upGrad Assignment – Credit EDA

NAME:- SOURAV DUTTA

BATCH NO:- C46

EMAIL:- SOURAVDUTTASV1999@GMAIL.COM

Problem Statements -

Introduction

This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

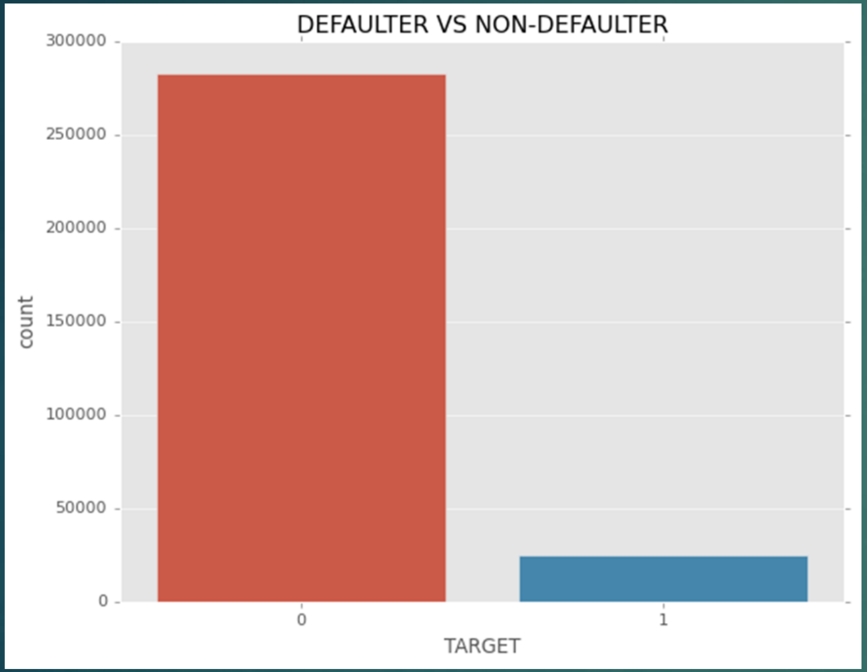
Business Understanding & it's Objectives

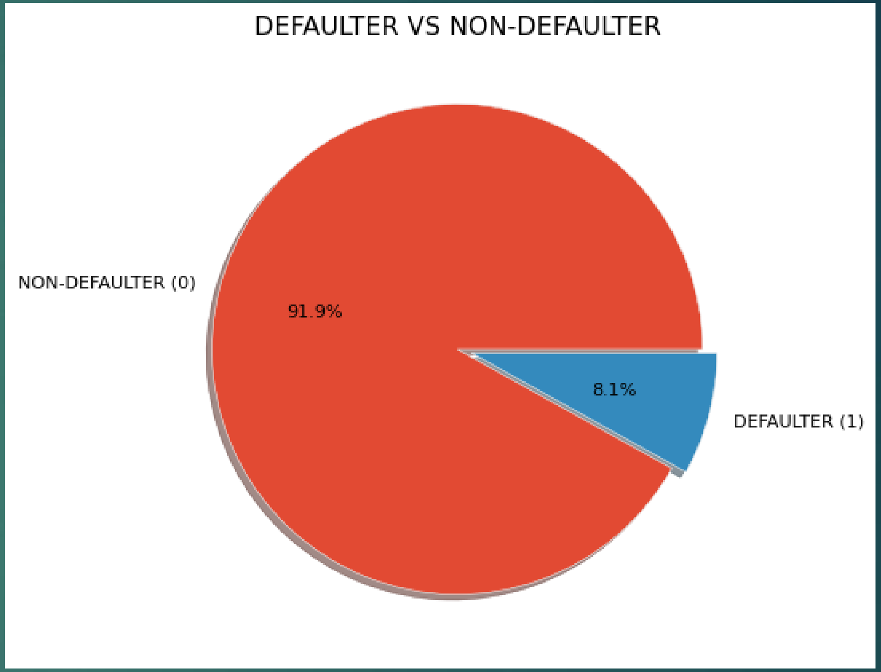
- Understanding the business & it's objectives is very crucial for further analysis
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Steps that will be followed in EDA process -

- Importing necassary liabraries
- Understanding the Data & Loading the Data
- Basic Sanity Checks
- Checking missing values / Handling missing values
- Outliers Detection / Handling Outliers
- Imbalance of Data
-
- Univariate Analysis – Categorical & Numerical
- Segmented Univariate Analysis
- Bivariate Analysis - Categorical & Numerical
- Top 10 Correlation

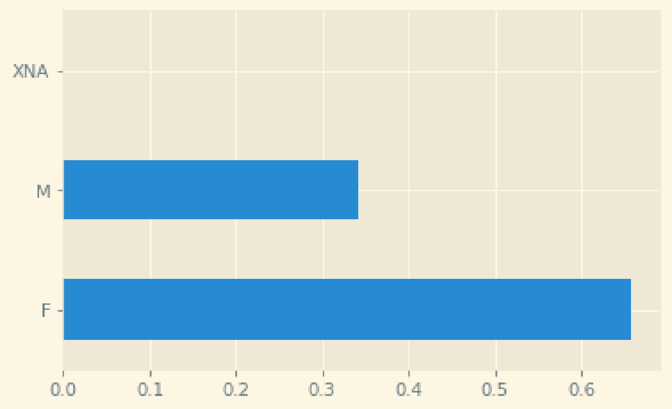
Imbalance of Data in Application Data

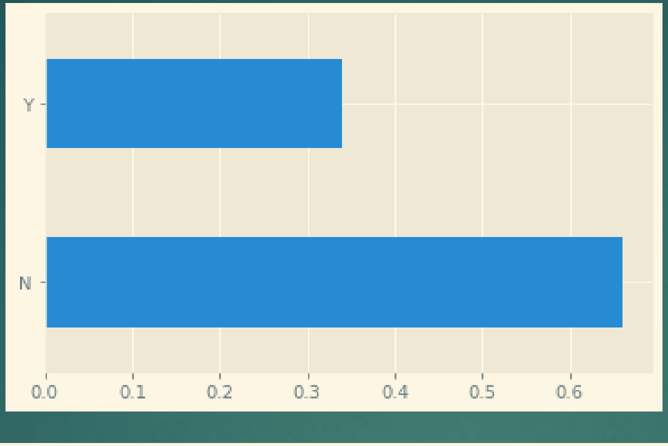


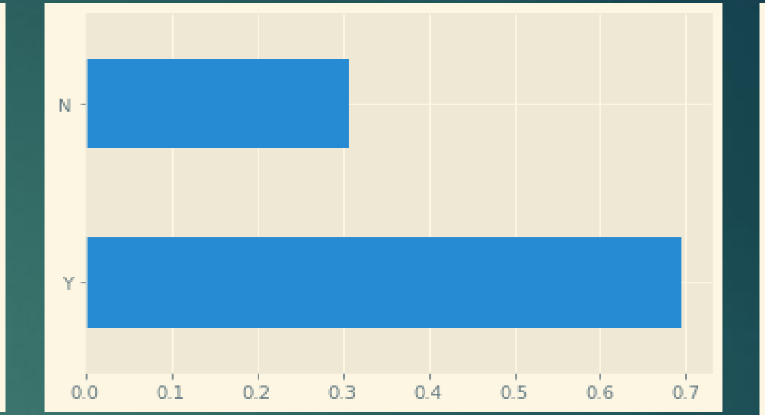


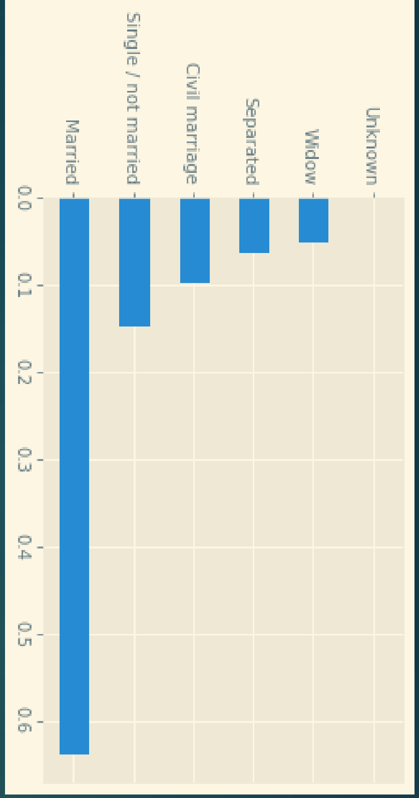
So here we can clearly see there is imbalance in Target variable between DEFAULTER & NON-DEFAULTER. Almost 92% peoples are NON-DEFAULTER and about 8% peoples are DEFAULTER.

