CREDIT
EDA - ASSIGNMENT

By:- Utkarsh Dixit

Index

Introduction

This assignment aims to give an idea of applying EDA in a real business scenario. In this assignment, apart from applying the EDA techniques we also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected

Business Objective

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments 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

Data Cleaning Approach

- ✓ Dropped columns with more than 40% missing values in both the data sets. For the remaining columns did missing value treatment. For ex. OCCUPATION_TYPE column, 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.
- ✓ Apart from the missing values there are many columns with XNA and XAP (Not available and Unknown). We can treat those values with mean /median / mode or keep it as it is, if the percentage is higher.
- ✓ There were columns with negative values for days (Birth, Employment, ID Publish). Converted them into positive and also into years for better analysis.
- ✓ 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

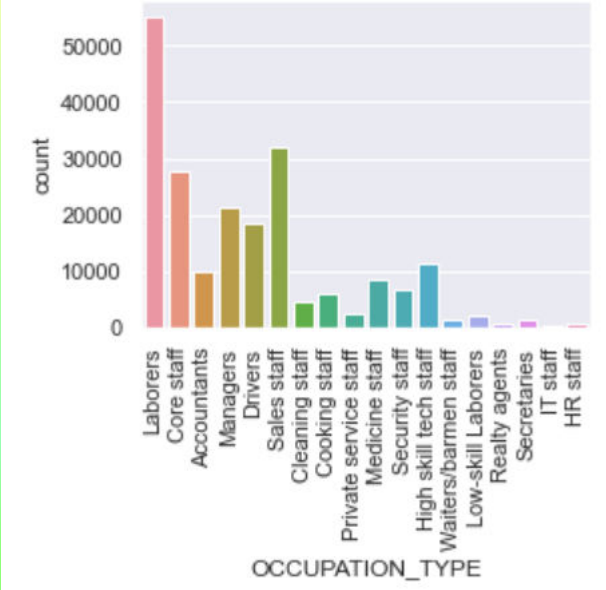
COMMONAREA_MEDI	69.87
COMMONAREA_AVG	69.87
COMMONAREA_MODE	69.87
NONLIVINGAPARTMENTS_MODE	69.43
NONLIVINGAPARTMENTS_AVG	69.43
NONLIVINGAPARTMENTS_MEDI	69.43
FONDKAPREMONT_MODE	68.39
LIVINGAPARTMENTS_MODE	68.35
LIVINGAPARTMENTS_AVG	68.35
LIVINGAPARTMENTS_MEDI	68.35
FLOORSMIN_AVG	67.85
FLOORSMIN_MODE	67.85
FLOORSMIN_MEDI	67.85
YEARS_BUILD_MEDI	66.50
YEARS_BUILD_MODE	66.50
YEARS_BUILD_AVG	66.50
OWN_CAR_AGE	65.99
LANDAREA_MEDI	59.38
LANDAREA_MODE	59.38
LANDAREA_AVG	59.38
BASEMENTAREA_MEDI	58.52
BASEMENTAREA_AVG	58.52
BASEMENTAREA_MODE	58.52
EXT_SOURCE_1	56.38
NONLIVINGAREA_MODE	55.18
NONLIVINGAREA_AVG	55.18
NONLIVINGAREA_MEDI	55.18
ELEVATORS_MEDI	53.30
ELEVATORS_AVG	53.30
ELEVATORS_MODE	53.30
WALLSMATERIAL_MODE	50.84
APARTMENTS_MEDI	50.75
APARTMENTS_AVG	50.75
APARTMENTS_MODE	50.75
ENTRANCES_MEDI	50.35
ENTRANCES_AVG	50.35
ENTRANCES_MODE	50.35
LIVINGAREA_AVG	50.19
LIVINGAREA_MODE	50.19
LIVINGAREA_MEDI	50.19
HOUSETYPE_MODE	50.18
FLOORSMAX_MODE	49.76
FLOORSMAX_MEDI	49.76
FLOORSMAX_AVG	49.76
YEARS_BEGINEXPLUATATION_MODE	48.78
YEARS_BEGINEXPLUATATION_MEDI	48.78
YEARS_BEGINEXPLUATATION_AVG	48.78
TOTALAREA_MODE	48.27
EMERGENCYSTATE_MODE	47.40
OCCUPATION_TYPE	31.35
EXT_SOURCE_3	19.83
AMT_REQ_CREDIT_BUREAU_HOUR	13.50
AMT_REQ_CREDIT_BUREAU_DAY	13.50
AMT_REQ_CREDIT_BUREAU_WEEK	13.50
AMT_REQ_CREDIT_BUREAU_MON	13.50

Previous Data

RATE_INTEREST_PRIVILEGED	99.64
RATE_INTEREST_PRIMARY	99.64
AMT_DOWN_PAYMENT	53.64
RATE_DOWN_PAYMENT	53.64
NAME_TYPE_SUITE	49.12
NFLAG_INSURED_ON_APPROVAL	40.30
DAYS_TERMINATION	40.30
DAYS_LAST_DUE	40.30
DAYS_LAST_DUE_1ST_VERSION	40.30
DAYS_FIRST_DUE	40.30
DAYS_FIRST_DRAWING	40.30
AMT_GOODS_PRICE	23.08
AMT_ANNUITY	22.29
CNT_PAYMENT	22.29
PRODUCT_COMBINATION	0.02

MISSING VALUE TREATMENT

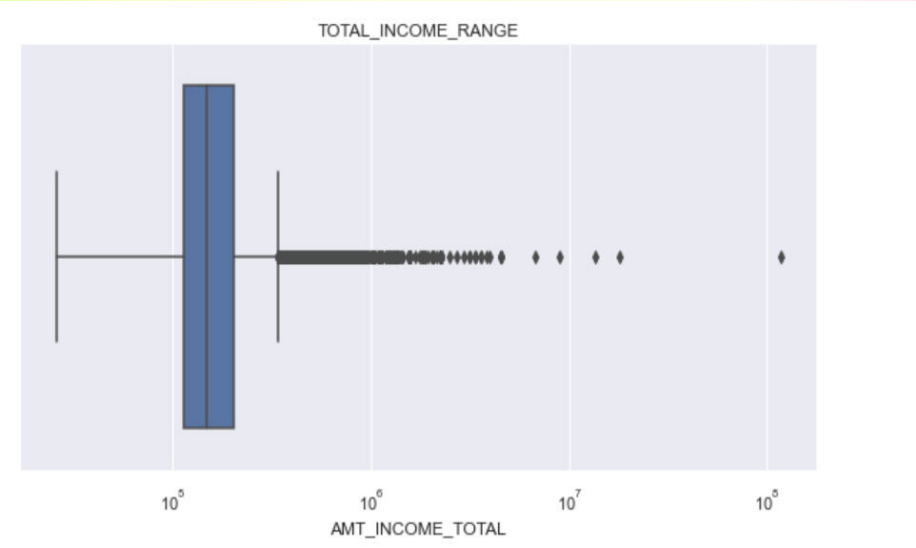
OCCUPATION_TYPE is categorical variable so we replace null with mode and here mode is Laborers so I have filled null



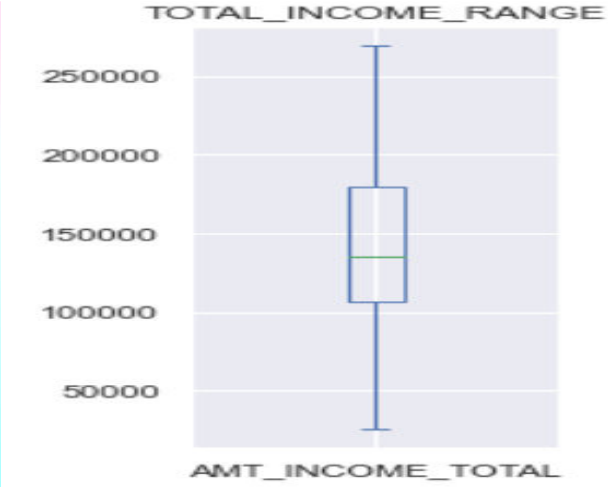
AMT_GOODS_PRICE is a continuous variable means numerical and there is many outliers in data set so we replaced with median not mean

