



CREDIT RISK ANALYSIS EDA CASE STUDY

BY
**SEYED JAVIDH &
VIVEK CHOWDHURY**

INTRODUCTION / PROBLEM STATEMENT

As there an exponential increase in the number of people taking loans over the year, it is important for the banks to stay a step ahead of the game. Increase in the amount of loans given out to people has always possessed a risk.

One of the most difficult challenges companies giving out loans face are:

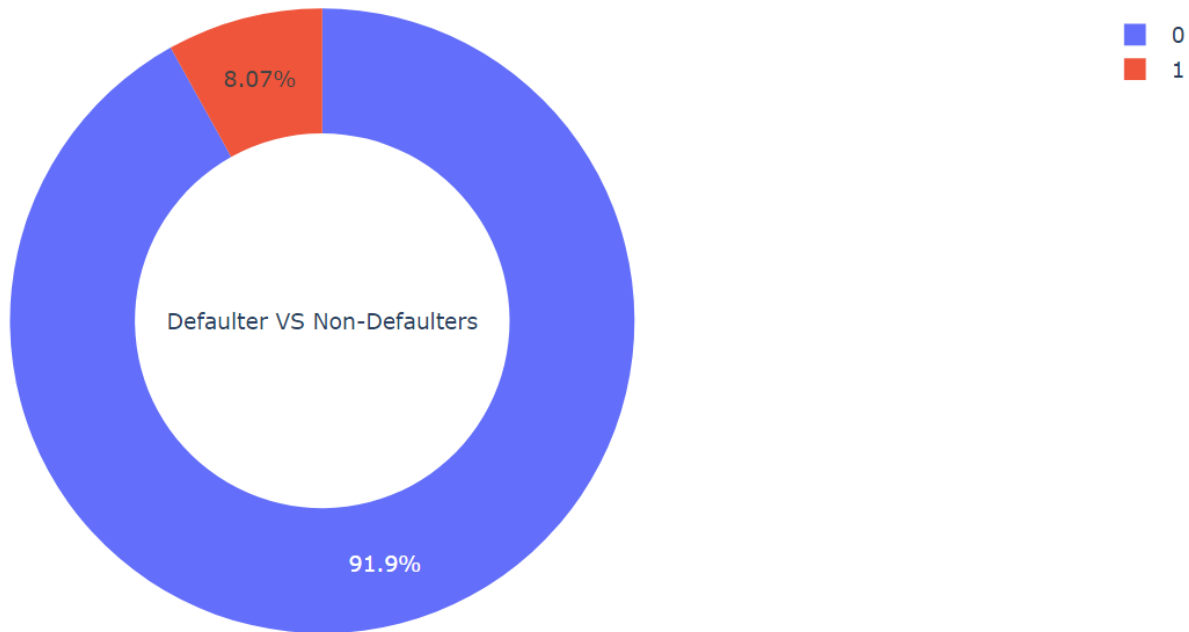
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Hence, with the help of exploratory data analysis we can use the rich arsenal of knowledge and technique at our disposal to overcome this obstacle.

How do we do that you ask? Take a look for yourself in the next slides...

DATA IMBALANCE

TARGET IMBALANCE



Inference

- There is a huge data imbalance with almost 92% of the data corresponding to the Non-Defaulter whereas only 8% belonging to the Defaulters.

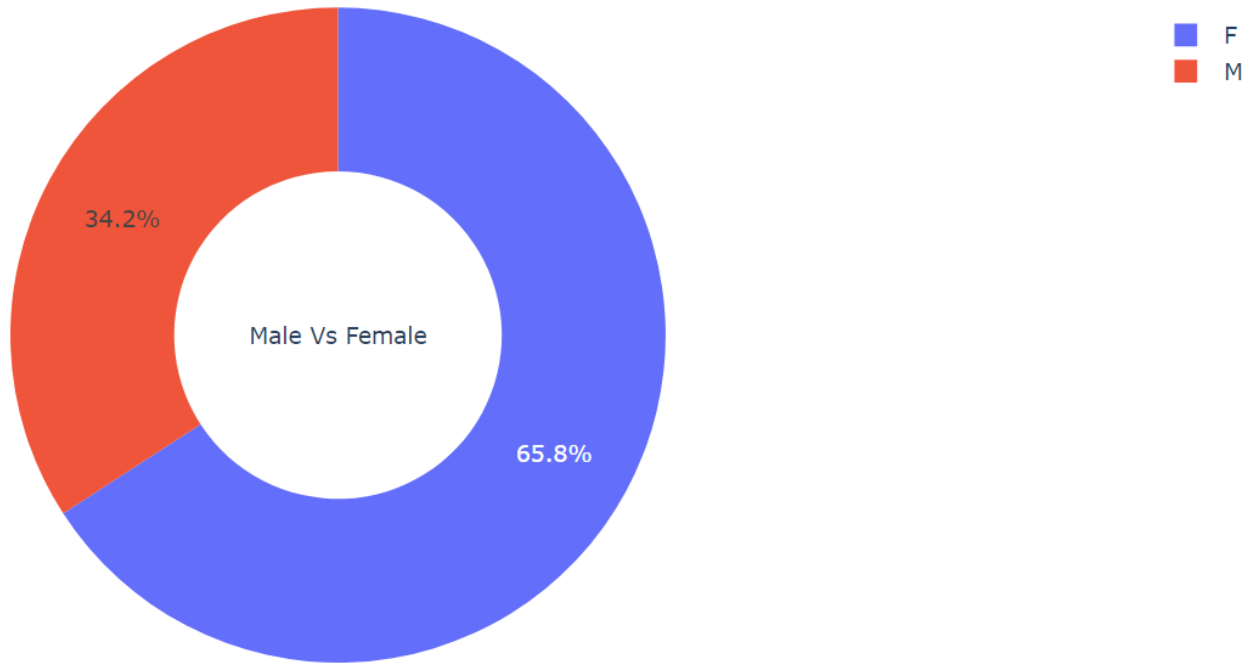
91.9%

NON - DEFAULTERS

8.07%

DEFAULTERS

GENDER IMBALANCE



Inference

- The majority of the loan applicants are female constituting around 66% whereas only 34% are male.

65.8%

FEMALE

34.2%

MALE

CORRELATIONS

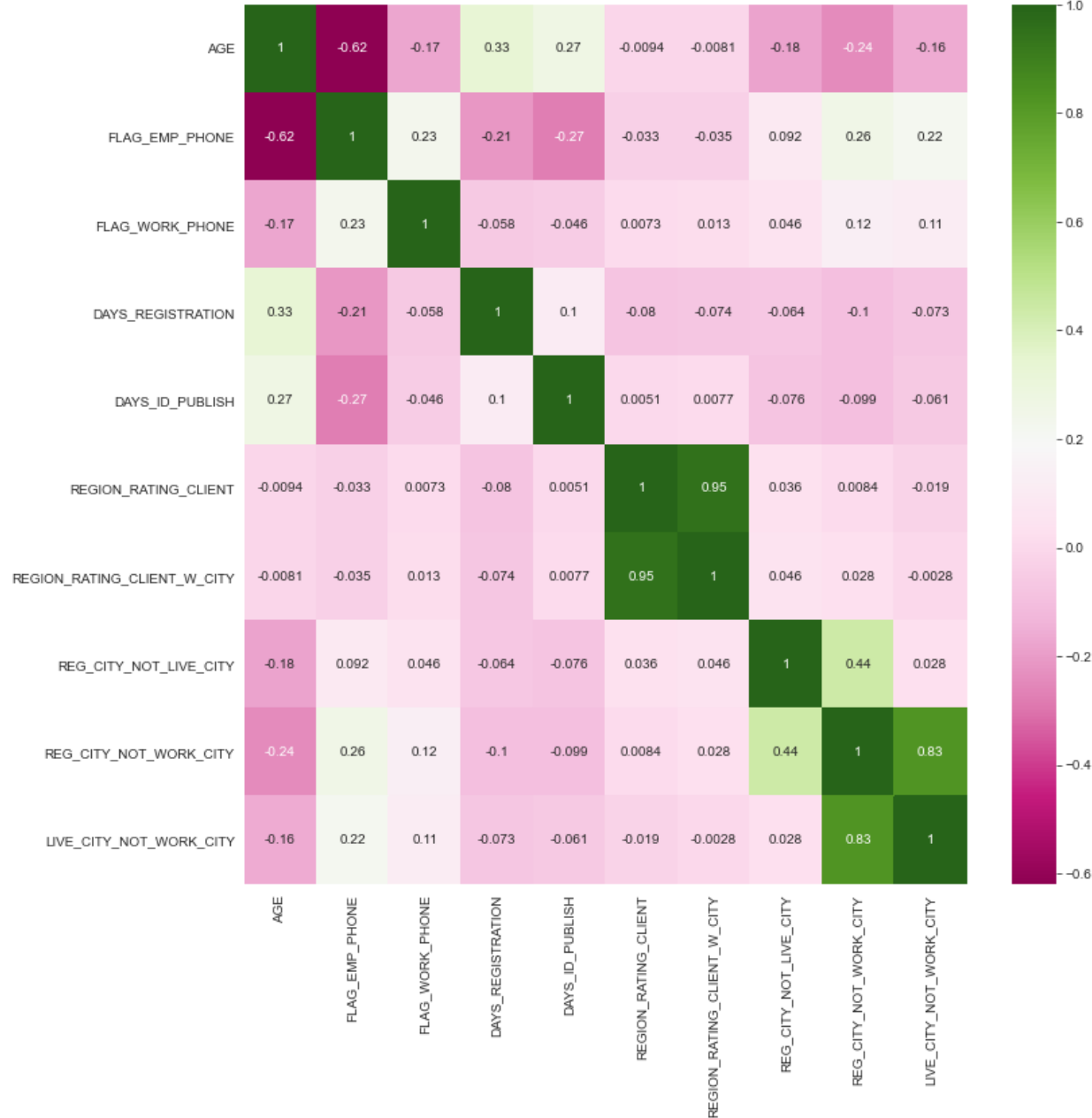
Correlation between variables



Inference

- Here, we can see a very strong correlation between the amount of goods price and the loan amount. From this, we can conclude like previously that the loan amount disbursed is mostly equal or slightly higher than the cost of article the client wishes to purchase.
- There is also a good correlation between the annuity amount and the loan amount as well as the good's price.
- Here there is a negative correlation between the client's region and the money he earns. This means that if a client is from a place with a higher rating, he or she will more likely earn less money.

