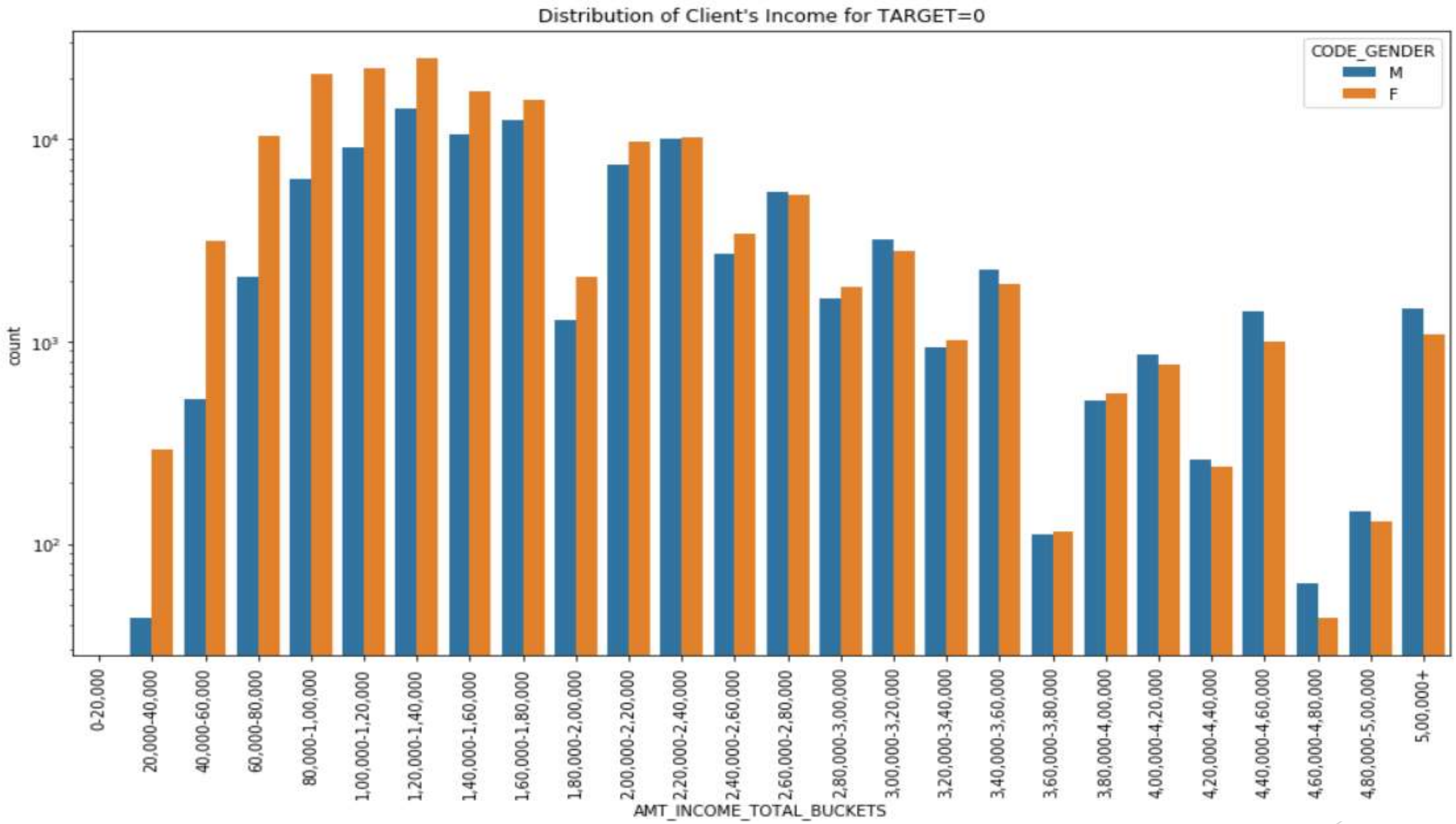
EDA CASE STUDY

BY AVLEEN SINGH & ATHULYA SOBHAN

FOR APPLICATION DATA DATAFRAME

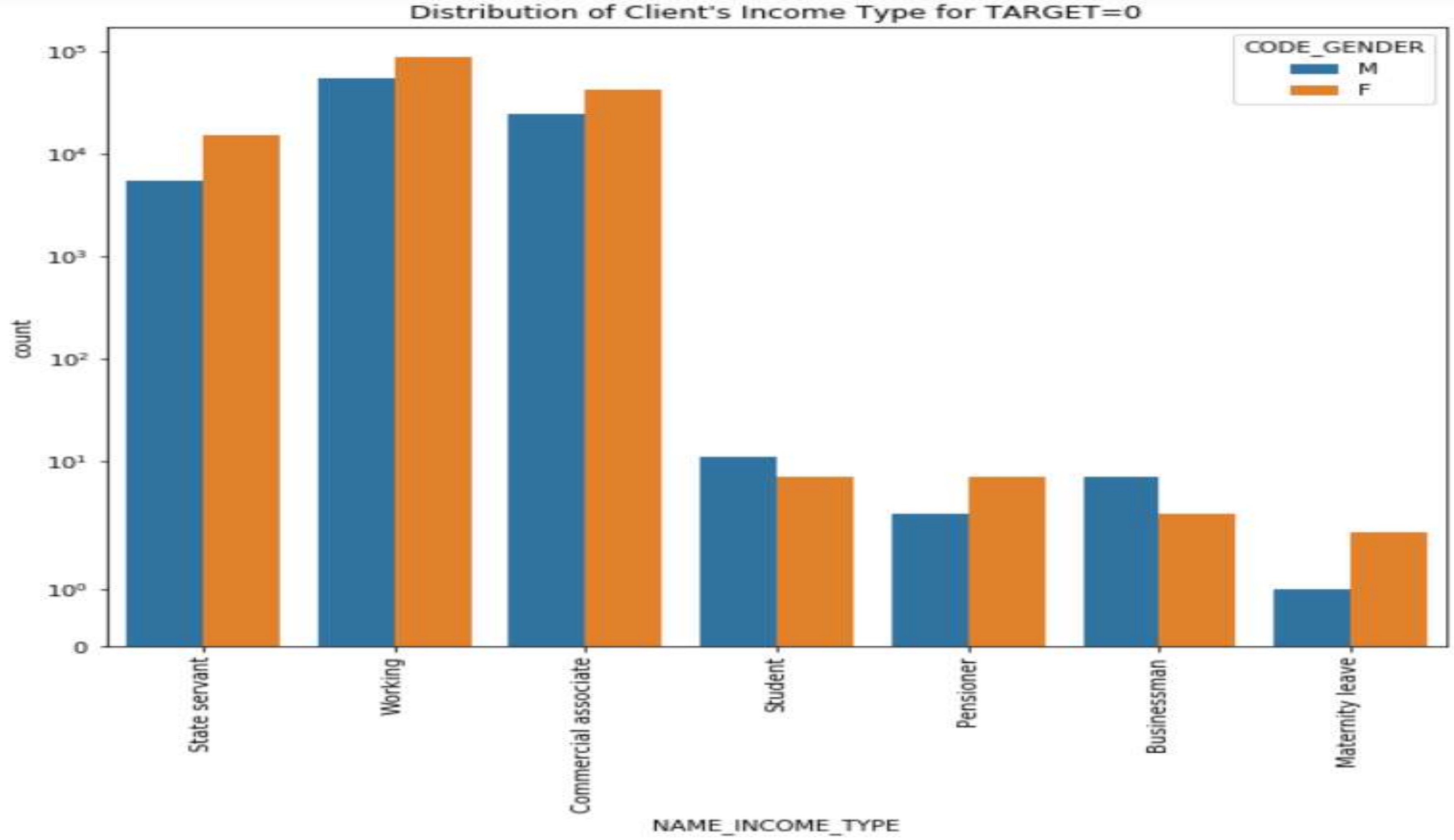
CATEGORIAL UNIVARIATE ANALYSIS FOR TARGET = 0



Countplot for TARGET=0 of Client's Income

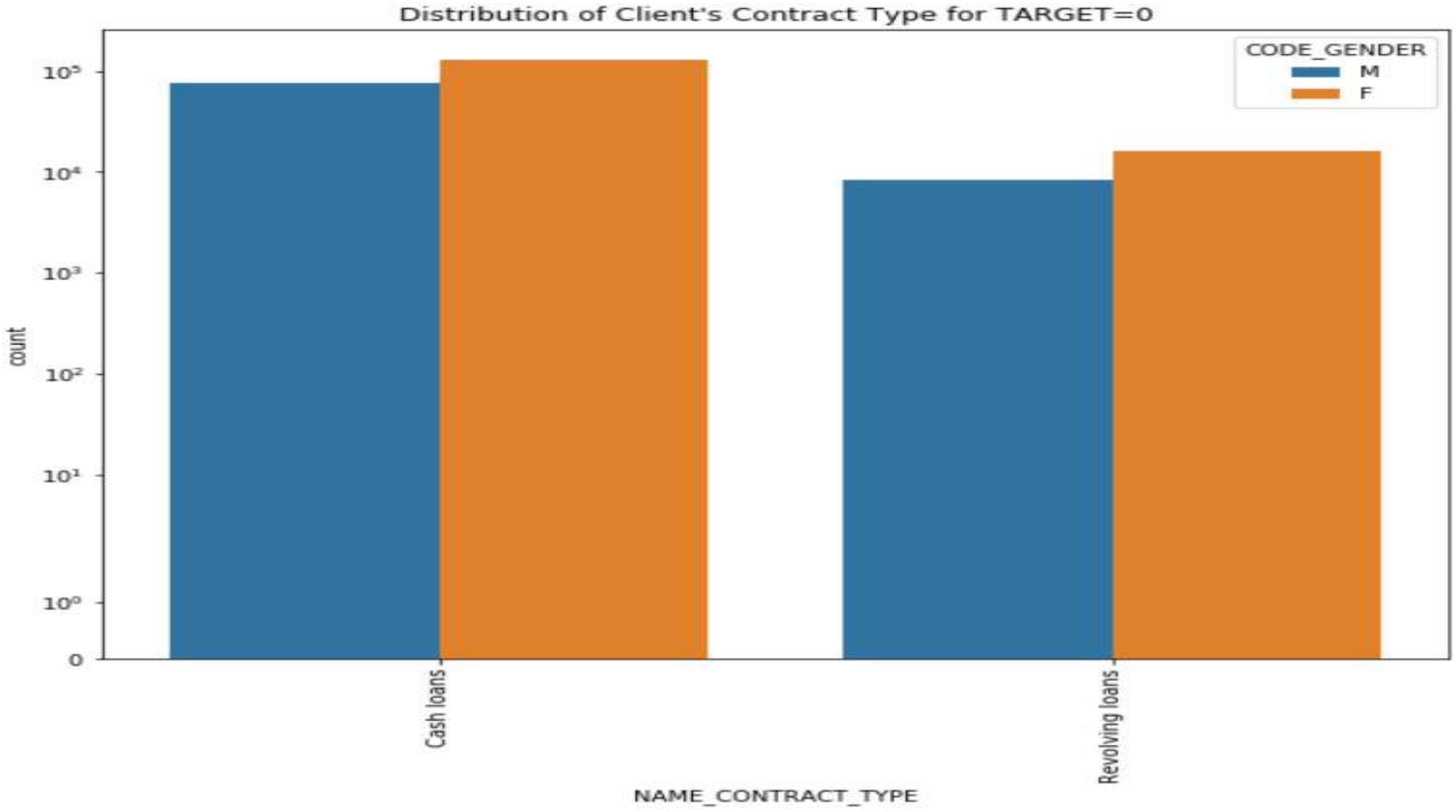
- Female count is more than the Male count in this graph.
  Maximum income of people ranges between 1.2 - 1.4 lakhs and in this, majority of the females are earning between this range as compared to the males.
- Very less people earn between 4.6 - 4.8 lakhs.

