Presentation on Credit EDA Case Study

Pradnya Gangurde

Problem Statement :

The loan providing companies find it hard to give loans to people due to their insufficient or non-existent credit history. Because of that, some customers use it as their advantage by becoming a defaulter. Suppose you work for a customer 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 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.

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 customers are capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

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.

Business Understanding

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, we will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

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.

Analysis of client's information at the time of application :

In order to perform data quality checks and handling missing values,
we have considered Application data set and performed necessary actions as explained below

To find out the missing values and handle it:

>We have observed there were many missing values in this data set, so using the indexing and count of missing values in each column, identified the % of missing values for each column.

>And dropped all columns from data frame for which missing values % is more than 50.We have also found few more columns which are having around 47% missing values. Since these are almost around 50% ,we have removed these columns as well.

>To extend the data enhancements and maintain a data frame with accurate values, we have identified columns with NULL values and even applied the same 47% logic and removed those columns from the data frame.

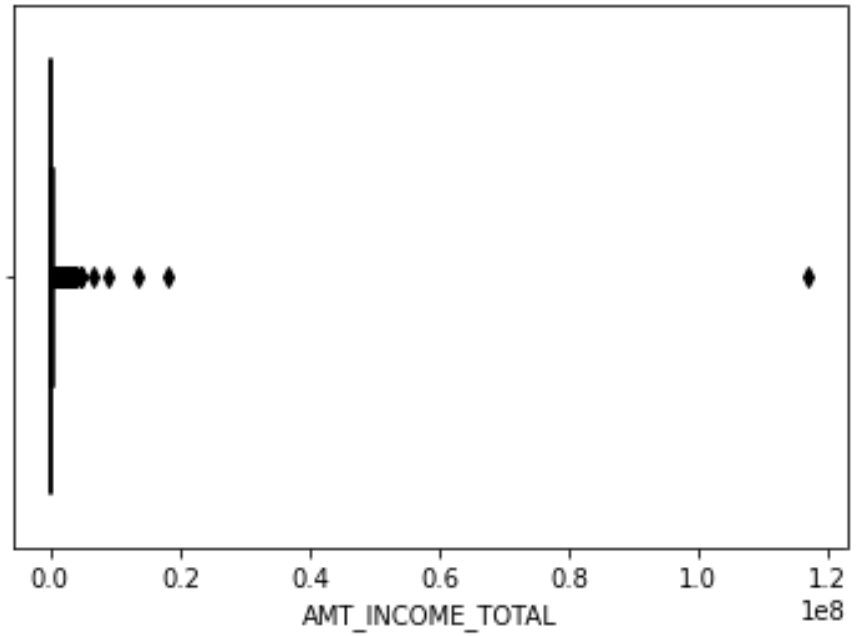
>After that there are still some columns left with some null values, we will impute the ones where null percentage is less than 1 %.

Checking the Data frame for outlier values and analyze them:

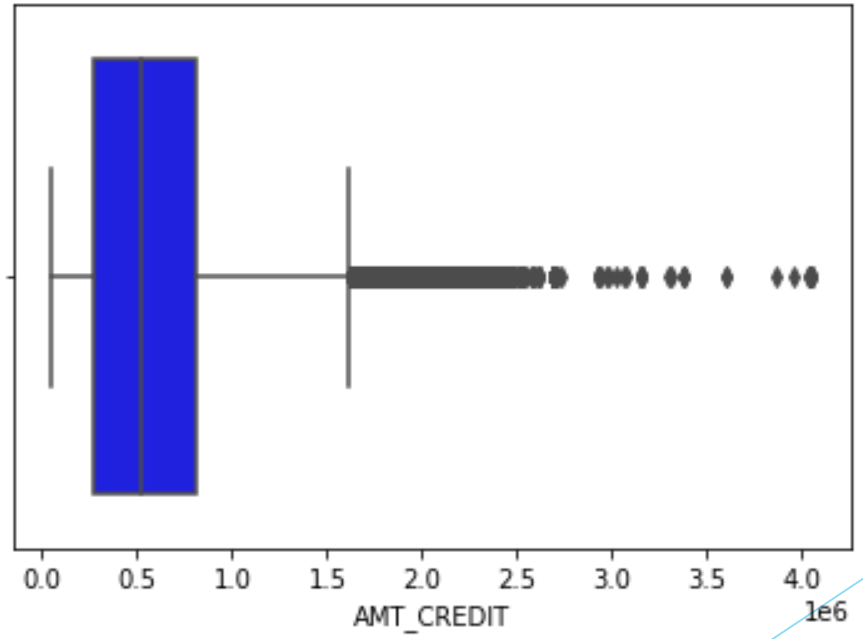
Checking the outliers for columns and understanding the reason to mention that as an outlier.

Here in our analysis to find out the outliers, we have considered few numerical columns and analyzed the statistics of them

We can see from the plot there is one value which is too high compared to others, hence it is an outlier

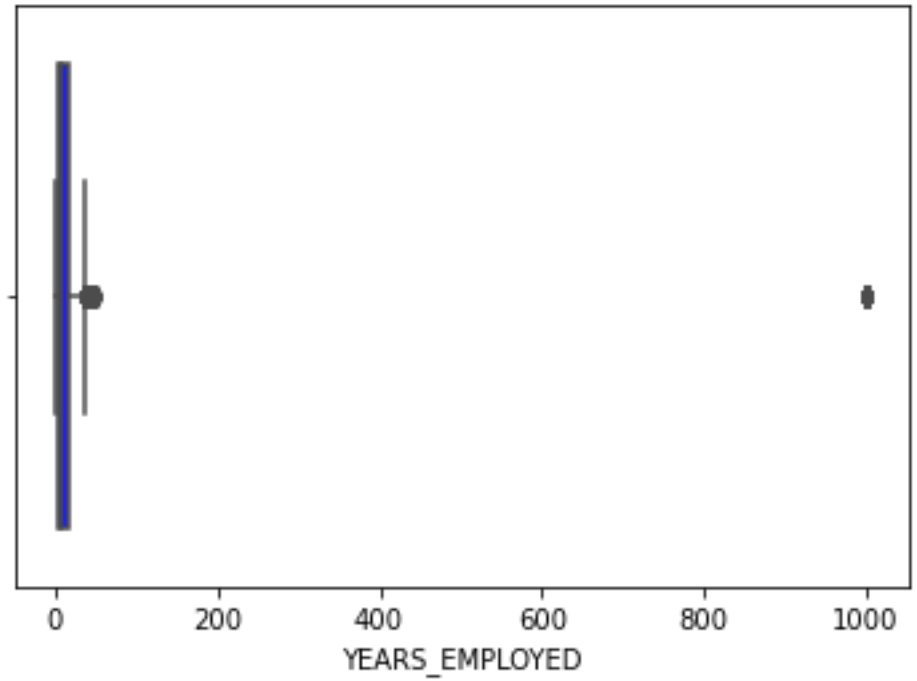


The AMT_CREDIT is greater than AMT_INCOME_TOTAL in all the cases



Outlier Analysis for YEARS_EMPLOYED

We can clearly see from the plot , the outlier value is 1000 yrs.

