CREDIT EDA ASSIGNMENT

upGrad

#LifeKoKaroLift

By:

Sanskruti D. Babar

Problem Statement

INTRODUCTION:

This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Problem Statement

BUSINESS UNDERSTANDING

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.
- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
- 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.
- The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios
- The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
- All other cases: All other cases when the payment is paid on time.
- When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

Approved: The Company has approved loan Application

Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want.

Refused: The company had rejected the loan (because the client does not meet their requirements etc.).

Unused offer: Loan has been cancelled by the client but at different stages of the process.

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency to default.

Problem Statement

BUSINESS OBJECTIVE

- This case study aims to 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.
- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.
- To develop your understanding of the domain, you are advised to independently research a little about risk analytics – understanding the types of variables and their significance should be enough.

Approach and Methodology

1. Checking and Reading the data set:

a. We load the data set provided for the analysis and use python library called ‘pandas’ to import the data set into a data frame for analysis.

b. We check the data frame using attributes such as head, shape, info, data types of all the columns, describe.

2. Data Correction and Cleaning:

a. In this step, we clean and correct the data frame of unwanted, unnecessary and irrelevant data.

b. This includes handling missing values through imputation and deletion, dealing with outliers, ensuring that each column has the correct data type, and that there are no duplicate records.

Approach and Methodology

3. Exploratory Data Analysis (EDA):

a. This step includes understanding the relationship between variables and data distribution.

b. This can be done through imbalance percentage, univariate analysis, bivariate analysis and multivariate analysis.

4. Identify key variables:

a. In this step, we determine which variables are strong indicators of loan defaulters.

b. This can done through correlation analysis, where variables with high correlation to the target variable are identified.

Approach and Methodology

6. Data Visualization:

a. This step includes visualizing data patterns and relationships to gain insights.

b. This can be done through histograms, box plots, scatter plots, count plots, heat maps and bar plots.

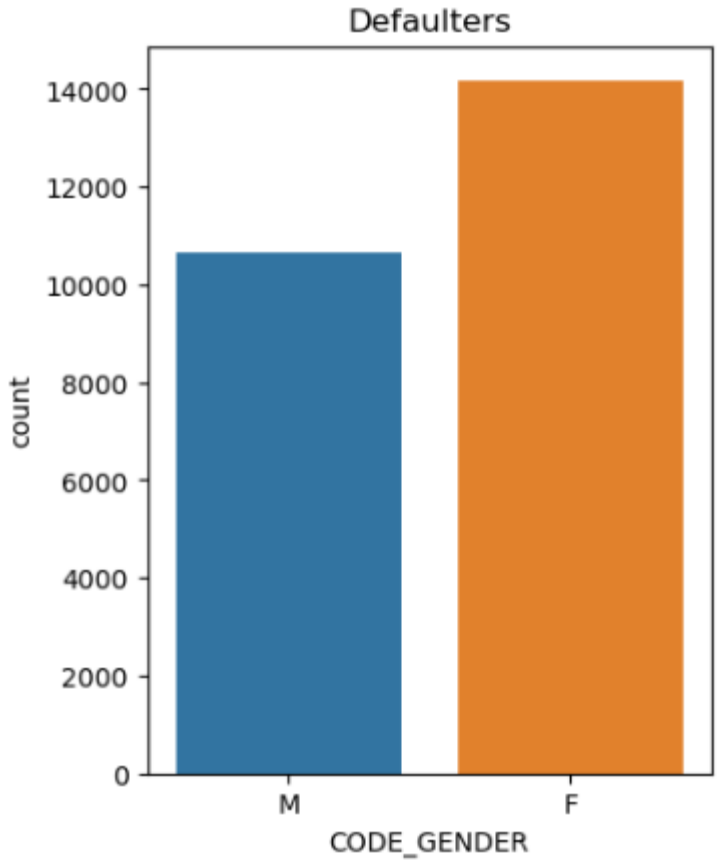
Graphs and Insights

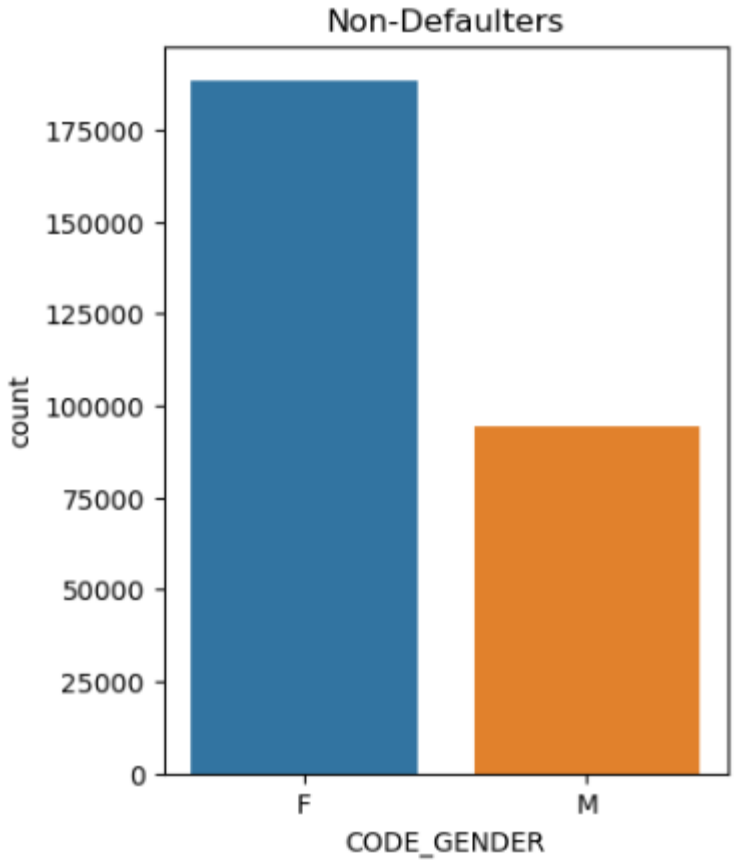
- **TARGET vs CODE_GENDER**

The above count plots indicate that there are around 14000 female defaulters whereas there are over 175000 female non-defaulters as compared to the male defaulters at around 11000 and male non-defaulters at around 100000 approximately. Therefore the percentage calculated for the above values are:

- percentage for male defaulters: (11000/105059)*100 = 10.7%
- percentage for female defaulters: (14000/202452)*100 = 6.9%
- percentage for male non-defaulters: (100000/105059)*100 = 89.3%
- percentage for female non-defaulters: 100-6.9 = 93.1%

This indicates that males have a higher percentage of defaulting loans than females


Defaulters
14000
12000
10000
8000
6000
4000
2000
0
count
M
F
CODE_GENDER

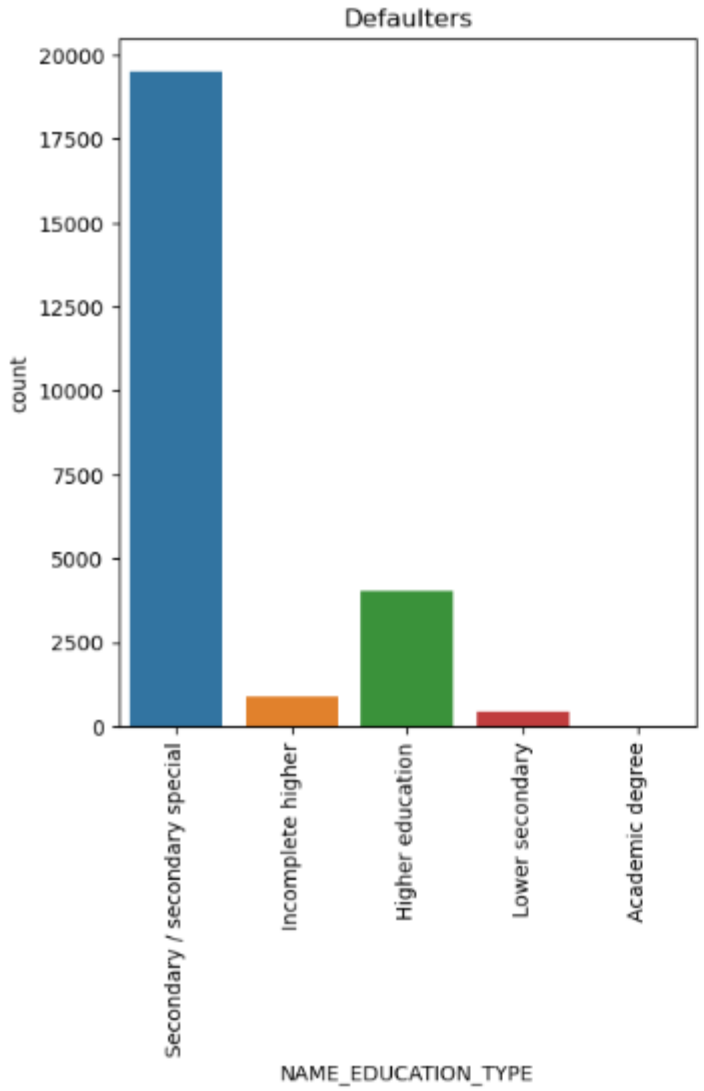

Non-Defaulters
175000
150000
125000
100000
75000
50000
25000
0
count
F
M
CODE_GENDER

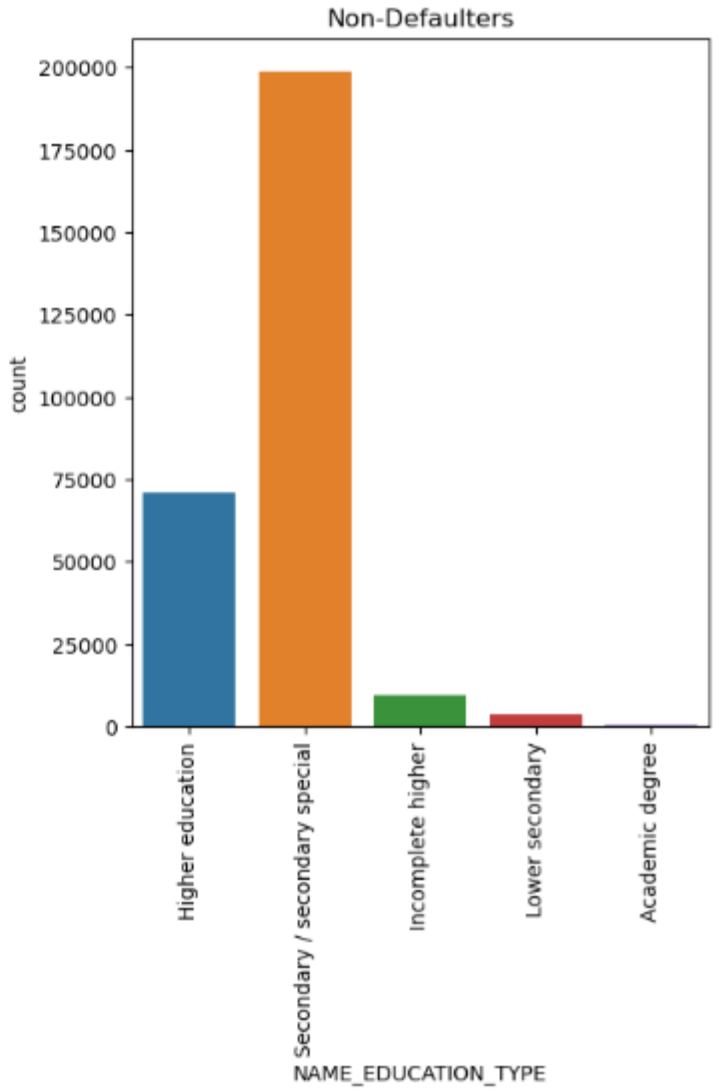
Graphs and Insights

- **TARGET vs EDUCATION_TYPE**

- From the above graphs, we observe that the max defaulters are in the education category of Secondary/Secondary Special (around 20000). This leads us to the result that people who have less education have low income which leads them take loans and then have trouble paying for them.
- For the non-defaulters the education category is also Secondary/Secondary Special (around 20000) it contradicts what the first graph tells us.

The analysis is for education type is very close for defaulters and non-defaulters

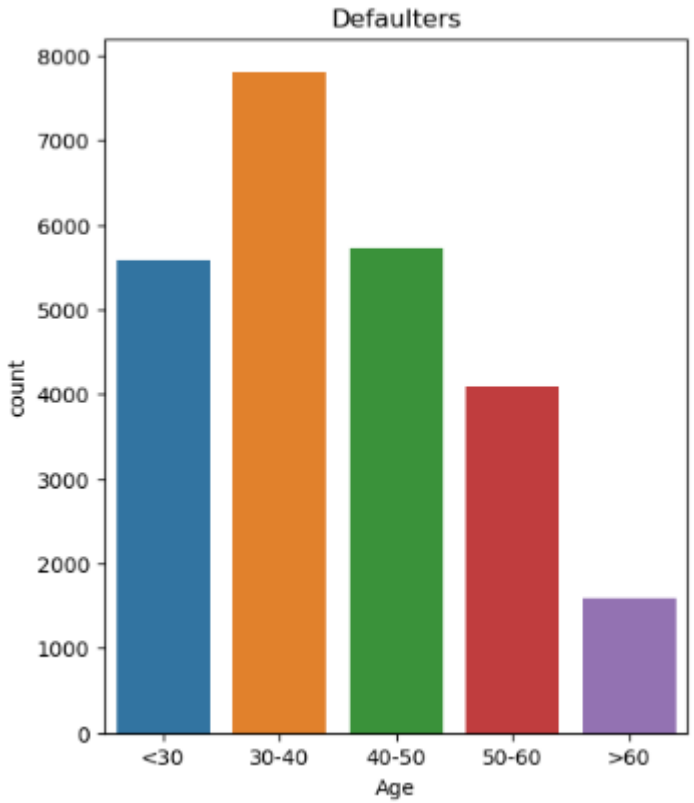

Defaulters
20000
17500
15000
12500
10000
7500
5000
2500
0
count
Secondary / secondary special
Incomplete higher
Higher education
Lower secondary
Academic degree
NAME_EDUCATION_TYPE

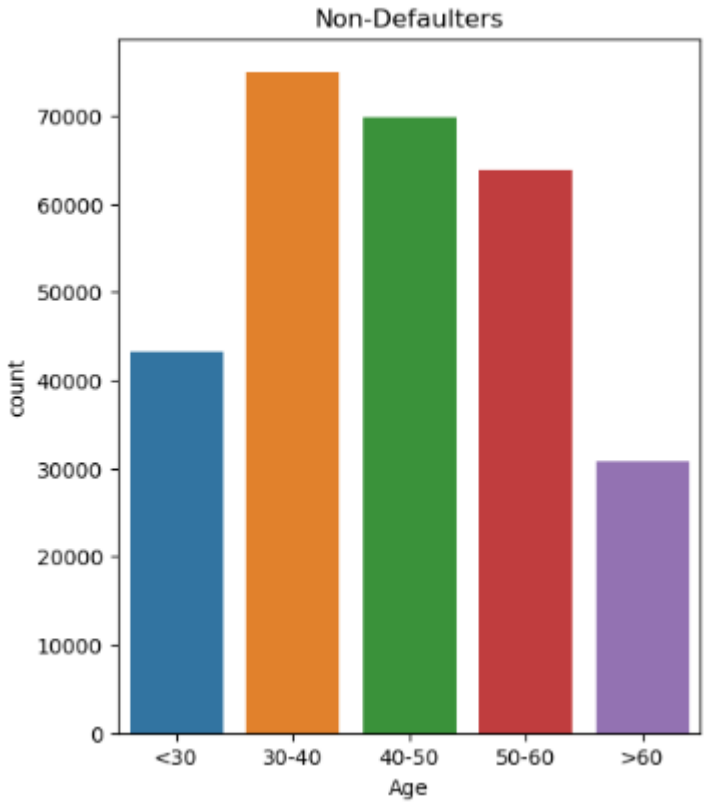

Non-Defaulters
200000
175000
150000
125000
100000
75000
50000
25000
0
count
Higher education
Secondary / secondary special
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE

Graphs and Insights

- **TARGET vs Age**

- From the above graphs, we observe that the age group 30-40 has the maximum number of people for both defaulters and non-defaulters.
- The second largest age group that follows in both defaulter and non-defaulter is the 40-50 age group.
- The age category for old age has the least amount of defaulters and non-defaulters.

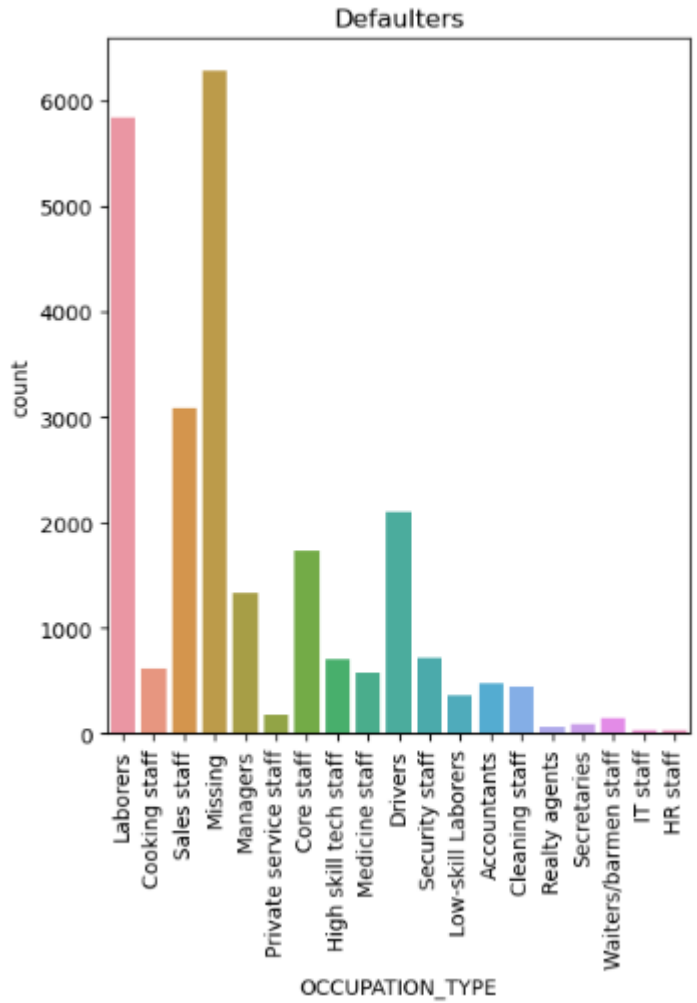


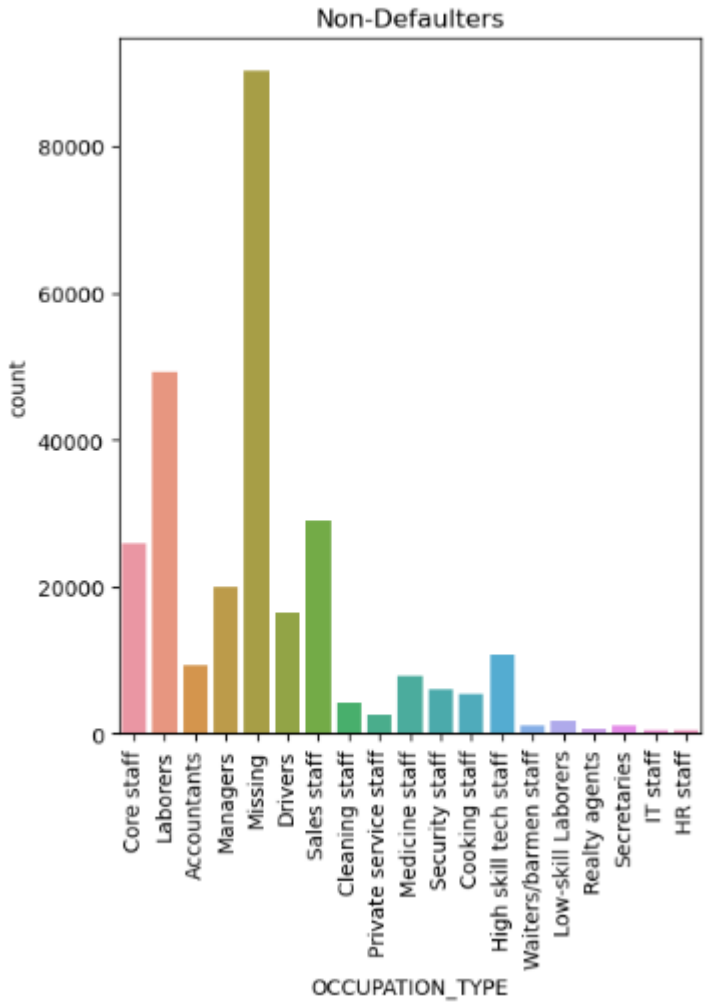


Graphs and Insights

- **TARGET vs OCCUPATION_TYPE**

The above plot indicates that Laborers are the most defaulters and non-defaulters

