Credit EDA Case Study

Submitted by
Niharika Chiluka
DS C68

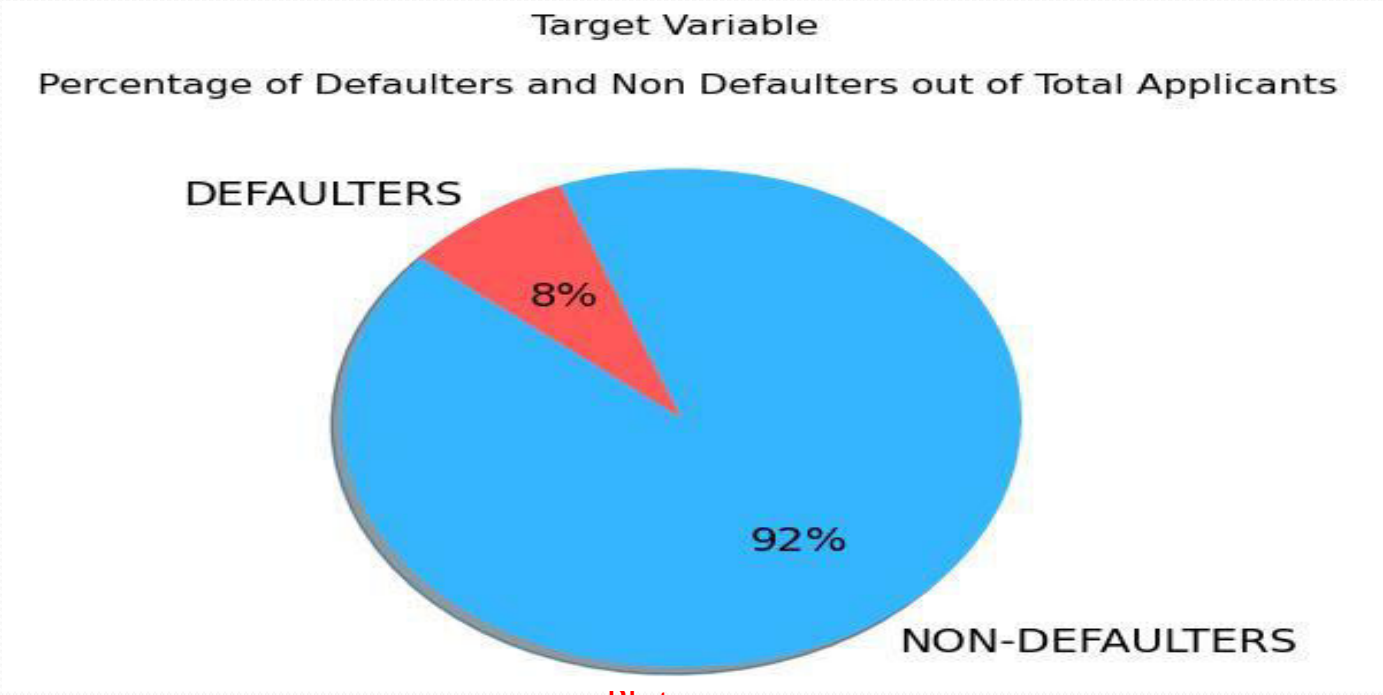
Problem Statement

- **Background**
  - A loan providing company which lends loans to the urban customers, processes loan application by verifying their capability to re-pay the loan.
- **Business Objective**
  - This case study aims to identify patterns which indicate if a client has difficulty paying their installments 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.

Understanding the data

- **Missing value check**
  - **For Application Dataset**
    - Removal of null values in data which are present more than 40%, as it can give outliers, which can result in incorrect analysis.
  - **For Previous Application**
    - Removal of null values in data which are present more than 30%, as it can give outliers, which can result in incorrect analysis.

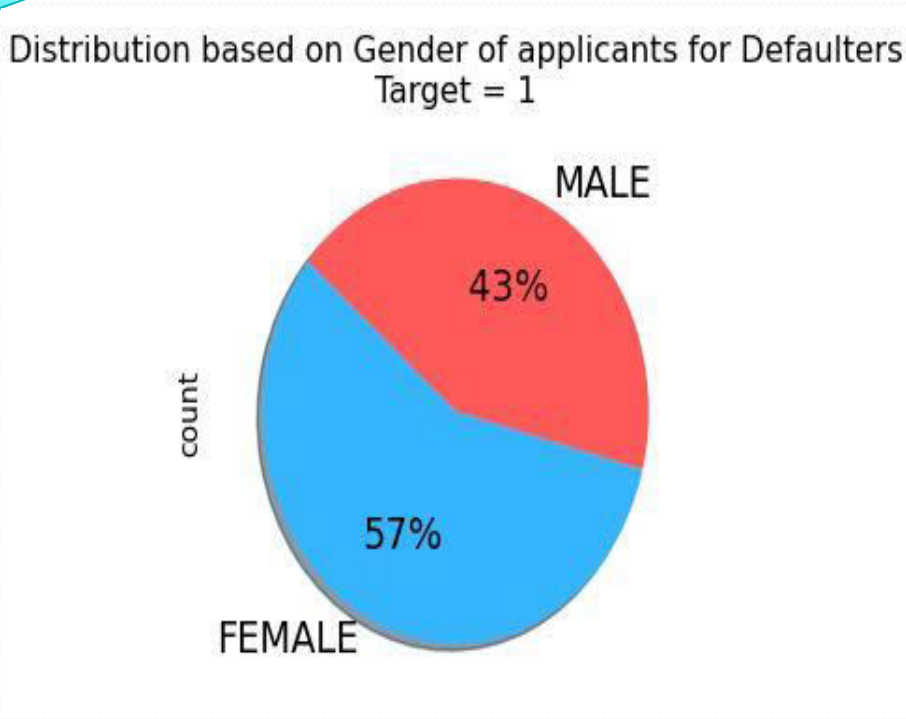
Imbalance Ratio

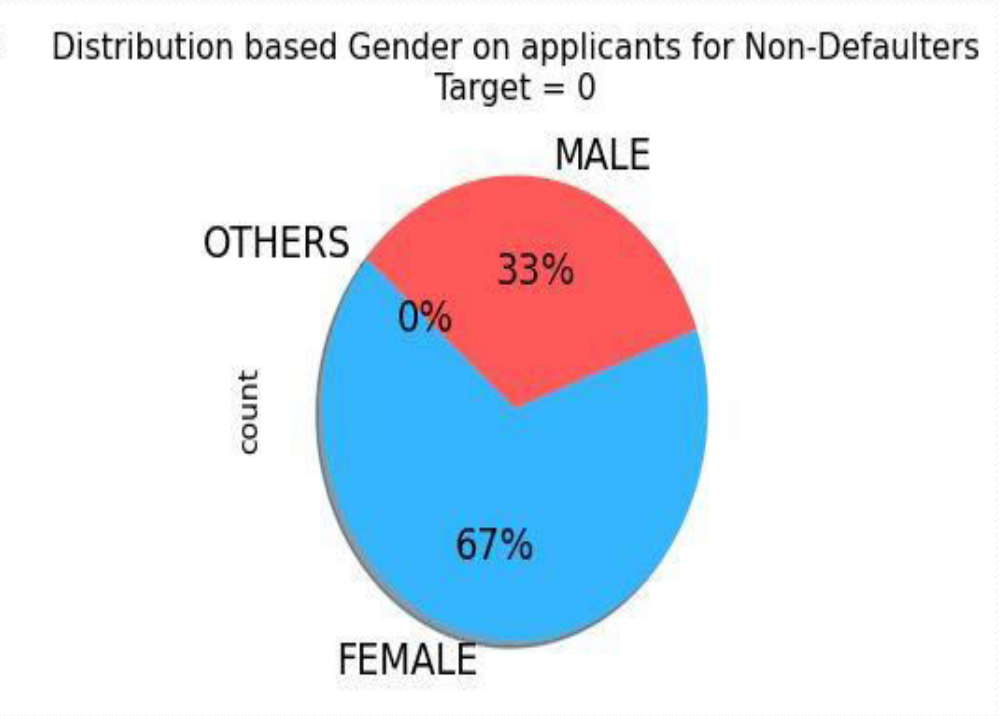


Plot: 1

The data after cleanup is highly imbalanced around data for loan defaulters (TARGET = 1) and remaining data for non-defaulters (TARGET = 0).

