CREDIT EDA ASSIGNMENT

BY - PALLAVI RAJ

Problem Statement

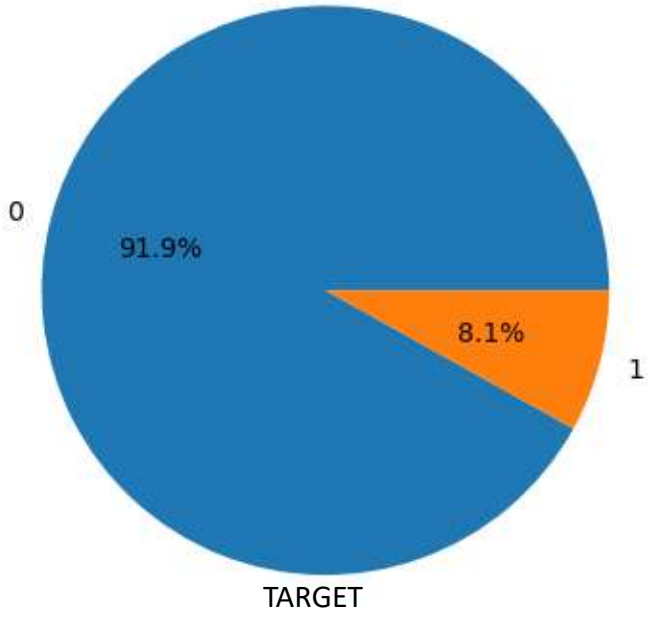
As an employee of a consumer finance company specializing in lending various types of loans to urban customers, you've observed a challenge in the lending process. The difficulty arises when potential borrowers have insufficient or non-existent credit histories, making it challenging for loan-providing companies to assess their creditworthiness accurately. Some consumers exploit this situation by intentionally becoming defaulters.

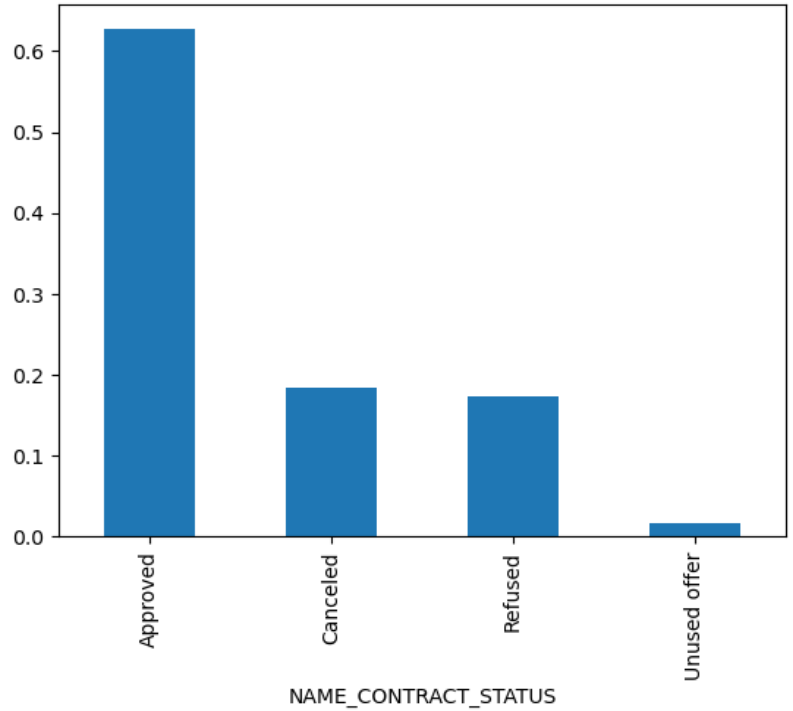
Our task is to address this issue by performing Exploratory Data Analysis (EDA) to analyze patterns present in the data. The goal is to develop insights that can help in identifying applicants who are capable of repaying the loan, thereby preventing the rejection of creditworthy individuals.

Methodology to perform EDA

1. Understanding the Structure of data
2. Data Quality check and handling missing values
3. Creating bins or categories to organize and analyze that data.
4. Identifying the data imbalance in the data and performing univariate , segmented univariate, bivariate analysis, etc
5. Merging the application data with previous application dataset
6. Performing Univariate, bivariate analysis

Insights





Data Imbalance

1.Default Status:

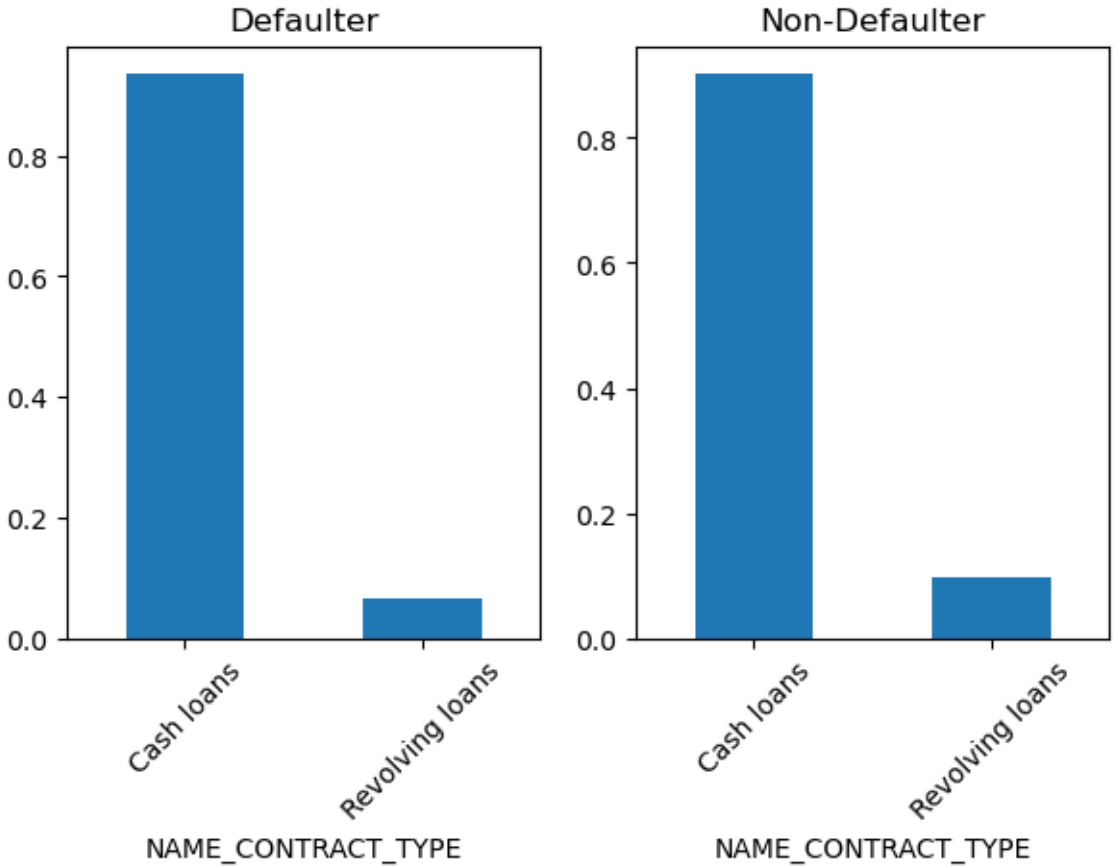
The target variable, representing defaulters and non-defaulters, is highly imbalanced, with only 8.1% of the data corresponding to individuals who default on loans.

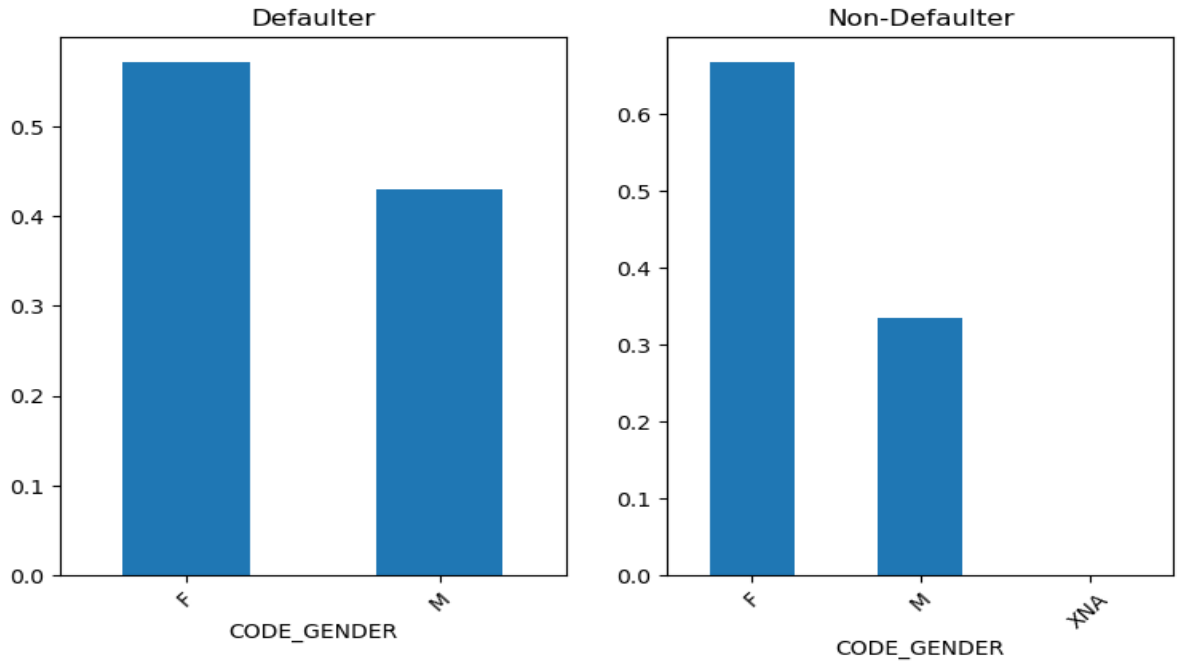
2.**LOAN APPROVAL STATUS**

Similarly, for the loan approval status variable, 63% of the data is associated with approved loans.

