EDA Case Study: Credit Risk Analytics

A Presentation by
Malavika V and Meera K

INTRODUCTION

This case study aims to give an idea of applying EDA in a real business scenario. In this case study, we will develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

BUSINESS UNDERSTANDING

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 consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study. The company can utilise this knowledge for its portfolio and risk assessment.To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

PROBLEM STATEMENT

The company has to decide for loan approval based on an applicant's profile. There are two types of risks associated with the bank's decision:

1) If the applicant is likely to repay the loan, then not approving the loan
results in a loss of business to the company.
2) 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

APPROACH

We perform Exploratory Data Analysis (EDA) on the given datasets with the help of
following information:

1)Loan payment status as per the 'application_data.csv' file
2) Decision on loan application as per the 'previous_application.csv' file

- Approved - by the Company
- Cancelled - by the Client
- Refused - by the Company
- Unused offer - by the Client

STEPS FOLLOWED:

1) Data Sourcing -
- Reading the data
- Structure of DataFrames

2) Data Cleaning -
- Handling Null Values
- Standardize Values
- Null Value Data Imputation
- Identifying Outliers

3) Data Analysis -
- Imbalance Data
- Univariate Analysis

Bivariate Analysis
- Merged Dataframe Analysis

DATA SOURCING

TASK - 1

Reading the data -

◯ The application_data.csv and previous_application.csv files are read into their respective dataframes using pd.read_csv().

DATA SOURCING

Structure of DataFrames -

◯ Basic inspection is performed on the dataframes to determine:

■ Column wise information using .info()

■ Dimension using .shape

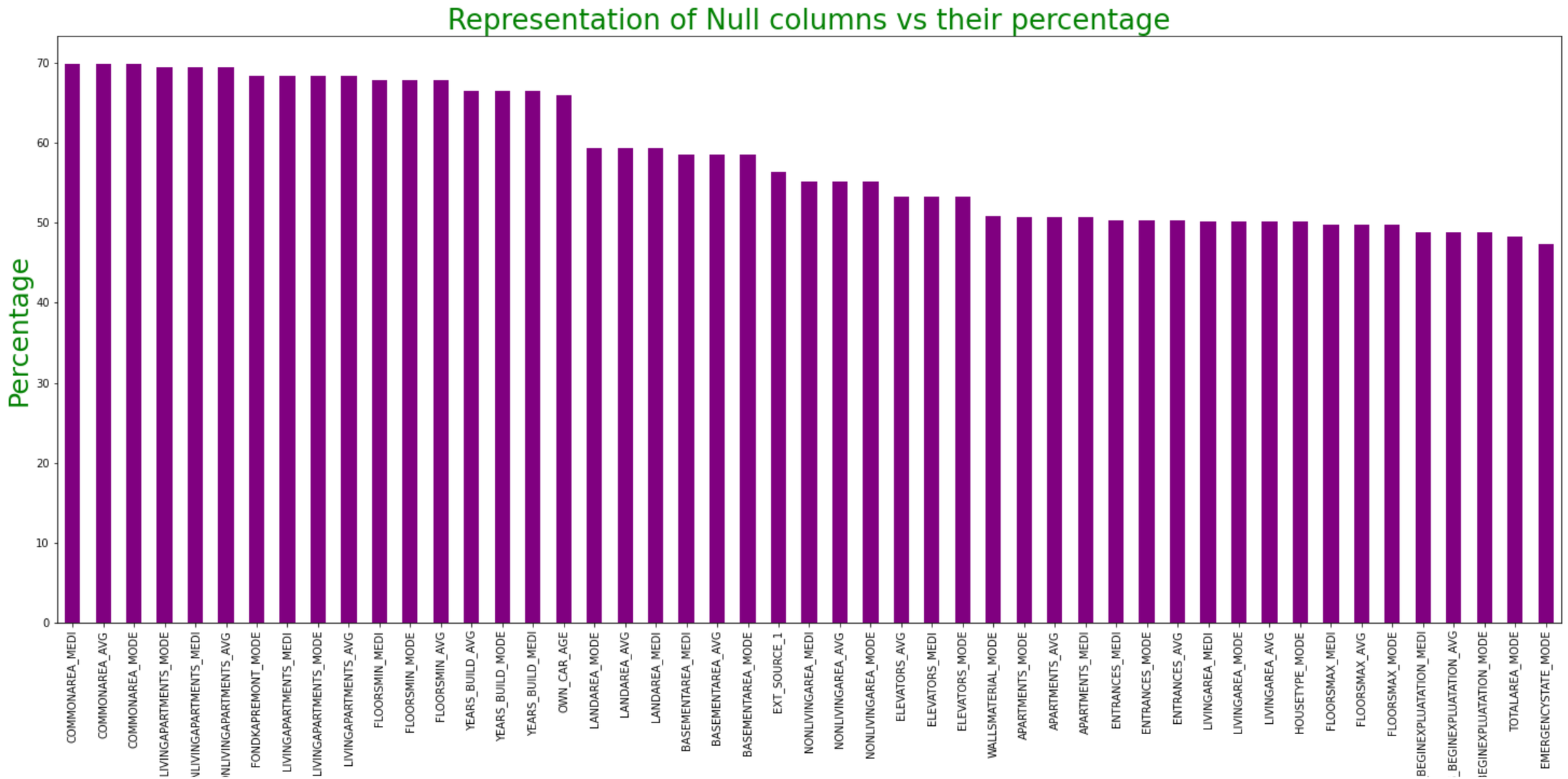
■ Summary statistics of numeric columns using .describe()

DATA CLEANING AND MANIPULATION

TASK - 2

It is observed that there are many columns in with high missing values (more than 47%, i.e. almost 50%). Such columns are handled by either dropping them or imputing values in them based on their relevance.