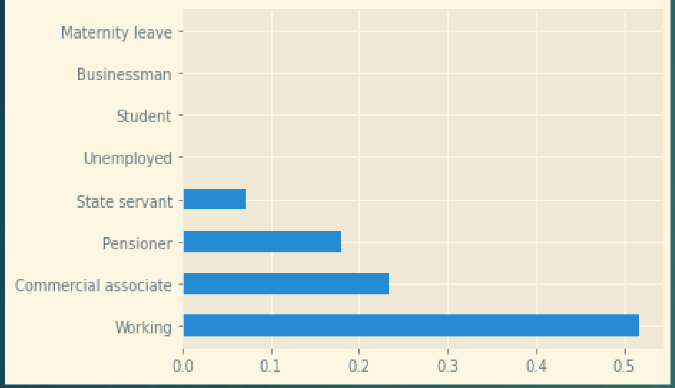
Univariate Analysis (Application Data)

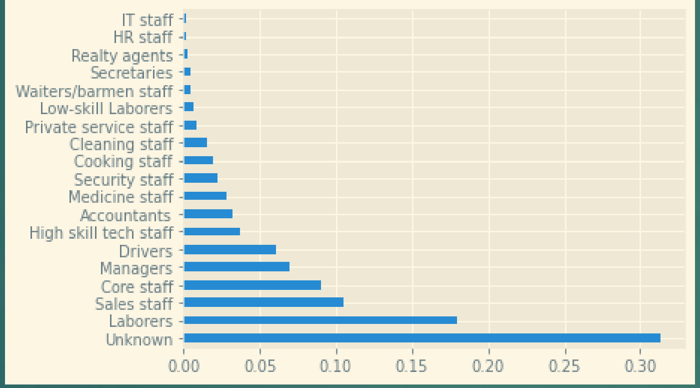


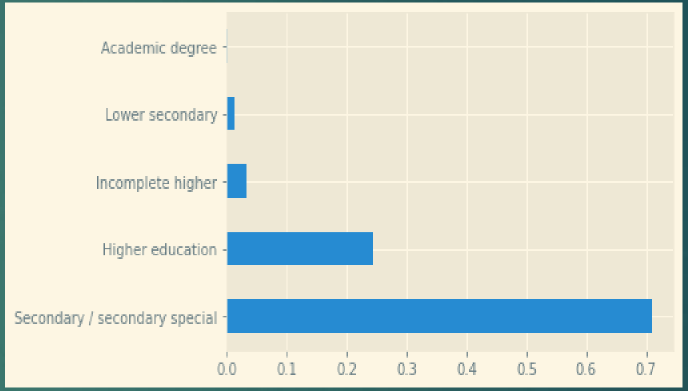


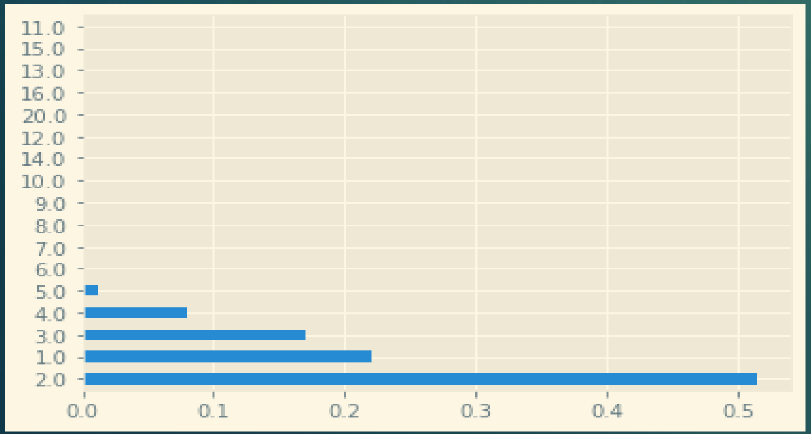


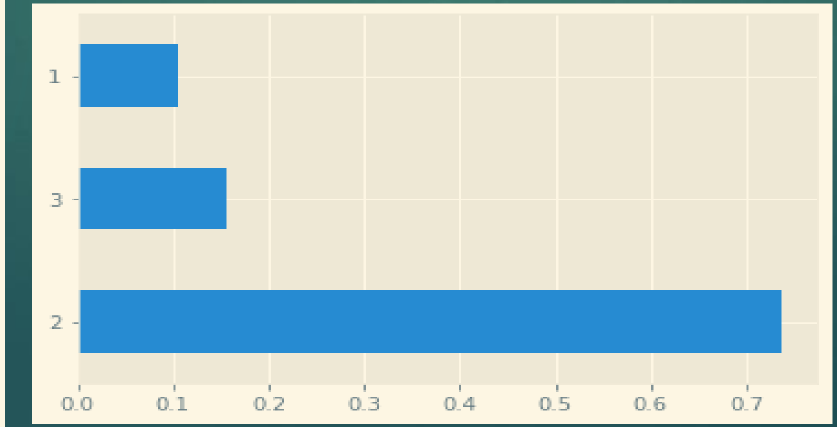


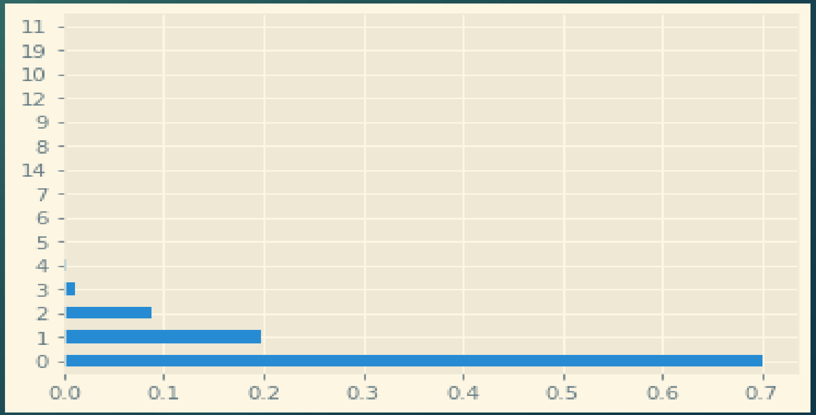








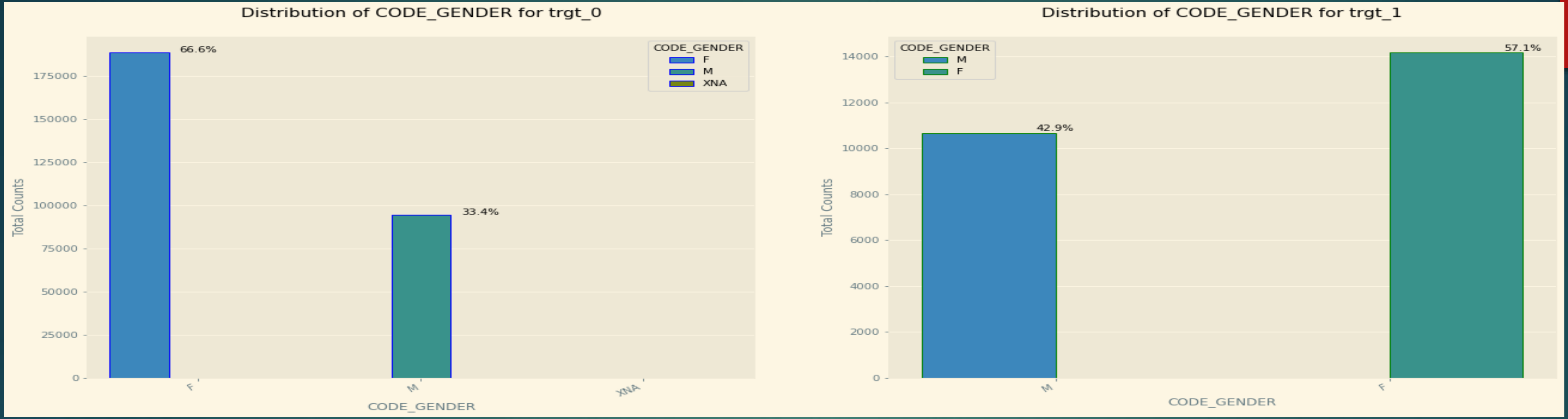




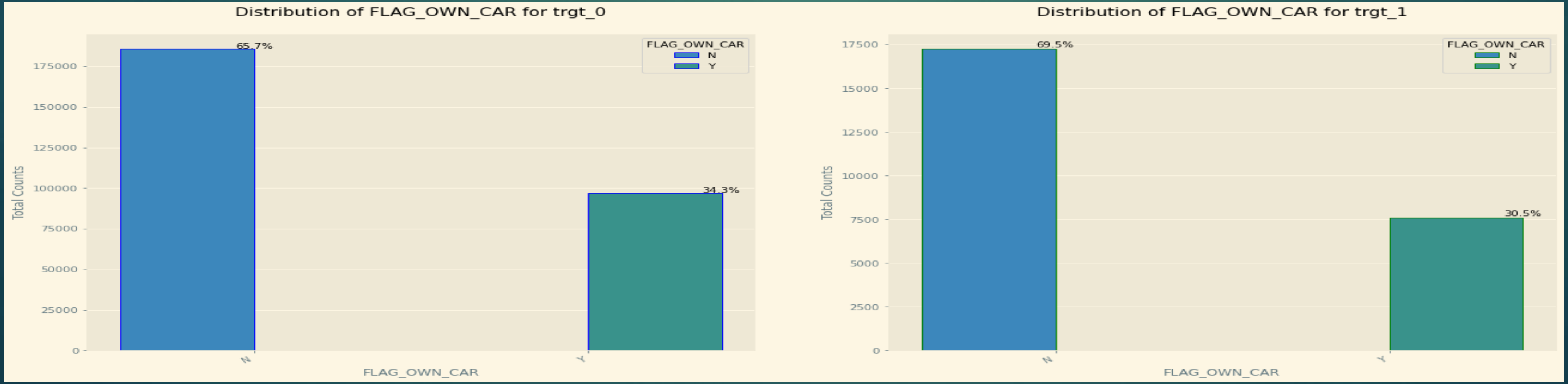
Observation from the above Univariate Analysis

- Count of female is high in applying loan
- Most number of people don't have car who applied for loan
- Most number of peoples have own House/Flat who are applying for loan
- We can infer that Married people's loan application is much higher than others
- We can clearly see Working peoples applied for loan higher than other categories
- Most numbers of loan application came from Unknown & Laborers categories
- Secondary & Higher education type peoples applying more loan
- Most number of peoples have less family members
- Most number of peoples have no children who are applying for loan
- Most peoples belongs from region with 2 rating

