Credit EDA Case Study

Dr. Deepa Jha
Batch – DSC40
Roll No – EDS210110080

Business Objective

1. To understand the driving factors behind the loan default and take actions to manage the risk

2. The key areas to be explored are:

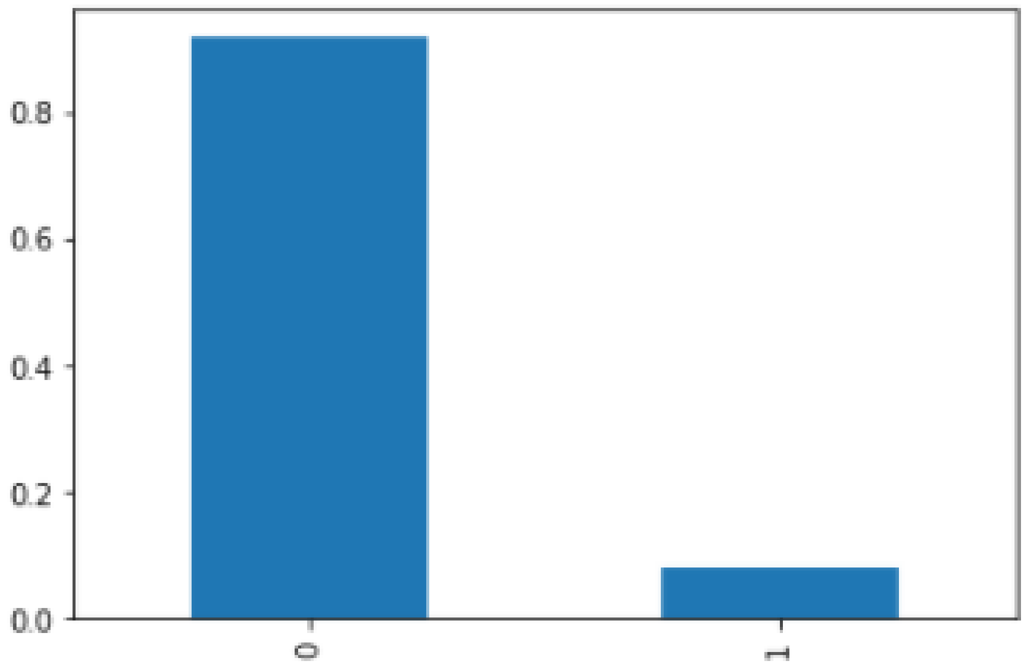
- ✓ Risk assessment of loan applications: based on gender, income, housing type, KYC (Know your customer) documents, previous applications, amount, type of loan etc.
- ✓ Understanding repayment and default patterns
- ✓ To get insight on denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate

Approach and Methodology

Approach and Methodology:

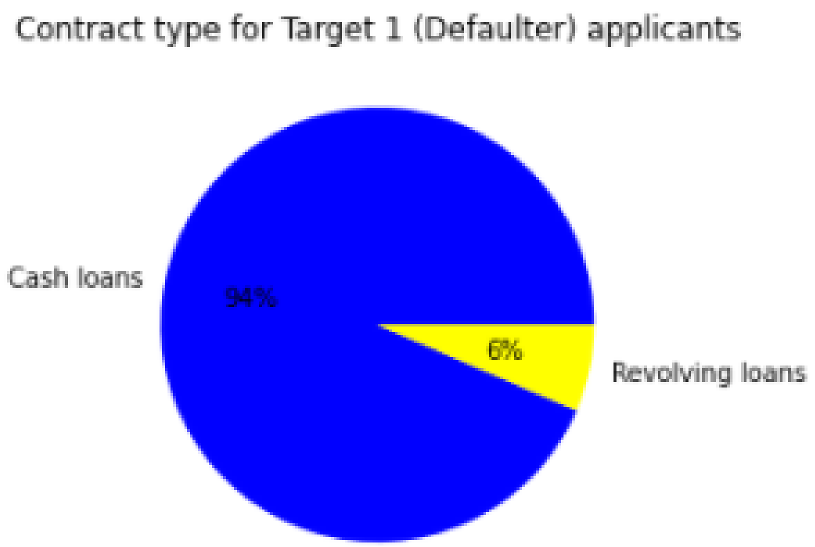
1. Extracting the data from both the given files (current and previous application)
2. Cleaning & Transforming the data like removing columns having high null values and suggesting the methods to handle the null values and identifying data outliers and data imbalance
3. Loading the transformed data
4. Using Python to visualize the data by plotting the relevant graphs

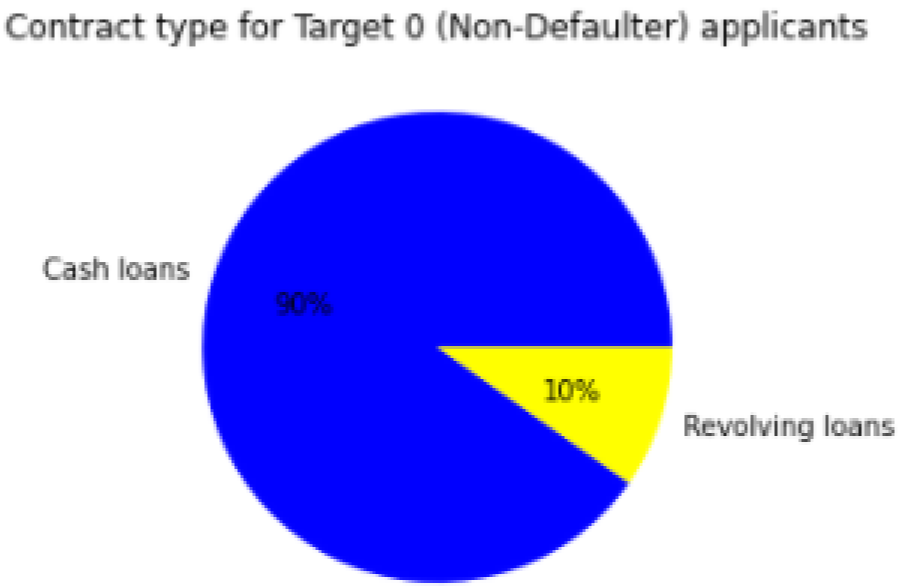
TARGET



Imbalance percentage has been detected in the 'TARGET' column. So, we will divide the dataset into 2 sets based on the value of this column.
If value of 'TARGET' = 1 then client has faced payment difficulties at some point of time (he/she had late payment)
If value of 'TARGET' = 0 for all other cases.

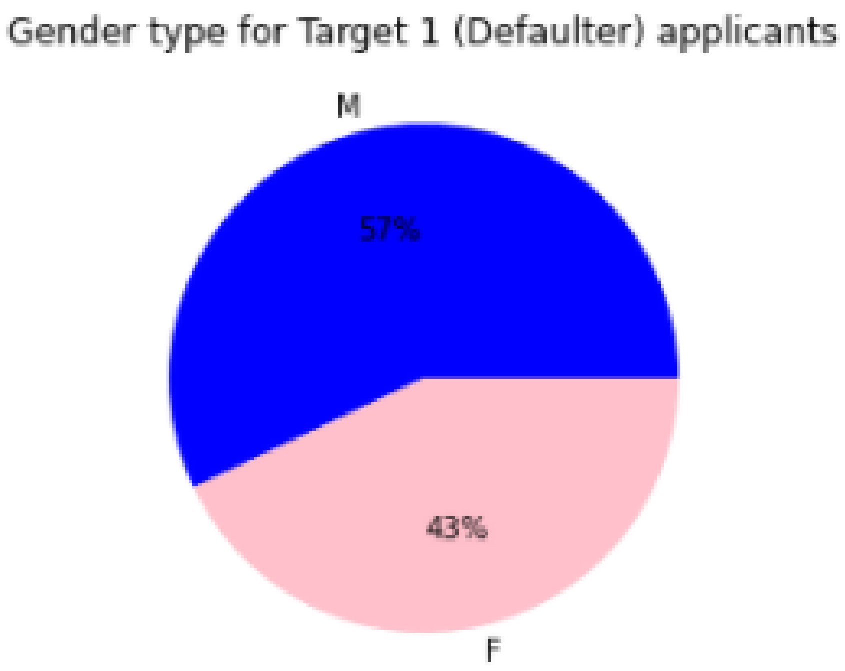
NAME_CONTRACT_TYPE

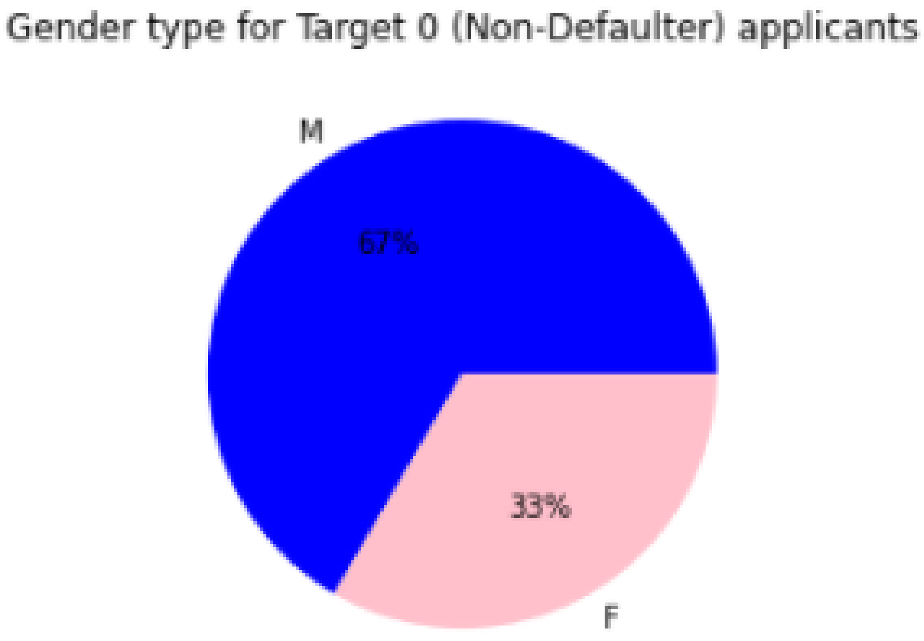




The majority of defaulters have cash loan
Revolving loans are less risky as 10% of non-defaulter comes from Revolving loans.
The data reveals that majority of loans are Cash loans.

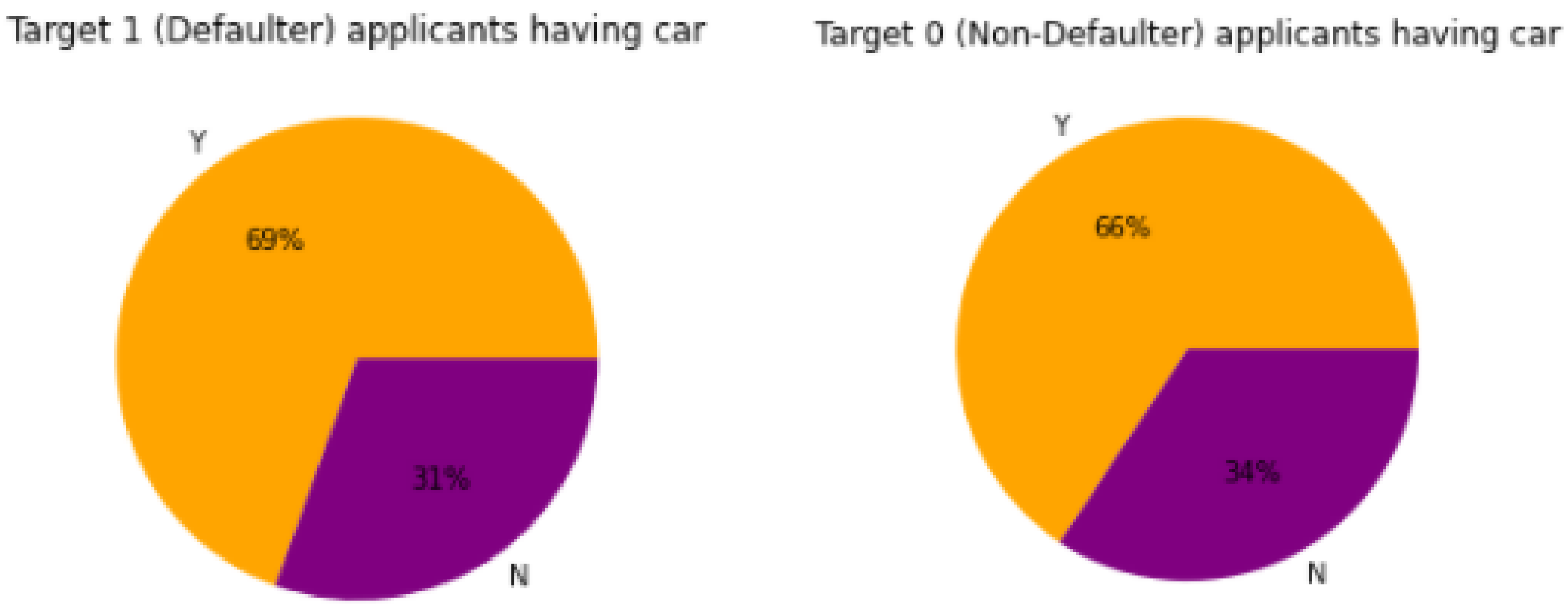
CODE_GENDER





Percentage of male defaulters is higher than the female.
2/3rd of non-defaulters are males while 1/3rd are females.

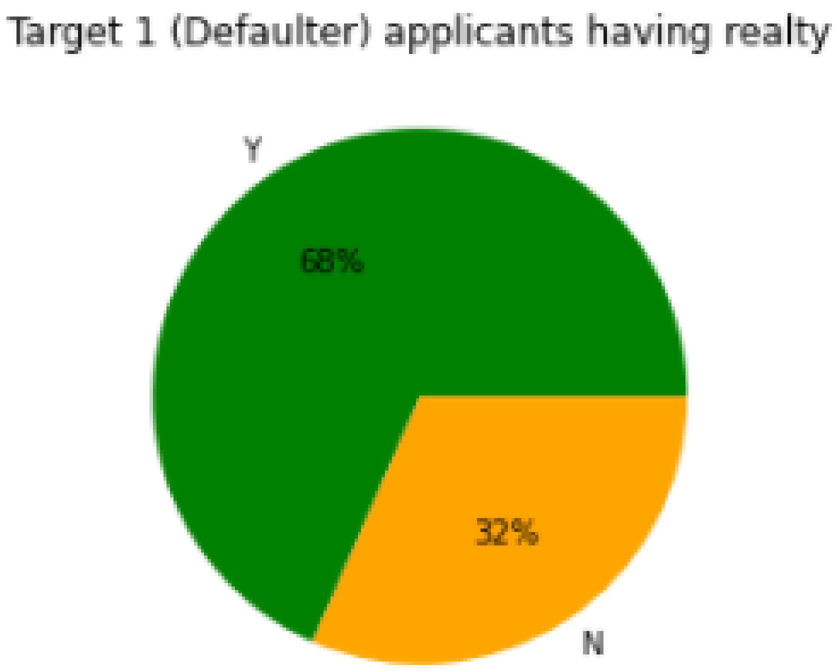
FLAG_OWN_CAR

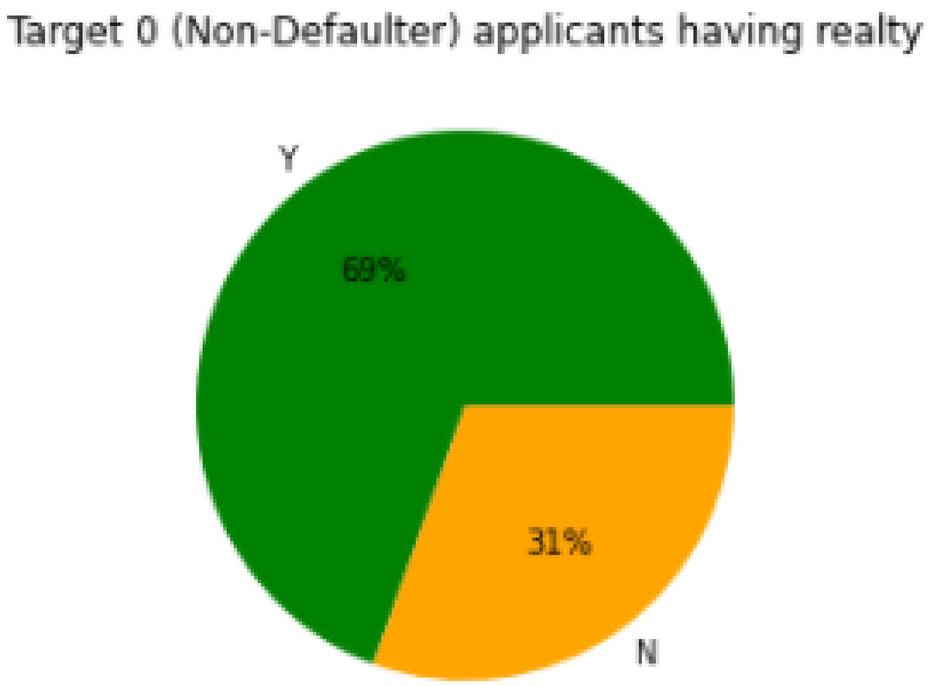


Both defaulters and non-defaulters have no significant difference in terms of car ownership and so this parameter should not be used for risk assessment.

1/3rd of both defaulters and non-defaulters don't own a car.

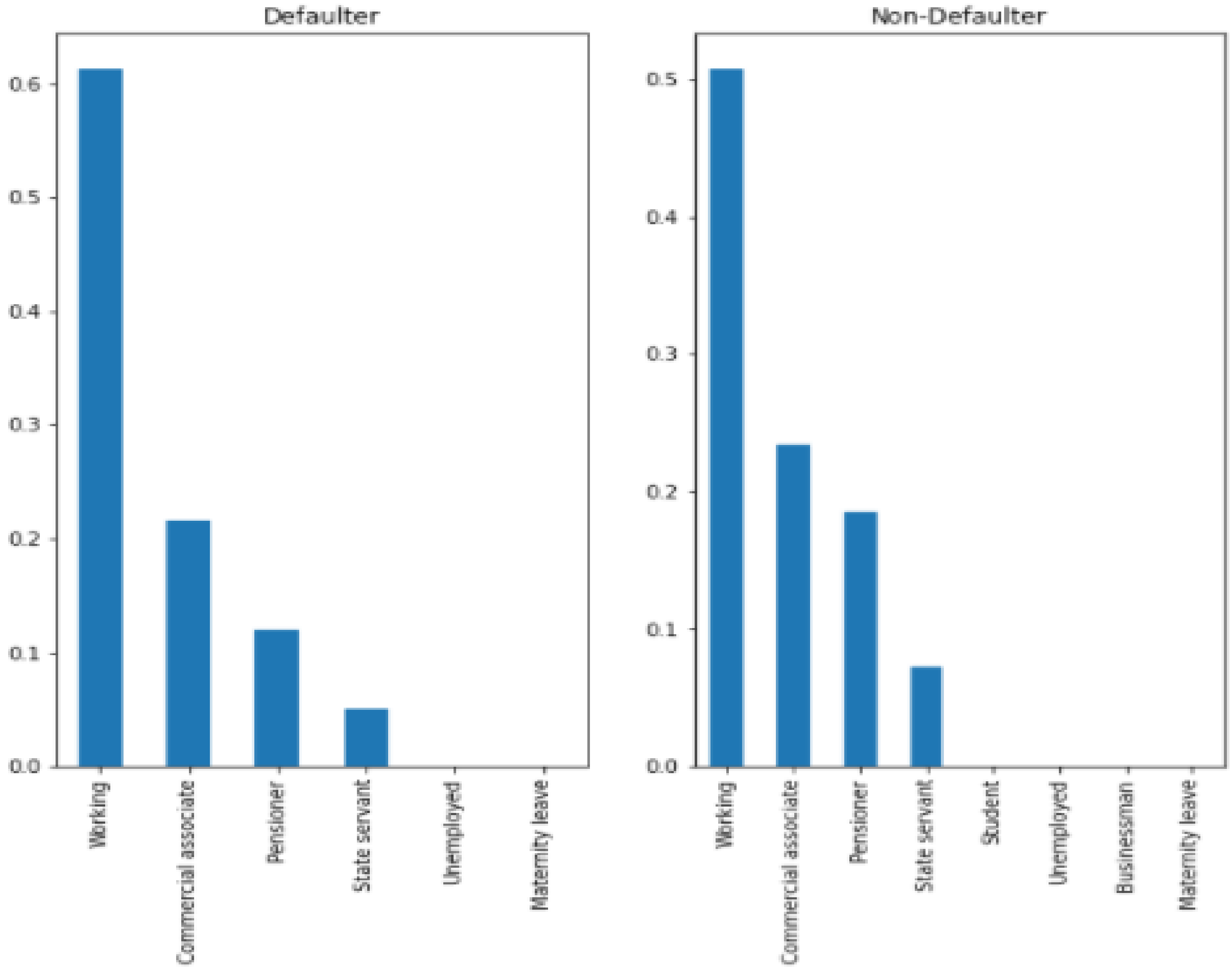
FLAG_OWN_REALTY





With the current data pattern, there is hardly any difference between defaulters and non-defaulters based on having realty.
Almost 1/3rd of both the defaulters and non-defaulters own a realty.

NAME_INCOME_TYPE



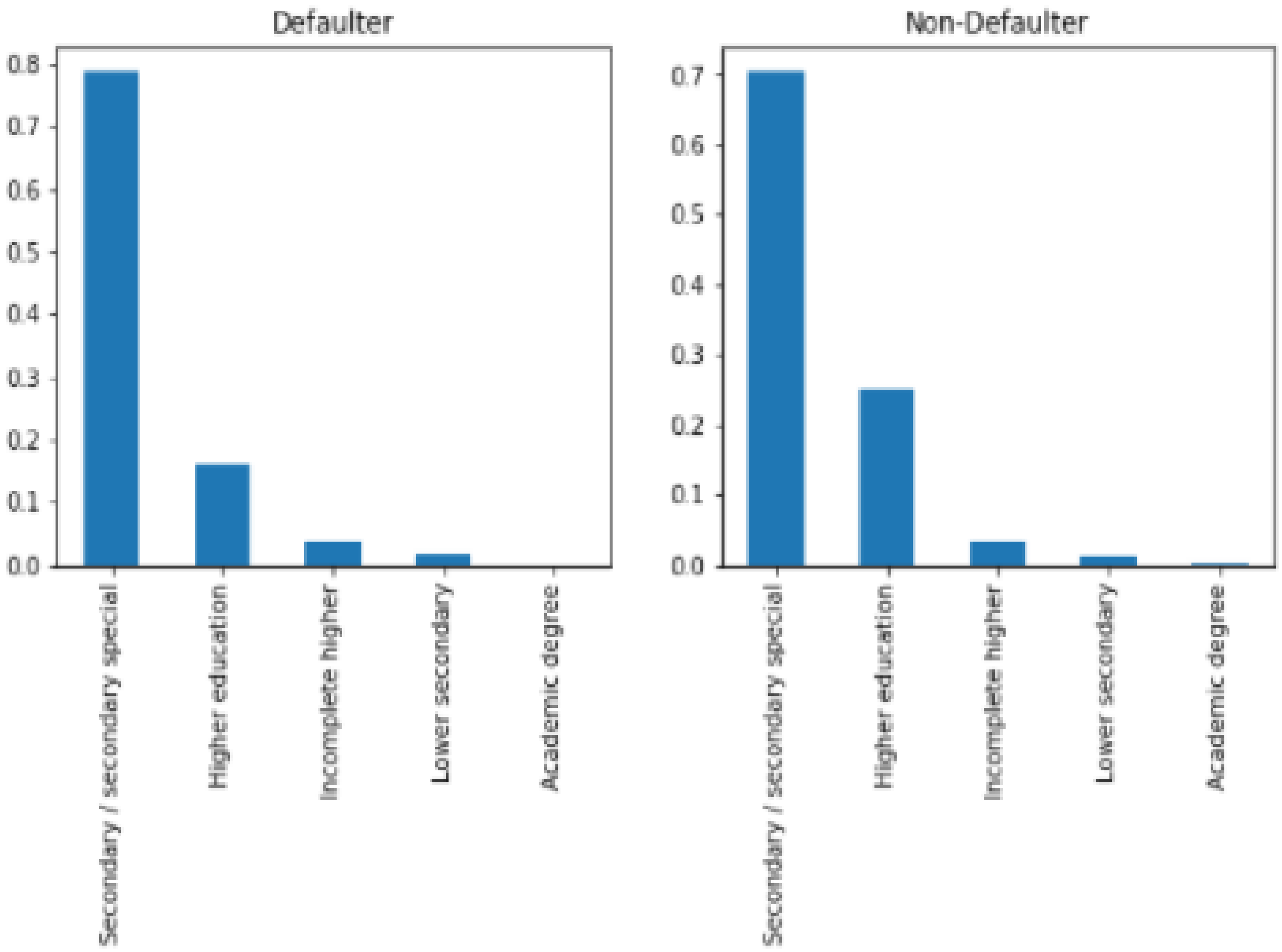
The majority of data set corresponds to working class.

The working class is on 1st rank both among defaulter and non-defaulter customers.

The commercial associates are better in the non-default category.

The Pensioners and state servants are better in non-defaulting category than defaulting category.

NAME_EDUCATION_TYPE



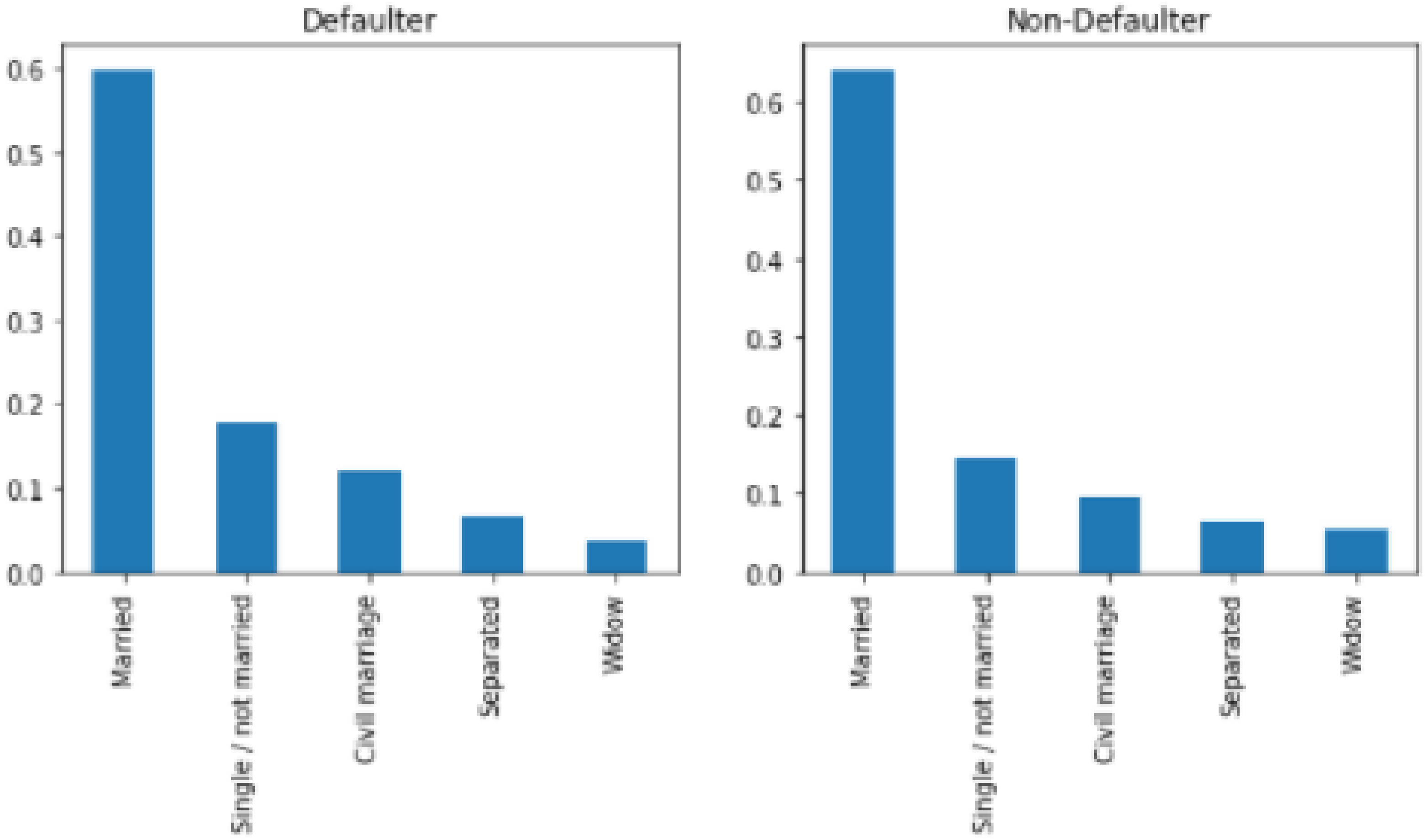
More than 3/4th defaulters and non-defaulters have secondary/secondary special education. This also underlines the fact that majority of applicants have secondary/secondary special education.

The percentage of non-defaulters with incomplete higher education is higher than defaulters with incomplete higher education.

The Higher education and lower secondary contribute very less among both defaulting as well as non-defaulting categories.

The Academic degree is insignificant for both the categories.

