BANK CREDIT ANALYSIS

SATVIK YADAV
ANTARA CHATTERJI

Summary: Key Findings on Crediting Loans

On performing the EDA for the banking datasets , the following findings were revealed:

- The proportion of defaulters is 8.0729% - The bank is in a healthy state with maximum number of non defaulters
- The bank lends more loans to females, the proportion of females in defaults in lower as compared to males – the bank can actively look for more male customers who satisfy other criteria as well.
- Proportion of working is highest in paying their loans on time, followed by commercial associate type
- People with higher education are the highest in loan paybacks – the bank should target them
- The Age group : 30-40 are highest in proportion for loan payments, followed by 40-60 age groups
- Clients with low credit amount tend to return the amount with ease followed by high and medium respectively – Bank should target the low credit promotion at first
- Bank can consider giving loan to Housing type - With Parents, as they are likely to give loan payment.
- Bank should get as many customers for loan with purpose 'Buying a home' and income type State Servant, as they are very likely to pay the loan
- Though 'Repair' Purpose have higher chance of loan repayment, but they also have highest chance of being a defaulter. So, bank should be more cautious in paying the loan for 'Repair' Purpose.

TABLE OF CONTENTS:

Problem Statement:

This case study aims to identify patterns which indicate if a client has difficulty paying their installments 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 applicants using EDA is the aim of this case study.

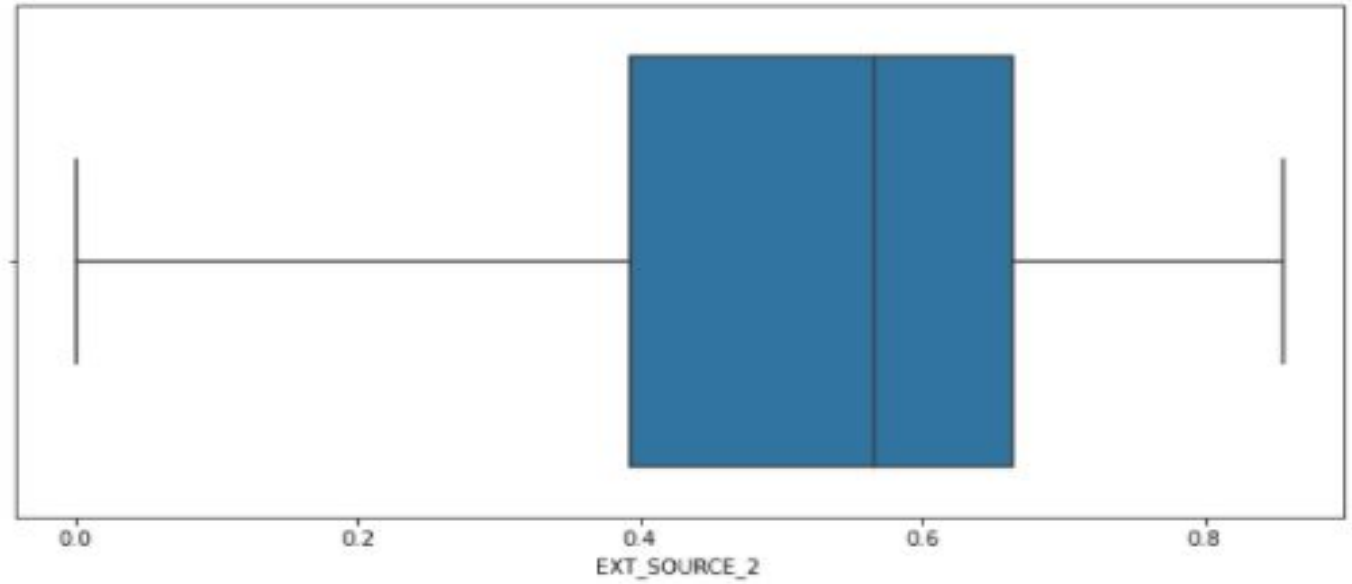
There are two types of risks associated with any loan request:

- H0 : If the applicant is likely to repay the loan , then not approving the loan results in a loss of the business to the company.

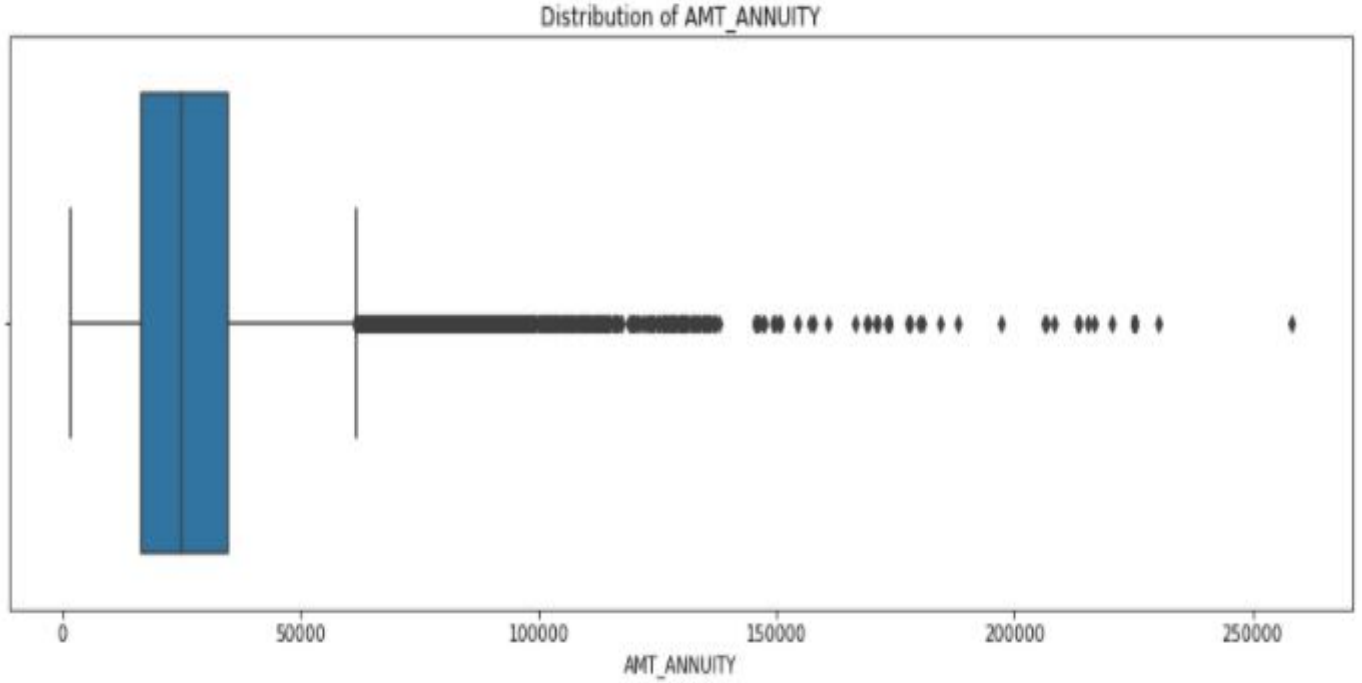
- H1: If the applicant is not likely to repay the loan(i.e. become a defaulter) , then approving the loan , may result in a financial loss to the bank .

Note: The detailed analysis of the datasets are done in a separate Jupyter Notebook .

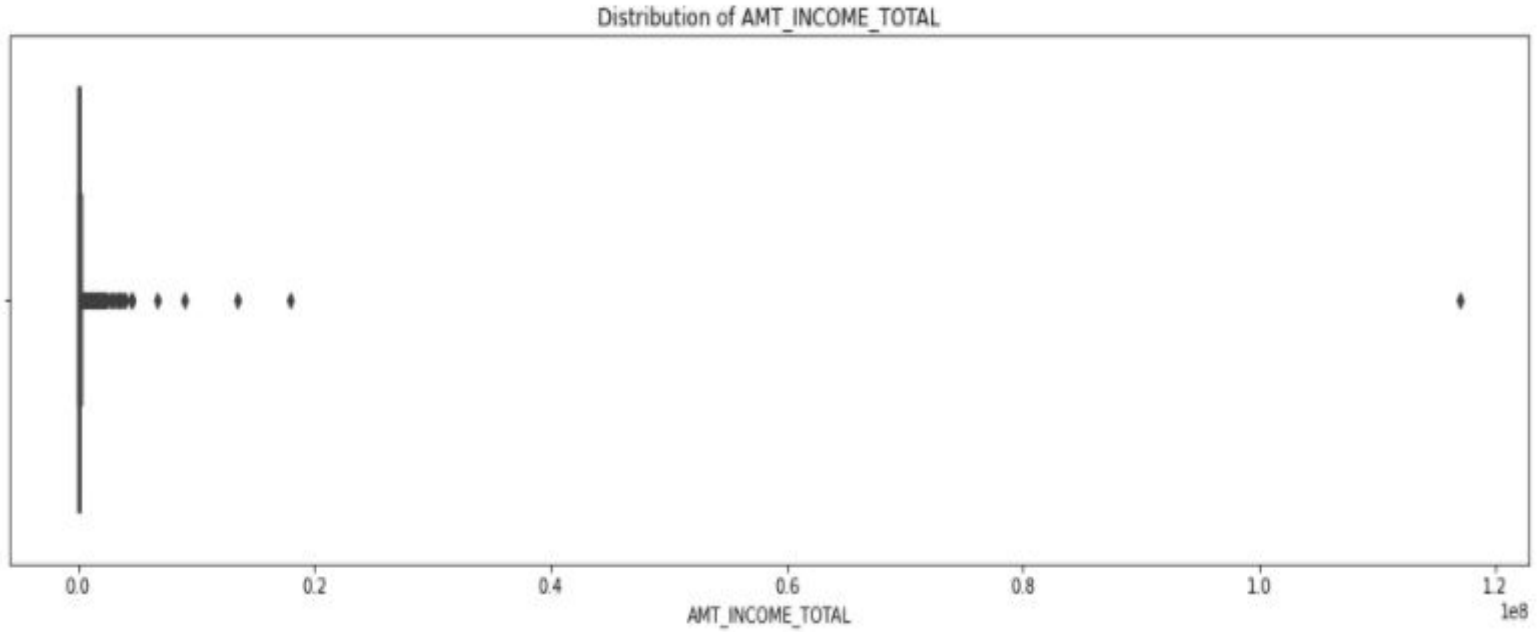
Dealing With Outliers



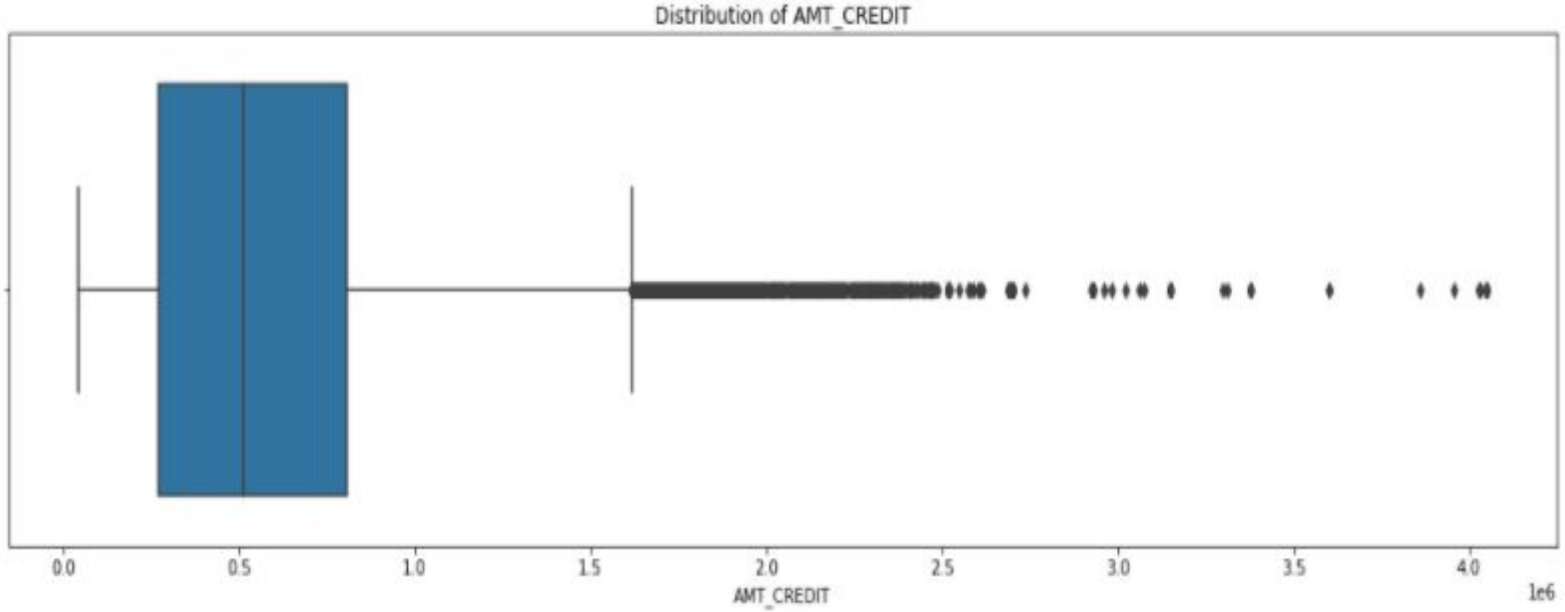
There are no outliers in the EXT_SOURCE_2 variable. The missing values are very less, so we can impute it with 0



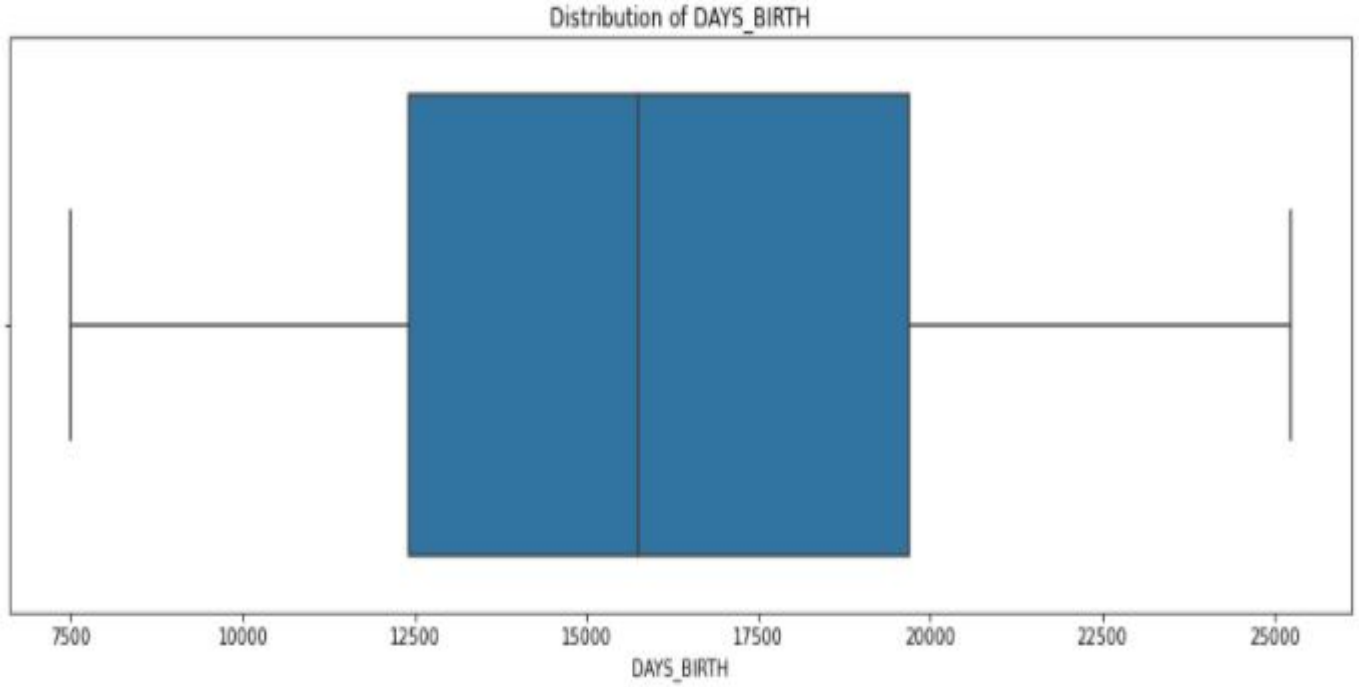
The outlier lie above 250,000. We can impute the outlier with Median of AMT_ANNUITY variable.



