# EDA

Assignment

Sai Srilekha Mutyala

# Business Objectives

- The aim of this case study is to find the patterns which tells if a consumer(client) has any difficulty in paying their installments, so that this information can be used to make decisions like denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This make sures that the clients who can repay the loan are not rejected. Identifying such type of applicants is the main aim of this particular case study

# Analysing the Data

## View the data information

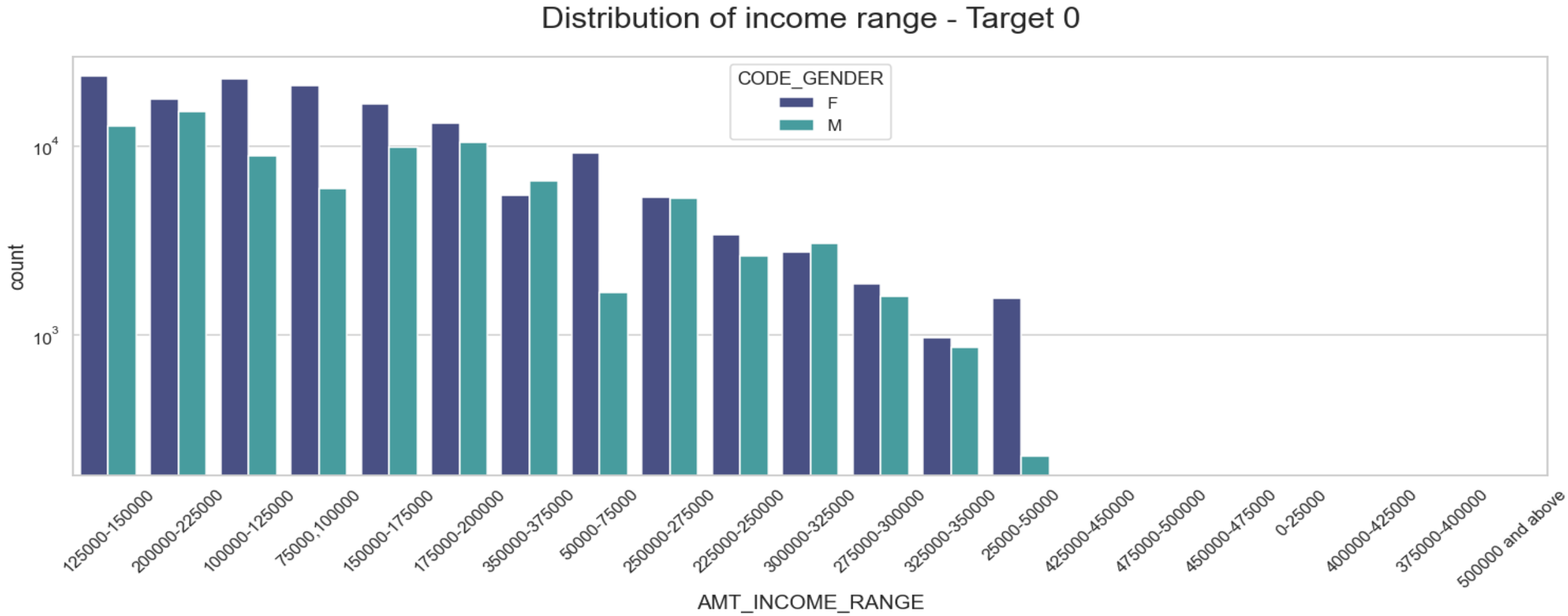
- After we load the data in to the pandas dataframe, we should start looking into the data using info(). As this function gives us list of all column names and number of columns, datatype of each column.
- By doing this we will get a initial insight of the data

# Analysing the Data

## Check the Data Type of the Features

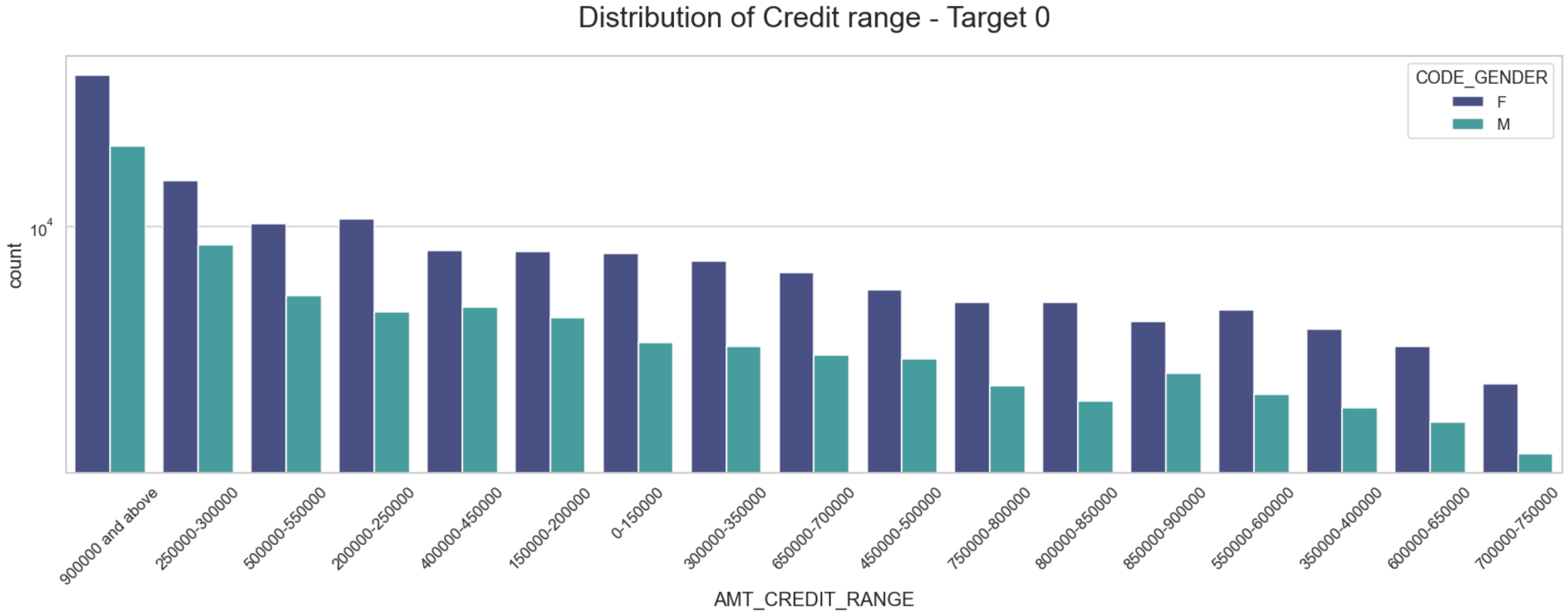
- After we load the data in to the pandas dataframe, we should start looking into the data using info(). As this function gives us list of all column names and number of columns, datatype of each column.
- By doing this we will get a initial insight of the data

## Handle Missing values and Outliers



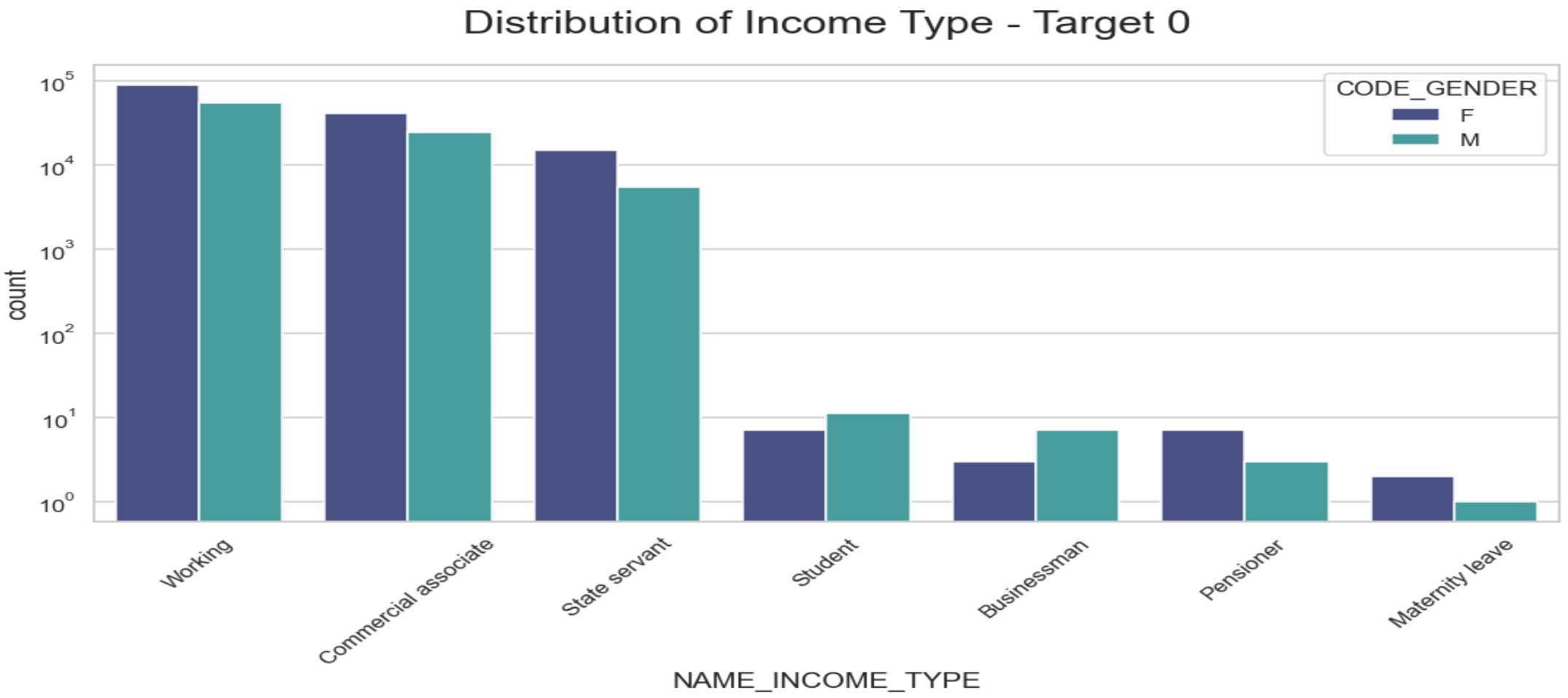

Inferences:

1) Female counts are higher than Male counts.
2) The income in the range of 100000 t0 200000 have more number of Counts.
3) There are no customer for income range greater than 375000.
4) For the income range > 300000, the Male count is greater than Female count.



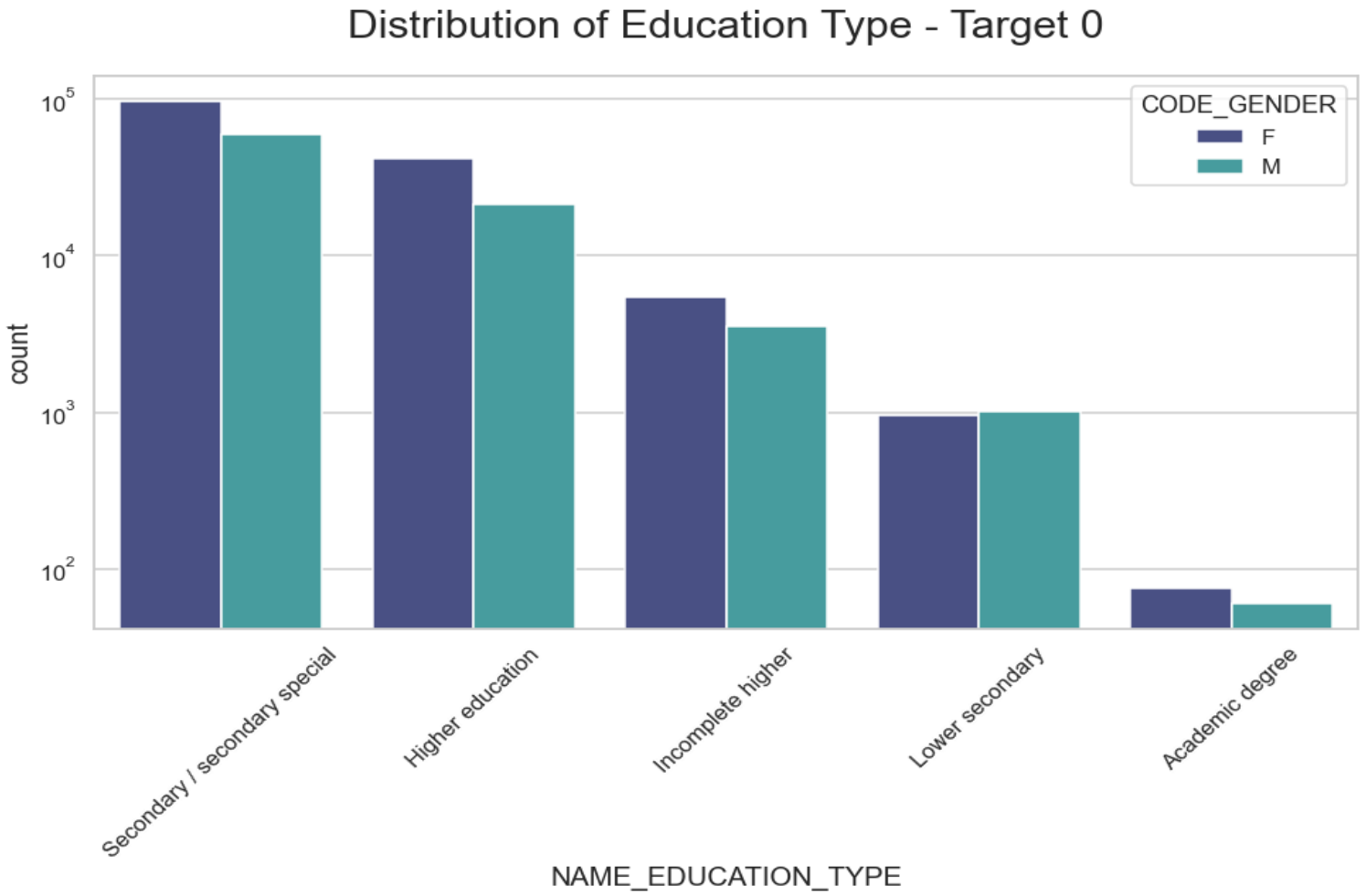

Inferences:

1) Female Counts are higher than Male counts
2) The Credit Range 900000 and above have maximum count



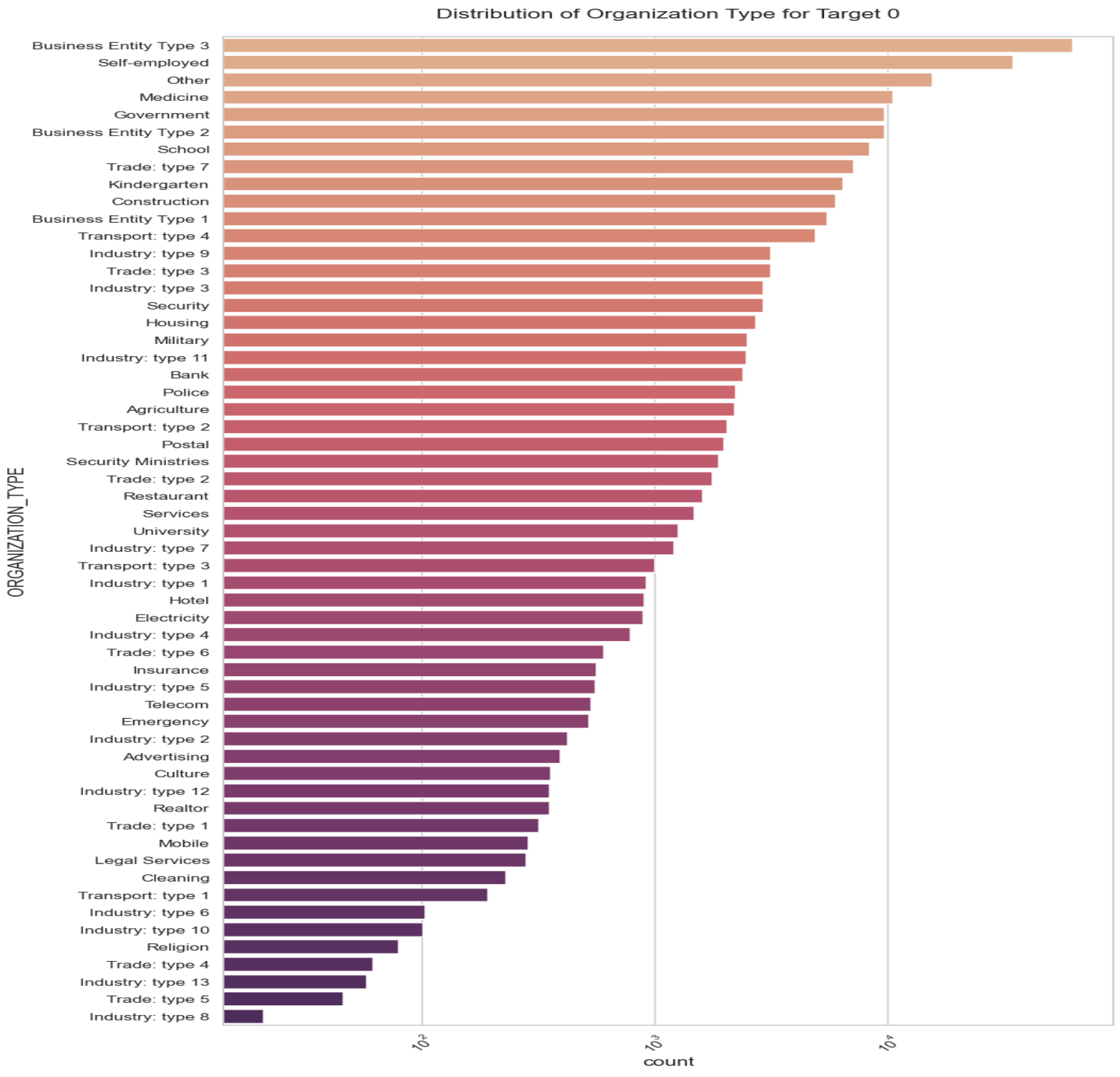

Inferences:

1) The max count is for the IncomeType Working, Commercial associate, StateServant.
2) Females are having maximum counts than males.
3) Business and student category having more male counts than female counts.
3) Less counts are for the income types Working, Commercial associate, StateServant.




Inferences:

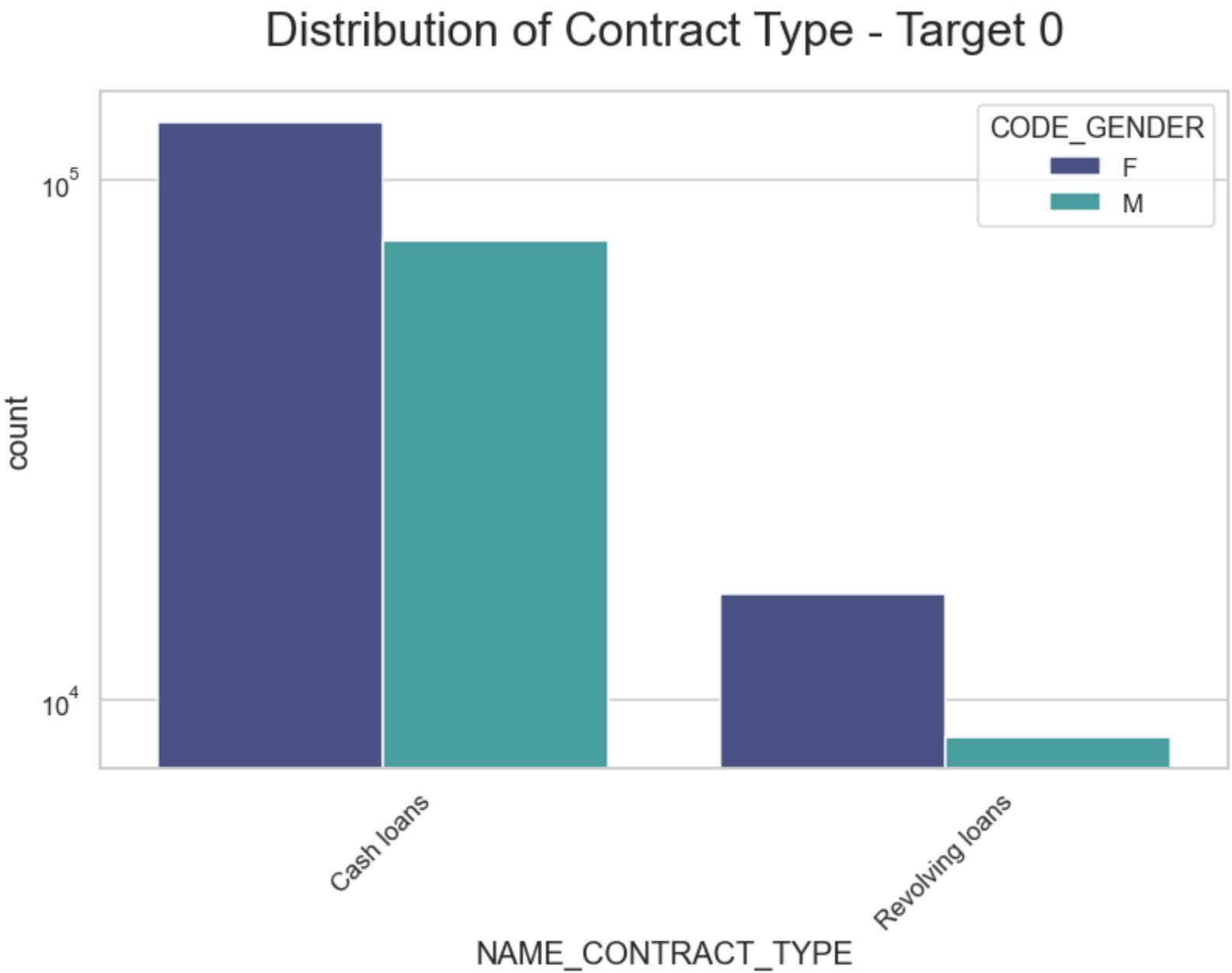
1) The Max count is for Secondary/secondary special Education Type
2) Female count is higher than male count
3) There is very less count for Academic degree




Inferences:

1) The organization type 'Business entity Type 3', 'Self employed', 'Other' , 'Medicine' and 'Government' - the most of the customers are from these Organization Types.

