Credit EDA Case Study

(Loan Defaulter)

Case Study Partners:

Nikhil Vilas Golambade / Yash Vijay Mhatre

Problem Statement

- **Introduction**

This case study aims to give us an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques of Exploratory Data Analysis (EDA), we will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

- **Business Understanding**

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants are capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

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

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
- **All other cases:** All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

1. **Approved**: The Company has approved loan Application
2. **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
3. **Refused**: The company had rejected the loan (because the client does not meet their requirements etc.).
4. **Unused offer**: Loan has been cancelled by the client but on different stages of the process.

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

Problem Statement

- **Business Objectives**

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

- **Data Understanding**

This dataset has 3 files as explained below:

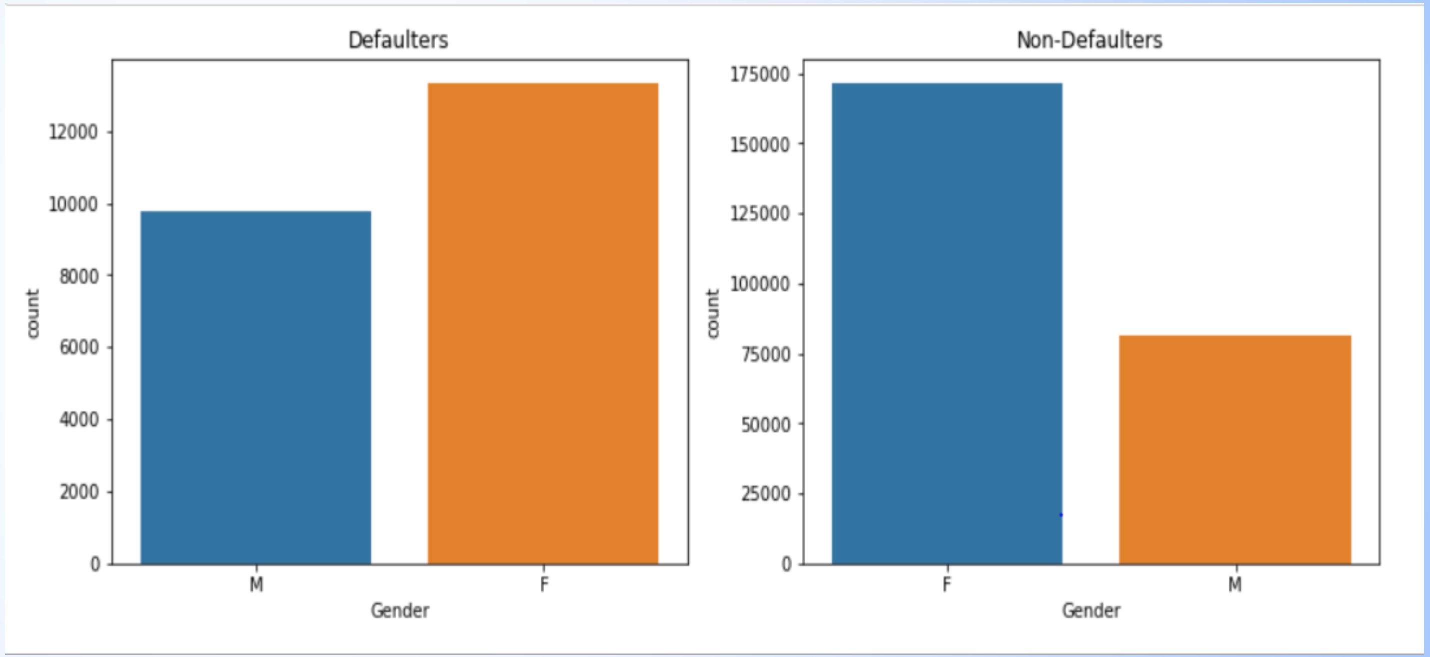
1. **application_data.csv** contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
2. **previous_application.csv** contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.
3. **columns_description.csv** is data dictionary which describes the meaning of the variables.

Current Applications

Univariate analysis for Unordered categorical variables

Count of defaulters and non-defaulters on the basis of gender

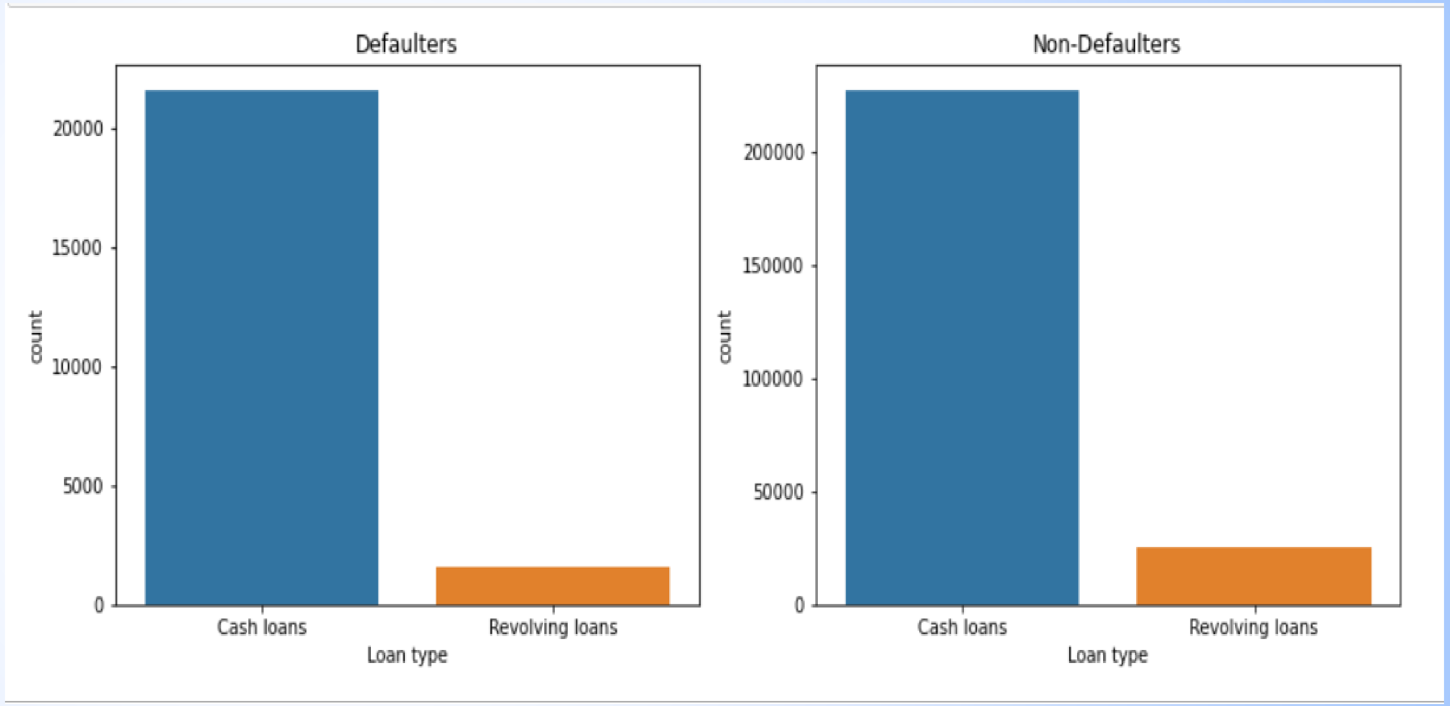
- Defaulters - We can see that females are slightly more in number of defaulters than male.
- Non-defaulters - The same pattern continues for non-defaulters as well. The females are more in number here than male.



Univariate analysis for Unordered categorical variables

Defaulters and non-defaulters on the basis of Loan type

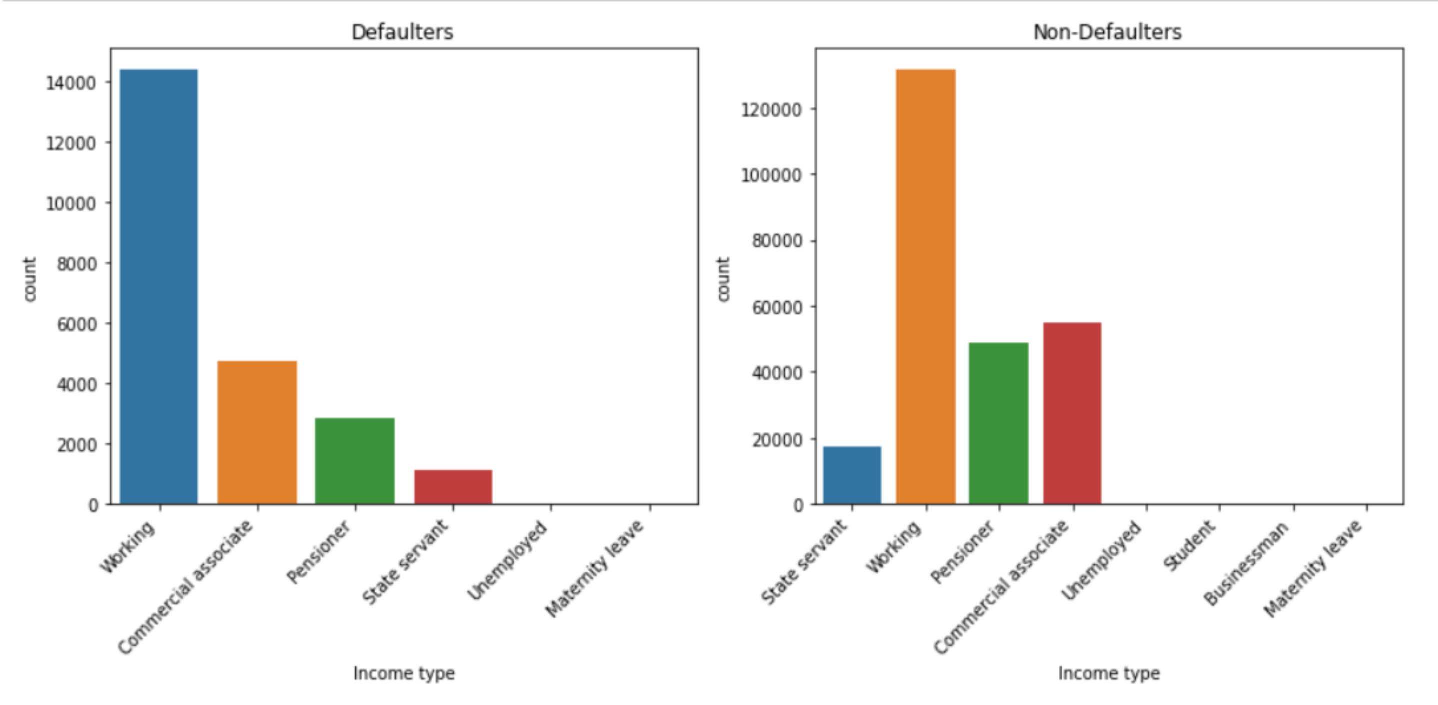
- We see in both the cases that Revolving loans are very less in number compared to Cash loans.



Univariate analysis for Unordered categorical variables

Defaulters and non-defaulters on the basis of Income type

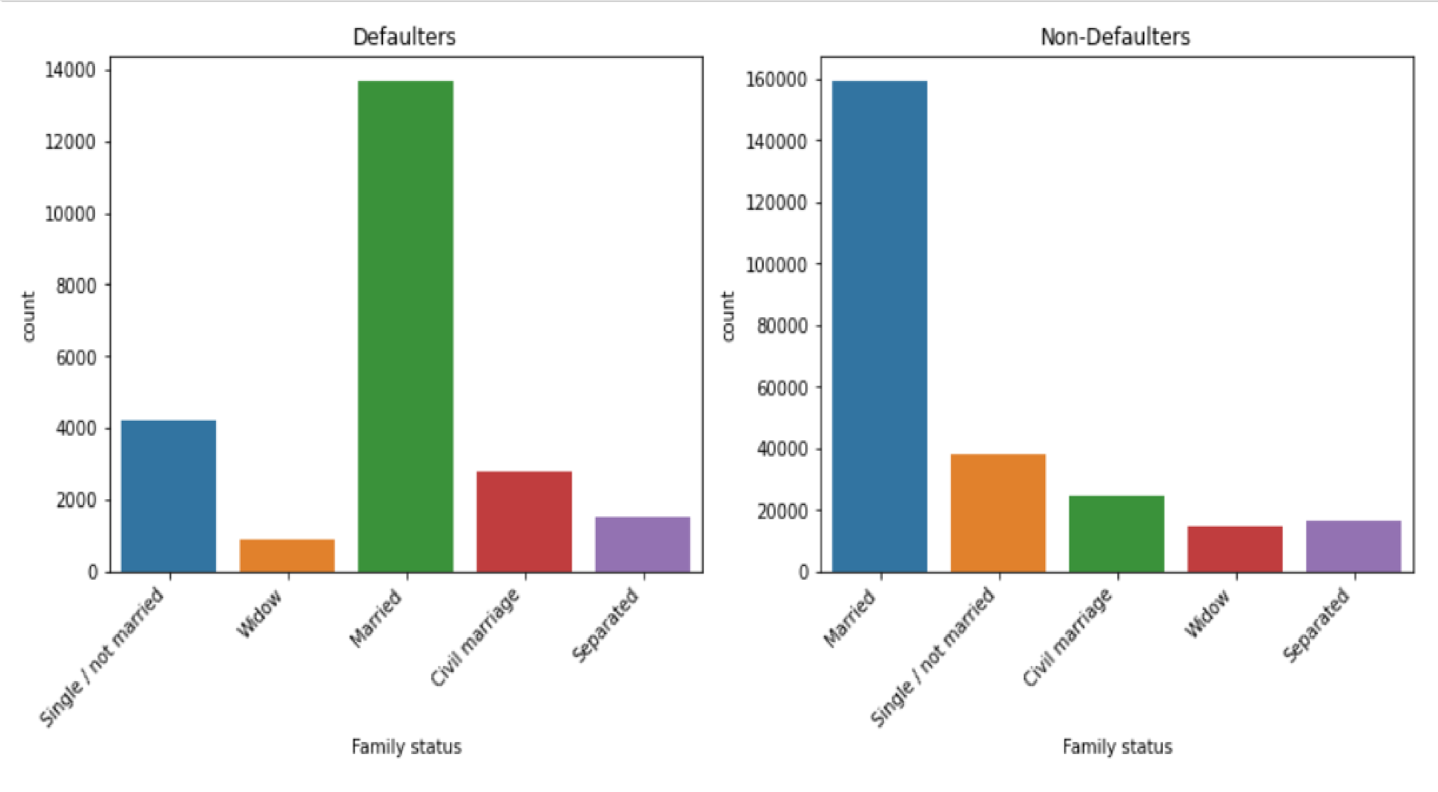
- Defaulters - Working people are mostly defaulted as their numbers are high with compare to other professions.
- Non-defaulters - Similarly here also working people are more in number who are not defaulted.



Univariate analysis for Unordered categorical variables

Defaulters and non-defaulters on the basis of Family status

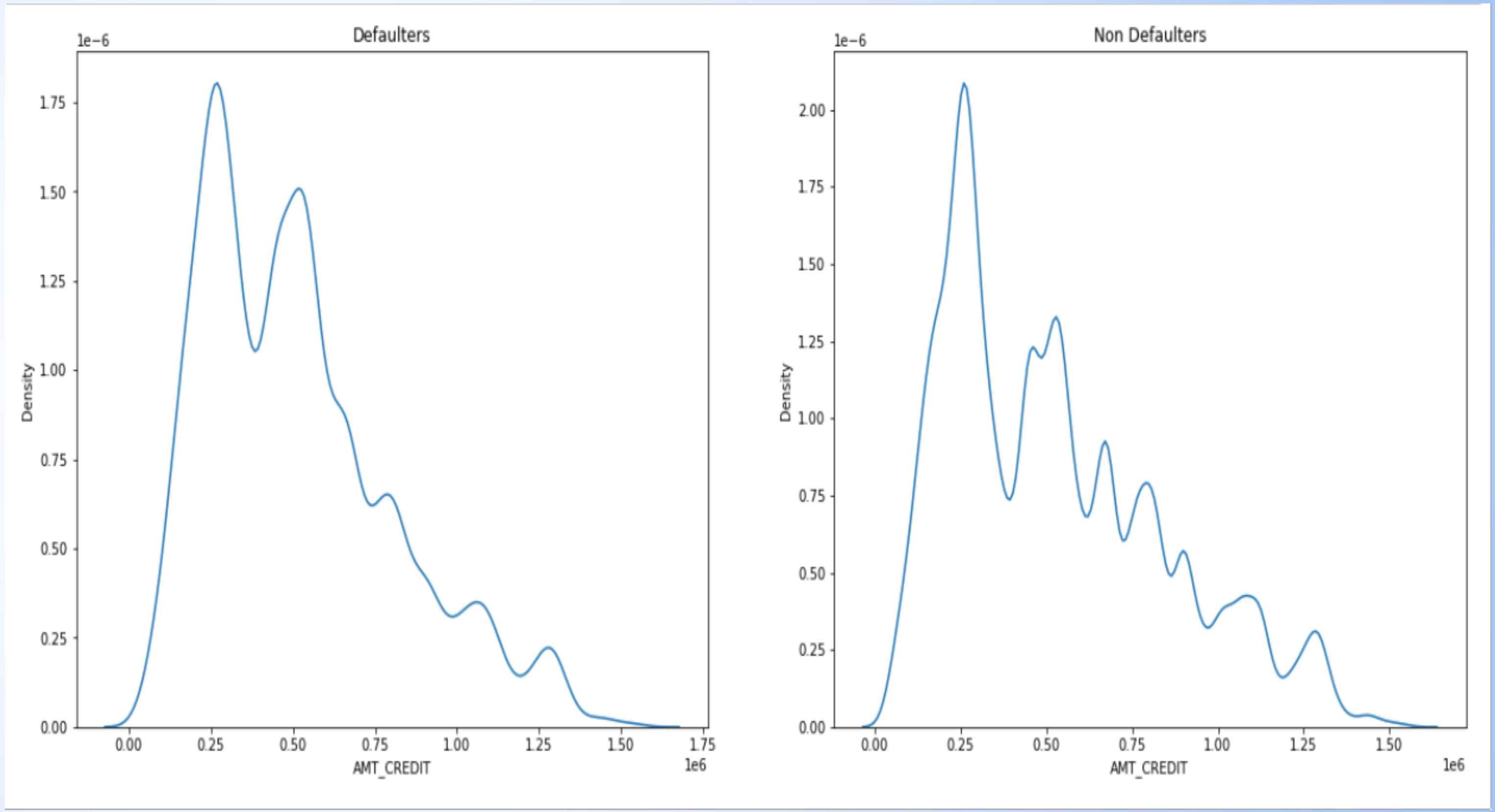
- For both the customers (defaulters and non-defaulters) married people are more in number compared with single, separated, widow etc.



Univariate analysis for continuous variables

Defaulters and non-defaulters on the basis of credit amount of the loan

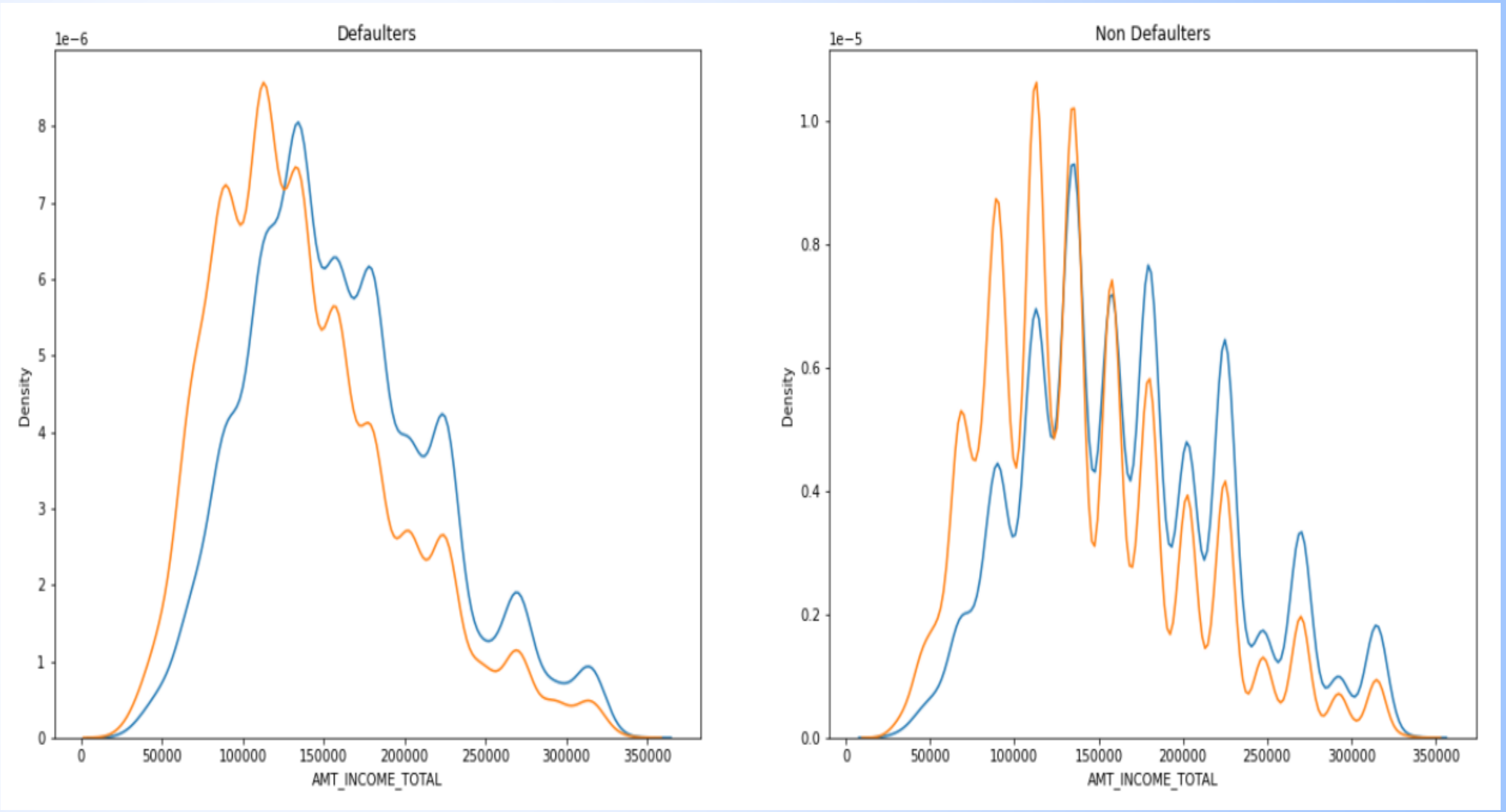
- Defaulters :- We can notice that the lesser the credit amount of the loan, the more chances of being defaulter. The spike is till 500000.
- Non defaulters :- If the credit amount is less, there is lesser chance of being defaulted. And gradually the chance is being decreased with the loan credit amount.



Univariate analysis for continuous variables

Defaulters and non-defaulters on the basis of gender and their total income

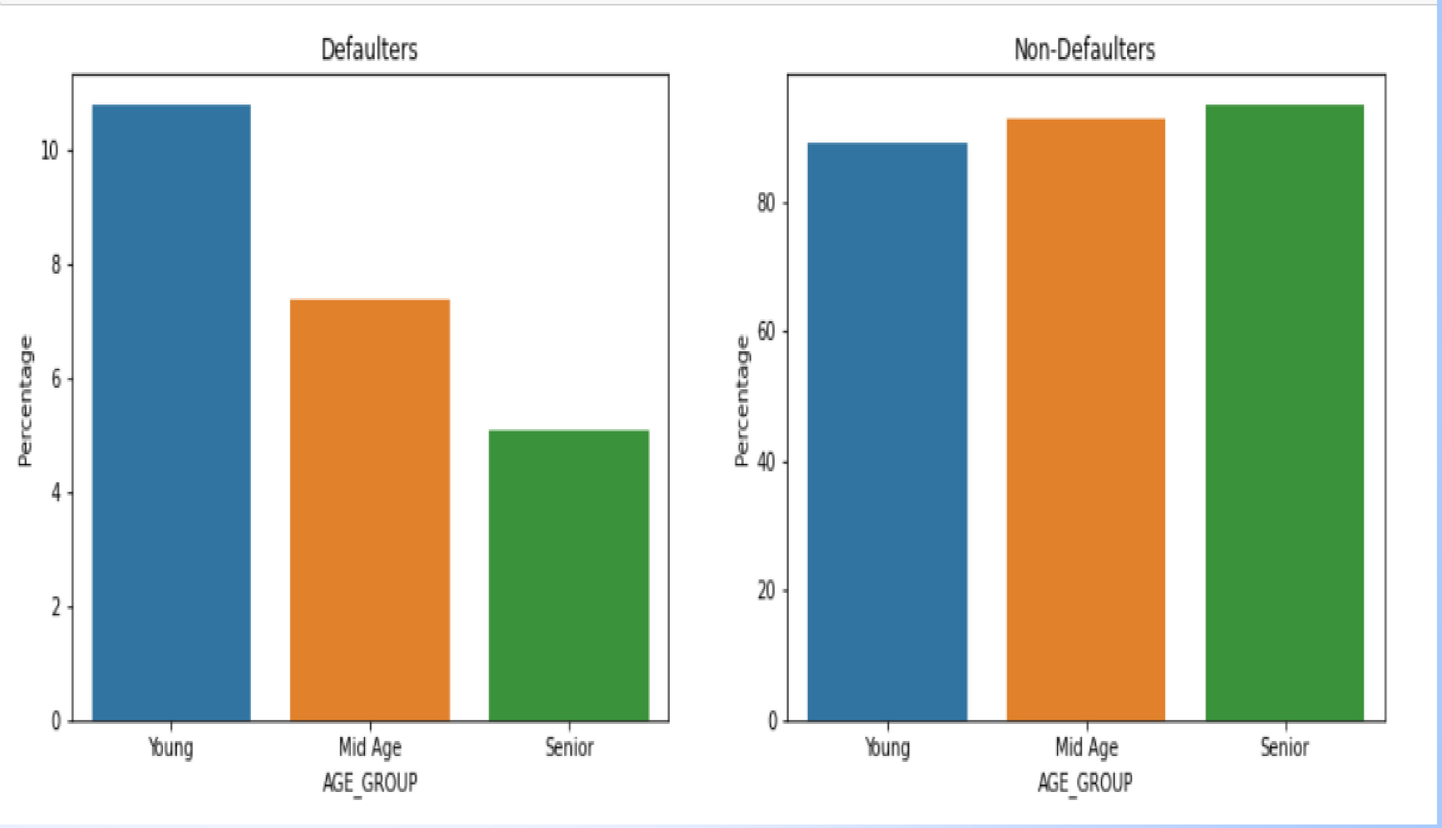
- Defaulters - We can notice by looking at the pattern that for being a defaulter both the genders (male and female) are almost equal in all income levels. The spike of being defaulters is from 50000 to 200000.
- Non defaulters - Here we see an interesting pattern. Females are more non defaulter on the lower income level but lesser non defaulter in higher income level. The spike is more for both the genders from 75000 to 150000.



