EDA- loan defaulter case study

BY-

Karthik Kini

Problem Statement

- This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which will 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 (banking companies) wants to understand the driving factors/variables behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Relative hypothesis

- **H0**: Customer who can repay the loan without default are chosen
- **H1**: Customer who are likely to default are chosen
- Errors which are likely to occur:
- **Type 1**: Reject Null Hypothesis i,e customer who can repay the loan are rejected, this might result in loss of business.
- **Type 2**: If the lender fails to reject false null hypothesis that is customer who are more likely to default are given loans, that will be a financial loss to the company.

Approach we used for our analysis:

We have been given 3 datasets namely:

- *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.**

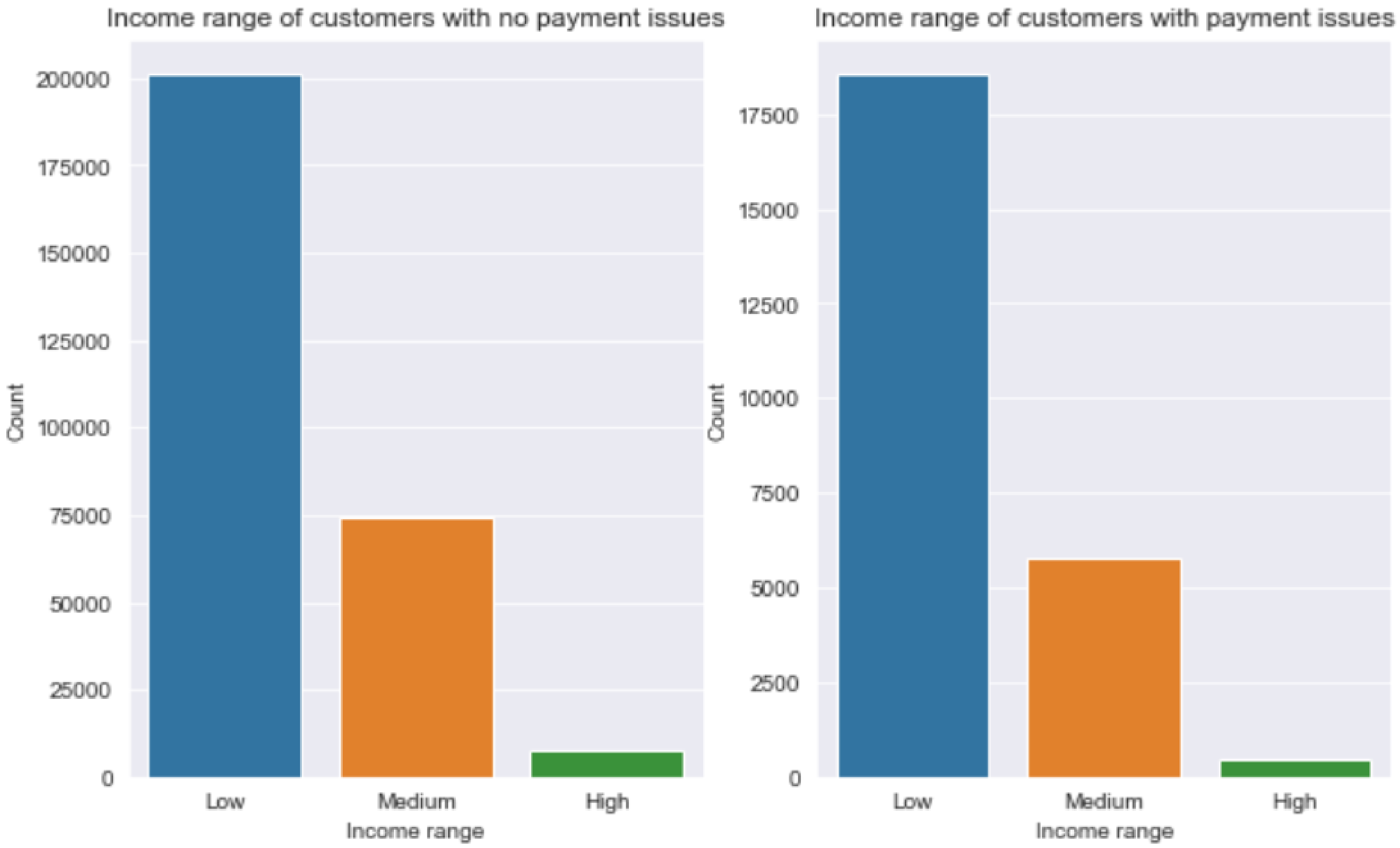
- *2. 'previous_application.csv'* contains information about the client's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

- *3. 'columns_description.csv'* is data dictionary which describes the meaning of the variables.

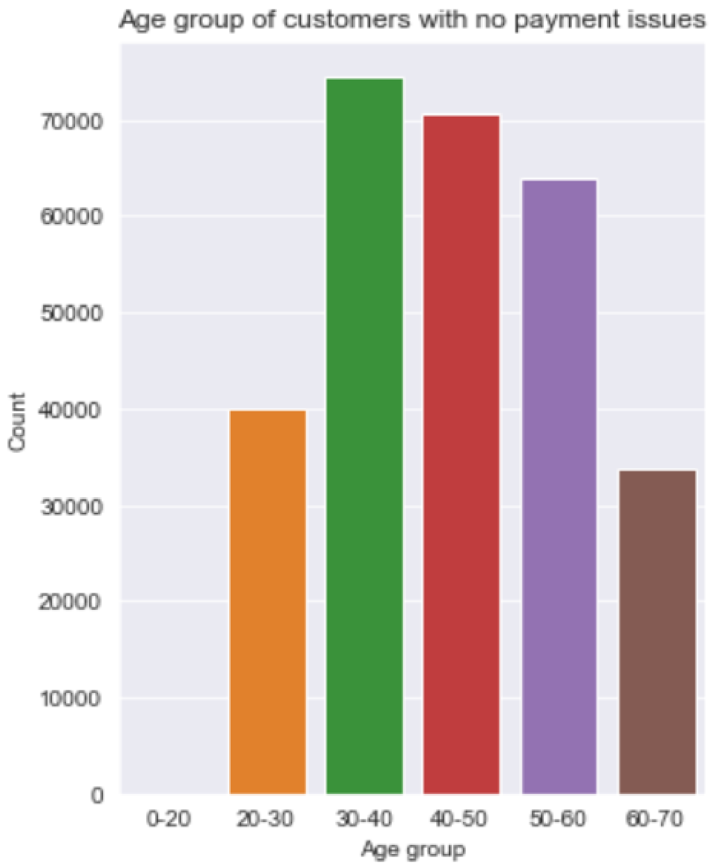
Steps we used for EDA

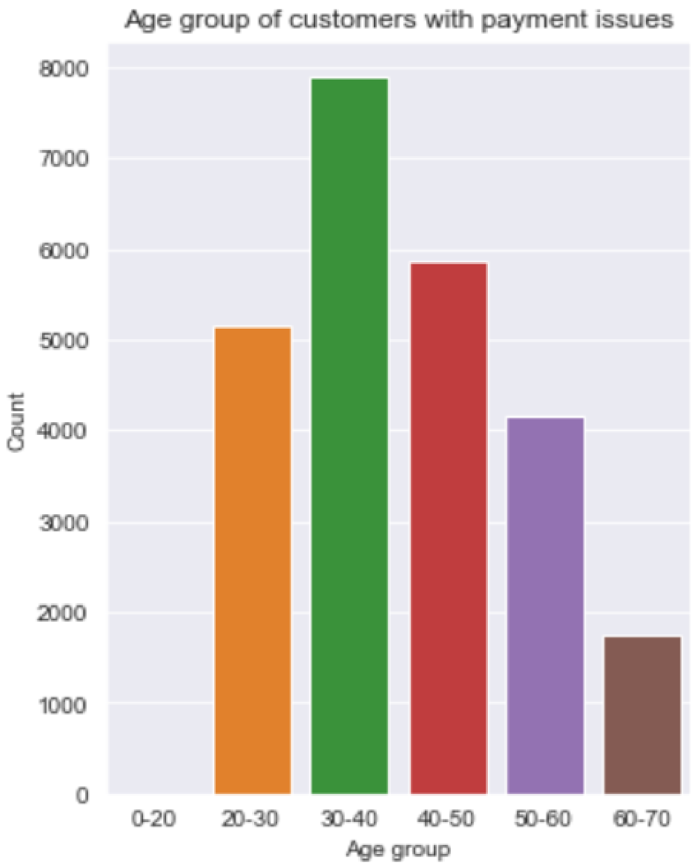
1. Data importing
2. Analysing structure of the data frame (shape, info, dtype etc)
3. Analysing numeric columns using describe
4. Finding out the null value percentage in each column and dropping columns having percentage greater than 50%
5. Finding out the important features(column)
6. Figuring out data imbalance
7. Univariate Analysis on the reduced data frame
8. Bivariate Analysis
9. Multivariate Analysis, Correlation

