



# **EDA CASE STUDY**

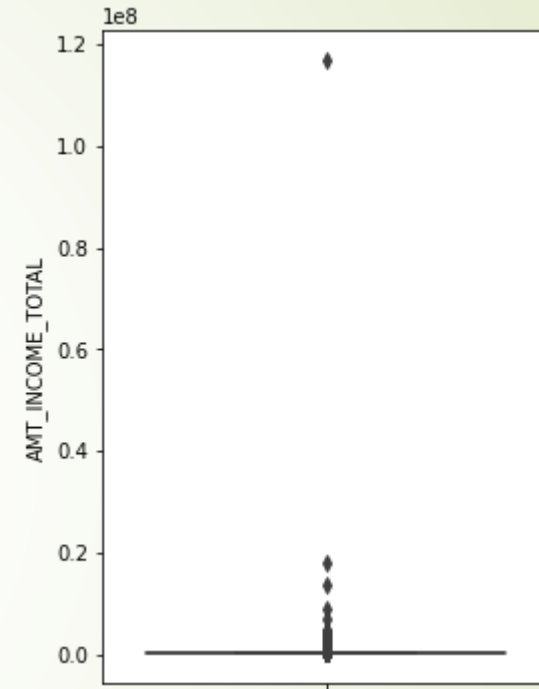
**NIMYA GEORGE**



# **OUTLIER TREATMENT / BOXPLOT ANALYSIS**

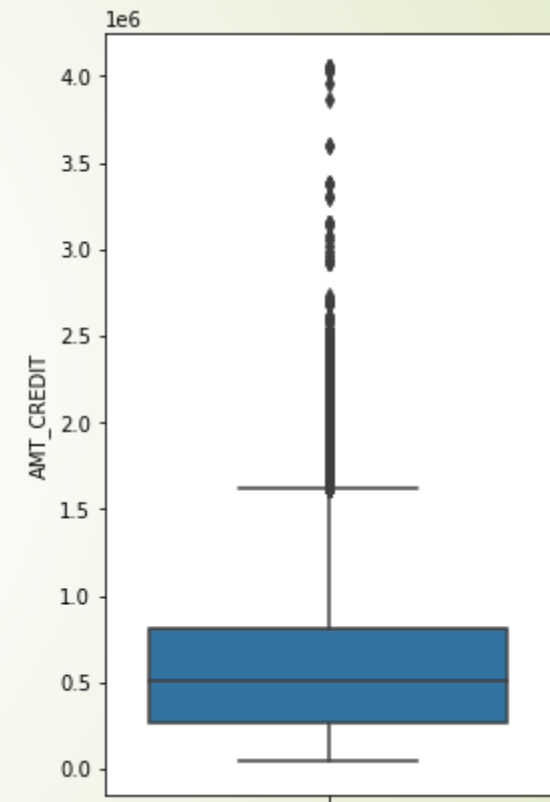
## BOXPLOT FOR AMT\_INCOME\_TOTAL COLUMN

AMT\_INCOME\_TOTAL contains outliers which can be a manual error because a labourer was having this maximum income and also he belongs to the category of defaulter. Thus this value was removed.



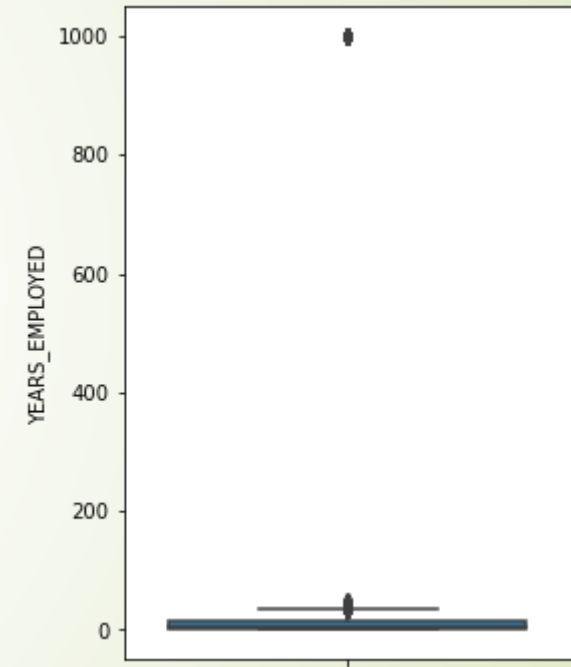
## BOXPLOT FOR AMT\_CREDIT COLUMN

In AMT\_CREDIT COLUMN there exist outliers.  
We removed outliers in order to get a better insight for further analysis.



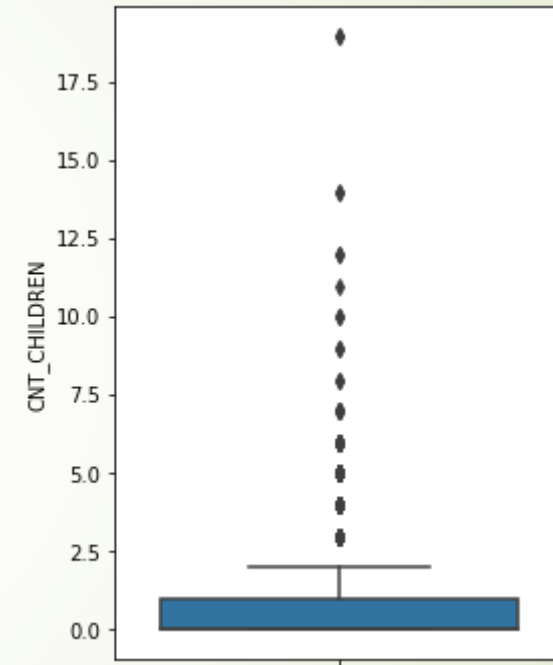
## BOXPLOT FOR YEARS\_EMPLOYED COLUMN

YEARS\_EMPLOYED COLUMN was having a value 1000.7, which was not even a possible value so that we need to treat it as an outlier and need to discard these rows.



## BOXPLOT FOR CNT\_CHILDREN

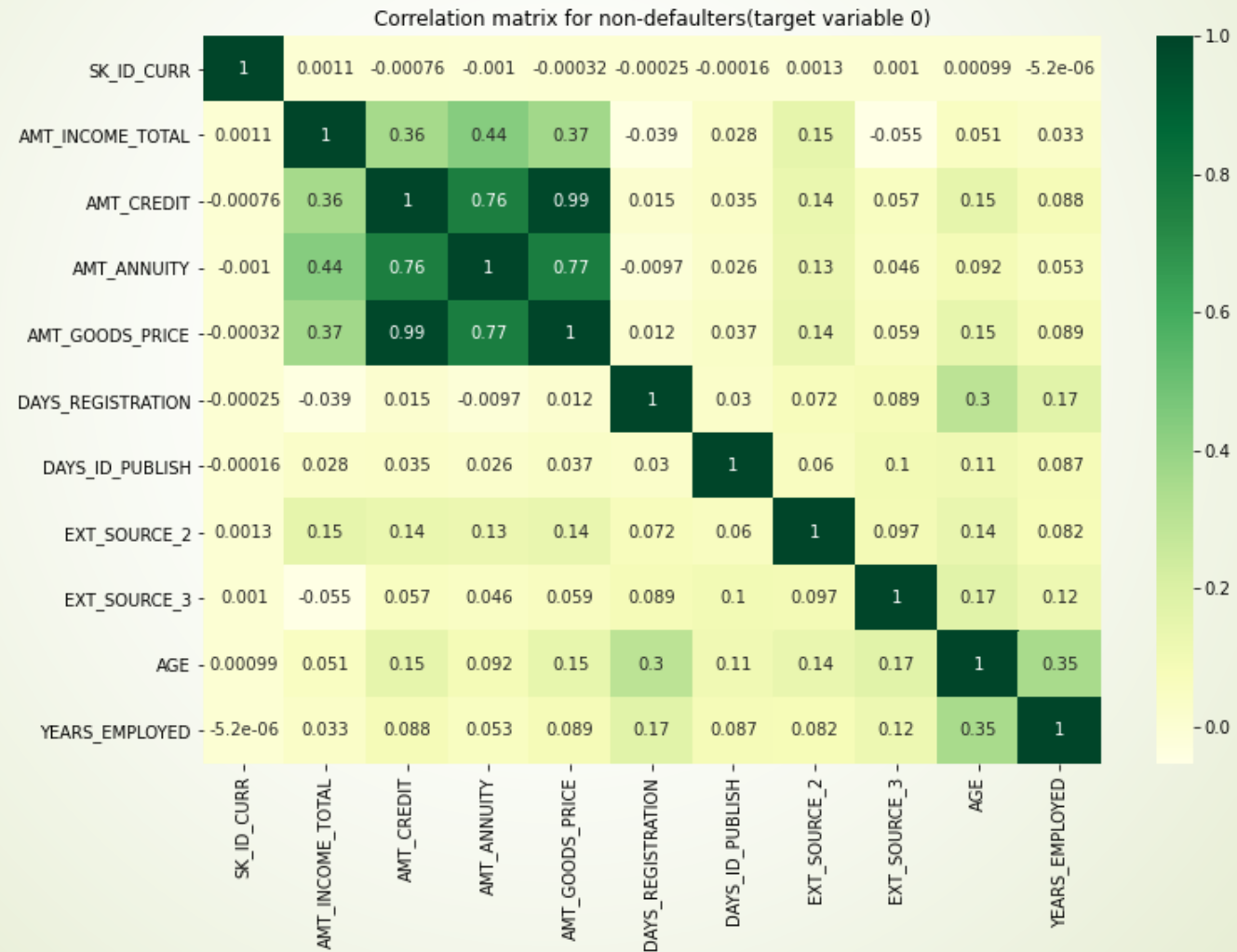
CNT\_CHILDREN have some outliers which having values up to 20+ , since it is impossible to have such number of children in normal circumstances. it might be a manual error occurred during data entry time