Univariate Analysis



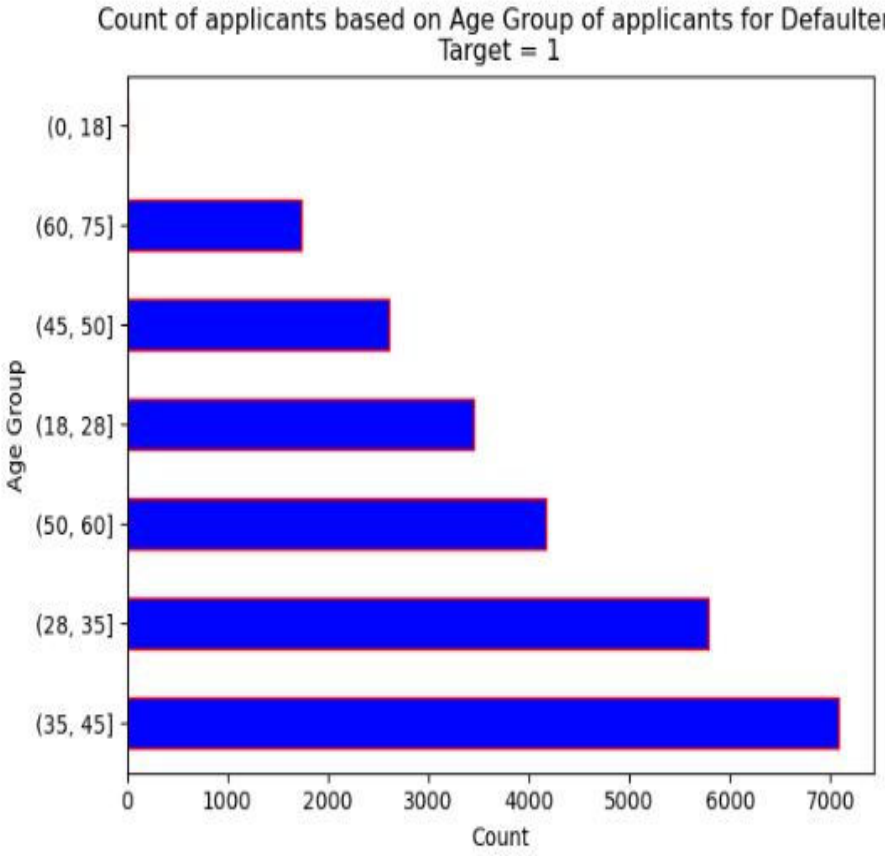


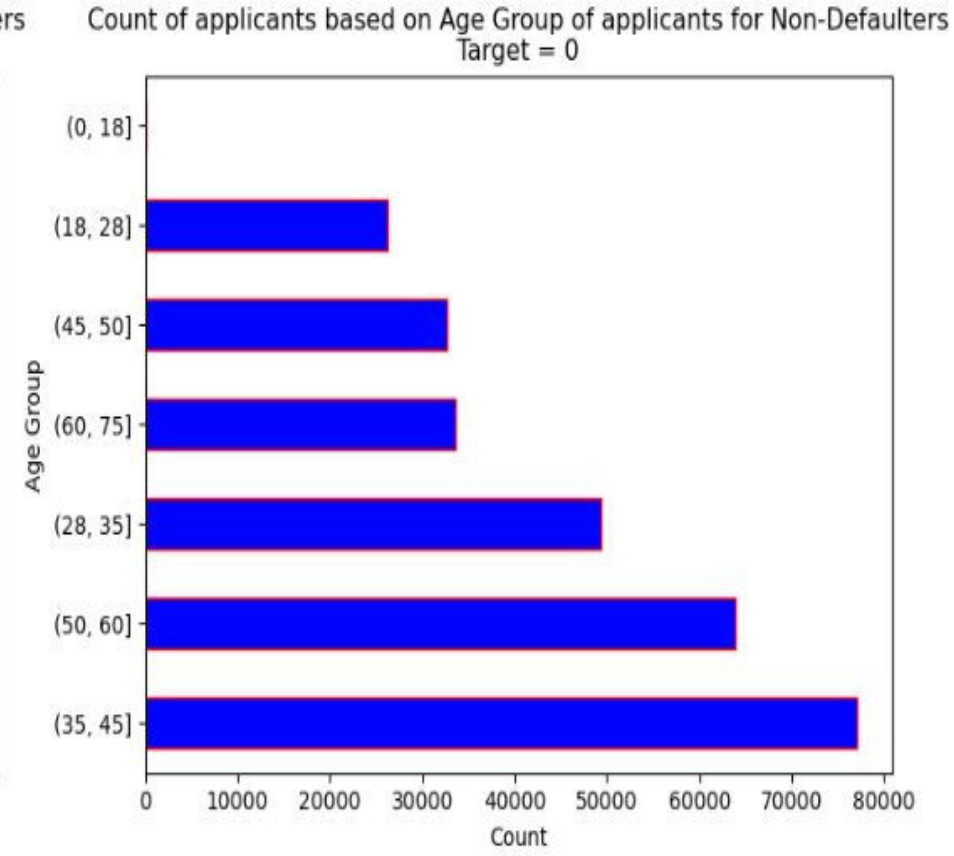
Plot: 2

Conclusion from the graph:

- Close to 60% of the applicants are Females in Defaulters
- Close to 70% of the applicants are Females in Non-Defaulters

It seems that there are more chances of women applicants being defaulters





Conclusion from the graph:

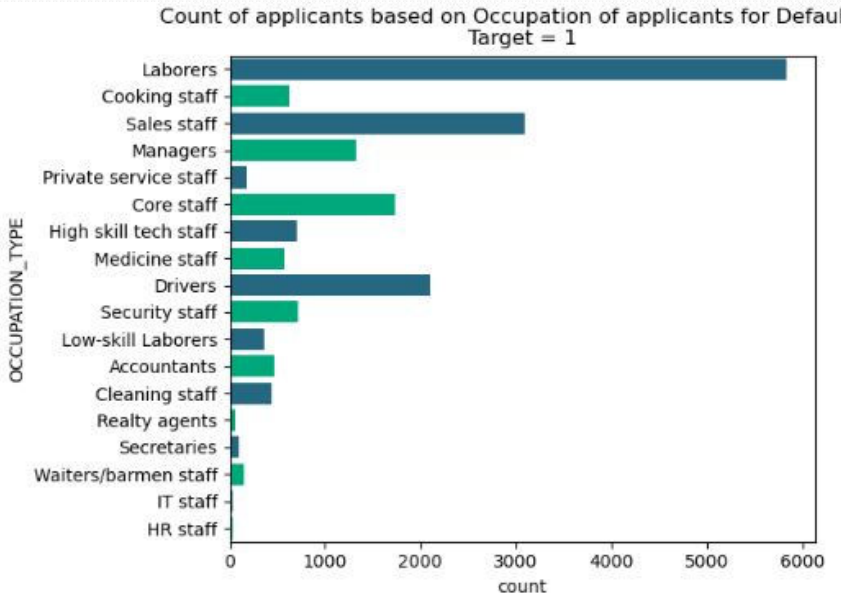
- Most of the applicants are falling under age group of 35-40 years in both defaulter and non-defaulter categories
- There are no applicant less than 18 years of age, i.e., no minor applicant

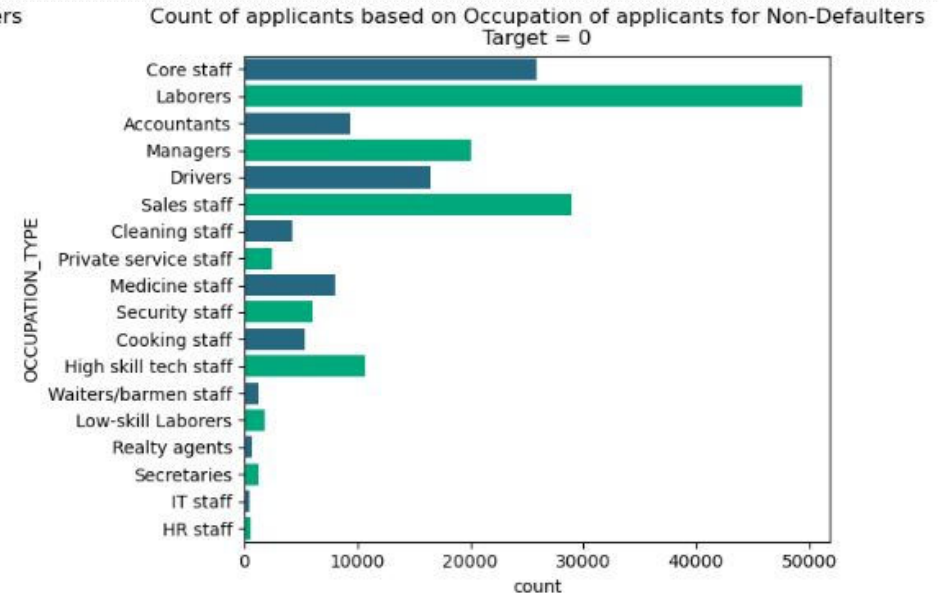
NON-DEFAULTER CATEGORY:

- 2nd highest applicants are of age group 50 - 60 years in non-defaulter category
- 3rd highest applicants are of age group 28 - 35 years in non-defaulter category

DEFAULTER CATEGORY:

- 2nd highest applicants are of age group 28 - 35 years in defaulter category
- 3rd highest applicants are of age group 50 - 60 years in defaulter category

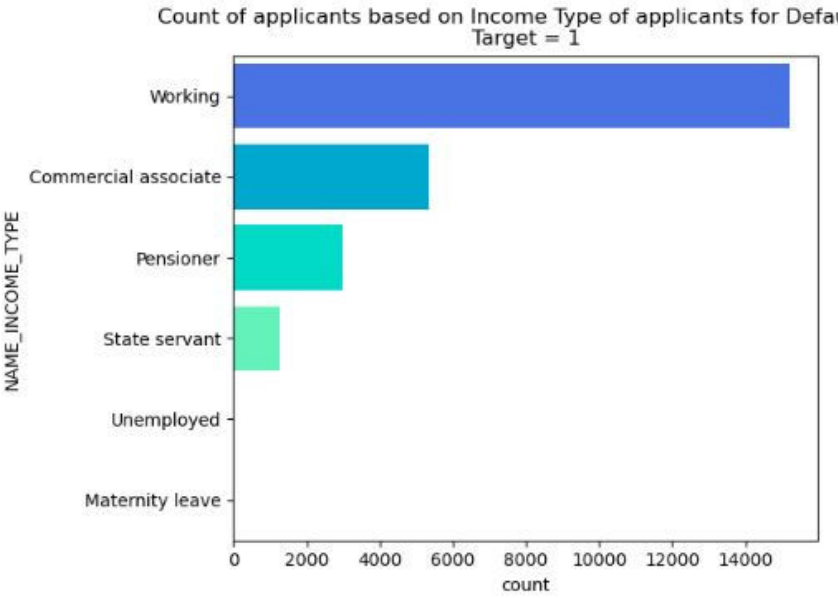


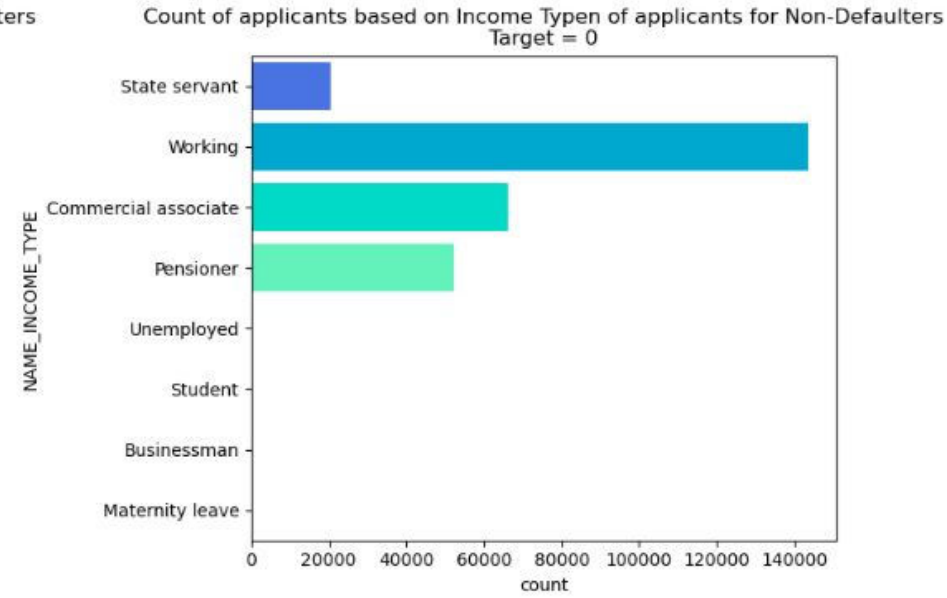


Plot: 4

Conclusion from the graph:

- Most of the applicants belong to Laborer as Occupation

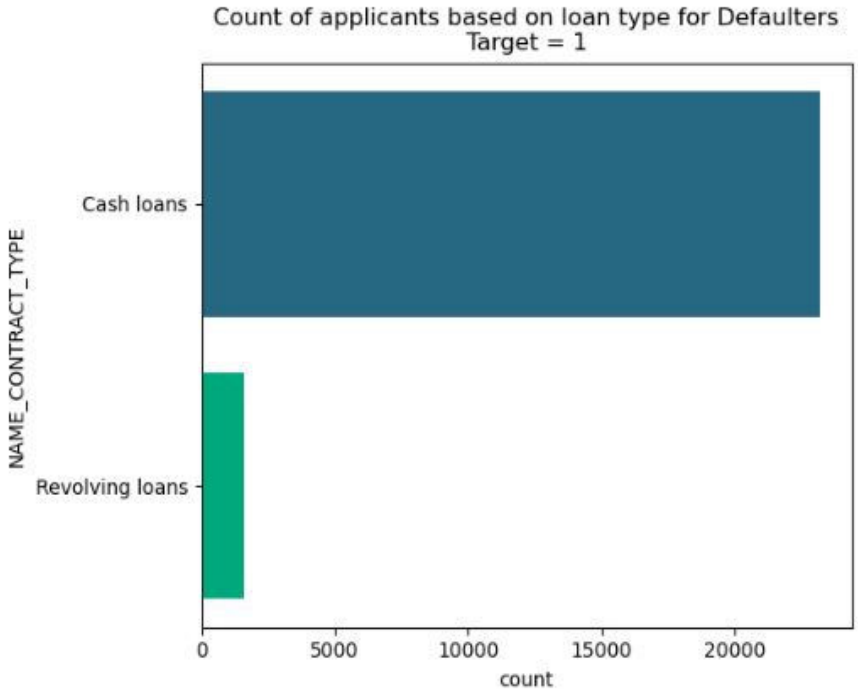


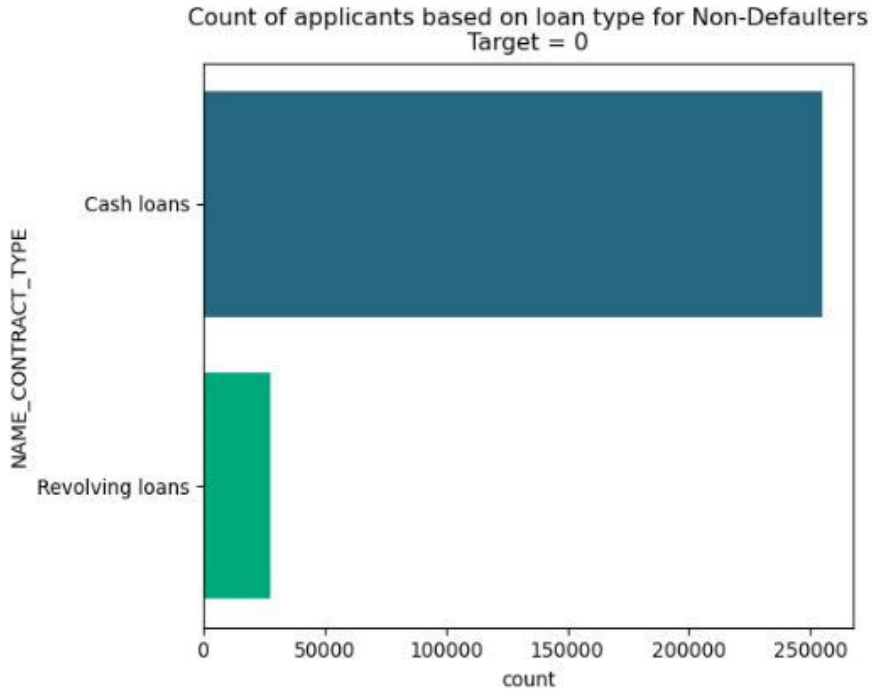


Conclusions from the graph:

- We can notice that the students are falling in non-defaulters. The reason could be they are not required to pay during their college tenure.
- We can also see that the Businessmen's are falling in non-defaulters.
- Most of the loans are distributed to working class people
- Pensioners are also good in number for applying loans and mostly they are non-defaulters as we can see in plots

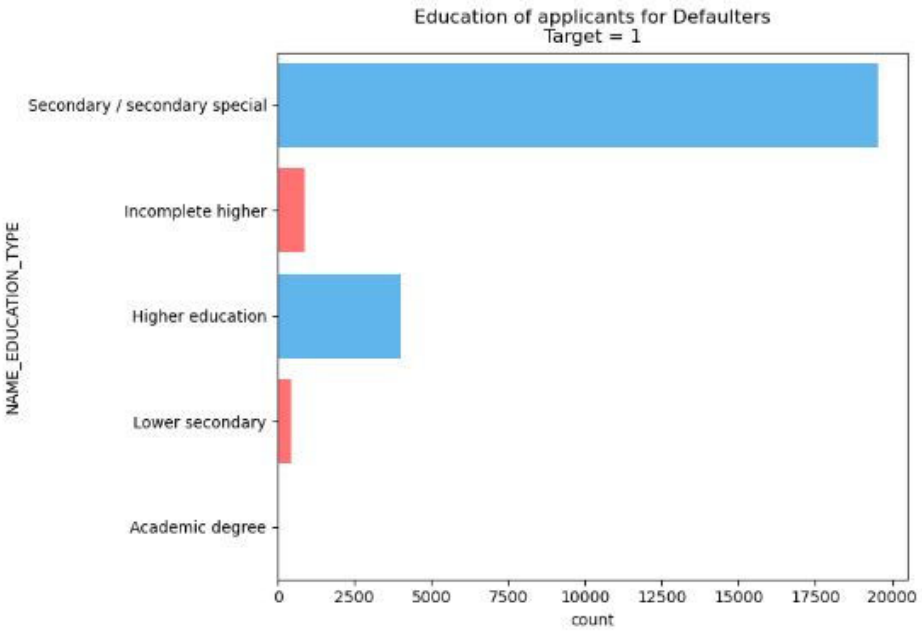
Plot: 5

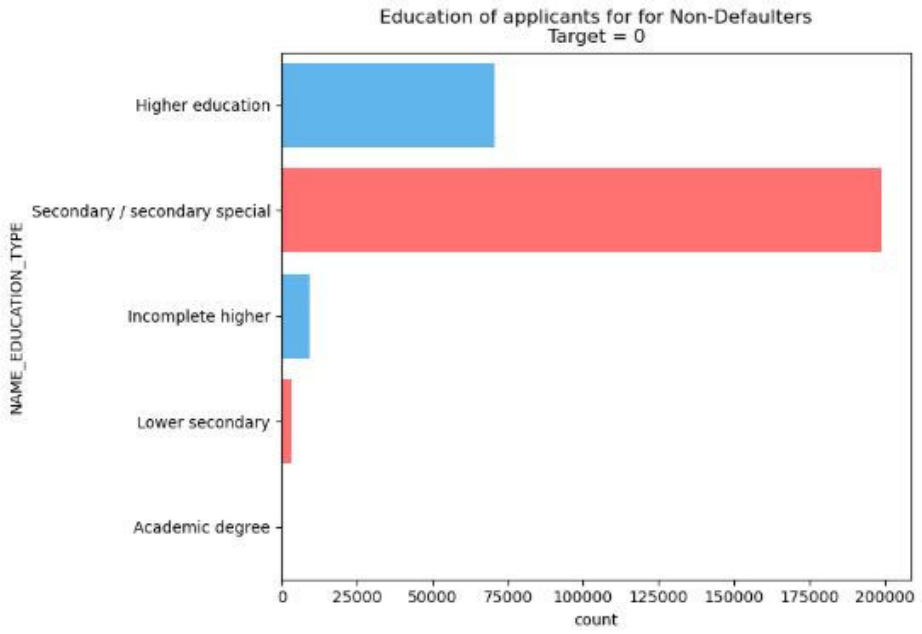




Conclusions from the graph:

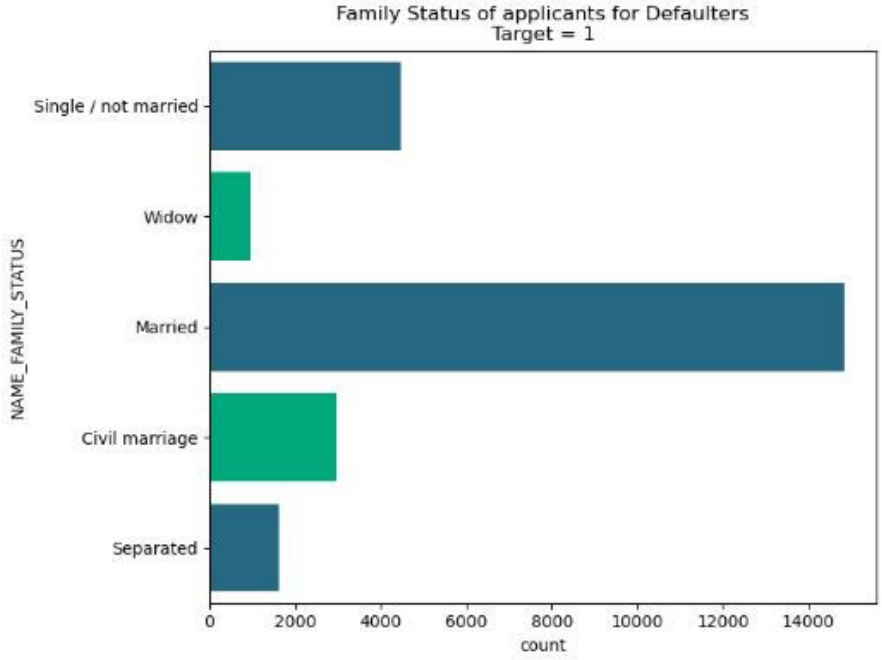
- Most of the loan types are Cash Loans
- Very less are Revolving loans

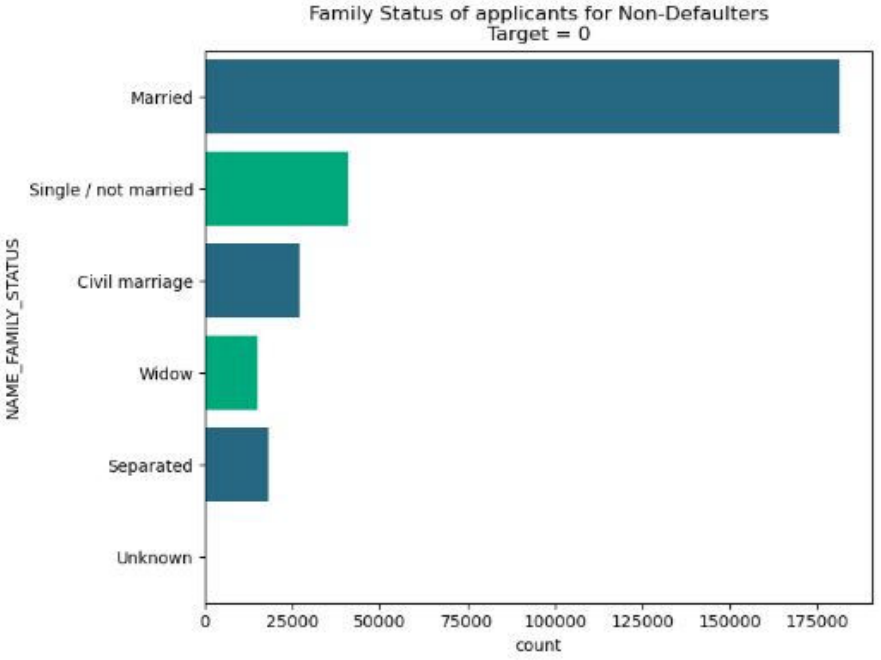




Conclusions from the graph:

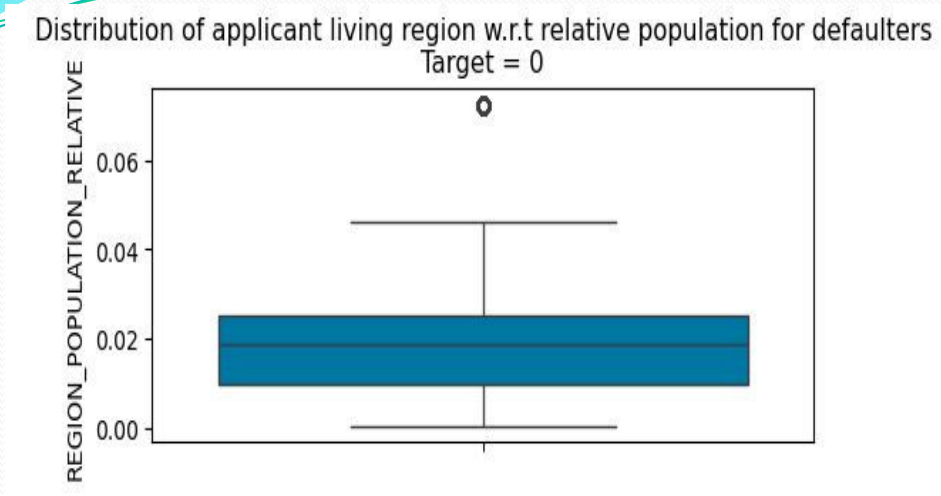
- Most of the the applicants have completed Secondary Education in both defaulter and non-defaulter categories
- Secondly many of the the applicants have completed Higher Education in both defaulter and non-defaulter categories
- Academic degree holders are almost neglible in number in both defaulter and non-defaulter categories

