upGrad

EDA Case Study

Bank Loan Risk Analysis

By - Sheza Waqar Beg

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 Analysis is done in Python Jupyter Notebook

Bank Loan Risk Analysis - Project (upGrad)

Data Cleaning

Handling missing values

Found out rows & columns with more than 60 null values

Imputing data with 13% of missing value

Used mode imputation technique

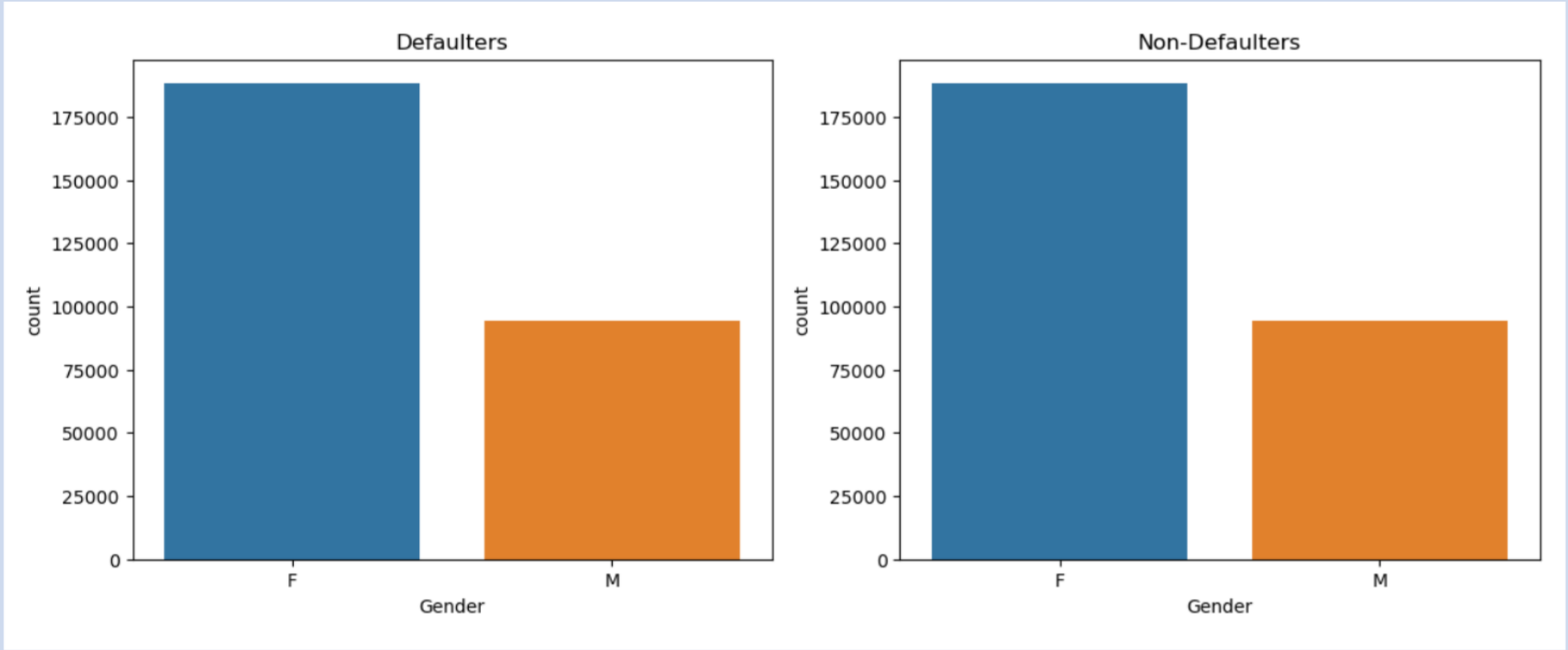
Check imbalance

Find outliers - technique used - box plot

Top factors that correlate

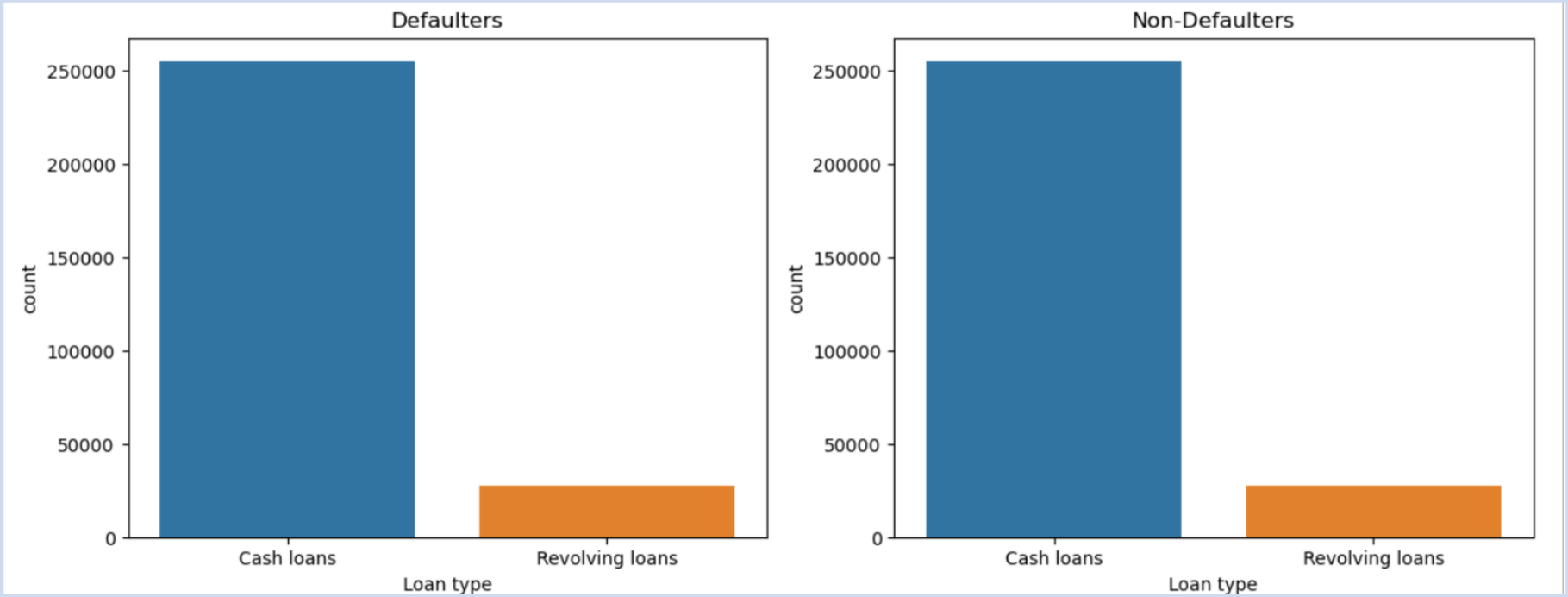
Factors like Age, AMT credited, Income, Gender that affects clients with difficulty during repayment

Univariate Analysis - Analysis of defaulters based **Gender type**



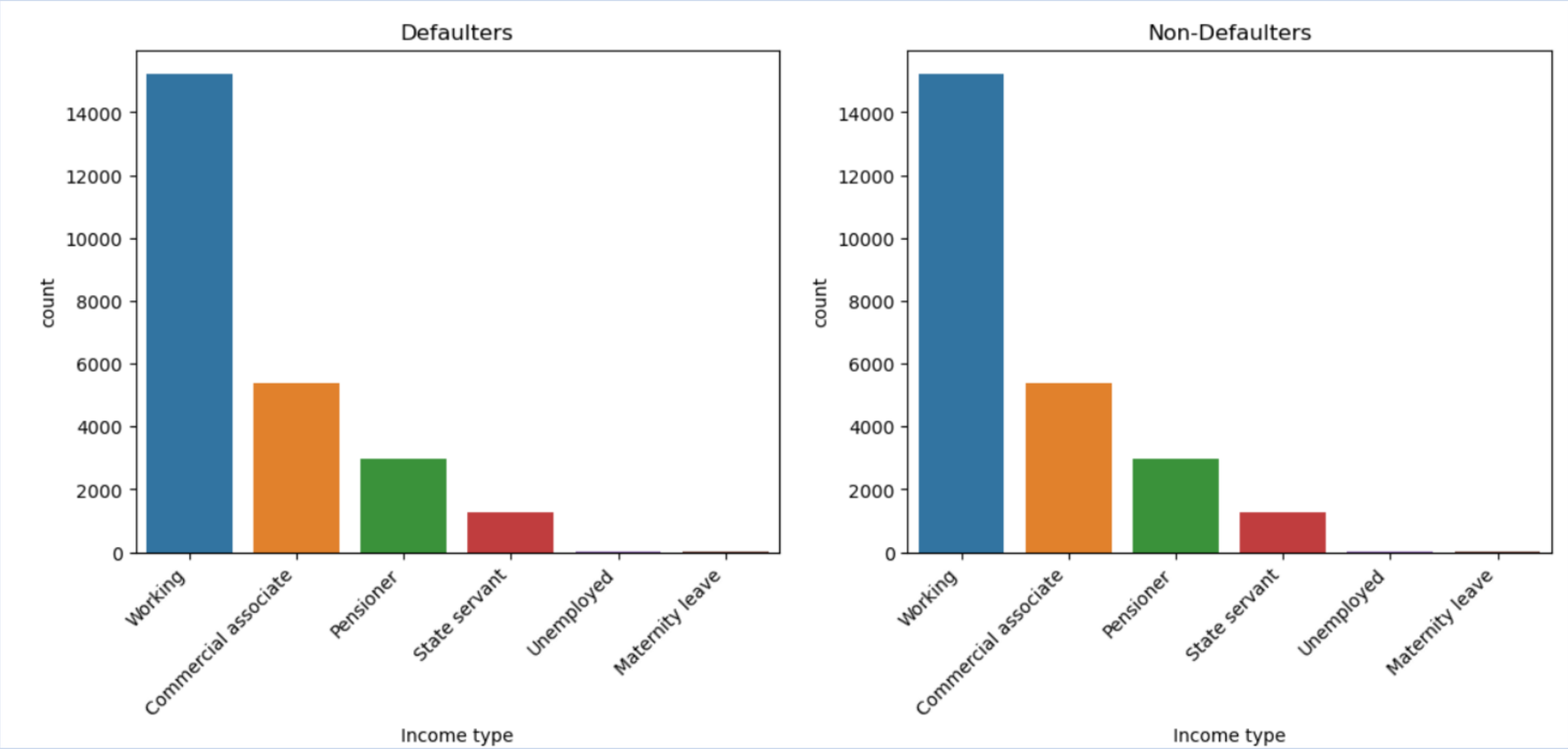
Defaulters - We can see that females default more in number than males.
Non-defaulters - The same pattern continues for non-defaulters as well.

Univariate Analysis - Analysis for defaulters basis **Loan type**



Revolving loans are very less in number compared to Cash loans in both the cases

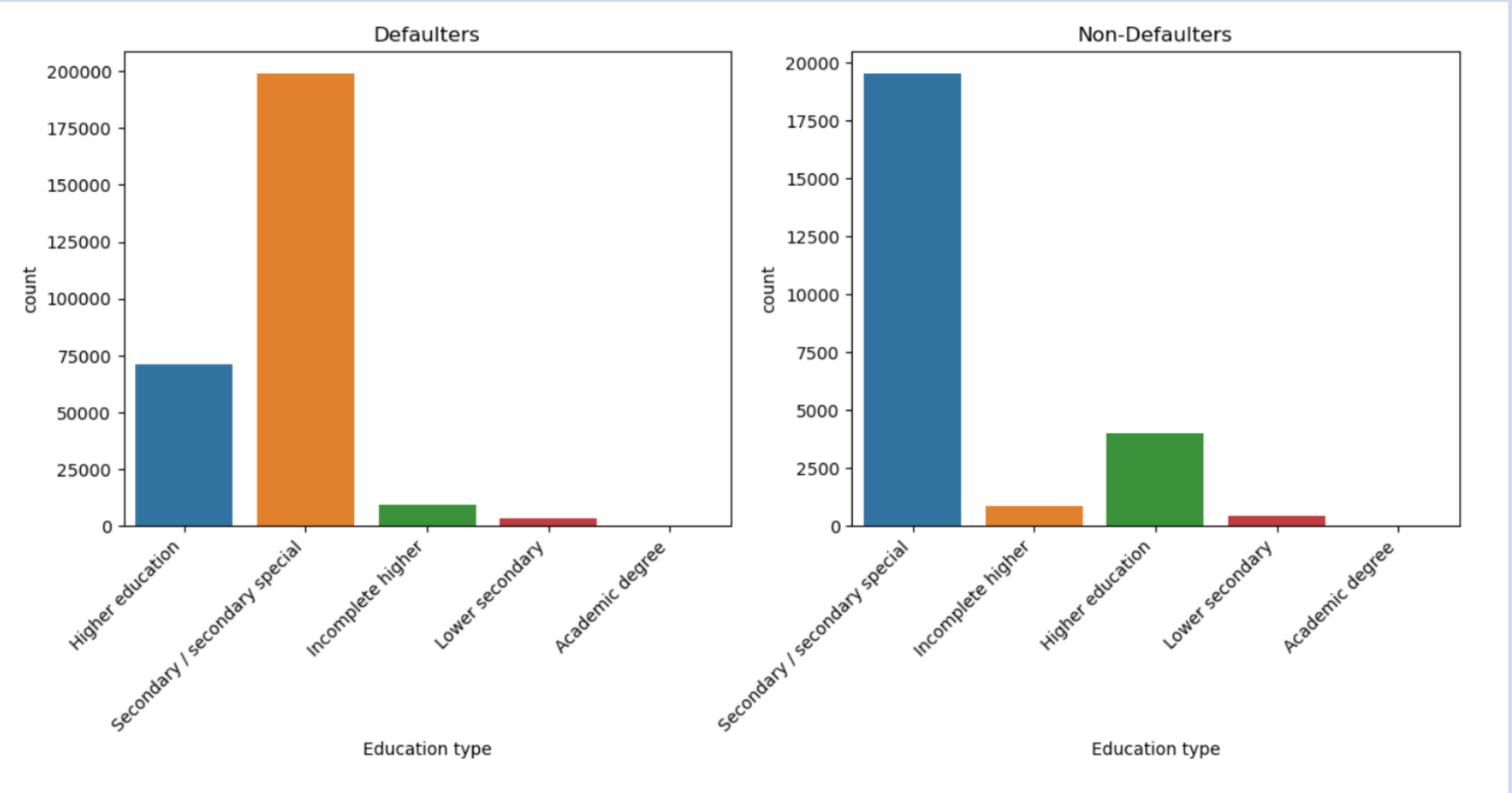
Univariate Analysis - Analysis of defaulters based **Income type**



Defaulters - We can see that working, followed by Commercial associate followed by Pensioner, followed by State servant default more

Non-defaulters - The same pattern continues for non-defaulters as well.

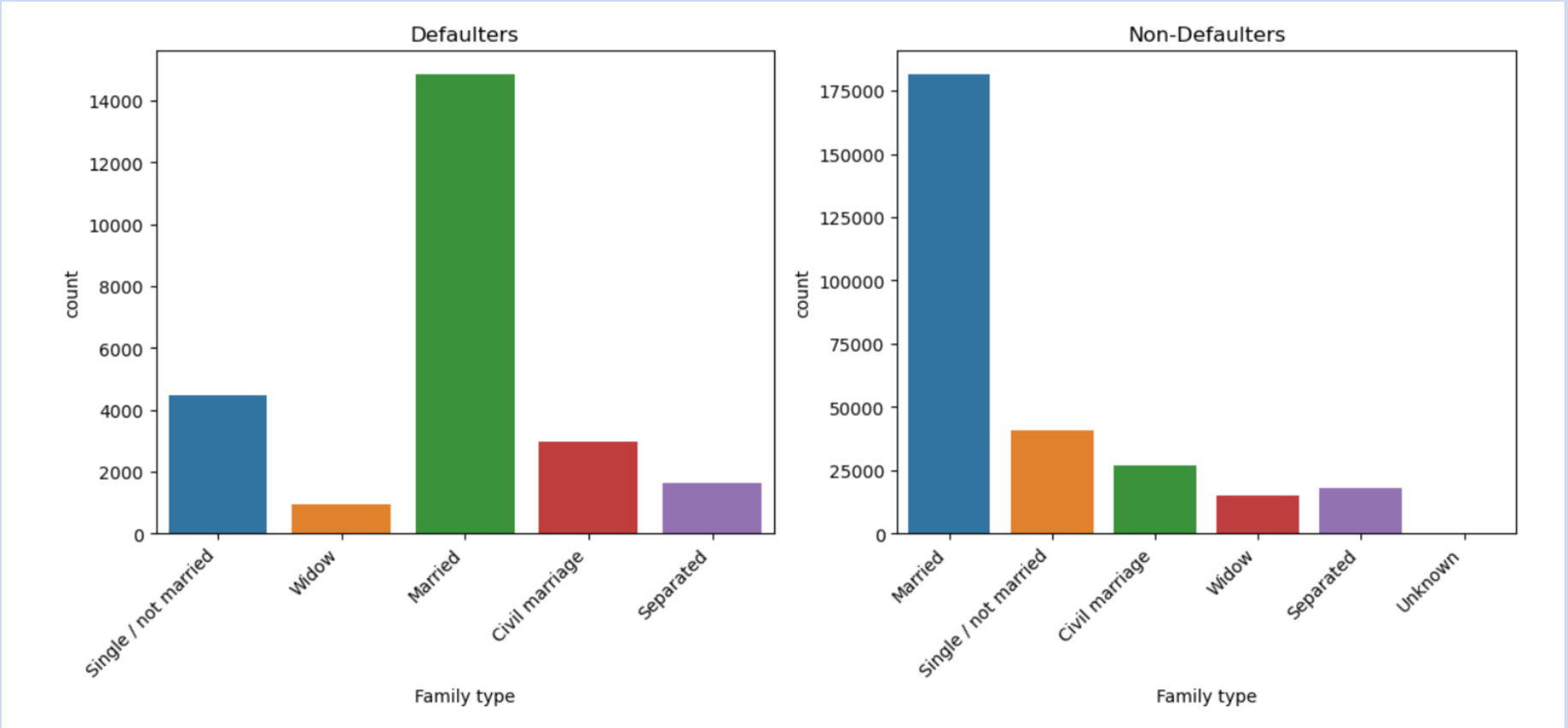
Univariate Analysis - Analysis of defaulters based **Education type**



Defaulters - We can see that **Secondary Education** defaults the most, followed by Higher Education, followed by Incomplete higher followed by Lower secondary.

