EDA CASE STUDY
CREDIT RISK ANALYSIS

By- Mohan Babu Uppu and Subham Kumar Behera

Problem Statement

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 company wants to understand the driving factors (or driver variables) behind loan default, i.e., the variables which are strong indicators of default

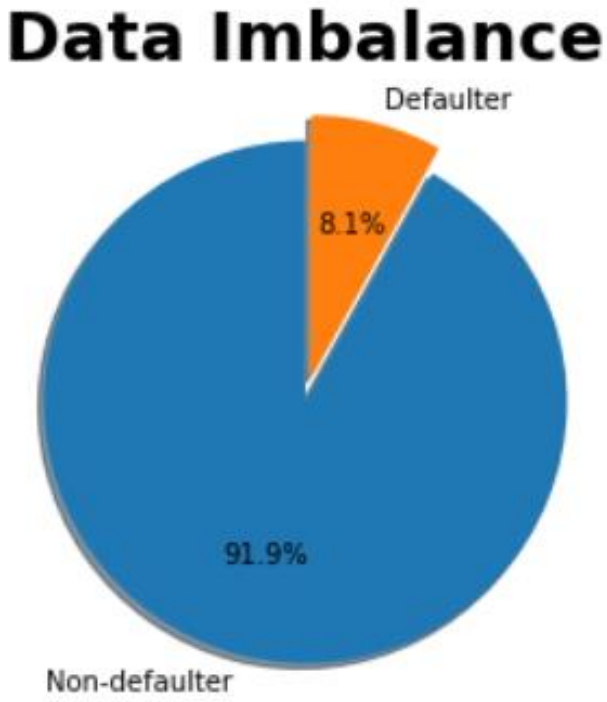
EDA Techniques Used :

- Importing the input data file i.e., “application_data.csv” and "previous_application.csv“ and assigning to data frames
- Checking the size and structure of the data
- Reading the data for number of columns, their data types and finding the unique values in them
- Performing data quality checks for null values and dropping columns with more than 50% of null values
- Further data quality checks to drop unrequired columns with insufficient information or no information related to the business problem (columns : NAME_TYPE_SUITE, TOTALAREA_MODE etc.)
- For columns which has less percentage(around 13%), imputing the missing /null values with mode or 0 (columns : AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY etc.)
- Converting the data types of columns as per requirement (columns : AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY etc.)
- Binning of continuous variables to categories for further analysis (columns : AGE_YEARS, AGE_BUCKETS, CHILDREN_BUCKETS, FAMILY_BUCKETS)
- Checking for outliers in the numerical columns like AMT_INCOME_TOTAL, AMT_CREDIT through box plots
- Analyzing data in next stages for outliers and trends through bar charts, pie charts, scatter plot and box plots for different columns by univariate, bivariate and multivariate analysis

Analysis:

Checking the imbalance percentage using Target variable

Inference : As we can see the percentage of Non defaulters in 91.9% and for defaulters is 8.1% .

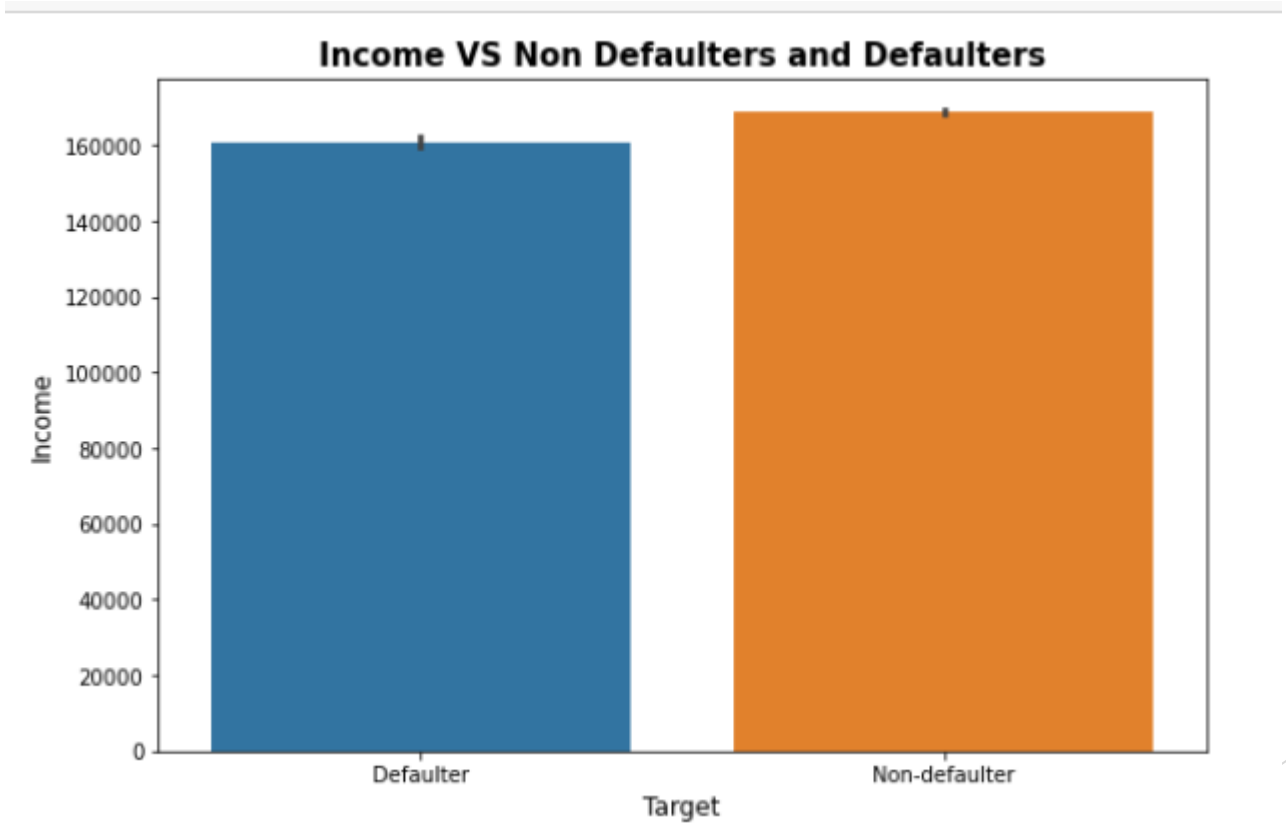


Splitting the data set into separate data frames for defaulters and non defaulters and performing further analysis

UNIVARIATE ANALYSIS

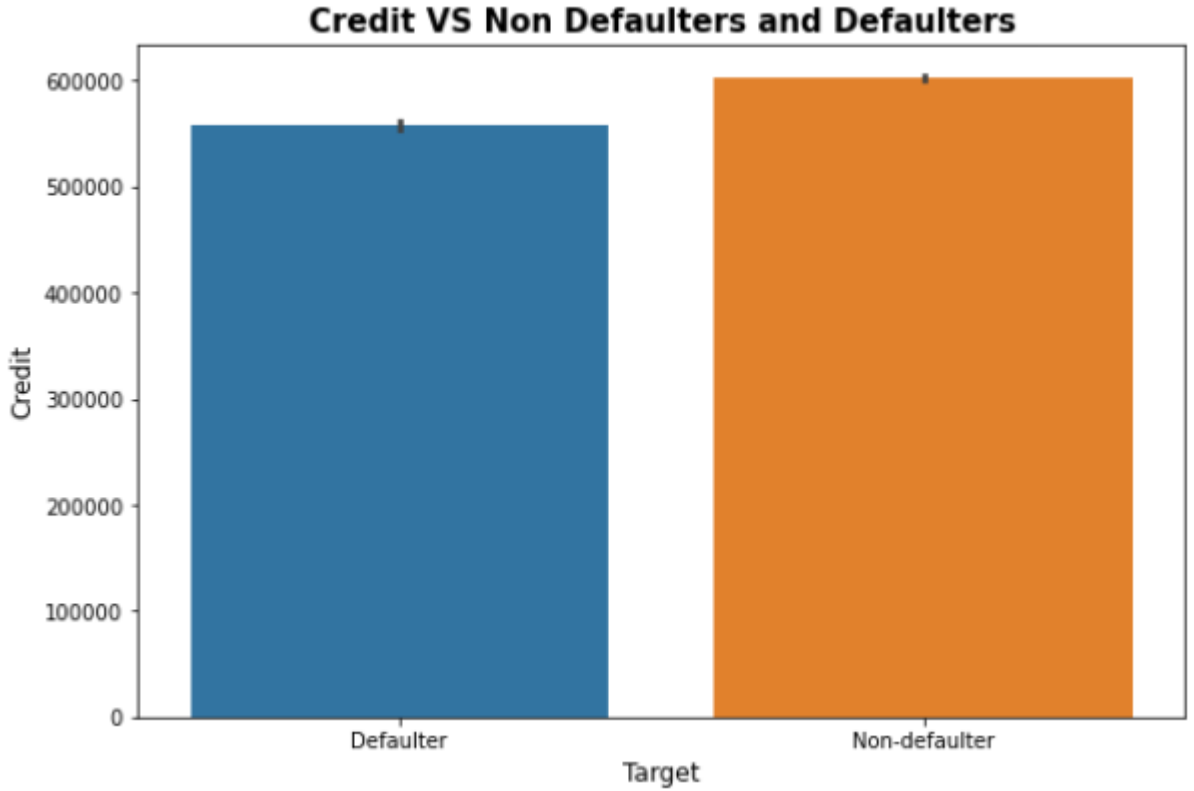
Plotting Income of Defaulters and Non-Defaulters

Non-Defaulters clearly have more income than defaulters and hence can repay their loans.



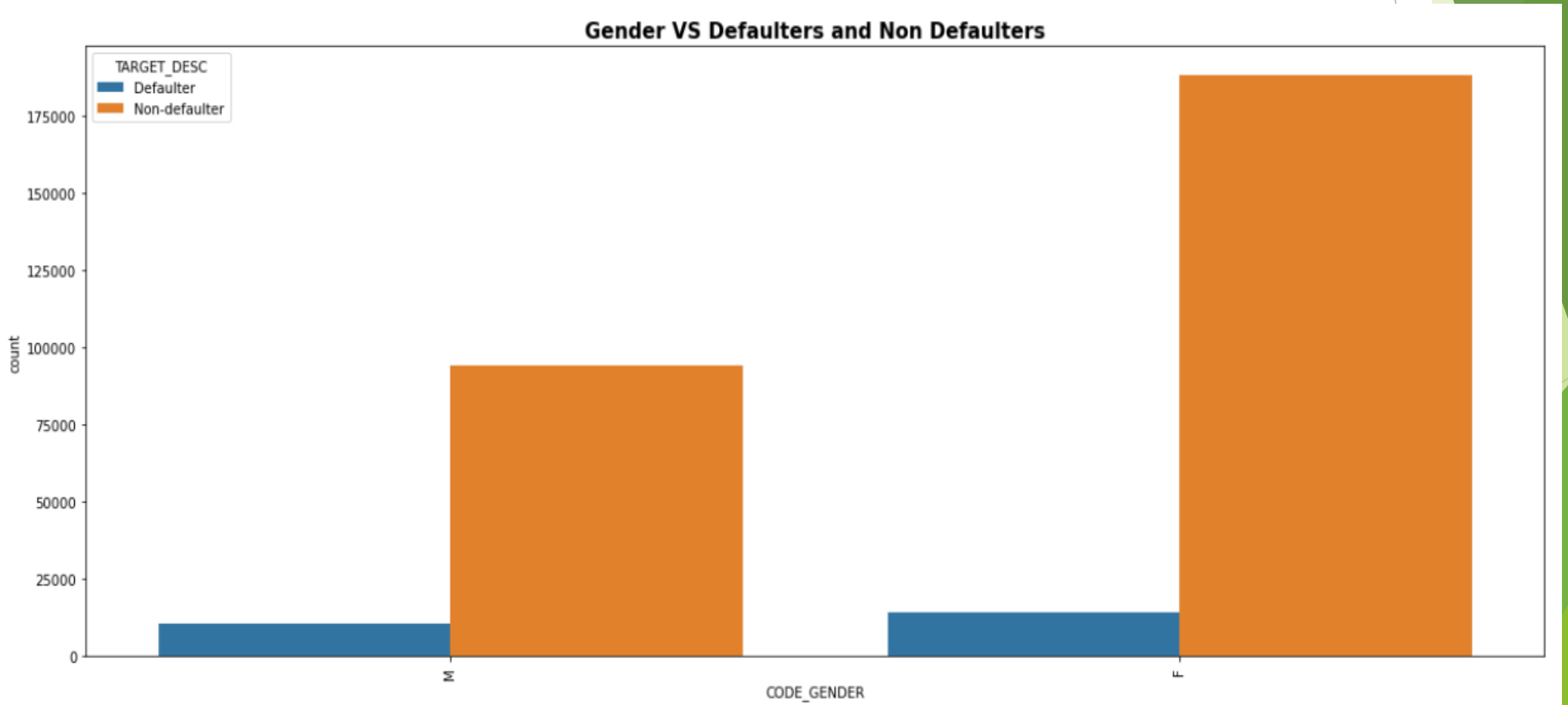
Plotting Credit of Defaulters and Non Defaulters

Credit taken by Non defaulters is higher than Defaulters as they have more potential to repay loans.



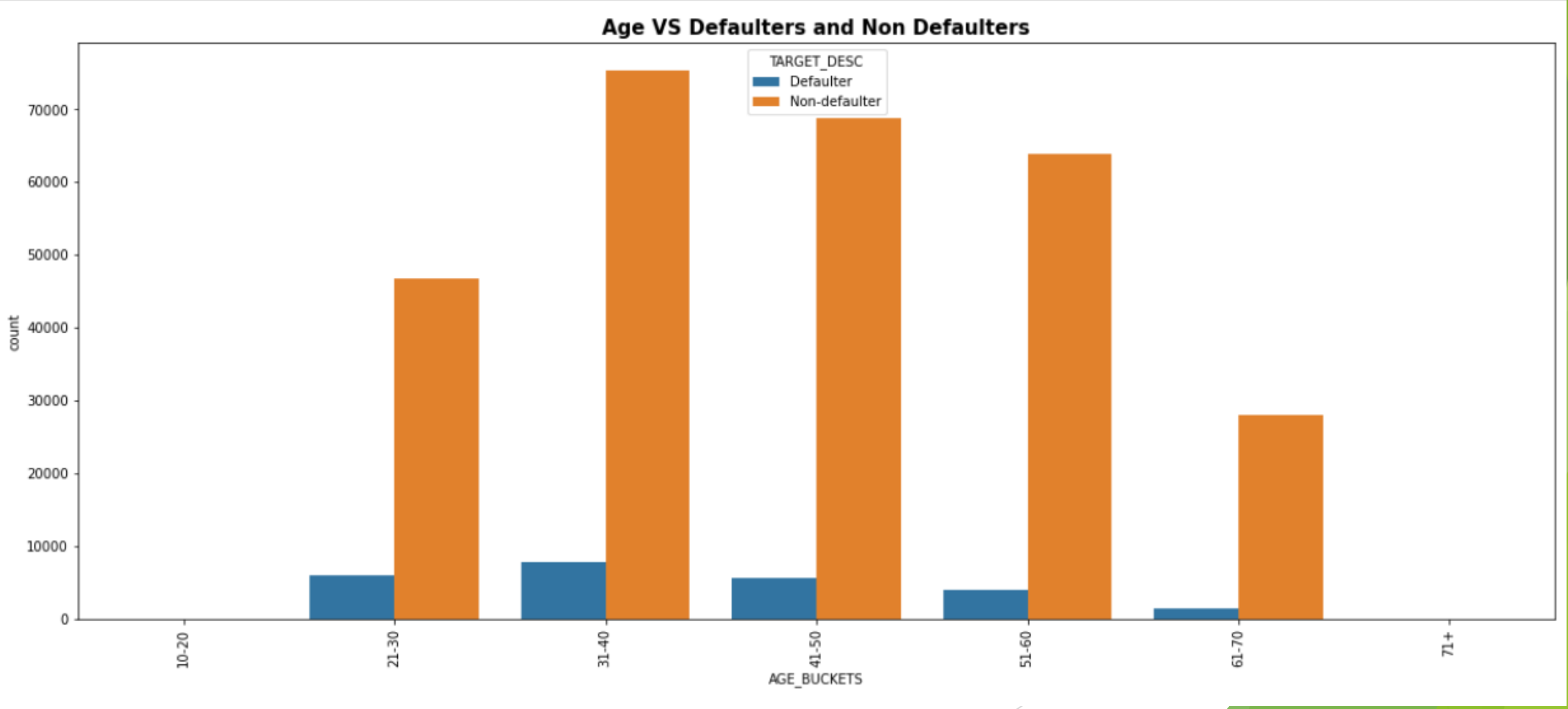
Studying the Gender of Defaulters and Non Defaulters

The number of females is higher in both the defaulters and non defaulters list. Also, More females have applied for loans than males. We can also conclude that bank has approved more loans for females than males.



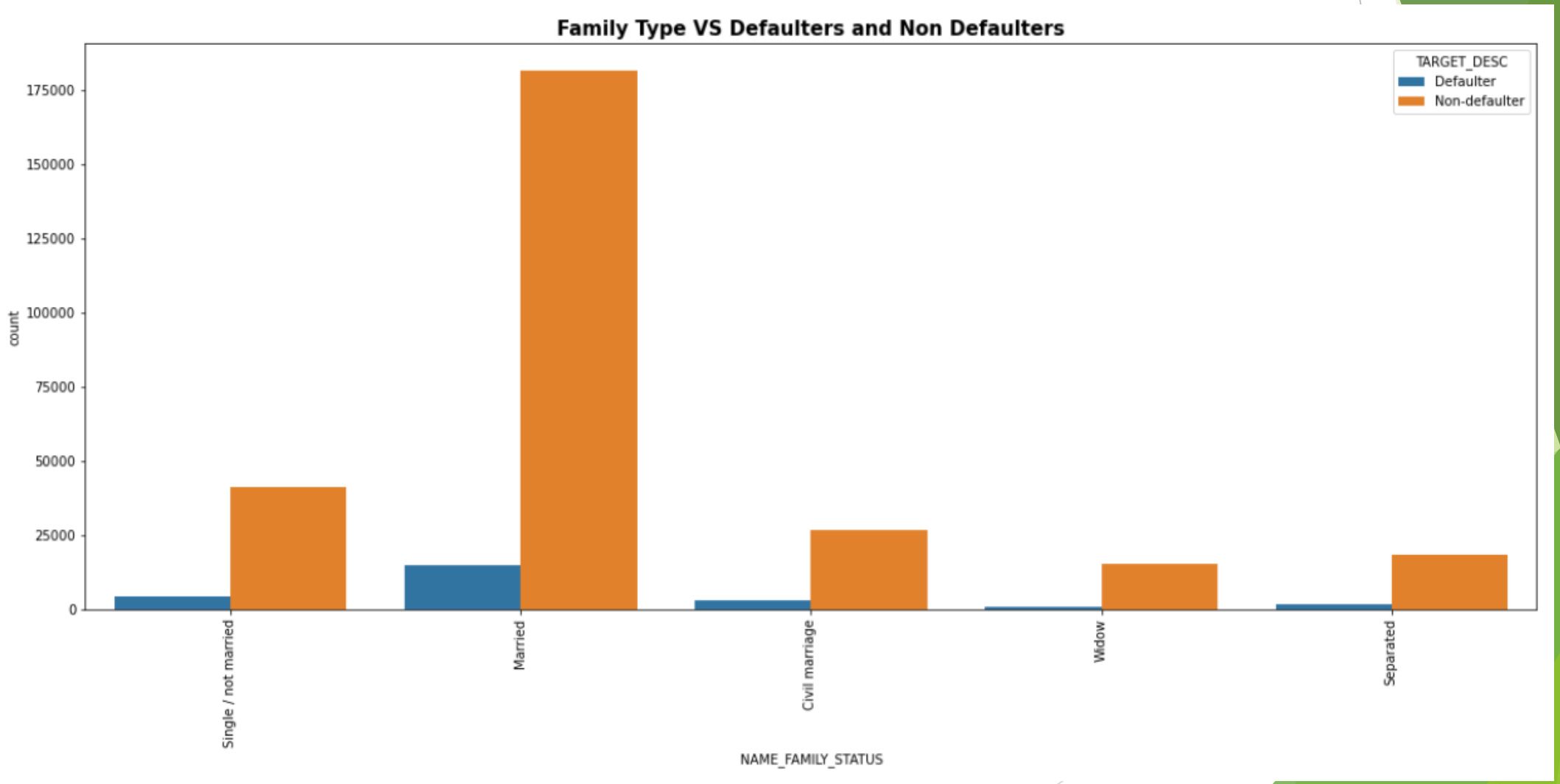
Studying the Age groups of Defaulters and Non Defaulters

The highest number of defaulters and non defaulters is in age band 31-40. One of the reasons might be that this age group has highest number of applicants. We have less number of defaulters and non defaulters for applicants in the age band 61-70 as the number of applicants of loans in old age group is very less as well as the requirement of loan at this age is also less. If we compare , the proportion of defaulters is almost same in the age bands 21-30 and 41-50.



