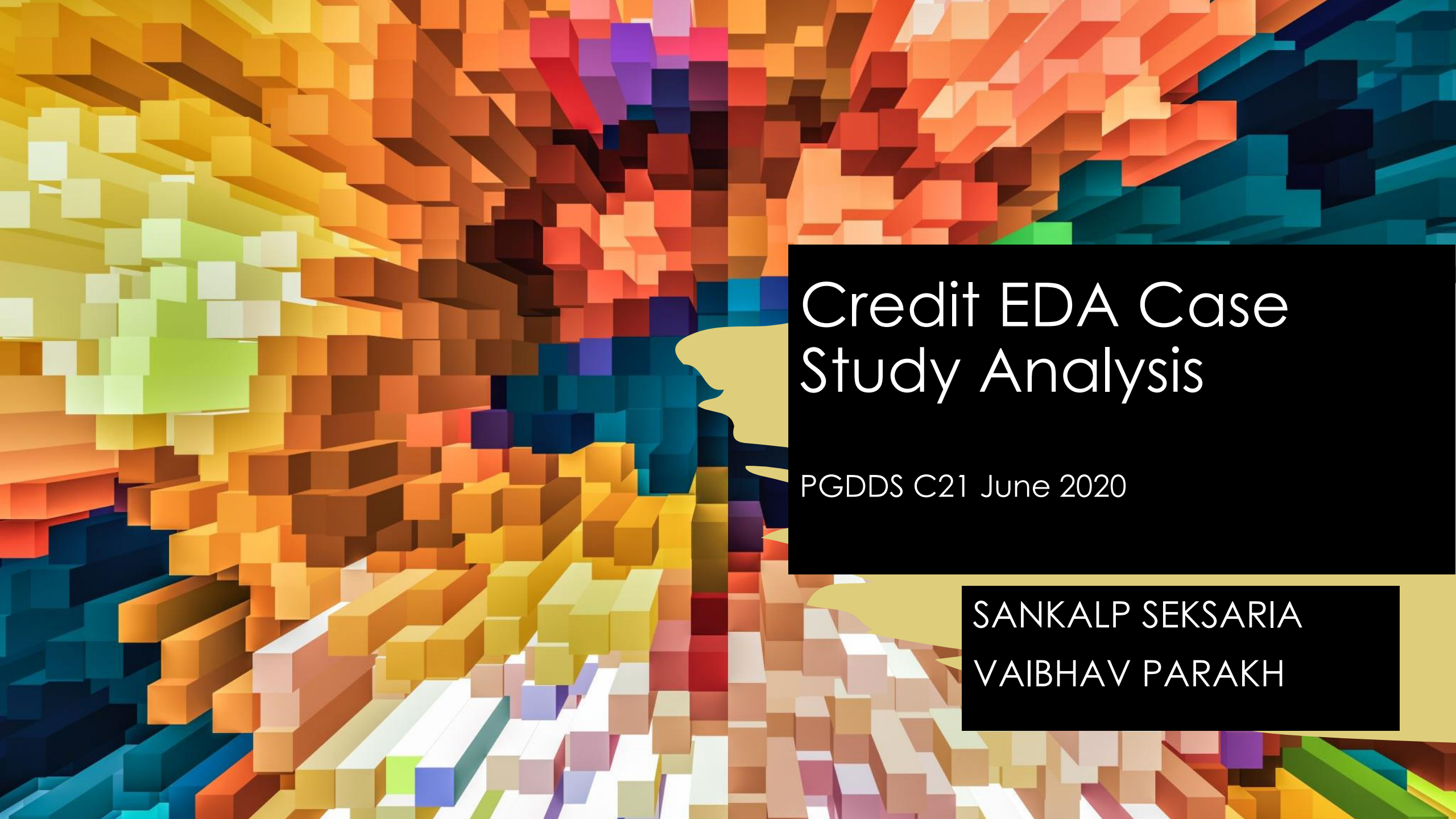


Credit EDA Case Study Analysis

PGDDS C21 June 2020

SANKALP SEKSARIA

VAIBHAV PARAKH

Credit EDA Analysis

- Objective
  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
  - Analyse, if the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.
- Data Set
  - Current applications
  - Previous applications
- Execution Plan
  - Step 1: Reading the data
  - Step 2: Inspecting and Understanding the data
  - Step 3: Data Cleaning of Current Application data
  - Step 4A: Analysis of current application dataset
  - Step 4B: Univariate Analysis for categorical variable application
  - Step 4C: Univariate Analysis for Numerical Variables of Current Application
  - Step 4D: Bivariate Analysis of Numerical Variables of Current application
  - Step 4E: Analysis of Categorical and Numerical Variables of Current Application
  - Step 4F: Multivariate Analysis of Target 0 and Target 1
  - Step 5: Merging Current Application and Previous Application
  - Step 6: Analysis for Categorical Variables

Step 1: Reading the data

Reading the data set – curr_app1 and prev_app1

Step 2: Inspecting and Understanding the data

Inspecting and understanding the data
- check few records for data set such as .head(), .shape, .info(), .describe().T

Step 3: Data Cleaning of Current Application data

Data Cleaning of Current application dataset
- Analyzed number of high null columns: 41
- Dropped columns having >50% of null values.
- Dropped another 21 columns after finding low null columns: 26

Description of Columns containing less than 50% of null values:
- Columns like OCCUPATION_TYPE, NAME_TYPE_SUITE, AMT_GOODS_PRICE, AMT_ANNUITY, CNT_FAM_MEMBERS.

From 26 columns we reduced to 5 columns with low % of null values.
- We do not impute the missing values here for the 5 columns
- Further dropped columns which will not be aiding our analysis.

After dropping we find
- DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH is given by number of days and convert to years.

Analysis/outcome after cleaning of Current Application Data – (Graphs displayed below)

- While reviewing the statistics again, we found that: 'CNT_CHILDREN' i.e the maximum count of the children is shown as 19 which is clearly an outlier.

- 'AMT_INCOME_TOTAL' has a maximum value of 117000000 which is a huge variation from the 75th percentile.

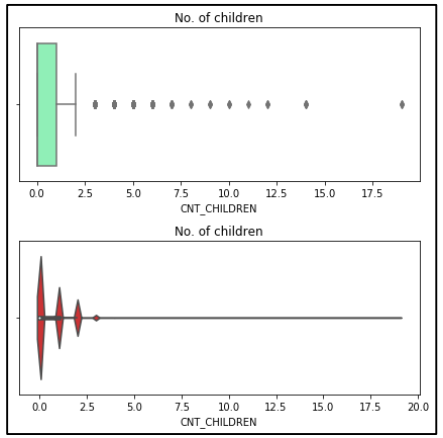
- 'AMT_CREDIT', here the maximum value of 4050000 is very far from the median and the 75th percentile.

- 'AMT_GOODS_PRICE', here the maximum value of 4050000 is very far from the median and 75th percentile value.

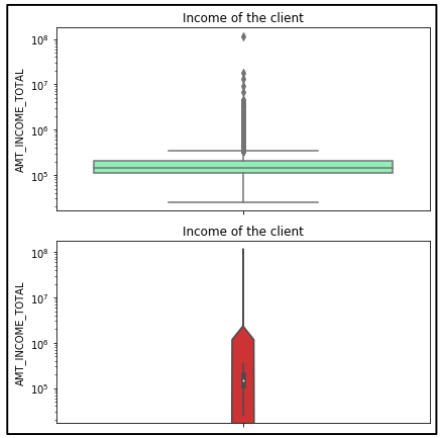
- 'WORK_EX_CLIENT' has a clear outlier with a maximum value of 1000.

- This is visually represented using boxplots/violin plots and distplot as shown below for 'AMT_INCOME_TOTAL', 'CNT_CHILDREN', 'AMT_CREDIT' and 'YEARS_EMPLOYED' in next slide.

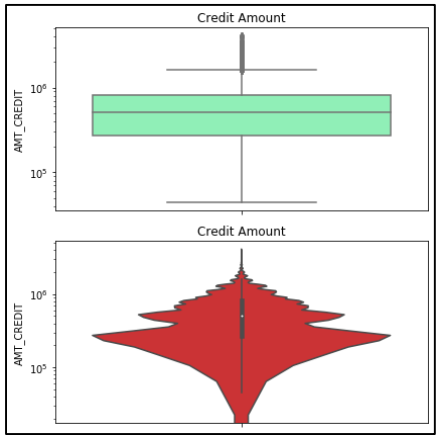
#1 CNT_CHILDREN



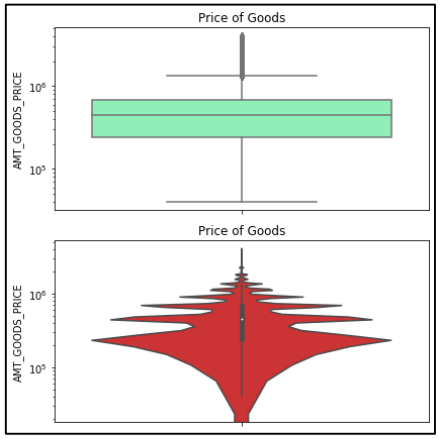
#2 AMT_INCOME_TOTAL



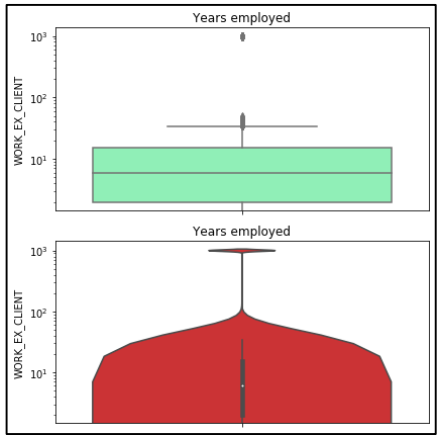
#3 AMT_CREDIT



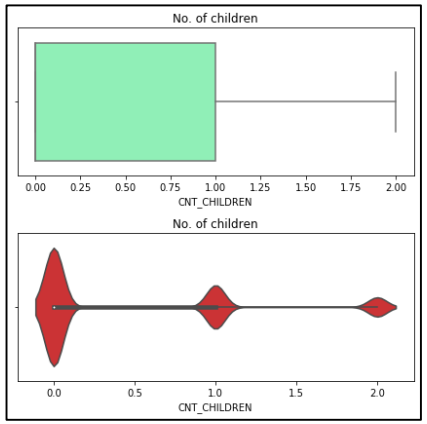
#4 AMT_GOODS_PRICE



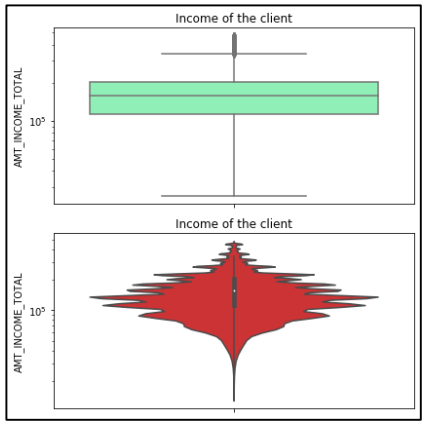
#5 'WORK_EX_CLIENT'