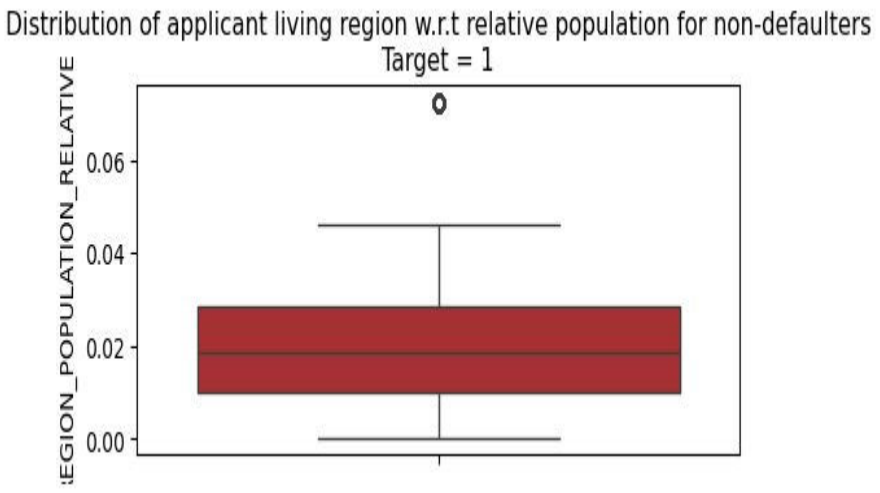




Conclusions from the graph:

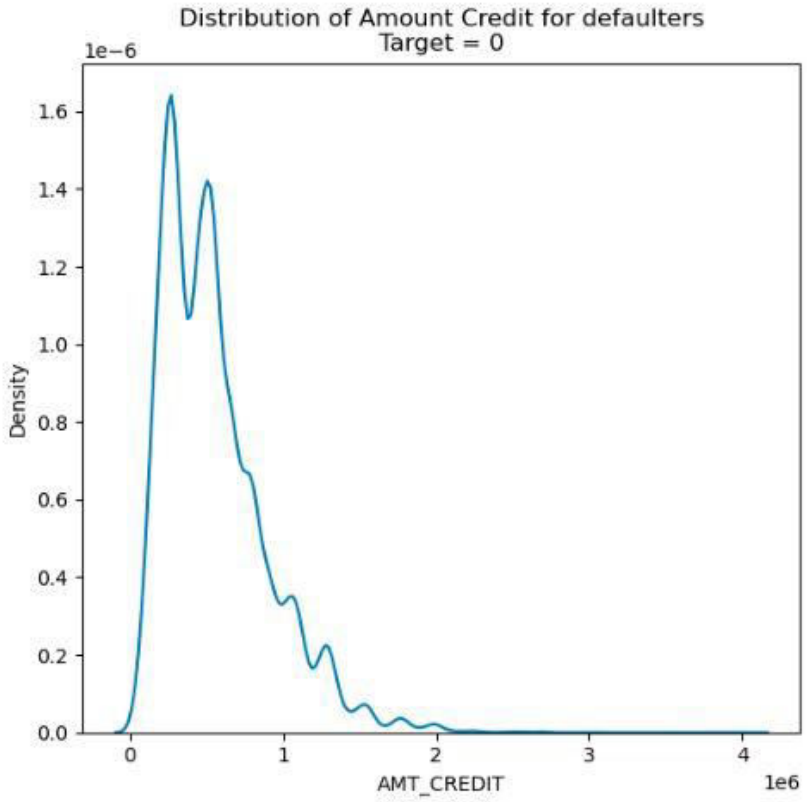
- Most of the the applicants are married
- 2nd highest applicants are Single/Non-married
- Most of the the applicants are married in both defaulter and non-defaulter categories
- Secondly many of the the applicants are Single/Non-married in both defaulter and non-defaulter categories

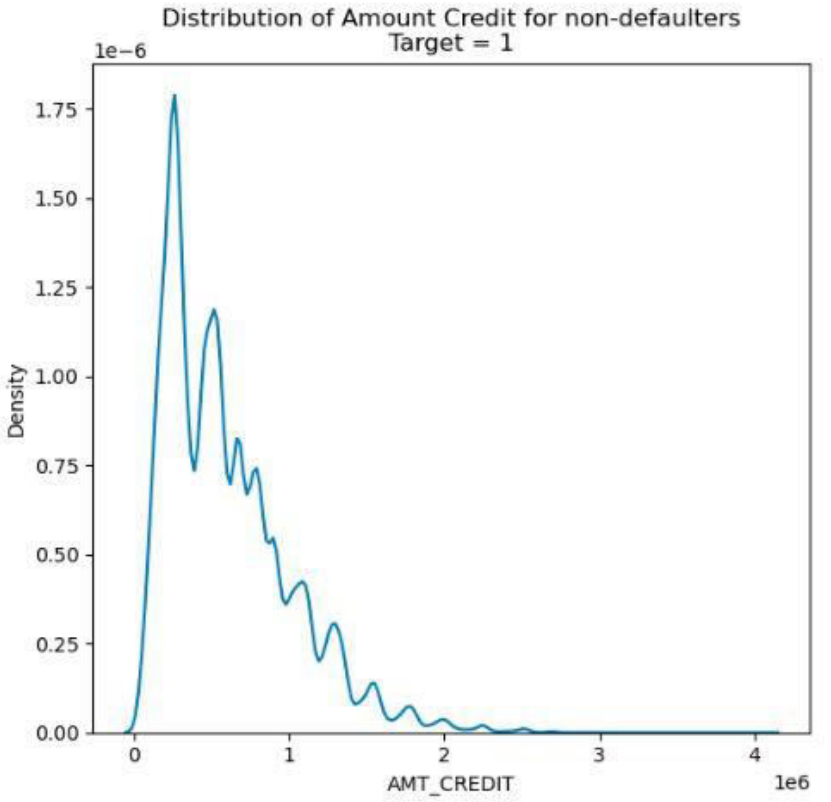




Conclusions from the graph:

- By looking at above box plot, we can see there is an outlier in this, which means one applicant is living in highly dense populated area
- For all other applicants, the trend falls within IQR and mean & median are almost same. This means that most population is living in average to medium dense populated areas

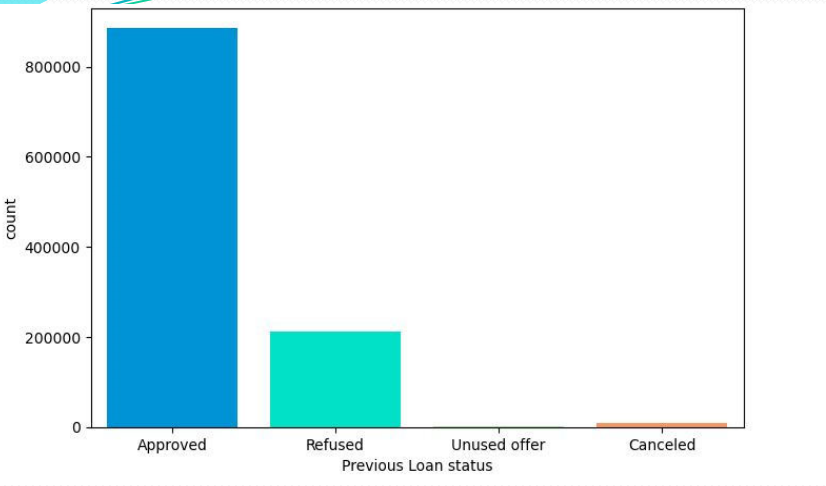


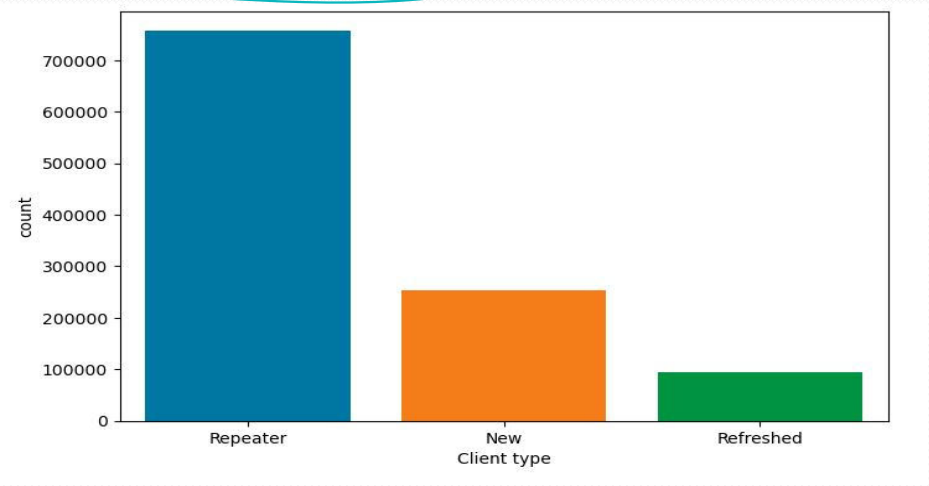


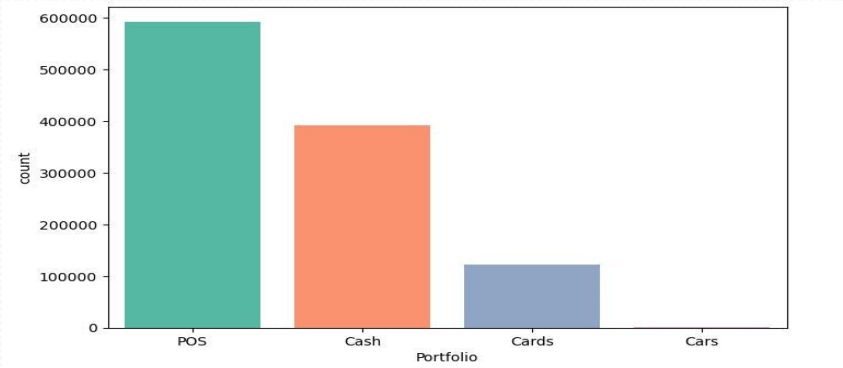
Conclusions from the graph:

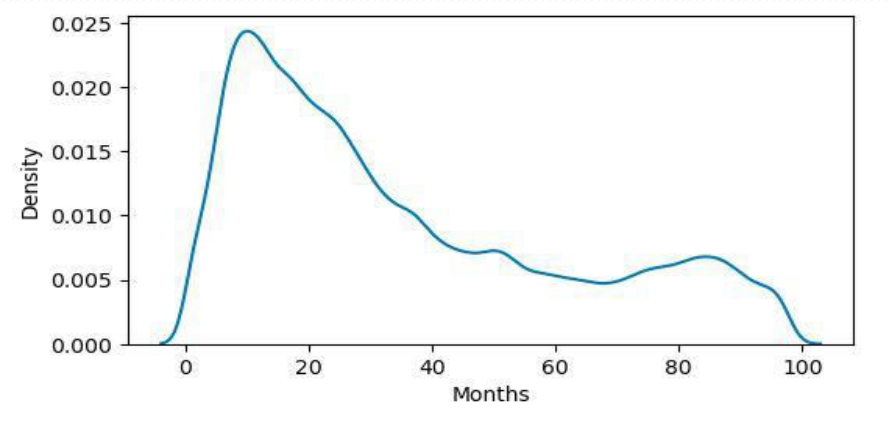
- Defaulters - We can notice that the lesser the credit amount of the loan, the more chances of being defaulter. The spike is till 500000.
- Non defaulters - If the credit amount is less, there is lesser chance of being defaulted. And gradually the chance is being decreased with the loan credit amount.

