CREDIT EDA - ASSIGNMENT

BY ANAND PRATAP

INDEX:

INTRODUCTION:

- This assignment aims to give an idea of applying EDA in a real business scenario. In this assignment, apart from applying EDA techniques, we also develop a basic understanding of risk analytics in banking and financial services. We understand how data is used to minimize money loss while lending to customers.
- Insufficient or non-existent credit histories make it difficult for loan providers to provide loans to people. Because of that, some consumers use it to their advantage by defaulting on their loans. Suppose you work for a consumer finance company that specializes in lending various types of loans to urban customers. we have to use EDA to analyze data patterns. The applicants will not be turned down if they are capable of repaying the loan.

BUSINESS OBJECTIVES:

- This case study aims to identify patterns which indicate if a client has difficulty paying their instalments. These patterns may be used for actions such as denying the loan, reducing the amount of the loan, lending (to risky applicants) at a higher interest rate, etc.
- This will ensure that consumers who are capable of repaying loans 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 that are strong indicators of default. Portfolio and risk assessments can be based on this information.

DATA CLEANING APPROACH:

- Dropped columns with more than 40% missing values in both data sets. For the remaining columns, we treated missing values. For example. OCCUPATION_TYPE column, despite having 31% missing values, gives some useful insights. Imputing missing values with mode will distort the data. So we will leave it as it is.
- Apart from the missing values there are many columns with XNA and XAP (Not available and Unknown). In case the percentage is higher, we can use mean/median/mode or keep it as is.
- There were columns with negative values for days (Birth, Employment, ID Publish). Converted them into positive and numerical years for better analysis.
- Outliers/aberrations were found in CNT_CHILDREN columns, for example. For a client in the 20-30 age range, there are 19 children. Most of our clients with more than 10 children are between the ages of 30 and 50.
- In reality, people who have been employed for 365243 days (1000+ years) are actually in the pensioner and unemployed income category. These are outliers.

MISSING VALUES:

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

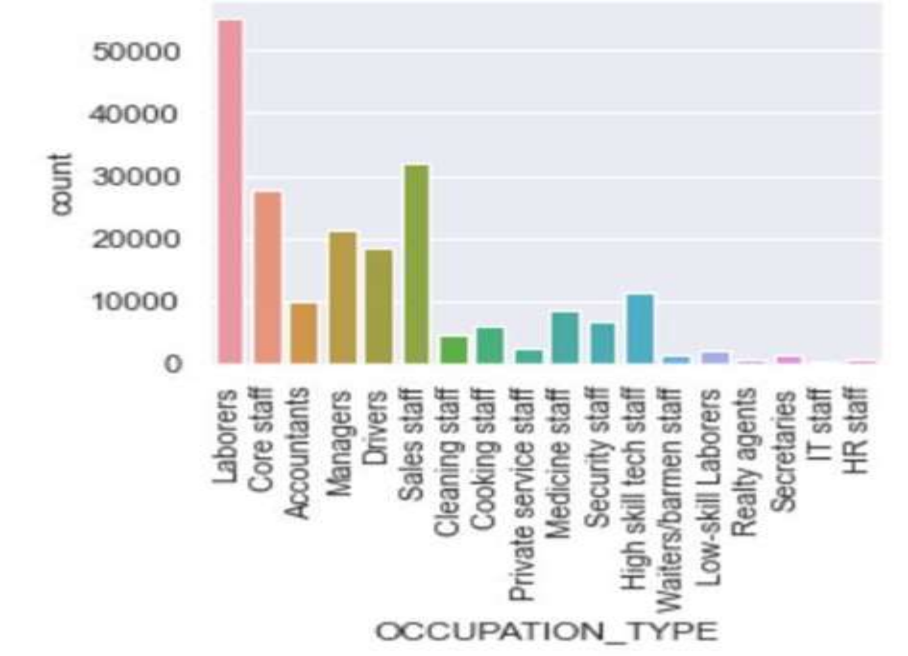
WALLSMATERIAL_MODE	50.84
APARTMENTS_MEDI	50.75
APARTMENTS_AVG	50.75
APARTMENTS_MODE	50.75
ENTRANCES_MEDI	50.35
ENTRANCES_AVG	50.35
ENTRANCES_MODE	50.35
LIVINGAREA_AVG	50.19
LIVINGAREA_MODE	50.19
LIVINGAREA_MEDI	50.19
HOUSETYPE_MODE	50.18
FLOORSMAX_MODE	49.76
FLOORSMAX_MEDI	49.76
FLOORSMAX_AVG	49.76
YEARS_BEGINEXPLUATATION_MODE	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

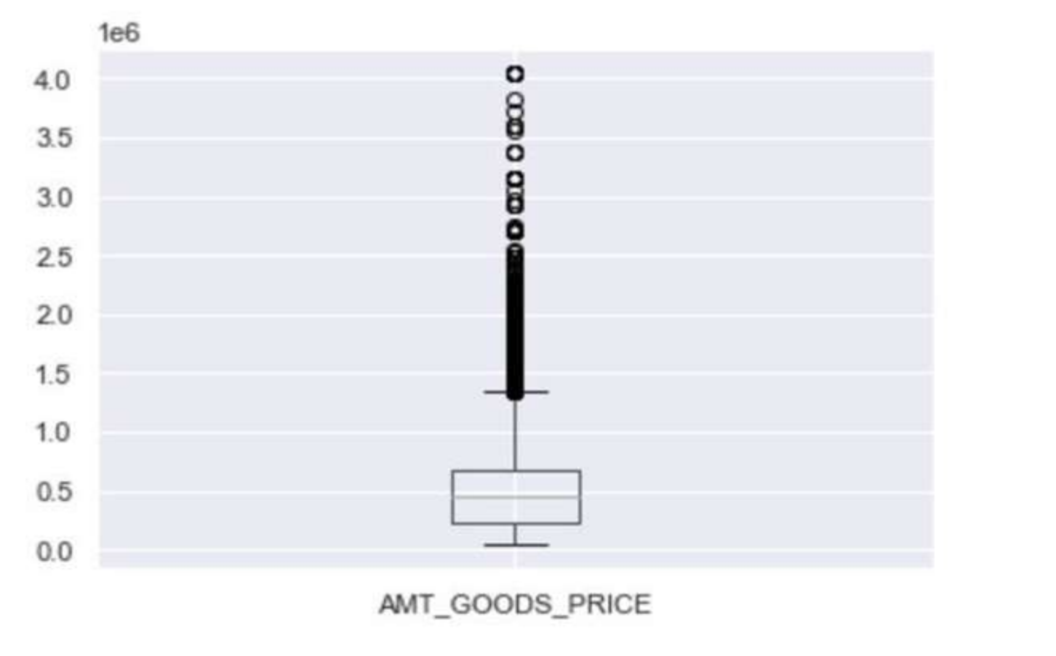
RATE_INTEREST_PRIVILEGED	99.64
RATE_INTEREST_PRIMARY	99.64
AMT_DOWN_PAYMENT	53.64
RATE_DOWN_PAYMENT	53.64
NAME_TYPE_SUITE	49.12
NFLAG_INSURED_ON_APPROVAL	40.30
DAYS_TERMINATION	40.30
DAYS_LAST_DUE	40.30
DAYS_LAST_DUE_1ST_VERSION	40.30
DAYS_FIRST_DUE	40.30
DAYS_FIRST_DRAWING	40.30
AMT_GOODS_PRICE	23.08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

▪ TREATING MISSING VALUES:

Since OCCUPATION_TYPE is a categorical variable, we replace null with mode and I've filled in null here as Laborers.

