CREDIT EDA CASESTUDY

DONE BY:
THARUN TEJ REDDY THODIMI

UNIVARIATE ANALYSIS on Application Data

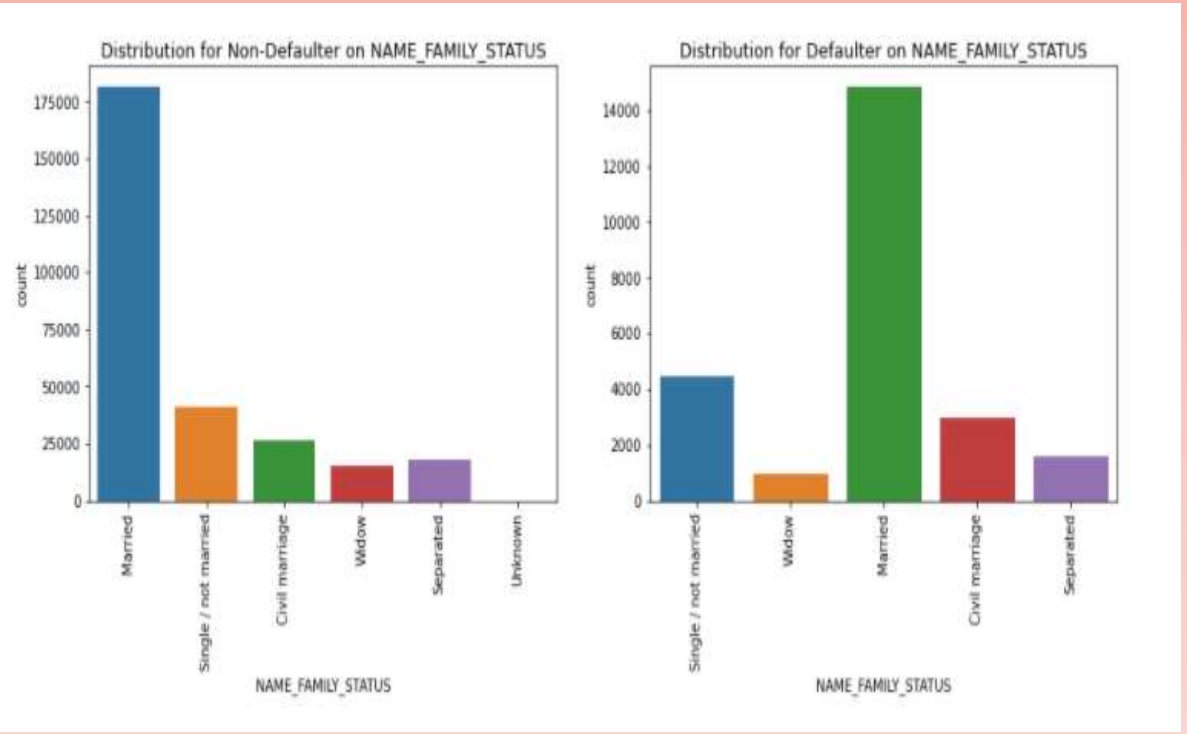
- Categorical Unordered UNIVARIATE analysis
- Categorical ordered UNIVARIATE analysis

UNIVARIATE Analysis helps in fetching insights from single variable which helps in Overall Analysis

Distribution on FAMILY STATUS –

Family status of the client

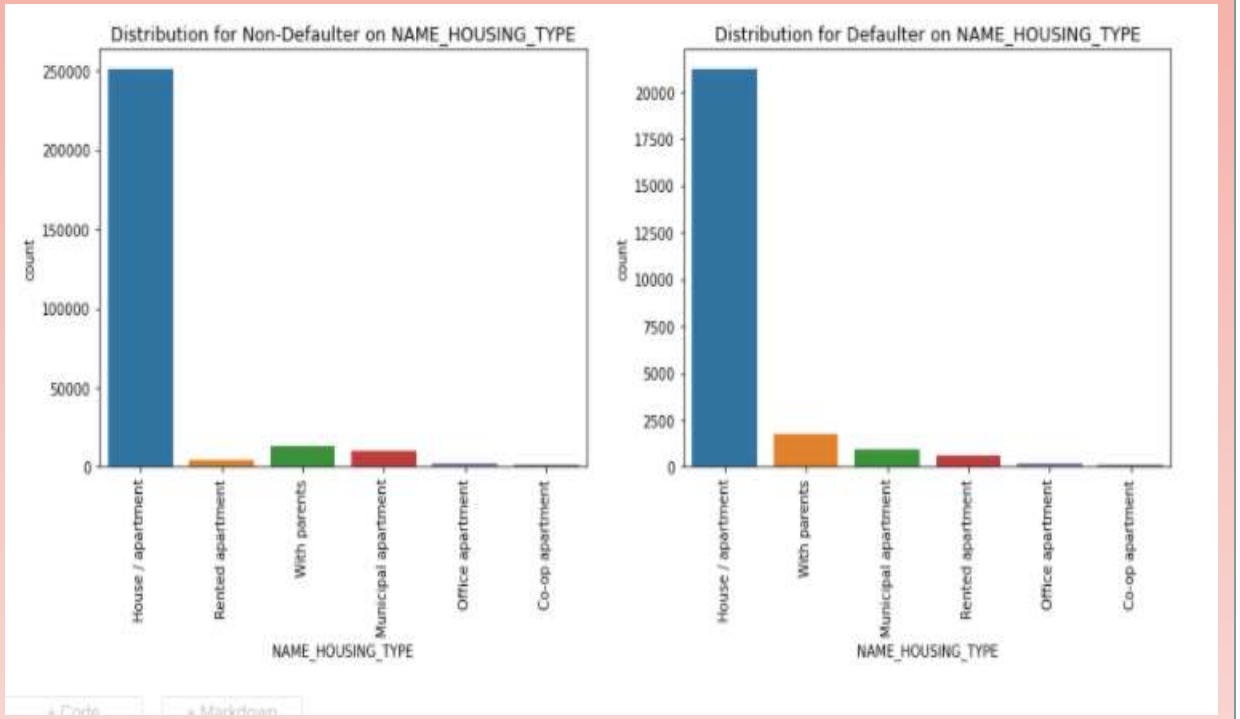
- Proportion of Married customers falling in default category is high when compared with all other categories .
- Single/Not married category has higher proportion count falling under default when compared with non-default count



Distribution on Housing Type –

Housing situation of the client

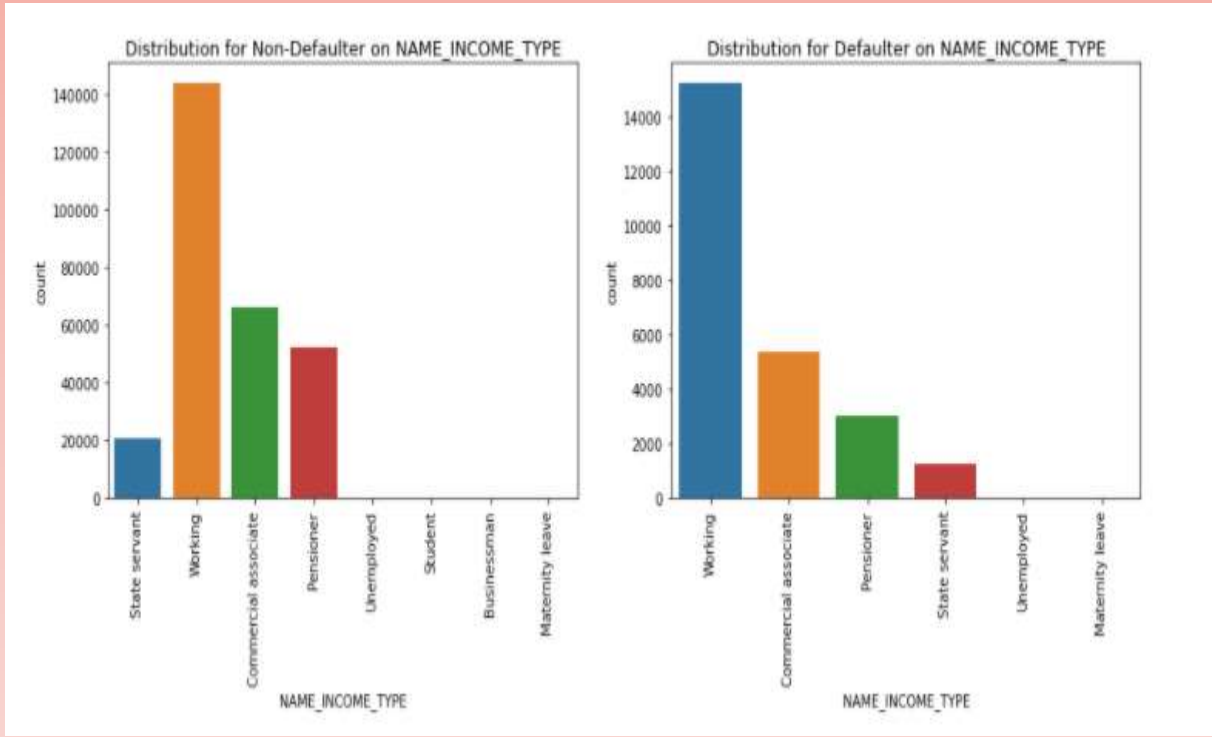
- House/Apartments category has a highest category of customers walking into bank for loans
- Rental apartment category has more defaulters than non-defaulters, Its a clear insight in the real world that there monthly expenses are going with house rents and which may lead to fall in defaulters list, This needs to be considered while giving loans
- If we observe the scale of the plots even with parents category has higher chances in falling under default category
- House/Apartments category has lot of defaulters its almost 8-10% when compared with the non-defaulters count



