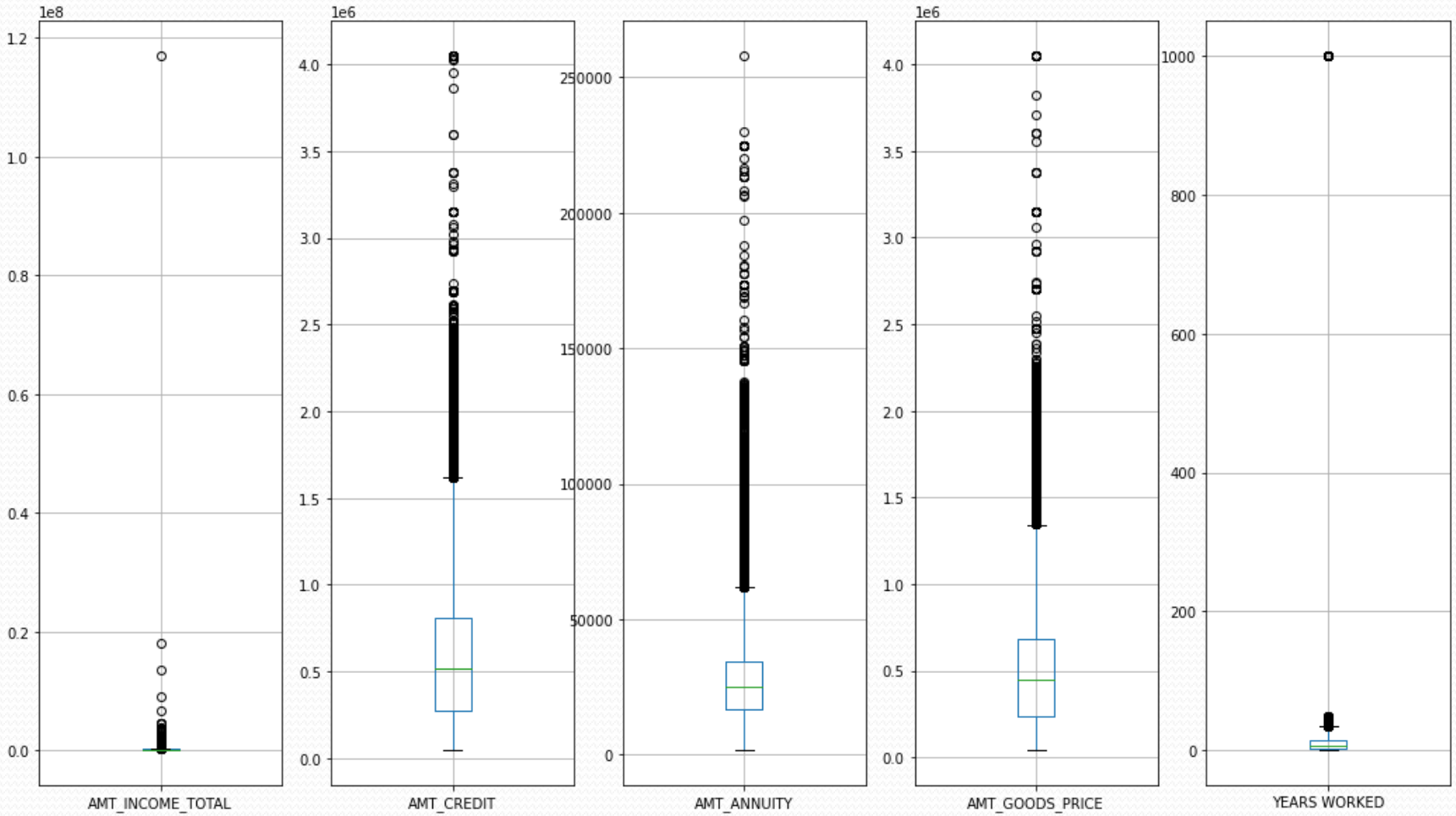
Objective:

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.

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

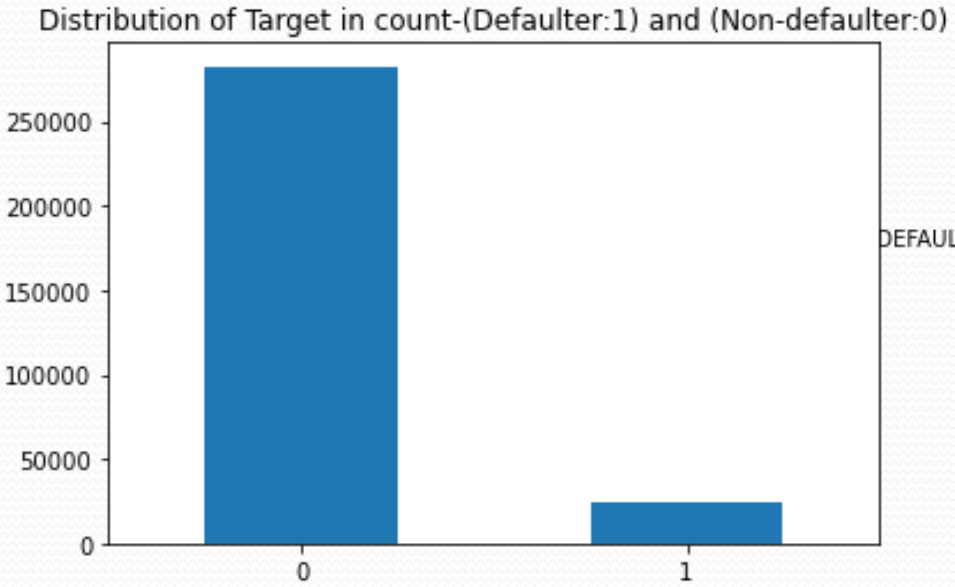
Outliers in Application data

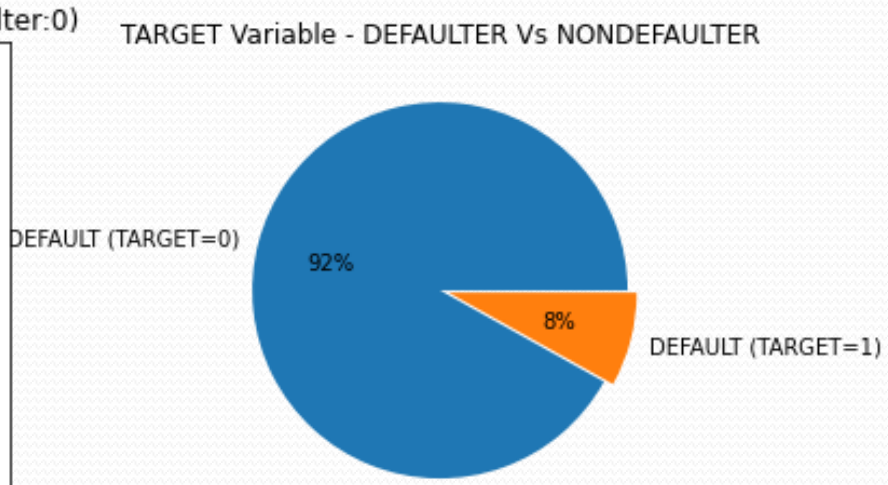


We found outliers in 5 numerical variables AMT_INCOME_TOTAL, AMT_CREDIT,AMT_ANNUITY, AMT_GOODS_PRICE AND YEARS WORKED

DATA IMBALANCE

From the below chart, we conclude that there is an imbalance in the provided dataset as 92% of data is with target-0 i.e. client with no payment difficulty while the remaining 8% are clients with difficulty in payment.





The ratio of non-default to default is 11.39.

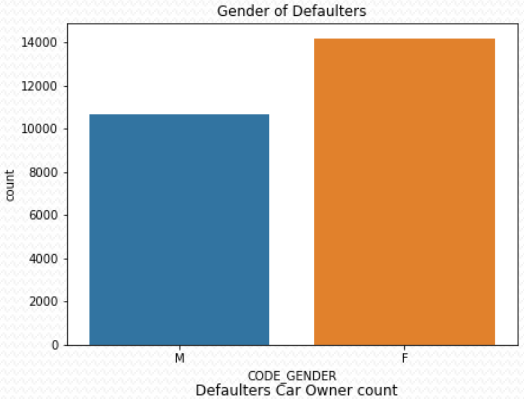
Univariate Analysis

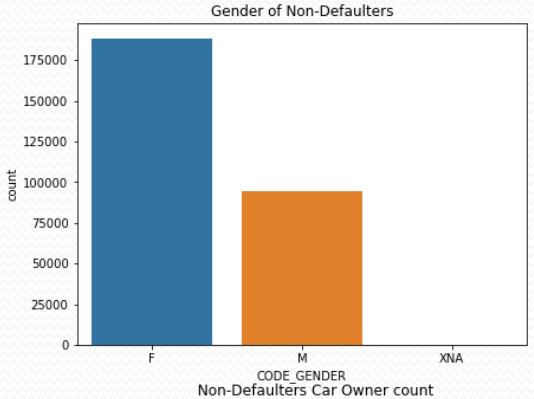
Categorical variables

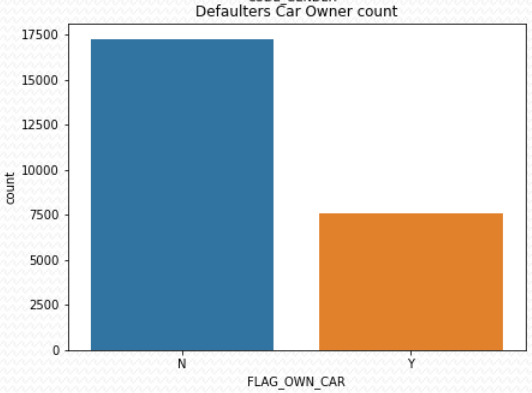
- Number of female clients are more in both the cases default/non-default. And the chances of default is lower among female applicants than that of the male.
- Clients that own a car are less likely to not repay the loan when compared to the ones that do not own a car.
- Client who do own real estate is higher than who has not. Applicants with no realty ownership has a higher propensity to default than the clients who own real estate.

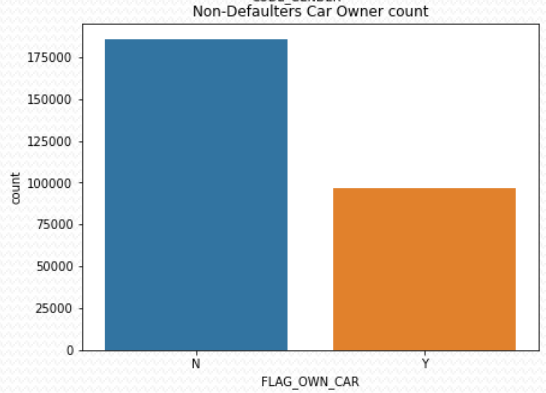
.

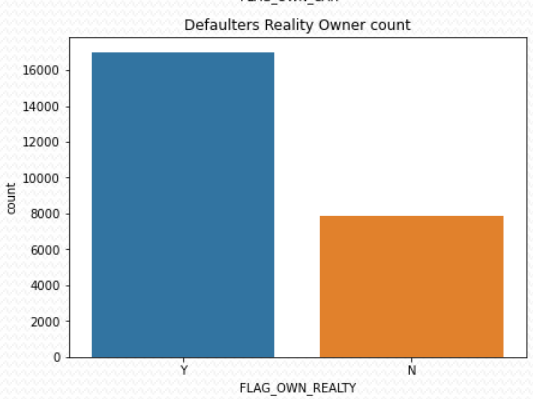
- Applicants who are married are among the highest number of defaulters and also non-defaulters. Whereas, widows are the lowest number of defaulters and non-defaulters. Which is interesting to see because you expect widows to not payback their loans but it is the opposite here.
- Applicants who have applied for credits are from most of the organization type Business entity Type 3 , XNA(unknown), Self employed, Medicine and Other for both defaulters and non defaulters