We replaced AMT_GOODS_PRICE, a continuous variable with median instead of mean because the data set has many outliers.

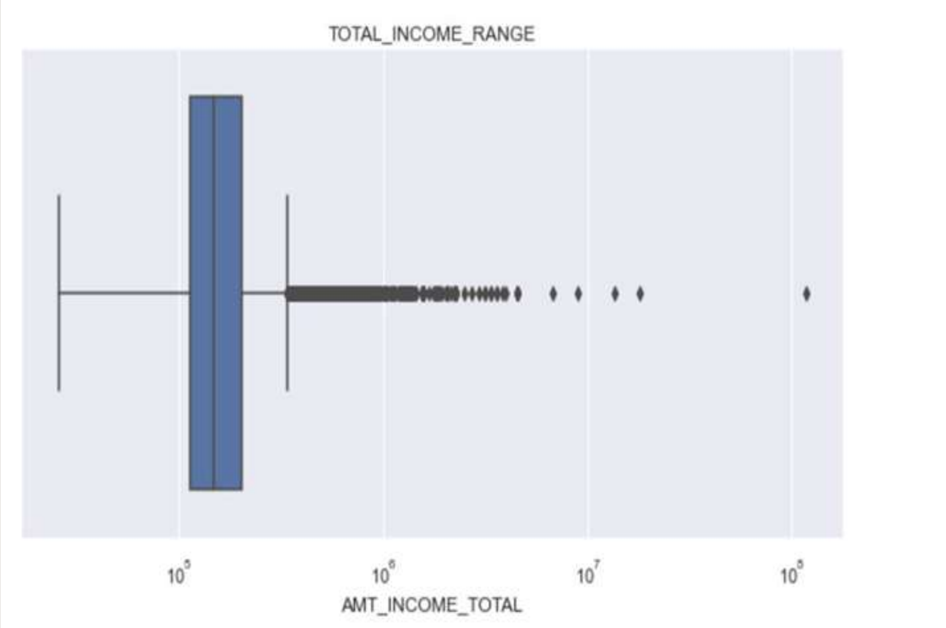


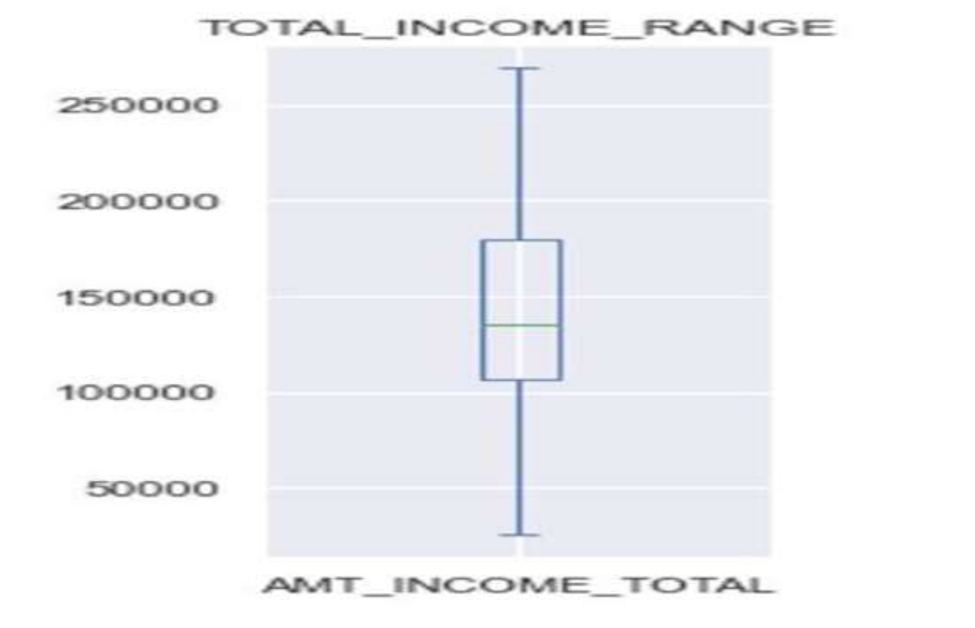


HANDLING OUTLIER:

- IN THIS GRAPH WE CAN SEE THERE ARE A LOT OF OUTLIERS
- AFTER CHECKING THE MEAN AND MEDIAN, THERE WAS A HUGE DIFFERENCE.
- THEREFORE, THE MEDIAN WILL BE USED TO FILL IN THE MISSING VALUES

- IN THIS CASE THE OUTLIERS HAVE TO BE TREATED BY TAKING VALUES LESS THAN 90 PERCENTILE.
- OTHER COLUMNS CAN ALSO BE TREATED SIMILARLY

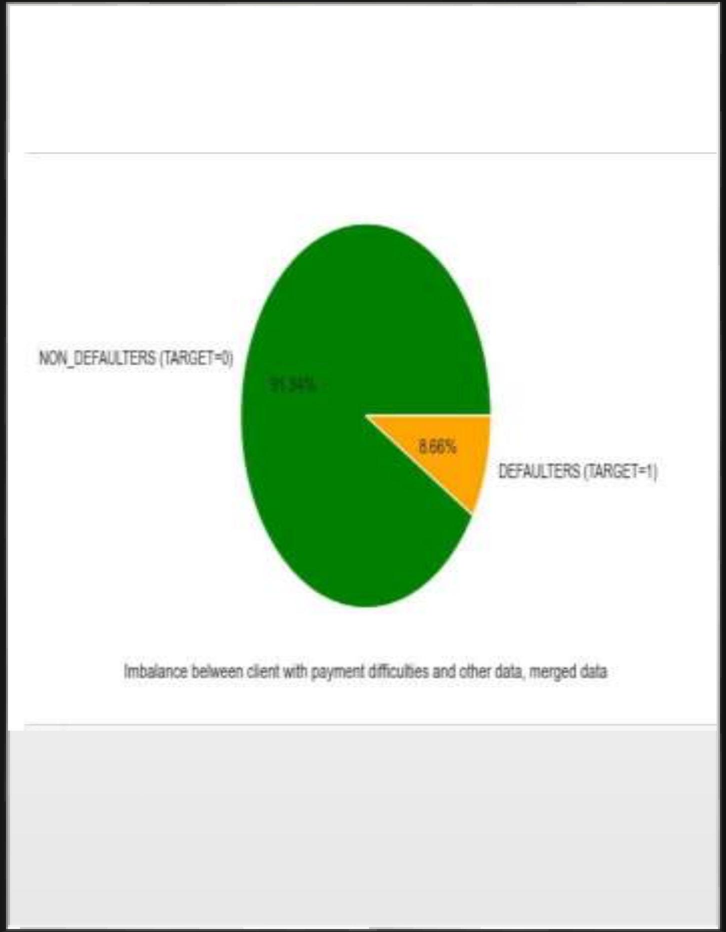




METHEDOLOGY:

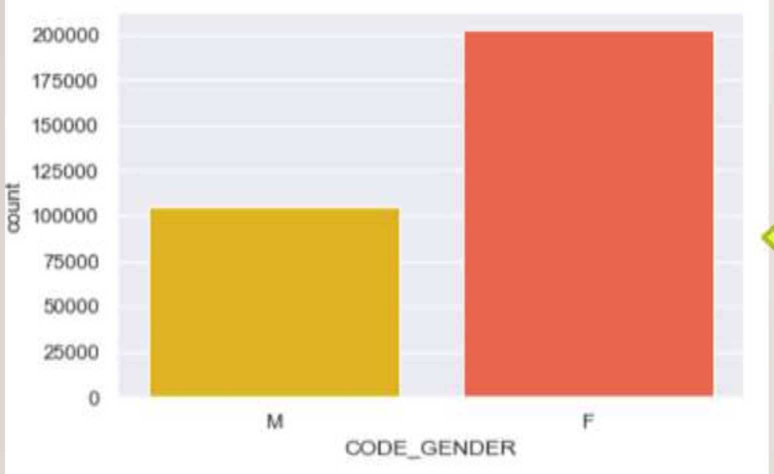
- Checked data imbalance in the Target variable, and found 11.4% imbalance. Due to data imbalance, we separated the application data into 2 datasets, with Target 0 and Target 1. We analyzed them separately with Pie charts and Count plots.

- Later merged Application Data set and Previous Application data set on common column SK_ID_CURR. It appears that there are duplicate entries of SK_ID in the current and previous applications, which indicates that the client have multiple loans. In the merged dataset we checked imbalance. The situation was similar.

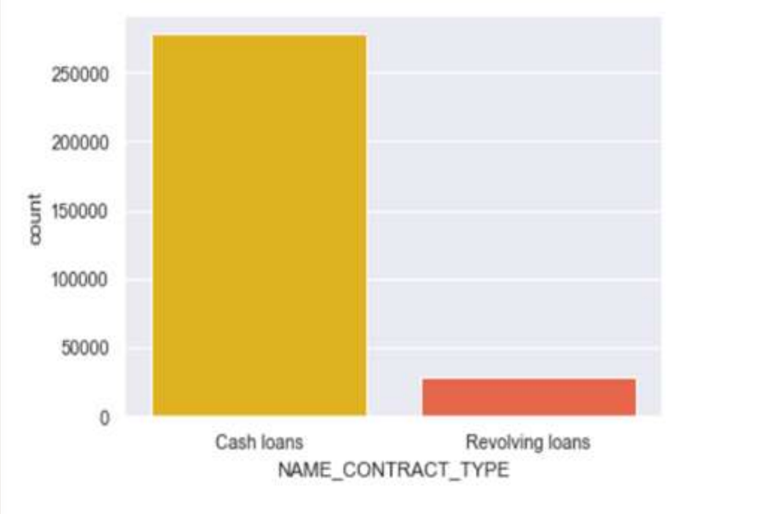


UNIVARIATE ANALYSIS:

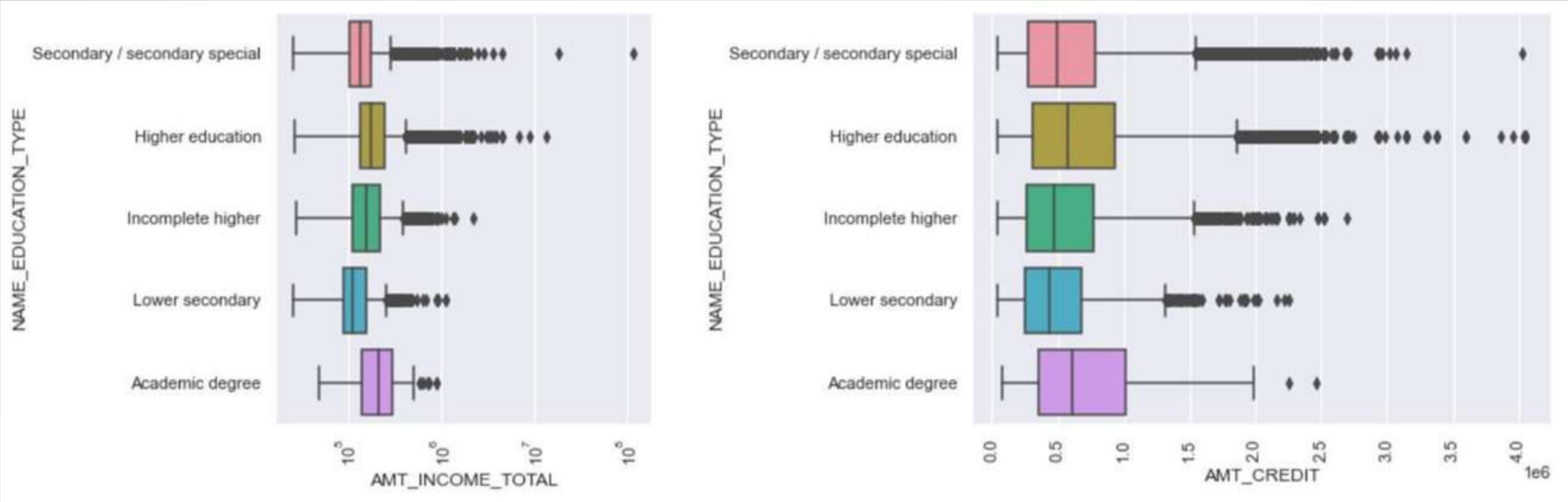
- Compared to revolving loans, cash loans are more common.



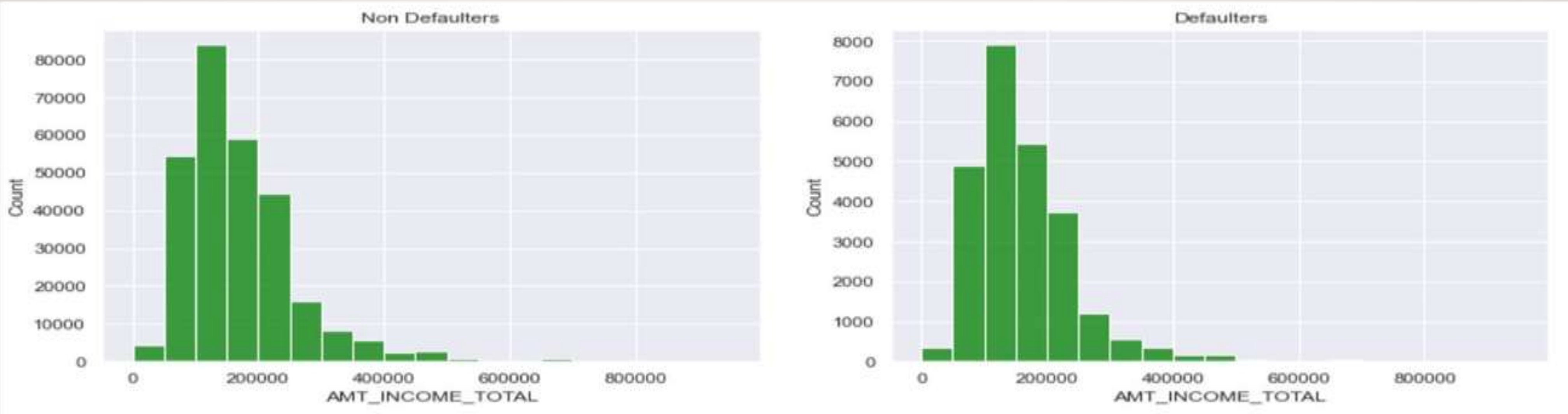
- According to the data female is the most dominating gender compared to male.
- Male population:- 100000 & female population:- 200000