Segmented Univariate analysis for ordered categorical variables

Percentage of age group applicants defaulted and not defaulted

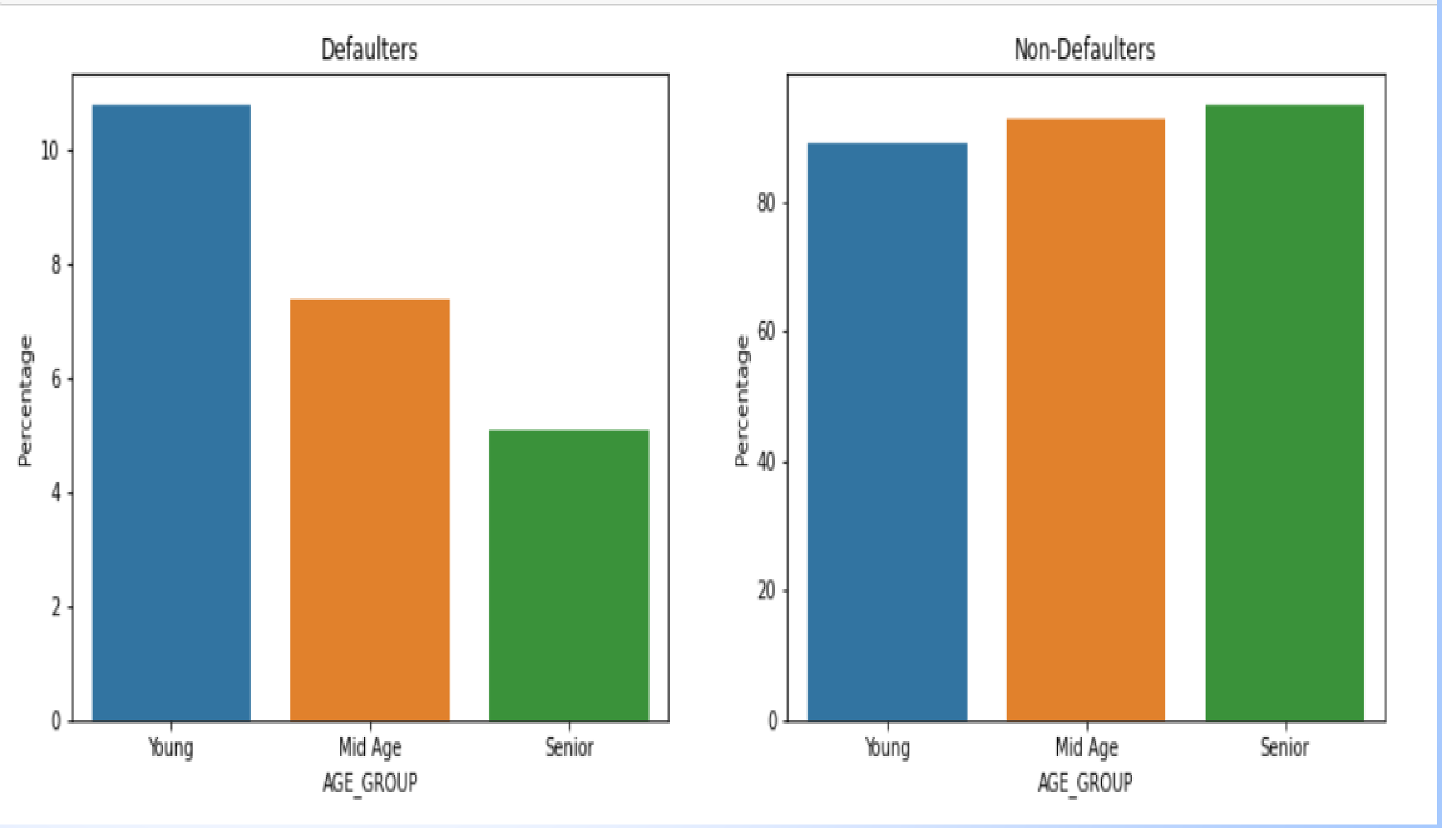
- The analysis below showed that the how much percentage of each age group(Young, Mid age and Senior citizen) applicants are defaulted and not defaulted.
- Defaulters - We see that Young people are more likely to default than other two age groups. Whereas, Senior citizens are less likely to default than others.
- Non defaulters - There is not much difference in the likelihood for non defaulters in the age groups.



Segmented Univariate analysis for ordered categorical variables

Credit amount group

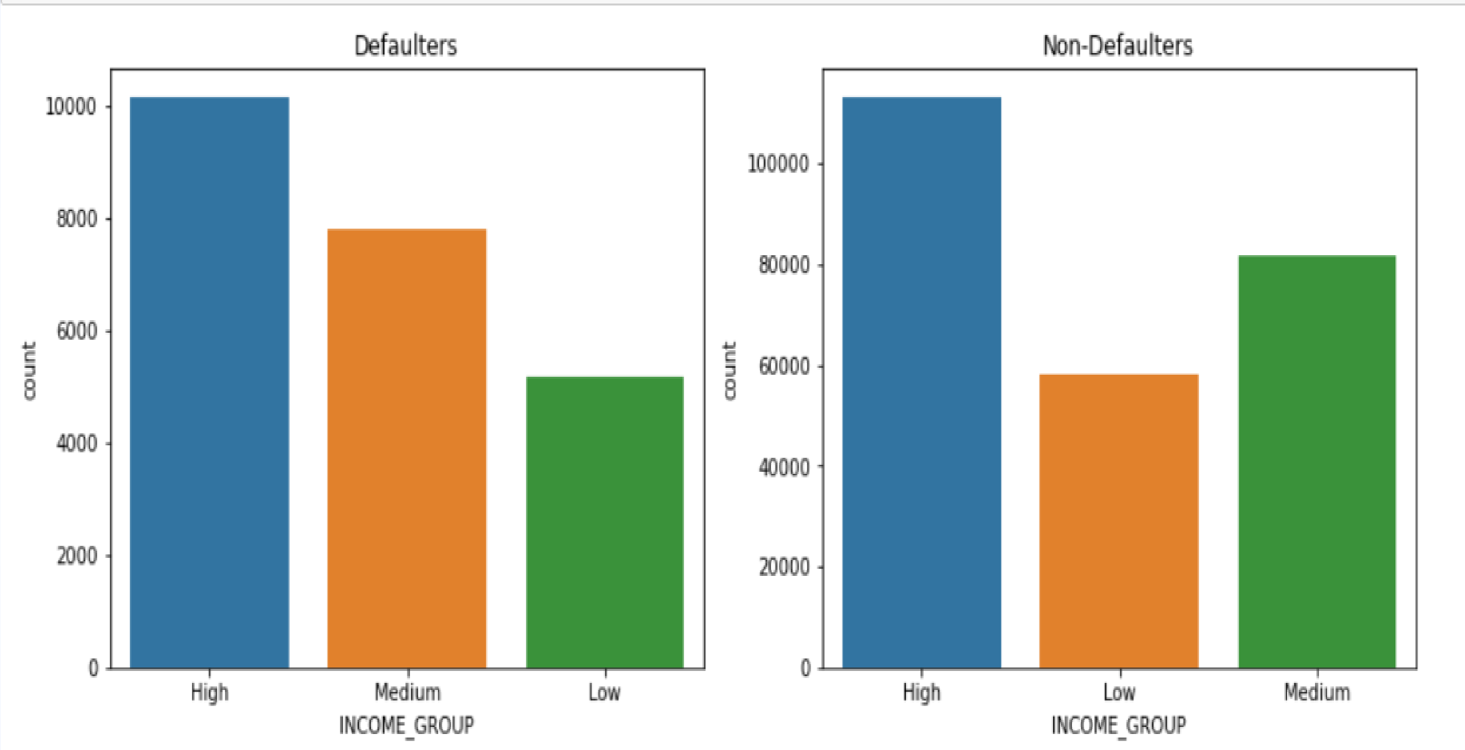
- Defaulters - Surprisingly low credited amount groups are more defaulters.
- Non defaulters - As expected low credit amount groups are more in number, who were not defaulted.



Segmented Univariate analysis for ordered categorical variables

Income group

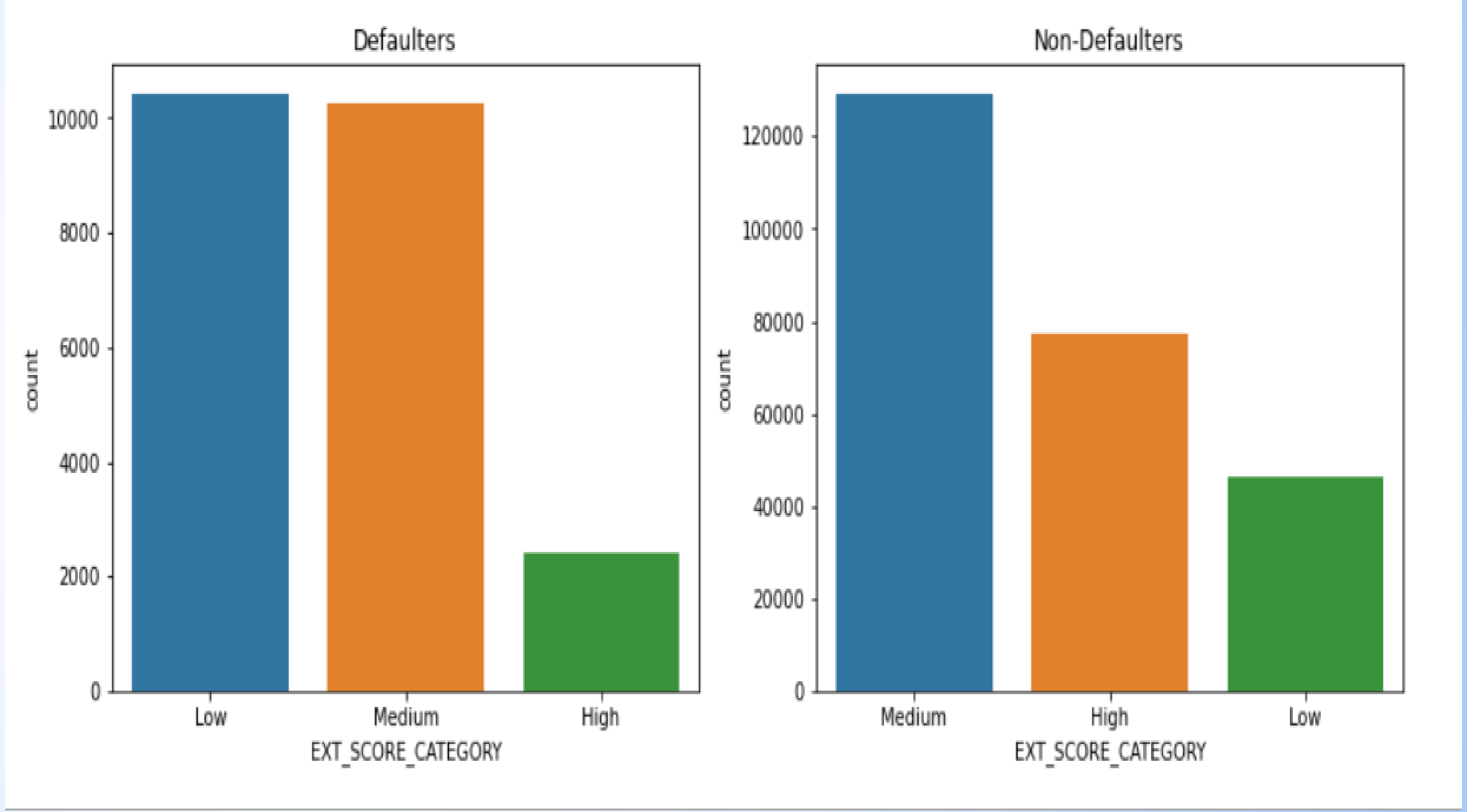
- Defaulters - Surprisingly the High income group is more in number to be defaulted, then Medium and then Low.
- Non defaulters - Here as expected the count of non defaulters more in High income group and less in low income group.



Segmented Univariate analysis for ordered categorical variables

Normalized score from external data source

- Defaulters - No surprise that low scorer from external data source are more defaulters. Also, the medium scorer are as likely defaulter as low scorer.
- Non defaulters - Medium scorers are no more defaulted than High scorer. As expected the Low scorers are lesser in number.

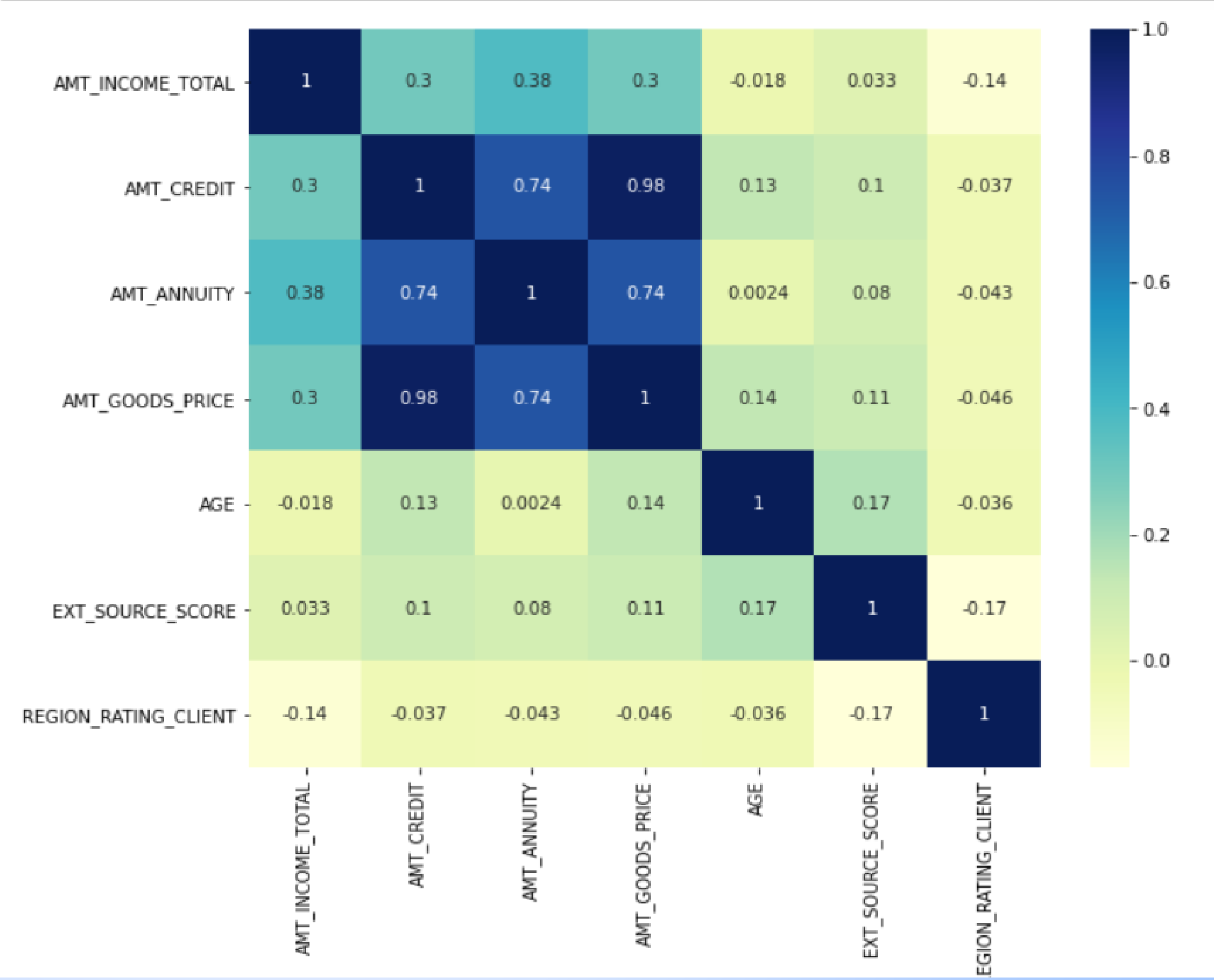


Bivariate analysis

Correlation of defaulters

Highly corelate columns for defaulters

- AMT_CREDIT and AMT_ANNUITY (0.74)
- AMT_CREDIT and AMT_GOODS_PRICE (0.98)
- AMT_ANNUITY and AMT_GOODS_PRICE (0.74)



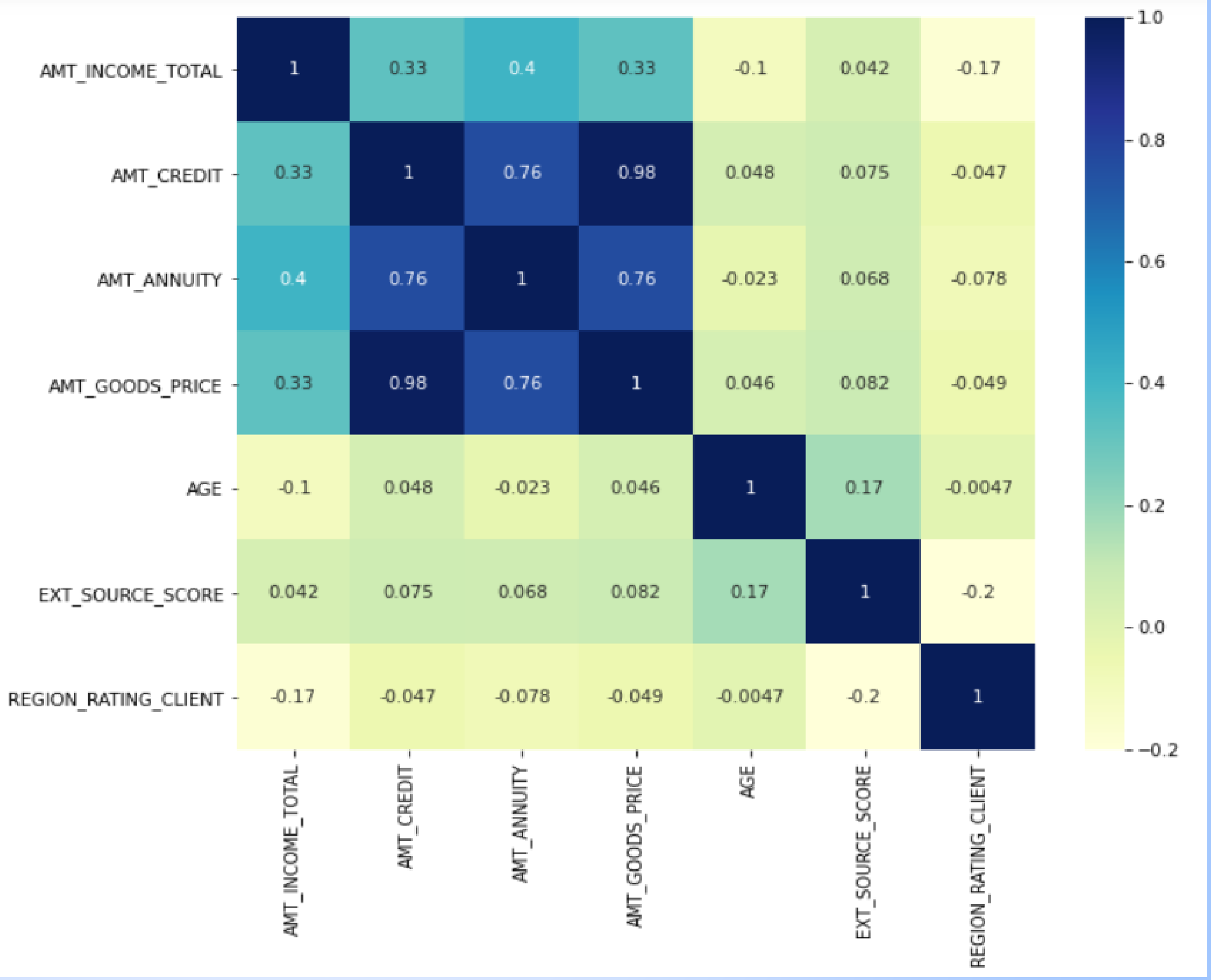
Bivariate analysis

Correlation of Non- defaulters

Highly corelate columns for defaulters

- AMT_CREDIT and AMT_ANNUITY (0.76)
- AMT_CREDIT and AMT_GOODS_PRICE (0.98)
- AMT_ANNUITY and AMT_GOODS_PRICE (0.76)

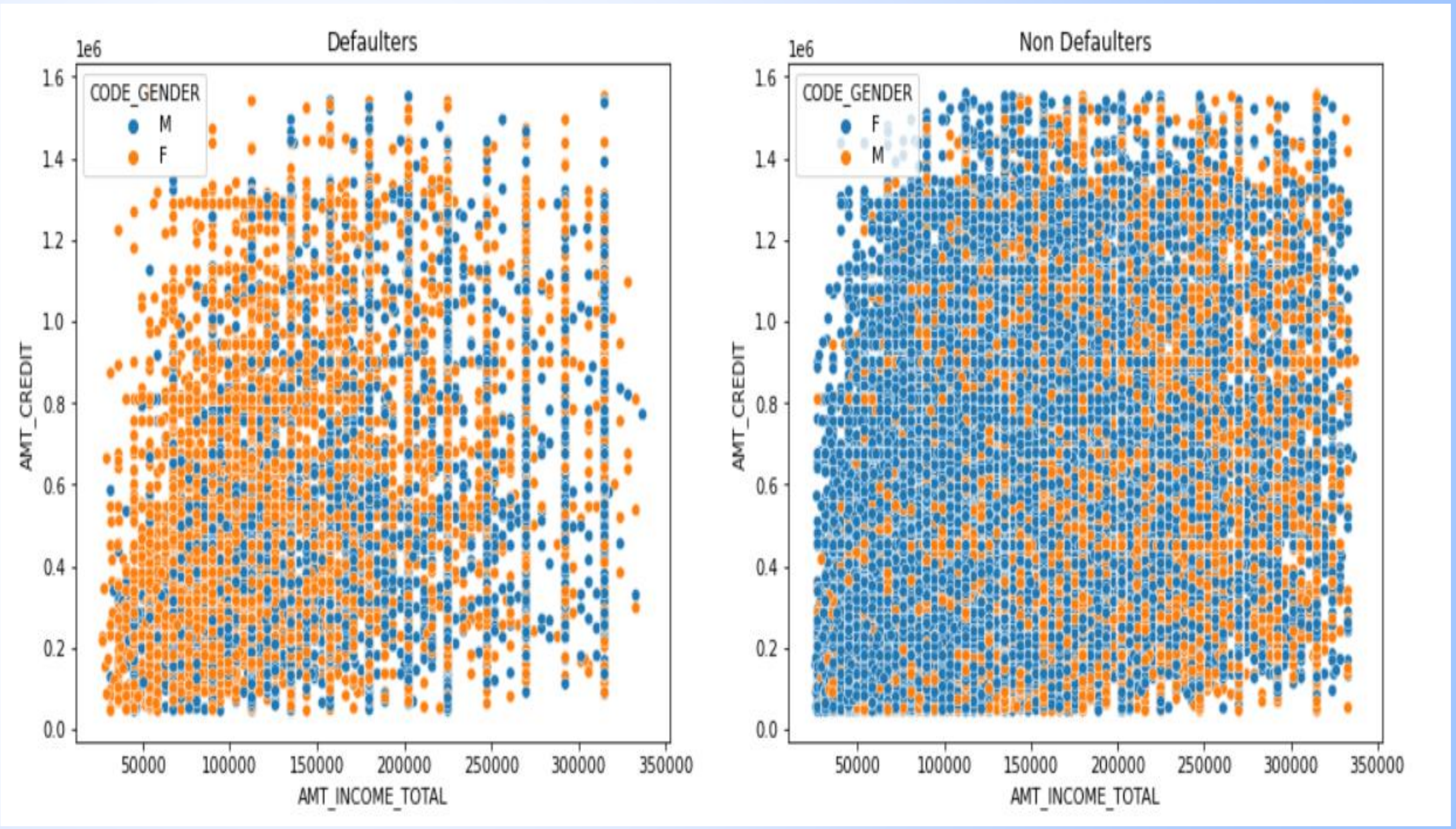
We can see that for both defaulters and non defaulters the same pairs of columns are highly corelated.