**FIND THE TOP 10 CORRELATIONS.**

## HEATMAP FOR NON-DEFAULTERS





## TOP 10 CORRELATIONS FOR TARGET-0 (NON-DEFAULTERS)

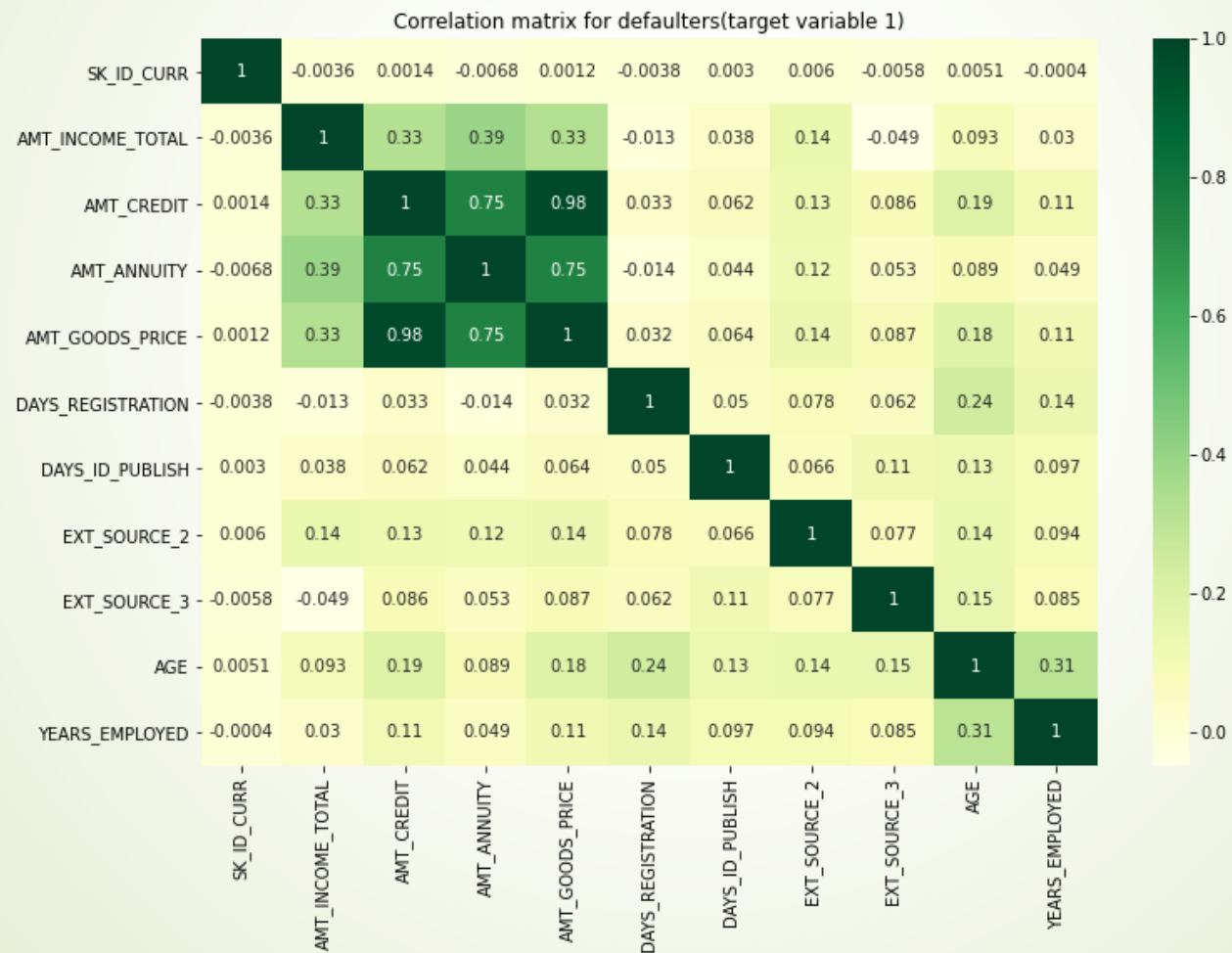
	Var1	Var2	Corr_Value
46	AMT_GOODS_PRICE	AMT_CREDIT	0.986646
47	AMT_GOODS_PRICE	AMT_ANNUITY	0.766231
35	AMT_ANNUITY	AMT_CREDIT	0.761335
34	AMT_ANNUITY	AMT_INCOME_TOTAL	0.437760
45	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.368447
23	AMT_CREDIT	AMT_INCOME_TOTAL	0.360781
119	YEARS_EMPLOYED	AGE	0.352425
104	AGE	DAYS_REGISTRATION	0.298857
107	AGE	EXT_SOURCE_3	0.174218
115	YEARS_EMPLOYED	DAYS_REGISTRATION	0.172101



## Inferences from target 0

- AMT\_GOODS\_PRICE and AMT\_CREDIT shows a highly correlated , which implies that if the price of the goods for which the loan is given is higher then the Credit amount of their loan is also higher.
- AMT\_ANNUITY and AMT\_INCOME\_TOTAL shows a high correlation, which implies that clients with high loan annuity has high annual income.
- AMT\_ANNUITY and AMT\_CREDIT are third highly correlated, that implies for high loan annuity, Credit amount of the loan is also higher.

# HEAT MAP FOR DEFAULTERS



## TOP 10 CORRELATIONS FOR TARGET VARIABLE-1 (DEFAULTERS)

	Var1	Var2	Corr_Value
46	AMT_GOODS_PRICE	AMT_CREDIT	0.982658
47	AMT_GOODS_PRICE	AMT_ANNUITY	0.748131
35	AMT_ANNUITY	AMT_CREDIT	0.747492
34	AMT_ANNUITY	AMT_INCOME_TOTAL	0.392363
45	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.330511
23	AMT_CREDIT	AMT_INCOME_TOTAL	0.326408
119	YEARS_EMPLOYED	AGE	0.306777
104	AGE	DAYS_REGISTRATION	0.241187
101	AGE	AMT_CREDIT	0.188889
103	AGE	AMT_GOODS_PRICE	0.184944



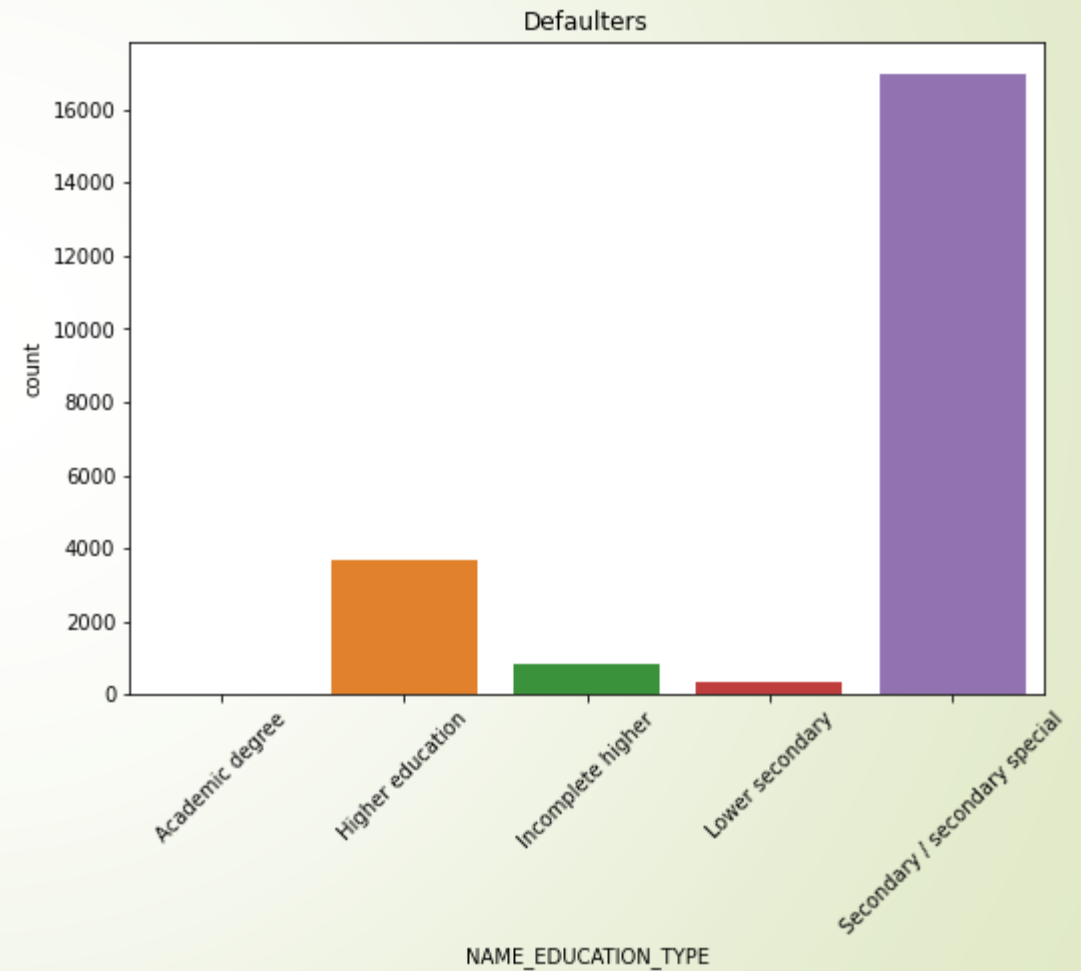
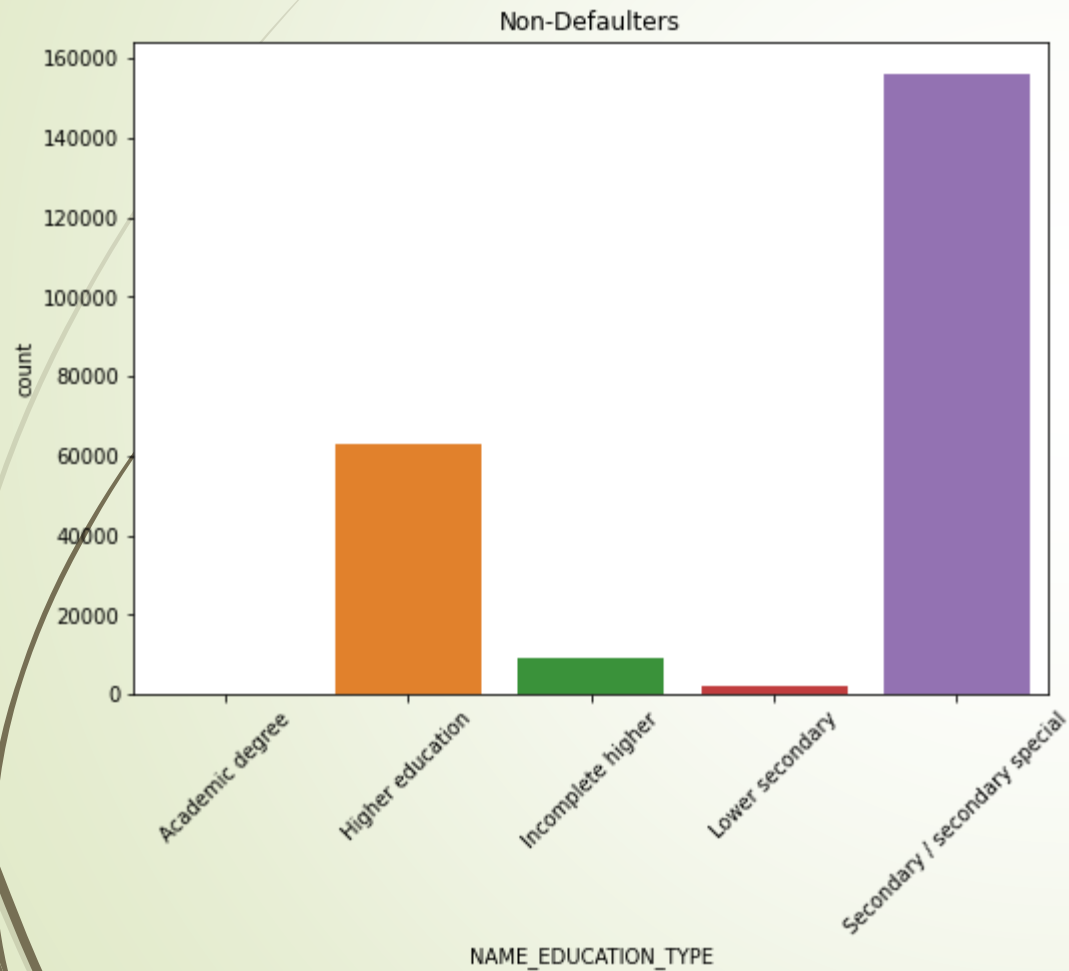
## Inference from target 1