Plotting "INCOME_TYPE" with "CODE_GENDER" for TARGET=0



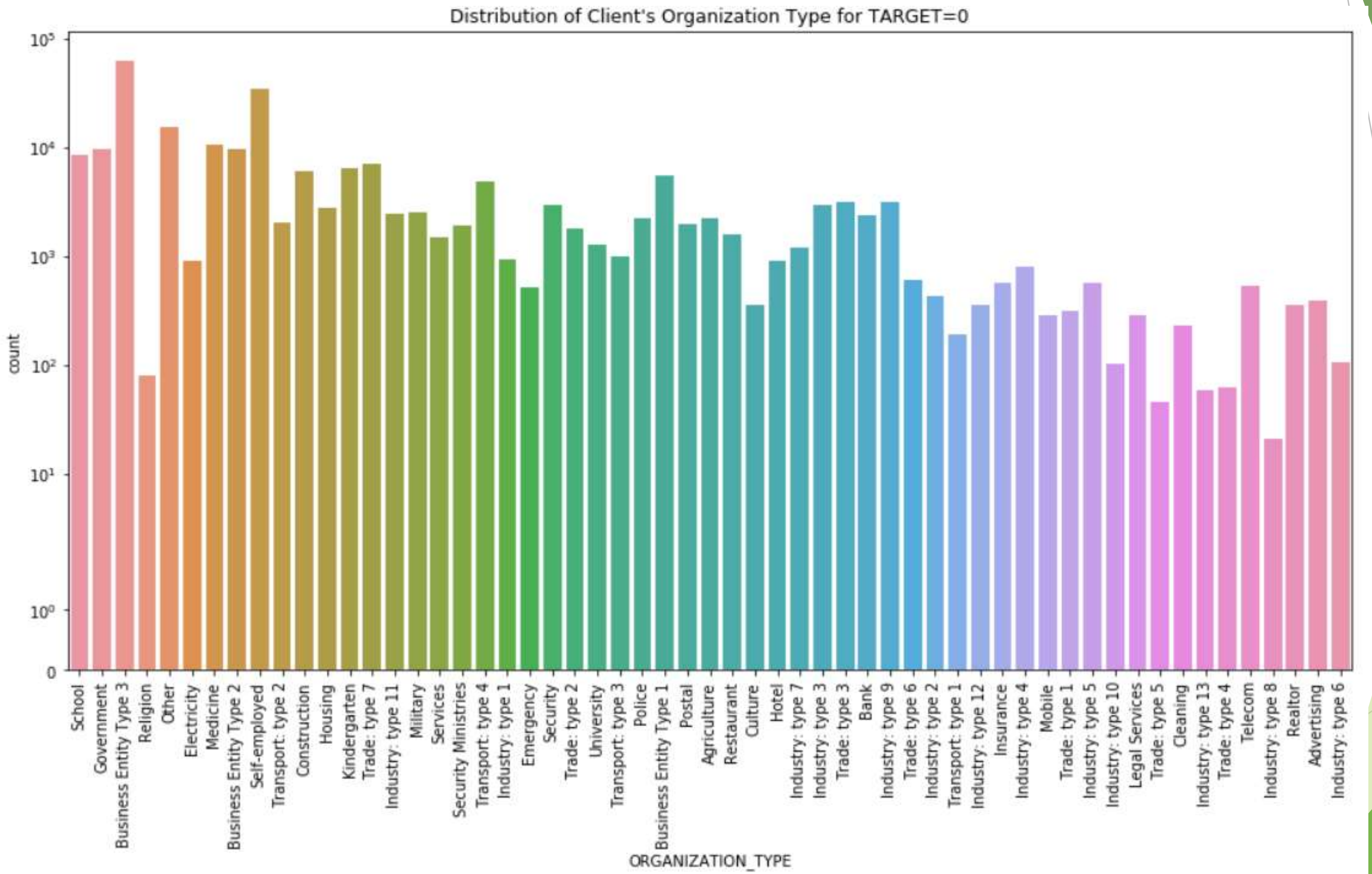
- The Working, State Servant and Commercial associates earn the maximum as compared to the others.
- The Student, Pensioner, Businessman and Maternity leave people earn less.
- In these categories, the females are the ones who are earning higher than the males.

Plotting the Contract_Type with respect to the Gender for TARGET=0



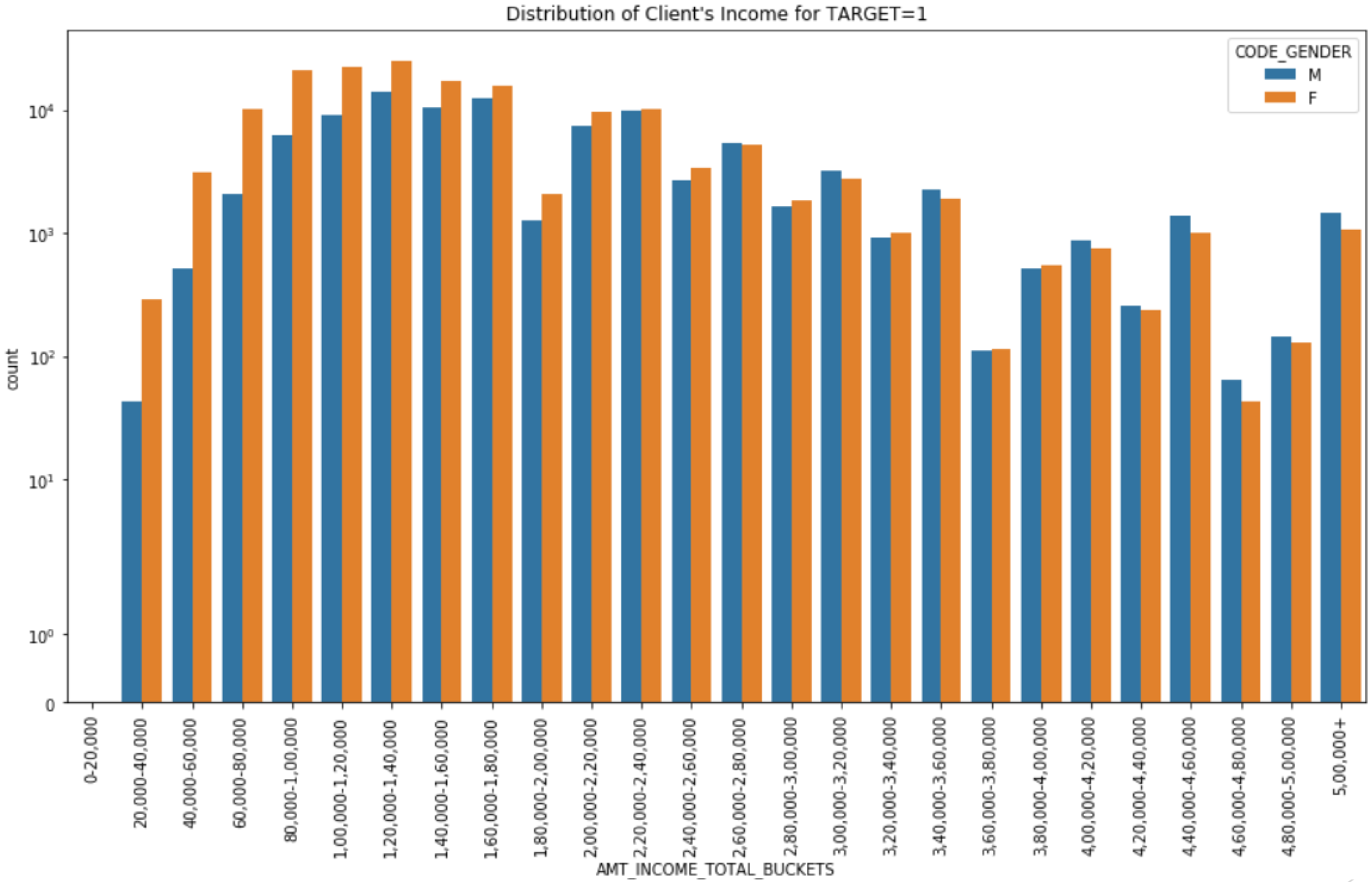
- We can clearly see that Contract Type : Cash Loans are having higher number of people as compared to that of Contract Type : Revolving Loans
- In both these Contract Types, Females are the majority than the males.

Plotting the Organization_type for TARGET=0



- Most number of people are working in Business Entity-Type3, Self employed, Other and Medicine categories.
- Least number of people are working in Industry: Type 8, Trade: Type 5, Trade: Type 4 and Industry: Type 13

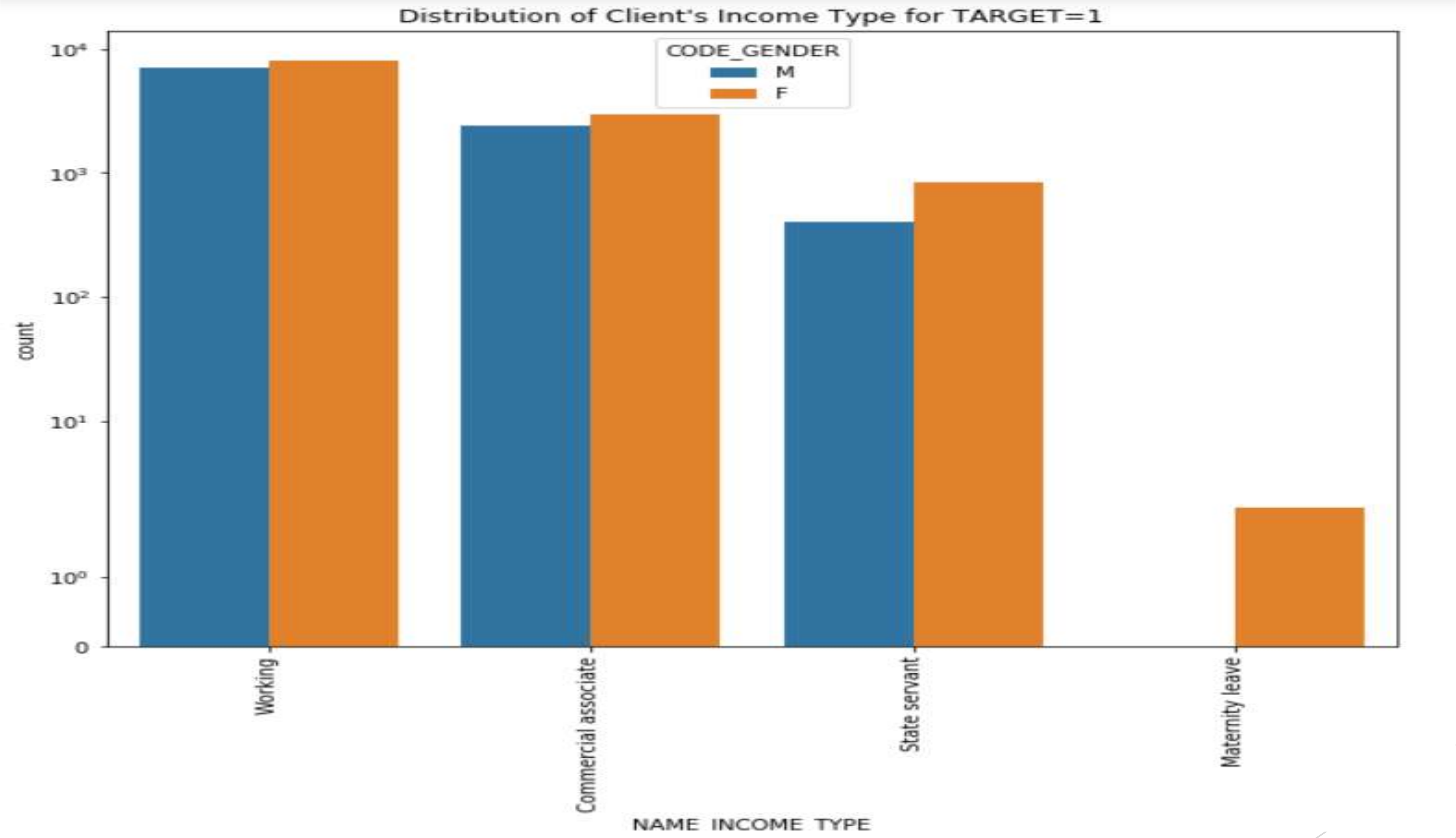
CATEGORIAL UNIVARIATE ANALYSIS FOR TARGET = 1