Visualizing the Percentage of the columns which is having Null values!



After inspecting the null columns and found out that there is no significance reason to keep the null columns more than 32%, so, dropping those columns

Handling null/missing values in each column:

1) we can impute the missing values of these columns with the mode value, which is '0'. We can see those columns are about the credit inquiries made by bank. So, imputing mode value seems to the good approach
2) Dropping the columns which starts with FLAG_DOCUMENT. Since, it doesn't contain any information about what type of docyments these are.

TASK - 3
HANDLING THE ERRORS

1) Checking the unique values of the columns starts with "DAYS" and handling the negative values.

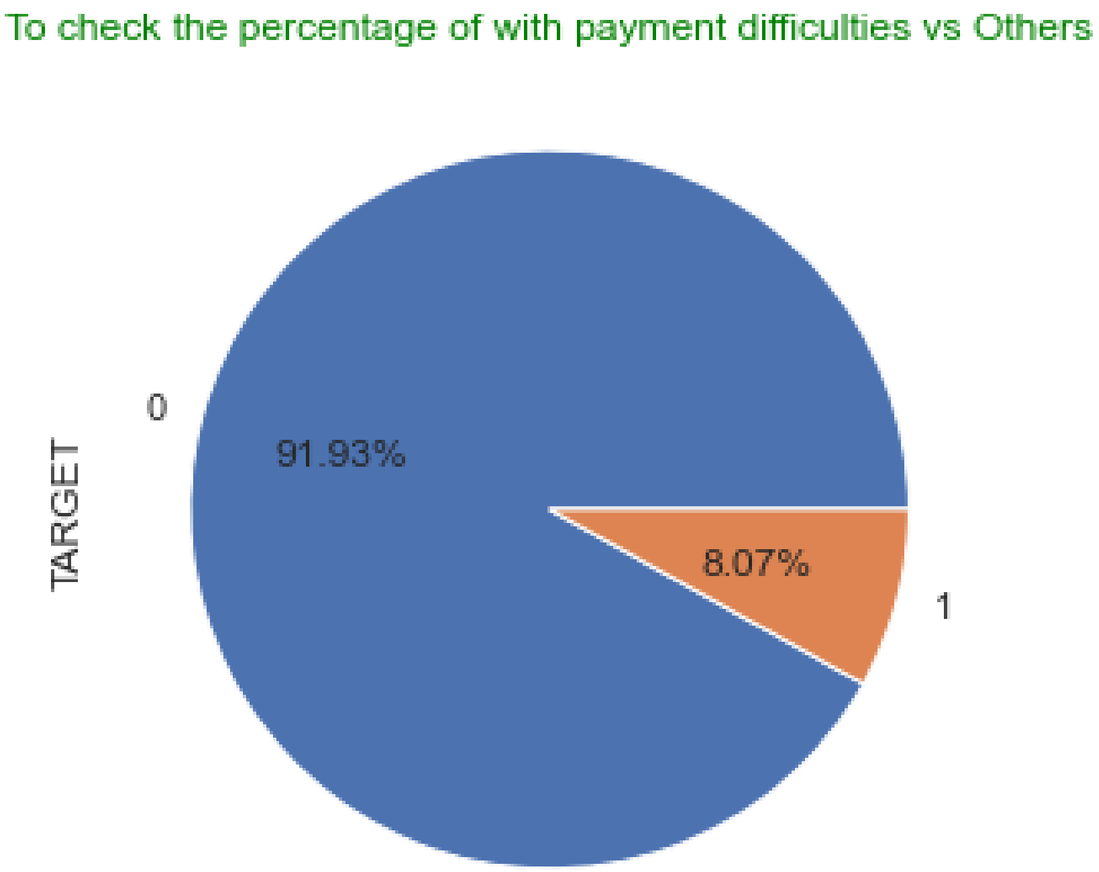
2) There are some columns where the value is mentioned as 'XNA' which means 'Not Available'. So we have to find the number of rows and columns and implement suitable techniques on them to fill those missing values or to delete them

TASK - 4

Analysis of Continuous variables and binning when required

our task is to analyze the continuous variable and creating the bins for the required categories-'Young','Adult', 'Middle_Aged', 'Senior_Citizen' and storing in a new variable called 'YEAR_BIRTH_BINNING'

Task - 5:
Data Analysis.