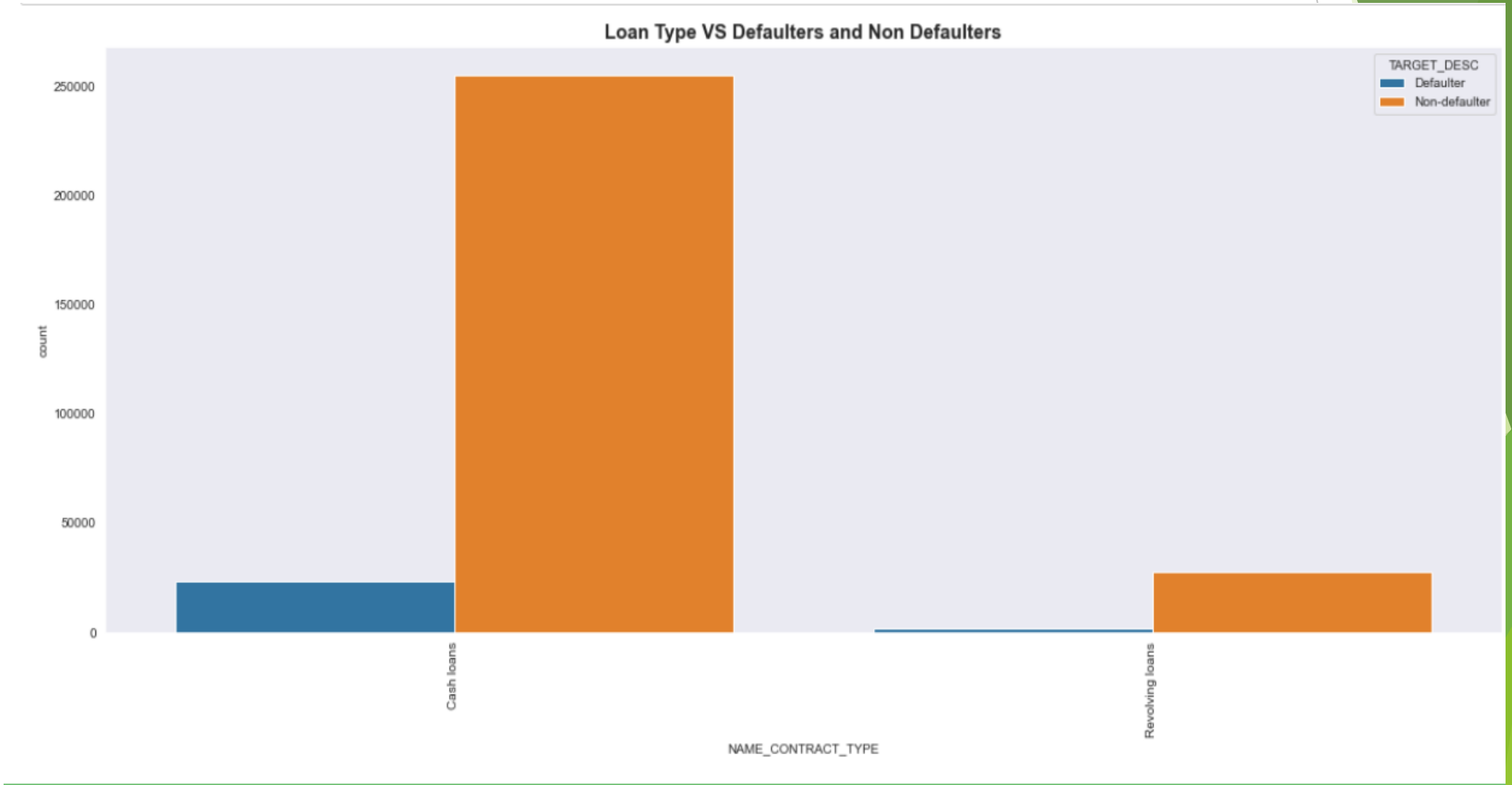
Studying the Family type of Defaulters and Non Defaulters

The number of married people in both defaulters and non defaulters is highest considering the need of loan is highest for this group. Single/not married people have more proportion of defaulters than non defaulters if we compare both the graphs hence we can conclude it is safer to provide loans to married applicants.



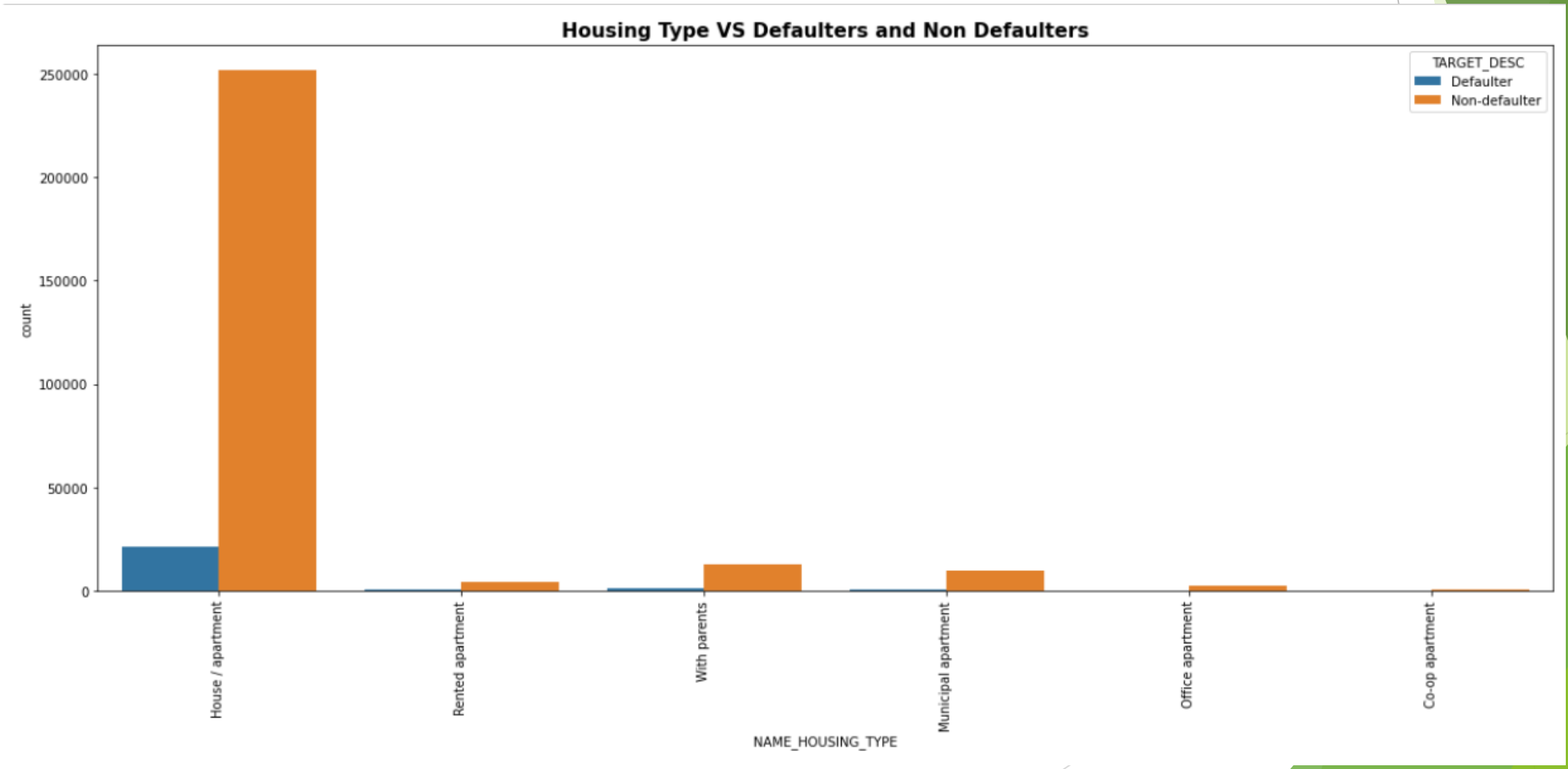
Finding the loan type of Defaulters and Non Defaulters

Majority of applicants are having cash loans. However, the number of non defaulters is more in Revolving loans hence we can conclude revolving loans are safer than cash loans for the bank since the credit limit is already fixed by the Bank. Credit cards are one of the example of Revolving loans.



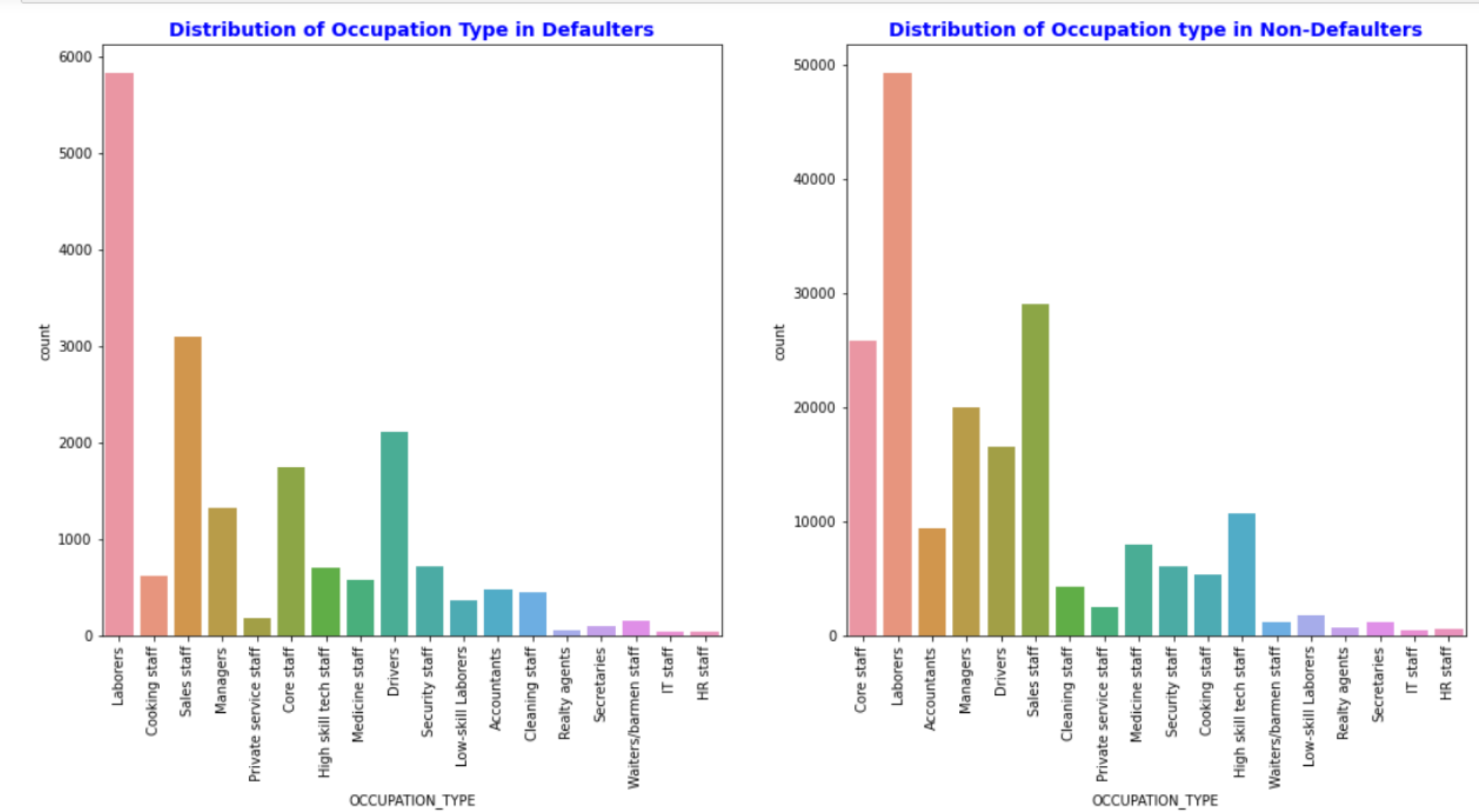
Housing Type of Defaulters and Non Defaulters

People living in Municipal apartments have more defaulters than non defaulters. Reason might be lower wages and hence unable to afford rented apartments or to own houses. People living in Rented apartments have more defaulters due to more investment every month for the house. We have most applicants of loans from people with House/Apartments.



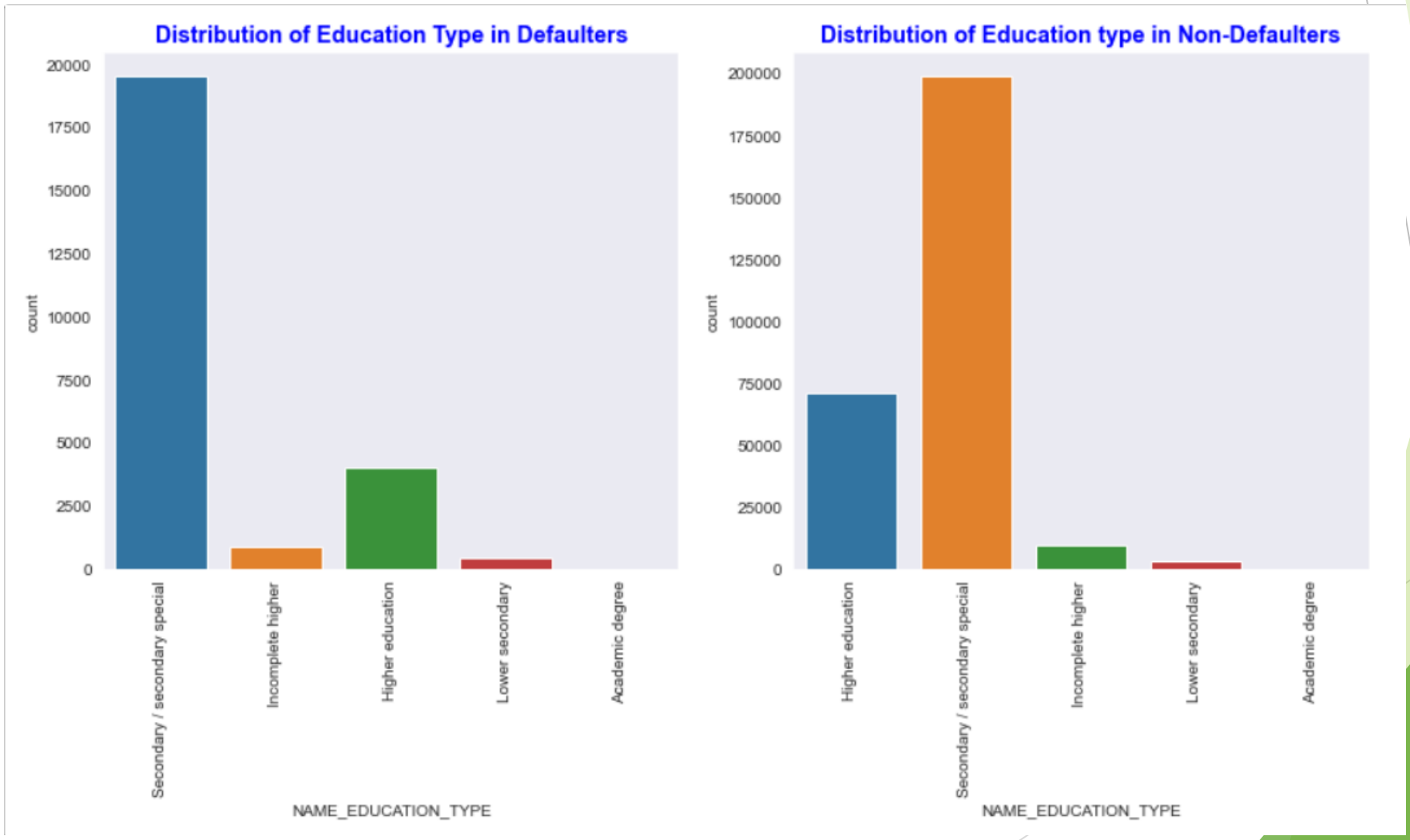
Occupation type of Defaulters and Non Defaulters

If we refer both the graphs then clearly laborers, sales staff, Drivers, core staffs etc have more applicants for loans irrespective of defaulters and non defaulters. This might be due to the lower wage than other high paying jobs like IT staff, HR staff, High Skill tech staff. Sales staff and core staffs have proportionally more non defaulters as well.



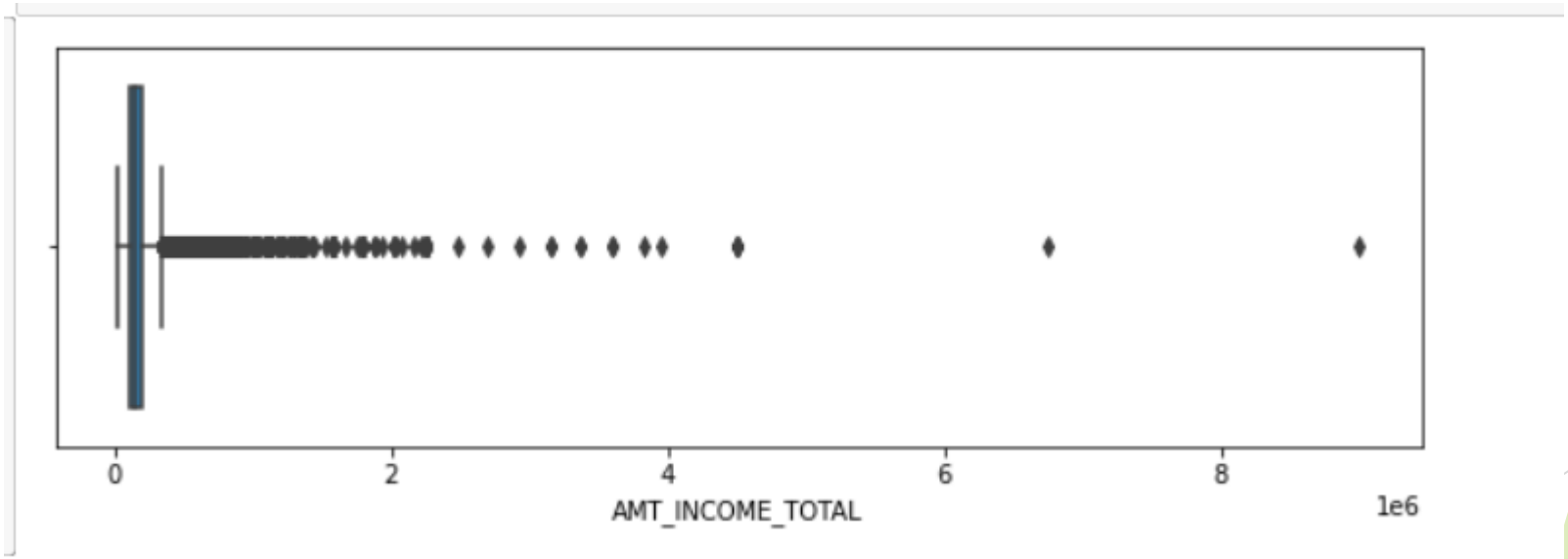
Education type of Defaulters and Non Defaulters

The number of applicants is highest in Secondary/secondary special education type as this group is likely to have less paying jobs for comparatively lower qualification and hence the defaulters are also highest in this group. Comparing both the graphs, applicants with Higher Education has more non defaulters.



