Credit EDA Case Study

- Rajesh Chhablani

Problem Statement

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 installments of the loan in our sample

- All other cases: All other cases when the payment is paid on time.

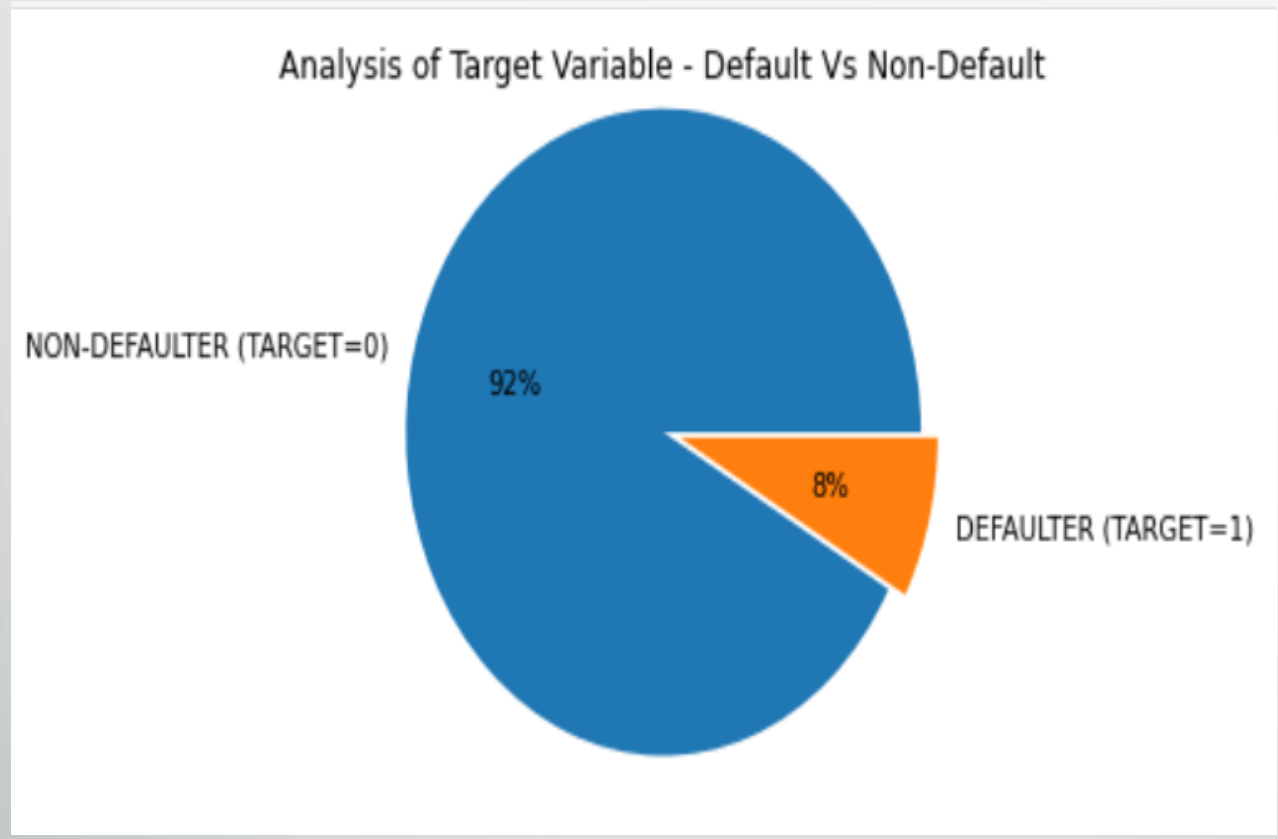
The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Solution Overall Approach

We have used the below approach for deriving the insights :

- The required libraries needed for data cleansing and visualisation are imported.
- We have done the data cleansing for columns wherever necessary and dropped the columns with majority of data as NA. Outliers are identified. Data imbalance is checked.
- Created new columns as per the requirements
- Univariate/Bivariate Analysis of the relevant Categorical/numerical is done and insights are derived
- Current and Previous application data is done to derive insights based on bank Approval loan status .

Data Imbalance

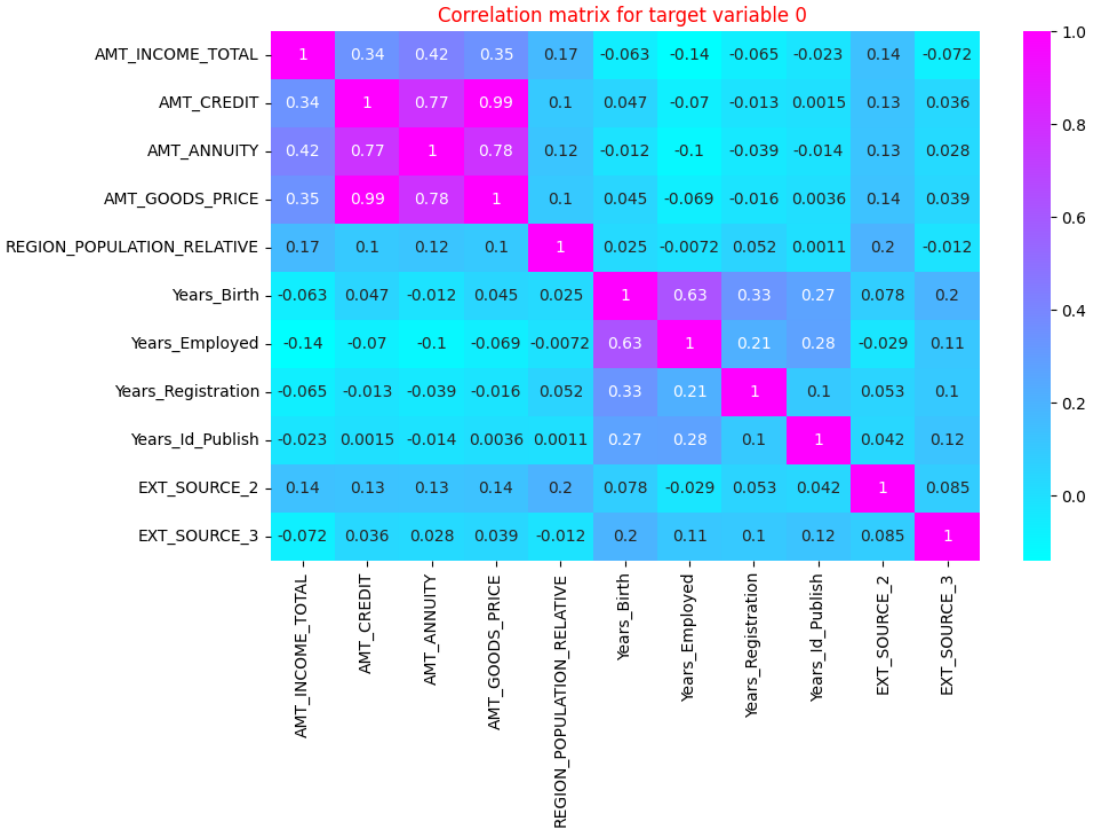


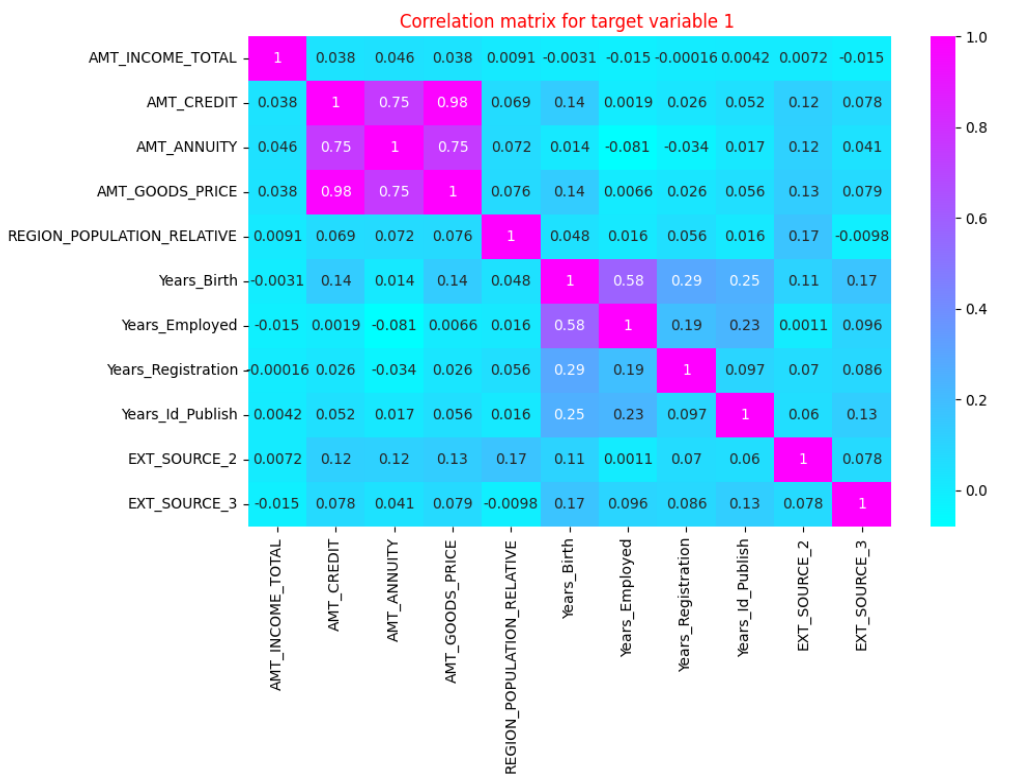
Ratio of Data Imbalance

11.39

In the given pie chart, we can conclude that 92% of people did not default compared to 8% of the people in the dataset, Therefore we can observe that there is a high imbalance between the defaulters and non-defaulters.