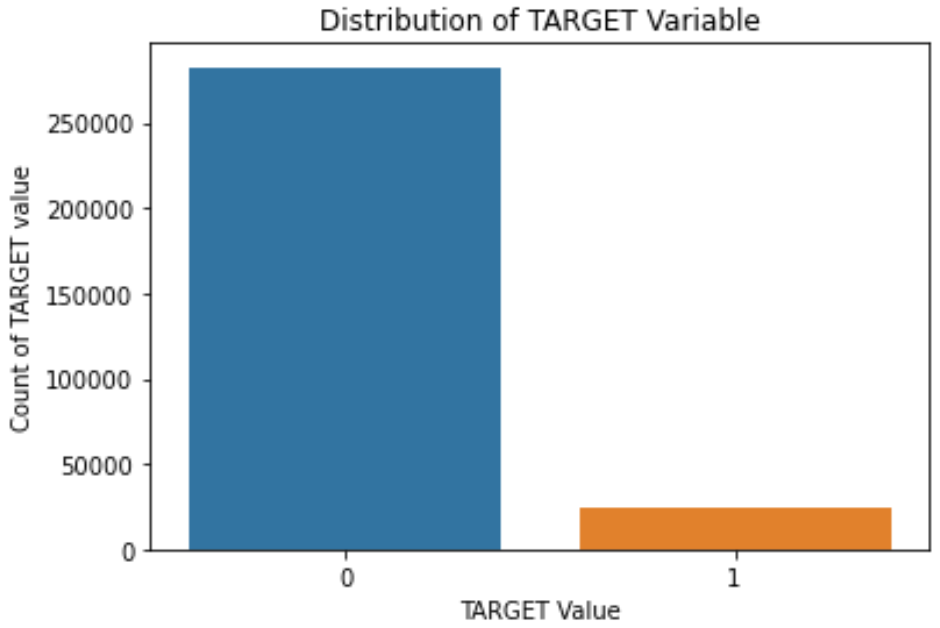


Data Analysis

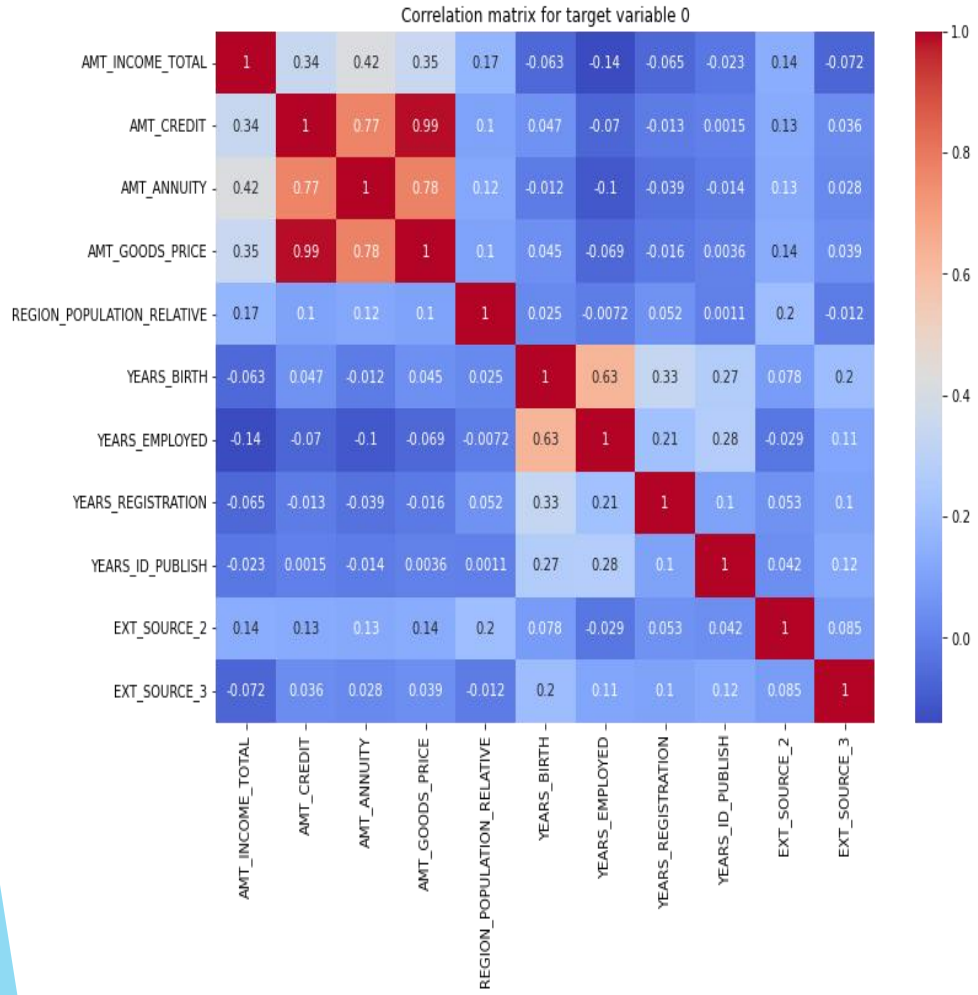
While working with the data frame based on Target variable,

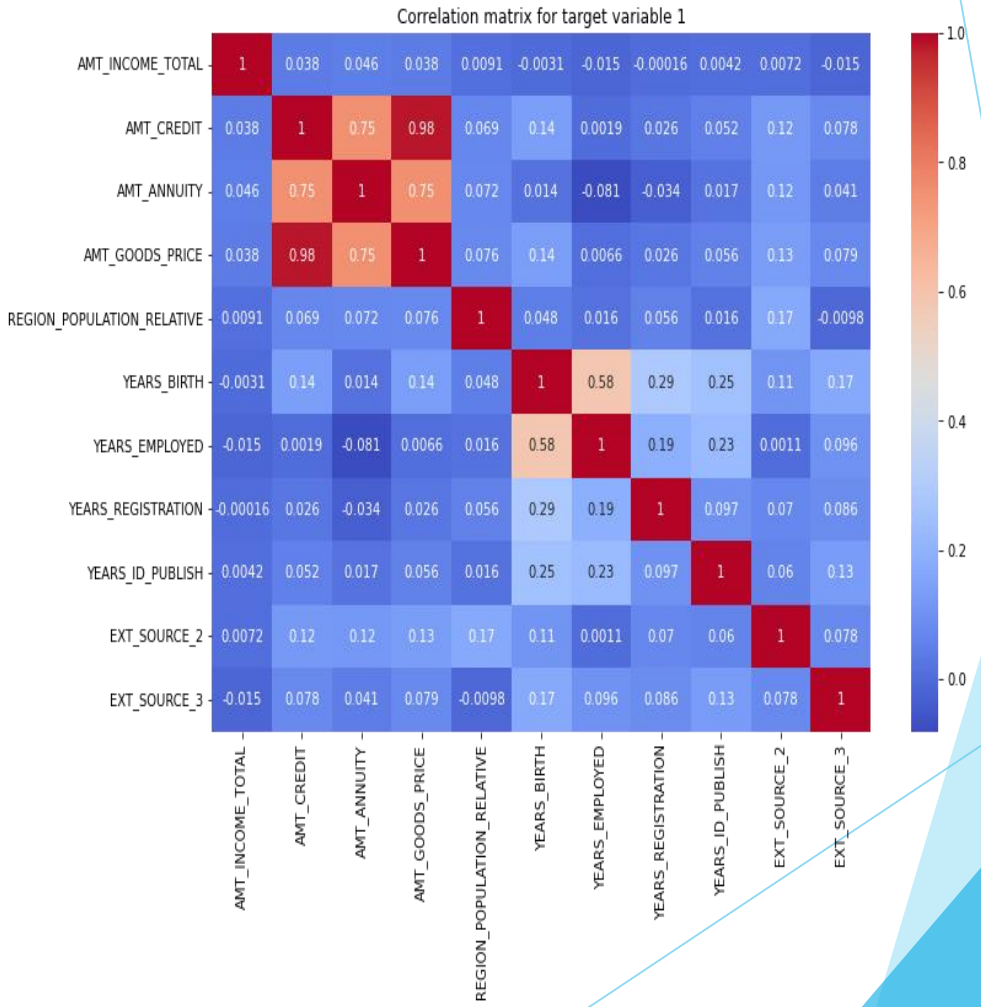
As shown in the image, we can clearly see the imbalance between target type 1 and 0

