EDA ASSIGNMNET
CREDIT CASE STUDY

RASHI GUPTA

OBECTIVE

EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not Rejected.

Type of risk

- If the applicant is likely to repay the loan, then not approving the loan results in a loss
- of business to the company
- 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.

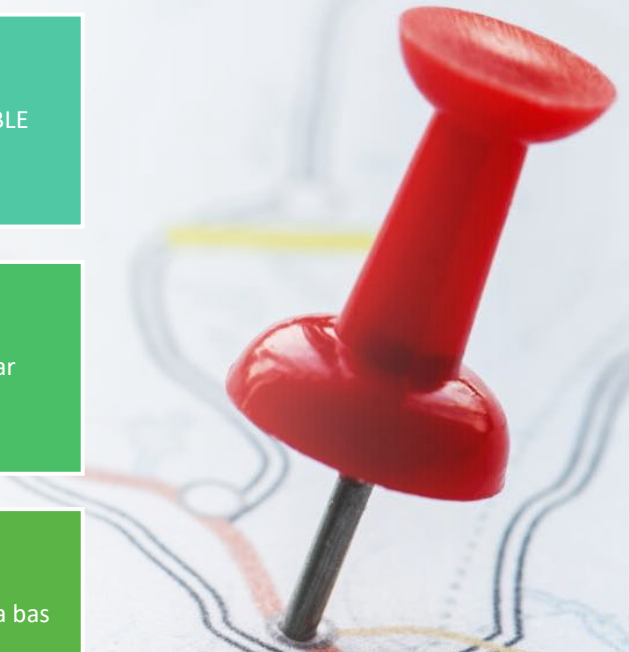
The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment

ROAD MAP

- Understanding Problem Statement
- Understanding Data
- Data Cleansing
- Outlier Analysis
- Data Analysis

ROAD MAP

- Analysisng Application_Data Frame data base
- Handling Missing values
- Dropping Null value collumn whose data more than 40%
- Impute the Numerical (countinous Variable) for missing values with the Median
- Imputing the missing value of Categorical variable
- looking for Outliers in Variable
- Univariate Analysis
- CONTINUOS VARIABLE
- CATEGORNICAL ANALYSIS
- Bivariate Anlysis
- Cont VS Cont
- TARGET VS Cont Var
- Cat vs Target
- MULTIVARIATE PLOTS
- Analysis done on second data base folowed same steps
- Merging both the data bas
- comparing all the APLICATION DF useful variable with the NAME_CONTRACT_STATUS
- PREVIOUS DATA useful variable with Target

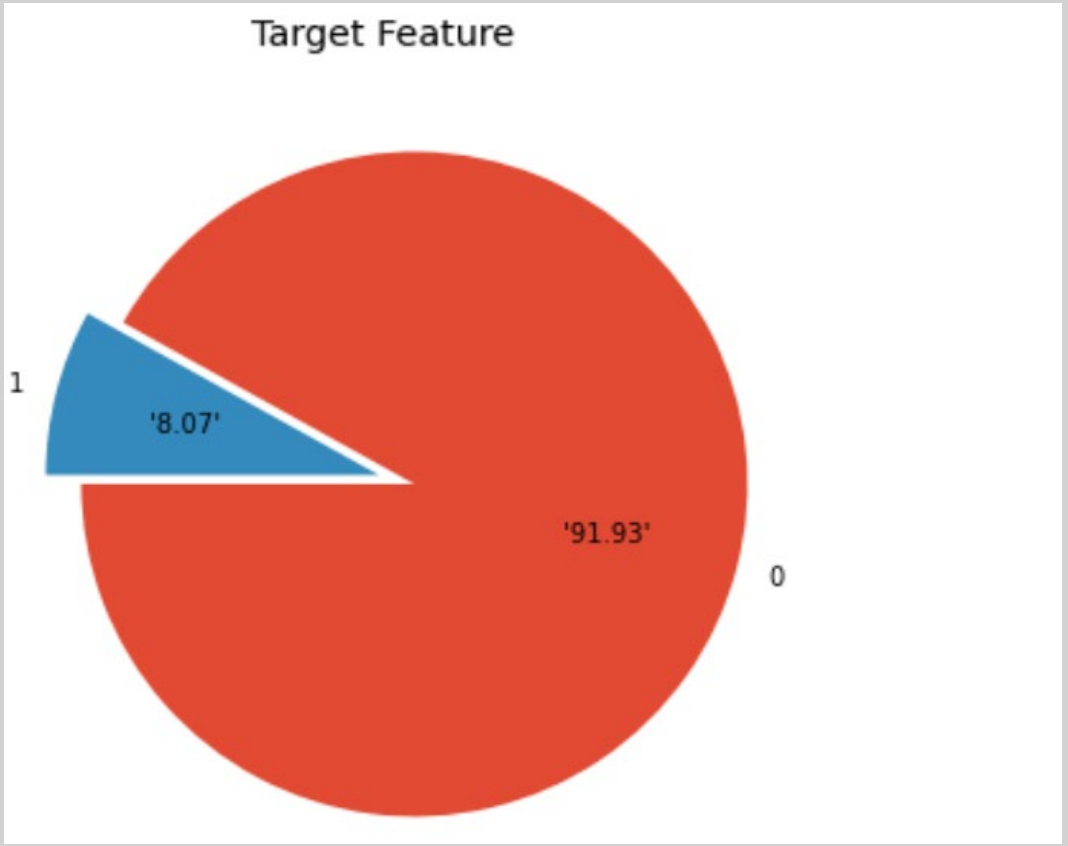


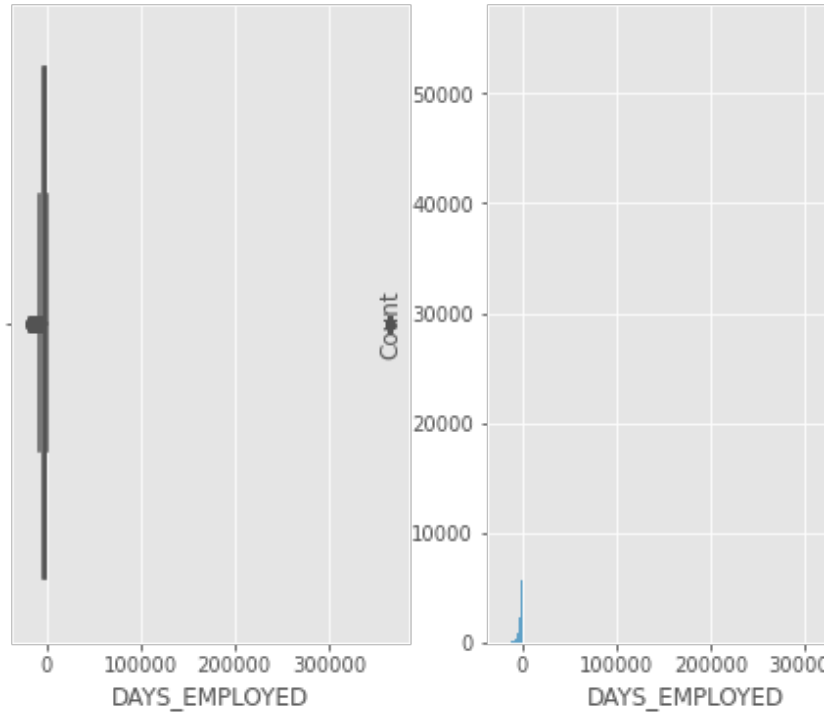
ANALYSIS ON APPLICATION DATA FRAME

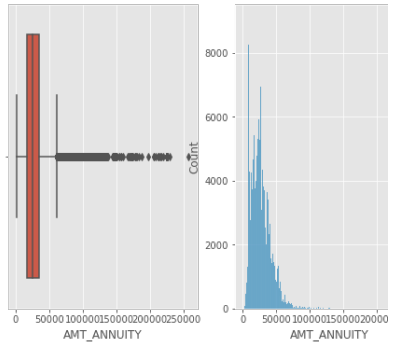
DATA is Highly Imbalance

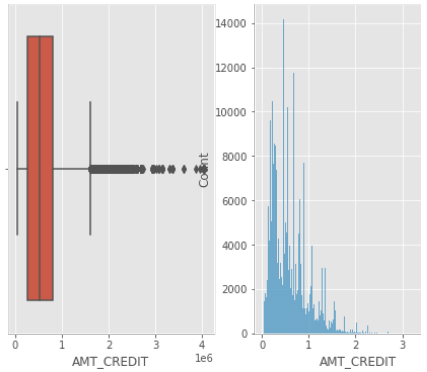
Insight From the above graph insight

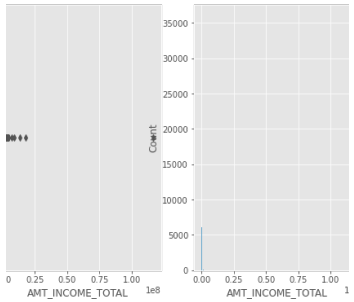
- Target Column is highly imbalance
- Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)
- Ratios of imbalance for Repayer and Defaulter in Percentage is: 91.93 and 8.07