HANDLING OUTLIER

- In this graph we can see there are lot of outliers
- After checking the mean and median there was a very huge difference.
- Hence the missing values will be filled by median

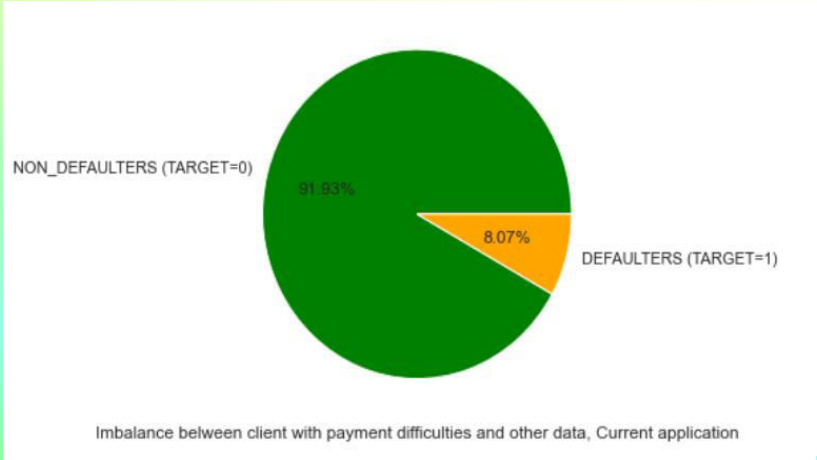


- In this case the outlier have be treated by taking values which are less than 90 percentile.
- Similarly we can treated for other column as well

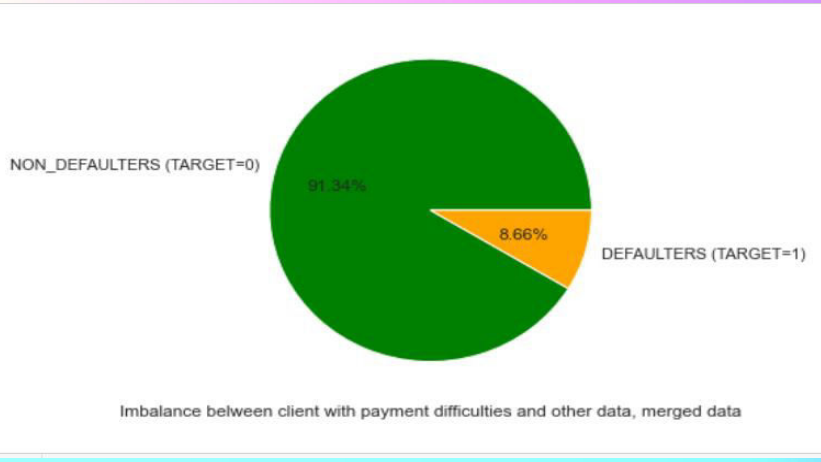


METHODOLOGY

- Checked data imbalance in the Target variable, we found 11.4% imbalance data. Due to data imbalance separated the application data into 2 datasets, with Target 0 and with Target 1 and analyzed separately with the help of Pie chart and Count plots.
- Later merged Application Data set and Previous Application data sets on common column SK_ID_CURR . The difference in unique entries of SK_ID in current and previous application shows that there are duplicate values , It means the client have multiple loans histories. In the merged dataset checked imbalance. It was similar.



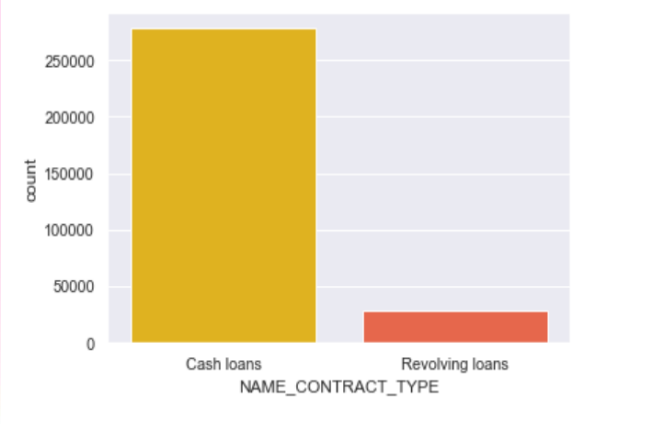
Imbalance between client with payment difficulties and other data, Current application

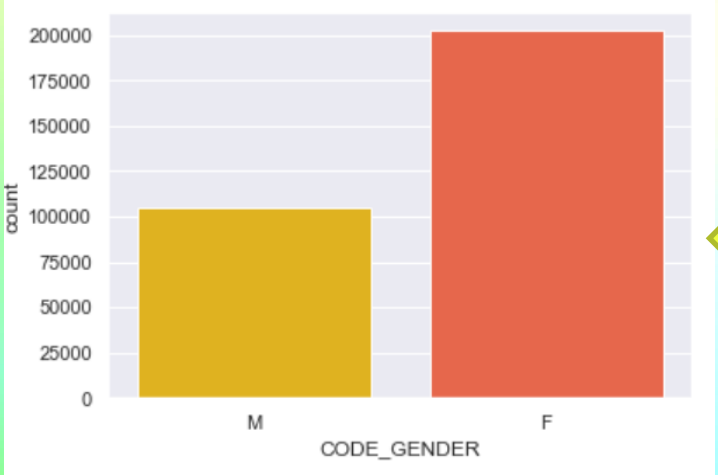


Imbalance between client with payment difficulties and other data, merged data

UNIVARIATE ANALYSIS

Cash Loans are more as compare to Revolving Loans





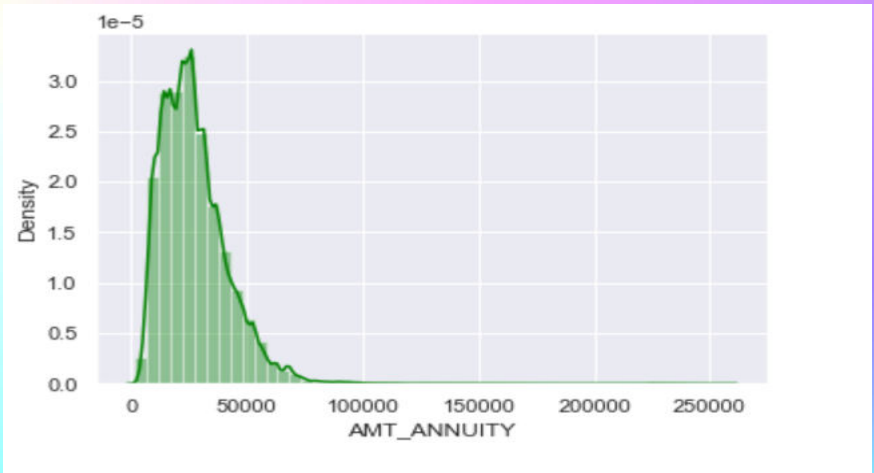
According to the data female is the most dominating gender as compare to male.

