Credit EDA Assignment

By: Deeksha Pant

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.
- 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

- This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

- In this case study, we will use EDA to understand how consumer attributes and loan attributes influence the tendency to default to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

Datasets

- Application_data: It contains information of clients at the time of application.
- Previous_application: Contains information about client's previous loan data
- Columns_description: Data dictionary with description of all columns.

Approach & Methodology

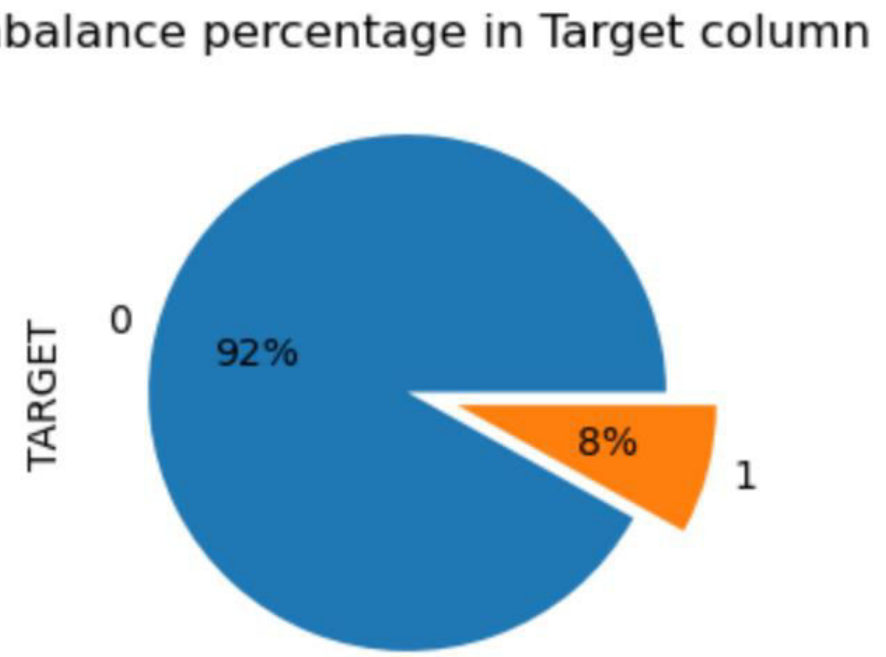
- Understanding the problem statement & columns significance using Data dictionary
- Loading the dataset provided in assignment using pandas functions
- Checking the structure of the data using methods like info(), shape(), describe() etc.
- Data cleaning
  - Fixing Missing value by Imputation of Removing rows/Columns
  - Handling the outliers
- Performing Univariate Analysis
- Bivariate & Multivariate Analysis

Data Imbalance

There is imbalance of 11.39% in application dataset for Target variable

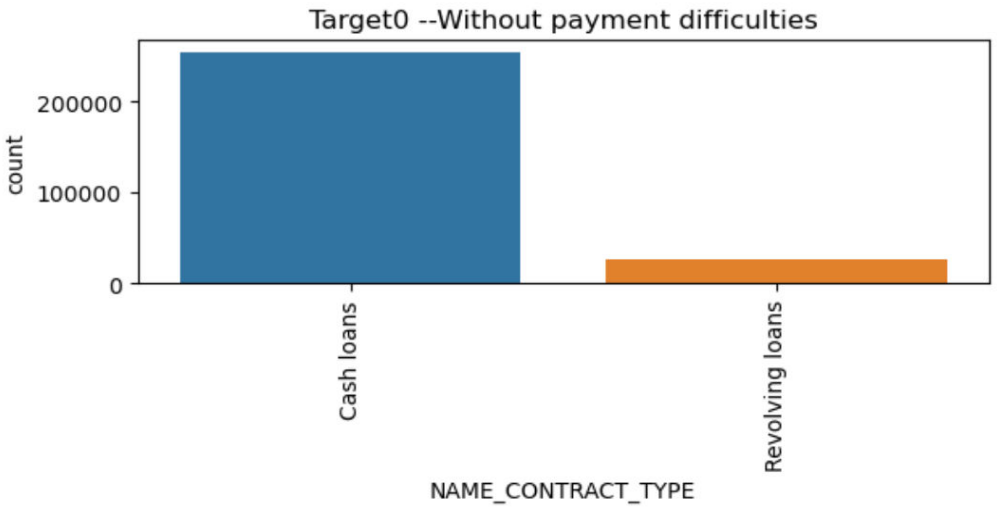
1 - client with payment difficulties,

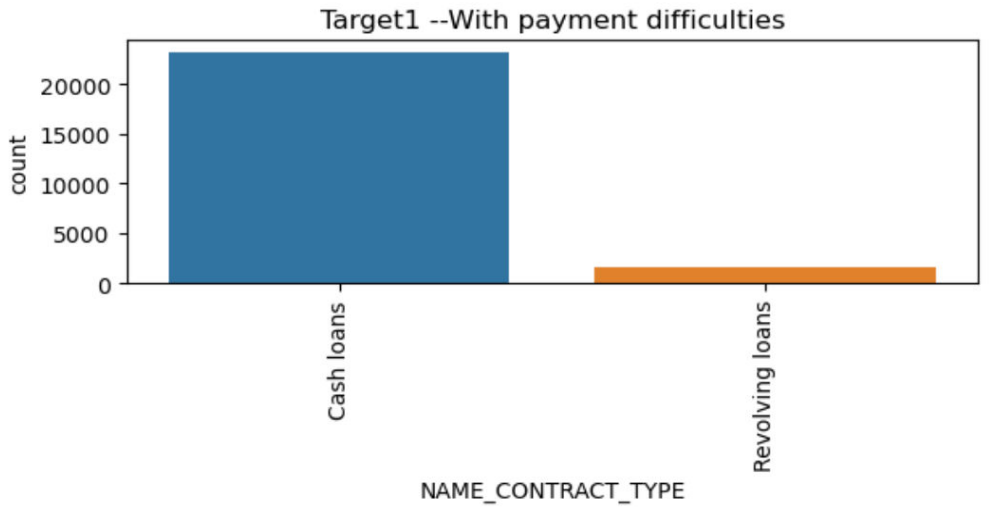
0 - all other cases



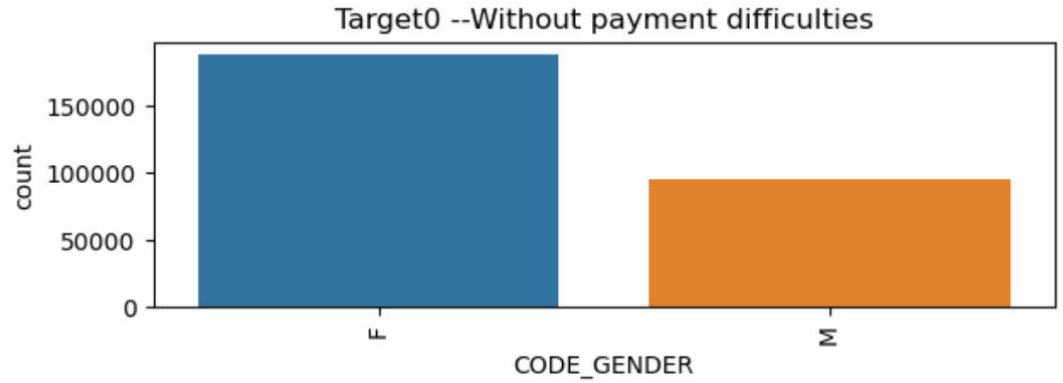
Application Dataset- Univariate Analysis

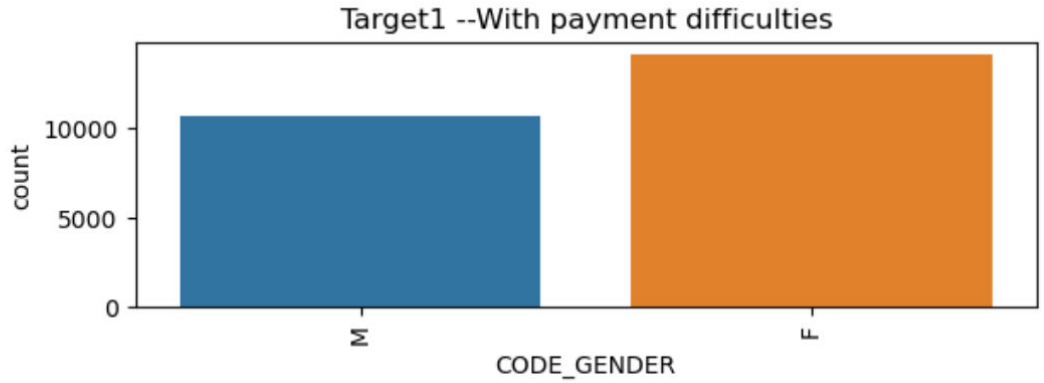
'Cash Loans' Contract Type has highest number of frequency irrespective of With/Without payment difficulties.





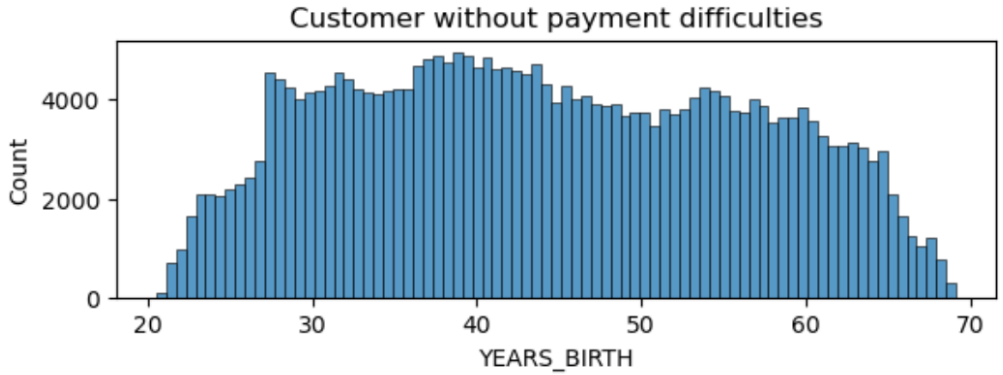
- Females are more in both With/Without payment difficulties cases.
- An increase is observed in the case of Male With Payment difficulties.