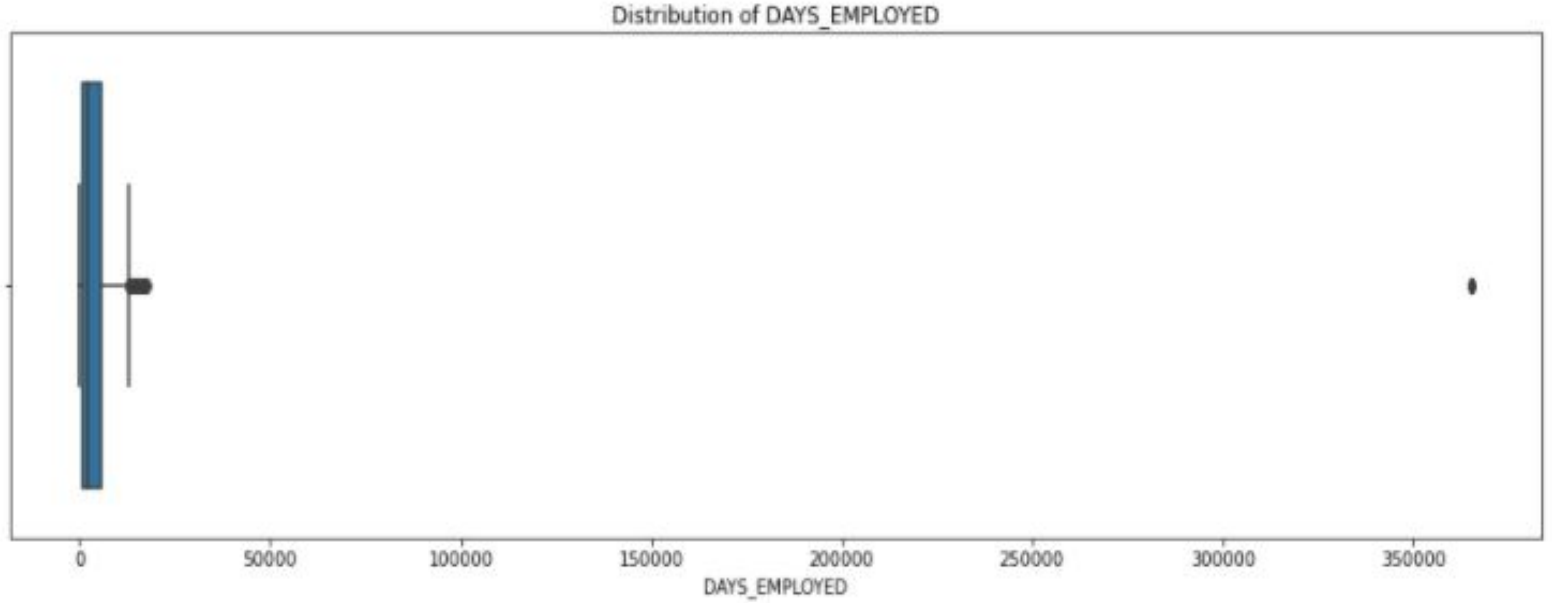
Here, we clearly see the outliers at extreme points i.e. max 1.700x10^8, which can bee seen from the description of variable and from the boxplot.



Outliers can be seen in the extreme ends near max and 98% quantiles.

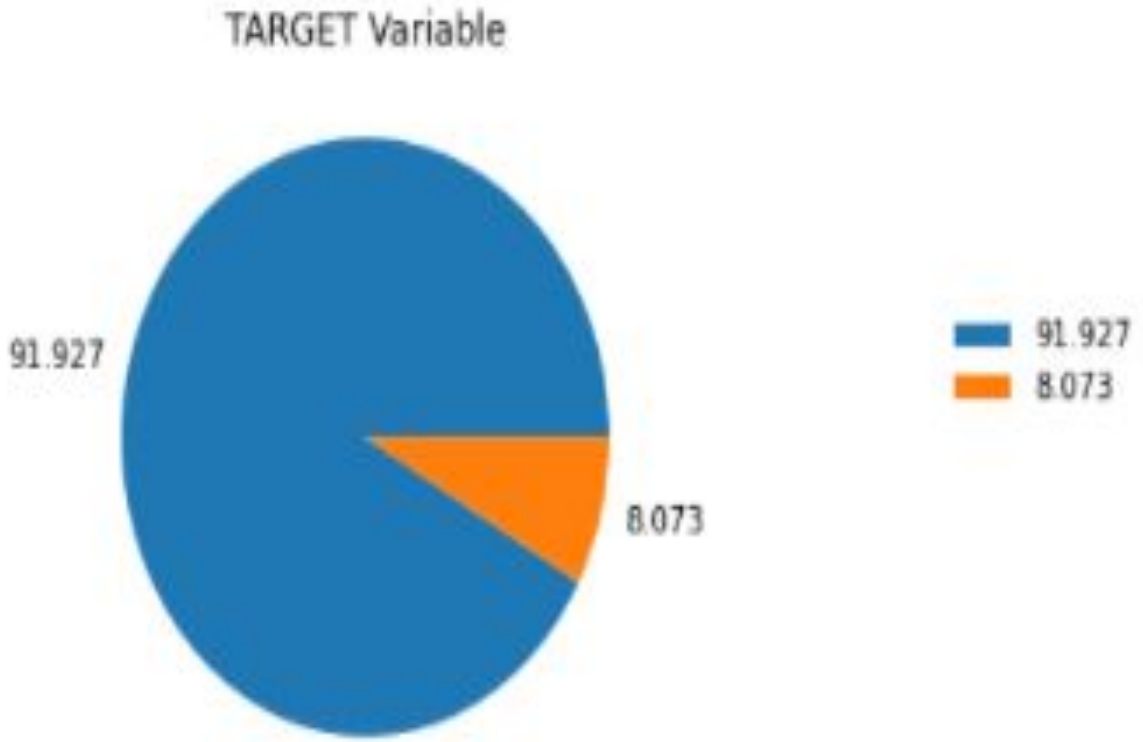


We can clearly see that there are no outliers. The Inter-Quartile Range(IQR) seems to properly distributed, which indicated that the DAYS_BIRTH variable has almost similar values.

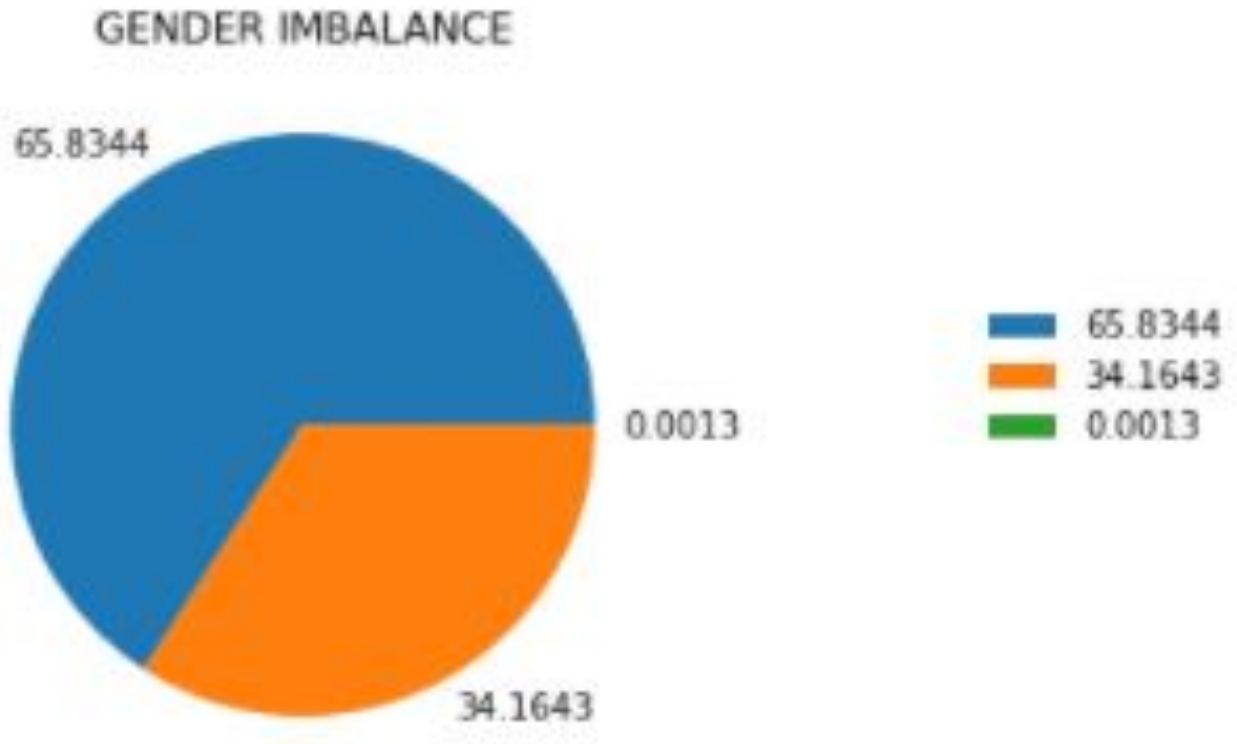


We can clearly see the outlier after 350,000 or near the max value i.e. 365,243. The DAYS_EMPLOYED shows the number of days a person is employed. It could be both High and Low. Here we can Cap the outliers.

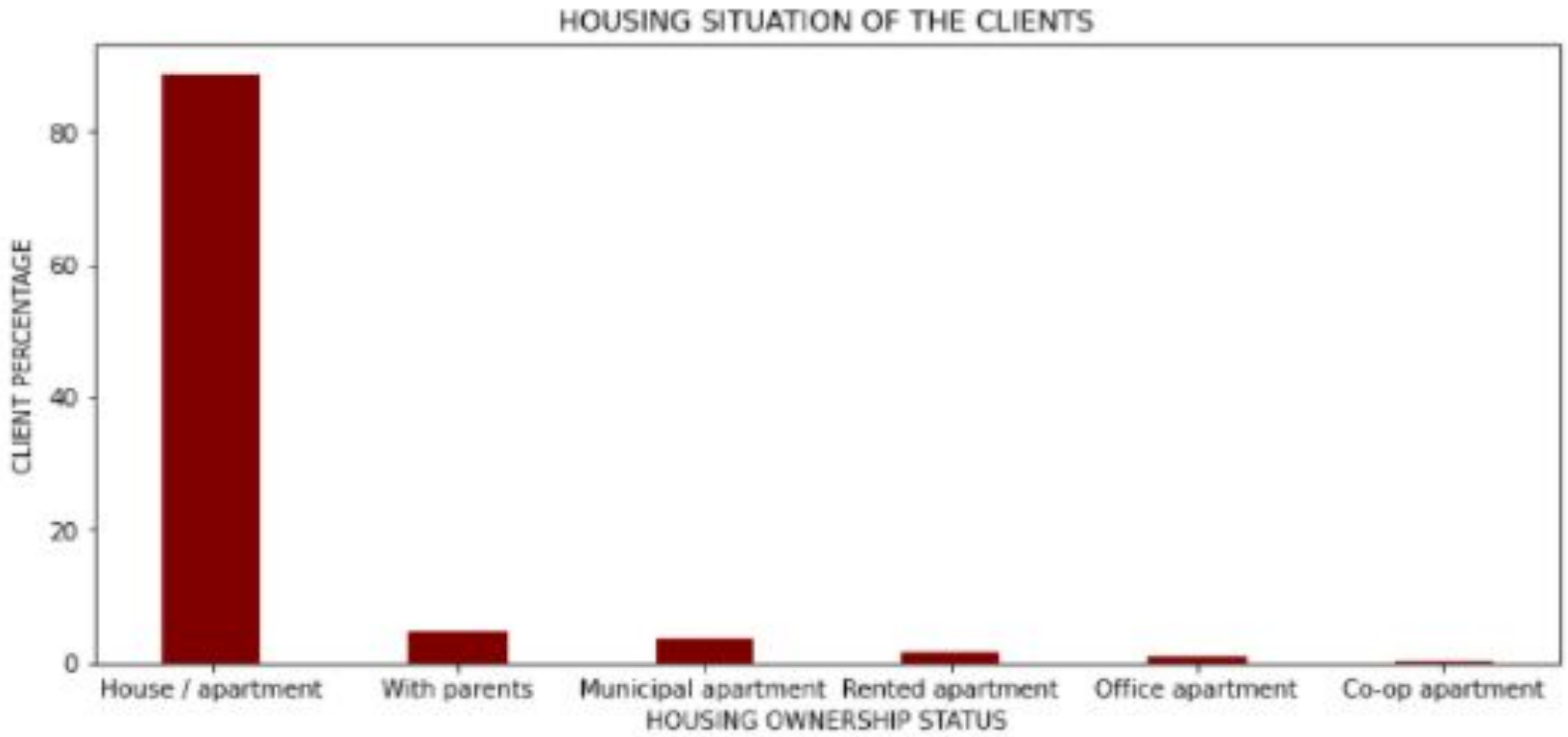
Checking the Imbalance Percentage



The above result indicated that 91.92% clients show no problem in payment, while the 8.0729% clients have problem in payment.



The above result indicated that 65.83% clients are females , while the 34.16% clients are males of the total customers. 0.0013% are junk data ,and is extremely small in proportion, hence can be ignored.



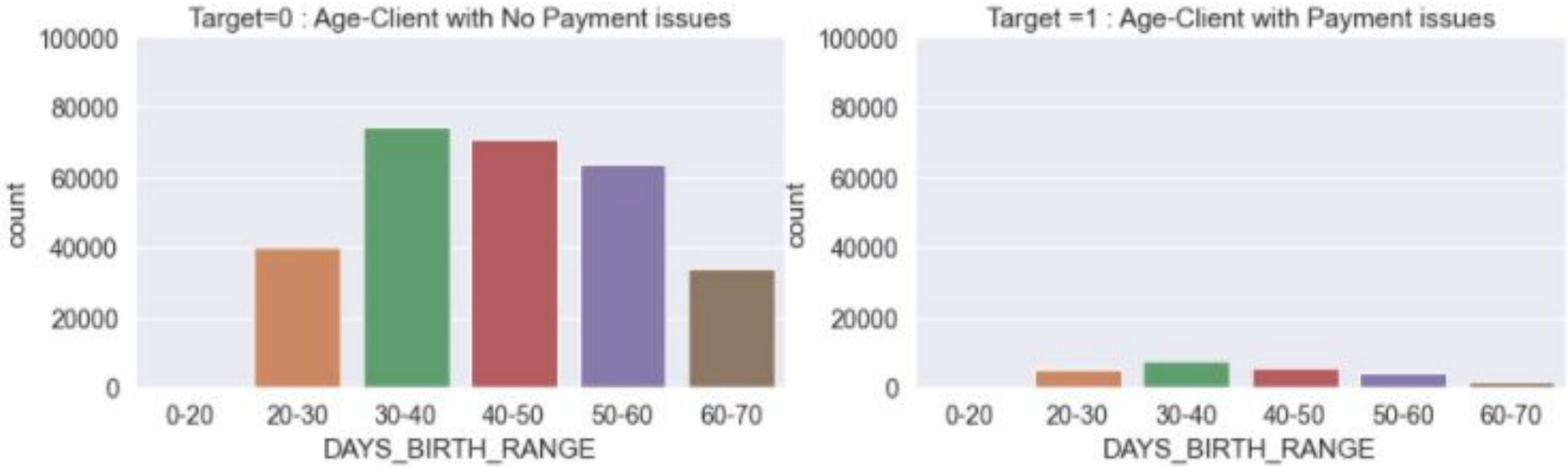
From the above plot, we get the below insights:

- House / apartment 88.7344%
- With parents 4.8258%
- Municipal apartment 3.6366%
- Rented apartment 1.5873%
- Office apartment 0.8510%
- Co-op apartment 0.3649%

Hence the bank can consider these percentages and target those groups without own house

Univariate Analysis on TARGET_1 and TARGET_0

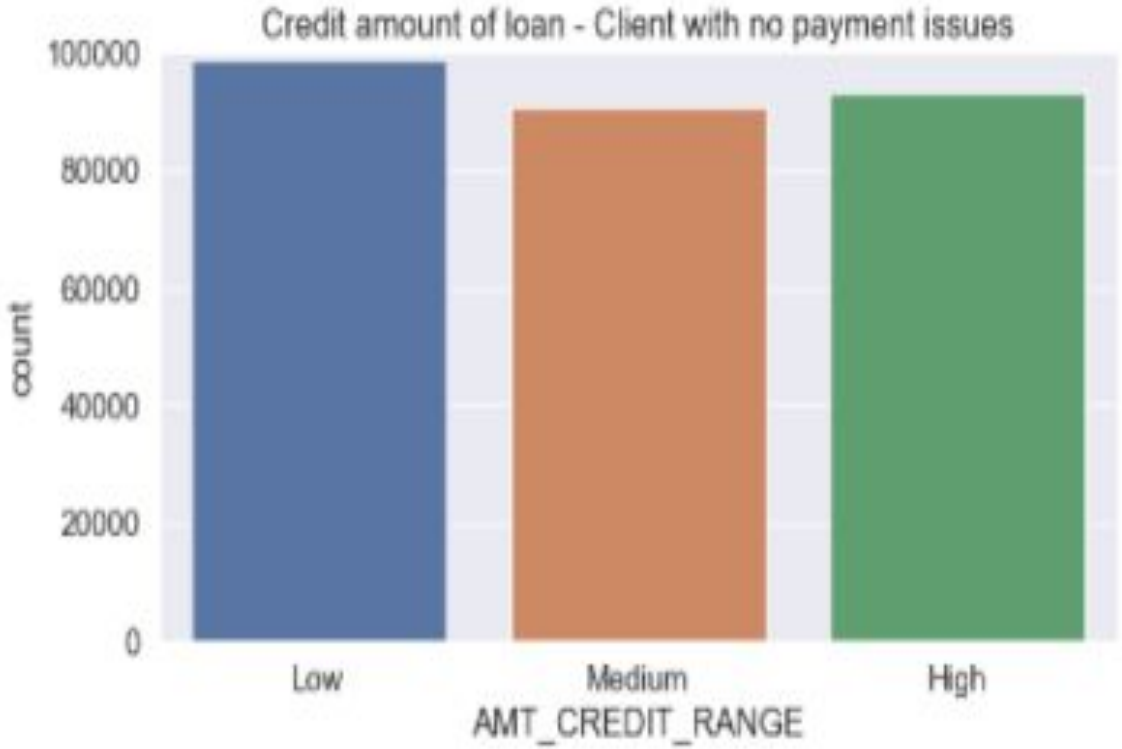
Analysis using AGE column(Binned Data):