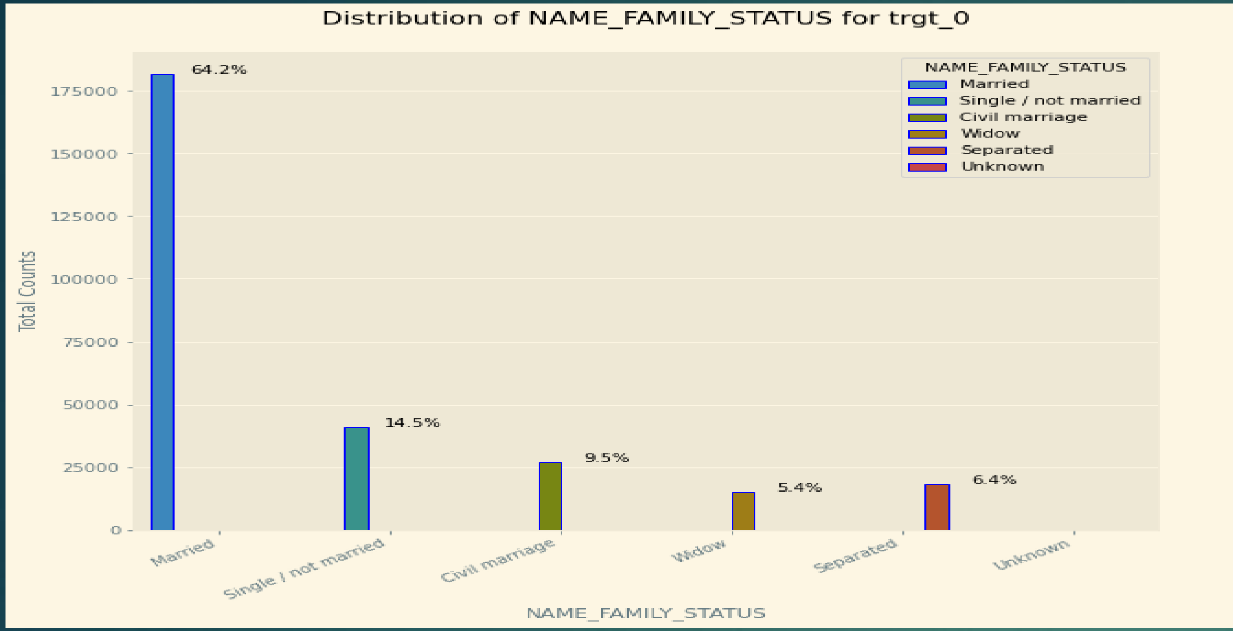
Segmented Univariate Analysis

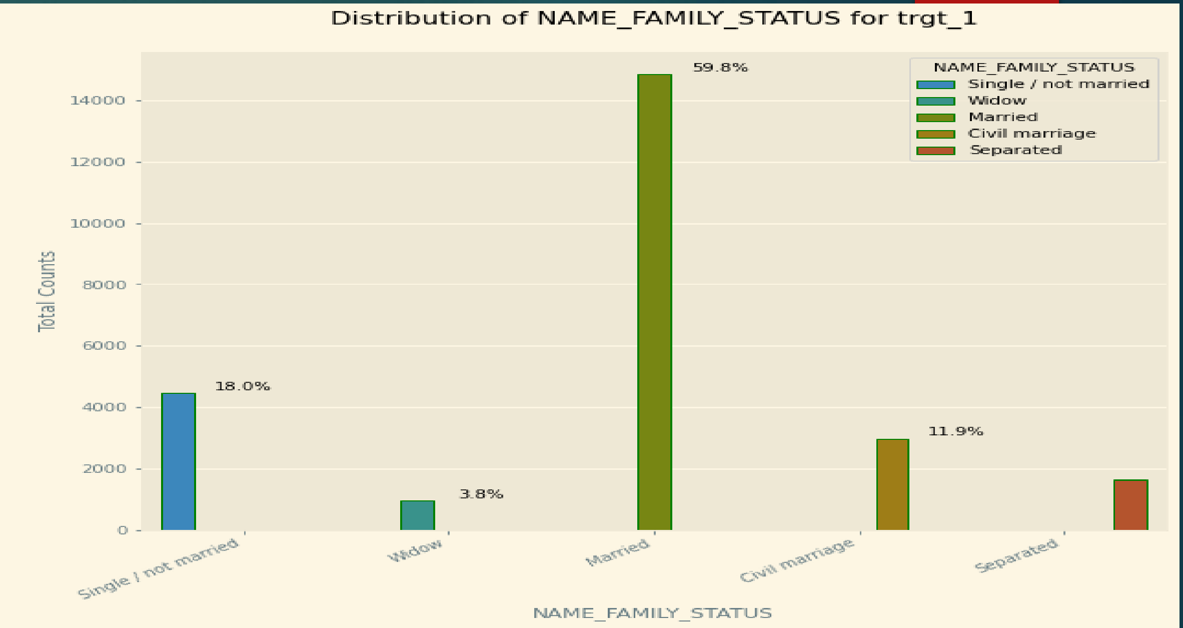


From this CODE_GENDER plot we can infer that females loan application is more than males, But the percentage of defaulter is higher in case of males.

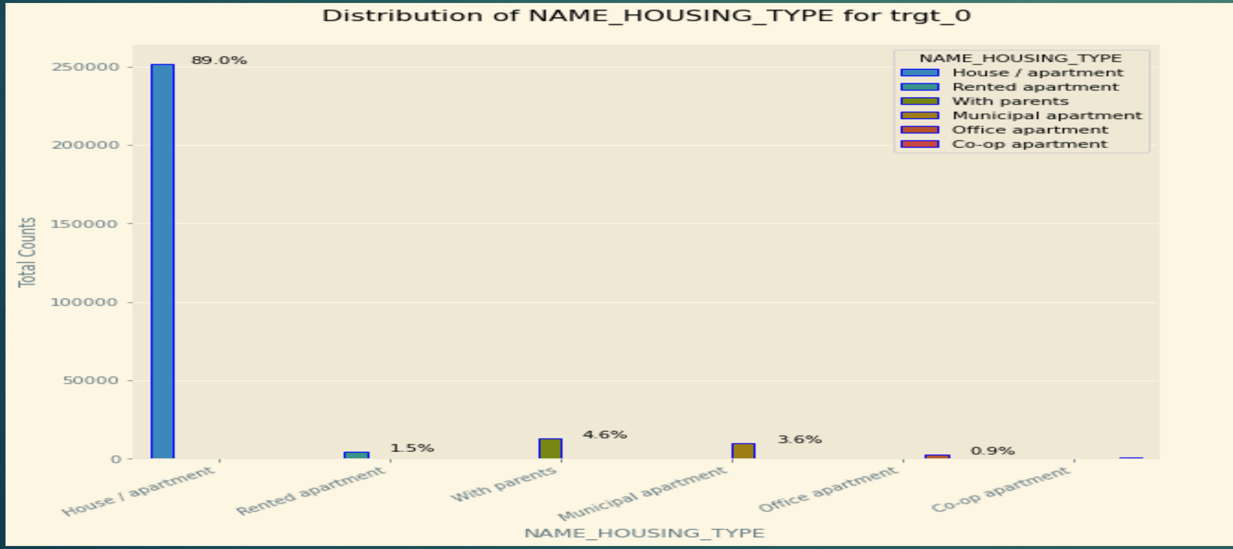


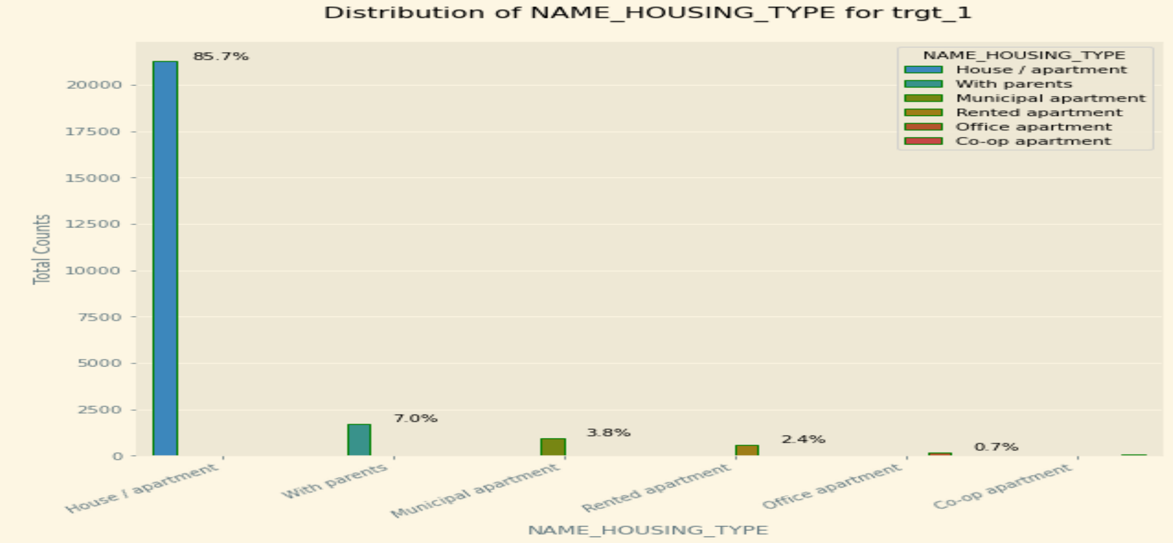
People who doesn’t own car chances of defaulter is higher