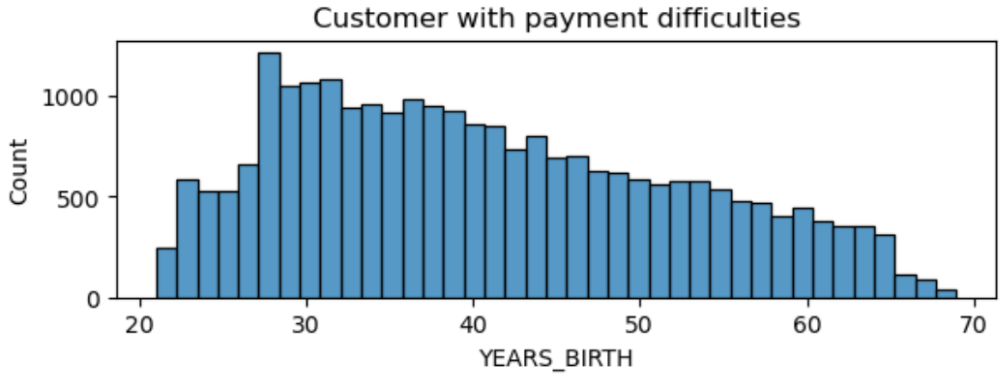




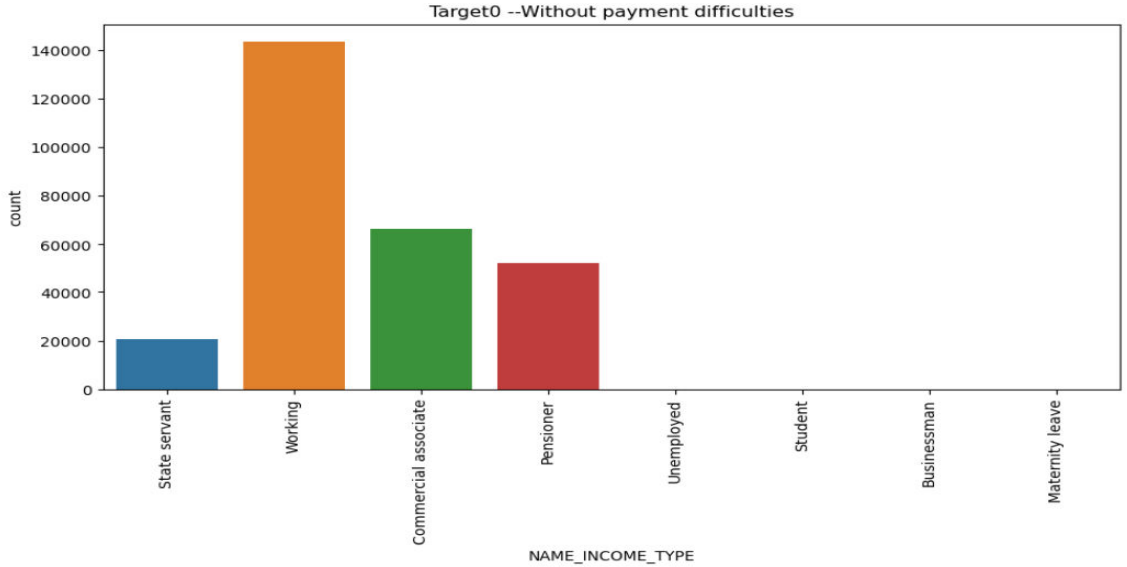
Univariate Analysis(Application Dataset) Contd..

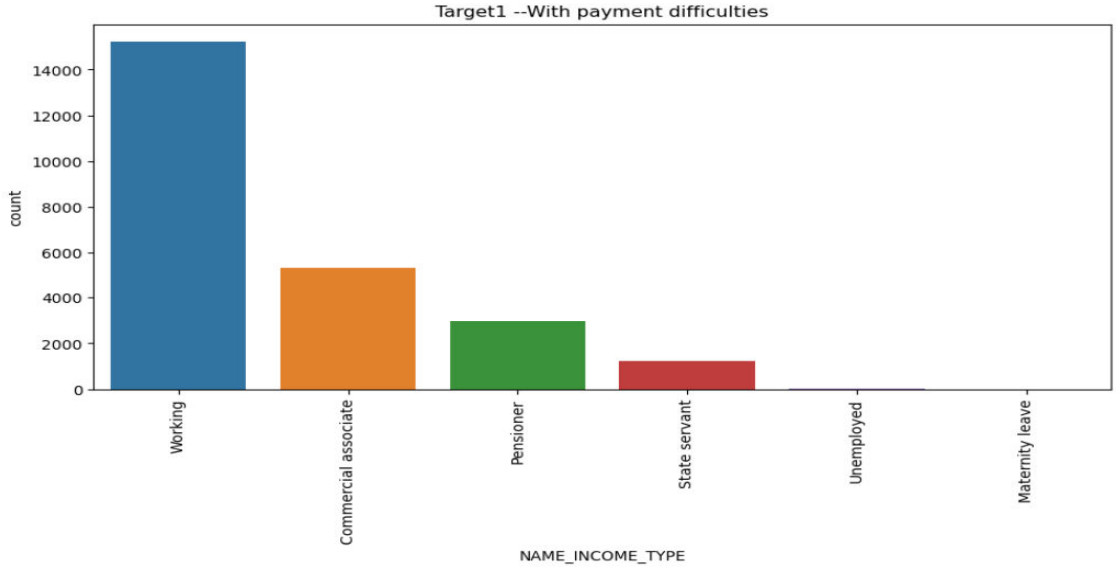
- Customers having no payment difficulties are spread across uniformly between 20-70 age group
- Majority of customers of age group <40 are having payment difficulties.





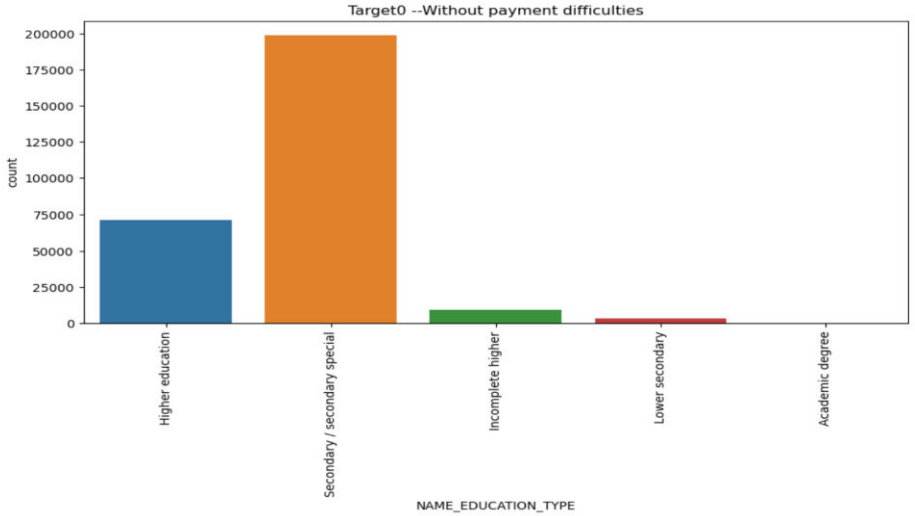
- 'Working' income type has highest cases in With/Without payment difficulties.
- There is decrease in percentage in 'State Govt', 'Commercial associate' & 'Pensioners' in case of 'With Payment difficulties'.

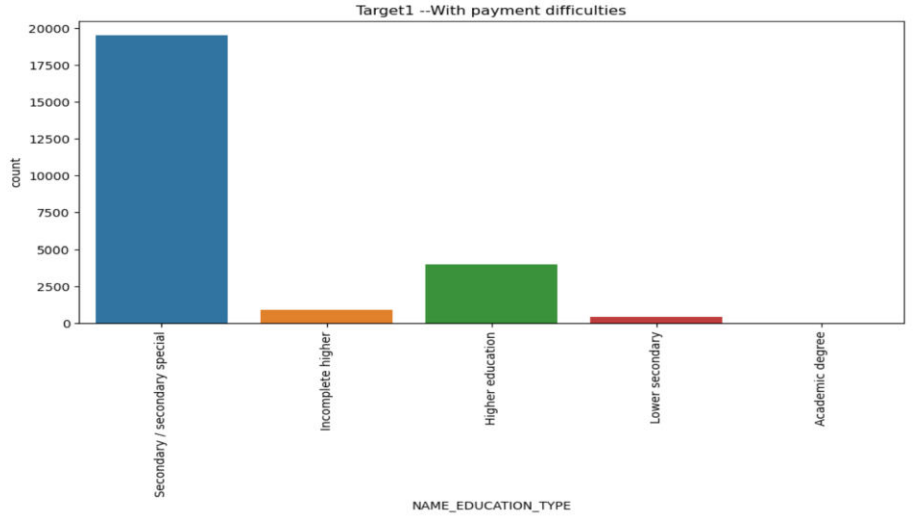




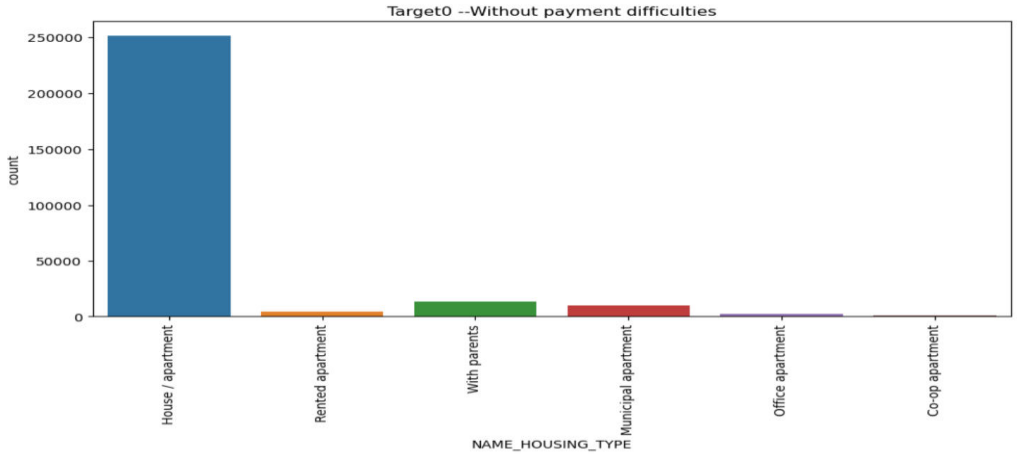
Univariate Analysis(Application Dataset) Contd..

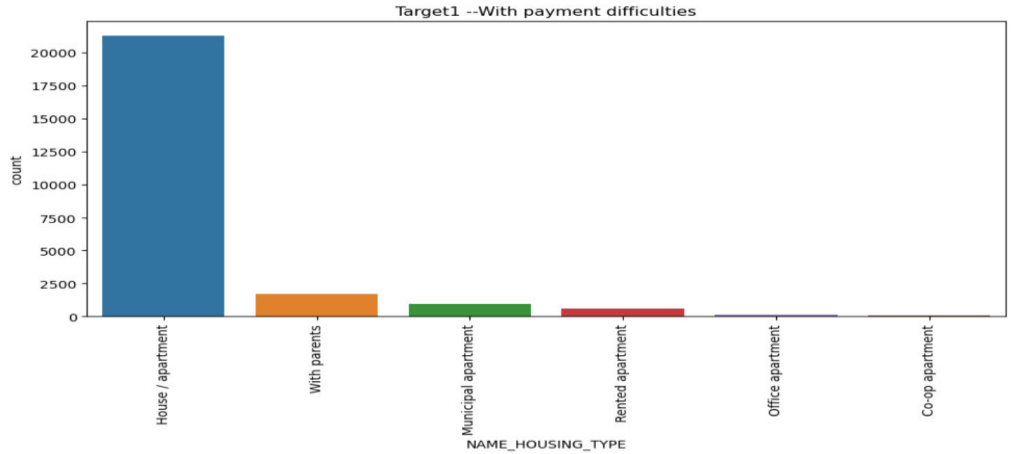
- Secondary Education type has the highest frequency irrespective of With/Without payment difficulties.
- Customers is having 'Higher education' is having less Payment difficulties.