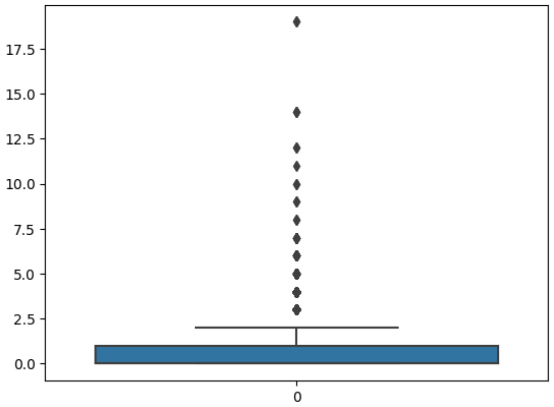
Correlation Mattrix / Heatmap

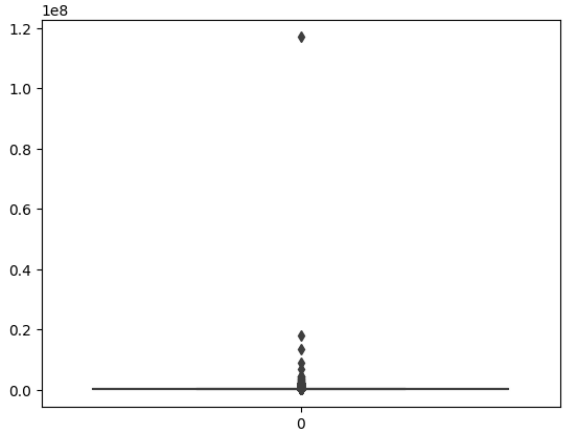


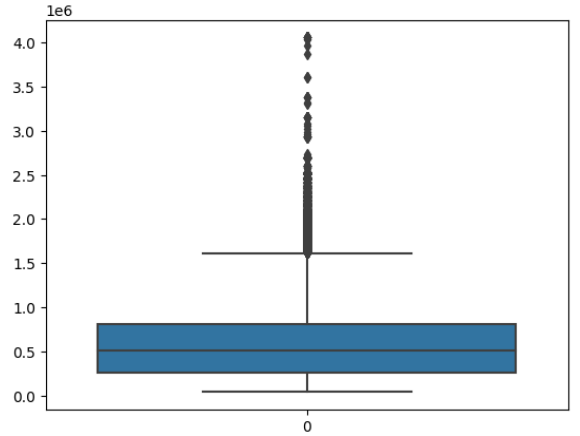


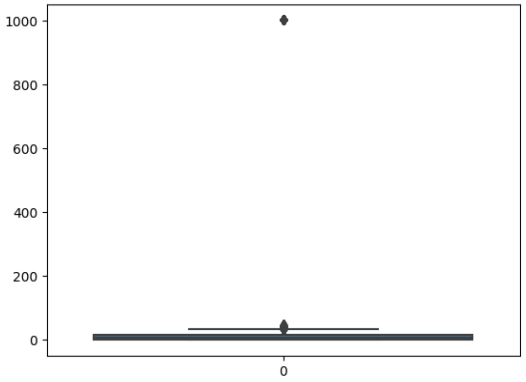
Through this correlation, using heatmap. we can conclude that correlations are almost same

