CREDIT EDA-ASSIGNMENT

By:- Taiba Khalid

OBJECTIVE

The main objective of this case study is to identify the pattern of a client and the difficulties he is facing in paying the loans, which will ensure that the consumers capable of repaying the loans are not rejected.

EDA is the technique used to identify such customers. 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

DATA CLEANING PROCESS

- Due to threshold theory more than 40% missing value not gives the accurate results, that's why I dropped the columns where more than 40% missing values

- DROPPING FLAG COLUMNS RELATED TO SINCE THEY ARE THE INFORMATION PROVIDED ARE NOT LIKE PHONE NUMBER, ADDRESS, DOCUMENTS

- REGION RATING CAN ALSO BE DROPPED BECAUSE THE REGION IN NOT RELEVANT VALUE FOR EVALUAION

- SOCIAL CIRCLE COLUMNS CAN ALSO BE DROPED

- ALL THE FLAG DATA IS BINARY VALUE SO WE CAN DROP IT.

- Though has 31% missing value it gives some useful insights. Imputing the missing values with mode will distort the data. So will leave it as it is.

- In CNT_CHILDREN columns there were outliers /aberrations found Ex. 19 children for client in 20-30 age range. clients with more than 10 children are mostly from 30-50 age range.

- The people who are showed as employed for 365243 days (1000+ Years) are actually in Pensioner and Unemployed income category. These are outliers.

MISSING VALUE

APPLICATION DATA

COMMONAREA_MEDI	69.696426
COMMONAREA_AVG	69.696426
COMMONAREA_MODE	69.696426
NONLIVINGAPARTMENTS_MODE	69.257147
NONLIVINGAPARTMENTS_AVG	69.257147
NONLIVINGAPARTMENTS_MEDI	69.257147
LIVINGAPARTMENTS_MODE	68.250838
LIVINGAPARTMENTS_AVG	68.250838
LIVINGAPARTMENTS_MEDI	68.250838
FONDKAPREMONT_MODE	68.195928
FLOORSMIN_AVG	67.679326
FLOORSMIN_MODE	67.679326
FLOORSMIN_MEDI	67.679326
YEARS_BUILD_MEDI	66.331230
YEARS_BUILD_MODE	66.331230
YEARS_BUILD_AVG	66.331230
OWN_CAR_AGE	65.992806
LANDAREA_MEDI	59.250087
LANDAREA_MODE	59.250087
LANDAREA_AVG	59.250087
BASEMENTAREA_MEDI	58.405146
BASEMENTAREA_AVG	58.405146
BASEMENTAREA_MODE	58.405146
EXT_SOURCE_1	56.370116

LIVINGAREA_MEDI	50.179838
HOUSETYPE_MODE	50.126069
FLOORSMAX_MODE	49.703598
FLOORSMAX_MEDI	49.703598
FLOORSMAX_AVG	49.703598
YEARS_BEGINEXPLUATATION_MODE	48.721943
YEARS_BEGINEXPLUATATION_MEDI	48.721943
YEARS_BEGINEXPLUATATION_AVG	48.721943
TOTALAREA_MODE	48.238959
EMERGENCYSTATE_MODE	47.355917
OCCUPATION_TYPE	31.289712
EXT_SOURCE_3	19.813530
AMT_REQ_CREDIT_BUREAU_HOUR	13.406995
AMT_REQ_CREDIT_BUREAU_DAY	13.406995
AMT_REQ_CREDIT_BUREAU_WEEK	13.406995
AMT_REQ_CREDIT_BUREAU_MON	13.406995
AMT_REQ_CREDIT_BUREAU_QRT	13.406995
AMT_REQ_CREDIT_BUREAU_YEAR	13.406995
NAME_TYPE_SUITE	0.412385

PREVIOUS DATA

|	
NAME_TYPE_SUITE	49.110719
NFLAG_INSURED_ON_APPROVAL	40.104452
DAYS_TERMINATION	40.104452
DAYS_LAST_DUE	40.104452
DAYS_LAST_DUE_1ST_VERSION	40.104452
DAYS_FIRST_DUE	40.104452
DAYS_FIRST_DRAWING	40.104452
AMT_GOODS_PRICE	22.960354
AMT_ANNUITY	22.209616
CNT_PAYMENT	22.209432
PRODUCT_COMBINATION	0.021492
NAME_PRODUCT_TYPE	0.000092
NAME_YIELD_GROUP	0.000092
NAME_SELLER_INDUSTRY	0.000092
SELLERPLACE_AREA	0.000092
CHANNEL_TYPE	0.000092
NAME_GOODS_CATEGORY	0.000092
NAME_PORTFOLIO	0.000092
NAME_CLIENT_TYPE	0.000092
CODE_REJECT_REASON	0.000092
NAME_PAYMENT_TYPE	0.000092