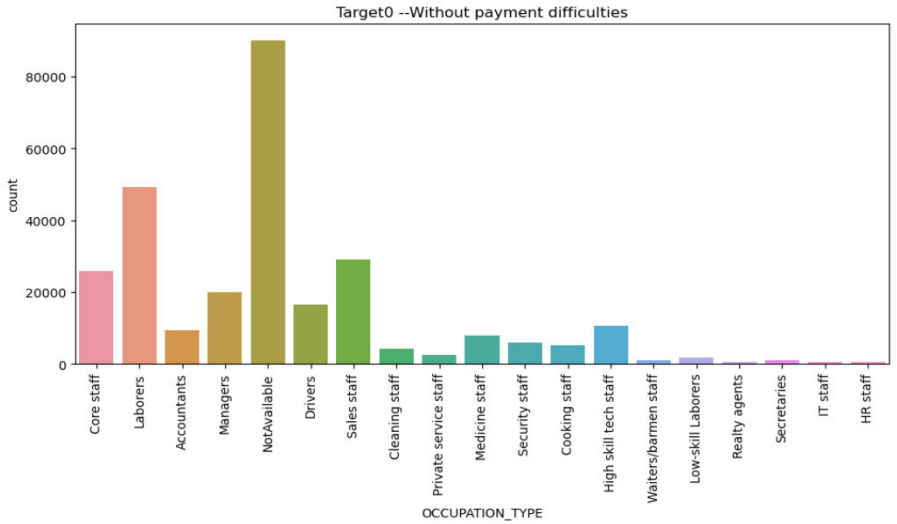
- Customers living in House/Apartment has the highest frequency irrespective of With/Without payment difficulties.
- Customers living with parents are having more Payment difficulties.

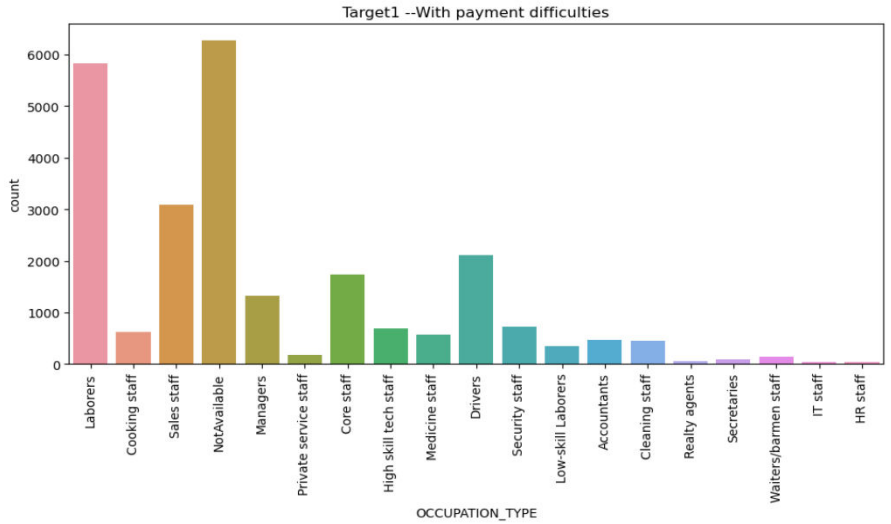




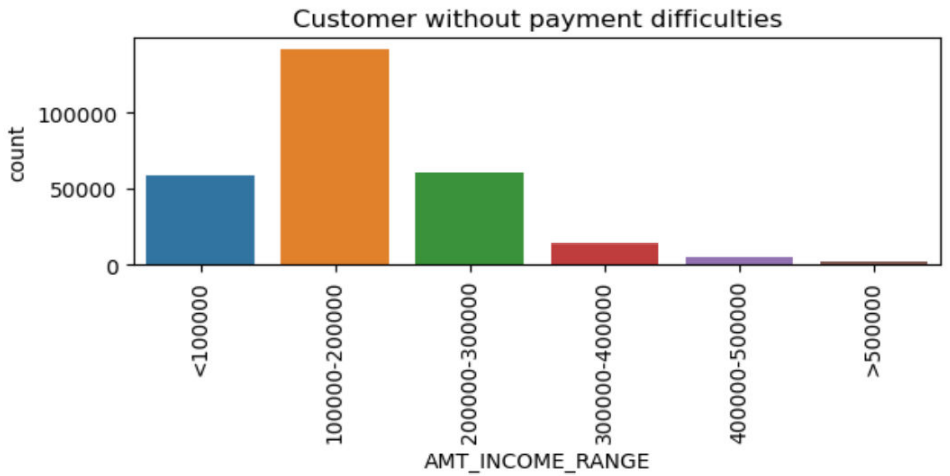
Univariate Analysis(Application Dataset) Contd..

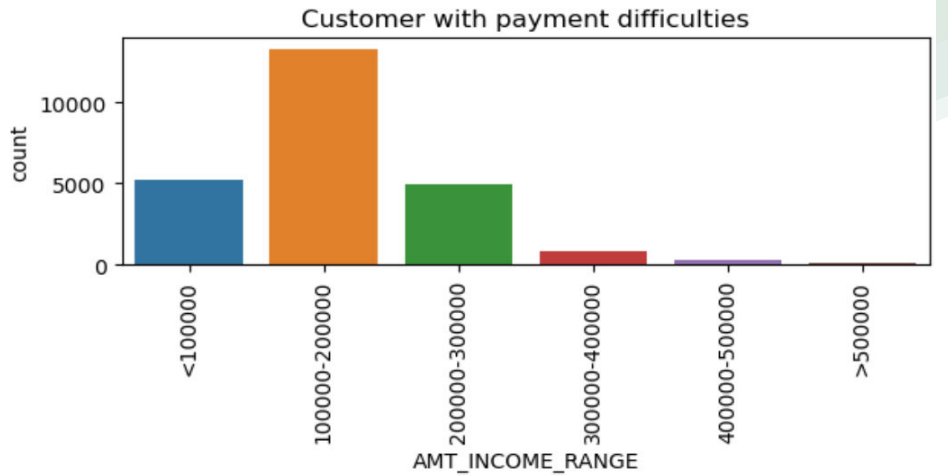
- Among non-missing data, 'Laborers' has the highest frequency irrespective of With/Without payment difficulties.
- In case of Payment difficulties ,their percentage is more compared to other occupation types.





- Low Income Customers are susceptible to both default and non-default scenarios.
- High income group are very less in both cases



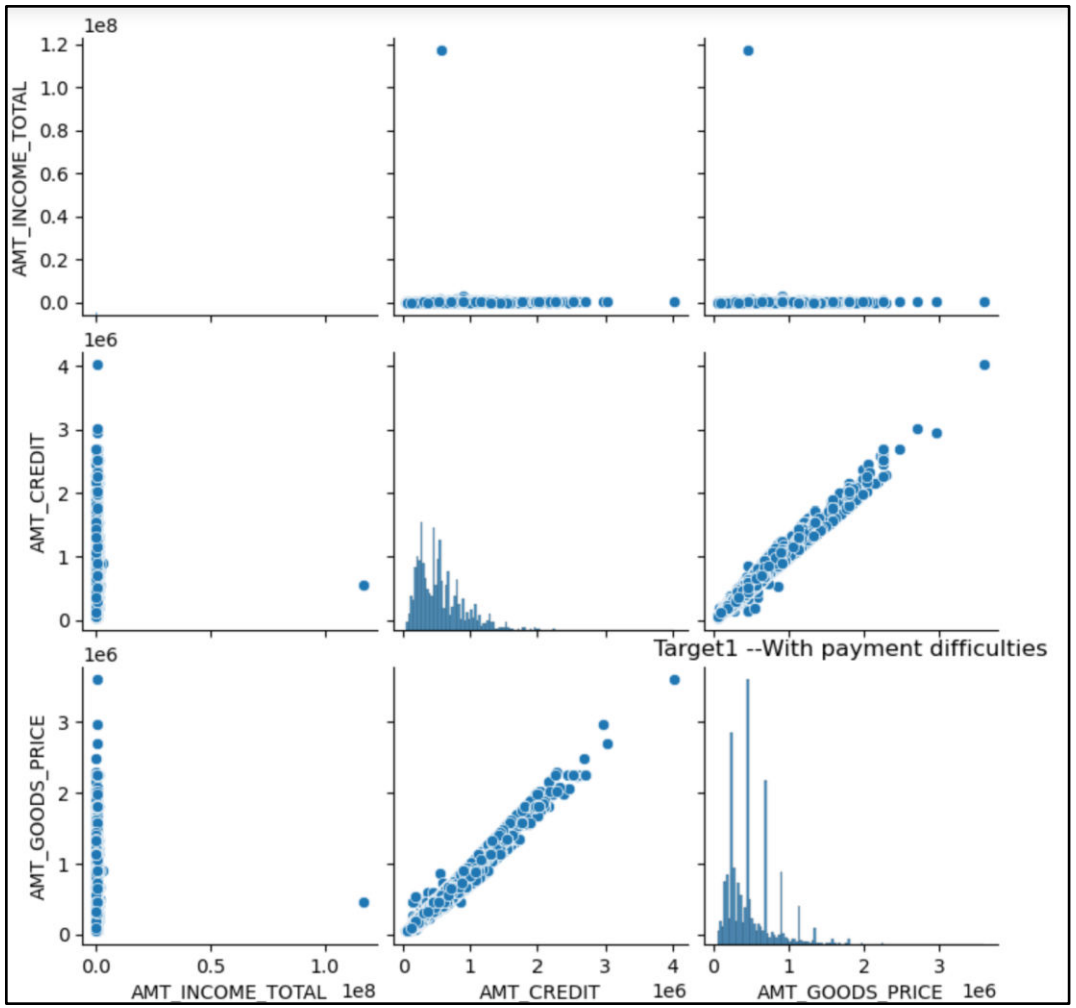


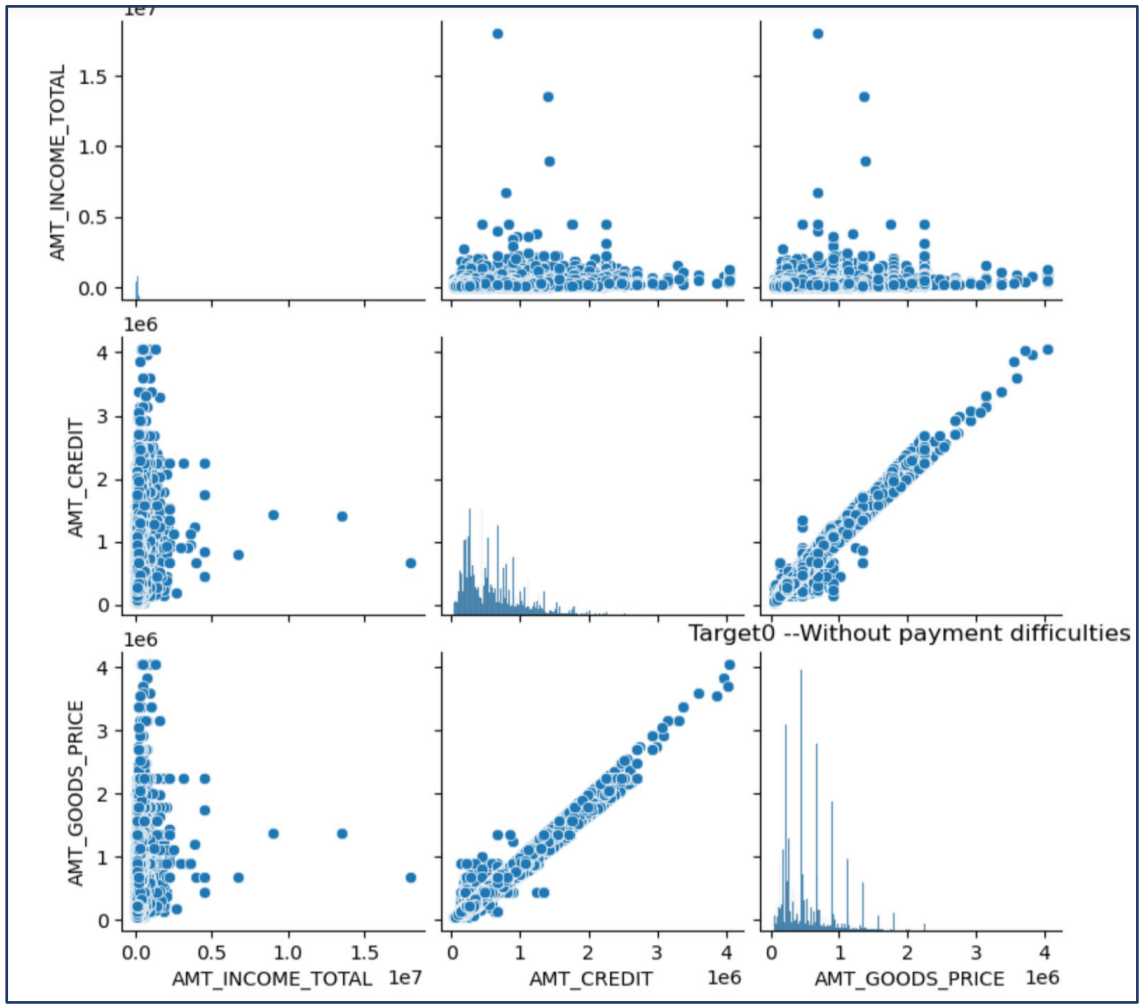
Bivariate Analysis (Numerical-Numerical)