Distribution on Income Type –

Clients Income type

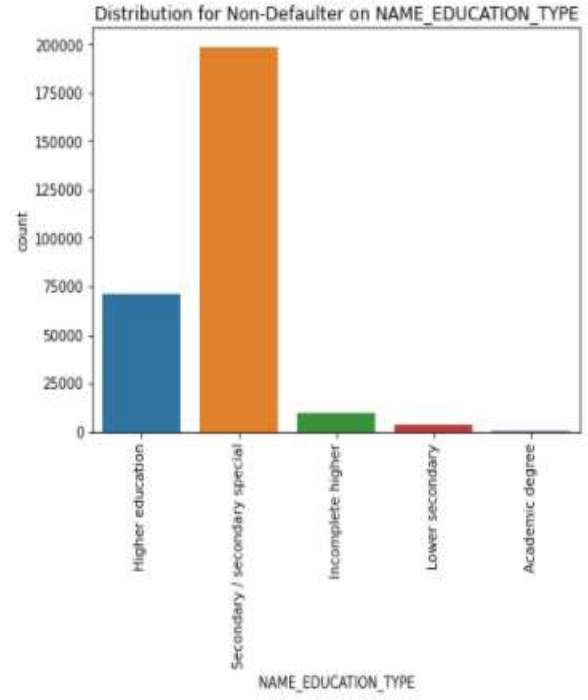
- The data implies almost same behaviour with all the categories with respect to defaulters and Non-defaulters
- On Concentrated observation we could see that percentage of Pensioner being in default is less when compared with other categories. So bank can concentrate in this category to generate profits by taking certain steps like reducing loan amount

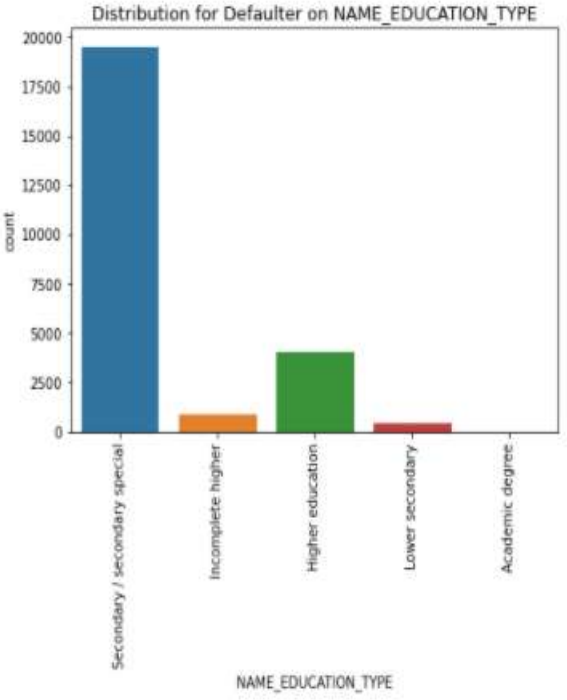


Distribution on Education Type –

Clients Highest Education type

- Academic degree looks more profitable for banks since the Defaulter percentage is very less in this category when compared with other categories.
- As the education level increases the Defaulters count is decreasing, This is quite realistic that customers might have settled with certain jobs and able to Repay the loans.
- Higher Education clearly implies less Defaulters to the bank. So bank can concentrate on giving loans accordingly.
- Minimal education level clearly implies that there is a chance of loss incurring lossed from those category of people





Distribution on Gender Type –

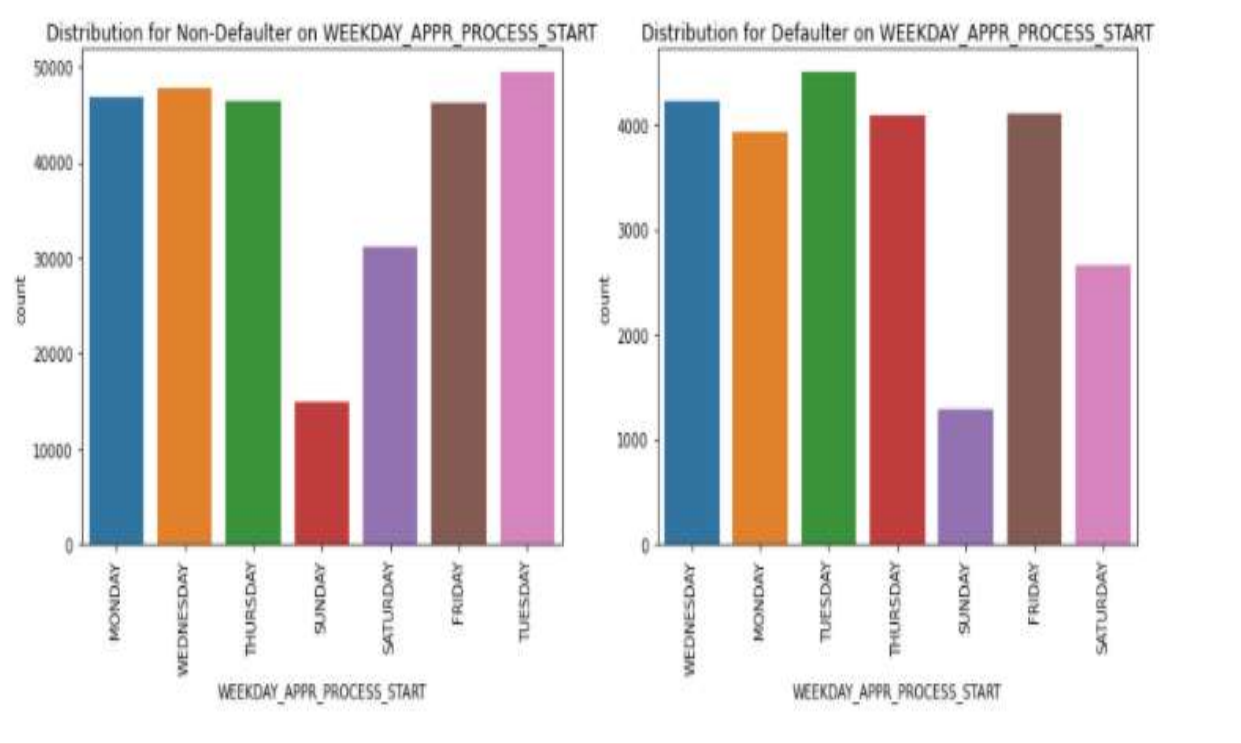
Type of Gender client Belongs to.

- From the two plots shown here we could observe that Female category in the dataset is twice as male category. In result to that we have plots mentioning that in defaulter and non-defaulter Female category has a high Count. Further analysis is done under bivariate analysis to infer results from GENDER category.

Distribution on Start Of WeekdayApproval –

The day on which clients applies for the loan.

- Day on which loan has been processed doesn't really imply the outcome/Target
- As it clearly implies real world cases Sunday and Saturday where loan Process is less and this doesn’t much imply in our analysis in finding Defaulters.