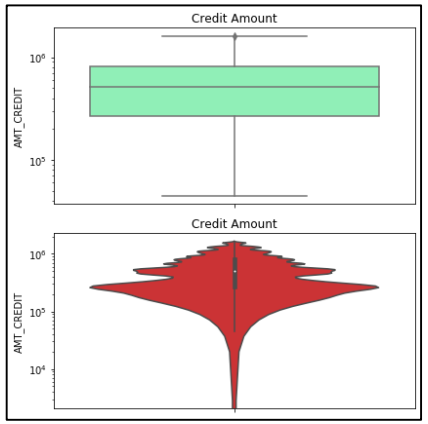
#1 CNT_CHILDREN



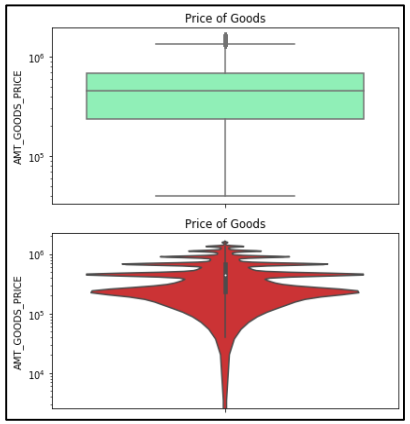
#2 AMT_INCOME_TOTAL



#3 AMT_CREDIT

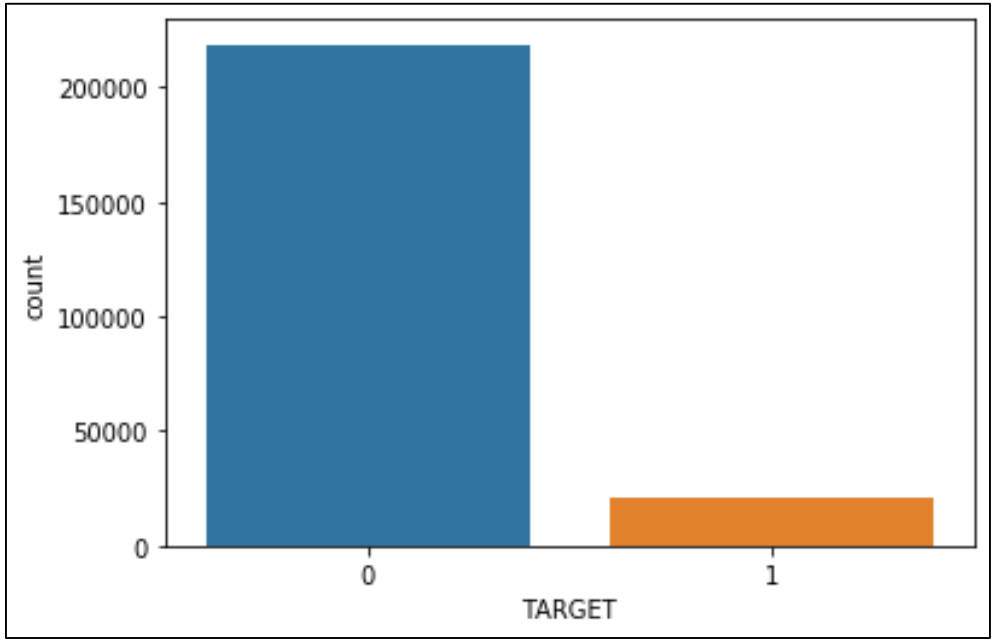


#4 AMT_GOODS_PRICE

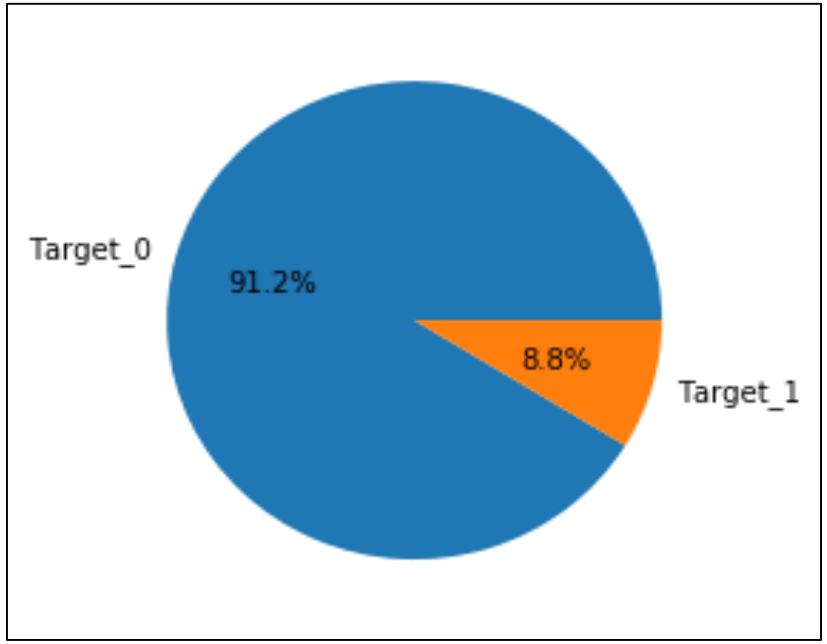


#5 'WORK_EX_CLIENT'

#Analyzing the count of Target Variable



To check imbalance percentage



Step 4A: Analysis of current application dataset

Target 0: Client without payment difficulties
Target 1: Client with payment difficulties

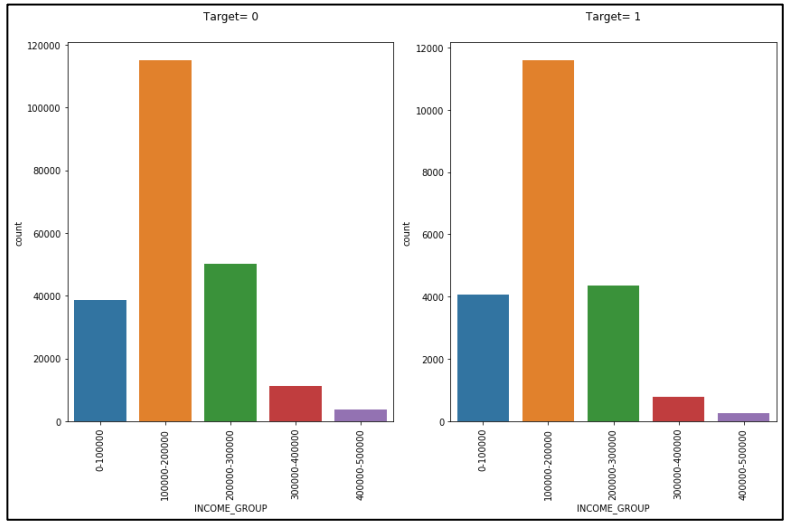
Since there is huge imbalance between the Target variables 0 and 1, it makes more sense to divide the data frame into two sub datasets and then continue our analysis.

In order to analyze the imbalance and various aspects of data we will perform various types of analysis such as:

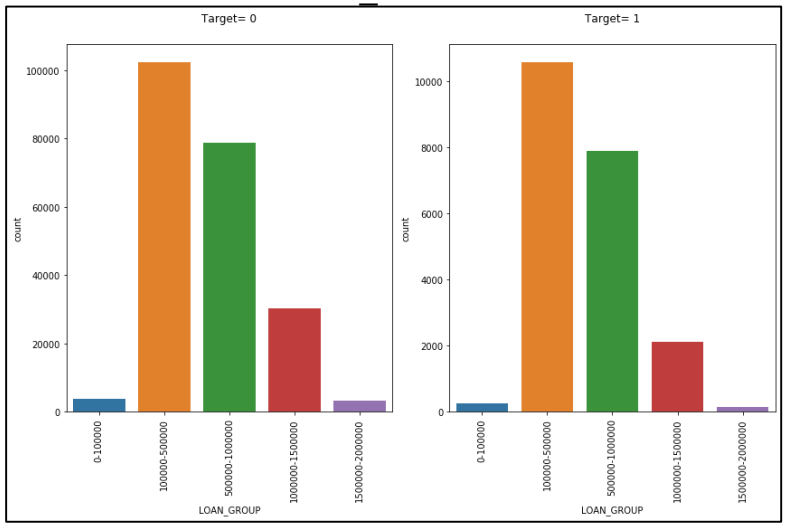
Univariate analysis
Bivariate analysis
Multivariate analysis

4B: Univariate Analysis for categorical variable application

INCOME_GROUP

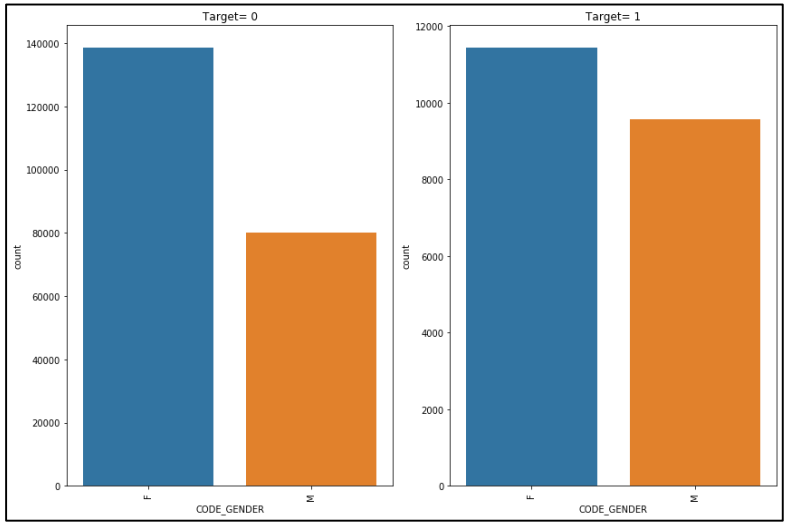


LOAN_GROUP

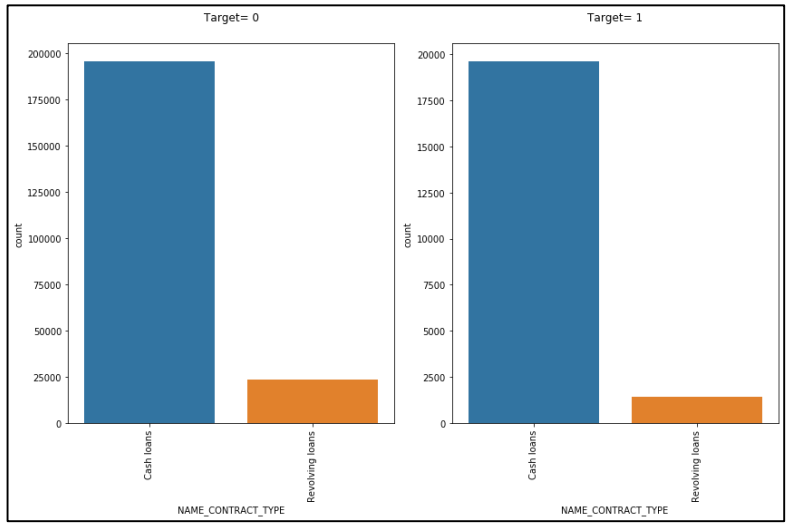


There is similar proportion of people belonging to different income groups who either have no payment difficulties or are having payment difficulties. Though, the count for people belonging to different payment groups having no payment difficulties far outweighs the count for people belonging to different payment groups having payment difficulties.

