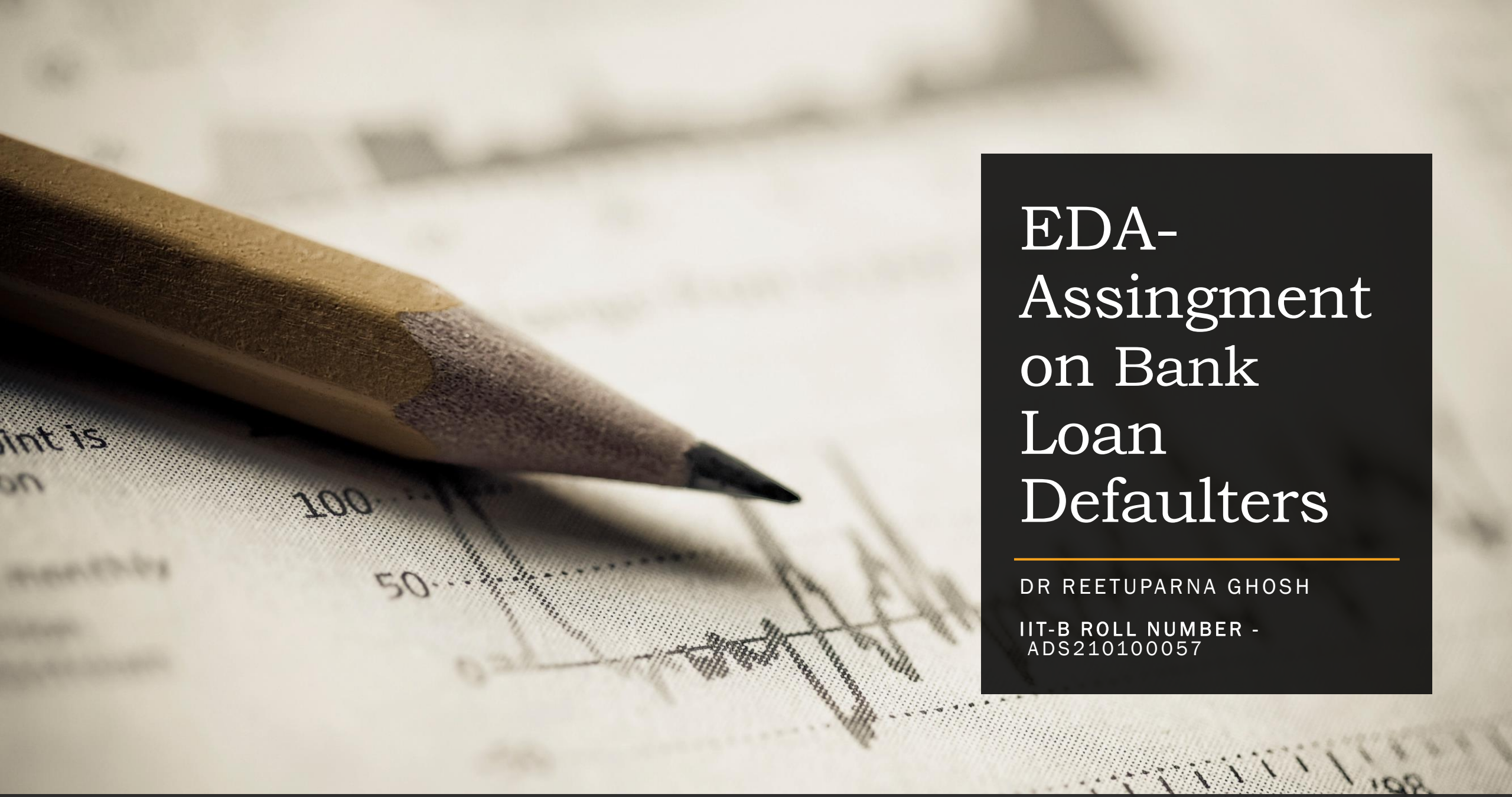


EDA-Assingment on Bank Loan Defaulters

DR REETUPARNA GHOSH

IIT-B ROLL NUMBER - ADS210100057

Purpose

1. To identify possible defaulters or clients facing payment issue along with possible parameters leading to becoming defaulters.
2. To identify potential customers for approving loan

Available Data set

Two data sets are provided, which are:

1. *'application_data.csv'* → contains all the information of the client at the time of application.
2. *'previous_application.csv'* → contains information about the client's previous loan data.

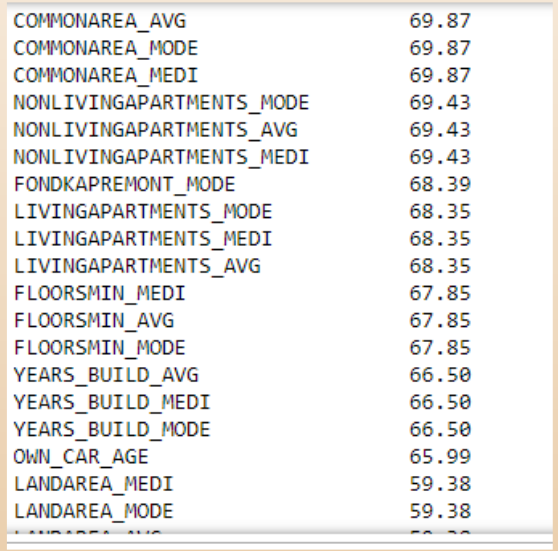
application_data.csv	
Rows	307511
Columns	122
Float-dtypes	65
Integer-dtypes	41
String dtypes	16

previous_application.csv	
Rows	1670214
Columns	37
Float-dtypes	15
Integer-dtypes	6
String dtypes	16

DATA QUALITY CHECK

Checking and treating columns with missing data in Application Data set

1. 64 columns have missing data
2. Of which 49 columns have missing data> 40 %→Dropping
3. Imputing values for categorical columns such as “OCCUPATION_TYPE”, “'NAME_TYPE_SUITE” with mode() function.
4. Dropping unwanted columns after analysis.
5. Remaining columns for analysis→24



COMMONAREA_AVG	69.87
COMMONAREA_MODE	69.87
COMMONAREA_MEDI	69.87
NONLIVINGAPARTMENTS_MODE	69.43
NONLIVINGAPARTMENTS_AVG	69.43
NONLIVINGAPARTMENTS_MEDI	69.43
FONDKAPREMONT_MODE	68.39
LIVINGAPARTMENTS_MODE	68.35
LIVINGAPARTMENTS_MEDI	68.35
LIVINGAPARTMENTS_AVG	68.35
FLOORSMIN_MEDI	67.85
FLOORSMIN_AVG	67.85
FLOORSMIN_MODE	67.85
YEARS_BUILD_AVG	66.50
YEARS_BUILD_MEDI	66.50
YEARS_BUILD_MODE	66.50
OWN_CAR_AGE	65.99
LANDAREA_MEDI	59.38
LANDAREA_MODE	59.38

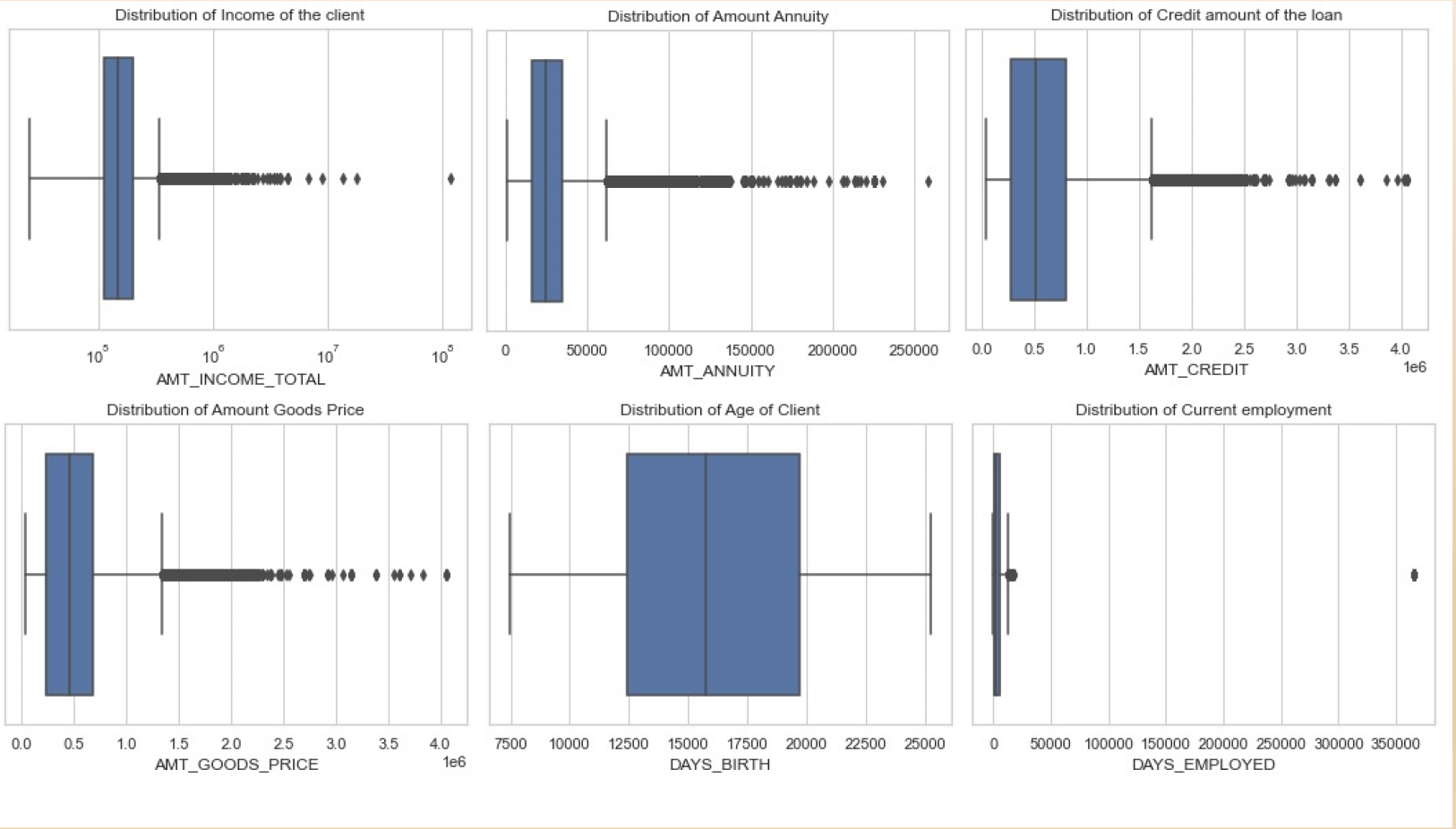
Checking and treating columns with missing data in Previous_application Data set

1. 15 columns have missing data
2. Of which 49 columns have missing data> 45 %→Dropping
3. Imputing values for numerical columns “AMT_GOODS_PRICE” with mean() function.
4. Dropping unwanted columns after analysis.
5. Remaining columns for analysis→22

Binning of Continuous Variables

- DAYS_BIRTH– age column of client
- DAYS_EMPLOYED – duration of employment of client
- AMT_INCOME – income of client
- AMT_CREDIT – loan amount credited to client

Checking outliers in the data

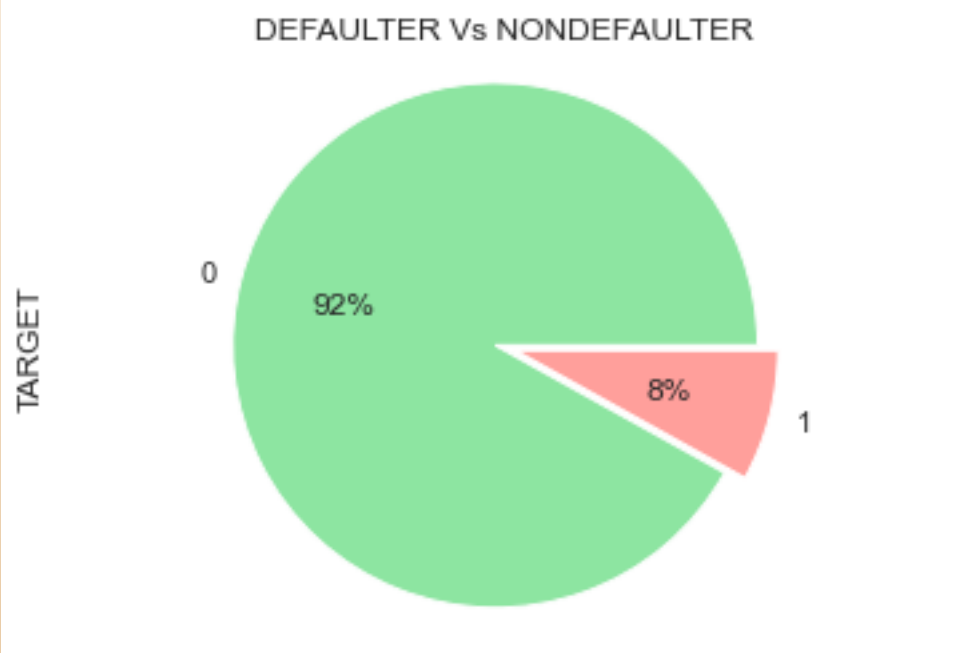


1. **AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE, and AMT_INCOME_TOTAL have high number of outliers** which indicating few loan applicants have high income in comparison the others.