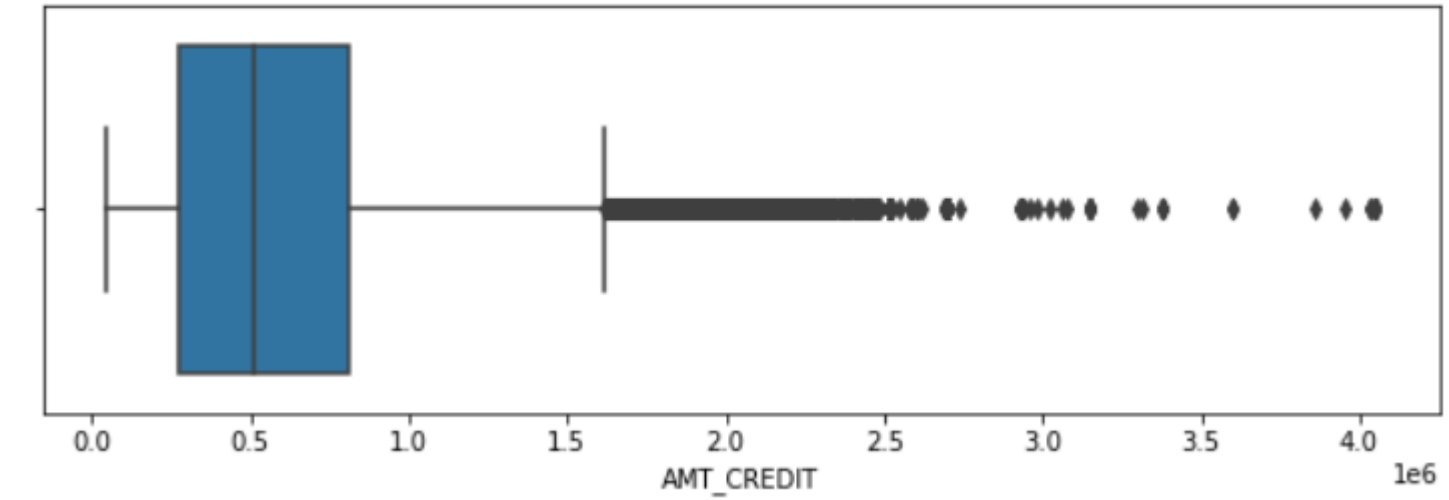
Checking for outliers in AMT_INCOME_TOTAL

We have many outliers in this column. This might be due to the applicants which have very high income. After observing the data one applicant with income 117000000 is skewing the over income analysis; for this exercise, we have modified the data to to 1170000 assuming this is a data error and also considering that an applicant with 117 million income would be defaulter requesting a loan of 562K



Checking for outliers in AMT_CREDIT

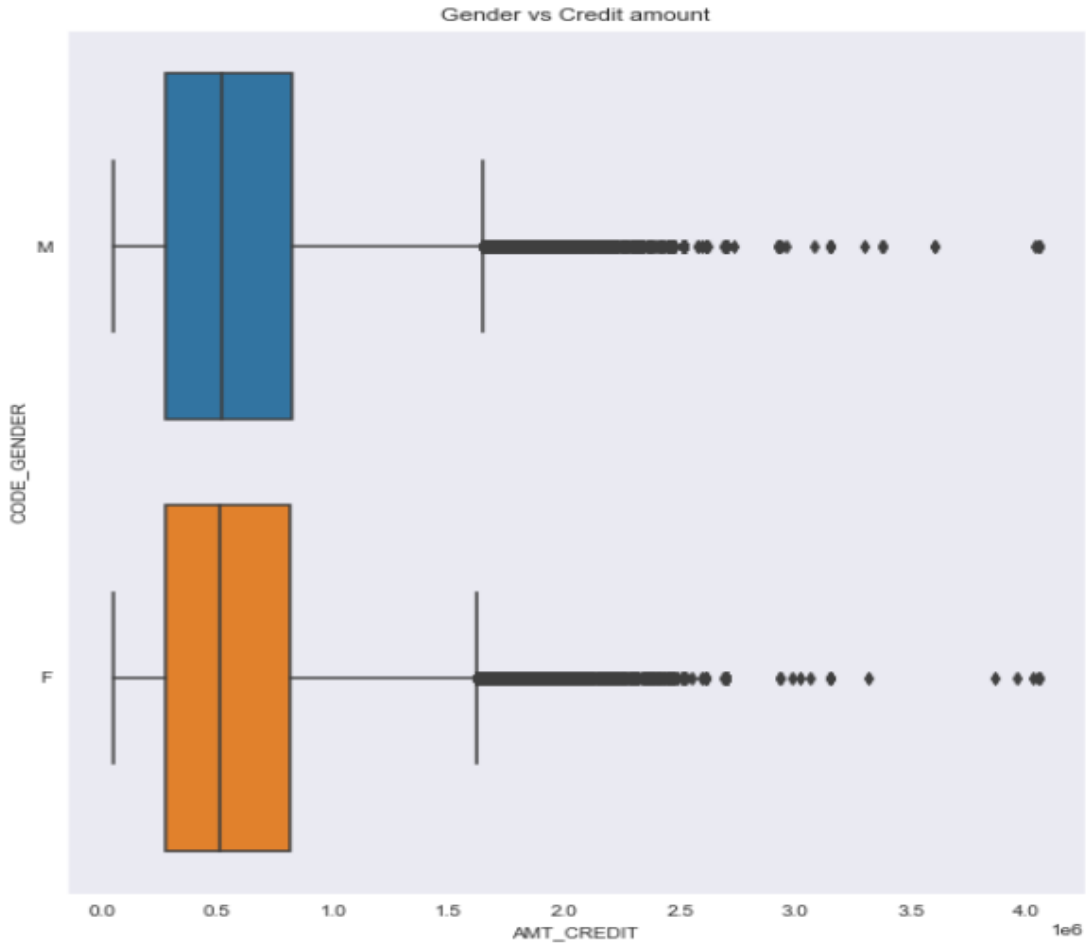
We can see a lot of outliers considering the high credit amount requested by applicants with high income.



BIVARIATE ANALYSIS

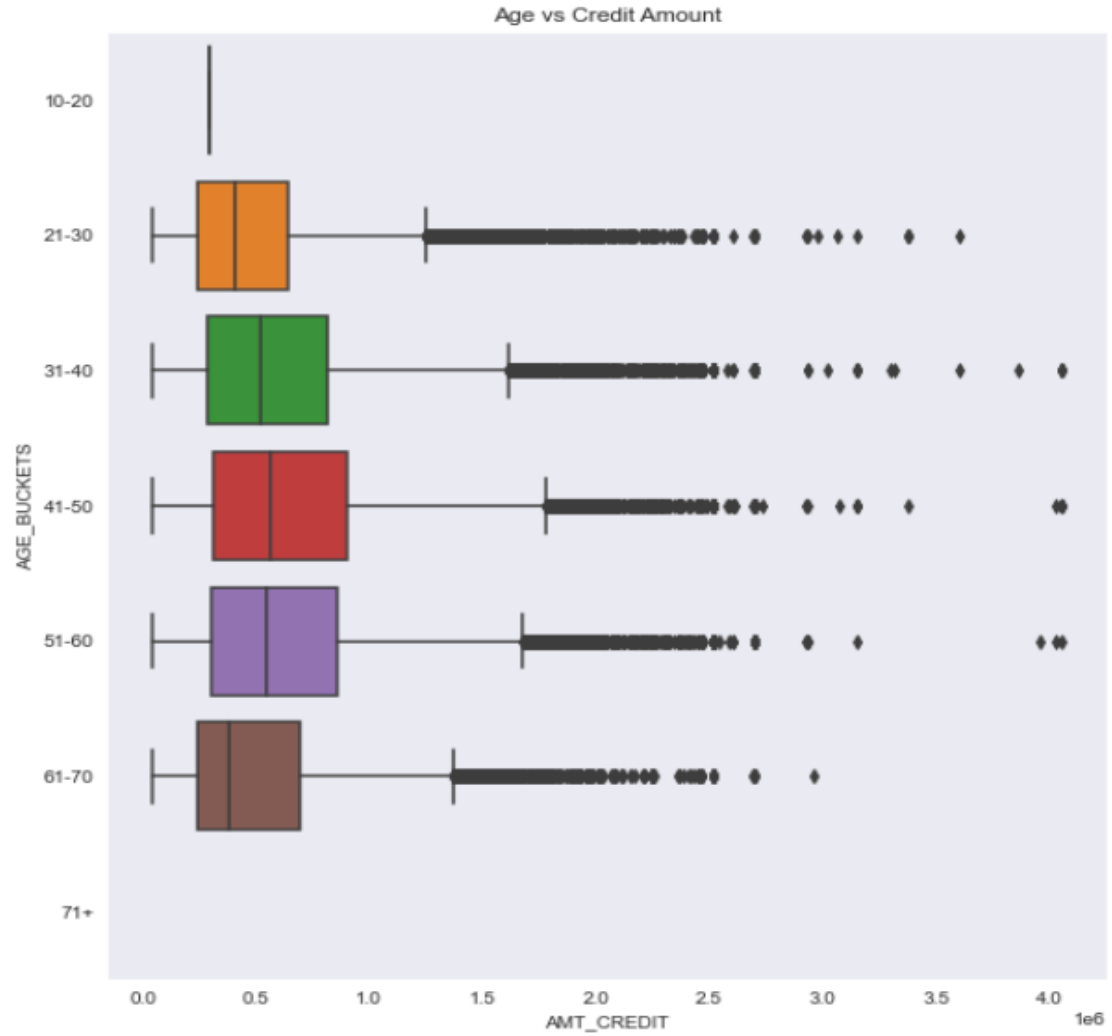
Graph of Gender VS Credit Amount

In case of Gender, the credit amount is having extreme values for both male and female. However, the females have more number of applicants towards the maximum Credit values. The median and quartiles are almost similar in both the genders.



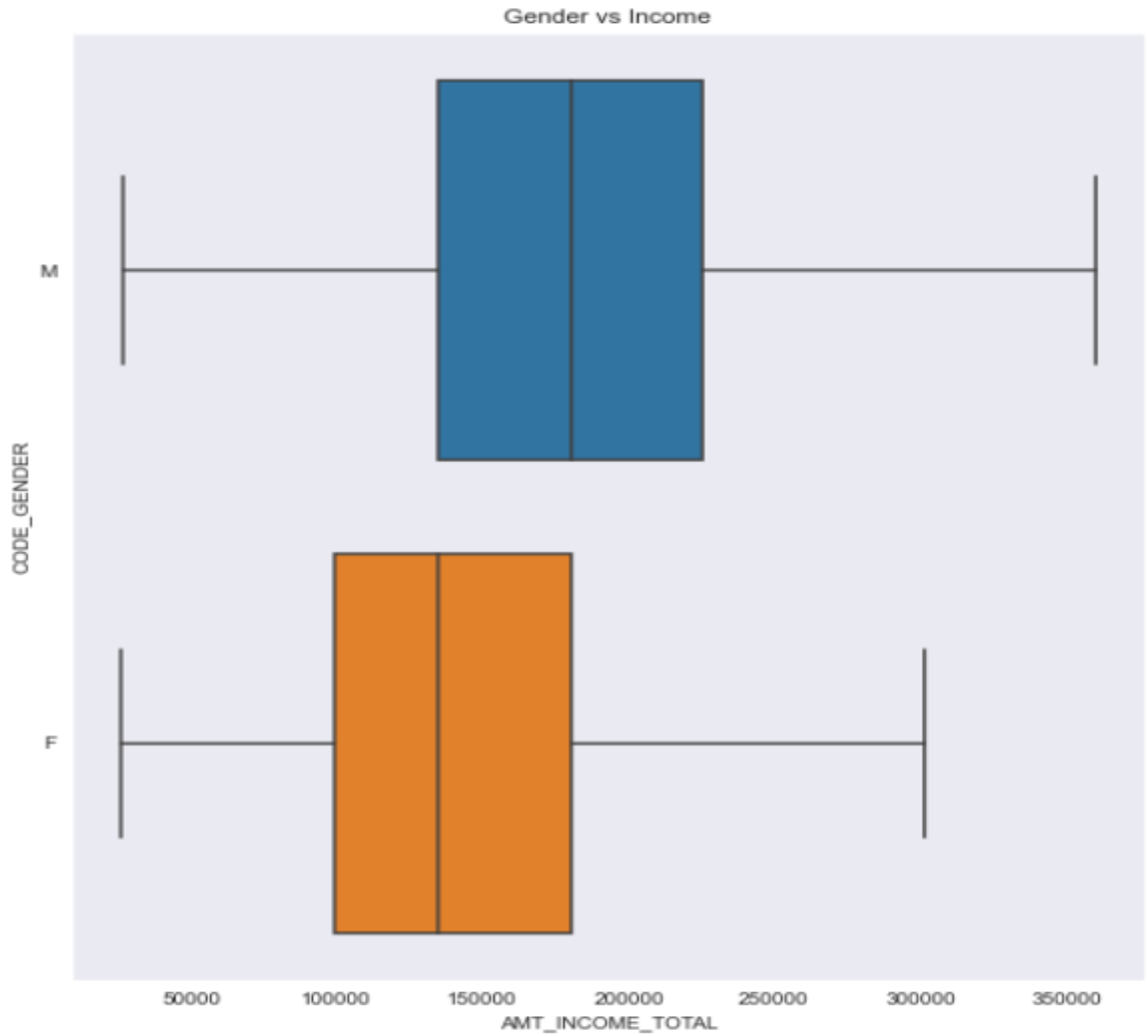
Graph of Age VS Credit amount

In case of Age groups, the least amount of Credit is in the 61-70 age group. However the median and quartile values are low hence, major proportion of the applicants have taken a lesser amount of credit compared to other age groups. The age groups 31-40, 51-60 and 41-50 have applicants with very high credit amount.



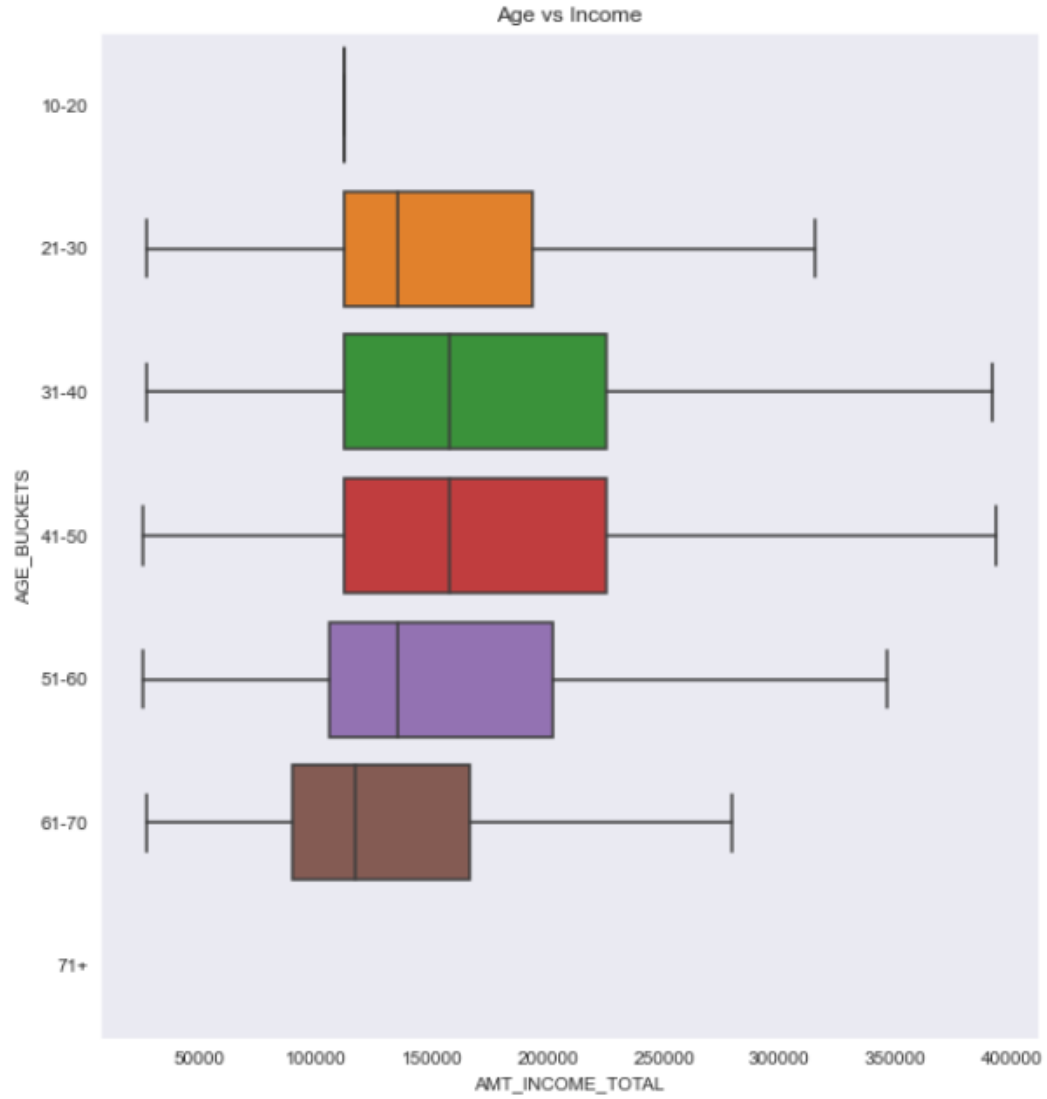
Graph of Gender Vs Income

Males are having more applicants with high income as compared to females.



Graph of Age Vs Income

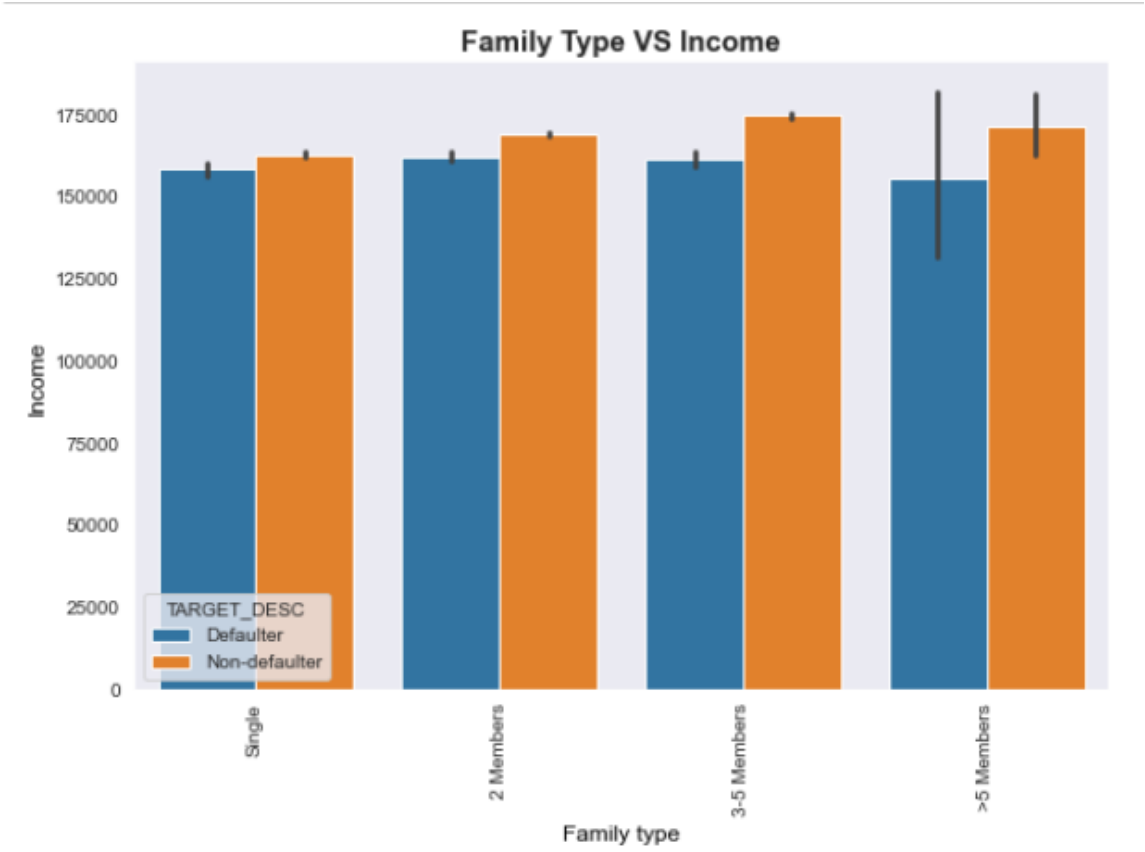
In case of age bands, 21-30 and 41-50 have comparable applicants with high income. The age band 61-70 is having least extreme income amount. The applicant with highest income amount seems to be in the 41-50 age band.