Insights from application dataset

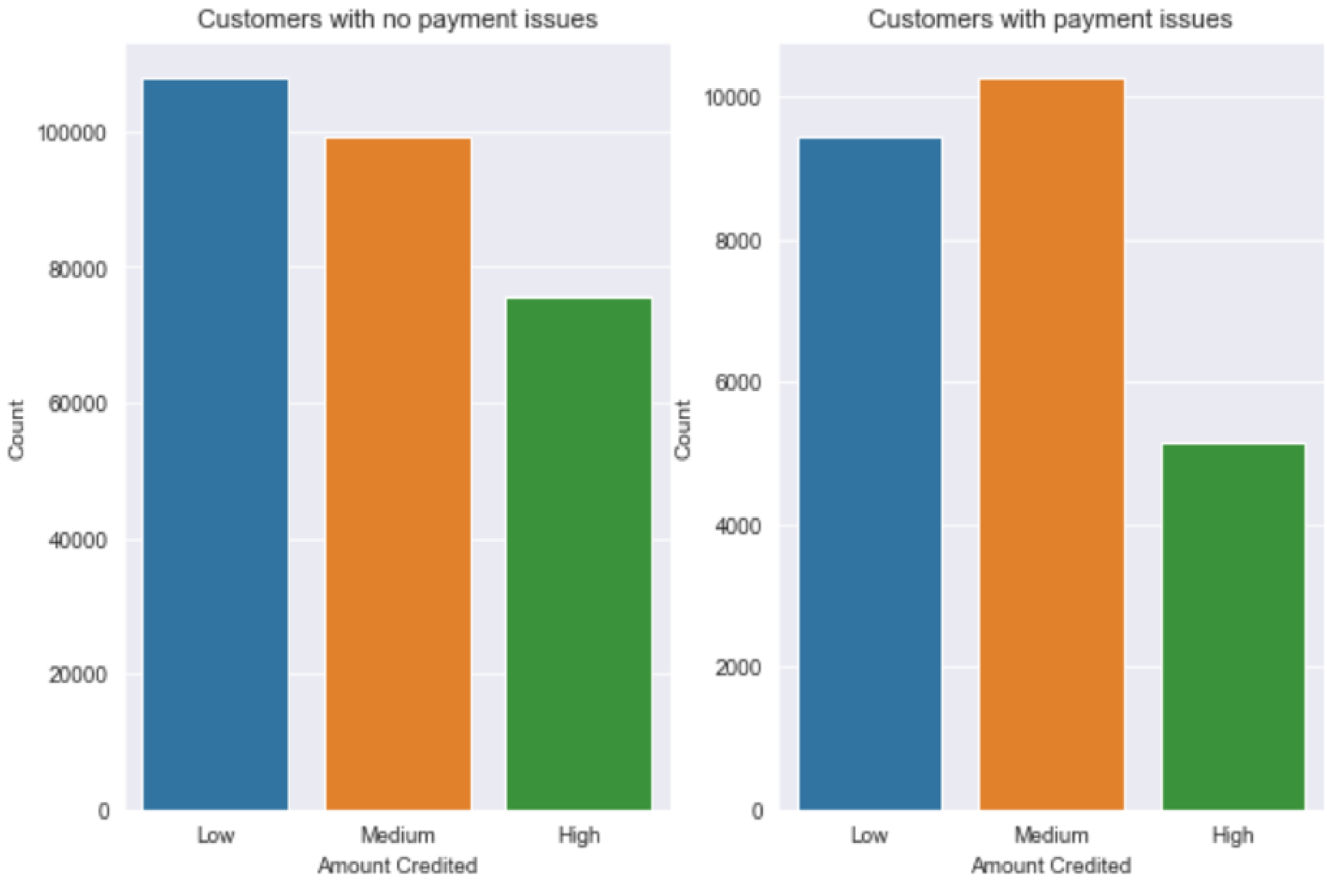


Most of the loan applicants have low to medium range incomes

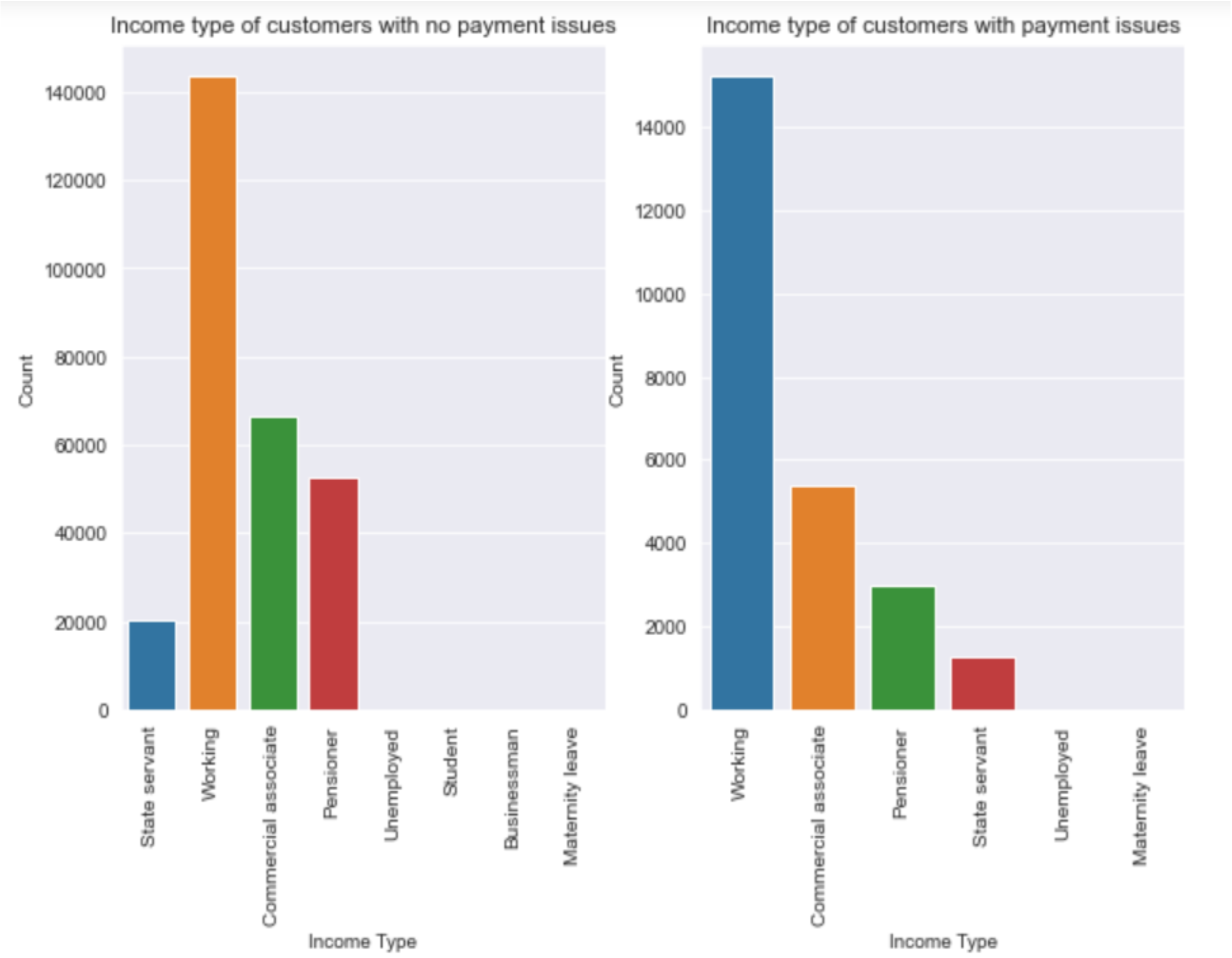




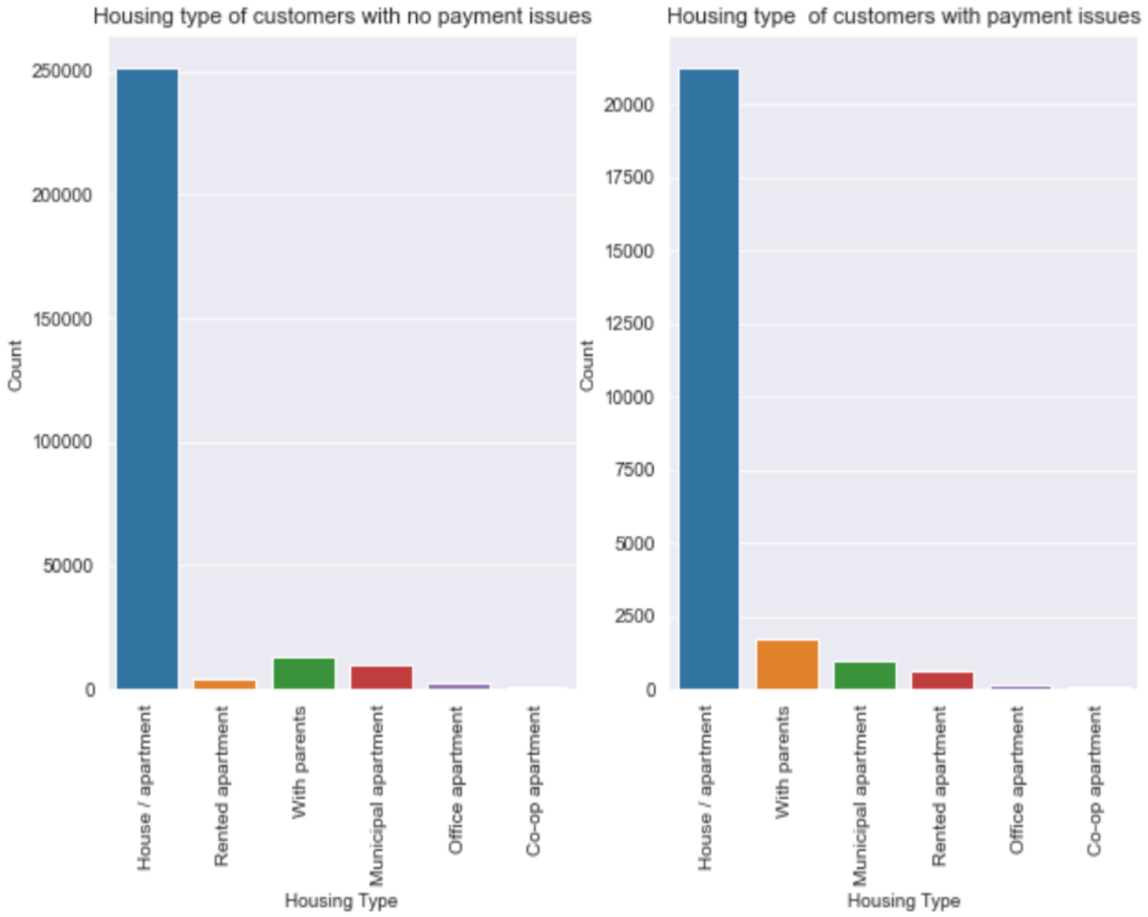
Loans should be given to customers in the age group of 30-60, but mostly in age group of 30-50



It can be seen that customers with less credit are more likely to make payment .Also, after low credit , medium credit amount loans can be considered

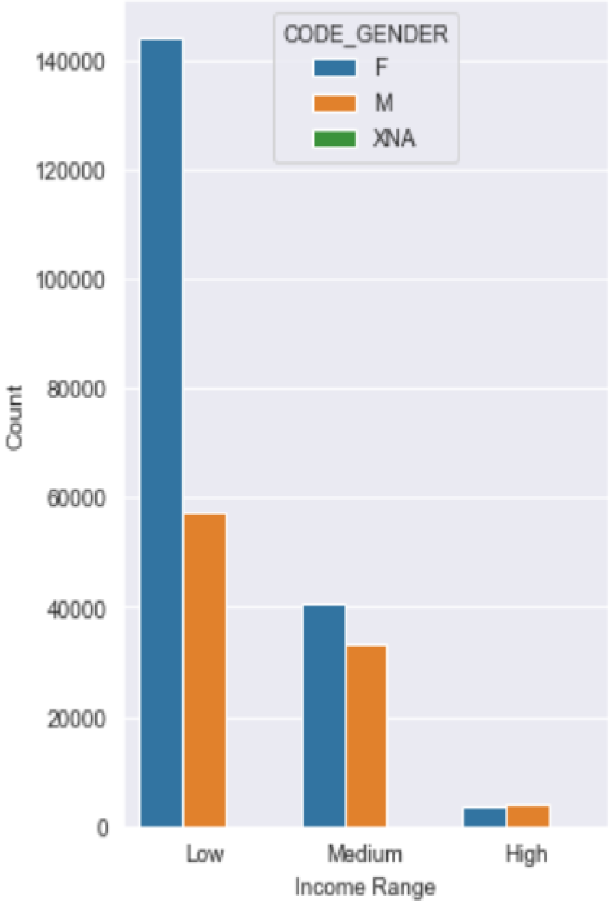


As we can see the count of customers who are working professionals are more likely to repay the loan.
Also , customers working as State Servants, Commercial Associate's and Pensioner's can be considered.

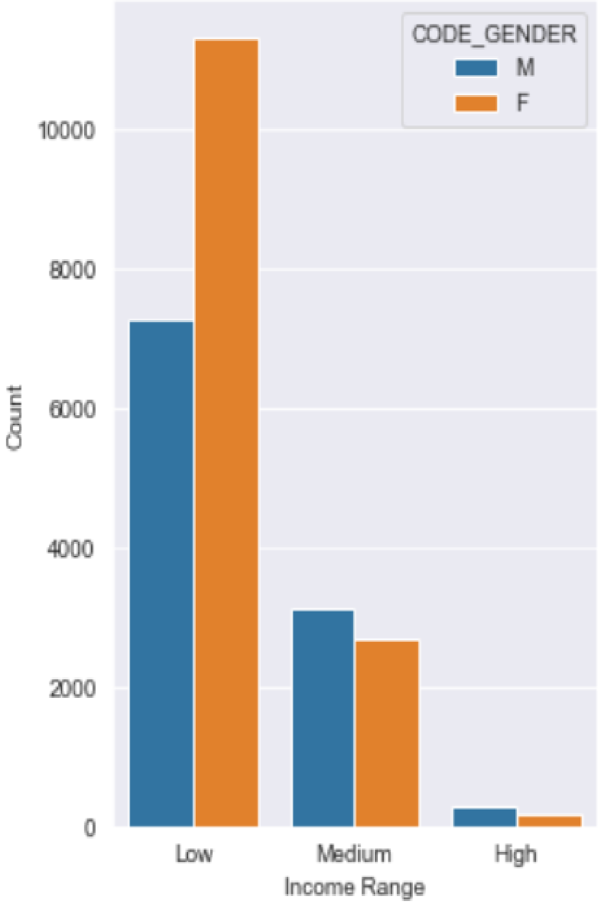


It can be seen that people with own houses can be targeted and also a small fraction of loan can be given to people
living with their parents as there are chances that the parents already have their own house

