CREDIT EDA CASE STUDY

Submitted by,
Sakshee Suryawanshi

EDA Steps

- Understanding and featuring the dataset by summarize its key concept.
- Data Cleaning : Identify missing values and outliers, binning , removal of unwanted variables, etc.
- Identify Data Imbalance
- Univariate Analysis
- Bivariate and Multivariate Analysis
- Merging application_data and previous_application datasets
- Conclusion of Analysis with Risk and Recommendation

Credit EDA Case Study

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 minimize the risk of losing money while lending to customers.

Business Objectives

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. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicant's using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

Problem Statement

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. 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:

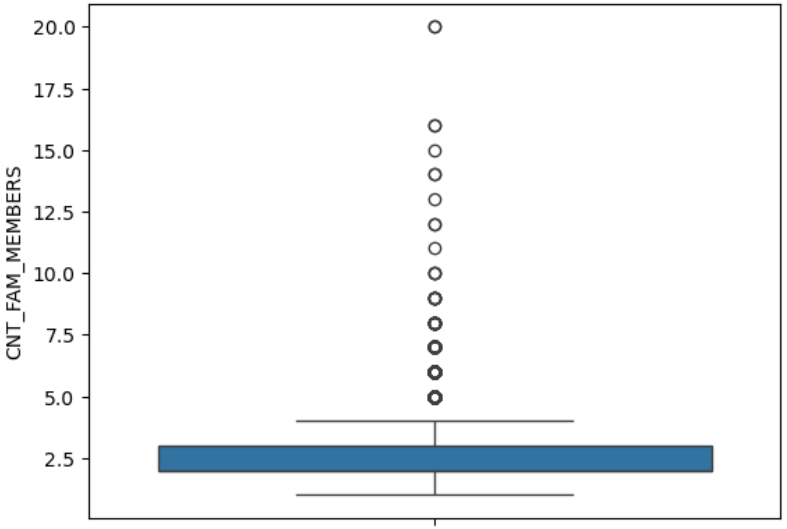
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

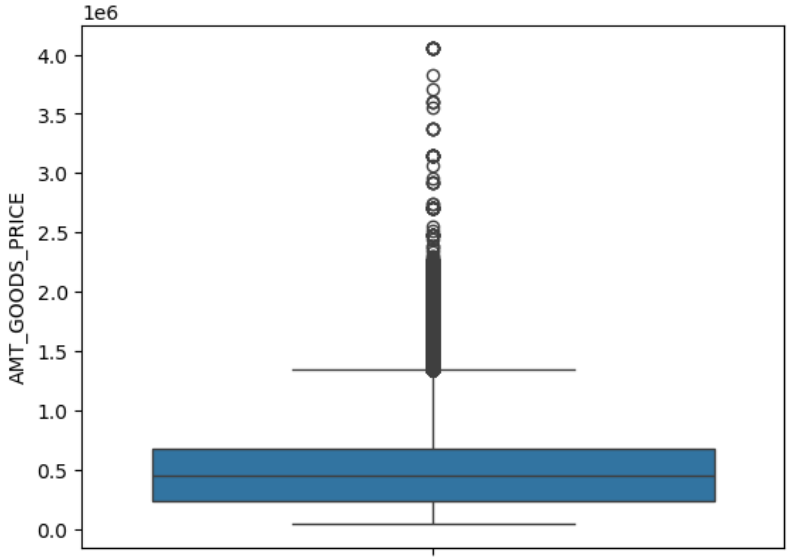
When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

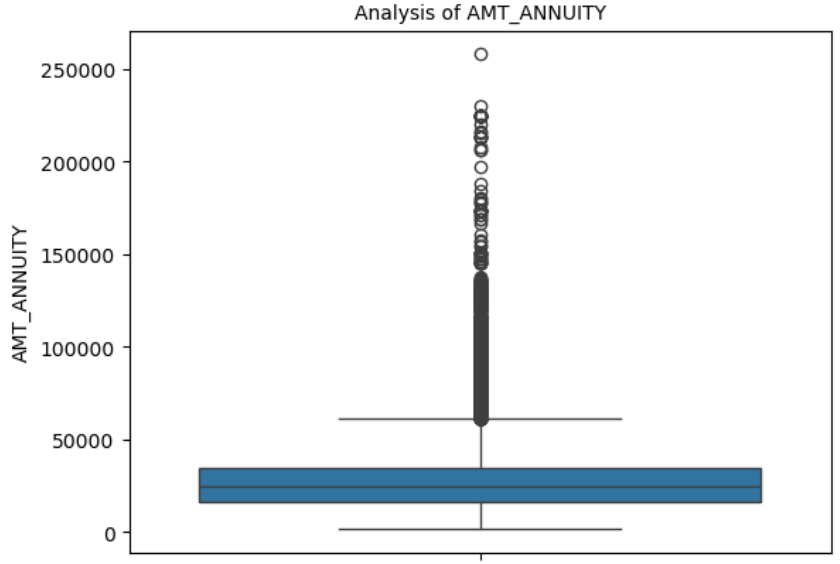
- **Approved:** The Company has approved loan Application
- **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want.
- **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
- **Unused offer:** Loan has been cancelled by the client but at different stages of the process.

Data Cleaning

- Identifying the missing values and treating them by finding the columns of dataset whose percentage of null values is greater than 45%.
- Dropping the unwanted columns having more than 45% null values in dataset.
- Impute values in columns with mean/median/mode for missing values with less percentage.
- Outliers : Identify the outliers of columns and check if they can be capped or removed. Following outliers identified in below columns like '*CNT_FAM_MEMBERS*' , '*AMT_GOODS_PRICE*' AND '*AMT_ANNUITY* '.





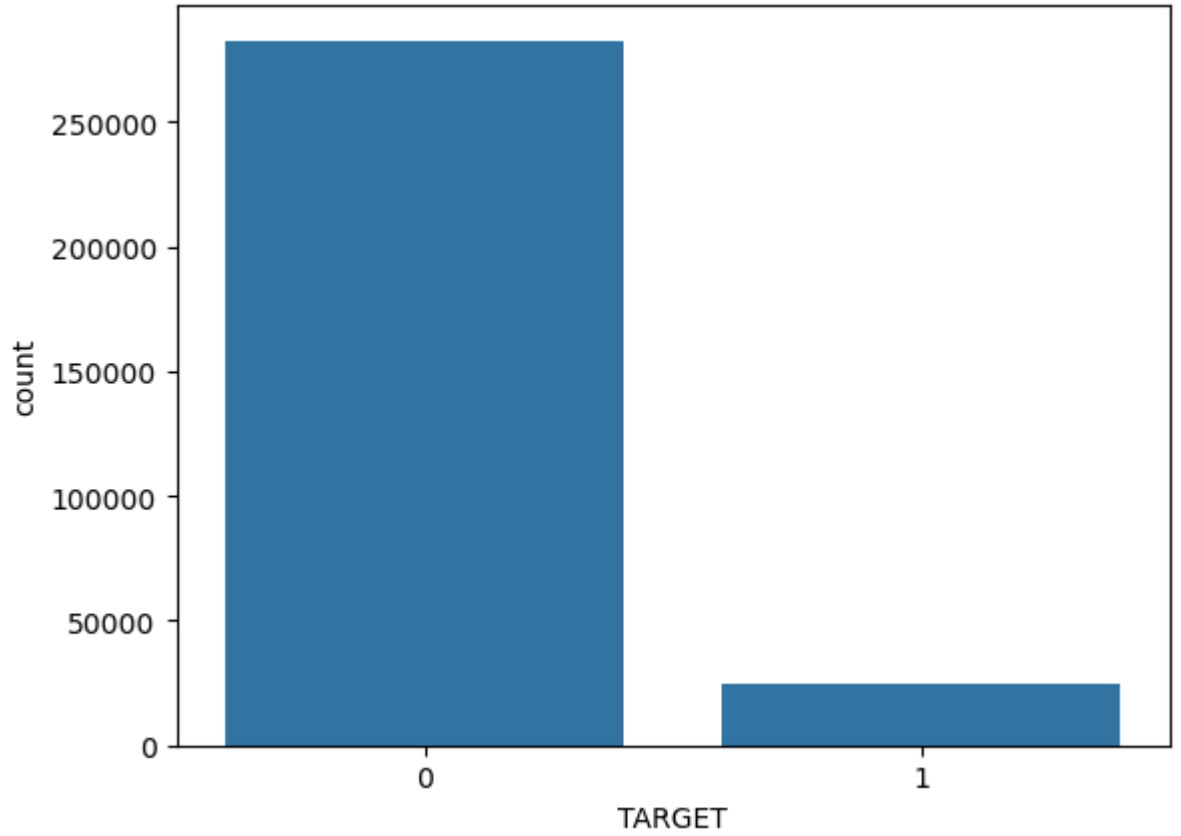


Data Imbalance

- After performing data distribution for column *'TARGET '*, we analyzed that the highly Data Imbalance where percentage for customers without payment difficulties (Target 0) is almost 91.9% and customers with payment difficulties (Target 1) is about 8.1% .

✓ Customers without payment difficulties = 0

✓ Customers with payment difficulties = 1



Univariate Analysis

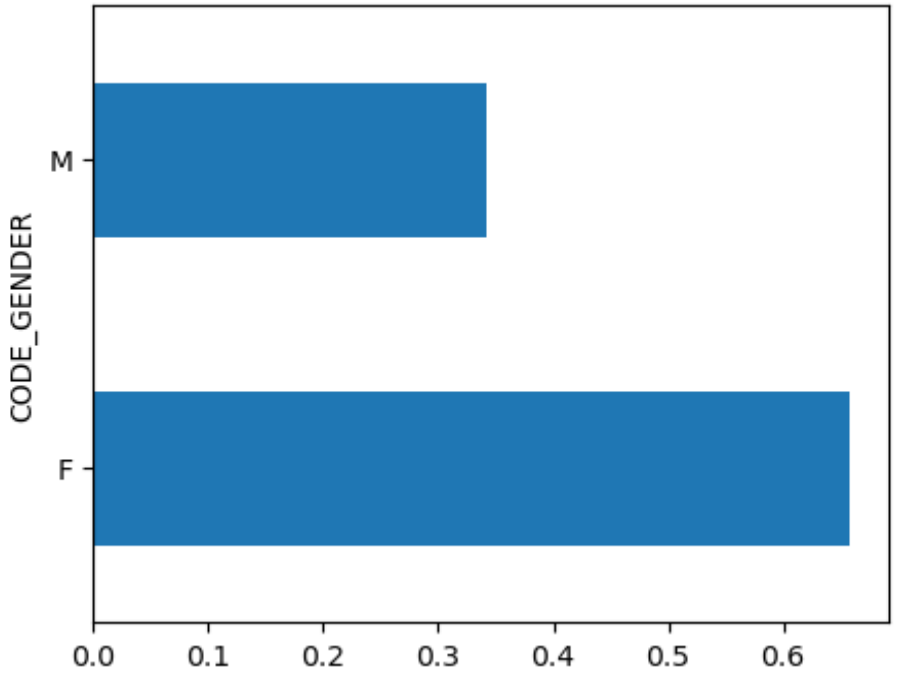
- We have performed Categorical Univariate Analysis for various columns of datasets wrt. ‘ TARGET ’ variable.
- You will find some graphs which explain the numerical/categorical variables differentiating the **clients with payment difficulties with all other cases.**
- Some of the depicted elements of graphs wrt. ‘ TARGET’ variables are below:

✓ Target-Customers without payment difficulties = 0

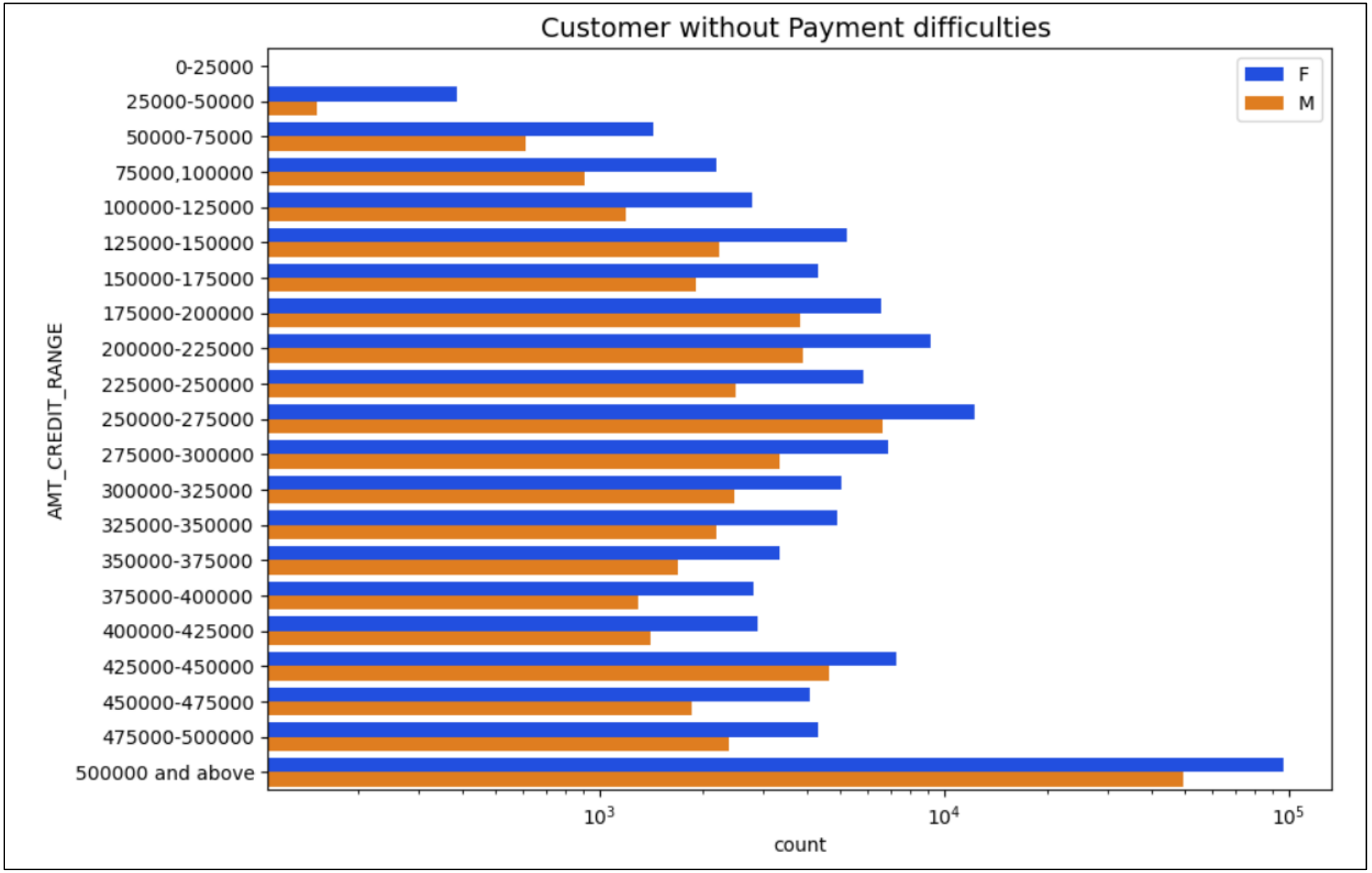
✓ Target-Customers with payment difficulties = 1

✓ Female Gender = ‘ F ’

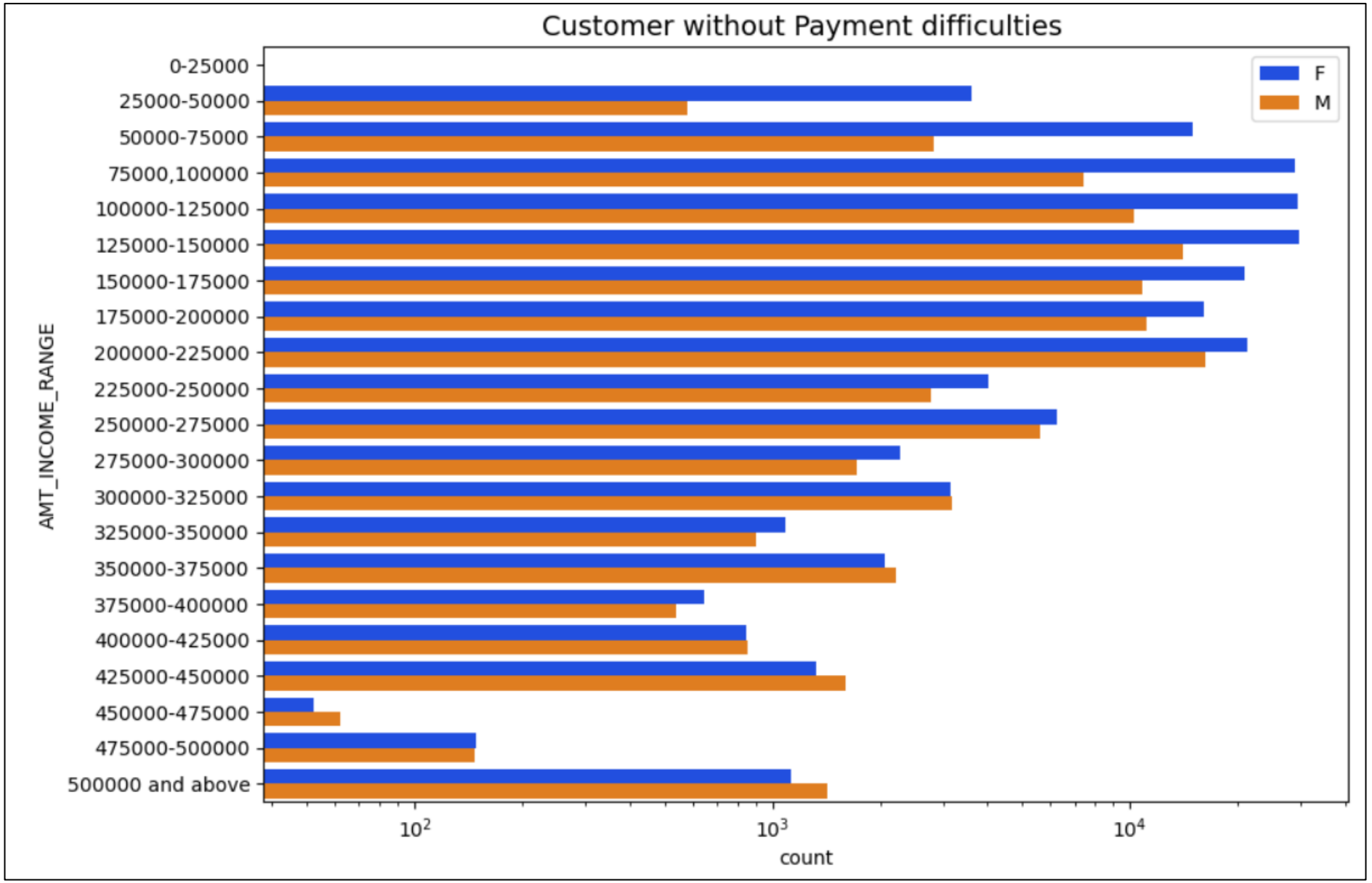
✓ Male Gender = ‘ M ’



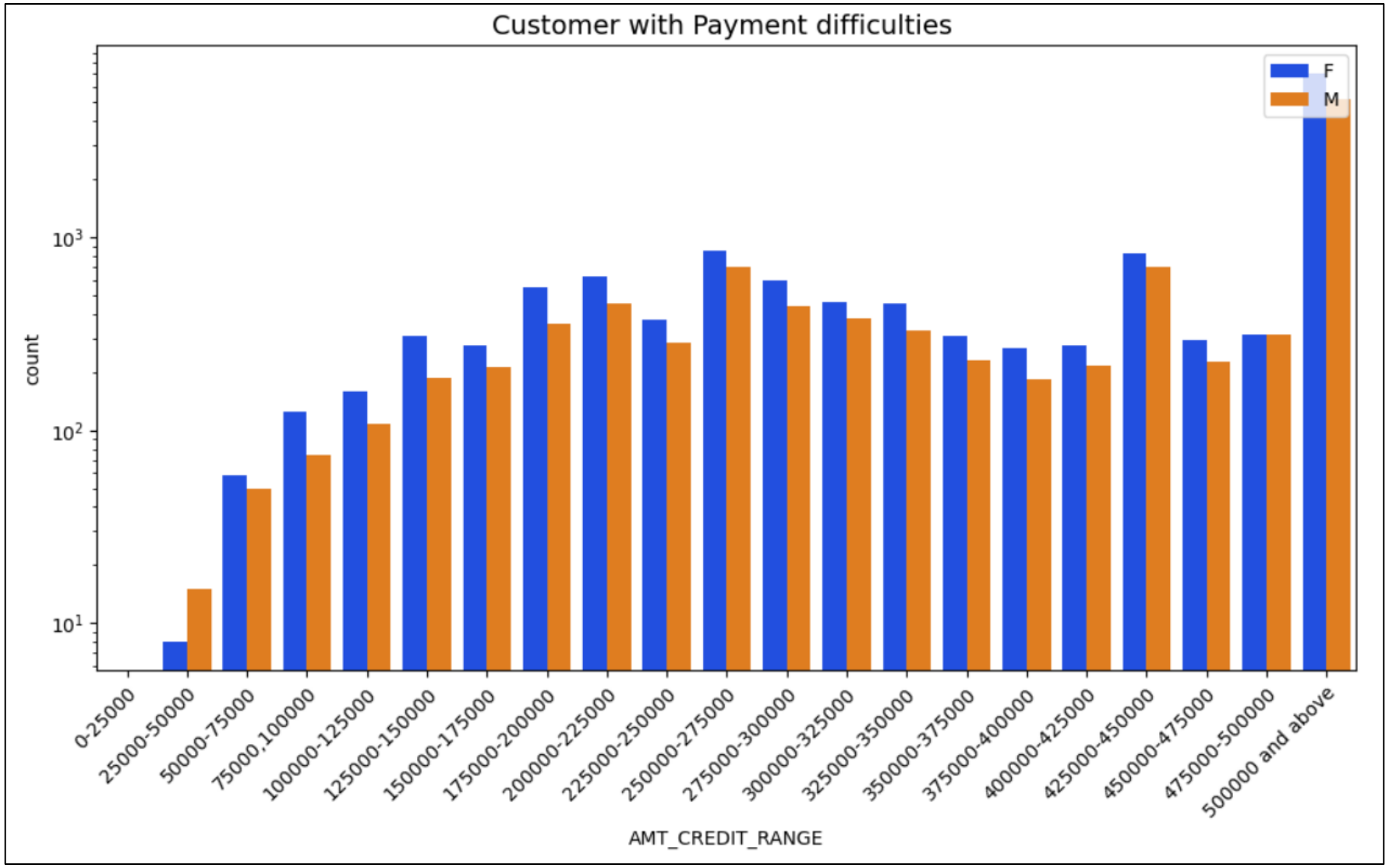
Analysis of Amount Credit Range variable wrt. TARGET - 0