Ratio is of 91.5 : 8.45



We will now find the correlation between different variables for both dataframes with target=1 and target=0

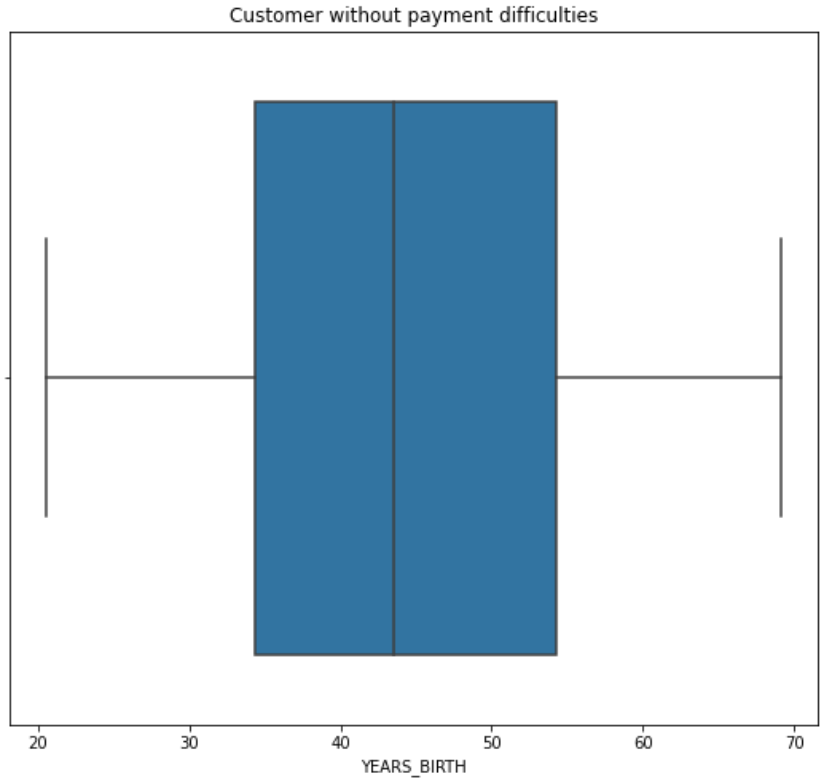


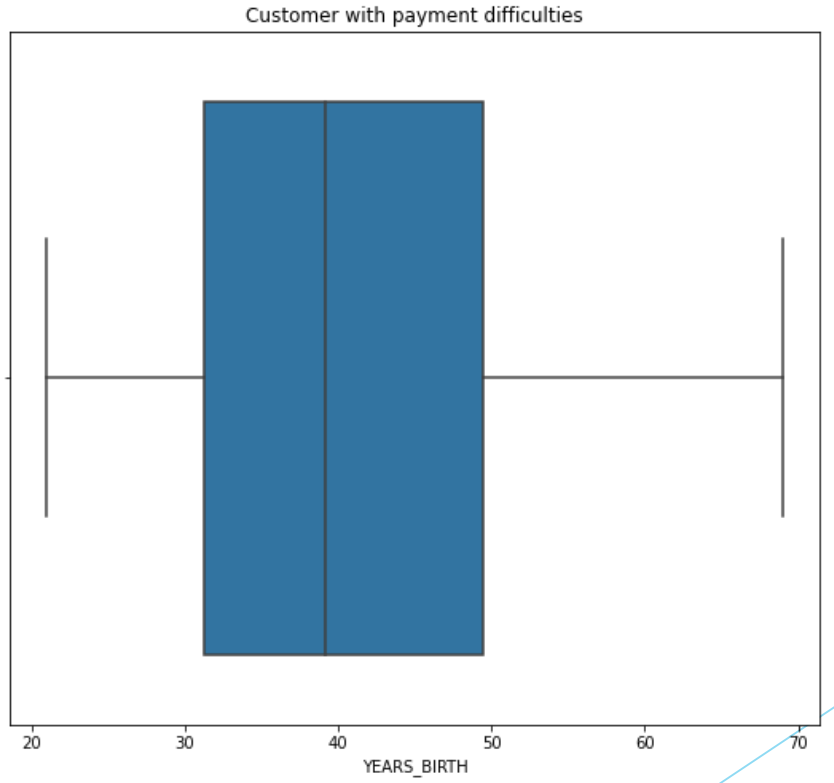


Univariate Analysis

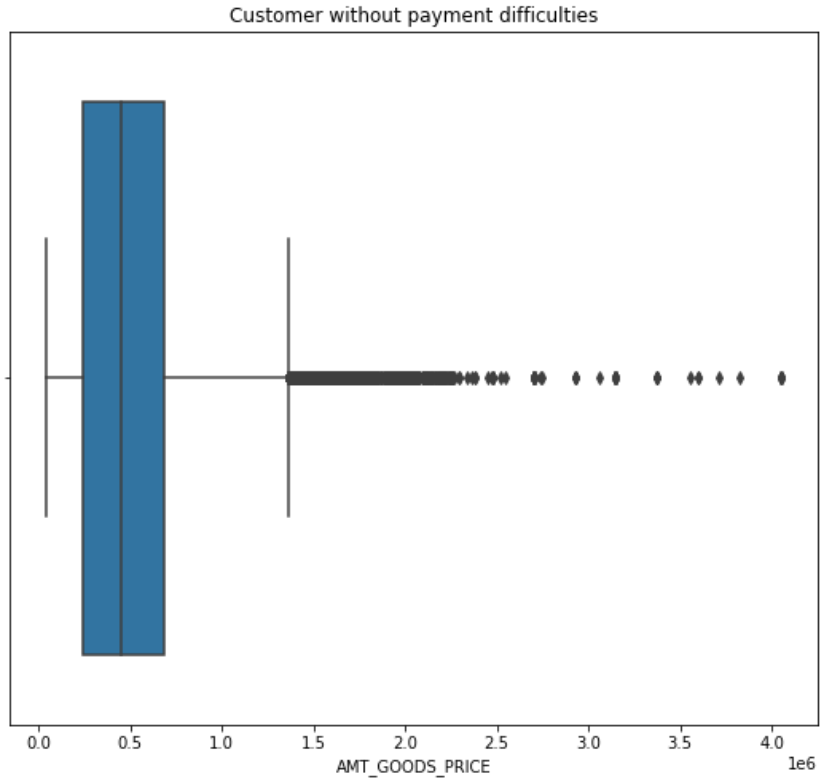
(A) For Numerical variables:

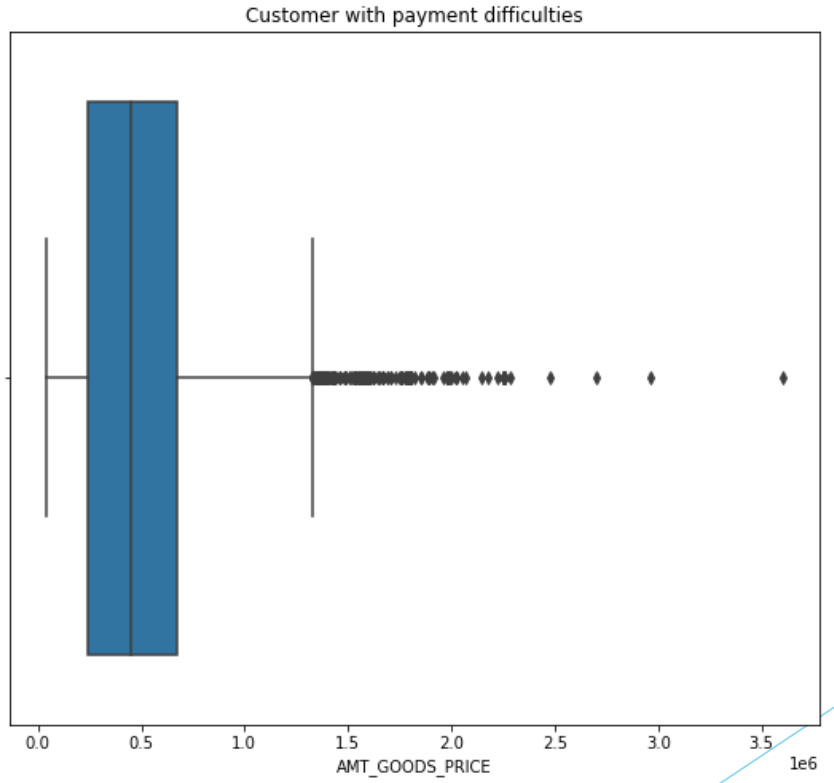
From the above box plot we can see that customer without payment difficulties is between 34 to 54 years , And customer with payment difficulties is between 31 to 50 years.





Here we can see that the customer without payment difficulties lies in between 0.3 to 0.7 and the customer with payment difficulties also lies in between 0.3 to 0.7. And also both are having the mid value about 0.5