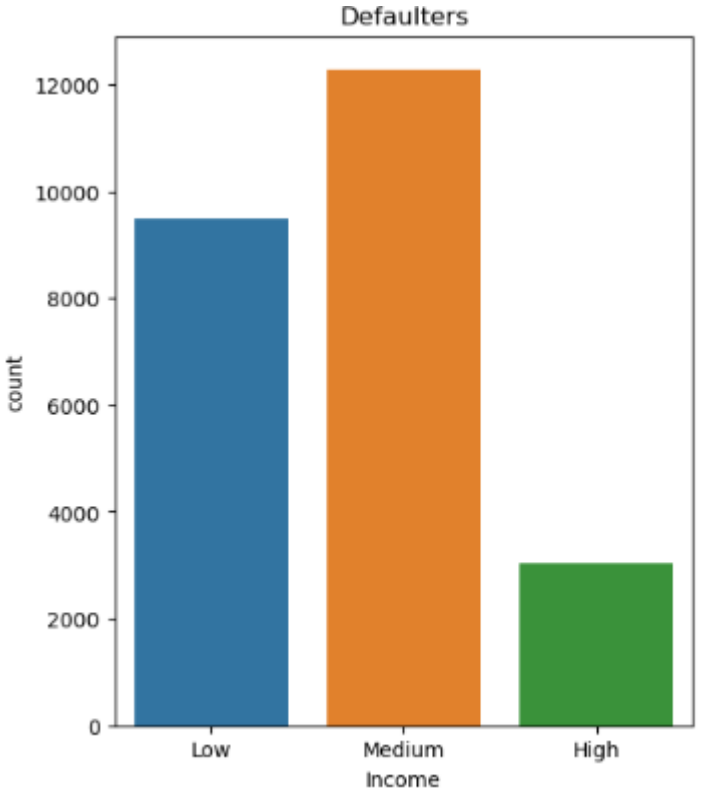


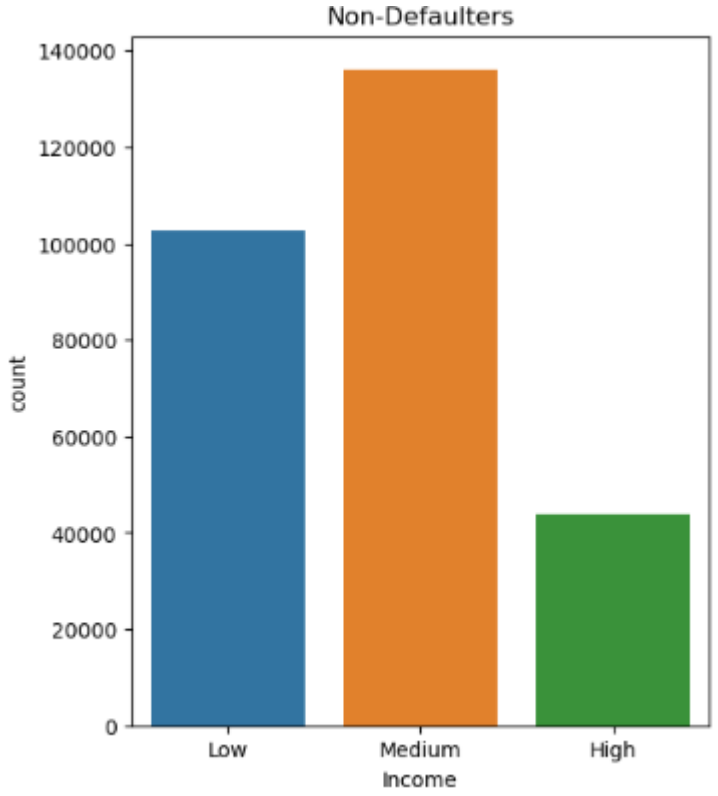


Graphs and Insights

- **TARGET vs Income**

The plot indicates that those with Medium income are the maximum for both defaulters and non-defaulters, followed closely by those with Low income. The plot for both Defaulters and Non-defaulters are very closely similar.

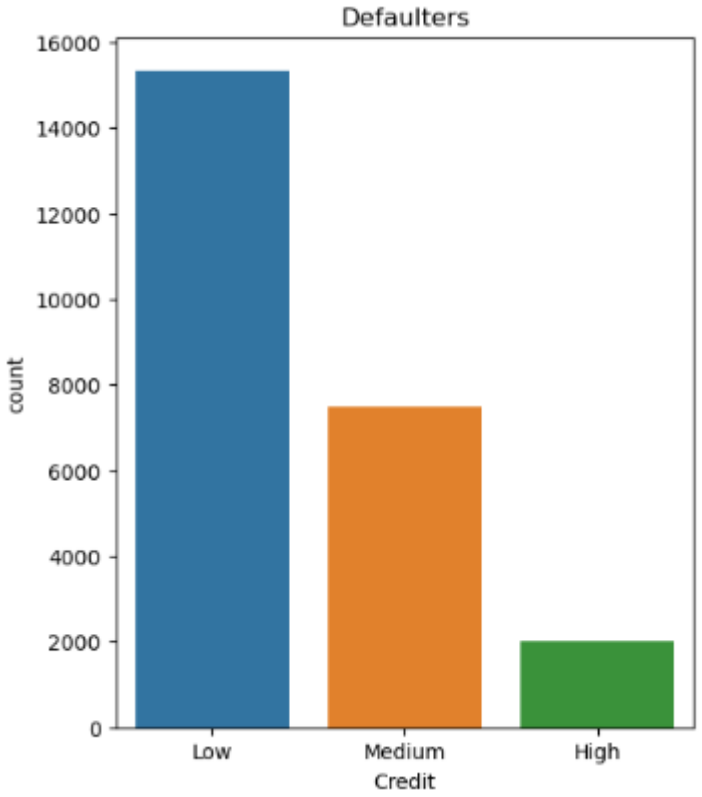


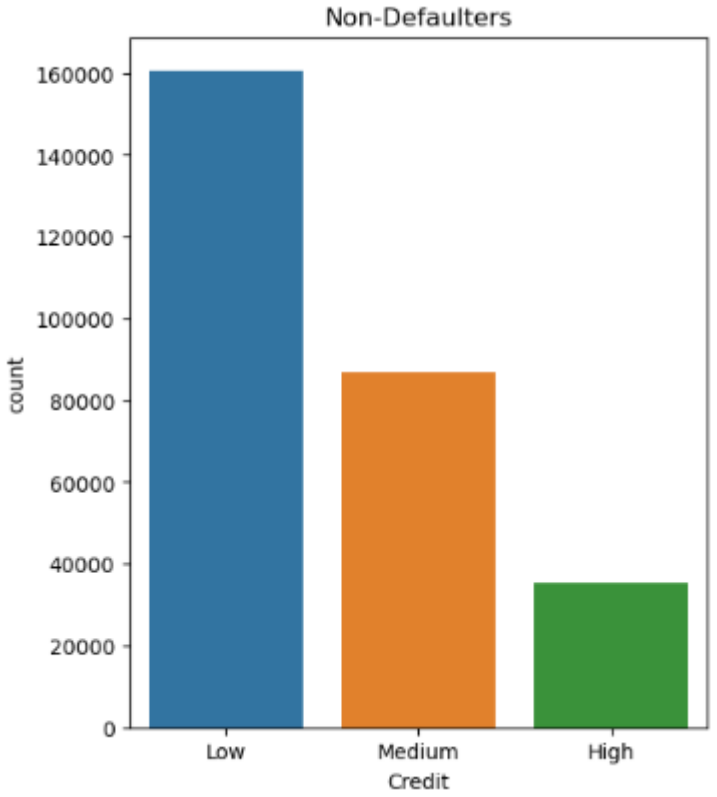


Graphs and Insights

- **TARGET vs Credit**

The plot indicates that people who have low credit amount are maximum in both defaulters and non-defaulters.

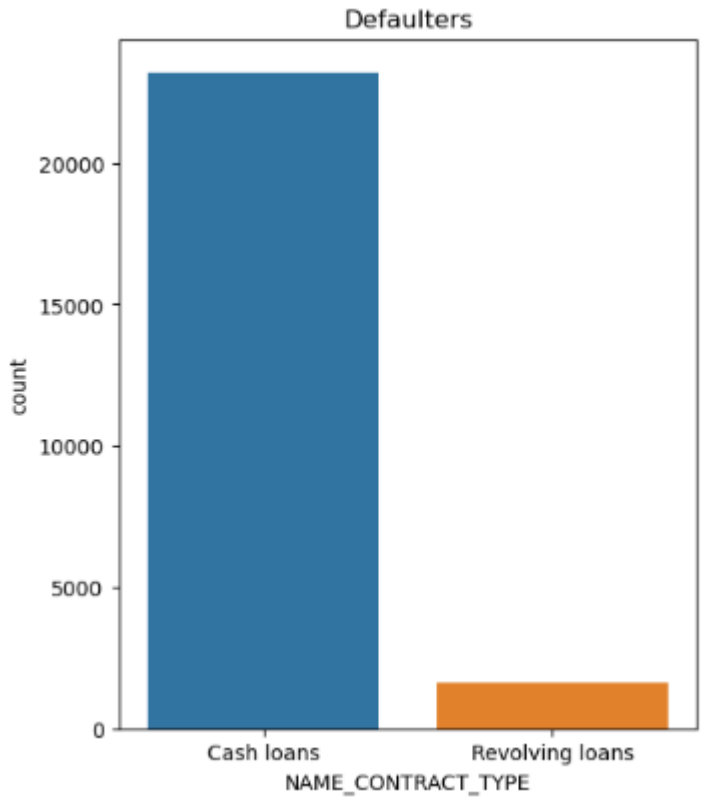


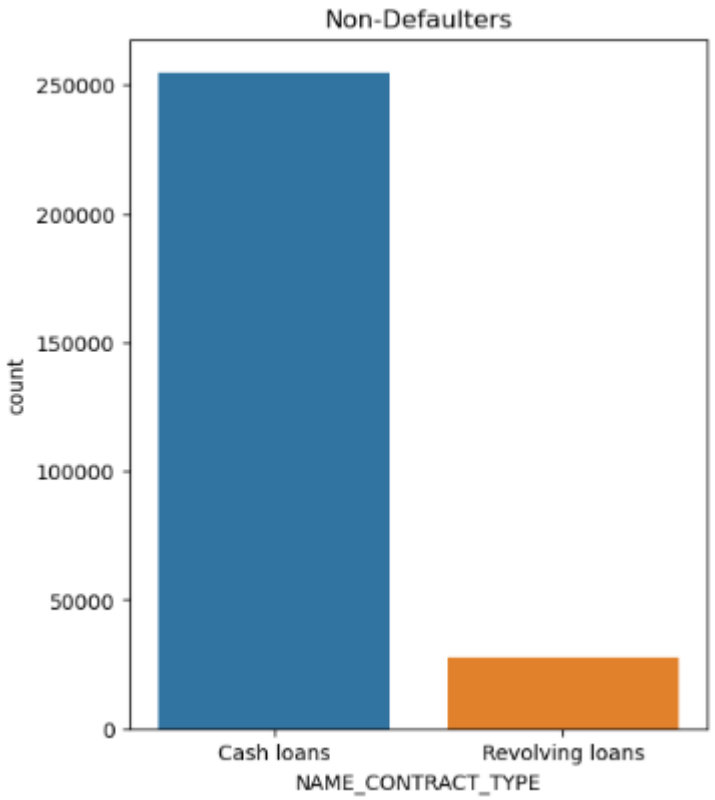


Graphs and Insights

- **TARGET vs Contract type**

The plot indicates that both defaulters and non-defaulters have taken cash loans



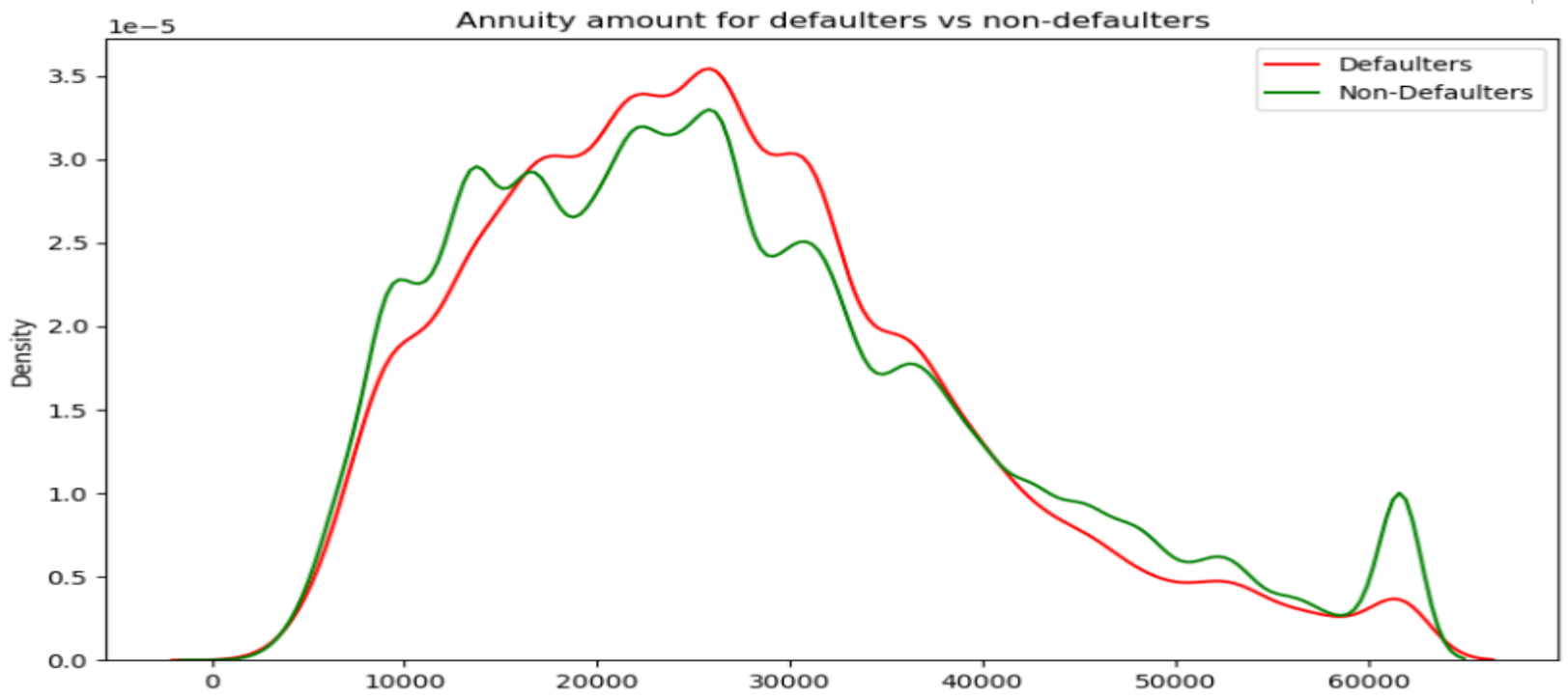


Graphs and Insights

- **Annuity amount for defaulters vs non-defaulters**

The above plot indicates that:

- there is a rapid rise of the spread of annuity between 150000 to 25000, and it further gradually decreases over increase off annuity value.
- The most defaulters of the loan payment thus come from 15000 to 30000. For the non-defaulters on the other hand, there is a peak in the spread between annuity values of 15000 to 30000, and the people who lie in this range are non-defaulters.

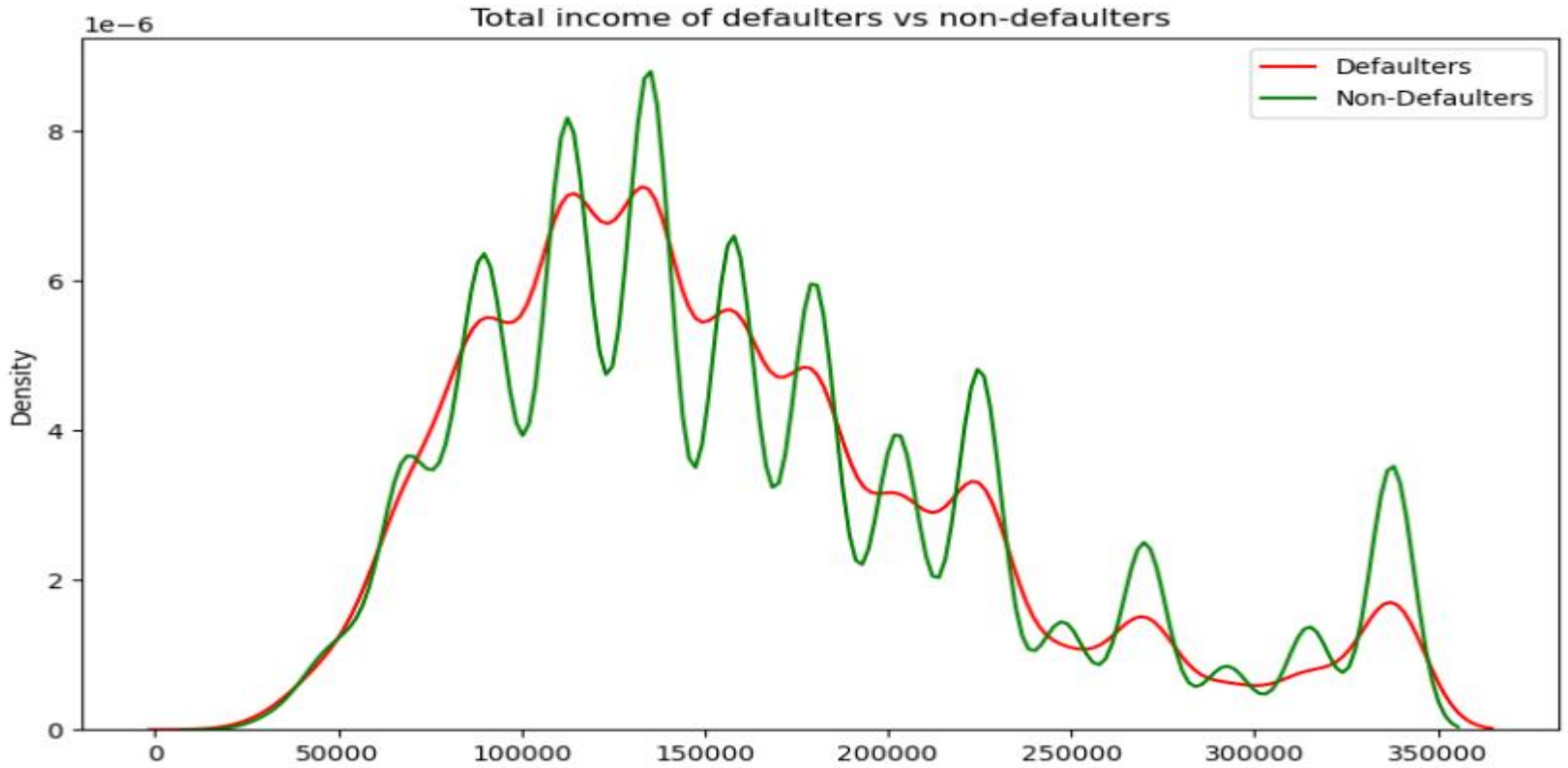


Graphs and Insights

- **Total income of defaulters vs non-defaulters**

The above plot indicates that:

1. the maximum defaulters are those people whose income lies in the range of 1 to 1.5 lakhs. And as the income increases the number o defaulters decreases.
2. on the other hand for non-defaulters, there is a mixed tread. People with either high or low income are non-defaulters, there is not set group for it.