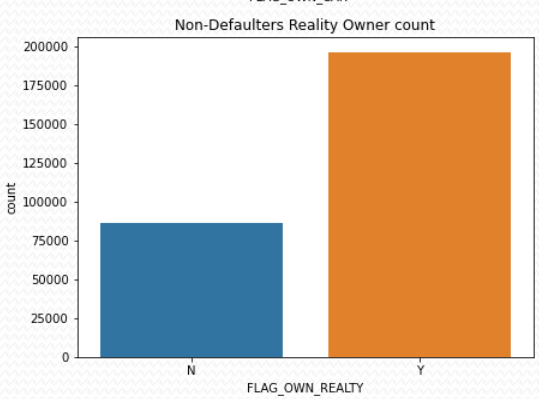






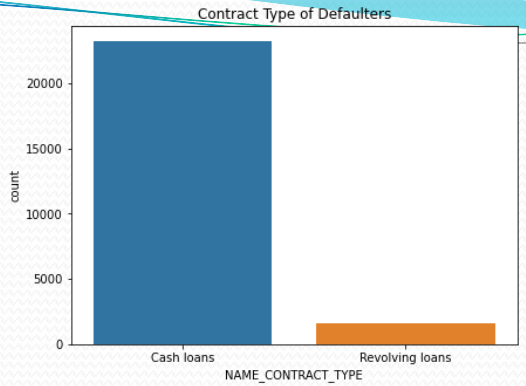


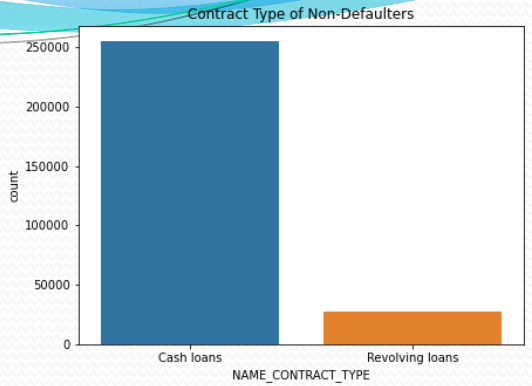


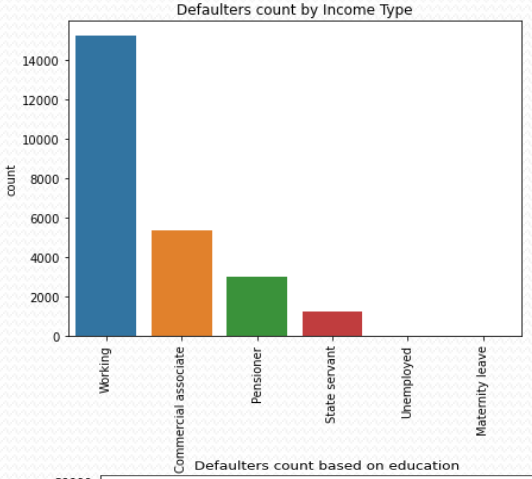


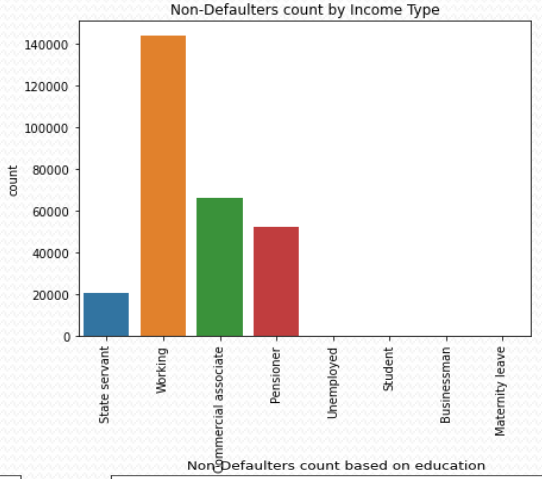
Categorical variables

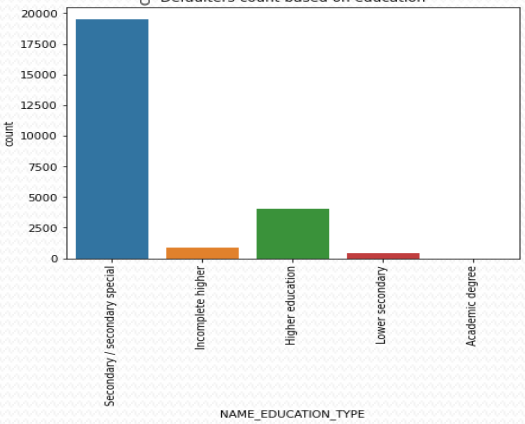
- Cash loans applicants are higher in both default/ non-default cases than revolving loans.
- Majority of the applicants are from working, commercial associate, pensioners and state servants. The remaining categories of income types are very small
- Applicants with secondary and higher secondary education are among the highest defaulters as well as not defaulters. Whereas, applicants with academic degrees are the smallest group of applicants that have applied for the loan and applicants from this background has no records of default. From the above figure, we see that a distinct pattern emerging. The chances of default is lower as the education level of the applicants increases.

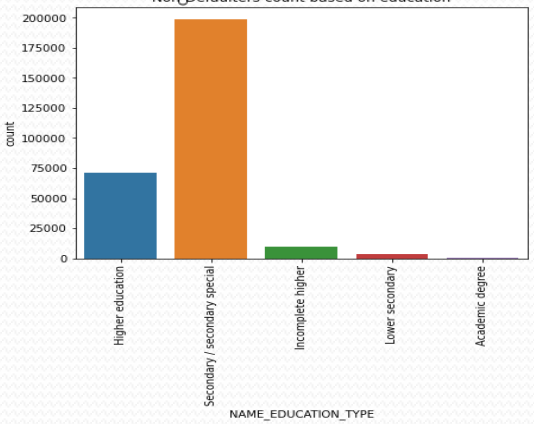








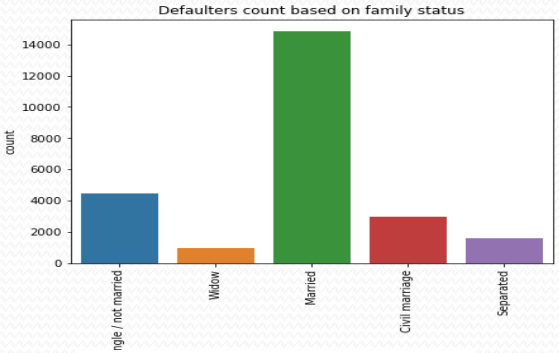


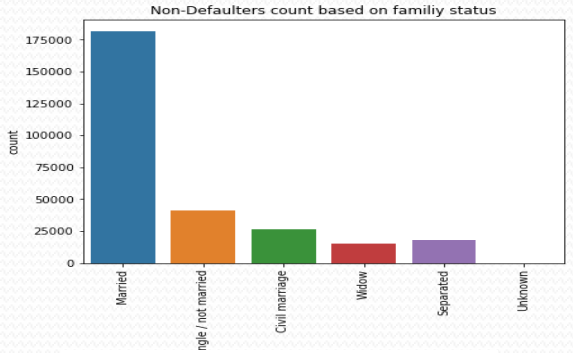


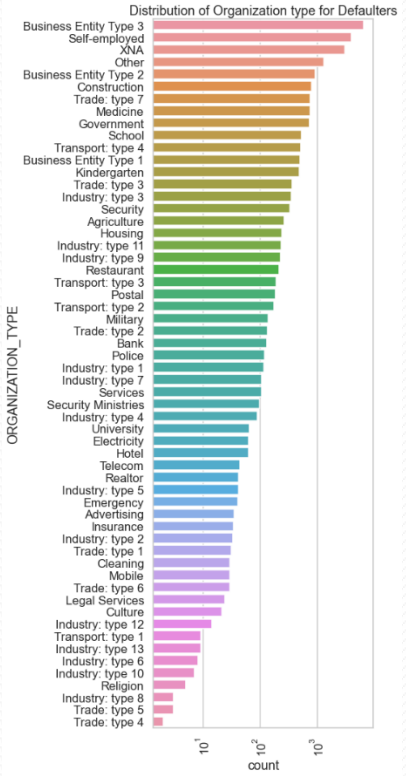
Categorical variables

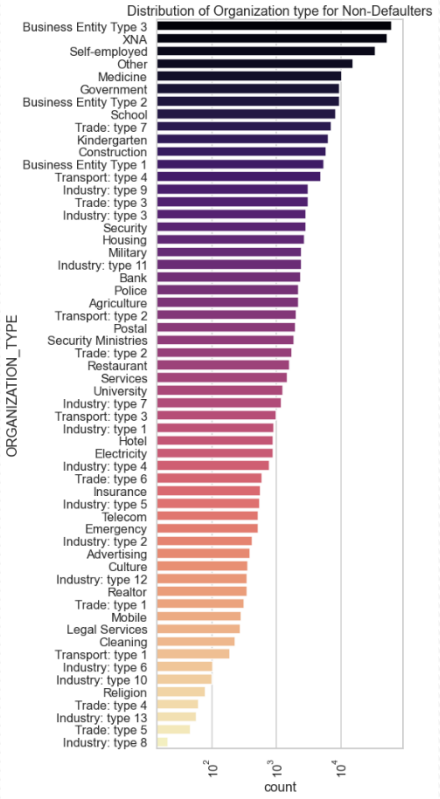
- Applicants who are married are among the highest number of defaulters and also non-defaulters. Whereas, widows are the lowest number of defaulters and non-defaulters. Which is interesting to see because you expect widows to not payback their loans but it is the opposite here.

- Applicants who have applied for credits are from most of the organization type Business entity Type 3 , XNA(unknown), Self employed, Medicine and Other for both defaulters and non defaulters



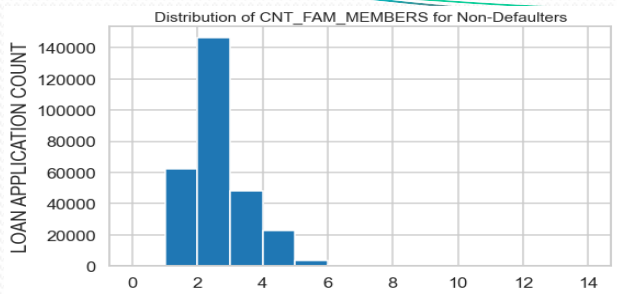


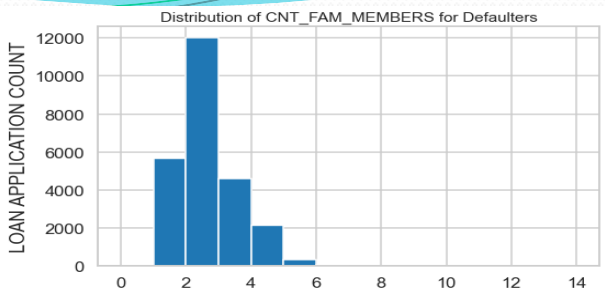


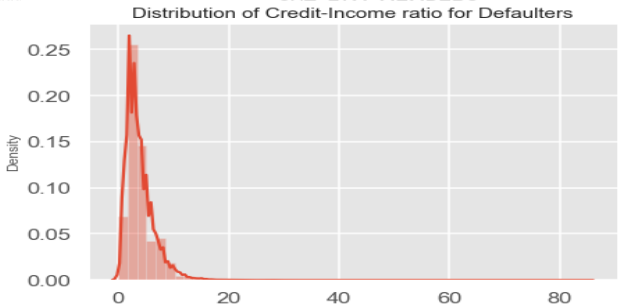


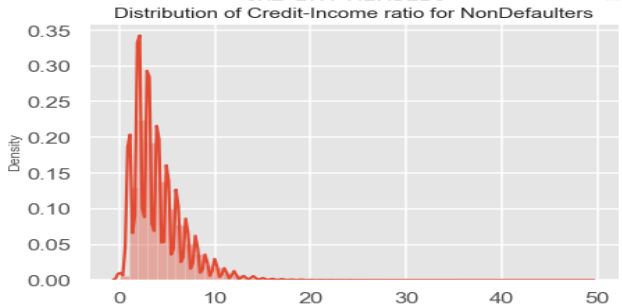
Continuous variable

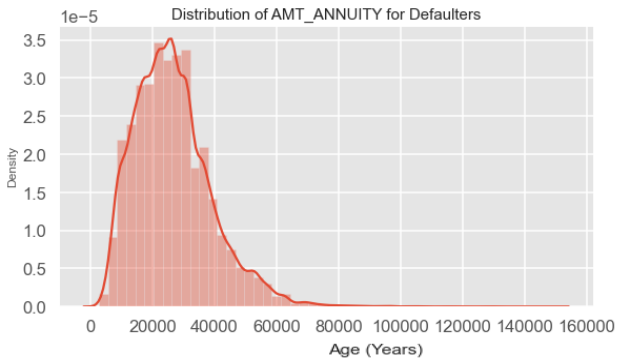
- A family of 3 applies loan more often than the other families
- When the CREDIT_INCOME_RATIO is more than 50, people default
- The annuity amount in defaulters is between 20000 and 35000 whereas for non defaulters it is between 30000 and 50000.The loan annuity is mostly concentrated within 10000 to 60000 range in both the cases
- Around 29 years to 40 years people are more defaulters. There is high chance to be defaulted of the young people.
- Lesser loan credit amount, the higher the default chances

