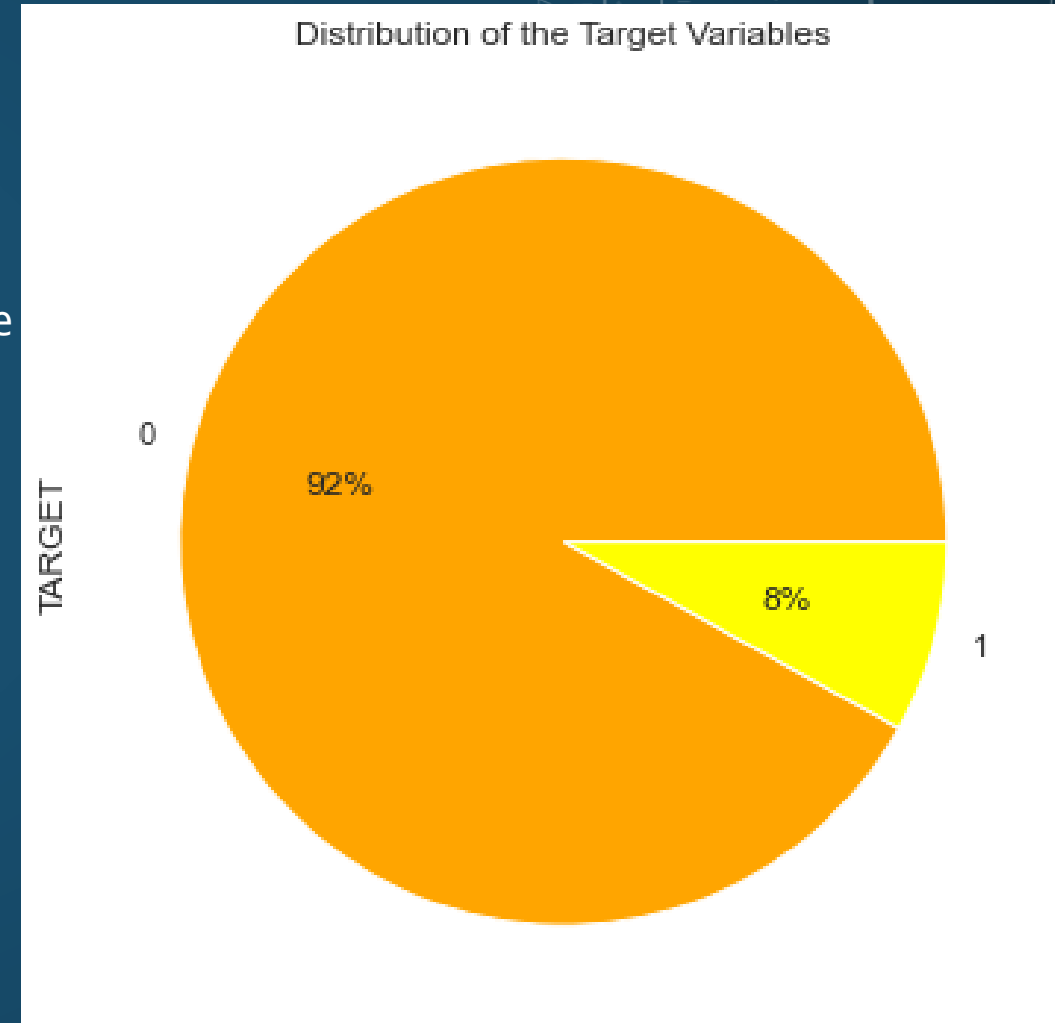# CREDIT EDA CASE STUDY

By-
- ❖ Vipin Panthri
- ❖ Abhisek De

# DISTRIBUTION OF TARGET VARIABLES

- We observe that the target variable 0 is dominating over target Variable 1.

- Which means that the number of defaulter people is less than the people who are non defaulters.

- **The imbalance ratio here is 11.38**



Distribution of the Target Variables

## Distribution of Income range

Points to be concluded from the graph on the right side.
- The high salary brackets are mostly male.
- People having salary<40 have a female majority who are paying on time.
- People having salary>60 have a male majority who are paying on time.
- After 600 income, the count of people are less.



Distribution Curve for Total income for Target Variable 0
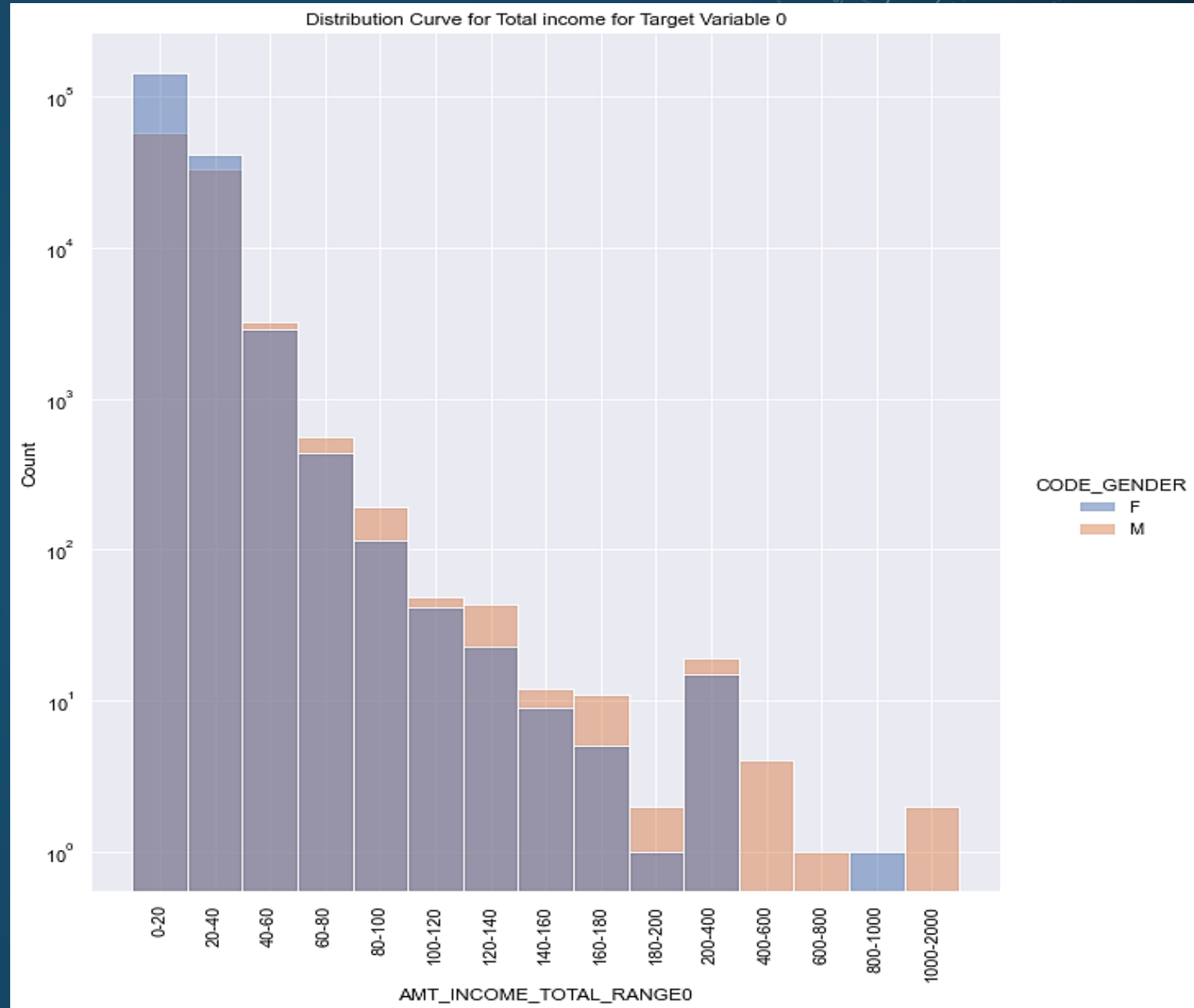
# DISTRIBUTION OF CREDIT AMOUNT
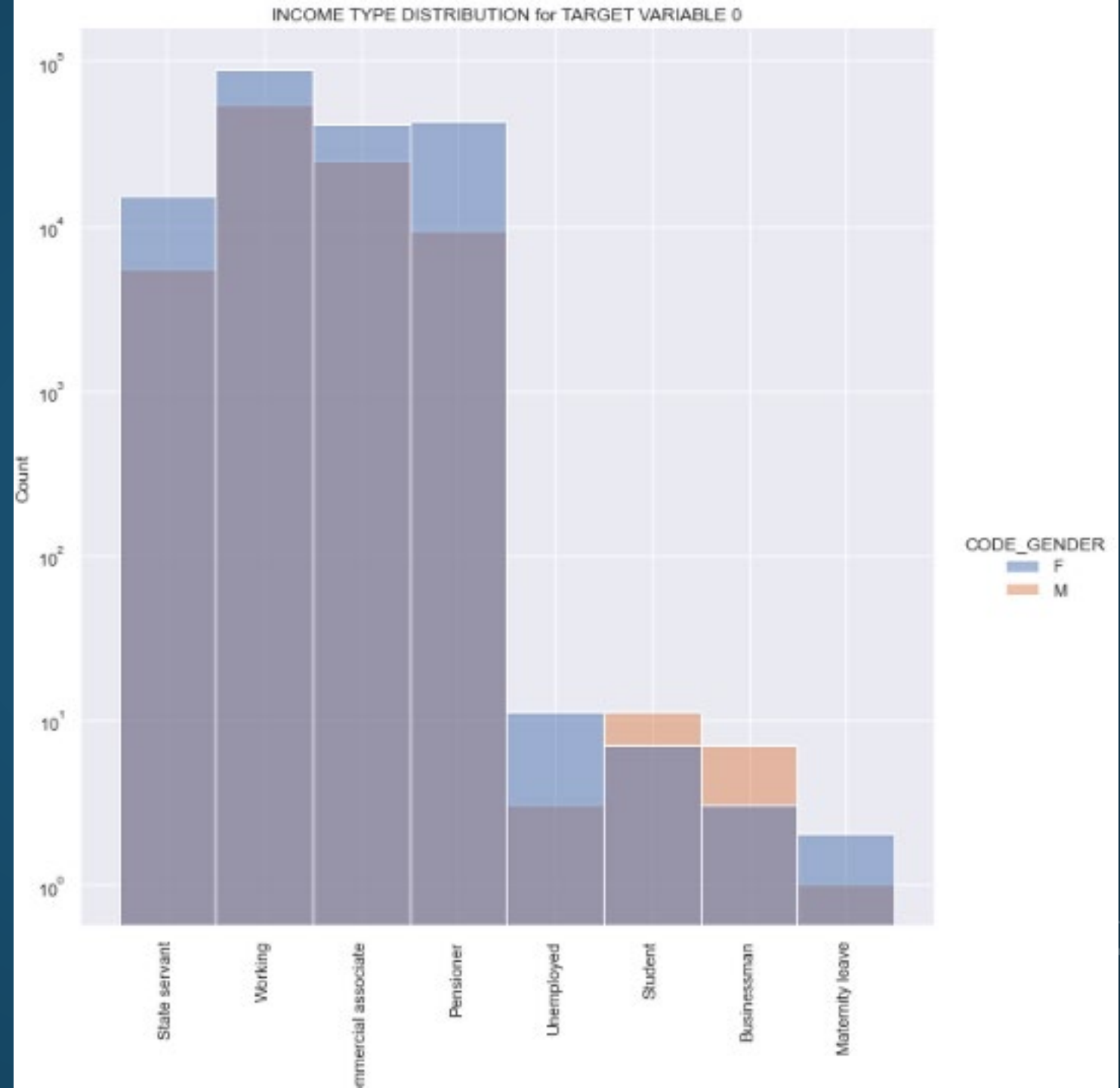
Inferences Drawn:

- The females are greater in count than the males in timely payments.
- Amongst the higher credit range bracket that is above 270 the males are more than female in timely payments.



Distribution Curve for Total credit amount for target 0
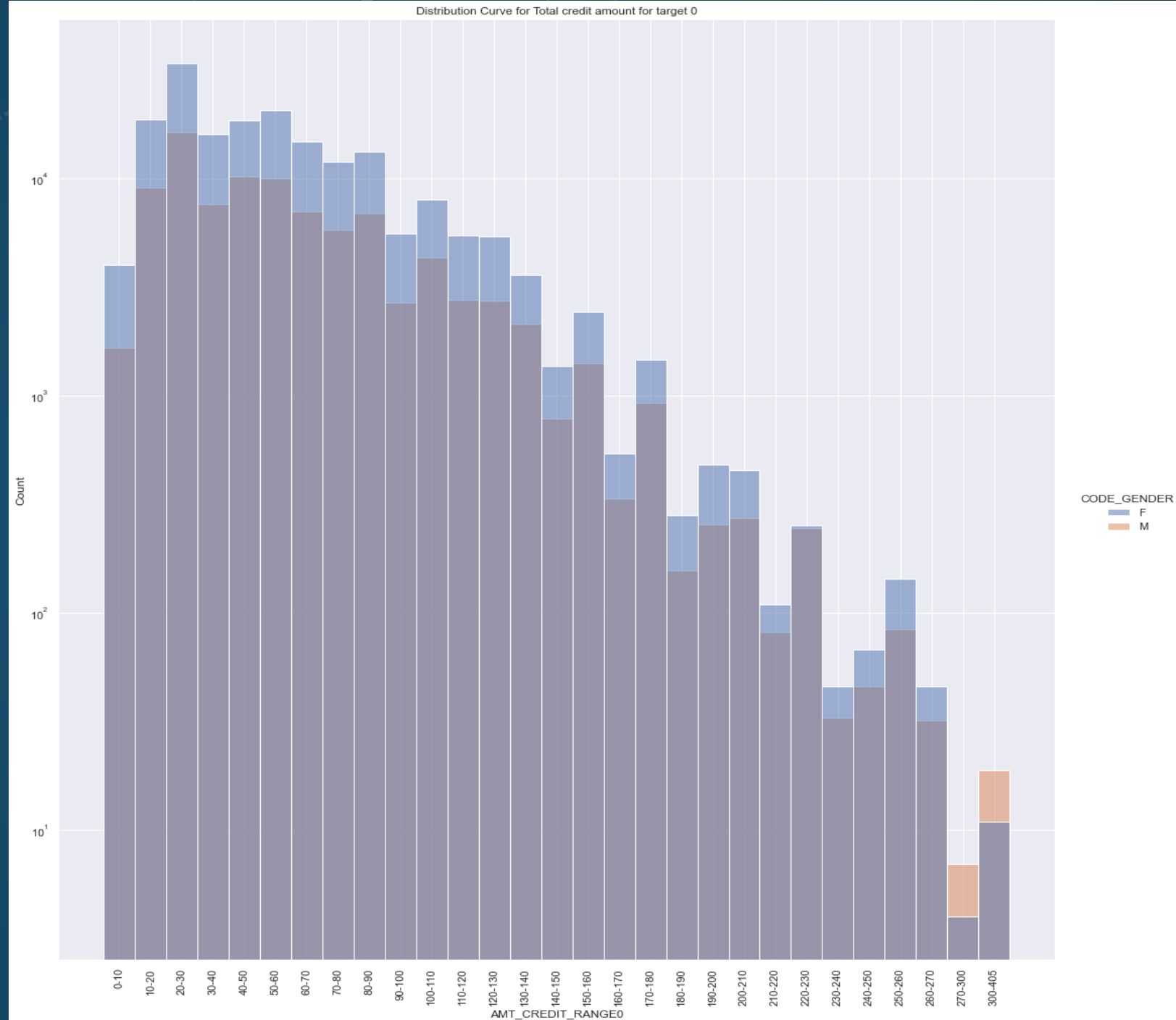
# DISTRIBUTION OF INCOME TYPE

Inference drawn:

- The people are higher from "Working", "Commercial associate", "Pensioner", and "State Servant" income type.