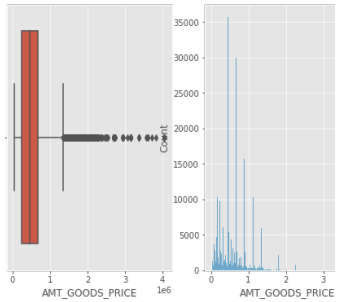


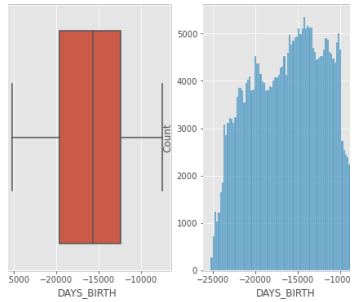








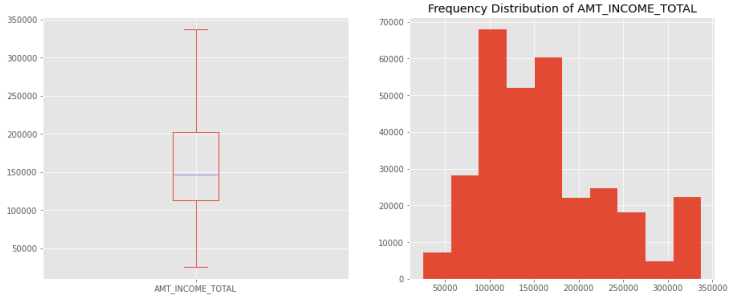


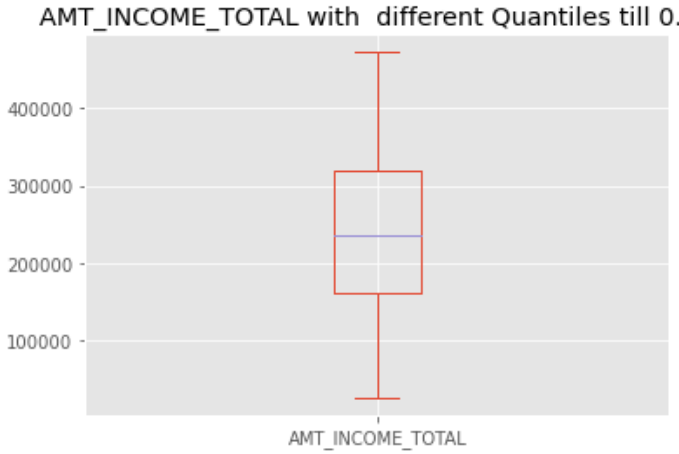


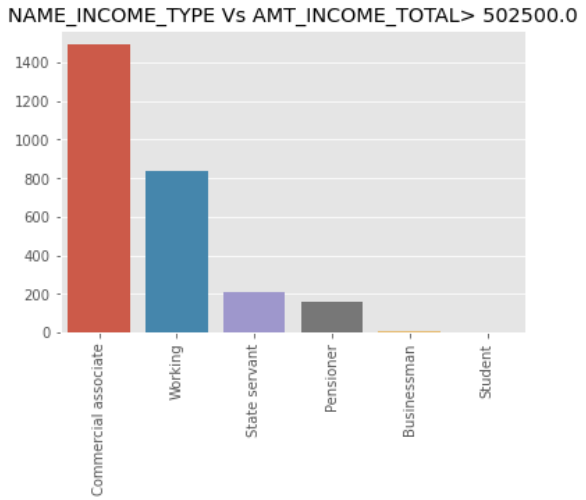
Outlier Analysis

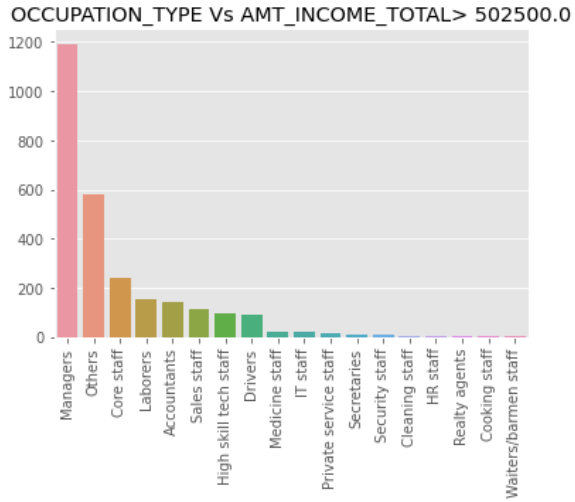
INFERENCE:-

- AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.
- AMT_INCOME_TOTAL has huge number of outliers which indicate that few of the loan applicants have high income when compared to the others.
- DAYS_BIRTH has no outliers which means the data available is reliable.
- DAYS_EMPLOYED,DAYS_BIRTH(DAYS_BIRTH,DAYS_EMPLOYED,having -ive value and converting in to years and take in to AGE Column.
- DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and hence this has to be incorrect entry.