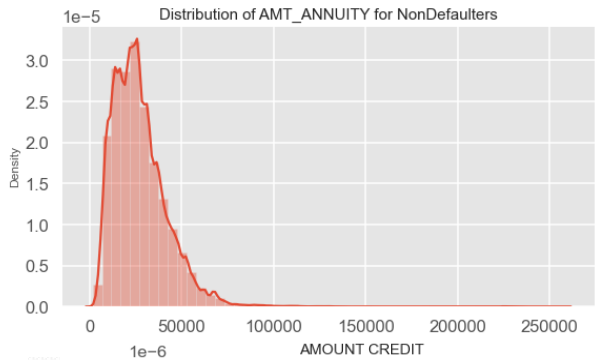


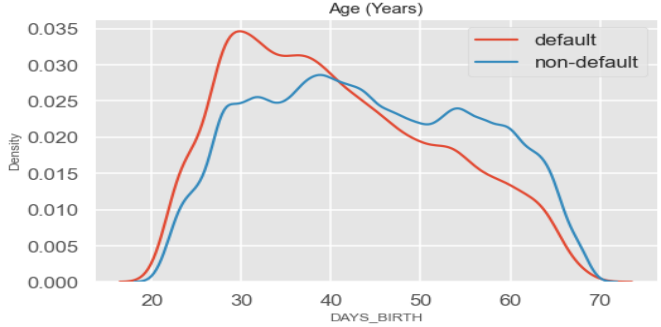


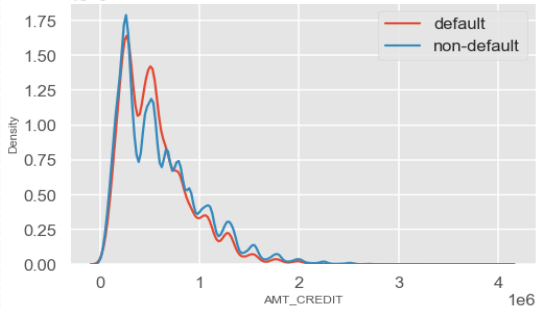








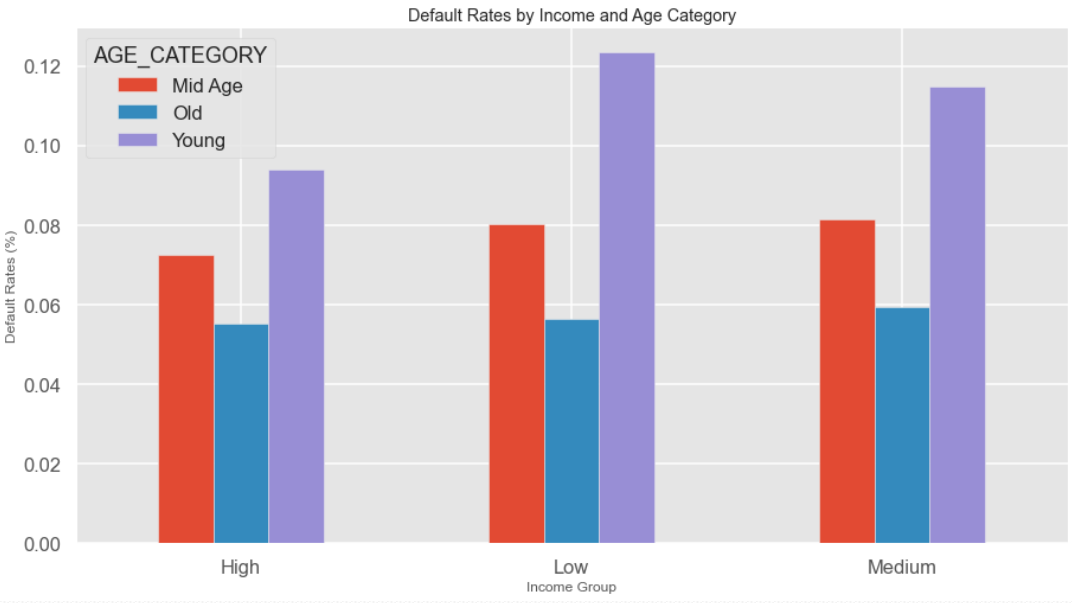


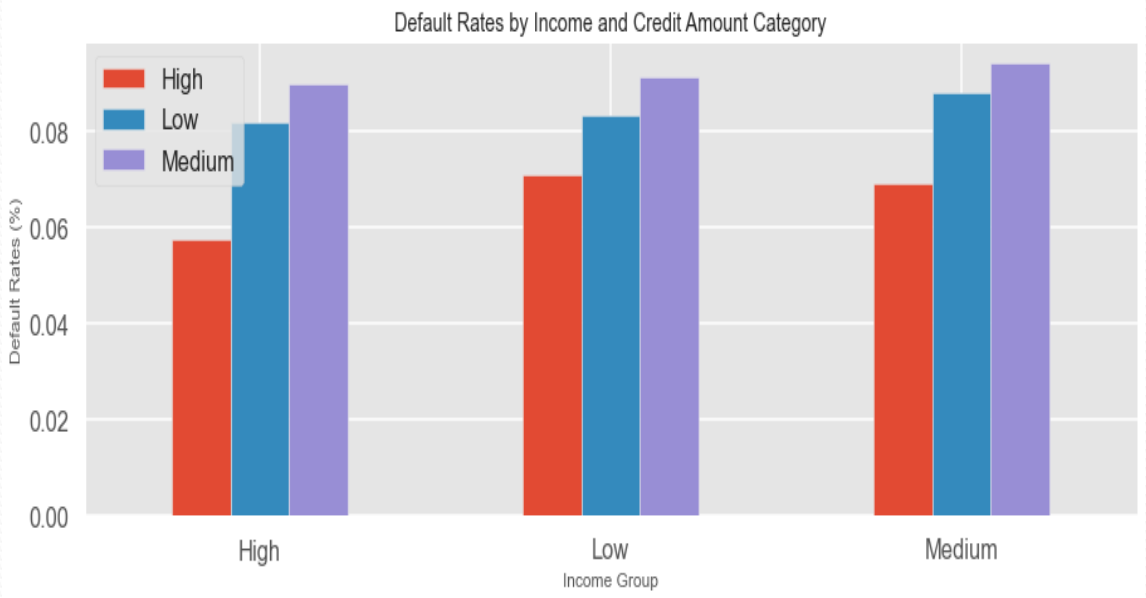


Segmented Univariate Analysis

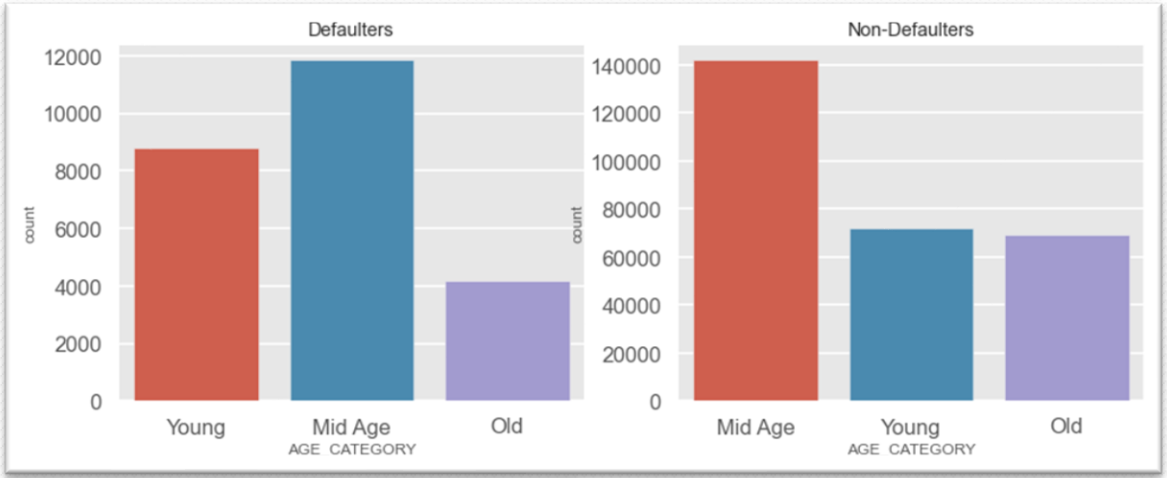
- Irrespective of the income groups, the chances of default decreases as the age of the applicants increases.

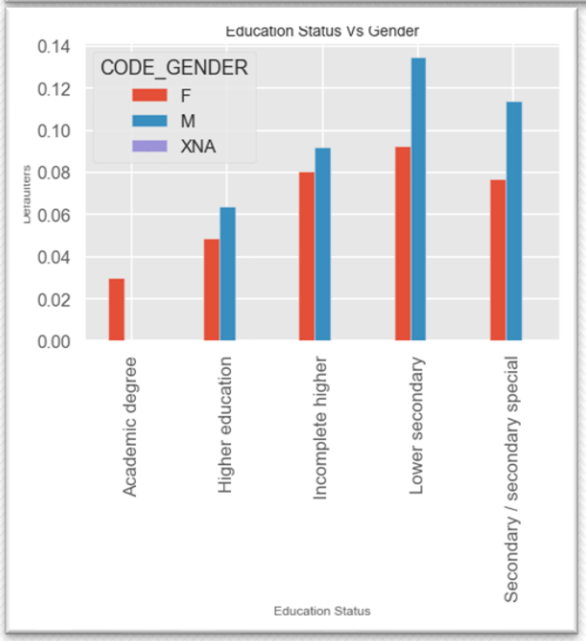
- Irrespective of the income group, the chances of default increases as the credit amount increases. Also if we compare credit amount categories by different income groups, then the default rates for all the three credit amount categories are lower in the high income group relative to the medium and low income groups

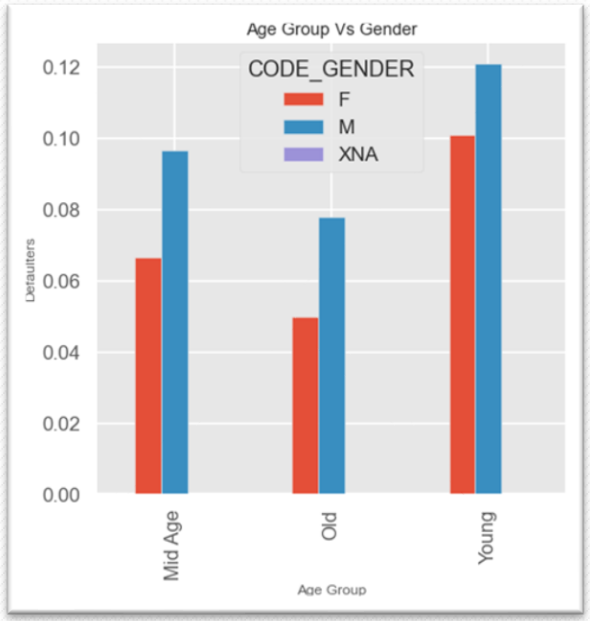




- Mid age (35-55) age group of people are more likely to be defaulted followed by the young people.
- Male with lower secondary education are more defaulted followed by Secondary /secondary special education.
- Young male clients are more in number to be defaulted.

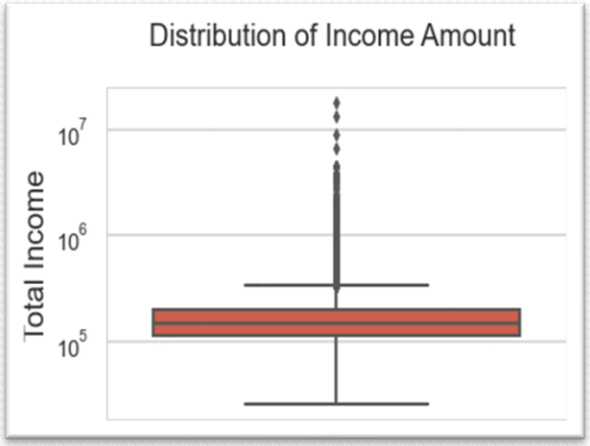


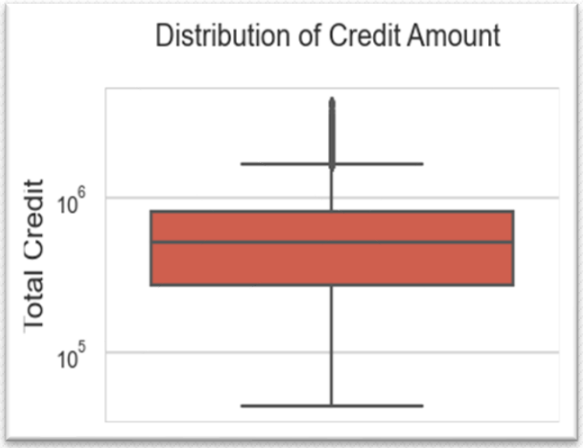


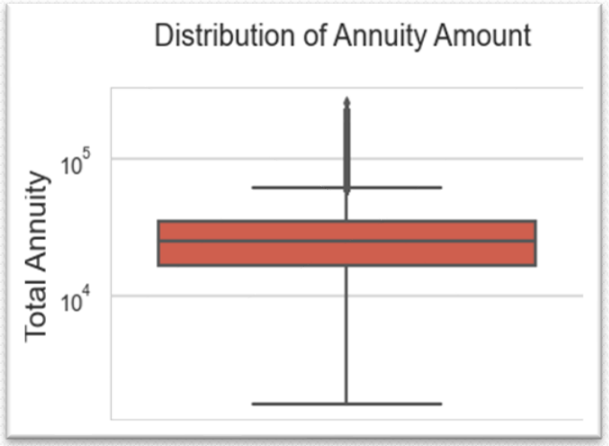


❖Finding Outliers in Defaulters and Non defaulters

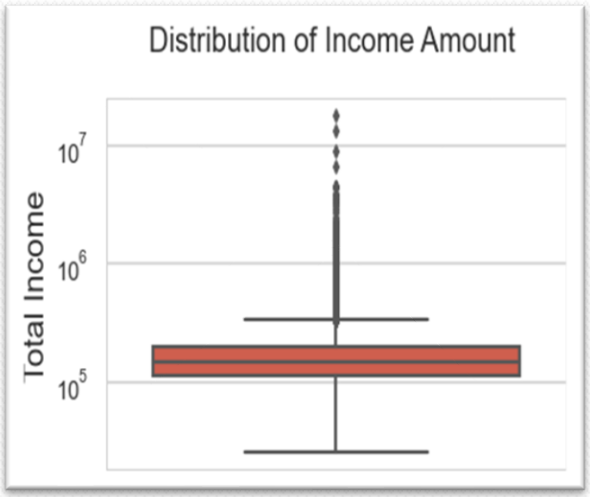
Univariate Analysis

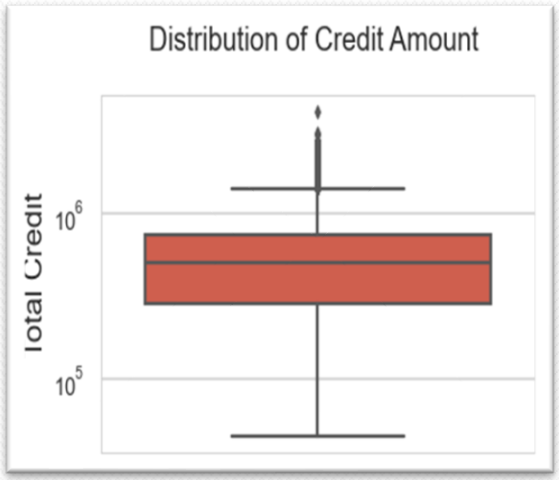


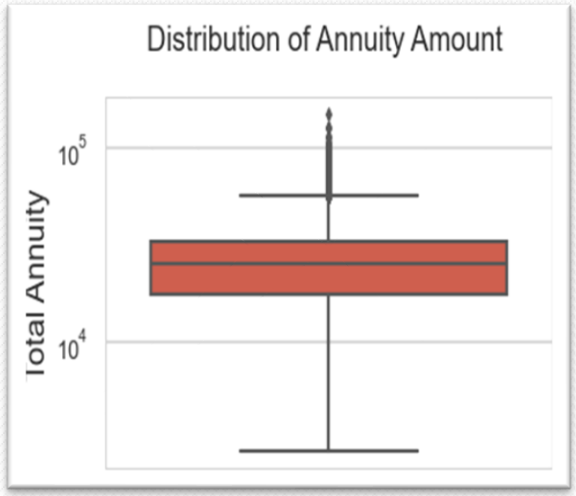




Non Defaulters (target = 0)