NAME_FAMILY_STATUS

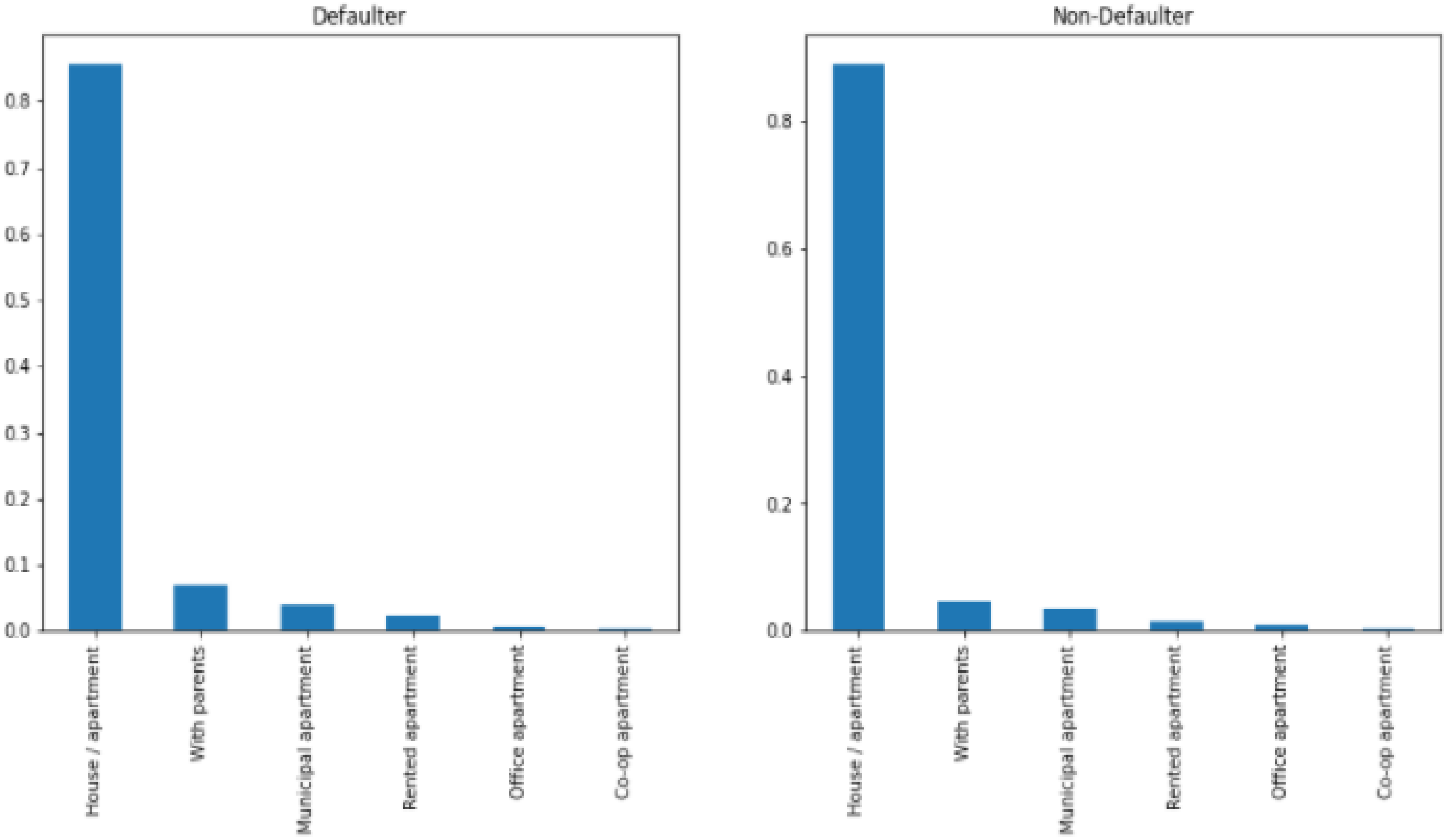


The ranking of defaulters and non-defaulters follow same sequence in terms of family status.

The Married and Single are top 2 categories in both the defaulter as well as non-defaulter categories.

The separated and widow are less than 10% in both the defaulter as well as non-defaulter categories.

NAME_HOUSING_TYPE

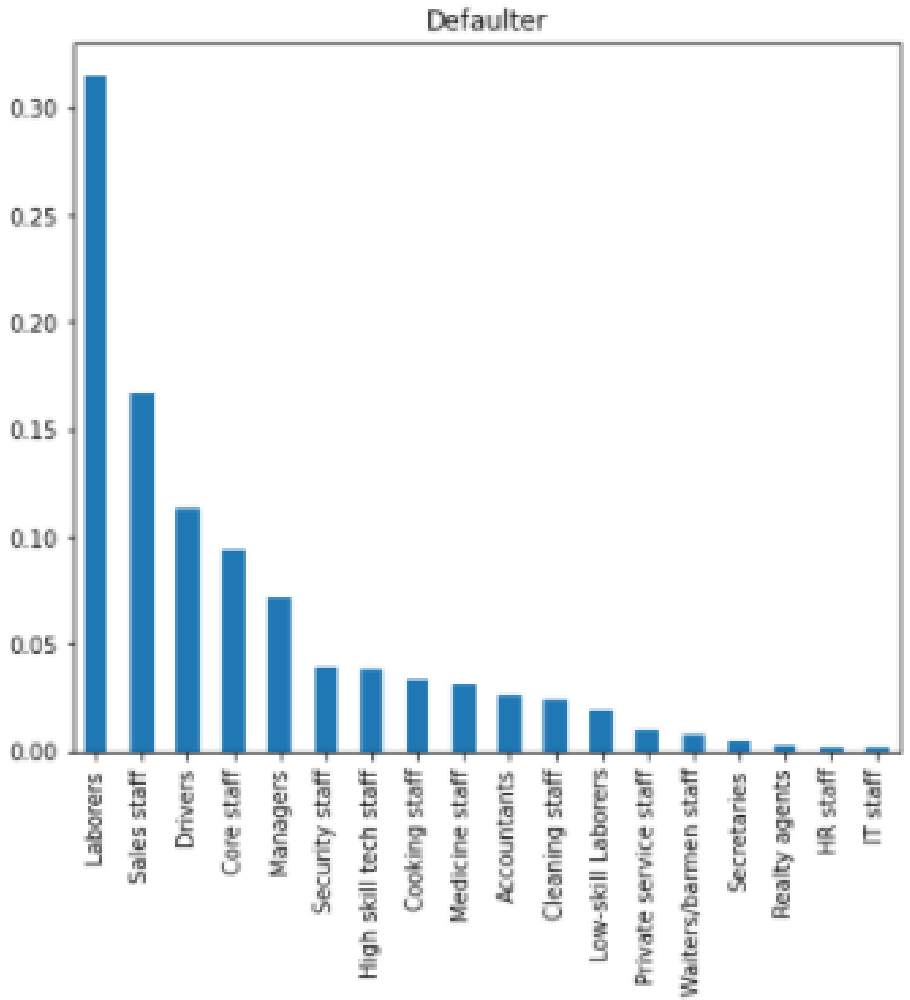


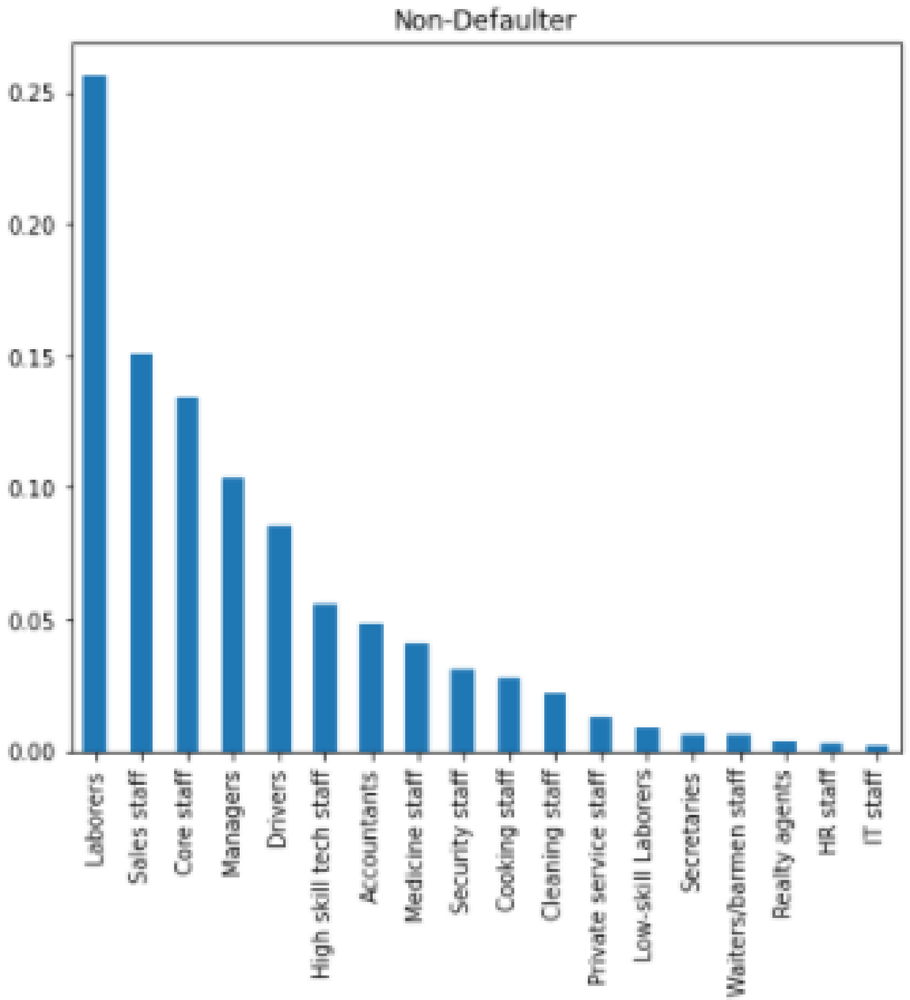
The Housing type is apartment for more than 80% of applications for both the defaulter as well as non-defaulter categories.

The all other Housing type except apartments are less than 10% for both the defaulting as well as non-defaulting categories.

The sequence of house type in terms of ranking is same in both the defaulting and non-defaulting categories.

OCCUPATION_TYPE



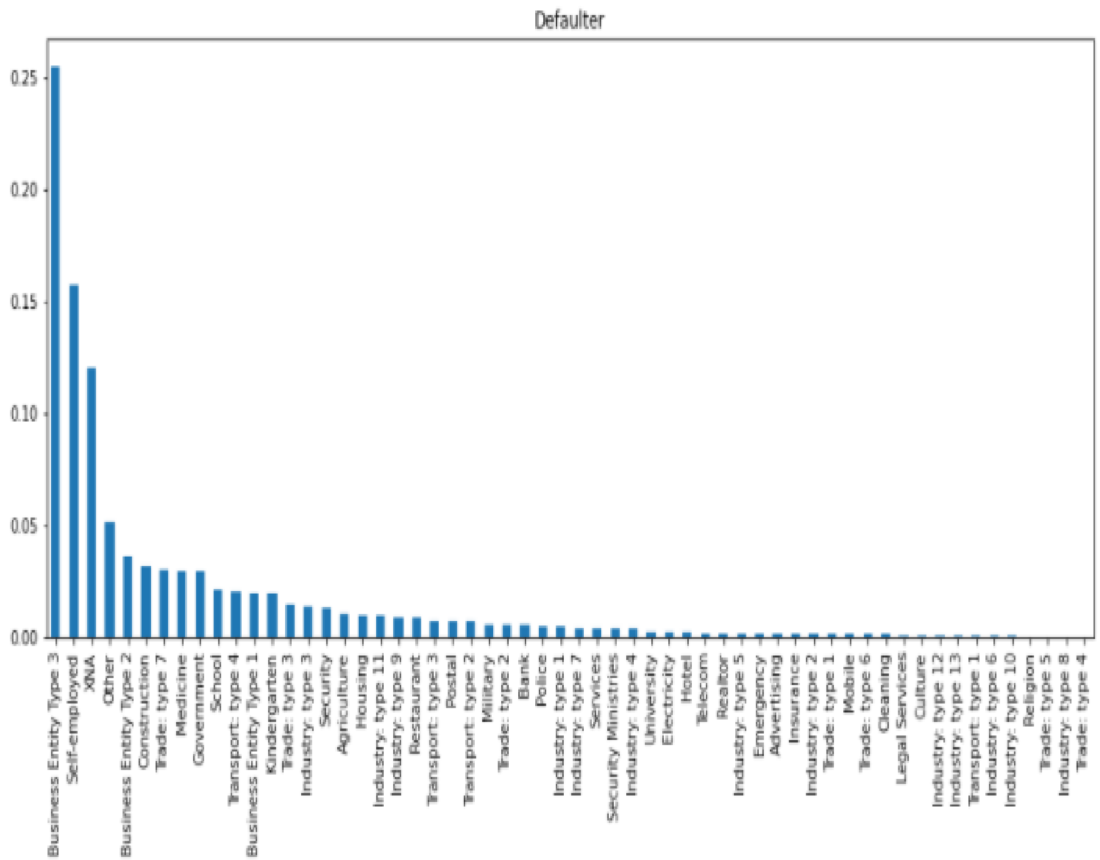


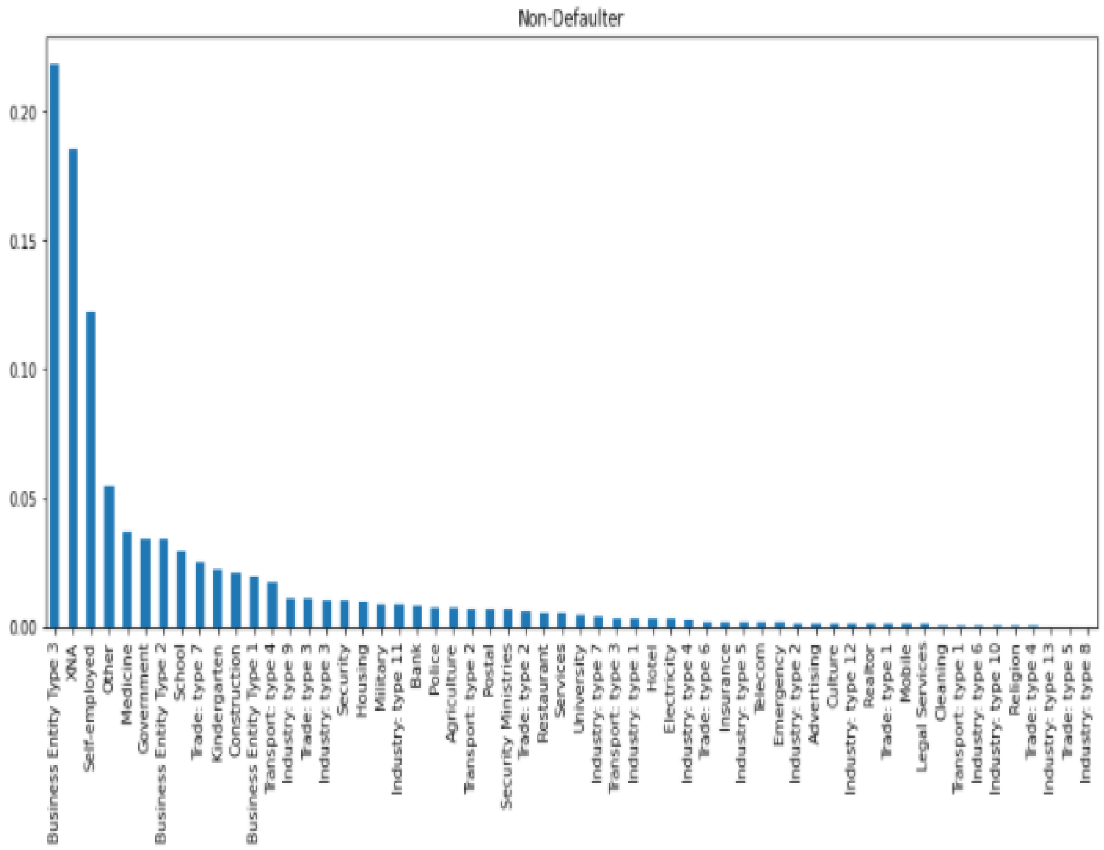
The default percentage is highest among laborers, sales staff and drivers with all these exceeding 10%.

The drivers have more rate of default than the non-default.
IT staffs are most secured in terms of lending.

There are 18 types of occupation among both the defaulters as well as non-defaulters.

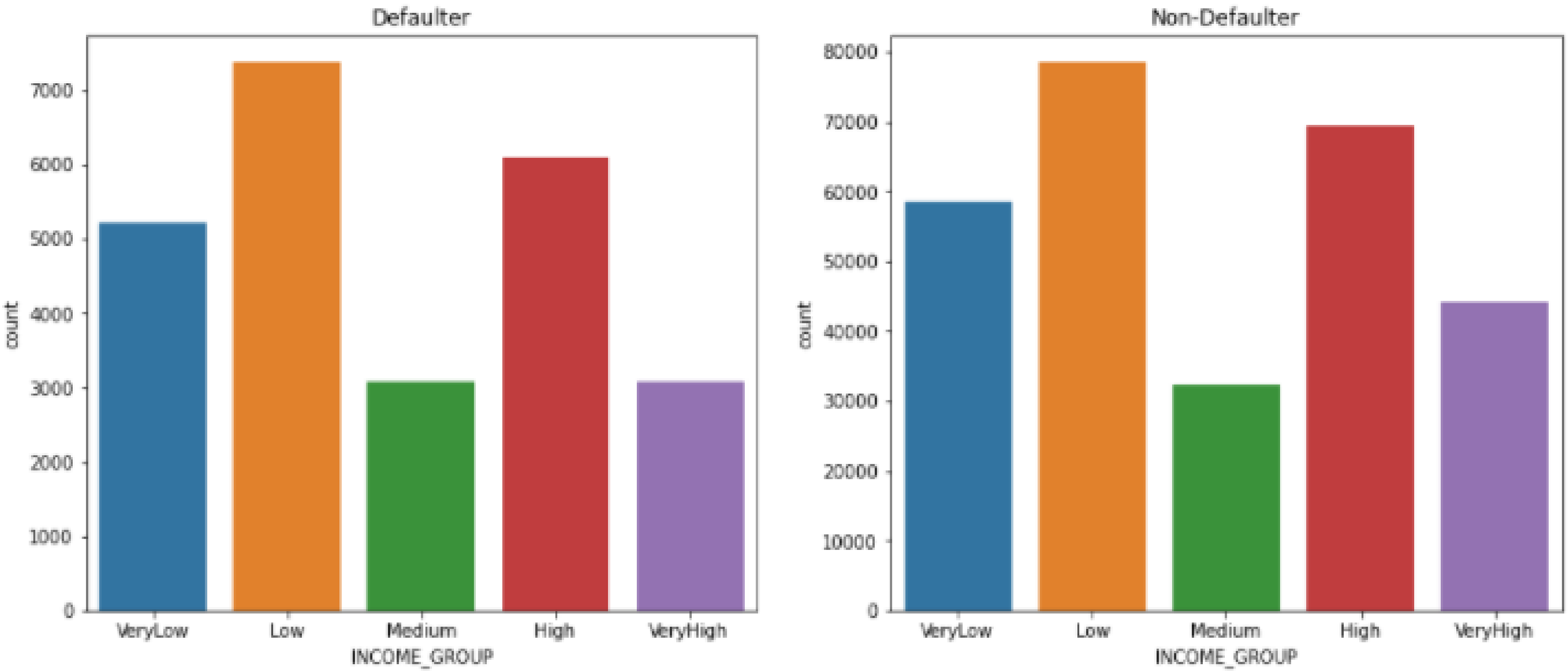
ORGANIZATION_TYPE





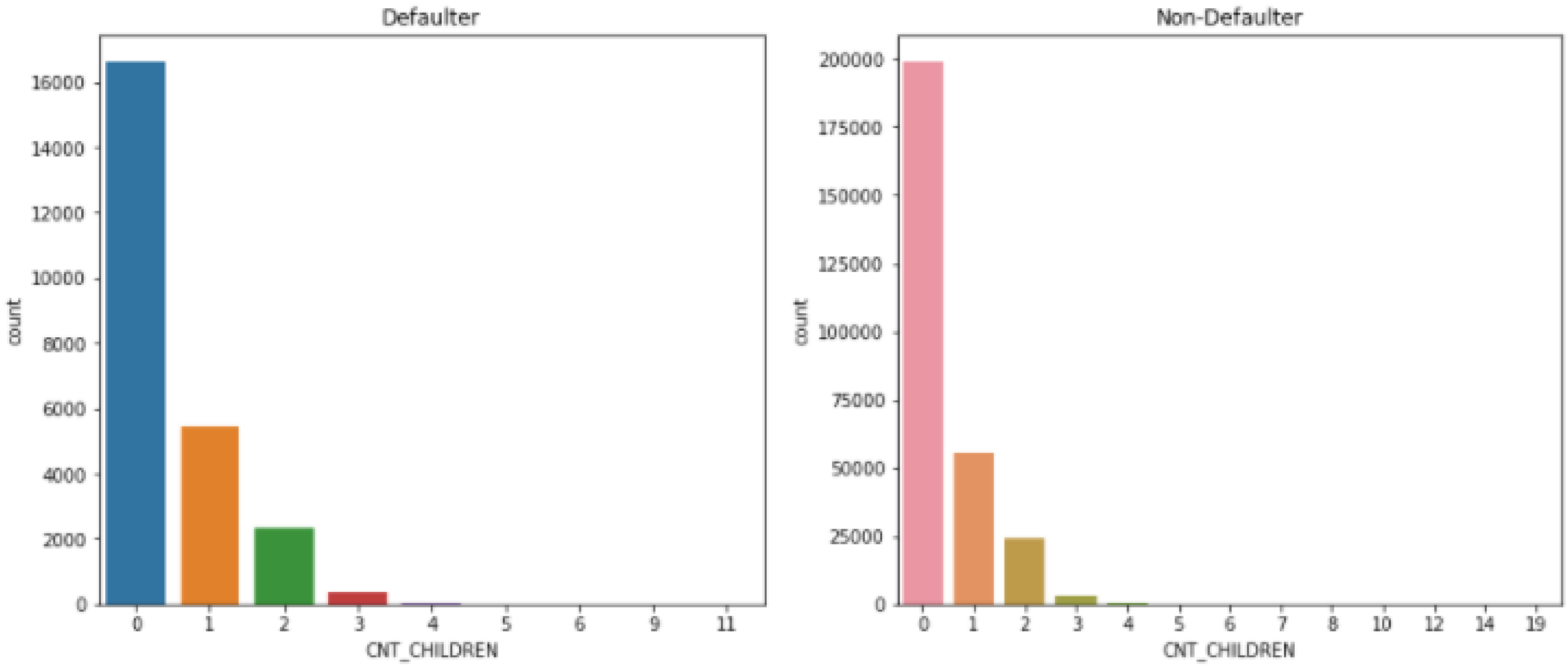
The Business Entity Type 3 is top most category in both the defaulters and non-defaulters exceeding 20% in both of them.
The self employed have more defaulting rate than non-defaulting.
The top 3 categories in both the defaulters and non-defaulters are same but their sequence is different.

INCOME_GROUP



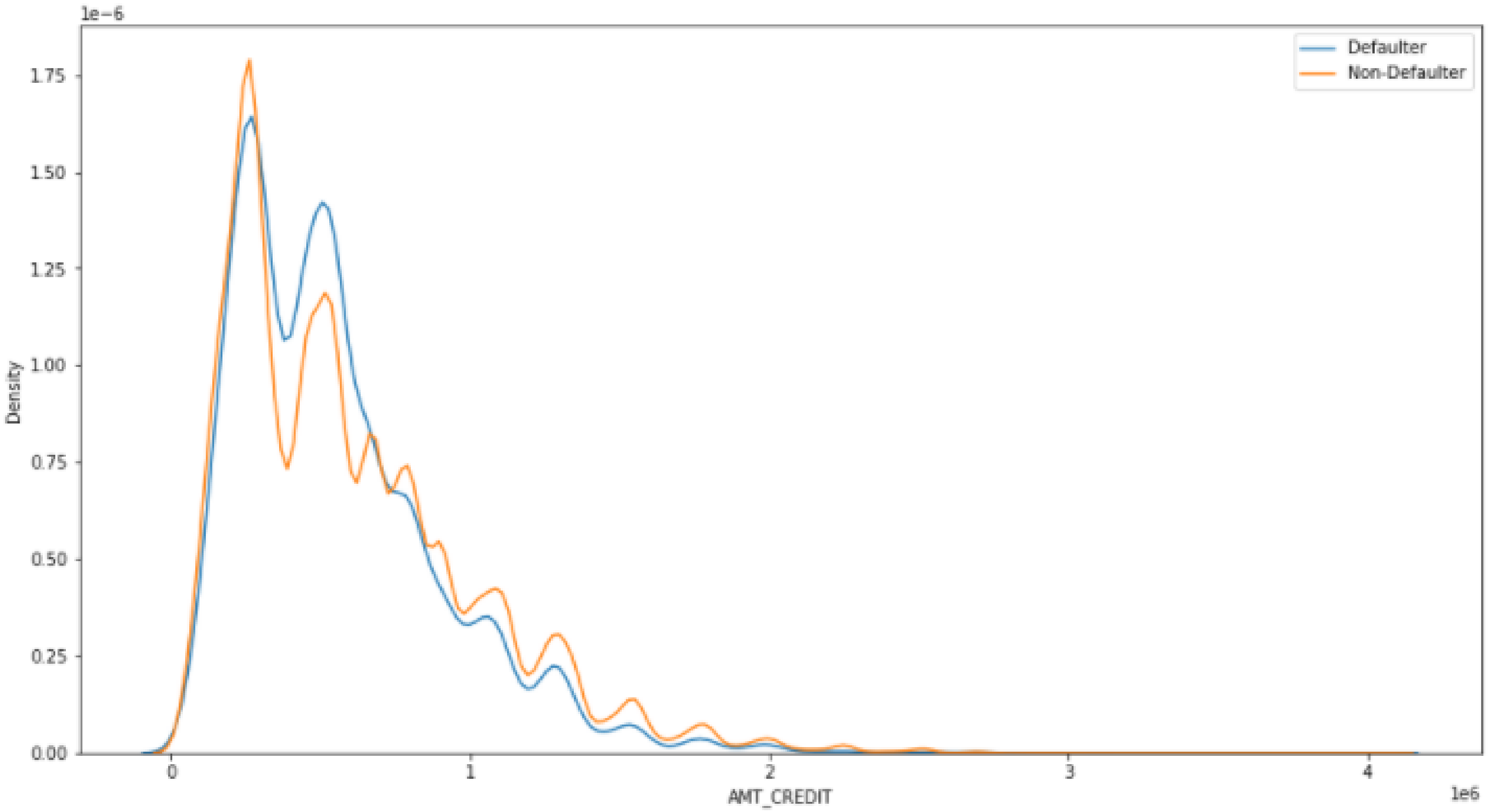
The majority of loan applicants are from "low" income group.
The medium income group seems to be most adverse to the concept of borrowing as their % is less in both defaulting and non-defaulting.

CNT_CHILDREN



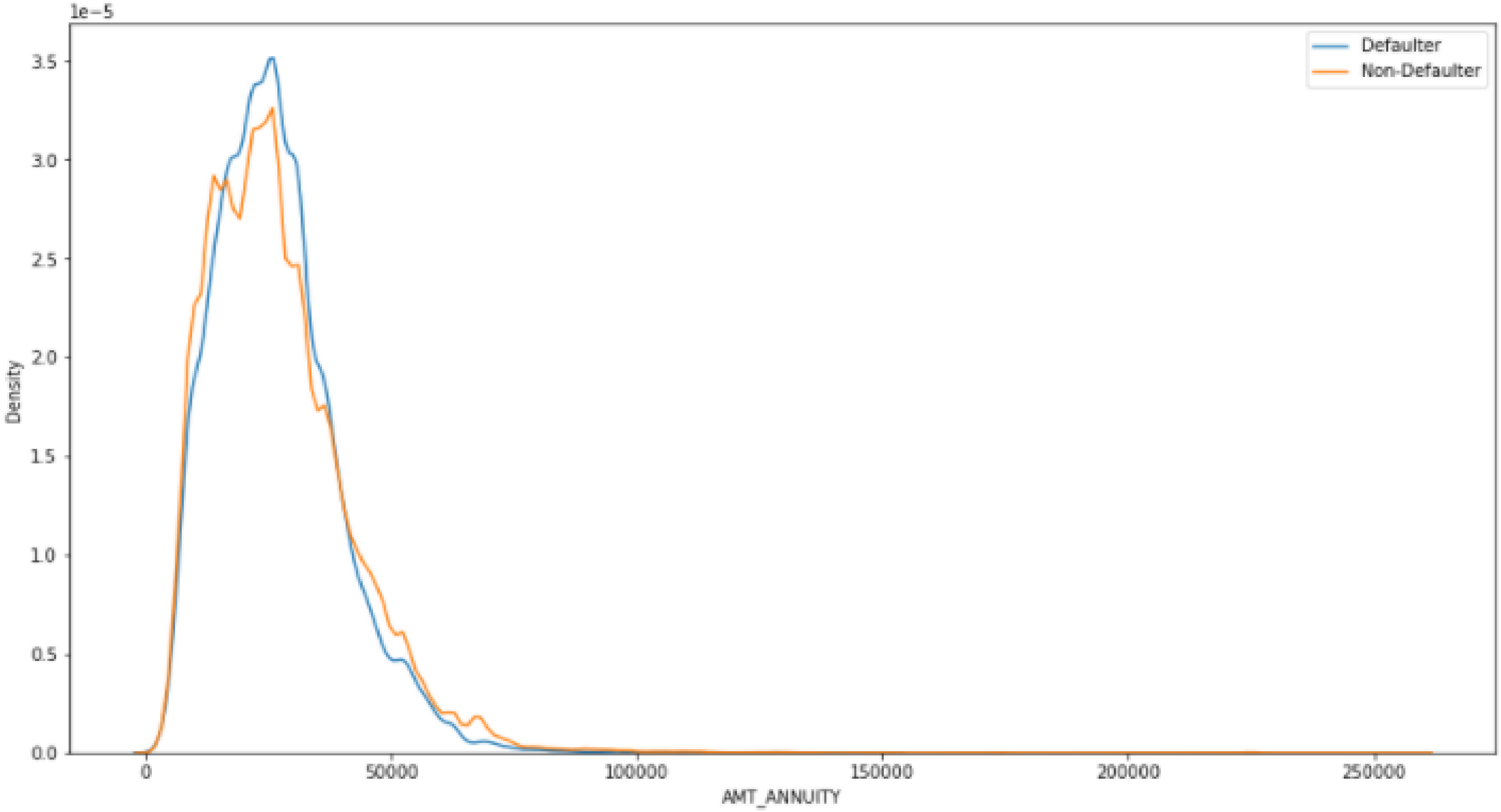
The no of children among defaulting and non-defaulting varies from 0-4 with the majority in both the categories have no child.
As both the defaulting and non-defaulting are leaned towards zero children, there can be non conclusive remarks on defaulters/ non-defaulters based on number of children.