Non-defaulters - We can see that Secondary Education defaults the most followed by Higher education, followed by Incomplte education, followed by Lower secondary

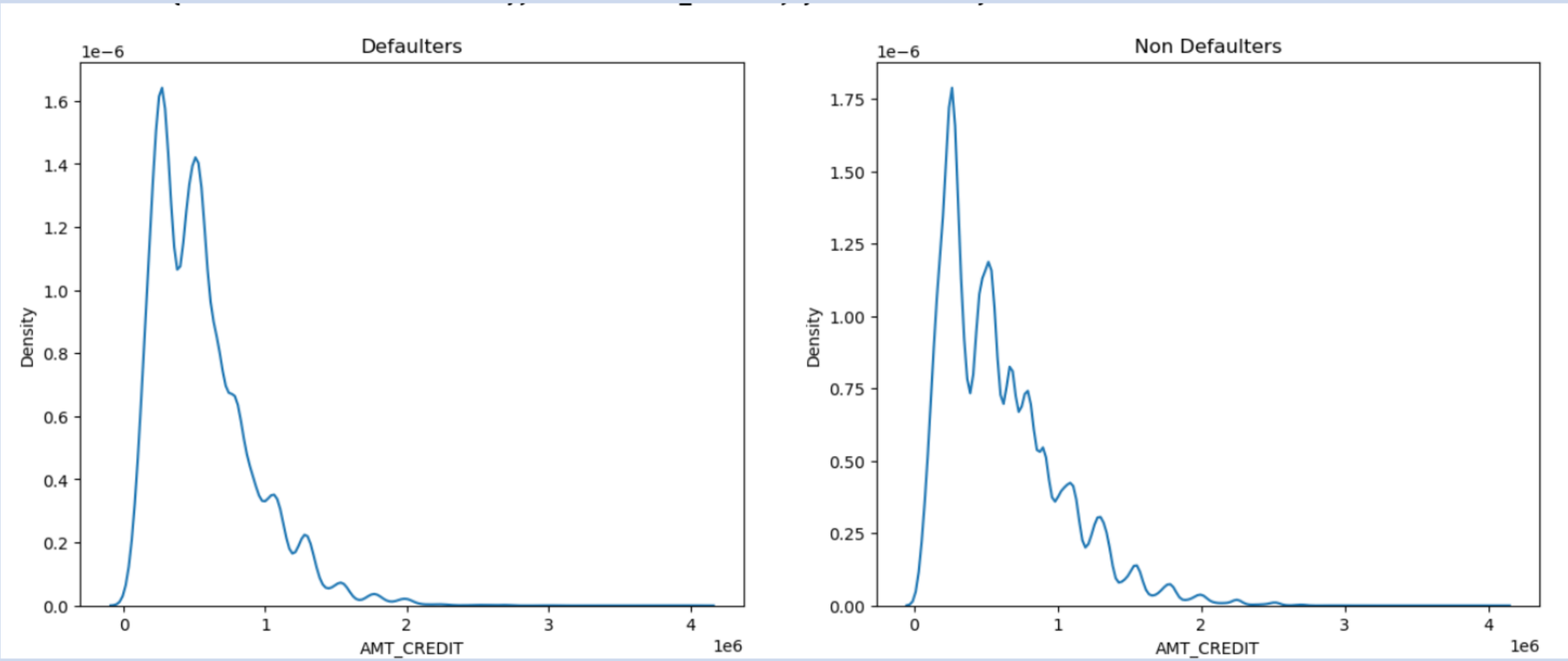
Univariate Analysis - Analysis of defaulters based **Family type**



Defaulters - We can see that **Married category** defaults the most, followed by Single followed by Civil marriage, followed by Separated, followed by Widow

Non-defaulters - We can see that Married category defaults the most, followed by Single followed by Civil marriage, followed by Separated, followed by

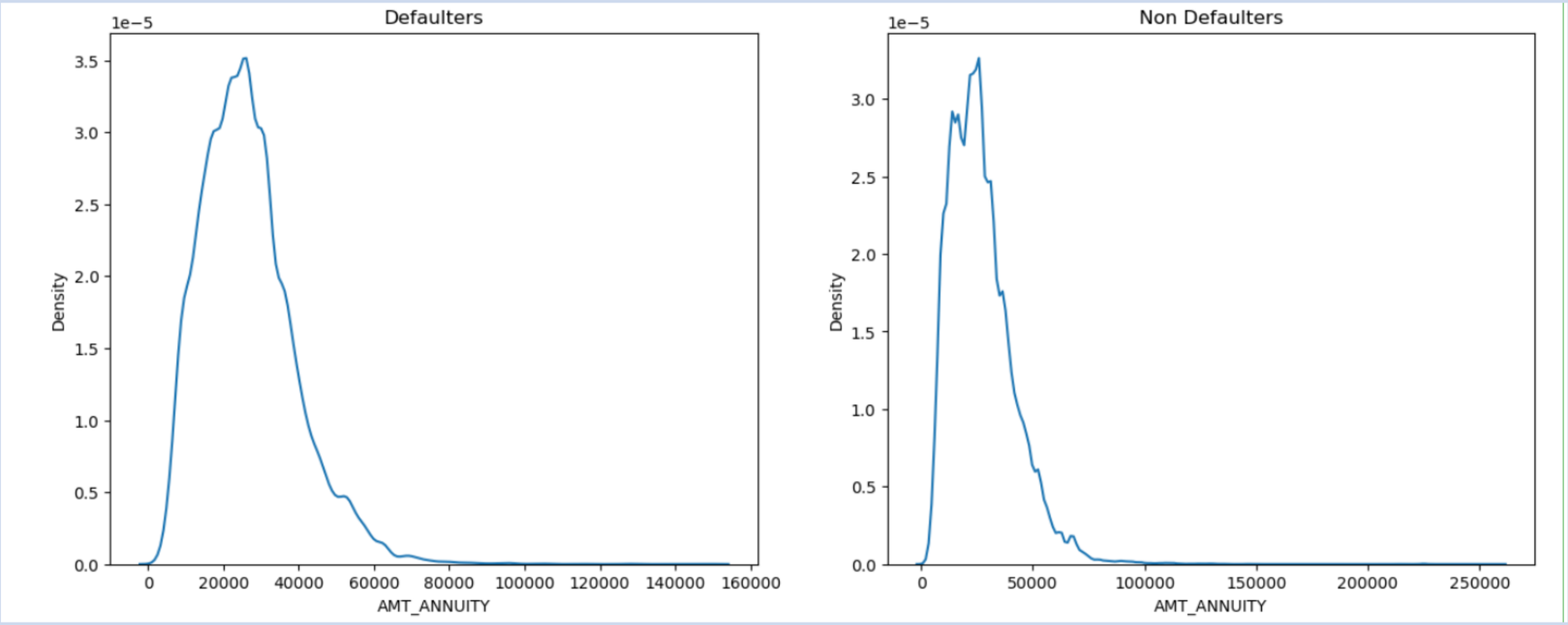
Analysis based **AMT CREDIT**



Defaulters - We can notice that the lesser the credit amount of the loan, the more chances of being defaulter. The spike is till 500000.

Non defaulters - If the credit amount is less, there is lesser chance of being defaulted. And gradually the chance is being decreased with the loan credit amount.

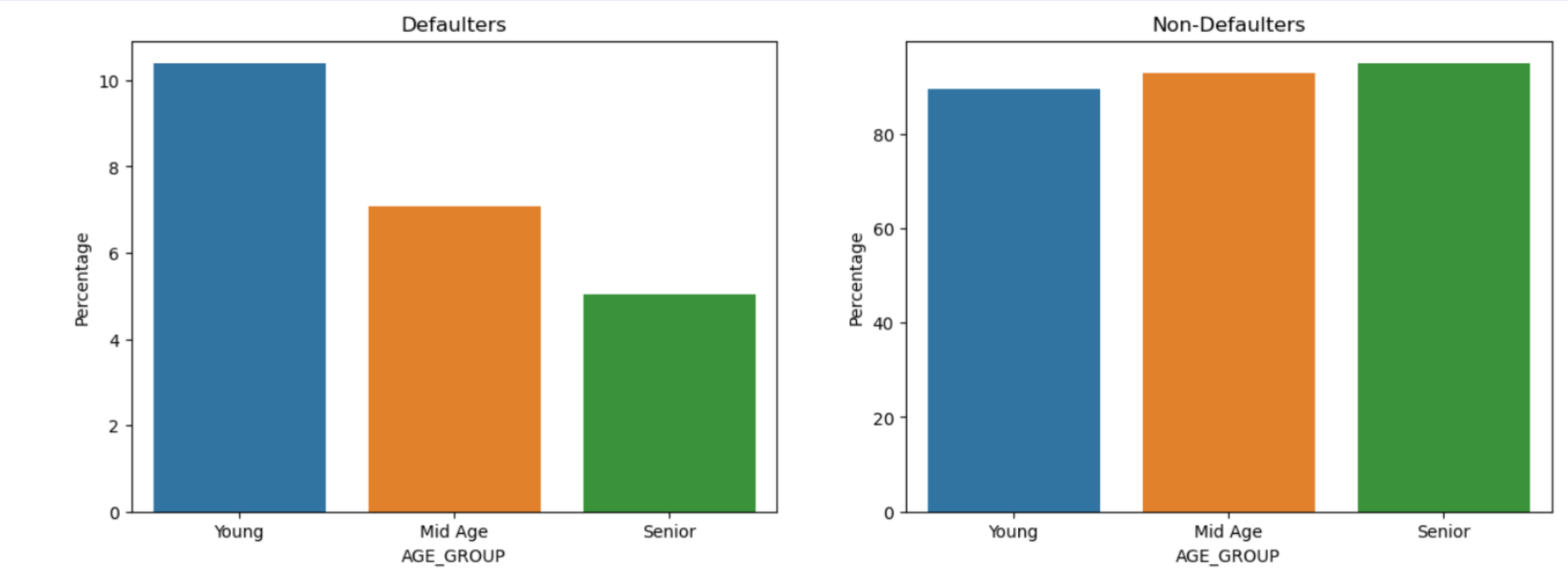
Analysis based **AMT_ANNUITY**



Defaulters - We can see that default more at 300000 .

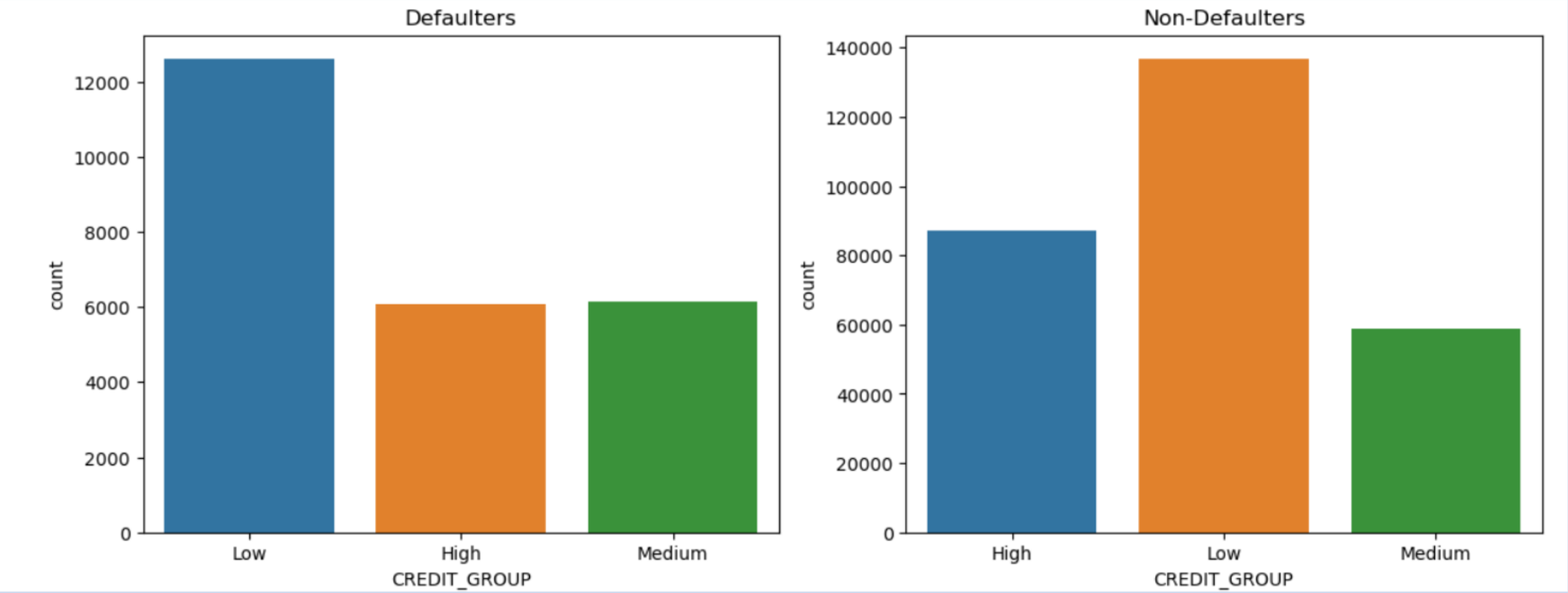
Non-defaulters - The same pattern continues for non-defaulters as well.