Checking data Imbalance

POINTS TO BE CONCLUDED:

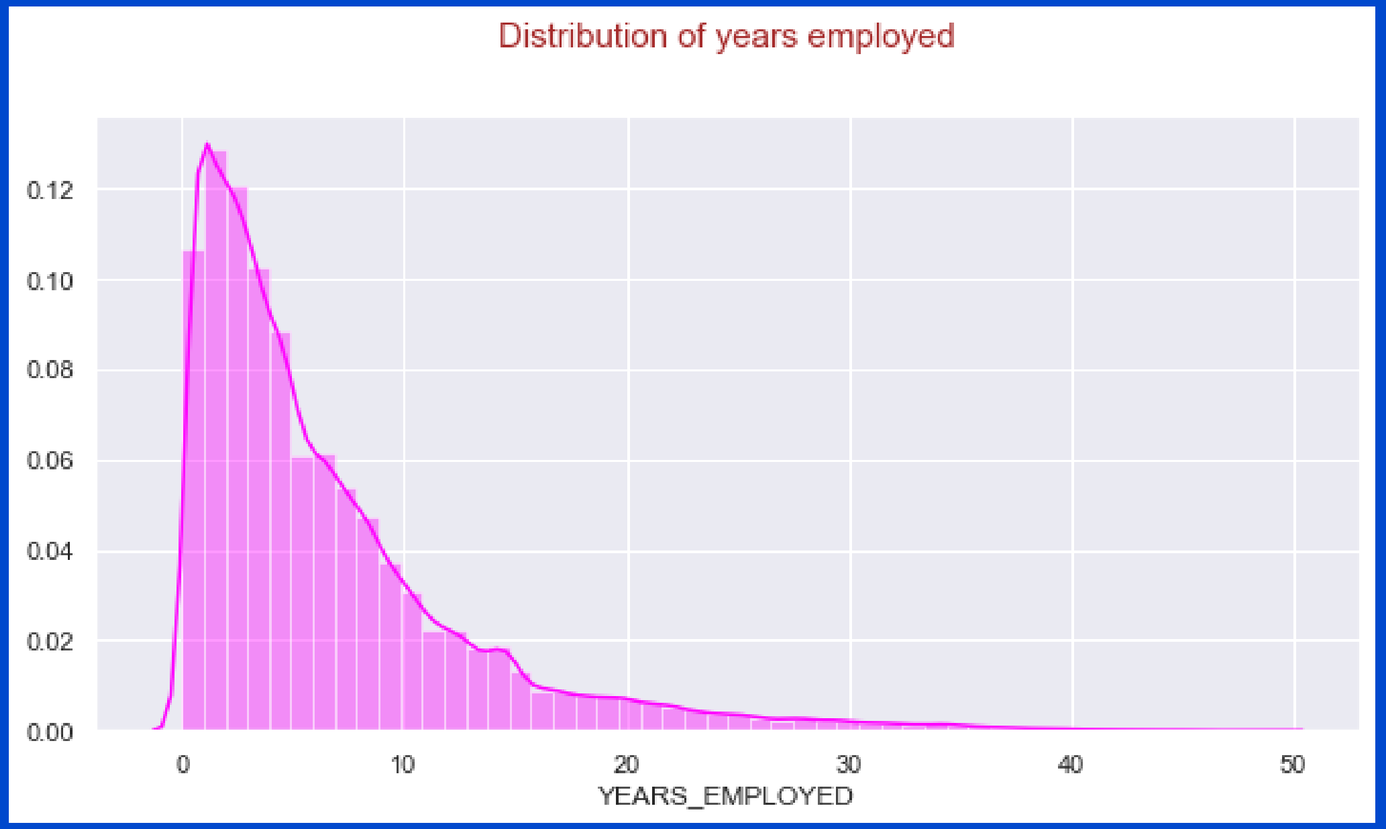
Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)

1)We can see from the above pie chart, the percentage of clients with payment difficulties is 8%

2)the percentage of clients with non-payment difficulties is 92%

1) Splitting Data with respect to TARGET=0 and TARGET=1
2)Checking distribution of important columns

