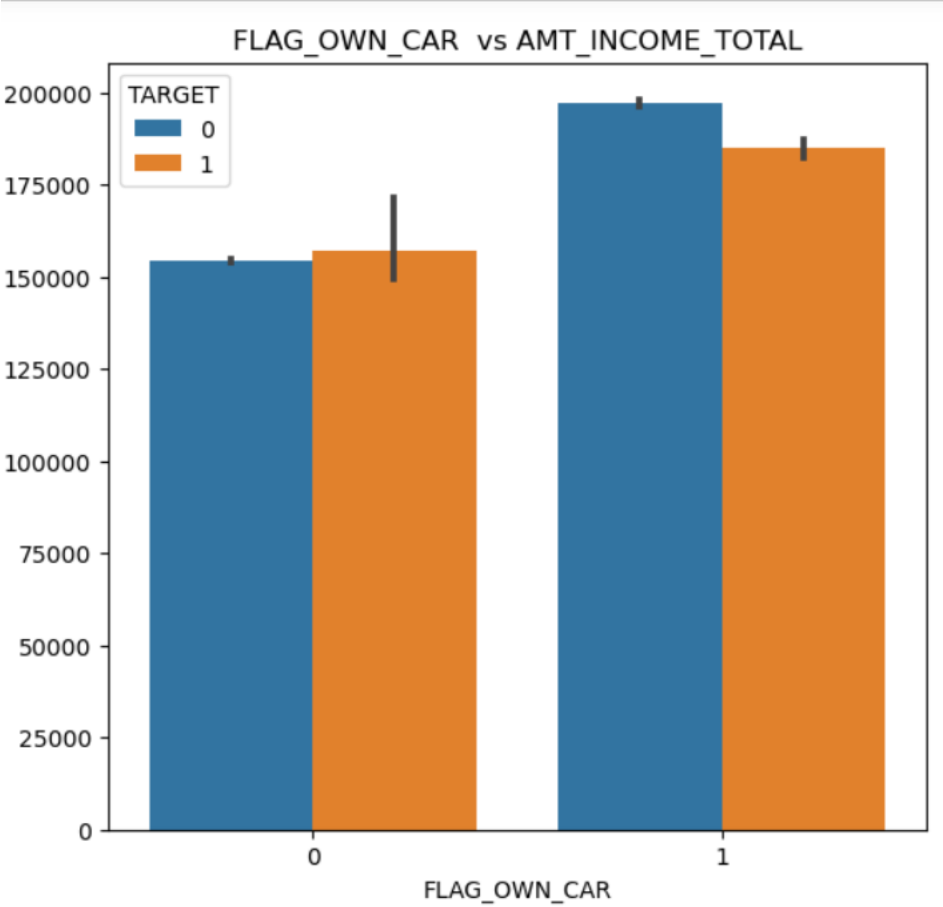
- NAME_EDUCATION_TYPE vs TARGET: Client’s whose qualification is lower secondary has difficulties in repaying the loan.



- NAME_CONTRACT_TYPE vs YEARS_BIRTH: Clients who are in mid 40s and who availed cash loans have more difficulties in repaying the loan.