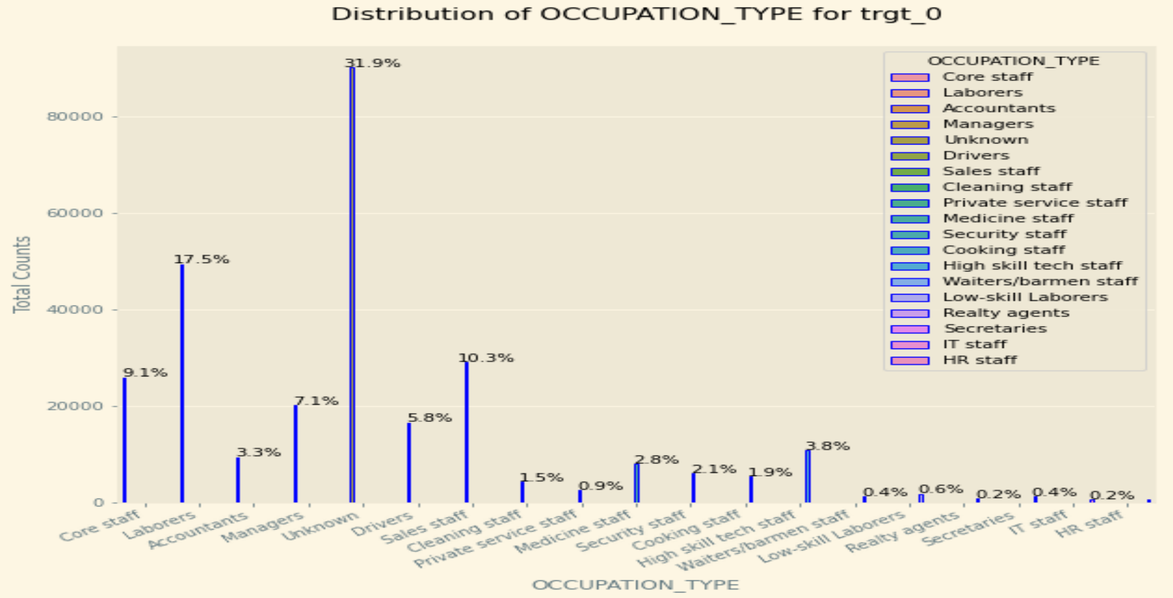


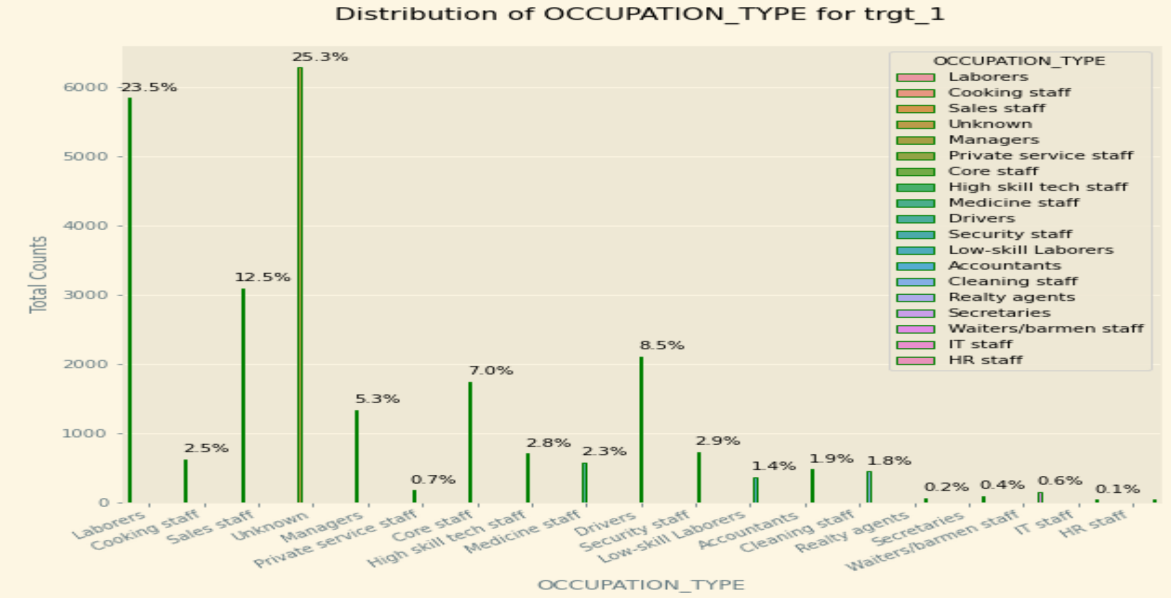
From this NAME_FAMILY_STATUS column we can infer that Unknown category not in defaulter, but Married people's loan application & defaulter count is higher than rest of the category.



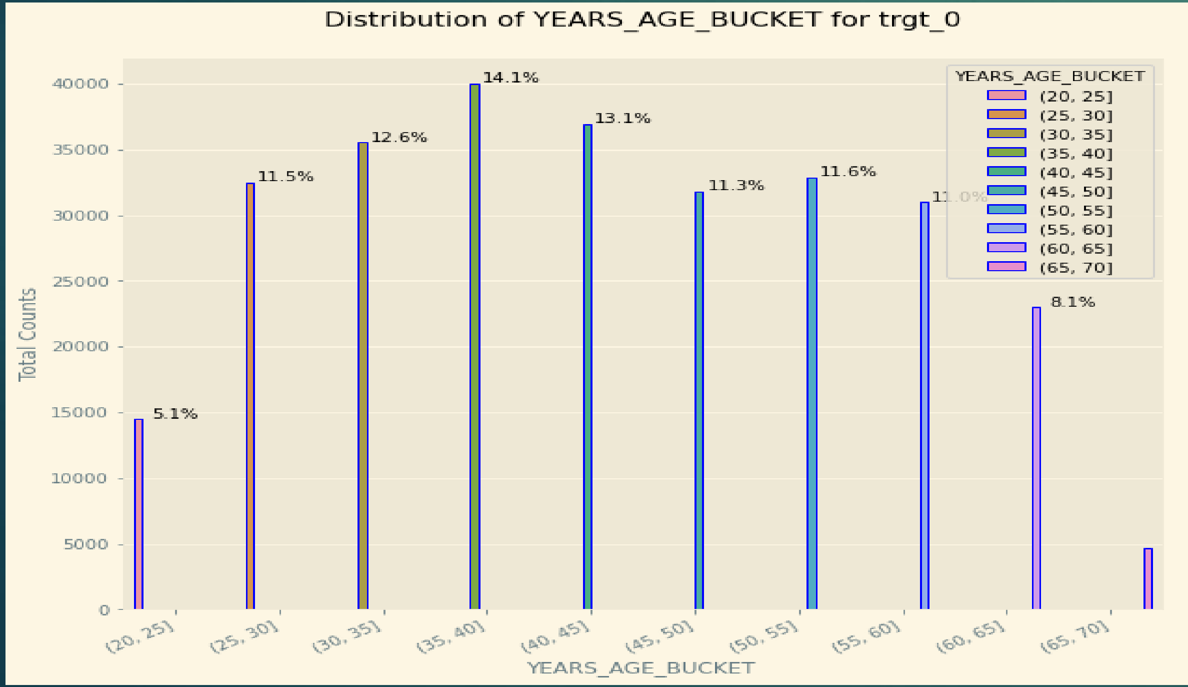


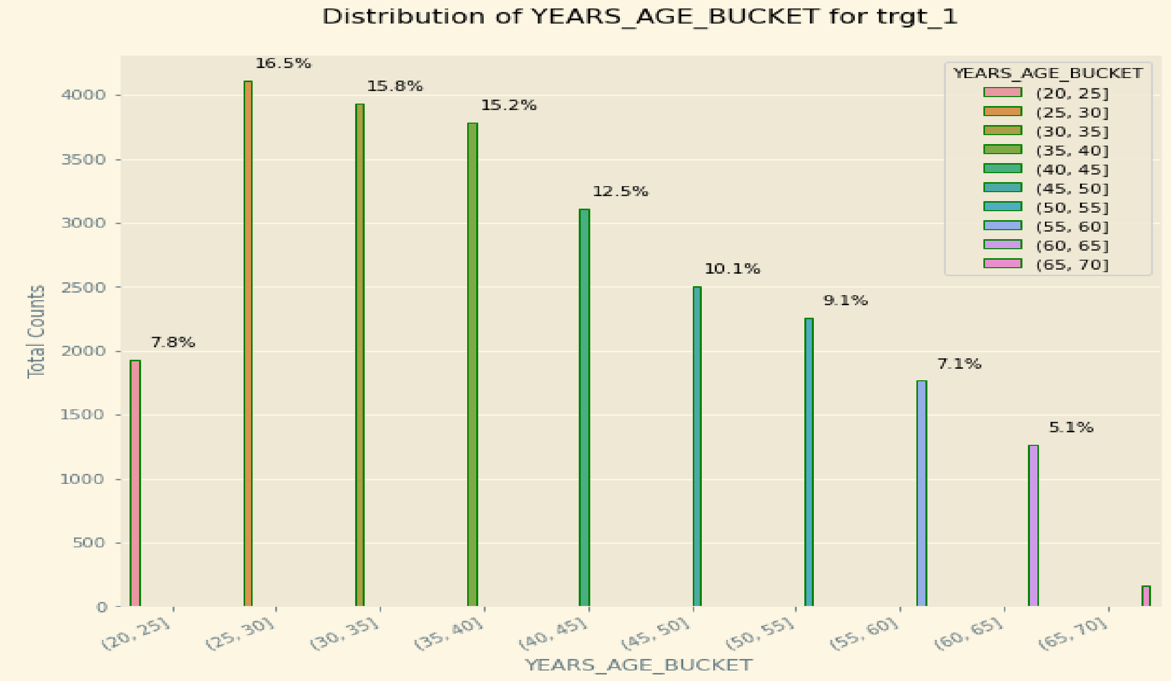
As we can see people with House/Apartment count of loan application is very higher than other & people who are living with parents count of defaulter is slightly higher.