- For both default and non-default categories, following columns has high linear co-relationship:

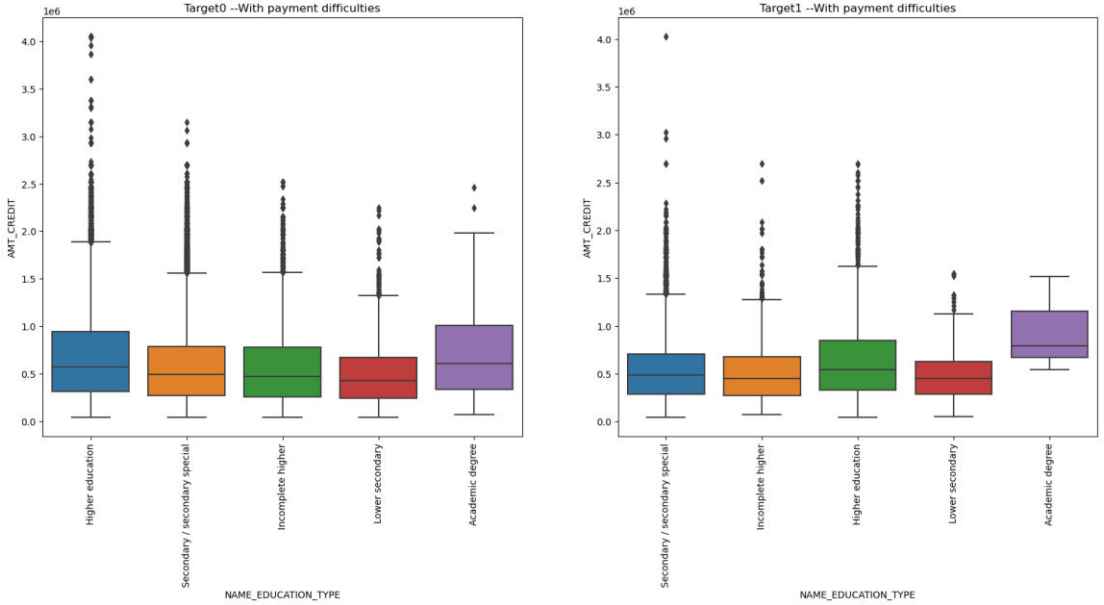
'AMT_CREDIT'-- 'AMT_GOODS_PRICE'

- Customers having larger Amount of Credit and more Goods Price have low probability to default
- For Customers with less income take more loan amount and more income have less difficulties in payments

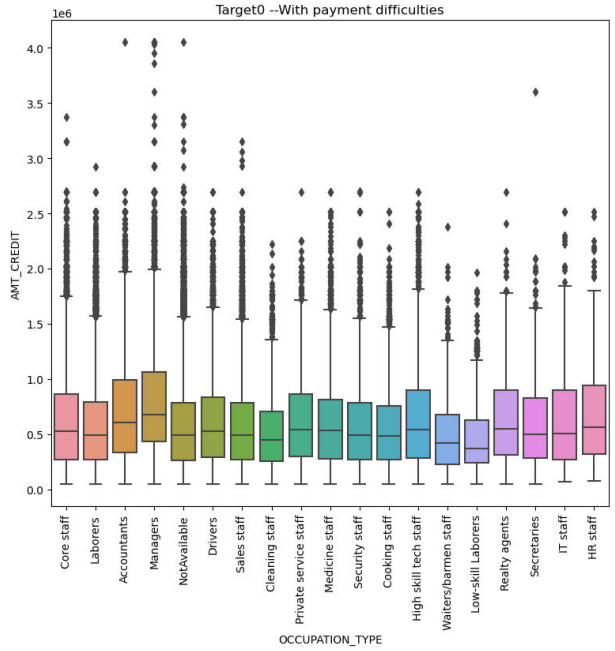


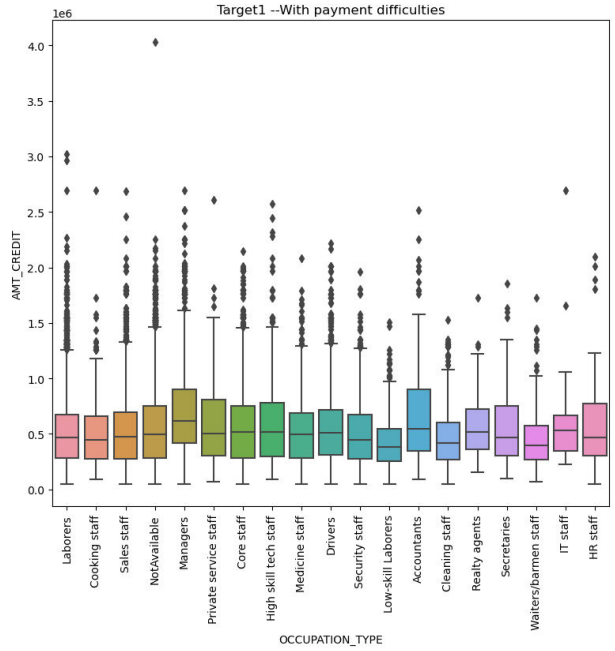


Bivariate Analysis(Numerical -Categorical)



- Amount Credit is highest/higher Quartile for 'Academic degree' & lowest/lower Quartile for 'Lower Secondary' Education Type for both default /non default cases.

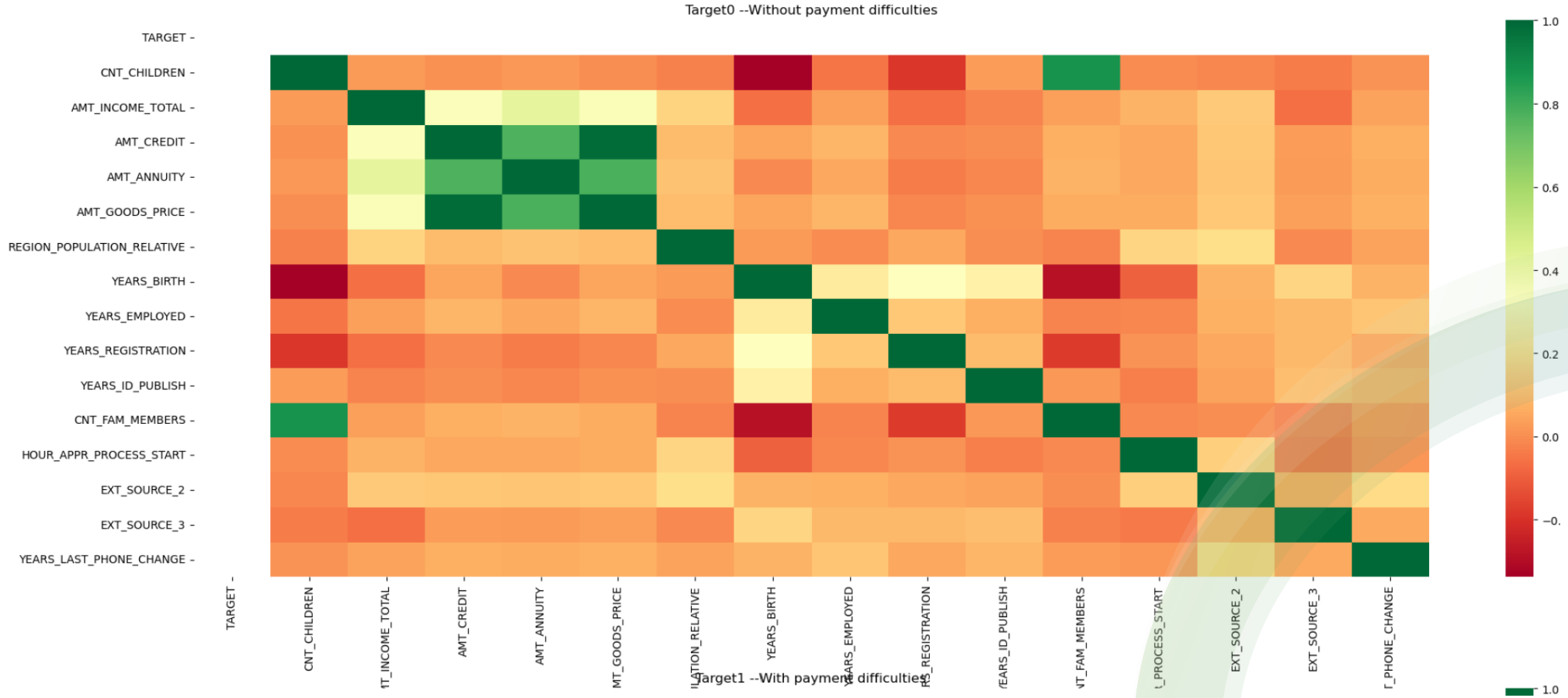




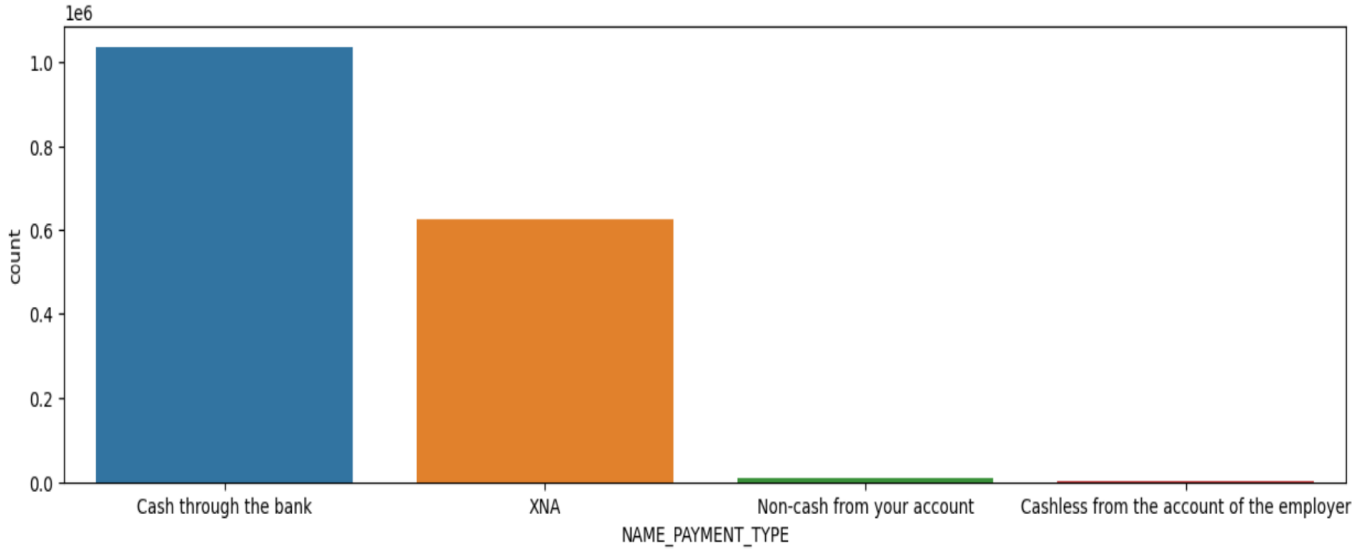
- Amount Credit is highest/higher Quartile among 'Managers' & 'Accountants' & lowest/lower Quartile for 'Low-Skill Laborers' Occupation Type for both default /non default cases.