Bivariate analysis on continuous variable

Credit amount of the loan on the basis of client income for both male and female

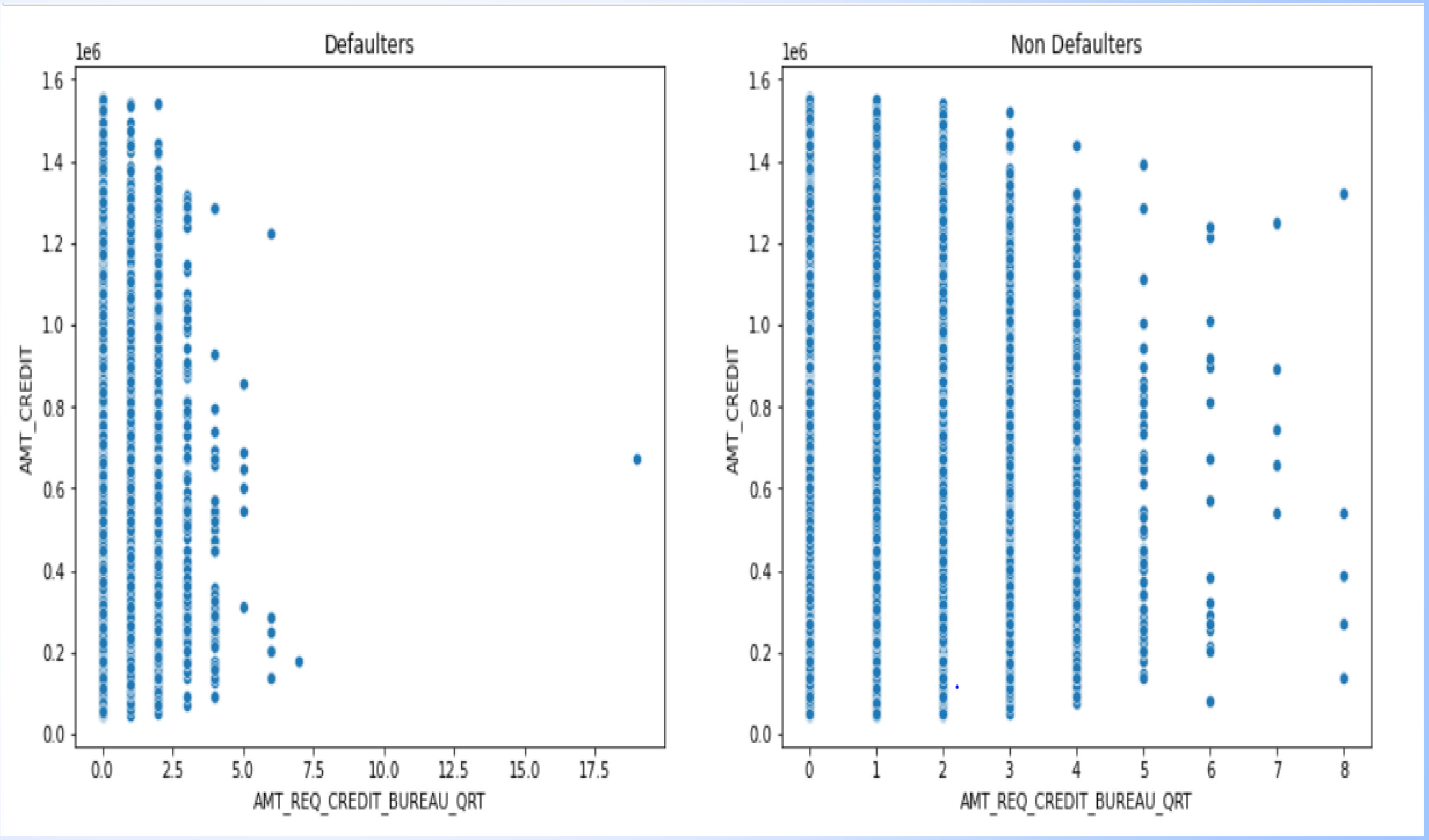
- Defaulters - We can slightly figure out that the values are more concentrated on the lower income and lower credit of the loan. That means as the income is increased, the amount of loan is also increased. This is true for both the genders.
- Non defaulters - We can hardly figure out any pattern out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



Bivariate analysis on continuous variable

Credit amount of the loan on the basis of Number of enquiries to Credit Bureau about the client

- We see that the more number of enquiries the lesser the amount of loan credited for both defaulters and non defaulters.

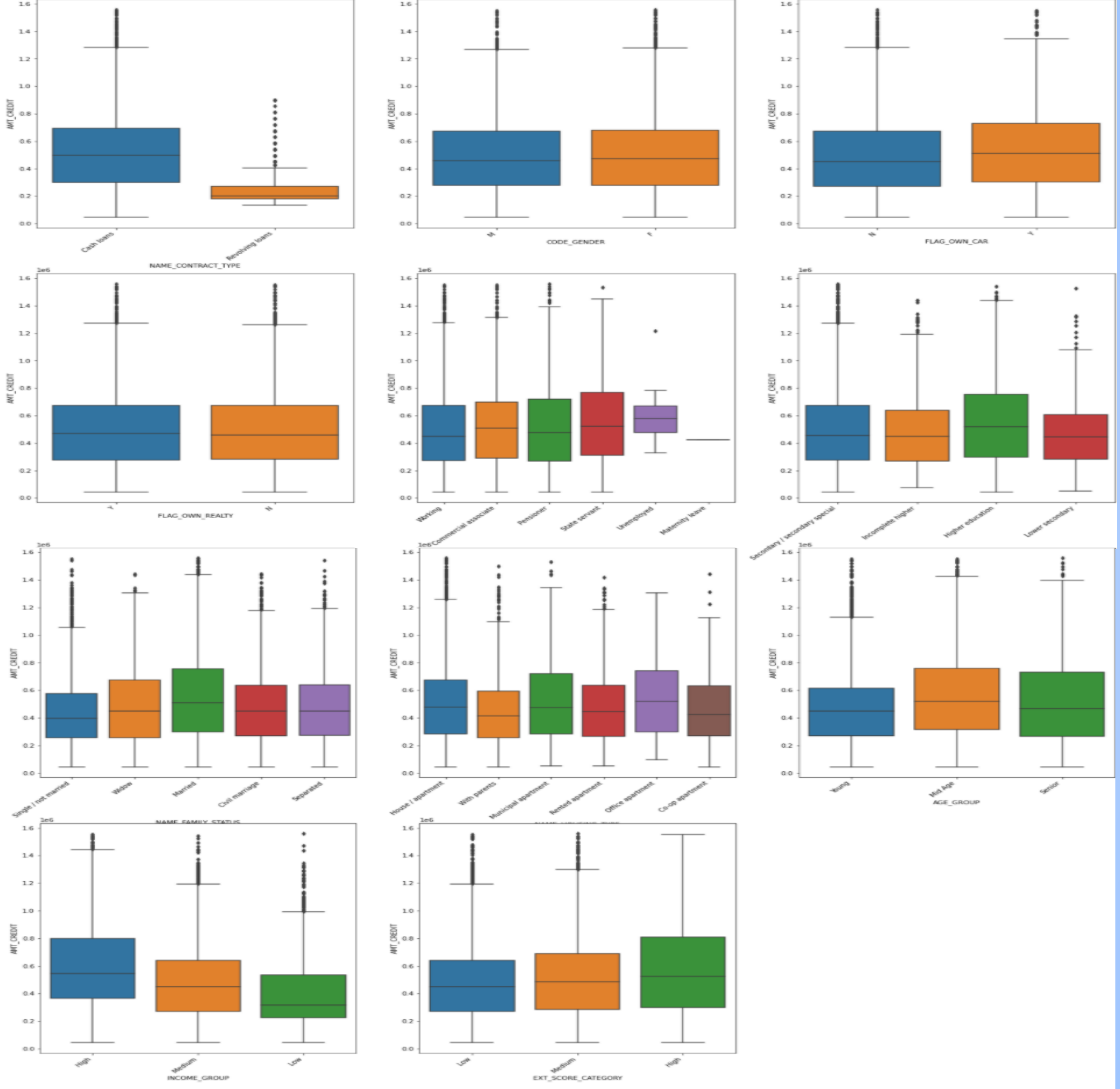


Bivariate analysis on categorical variable

Credit amount of the loan of various categories

Defaulter:

- Credit amount of the loans are very low for Revolving loans
- There is no credit amount difference between genders, client owning cars or realty.
- The Young age group got less amount of loan credited compared to mid age and senior citizen.
- Higher income group have more loan amount credited.
- Clients having higher external score have more loan amount.

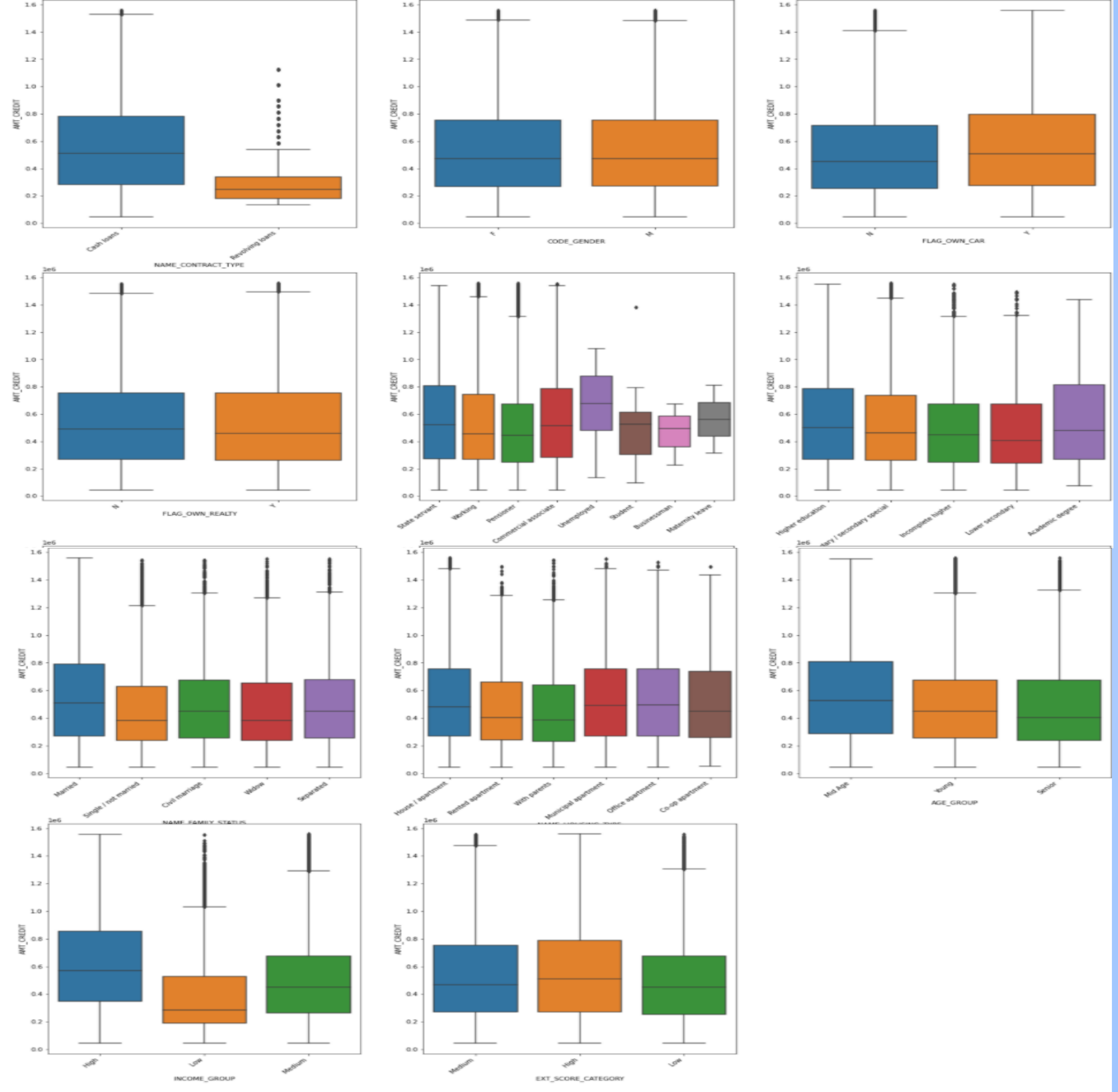


Bivariate analysis on categorical variable

Credit amount of the loan of various categories

Non-Defaulter:

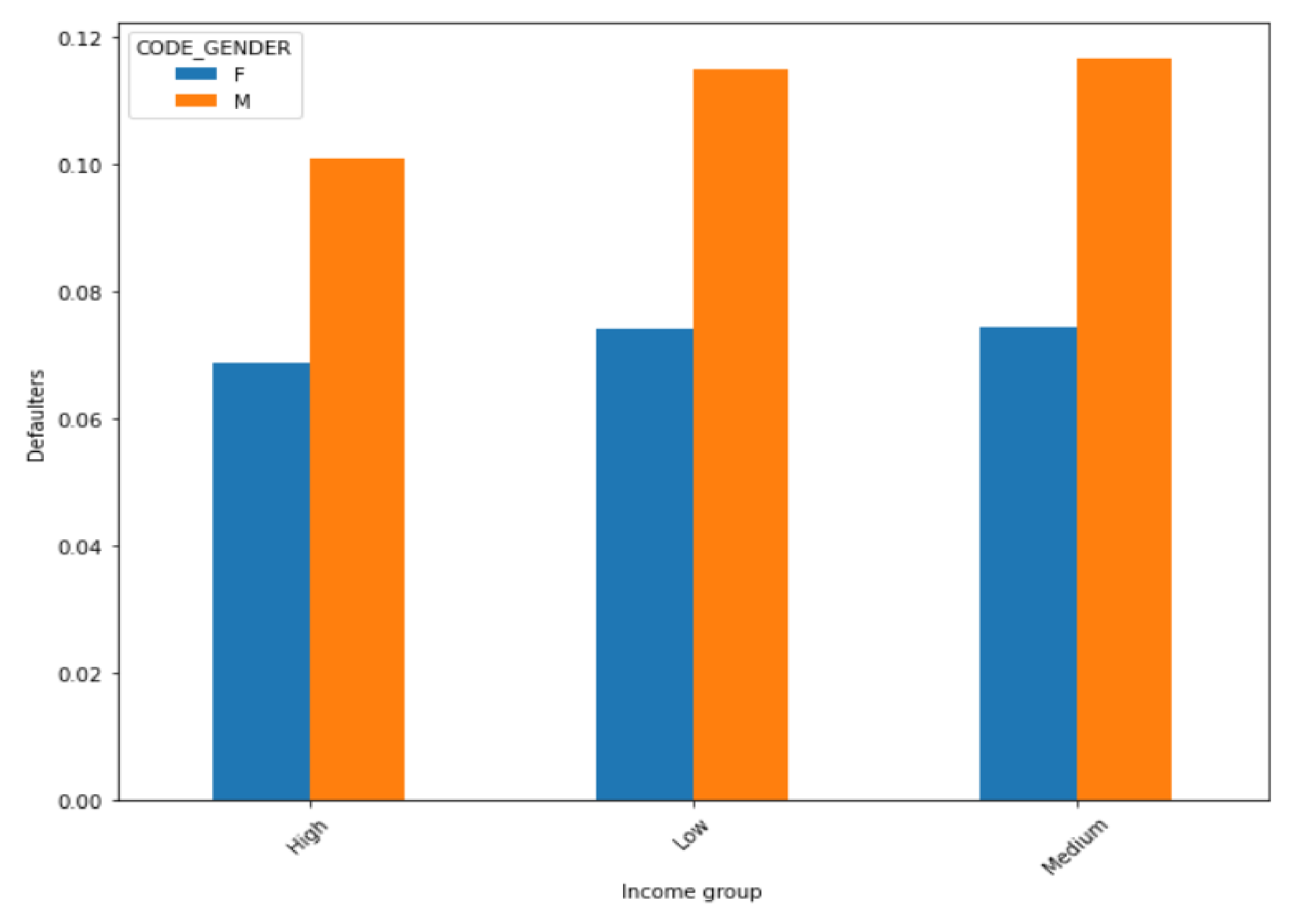
- Credit amount of the loans are very low for Revolving loans
- There is no credit amount difference between genders, client owning cars or realty.
- The mid age group got more amount of loan credited compared to young and senior citizen.
- Higher income group have more loan amount credited and lower the lowest.
- Clients having higher external score have more loan amount.
- Surprisingly the unemployed people have spike in credit amount of loan
- The Married people have more loan amount credited.



Analysis of two segmented variables

Income group and gender

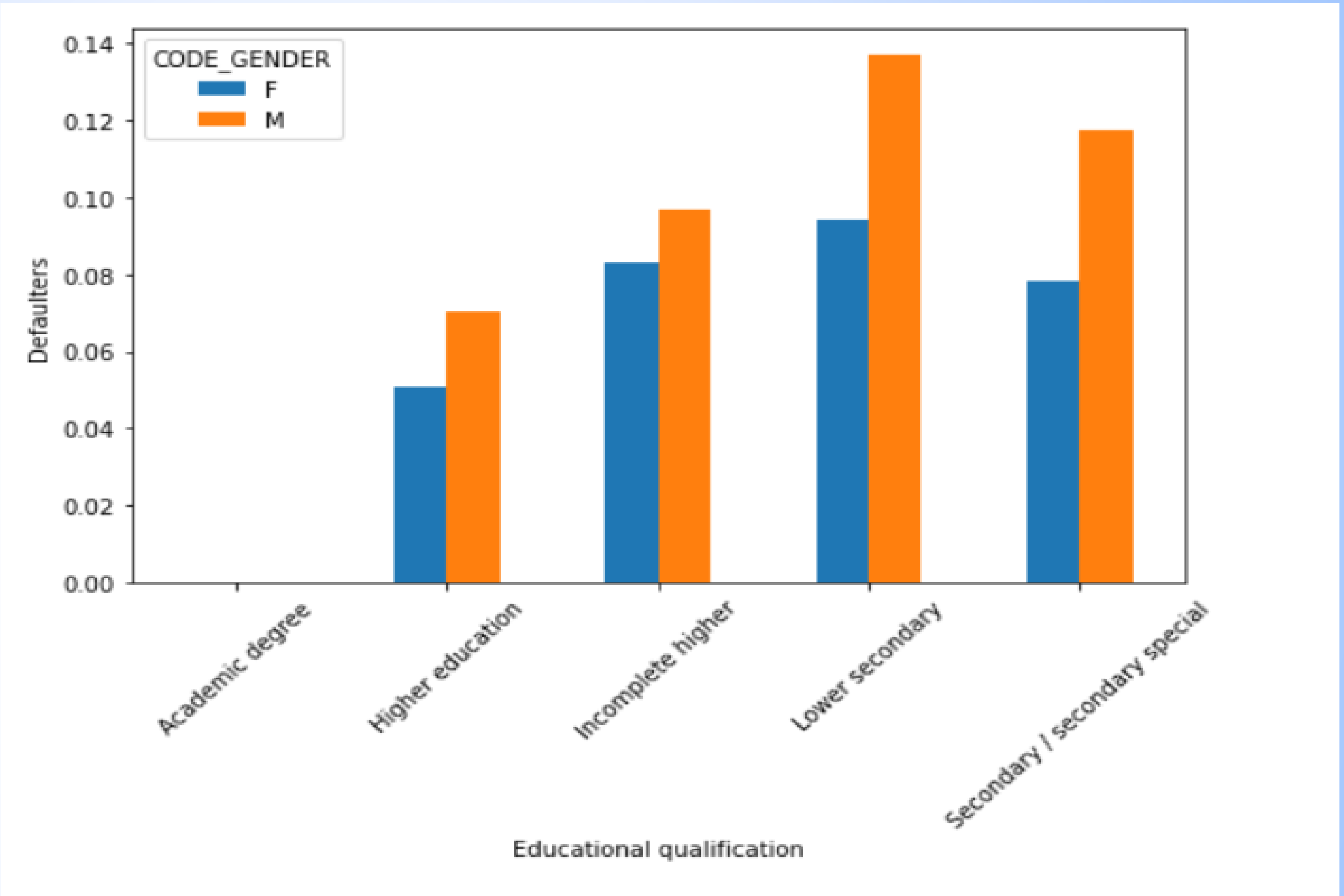
- We can see that Males are more likely defaulted than Females across all income groups.



Analysis of two segmented variables

Education and gender

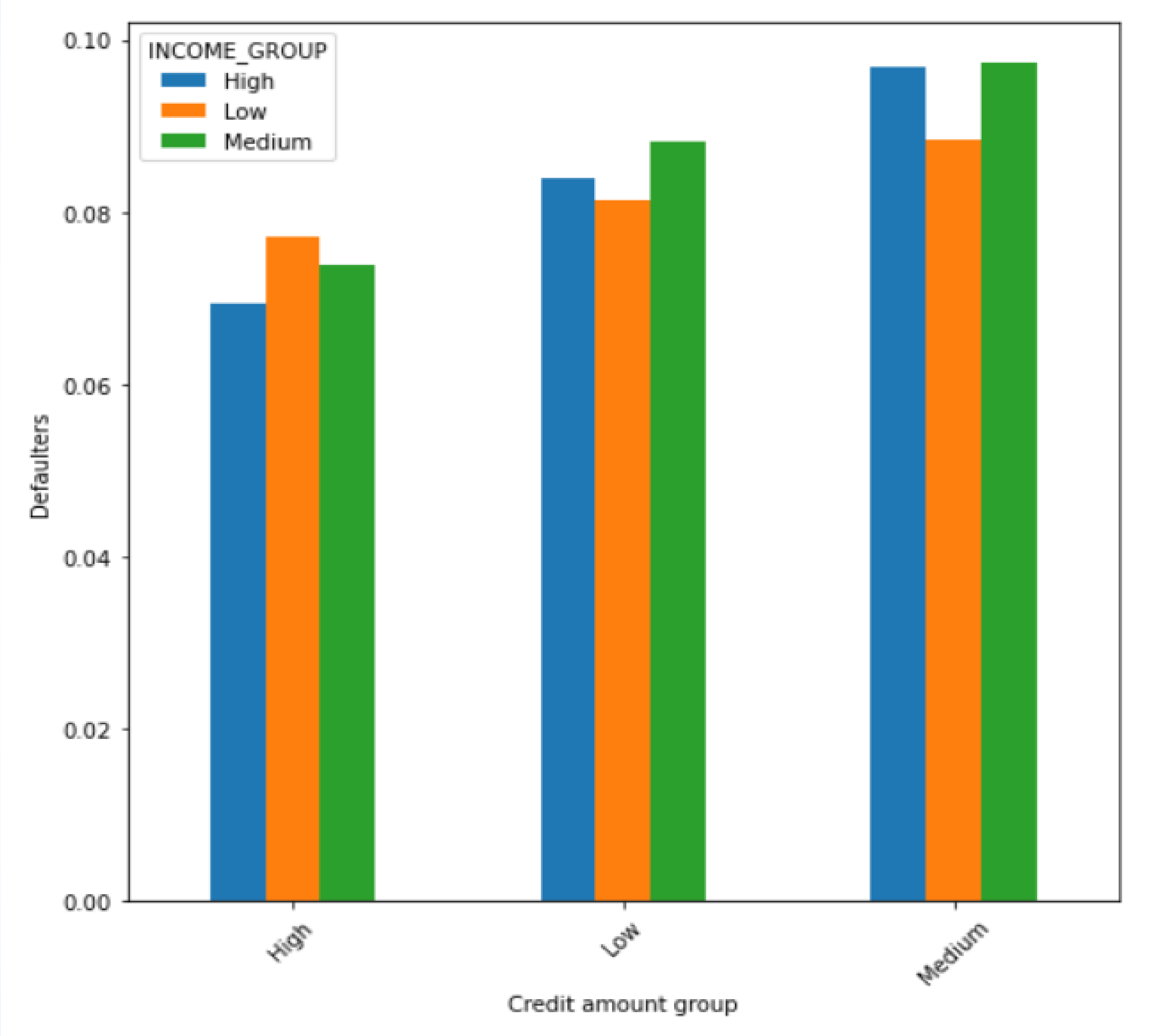
- Lower secondary educated clients are more defaulted followed by Secondary and Incomplete higher educated clients.
- The Higher educated people are less defaulted.
- Across all educated level Females are less defaulted than male.



Analysis of two segmented variables

Credit amount group and Income group

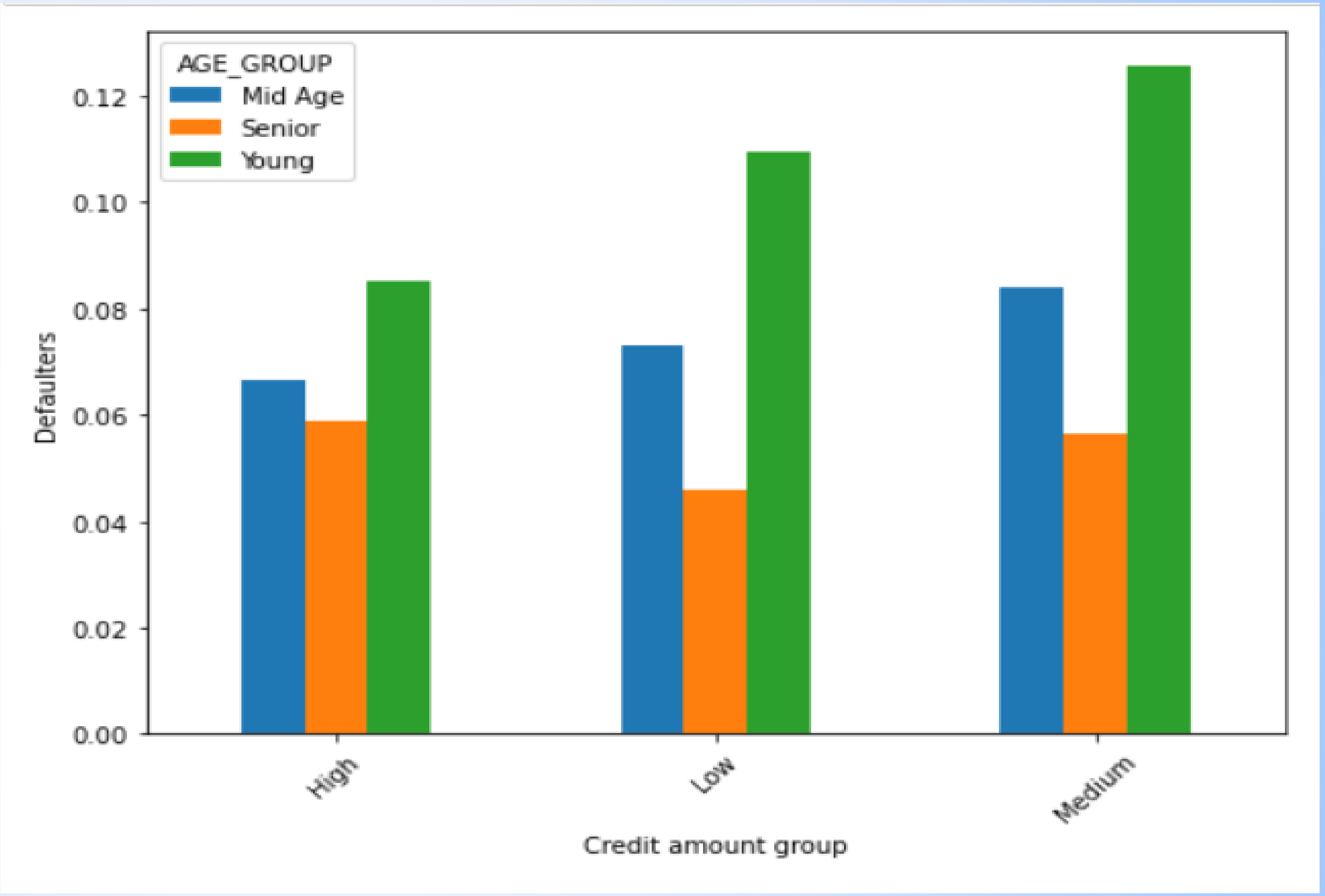
- Medium credit amount group are highly defaulted in all income groups.
- High credit amount groups are less likely to default in all income groups.