Univariate analysis for categorical columns

Application data





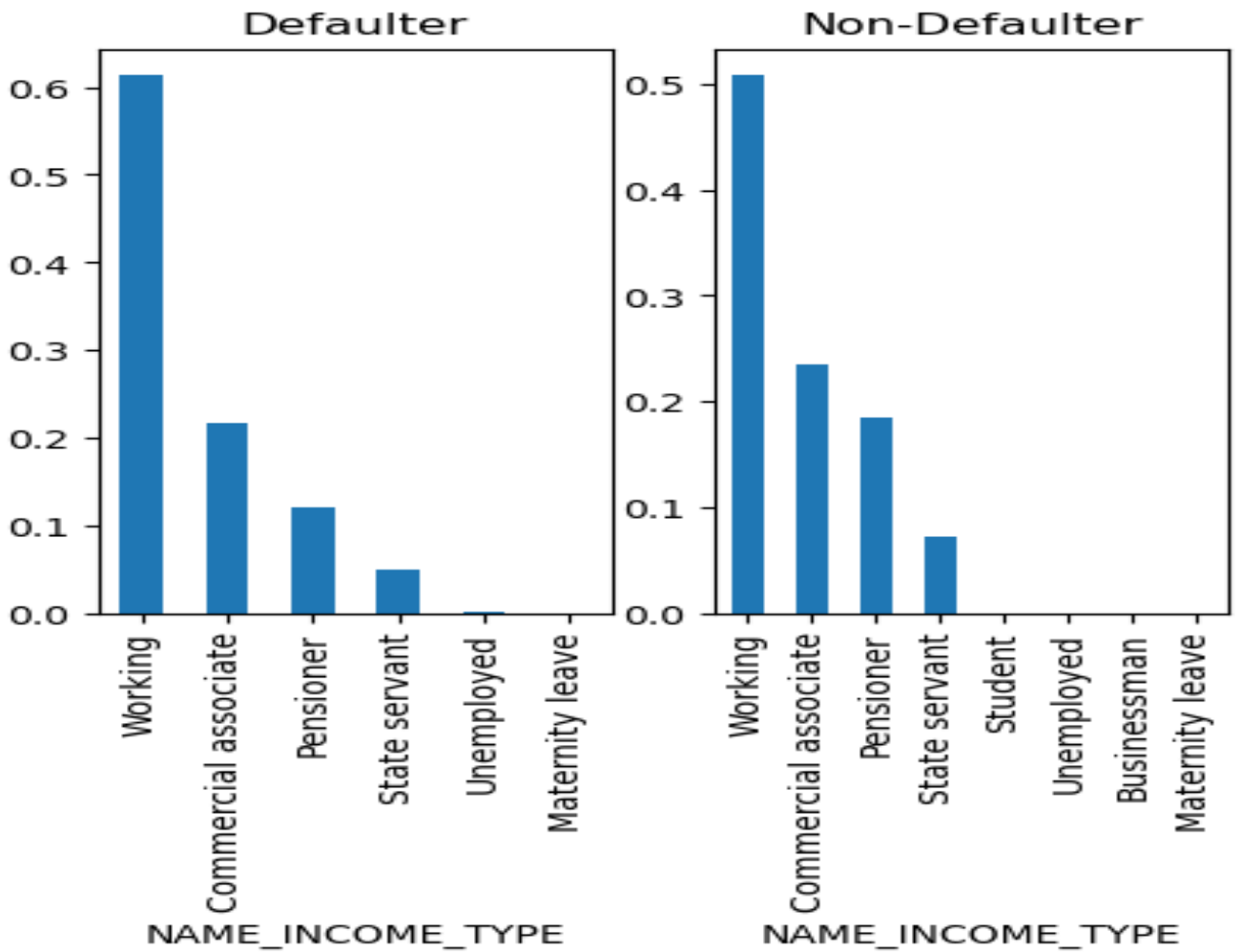
1. We conclude that the 80% of loan are the cash loan whereas the revolving loans are 20% for both.
2. Percentage of Revolving loans for non-defaulter are high.

1. Females are taking more loans as compared to male.
2. 65% Females have taken loans in comparison to 34% mal

Distribution of INCOME TYPE

Conclusion:

Individual from Income type “Working” and “Commercial associate”, the percentage are higher for both defaulter and Non-defaulter and hence they seek for more no. of loans although bank also focus on these type of Income type.

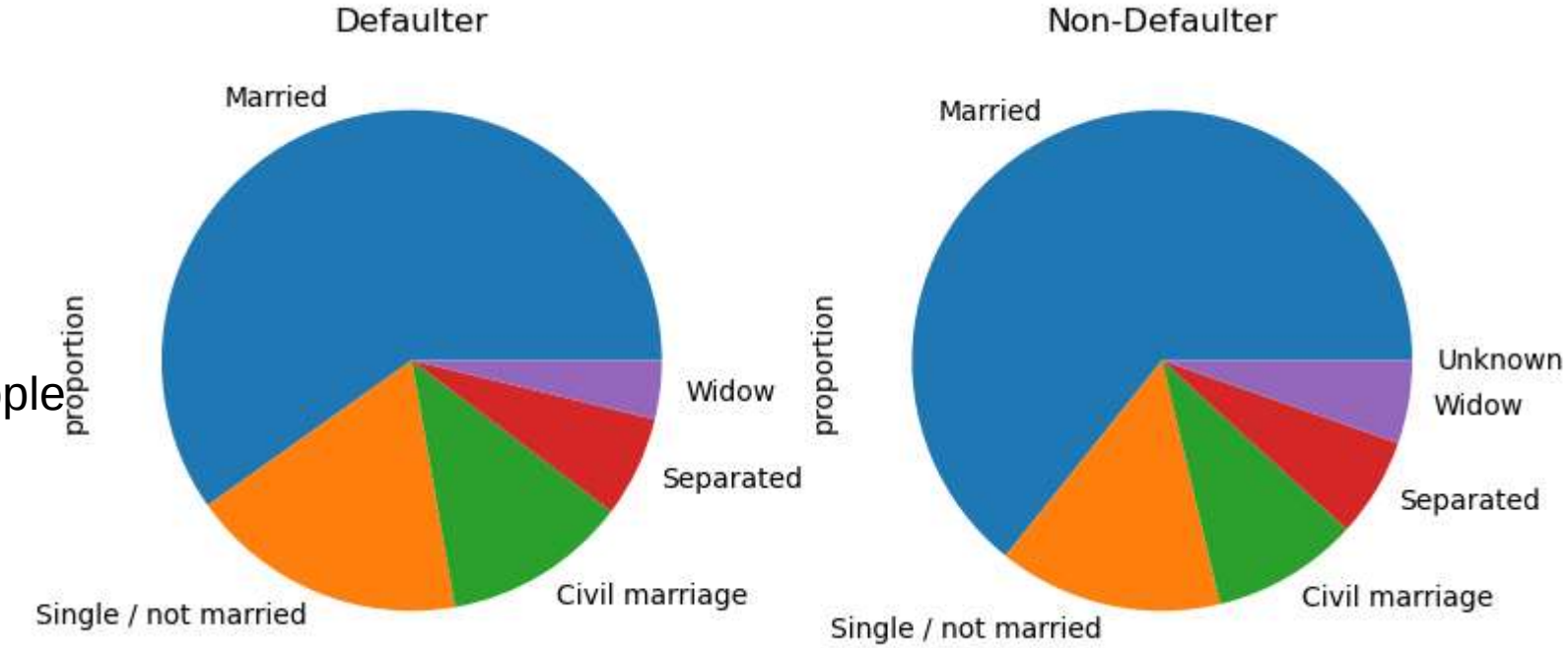


Distribution of FAMILY STATUS

Conclusion:

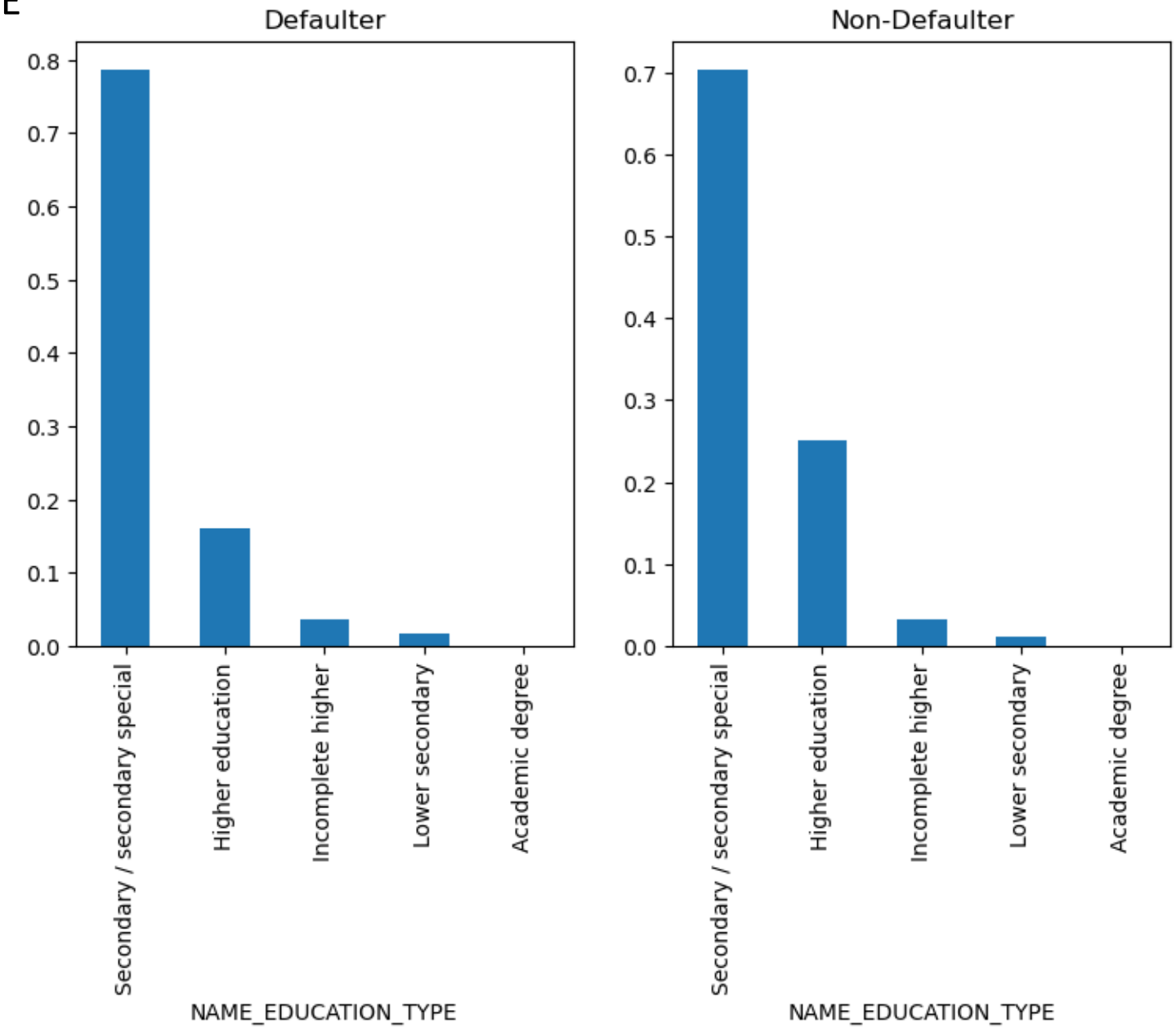
1. Married people tends to take more loan compared to other categories for both Target0 and Target1.

2.Defaulter has a higher percentage of people who are single/not married.



NAME_FAMILY_STATUS

Distribution of EDUCATION TYPE



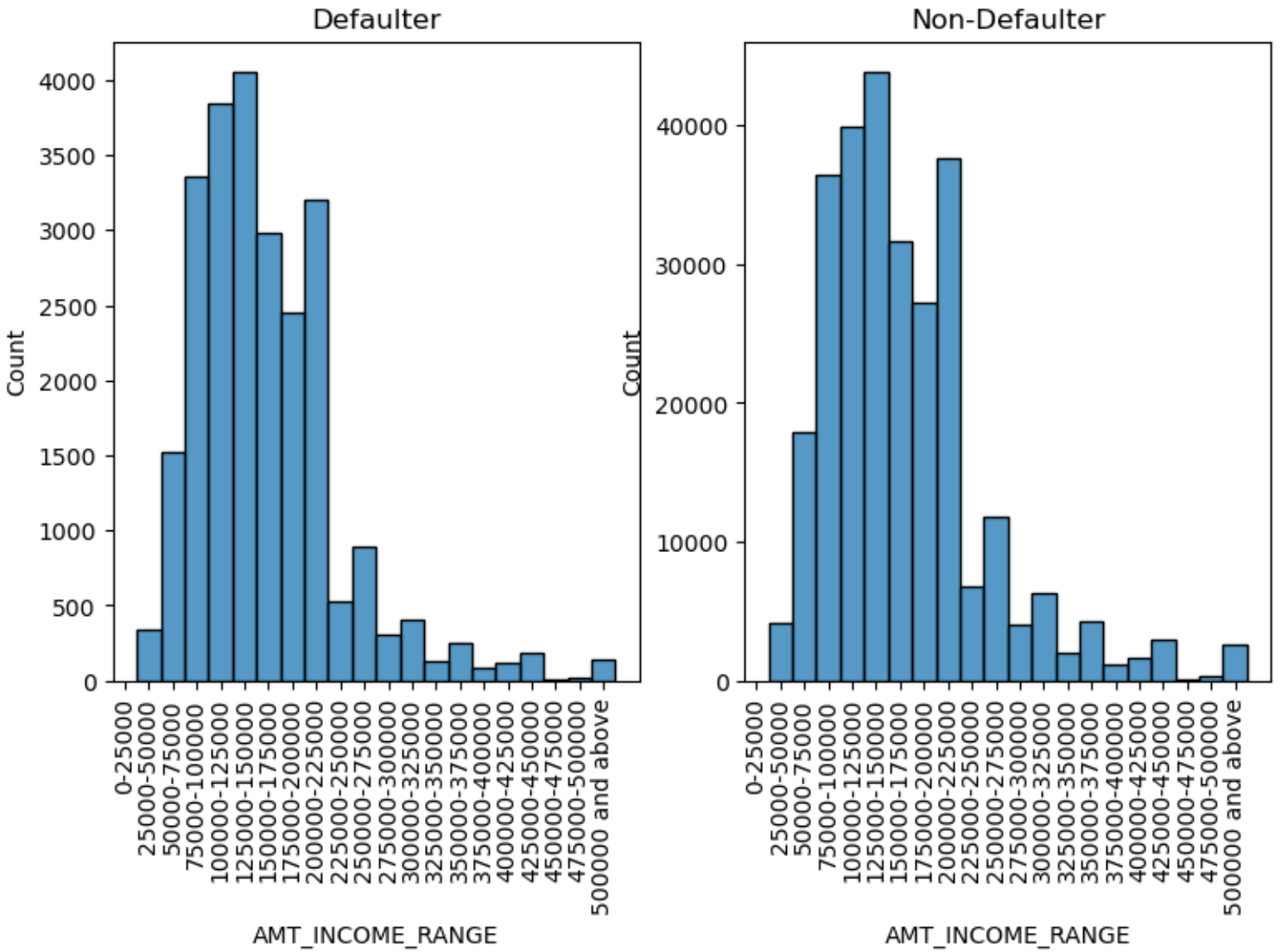
Conclusion:

1. ‘Secondary /secondary special’ people take majority of the loan and the least loan taker are from ‘Academic degree’ for both the cases.

2. Defaulter has higher percentage of Customers from ‘Higher education’.

Univariate analysis for Numerical columns

Distribution of AMT INCOME RANGE



Conclusion-

1.Peope from Income range 100000 to 225000 are taking more loan for both defaulter and non-defaulter.