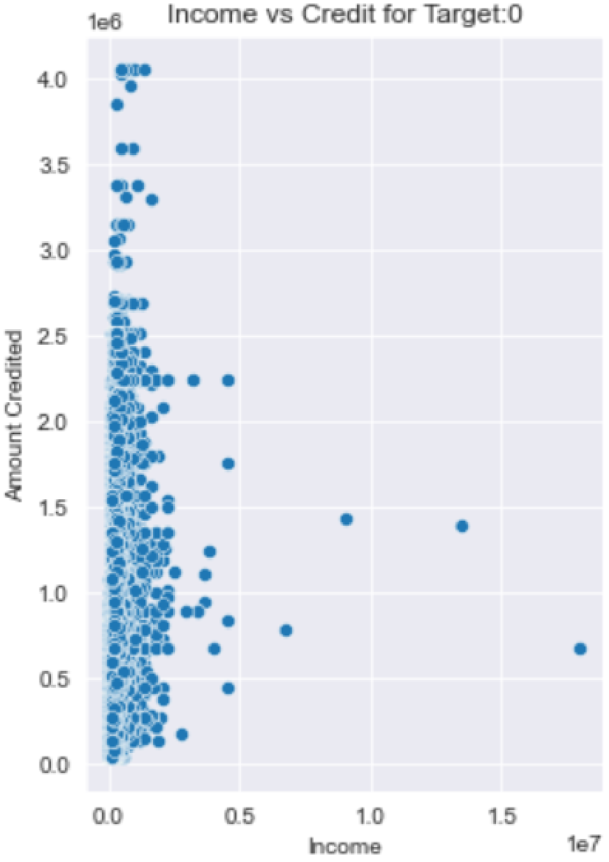
Income range comapred with Gender(With timely payments)

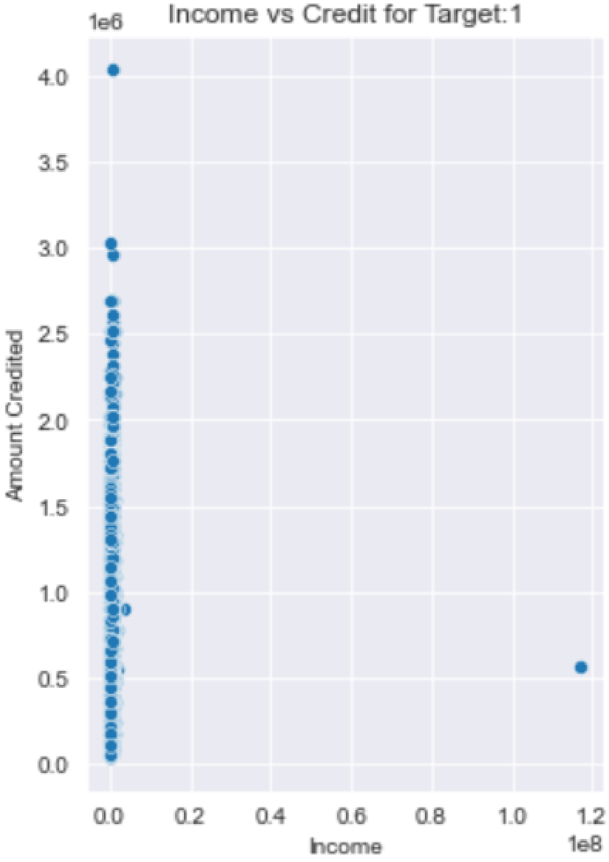


Income range compared with Gender(With payment difficulty)

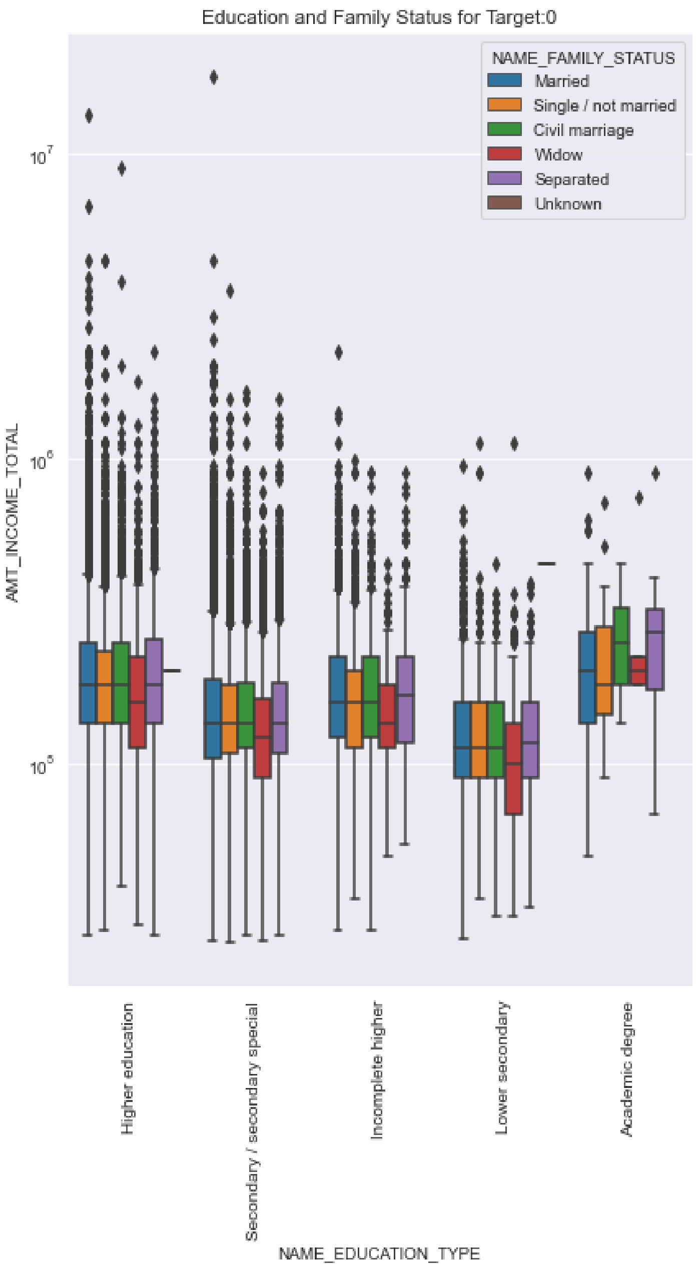


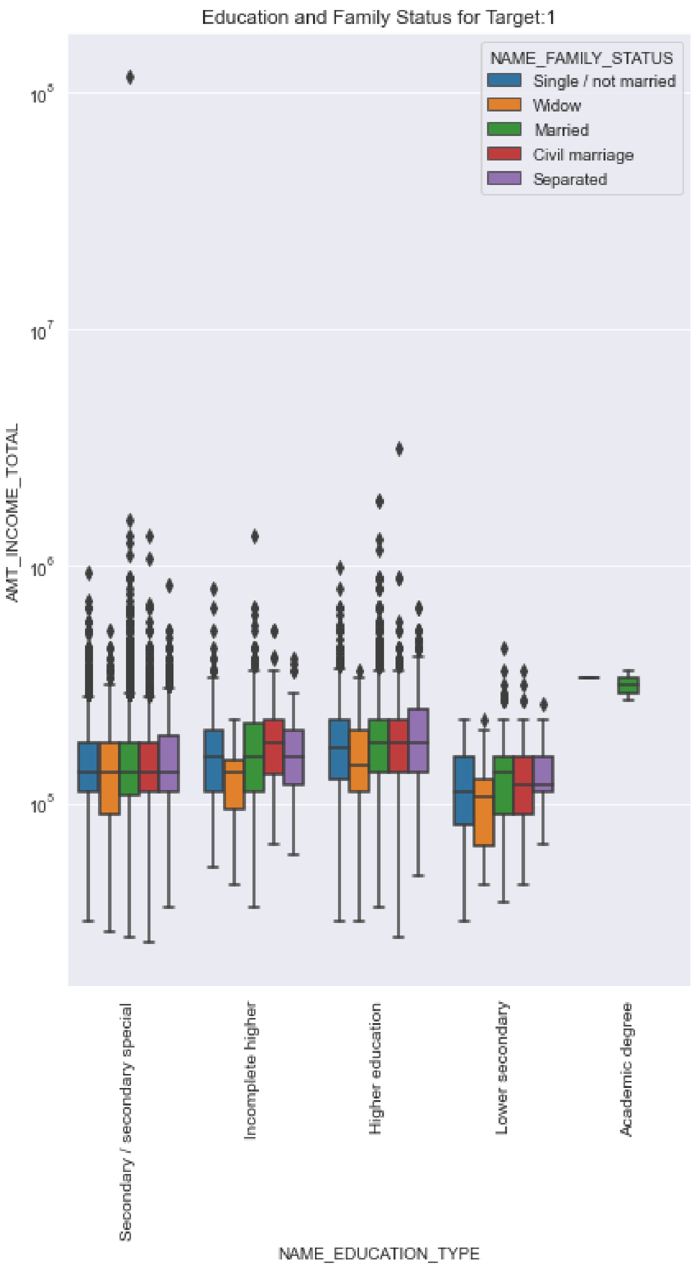
Females with low income are likely to make timely payments and can be targeted.





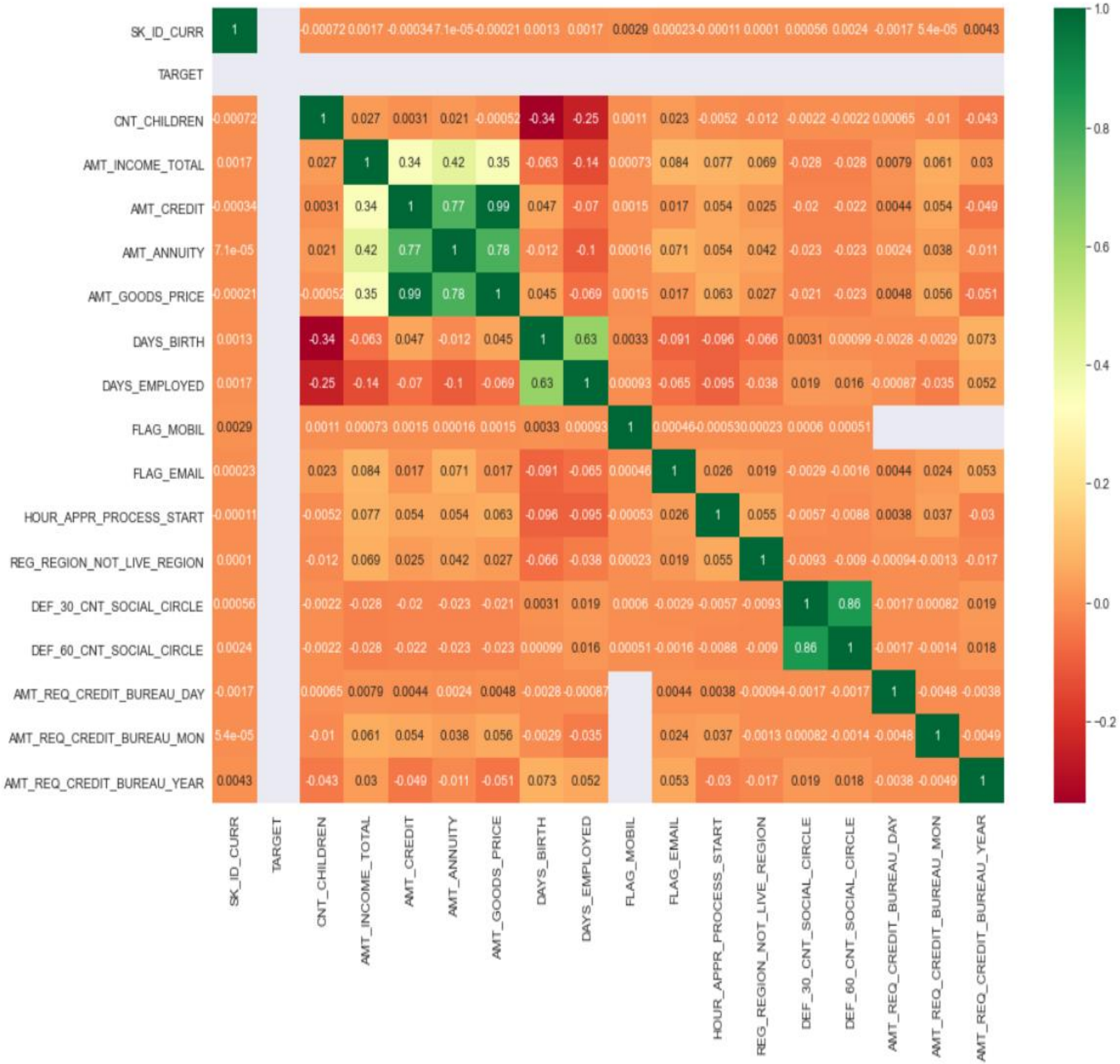
Although not very concrete correlation can be seen between the Income and credit loan but roughly, people with more income apply for more amount of credit and do not have payment issues as compared to people having low income