From perspective of red flags, highest number of defaulters belong to a loan group of Rs 1 Lac to Rs 5 Lac and the second highest number of defaulters belong to a loan group of Rs 5 Lac to Rs 10 Lac.



CODE_GENDER



NAME_CONTRACT_TYPE

For both Target 0 and Target 1, Female application count is more than Male application count. Owing to the data imbalance, the count for females having no payment difficulties far outnumbers the count for females having payment difficulties.
Keeping the statistic in mind, we can make a proposition for preferring females over males when it comes to seeking recipients for loan.

In name contract type of loans, cash loans are more than revolving loans for both Target 0 and Target 1 and in the case of Target 1, people find difficulties in paying loan back involving cash in comparison with revolving loans.

4B: Univariate Analysis for categorical variable application

In both group of application male are more than females , even in default list males' default more than female.

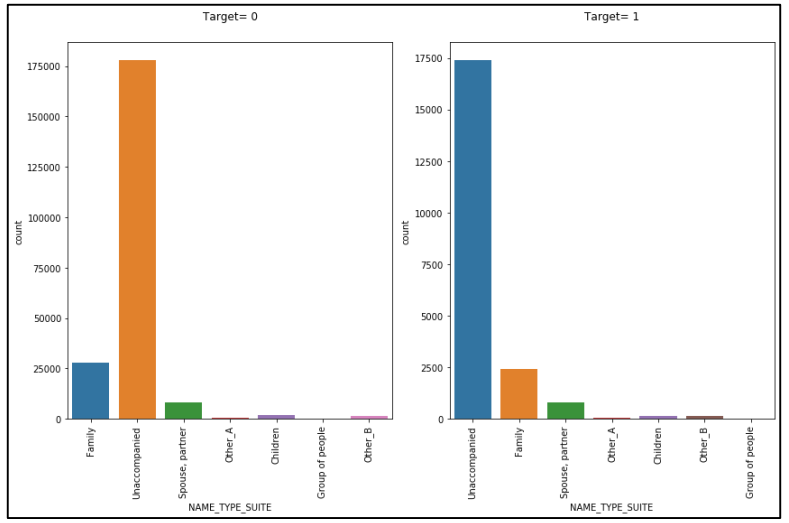
In name contract type cash loan cash loans are more than revolving loans and people find difficulties in paying loan in cash loans.

Number of clients who owned houses are almost double than number of people who do not. Same is in case of people who fail to pay loan back on time.

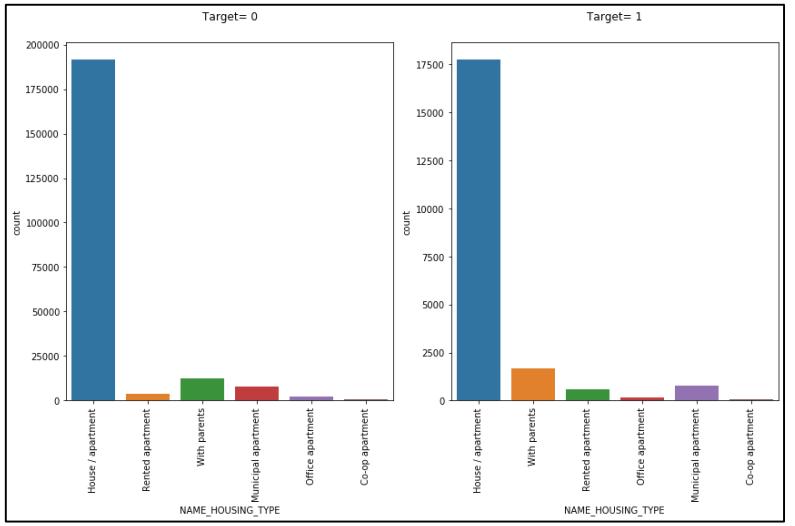
People who do not own a car are almost double than people who own the car and similar in case of people who fail to pay loan back on time.

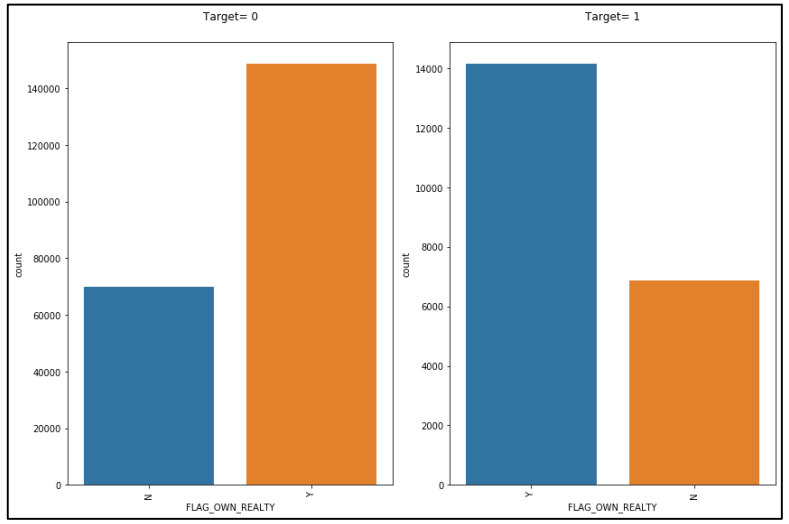
From the graphs, taking into account Target 1 statistics, not having a car and possessing a realty is a greater predictor of failing to pay the loan back on time.