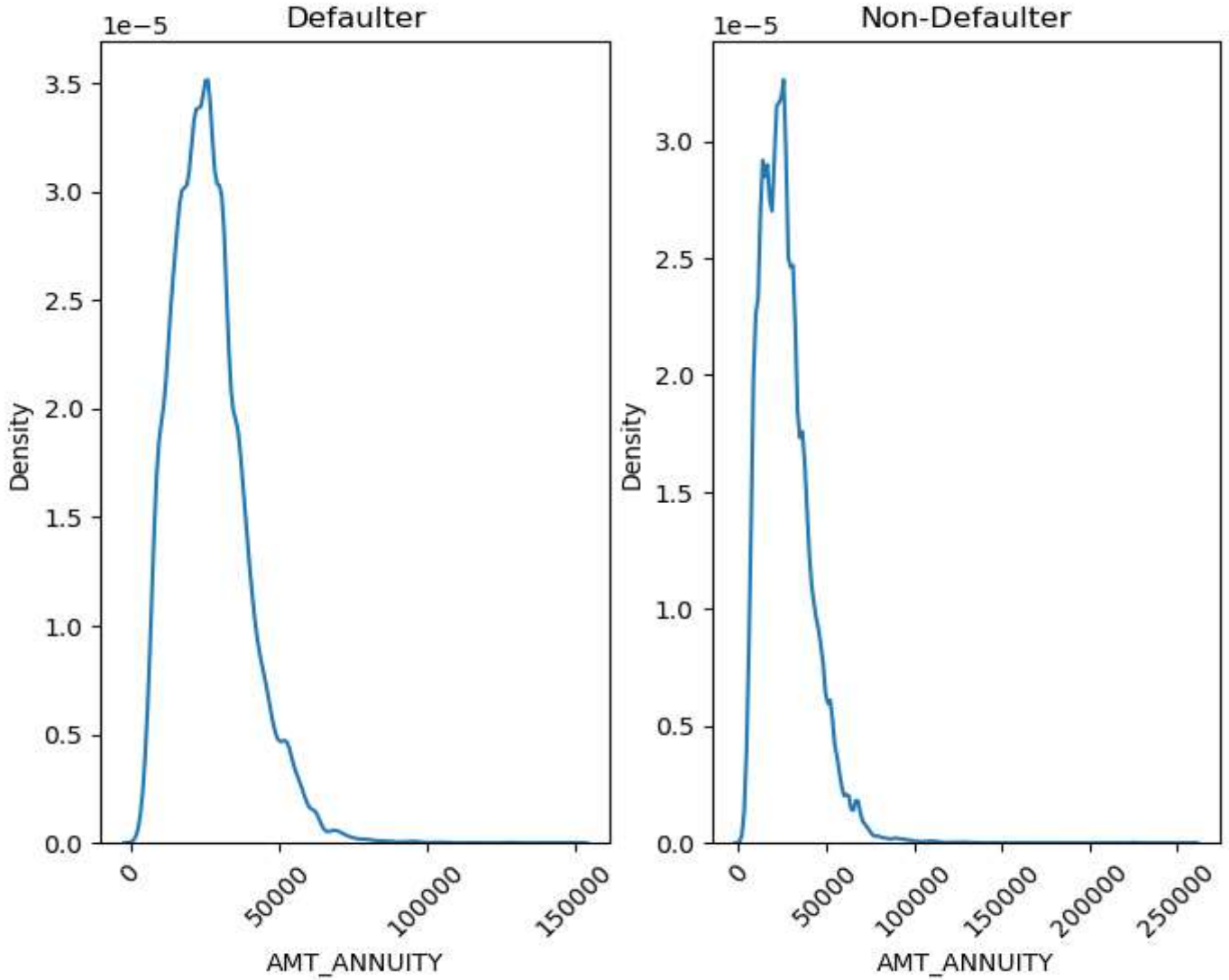
2.Very less people take loan for income range 400000 and above.

Distribution of AMT_ANNUITY

Conclusion:

1. People who has low amount of annuity takes more loan for both defaulter and non-defaulter.

2. Majority loan takers are paying in the interval 0-100000 in both the cases.



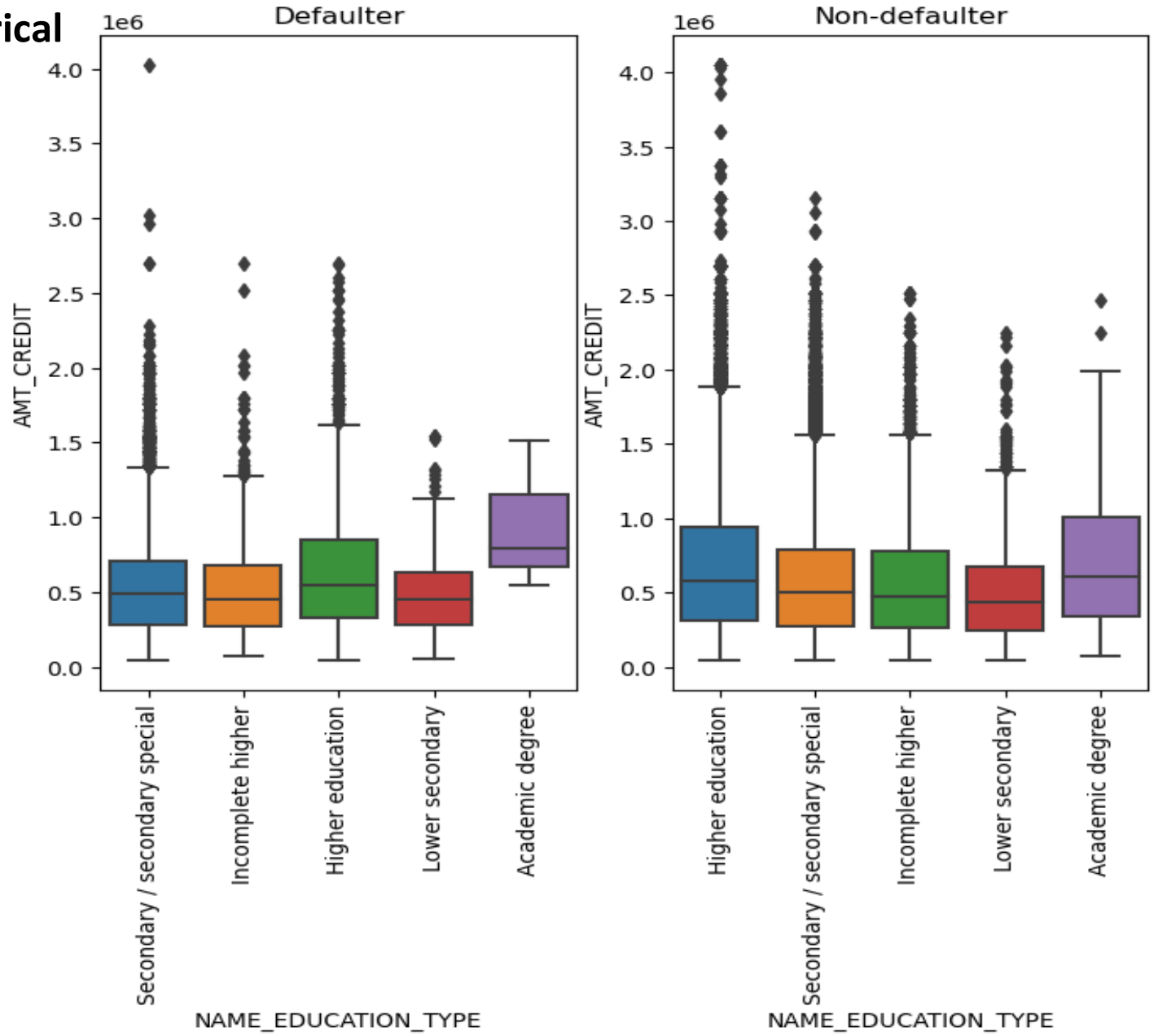
Bivariate analysis between numerical and categorical Column

NAME_EDUCATION_TYPE and AMT_CREDIT

Conclusion:

1. Outliers present across all education types except Academic degree for Defaulter, suggesting the presence of some unusually high or low values.