BIVARIATE ANALYSIS:



SEGMENTED ANALYSIS:

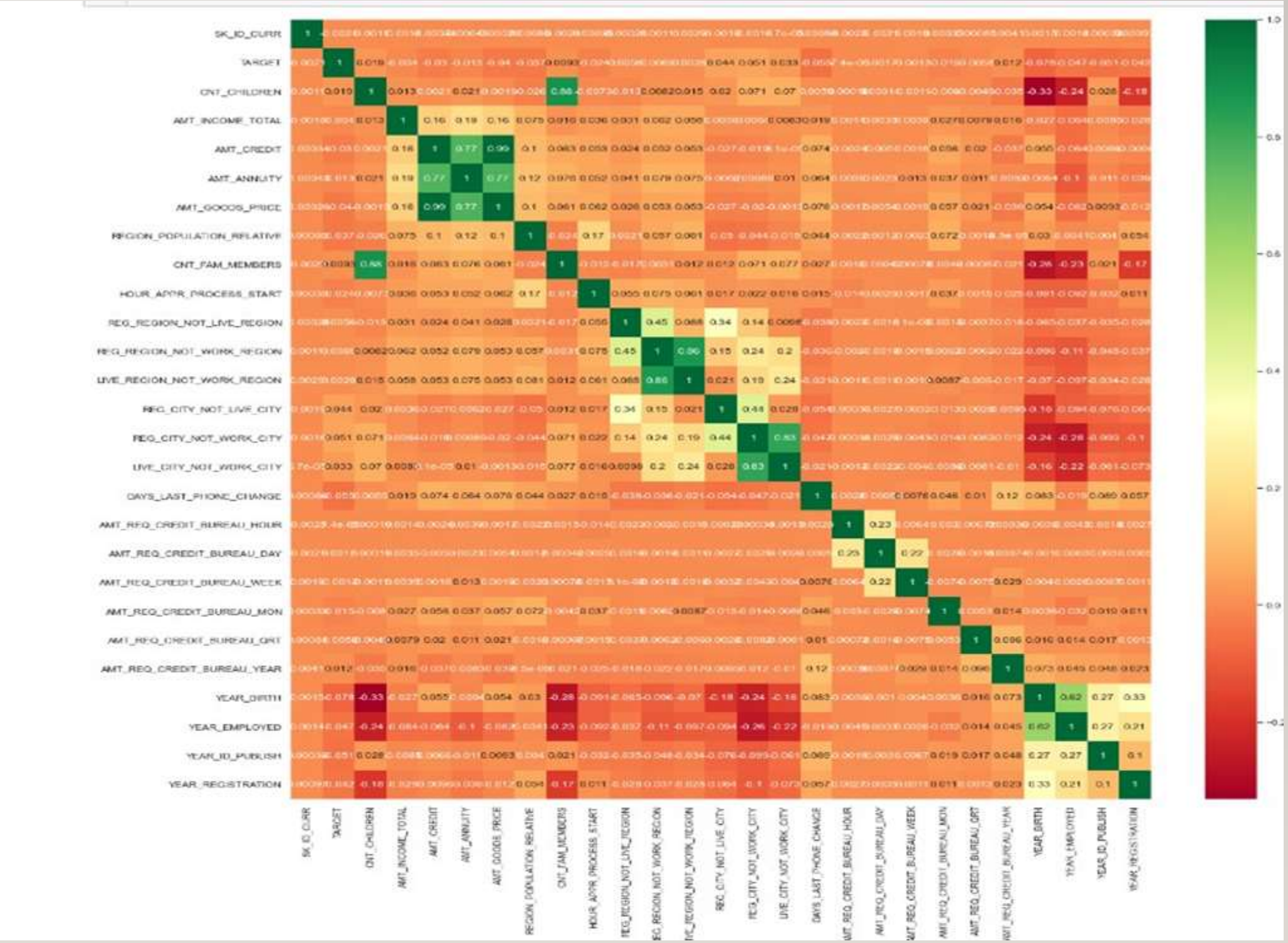


AMT_INCOME_TOTAL for Defaulters And Non-Defaulters.

MULTIVARIATE ANALYSIS

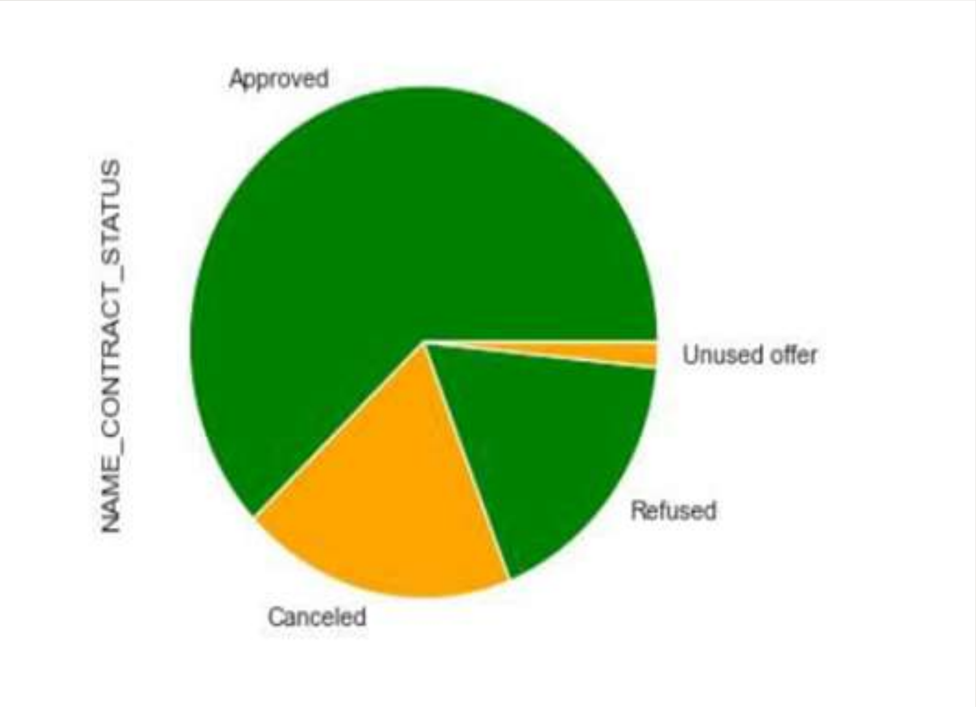
- According to the heat map, there is a strong association between AMT_CREDIT and AMT_GOODS_PRICE.

- A substantial correlation of 0.88 exists between CNT_CHILDREN and CNT_FARM_MEMBERS.