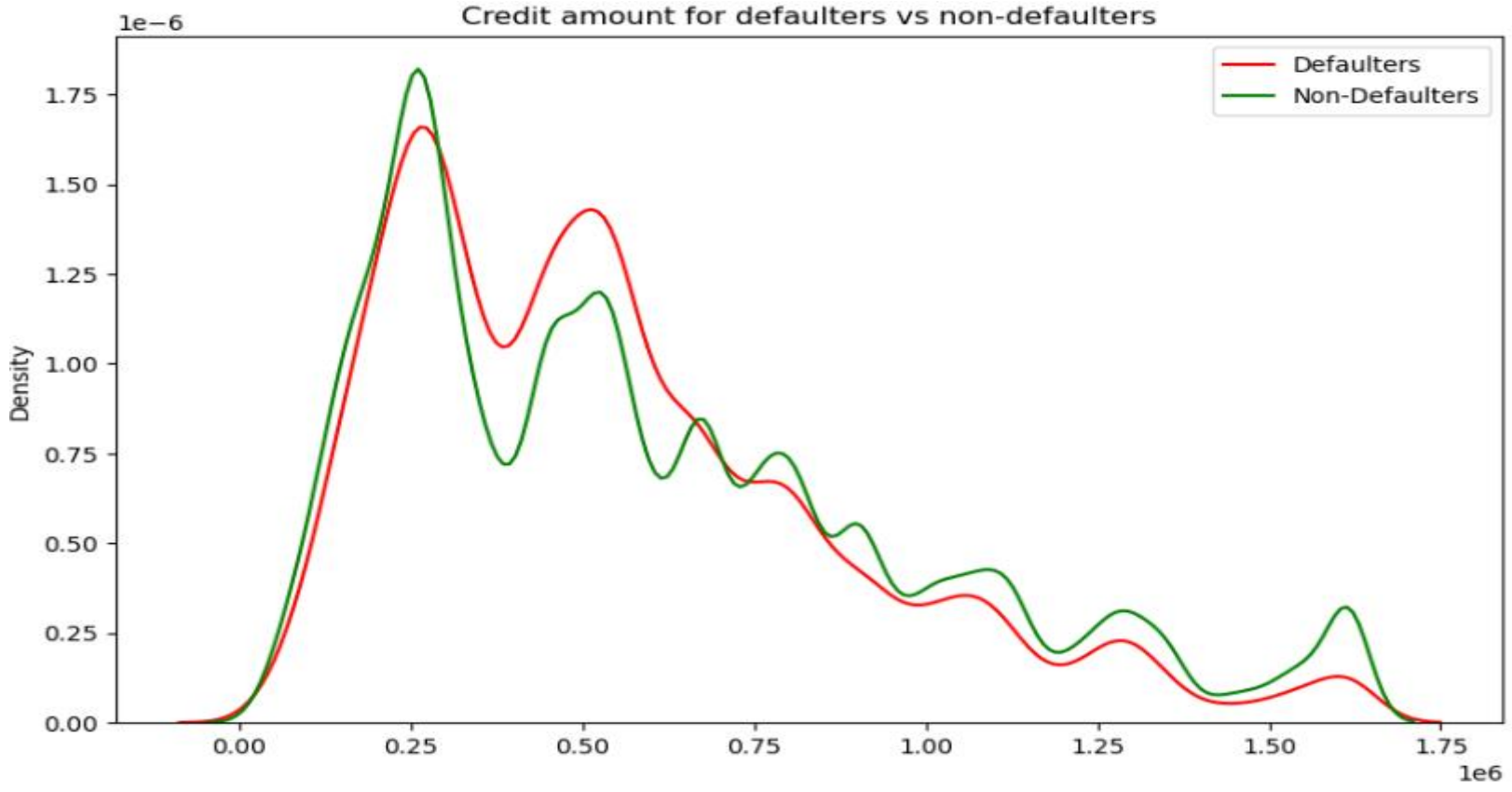


Graphs and Insights

- **Credit amount for defaulters vs non-defaulters**

The above plot indicates that:

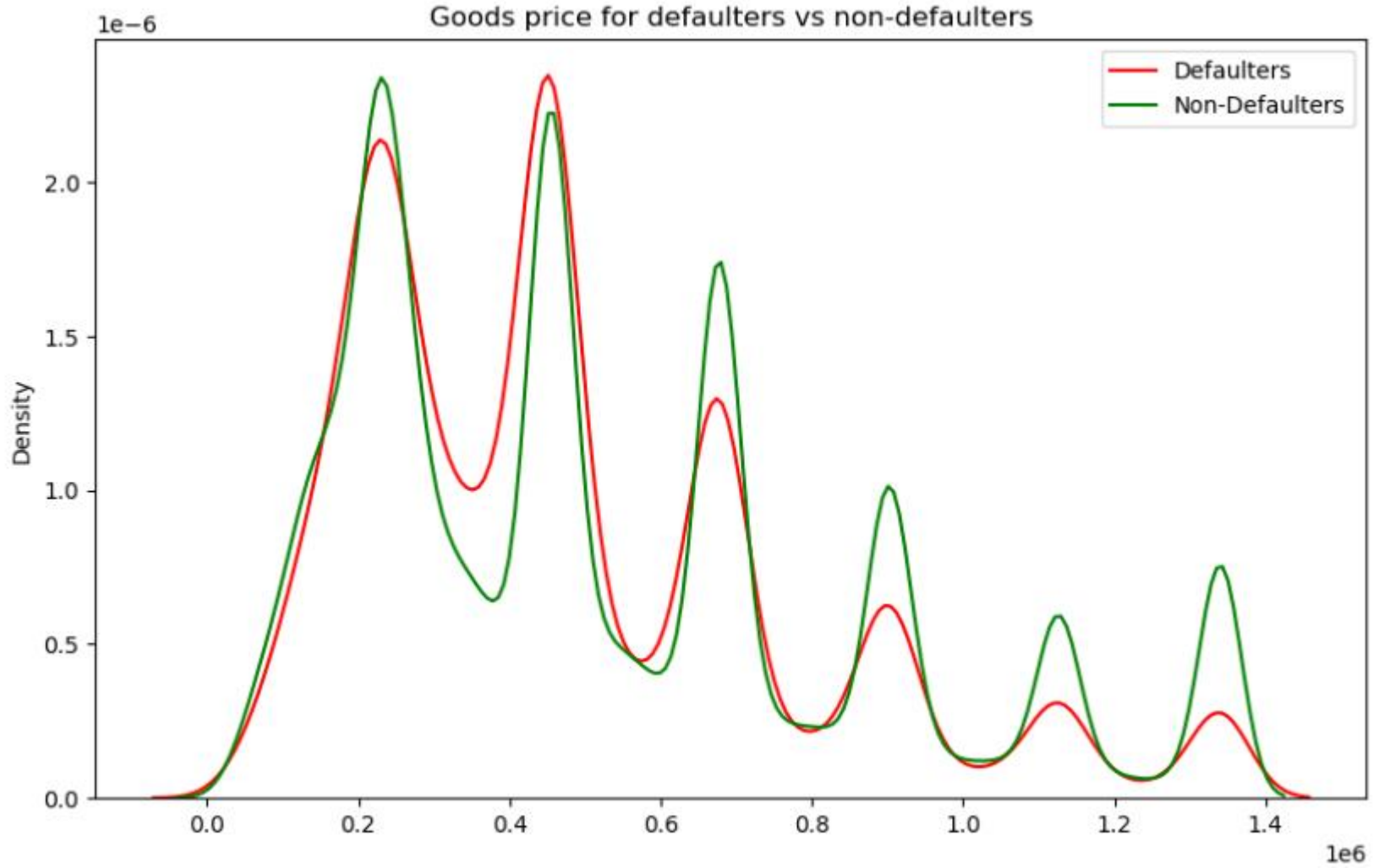
- credit value peaks at 250000 and 500000 for defaulters>
- credit value peaks at 250000 and 500000 and still keeps having rise and fall of the peaks at 750000, 1250000 and then gradually decreasing for non-defaulters.



Graphs and Insights

- **Goods price for defaulters vs non-defaulters**

The plot indicates that both defaulters and non-defaulters show a similar trend when comparison is made on the basis the goods price.



Graphs and Insights

- **Bivariate analysis between income type and income total**

The plot indicates that: For defaulters:

The income bracket for Commercial associates is the highest followed by State servants, Working class, Pensioner, Unemployed and Maternity leave. For non-defaulters:

The income bracket for Businessmen is the highest followed by Maternity leave, Commercial associates, State servants, Working class, Students, Unemployed and finally Pensioners.

The conclusion is that defaulters do not businessmen and student as its categories. They are trusted by banks to pay their loans.

Graphs and Insights

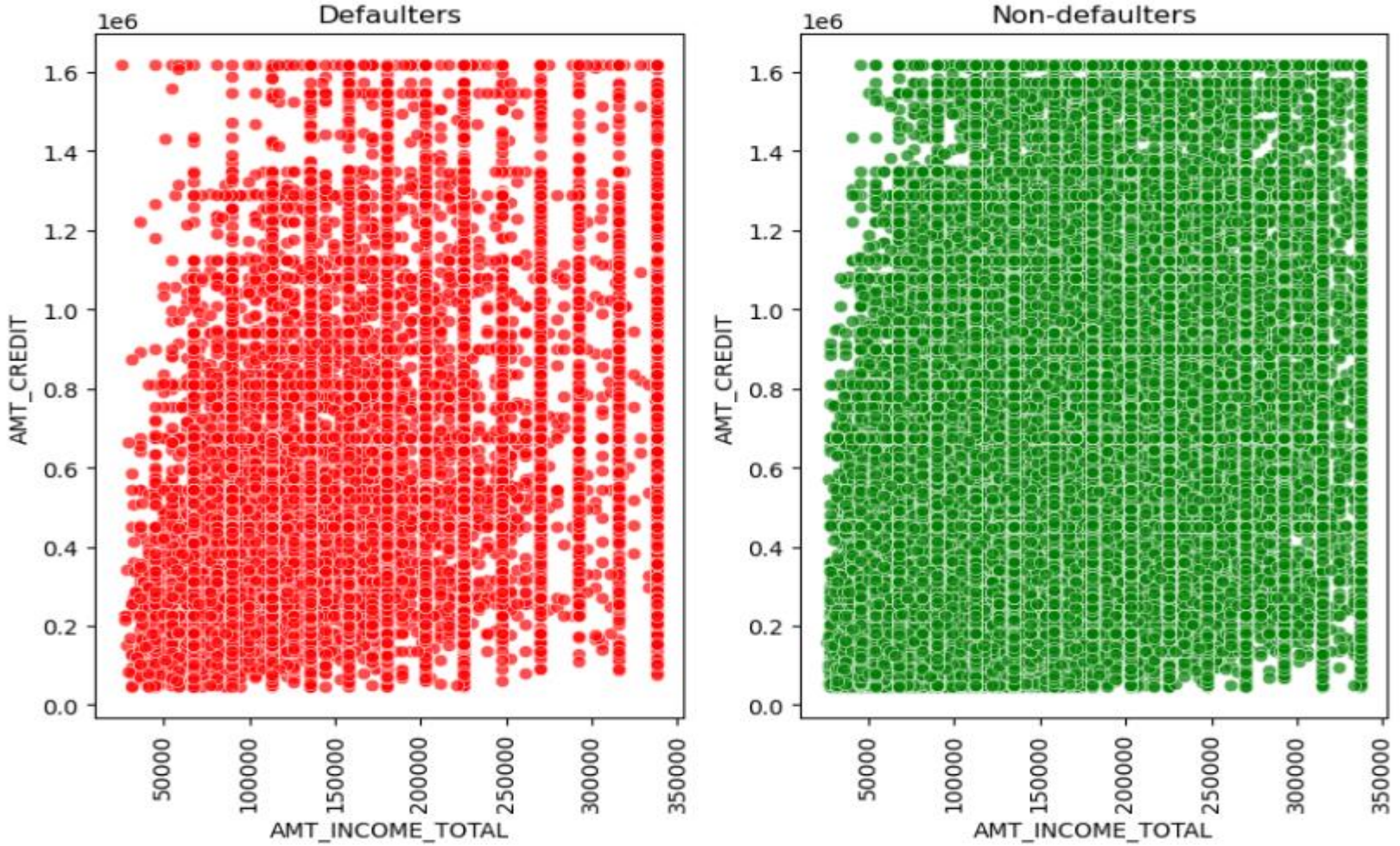
- **Correlation between Total income and credit amount for defaulters and non-defaulters:**

The above scatterplot indicates that:

- for defaulters, the correlation between income and credit is not good, there is a positive relation between them, but no linear relation.
- for non-defaulters, the correlation

between income and credit is not

linear

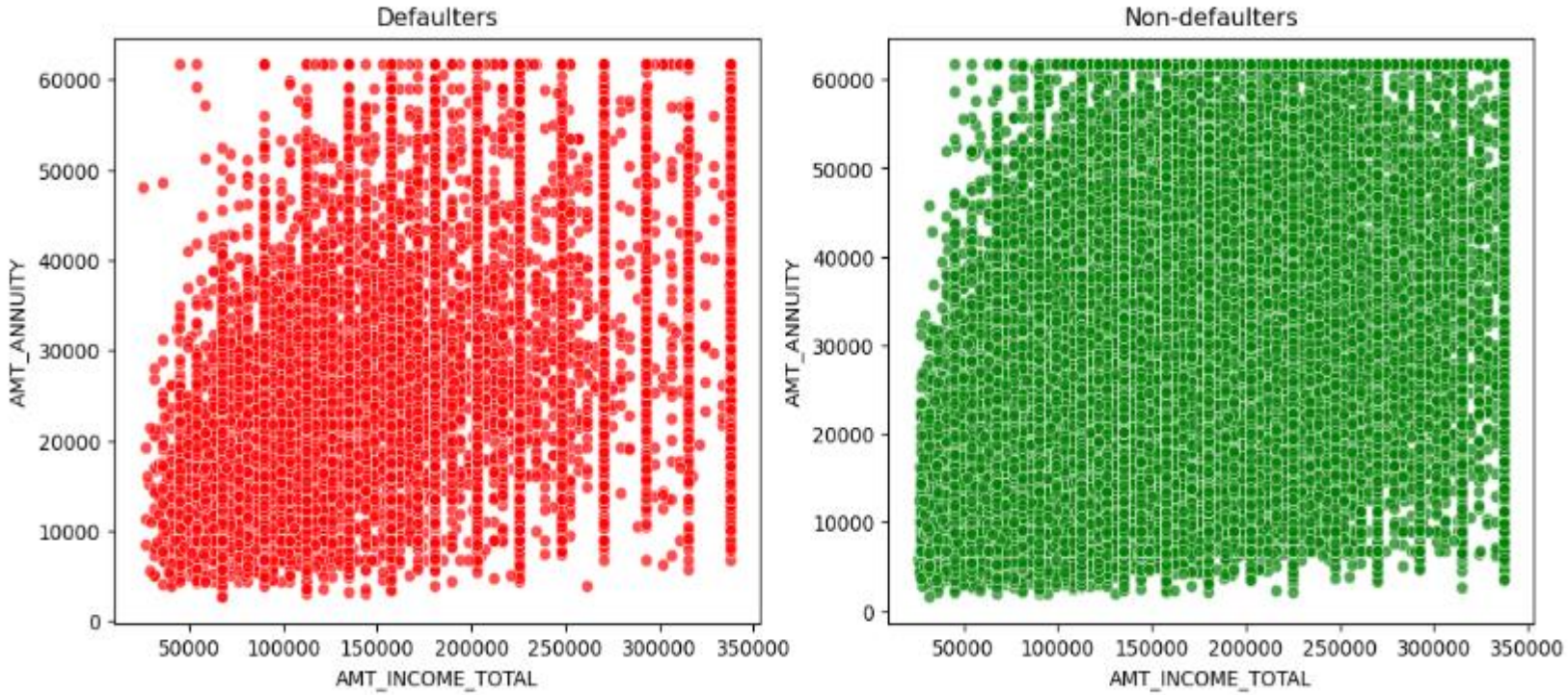


Graphs and Insights

- **Correlation between Annuity and Total income for defaulters and non-defaulters:**

The above plot indicates that:

- For defaulters, the correlation between annuity and total income is not that great, it is positive but there is no linear relation between them.
- For non-defaulters, the correlation between annuity and total income does not have any linear relationship

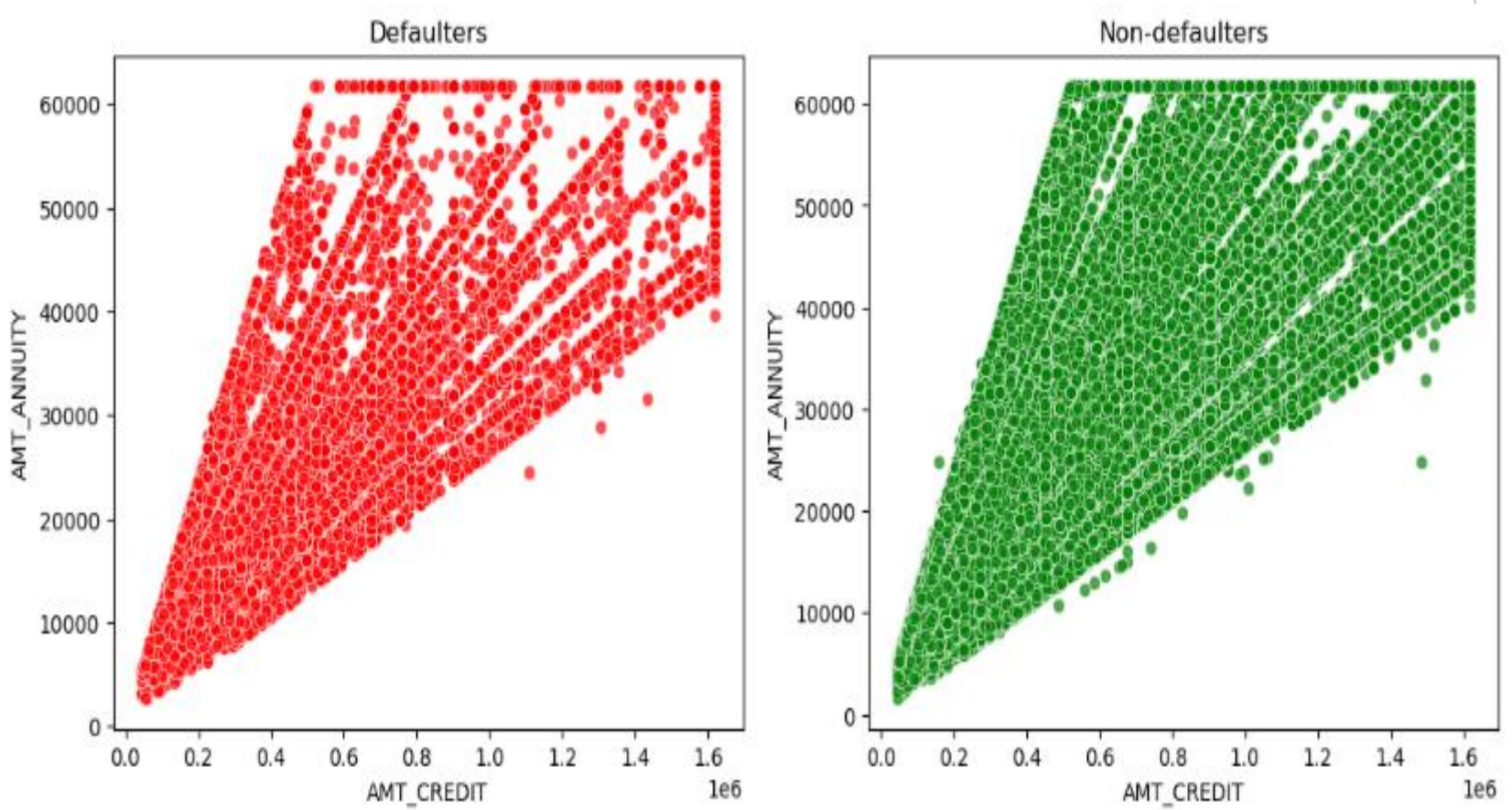


Graphs and Insights

- **Correlation between Annuity and Credit amount for defaulters and non-defaulters:**

The above plots indicates that:

- The correlation between annuity and credit amount is positive and it have a linear relation between the two variables for both defaulters and non-defaulters category.