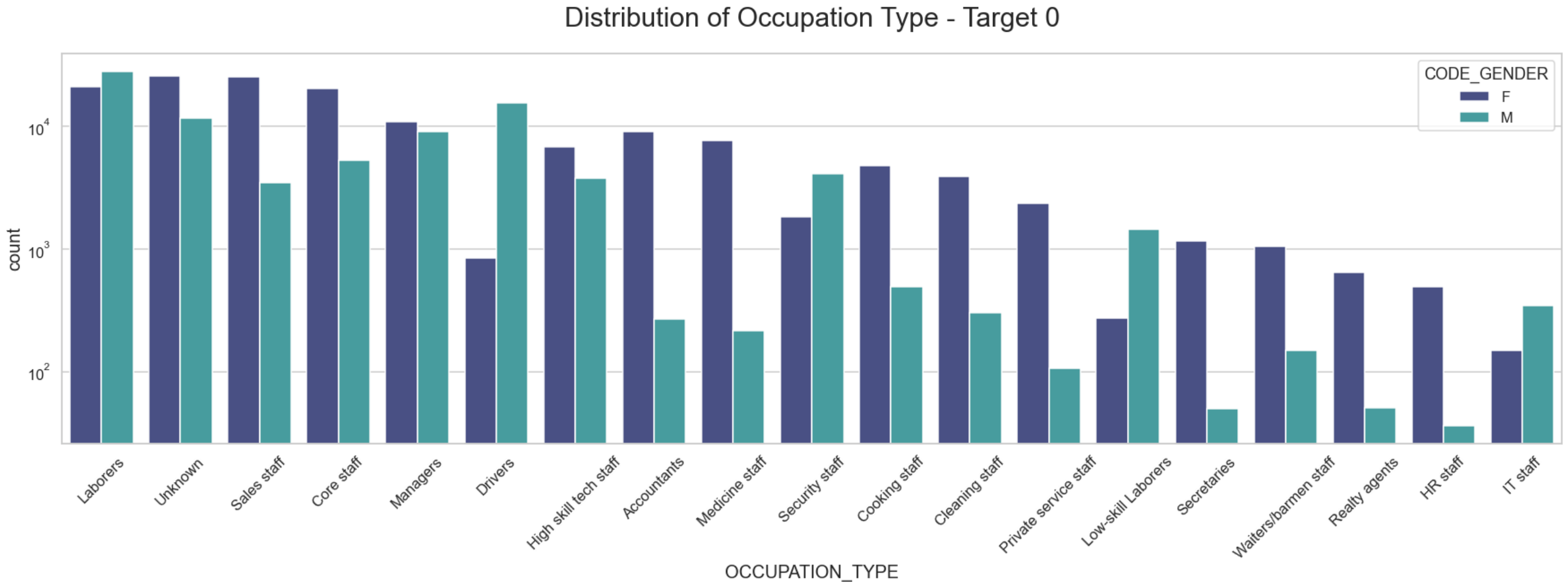
2) Industry type 8,type 13,type 10, 'religion' and trade type 5, type 4 - less customers are from these organization Types.




Inferences:

1)Cash loans have higher count than Revolving loans

2) Female has more count than Male

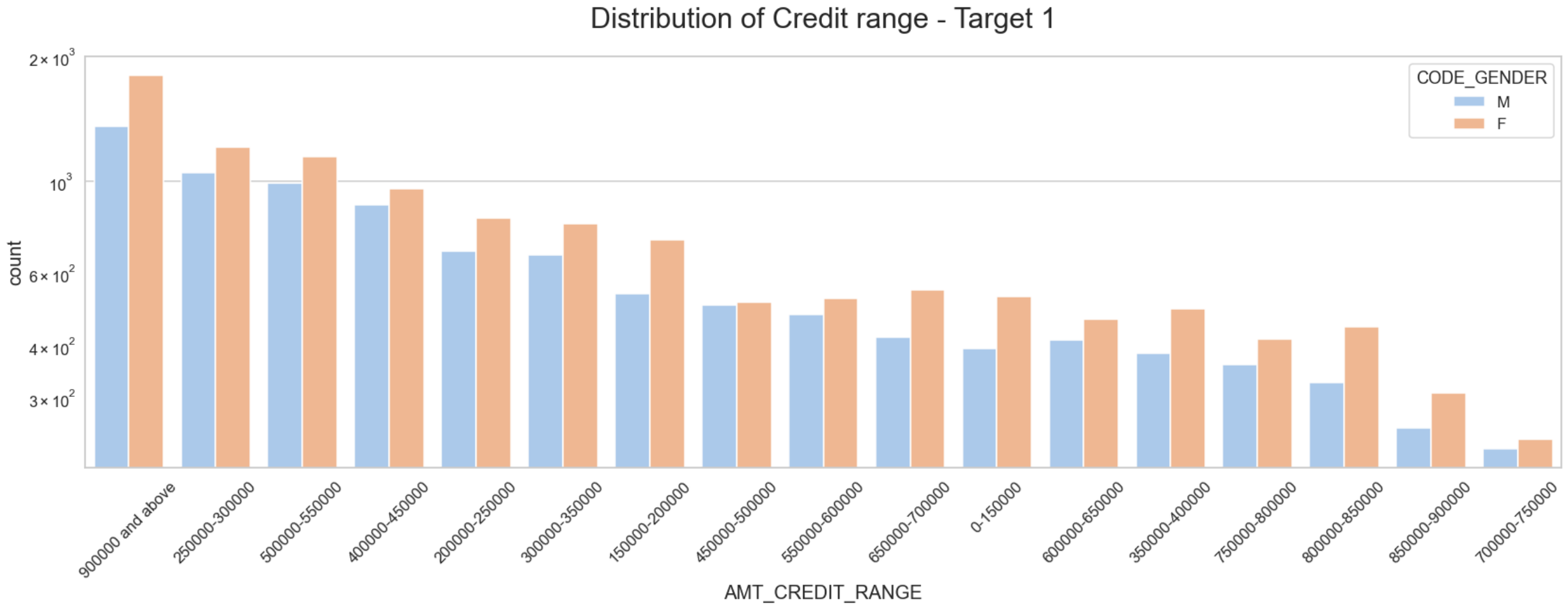



Inferences:

1) The Least count is for IT and HR Staff
2) The max count is for Laborers, Unknown, sales staff, core staff, Mangers...
3) Females count can be seen as high than Males


Inferences:

1) Male counts are higher than females
2) The income in the range of 100000 t0 200000 have more number of Counts
3) There are no customer for income range greater than 375000
4) For the income range < 175000, the Female count is greater than male count.

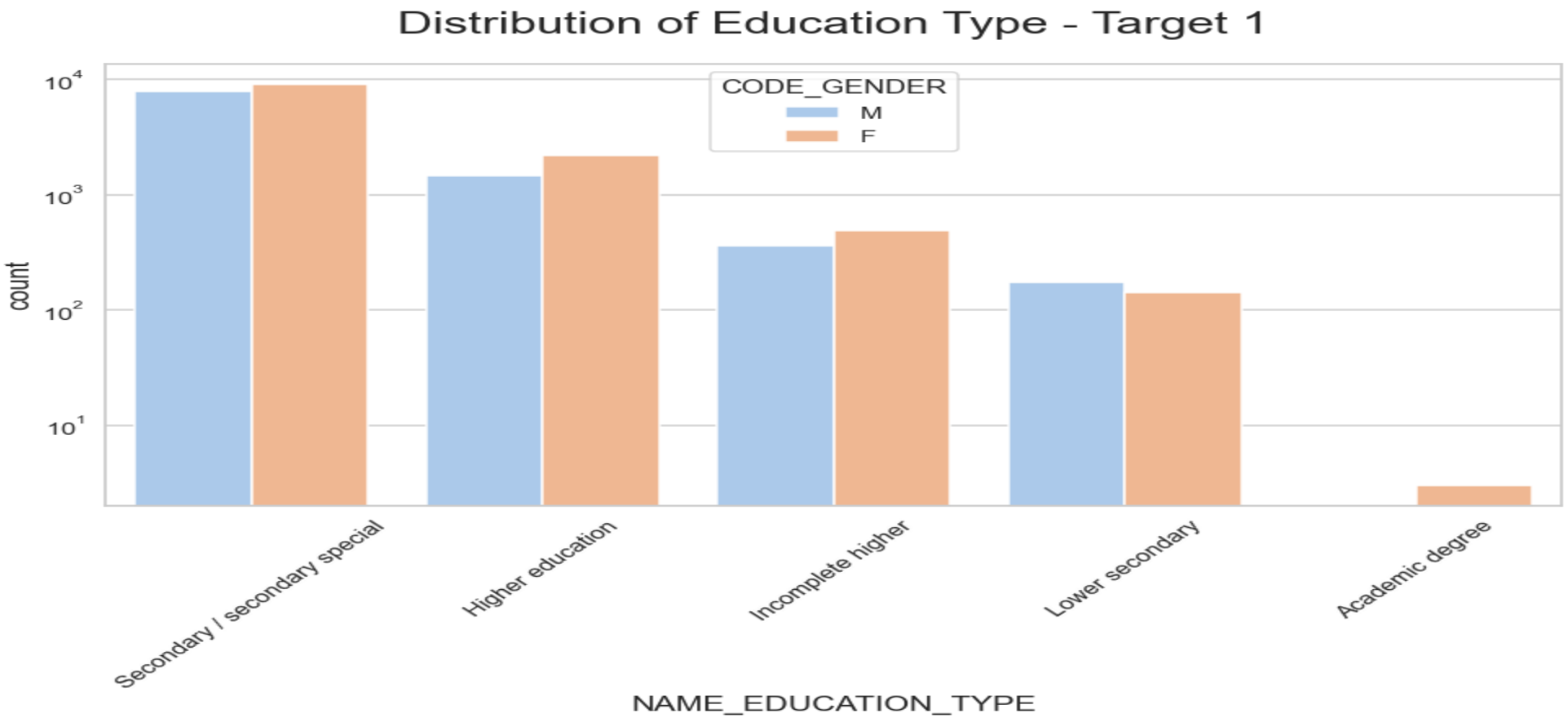



Inferences:

1) Female count is more than male count


Inferences:

1) The income type 'working', 'commercial associate', and 'State Servant' have more count than 'Maternity leave'
2) Here female count is greater than male
3) we don’t have income type Student, Business man, Pensioner in Target=1 data, so there is no late payments from them.
4) less number of counts for Maternity leave



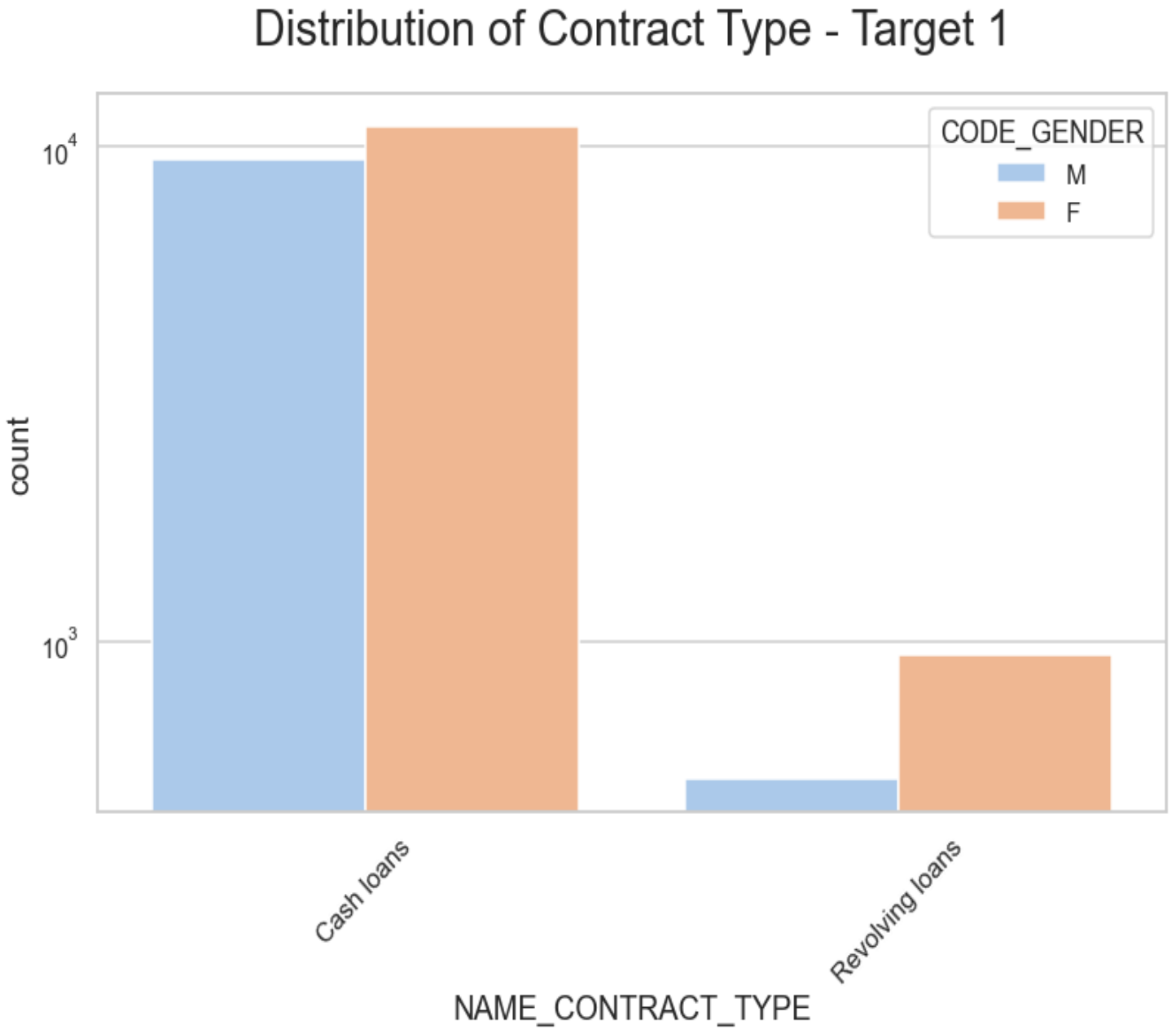

Inferences:

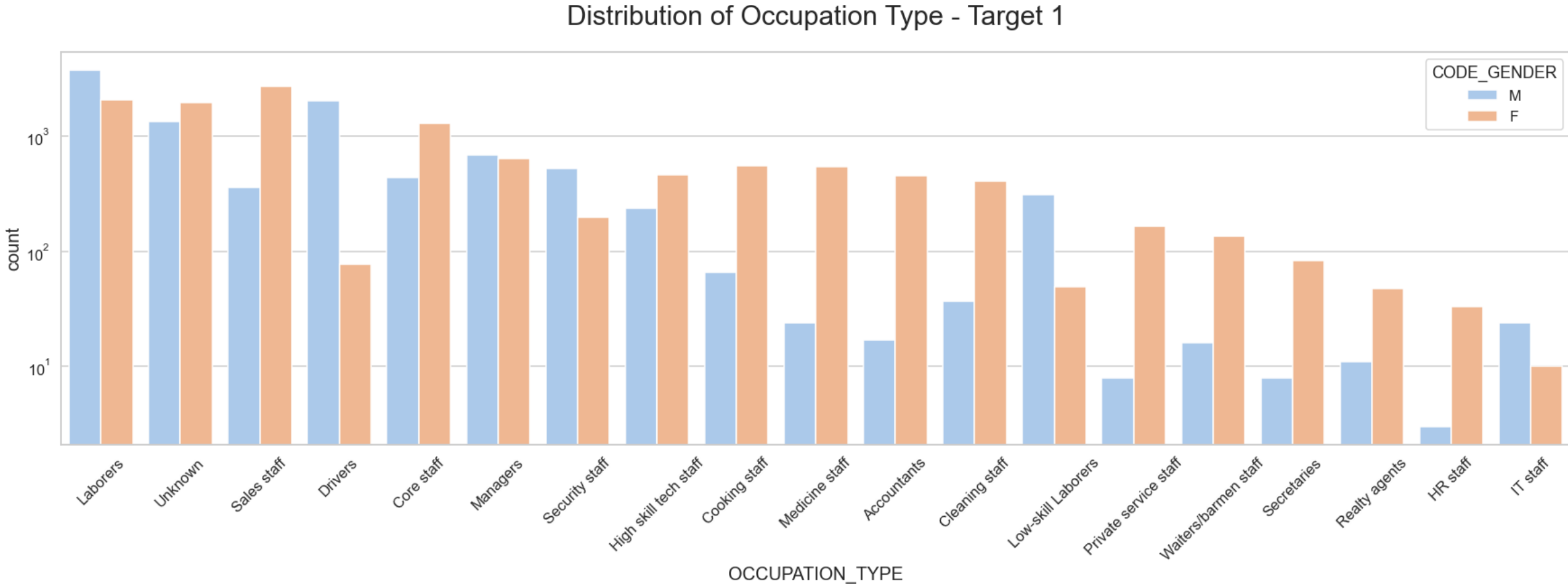
1) Female count is higher than Male count
2) Less count for Academic degree
3) Secondary/Secondary special Education type has mare counts

Inferences:

1) Cash Loans are in Higher amounts than Revolving loans

2) Female count is more than Male count






Inferences:

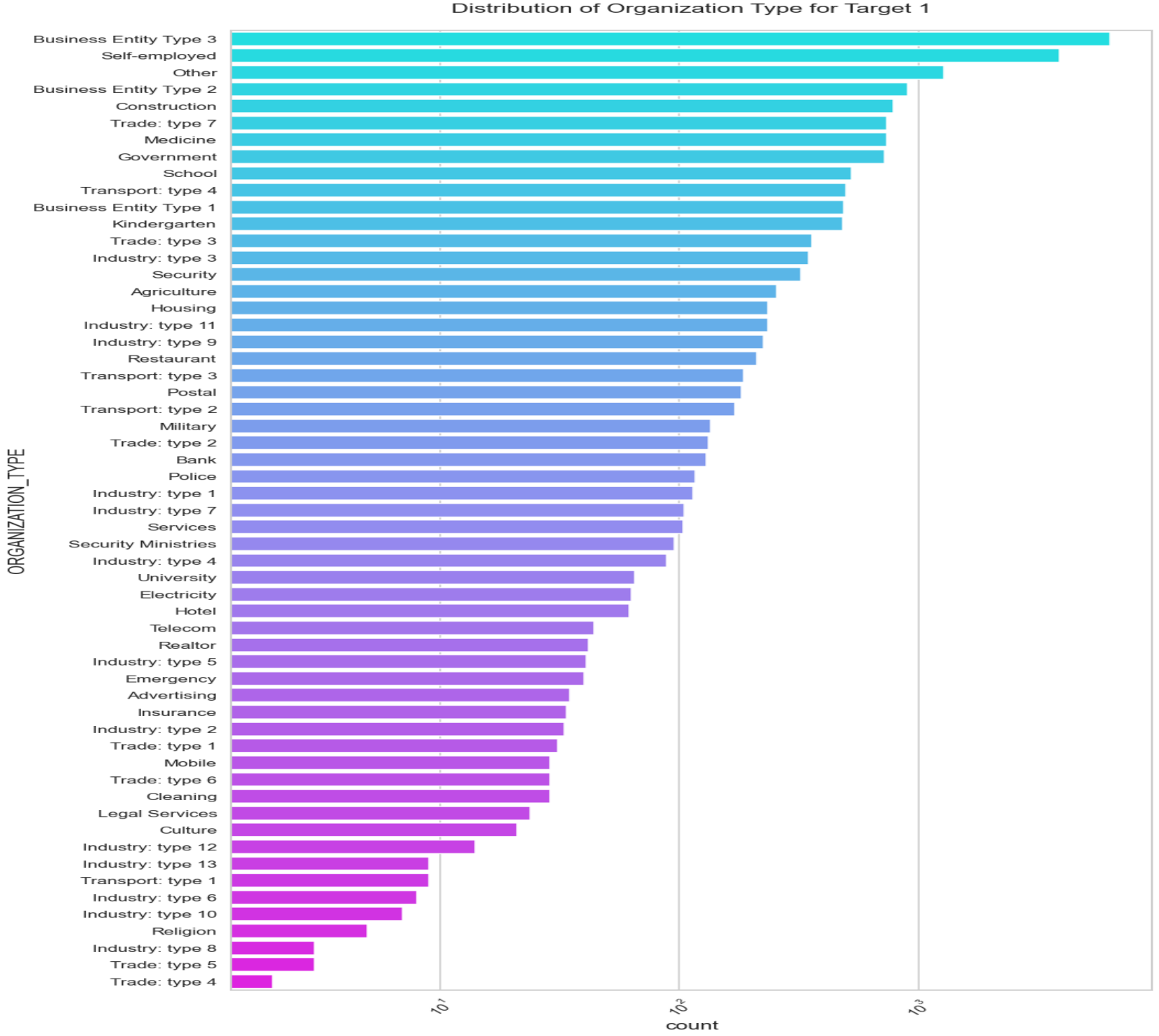
1) Male count is more for laborers, Drivers, Security staff, Lowskill laborers, IT staff

2) Less count for IT, HR staff, Realty agents

Inferences:

1) The organization type 'Business entity Type 3', 'Self employed', 'Other' , ‘Construction' and Targe – Type 7' - the most of the customers are from these Organization Types.

2) Industry type 8,type 6,type 10, 'religion' and trade type 5, type 4 - less customers are from these organization Types.