AMT_CREDIT



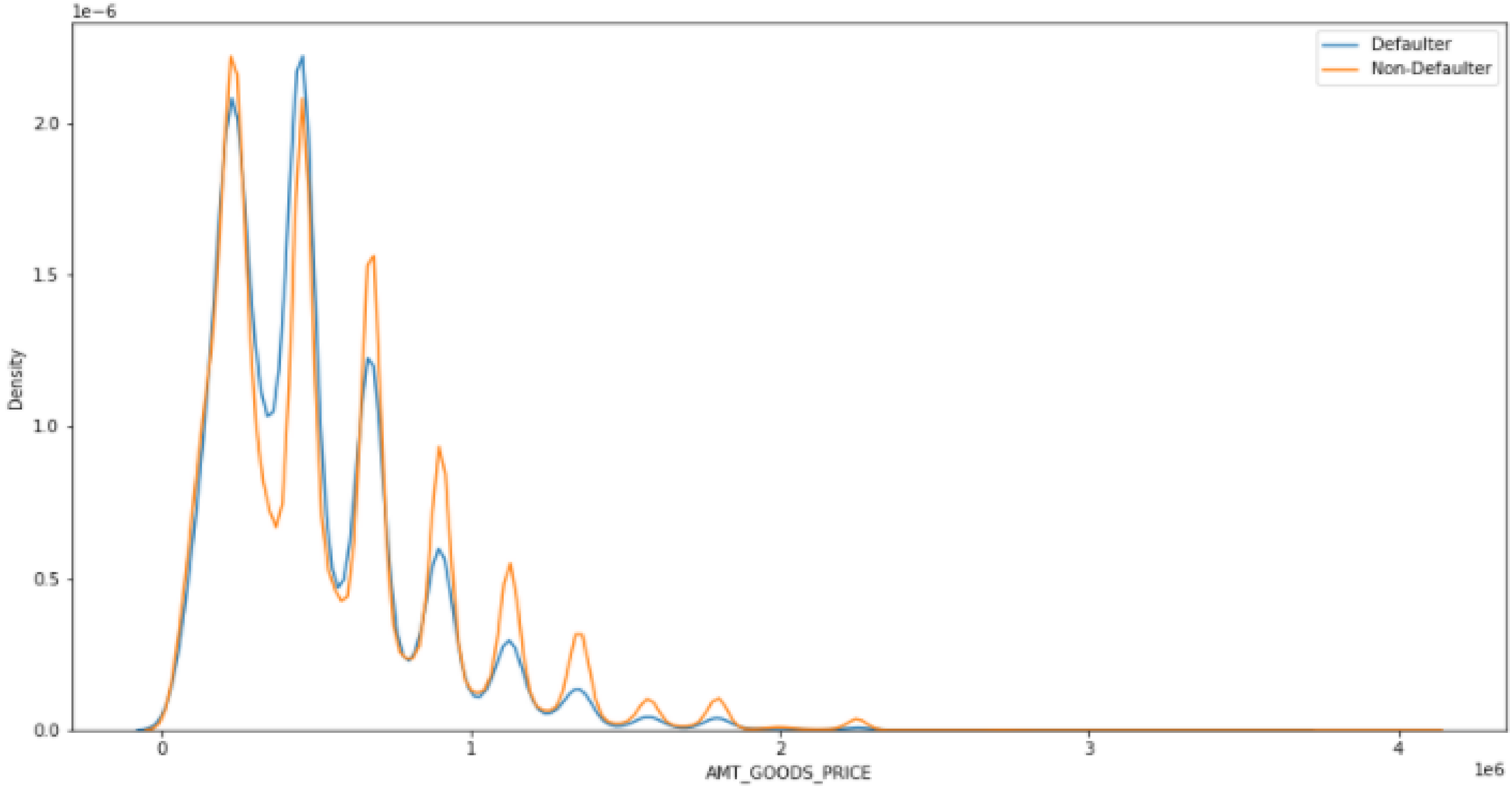
The Average amount credited is more for non defaulters than defaulters.
The minimum and maximum amount of loan credited are similar in both the cases.

AMT_ANNUITY



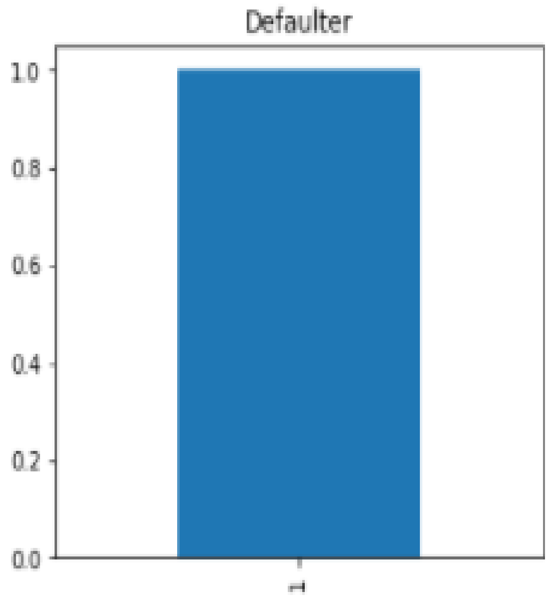
The Average annuity amount is more for non defaulters
Both the curves attain peak at similar annuity amount but defaulter has higher density.

AMT_GOODS_PRICE

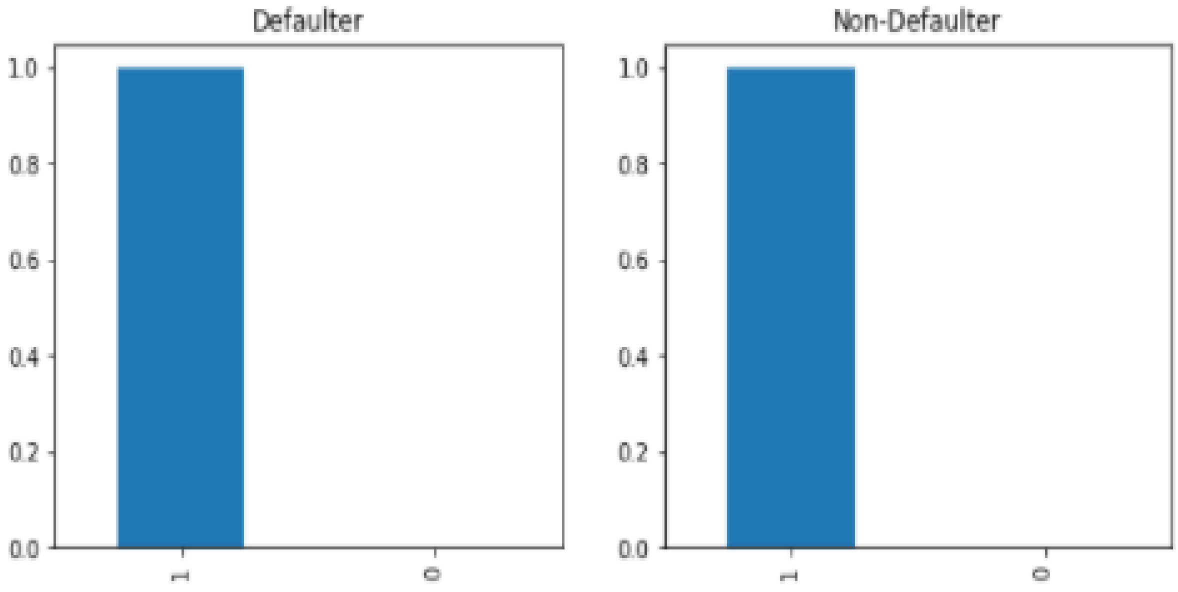


The Average goods price amount is more for non defaulters
At most data points, the non-defaulters have higher value of density compared to defaulters

FLAG_MOBIL

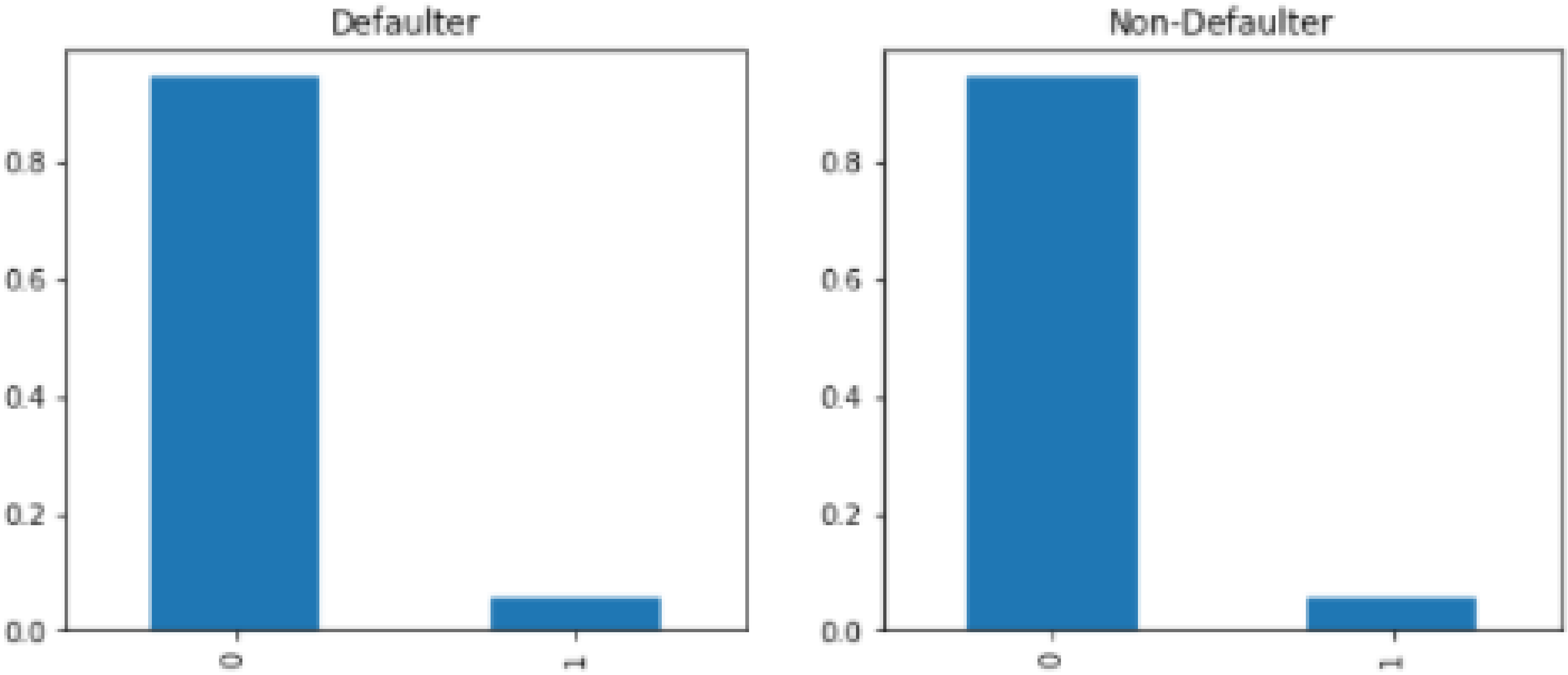


FLAG_CONT_MOBILE



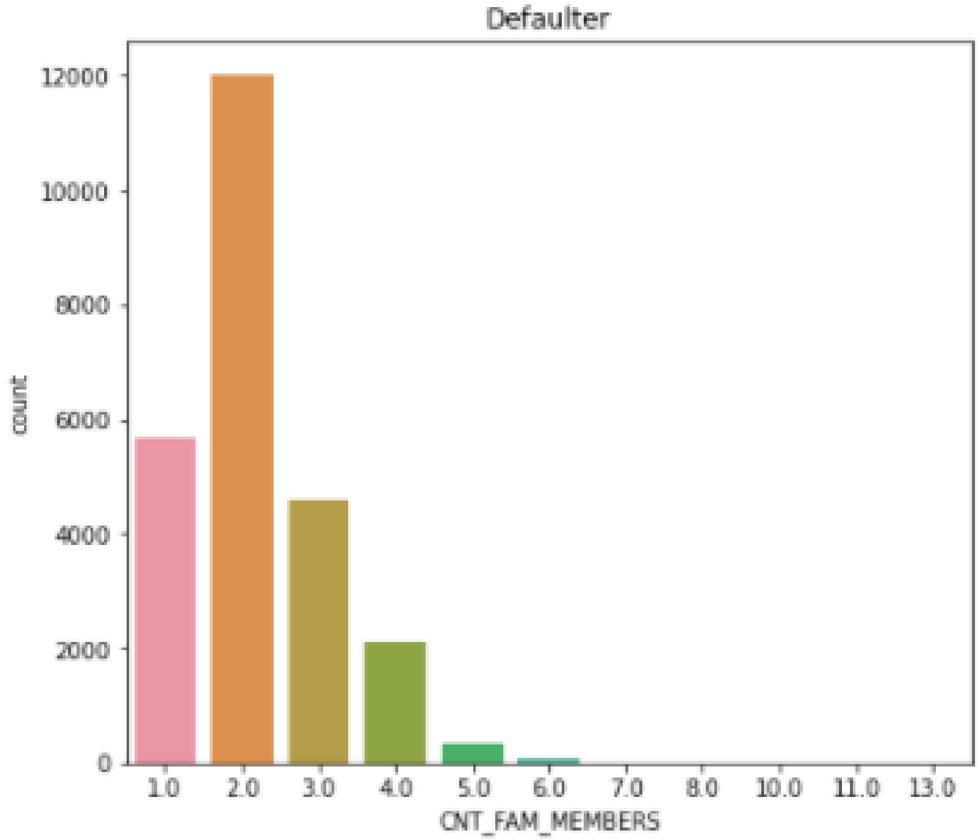
There is no significant value difference observed between defaulters and non-defaulters on the basis of accessibility of mobile phones.

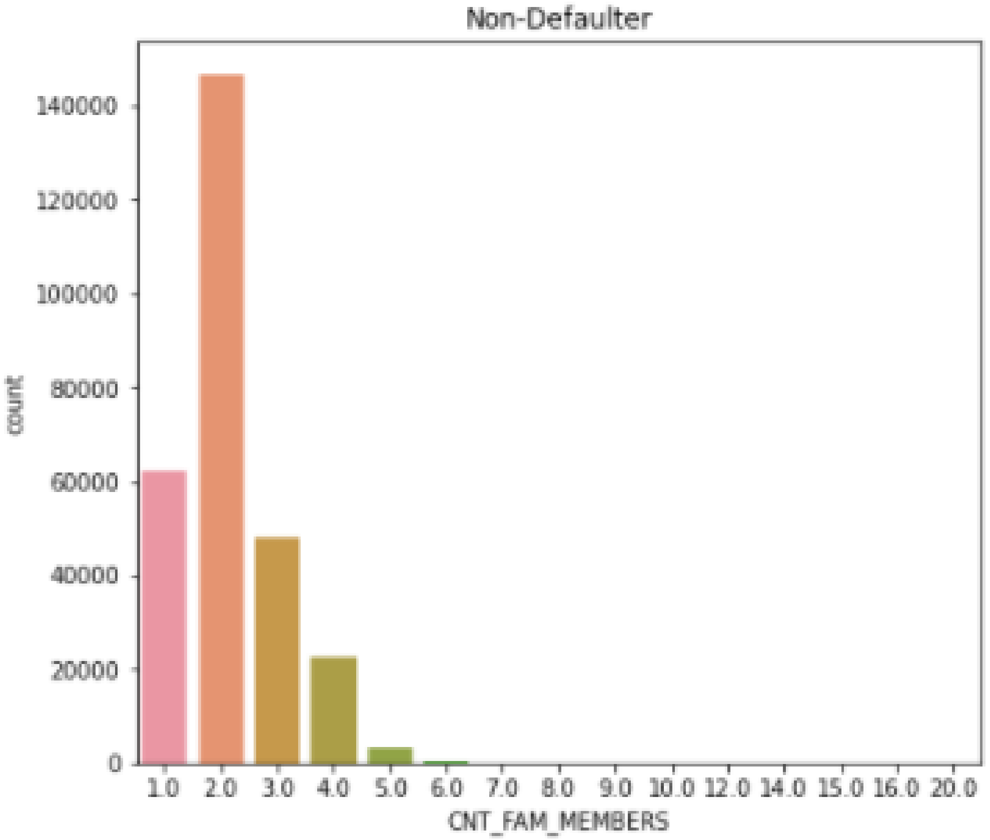
FLAG_EMAIL



The data reveals that more than 90% applicants in both the defaulter and non-defaulter did not provide the email id.

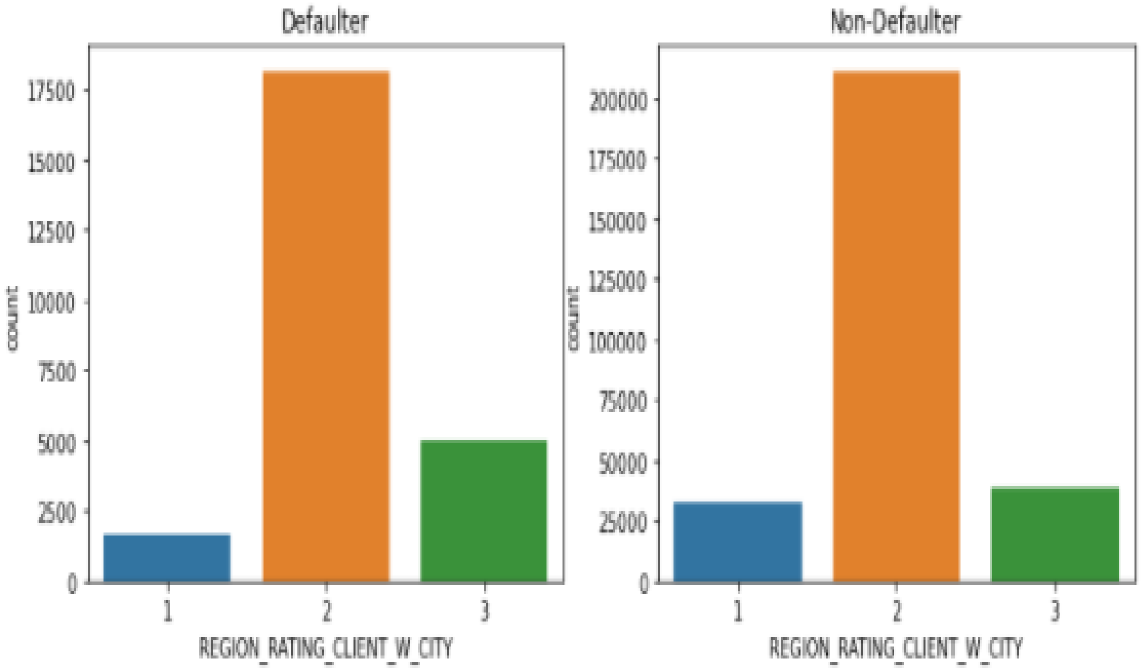
CNT_FAM_MEMBERS





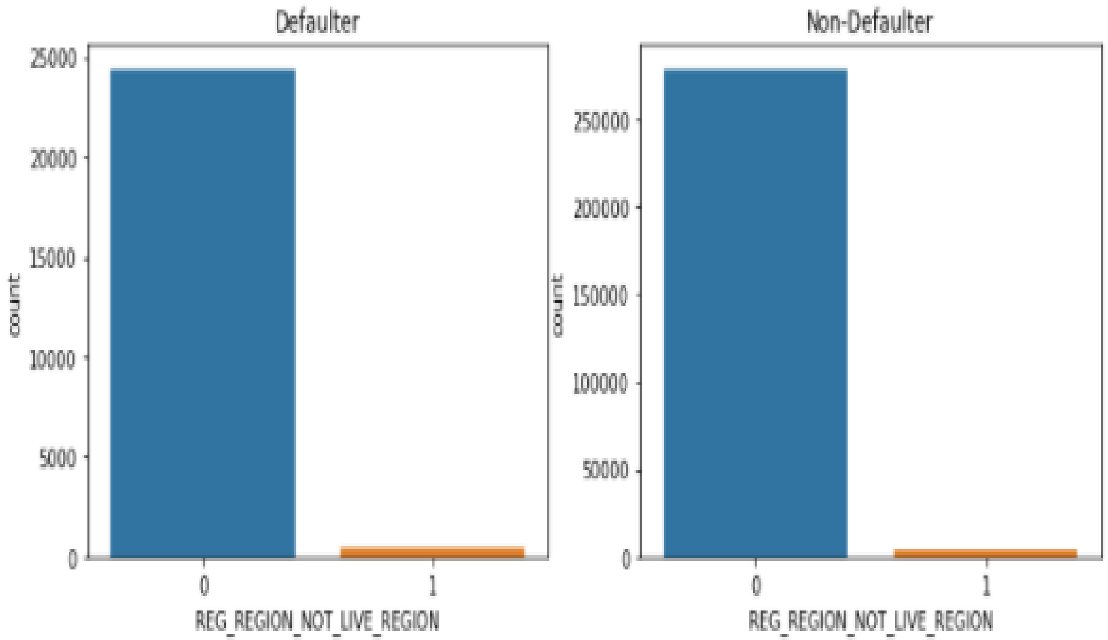
The count of family members is 2 for majority of the defaulting and non-defaulting loan applicants.
Count of family members greater than 5 is insignificant for both the categories.
The sequence of count of family is same for both the defaulting and non-defaulting categories.

REGION_RATING_CLIENT_W_CITY



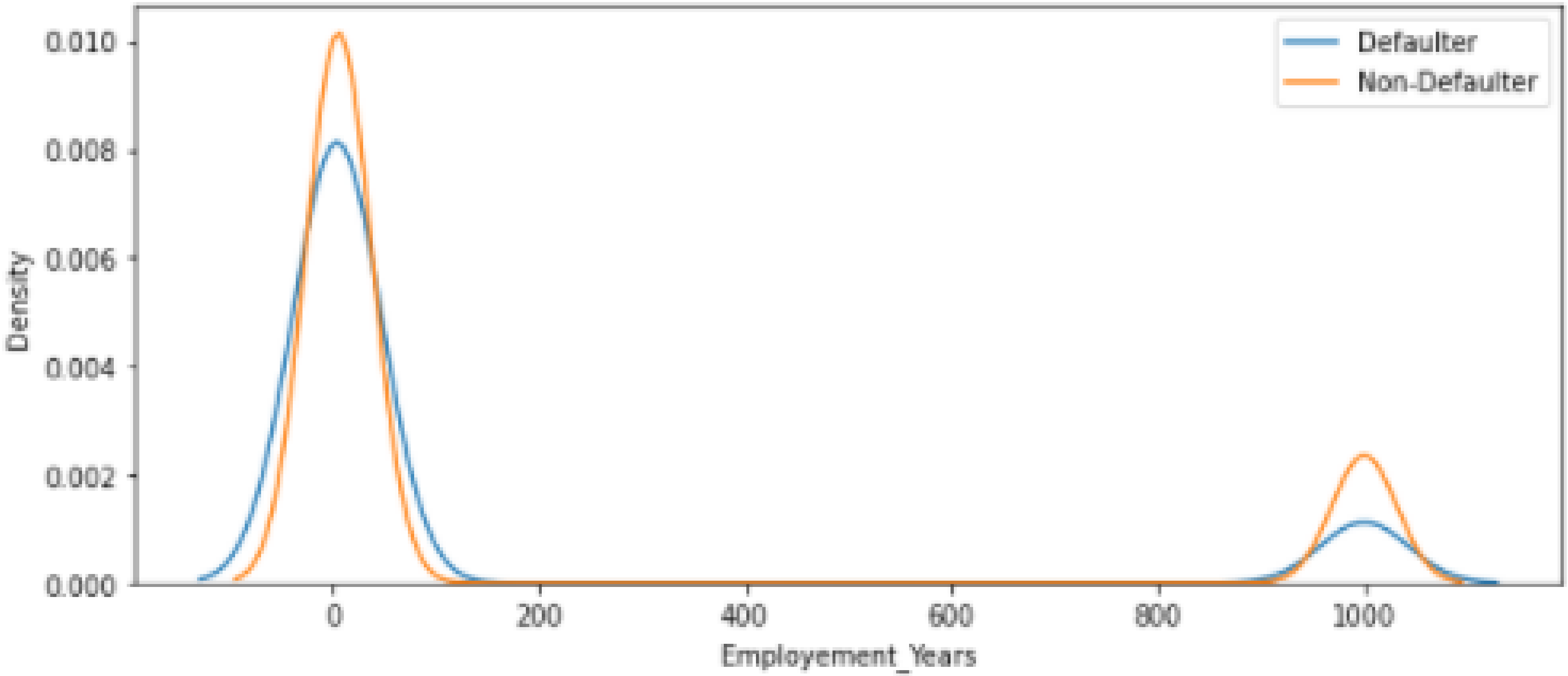
Clients with Region rating = 1 are comparatively more likely to be non defaulters
Clients with Region rating = 3 are comparatively more likely to be defaulters.

REG_REGION_NOT_LIVE_REGION



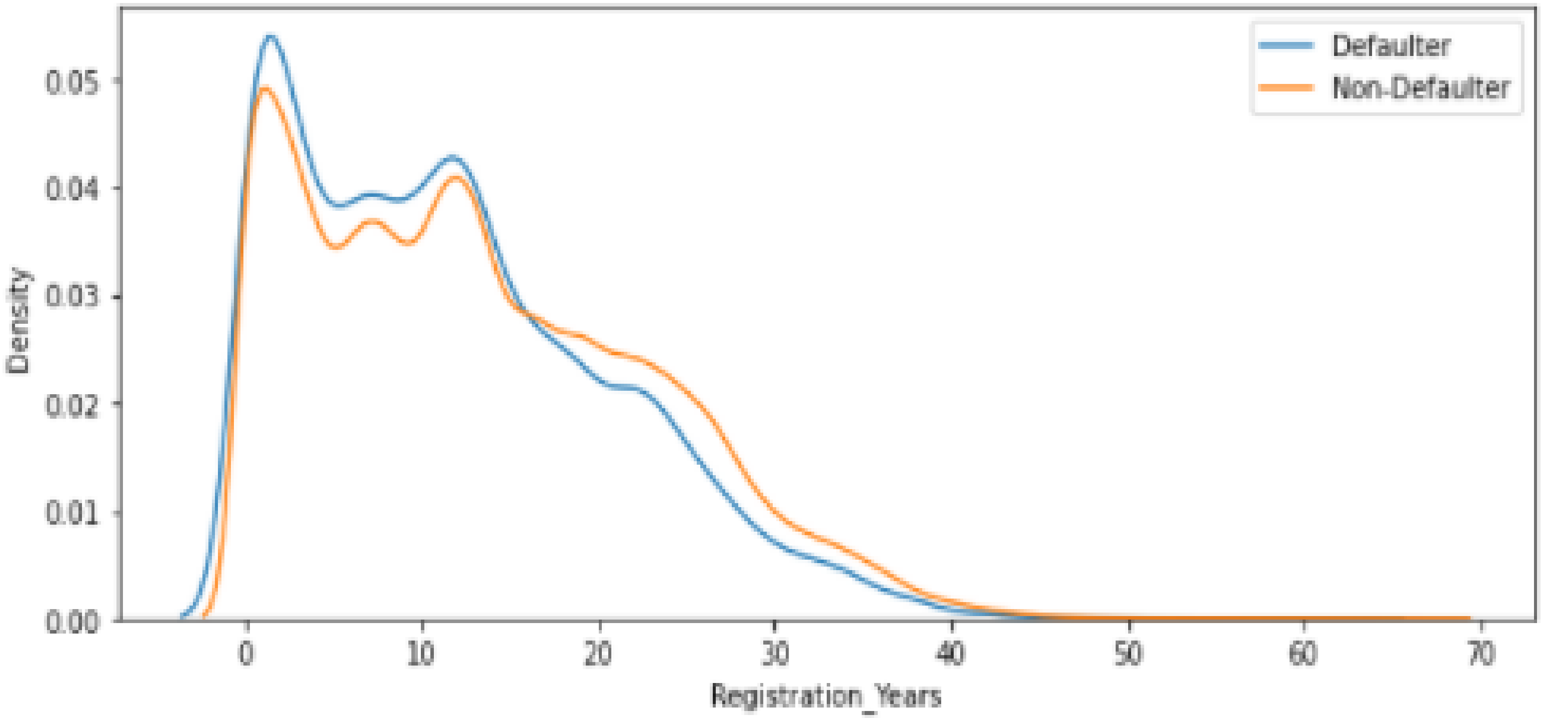
There is no significant value difference observed in between defaulters and non-defaulters on the basis of their permanent address and contact address

Employement_Years



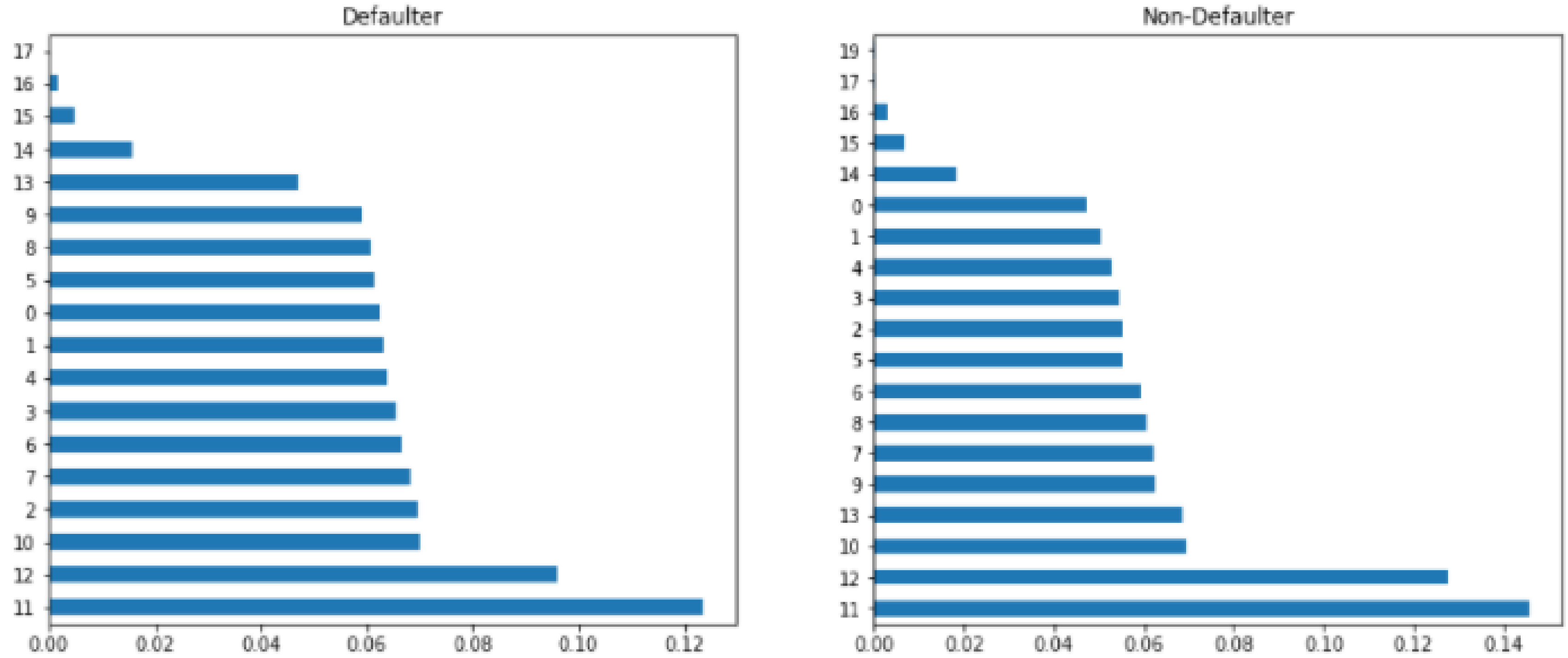
The non defaulter clients have a greater number of years of employment than defaulters
The Average value of number of years employed is more for non defaulters