Analysis based **AGE_GROUP**



Defaulters - We can see that **Young** default more than Mid Age followed by Senior
Non-Defaulters - We see that Senior people are more likely to not default than other two age groups. Followed by Mid Age, followed by Young

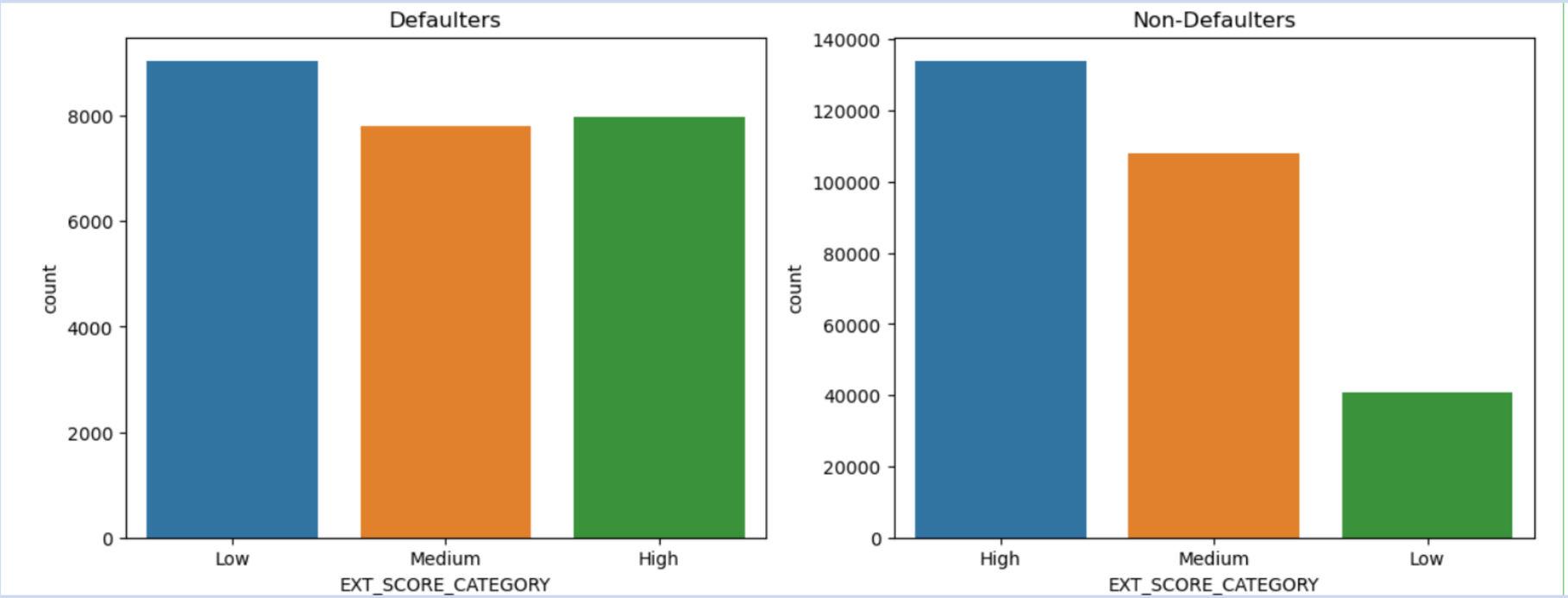
Analysis based **CREDIT_GROUP**



Defaulters - We can see that **LOW** credit group default more than high & medium

Non-Defaulters - We see the same trend

Analysis based **EXT_SCORE_CATEGORY**

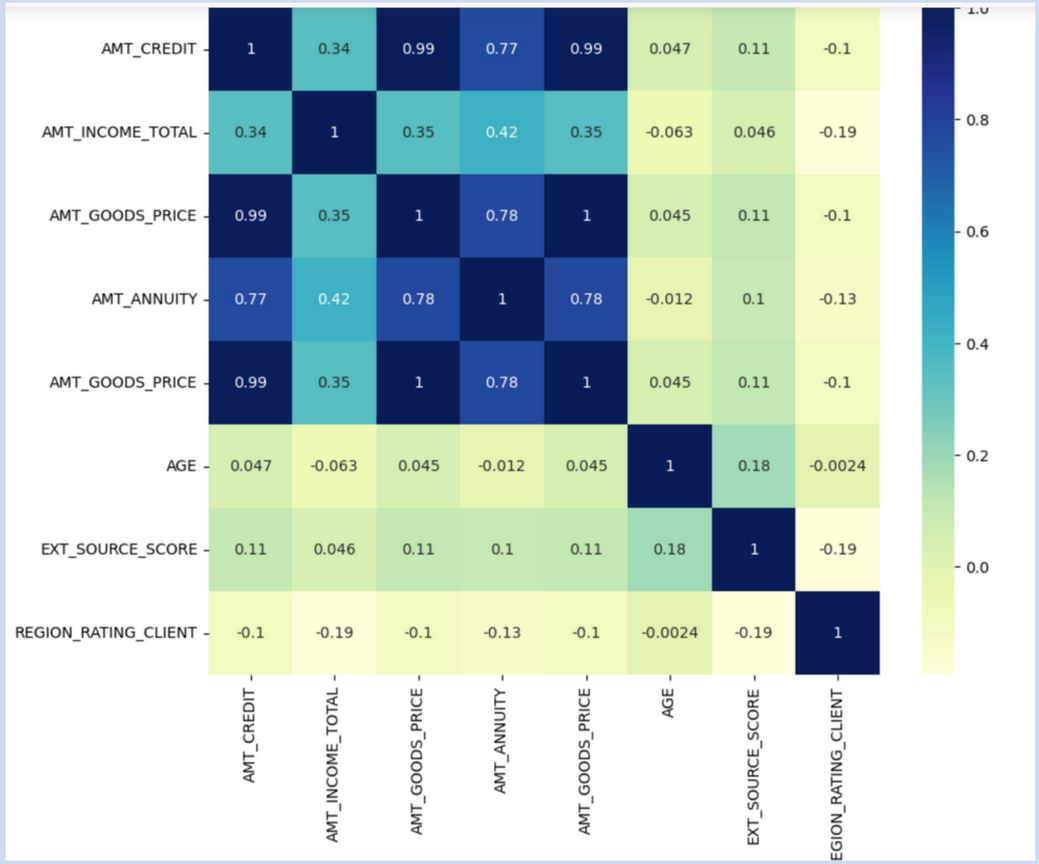


Defaulters - **Low score** group defaults more than High and Medium Ext Score.
Non defaulters - Non defaulters more in High EXT_SCORE followed by low EXT score.

Correlation Analysis - Analysis for all factors that correlate for defaulters

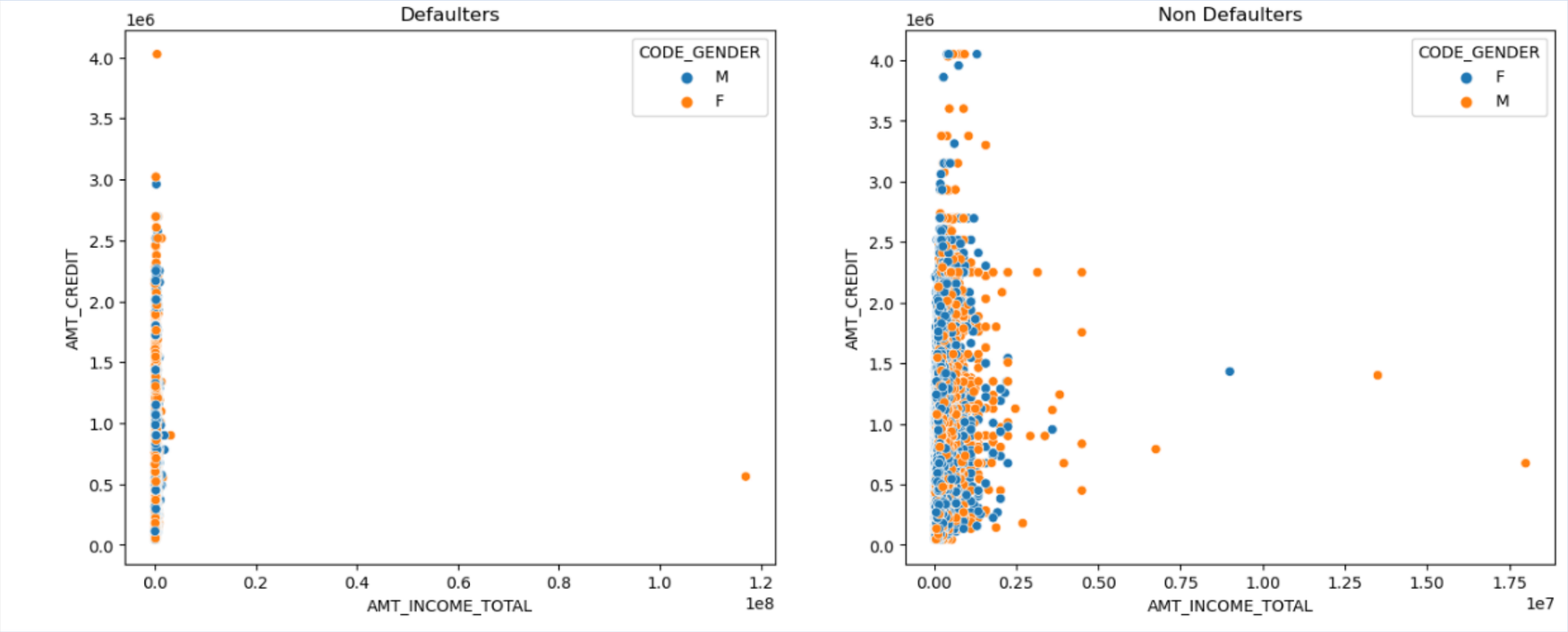
AMT_CREDIT and AMT_ANNUITY (0.75)
AMT_CREDIT and AMT_GOODS_PRICE (0.1)
AMT_ANNUITY and AMT_GOODS_PRICE (0.75)

Correlation Analysis - Analysis for all factors that correlate for non defaulters



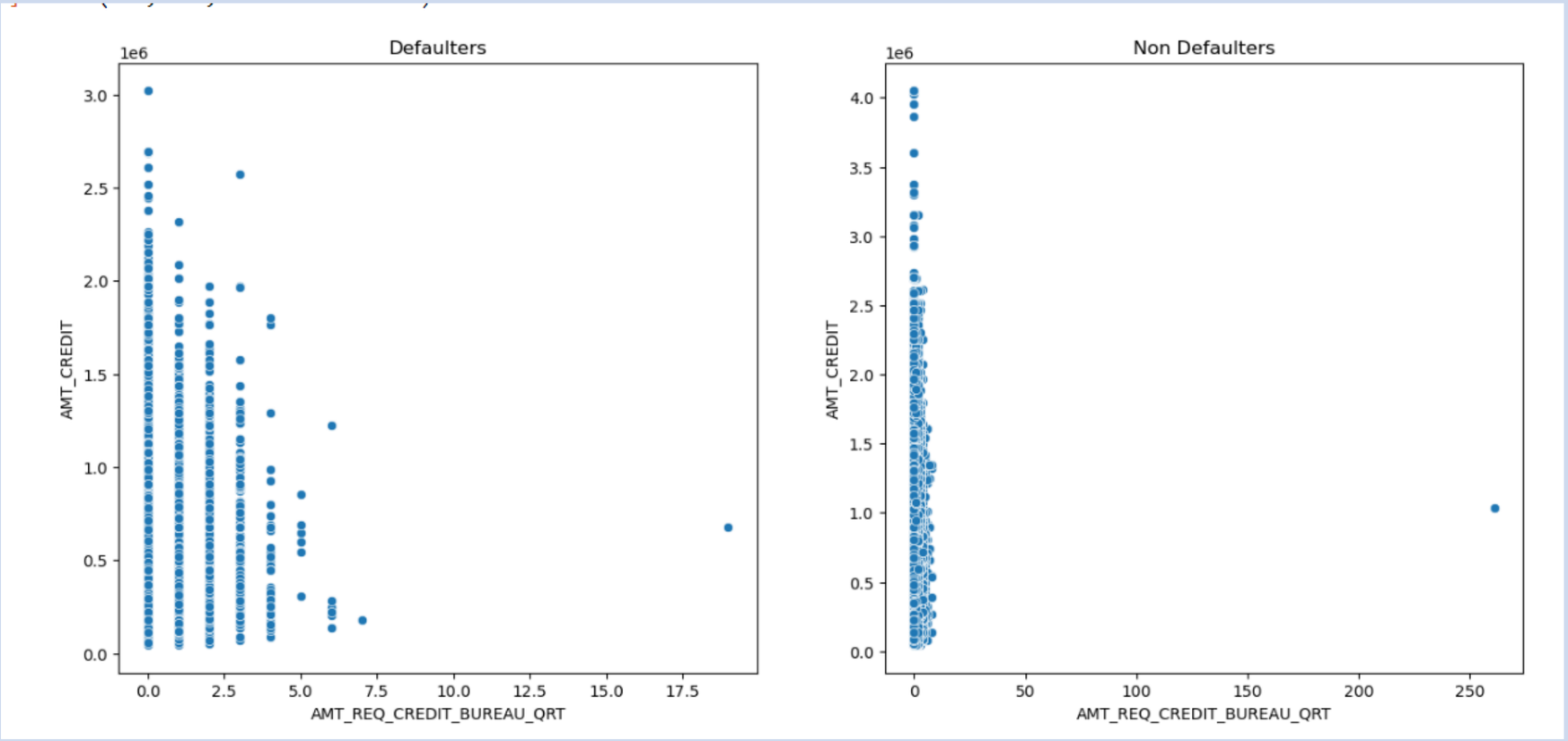
AMT_CREDIT and AMT_ANNUITY (0.78)
AMT_CREDIT and AMT_GOODS_PRICE (1)
AMT_ANNUITY and AMT_GOODS_PRICE (0.78)

Bivariate analysis on continuous variable



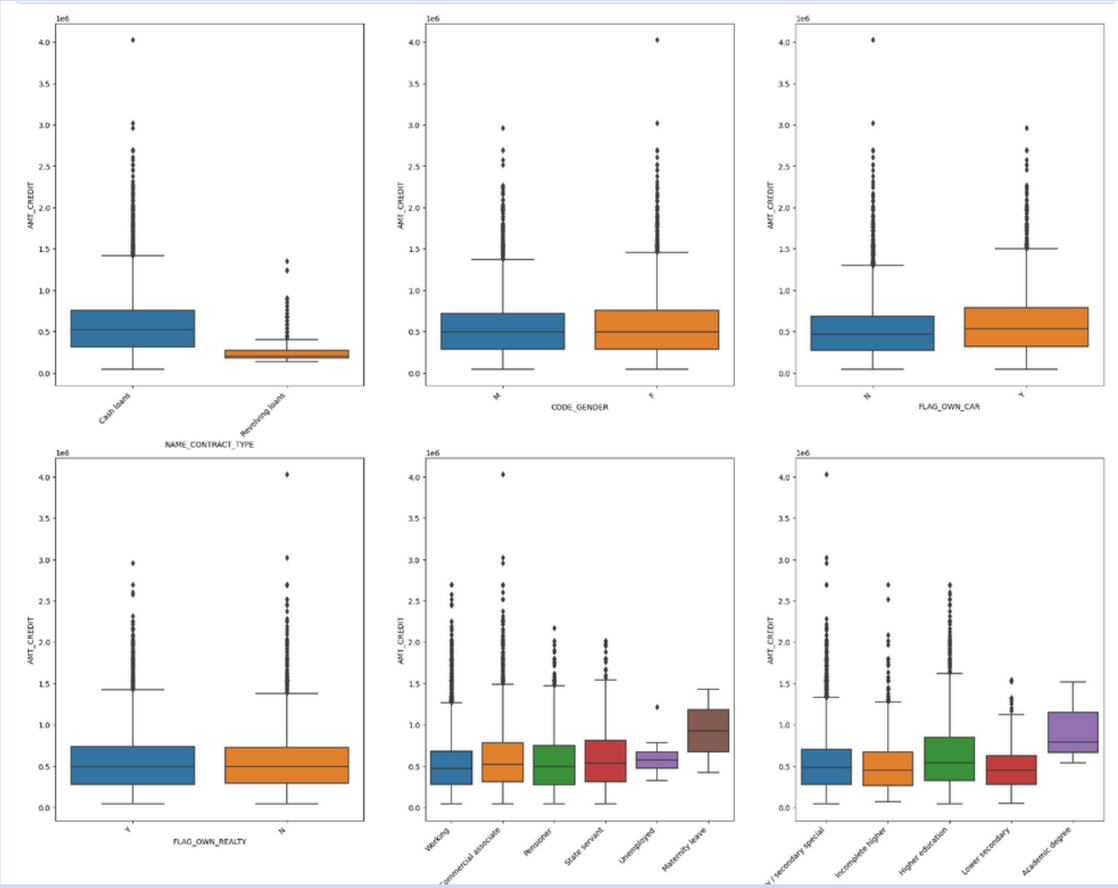
Defaulters - We can slightly figure out that the values are more concentrated on the lower income and lower credit of the loan. That means as the income is increased, the amount of loan is also increased. This is true for both the genders.
Non defaulters - We can hardly figure out any pattern out of this.

Bivariate analysis on continuous variable

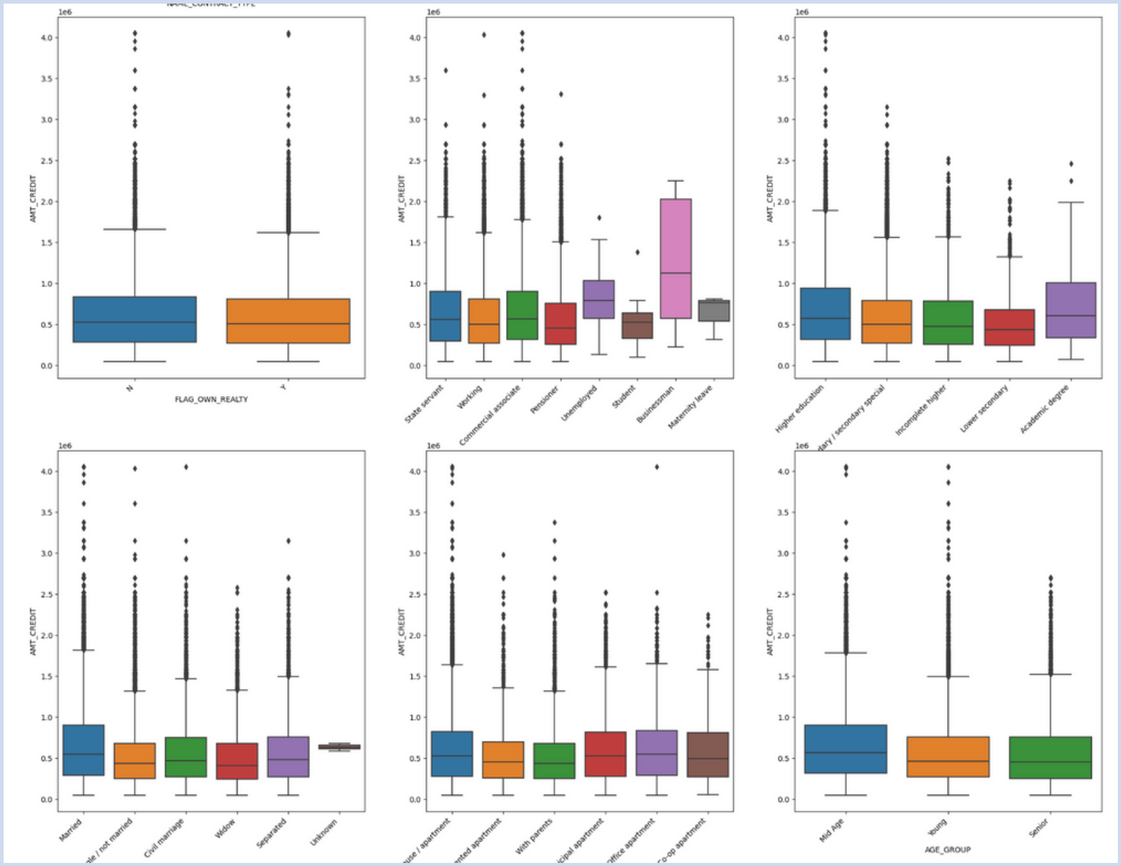


Defaulters - We can slightly figure out that the values are more concentrated on the lower income and lower credit of the loan. That means as the income is increased, the amount of loan is also increased. This is true for both the genders.
Non defaulters - We can hardly figure out any pattern out of this.