Inferences:

1)IncomeAmount and children count has negative corelation it means, both are inversly proportional to each other i.e. the more the children, the less the income and viceversa

2) CreditAmount and children count has negative corelation it means, both are inversly proportional to each other i.e. the more the children, the less the credit amount and viceversa

3) Children count and RegionpopulationRelative has negative corelation it means, both are inversly proportional to each other i.e. the more the densly populated are, the min the number of children and viceversa

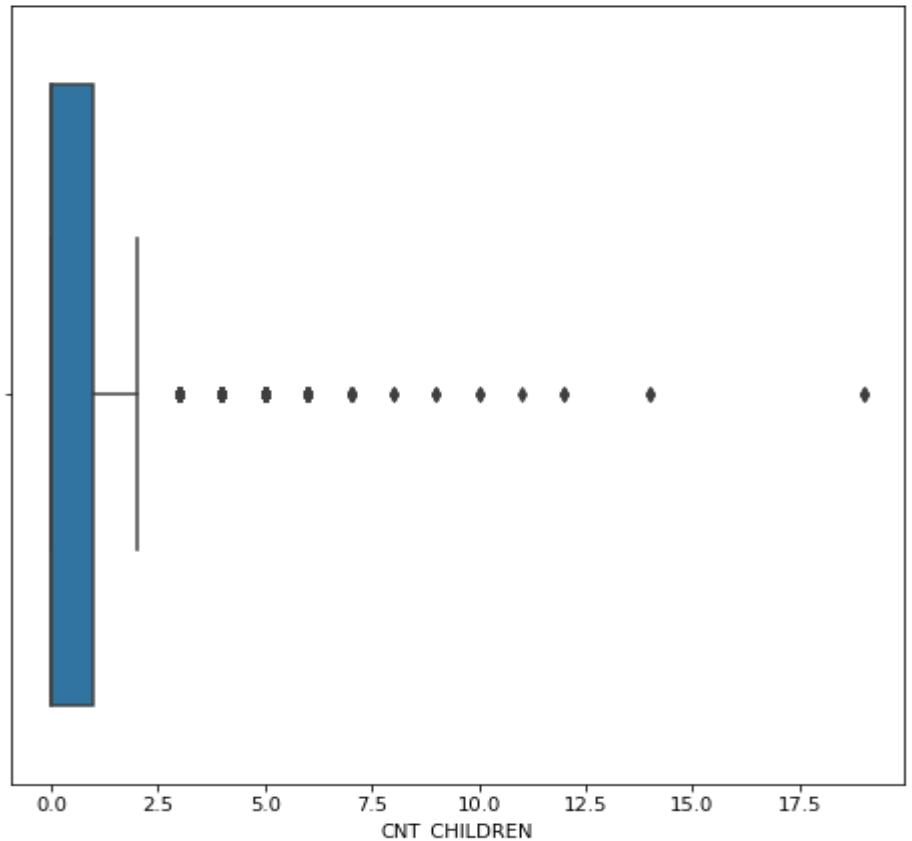
4) The more the children count, the more than count of family members

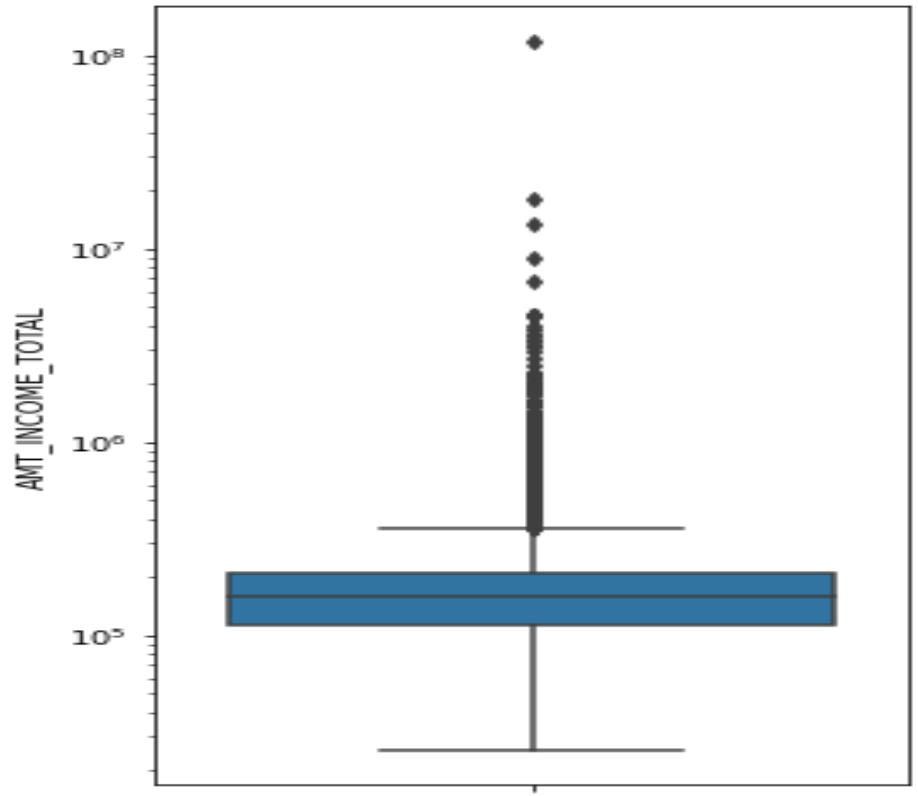
5) The income is higher in the densly populated area(corelation is 0.031628)


Inferences:

The Heap map for Target1 has almost similar observations as we can see in Heat Map for Target 0

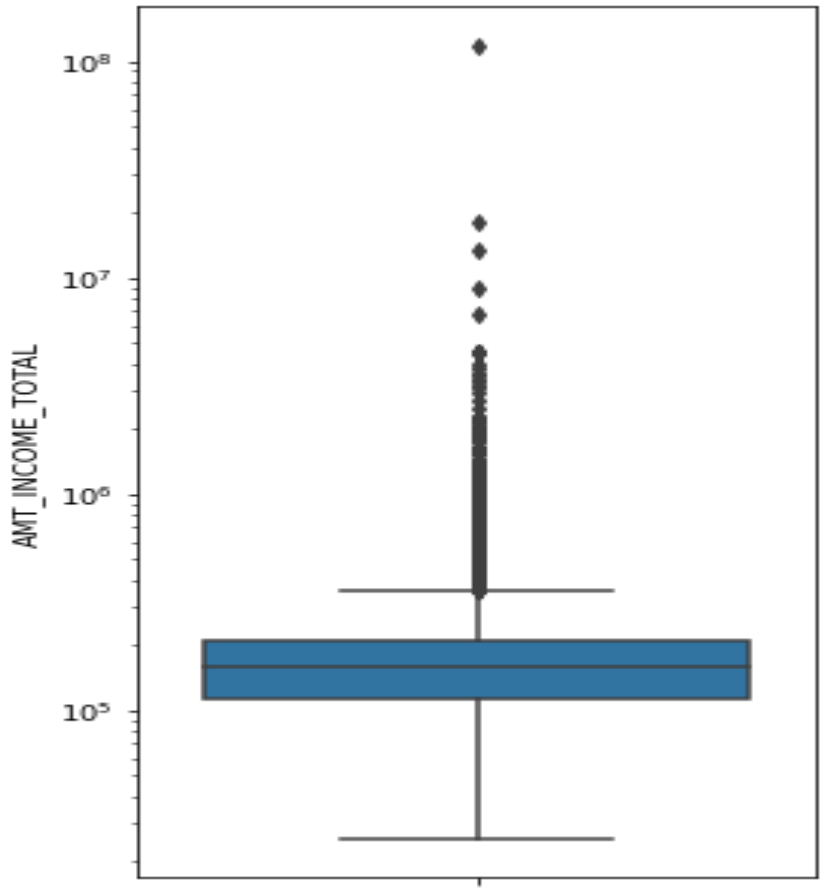
Univariate Analysis for Numerical Variables – Outliers Check

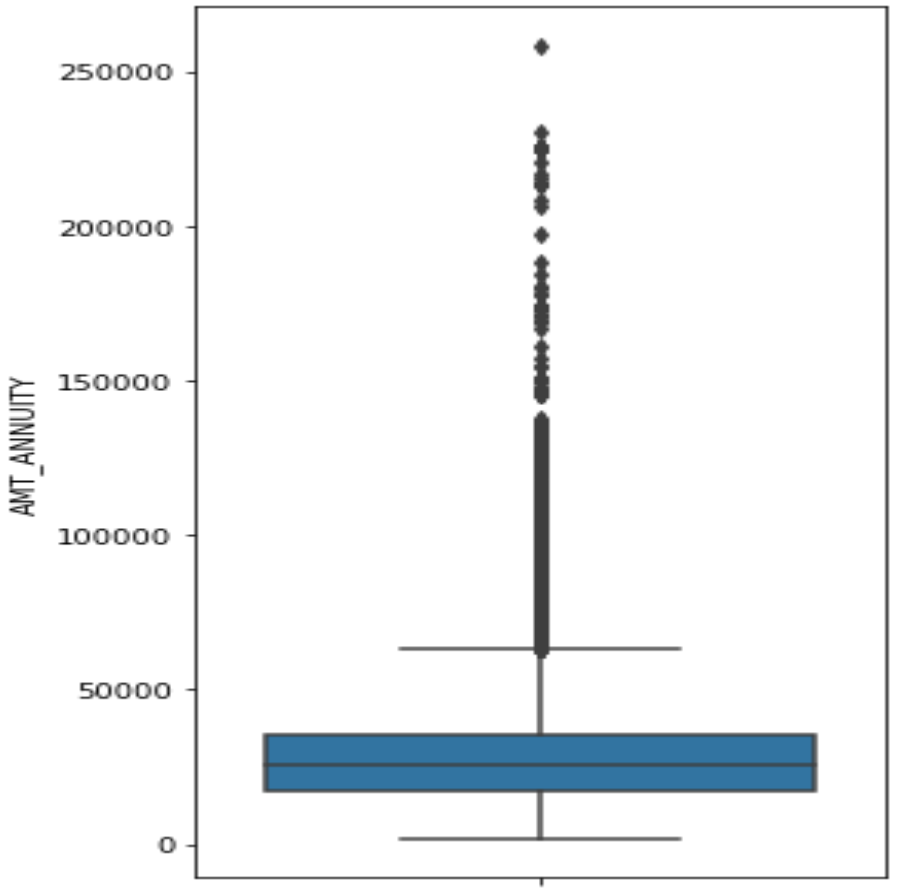



Inferences:

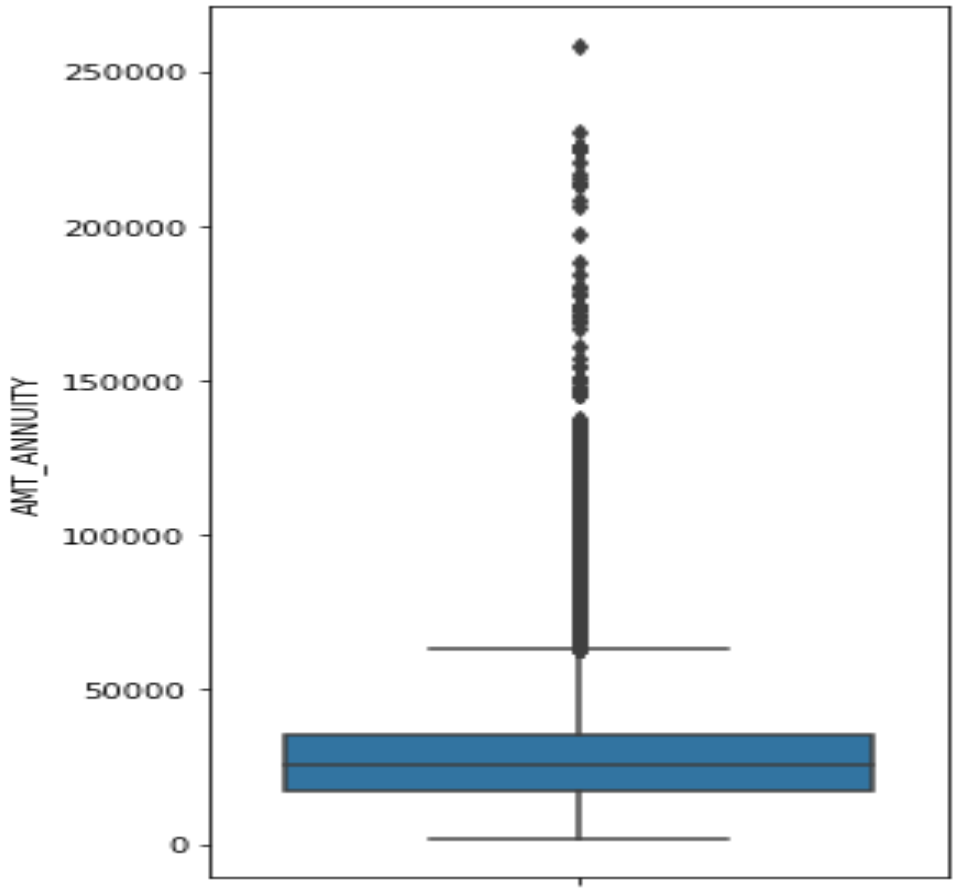
Outliers where present and handled by using capping

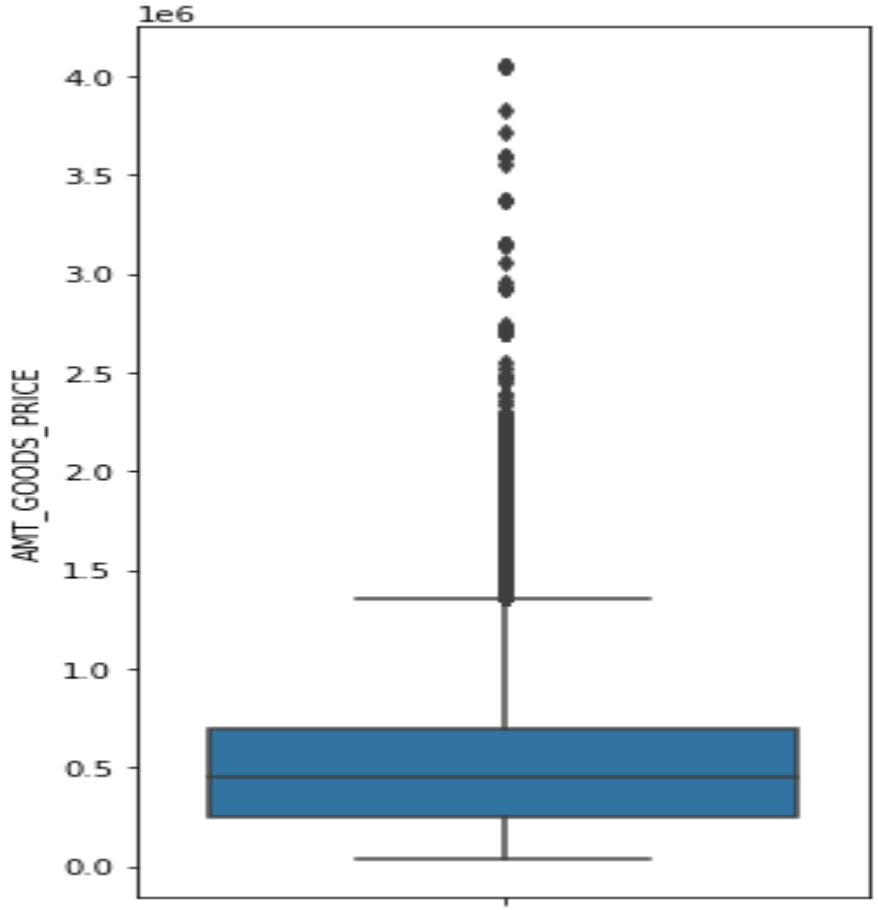



Inferences:

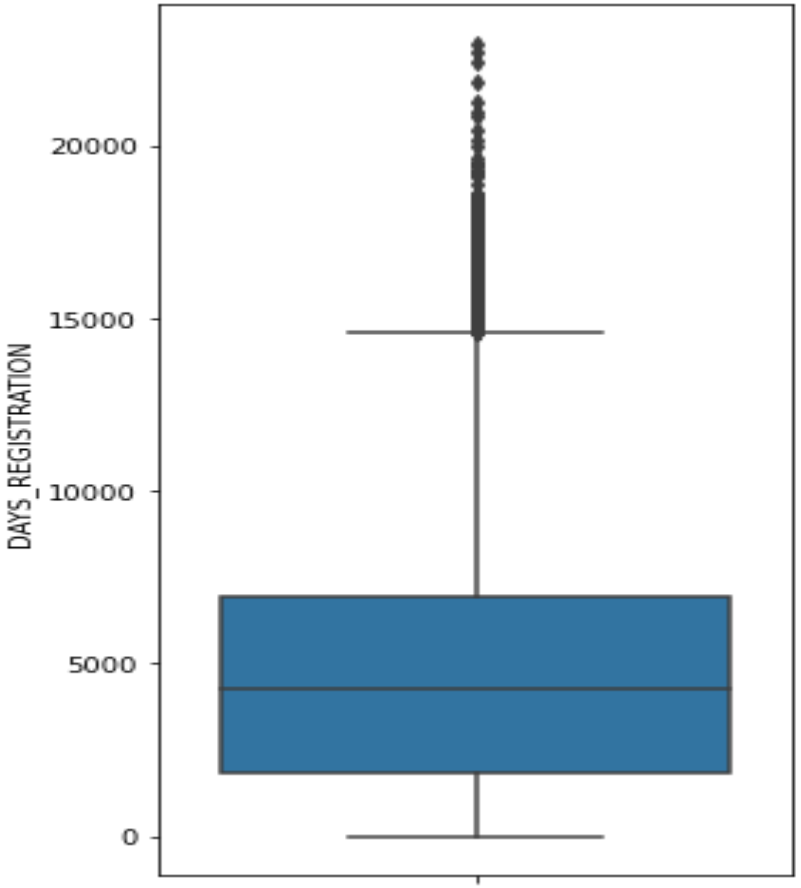
Outliers where present And Handled by using Capping.



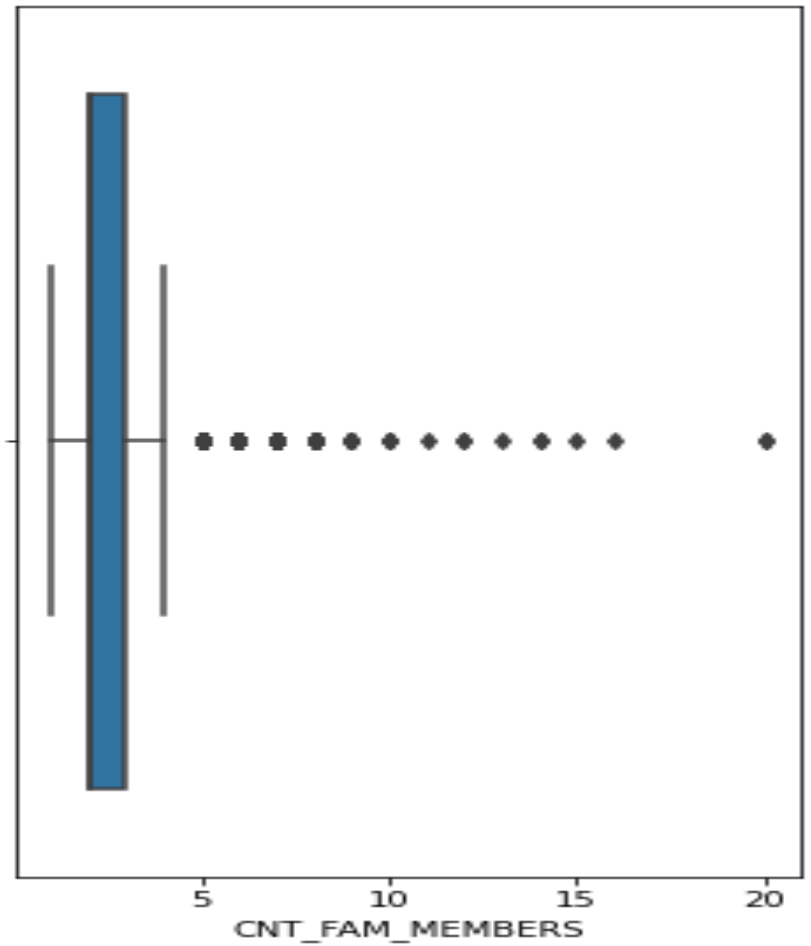

Inferences:

Outliers where present And Handled by using Capping.



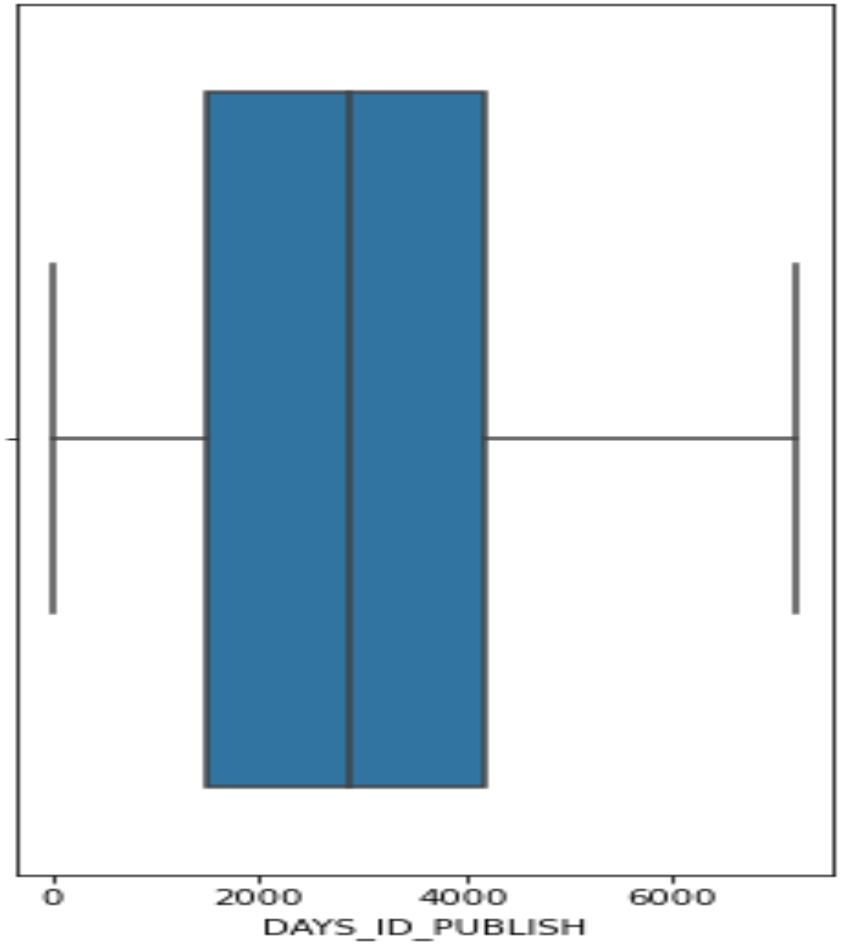

Inferences:

Outliers where present And
Handled by using Capping.



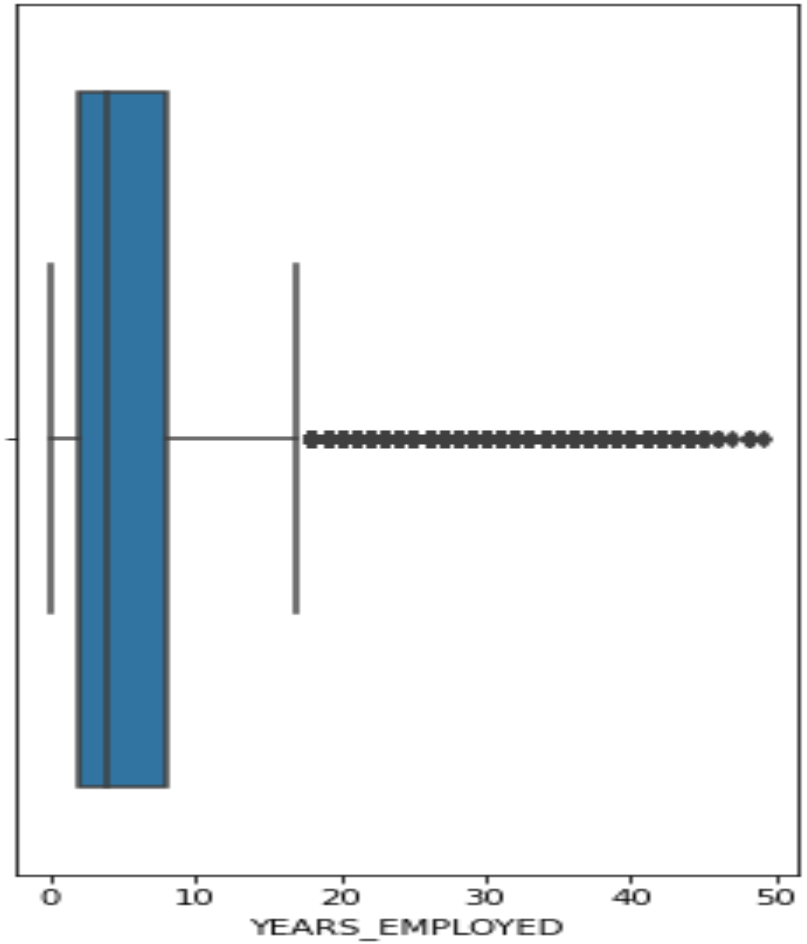

Inferences:

Outliers where present And
These not handled because there can more family members too




Inferences:

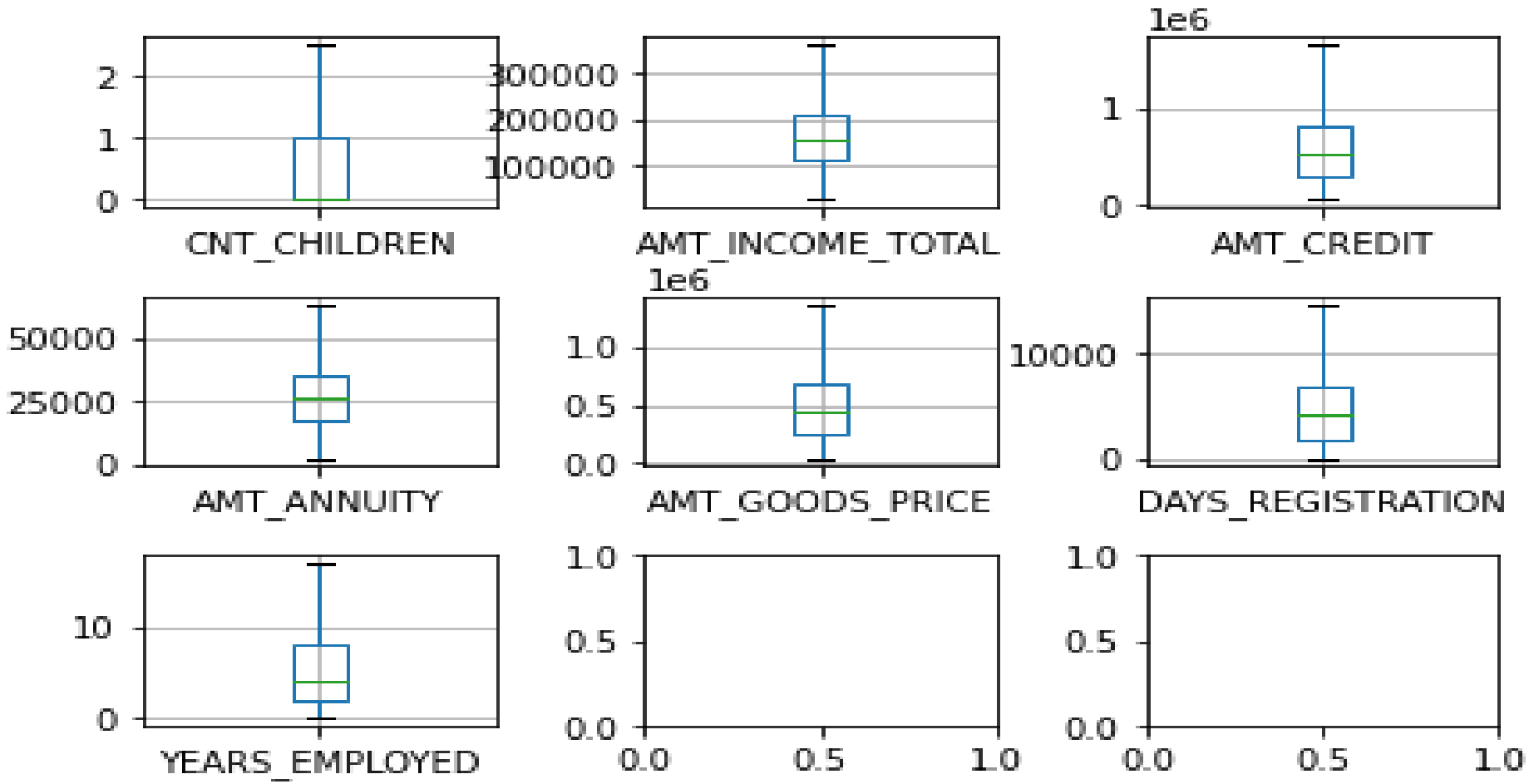
No Outliers



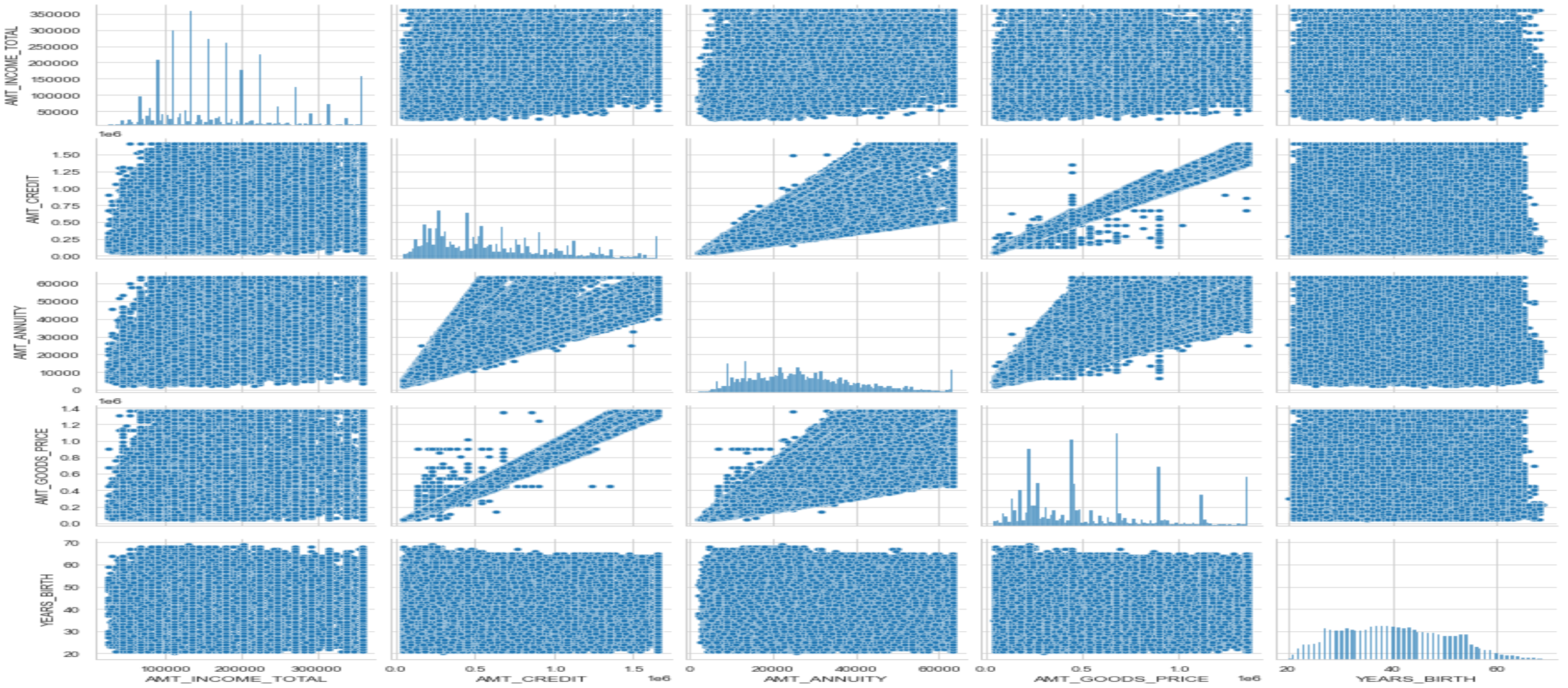

Inferences:

Outliers where present And
Handled by using Capping.

After Handling Outliers



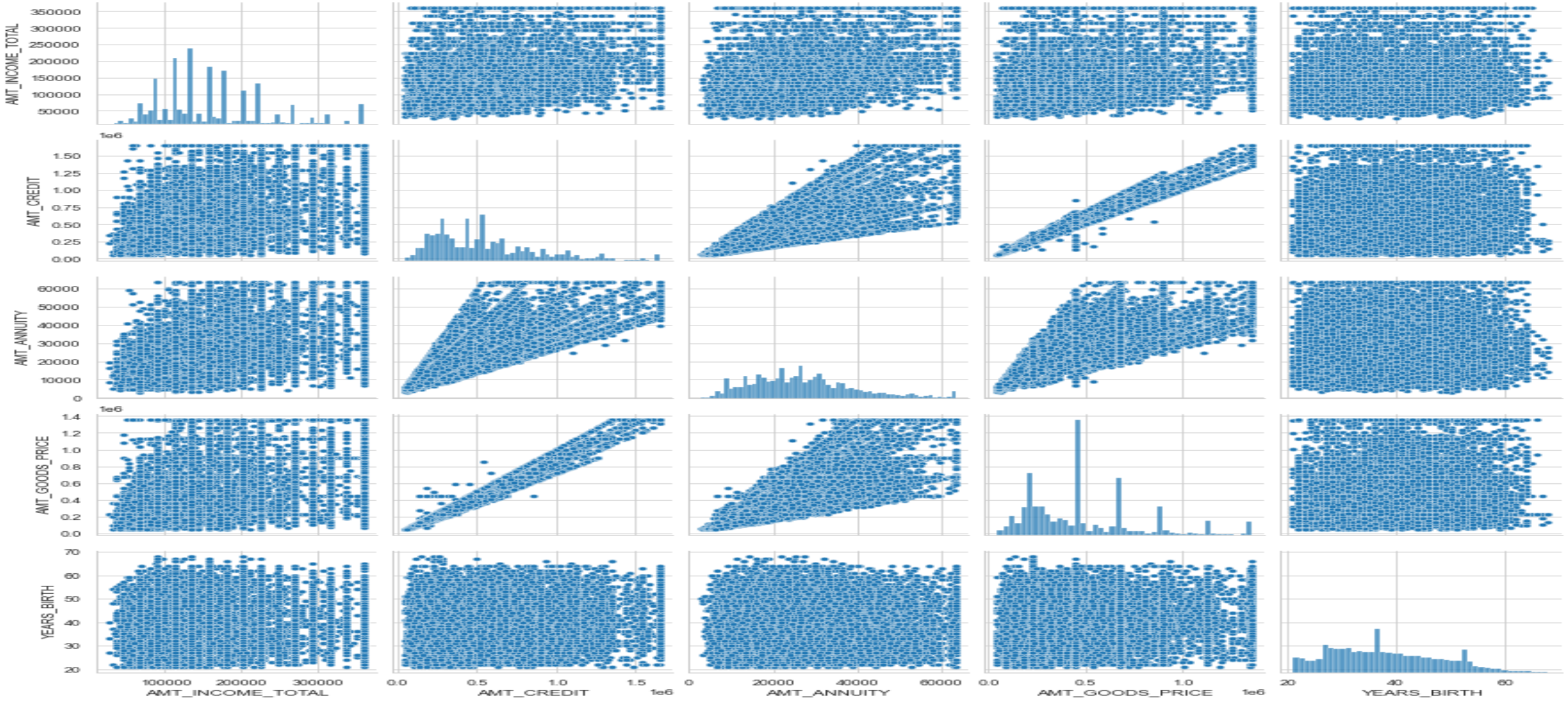
# Numerical vs Numerical - pair plot of all numerical columns that mentioned below # pair plot for target 0




Inferences:

1) AMT_GOODS_PRICE vs AMT_CREDIT , AMT_GOODS_PRICE vs AMT_ANNUITY, AMT_CREDIT vs AMT_ANNUITY - Linear Corelation present

# Numerical vs Numerical - pair plot of all numerical columns that mentioned below # pair plot for target 1

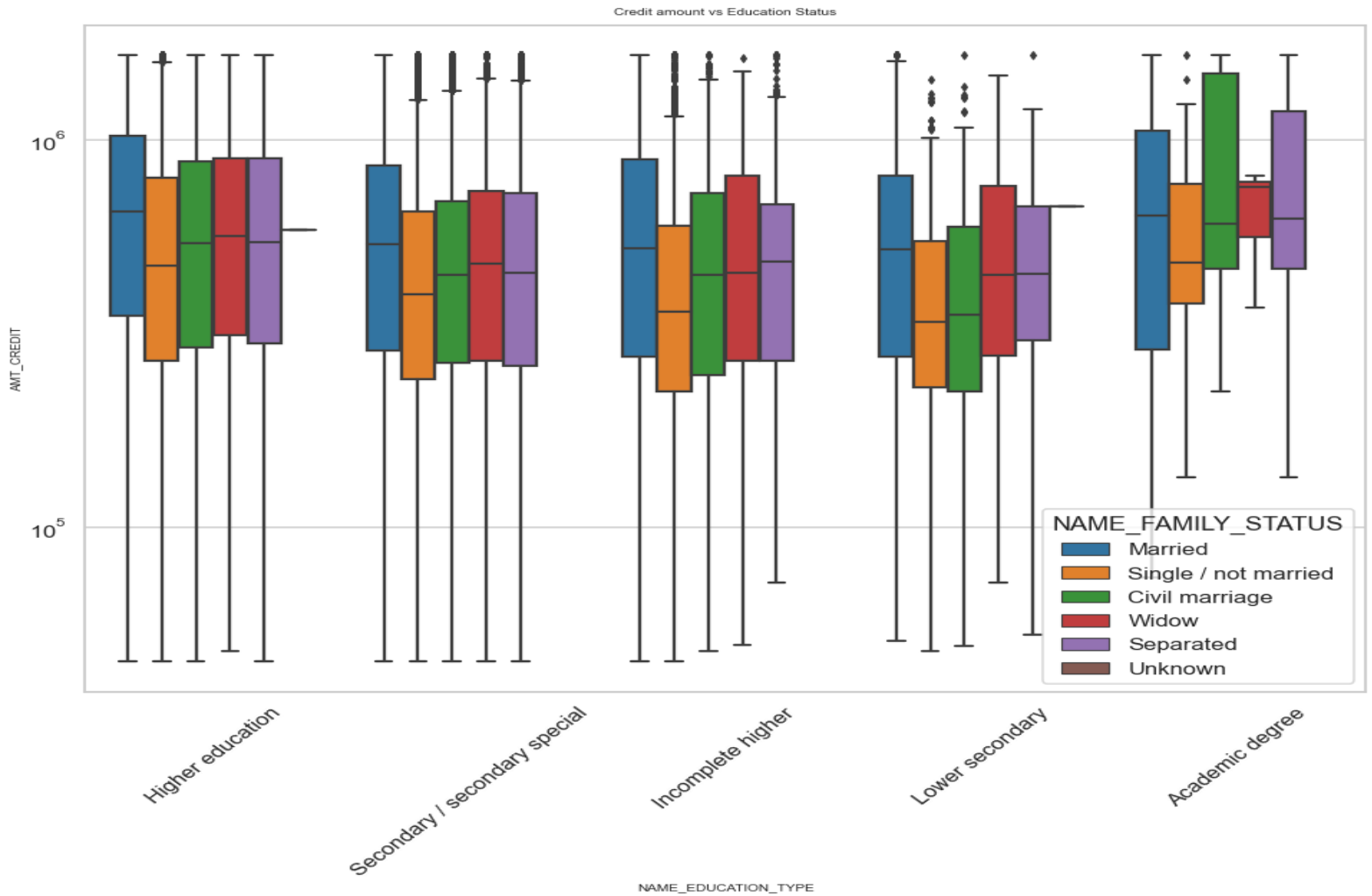



Inferences:

1) AMT_GOODS_PRICE vs AMT_CREDIT , AMT_GOODS_PRICE vs AMT_ANNUITY, AMT_CREDIT vs AMT_ANNUITY - Linear Corelation present

# Bivariate Analysis – Numerical vs Categorical

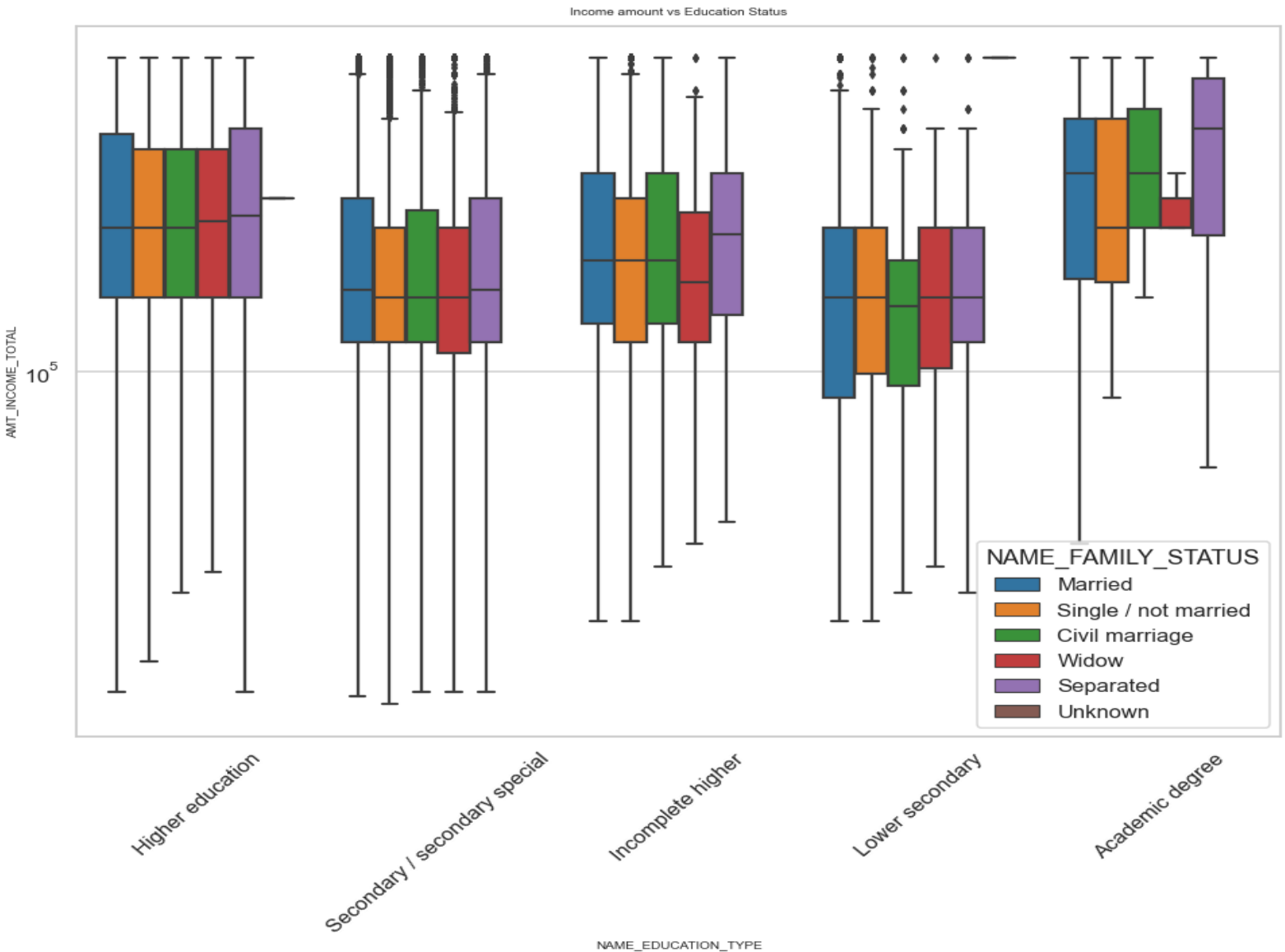
## Credit amount vs Education Status – Target 0




Inferences:

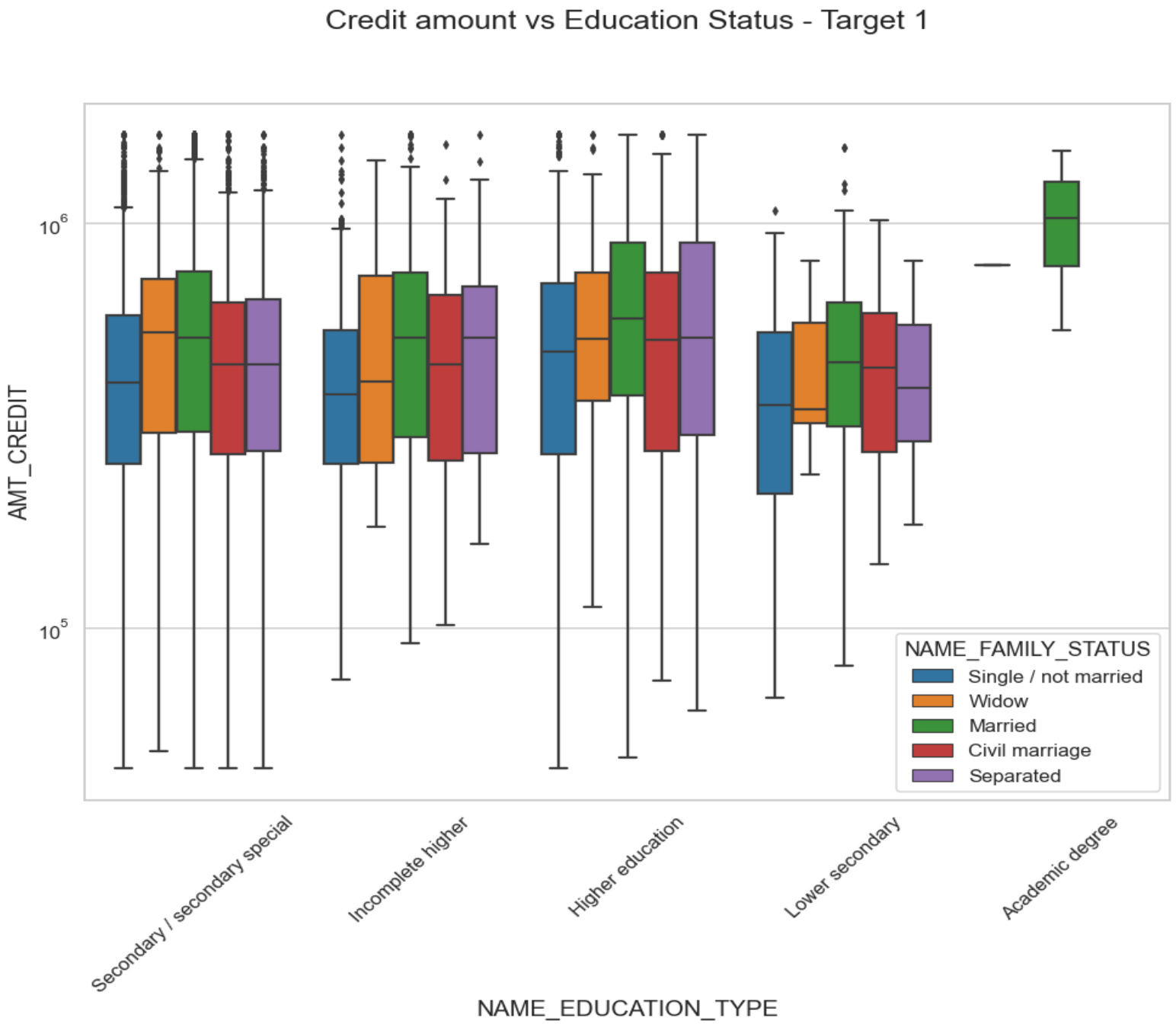
From the above plot we can conclude that, married, civil marriage, separated customers who are having academic degree are having high number of credit amounts than others.