Registration_Years



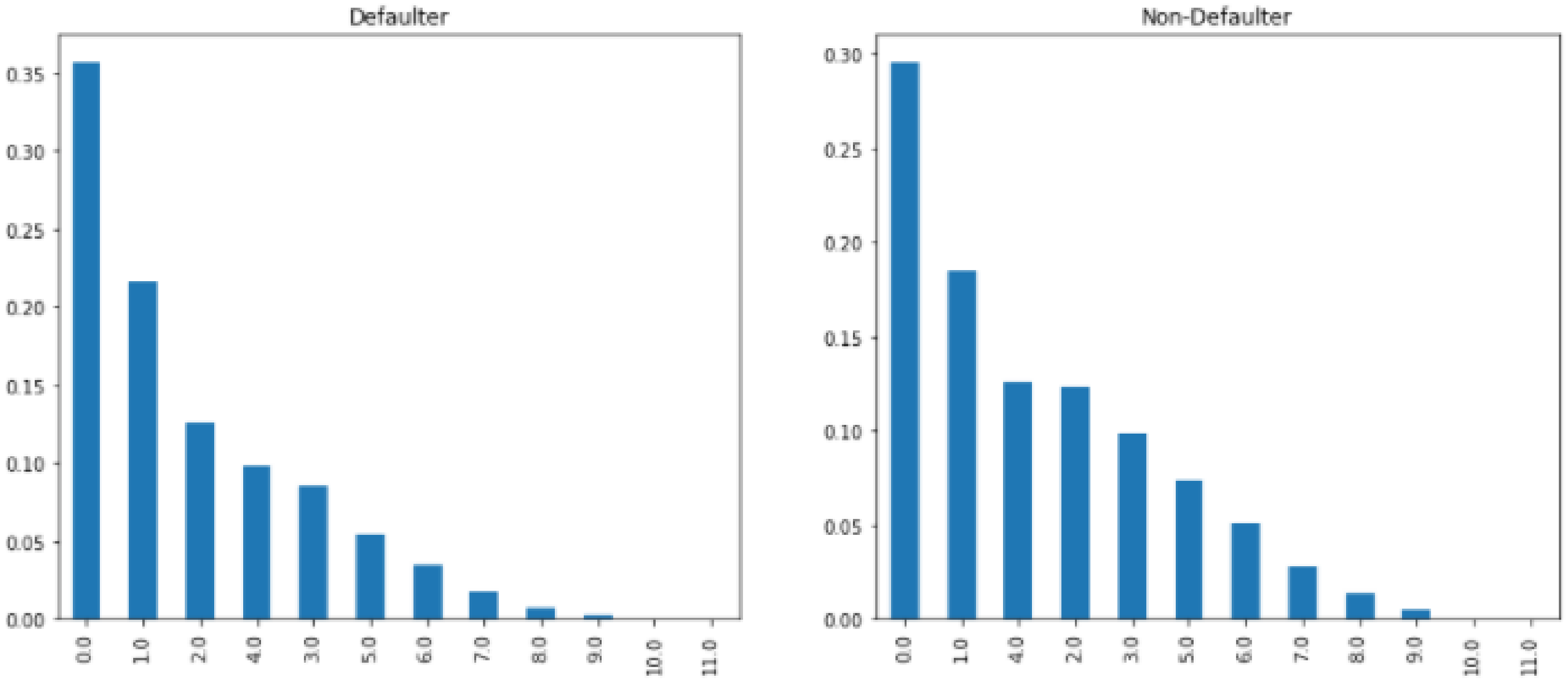
The density for registration years for both defaulter and non-defaulters converge at zero for values greater than 40 years.
The density for defaulters have higher value for 0-15 years
The density for non defaulters have higher value for 15-40 years

ID_Years



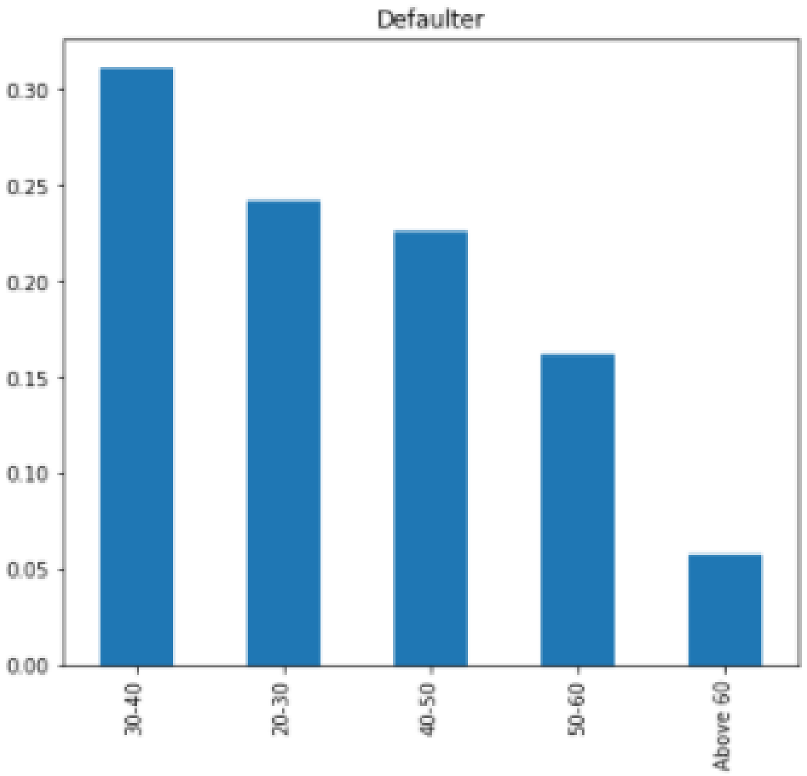
The Id years for 2-13 range is less than 6% .
The maximum Id years is 11 for both the defaulting and non-defaulting years.

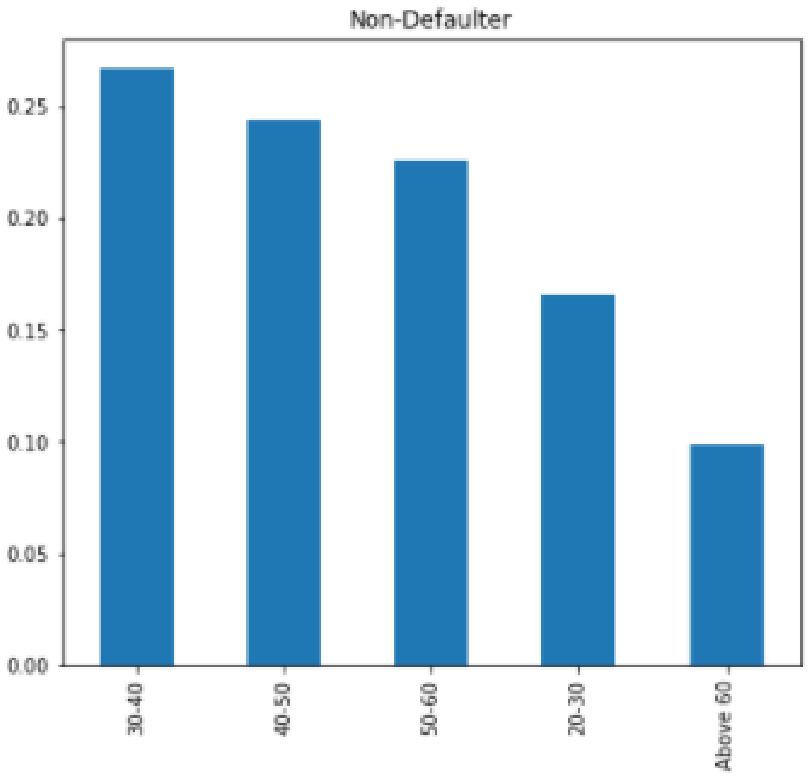
YEARS_LAST_PHONE_CHANGE



The majority of defaulters and non defaulters have changed their phone within 1 year.
Almost half of the applicants have changed their phone no in last 1 year for both the defaulting and non-defaulting categories.

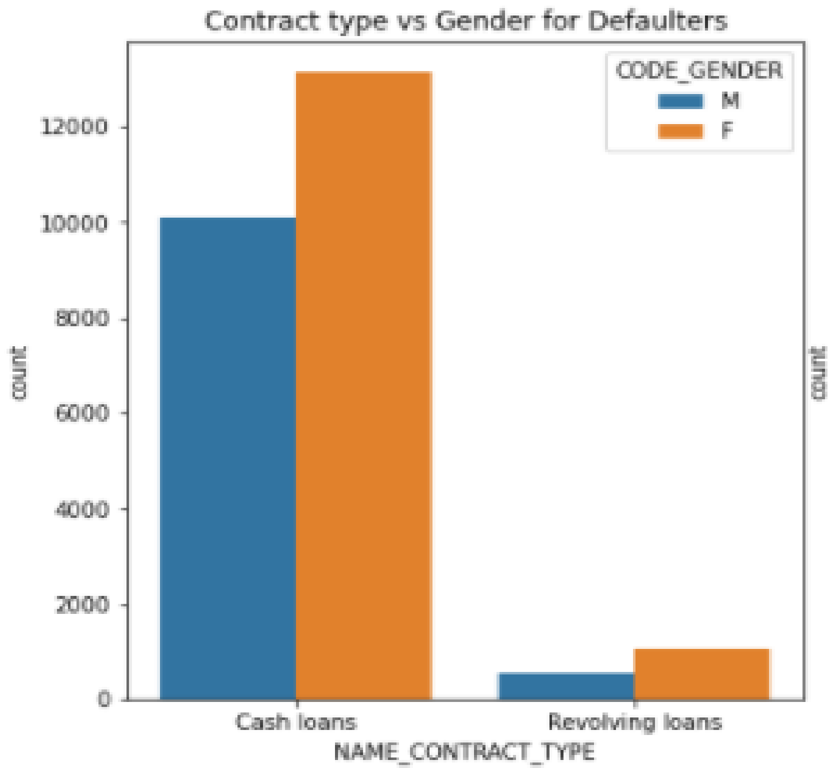
AGE_GROUP

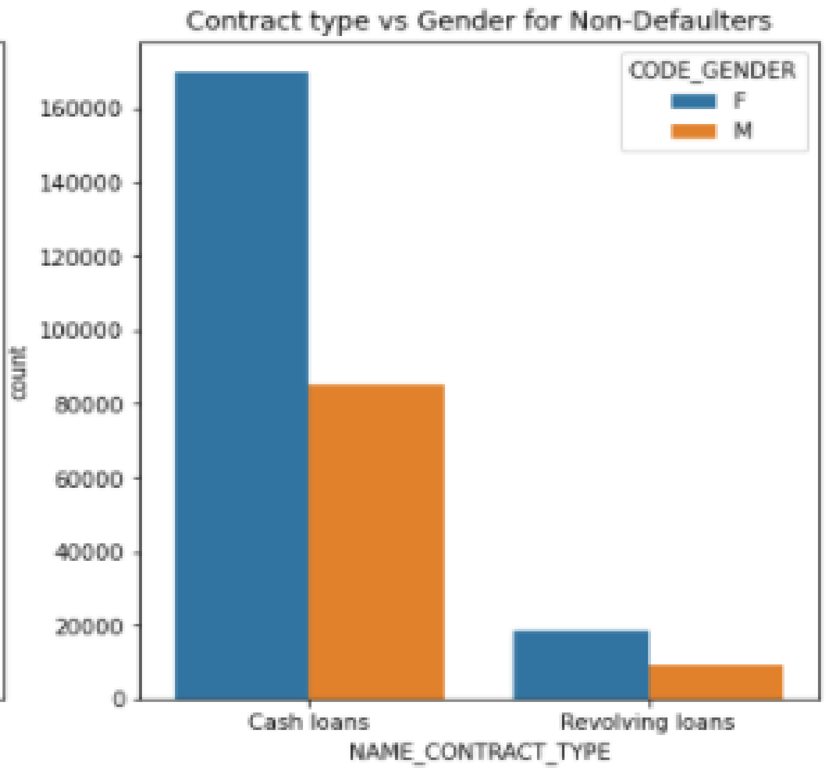




The Age Group 30-40 consist of majority of loan applicants.
The 20-30 age group applicants are more likely turn to be defaulters.
The above 60 age group is the only category falling below 5% among defaulters.
More than 50% of defaulter as well as non-defaulter comes from 2 age groups i.e. 30-40 and 20-30.

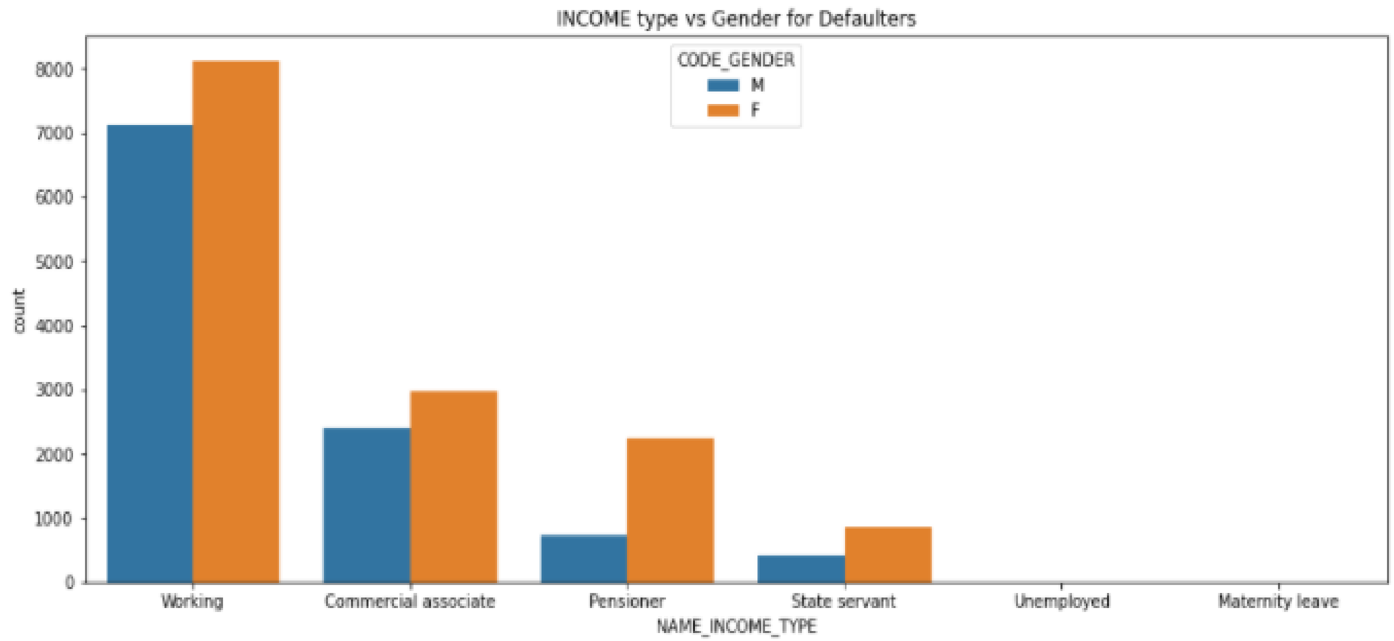
NAME_CONTRACT_TYPE VS CODE_GENDER

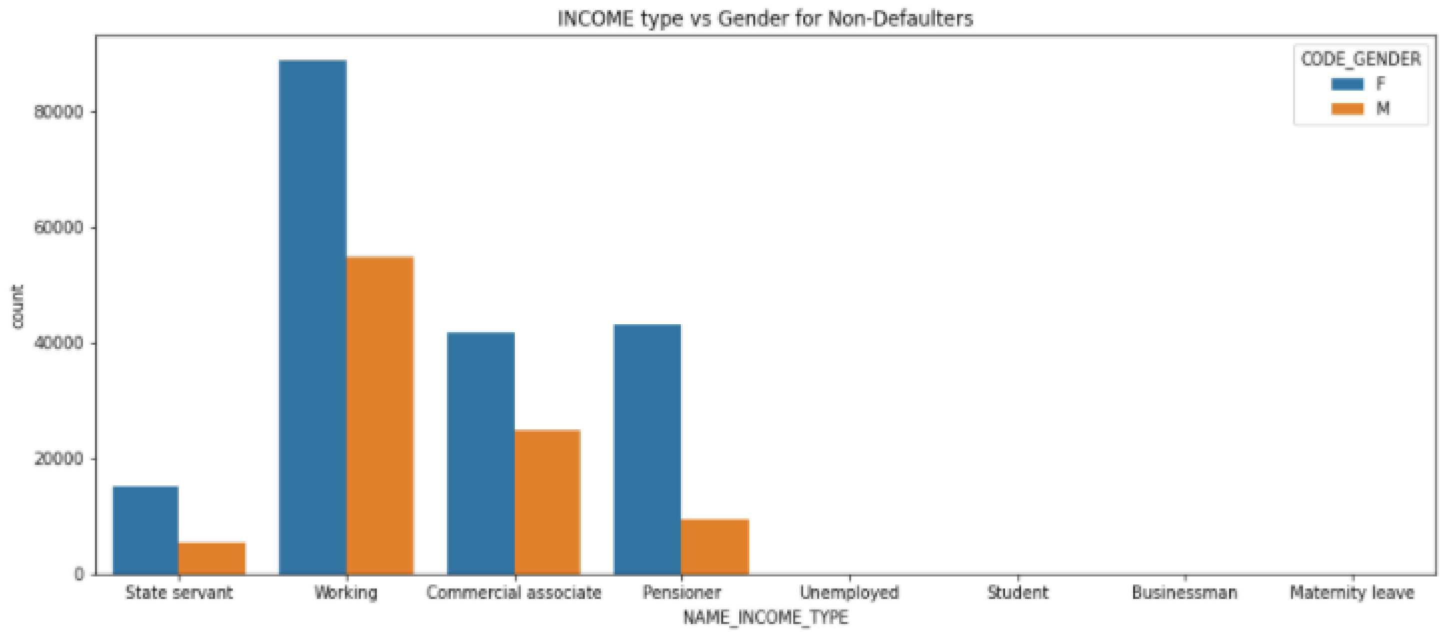




Cash loans consist of bulk of loans as per the data set.
Females taking cash loans are more prone to default than males.
The difference between males and females in cash loans in defaulting is more than non-defaulting categories.

NAME_INCOME_TYPE VS CODE_GENDER



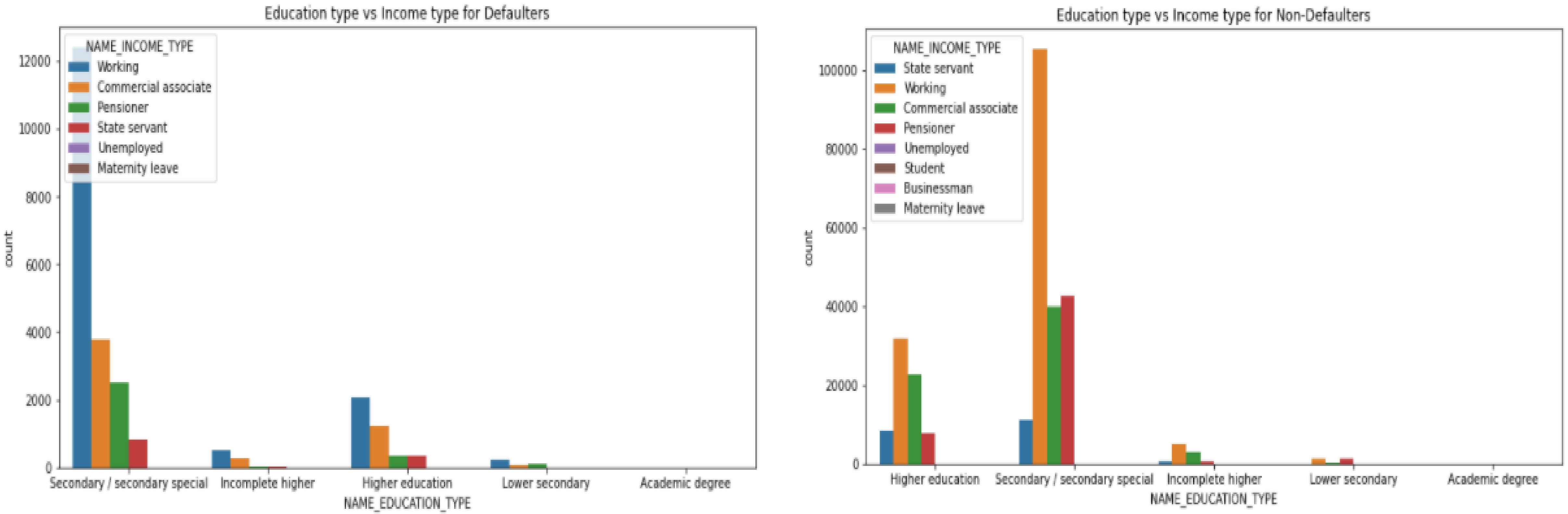


Cash loans consist of bulk of loans as per the data set.

Females taking cash loans are more prone to default than males.

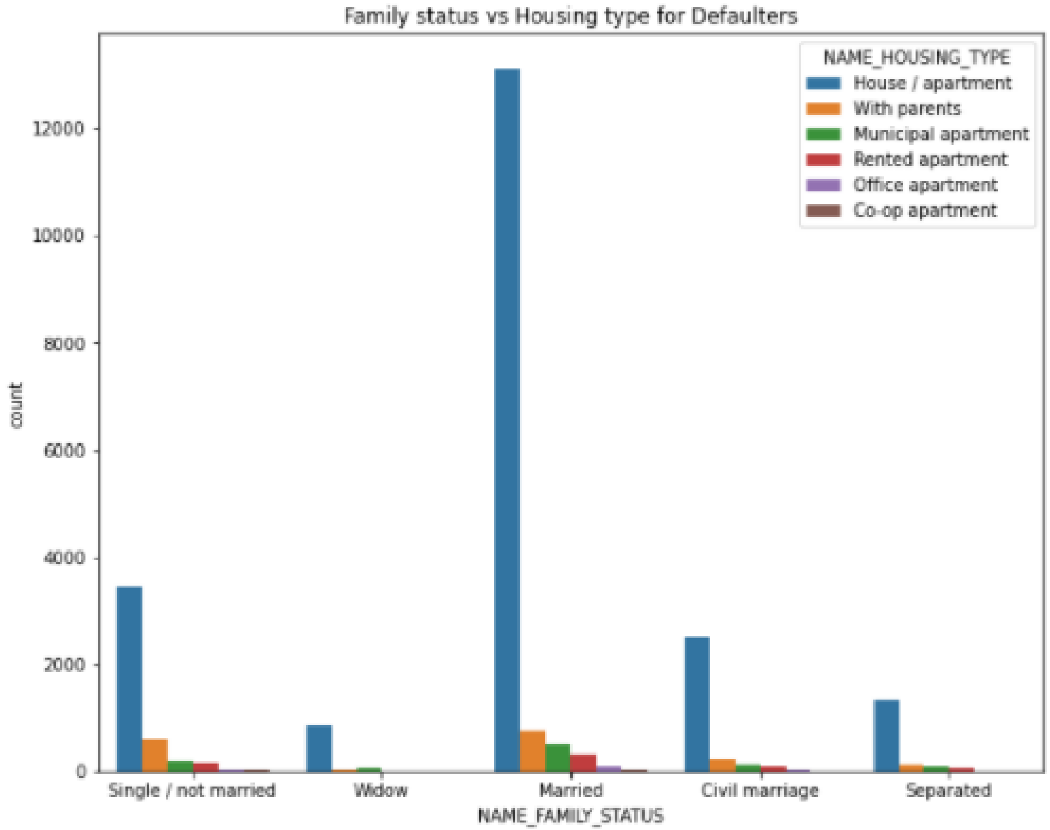
The difference between males and females in cash loans in defaulting is more than non-defaulting categories.

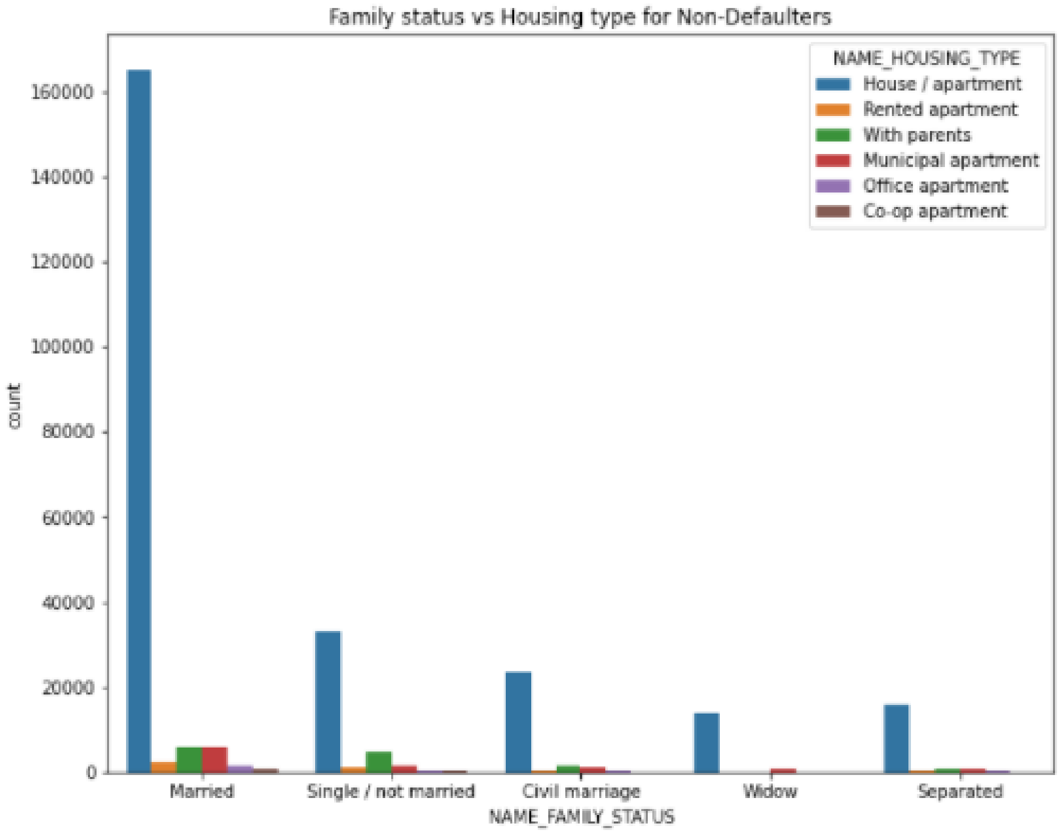
NAME_EDUCATION_TYPE vs NAME_INCOME_TYPE



Working applicants with secondary education are the majority in both the defaulter and non-defaulter categories.
For defaulters the secondary education background is for working, commercial associates, pensioners and state servant.
For non defaulters also, the secondary education background is for working, commercial associates, pensioners and state servant though in a different order.