- Less number of people are from "Maternity Leaves", "Businessman" and "Student".

- Females are dominant over males in taking loans in "Student" and "Businessman" Income type

# DISTRIBUTION FOR CONTRACT TYPE

Inferences Drawn:

- Majority of the people who don't have late payments are taking cash loans rather than revolving loans.
- Female dominates over men in both type of loans.
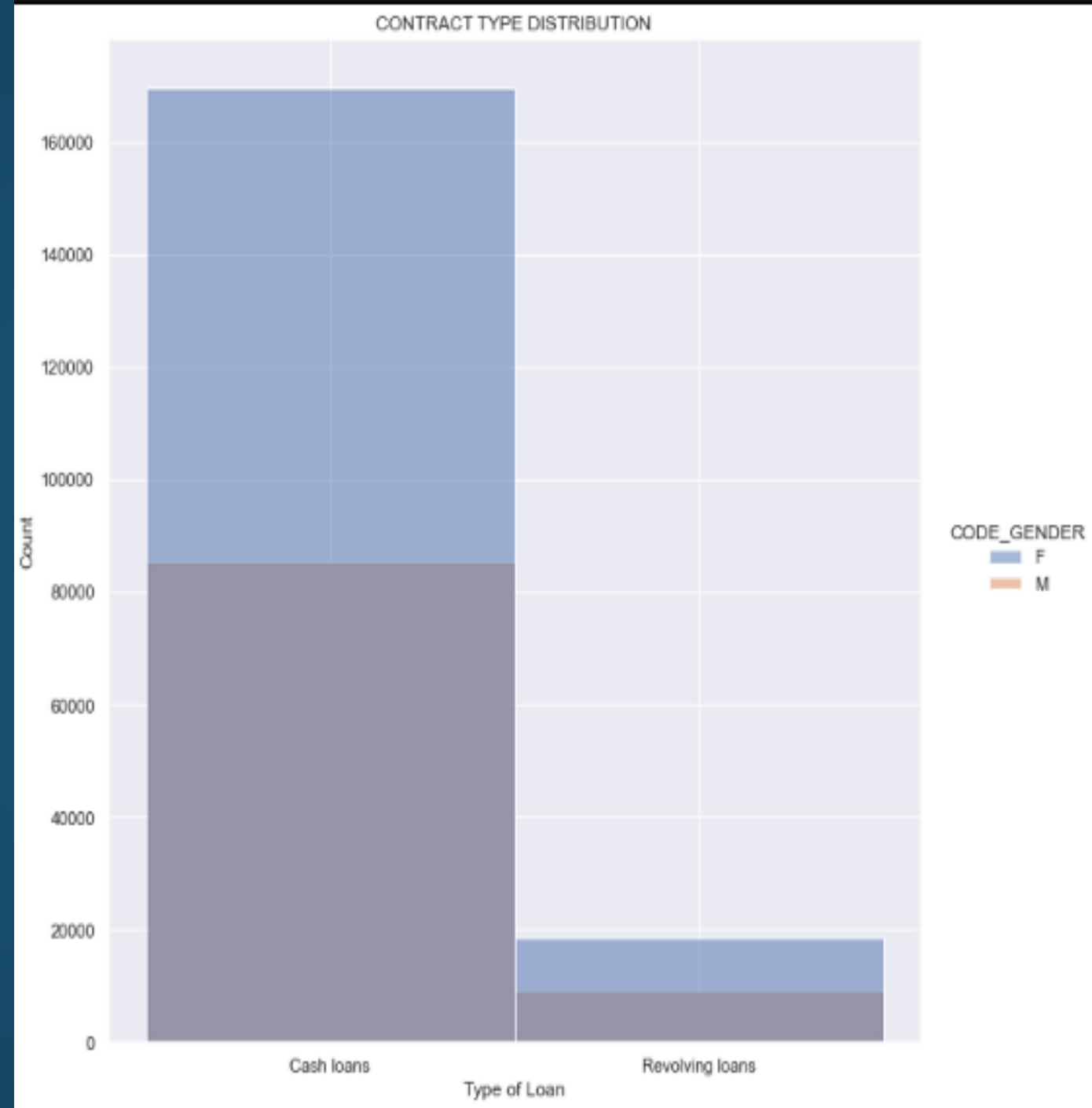
# DISTRIBUTION OF ORGANIZATION TYPE ON A LOG SCALE

Inferences Drawn:

- The people who have paid their loans on time are majorly from Business Type 3, Self-Employed, Other, Medicine and unknown whose data are not mentioned in the data probably because those customers has not filled in this data.
- There are less customers from Industry type 8,type 10, type 13, religion and trade type 5, type 4



ORGANIZATION TYPE DISTRIBUTION
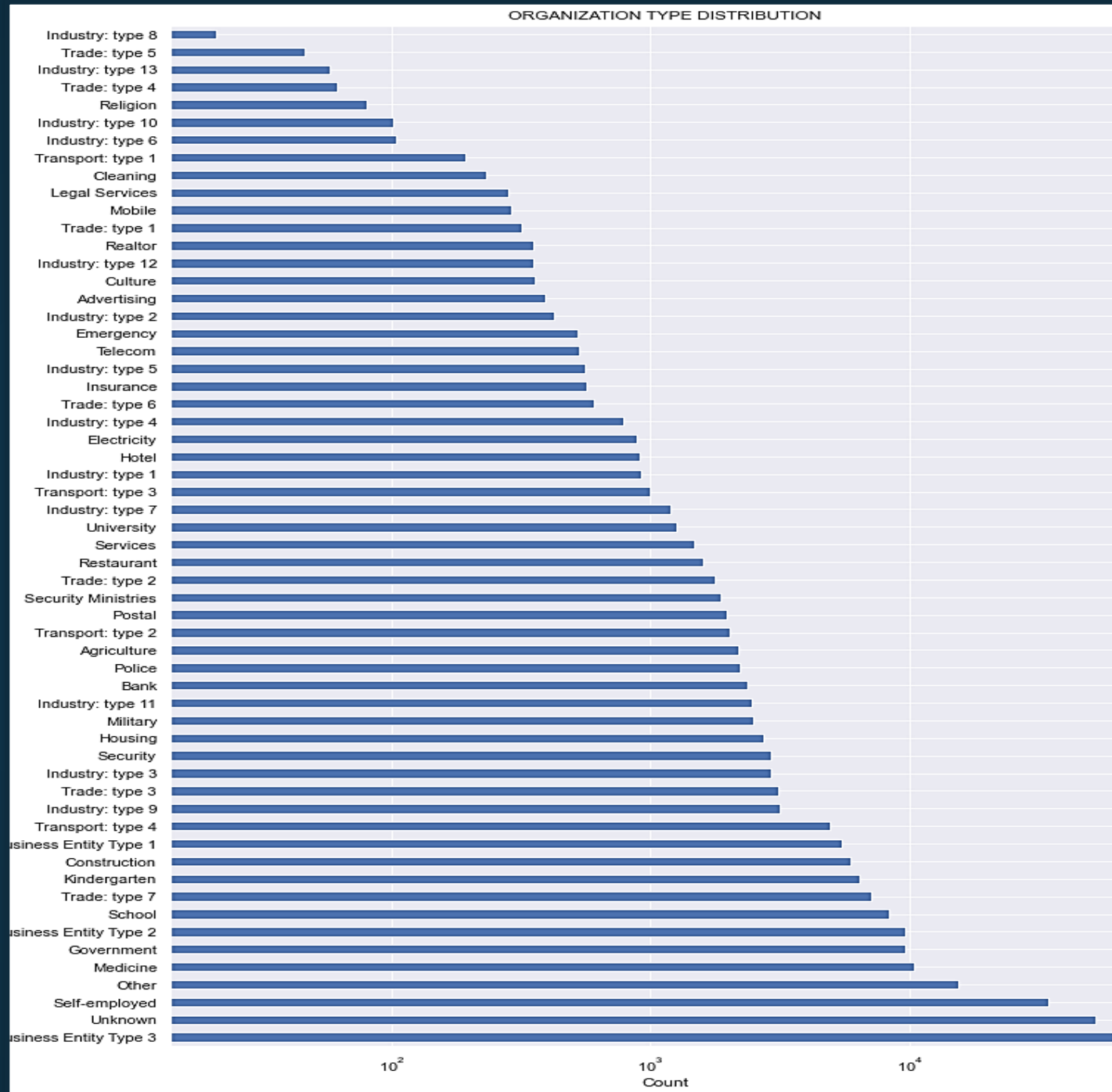
# CATEGORICAL UNIVARIATE ANALYSIS FOR VARIABLES TARGET 0

## Boxplot for Total Income Amount

- Few points can be concluded from the graph.
- Some outliers are noticed in income amount.
- The third quartiles is very slim for income amount.
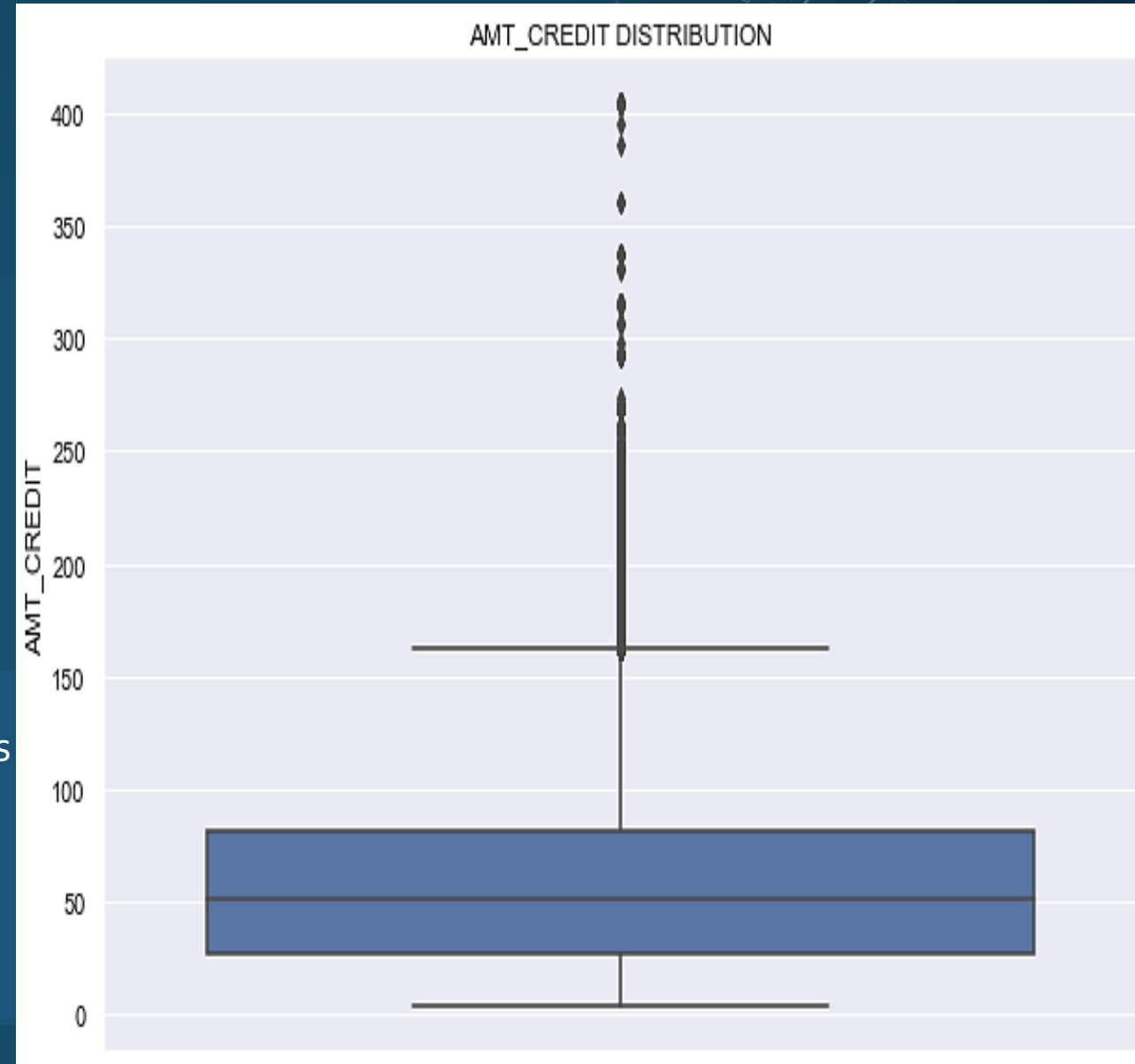


AMT_INCOME_TOTAL DISTRIBUTION

# BOXPLOT FOR CREDIT AMOUNT

Inferences Drawn:

- There are outliers in the "AMT_CREDIT" column.

- The third quartile is wider than the first quartile indicating that most of the values are in the third quartile.

- 75% of the people who had late payments have credits less than 81.