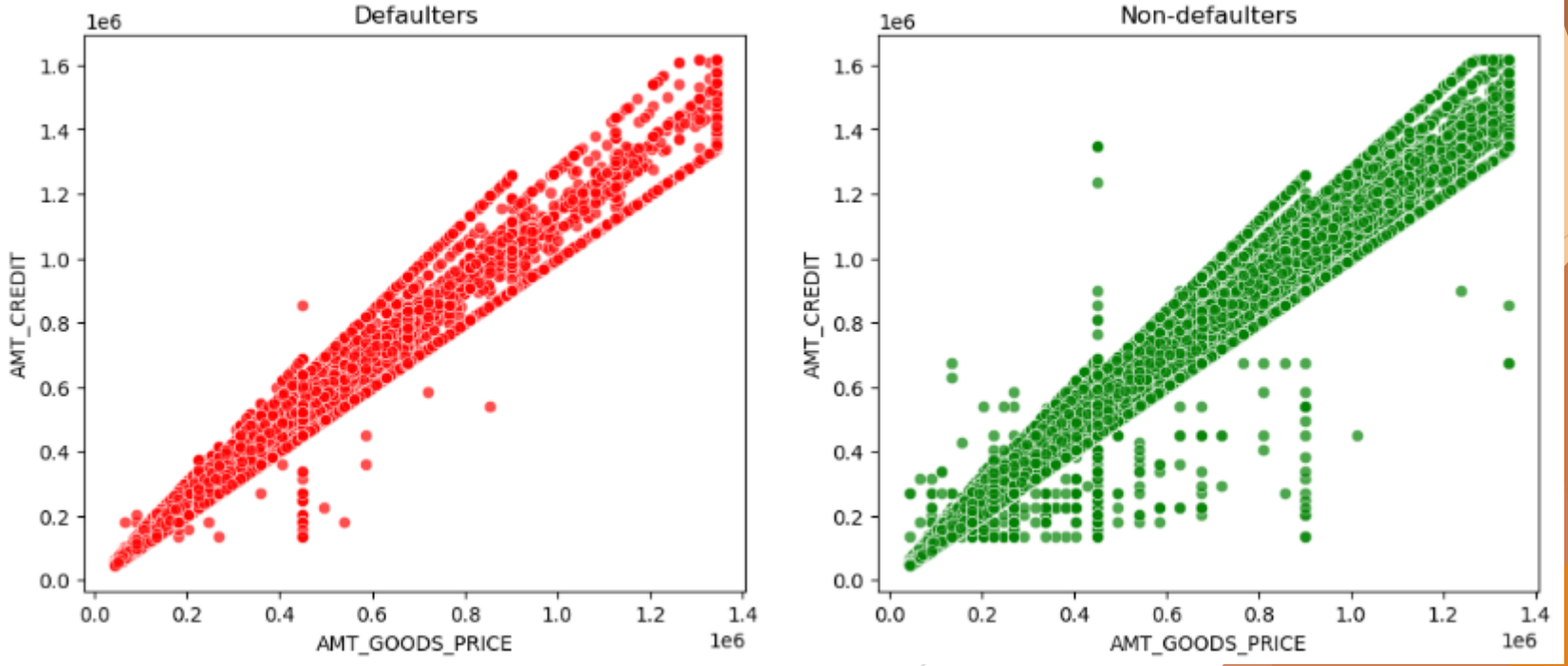


Graphs and Insights

- **Correlation between Goods price and Credit amount for defaulters and non-defaulters:**

The above plot indicates that:

- There is a very high correlation between credit amount and goods price under defaulters category, it is a positive and linear correlation. For non-defaulters, though there is high correlation between credit amount and goods price, it amounts to 97-98% approximately.
- There is causation between credit amount and goods price, i.e., when goods price increase, so does the credit amount.



Graphs and Insights

- **Bivariate analysis between Occupation type and Credit amount**

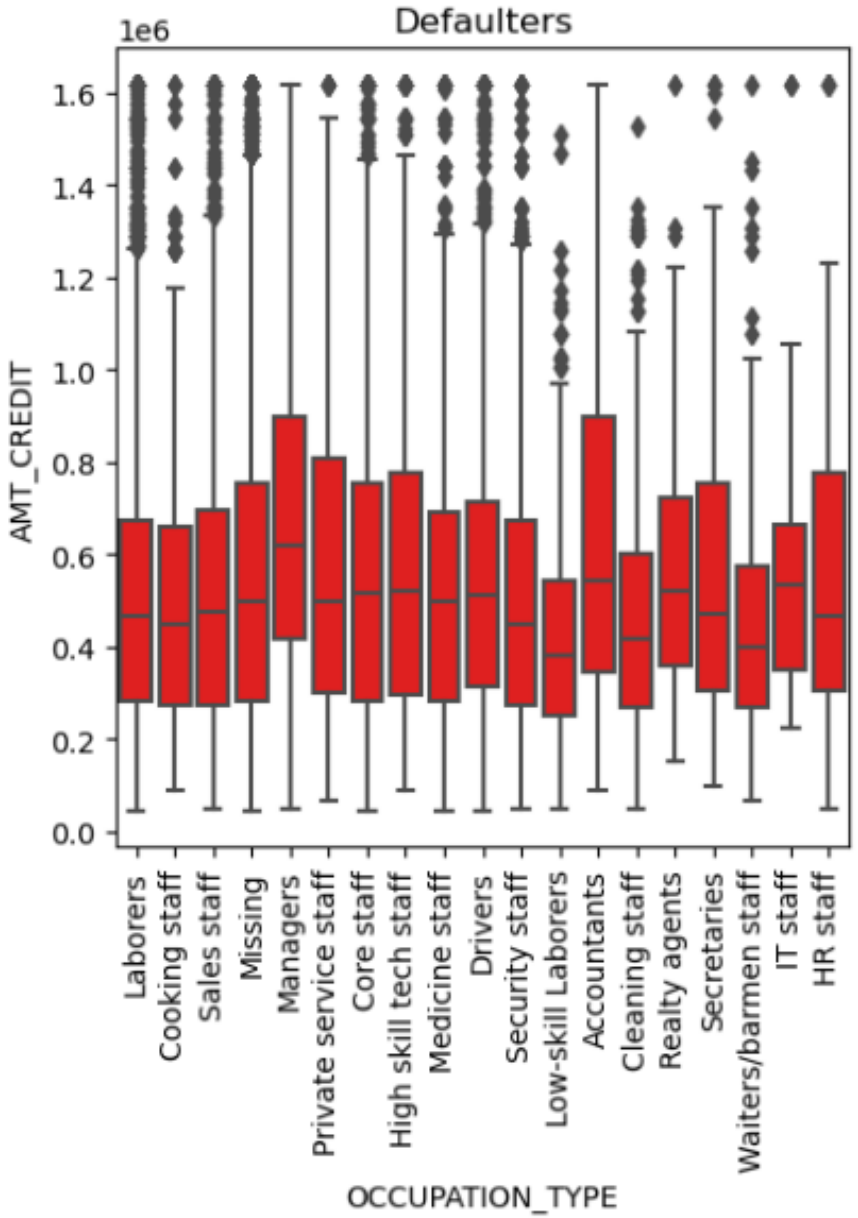
The above plot indicates that:

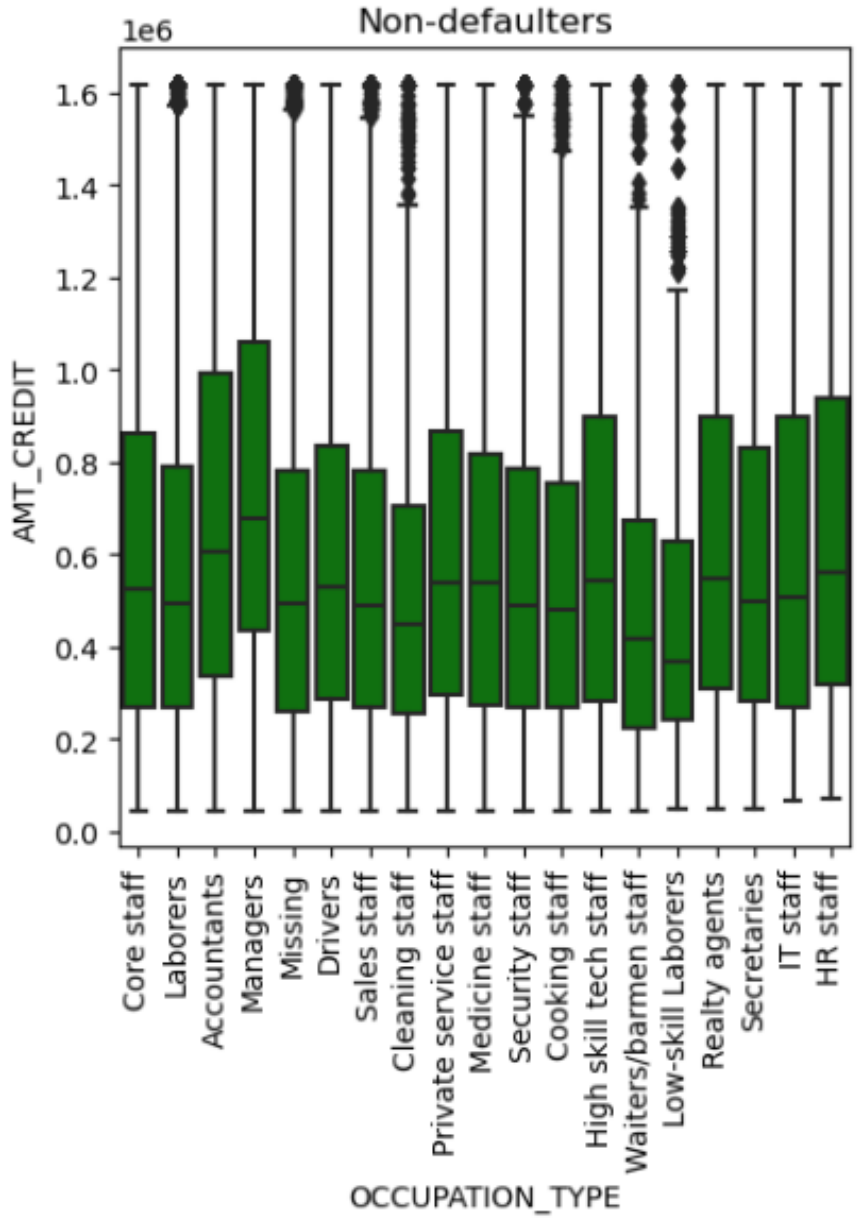
1. For Defaulters:

- Managers and Accountants have more credit than other occupation types, and on top of that they are the defaulters of loan too. Banks should be more cautious when accepting loans for these two categories.\
- Managers and Accountants are followed by Private service staff->High skill tech staff->HR staff->Secretaries and finally Drivers. Banks should also be cautious and careful with these categories.
- The others, while also being defaulters have less credit with them.

2. Non-defaulters:

- The boxplot for non-defaulters indicates that Core staff, Accountants, Managers, Drivers, Private service staff, Medicine staff, High skill staff, Realty agents and Secretaries are the ones who have taken maximum loans and paid them back on time too.


Defaulters
1e6
AMT_CREDIT
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Laborers
Cooking staff
Sales staff
Missing
Managers
Private service staff
Core staff
High skill tech staff
Medicine staff
Drivers
Security staff
Low-skill Laborers
Accountants
Cleaning staff
Realty agents
Secretaries
Waiters/barmen staff
IT staff
HR staff
OCCUPATION_TYPE


Non-defaulters
1e6
AMT_CREDIT
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Core staff
Laborers
Accountants
Managers
Missing
Drivers
Sales staff
Cleaning staff
Private service staff
Medicine staff
Security staff
Cooking staff
High skill tech staff
Waiters/barmen staff
Low-skill Laborers
Realty agents
Secretaries
IT staff
HR staff
OCCUPATION_TYPE

Graphs and Insights

- **Bivariate analysis between annuity and income type**

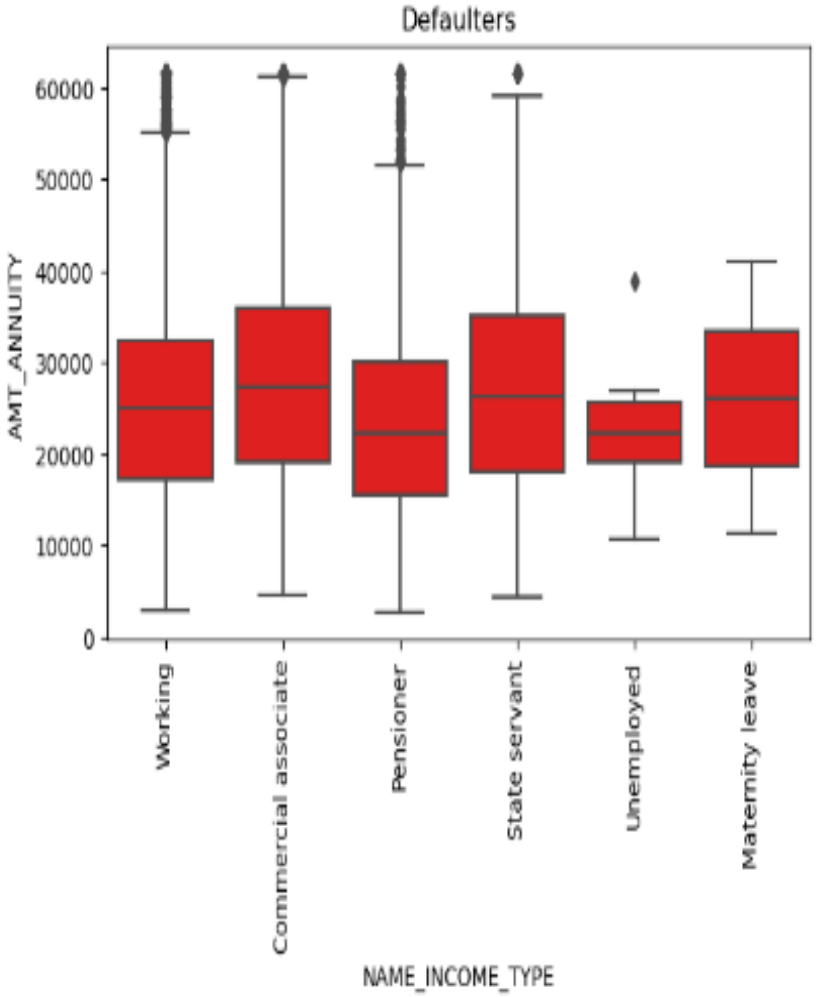
The above plot indicates that:

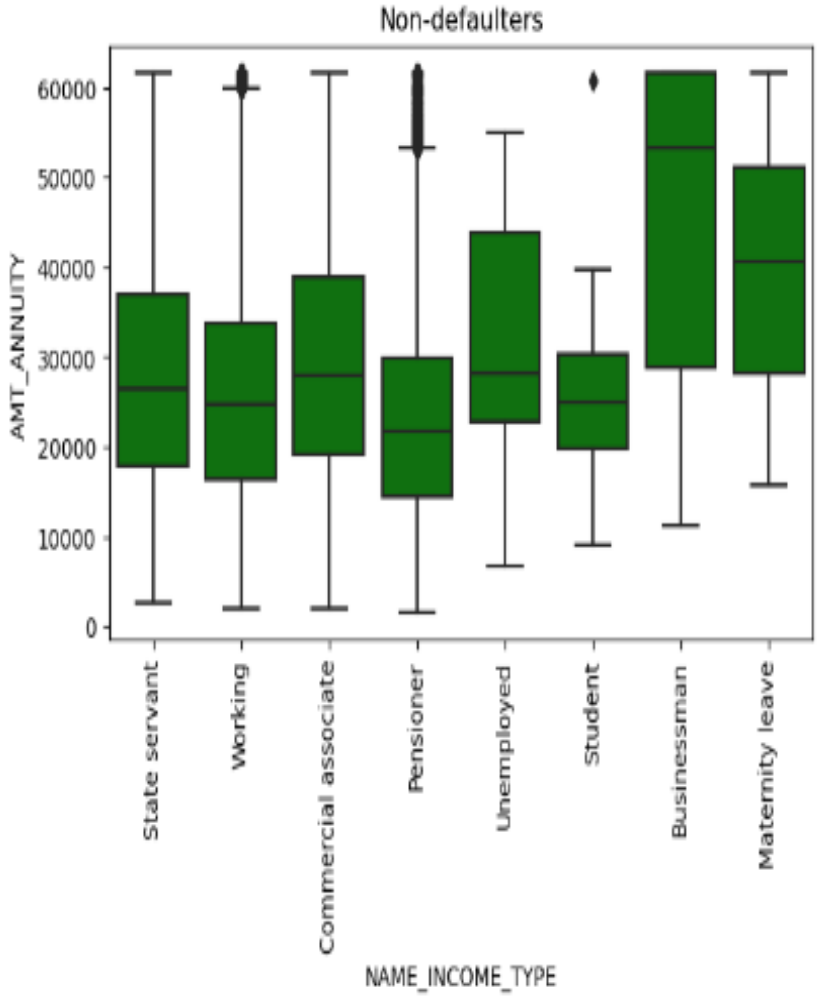
1. For defaulters:

- In the plot, Commercial associate and State servant have similar amount of Annuity. These two categories have high amount of annuity, around 35000, with a median value of 25000.
- Maternity leave also has high amount of annuity, around 350000.
- The annuity of unemployed is the least among all categories.
- Pensioners comprised of old people also have high annuity.
- Working category comes under the middle bracket, having an annuity of around 30000 and a median value being slightly above that of Pensioners.

2. For non-defaulters:

- In non-defaulters plot, the businessmen are in the lead for annuity amount, and are also non-defaulters as well.
- Surprisingly, those who are Unemployed also have more annuity with an amount of around 450000.
- State servants also enjoy a good annuity amount, which are then followed by the Pensioners who have the least amount of annuity amount.
- Working class and Students have the same median value, but working people have more annuity amount as compared to the Students category.
- Maternity leave category also have a high amount of annuity.
- Commercial associates have similar amount of annuity as Unemployed.