Countplot of Client's Income for TARGET=1

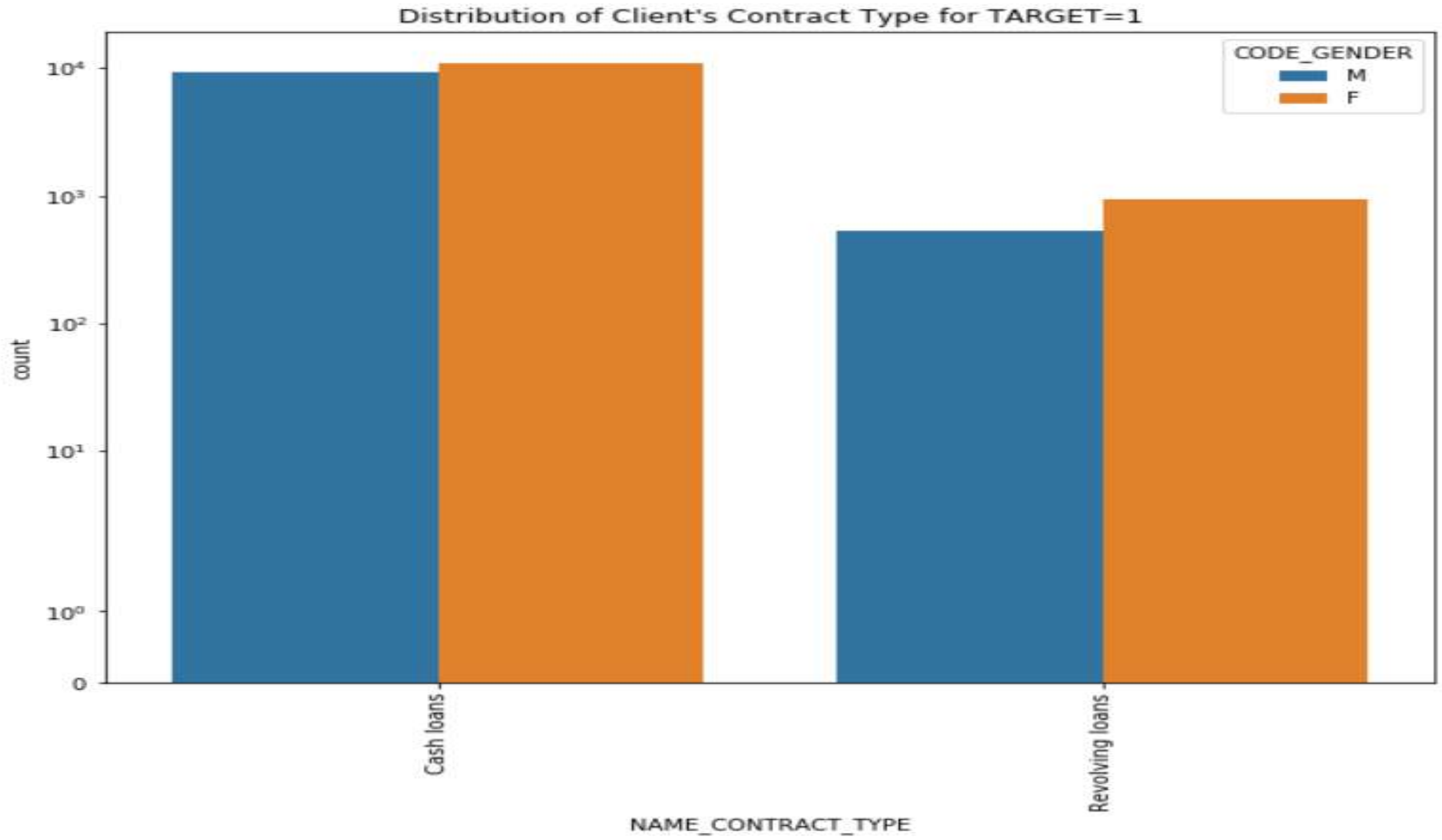
- In this, the females are a slightly higher than the males and not as a huge difference that we had seen in TARGET=0 dataframe.
- Maximum people's income is between 1.2 - 1.4 lakhs and the least income is between 4.6 - 4.8 lakhs.

Plotting "INCOME_TYPE" with "CODE_GENDER" for TARGET=1



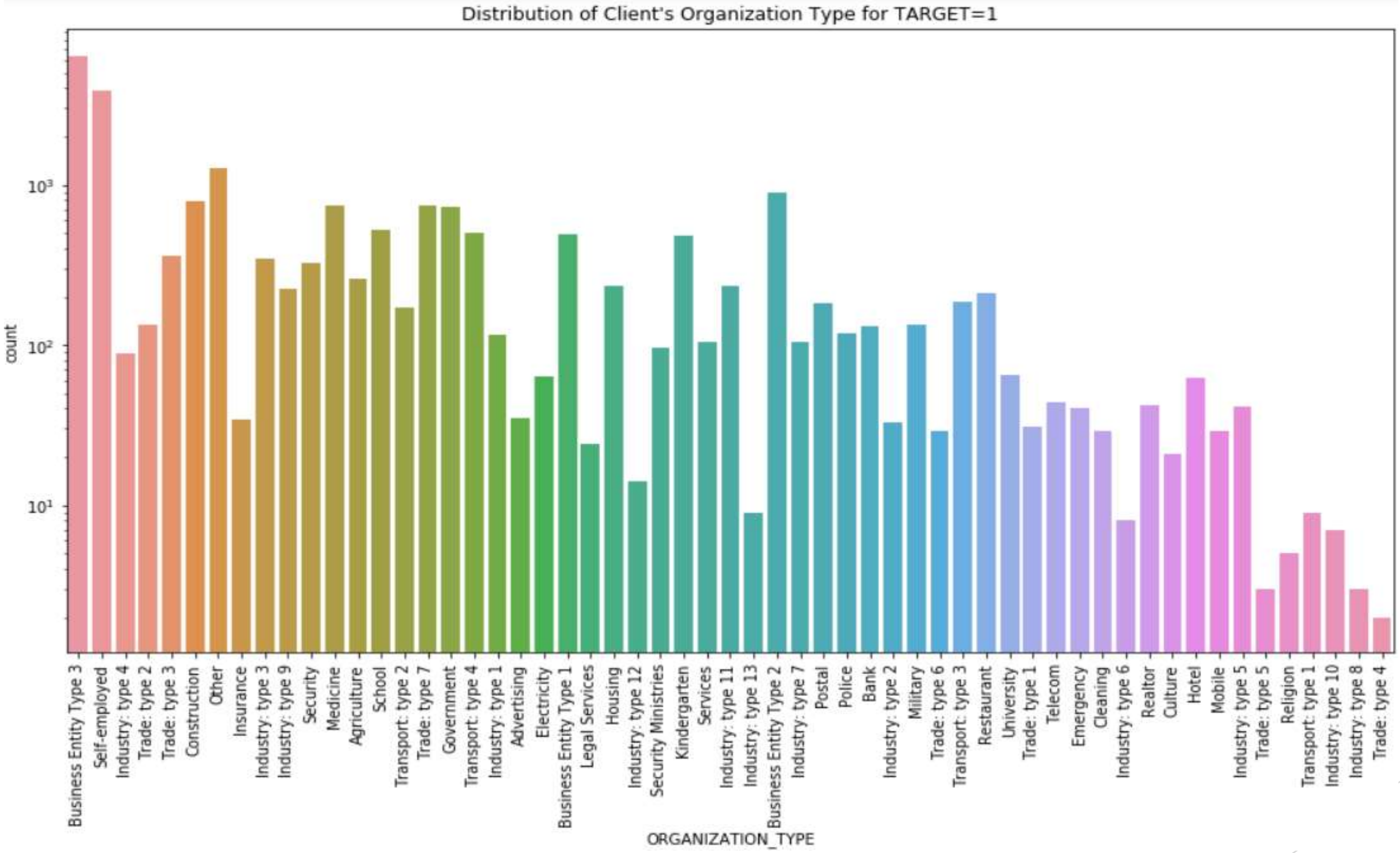
- *W*e can say that Working, Commercial associate, State servant and Maternity leave people are the ones who are earning.
- In these columns, female counts is higher than the male.
- In this, there is no income type student , pensioner and Business man, which means they don’t do any late payments.

Plotting the Contract_Type with respect to the Gender for TARGET=1



- In this, the count of Cash loans and Revolving in target=1 is reduced as compared to that of target=0, who are difficulty in paying loans on time.
- In this, the females are the majority as compared to that of males.

Plotting the Organization_type for TARGET=1



- Most number of people are working in Business Entity-Type3, Self employed, Other and Medicine categories.
- Least number of people are working in Industry: Type 8, Trade: Type 5, Trade: Type 4 and Industry: Type 13.
- It is almost same as that of target=0, except the count of people is reduced.