- AMT\_GOODS\_PRICE and AMT\_CREDIT are highly correlated, which means that when credit amount of the loan is higher the amount of goods price for which loan is given is also higher.
- Second highest correlation is seen between AMT\_GOODS\_PRICE and AMT\_ANNUITY which implies that for higher amount of goods price loan annuity is also higher.
- Third highest correlation is seen between AMT\_ANNUITY and AMT\_CREDIT which implies for higher loan annuity, credit amount of the loan is also higher



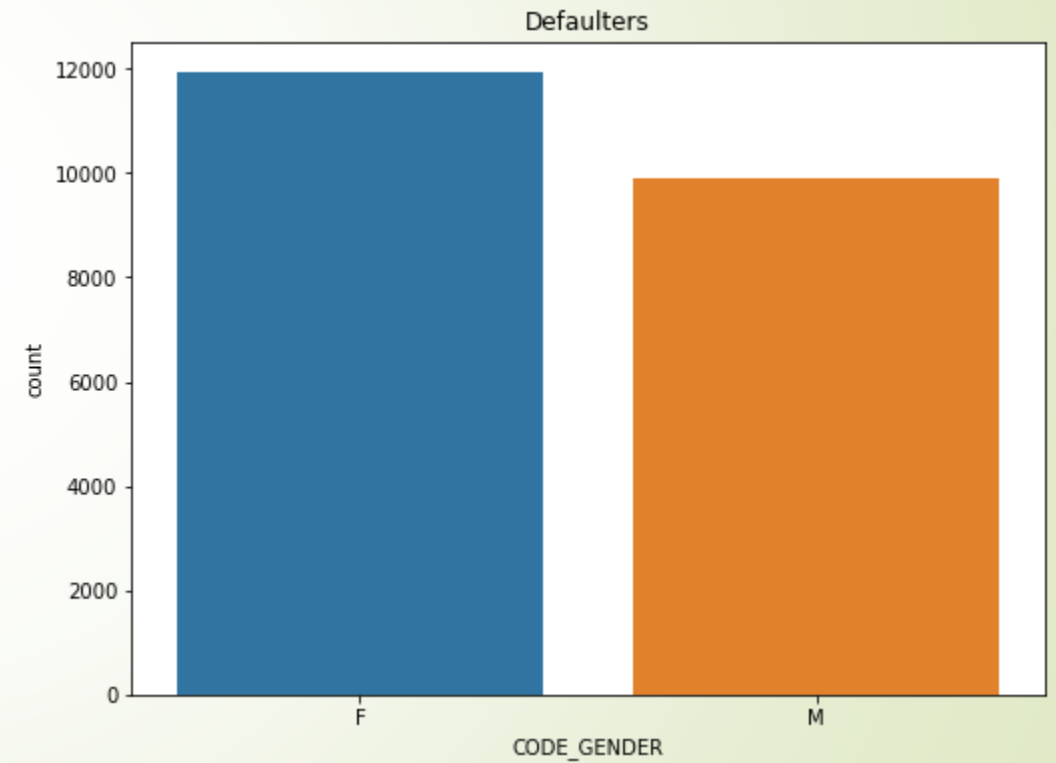
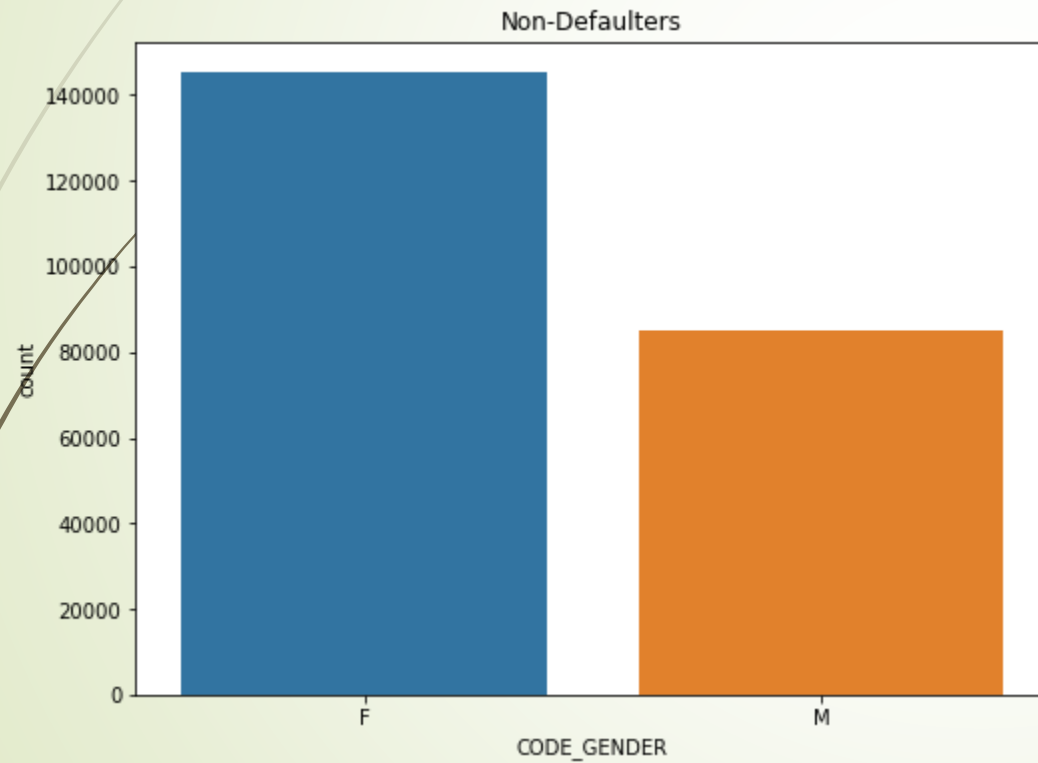
# **UNIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES**

From studying NAME\_EDUCATION\_TYPE graph, we can see that people who have secondary/secondary special takes more loans compared to others while lower secondary take less loans.