Defaulters
AMT_ANNUITY
60000
50000
40000
30000
20000
10000
0
Working
Commercial associate
Pensioner
State servant
Unemployed
Maternity leave
NAME_INCOME_TYPE


Non-defaulters
AMT_ANNUITY
60000
50000
40000
30000
20000
10000
0
State servant
Working
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
NAME_INCOME_TYPE

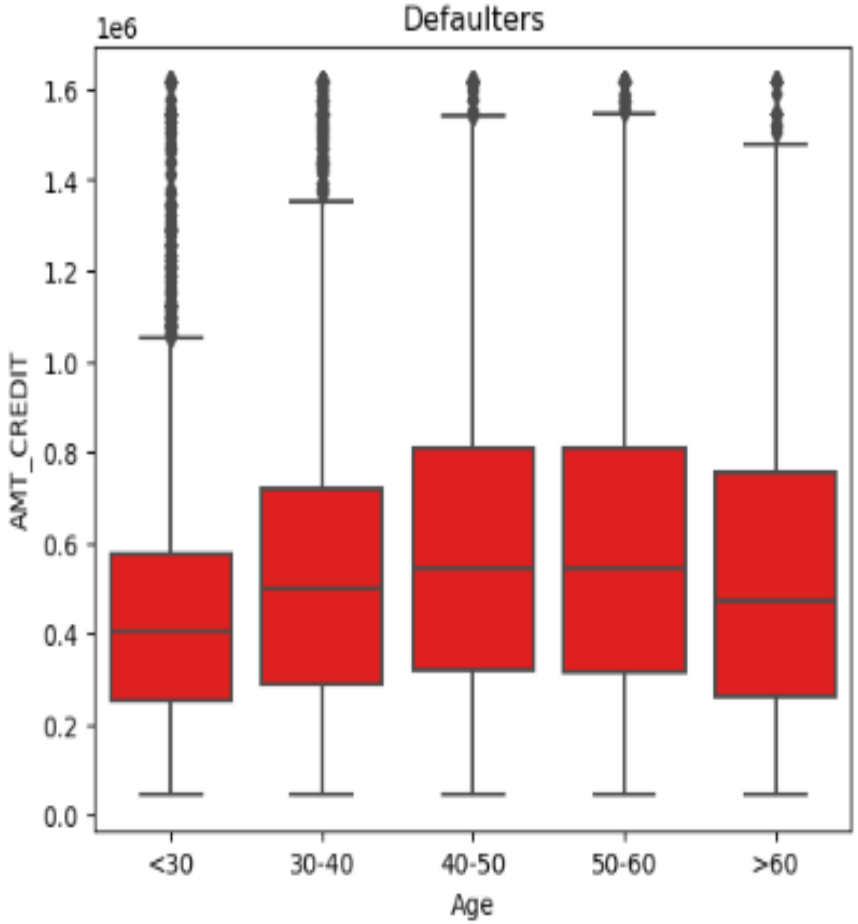
Graphs and Insights

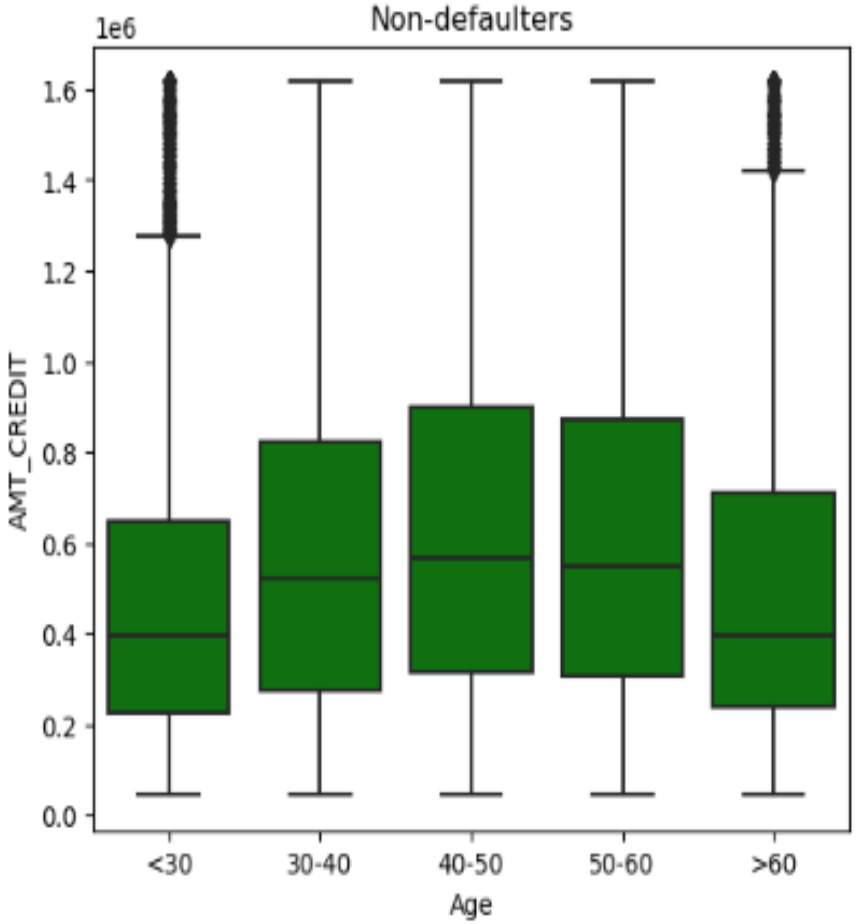
- **Bivariate analysis between Age and Credit**

The above plot indicates that:

- For the defaulters, the maximum credit has been taken by the 40-50 and 50-60 age group, followed by >60 and 30-40 age group.
- Those younger than 30 years have very less credit for both defaulters and non-defaulters.
- For non-defaulters, the age group of 40-50 tend to take more loans followed by those who are older than 50, i.e., the 50-60 and >60 age groups, who are then finally followed by 30-40 age group.

The conclusion is that, those younger than 30 years of age are trusted more because they take up less credit and also pay back the loans on time. The same can be said about those older than 60 years of age, as they too are trusted by banks to pay back the loans as they have quite a lot of savings by the time they retire.


Defaulters
1e6
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
AMT_CREDIT
<30
30-40
40-50
50-60
>60
Age

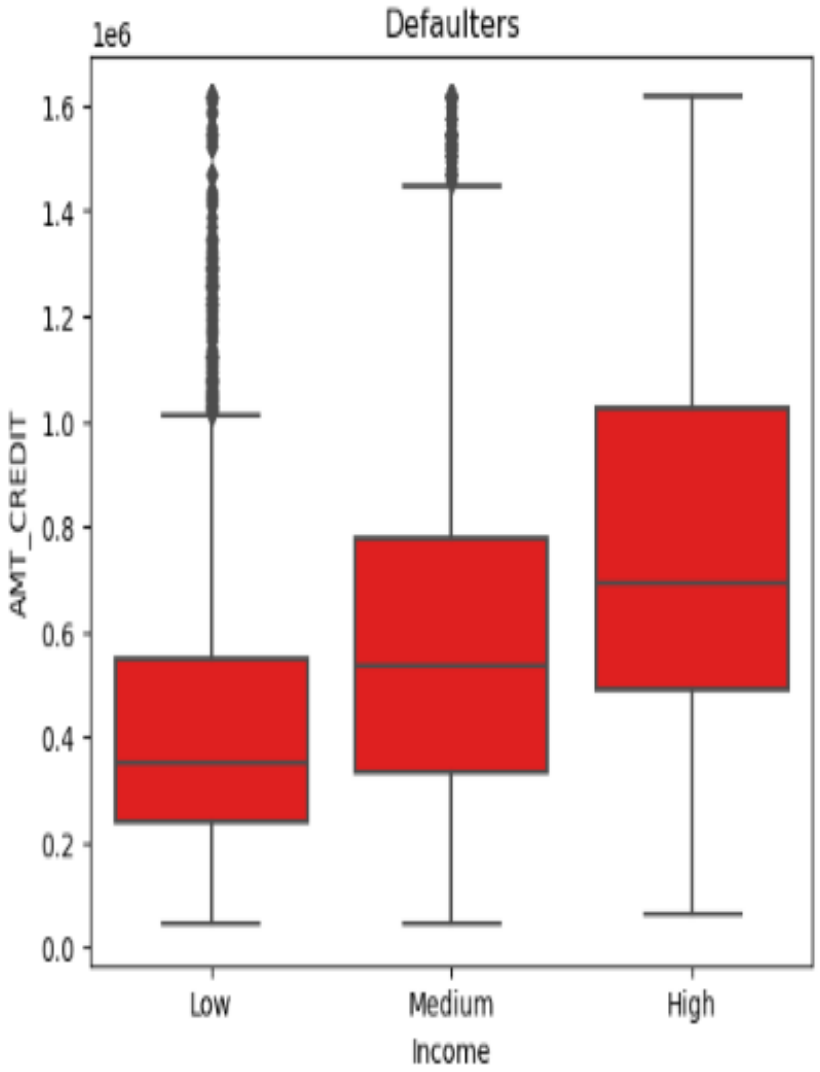

Non-defaulters
1e6
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
AMT_CREDIT
<30
30-40
40-50
50-60
>60
Age

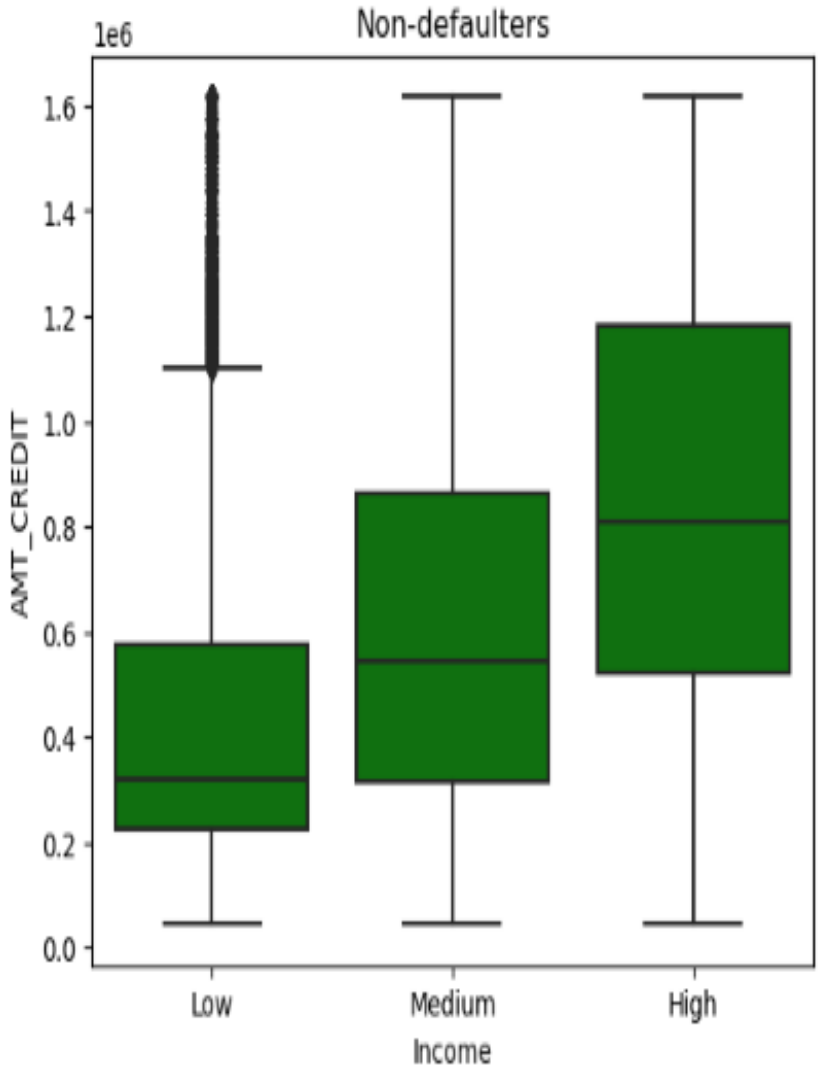
Graphs and Insights

- **Bivariate analysis between income and credit amount**

The above plot indicates that:

- In the defaulters category, people with high income have a high credit amount, while people with low income have a low credit amount.
- The same trend can be seen in the non-defaulters category too, but the different is that those with high income have been given more credit as compared to those of people who have high income in defaulters category. The conclusion is that the bank trusts a a group of people who have been rewarded for this trust.


Defaulters
1e6
AMT_CREDIT
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Low
Medium
High
Income


Non-defaulters
1e6
AMT_CREDIT
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Low
Medium
High
Income

Graphs and Insights

- **Bivariate analysis between credit amount and income type**

The plot indicates that:

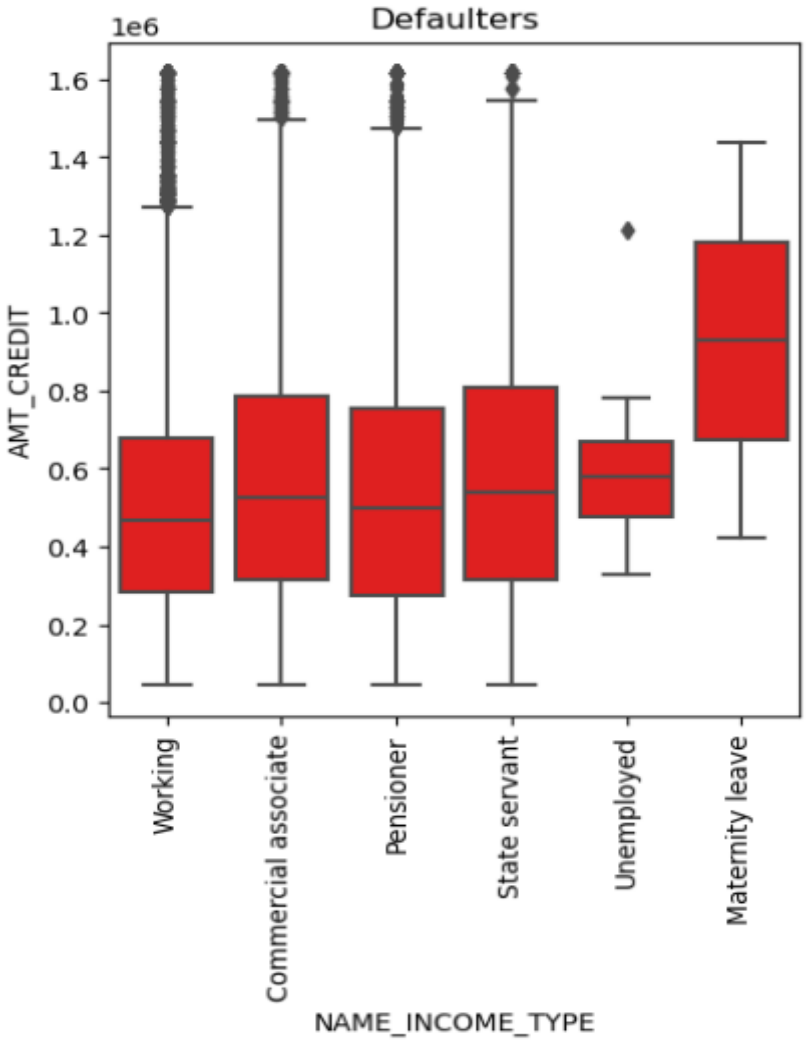
1. Defaulters:

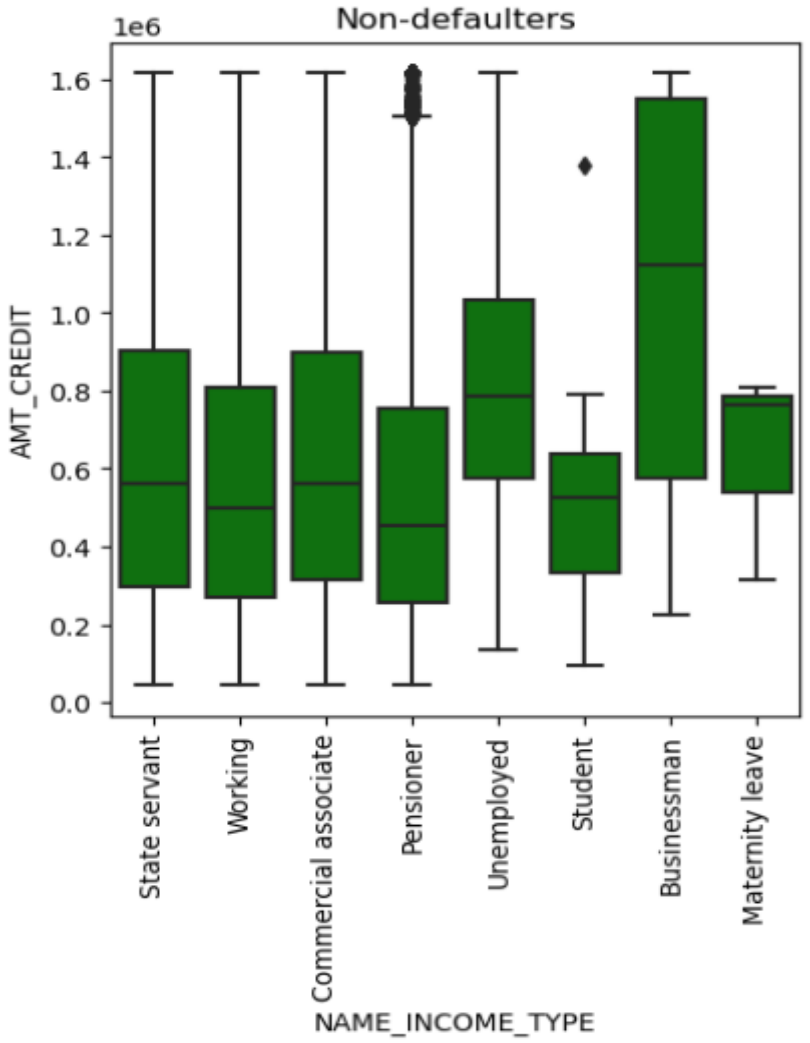
- People who take more loans have more credit are under the maternity leave category. On other hand Commercial associates and State servants have nearly the same credit values., while the Working class have low credit value.
- Pensioners fall under the same category as Commercial associates and State servants.
- The reason for this could be family planning, thus taking maternity leave, thus taking more credit.

2. Non-defaulters:

- The businessmen have the highest credit amount among all the categories.
- They are then followed closely by Unemployed people who require financial aid for living, but they do pay their loans back.
- Similar to defaulters, Commercial associates and State servants follow the same credit pattern, who are then followed by the Working class.
- Pensioners in non-defaulters take less credit as compared to their defaulter counterpart.

Businessmen require more credit to run their businesses, hence their high loan amounts. Unemployed take loans to meet their ends. Pensioners take up less credit as they have their own savings to depend on.


