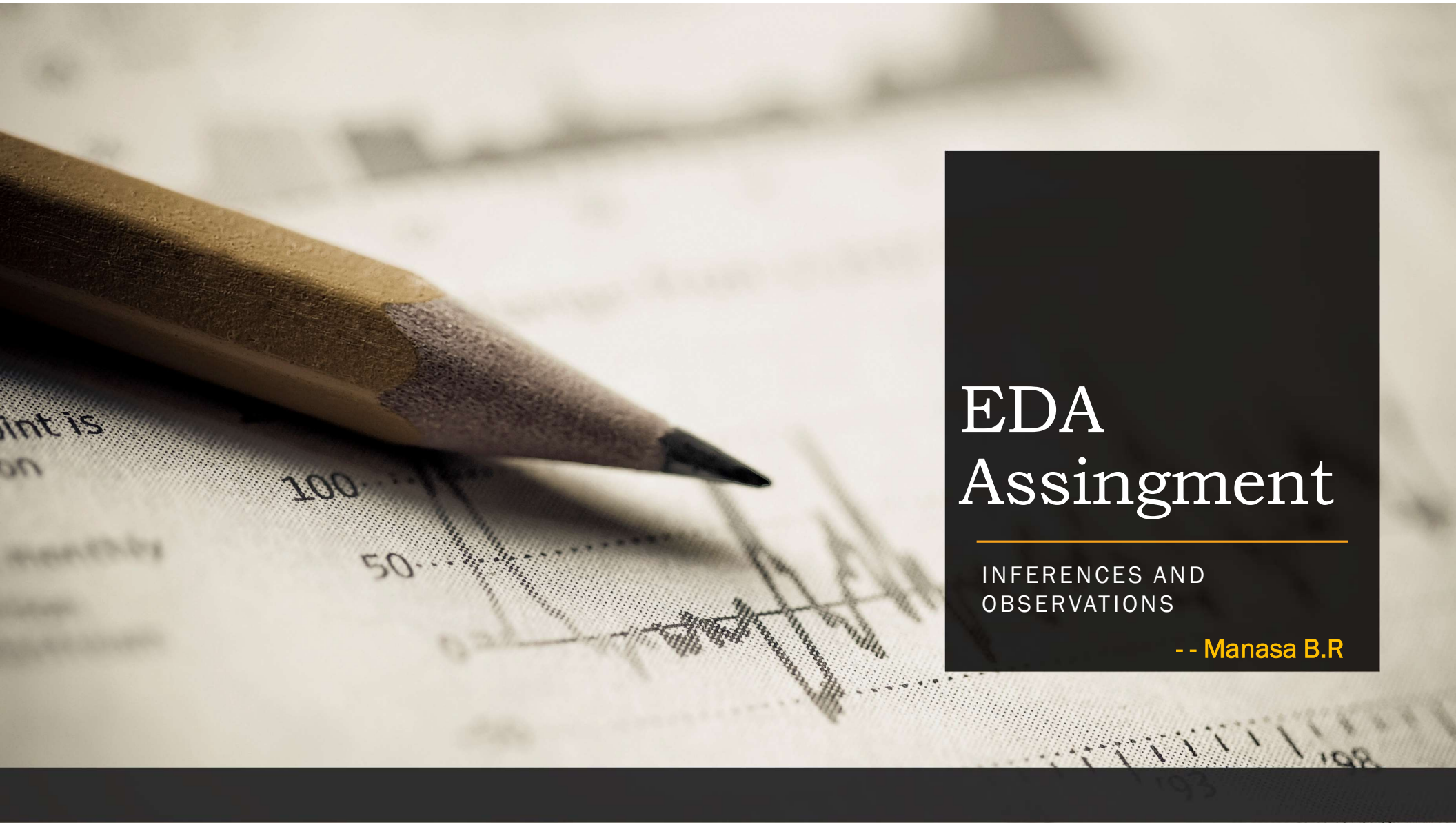


EDA Assingment

INFERENCES AND OBSERVATIONS

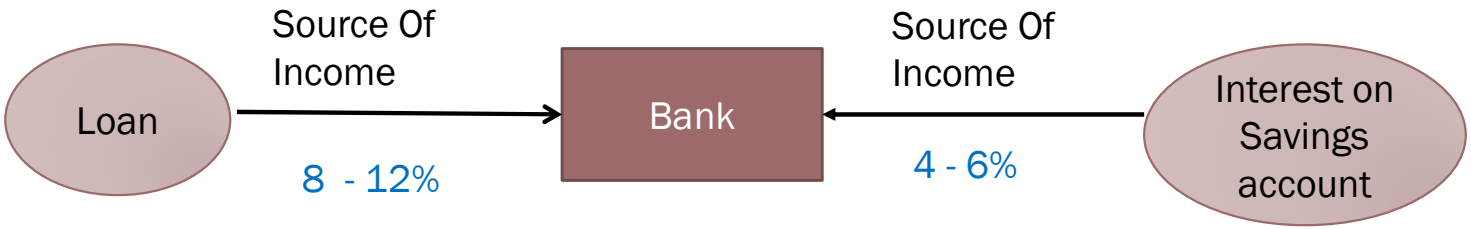
- - Manasa B.R

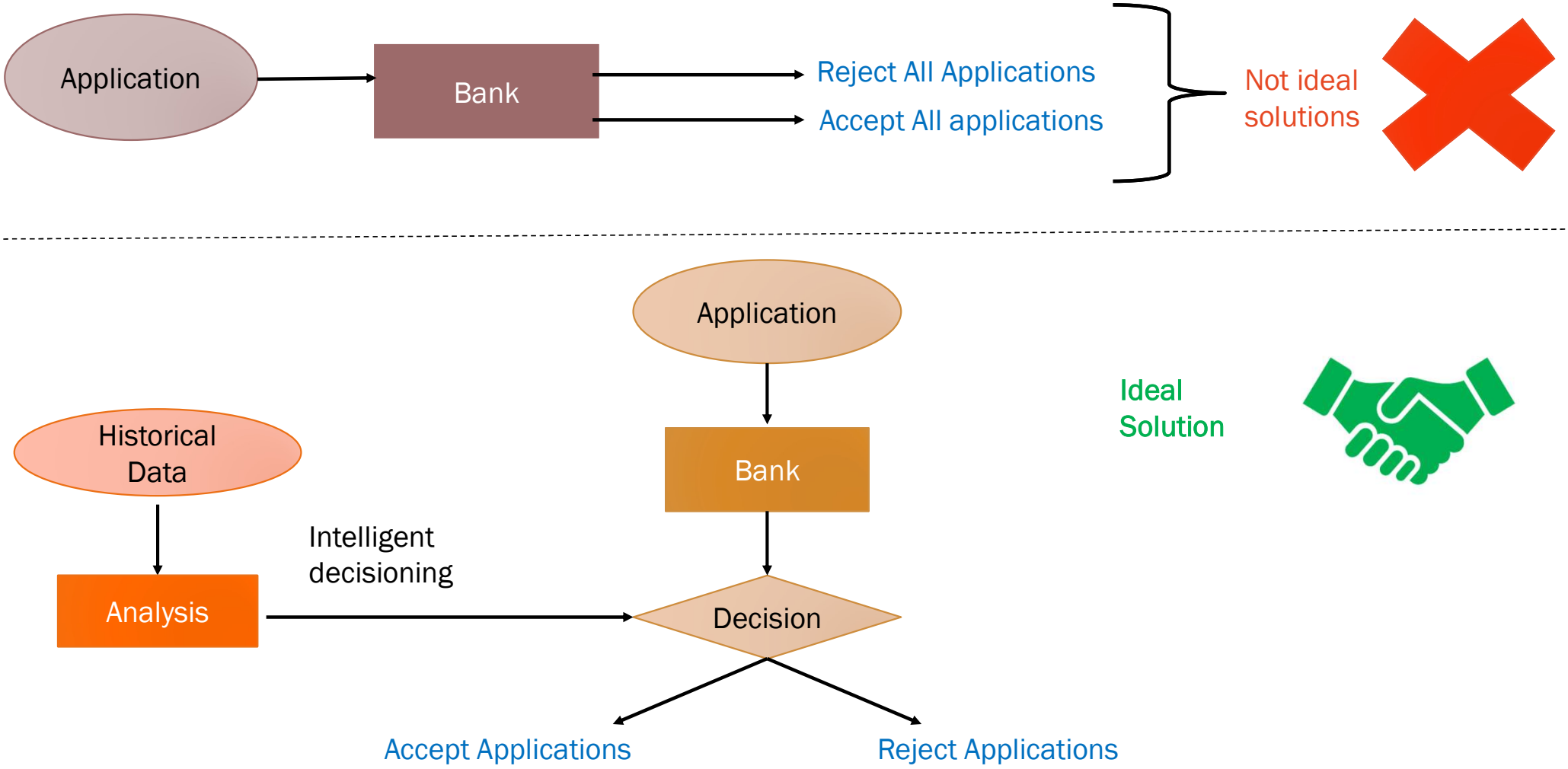
Agenda

- Problem and business understanding
- Dataset
- Aim
- Inferences after EDA
- Conclusion

Problem and Business Understanding

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter.
- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile.




Application
Bank
Reject All Applications
Accept All applications
Not ideal solutions
Application
Historical Data
Bank
Intelligent decisioning
Analysis
Decision
Accept Applications
Reject Applications
Ideal Solution

Risks involved

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.

Dataset

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 on 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.

Aim:

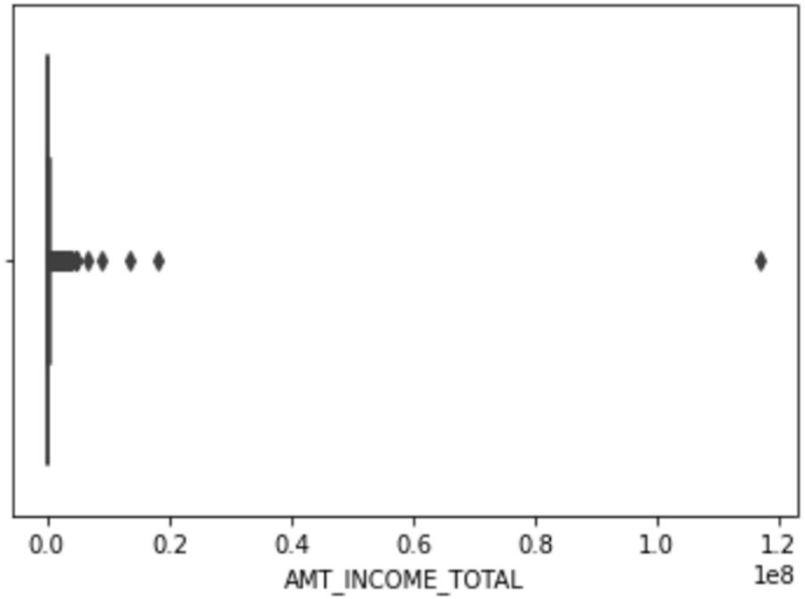
Using EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

Inferences:

Dataset 1→ Application Data

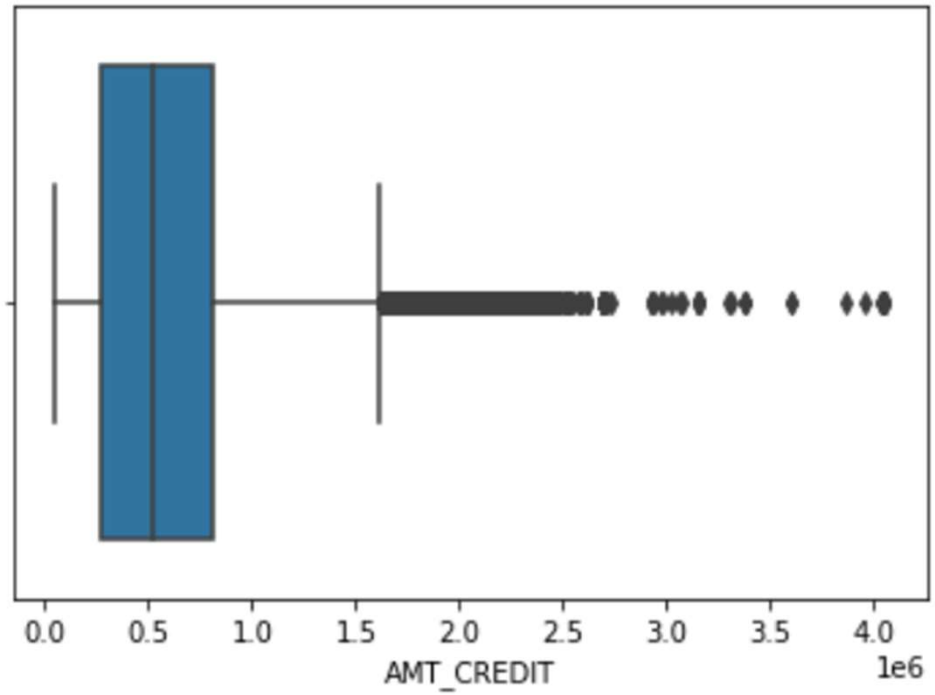
❖**Checking for Outliers:**

❑**For - AMT_INCOME_TOTAL -** this varibale indictes the Income of the client**.**



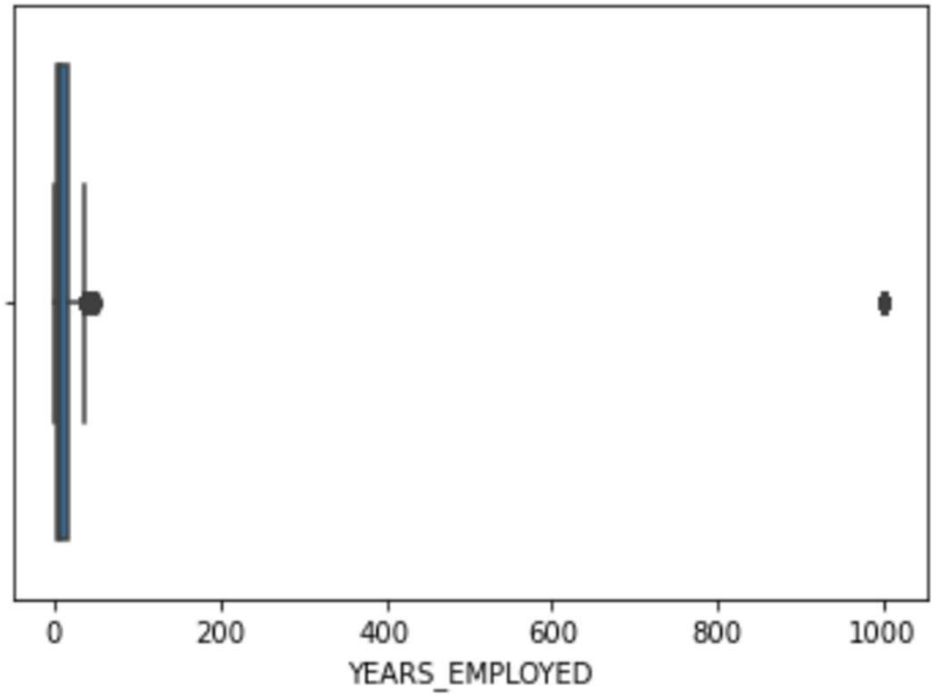
Inference: There is one value which is too high compared to others.hence it is an outlier.