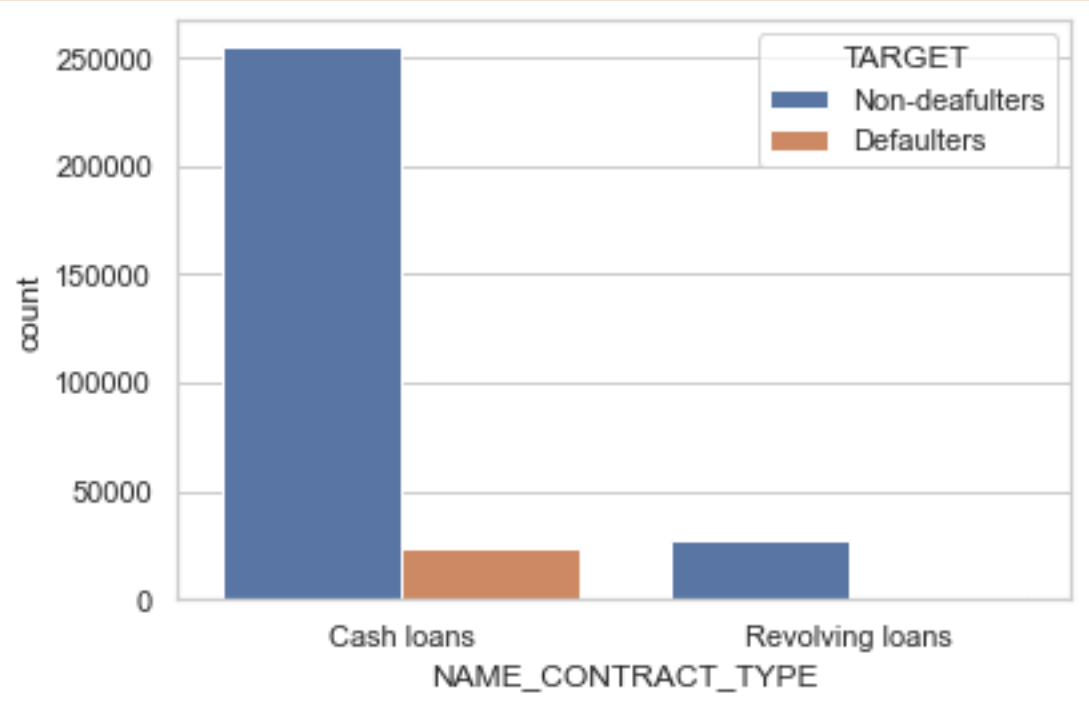
3. **DAYS_BIRTH has no outliers** meaning available **data is reliable**.

4. **DAYS_EMPLOYED has outlier** value above 350000(days) which is **impossible** and hence this must to be **incorrect entry.**

Checking Imbalance



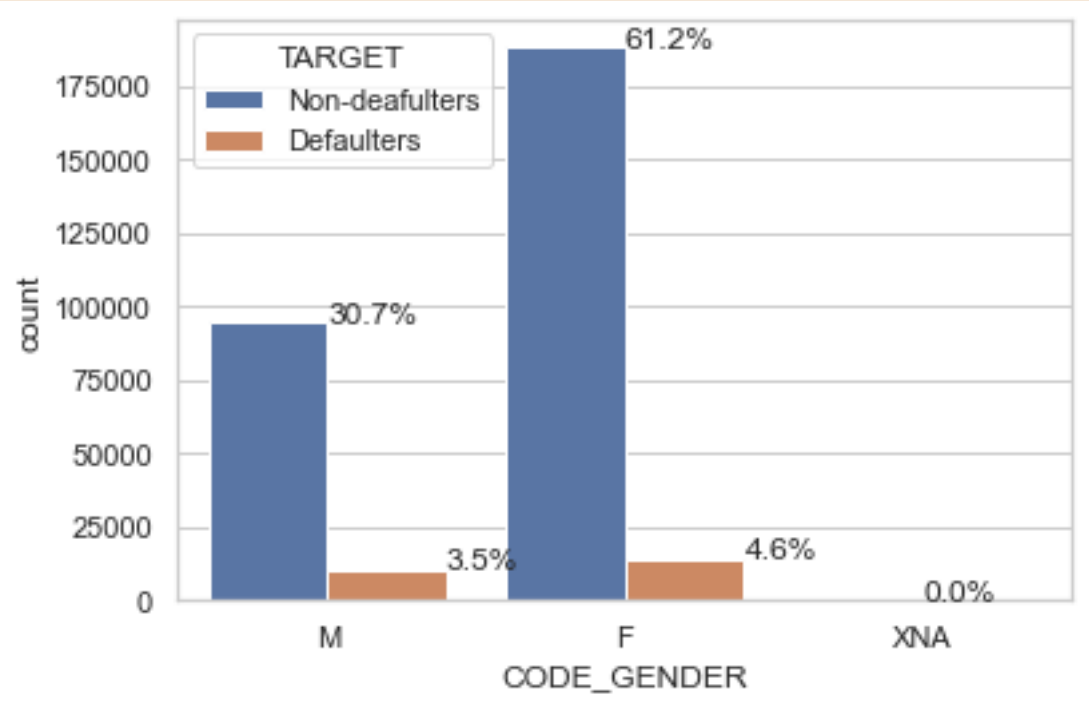
- 92 % loan applicants are non-defaulters
- 8 % loan applicants are defaulters



- Cash loans have defaulters
- Revolving loans do not have defaulters

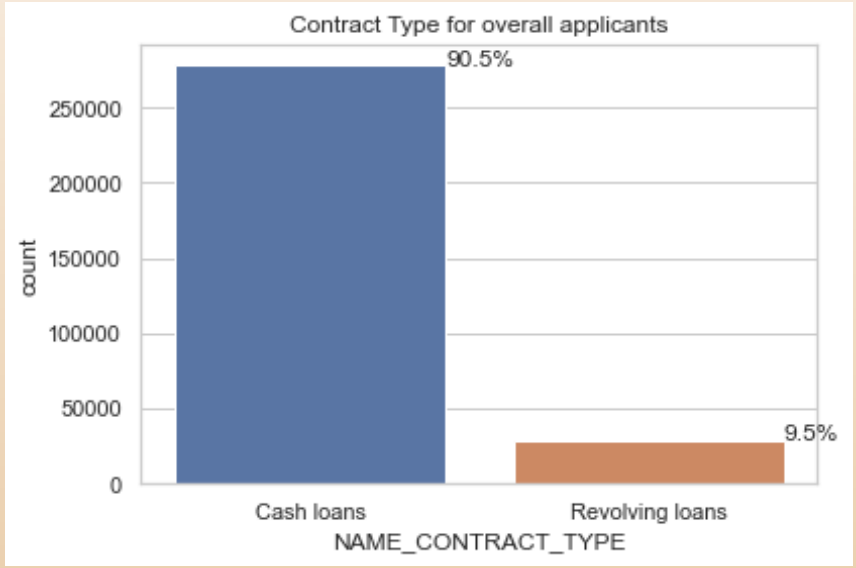
UNIVARIATE ANALYSIS- Categorical Data

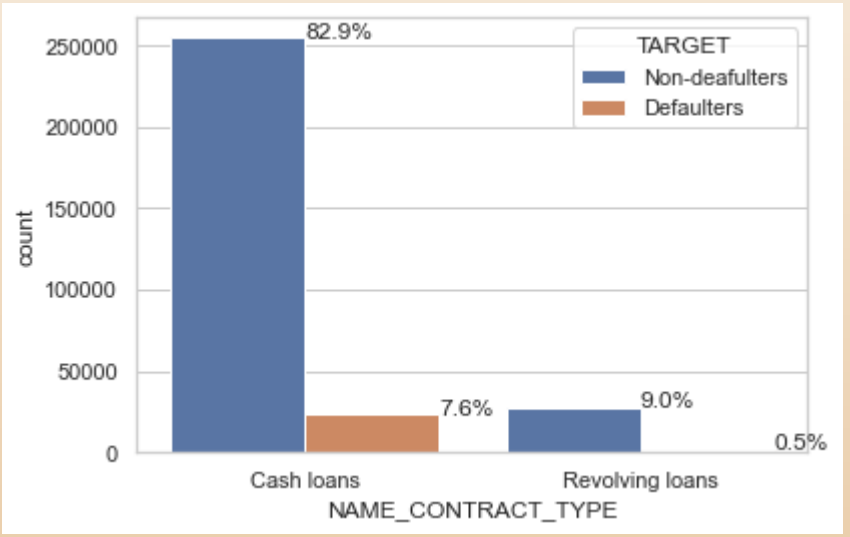
Distribution based on Gender



1.Clearly Females more likely to apply for loans.
2. Female are most likely to be defaulters

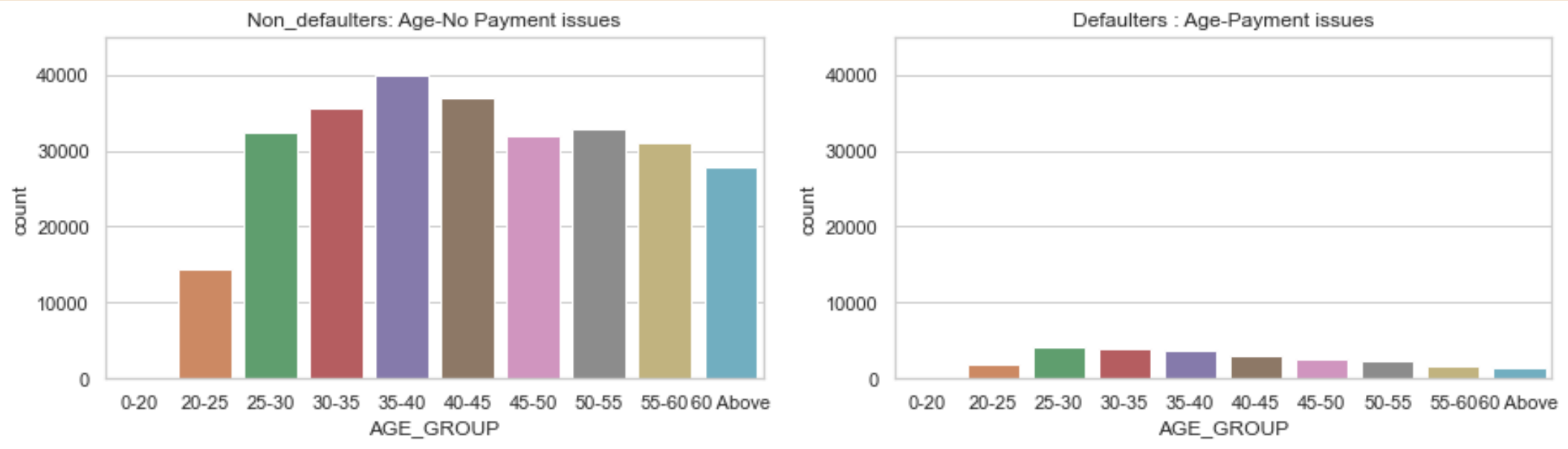
Product Distribution





1.Bank offers two kinds of loans, viz., Cash loans and Revolving loans.
2. Cash loans have much higher % than Revolving loans in general
3. Cash loans have defaulters
4. Revolving loans have no defaulters

Distribution based on Age

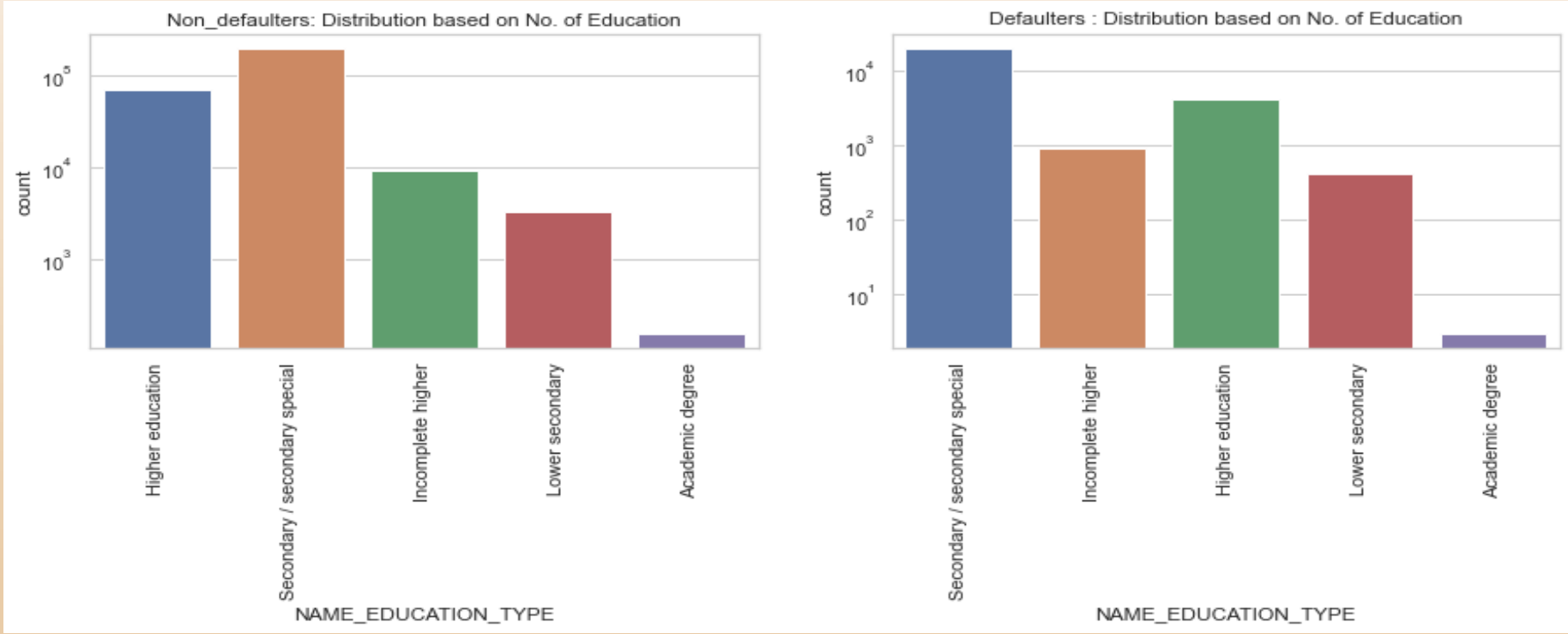


1.Customers from age 35-45 most likely to make payment.

