CREDIT EDA CASE STUDY

VINAYAK RANE

Business Problem Statement

Identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

INDEX

Data Understanding and Preparation.

1. Read the data from input files.
2. Read each column data types and null value counts.
3. Read the shape of data.
4. Describe the data.
5. Read the data columns description or meaning from excel.

Handling Null Values

1. Identify percentage of null values for each column [inpputdf.isnull().sum()/len(app)*100]
2. Identify columns with null values >40 % from both input data sets. nullcol[nullcol.values>(40/100*len(app))]
3. Drop all the null value columns which are >40%. app.drop((list(nullcol.index)),axis=1,inplace=True)
4. Read null columns one by one by keeping outliers in mind and fill them with median and mode Categorical data with mode and Numerical data with median.
5. Drop the XNA,XAP data from history dataset.

Data Transformation

1. Verify each and every column for incorrect data.
2. Fix the column values XNA with mode of the columns such as CODE_GENDER with median.
3. DAYS* columns should be converted into + values with abs function
4. AGE will be calculated from BIRTH column by converting days into year.
5. Convert Y/N values into 1/0 for all the Flags columns for calculation.
6. Create bins or bucket of categorial data for data distribution such convert age to AGE group, AMT_INCOME_TOTAL to AMT_INCOME_GROUP by using pd.cut or pd.qcut
7. Drop the unnecessary columns from the input data.
8. Convert datatypes into appropriate data type of each column.

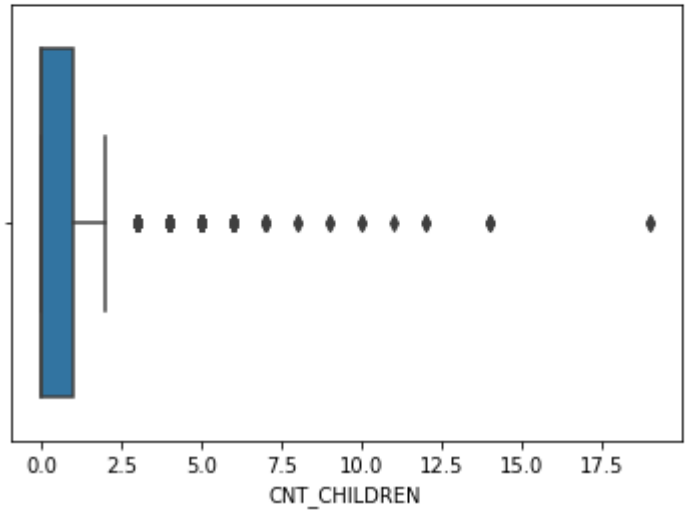
Outlier Analysis

1. Verify number columns for outlier verification.
2. Identify Outlies for each column by using boxplot.
3. Ignore the outliers if not impacting analysis

```
sns.boxplot(app['CNT_CHILDREN'])
```

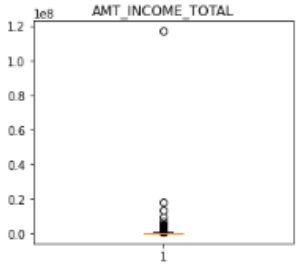
```
C:\Users\Lenovo\anaconda3\lib\site-packages\seaborn\_dec
ing variable as a keyword arg: x. From version 0.12, the
`, and passing other arguments without an explicit keywo
on.
  warnings.warn(
```

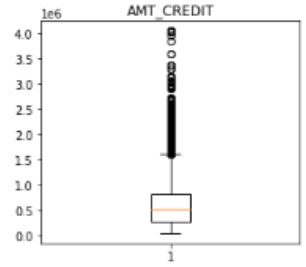
```
<AxesSubplot:xlabel='CNT_CHILDREN'>
```

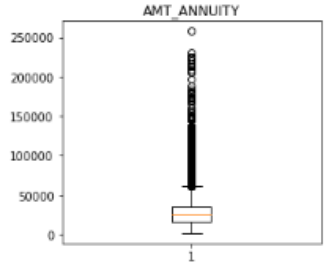


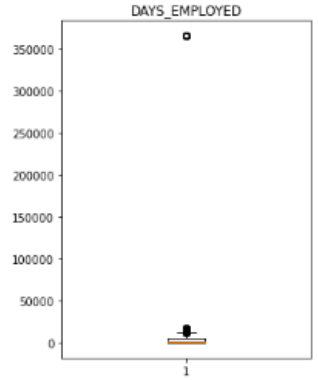
Observation
CNT_CHILDREN column has an outlier, some applicant with count of children's more than 5 which is consider as outlier

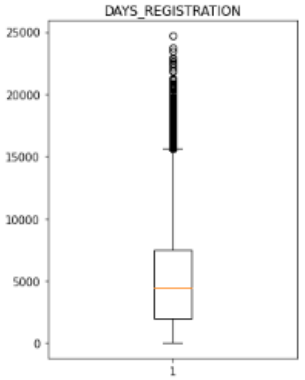
Outlier Analysis - Conti

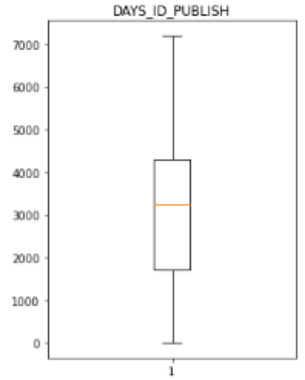


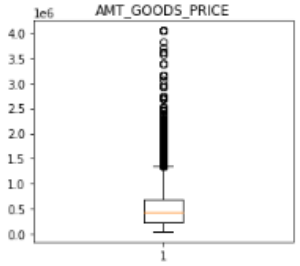


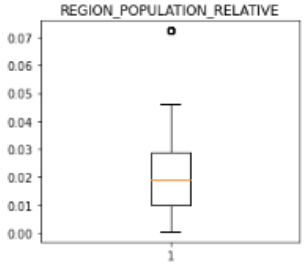


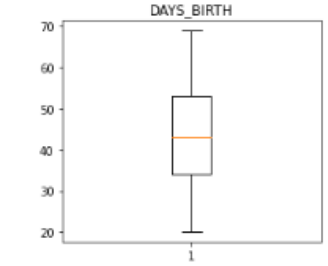


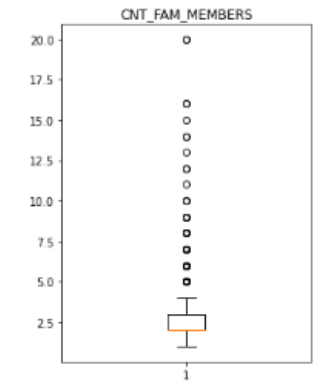


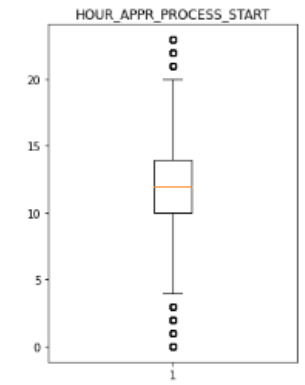


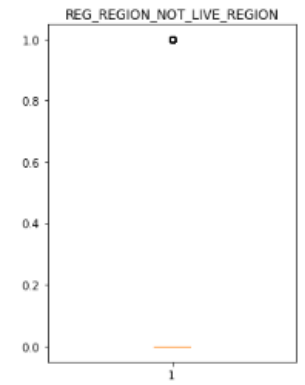


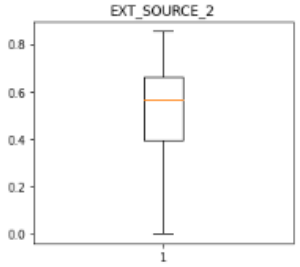


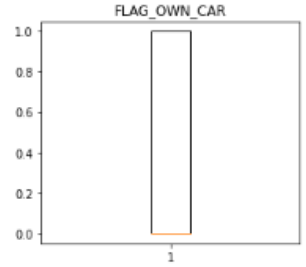












Outlier Analysis - Conti

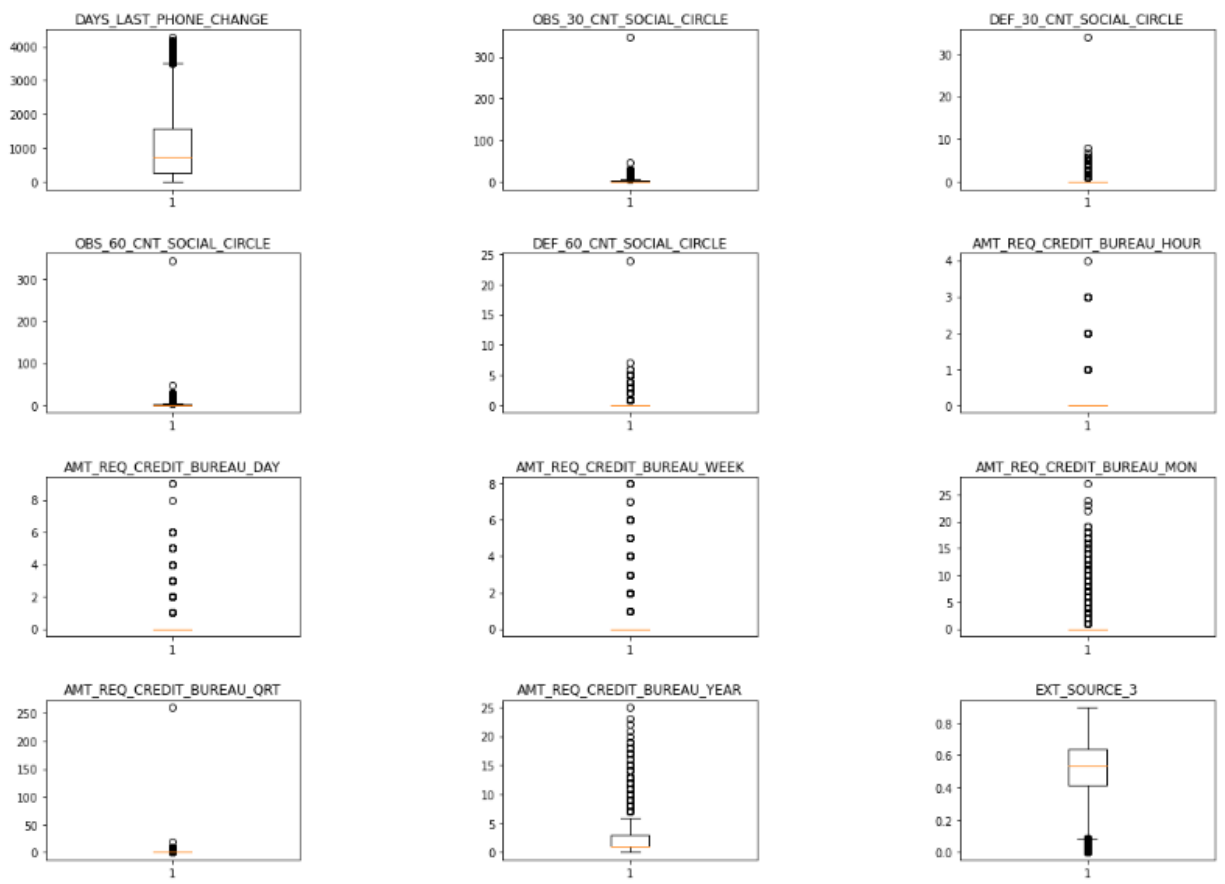
AMT_INCOME_TOTAL, AMT_CREDIT ,AMT_ANNUITY, AMT_GOODS_PRICE, REGION_POPULATION_RELATIVE has an outlier values

DAYS_LAST_PHONE_CHANGE , DAYS_EMPLOYED ,OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE,OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE,AMT_REQ_CREDIT_BUREAU_HOUR,AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK,AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT and AMT_REQ_CREDIT_BUREAU_YEAR has a large number of outliers

FLAG_OWN_REALTY and FLAG_OWN_CAR doesn't have First and Third quantile and values lies within IQR , we can conclude that most of the clients own a car and House

DAYS_EMPLOYED, DAYS_REGISTRATION, CNT_FAM_MEMBERS has an outlier values

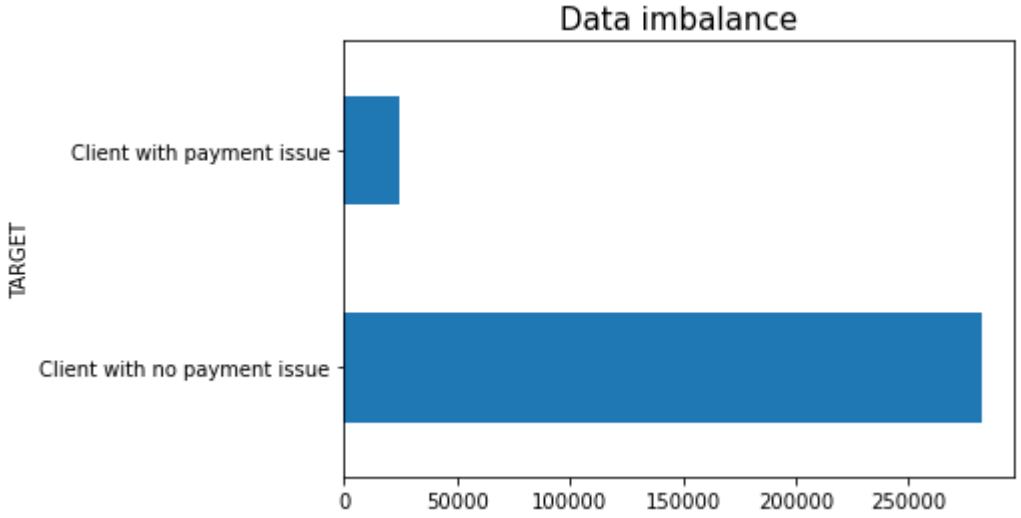
DAYS_BIRTH ,DAYS_ID_PUBLISH and EXT_SOURCE_2,EXT_SOURCE_3 don't have any outliers.

