# BANK LOAN ANALYSIS

In this project we are working with the dataset of loan providing company which specialises in lending various types of loans to urban customers. In this Project we have learnt the Practical application of EDA in a business environment. In this Project we have also gain a basic grasp of risk analytics in banking and financial services, as well as how data is utilized to reduce the risk of losing money when lending to consumers.

Approach:

The dataset provided is very large and have many unneeded columns that have many missing values and some doesn't have any relation to our analysis. So we first started with deleting all those columns which have more than 40percent of missing values and then began performing univariate and bivariate analysis using various type of graph.

**Problem Statement:**

The main Problem for any loan providing companies is the defaulters due to their insufficient or non-existent credit history companies find it hard to give the loan to people. 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.

So for identify Patterns which indicates whether the particular client is likely to be s defaulter or not we will apply EDA to understand how consumer attributes and loan attributes influence the tendency of default.

We are provided with two enormous data set "application data" (contains all the information of client's at the time of application) and "previous data" (contains the data of the client's previous loans). Both sets of data contained many undesired columns that will not be used for risk analytics, as well as many blank columns.

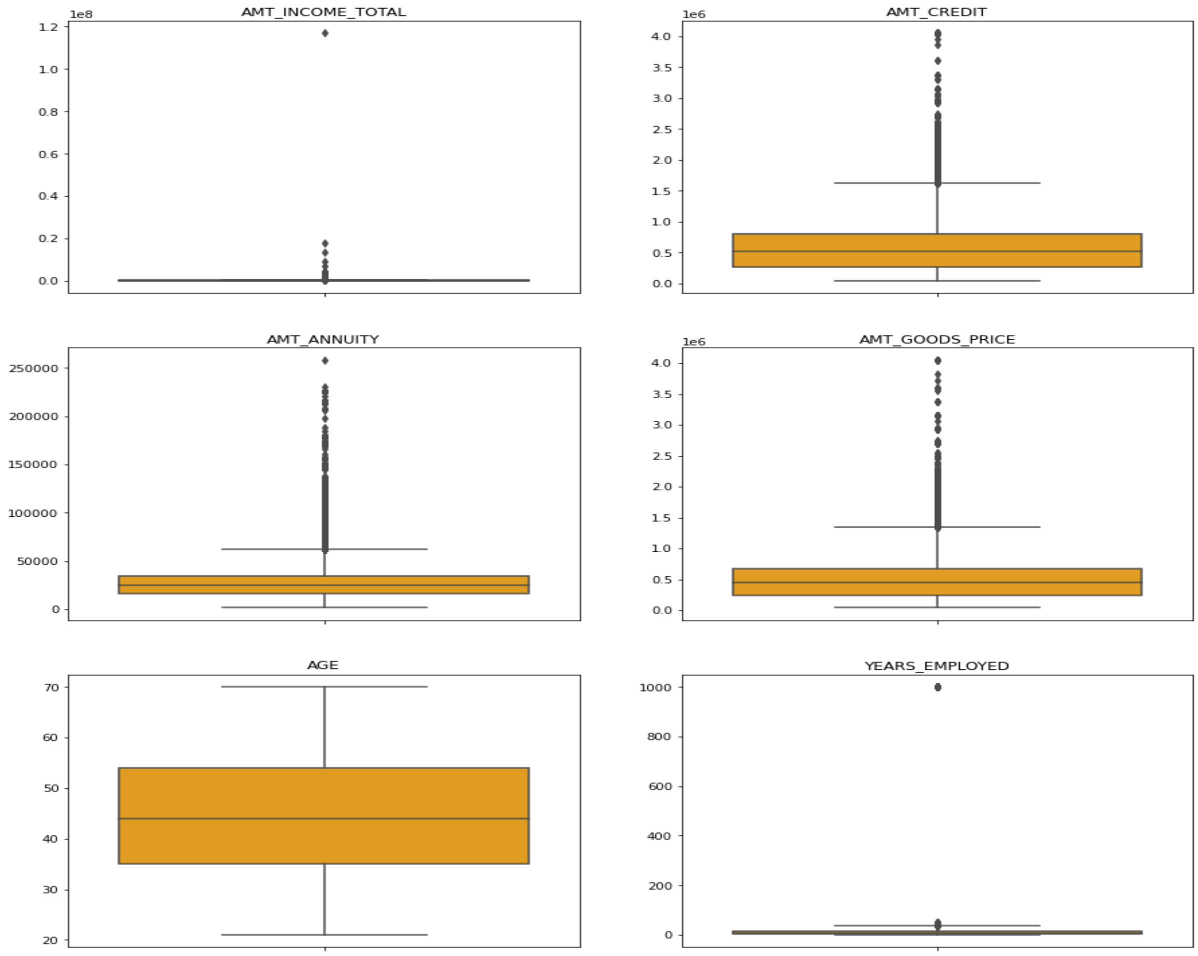
So firstly we have to deal with those missing Values and clean our data for analysis purpose.

**Missing Data**

We have deleted the columns which have more than 40 percent of missing of values in both application and previous dataset. Then we have deleted the irrelevant columns from our dataset which are not going affect in risk analytics.

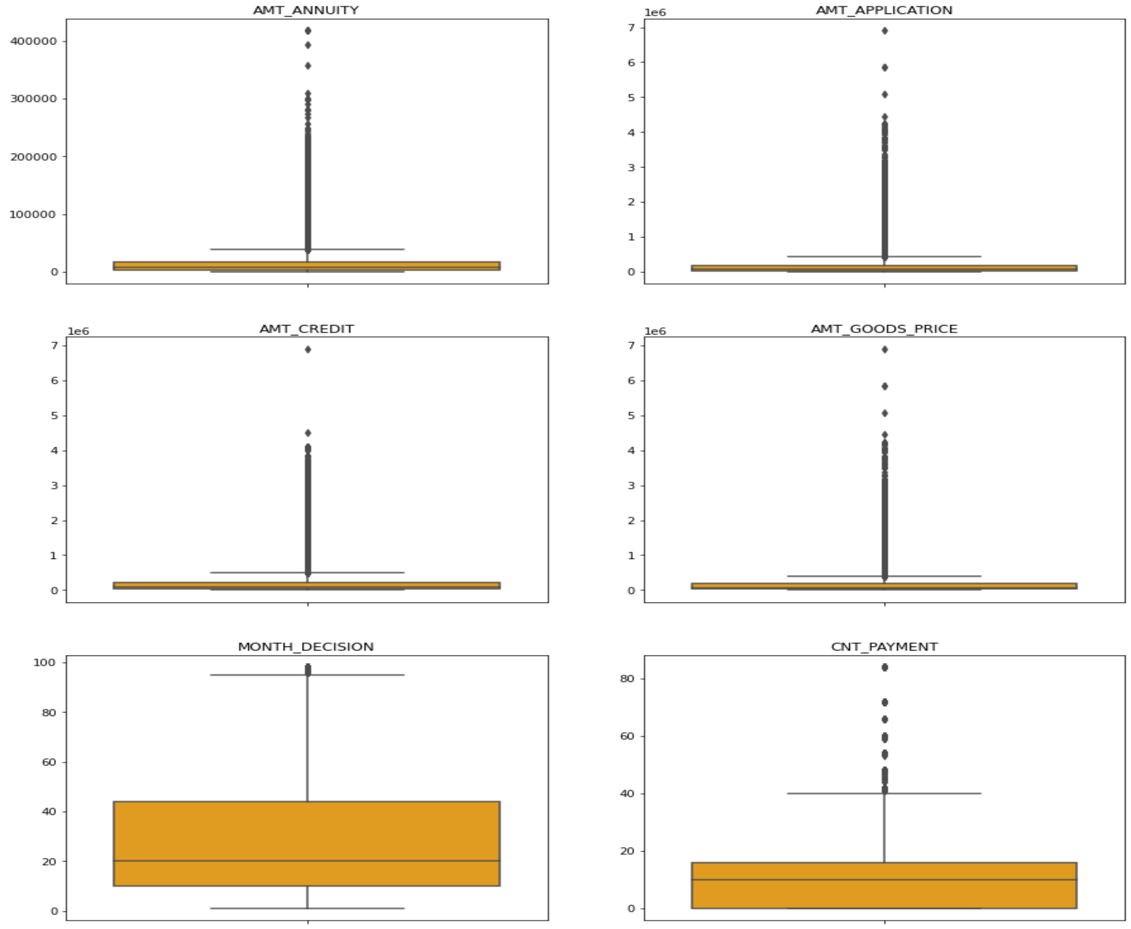
The remaining dataset is the required dataset in which we to perform our analysis. But before proceeding we have to fill the missing values if any in the dataset with the appropriate values and check for **outliers** present in the dataset.

### Checking Outliers in application dataset



- AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE have some number of outliers.
- AMT_INCOME_TOTAL has huge number of outliers which indicate that few of the loan applicants have high income when compared to the others.
- AGE has no outliers which means the data available is reliable.
- YEAR_EMPLOYED has outlier values around 1000 years which is impossible and hence this has to be incorrect entry.