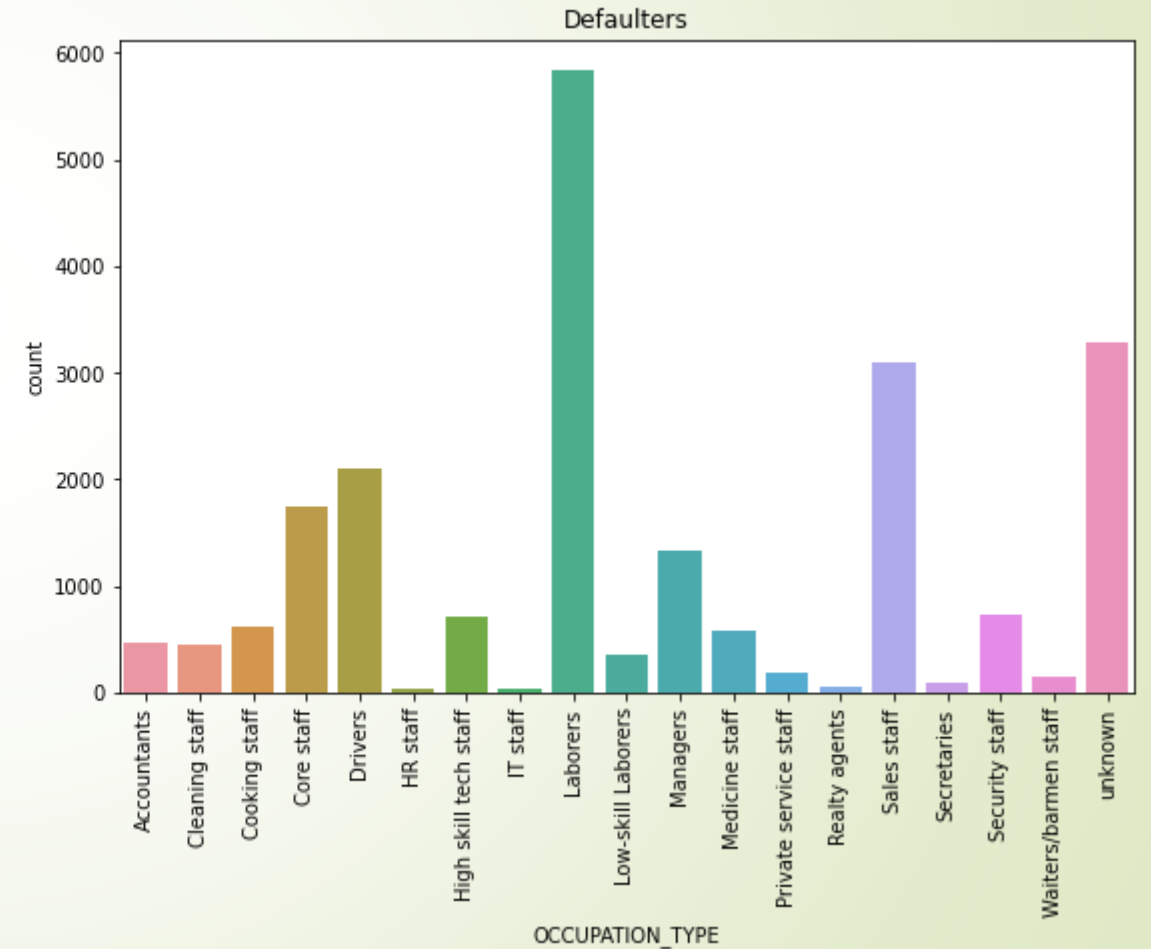
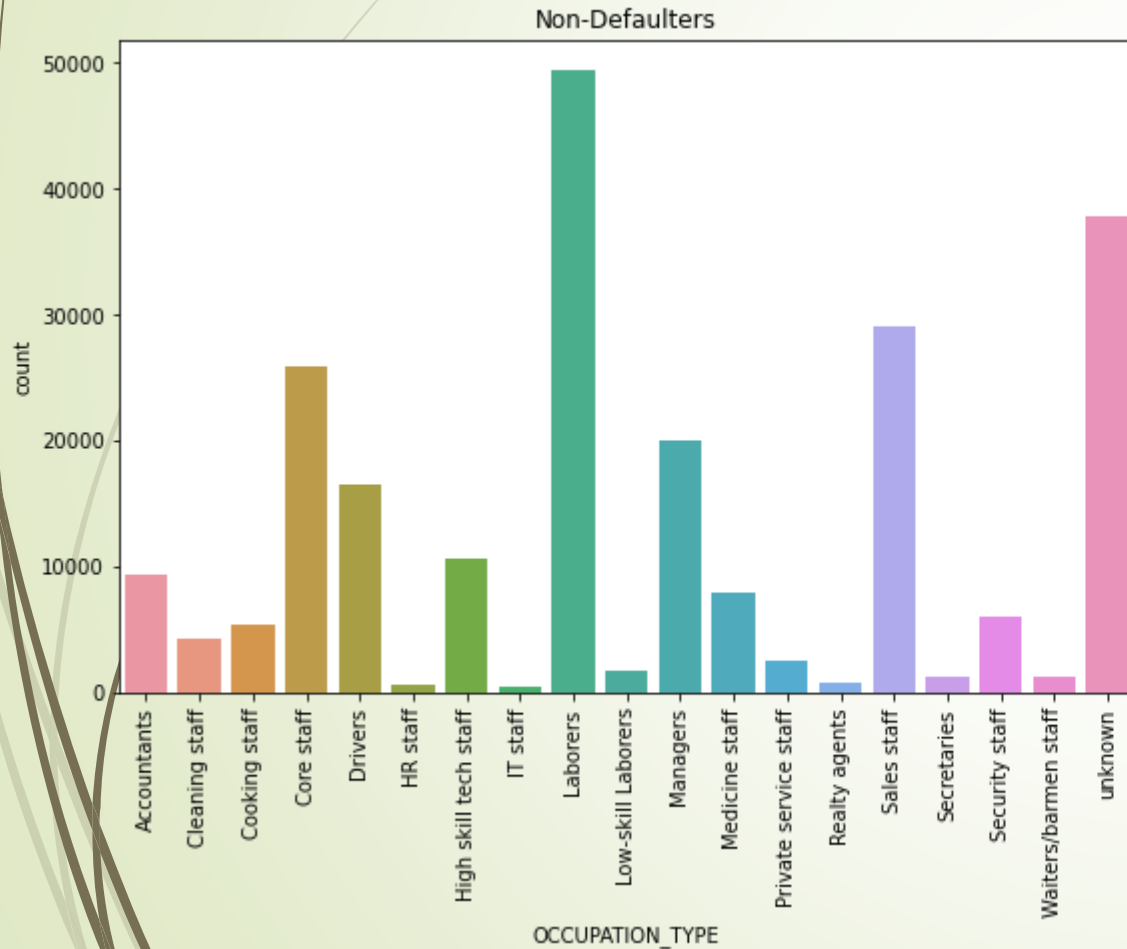


From the graph it is clear that female take more loan than male but the repaying capacity when compared with male it is higher for female.

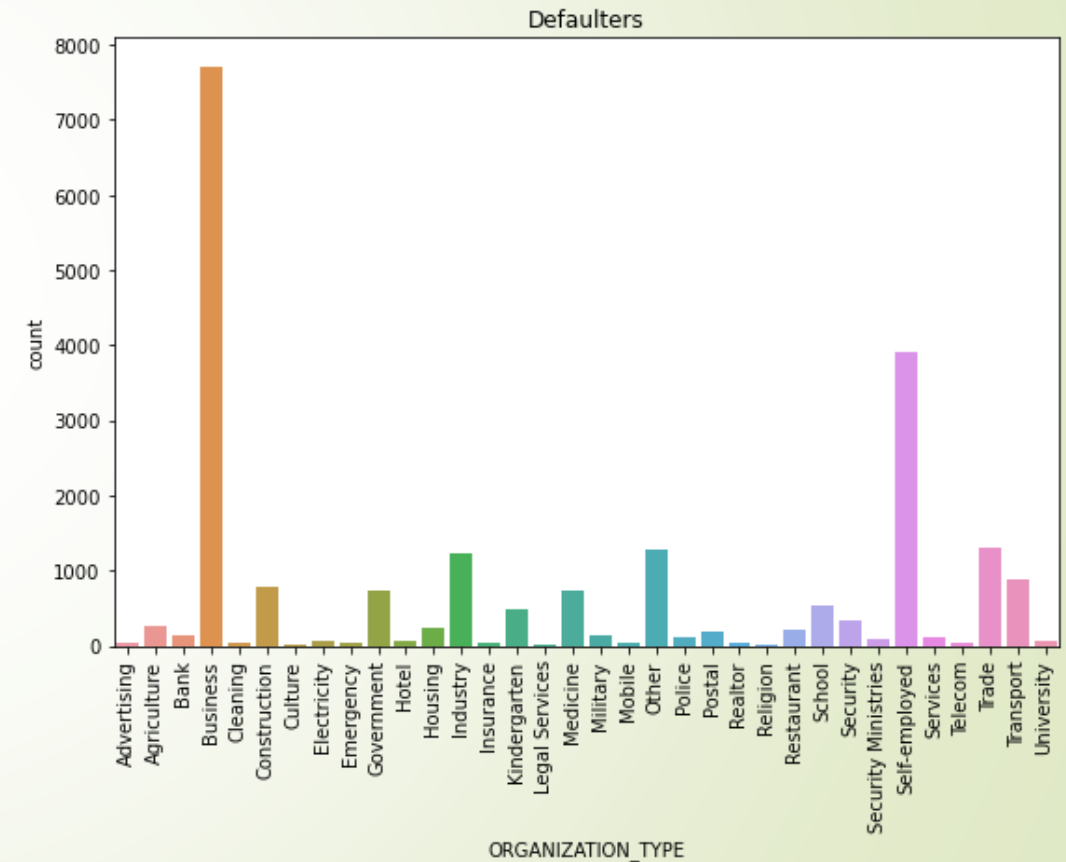
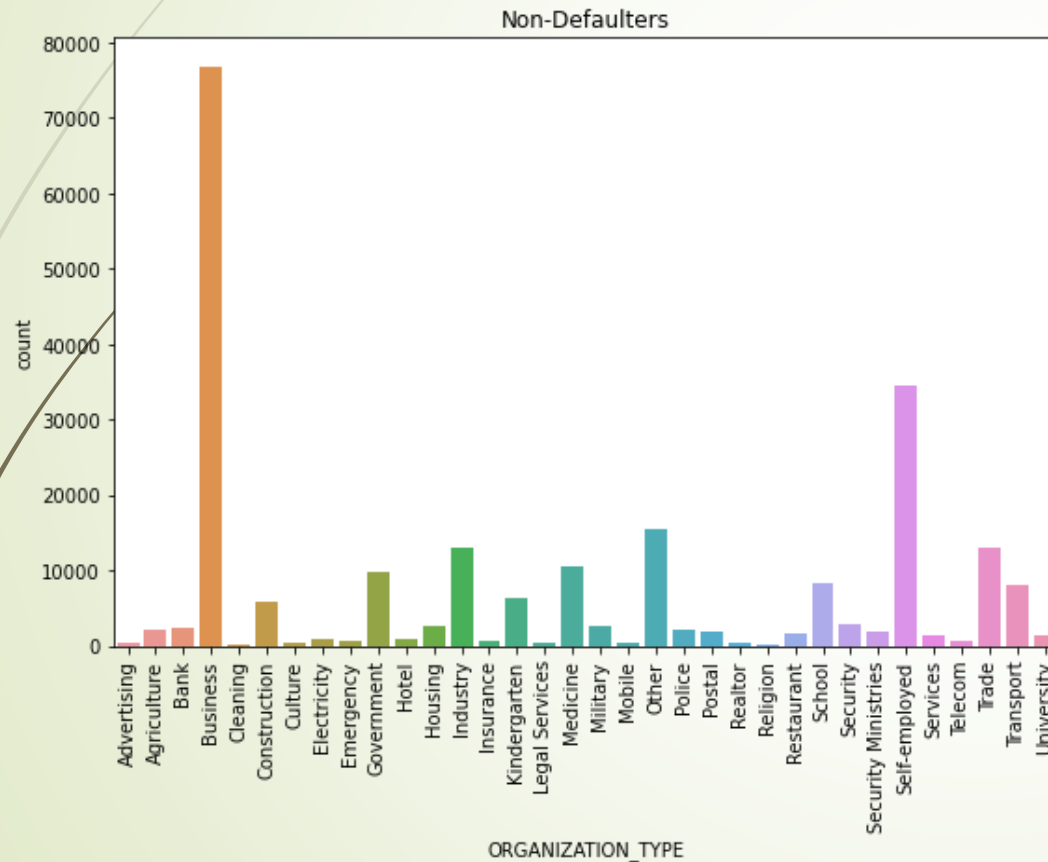




From analyzing OCCUPATION\_TYPE we can see that people who belongs to labourers, sales staff, core staff take more loans and as a result they tend to be defaulters as their income is not consistent.