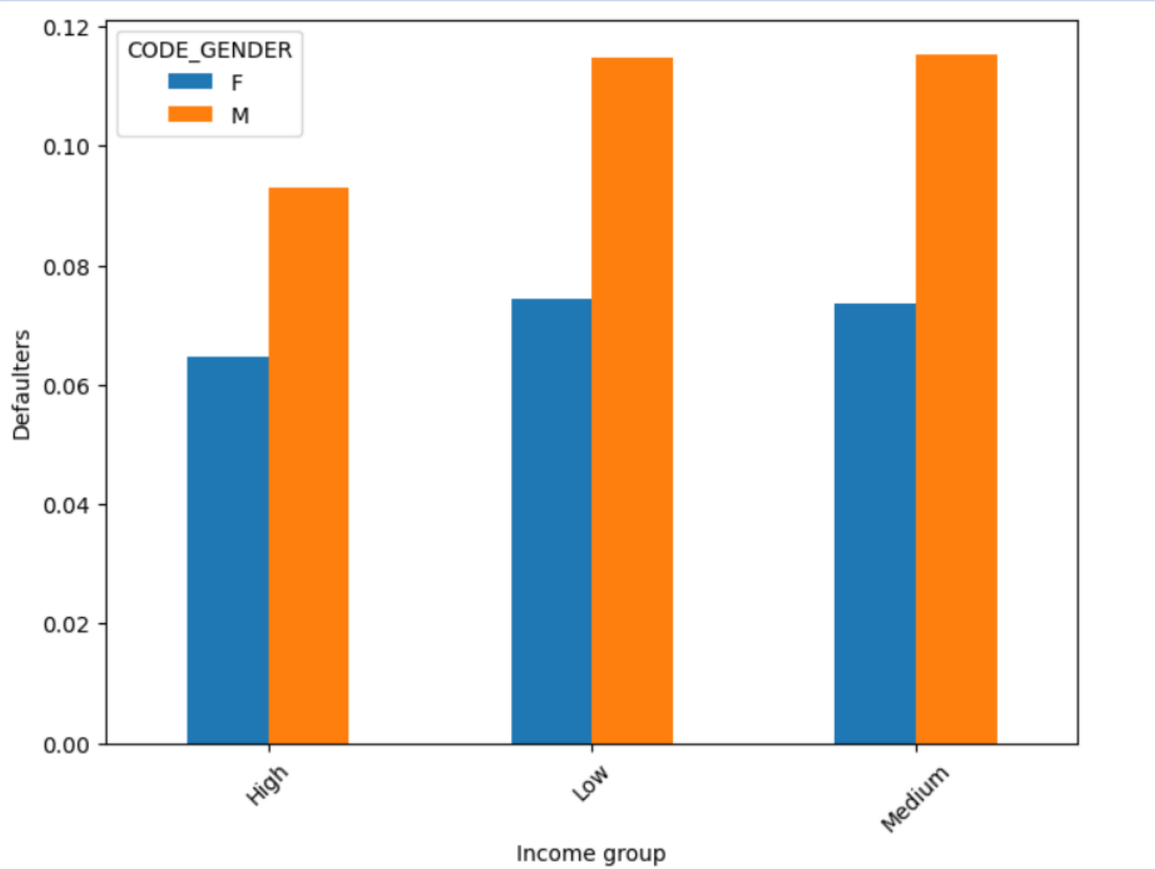
Bivariate analysis of continuous variables
Credit amount of the loan of various categories



Bivariate analysis of continuous variables
Credit amount of the loan of various categories (contd)

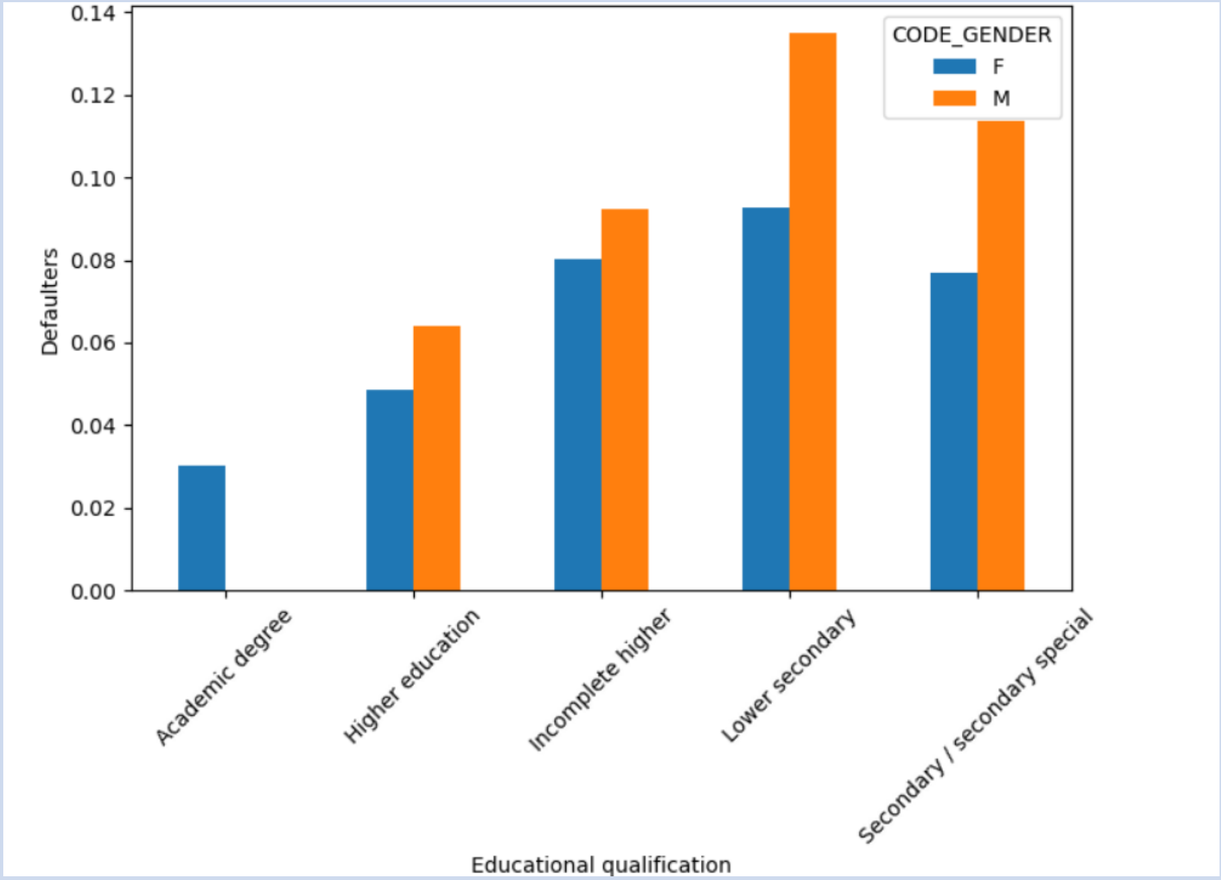


Analysis of two segmented variables for defaulters - **Income group & Gender**



We can see that Males are more likely default than Females accross all income groups.

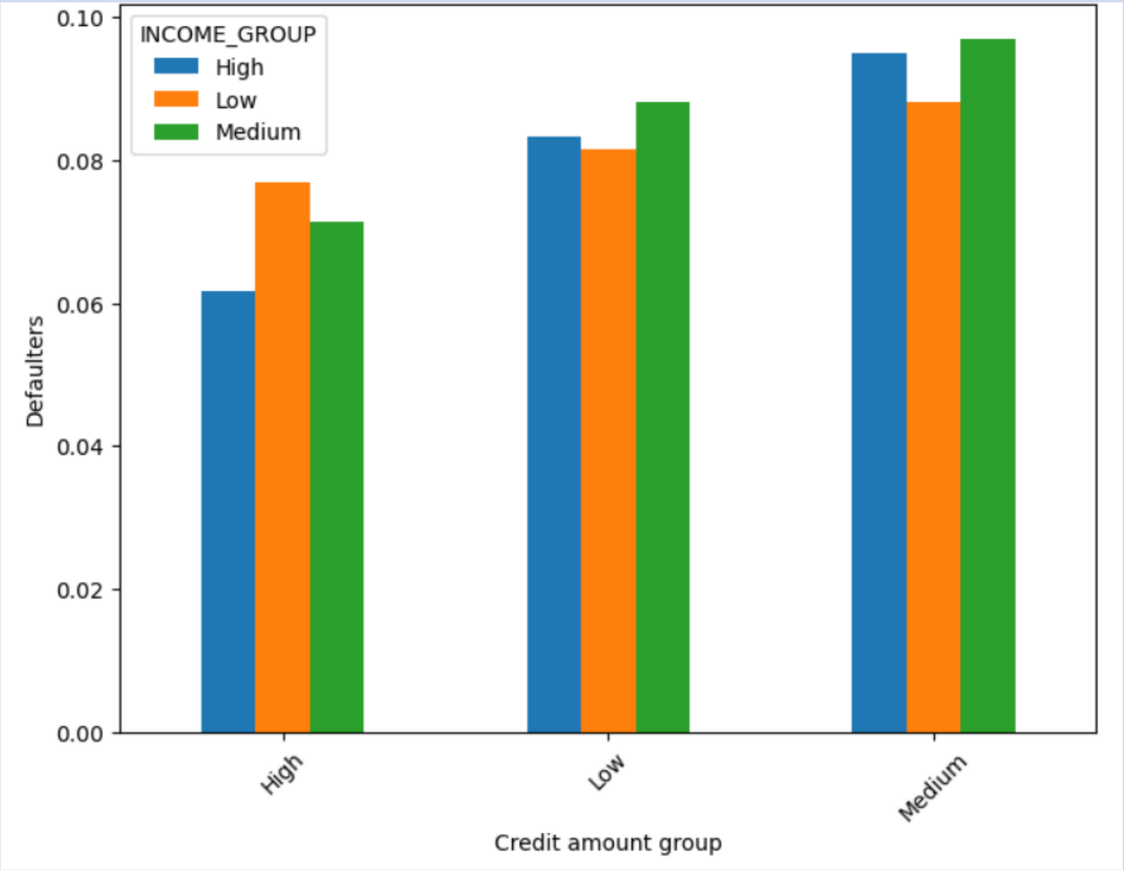
Analysis of two segmented variables for defaulters - **Education & Gender**



Defaulters males — We can see that lower secondary education males have the highest defaults followed by secondary special, followed by incomplete higher, followed by higher education

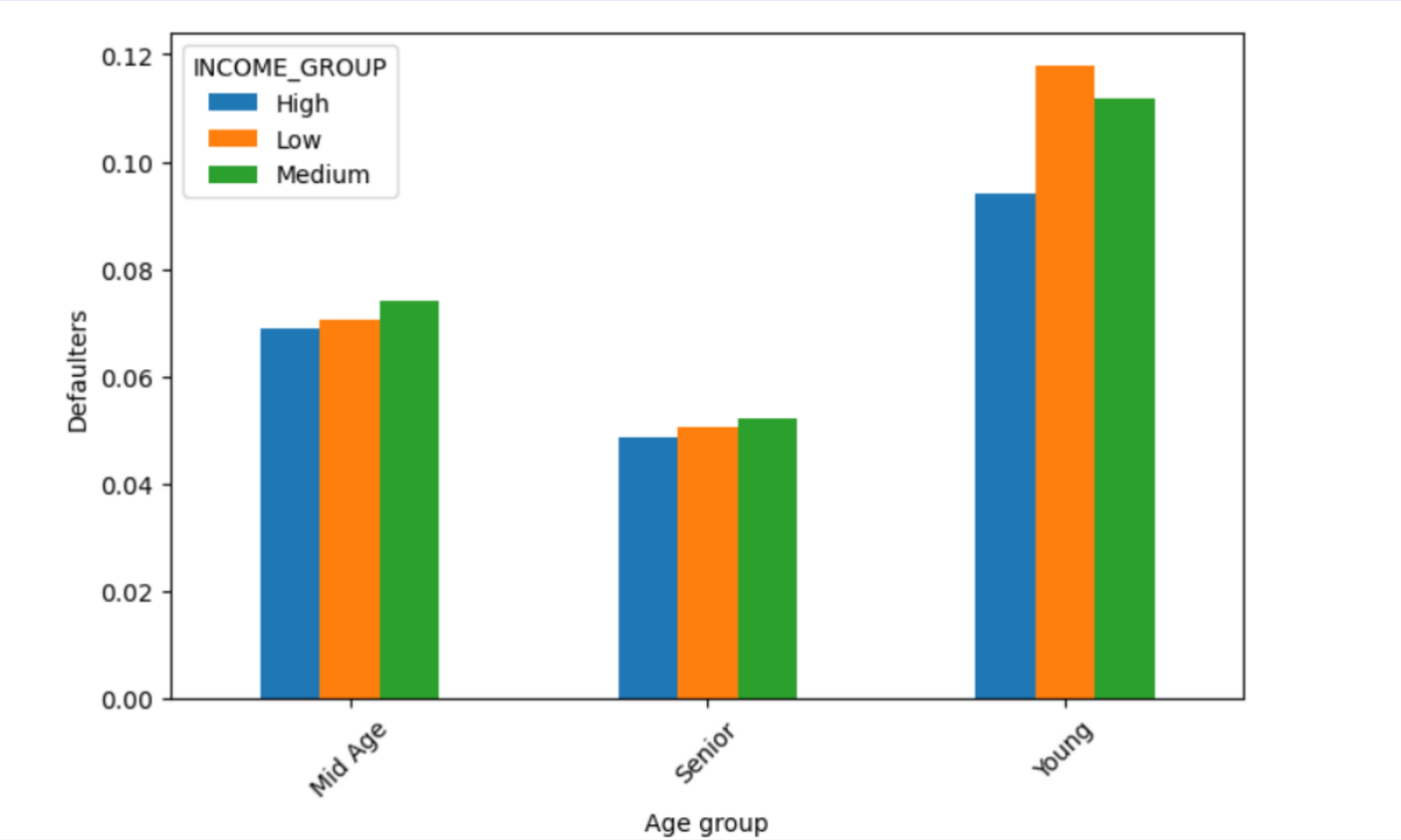
Non-Defaulters females — The same pattern continues for non-defaulters as well.