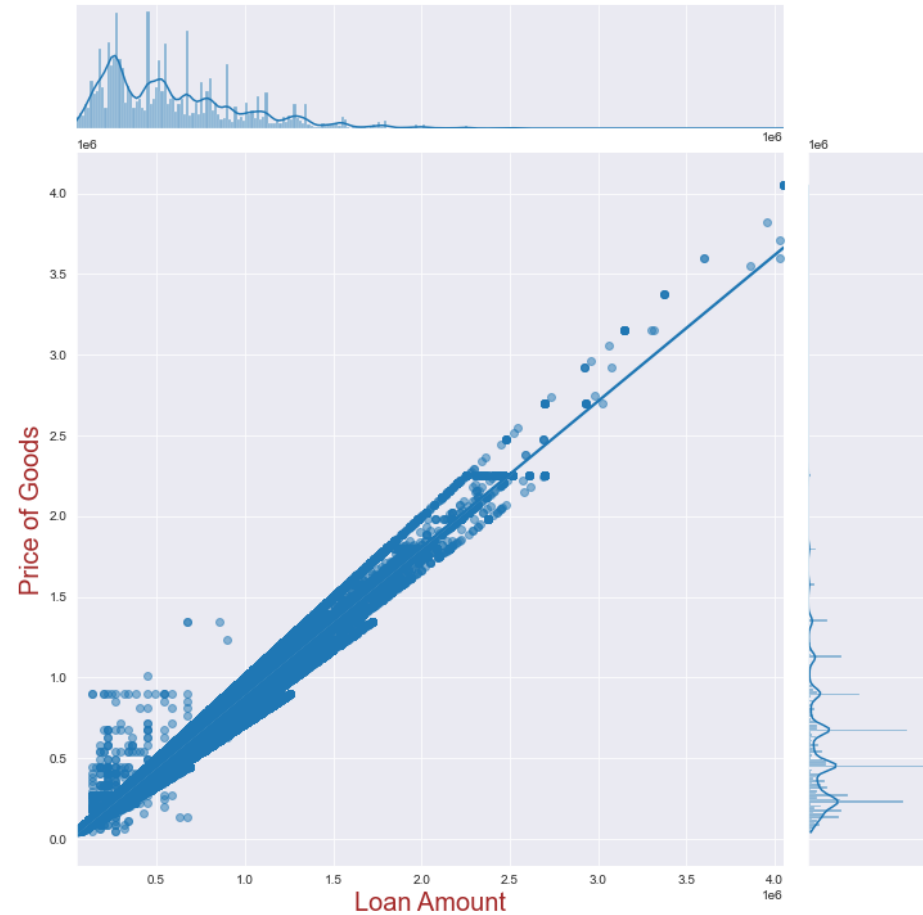
Correlation between variables



Inference

- Here we see a strong negative correlation between employee phone number and age.
- There is a positive correlation between the number of days before which client changed his registration with respect to age. This goes to show that elderly people are less likely to make changes to their registration prior to applying for loan.
- Clients that do not provide their phone numbers are also less likely to provide incorrect permanent and work address.

Correlation between the Loan amount and the price of goods for which loan was given

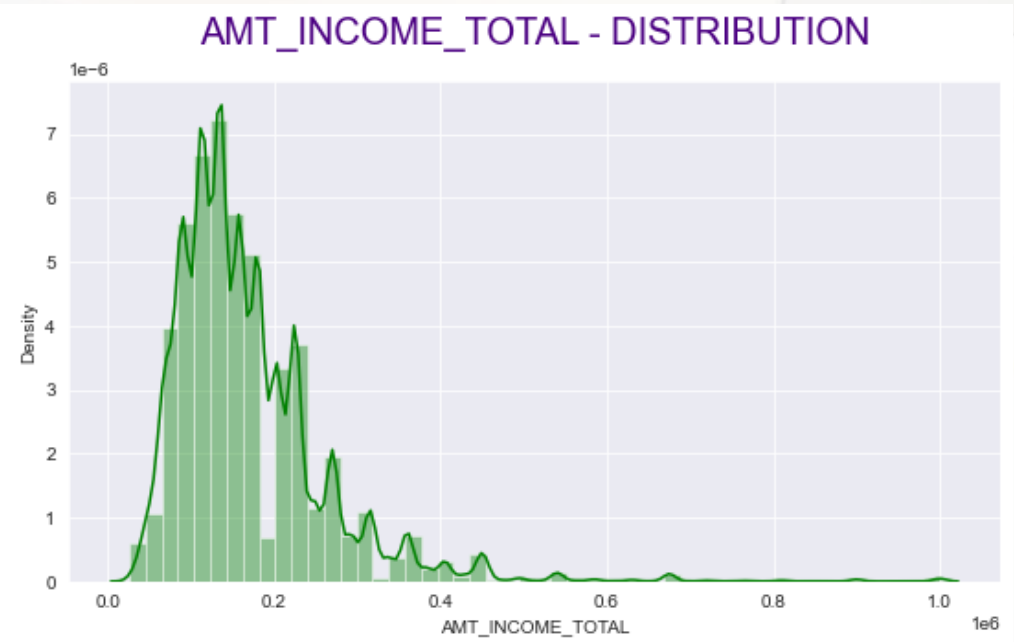
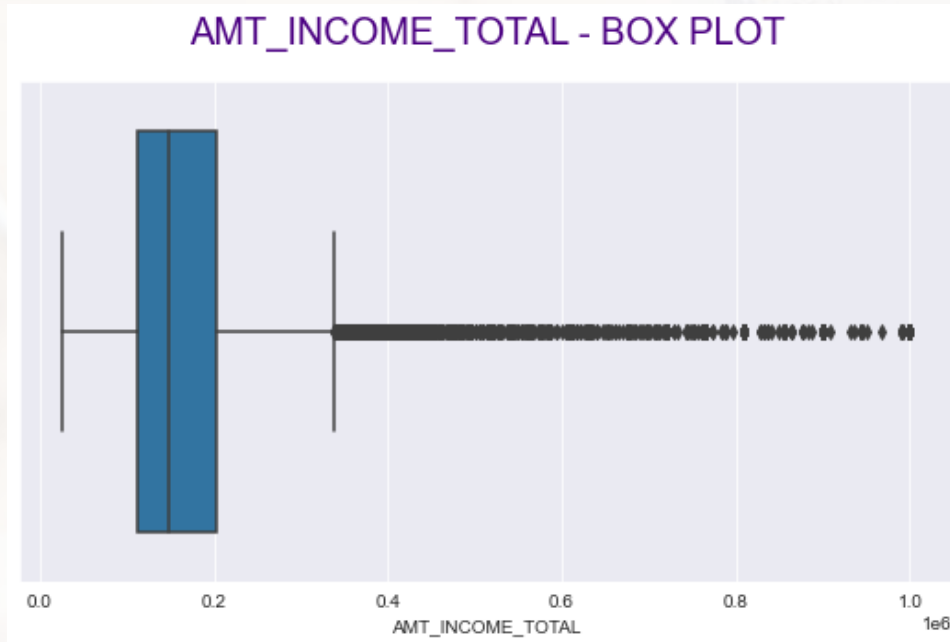


Inference

- Since there is a very linear and positive correlation between the Loan Amount and the Good's price, we can assume that, in most cases the loan amount demanded by the customer is slightly more than but mostly equal to the price of the article he/she wishes to purchase

UNIVARIATE ANALYSIS

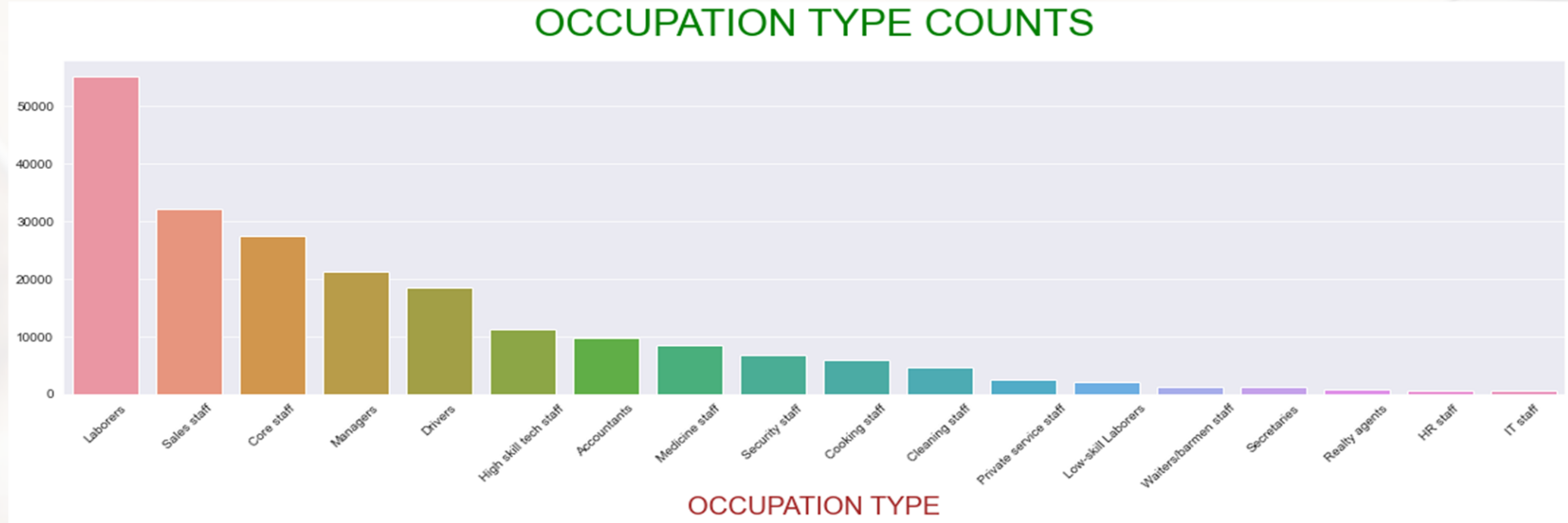
DISTRIBUTIONS OF INCOME



Inference

- It can be inferred that, most of the people earn around 1-2 lakh annually.
- There are of course people who earn a lot more, but they are present in mere numbers up to 10lakhs.
- Largely the bigger part of the population, applying for loan is concentrated near the 20 thousand to 4 lakh bucket