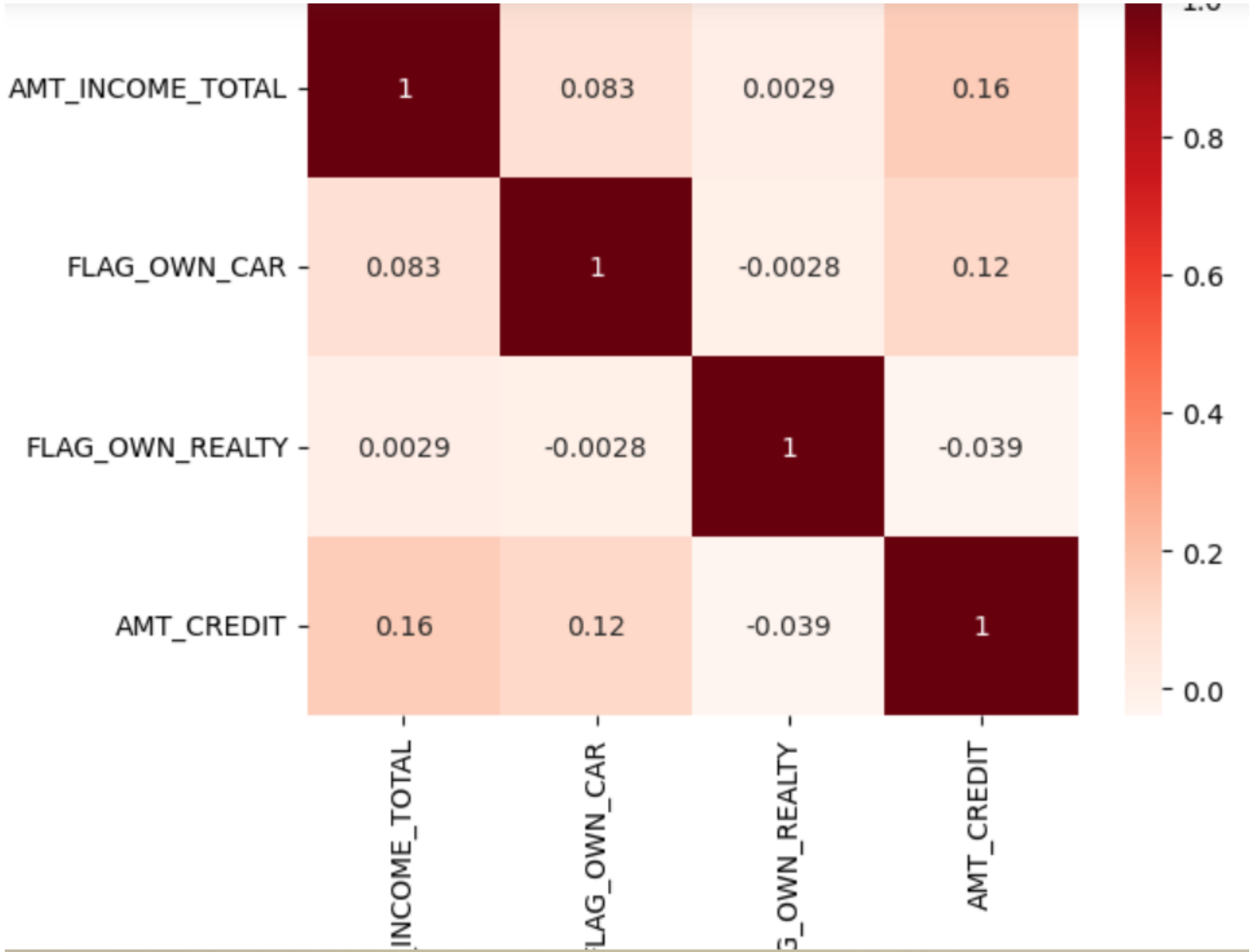


- FLAG_OWN_CAR vs AMT_INCOME_TOTAL: Clients with higher income and who owns a car has difficulties in paying the loan.



Session-4 Multivariate Analysis

- AMT_INCOME_TOTAL vs FLAG_OWN_CAR vs FLAG_OWN_REALTY vs AMT_CREDIT: This is the correlation matrix of client's income, whether they own car and house/flat.



	INCOME_TOTAL	LAG_OWN_CAR	G_OWN_REALTY	AMT_CREDIT
AMT_INCOME_TOTAL	1	0.083	0.0029	0.16
FLAG_OWN_CAR	0.083	1	-0.0028	0.12
FLAG_OWN_REALTY	0.0029	-0.0028	1	-0.039
AMT_CREDIT	0.16	0.12	-0.039	1