# Income amount vs Education Status – Target 0



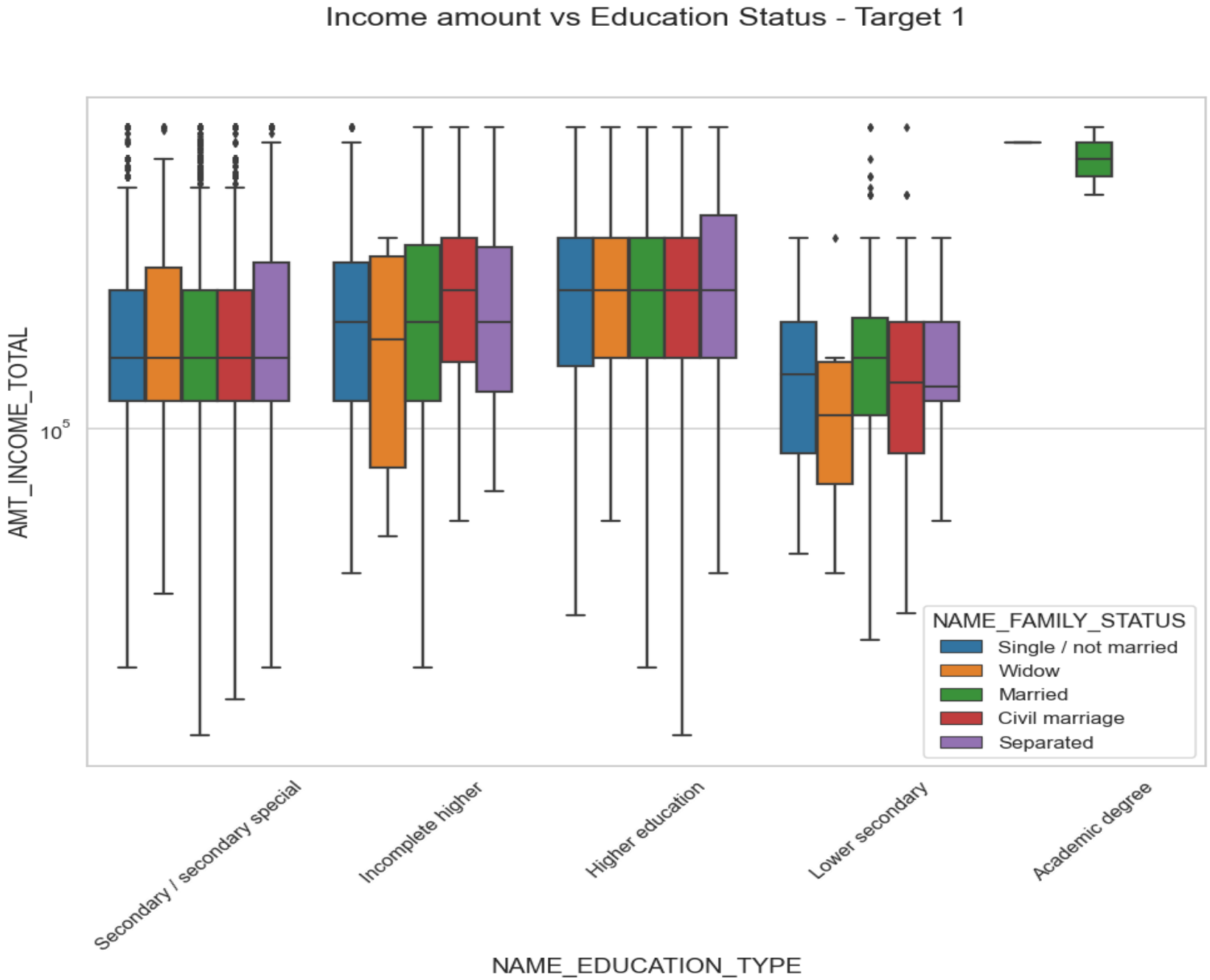

Inference:

Education Type - 'Higher Studies', the income amount almost equal to all familyStatus.




Inferences:

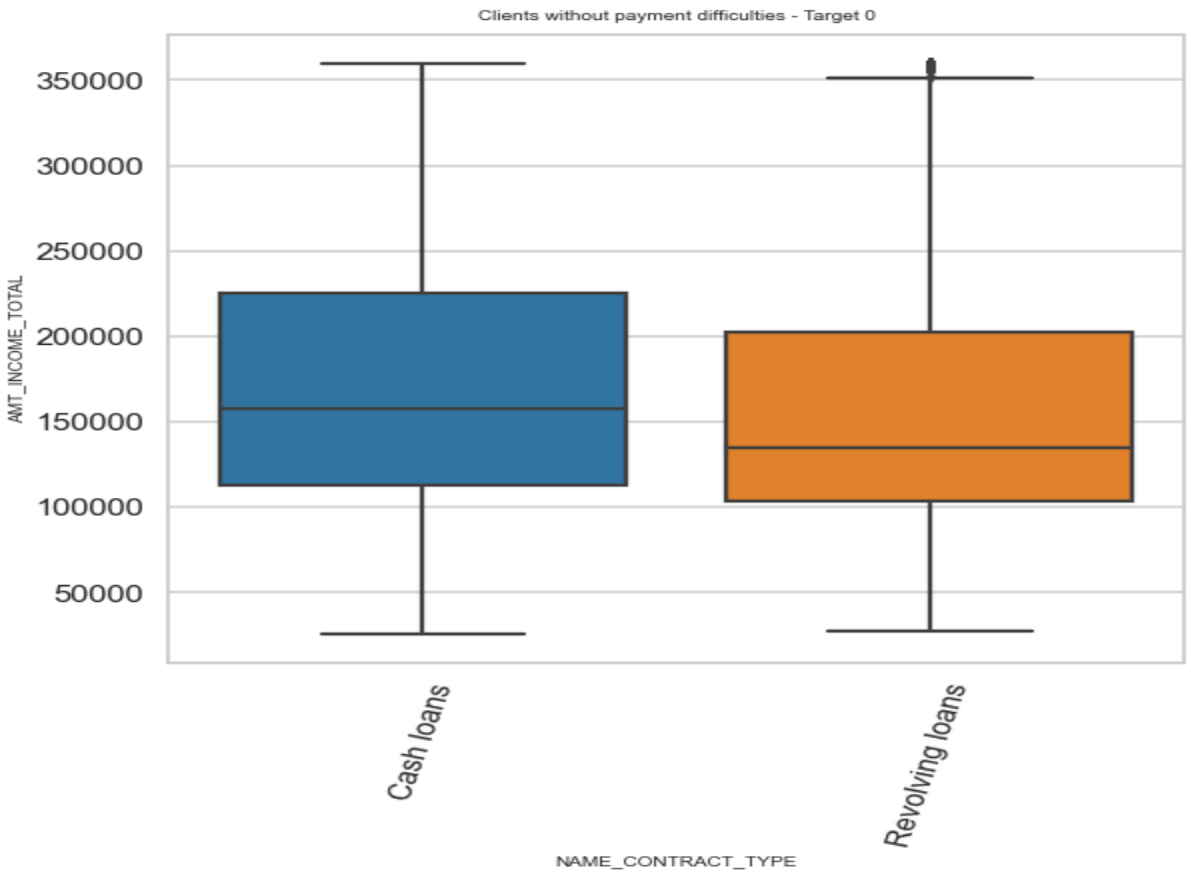
From the above plot we can conclude that, married, civil marriage, separated customers who are having Higher Education are having high number of credit amounts than others.

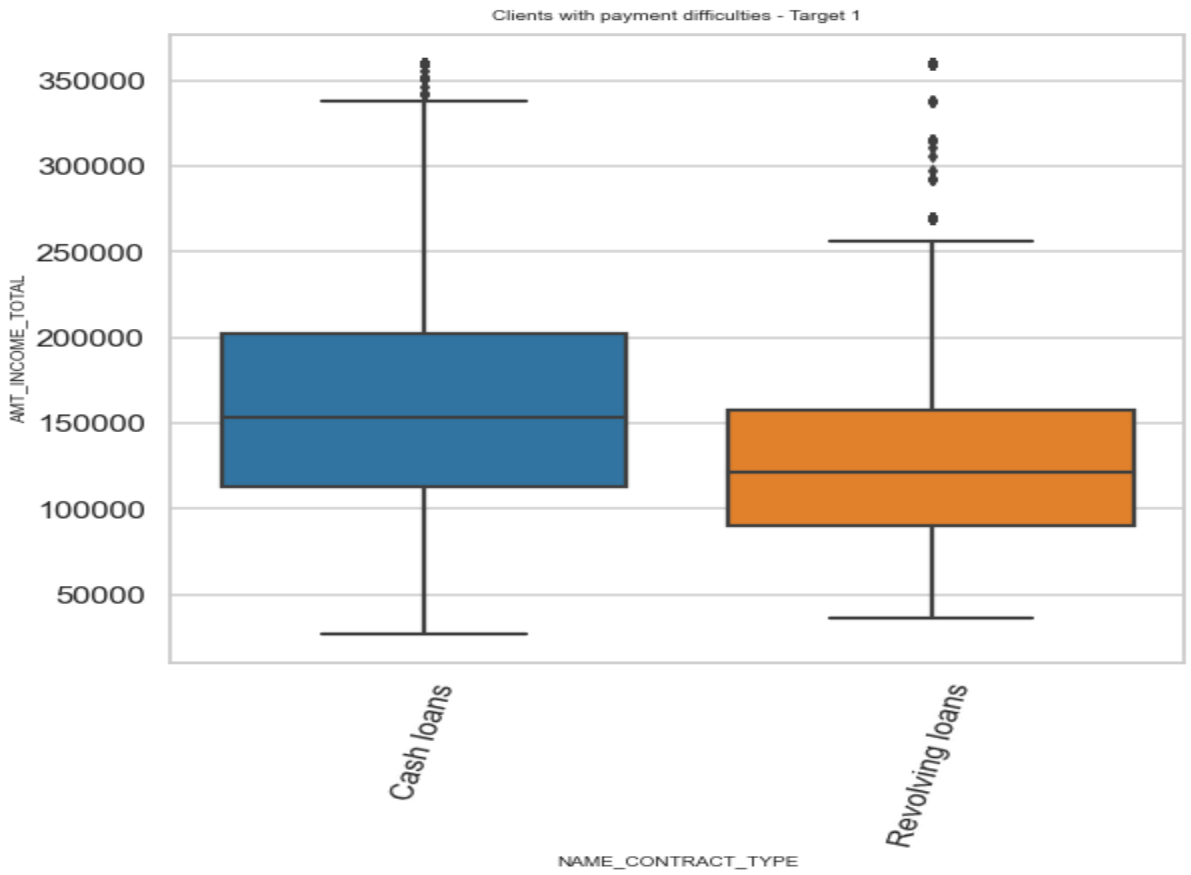



Inference:

Education Type - 'Higher Studies', the median, 75% quartile of income amount almost equal to all family Status except separated

# AMT_INCOME_TOTAL vs NAME_CONTRACT_TYPE

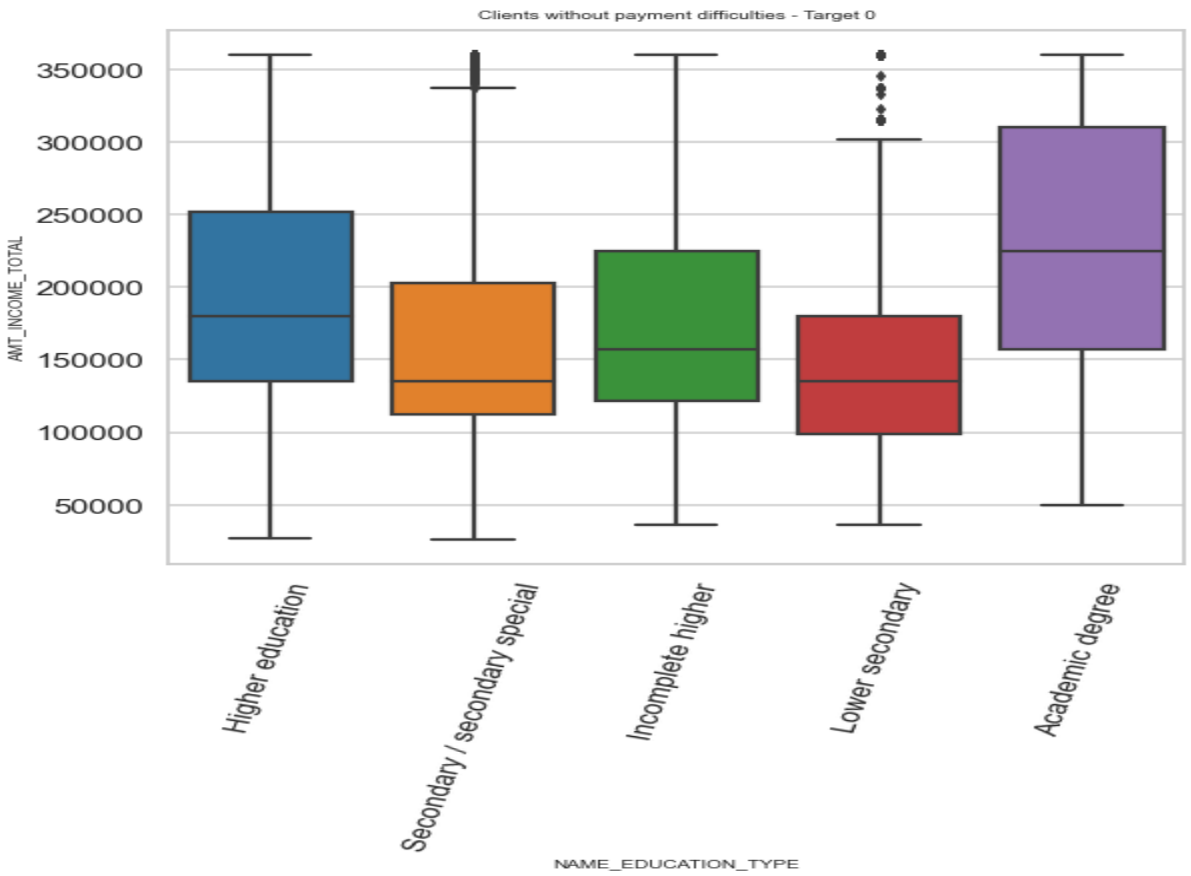


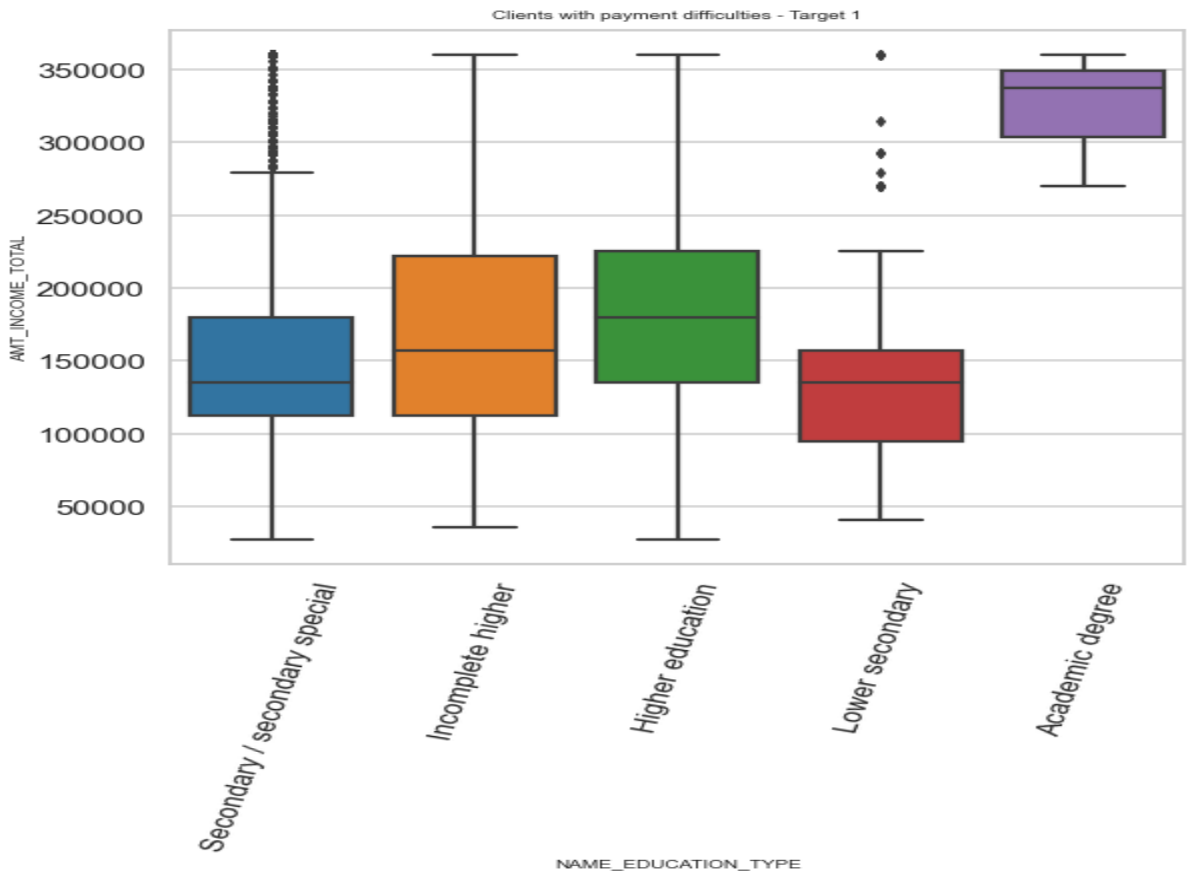




Inferences:

1) Median is almost similar for both Target 0&1 Cash loans
2) Some outliers present in Customer in Payment difficulties(Target1)
3) There is an huge difference between 75$^{th}$ Quartile of Target 0 and 1.
4) Income from the clients who opted for Cash loans are higher for both Target0 and 1 than revolving loans

# AMT_INCOME_TOTAL vs NAME_EDUCATION_TYPE

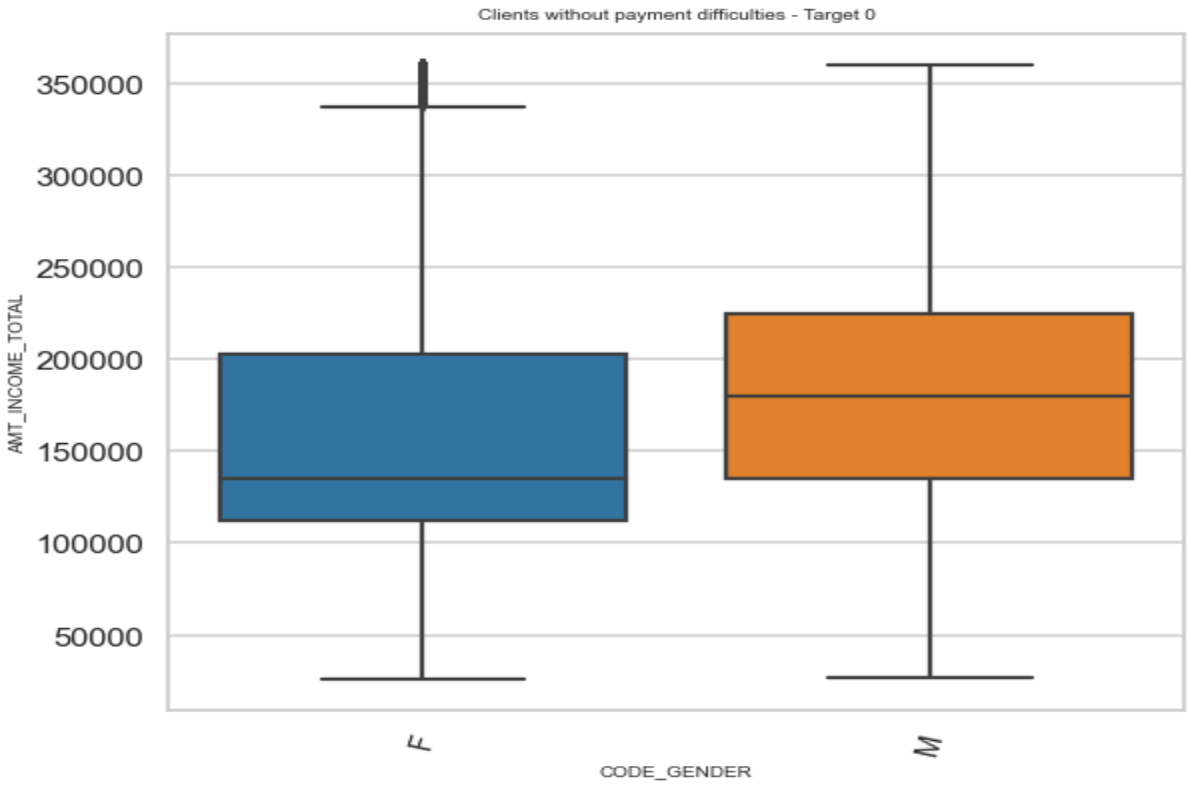


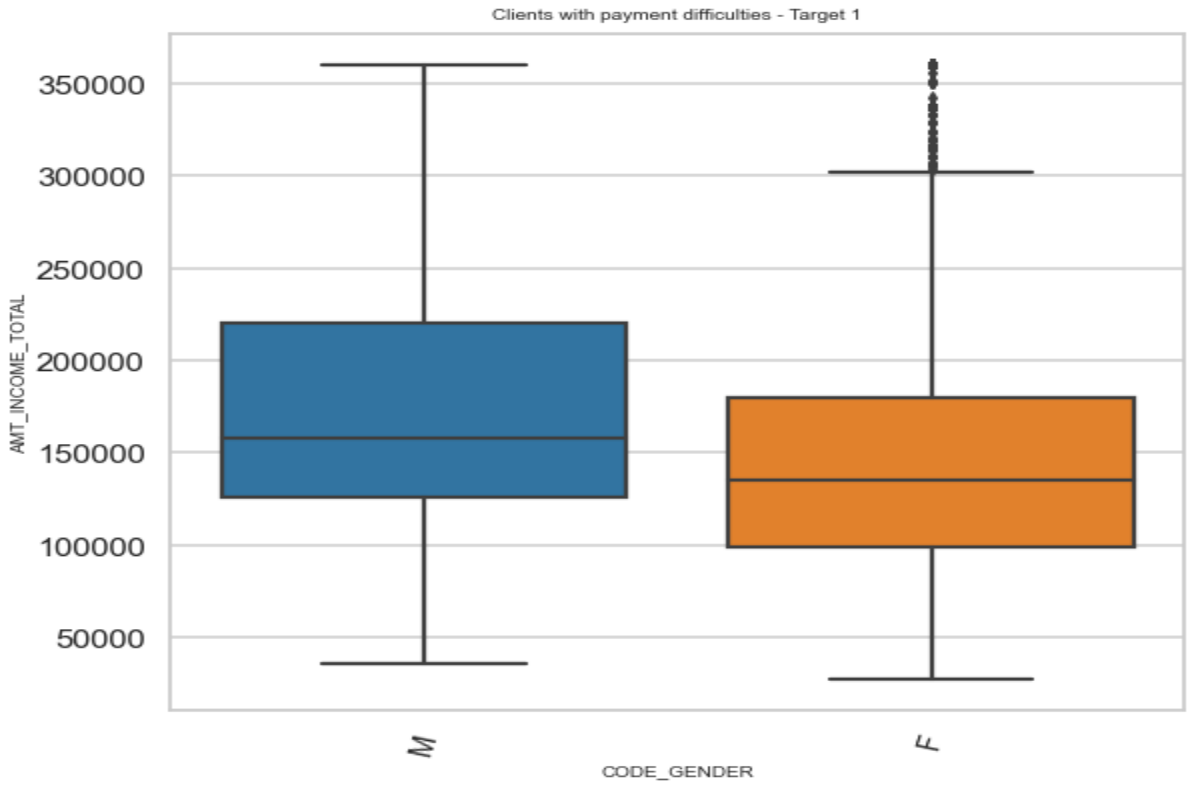




Inferences:

1) Income is higher for the customers who have Academic degree
2) Academic degree has high median, lower Secondary has the lowest median.