❖ **For - AMT_CREDIT -** this variable indicates Credit amount of the loan.



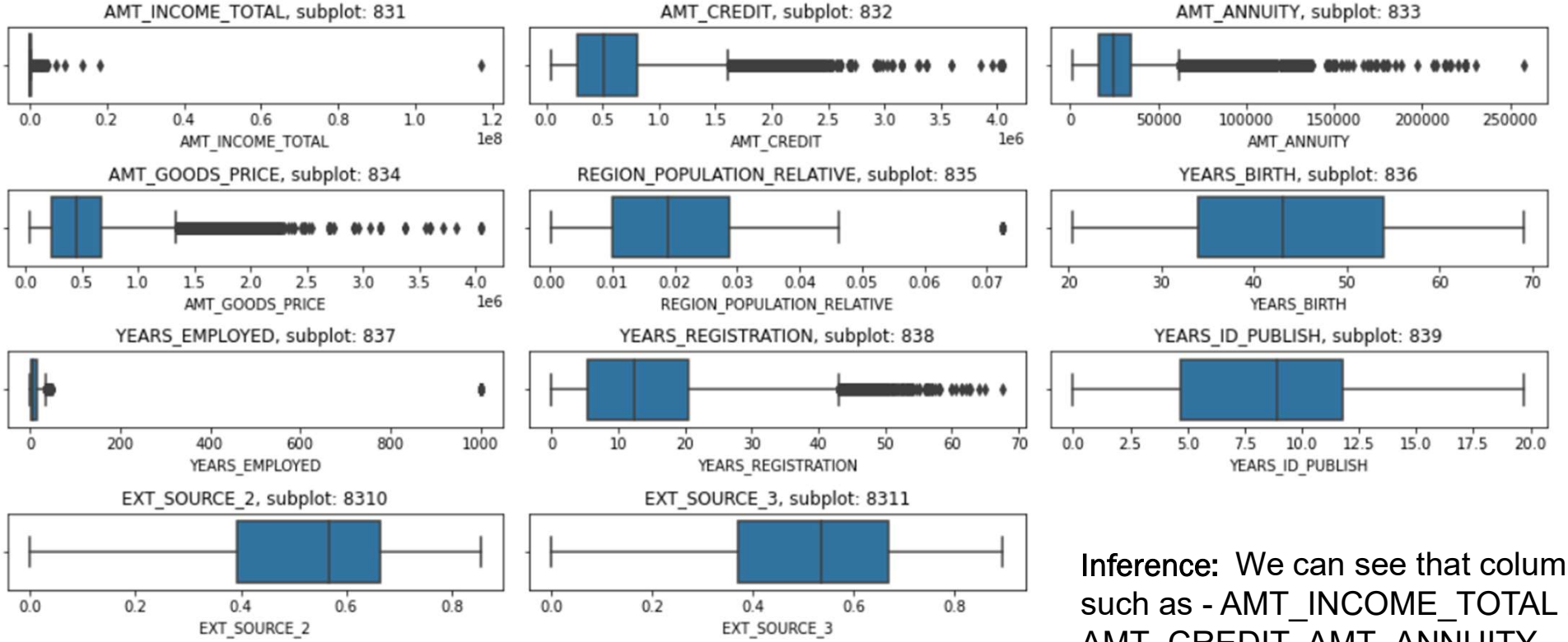
Inference: we can see from the graph there are few outliers. We will check these values to confirm. After checking ,we can see from the values , the AMT_CREDIT is greater than AMT_INCOME_TOTAL in all the cases and then its greater than most values

❖ For variable - YEARS_EMPLOYED - this variable indicates How many years before the application the person started current employment



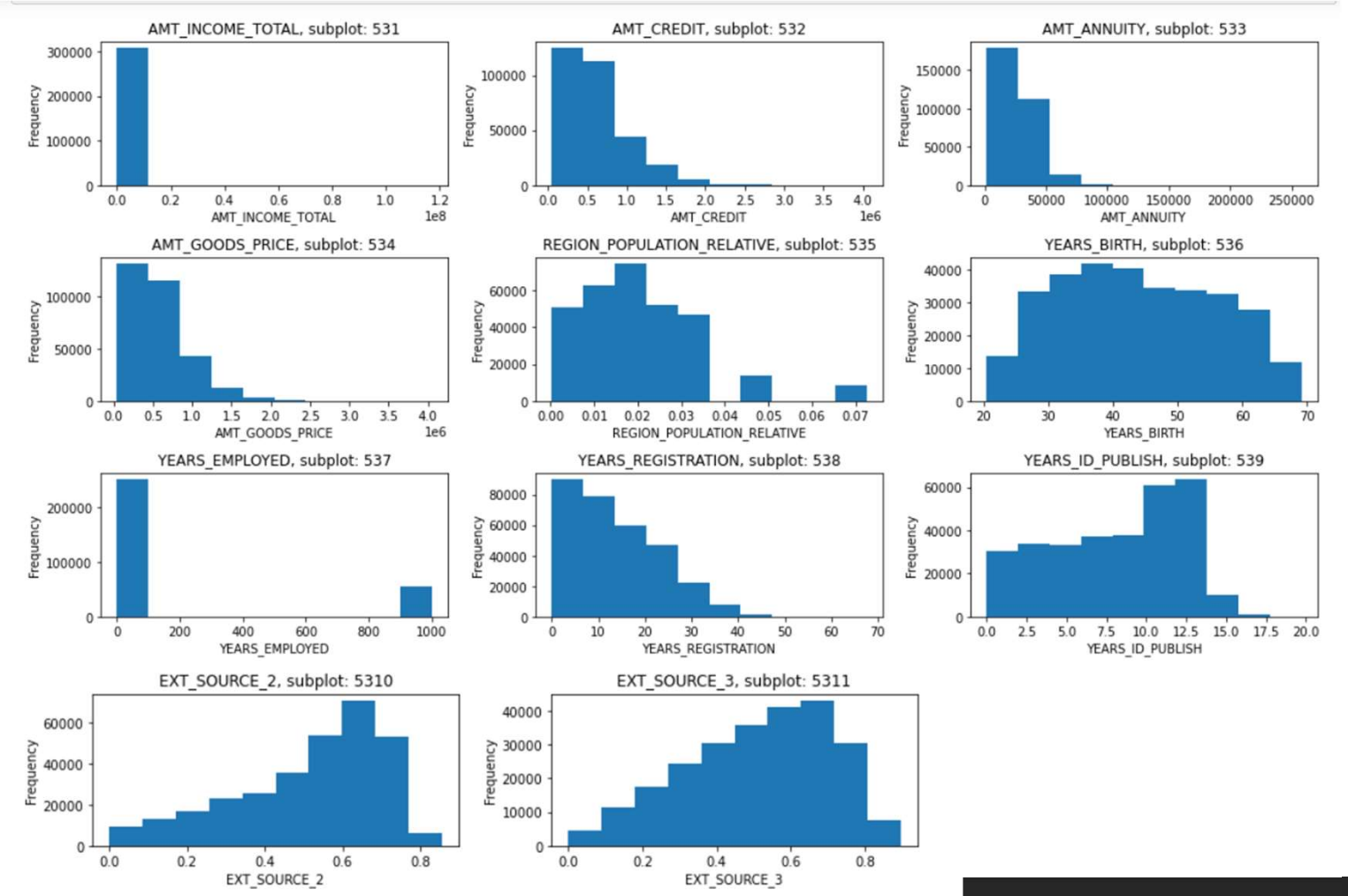
Inference: the outlier value is 1000 yrs.
Which makes the case for it being an outlier

❖ Analysing outliers using boxplots



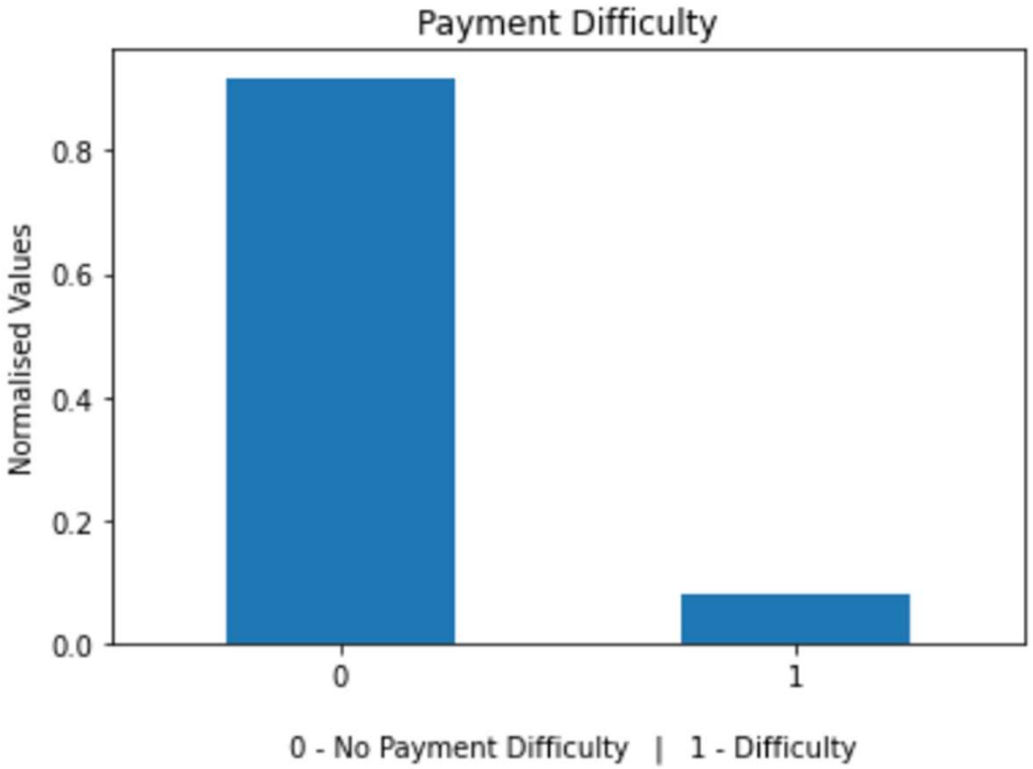
Inference: We can see that columns such as - AMT_INCOME_TOTAL , AMT_CREDIT ,AMT_ANNUITY, REGION_POPULATION_RELATIVE , YEARS_EMPLOYED have outliers