Correlation for TARGET = 0

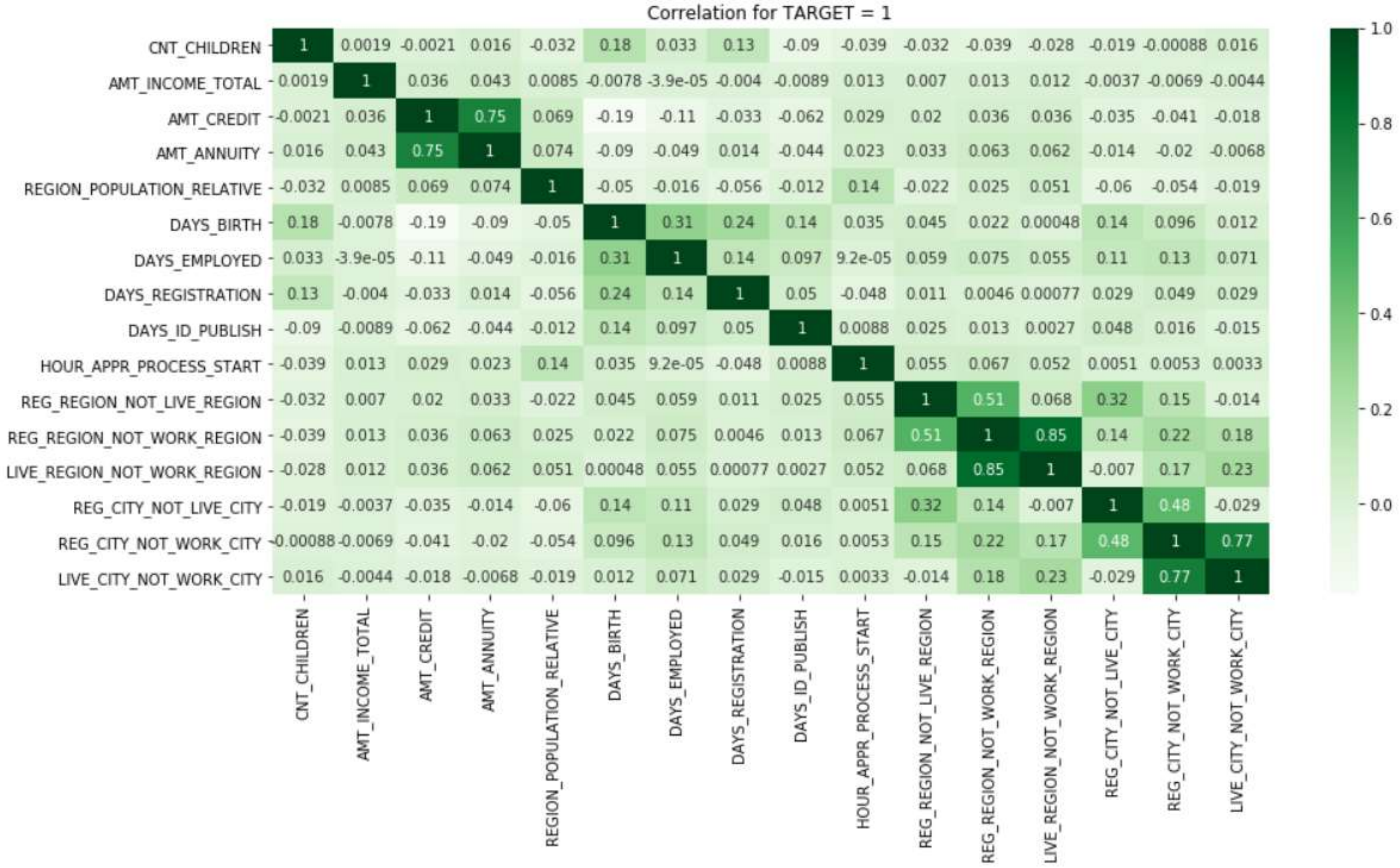
Correlation for TARGET = 0

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	HOUR_APPR_PROCESS_START	REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	REG_CITY_NOT_LIVE_CITY	REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY
CNT_CHILDREN	1	-0.0098	-0.019	-0.0076	-0.03	0.24	0.063	0.16	-0.12	-0.033	-0.023	-0.017	-0.0069	-0.0016	0.01	0.018
AMT_INCOME_TOTAL	-0.0098	1	0.33	0.4	0.17	-0.046	-0.03	0.035	-0.026	0.056	0.065	0.13	0.12	-0.0043	-0.02	-0.011
AMT_CREDIT	-0.019	0.33	1	0.76	0.1	-0.15	-0.088	-0.015	-0.035	0.04	0.021	0.051	0.052	-0.038	-0.039	-0.015
AMT_ANNUITY	-0.0076	0.4	0.76	1	0.12	-0.092	-0.052	0.0098	-0.026	0.034	0.04	0.076	0.072	-0.019	-0.029	-0.014
REGION_POPULATION_RELATIVE	-0.03	0.17	0.1	0.12	1	-0.041	0.0087	-0.06	-0.004	0.17	-0.0033	0.066	0.092	-0.054	-0.049	-0.017
DAYS_BIRTH	0.24	-0.046	-0.15	-0.092	-0.041	1	0.35	0.3	0.11	0.049	0.058	0.039	0.014	0.17	0.11	0.031
DAYS_EMPLOYED	0.063	-0.03	-0.088	-0.052	0.0087	0.35	1	0.17	0.087	0.022	0.057	0.082	0.063	0.11	0.13	0.075
DAYS_REGISTRATION	0.16	0.035	-0.015	0.0098	-0.06	0.3	0.17	1	0.03	-0.032	0.021	0.018	0.0093	0.047	0.055	0.031
DAYS_ID_PUBLISH	-0.12	-0.026	-0.035	-0.026	-0.004	0.11	0.087	0.03	1	0.0054	0.027	0.021	0.0089	0.053	0.033	0.0018
HOUR_APPR_PROCESS_START	-0.033	0.056	0.04	0.034	0.17	0.049	0.022	-0.032	0.0054	1	0.055	0.072	0.058	0.011	-0.0015	-0.0051
REG_REGION_NOT_LIVE_REGION	-0.023	0.065	0.021	0.04	-0.0033	0.058	0.057	0.021	0.027	0.055	1	0.46	0.09	0.34	0.14	0.0035
REG_REGION_NOT_WORK_REGION	-0.017	0.13	0.051	0.076	0.066	0.039	0.082	0.018	0.021	0.072	0.46	1	0.86	0.15	0.22	0.18
LIVE_REGION_NOT_WORK_REGION	-0.0069	0.12	0.052	0.072	0.092	0.014	0.063	0.0093	0.0089	0.058	0.09	0.86	1	0.015	0.17	0.22
REG_CITY_NOT_LIVE_CITY	-0.0016	-0.0043	-0.038	-0.019	-0.054	0.17	0.11	0.047	0.053	0.011	0.34	0.15	0.015	1	0.44	0.012
REG_CITY_NOT_WORK_CITY	0.01	-0.02	-0.039	-0.029	-0.049	0.11	0.13	0.055	0.033	-0.0015	0.14	0.22	0.17	0.44	1	0.82
LIVE_CITY_NOT_WORK_CITY	0.018	-0.011	-0.015	-0.014	-0.017	0.031	0.075	0.031	0.0018	-0.0051	0.0035	0.18	0.22	0.012	0.82	1

1.0
0.8
0.6
0.4
0.2
0.0

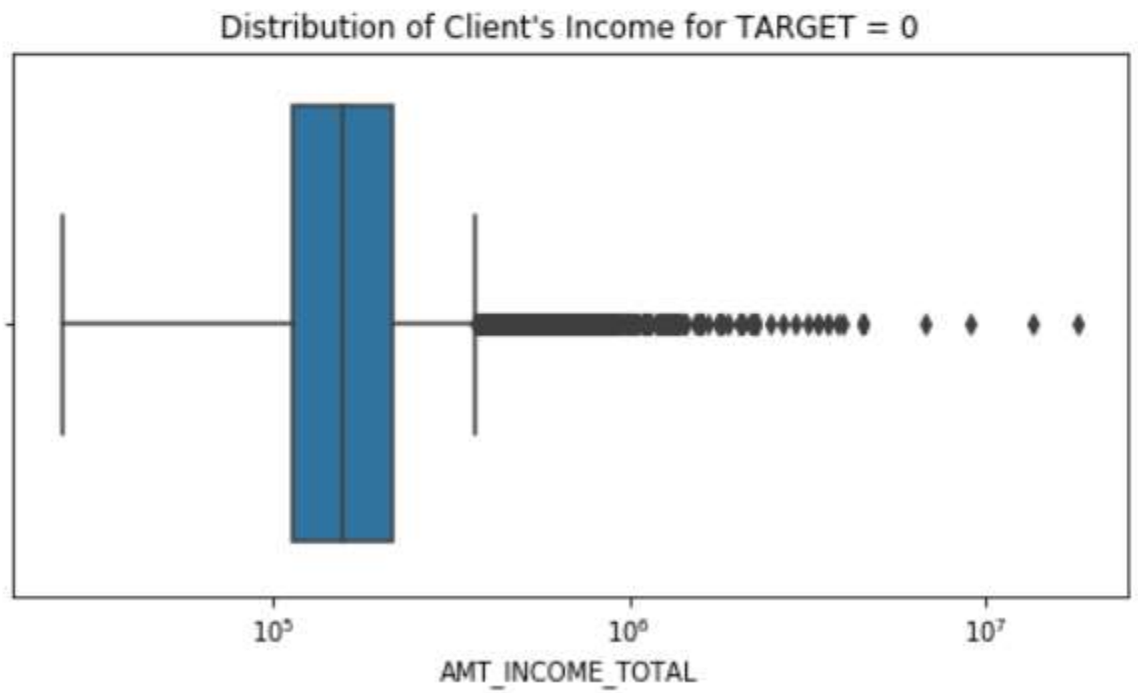
- AMT_CREDIT is inversely proportional to the DAYS_BIRTH.
- AMT_CREDIT is inversely proportional to the number of children the client has, i.e. CNT_CHIDREN.
- AMT_INCOME_TOTAL is inversely proportional to the number of children the client has, i.e. CNT_CHILDREN.

Correlation for TARGET = 1



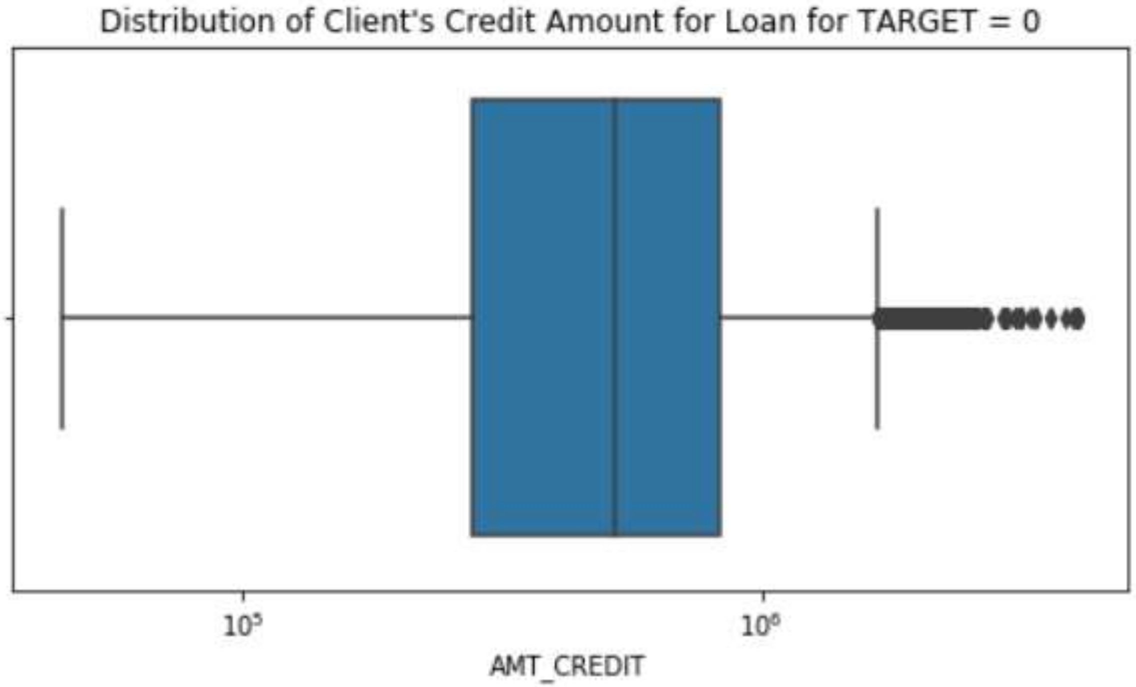
- The client's permanent address (REG_REGION_NOT_LIVE_REGION) does not match contact address are having less children (CNT_CHILDREN) and vice - versa.
- The client's permanent address (REG_REGION_NOT_LIVE_REGION) does not match work address are having less children (CNT_CHILDREN) and vice-versa.