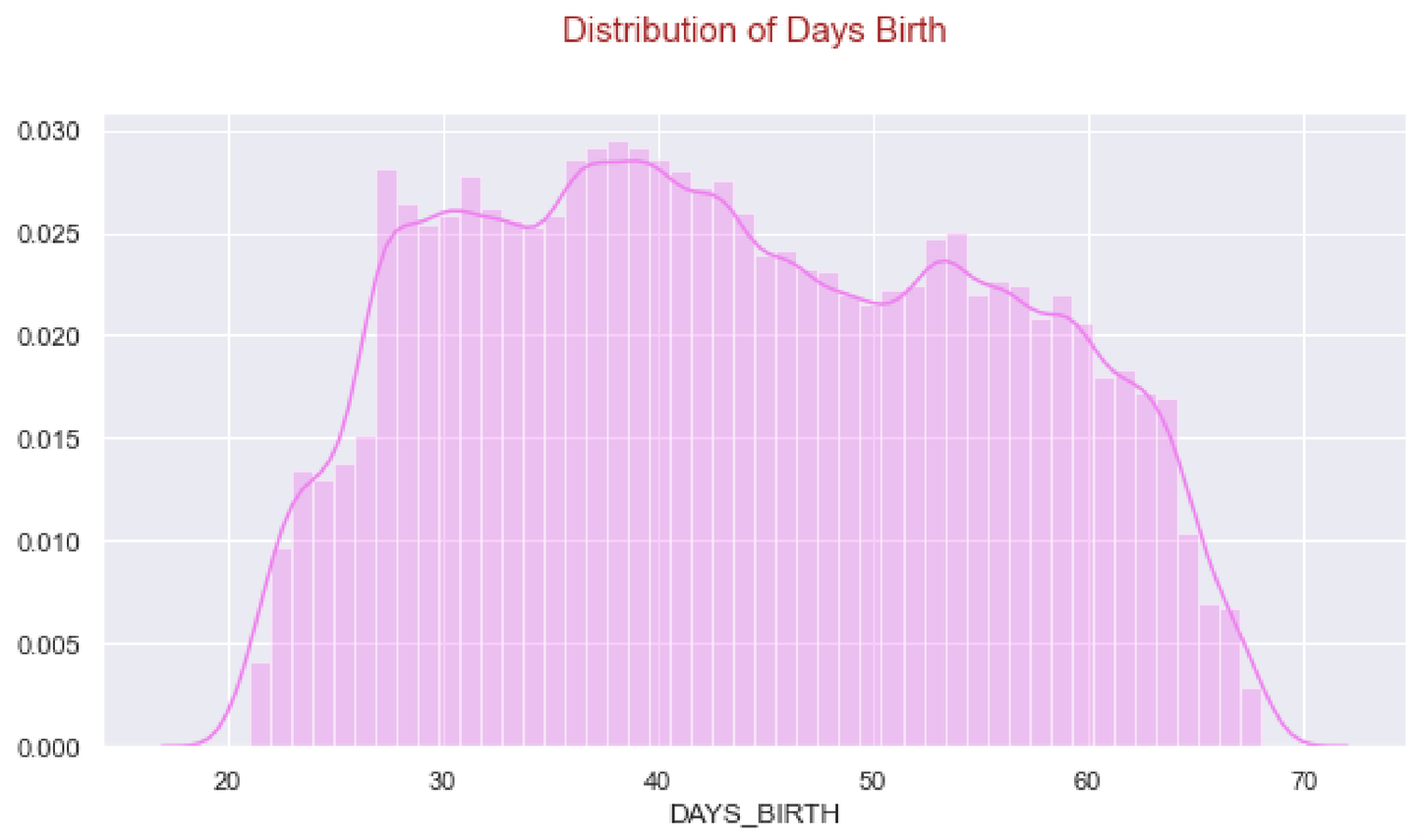
2.1)Checking Distribution of 'YEARS_EMPLOYED' column



Points to be concluded from the graph

1)We can observe that the column "YEARS_EMPLOYED" is normally distributed.
2)The experience/years employed ranges from 0 to 50 years