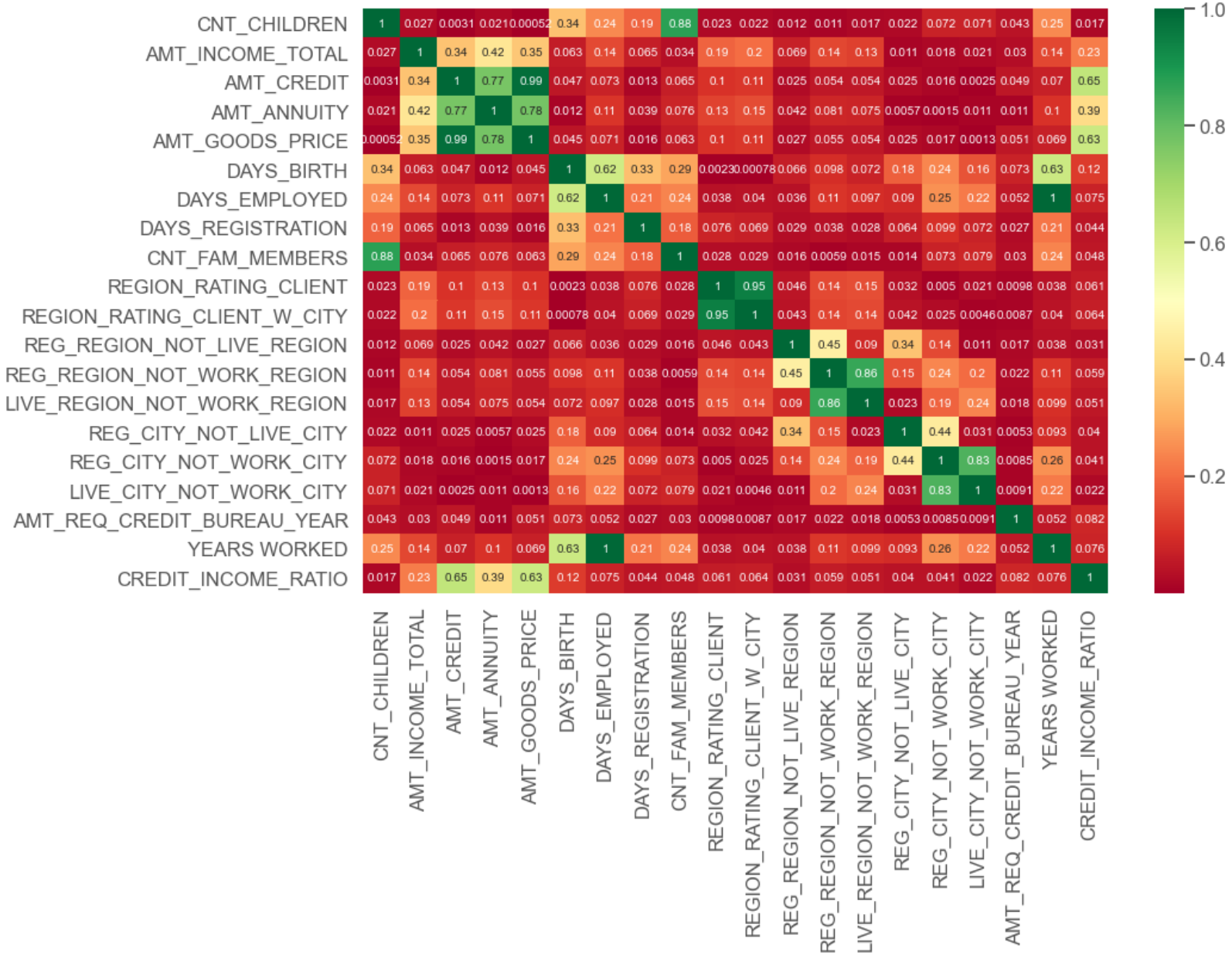




For Defaulters (target=1)

Bivariate analysis

Correlation for Non-Defaulters



Findings: Correlation for Non defaulters

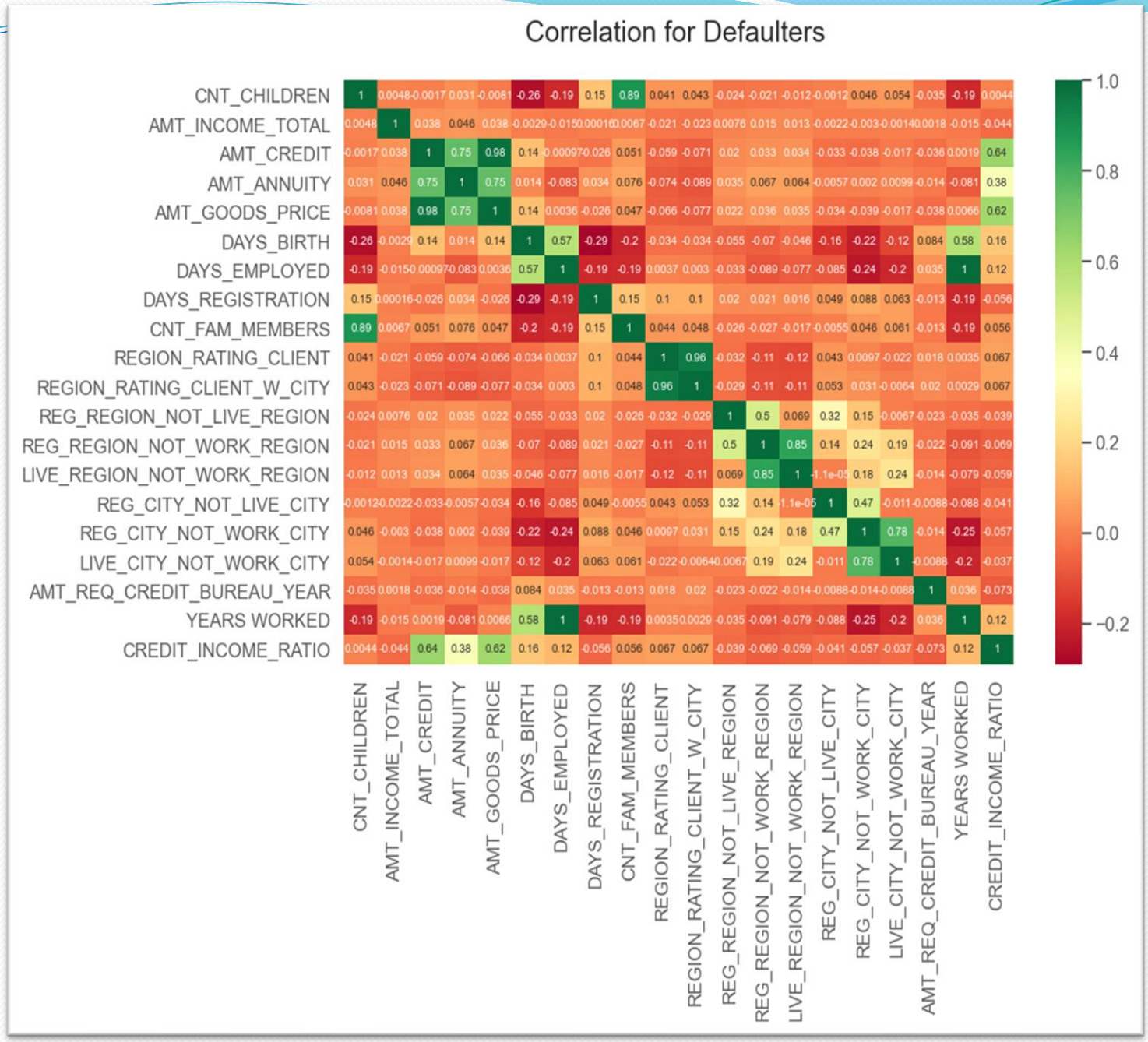
- **Credit amount is inversely proportional to the date of birth, which means Credit amount is higher for low age and vice-versa.**

- **Credit amount is inversely proportional to the number of children client have, means Credit amount is higher for less children count client have and vice-versa.**

- **Income amount is inversely proportional to the number of children client have, means more income for less children client have and vice-versa.**
- **Less children client have a densely populated area.**

- **Credit amount is higher to densely populated area.**

- **The income is also higher in densely populated area**


Correlation for Defaulters
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
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
AMT_REQ_CREDIT_BUREAU_YEAR
YEARS WORKED
CREDIT_INCOME_RATIO
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Findings: Correlation for Defaulters

This heat map for Defaulters is also having quite a same observation just like non defaulters. But for few points are different. They are listed below.

- The client's permanent address does not match contact address are having less children and vice-versa

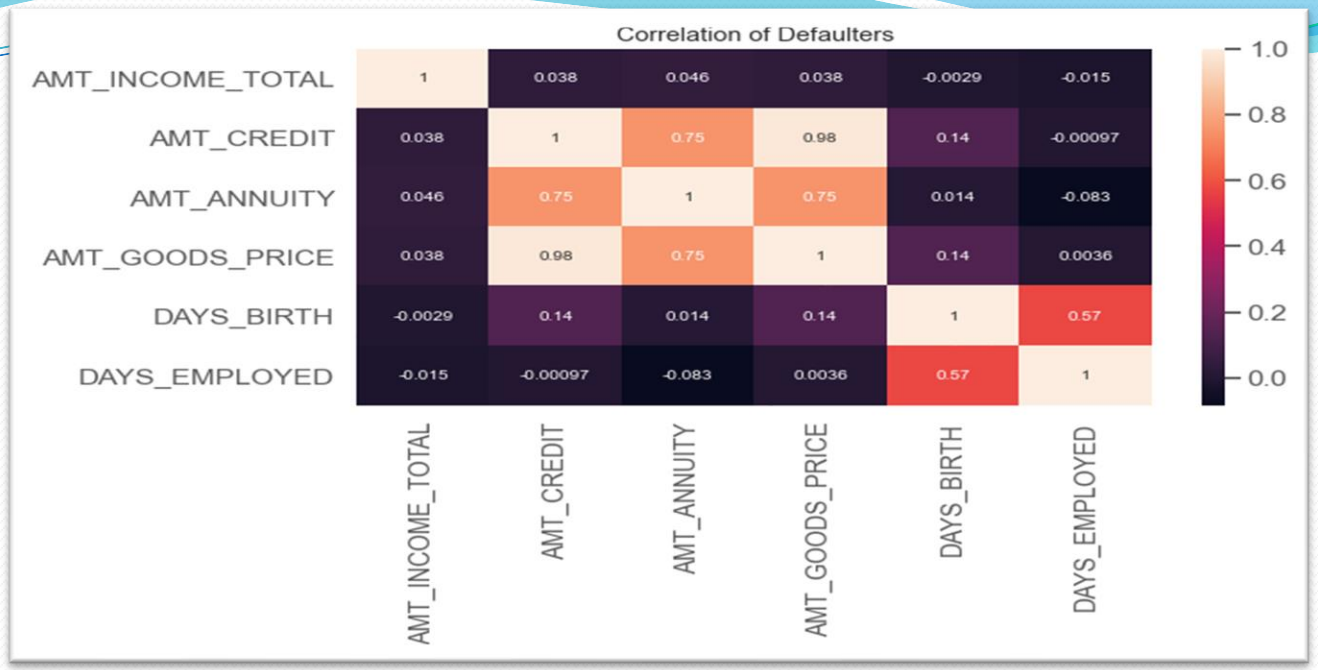
- The client's permanent address does not match work address are having less children and vice-versa

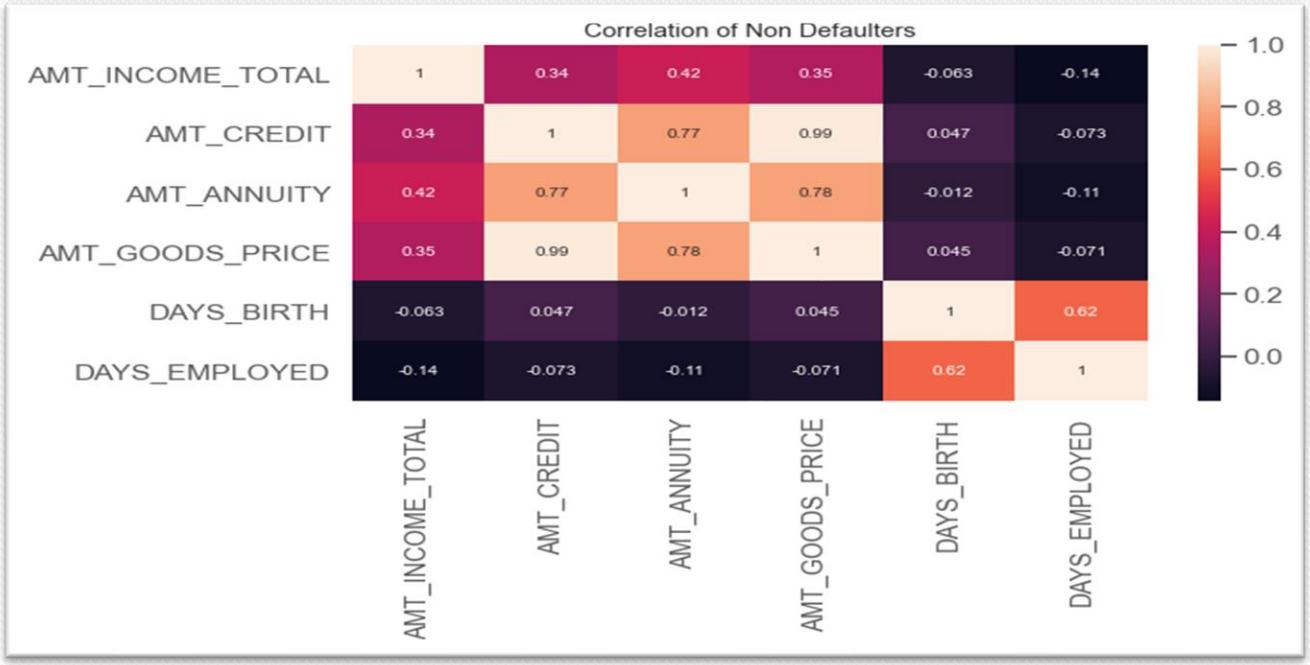
Top 10 Correlation

Below are the top 10 variables, where we can see high correlation in Defaulters and non defaulters.
We can see that GOODS_PRICE and AMT_CREDIT, AMT_ANNUTY and AMT_AMT_CREDIT are highly correlated

Defaulters		
Variable 1	Variable 2	Correlation
YEARS WORKED	DAYS_EMPLOYED	1
AMT_GOODS_PRICE	AMT_CREDIT	0.98
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.96
CNT_FAM_MEMBERS	CNT_CHILDREN	0.89
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.85
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.78
AMT_GOODS_PRICE	AMT_ANNUITY	0.75
AMT_ANNUITY	AMT_CREDIT	0.75
CREDIT_INCOME_RATIO	AMT_CREDIT	0.64
CREDIT_INCOME_RATIO	AMT_GOODS_PRICE	0.62

Non Defaulters		
Variable 1	Variable 2	Correlation
YEARS WORKED	DAYS_EMPLOYED	1
AMT_GOODS_PRICE	AMT_CREDIT	0.99
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.95
CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.86
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.83
AMT_GOODS_PRICE	AMT_ANNUITY	0.78
AMT_ANNUITY	AMT_CREDIT	0.77
CREDIT_INCOME_RATIO	AMT_CREDIT	0.65
CREDIT_INCOME_RATIO	AMT_GOODS_PRICE	0.63