MALE POPULATION:- 100000 & FEMALE POPULATION:- 200000

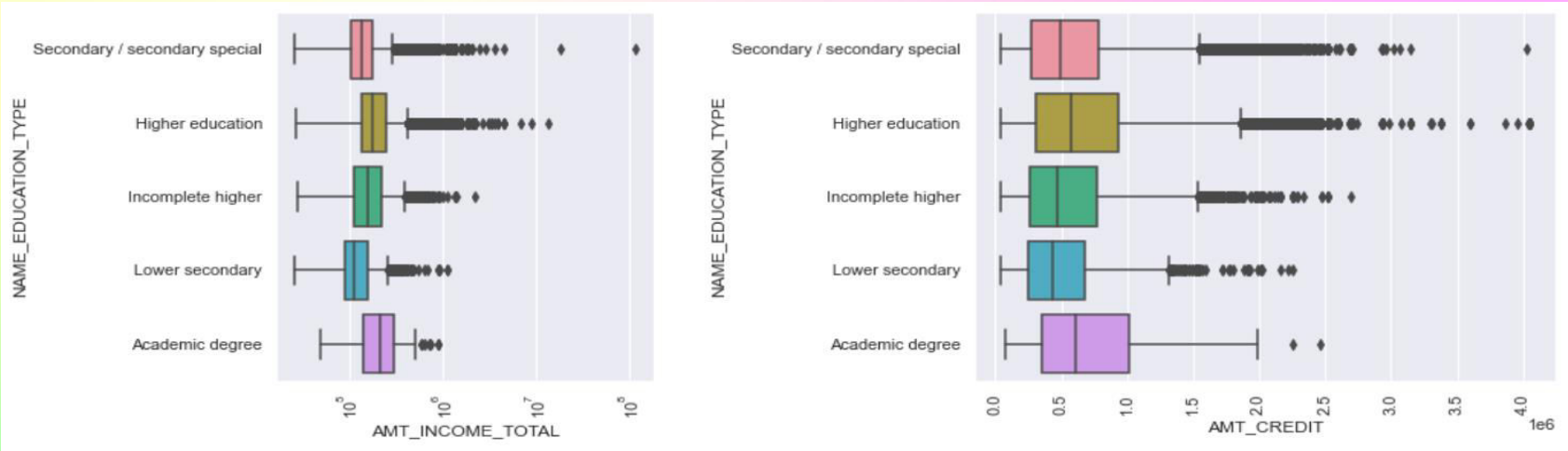
UNIVARIATE ANALYSIS NUMERICAL DATA

In AMT_GOODS_PRICE outliers are found which can be handled by taking median of the values

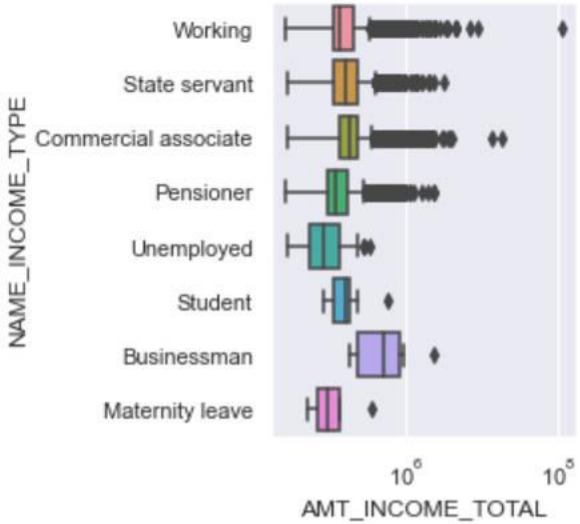
The AMT_ANNUITY has a positive Skew of 1.58 and reaches the Maximum density between 0-5000

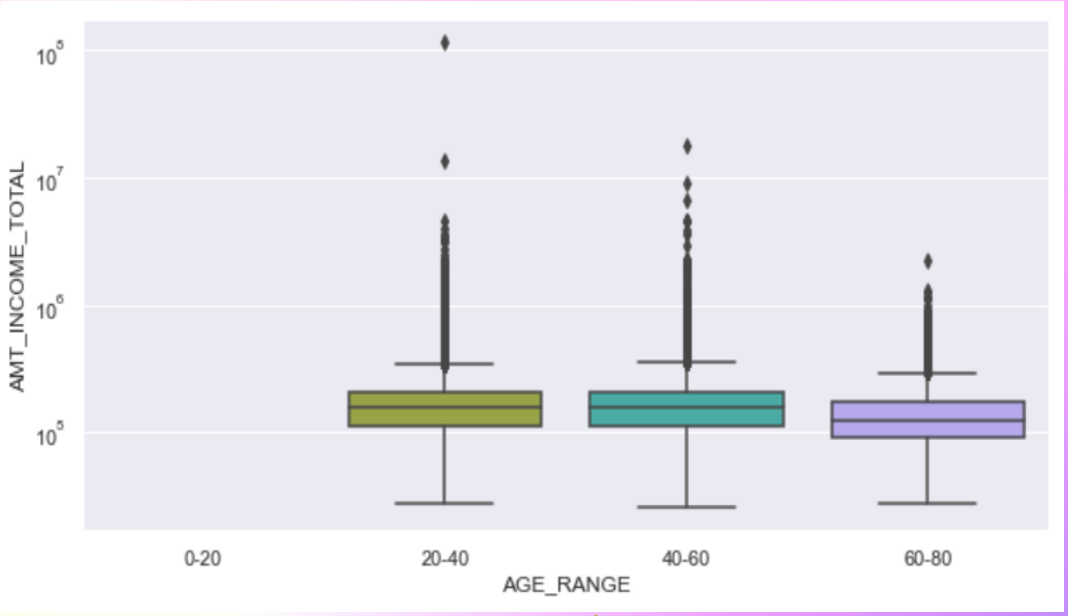


BIVARIATE ANALYSIS



NAME_EDUCATION_TYPE vs AMT_INCOME_TOTAL and AMT_CREDIT
The AMT_INCOME_TOTAL is highest for Academic degree and AMT_CREDIT is also highest for Academic degree

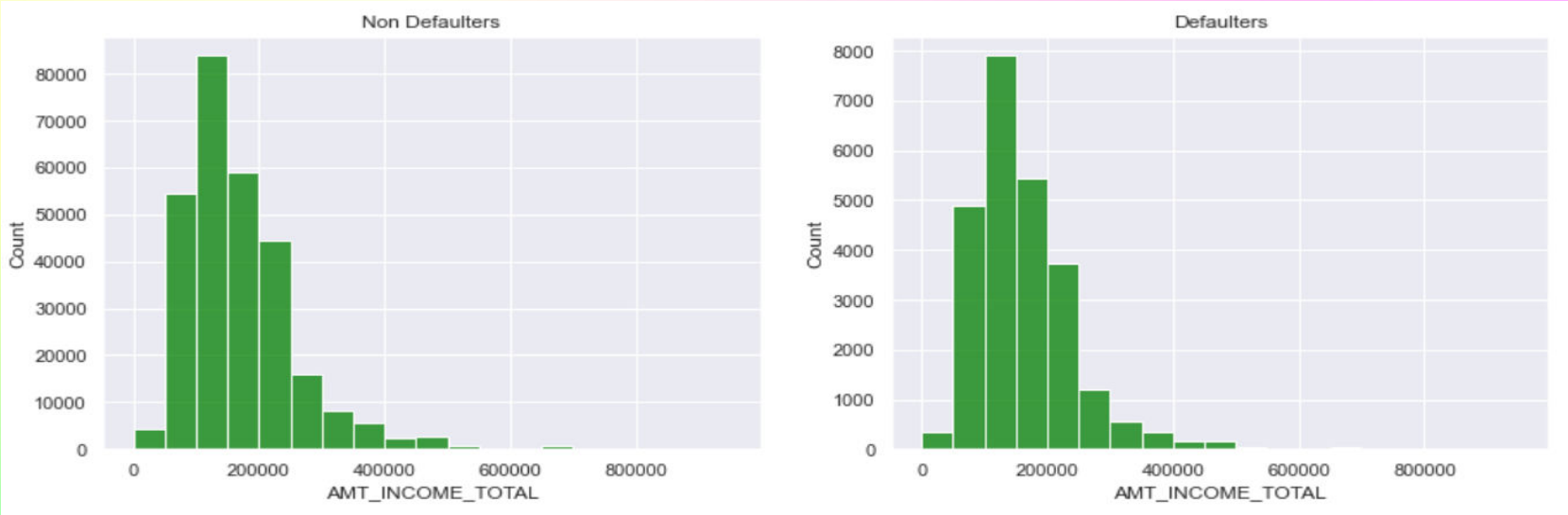




NAME_INCOME_TYPE vs NAME_ INCOME_TOTAL
The AMT_INCOME_TOTAL is Highest for business and least For Unemployed