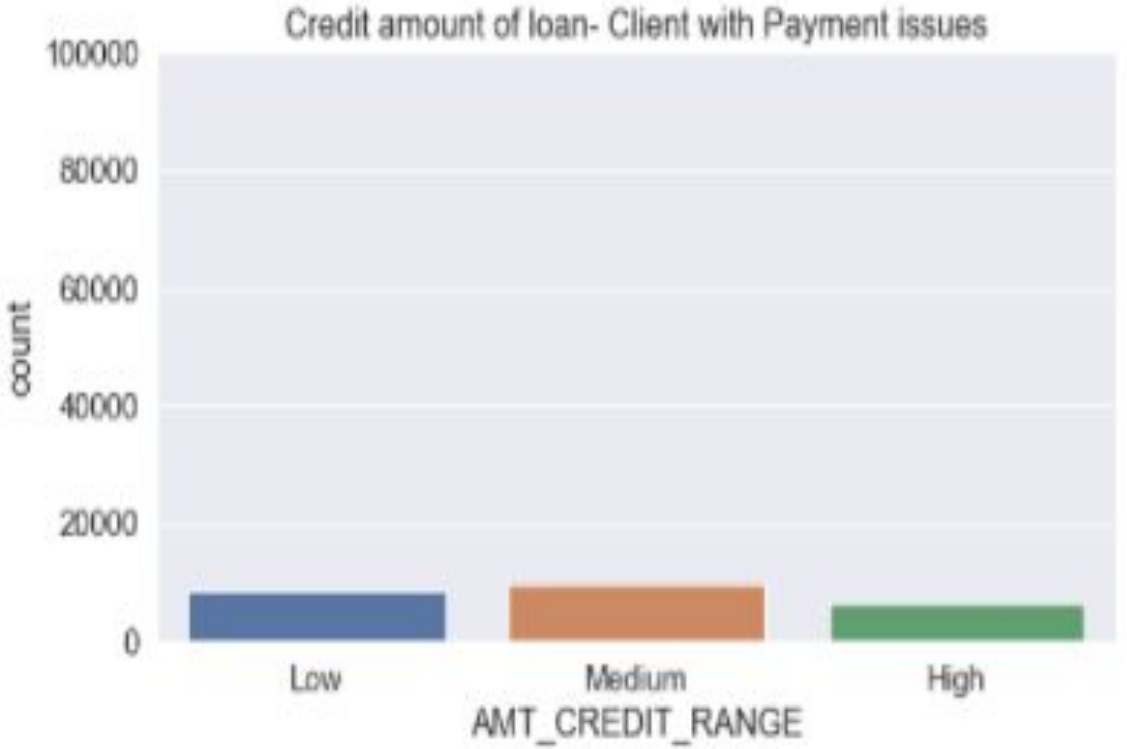


From the above plot, we can infer that the age group 30-40 have paid back the loans in the specified time without any issues. Hence this age range can be considered by the bank while lending loans.

Added to it, the graphs for the age groups 40-60 are on the higher side as well. Hence based on the age factor, this group can also be considered while lending loan.

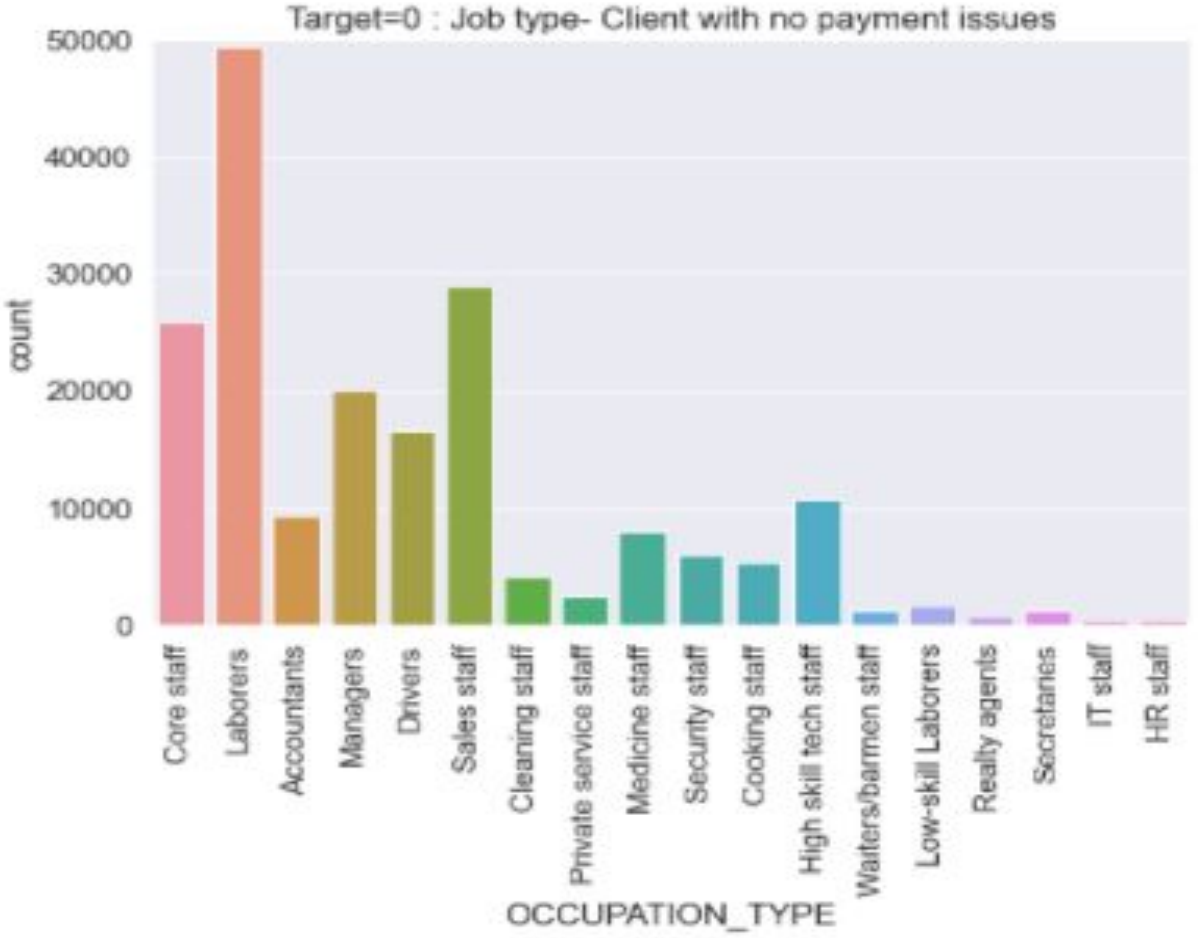
Analysis using AMT_CREDIT_RANGE column:

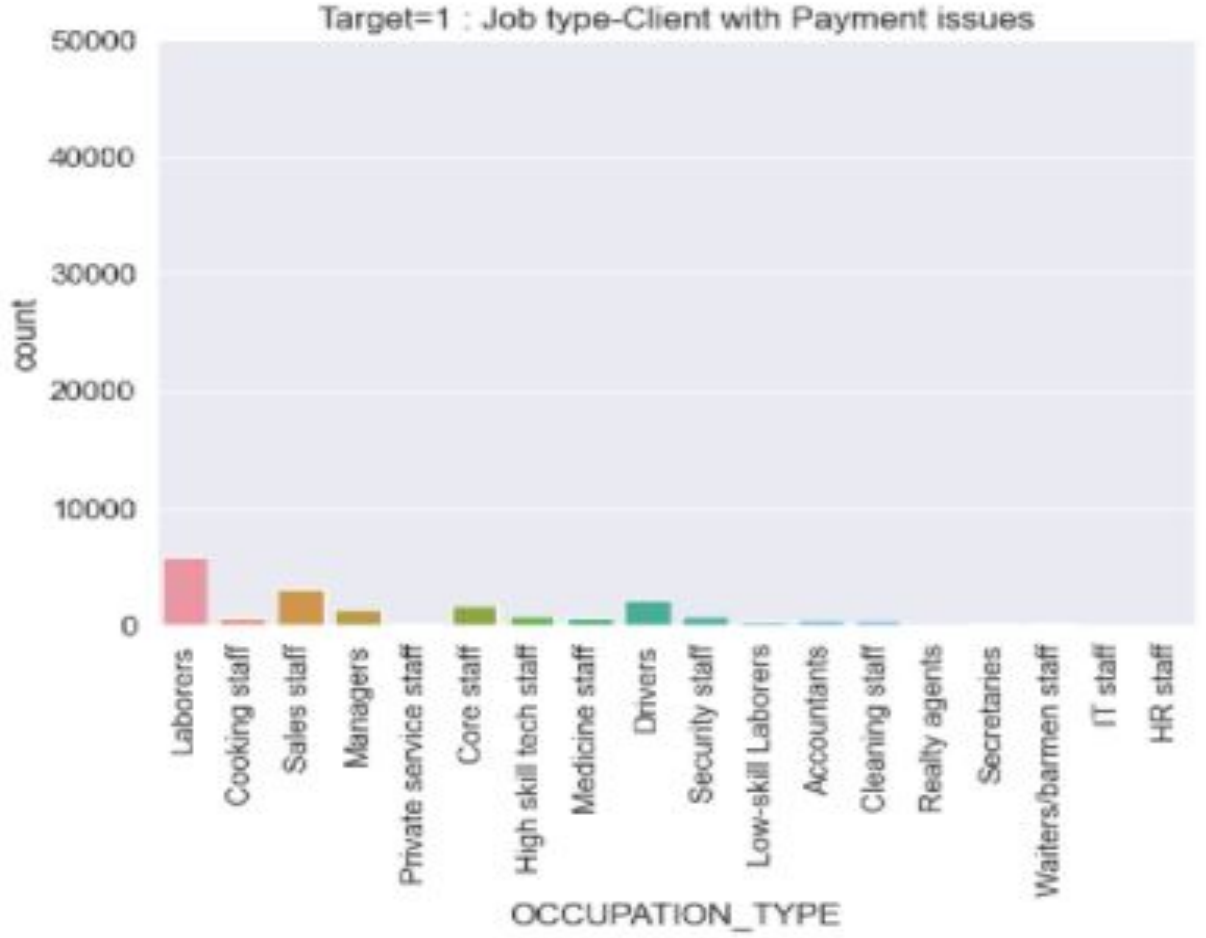




Inferring from the graph, Clients with low credit amount tend to return the amount with ease followed by high and medium respectively. Hence, all the three categories (prioritising the low credit amount) should be considered while lending the loans.

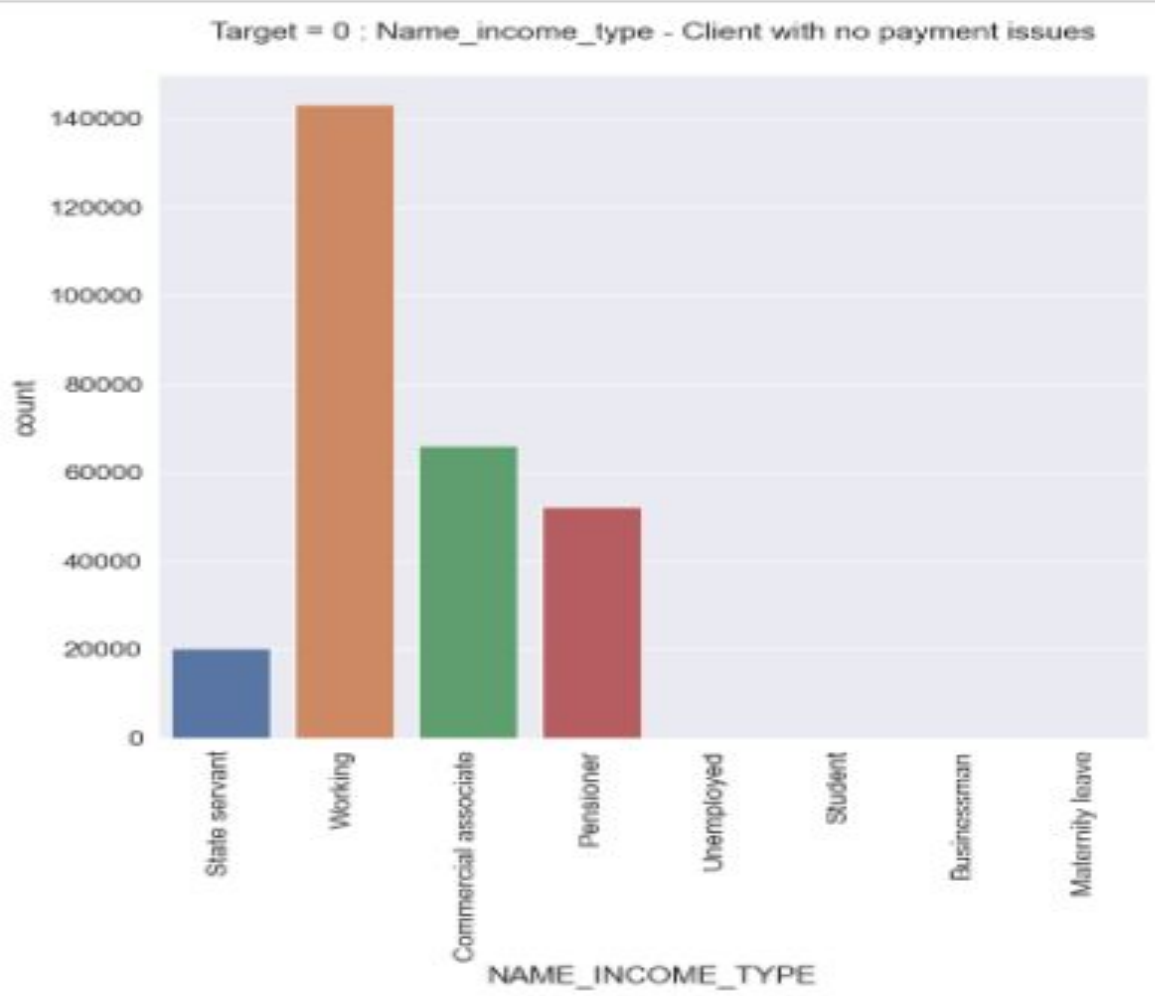
Analysis on Categorical Variables using OCCUPATION_TYPE column:

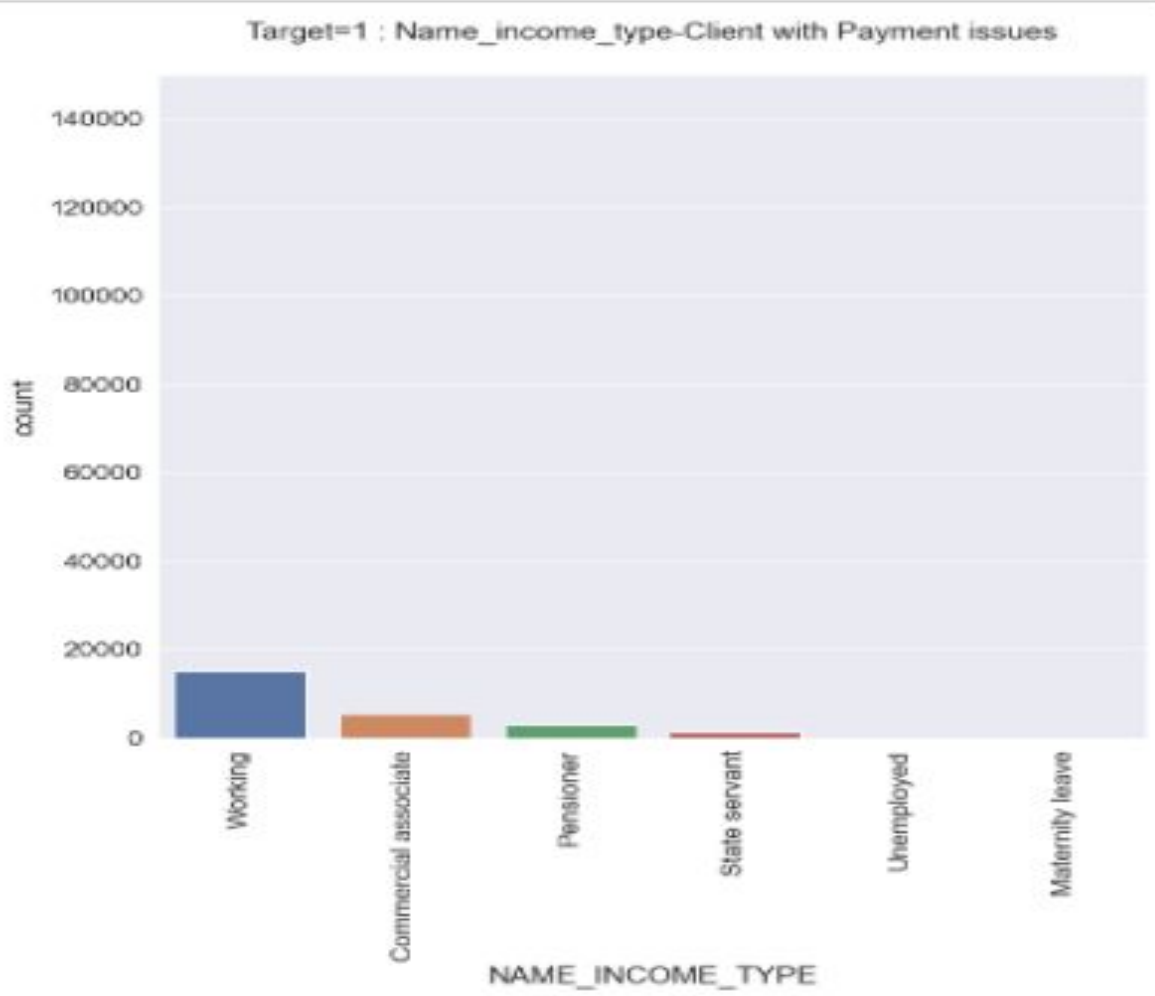




From the graph, it can be inferred that the laborers are most likely to pay the loans on time whereas the HR staffs are the least in number to pay the loan on time. Hence, the bank can consider the occupation type: labourer while lending loans.