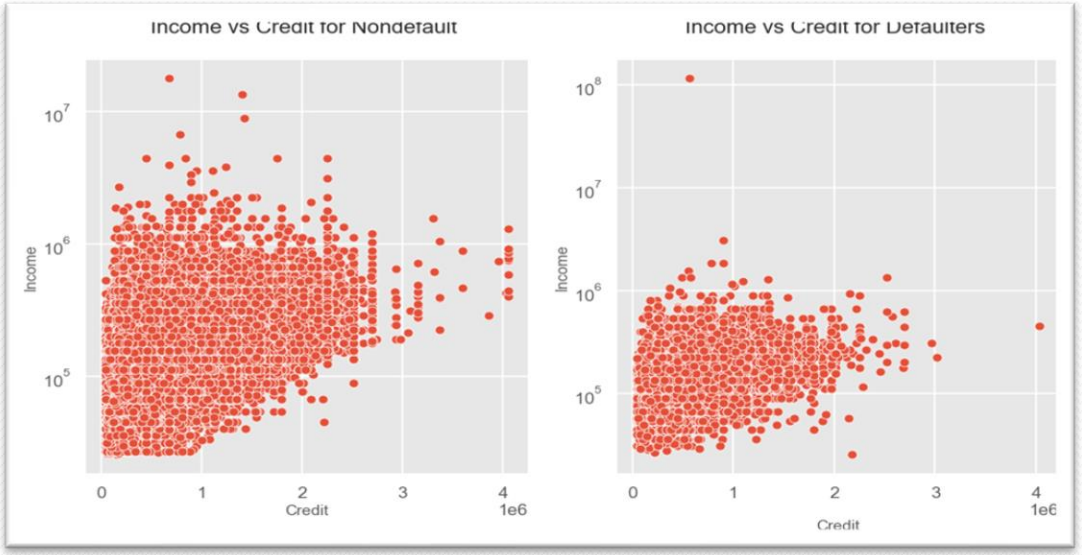

Correlation of Defaulters
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
DAYS_BIRTH
DAYS_EMPLOYED
1 0.038 0.046 0.038 -0.0029 -0.015
0.038 1 0.75 0.98 0.14 -0.00097
0.046 0.75 1 0.75 0.014 -0.083
0.038 0.98 0.75 1 0.14 0.0036
-0.0029 0.14 0.014 0.14 1 0.57
-0.015 -0.00097 -0.083 0.0036 0.57 1
1.0
0.8
0.6
0.4
0.2
0.0

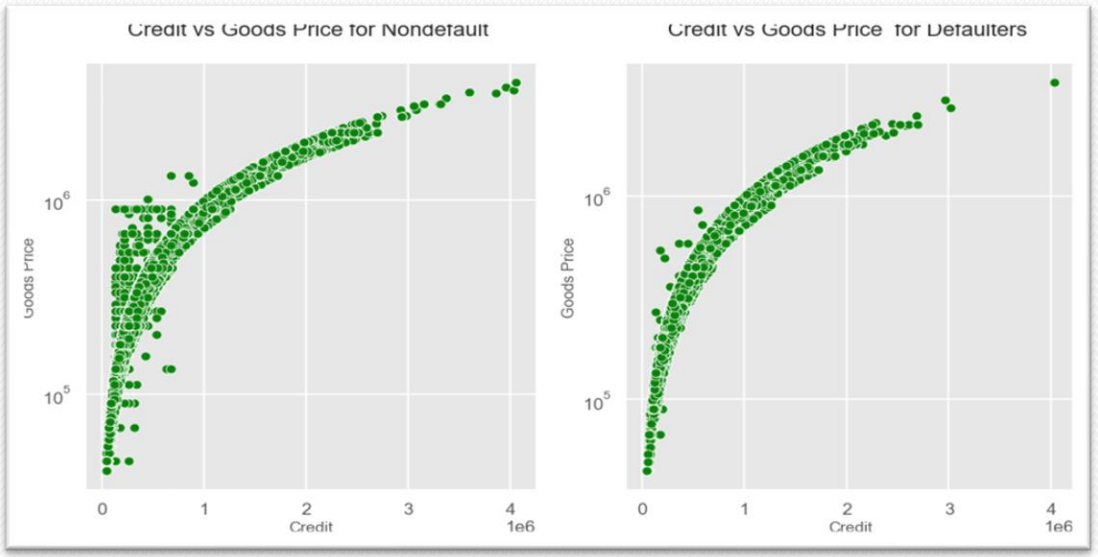

Correlation of Non Defaulters
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
DAYS_BIRTH
DAYS_EMPLOYED
1 0.34 0.42 0.35 -0.063 -0.14
0.34 1 0.77 0.99 0.047 -0.073
0.42 0.77 1 0.78 -0.012 -0.11
0.35 0.99 0.78 1 0.045 -0.071
-0.063 0.047 -0.012 0.045 1 0.62
-0.14 -0.073 -0.11 -0.071 0.62 1
1.0
0.8
0.6
0.4
0.2
0.0

Correlation on Continuous variables

- There is a positive correlation between Income and Credit since the plotted points are distributed from lower left corner to upper right corner

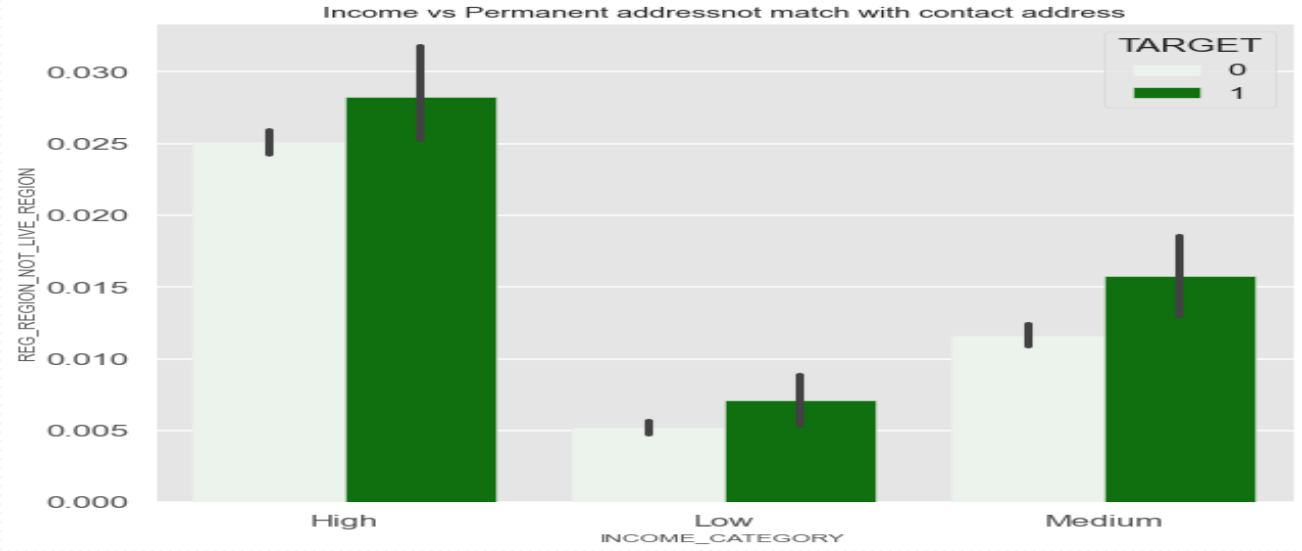
- Since the plots are scattered from left corner to upper right corner , we can conclude that there is a positive correlation between the two variables which means if there is increase in goods price, the credit also increases directly and vice versa.



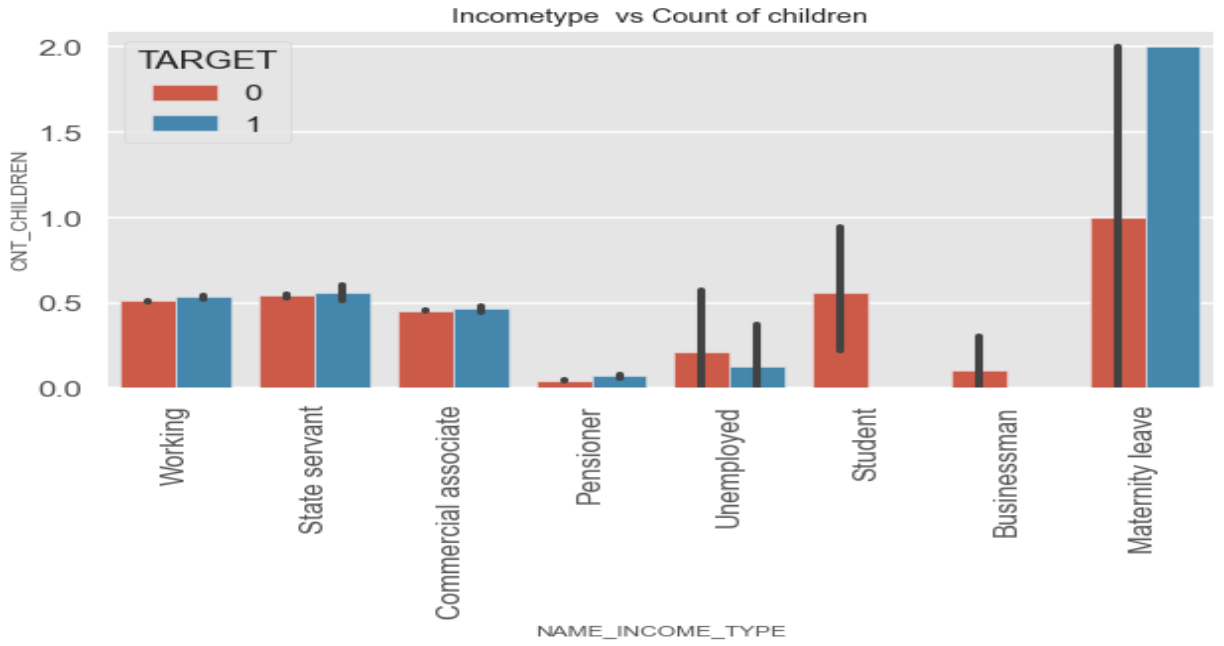


Findings in Bivariate analysis on Categorical variables

- When Client gets Extremely lower salary and if his/her address doest match, then there is a Higher chance for him/her to be defaulter



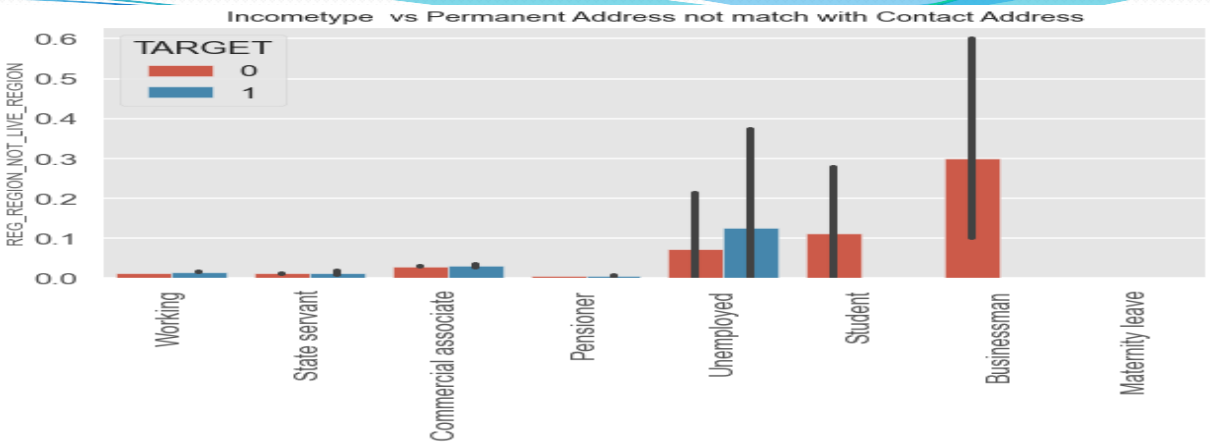
- It is interesting to note that applicants who are going on maternity leave can be defaulters and also non defaulters. Probably because the income is via Maternity Leave and tends to be more Defaulter when they have more children

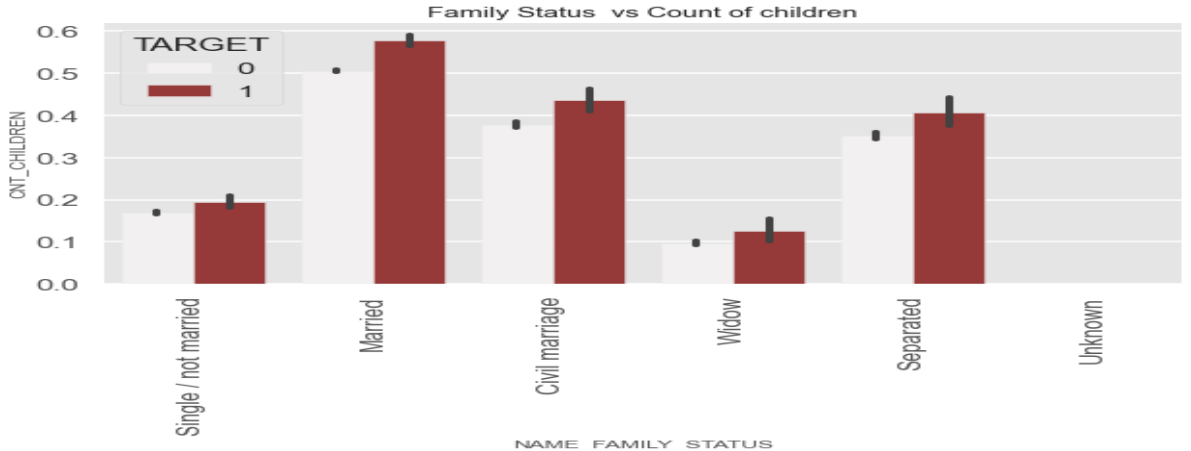


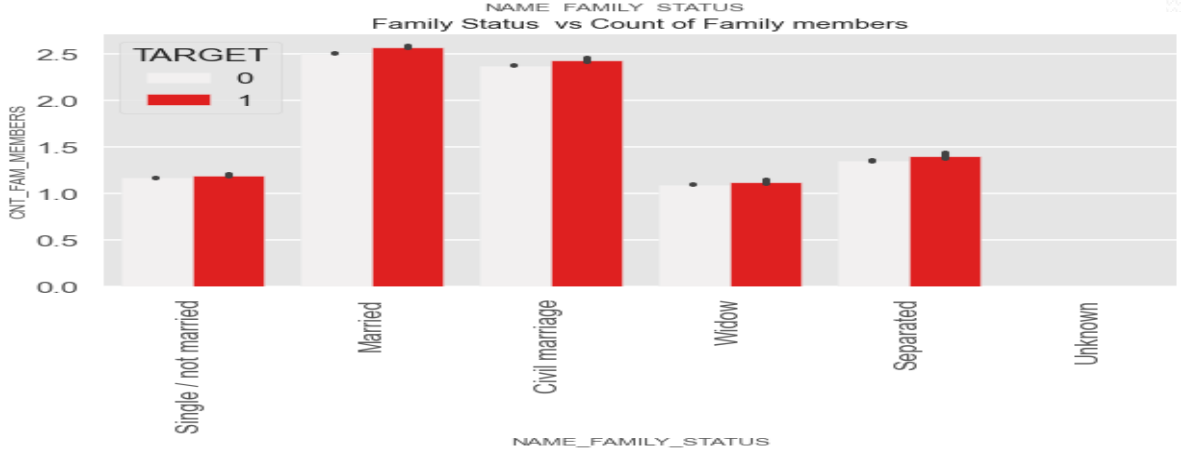
Categorical variables

- Applicants who are Unemployed has more chance to be a defaulter , when their Permanent Address does not match with the Contact Address in the Regional Level