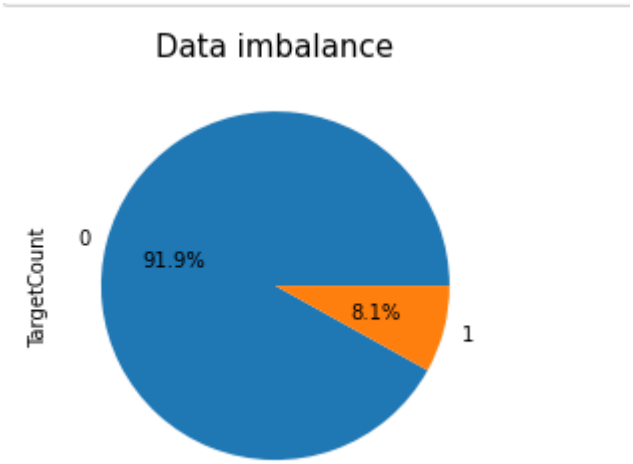


DATA ANALYSIS

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) So we can divide the whole data into 2 parts, 1 with clients without payment difficulties and other with late payments and calculate data imbalance

Data is highly imbalanced. Population with payment issue is 8.1 % and population without payment issue is 91.9% with Ratio 11.39



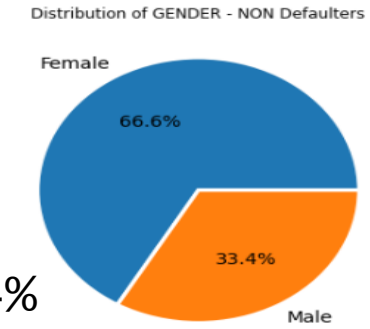


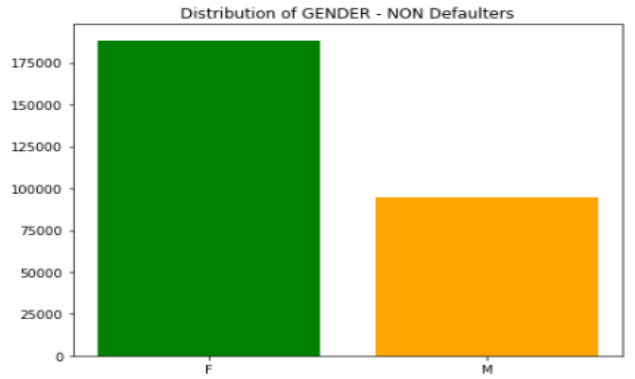
Univariate, Bivariate Analysis and Multivariate Analysis

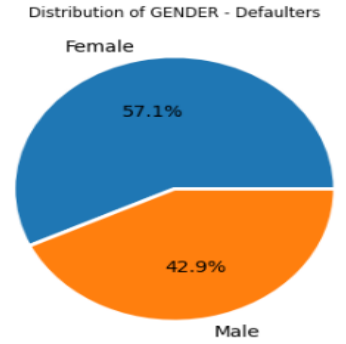
Female clients applied higher than male clients for loan

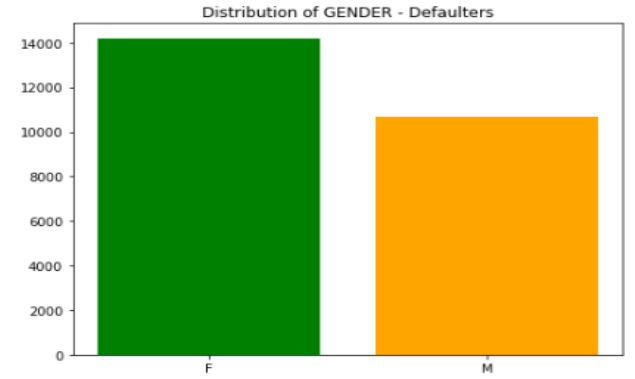
66.6% Female clients are non-defaulters while 33.4% male clients are non-defaulters

57% Female clients are defaulters while 42% male clients are defaulters.



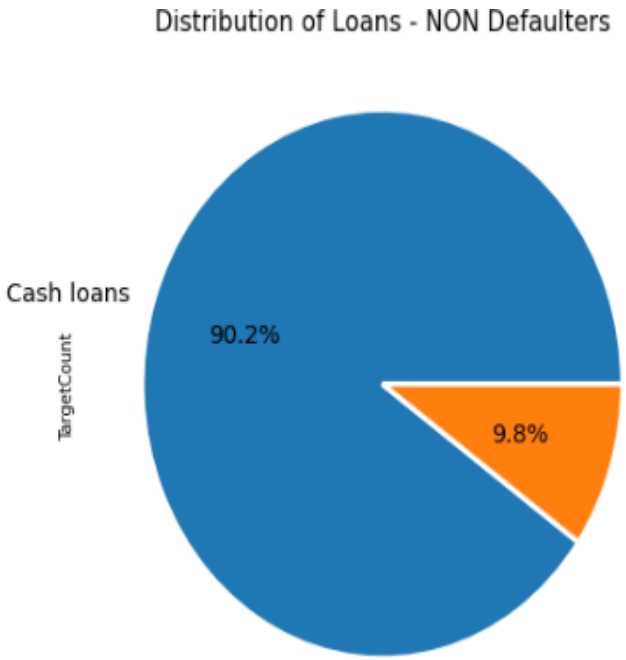


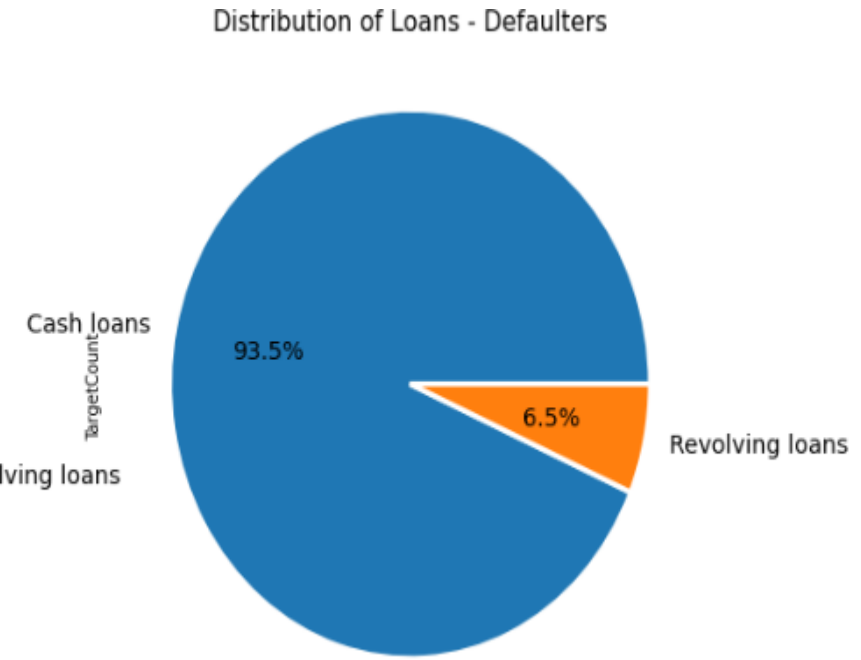




Univariate, Bivariate Analysis and Multivariate Analysis - Conti

Non Defaulters - 90.2% clients applied for cash loans while 9.8 % clients applied for Revolving loans
Defaulters - 93.5% clients applied for cash loans while 6.5 % clients applied for Revolving loans

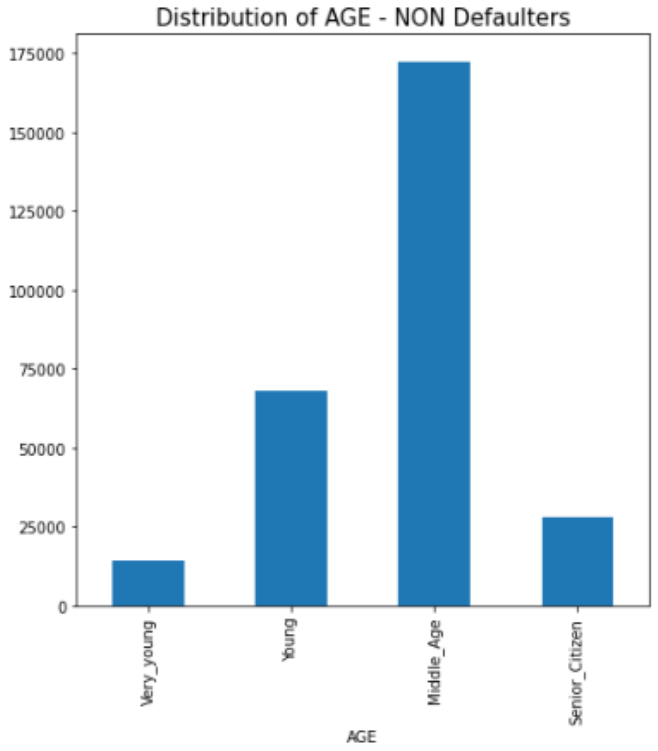


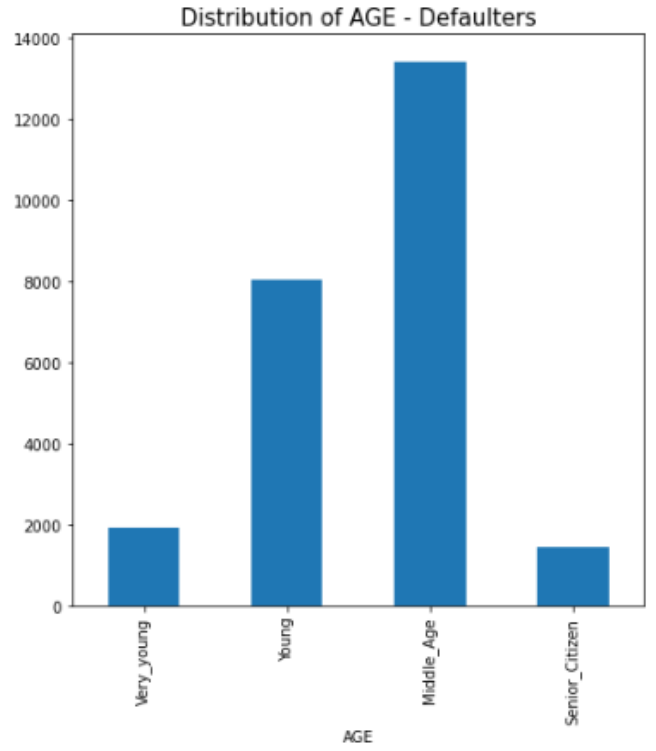


Univariate, Bivariate Analysis and Multivariate Analysis - Conti

Middle age (35-60) clients seems to be applied higher than any other age group in case if both defaulters and non defaulters