Multivariate Analysis

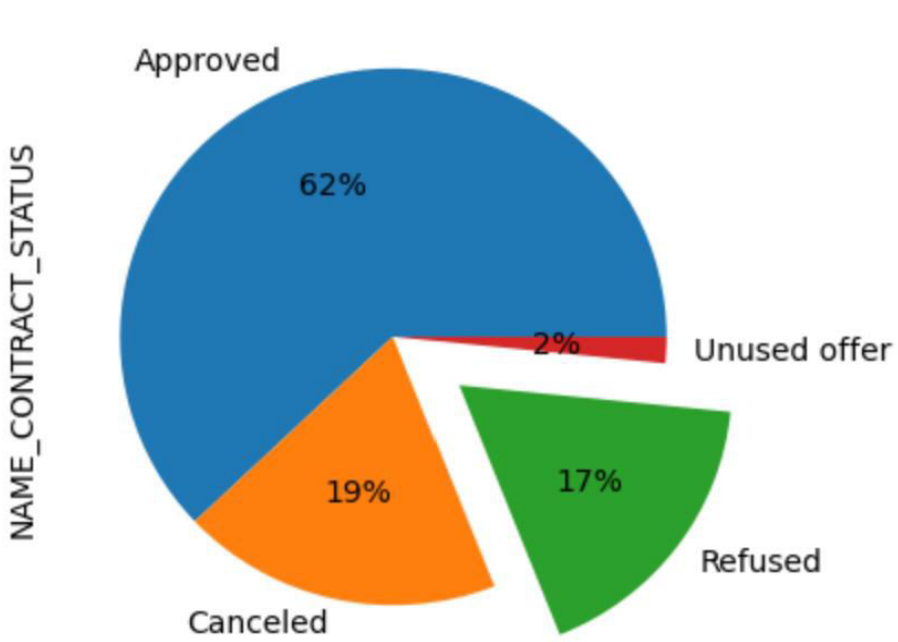
- For Non defaulters, there is
- strong co-relation between 'AMT_CREDIT'-- 'AMT_GOODS_PRICE'
- good co-relation between 'AMT_ANNUITY'-- 'AMT_GOODS_PRICE' & 'AMT_ANNUITY'--'AMT_CREDIT' & 'CNT_FAM_MEMBERS'—'CNT_CHILDREN'
- & low relation between 'CNT_CHILDREN'—'YEARS_BIRTH' & 'CNT_FAM_MEMBERS'—'YEARS_BIRTH'



Data Analysis on Previous Application Dataset



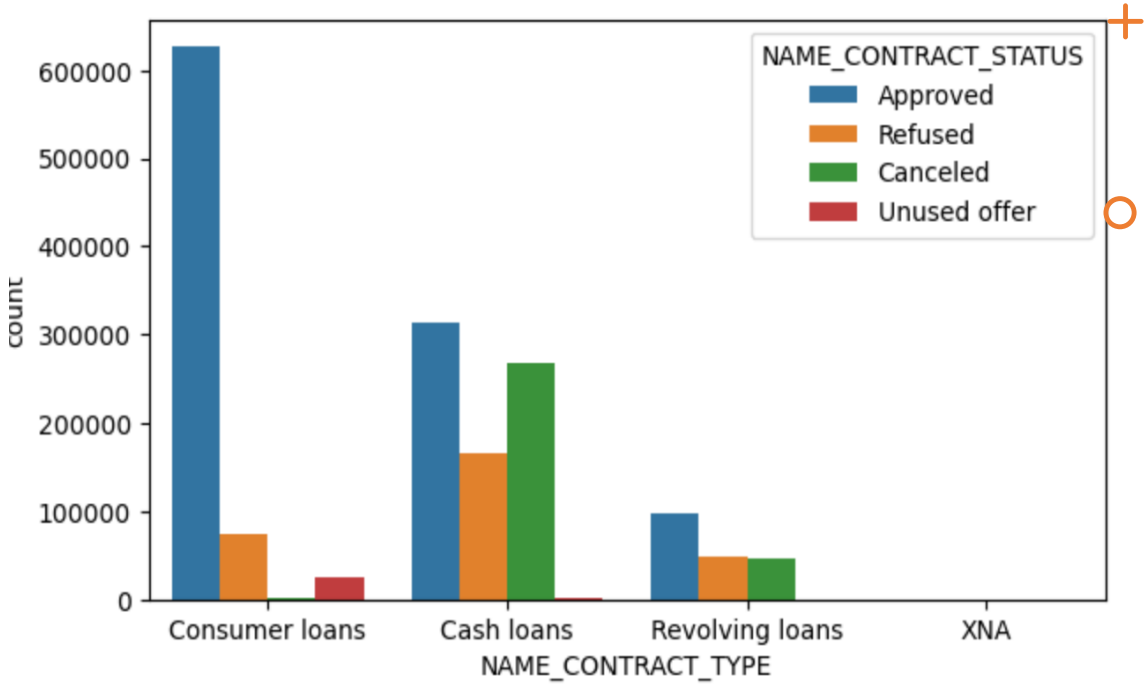
- Highest percentage of NAME_CONTRACT_STATUS in applications are 'Approved' and lowest are 'Unused Offer'.

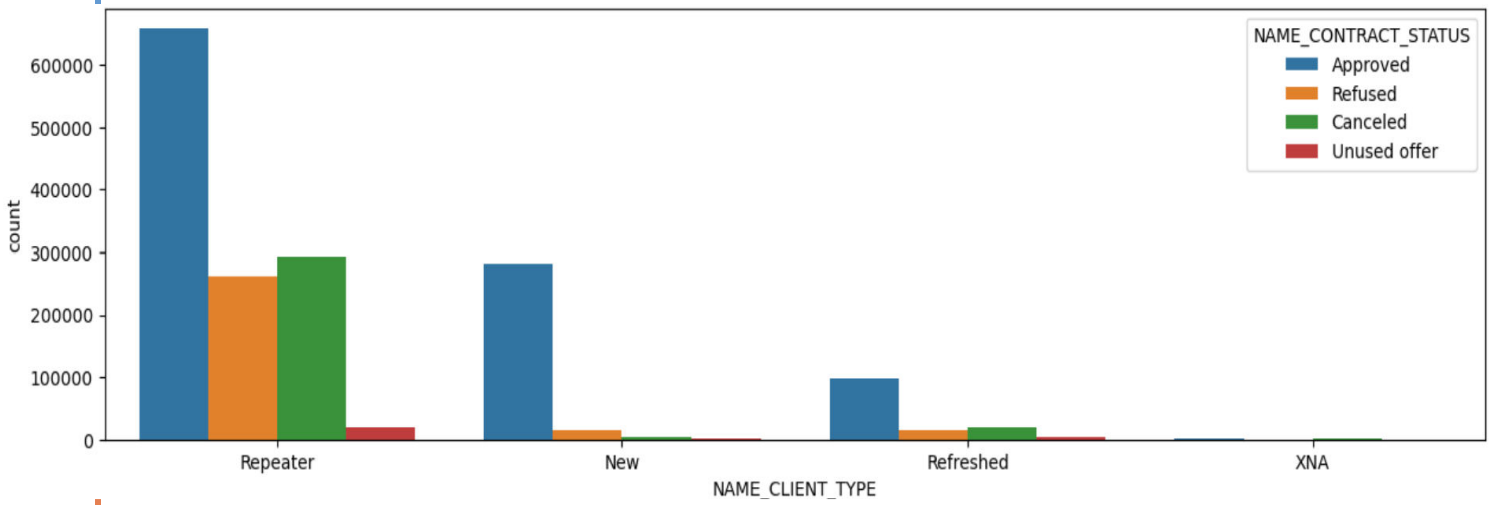


- Maximum Number of Payments in Previous Applications are done with 'Cash through Bank' and lowest with 'Cashless from account'.

Bivariate Analysis:

- Major chunk of consumer loans were approved compared to other contract status & also types of loans.
- In case of Cash loans, ratio of Approvals to Cancellation is more compared to other loans