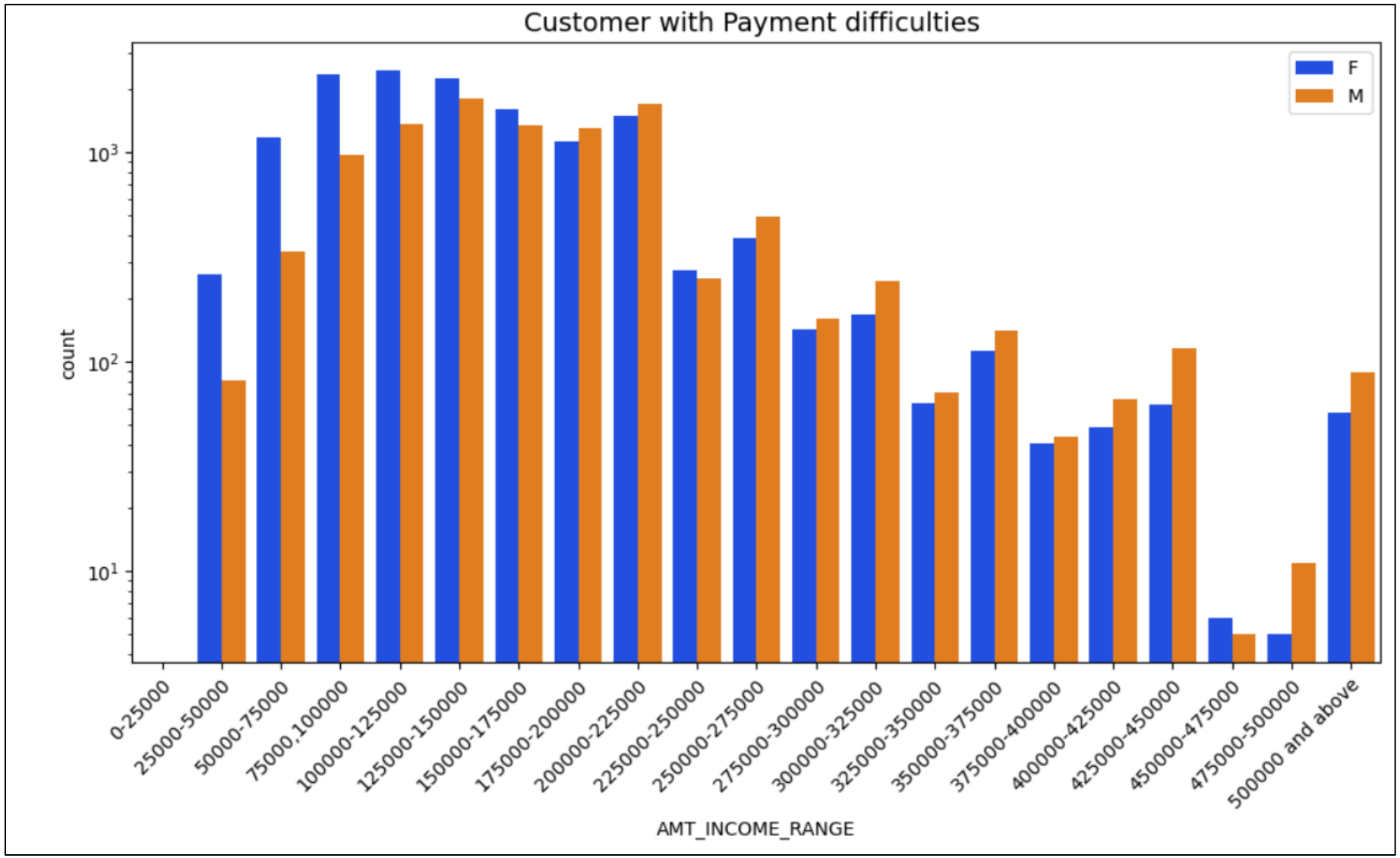
Analysis of Amount Income Range variable wrt. TARGET - 0



Analysis of Amount Credit Range variable wrt. TARGET - 1



Analysis of Amount Income Range variable wrt. TARGET - 1



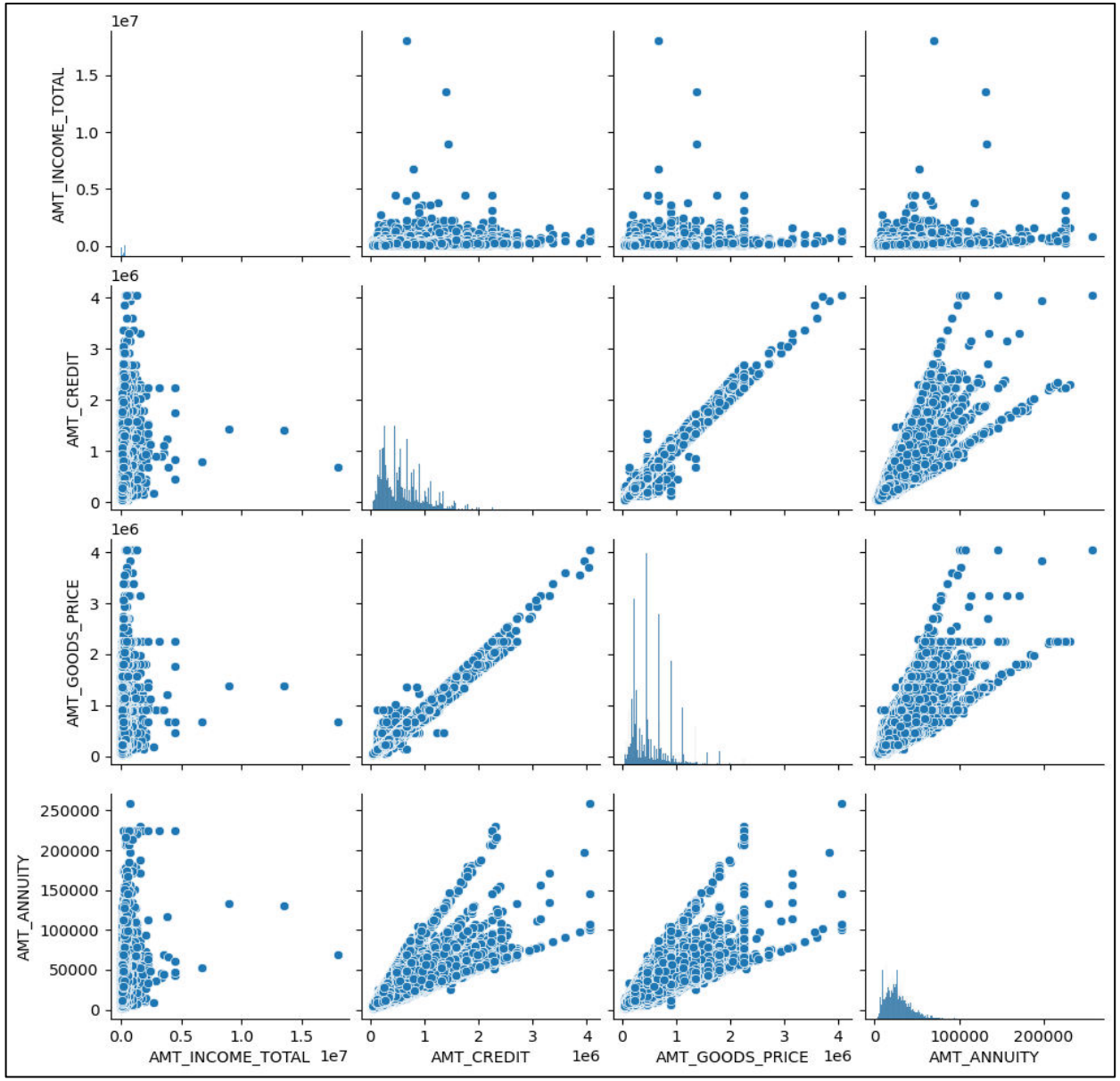
Bivariate and Multivariate Analysis

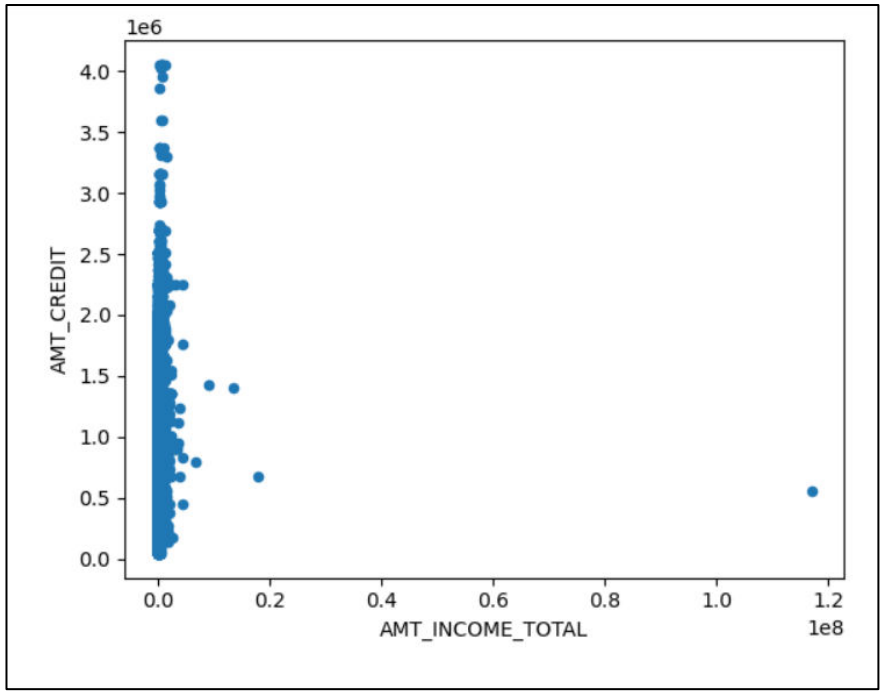
- We have performed Bivariate and Multivariate Analysis by combining multiple variables to visualize the data set.
- You will find some graphs which explain the numerical/categorical variables differentiating the **clients with payment difficulties with all other cases.**

- Here, we performed various types of bivariate and multivariate analyses :

1. Numerical – Numerical Analysis : Analysis between two numerical variables using correlation and scatter plots .
2. Numerical – Categorical Analysis : Analysis between numerical and categorical variables using countplot , barplot and boxplot

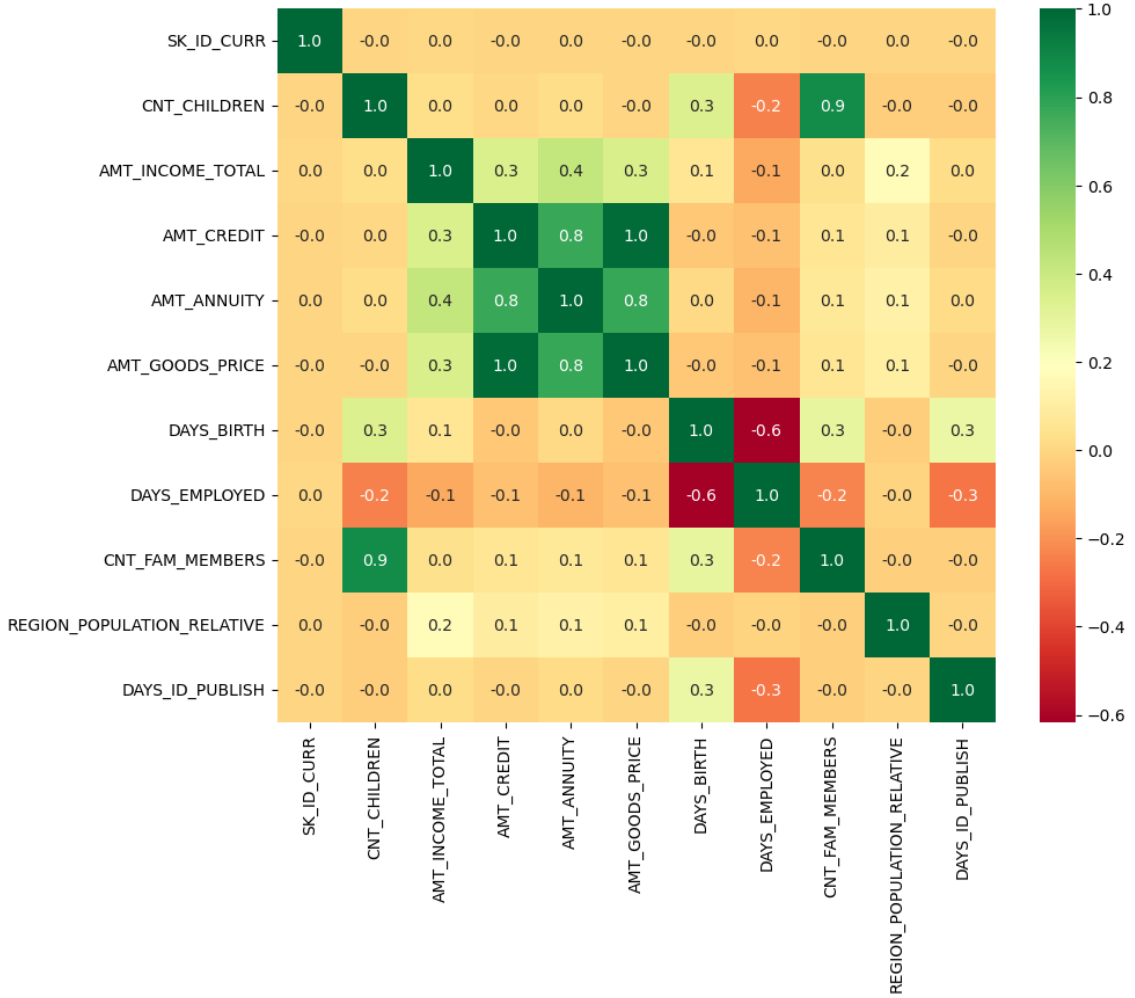
Bivariate and Multivariate Analysis of various numerical variables