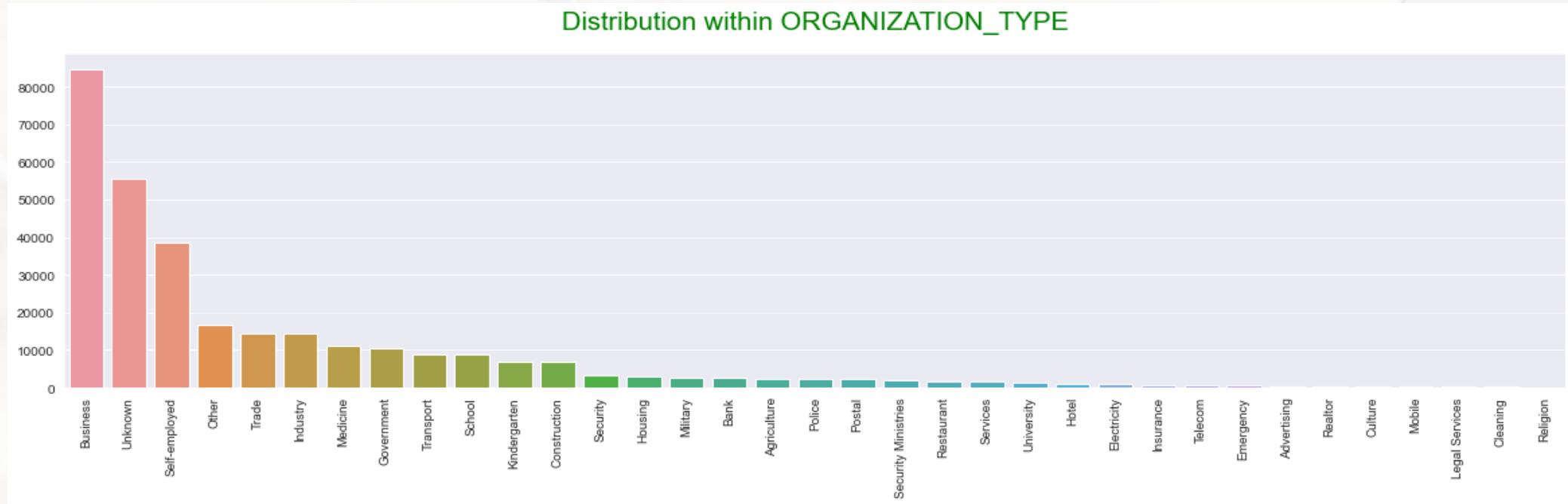
DISTRIBUTIONS OF OCCUPATION TYPE



Inference

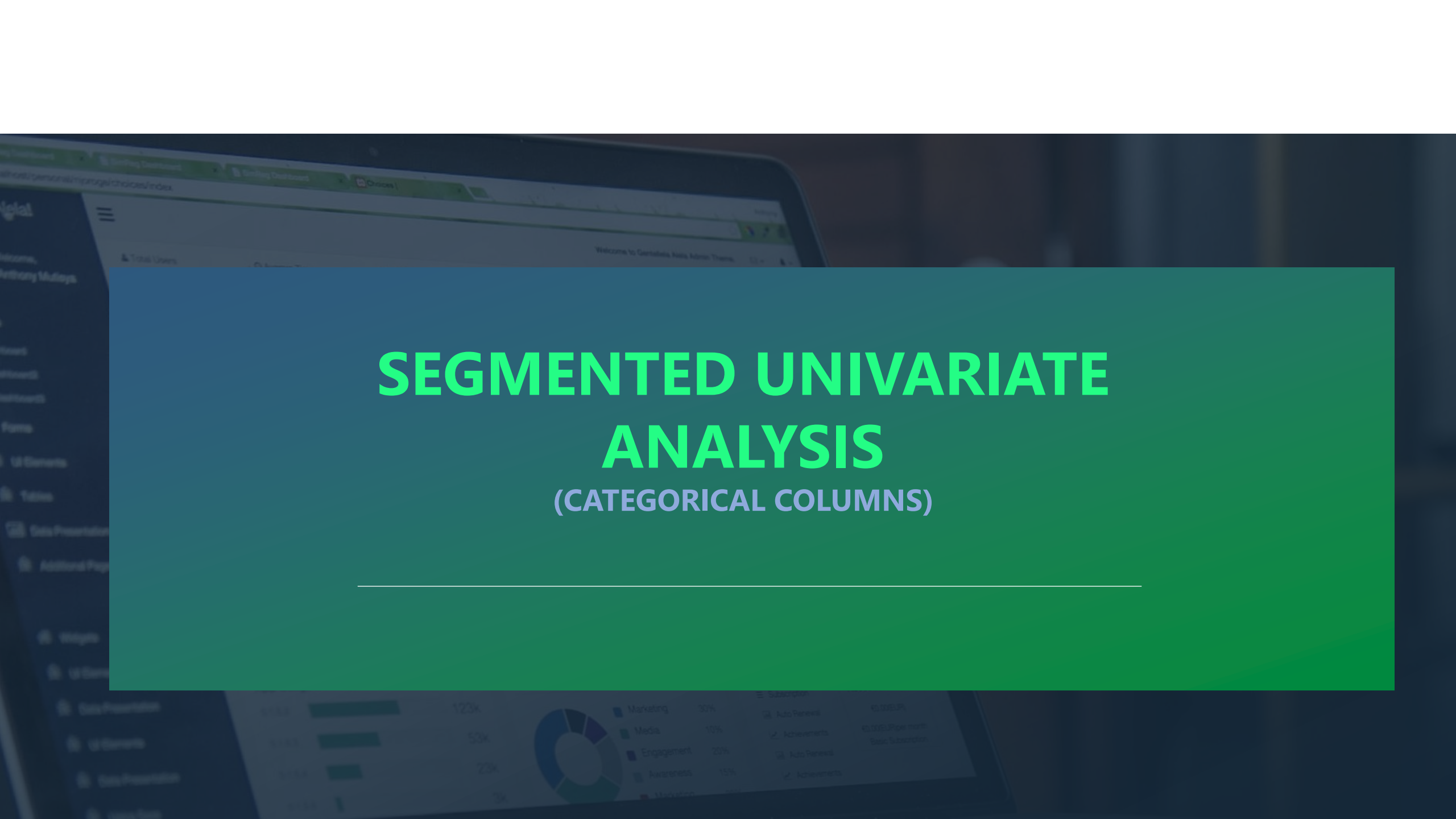
- Laborers and Sales Staff constitute the majority whereas IT Staff and HR staff are on the lower side.

DISTRIBUTIONS OF ORGANIZATION TYPE



Inference

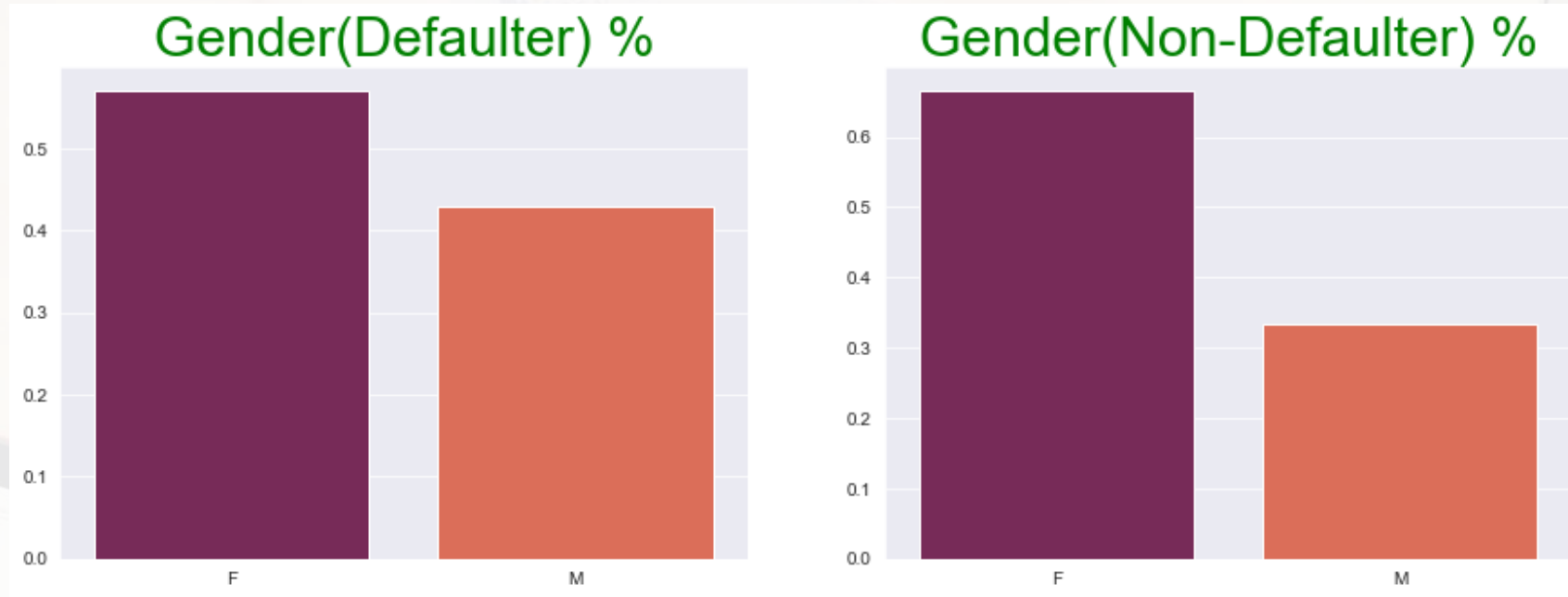
- people who are in business field applied more in number for the loan compared to the other fields.