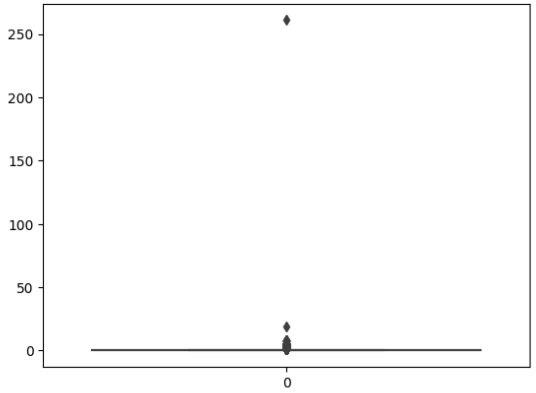
Outlier Identification





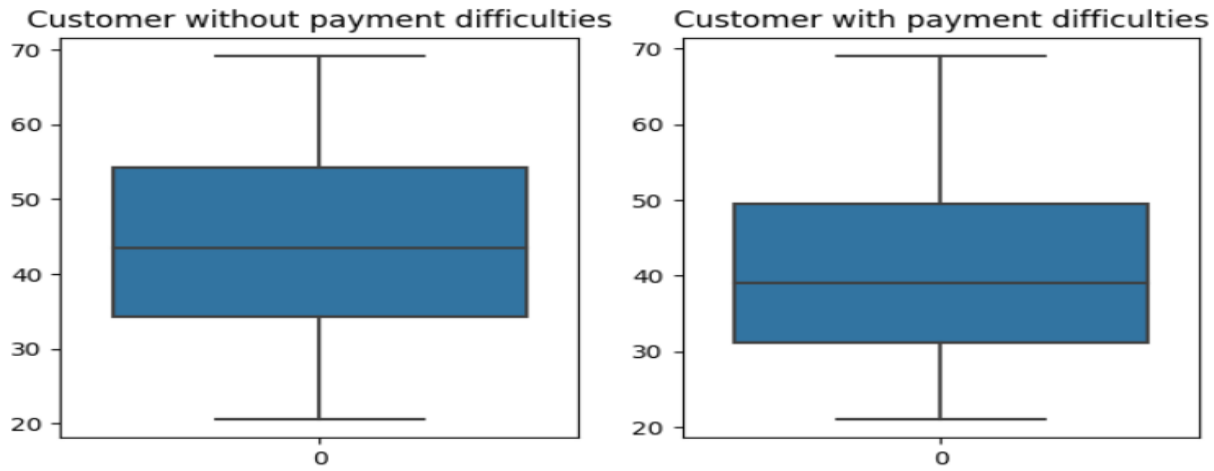




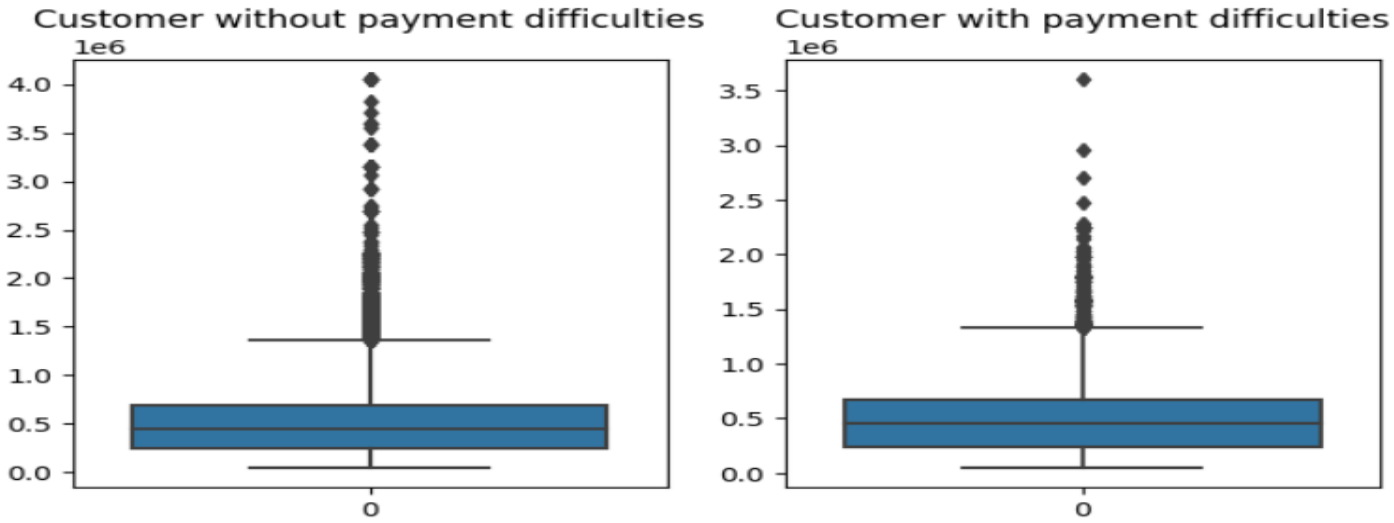


Univariate Analysis

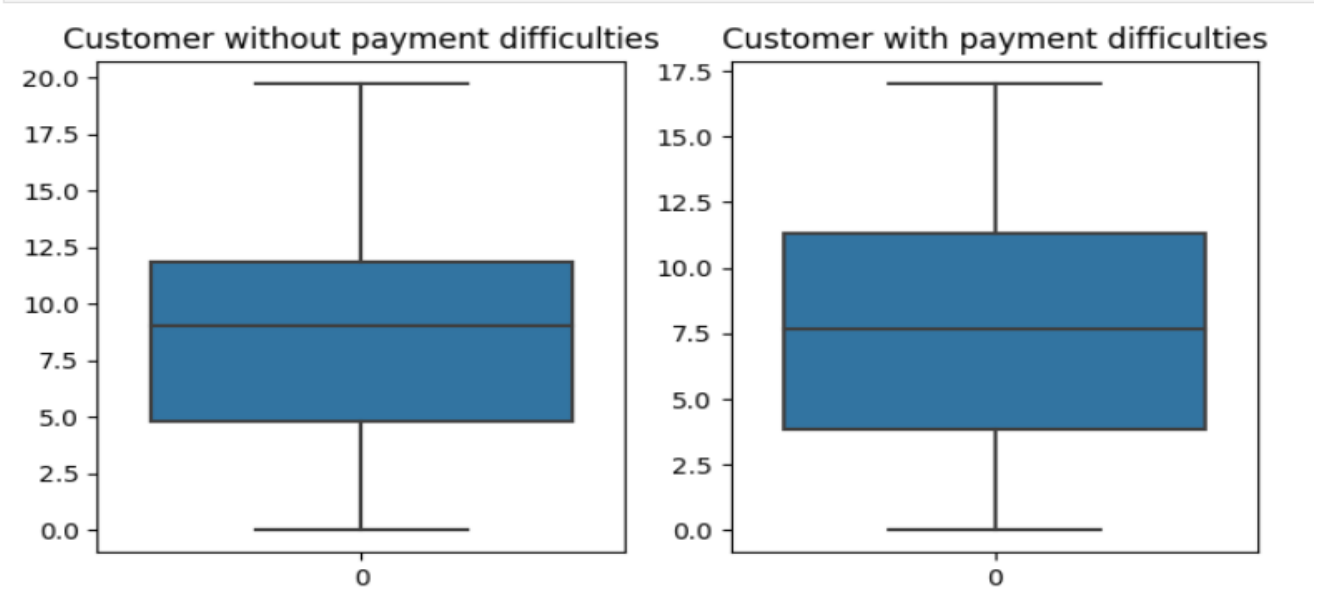
Numerical Variables



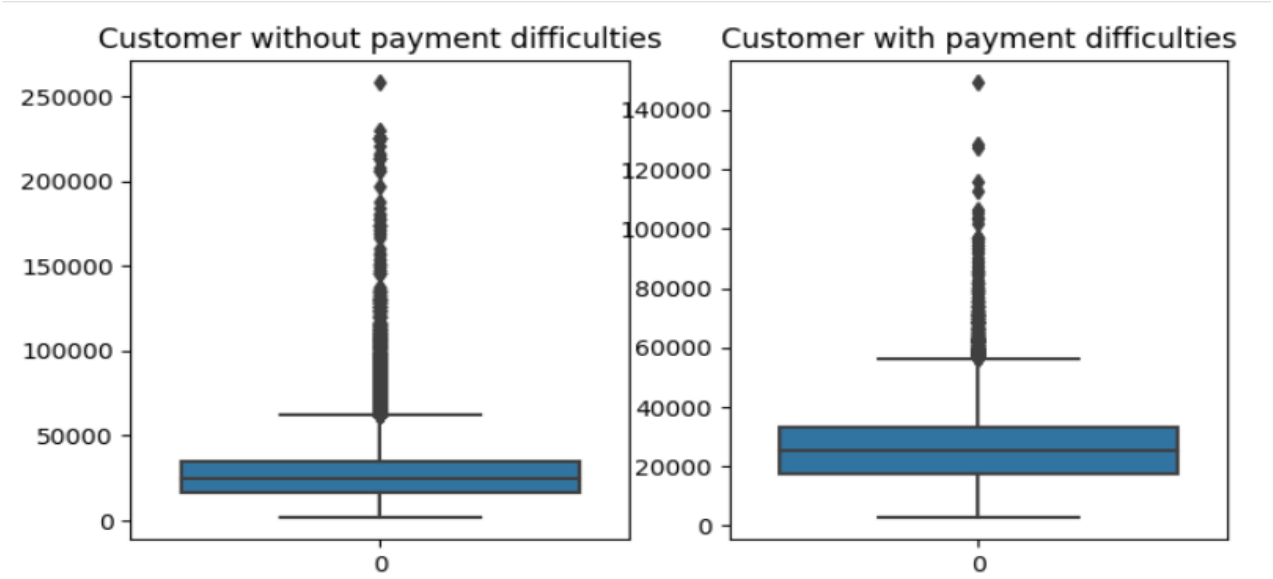
From the above box plot we can note that customer without payment difficulties having year in between 34 to 54 years , And customer with payment difficulties having in between 31 to 50 years.



From the above box plot we can note that customer without payment difficulties having AMT_GOODS_PRICE in between 0.3 to 0.7, and customer with payment difficulties having AMT_GOODS_PRICE in between 0.3 to 0.7

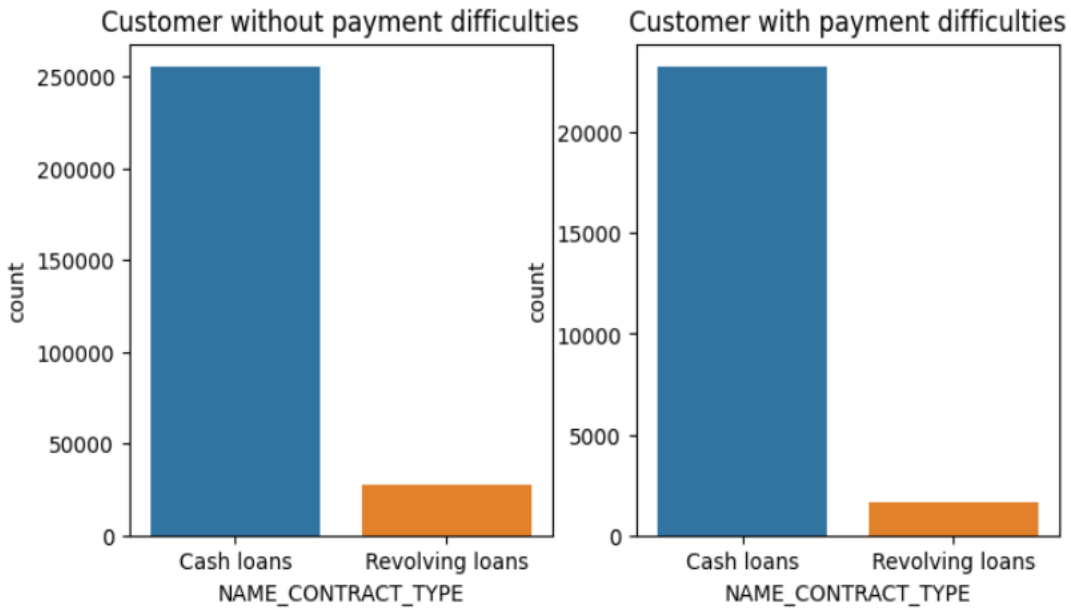


From the above box plot we can note that customer without payment difficulties having YEARS_ID_PUBLISH in between 5 to 11, and customer with payment difficulties having AMT_GOODS_PRICE in between 3 to 11 years.

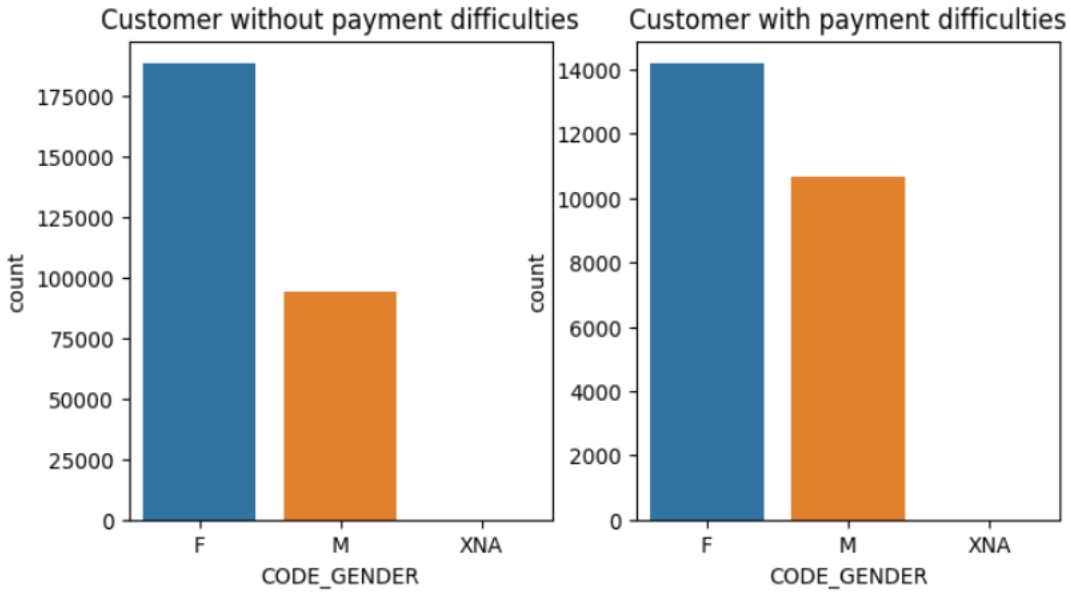


From the above box plot we can note that customer without payment difficulties having AMT_ANNUITY in between 20000 to 30000, and customer with payment difficulties having AMT_GOODS_PRICE in between 20000 to 30000 years.