Distribution of OCCUPATION_TYPE for trgt_0

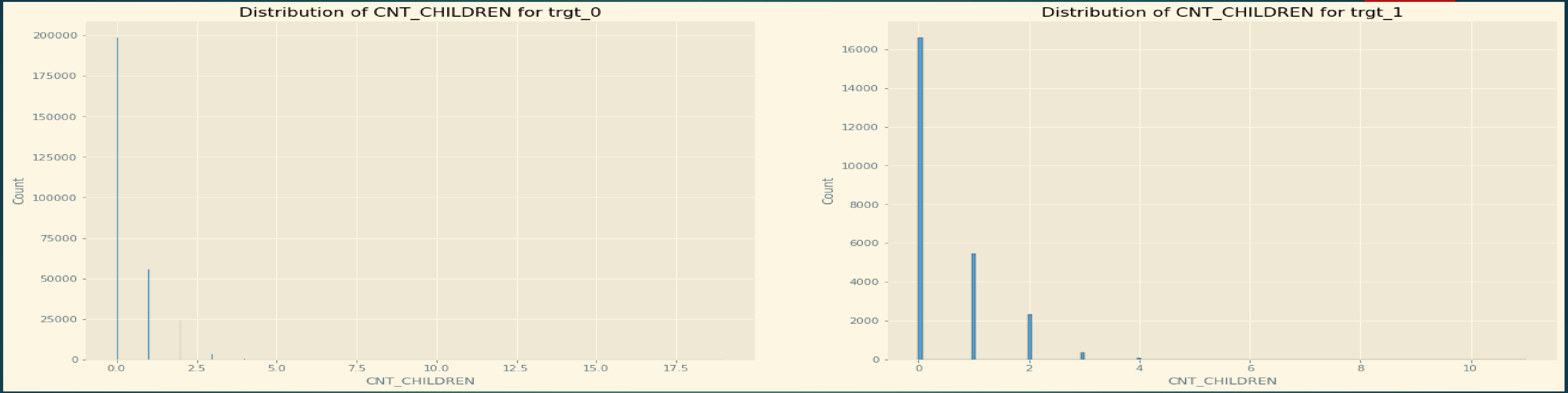

Distribution of OCCUPATION_TYPE for trgt_1

From this OCCUPATION_TYPE column we can see Laborers,Drivers,Sales staff peoples chances of getting defaulter is high

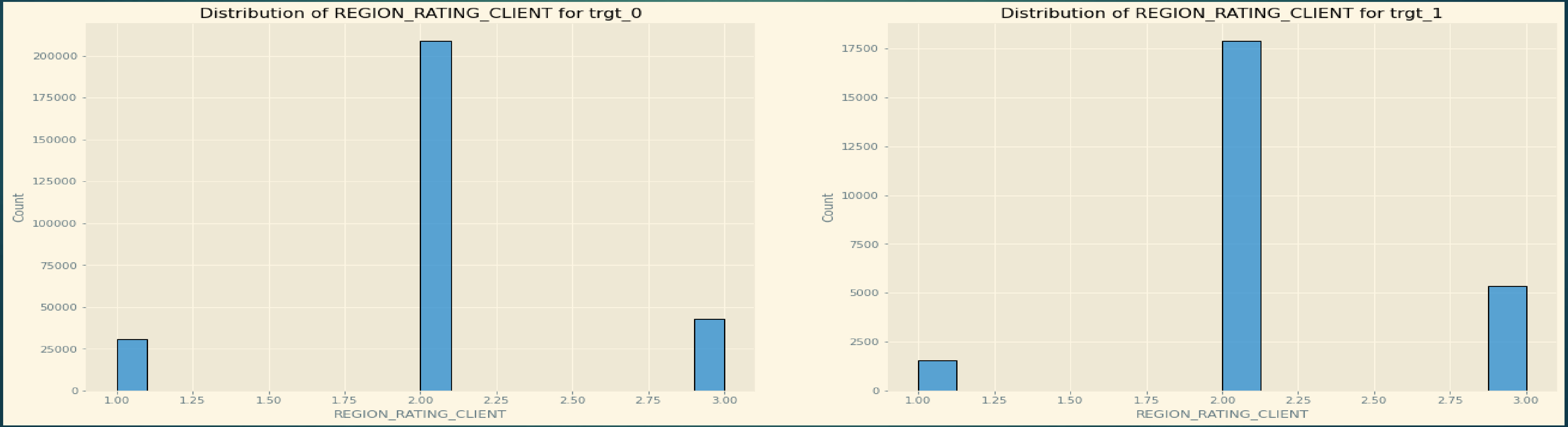

Distribution of YEARS_AGE_BUCKET for trgt_0


Distribution of YEARS_AGE_BUCKET for trgt_1

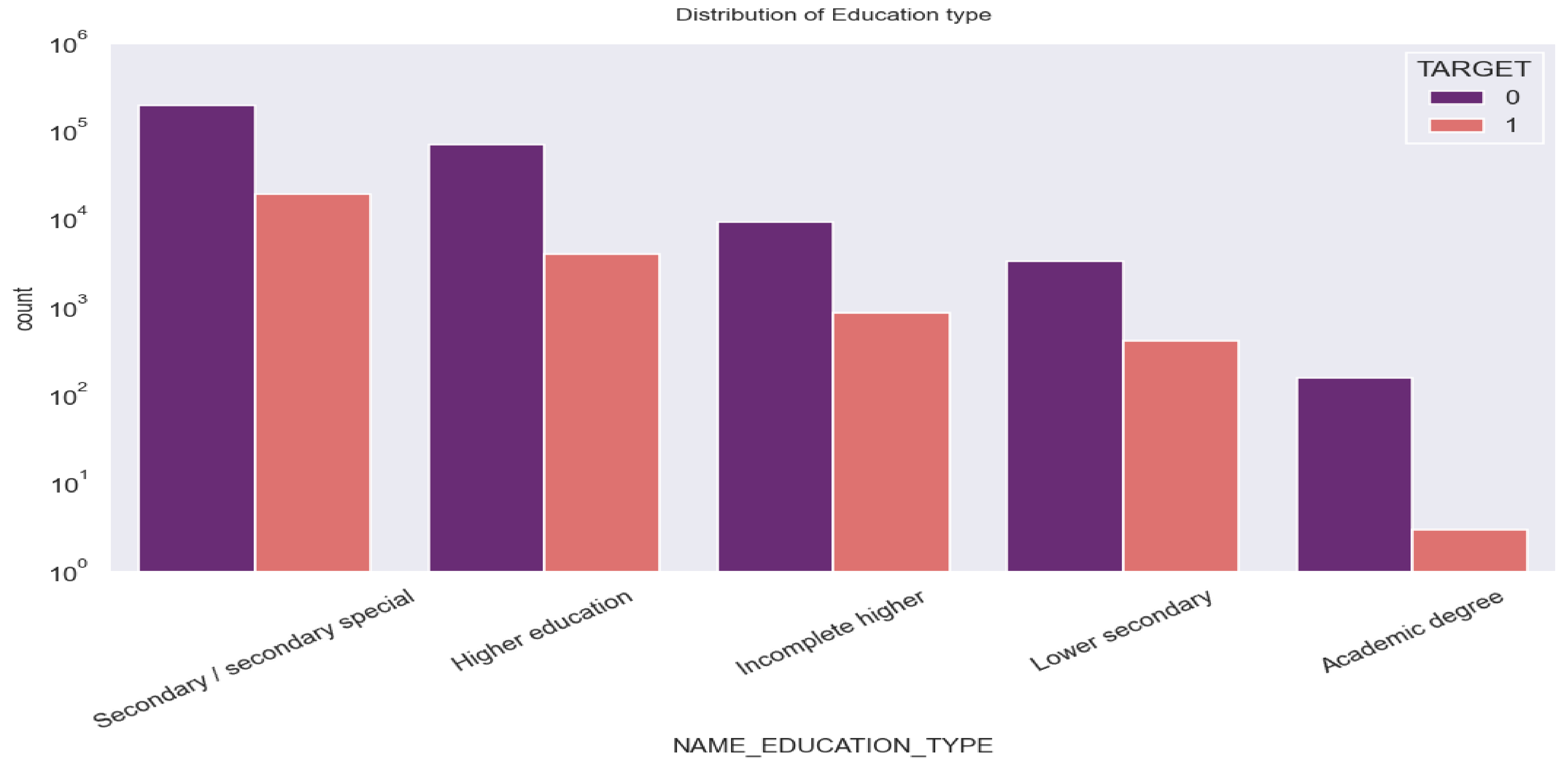
We can infer that the age group between [20,25],[25,30],[30,35],[40,45] has more risk of getting defaulter.



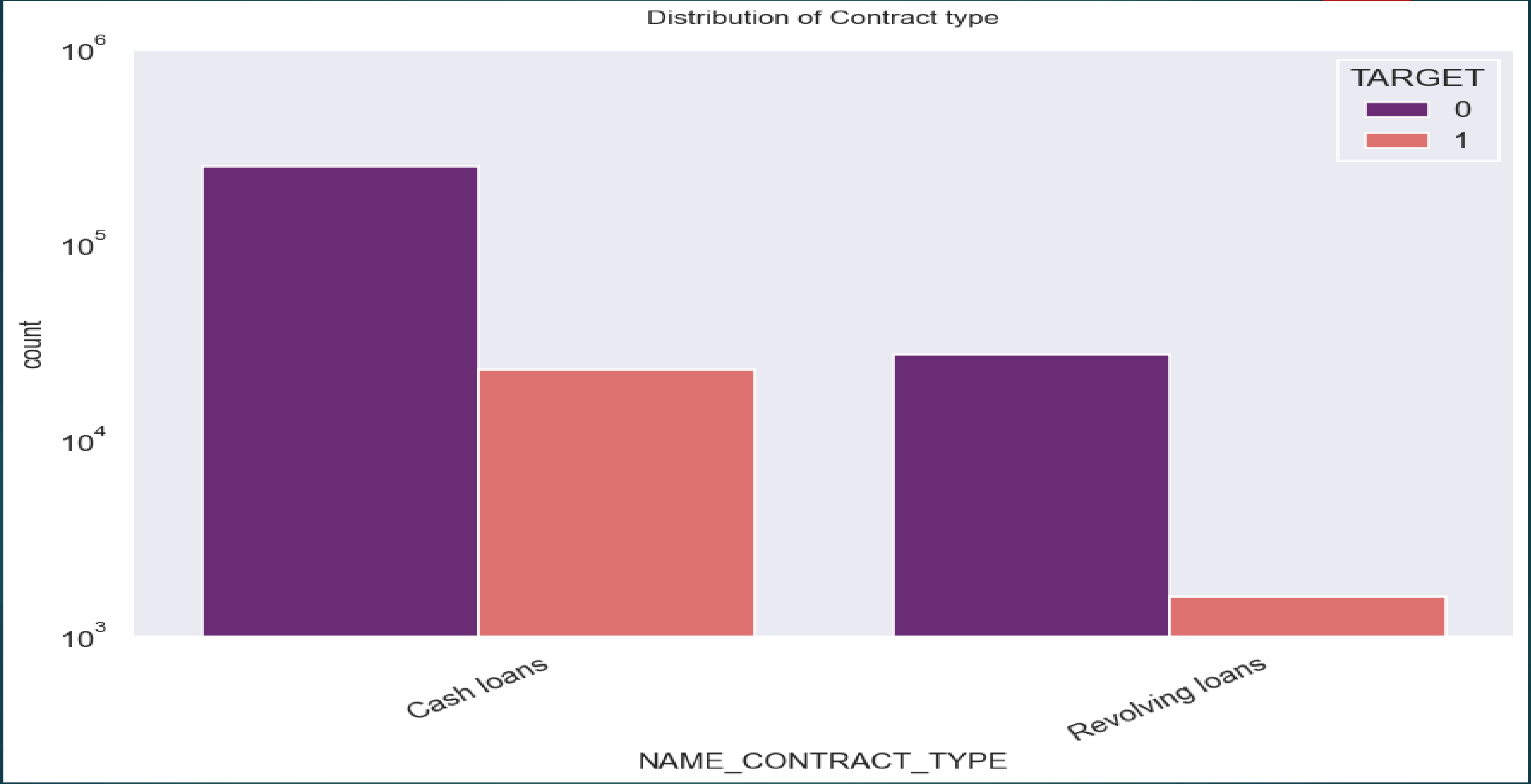
Applicats with 2 or more numbers of children having high chances of defaulter