❖ Analysing the same using histograms



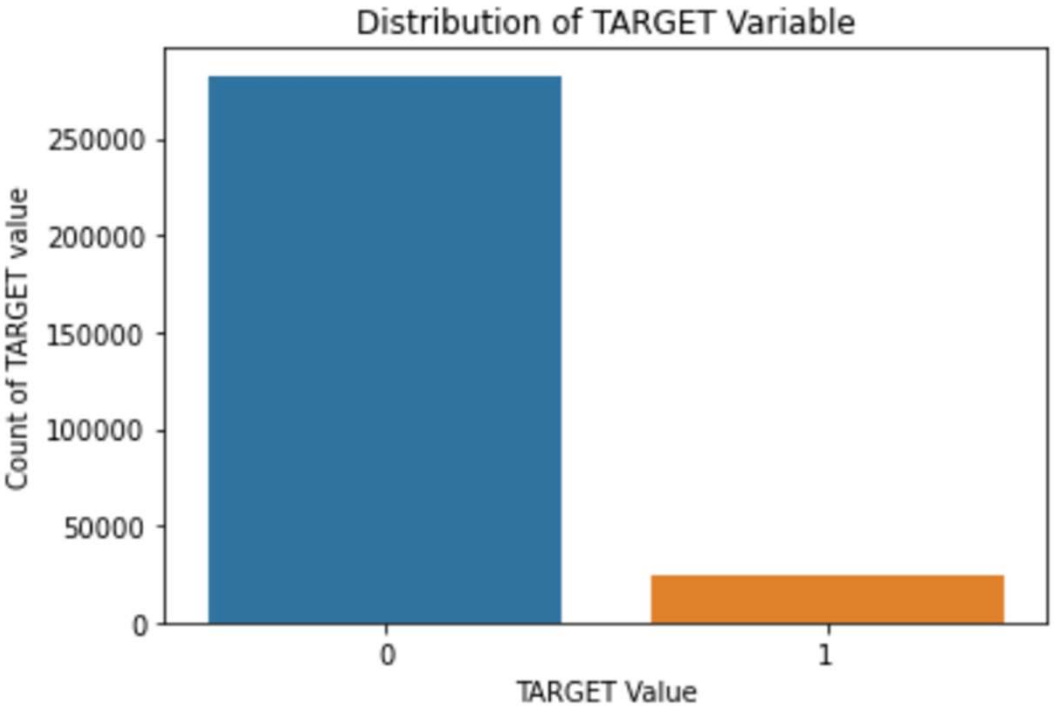
- Univariate Analysis

- Checking for imbalance data



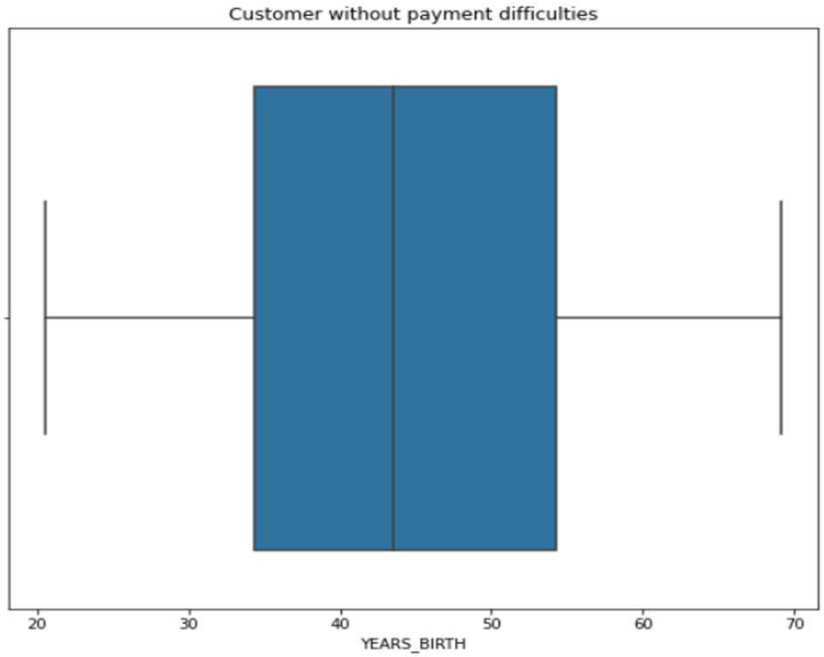
Inference: Highly imbalanced

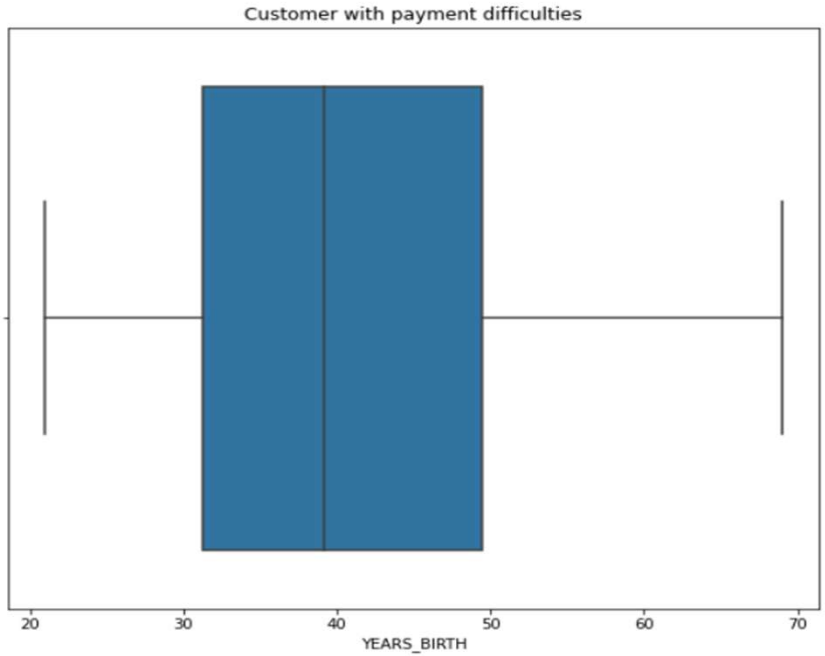
❖ Distribution of target variable



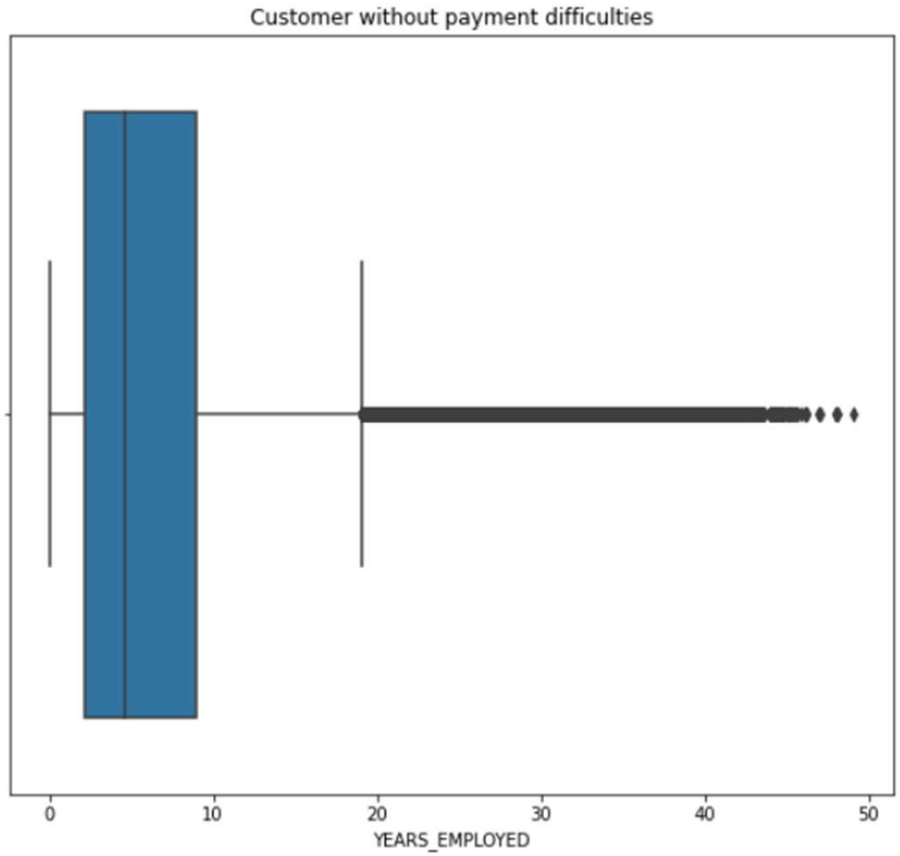
Inference: Further after finding the ratio between the target variables, it indiactes that for every 1 there are almost 11 number of 0's.i.e.,1 in every 11 applicant has payment difficulty. this is a highly imbalanced data set

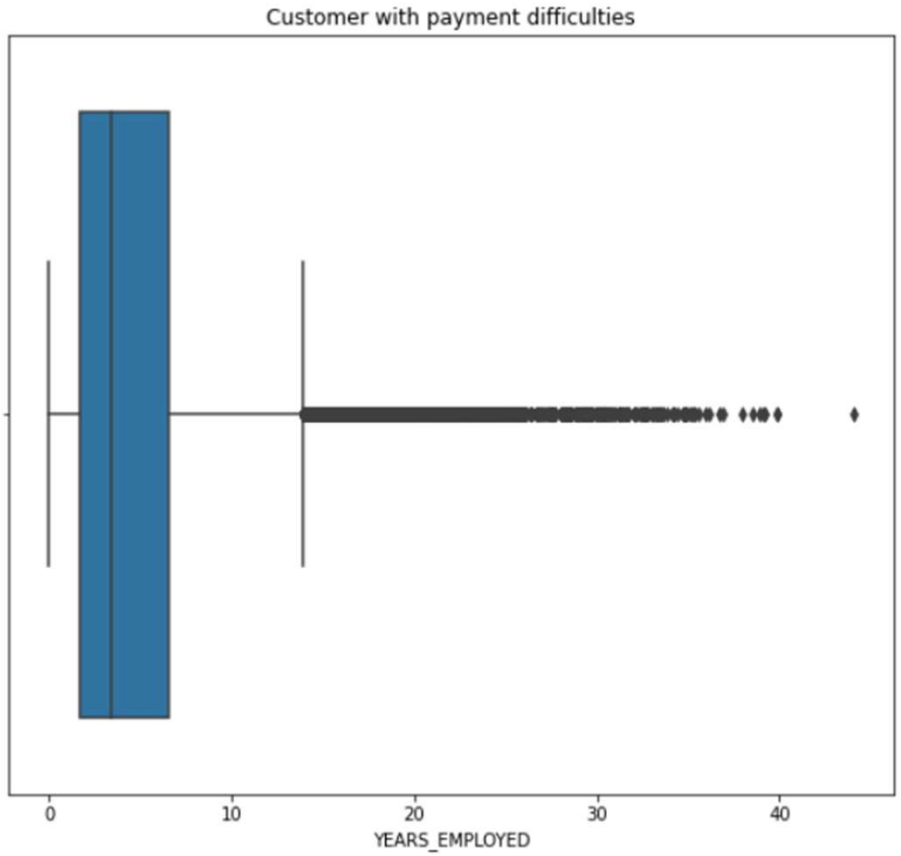
❖ Numeric Variables

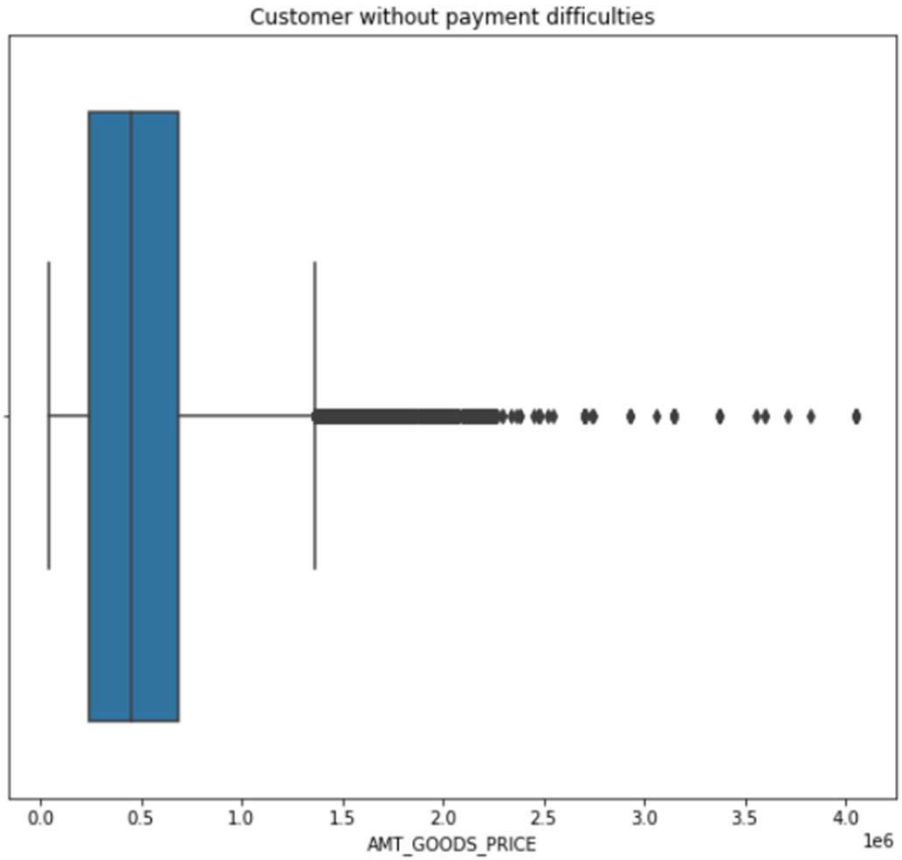


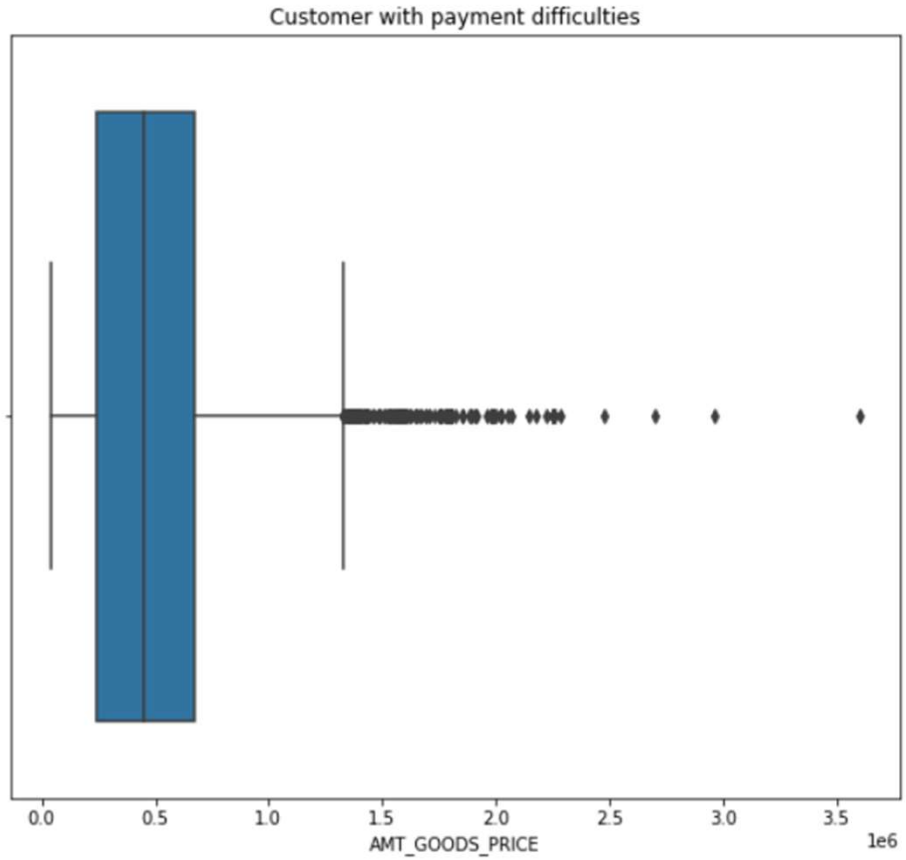


Inference:Customer without payment difficulties having year in between 34 to 54 years , And customer with payment difficulties having in between 31 to 50 years.