MULTIVARIATE ANALYSIS

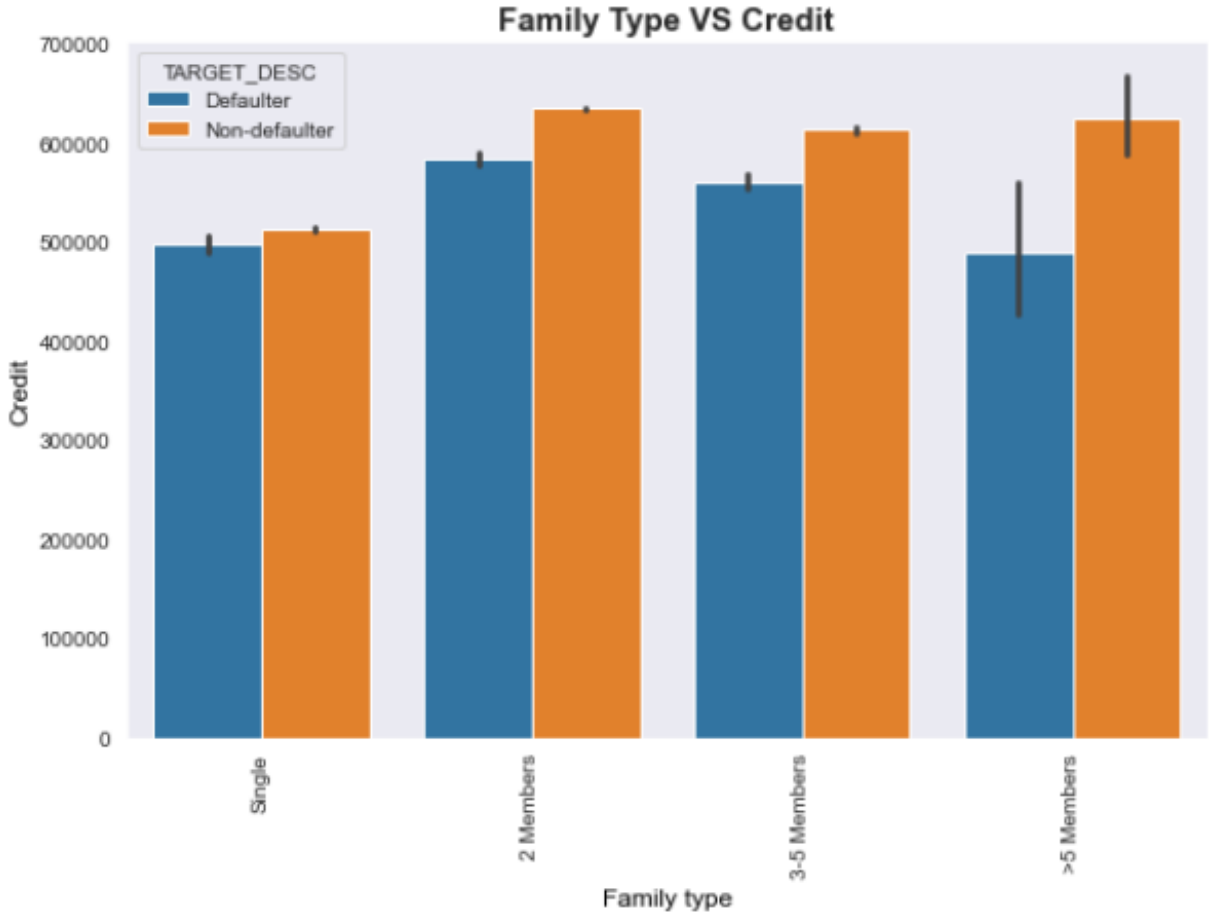
Studying Family type and Income across Defaulters and Non Defaulters

The income of non defaulters is higher than defaulters. Income is proportionally higher in families having 3-5 members and the non defaulters are also high in this group. The income of defaulters is lowest in case the number of family members is more than 5.



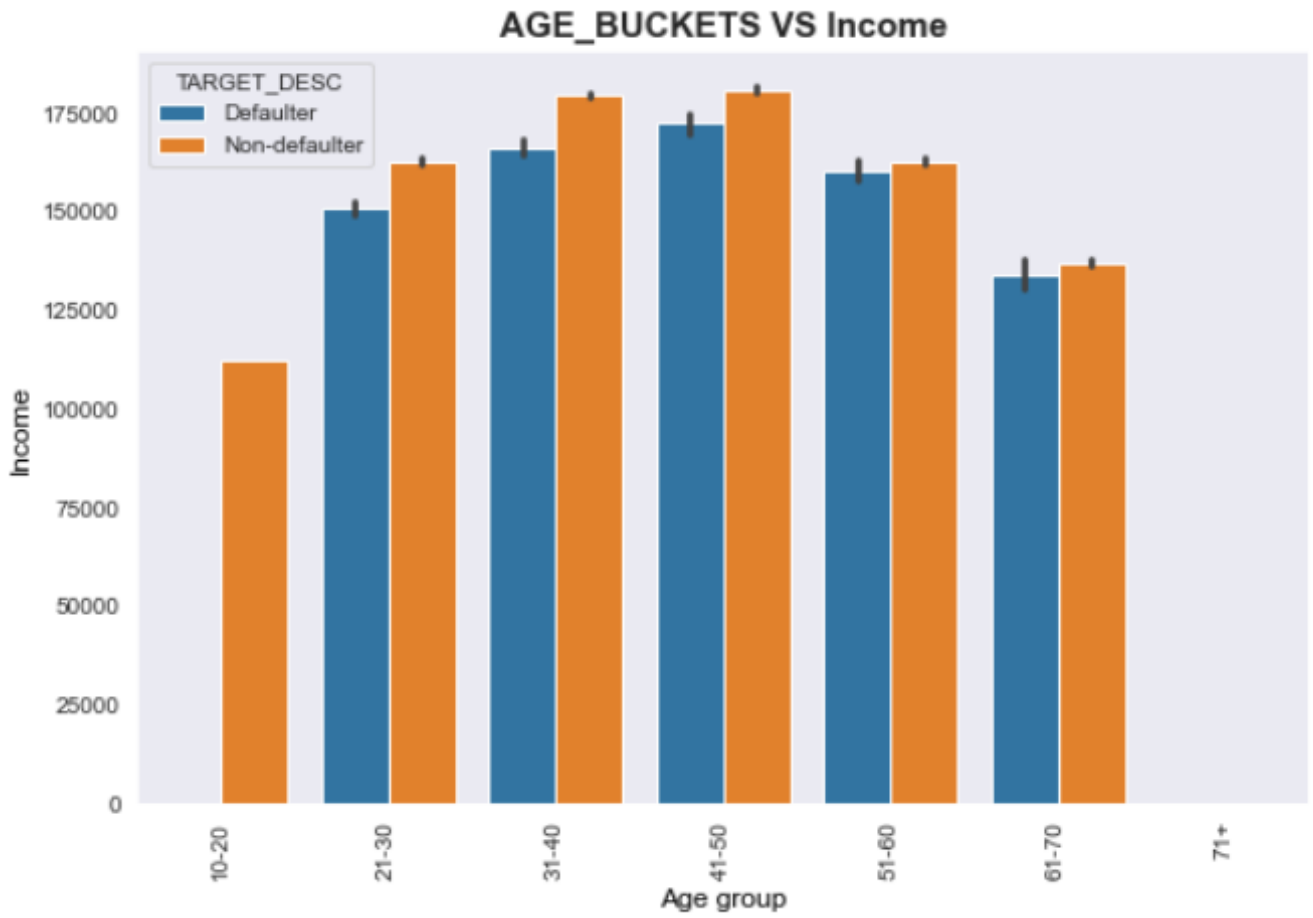
Plotting Family type and Income across Defaulters and Non Defaulters

families of 2 members is having highest credits and the number of non defaulters is also high. The lowest number of defaulters is in the group having more than 5 family members. Also the number of non defaulters is almost same for applicants having 2 or more than 2 family members. Single applicants have less credits than applicants having families irrespective of being Defaulters and Non defaulters



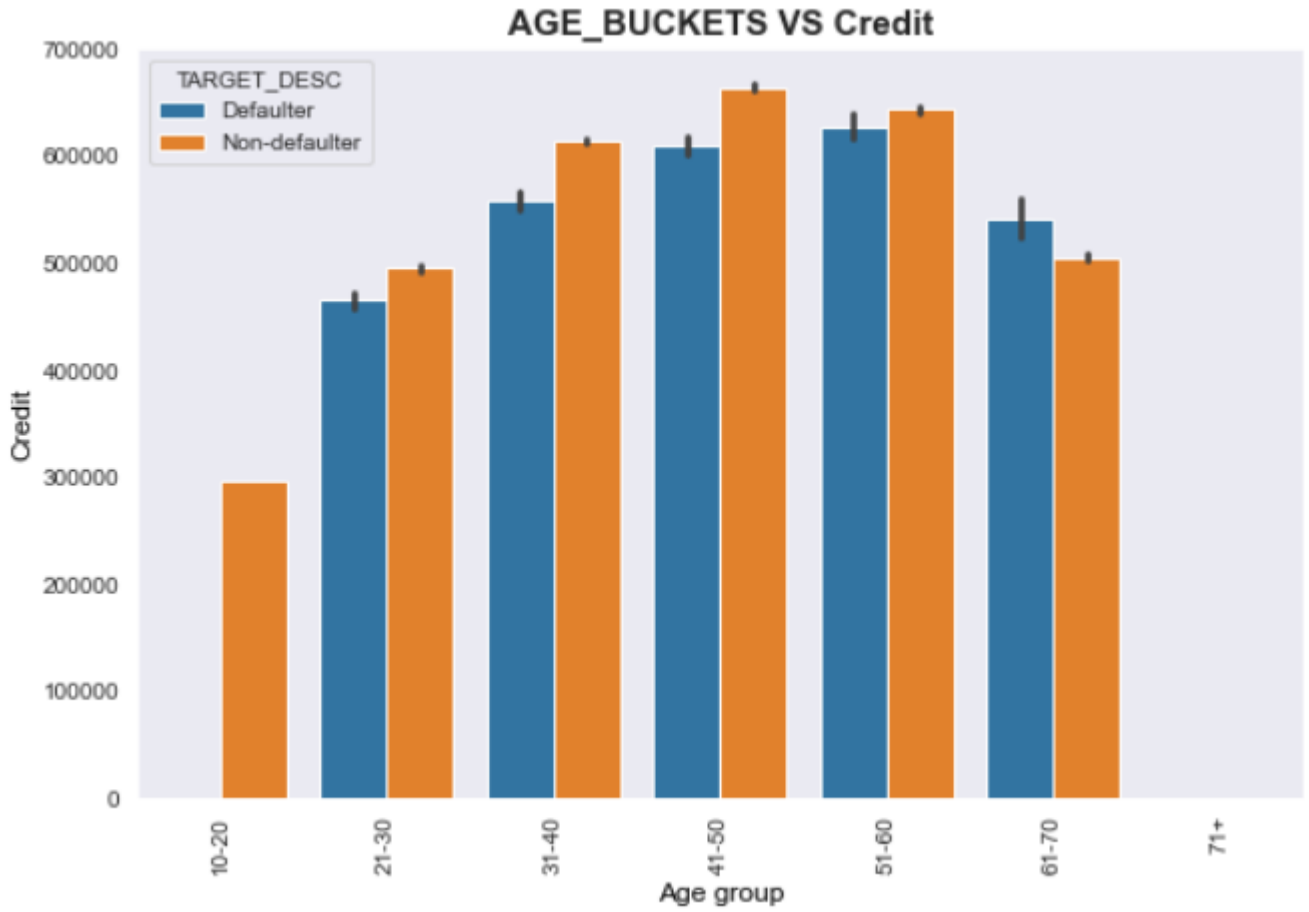
Plotting Age and Income across Defaulters and Non Defaulters

The applicants in age group 31-40 and 41-50 have maximum income. However the defaulters are more in the 41-50 age group. The number of defaulters and non defaulters are almost same in the 51-60 age group if we consider incom



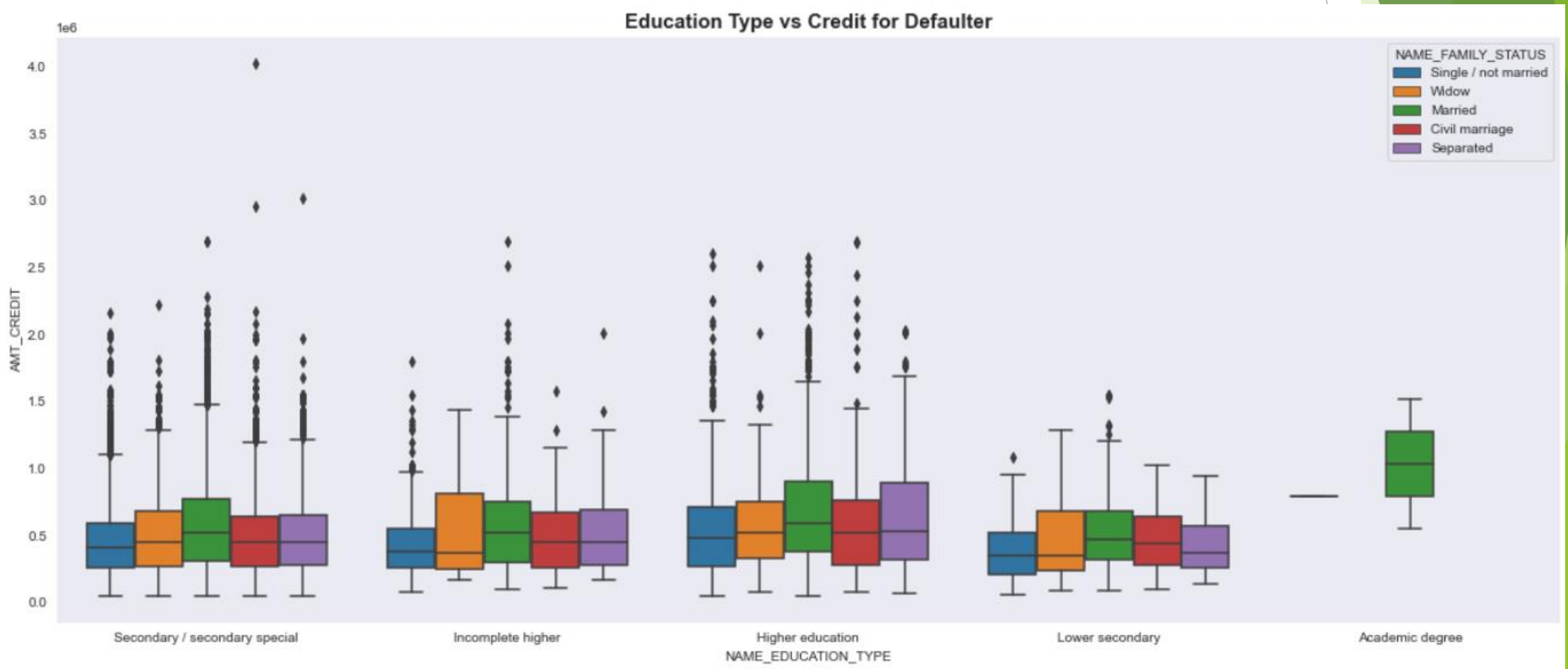
Plotting Age and Credit across Defaulters and Non Defaulters

In case of credit, the highest number of defaulters are in the age group 51-60 and the highest number of non defaulters are in 41-50 age group. We have more defaulters than non – defaulters in the 61-70 age group.



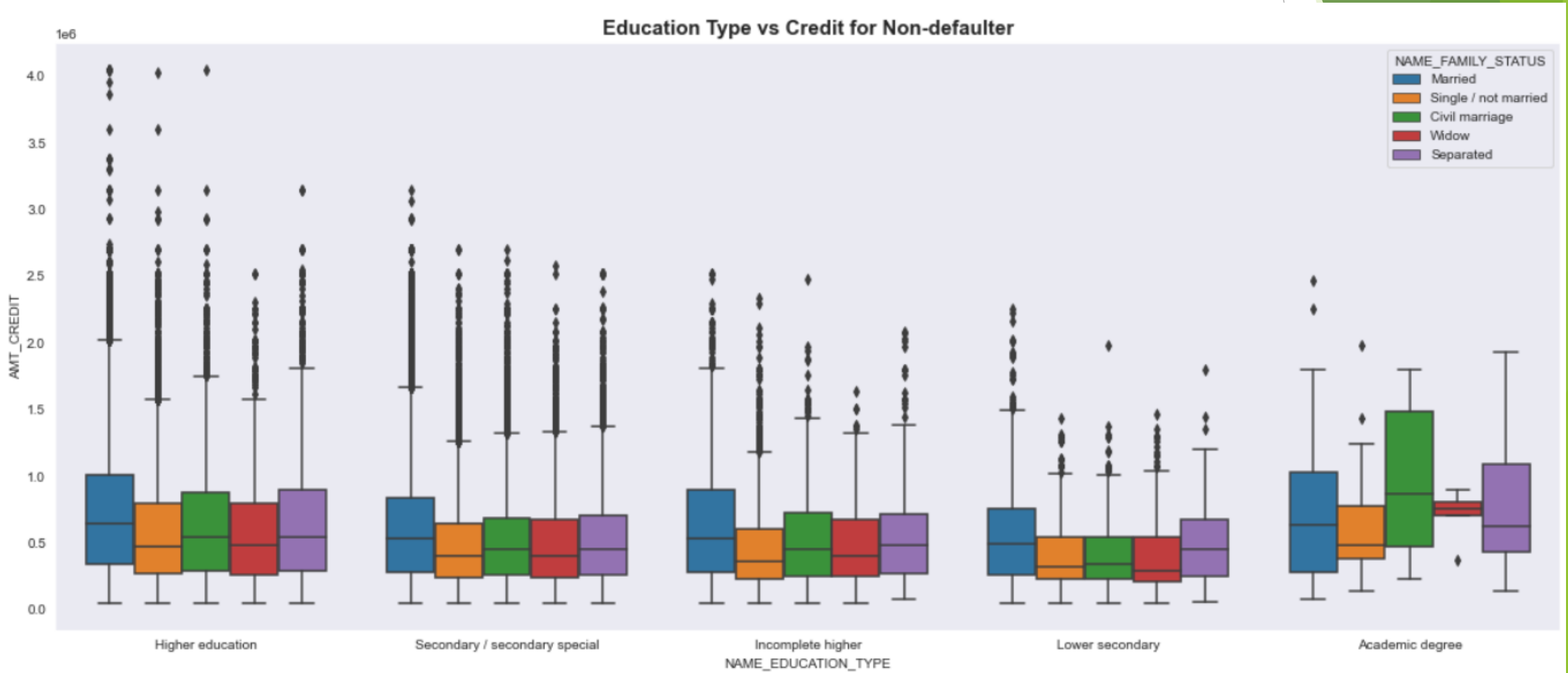
Plotting the education type VS credit of defaulters based on family type

Among defaulters, We have highest credit for applicants with Secondary , Incomplete higher and higher education. The number of applicants and amount of credit is very low for Lower secondary and Academic degrees among defaulters. Most of the defaulters are married across all education groups. Proportionally, Separated groups are having least defaulters with extreme amount of credits.