Analysis of two segmented variables

Credit amount group and Age group

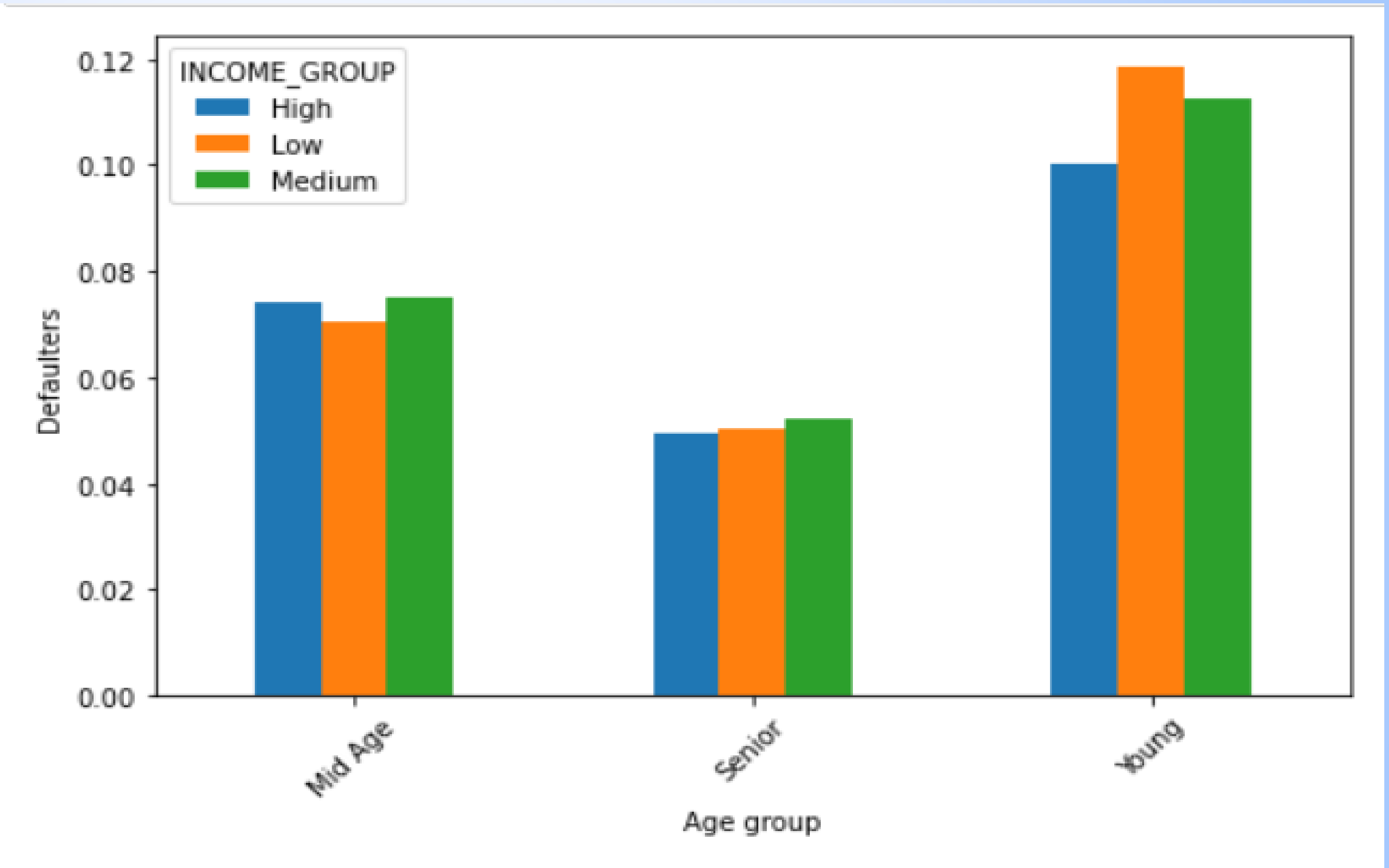
- Young clients with medium and low credit amount group are highly defaulted.
- Senior citizens across all credit amount groups are less likely defaulted.



Analysis of two segmented variables

Age group and Income group

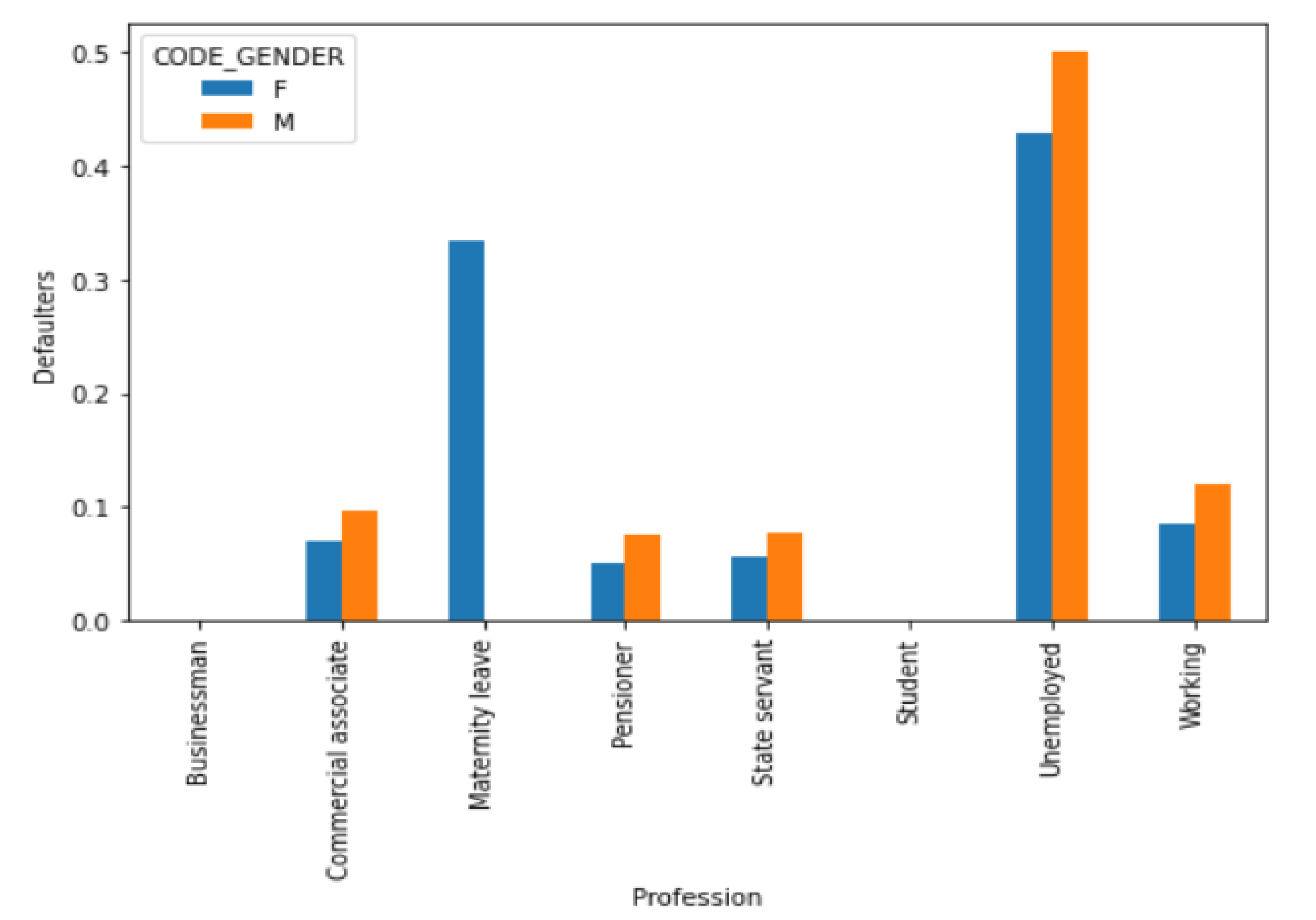
- Young clients are more defaulted than Mid age and senior.
- Young low income people are more defaulted.
- For Mid age and senior people the default rate is almost same in all income group.



Analysis of two segmented variables

Profession and Gender

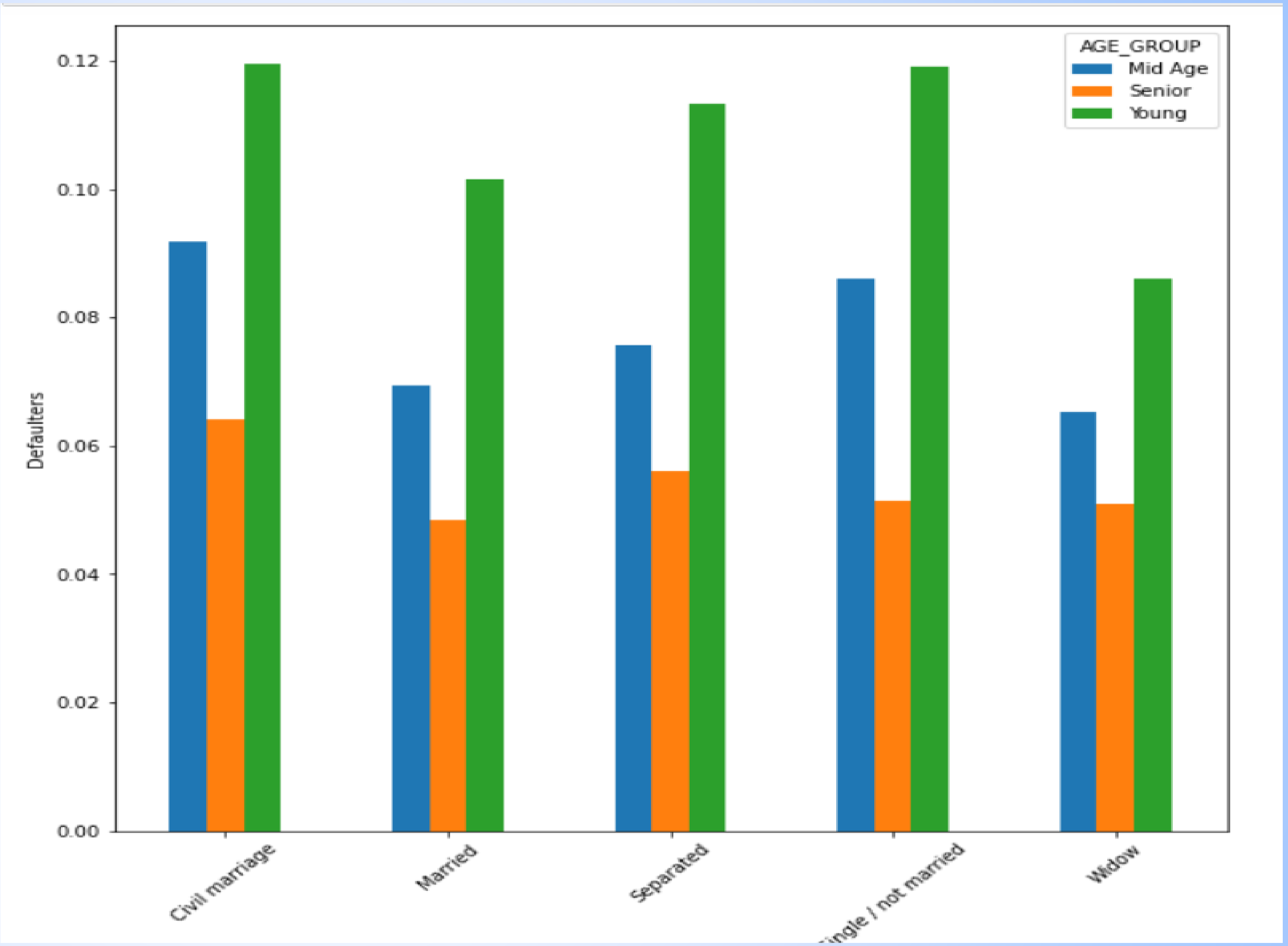
- No surprise the unemployed clients are more defaulted.
- Clients with maternity leave are expected to be defaulted more.
- The default rate is lesser in all other professions.
- Males are more defaulted with their respective professions compared to females.



Analysis of two segmented variables

Family status and age group

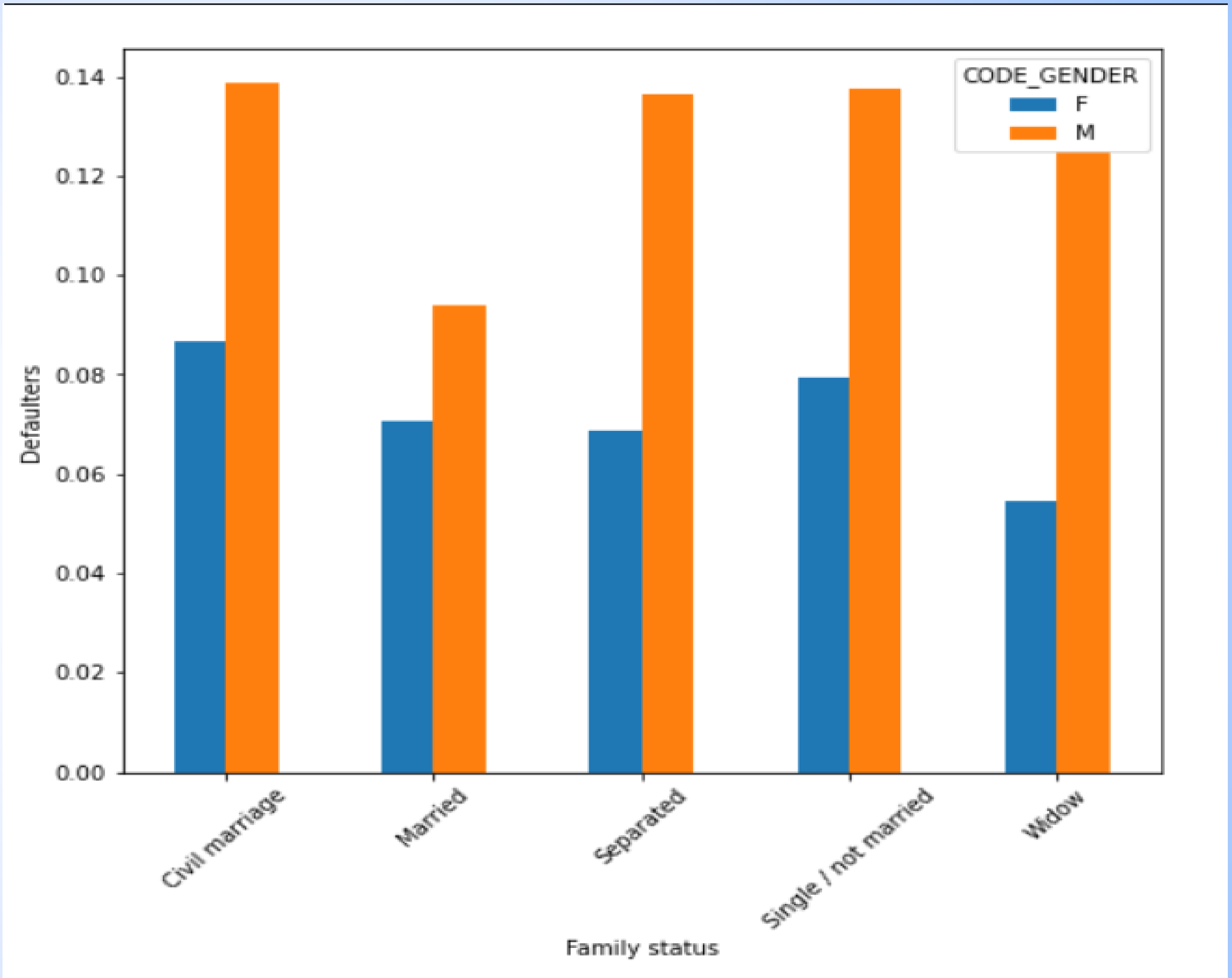
- Across all family status the Young clients are more defaulted and Senior citizen are less.



Analysis of two segmented variables

Family status and gender

- Across all family status the Male clients are more defaulted than Female.



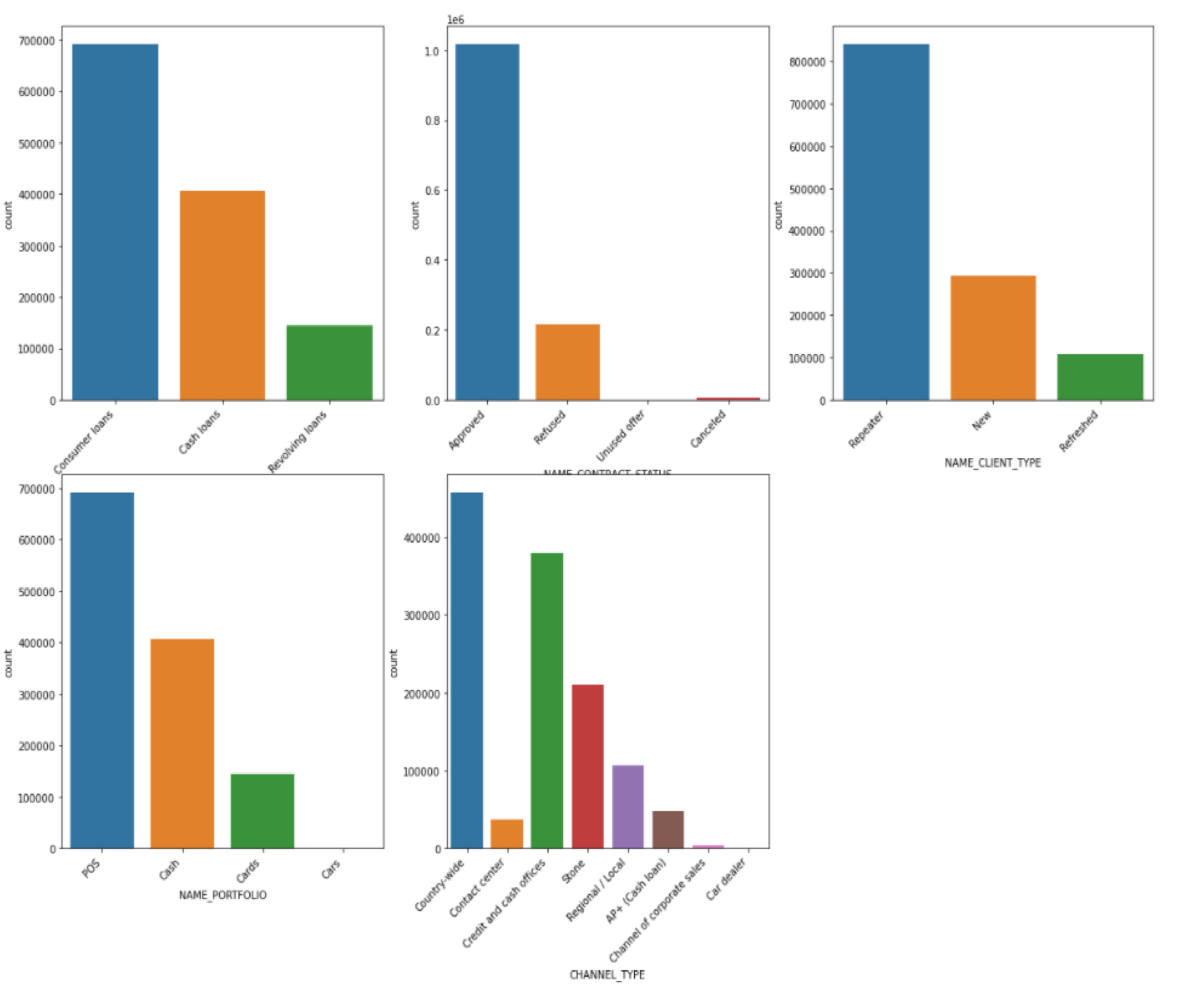
Previous Applications

Checking data imbalance

Family status and gender

We can see that there is data imbalance in below columns:-

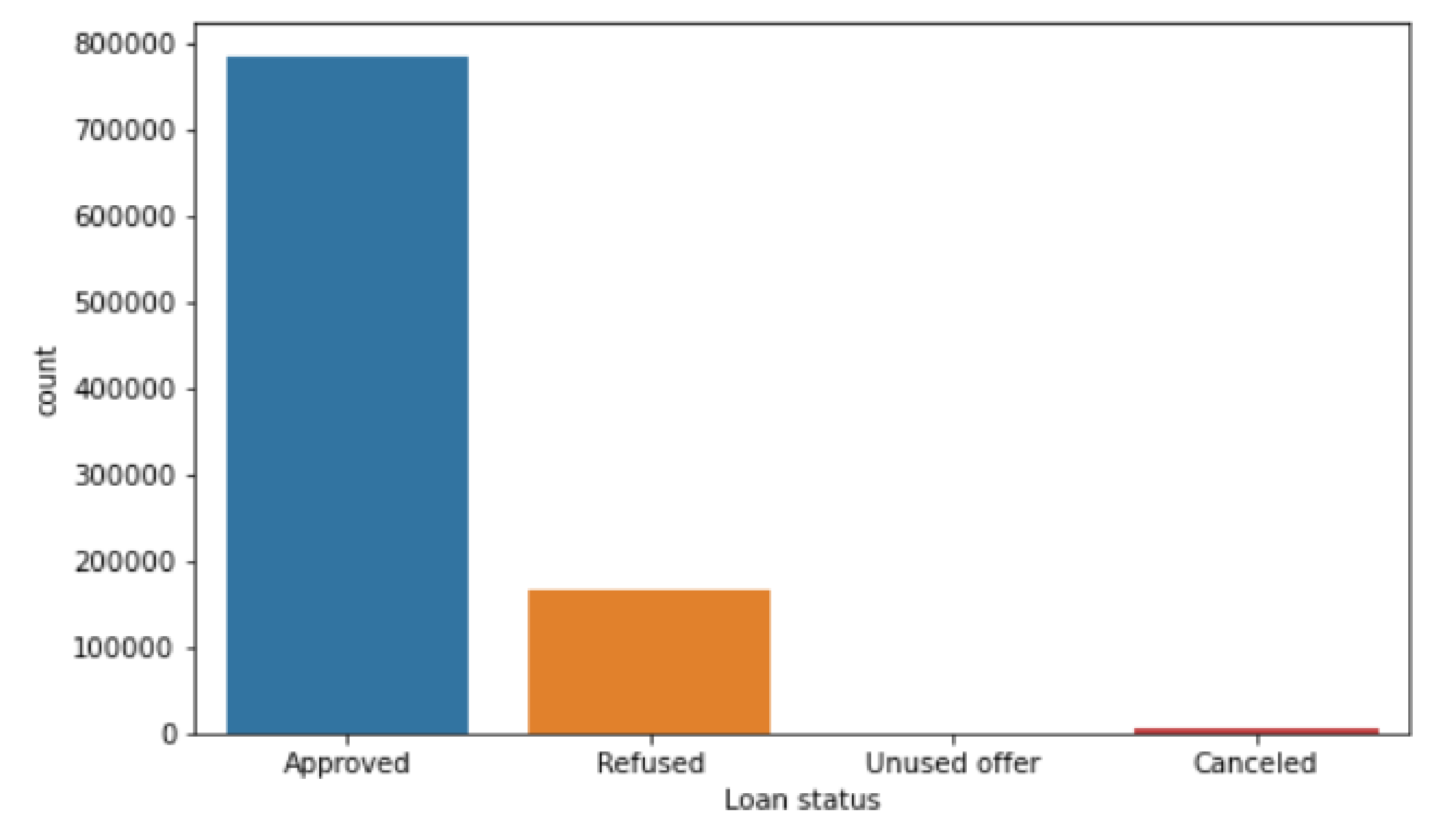
- NAME_CONTRACT_TYPE - There are very few Revolving Loans
- NAME_CONTRACT_STATUS - There are very few Refused loans. Almost negligible Canceled loans.
- NAME_CLIENT_TYPE - There are very few New applicant. Even fewer Refreshed applicants.
- NAME_PORTFOLIO - Very few application for Cards and Cars
- CHANNEL_TYPE - Except Country-Wide, Credit and Cash offices and Store all other channels are very few in number.