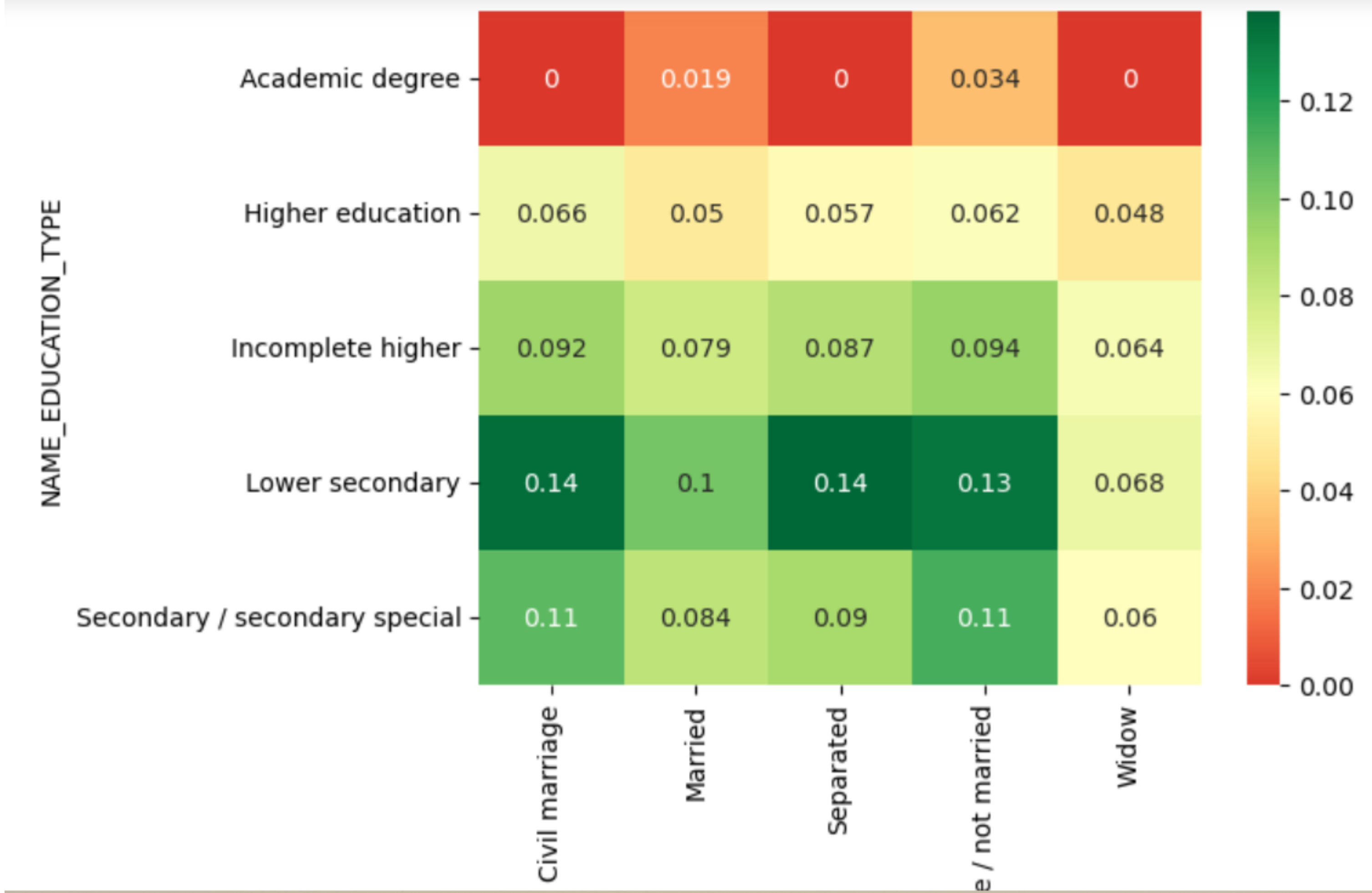
- NAME_EDUCATION_TYPE vs NAME_FAMILY_STATUS vs TARGET: Client’s whose educations is lower secondary