Finding Outliers for "AMT_INCOME_TOTAL" for TARGET=0



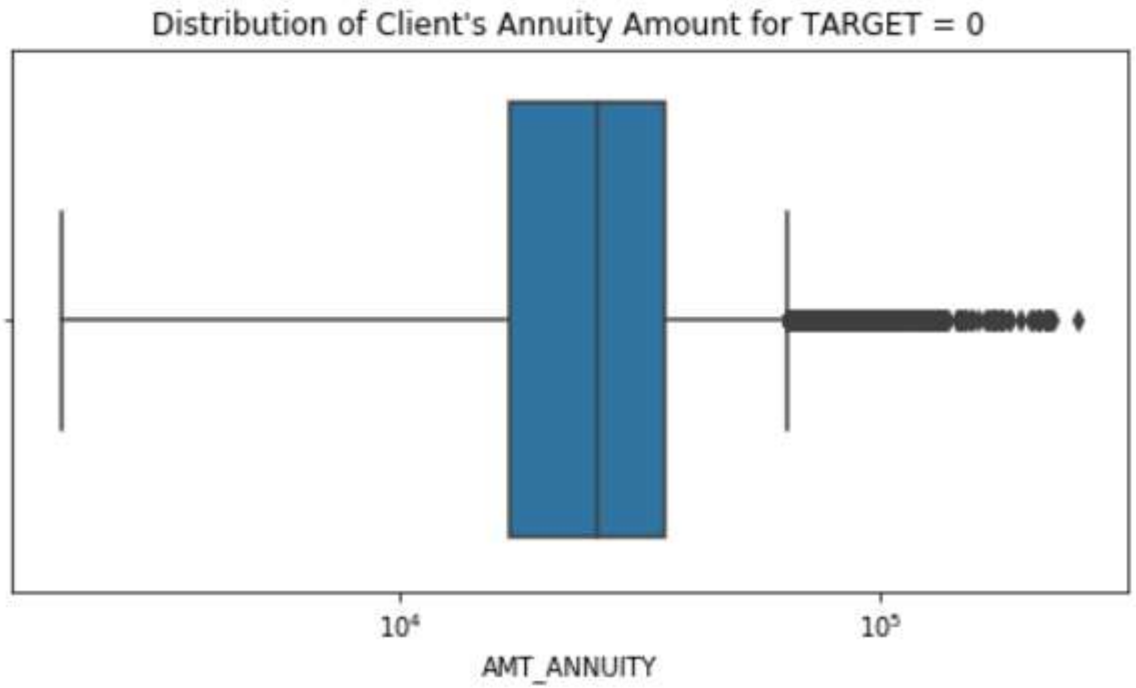
- There are some outliers that are spotted from the above graph.
- The 75th Quartile is smaller as compared to that of 25th Quartile range.

Plotting a boxplot to find outliers for "AMT_CREDIT" for TARGET = 0



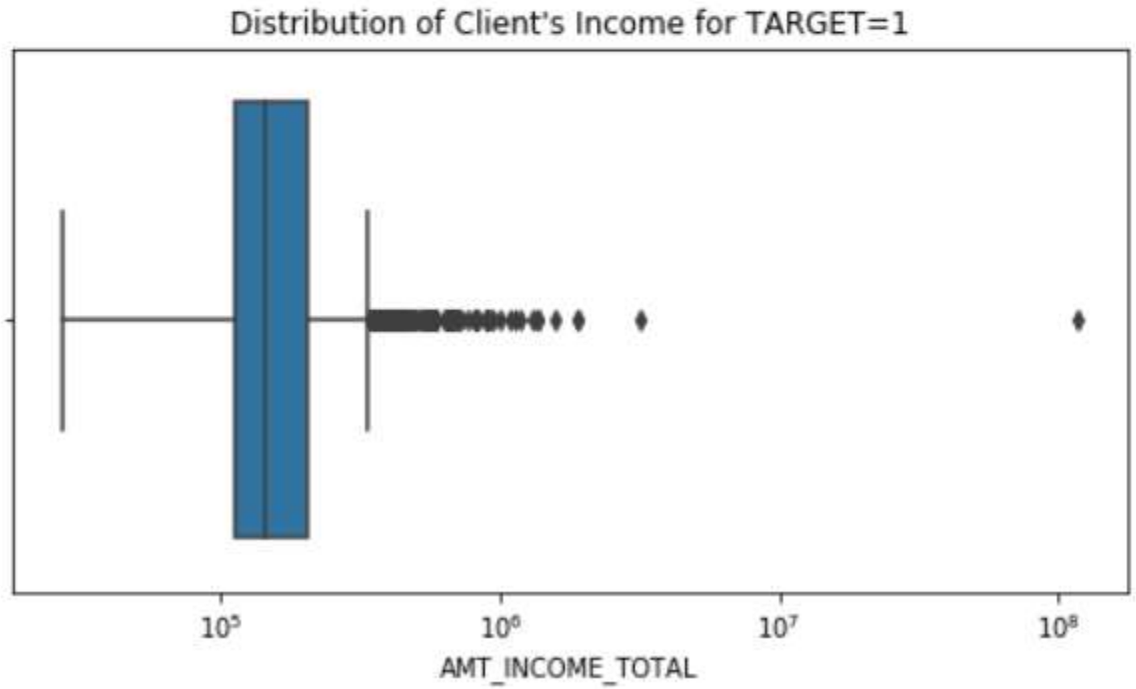
- There are some outliers that are spotted from the above graph.
- The 75th Quartile is smaller as compared to that of 25th Quartile range.

Plotting a boxplot to find outliers for "AMT_ANNUITY" for TARGET = 0



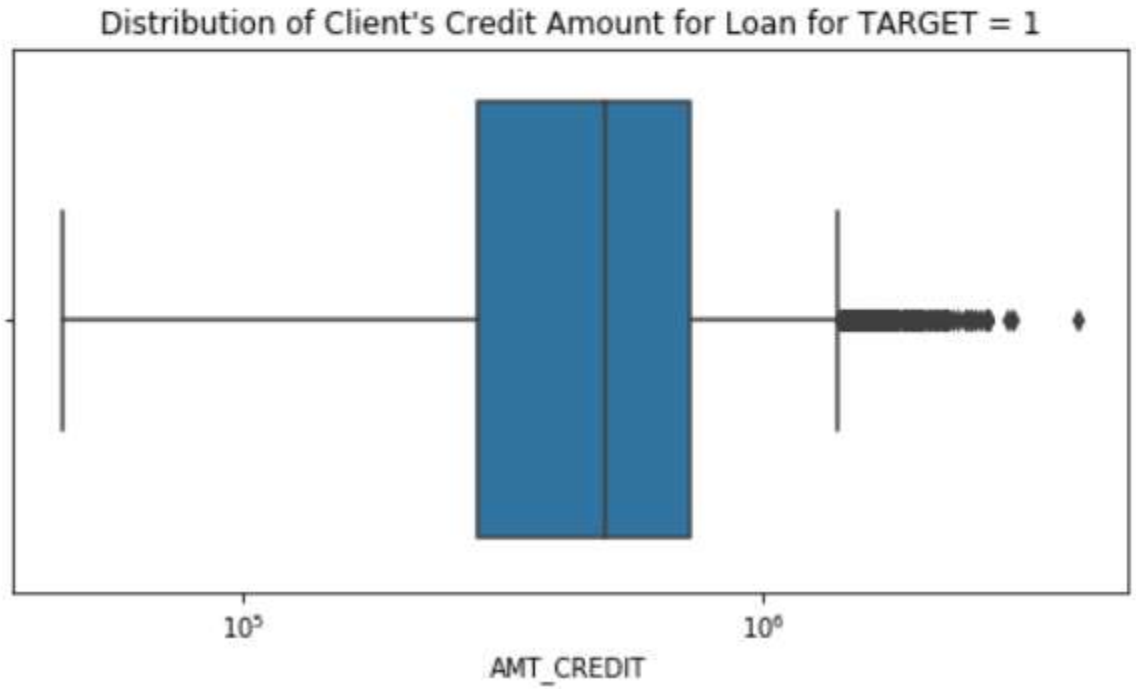
- There are some outliers that are seen from the above graph.
- In this, the 25th Quartile is bigger than the 75th Quartile range.

Plotting a boxplot to find outliers for "AMT_INCOME_TOTAL" for TARGET = 1



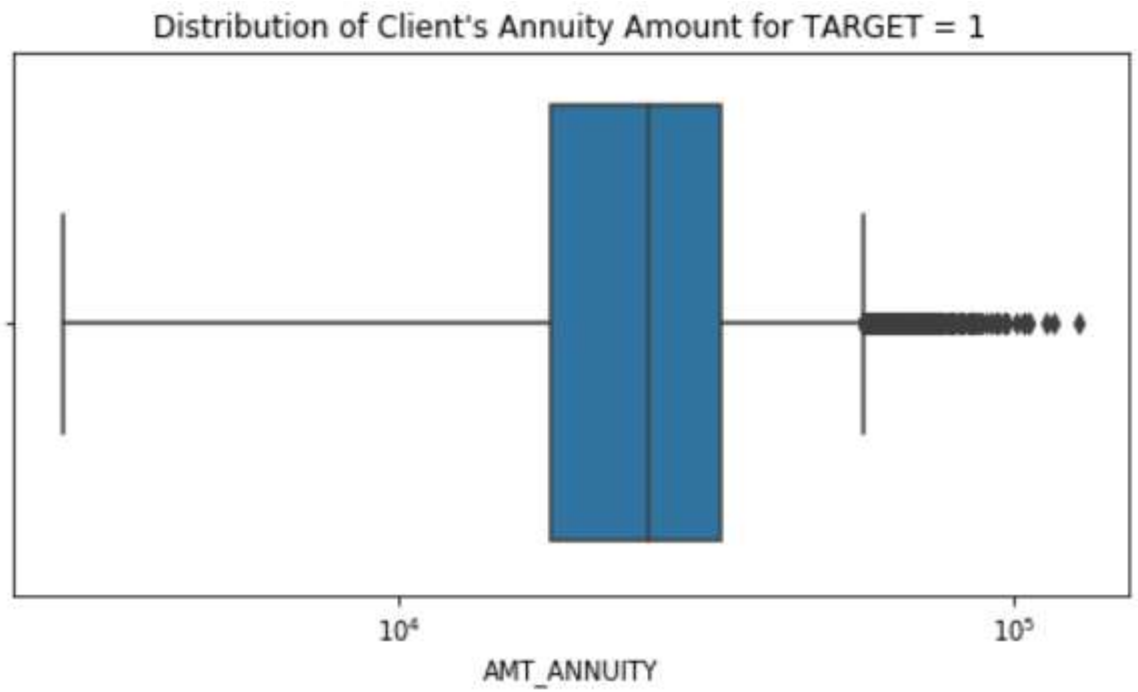
- There are some outliers that are found from the above graph.
- In this, the 25th Quartile is bigger than the 75th Quartile range.

