ASSIGNMENT ON EDA (BANK CREDIT DATA SET)

SUBMITTED BY

DHEERAJ MISHRA

BATCH = DS C40

Program : upGrad & IIITB | Data Science Program – January 2022

Data : Banking and finance service data set

Objective :

1. To get a pattern which indicates if a client has difficulty in loan payments.
2. Key parameters that can affect the loan repayment.

Dataset provided :

This dataset has 3 files as explained below:

1. 'application_data.csv' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.

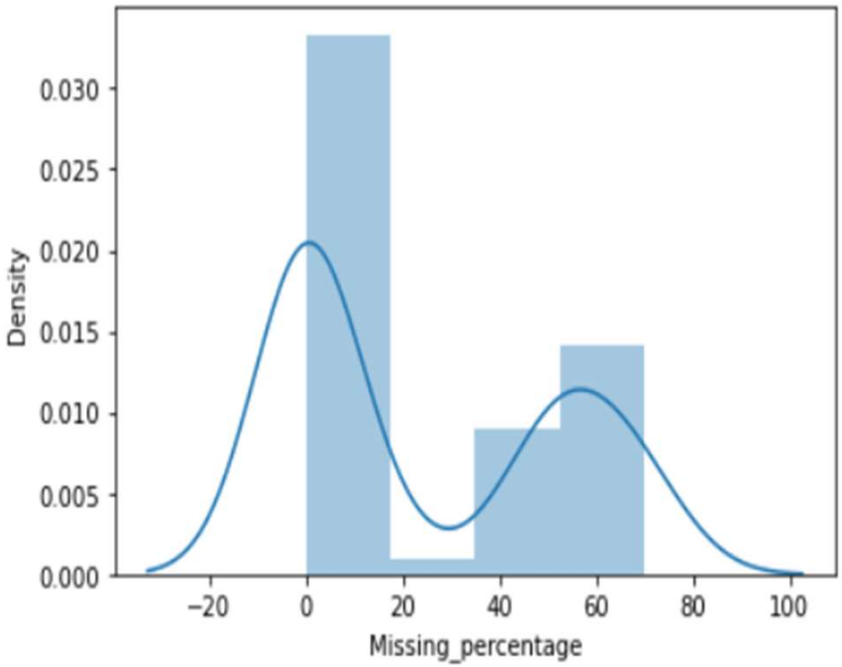
2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

3. 'columns_description.csv' is the data dictionary which describes the meaning of the variables.

Understanding & cleaning of data:

1. Necessary checks are done on the data using functions like info(), describe(), shape, .head() etc. on both the data set.
2. Data cleaning is then applied to the data set
3. Check for NA values and necessary action is then taken on those
4. Imputation is done on some of the column features
5. Check for correct data types and standardizing data
6. Check for outliers

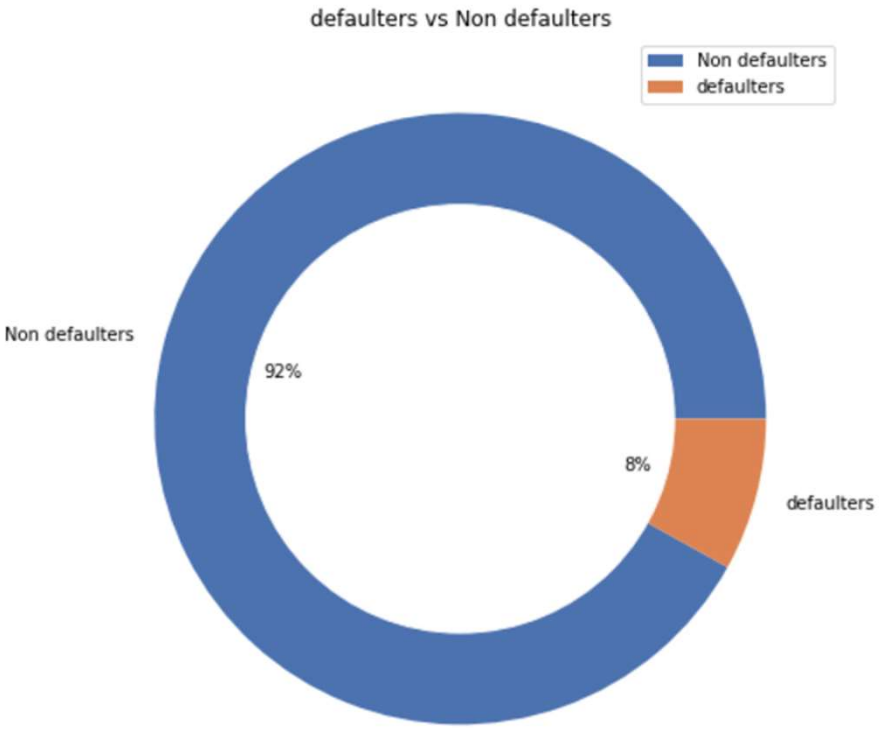
Null Values percentage in application data set



we can see two peaks in this graph
1.fist peak shows that most of the missing value percentage is between 0% to 16%.

2.second peak shows that second most missing value are in between 50% to 75%.

ANALYSIS



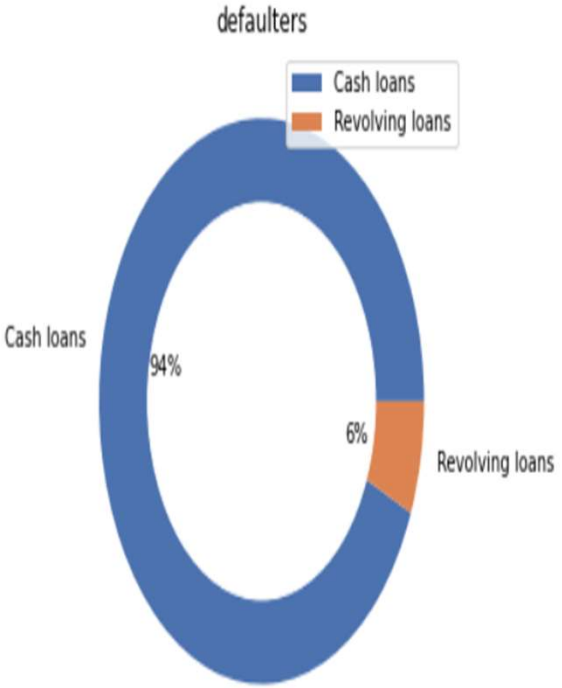
This is a Graph Which Show Imbalance

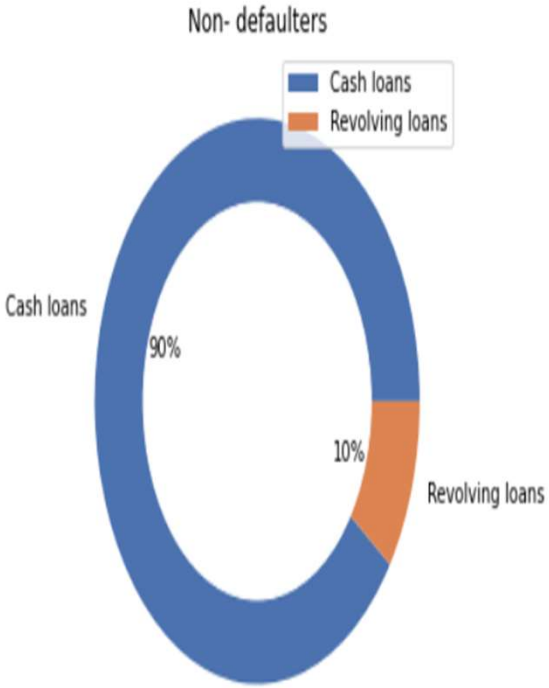
As we can see the imbalance is high between defaulters (8%) and non defaulters(92%) the ratio is 0.086

Percent chart elements
Defaulters
Non-Defaulters

Ratio =0.086

Univariate Analysis (Categorical)
Target : NAME_CONTRACT_TYPE
Objective: to understand the loan Type of defaulters vs Non- defaulters