NAME_TYPE_SUITE

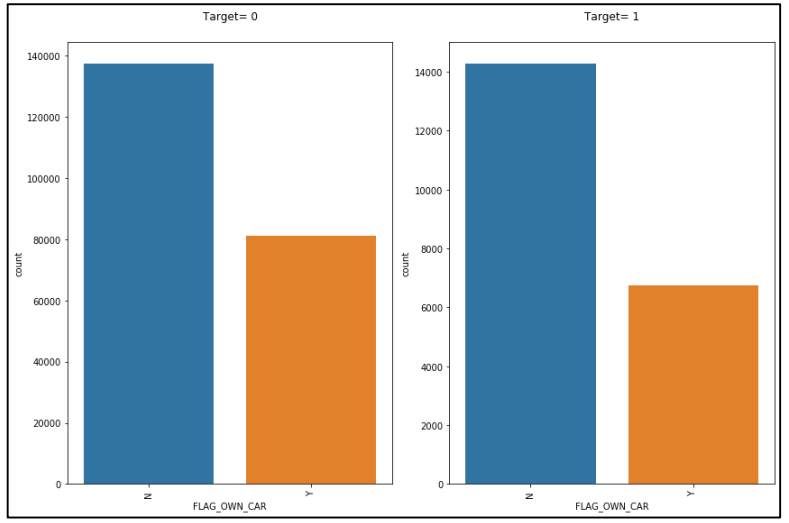


NAME_HOUSING_TYPE





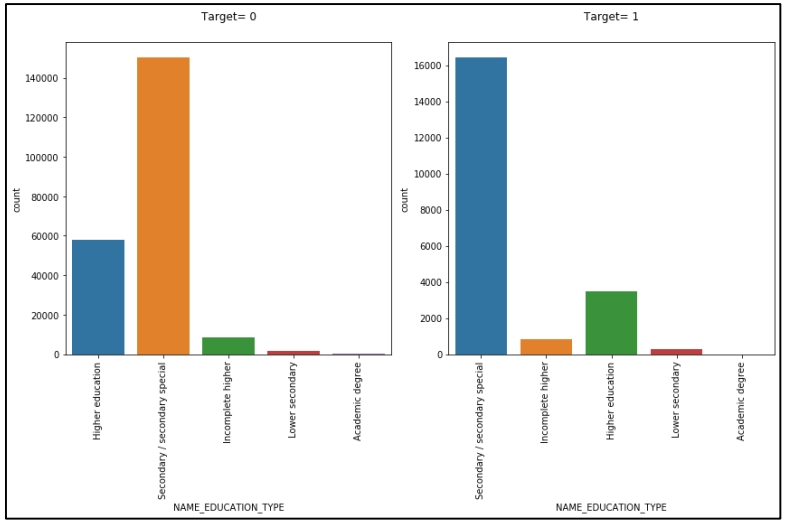
FLAG_OWN_REALTY



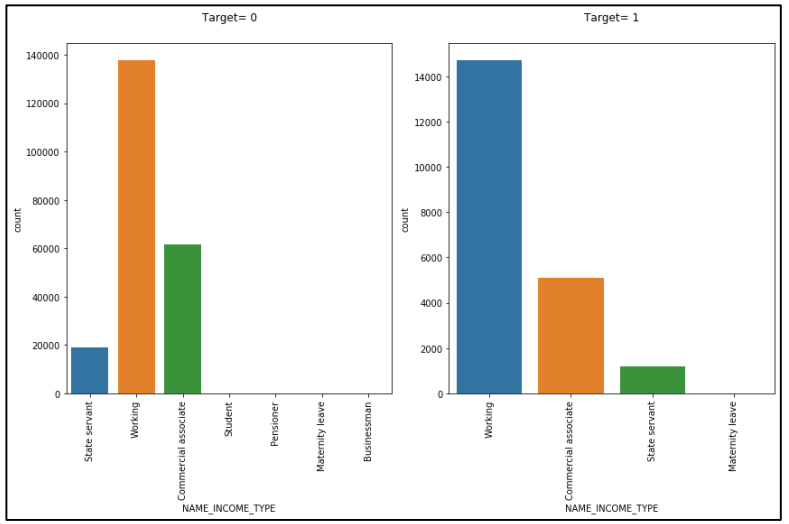
FLAG_OWN_CAR

4B: Univariate Analysis for categorical variable application

NAME_EDUCATION_TYPE

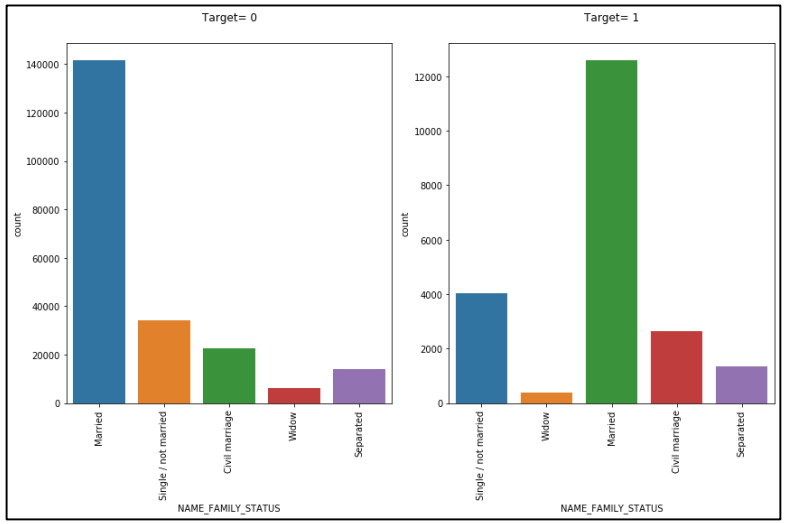


NAME_INCOME_TYPE

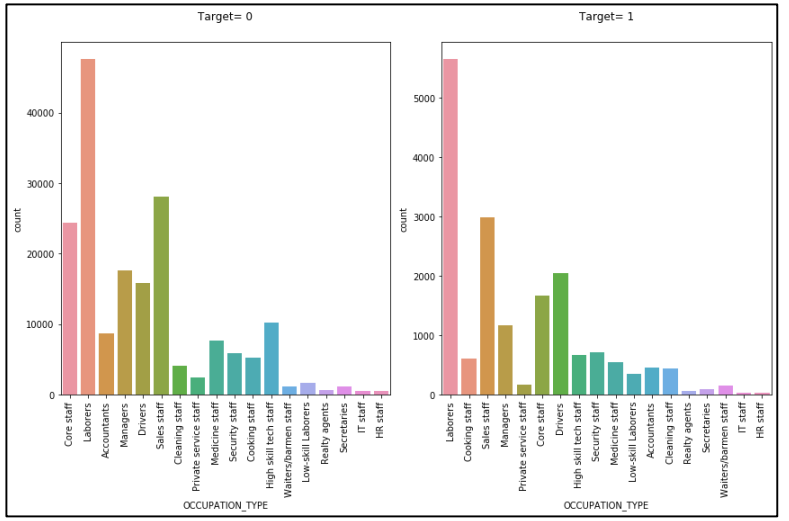


People with academic degree rarely apply for loan whereas people with secondary education are highest in application loan and default of loan payment.

State servants are least in Name Income type to apply and default on loan whereas working professionals are the ones who apply most.



NAME_FAMILY_STATUS



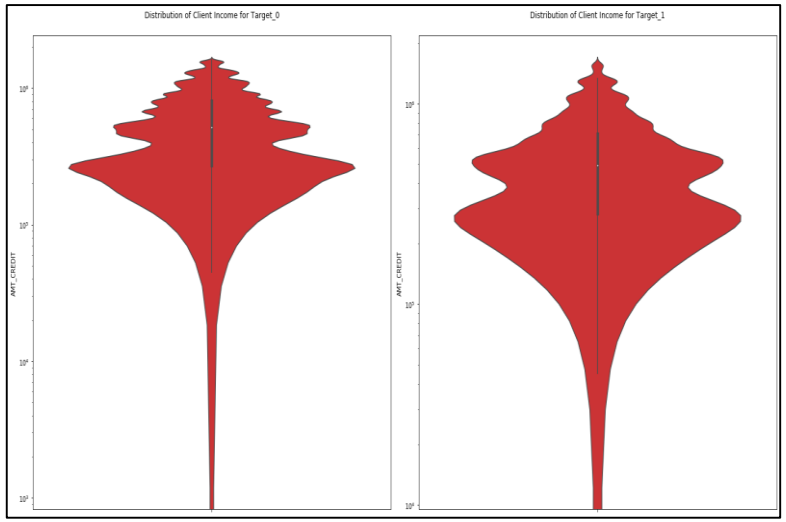
OCCUPATION_TYPE

Number of clients who owned houses are double than number of people who don't. Same is in case of people who default to pay loan on time.

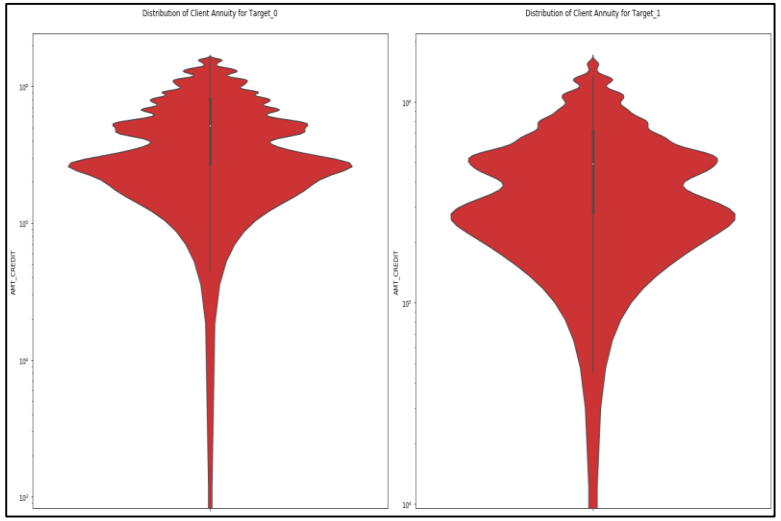
People who do not own a car are double than people who own the car and similar in case of people who default to pay loan on time.

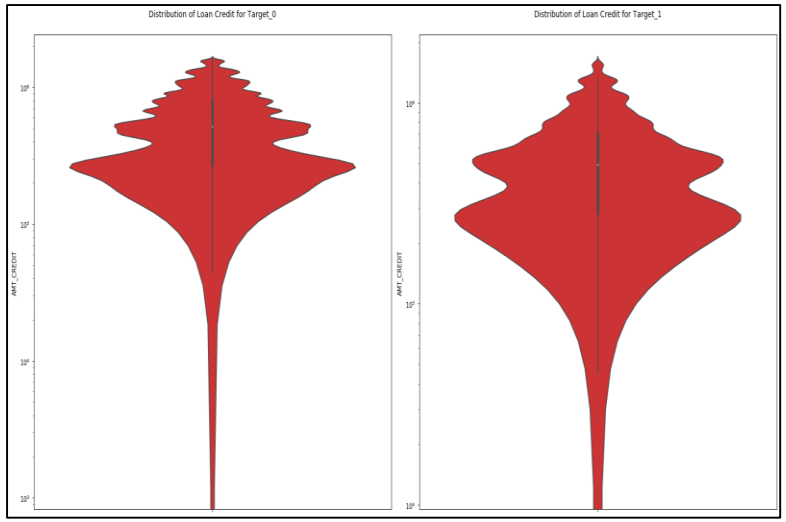
STEP 4C: Univariate Analysis for Numerical Variables of Current Application

AMT_INCOME_TOTAL

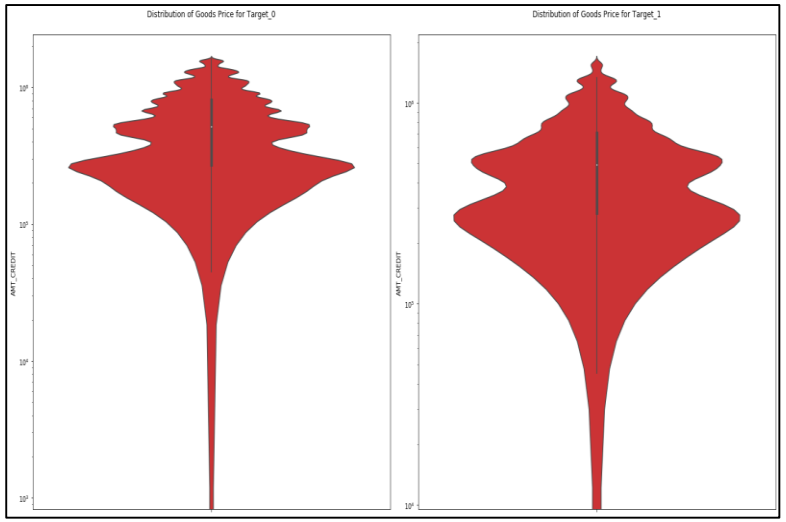


AMT_ANNUITY





AMT_CREDIT



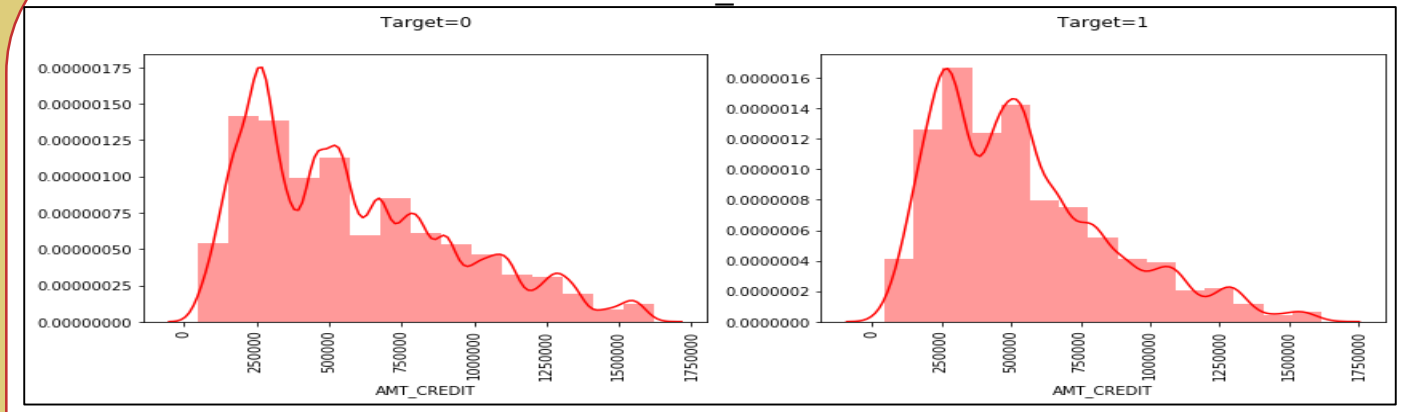
AMT_GOODS_PRICE

People with target one has largely staggered income as compared to target zero. Violin plot clearly shows that the shape in Income total, Annuity, Credit and Good Price are similar for Target 0 and similar for Target 1.