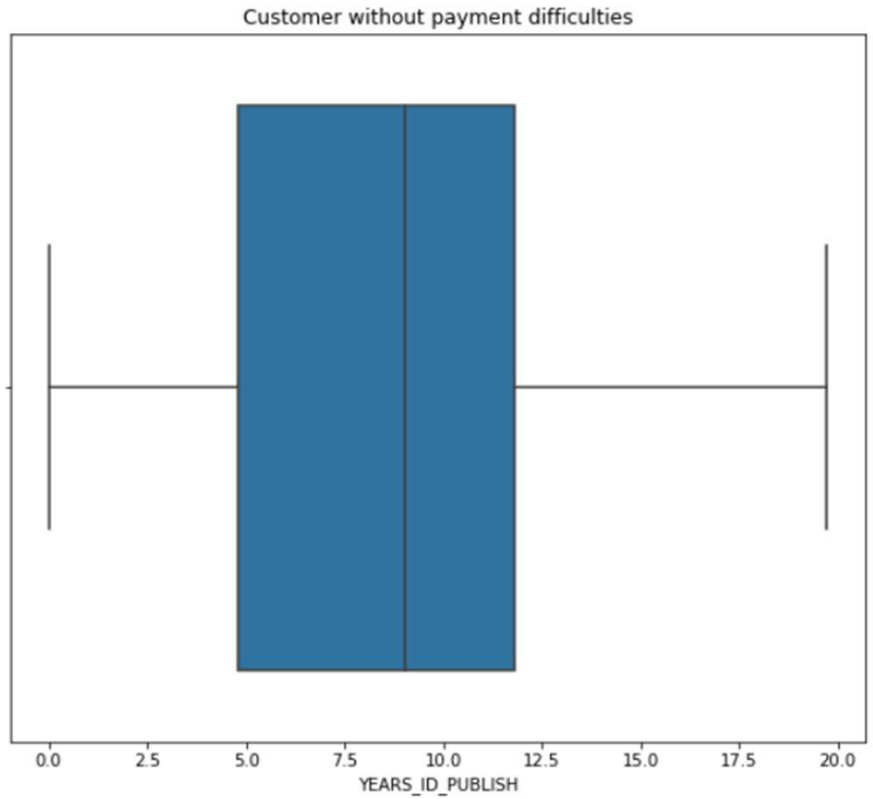

Customer without payment difficulties
0
10
20
30
40
50
YEARS_EMPLOYED

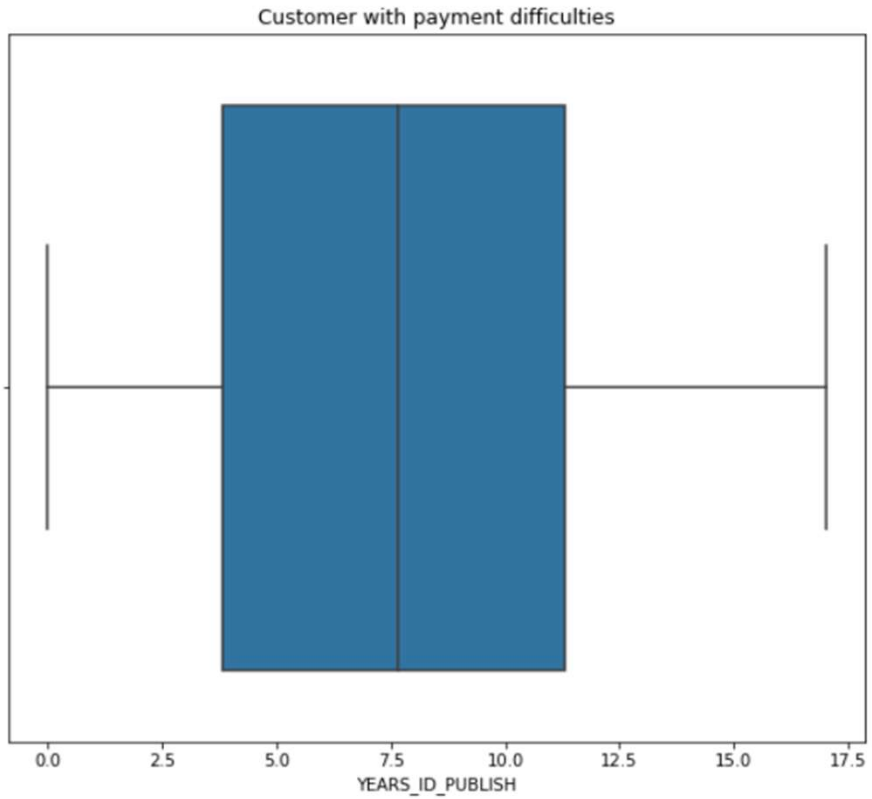

Customer with payment difficulties
0
10
20
30
40
YEARS_EMPLOYED



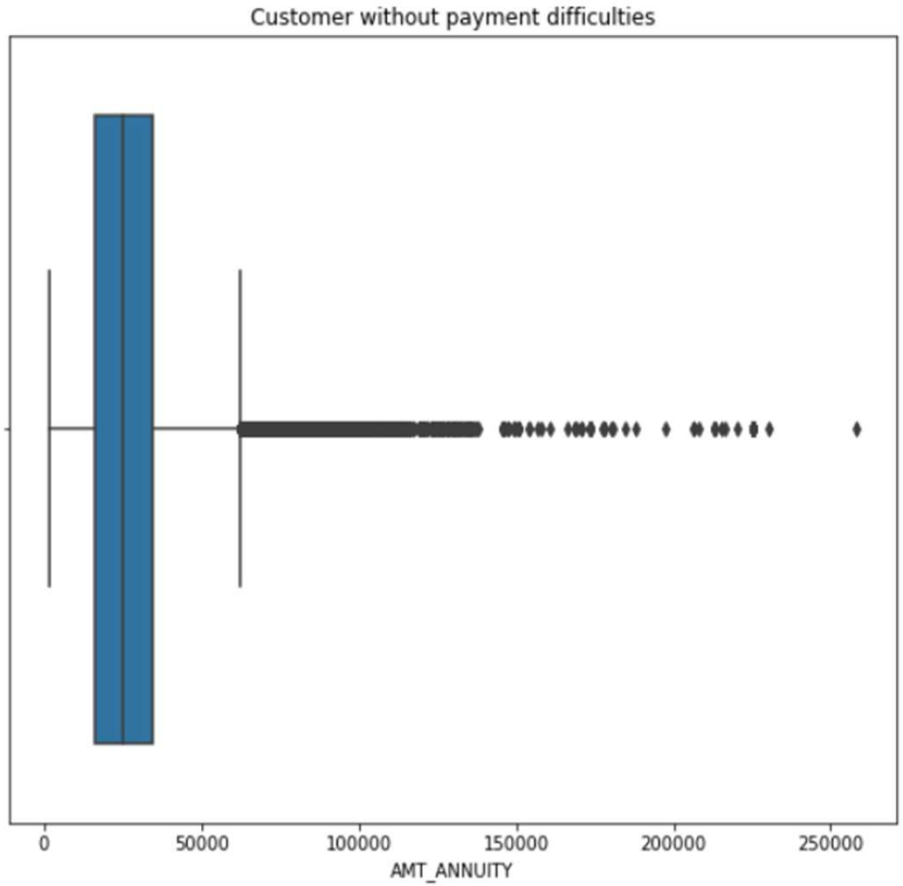


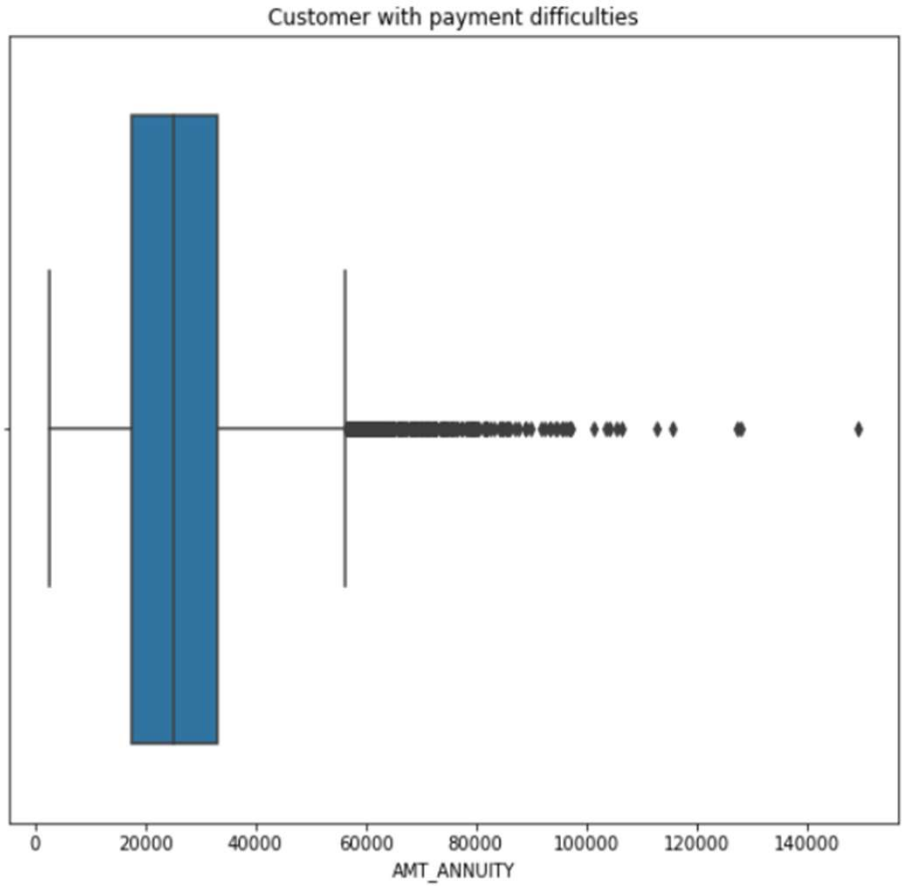
Inference: Customer without payment difficulties lies in between 0.3 to 0.7 and the customer with payment difficulties lies in between the same as of the without payment 0.3 to 0.7. And also both are having the mid value about 0.5.



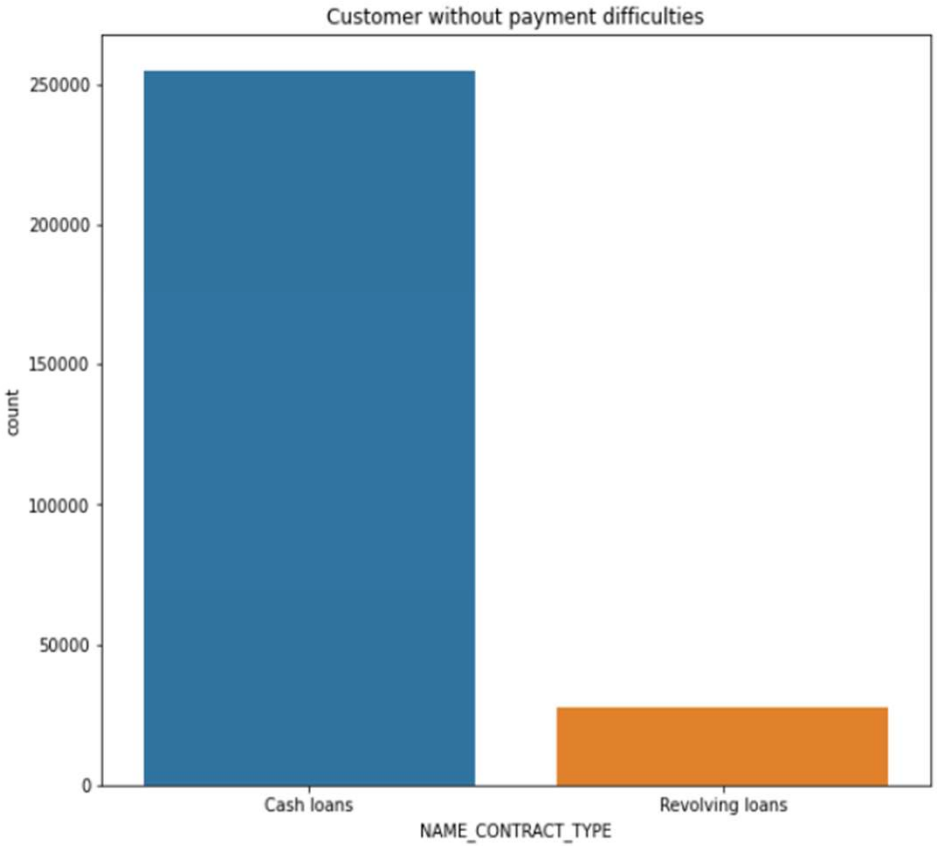


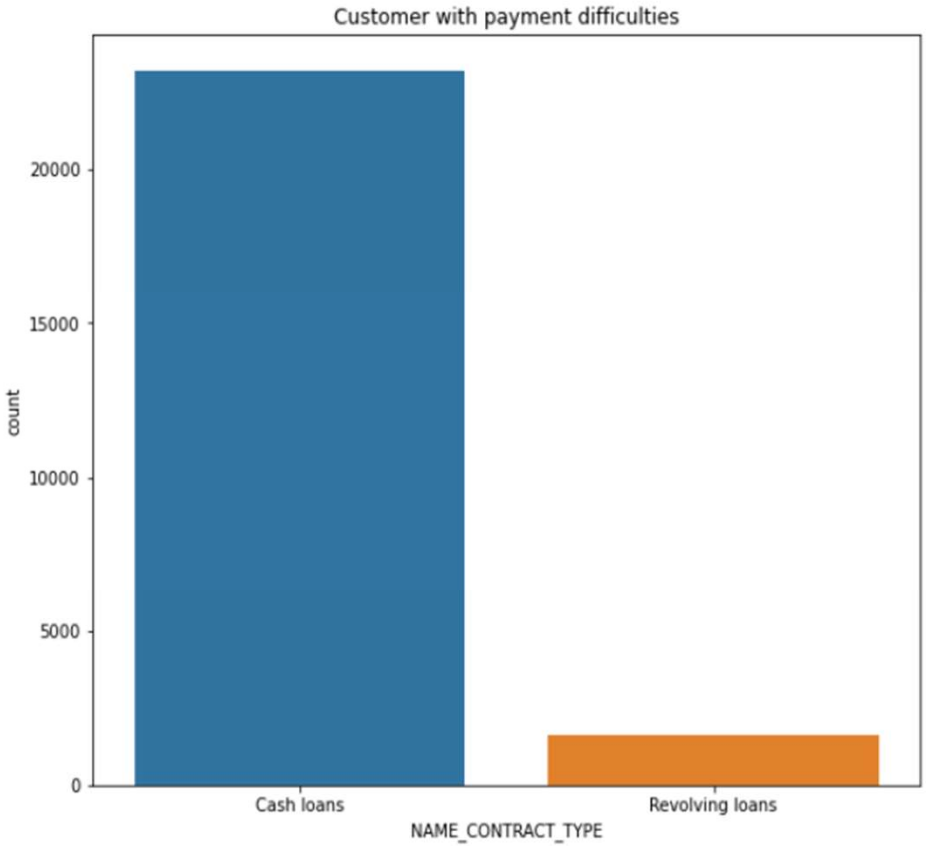
Inference: Customer without payment difficulties lies in between 5 to 11 and Here we can see that the customer with payment difficulties lies in between 3 to 11


Customer without payment difficulties
0
50000
100000
150000
200000
250000
AMT_ANNUITY

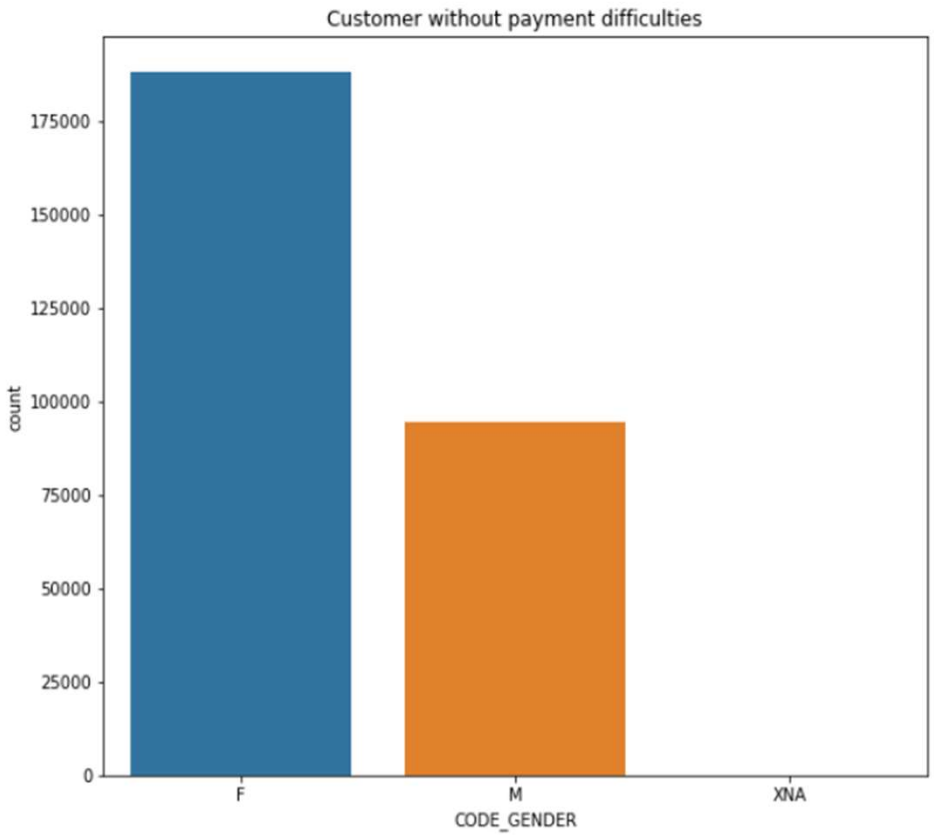

Customer with payment difficulties
0
20000
40000
60000
80000
100000
120000
140000
AMT_ANNUITY