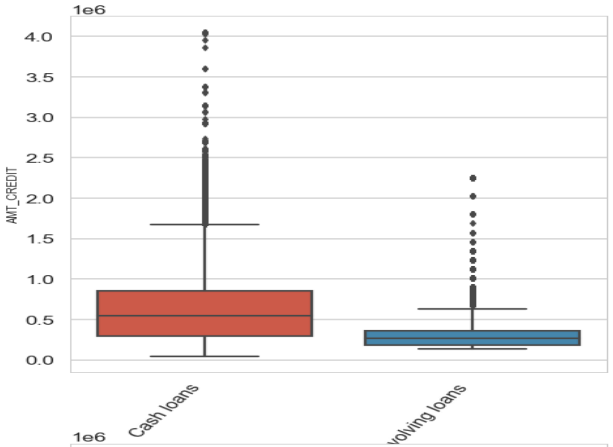
- Married applicants with children (5+), chances to be a defaulter in High. This may be due to the Economic situation of their family, because of more children
- Applicants who are married with family of 3 or more tend to be defaulters. This may be due to the Economic situation of their family, because of more members

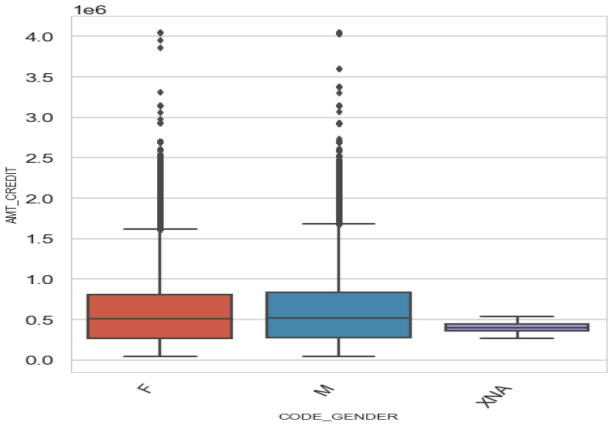


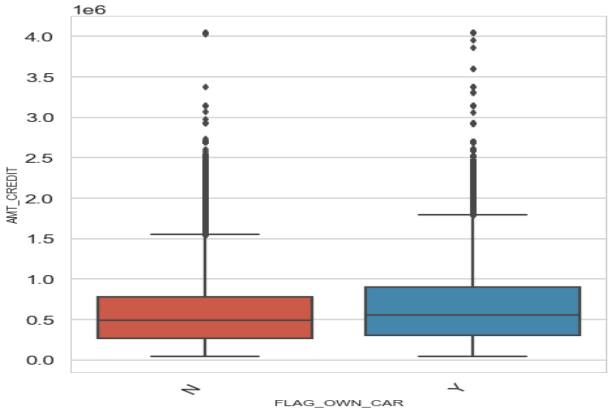


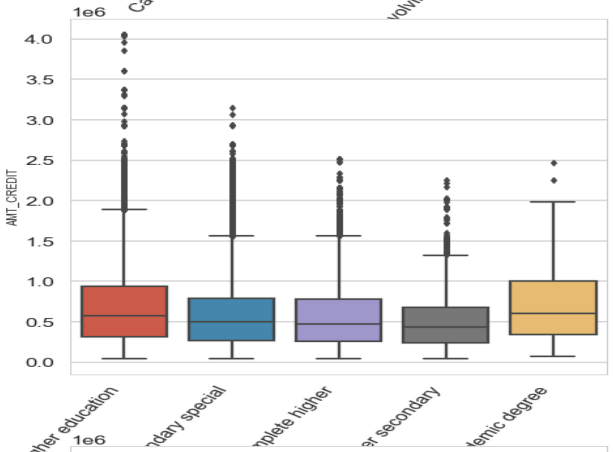


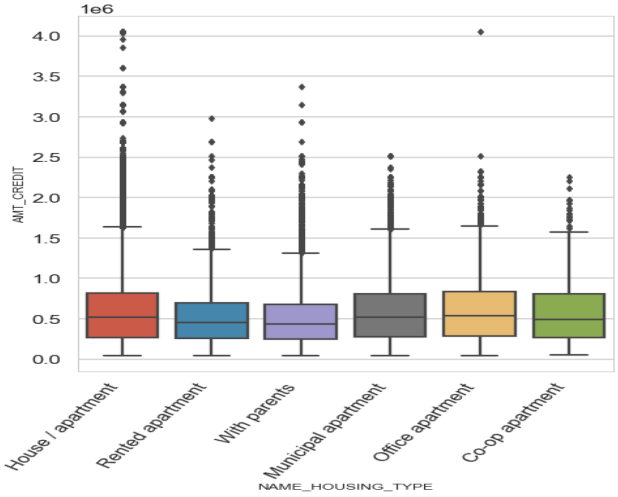
Plotting boxplot for Defaulters

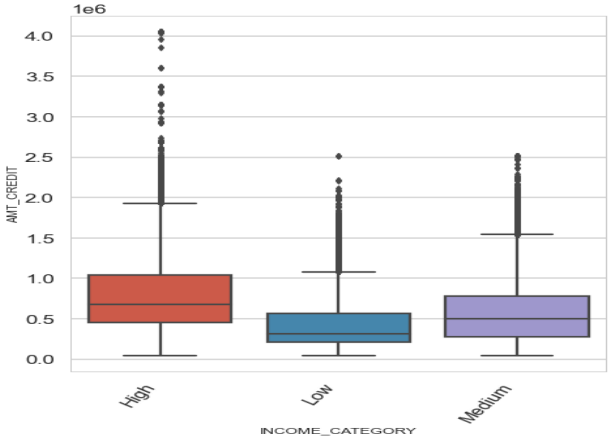


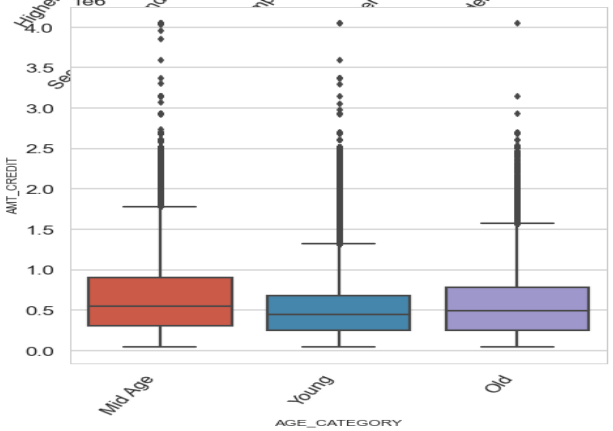












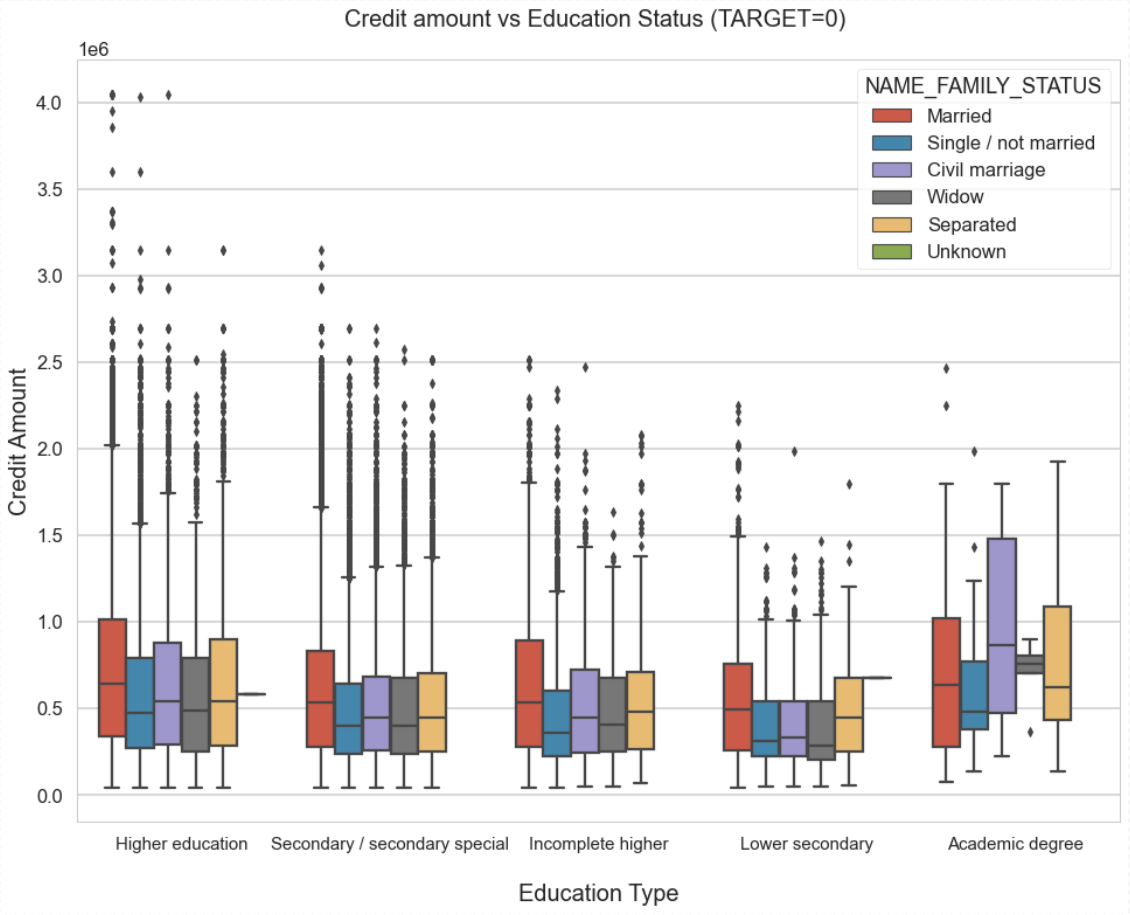
- Cash loans are more credited.
- Those who are female and own car they got little more number of loans
- State servant got more number of loans
- Higher education got more loans
- Applicants who are living in municipal and office apartment, got more number of loans
- High income group people got more loans.
- Mid age people tend to get more number of loans

Multi Variate Analysis

For Non Defaulters

- 'Higher education' the income amount is mostly equal with family status. It does contain many outliers.

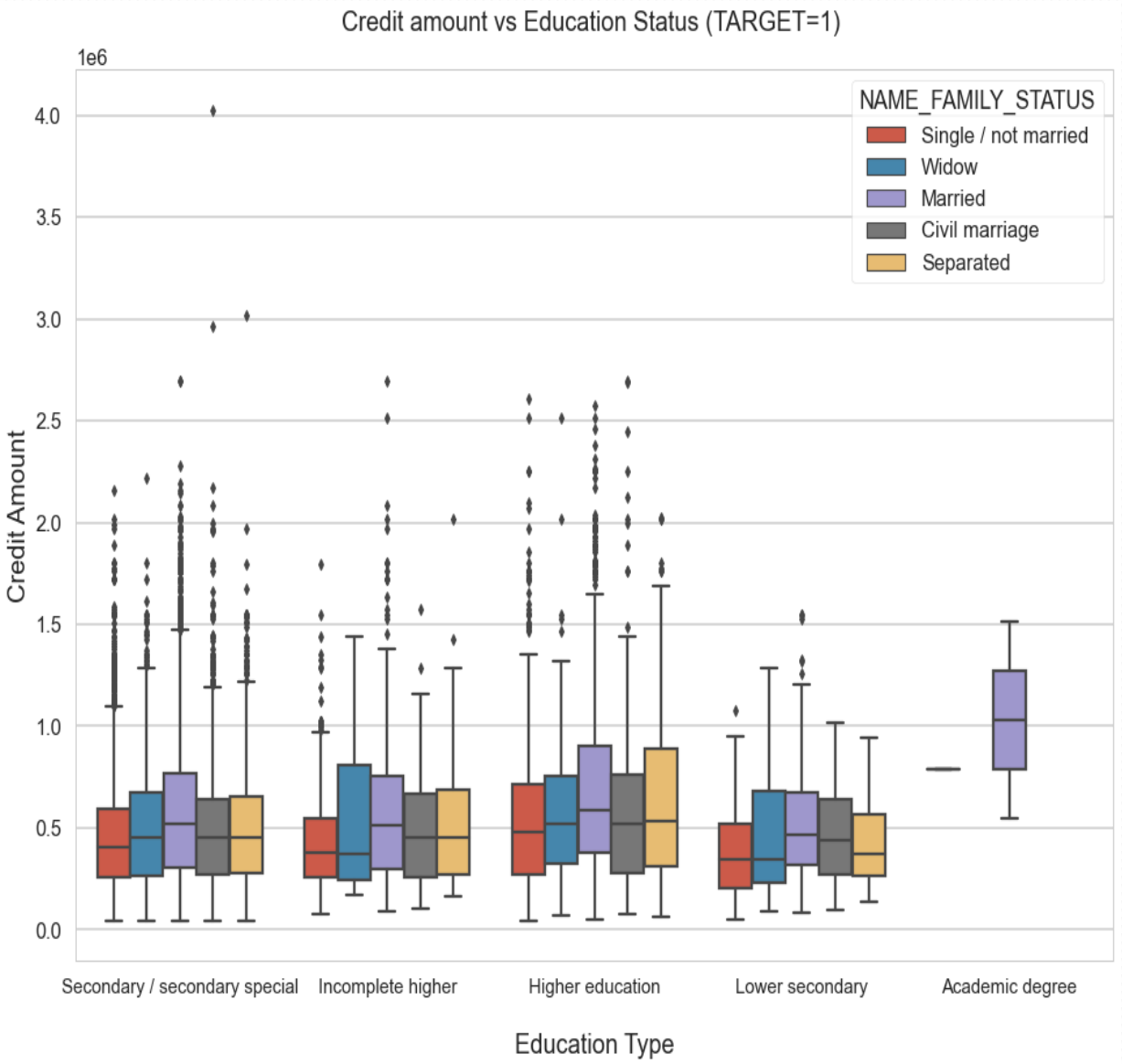
- Less outlier are having for Academic degree but there income amount is little higher that Higher education.
- Lower secondary of civil marriage family status are have less income amount than others.