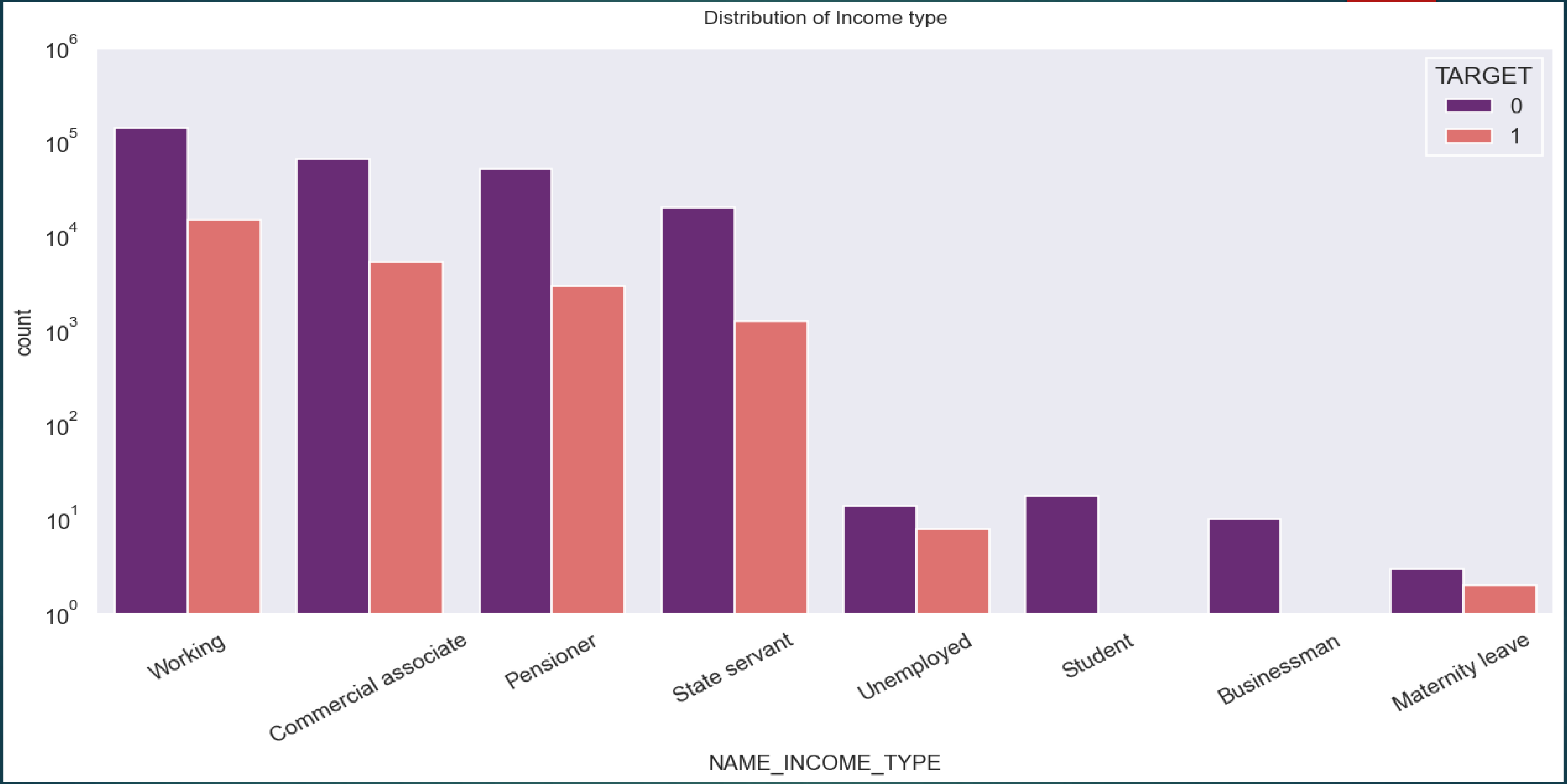
Applicants with Region Rating 3 having high chances of default



In Academic degree education type risk of Defaulter is lower

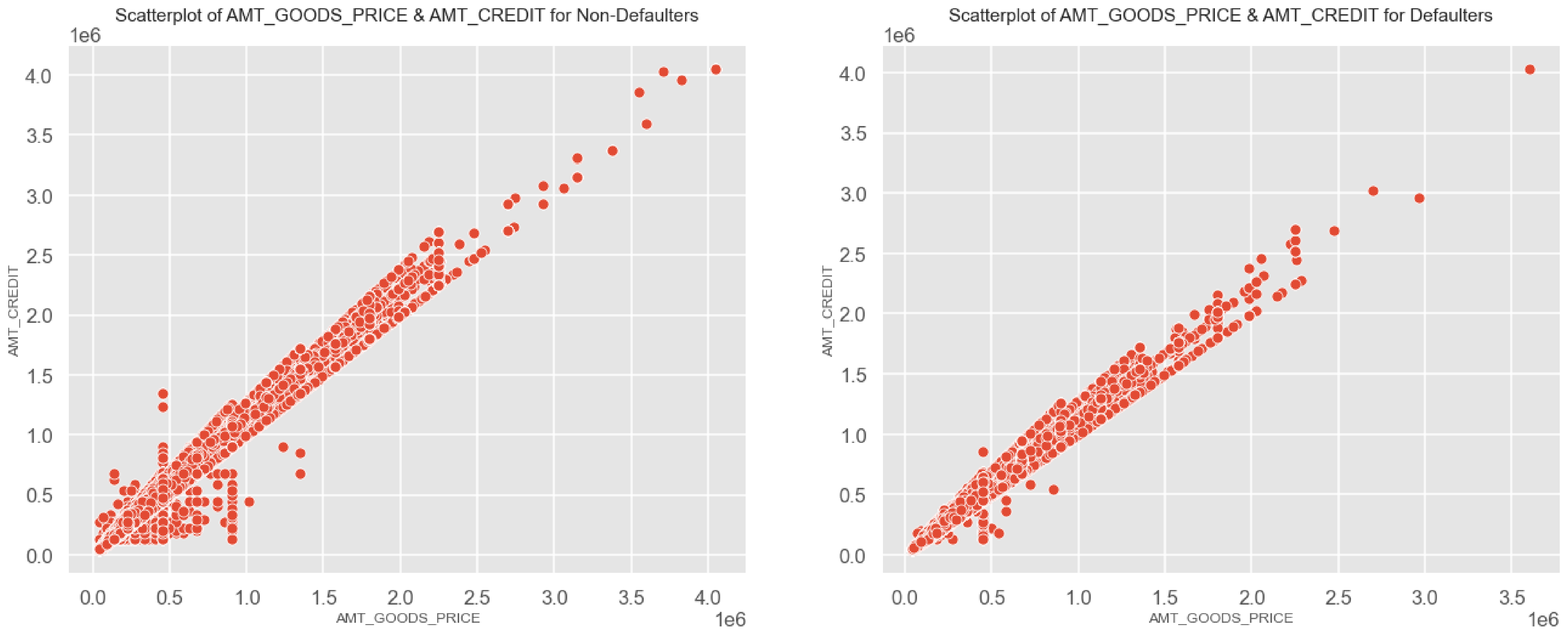


In the cash loan type chances of Defaulter is high, we can focus on Revolving loans for more approval.



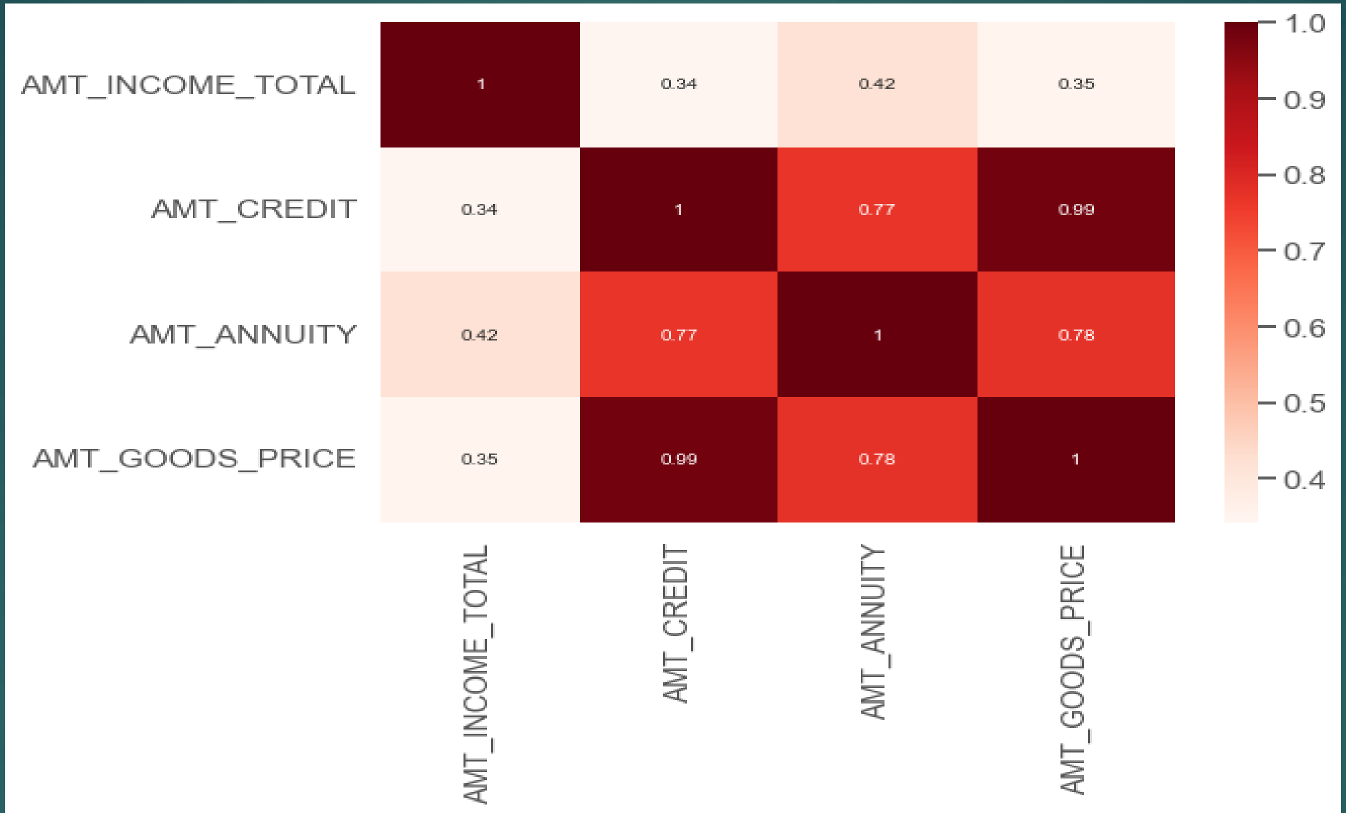
From this we can see Students & Businessman category has no default value. So we should focus on approving loan of these categories cause very low chances of default.