### Checking Outliers in Previous Application Dataset

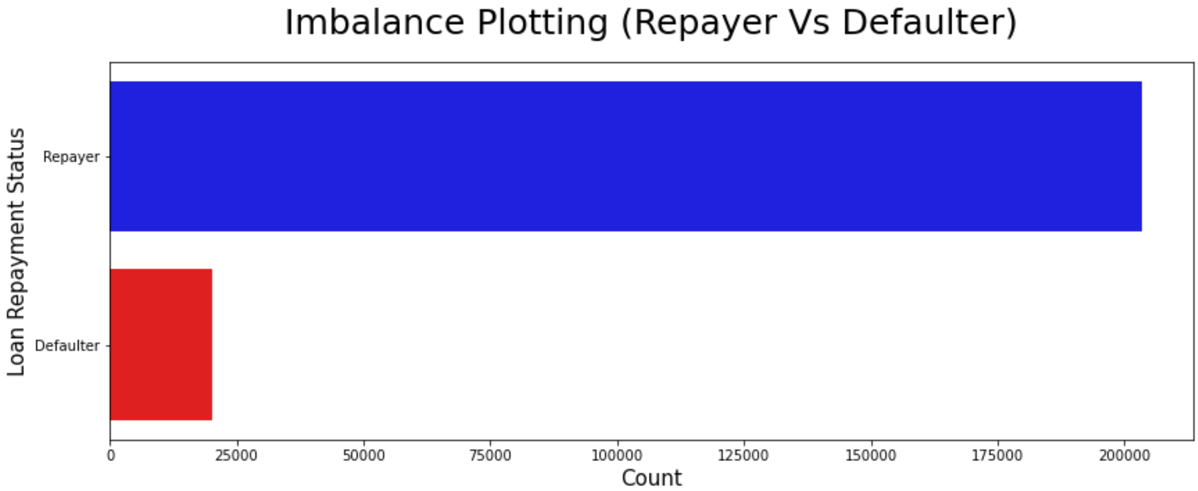


We can see by the above boxplot outlier present in the numerical columns of the previous dataset.

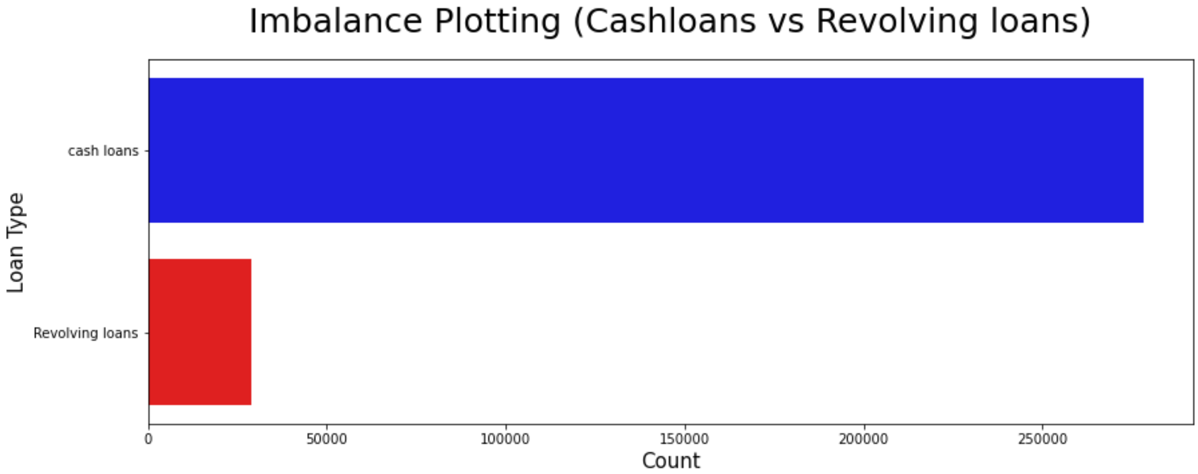
- AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE have huge number of outliers.
- CNT_PAYMENT has few outlier values.
- MONTH_DECISION has little number of outliers indicating that these previous applications decisions were taken long back.

# Data Imbalance

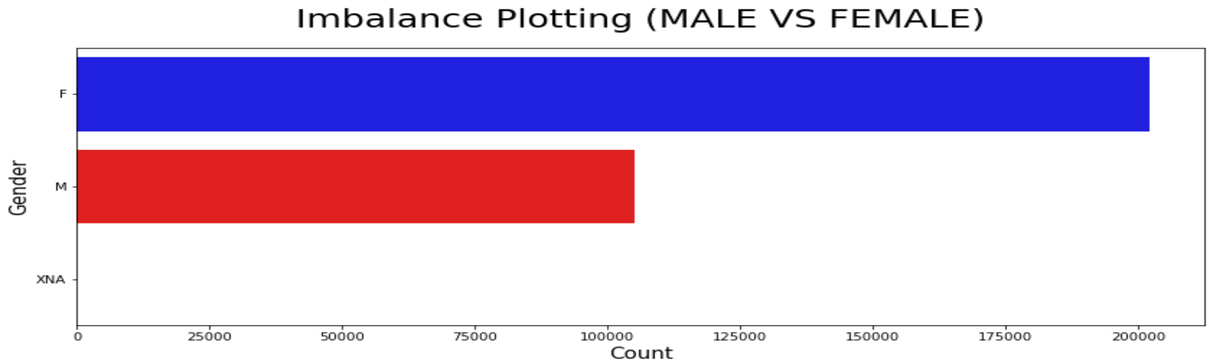
### Application Data



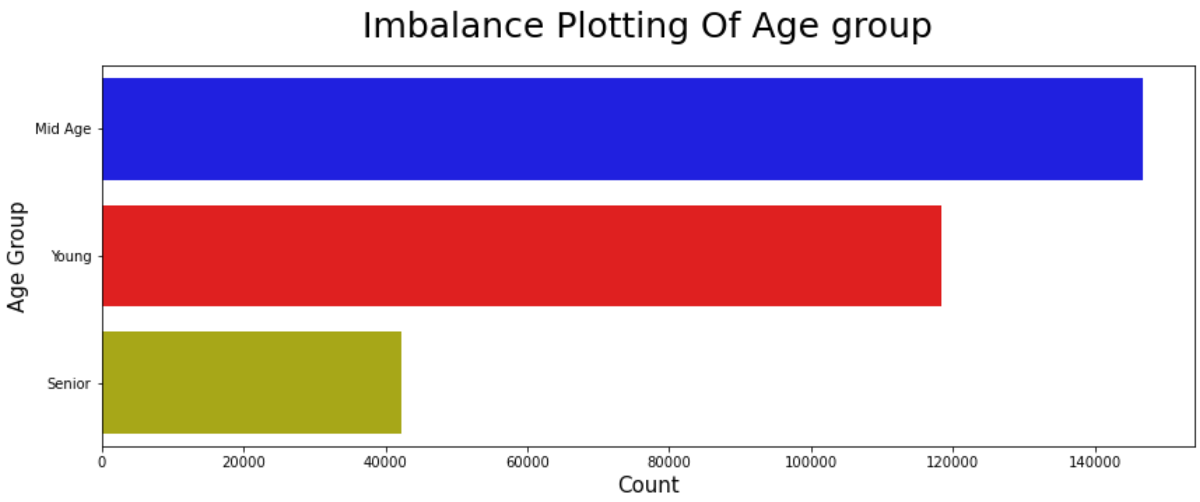
Repayer Percentage is 91.93%
Defaulter Percentage is 8.07%
Imbalance Ratio with respect to Repayer and Defaulter is: 11.39/1(appox).



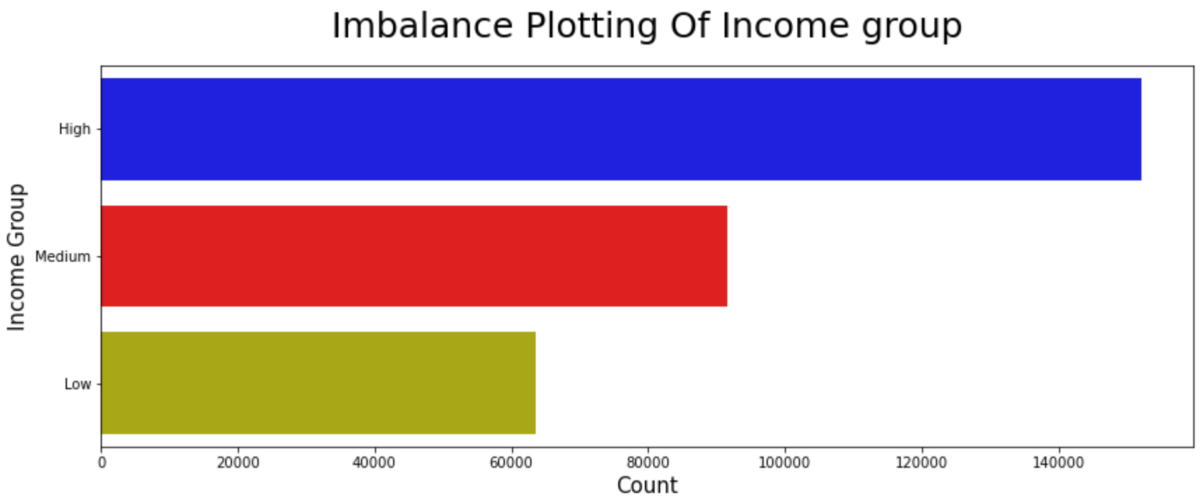
cash loan Percentage is 90.56%
revolving loan Percentage is 9.44%
Imbalance Ratio with respect to cash and revolving loan is: 9.59/1 (approx)



Female Percentage is 65.83%
Male Percentage is 34.17%
Imbalance Ratio with respect to Female and Male is: 1.93/1 (approx)