2.Customers in age 25-30 are most likely to be defaulters

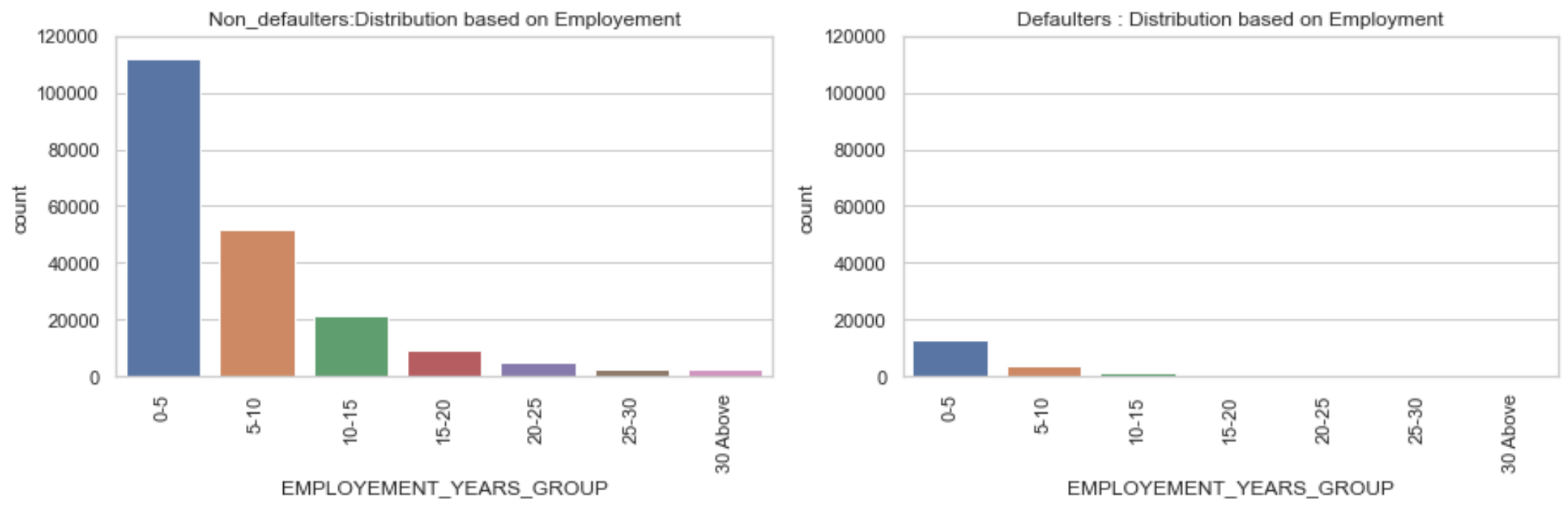
3. Defaulters number reduces from 45 years

Distribution based on Education



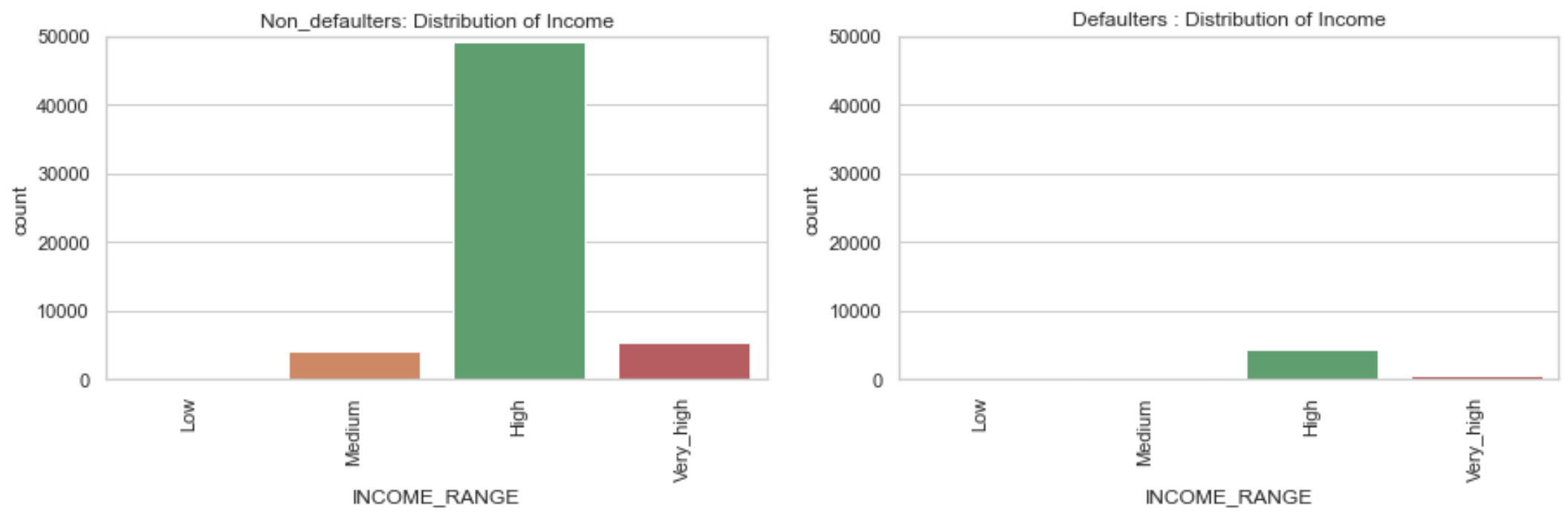
People with Secondary/secondary special are more likely to be defaulters.

Distribution based on Employment-Years



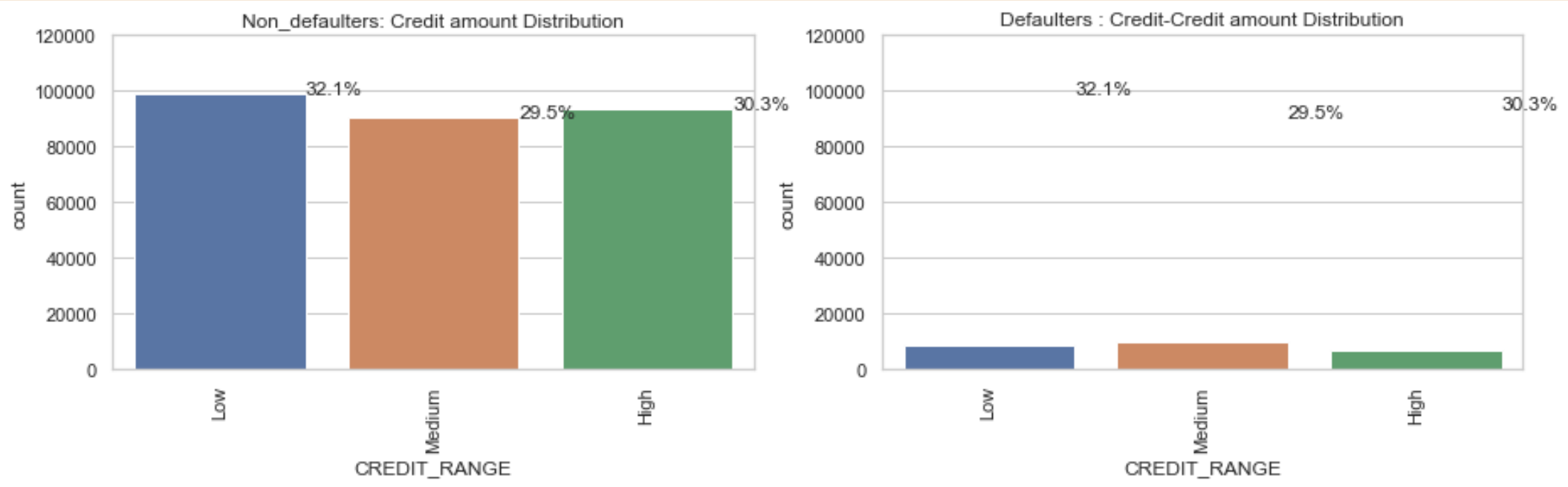
1. **People employed from 0-5 years take more loans**
2. **People employed for 0-5 years are more likely to be defaulters**

Distribution based on Income Range



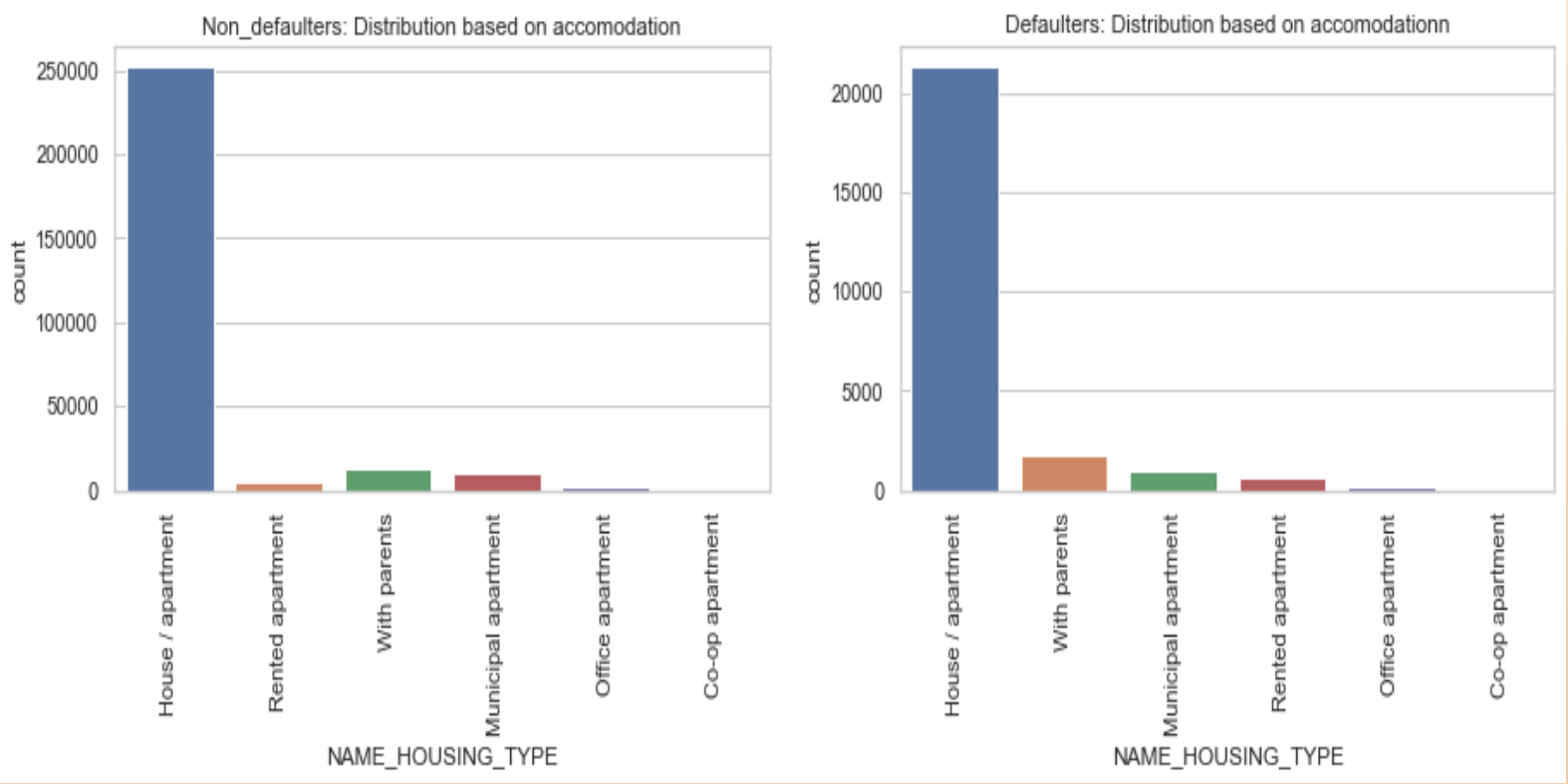
Clearly defaulters have significantly lesser income than non-defaulters.