  and marital status as separated

  or civil marriage have more

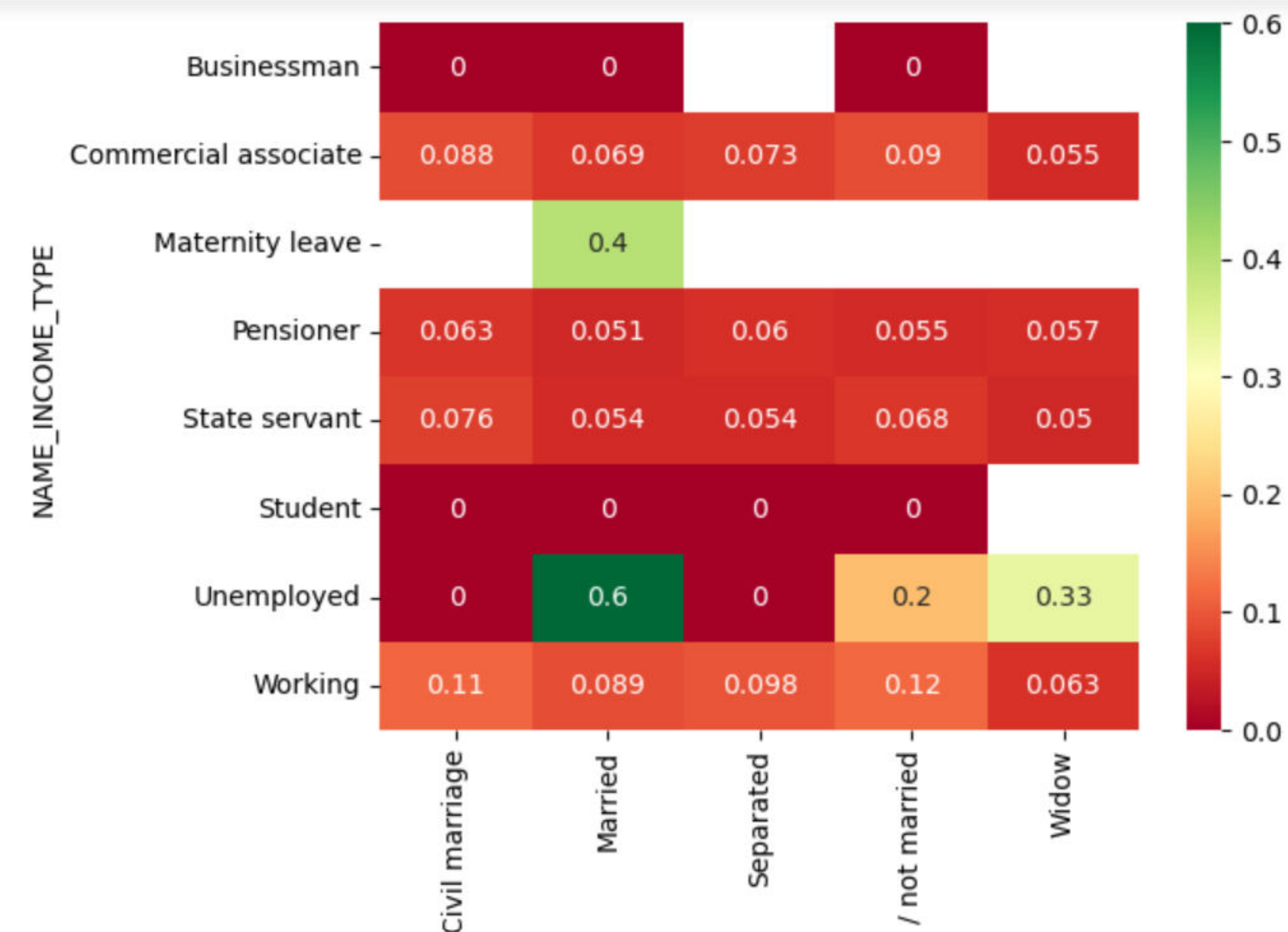
  difficulties in repaying the

  loan.