The plots are also highlighting that people who have difficulty in paying back loans with respect to their income, loan amount, price of goods against which loan is procured and Annuity.

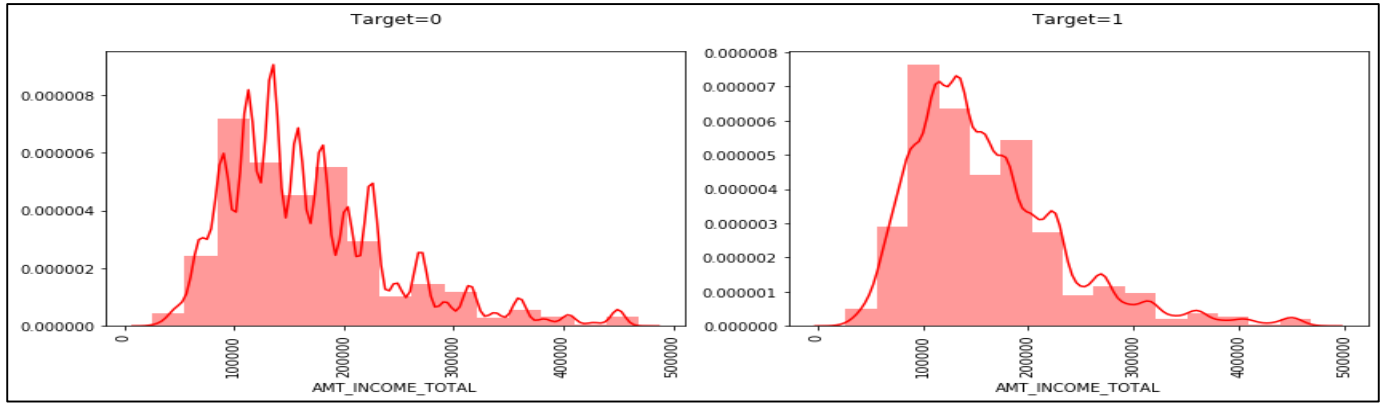
Violin plot highlights the curve shape which is wider for Target 1 in comparison to Target 0 which is narrower with well defined edges.

AMT_CREDIT



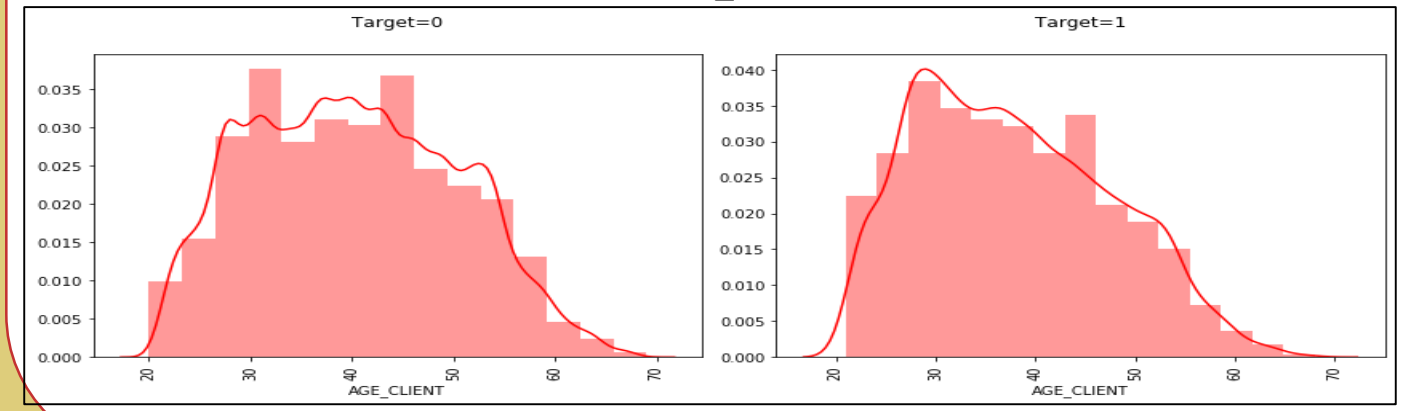
The highest number of applicants apply for loans between Rs 2.5 Lac and Rs 5 Lac with some increase applications on Rs 7.5 Lac whereas difficulty in paying ranges for applicants who have taken loan between Rs 2.5 Lac and Rs 7.5 Lac.

AMT_INCOME_TOTAL



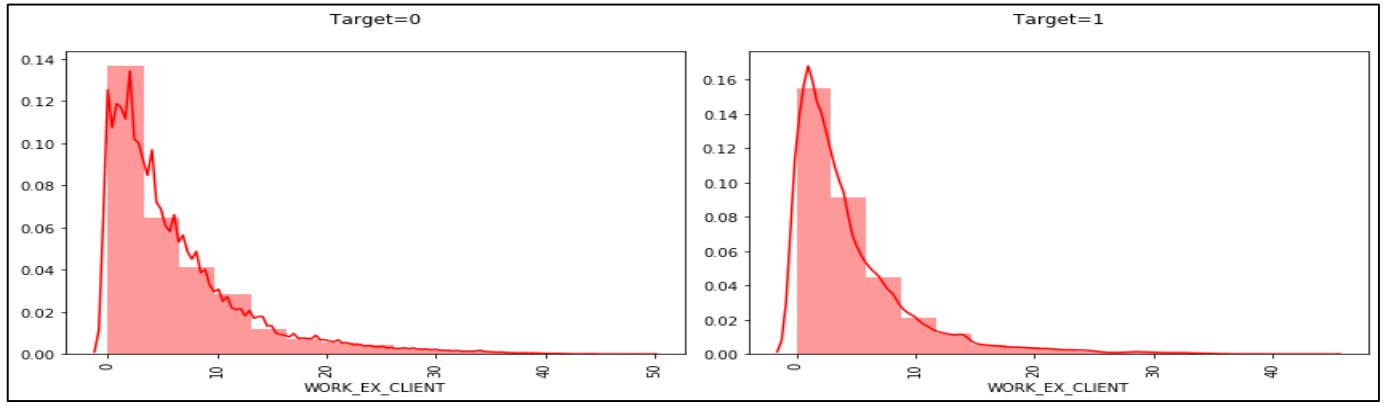
Applicants who have income between Rs1Lac and Rs 1.5 Lac are the one's who apply for the loan in high number and are the same number of people who default in loans.

AGE_CLIENT



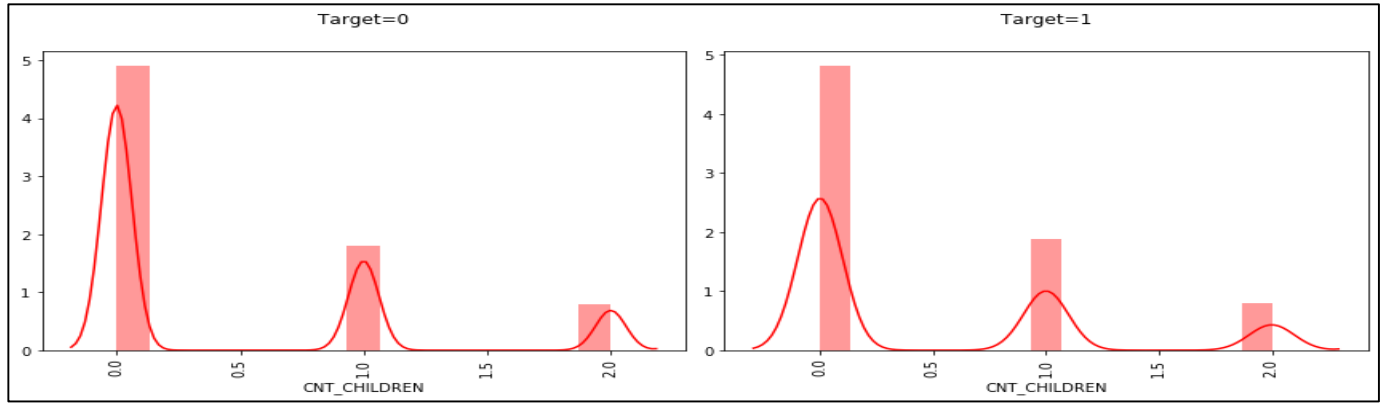
The highest number of clients who apply for the loan are between the age of 30 and 50. Although the highest number of clients who tend to default in their repayment of the loan are in their mid-20's.

WORK_EX_CLIENT



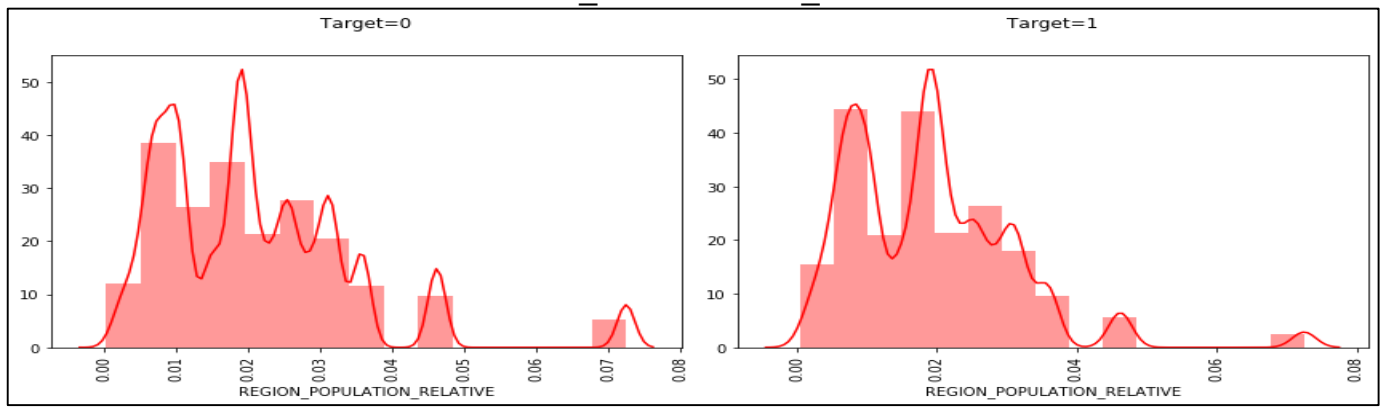
The highest applicants are between 0-10 in work ex client and are the same who are most likely to default in the payments they make.

CNT_CHILDREN



The highest applications for loan are the ones who have no children and are the same who are likely to default on the payment of the loan.