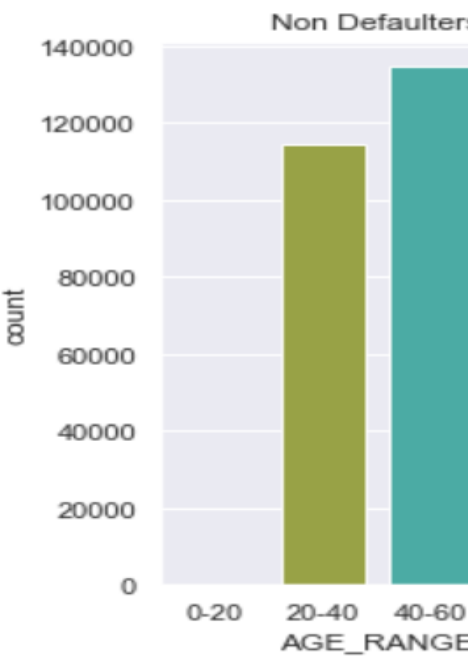
NAME_INCOME_TYPE vs AGE_RANGE
The AMT_INCOME_TOTAL is approximately same for the AGE_RANGE 20-40, 40-60

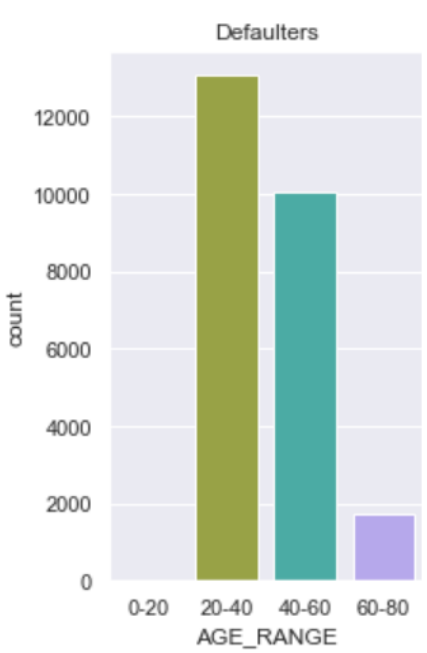
SEGMENTED ANALYSIS

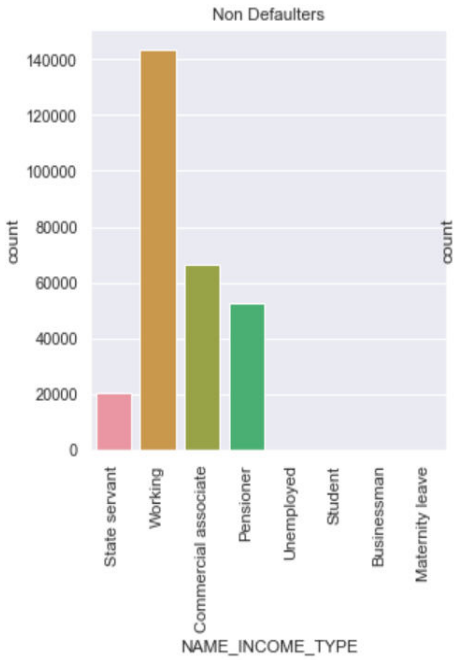


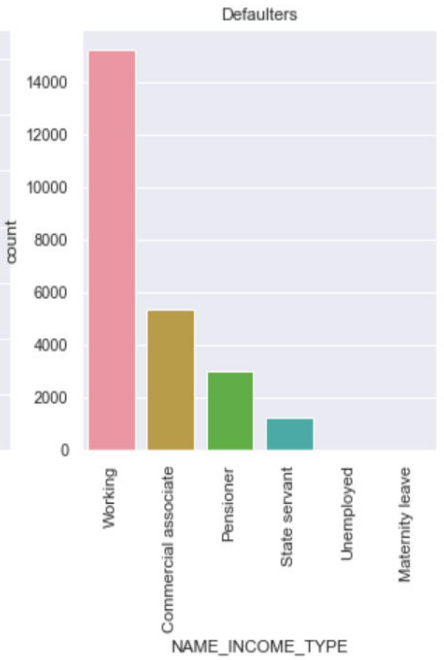
AMT_INCOME_TOTAL for Defaulters And Non-Defaulters.

- The both chart is similar.
- The Income is similar for both defaulters and Non- defaulters.
- The two Data income is maximum as 10000



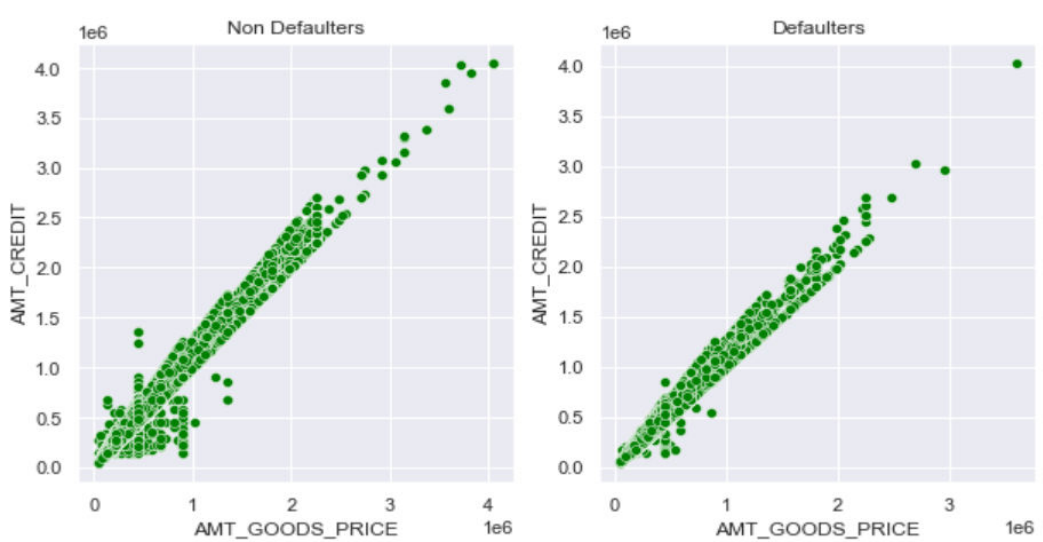






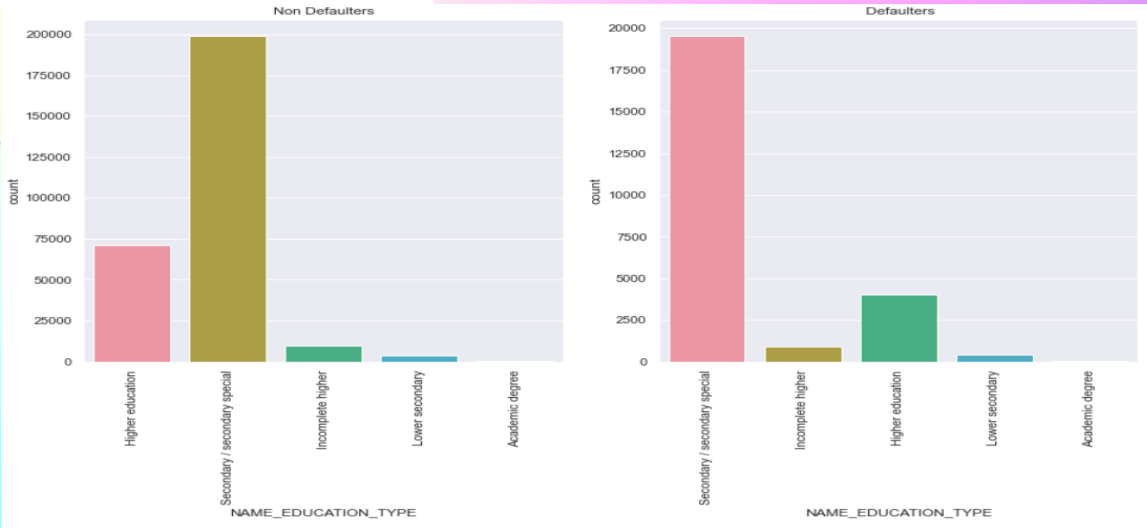
The age range of 20-40 have the highest rate of defaulters where as 40-60 have non defaulters