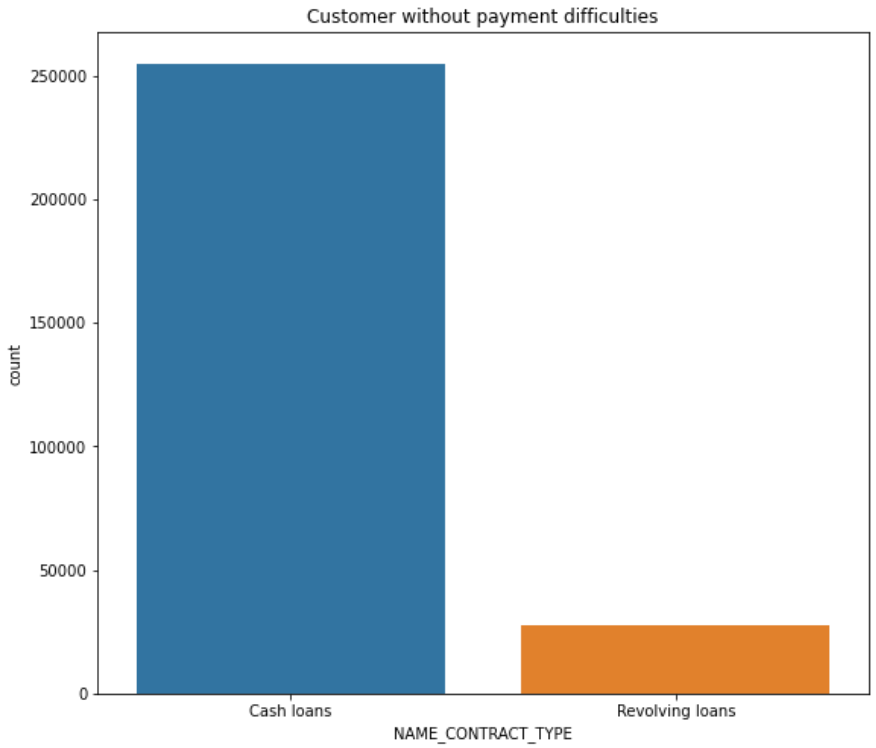


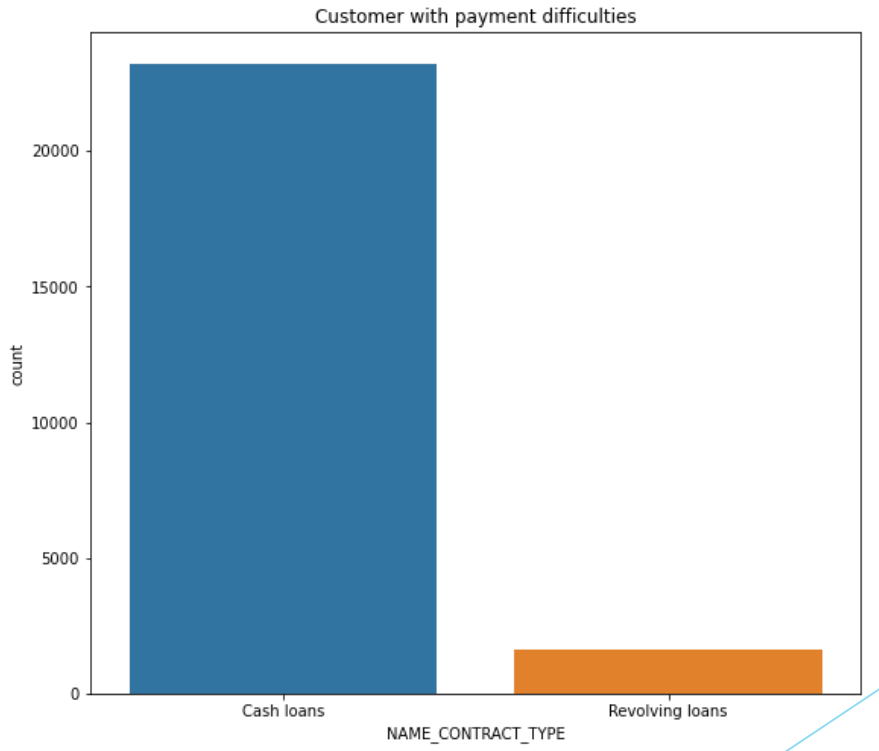


Univariate Analysis

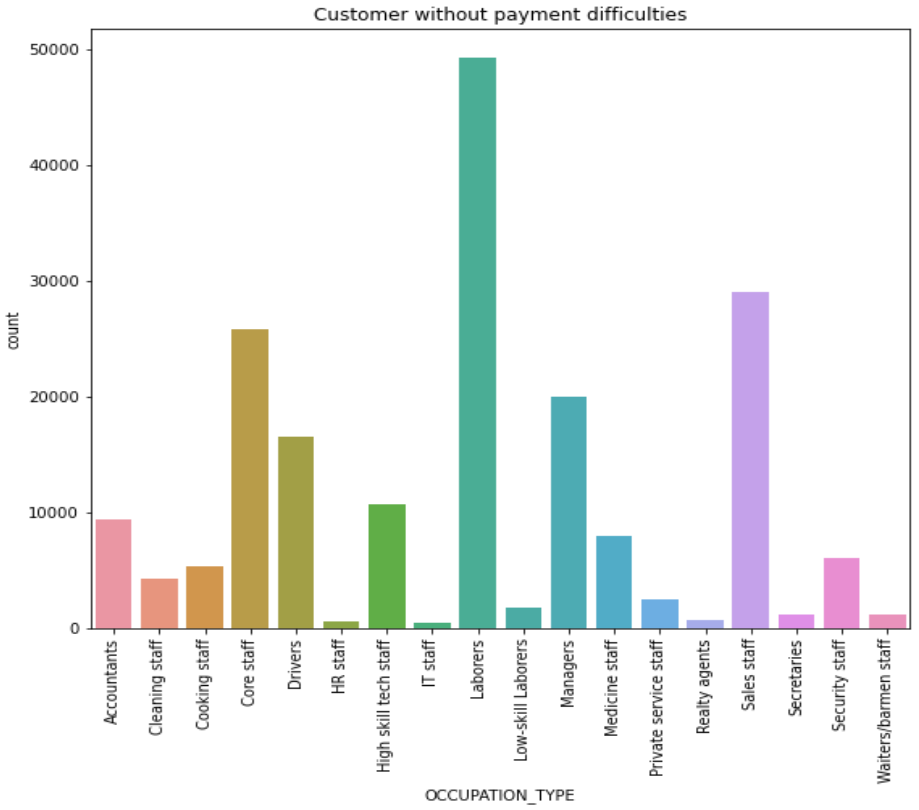
(B) For Categorical Variables

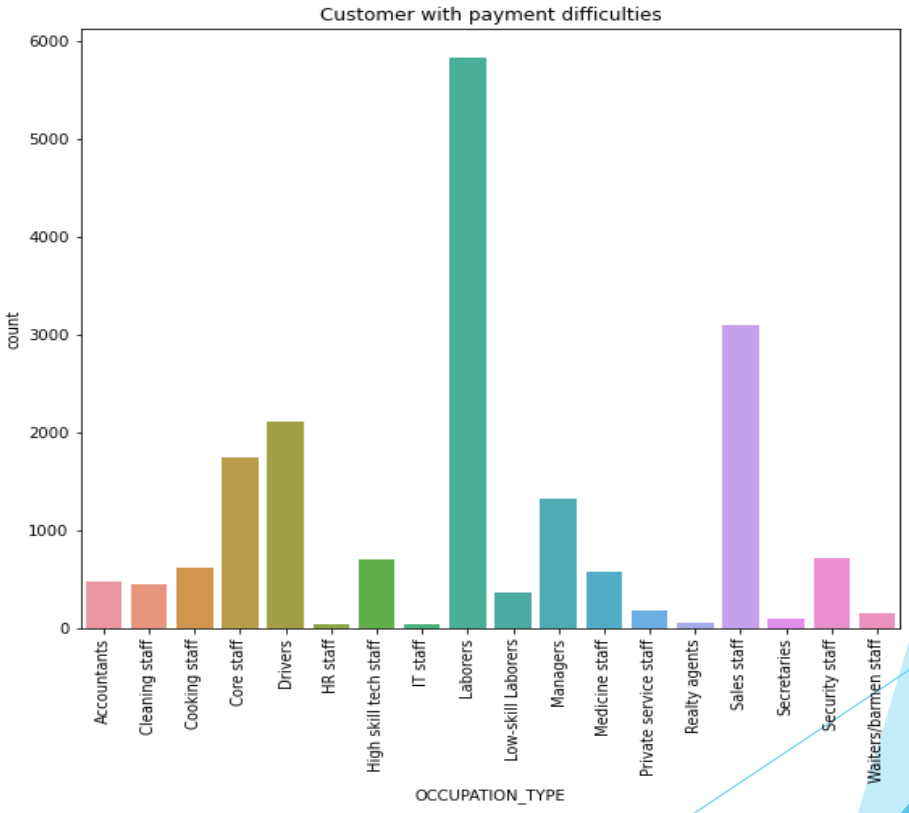
Here we can see that the customer without payment and customer with payment difficulties both are taking cash loans.





Here we can see that laborers are having more difficulties in repaying the loan and also the core staff and the sales staff. But in the case of laborers those who have without payment are way more than with having the payment.