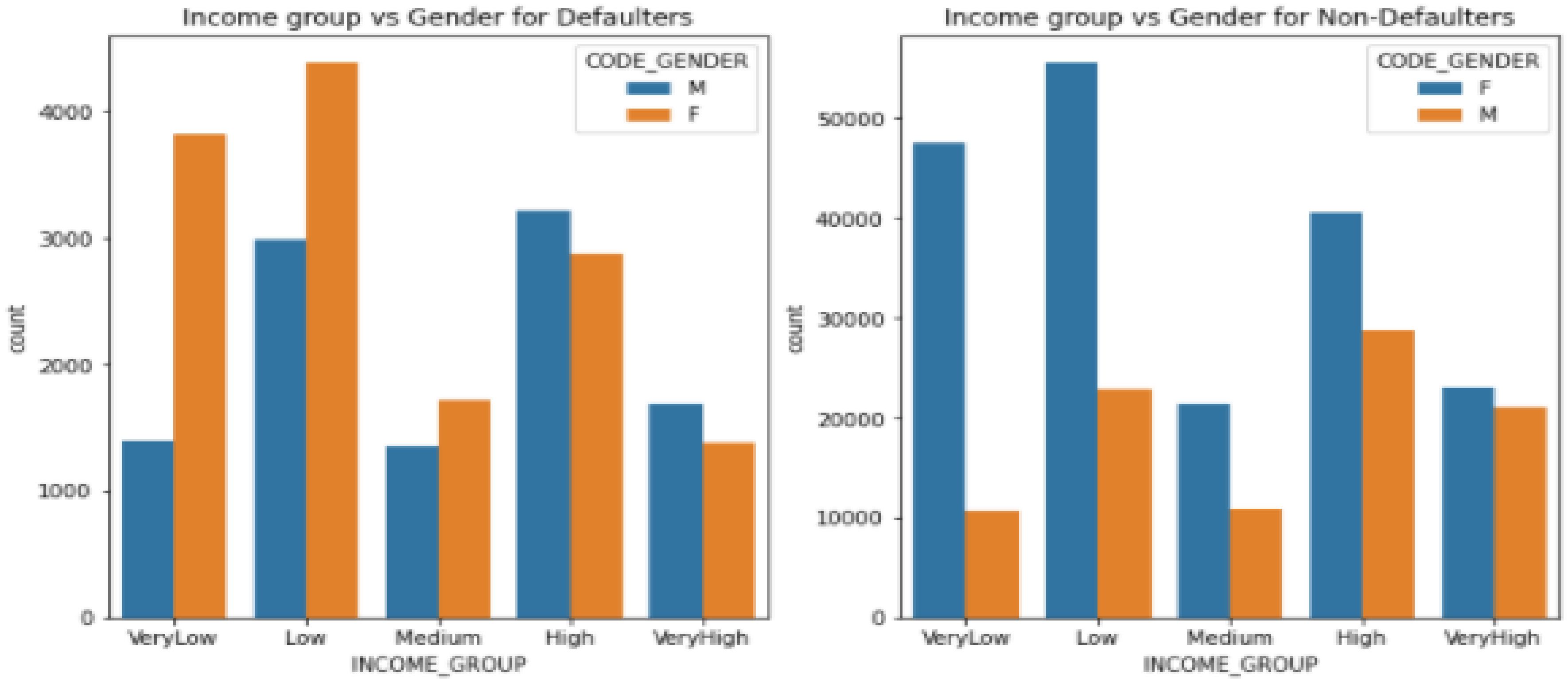
NAME_FAMILY_STATUS vs NAME_HOUSING_TYPE





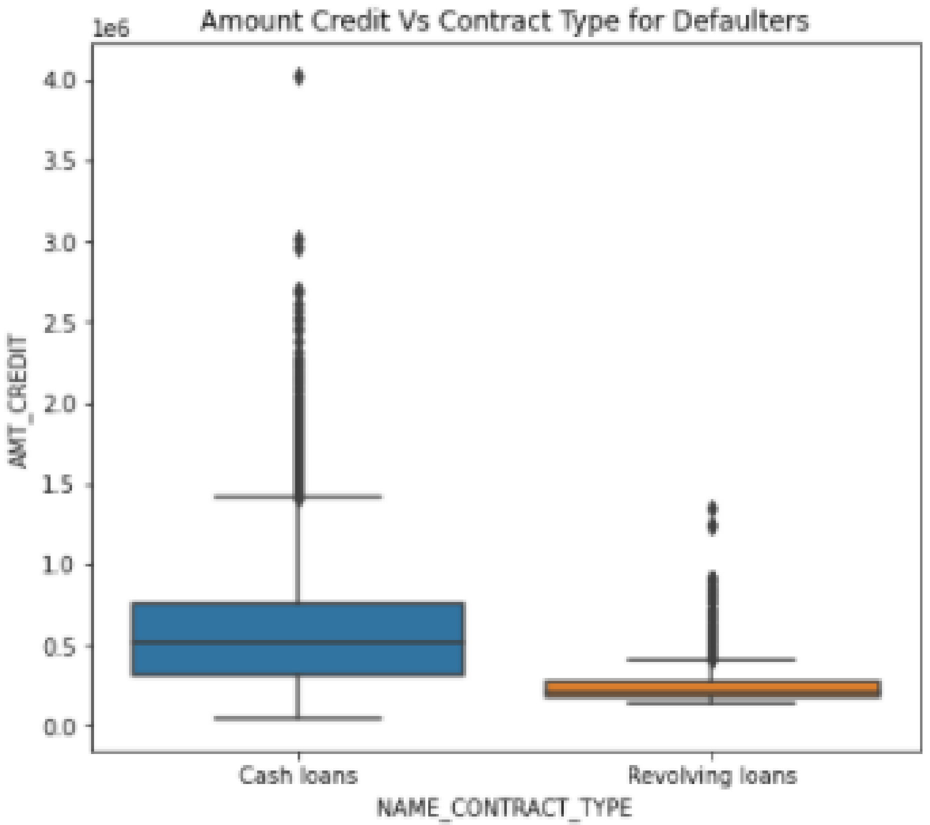
In all sub categories of defaulters applicants living in apartments are in majority
The married applicants living in apartments are more than the sum of all non-married applicants in the defaulters categories
The defaulter and non-defaulters follow same pattern in family status and housing type

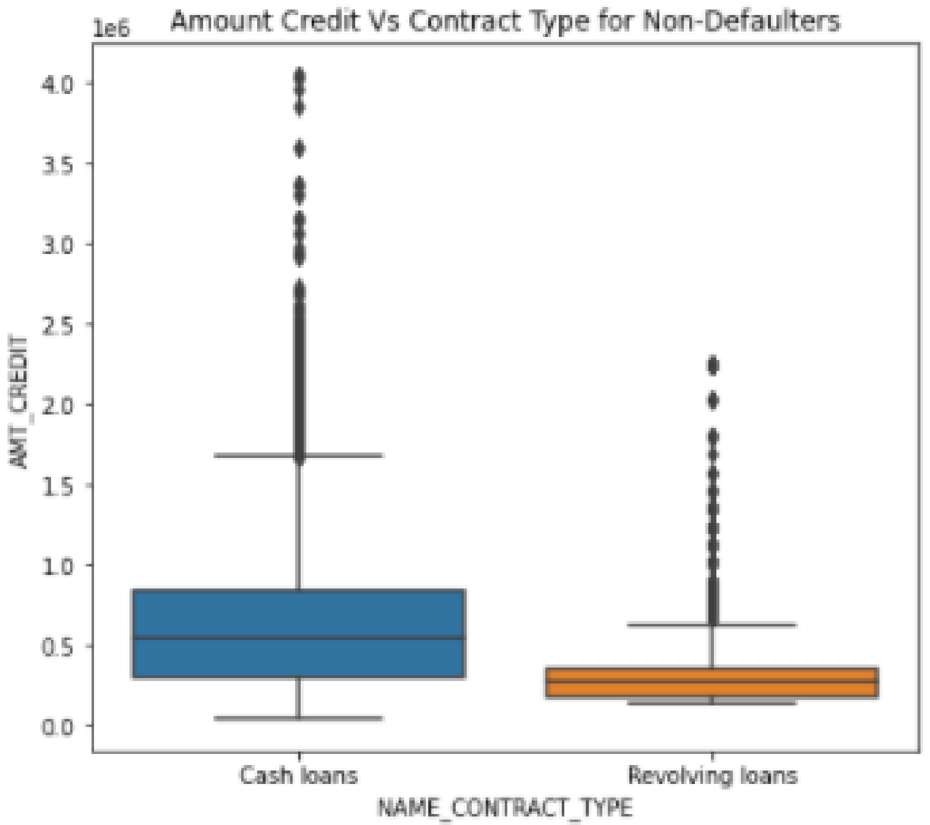
INCOME_GROUP vs CODE_GENDER



For defaulters among very low, low and medium categories females have higher proportion than the males
For defaulters among High and Very high categories females have lower proportion than the males
For all income groups females are better places compared to males in the non-defaulter category
For very low income group the difference between males and females is highest in the non defaulting category

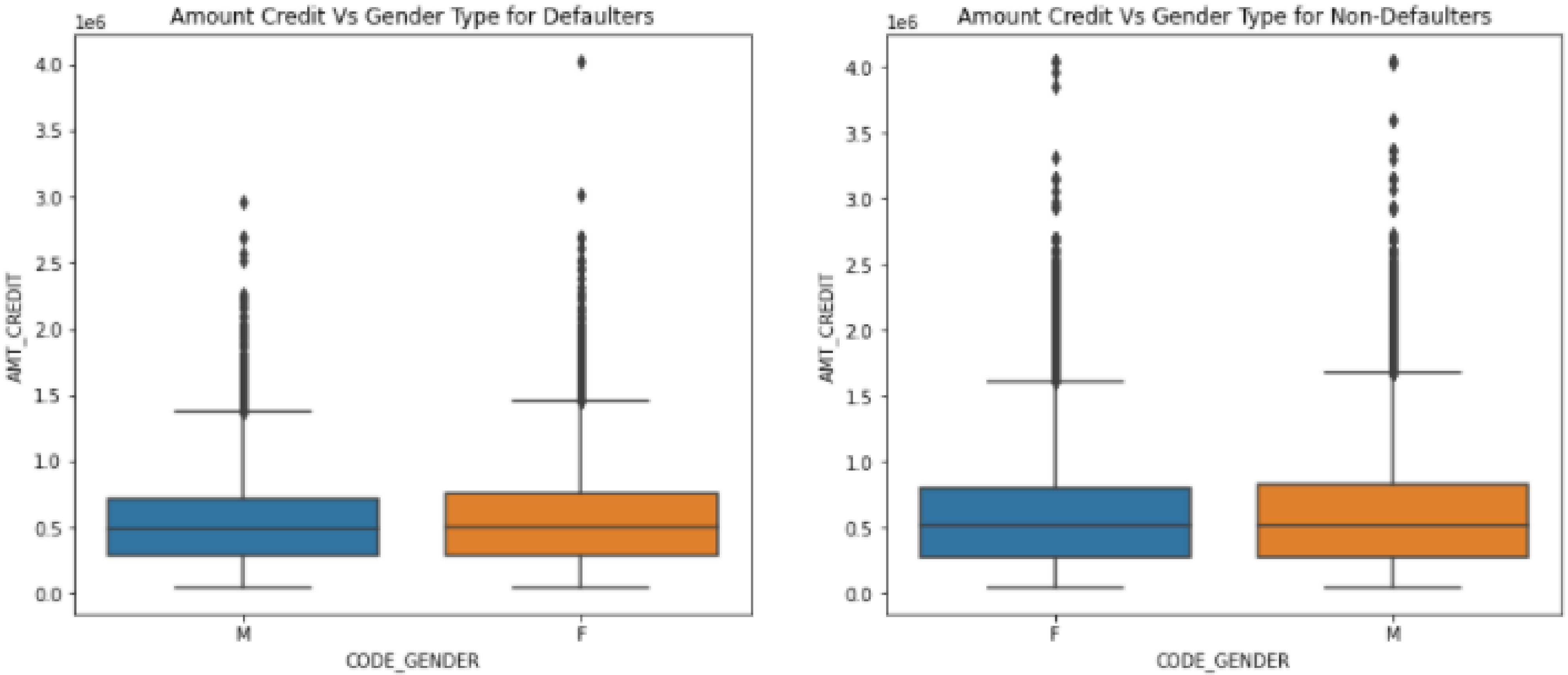
NAME_CONTRACT_TYPE vs AMT_CREDIT





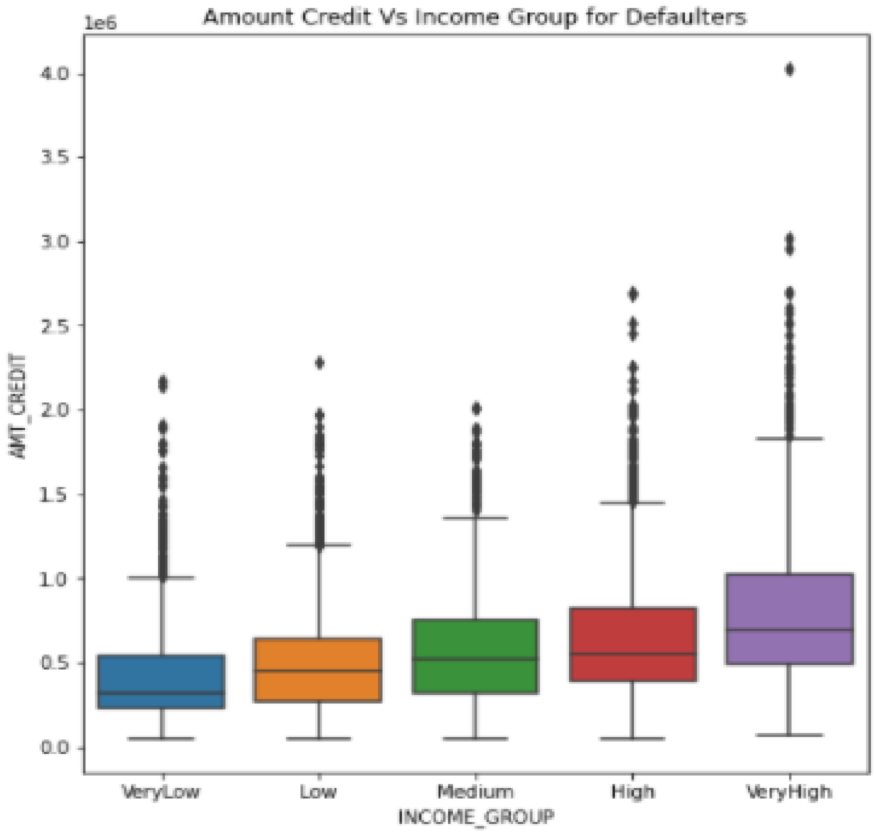
The amount credited is higher for cash loans in both the defaulting and non defaulting category
The bulk of the cash loans for defaulter is less than 1
The cash loans proportion is higher for non defaulters than defaulters as far as amount credit is concerned

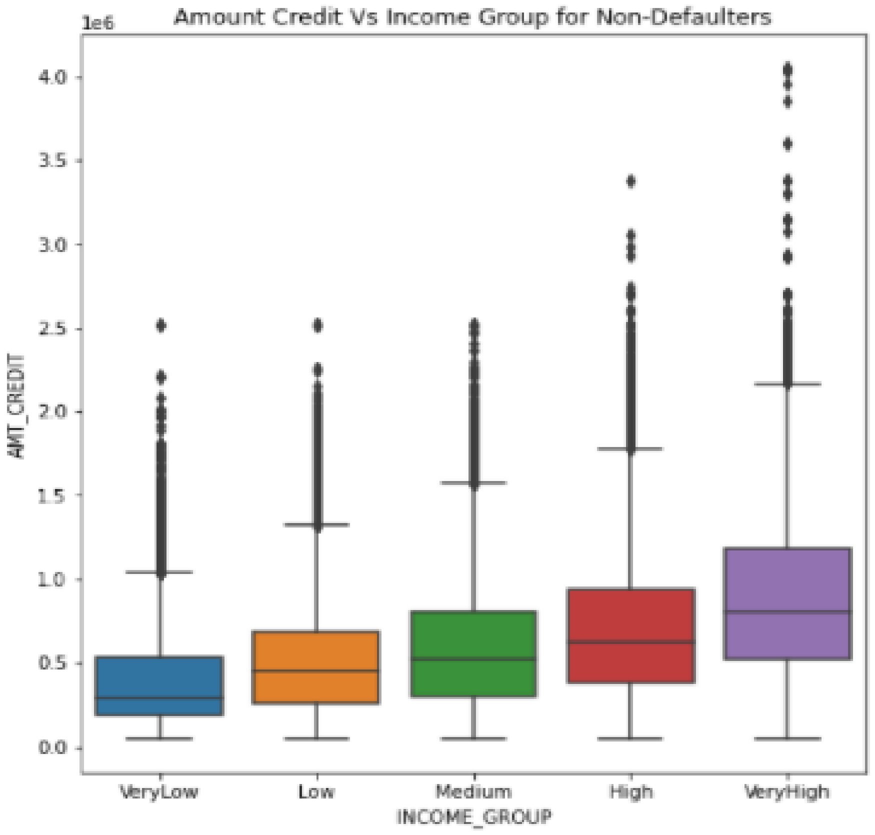
CODE_GENDER vs AMT_CREDIT



The amount credited for male and female in both defaulting and non-defaulting categories are almost same for majority of applications
The amount credited for both the males and females for non defaulting category is higher than defaulting category

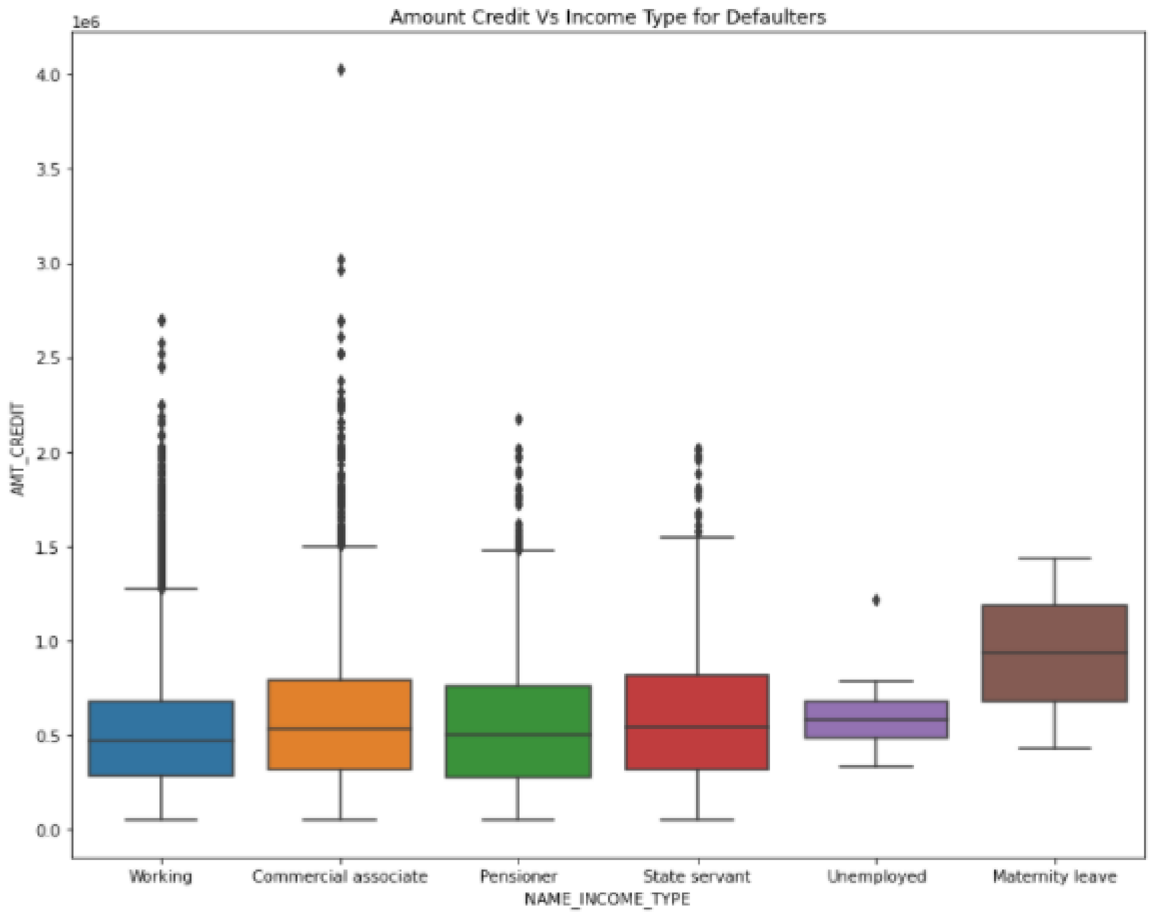
INCOME_GROUP vs AMT_CREDIT

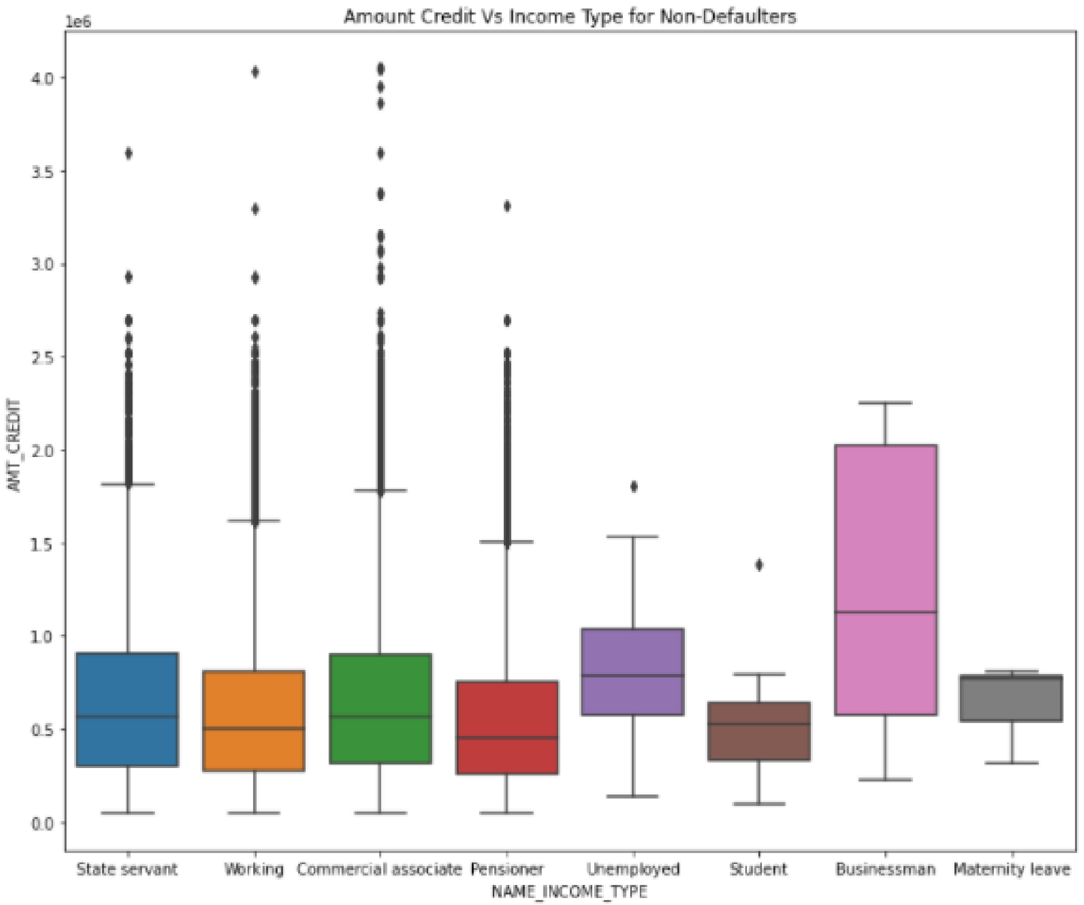




The amount credit for both the defaulters and non defaulters follow the same pattern with Very low income group at bottom and very high at the top.
The non defaulter category has almost continuous coverage for very high income group till the maximum amount.

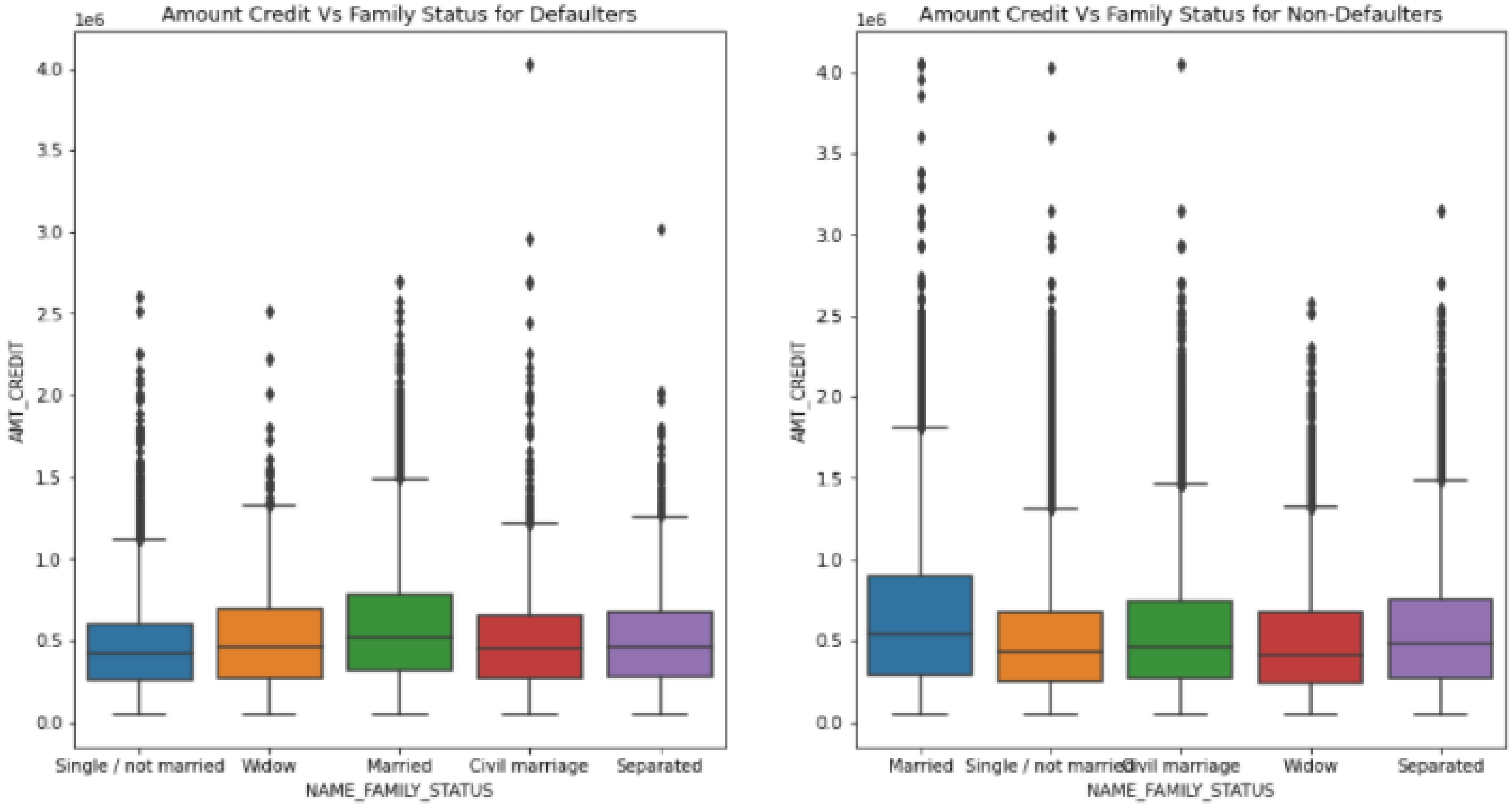
NAME_INCOME_TYPE vs AMT_CREDIT



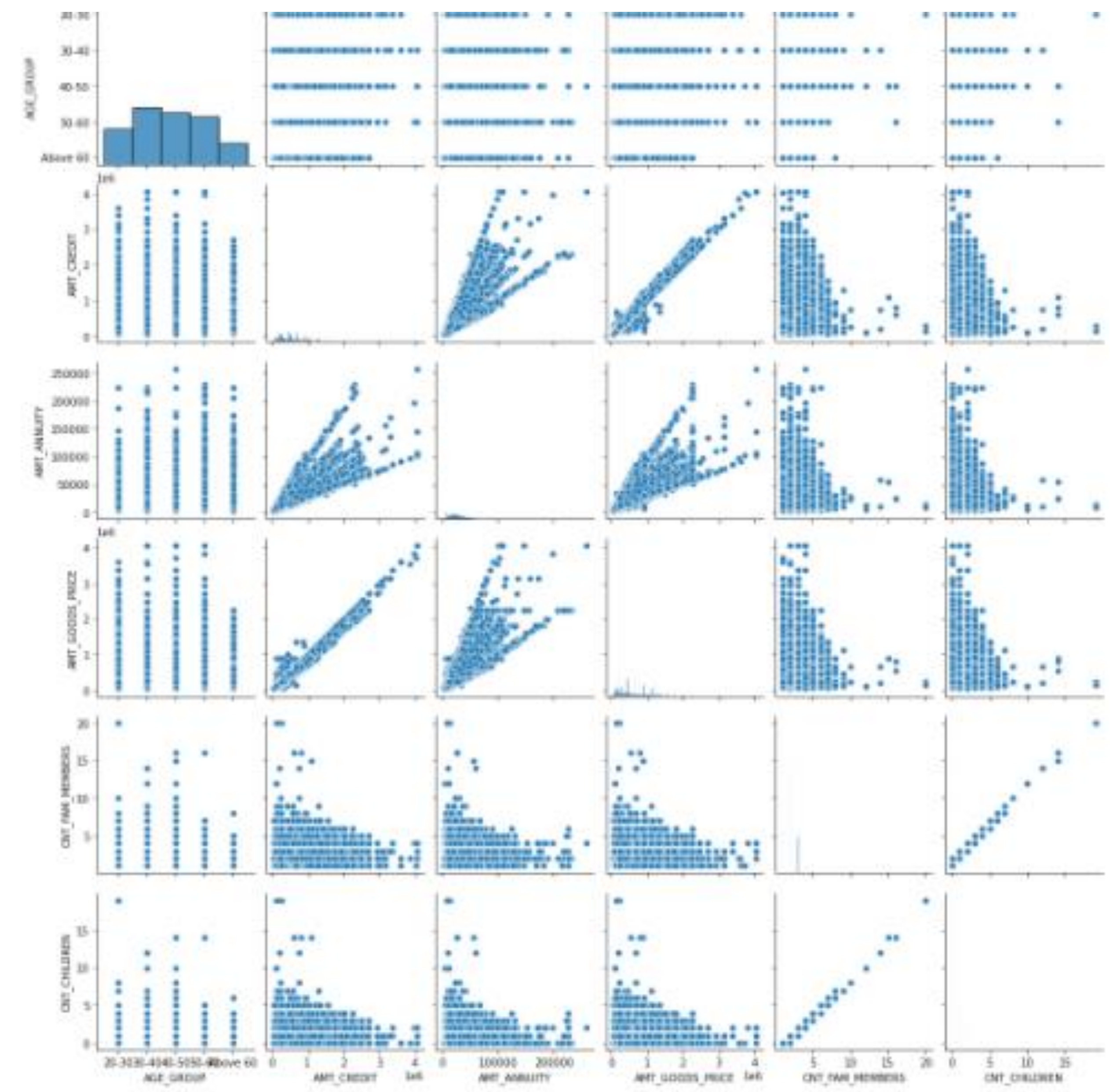


The income type for non-defaulters have more sub categories including the student and businessman which are not significant for defaulters
The maternity leaves category has higher default though the amount credit is lower than other segments
The data reveals that businessman are less prone to default compared to other categories though the amount credit for them is not the highest

NAME_FAMILY_STATUS vs AMT_CREDIT



The defaulting proportion of married applicants is higher along with the average amount credit
The amount credit is highest for civil marriage in both the defaulting and non defaulting category
The sub categories covered in both the defaulting and non defaulting are same