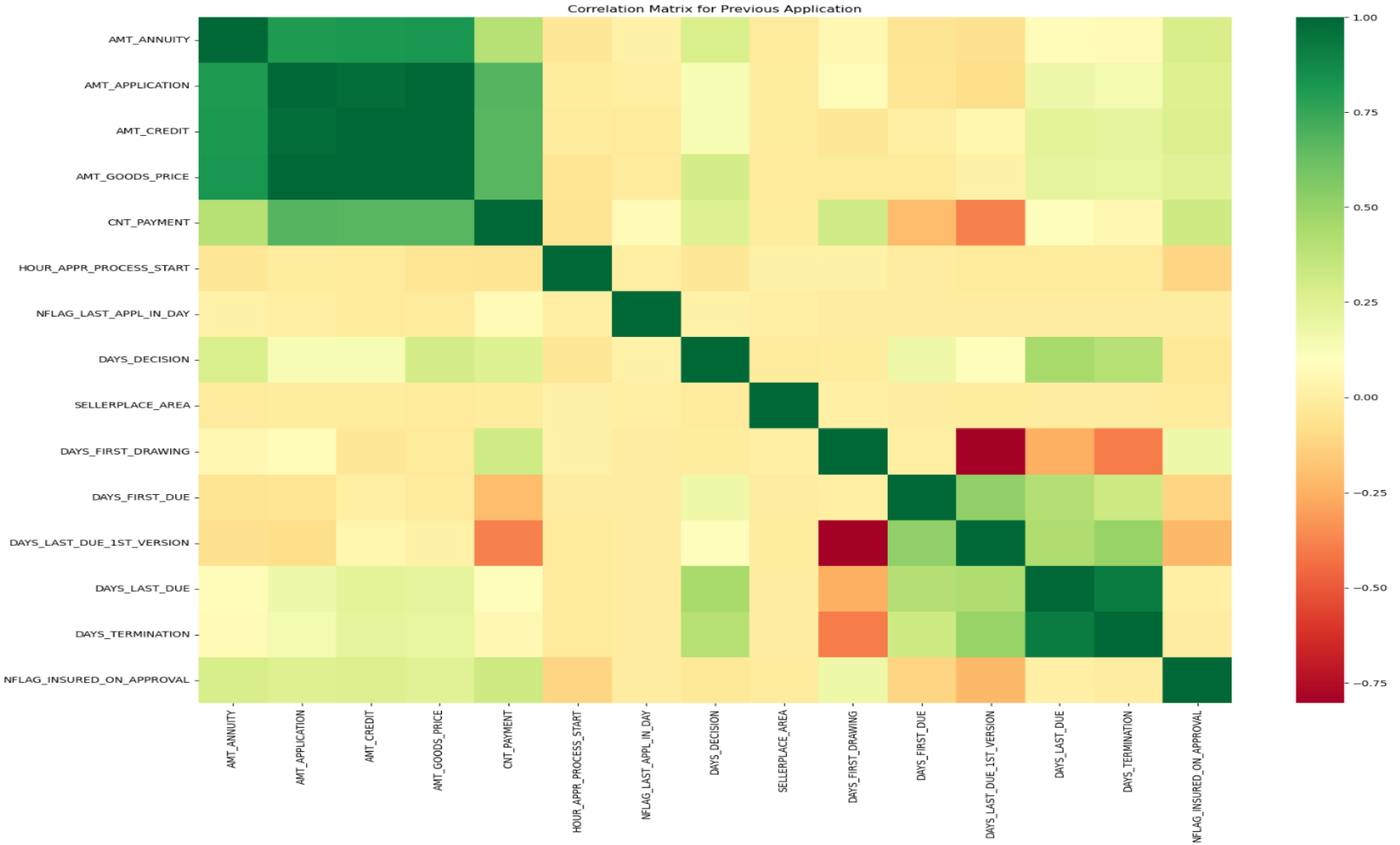




- Among all types of clients, repeaters are highest. Also, cancellations are more than loan refusals.
- Leaving XNA category, Refreshed client type has the lowest percentage among all client types.

Multivariate Analysis

- Highest co-relationship is between 'AMT_CREDIT','AMT_GOODS_PRICE' and 'AMT_APPLICATION'.
- DAYS_LAST_DUE & DAYS_TERMINATION have good correlation.
- DAYS_LAST_DUE_1ST_VERSION & DAYS_FIRST_DRAWING has poor correlation between them.



Top & Bottom Co-Related columns are:

Application Dataset:

OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998382
AMT_CREDIT	AMT_GOODS_PRICE	0.986736
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950842
CNT_CHILDREN	CNT_FAM_MEMBERS	0.878869
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.860627
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.858342
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.825574
AMT_ANNUITY	AMT_GOODS_PRICE	0.774835
	AMT_CREDIT	0.770124
dtype: float64		
HOUR_APPR_PROCESS_START	REGION_RATING_CLIENT	-0.285697
EXT_SOURCE_2	REGION_RATING_CLIENT_W_CITY	-0.288011
REGION_RATING_CLIENT	EXT_SOURCE_2	-0.292619
CNT_CHILDREN	YEARS_BIRTH	-0.331448
FLAG_DOCUMENT_8	FLAG_DOCUMENT_3	-0.465743
FLAG_DOCUMENT_6	FLAG_DOCUMENT_3	-0.486252
REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT_W_CITY	-0.531542
	REGION_RATING_CLIENT	-0.532882
FLAG_EMP_PHONE	FLAG_DOCUMENT_6	-0.597731
YEARS_BIRTH	FLAG_EMP_PHONE	-0.619890

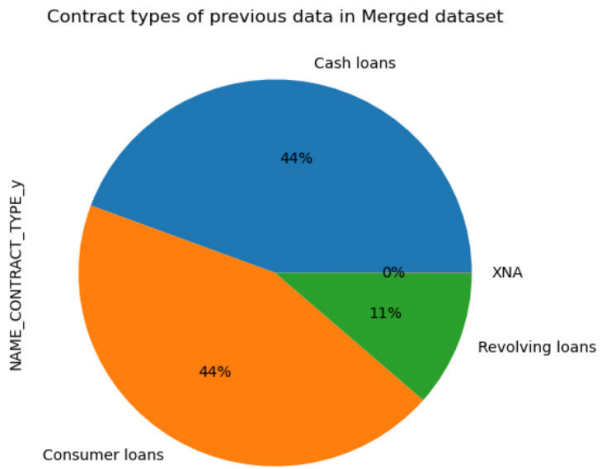
Previous Application Dataset:

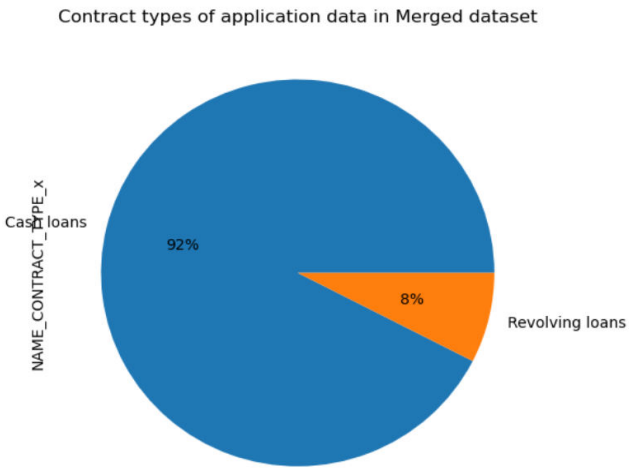
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998270
AMT_GOODS_PRICE	AMT_CREDIT	0.982783
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637
CNT_CHILDREN	CNT_FAM_MEMBERS	0.885484
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.869016
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.847885
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.778540
AMT_ANNUITY	AMT_GOODS_PRICE	0.752295
	AMT_CREDIT	0.752195
dtype: float64		
CNT_CHILDREN	YEARS_BIRTH	-0.259109
HOUR_APPR_PROCESS_START	REGION_RATING_CLIENT_W_CITY	-0.275703
REGION_RATING_CLIENT	HOUR_APPR_PROCESS_START	-0.293908
REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT	-0.443236
	REGION_RATING_CLIENT_W_CITY	-0.446977
FLAG_DOCUMENT_6	FLAG_DOCUMENT_3	-0.475807
FLAG_DOCUMENT_3	FLAG_DOCUMENT_8	-0.528927
FLAG_EMP_PHONE	YEARS_BIRTH	-0.578519
FLAG_DOCUMENT_6	FLAG_EMP_PHONE	-0.617421

Merged Dataset (Application & Previous)

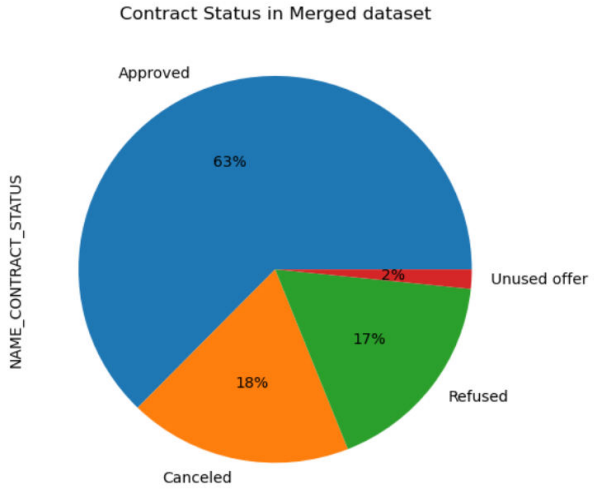
Univariate Analysis:

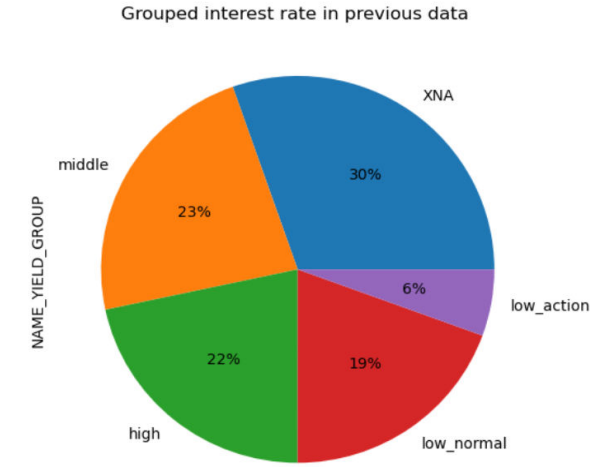
- Cash Loans are more in application dataset than Previous dataset.
- Consumer Loans are not present in application dataset.
- Compared to previous dataset, Revolving Loans have decreased in application dataset.





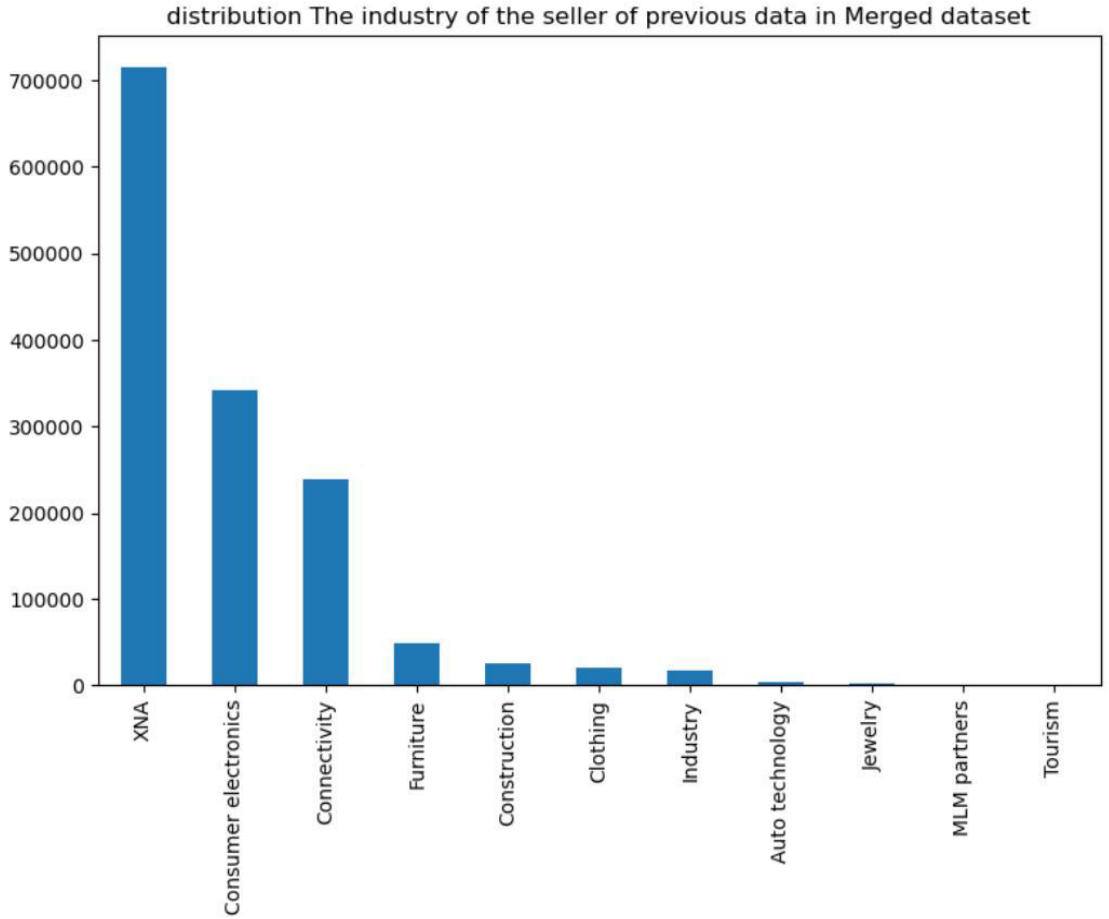
- For status of Contracts, 'Approved' status has the highest occurrence and 'Unused offer' has lowest.
- Among 'Grouped interest rate' medium category has highest percentage in the merged dataset.