Univariate analysis on unordered categorical variable

Previous Loan status

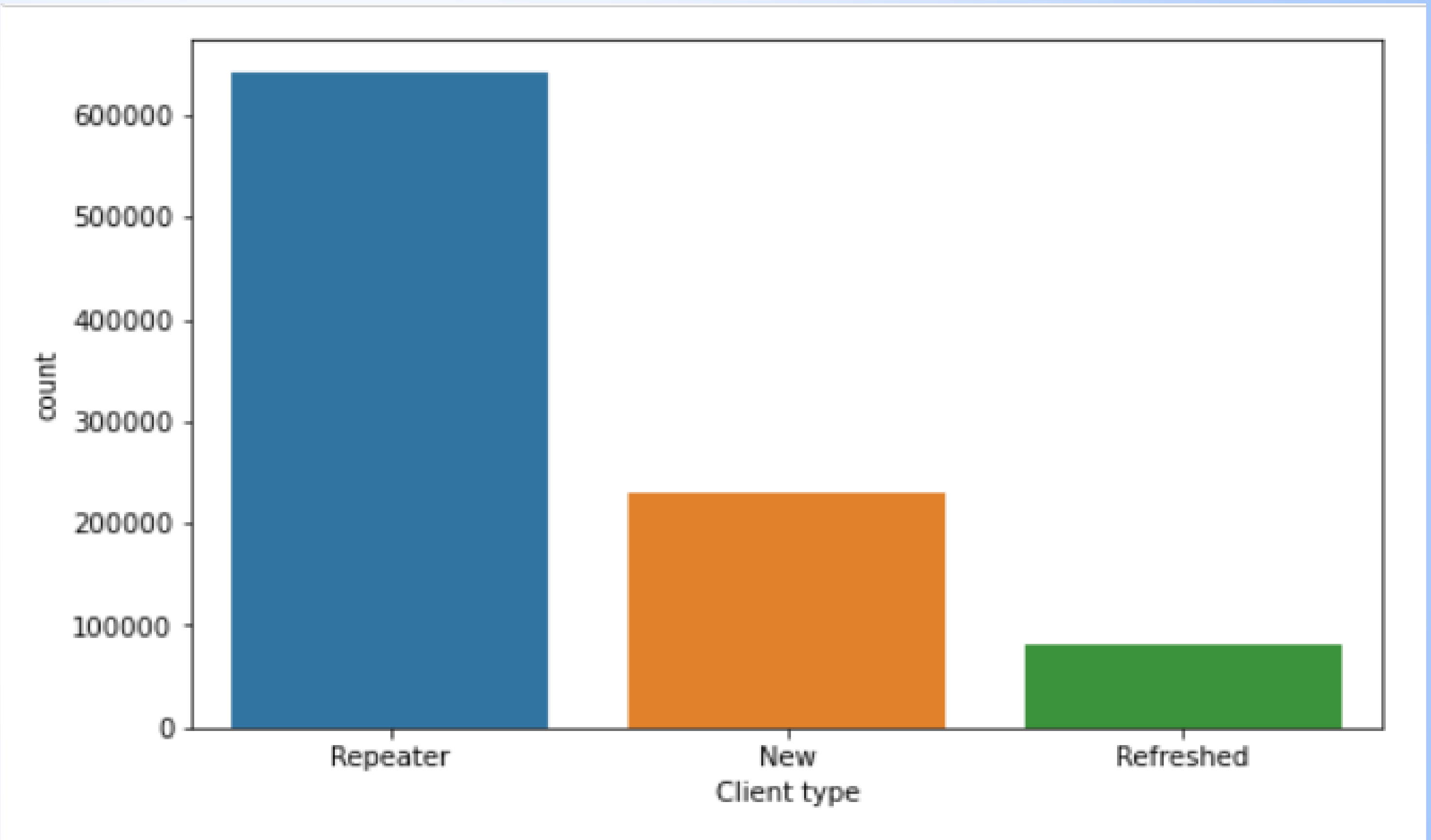
- There are huge number of Approved loan than Refused. Hardly, there are any Canceled or Unused offer loan.



Univariate analysis on unordered categorical variable

Client type

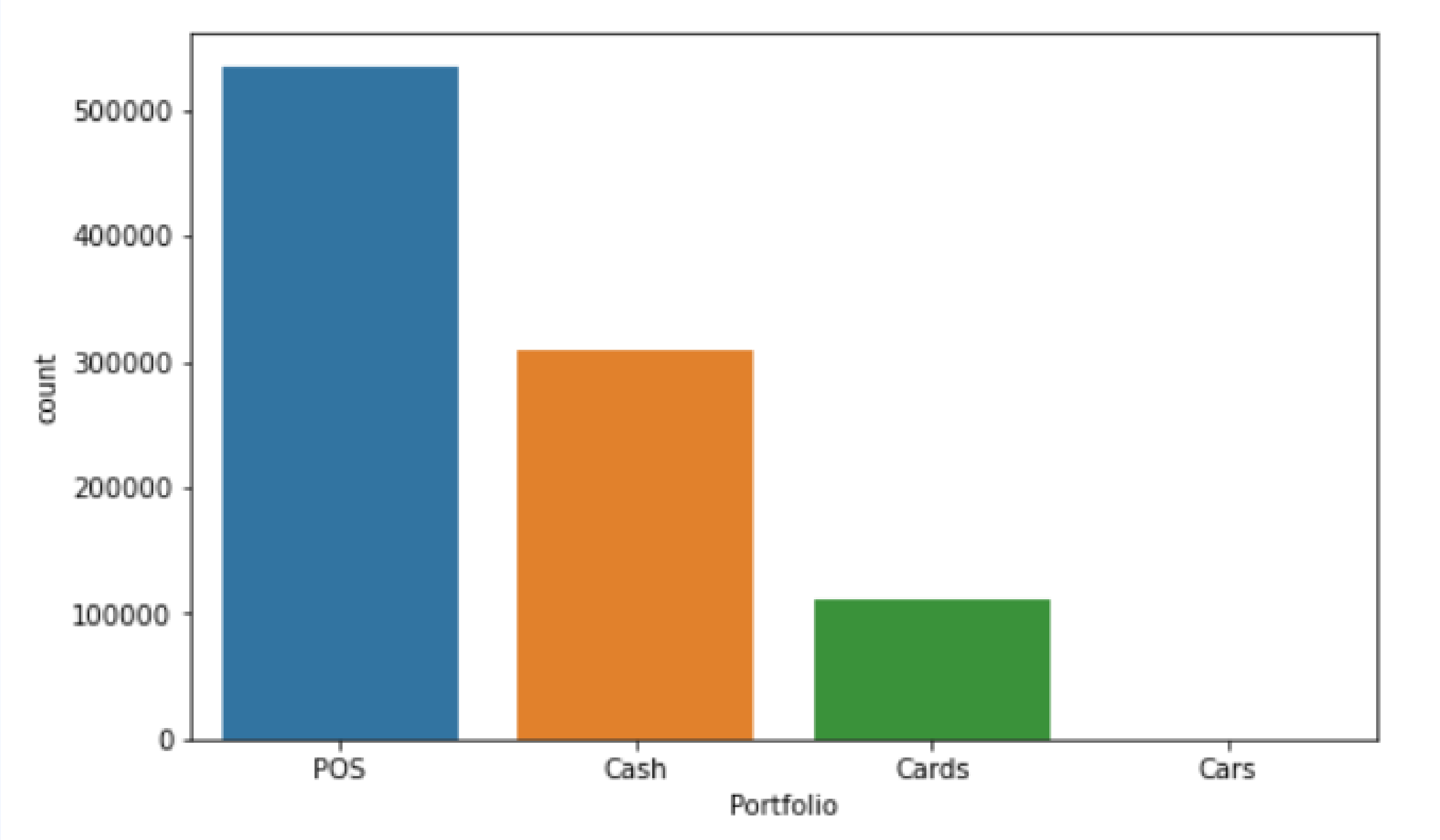
- Mostly the applicants were Repeater



Univariate analysis on unordered categorical variable

Portfolio of the previous applications

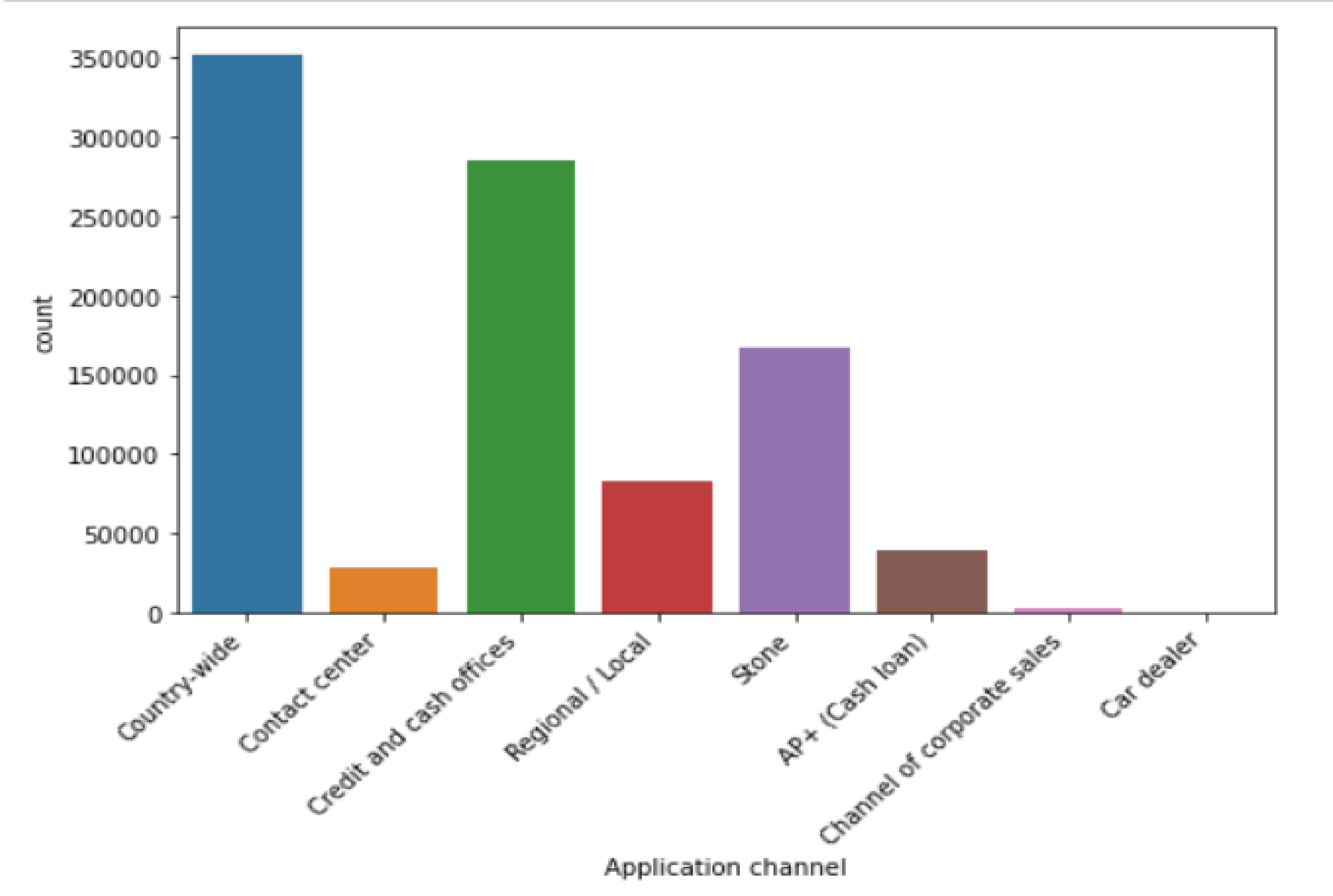
- The highest number of the previous applications was for POS. Applications for Cash also has good number. Applications for Cards were very few.



Univariate analysis on unordered categorical variable

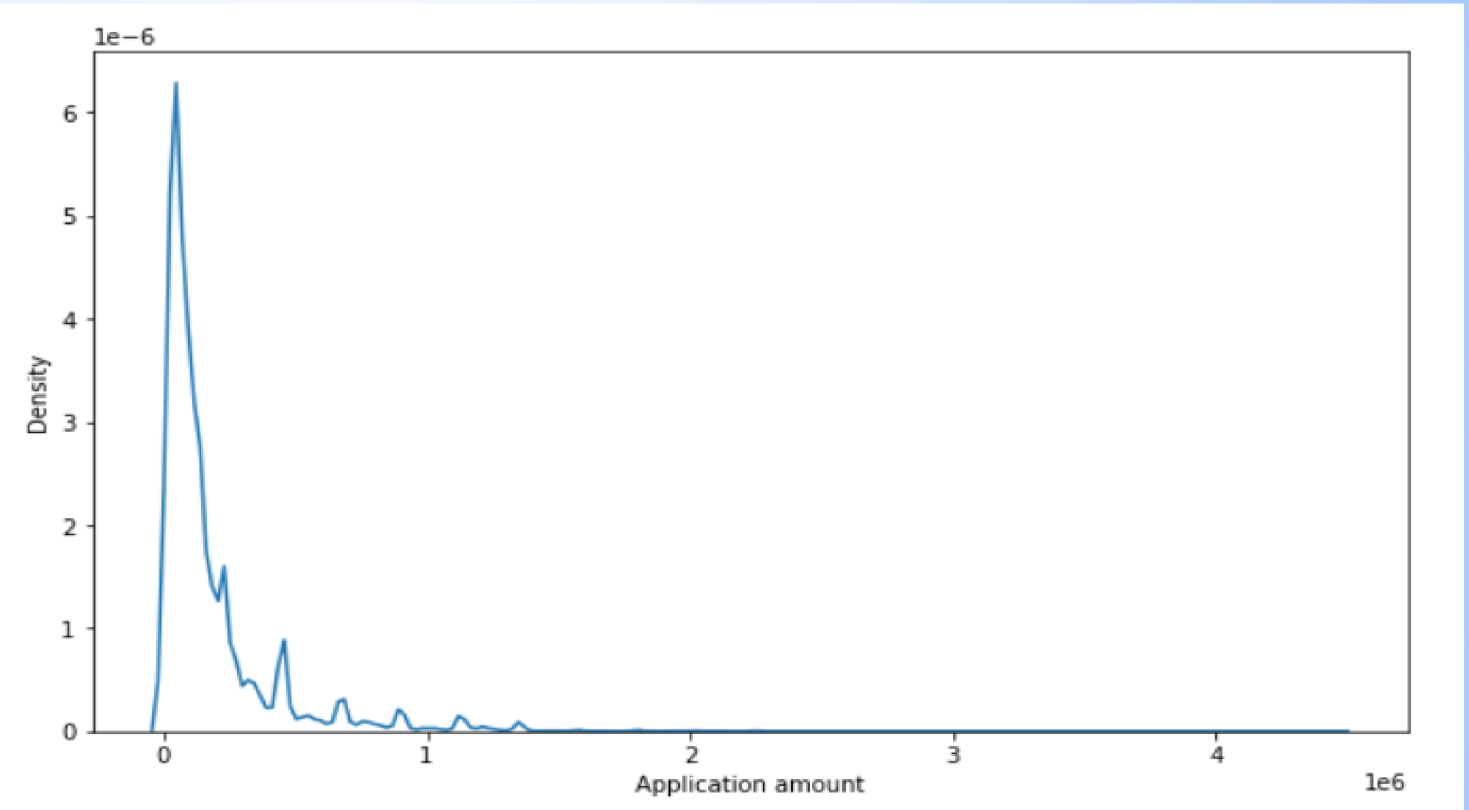
Application channel type

- We see that Country-wide was heavily used for previous applications followed by Credit and Cash offices, Stone and Regional. Rest other channels are hardly used.



Univariate analysis for continuous variables

Applied loan amount

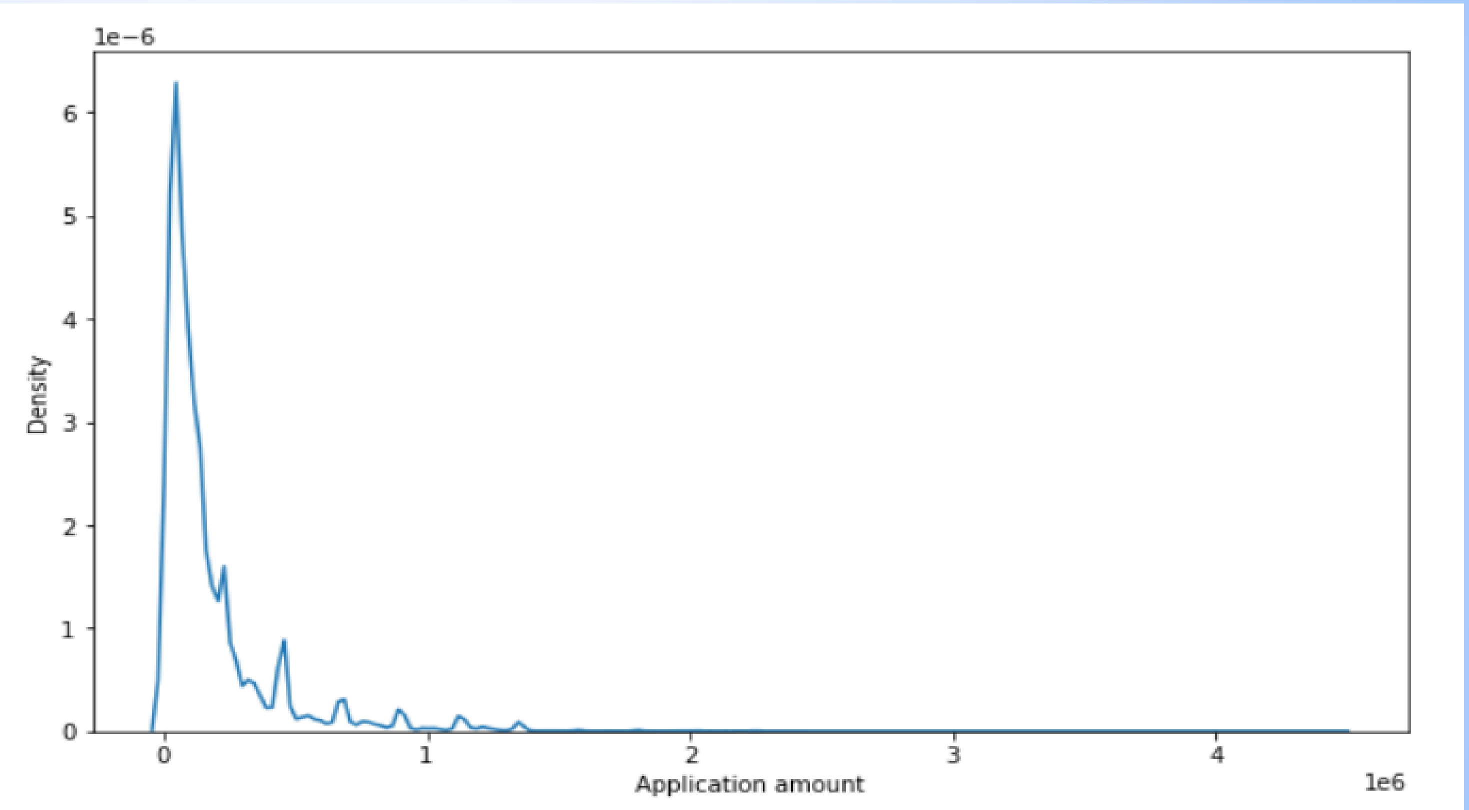


- Most of the applications were for the amount of below 250000 as we see from the above distribution.

Univariate analysis for continuous variables

Credited loan amount

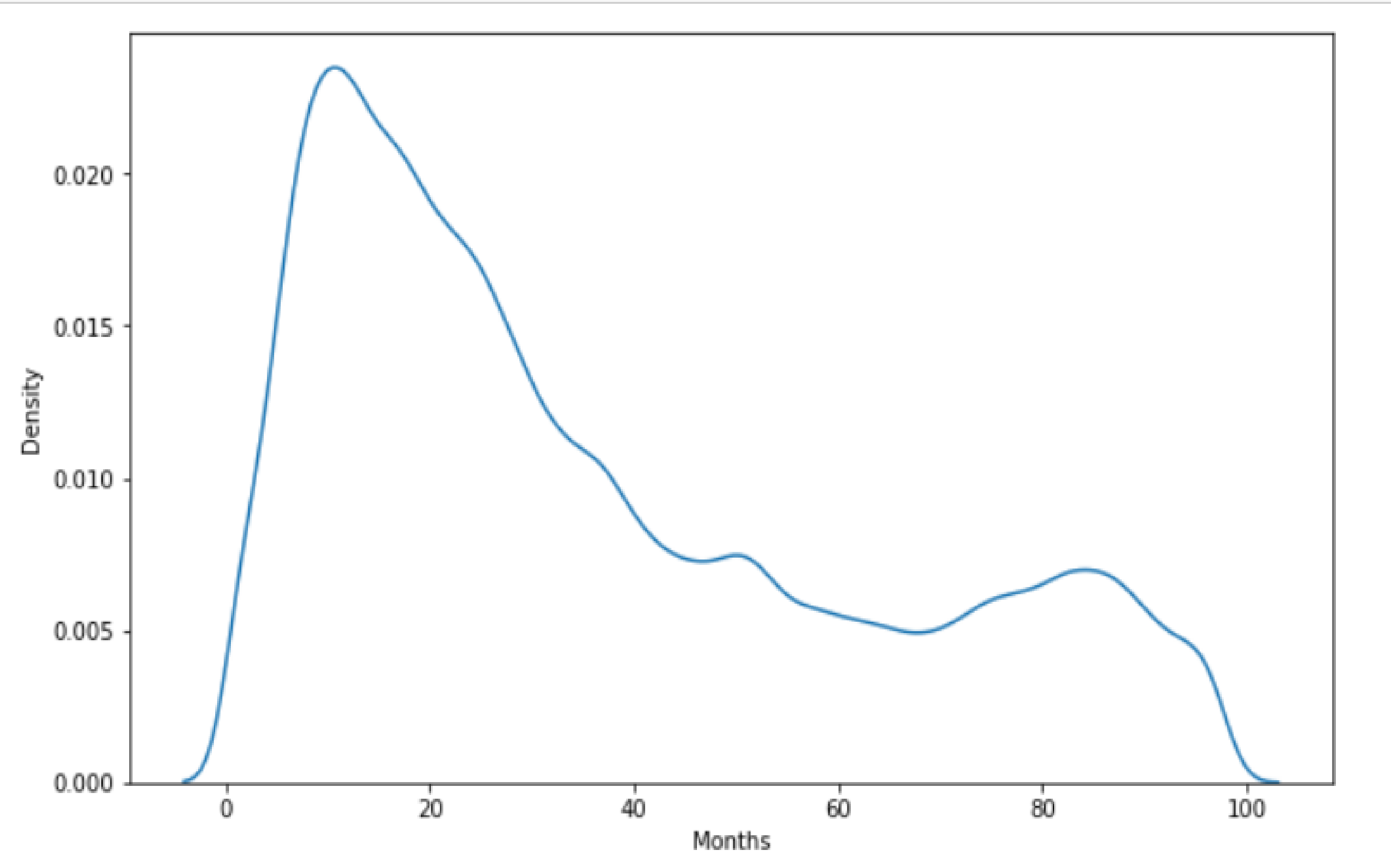
- The distribution of the credited amount of the loan was mostly in 250000 range.



Univariate analysis for continuous variables

Months took for the pervious application decision relative to the current application

- We can see that most of the applications decision took approximately 30 months. The time taken spread upto 100 months.