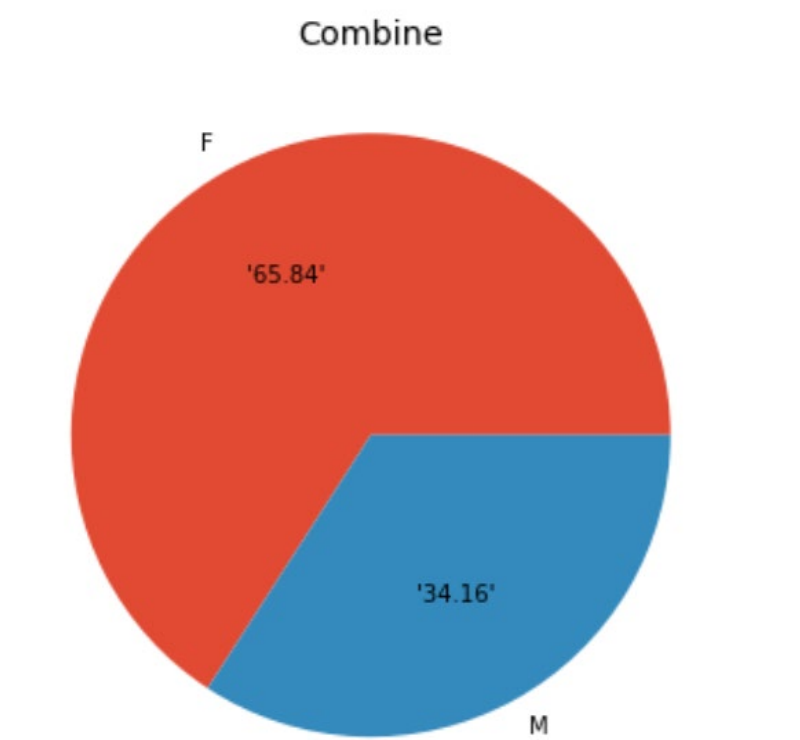


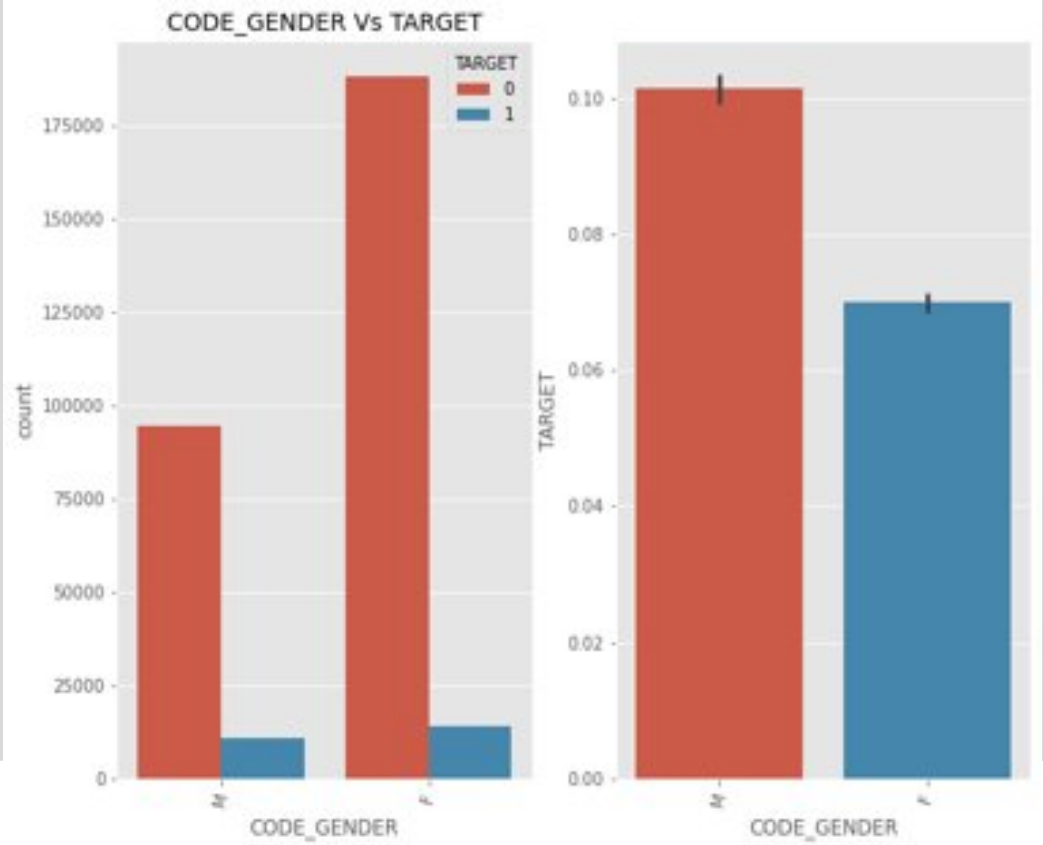


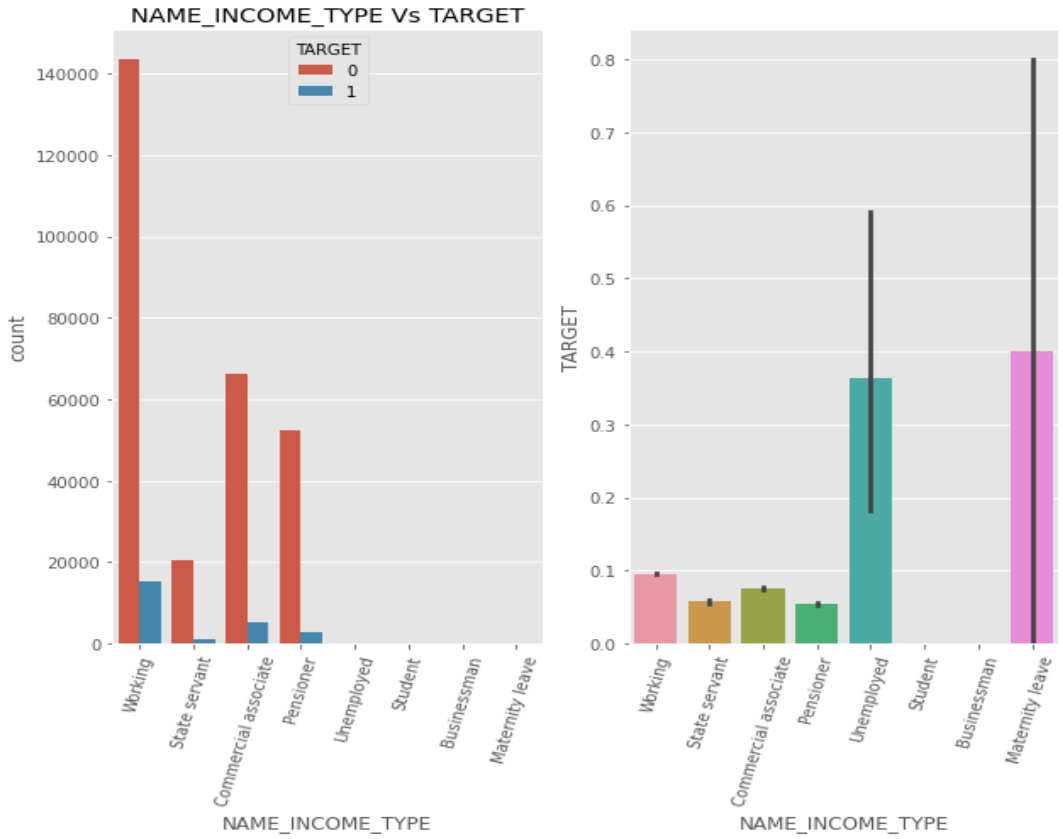
OUTLIER TREATMENT: one way to deal with OUTLIER is to CAP created function to capping to IQR the outlier , Analyse it from all respect

- Code Gender:--> Female ratio is very High more then 50% and XNA is wrongly input data which is not possible(only 4 in count so changing it to Mode value

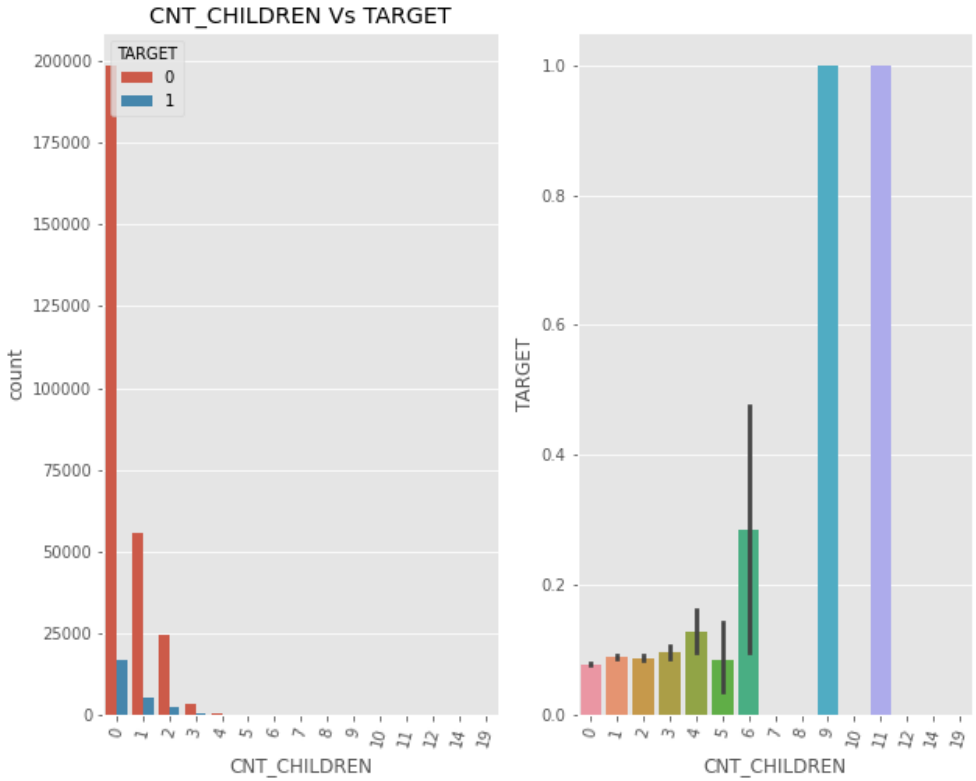
- As compare to female, Male are at default