Plotting a boxplot to find outliers for "AMT_CREDIT" for TARGET = 1



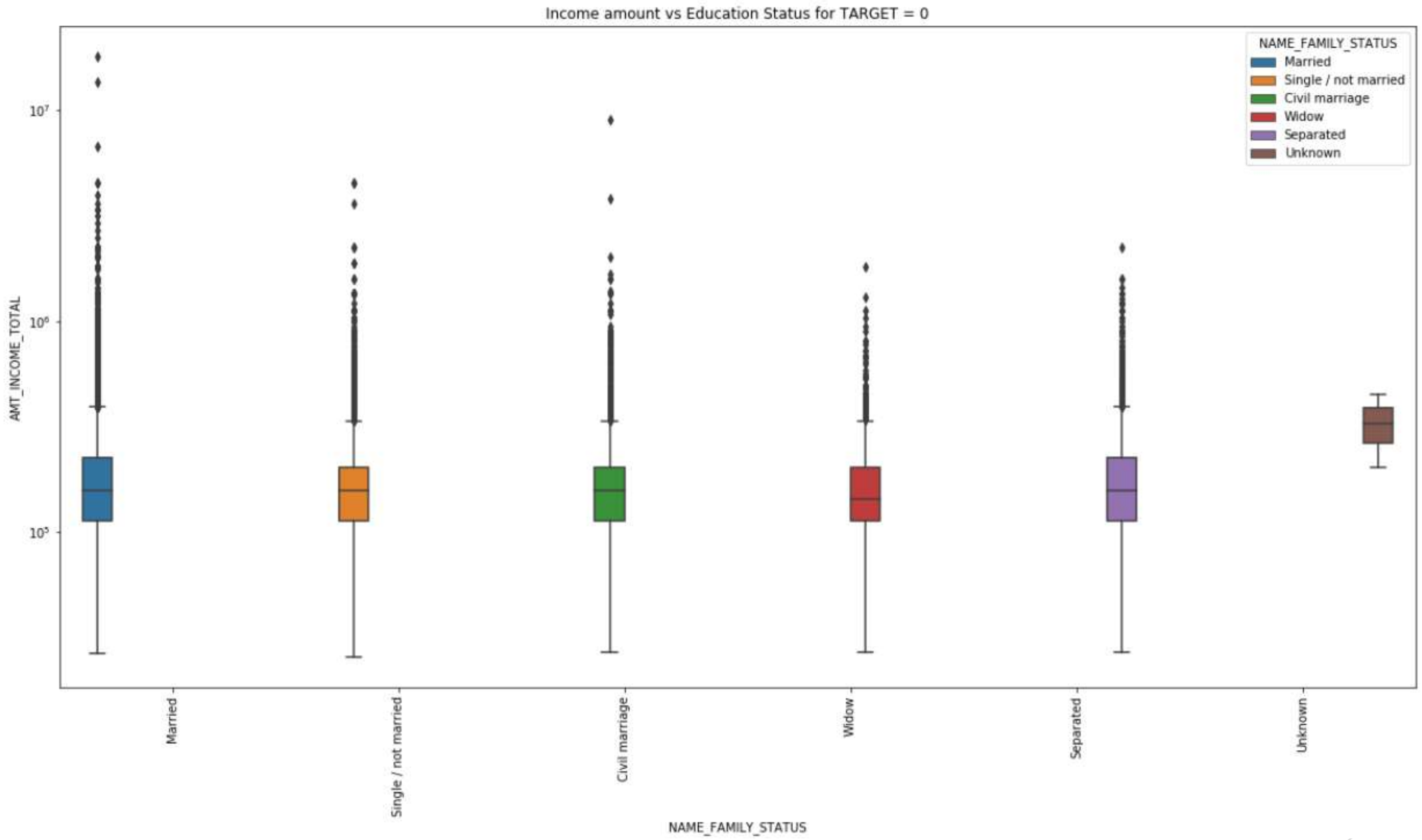
- There are some outliers that are spotted from the above graph.
- The 75th Quartile is smaller as compared to that of 25th Quartile range.

Plotting a boxplot to find outliers for "AMT_ANNUITY" for TARGET = 1



- There are some outliers that are found from the above graph.
- In this, the 25th Quartile is bigger than the 75th Quartile range.

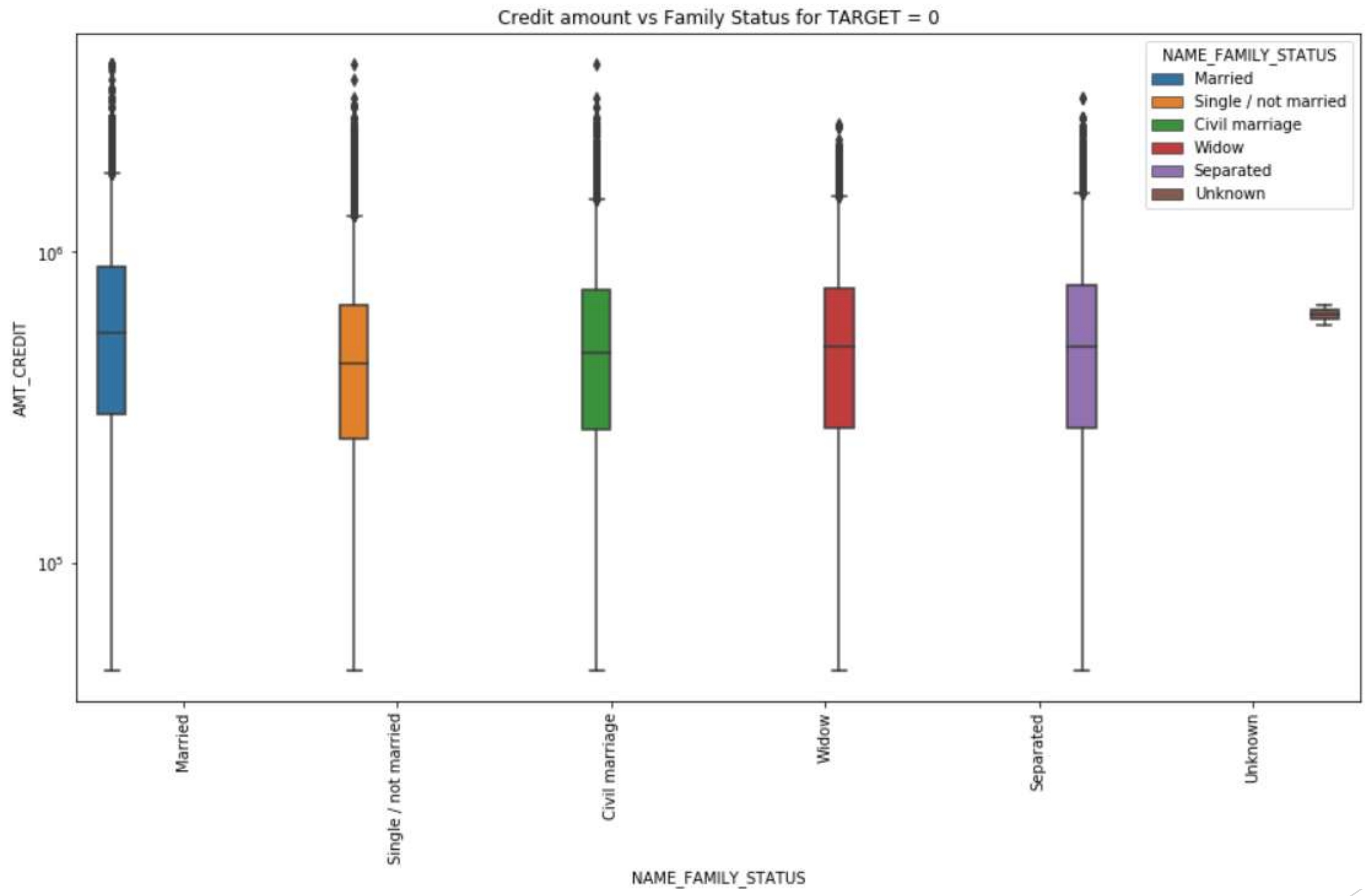
Bivariate Analysis for TARGET = 0



Box plotting for Income amount vs Education Status for TARGET = 0

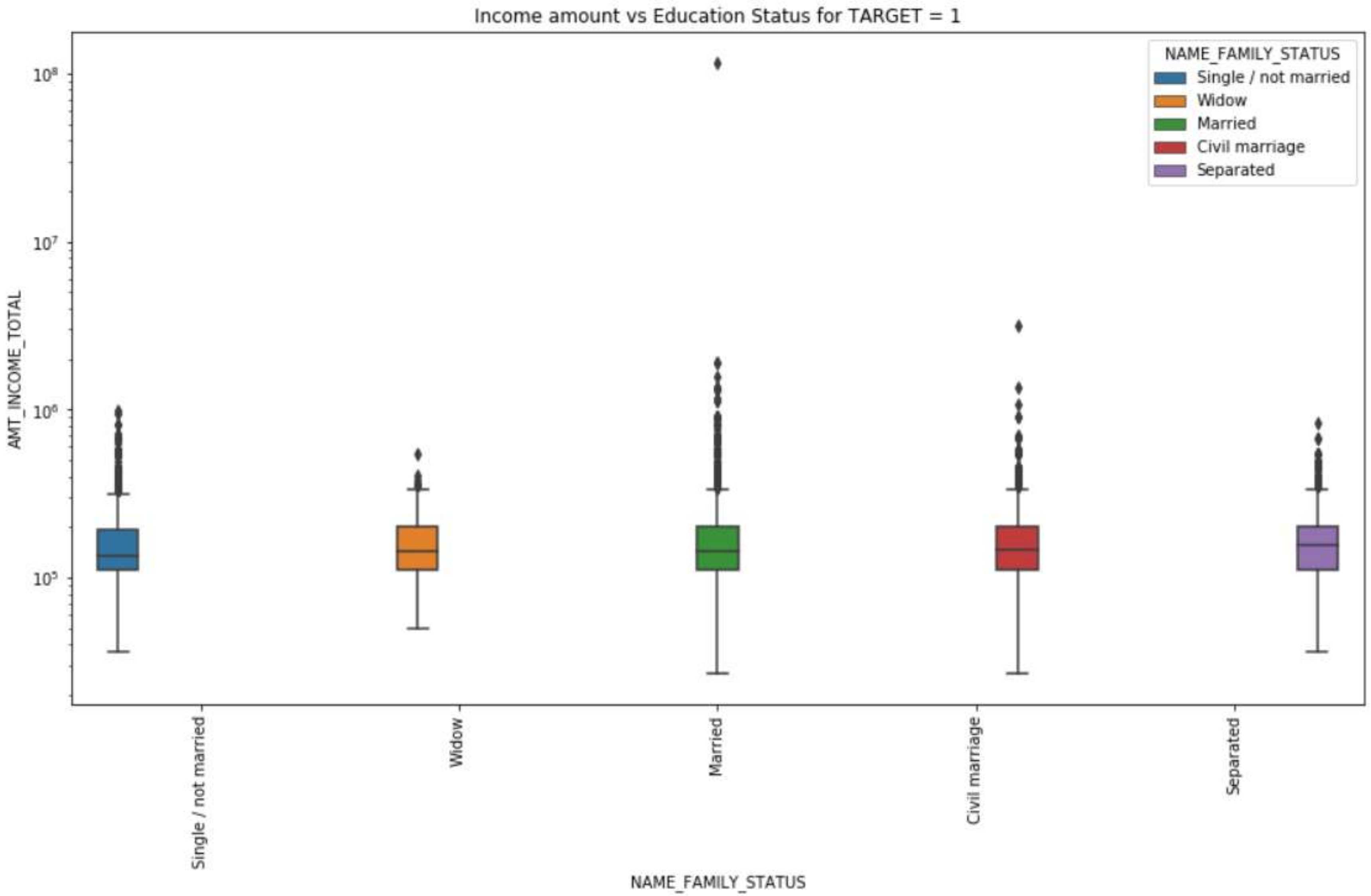
- Married status are having most outliers as compared to the other family status.
- Unknown status are very less in number and do not have any outliers.