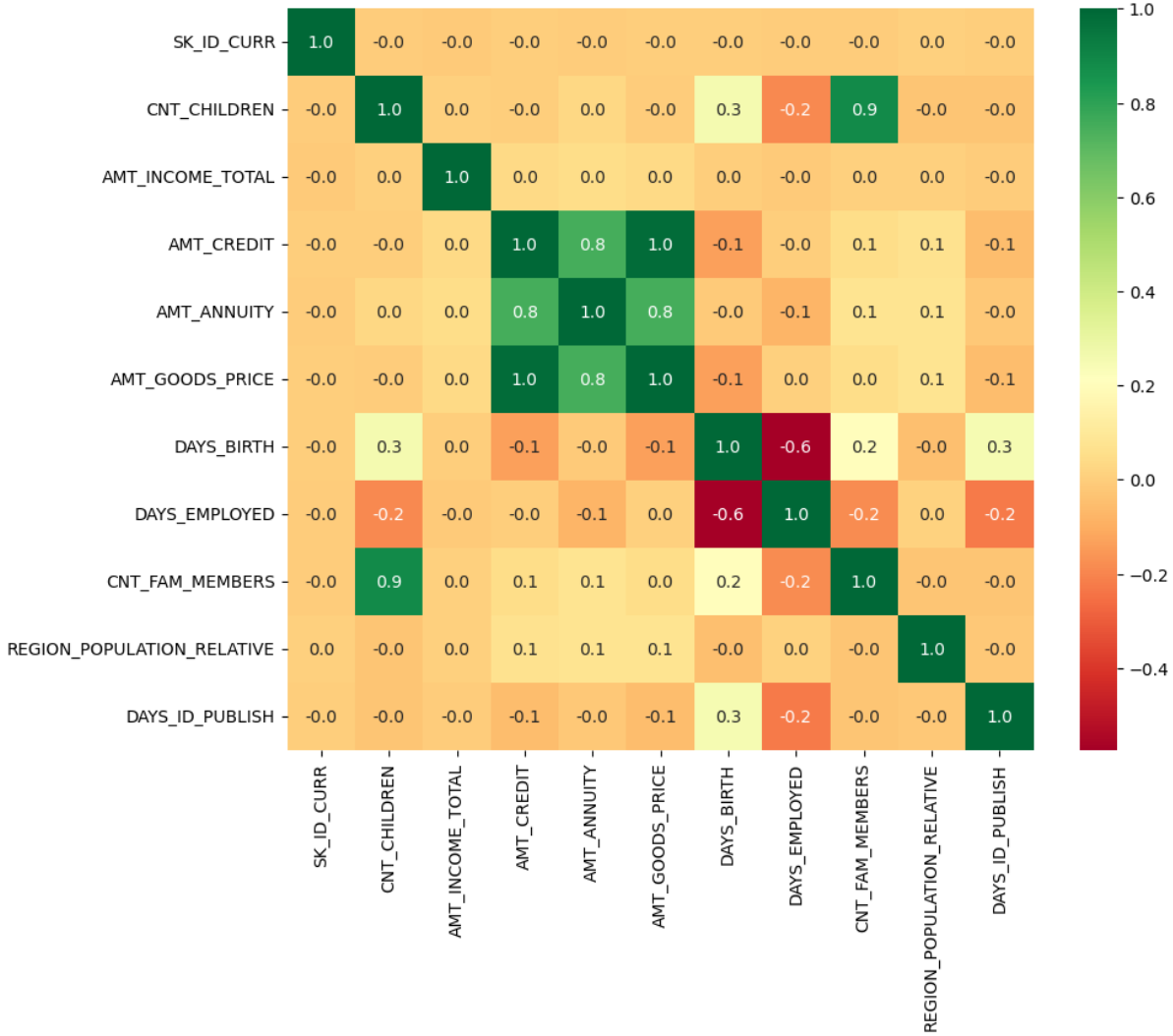


Correlation Matrix of some numeric variables wrt. Both Target variables

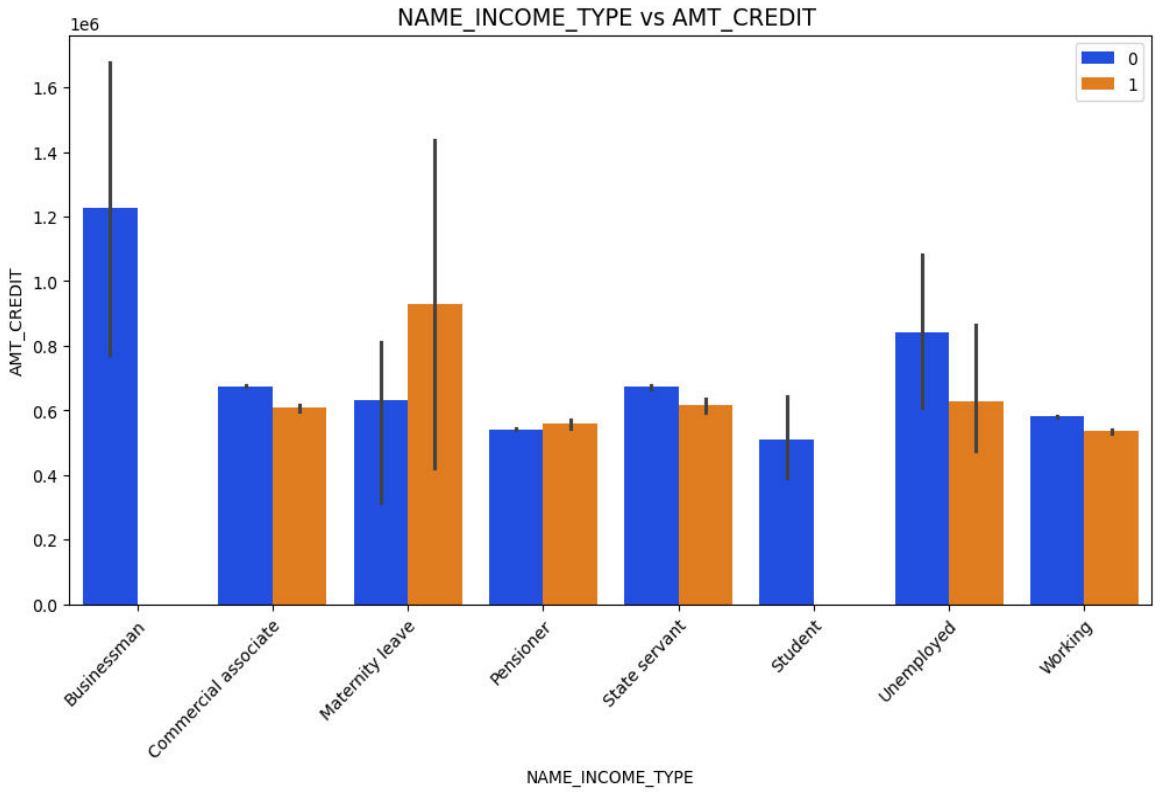
Correlation Matrix for (Target 0) Customer without Payment difficulties