Univariate Analysis of merged data





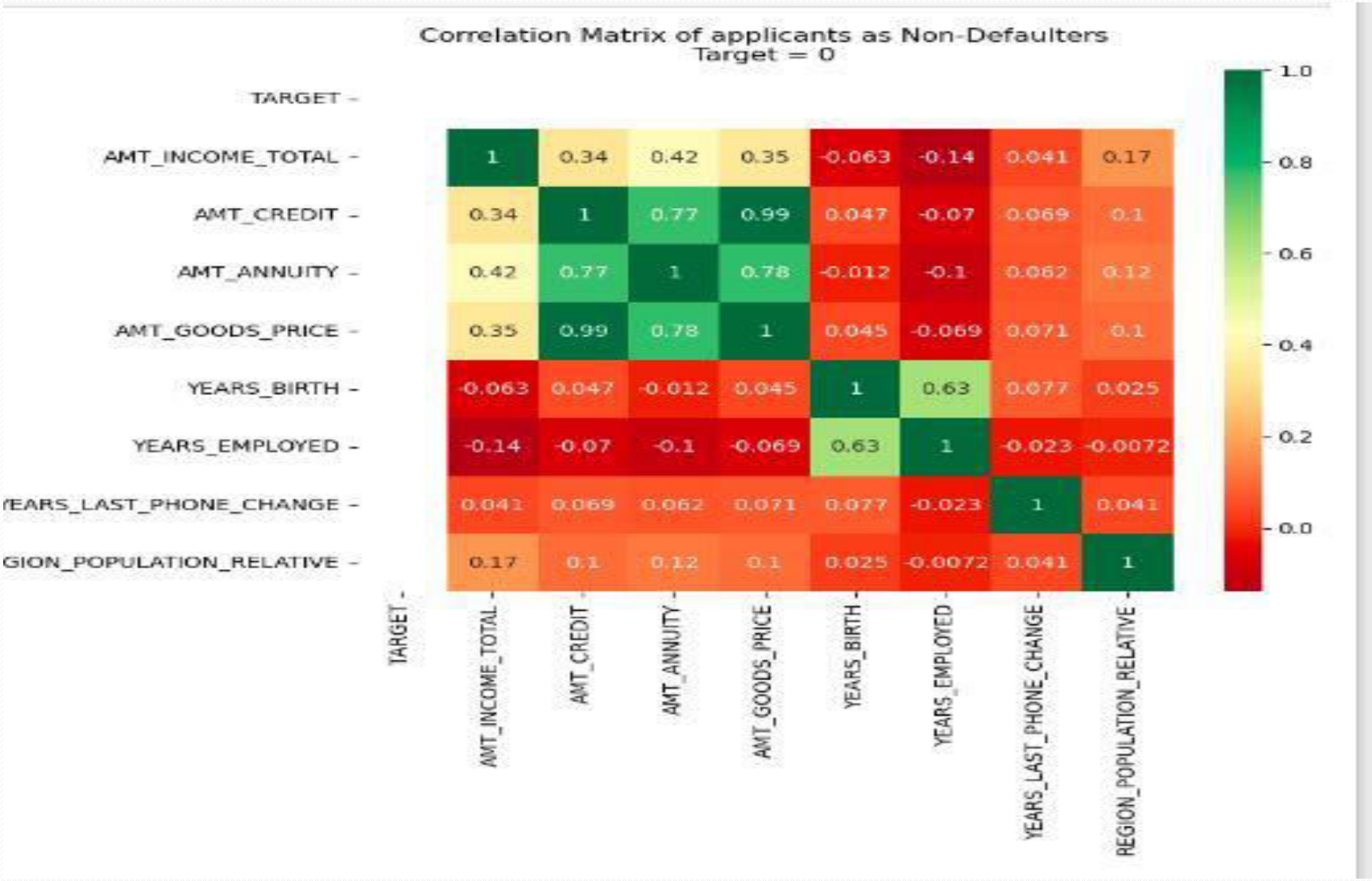




Conclusions from the graph

- There are huge number of Approved loan than Refused. Hardly, there are any Canceled or Unused offer loan.
- Mostly the applicants were Repeater
- The highest number of the previous applications was for POS. Applications for Cash also has good number. Applications for Cards were very few.
- We can see that most of the applications decision took approximately 30 months. The time taken spread upto 100 months.

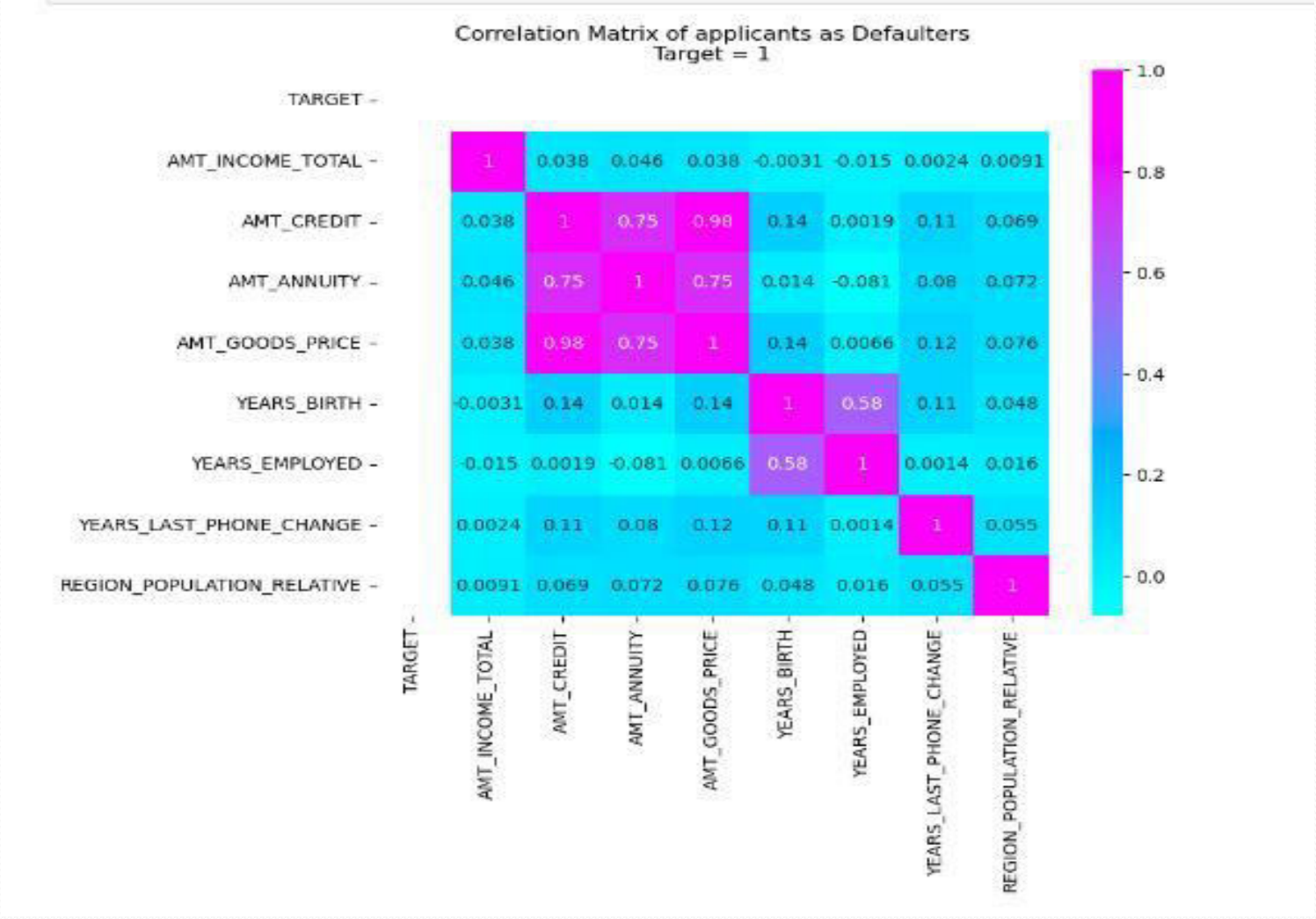
Correlation for Target = 0



Plot: 10

In Continuation...

Correlation for Target = 1



in continuation...

Conclusions from Correlation matrix for Target 0 and Target 1

Highly correlate columns for defaulters

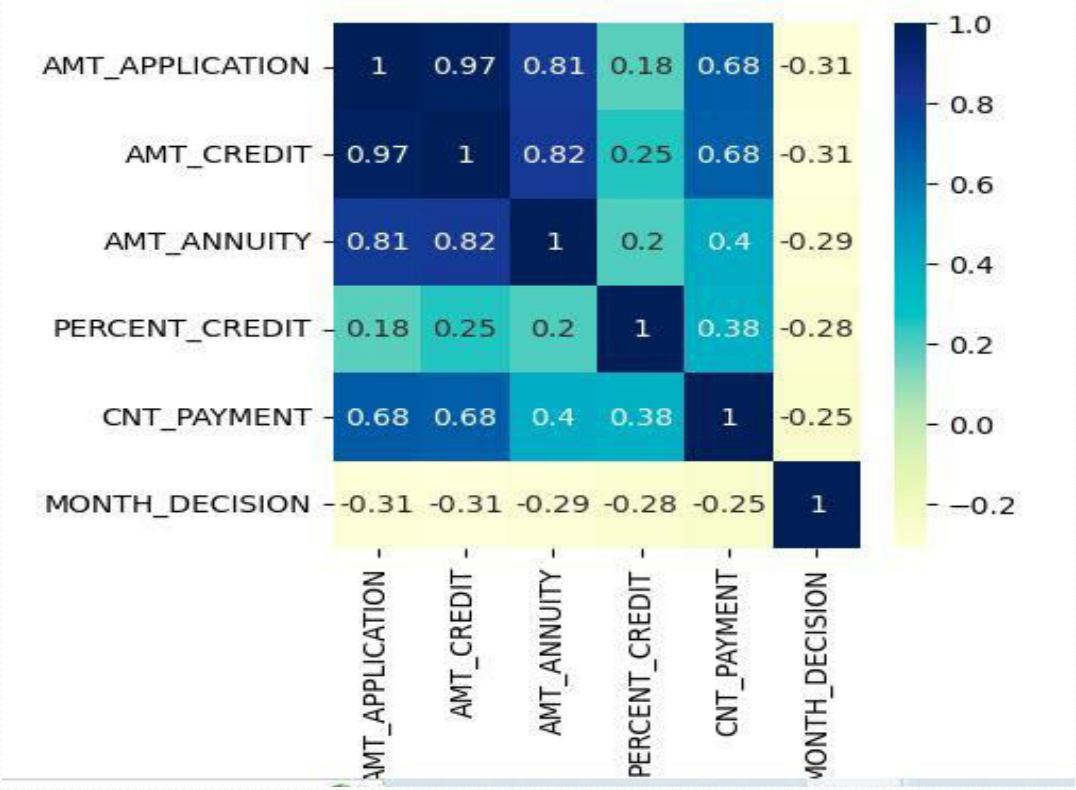
- AMT_CREDIT and AMT_ANNUITY (0.74)
- AMT_CREDIT and AMT_GOODS_PRICE (0.98)
- AMT_ANNUITY and AMT_GOODS_PRICE (0.74)