Analysis of two segmented variables for defaulters - **Credit amt group & Income Group**



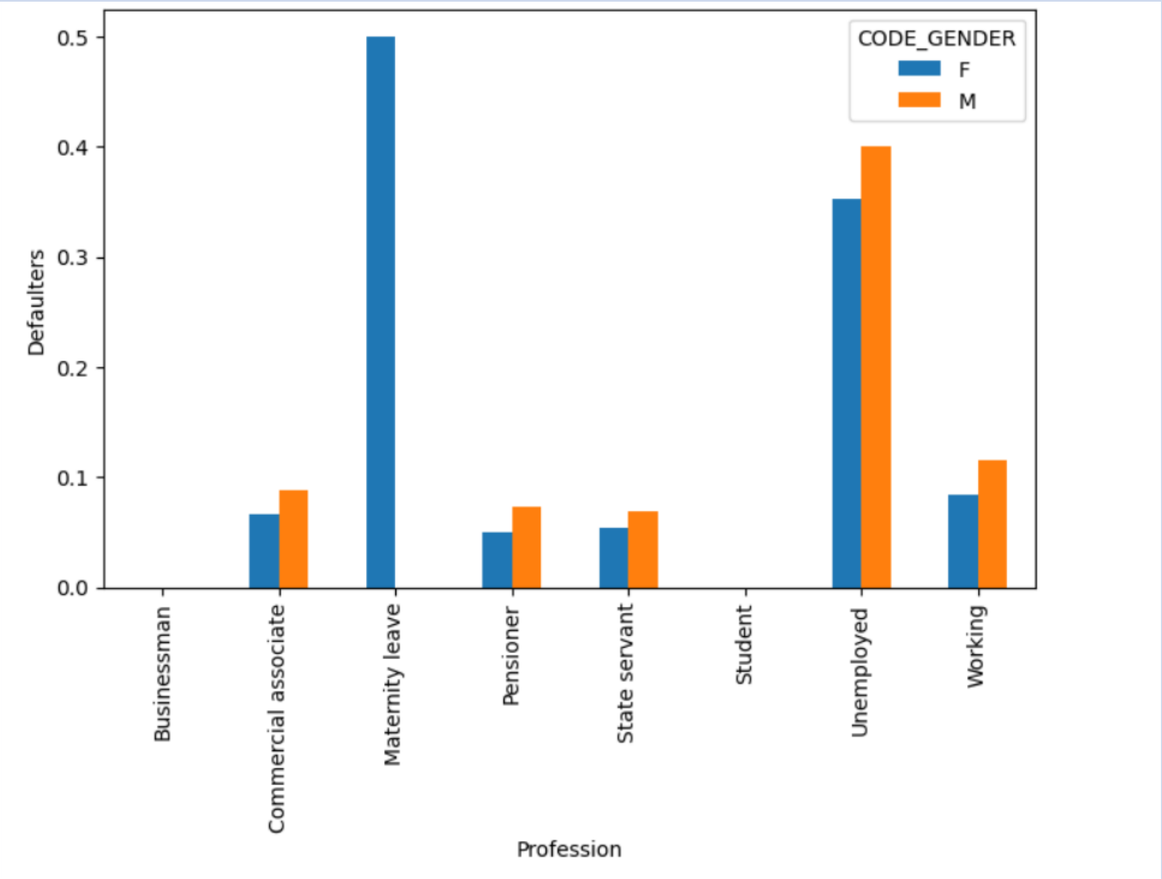
1)Medium credit amount group are defaulted the most in all income groups.
2)High credit amount groups are less likely to default in all income groups.

Analysis of two segmented variables for defaulters - **Age group & Income Group**



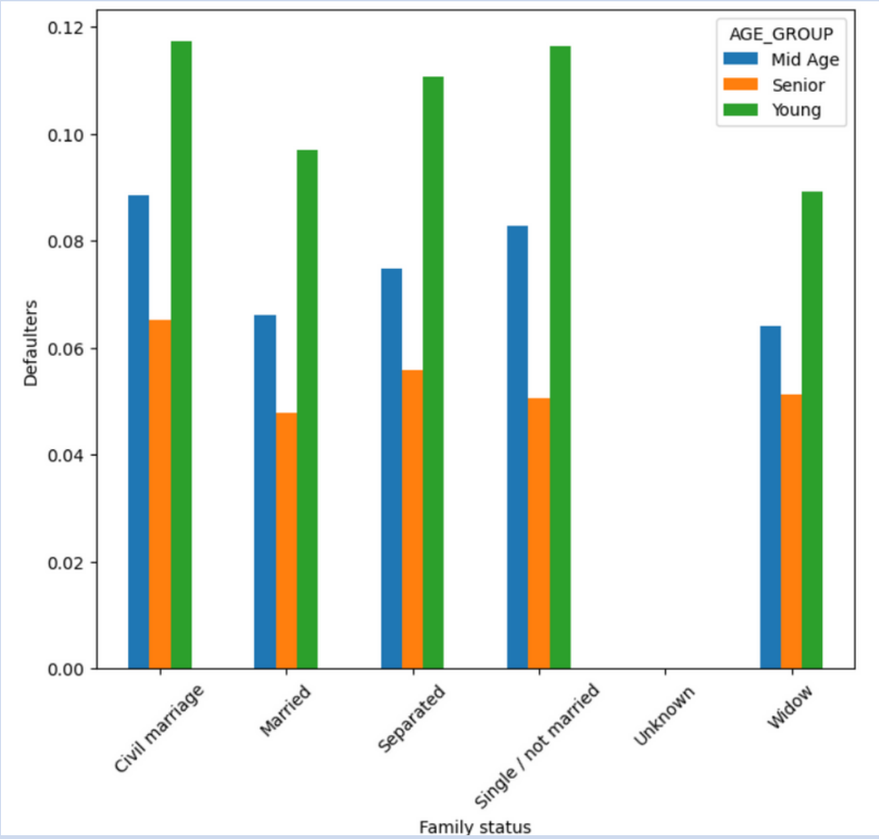
Young default more than Mid Age, followed by Senior age group

Analysis of two segmented variables for defaulters - **Profession & Gender**



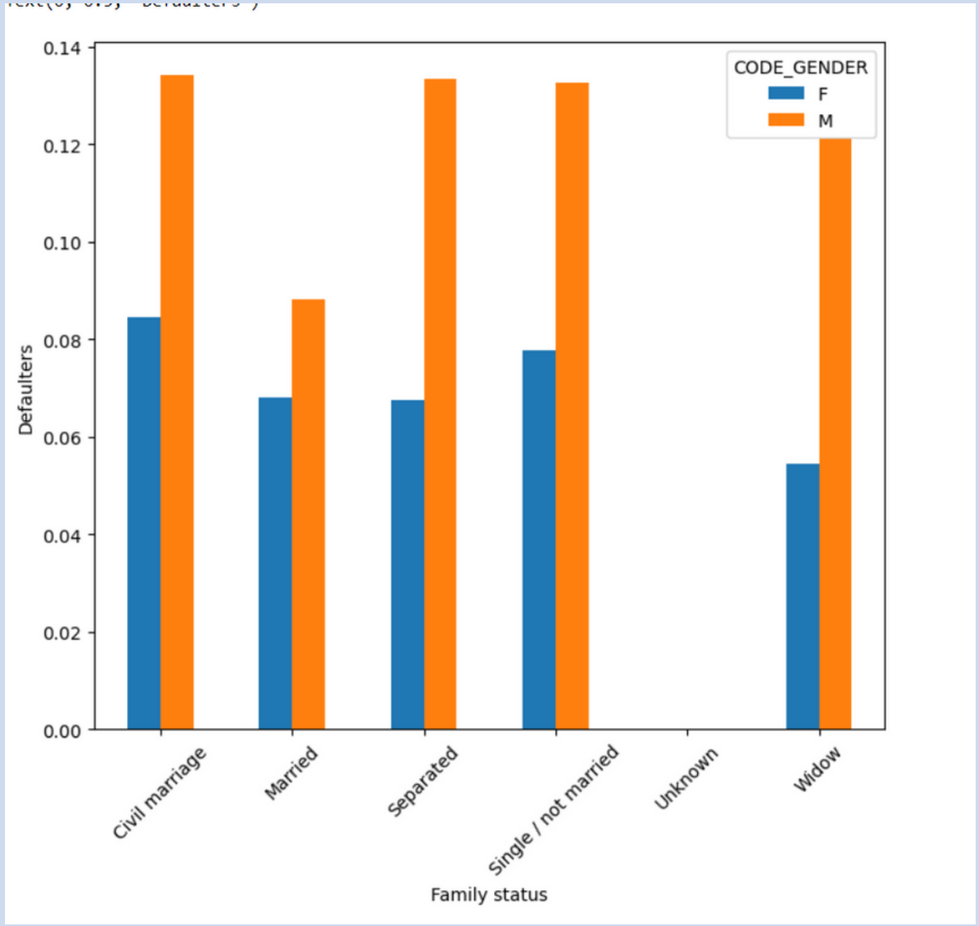
Maternity females default the most, followed by unemployed males, followed by unemployed females

Analysis of two segmented variables - **Family status & Age group**



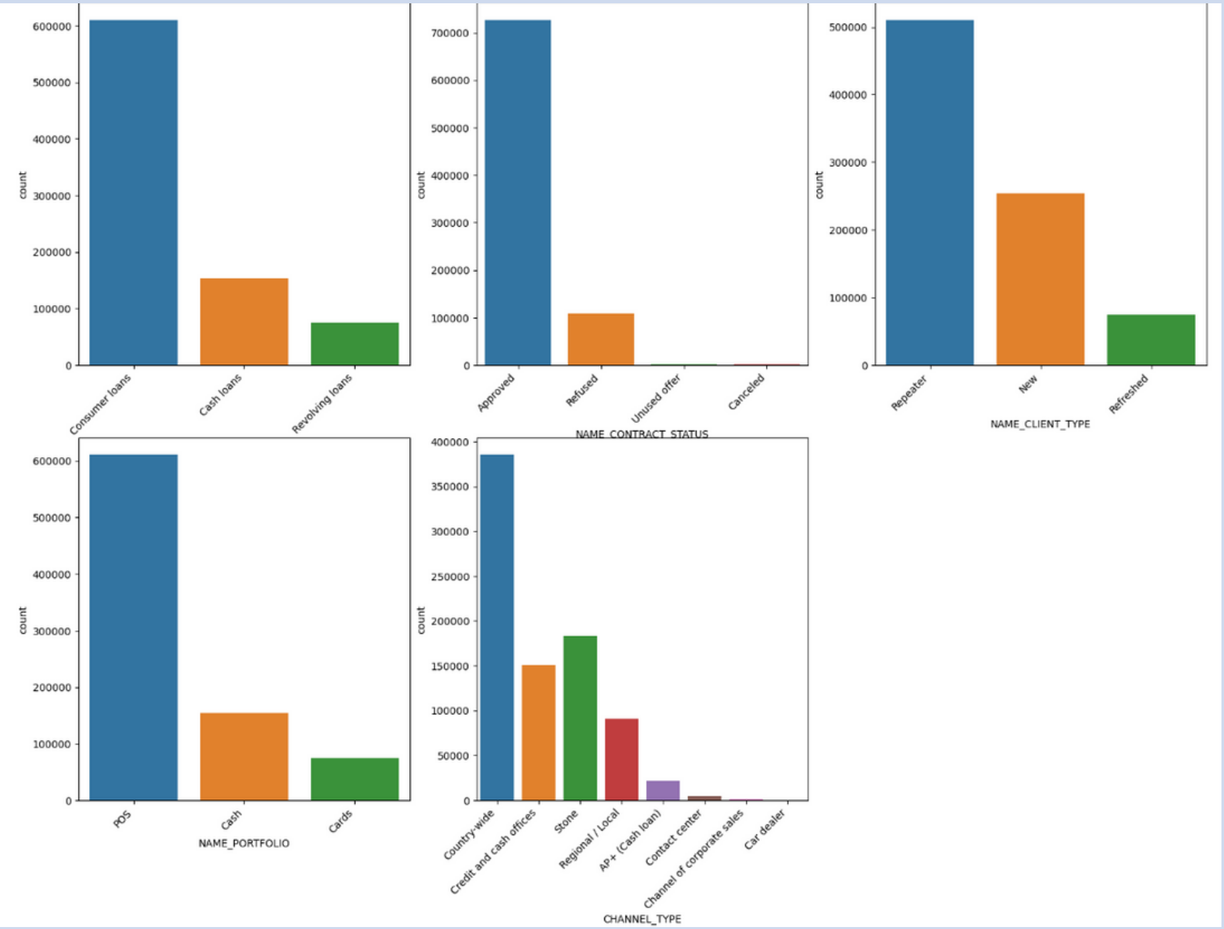
Civil marriage young couples default the most followed by singles.