Plotting the education type VS credit of Non defaulters based on family type

We have maximum extreme credit values for credit in Married group of applicants with Higher education and least are in applicants with Academic degree. Most of the non defaulters with Academic degree have civil marriage. We can also conclude that the highest number of applicants are from higher education and lowest are from Academic degree.

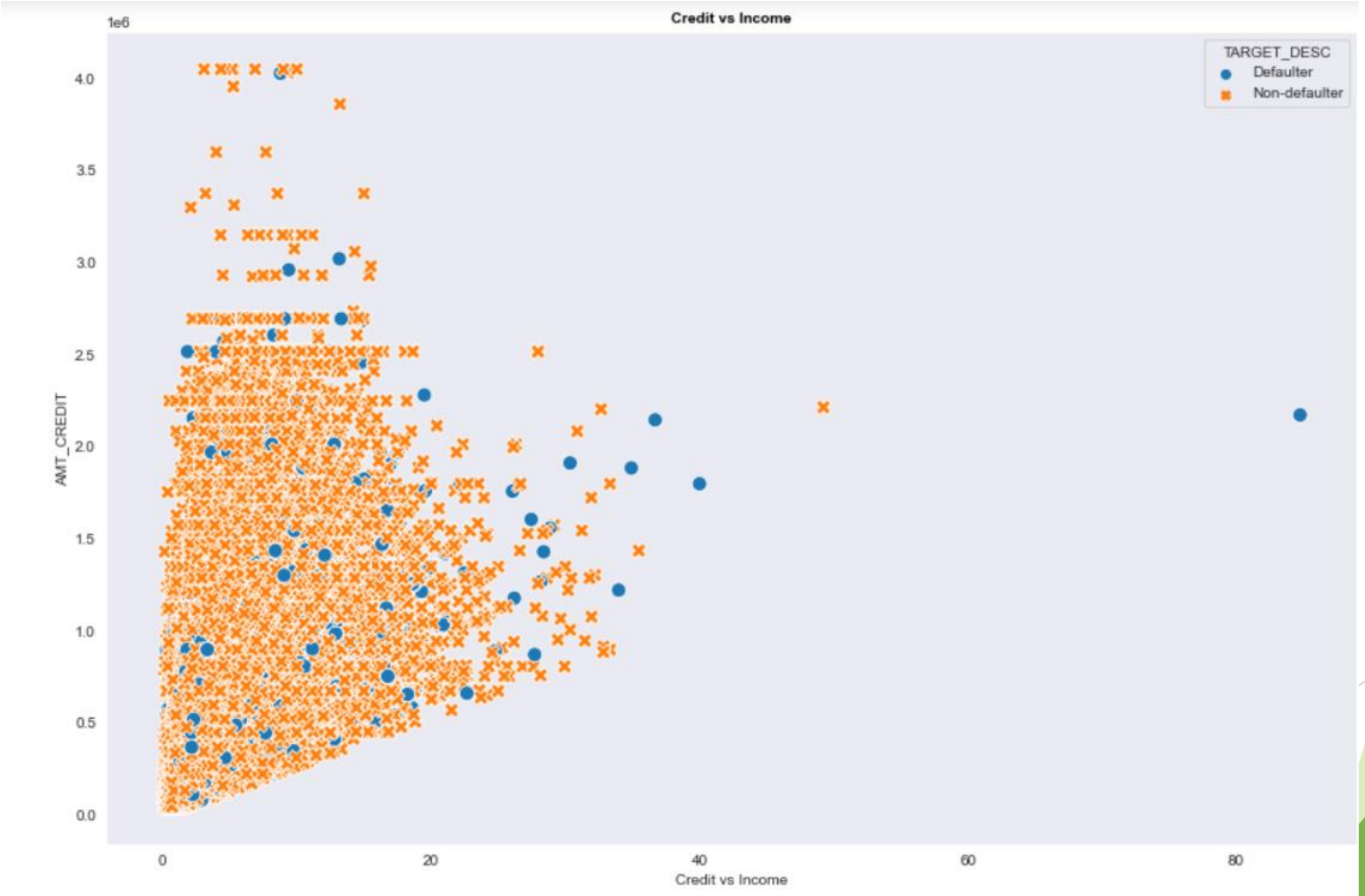


Analyzing and visualizing the Credit to Income Ratio

- We have calculated the Credit to Income Ratio which is one of the key measures for the bank to decide if an application should be refused or accepted. If the Ratio is more than 35% then the chances of defaulting is higher for that applicant.
- We have checked for the applicants who have more annuity than Income and their loans should be actually rejected. However, there are other factors such as spouse income , assets etc. We can see both defaulters and non defaulters in this list.
- Further we have binned the applicants to HIGH, MEDIUM, LOW and VERY HIGH for the calculated Credit vs Income ratio .
- We have then analyzed the Credit to Income Ration against the amount of credit for both defaulters and non defaulter in a scatter plot.

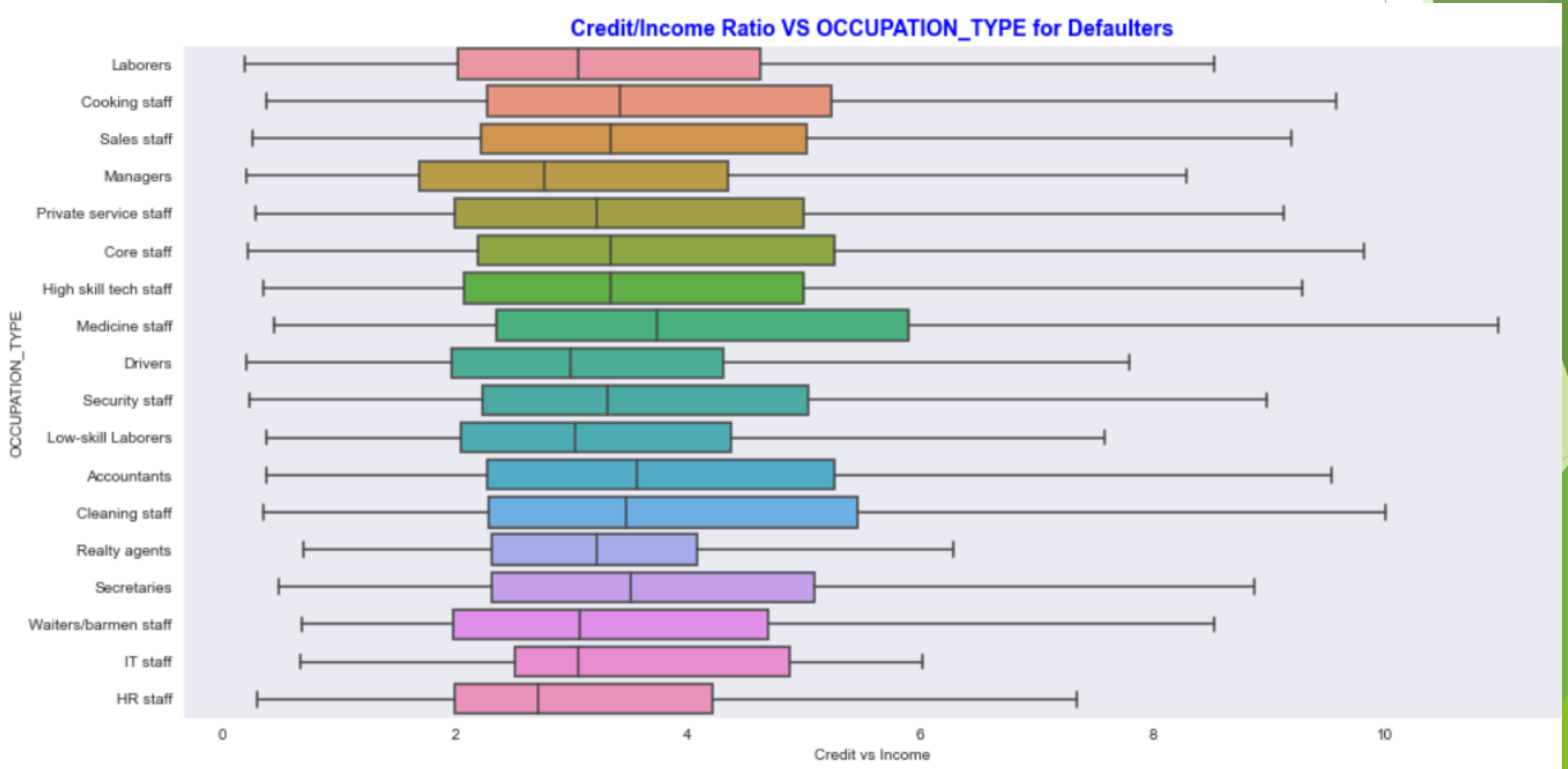
Scatter plot of Credit to Income Ratio Vs Credit

As we can see from the below plot the number of defaulters for high credit amount is low. Also, there are many applicants who have been given loan instead of having Credit to income ration more than 35 %. There are some defaulters as well above the 35/% mark which should be taken into consideration before approving loans. There is one defaulter applicant with very high credit to income ratio and high credit amount.



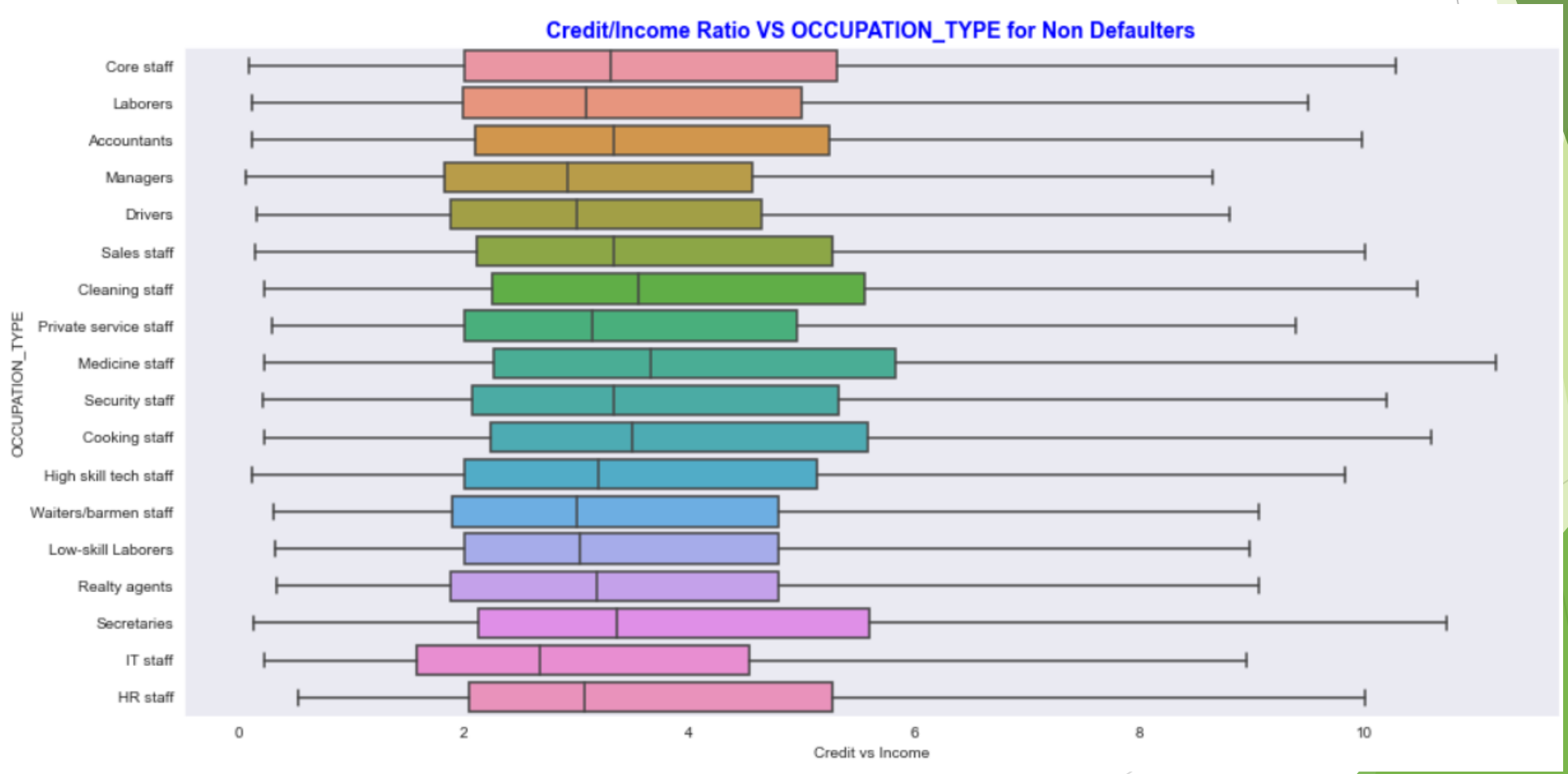
Credit/Income Ratio VS OCCUPATION_TYPE for Defaulters

In Defaulters, the Credit to Income Ratio for Medicine Staff is the highest followed by Cleaning staff, cooking staff, laborers etc. If we observe, the maximum number of defaulters across all occupations have credit to income ration are spread over more than 35% mark. Least number of defaulters are from IT staff, HR staff etc.



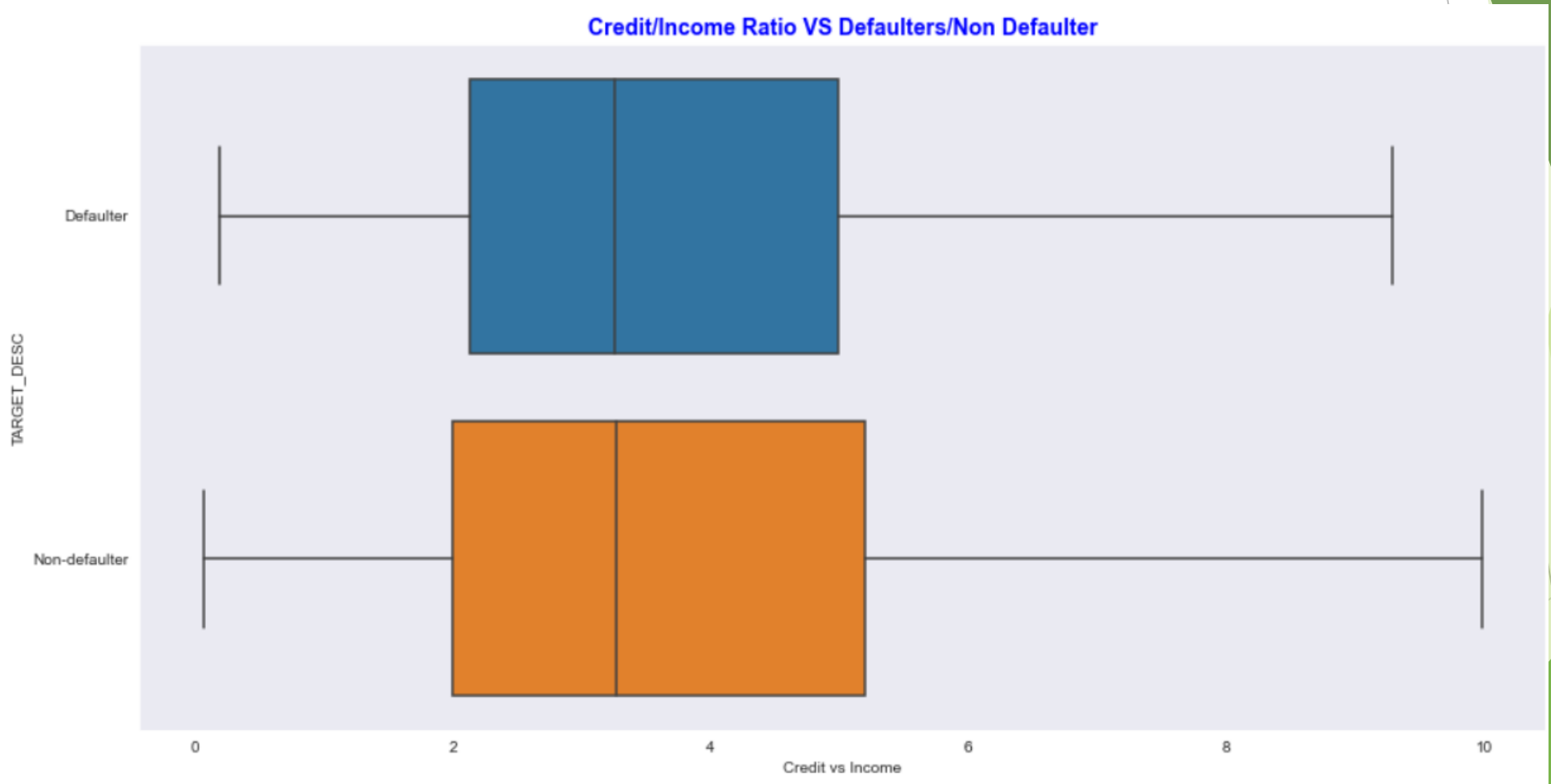
Credit/Income Ratio VS OCCUPATION_TYPE for Non Defaulters

The credit to Income ratio is almost comparable across all the occupation types except some jobs such as HR staff, cooking staff, Medicine staff, accountants , core staffs etc. This might be due to either very high Credit or very high income of applicants in these jobs. However, they are non defaulters as Credit to Income ratio is not the only parameter which determines the repayment of loan as mentioned previously.



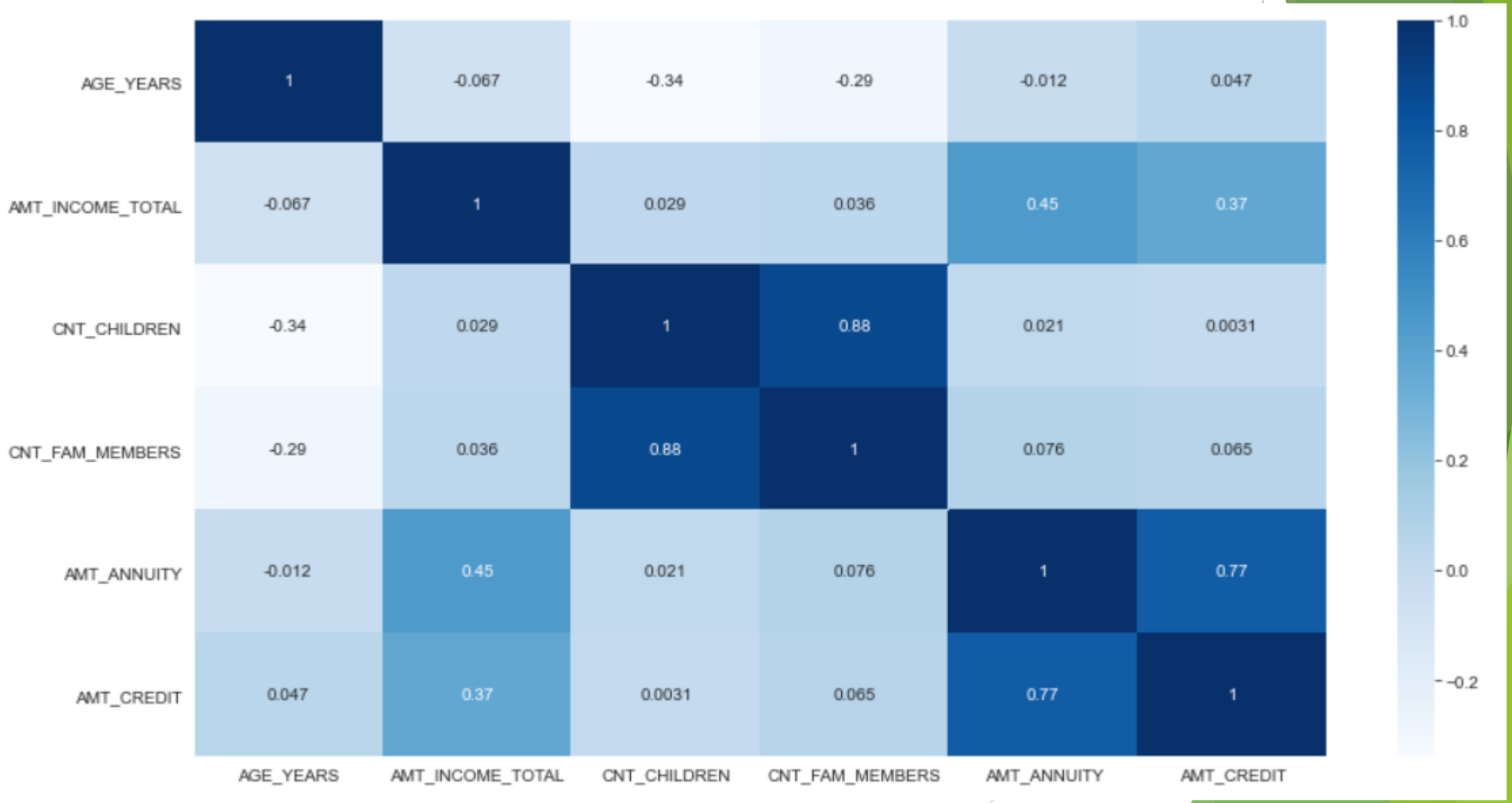
Credit/Income Ratio VS Defaulters/Non Defaulter

Credit to income ratio for Non Defaulters is high but ideally there are other parameters which are considered to conclude if an applicant will default.