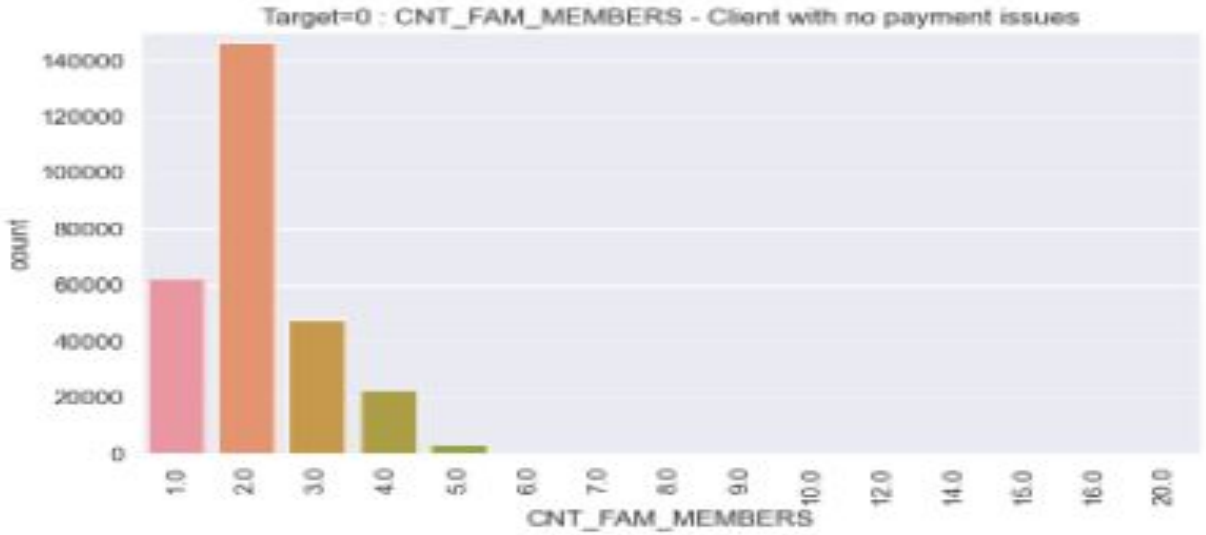
Analysis on Categorical Variables using INCOME_TYPE column:





From the graph we can clearly see that the working group are highest in number while making the payment on time without any issues, followed by the commercial associate group. Hence, the bank should give higher priority to the working group while lending loan.

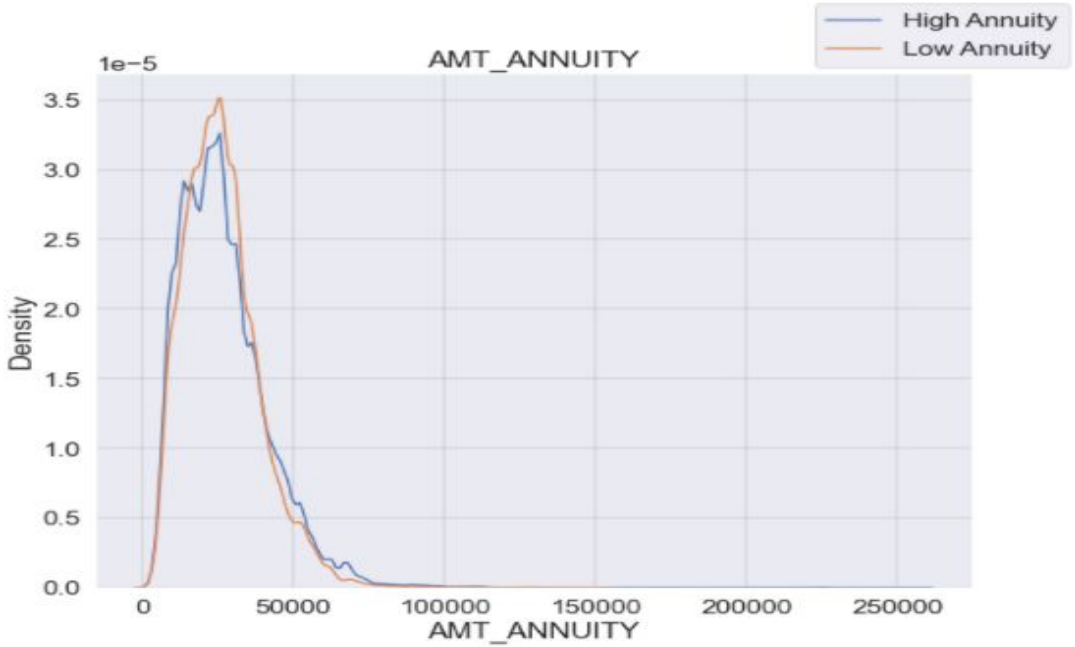
Analysis on CNT_FAM_MEMBERS VARIABLE :



As can be seen from the graph, the clients with family members as 2 have the highest count in payment being done on time. The count is least for those with 6,7 or 8 family members. Also, when we checked the imbalance percentage, it was found that the number of clients was almost = 50% of the total number of clients in the category CNT_FAM_MEMBERS=2.There might be chances that the data is getting baised. Hence, the bank may consider targetting more clients in other groups as well.

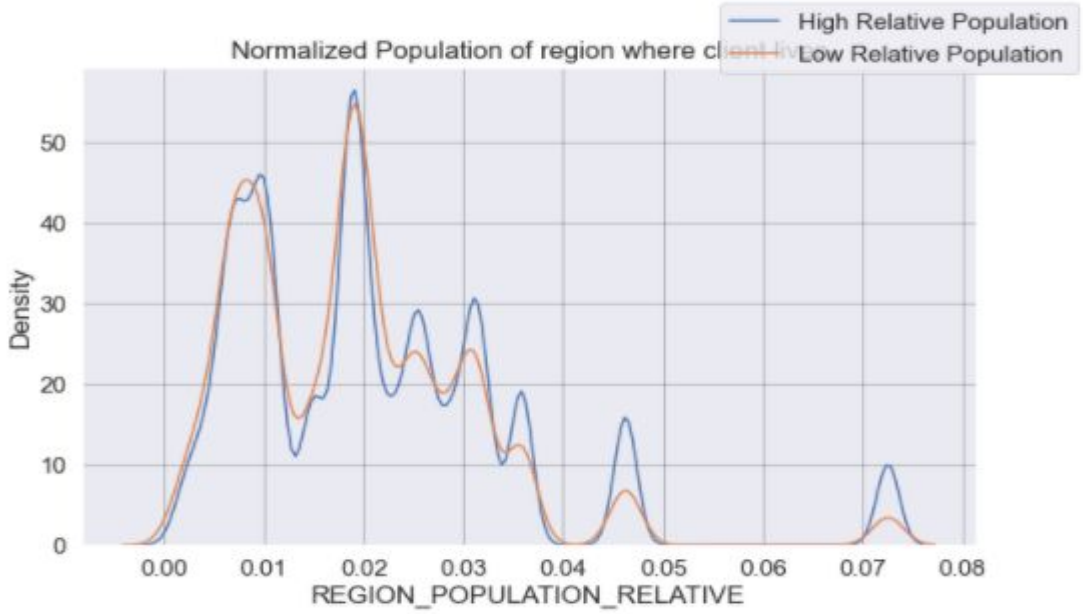
Analysis of the continuous columns with respect to the target column

The graph almost overlaps for both high amount credits as well as the low amount credits. However with higher loan amount credit the return of loans also increases.
Column Used: AMT_CREDIT



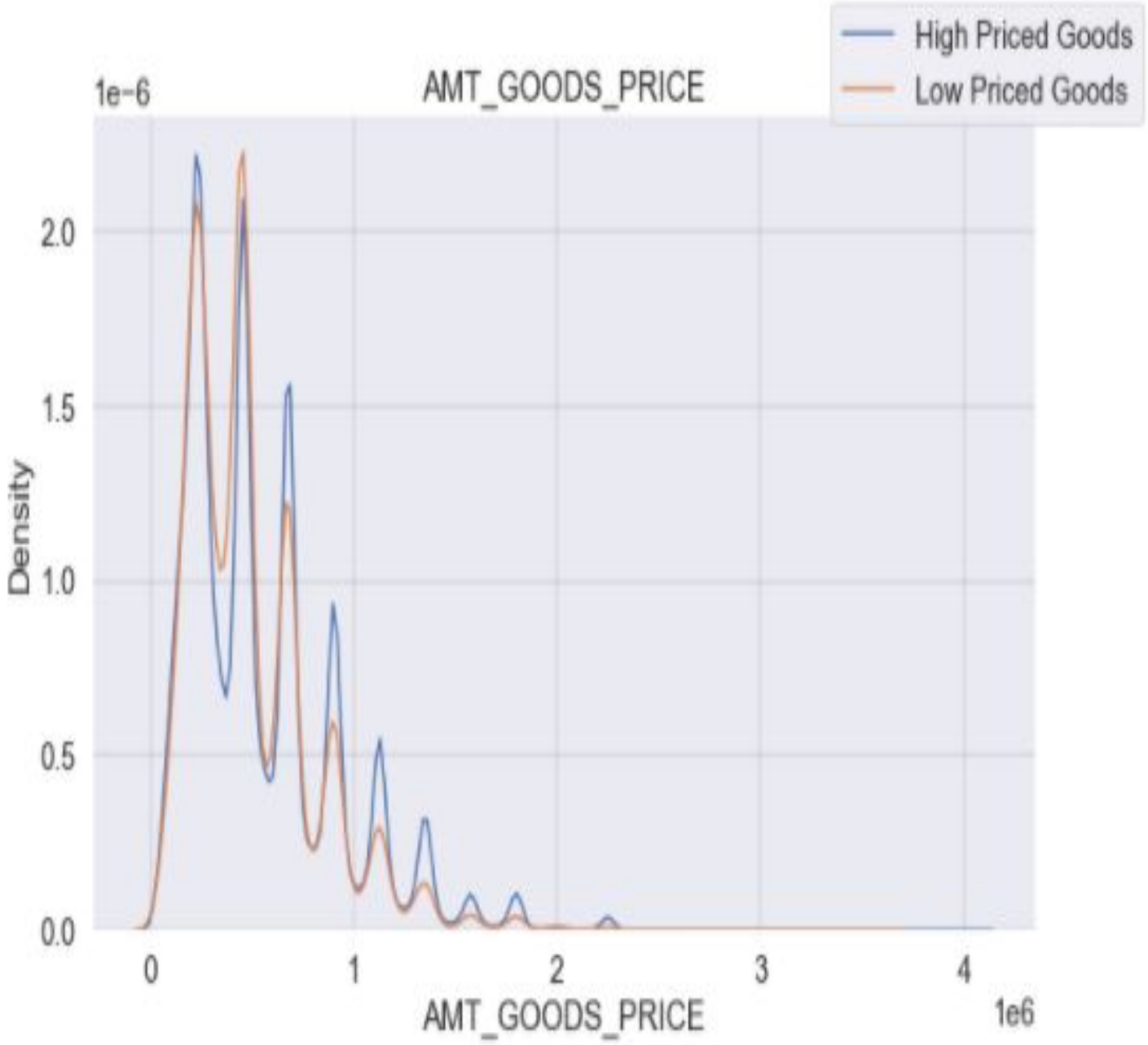
As seen from the plot, the graph almost overlaps for both High loan annuity as well as for low loan annuity. However, we can see that as the annuity amount increases the return of loan also becomes better.
Column Used: AMT_ANNUITY

From the graph it can be inferred that, with the increase in goods price the return of loans is better as compared to the low priced goods.
Column Used: AMT_GOODS_PRICE



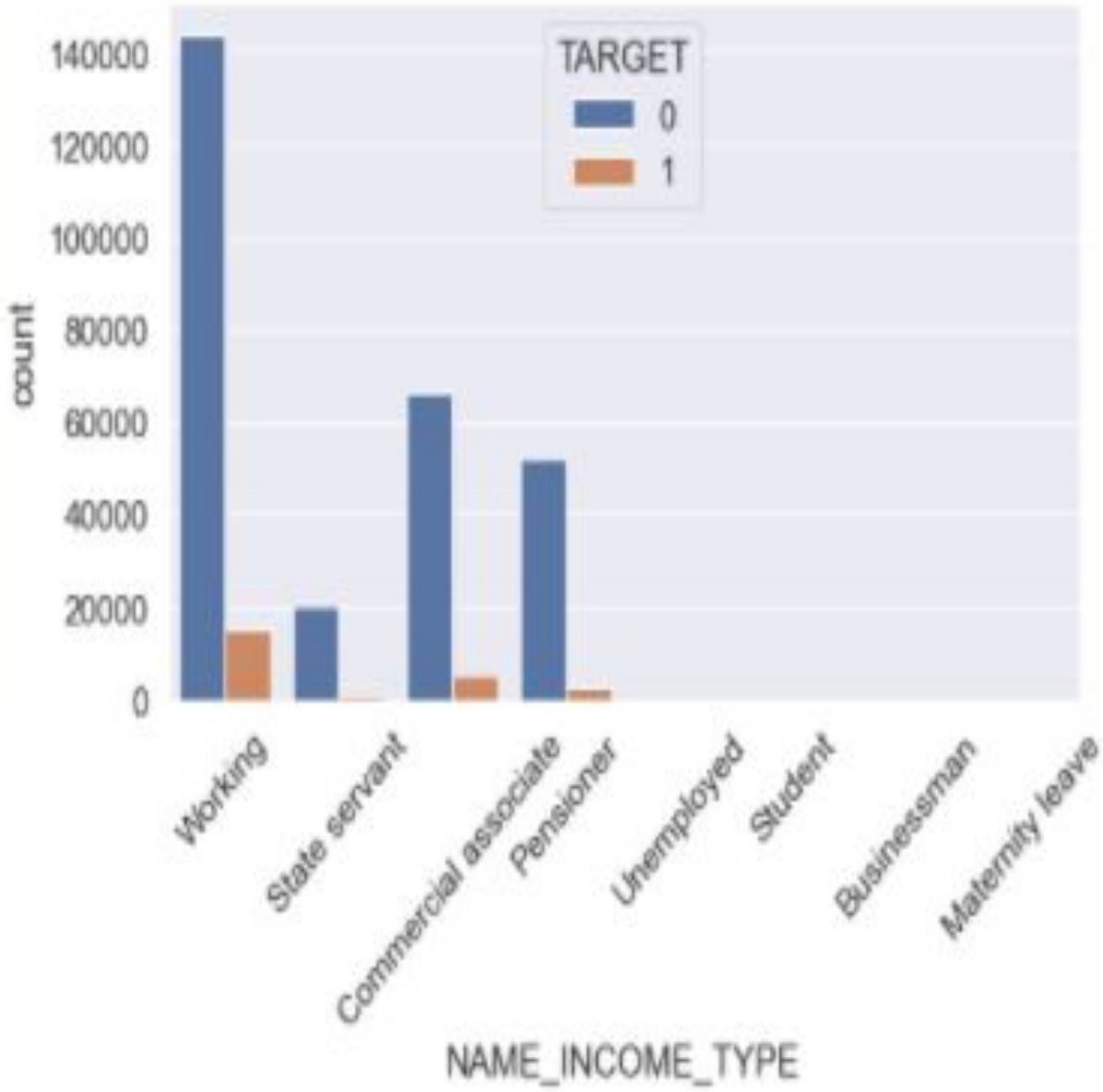
From the above plot, we can see that the return of loans are better for higher relative region population
Column Used: REGION_POPULATION_RELATIVE