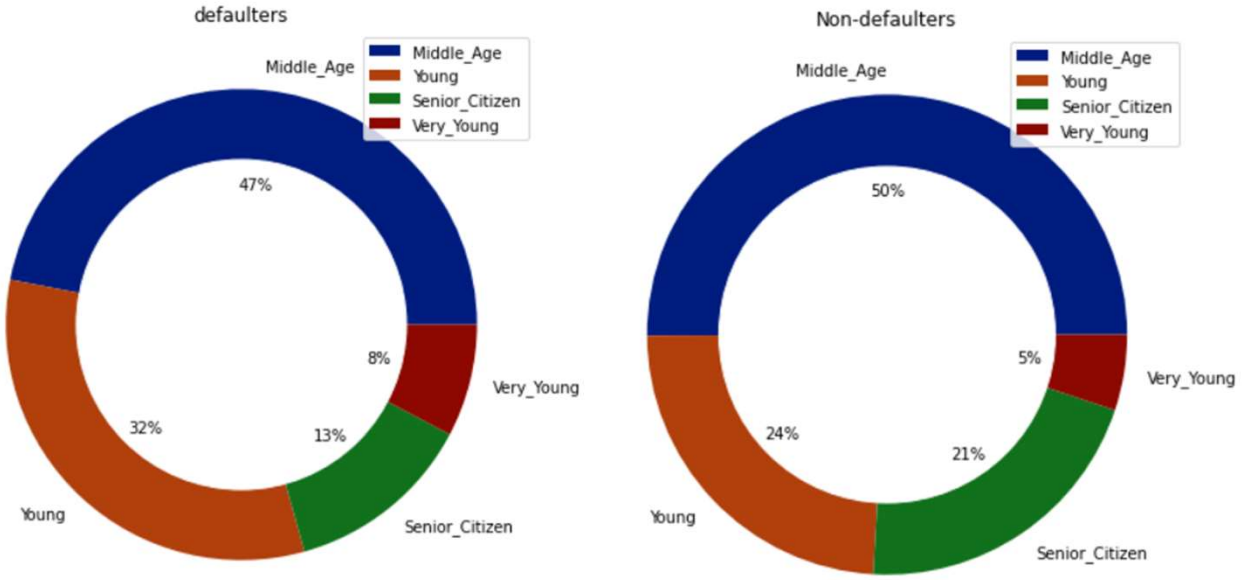




Observation

1. In both Defaulters and non Defaulters people prefer cash loans.
2. But in Defaulters this can be seen that they avoid revolving loans.

Univariate Analysis (Categorical)
Target :Age category
Objective: to understand the Age range of the client in defaulters vs Non- defaulters

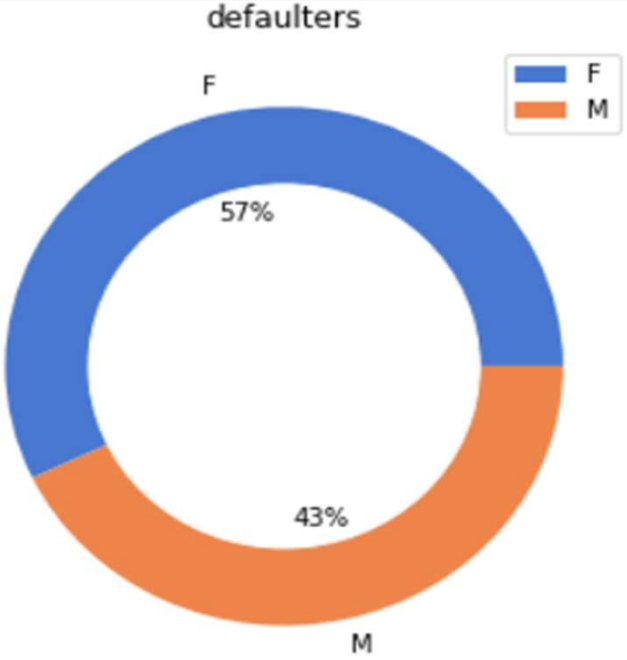


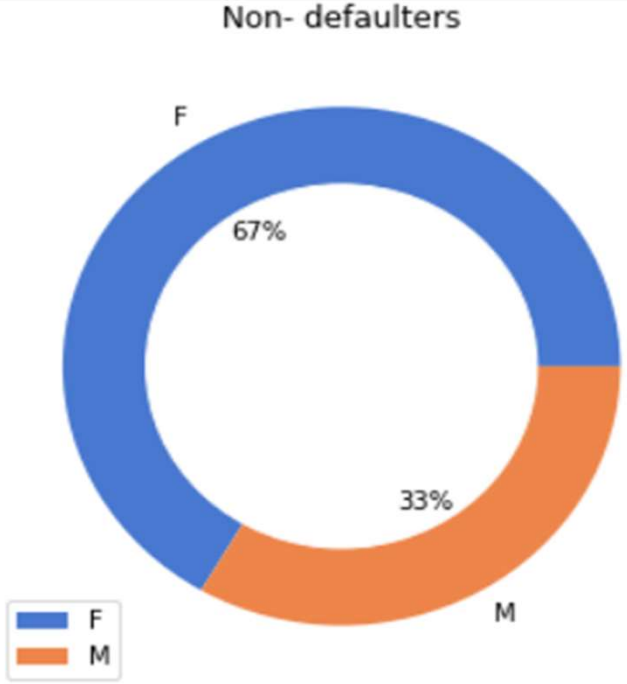
Observation

1. In both Defaulters and non Defaulters Middle Age client are more.

2. In defaulters young client are more as compare to non defaulters and senior Citizen are less.

Univariate Analysis (Categorical)
Target : Gender
Objective: to understand the gender distribution in defaulters vs Non- defaulters