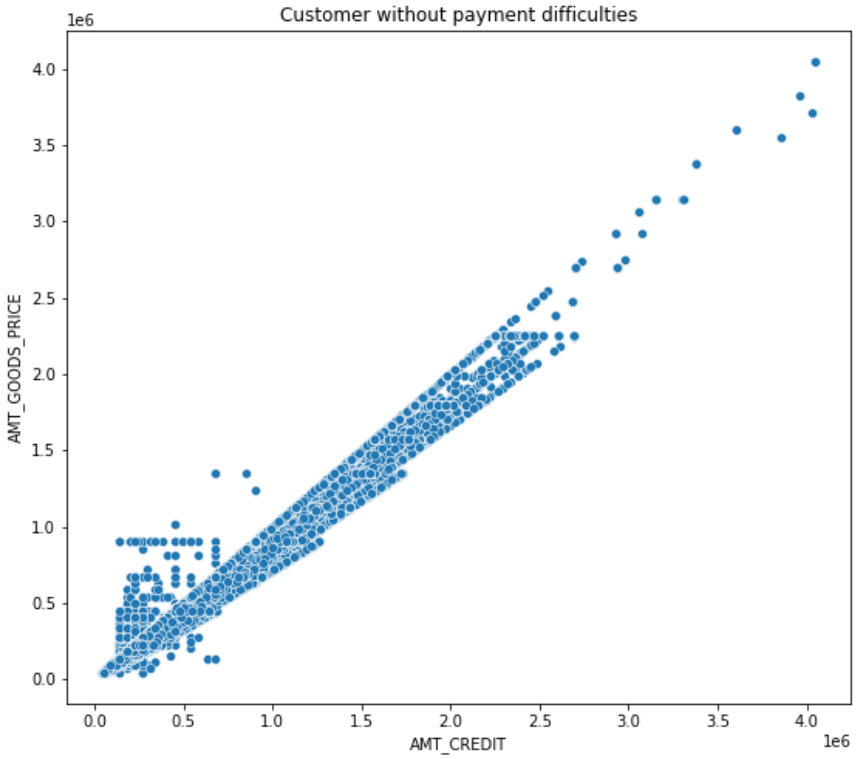


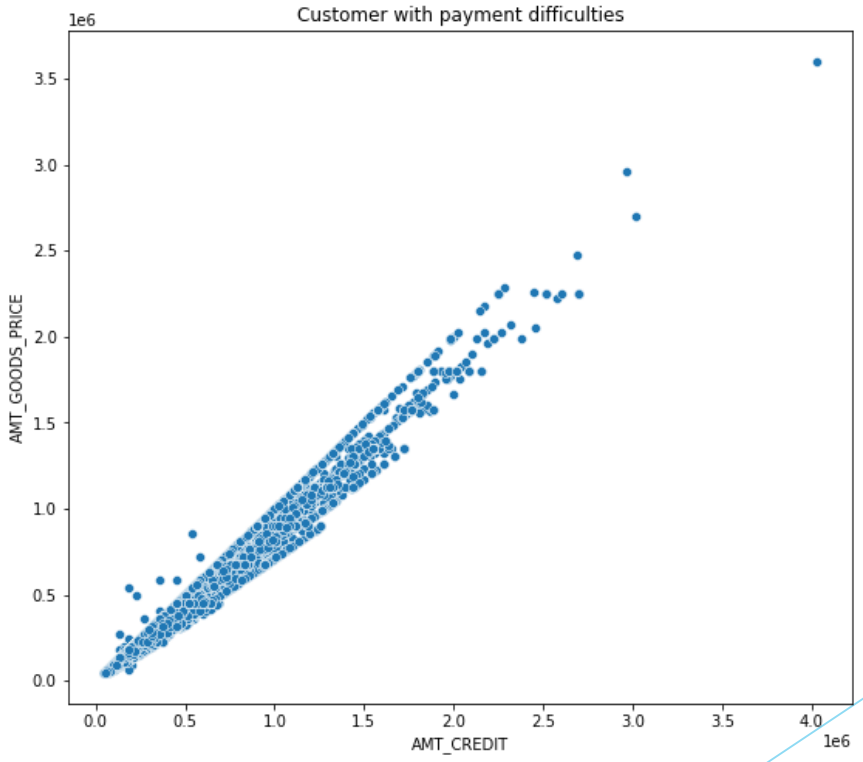


Bivariate Analysis

A)Numerical-Numerical bivariate analysis

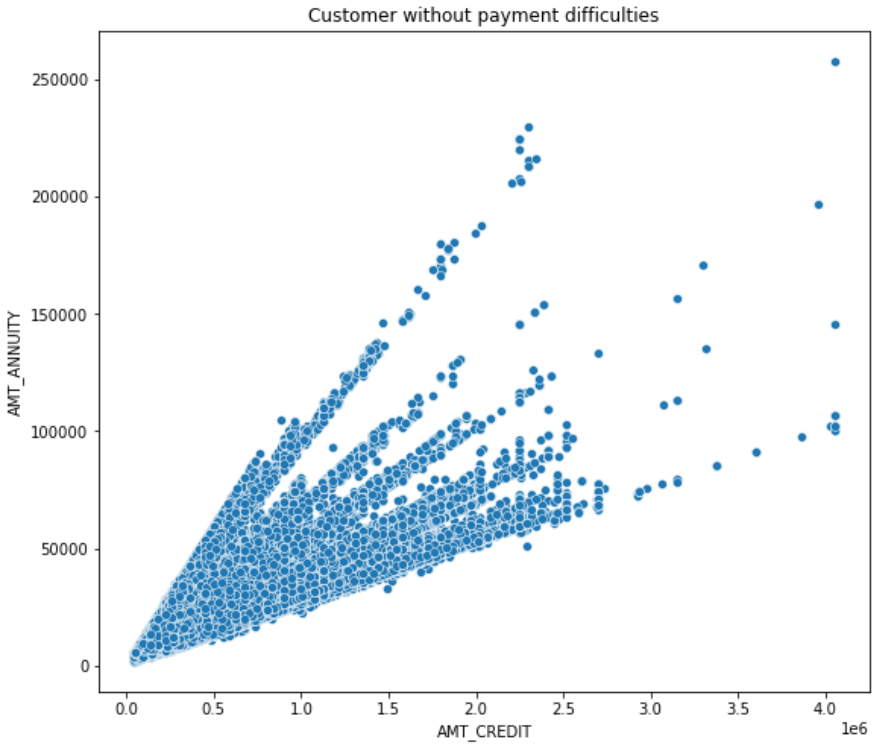
We can see below that goods price is positively correlated with credit amount.

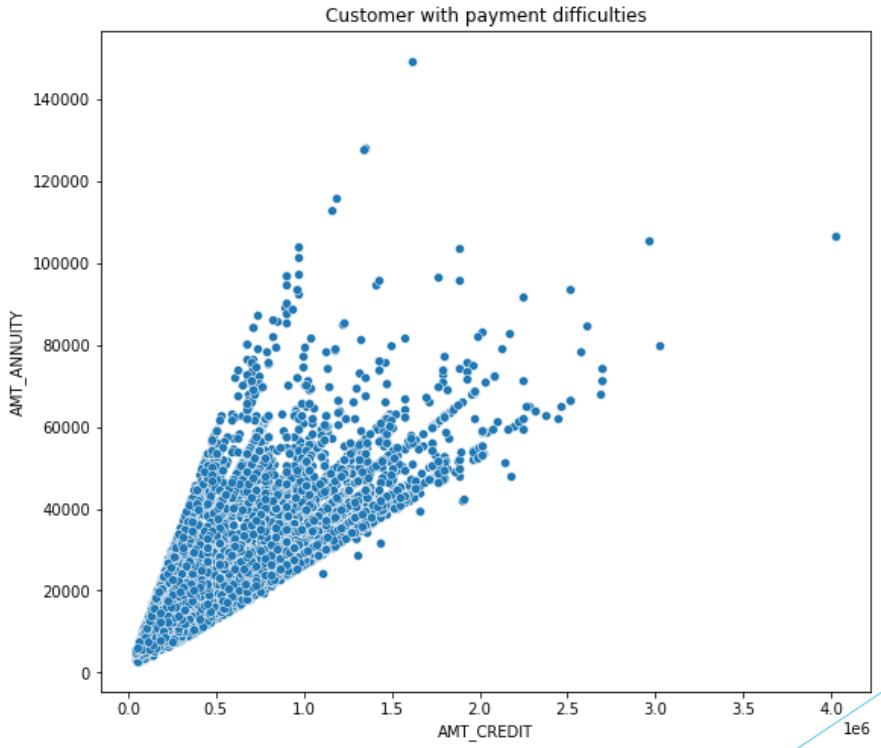




Bivariate Analysis

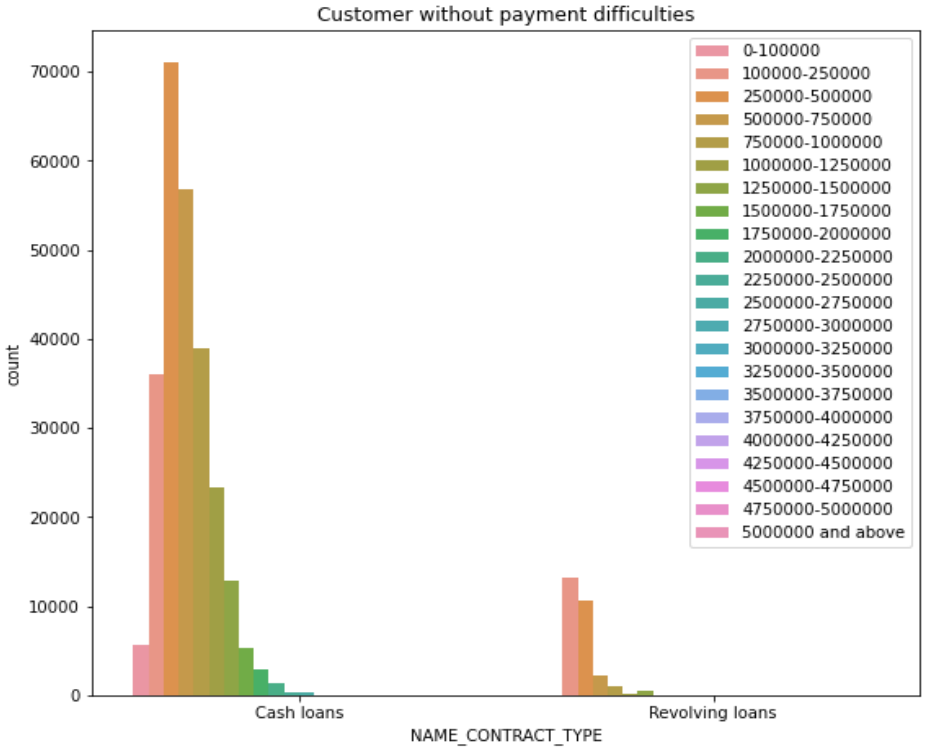
People without payment difficulties take more credit for the annuity