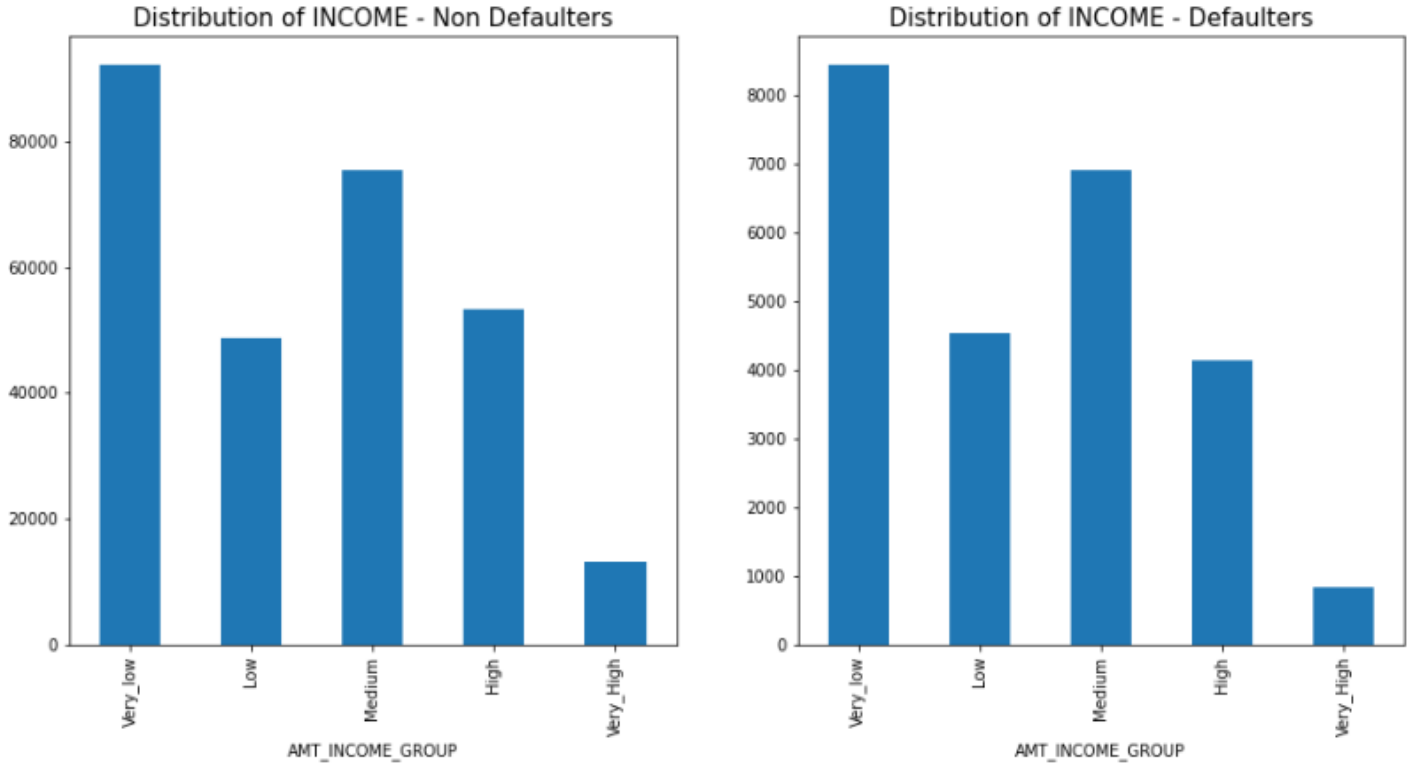
Middle age group having lot of paying difficulties but at the same time Very young (18-25) and senior citizens (60>) do not face payment issues





Univariate, Bivariate Analysis and Multivariate Analysis - Conti

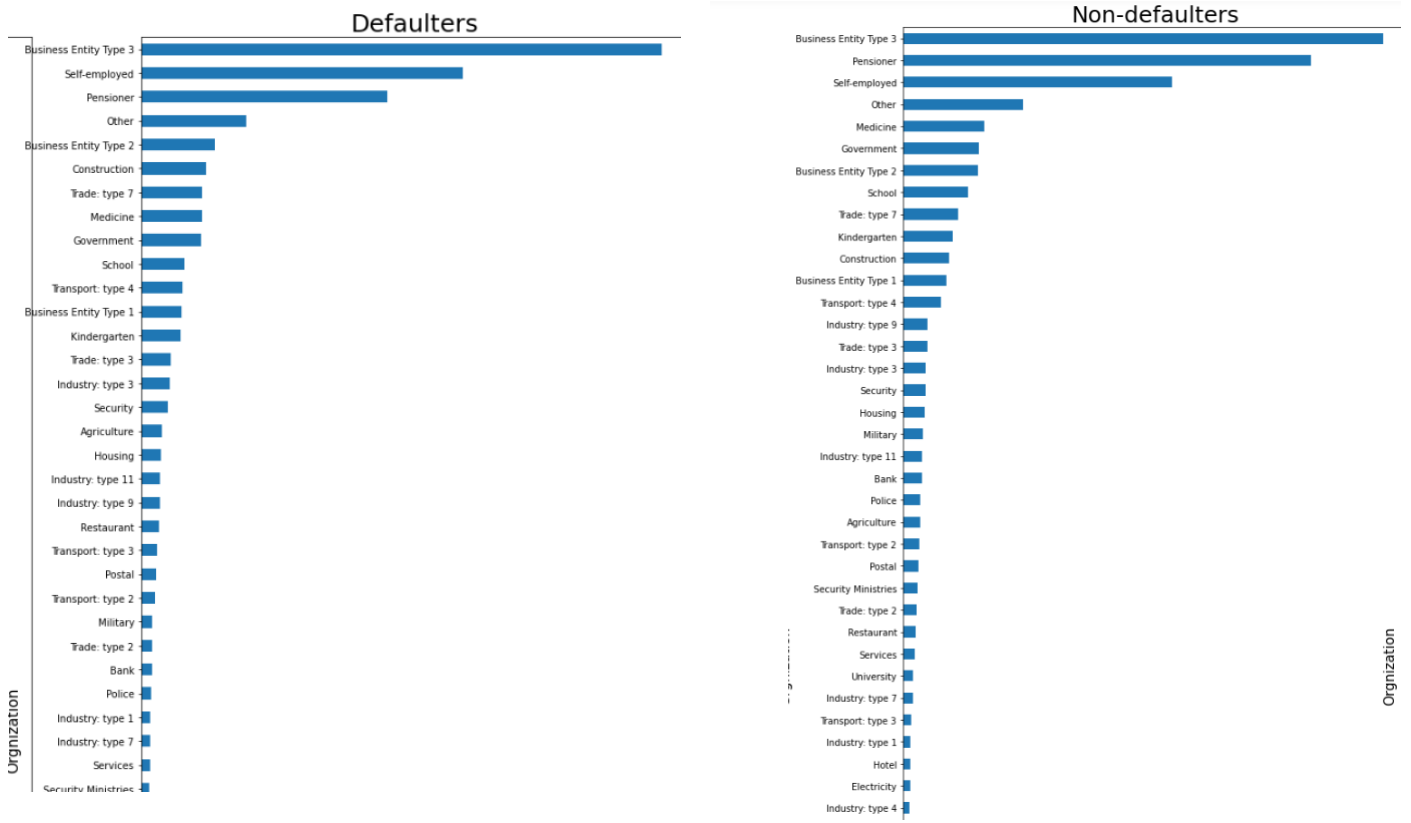
We have high number of clients with very low income in case of both defaulters and non defaulters data
Medium and Low salary have high risk to default



Univariate, Bivariate Analysis and Multivariate Analysis - Conti

Clients with ORGANIZATION_TYPE Business Entity Type 3, Self-employed, Other ,Medicine, Government, Business Entity Type 2 applied the most for the loan as compared to others in both the cases

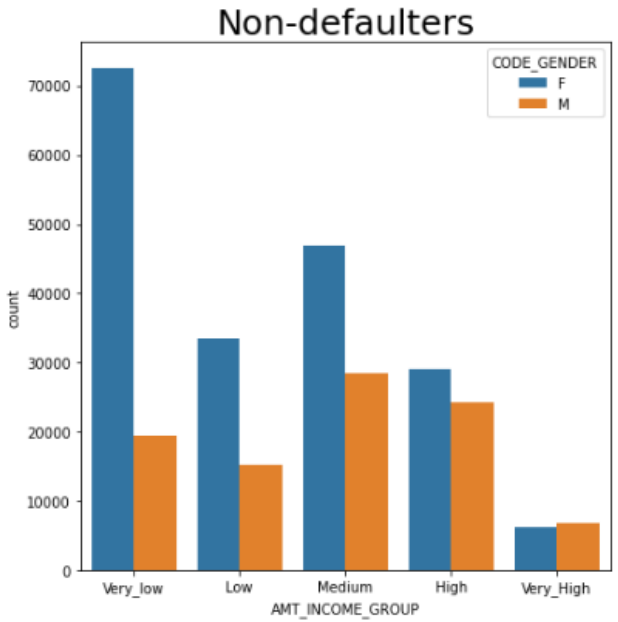
Clients having ORGANIZATION_TYPE Industry: type 13, Trade: type 4, Trade: type 5, Industry: type 8 applied lower for the loan as compared to others.

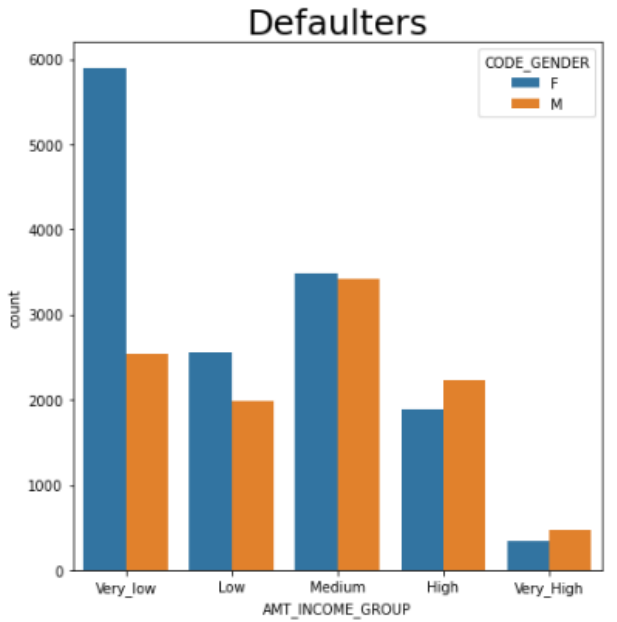


Univariate, Bivariate Analysis and Multivariate Analysis - Conti

We have Female clients with very low income who applied for loans with both defaulter and non defaulter cases

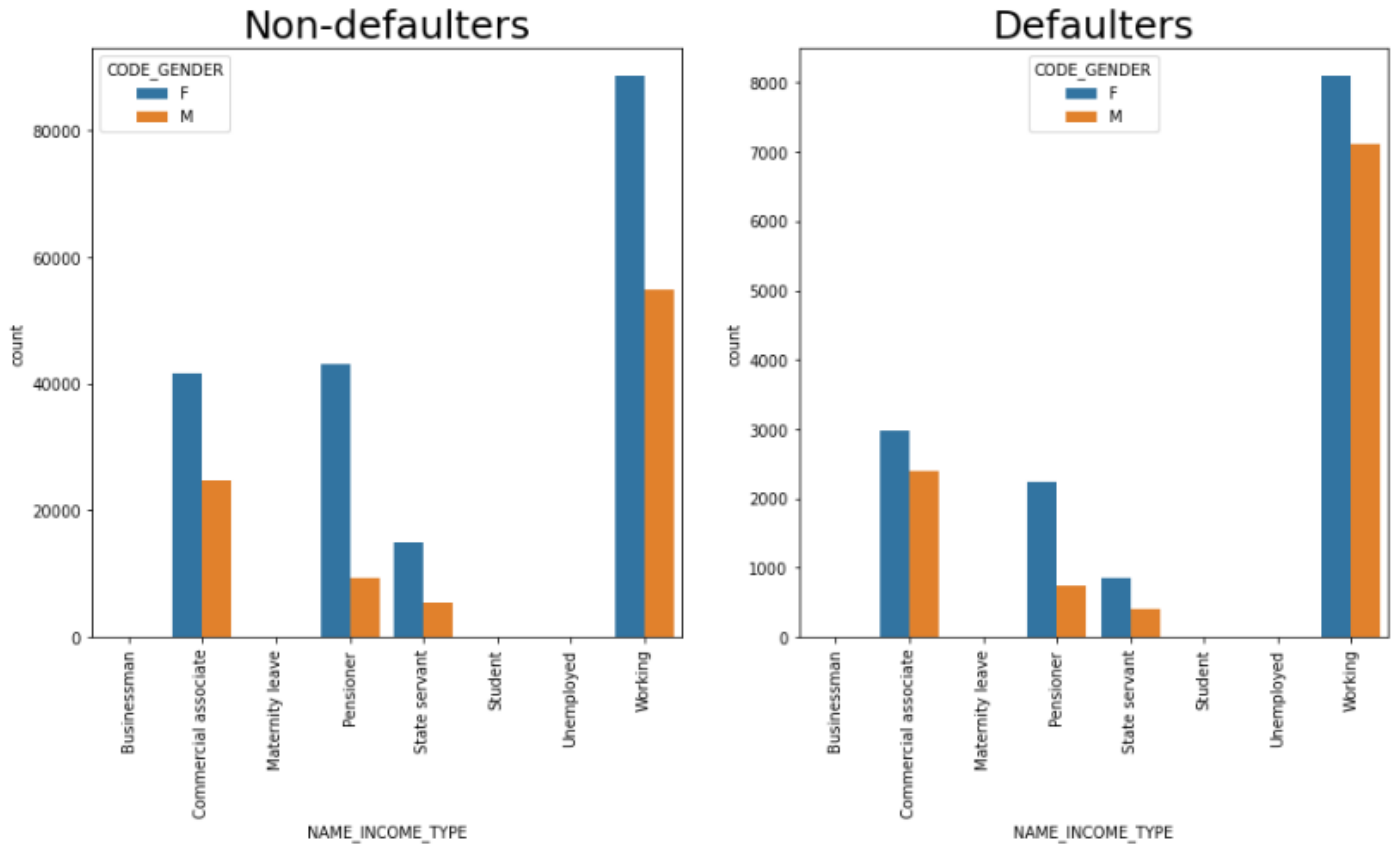
We have very less population with very high income who applied for loan