1.People with higher education are most likely to make clean payments

2.Combination of higher education and married marital status might result in successful deals



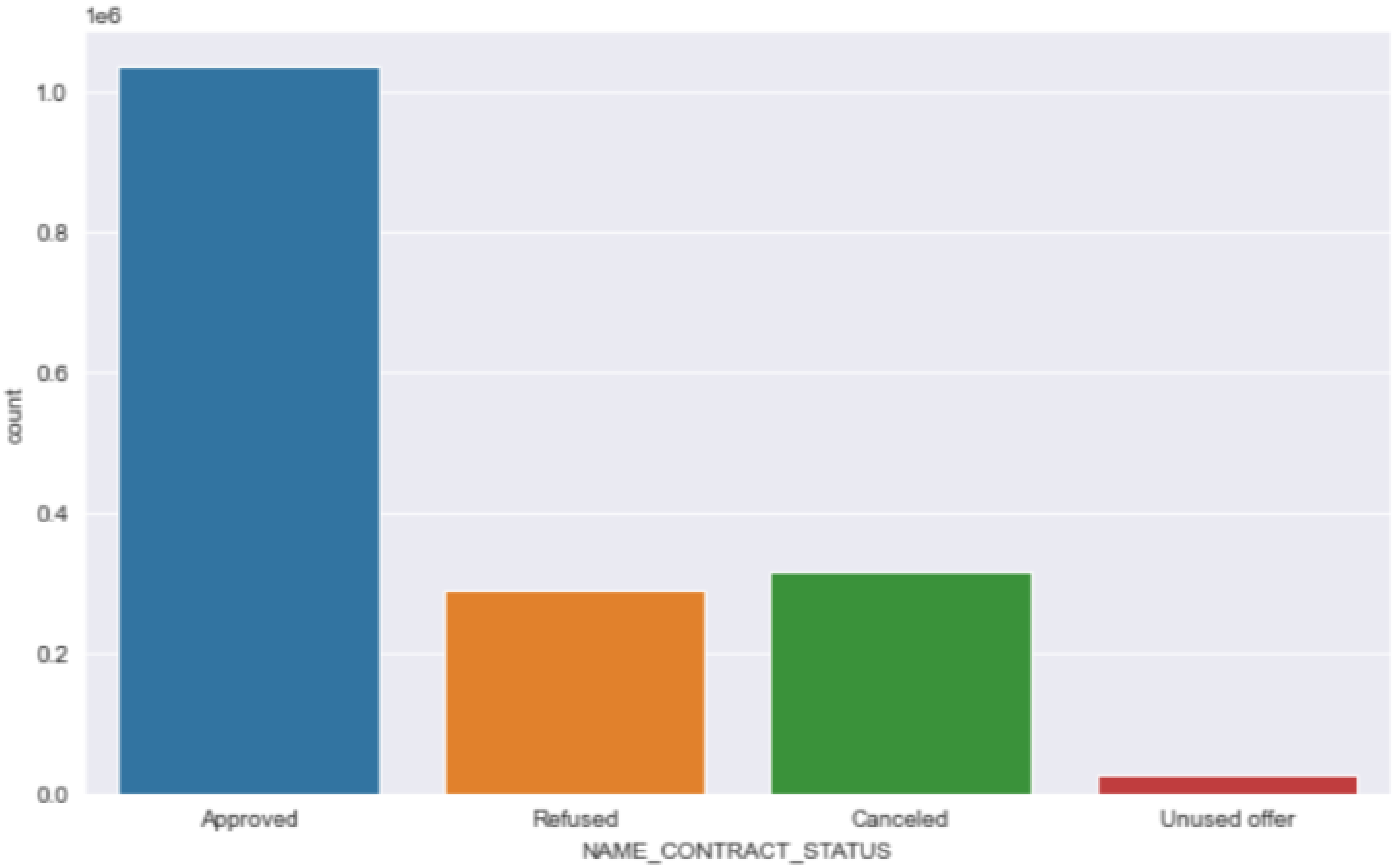
From the above we can find out following:

1.AMT_CREDIT AND AMT_GOODS_PRICE are highly correlated

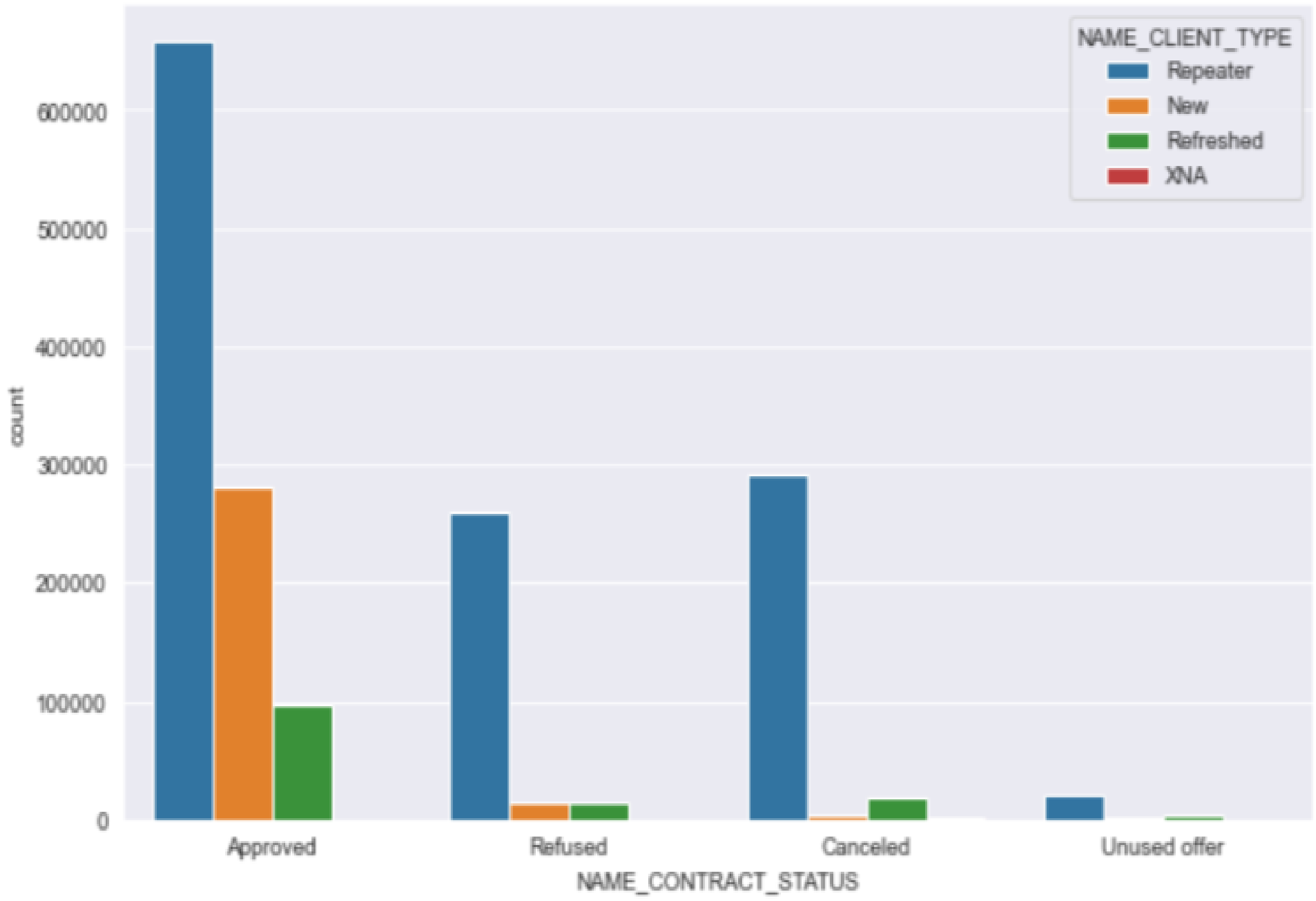
2.DEF_30_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE have strong correlation.