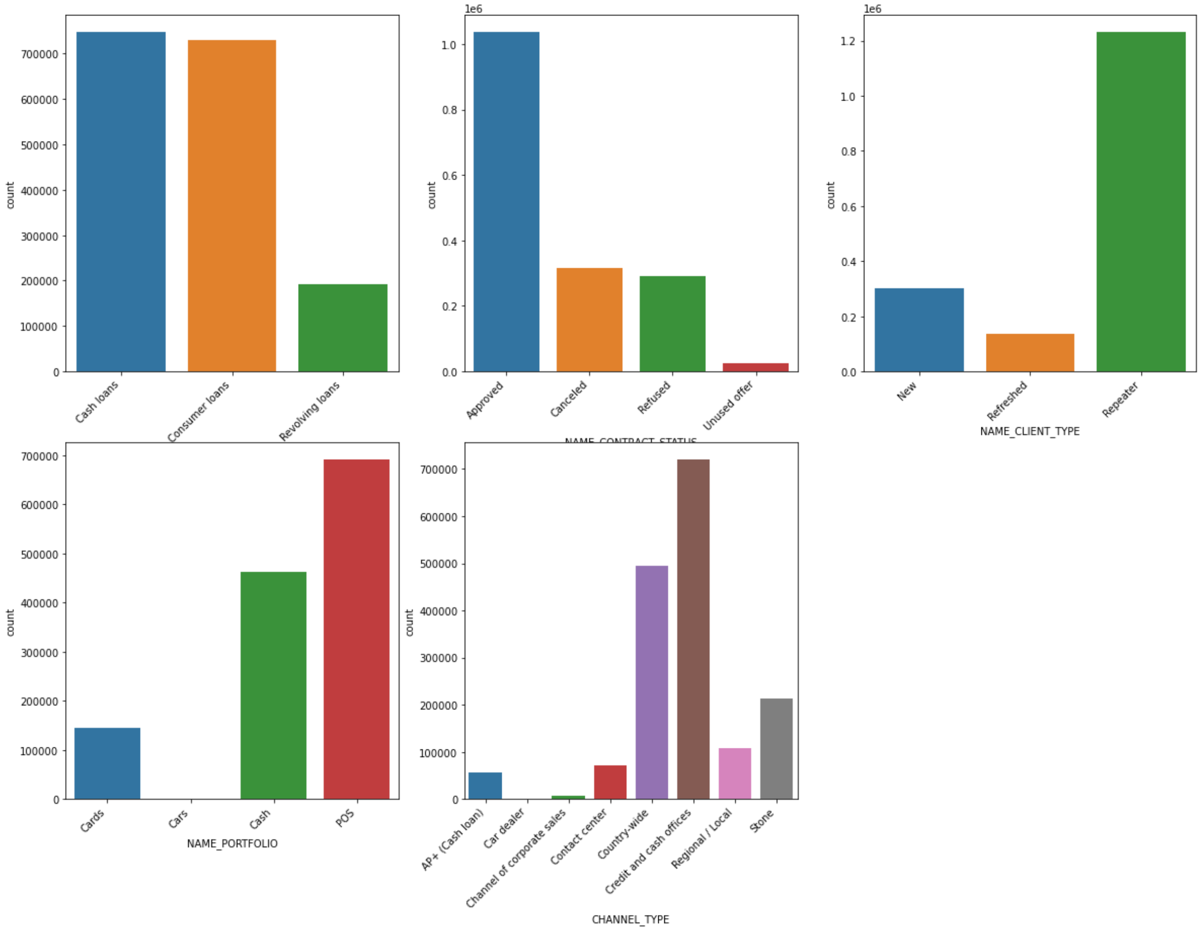
Mid Age Percentage is 47.73%
Young Age Percentage is 38.54%
Senior Percentage is 13.73%



High Income group Percentage is 49.52%
Medium Income group Percentage is 29.78%
Low Income group Percentage is 20.7%.

These are the few columns which are imbalance. Highly unbalance dataset affect the model and their result. The highest imbalance ratio occur in the target column which is around 11. 39 per cent. This is not very high so we can perform our analysis.

**Previous Application Data**

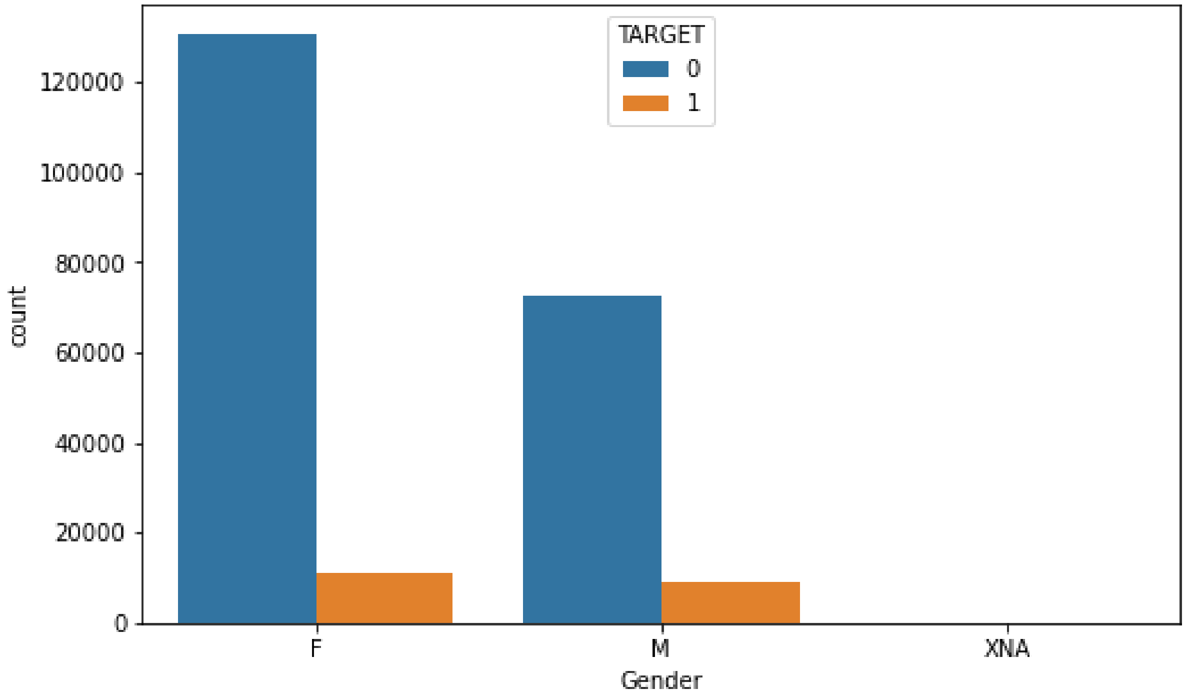


We can see that there is data imbalance in below columns:-

- NAME_CONTRACT_TYPE - There are very few Revolving Loans
- NAME_CONTRACT_STATUS - There are very large Approved status loan and almost negligible Unused offer.
- NAME_CLIENT_TYPE - There are very few Refreshed applicant. Even new applicants.
- NAME_PORTFOLIO - Very few application for Cards and Cars
- CHANNEL_TYPE - Except Country-Wide, Credit and Cash offices and Stone all other channels are very few in number.

# Univariate Analysis on Categorical Data

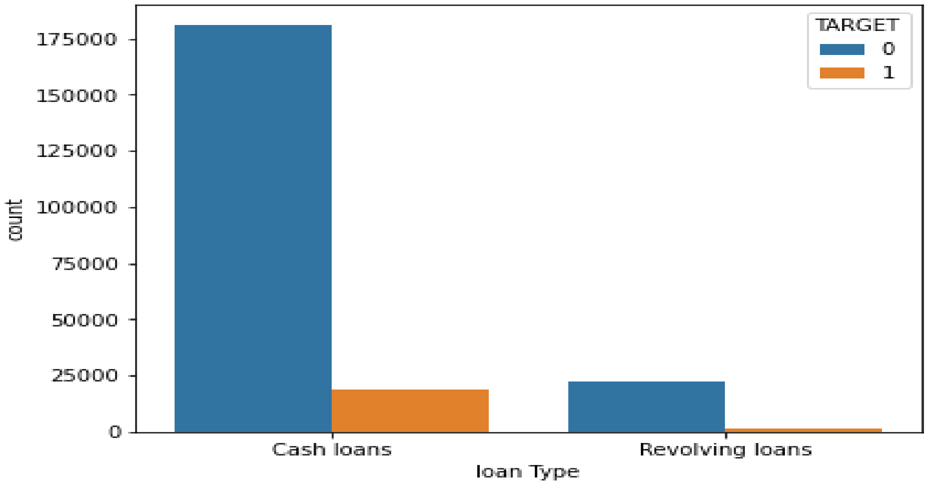
## Defaulter and non-defaulter on the basis of Gender



Analysis

The Females are high in number compare to Male in terms of repaying the loan.