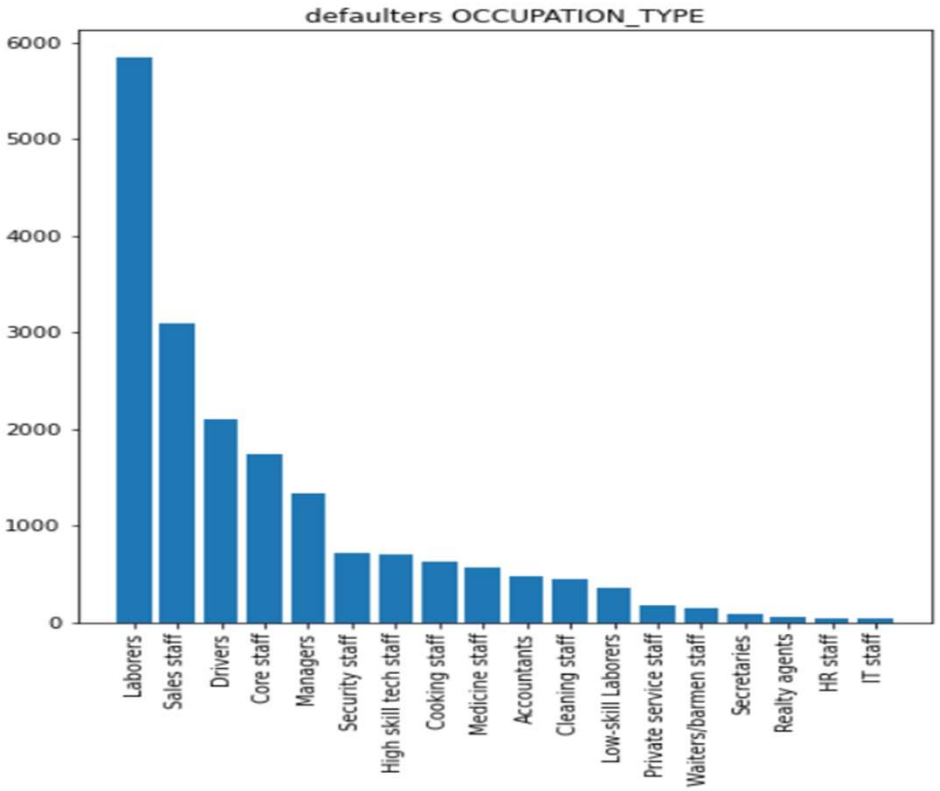


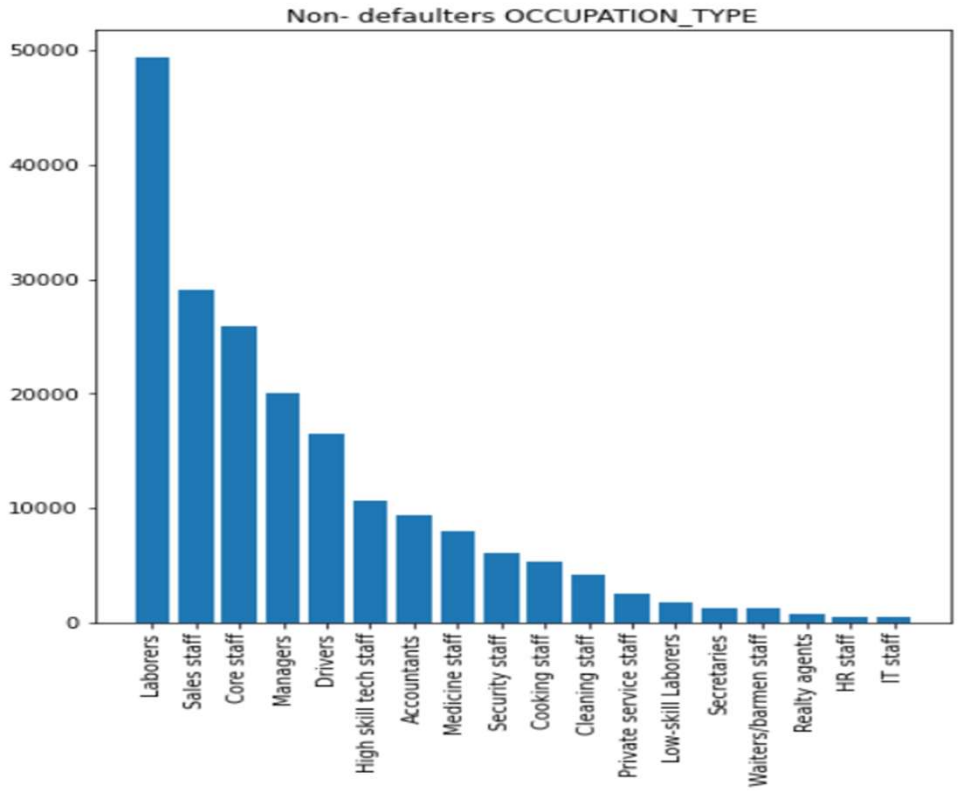


OBSERVATION

1. In both Defaulters and non Defaulters percentage of females is on higher side.
2. But in Defaulters this can be seen that male percentage is also high as compare to non defaulters.

Univariate Analysis (Categorical)
Target : OCCUPATION_TYPE
Objective: to understand the gender distribution in defaulters vs Non- defaulters