2. Major distribution of Academic degree lies above mean and has a higher number of credits.



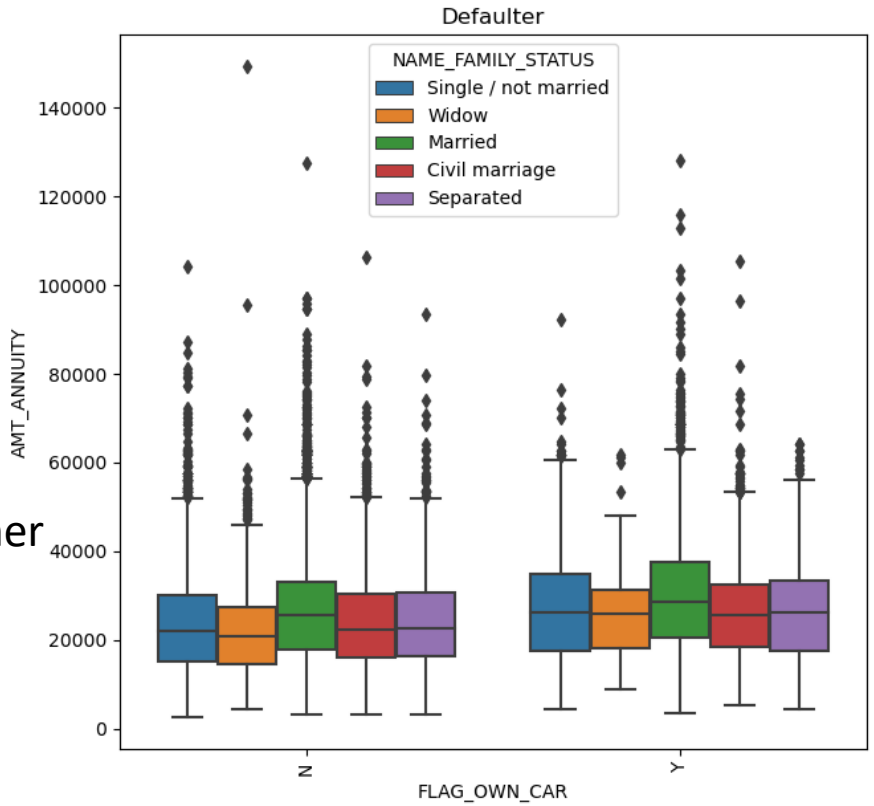
AMT_ANNUITY AND FLAG_OWN_CAR AND NAME_FAMILY_STATUS

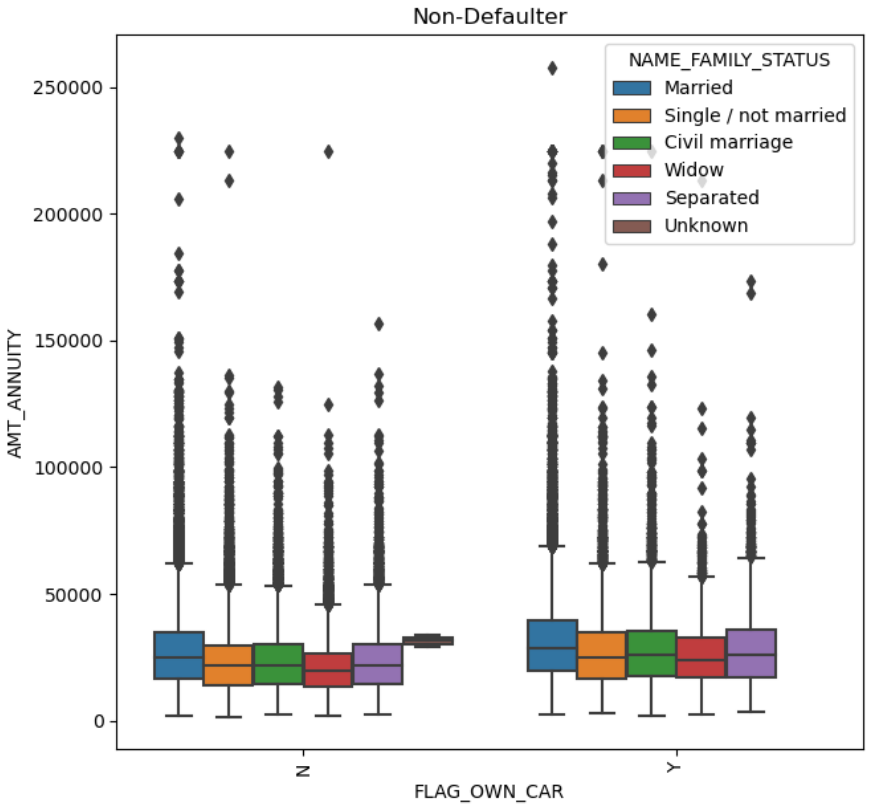
Conclusion:

1. Family status ‘Married’ has higher number of outlier for both defaulter and non-defaulter

2. Married people having car has a higher Number of defaulter.

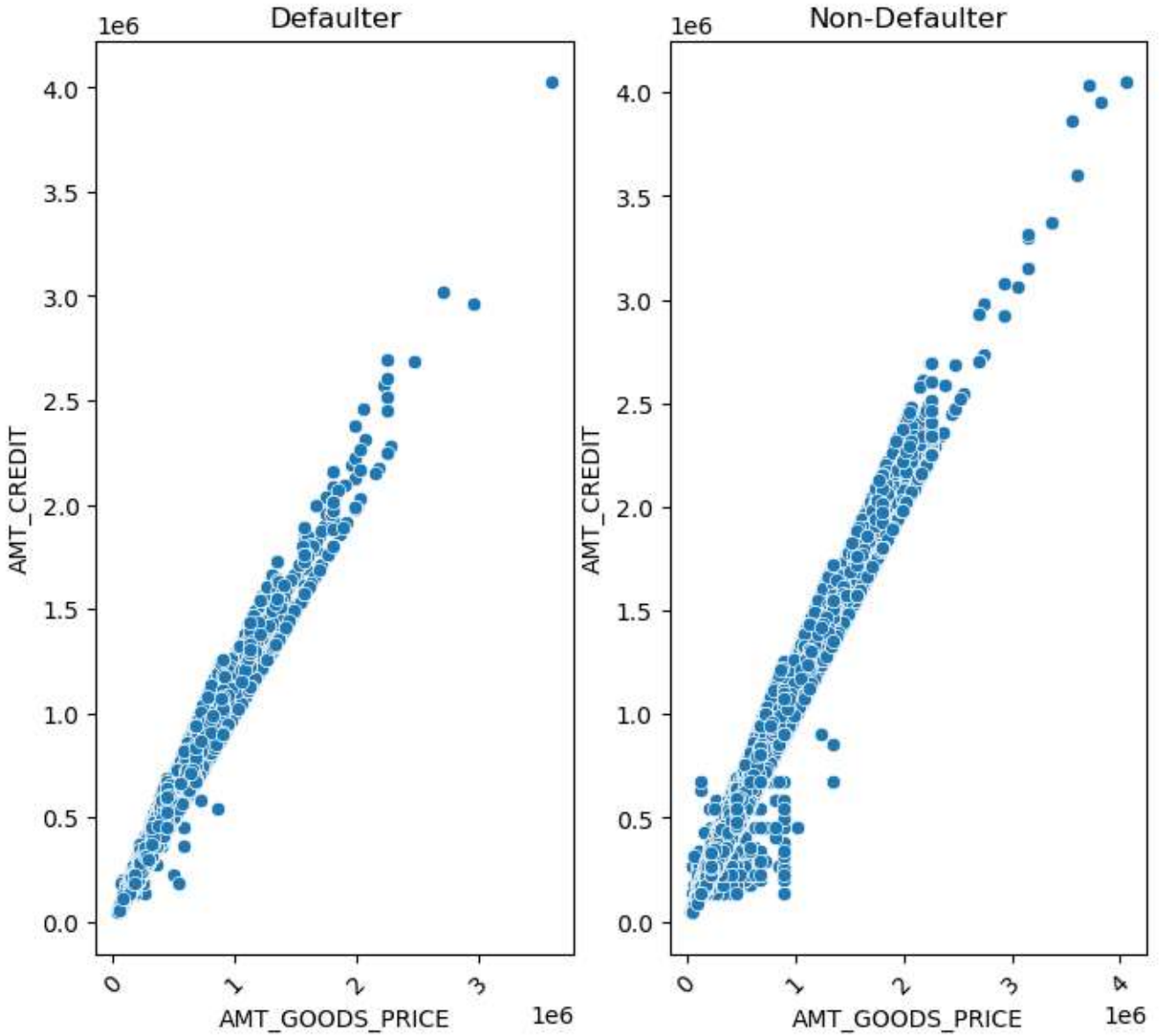
3. Amt anuuity are higher for married people