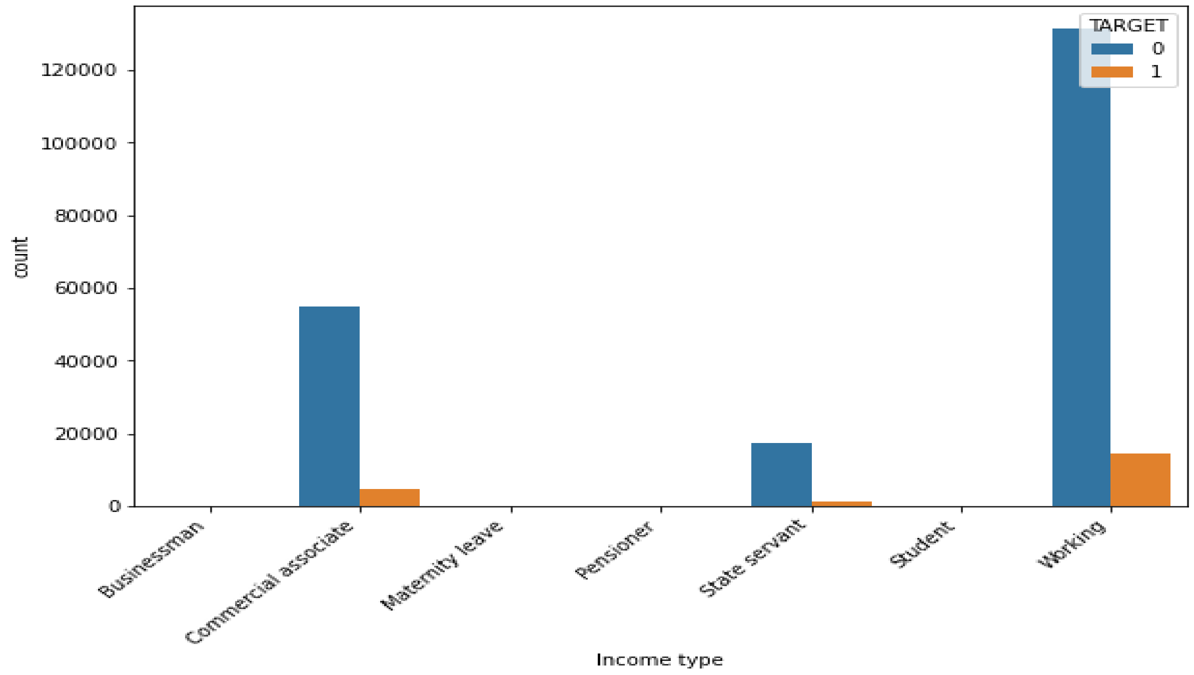
## Defaulter and Non Defaulter on the basis of Loan Type



## Analysis

As we have seen the imbalance in the cash loans and revolving loans. The large number loan application is for cash loans.

### Defaulters and non-defaulters on the basis of Income type



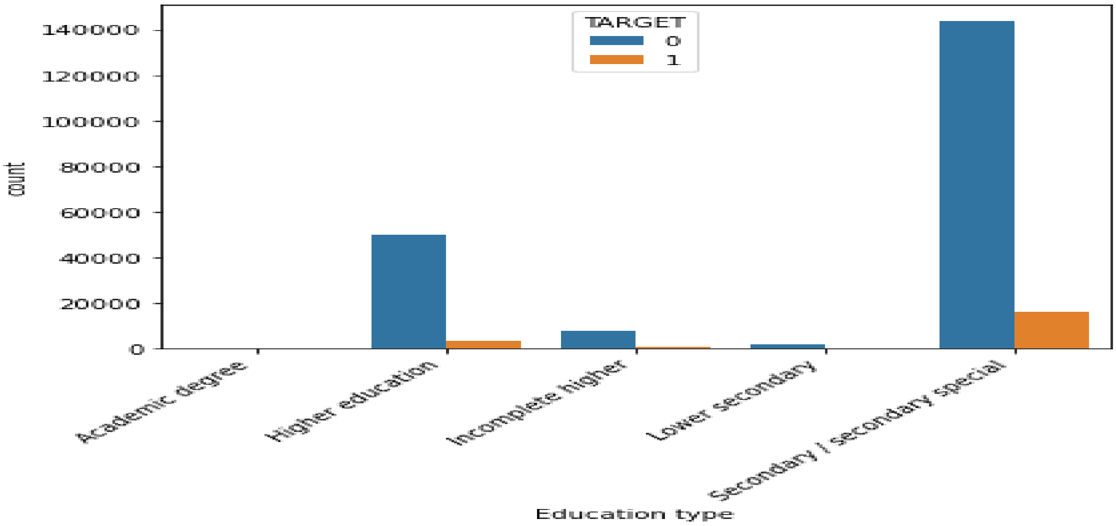
Profession Count

```
Working                 145578
Commercial associate     59564
State servant            18635
Student                     16
Pensioner                    9
Businessman                  3
Maternity leave              3
```

## Analysis

As we can see that the working professional is very high in number who is taking the loan from the above bar graph. Working Professional are higher on compare to other profession who is mostly defaulter and repayer both. As the data is imbalance so we can’t predict anything from this.

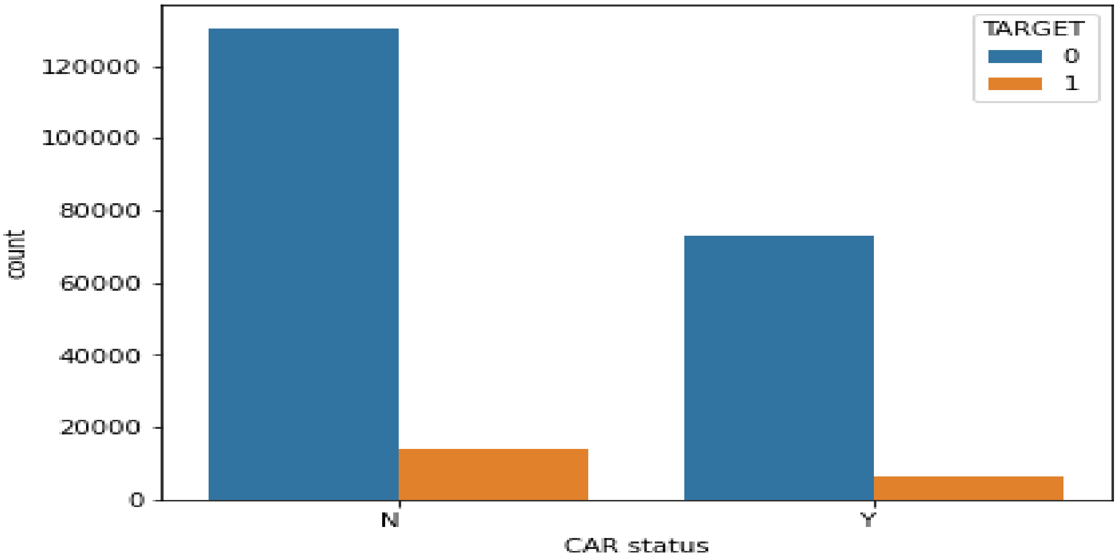
### Defaulters on the basis of Education type

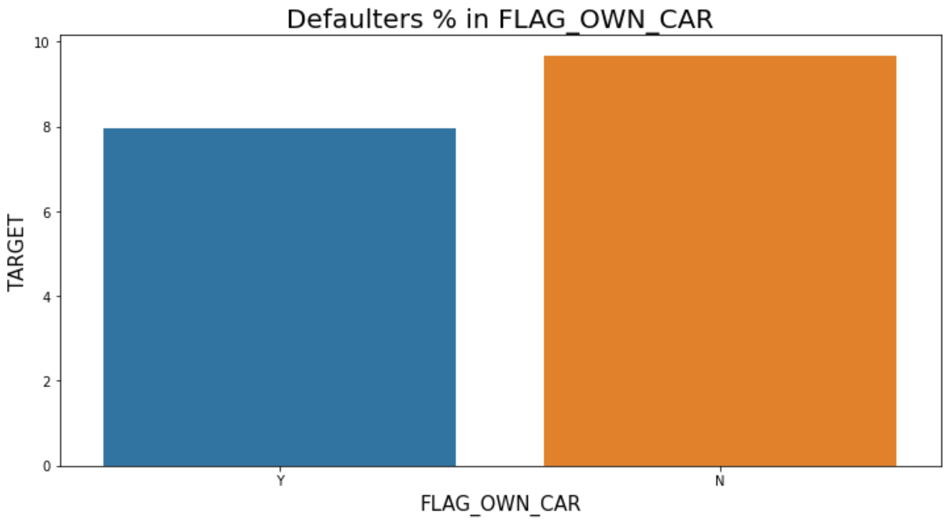


### Analysis

- Majority of clients have Secondary/secondary special education, followed by clients with Higher education.
- Very few clients have an academic degree.
- Lower secondary category have highest rate of defaulting around 11%.
- People with Academic degree are least likely to default.