AMT_CREDIT DISTRIBUTION

# Boxplot For Annuity Amount

Inferences Drawn:

- There are outliers present in this column.
- 75% of the people who had late payments have annuity less than or equal to 34.79K
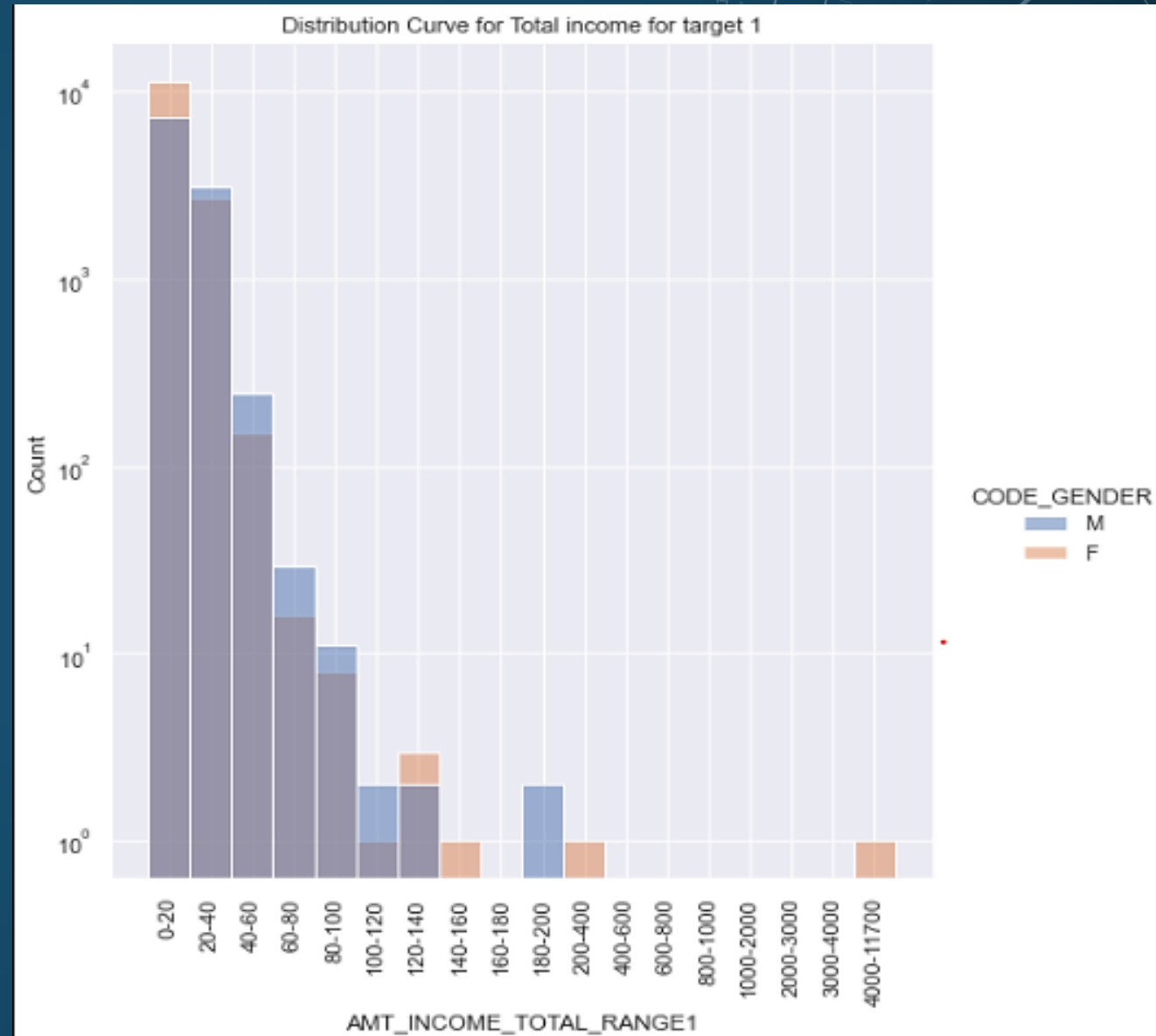- A good amount of people have high annuity amount.
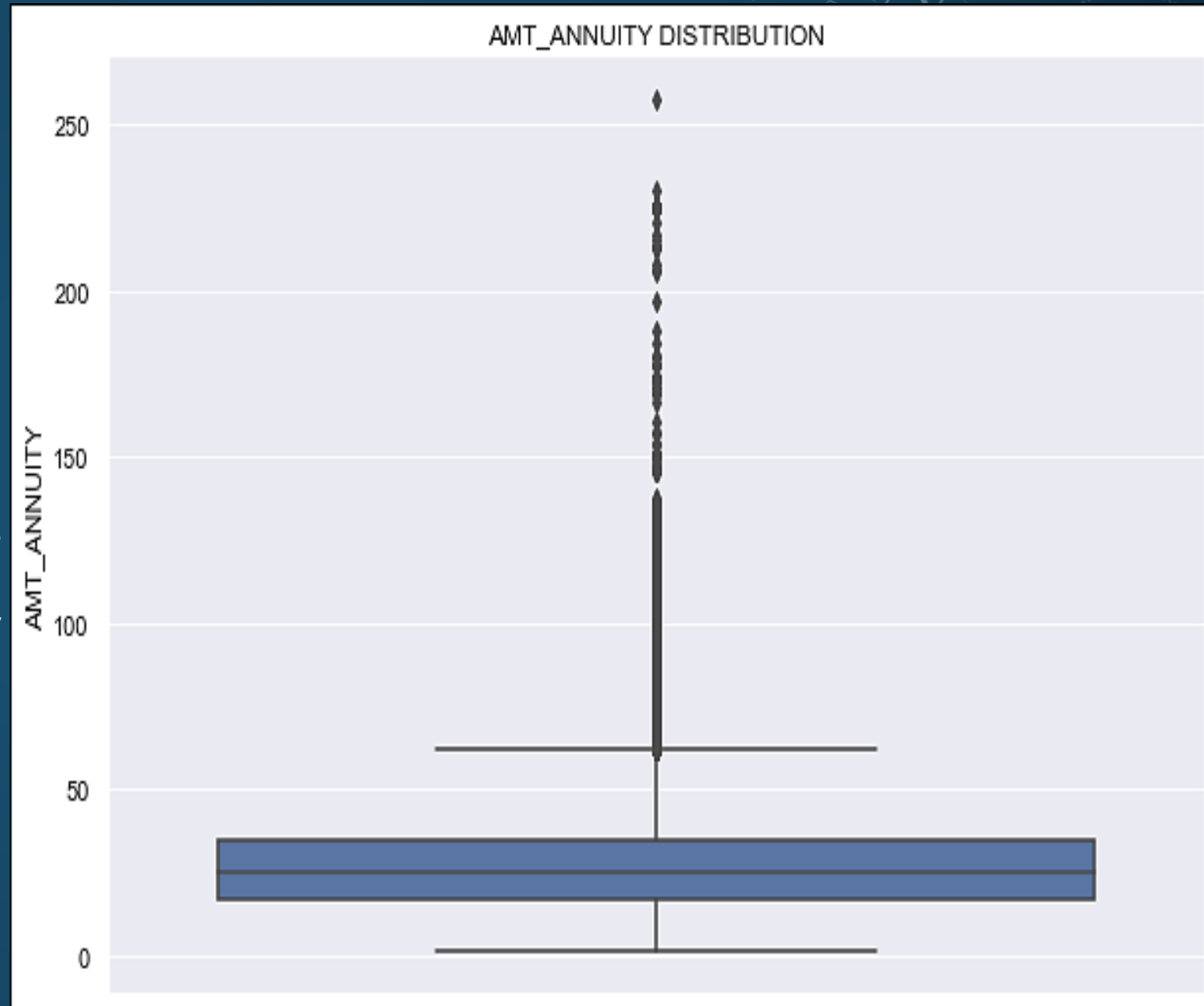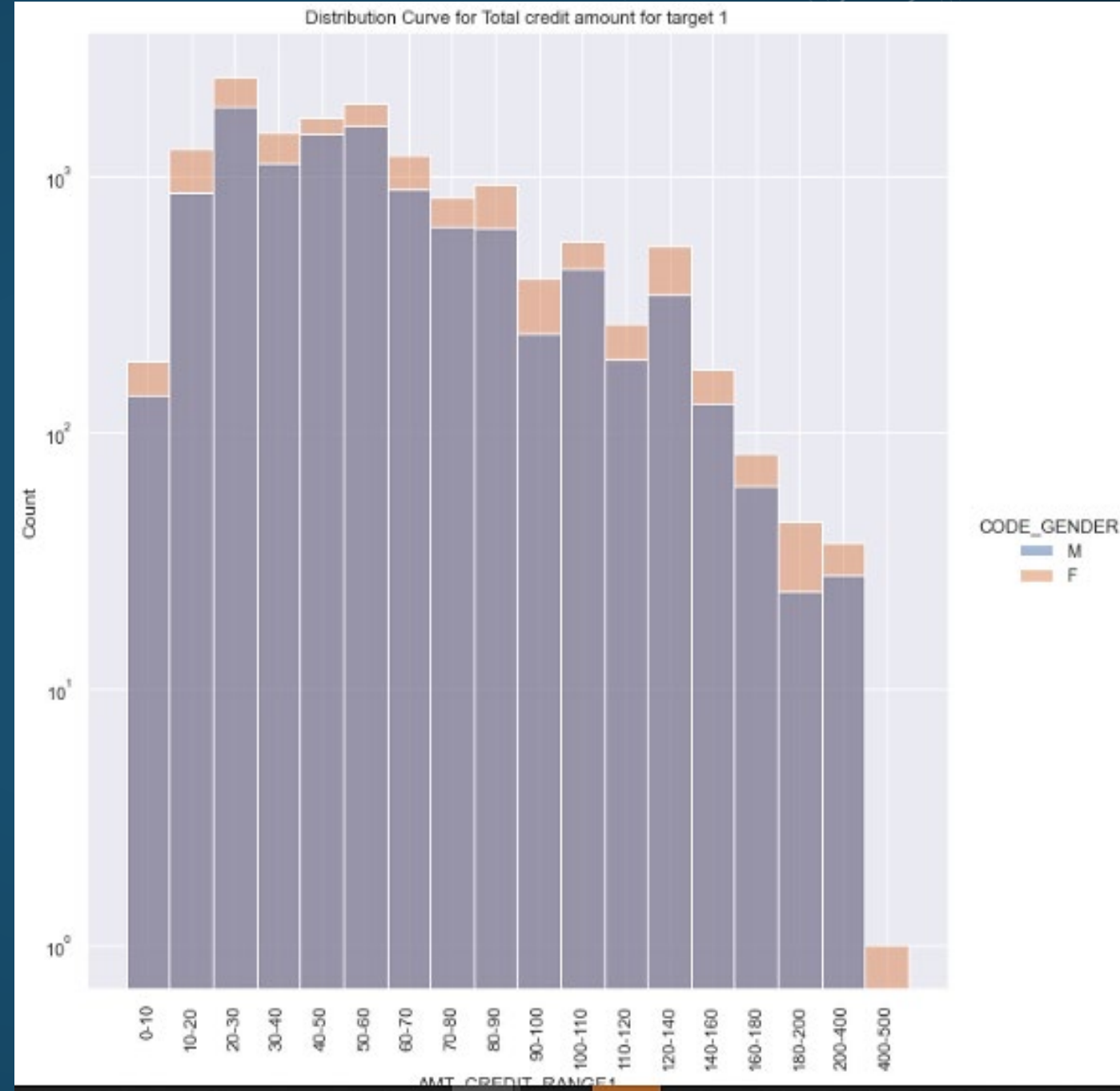
## Distribution Of Income Range

Inferences Drawn:

- Males Count is greater than female amongst those people who have delayed payments.
- The count is less for salary>60. In that part males are more.
- More of females having salary<20 are having late payment problems.



Distribution Curve for Total income for target 1

# DISTRIBUTION OF CREDIT AMOUNT
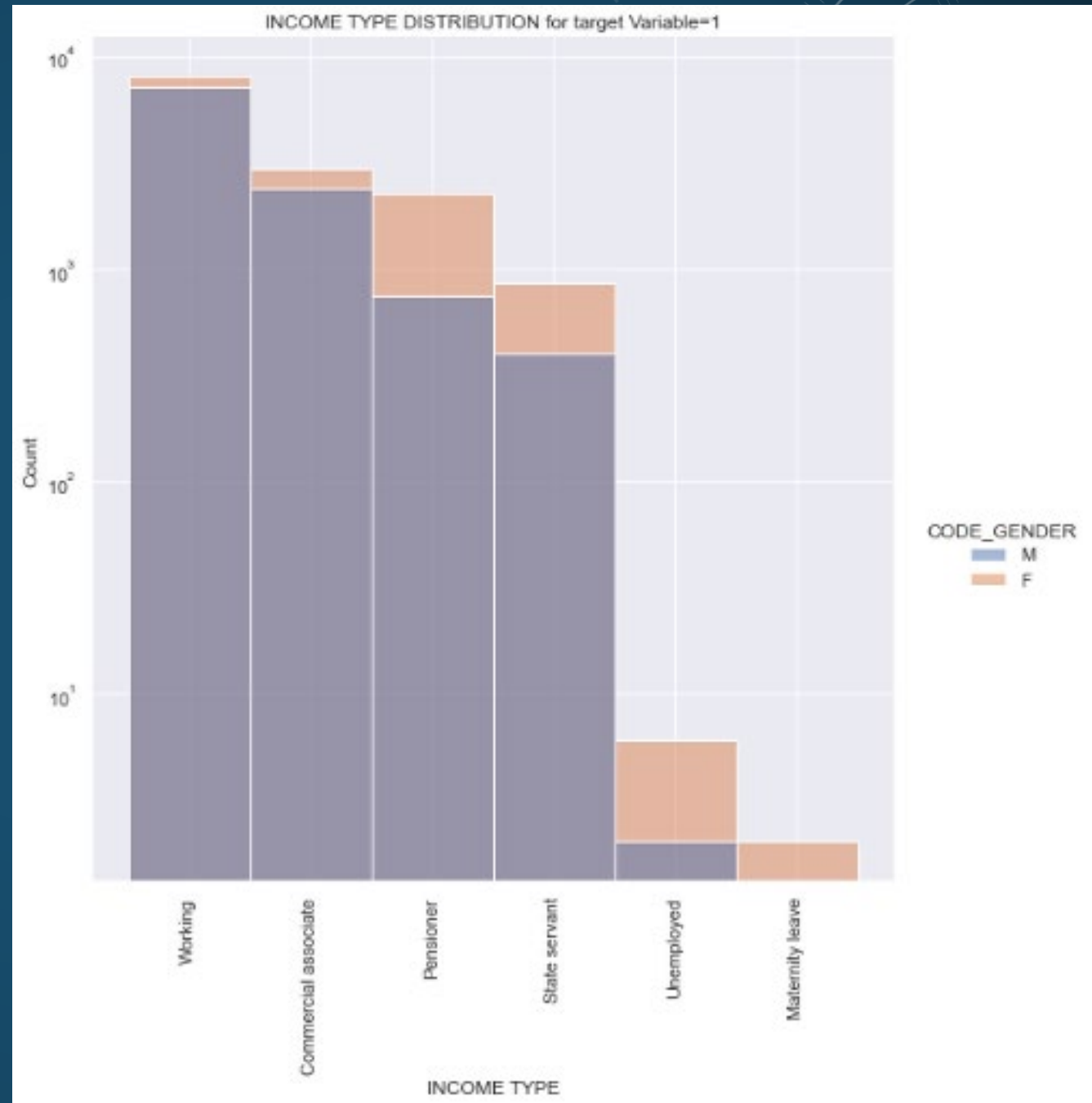
Inference Drawn:

- The delayed payments is the highest amongst the people who have taken a credit loan from the bracket 20-30 followed by 50-60 bracket.
- Females are predominantly more in this segment.