Distribution based on Credit Amount



1.Low credit amount have high distribution in non-defaulters, followed by High credit amount and then Medium.
2. While clients with low credit amount showed more susceptibility towards defaulting.

Distribution based on Housing Type

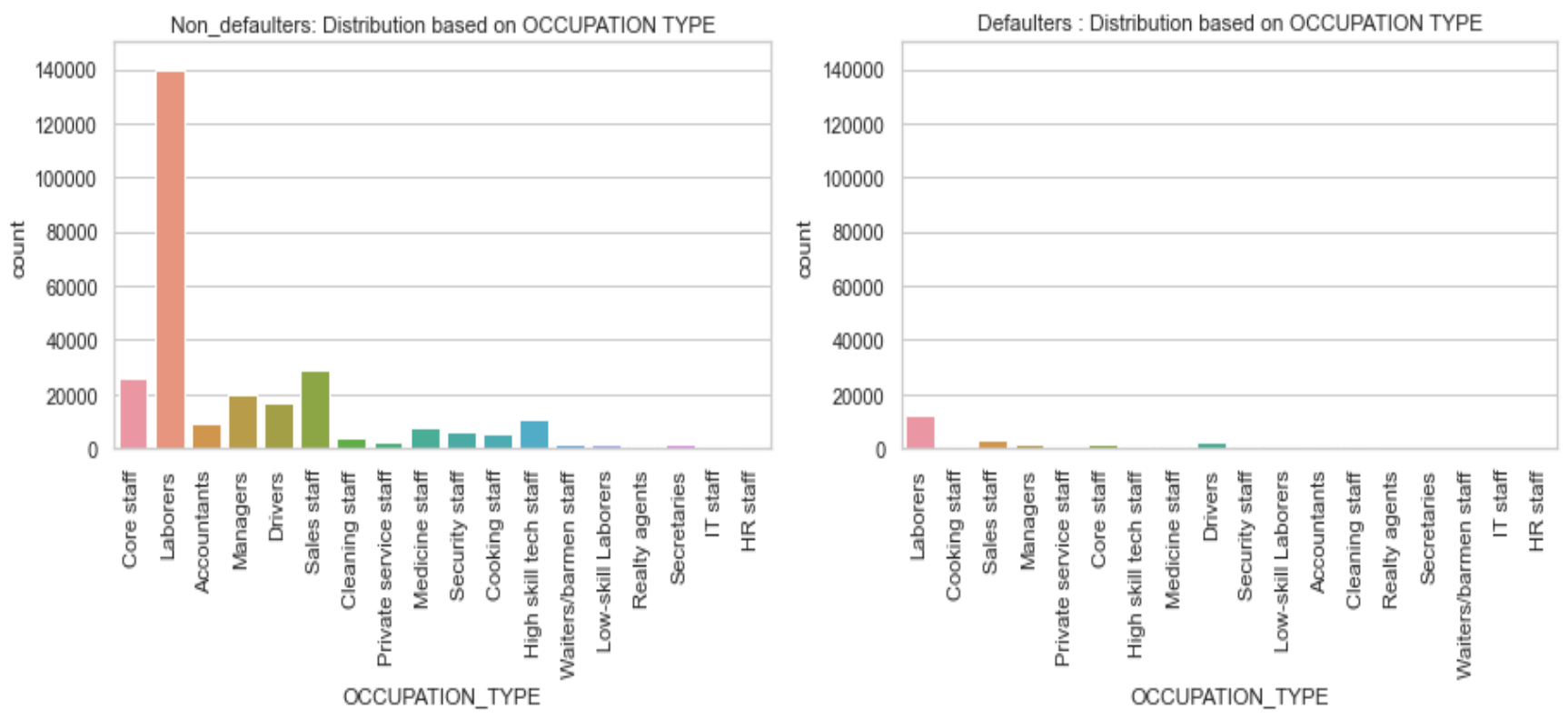


1.People with own apartments are given max loan, making them more prone to defaulter status

2.People in rented apartment are not likely to be defaulters

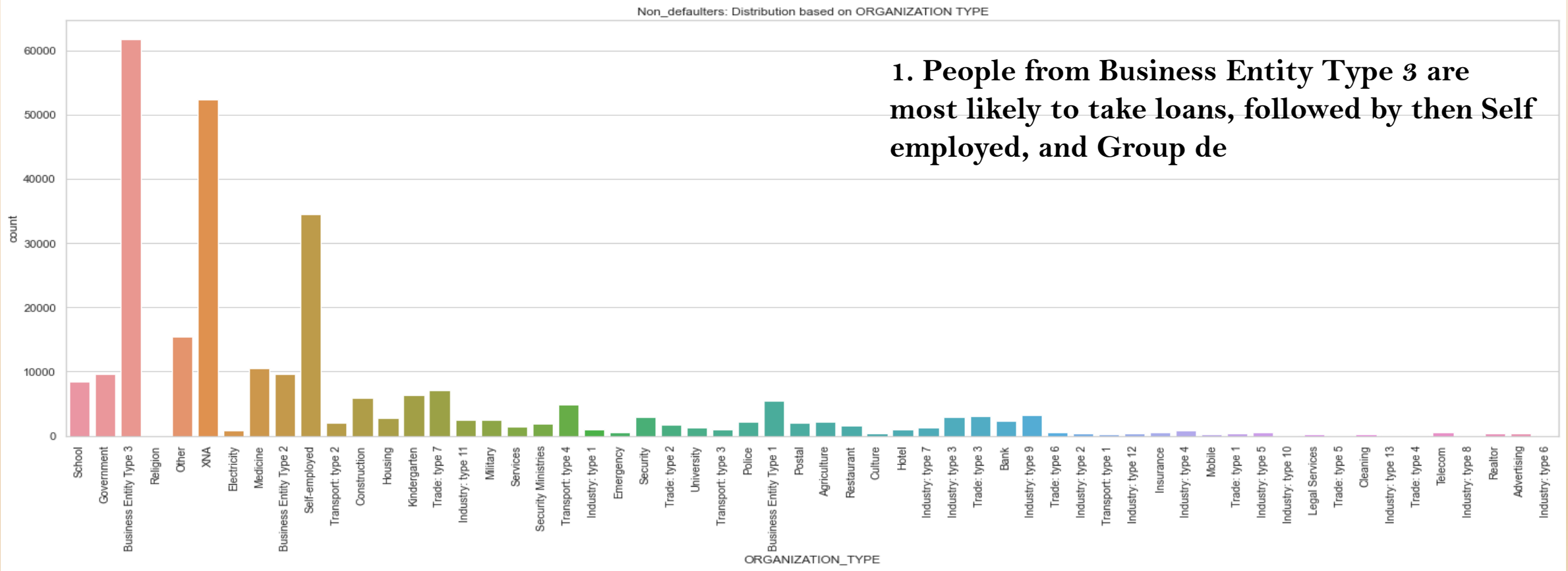
3.People living with parents have more chances of being defaulters

Distribution based on Occupation Type



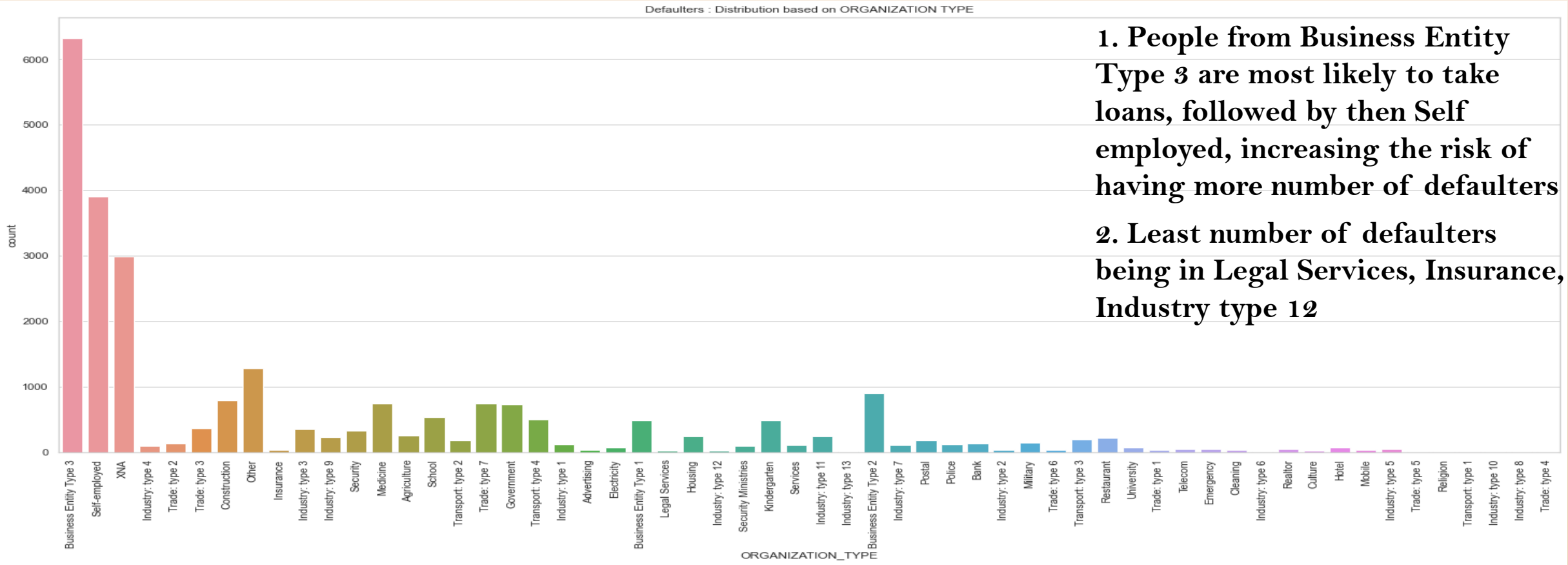
1. Most of the loans are taken by Laborers, followed by Sales staff and Core staff.
2. IT staff, HR staff and Realty staff are less likely to apply for Loan.
3. Category with highest percent of defaulters are laborer's followed by Drivers and Sales staff, Managers and Core staff.

Distribution based on Organization Type –Non-defaulters



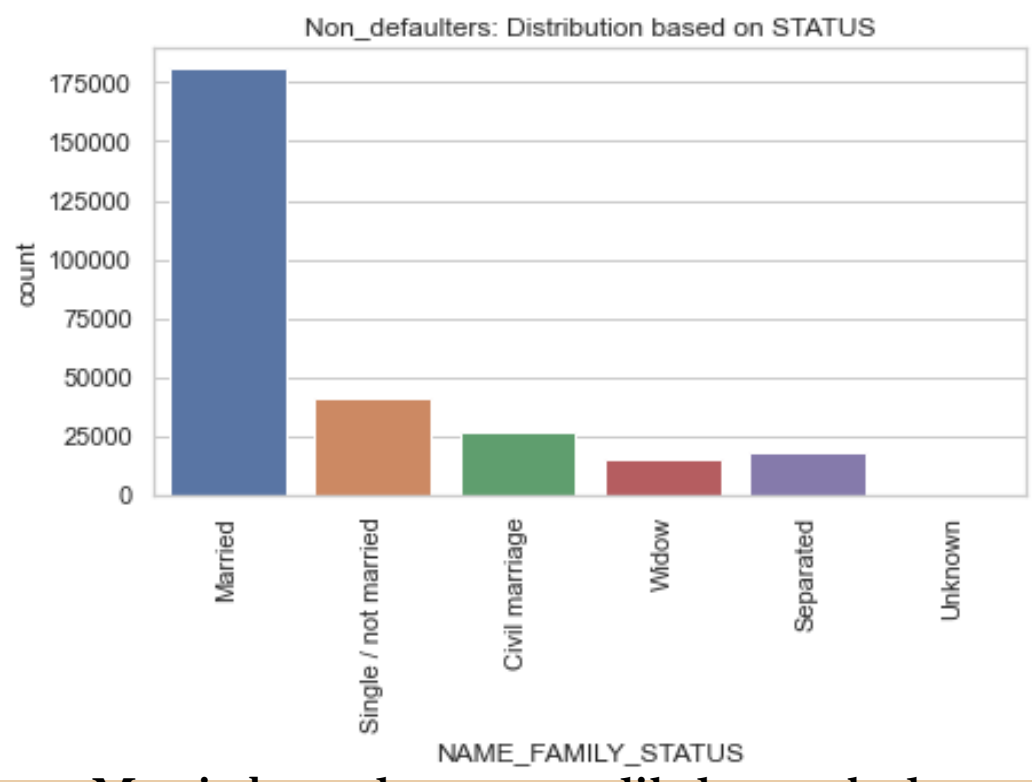
1. People from Business Entity Type *3* are most likely to take loans, followed by then Self employed, and Group de