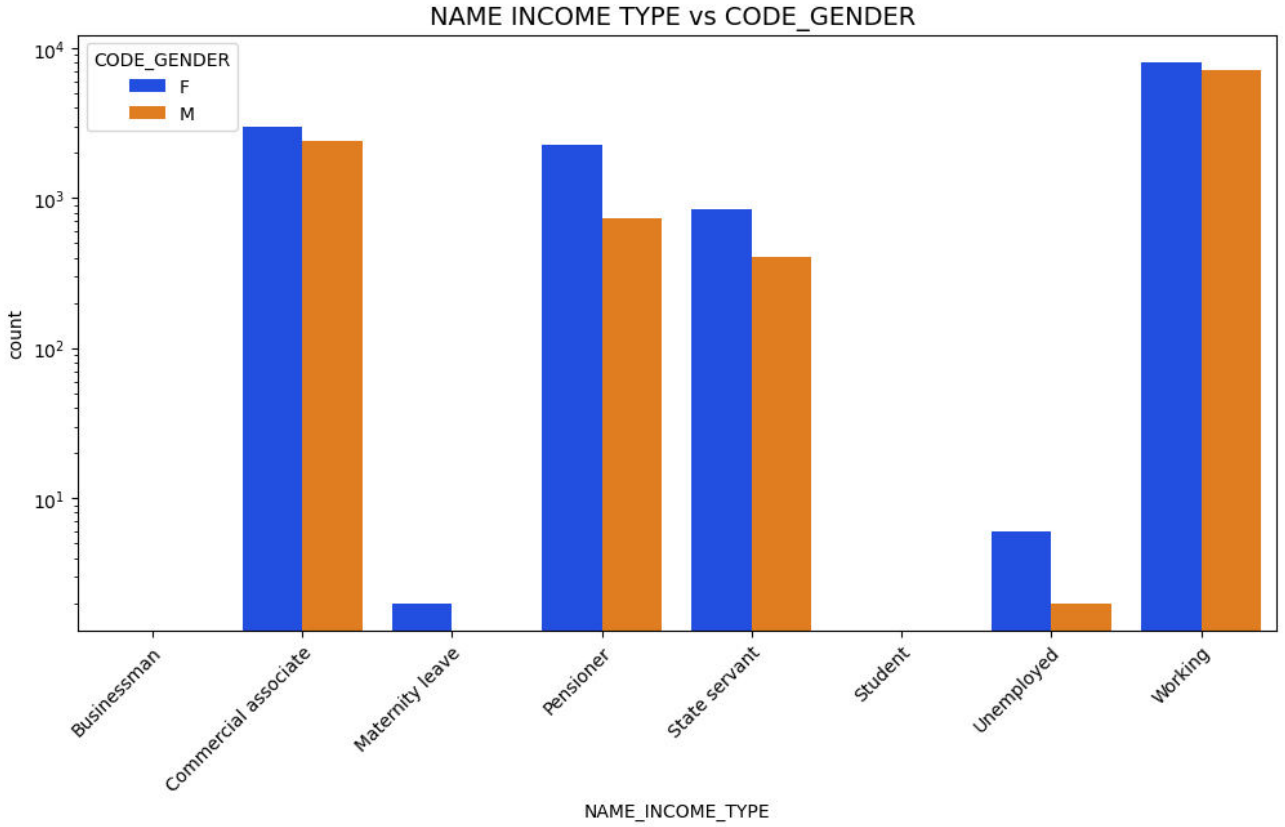


Correlation Matrix for (Target 1) Customer having Payment difficulties

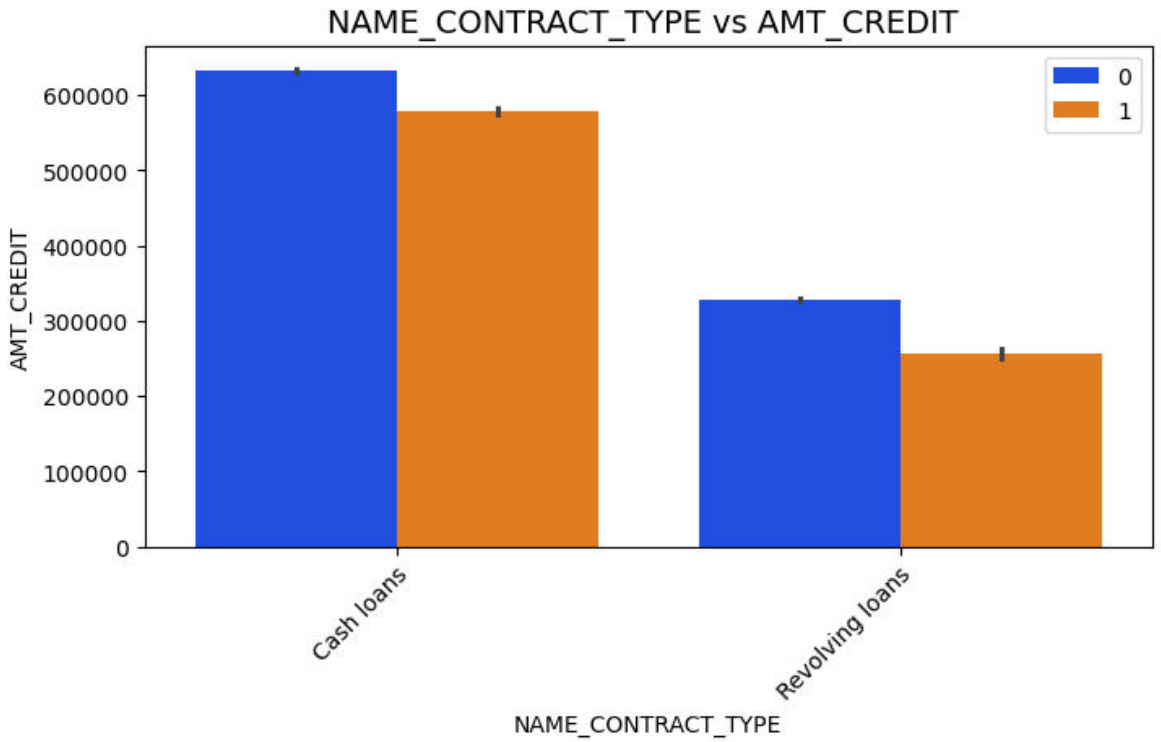


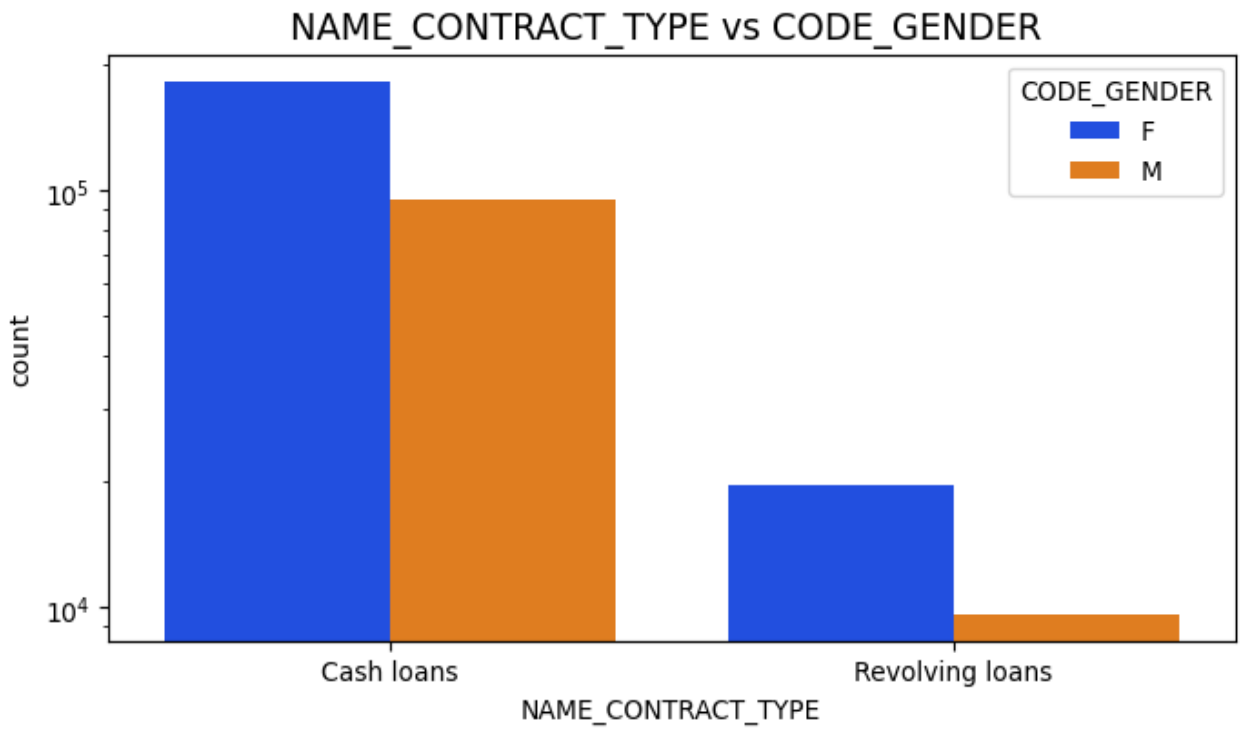
Numeric- Categorical Analysis of NAME_INCOME_TYPE variable wrt. both TARGET datasets



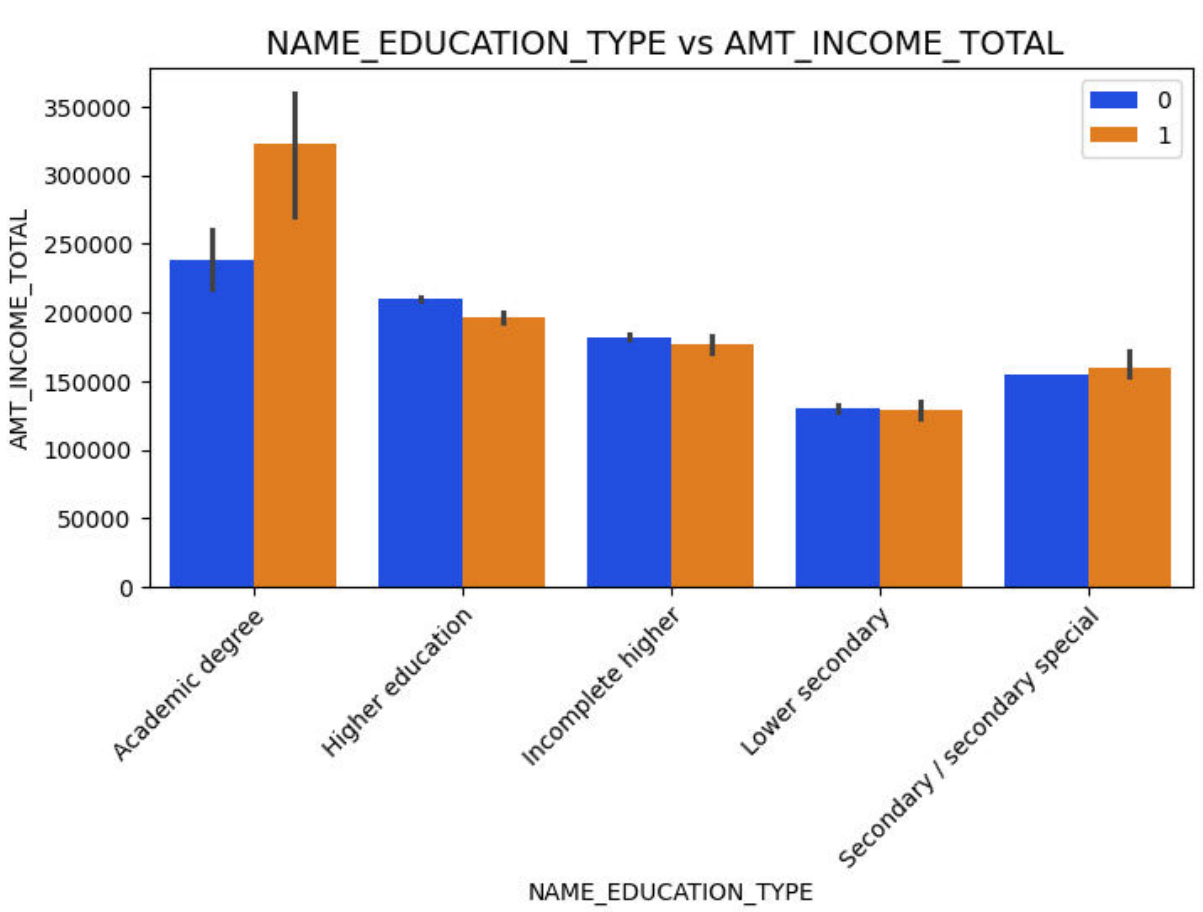


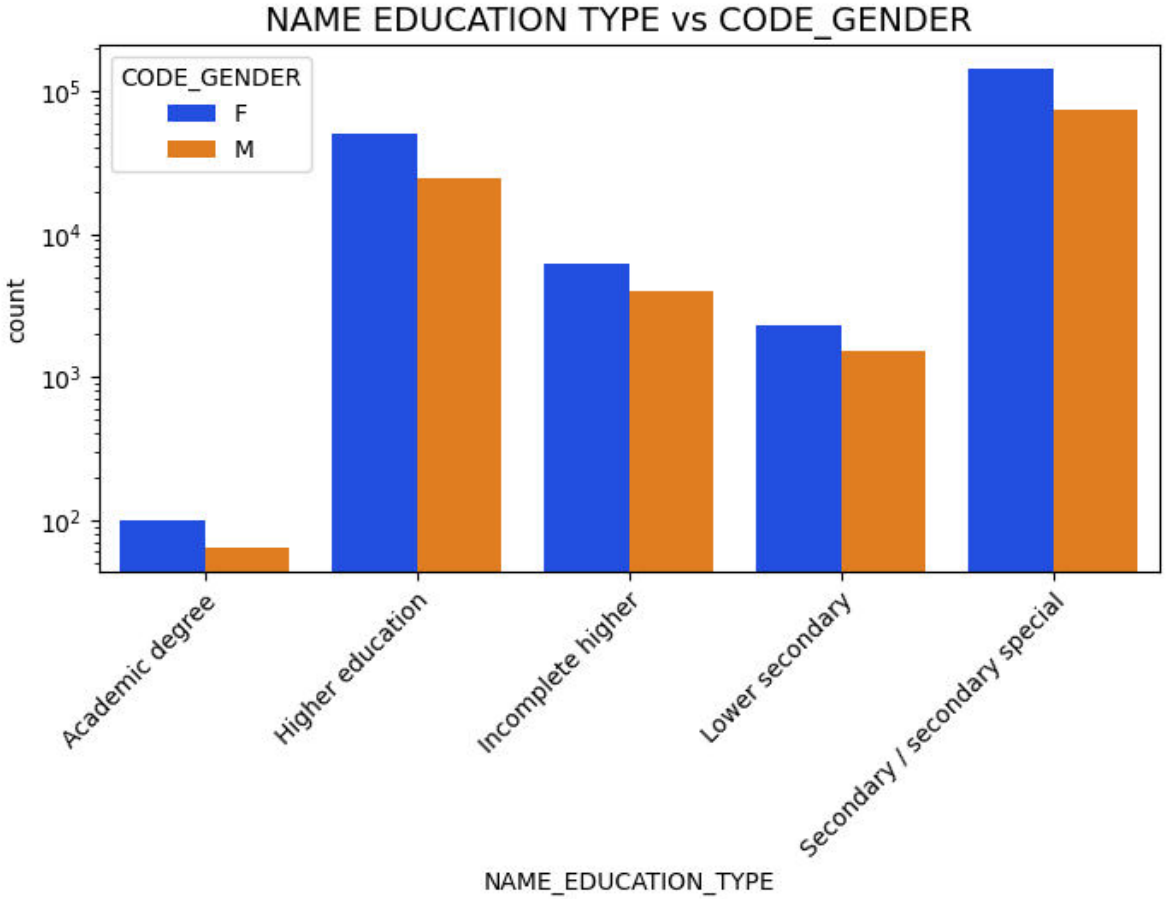
Numeric- Categorical Analysis of NAME_CONTRACT_TYPE variable wrt. both TARGET datasets