SEGMENTED UNIVARIATE ANALYSIS

(CATEGORICAL COLUMNS)

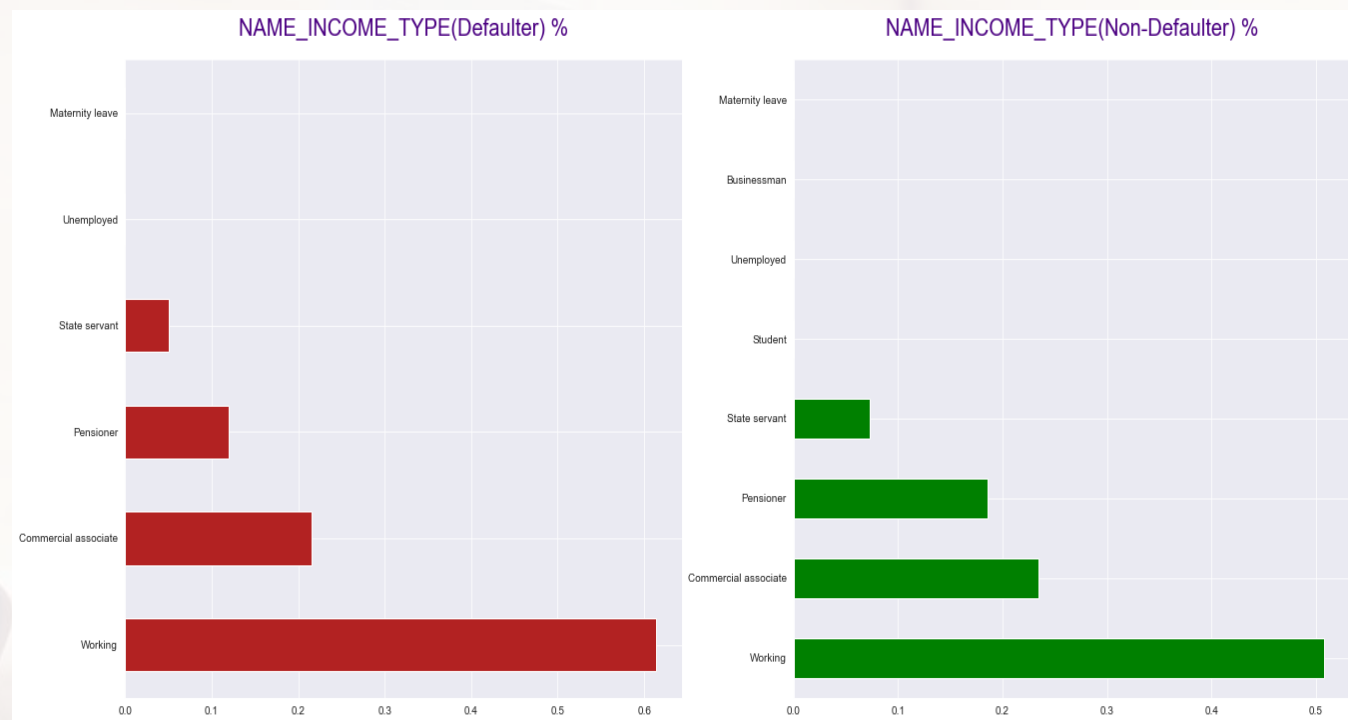
DEFAULTER PROPORTION BY GENDER



Inference

- Here we see that the male % has increased almost by 10% from non-defaulter to defaulter.
- In-case of female, we can see that there is also a similar 10% decrease from defaulter to non-defaulter.
- We can imply that, men are more likely to default a loan than women.

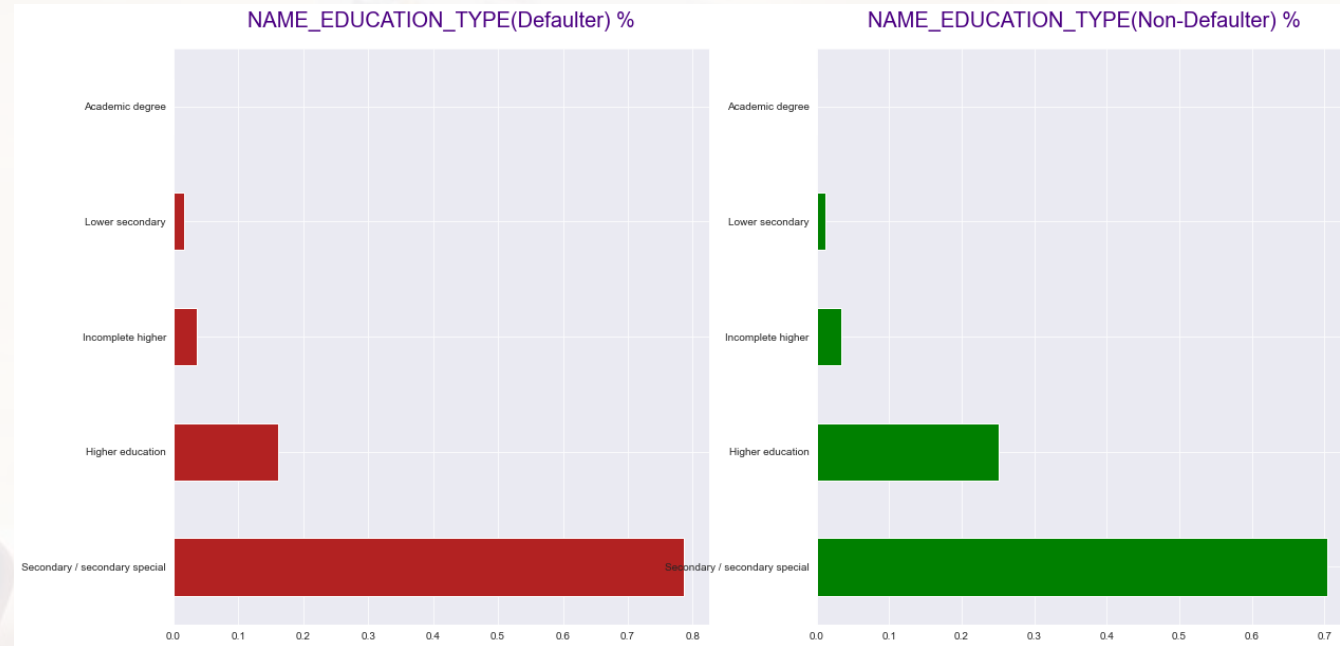
DEFAULTER PROPORTION BY INCOME TYPE



Inference

- We can notice that "Students" do not appear on the defaulters as they don't have to pay when they study. So they are a very good client to target.
- Also businessmen don't default much like students category
- Also, we see more than 10% increase in the number of "Working" category people who default loans.

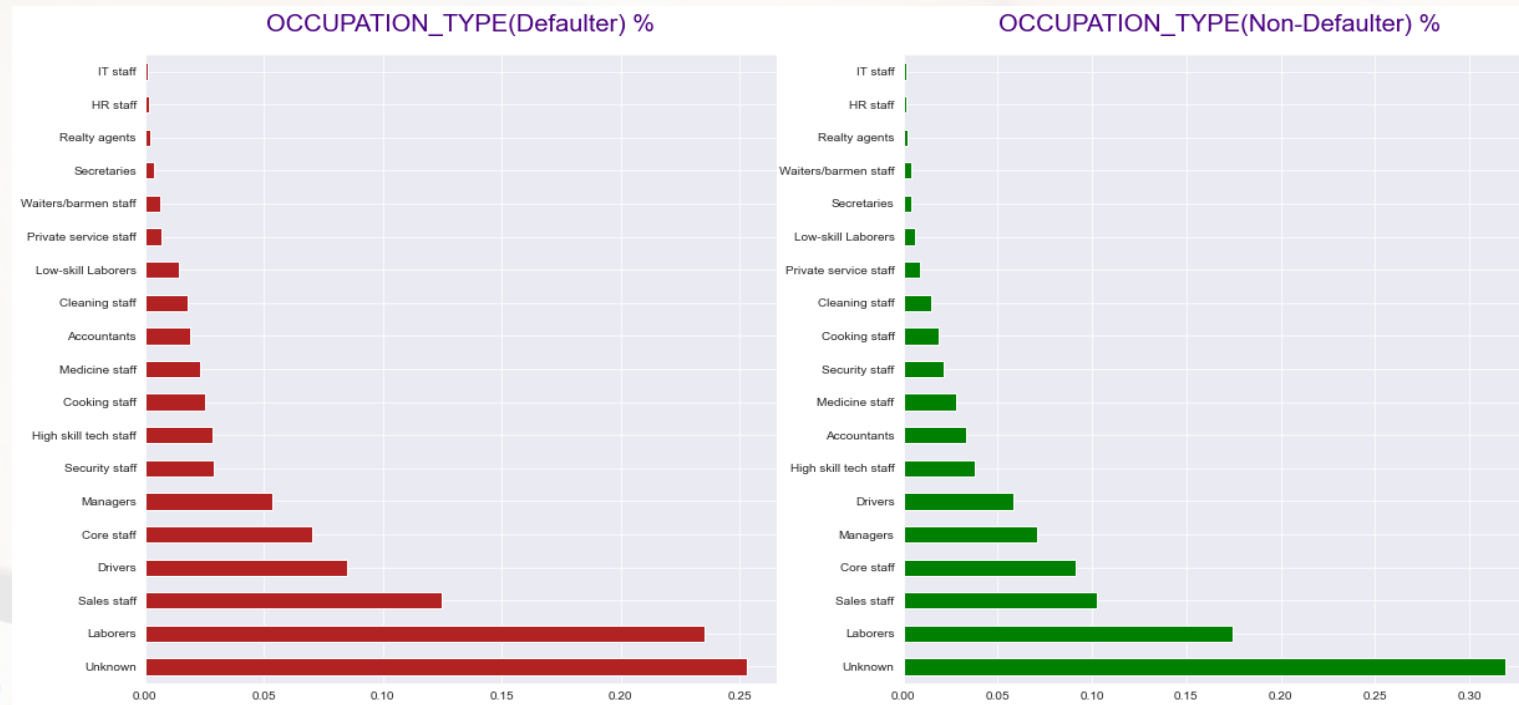
DEFAULTER PROPORTION BY EDUCATION TYPE



Inference

- From the above graphs we can make out that people who pursue "Higher Education" are less likely to default loans.
- Client who have attained only "Secondary education" are more likely to default.