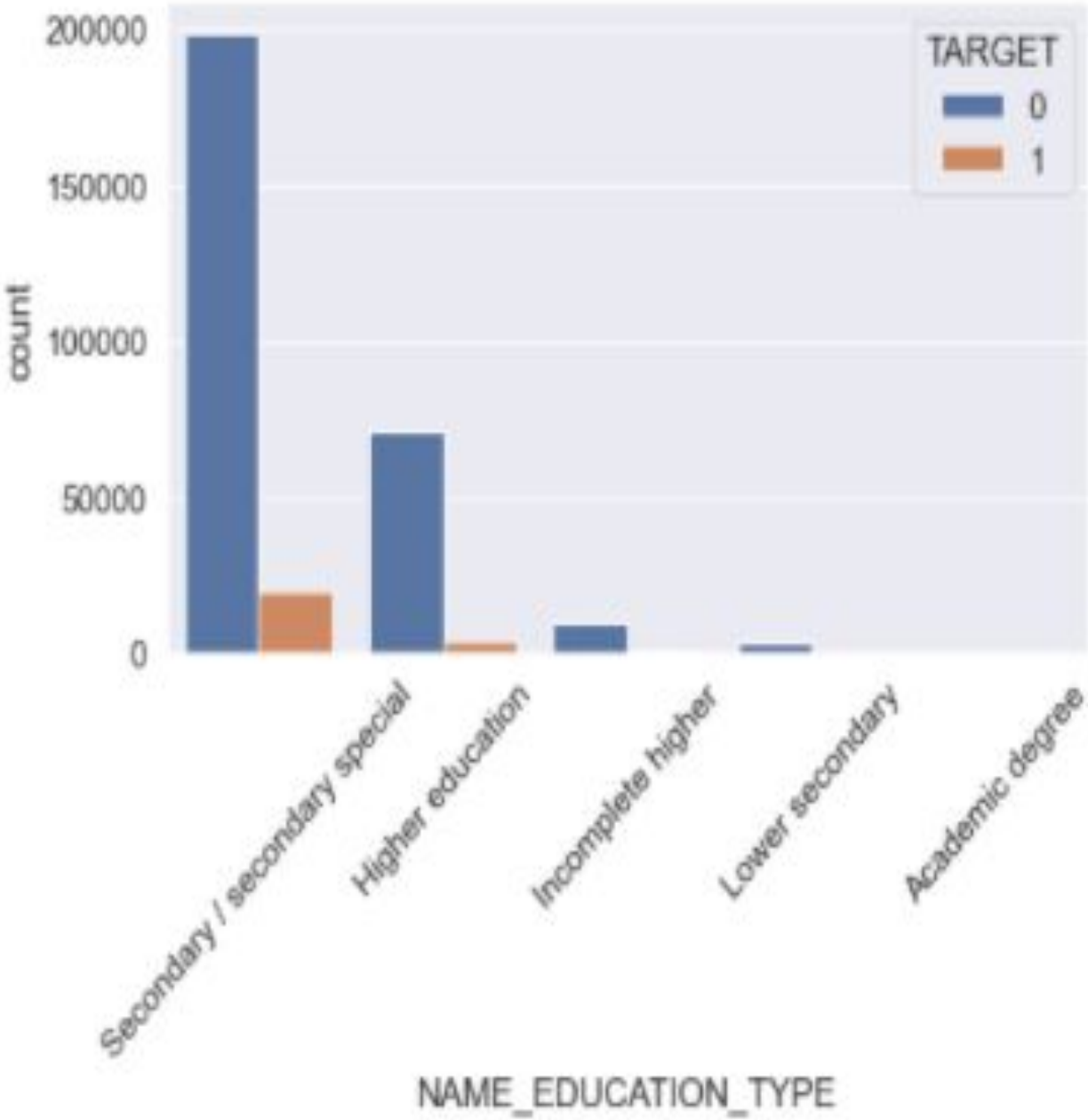
From the above plot, we can see that the return of loans are better for higher relative region population

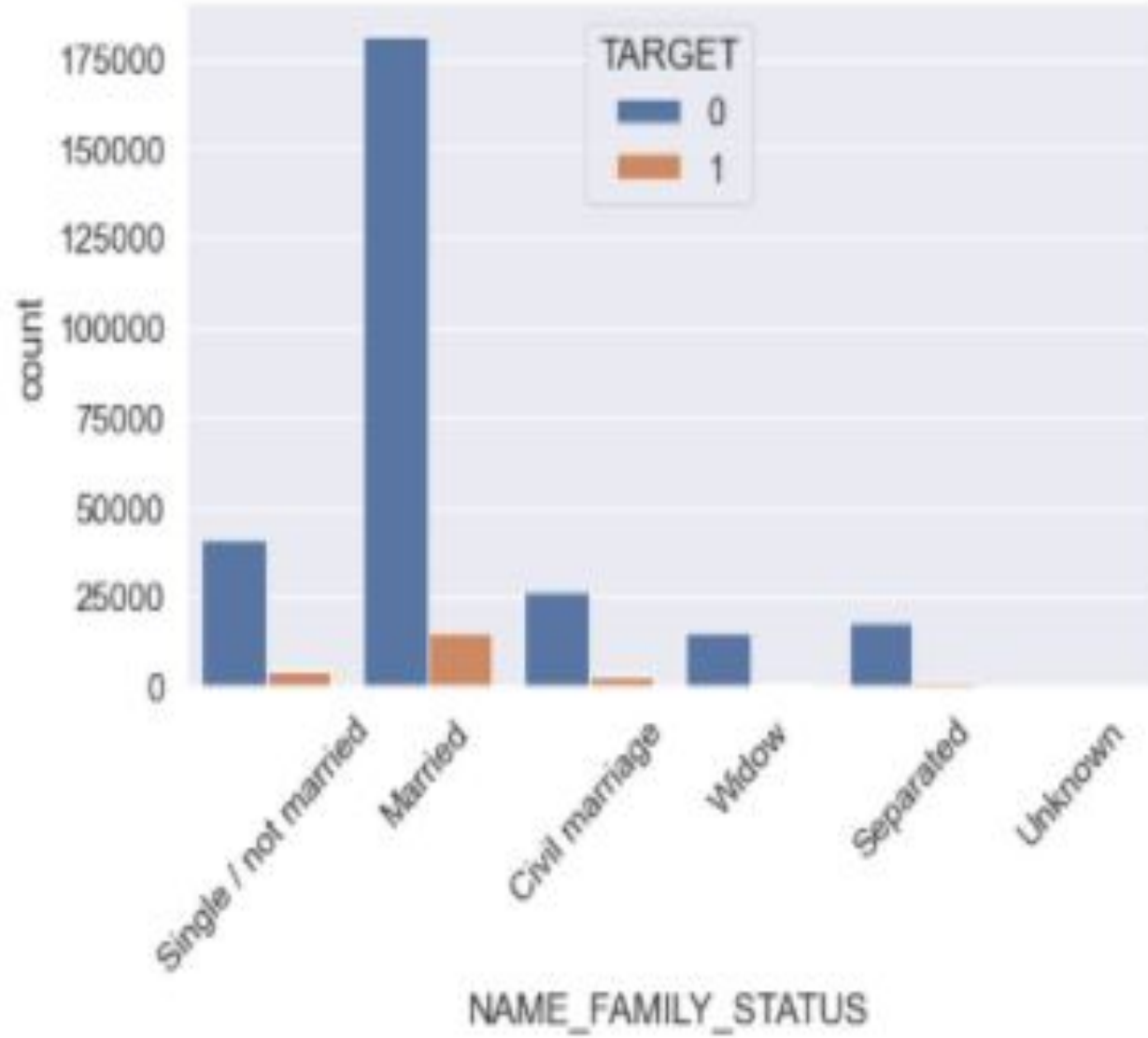


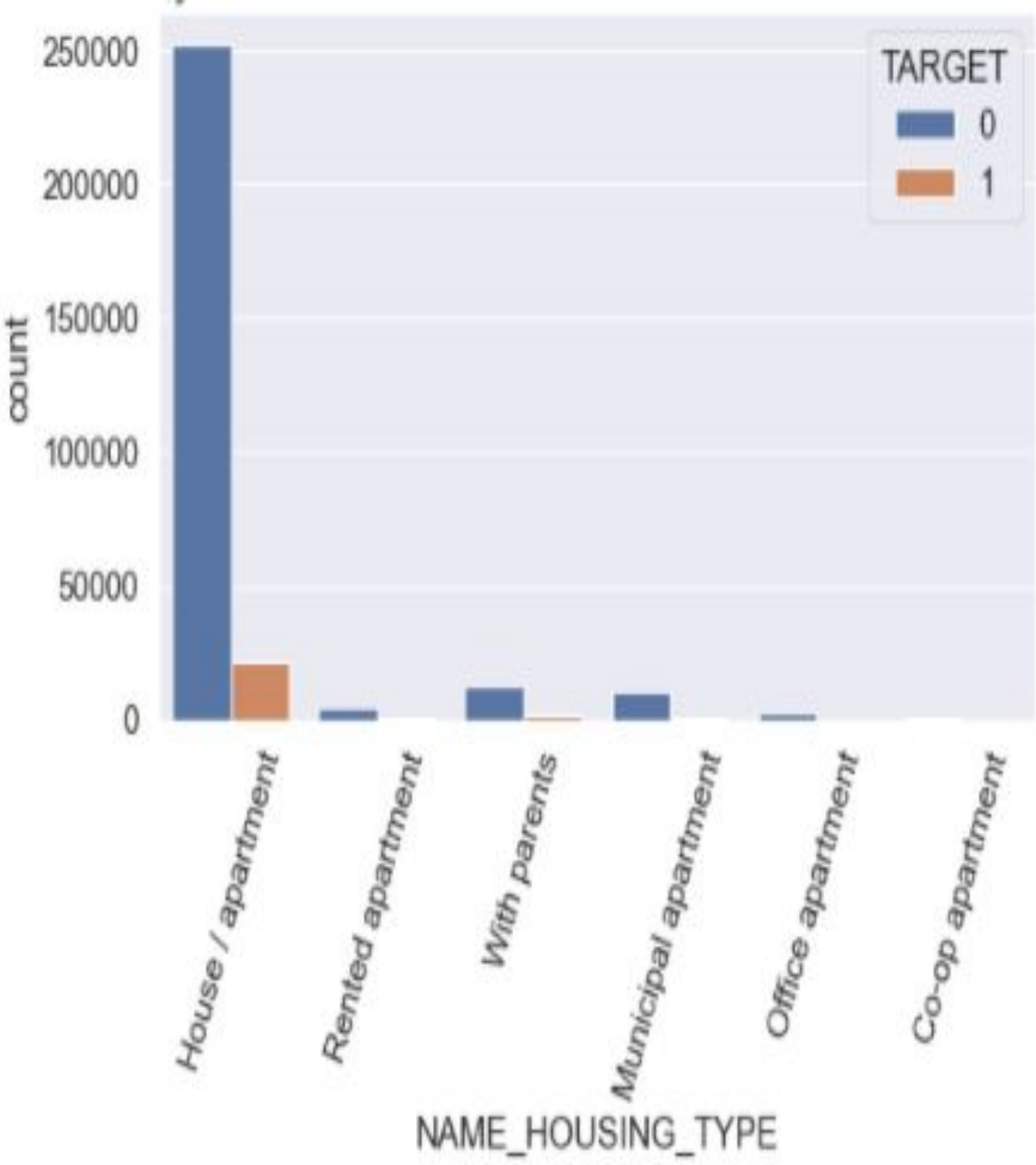
From the graph not much can be inferred, as we can see that the densities/count for both high priced goods and low priced goods almost overlap each other.
Column Used: AMT_GOODS_PRICE

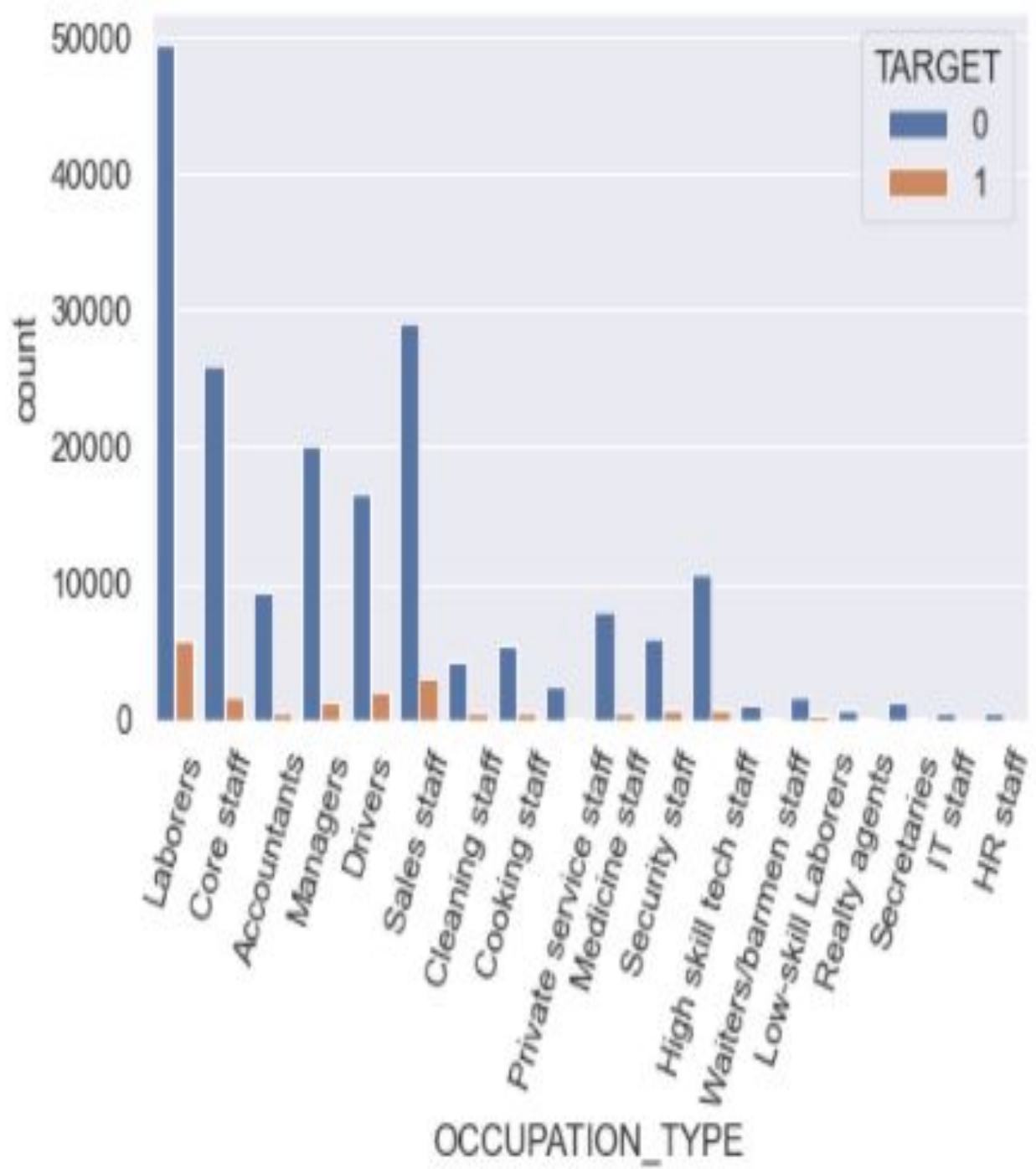
Analysis of multiple Categorical variables with respect to Target variable:




TARGET
0
1
count
200000
150000
100000
50000
0
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE


TARGET
0
1
count
175000
150000
125000
100000
75000
50000
25000
0
Single / not married
Married
Civil marriage
Widow
Separated
Unknown
NAME_FAMILY_STATUS


TARGET
0
1
250000
200000
150000
100000
50000
0
count
House / apartment
Rented apartment
With parents
Municipal apartment
Office apartment
Co-op apartment
NAME_HOUSING_TYPE


TARGET
0
1
50000
40000
30000
20000
10000
0
count
Laborers
Core staff
Accountants
Managers
Drivers
Sales staff
Cleaning staff
Cooking staff
Private service staff
Medicine staff
Security staff
High skill tech staff
Waiters/barmen staff
Low-skill Laborers
Realty agents
Secretaries
IT staff
HR staff
OCCUPATION_TYPE

The following inferences can be drawn from the above 6 subplots:

- From the first plot, it can be clearly seen that the female are much more punctual in paying their loan amounts as compared to the males. Hence the bank can target the female group to lend the loans.

- Using the second plot, we can conclude that the working group have higher percentage of paying their loan amounts on time. Hence, this group should be targeted by the bank for lending the loans.

- From the third plot, we can see that the secondary education group have the highest percentage of paying their loans on time, showing that these groups can also be targeted by the bank.

- From the fourth plot, the married customers have timely paid their loans without being defaulters, hence these groups can also be targeted by the bank. On the other hand, the window group has least percentage of loan payment record.