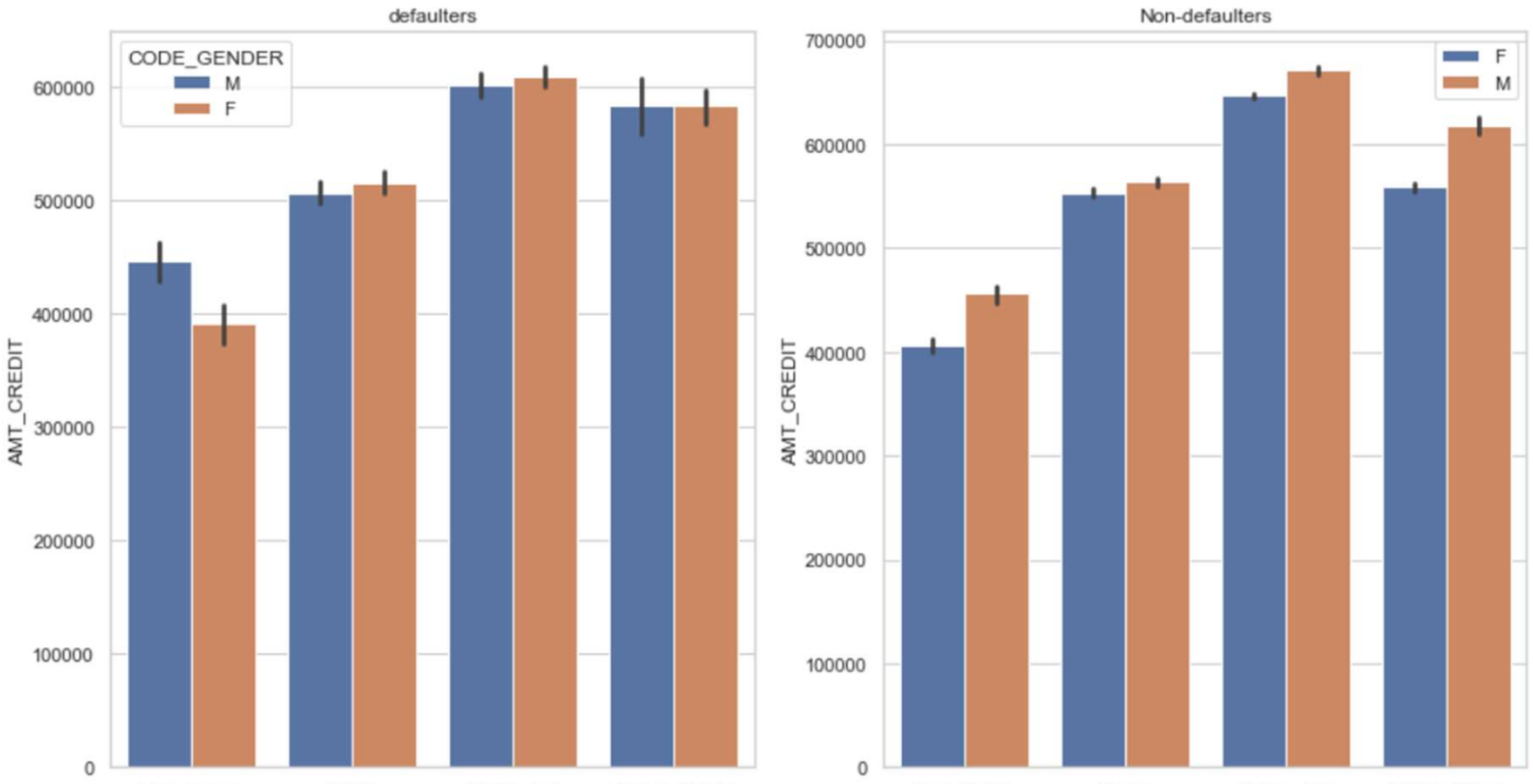




OBSERVATION

1. In both Defaulters and non Defaulters majority are Laborers.

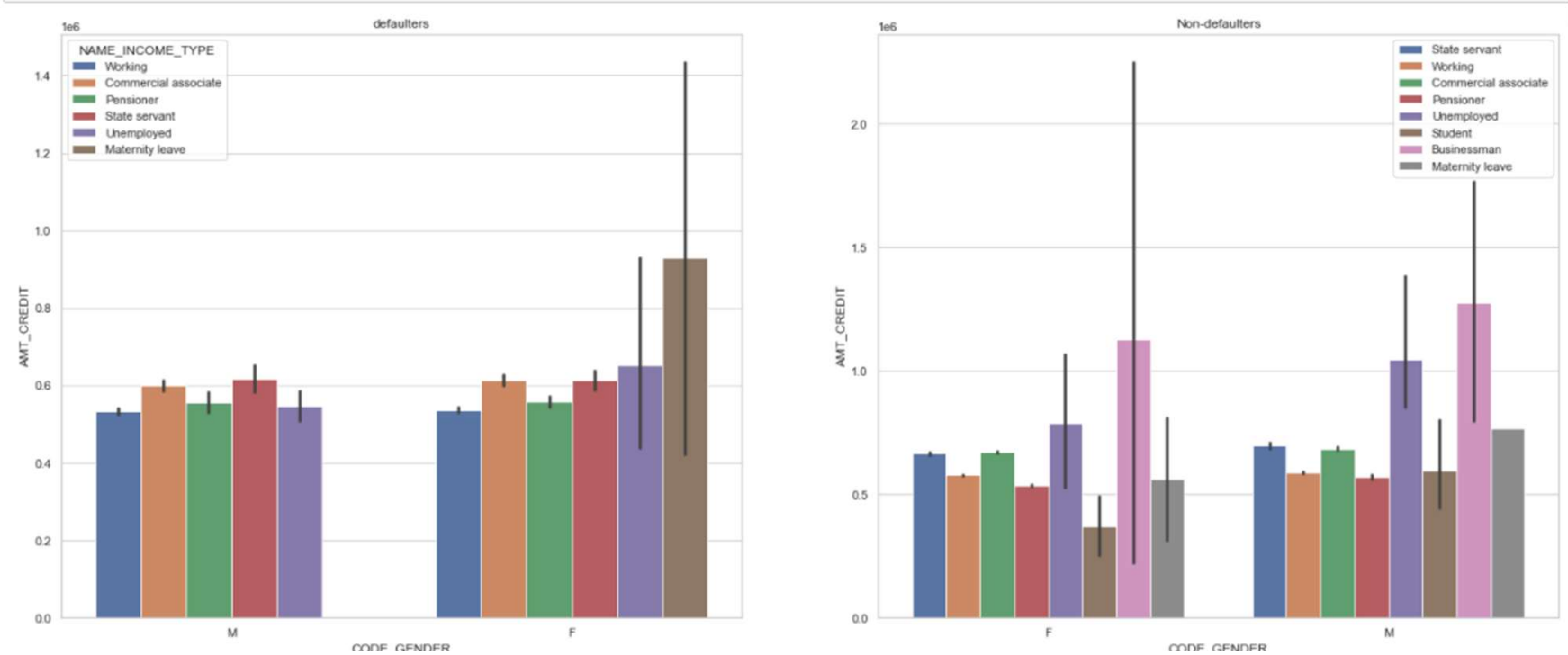
bivariate Analysis
Target :AMT_CREDIT vs Age category
Objective: to understand the AMT_CREDIT vs Age category of the client in defaulters vs Non- defaulters



Observation

1. In Defaulters male(very young) are in higher no then in female(very young).

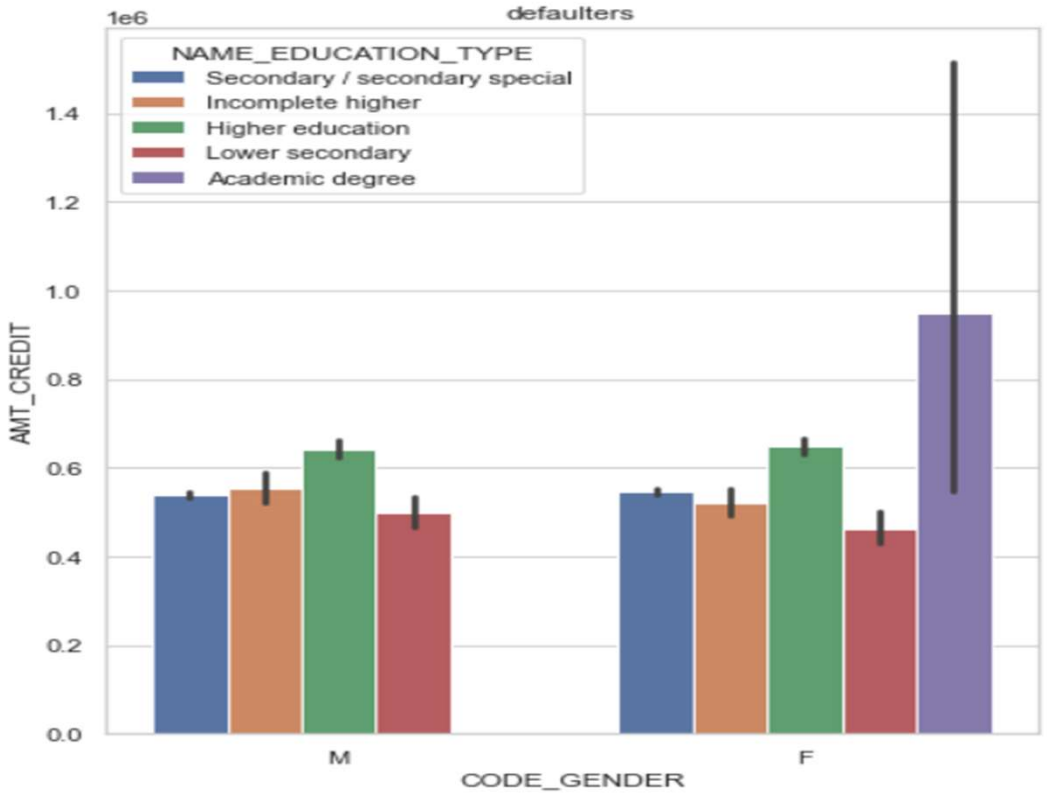
bivariate Analysis
Target :AMT_CREDIT vs CODE_GENDER and Income source
Objective: to understand the AMT_CREDIT vs CODE_GENDER and Income source of the client in defaulters vs Non- defaulters

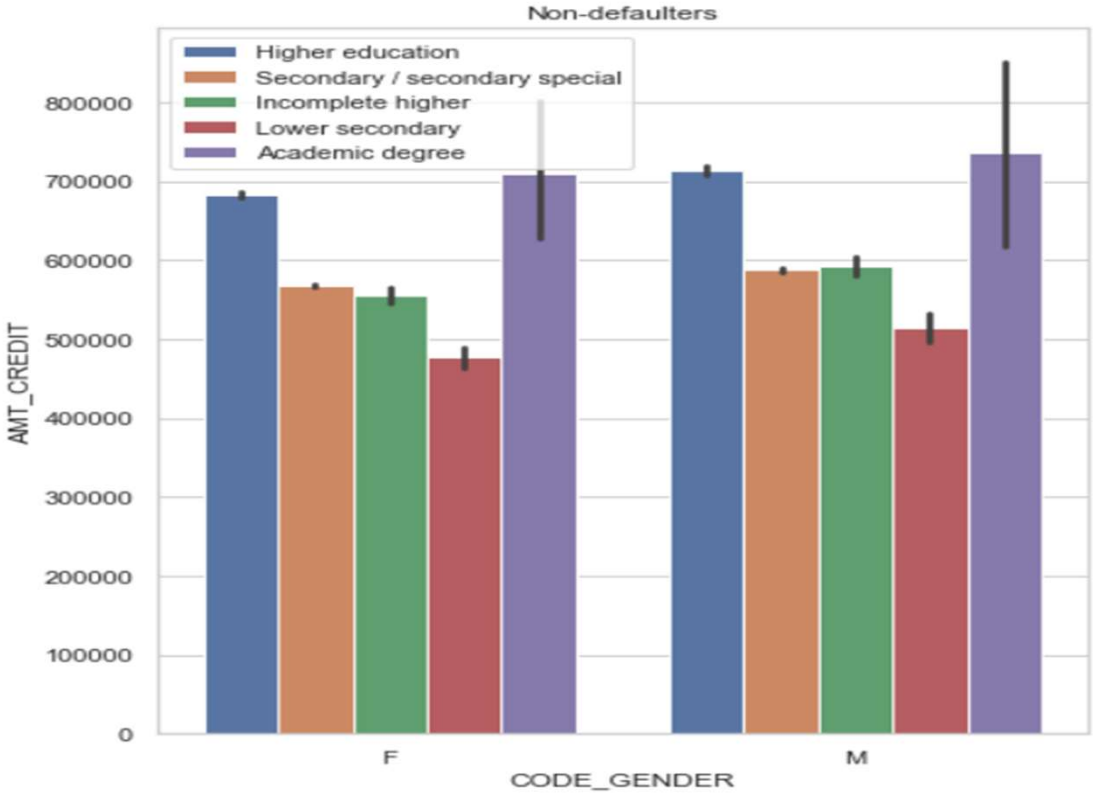


Observation

1. In Defaulters Female(Maternity leave) are having difficulties in paying their loan.

bivariate Analysis
Target :AMT_CREDIT vs CODE_GENDER and Education type
Objective: to understand the AMT_CREDIT vs CODE_GENDER and Education Type of the client in defaulters vs Non- defaulters