Students and Businessman are no payment default because they are no listed in defaulters category



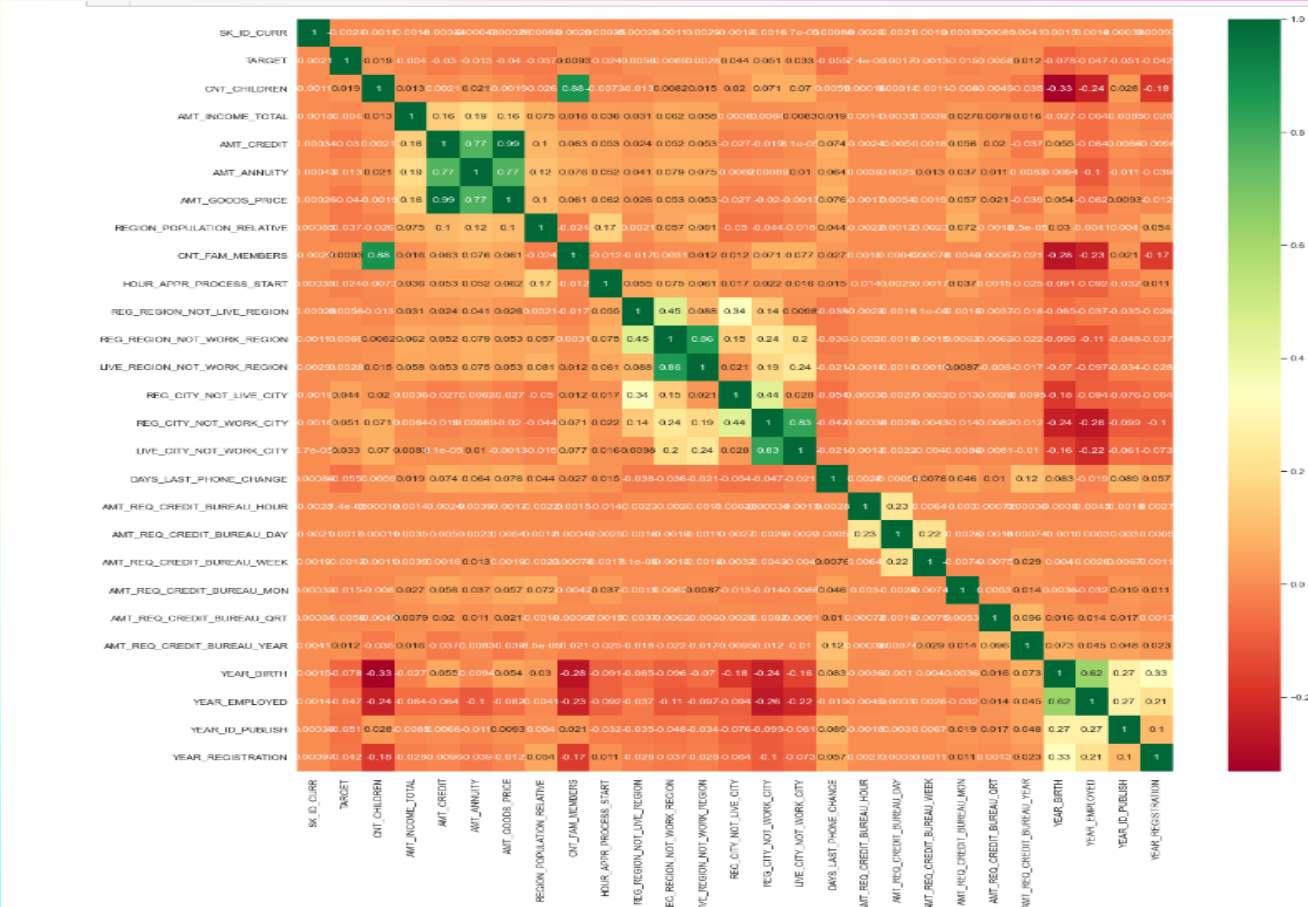
It is seen that the Goods Price Increases so the AMT_C also

In the defaulter rate of secondary / secondary special is more in compare to others and Academic degree holders have less defaulters same as non defaulters.



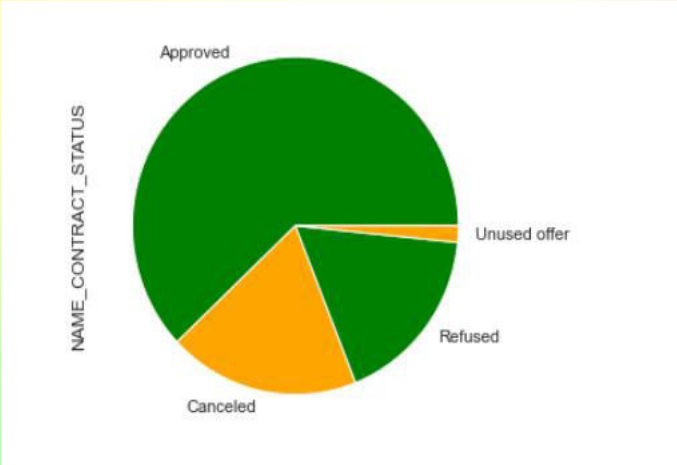
MULTIVARIATE ANALYSIS

There is a highly correlation between AMT_CREDIT and AMT_GOODS_PRICE as the heat plot. CNT_CHILDREN and CNT_FARM_MEMBERS have high correlation about 0.88



ANALYSIS OF PREVIOUS DATA

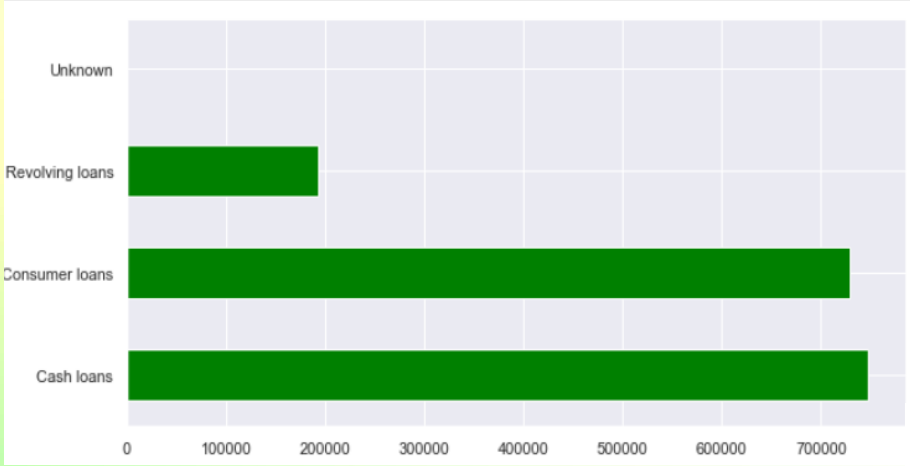
PREVIOUS DATA



As per the previous data csv file the Target column is NAME_CONTRACT_STATUS.