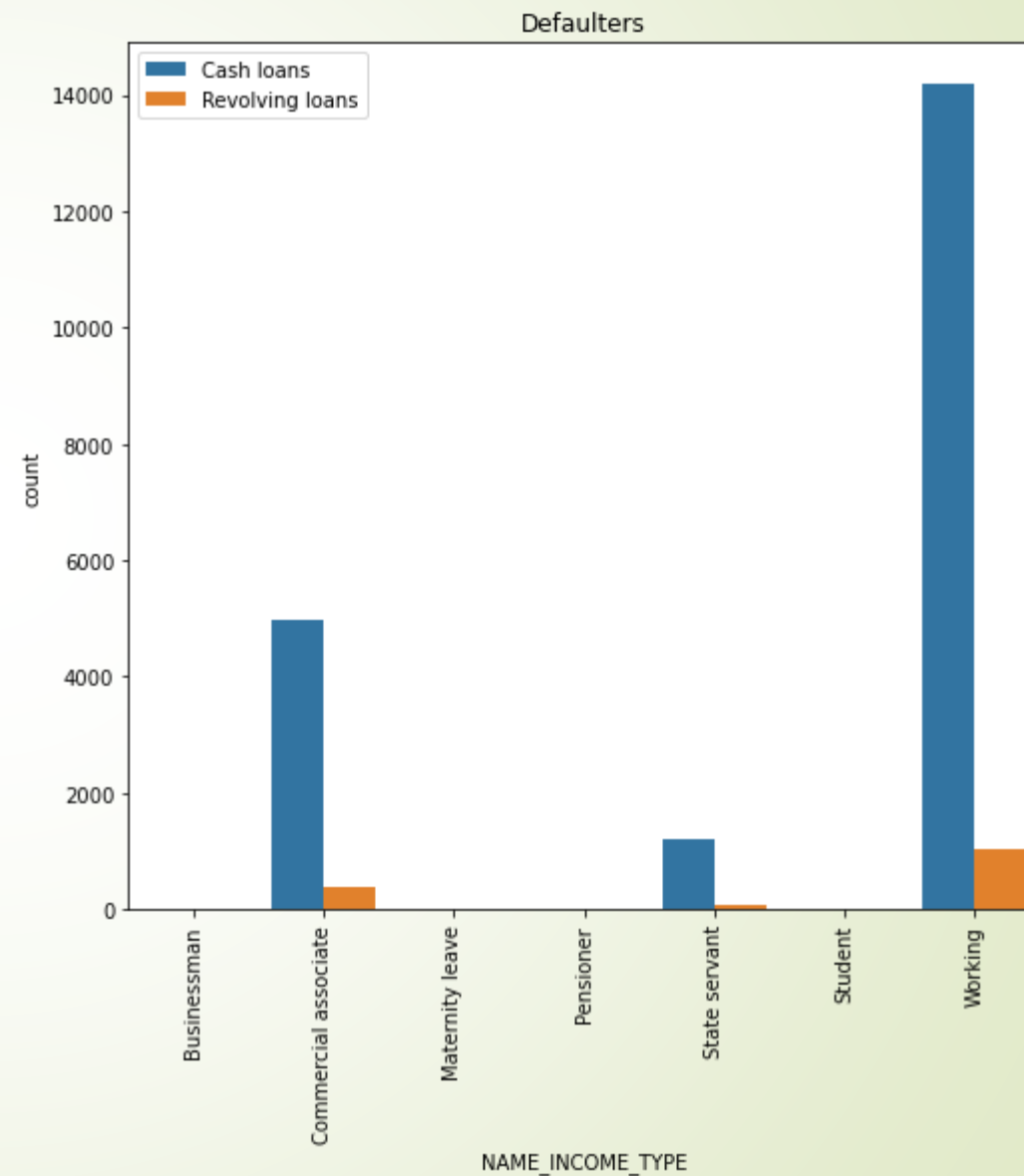
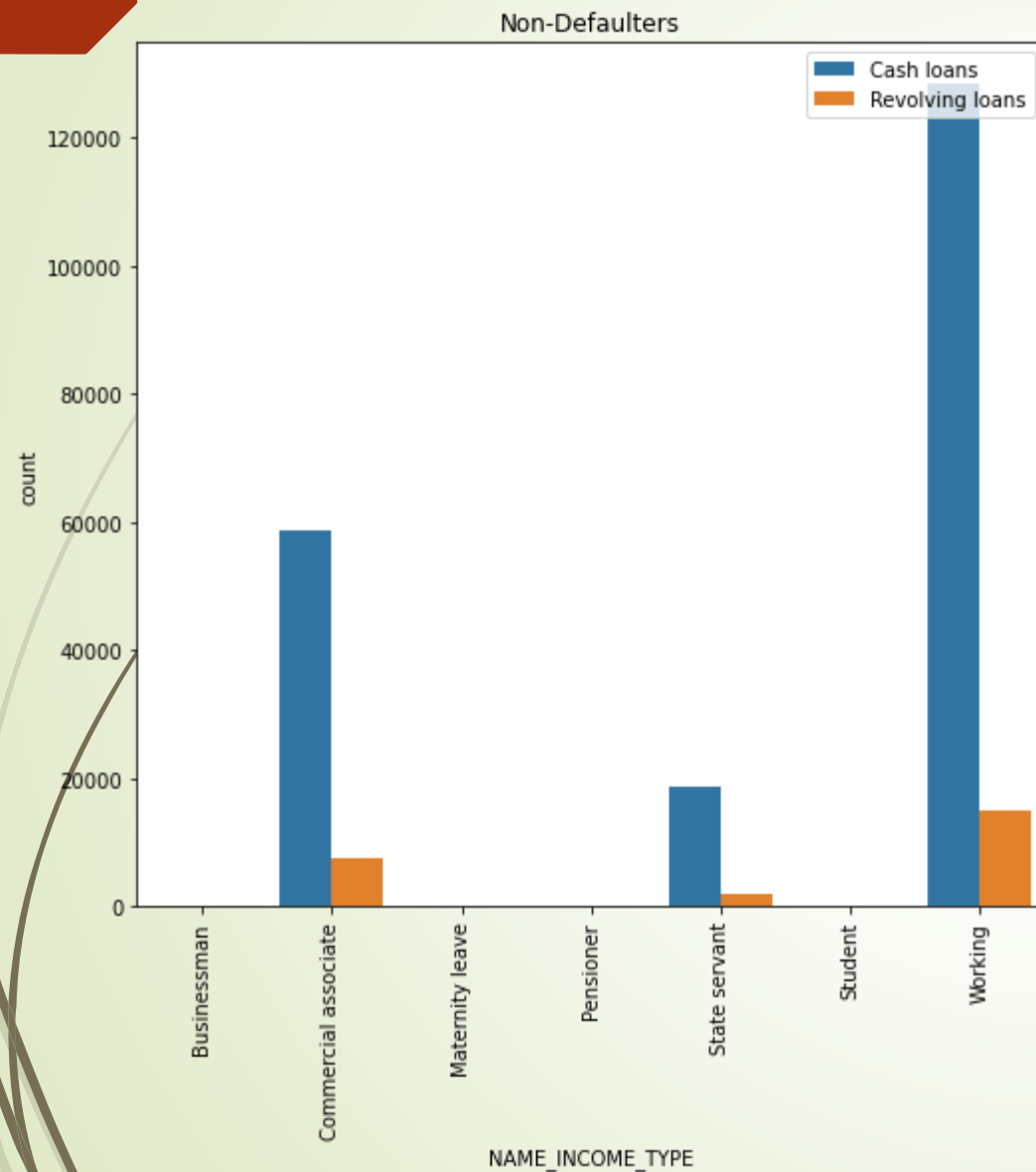
From the graph ORGANIZATION\_TYPE , it is visible that a high percentage of defaulters are present in self-employed and business categories





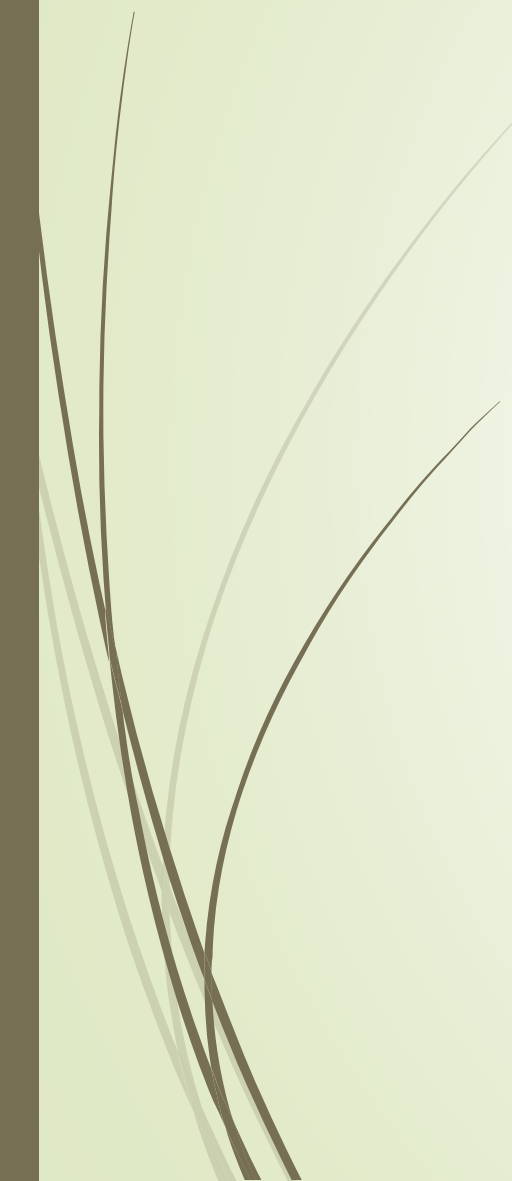
# **Bivariate Analysis for Categorical variables**