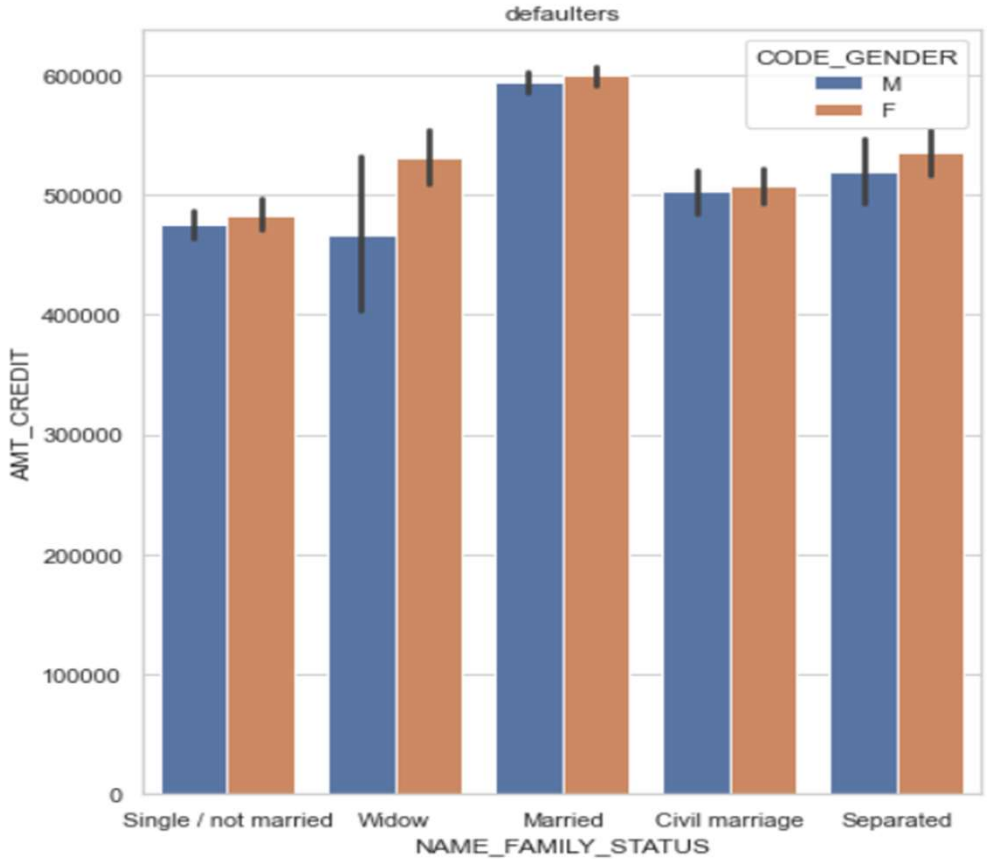


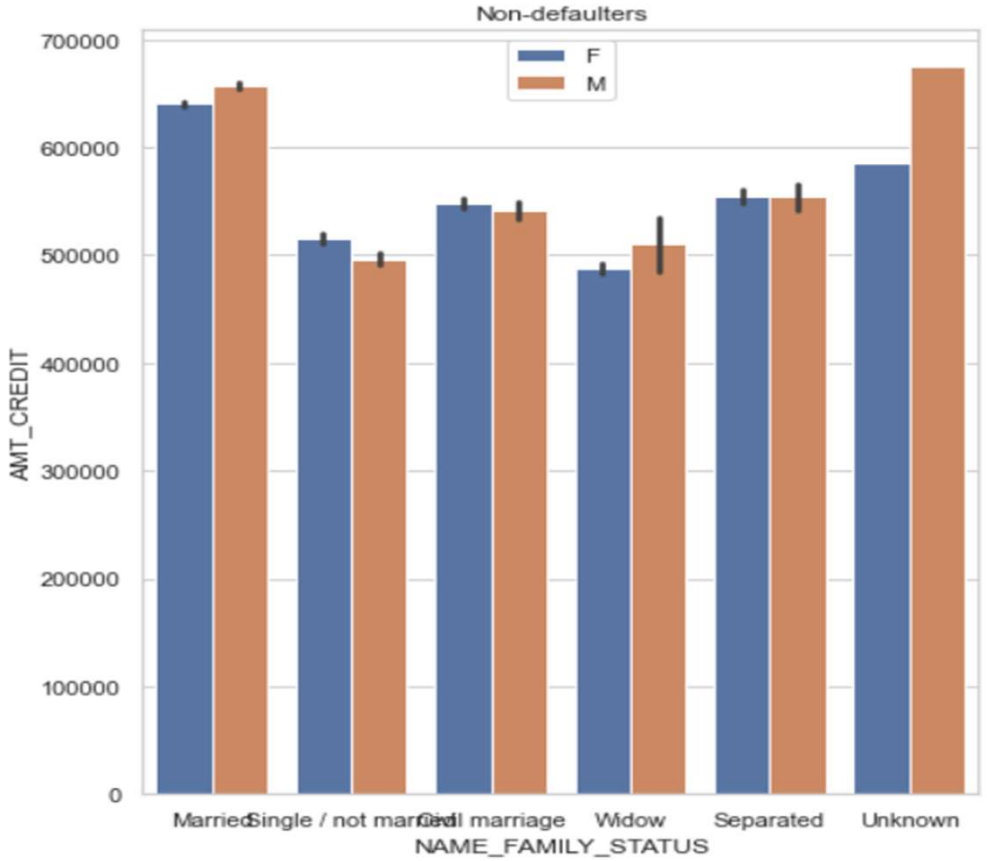


Observation

1. In Defaulters male there are no academic degree client (they don't have any problems in paying their loan) .

bivariate Analysis
Target :AMT_CREDIT vs CODE_GENDER and Family status
Objective: to understand the AMT_CREDIT vs CODE_GENDER and Family status of the client in defaulters vs Non- defaulters