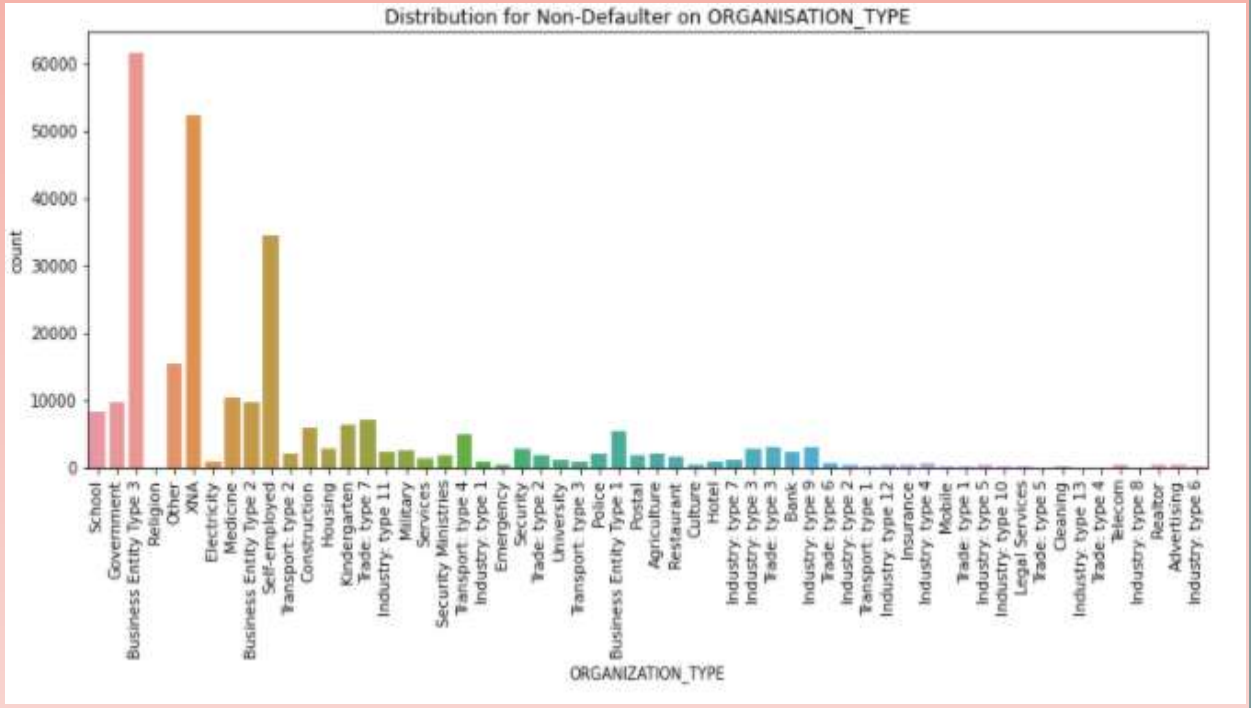


Distribution of Non-Defaulter on Organisation Type –

Type of Organisation client Belongs to.

- We could observe that most of the categories is having 10 Percent defaulters when compared with non-defaulters.
- Business Entity Type3 and `self-employed` has higher percentage of defaulter count
- Business Entity Type 1 has more defaulter count when compared with overall percentage of that category

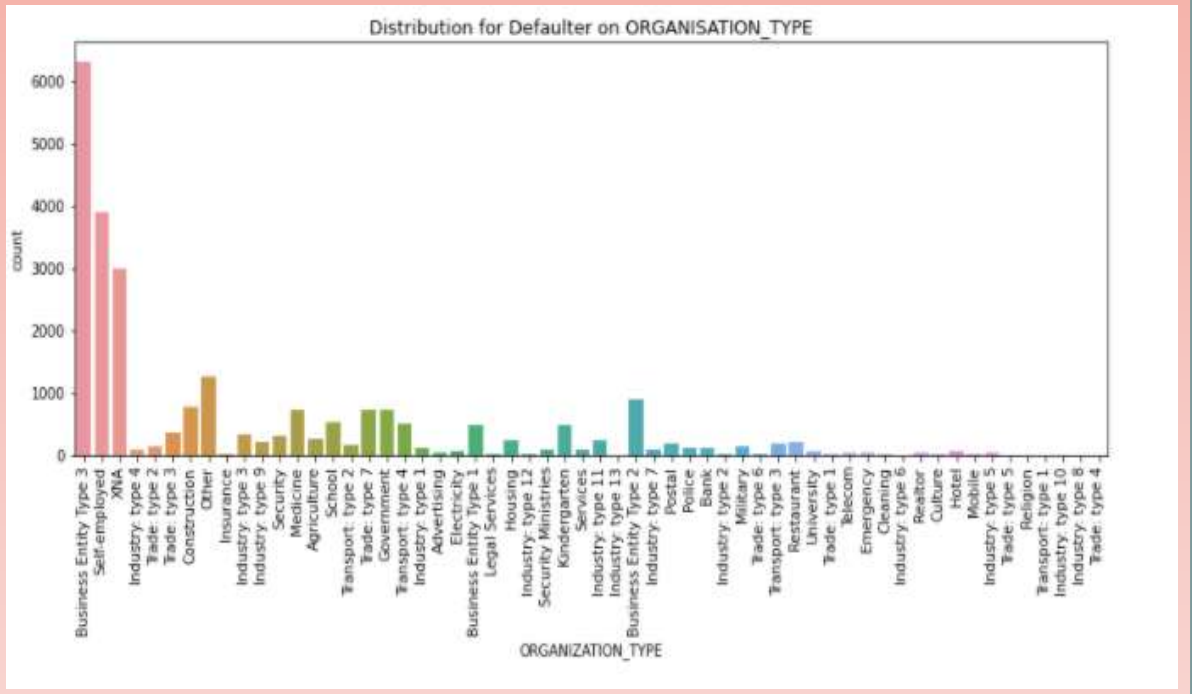
Plot-I on Non-Defaulters Category



Plot-II on Defaulters Category

Distribution of Defaulter on Organisation Type –
Type of Organisation client Belongs to.

- We could observe that most of the categories is having 10 % Defualters when compared with non-defaulters.
- Business Entity Type3 and `self-employed` has higher percentage of defaulter count
- Business Entity Type 1 has more defaulter count when compared with overall percentage of that category



BI-VARIATE ANALYSIS on Application Data

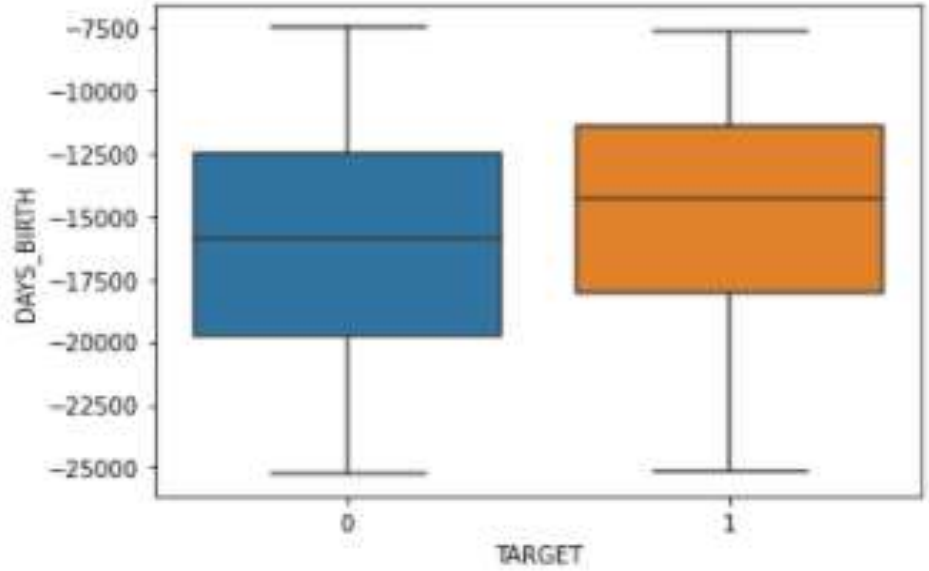
- Numeric – Numeric analysis
- Numerical – Categorical analysis
- Categorical – Categorical analysis

BIVARIATE Analysis helps in fetching insights by looking into multiple variables which helps in gaining Overall Insights from the hidden data

Age *versus* Target –

Client Age versus Defaulter and Non-defaulter.