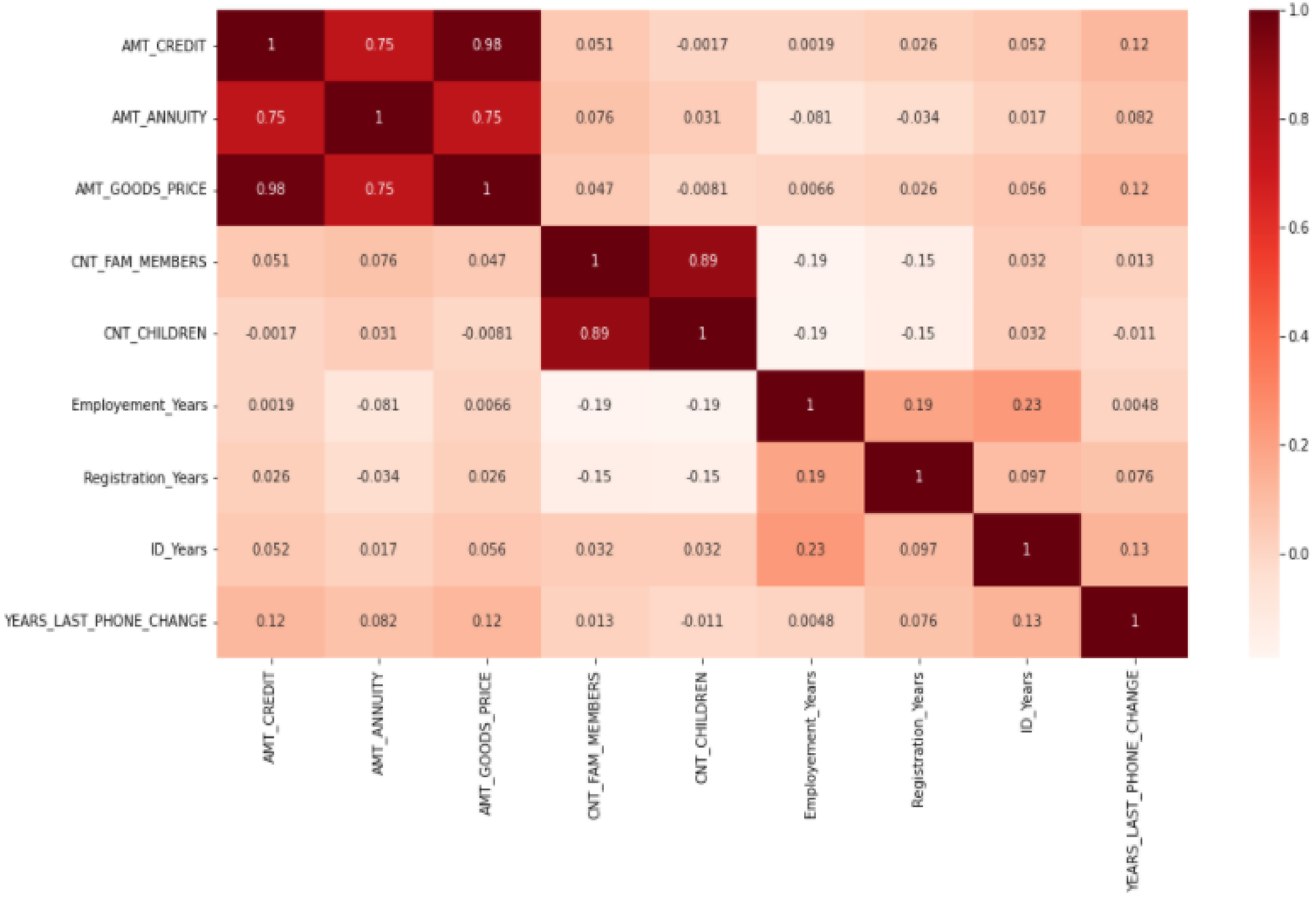
Bivariate Analysis of Numerical Columns:

'AGE_GROUP’,
'AMT_CREDIT', 'AMT_ANNUITY',
'AMT_GOODS_PRICE’,
'CNT_FAM_MEMBERS',
'CNT_CHILDREN'

Multivariate analysis: correlation matrix for defaulters

	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	CNT_FAM_MEMBERS	CNT_CHILDREN	Employement_Years	Registration_Years	ID_Years	YEARS_LAST_PHONE_CHANGE
AMT_CREDIT	1.000000	0.752195	0.983103	0.051224	-0.001675	0.001936	0.025949	0.052010	0.117117
AMT_ANNUITY	0.752195	1.000000	0.752699	0.075711	0.031257	-0.081213	-0.034047	0.016624	0.081911
AMT_GOODS_PRICE	0.983103	0.752699	1.000000	0.047388	-0.008112	0.006644	0.025834	0.055899	0.123883
CNT_FAM_MEMBERS	0.051224	0.075711	0.047388	1.000000	0.885484	-0.186516	-0.145576	0.031652	0.012709
CNT_CHILDREN	-0.001675	0.031257	-0.008112	0.885484	1.000000	-0.192866	-0.149029	0.031792	-0.011222
Employement_Years	0.001936	-0.081213	0.006644	-0.186516	-0.192866	1.000000	0.192537	0.229150	0.004767
Registration_Years	0.025949	-0.034047	0.025834	-0.145576	-0.149029	0.192537	1.000000	0.096846	0.075953
ID_Years	0.052010	0.016624	0.055899	0.031652	0.031792	0.229150	0.096846	1.000000	0.131512
YEARS_LAST_PHONE_CHANGE	0.117117	0.081911	0.123883	0.012709	-0.011222	0.004767	0.075953	0.131512	1.000000

Multivariate analysis: Plot the correlation matrix for defaulters (Heatmap)

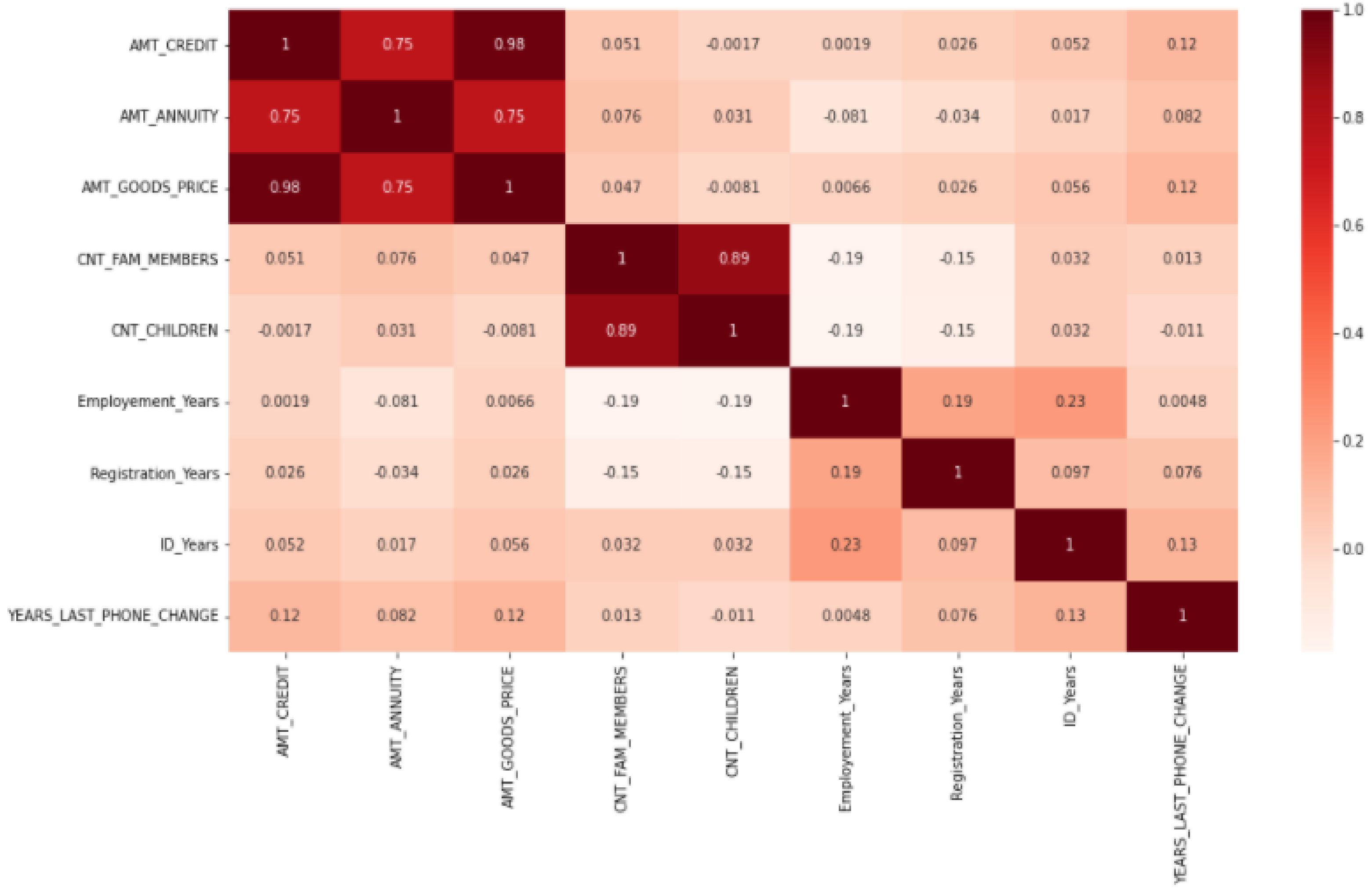


The amount credit shows very high correlation with the amount good price for defaulters which is expected.
The amount credit and no of employments years have no significant correlation for defaulters
The amount of annuity is correlated with amount credit for defaulters
The count of children are highly correlated with count of family members for defaulters which is inline with the expectation
The count of children have negative correlation with the employment years for defaulters
The count of children has low correlation with most of the factors for defaulters
The registration years has low correlation with most of the factors for defaulters
The Id-years has low correlation with the employment years for defaulters

Multivariate analysis: correlation matrix for non-defaulters

AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	CNT_FAM_MEMBERS	CNT_CHILDREN	Employement_Years	Registration_Years	ID_Years	YEARS_LAST_PHONE_CHANGE	
AMT_CREDIT	1.000000	0.771309	0.987250	0.064536	0.003081	-0.070105	-0.013375	0.001778	0.072697
AMT_ANNUITY	0.771309	1.000000	0.776686	0.075787	0.020905	-0.104983	-0.039304	-0.013916	0.063098
AMT_GOODS_PRICE	0.987250	0.776686	1.000000	0.062814	-0.000525	-0.068609	-0.015830	0.003983	0.073979
CNT_FAM_MEMBERS	0.064536	0.075787	0.062814	1.000000	0.878571	-0.238301	-0.175702	0.020442	0.027819
CNT_CHILDREN	0.003081	0.020905	-0.000525	0.878571	1.000000	-0.245173	-0.185818	0.028867	0.007724
Employement_Years	-0.070105	-0.104983	-0.068609	-0.238301	-0.245173	1.000000	0.214525	0.276254	-0.020674
Registration_Years	-0.013375	-0.039304	-0.015830	-0.175702	-0.185818	0.214525	1.000000	0.099681	0.055187
ID_Years	0.001778	-0.013916	0.003983	0.020442	0.028867	0.276254	0.099681	1.000000	0.087370
YEARS_LAST_PHONE_CHANGE	0.072697	0.063098	0.073979	0.027819	0.007724	-0.020674	0.055187	0.087370	1.00000

Multivariate analysis: Plot the correlation matrix for non-defaulters (Heatmap)

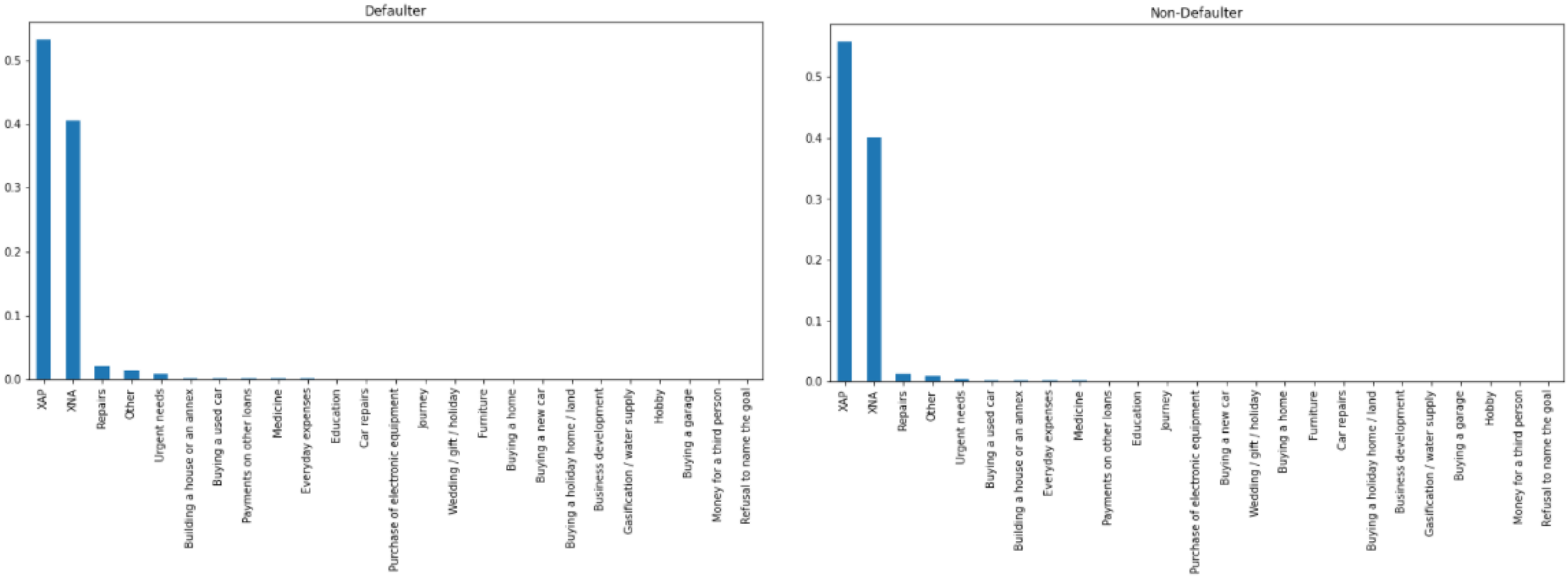


The amount credit shows very high correlation with the amount good price for non defaulters.
The amount credit and no of employments years have no significant correlation for non defaulters
The amount of annuity is correlated with amount credit for non defaulters
The count of children are highly correlated with count of family members for non defaulters
The count of children have negative correlation with the employment years for non defaulters
The count of children has low correlation with most of the factors for non defaulters
The registration years has low correlation with most of the factors for non defaulters
The Id-years has low correlation with the employment years for non defaulters

Analysis of the merged data set: Previous application and Application data

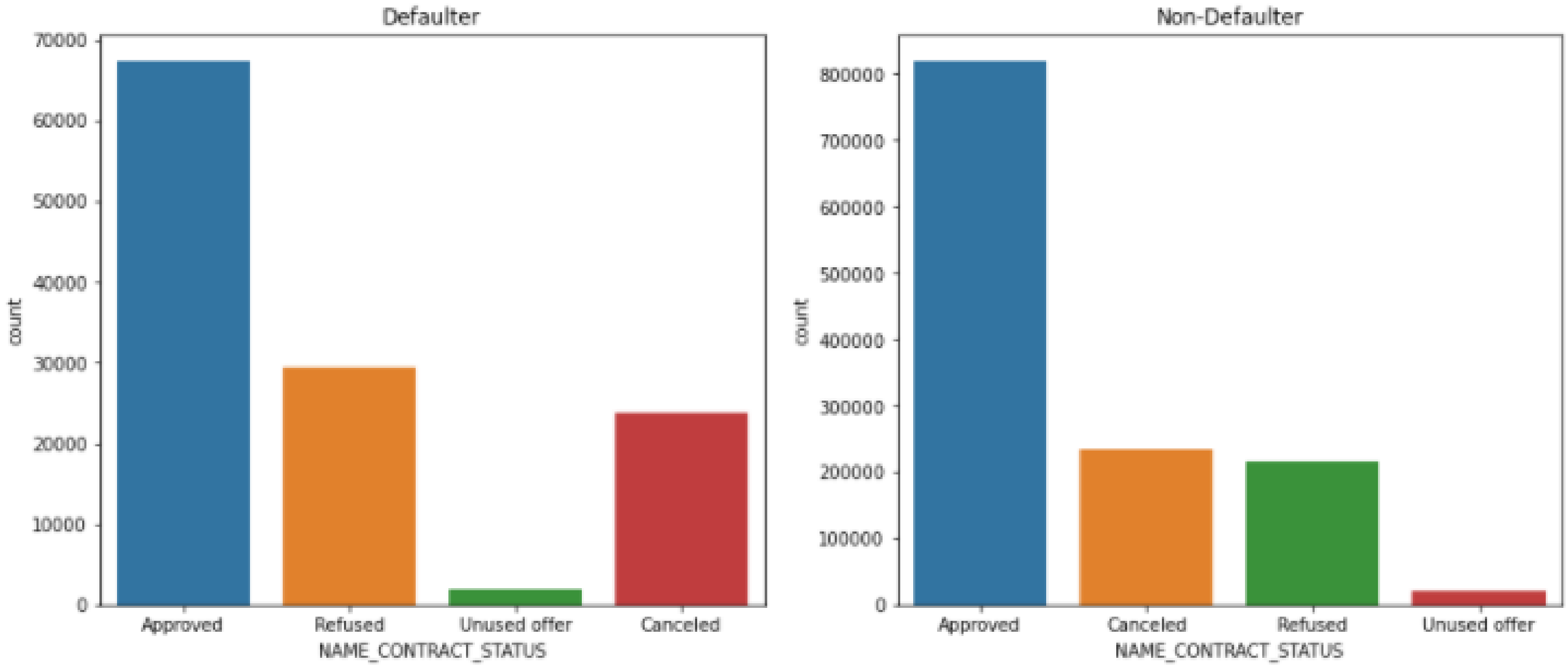
```
# df_merge_d : Defaulters
# df_merge_nd : Non-Defaulters
df_merge_d = app_merge[app_merge['TARGET']==1]
df_merge_nd = app_merge[app_merge['TARGET']==0]
```

NAME_CASH_LOAN_PURPOSE



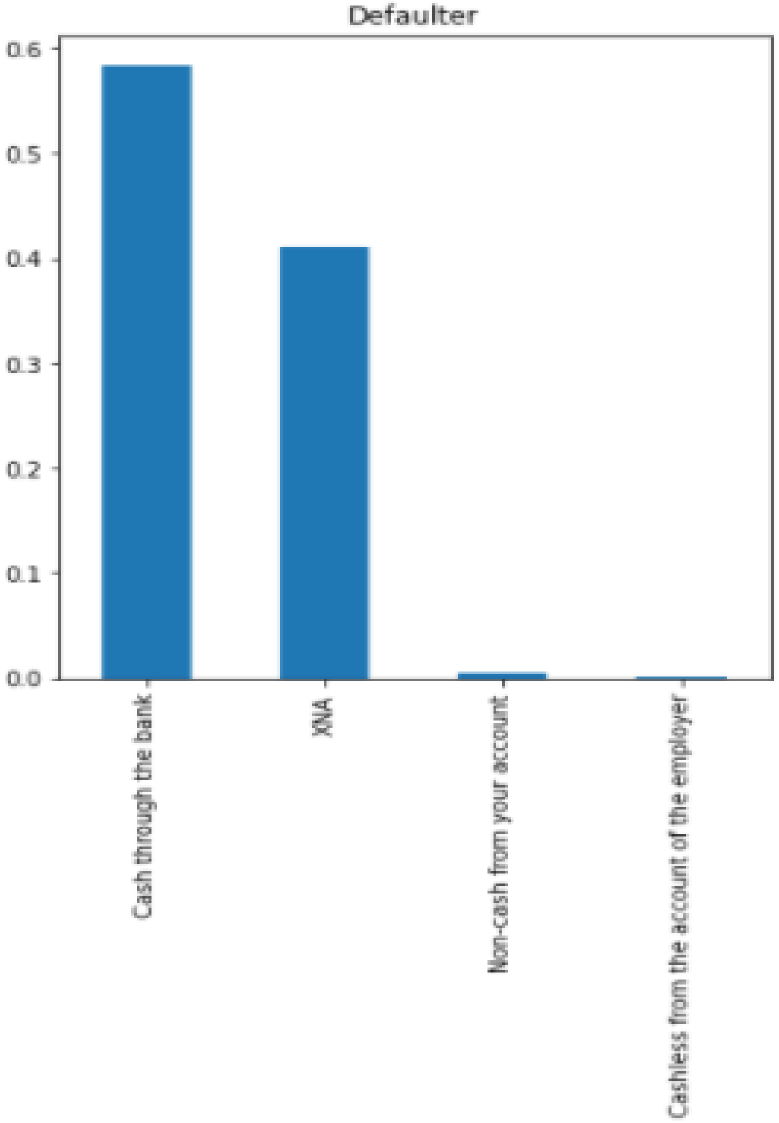
The XAP and XNA are majority in both the defaulter as well as non defaulter segment
The repairs is the 3rd largest segment in both the defaulter and non defaulter segment
There are 25 categories of cash loan purpose for both the defaulter and non defaulter group

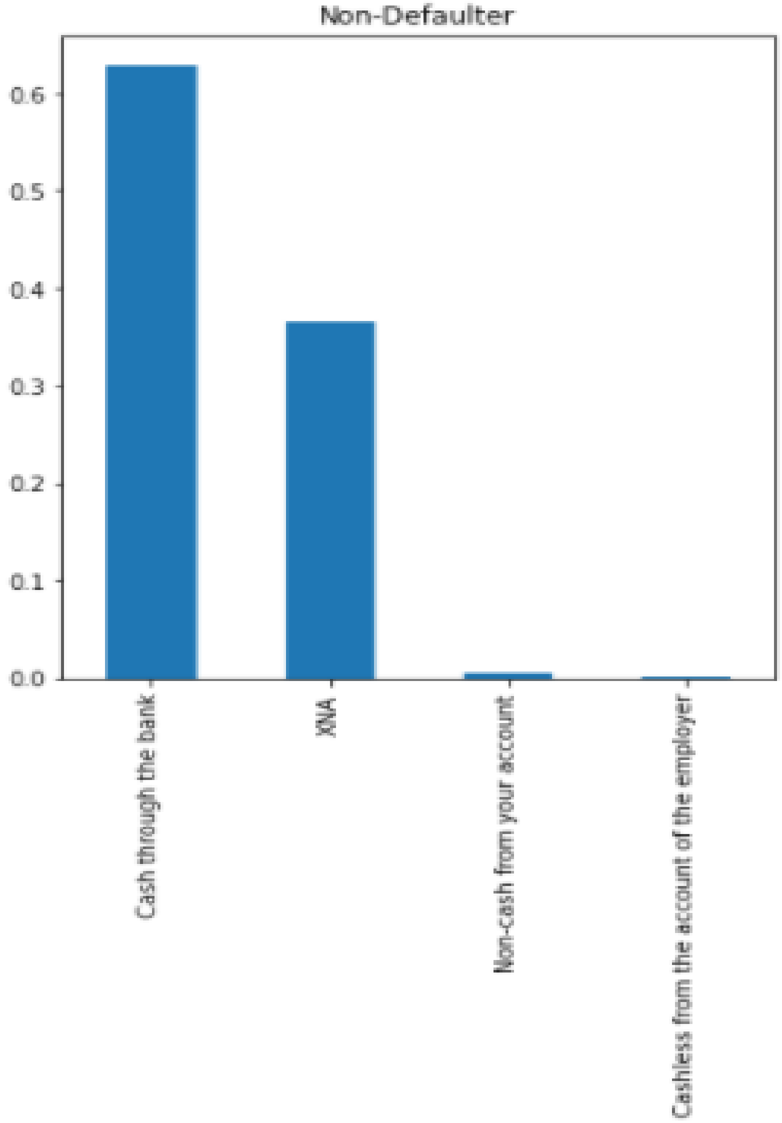
NAME_CONTRACT_STATUS



The approved is the largest category in both the defaulting and non defaulting segment
The refused is 2nd largest category for defaulters while the cancelled is 2nd largest category for non defaulter segment
Unused offer is the least contribution category in both the defaulting and non defaulting segment

NAME_PAYMENT_TYPE



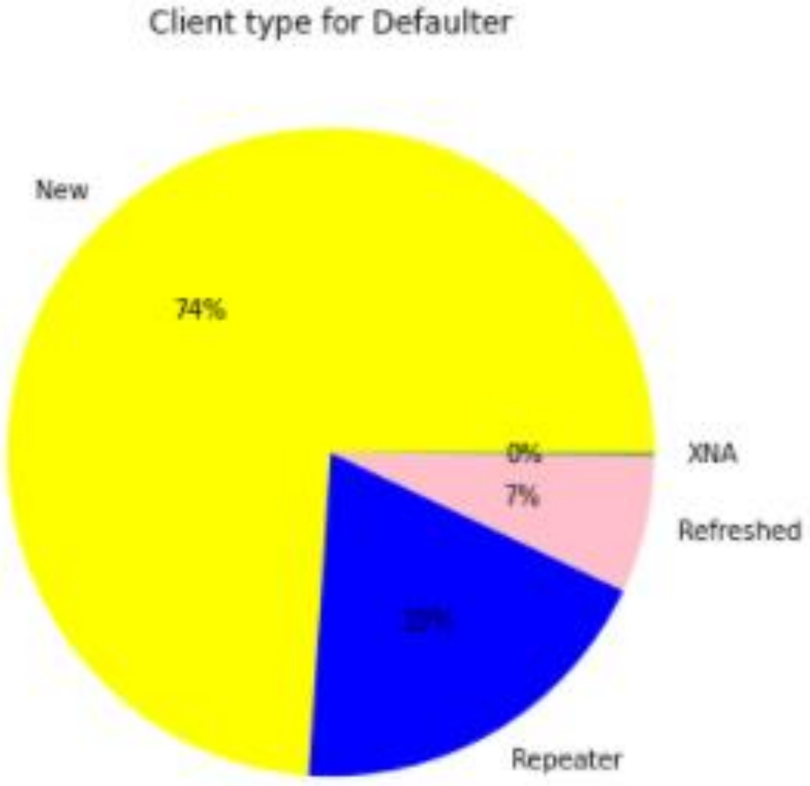


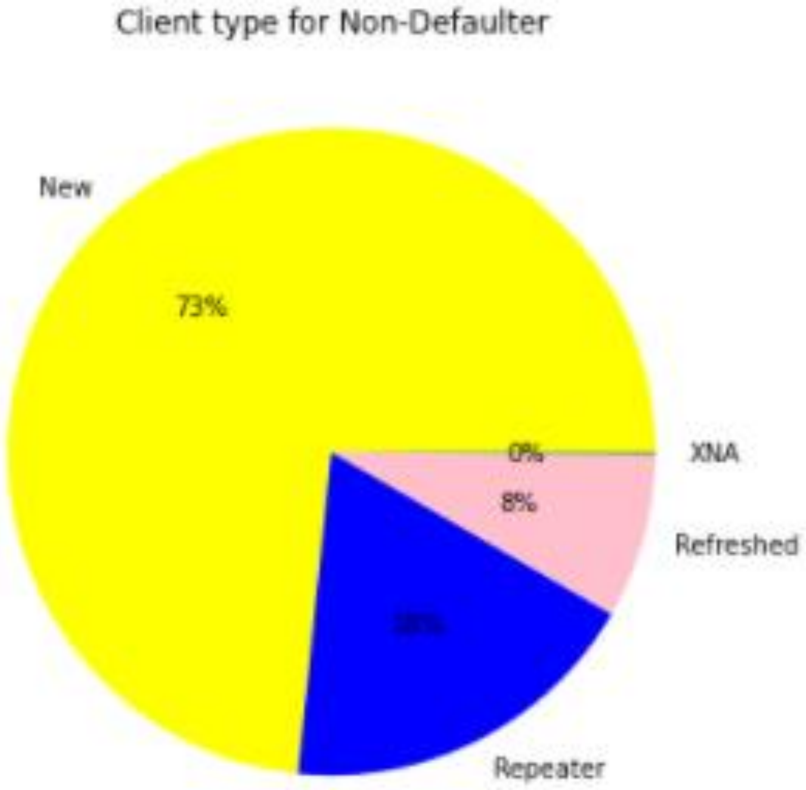
The cash through the bank is the highest contributor for both the defaulting and non defaulting category

The XNA is the 2nd largest category for both the segments

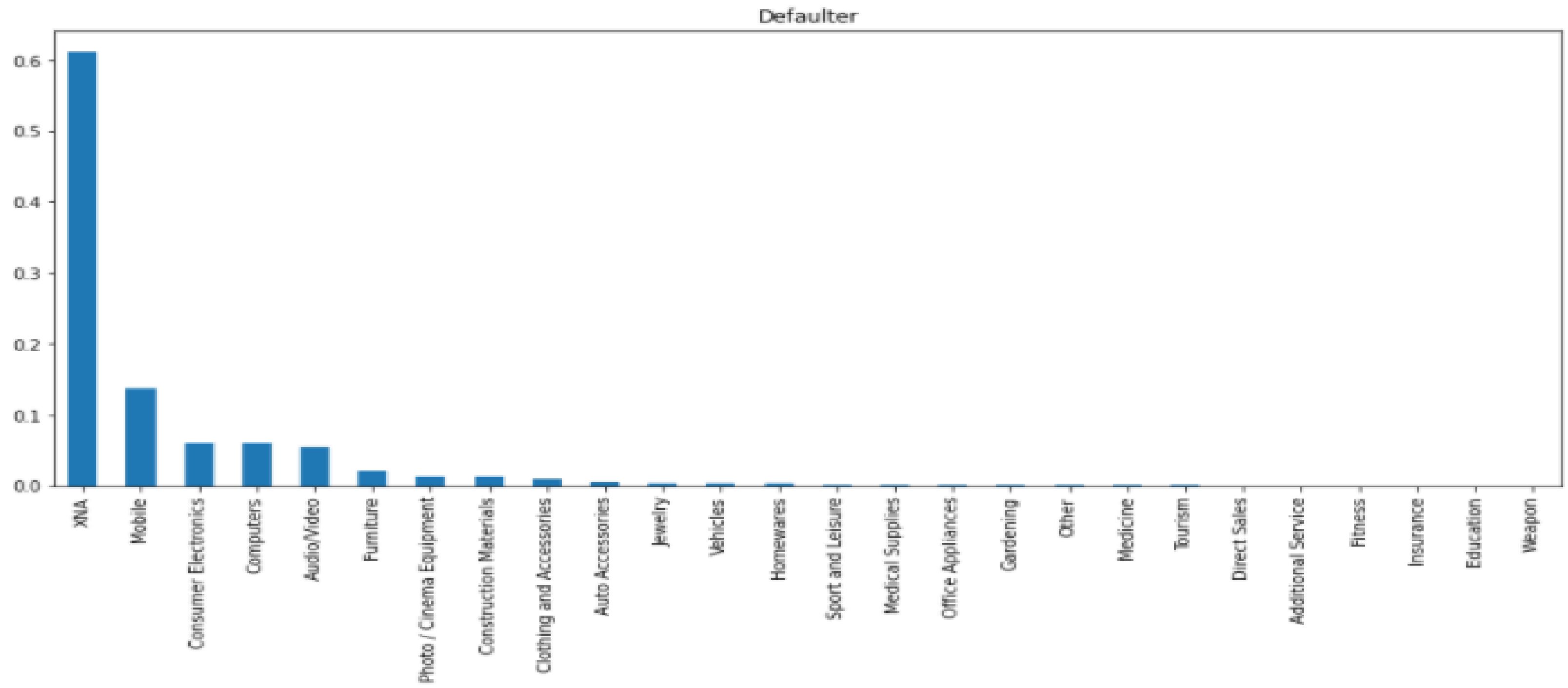
The rest other categories contribute insignificantly to both the segments