Previous application Dataset analysis

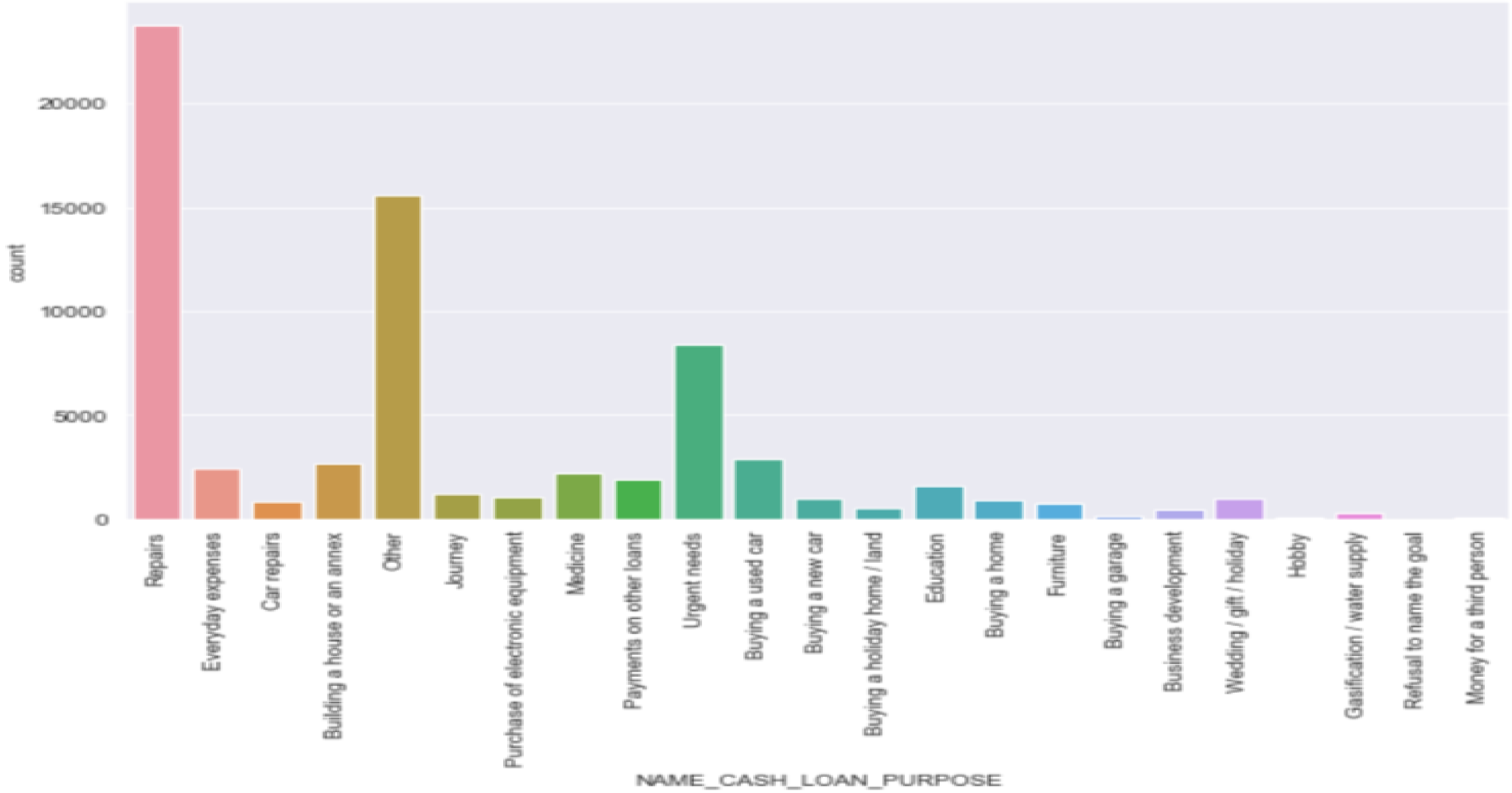
Previous_application dataset analysis



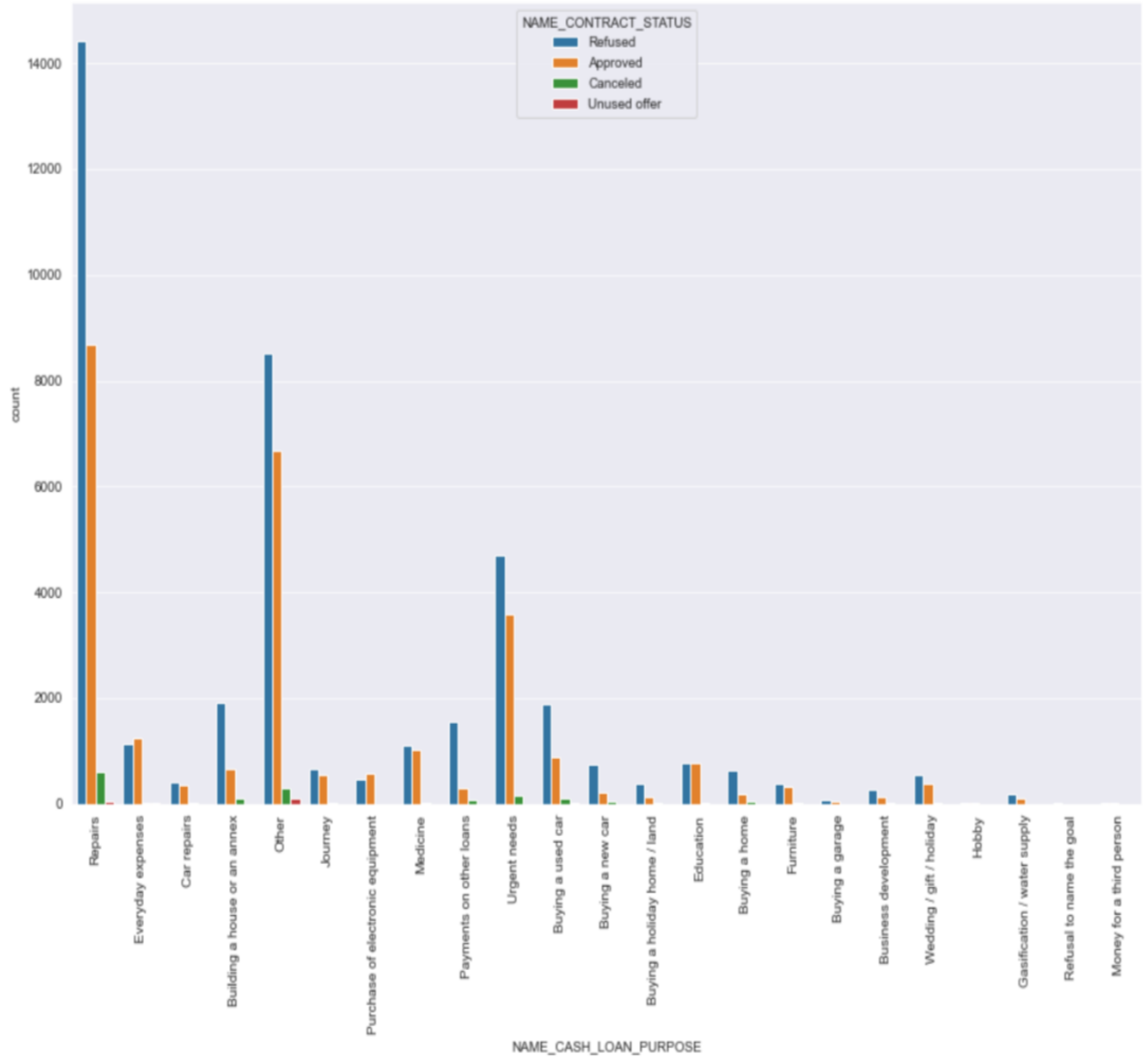
Majority of the loans were approved in the Previous_application dataset as well



Clients who had previously applied for loan got a good number of approval as compared to new or refreshed applications

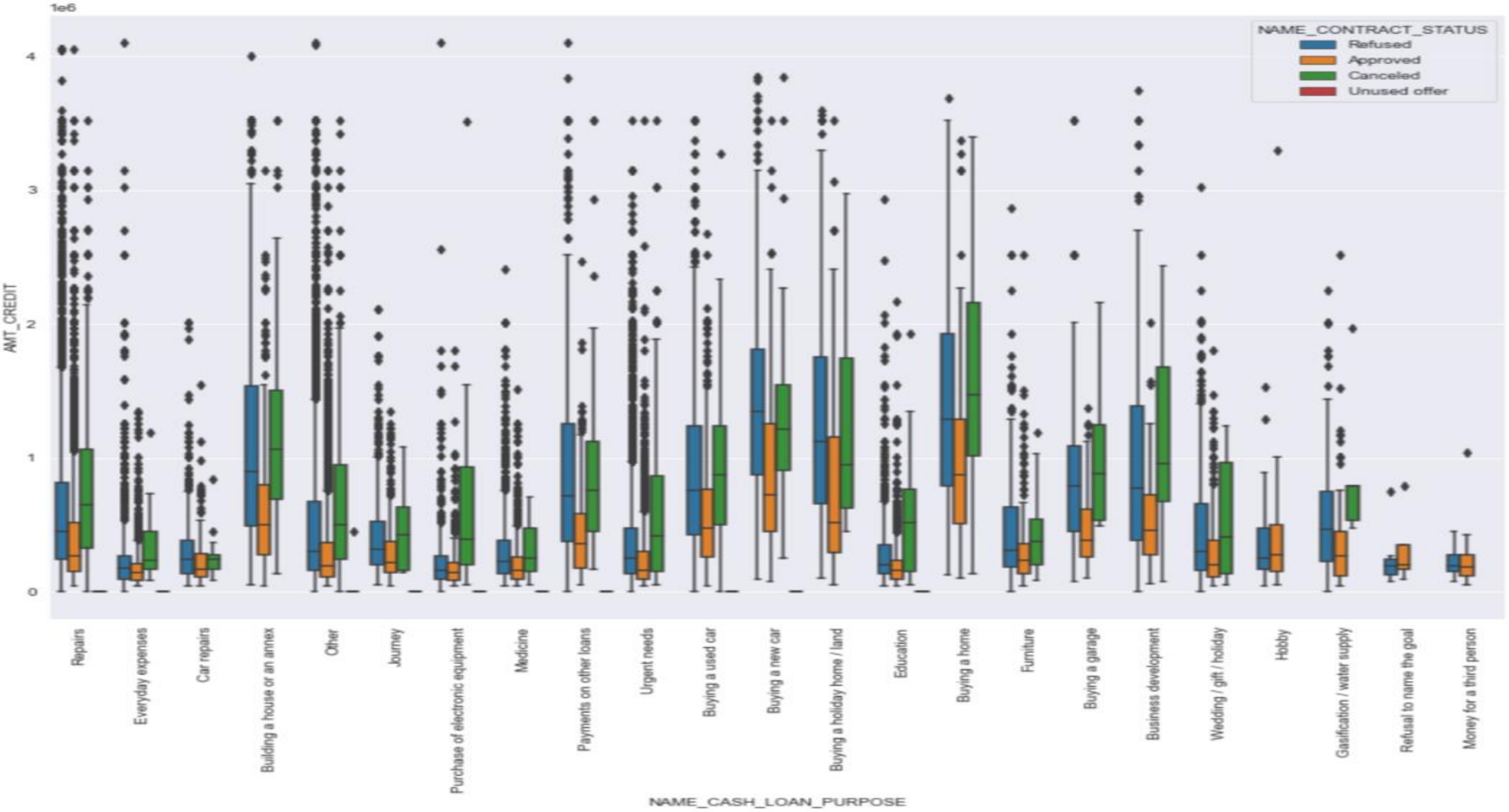


The most common purpose for applying for a loan is Repairs very clearly



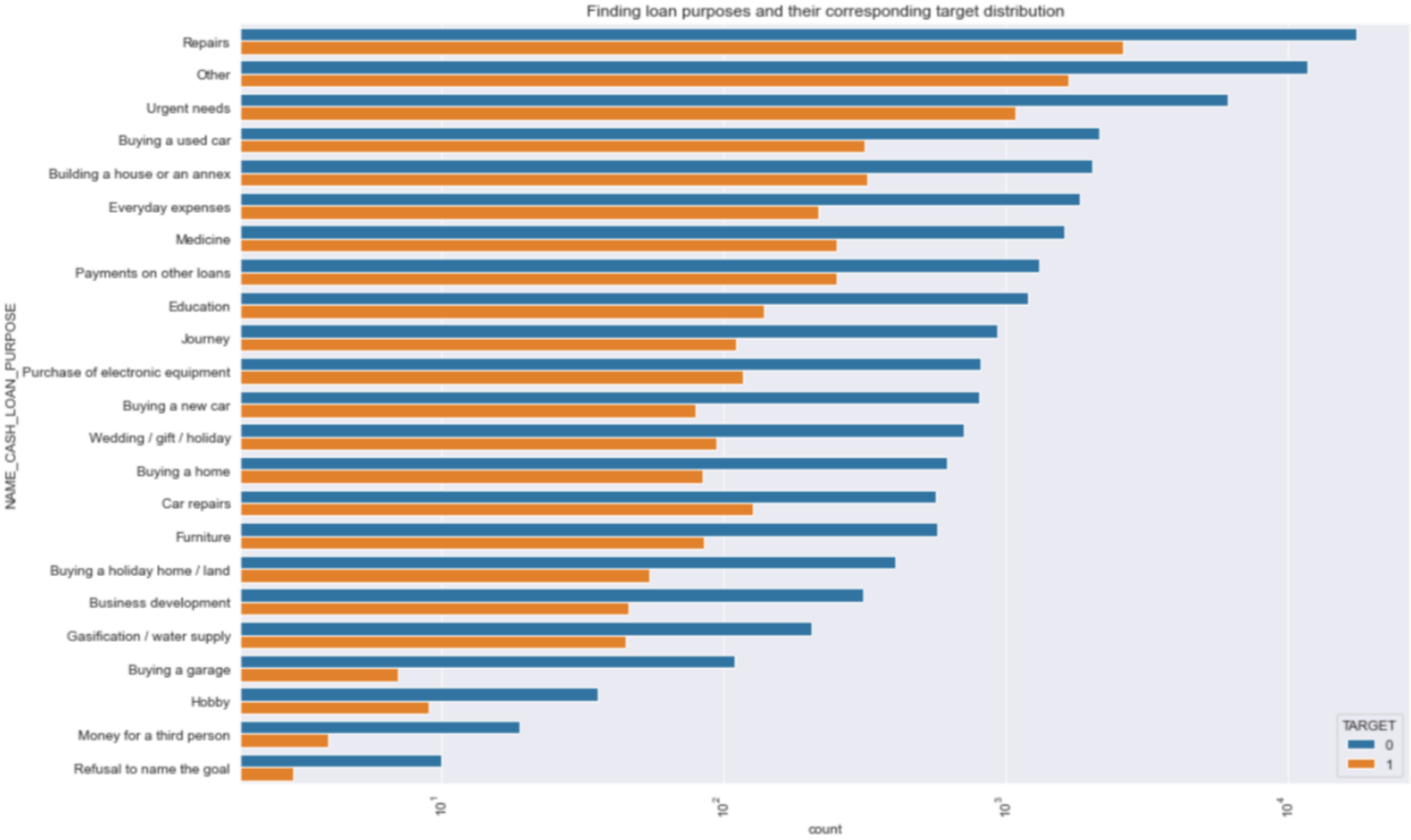
1.Loans for repair were refused the highest