UNIVARIATE OF APPLICATION DATA
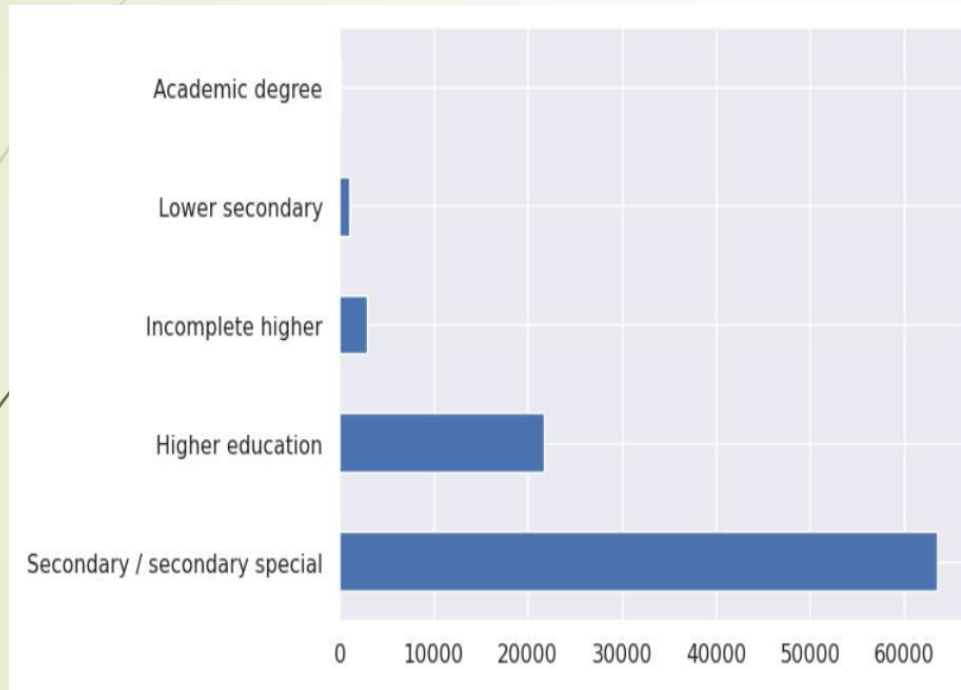
Observation :- People with higher secondary school are more in comparison to others
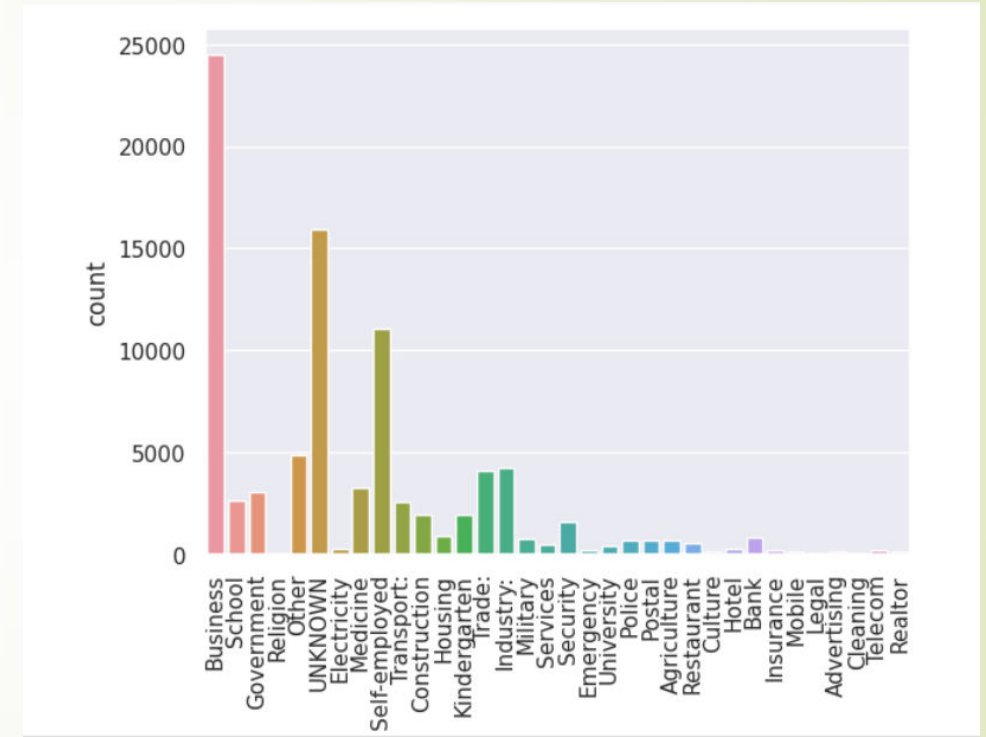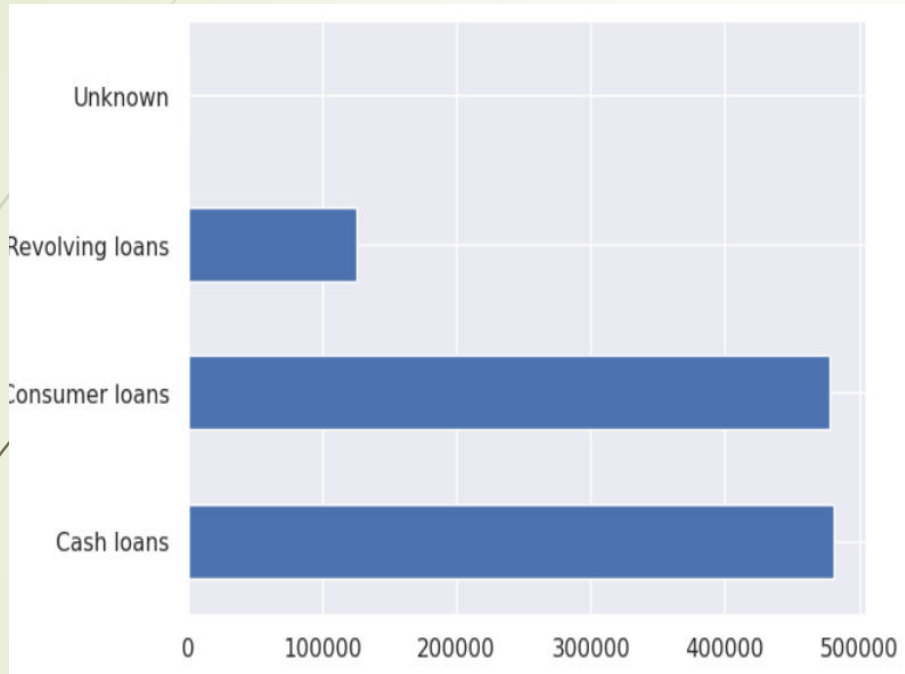