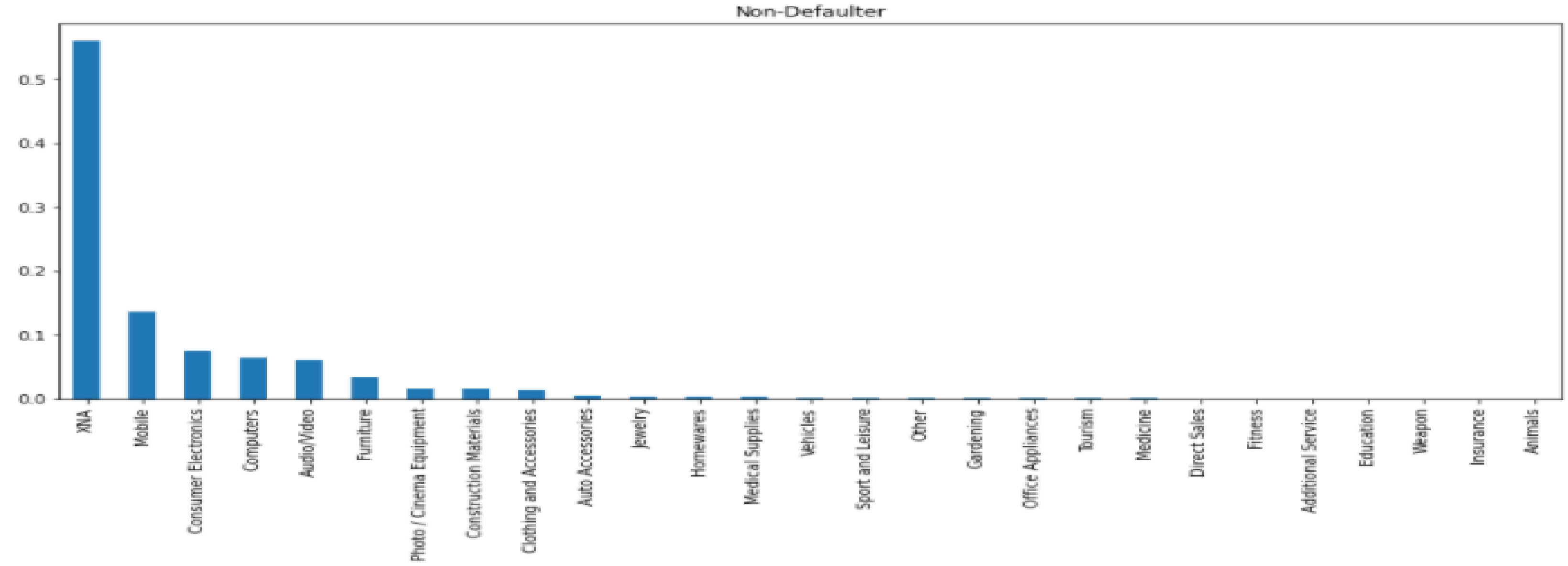
NAME_CLIENT_TYPE





The client type for both the defaulting and non defaulting segment follow same pattern
1/5th of the applicants in both defaulting and non defaulting segment are repeaters
3/4th of the applicants in both defaulting and non defaulting segment are new



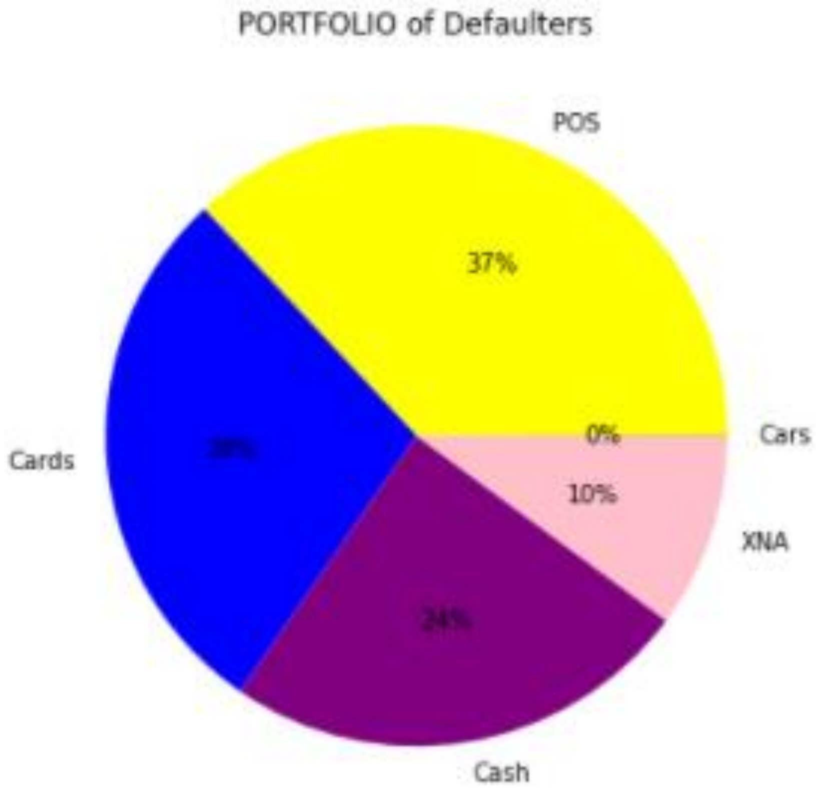


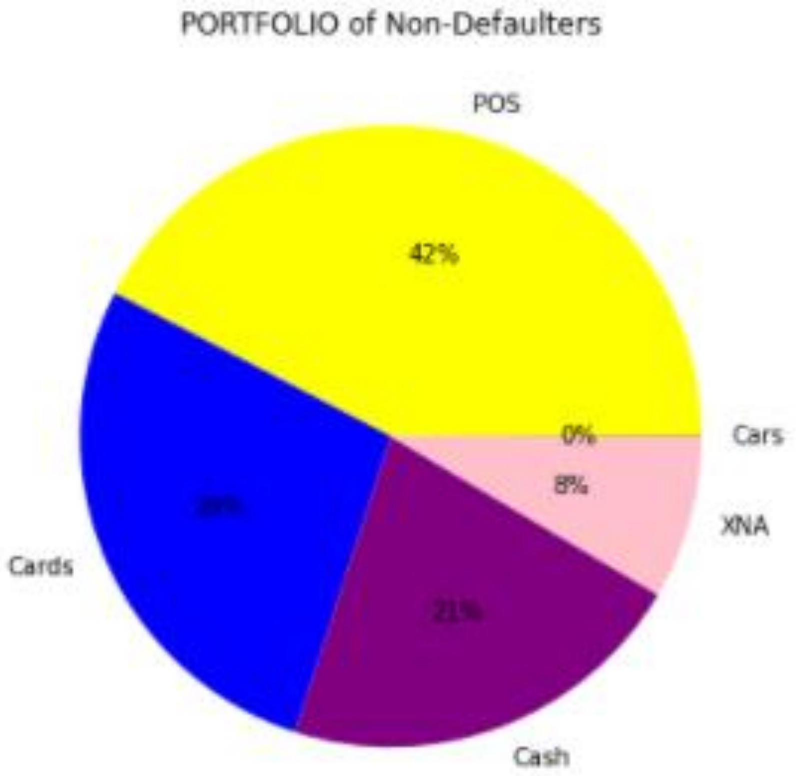
NAME_GOODS_CATEGORY

Apart from XNA, the mobile and electronics items cover the major portion in both the defaulting and non defaulting segment

There are more than 25 categories in both the defaulting and non defaulting segment in the Goods category

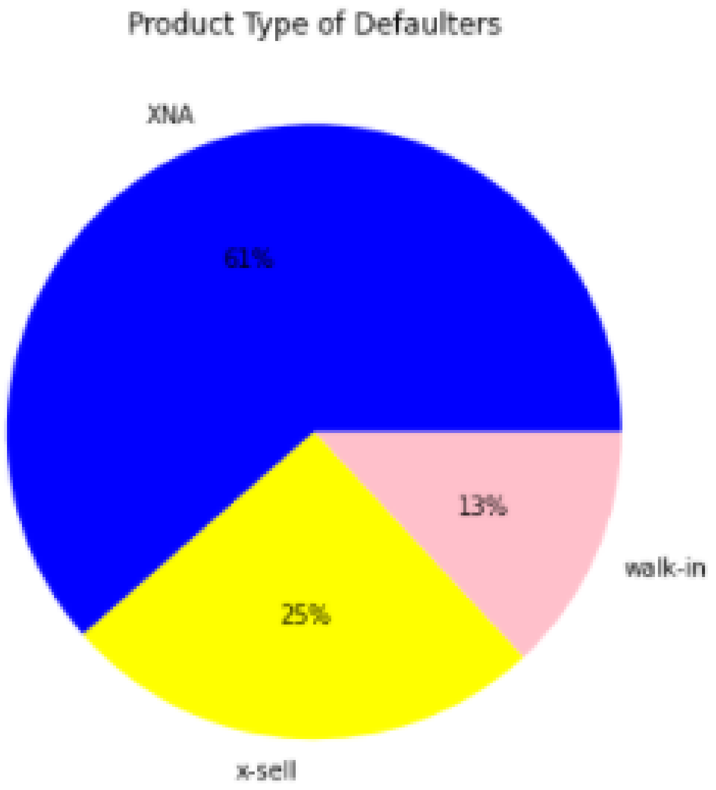
NAME_PORTFOLIO

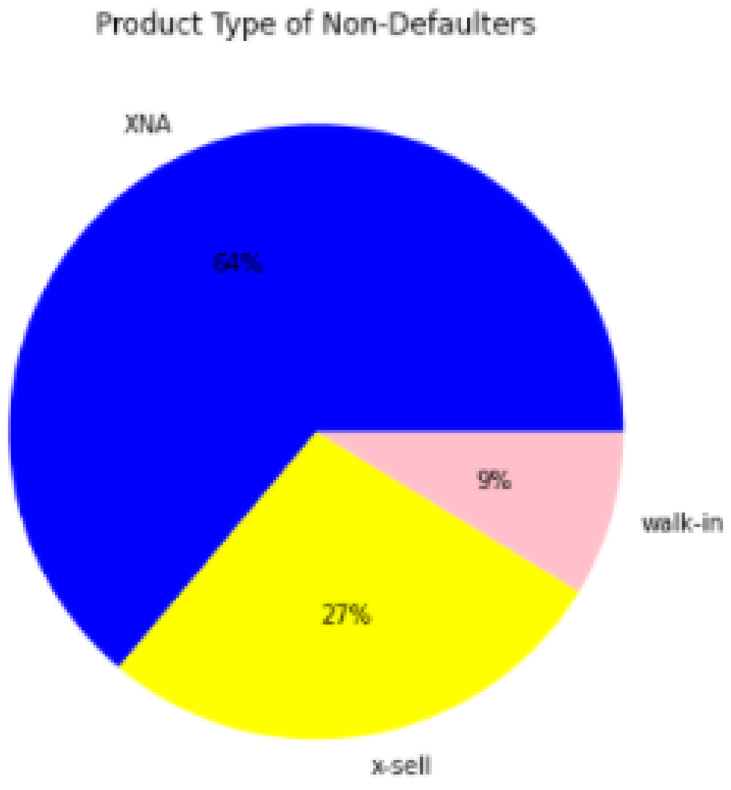




The POS, Cards and Cash remain the top 3 categories with cumulative contribution of around 90% for both the segments-defaulting and non-defaulting.
The POS is better compared to cards and cash in terms of repayment as it has higher % of non-defaulters.

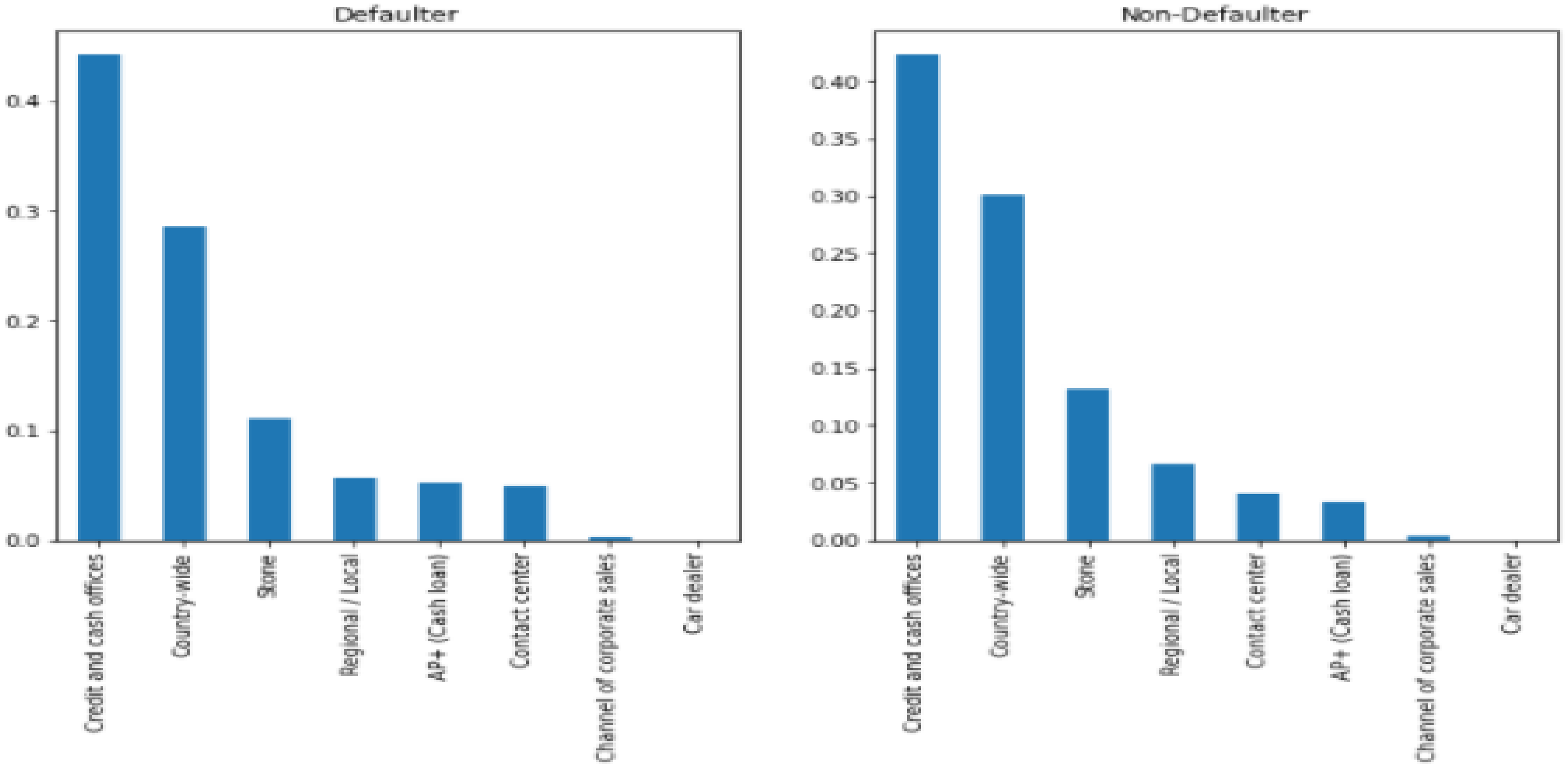
NAME_PRODUCT_TYPE





The product types of defaulters and non defaulters follow the same pattern
1/4th of the defaulters and non defaulters are x-sell
The proportion of defaulters in walk-in is higher compared to non defaulters

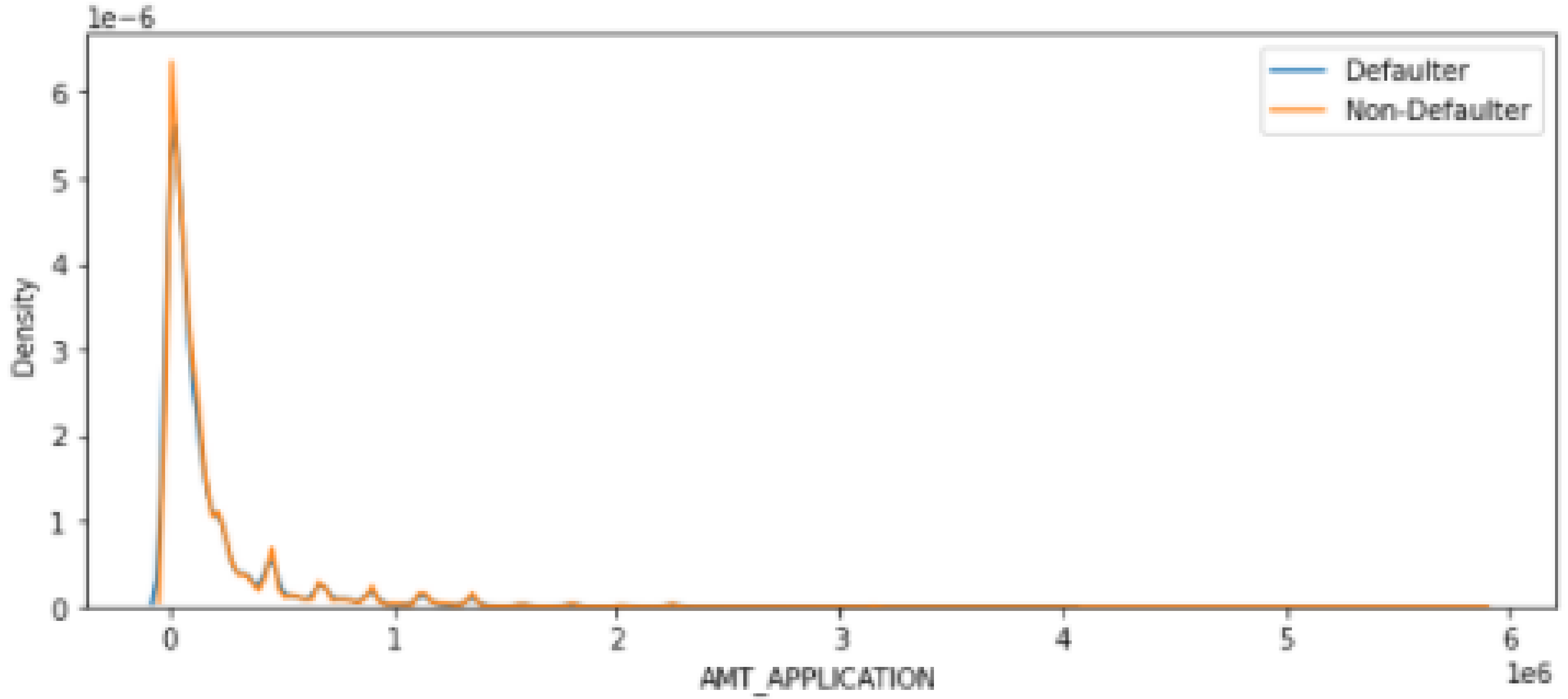
CHANNEL_TYPE



The Credit and cash offices, country wise and store remains the top 3 categories for both the defaulting and non-defaulting segments.

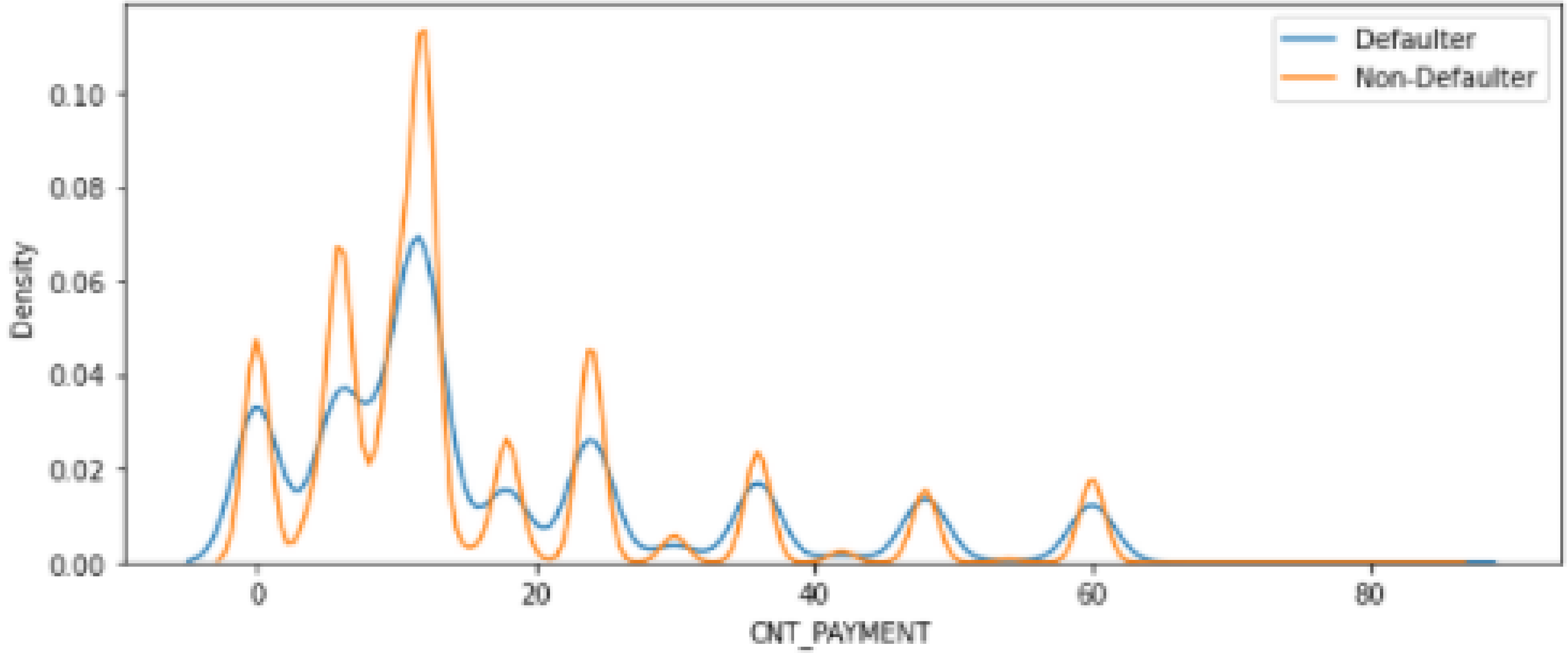
Both the defaulters and non defaulters have same 8 categories with the similar contribution from each of the categories

AMT_APPLICATION



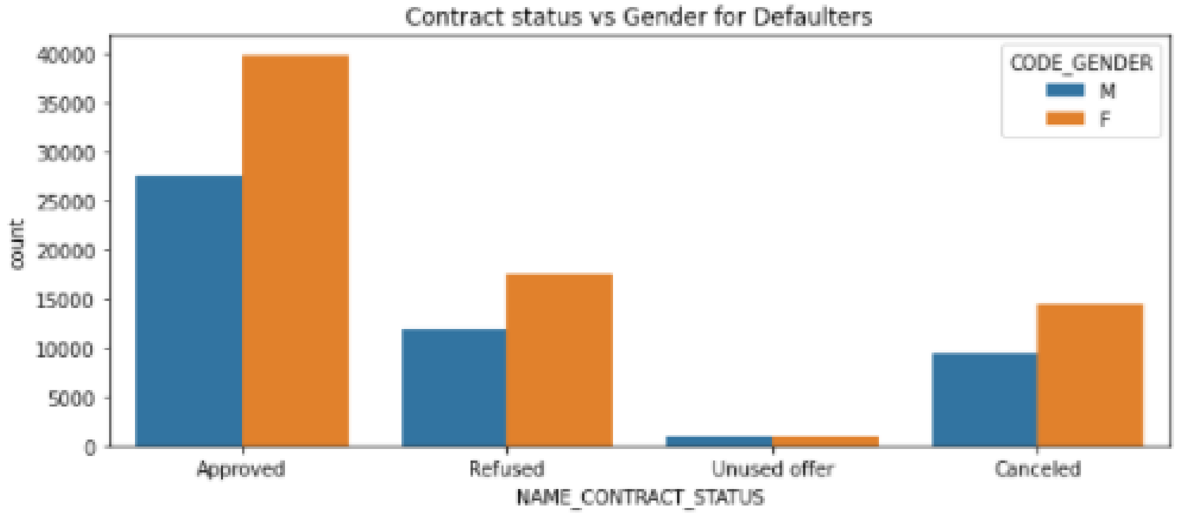
The density is highest for lower amount of loan application for no defaulters
The density is almost zero for amount application greater than 1

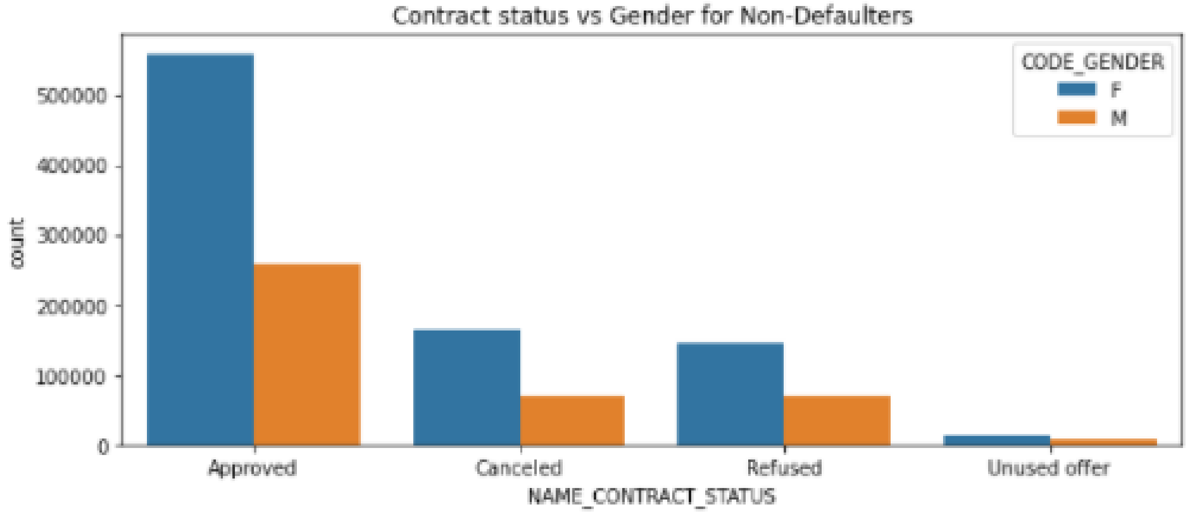
CNT_PAYMENT



The density is higher for non defaulter than defaulter type of clients in terms of control payments

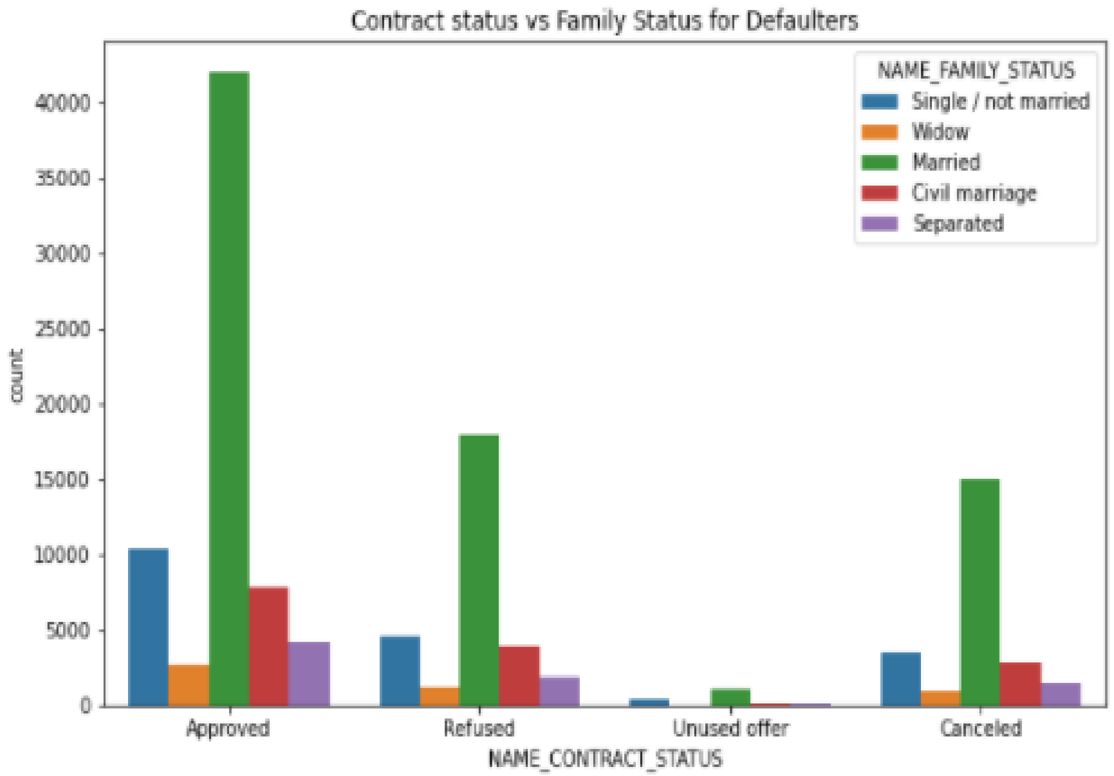
CODE_GENDER vs NAME_CONTRACT_STATUS

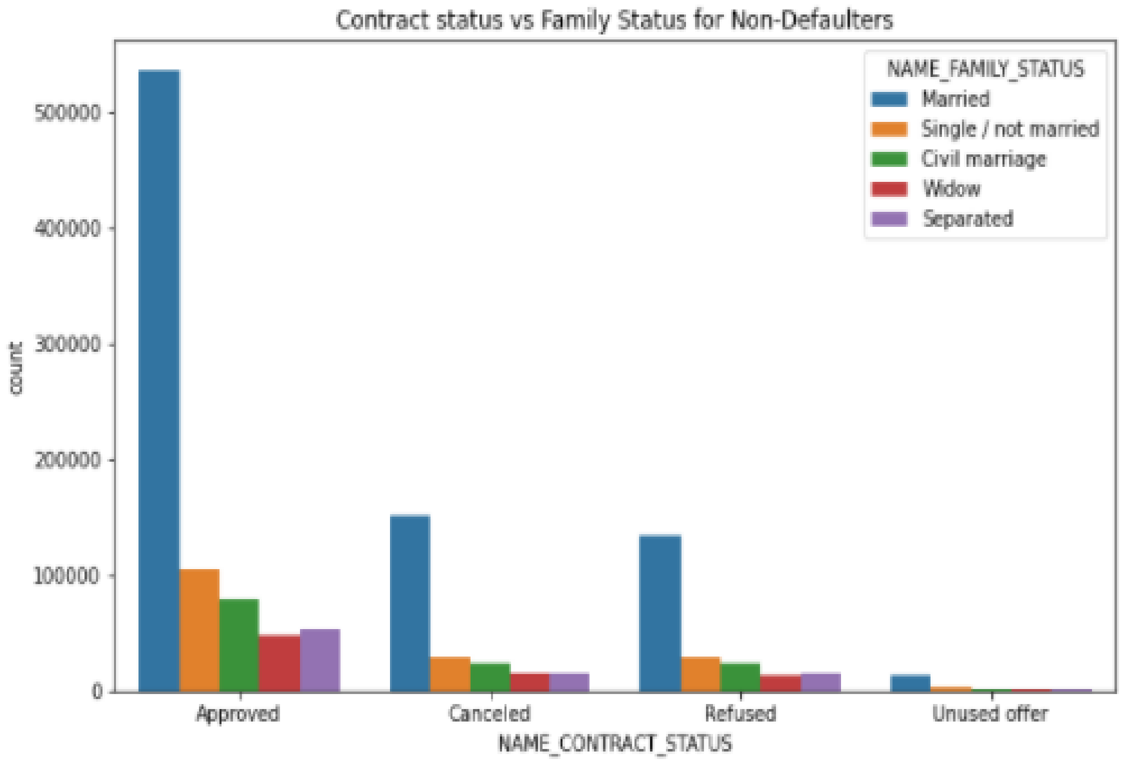




Number of females clients for both defaulter and non-defaulter categories are high in all four types of contract status.