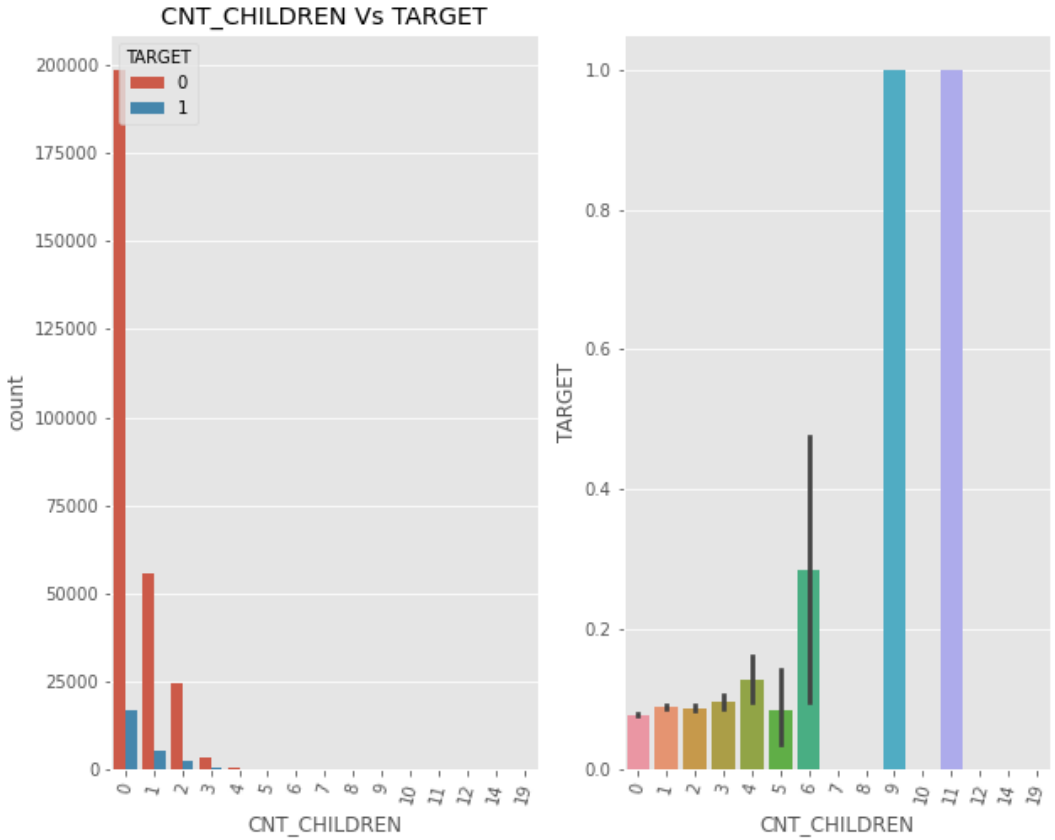


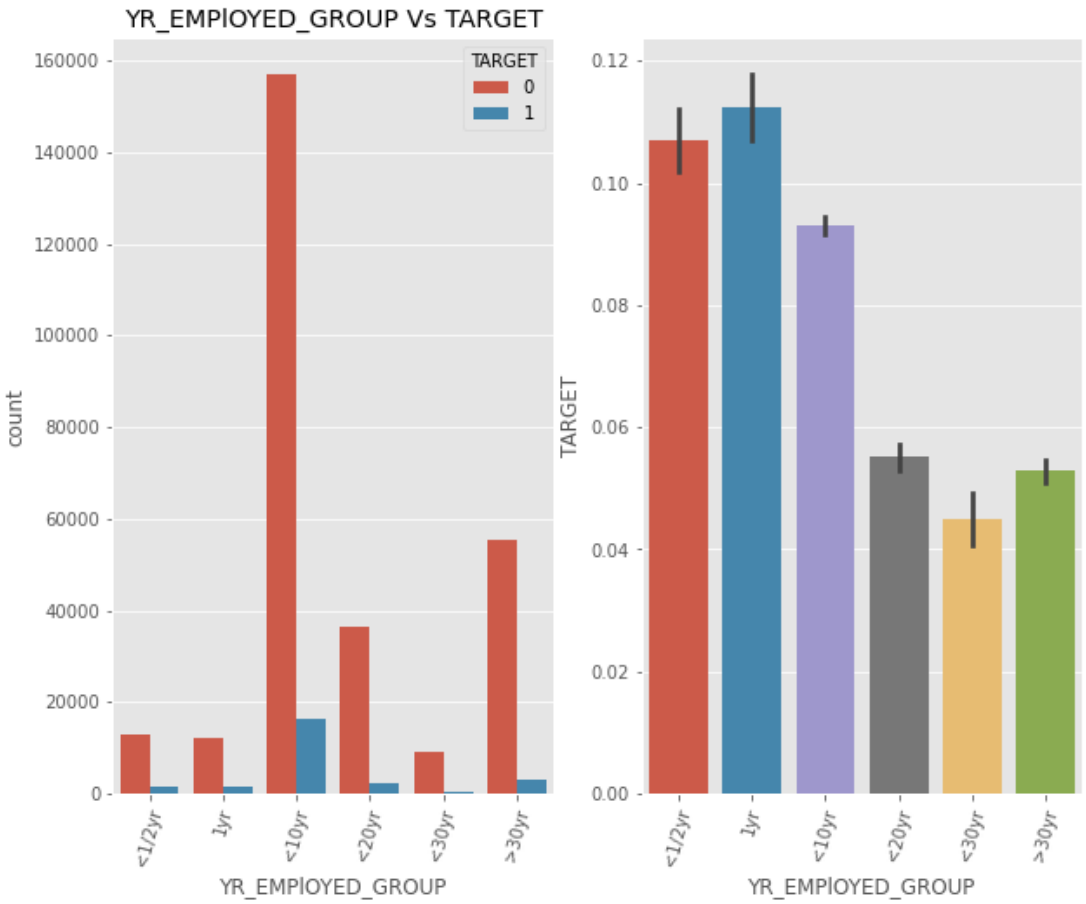


- Client who are UNEMPLYED
- Ratio of them to be defaulter are very high



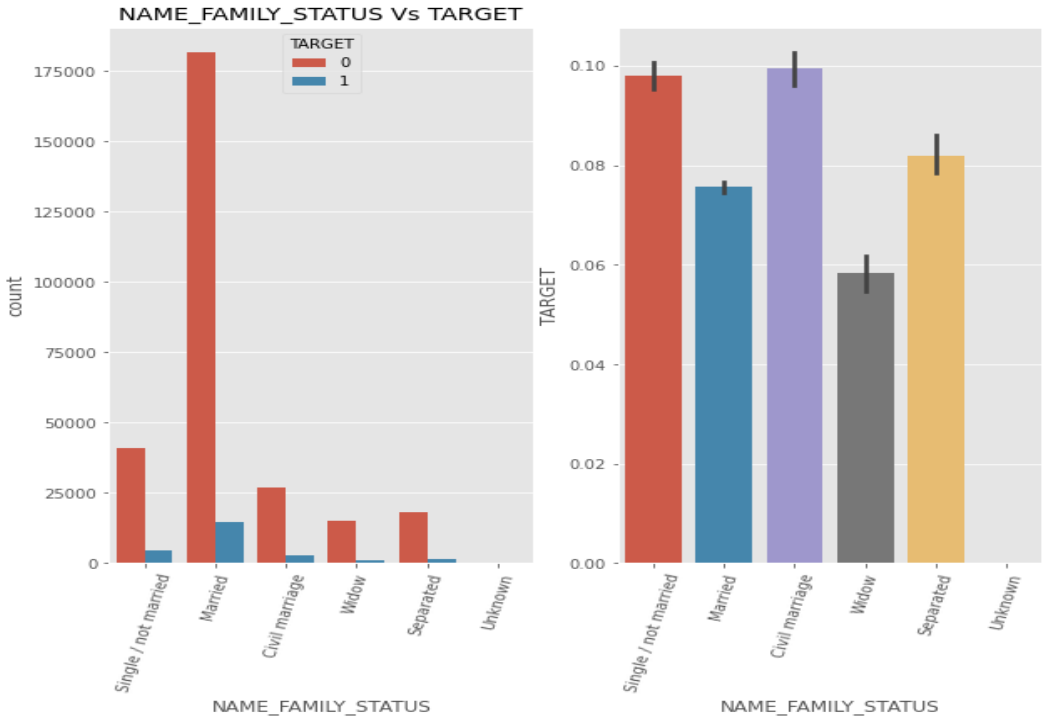
- Client who have less kids are more like to repay as compare to the kids more than 2

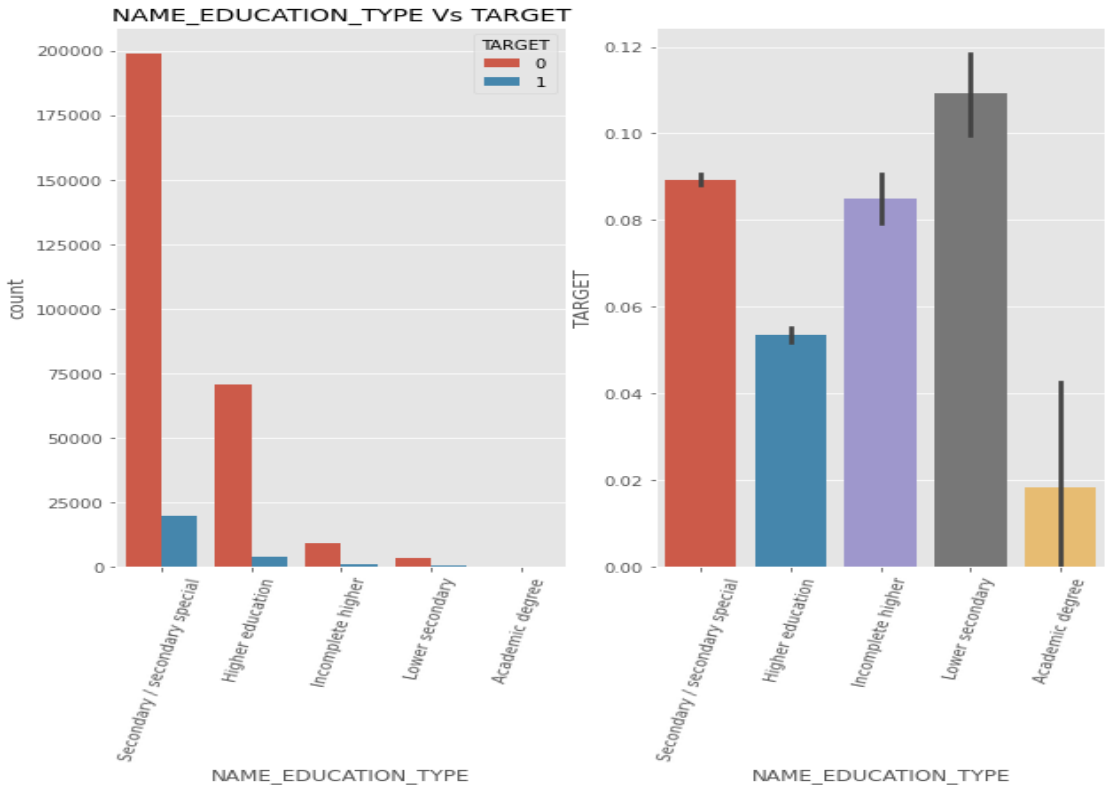




As the AGE increase Client is mst like to do less faulty, client below than 30 years of age likely to do more fault