Distribution Curve for Total credit amount for target 1

# DISTRIBUTION OF INCOME TYPE

Inferences Drawn:

- We see that the Working, Commercial Associate, Pensioner are higher than the Maternity and the Unemployed Columns.
- "Students" and "Businessman" categories are not available which means that that these categories of people are paying their loans on time. Hence loan to such categories of people would be apt.
- We see that the females are higher in number than males in taking loans and in all income types.

# DISTRIBUTION FOR CONTRACT TYPE AND COMPARING IT WITH THE PREVIOUS ONE

Inferences Drawn:

- Here also, people who have made late payments are mostly taking cash loans more than revolving loans.
- When we compare the above 2 graphs, we see that the no of people who have taken revolving loans and cash loans is more with the people who have paid their payments on time than the late payments people.
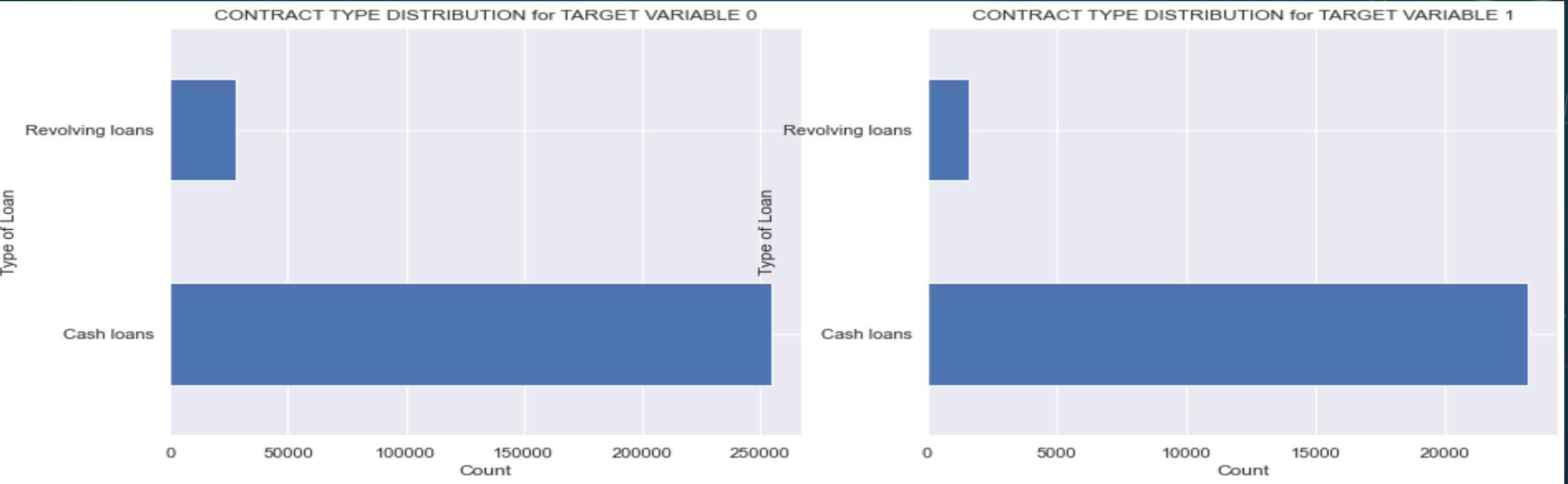
# DISTRIBUTION OF ORGANIZATION TYPE ON A LOG SCALE

Inferences Drawn:

- The people who have a history of late payments are frequently from Business Entity: Type 3, Self-Employed, Other, Unknown and Business Entity: Type 2.
- The people in the Business Entity: Type 3, Self-Employed are < ($10^4$) in late payments but > ($10^4$) in timely payments. So a majority of the people in these groups pay on time than make late payments.
- It's more or less same like Target Variable 0.



ORGANIZATION TYPE DISTRIBUTION for target Variable=1
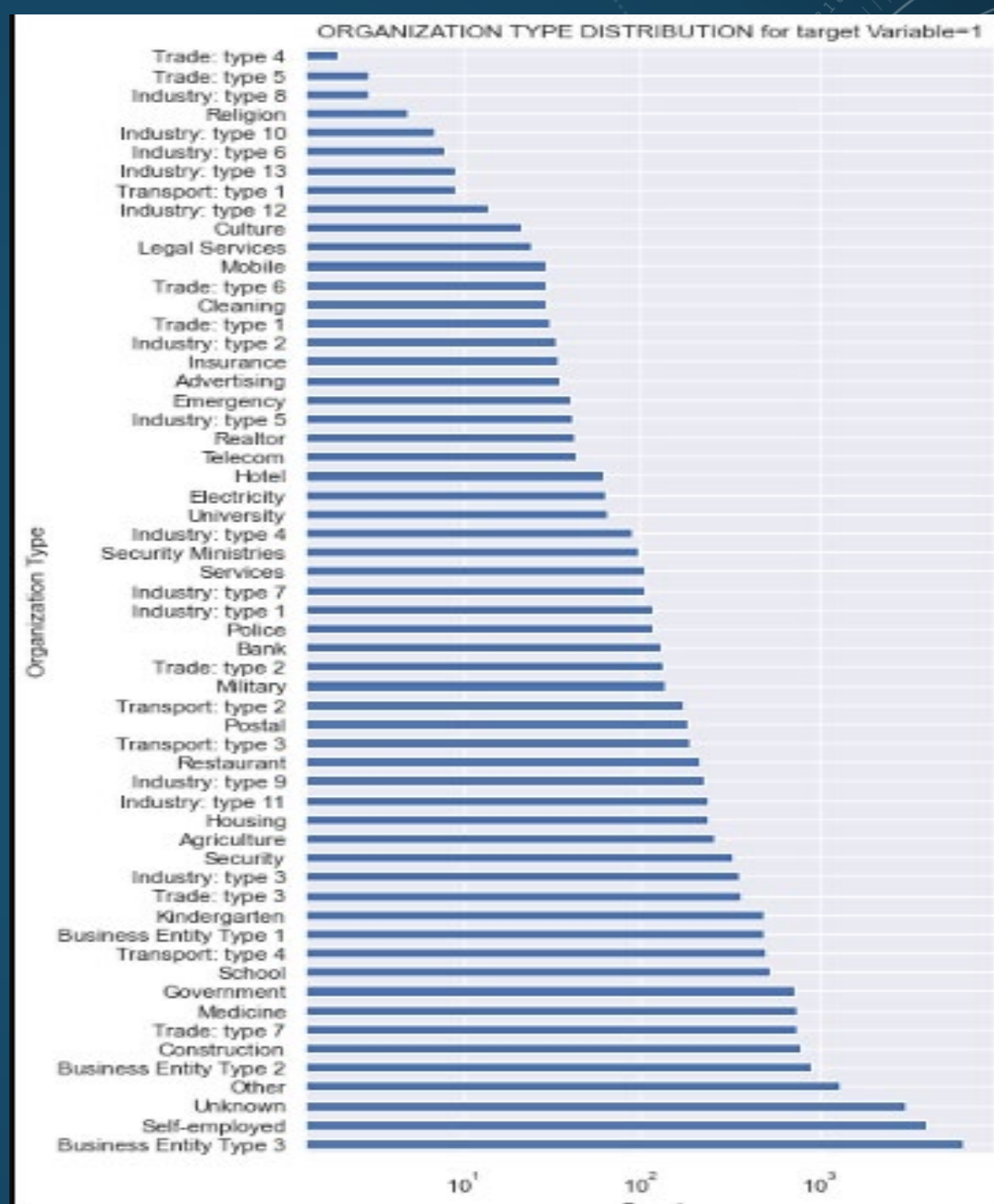
# CATEGORICAL UNIVARIATE ANALYSIS FOR VARIABLES TARGET 1

## Boxplot for Total Income Amount

Inferences Drawn:

- There are outliers in this column.
- The first quartile is less wider as compared to the third quartile.
- Most of the people have income in the third quartile.



AMT_INCOME_TOTAL DISTRIBUTION OF TARGET VARIABLE 1
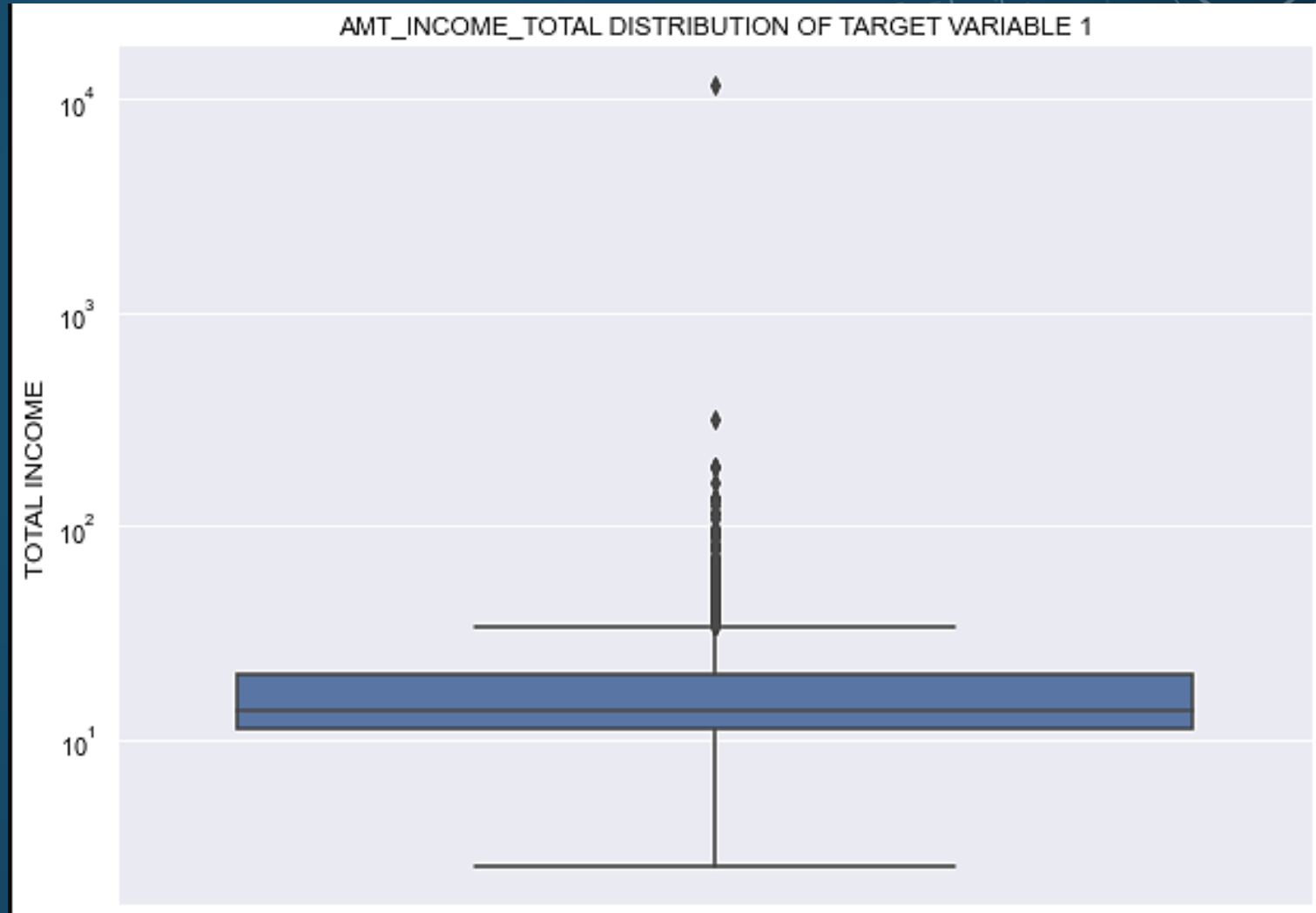
# BOXPLOT FOR CREDIT AMOUNT

Inferences Drawn:

- There are outliers in this column.
- The third quartile is slightly wider than the first quartile.



AMT_CREDIT DISTRIBUTION OF TARGET VARIABLE 1

# Boxplot For Annuity Amount

Inferences Drawn:

- There are outliers in this column.
- The first and the third quartile is quite similar.