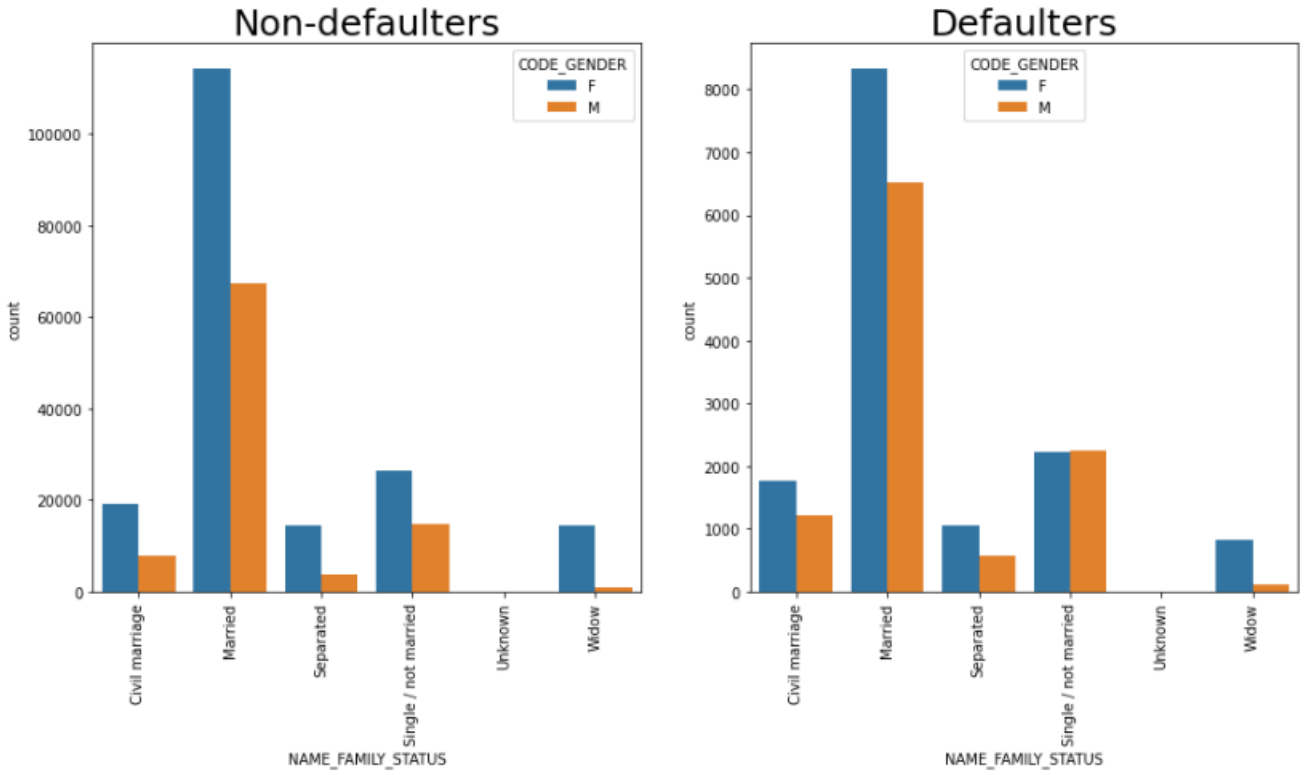
Univariate, Bivariate Analysis and Multivariate Analysis - Conti

We have Female and male clients as a working professionals, applied for loan in defaulter and non defaulter cases
we have very less population as a student and unemployed who applied for loan.
Working category have high risk to default.
State Servant is at Minimal risk to default



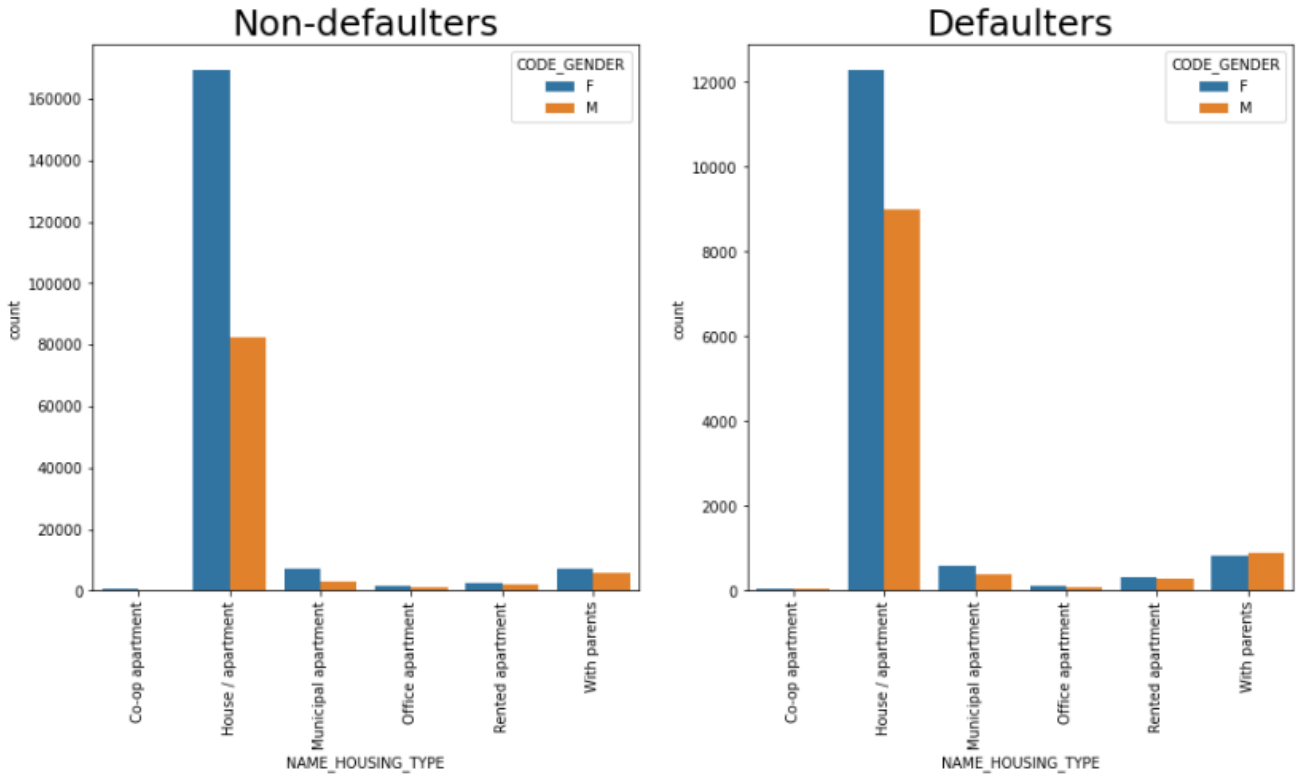
Univariate, Bivariate Analysis and Multivariate Analysis - Conti

We have Married Clients seems to be applied most for loan in defaulter and non defaulter cases
In case of Defaulters, Clients having single relationship are less risky
Widows are with Minimal risk.



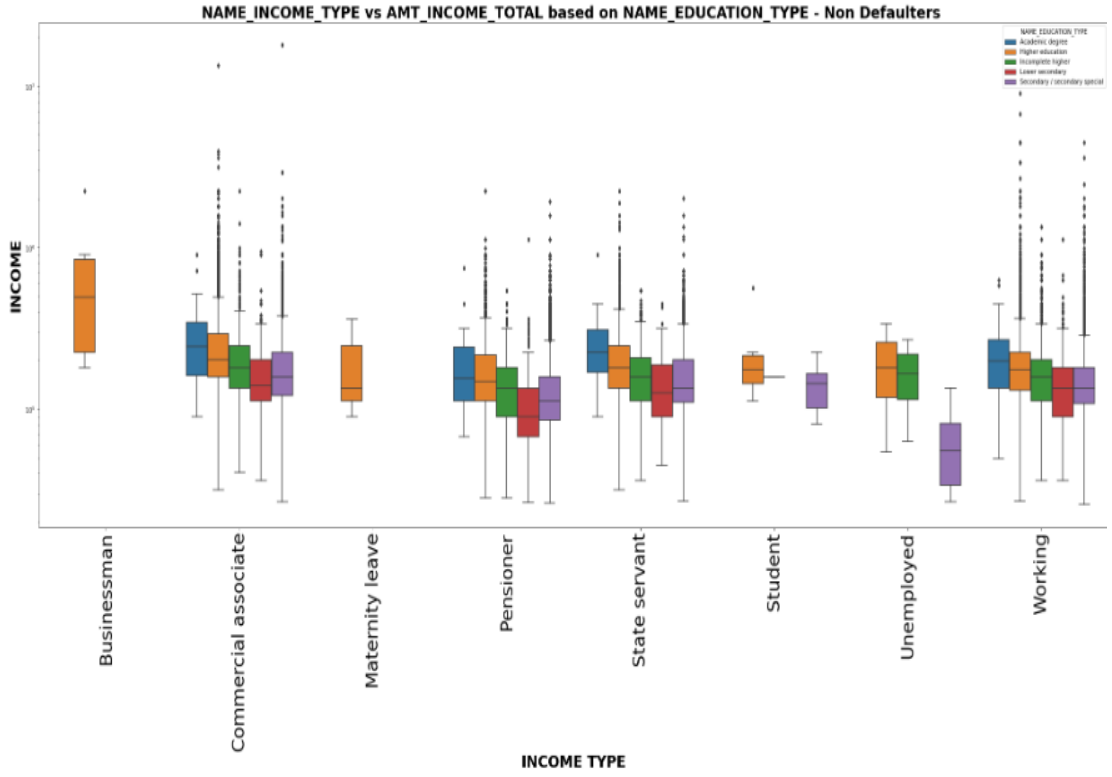
Univariate, Bivariate Analysis and Multivariate Analysis - Conti

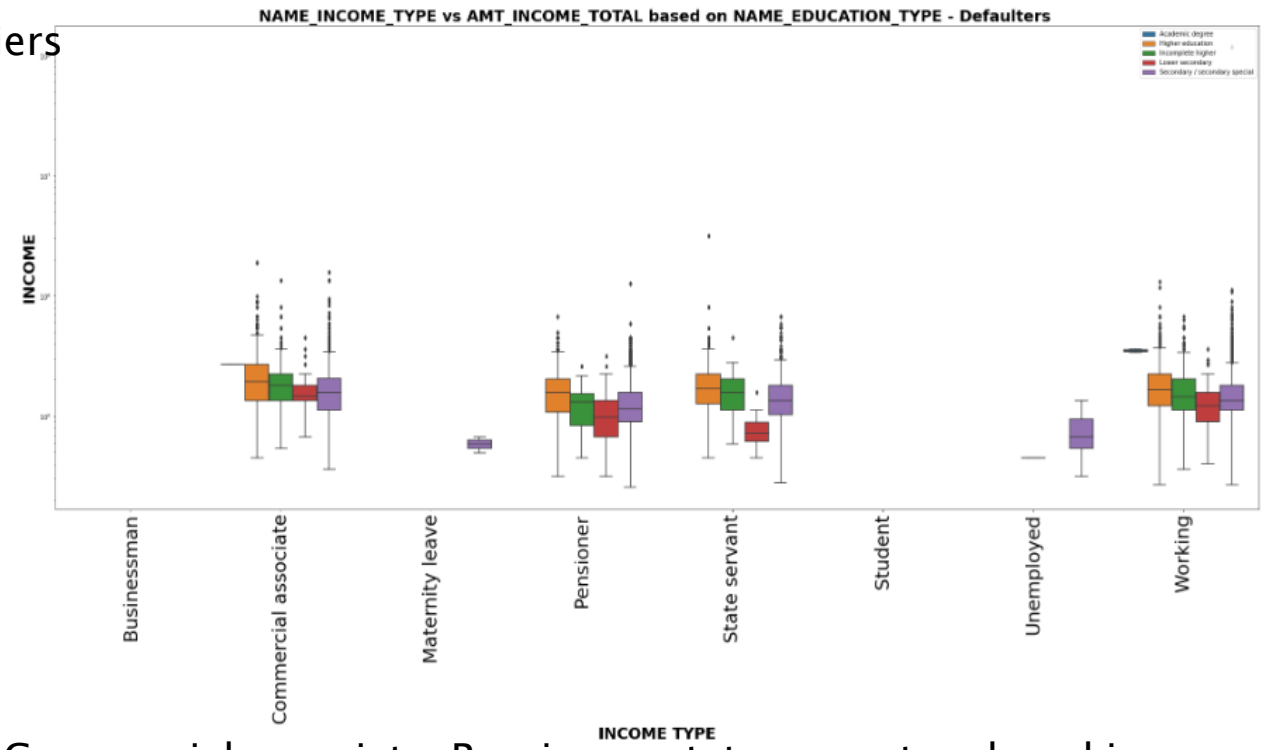
We have clients living in House or apartment to be applied most for loan in defaulter and non defaulter cases
In case of Defaulters, Clients with parents , office, municipal and rented apartment are less risky
Clients living in House or apartment are highly risky