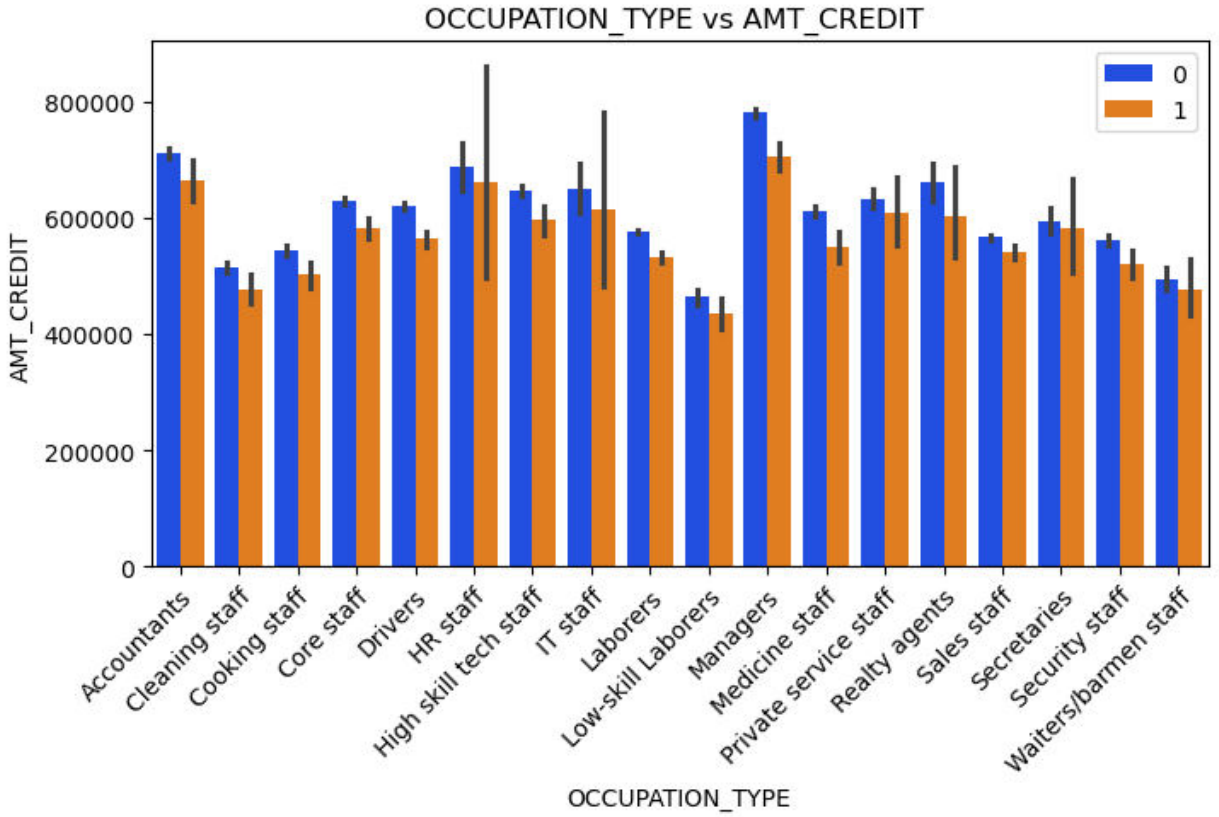


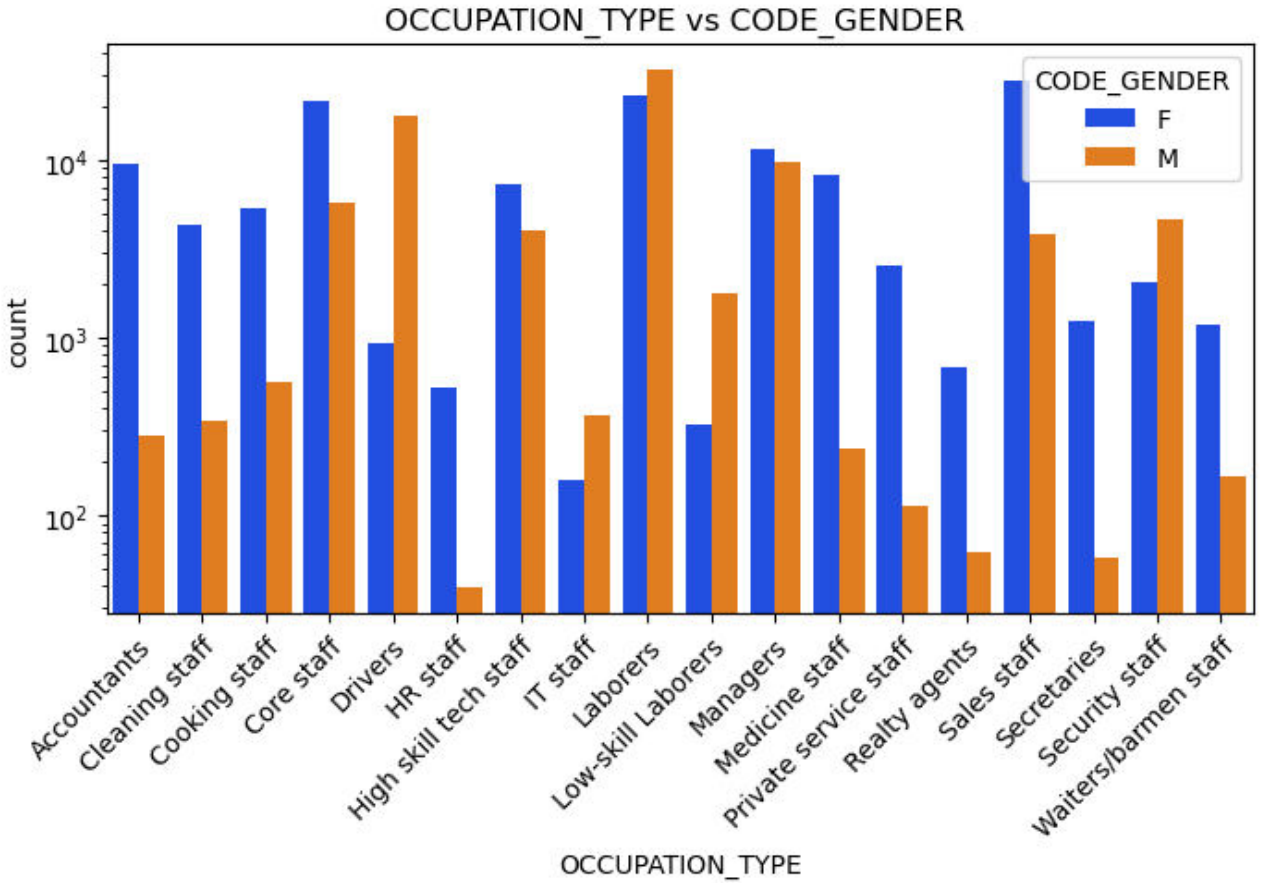
Numeric- Categorical Analysis of NAME_EDUCATION_TYPE variable wrt. both TARGET datasets