For Defaulters

We can say that Family status of 'civil marriage','marriage' and 'separated' of Academic degree education are having higher number of credits than others.

Most of the outliers are from Education type 'Higher education' and 'Secondary'. Civil marriage for Academic degree is having most of the credits in the third quartile.



Analysis on Merged dataset (Previous and current Application)

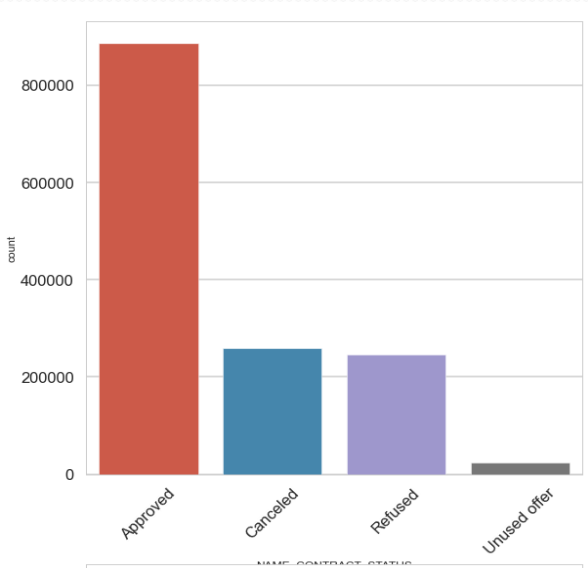
Univariate analysis on categorical columns

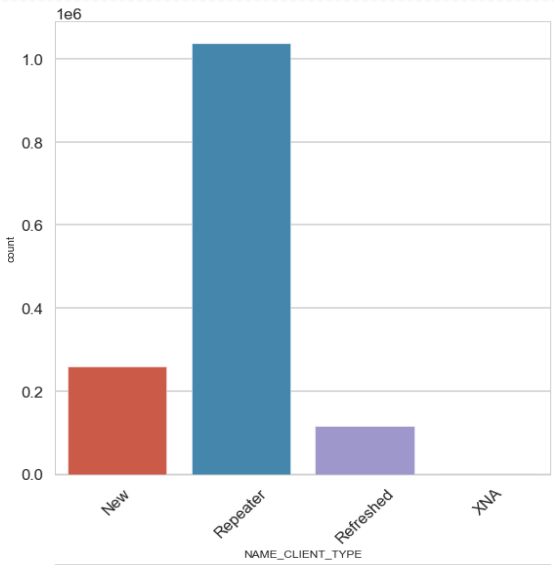
1.Approved loan status is huge than rejected or canceled.

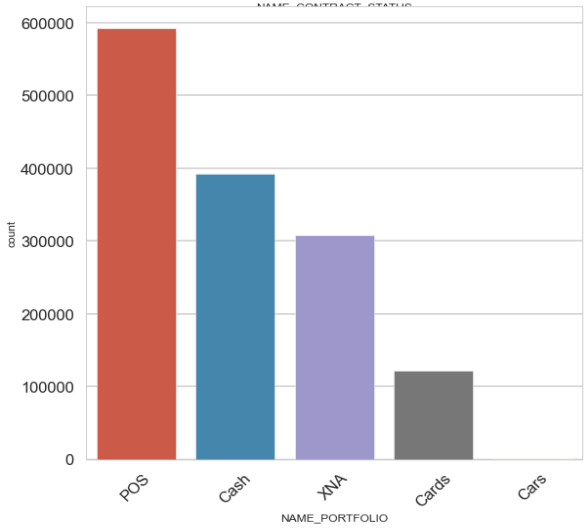
2. Repeater clients are highest in number than new client.

3.POS loans are highest rather than cash loans.

4.Credit and cash offices channel type is the most used channel followed by Country-wide channel







Univariate analysis on continuous columns

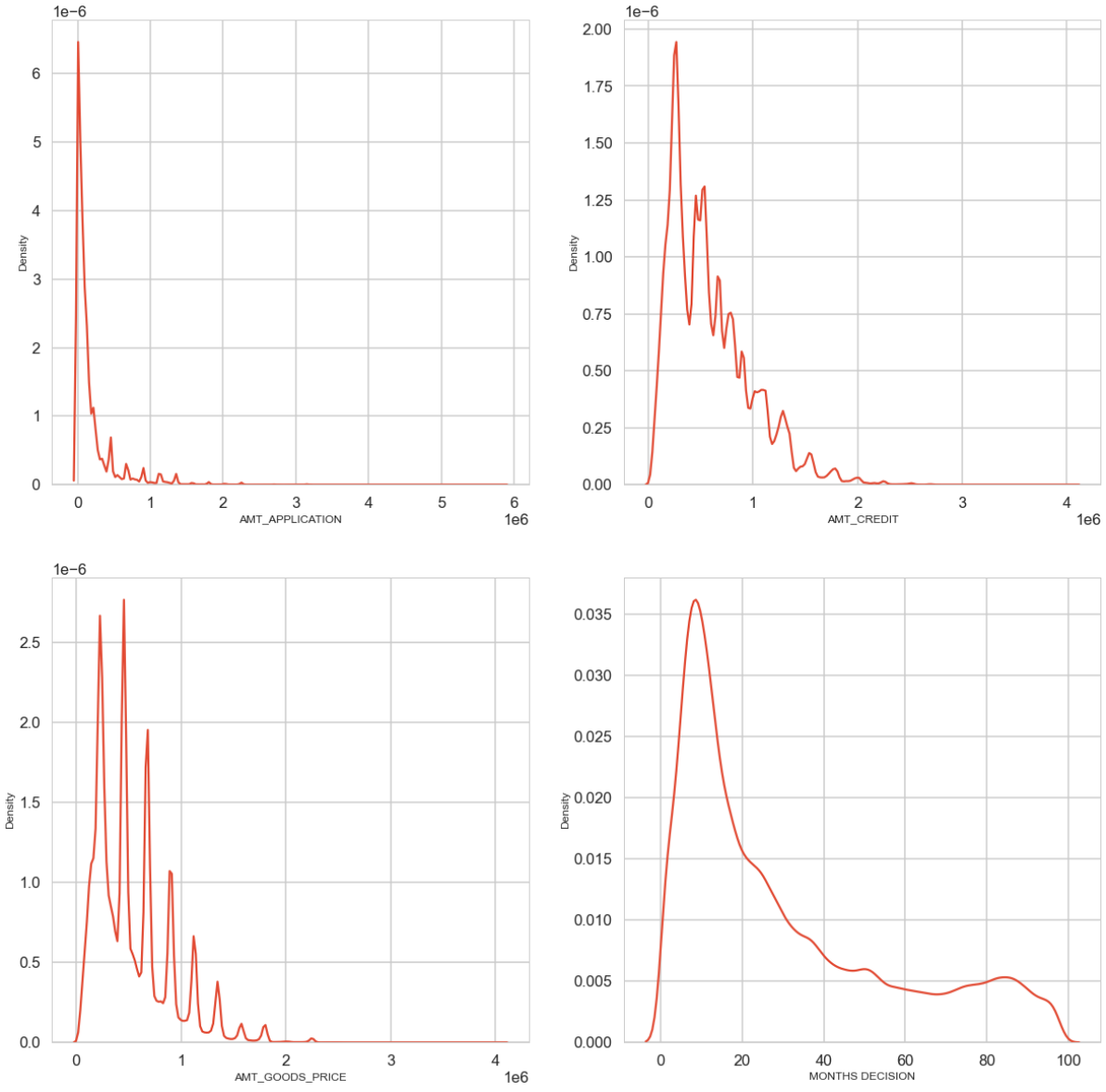
Analysis

1.Most of the loan application amount were below 500000, we can see a huge spike around 100000 amount.

2.Amount credited, is also following the pattern of loan application. We already saw that most of the application was approved in previous plots.

3.Amount of the goods price is also following the same distribution like application amount and amount credited. Because, based on the price of the goods, the loan was approved and amount was credited.

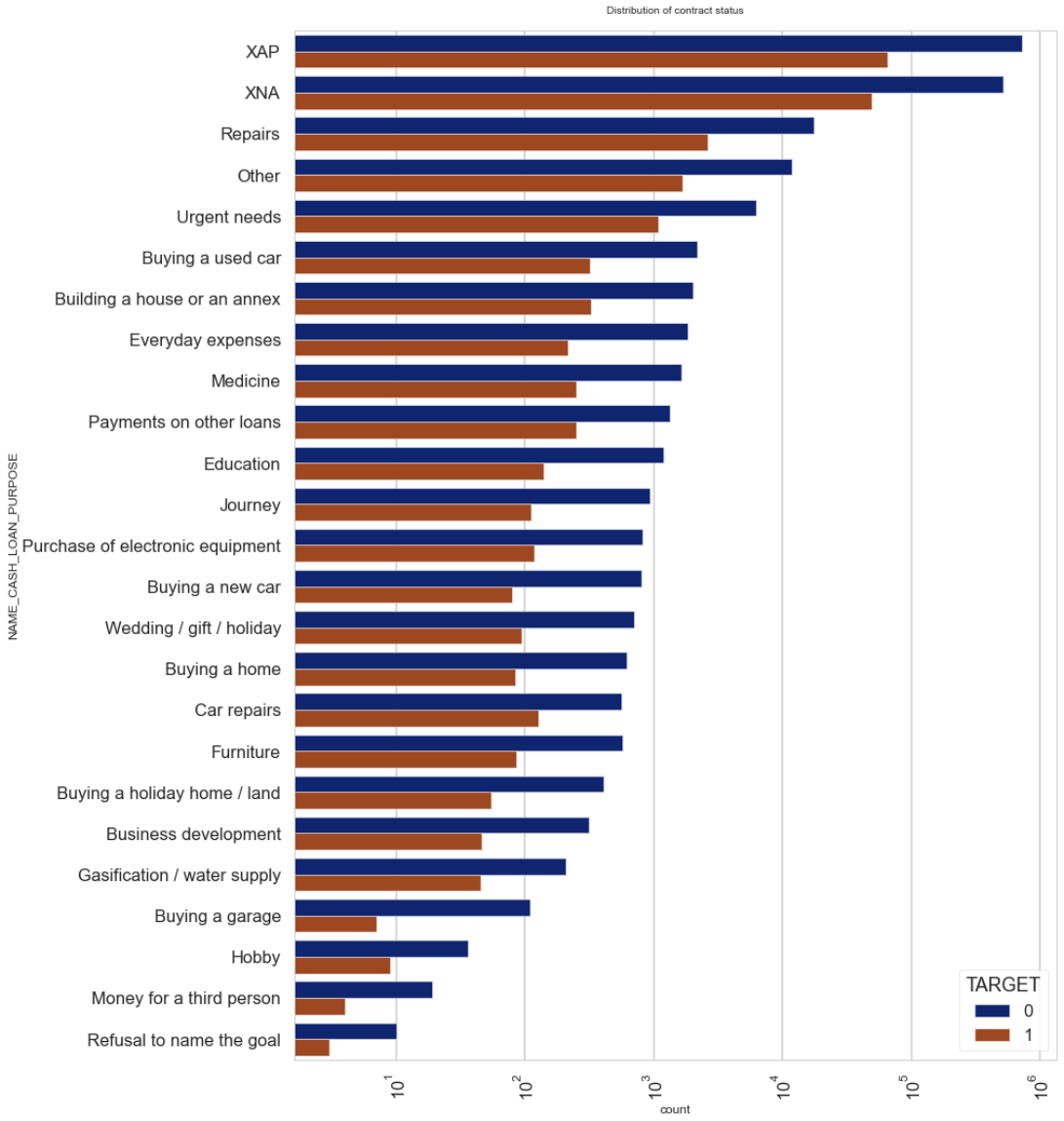
4.Most of the applications decision took around 10 to 20 months



Plotting on contract status vs target

Loan purposes with 'Repairs' are facing more difficulties in payment on time.

2. There are few places where loan payment is significant higher than facing difficulties. They are 'Buying a garage', 'Business development' , 'Buying land',' Buying a new car' and 'Education' Hence we can focus on these purposes for which the client is having for minimal payment difficulties.



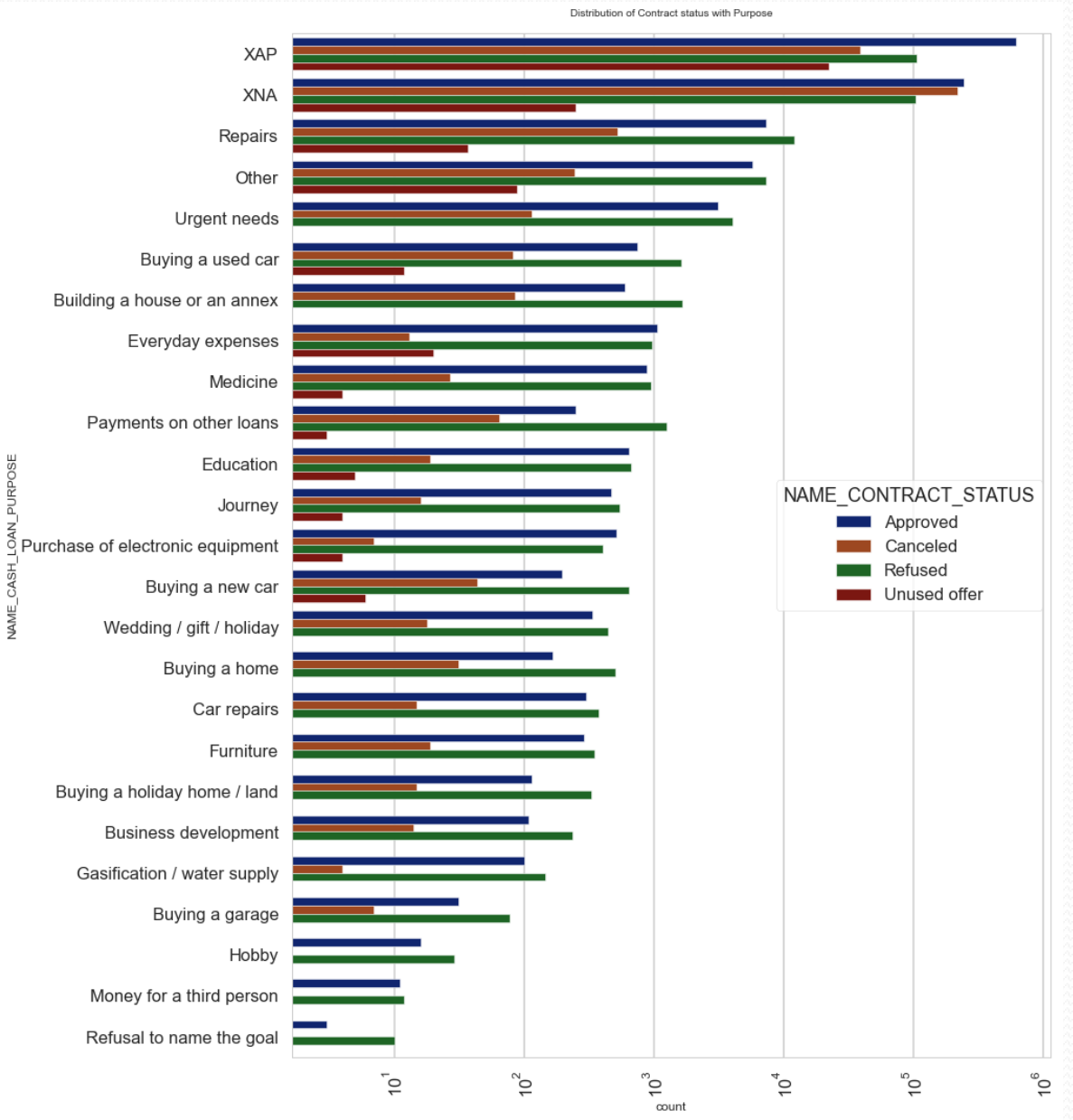
Plotting on contract status vs target with purpose

1.Most rejection of loans came from purpose 'Repairs'.