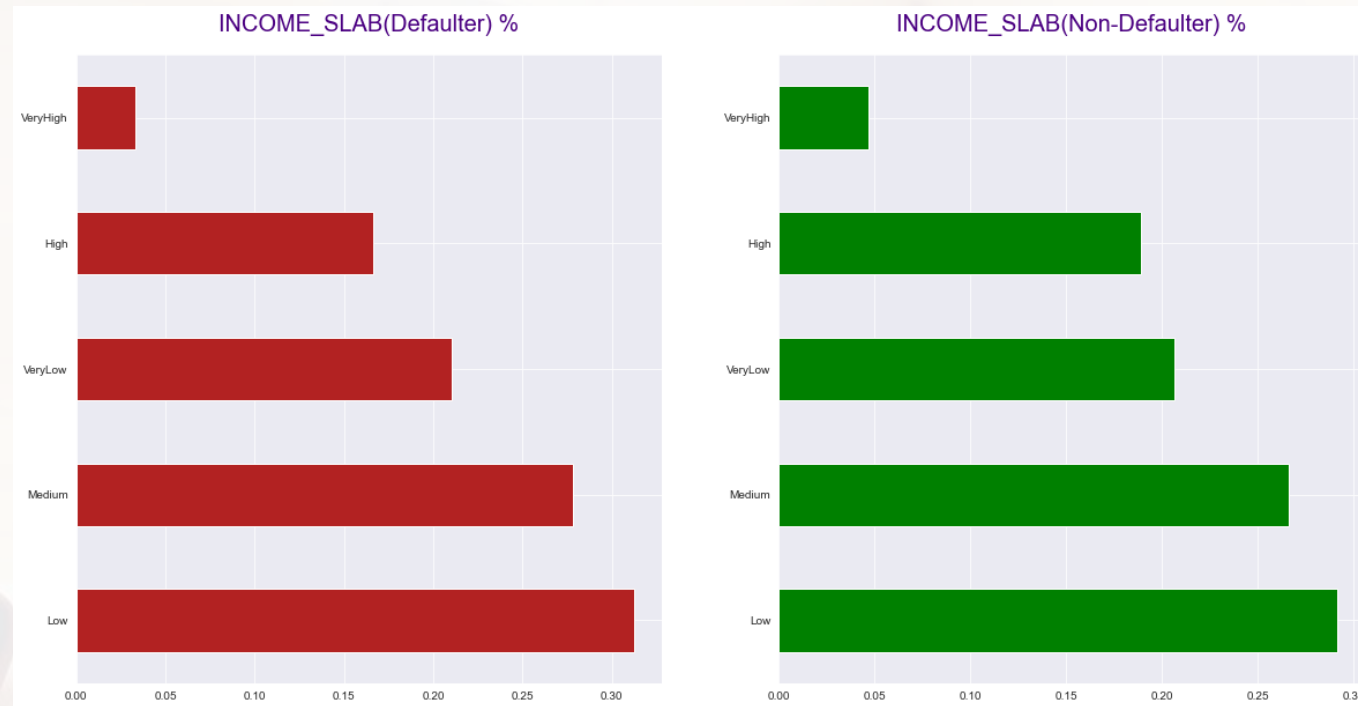
DEFAULTER PROPORTION BY OCCUPATION TYPE



Inference

- From the above graph, we can understand that, Laborers, Sales staff, drivers, cleaning staff, low-skill labors are more likely to default a payment of the loan.
- The best clients to target in this case would be Managers, core staff, high skill tech staff.

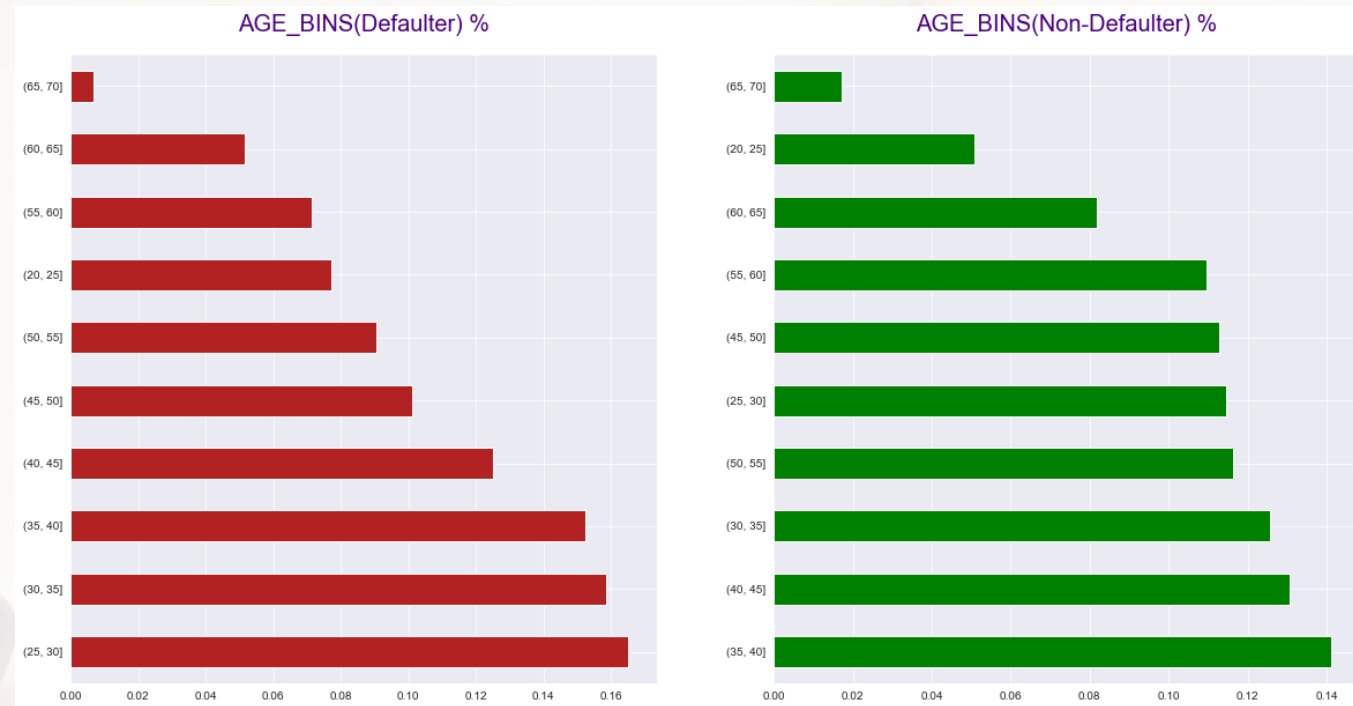
DEFAULTER PROPORTION BY INCOME GROUPS



Inference

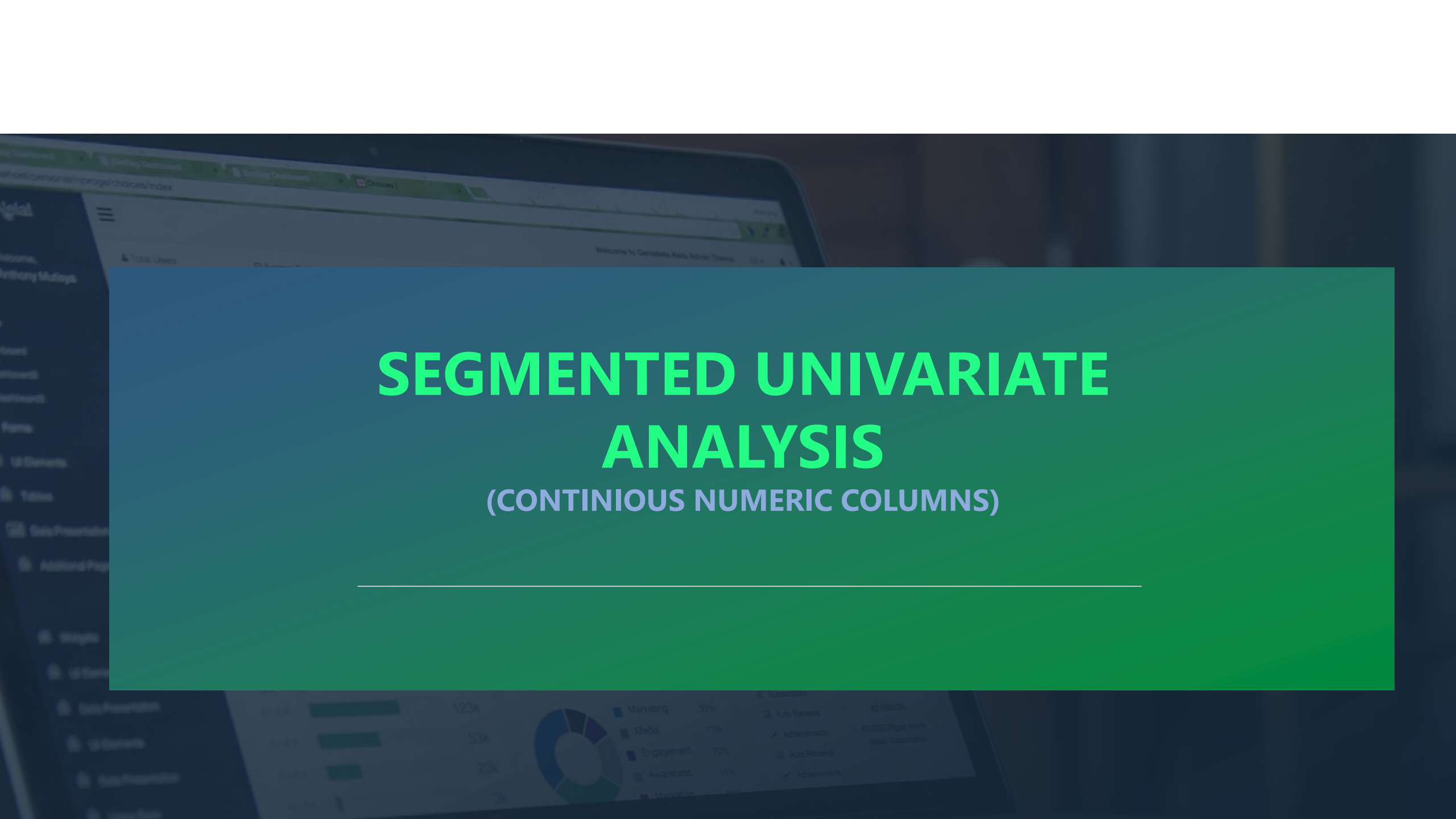
- Here people belonging to very high, high income slab do not face much difficulty with loan repayment.
- However, people with low income struggle to make payment and are likely to default.

DEFAULTER PROPORTION BY AGE GROUPS



Inference

- From the above graph we can infer that the age bin from 25 to 40 are more likely to default a loan payment.
- People above 40 are less likely to default.
- With increasing age group, people tend to default less.