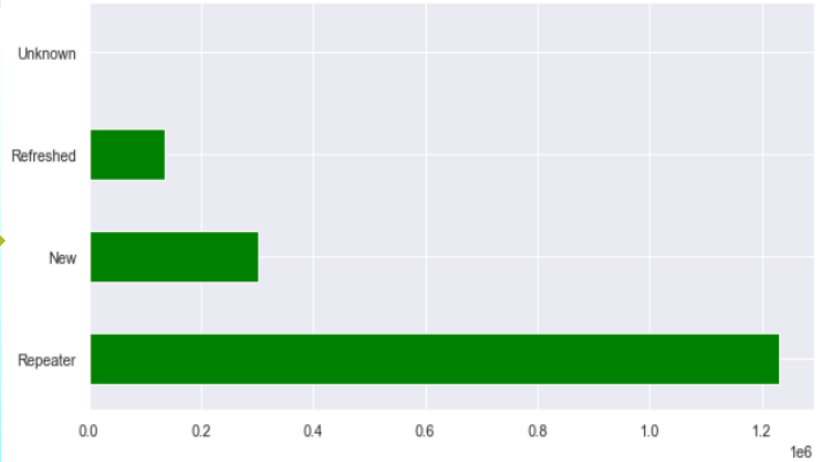
As per the pie chart The rate of APPROVAL is higher as compare to any other type of status.

UNIVARIATE ANALYSIS



The cash loan are more as compare to other.

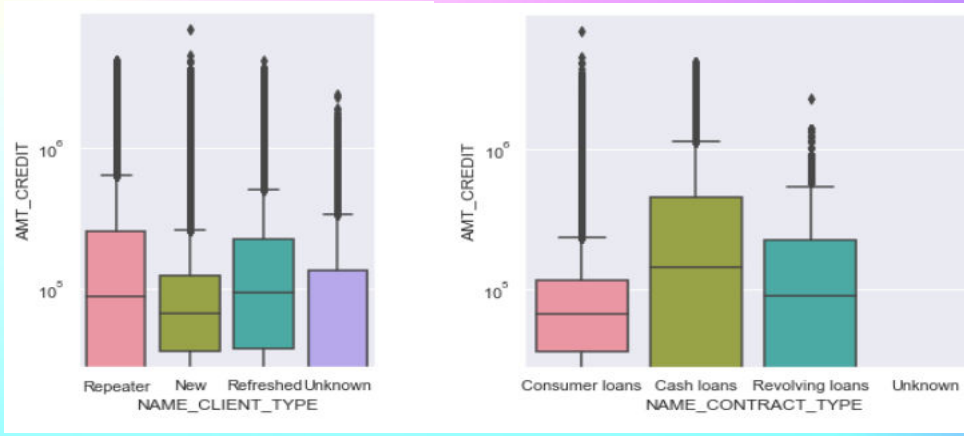
Repeater clients are more as compare to others.

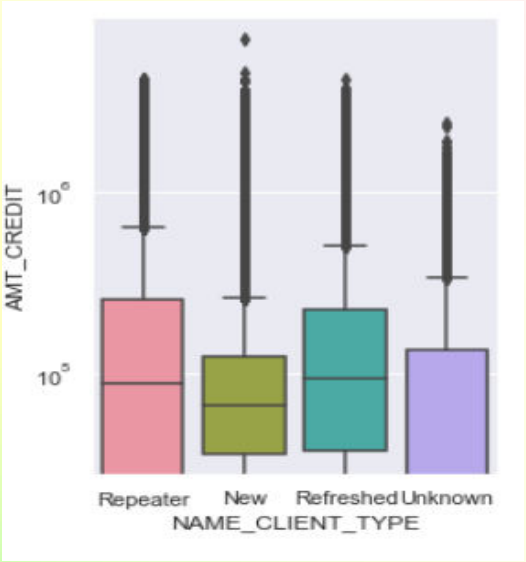


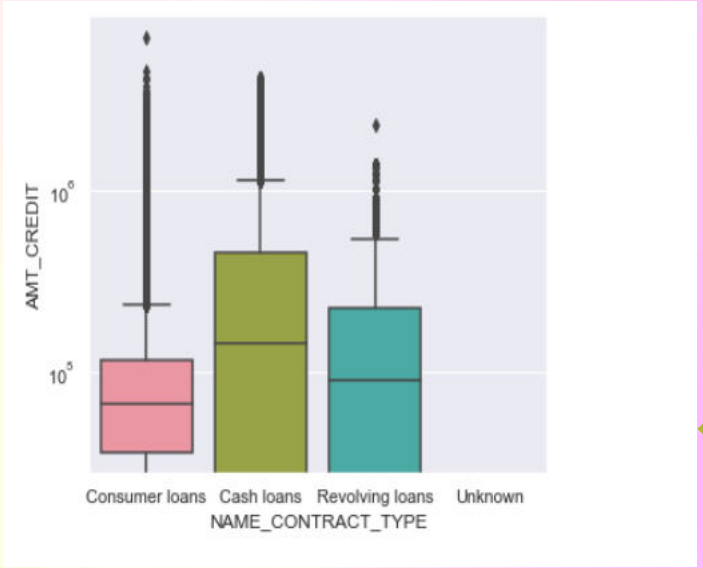
BIVARIATE ANALYSIS

Repeaters have the highest AMT_ GOODS_PRICE and Cash loans are also more compare to other loans

Repeaters and Cash loan have the highest value in comparison to AMT_CREDIT





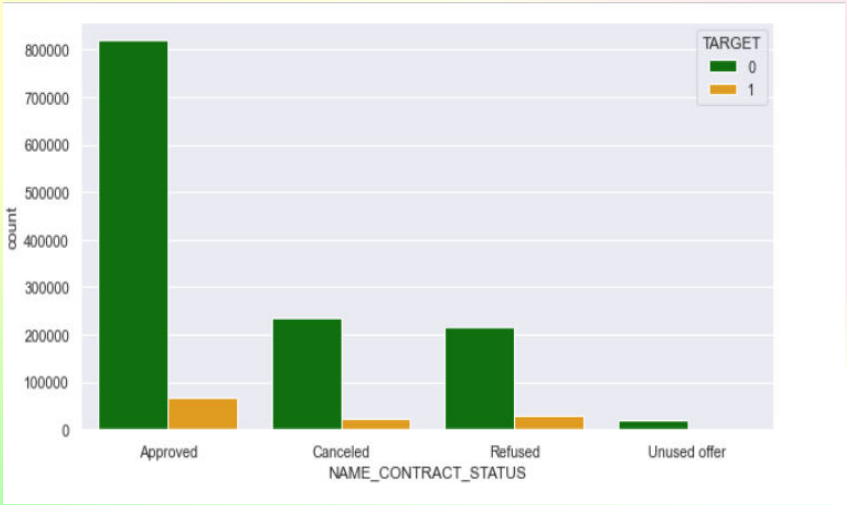


Repeaters and Cash loan have the highest value as compare to AMT_CREDIT

Repeater have high rate of APPROVAL then comes New Clients when compared it with Repeaters.