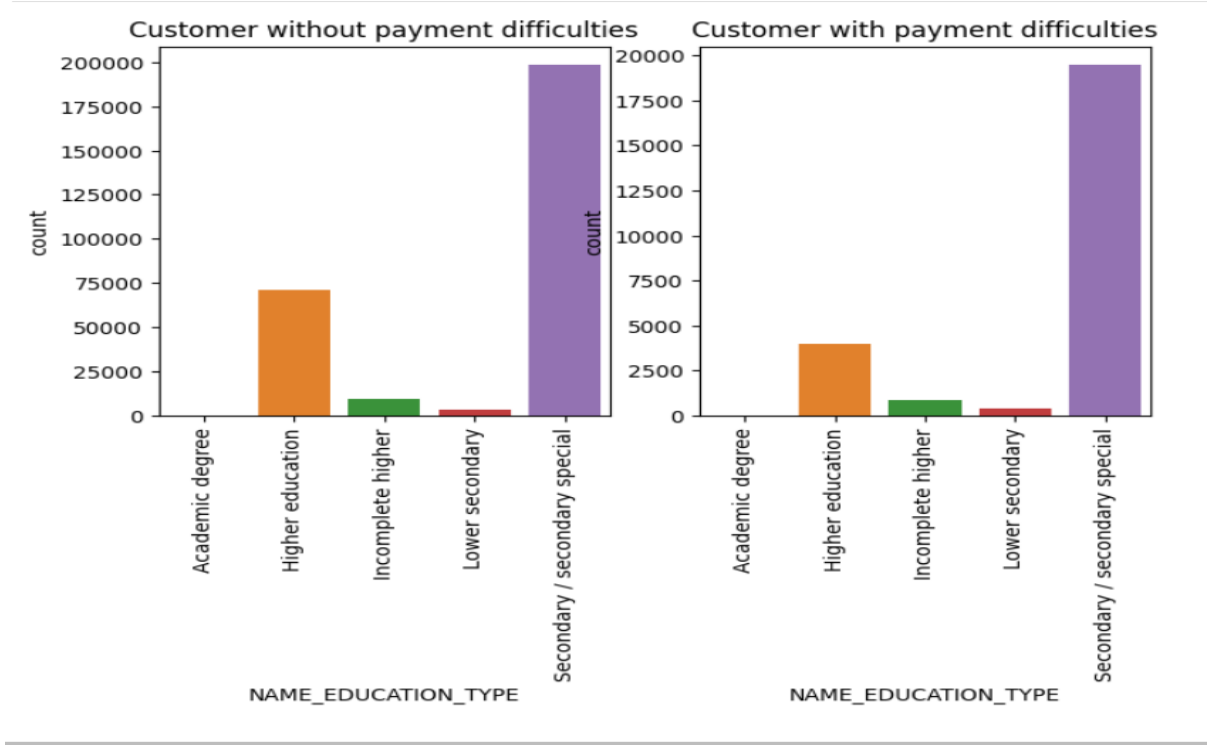
Categorical Variables



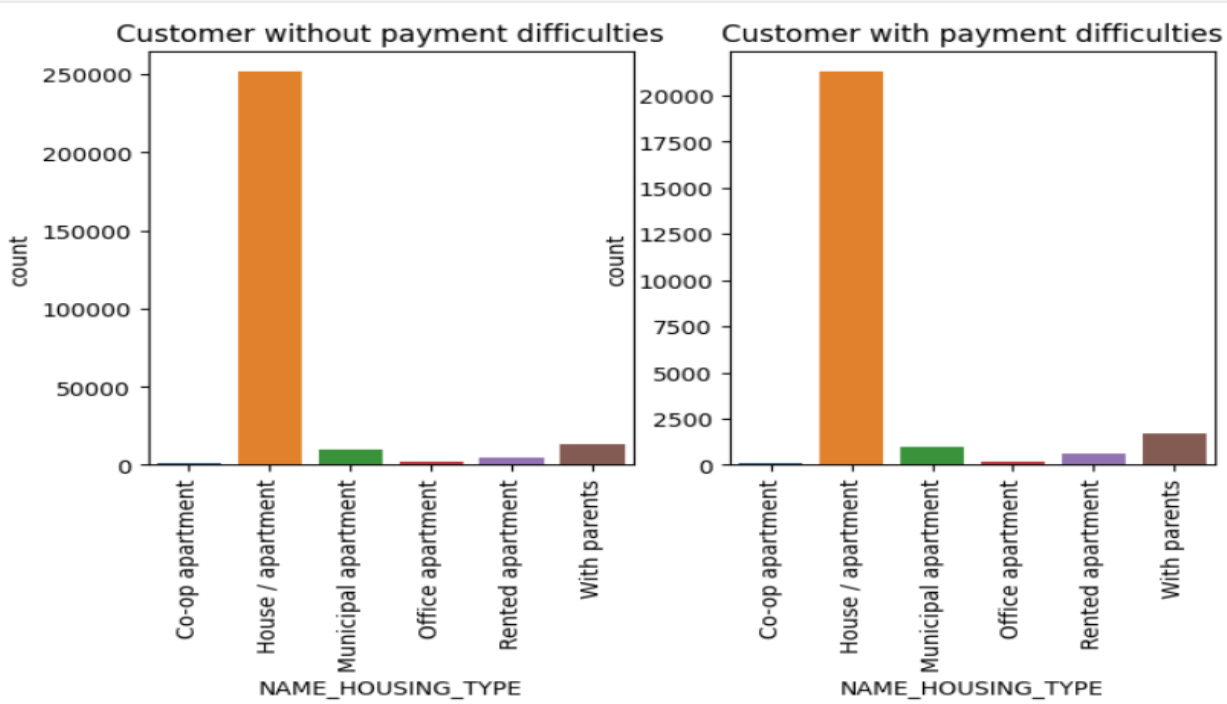
Here we can see that customer without payment difficulties and customer with payment difficulties are taking cash loans



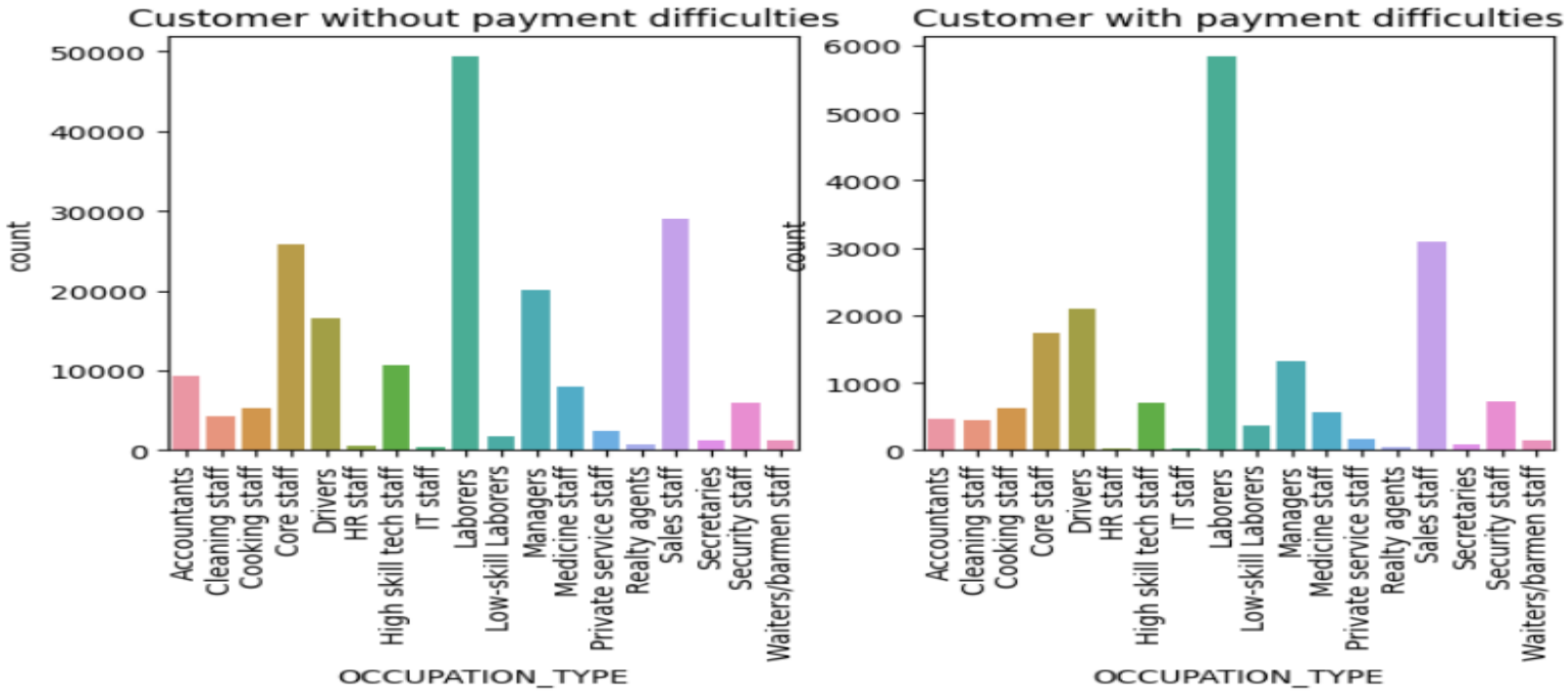
Here we can see that Females are more in number, in both cases