### Defaulter and Non Defaulter on the basis of Own Car status



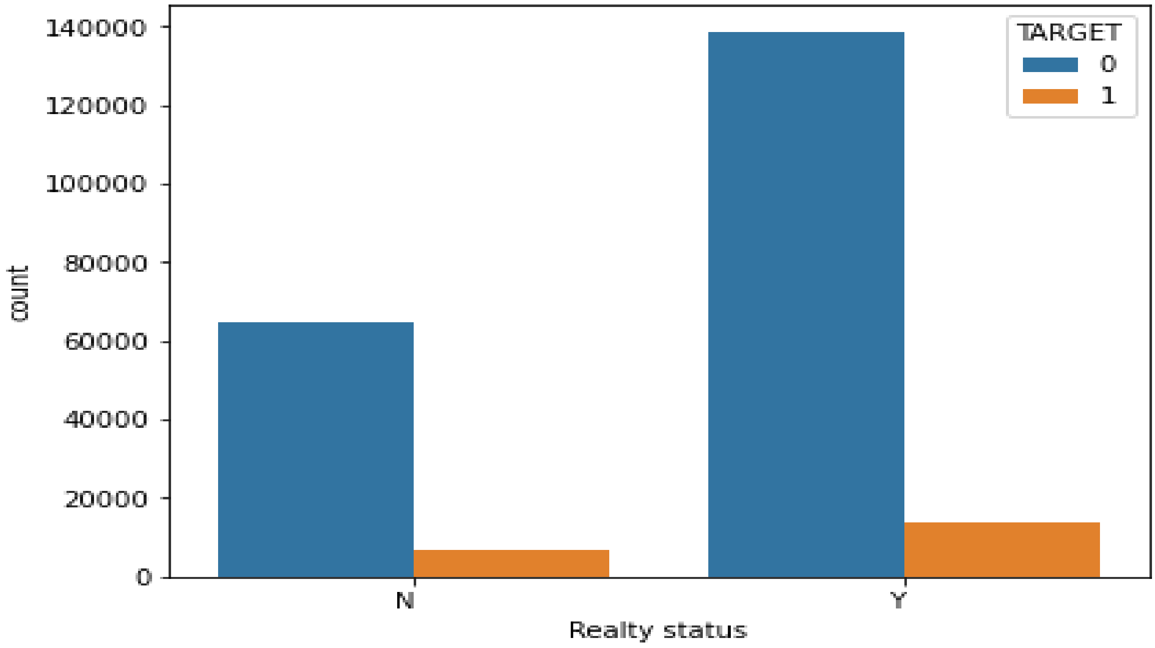


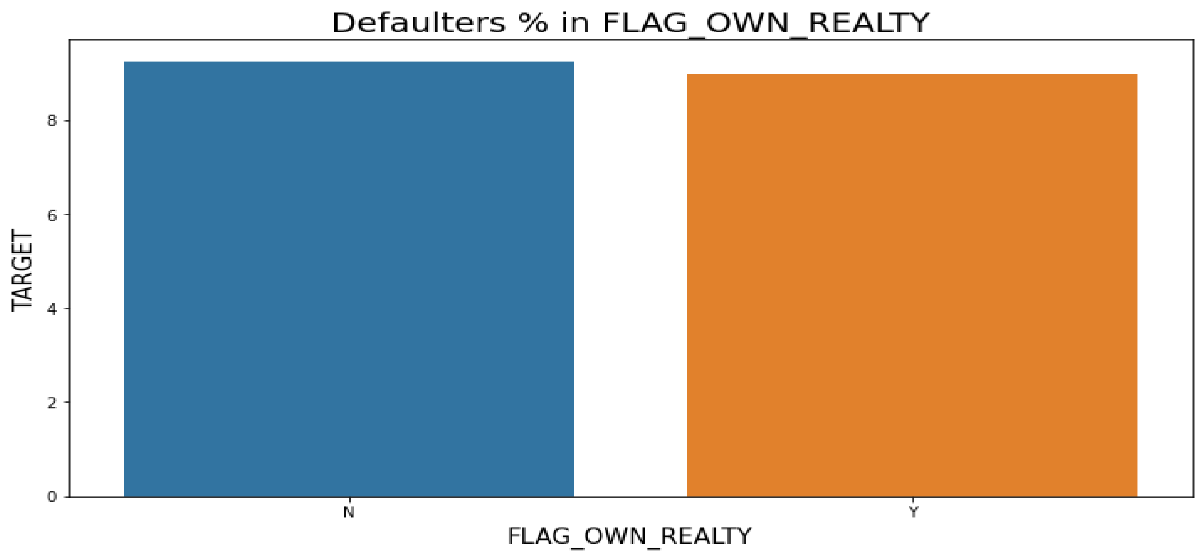
## Analysis

Most of People who have taken the loans are someone who doesn't have Cars

People who don’t have cars have slightly higher chance of being the defaulter than with someone who has cars.

**Defaulter and Non Defaulter on the basis of Own real state status**

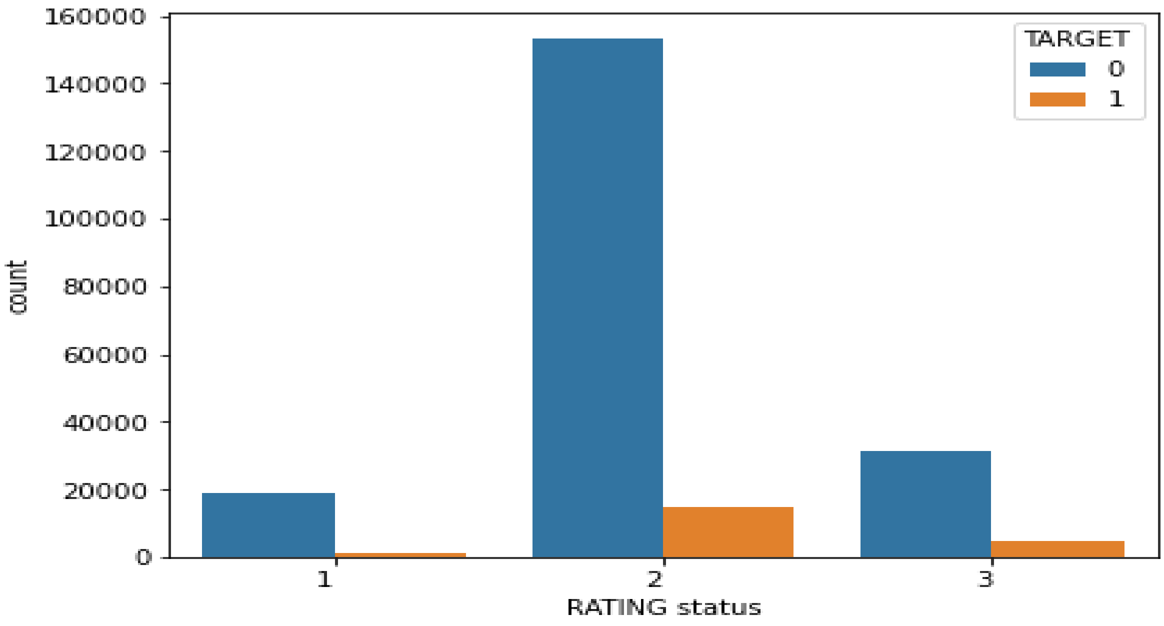


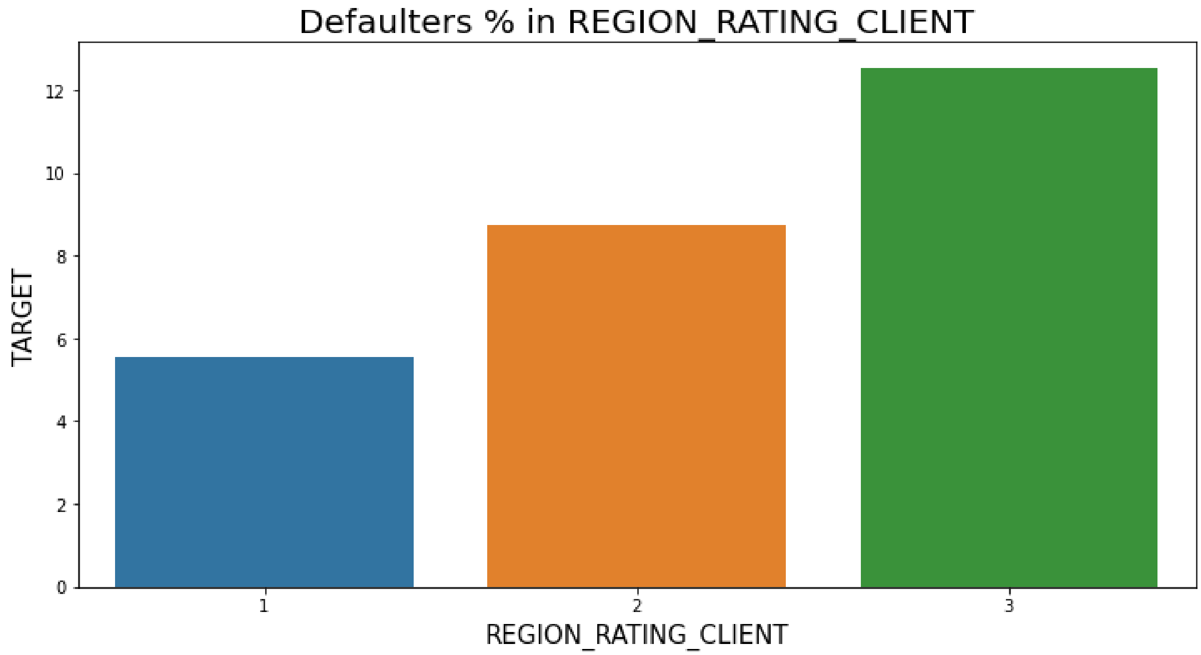


Analysis

- The clients who own real estate are more than double of the ones that don't own.
- The defaulting rate of both categories are around the same (~8%). Thus we can infer that there is no correlation between owning a reality and defaulting the loan.