AMT_ANNUITY DISTRIBUTION OF TARGET VARIABLE 1
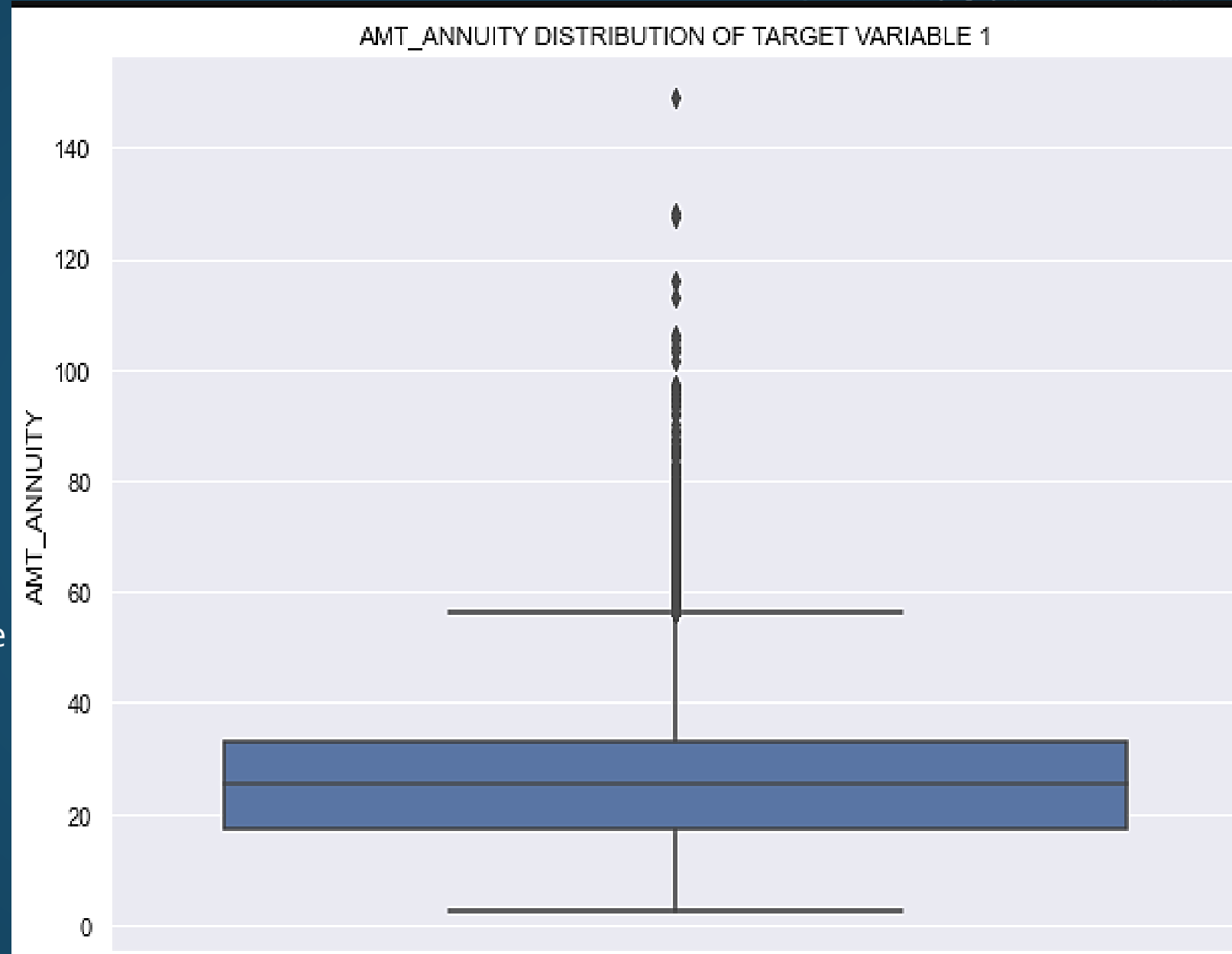
# BIVARIATE ANALYSIS FOR TYPE 0

## Credit amount vs Total Income Range Scatter Plot

Inferences Drawn:

- According to the scatter plot, it is evident that the people having income of 0-140 are actively taking loans be it big or small and also are able to pay the loan in time.
- There are more females than males in the set of people who have been taking successful loans.
- Percentage of males taking a higher amount of loan(>300) is almost equal to the percentage of women taking the same amount of loan.



Credit Amount v/s Total Income for target variable0
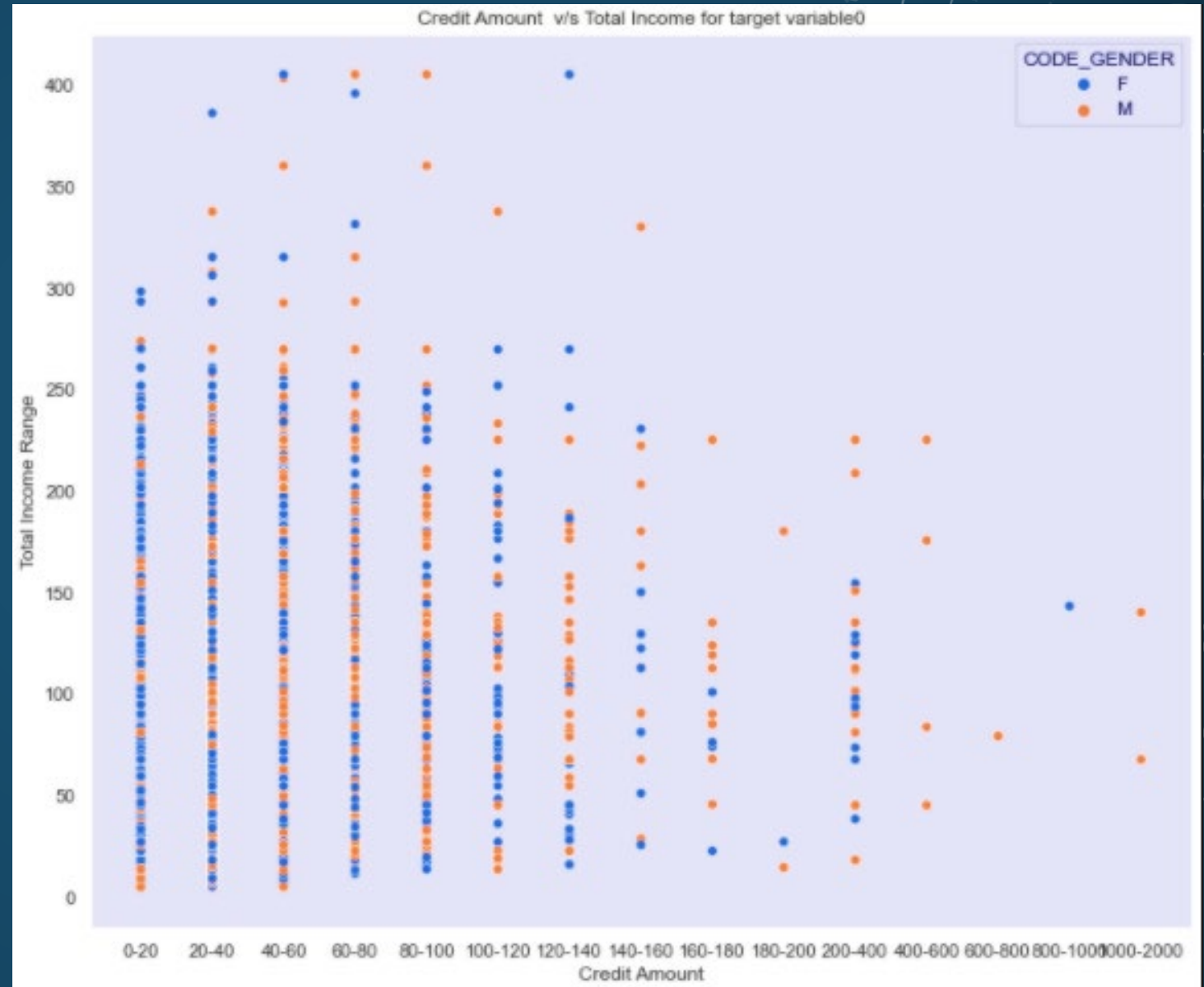
# Bar plot for Total Income v/s Income Type for Target Variable 0

- A bar plot shows only the mean(default) or other estimator value. So the Y- Axis isn't the count of the people in the respective Income type brackets. It just represents an estimate of central tendency for a numeric variable.

- Inferences Drawn:

- Highest Income is there in the the type "Businessman"
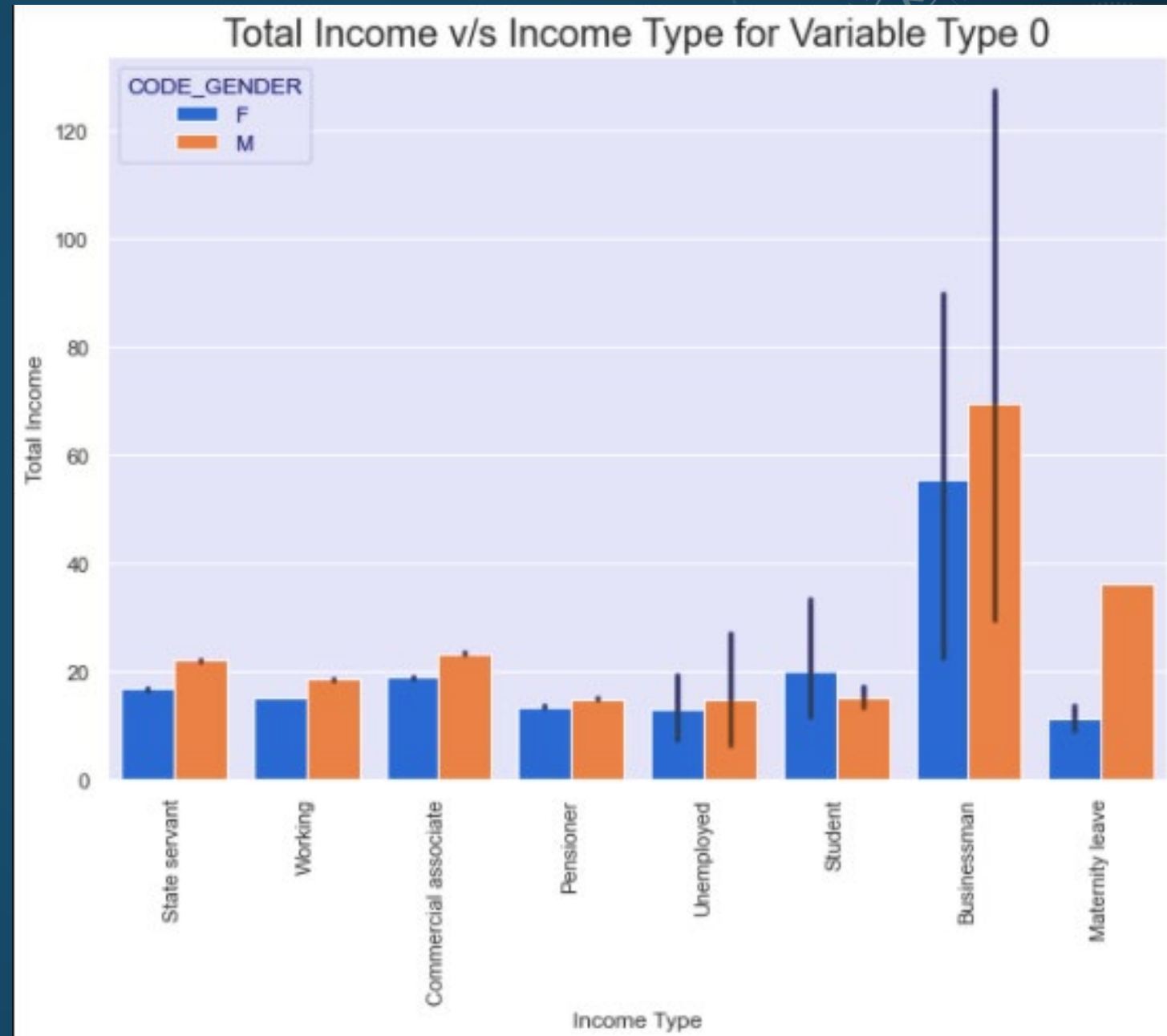
## CREDIT AMOUNT VS EDUCATION STATUS FOR TARGET VARIABLE=0

Inferences Drawn:

- Married people in any type of education status has the maximum credit amount.
- Family status of "Civil marriage" for "Academic degree" is having most of the credit amount in the third quartile.
- Family status of "Married", "Civil Marriage" and "Seperated" have higher amount of credits than any other set of people in that particular education status bracket.
- Family Status of "Married", "Single/Not Married", "Civil Marriage" are having high outliers which means that the credit amount is high in "Higher Education"
- There are no extreme values in family status of "Civil Marriage" and "Seperated" in the Education status of "Academic Degree".