Here, we can see that secondary/secondary special are more in number in both cases



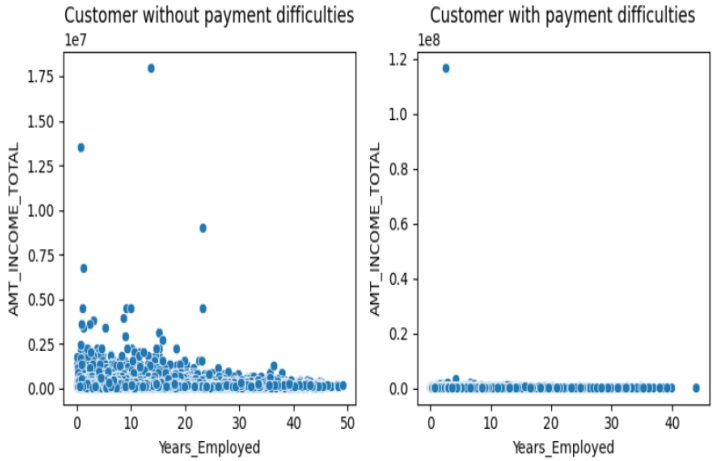
Here, we can see that House/Apartment category are more in number in both cases customers take more in number for House/Apartment category in both cases

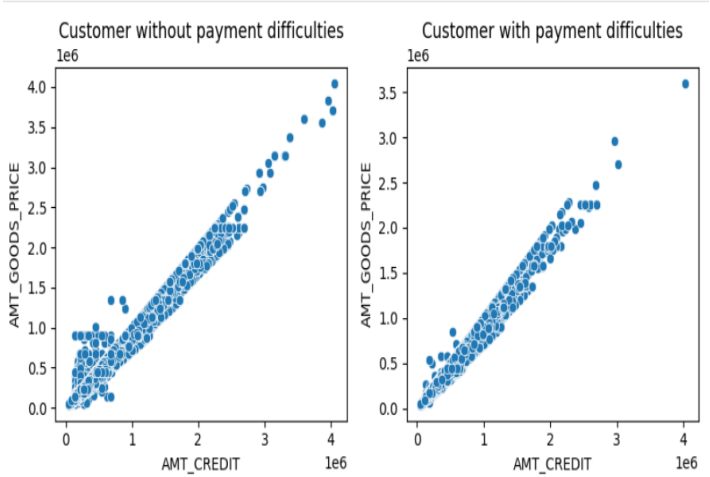


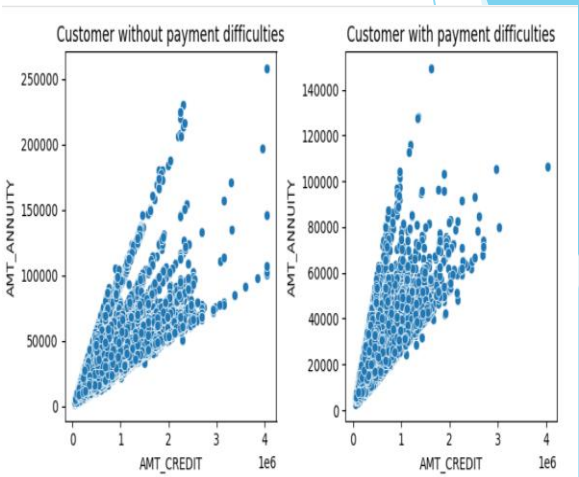
Here, we can see laborer are facing more difficulty in paying loans, sales staff and core staff too facing difficulty in paying loans laborers are more in number in paying of loans too