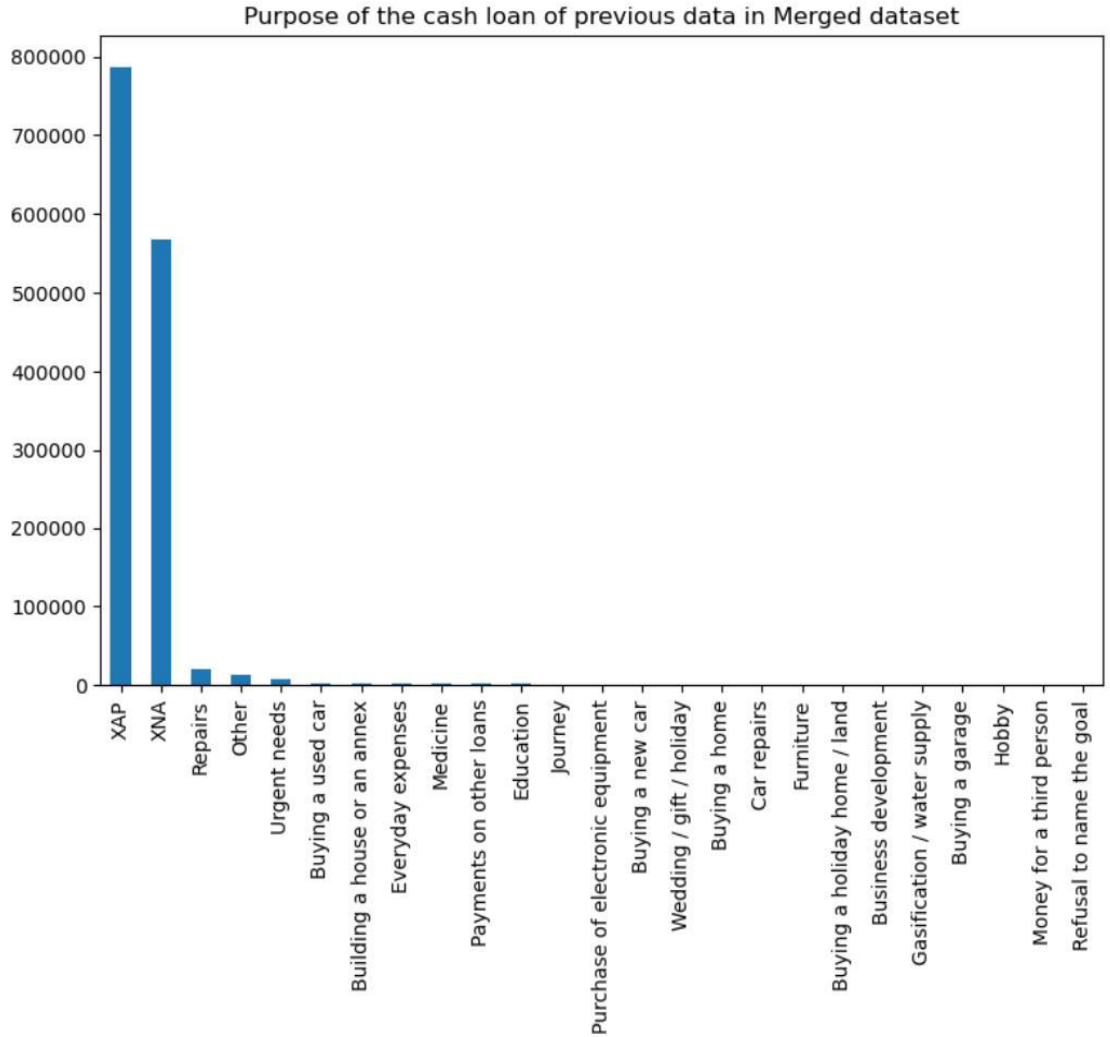




Univariate Analysis(Merged Dataset) Contd..

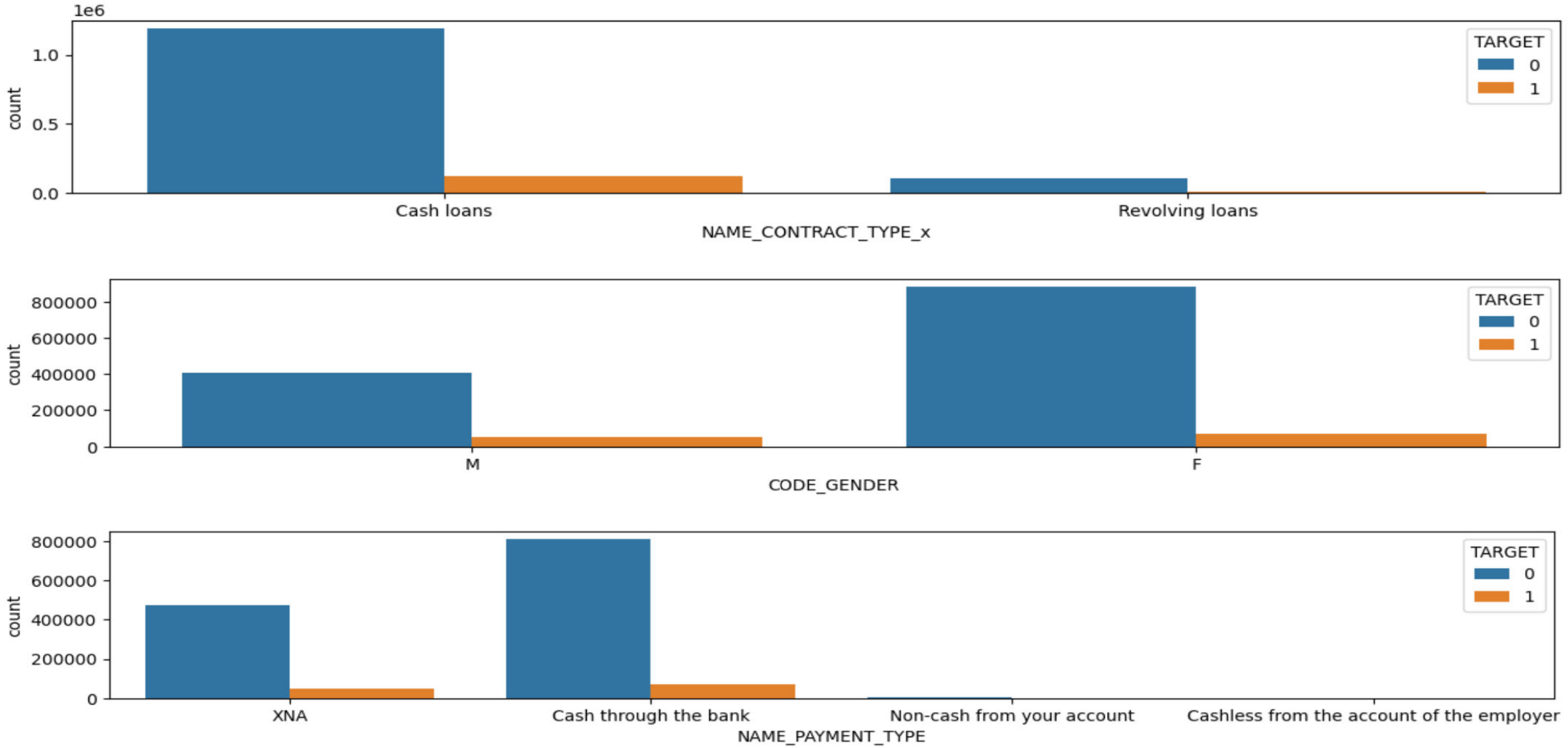
- 'Consumer Electronics' is on top & 'Tourism' is lowest in 'industry of the seller‘.
- Customers top Purpose of the cash loan is 'Repairs‘.



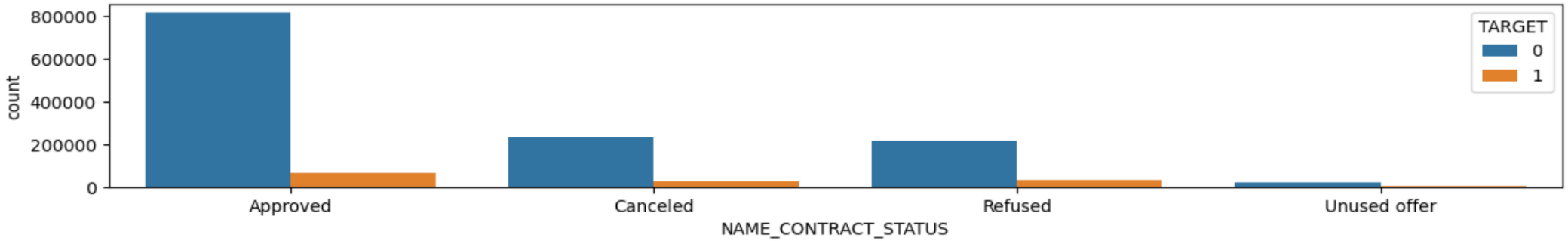


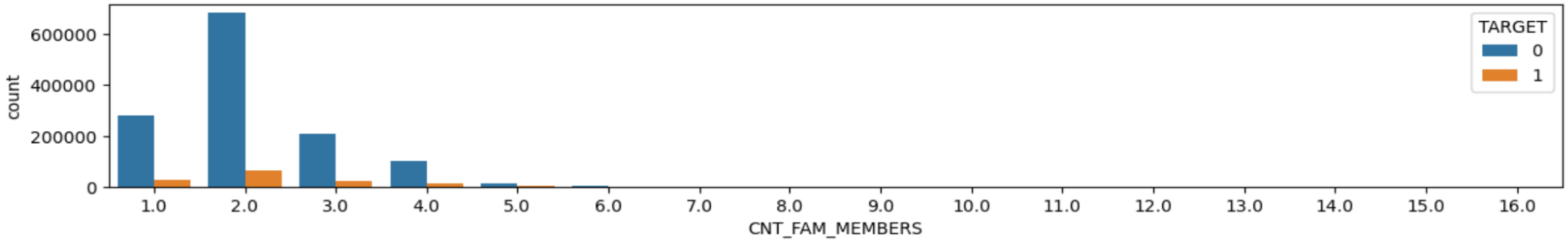
Bivariate Analysis(Merged Dataset):