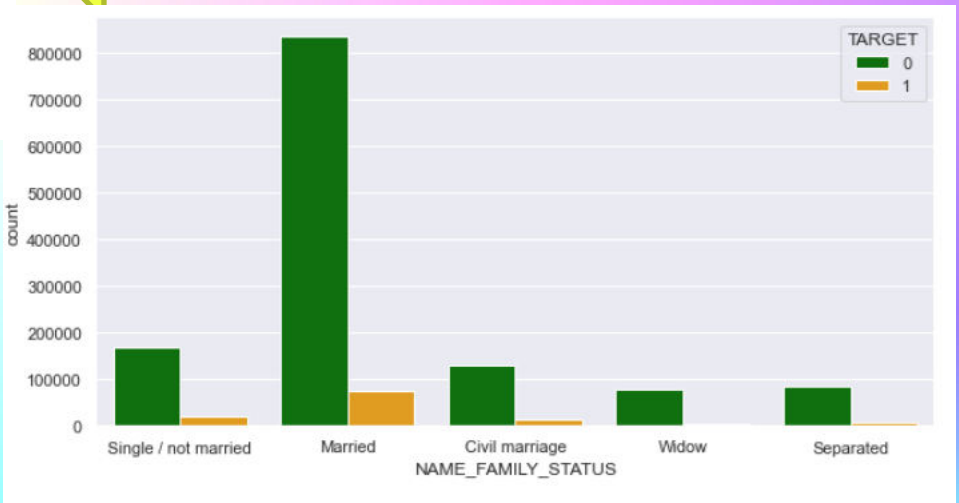
ANALYSIS OF MERGED DATA

MERGE DATA FRAME OF PREVIOUS AND CURRENT APPLICATION

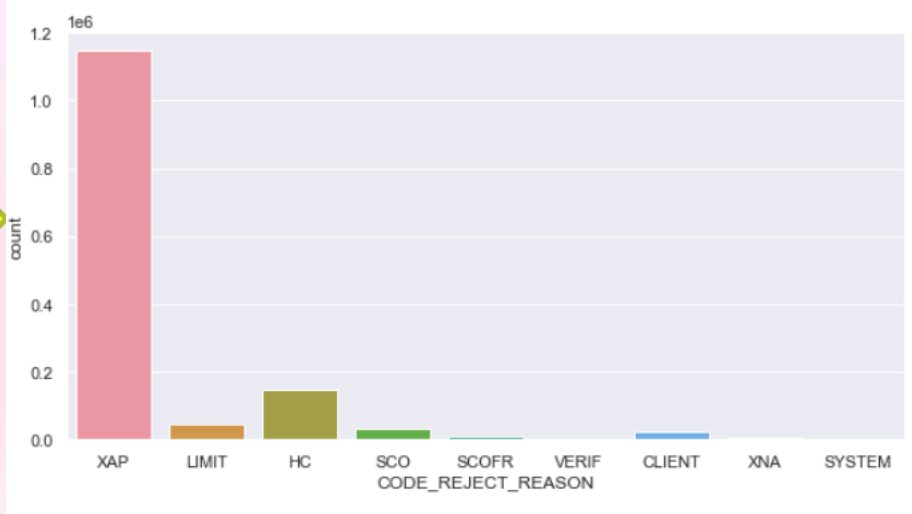


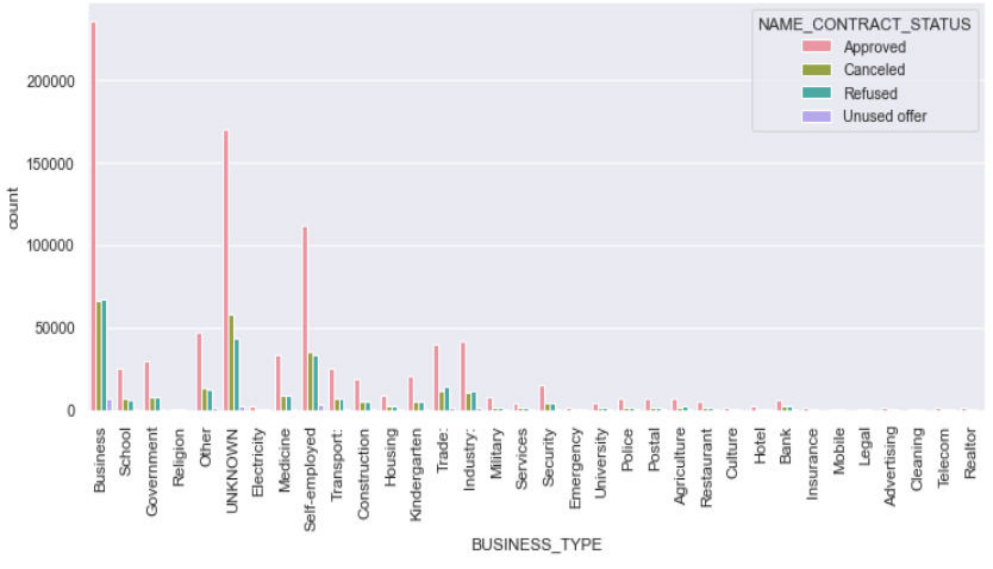
Here showing that Approved status is high of Non Defaulters

Married people are more likely to get Loan approved in compare to any other marital status of people.

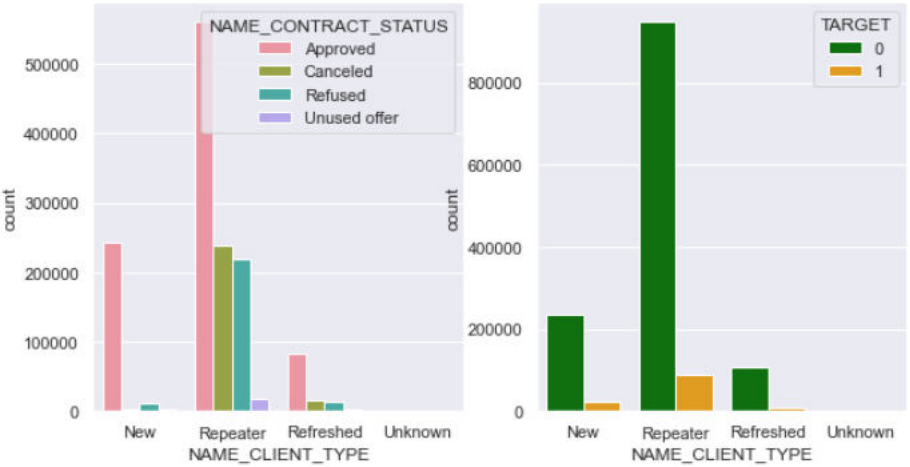


XAP or Unknown variable has the higher rejection rate.



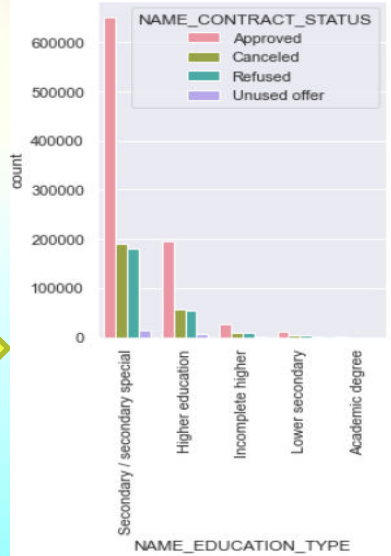


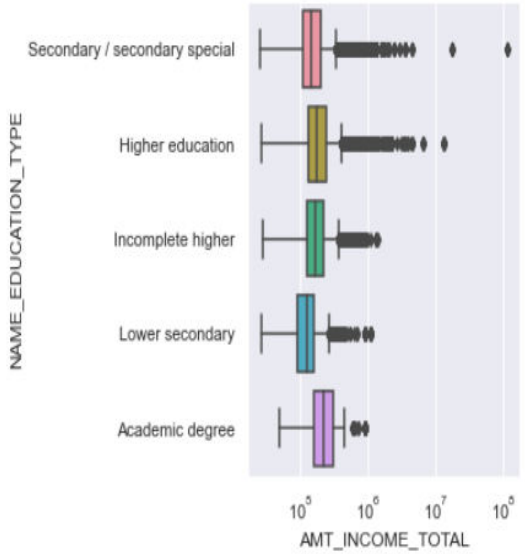
Business , Unknown and Self -Employed are the top 3 occupation where approval rate is more than any other occupation.