ANALYSIS OF PREVIOUS DATA

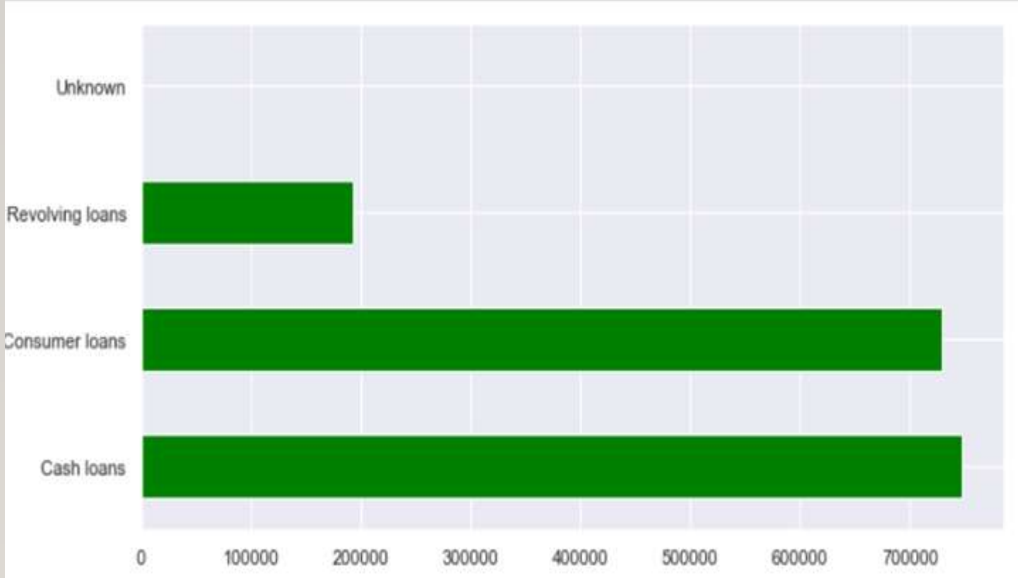
PREVIOUS DATA:

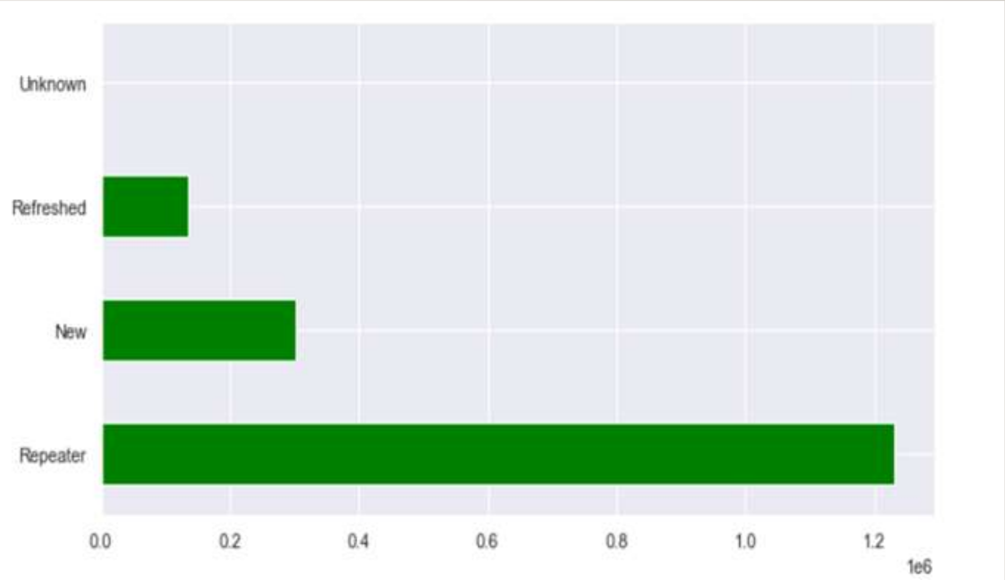


- The target feature found in the previous data csv fi le is NAME_CONTRACT_STATUS.
- The acceptance rate of APPROVAL is higher than all other kinds of status, according to the pie chart .

UNIVARIATE ANALYSIS:

- The cash loan are more as compare to other.
- Repeater clients are more as compare to others.