- From the fifth graph, we see that the customers staying in house/apartments are most likely to pay their loans on time as the graph shows the highest peak for this category. Apparently, the co-co apartment category has the least percentage of loan payment records.

- For the sixth graph, the labourer occupation type shows the highest percentage of loan payment record making them most eligible for getting targeted by the bank to offer future loans. On the other hand, the HR staff department shows least percentage in terms of loan payment.

TARGET-0: Correlation matrix [Analysis] - for top10 correlations

	COL1	COL2	Correlation
440	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
154	AMT_GOODS_PRICE	AMT_CREDIT	0.99
302	CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
466	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86
155	AMT_GOODS_PRICE	AMT_ANNUITY	0.78
129	AMT_ANNUITY	AMT_CREDIT	0.77
233	DAYS_EMPLOYED	DAYS_BIRTH	0.63
128	AMT_ANNUITY	AMT_INCOME_TOTAL	0.42
153	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.35
202	DAYS_BIRTH	CNT_CHILDREN	0.34

INFERENCE:

From the matrix(for TARGET_0), it is pretty clear that the columns, OBS_60_CNT_SOCIAL_CIRCLE and OBS_30_CNT_SOCIAL_CIRCLE are most correlated as they hold a correlation value=1. Also, AMT_GOODS_PRICE and AMT_CREDIT hold a correlation value of 0.99 , which is also pretty high and it can be stated that these columns are highly correlated. AMT_GOODS_PRICE and AMT_CREDIT hold a correlation value of 0.78

TARGET-1: Correlation matrix [Analysis] - for top10 correlations

	COL1	COL2	Correlation
440	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
154	AMT_GOODS_PRICE	AMT_CREDIT	0.98
302	CNT_FAM_MEMBERS	CNT_CHILDREN	0.89
466	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87
129	AMT_ANNUITY	AMT_CREDIT	0.75
155	AMT_GOODS_PRICE	AMT_ANNUITY	0.75
233	DAYS_EMPLOYED	DAYS_BIRTH	0.58
441	OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.34
415	DEF_30_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.33
258	DAYS_REGISTRATION	DAYS_BIRTH	0.29

INFERENCE:

From the matrix (for TARGET_1), it is pretty clear that the columns, OBS_60_CNT_SOCIAL_CIRCLE and OBS_30_CNT_SOCIAL_CIRCLE are most correlated as they hold a correlation value=1. Also, AMT_GOODS_PRICE and AMT_CREDIT hold a correlation value of 0.98 , which is also pretty high and it can be stated that these columns are highly correlated. (AMT_ANNUITY and AMT_CREDIT) and (AMT_GOODS_PRICE and AMT_ANNUITY) both these combination of columns hold a correlation value of 0.75.

INFERENCE using the correlation values of TARGET_0 & TARGET_1 dataframes :

TARGET-0: Correlation matrix [Analysis] - for top10 correlation

	COL1	COL2	Correlation
440	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
154	AMT_GOODS_PRICE	AMT_CREDIT	0.99
302	CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
466	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86
155	AMT_GOODS_PRICE	AMT_ANNUITY	0.78
129	AMT_ANNUITY	AMT_CREDIT	0.77
233	DAYS_EMPLOYED	DAYS_BIRTH	0.63
128	AMT_ANNUITY	AMT_INCOME_TOTAL	0.42
153	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.35
202	DAYS_BIRTH	CNT_CHILDREN	0.34

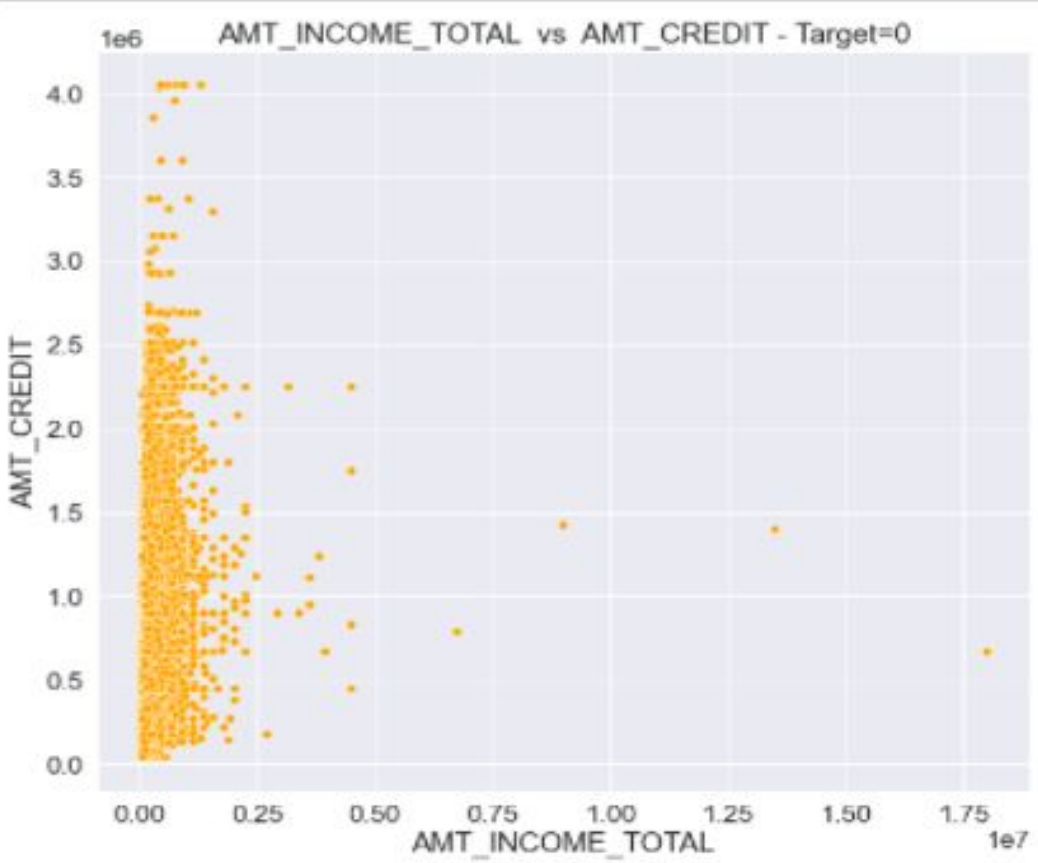
TARGET-1: Correlation matrix [Analysis] - for top10 correlation

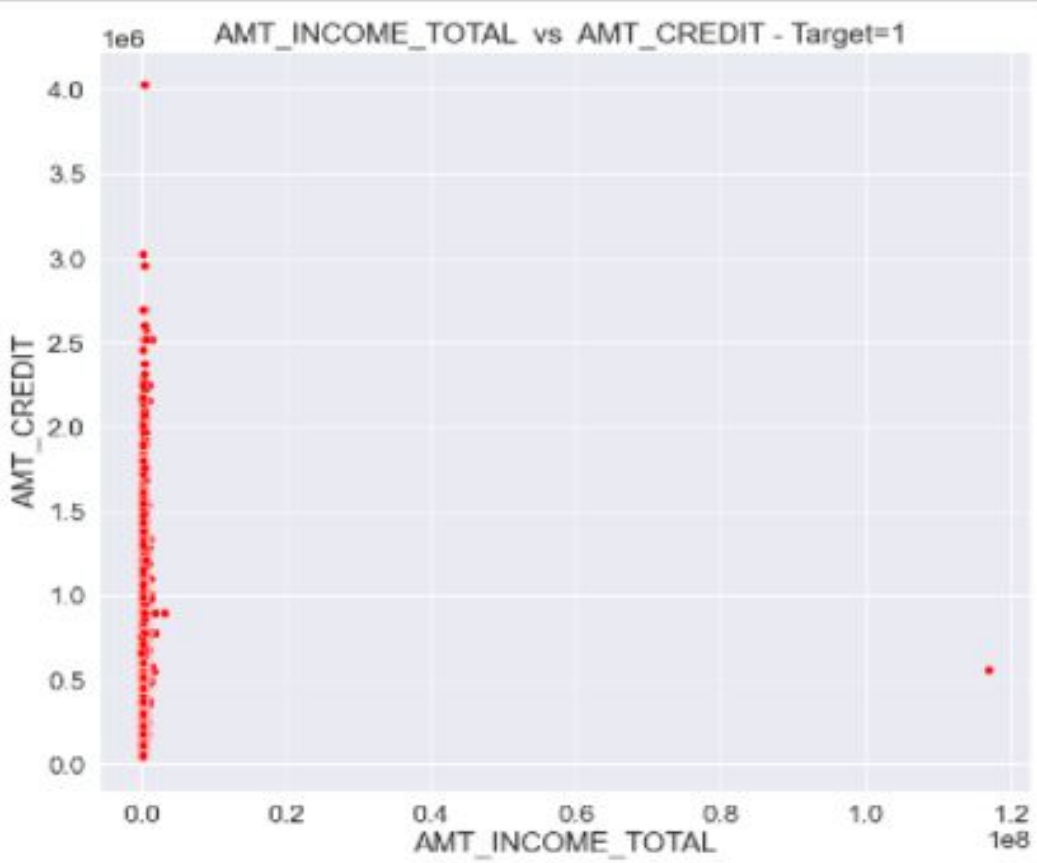
	COL1	COL2	Correlation
440	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
154	AMT_GOODS_PRICE	AMT_CREDIT	0.98
302	CNT_FAM_MEMBERS	CNT_CHILDREN	0.89
466	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87
129	AMT_ANNUITY	AMT_CREDIT	0.75
155	AMT_GOODS_PRICE	AMT_ANNUITY	0.75
233	DAYS_EMPLOYED	DAYS_BIRTH	0.58
441	OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.34
415	DEF_30_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.33
258	DAYS_REGISTRATION	DAYS_BIRTH	0.29

Hence, if we compare the top 10 variables of TARGET_0 and TARGET_1 dataframes, we can clearly state that the variables in both the dataframes are same with slight differences in their correlation values.

Bivariate Analysis for TARGET_0 and TARGET_1

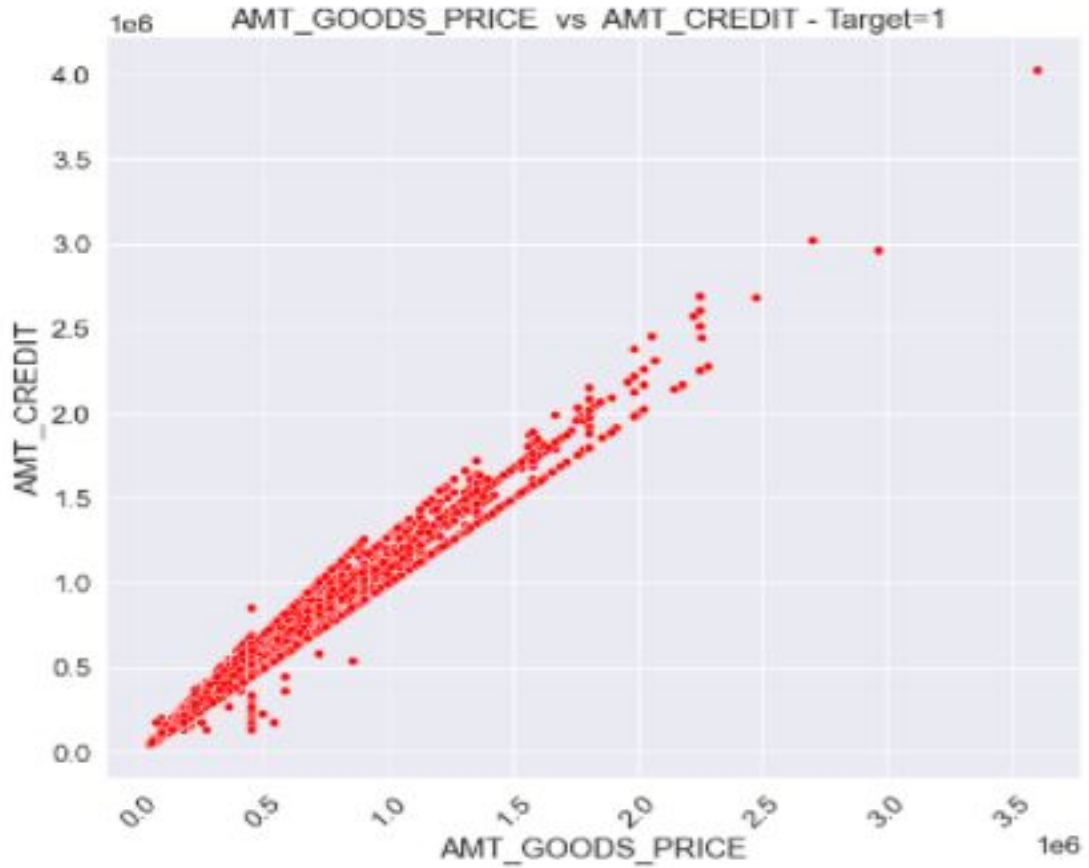
AMT_INCOME vs AMT_CREDIT:





From the above graphs, it can be inferred that, those who have paid the loan amount on or within the specified timeline, are more likely to get higher credits than those who failed to pay or did late payments .

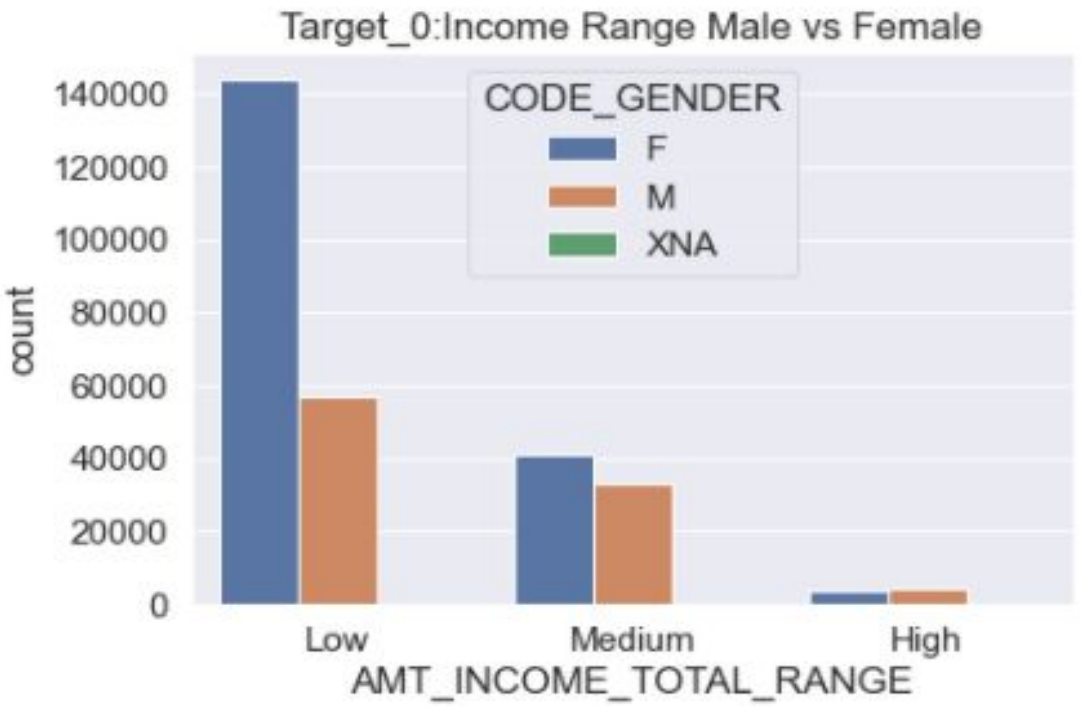
AMT_GOODS_PRICE vs AMT_CREDIT :

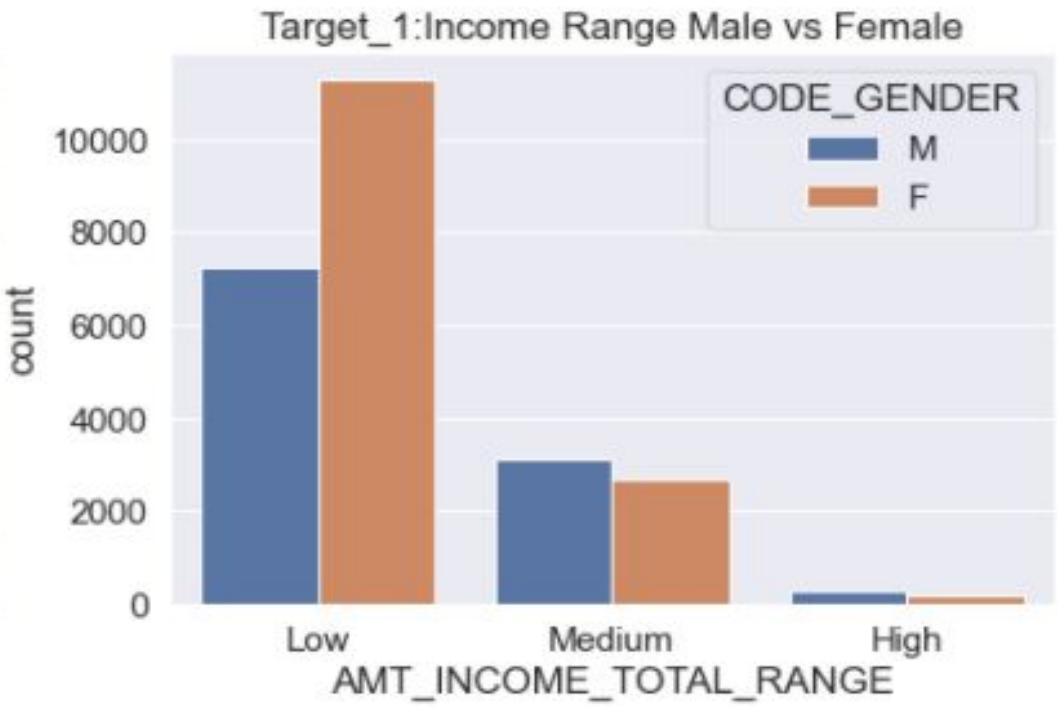


Also, the customers who have higher goods price and have made payments on time have higher credits than those with higher goods price but failed to pay the loan and became defaulters.