Credit amount w/s Education Status for target variable0
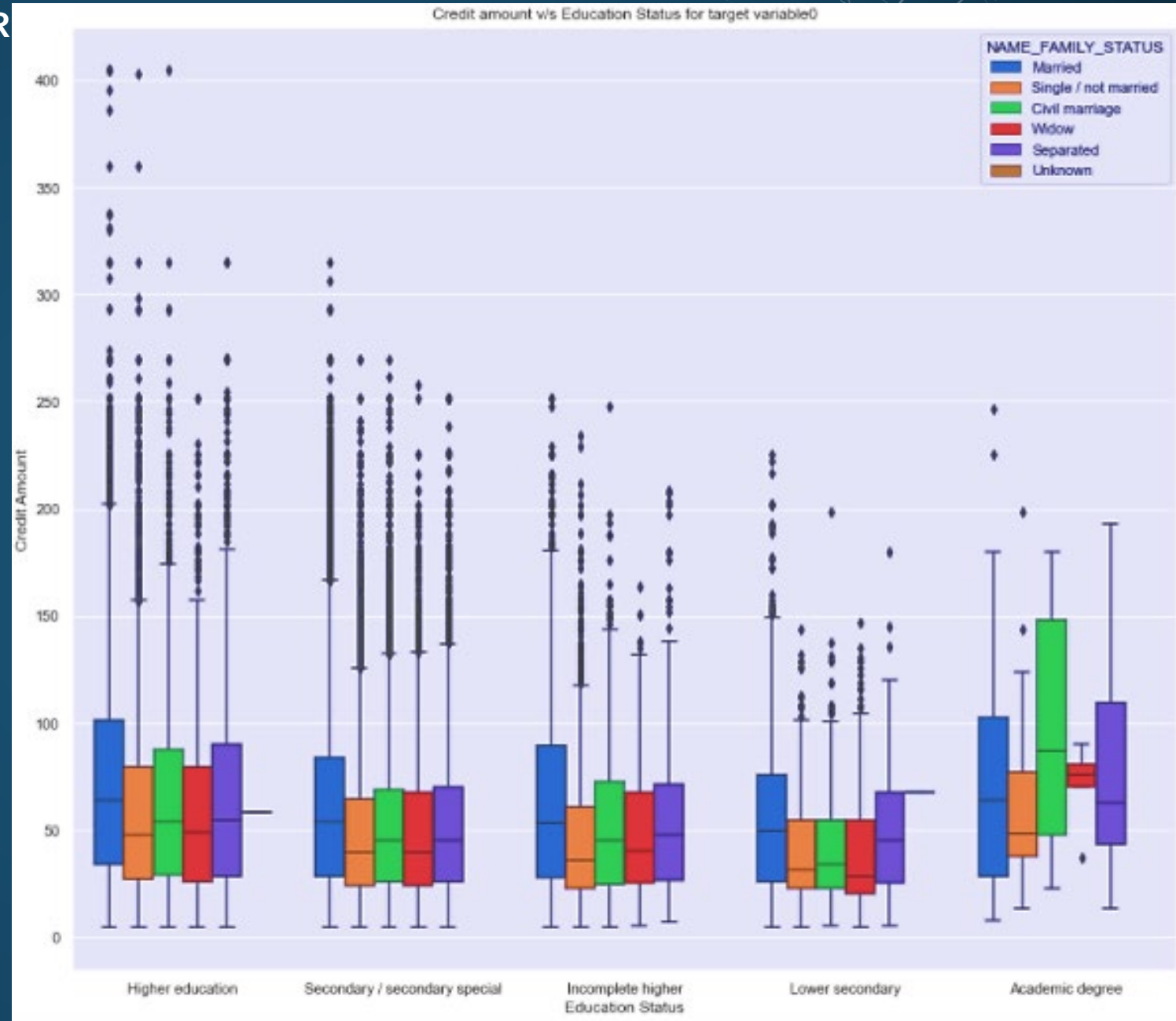
# INCOME AMOUNT VS EDUCATION STATUS FOR TARGET VARIABLE=0
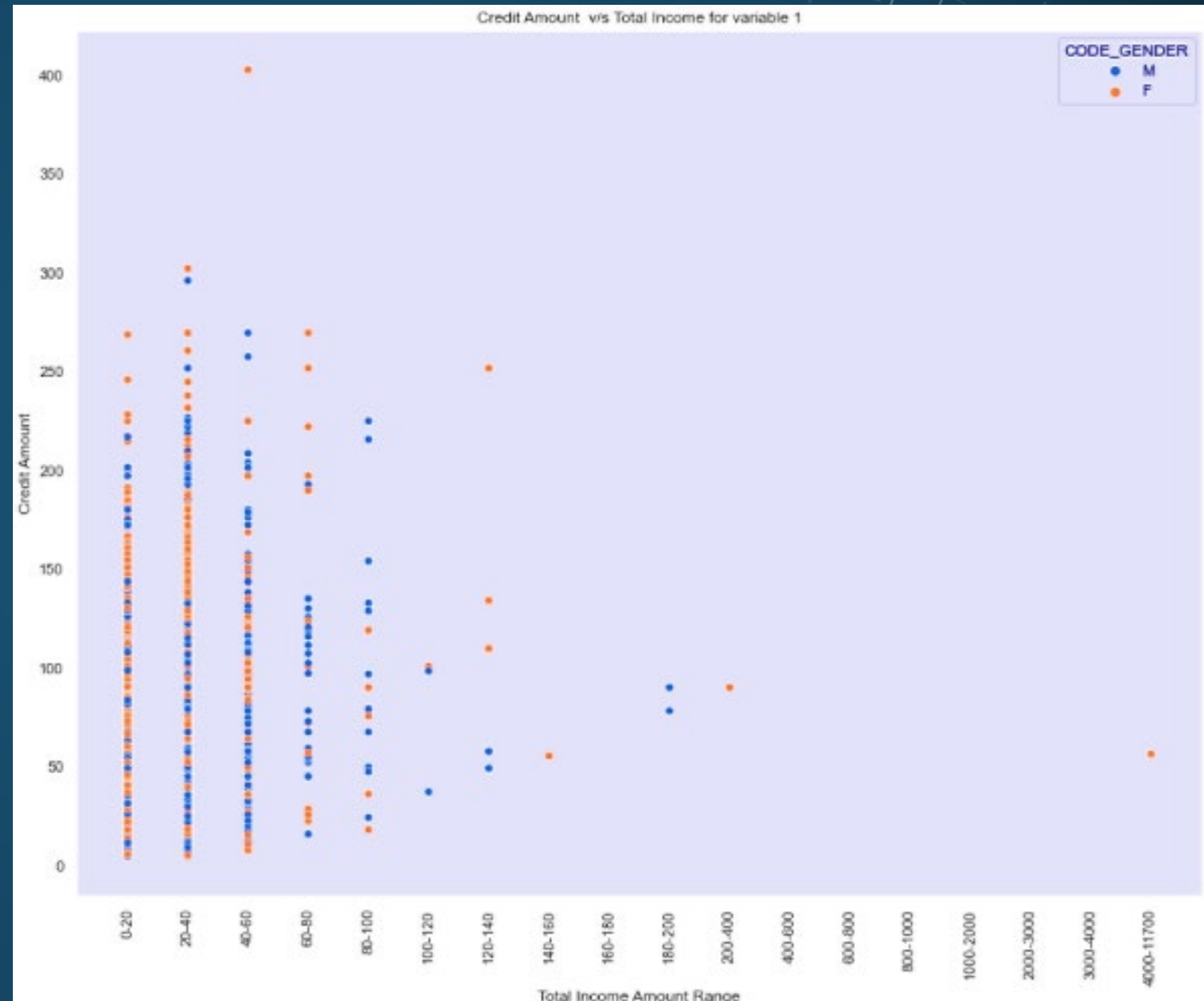
Inferences Drawn:

- Family status of "Married" having "Lower Secondary" education, "Single/Not Married" having "Secondary/Secondary Special" are having the lowest income.
- Married people in "Higher Education" and "Secondary/Secondary Special" are having higher income in their respective Education Status brackets.
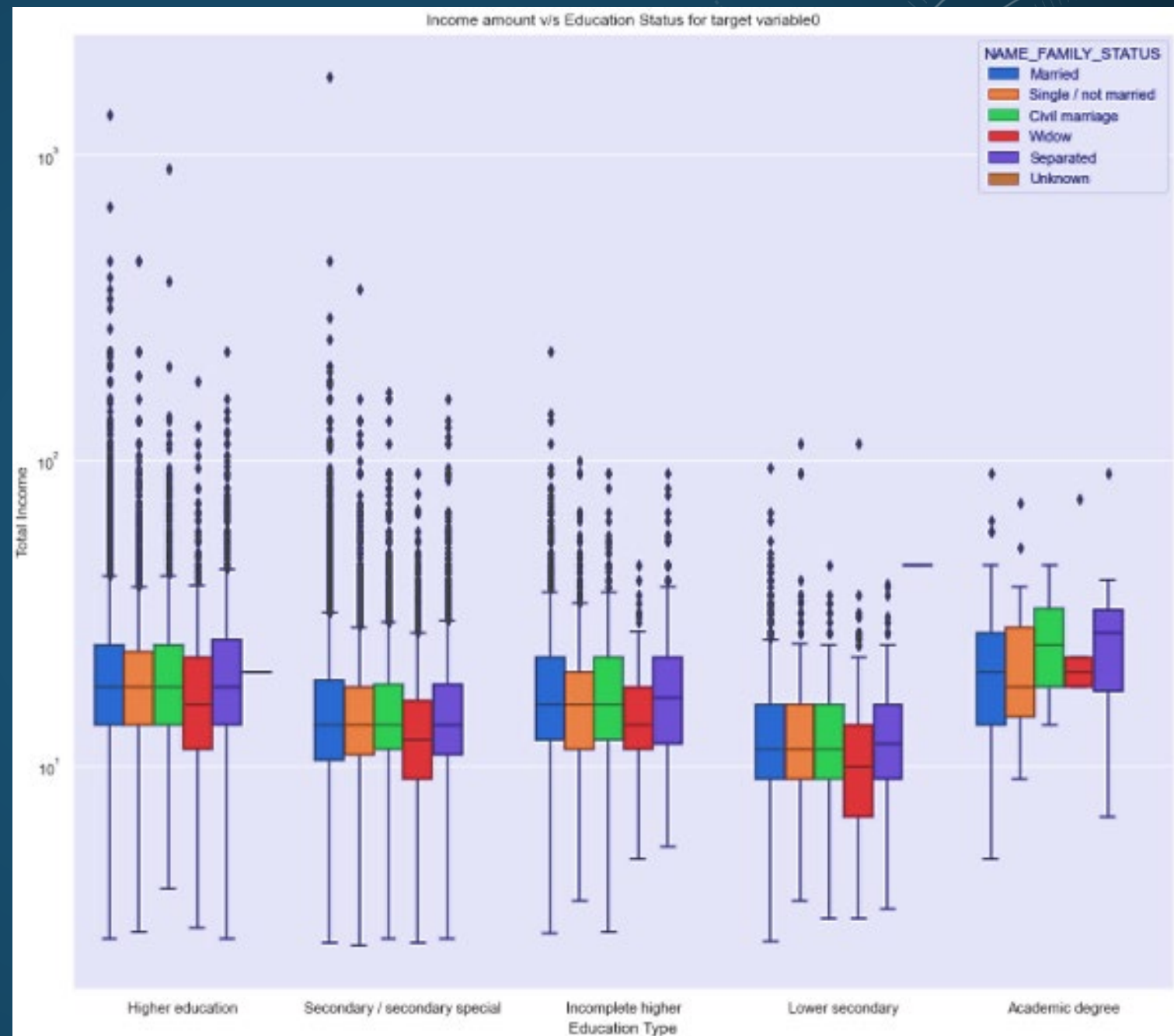
# BIVARIATE ANALYSIS FOR TYPE 0
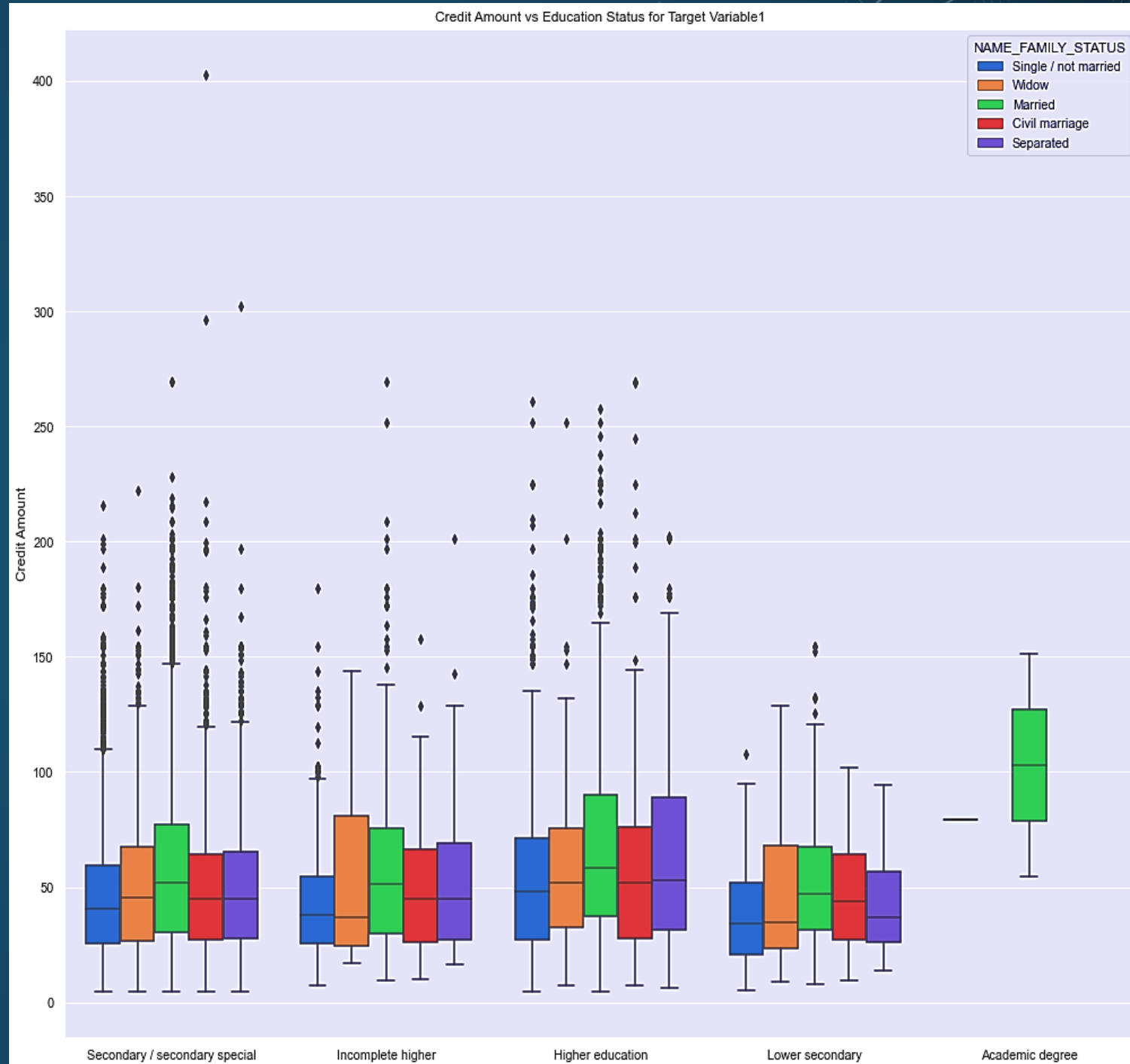
Inferences Drawn:

- There are less values in this category than the previous one(Target=0)
- One major difference between the above plot vs that of target variable 0 is that the Total Income Amount is that majority of these people have low to moderate to high credit amount values till 300 in the Total Income Amount Range of 0-100.
- Many people who have less Total Income lower than 40 are having loan amount of more than 150 which is the result of them not being able to pay the loan on time.



Credit Amount v/s Total Income for variable 1

## CREDIT AMOUNT VS EDUCATION STATUS FOR TARGET VARIABLE=1

Inferences Drawn:

- All Family status people having Education Status of "Higher Education" are having considerably higher Credit Amount value.
- People having Family Status of "Civil Marriage" in "Higher Education" and "Secondary/Secondary Special" are having many outliers and are high in both of the groups.
- People in "Incomplete Higher" Education are having Credit Amount greater than 90.
- The same goes with "Lower Secondary" Education people as well.



Credit Amount vs Education Status for Target Variable1

# INCOME AMOUNT VS EDUCATION STATUS FOR TARGET VARIABLE=1

Inference Drawn:

- The "Incomplete Higher" Education Status people are having a Total income lesser than 50 but higher credit amount resulting in being a defaulter.
- People having a Family Status of "Widow" are having a small lower quartile range. 25% of them are having Total Income of less than 10 in most cases.