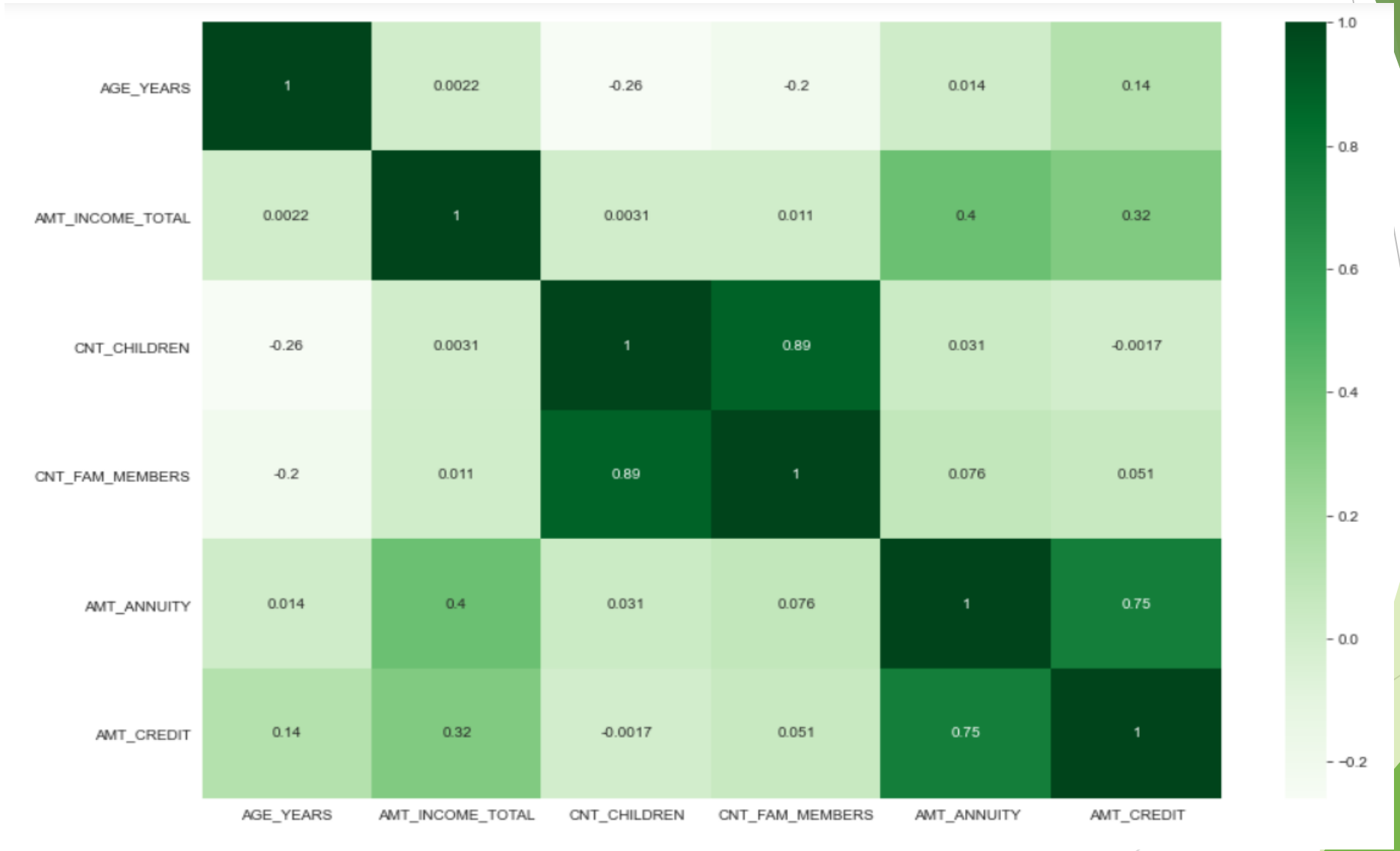


The next step is to find the correlation between different numerical variables of both the defaulters and non defaulters. The correlations are plotted in a heat map visualize and draw conclusions.

Heat map of correlation for numerical columns of Non Defaulters



Heat map of correlation for numerical columns of Defaulters



Conclusions on correlations from the heat maps

For Non defaulters :

- Count of children with the count of family members has highest correlation. There are also considerable amount of applications from the group of people having more children if we look into the previous graphs.
- Amount of Credit and Loan annuity has a very high correlation. Since the credit is higher , the annuity to be paid every year will be high.
- Income has high correlation with amount of credit as well as Annuity

For Defaulters :

- Count of children with the count of family members has highest correlation.
- Amount of Credit and Loan annuity has a very high correlation. Since the credit is higher , the annuity to be paid every year will be high.
- In case of defaulters, Income has high correlation with amount of credit as well as Annuity

Conclusion : From the above observations, we can conclude that the most of the variables are having same correlation for both defaulters and non defaulters.

Analyzing the Previous Application and Application data

At this stage, we have merged the previous application file with the application data to have the information of applicants who had previously applied for loans for further analysis.

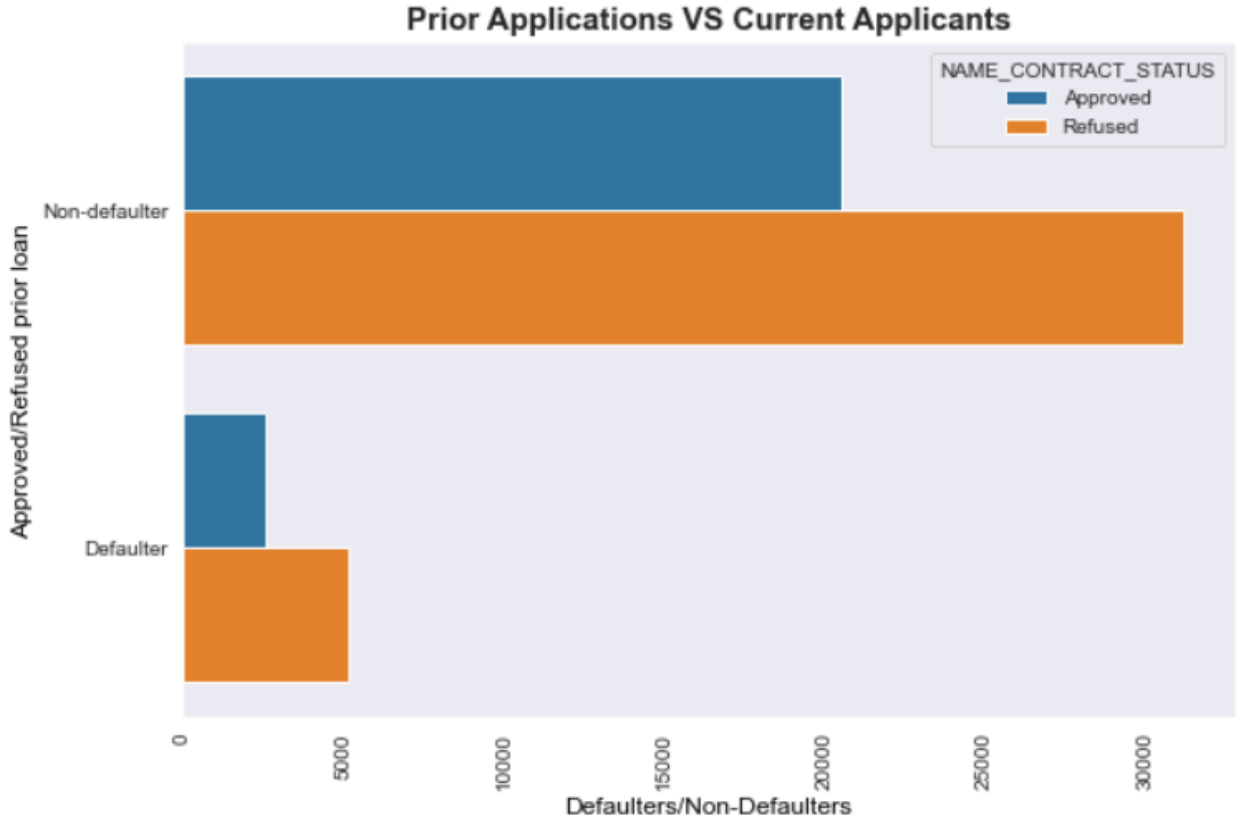
We perform similar data quality and cleaning steps as for Application data followed by univariate and bivariate analysis.

Our objective is also to investigate if there are applicants who were previously rejected for loans have been approved now in current application and vice versa.

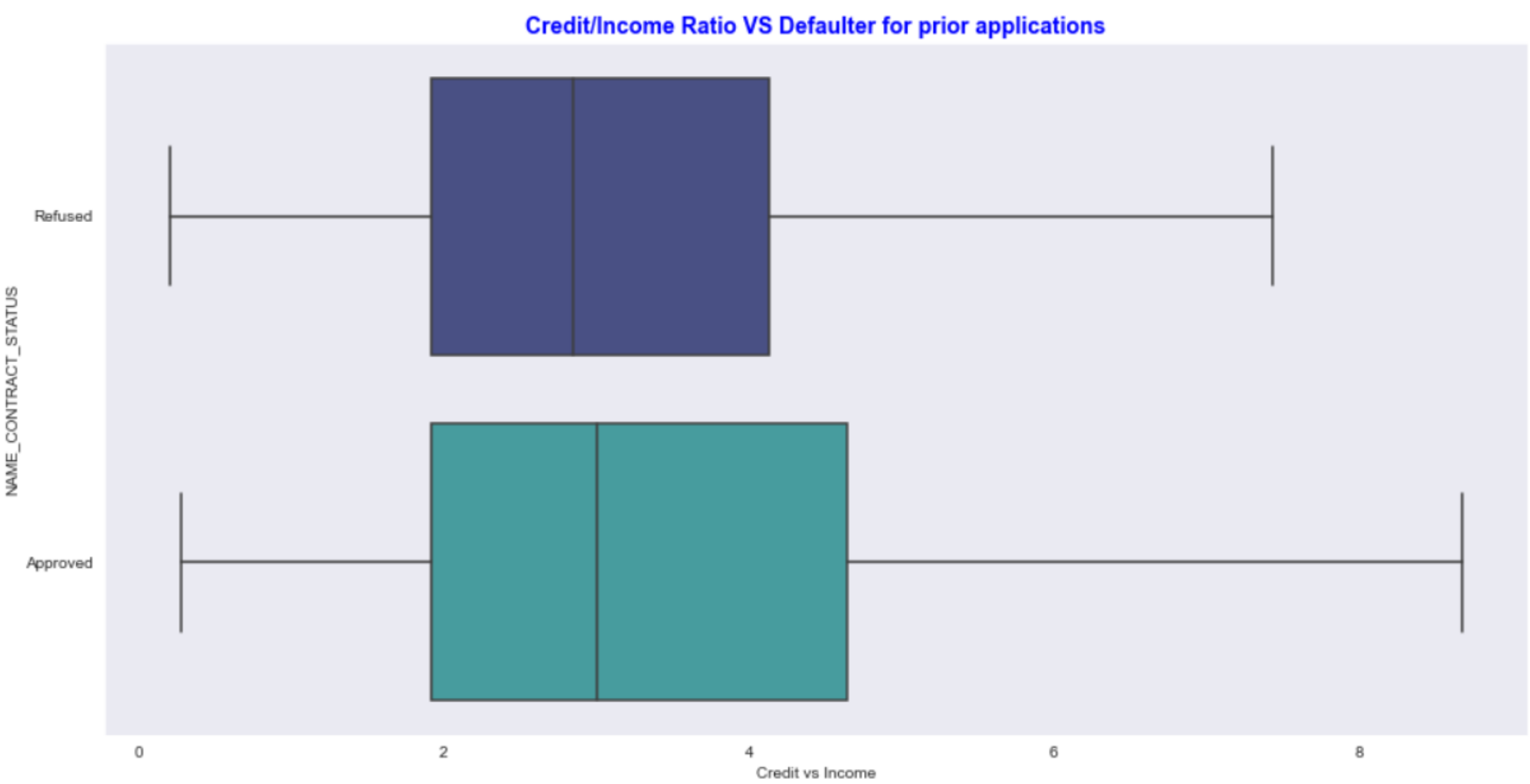
Also, we have done analysis based on the credit/income ratio for both previous application and current application on both defaulters and non defaulters.

Plotting the status of previous application for current defaulters and non defaulters

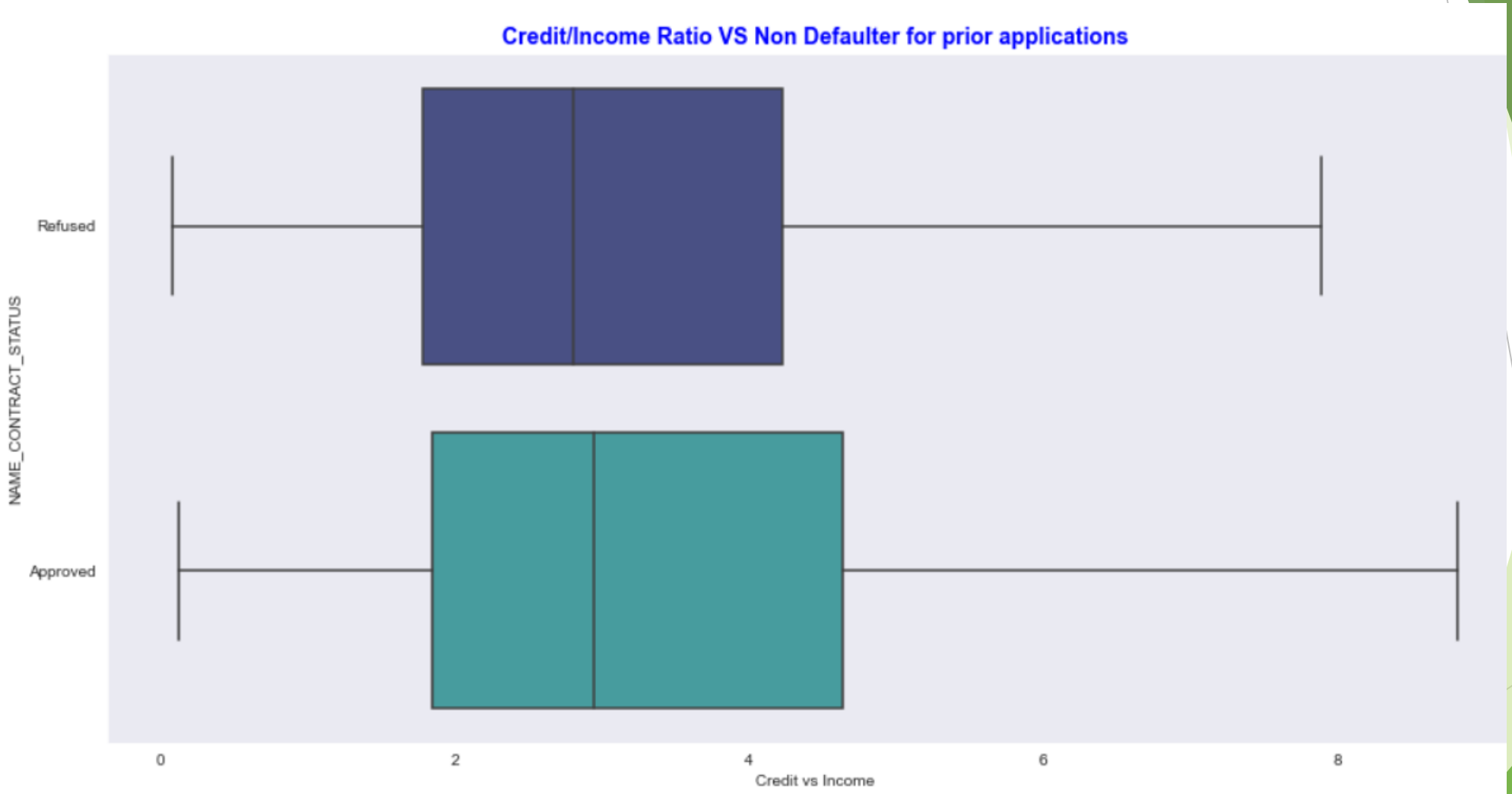
We can observe that most of the applications that were refused last time are non-defaulters in their current application. Further analysis is need for the defaulter with approved vs refused on other categories like income, credit, age, education and occupation.