Bivariate Analysis



From here we can clearly judge that there is positive relation between AMT_CREDIT & AMT_GOODS_PRICE

Correlation using Heatmap



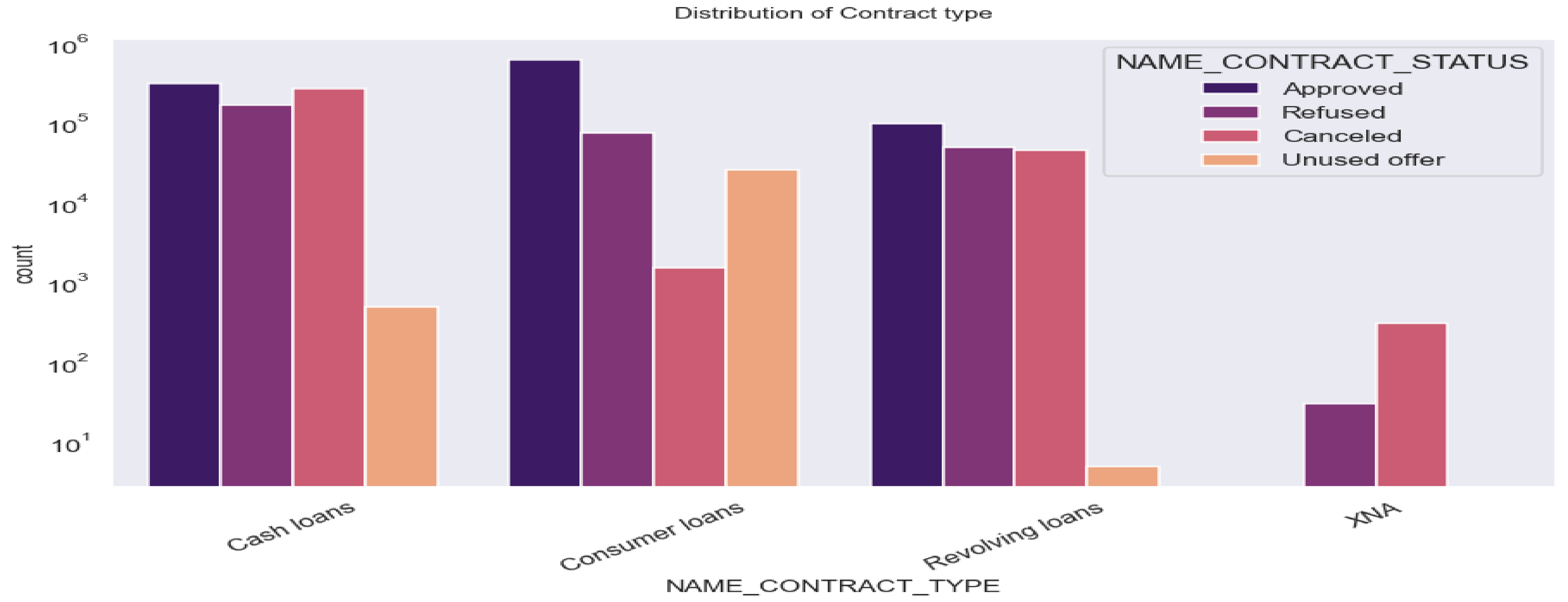
	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
AMT_INCOME_TOTAL	1	0.34	0.42	0.35
AMT_CREDIT	0.34	1	0.77	0.99
AMT_ANNUITY	0.42	0.77	1	0.78
AMT_GOODS_PRICE	0.35	0.99	0.78	1

- ✓ From this heatmap we can infer that apart from the diagonal between AMT_CREDIT & AMT_GOODS_PRICE there is huge positive correlation
- ✓ Between AMT_CREDIT & AMT_ANNUITY slightly positive correlation
- ✓ Between AMT_ANNUITY & AMT_GOODS_PRICE slightly positive correlation

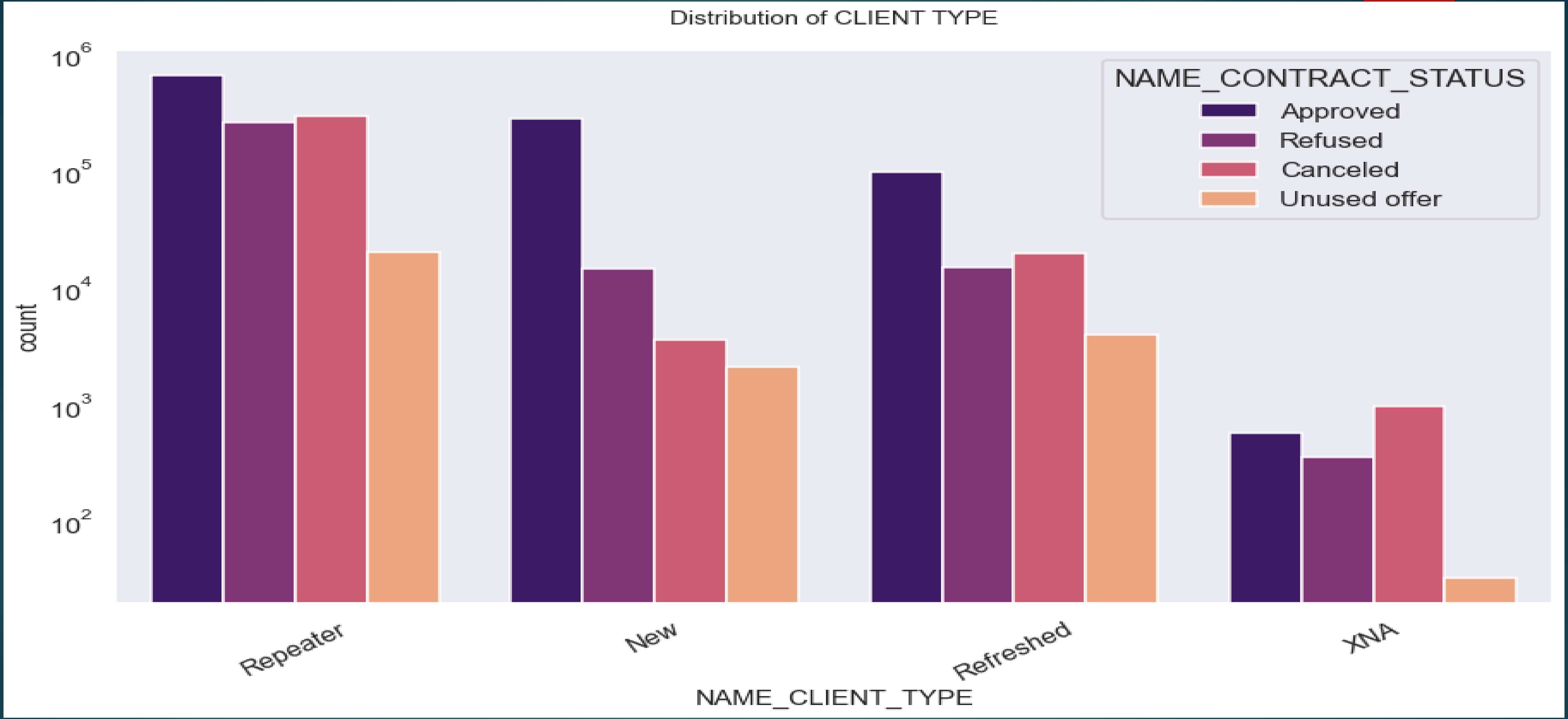
Top 10 Correlation

		0
CNT_CHILDREN	CNT_FAM_MEMBERS	0.879
AMT_CREDIT	AMT_ANNUITY	0.770
DAYS_BIRTH	DAYS_EMPLOYED	0.616
REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT	0.533
DAYS_BIRTH	DAYS_REGISTRATION	0.332
CNT_CHILDREN	DAYS_BIRTH	0.331
DAYS_BIRTH	CNT_FAM_MEMBERS	0.279
	DAYS_ID_PUBLISH	0.273
DAYS_EMPLOYED	DAYS_ID_PUBLISH	0.272
CNT_CHILDREN	DAYS_EMPLOYED	0.240