Defaulters
1e6
AMT_CREDIT
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Working
Commercial associate
Pensioner
State servant
Unemployed
Maternity leave
NAME_INCOME_TYPE

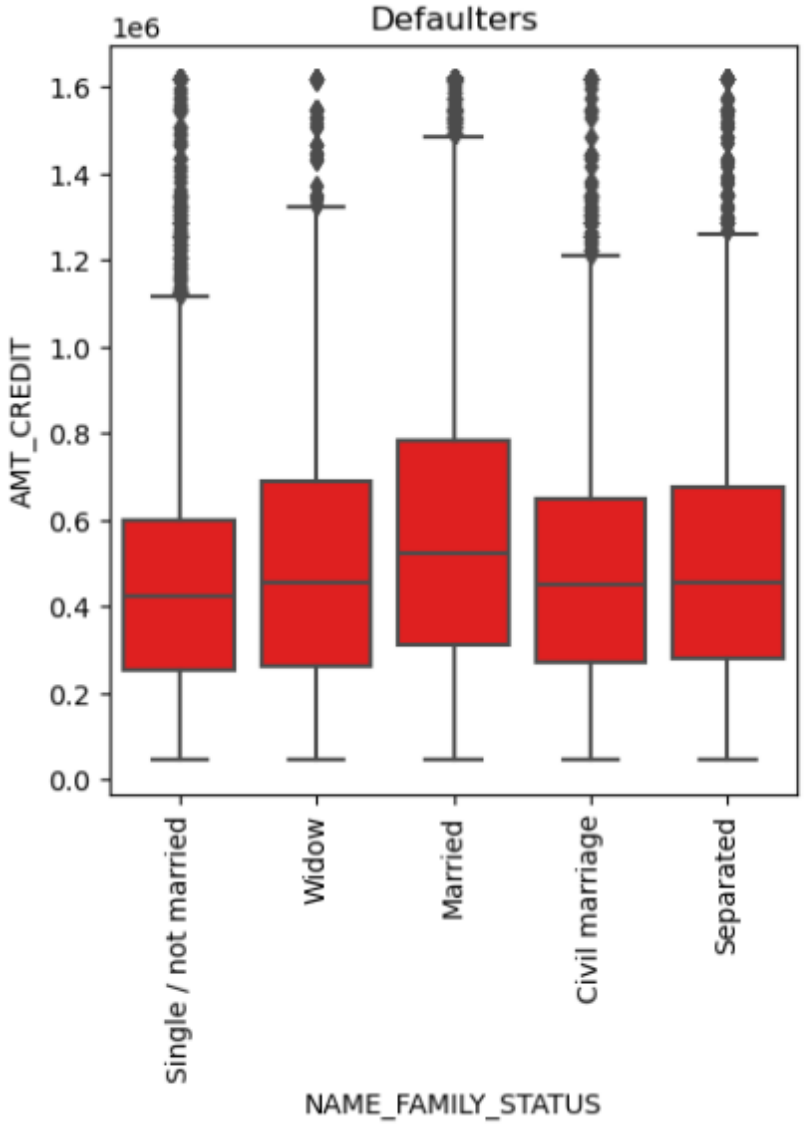

Non-defaulters
1e6
AMT_CREDIT
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
State servant
Working
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
NAME_INCOME_TYPE

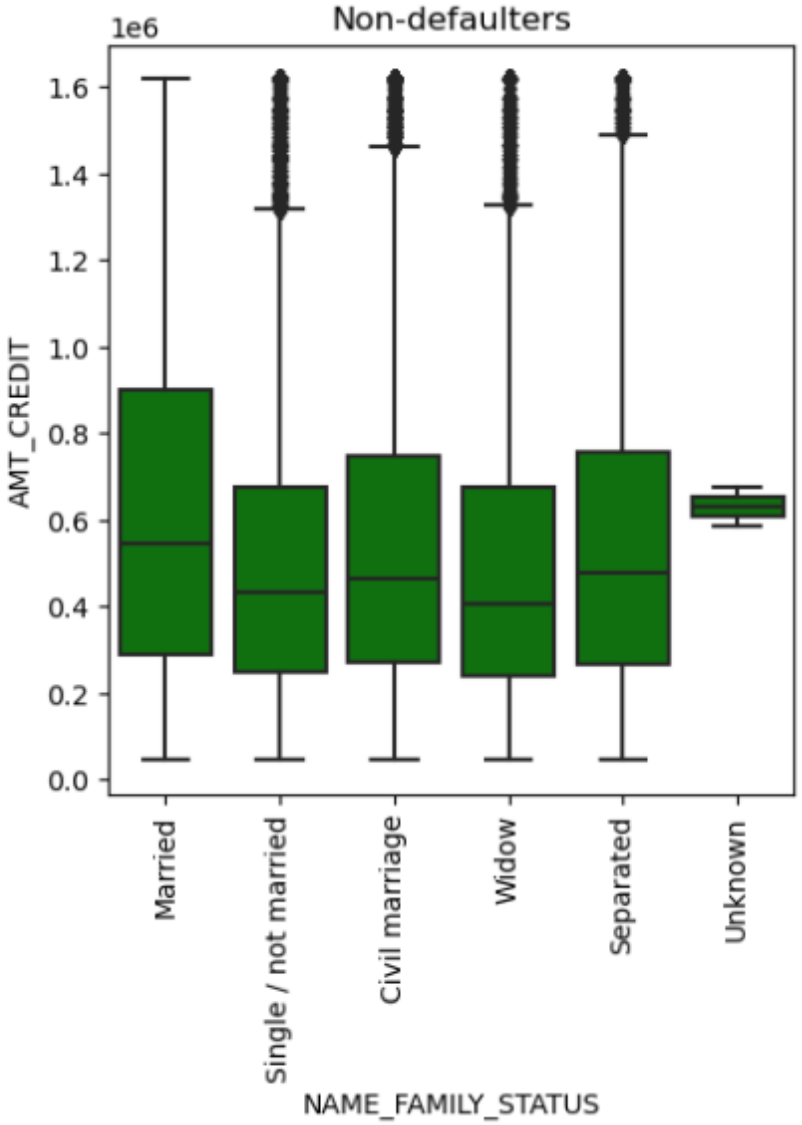
Graphs and Insights

- **Bivariate analysis between credit amount and family status**

The plot indicates that:

1. Defaulters:
    - Married people take more credit and are defaulters as well.
    - The people who are single or have a civil marriage have less credit amount, with the median being at the same credit amount.
    - Windows and Separated have the same loan amounts.
2. Non-defaulters:
    - Married people have the highest credit amount.
    - People who have done civil marriage or are Separated have the same credit amount, with their median values being the same, along with the window category too.


Defaulters
1e6
AMT_CREDIT
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Single / not married
Widow
Married
Civil marriage
Separated
NAME_FAMILY_STATUS

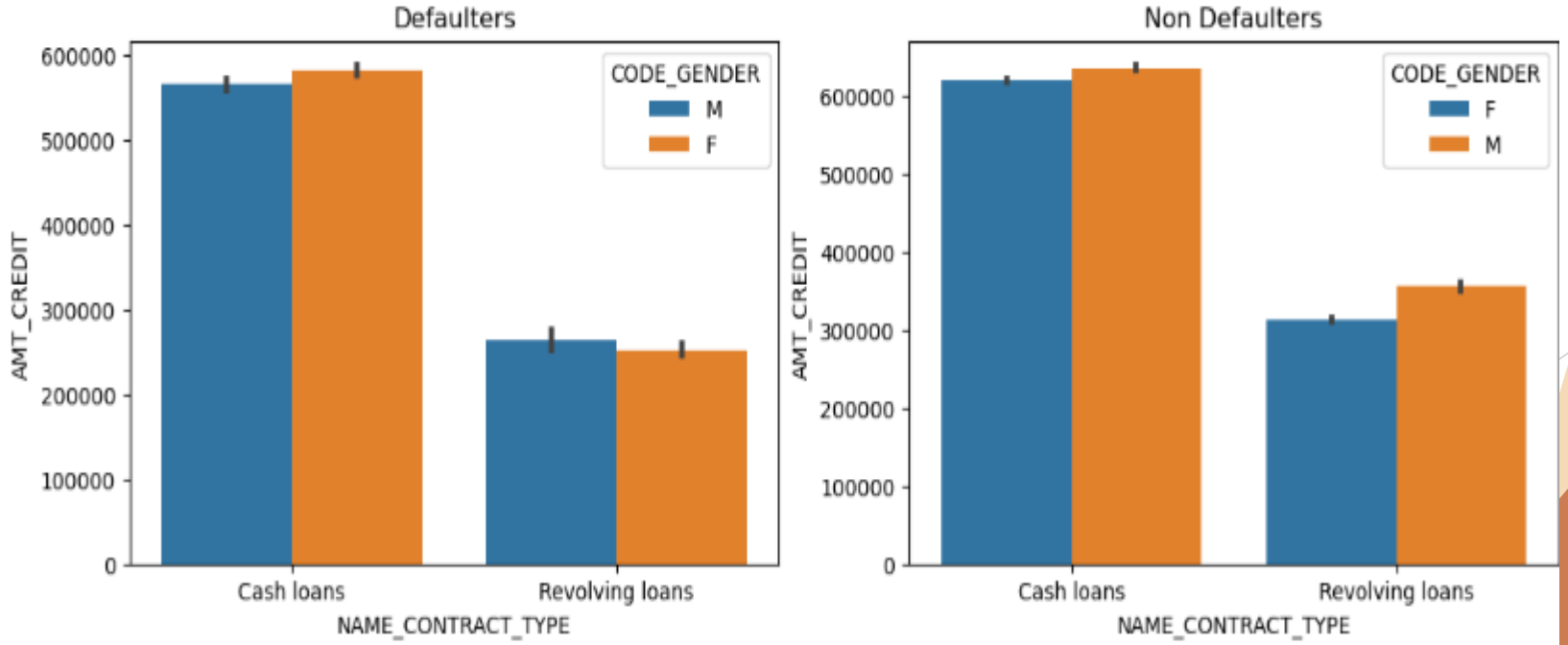

Non-defaulters
1e6
AMT_CREDIT
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Married
Single / not married
Civil marriage
Widow
Separated
Unknown
NAME_FAMILY_STATUS

Graphs and Insights

- **Bivariate analysis between Gender and Contract type**

The above plot indicates that:

- There are more females who get cast loans as compared to their male counterparts in the defaulter category. Whereas, revolving loans are offered more to males than females.
- In the non-defaulters plot, we see that both cash loans and revolving loans are given more to male clients than females clients
- From this we can conclude that, banks trust male clients more to pay their loans back than their female counterparts



Graphs and Insights

- **Bivariate analysis between income type and income total**

The plot indicates that:

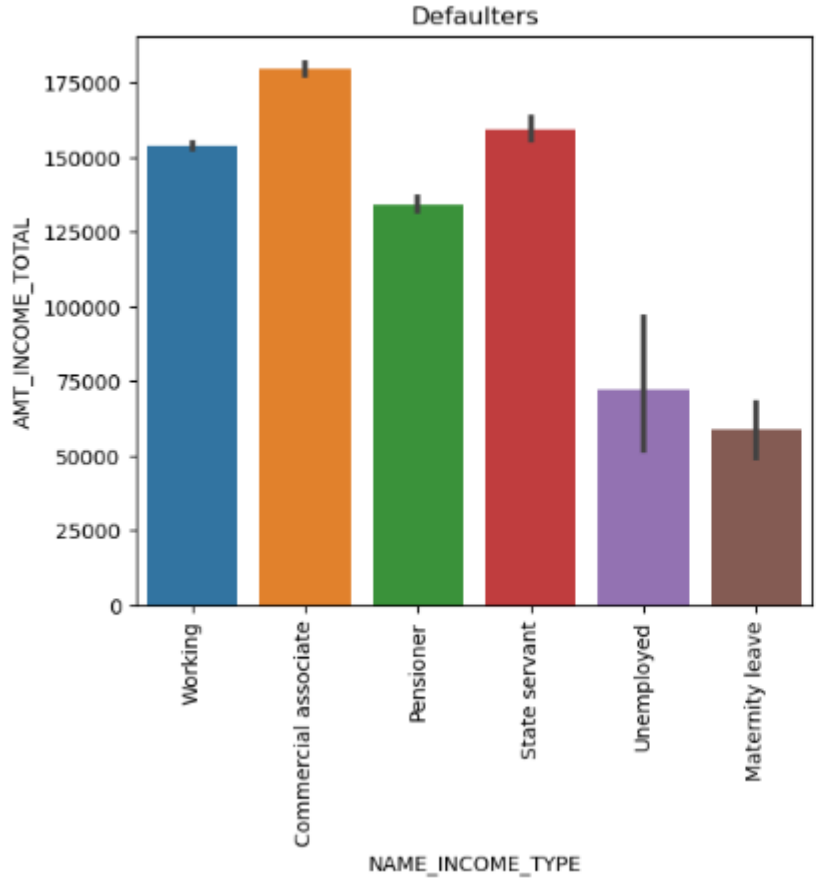
For defaulters:

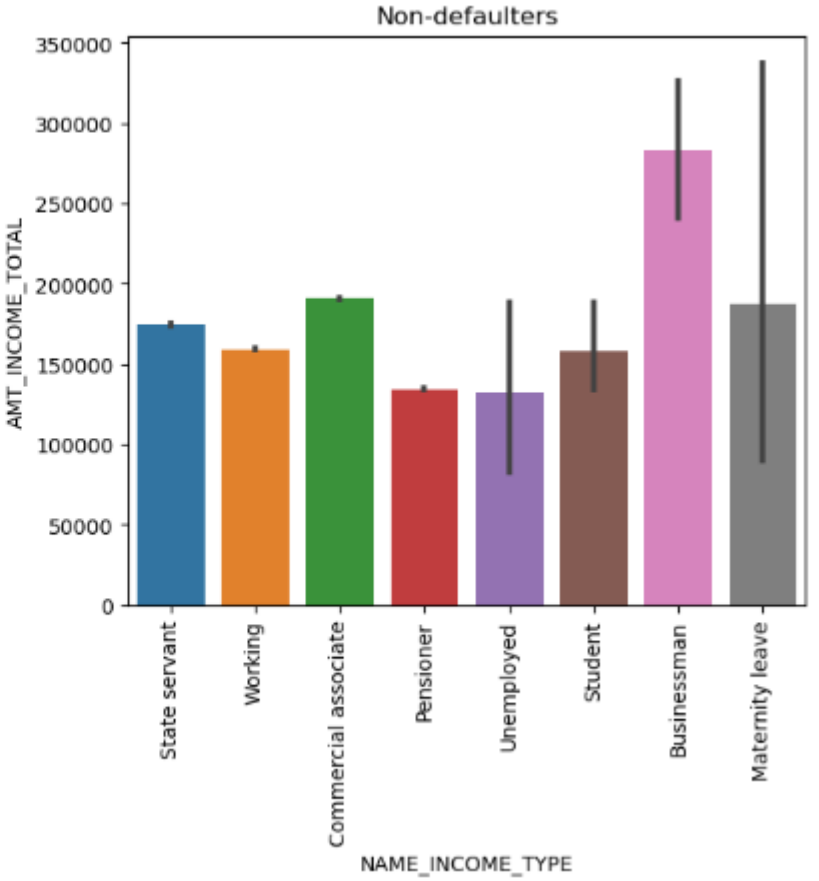
- The income bracket for Commercial associates is the highest followed by State servants, Working class, Pensioner, Unemployed and Maternity leave.

For non-defaulters:

- The income bracket for Businessmen is the highest followed by Maternity leave, Commercial associates, State servants, Working class, Students, Unemployed and finally Pensioners.

The conclusion is that defaulters do not businessmen and student as its categories. They are trusted by banks to pay their loans.


Defaulters
AMT_INCOME_TOTAL
175000
150000
125000
100000
75000
50000
25000
0
Working
Commercial associate
Pensioner
State servant
Unemployed
Maternity leave
NAME_INCOME_TYPE


Non-defaulters
AMT_INCOME_TOTAL
350000
300000
250000
200000
150000
100000
50000
0
State servant
Working
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
NAME_INCOME_TYPE

Graphs and Insights

- **Bivariate analysis between gender and credit amount on the basis of income type**

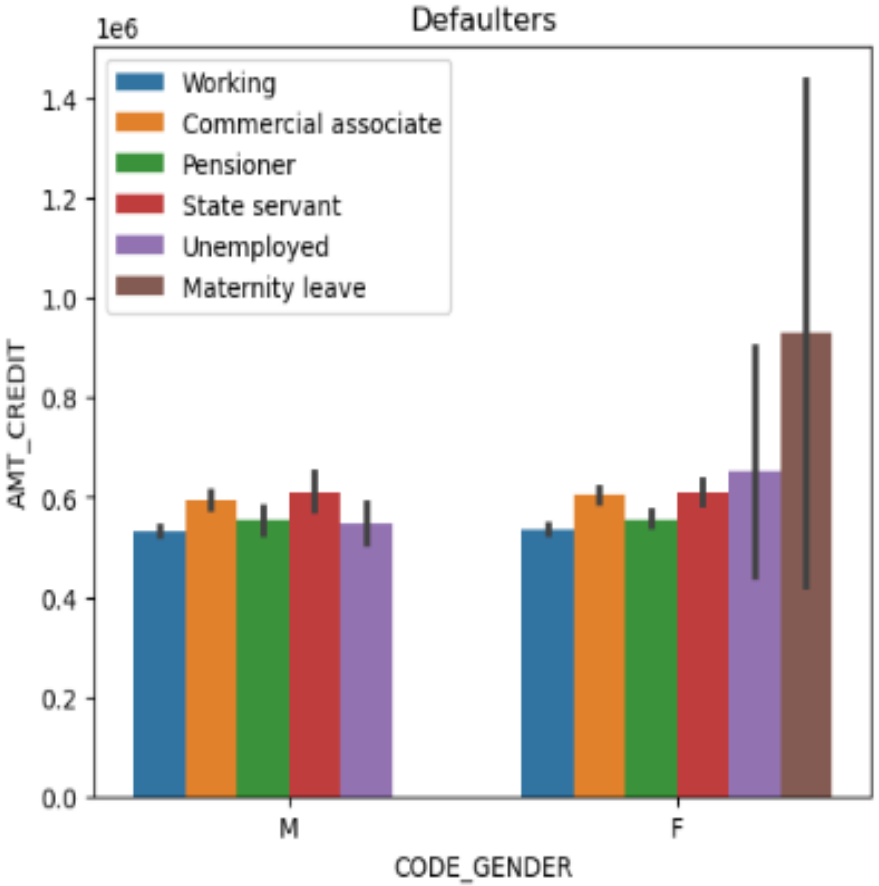
The plot indicates that:

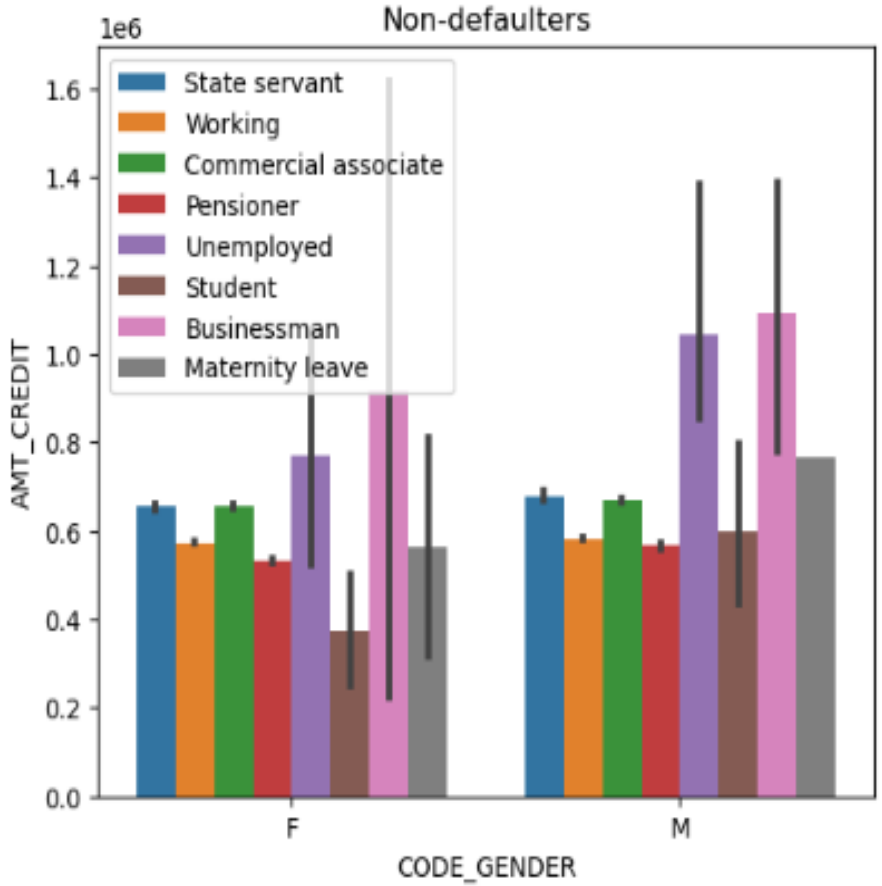
Defaulters:

- Females on maternity leave take up larger amounts of credit than others.
- The males have credit amount similar to each other. Conclusion: Banks should be careful with giving loans to pregnant women.