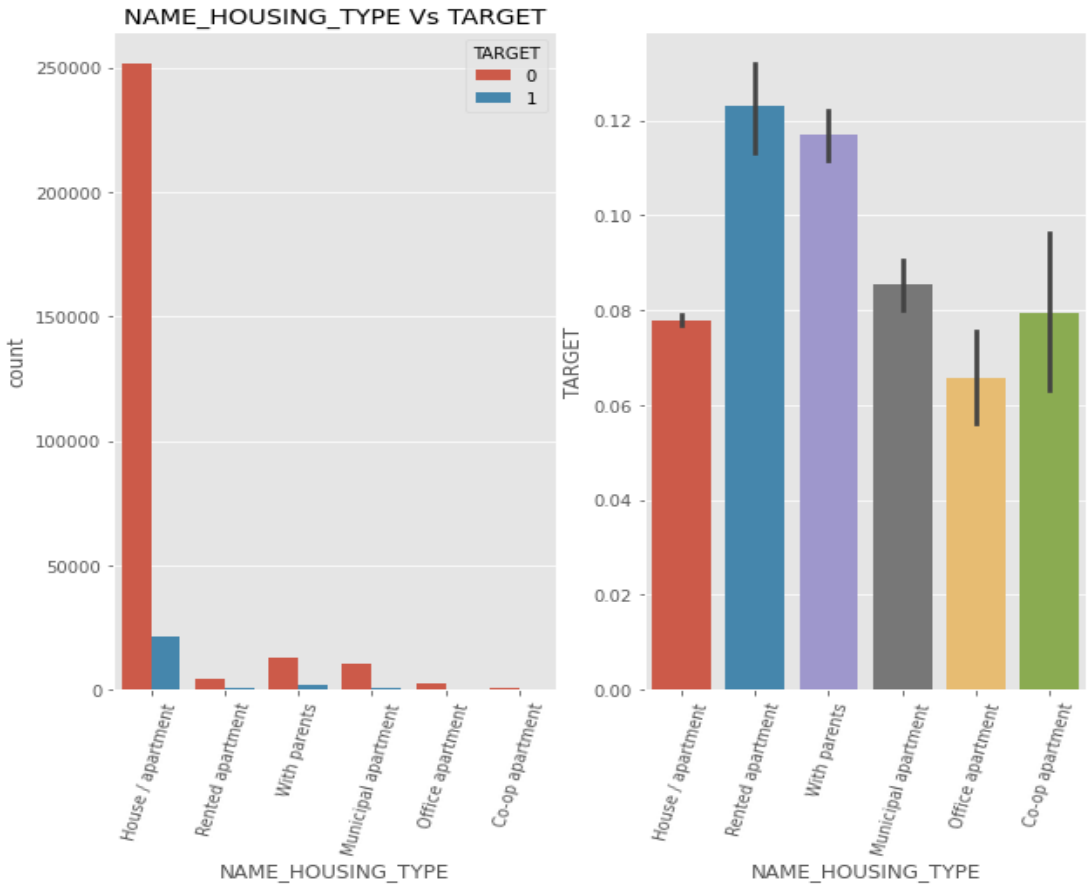
client having less experience are more like to be at fault, as
the experince increase clinet more likely to repay loan