# AMT_INCOME_TOTAL vs CODE_GENDER

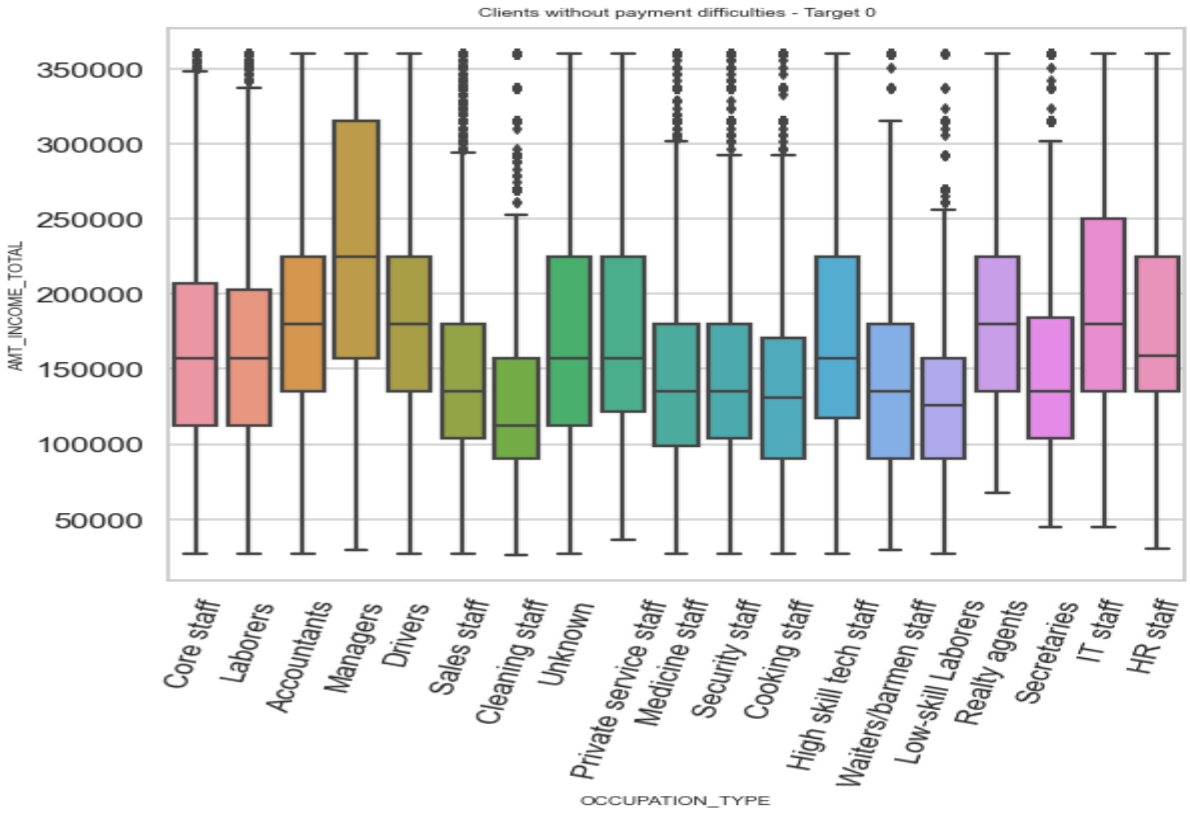


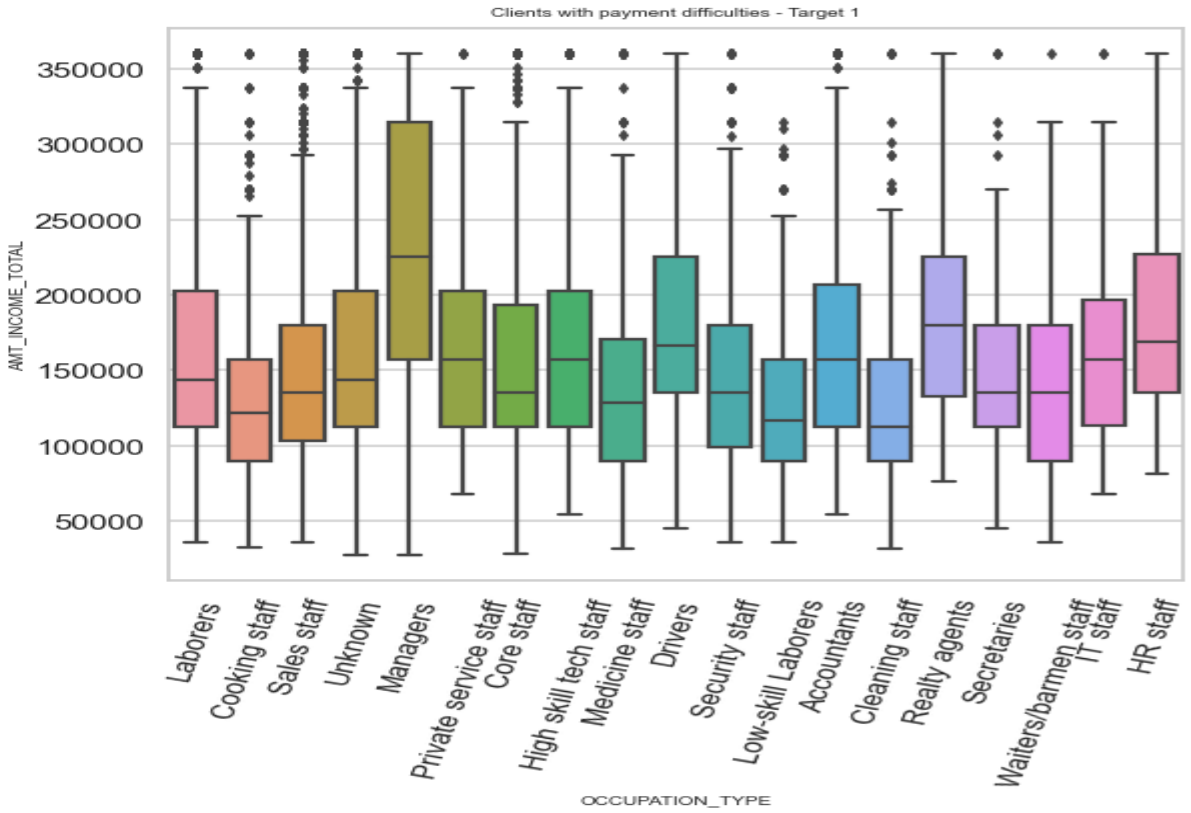




Inferences:

1) Median is similar for both Target 0 and 1 for code gender 'F'
2) No Outliers are present for code gender 'M'
3) Gender Code 'M' has more salary than 'F'

# AMT_INCOME_TOTAL vs OCCUPATION_TYPE

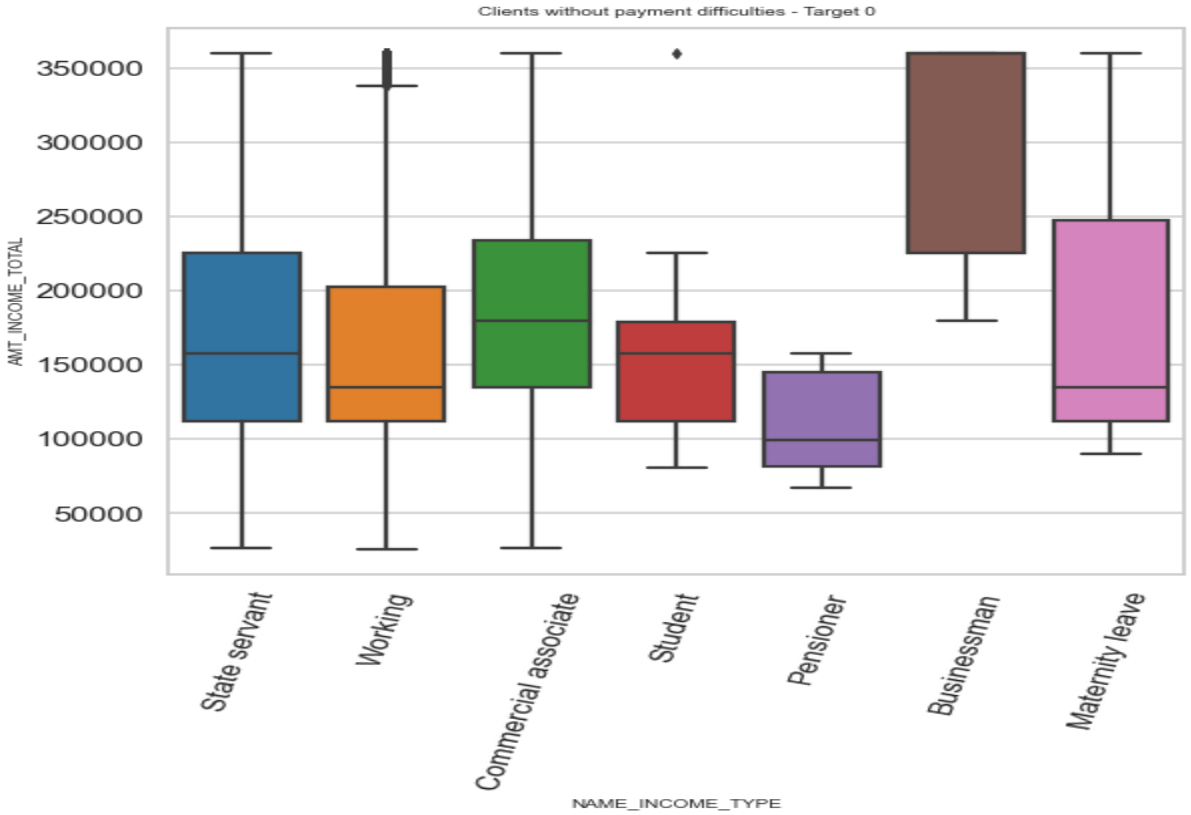


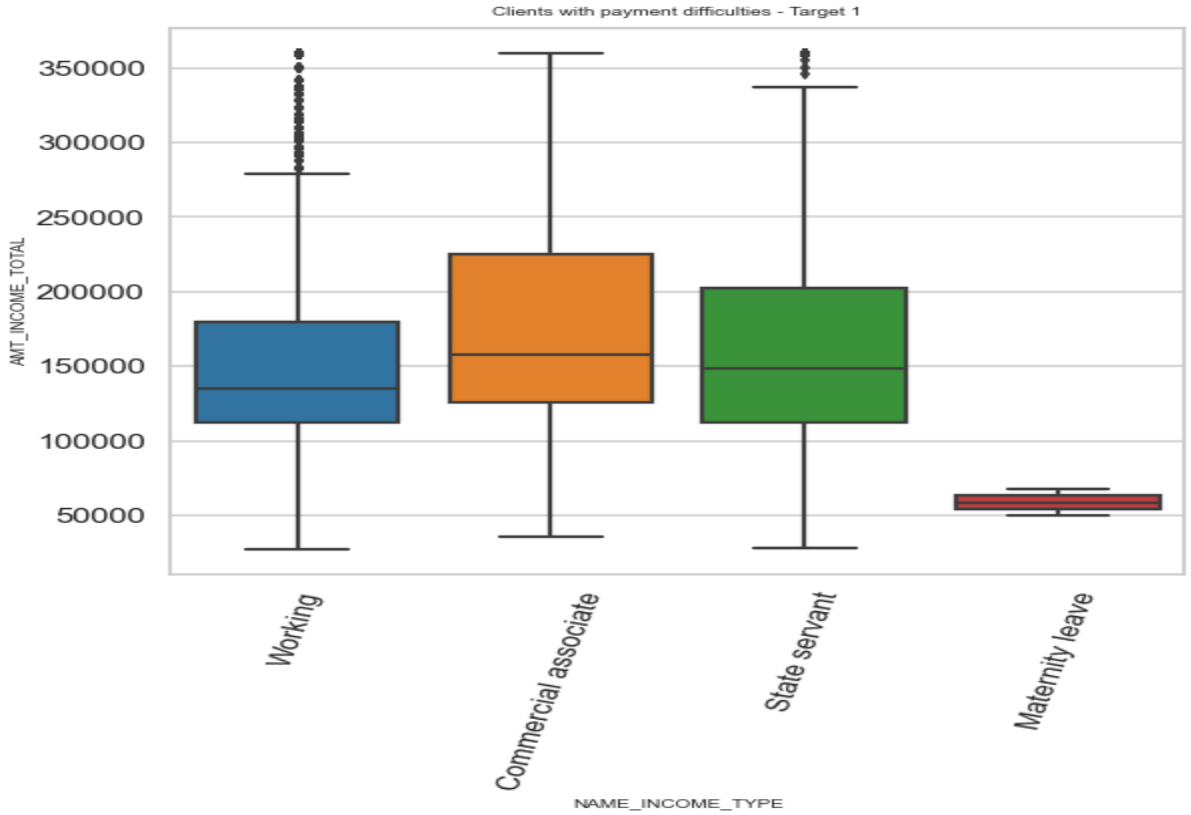




Inferences:

1) Managers have highest income
2) Cleaning staff, cooking staff and lowSkillLaborers have the lowest income

# AMT_INCOME_TOTAL vs NAME_INCOME_TYPE



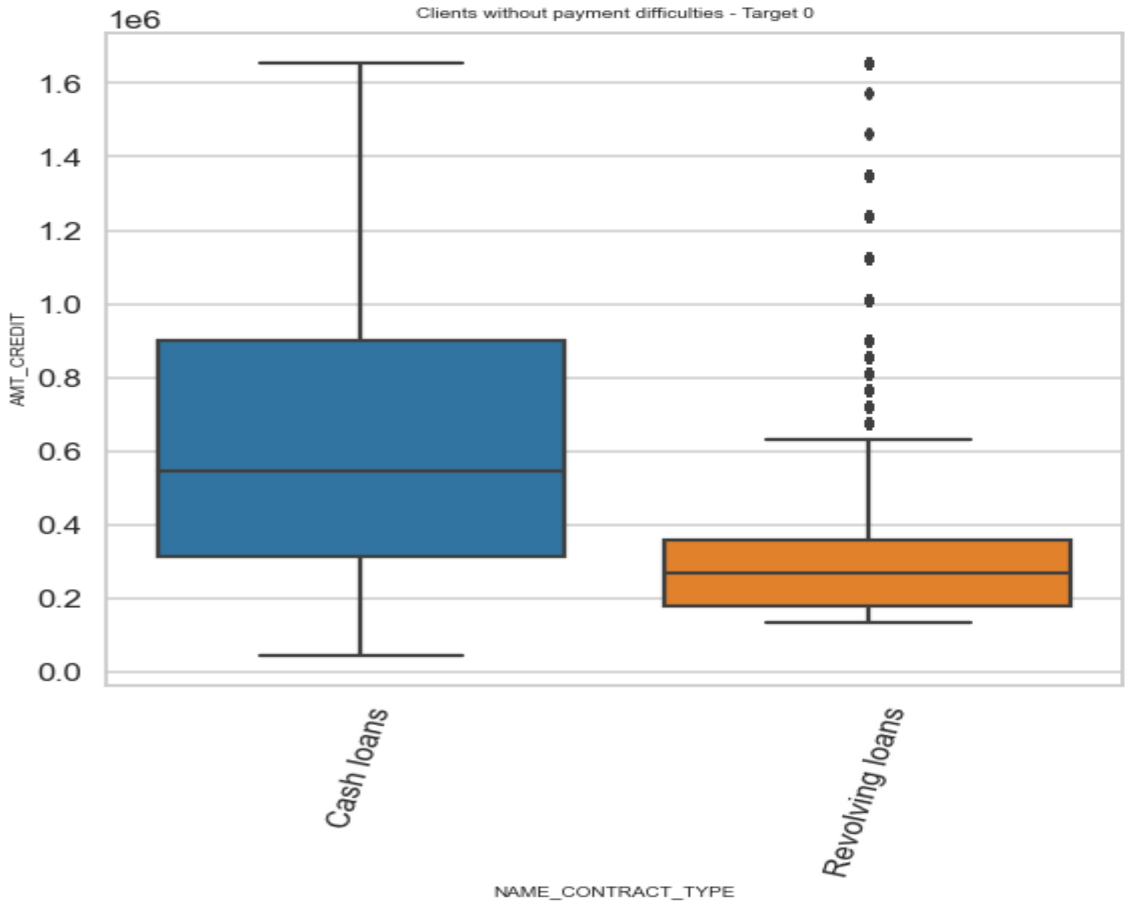



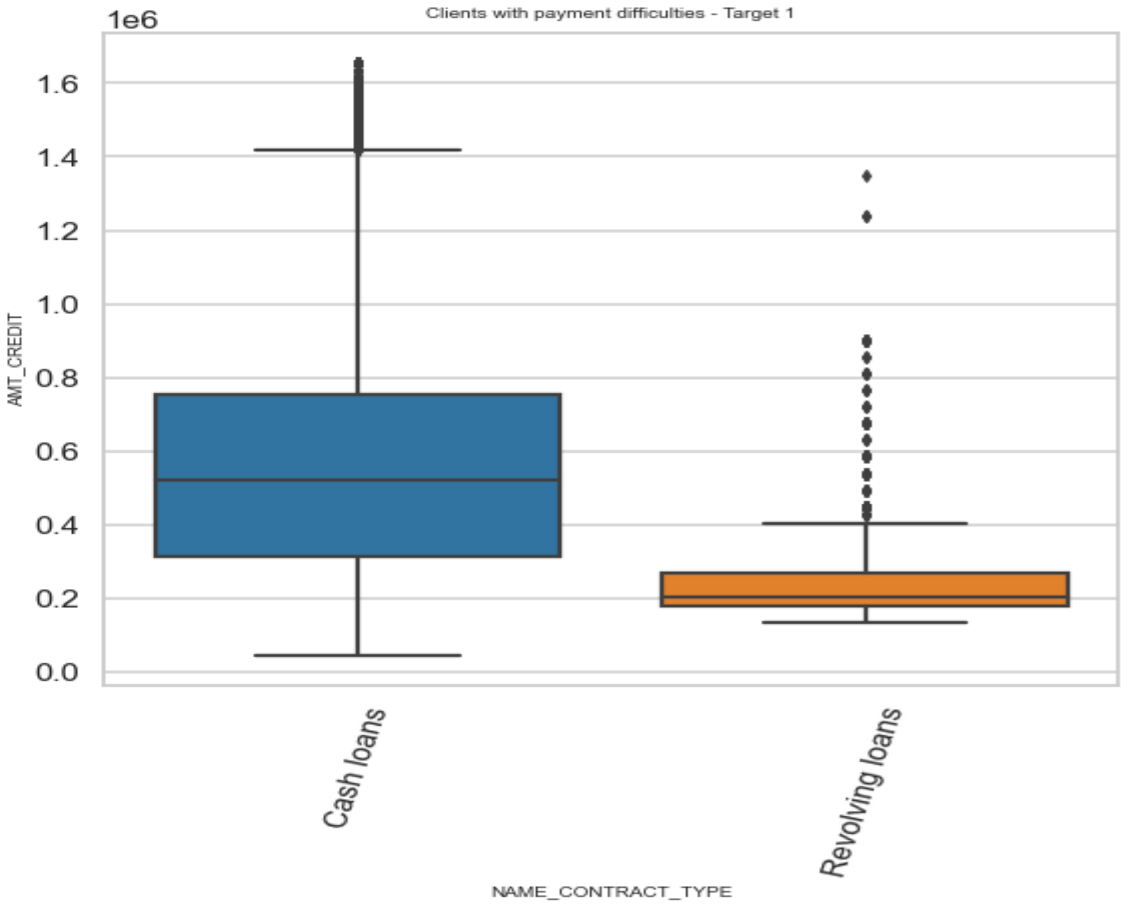

Inferences:

1) Commercial Associate have highest Income in both Target 0 and Target 1

2) Total Income is less for Student and pensioner incometypes for the customers who doesn't have payment difficulties

# AMT_CREDIT vs NAME_CONTRACT_TYPE

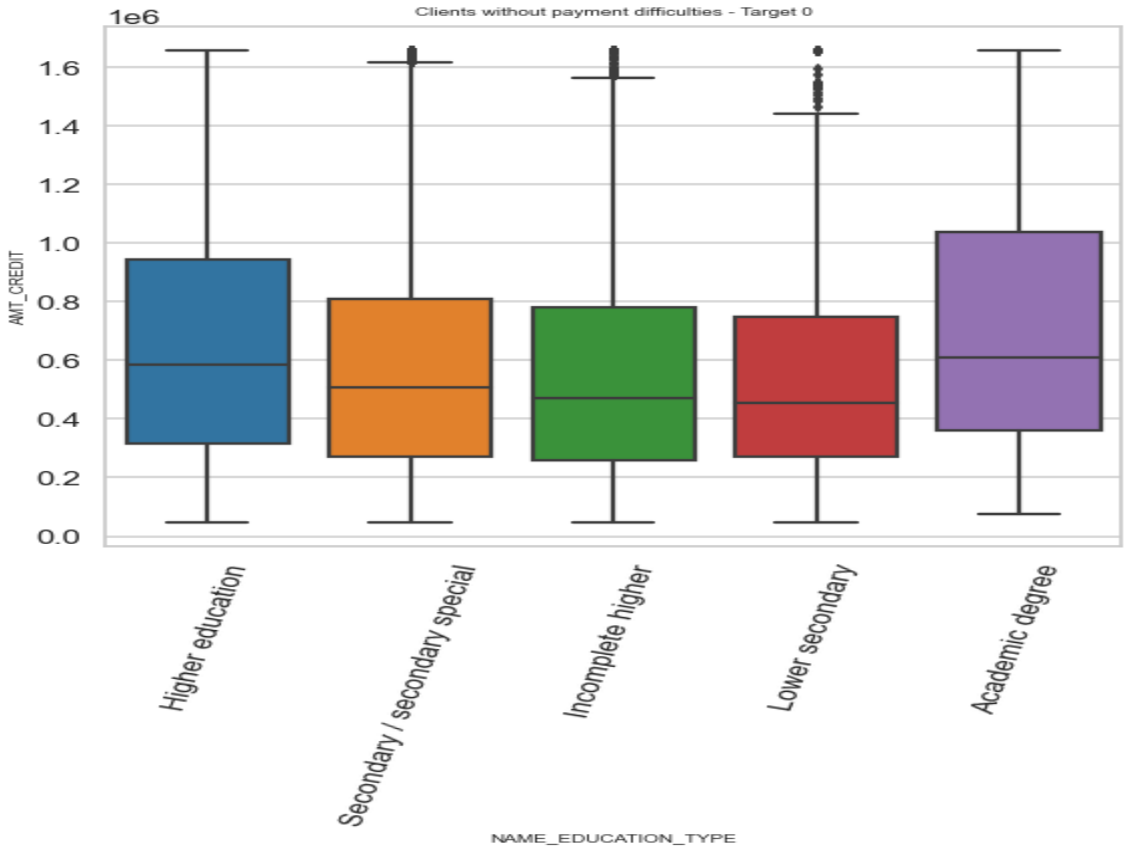


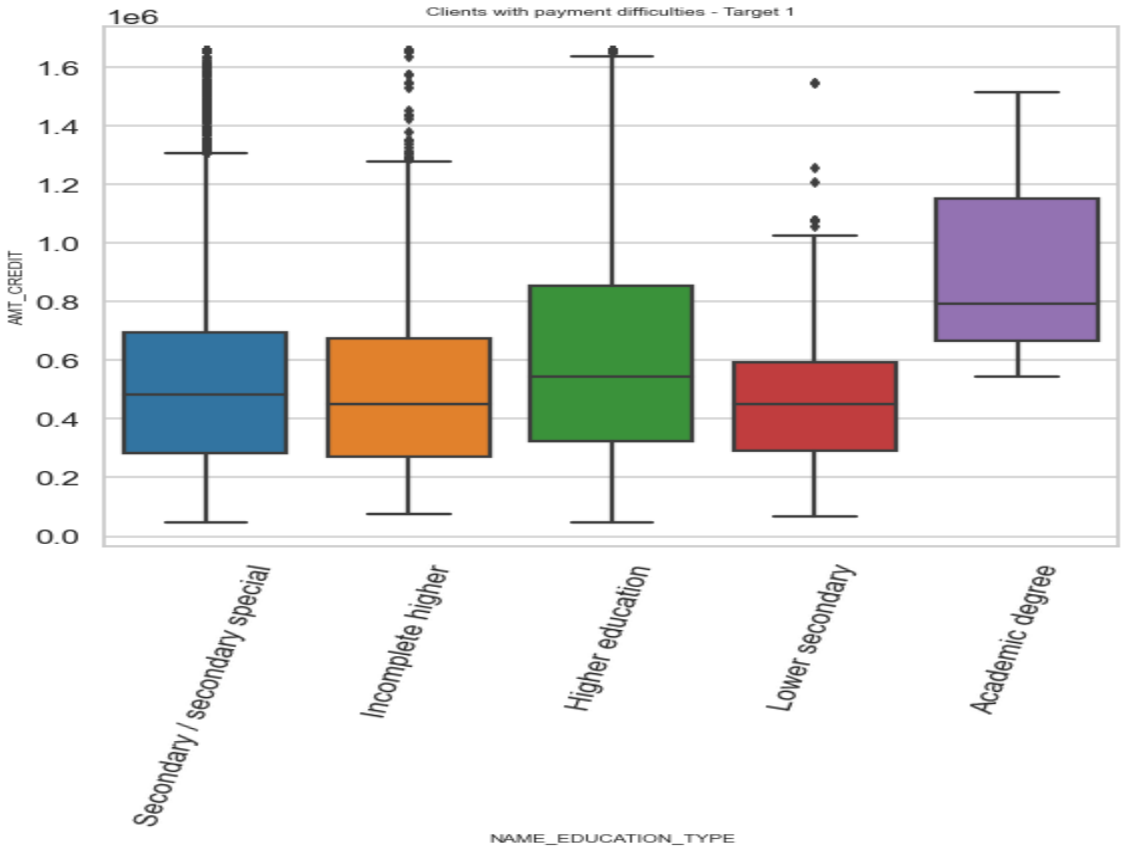




Inferences:

1) Credit amount is higher for clients who opted for cash Loans
2) More number of Outliers are present for Revolving loans

# AMT_CREDIT vs NAME_EDUCATION_TYPE

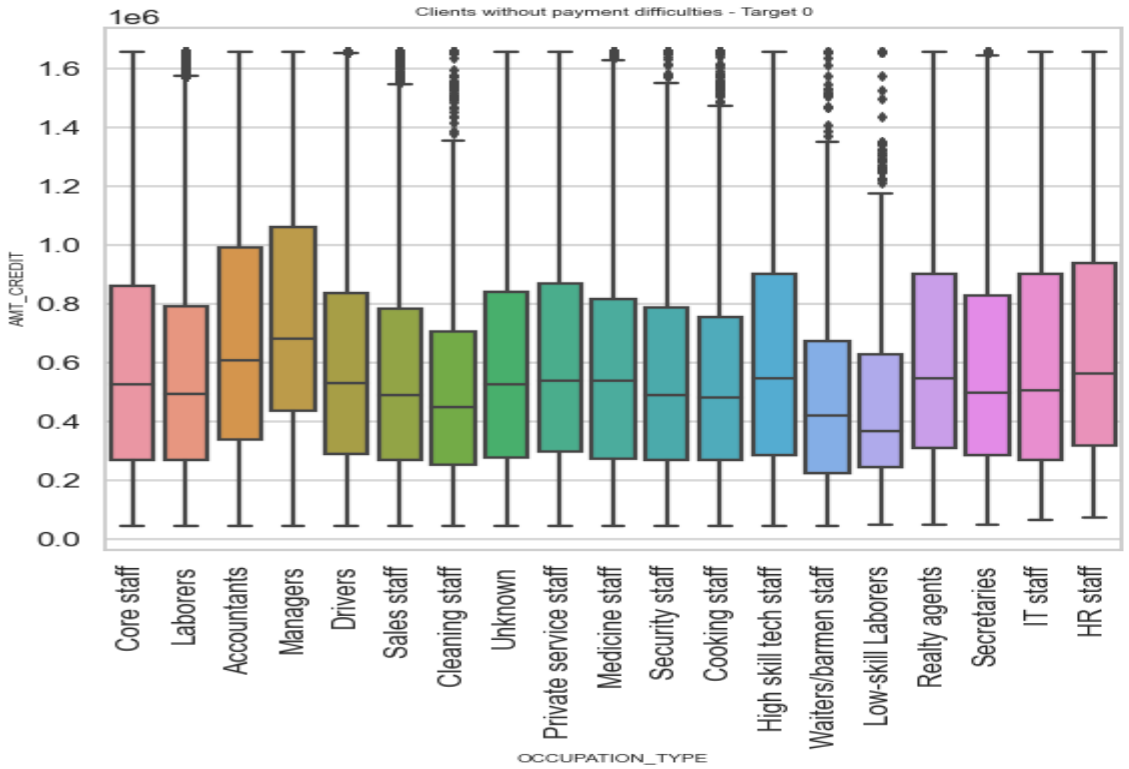


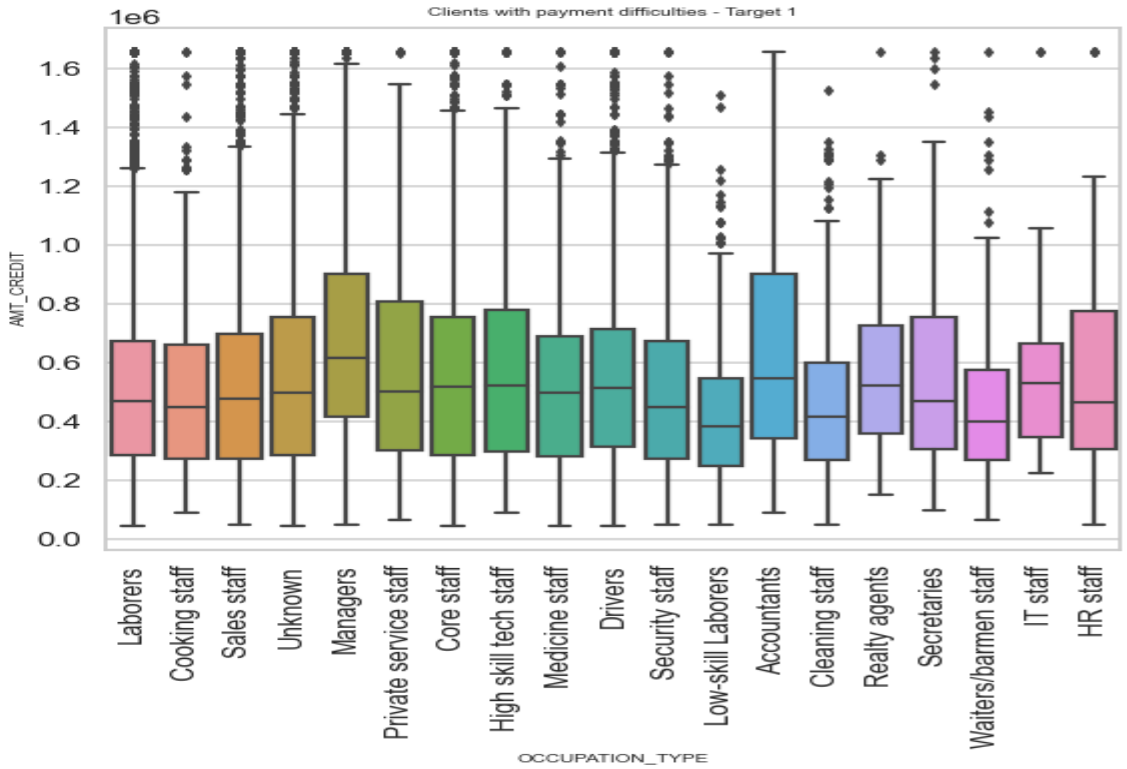




Inferences:

1) Higher Education, Academic degree, Secondary Education holders have high credit amount
2) Median for Academic degree is higher, lowest is for lower secondary and income higher for both Target1 and Target2

# AMT_CREDIT vs OCCUPATION_TYPE

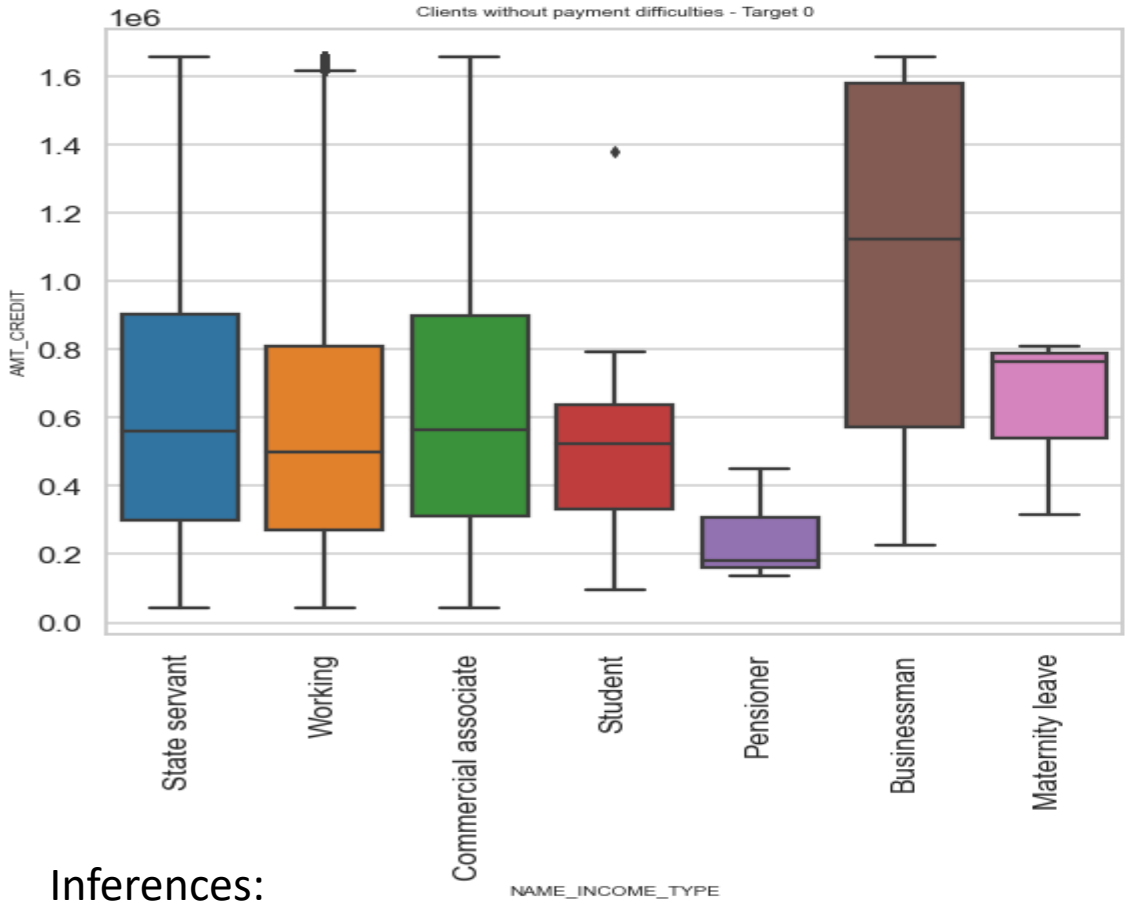


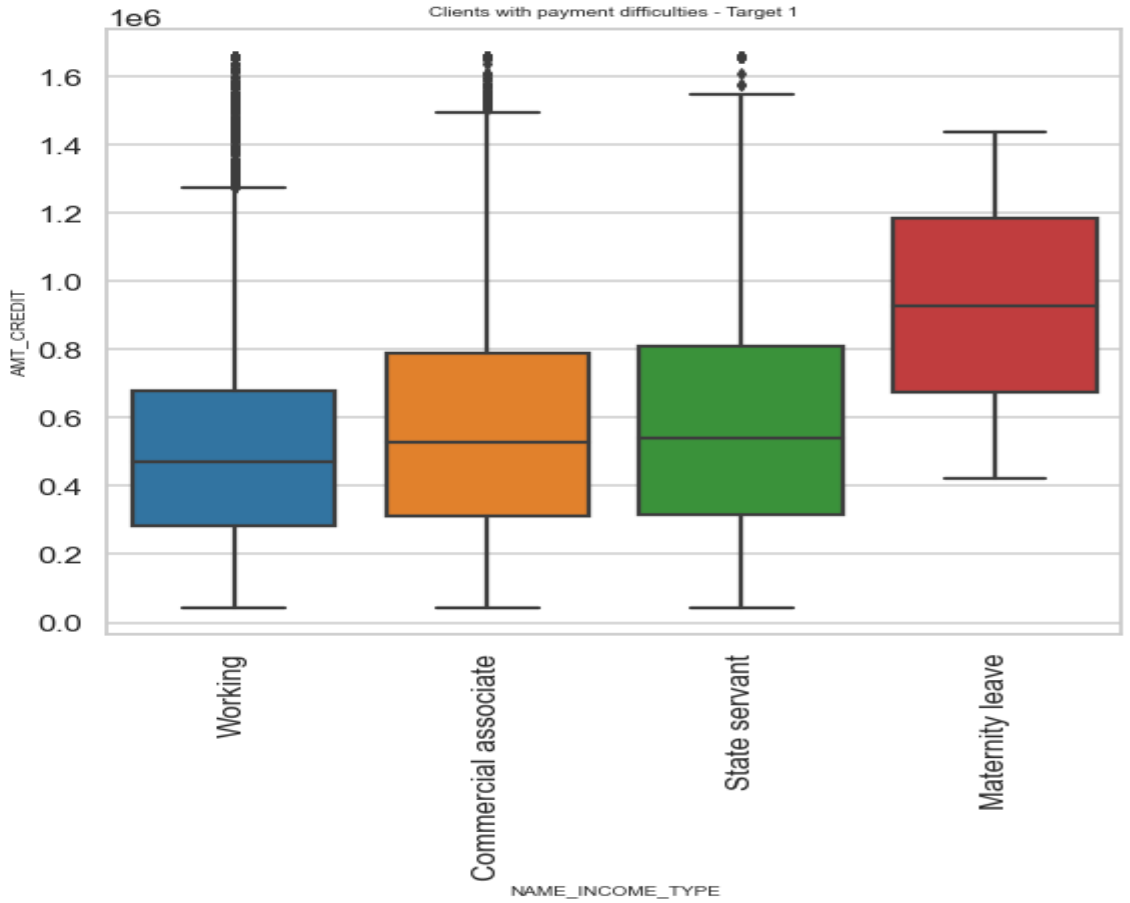




Inferences:

1) Median for Managers is highest for both
2) LowSkillLaborers have lowest credit amount
3) More number of Outliers are present for Customers who has Payment difficulties

# AMT_CREDIT vs NAME_INCOME_TYPE

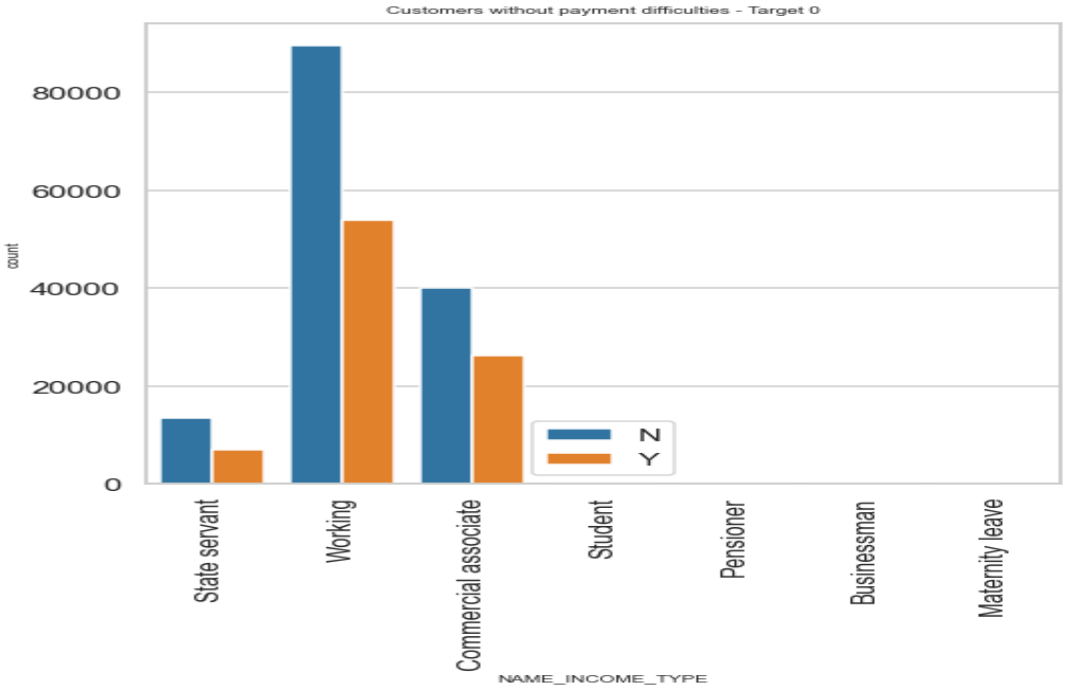


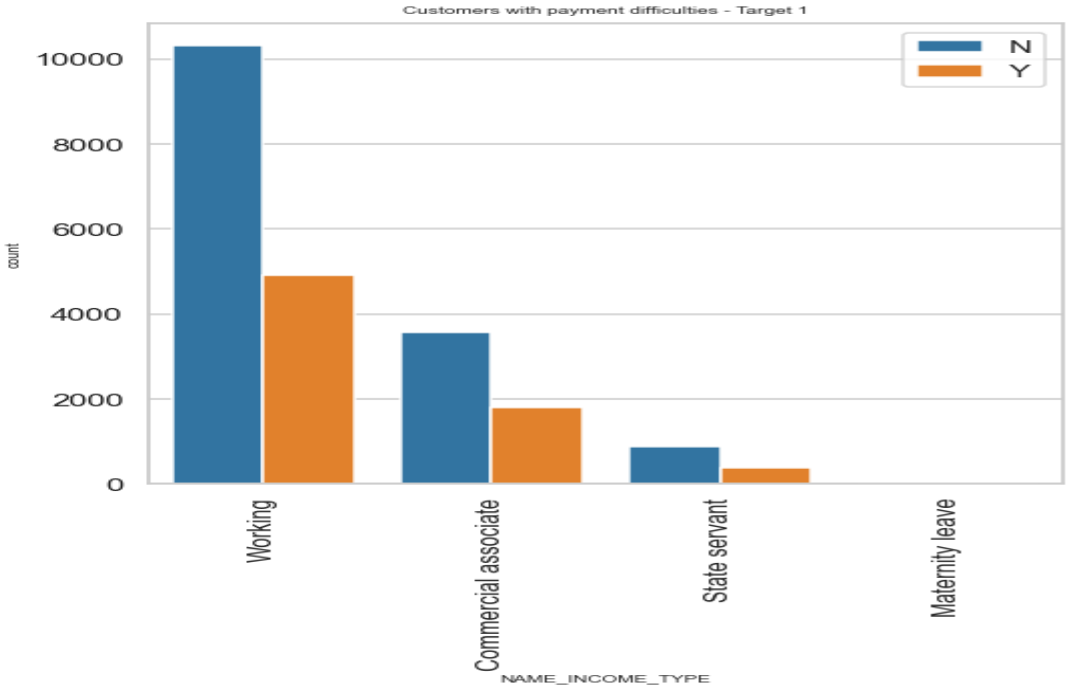




Inferences:

1) Number of Outliers are more for customers who are not facing any payment difficulties
2) under Target 0 - median is higher for Businessman
3) under Target 1 - median is higher for Maternity leave
4) Not facing any payment difficulties - Businessman, Pensioner and Students

# NAME_INCOME_TYPE vs FLAG_OWN_CAR

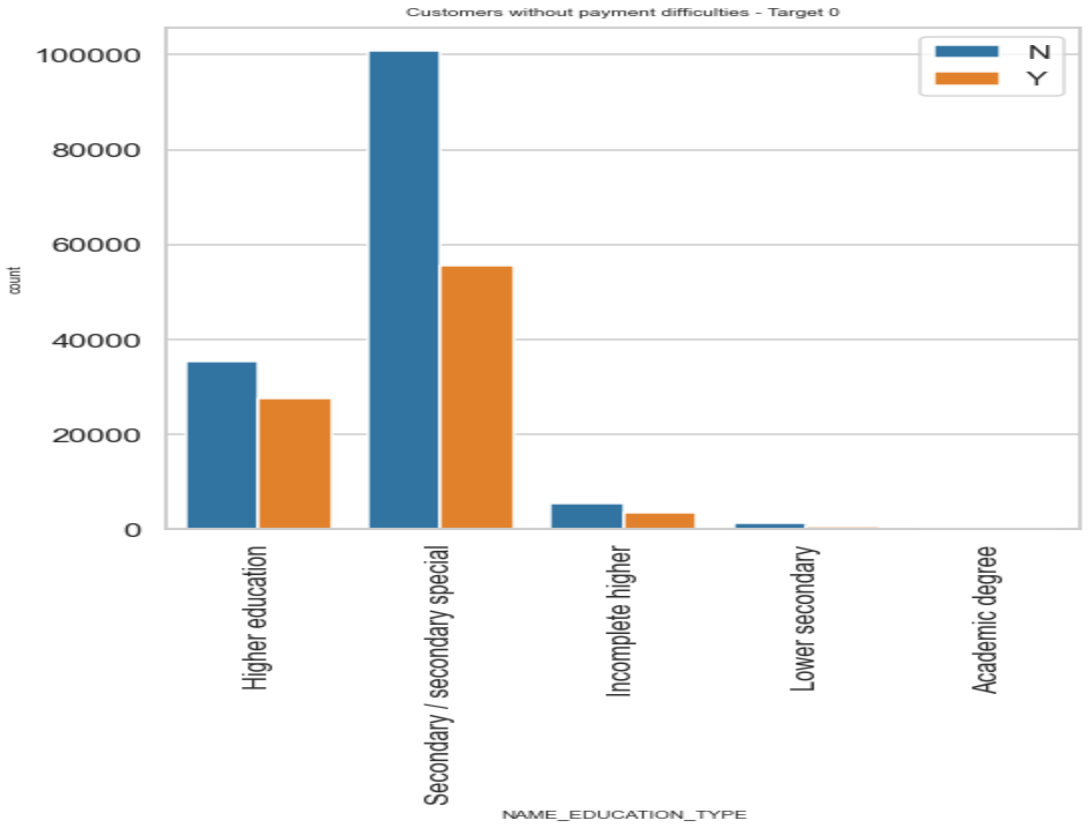


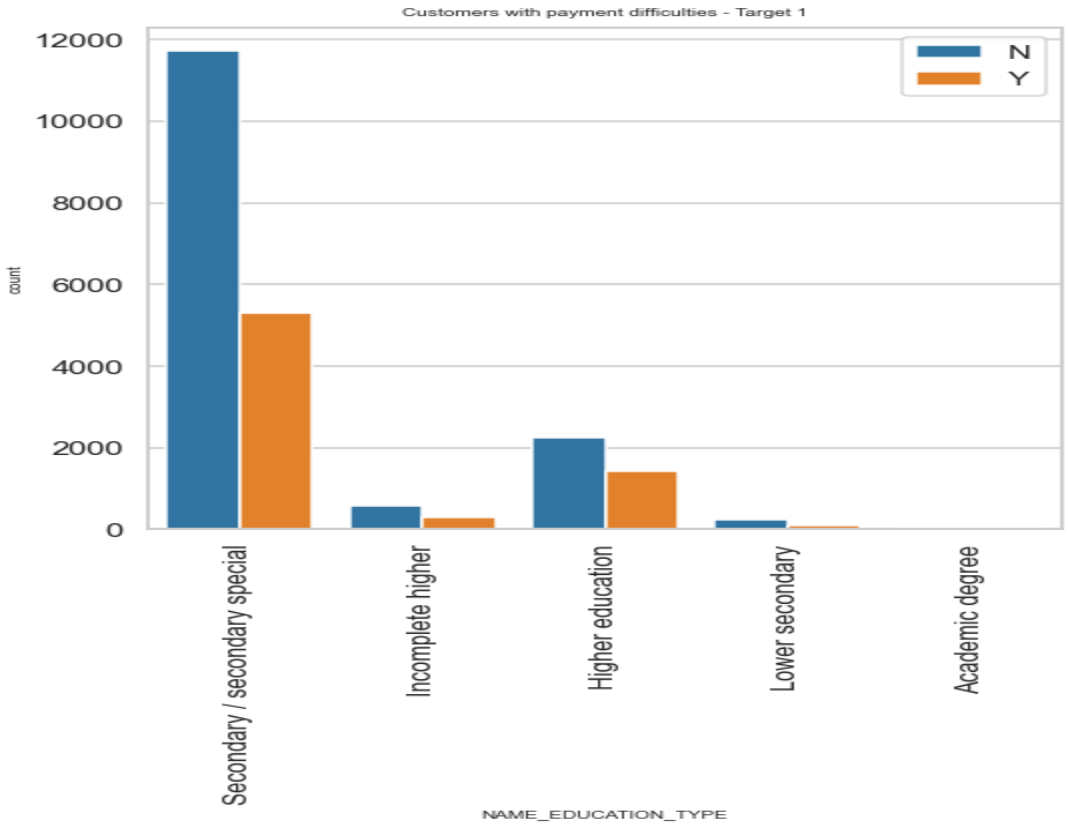




Inferences:

1) The customers who has Income Type Working having highest number of cars
2) Income Type - Student, Pensioner, Businessman, Maternity leave doesn't have cars

# NAME_EDUCATION_TYPE vs FLAG_OWN_CAR



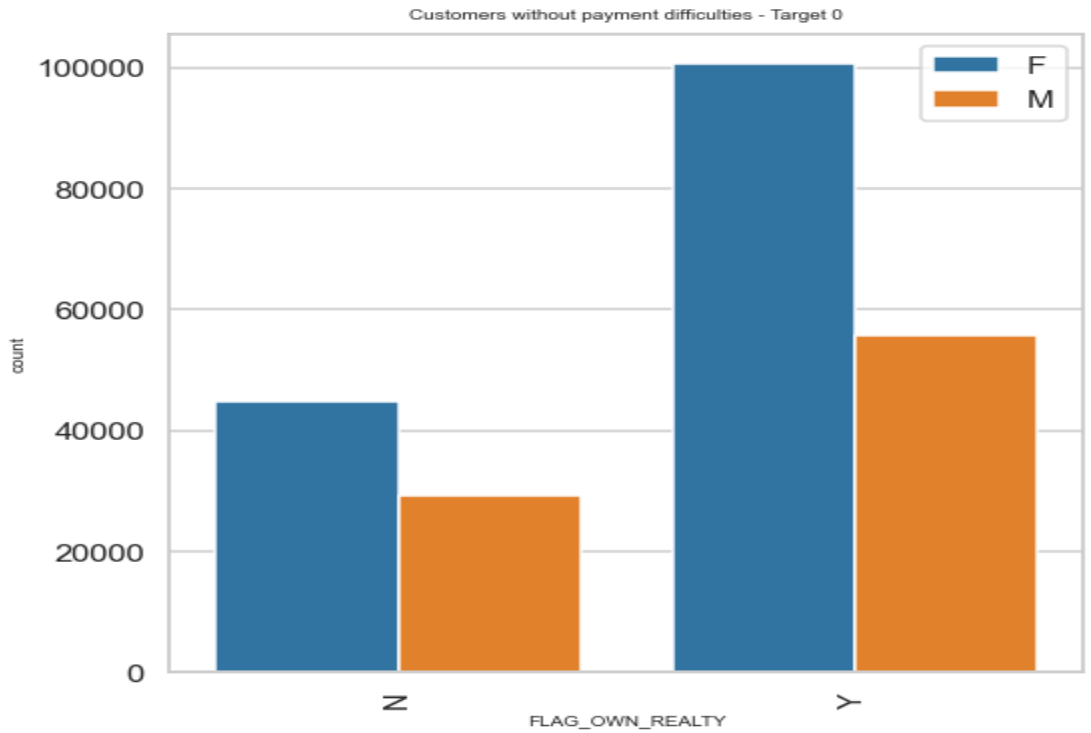



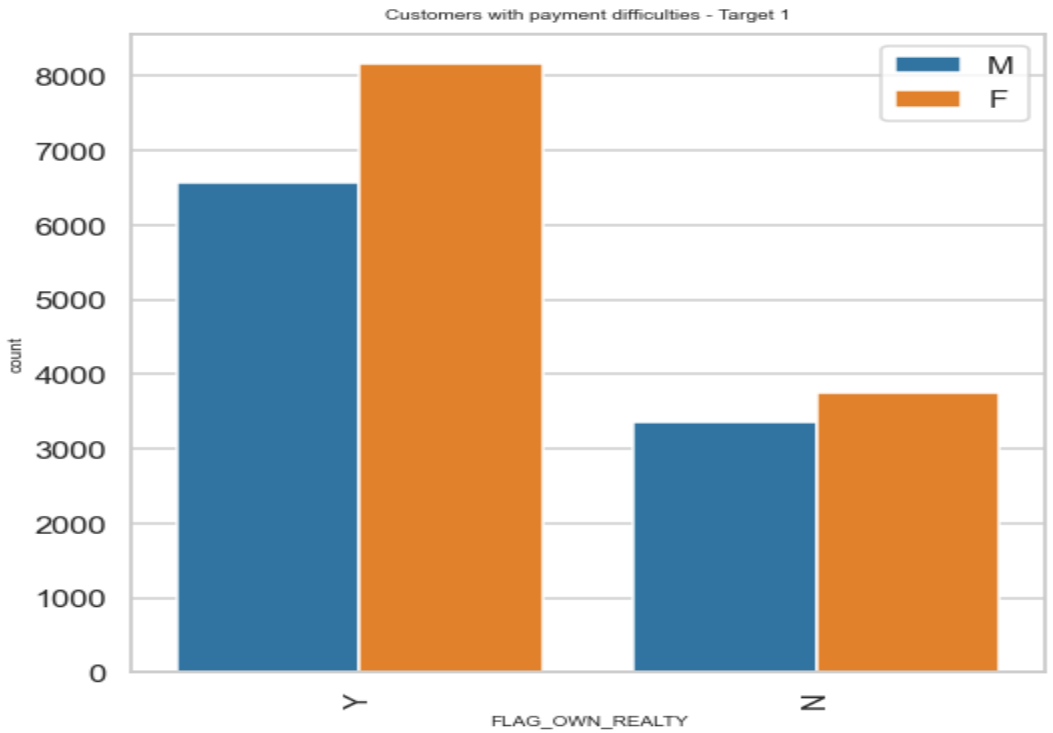

Inferences:

1) Academic degree customers have no cars

2) Secondary/ Secondary Special Education Customers have highest number of cars, next comes the Higher Education customers

# 'FLAG_OWN_REALTY' vs 'CODE_GENDER'




Inferences:

1) Female customers are more who owns reality than male for both Customer without payment difficulties and Customer with payment difficulties

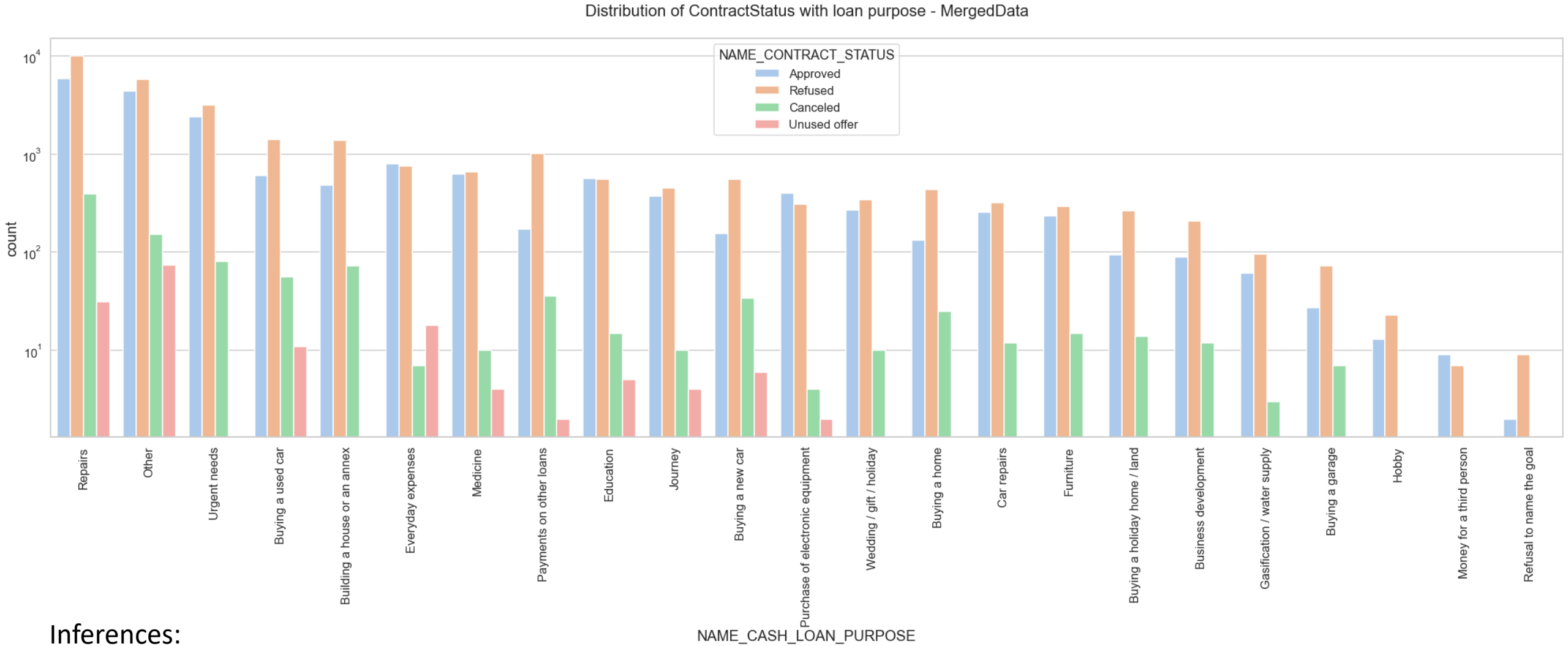
2) There is a huge difference between males and Females for both Customer without payment difficulties and Customer with payment difficulties

EDA on Previous Application Data:

Step1 → Load the Dataset
Step2 → Handle the Missing values

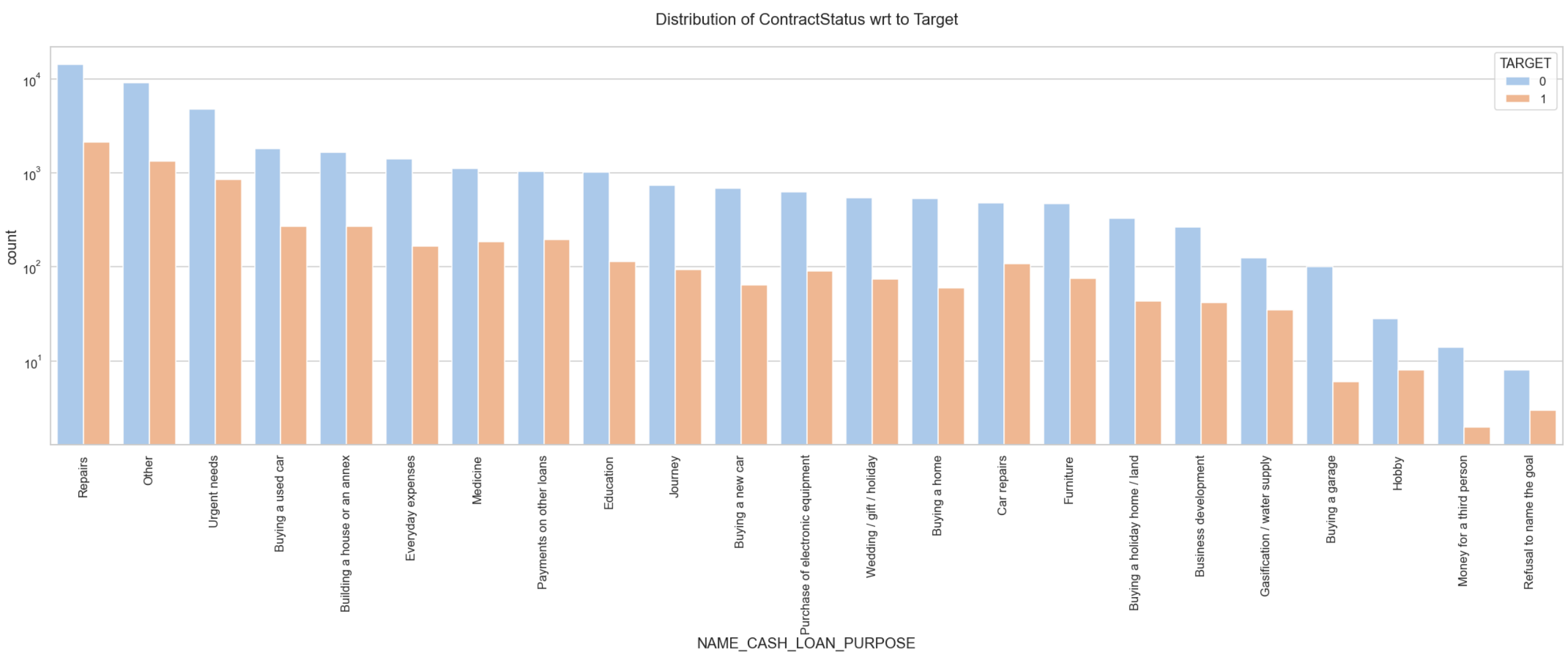
- Dropping the columns having null values greater than 30% in the dataset.
- Handling ‘XNA’ and ‘XAP’ values either by creating a new category as unknown if there is a high percentage of those values  but that column is important for analysis. And also deleting the rows which are having these values if the percentage is very less

Step3 → Handling Outliers – using Capping and Flooring.
Step4 → Performing Univariate Analysis
Step5 → Performing Bivariate/Multivariate Analysis



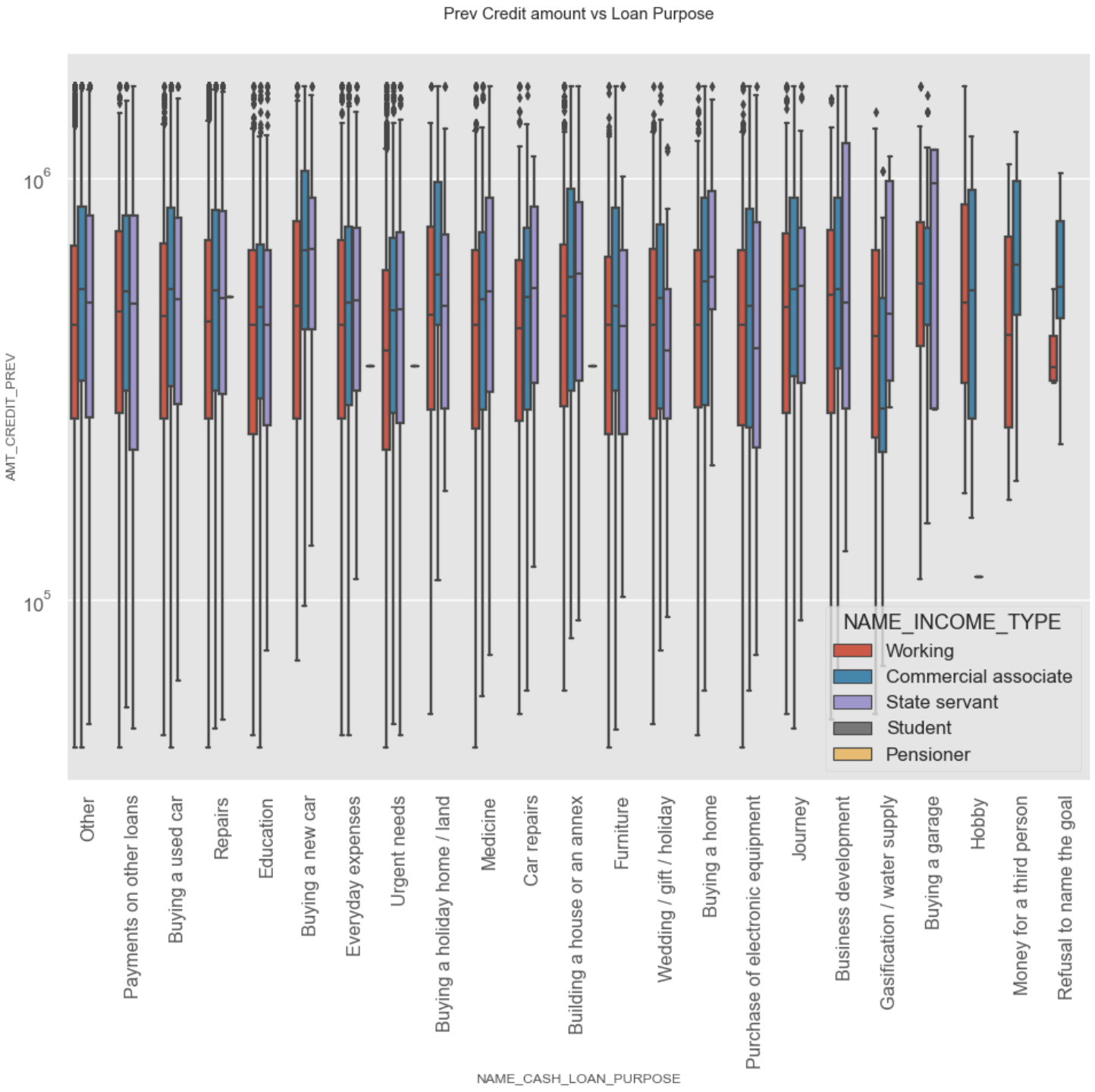

Inferences:

1) Maximum rejection of loans are from Repairs
2) Education Purpose - we have equal number approves and rejections
3) Maximum number of rejections than approvals - Payment of other loans, Buying a new car, Refusal to name the goal
4) Maximum cancelled loans are for Repairs, Urgent needs, other categories




Inferences:

1) Difficulty in payment on time - Repairs.



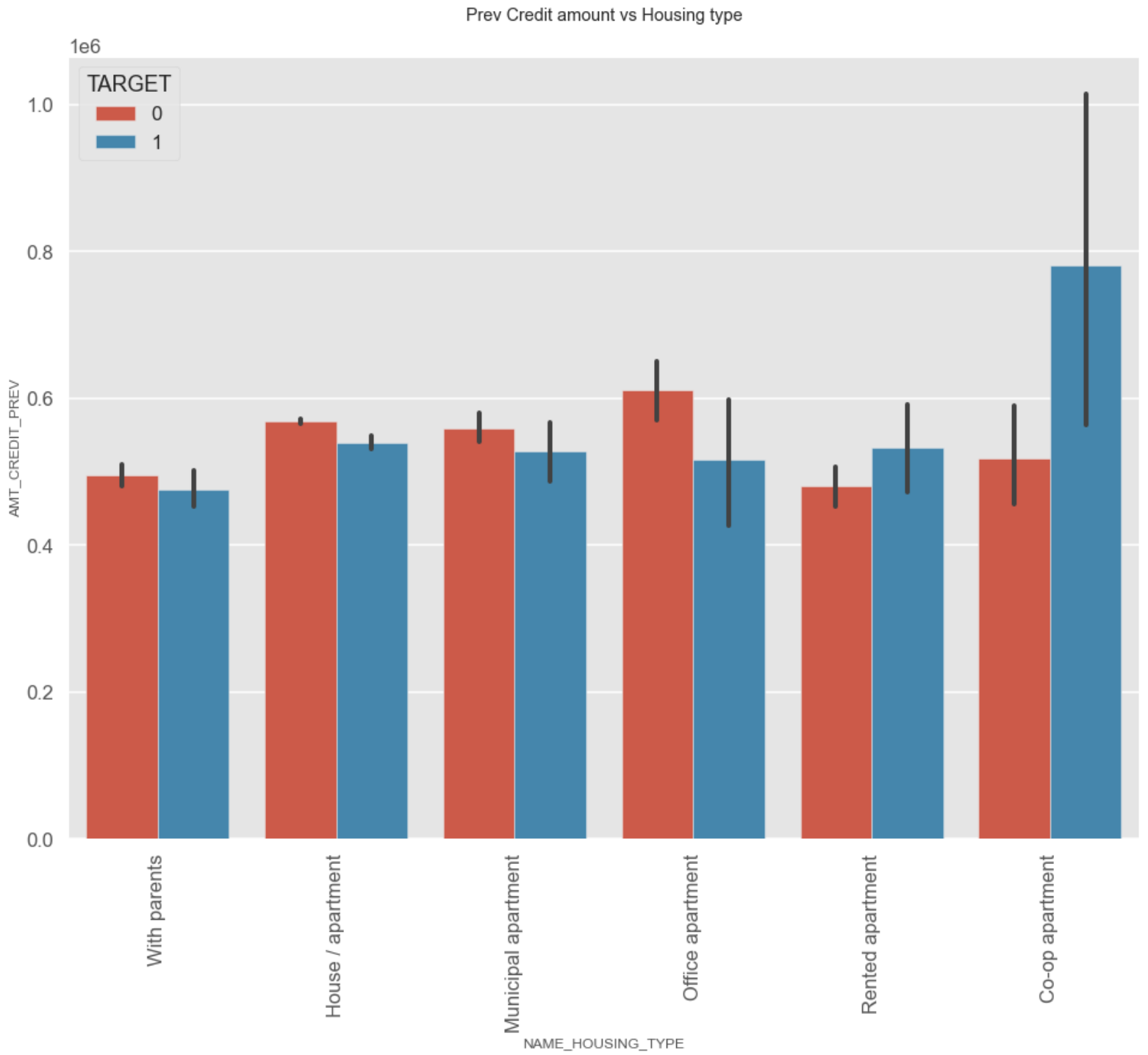

Inferences:

1) The Loan purposes 'Buying a home','Buying a land','Buying a new car' and'Building a house' have credit amount higher.

2) There is a significant amount of credit applied for the Income type of state servants.

Inferences:

Office apartment, Housing type is having higher credit of target 0
and
co-op apartment is having higher credit of target 1.
Conclusion:
→Focus should be on housing type with parents or House\apartment or municipal apartment for successful payments.
→Bank should avoid/reduce giving loans to the customers who are having housing type as co-op apartment as they have difficulties in payment.



**Analysis Conclusion :**

→ Clients Occupation Type plays a major role in deciding whether a customer can repay a loan or not. We can see like IT Staff, Business, Managers etc. have less difficulty in payment whereas Laborers, cleaning staff, Drivers they have difficulty in repaying the loan.

→ Banks focus should be more on contract type ‘Student’, ‘Pensioner’ and ‘Businessman’ with housing ‘type other than ‘Co-op apartment’ for successful payments.

→ Analysis says that clients having income type Businessman, Maternity leave, Pensioner and Student are less likely tending to applying for the loan. Where as Working and commercial associates are the highest to apply for the loan.

→ Banks focus should be less on income type ‘Working’ as they are having most number of unsuccessful payments.