Bivariate analysis between Numerical Columns

AMT_CREDIT and AMT_GOODS_PRICE



Conclusion

1. From the graph, we conclude that there are positive correlation between AMT_GOODS_PRICE and AMT_CREDIT.

2. Higher the goods price, higher the credit amount for defaulter and non-defaulter.

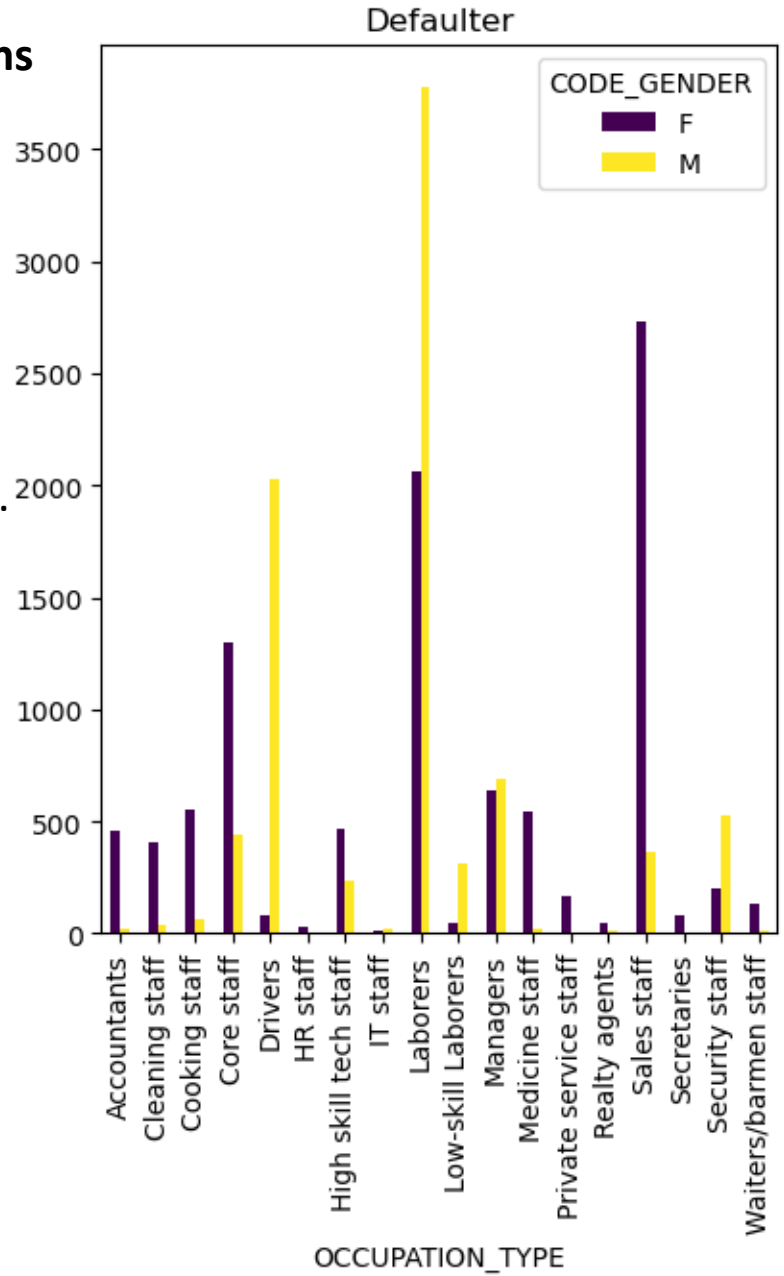
Bivariate analysis between Categorical Columns

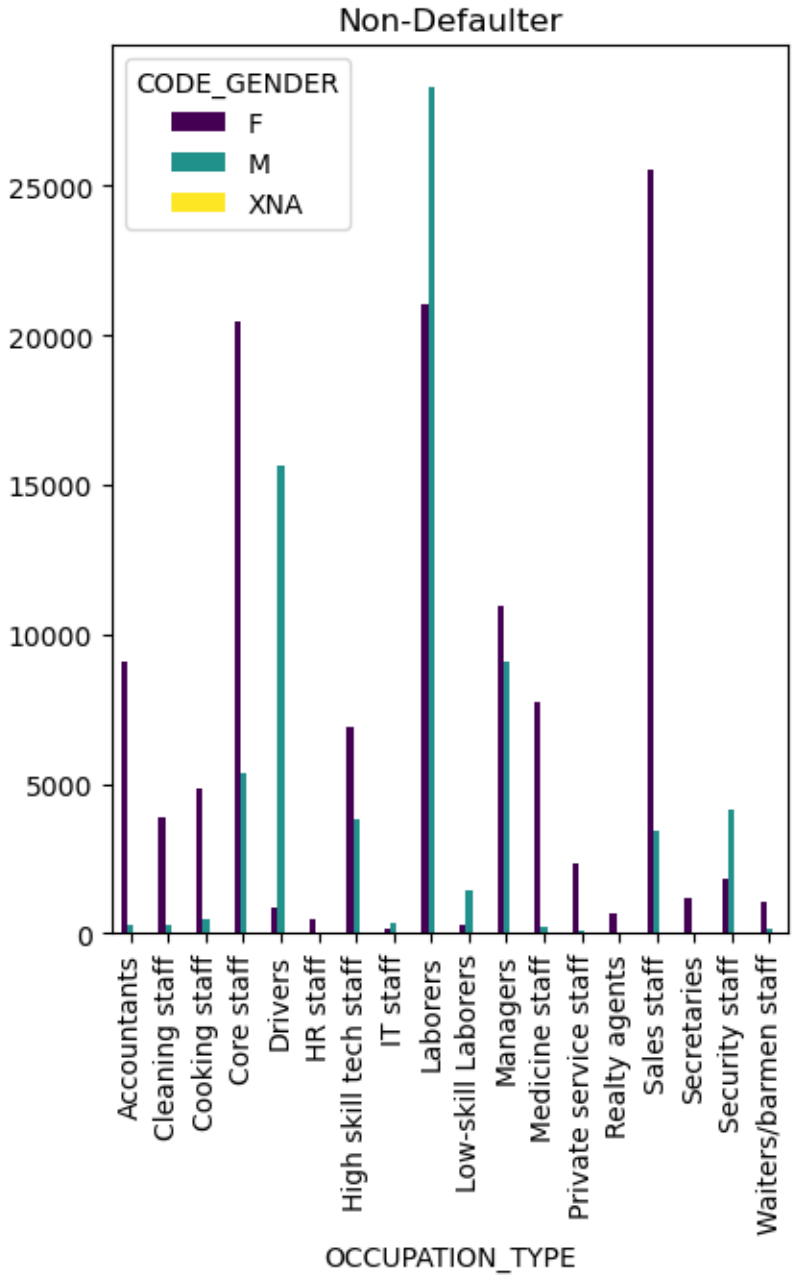
OCCUPATION_TYPE AND CODE_GENDER

Conclusion:

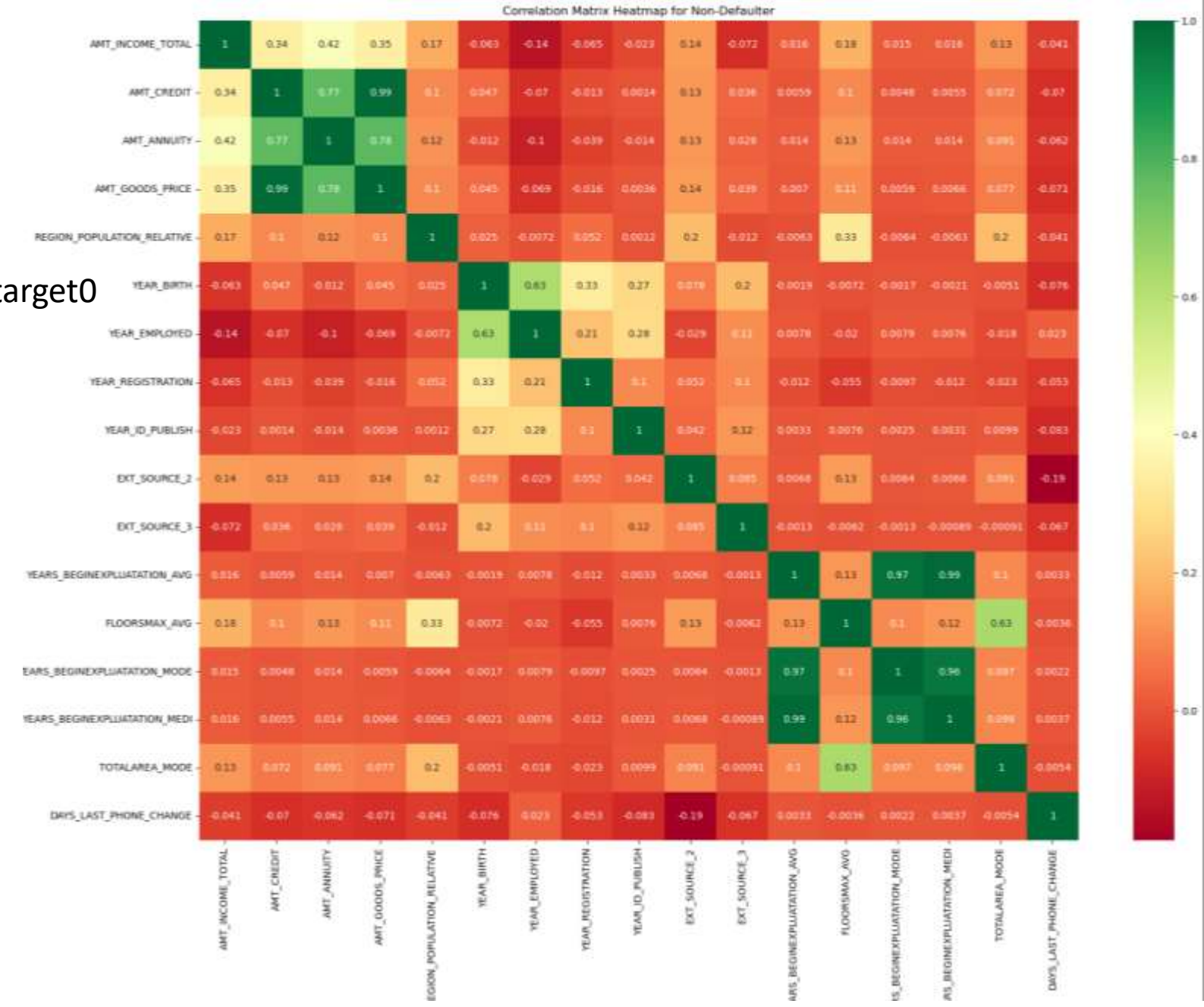
1. Male counts are more compared to female and male are more in defaulter than female.