2.2) Checking Distribution of 'DAYS_BIRTH' column

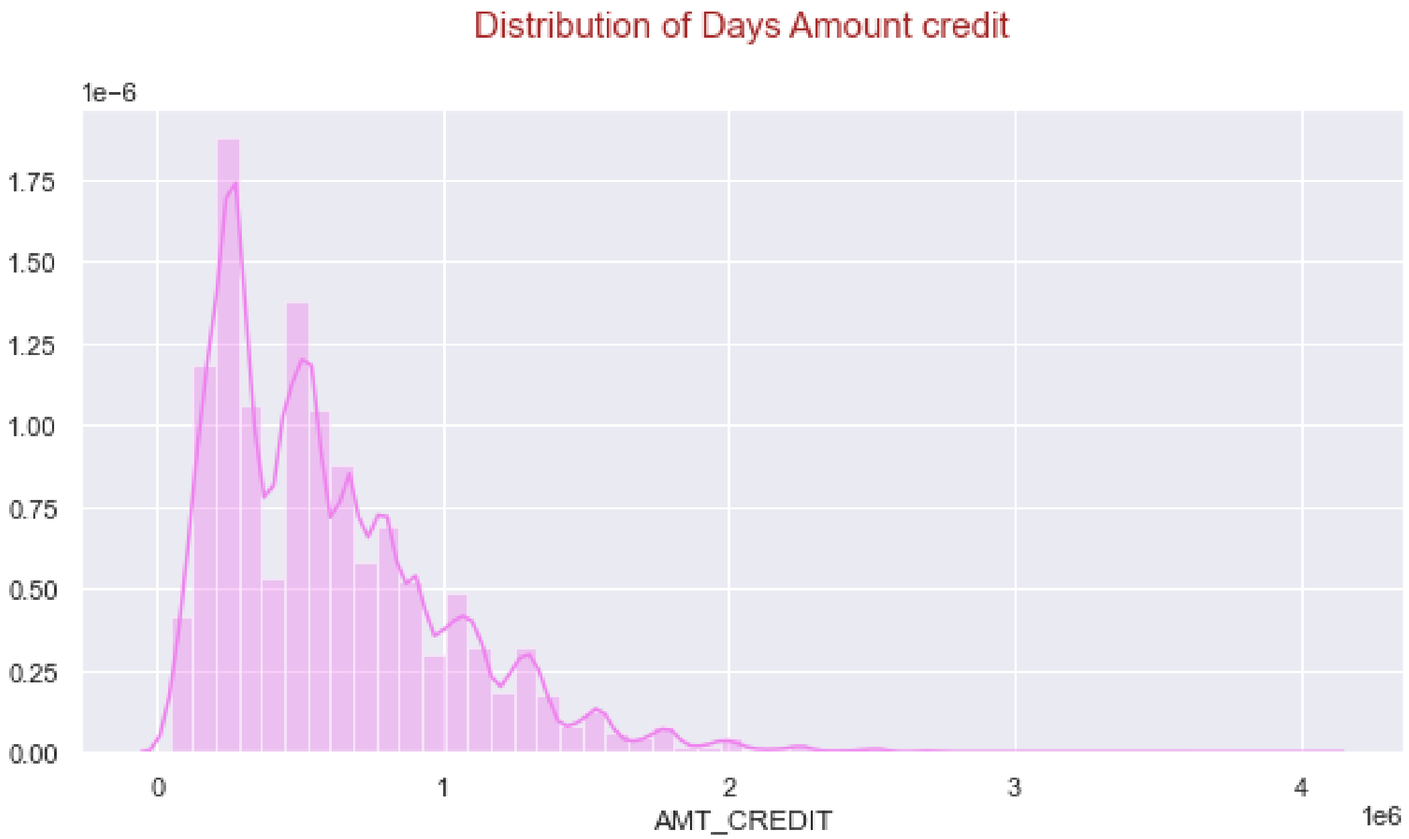


Points to be concluded from the graph

1)From the above graph, we can observe that, the column "DAYS_BIRTH" is normally distributed.

2)The age ranges from 25 to 70 years

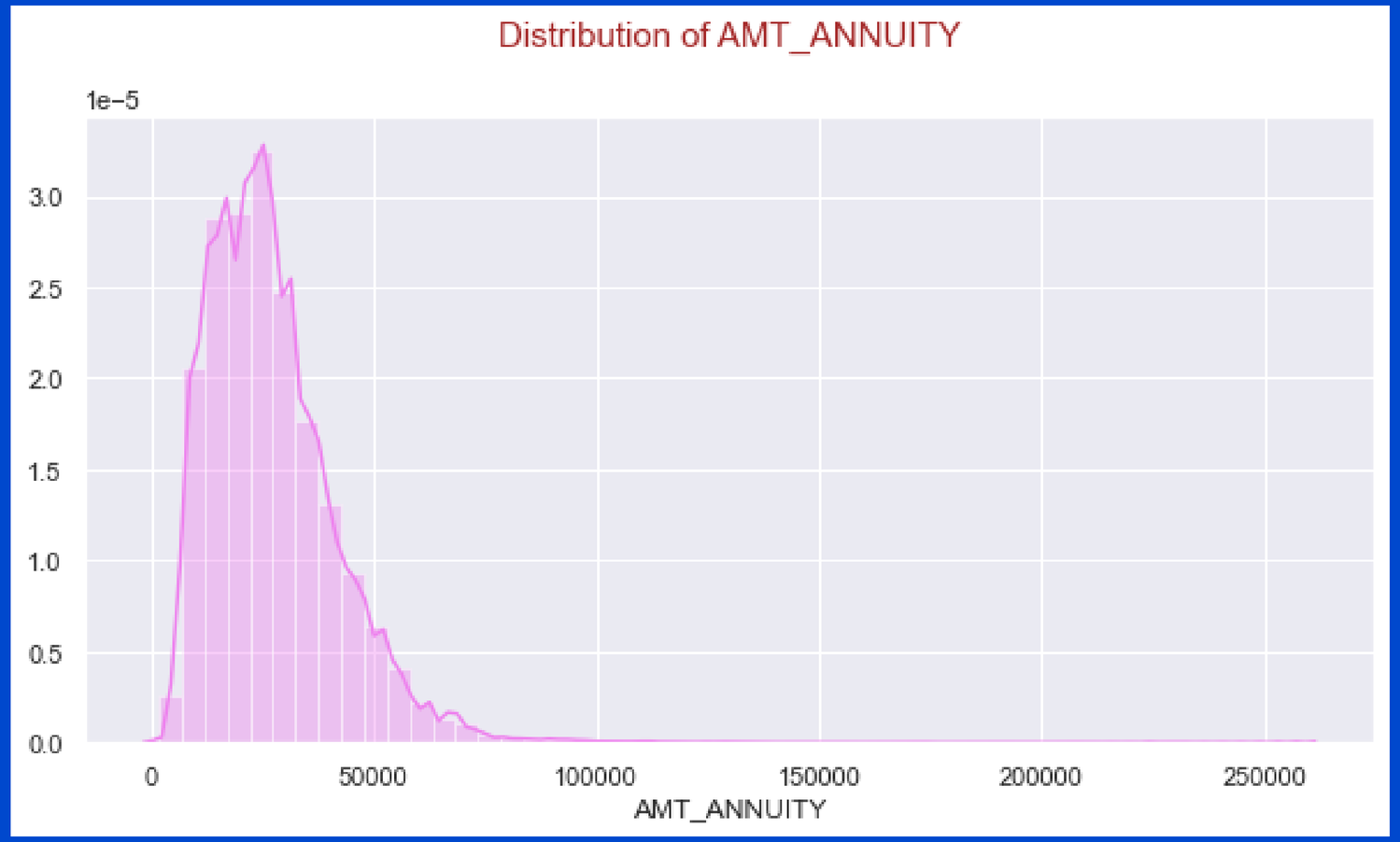
2.3) Checking Distribution of 'AMT_CREDIT' column



Points to be concluded from the graph

1)From the above graph, we can observe that, the column "AMT_CREDIT" distribution curve does not appear to be normal or bell curve.