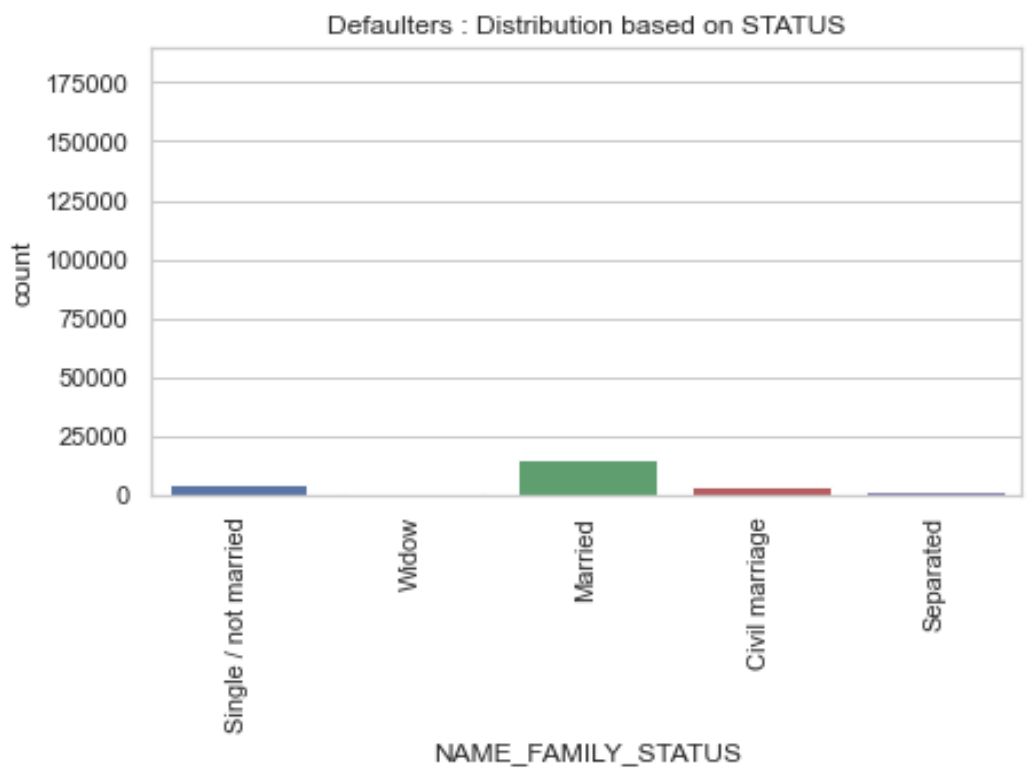
Distribution based on Organization Type –Defaulters



1. People from Business Entity Type 3 are most likely to take loans, followed by then Self employed, increasing the risk of having more number of defaulters
2. Least number of defaulters being in Legal Services, Insurance, Industry type 12

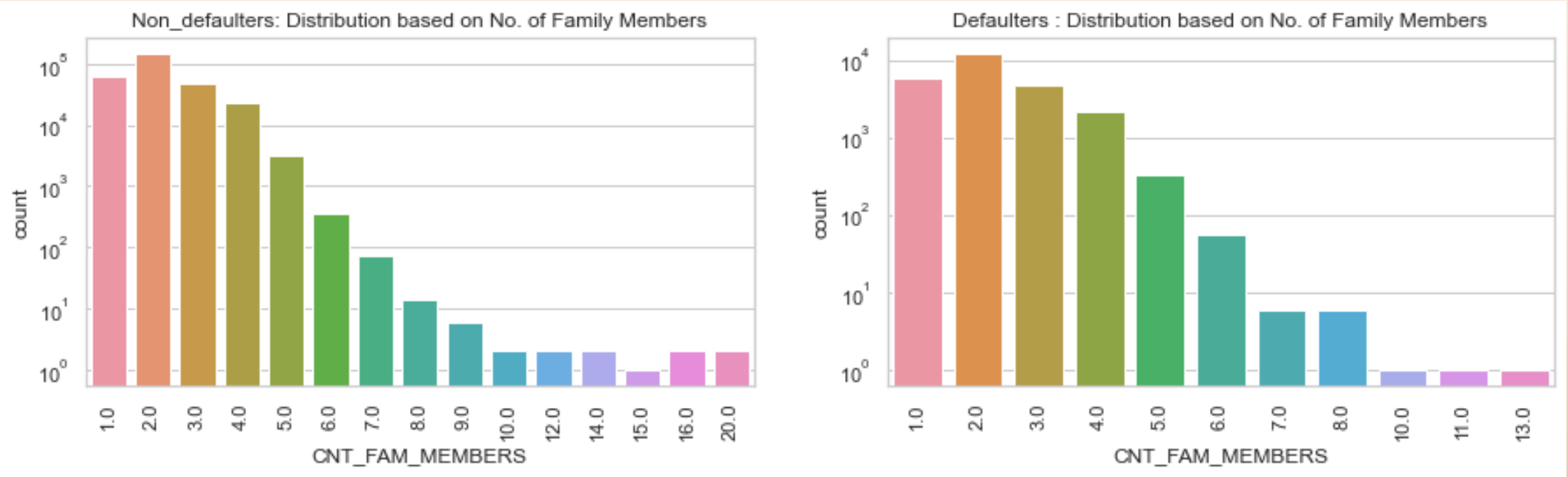
Distribution based on Status





1. **Married people are most likely to take loans**
2. **Married people are more likely to be defaulters followed by Single, Civil Marriage and separated**
3. **Windows are most reliable to give loans**