2. Male people from occupation type like “Laborers” and “Drivers” are doing more default.





Correlation matrix for Non-Defaulter



Conclusion

These columns having high correlation value for target0

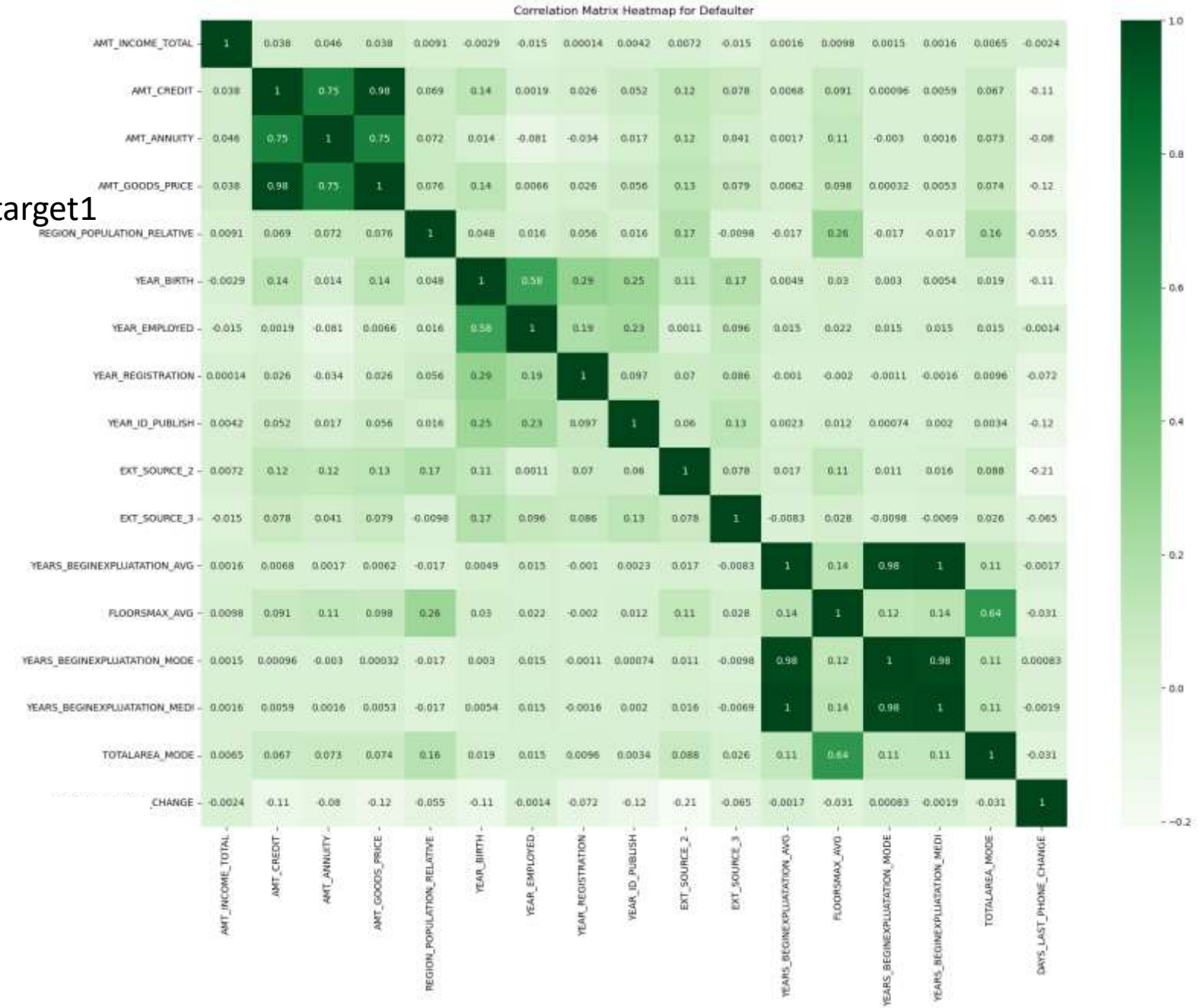
- AMT_GOODS_PROCE and AMT_CREDITS
- YEARS_BEGINEXPLUATATION_MEDI and YEARS_BEGINEXPLUATATION_AVG
- YEARS_BEGINEXPLUATATION_MODE and YEARS_BEGINEXPLUATATION_AVG
-YEARS_BEGINEXPLUATATION_MODE and YEARS_BEGINEXPLUATATION_MEDI
- ANT_ANNUITY and AMT_CREDIT

Correlation matrix for Defaulter

Conclusion

These columns having high correlation value for target1

-AMT_GOODS_PROCE and AMT_CREDITS
-YEARS_BEGINEXPLUATATION_MEDI and YEARS_BEGINEXPLUATATION_AVG
-YEARS_BEGINEXPLUATATION_MODE and YEARS_BEGINEXPLUATATION_AVG
-YEARS_BEGINEXPLUATATION_MODE and YEARS_BEGINEXPLUATATION_MEDI
-ANT_ANNUITY and AMT_CREDIT



Univariate analysis for merged dataset

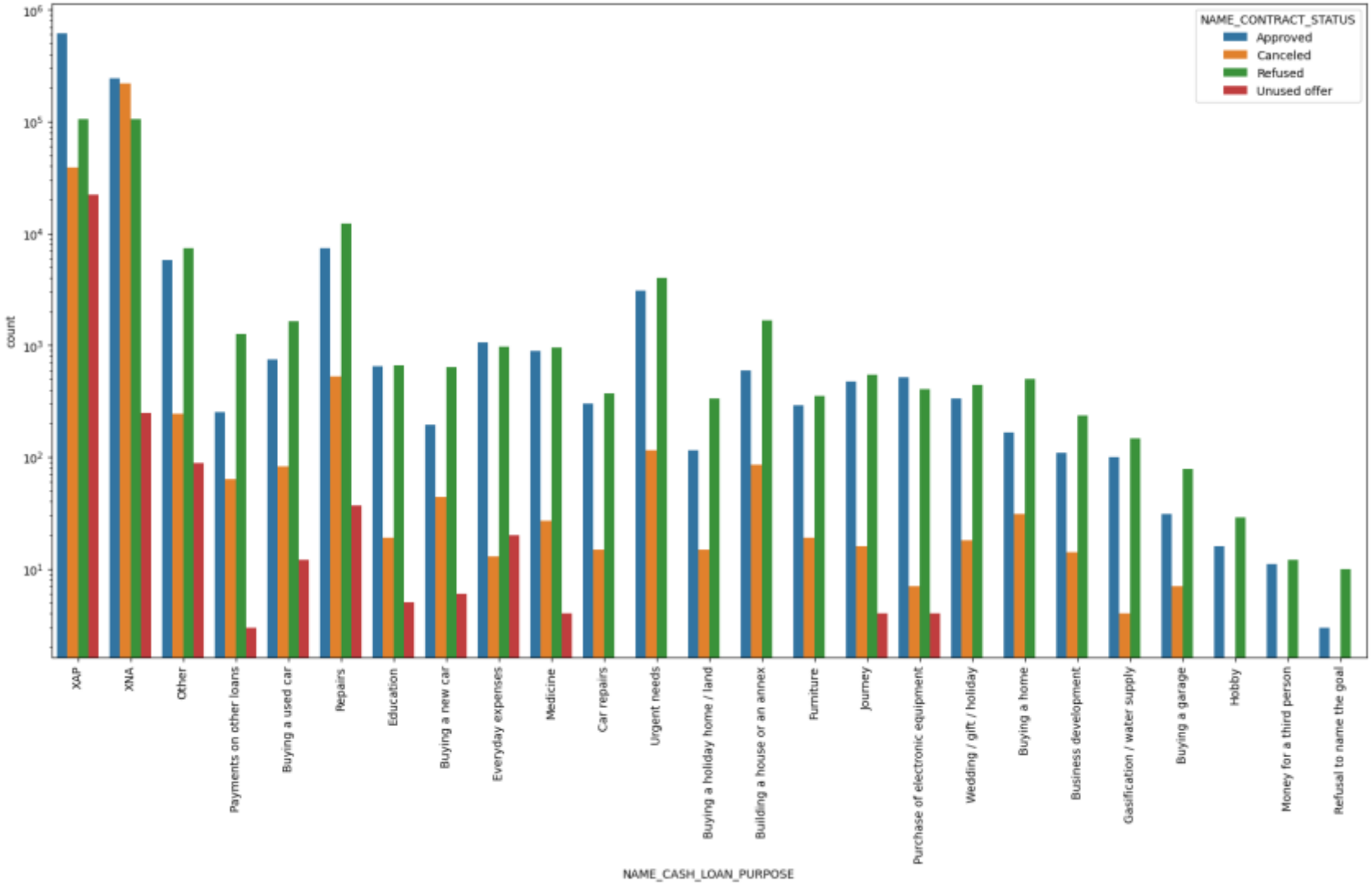
NAME_CONTRACT_STATUS & CASH_LOAN_PURPOSE