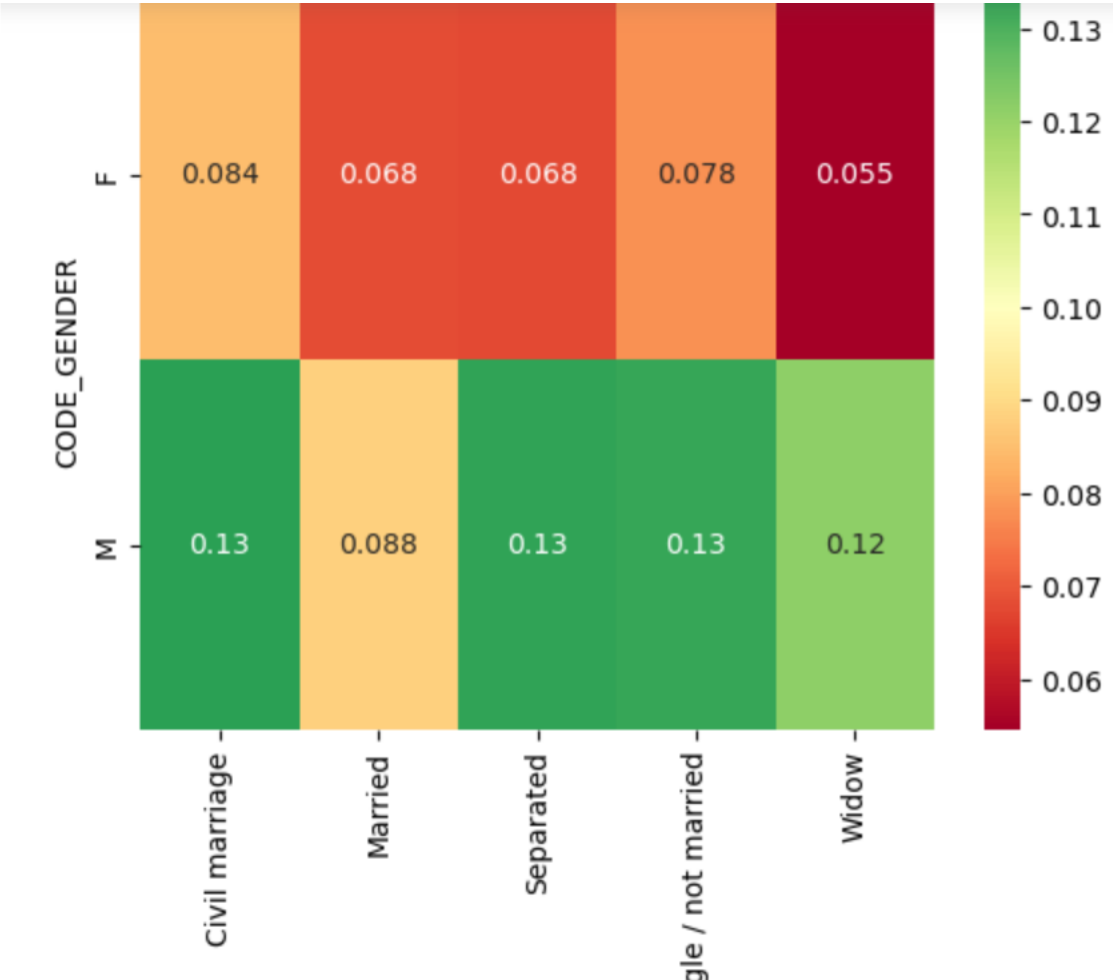


- NAME_INCOME_TYPE vs NAME_FAMILY_STATUS vs TARGET: Clients who are unemployed and married have

more difficulties in repaying the loan.



- CODE_GENDER vs NAME_FAMILY_STATUS vs TARGET: Males seems have more difficulties in repaying the loan and whose family status is civil marriage, Separated, Single/ not married.