Income amount v/s Education Status
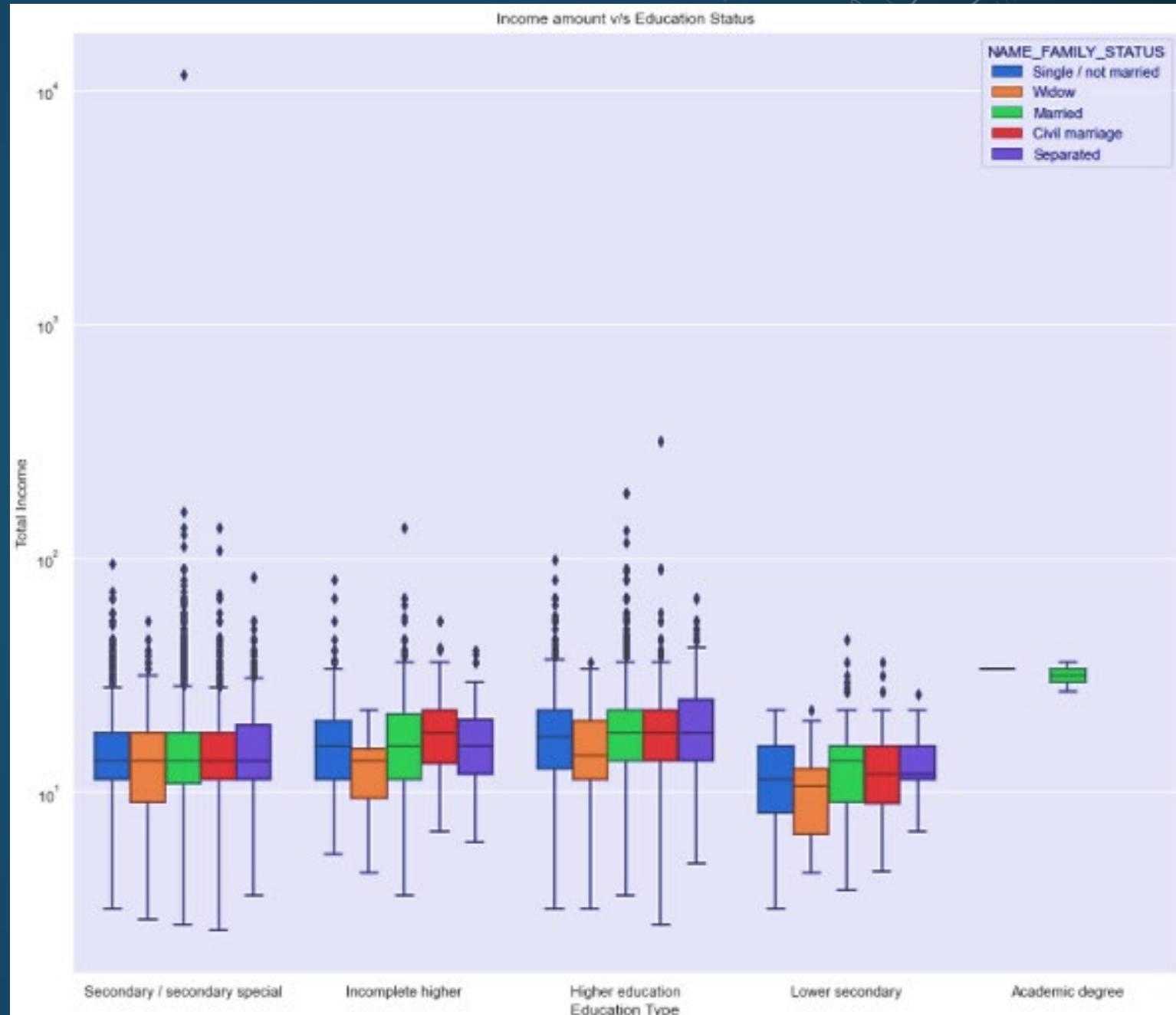
# MULTIVARIATE ANALYSIS OF TARGET VARIABLE=0

Correlation of target 0
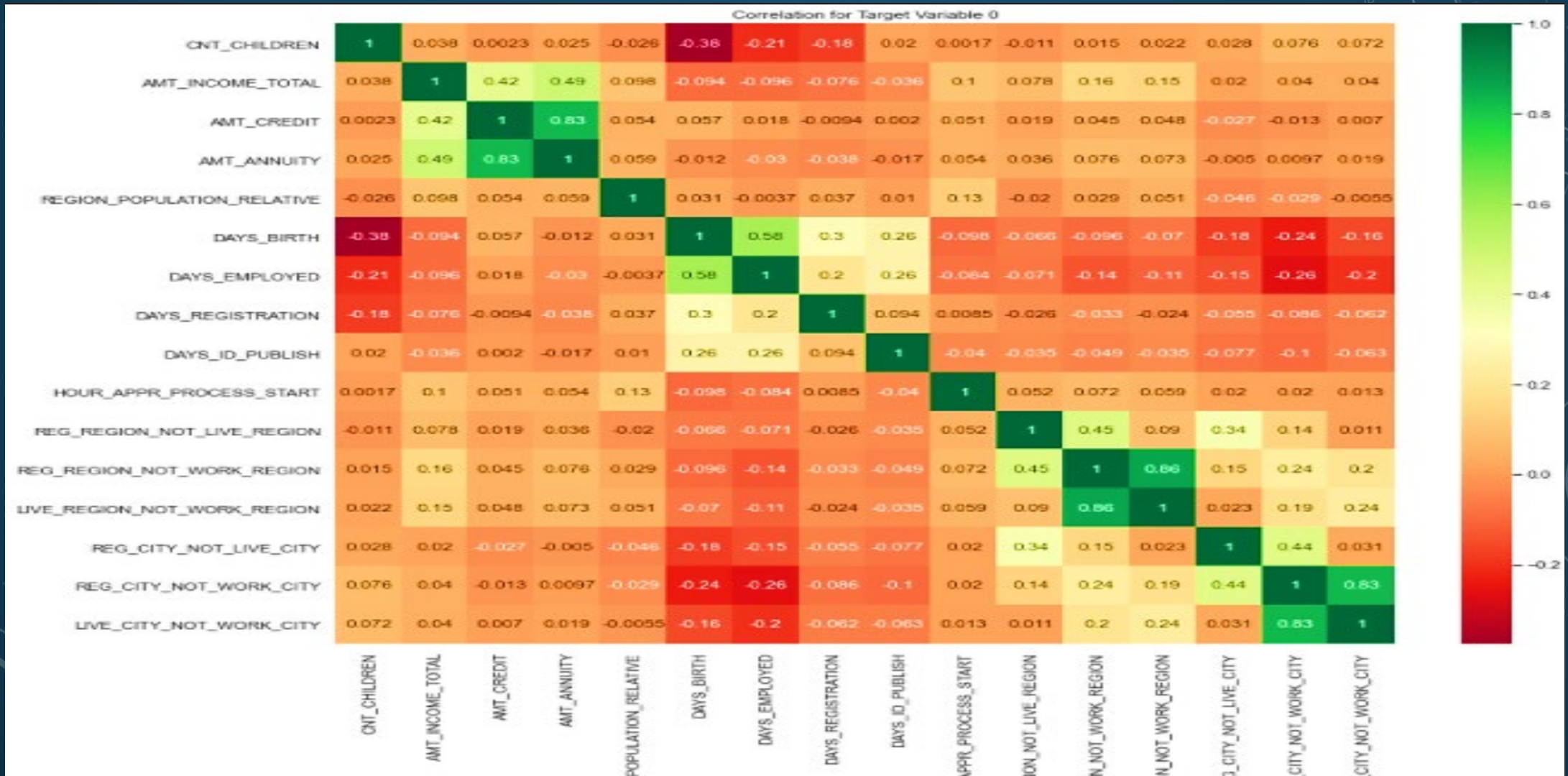
# CORRELATION FOR TARGET 0

Inferences Drawn:

- Income amount is inversely proportional to the number of children client have, which goes to say that more income then the children count will be less and vice-versa.
- Similarly, Credit Amount is inversely proportional to the number of children client have, which goes to say that more Credit amount then the children count will be less.
- Credit Amount is inversely proportional to the date of birth which goes to say that the Credit Amount is higher for lower age and vice-versa.
- The income is higher in densely populated area.
- Credit amount is higher as well in highly populated area.
- These segment of people have less children in highly populated area.

# MULTIVARIATE ANALYSIS OF TARGET VARIABLE=1

Correlation of target 1
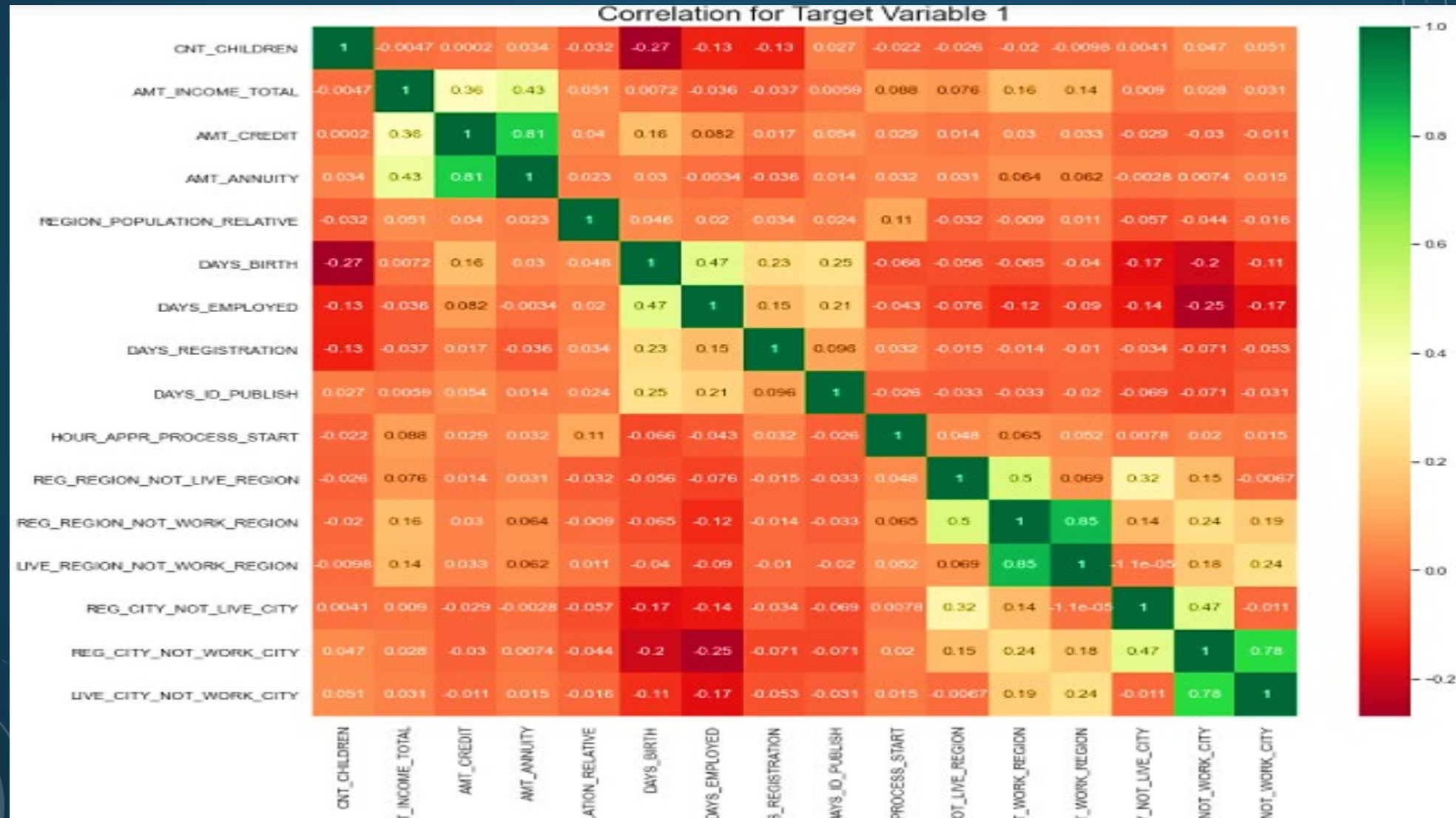
Inferences Drawn:

- Those client's whose permanent address does not match with the contact address are having less children and vice-versa.
- Those client's whose permanent does not match work address are having less children and vice-versa
- Almost similar to the correlation of Target Variable =o

# UNIVARIATE ANALYSIS AFTER MERGING PREVIOUS DATA

Distribution of contract status with purposes

Inferences Drawn:

- The bank has rejected loans frequently of people from 'repairs'.
- Reasons like "Everyday Expenses", "Purchase of "Electronic Equipment" have more approvals than rejections.
- There is a high rate of cancellation in "Repairs" as well.
- For the field education it seems that there is an equal number of approves and rejection
- "Payment of other loans", "Buying a new car" and "Buying a new home" are having higher rejection than approvals.

Distribution of contract status with purposes
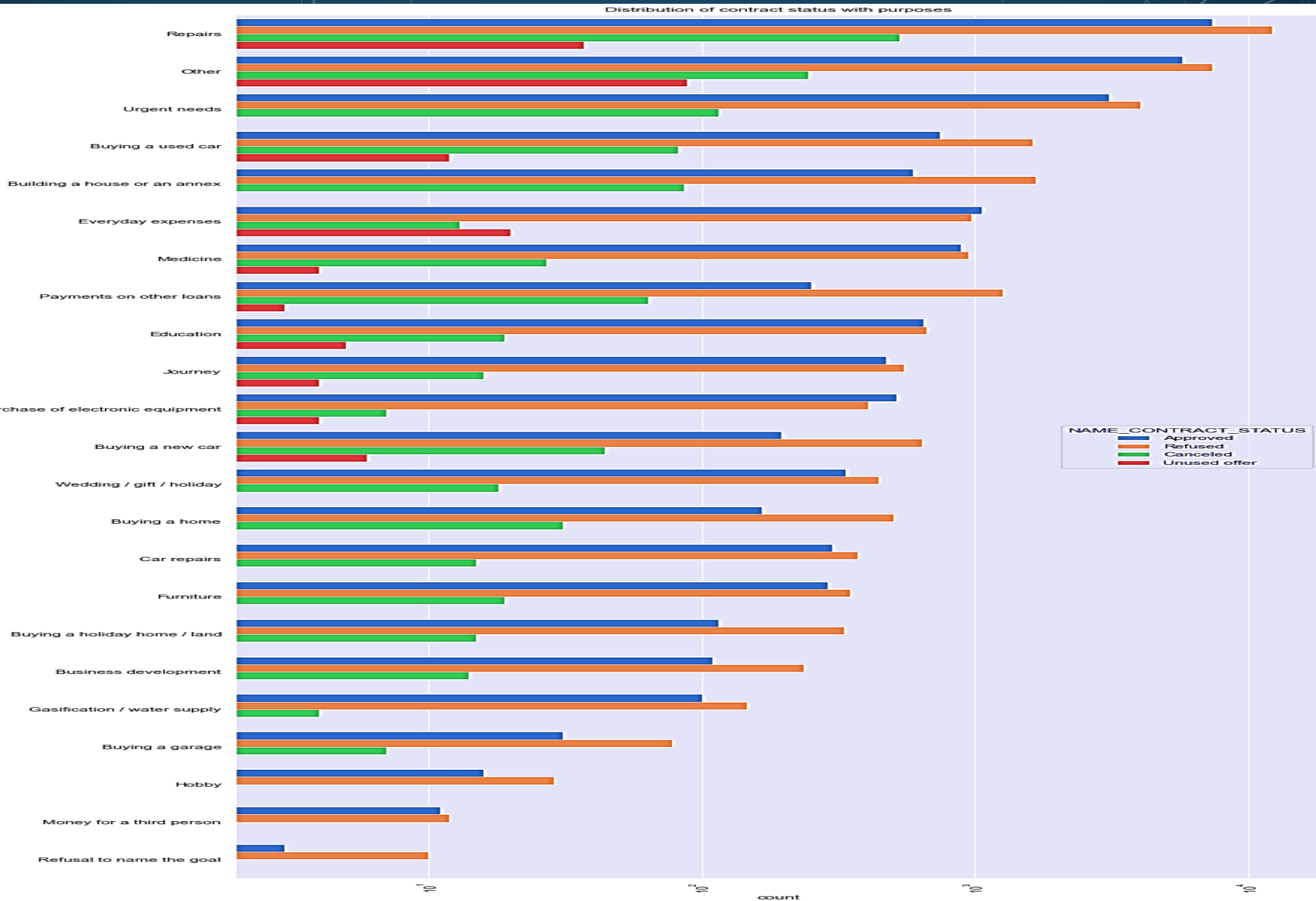
# DISTRIBUTION OF PURPOSES WITH TARGET

Inferences Drawn:

- Loan Purposes like "Repairs" are having the max trouble in timely payments.
- People having no issues in payment is more than the people having issues in all the Loan Purposes.
- Purposes like "Buying a Garage", "Buying a home/Land", "Buying a new Car", "Education", "Journey", "Everyday Expenses" have a better successful payment rate than the delayed payments rate. So we can focus at these segments where there are less payment issues.