Conclusion:

1. Most approval of loan came from repairs.

2. Most rejection of loan too came from Repairs.

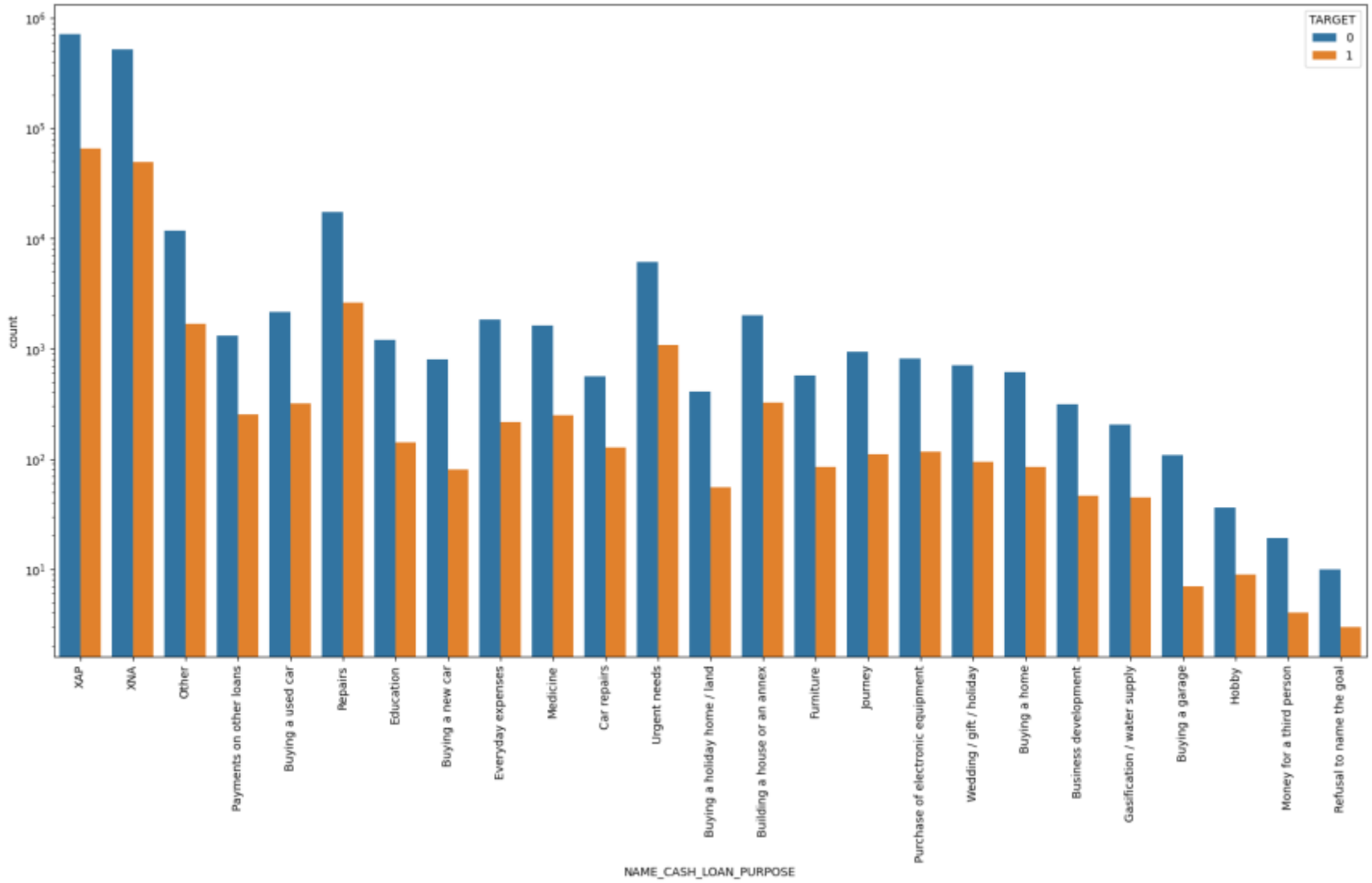
3. Equal number of approval and rejection came from Education, everyday expenses, Medicine and Car repairs.



NAME_CASH_LOAN_PURPOSE & TARGET

Conclusion:

From the graph we conclude that loan purpose “Repairs” and “Urgent needs” are facing more difficulties in paying the loan on time.

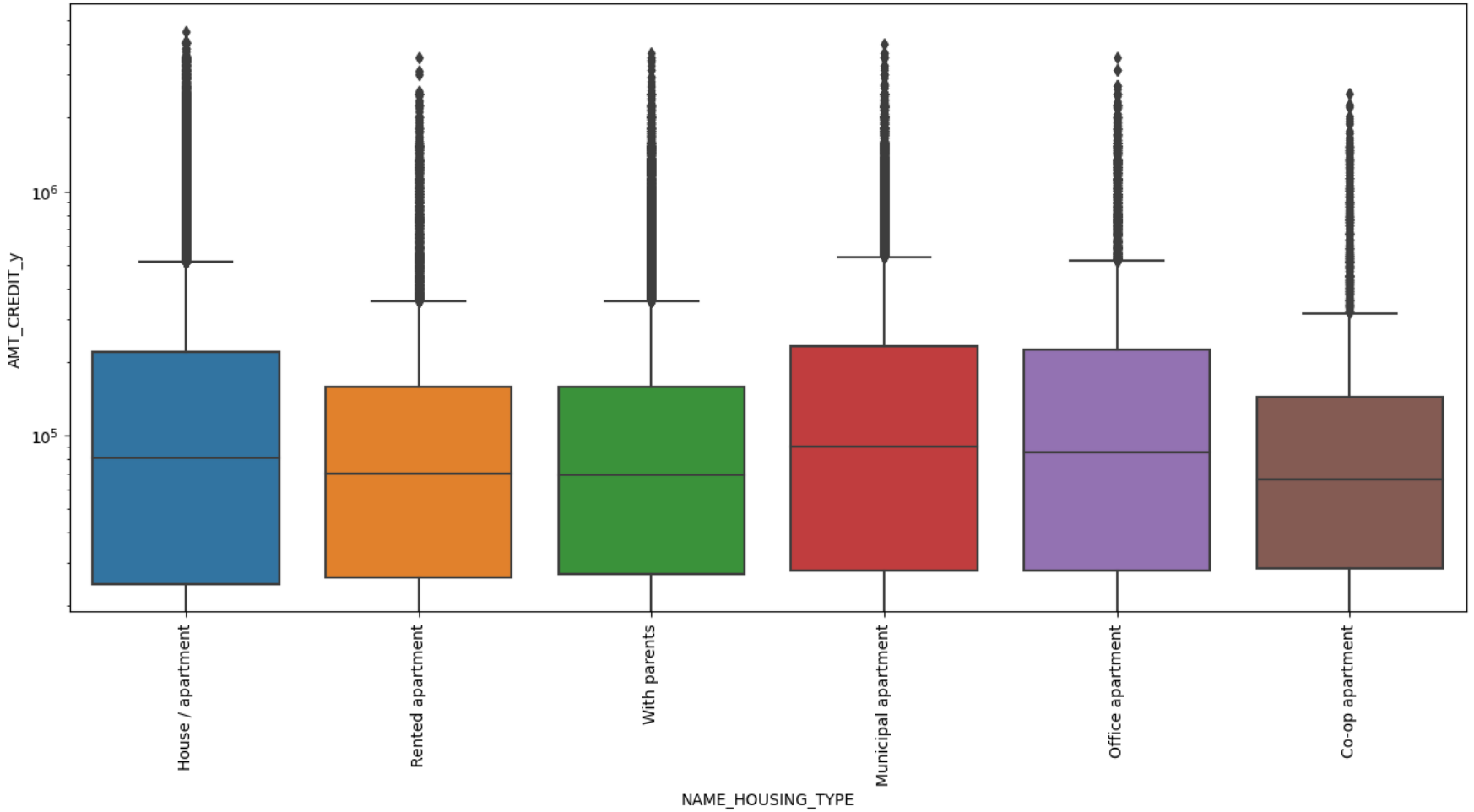


Bivariate Analysis

AMT CREDITy & NAME_HOUSING_TYPE

Conclusion:

From the graph we can conclude people who live in apartment have taken higher credit.



Conclusions for the data set provided

- 8.1 of the proportions of individuals who takes loan are defaulters
- Female takes 1.9 times more loans as compared to male.
- The percentage of revolving loans are low as compared to cash loan.
- Married and single people tends to take more loan.
- Defaulter has higher percentage of customers from “Higher Education.
- People who has low annual income tends to take more loans and falls under defaulter category.
- Higher the goods price, higher the credit amount for both defaulter and non-defaulter.
- Male people from occupation type “Laborers” and “Drivers” are doing more default.
- Customers who are from income type “Working” are facing difficulties for paying the loan.
- Low income people with low credit range are more likely to default

THANK YOU!