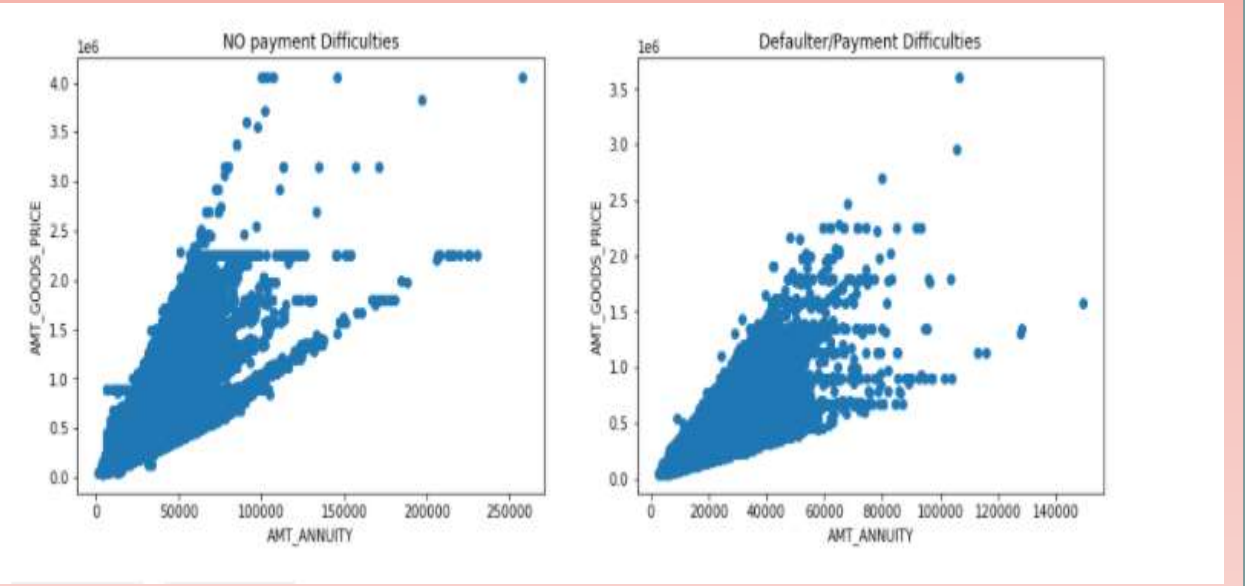
- As we can see that Younger age group has more Payment difficulties when compared with older age group
- We can also observe that Age group greater than 47 (17500/365) has very less payment difficulties
- Considering this we also saw that Customer belonging to Pensioner Category is beneficial for bank.



Amount Annuity *versus* Amount Goods Price –

Amount Annuity versus Amount Of Goods Purchased

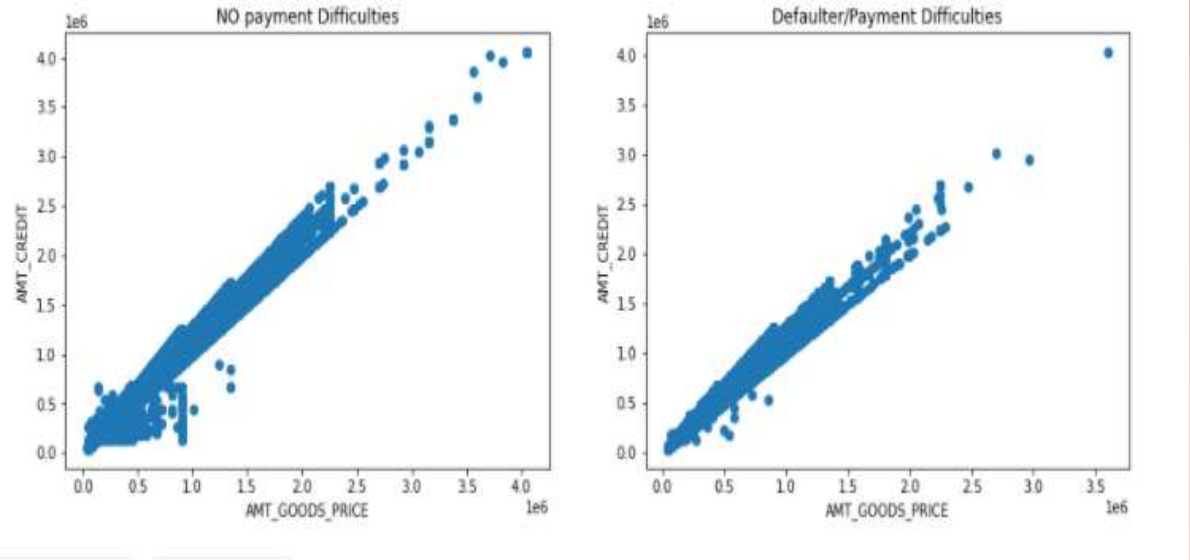
- This is quite Obvious that sum of Amount ANNUITY(Term repayments) is equal to the loan amount and our plot depicts the same
- When both plots scales are compared it says that the AMT_ANNUITY scale is less than 140000, when compared with No payment difficulties it says that the higher amount installments don't have much much defaulters which is one way profit to the bank with regular payments from customers



Amount Of Goods *versus* Amount Credit –

Amount Goods versus Amount Credit

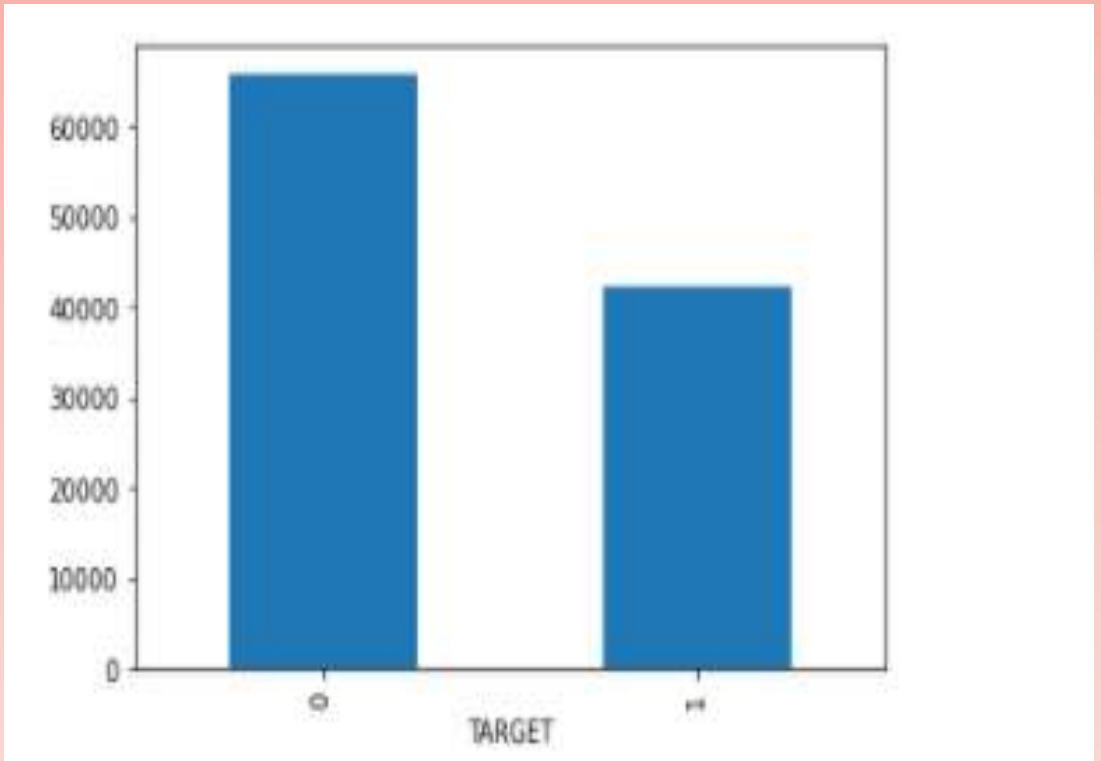
- 'AMT_GOODS_PRICE' and 'AMT_CREDIT' are linearly related
- As observed from defaulters list there are some customers under defaulter category at higher Goods price. Its better to analyse other variables and reduce the loan amount to such customers to reduce the loss to the bank