Numerical categorical analysis

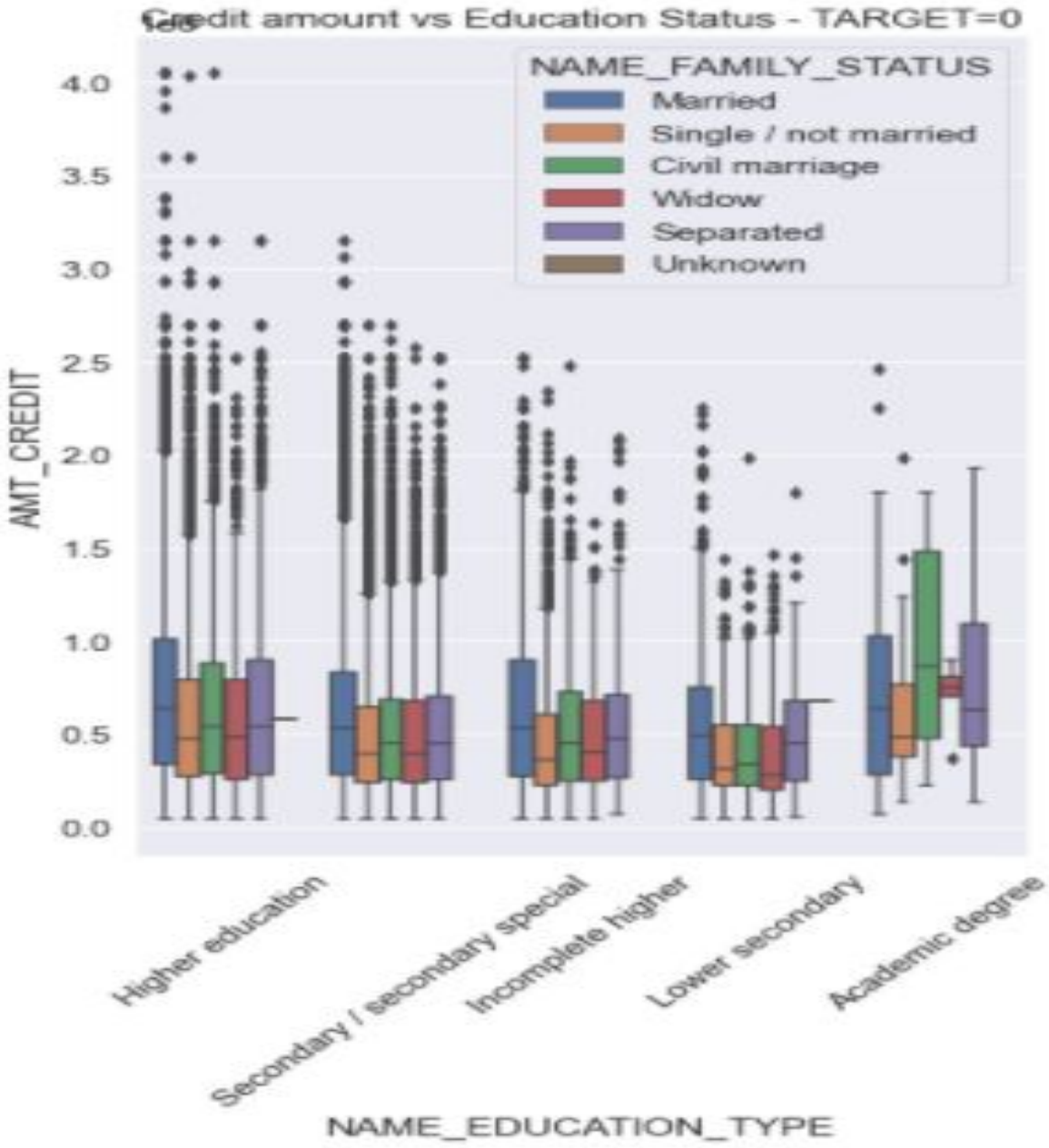
Analysis on Income range- Gender:

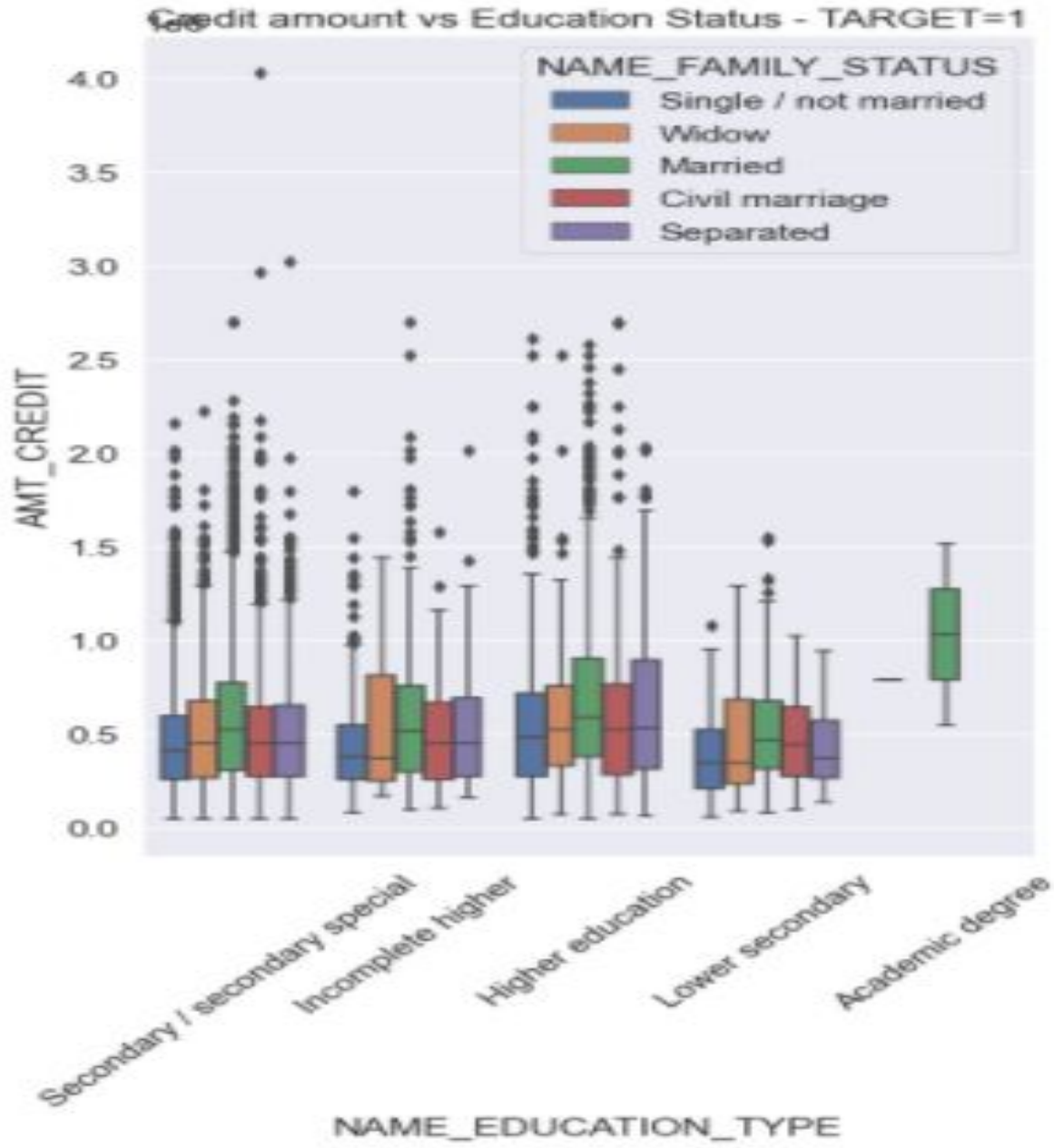




It can be clearly seen that female with low income range have higher percentage of loan payment record as compared to male. On the other hand, with reference to the second plot, males with low income range have higher percentage records i.e. more defaulters are there in this case.

Analysis of Credit amount vs Education Status:





INFERENCES:

Based on the numerical analysis done on (Credit amount vs Education Status) it can be inferred that:

1. Some customers belonging to the academic degree, married category are having higher credits than the others in the same category.

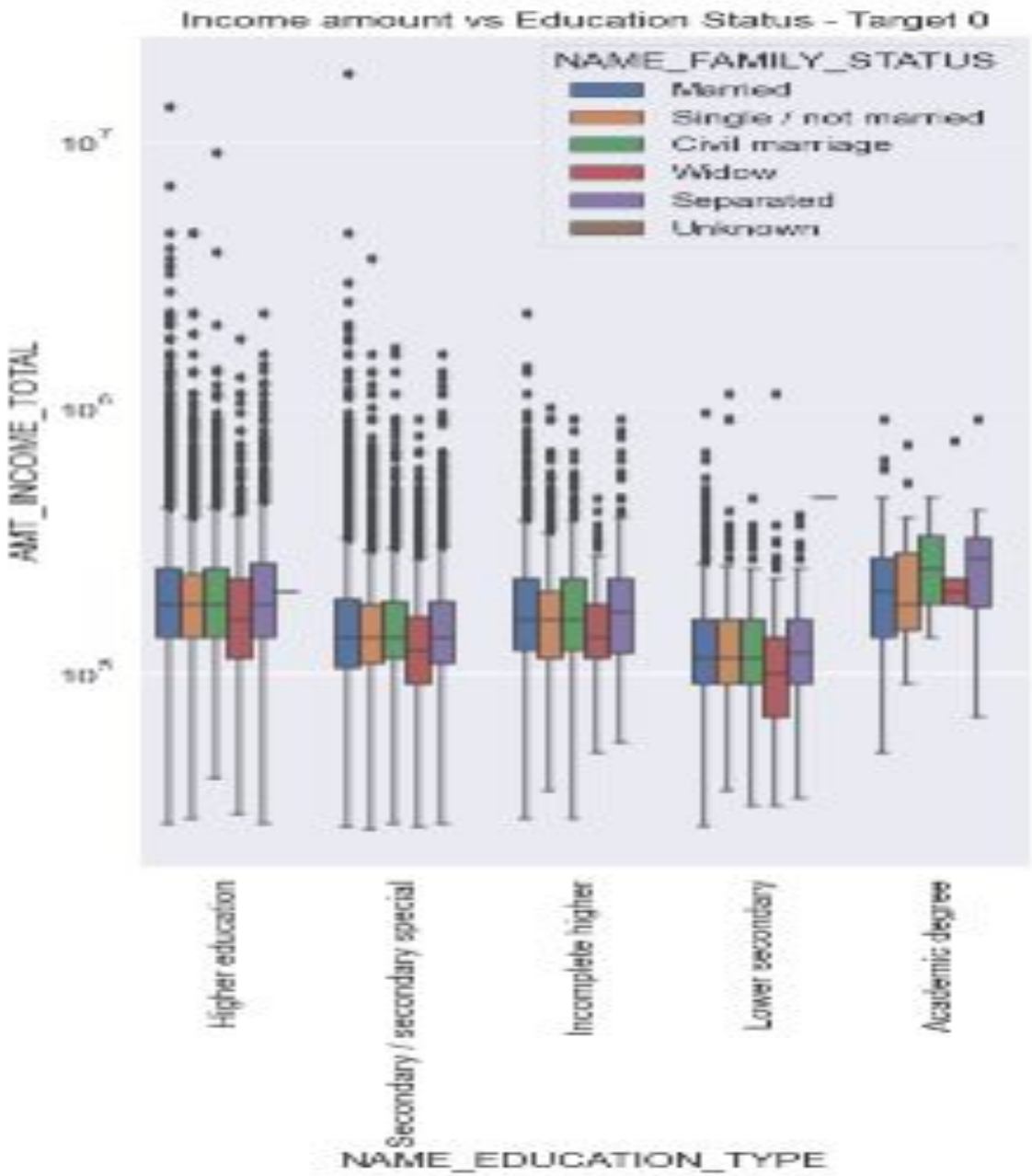
2. Similarly, some of the highly educated, married customers are having credits higher than those who have done lower secondary education.

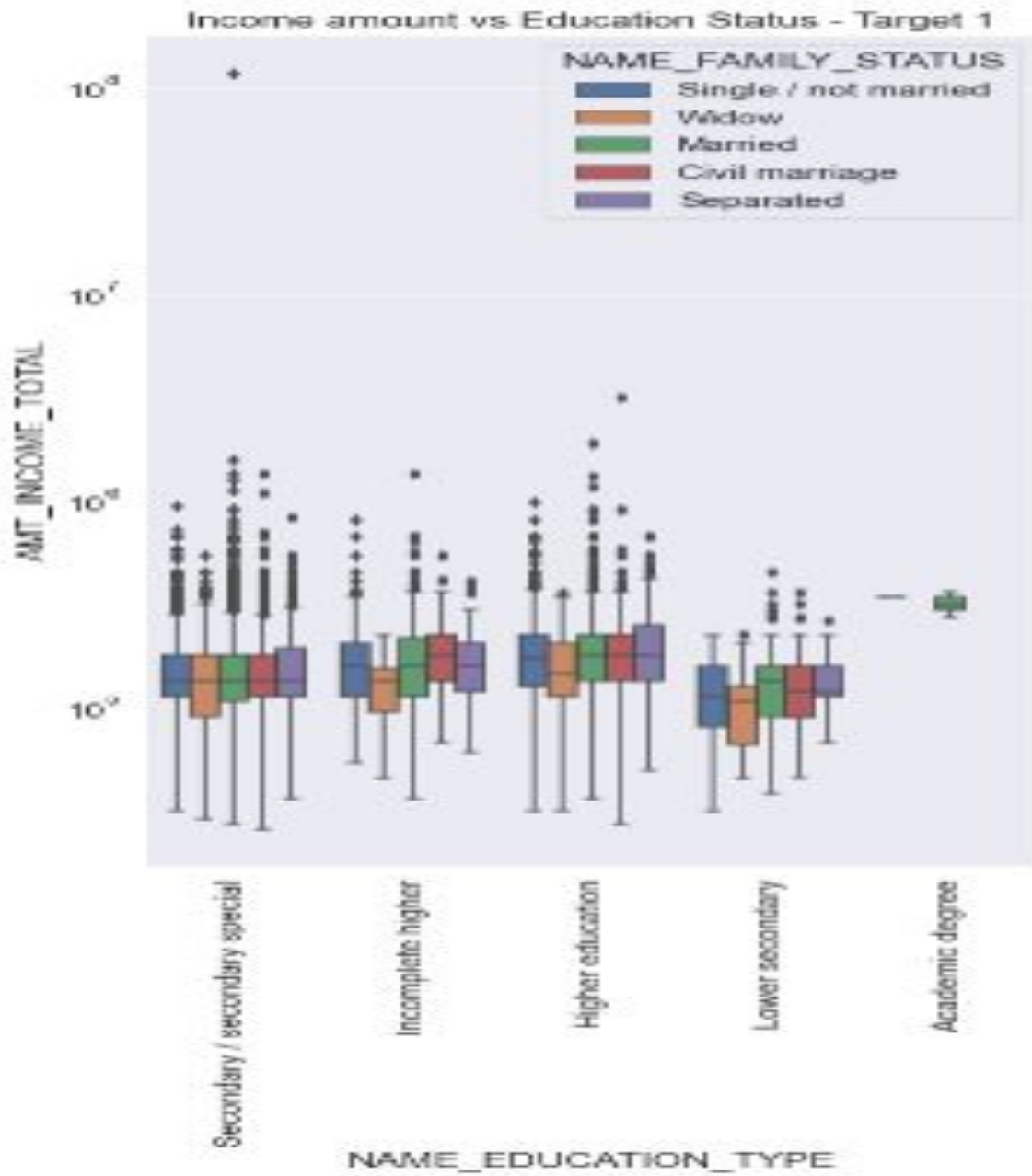
3. It is also observed that the customers with higher education have higher credits and are more likely to make payments on time.

4. Maximum number of outliers are seen in the category higher education.

5. Also , it is seen from the graph that, the people with secondary and secondary special education are less likely to make the loan payments on time.