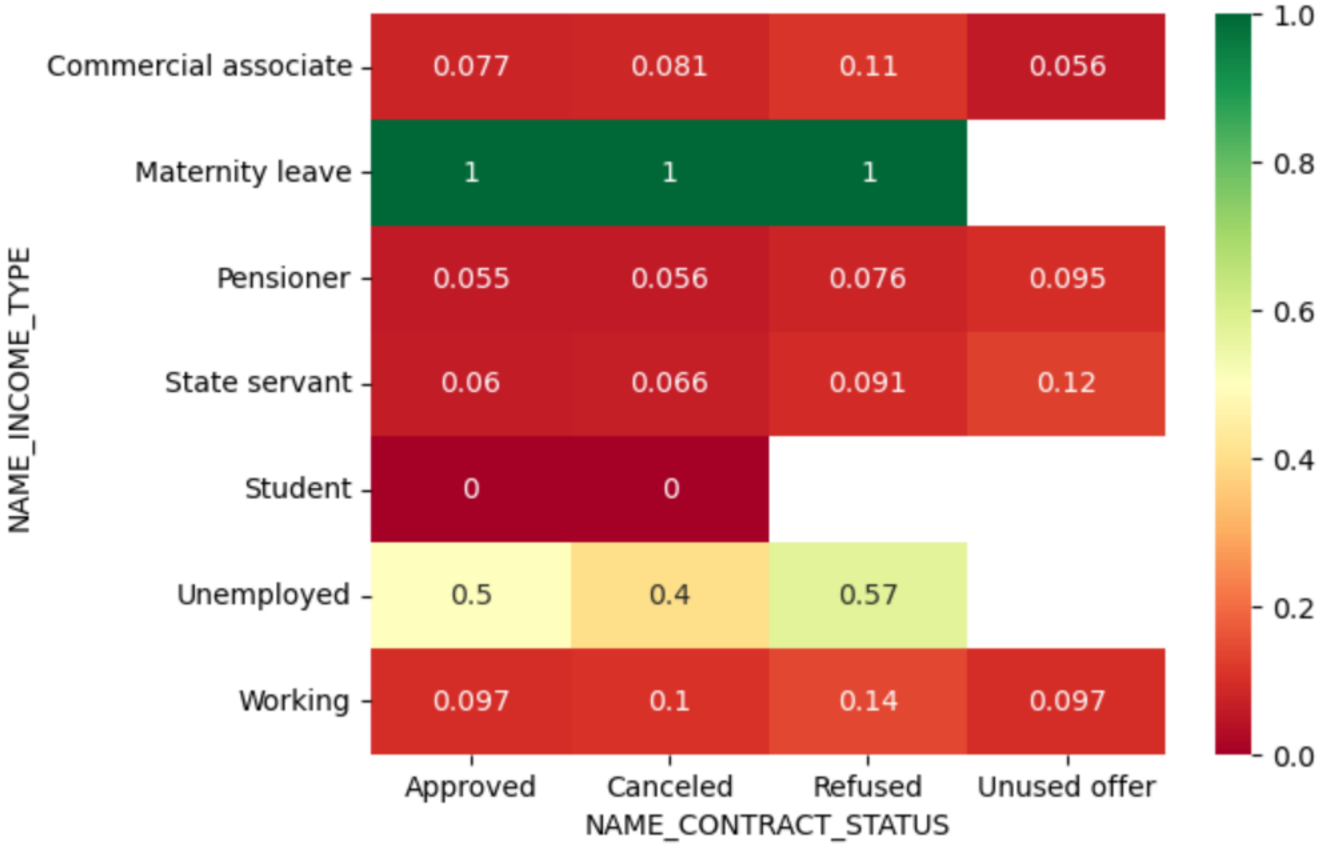
Session-5 Previous application dataset

Business Objective: Clients capable of repaying the loan but applications are rejected.

(Analysing previous_application dataset