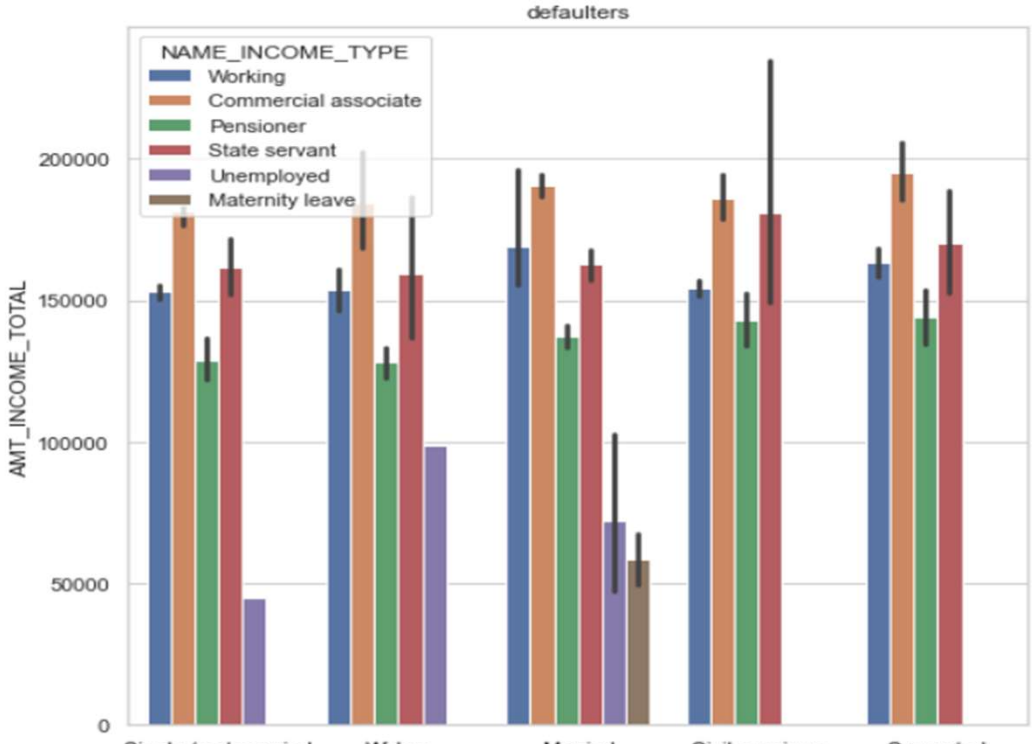


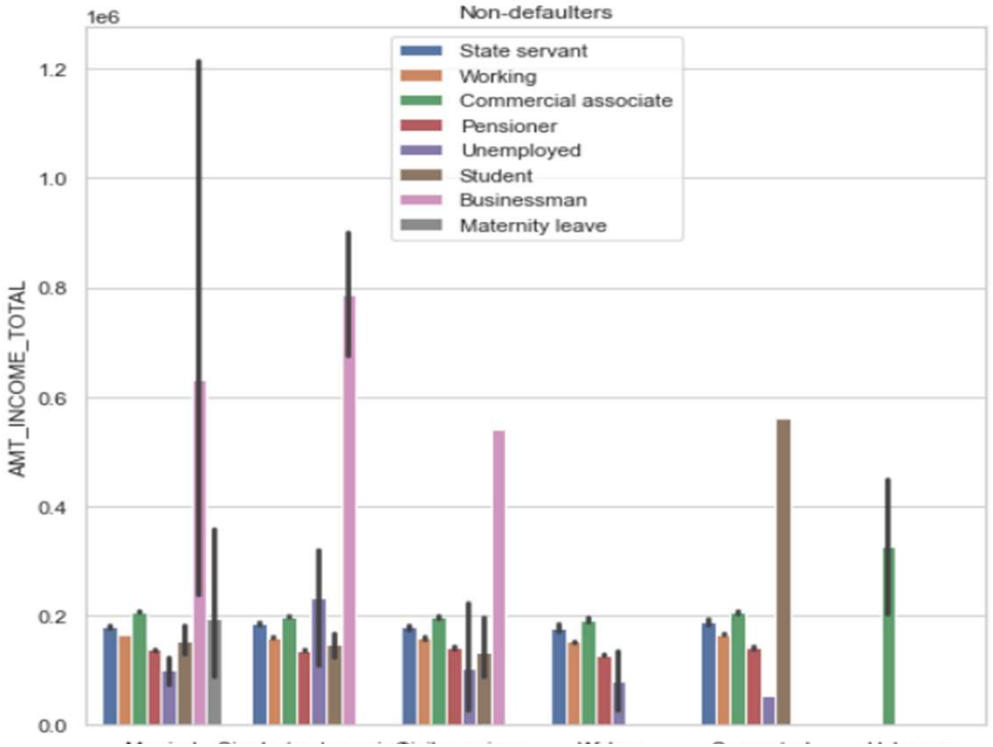


Observation

1. In Defaulters both male and female those are married have problems in paying their loan.
2. In defaulters female(widow) have problems in paying their loan as compare to male (widow).

bivariate Analysis
Target :Source of Income vs Total Income and Family status
Objective: to understand the Source of Income vs Total Income and Family status of the client in defaulters vs Non- defaulters





Observation

1. In Defaulters Separated(Commercial Associate) are highest salaried client (it means less payment difficulties)
2. In defaulters married(maternity leaves) have problems in paying their loan. 3. In defaulters not married(unemployed) have problems in paying their loan.

Top Correlation

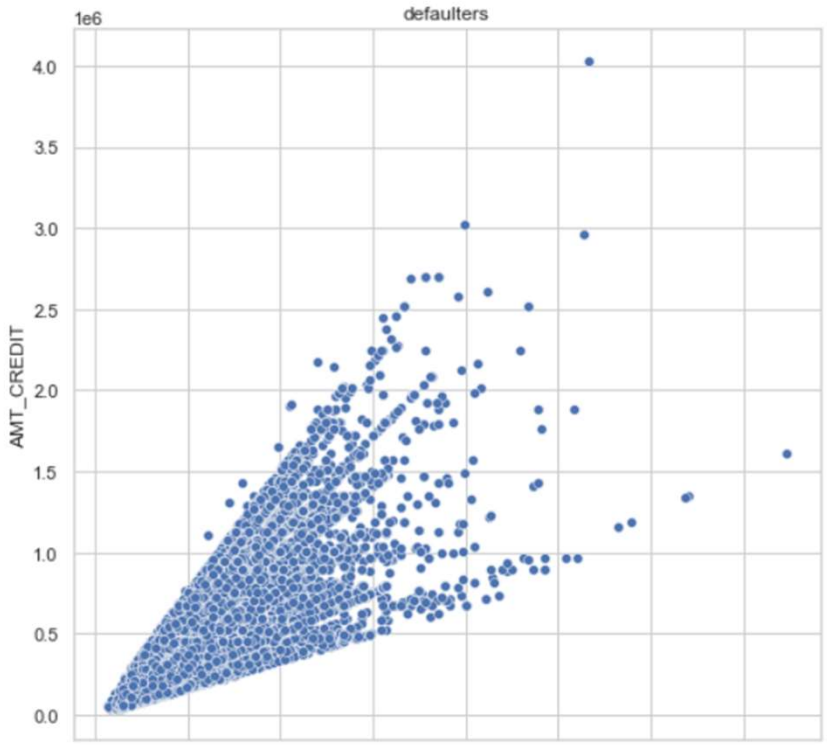
1. for Defaulters

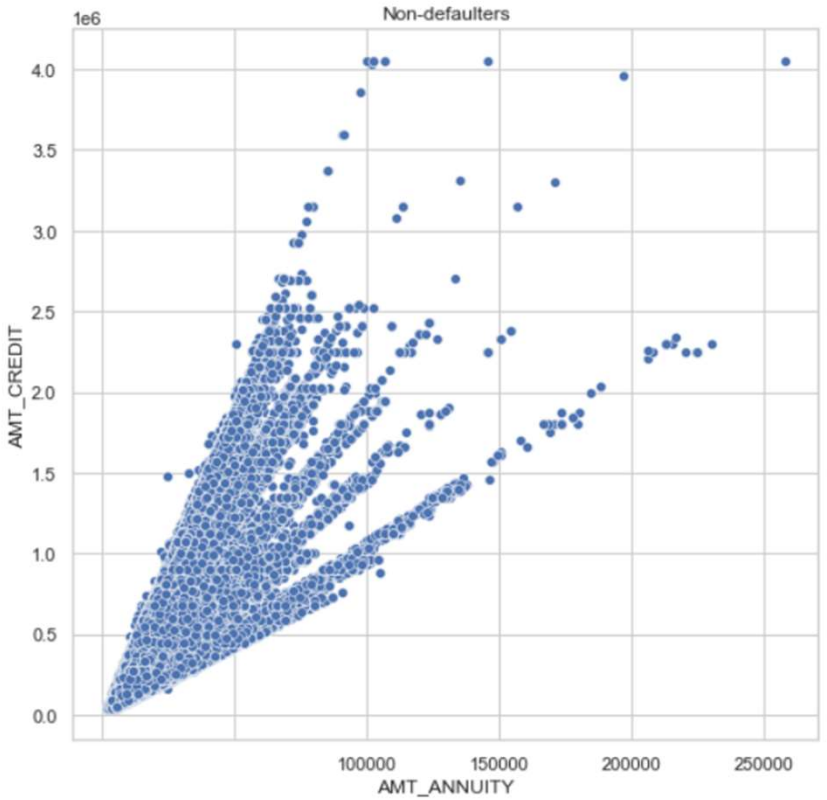
AMT_CREDIT	AMT_GOODS_PRICE	0.98278
CNT_FAM_MEMBERS	CNT_CHILDREN	0.88548
AMT_GOODS_PRICE	AMT_ANNUITY	0.75230
AMT_CREDIT	AMT_ANNUITY	0.75219
Age	DAYS_EMPLOYED	0.58244
DAYS_BIRTH	DAYS_EMPLOYED	0.58219
Age	DAYS_REGISTRATION	0.28912
DAYS_REGISTRATION	DAYS_BIRTH	0.28911
DAYS_BIRTH	DAYS_ID_PUBLISH	0.25286
DAYS_ID_PUBLISH	Age	0.25226