Bivariate analysis of previous application

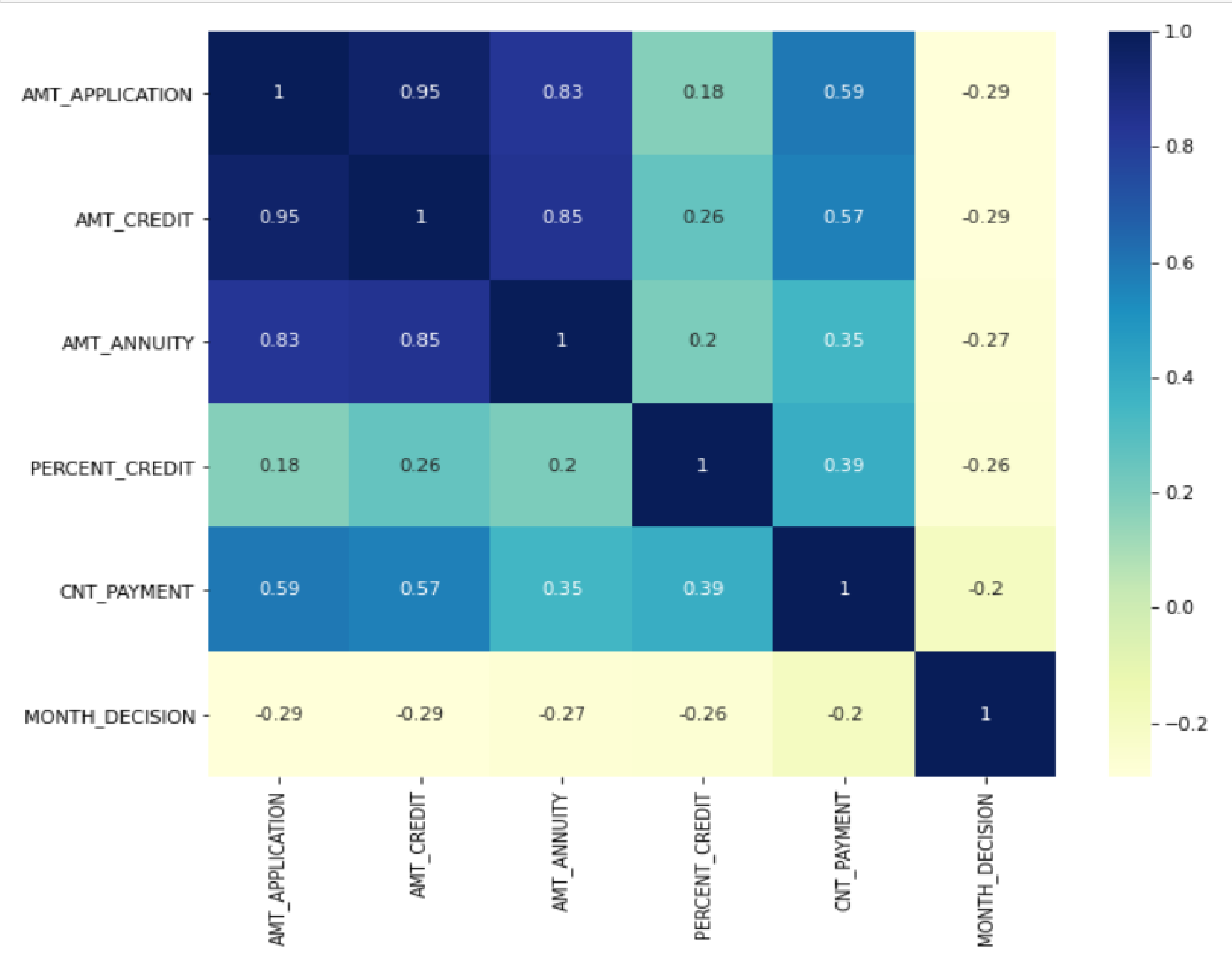
Correlation of relevant numerical columns

Highly corelate columns

- AMT_APPLICATION and AMT_CREDIT
- AMT_APPLICATION and AMT_ANNUITY
- AMT_CREDIT and AMT_ANNUITY

Moderately corelated columns

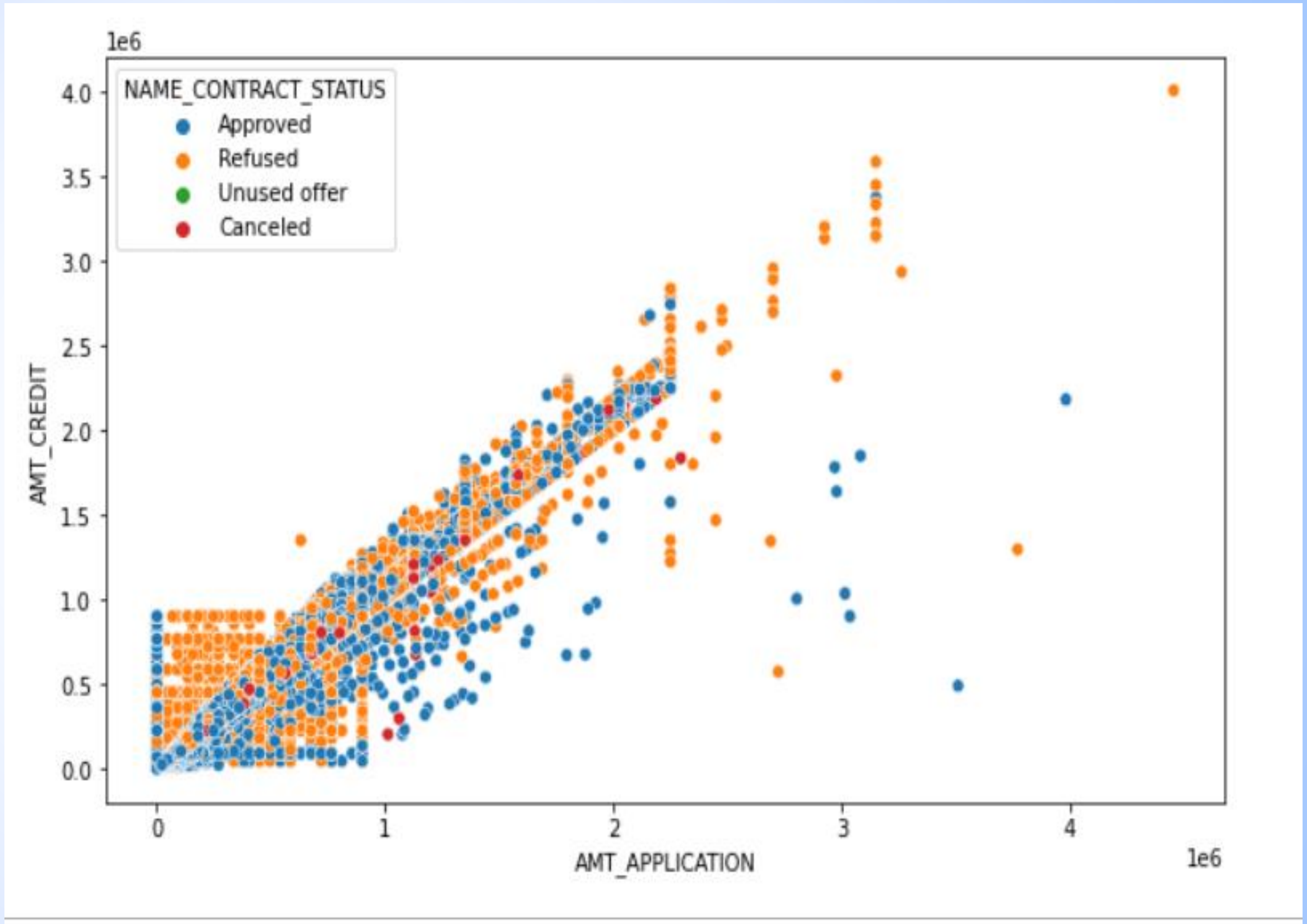
- AMT_APPLICATION and CNT_PAYMENT
- AMT_CREDIT and CNT_PAYMENT



Bivariate analysis on continuous variable

Application amount and credited amount

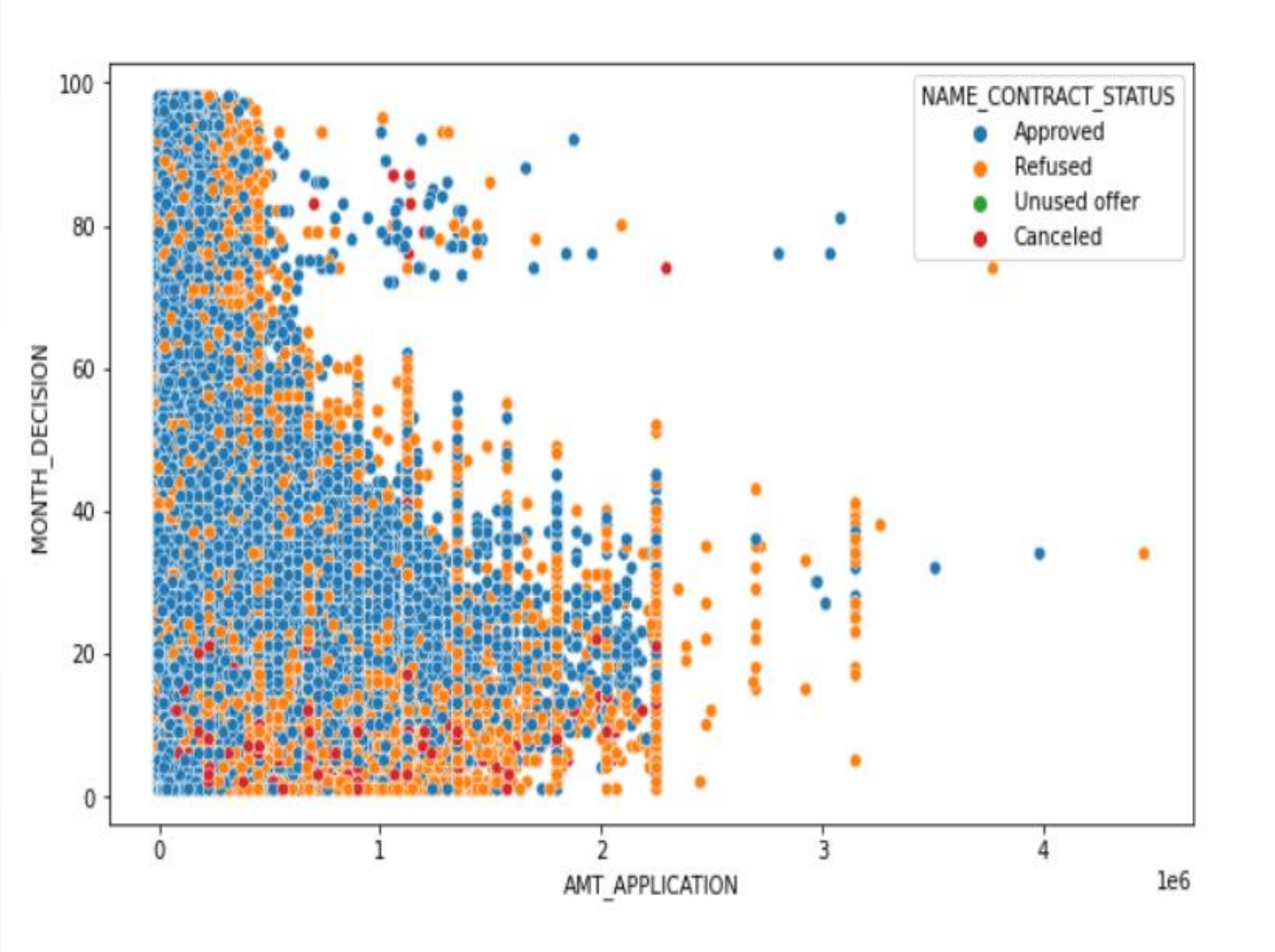
- We can see that the applications are more concentrated on the lesser amount and so as the credited amount. Also, the credited amount is increased with respect to the application amount.



Bivariate analysis on continuous variable

Application amount and the month taken to take decision related to current application

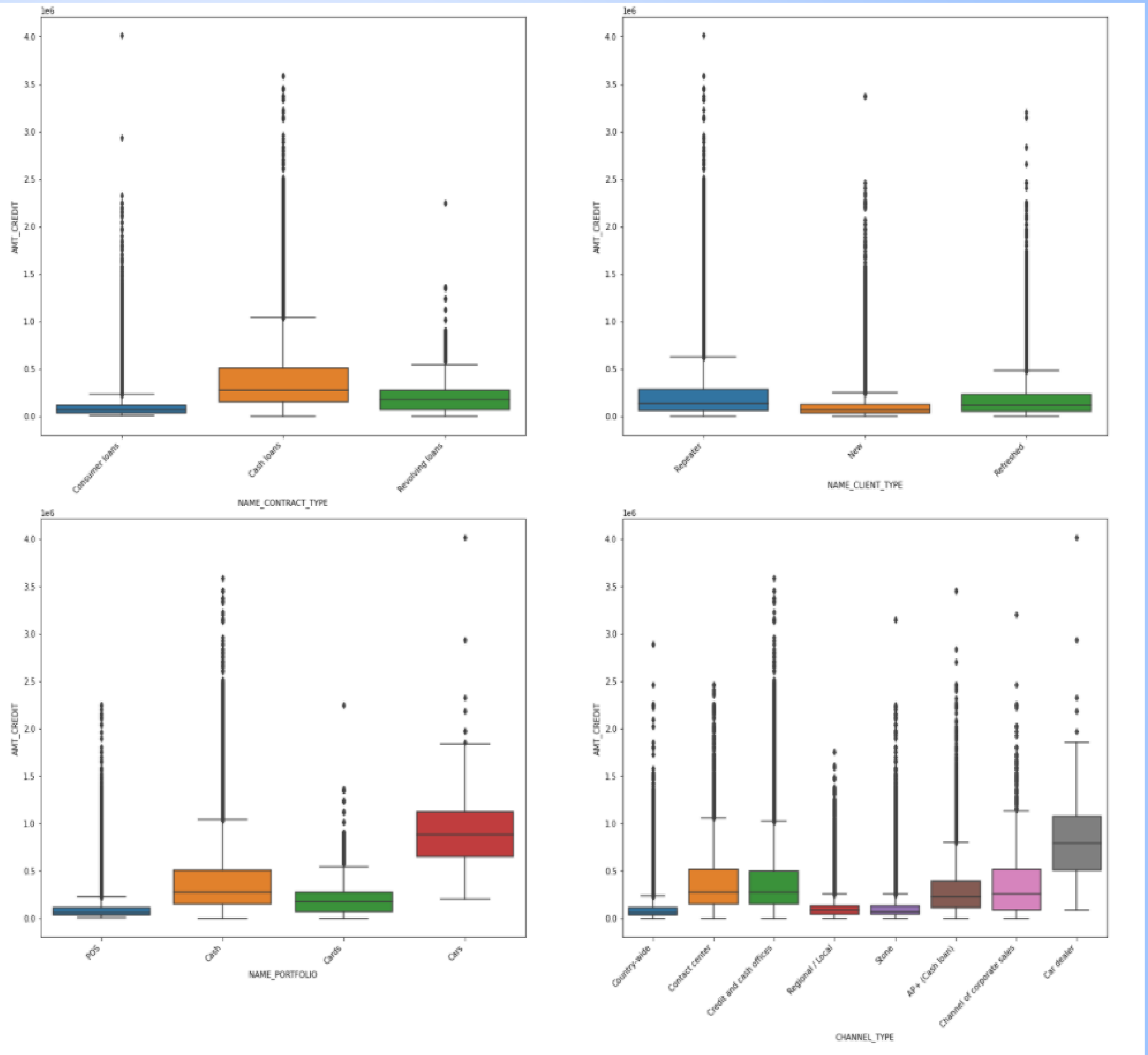
- We can see a pattern here that the more the application amount of the loan, the lesser the months taken prior to current application. That means, most of the higher amount of the loan application decision made in the recent time compared to the lower loan amount application.



Bivariate analysis on categorical variable

Credit amount of the loan of various categories

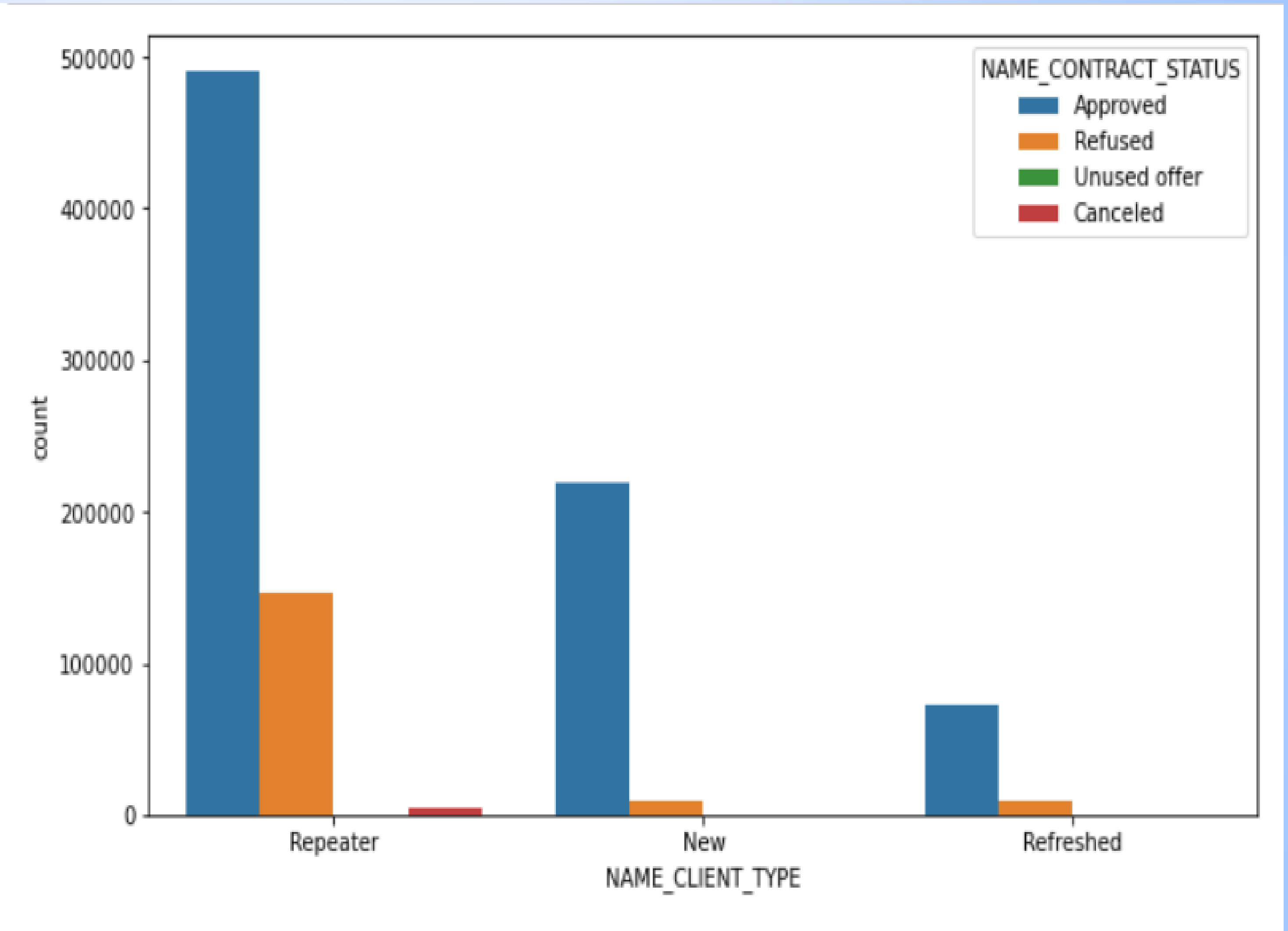
- Cash loans are more credited in amount than Revolving and Consumer loans.
- Repeater clients get more amount loan than new and refreshed clients.
- The loan with portfolio Cars are more amount credited followed by Cash.
- The credit amount of the loan is more from the application channel type as car dealer followed by Channel of corporate sales, Credit and cash offices and Contact center. The amount is very less for Regional, Stone and Country-wide channels.



Bivariate analysis on categorical variable

Status and Client type

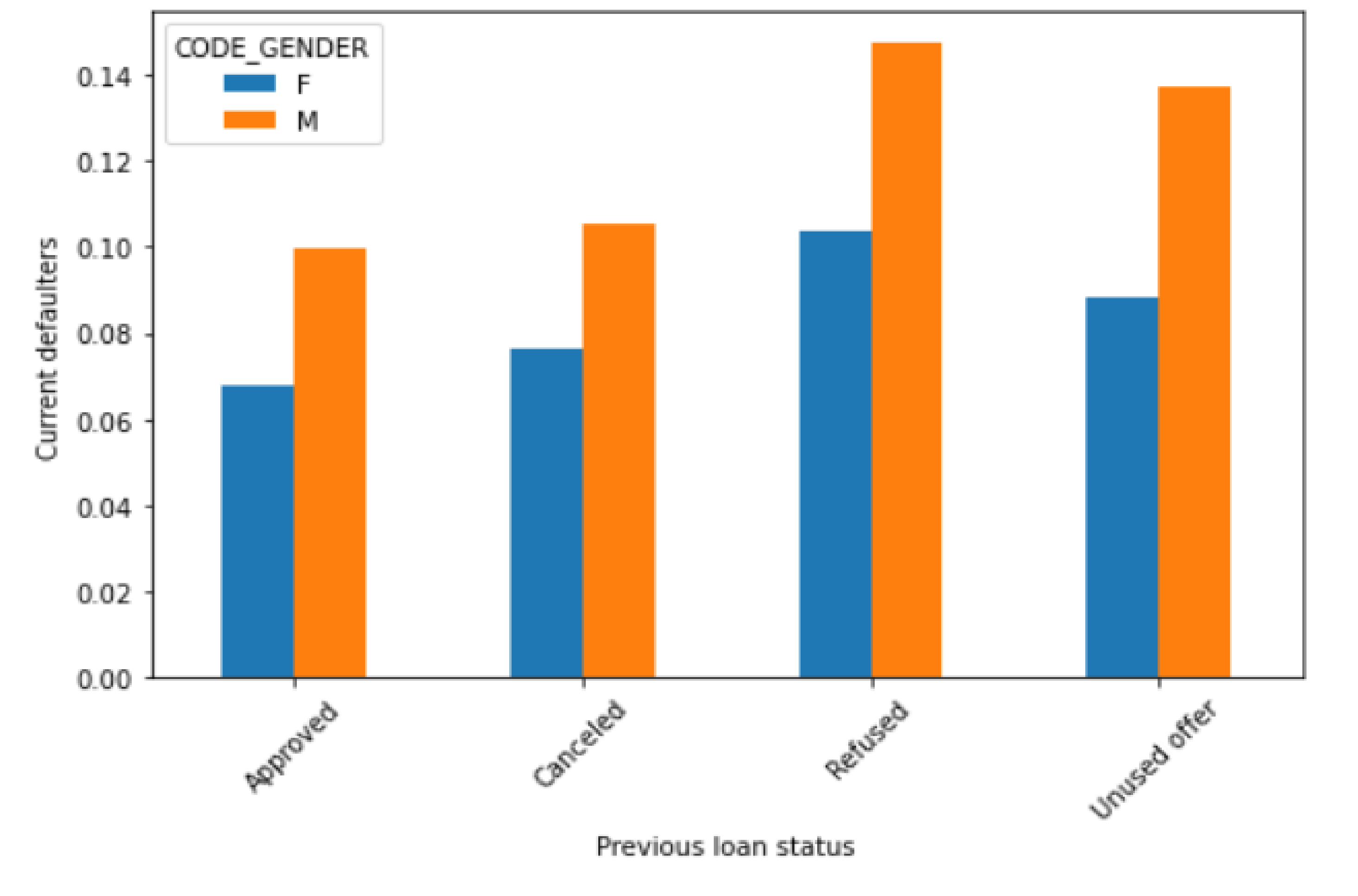
- We see that the Repeater clients have more approved loans than New and Refreshed clients.



Bivariate analysis on categorical variable

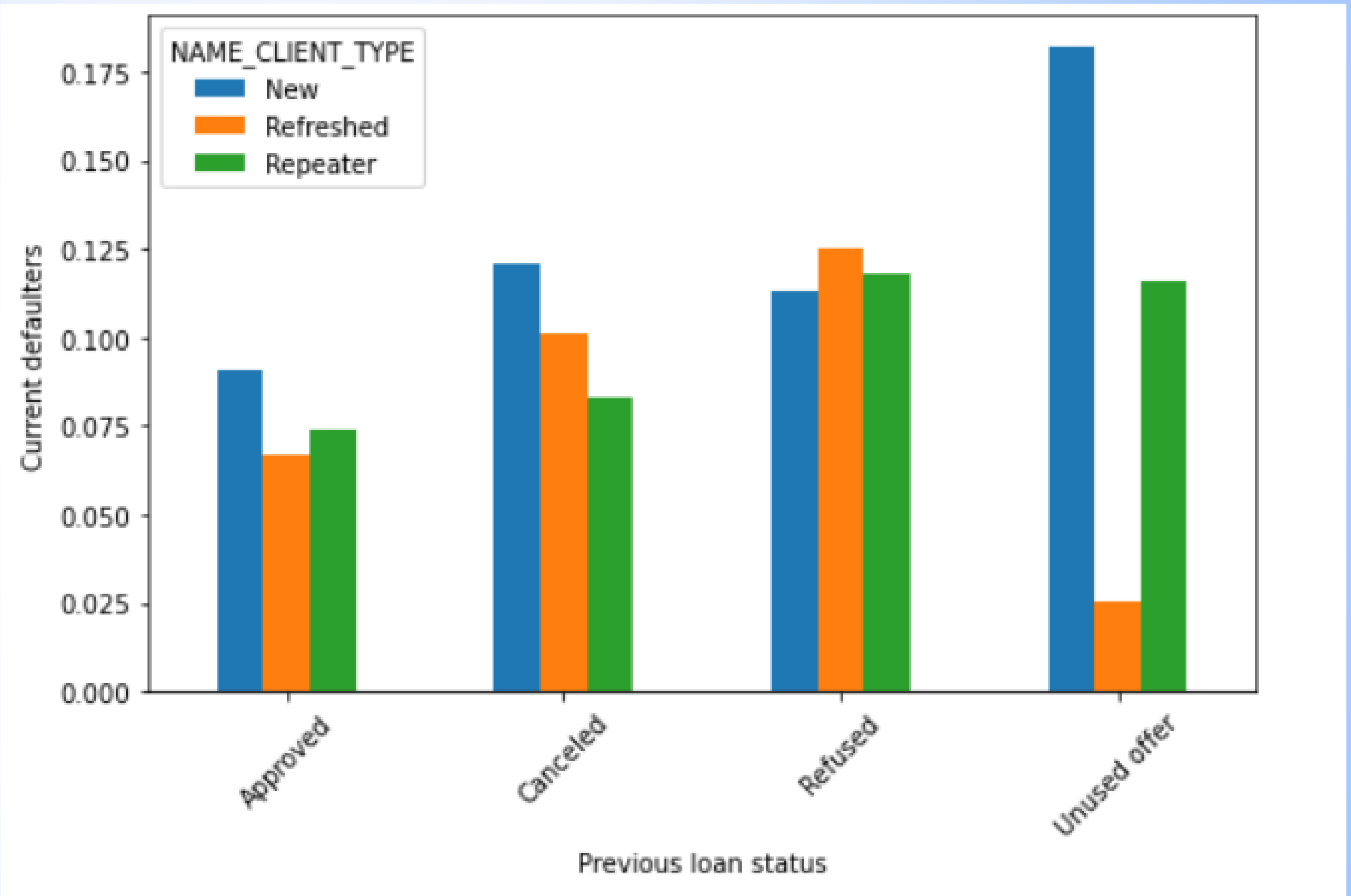
Current loan defaulter status with respect to previous loan application status

- We see that previously Refused client is more defaulted than previously Approved clients. Also, in all the cases the Males are more defaulted than Females.