Observation :- Business , Unknown and Self-Employed are the top most occupation

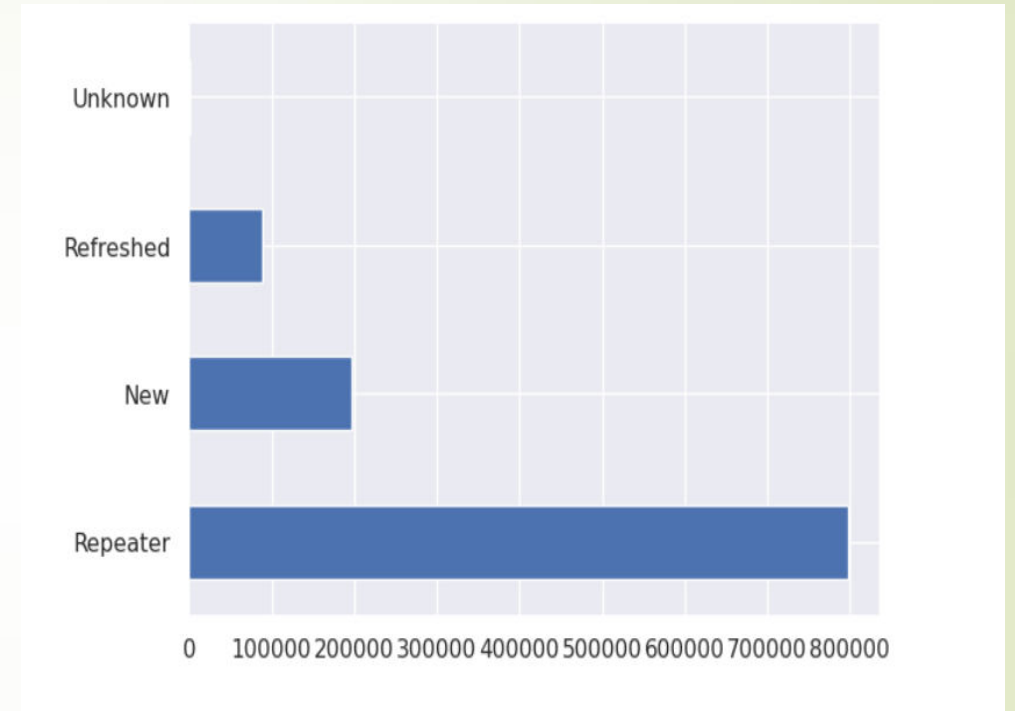# BIVARIATE OF APPLICATION DATA
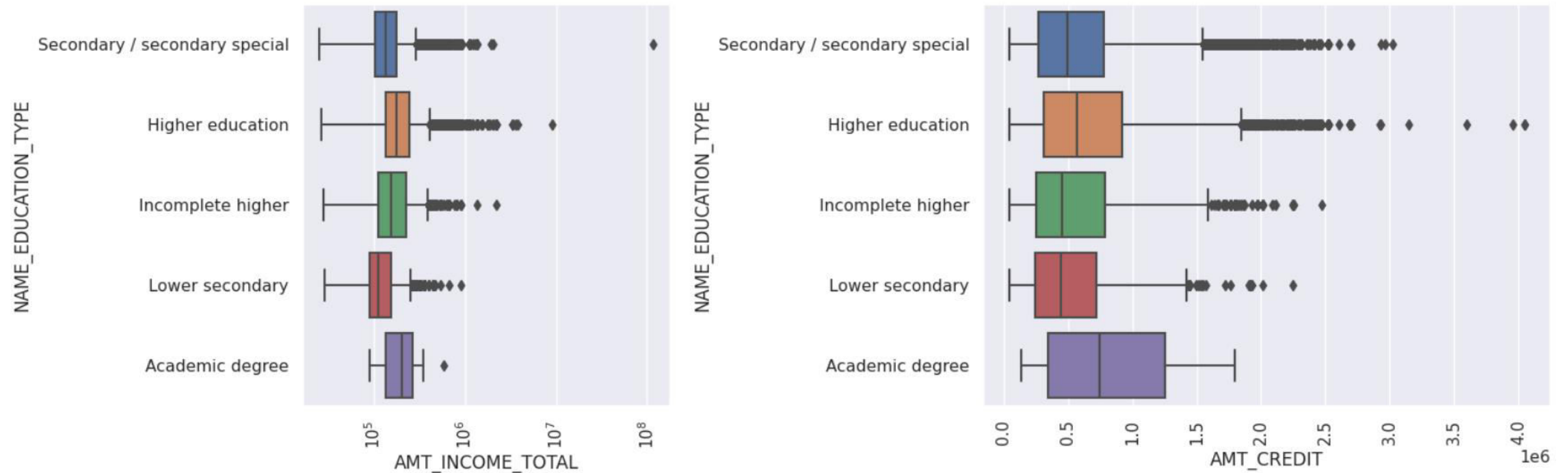
Here is seen that the people with Academic degree have more income as compare to other.
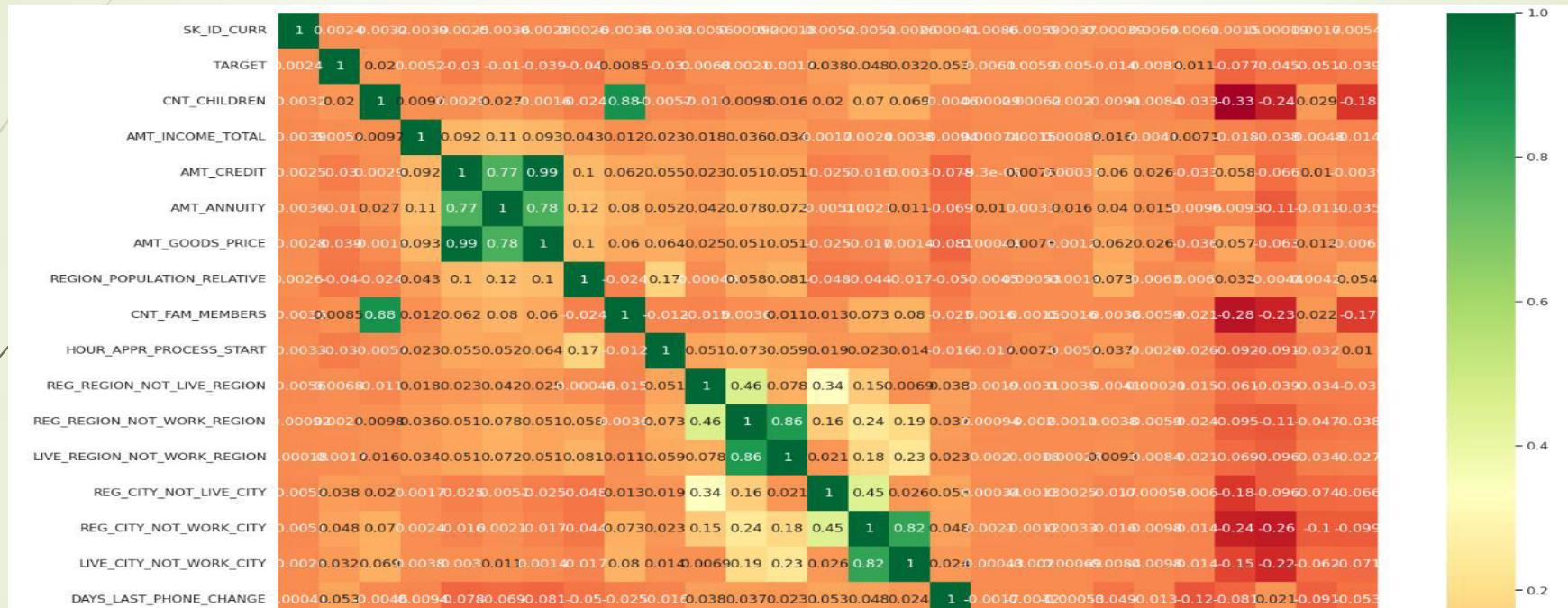
BIVARIATE OF APPLICATION DATA