Univariate, Bivariate Analysis and Multivariate Analysis - Conti

Business clients with Higher education having very few outliers in data followed by Maternity, student, unemployed clients

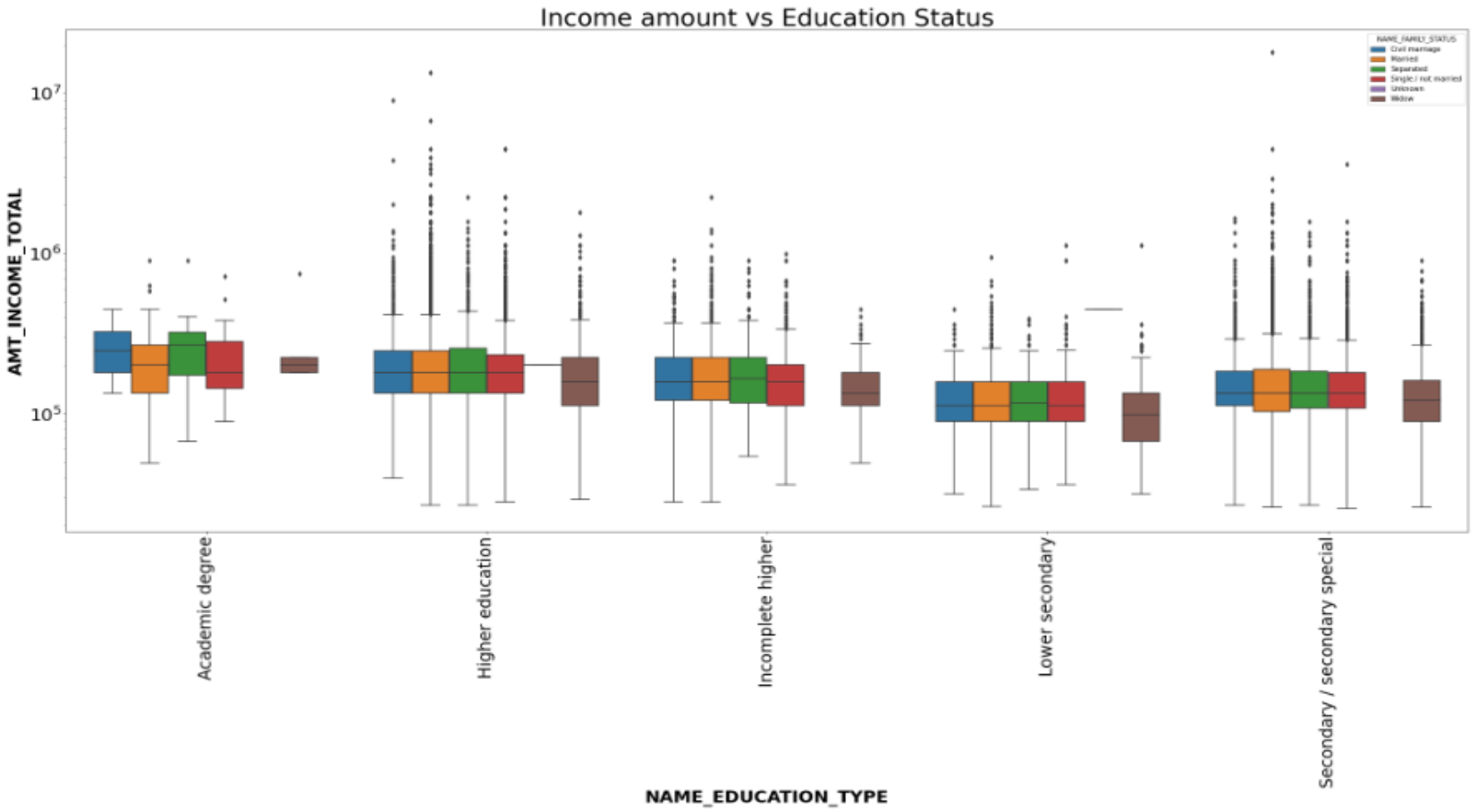




Commercial associate, Pensioner, state servant and working clients applied most with high number of outliers
Very less chances when business clients and students will get default in loan repayment

Univariate, Bivariate Analysis and Multivariate Analysis - Conti

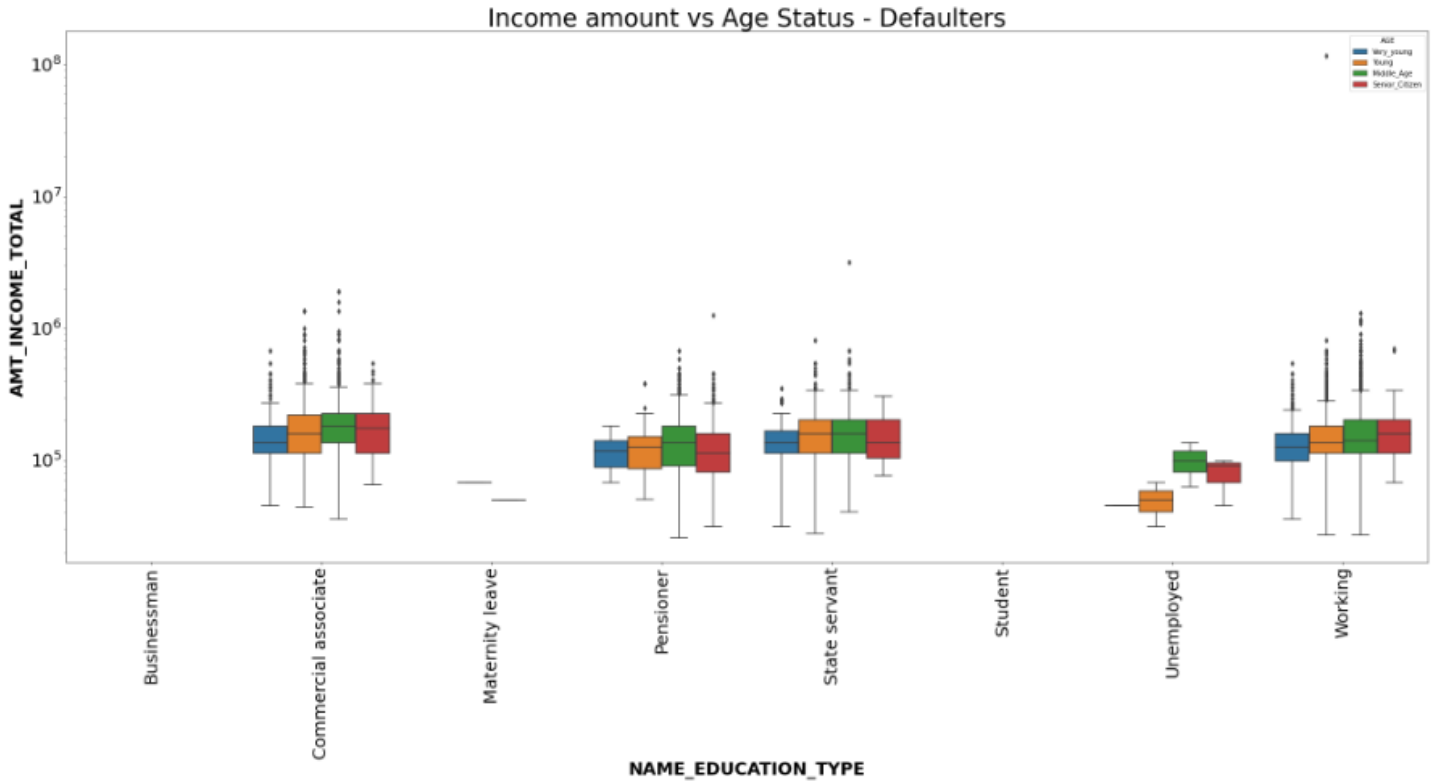
Clients having Higher Education have the highest income compared to others.
Also maximum outliers we can see in higher education client data
Some of the clients who haven't completed their Higher Education and with Secondary/Secondary Special Education also has higher incomes and outliers



Univariate, Bivariate Analysis and Multivariate Analysis - Conti

Clients having working, commercial, pensioner and state servant with all age group contain outliers

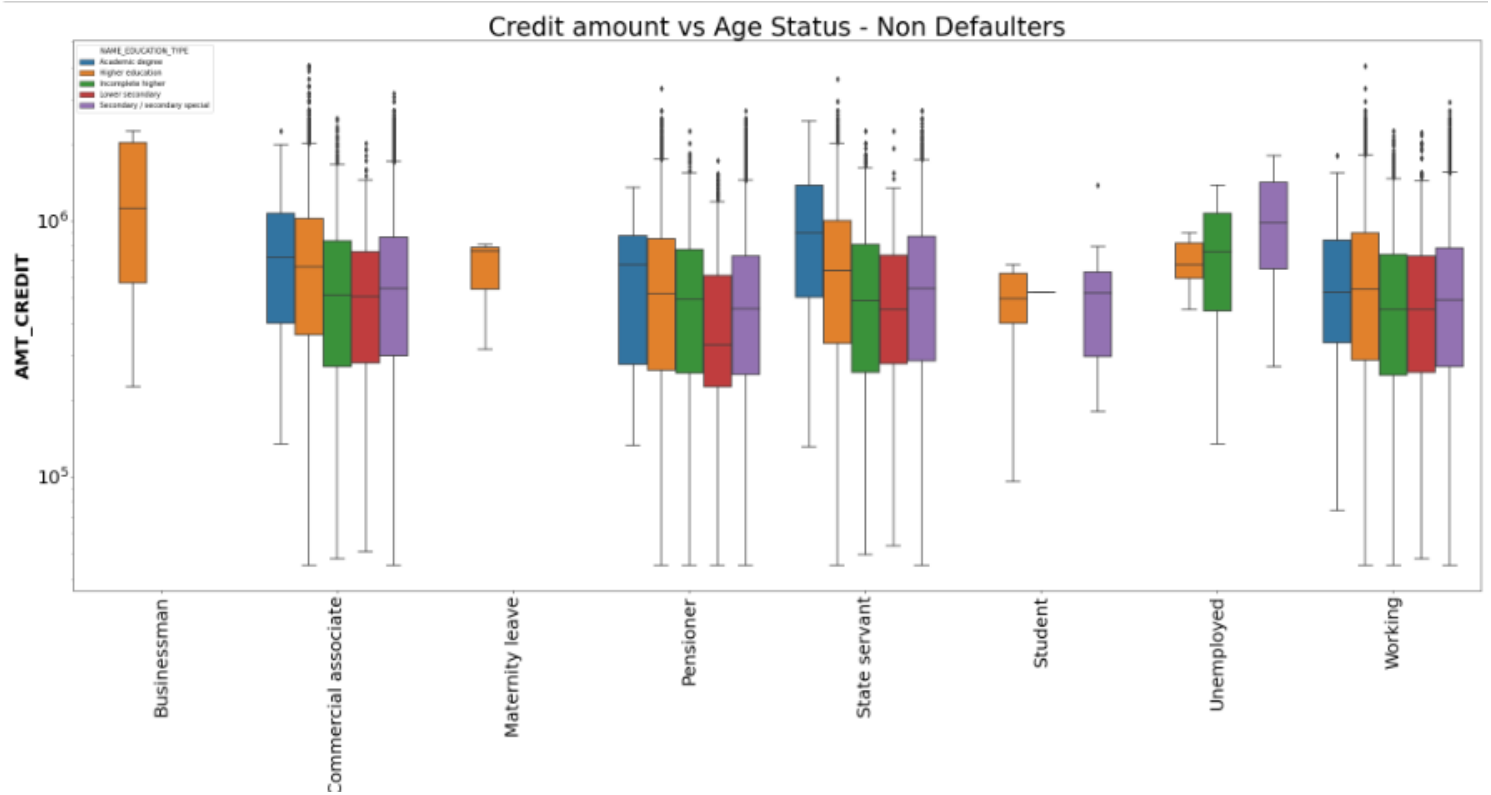
Business Man and Students are not possible to make late payments in any age group