Box plotting for Credit amount vs Family Status for TARGET = 0



- The Married status count lies in the particular range and does not have much outliers too.
- Single / not married have less count and have more outliers in them.

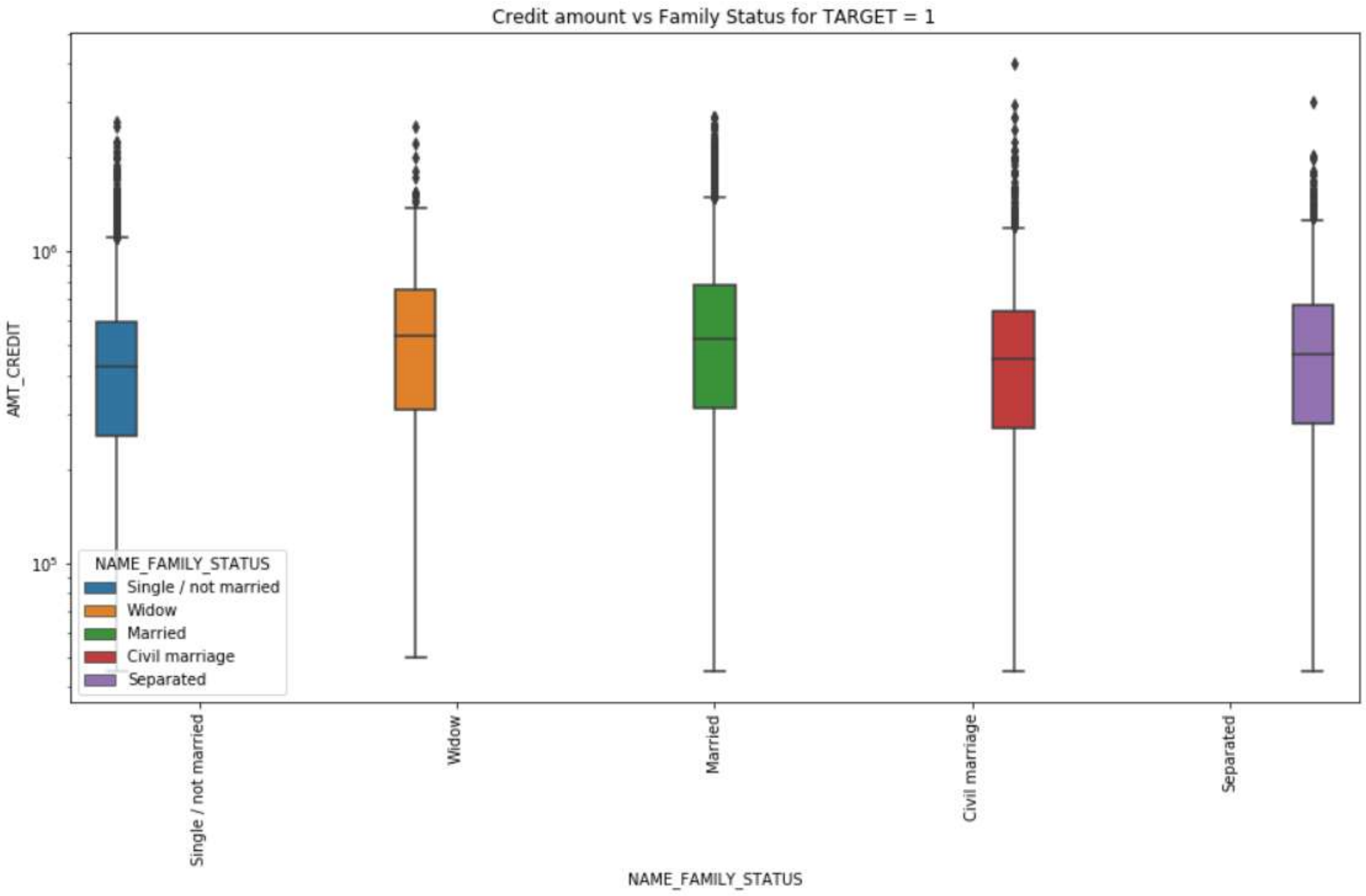
Bivariate Analysis for TARGET = 1



Box plotting for Income amount vs Education Status for TARGET=1

- Married status have the maximum number of outliers as compared to that of other family status.
- There is only change from the TARGET = 1 graph, i.e., Unknown status is not there in the above graph.

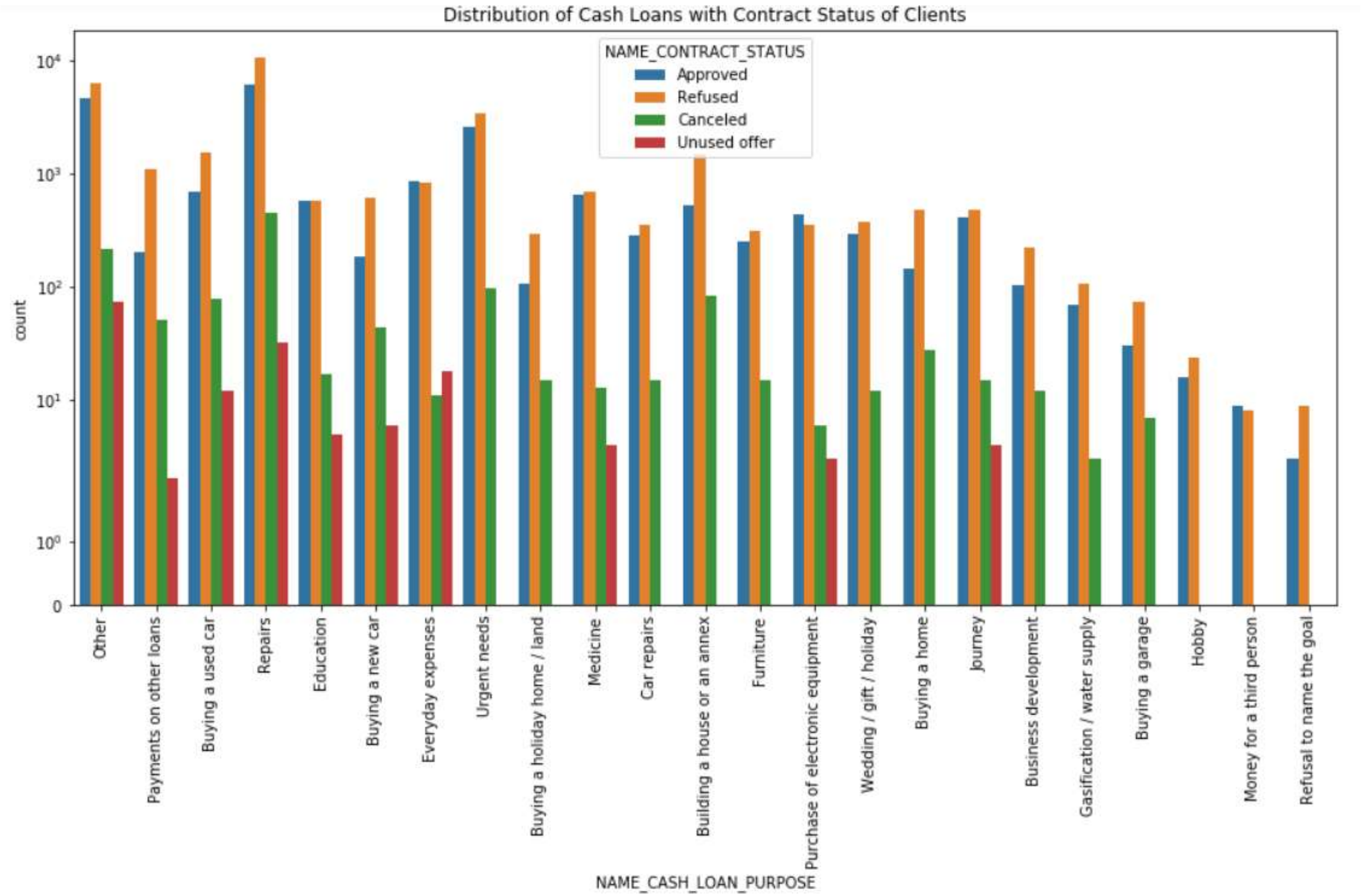
Box plotting for Credit amount vs Family Status for TARGET = 1



- Civil marriage status have the maximum number of outliers as compared to that of other family status.
- There is only change from the TARGET = 1 graph, i.e., Unknown status is not there in the above graph.