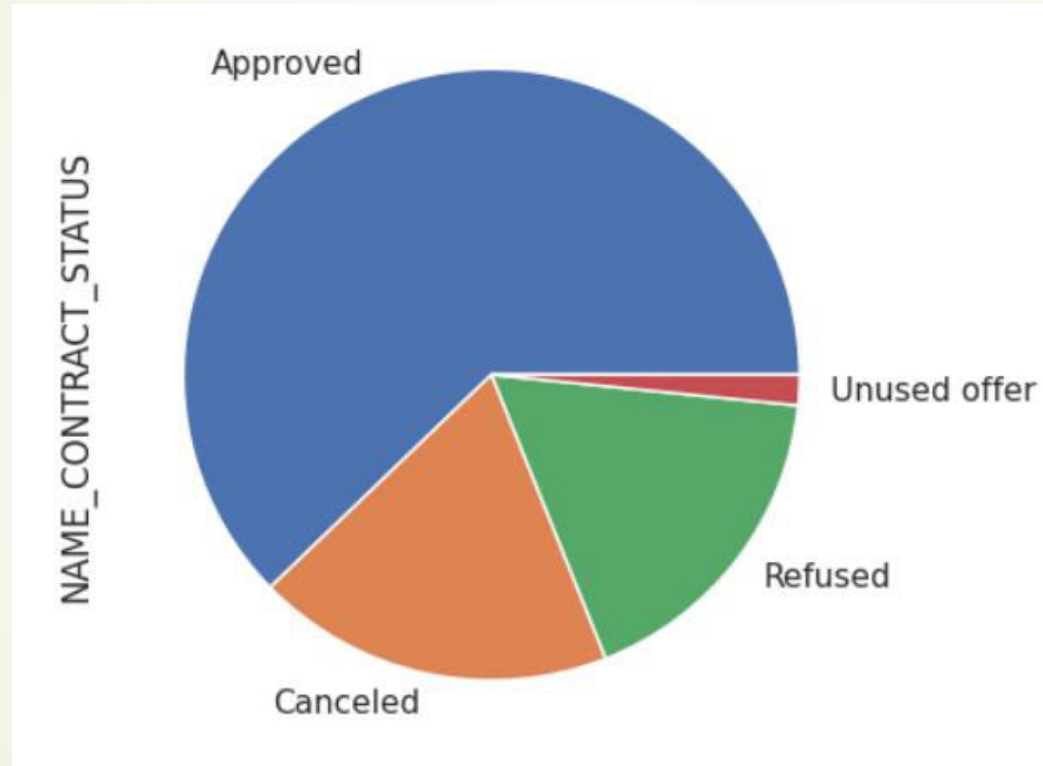Observation :- The CASH LOAN ARE MORE IN COMPARISON TO ANY OTHER TYOE OF LOANS

Observation :- Repeater clients are more in comparison to others

MULTIVARIATE ANALYSIS

Observation :- THERE IS A HIGH CORELATION BETWEEN AMT_CREDIT AND AMT GOODS PRICE AS PER THE HEAT PLOT. CNT_CHILDREN AND CNT_FAM_MEMBERS HAVE A HIGH CORELATION ABOUT 0.88

ANALYSIS OF PREVIOUS DATA

Observation :- The rate of APPROVAL is higher in comparison to any othrt type of status