Credit/Income Ratio VS Defaulter for prior applications



Credit/Income Ratio VS Non Defaulter for prior applications



Comparing Credit to Income ratio with Target column

C vs I Buckets	TARGET_DESC	Applicants Count
Very Good CTI	Non-defaulter	275947
Very Good CTI	Defaulter	24324
Good CTI	Non-defaulter	6571
Good CTI	Defaulter	480
Average CTI	Non-defaulter	167
Average CTI	Defaulter	18
Poor CTI	Defaulter	2
Very Poor CTI	Defaulter	1
Very Poor CTI	Non-defaulter	1
Total applicants		307511

Note: applicants_new['C vs I Buckets']=applicants_new["Credit vs Income"].apply(lambda x: 'Very Good CTI' if x<=10 else ('Good CTI' if x>10 and x<=20 else
('Average CTI' if x>20 and x<=35 else
('Poor CTI' if x>35 and x<=40
else 'Very Poor CTI'))))

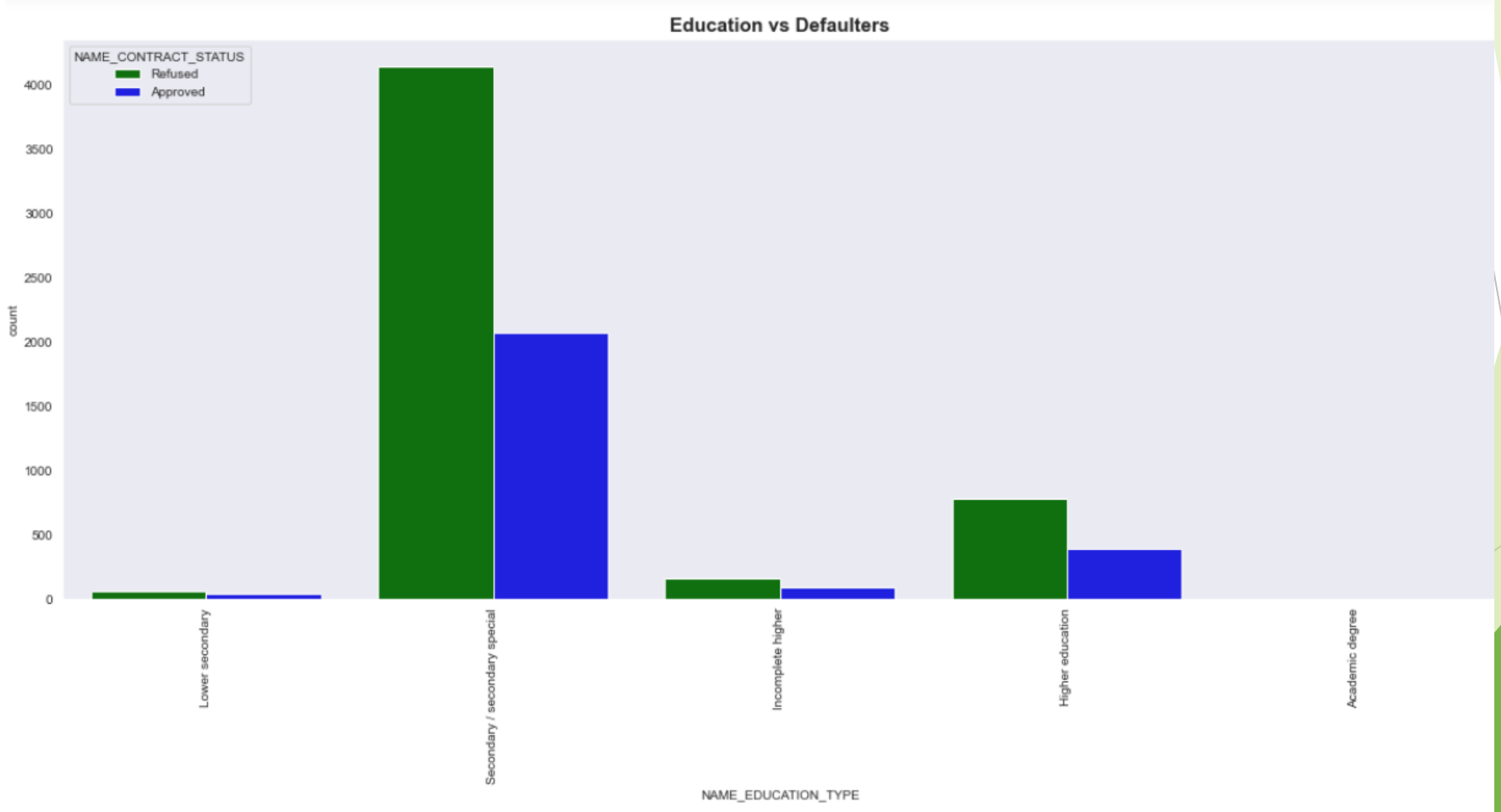
From the above two slides we can conclude:

- Ignoring outliers, the defaulters whose loans were approved have higher credit to income ratio
- Ignoring outliers, the refused applications have moderate credit to income ratio

Since these results are contradicting, we have to investigate the Refused vs approved with other category information

Graph of previously refused/approved loans vs Defaulters of current application across Education types

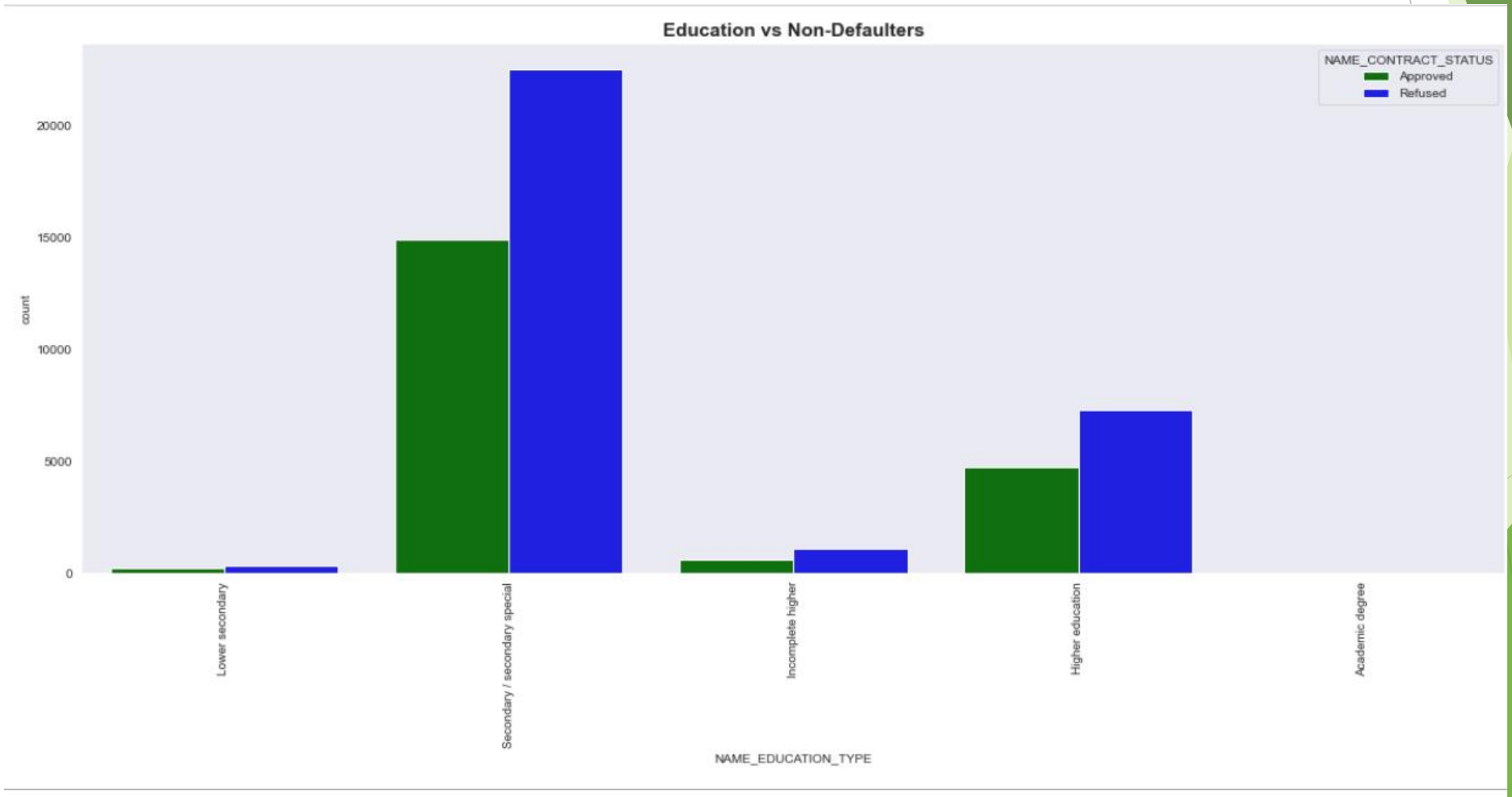
The highest number of defaulters were also previously refused from secondary education type the highest. This matches with our observation from Application data. Most applicants from higher education and higher education who were previously approved have defaulted now.



Graph of previously refused/approved loans vs Non Defaulters of current application across Education types

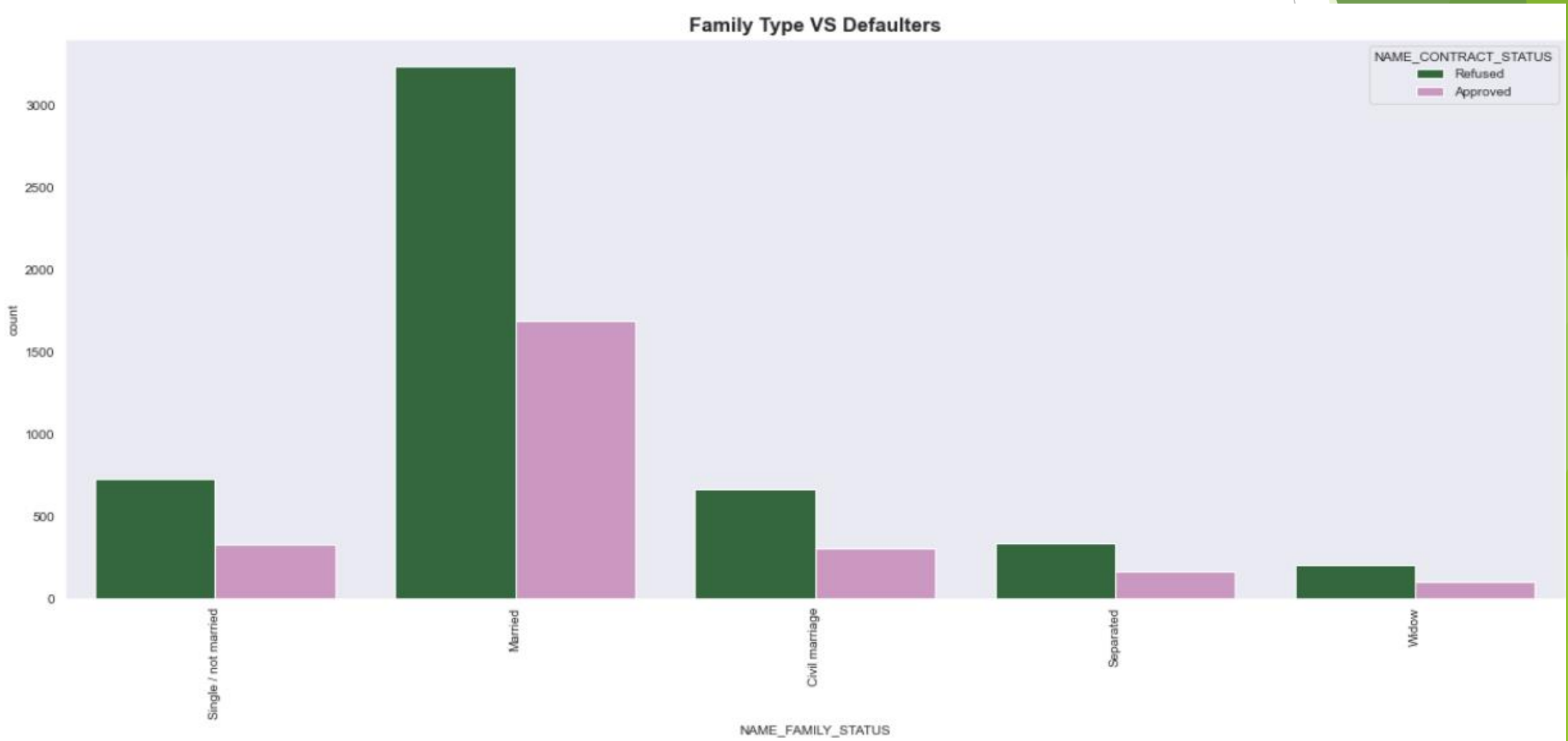
Secondary education type has the highest applicants whose loans were previously approved as well as are Non defaulters now followed by Higher education. This matches with our observation from applicants data.

We can also conclude that most of the applications are from applicants with Secondary , higher education .



Graph between Family type and Defaulters for Refused/Approved loans

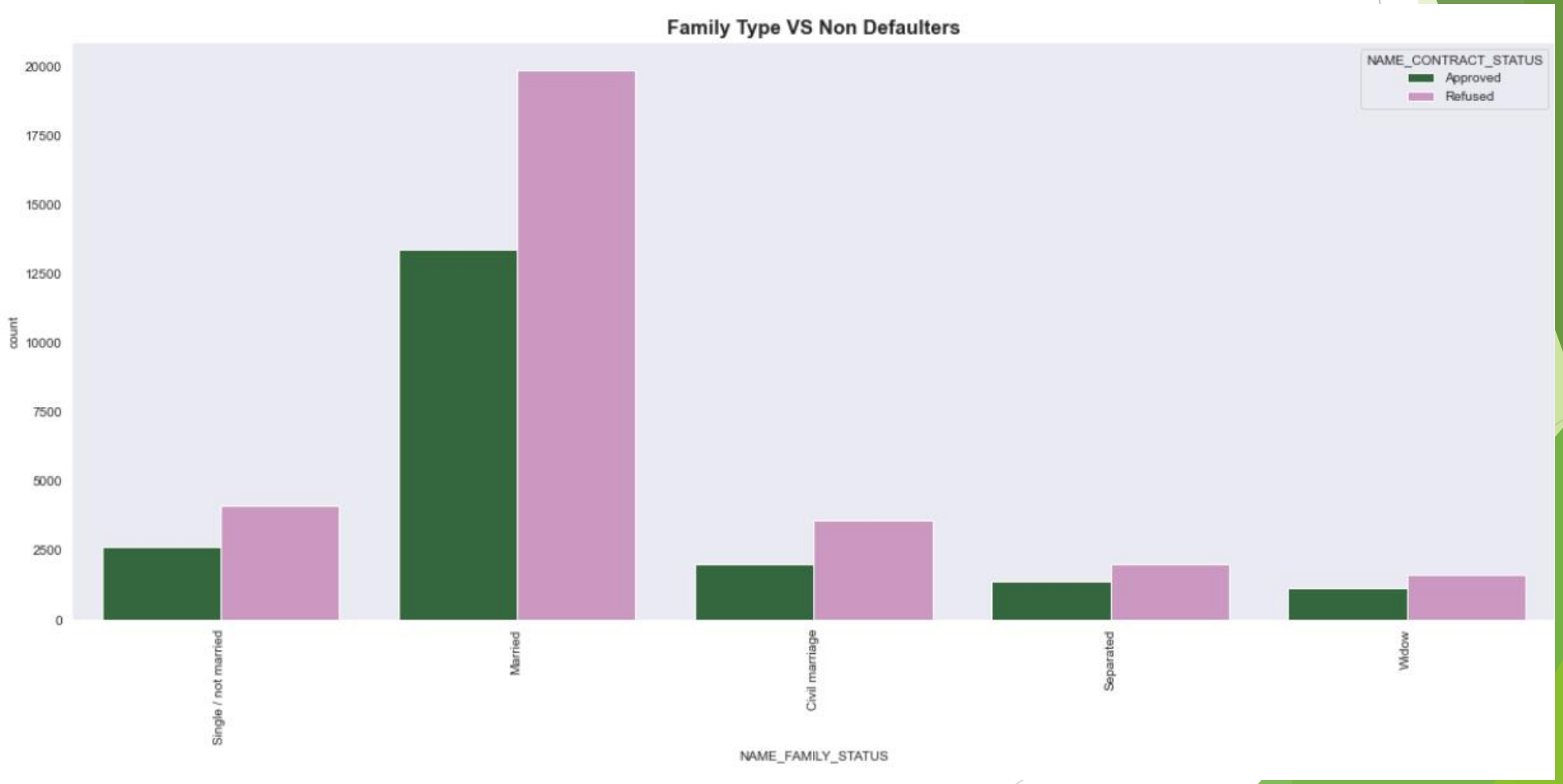
Married applicants have the highest refused applications as compared to other groups among defaulters. Widows have least loan applications. Single and Civil Marriage applicants have almost similar refusal and approval for defaulters.



Graph between Family type and Non Defaulters for Refused/Approved loans

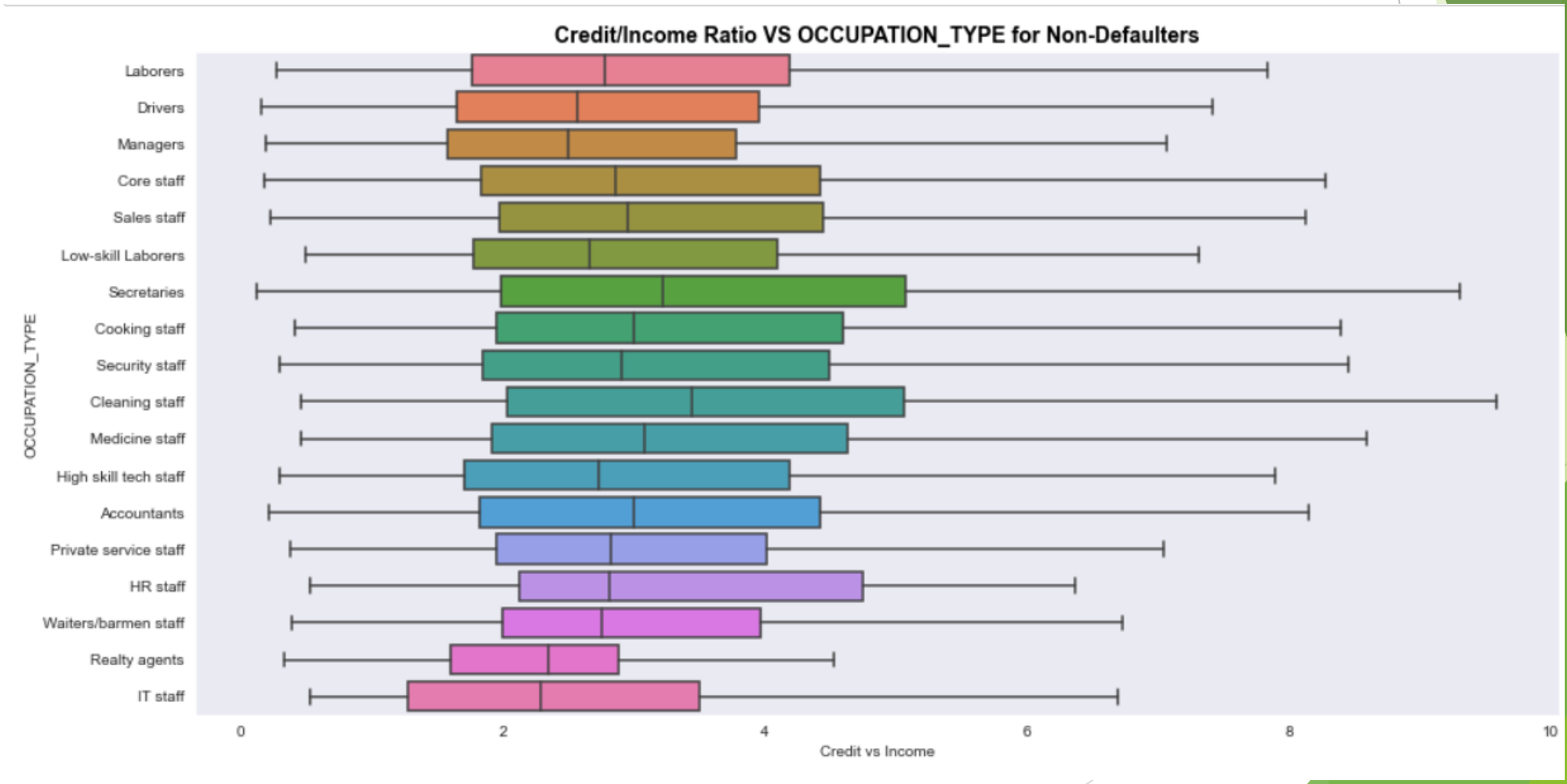
Most of the applicants from married group which were previously refused are now non defaulters. The refusal and approval of current non defaulters from Separated and Widow group are almost similar.

We can conclude that the married groups have more requirement of loans and have to be checked on other parameters for approval of loans.