Univariate, Bivariate Analysis and Multivariate Analysis - Conti

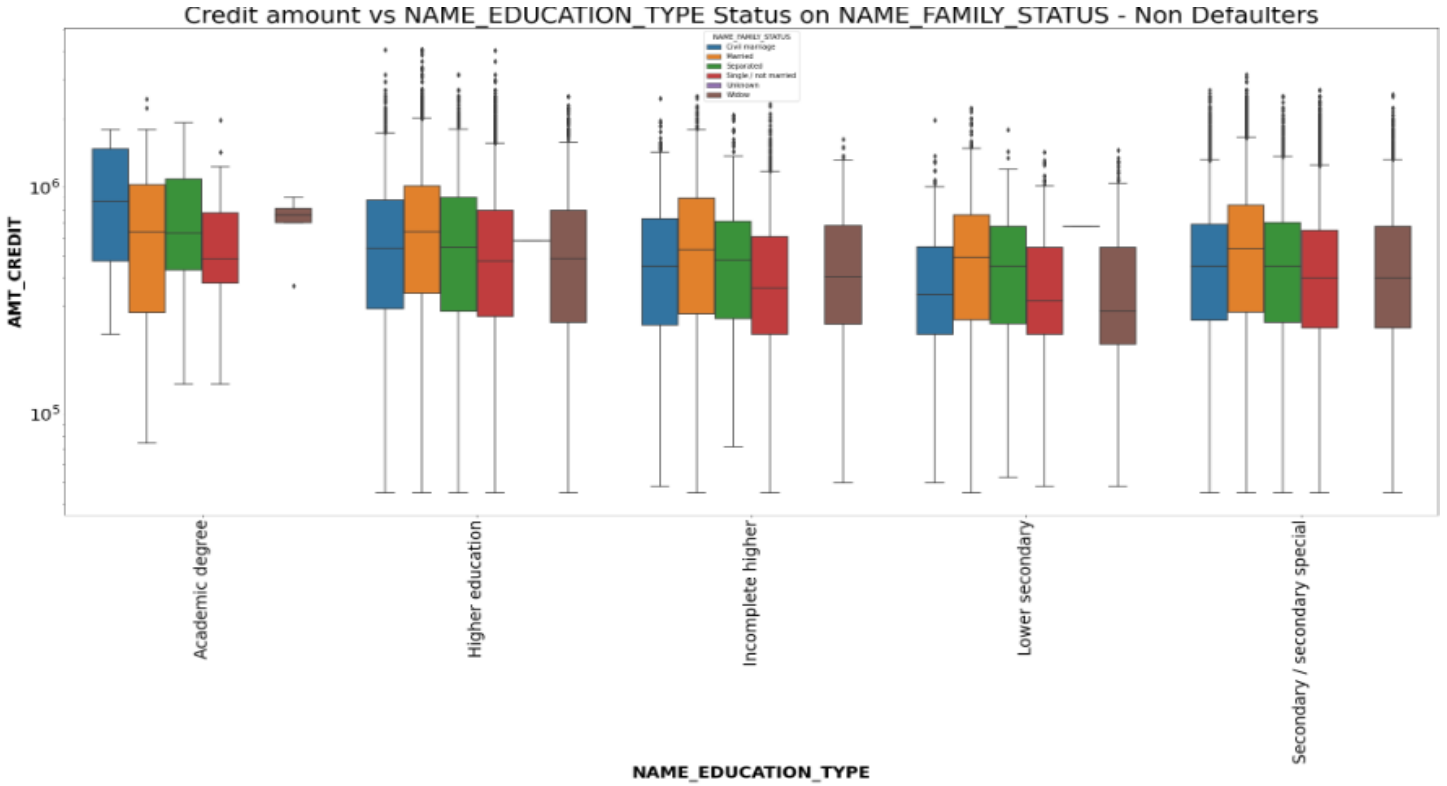
Students with higher education are likely to take less credit loan and with no outlier with Businessman having higher education

Working professionals along with Commercial associate, Pensioner and state servant holds highest outlier

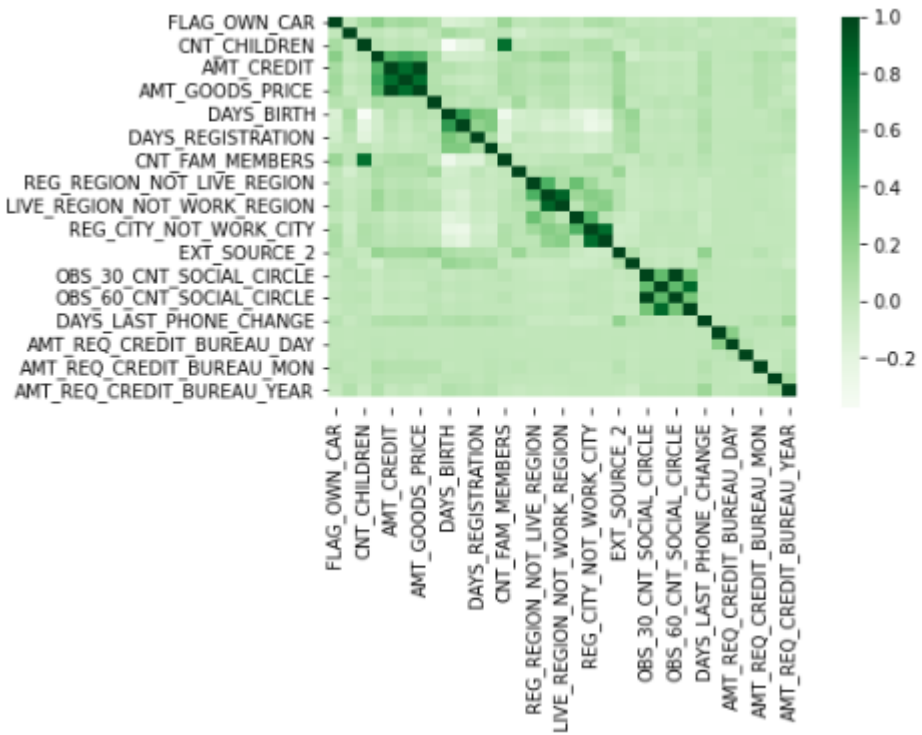


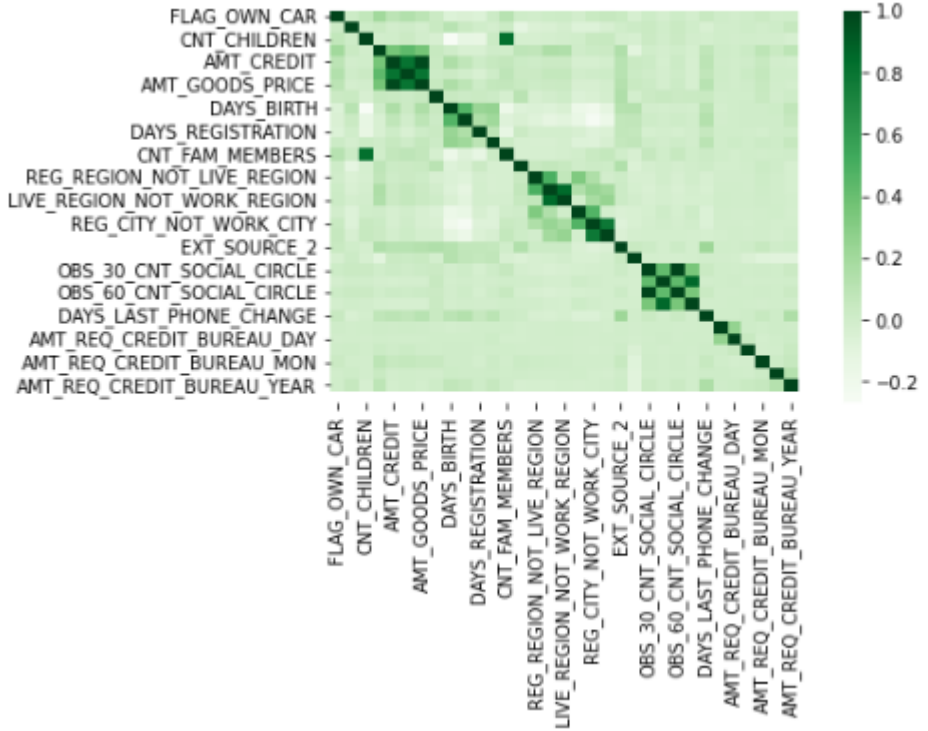
Univariate, Bivariate Analysis and Multivariate Analysis - Conti

Clients with Higher Education, Incomplete Higher Education, Lower Secondary Education and Secondary/Secondary Special Education are more likely to take a high amount of credit loans.



Univariate, Bivariate Analysis and Multivariate Analysis - Conti





Univariate, Bivariate Analysis and Multivariate Analysis - Conti

Credit amount is higher to densely populated area

The income is also higher in densely populated area.

Income amount is inversely proportional to the number of children client have, means more income for less children client have and vice-versa.

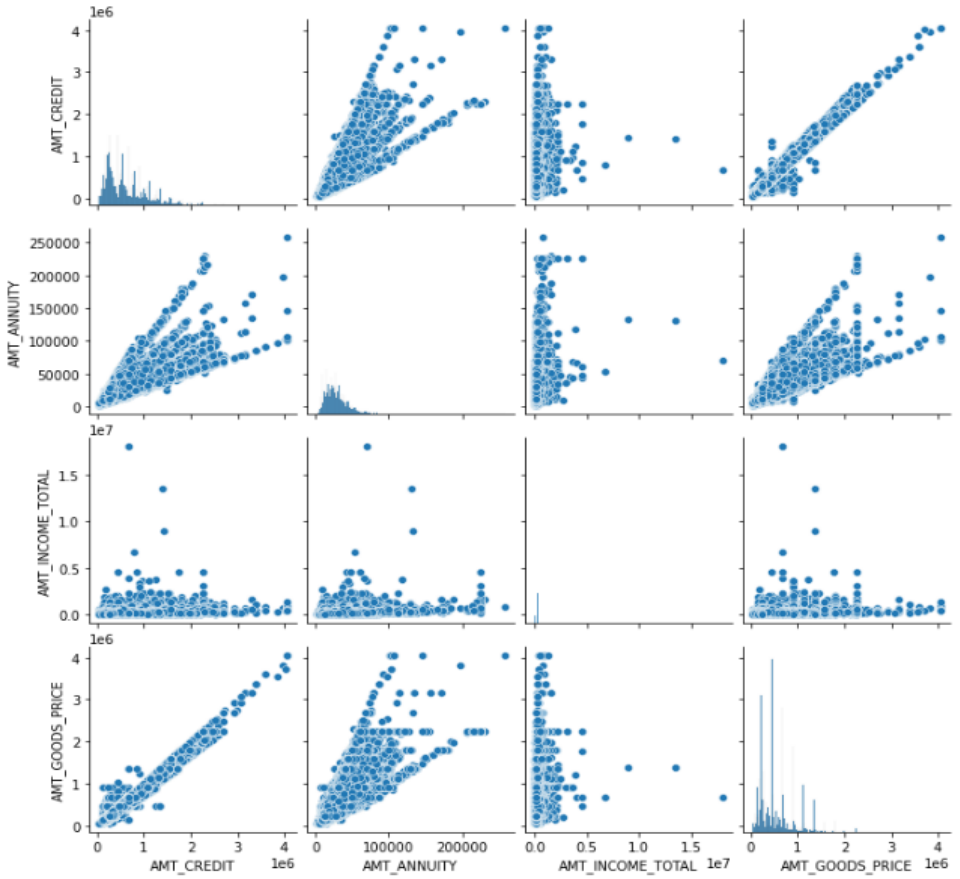
Credit amount is inversely proportional to the number of children client have, means Credit amount is higher for less children count client have and vice-versa.

AMT_GOODS_PRICE is proportional to AMT_CREDIT

Univariate, Bivariate Analysis and Multivariate Analysis - Conti

AMT_CREDIT and AMT_GOODS_PRICE are highly correlated with each other, one increases other
AMT_CREDIT and AMT_ANNUITY are also highly correlated with each other
AMT_GOODS_PRICE and AMT_ANNUITY are highly correlated

Increase in Credit amount will increase EMI amount at the same time increase in property price will increase EMI and CREDIT

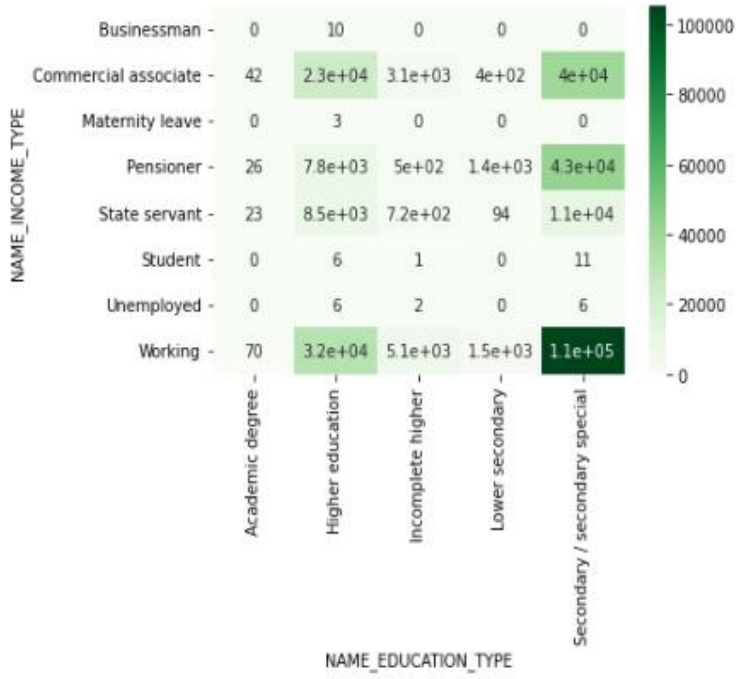


Univariate, Bivariate Analysis and Multivariate Analysis - Conti

we can decide how may application we received for each group and we should proceed with them or not?