Bivariate Analysis

Numerical-Numerical bivariate analysis

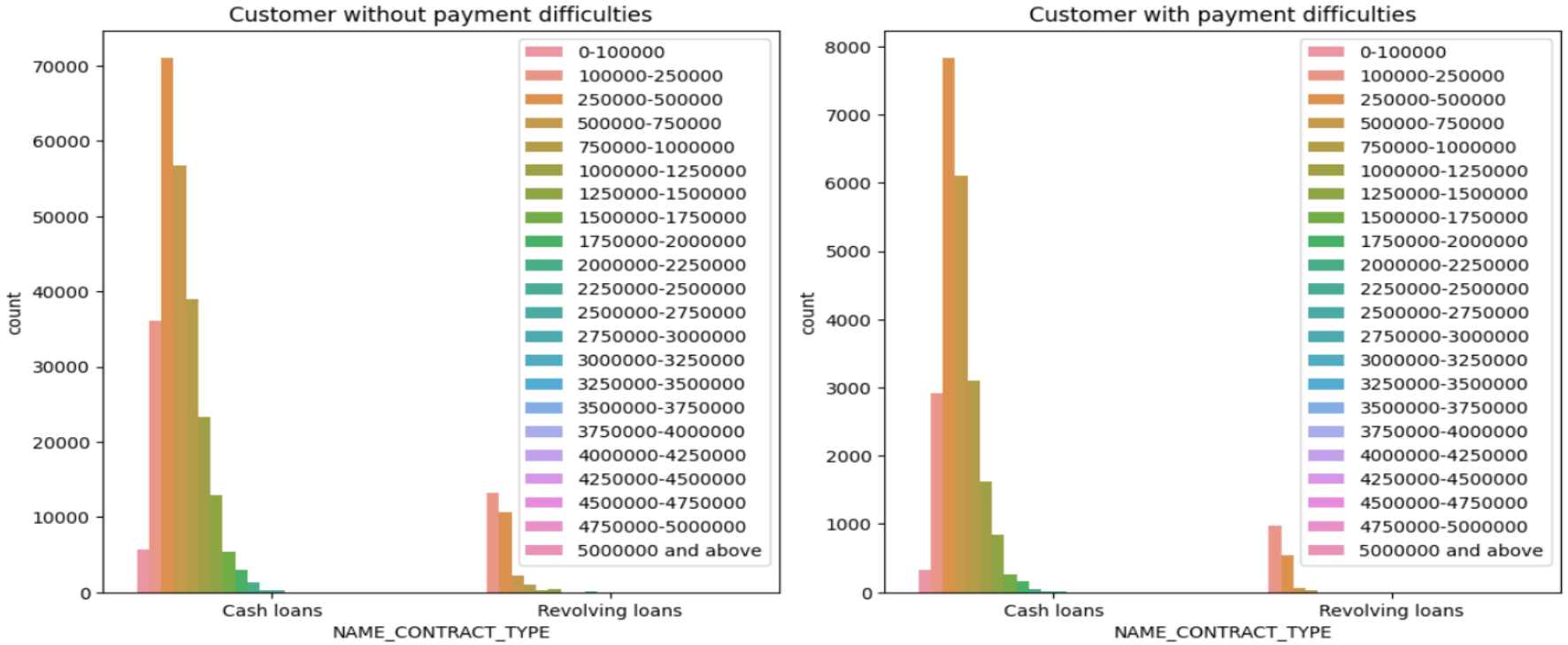




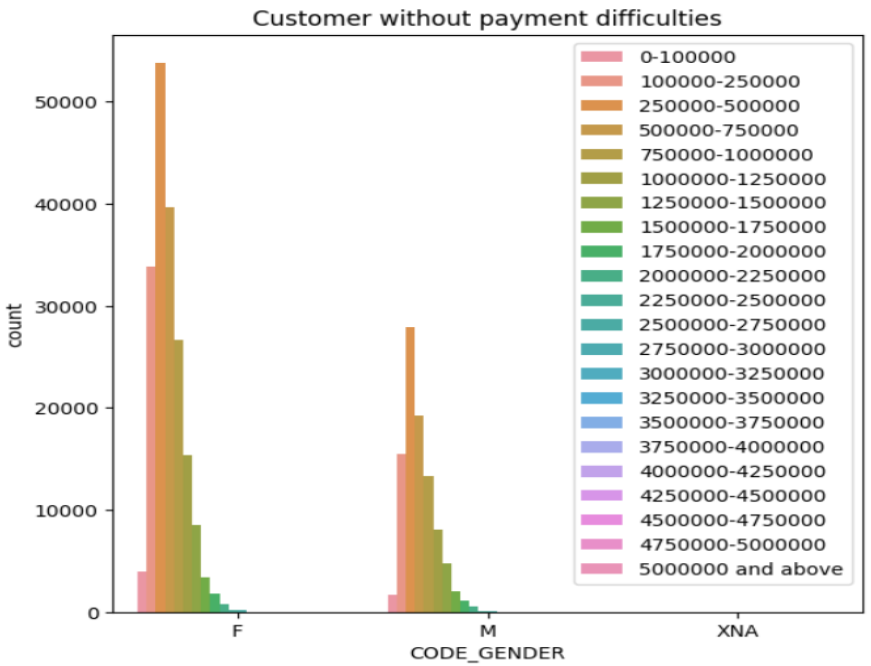


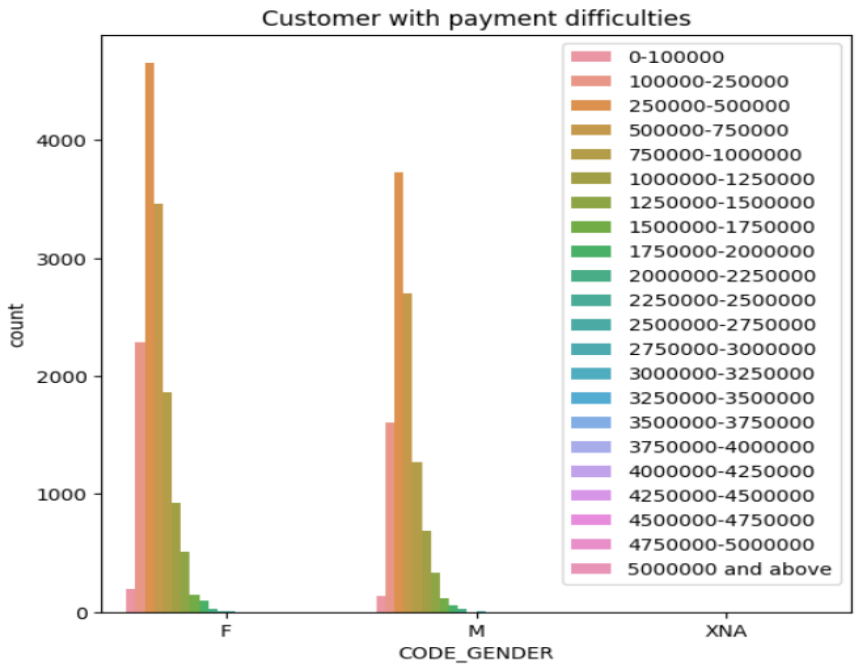
We can conclude that, people without payment difficulties take more credit for the annuity

categorical - categorical bivariate analysis



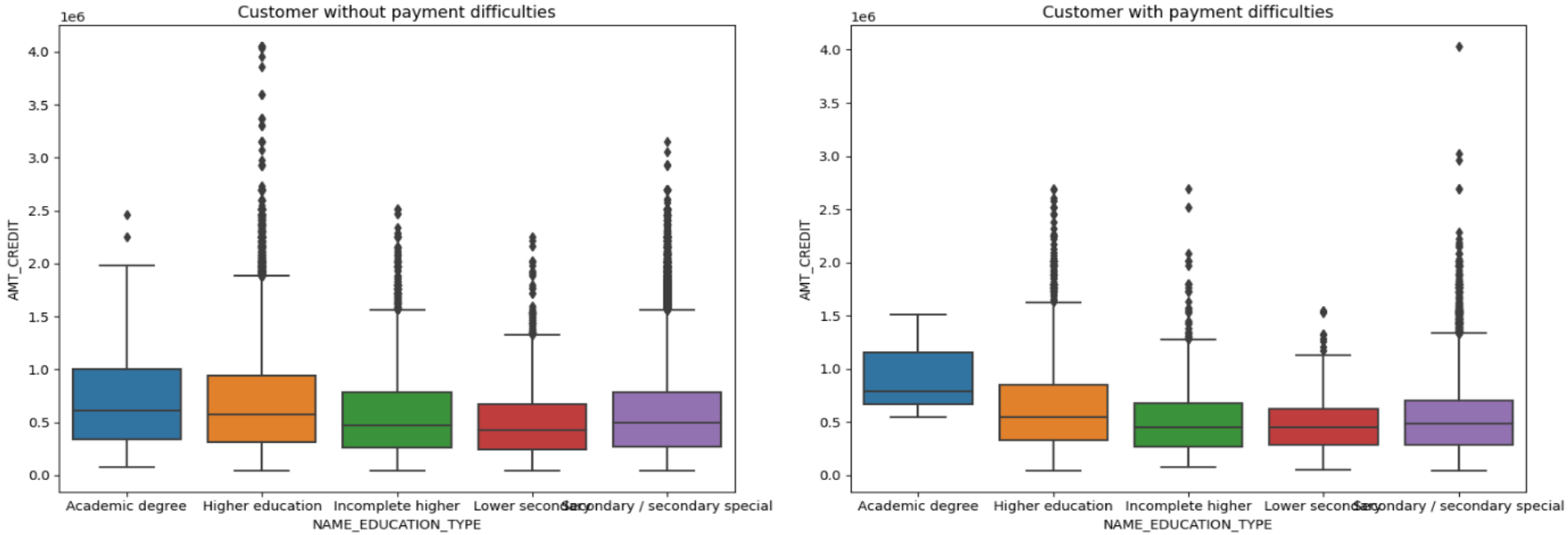
Here we can see ,most of the people taking cash loans and people without facing payment difficulties have revolving loans more than other case



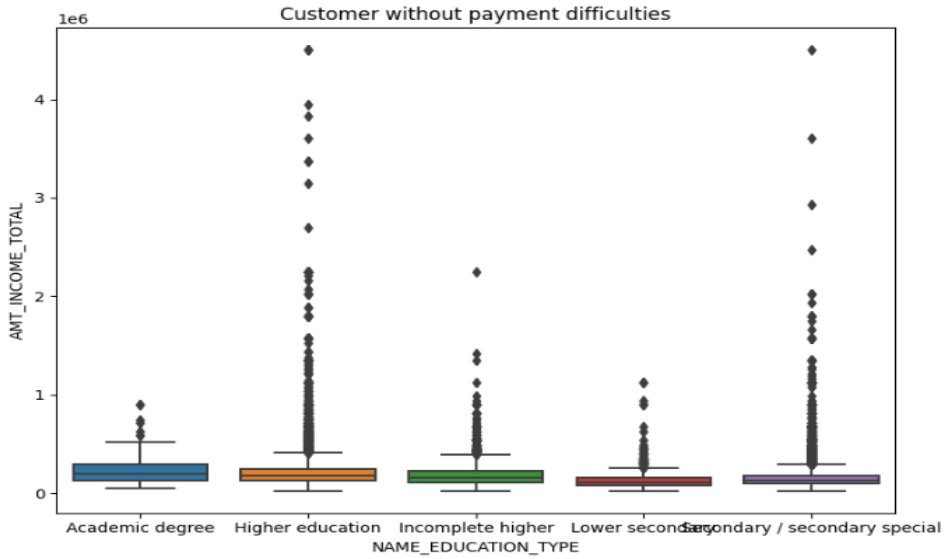


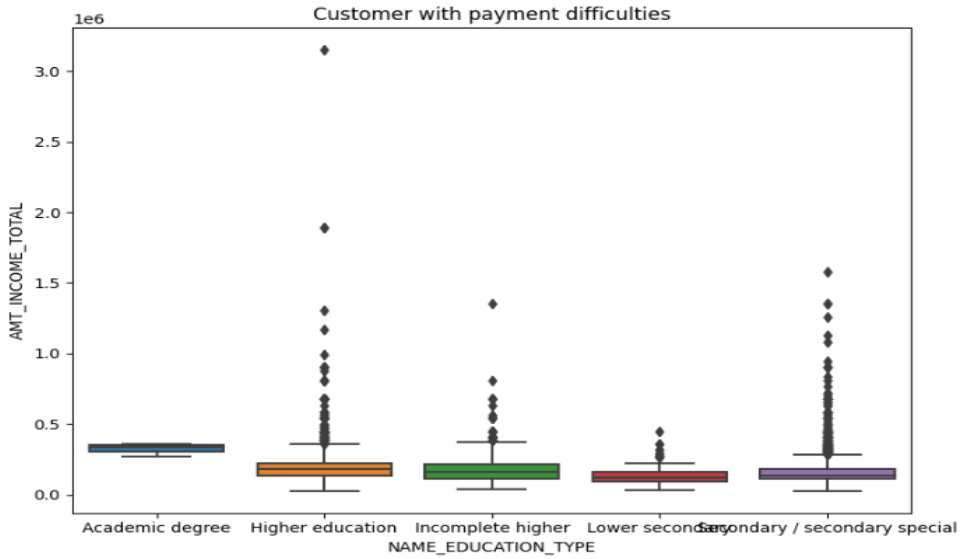
here we can see that, Females are more in number taking loans and females have more amount credit range in both cases