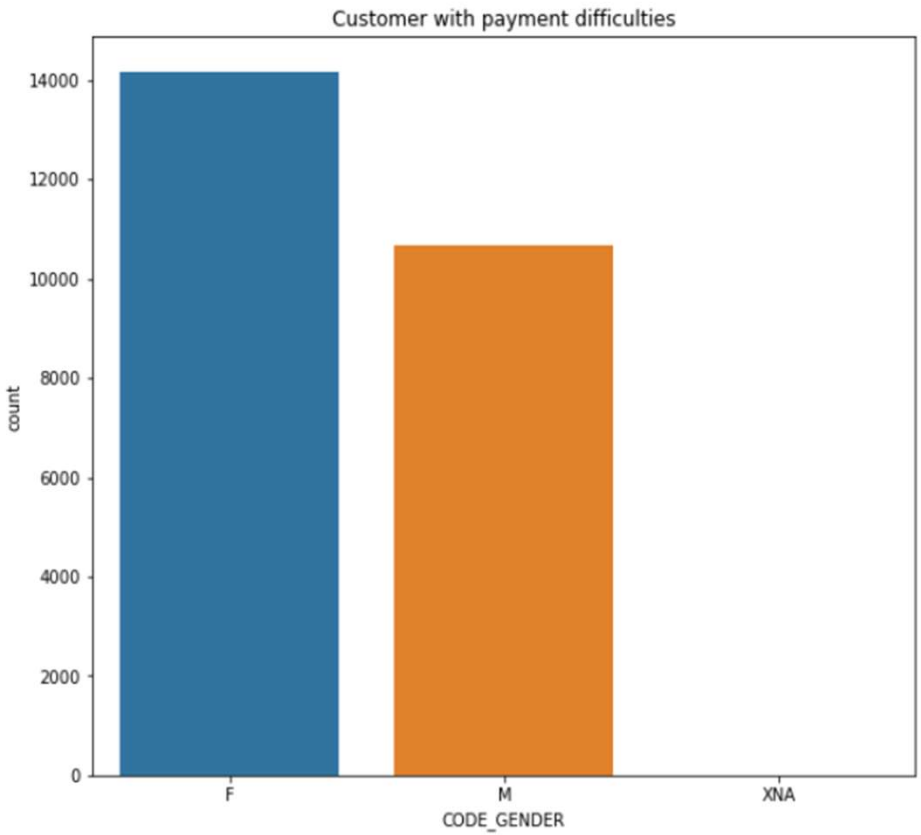
❖ Categorical Variables



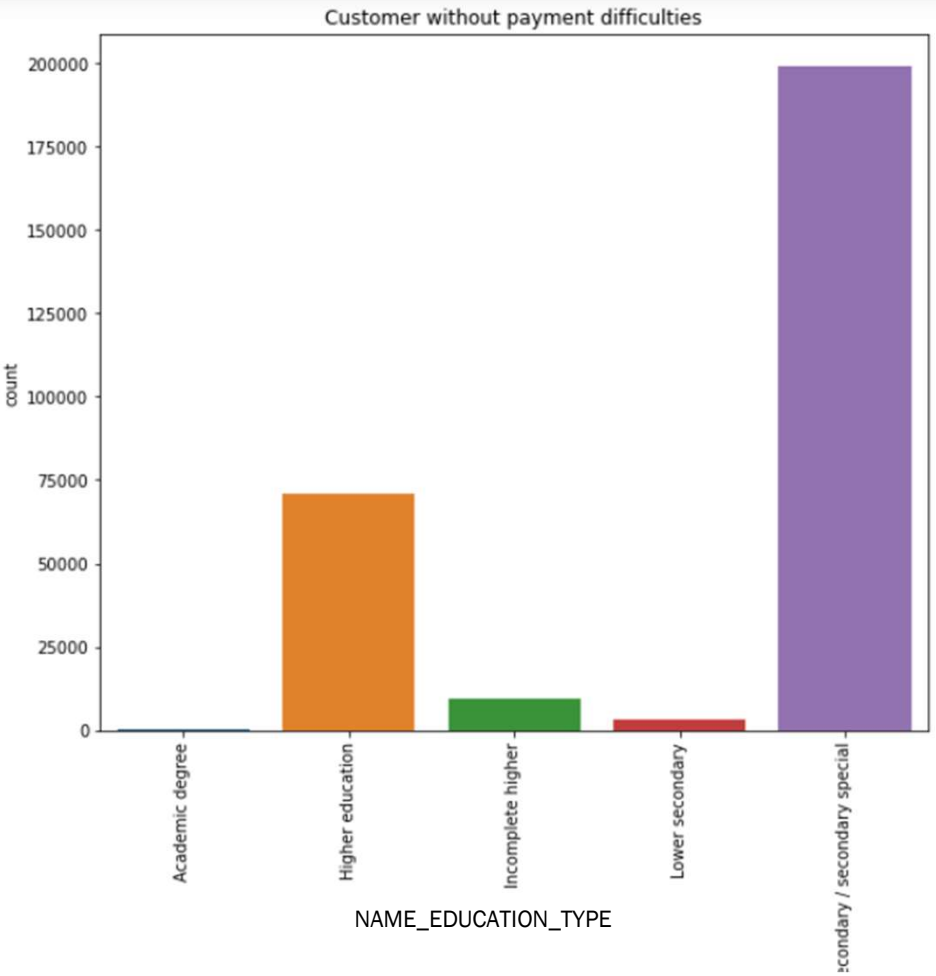


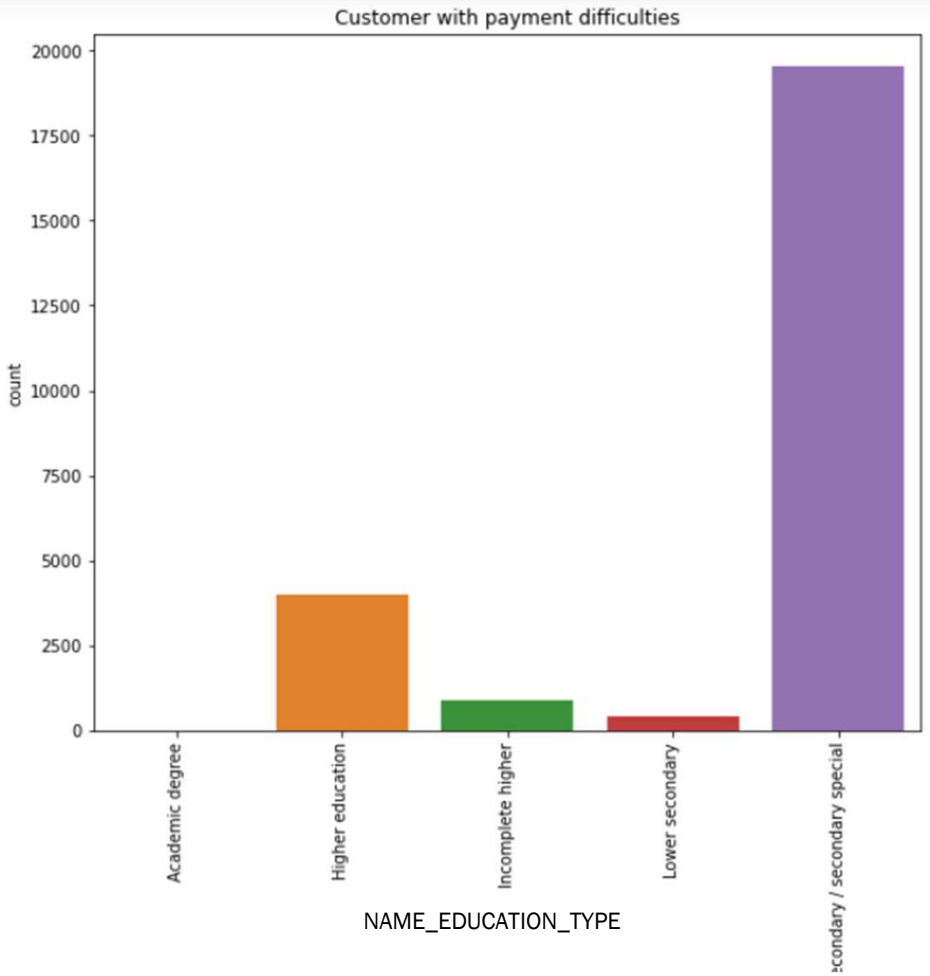
Inference: Customer without payment and customer with payment difficulties both are taking cash loans



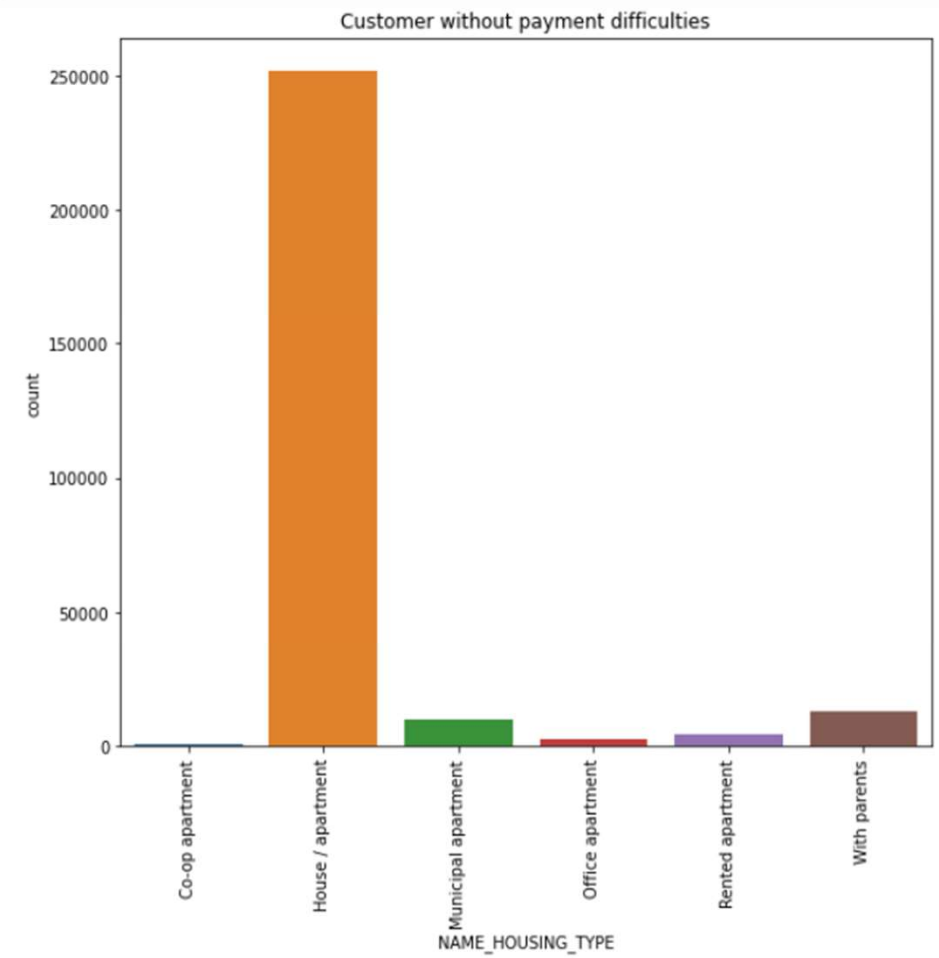


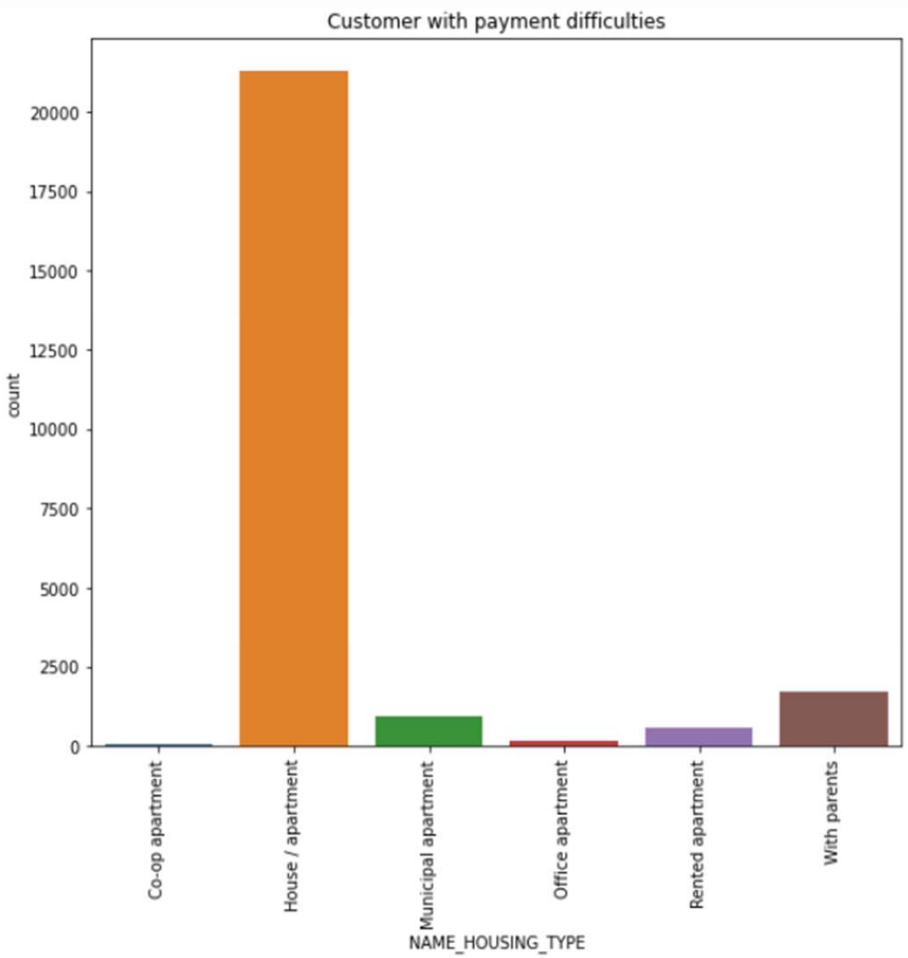
Inference: Female customers are having highest count as compare to male customers in both the cases.