Bivariate analysis on categorical variable

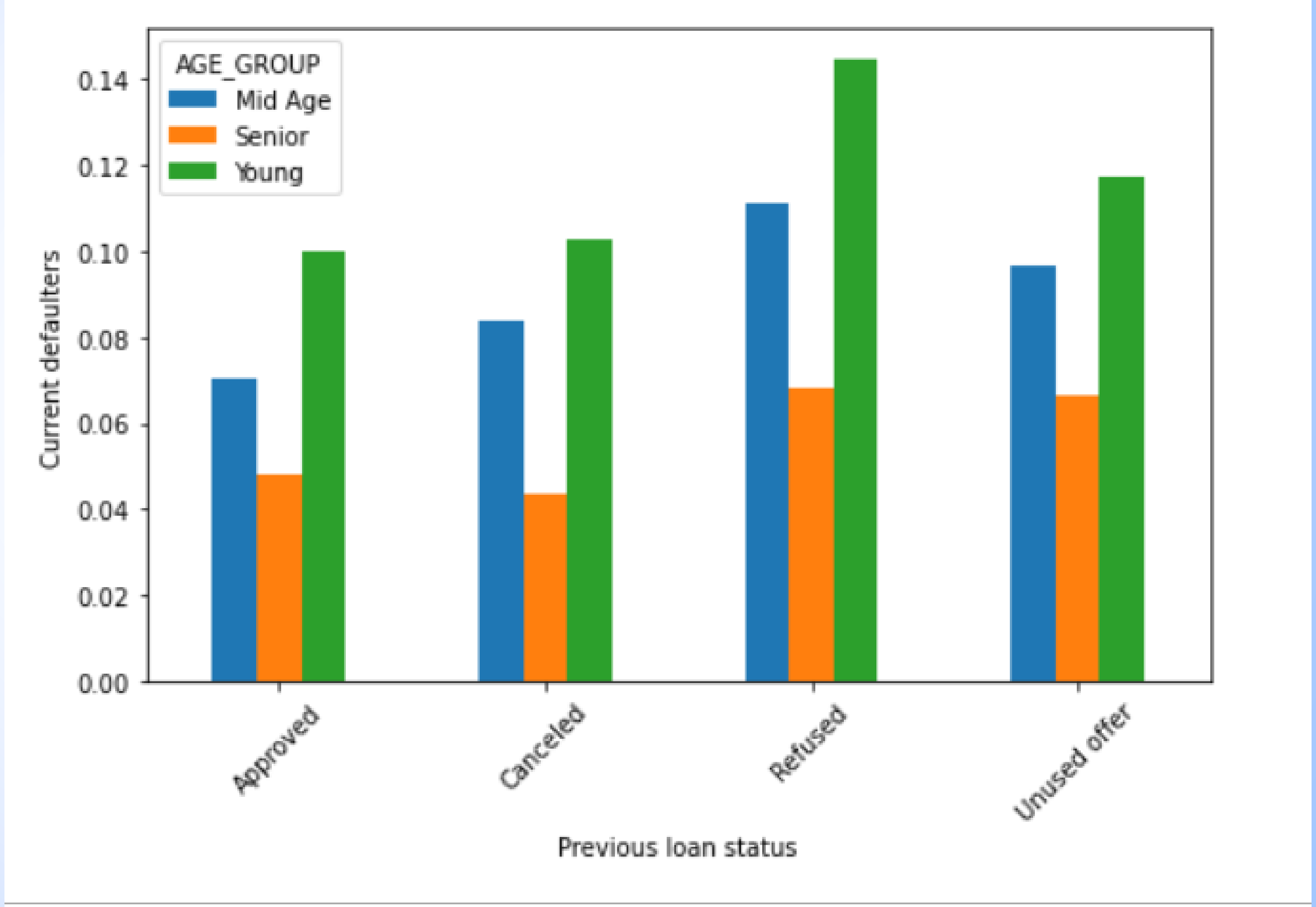
Current loan defaulter status with respect to previous loan application status and client types



- We can see that the Defaulters are more for previously Unused offers loan status clients, who were New.
- For previously Approved status the New clients were more defaulted followed by Repeater.
- For previously Refused applicants the Defaulters are more Refreshed clients.
- For previously Canceled applicants the Defaulters are more New clients.

Bivariate analysis on categorical variable

Current loan defaulter status with respect to previous loan application status and age group

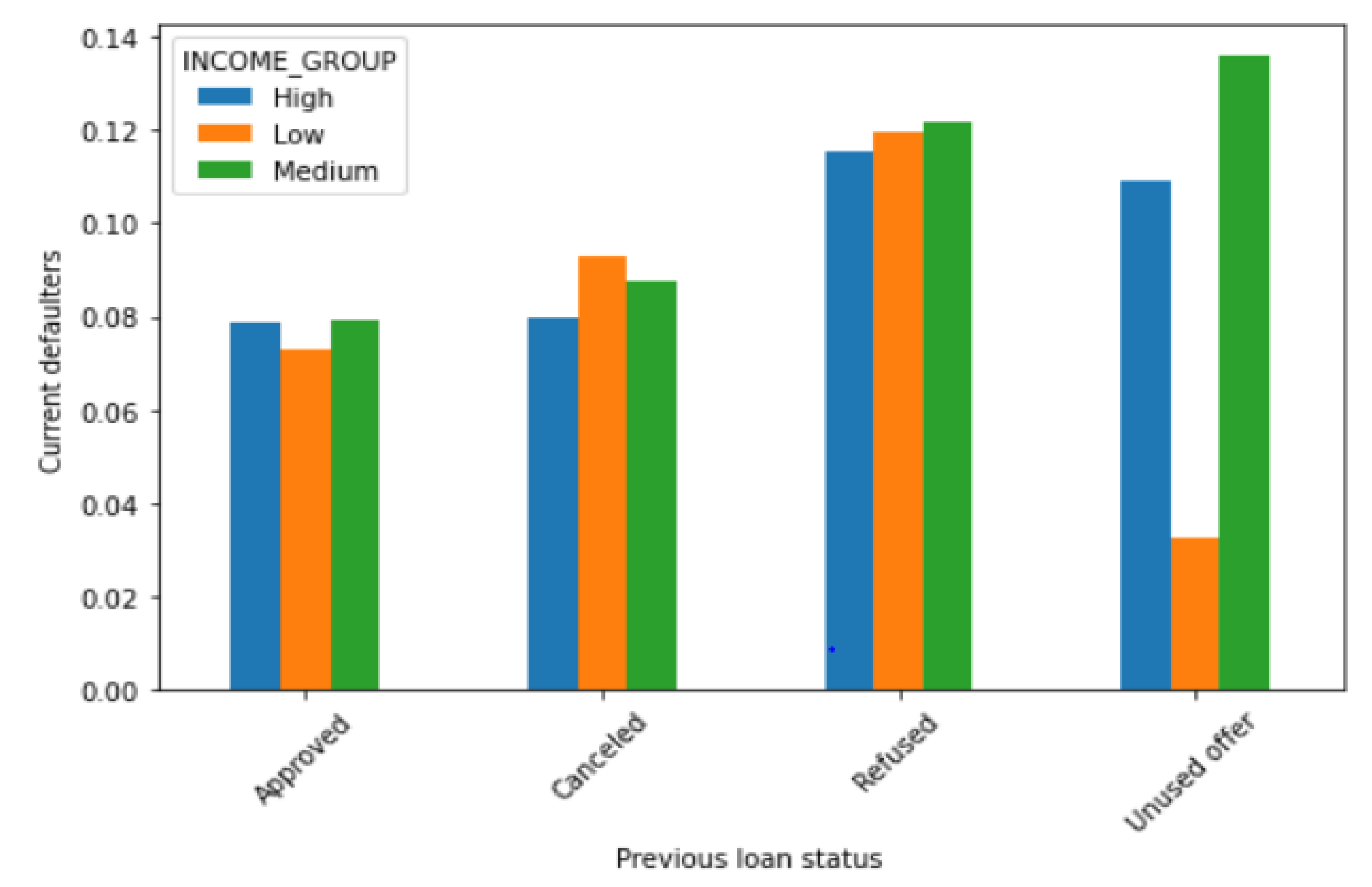


- For all the previous status Young applicants are more defaulted.
- For all the previous status Senior applicants are less defaulted compared to others.

Bivariate analysis on categorical variable

Current loan defaulter status with respect to previous loan application status and income group

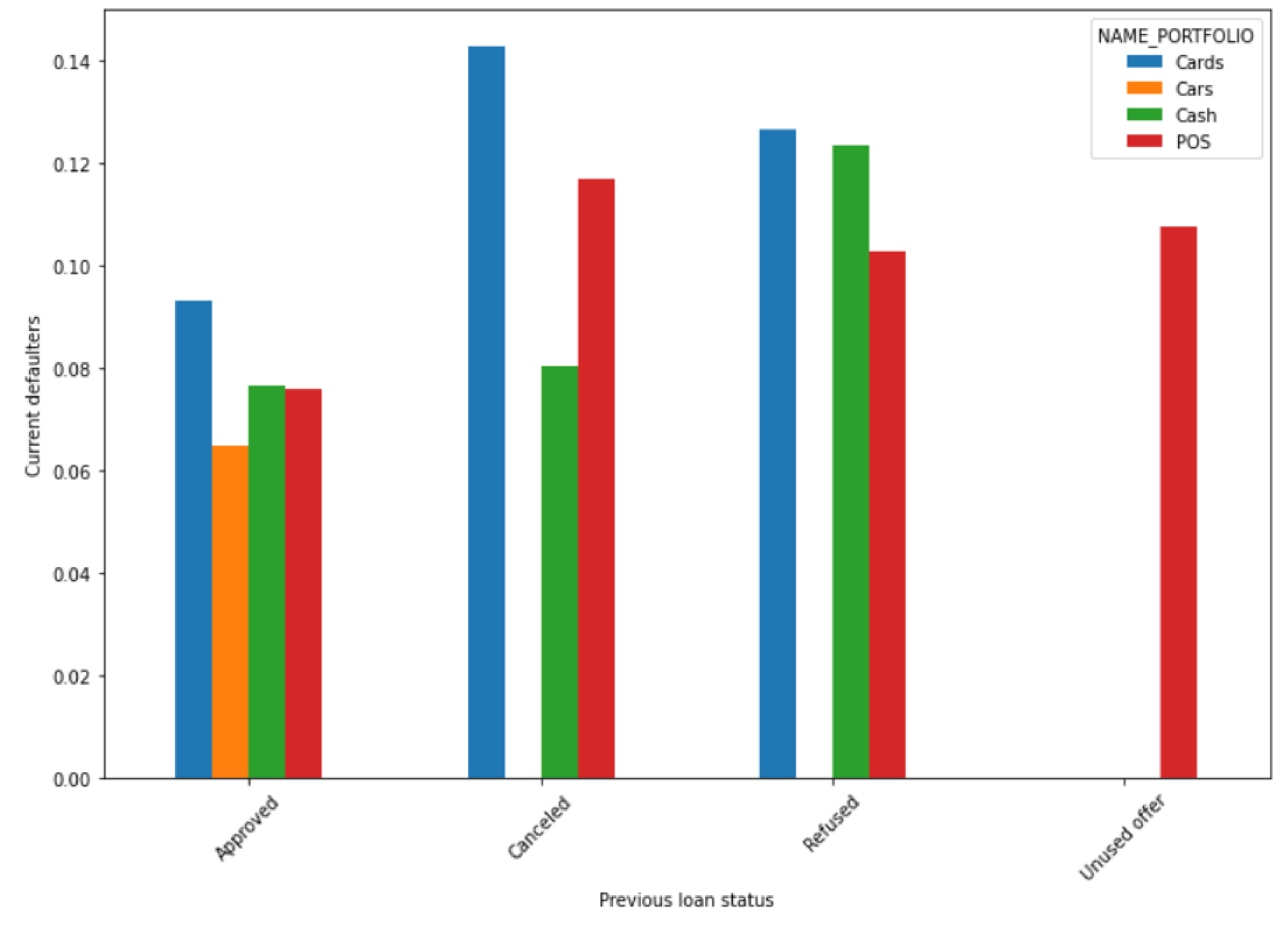
- For previously Unused offer the Medium income group was more defaulted and Low income group is the least.
- For other application status more or less all the income groups are equally defaulted.



Bivariate analysis on categorical variable

Current loan defaulter status with respect to previous loan application status and portfolio of the loan

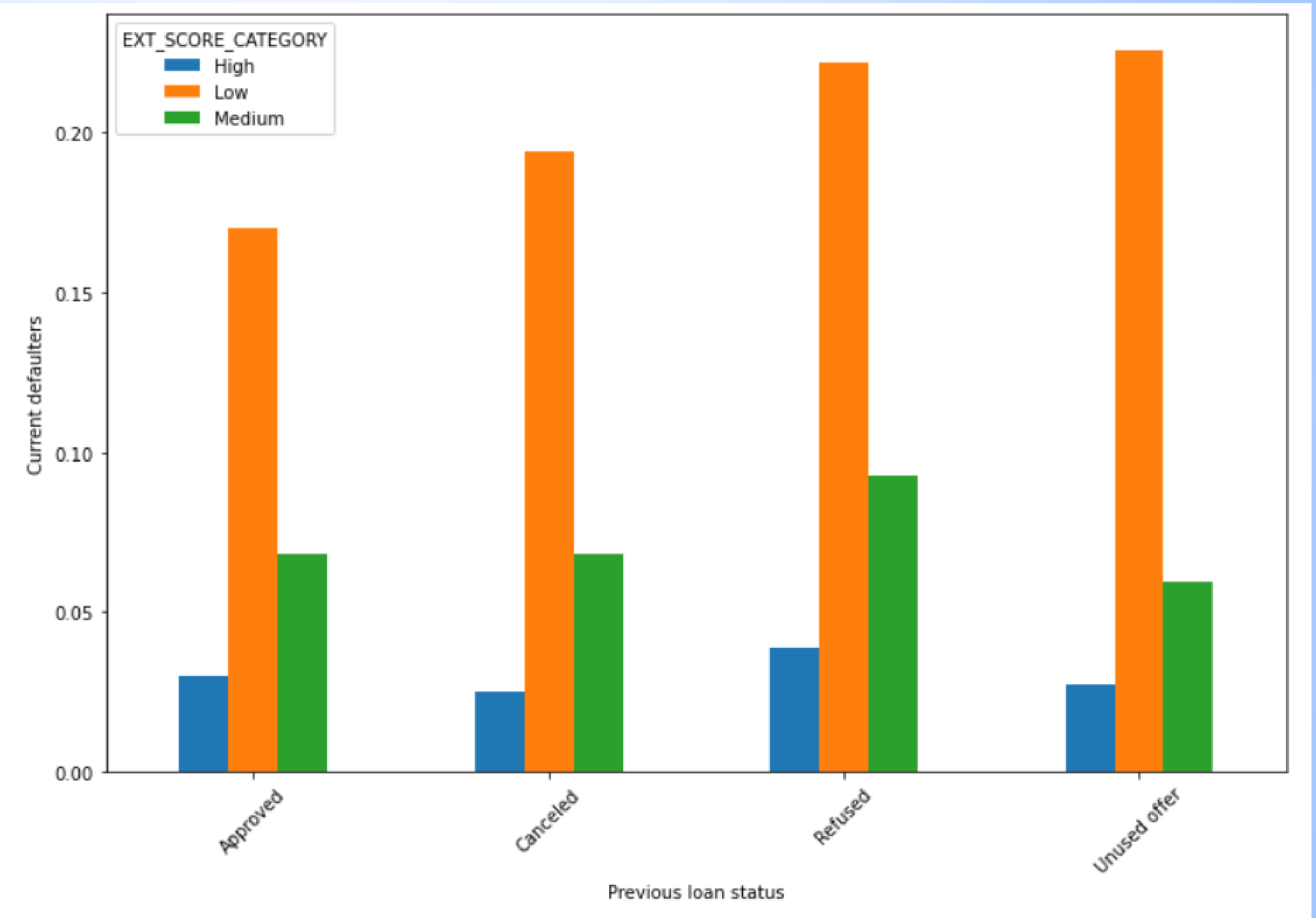
- Most of the clients were defaulted, who previously applied loan for Cards.
- For approved loan status\the clients applied for Cars are less defaulted.
- For Refused loan status the clients applied for POS are less defaulted.



Bivariate analysis on categorical variable

Current loan defaulter status with respect to previous loan application status and external source score category

- Applicants with low external source score are highly defaulted.
- Higher scorer applicants are very unlikely to default irrespective of their previous loan status.



Business Recommendation

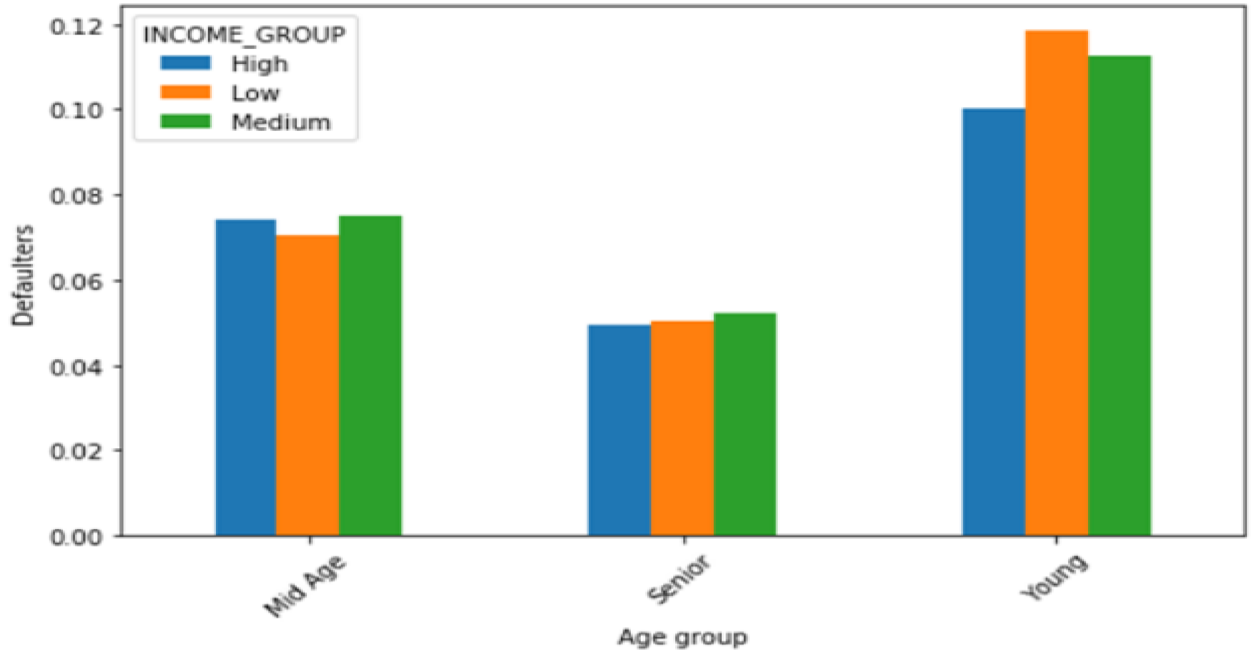
Current applications

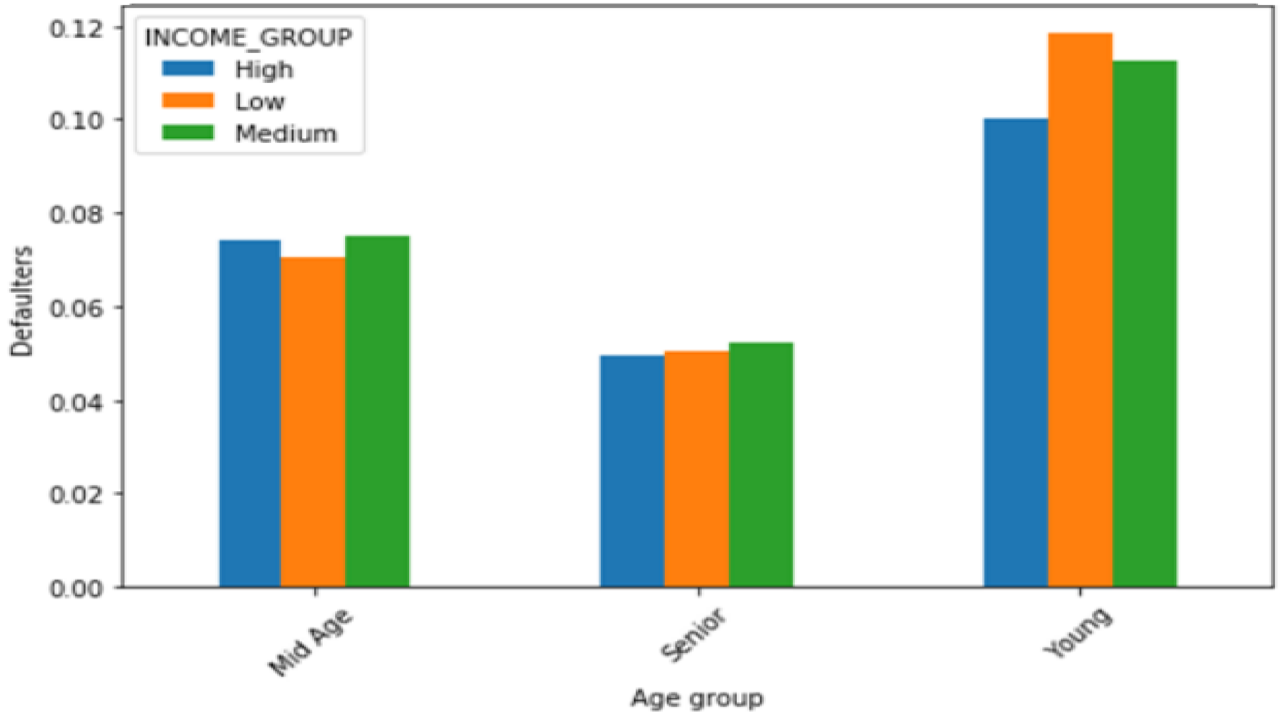
Observation

- High income groups are less defaulter than comparatively lower income groups.
- Mid age and senior people with all income groups are less defaulted.

Recommendation

- Safer to grant loan for mid age and senior citizen clients with higher income.
- Risky to grant loans for young people with low income groups.





Business Recommendation

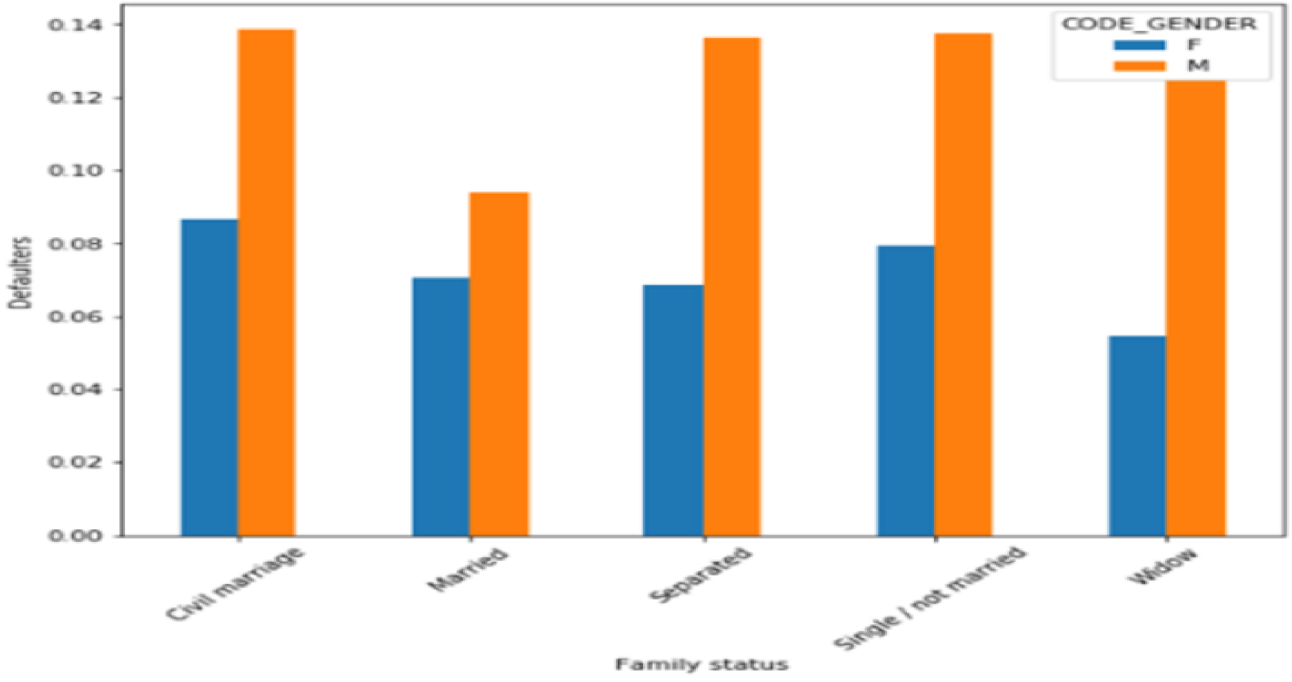
Current applications

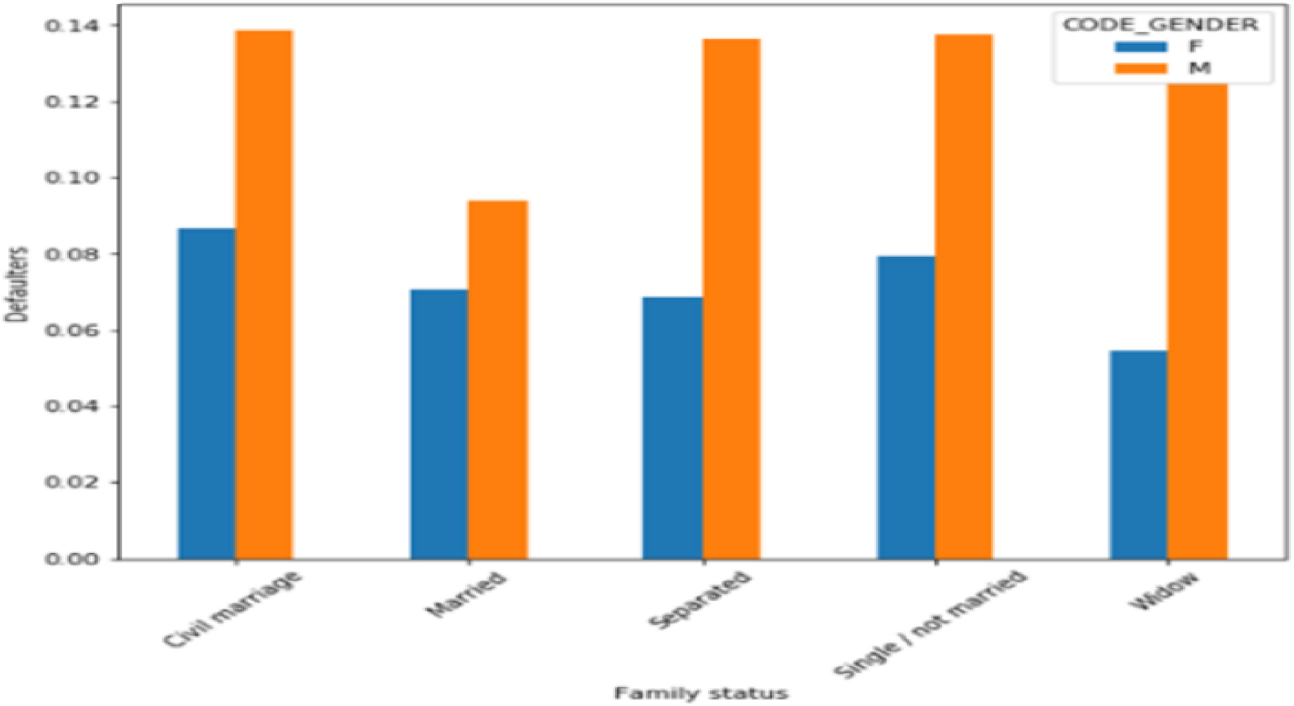
Observation

- Senior people irrespective of family status are less likely to be defaulted.
- Young people are more likely to be defaulted in all family status.
- Males are more like to be defaulted than females.