Out[64]:

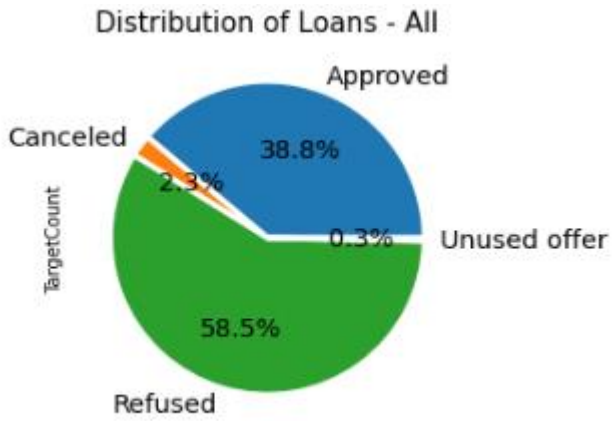
NAME_EDUCATION_TYPE	Academic degree	Higher education	Incomplete higher	Lower secondary	Secondary / secondary special
NAME_INCOME_TYPE					
Businessman	0.0	10.0	0.0	0.0	0.0
Commercial associate	42.0	22782.0	3121.0	404.0	39908.0
Maternity leave	0.0	3.0	0.0	0.0	0.0
Pensioner	26.0	7848.0	495.0	1427.0	42584.0
State servant	23.0	8519.0	718.0	94.0	11100.0
Student	0.0	6.0	1.0	0.0	11.0
Unemployed	0.0	6.0	2.0	0.0	6.0
Working	70.0	31680.0	5068.0	1474.0	105258.0

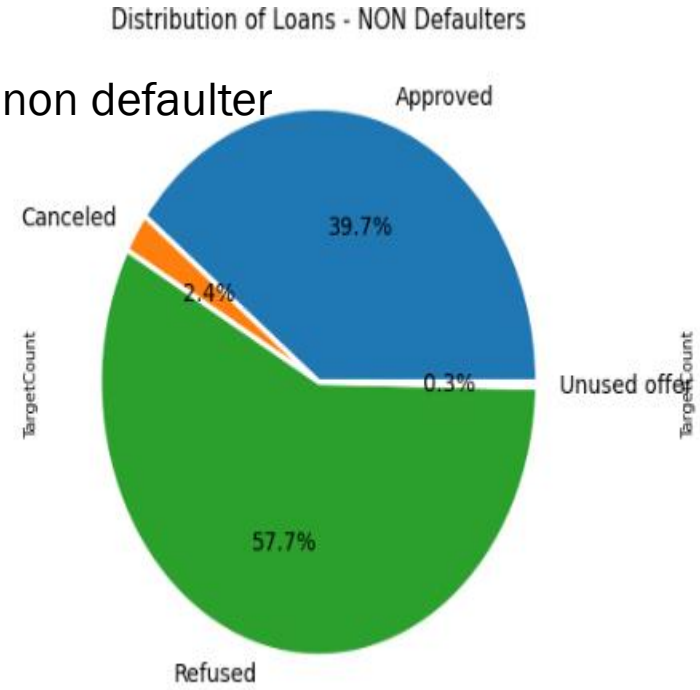


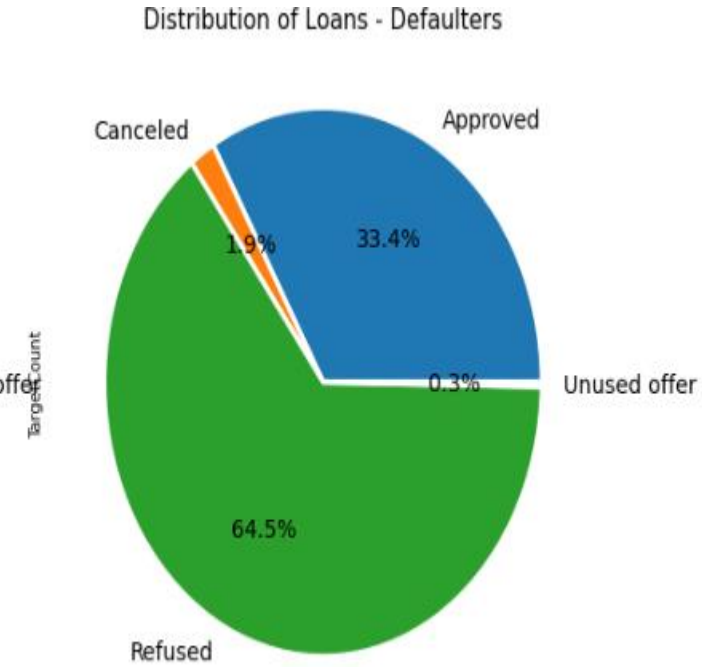
Univariate, Bivariate Analysis and Multivariate Analysis - Conti

Previously 38.8% loans are approved, 2.3% are cancelled ,
58.5% are Refused and 0.3% are Unused

39.7% of clients whose loans are approved are non defaulter and 33.4% are defaulters

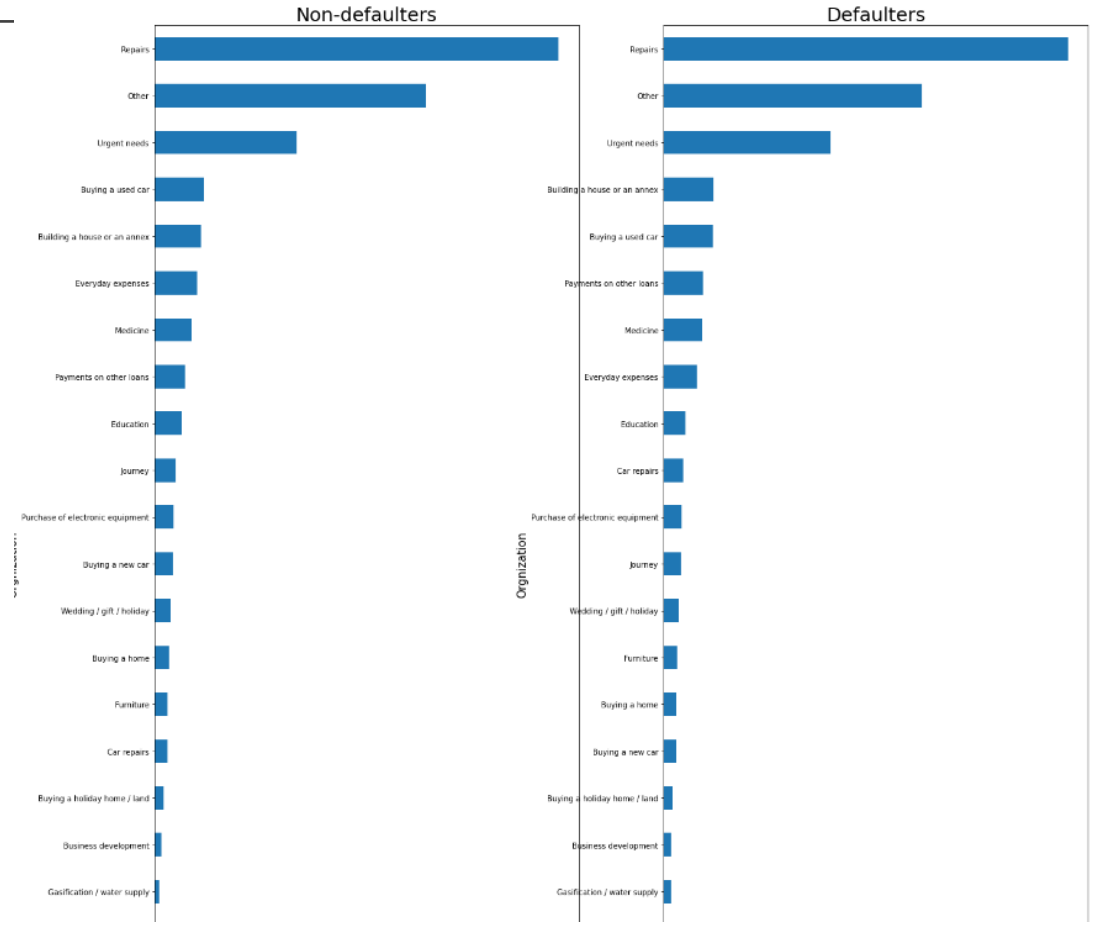






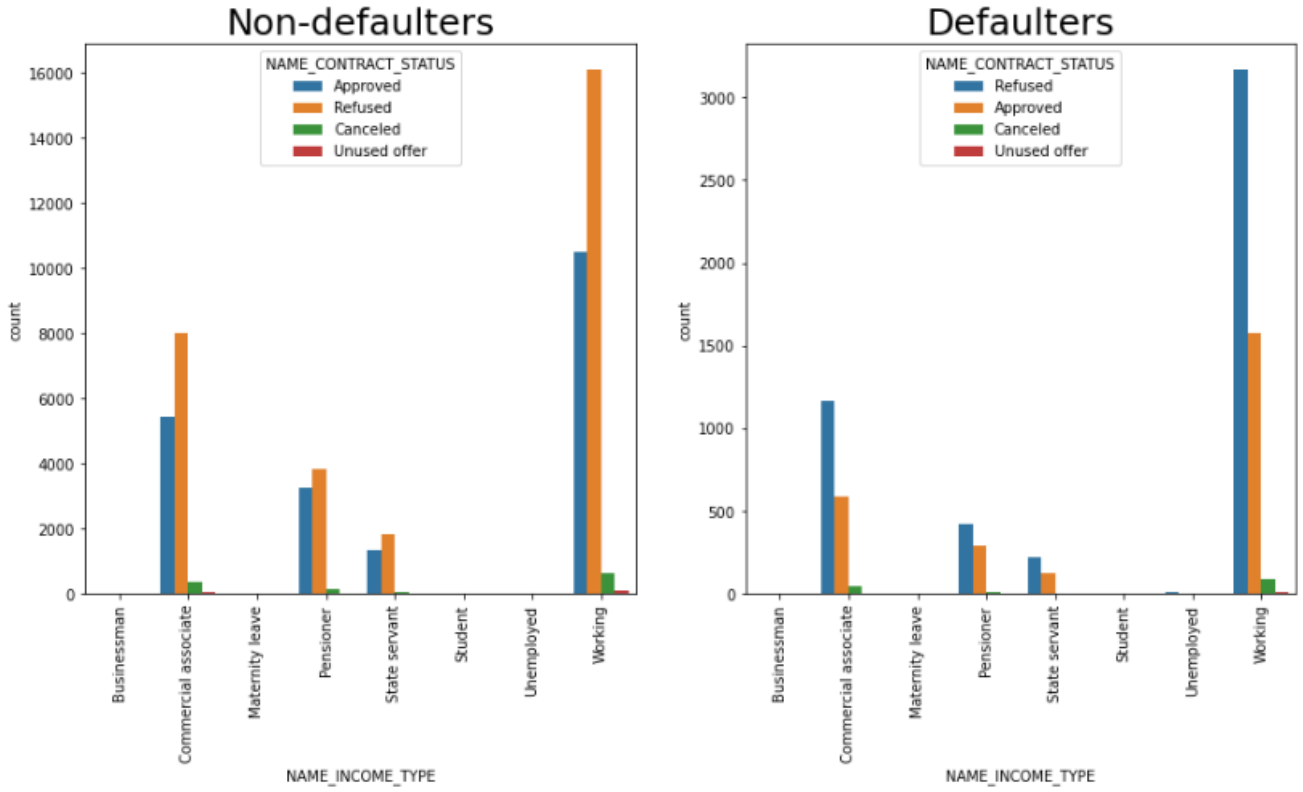
Univariate, Bivariate Analysis and Multivariate Analysis - Conti