2.For education purposes we have equal number of approves and rejection.

3.There is more rejection than approval for buying used car

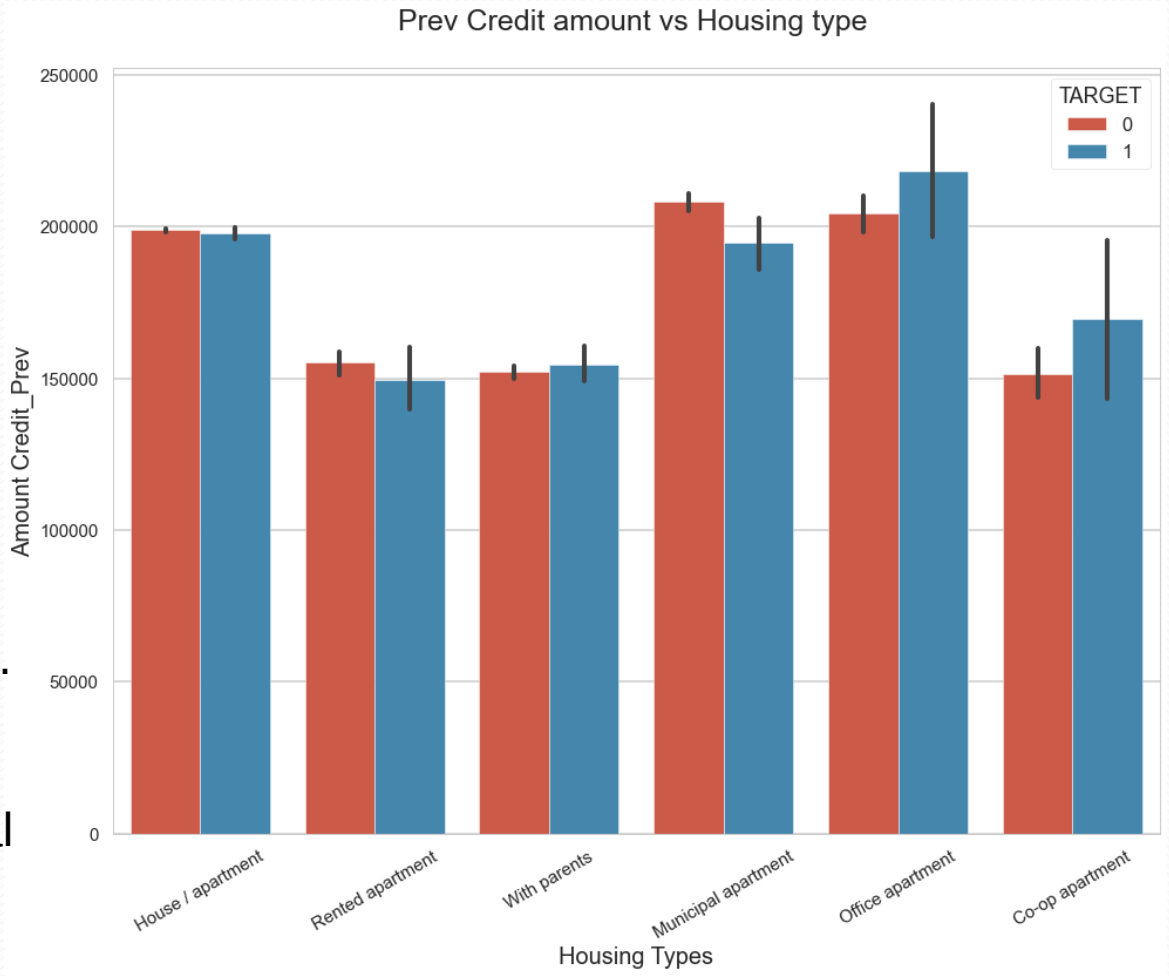
4.Paying other loans and buying a new car is having significant higher rejection than approves



❖Bivariate Analysis on Merged data

- Here for Housing type, Municipal is having higher credit of target 0 and office apartment is having higher credit of target=1.

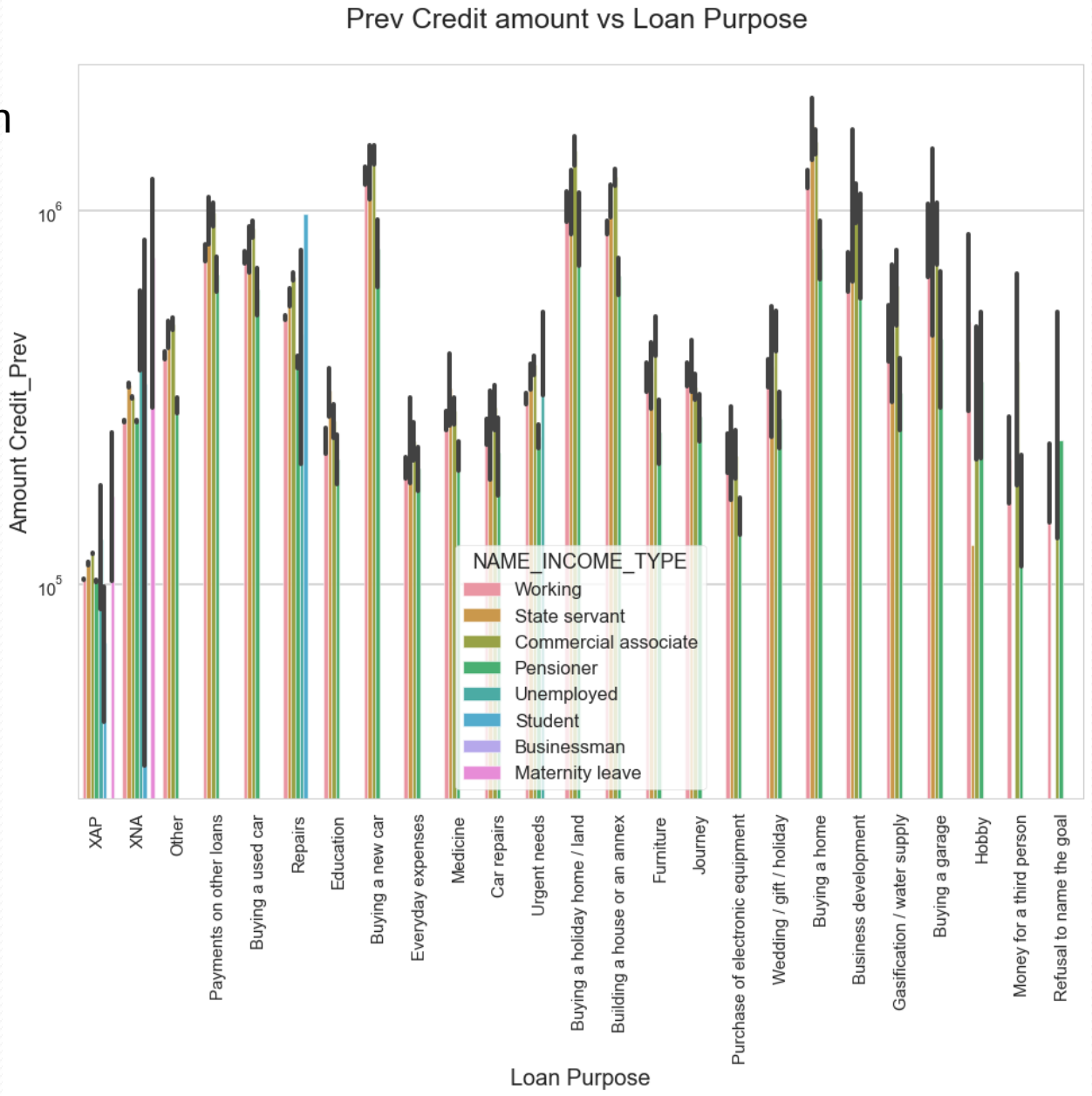
- So, we can conclude that bank should avoid giving loans to the housing type of office apartment as they are having difficulties in payment. Bank can focus mostly on housing type with parents or rented apartment or muncipal apartment for successful payback of loan.



1.The credit amount of Loan purposes like 'Buying a home','Buying a land','Buying a new car' and'Building a house' is higher.

2.Income type of state servants have a significant amount of credit applied

3.Money for third person or a Hobby is having less credits applied for.



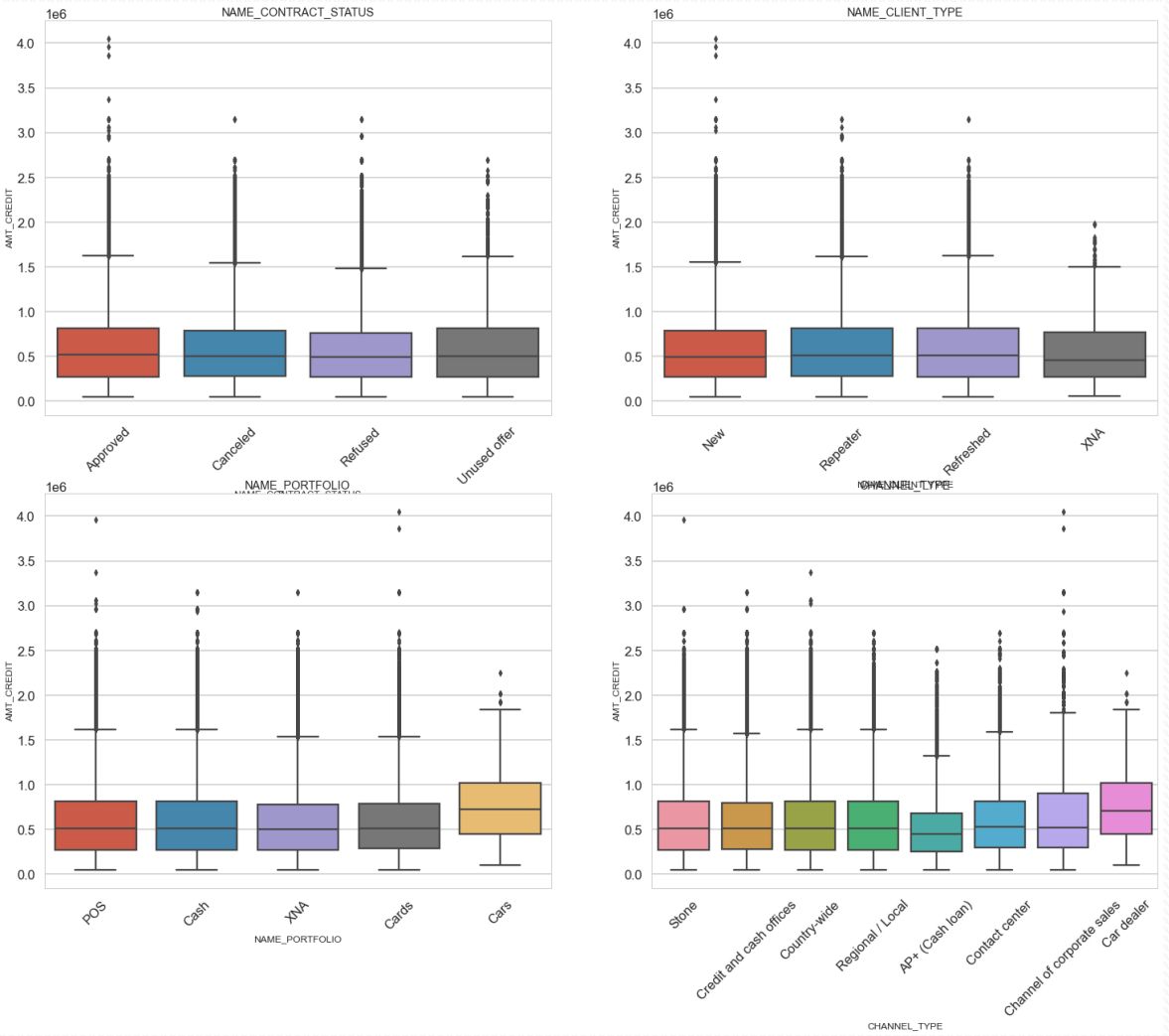
Bivariate analysis on categorical columns

1.There were more unused offer, more outliers noticed in approved application who have taken larger credit

2.Outliers amore in new client type, Refreshed and Repeater client have same median

3.Car loan got more credited of all portfolio.

4.Through the car dealer more loan got credited, followed by channel of corporate sales



Conclusion

1) Bank can focus mostly on housing type with parents or rented apartment or municipal apartment for successful payback of loan.

2) Banks should provide loans to Repairs and other purposes.

3) Banks should provide loans to the Business Entity Type-3 and Self-Employed persons.

4) Working people especially female employers are the best to target for the loans