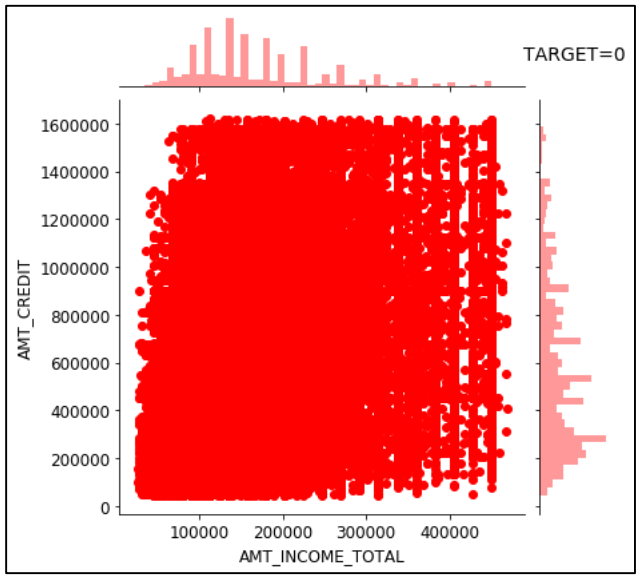
REGION_POPULATION_RELATIVE



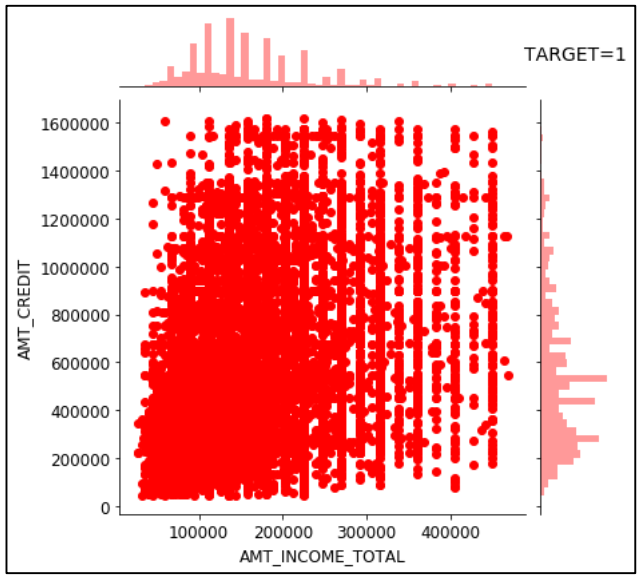
Many applicants who stay in less populated areas have applied for a loan and the same can be defaulter for the loan, as in less populated area employability and wages are less as compare to high populated area.

Step 4D: Bivariate Analysis of Numerical Variables of Current application

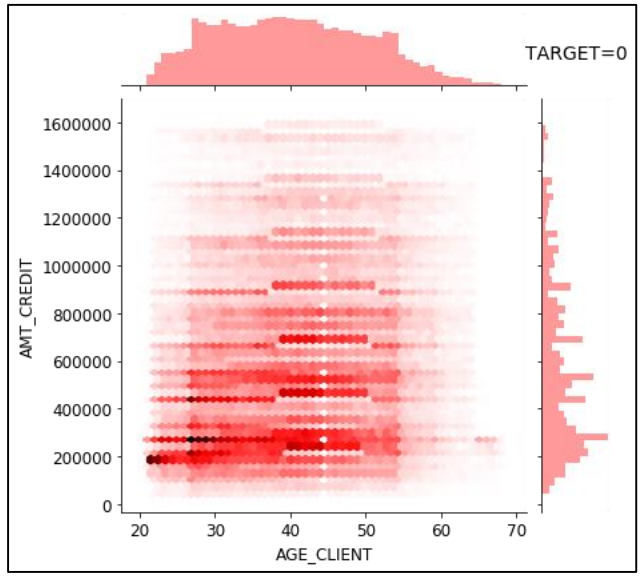
AMT_INCOME_TOTAL



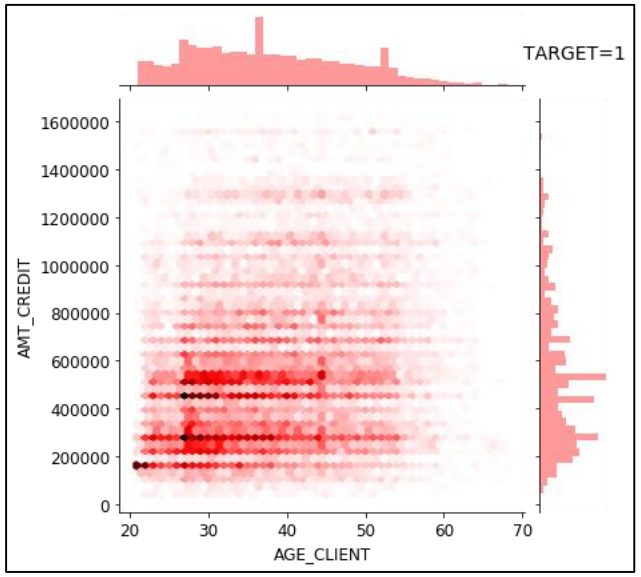
AMT_INCOME_TOTAL



Applicants having income between Rs 1 Lac and Rs 2 Lac apply for double amount of loan i.e. between Rs 2 Lac to Rs 4 Lac.
Similarly applicants between loan amount of Rs 2 Lac to 4 Lac with income of Rs 1 Lac to Rs 2 Lac tend to default.



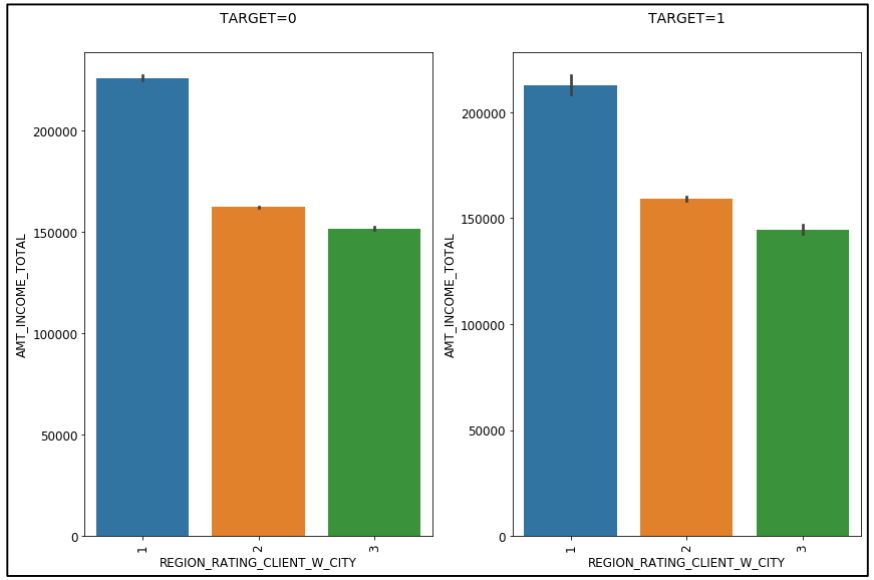
AGE_CLIENT



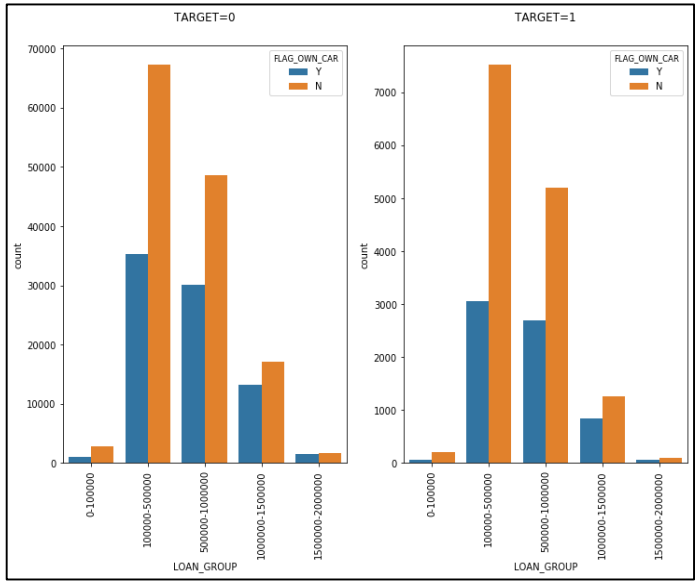
Clients between the age of 20-50 apply for loan ranging between Rs 2 lac to Rs 4 Lac whereas defaults are more between the age of 35-40 for amount between Rs 4 Lac to Rs 6 Lac.
On second number defaulters are between Rs 2 Lac and Rs 4 Lac & between age of 30-35.

Step 4E: Analysis of Categorical and Numerical Variables of Current Application

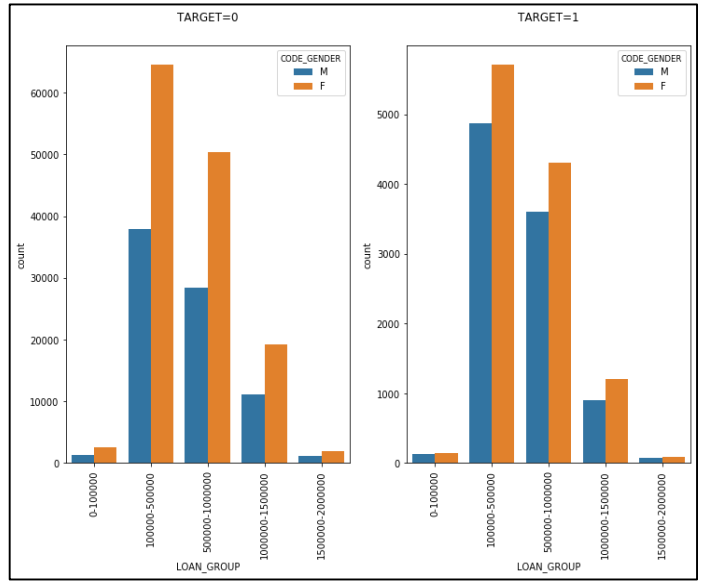
REGION_RATING_CLIENT_W_CITY vs AMT_INCOME_TOTAL