**Defaulter and Non Defaulter on the basis of Region Rating client**



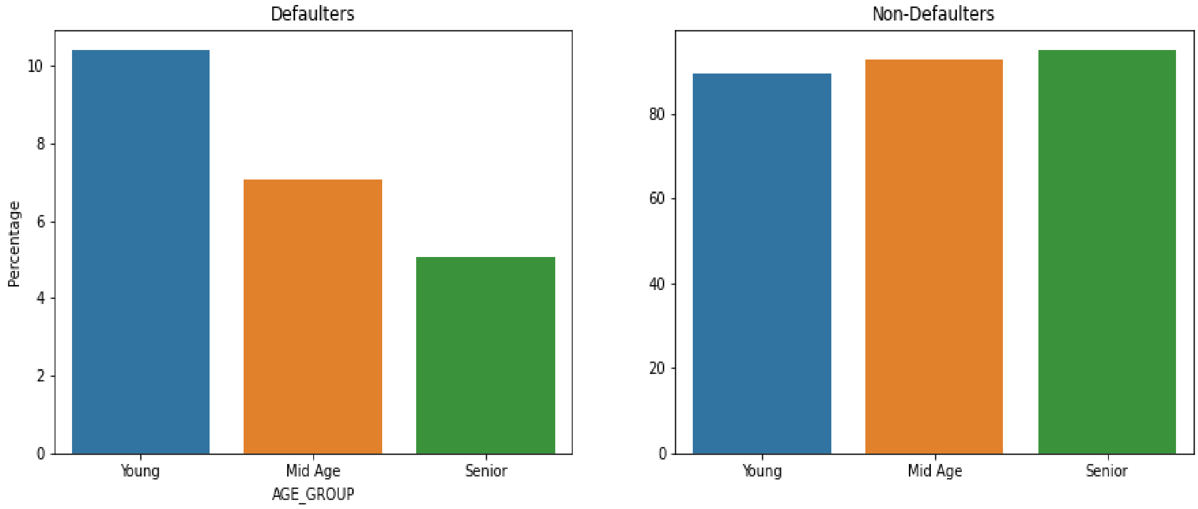


**Analysis**

- Most of the applicants are living in Region with Rating 2 place.
- Region Rating 3 has the highest default rate around(12%).
- Applicant living in Region_Rating 1 has the lowest probability of defaulting, thus safer for approving loans.

## Segmented Univariate Analysis

### Defaulter and Non defaulter based on age group



Defaulter Percentage

| AGE_GROUP | Percentage |
|---|---|
| Young | 10.40 |
| Mid Age | 7.07 |
| Senior | 5.04 |

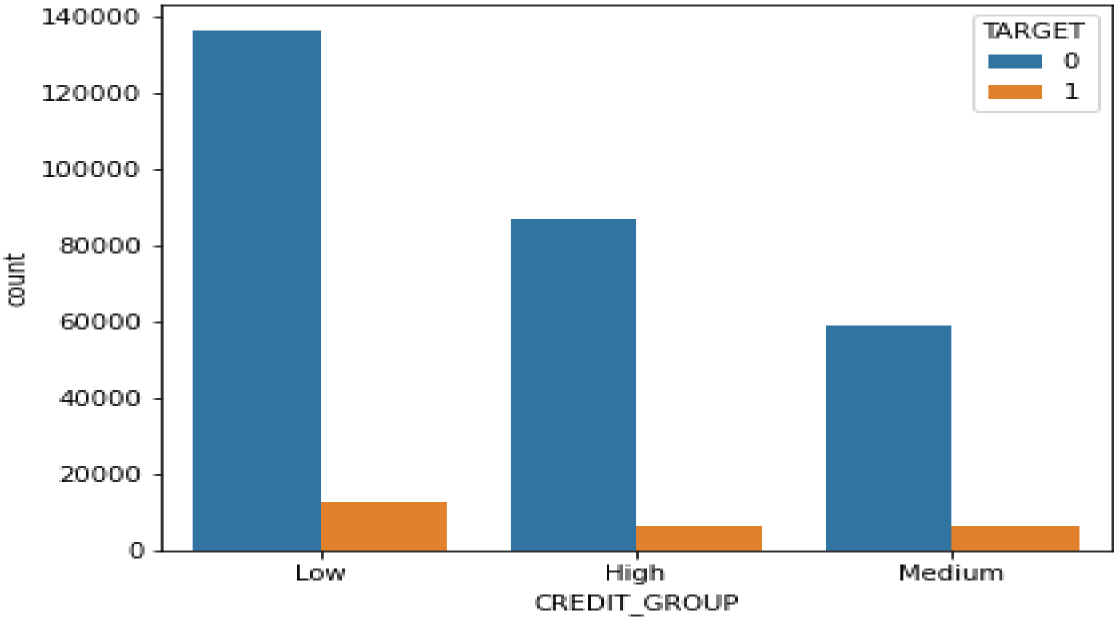
Non Defaulter Percentage

| AGE_GROUP | Percentage |
|---|---|
| Young | 89.60 |
| Mid Age | 92.93 |
| Senior | 94.96 |

## Analysis

We can see that the Young people are more likely to dafault and senior category people more punctual towards their loan repayment.

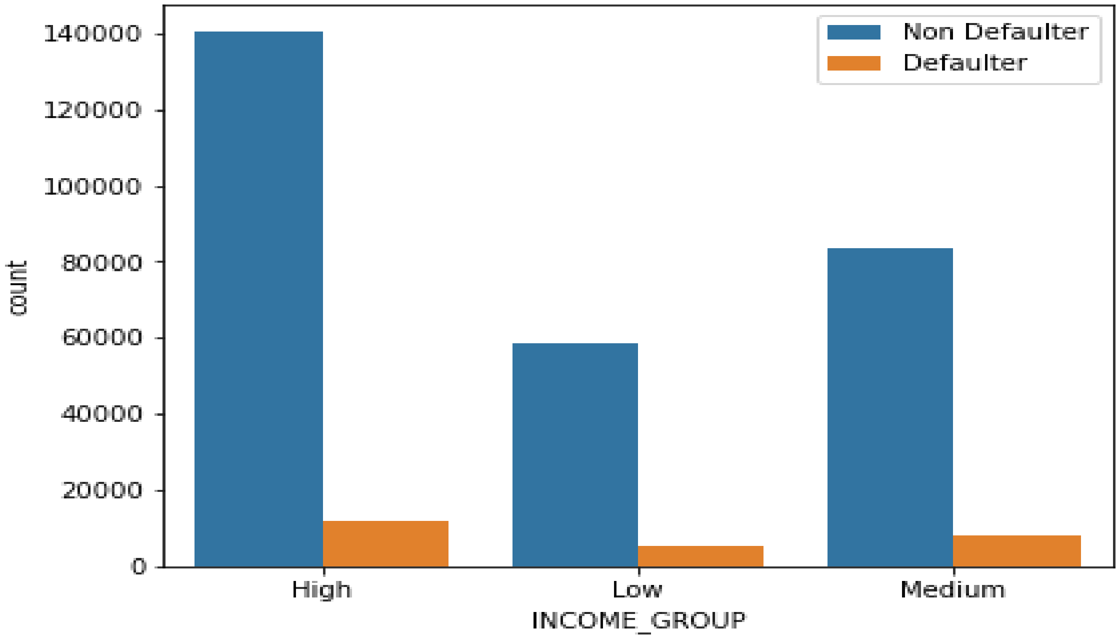
**Defaulter and Non Defaulter based on credit group**



## Analysis

As expected low credit amount groups are more in number, who were not defaulted whereas surprisingly **low credit amount group are slightly higher in number than other in defaulted Count .**

### Defaulter and Non defaulter based on Income Group

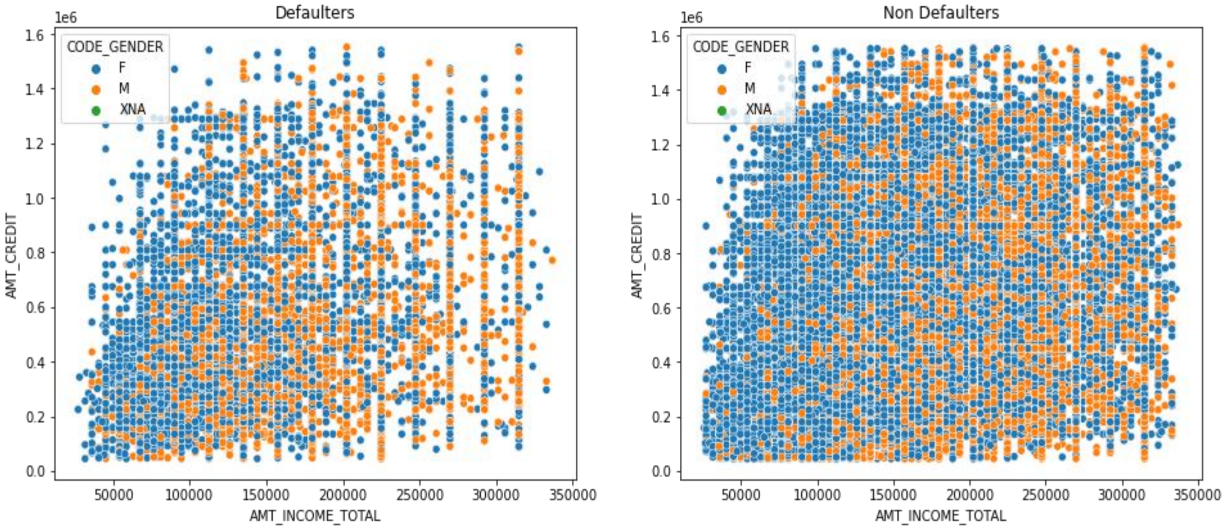


## Analysis

As expected High Income group people are likely to repay the loan on timely basis.