Analysis of two segmented variables - **Family status & Gender**

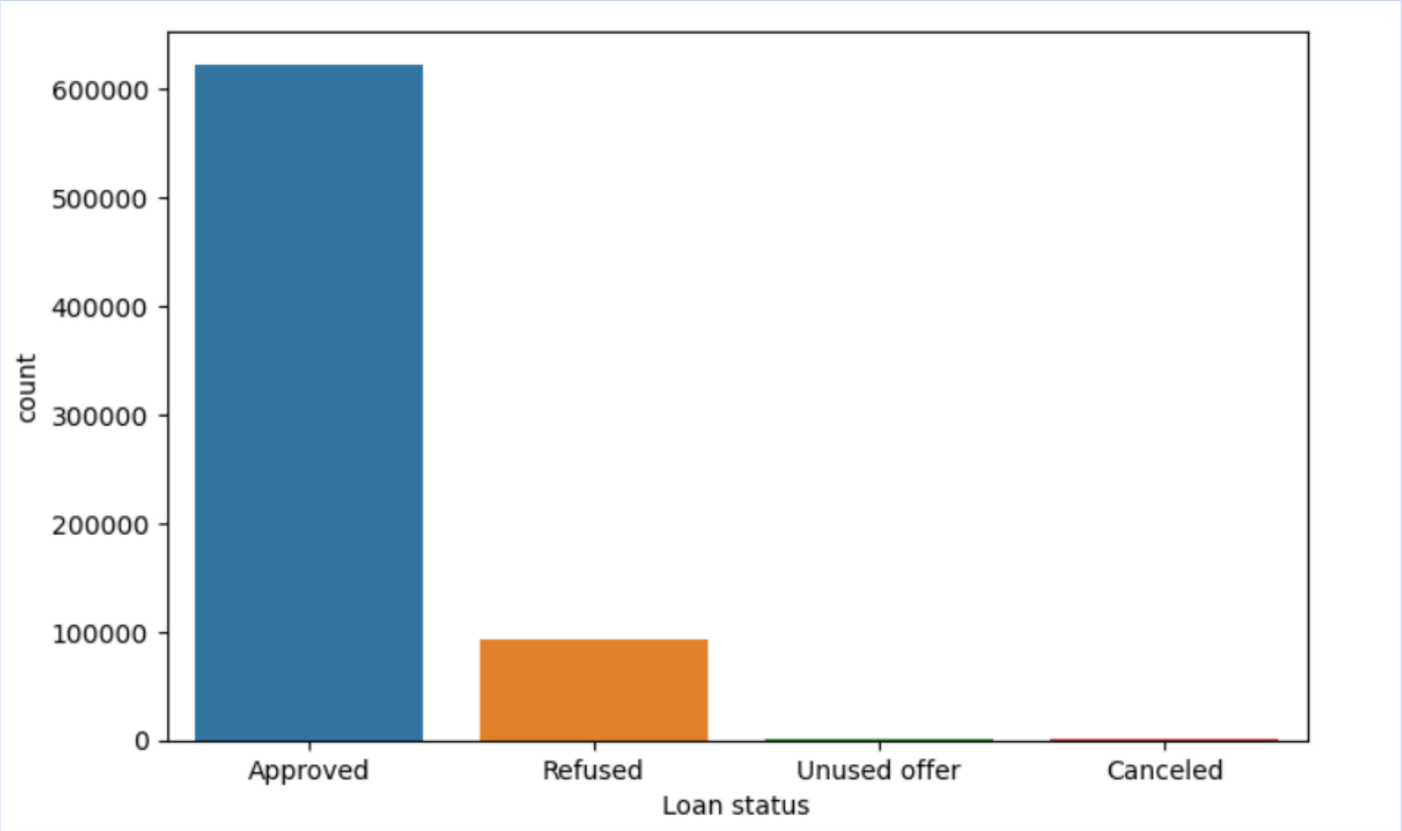


Civil marriage, Separated and Single men default the most

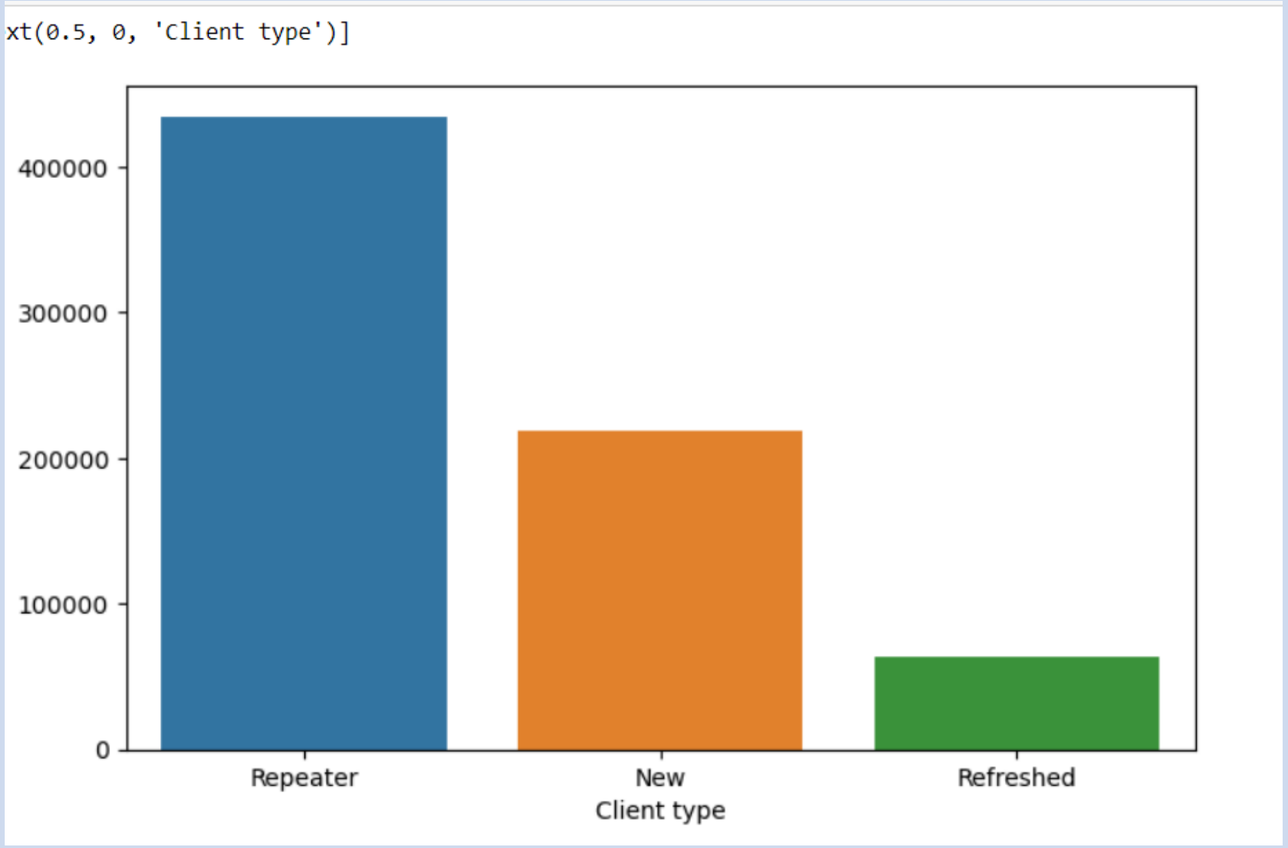
Univariate Analysis on multiple factors



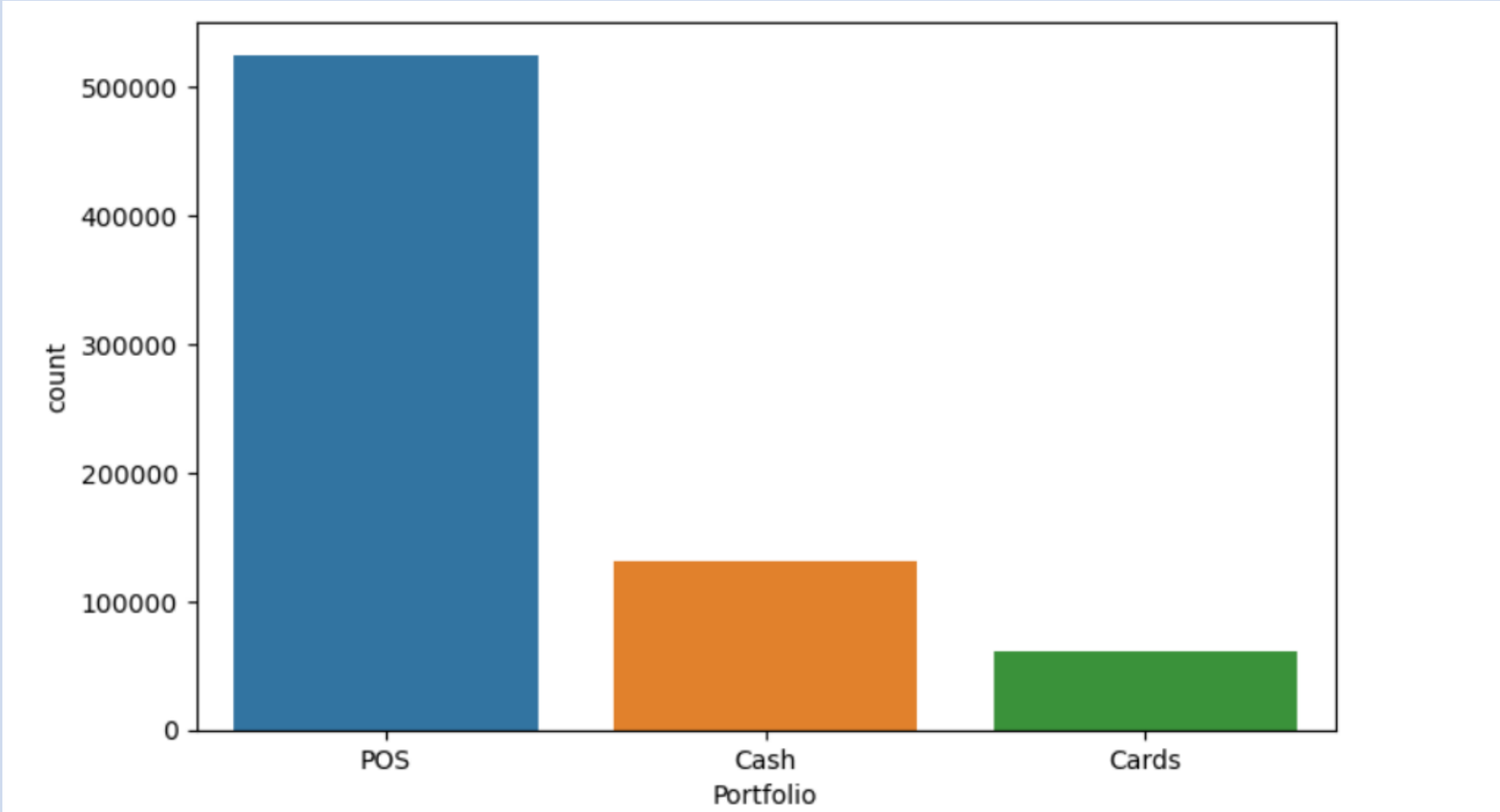
Univariate analysis on unordered categorical variable- **Previous Loan status**



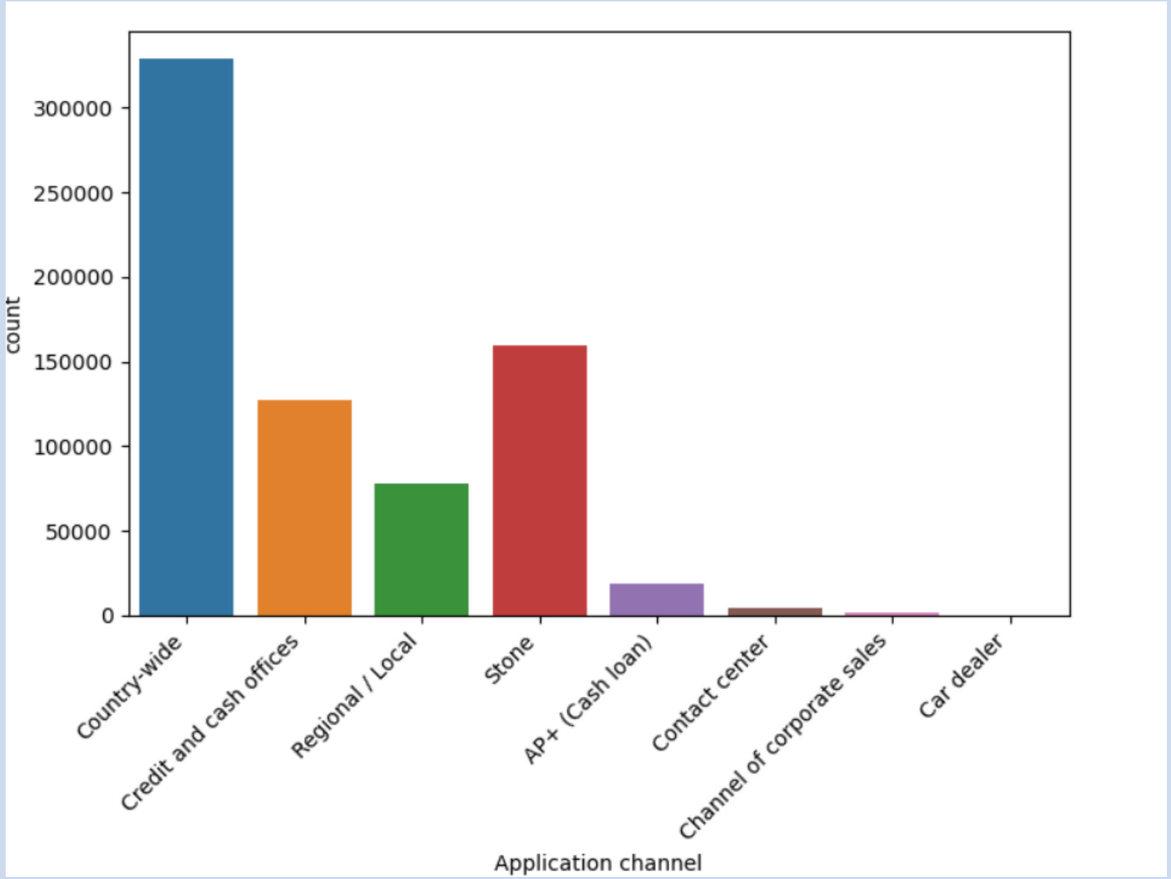
Univariate analysis on unordered categorical variable- **Client Type**



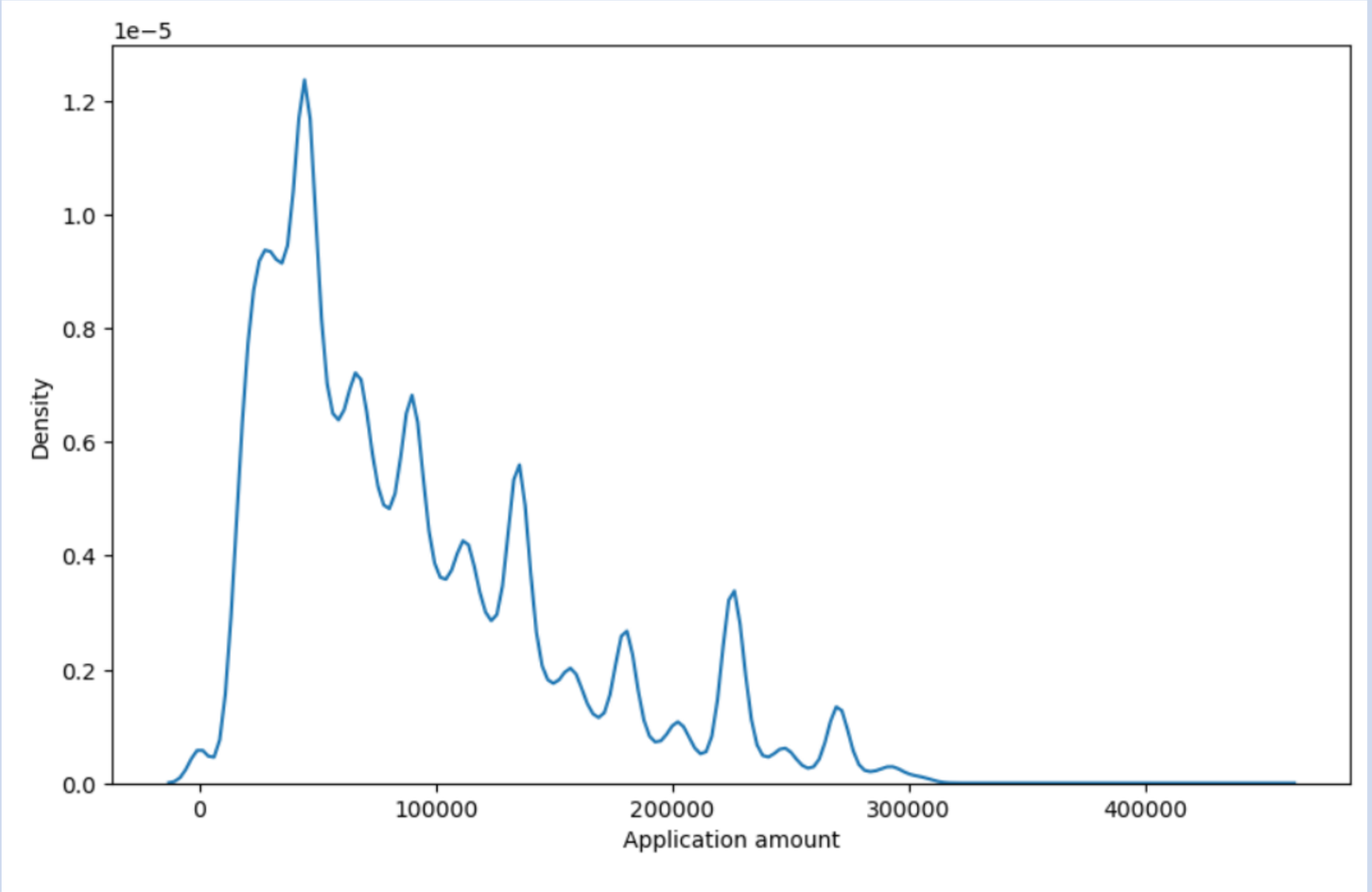
Univariate analysis on unordered categorical variable - **Portfolio Type**



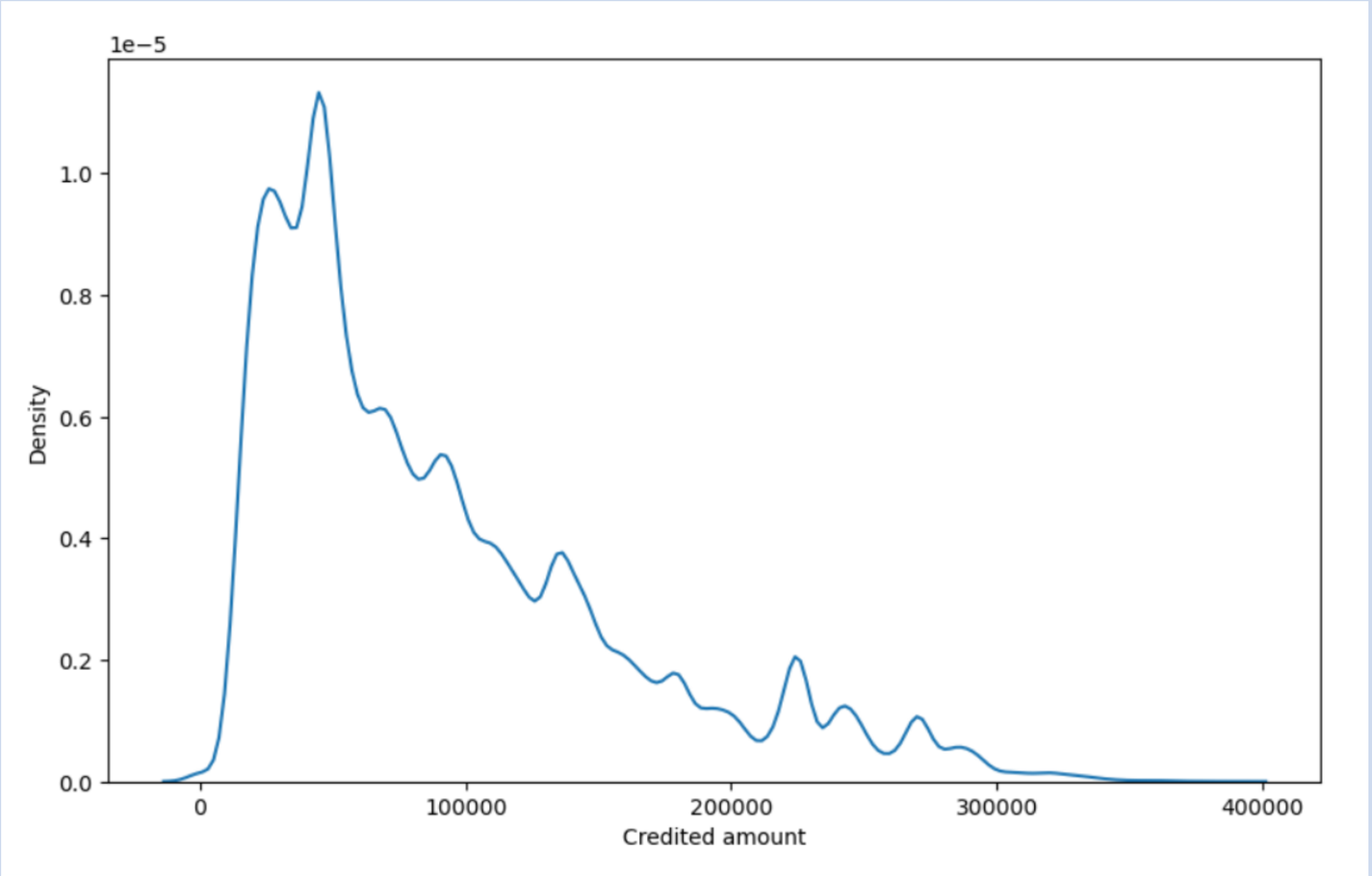
Univariate analysis on unordered categorical variable - **Application channel**