Distribution on DaysEmployeed –

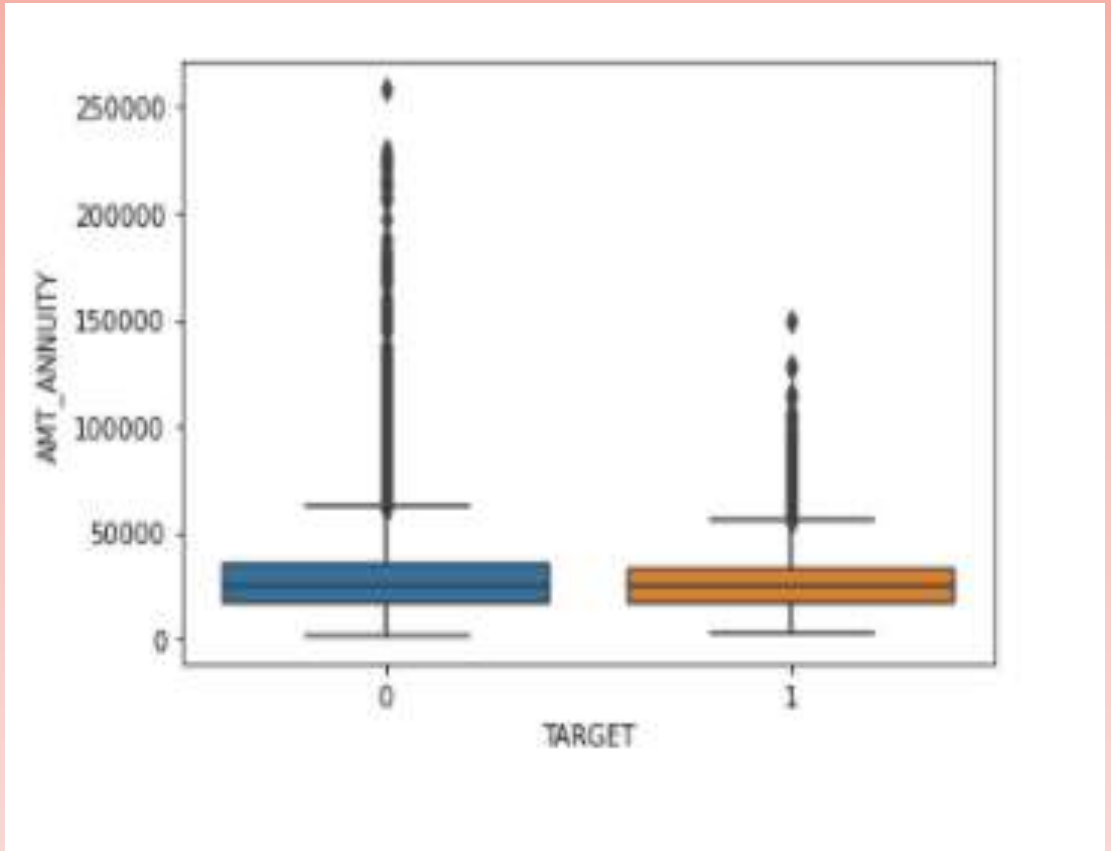
Days before the application the person started current employment

- As observed from the Plot we could see that the Customer working in the company for longer duration of time has less chance of falling under defaulter, Which can be considered while giving loan to make benefits to the bank.
- This Variable needs to be considered while giving loan, Since its implying the outcome.



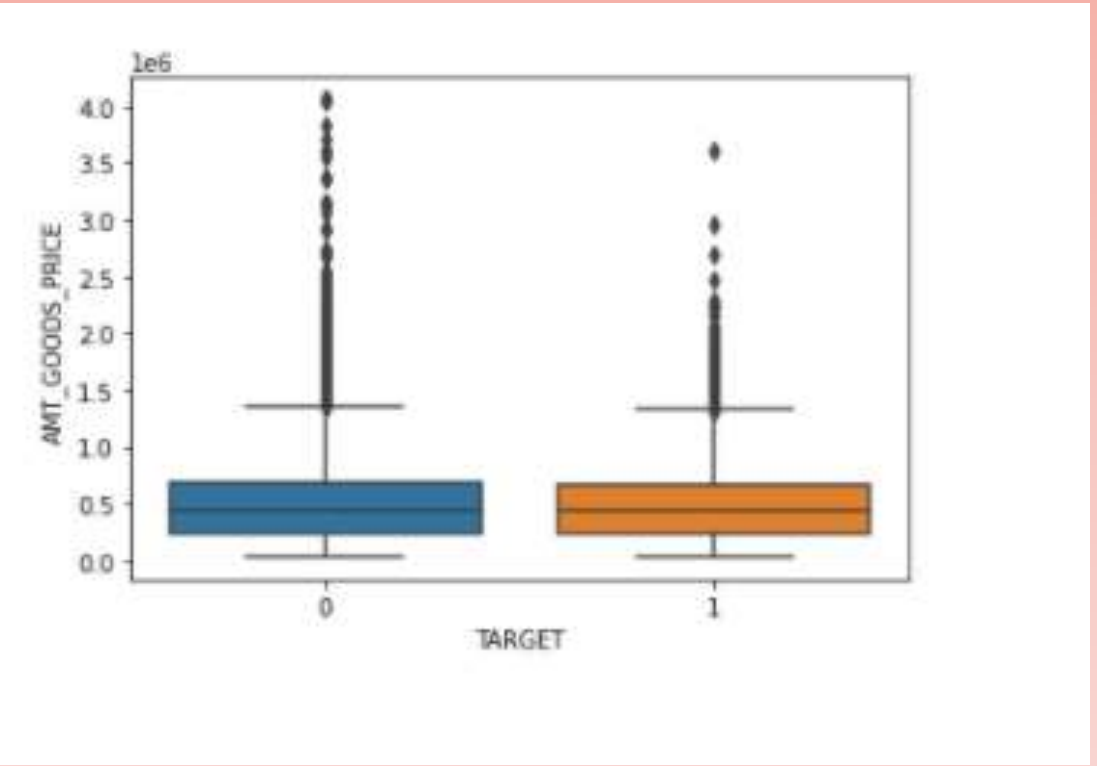
Analysing Spread of Data using BoxPlot on *Amount Annuity*

- AMT_ANNUITY tells us about the series of payments made by the customer, We could observe that Higher instalments are getting paid on time as we can see higher Annuity payments fall under non-defaulter's list which is making profits to the bank.
- But we could observe that were amount of Instalments is <150000. Bank should behave intelligently to make profits.



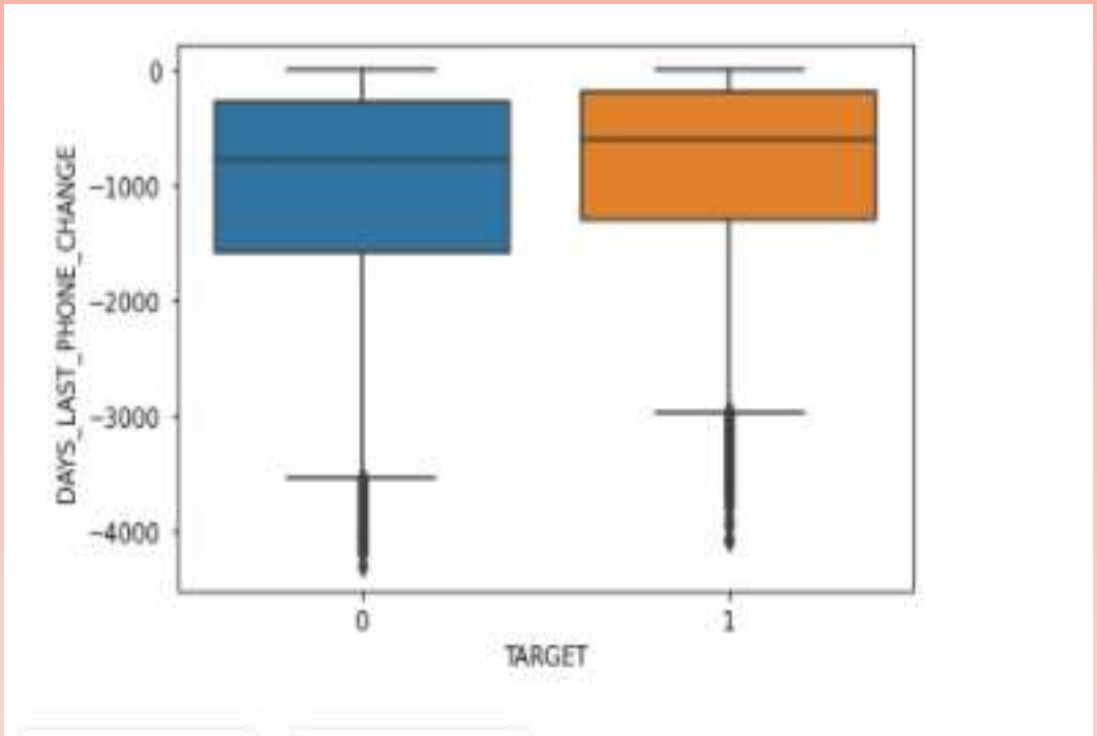
Analysing Spread of Data using BoxPlot on *Amount Goods_price*

- AMT_GOODS_PRICE tells us about the price of goods for which amount of loan is given by bank, We could observe that Higher Goods price for which loan was given is under non-default category(0) which is profit for bank.
- Plot2 shows that there are huge set of customers who fall under Non-defaulter category, Where bank is giving loans for higher amount of Goods.This should be taken care to make profits