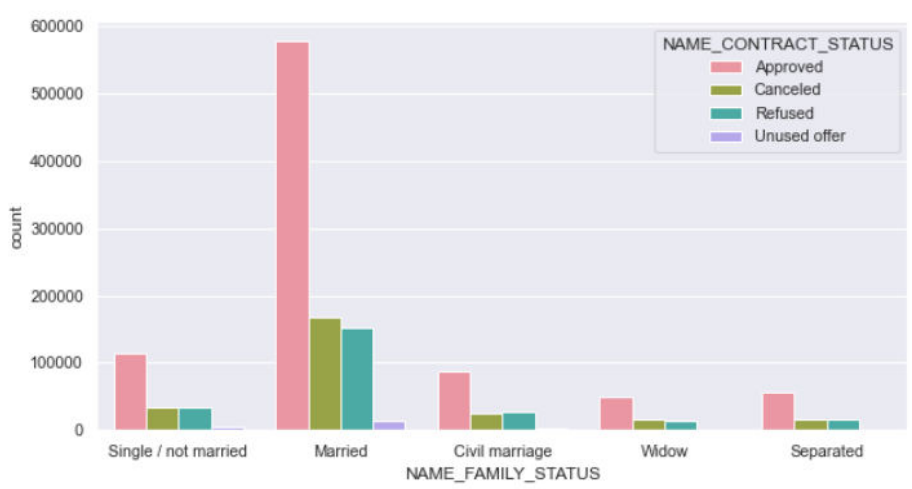
Repeater are more refused as compared to any other considering their Credit history

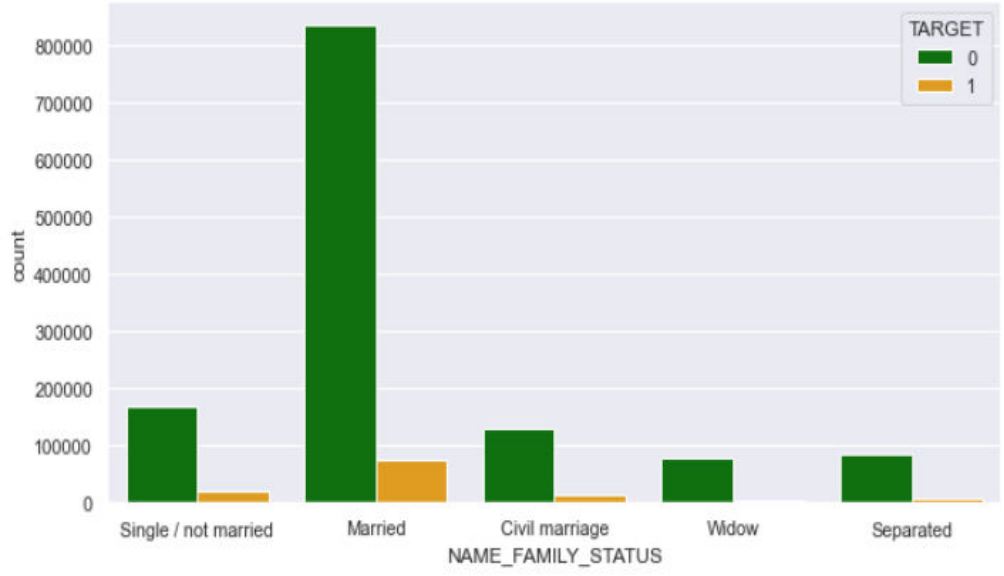
Secondary Education has the Highest Approval rate ,although the Income of Academic degree holder are more in comparison Secondary education still the approval rate of Secondary Education is more than Academic Degree holders



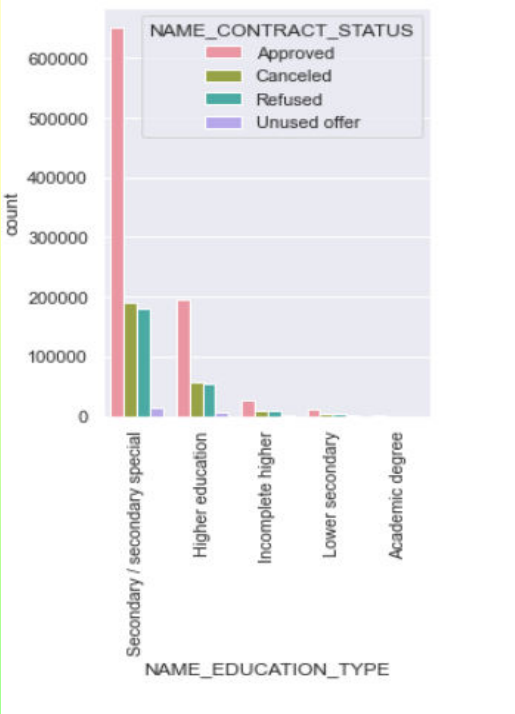


Married people are more likely to get Loan approved in comparison to any other marital status of people with respect to NAME_CONTRACT_STATUS

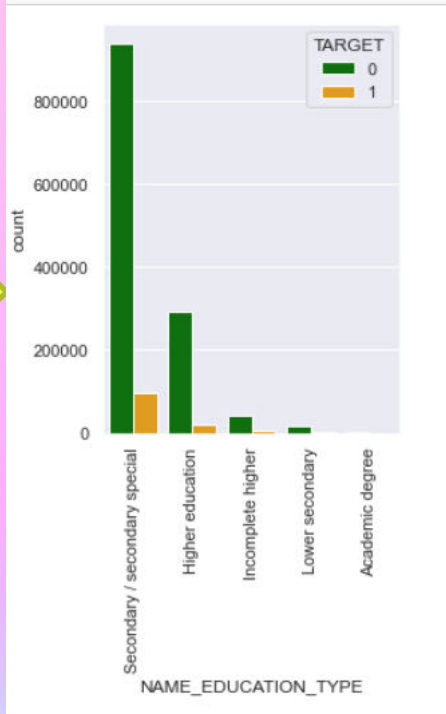


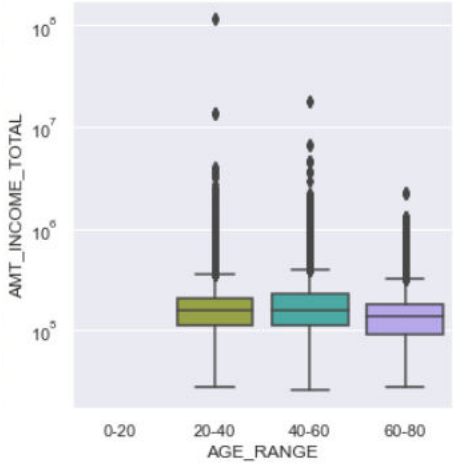


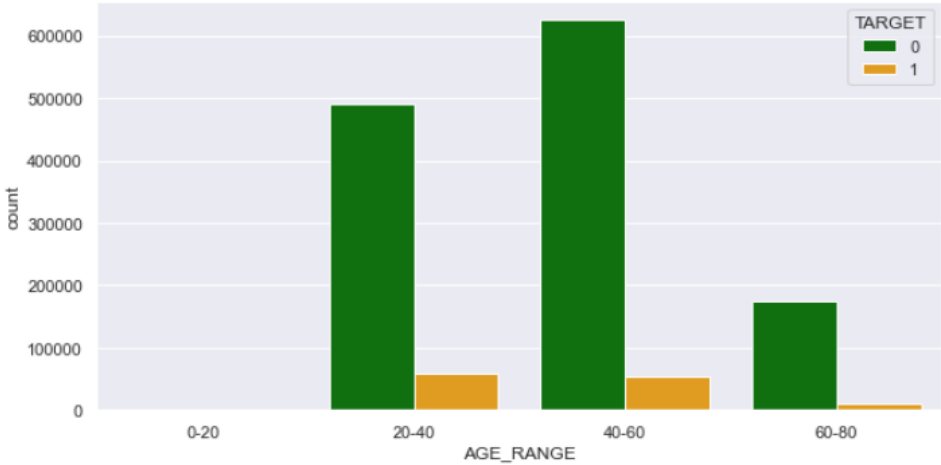
Married people are more likely to get Loan approved in compare to any other marital status of people with respect to TARGET.



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







When analyzing the age range , it is noticed that there is not much of difference in the income of people of age range 20-40, 40-60, further more the approval rate of people with age range 40-60

FINAL RECOMMENDATIONS :-

- ✓ Target variable for Application dataset - "TARGET"
- ✓ Target variable for Previous dataset - "NAME_CONTRACT_STATUS"
- ✓ The rate of defaulters are less in the age range of 40-60 are good target audience.
- ✓ Laborers , Core and Sales Staff is the occupation type that has the loan approved and has the highest non defaulter rate.
- ✓ Married people are more likely to get loan approved as compared to any other Marital Status of the people so this is also a good target audience .
- ✓ 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