Non-defaulters:

- Whether male or female, businessmen are the ones who take up most loans and are also non-defaulters. This category is a safe bet for the banks to make and can be trusted.
- Important note is that, both businessmen and student categories are not present under defaulters.


Defaulters
1e6
Working
Commercial associate
Pensioner
State servant
Unemployed
Maternity leave
AMT_CREDIT
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
M
F
CODE_GENDER


Non-defaulters
1e6
State servant
Working
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
AMT_CREDIT
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
F
M
CODE_GENDER

Graphs and Insights

- **Bivariate analysis between gender and credit amount on the basis on age**

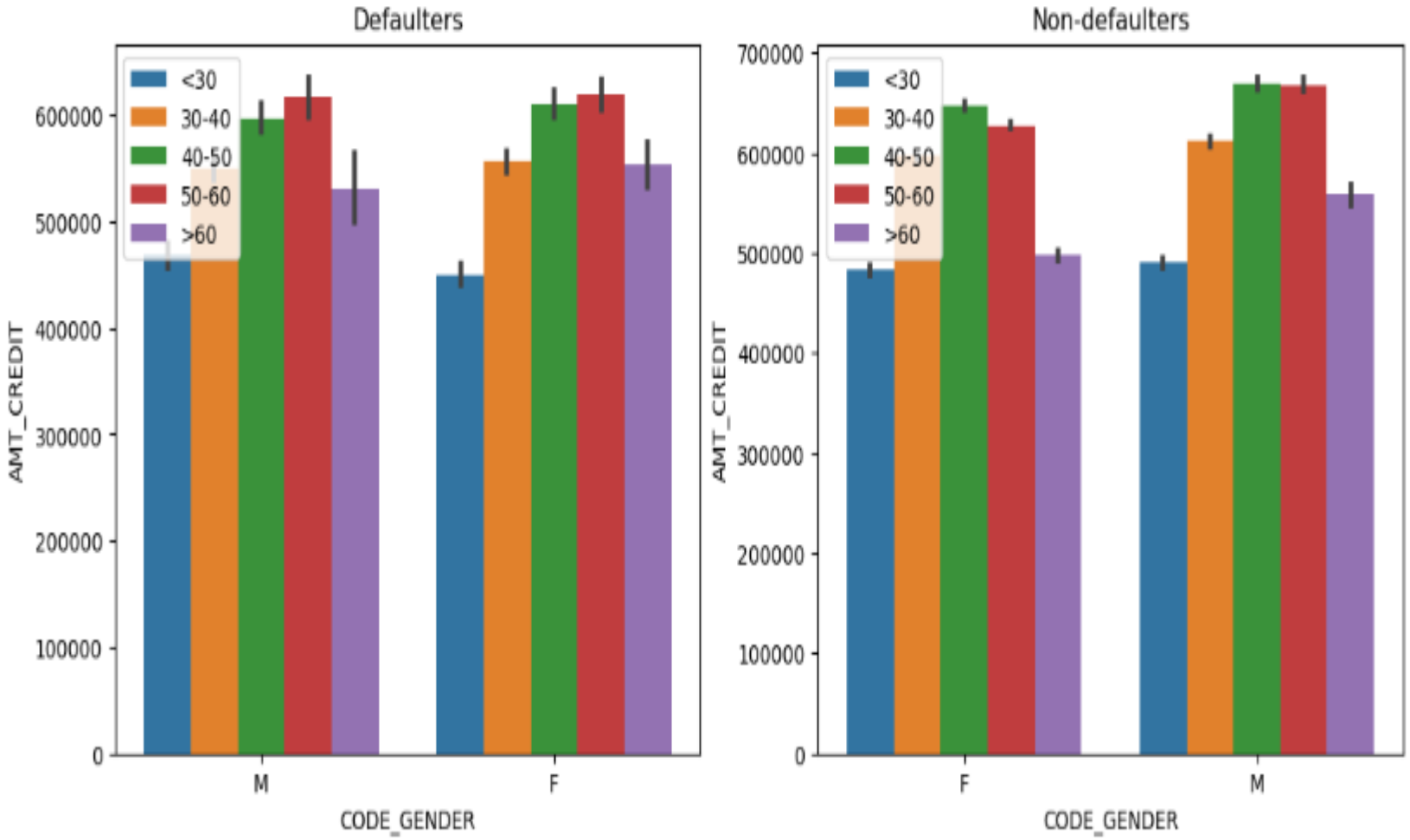
The above plot indicates that:

Defaulters:

- Males in the age category of 40-50 and 50-60 take up more credit amounts and are defaulters as well.
- Females in the age category of 40-50 and 50-60 also take up more crediit amount than those in >60 adn 30-40 age category.
- The safest bet to give loans is the 30-40 age category for both males and females.

Non-defaulters:

- The maximum loans is withing the age category of 40-50 for both males and females, followed by 30-40 and >60.

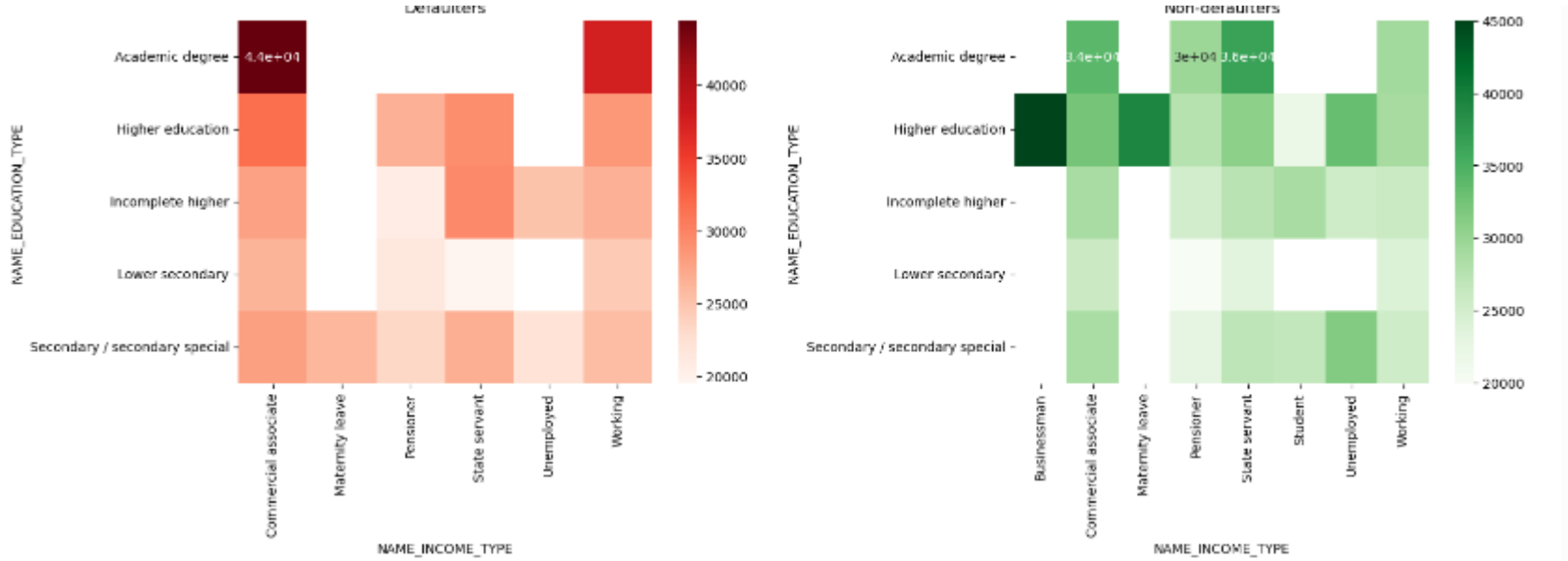

Defaulters
<30
30-40
40-50
50-60
>60
AMT_CREDIT
0
100000
200000
300000
400000
500000
600000
M
F
CODE_GENDER
Non-defaulters
<30
30-40
40-50
50-60
>60
AMT_CREDIT
0
100000
200000
300000
400000
500000
600000
700000
F
M
CODE_GENDER

Graphs and Insights

- **Multivariate analysis for Education type vs Income type vs annuity**

The above plot indicates:

1. Defaulters:
   - The heatmap shows the relation between academic degree and commercial associates is highly correlated where as the lower section is the lest correlated with any of the income types.
2. Non-defaulters:
   - The heatmap shows that higher education is positively correlated with Businessmen



Graphs and Insights

- **Multivariate analysis for Housing type vs Family status vs credit amount**

The above plot indicates that:

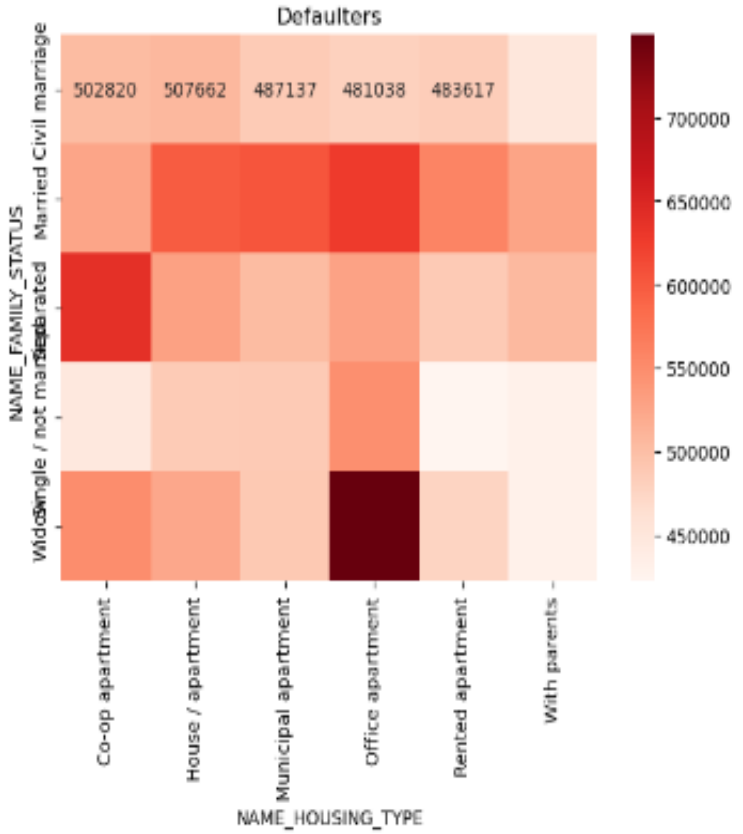
Defaulters:

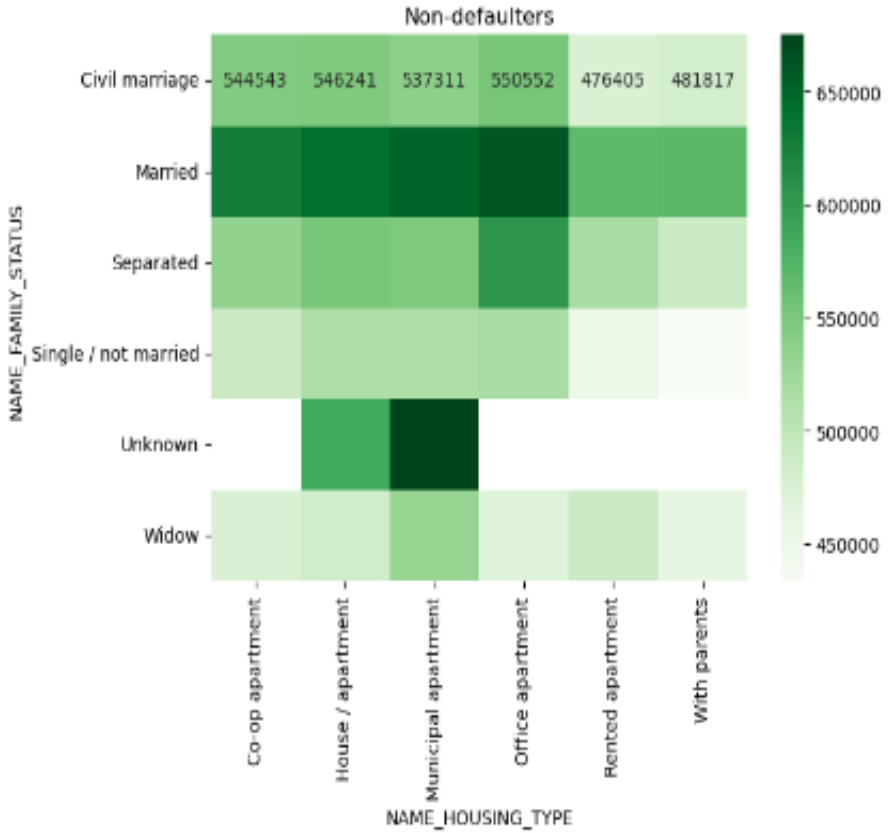
- The heatmap shows that the widows and office apartment housing type has a very strong correlation, followed by Separated and co-cop housing type.
- Married people also have a positive correlation with all the housing types.

Non-defaulters:

- The strongest correlation is between married people and office apartment housing type.

Similar to defaulters, there is a strong correlation of married people to all housing types.


Defaulters
NAME_FAMILY_STATUS
Civil marriage
Married
Separated
Single / not married
Widow
502820
507662
487137
481038
483617
700000
650000
600000
550000
500000
450000
Co-op apartment
House / apartment
Municipal apartment
Office apartment
Rented apartment
With parents
NAME_HOUSING_TYPE


Non-defaulters
NAME_FAMILY_STATUS
Civil marriage
Married
Separated
Single / not married
Unknown
Widow
544543
546241
537311
550552
476405
481817
650000
600000
550000
500000
450000
Co-op apartment
House / apartment
Municipal apartment
Office apartment
Rented apartment
With parents
NAME_HOUSING_TYPE

Graphs and Insights

- **Multivariate analysis for occupation type vs income type vs credit amount**

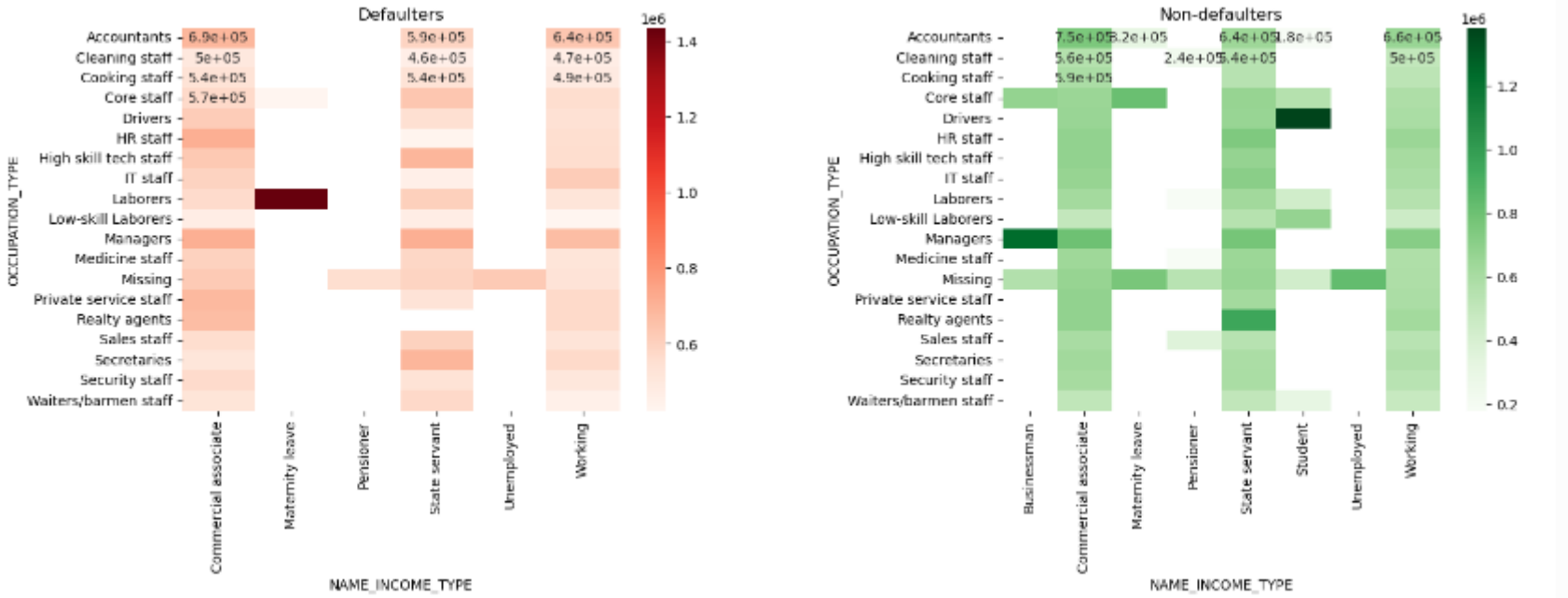
The above plot indicates that:

Defaulters:

- Laborers have a string correlation with maternity leave income type.

Non-defaulters:

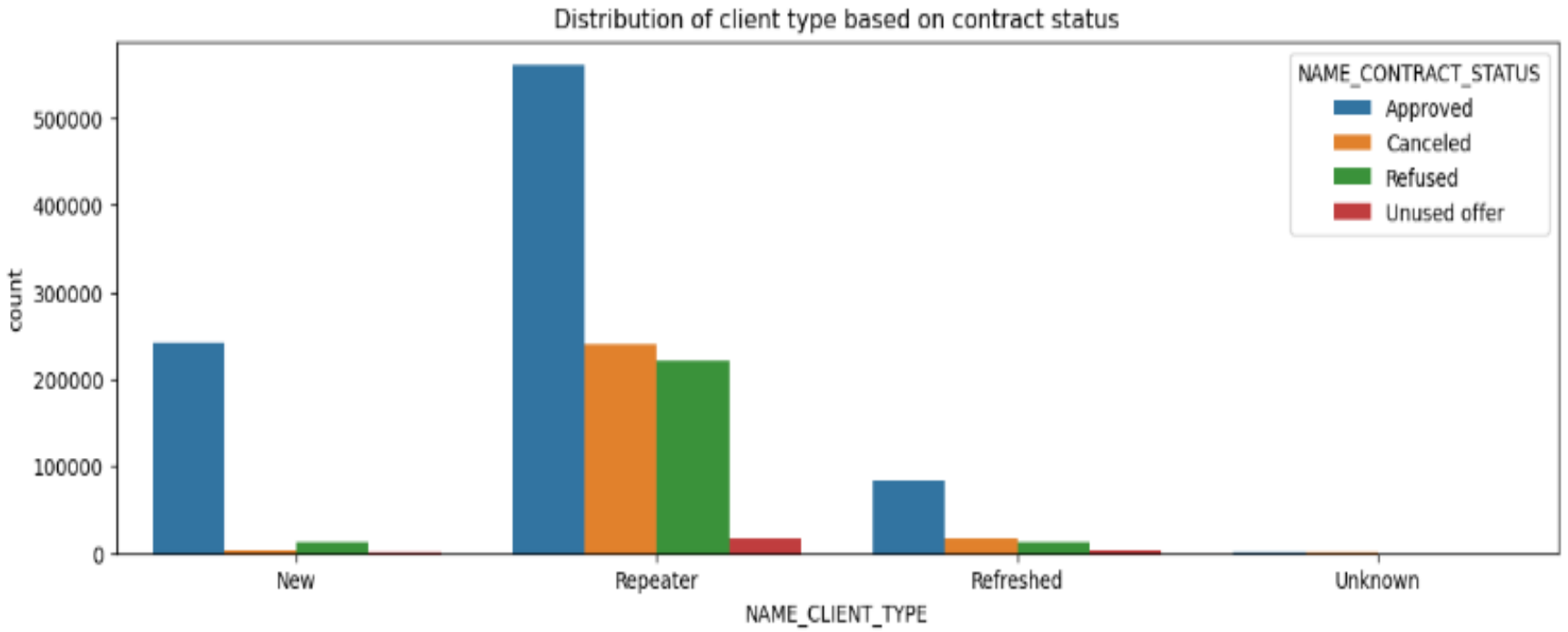
- High correlation is seen between drivers occupation type and student income type, followed by managers occupation type and commercial associates occupation type.