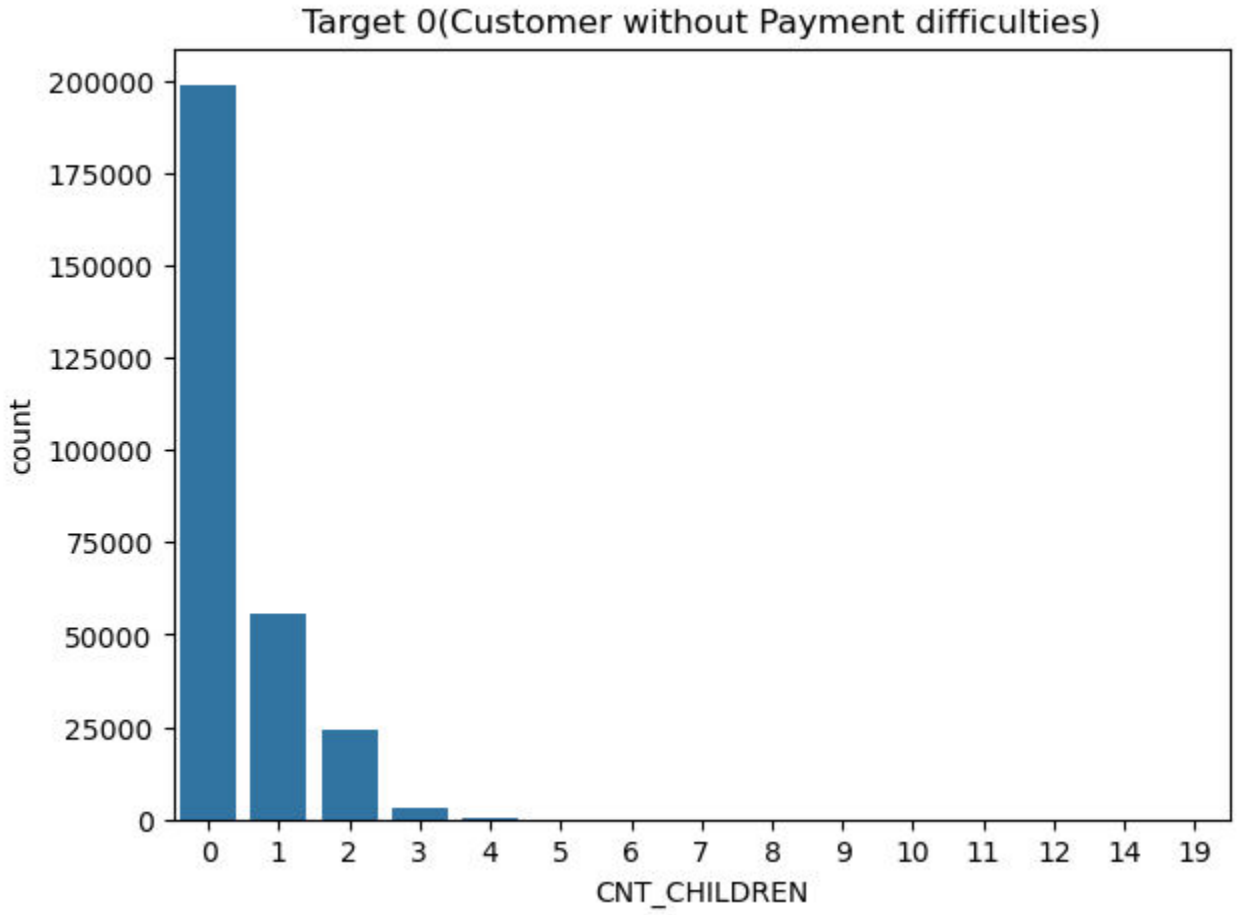


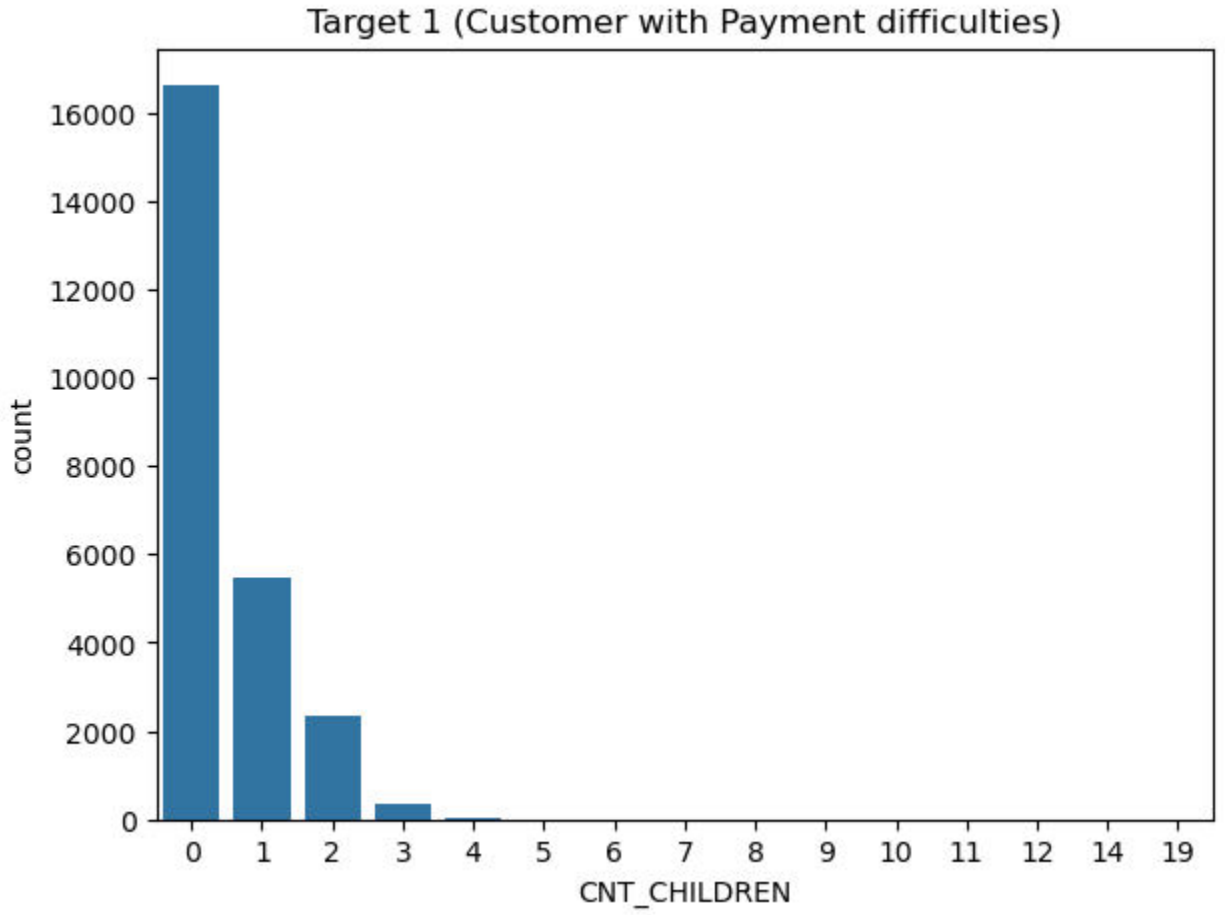
Numeric- Categorical Analysis of OCCUPATION_TYPE variable wrt. both TARGET datasets



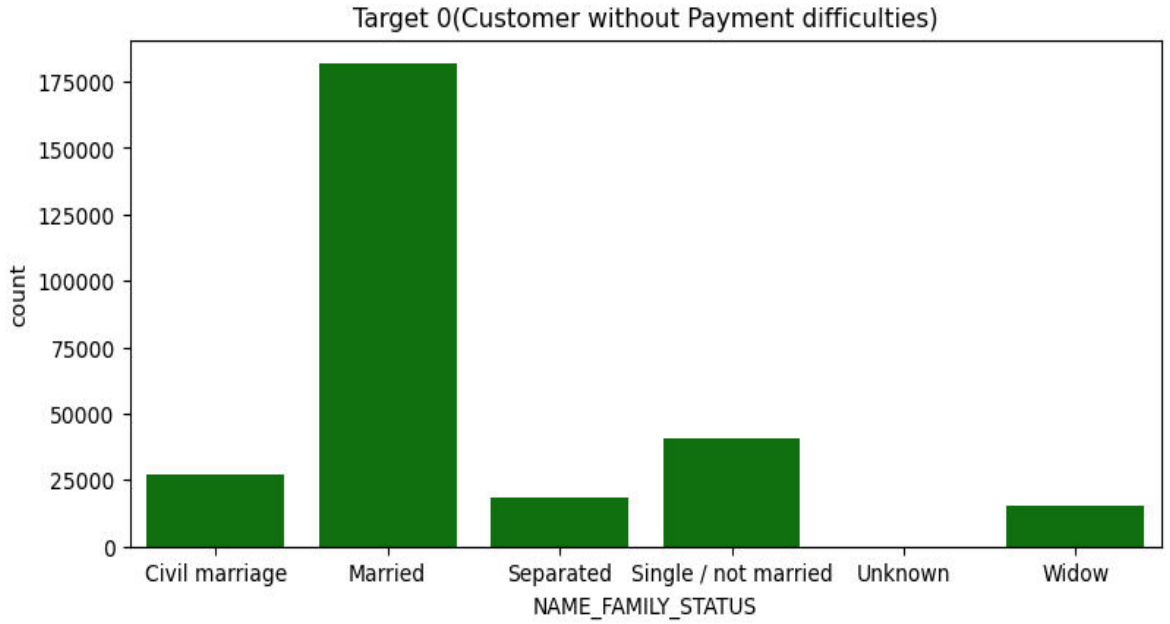


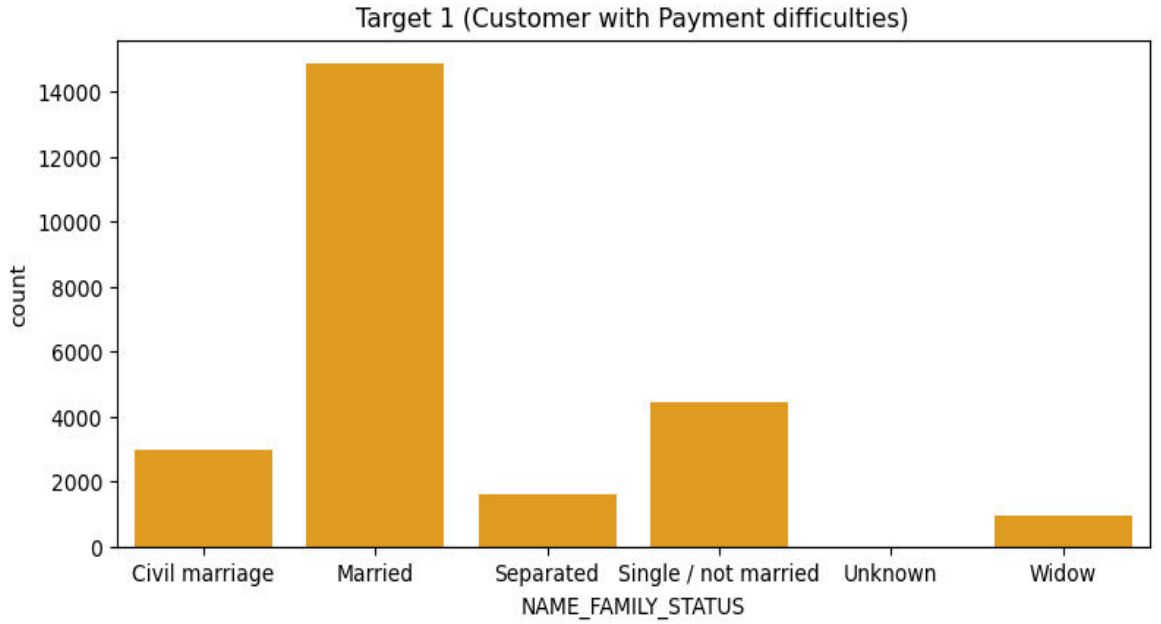
Numeric- Categorical Analysis of CNT_CHILDREN variable wrt. both TARGET datasets



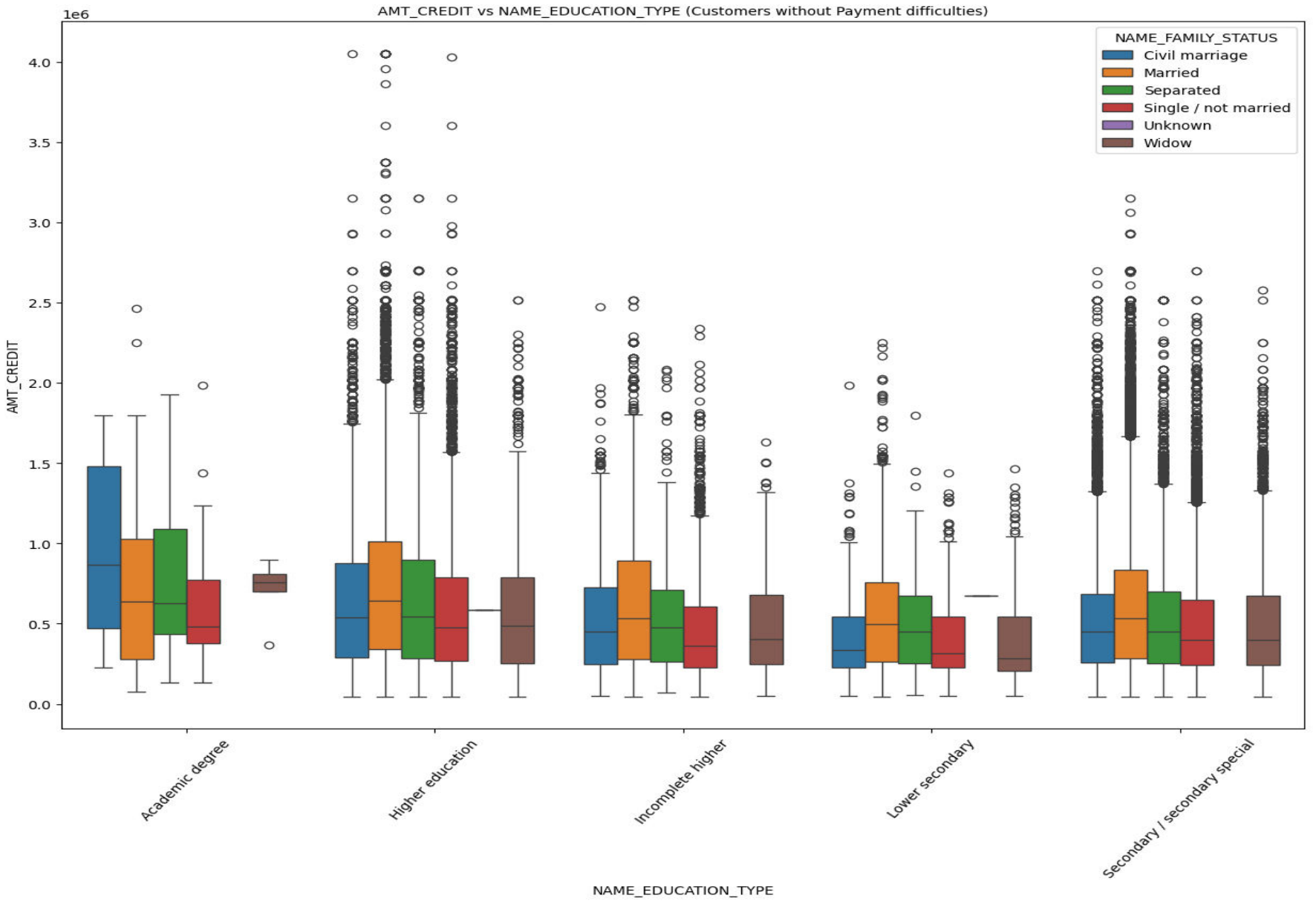


Numeric- Categorical Analysis of NAME_FAMILY_STATUS variable wrt. both TARGET datasets