Numerical - Categorical bivariate analysis

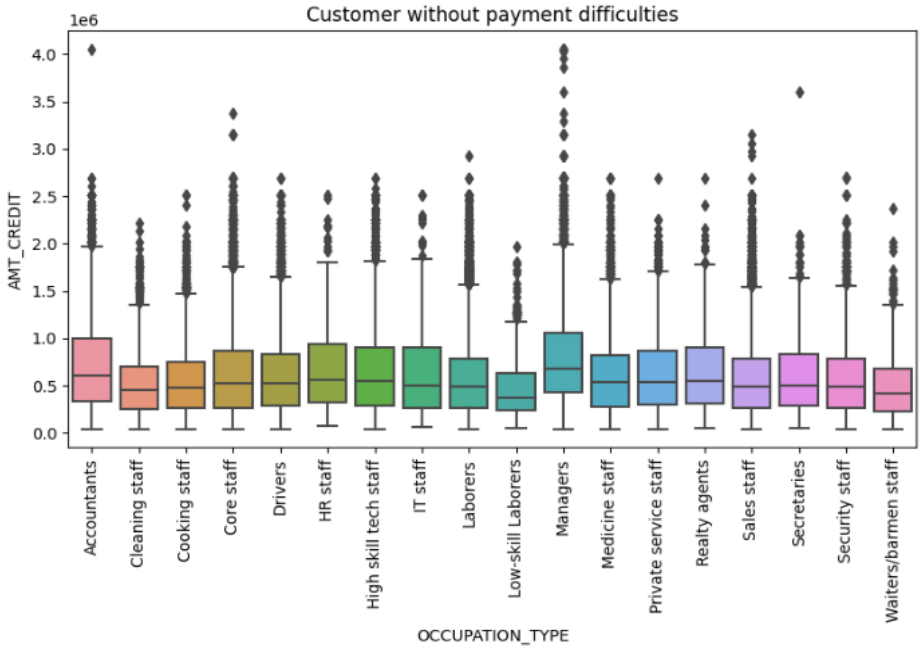


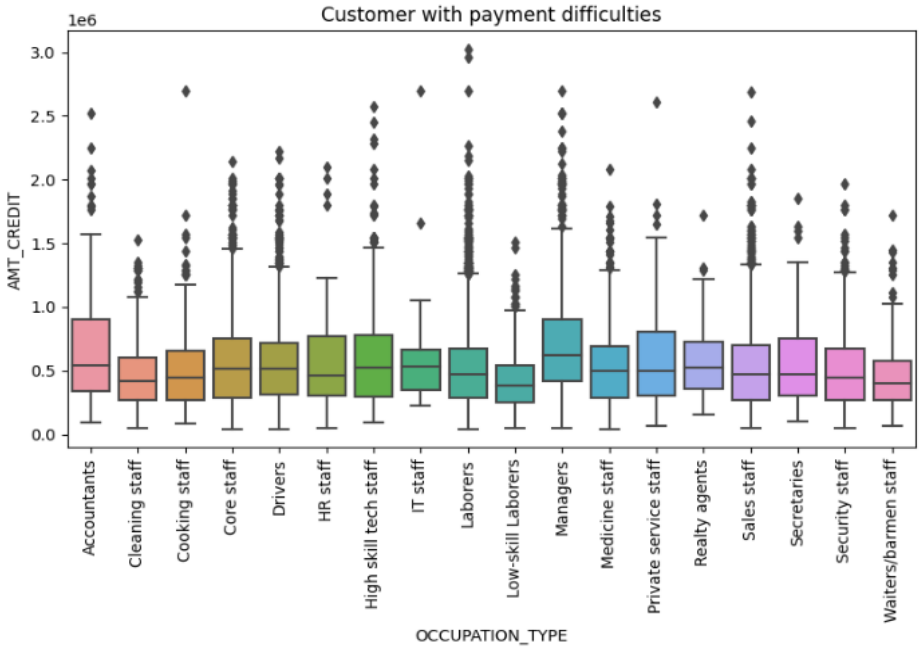
Here we can see that the range of customers with payment difficulties of academic degree is higher than without payment difficulties rest of education type is similar to each other





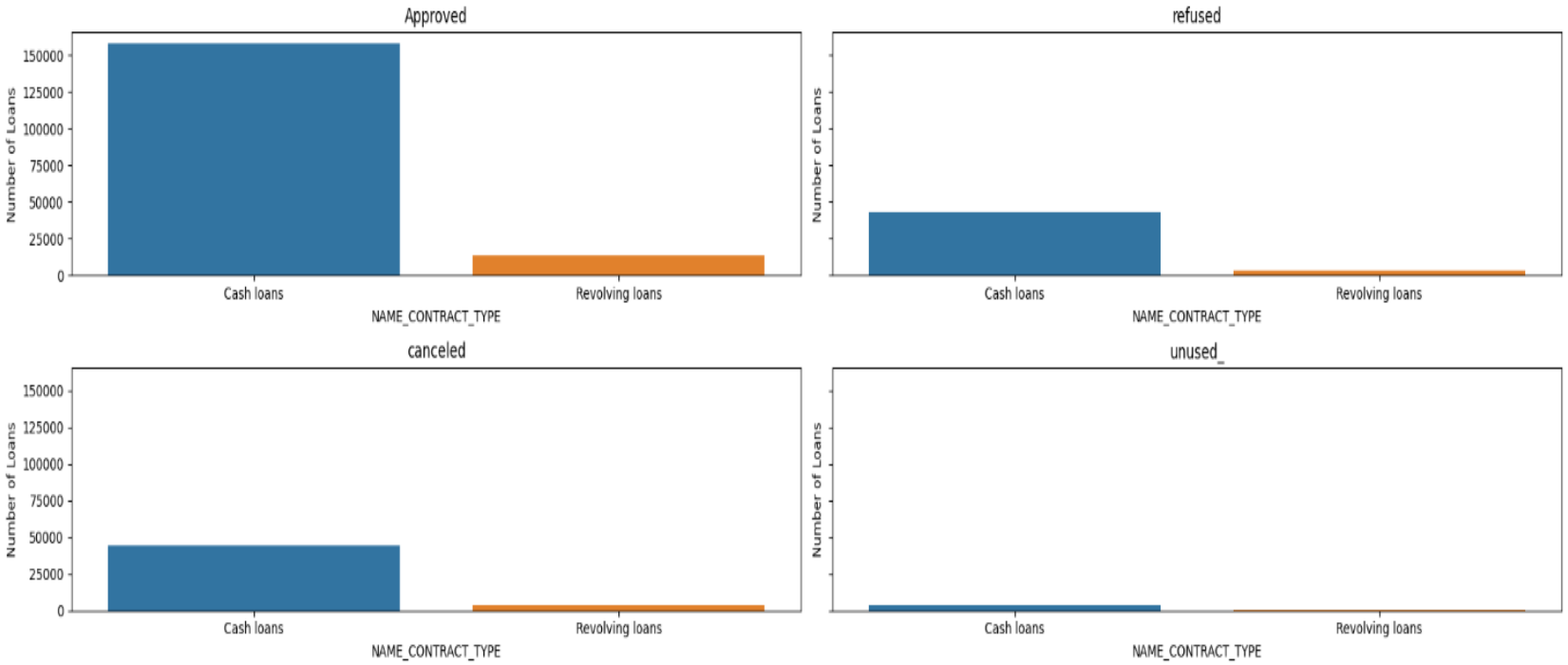
here, we can see that customer without payment difficulties are having more outliers than compared to customer with payment difficulties