with application_data dataset.)

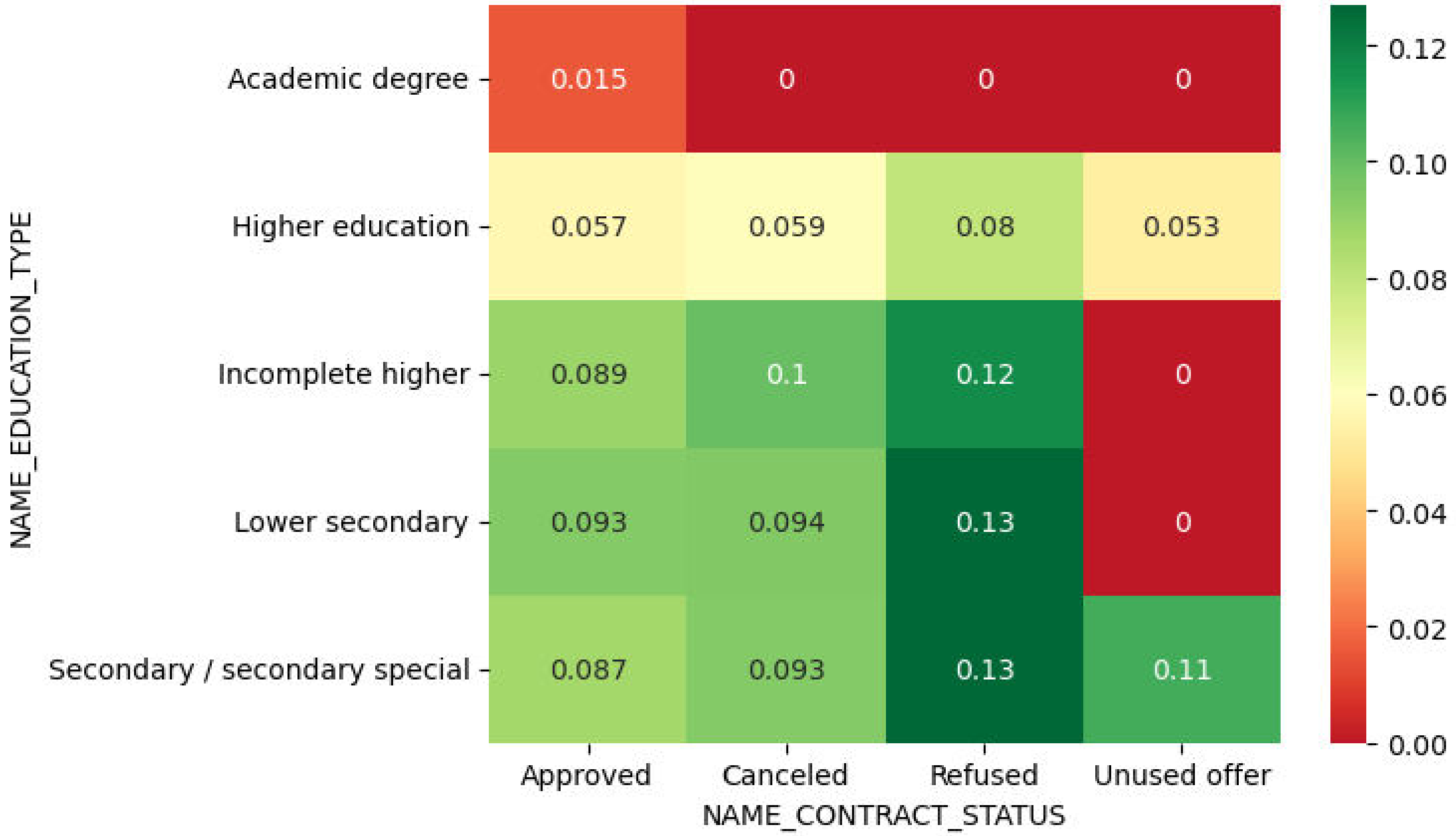
> Extract the required variables from previous_application dataset and merge with application_data dataset to understand the patterns.