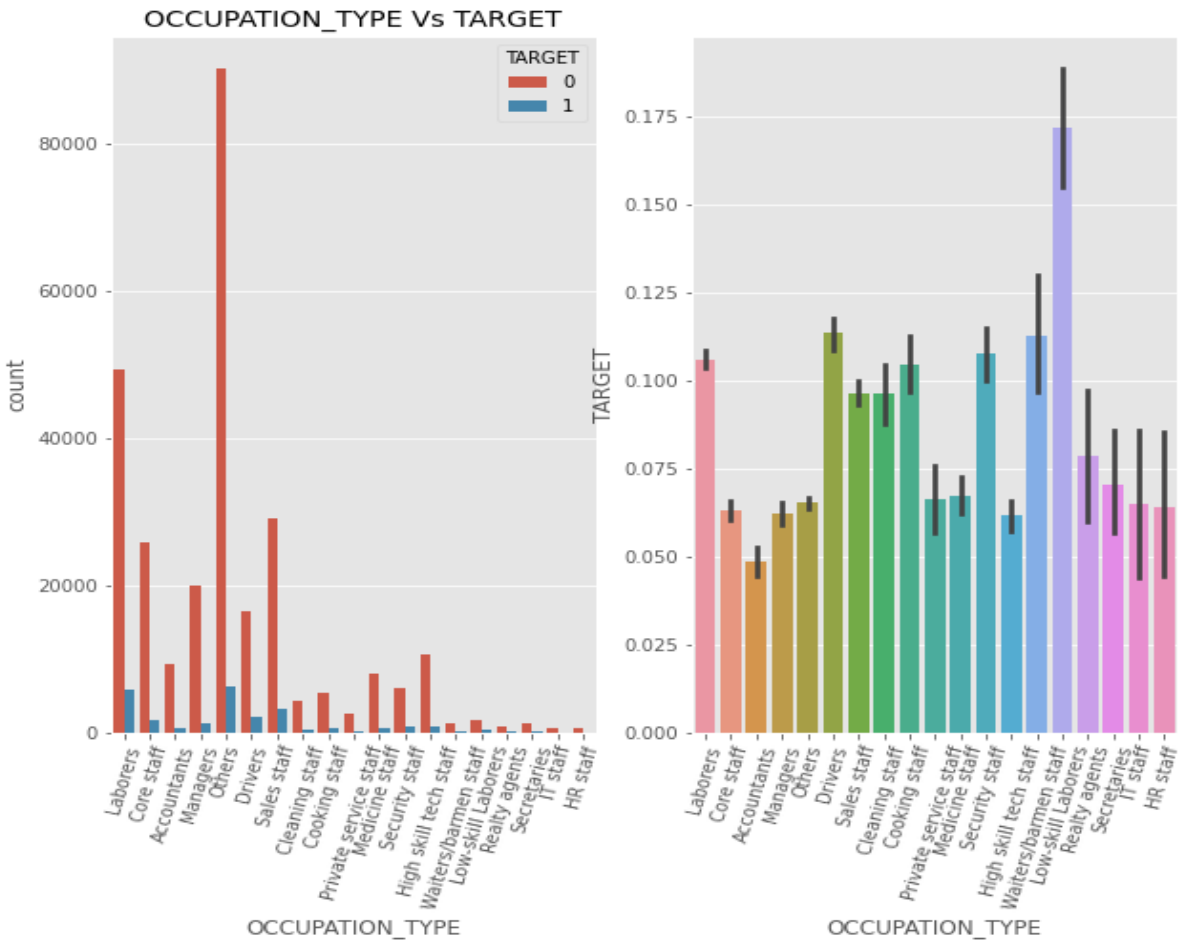




Client who are Single/unmarried and have done Civil Marriage are Defaulter

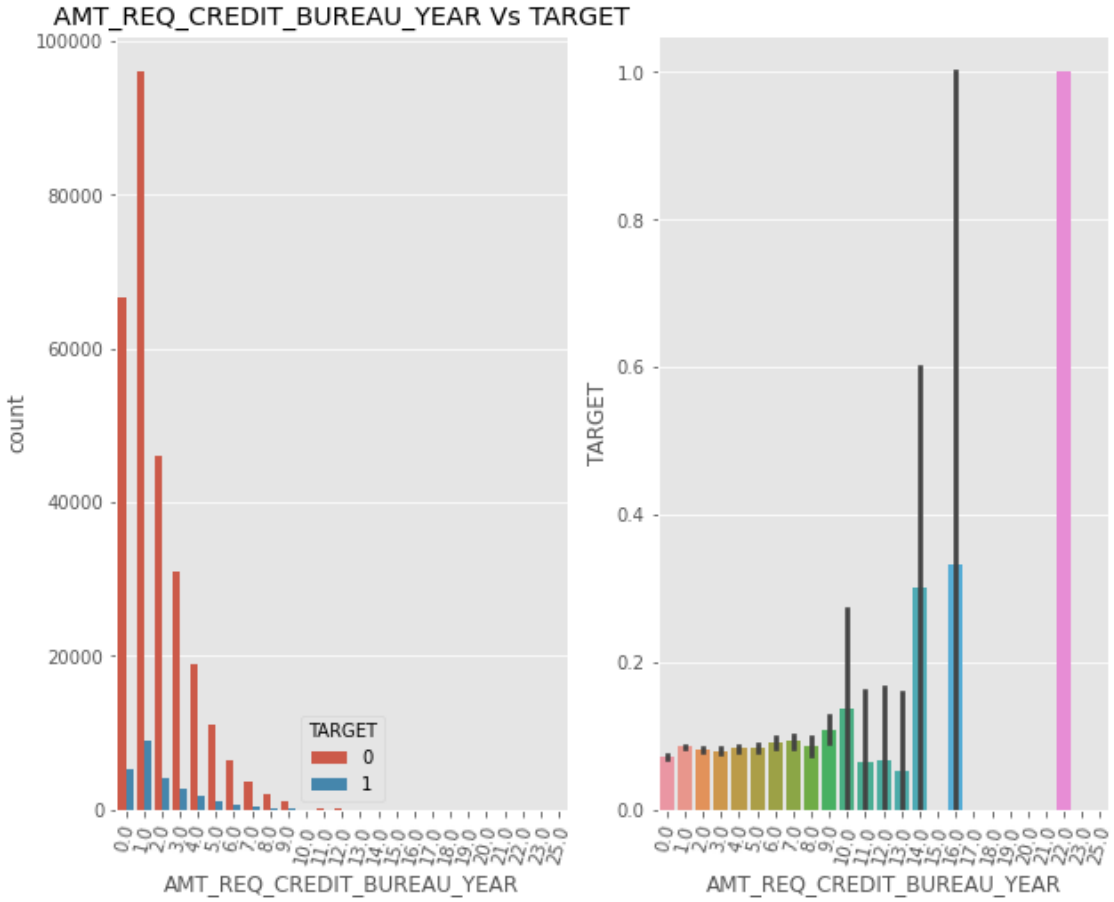
As the Education level increase clients are more likely to repay loan





Client who live in Office Apartment and have own Apartment or House are most likely to repay. whereas Clients who are on Rent can be a Defaulter

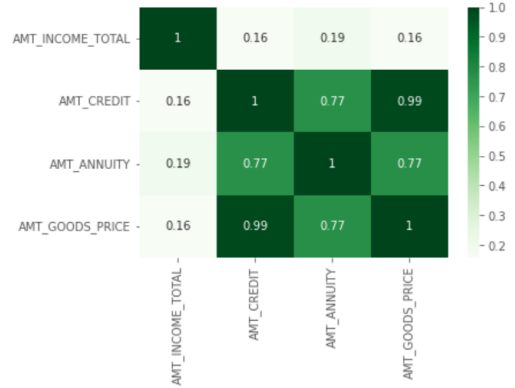
laborers class have higher defaulter rate Accountants and manager are most likely to repay



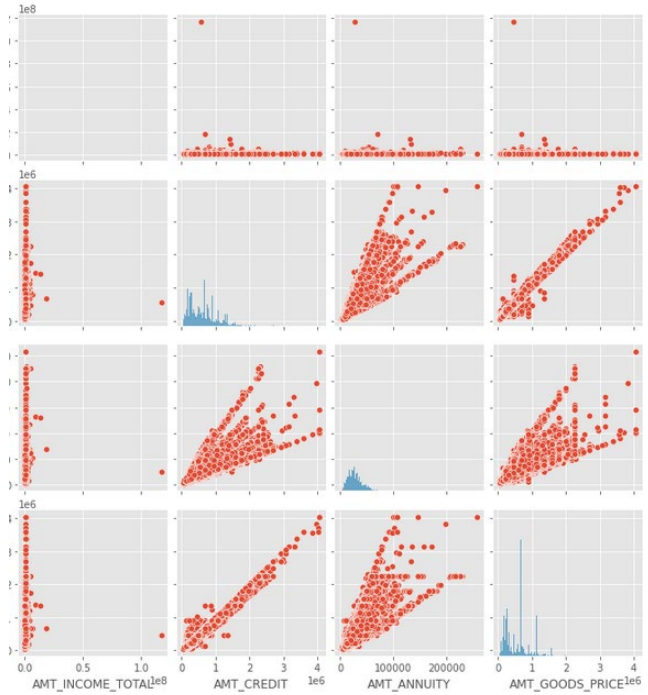
AMT_REQ_CREDIT_BUREAU:: Higher the number of request higher the chance of client to be at Fault

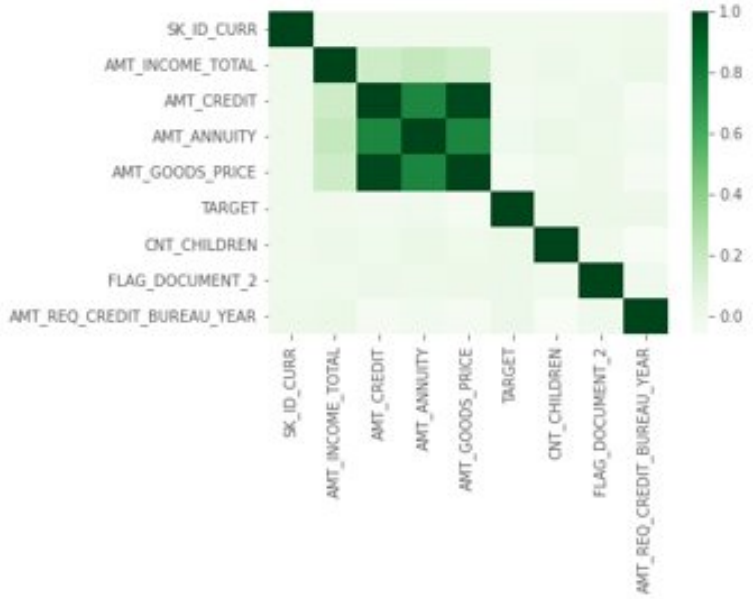
Pair Plot and Heat Plot

Inferences: Very high correlation between AMT_CREDIT and AMT_GOODS_PRICE - Applicants owning goods of high value can take loans of higher amounts.

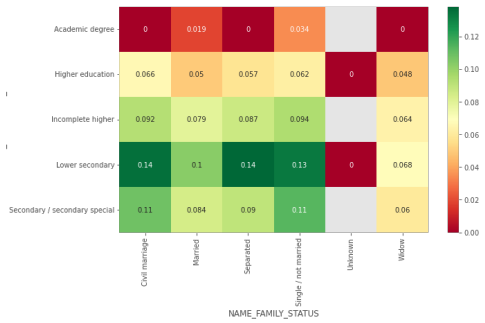


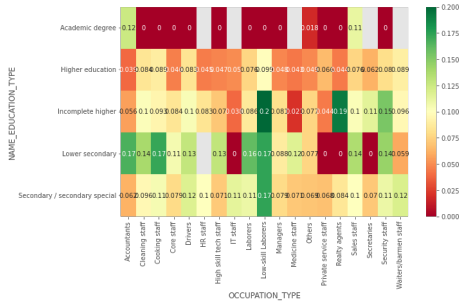
AMT_CREDIT have high corelation with the AMT_GOOD_PRICE()0.99 and with Amt_ANNUITY(0.77