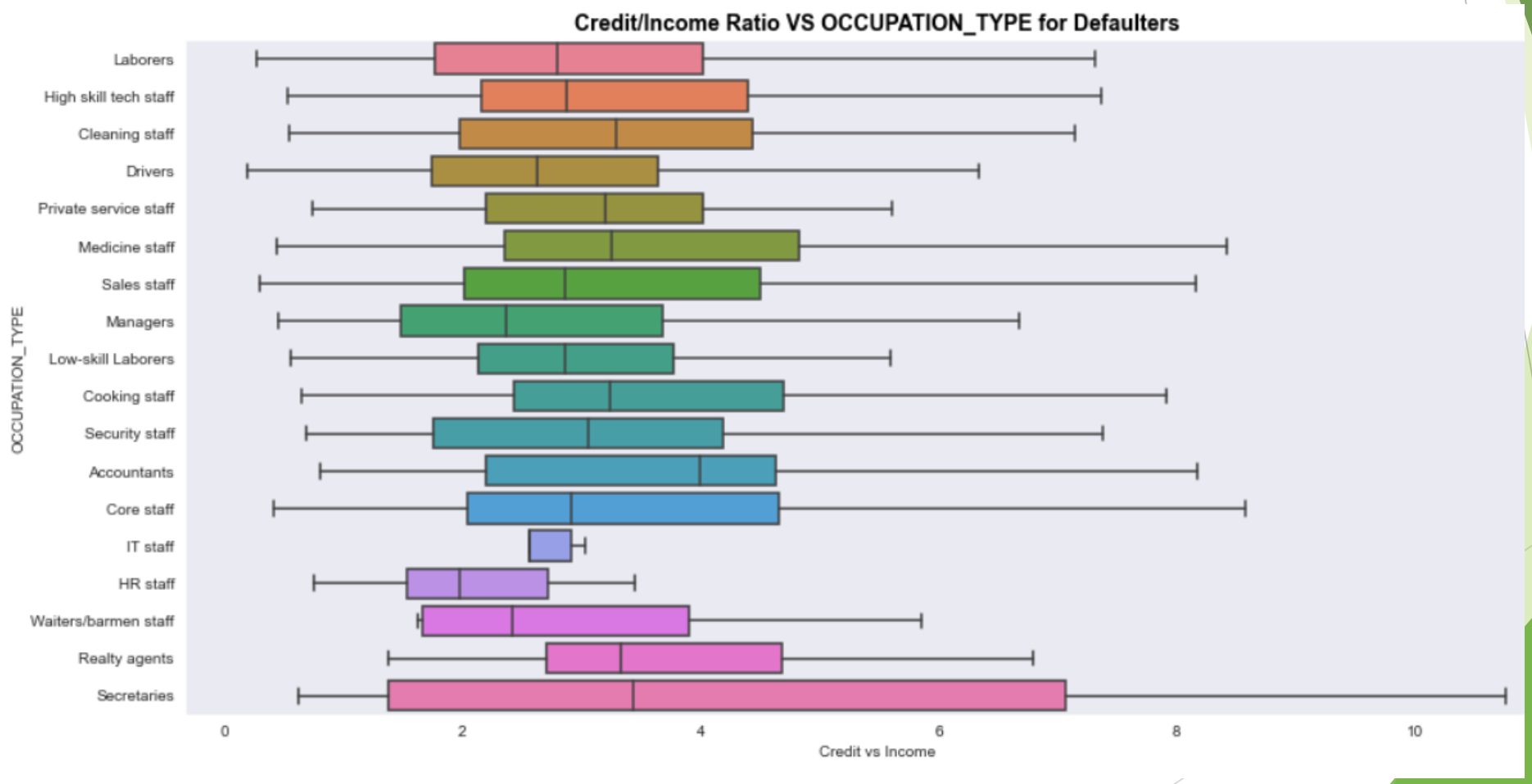
Credit/Income Ratio VS OCCUPATION_TYPE for Non-Defaulters

The previous Credit/Income ratio of most of the current non defaulters is comparable except some groups such as laborers, secretaries, cleaning staffs which have high values , more number of applicants in these groups and higher credit can be one of the reasons for this.



Credit/Income Ratio VS OCCUPATION_TYPE for Non-Defaulters

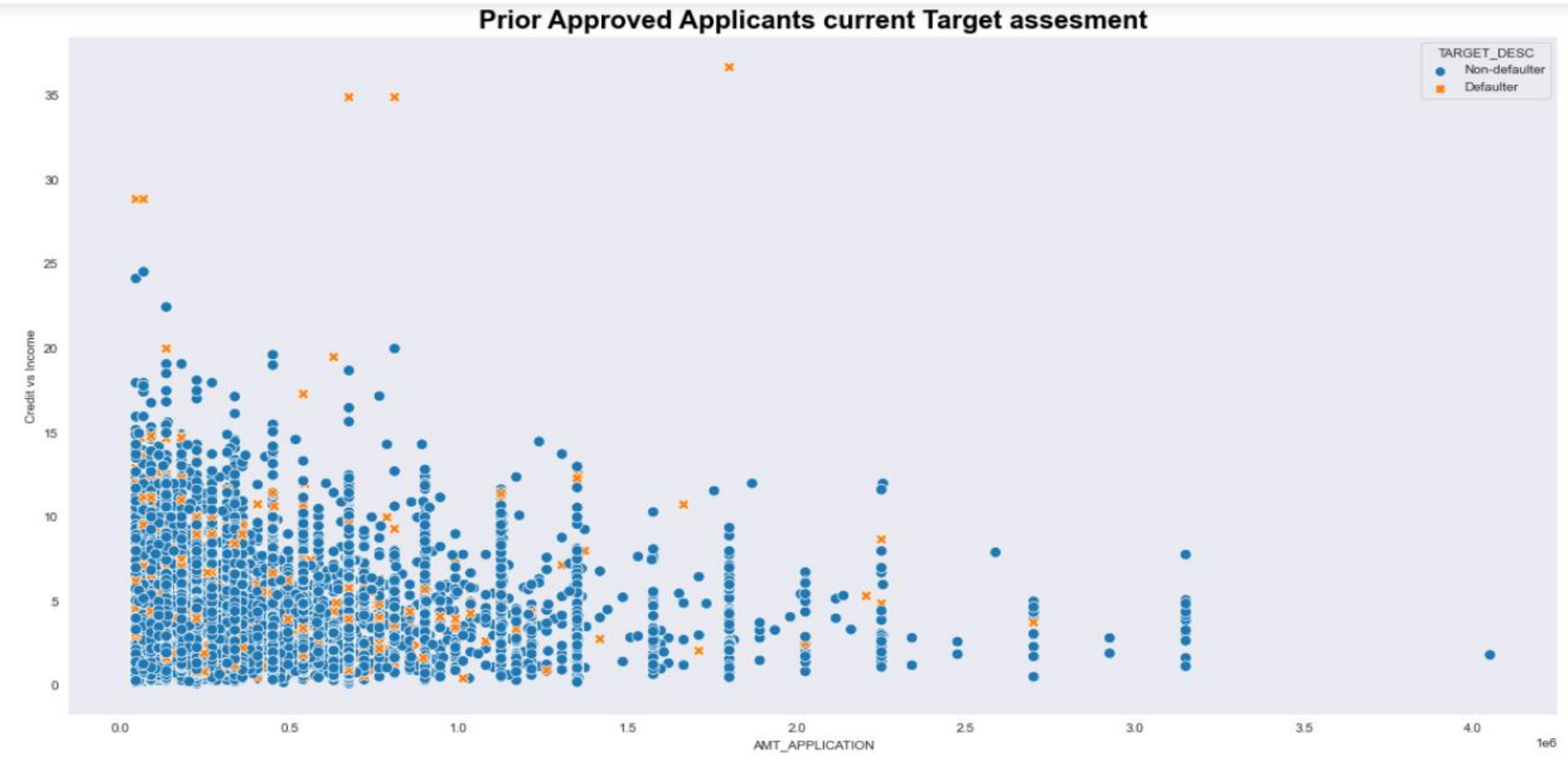
We have very less previous Credit/Income ratio yet more current defaulters for Private service staffs, Core staff, Secretaries, Medicine staffs etc. These groups must be checked for other parameters before approving loans. IT Staff and HR staff have least Credit/income ratio and number of defaulters.



To find the patterns between current defaulters/non defaulters and previously approved/refused loans, we plot a series of scatter plots based on previous credit/income ratio and previous application amount which are explained in the following slides.

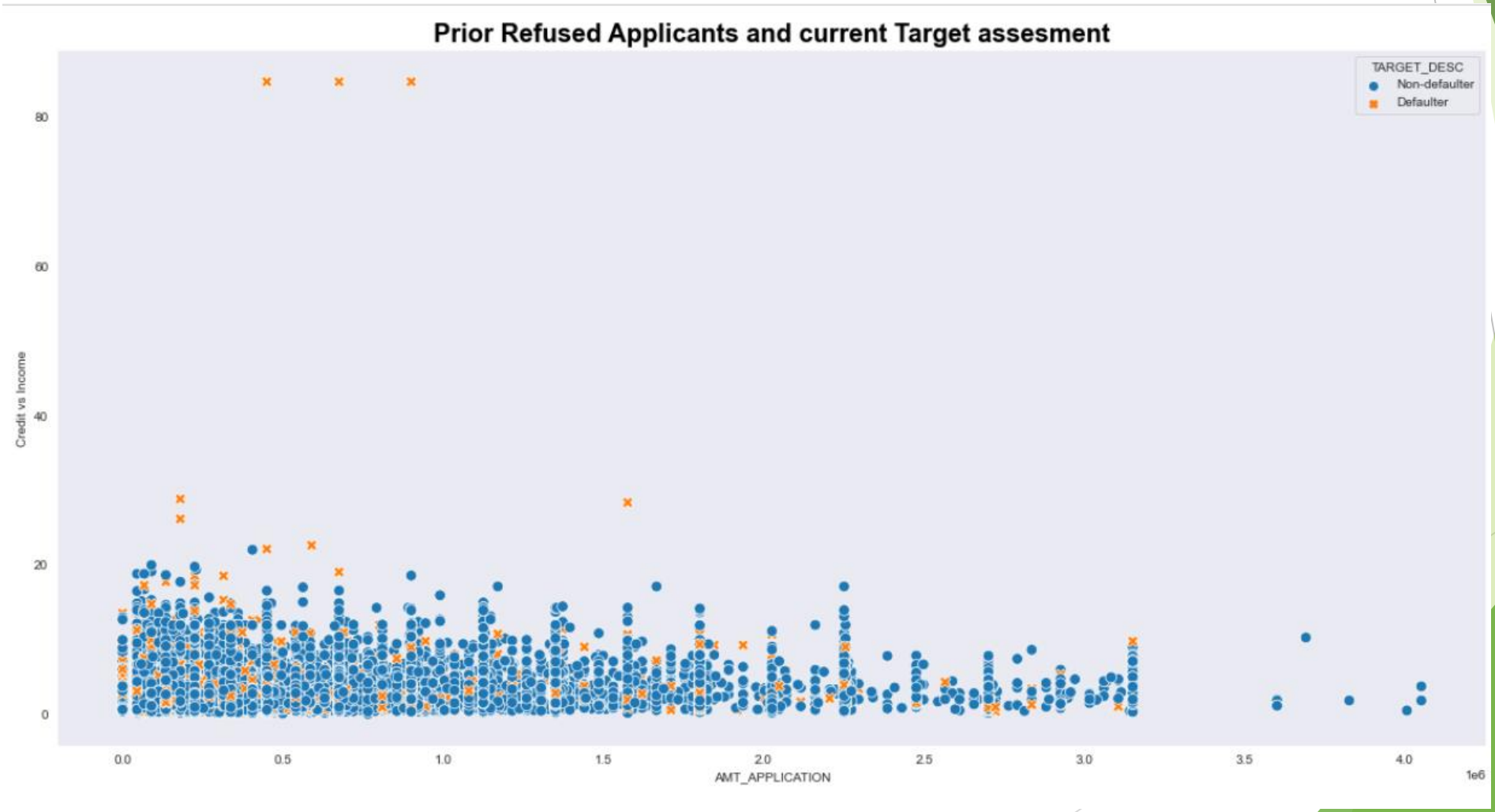
Scatter plot of the previously Approved applicants considering their previous Credit/income ratio and previous application amount

In this plot the applicants are classified as per current defaulters and non defaulters. We can observe that most of the non defaulters are having credit/income ratio less than 30% . We have negligible non defaulters for high application amounts. All the applicants with credit/Income ratio close to 30% and above are currently defaulters.



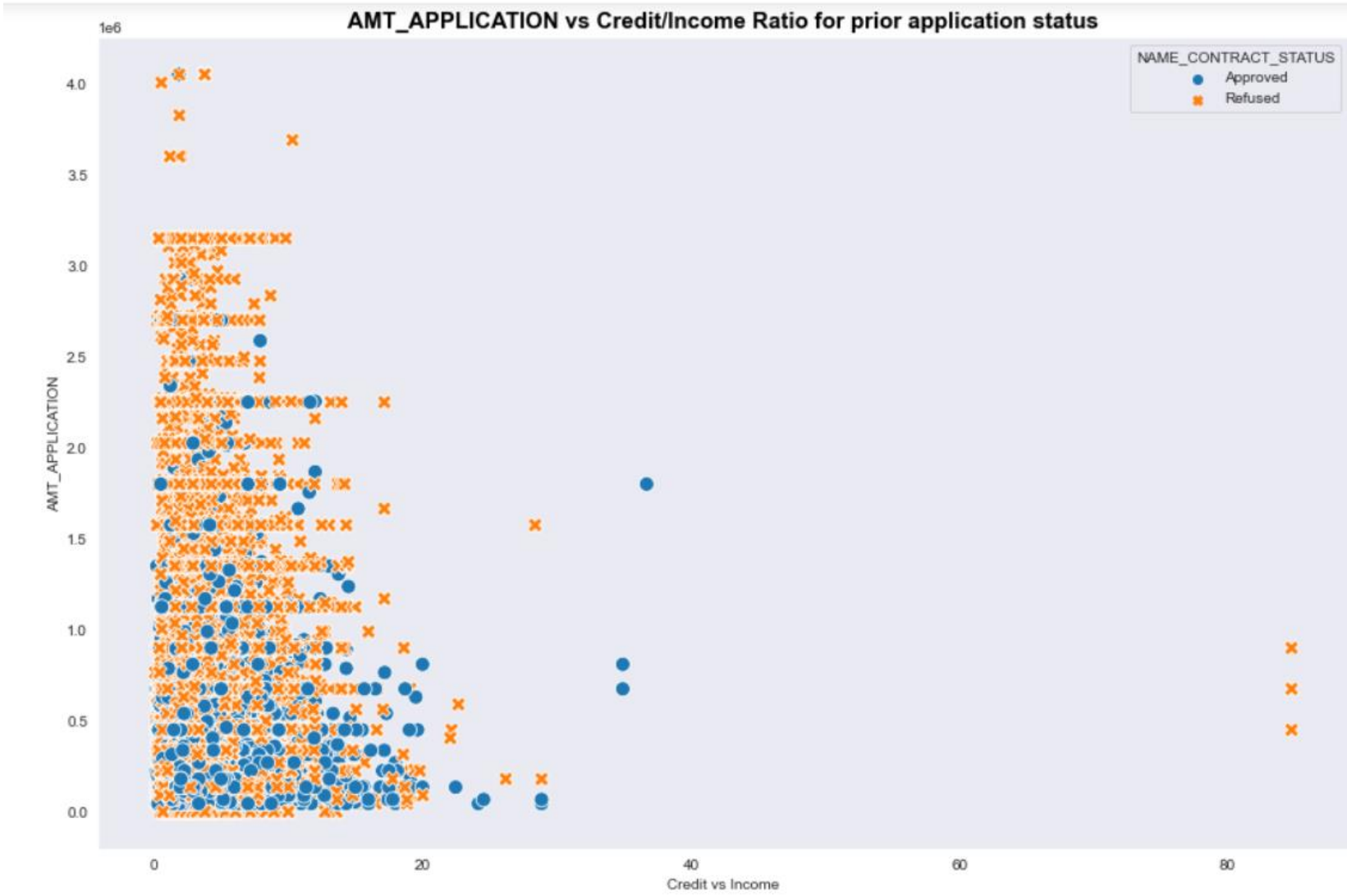
Scatter plot of the previously Refused applicants considering their previous Credit/income ratio and previous application amount

In this plot the applicants are classified as per current defaulters and non defaulters. We can observe that all the previously refused applicants having credit/income ratio less than 30% are non defaulters in their current application. All the applicants with credit/Income ratio 30% and above are currently defaulters.



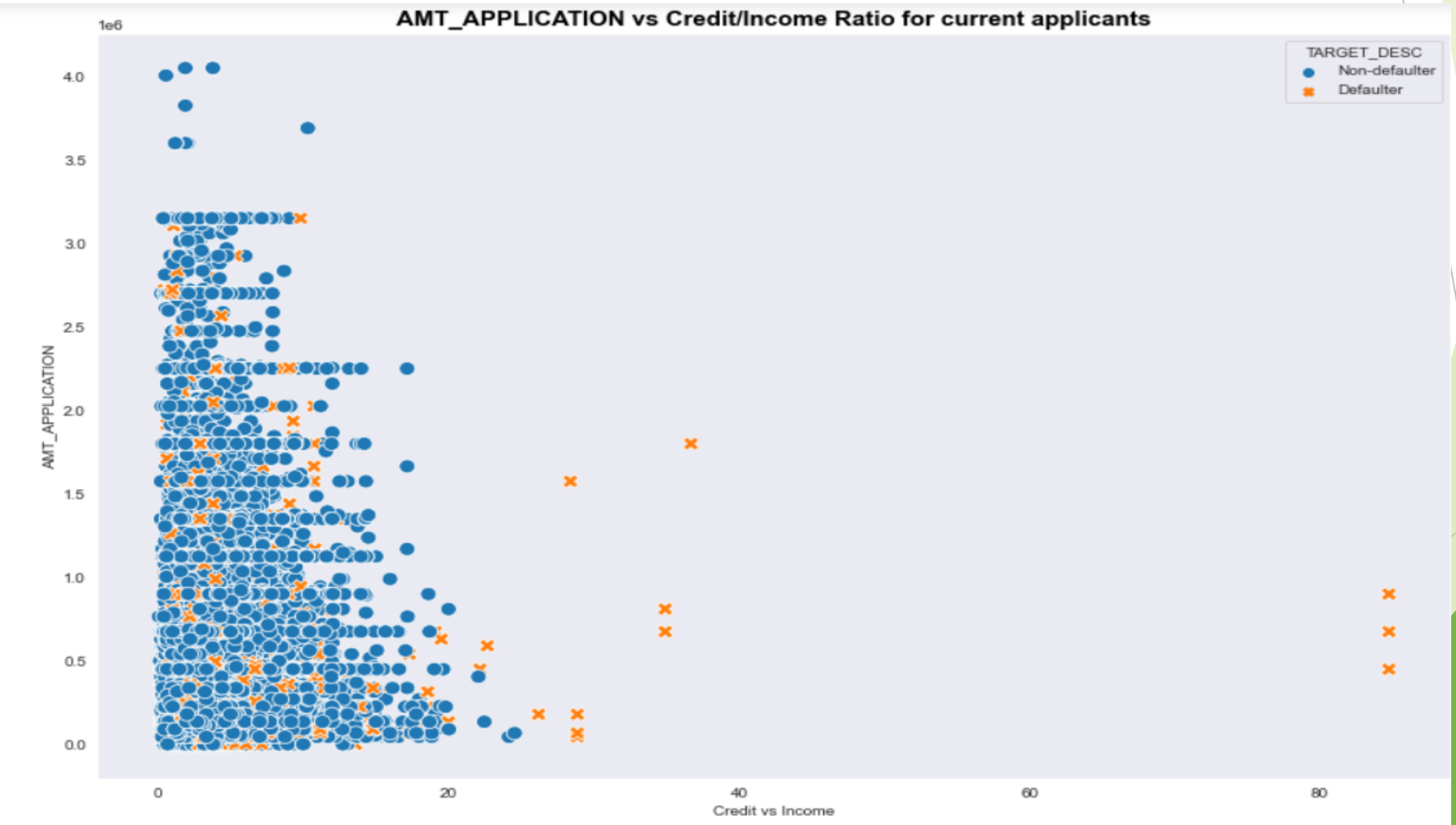
Plotting the previous application amount against credit/income ratio

In this plot , we have classified the scatter plot based on the result if the applications are refused or approved. We can observe that higher proportion of the loans were refused when the amount application was high.



Comparing the previous graph with current application by previous application amount against credit/income ratio

To compare the previous graph with the current application , we have now plotted the previous amount application against credit/income ratio and classified the scatter plot based on the result if the applicants are now defaulters or non defaulters. We can observe that most of the applicants are now non defaulters who were previously refused for loans. However, this does not give a conclusion since in the current application, these applicants can apply for a different amount of credit after getting refused in previous applications which gives them better option of repaying their loan.



Conclusion

Based on our analysis of both applications data and previous applications, we can conclude the below points. However, the data is subject to further analysis.

- The proportion of defaulters is less as compared to non defaulters.
- Most of the applicants are females and have the highest defaulters and non defaulters as well. It can also be concluded that banks approve more loans of females.
- Applicants with high income default less.
- Married applicants tend to apply more for loans and also have highest number of defaulters and non defaulters. This group can default based on other parameters.
- Applicants in age groups 31- 50 with high income and high education tend to default less.
- Revolving loans are safer than cash loans.
- If applicants have higher income and higher credit, then they tend to default less.
- Credit to income ratio plays a crucial role in determining if applicants will default. It should be less than 35%.
- There can be applicants who have more annuity than Income and their loans should be actually rejected. However, there are other factors such as spouse income , assets etc. that are taken into consideration for approving loans.

THANK YOU