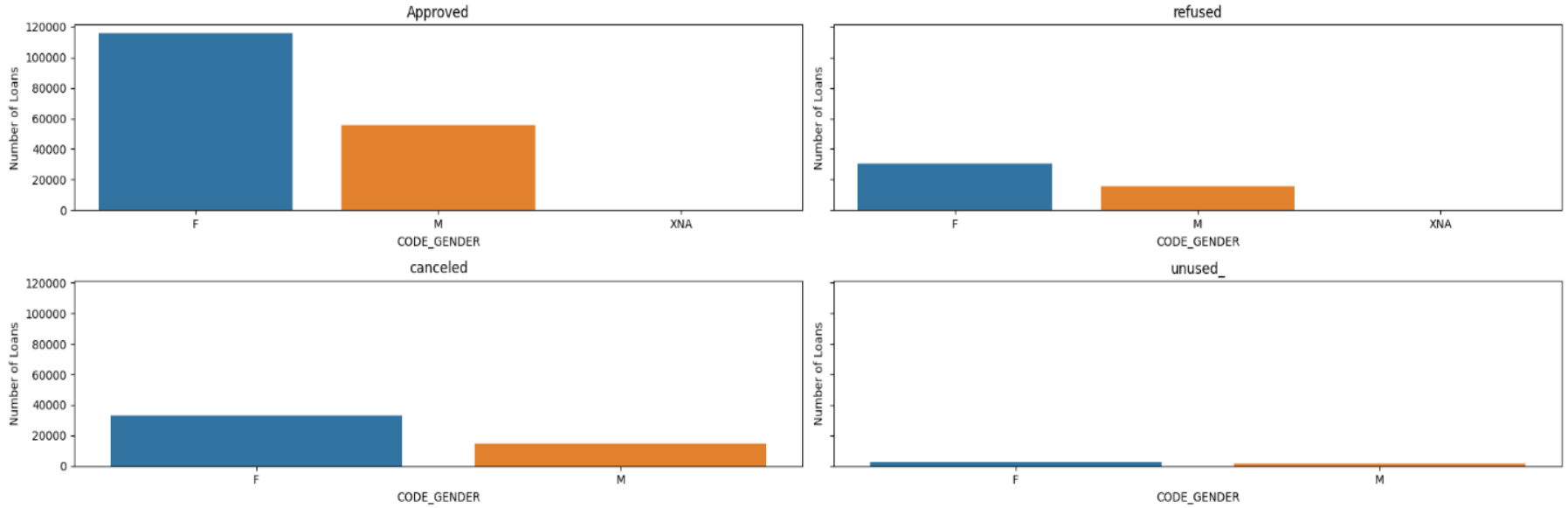


Here we can see that the range of the customers without payment more as compared to the customers with payment.

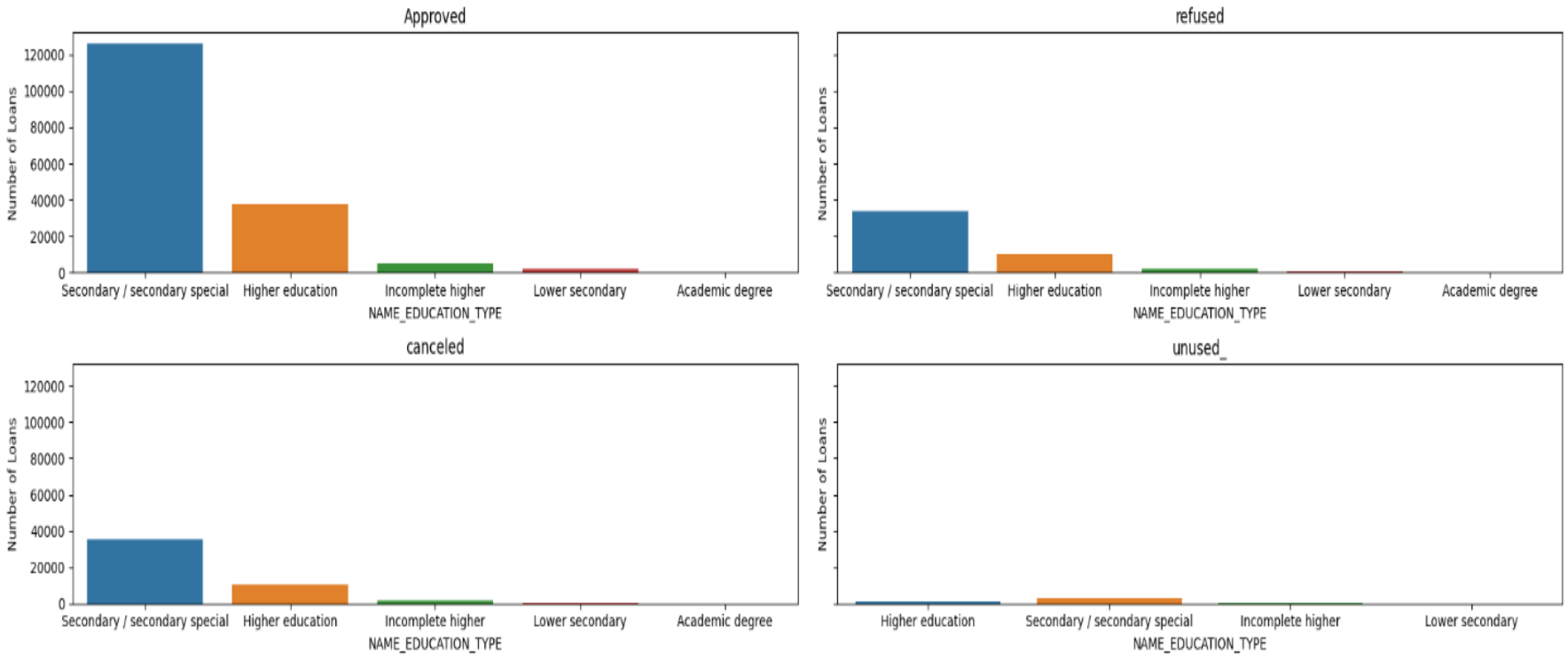
Merge Data Analysis



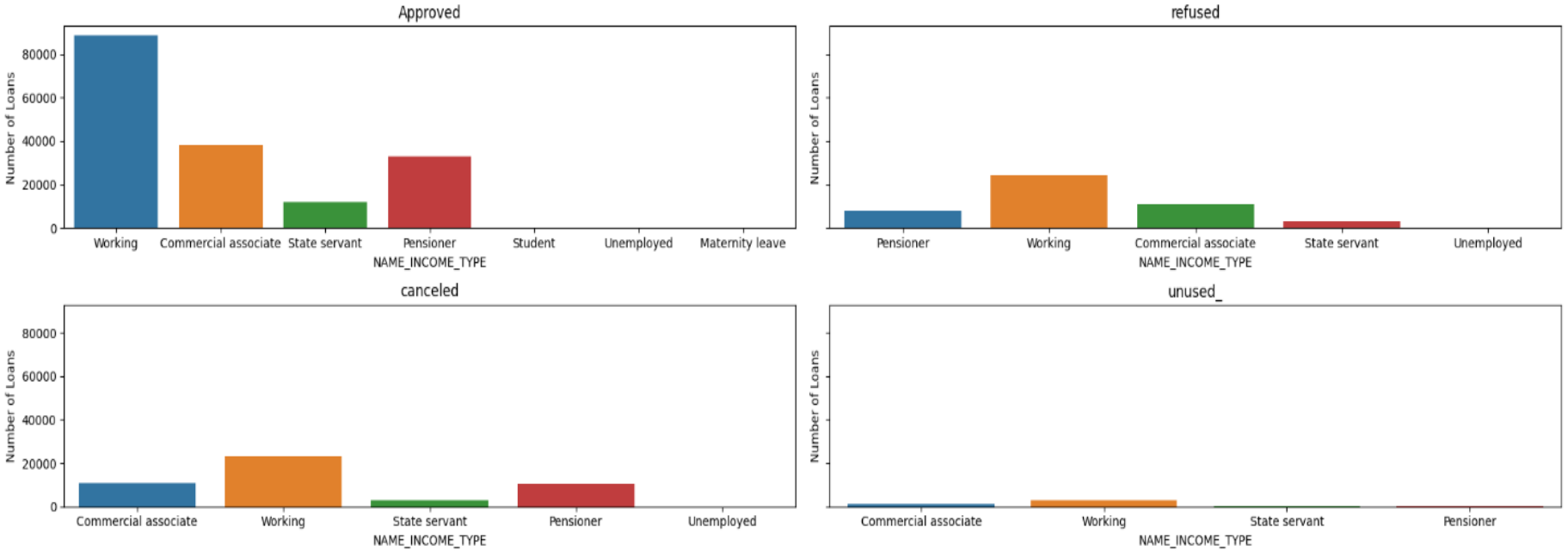
Here we can see that the Revolving loan is much more acceptable as compared to the cash and consumer loans



Here we can see that Female is getting more Refused more approved more canceled more unused, but in case of male it is having average in every category.



Here we can see that Secondary/ Secondary special is more effective in every case



Here we can see that the working type people are applying more loans as compared to others and Commercial associates' people are taking more loans.

Conclusion

- Banks should focus more on education type 'Higher education' and avoid Secondary/secondary special, incomplete higher or lower secondary as they face paying difficulties.
- Avoid income type of 'Working' clients as they have high percentage of paying difficulties. Instead focus on Commercial associate, pensioner and State servant.
- Focus on clients from housing type ‘House/apartment’ as they are having fewer paying difficulties.
- In Genders, 'Females' are more in number for applying loans.
- Banks should focus on 'Students' ,'Pensioner' for successful repayments.