Maximum clients who did repair work previously are defaulters



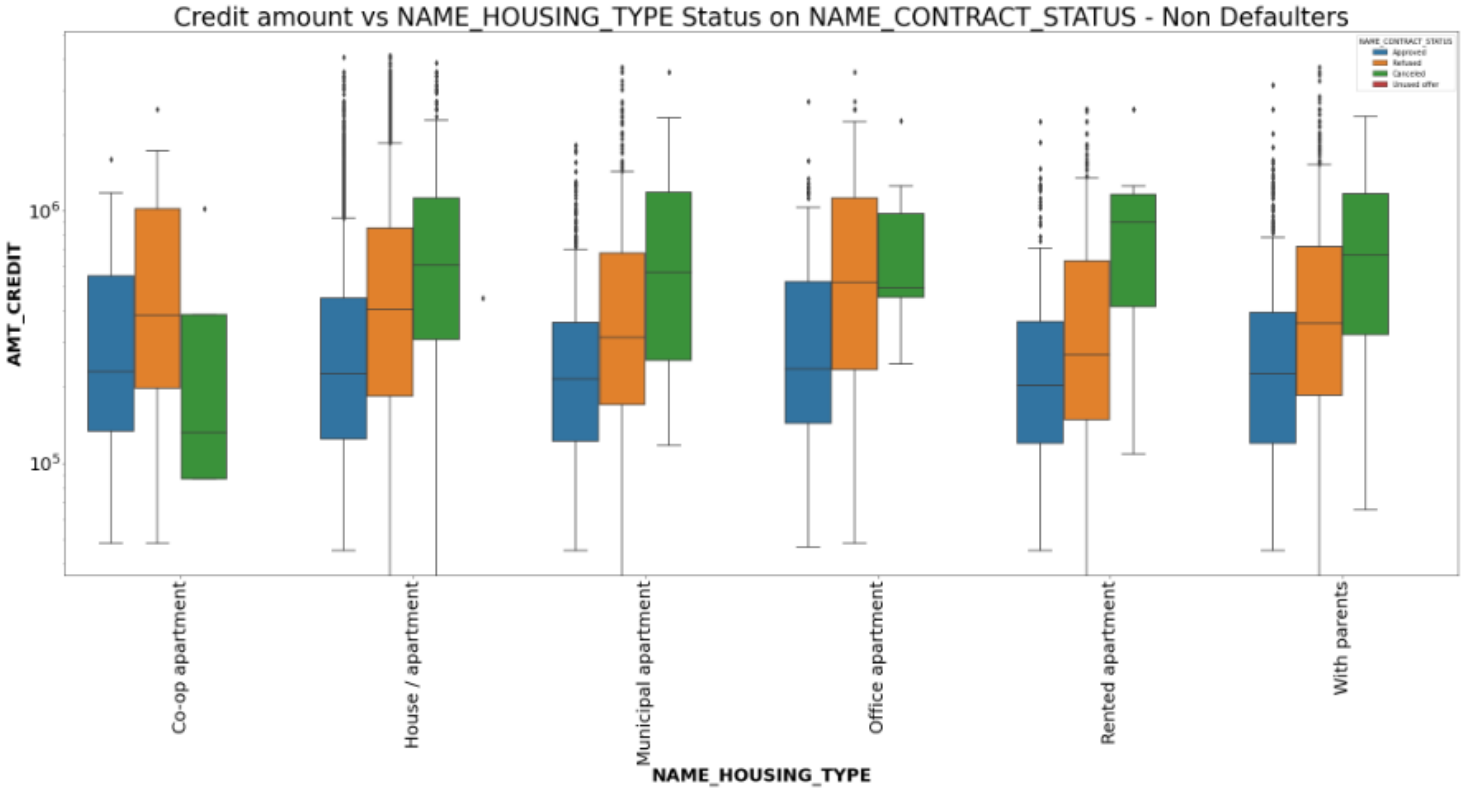
Univariate, Bivariate Analysis and Multivariate Analysis - Conti

Maximum number of application are refused for Working clients
All the working clients are very high risk clients



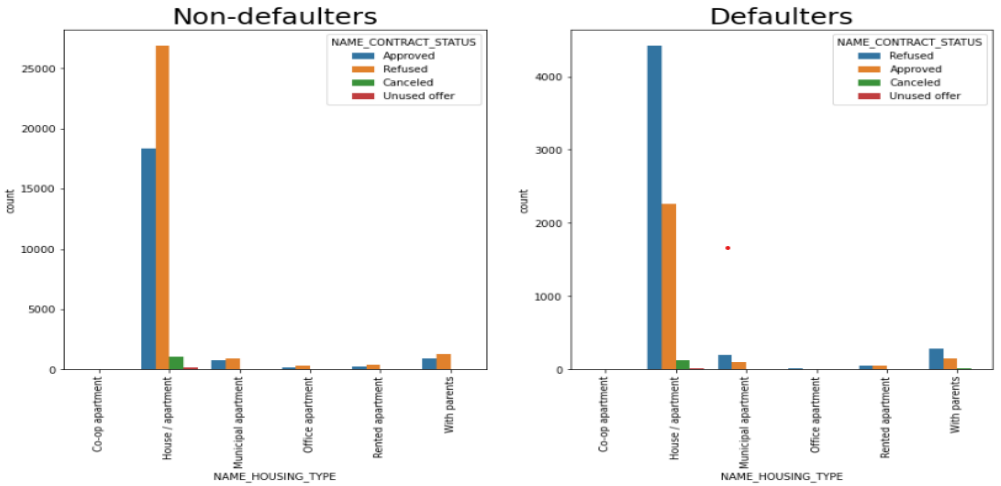
Univariate, Bivariate Analysis and Multivariate Analysis - Conti

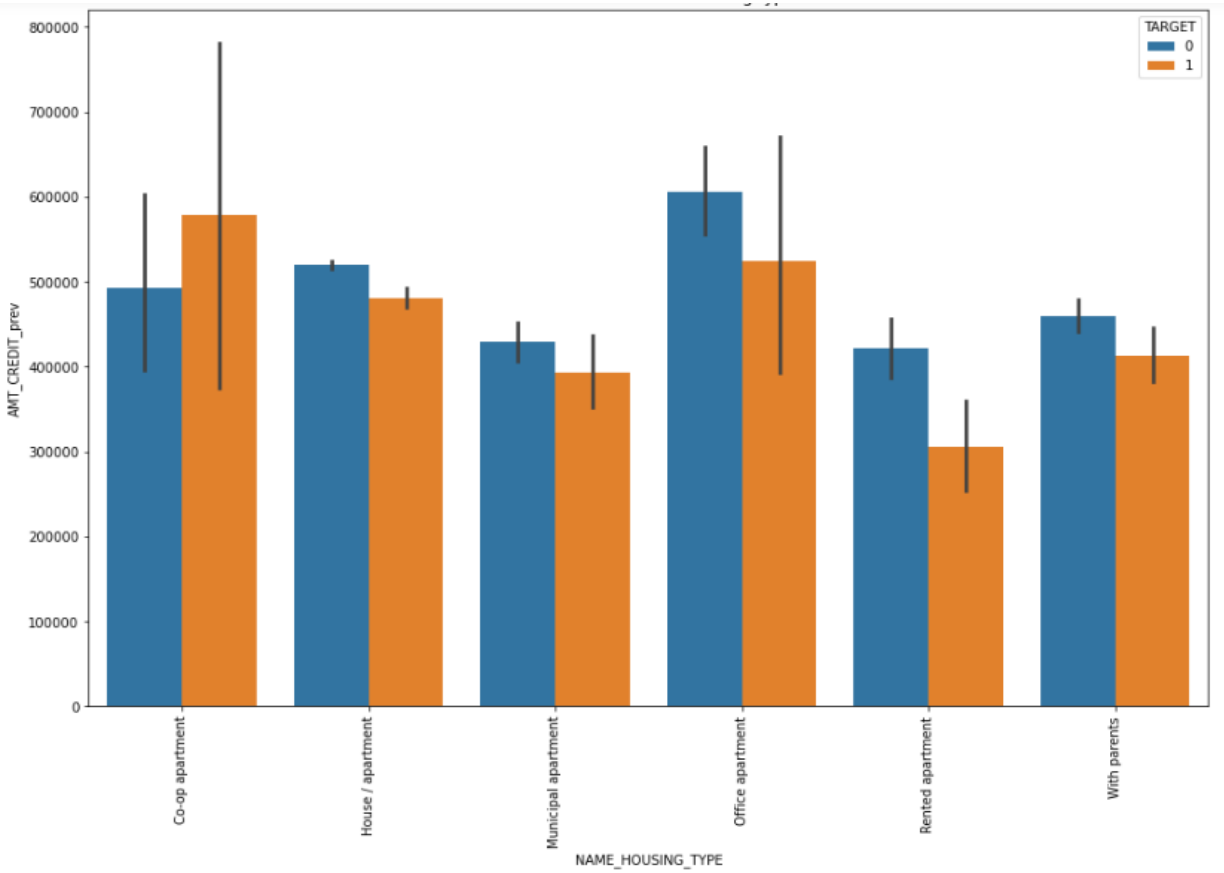
House and Municipal apartment clients has maximum number of refused and approved applicants with Outliers



Univariate, Bivariate Analysis and Multivariate Analysis - Conti

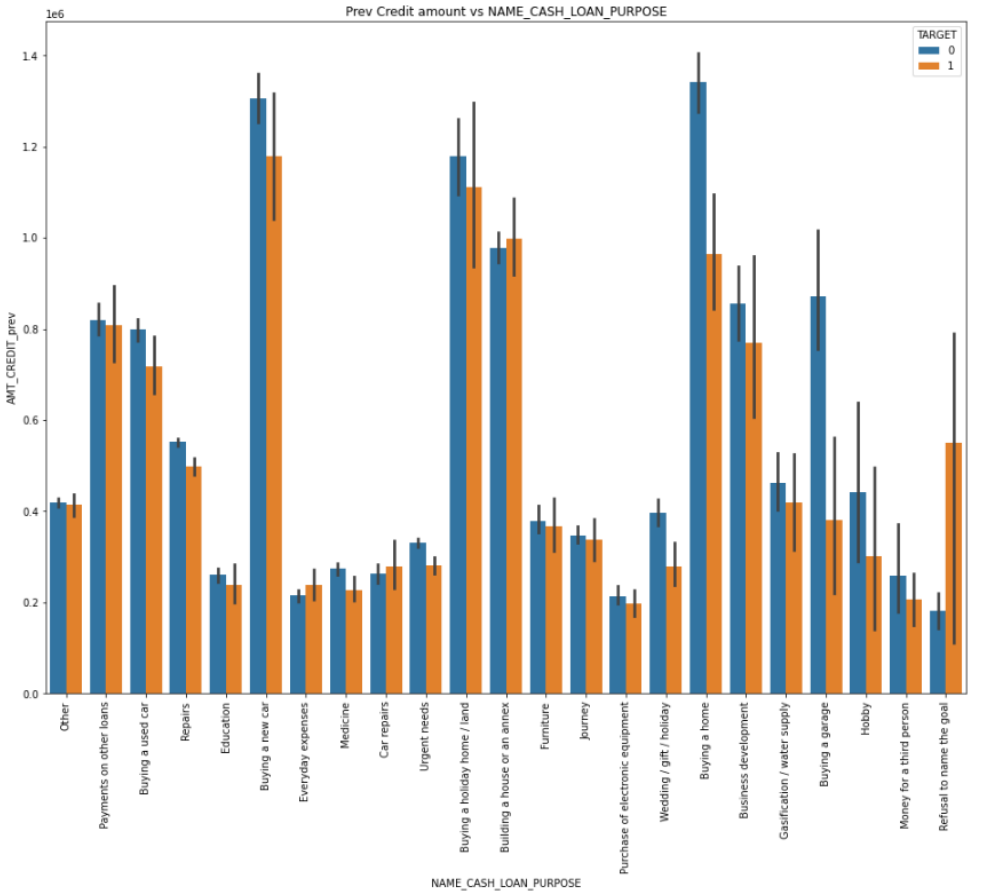
Client living in co, house/ Apartment got maximum number of application refused and they are the high risk clients
Clients living with parents and in municipal or office apartment are relatively low risk clients





Univariate, Bivariate Analysis and Multivariate Analysis - Conti

Client with building a house are the high risk clients who are mainly facing payment difficulties followed by refusal to the name goal.
Loan with the purpose of Buying a garage clients are paying EMI on time comparing to others.
Buying a new car, holiday home are also risky clients in terms of repayment.
Clients with cash purpose of Hobby are relatively less risk clients.



Conclusion

1. Bank received maximum number of applications from INCOME_TYPE - Working clients, but looking at the data we can say that, working clients holds the maximum number of records in late payments, they are high risk clients which are likely to get default.

2. Also Very low-income clients are likely to get default easily so Bank should check credit amount, Income and EMI while considering application for approval.

3. Client with single relationship are less risky to get default but married clients are applied the most with the high rate of late payments.

4. Clients with housing type "with Parents" and "Municipal apartment" are less risky so Bank should focus more on them.

5. Middle age clients having lot of paying difficulties but very young clients do not face any payment issues.

6. Loan purpose Repair is having high number of defaulters.

7. Previous credit mainly raised for repairs, buying new car, holiday home and house with high number of defaulters.

8. Previously 38.8% loans are approved, 2.3% are cancelled, 58.5% are Refused and 0.3% are Unused. Maximum number of application are refused for Working clients.

9. Bank can focus more on Pensioner, Students, State Servant and Businessman clients. Can asked high ROI or EMI for clients with middle or high risk and offer low ROI for the low risk, well educated, non-defaulters , young and middle to high income clients.