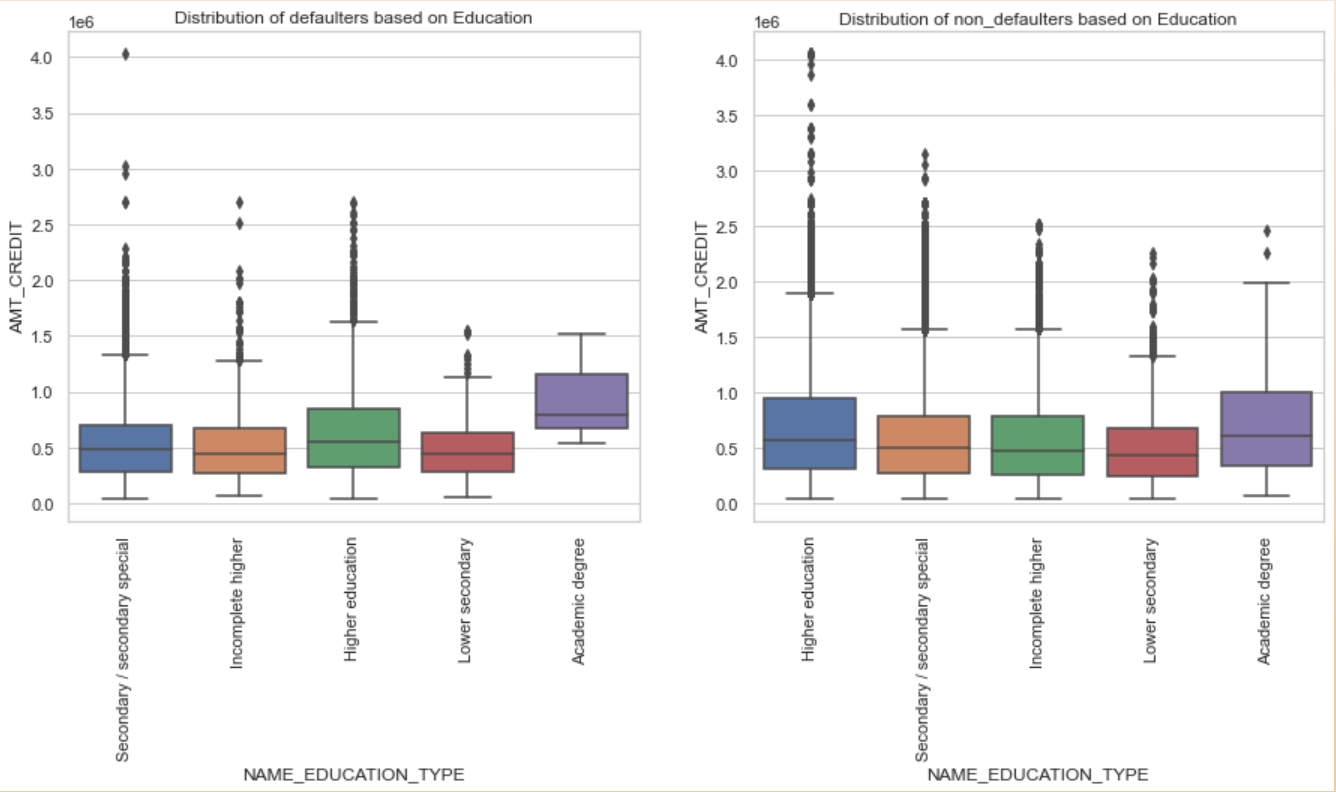
Distribution based on Family members



Number of family members has no impact on defaulters and non-defaulters

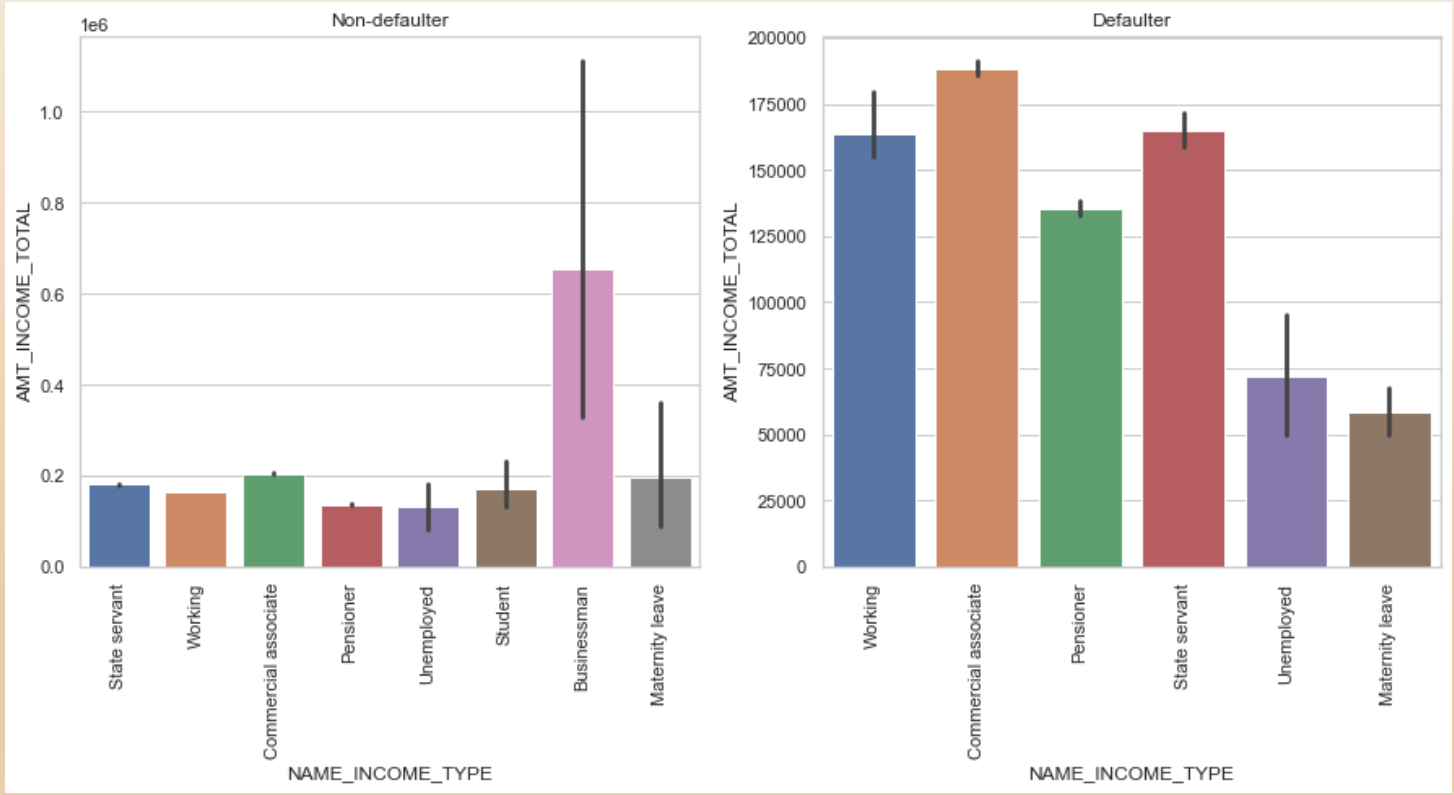
BIVARIATE ANALYSIS-Categorical Data

Distribution of Loan amount vs Education



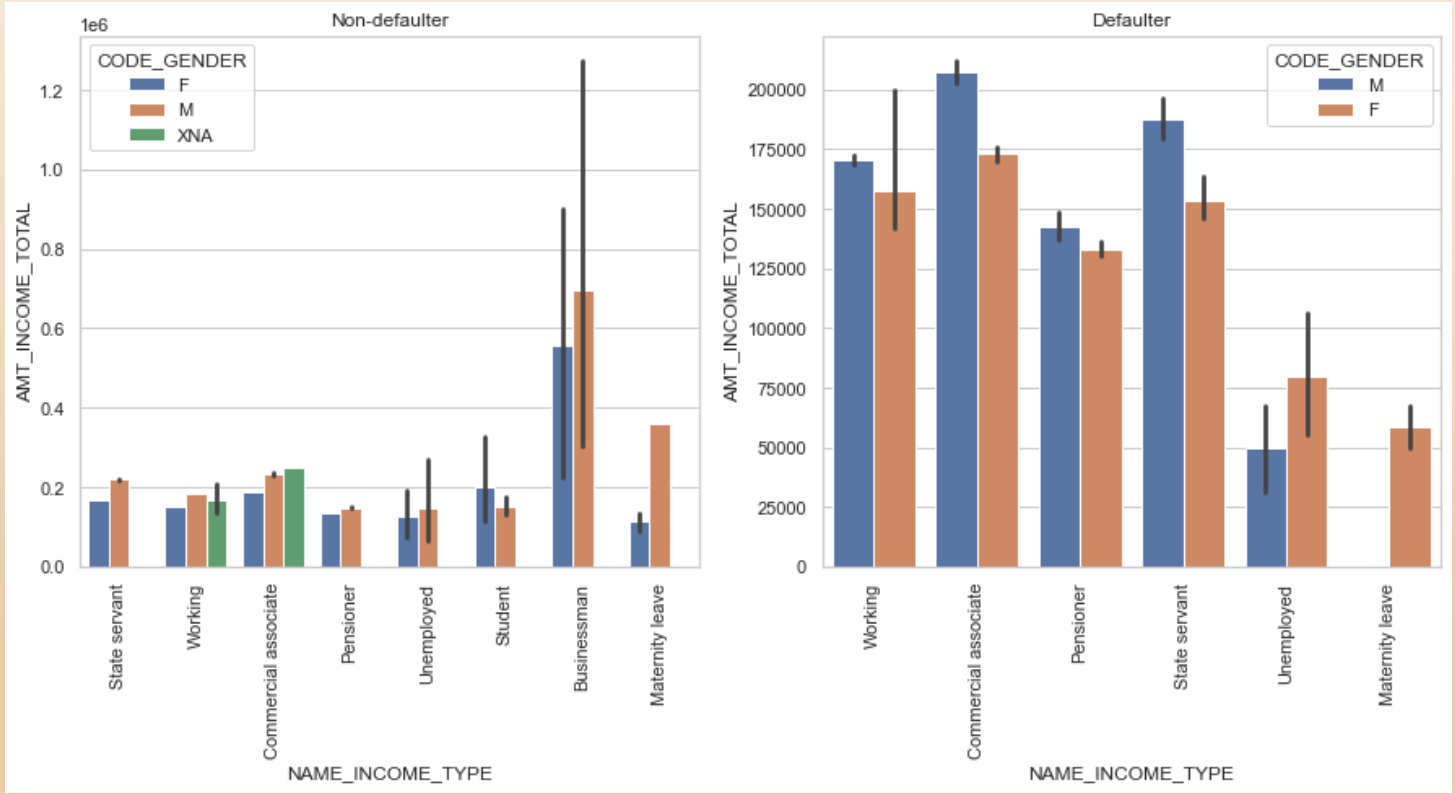
1. People with Academic degree and highest degree have higher loan credit amount

Distribution of Income amount vs Type of Income



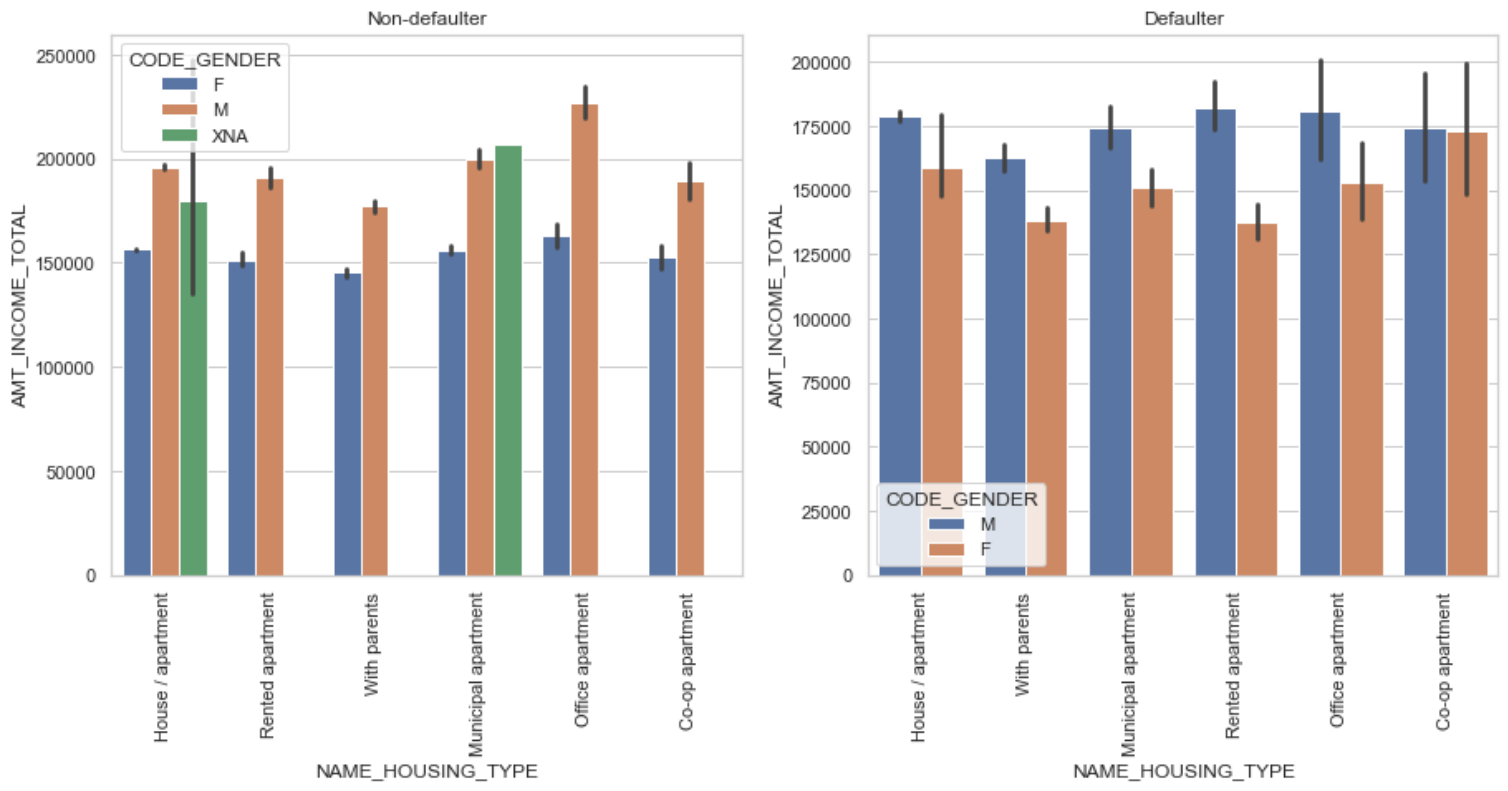
1. **It can be seen Businessman have the highest income and the estimated range seem to indicate that the income of a Businessman could be in the range of slightly close to 6 lakhs and slightly above 10 lakhs and are non defaulters.**

Distribution based on Income Type Vs Income on grounds of Gender



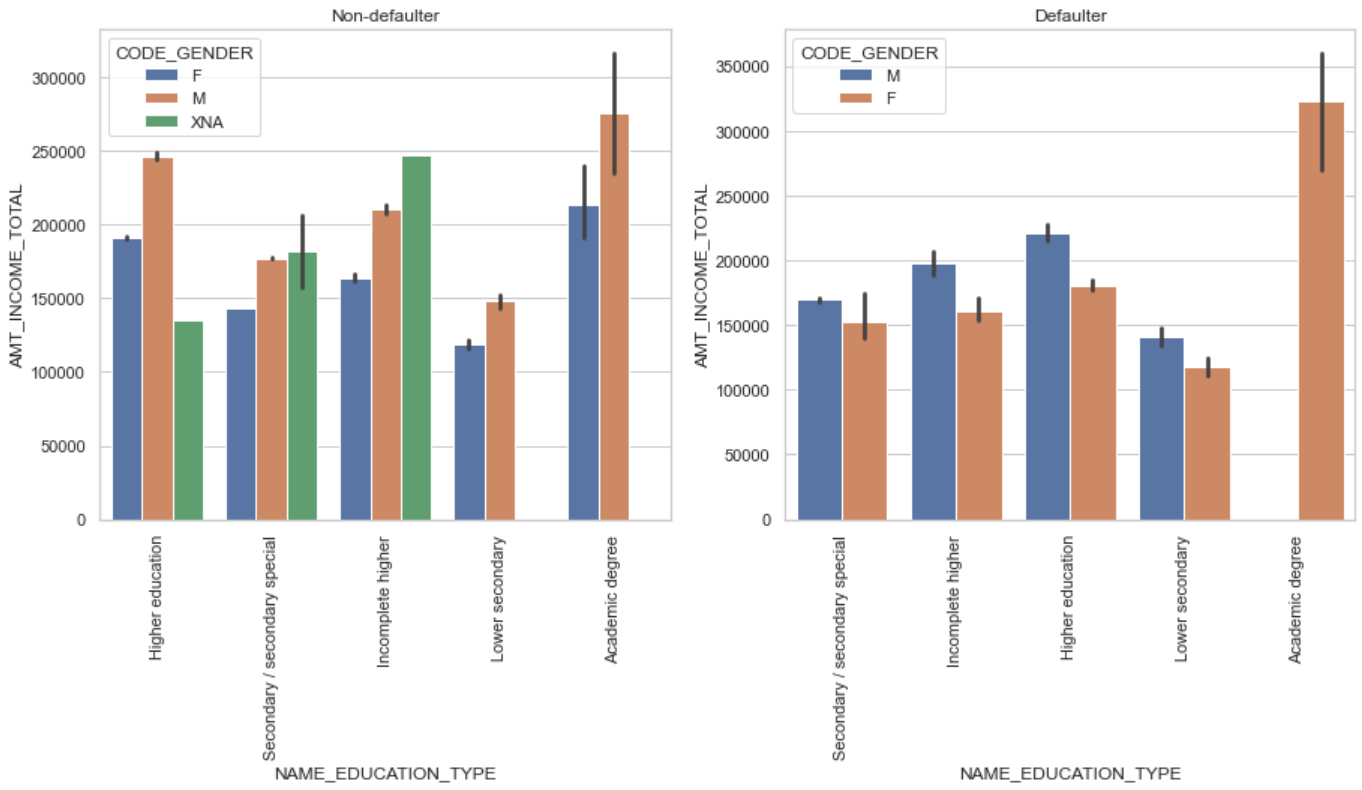
1. **Male & female under the Businessman & student category are non-defaulters**
2. **Males from Commercial associates are most likely to be defaulters, followed by State servant and Working class**
3. **Unemployed females are most likely to be defaulters**
4. **Females on maternity leave are more likely to be defaulters**

Distribution based on Housing Type Vs Income on grounds of Gender



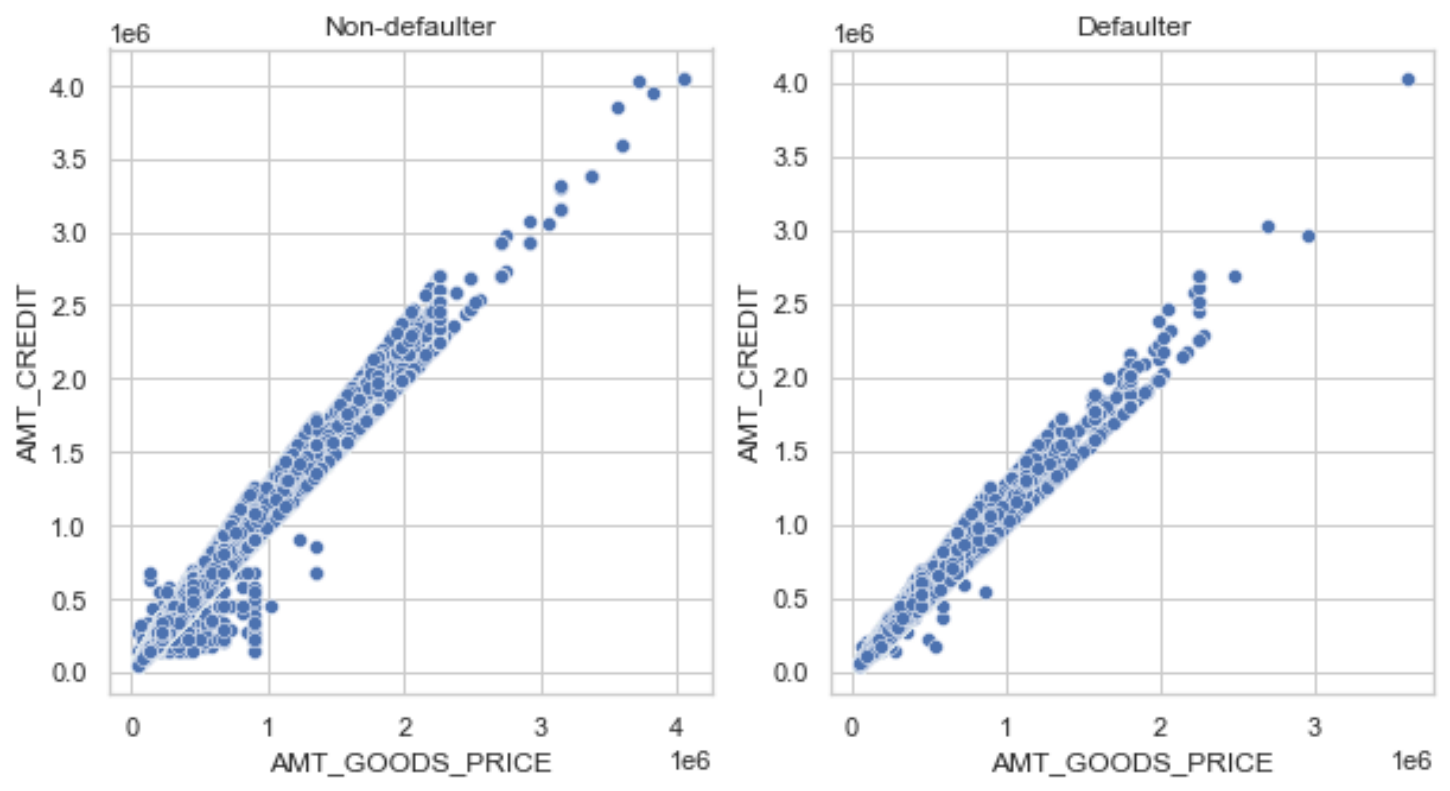
1. **Females** with **office apartment** are most likely to be **non-defaulters.**
2. **Male** owning **rented apartment, office apartment, own apartment** are most likely to be **defaulters**