## NAME\_INCOME\_TYPE vs NAME\_CONTRACT\_TYPE

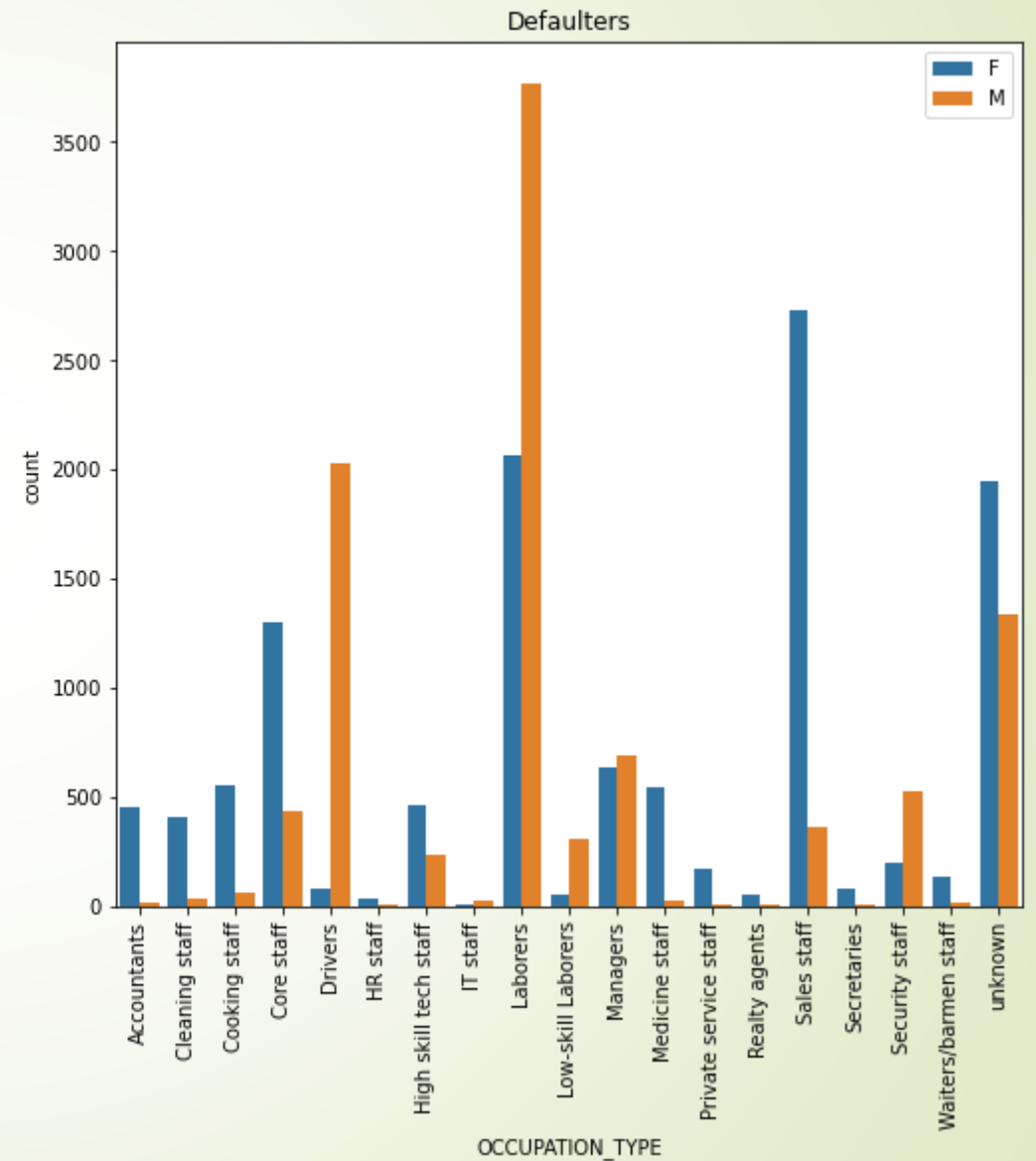
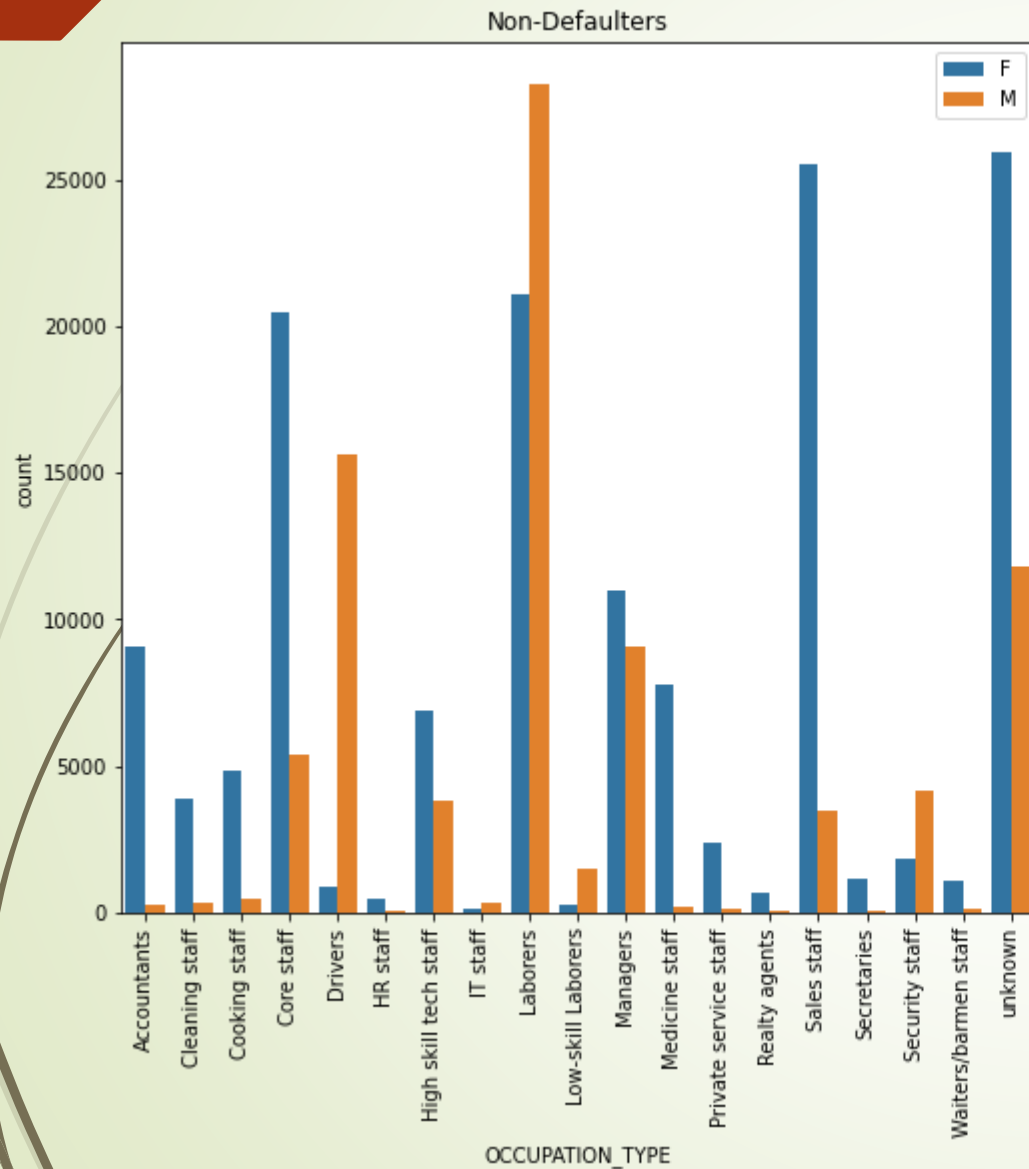




## Inference

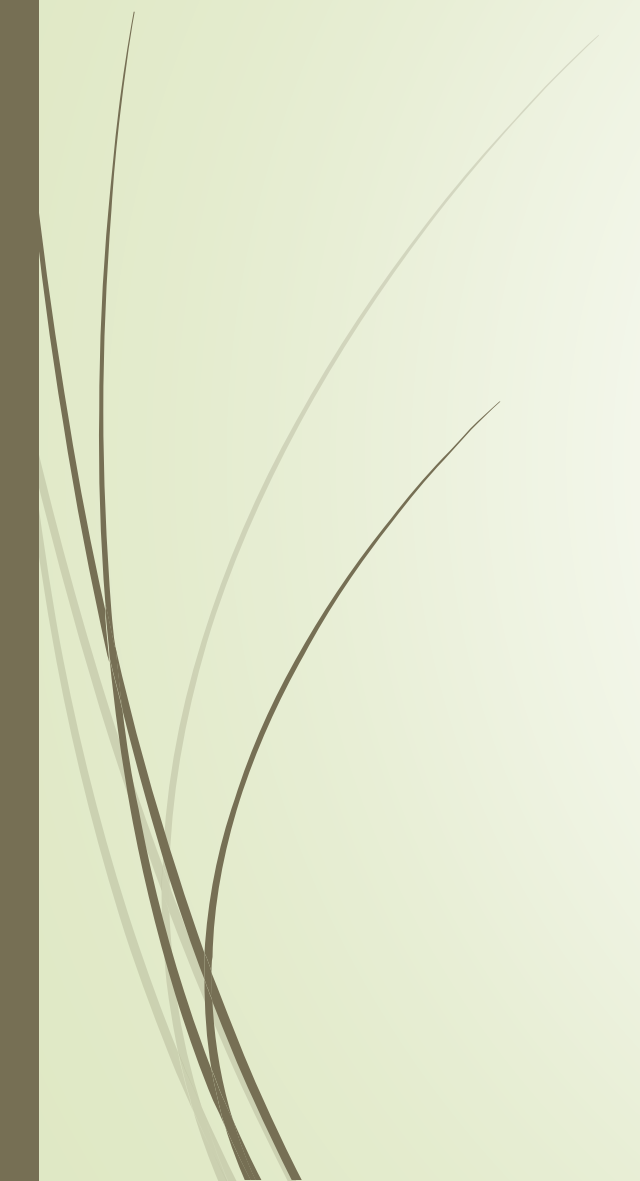
- There is no any correlation between income type Businessman , Maternity leave, Pensioner , Student and the contract type as no data available for these categories.
  - Working people tend to take more loans but they are more prone to become a defaulter.
- 