Recommendation

- Better to grant loan for senior citizen of all family status.
- It is risky to grant loan for single, separated and civil marriage young men.





Business Recommendation

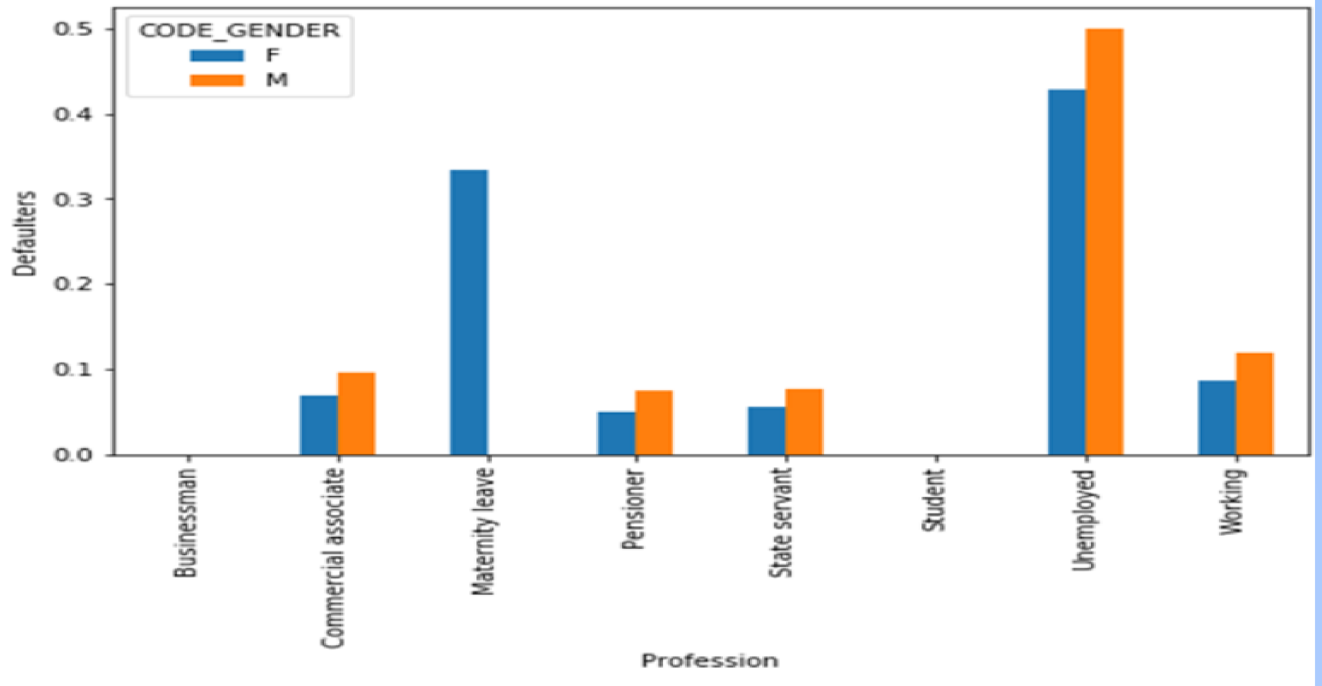
Current applications

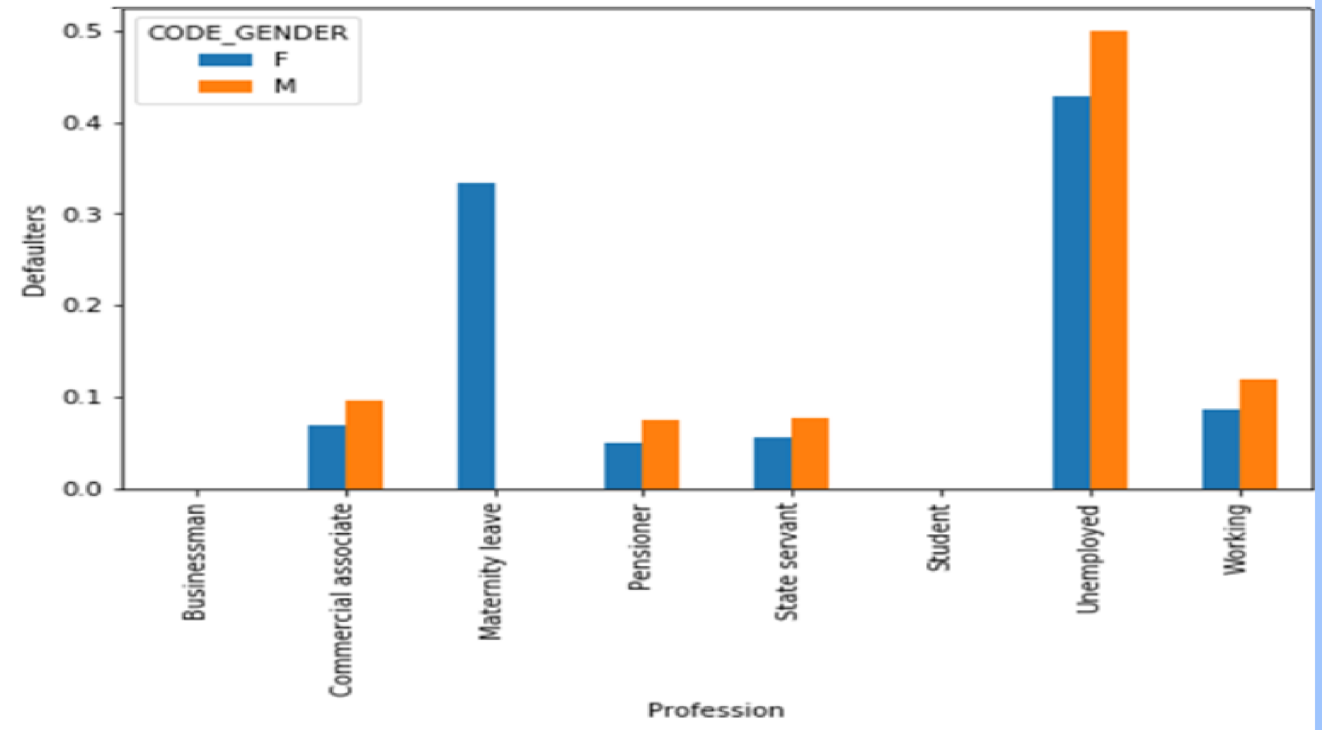
Observation

- Higher educated people are less defaulted and lower secondary educated people are more.
- Unemployed clients along with clients with maternity leave are heavily defaulted.

Recommendation

- Safe to grant loans to higher educated clients across all profession except unemployed and women with maternity.





Business Recommendation

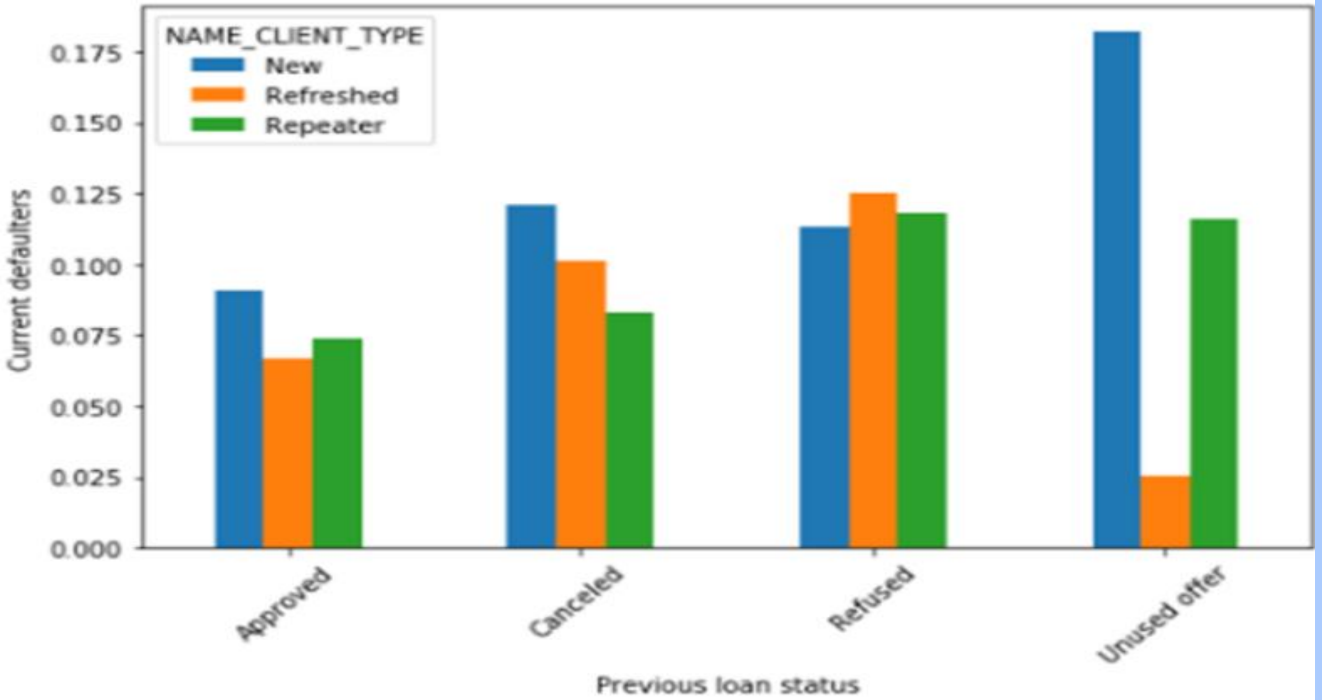
Loan application status relations - Current and Previous

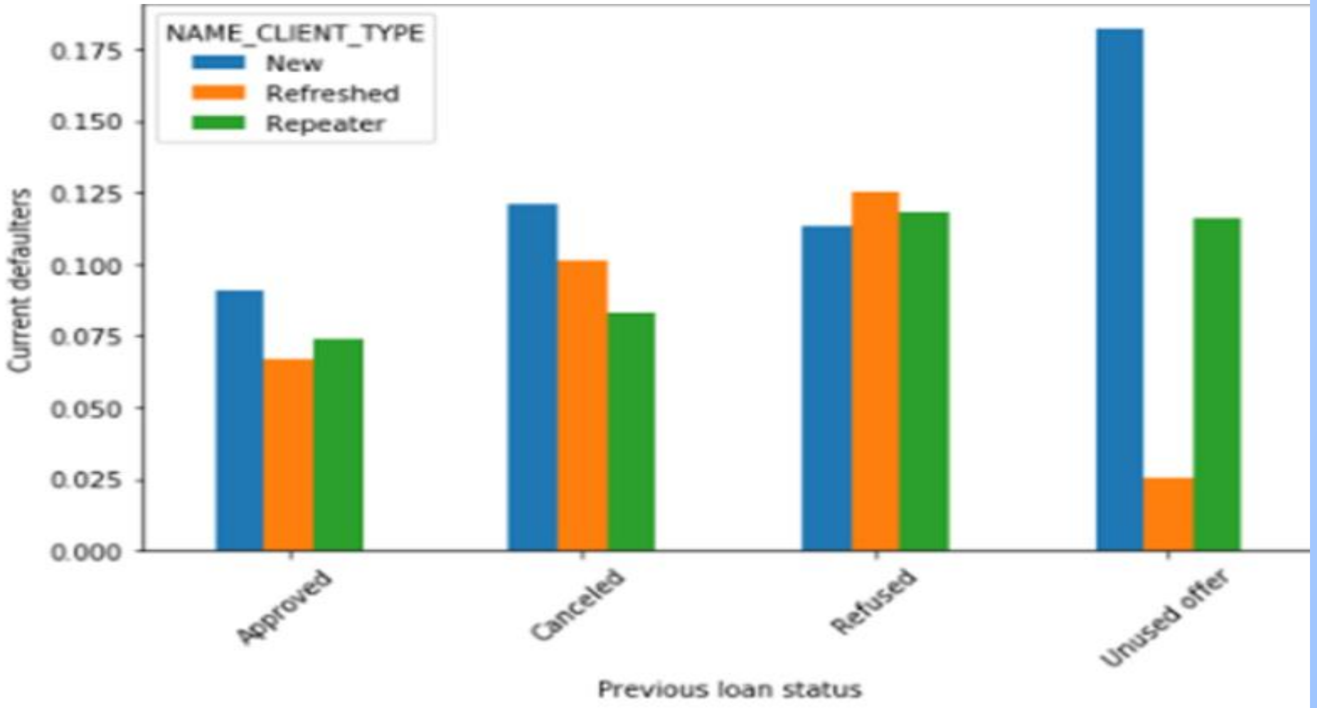
Observation

- Previously refused and unused offer applications were more defaulted in male.
- New clients with previously unused offer are more defaulted.

Recommendation

- It is recommended to provide loans to previously approved females.
- There is a risk to grant loans for clients, whose applications were refused or unused previously.





Business Recommendation

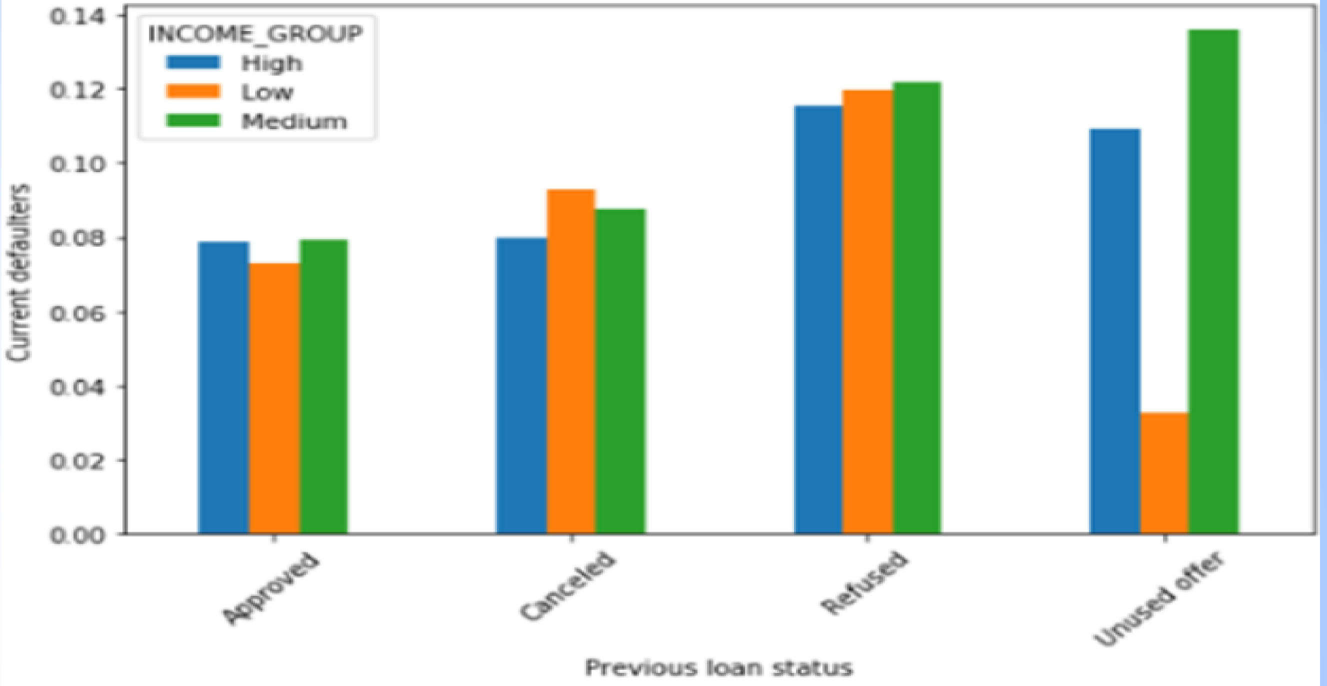
Loan application status relations - Current and Previous

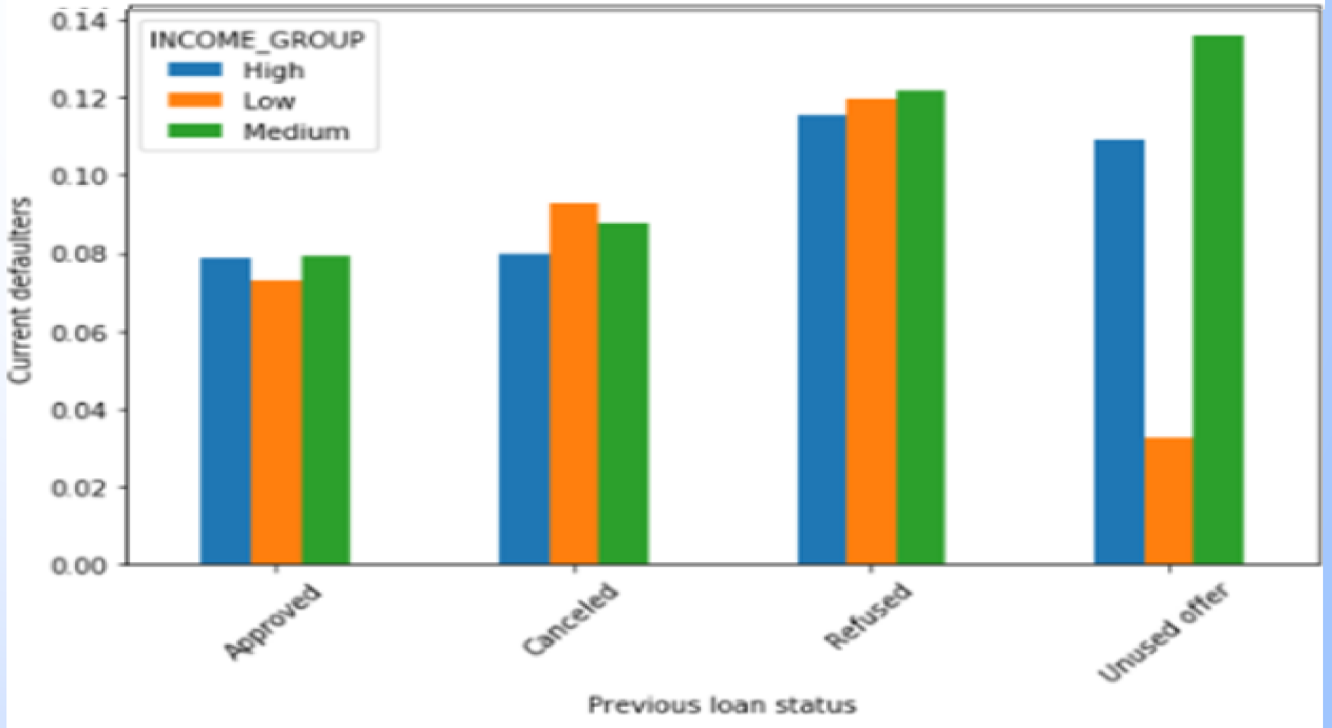
Observation

- Young people, who were previously refused are mostly defaulted.
- The senior citizens are less defaulted irrespective of their previous loan status.
- In all income groups previously refused applicants are more defaulted.

Recommendation

- Safer to grant loans for senior citizen.
- Lesser risk to grant loans for approved applicants to all income groups.





Business Recommendation

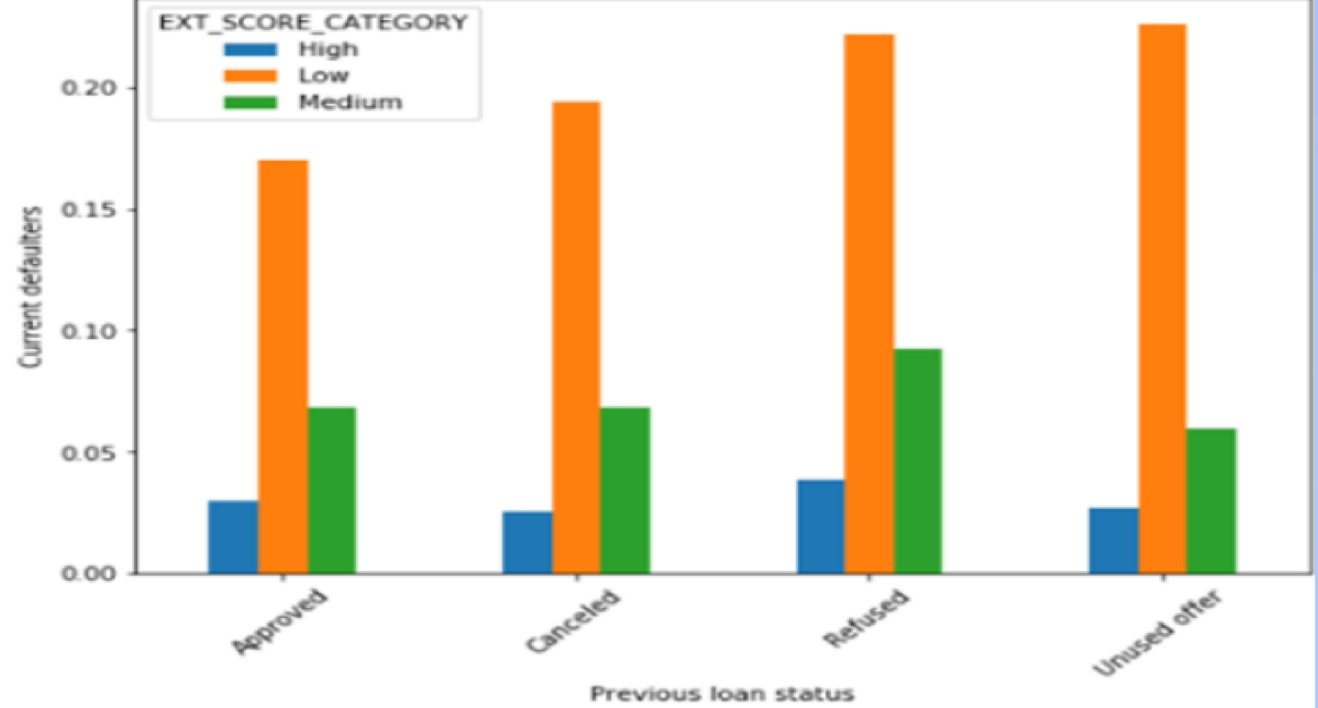
Loan application status relations - Current and Previous

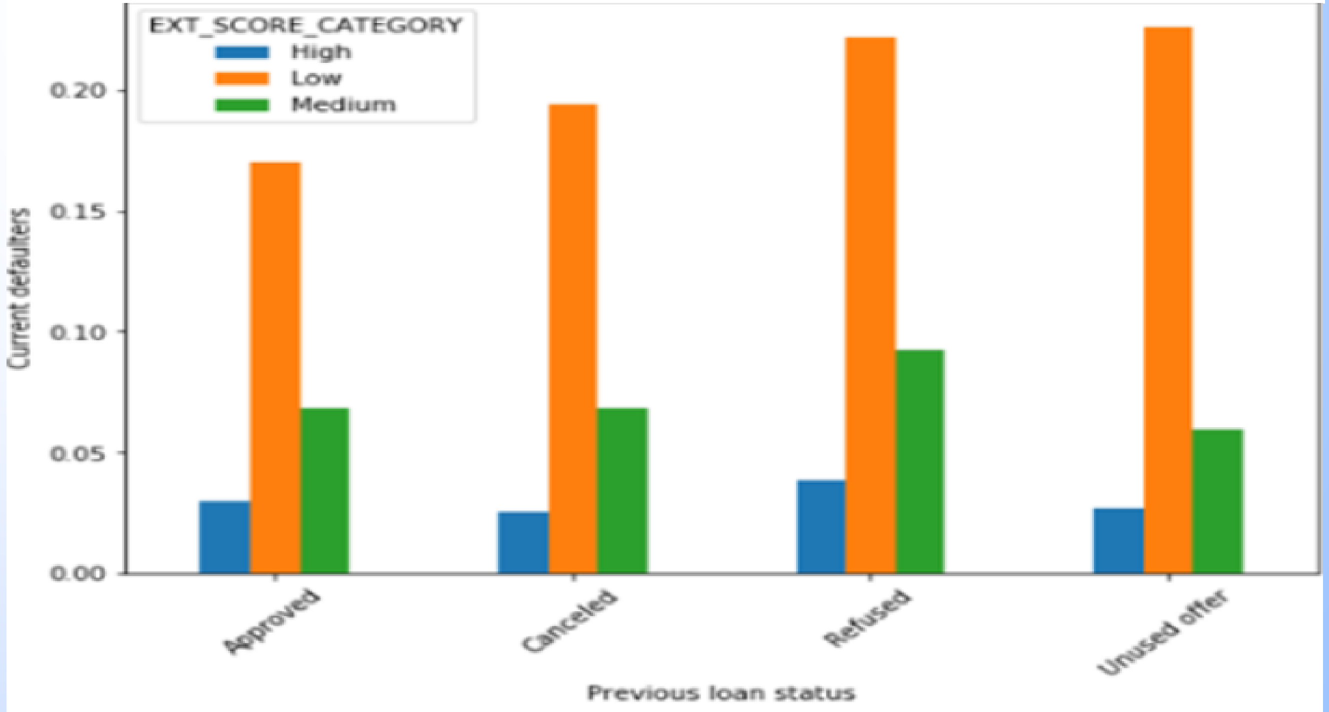
Observation

- The previous applications for portfolio Cards and POS are mostly defaulted .
- Previously refused applications for Cash are also defaulted in higher rate.
- Low external source scorer are highly defaulted irrespective of their previous loan status.

Recommendation

- It is safer to grant loans for any portfolio for previously approved applicants.
- It is high risk to grant loans for applicants, who have poor external source score specially whose loan were previously refused, unused or cancel.





Conclusion

Highly recommended groups:-

- Approved clients in their previous applications.
- Highly educated clients with higher income.
- Clients with higher external source score.
- Senior citizens in all categories.
- Married clients compared to other family status.
- Females are comparatively favorable than male.

High risk groups:-

- Previously refused, cancelled or unused offer clients.
- Low income groups with previously refused status.
- Unemployed clients.
- Poor external source scorer.
- Young clients are comparatively riskier than mid age clients and senior citizens.
- Lower secondary and secondary educated clients.