2.4) Checking Distribution of 'AMT_ANNUITY' column

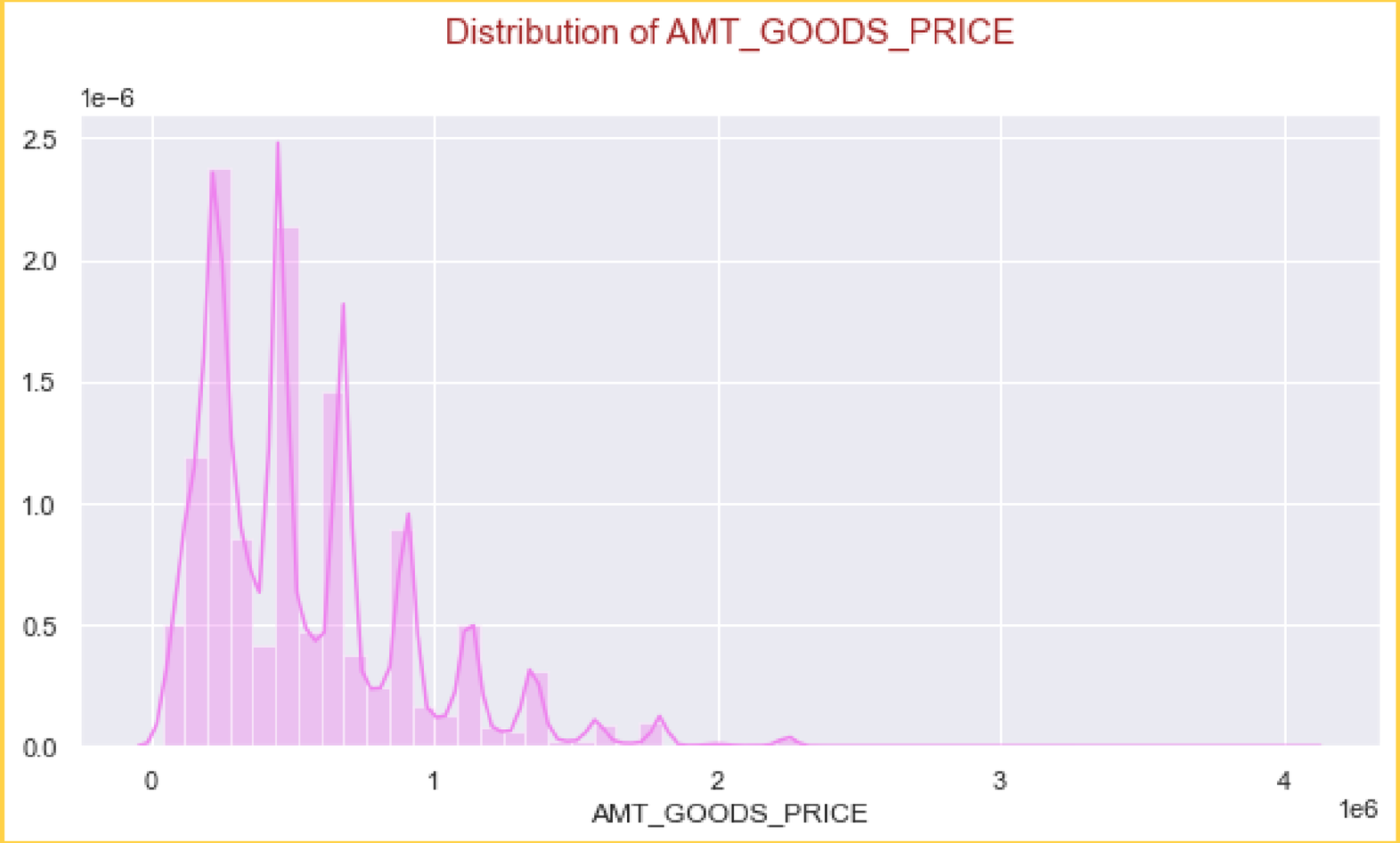


Points to be concluded from the graph

1)From the above graph, we can observe that, the column "AMT_ANNUITY" distribution curve appear to be normal