- NAME_INCOME_TYPE vs NAME_CONTRACT_STATUS vs TARGET: Clients of Pensioner and State Servant, whose applications were rejected earlier are capable of repaying the loan.

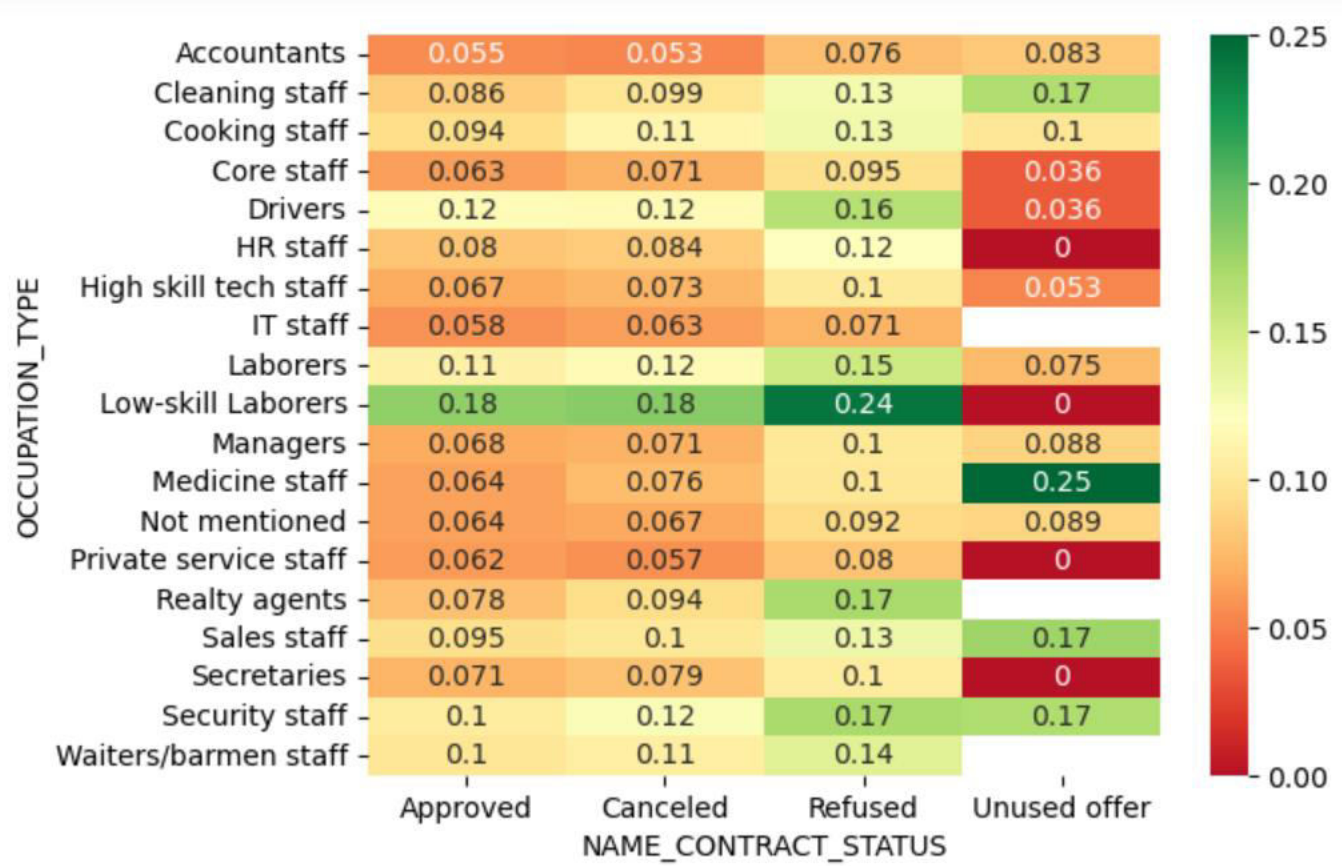


- NAME_EDUCATION_TYPE vs NAME_CONTRACT_STATUS vs TARGET: Clients of Higher education, whose applications were rejected earlier are capable of repaying the loan compared to

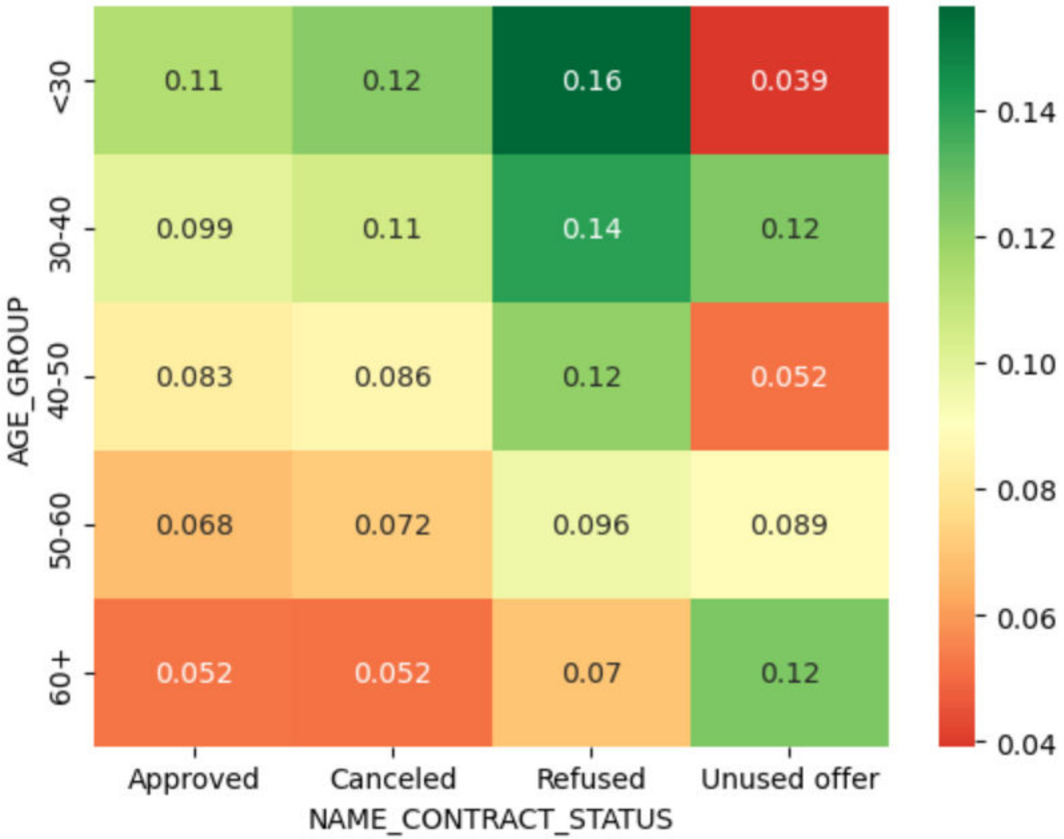
  clients with other qualifications



- OCCUPATION_TYPE vs NAME_CONTRACT_STATUS vs TARGET: IT staff and Accountants, whose applications were rejected earlier are capable of repaying the loar



- AGE_GROUP vs NAME_CONTRACT_STATUS vs TARGET: Age group of 60+, whose applications were rejected earlier are capable of repaying the loan.


AGE_GROUP
<30
NAME_CONTRACT_STATUS
Approved
Canceled
Refused
Unused offer