Analysis of Income vs Education Status:





INFERENCES:

Based on the analysis done on **Income vs Education Status columns,** it can be inferred that:

1.The category, Higher education has the highest number of outliers followed by Secondary/secondary Special category.

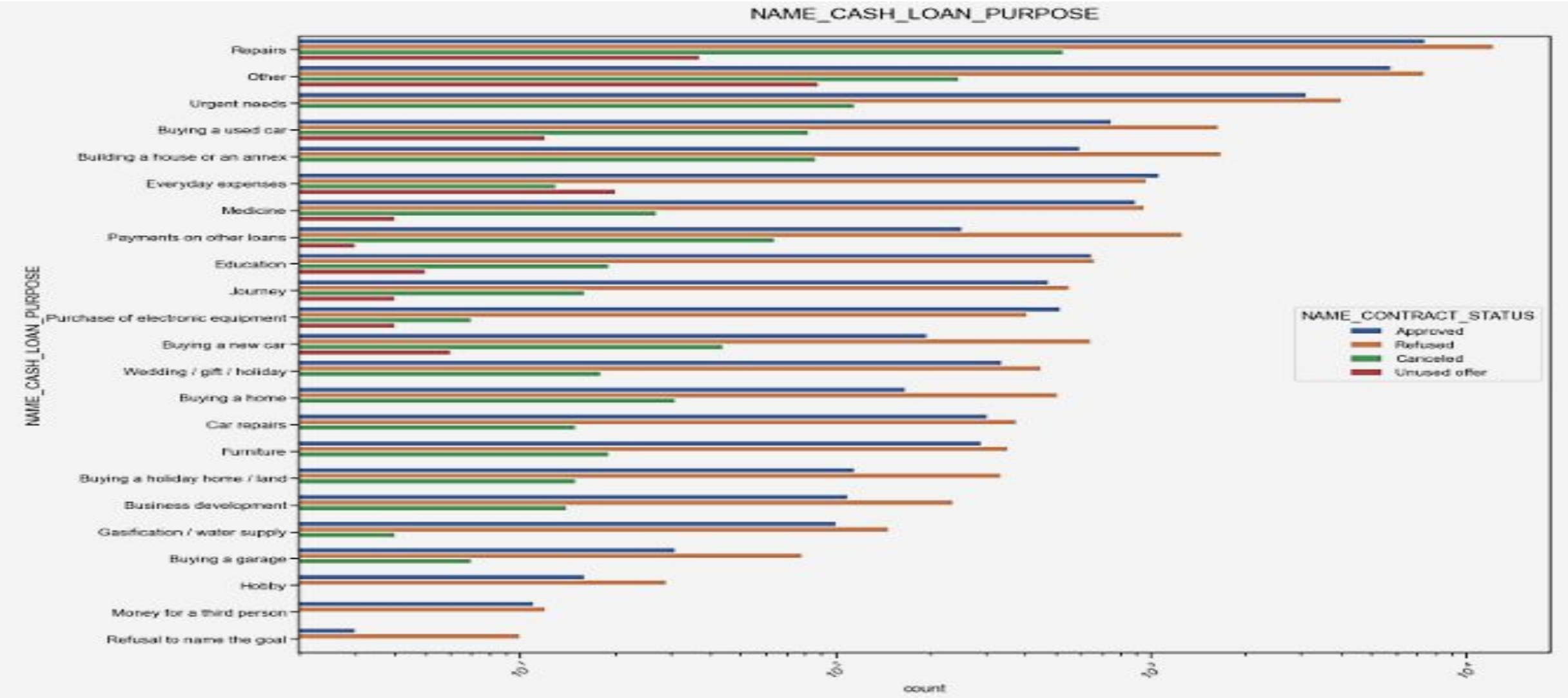
2. We can also see that the customers with higher education have higher income and dont have any difficulties in paying the loan amount.

3. However, the customers with higher education who have lesser income are unable to pay the loan and become defaulters.

Hence it can be concluded that, the people with higher education have higher chances to pay the loan on time and hence the bank can target this category.

Univariate Analysis on the Merged Dataframe

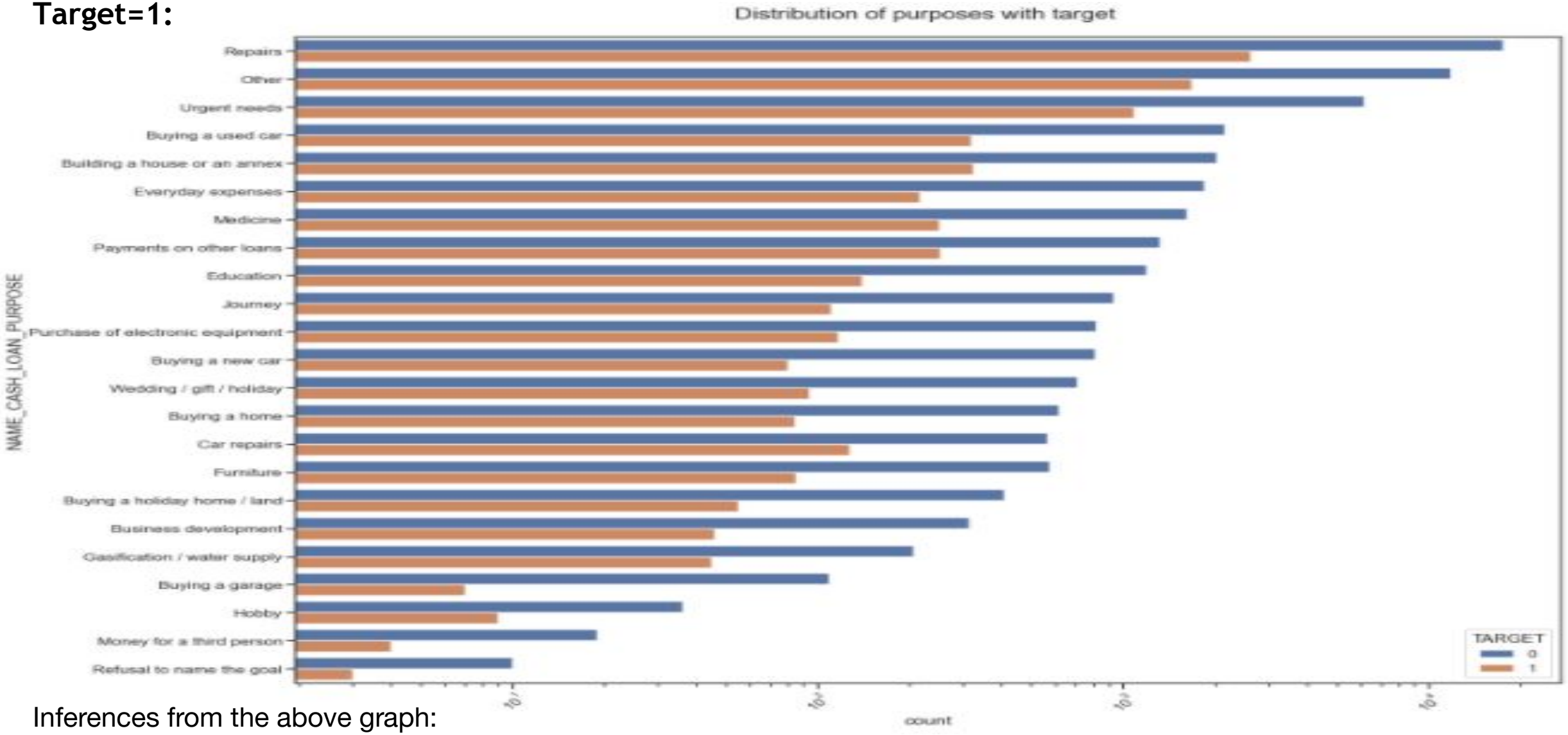
Distribution NAME_CONTRACT_STATUS:



Inferences from above graph:

Maximum of the loans are rejected which have load purpose as 'Repairs'. 'Everyday Expenses' has approval slightly greater than rejection, while, the cancelation is very low. 'Payments of other loans' has much higher rejection than approval. 'Education' purpose has almost equal rates of approvals and rejections.

Distribution of loan purpose with Target=0, Target=1:



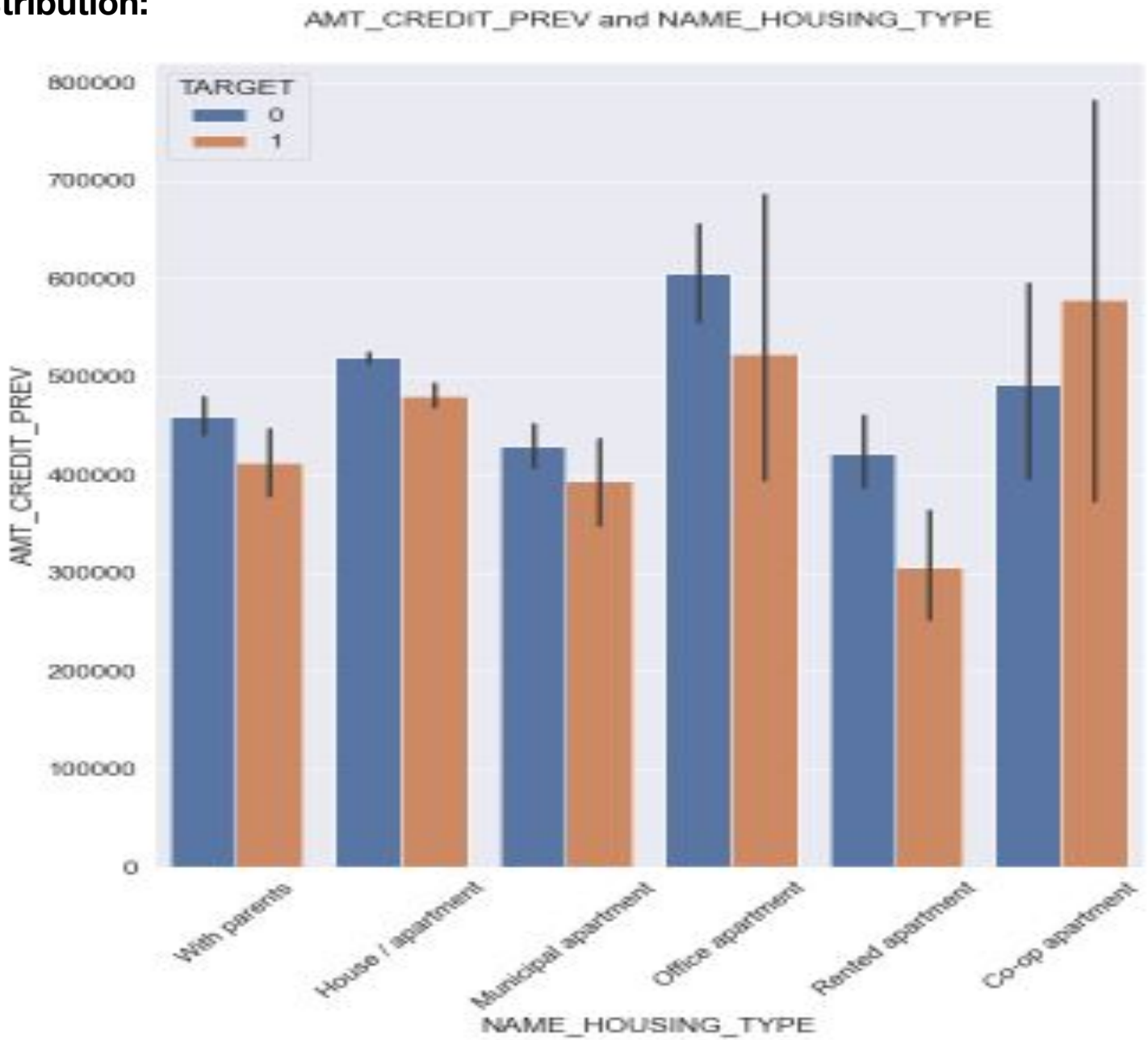
Inferences from the above graph:

1. 'Repairs' purpose is having the most difficulty in loan payment.
2. 'Education' purpose is having significantly high loan payment. Hence this category can be considered while giving loan.
3. 'Car Repairs' and 'Furniture' is having almost similar stats of loan payment.
4. 'Buying a new car' is having high loan payment compared to its chances of being a defaulter.

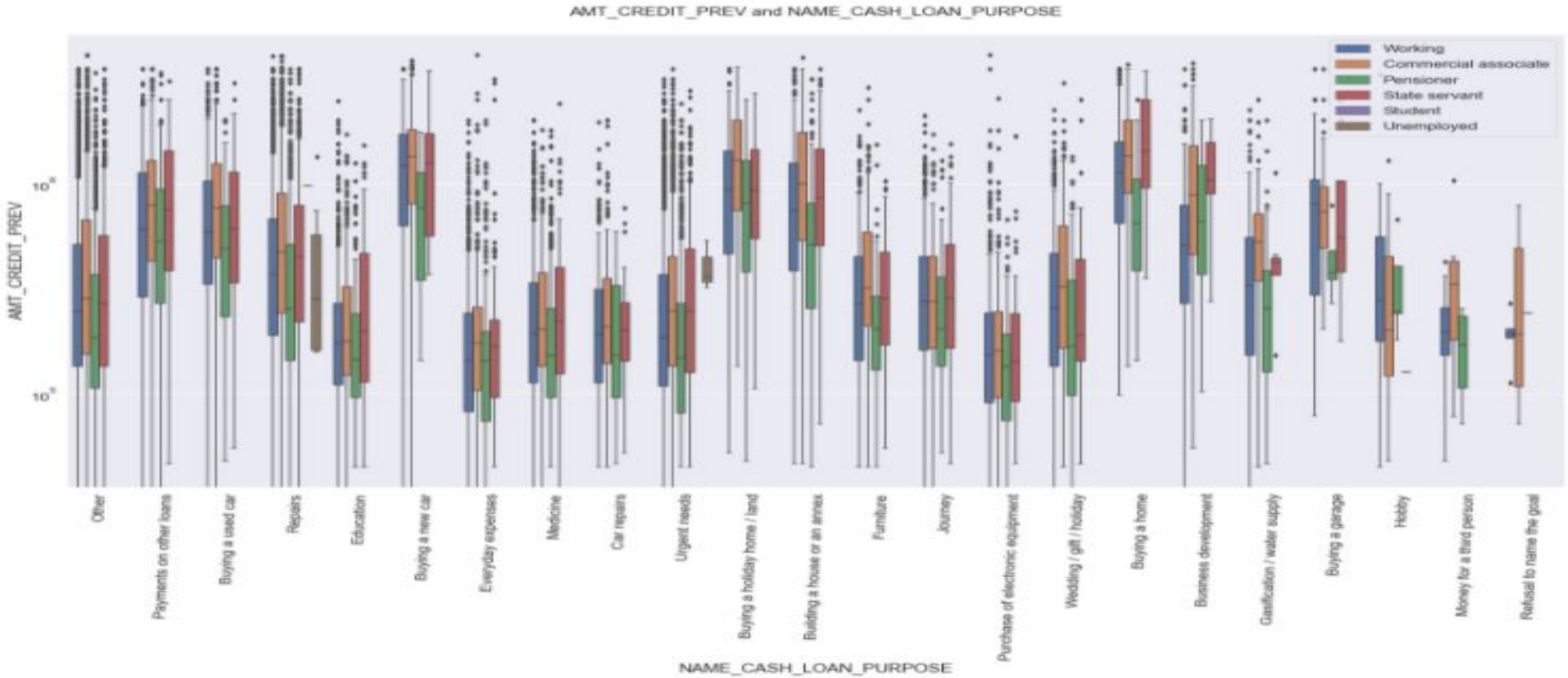
Bivariate analysis on the Merged Dataframe:

AMT_CREDIT_PREV and NAME_HOUSING_TYPE distribution:

In NAME_HOUSING_TYPE, Municipal Apartments has the highest credit for Target=0 and Office Apartments has highest credit for Target=1. While providing loans, bank can consider House/Apartments or Parents Housing_Type as they provide successful loan payment.



AMT_CREDIT_PREV and NAME_CASH_LOAN_PURPOSE:



The AMT_CREDIT_PREV is much higher in purposes like 'Buying a new car', 'Buying a holiday home/land', 'Building a house or annex', 'Buying a home', 'Business development'. Bank can provide loan to such customers. Even in these purposes Commercial Associate/State Servant has a higher credit as compared to others of same purpose.
Most lower credit is of a Pensioner having a purpose 'Purchase of electronic equipment'. So bank can avoid/be alert about such customers.

OVERALL CONCLUSIONS

1. While giving loans, Bank can consider Commercial Associate/State Servant/Student NAME_INCOME_TYPE as they mostly pay back the loan.

2. Bank can also consider Working NAME_INCOME_TYPE in some purposes like 'Buying a new car' ,'Buying a holiday home/land' as they have very small chance of being a defaulter.

3. Though 'Repair' Purpose have higher chance of loan repayment, but they also have highest chance of being a defaulter. So, bank should be more cautious in paying the loan for 'Repair' Purpose.

4. Bank can consider giving loan to Housing type - With Parents, as they are likely to give loan payment.

5. Bank should get as many customers for loan with purpose 'Buying a home' and income type State Servant, as they are very likely to pay the loan.