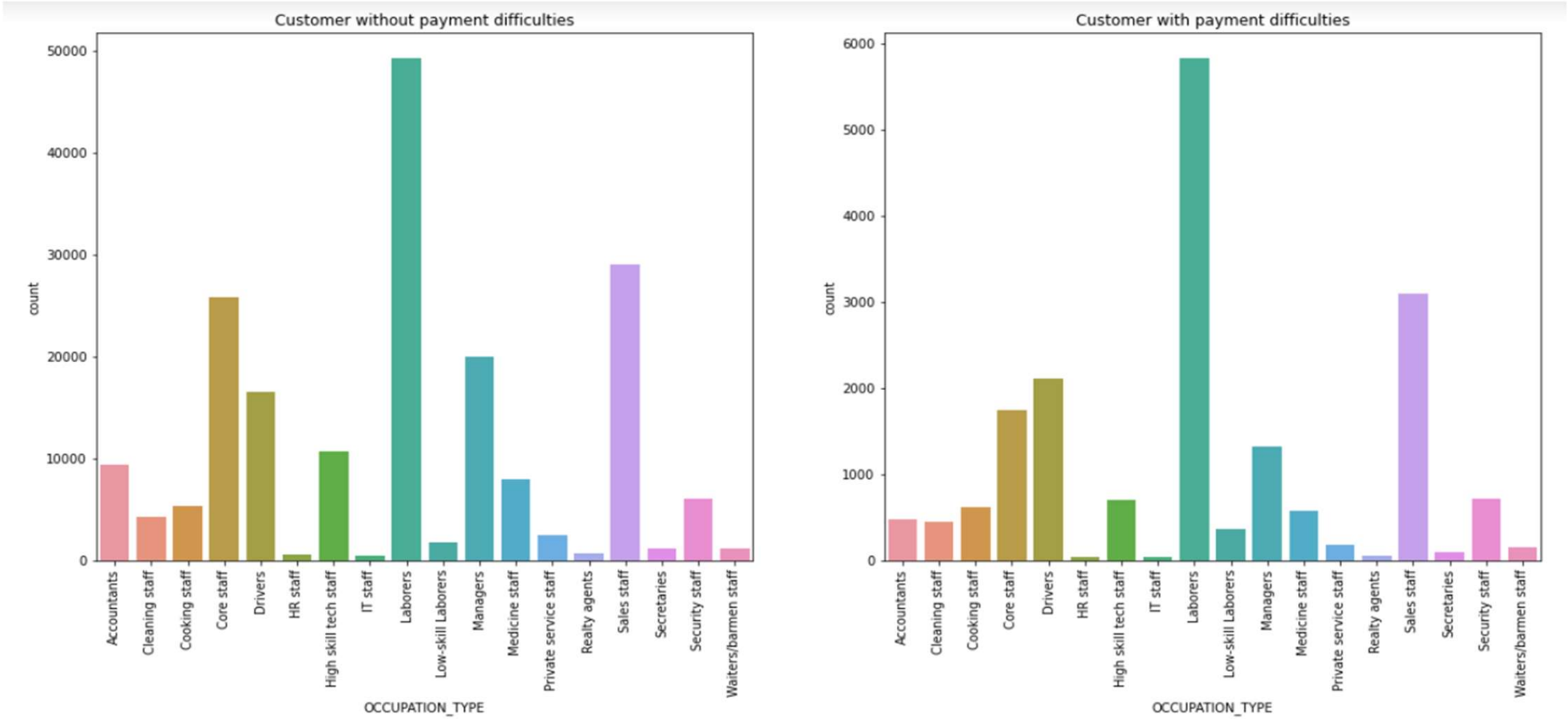


Inference: Customer having payment difficulties in secondary/ secondary special in both the cases.





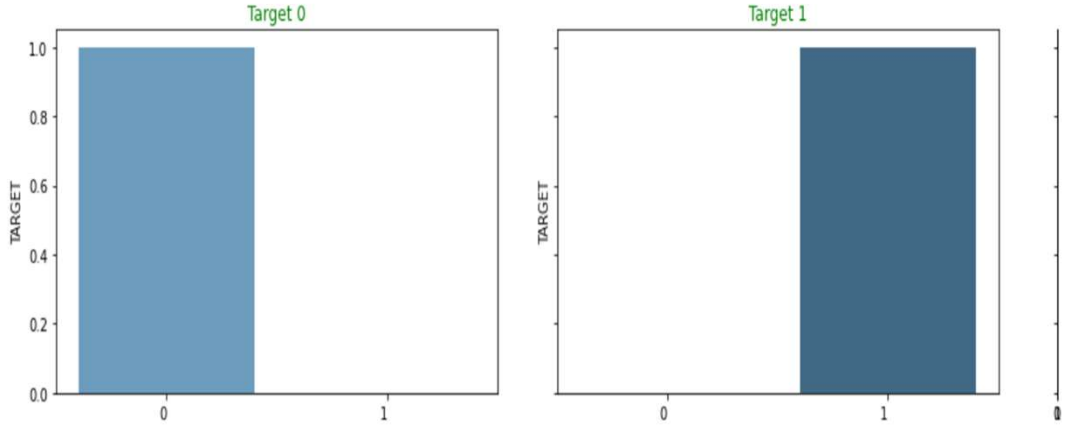
Inference: Payment difficulties in home/ apartment in both the cases. And we can also say that customers take loan for house/ apartment in compare to others



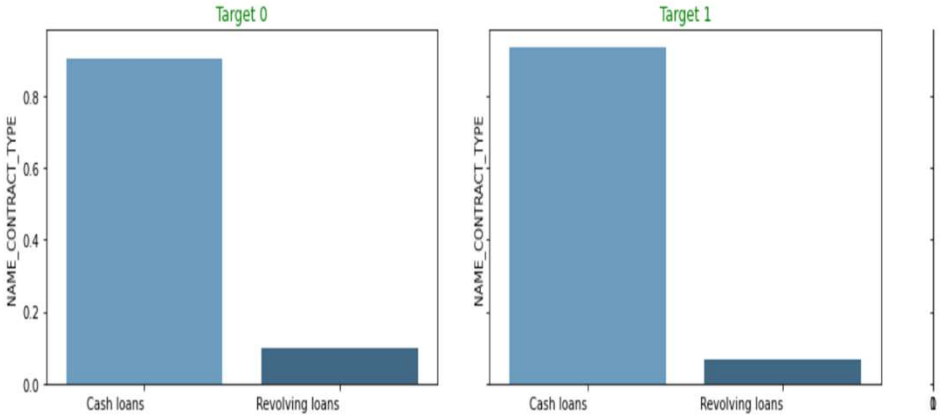
Inference: Laborers are having more difficulties in repaying the loan and also the core staff and the sales staff. But in the case of laborers those wo have without payment is way more then with having the payment.

❖ Categorical Ordered

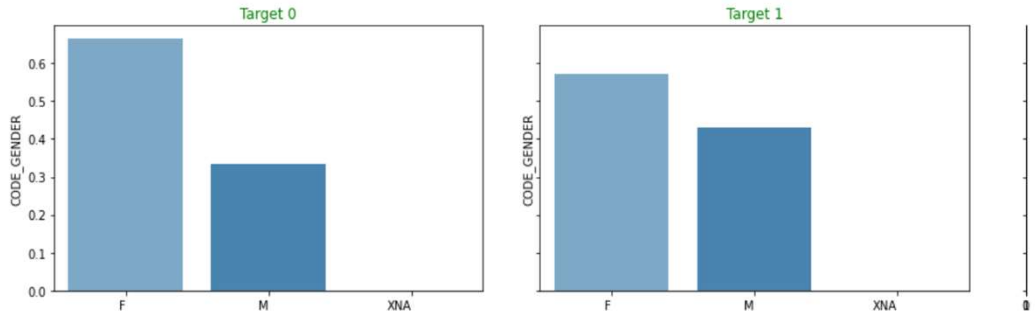
Graph for : TARGET



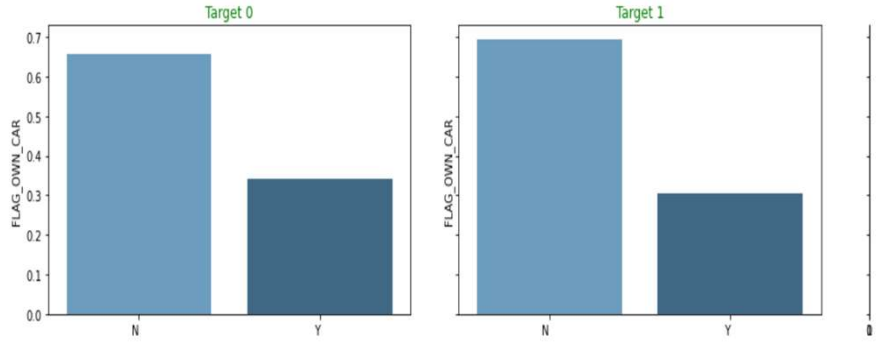
Graph for : NAME_CONTRACT_TYPE



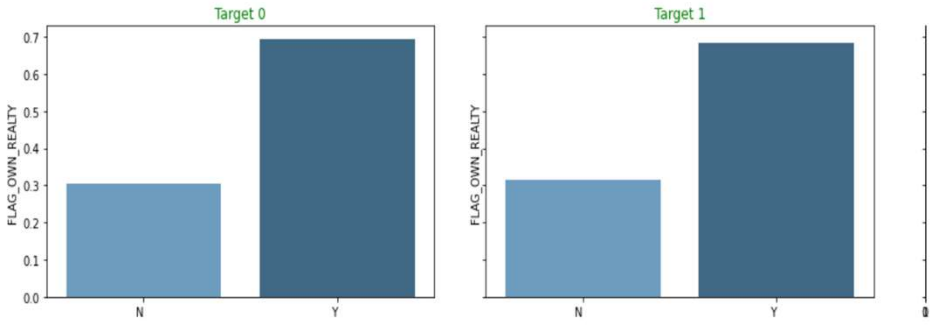
Graph for : CODE_GENDER



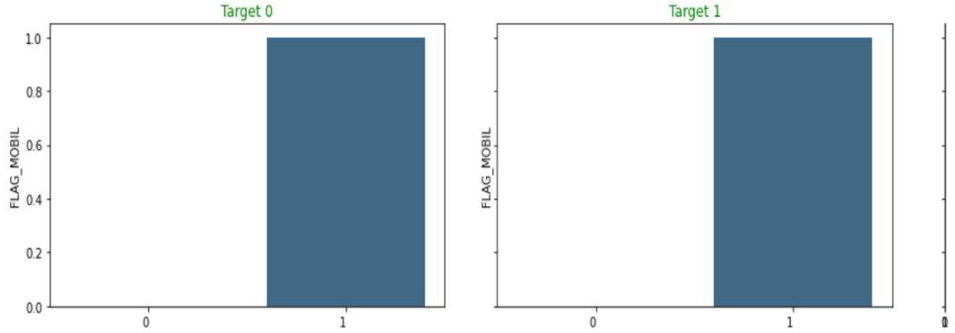
Graph for : FLAG_OWN_CAR



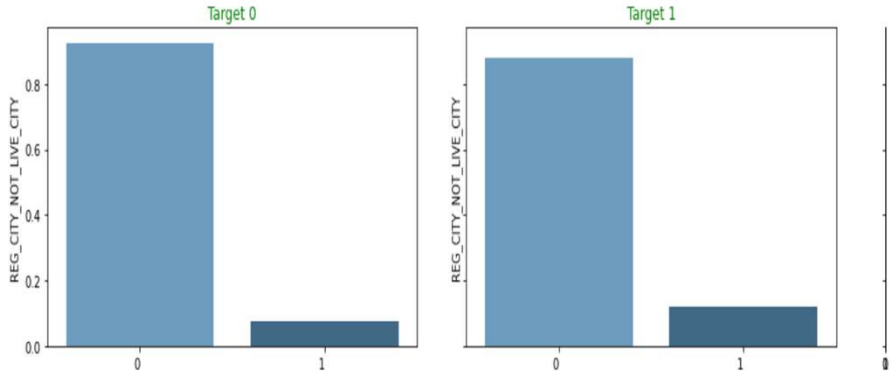
Graph for : FLAG_OWN_REALTY



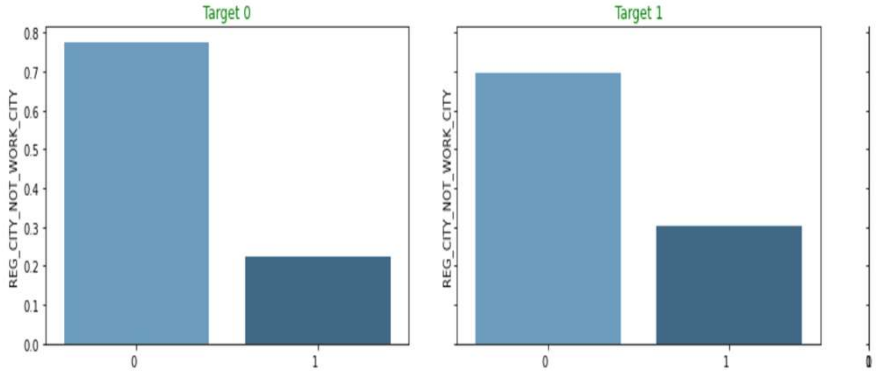
Graph for : FLAG_MOBIL



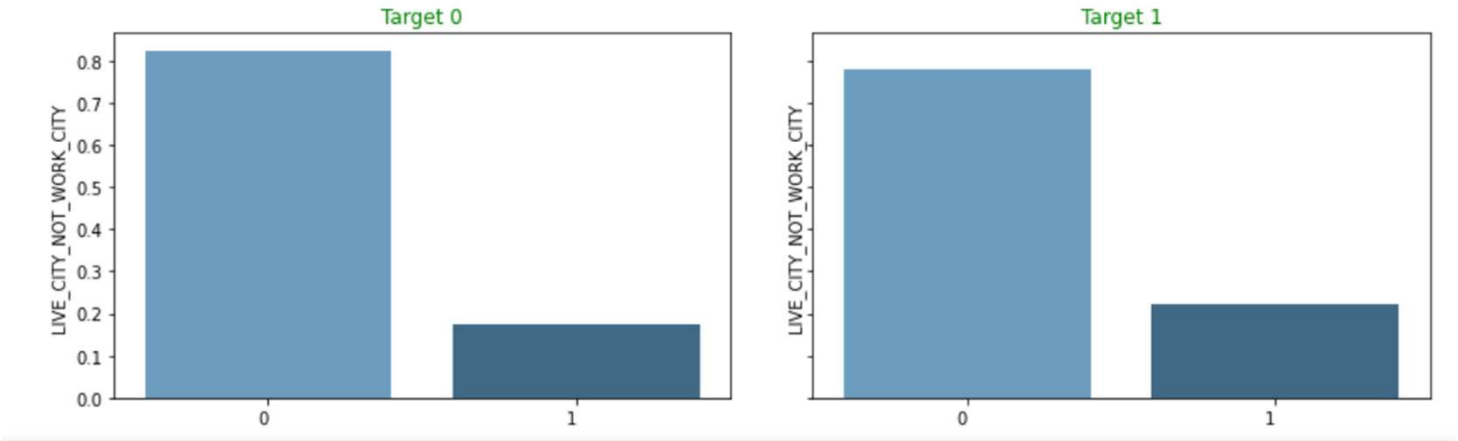
Graph for : REG_CITY_NOT_LIVE_CITY



Graph for : REG_CITY_NOT_WORK_CITY



Graph for : LIVE_CITY_NOT_WORK_CITY



Inferences:

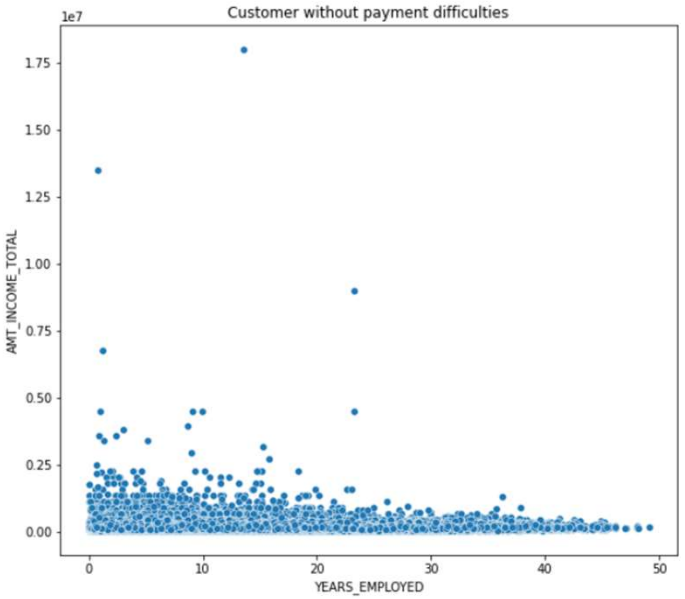
-->REG_CITY_NOT_LIVE_CITY, REGION_NOT_WORK_CITY, LIVE_REGION_NOT_WORK_CITY- For both Target 0 and Target 1 out of Region, ie 1 is very low and does not seem to affect the default rate

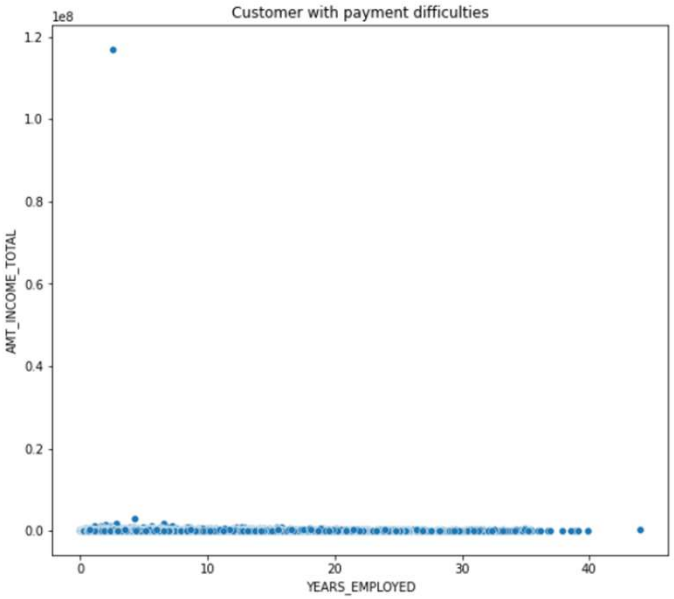
--> REG_CITY_NOT_LIVE_CITY, LIVE_CITY_NOT_WORK_CITY - Default ratio is higher for 1, ie different from permanent address

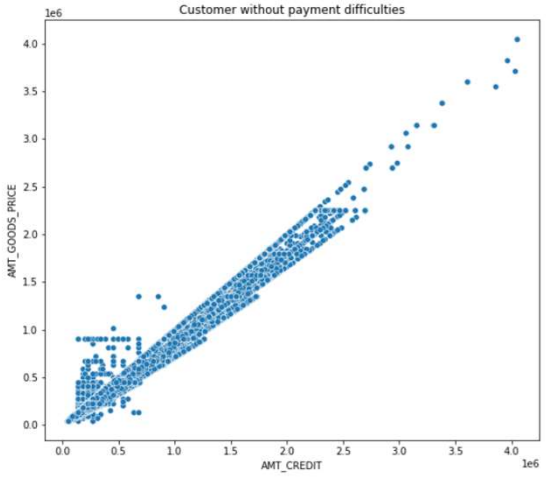
--> CODE_GENDER - Ratio of F to M in Target 0 is 2.3 and F to M in Target 0 - 1.3. indicatign that MEN are defaulting more than Women

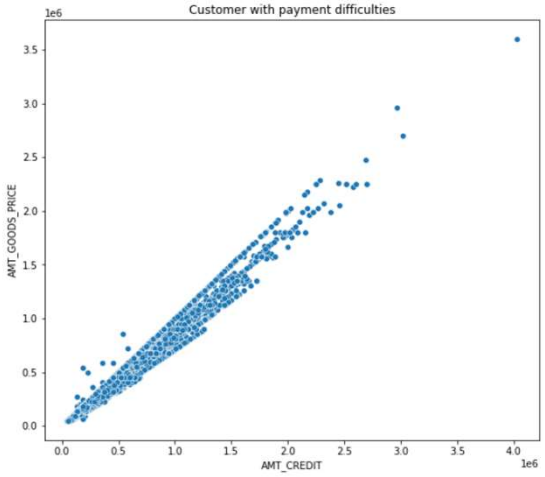
- Bivariate Analysis

- Numeric-Numeric

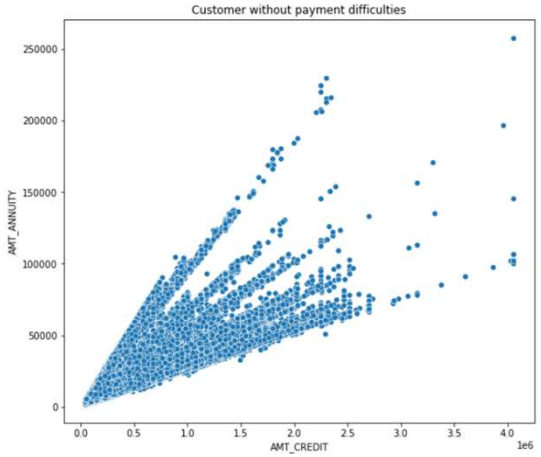


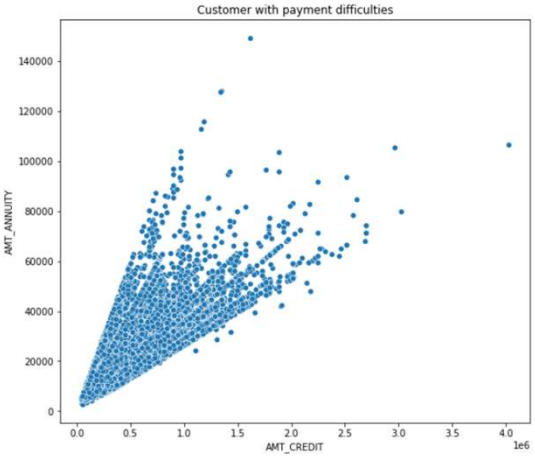






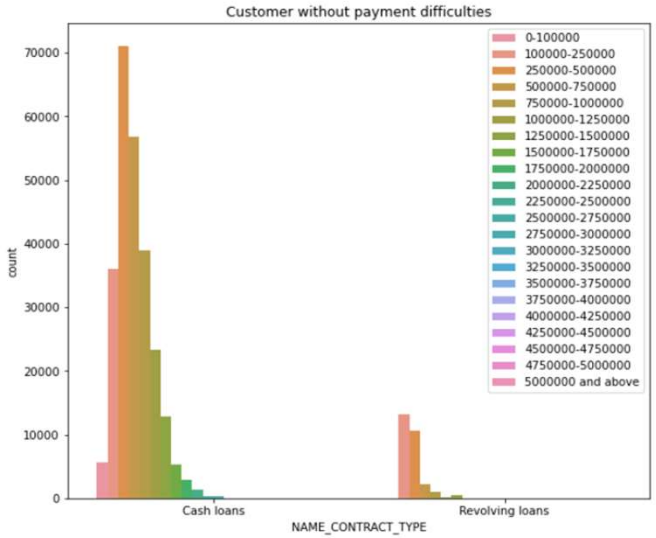
Inference:Goods price is positively correlated with credit amount

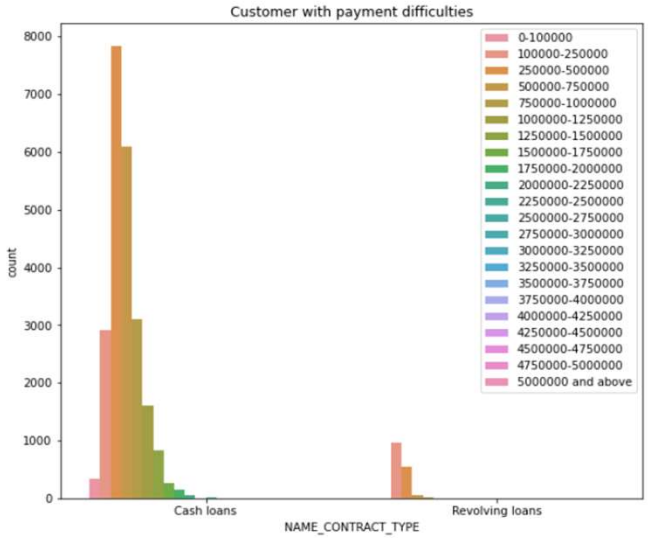


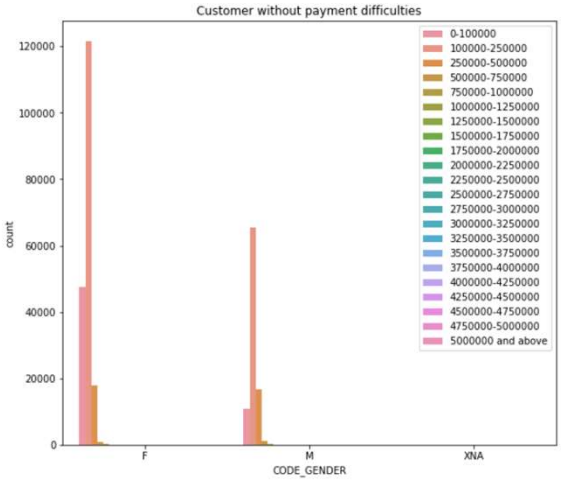


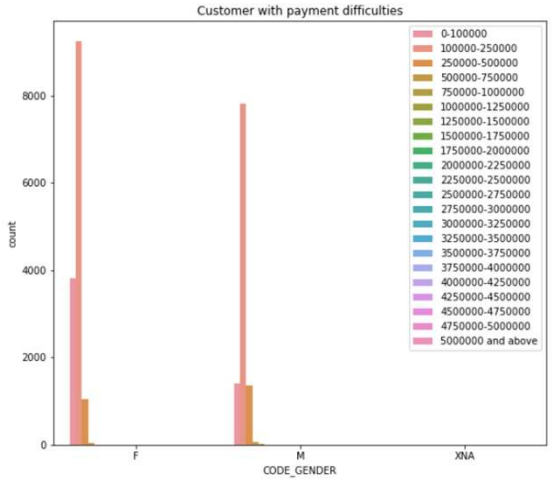
Inference:People without payment difficuties take more credit for the annuity that they have

❖ Categorical- Categorical



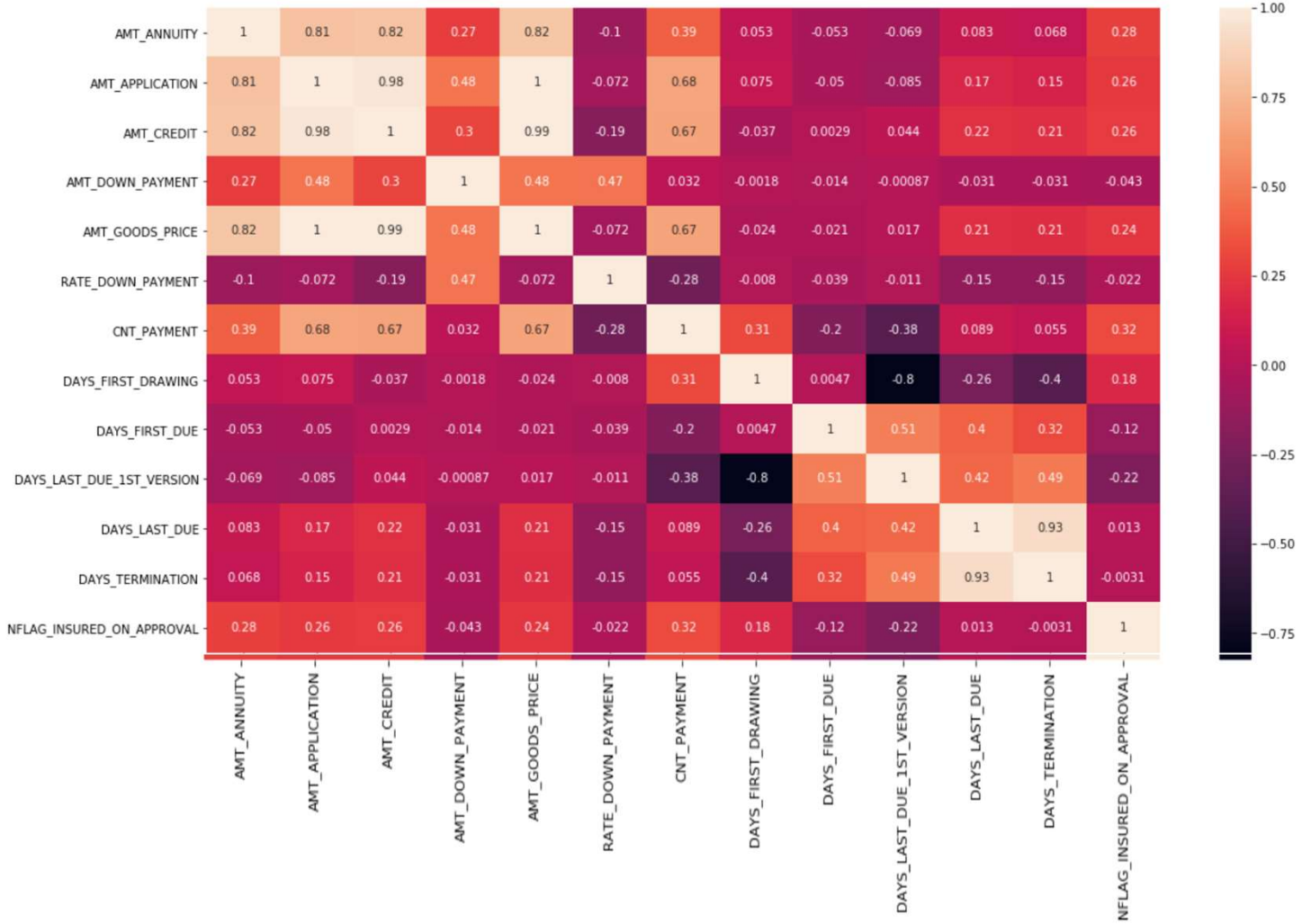






Merging the Application data with Previous data to get Final data frame

- Checking correlation between numeric features of previous application data



Inference:

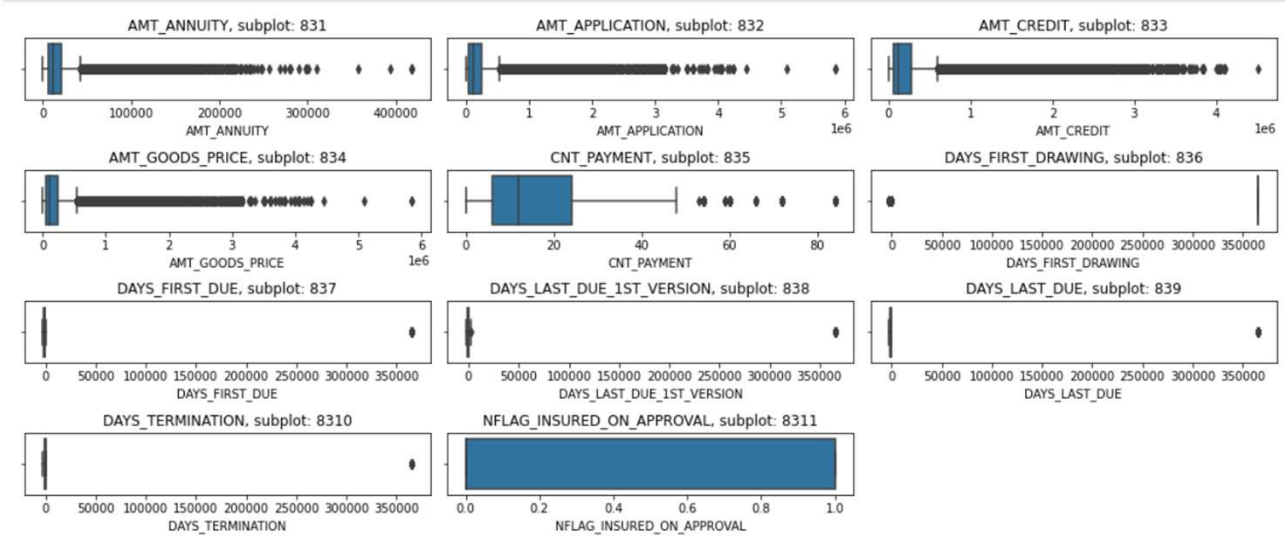
•'DAYS_LAST_DUE' and 'DAYS_TERMINATION' are highly correlated

•'DAYS_FIRST_DRAWING' and 'DAYS_LAST_DUE_1st_VERSION' have high negative correlation

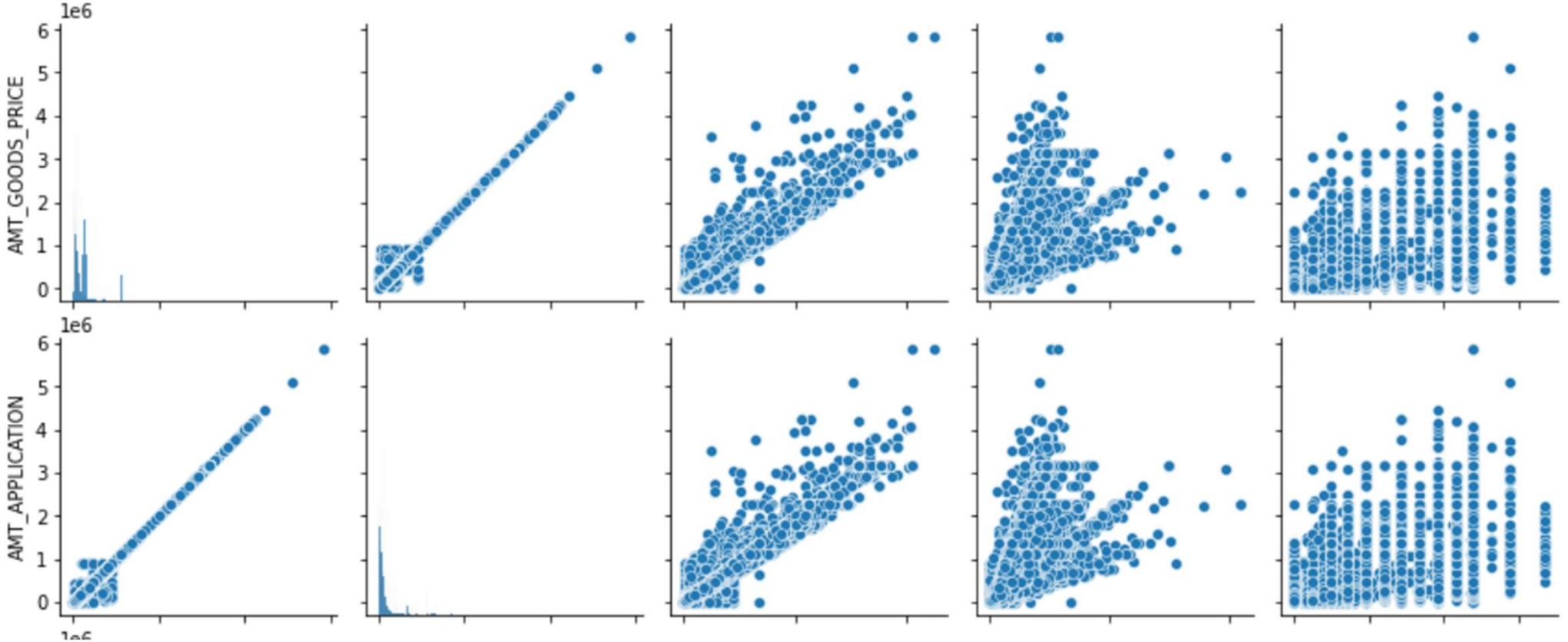
•'AMT_ANNUITY','AMT_APPLICATION','AMT_CREDIT','AMT_GOODS_PRICE' are highly correlated

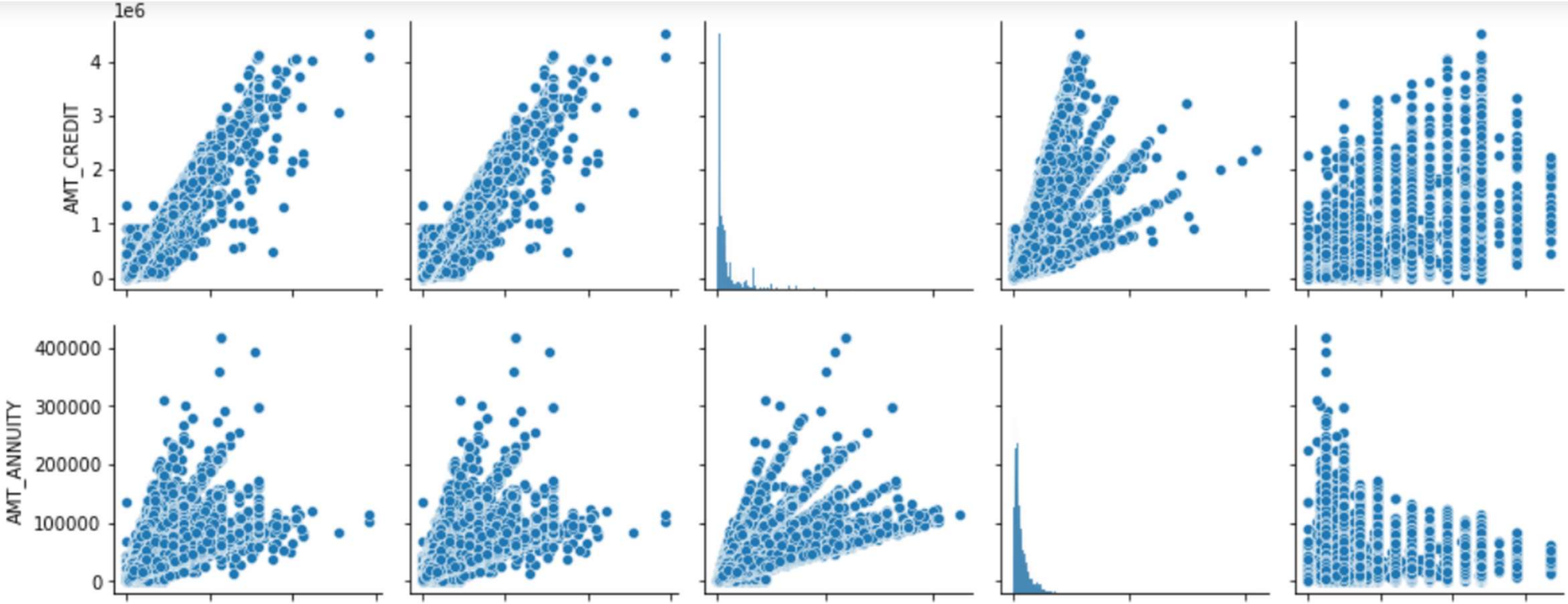
The features can be removed before modelling this data, as they would cause collinearity **'DAYS_TERMINATION','DAYS_LAST_DUE_1st_VERSION','AMT_APPLICATION','AMT_CREDIT','AMT_GOODS_PRICE'** For EDA purpose we are not removing them.

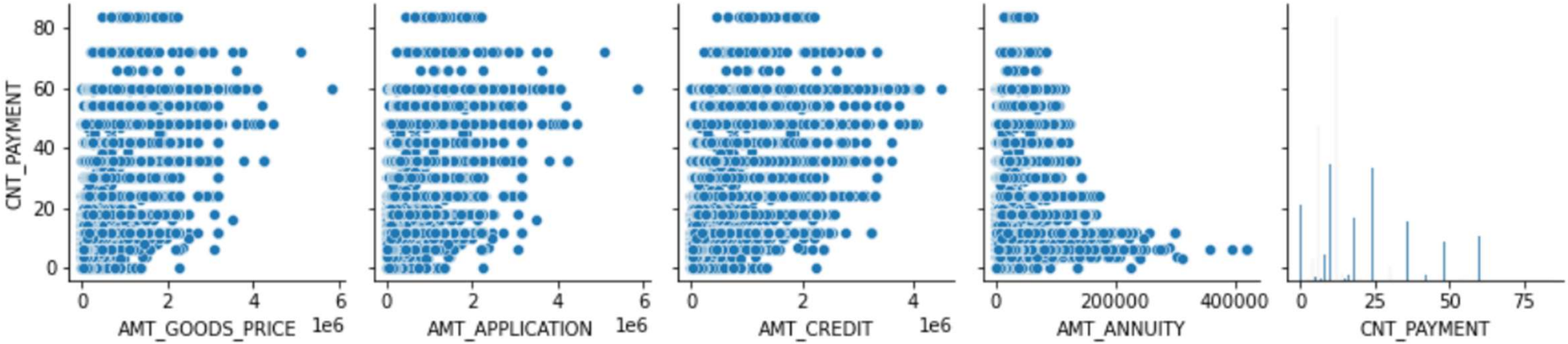
❖ Univariate, Bivariate and Multivariate analysis



Inference: Continuous Variables seem to have high percentage of outliers. Checking the distribution


1e6
AMT_GOODS_PRICE
AMT_APPLICATION
0
1
2
3
4
5
6


1e6
AMT_CREDIT
0
1
2
3
4
AMT_ANNUITY
0
100000
200000
300000
400000



Inference:-->AMT_GOODS_PRICE, AMT_ANNUITY, AMT_APPLICATION - as expected have high correlation.
Higher the value of good purchased more there will be need of loan and surely all these will correlate
-->Similary, AMT_Credit to AMT_GOOD_PRICE also the correlation is high
--> Column CNT_Payment ideally should have had a high correlation with AMT_credit, ie higher credit, more the term of loan. But no such correlation can be seen.

Correlation

Inference:

-->Unused offer application amount is low
-->Cancelled application amount is high. The bank may be refusing these possibly as the Debt liabiltiy ratio of consumer must be going high due to the high amount and thus credit default risk.
-->Repeater's application amount is height than the New customers. This may indicate that the bank has more conducive policies/rate of interest etc for repeat applicants

- Note: More analysis and inferences are made , refer python notebook for more

Conclusions:

-->There are feature columns in the dataset that are highly correlated to each other. Which means both will have similar impact on the target value. Those features can be removed before feeding this data to a model to avoid collinearity.

-->Feature columns with 50% or more missing data can be dropped.

-->This dataset is highly imbalanced

-->The applicants whose previous loans were approved are more likely to pay current loan in time, than the applicants whose previous loans were rejected.

-->NAME_CONTRACT_STATUS is an important feature.

-->7% of the previously approved loan applicants that defaulted in current loan

-->90 % of the previously refused loan applicants that were able to pay current loan 'SCO', 'LIMIT' and 'HC' are the most common reason of rejection.

-->Most of the people did not request insurance during previous loan application.

-->For "Cars" defaulter percentage is highest .

-->'NAME_PORTFOLIO' is an important feature for analyzing 'TARGET' variable.

-->15% loan applicatiant defaulted for AP+ (Cash Loan).

-->'CHANNEL_TYPE' is an important feature for analyzing 'TARGET' variable.

Credible Applications refused:

-Unused applications have lower loan amount. Is this the reason for no usage?

-Female applicants should be given extra weightage as defaults are lesser.

-60% of defaulters are Working applicants. This does not mean working applicants must be refused. Proper scrutiny of other parameters needed

-Previous applications with Refused, Cancelled,Unused loans also have cases where payments are coming on time in current application. This indicates that possibly wrong decisions were done in those cases.

-->Other IMPORTANT Factors to be considered

-Days last phone number changed

- Lower figure points at concern

-No of Bureau Hits in last week. Month etc – zero hits is good

-Amount income not correspondingly equivalent to Good Bought

– Income low and good value high is a concern

-Previous applications with Refused, Cancelled, Unused loans also have default which is a matter of concern.

This indicates that the financial company had Refused/Cancelled previous application but has approved the current and is facing default on these.

Summary

-->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.