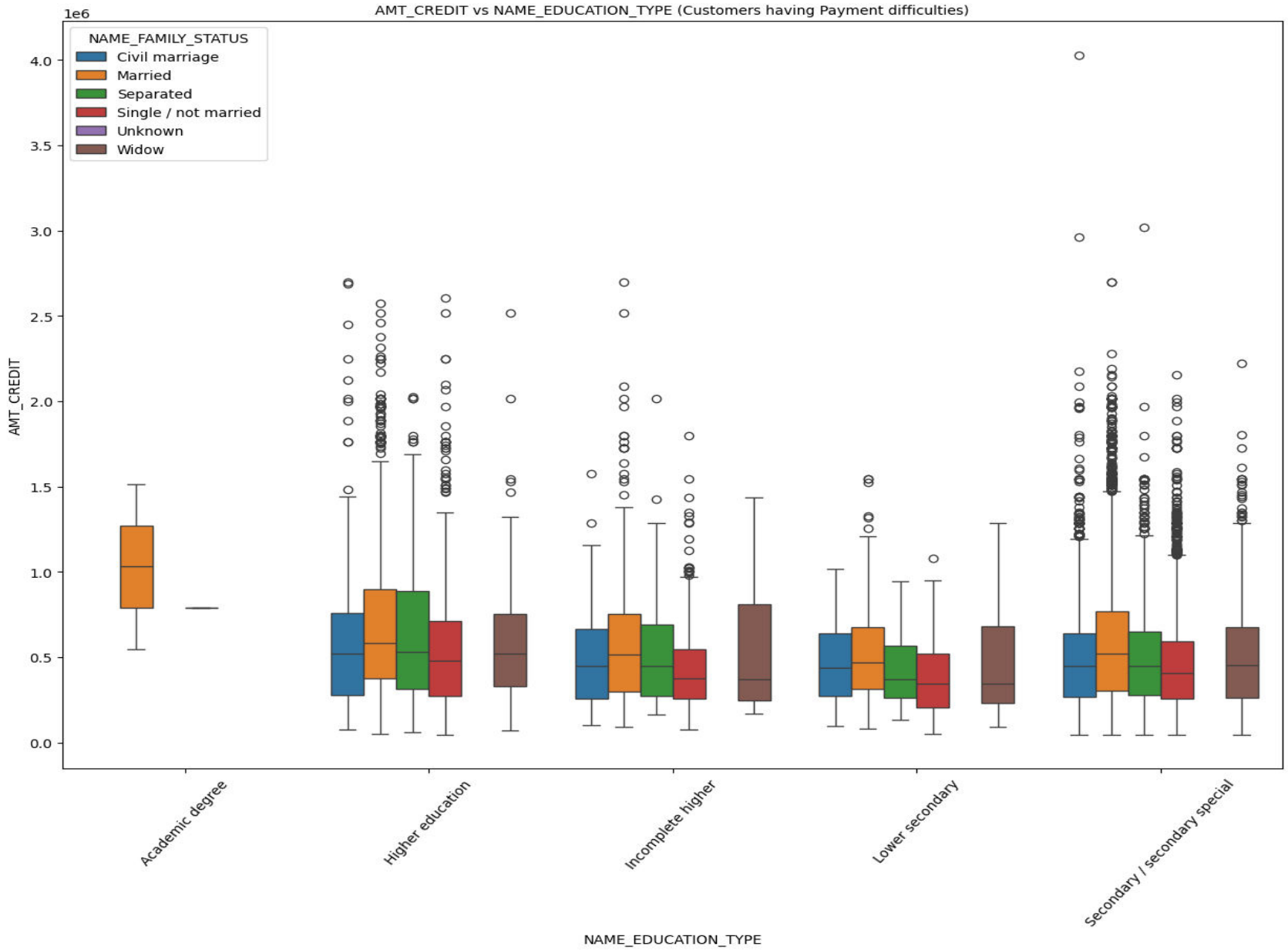




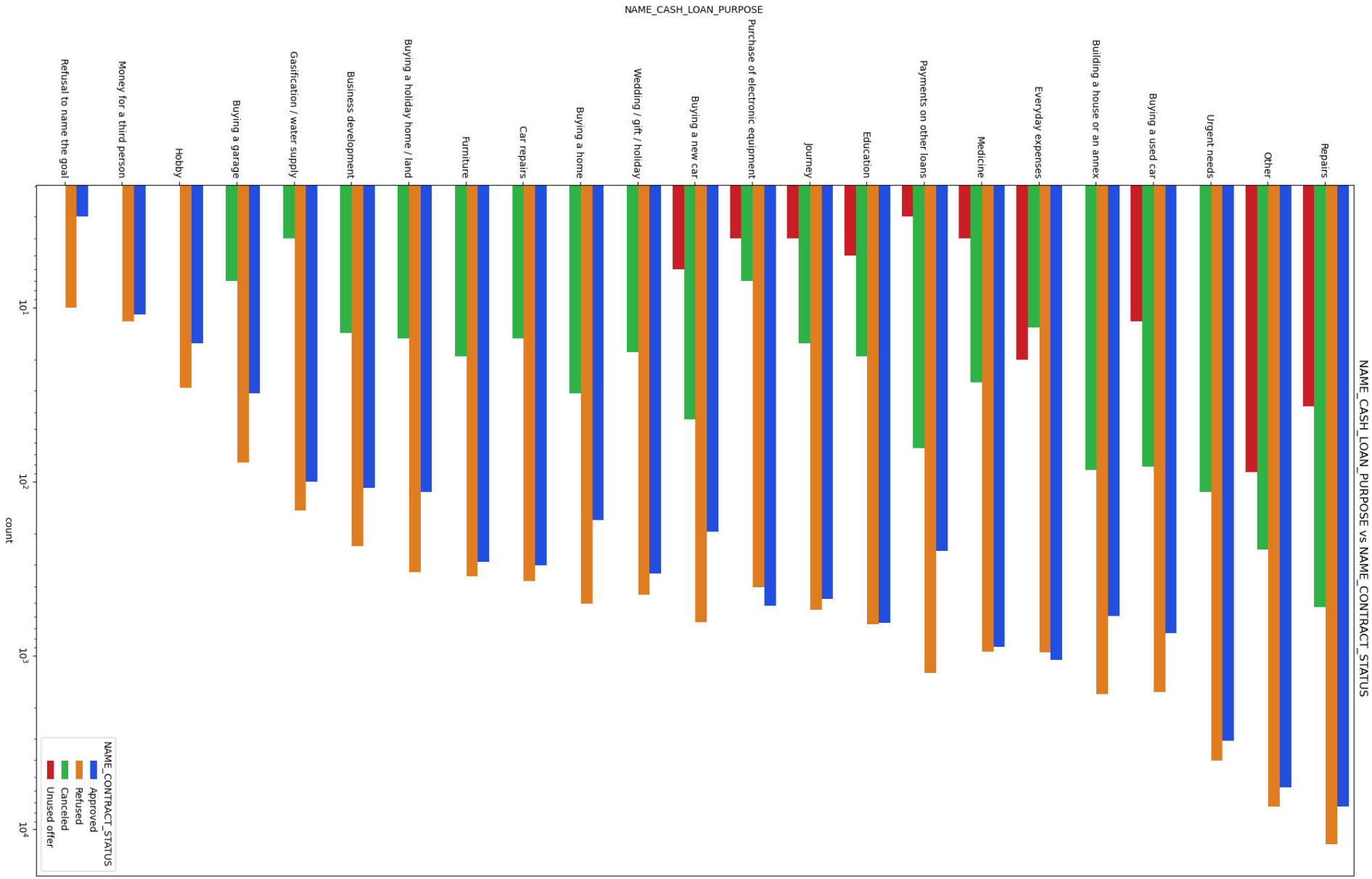
Analysis of AMT_CREDIT vs NAME_EDUCATION_TYPE variable wrt. TARGET – 0



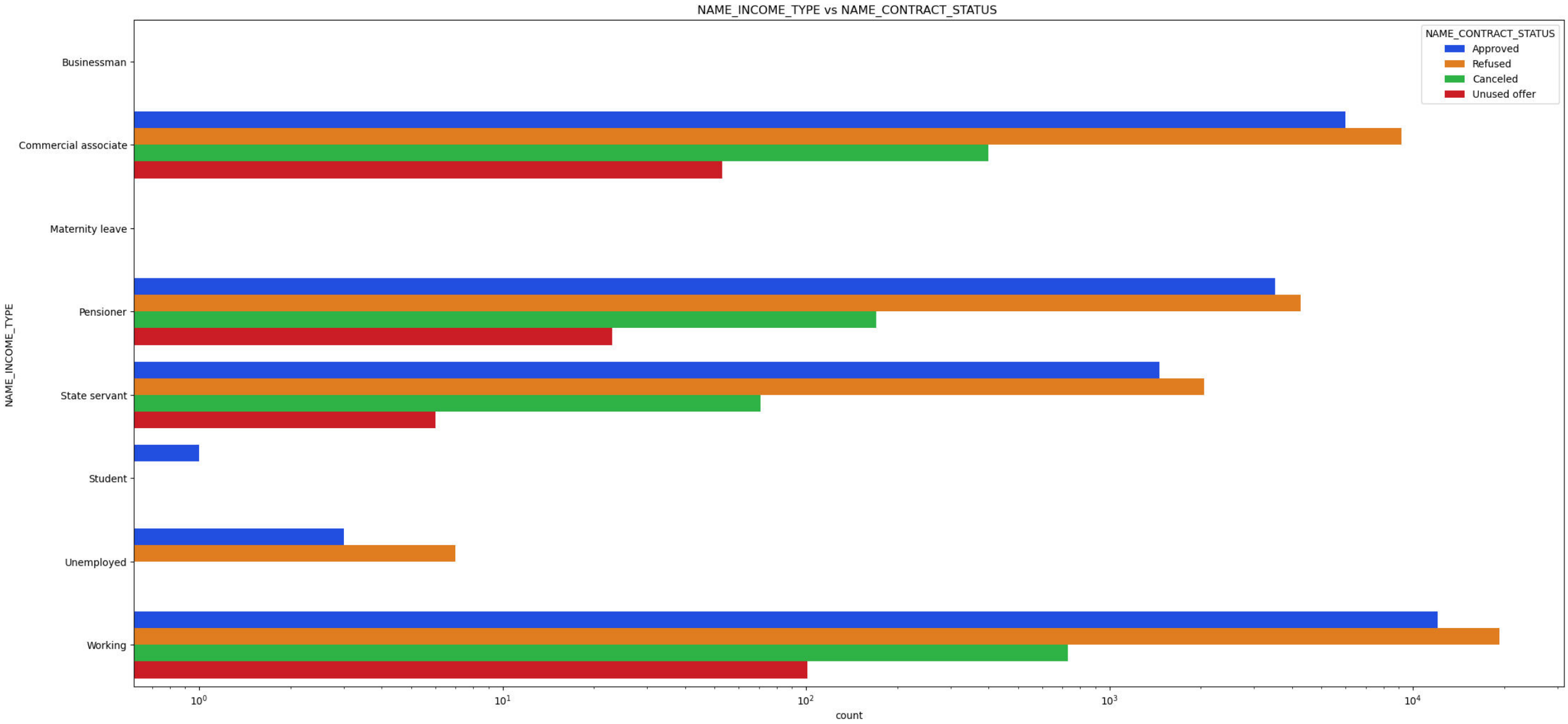
Analysis of AMT_CREDIT vs NAME_EDUCATION_TYPE variable wrt. TARGET – 1



Analysis of NAME_CASH_LOAN_PURPOSE vs NAME_CONTRACT_STATUS variable



Analysis of NAME_INCOME_TYPE vs NAME_CONTRACT_STATUS variable



Conclusion

Following points were concluded after performing Univariate , Bivariate and Multivariate Analysis for ***application data*** and ***previous application*** dataset :

- Female Gender are more likely to take Loans than Male gender. Therefore, females can become best customer for bank with Cash loans type.
- Bank must focus on providing more loans to contract type 'Student' , 'Pensioner'and 'Businessman' with more credit as they have good approval status.
- Loan Payment Risk with Cash loans are more than Revolving loans.
- Bank must avoid giving loans to the Customers having more credit with Maternity Leave
- Only Married Customers with Academic degree are facing difficulties to pay loans.
- Loan purpose ‘Repair’ has higher number of unsuccessful payments on time.
- Income Type 'Working', ' State Servant' and 'Pensioner'customer has higher Rejection than Approval and so having more unsuccessful payments.

THANK YOU