SEGMENTED UNIVARIATE ANALYSIS

(CONTINUOUS NUMERIC COLUMNS)

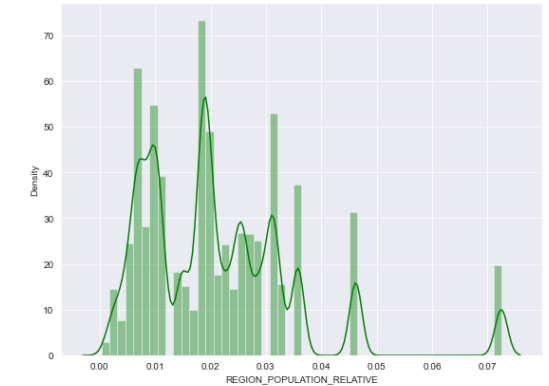
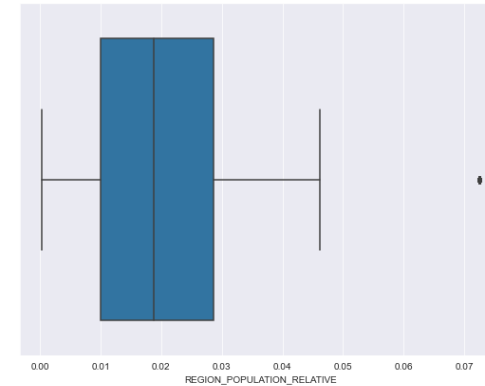
DEFAULTER PROPORTION BY REGIONAL POPULATION OF CLIENT

Inference

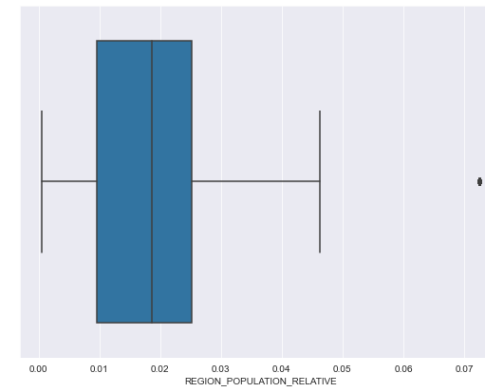
- From the above histograms, we can see that people who live in a place which is not so populated, like village or small towns, have difficulty in repaying loan amount.
- We can also see that people who live in cities which are more populated, do not face much difficulty with loan payments.

REGION_POPULATION_RELATIVE for Defaulters and Non-Defaulters

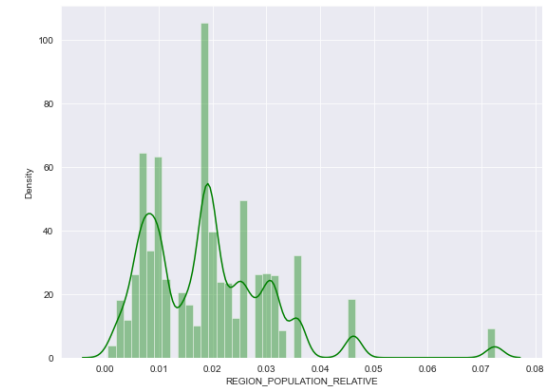
REGION_POPULATION_RELATIVE - BOXPLOT(Non-Defaulter) REGION_POPULATION_RELATIVE - DISTRIBUTION(Non-Defaulter)



REGION_POPULATION_RELATIVE - BOXPLOT(Defaulter)



REGION_POPULATION_RELATIVE - DISTRIBUTION(Defaulter)



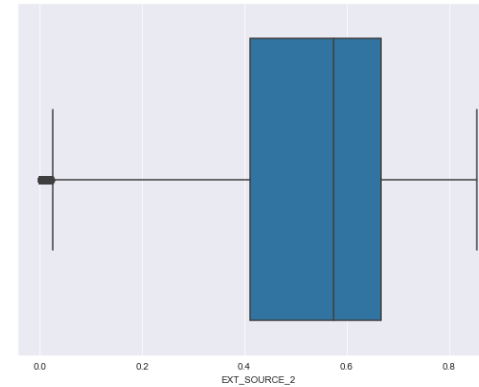
DEFAULTER PROPORTION BY EXTERNAL DATA SCORE

Inference

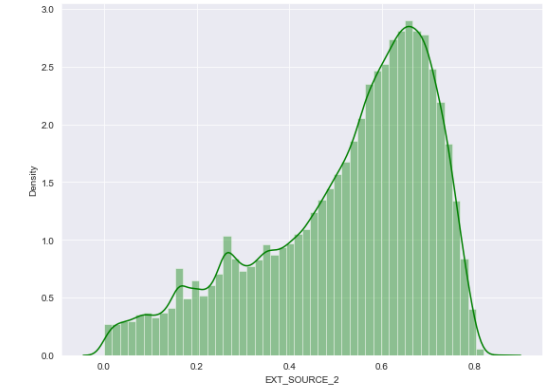
- We can see that people who face difficulty in paying a loan back are the ones, whose external source score are below 2.0.
- On the other hand, people with score above 2.0 are less likely to default on a loan payment.

EXT_SOURCE_2 for Defaulters and Non-Defaulters

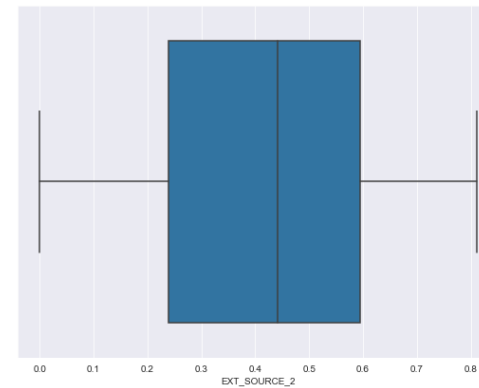
EXT_SOURCE_2 - BOXPLOT(Non-Defaulter)



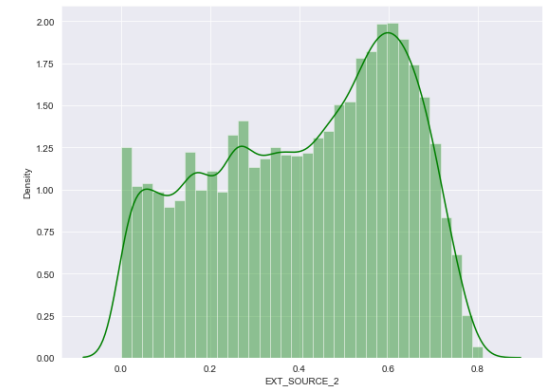
EXT_SOURCE_2 - DISTRIBUTION(Non-Defaulter)



EXT_SOURCE_2 - BOXPLOT(Defaulter)



EXT_SOURCE_2 - DISTRIBUTION(Defaulter)

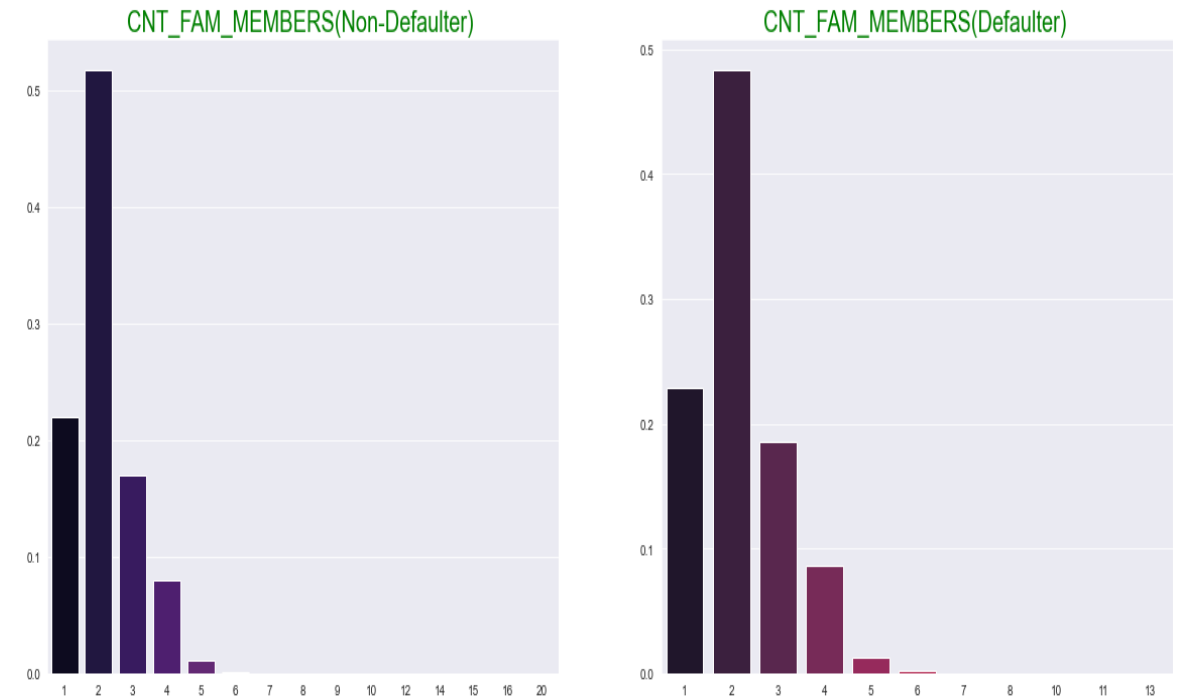


DEFAULTER PROPORTION BY FAMILY MEMBER COUNT

Inference

- We do not observe any significant impact of the number of family members of a client on defaulting.
- We do however, see a very small trend that, clients who default might have more than 4 family members.