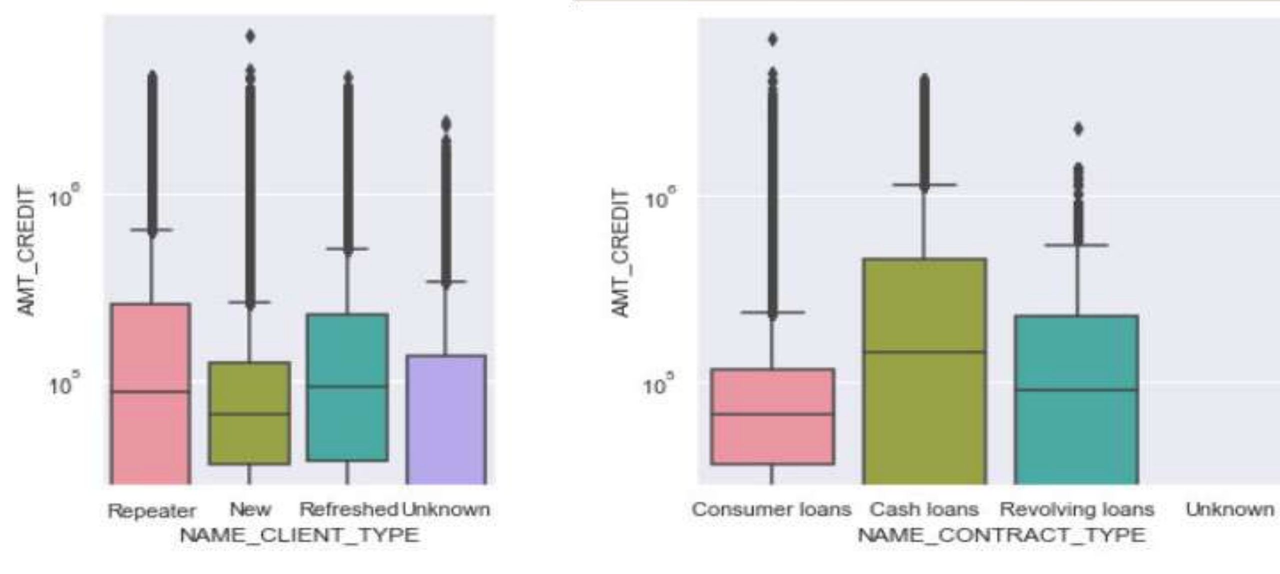




BIVARIATE ANALYSIS

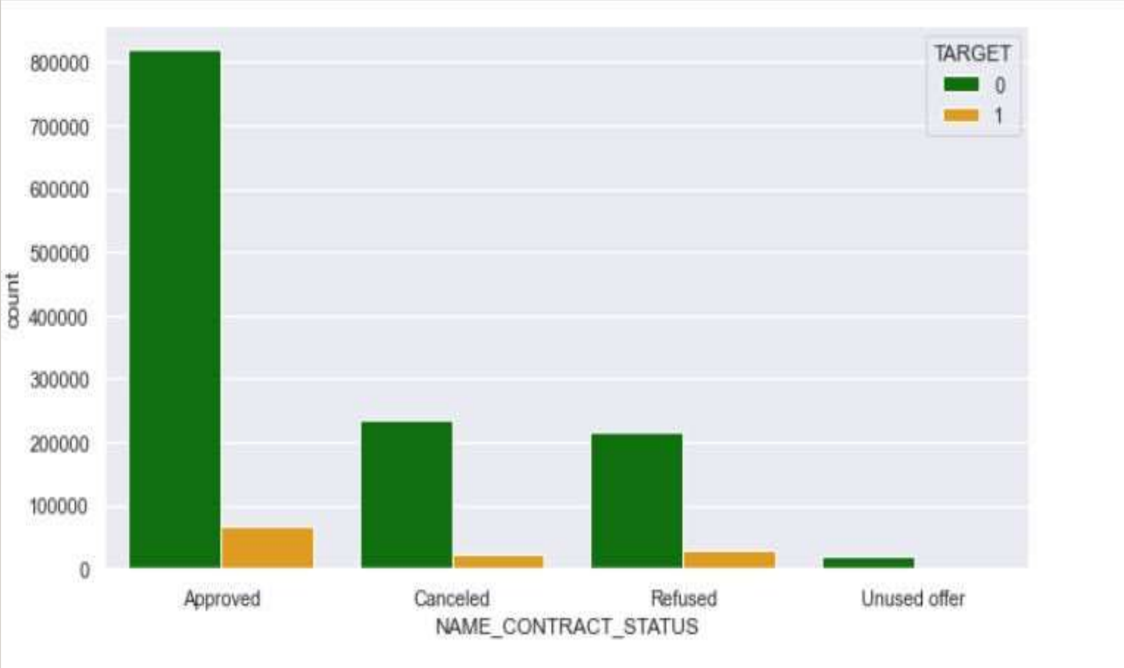
- ➢ Repeaters have the highest AMT_ GOODS_PRICE and Cash loans are also more compare to other loans

- ➢ When compared to AMT_CREDIT, repea ters and cash loan are worth the most.