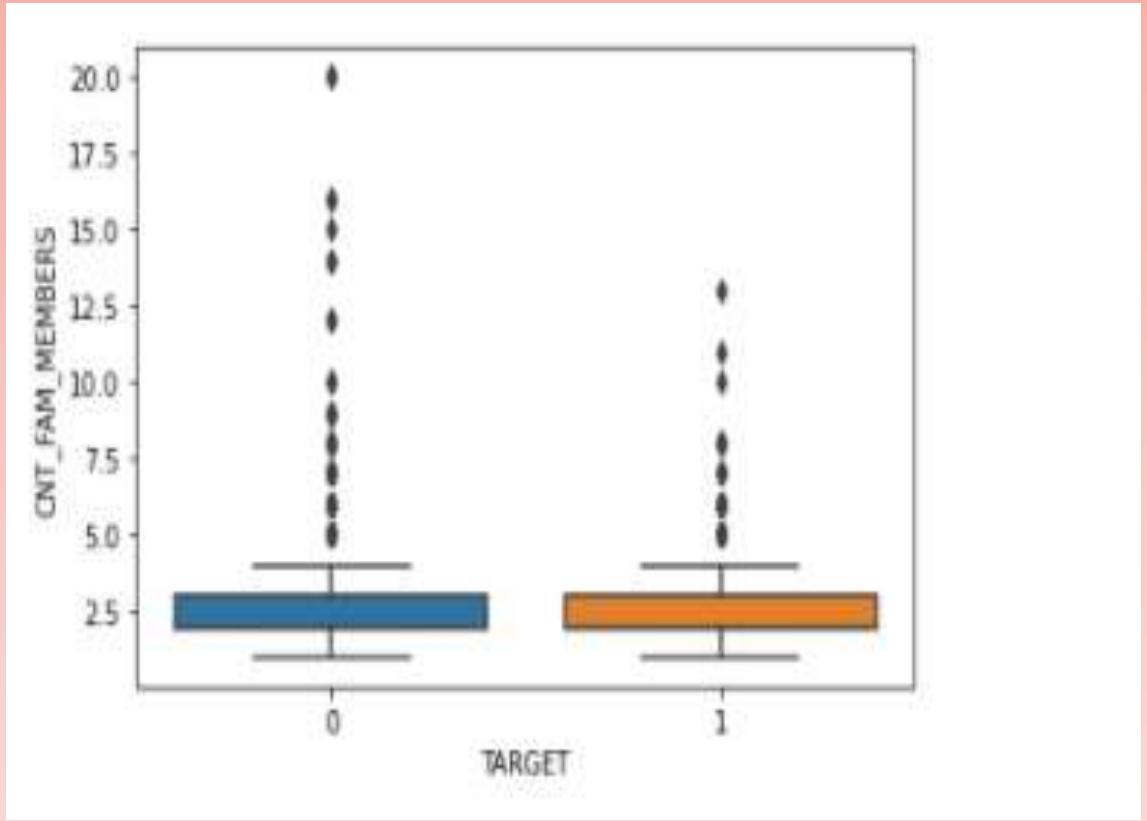
Analysing Spread of Data using BoxPlot on *DaysBeforeCustomer changes phonenumber*

- In defaulter list we could see that there is a change in the phoneNumber in the recent days falling in defaulter's list.
- Even 75th percentile is close to the loan processing day. So variable is considered for further analysis to detect Loan-defaulters.



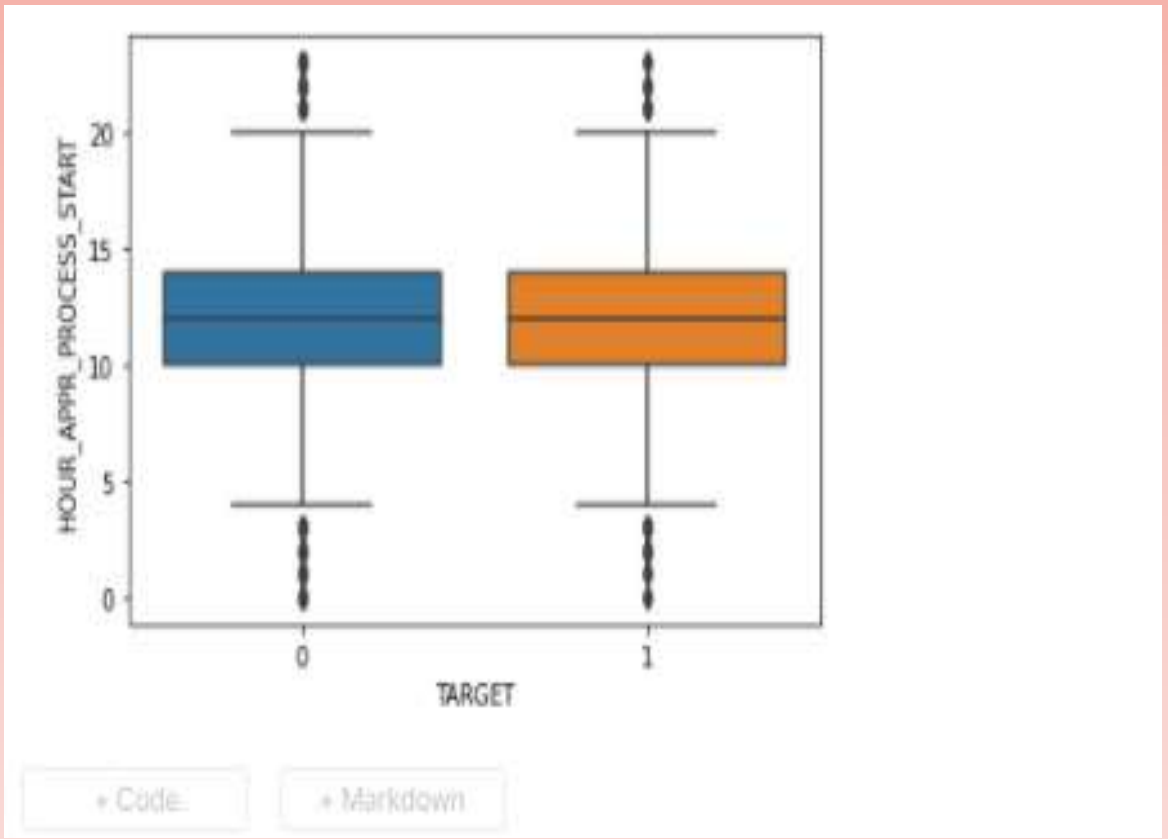
Analysing Spread of Data using BoxPlot on *Count of family members in a client*

- From the plot shown we could see that increase in count of family members doesn‘t really impact the final target variable
- But further analysis can be made using bivariate analysis. To incur more insights from this variable.



Analysing Spread of Data using BoxPlot on *Approximately at what hour did the client apply for the loan*

- As observed from the above plot ‘HoursApprovalProcess Start’ doesn't make much difference in defaulters and Non-defaulters.
- This doesn’t seriously imply in analysing defaulters or non-defaulters.



Top 10 Correlated Variables –

High correlation between variables helps in understanding relation between variables

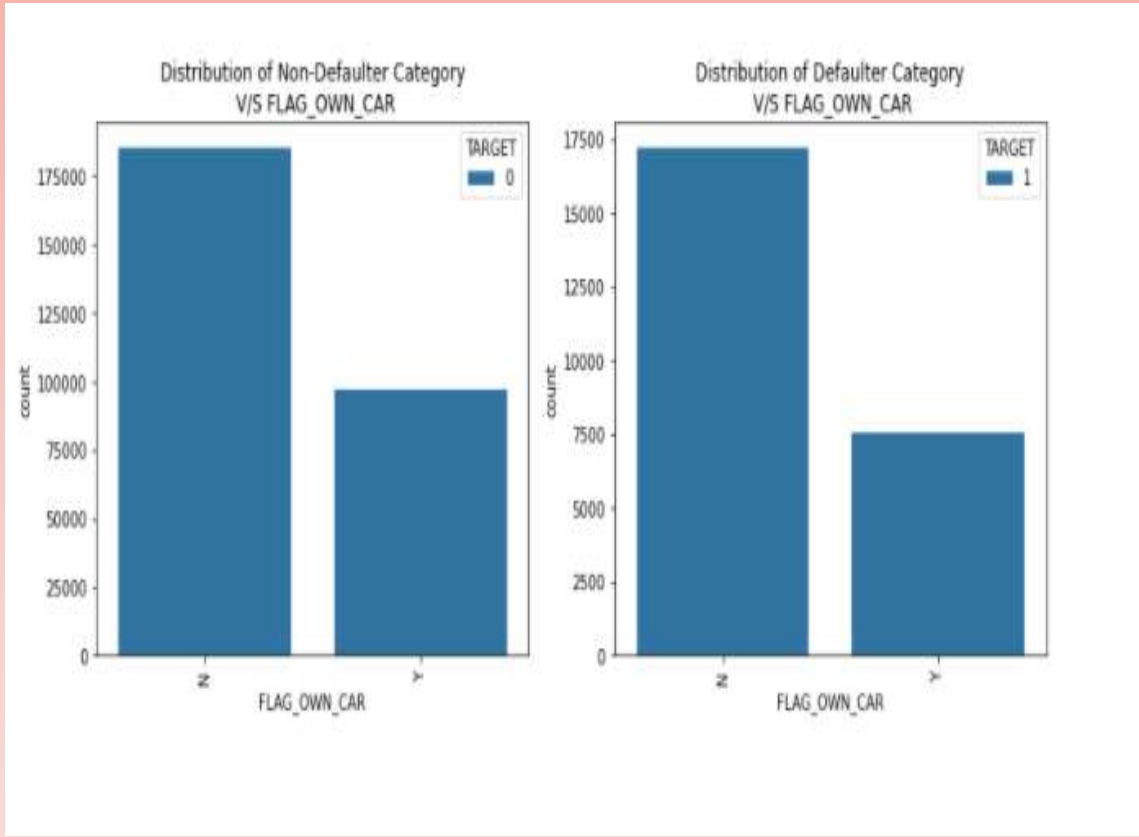
- As shown in the image, we should consider these variables on high priority in analysing whether a customer can become a defaulter/Customer having payment difficulties or Non-defaulter(All other categories)
- These variables includes both Positive Correlation and negative correlation between them.