Distribution based on Education Type Vs Income on grounds of Gender



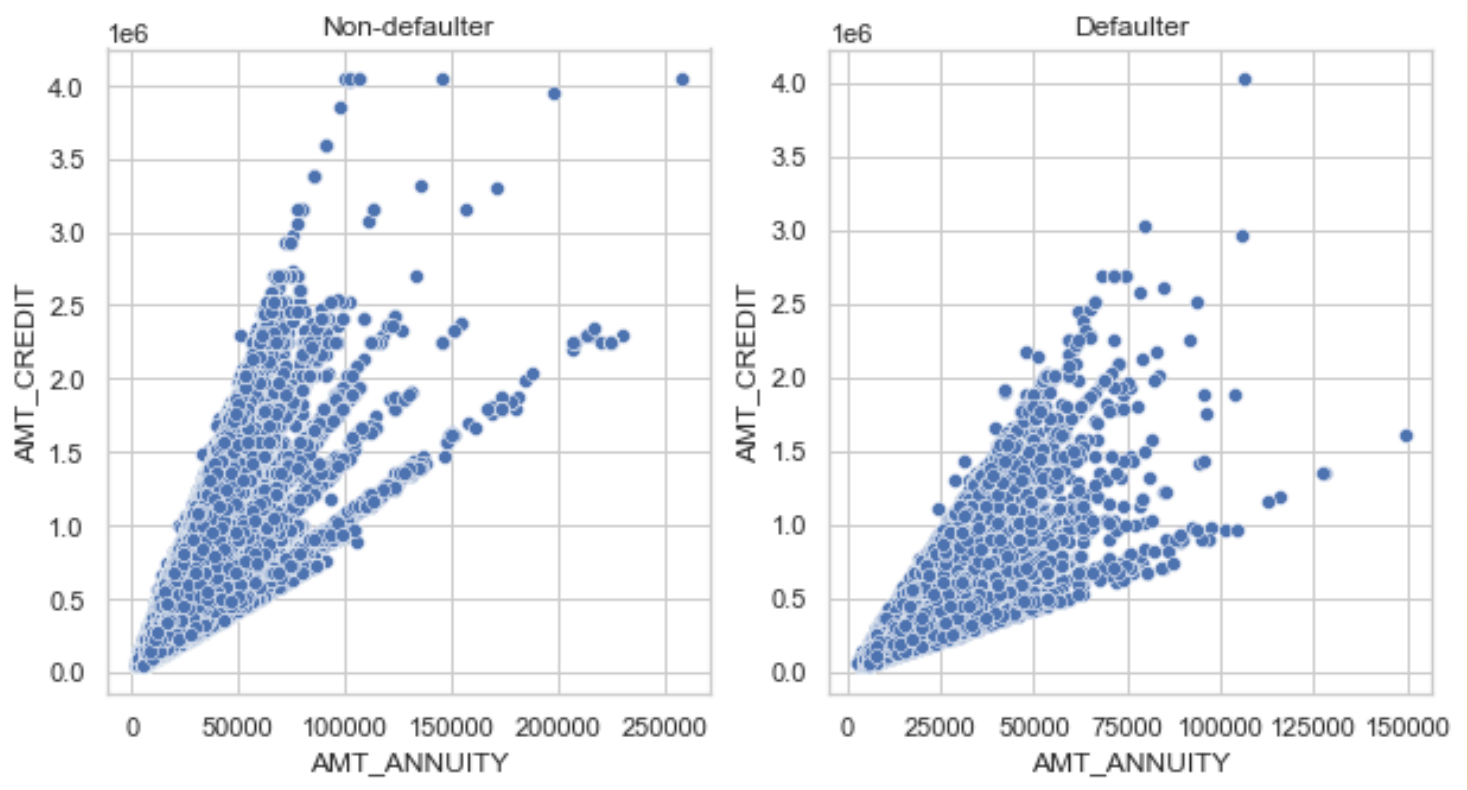
1. **Females** with **academic degree,** followed by **Higher education** are most likely to be **defaulters as well as non-defaulters.**
2. **Male** owning **Higher education, followed by incomplete education** are most likely to be **defaulters.**

Distribution of Loan amount vs Type of Goods price amount



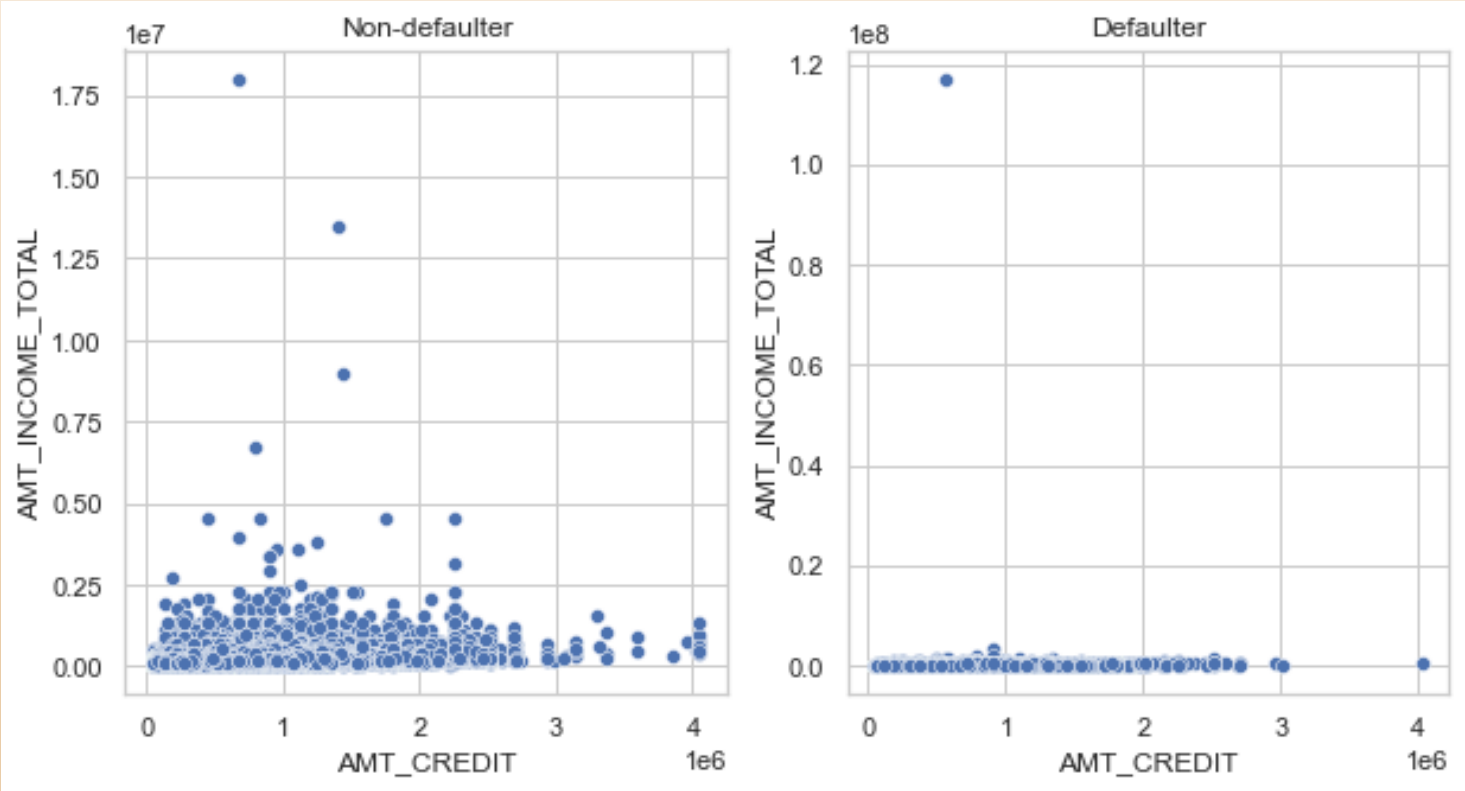
1. **Loan amount and Goods price amount show linear dependency.**
2. **Lesser defaulters than non-defaulters are there for lower range of Loan amount and Goods price amount.**
3. **Higher Goods price amount shows very less Loan amount in defaulters.**

Distribution of Annuity amount vs Credit amount



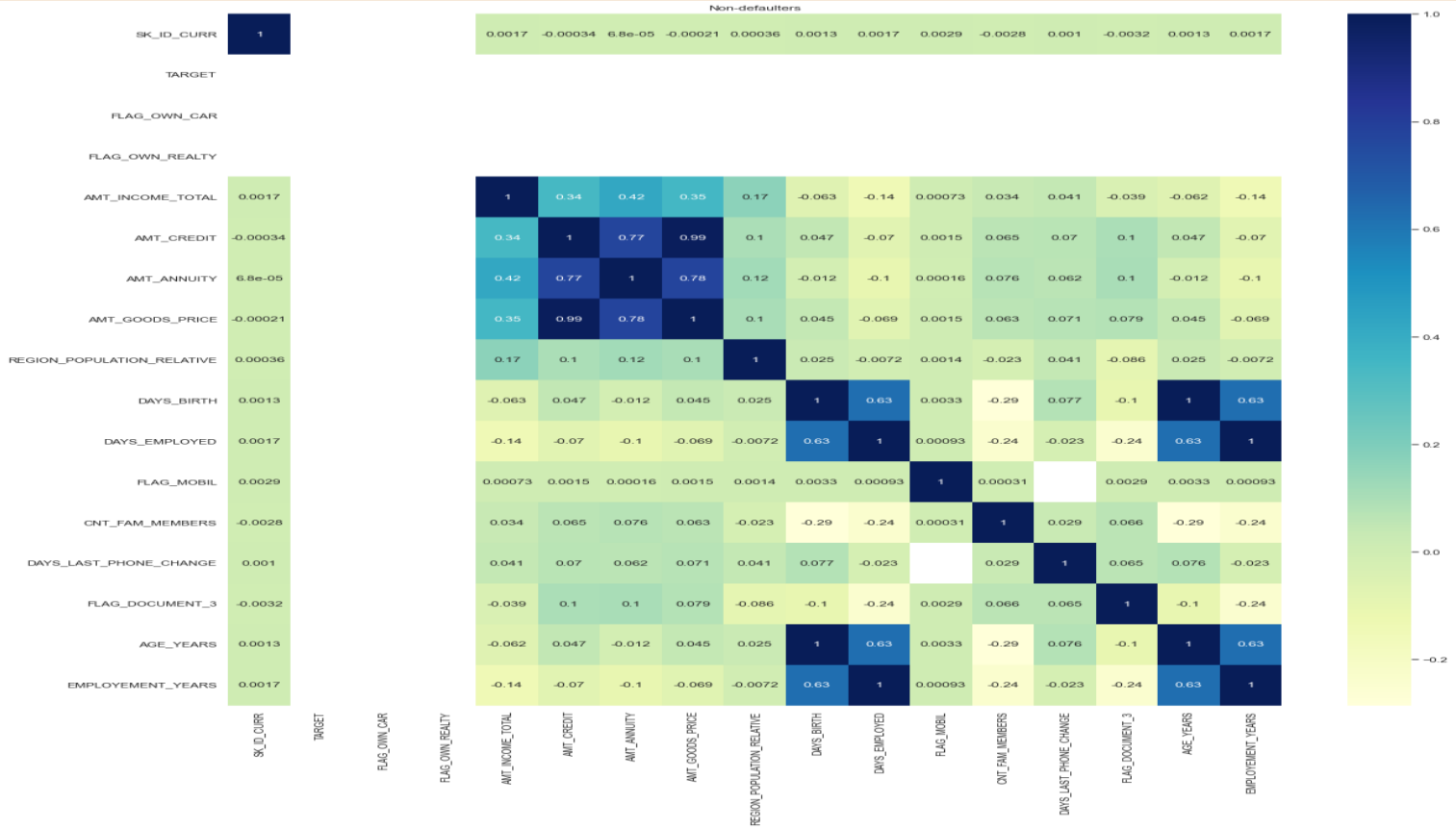
1. **Loan amount and Annuity amount show linear dependency for both defaulters and non-defaulters.**