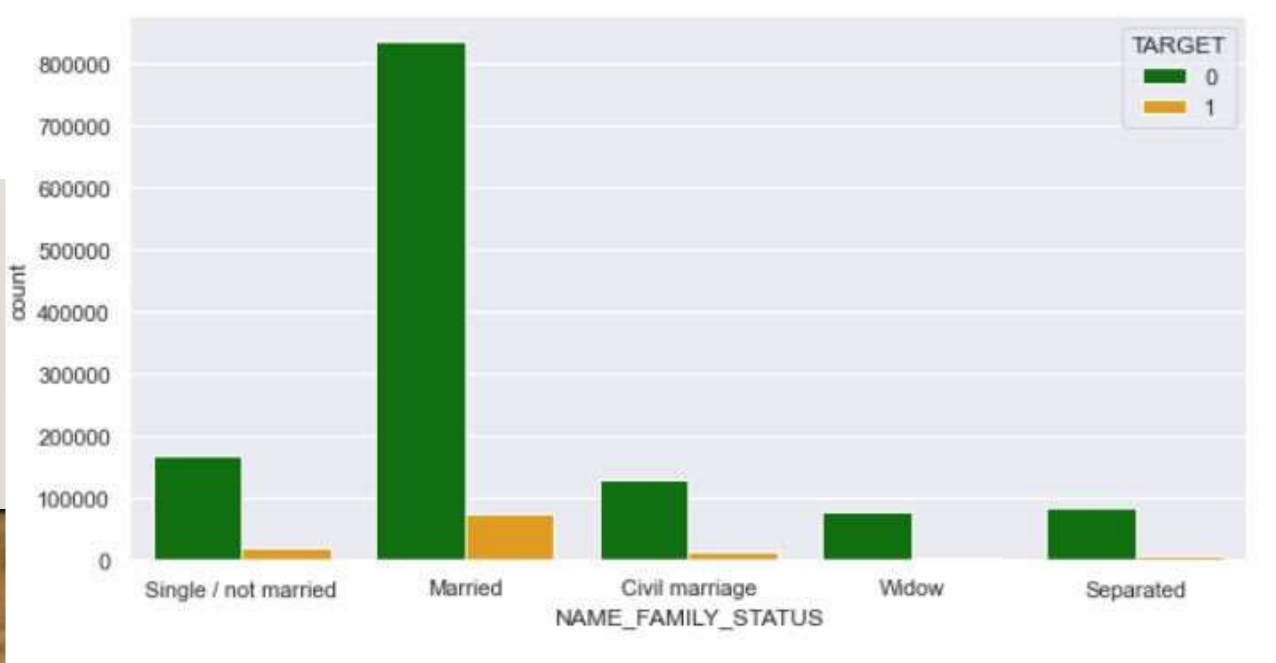
ANALYSIS OF MERGED DATA

MERGE DATA FRAME OF PREVIOUS AND CURRENT APPLICATION

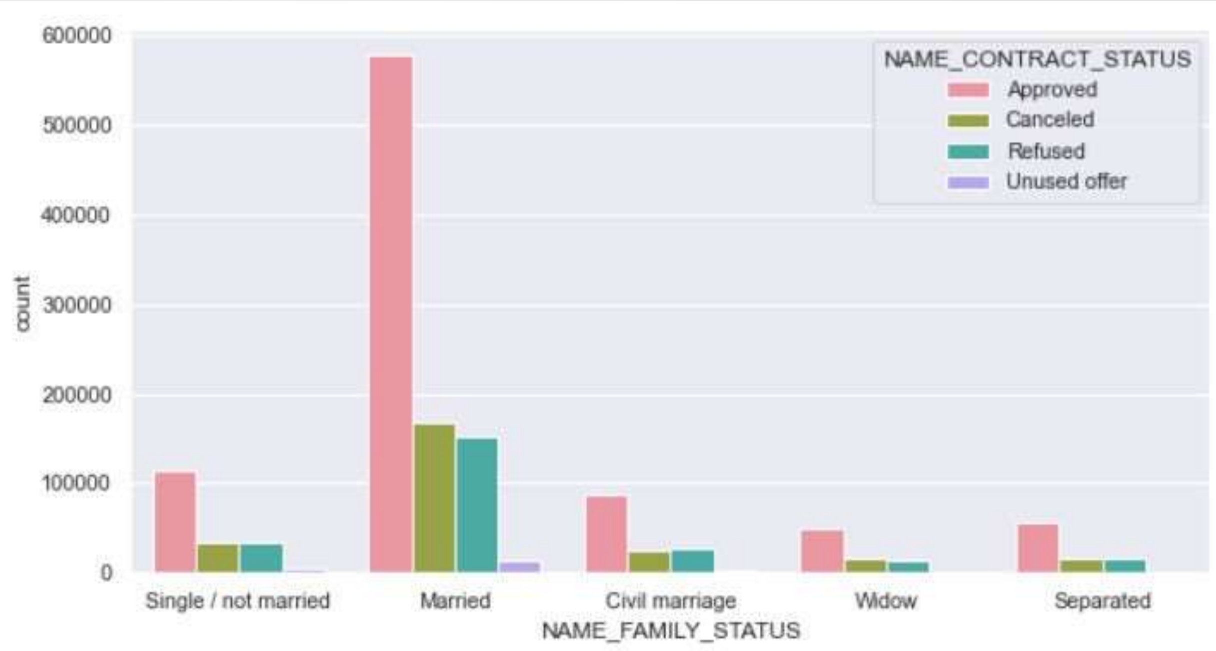


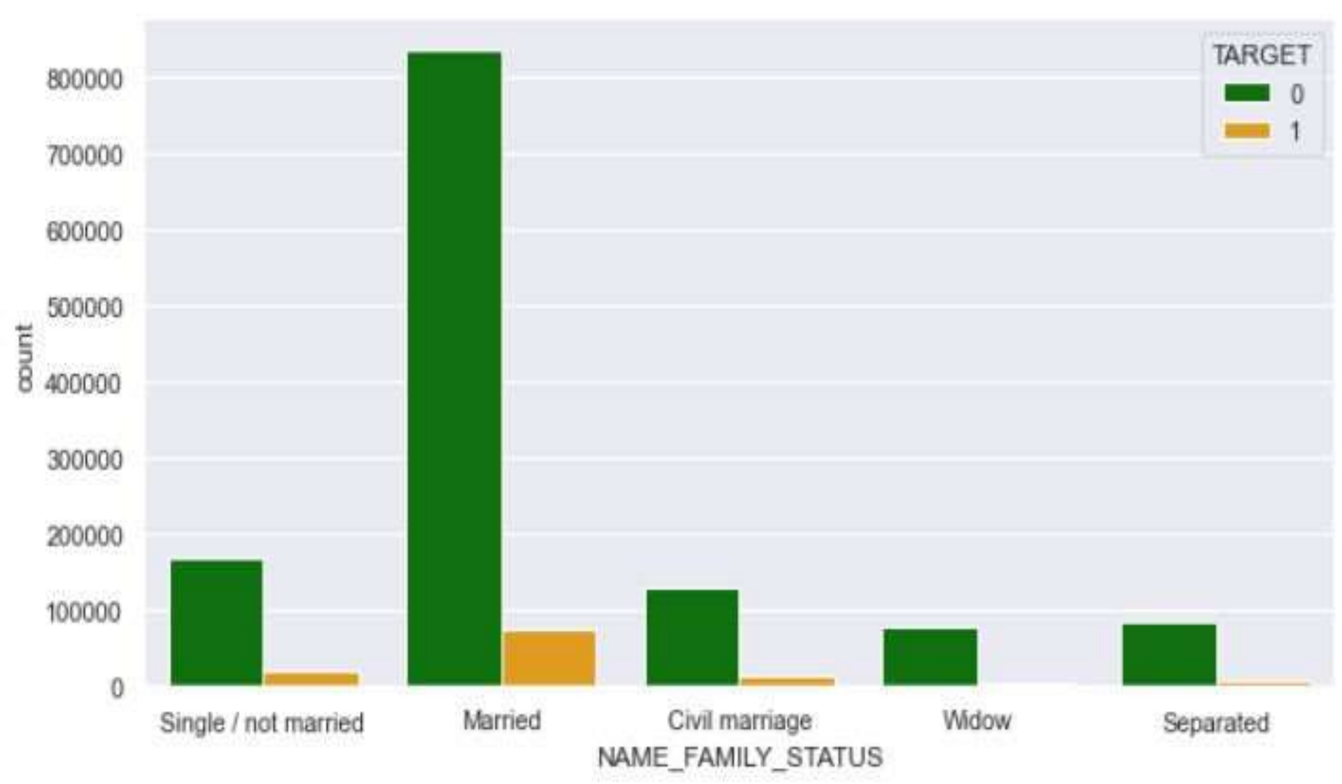
- Here showing that approved status is high of non defaulters

- In comparison to any other marital stat us, married people have a higher chan ce of getting the loan request approved .

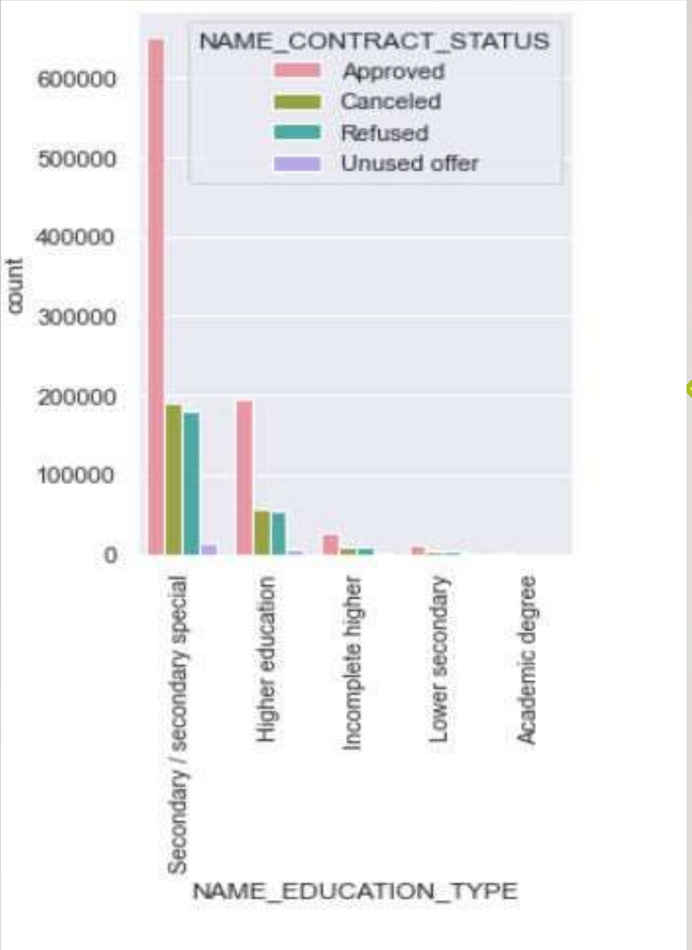


- Regarding NAME_CONTRACT_STATUS, married individ uals are more likely to have their loan request acknowle dged than people
with any other marital status.