## Bivariate And Multivariate Analysis

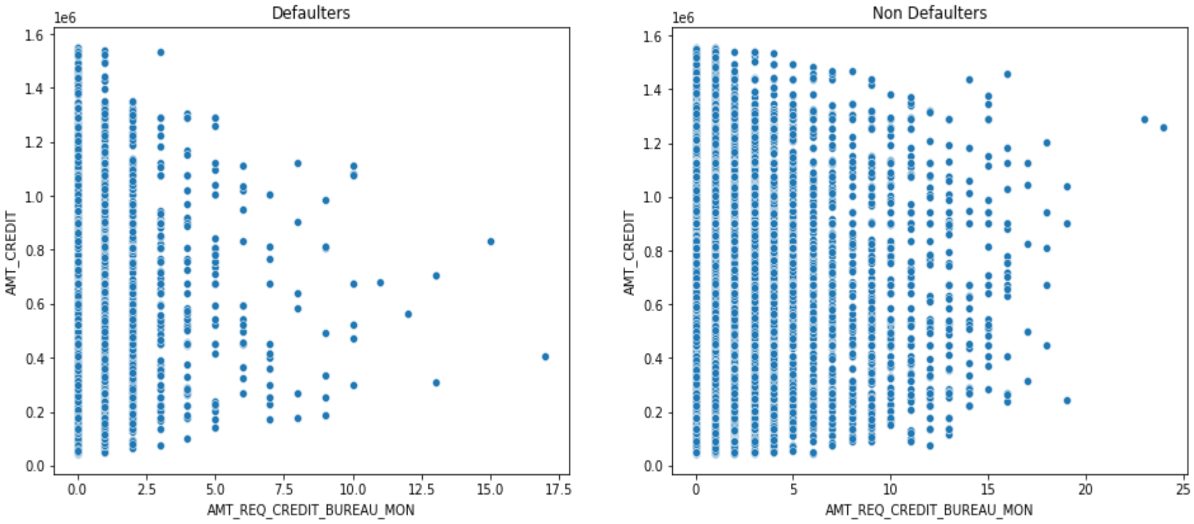
### Credit amount of the loan on the basis of client income for both male and female



## Analysis

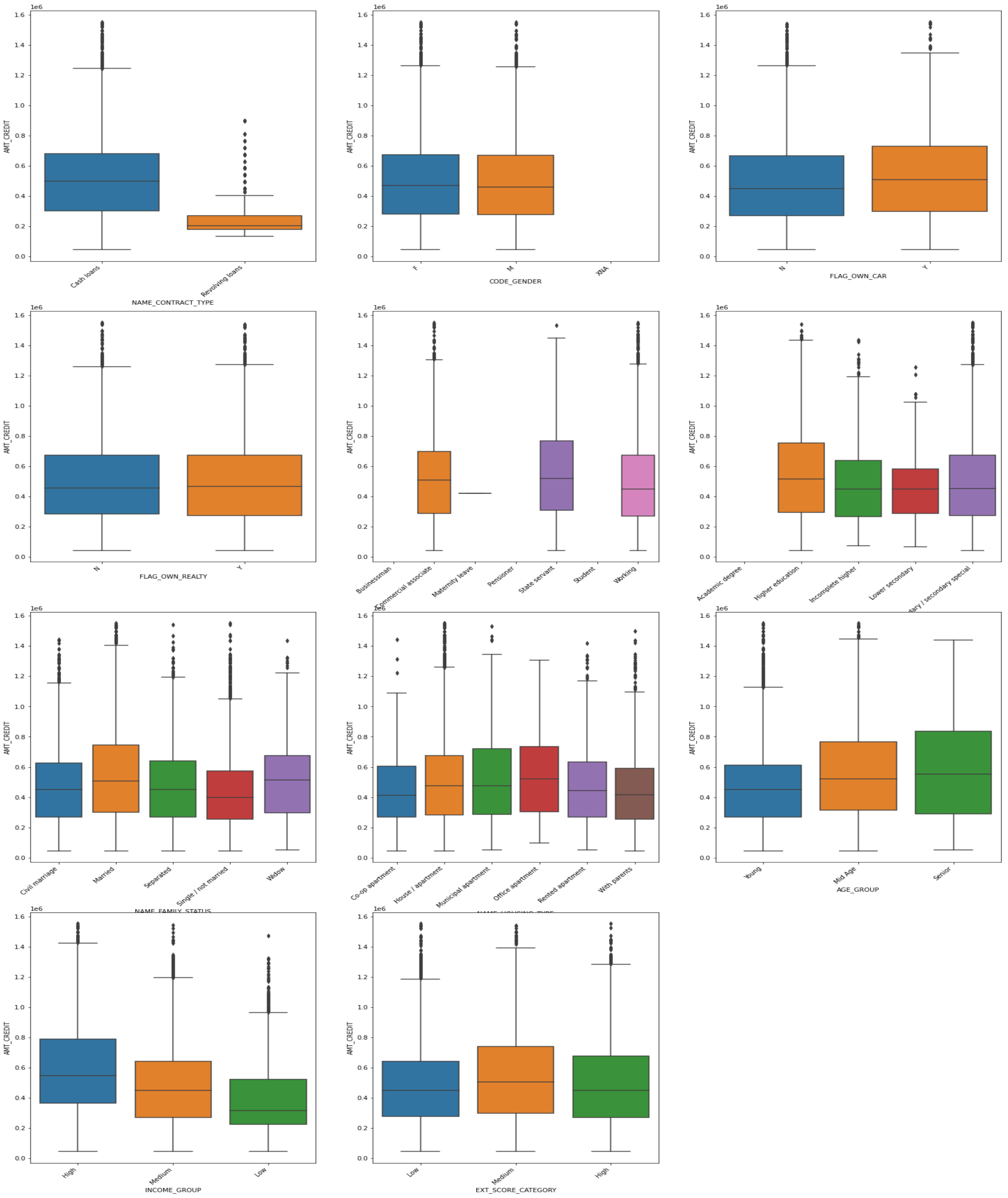
- From the above scatter plot we can see that values are highly concentrated at lower income and lower credit group and the concentration of female at lower credit and income group is high for both defaulter and non-defaulter plot.
- Large of number of female defaultees are from lower income group.
- Male Deaultees are in Higher Income group.

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



We can see from the above scatter plot that as the number of enquiries increasing the concentration started decreasing i.e. larger the number of enquiries the chances of loan getting approved is less. The amount credit is also decreasing with increase in credit inquiry.