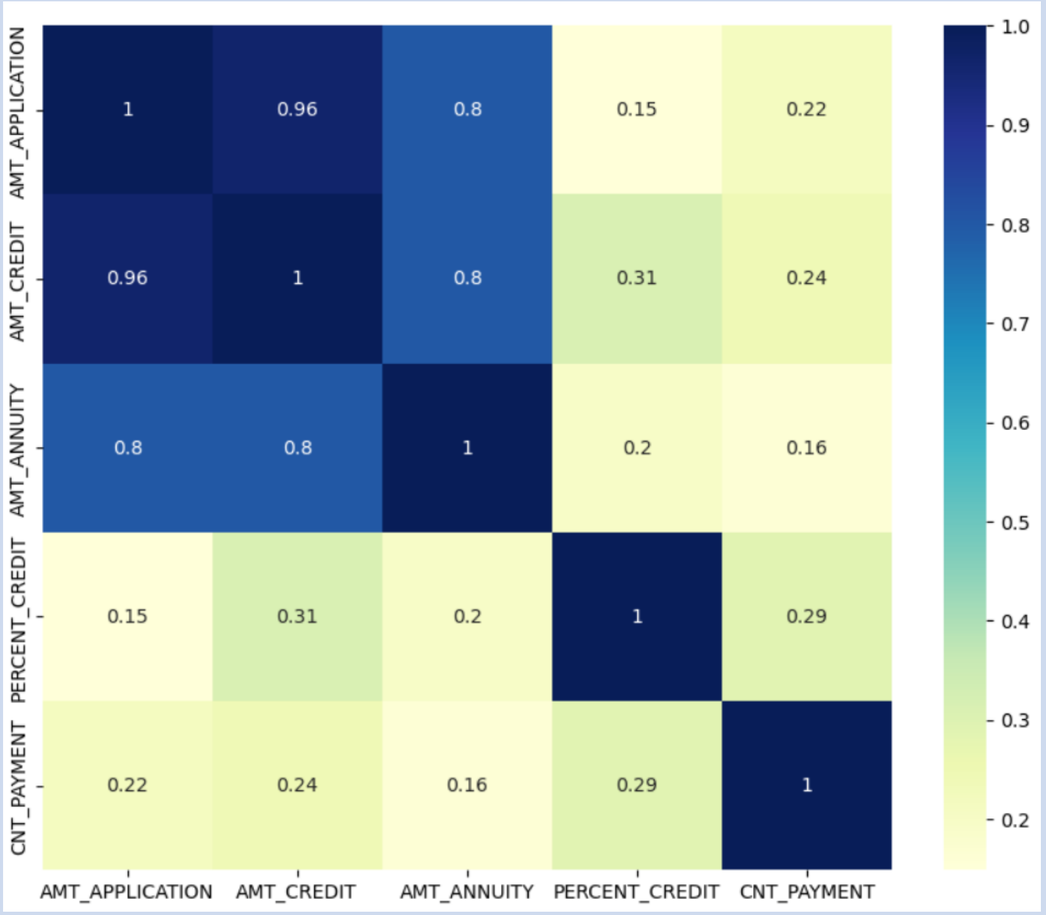
Univariate analysis for continuous variables **- Applied loan amount**



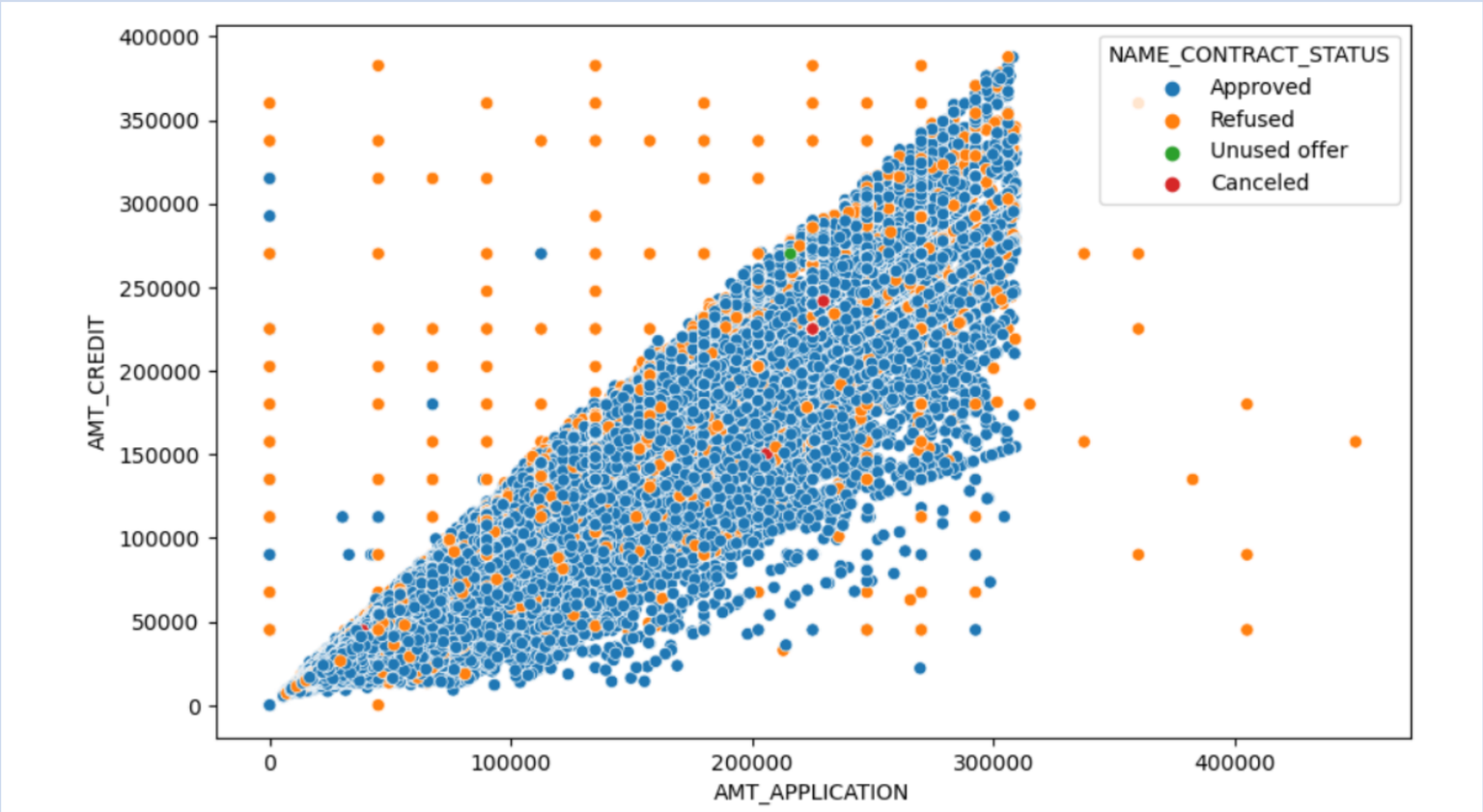
Univariate analysis for continuous variables **- Credit loan amount**



Correlation Analysis - Analysis for all factors that correlate for **defaulters**



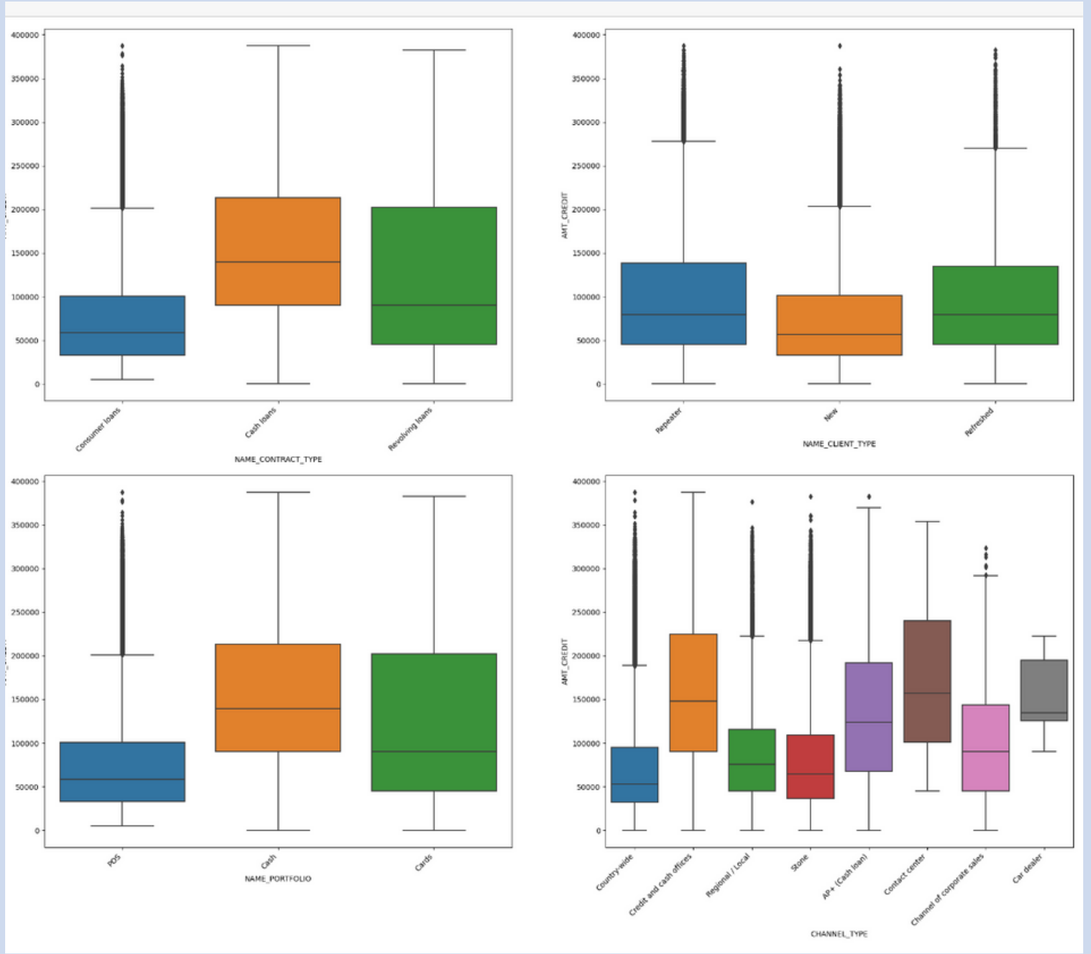
Correlation Analysis - Analysis for all factors that correlate for **defaulters**



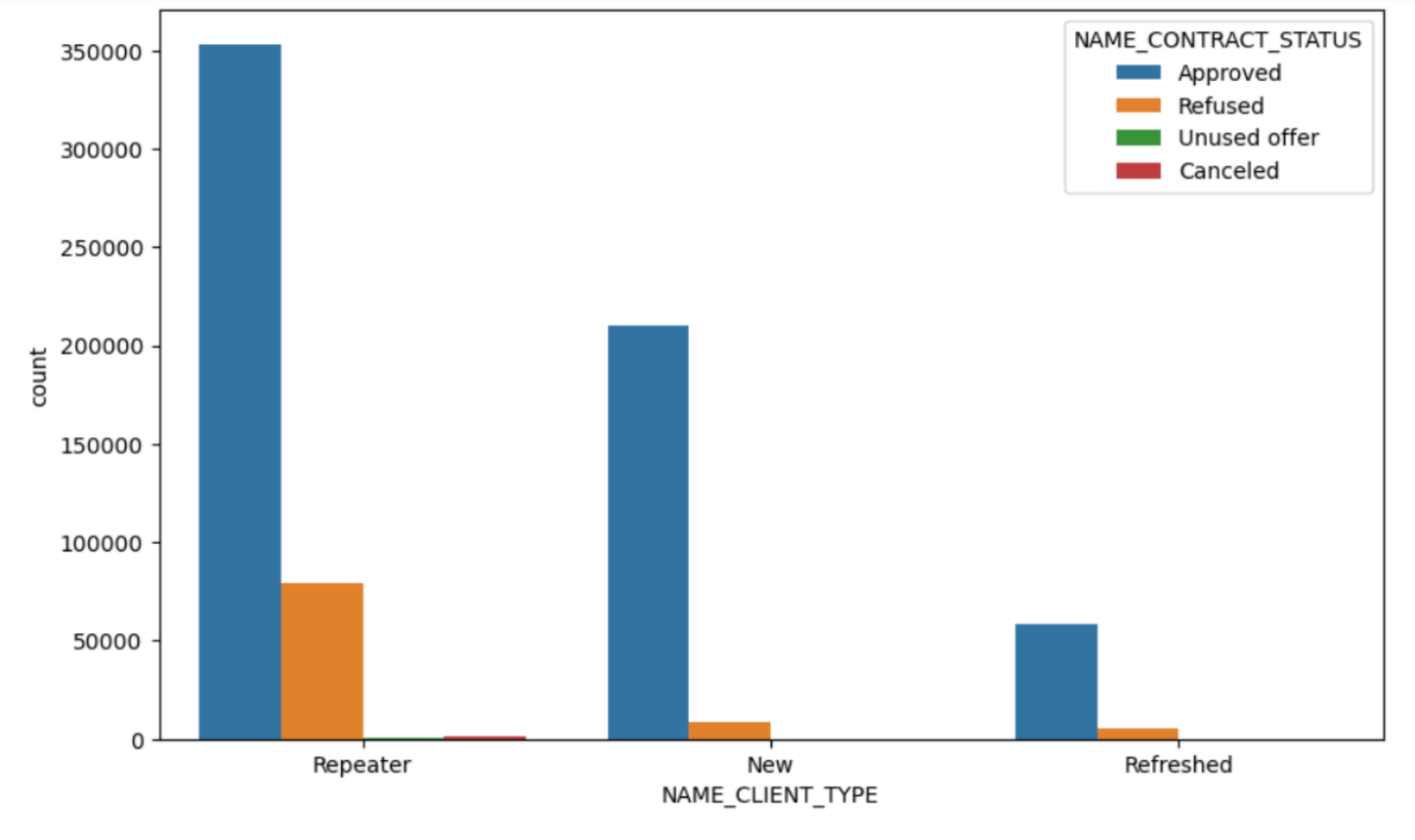
We can see that the applications are more concentrated on the lesser amount and so as the credited amount.

Bivariate analysis on categorical variable - Credit amount of the **loan of various categories**

1) Cash loans are credited more than Revolving and Consumer loans.
2) Repeater clients get more amount loan than new clients and refreshed clients.
3) The loan with portfolio Cars are credited more followed by Cash.
4) The credit amount of the loan is more from the application channel type

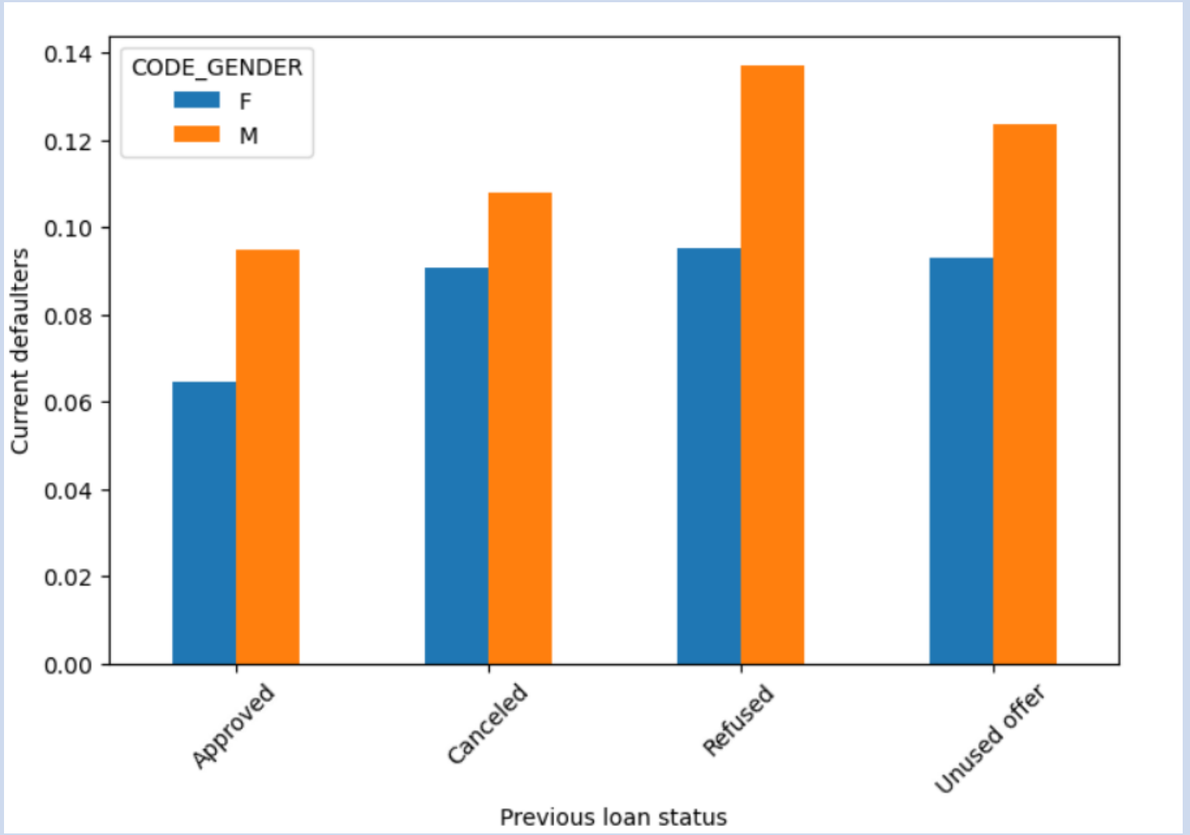


Analysis of two segmented variables - **Client type & Contract Status**



We see that the Repeater clients have more approved loans than New and Refreshed clients.