Distribution of purposes with target
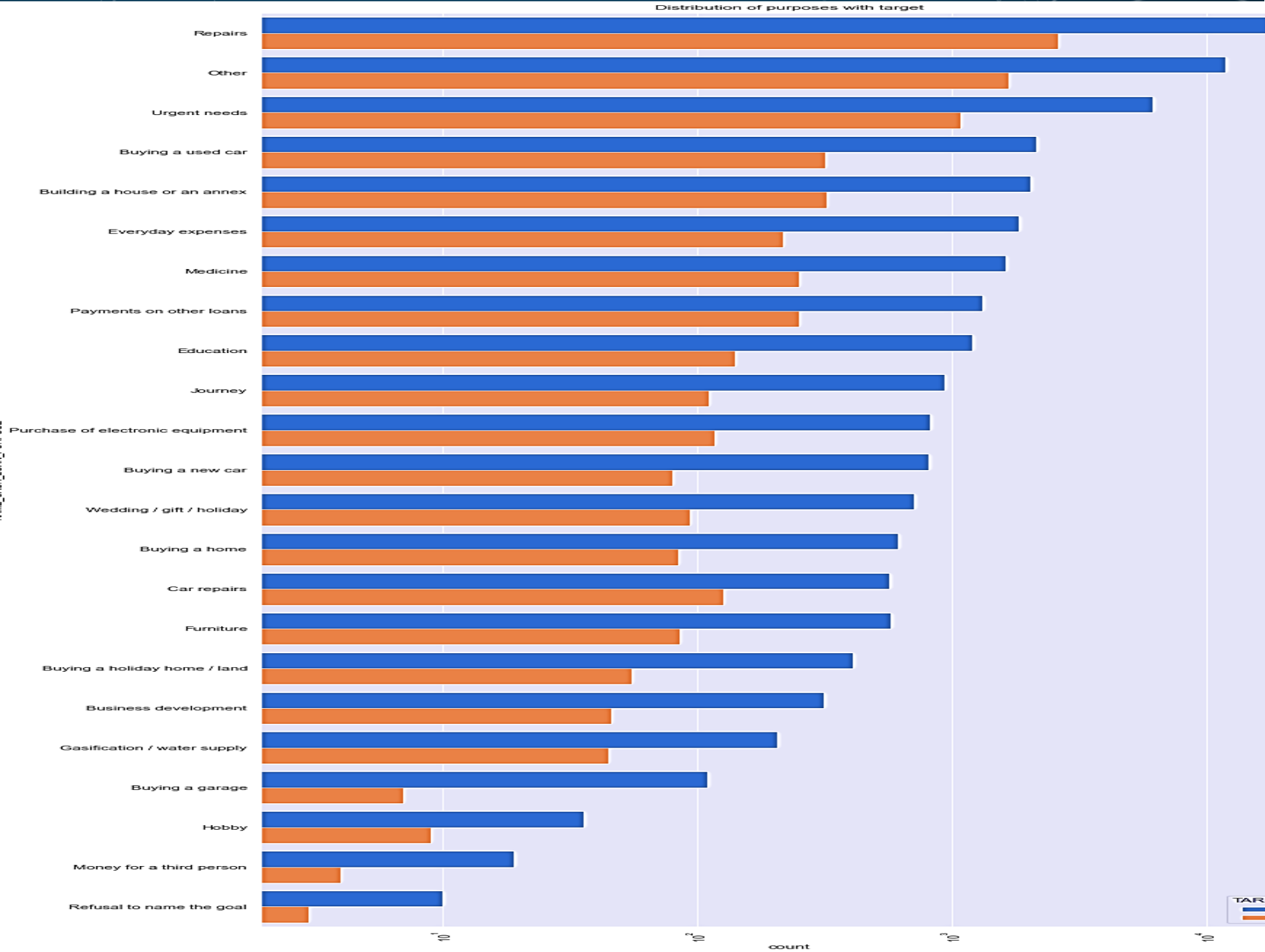
# Performing bivariate analysis



Prev Credit amount v/s Loan Purpose
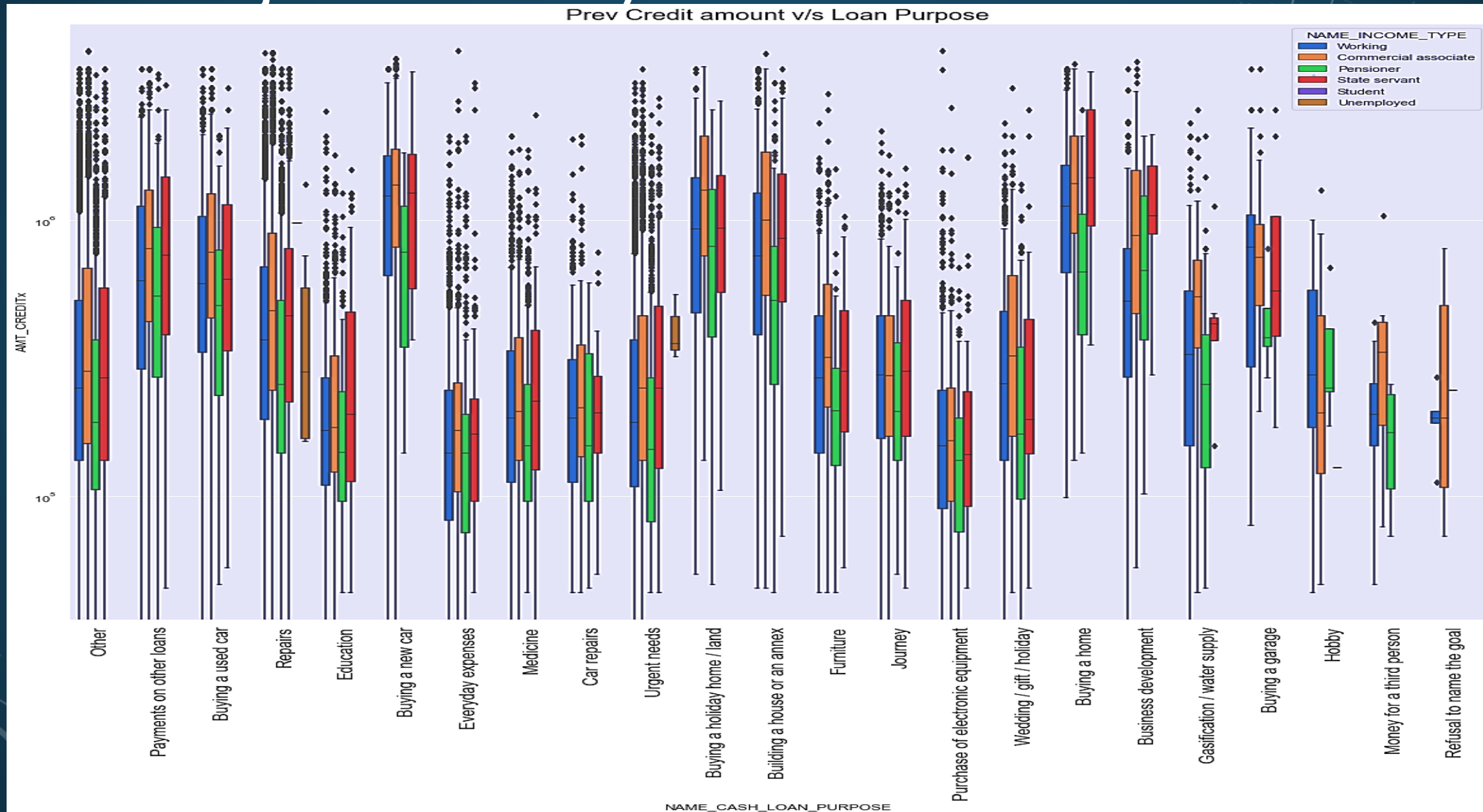
**Inferences Drawn:**

- The credit amount of Loan purposes like 'Buying a home', 'Buying a holiday home/ land', 'Buying a new car' and 'Building a house or annex' is higher.
- Money for third person or money for a Hobby is having less credits.
- Tax for State Servants are higher as they have a higher upper quartile, and outliers in all the segments.

## Prev Credit amount vs Housing type

Inferences Drawn:

- We see that office apartment has higher credit for Target Variable 0 and Co-op Apartment has highest credit for Target Variable 1.
- Hence the banks should not approve loans from housing type of "CO-op Apartment".
- Banks should approve loans from housing type of "With Parents", "House/Apartment,"Municipal Apartment", "Office Apartment", "Rented Apartment" in order to not get into losses.



Previous Application Credit amount v/s Housing type
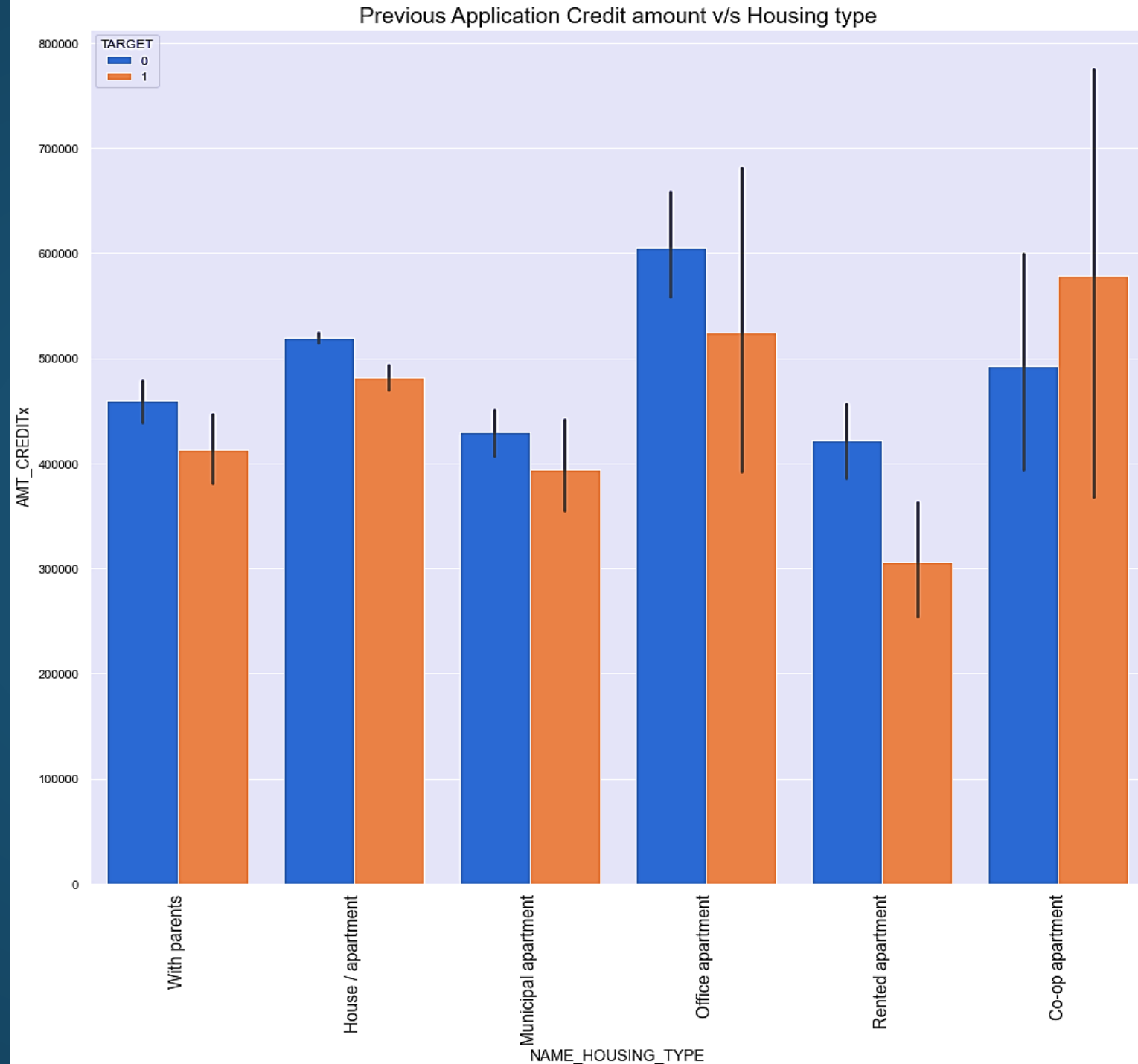
# CONCLUSION

- Banks should focus more on contract type 'Student' and 'Businessman' with housing type other than 'Co-op apartment' for successful payments.

- Banks should focus less on income type Working as they are having most number of unsuccessful payments.

- Loan Purposes like "Repair" is having higher number of unsuccessful payments on time.

- Get as much as clients from housing type "With parents" as they are having least number of unsuccessful payments.

- Bank should not approve loans from females having a salary of 200K-300K as the bank.

# THANK YOU