Distribution of Loan amount vs Type of Goods price amount



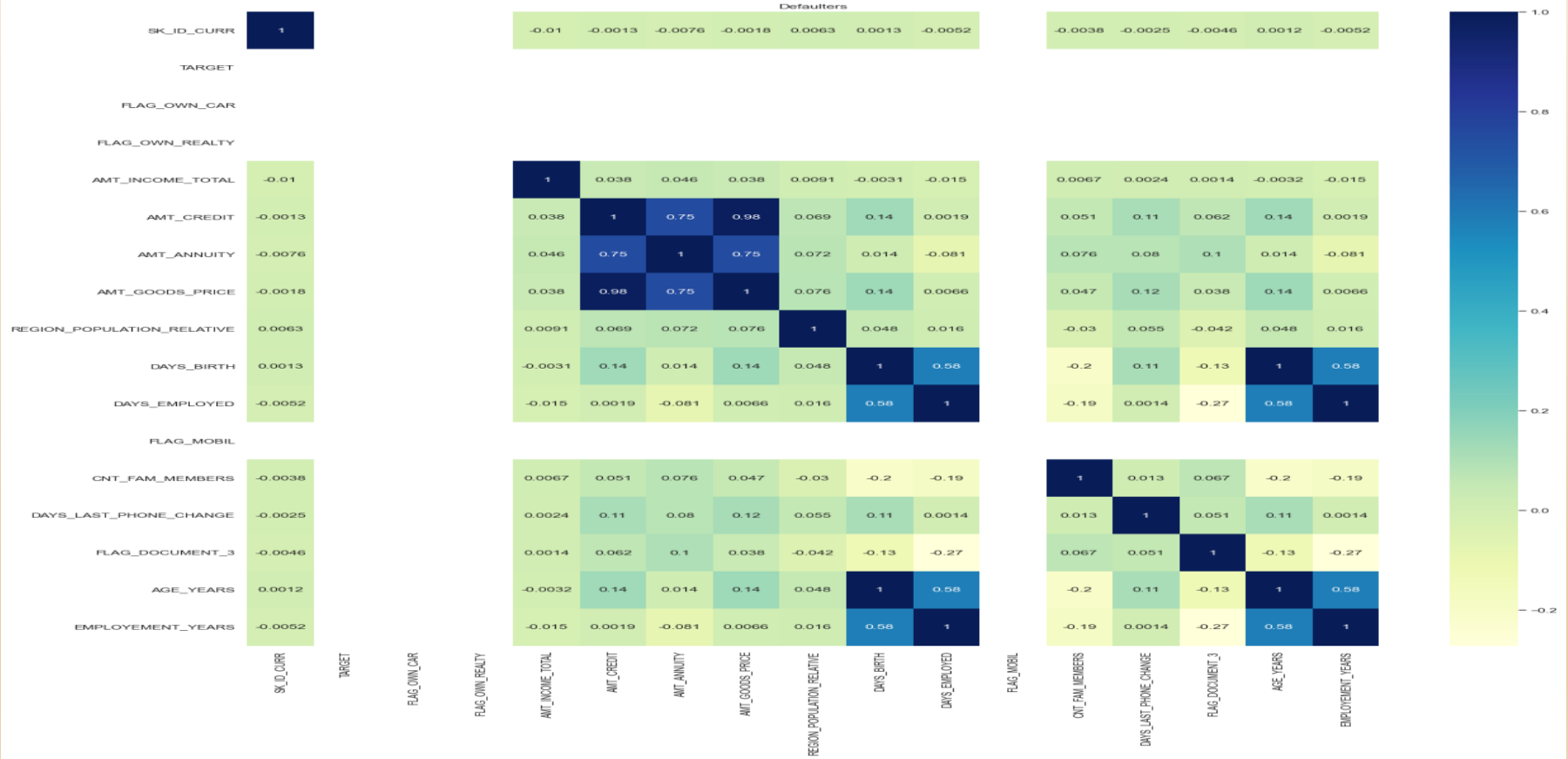
1. **Credit Amount does not vary much with Income in case of defaulters. Rather there is no relationship between Credit amount and Income**

Correlation in case of Non-defaulters



1. **Credit amount is highly correlated with: Goods Price Amount(0.98), Loan Annuity(0.75), Income (0.34)**
2. **Income has high correlation with: Credit amount(0.34), Annuity amount(0.42) and Goods Price Amount(0.35)**

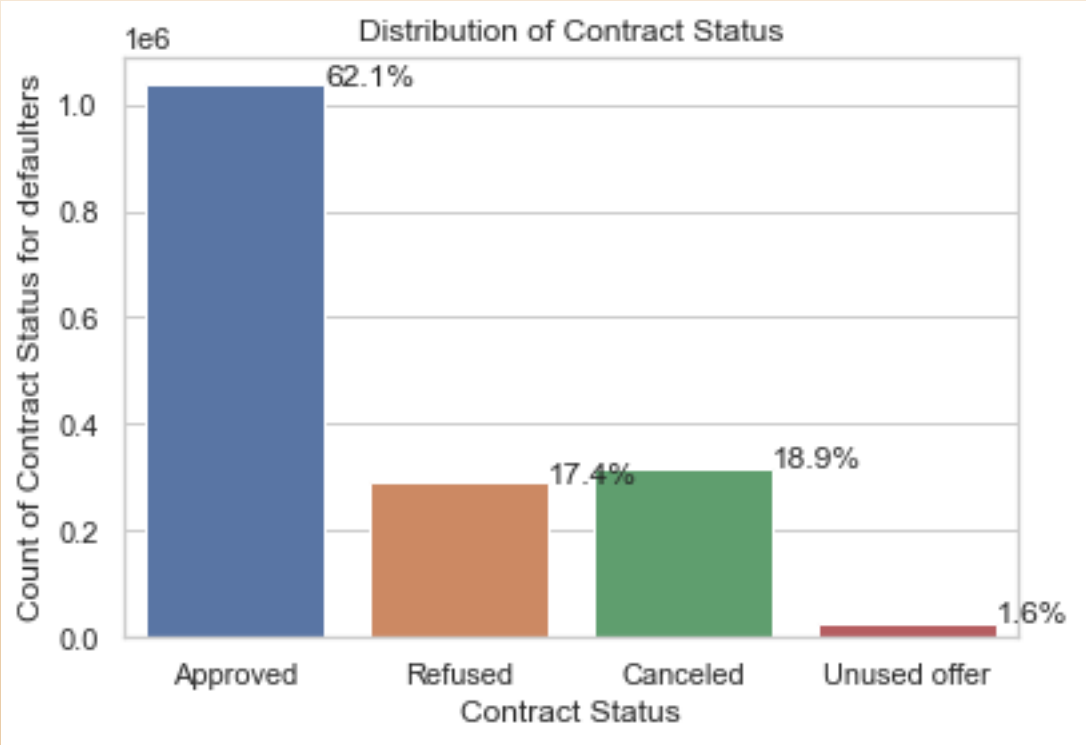
Correlation in case of Defaulters



1.Credit amount is highly correlated with: Goods Price Amount(0.98) & Loan Annuity(0.75)

2.Annuity amount shows high correlation of 0.75 with Goods price amount as well as Credit amount

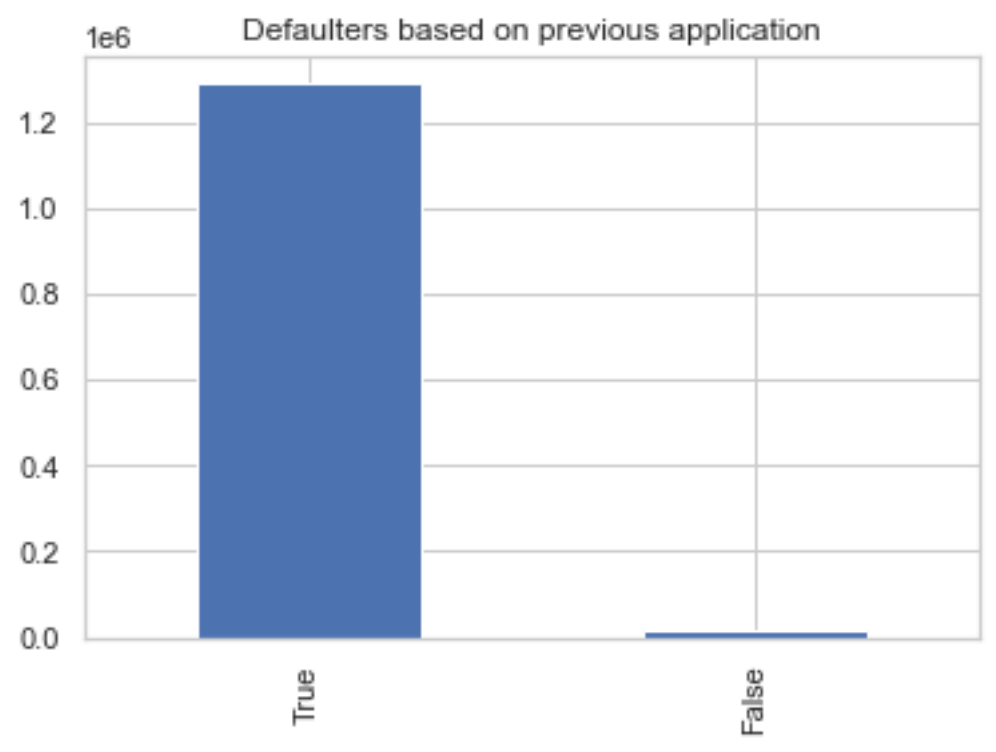
Status of Application extracted from previous_application.csv



1. **62.1 % of previous application are Approved**
2. **18.9 % are canceled**
3. **17.4 % is refused**
4. **1.6 % is unused**

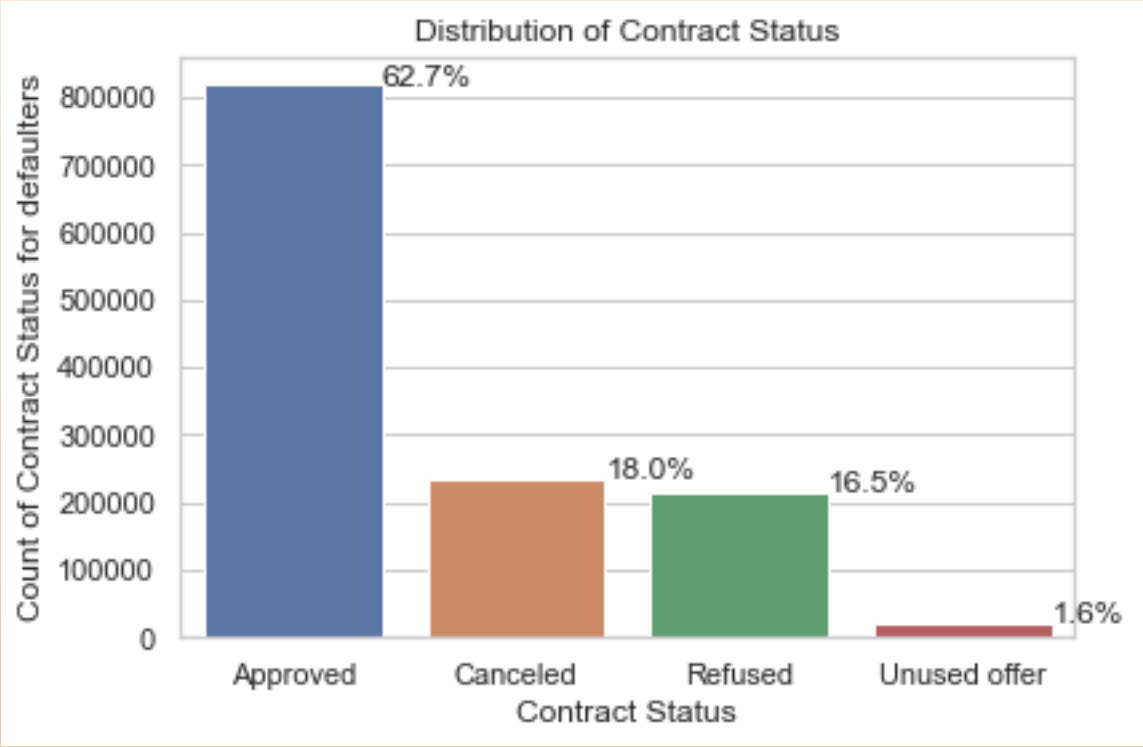
COMBINED DATASET

Status of Defaulters



- Great number of previous application are defaulters

Status of Application in defaulters



1. **62.7 % of approved are previous defaulters**
2. **18 % are canceled**
3. **16.4 % is refused**
4. **1.6 % is unused**

Conclusion

- Mostly females apply for loan, owing to which their number as defaulters is more
- Bank offers two kinds of loans: Cash loans and Revolving loans. Mostly Cash loans are offered, and defaulters are observed in cash loans. Less percentage of Revolving loans are offered, however, no defaulters are observed.
- Students and Businessman are not defaulters.
- Client with Higher academic degree as non-defaulters.
- Married client are more defaulters, while widows are non-defaulters
- Females with office apartment are most likely to be non-defaulters.
- Male owning rented apartment, office apartment, own apartment are most likely to be defaulters
- Customers from age 35-45 most likely to be non-defaulters.
- Customers in age 25-30 are most likely to be defaulters, Defaulters number reduces from 45 years
- Great number of previous application are defaulters

THANK YOU!!!!!