# OCCUPATION\_TYPE vs CODE\_GENDER

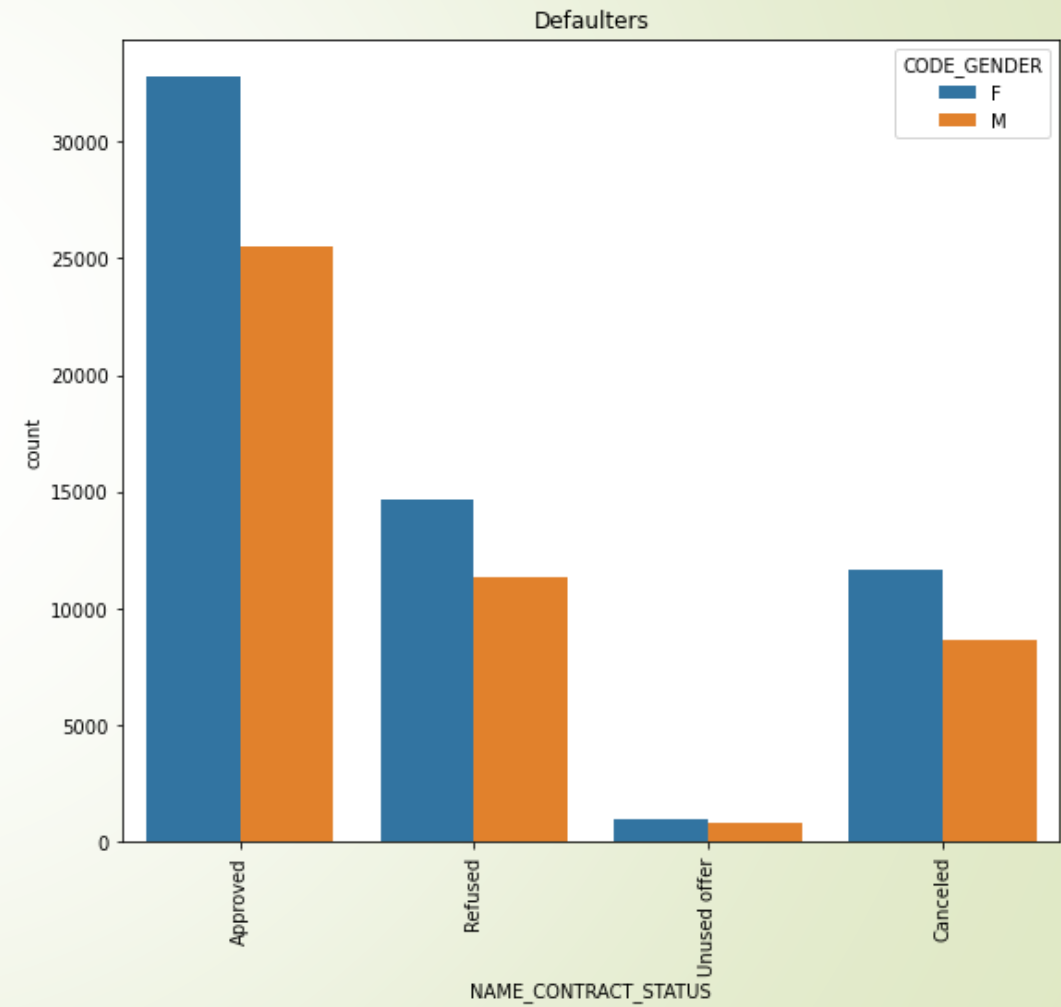
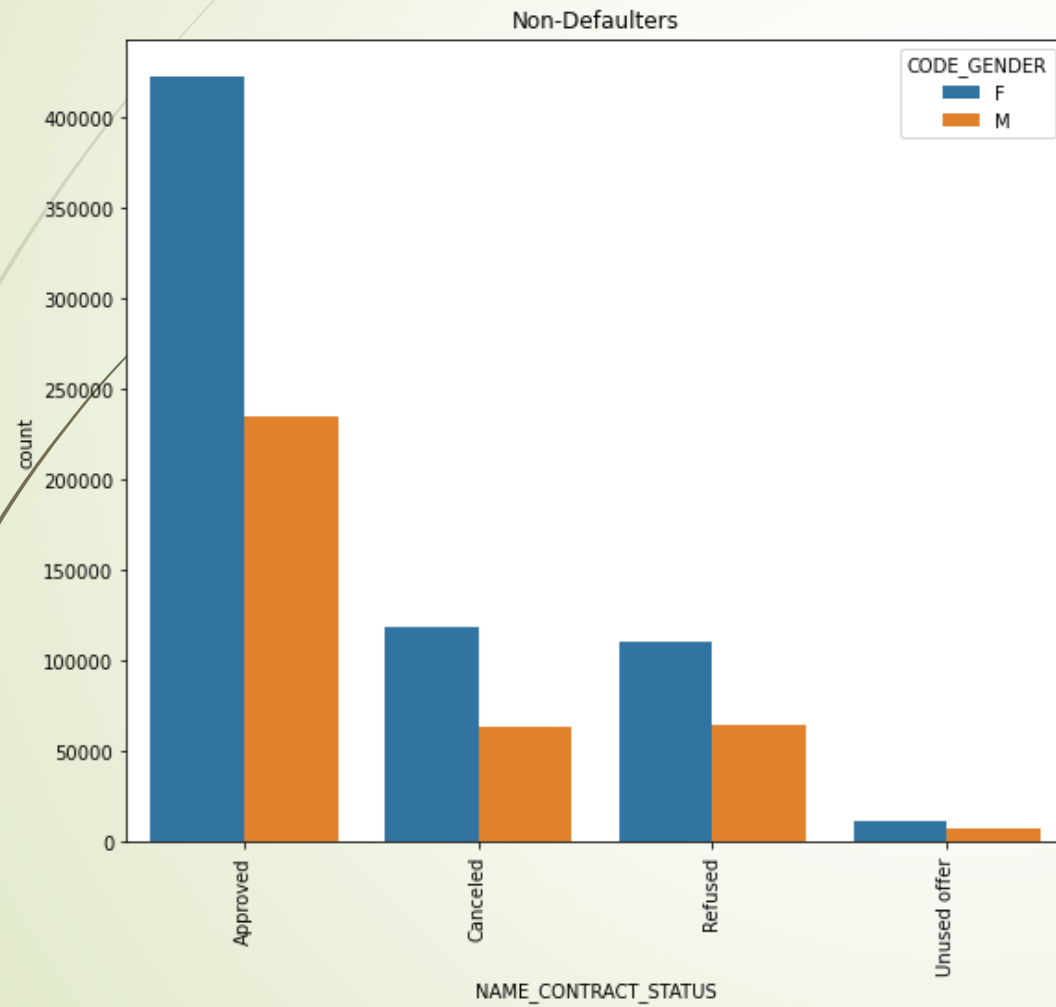




## Inferences

- Among defaulters in sales staff females are higher in count.
  - Among Non-Defaulters for except drivers and labourers female stood higher in proportion with male.
  - Proportion of female clients is more having occupation type Core staff, Sales staff , Accountants , cleaning staff , cooking staff and unknown.
  - Number of male clients is more in Drivers, Security staff and Laborers.
- 

## NAME\_CONTRACT\_STATUS vs CODE\_GENDER







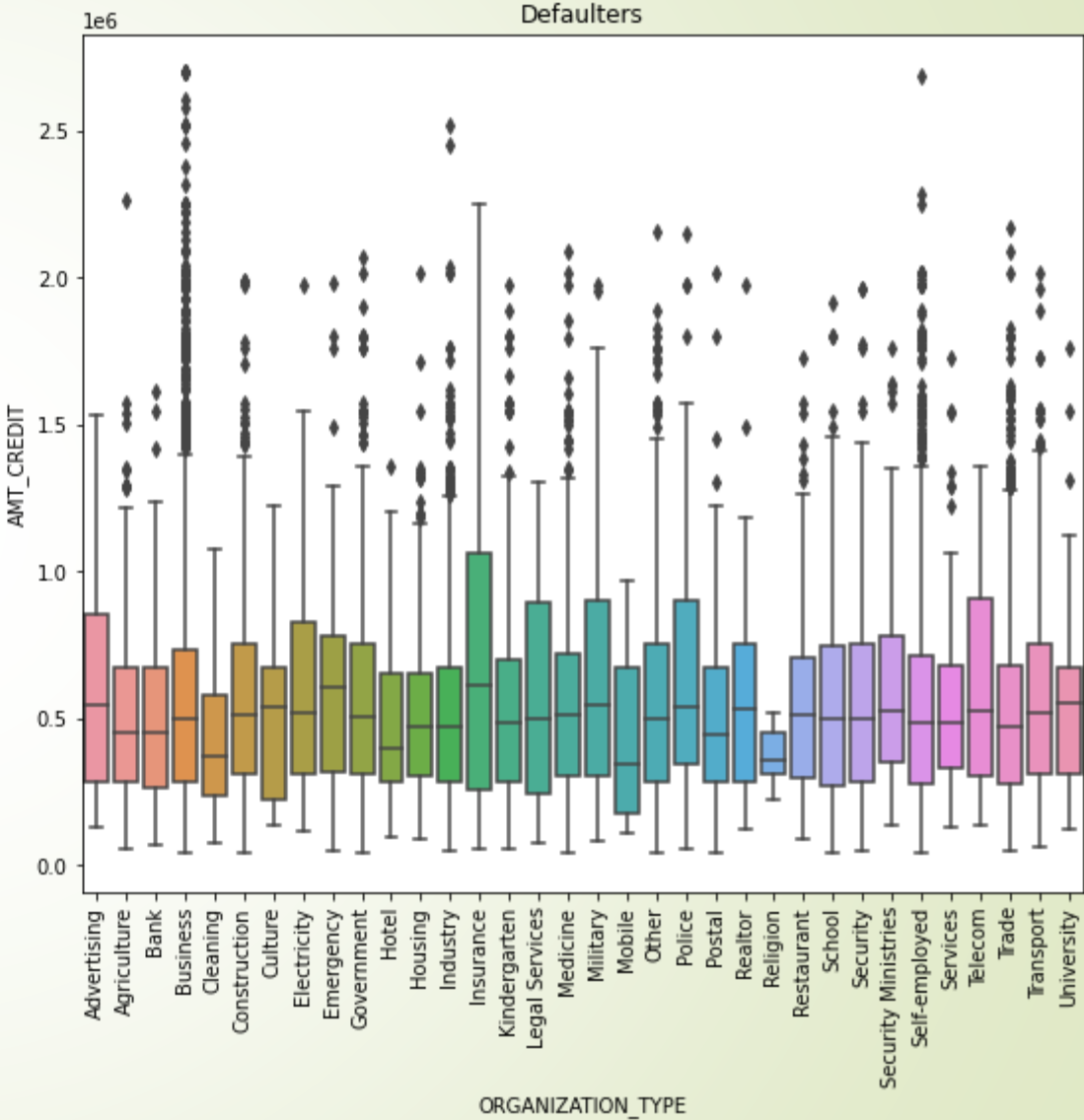
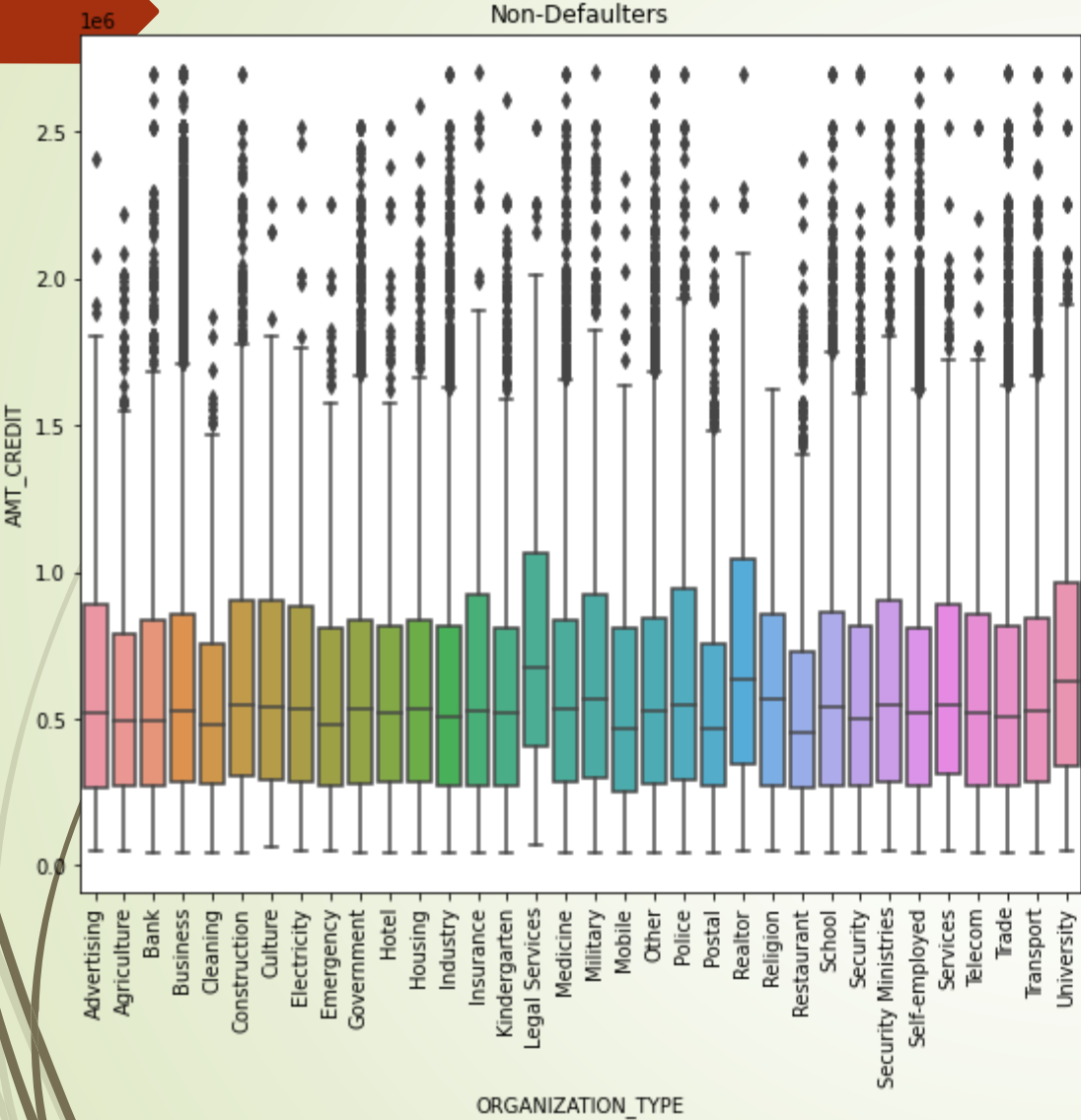
## Inference