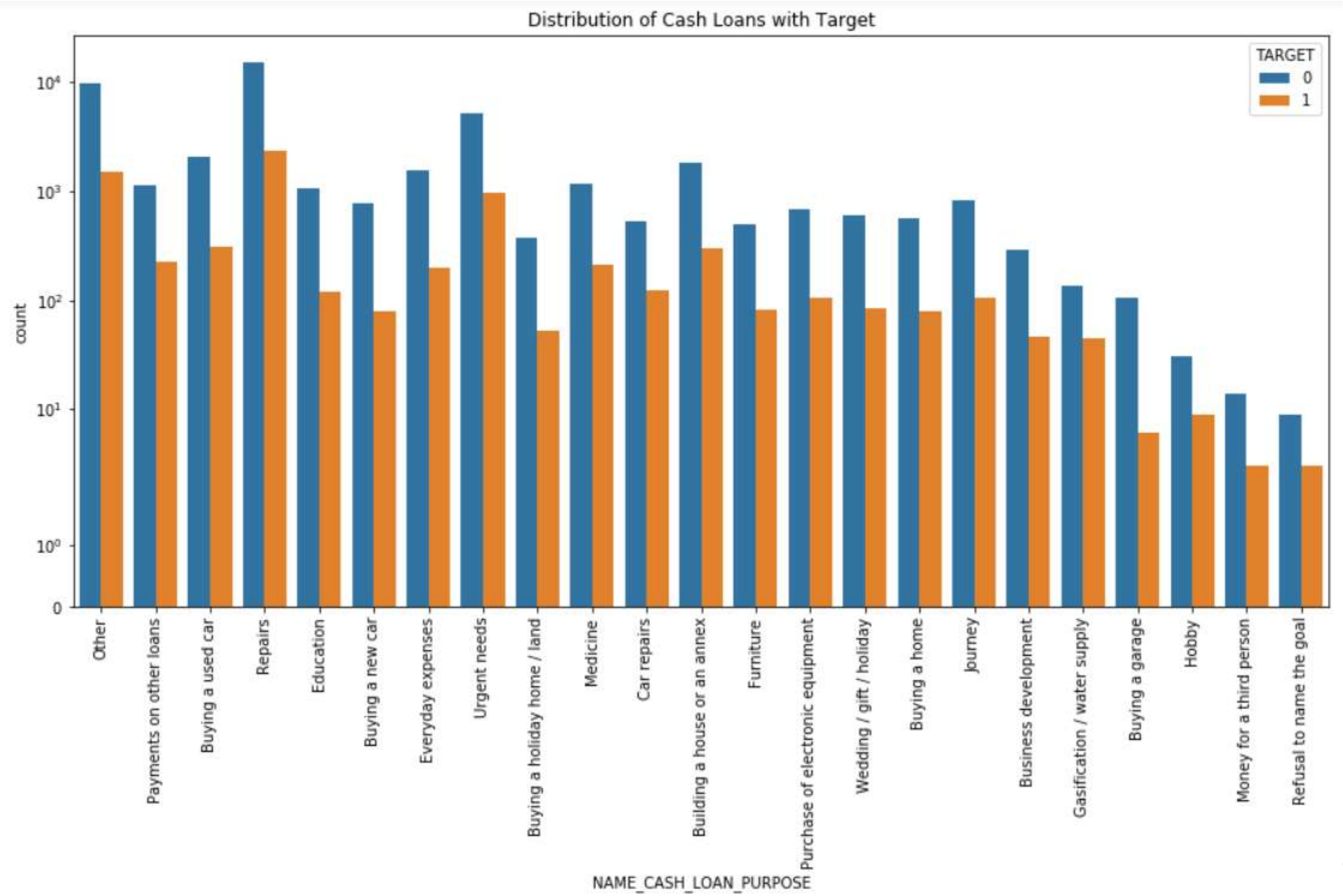
FOR PREVIOUS APPLICATION DATAFRAME

Countplot of Cash Loans with Contract Status of Clients



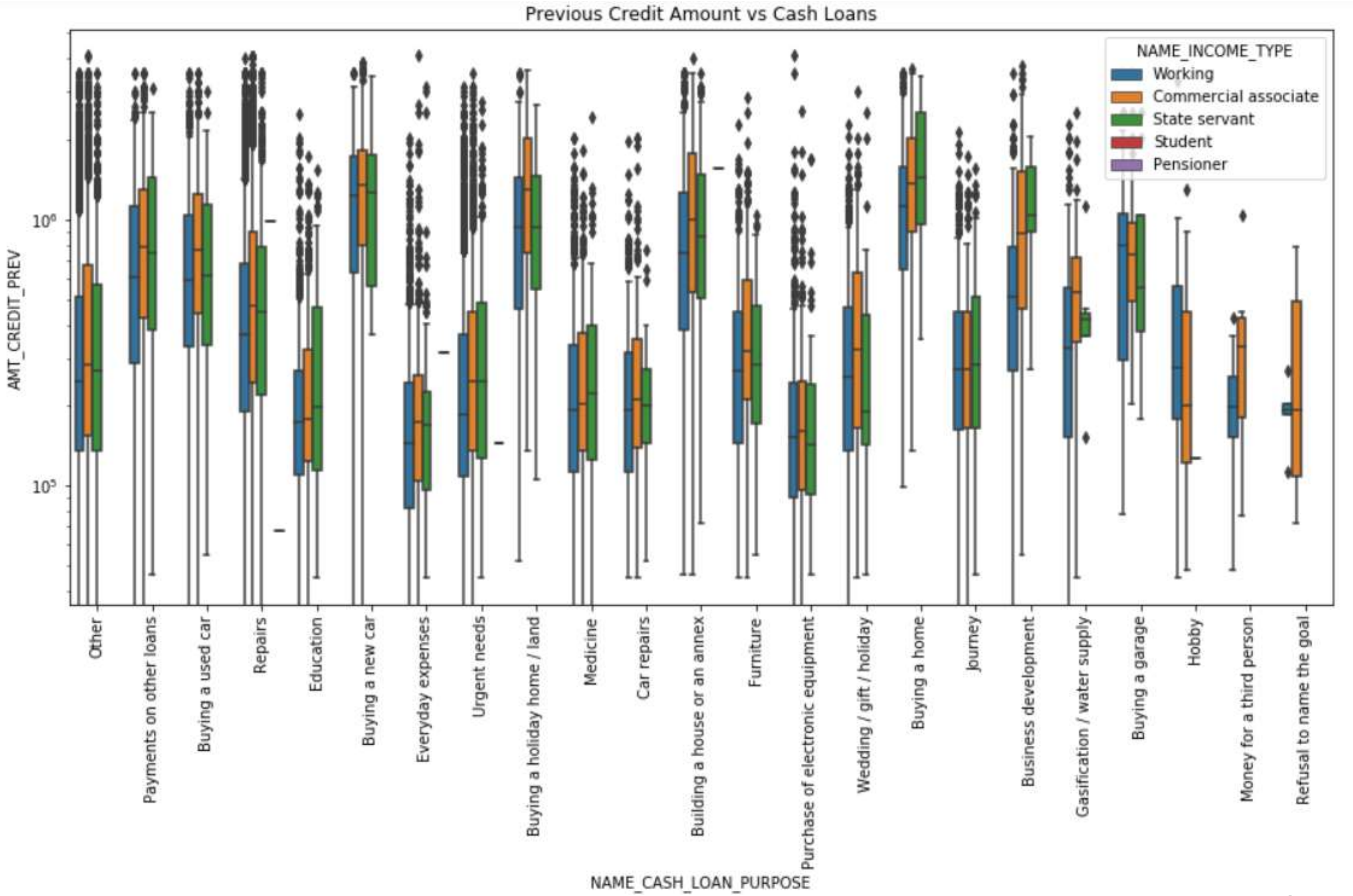
- We can say that the loans are majorly refused by the bank to give it to the clients.
- The loans that are most refused are from the Repairs section.
- In Education section, we can see almost equal number of loans are being approved and refused.

Countplot of Cash Loans with TARGET



- We can say that the Repairs section are having the least difficulty in paying the loans.
- In majority of all the sections, they are not having any difficulties in paying the loans.

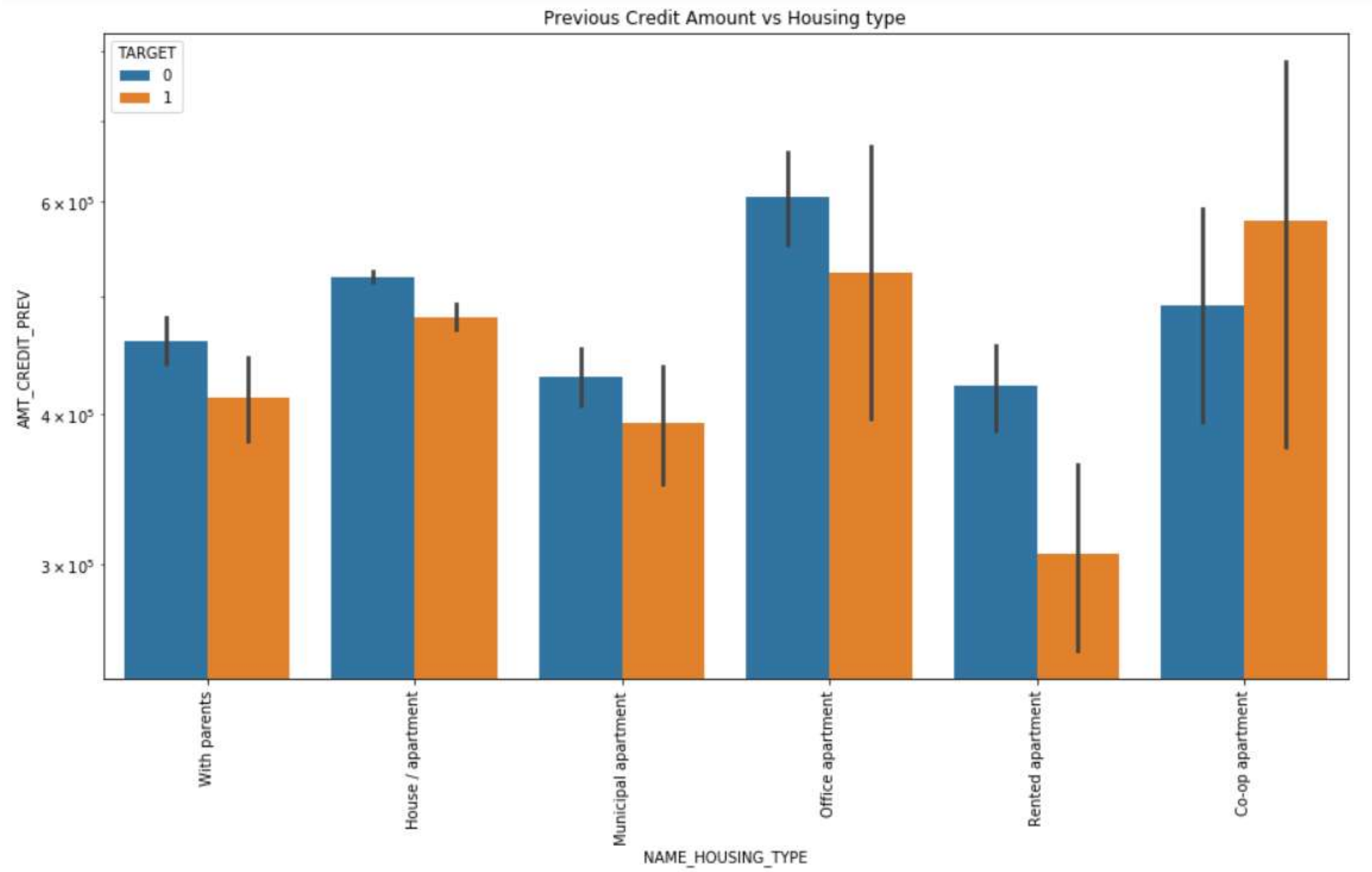
Bivariate Analysis for Inp2 Dataframe



Plotting for Previous Credit Amount vs Cash Loans

- We can say that Buying a new car, Buying a holiday home / Land, Buying a home are the categories that are having maximum previous Credit Amount of the loan.
- Everyday expenses and Purchase of electronic equipment are the categories that are having least previous Credit Amount of the loan.

Plotting for Previous Credit Amount vs Housing type



- We can say that Office apartment is having highest previous Credit Amount and are also able to pay the loans without any difficulty.
- Municipal apartment is having the least previous Credit Amount.
- Co-op apartment is having maximum difficulty as compared to the other Housing Types in paying the loans.

THANK YOU