DAYS_EMPLOYED	FLAG_EMP_PHONE	0.999755
FLAG_EMP_PHONE	DAYS_EMPLOYED	0.999755
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998490
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998490
AMT_GOODS_PRICE	AMT_CREDIT	0.986968
AMT_CREDIT	AMT_GOODS_PRICE	0.986968
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950842
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950842
CNT_FAM_MEMBERS	CNT_CHILDREN	0.879161
CNT_CHILDREN	CNT_FAM_MEMBERS	0.879161
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.860627
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.860627
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.860517
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.860517
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.825575
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.825575
AMT_ANNUITY	AMT_GOODS_PRICE	0.775109
AMT_GOODS_PRICE	AMT_ANNUITY	0.775109
AMT_CREDIT	AMT_ANNUITY	0.770138
AMT_ANNUITY	AMT_CREDIT	0.770138
FLAG_EMP_PHONE	DAYS_BIRTH	0.619888
dtype: float64		

FlagOwningaCar *versus* Target –

Client having car versus Defaulter and Non-defaulter.

- As observed percentage of customers owning a car and not owning a car is almost half, Which is quite realistic.
- Percentage of customer owning a car and falling in defaulters is quite less but doesn't imply drastic difference.



UNIVARIATE ANALYSIS on Previous Application Dataset

- Categorical Unordered UNIVARIATE analysis
- Categorical ordered UNIVARIATE analysis

UNIVARIATE Analysis helps in fetching insights from single variable which helps in Overall Analysis

BI-VARIATE ANALYSIS on Previous Application Dataset

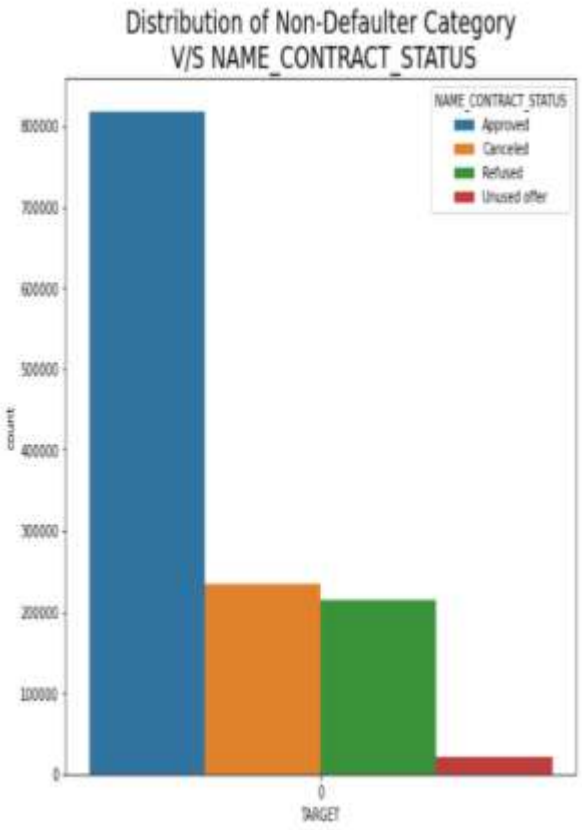
- Numeric – Numeric analysis
- Numerical – Categorical analysis
- Categorical – Categorical analysis

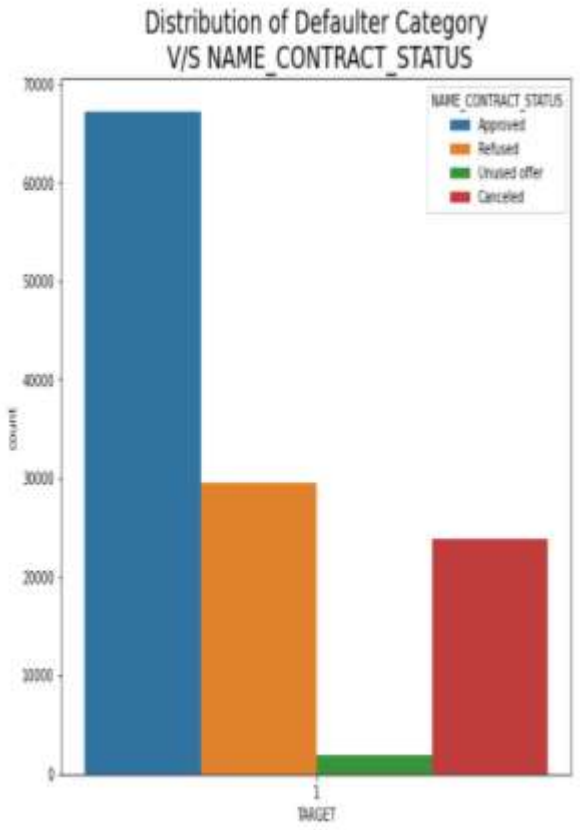
BIVARIATE Analysis helps in fetching insights by looking into multiple variables which helps in gaining Overall Insights from the hidden data

Distribution on Contract Status–

status of previous application submitted by the client

- we could see some good insights after combining both the Datasets that almost 65000(plot2) loans has been approved by the bank and which are falling under Default category, which is loss for the bank.
- Plot1 says about non-defaulter category where we could see the company has refused almost 200000 loans, but the customer is capable of repaying the loans, This should be avoided to get profits to the bank

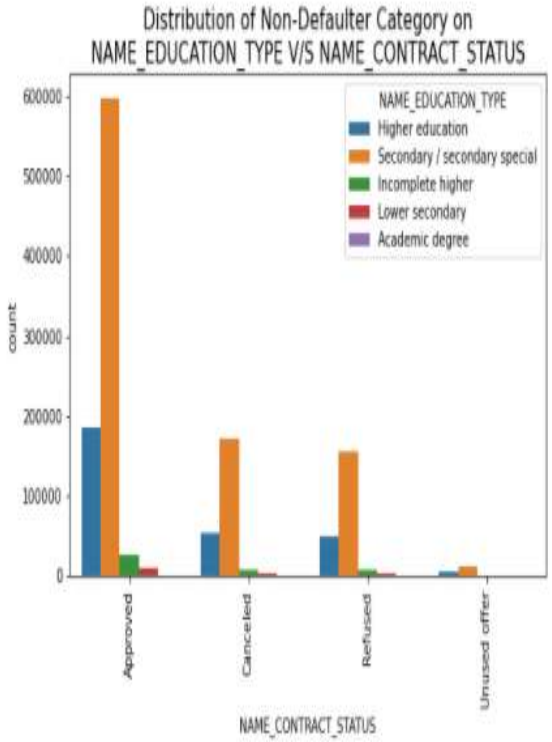


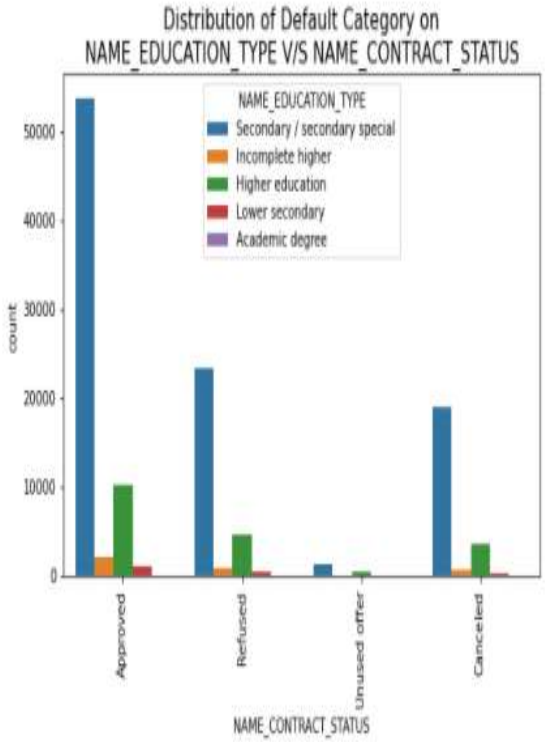


Distribution of Contract Status *versus* Education type on target –

previous application status versus Highest Education by the client

- Bank should consider ‘Secondary/Secondary special’ category since the customers are more from this category in both defaulter and non-default list.
- Education looks greater impact in the outcome.
- Higher the education less the customer being in defaulter