- Females who got loan approved are low tendency to be a defaulter where as male is more prone to be a defaulter.
- Among those who cancelled the loan in the previous application, females high chances of not to be a defaulter



## **BIVARIATE ANALYSIS FOR NUMERICAL vs CATEGORICAL VARIABLES**

# AMT\_CREDIT vs ORGANIZATION TYPE





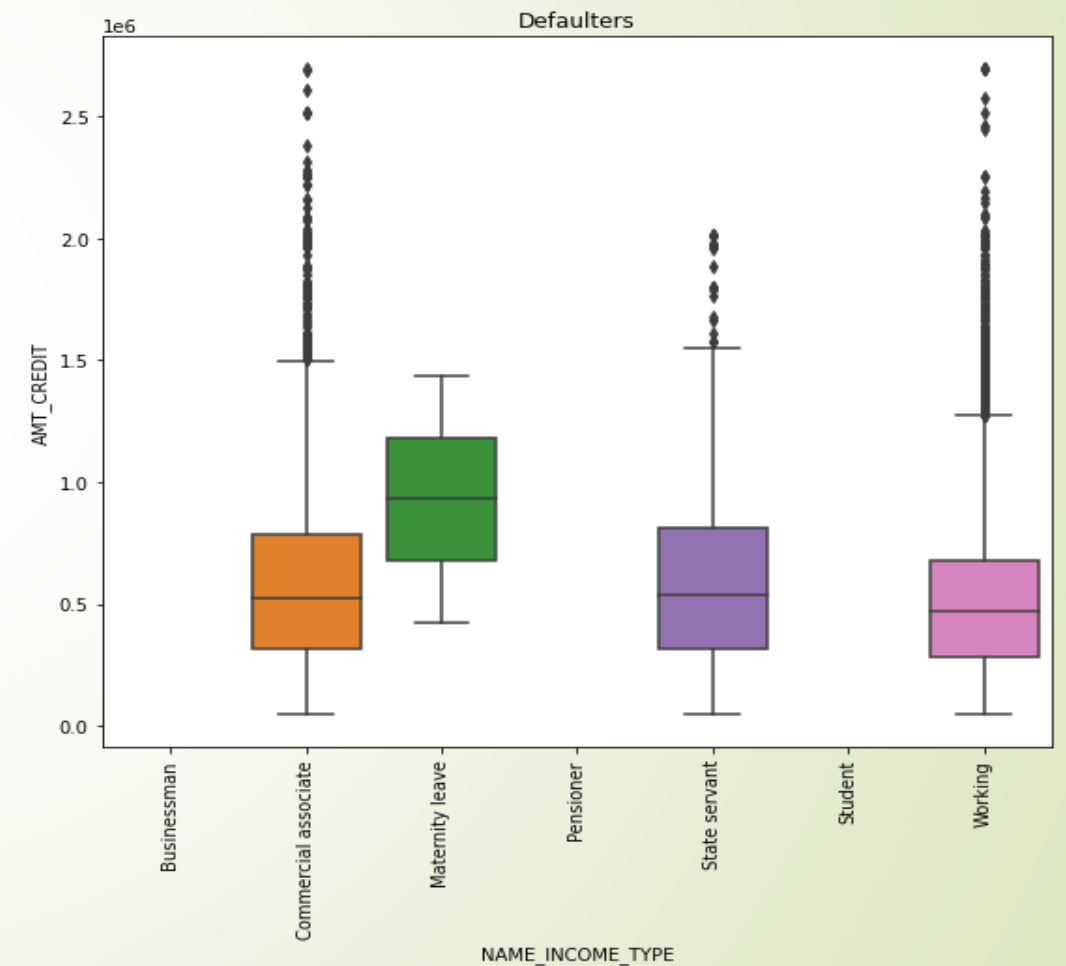
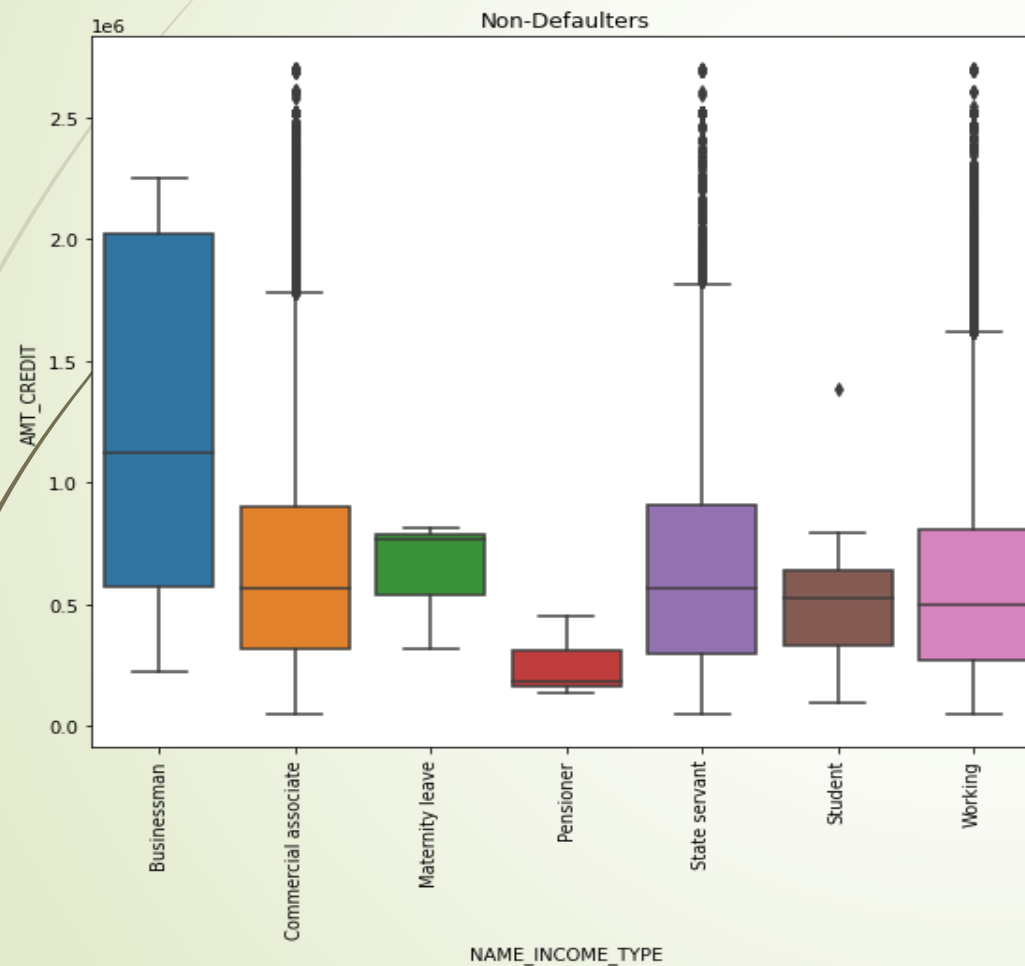
## **Inference**

AMT\_CREDIT is higher for clients from legal service among Non-Defaulters.

AMT\_CREDIT is higher for clients from insurance among Defaulters and also it doesn't have any outliers.


Outliers are higher among Non-Defaulters

## AMT\_CREDIT vs NAME\_INCOME\_TYPE





## Inference

- Businessman, Pensioner and Students are not facing any payment difficulties that is any of them belongs to Defaulters
  - Number of outliers are more for Non-Defaulters as compared to Defaulters.
  - Median is highest for Businessman for Non-Defaulters
  - Median is highest for Maternity leave for Defaulters
- 



## CONCLUSION

- High percentage of defaulters are present in self-employed and business organizational categories and as a result banks should be careful when giving loans to people who belongs to this sector .
- People in the age category 18-35 having high-income is most likely to be default . So banks should be very keen while giving loans to them.
- Banks should encourage giving more loans to married women as the percentage of unsuccessful payment is much less compared to others.
- Banks should focus less on income type “working” as they are having most number of unsuccessful payments.



**THANK YOU**