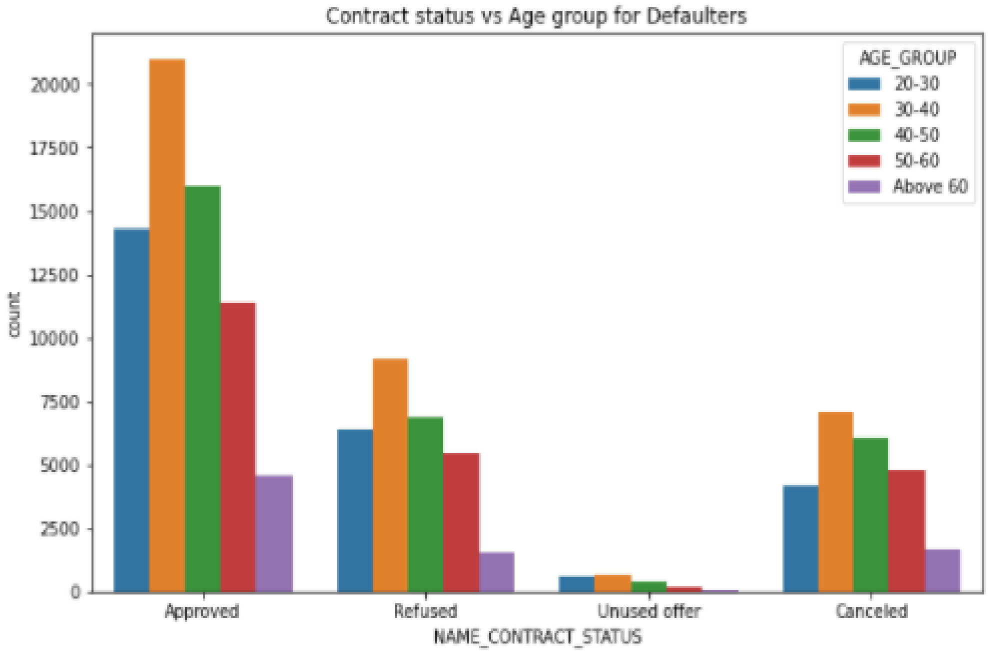
NAME_FAMILY_STATUS vs NAME_CONTRACT_STATUS

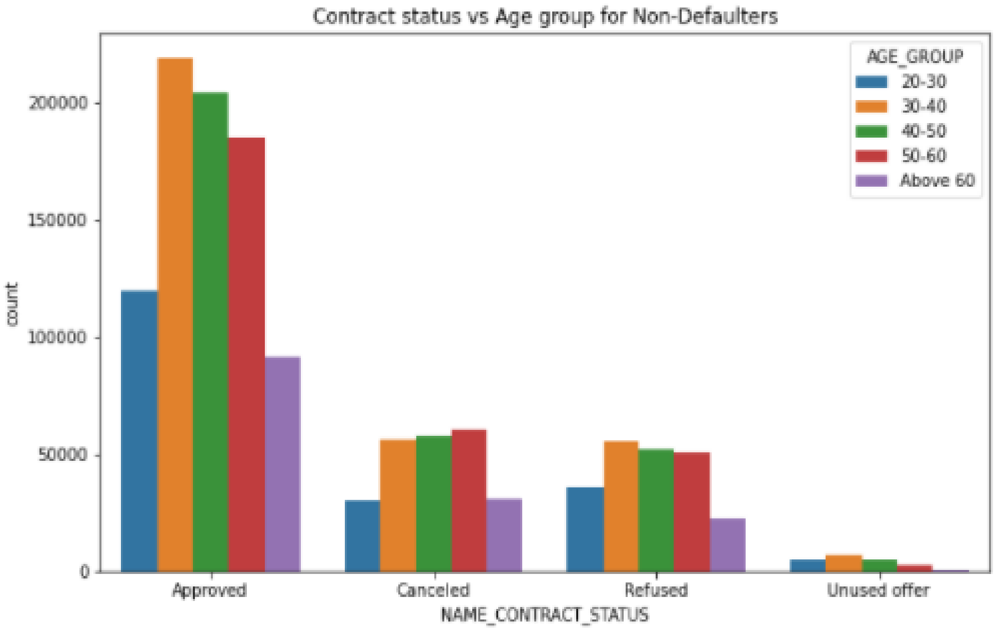




Number of married category of clients for case of both defaulters and non-defaulters are highest in all four types of contract status.
Number of single/not married category of clients for case of both defaulters and non-defaulters are highest in all four types of contract status.

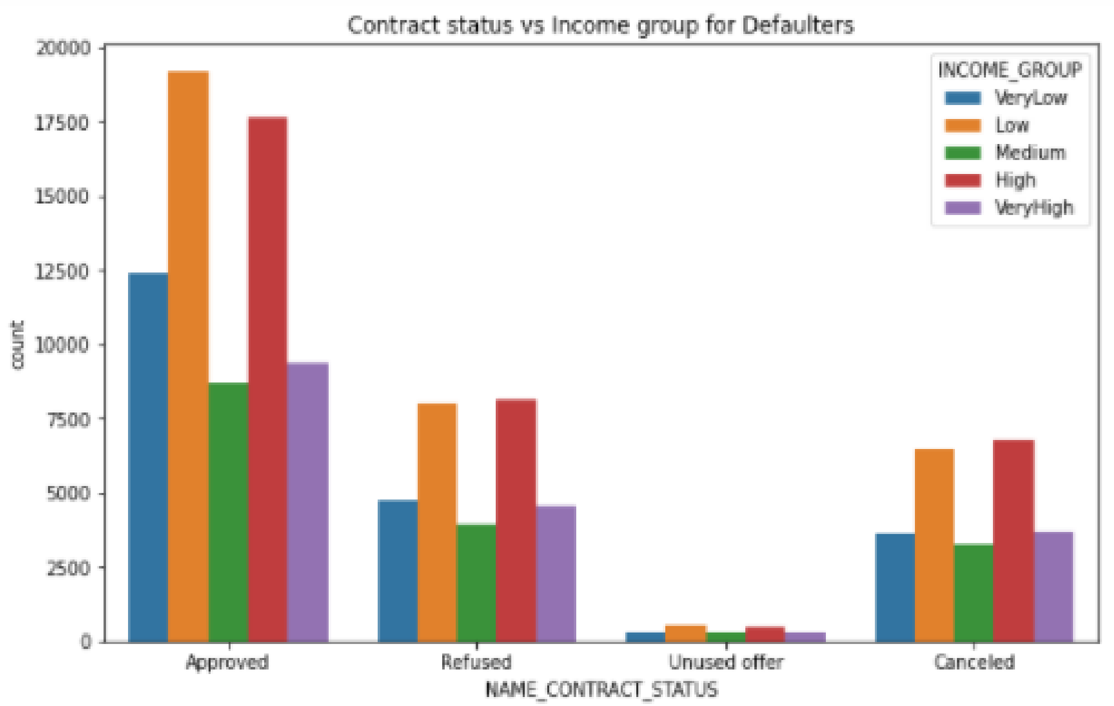
NAME_CONTRACT_STATUS vs AGE_GROUP

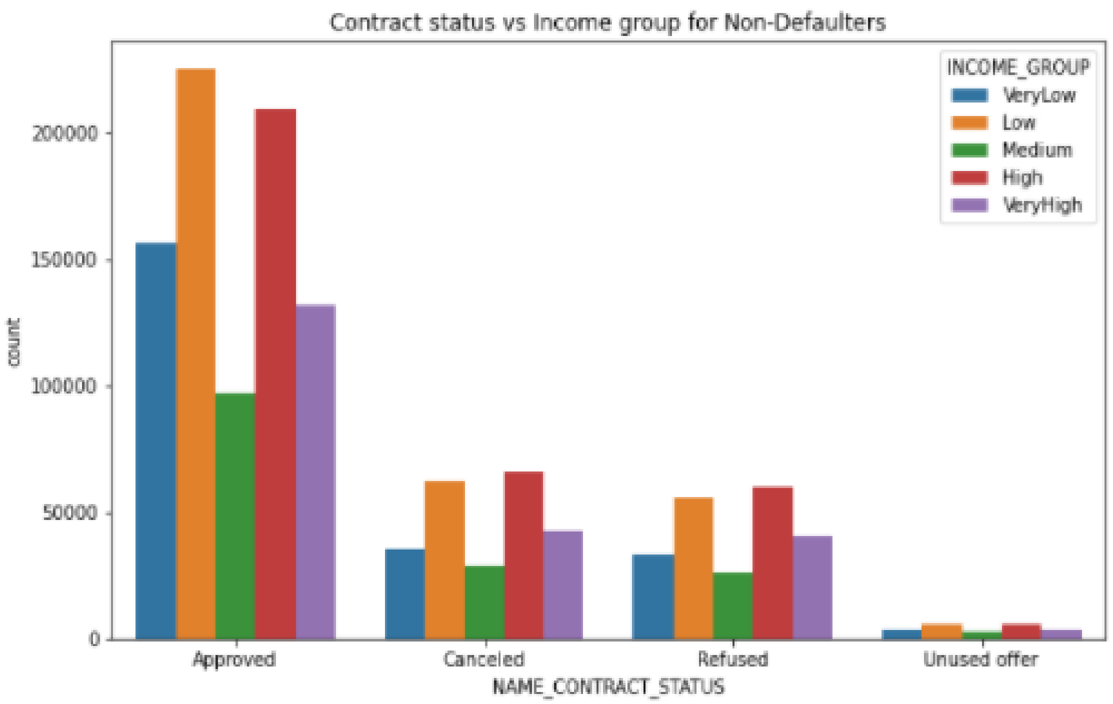




Defaulter clients having 30-40 age group are maximum in all types of contract status. Almost similar trend has observed in case of non defaulter type of clients.

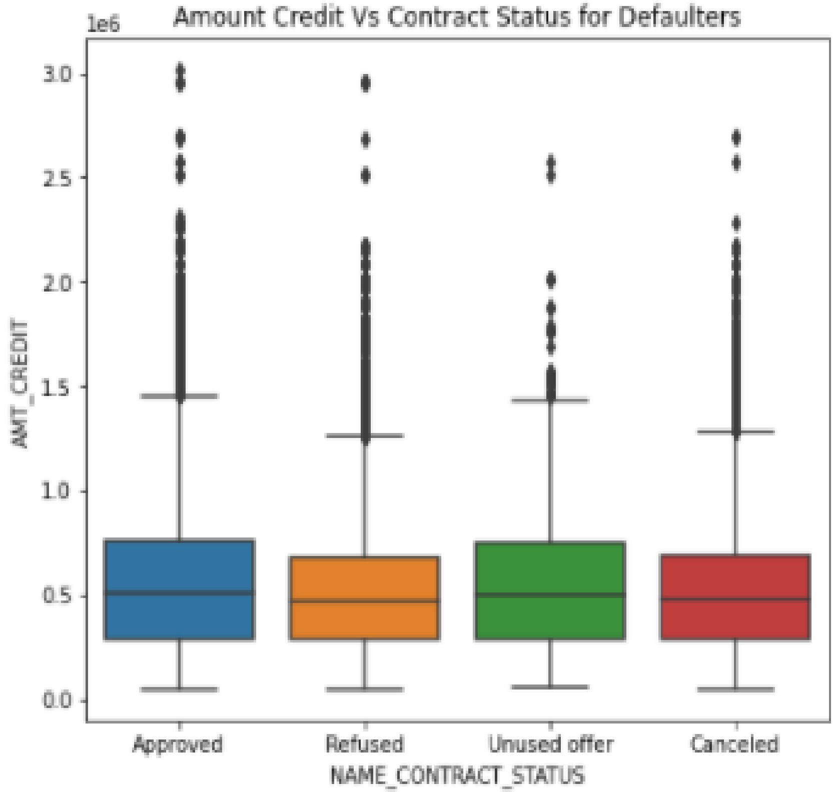
NAME_CONTRACT_STATUS vs INCOME_GROUP

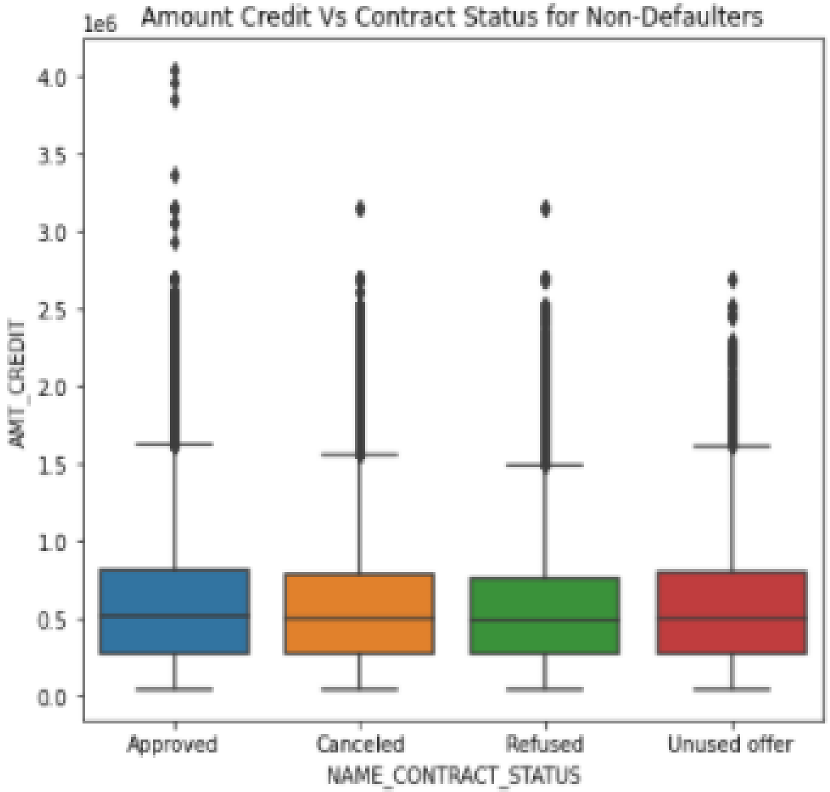




Low income group clients are maximum in number in case of approved category for both the defaulters and non-defaulters followed by high income category clients.

AMT_CREDIT vs NAME_CONTRACT_STATUS

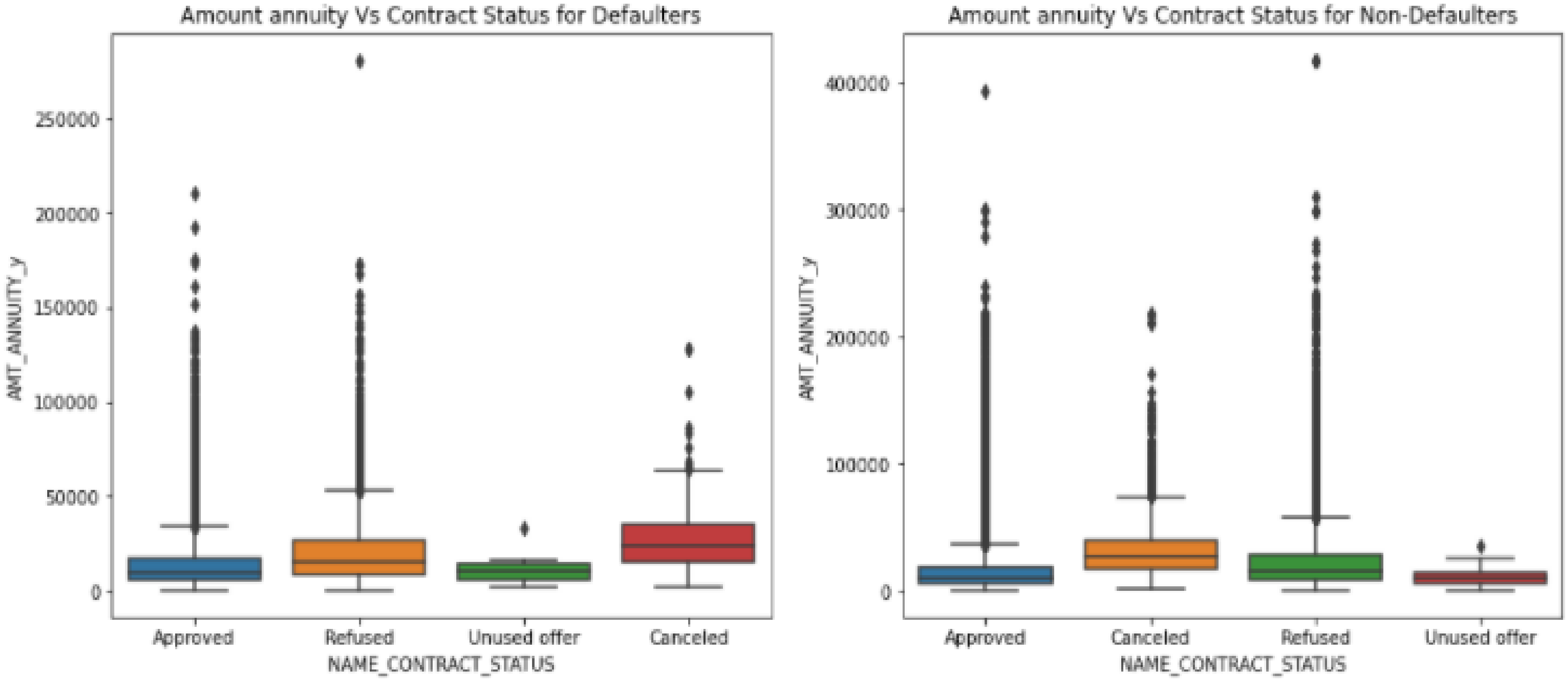




Amount credited in case of approved category of clients has high number of outliers for both defaulters and non defaulters.

Median values of the credited amount are almost similar for all category of defaulter and non defaulter clients.

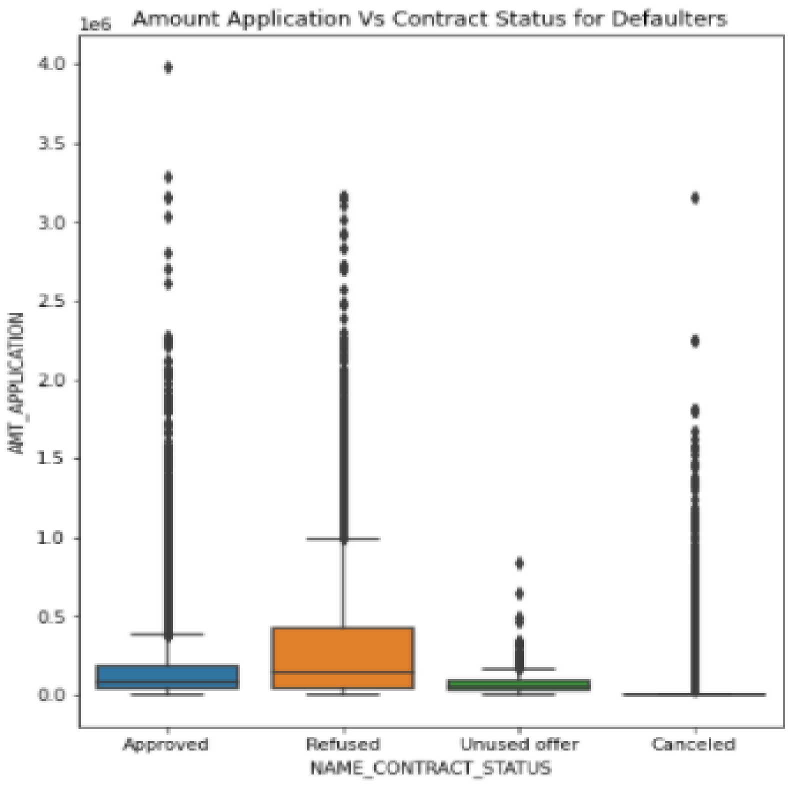
AMT_ANNUITY_y vs NAME_CONTRACT_STATUS

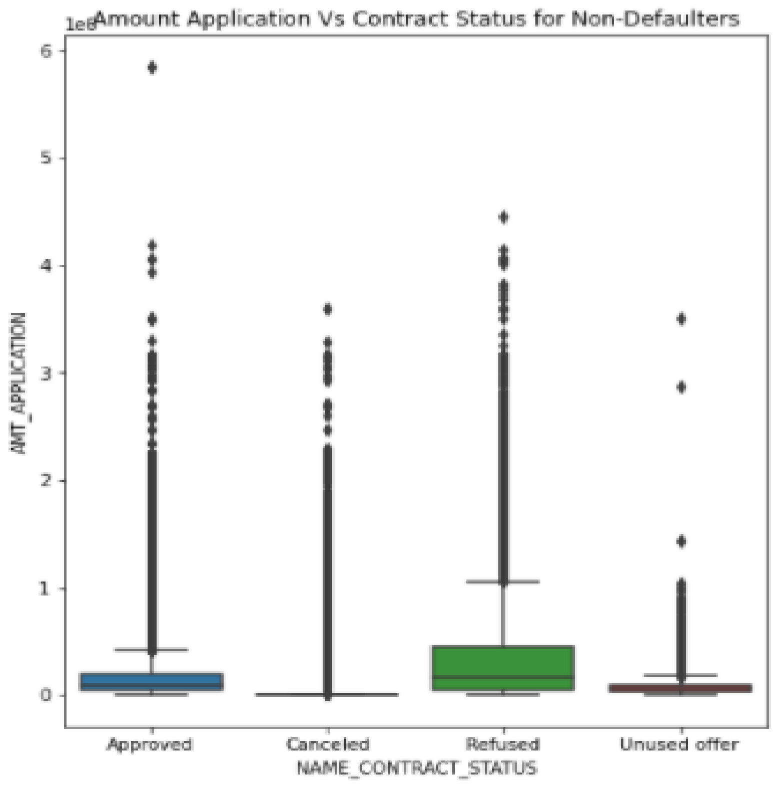


Annuity amount in case of refused and approved category of clients has high number of outliers for both the defaulters and non defaulters.

Annuity amount is highest in case of cancelled type of clients for both the defaulters and non defaulters.

AMT_APPLICATION vs NAME_CONTRACT_STATUS





High application amount has been refused in case of both defaulters and non defaulters

Results & Conclusion

UNIVARIATE ANALYSIS

- The majority of defaulters are associated with cash loan.
- Revolving loans are less risky as 10% of non-defaulter comes from Revolving loans.
- Both defaulters and non-defaulters have no significant difference in terms of car ownership and so this parameter should not be used for risk assessment.
- With the current data pattern, there is hardly any difference between defaulters and non-defaulters based on having realty.
- The percentage of non-defaulters with incomplete higher education is higher than defaulters with incomplete higher education.
- The Housing type is apartment for more than 80% of applications for both the defaulter as well as non-defaulter categories.
- The default percentage is highest among laborers, sales staff and drivers with all these exceeding 10%.

Results & Conclusion

UNIVARIATE ANALYSIS

- IT staff are most secured in terms of lending.
- The majority of loan applicants are from "low" income group.
- The medium income group seems to be most adverse to the concept of borrowing as their % is less in both defaulting and non-defaulting.
- The number of children among defaulting and non-defaulting varies from 0-4 with the majority in both the categories having no child.
- The Average amount credited is more for non defaulters than defaulters.
- The Average goods price amount is more for non defaulters.
- There is no significant value difference observed between defaulters and non-defaulters on the basis of accessibility of mobile phones.
- The data reveals that more than 90% applicants in both the defaulter and non-defaulter did not provide the email id.

Results & Conclusion

UNIVARIATE ANALYSIS

- Clients with Region rating = 1 are comparatively more likely to be non defaulters
- Clients with Region rating = 3 are comparatively more likely to be defaulters.
- There is no significant value difference observed in between defaulters and non-defaulters on the basis of their permanent address and contact address
- The non defaulter clients have a greater number of years of employment than defaulters
- The Age Group 30-40 consists of the majority of loan applicants.
- The 20-30 age group applicants are more likely to turn to be defaulters.
- More than 50% of defaulters as well as non-defaulters come from 2 age groups i.e. 30-40 and 20-30.

Results & Conclusion

BIVARIATE ANALYSIS

- Females taking cash loans are more prone to default than males.
- The married applicants living in apartments are more than the sum of all non-married applicants in the defaulters categories
- The defaulter and non-defaulters follow same pattern in family status and housing type
- For defaulters among very low, low and medium categories females have higher proportion than the males
- For defaulters among High and Very high categories females have lower proportion than the males
- For all income groups females are better places compared to males in the non-defaulter category
- The amount credited is higher for cash loans in both the defaulting and non defaulting category
- The income type for non-defaulters have more sub categories including the student and businessman which are not significant for defaulters
- The maternity leaves category has higher default though the amount credit is lower than other segments

Results & Conclusion

BIVARIATE ANALYSIS

- The data reveals that businessman are less prone to default compared to other categories though the amount credit for them is not the highest
- The defaulting proportion of married applicants is higher along with the average amount credit

Results & Conclusion

MULTIVARIATE ANALYSIS

- The amount credit shows a very high correlation with the amount of good price for defaulters which is expected.
- The amount credit and no of employments years have no significant correlation for defaulters
- The amount of annuity is correlated with amount credit for defaulters
- The count of children are highly correlated with count of family members for defaulters which is inline with the expectation
- The count of children has low correlation with most of the factors for defaulters
- The amount credit shows a very high correlation with the amount of good price for non defaulters.
- The registration years has low correlation with most of the factors for non defaulters
- The Id-years has low correlation with the employment years for non defaulters

Results & Conclusion

MERGED APP DATA WITH PRE APP DATA

- The repairs is the 3rd largest segment in both the defaulter and non defaulter segment
- There are 25 categories of cash loan purpose for both the defaulter and non defaulter group
- The approved is the largest category in both the defaulting and non defaulting segment
- The refused is 2nd largest category for defaulters while the canceled is 2nd largest category for non defaulter segment
- Unused offer is the least contributing category in both the defaulting and non defaulting segment
- The cash through the bank is the highest contributor for both the defaulting and non defaulting category
- 1/5th of the applicants in both defaulting and non defaulting segment are repeaters
- 3/4th of the applicants in both defaulting and non defaulting segment are new
- The POS, Cards and Cash remain the top 3 categories with cumulative contribution of around 90% for both the segments-defaulting and non-defaulting.

Results & Conclusion

MERGED APP DATA WITH PRE APP DATA

- The POS is better compared to cards and cash in terms of repayment as it has a higher % of non-defaulters.
- 1/4th of the defaulters and non defaulters are x-sell
- The proportion of defaulters in walk-in is higher compared to non defaulters
- The Credit and cash offices, country wise and store remain the top 3 categories for both the defaulting and non-defaulting segments.
- Both the defaulters and non defaulters have same 8 categories with the similar contribution from each of the categories
- Number of married categories of clients for case of both defaulters and non-defaulters are highest in all four types of contract status.
- Low income group clients are maximum in number in case of approved category for both the defaulters and non-defaulters followed by high income category clients.
- Amount credited in the approved category of clients has a high number of outliers for both defaulters and non defaulters.