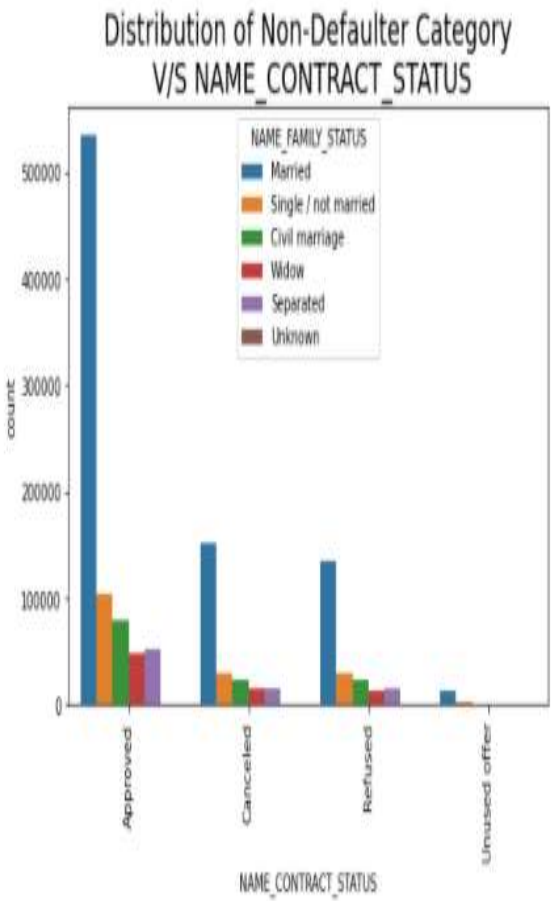


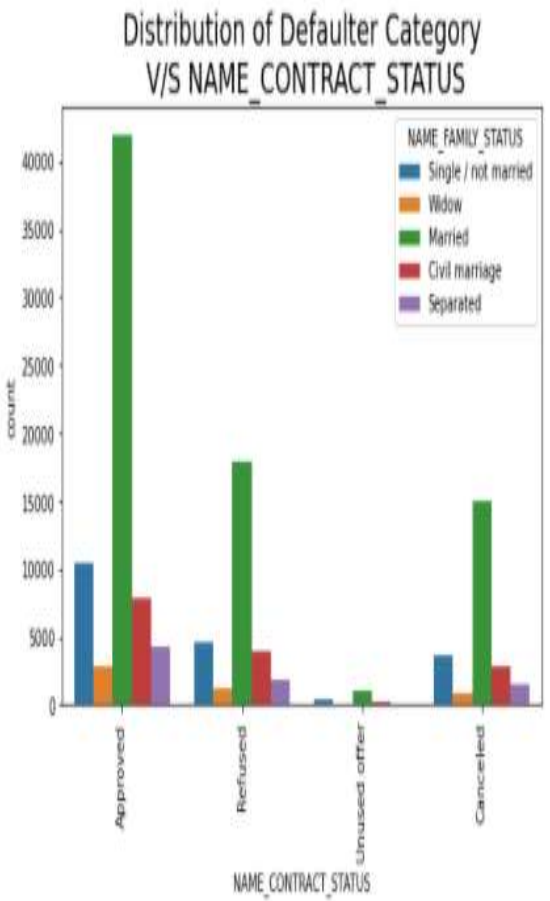


Distribution of Contract Status *versus* Family status on target –

previous application status versus Family status by the client and its impact on target

- From plot1 we can understand that some of the loans has been refused across different categories, But those category of people are capable of repaying the loan
- From plot2 we can say that ‘Married’ category has high count in refused and cancelled loans when compared with other categories

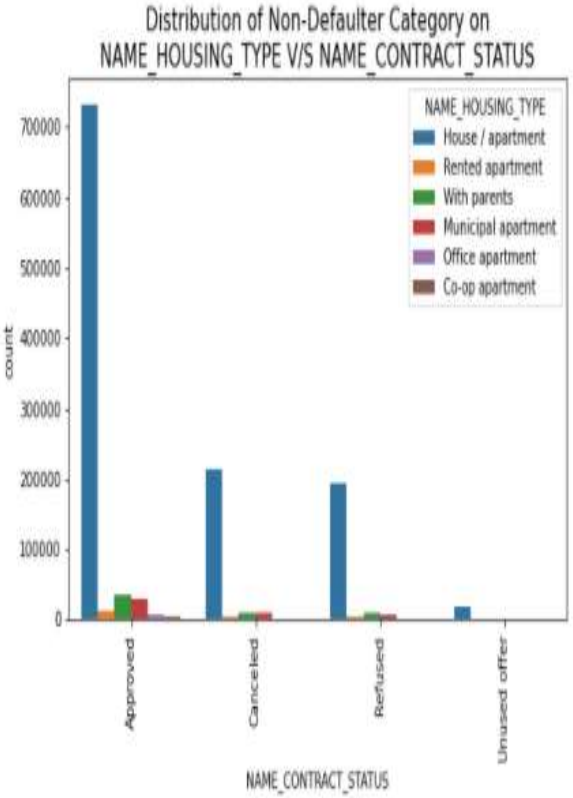


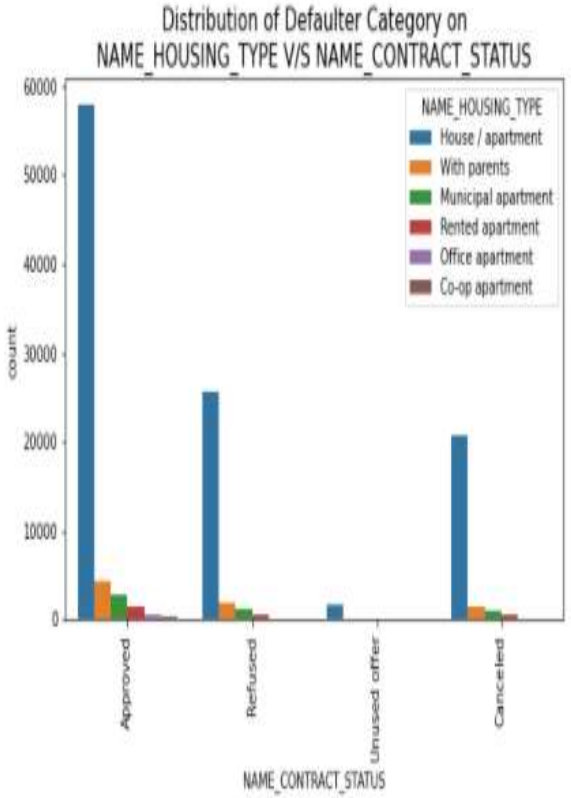


Distribution of Contract Status *versus* housing type on target –

previous application status versus housing type and its impact on target

- The bank has refused some of the loans across different categories but actually they are capable of repaying. So bank may take another steps by reducing the loan amount and pricing the loan to such customers to incur profits.
- Plot 2 shows high Loan approvals on `House/Apartment` category but the customers are falling under default category, Which results in loss for bank. So bank should consider that having 'House/Apartments' doesn't really imply that the customer will be able to repay loan on time.





CONCLUSION

- Bank should concentrate on following categories while providing loans:

- Banks can take risk in providing loans to >47 in generating profits. Since from our analysis we found that higher age group people are less in number under defaulters, When compared with other categories of age.
- Lesser the education of Client, High chances of falling in Defaulter's list. This can be considered to gain benefits.
- Employees staying in their current jobs for longer period of time, Chances of falling under defaulter is less, This should be considered while providing loan.
- By analysing the data we observed that clients belonging to Apartment/House are falling under defaulters which needs to be considered to incur profits.
- Bank should reduce their focus on ‘working’ category where we say huge percentage of people are falling under defaulters.
- Pensioner category are good for benefitting profits.