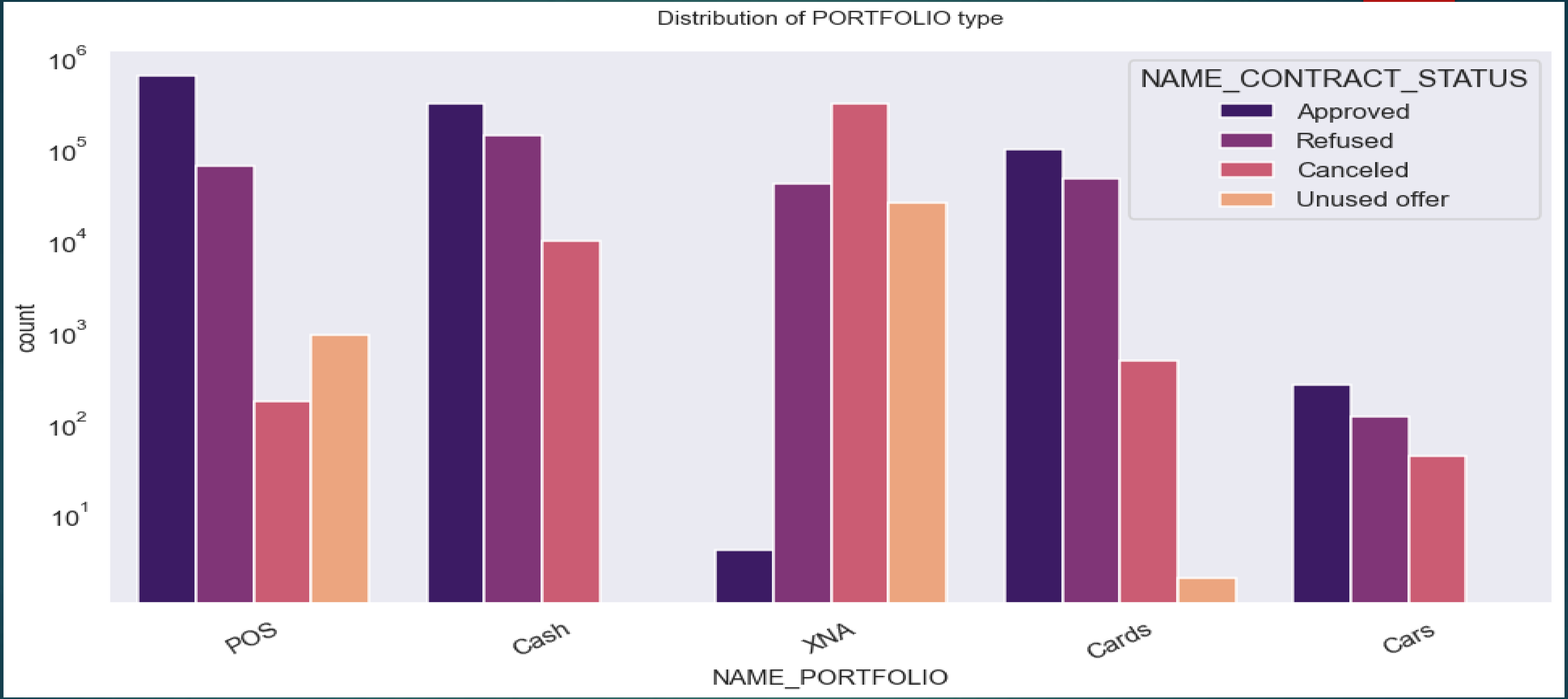
Univariate Analysis on Previous Application Data



Consumer loan approved more followed by cash loan & revolving loan

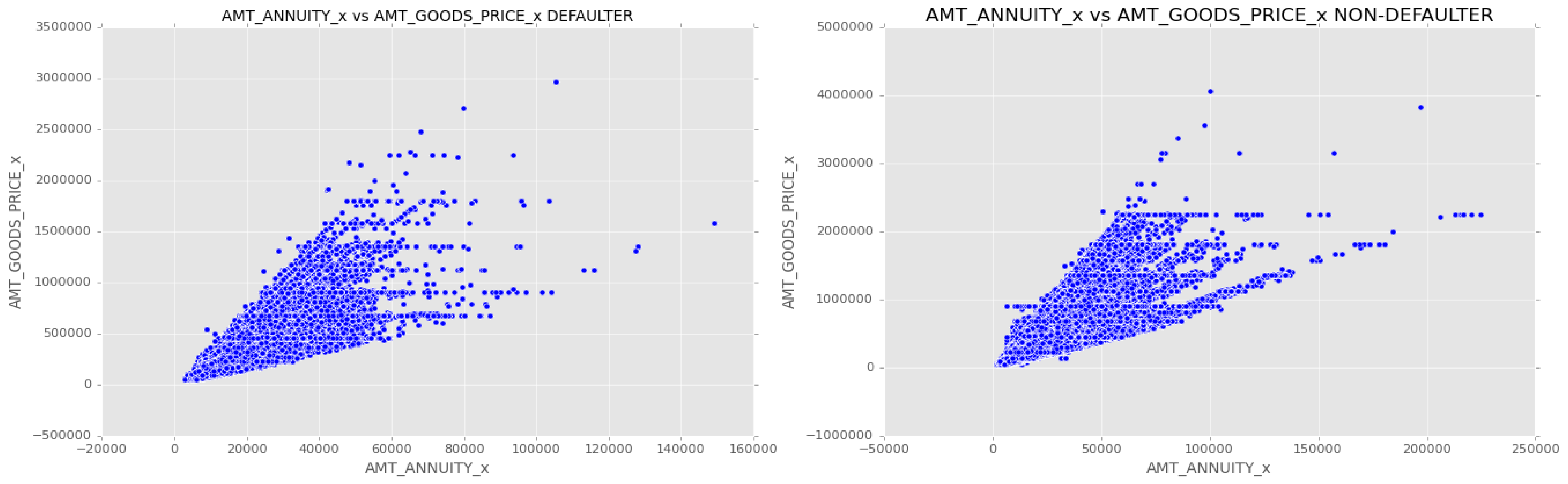


We can clearly judge that Repeater category has the highest number of approval & New category with low chances of Refused & cancelled



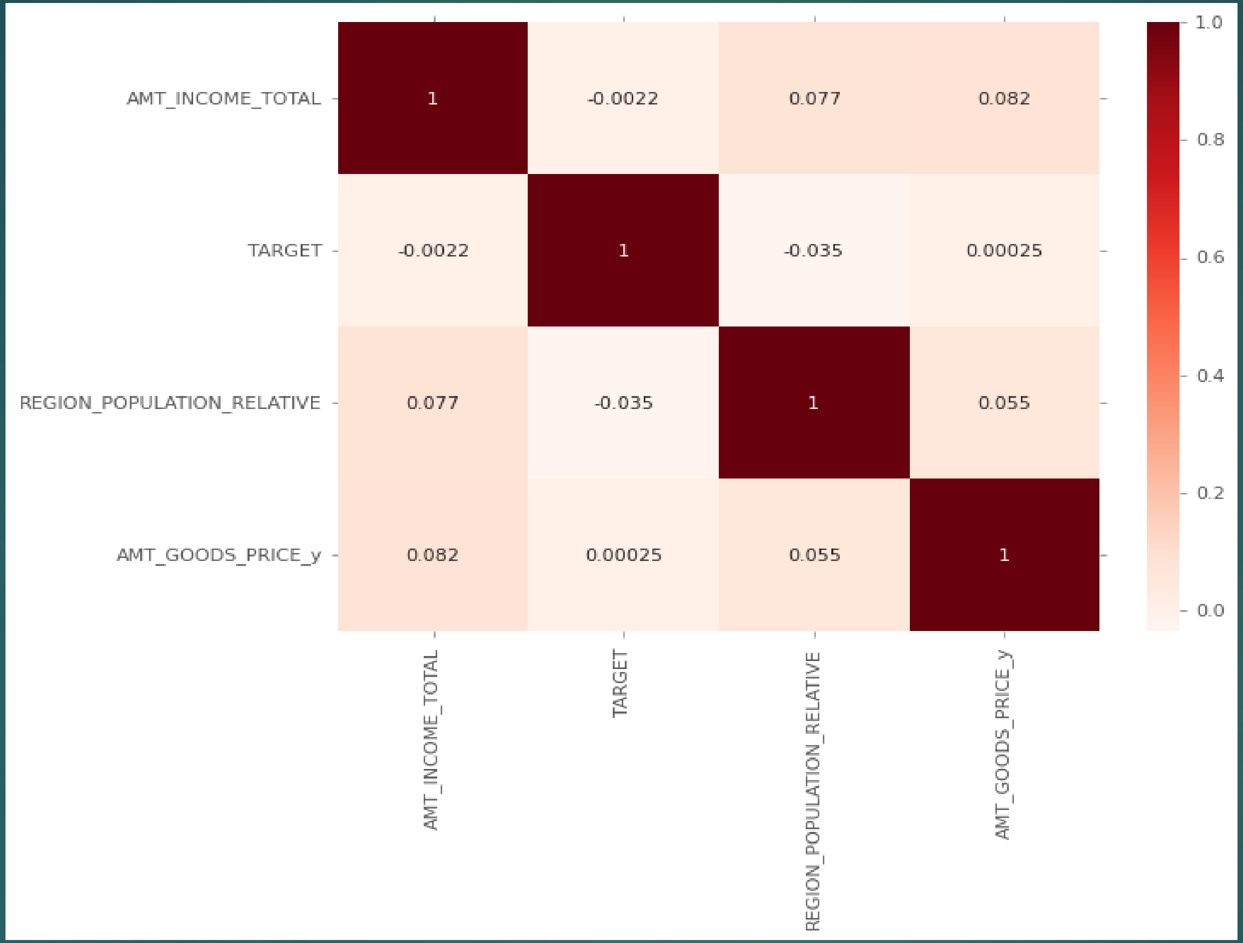
POS has the highest number of approval with lowest canceled chance

Merge Data Set – Bivariate Analysis



Between AMT_ANNUITY_x & AMT_GOODS_PRICE_x highly Positive correlation

Correlation using Heatmap



We can see no such strong correlation, but there is slightly negative correlation between [REGION_POPULATION_RELATIVE & TARGET] & [AMT_INCOME_TOTAL & TARGET].