UNIVARIATE OF PREVIOUS DATA
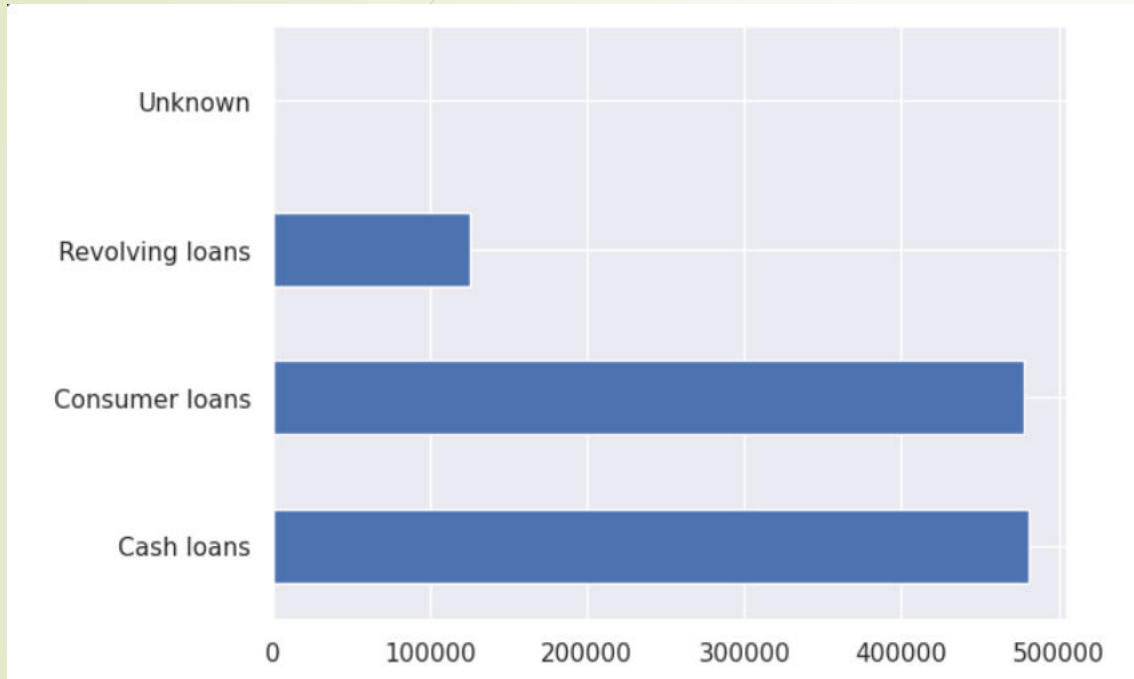
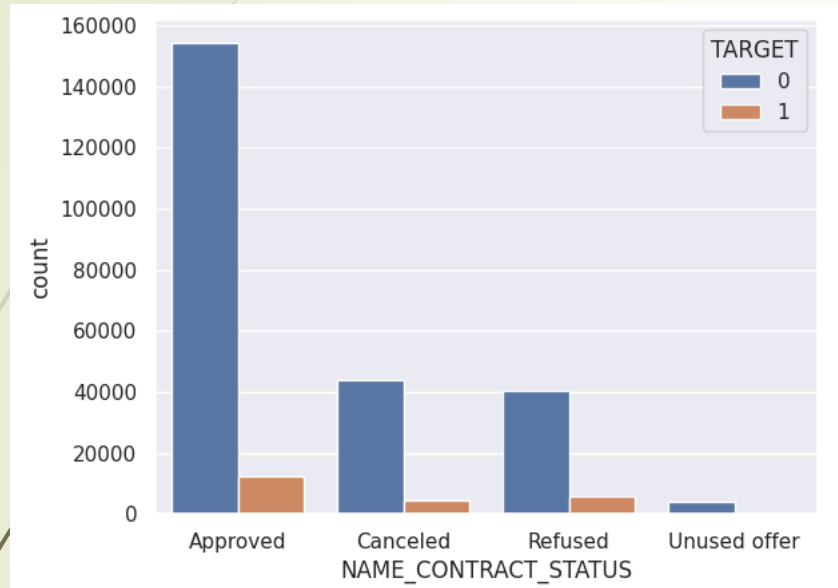Observation :- The CASH LOAN ARE MORE IN COMPARISON TO ANY OTHER TYOE OF LOANS
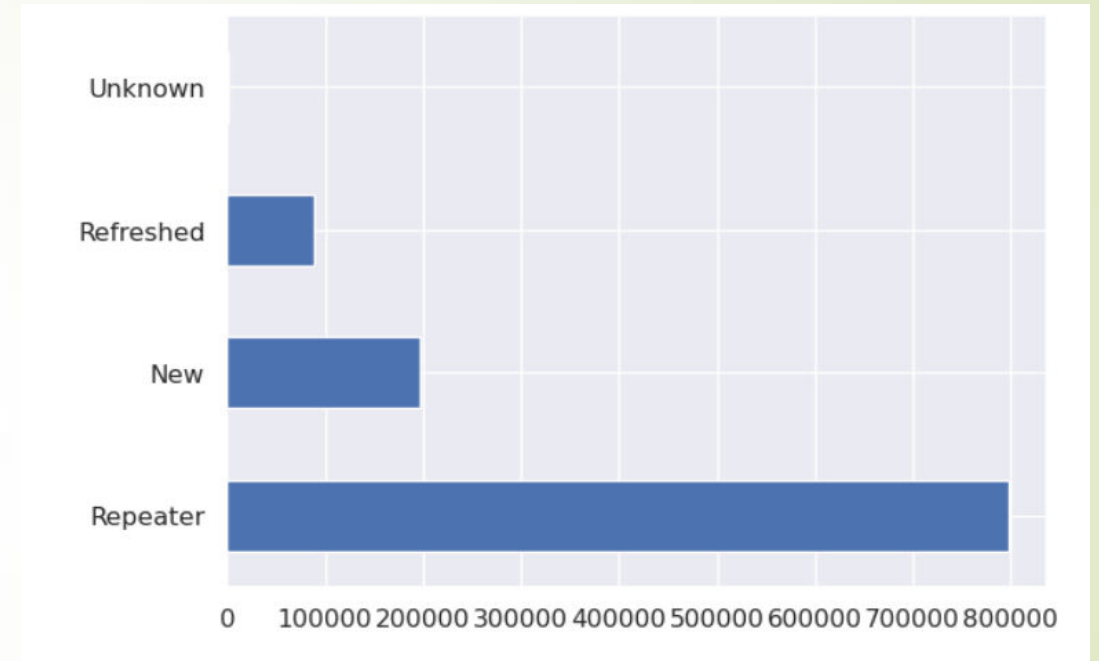
Observation :- Repeter clients are more in comparison to others
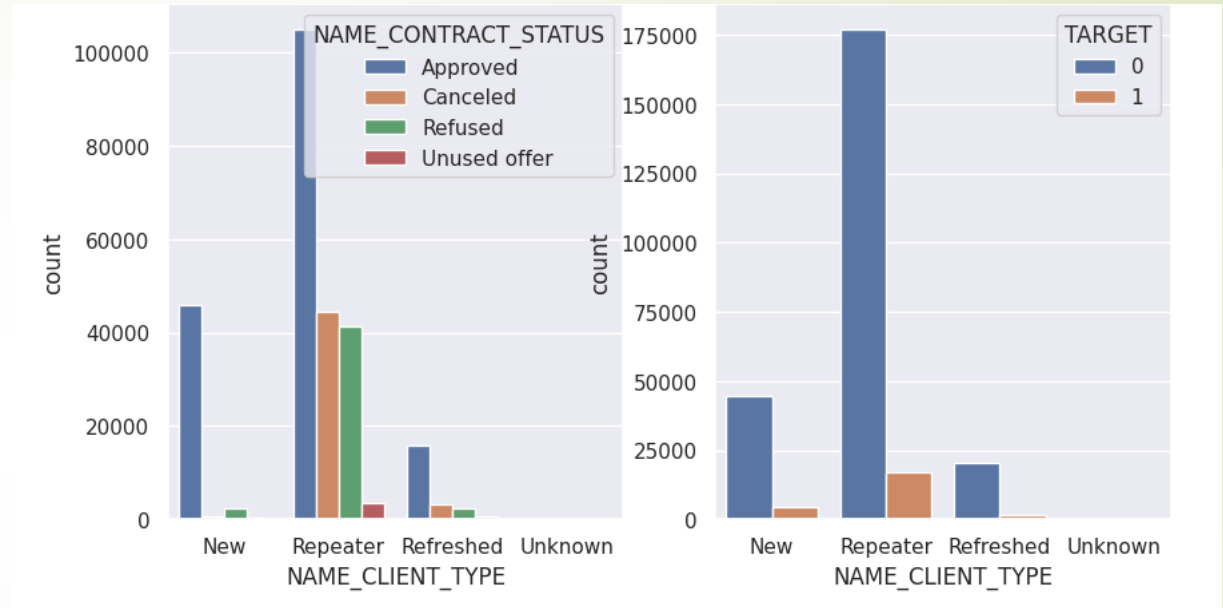
BIVARIATE OF PREVIOUS DATA
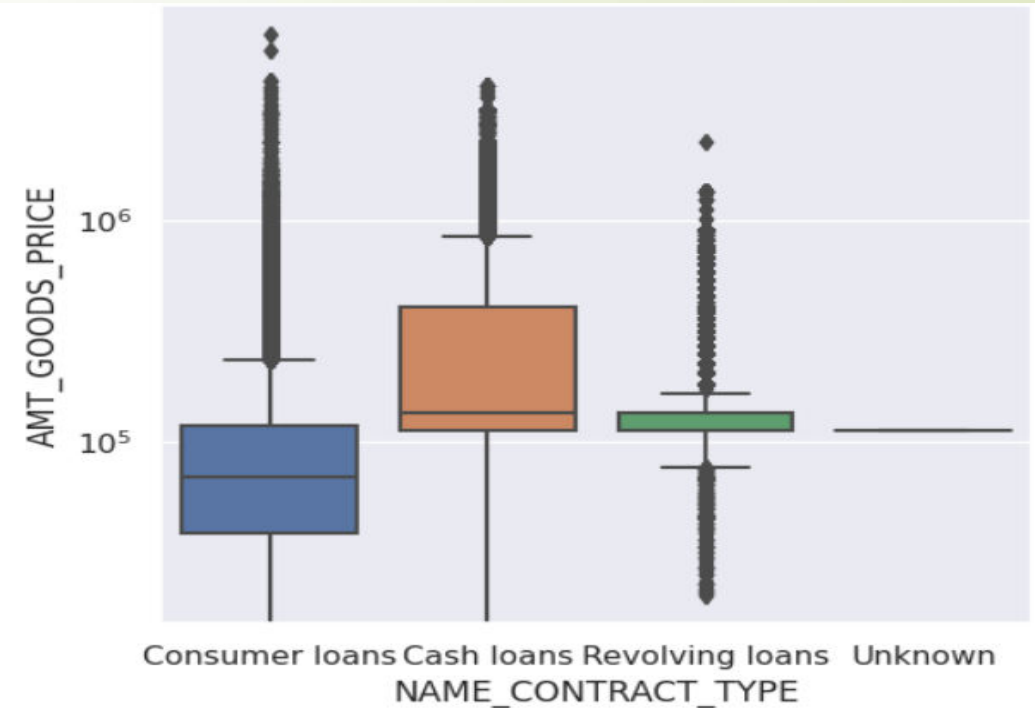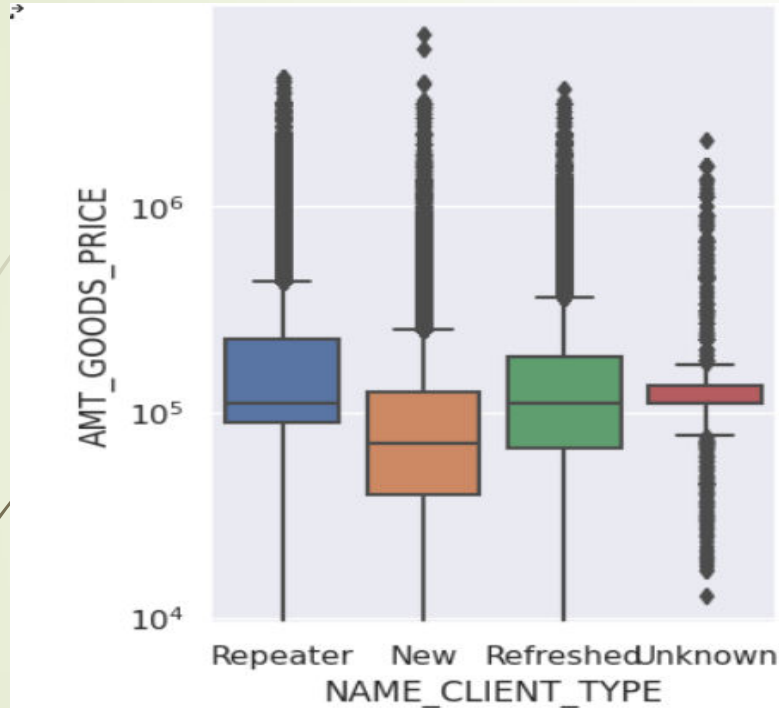
Observation :- Repeaters have the highest AMT_GOODS_PRICE CASH LOANS are also more in comparison to other loans
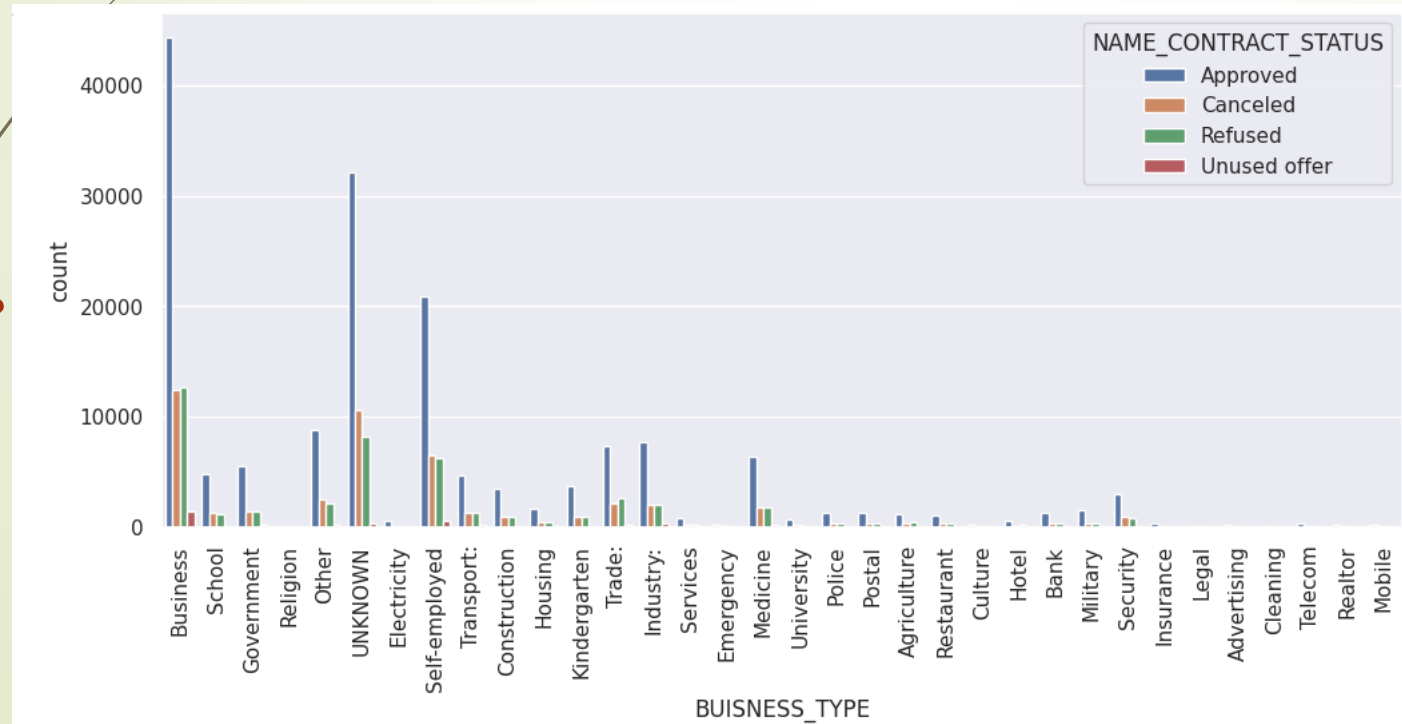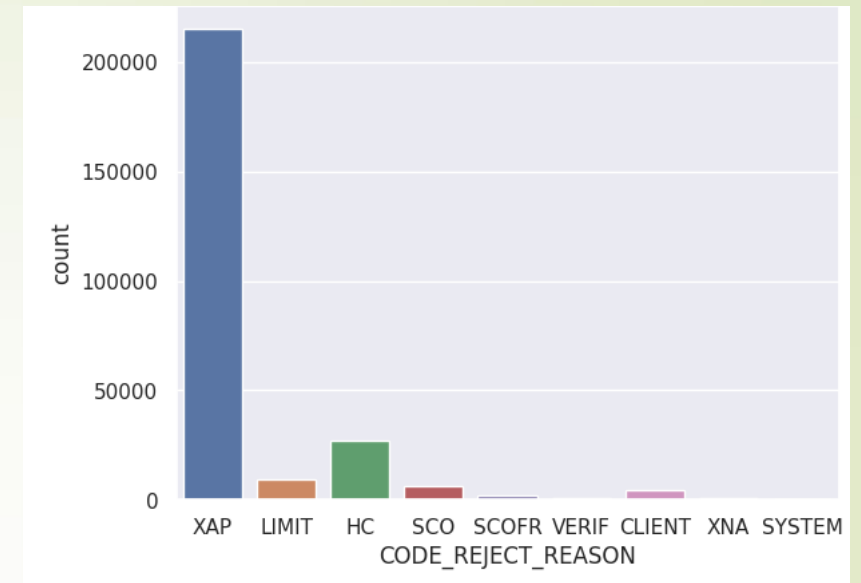
UNIVARIATE OF MERGE DATA