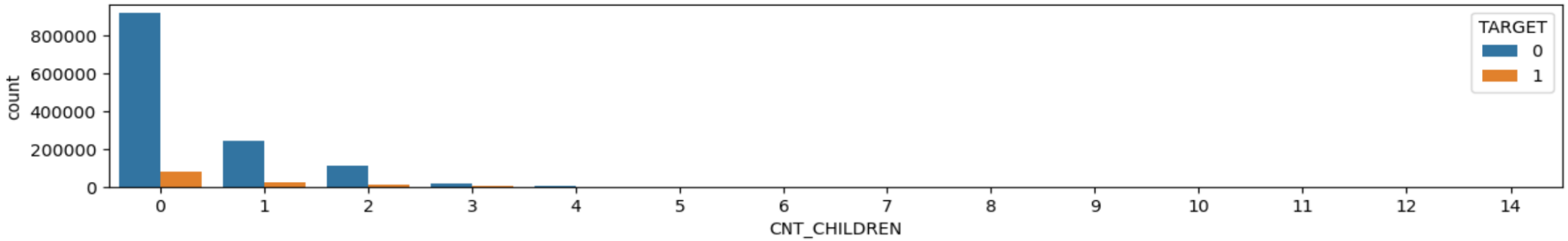
- Customers with/without payment difficulties prefer cash loans over revolving loans
- Female clients are the best repayers of loan compared to Males.
- Customers are mostly 'Working' in both with/without payment difficulties



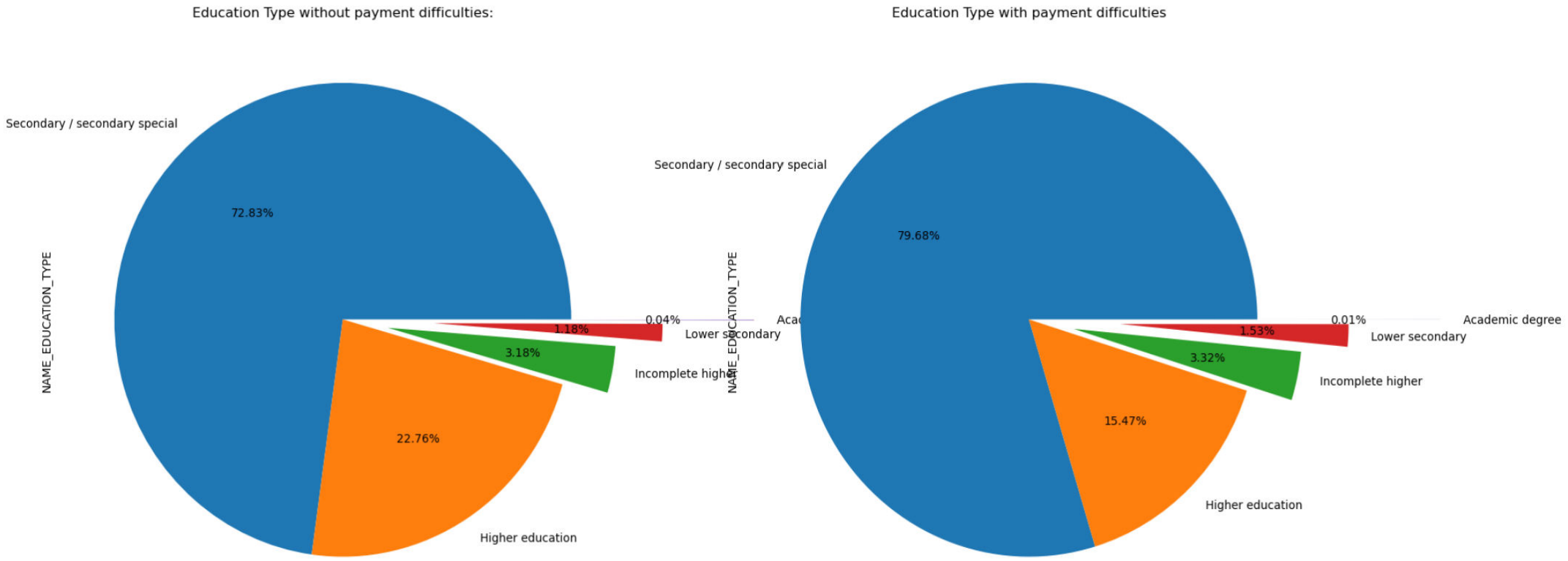
- Customers having approved previous loan have less payment difficulty ,so are better candidates for new loan.
- Customers whose earlier loan was Refused/Cancelled have more chances to default.
- Customers with/without payment difficulties have majorly 2 family members.
- The majority in both cases of with/without payment difficulties, have zero children.
- Customers with no children have less difficulty in making payments.





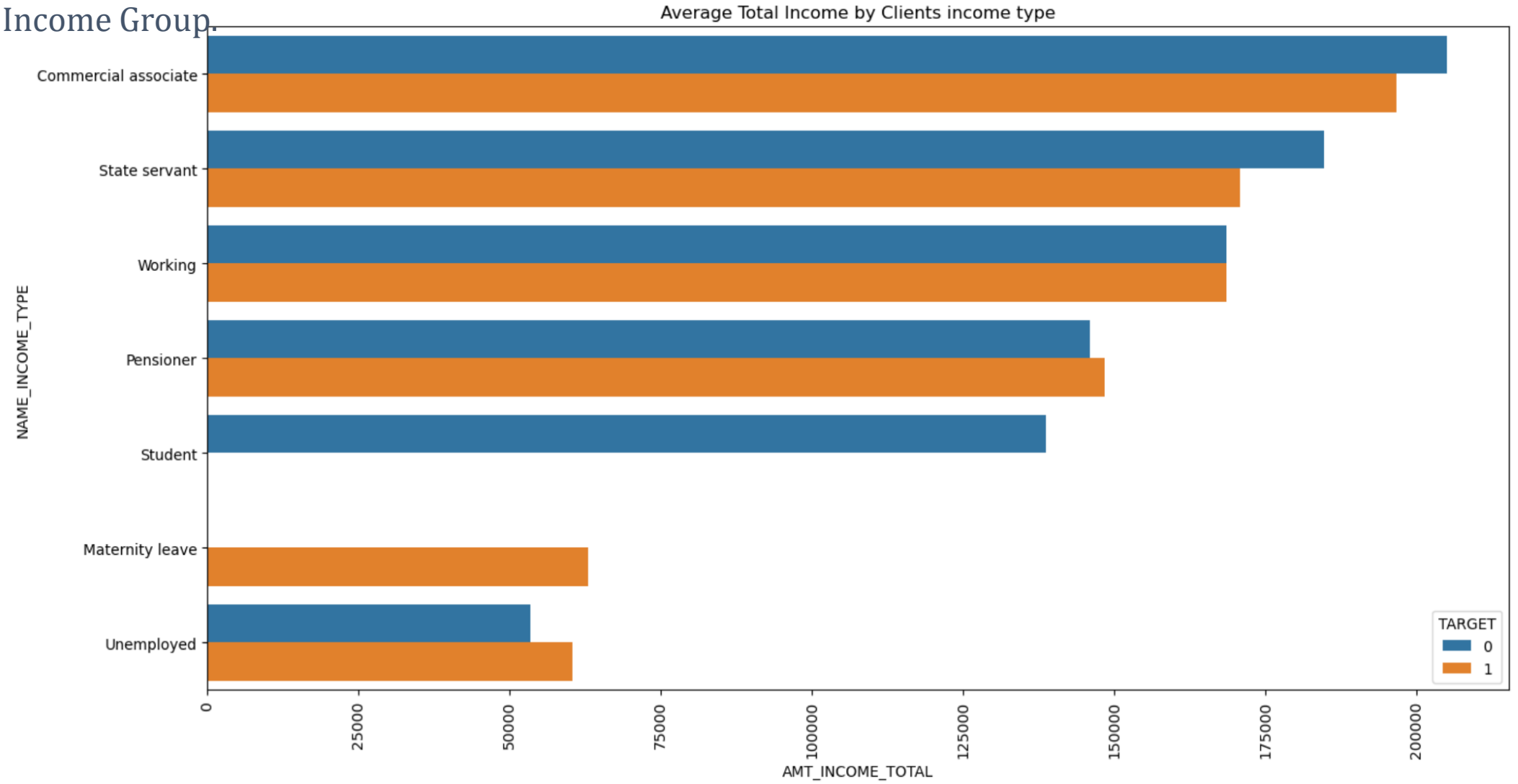


Bivariate Analysis(Merged Dataset) Contd..



- Customers with Secondary Education, are more likely to have payment difficulties more than other Education types.
- Customers are less in both with/without repayment difficulties in lower secondary and academic degree.

- 'Commercial Associates' have majority in cases of With/Without Payment difficulty.
- Students don't have any payment difficulties
- All Customers on Maternity leave are having Payment difficulty.
- Customers having Lower Income (<60K) are facing payment difficulties more compared to Higher Income Group.



Multivariate Analysis(Merged Dataset)

Top & Bottom Co-Related columns(Merged Dataset)

```
SK_ID_CURR                   SK_ID_CURR                   1.000000
AMT_GOODS_PRICE_y            AMT_APPLICATION              0.999871
OBS_60_CNT_SOCIAL_CIRCLE     OBS_30_CNT_SOCIAL_CIRCLE     0.998502
AMT_CREDIT_y                 AMT_GOODS_PRICE_y            0.993201
AMT_GOODS_PRICE_x            AMT_CREDIT_x                 0.986116
AMT_APPLICATION              AMT_CREDIT_y                 0.975683
REGION_RATING_CLIENT_W_CITY  REGION_RATING_CLIENT         0.945596
DAYS_LAST_DUE                DAYS_TERMINATION             0.927738
CNT_FAM_MEMBERS              CNT_CHILDREN                 0.878959
REG_REGION_NOT_WORK_REGION   LIVE_REGION_NOT_WORK_REGION  0.875505
dtype: float64
DAYS_LAST_DUE_1ST_VERSION   CNT_PAYMENT                  -0.377533
DAYS_TERMINATION            DAYS_FIRST_DRAWING           -0.396472
FLAG_DOCUMENT_8             FLAG_DOCUMENT_3              -0.476653
REGION_POPULATION_RELATIVE  REGION_RATING_CLIENT_W_CITY  -0.515888
REGION_RATING_CLIENT        REGION_POPULATION_RELATIVE   -0.516748
FLAG_DOCUMENT_3             FLAG_DOCUMENT_6              -0.526761
FLAG_DOCUMENT_6             FLAG_EMP_PHONE               -0.573793
FLAG_EMP_PHONE              YEARS_BIRTH                  -0.628899
DAYS LAST DUE 1ST VERSION   DAYS FIRST DRAWING           -0.809048
```

- There is high correlation between AMT_GOODS_PRICE--AMT_APPLICATION & AMT_CREDIT--AMT_GOODS_PRICE for previous applications.
- CNT_FAM_MEMBERS & CNT_CHILDREN are also highly correlated implying customer having more family members are having more children as well.
- 'client's permanent address does not match work address' is related well to 'client's contact address does not match work address'
- Relative to application date of current application there is Low relationship between when was the first due and first disbursement of the previous application

Insights

- All Female customers on maternity leave have payment difficulties and are highly doubtful to default ,therefore should NOT be targeted.
- Commercial Associates, State Servants & Students have less payment difficulties and so should be targeted for loans.
- Females should be given preference as they have good repayment percentage over Males.
- Customers having Revolving loans have very negligible payment difficulties, making them GOOD target.
- Customers with no children have less difficulty in making payments and should be targeted MORE by the bank.
- Customers having approved previous loan have less payment difficulty ,so are better candidates for new loan.
- Customers whose earlier loan was Refused/Cancelled have more chances to default.
- Majority of customers of age group <40 are having payment difficulties so Bank should target older people i.e. >40 as they are good in repayments.
- Customers with Secondary Education, are likely to have payment difficulties more than other Education types.
- Customers having Lower Income(<60K) are facing payment difficulties more compared to Higher Income Group.

Thank You