2.Education loans have equal approvals and rejections



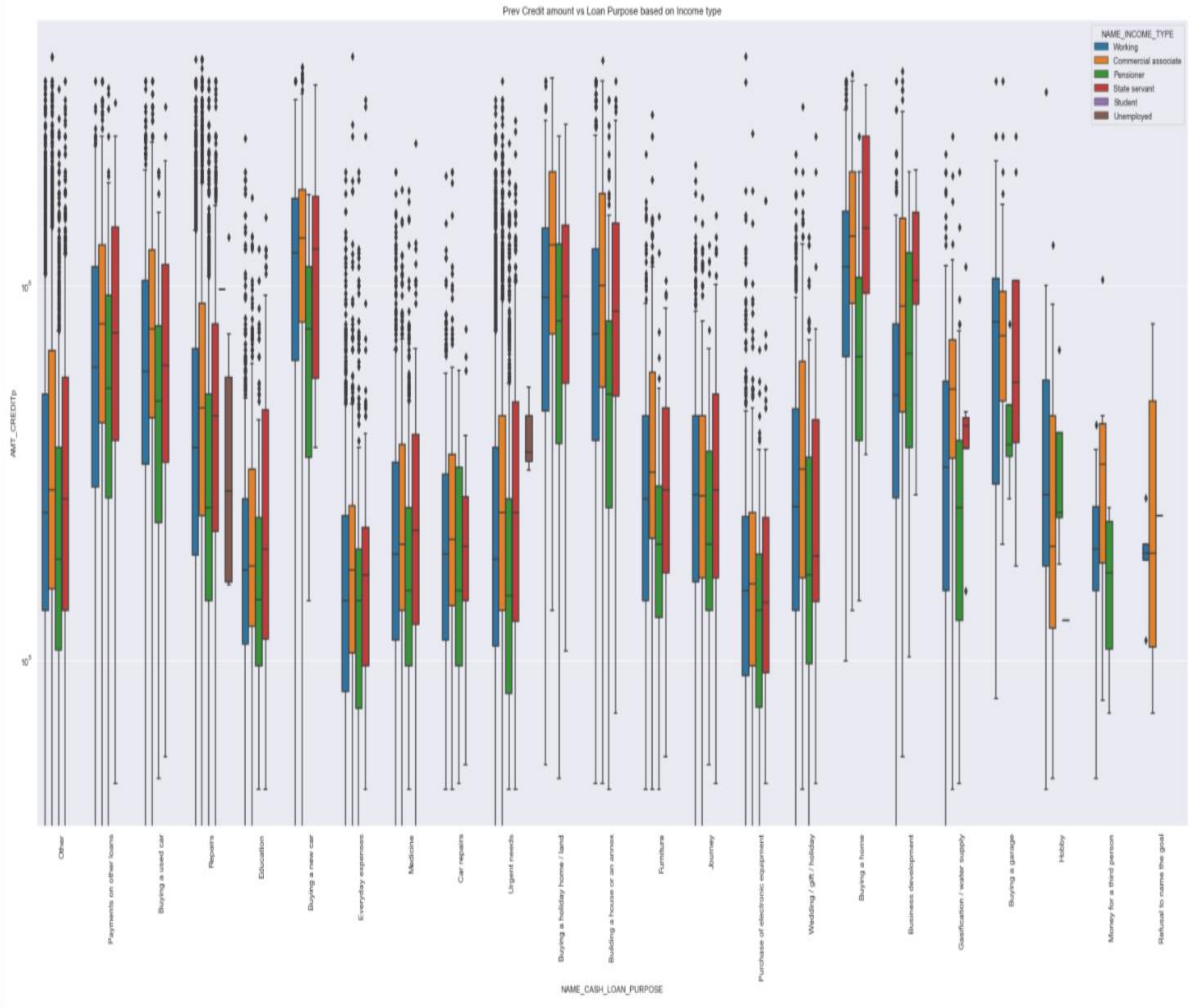
Buying a new car /Land /Home have highest approval rated and equally ranging credit amount

Analysis on the merged dataset

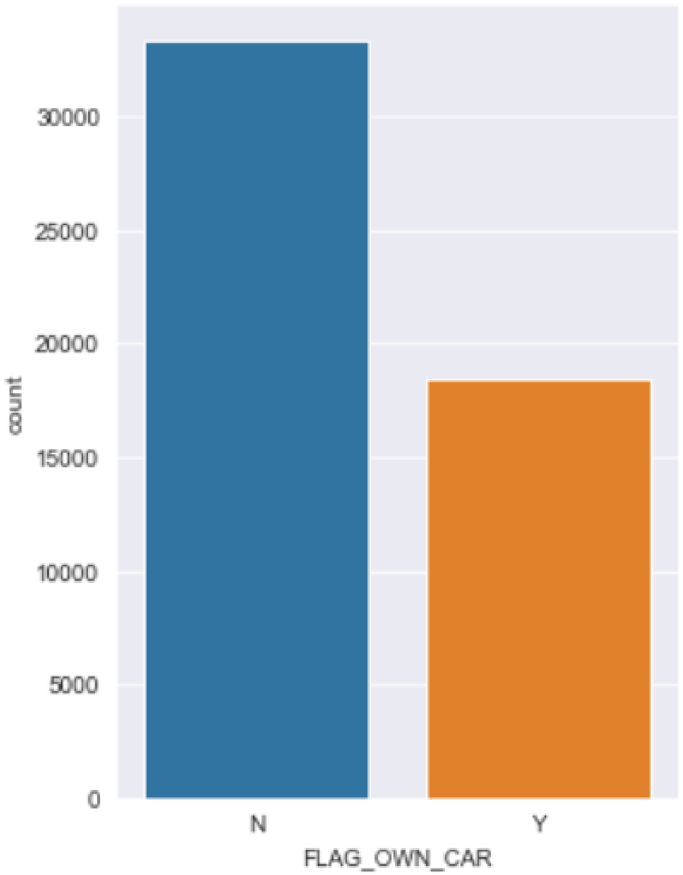


1.We can see that people who taken loan for repairs are facing difficulties.

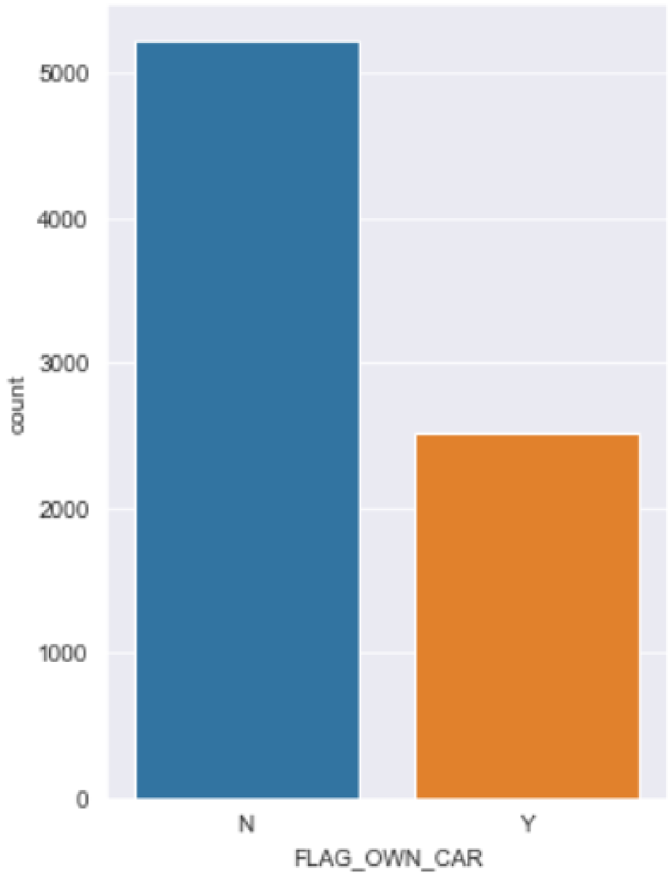
2.Loans should be given for Buying land/home , business development ,buying car.



It can be seen that majority of people from all income types excluding student and unemployed are looking for buying a car, buying a land, Building a house ,Buying a home and the amount credited is also high . Commercial associates did apply for a loan but didn't state the purpose and credit range is also at lower side for them



Distribution of whether the non-defaulters own a car

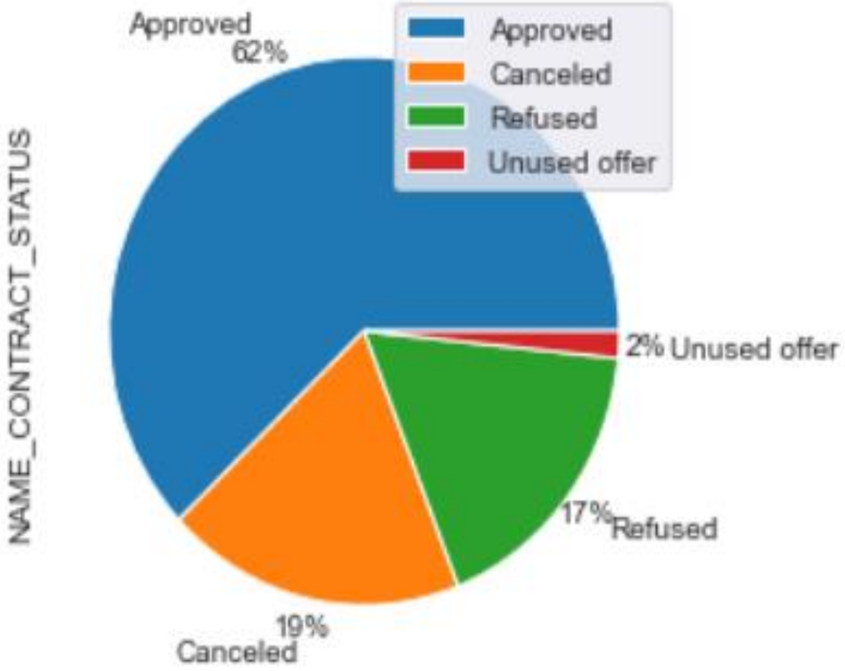


Distribution of whether the defaulters own a car

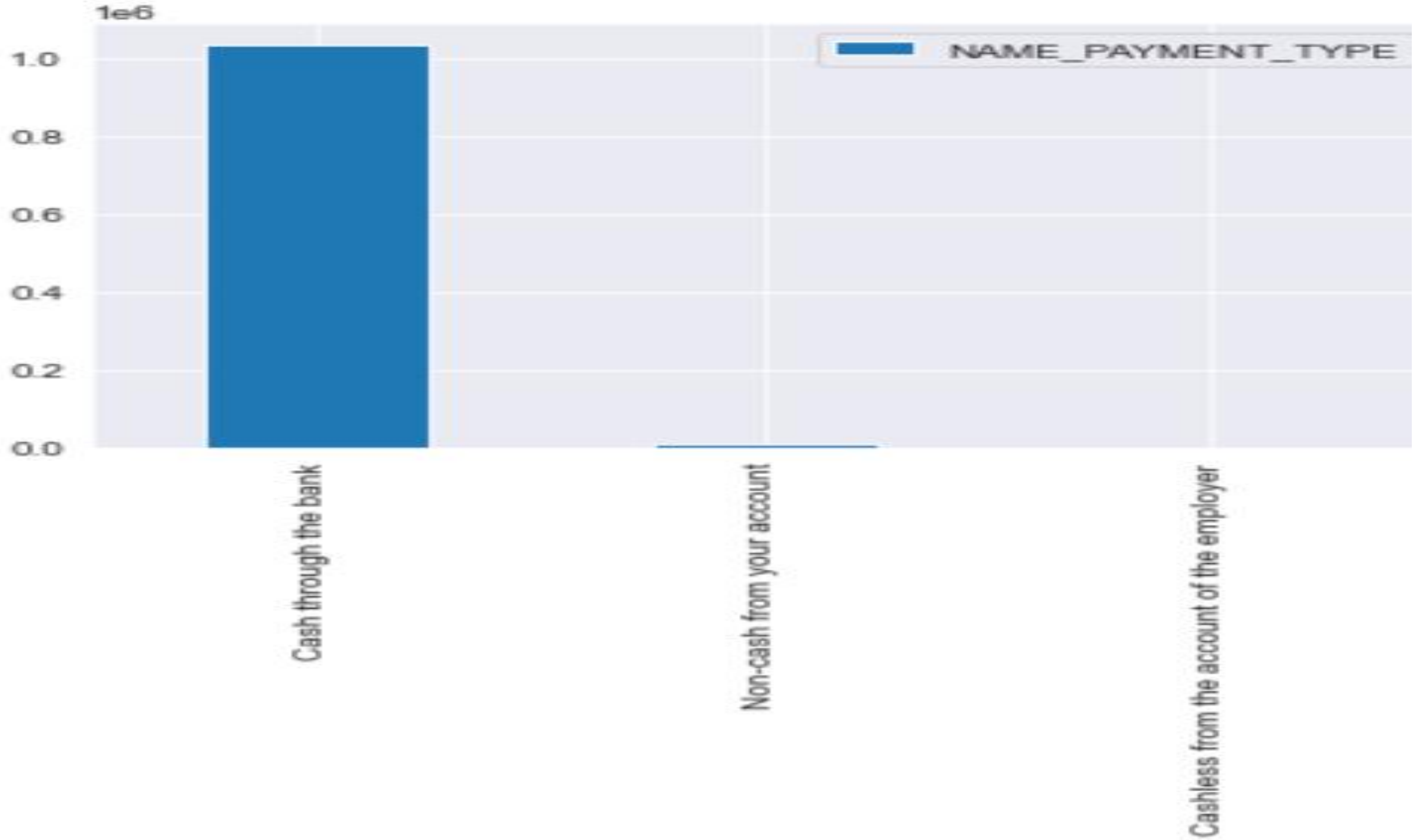
Those who do not own a car can be targeted more