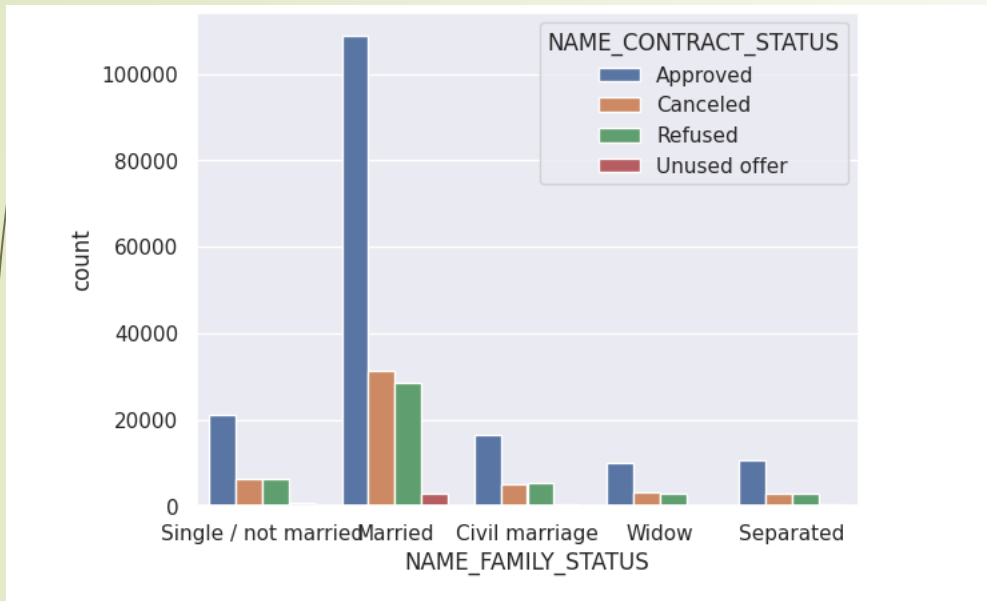Observation :- Approved status is high of Non Defaulters