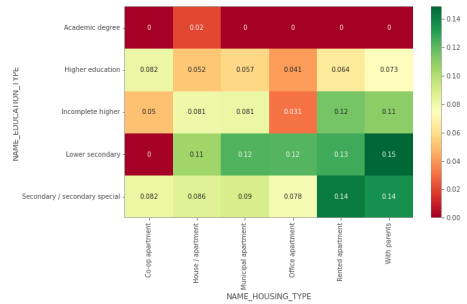


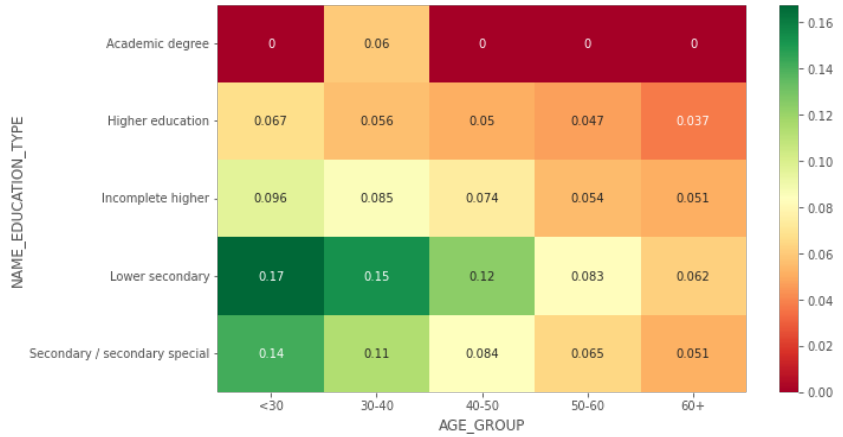


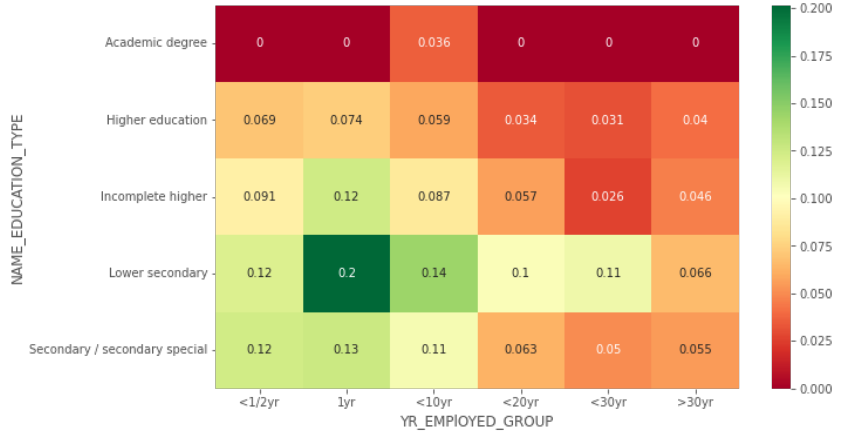
MULTI VARIATE





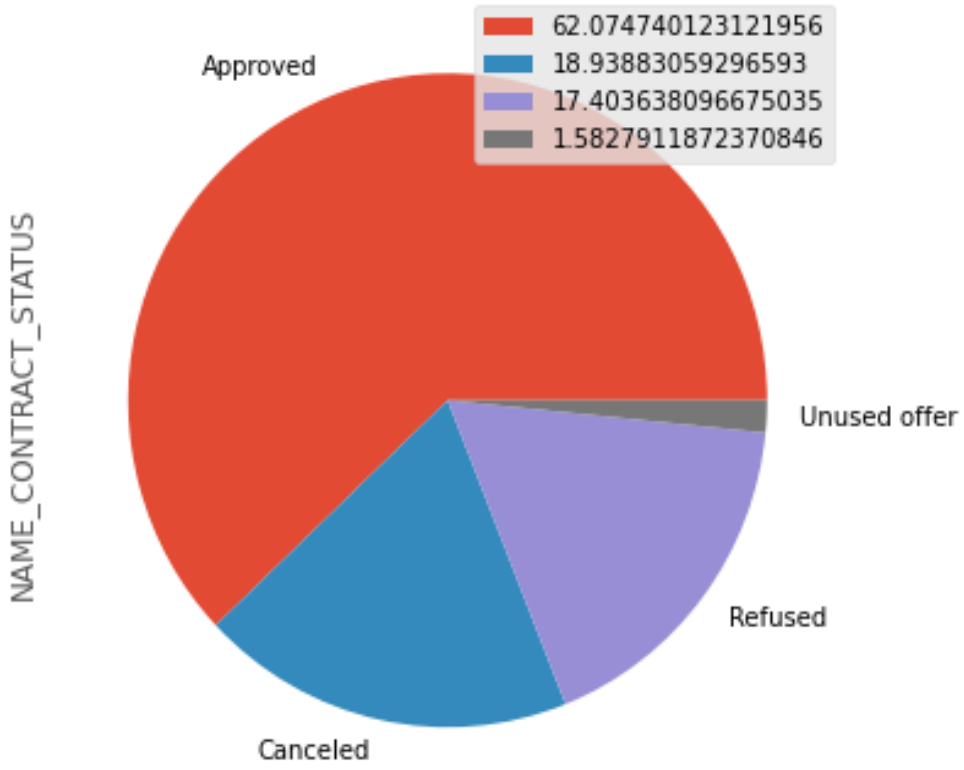


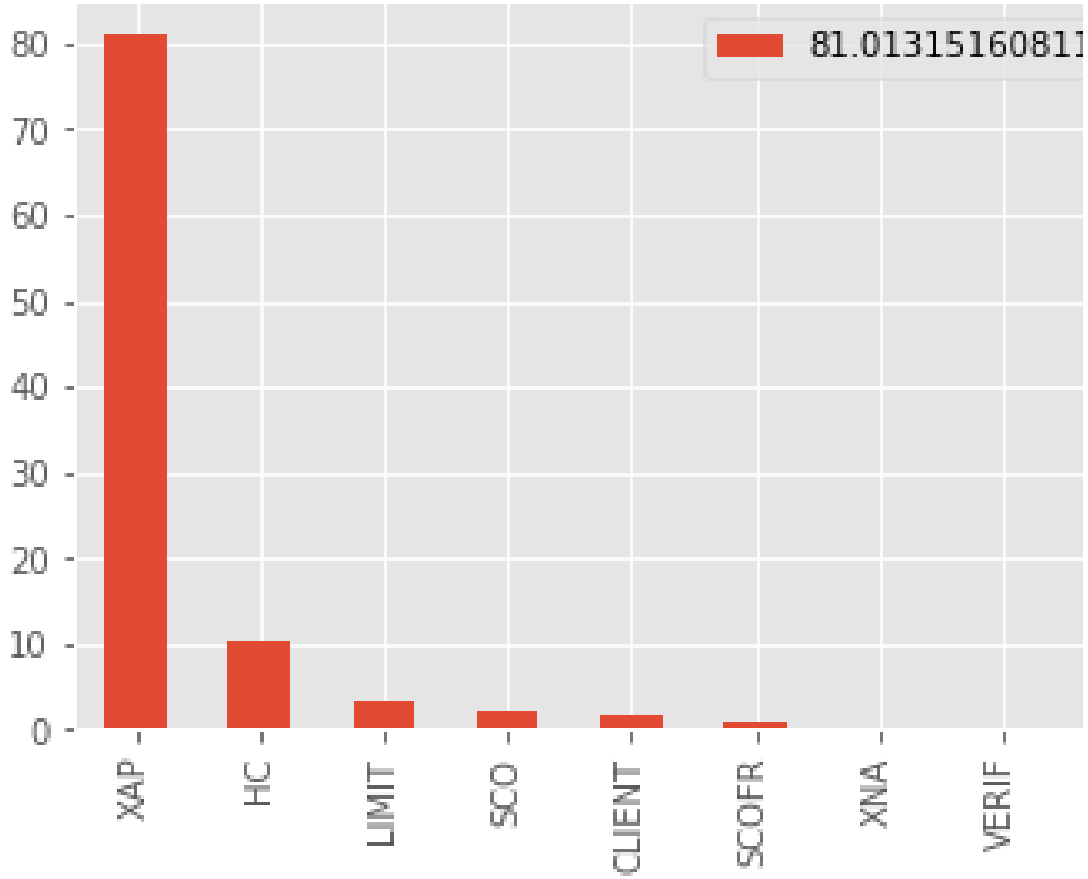




ANALYSIS ON PREVIOUS DATA FRAME

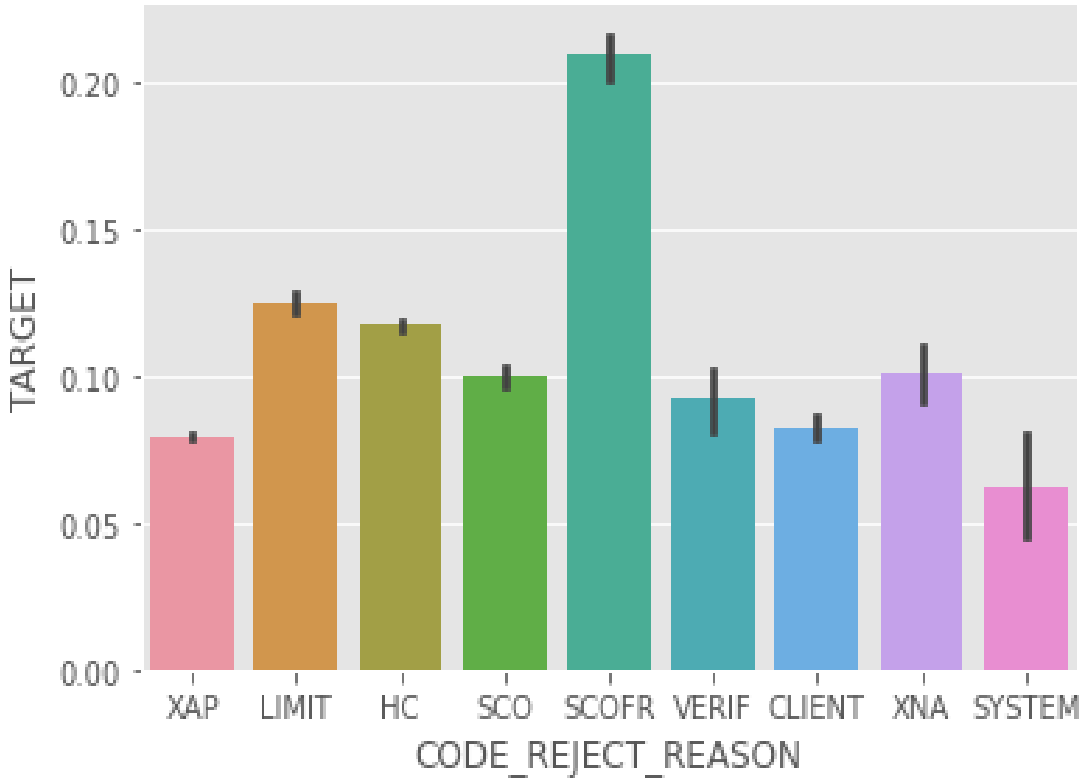
PREVIOUS DATA BASE



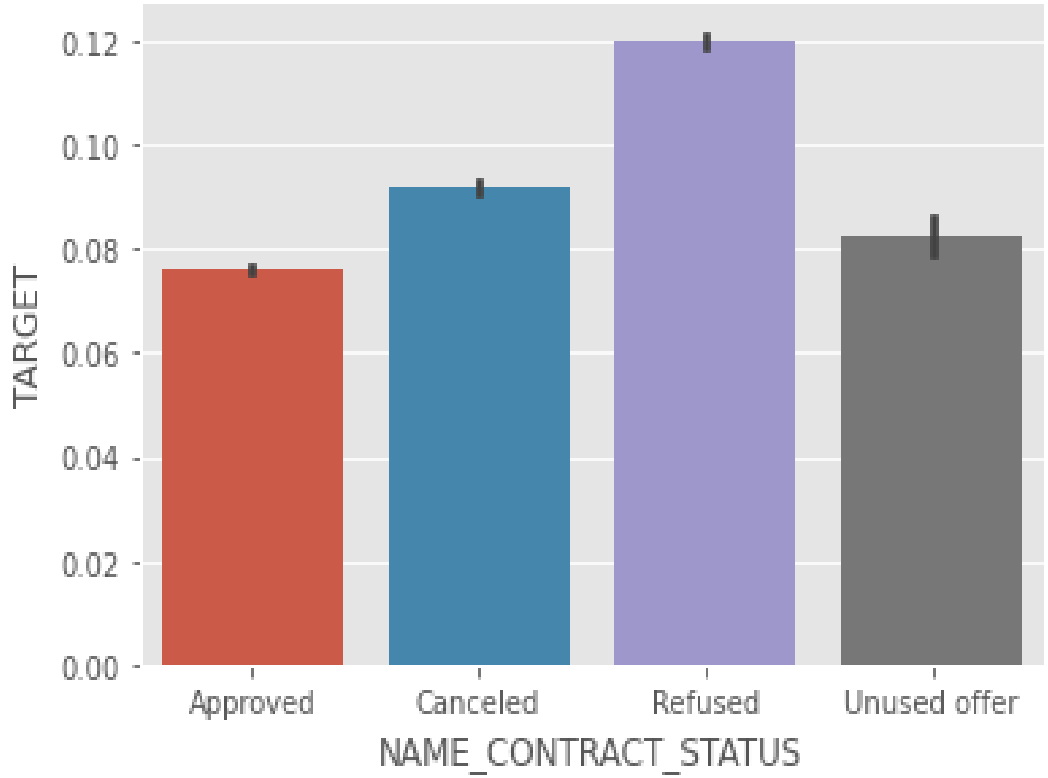


ANALYSIS AFTER MERGING

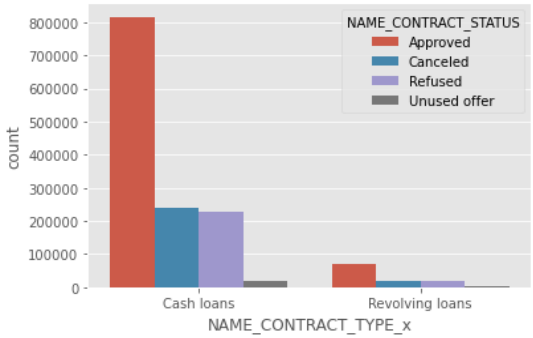
Infer After Merging the DataFrames

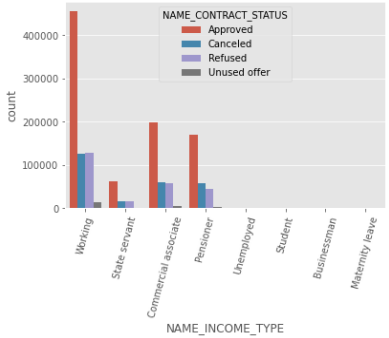


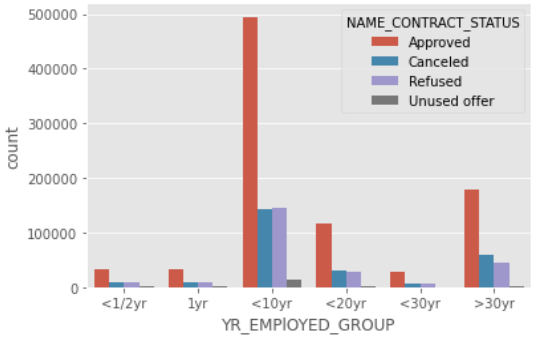
Client whose loan got reject due to SCOFR are the highest defaulter

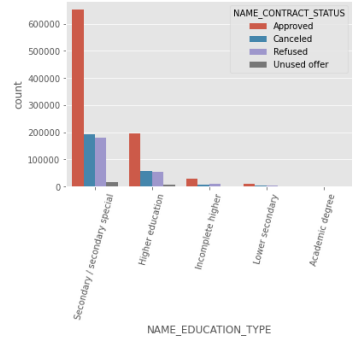


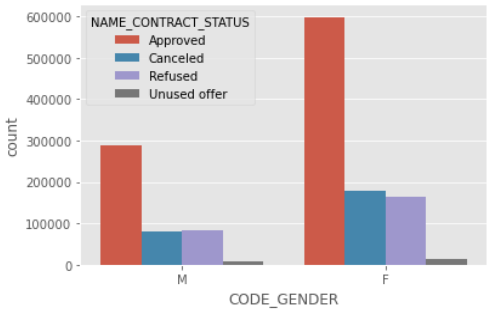
Client whose Loan got Refused and canceled earlier are high defaulter

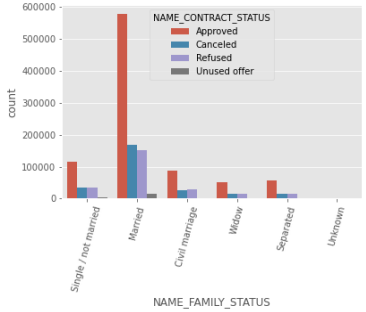


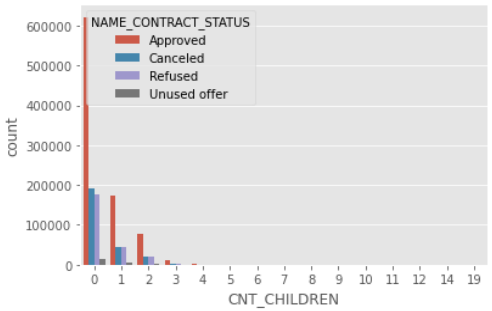


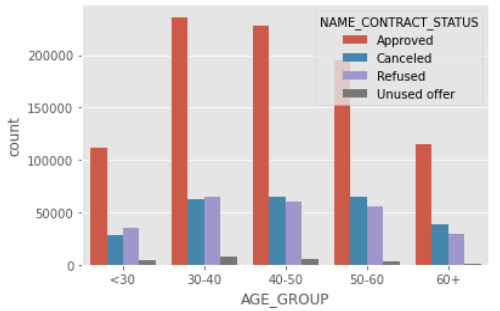


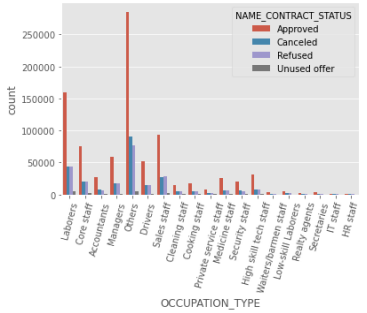












whatever is analyzed in the Application data Frame it is same i got after the merging, same non defaulters approved rate are high

CONCLUSION

- As compare to female Males are at defaulter
- Client who have less kids are more like to repay
- As the AGE increase Client is most like to do less fault, client below than 30 years of age likely to do more fault
- client having less experience are more like to be at fault, as the experience increase client more likely to repay loan
- Client who are UNEMPLOYED to be defaulter are very high
- As the Education level increase clients are more likely to repay loan
- Client who are Single/unmarried and have done Civil Marriage are Defaulter
- Office Apartment ppl are most likely to repay.
- Laborer class are higher defaulter rate Accountants and manager are most likely to repay
- AMT_REQ_CREDIT_BUREAU:: Higer the number of request higher the chance of client to be at Fault.
- Client whose loan got reject due to SCOFR are the highest defaulter
- Client whose Loan got Refused and canceled earlier are high defaulter