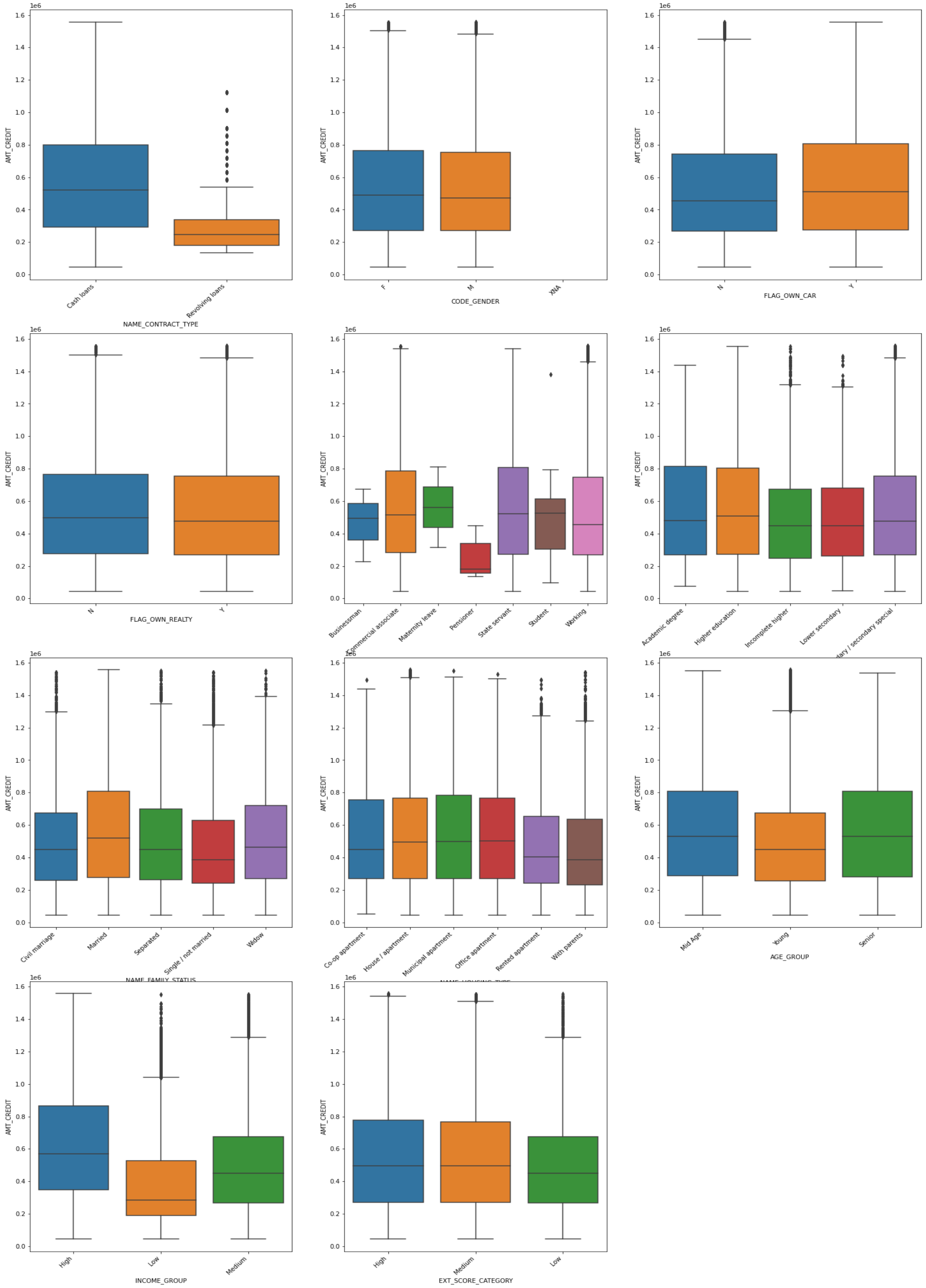
# Categorical Variable  : Defaulter



## Analysis

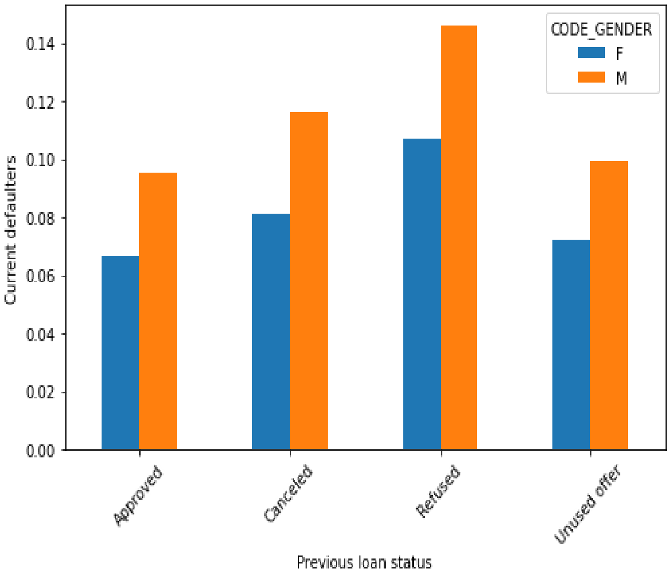
- Credit amount of the loans are very low for Revolving loans.
- There is no credit amount difference between genders, client owning cars or realty.
- The Young age group got less amount of loan credited compared to mid age and senior citizen.
- Higher income group have more loan amount credited.
- Clients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

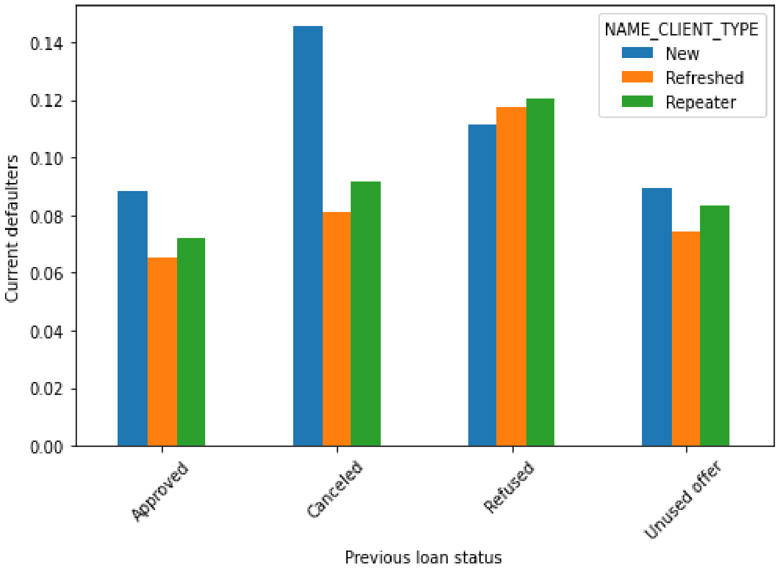
# Non Defaulter



## MERGED DATA ANALYSIS

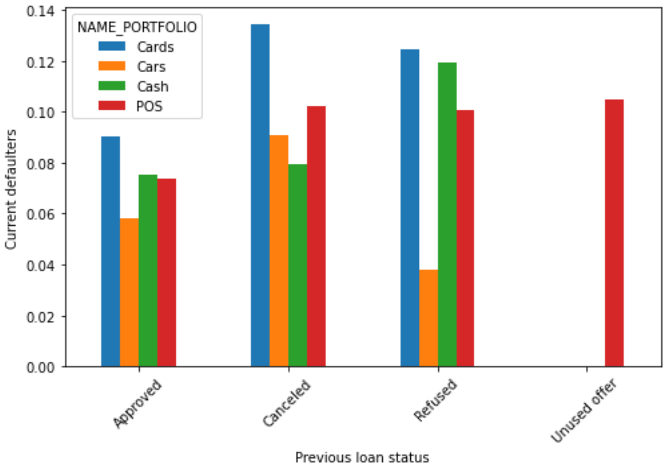
### Previous Loan Status

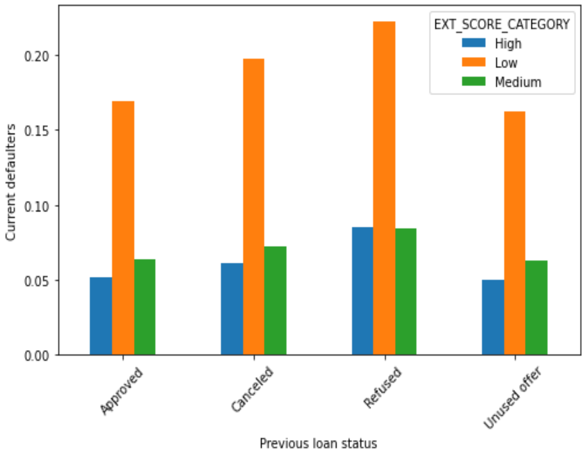




## Analysis

- Previously refused and unused offer applications were more defaulted in male.
- New clients with Previously unused offer and canceled are more defaulted.
- For previously Refused applicants the Defaulters are more Refreshed clients.





### Analysis

- The application for portfolio cards are mostly defaulted.
- For unused offer Client applied for POS are only defaulted.
- Low Ext Source scorer is highly defaulted.

## Top 10 correlation for the Client

### Payment Difficulties

| Var 1 | Var 2 | Correlation |
|---|---|---|
| AMT_GOOD_PRICE | AMT_CREDIT | 0.983103 |
| AMT_GOOD_PRICE | AMT_ANNUITY | 0.752699 |
| AMT_ANNUITY | AMT_CREDIT | 0.751957 |
| YEARS_EMPLOYED | AGE | 0.582545 |
| EXT_SOURCE_SCORE | AGE | 0.184311 |
| EXT_SOURCE_SCORE | AMT_GOODS_PRICE | 0.139405 |
| AGE | AMT_GOODS_PRICE | 0.135578 |
| AGE | AMT_CREDIT | 0.135084 |
| EXT_SOURCE_SCORE | AMT_CREDIT | 0.132023 |
| EXT_SOURCE_SCORE | AMT_ANNUITY | 0.104467 |

These columns are highly correlated for the data of the customer which are having the payment difficulties.

### Analysis

Credit amount is highly correlated with amount good price and amount annuity.

### Non Defaulter

| Var1 | Var2 | Correlation |
|---|---|---|
| AMT_GOODS_PRICE | AMT_CREDIT | 0.987253 |
| AMT_GOODS_PRICE | AMT_ANNUITY | 0.776686 |
| AMT_ANNUITY | AMT_CREDIT | 0.771113 |
| YEARS_EMPLOYED | AGE | 0.626069 |
| AMT_ANNUITY | AMT_INCOME_TOTAL | 0.418745 |
| AMT_GOODS_PRICE | AMT_INCOME_TOTAL | 0.349461 |
| AMT_CREDIT | AMT_INCOME_TOTAL | 0.342580 |
| EXT_SOURCE_SCORE | AGE | 0.183079 |
| YEARS_EMPLOYED | AMT_INCOME_TOTAL | 0.140504 |
| EXT_SOURCE_SCORE | AMT_GOODS_PRICE | 0.113823 |

### Analysis

- Credit Amount in both Repayer and Defaulter case are highly correlated with amount good Prices.
- We can see that the correlation between the age and years employed is higher in case of repayer than Defaulter.
- We can see that the correlation between the amount credit and amount annuity is higher in case of rapayer than defaulter. There is a small drop of 0.019156.

## Result

After analysing all the features we can conclude that there are few attribute of a client on which bank would identify if they will repay the loan or not.

1. Bank should focus more on Income types  student, businessman and pensioner who have no default percent or have very less default chances. Bank should refrain from giving loan to someone who is on maternity leave and who are unemployed as they have high default rate.
2. Bank should focus more on someone who have Academic degree and atleast completed their higher education has they have less default rate. Bank should refrain from giving loan to people who have completed lower secondary.
3. Region rating of client also influencing their chances of getting defaulter. Region rating 1 clients repay the loan on time whereas rating 3 clients are mostly defaulters.
4. Bank should check occupation of client before giving the loan. Low skill labourers, labourers and drivers are mostly the defaulters.
5. Since almost 90 % of applicants have Income less than 3L and they have high probability of defaulting, they could be offered loan with Higher Interest compare to others category.
6. People who get loan between 5-8Lakh in medium credit group are tend to default more and hence having the higher interest rate for this credit range would be ideal.
7. It is recommended to provide loans to previously approved females.
8. Clients whose applications are refused or unused previously tend to be more risky hence having higher interest rate for that clients would be ideal.