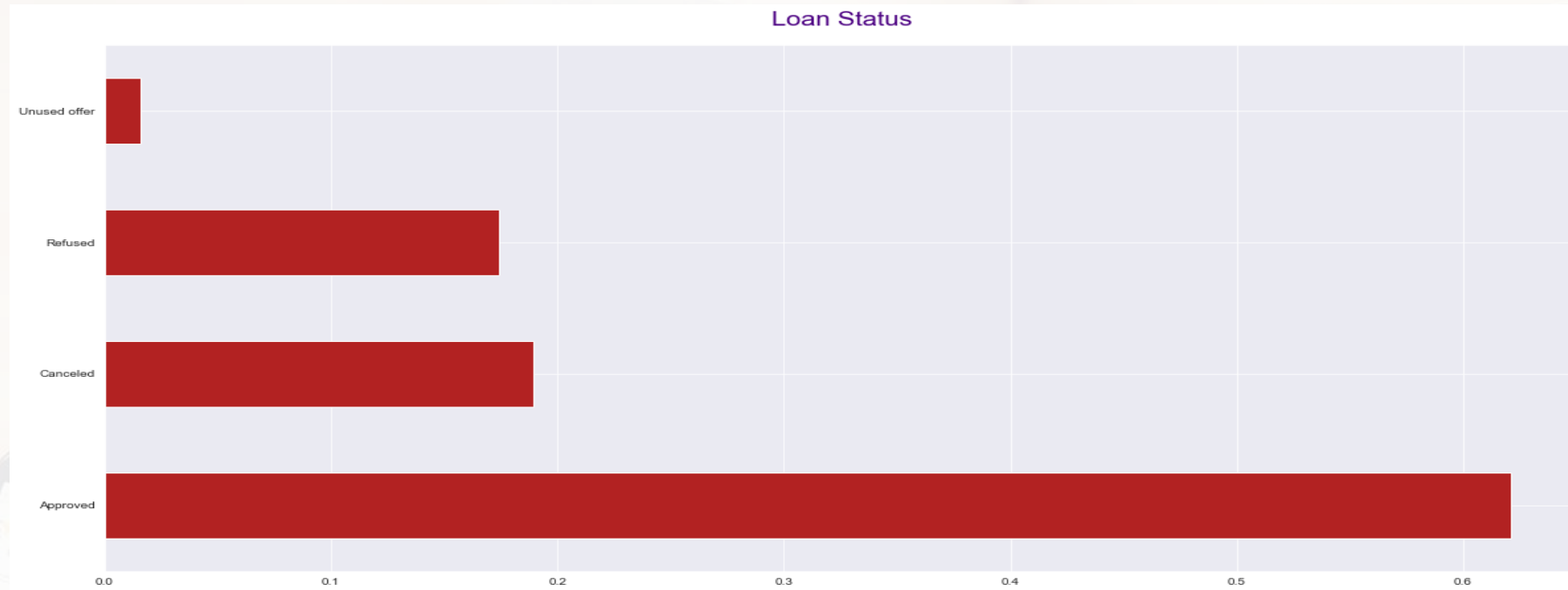
CNT_FAM_MEMBERS for Defaulters and Non-Defaulters





PREVIOUS APPLICATION DATA ANALYSIS

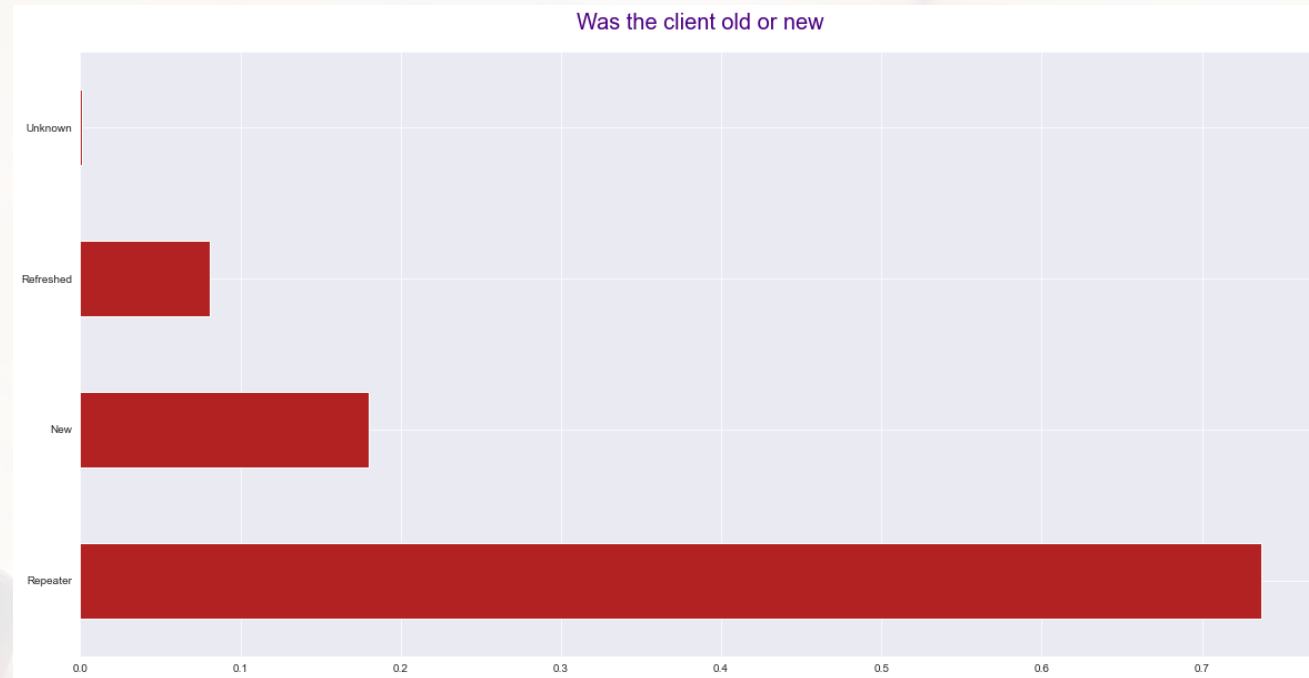
ANALYSIS ON LOAN STATUS



Inference

- About 62% of the loans were approved by the bank.
- 19% were cancelled and 17% were refused.

ANALYSIS ON CLIENT TYPE



Inference

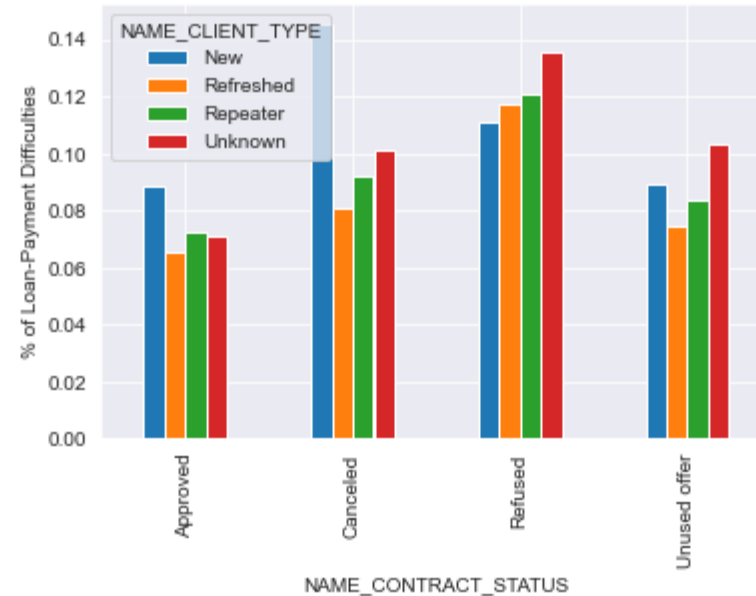
- From the plot above we can clearly see that most of the customers are repeaters.
- Only about 19% of the customers are new.



BIVARIATE / MULTIVARIATE ANALYSIS

CONTRACT STATUS VS CLIENT TYPE

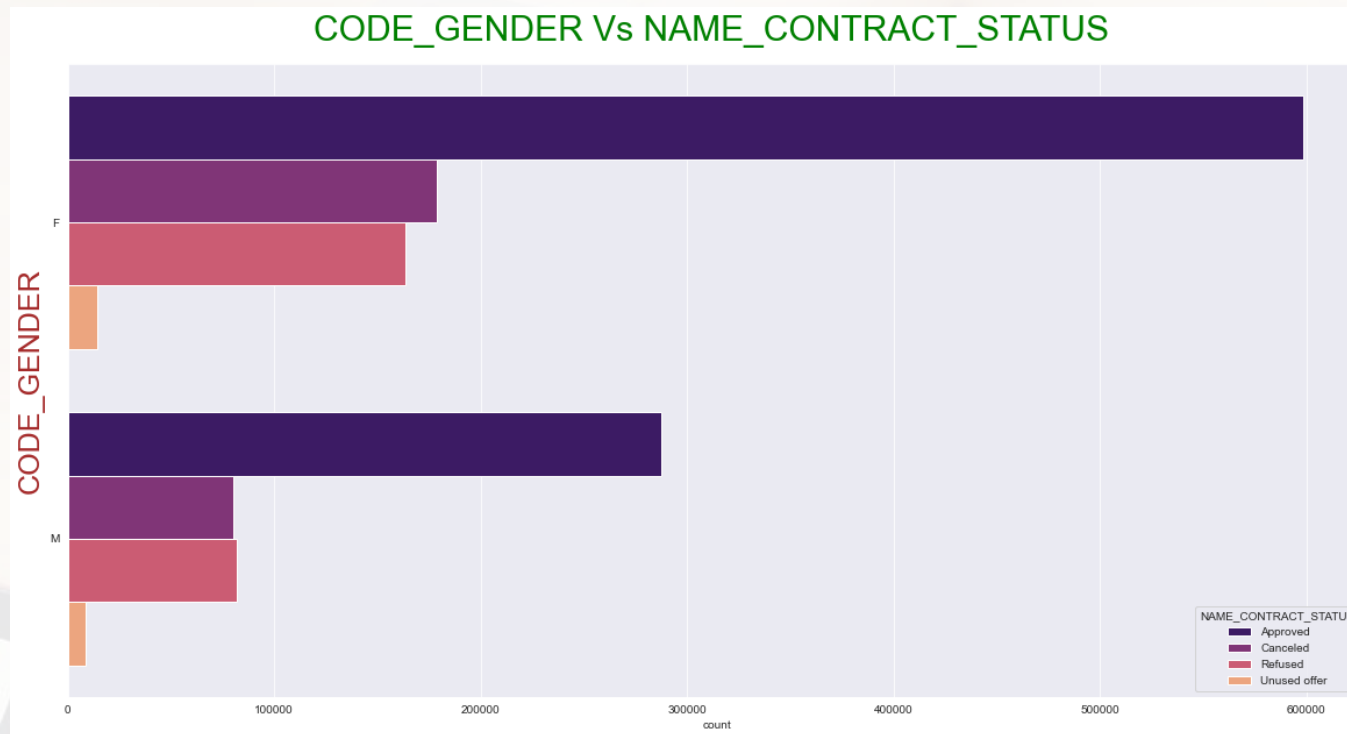
% of Loan Payment Difficulties for NAME_CONTRACT_STATUS and NAME_CLIENT_TYPE



Inference

- From the above data we can infer that new clients are more likely to cancel loans.
- Also, new clients are more likely to get their loan amount refused.
- Repeater clients are more likely to get a loan refused.