Other insights which we could make:

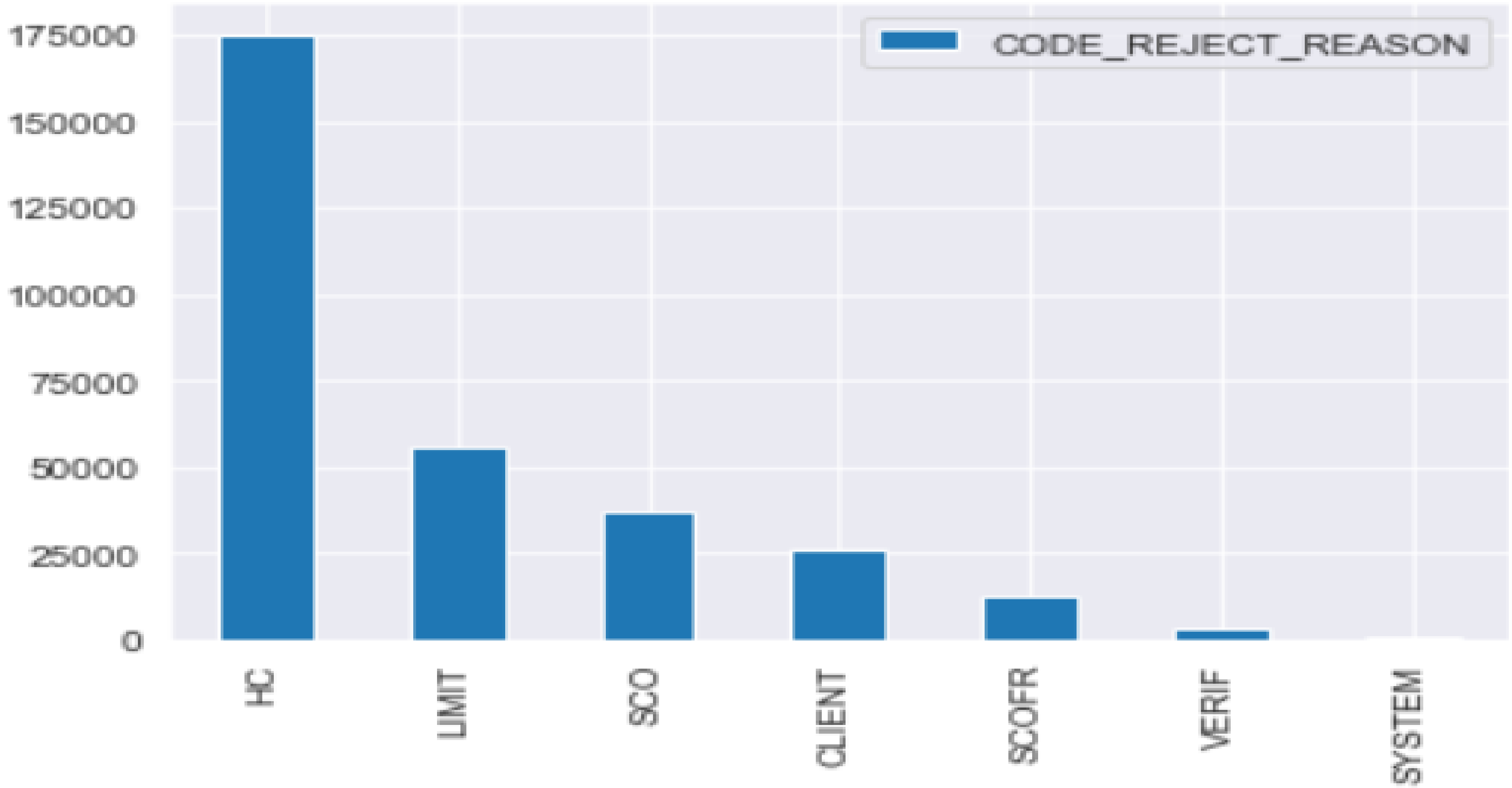
Other useful insights



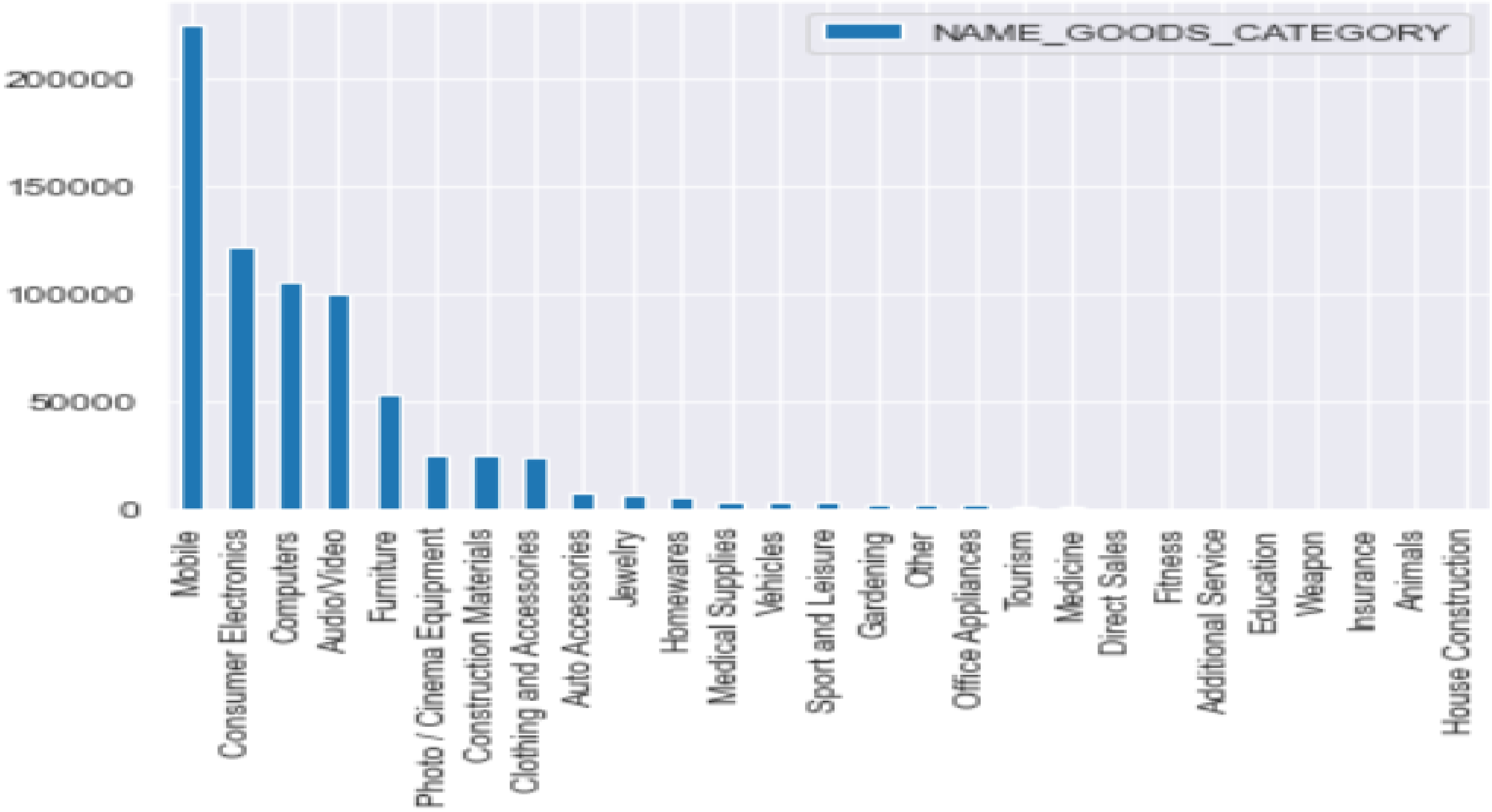
An overall percentage view of applications in previous_application dataset. Hence we observe that most of the loans fall under approved category



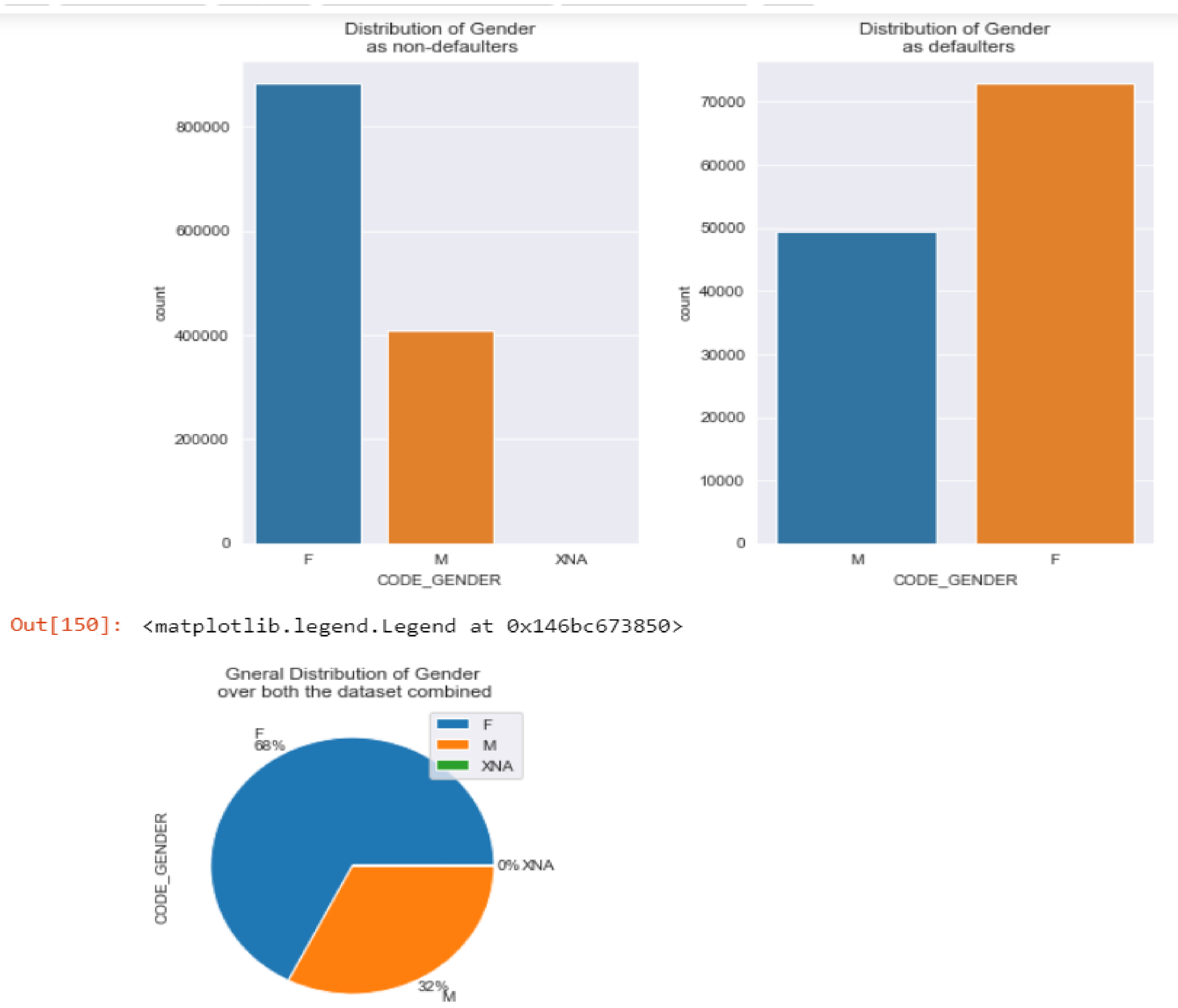
The most common payment type in previous application was cash through bank