2. for Non-Defaulters

AMT_CREDIT	AMT_GOODS_PRICE	0.98702
CNT_CHILDREN	CNT_FAM_MEMBERS	0.87857
AMT_ANNUITY	AMT_GOODS_PRICE	0.77642
AMT_ANNUITY	AMT_CREDIT	0.77130
DAYS_EMPLOYED	DAYS_BIRTH	0.62611
Age	DAYS_EMPLOYED	0.62603
AMT_INCOME_TOTAL	AMT_ANNUITY	0.41895
AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.34943
AMT_CREDIT	AMT_INCOME_TOTAL	0.34280
DAYS_BIRTH	DAYS_REGISTRATION	0.33315

bivariate Analysis (numerical vs numerical
Target :AMT_ANNUITY vs AMT_CREDIT
Objective: to understand the correlation in defaulters vs Non- defaulters

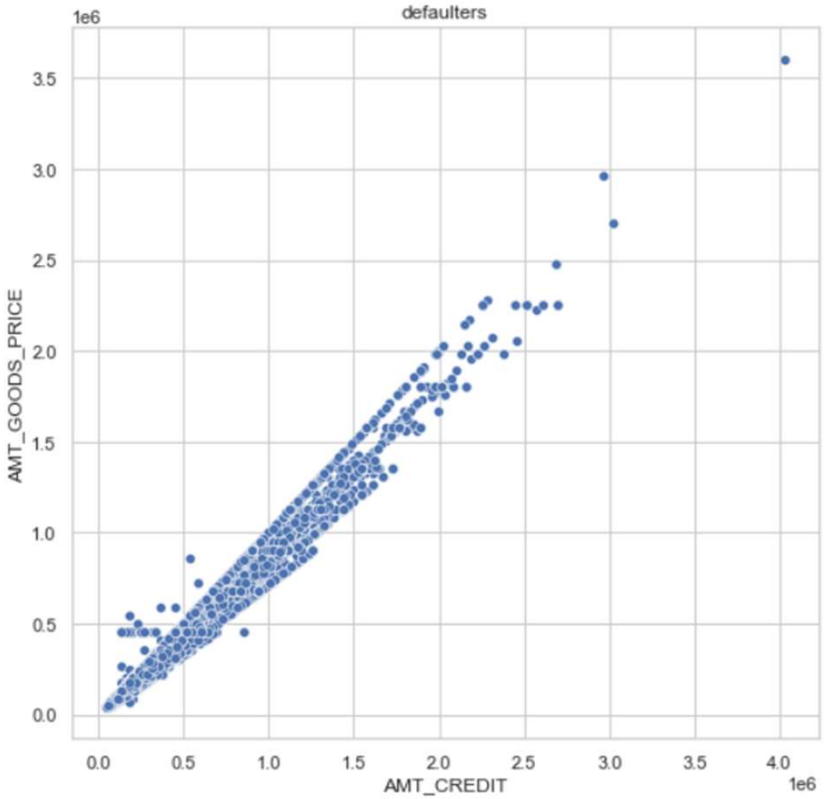


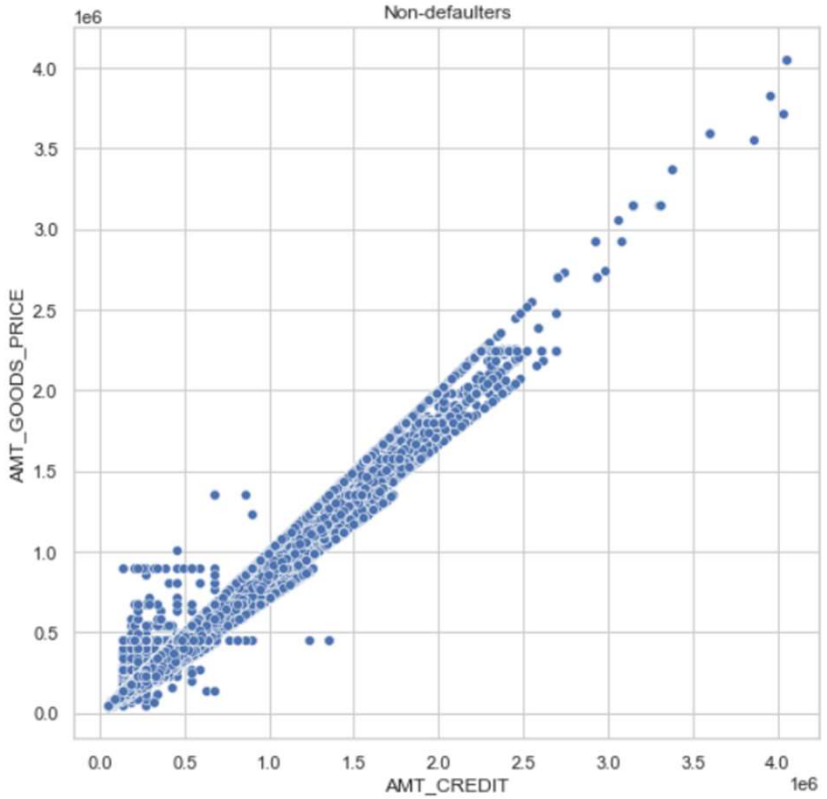


Observation

1. As we can see AMT_ANNUITY and AMT_CREDIT are positive correlated in both the cases defaulters and non-defaulters.
2. so as loan annuity increases the amount of credit also increases.

bivariate Analysis (numerical vs numerical
Target :AMT_GOODS_PRICE vs AMT_CREDIT
Objective: to understand the correlation in defaulters vs Non- defaulters