.2)And the curve is skewed to the right side of the graph

2.5) Checking Distribution of 'AMT_GOODS_PRICE' column

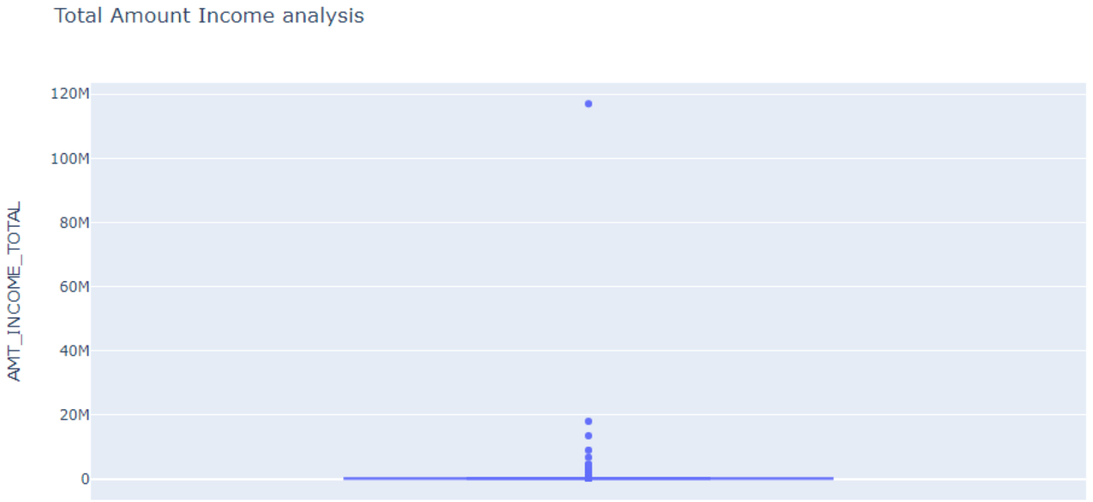


Points to be concluded from the graph

1)From the above graph, we can observe that, the column "AMT_GOODS_PRICE" distribution curve does not appear to be normal or bell curve.

2)However the curve is skewed to the right side of the graph

Analysis on coulmn AMT_INCOME_TOTAL to find outliers

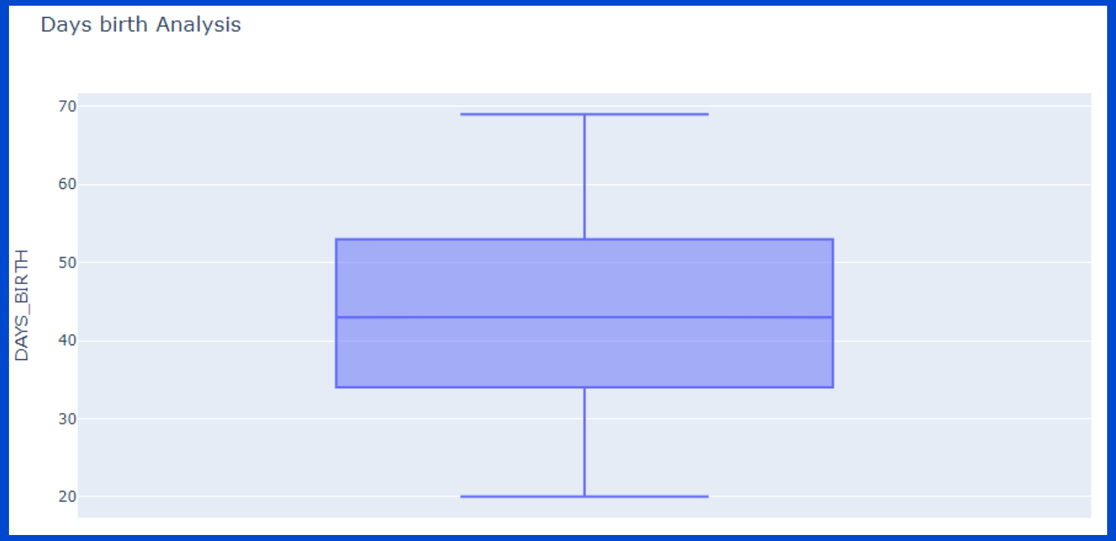


Some outliers are noticed in income amount. The third quartiles is very slim for income amount. We can conclude that, the amount 117M observed from the box plot is an outlier

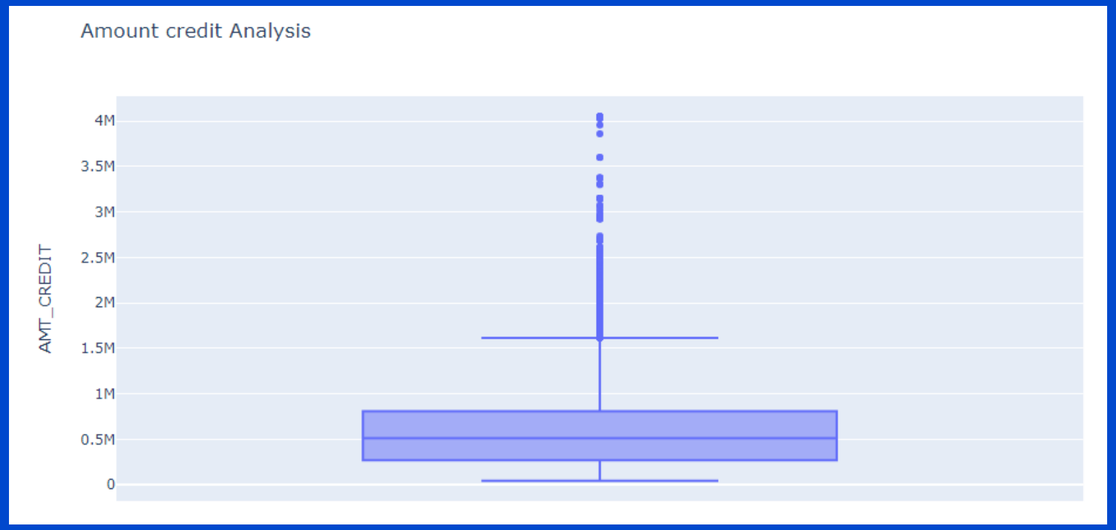
Here, in the coloumn 'DAYS_EMPLOYED' which tells how many days before the application the person started current employment.