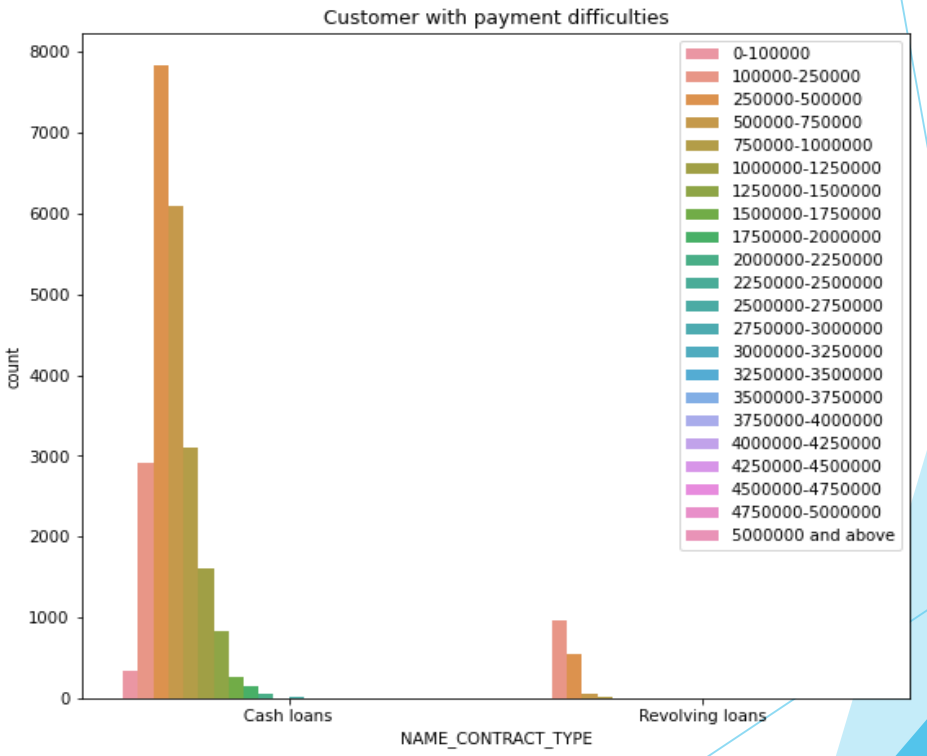




Bivariate Analysis

B) Categorical - Categorical bivariate analysis

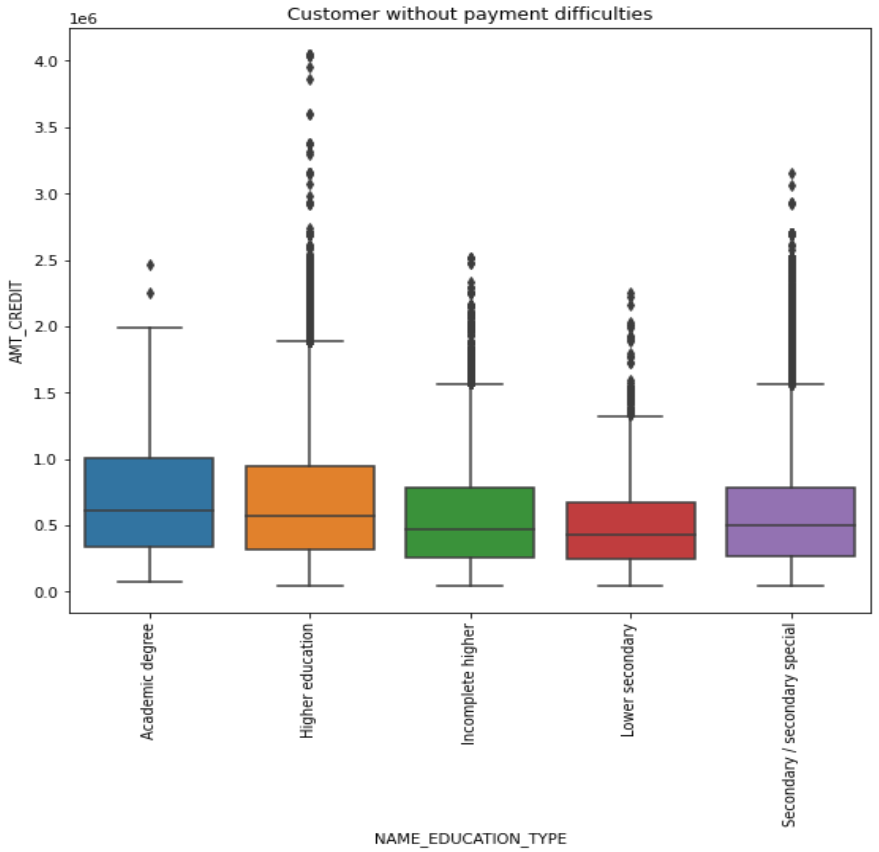


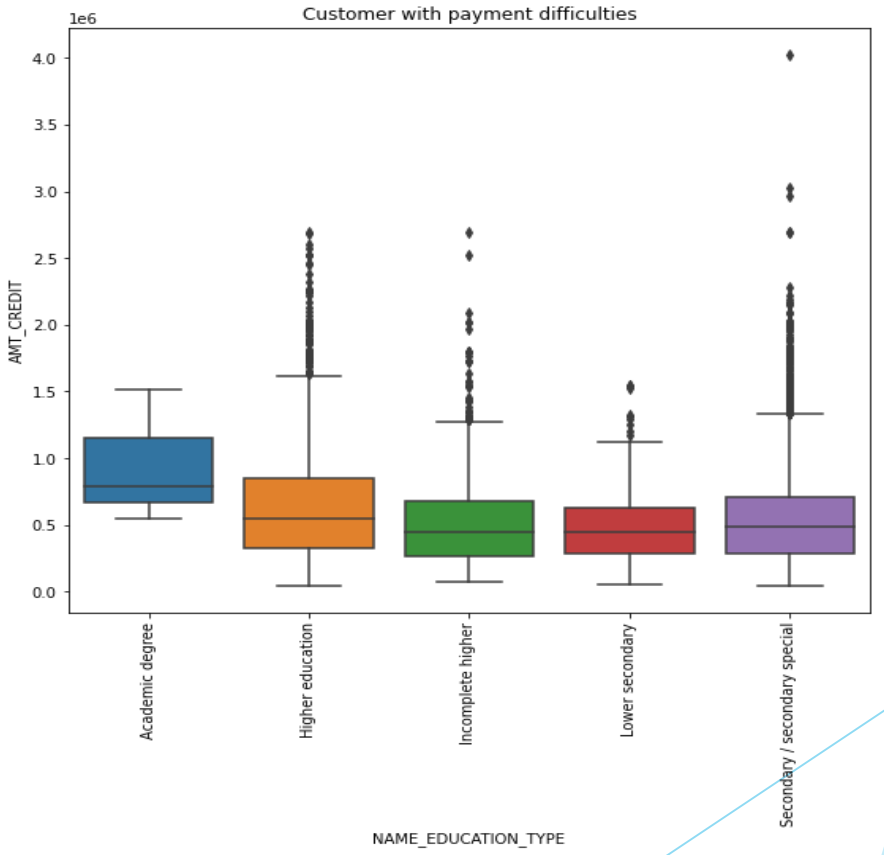


Bivariate Analysis

C) Numerical - Categorical bivariate analysis:

Here we can see that the range of customers without payment from Academic degree is higher than the customer of with payment. And the rest of the Education type is almost same for both the cases.





Merging the application data with previous application data

Took sample out of the output data and Started visualization.