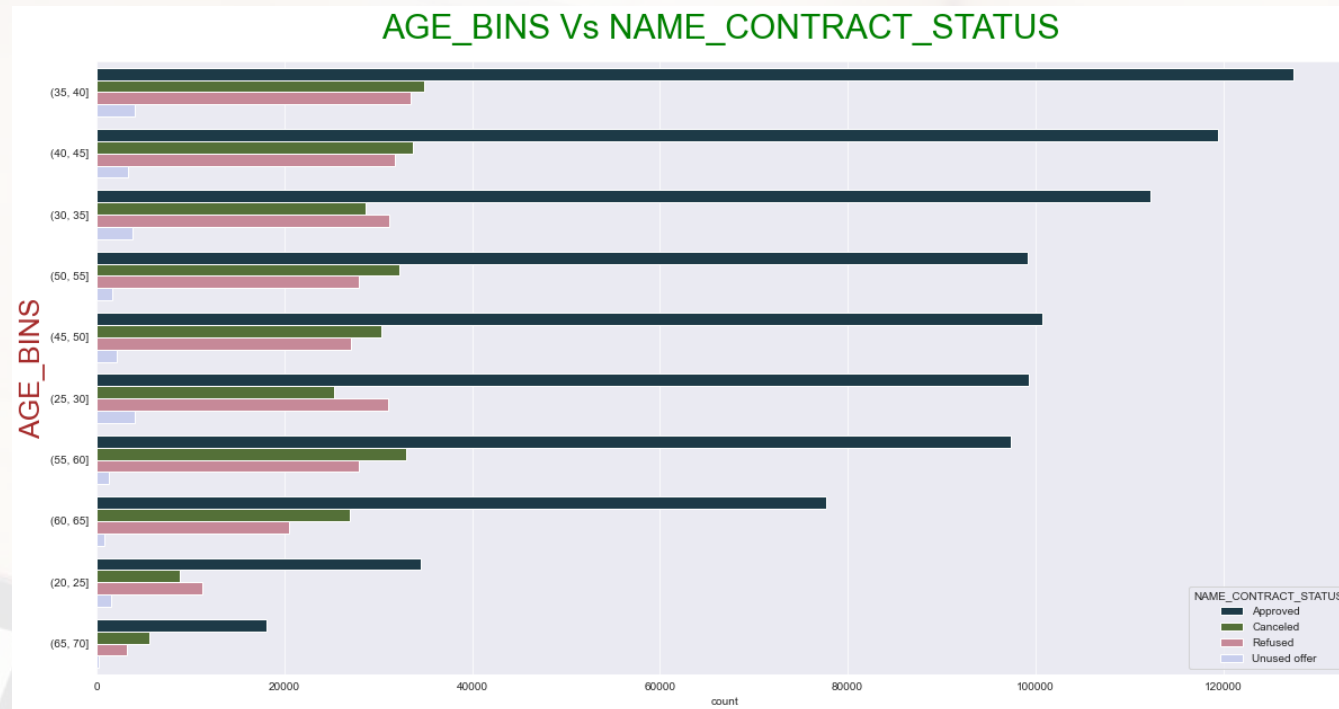
CONTRACT STATUS VS GENDER



Inference

- It is observed that Female clients are more successful in terms of having their loans approved.
- However male clients are not so successful and do see an increase in the number of times their loans get refused.

CONTRACT STATUS VS AGE GROUPS



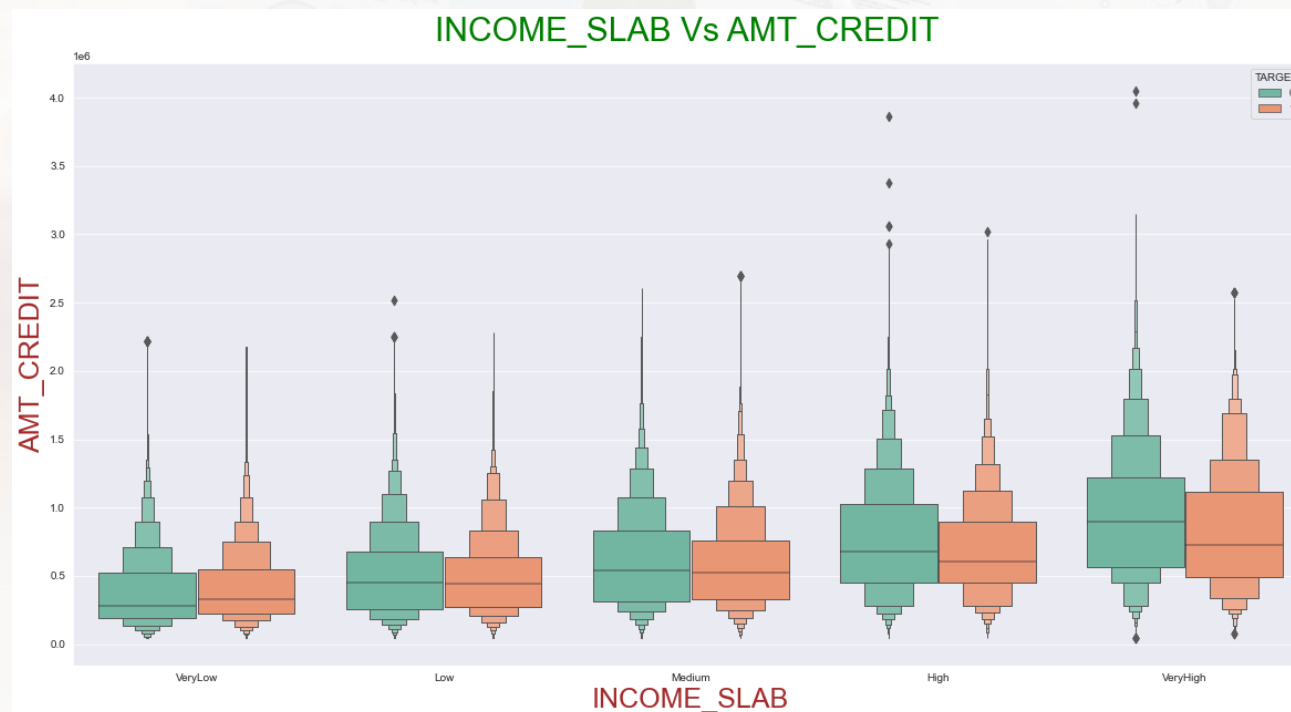
Inference

- In the age category of 20 to 35 years of age, we see a lot of rejection of loan. These group of people are also more likely to default as per our previous inferences and conclusions.
- Age group of people above 45 are less likely to default and also they see less rejection and cancellation of loan amounts.

DEFAULTERS BY INCOME RANGE

Inference

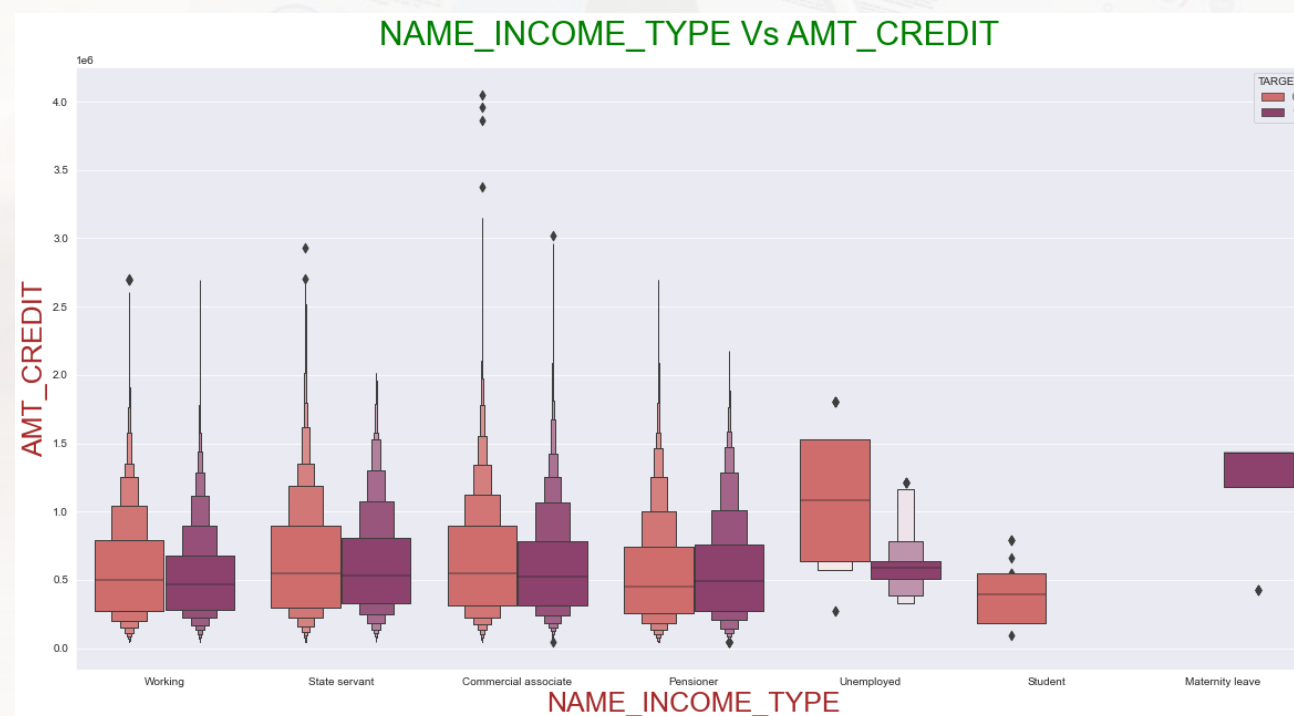
- We can see that as the income slab increases, the credit amount of the loan also increases.
- We can also see that the people in the low and very low income slabs are likely to default more.



DEFAULTERS BY INCOME TYPE

Inference

- People involved in business and unemployed people ask for more loan and repay better.



CONCLUSION AND RECOMMENDATIONS :

The application dataset and the previous application dataset were analyzed, cleaned and inferences/correlations were drawn. We have thoroughly observed these datasets and here are our observations and comments about the same:

- ✓ Banks can give away loans to Students, pensioners and people with higher education degrees, as they are very less likely to default loan payments.
- ✓ We understood that, Laborers, Sales staff, drivers, cleaning staff, low-skill labors are more likely to default a payment of the loan.
- ✓ The best clients to target in this case would be Managers, core staff, high skill tech staff.

- ✓ **People in the age group of 20 to 30 are more likely to default. People above the age group of 40 do not default on their payments as much.**
- ✓ **It was also observed that, people belonging to low and very low income slabs were showing strong indicating signs of defaulting.**
- ✓ **We also observed that people who live in a place which is not so populated, like village or small towns, have difficulty in repaying loan amount.**
- ✓ **Clients who are more likely to default loans are more likely to change their registration, few days prior to applying for the loan.**
- ✓ **Keeping these points in mind, if a customer can be evaluated based on the above parameters, the bank would see less default payments**

**THANK
YOU**