Observation :- Repeater are more refused as compared to any other considering their Credit history
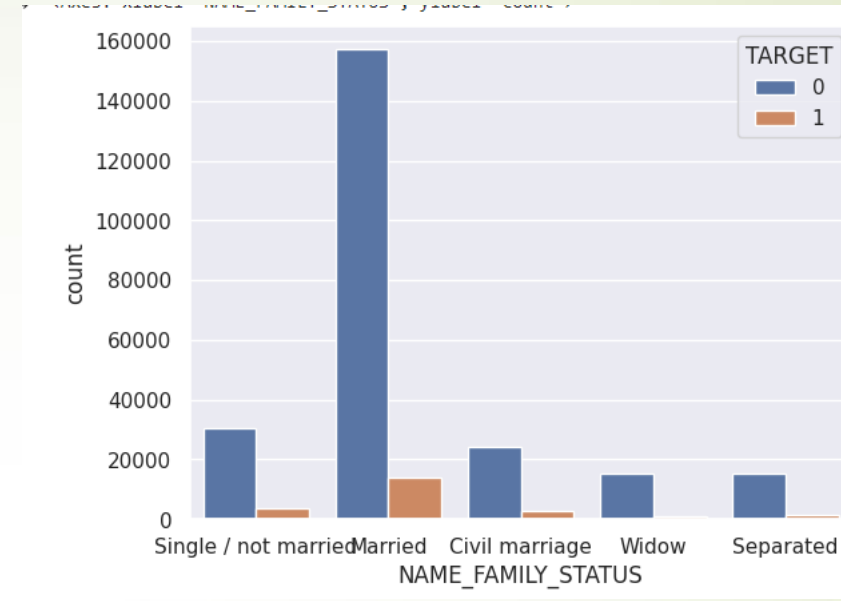
Observation :- XAP AN UNKNOW VARIABLE ARE HAS THE HIGHEST REJECTION RATE