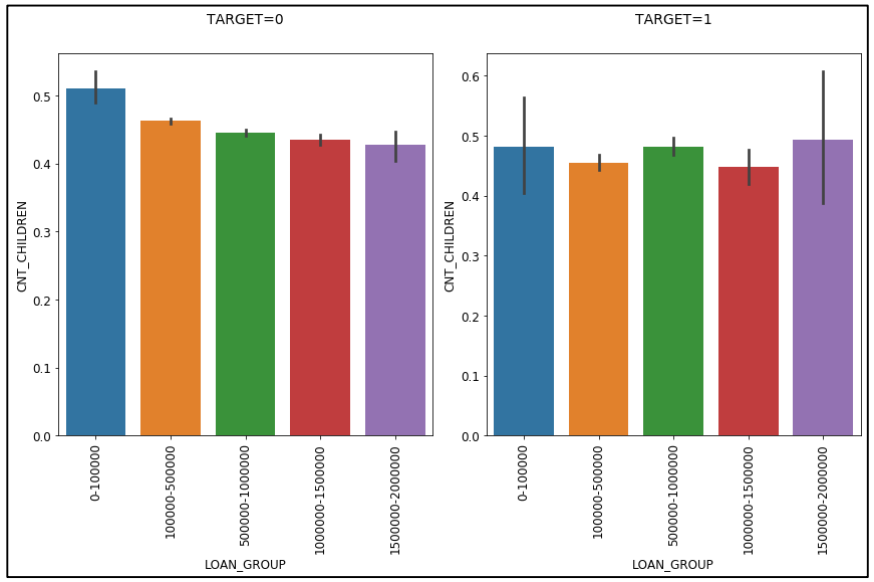


FLAG_OWN_CAR

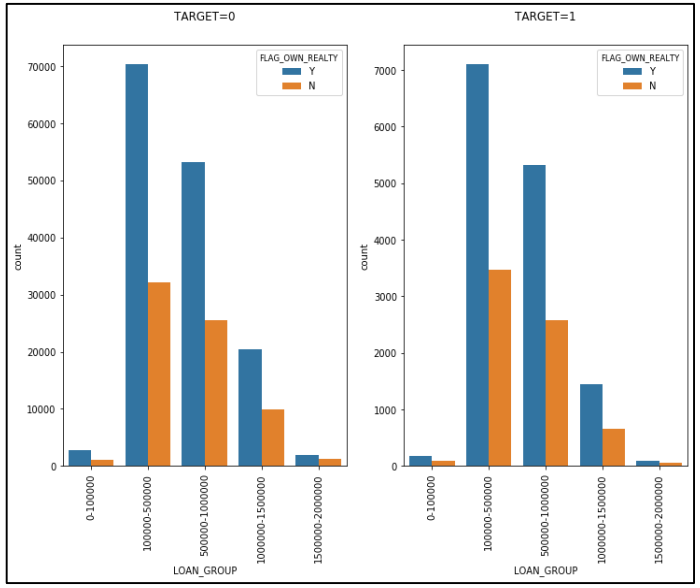


LOAN_GROUP





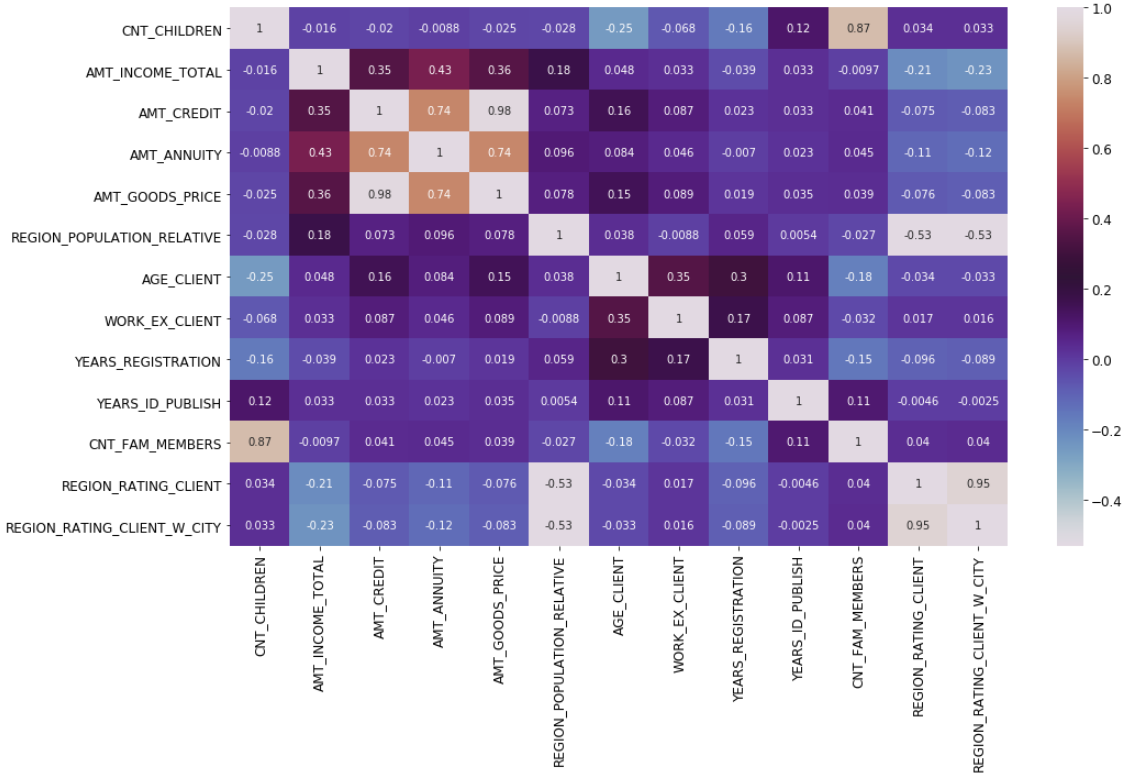
LOAN_GROUP vs CNT_CHILDREN

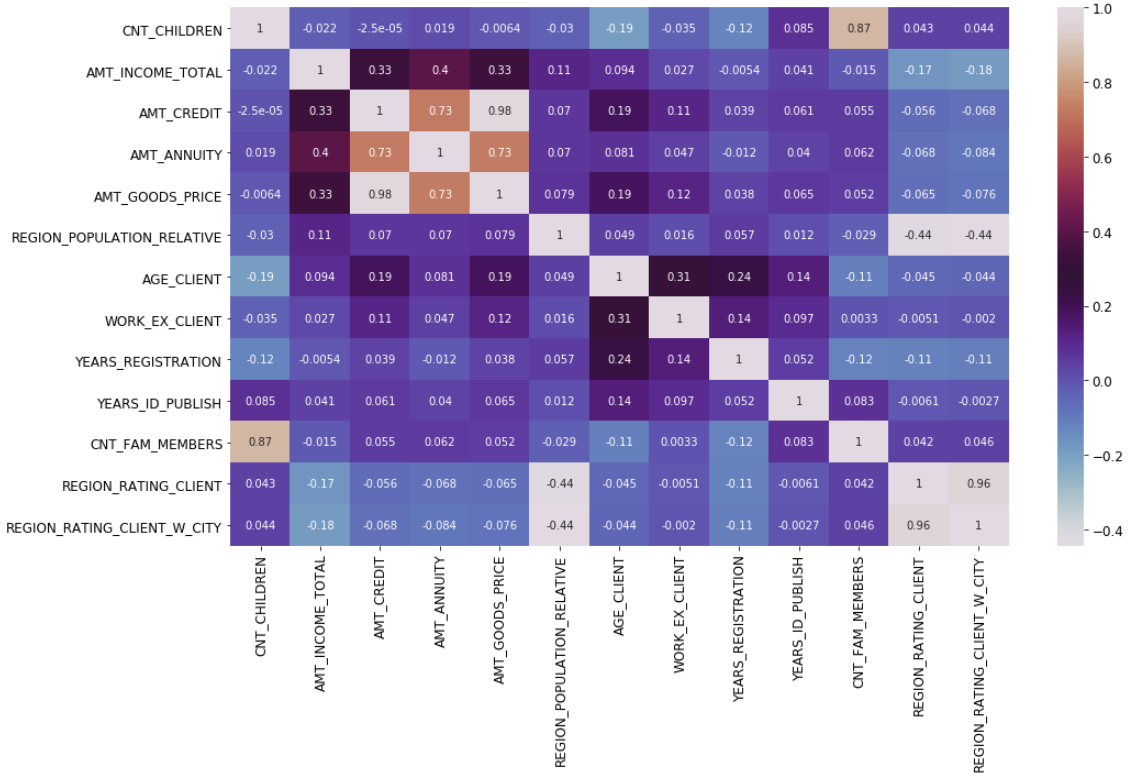


FLAG_OWN_REALTY

- In Region rating, more clients apply for loan in city 1 as compared to 2 & 3 and default also in city 1 may be because of more application.
- In Loan group clients from with more children and less income apply for the loan where in defaults clients with high income and more children default.
- Here also the outcome is like previous analysis where flag own car, flag own realty and loan group have similar applications and defaulters.

Step 4F: Multivariate Analysis of Target 0 and Target 1





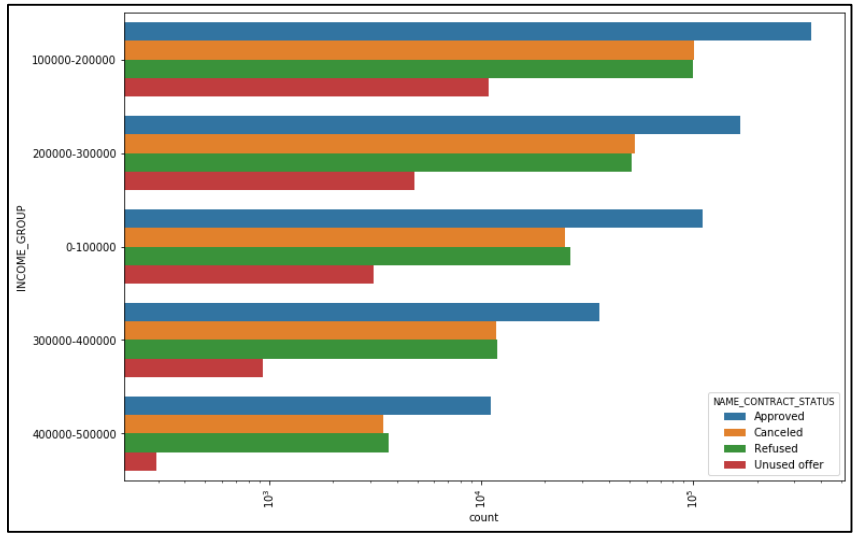
VAR1	VAR2	Correlation
CNT_FAM_MEMBERS	CNT_CHILDREN	0.871534959
AMT_ANNUITY	AMT_CREDIT	0.737342397
AMT_GOODS_PRICE	AMT_ANNUITY	0.741637859
AMT_INCOME_TOTAL	AMT_ANNUITY	0.434437641
AMT_CREDIT	AMT_GOODS_PRICE	0.983612384
REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	0.529495682
WORK_EX_CLIENT	AGE_CLIENT	0.353443870
AMT_GOODS_PRICE	AMT_CREDIT	0.983612384
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.949002598
YEARS_REGISTRATION	AGE_CLIENT	0.299787592

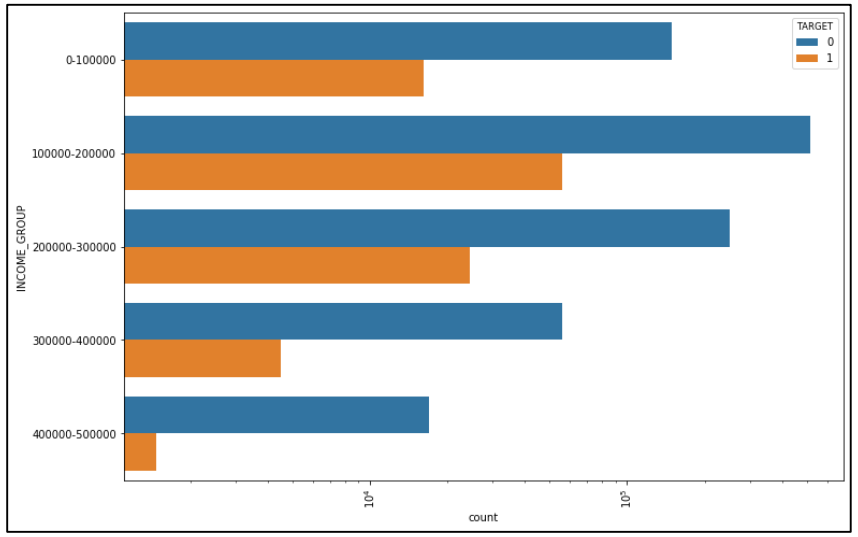
Correlation of CNT_FAM_Members is strong with CNT_CHILDREN, AMT_GOODS_Price is highish correlated with AMT_Credit strong in Target 0.