Highly correlate columns for non defaulters

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

Through this correalation ,using heatmap. We can conclude that correaltions are almost same while seeing Target 0 & 1

Correlation of merged data



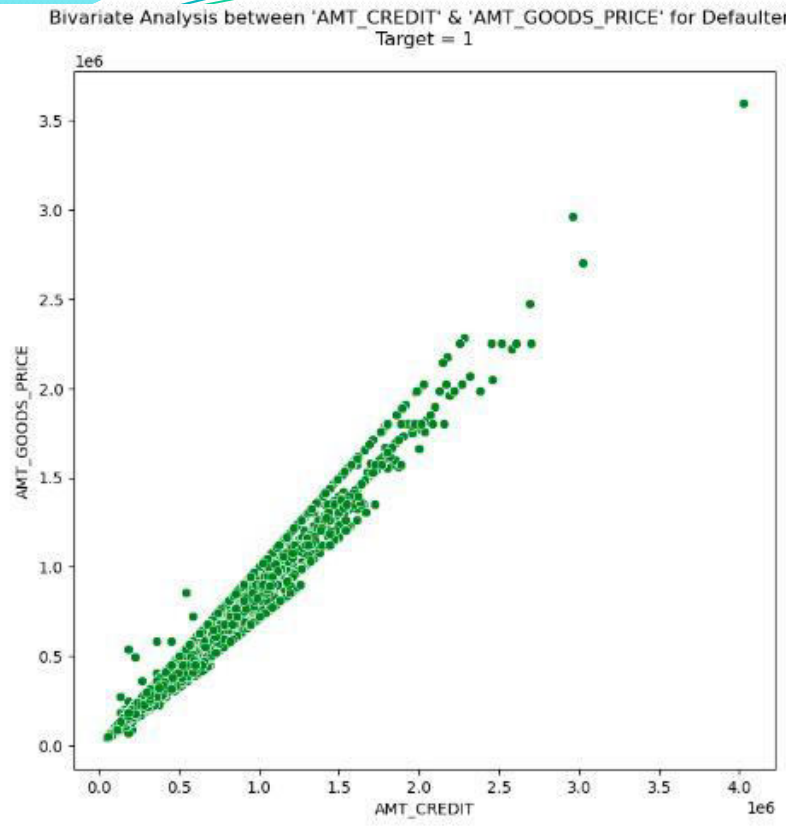
Highly corelate columns

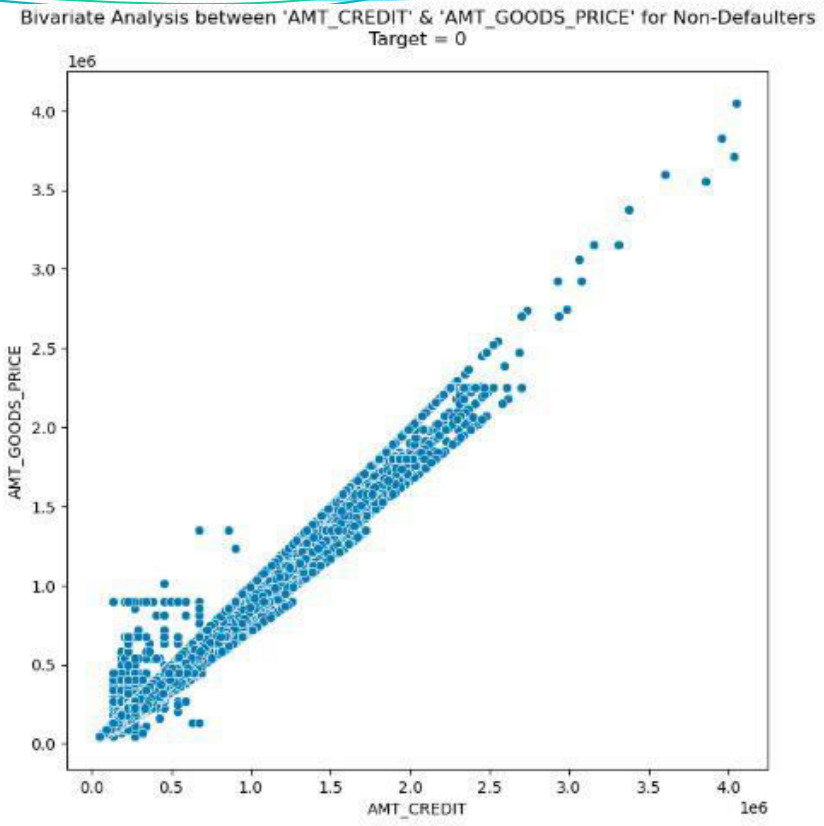
- AMT_APPLICATION and AMT_CREDIT
- AMT_APPLICATION and AMT_ANNUITY
- AMT_CREDIT and AMT_ANNUITY

Moderately corelated columns

- AMT_APPLICATION and CNT_PAYMENT
- AMT_CREDIT and CNT_PAYMENT

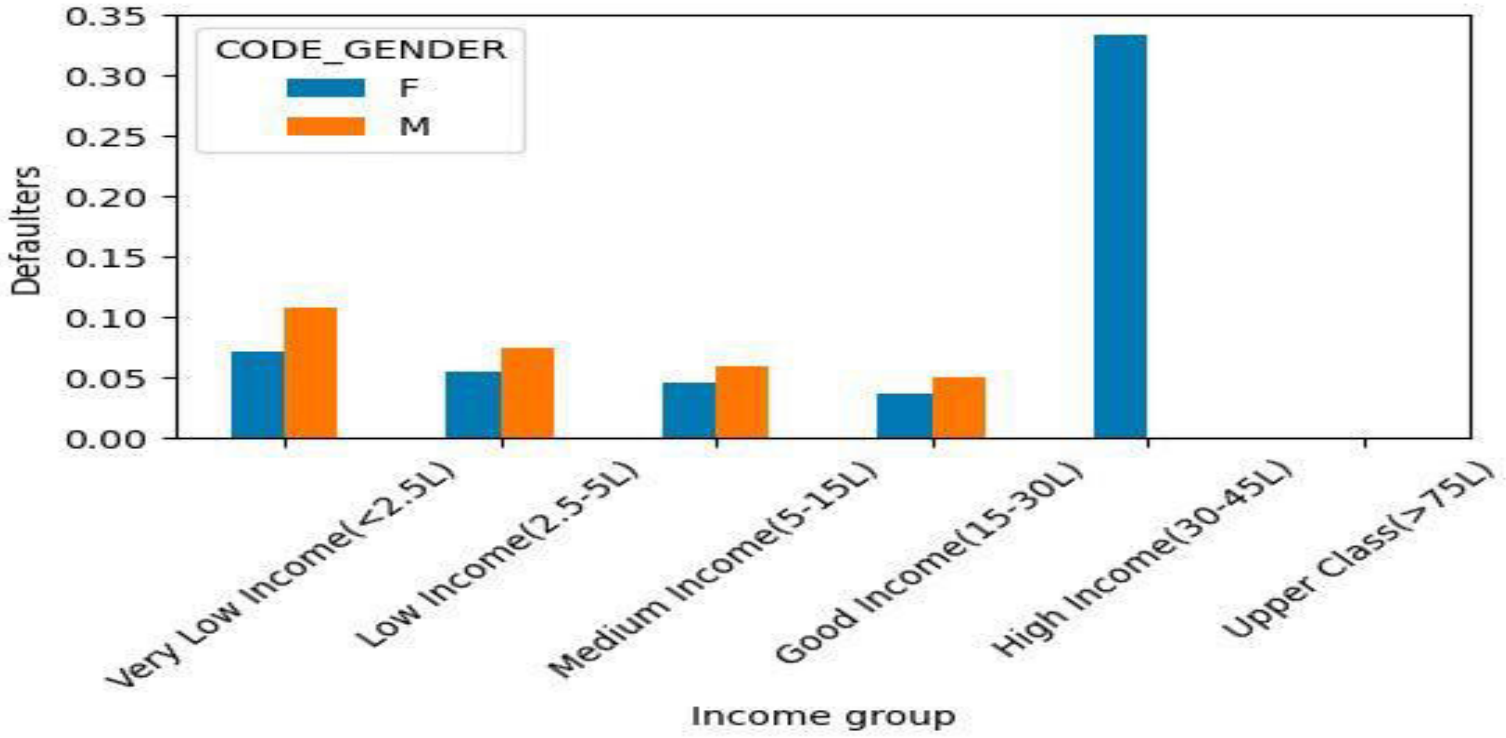
Bivariate Analysis





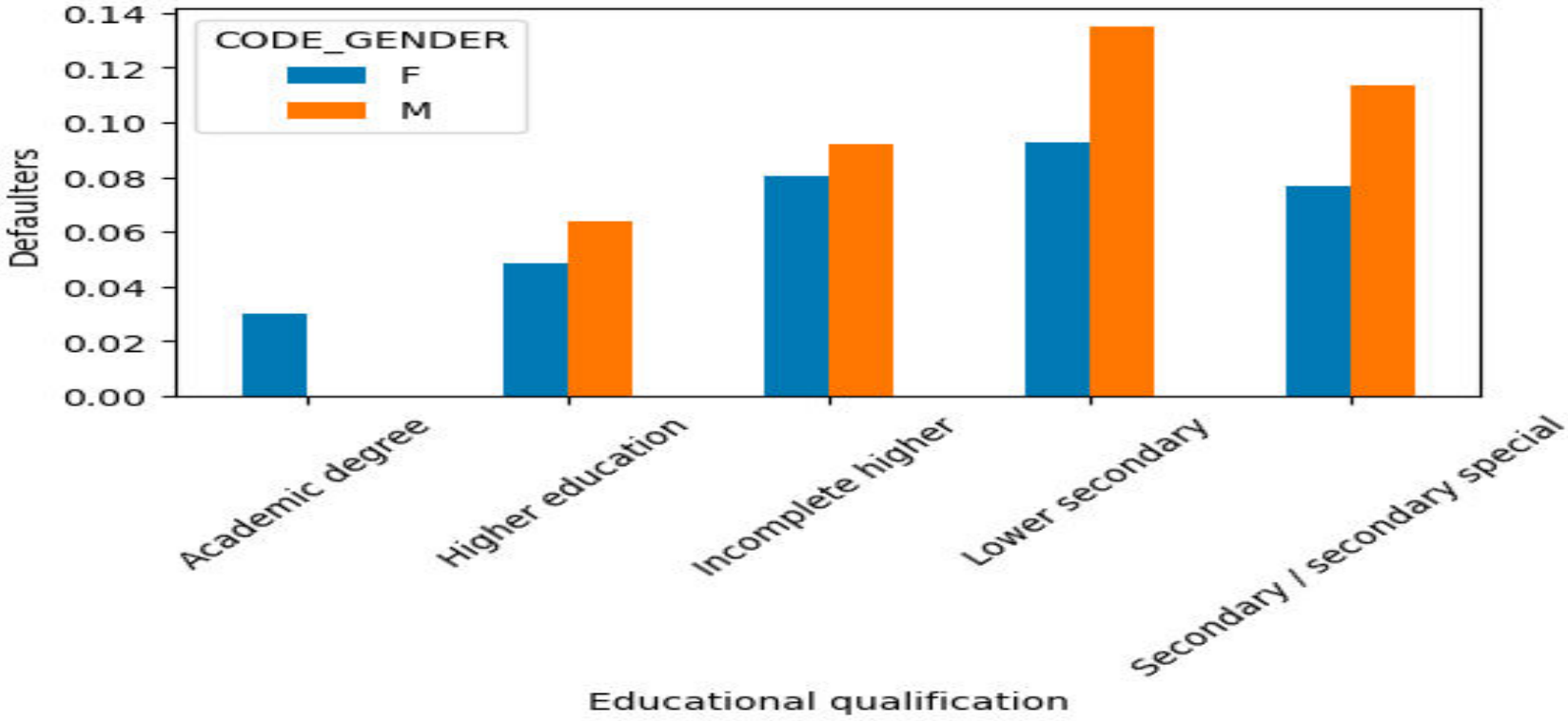
Conclusions from the graph:

- Amount Credit and Amount of Good price are showing same trend and which is mostly true as credit amount may be same or less than goods price



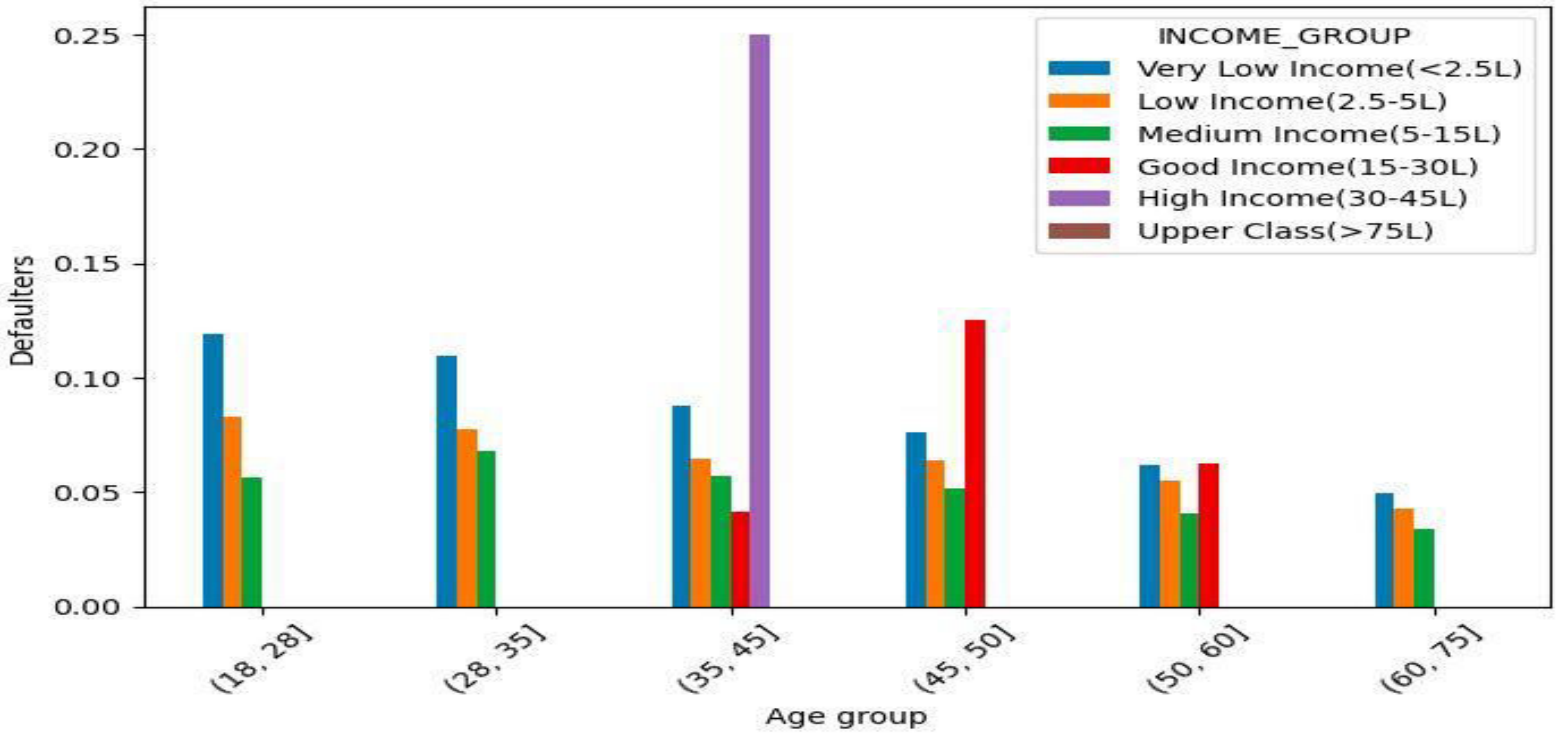
Conclusion from graph:

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



Conclusions from the graph:

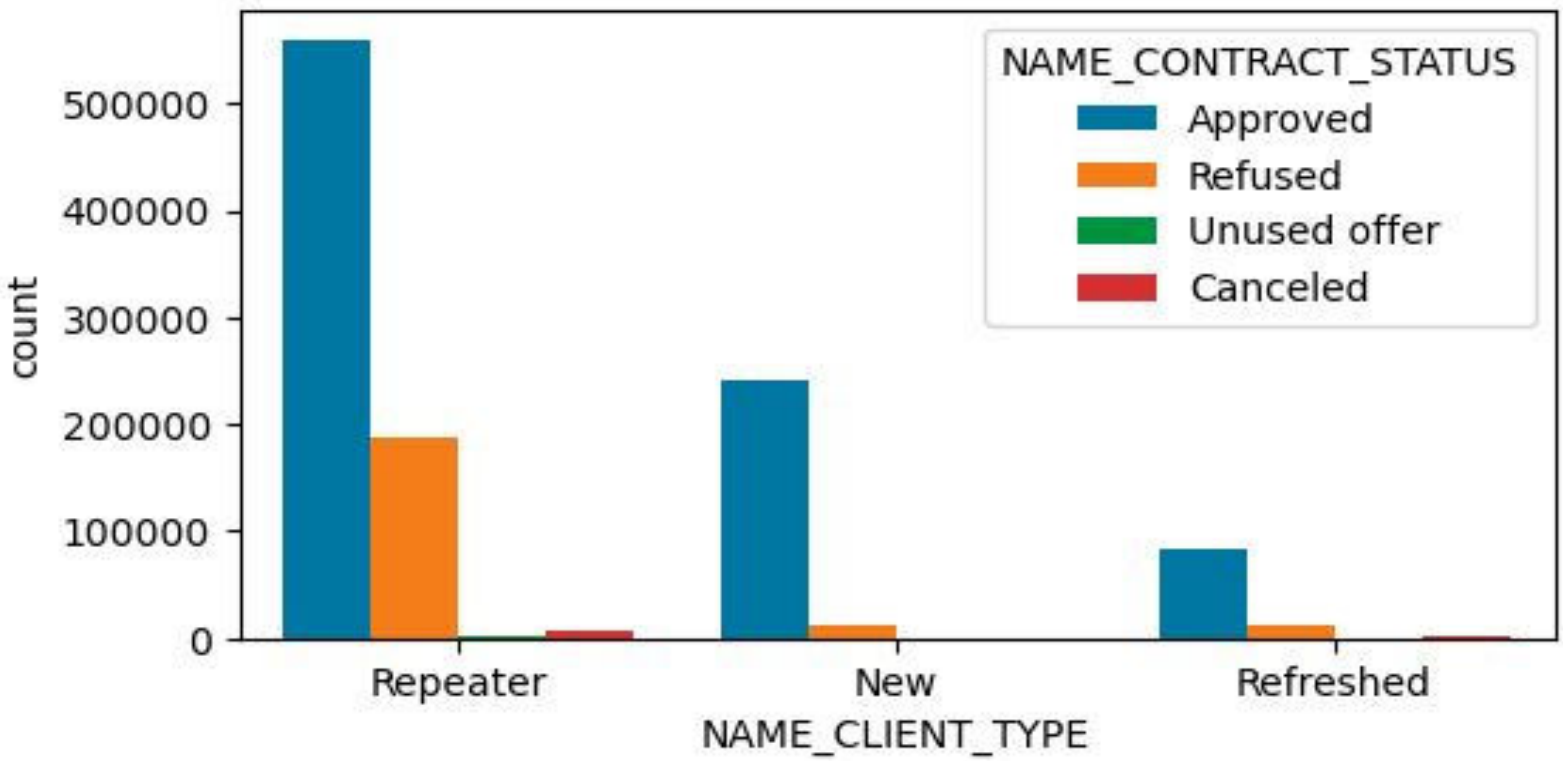
- Lower secondary educated clients are more defaulted followed by Secondary and Incomplete higher educated clients.
- The Higher educated people are less defaulted.
- Across all educated level Females are less defaulted than male.



Conclusions from the graph:

- 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.

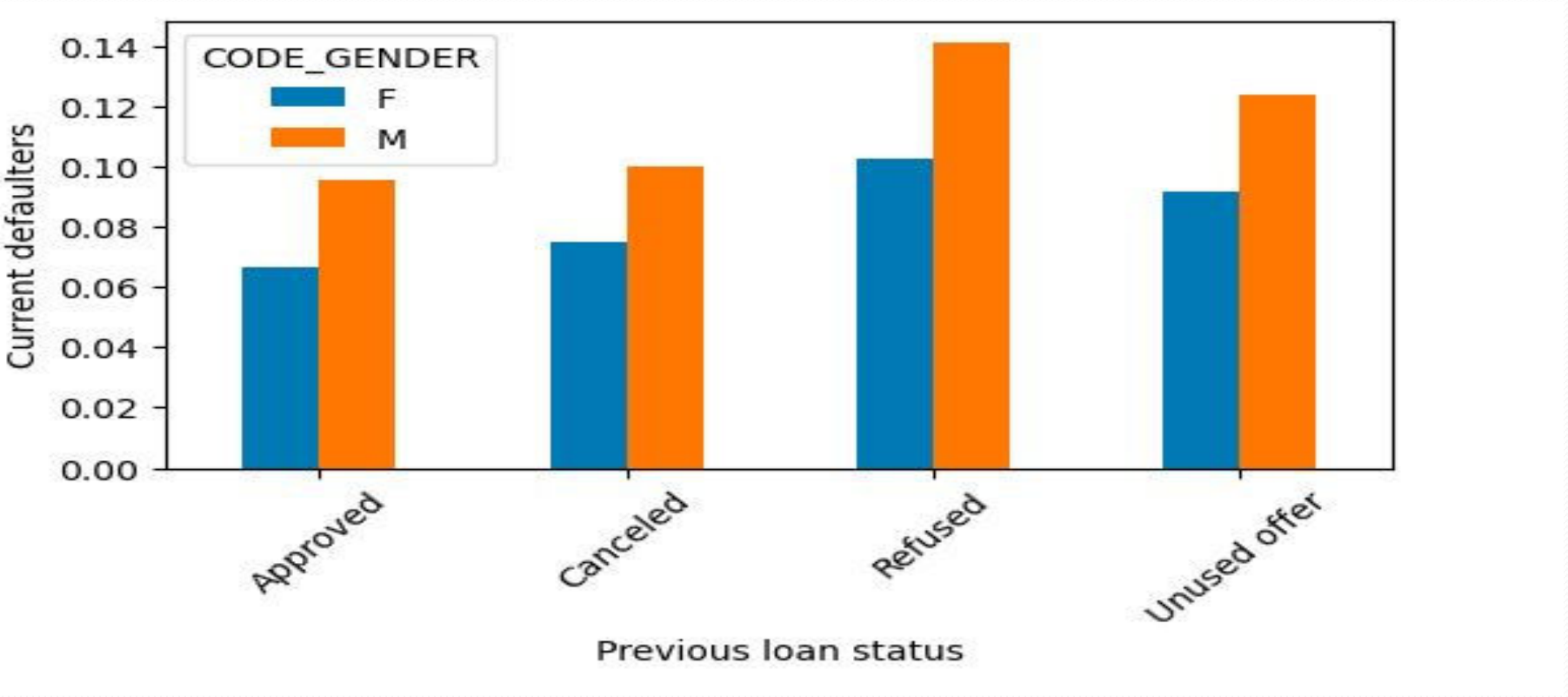
Status and Client type of merged data



Conclusions from the graph:

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

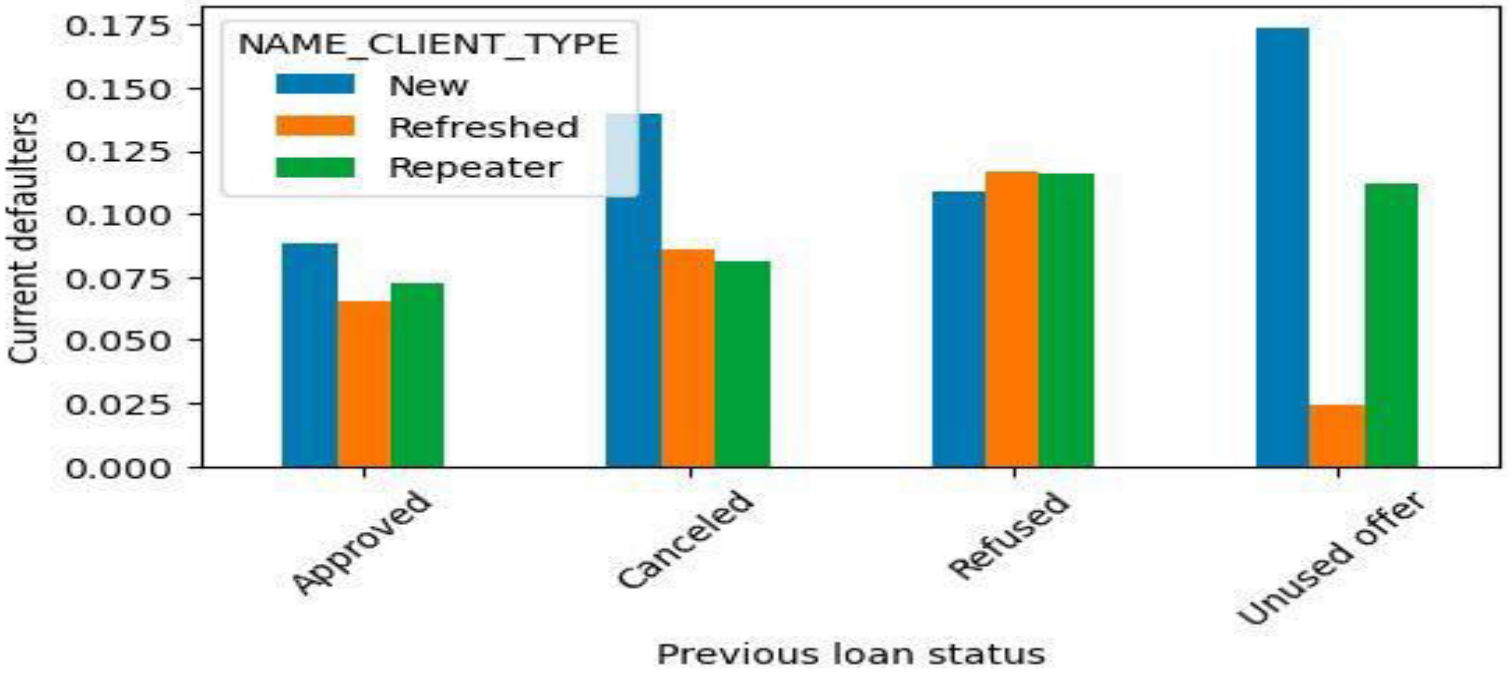
Current loan defaulter status with respect to previous loan application status



Conclusions from the graph:

- 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.

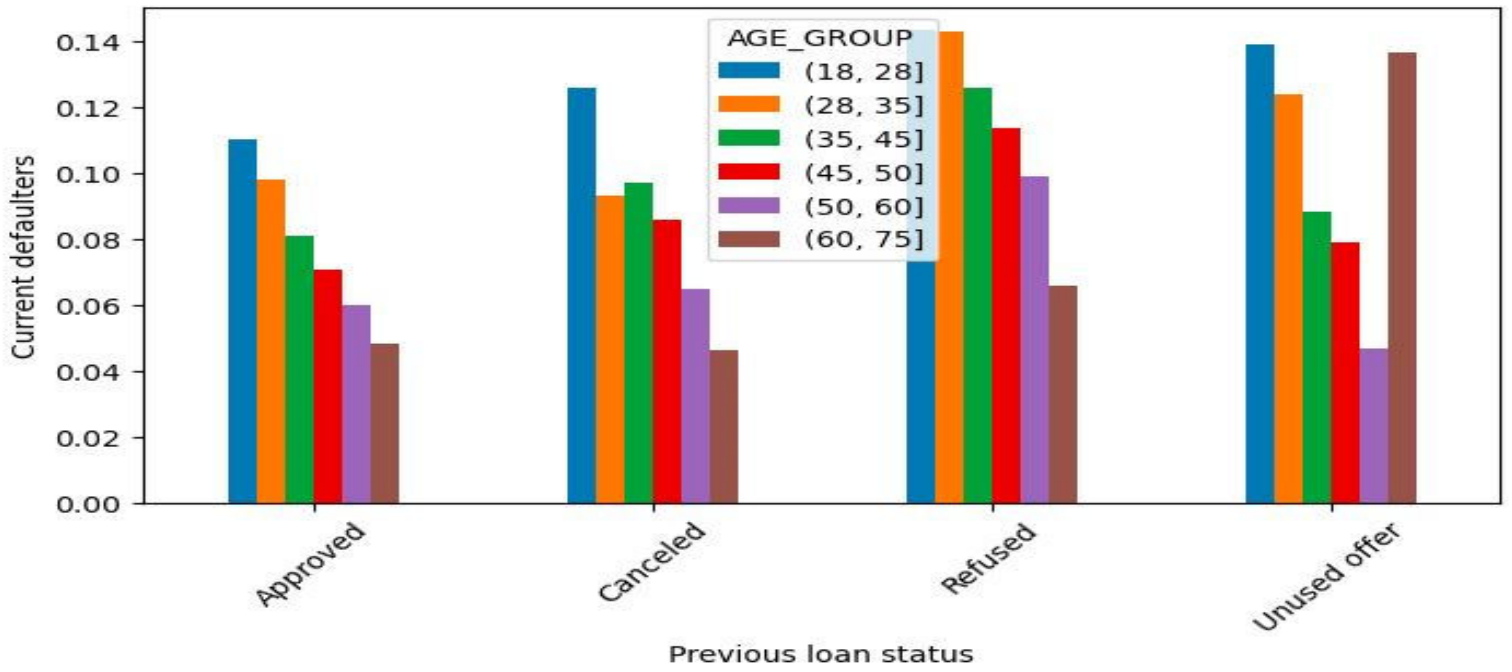
Current loan defaulter status with respect to previous loan application status and client types



Conclusions from the graph:

- 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.

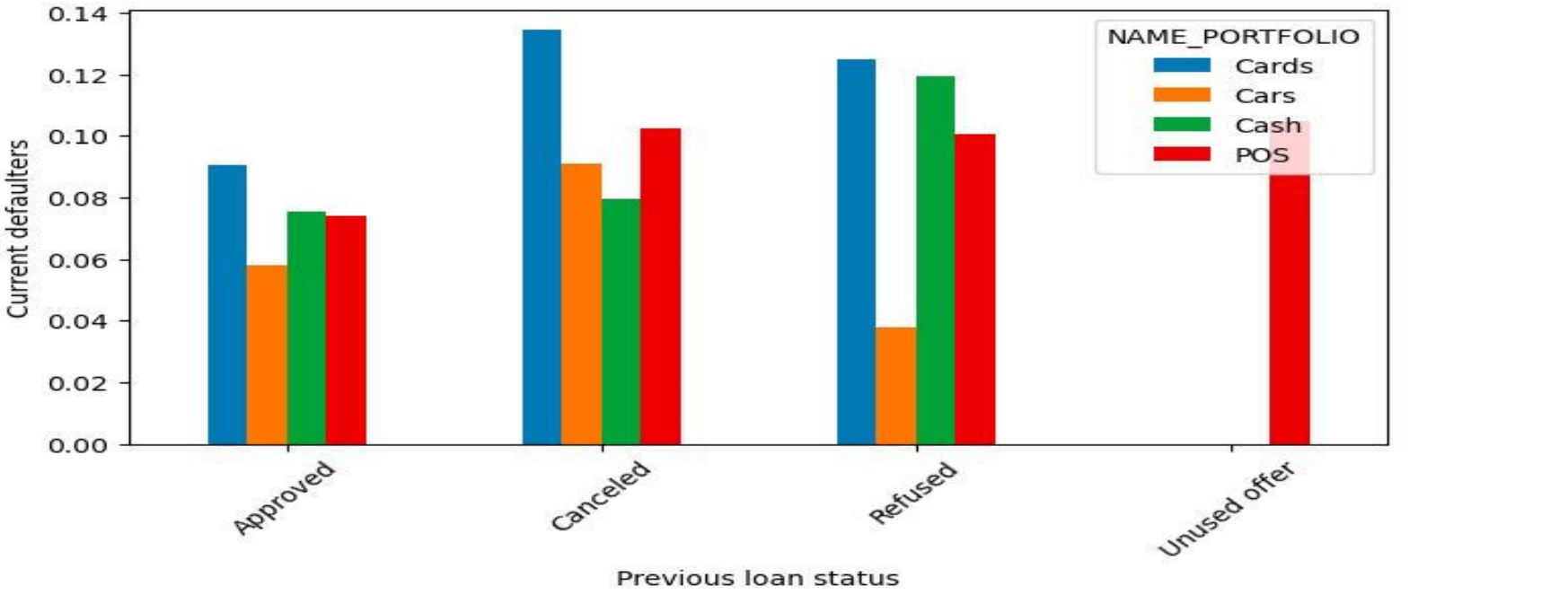
Current loan defaulter status with respect to previous loan application status and age group



Conclusions from the graph:

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

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



Conclusions from the graph:

- 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.

Conclusion on the analysis of the data

- **Banks should approve loans more for Office apartment, Co-Op apartment housing type as there are less payment difficulties.**
- **Banks should provide loans to 'Repairs' & 'Others' purposes.**
- **Banks should provide loans to the 'Business Entity Type-3' and 'Self-Employed' persons.**
- **'Working' people especially female employers are the best to target for the loans.**