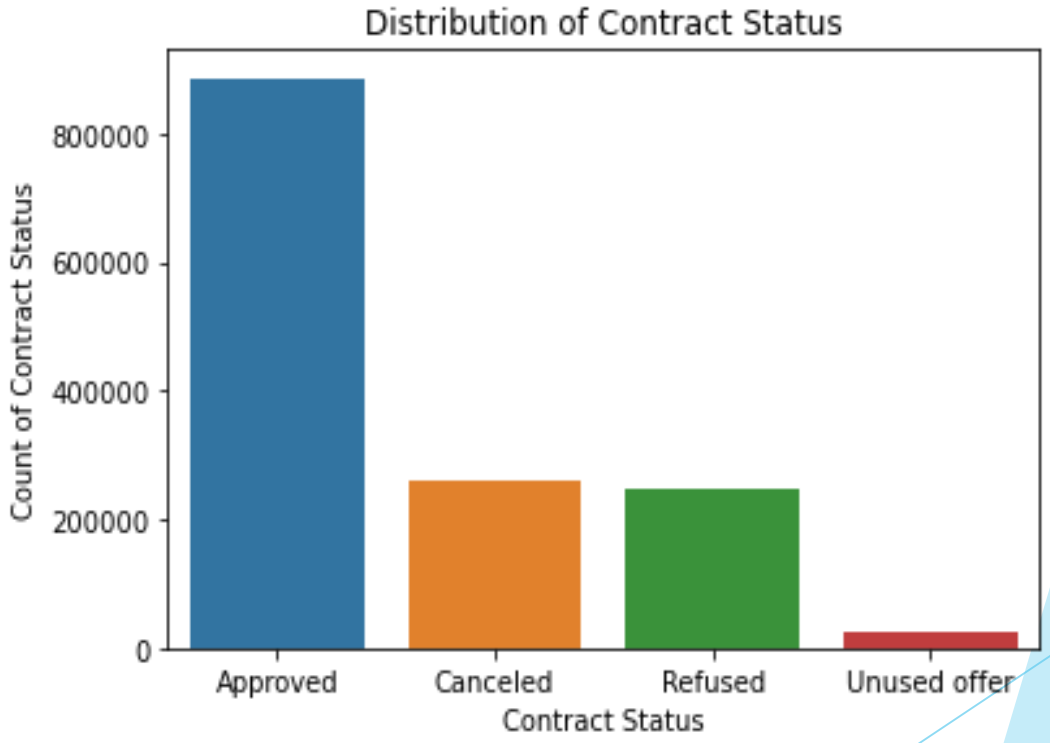
Contract was approved or not in previous application:

Approved: 62.1 % times

Cancelled: 18.9 % times

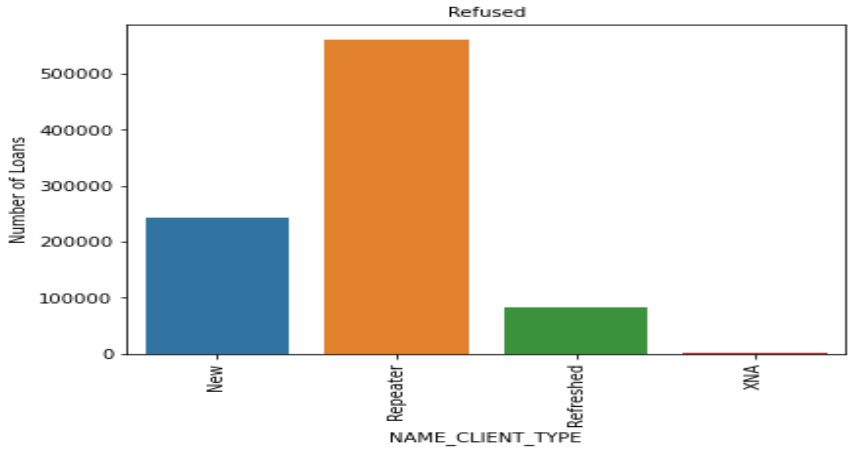
Refused: 17.4 % times

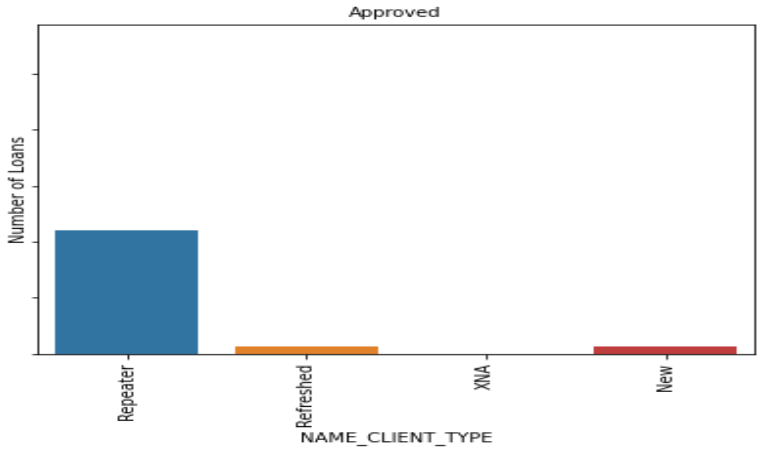
Unused offer: 1.58 % times

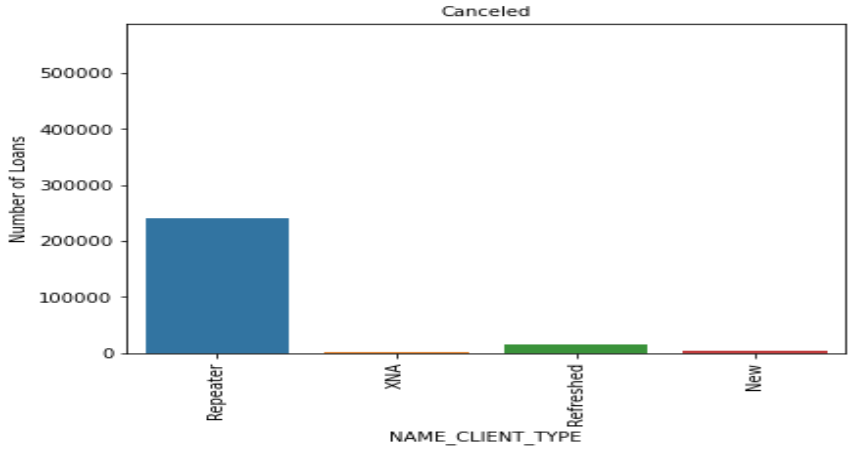


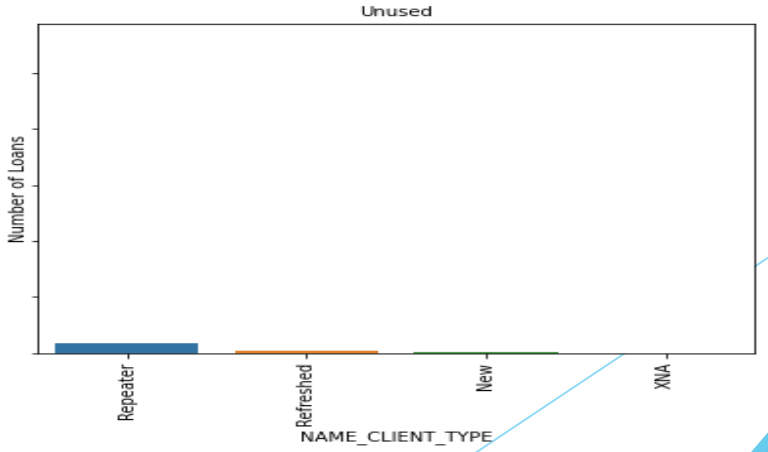
Visualization using multiplot

Here we can see that the Repeater is getting more Refused but also we can see that the it also getting more approved and even that it is getting more canceled and more unused.









Conclusion

*Banks should focus more on contract type ‘Student’ ,’pensioner’ and ‘Businessman’ with housing ‘type other than ‘Co-op apartment’ for successful payments.
*Banks should focus less on income type ‘Working’ as they are having most number of unsuccessful payments.
*Also with loan purpose ‘Repair’ is having higher number of unsuccessful payments on time. Get as much as clients from housing type ‘With parents’ as they are having least number of unsuccessful payments.

Client categories to be targeted for providing loan :

- Clients who are employed for more than 19 years
- Clients in the age range 30-40 and 40-50
- Clients who are Married
- Male clients with Academic degree
- Students and Businessman
- Repeater clients