Observation:- Business , Unknown and Self - Employed are the top 3 occupation where approval rate is more than any other occupation

Observation :- Married people are more likely to get Loan approved in comparison to any other marital status of people.

Observation :- Married people are more likely to get Loan approved in comparison to any other marital status of people.
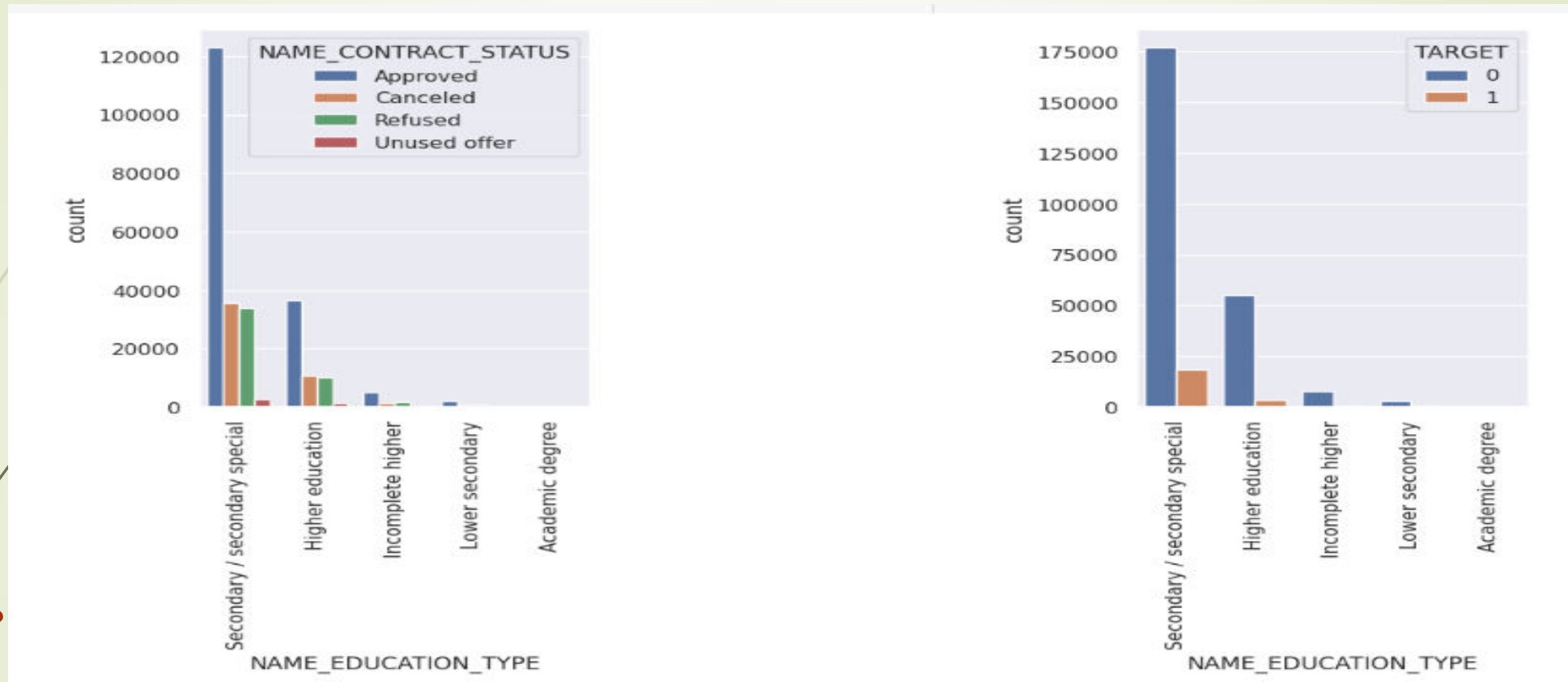
According to the both graphs secondary Education type has the least number of defaulters and the highest number of approval rate, so as per Education background this will be best target audience

FINAL OBSERVATION

Target variable for Application dataset - "TARGET"
Target variable for Previous dataset - "NAME_CONTRACT_STATUS"

1) The rate of defaulters are less in the range of 40-60 are good target audience.

2) Laborers , Core and Sales Staff is the occupation type that has the loan approved and has the highest non defaulter rate.

3) Married people are more likely to get loan approved in compression to any other Marital Status of the people so this is
also a good target audience .

4) Secondary Education has the Highest Approval rate ,although the Income of Academic degree holder are more in
comparison Secondary education still the approval rate is more than Academic Degree holders.

Thank You