Region client with city is closely correlated with region rating client in both Target audience.

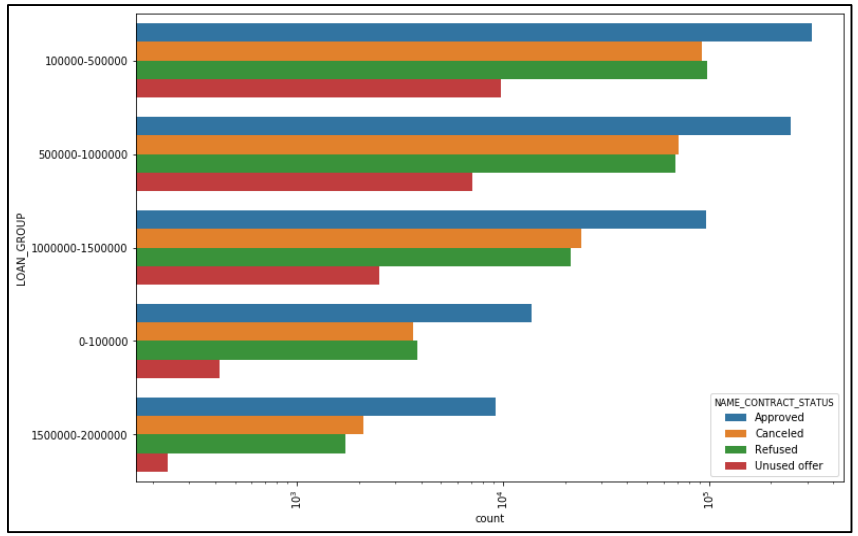
VAR1	VAR 2	Correlation
CNT_FAM_MEMBERS	CNT_CHILDREN	0.868061321
AMT_ANNUITY	AMT_CREDIT	0.727842772
AMT_GOODS_PRICE	AMT_ANNUITY	0.729308895
AMT_INCOME_TOTAL	AMT_ANNUITY	0.404922843
AMT_CREDIT	AMT_GOODS_PRICE	0.979764225
REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	0.439030852
WORK_EX_CLIENT	AGE_CLIENT	0.307859892
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.332801246
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956614091
YEARS_REGISTRATION	AGE_CLIENT	0.240583284

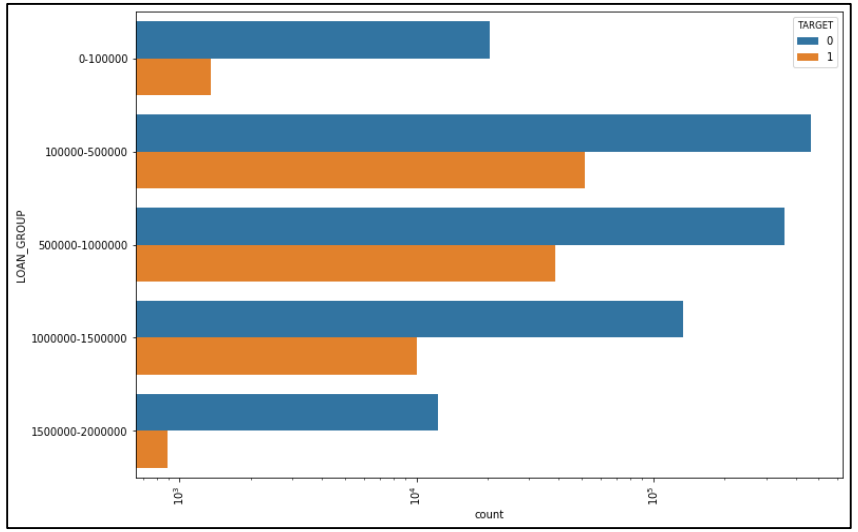
Step 6: Analysis for Categorical Variables





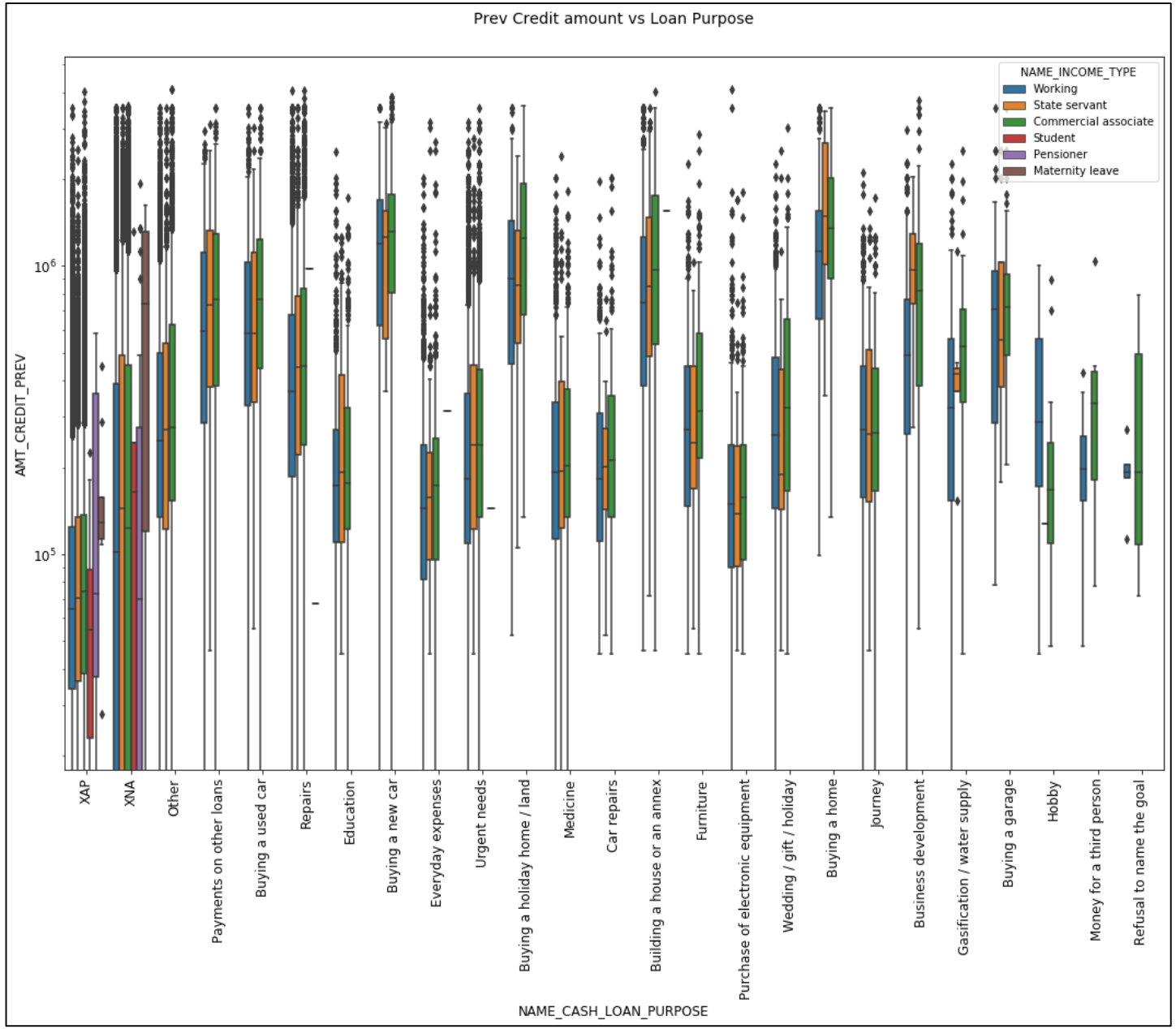
Clients with income between Rs 1 Lac and Rs 2 Lac have the highest number of refused offers and equally high number of cancelled loans compared to other belonging to different income groups.
Those with payment difficulties are the highest in number for income group between Rs 1 Lac and Rs 2 Lac and lowest for income group between Rs 4 Lac and Rs 5 Lac.





Clients with highest loan requirement along with similar income group faced the same challenges of high refusal and cancellations of loan.

Data is quite similar in Income group and loan group when it is compared between two targets 0 and 1.



Loan for buying a new car, buying holiday home/land, buying house journey, payments on other loans are higher compared to other loan purposes. This highlights people are applying for loan when they need to purchase something new.

Students and pensioners income type people have very limited credit.

XNA and XAP have data collection highlights loans are applied by all category of clients and even the loan amount is for basic loan between Rs 1 Lac - Rs 2 Lac.

This also suggests that the data collection system of bank can be improved in order to get clearer understanding of loan disbursal.

Overall working category, state servants and commercial associates are the clients bank can focus on in terms of ensuring proper loan facility as they are the consumers of loan and can create income opportunity for bank.

Conclusion/ Suggestions

From above analysis, following needs to be considered for approving or rejecting the loan.

- Quite few applications were rejected because clients applied for consumers loans.
- Bank should focus on less number of working type clients as they are the highest defaulters of loan.
- Bank can avoid giving loans for repairs as it is show clients applying for repairs loan have difficulty in repaying the loan.
- Also focus on clients living with parents as they have less difficulties in making payments on time.
- High number of loan applications are rejected for clients who have applied for loan between Rs 1 Lac and Rs 5 Lac.
- Correlation between amount credit and amount good price is high and similar in both target clients of zero and one.
- Bank can focus on giving more loan to people who have no children and staying with parents as they have high repayment and less defaults.

THANK YOU