Analysis of two segmented variables - **Previous Loan Status & Gender**



We can see that the Refused Males have the highest defaluters, followed by Unused Offer, followed by Cancelled, followed by Approved. Females follow the same trend

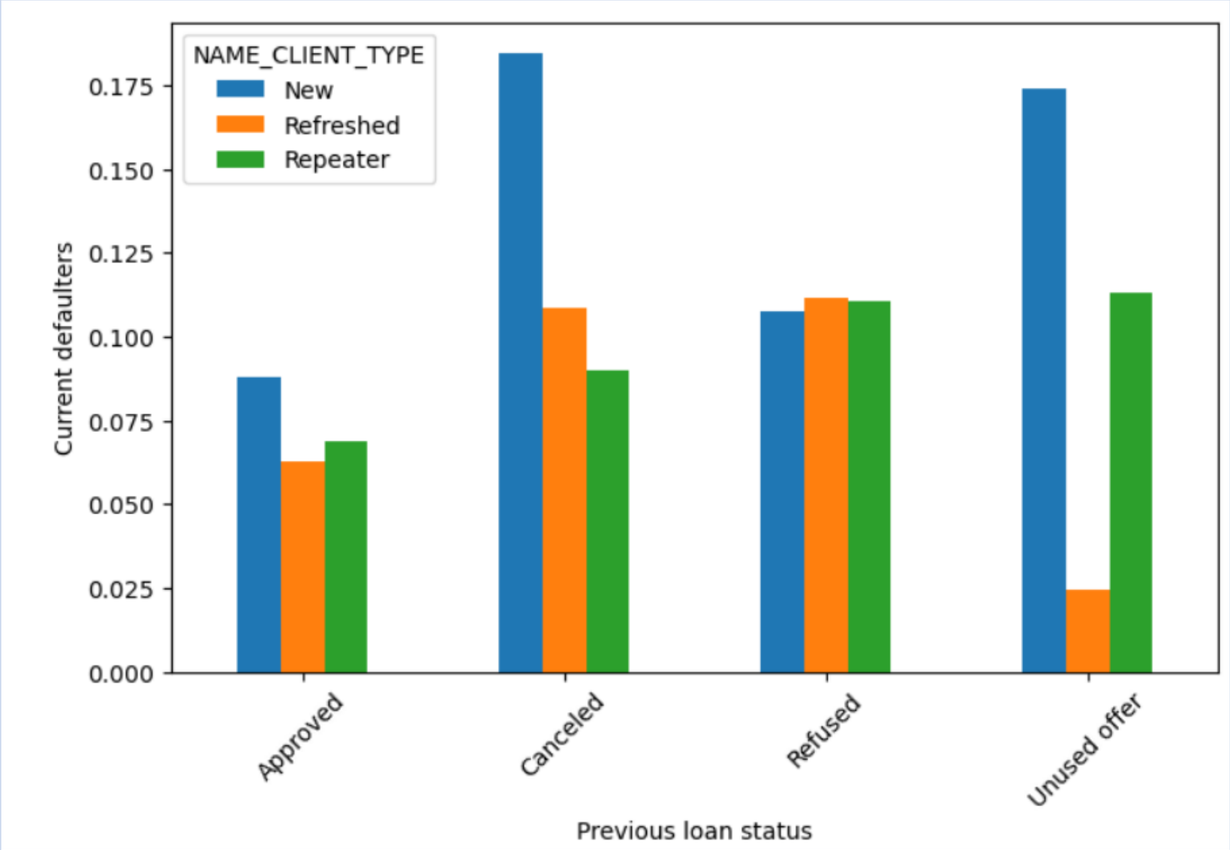
Analysis of two segmented variables - **Previous Loan Status & Client Type**

1)We can see that the Defaulters are more for Cancelled status who are New users

2)For Unused Offer status - the New clients were more defaulted followed by Repeater followed by Refreshed users.

3)For previously Refused applicants the Defaulters are more Refreshed clients.

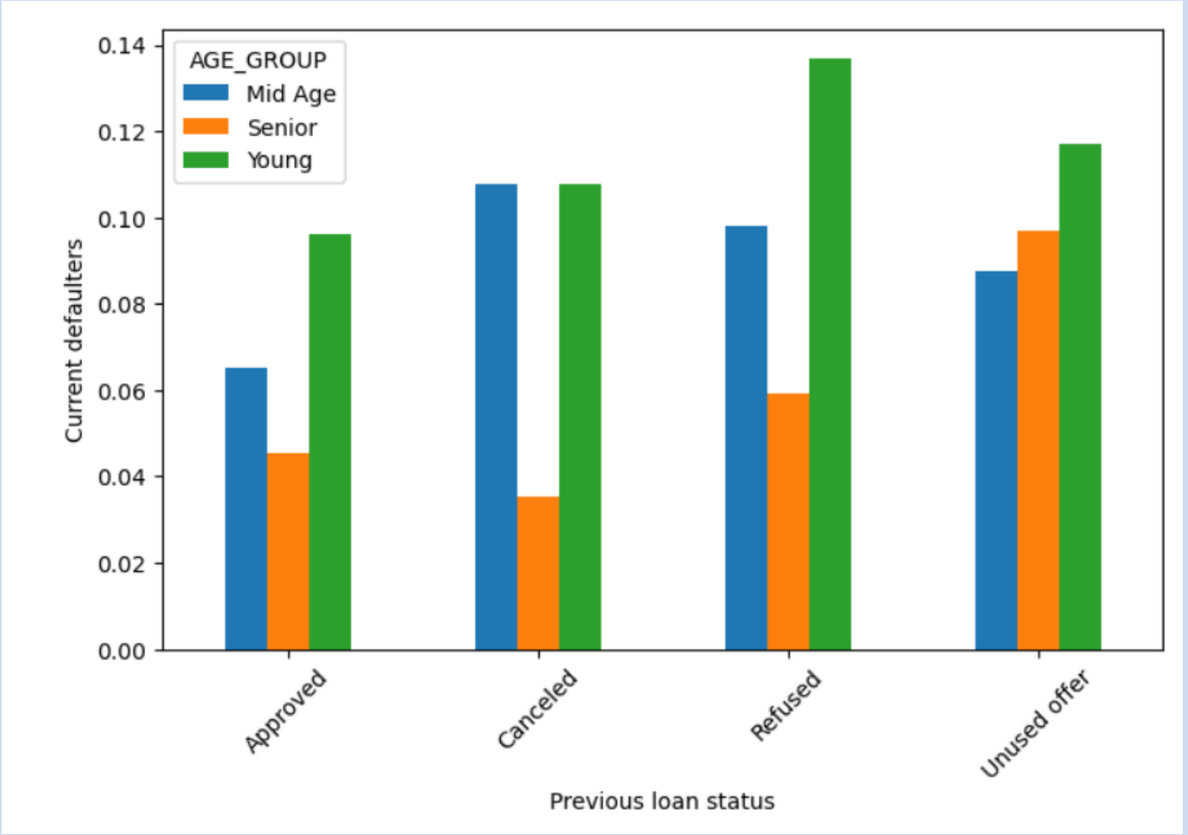
4)For previously Canceled applicants the Defaulters are more New clients.



Analysis of two segmented variables - **Previous Loan Status & Age Group**

1) For all the status Young applicants are more defaulted.

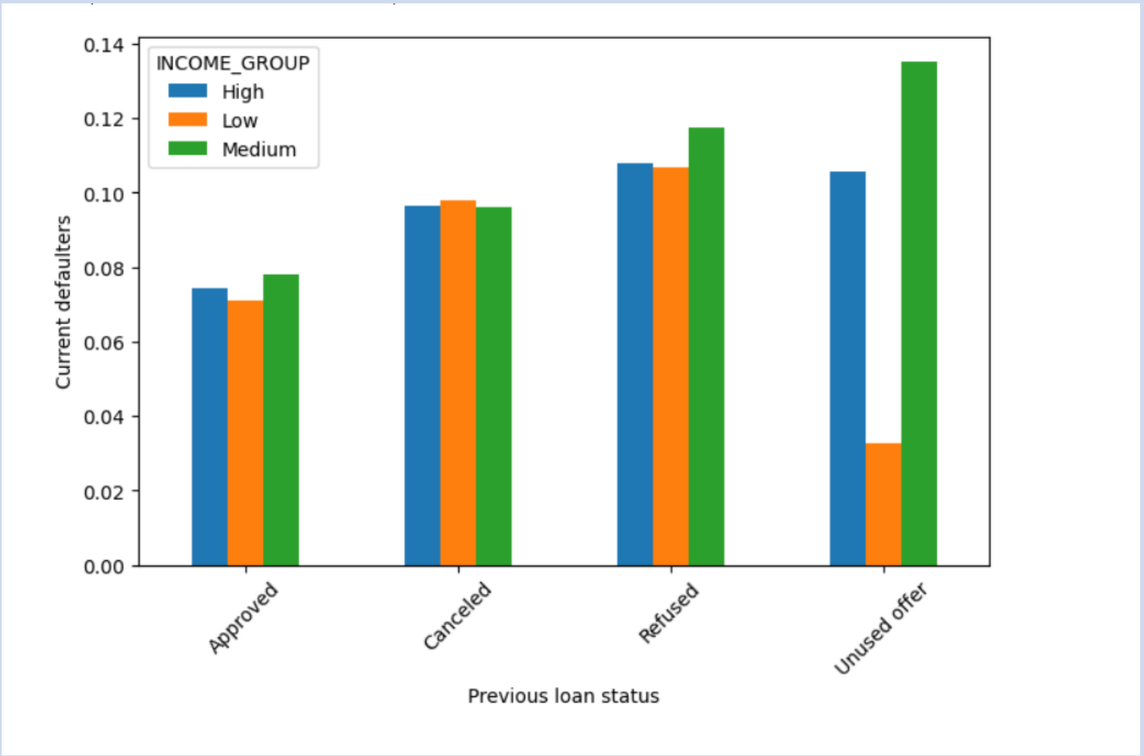
2) For all the previous status Senior applicants are less defaulted compared to others.



Analysis of two segmented variables - **Previous Loan Status & Income Group**

1) For previously Unused offer the Medium income group was more defaulted and Low income group is the least.

2) For other application status more or less all the income groups are equally defaulted.

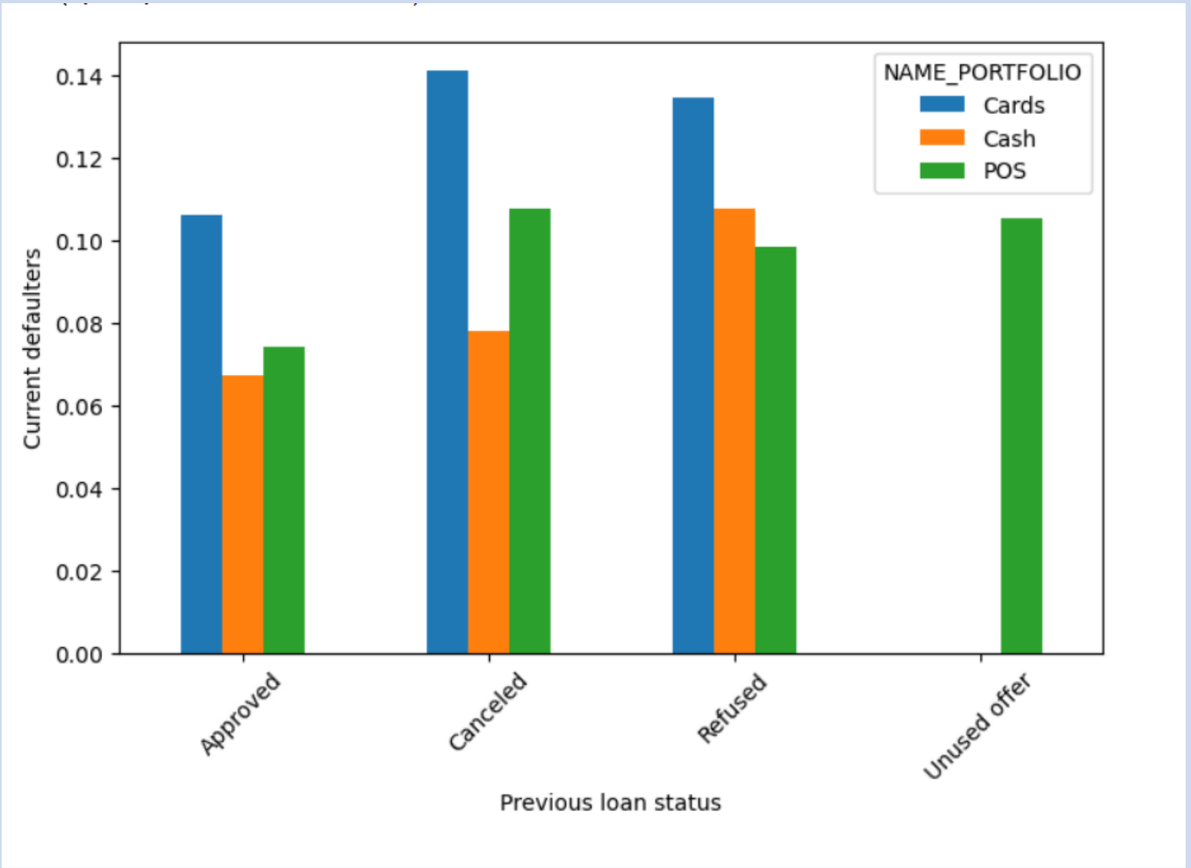


Analysis of two segmented variables - **Previous Loan Status & Name Portfolio**

1) Most of the clients were defaulted, who previously applied loan for Cards.

2) For approved loan status the clients applied for Cars are less defaulted.

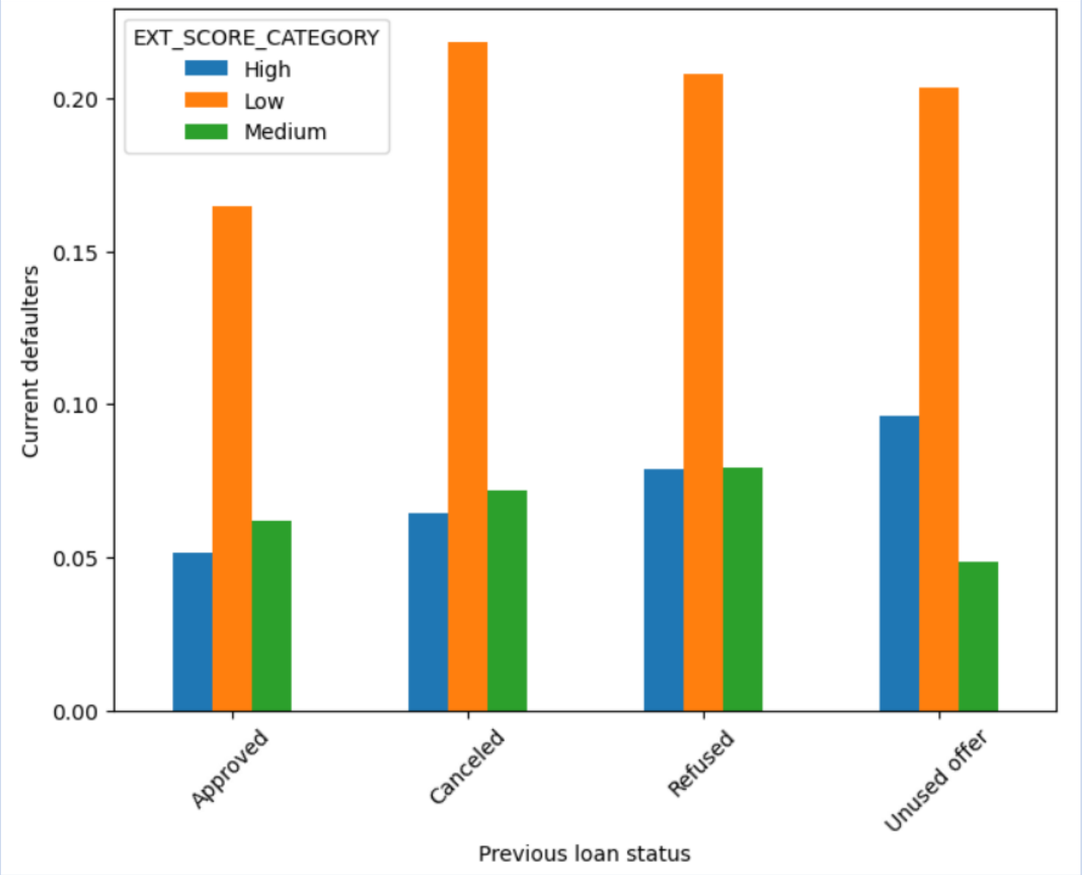
3) For Refused loan status the clients applied for POS are less defaulted.



Analysis of two segmented variables - **Previous Loan Status & Score Category**

1) Applicants with low external source score defaulted the most.

2) Higher scorer applicants are very unlikely to default irrespective of their previous loan status.



Sheza Waqar Beg

Thank you!

Email: shezabeg@gmail.com