--The most common reason of rejecting the application in previous applications was HC

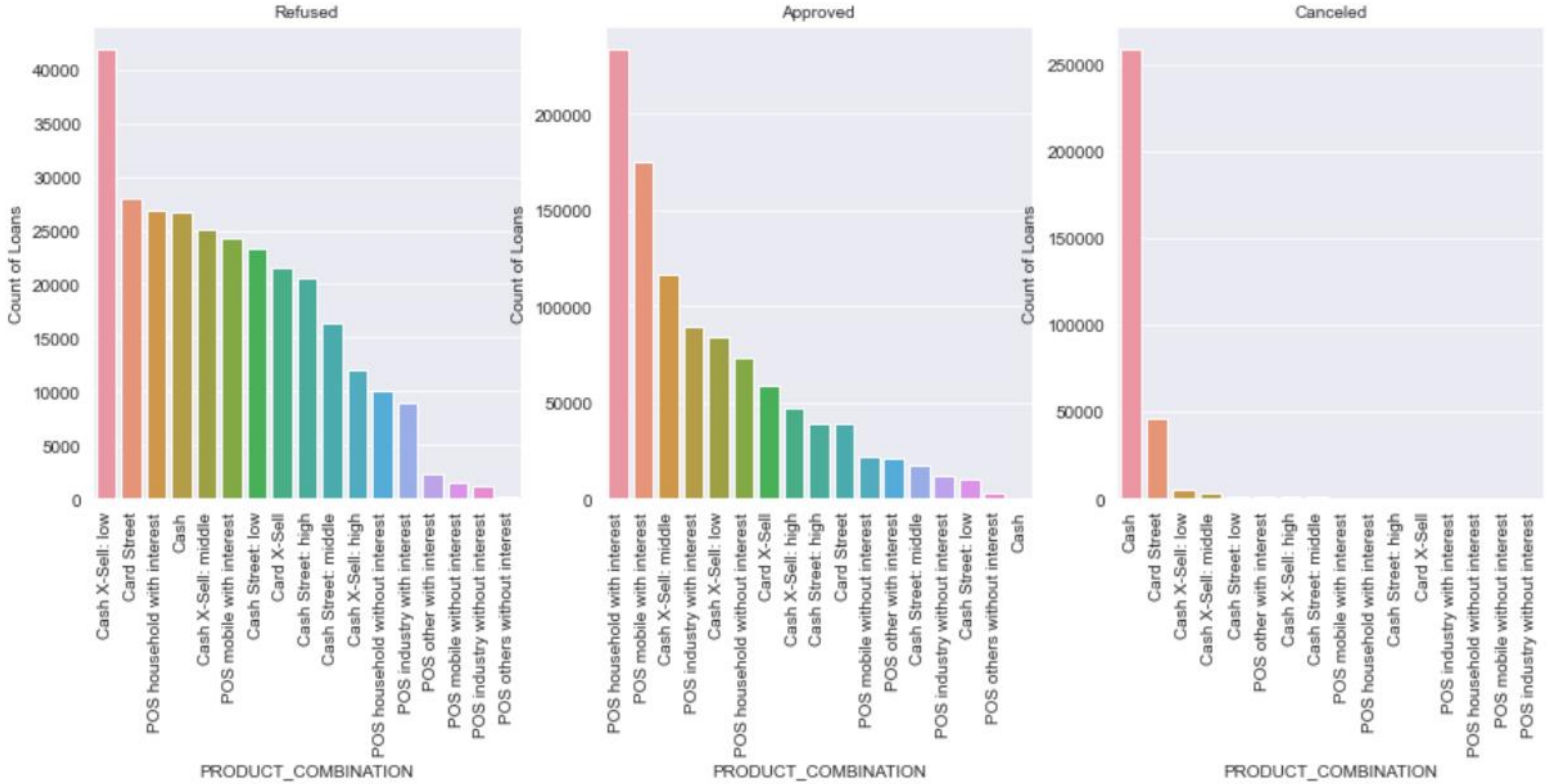


--Clearly a very good number of people took loan to buy mobile phones

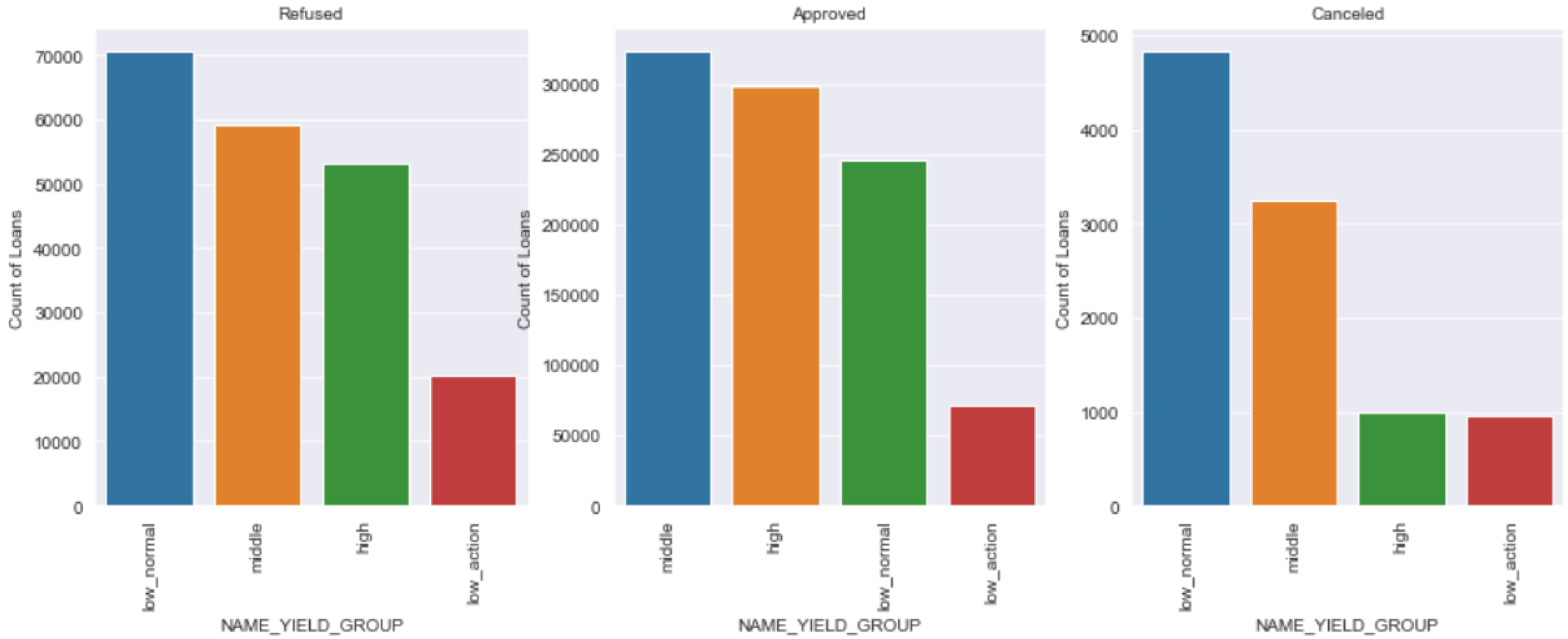


```
Out[150]: <matplotlib.legend.Legend at 0x146bc673850>
```

- The graphs shows gender imbalance and their percentages when it comes to defaulting



- We observe most number of loans were approved for POS household with interest.
- Most number of refused loans were of Cash X-Sell: Low Product combination
- Most Cancelled loans were Cash loans



- Most approved loans were from **Middle** Yield Group
- Most refused loans were from Yield Groups Not specified

Insights and Inferences:

Insights and Inferences:

- **Decisive variables that a applicant does not default on repayment:**

- NAME_EDUCATION_TYPE: Academic degree has less defaults.
- DAYS_BIRTH: People above age between 30-50 have low probability of defaulting
- DAYS_EMPLOYED: Clients with 40+ year experience have less chance of defaulting
- AMT_INCOME_TOTAL: Applicants with income more than 700,000 are less likely to default
- NAME_CASH_LOAN_PURPOSE: Loans bought for Hobby, Buying garage, home, car are being repayed mostly.

- **Decisive Factor that the applicant will become a Defaulter:**

- CODE_GENDER: Men are at relatively higher default rate
- NAME_FAMILY_STATUS : People who have civil marriage or who are single default a lot.
- NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education
- DAYS_BIRTH: Avoid young people who are in age group of 20-30 as they have higher probability of defaulting
- DAYS_EMPLOYED: People who have less than 5 years of employment have high default rate.
- AMT_GOODS_PRICE: When the credit amount goes beyond a certain limit, there is an increase in defaulters.

- **The following attributes indicate that people from these category tend to default but they are high in number hence the bank's risk management group should come up with ideas to mitigate the risk of offering them loan .Eg: increasing the interest rate.**

- NAME_HOUSING_TYPE: Our analysis suggests that High number of loan applications are from the category of people who live in Rented apartments & living with parents.
- AMT_CREDIT: People who get loan for more than 300k tend to default more hence by increasing the interest we can mitigate the risk.
- AMT_INCOME: Most of the population are having salary less than 300,000 make defaults in payment, we might consider to increase their interest to avoid any business loss.
- NAME_CASH_LOAN_PURPOSE: Loan taken for the purpose of Repairs seems to have highest default rate. A very high number applications have been rejected by bank or refused by client in previous applications as well which has purpose as repair or other. This shows that purpose repair is taken as high risk by bank and either they are rejected, or bank offers very high loan interest rate which is not feasible by the clients, thus they refuse the loan. The same approach could be followed in future as well.

Suggestions to the bank based on the EDA:

How can the bank ensure less financial losses:

- Target more female customers as it was observed that they apply for the loans mostly and also they have less payment difficulties.
- It was observed that most of the customers fall within Low-Medium income ranges and also they had no payment difficulty. If we combine this with age , we can target people and lend loans to people in age group of 30-50.
- Bank should give low-medium ranged loans .
- As we saw the plots for Income_type , we can see that working professionals are more likely to have less payment difficulties and can be targeted more .After them State servants and pensioners can be give a small fraction of loan in the banks asset book.
- Customers who have no liabilities like own car/ house can be targeted
- From merged dataset ,we saw that customers who apply loans for ‘Repairs’ as the purpose they were mostly rejected .Banks should give loans to people who want to buy a house/land ,buy home , buy a car
- We saw that mobiles were highly purchased and as mobile can be seen as short term loans .Banks can run a offer for people who are looking to buy a mobile in order to draw more customers
- It was seen that people who also had previously applied for loans and with status as approved are high in number .The bank can give them loans as they would have the credit history

- According to us the company should not give loans to people who have payment difficulties .They can refer all the insights given in previous slide .If banks give loans to likely defaulters , it will surely result in financial loss to the company .
- Bank should try to reduce the Type 2 error . Even if they loose on this customer(likely to default) there is 50% probability that they might get a non defaulter as their next customer.
- In order to increase the probability of getting good customers banks can make sales strategy using the given insights and their combinations .

Note:

- In all we made two notebooks namely Final_Submission in which we kept a strict approach and the other named as Extra_insights .
- In Extra_insights we tried to divide the data with a different approach so that we don’t miss out on insights.
- The analysis was done over a week , hence we had to save the datasets and reimport them again on daily basis so in some cells you might see to.csv and read.csv code used .