Graphs and Insights

- **Count plot for Client type on basis of Contract status**

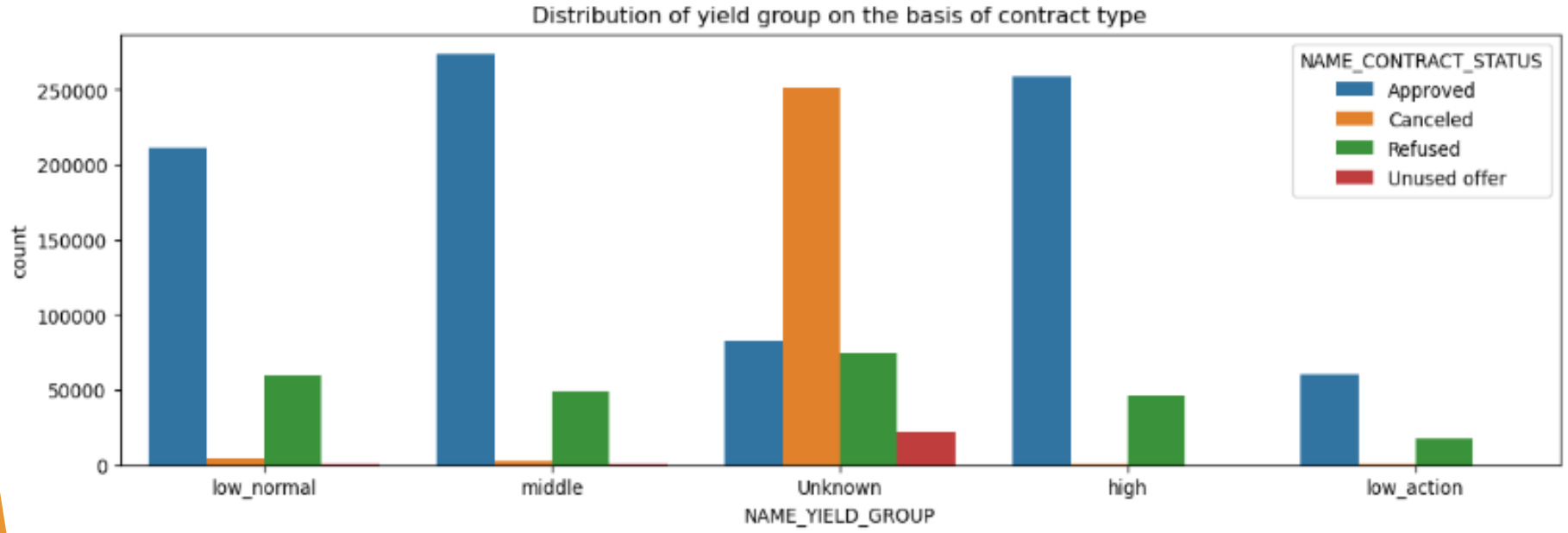
The above plot indicates that the people who are in the 'Repeater' category, their loans have been approved as the count of such loans is the maximum.



Graphs and Insights

- **Count plot for Yield group on basis of Contract status**

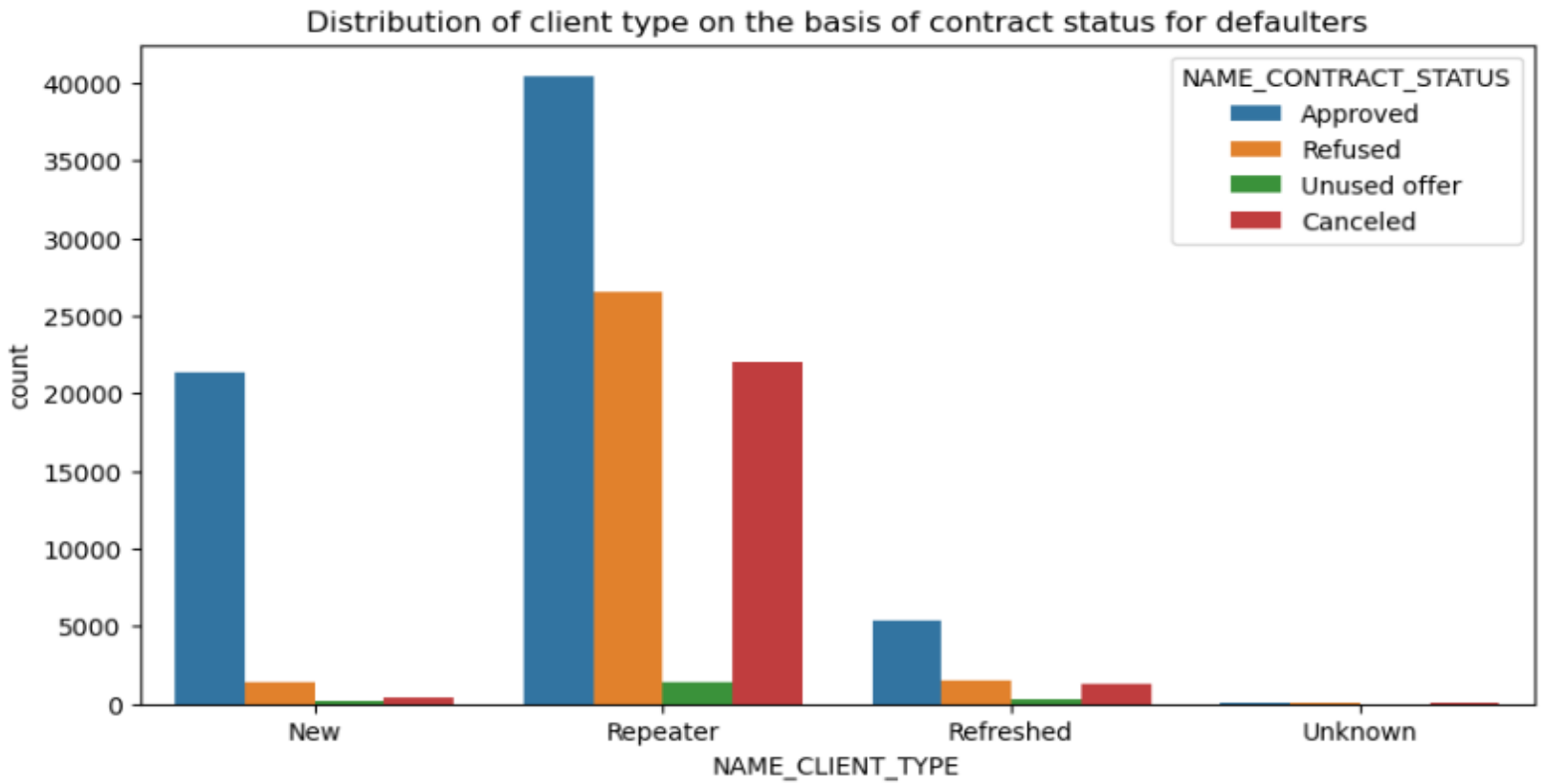
The above plot indicates that the yield group in the middle are ones whose loans are being approved the most.



Graphs and Insights

- **Distribution of Client type on the basis of Contract status for defaulters**

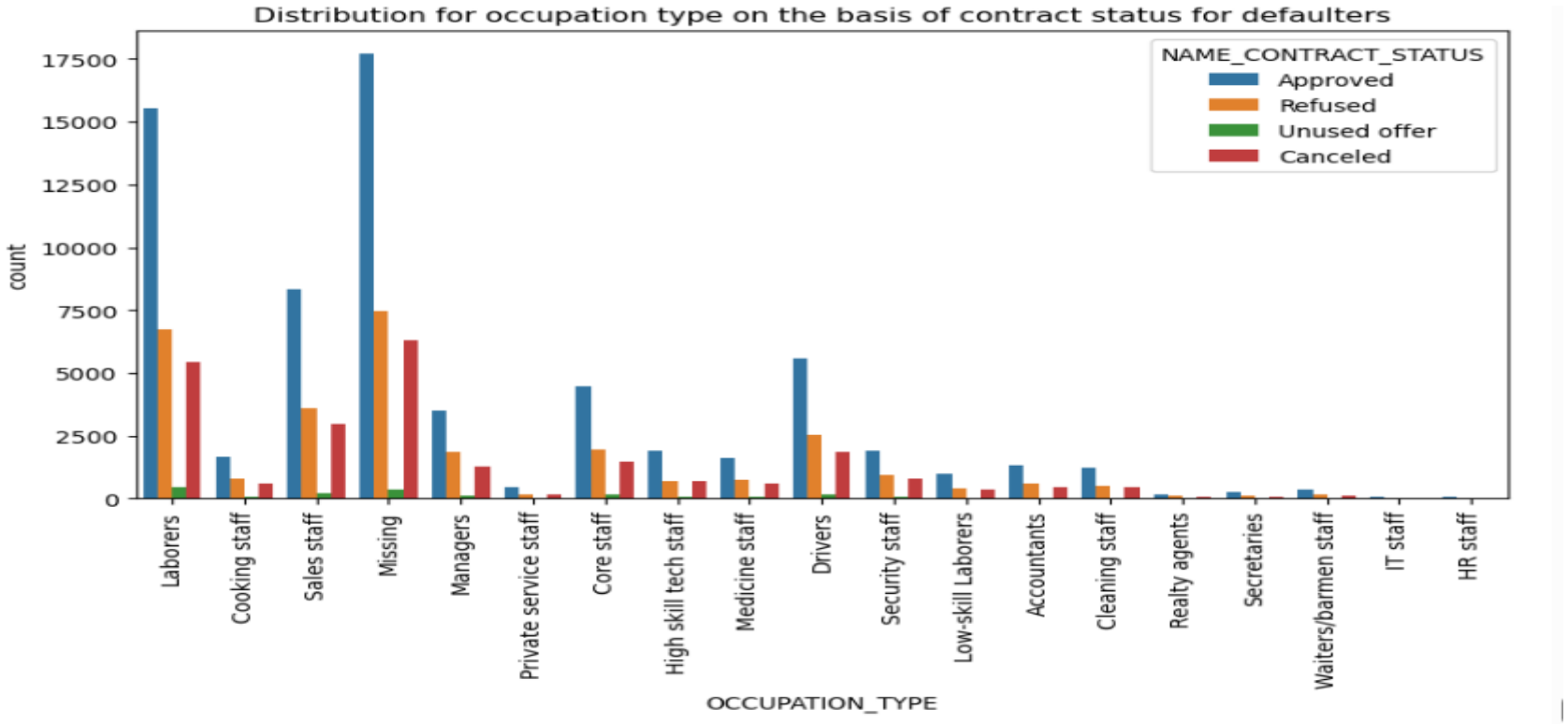
The above plot indicated that there are more defaulters from the 'Repeater' category who have applied for loans previously in the current data.



Graphs and Insights

- **Distribution of Occupation type on the basis of Contract status for defaulters**

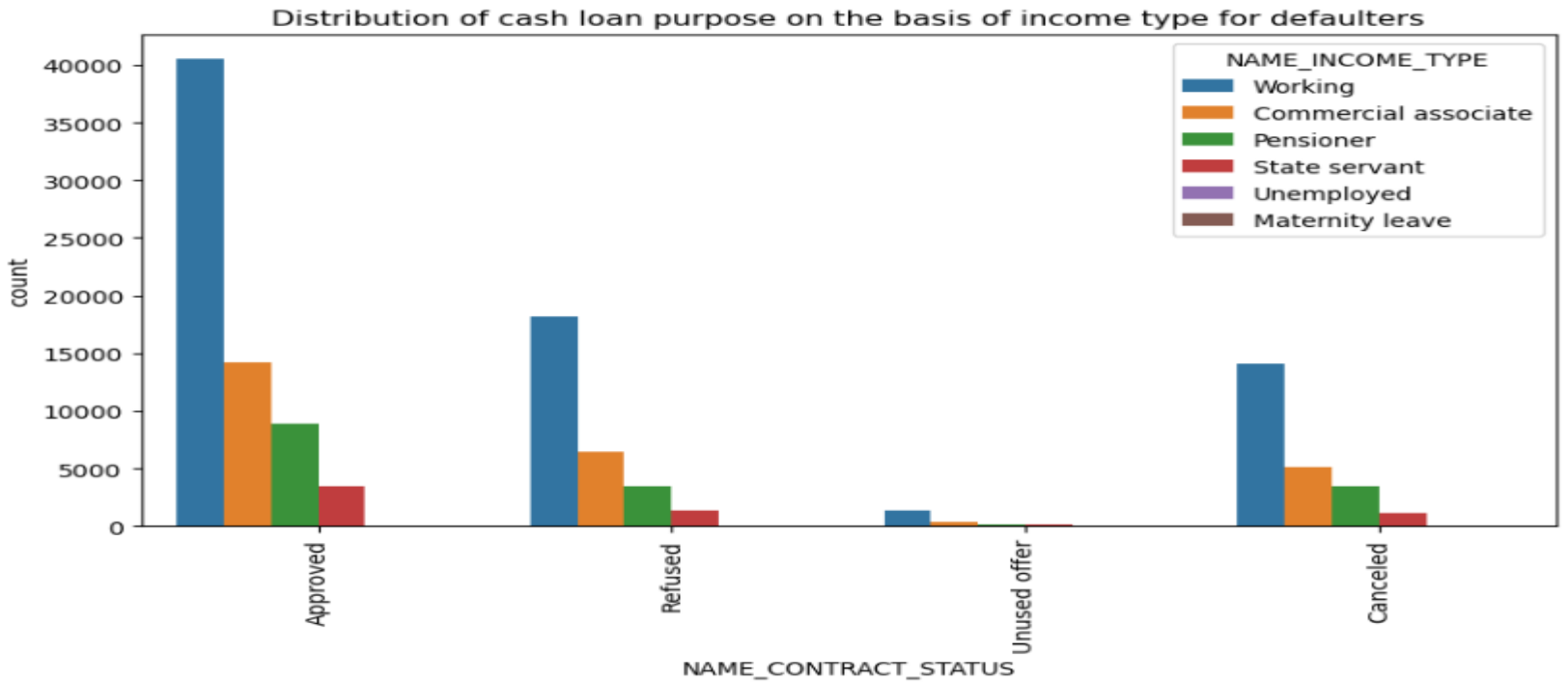
Even though laborers make up a large amount of defaulters, their loans are being approved more than they are refused.



Graphs and Insights

- **Distribution of Cash loan purpose on the basis of Income type for defaulters**

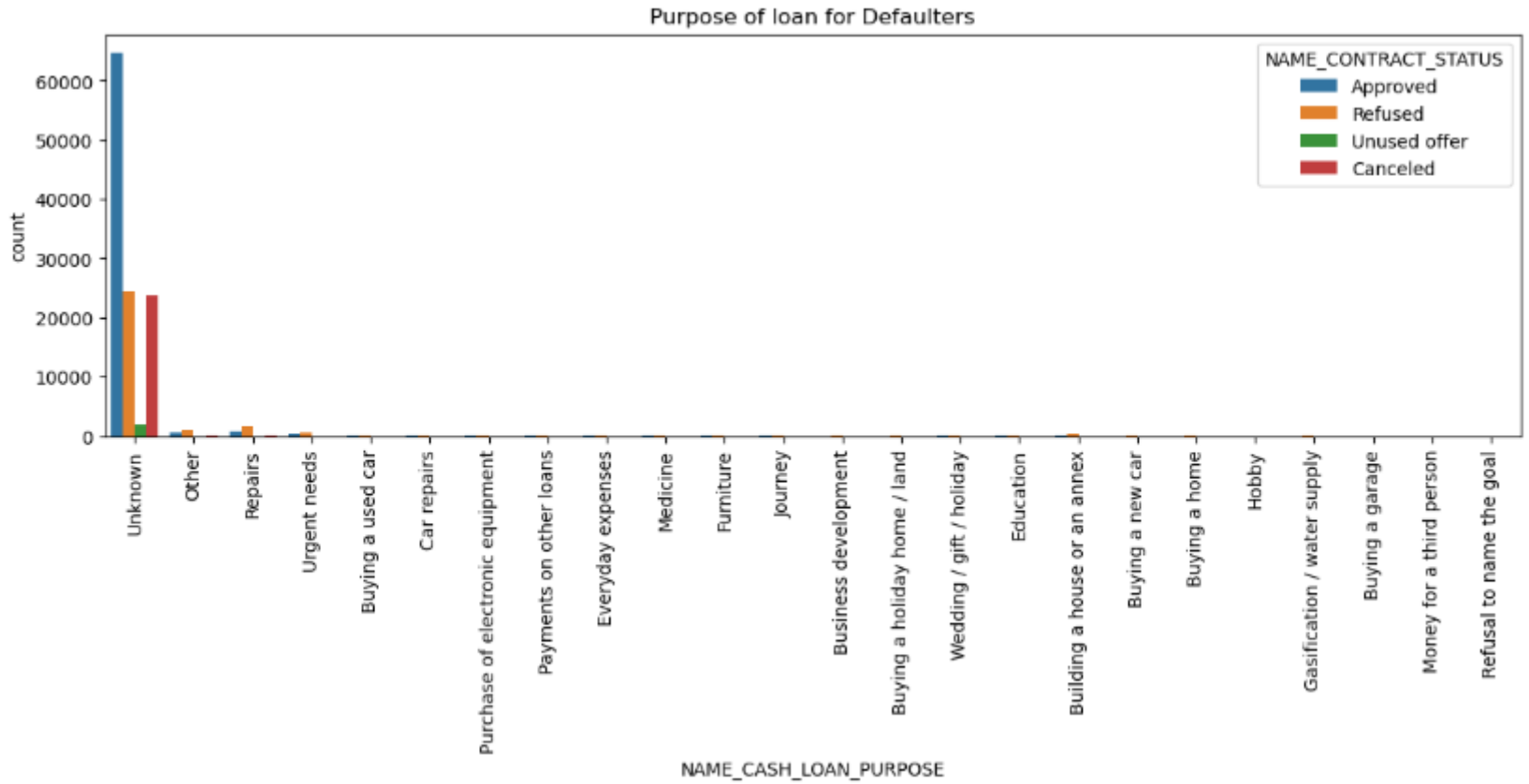
The above plot indicates that 'Working' class has the most approved loans.



Graphs and Insights

- **Purpose of loan for defaulters**

Most refused loans were rejected for 'Repairs'



Conclusion

1. Banks must target on occupation type 'Student', 'Pensioner' and 'Businessmen' for profitable business as these occupation types pay their loans back in a timely manner.
2. Banks must focus a less on income types 'Working' as they make up the most defaulters and leads to the banks to a financial loss.