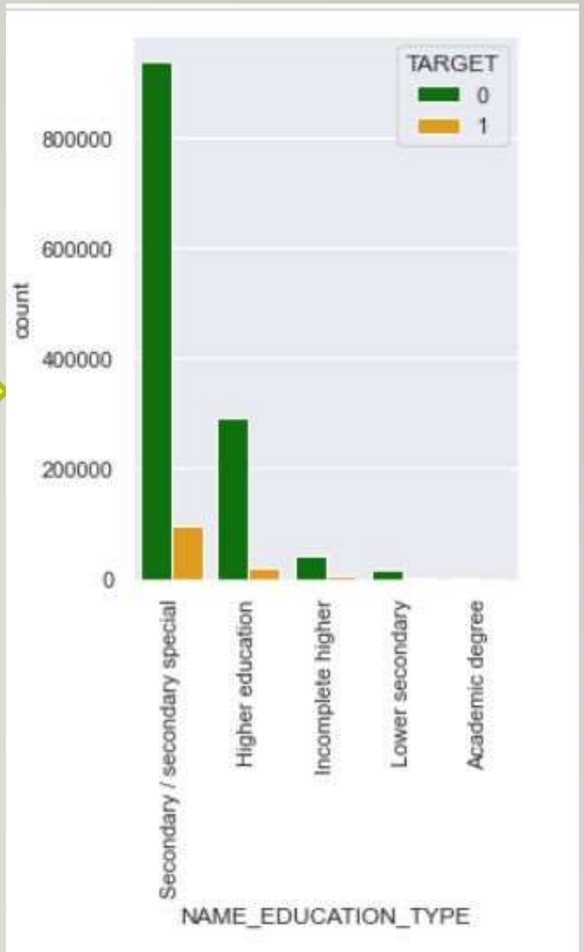


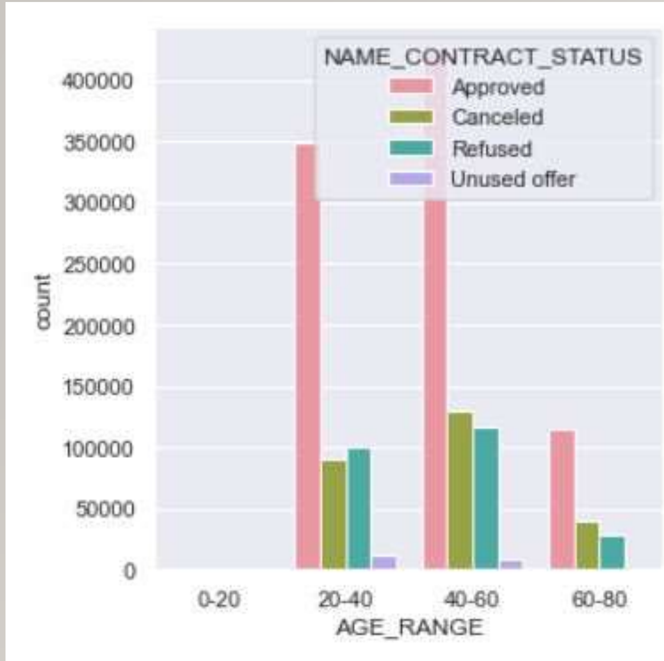


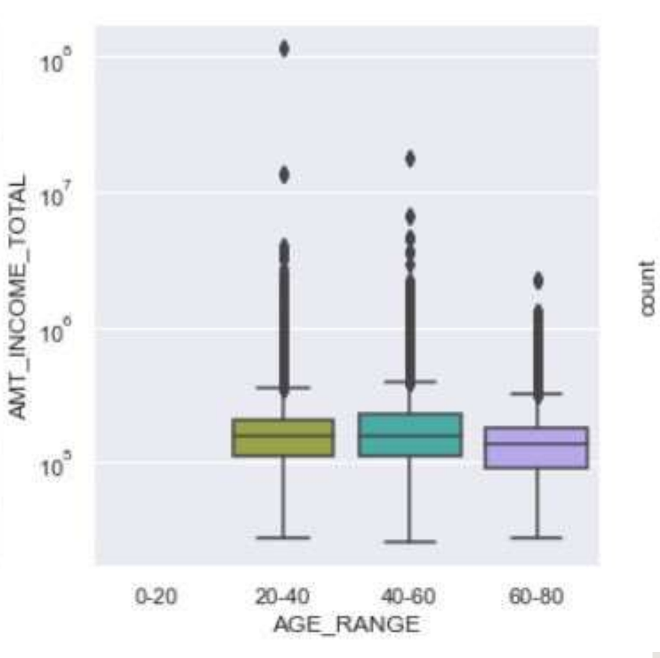
According to TARGET, married individuals have a higher chance of getting a loan granted than people in any other marital situation.

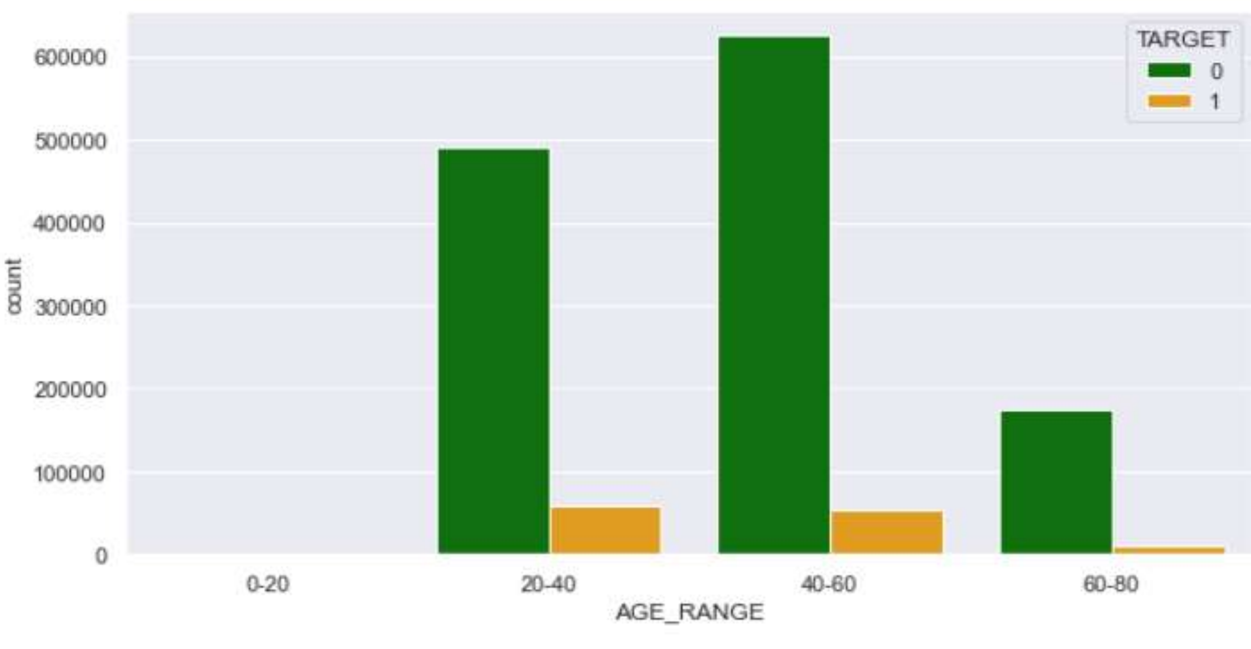


- According to the two graphs, secondary
- education has the fewest defaulters and the hi ghest approval percentage, consequently this will be the audience to target if you possess a h istory in education.









When the age range is analyzed, it is found that there is little difference in the financial standing of those between the ages of 20 and 40 and those between the ages of 40 and 60.

FINAL RECOMMENDATION:

- "TARGET" is a desired variable for the application dataset.
- "NAME_CONTRACT_STATUS" is the desired variable for the previous dataset.
- The age group of 40 to 60 is a strong demographic to target because there are fewer defaulters in that group.
- The occupations with the highest non-defaulter rates include laborer's, core staff members, and sales staff.

- This is also an excellent target group because married people are more likely to have a loan authorised than persons in any other marital status.
- Although secondary schooling has the greatest approval percentage, academic degree holders' earnings are higher in contrast. The approval rate for secondary education is still higher than that for those with academic degrees.

Thank you