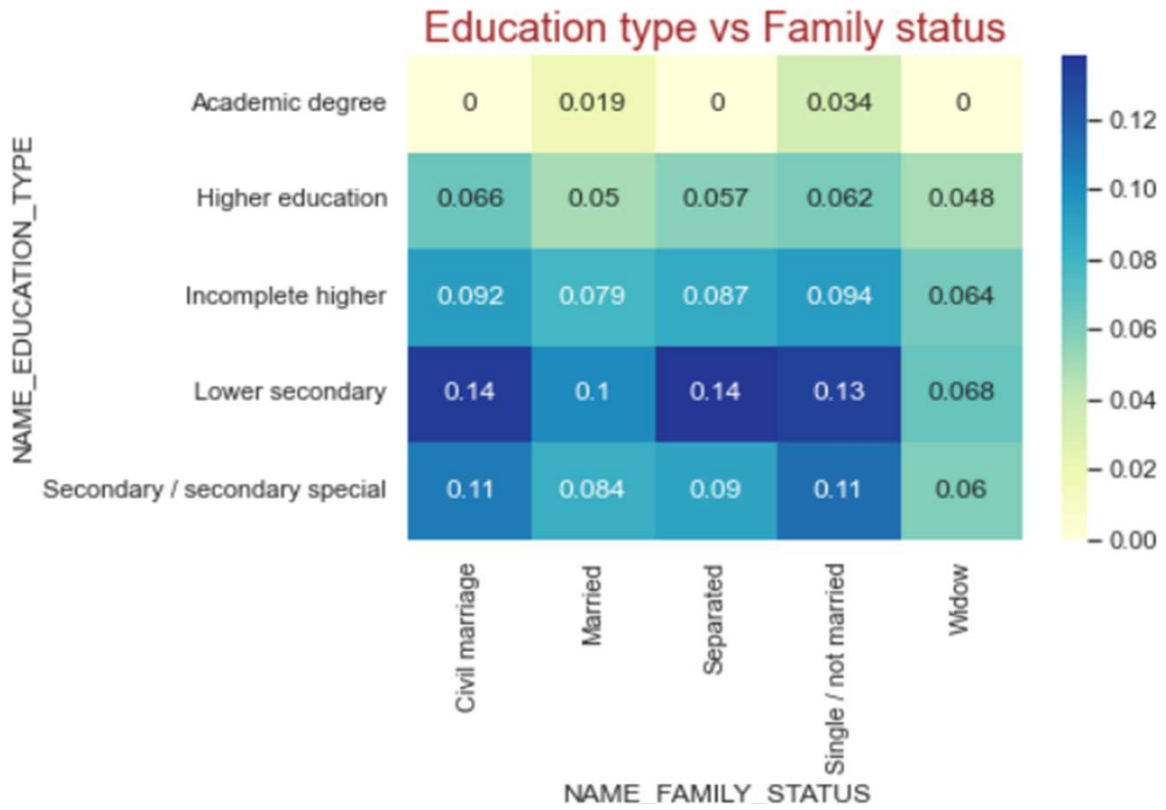


Observation

1.As we can see AMT_GOOD_PRICE and AMT_CREDIT are positive correlated in both the cases defaulters and non-defaulters.

2.so as goods price increases the amount of credit also increases.

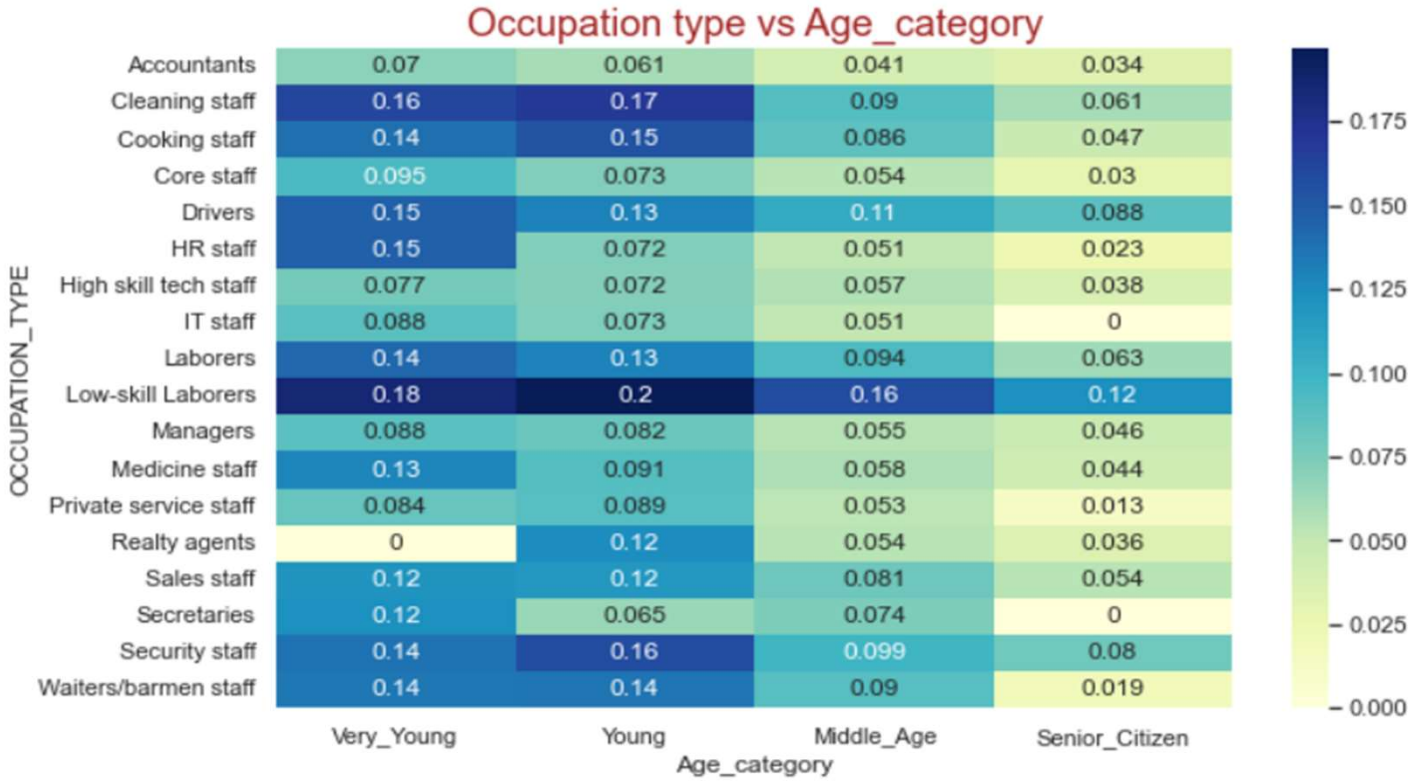
Multivariate Analysis



Observation---→As we set the center mean of the target variable (if the correlation is near to 1 then chances to become defaulter is more)

1. chances of becoming defaulter is more of (separated with lower secondary and civil marriage with lower secondary)

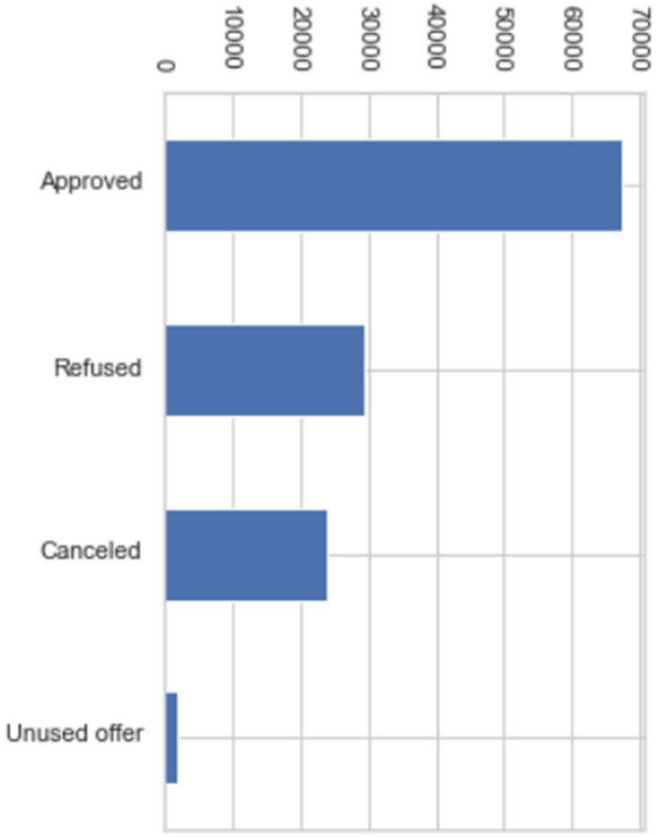
Multivariate Analysis



Observation

AS we can see the low skill laborers correlation is on higher side chances of becoming defaulters is high for them

Analysis after merging



Observation

1.most of the client with previously approved loans (67243) having problems in paying their loan.

Conclusion

From the analysis so far conducted we can conclude the following

1. In both Defaulters and non Defaulters people prefer cash loans But in Defaulters this can be seen that they avoid revolving loans.

2. It is important to check the gender and the age of the client as we see in Defaulters male percentage is also high as compare to non defaulters.

3. In defaulters married(maternity leaves) have problems in paying their loan. And In defaulters not married(unemployed) have problems in paying their loan.

4. In defaulters female(widow) have problems in paying their loan.

5. low skill laborers correlation is on higher side chances of becoming defaulters is high for them.

6. chances of becoming defaulter is more of (separated with lower secondary and civil marriage with lower secondary).

7. The clients with previously approved loans are most of them(67243) having more no of difficulties(now become defaulters).

Thank You