We don't see any outliers in this case, as we have already handled a, having the value 1000.

(continued)

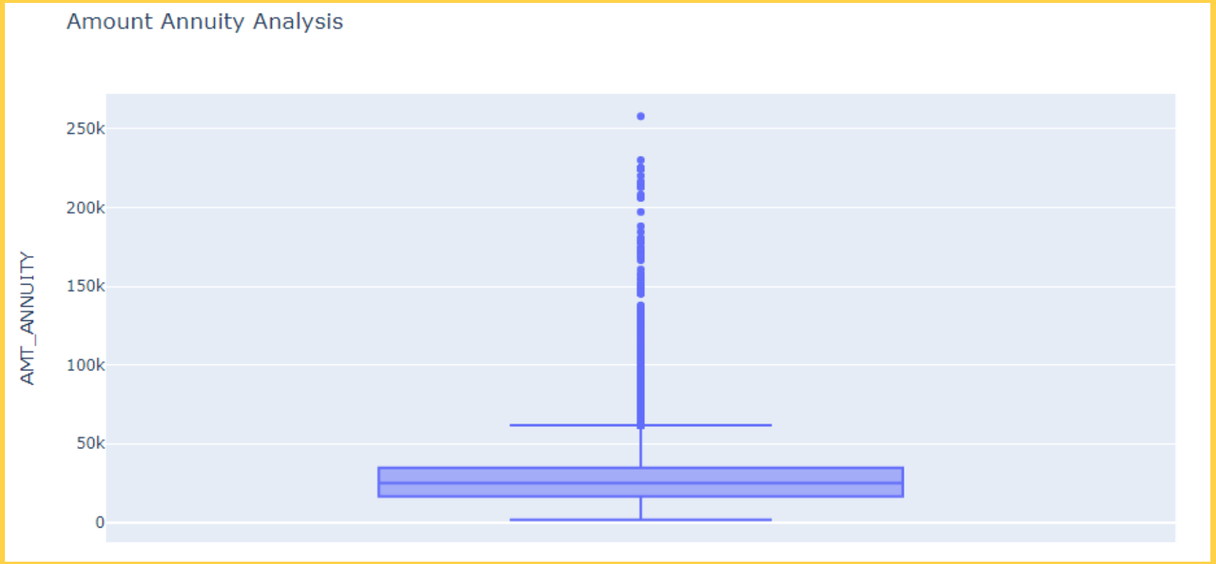


There is no outliers present in the age column.

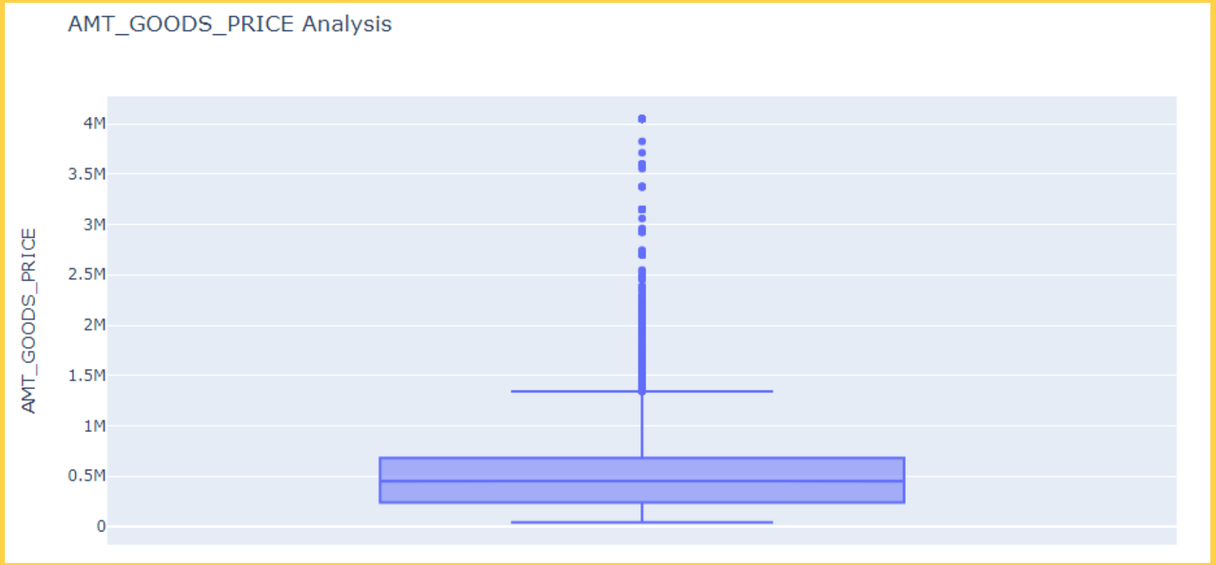


1) Some outliers are noticed in credit amount.

2) The first quartile is bigger than third quartile for credit amount which means most of the credits of clients are present in the first quartile.

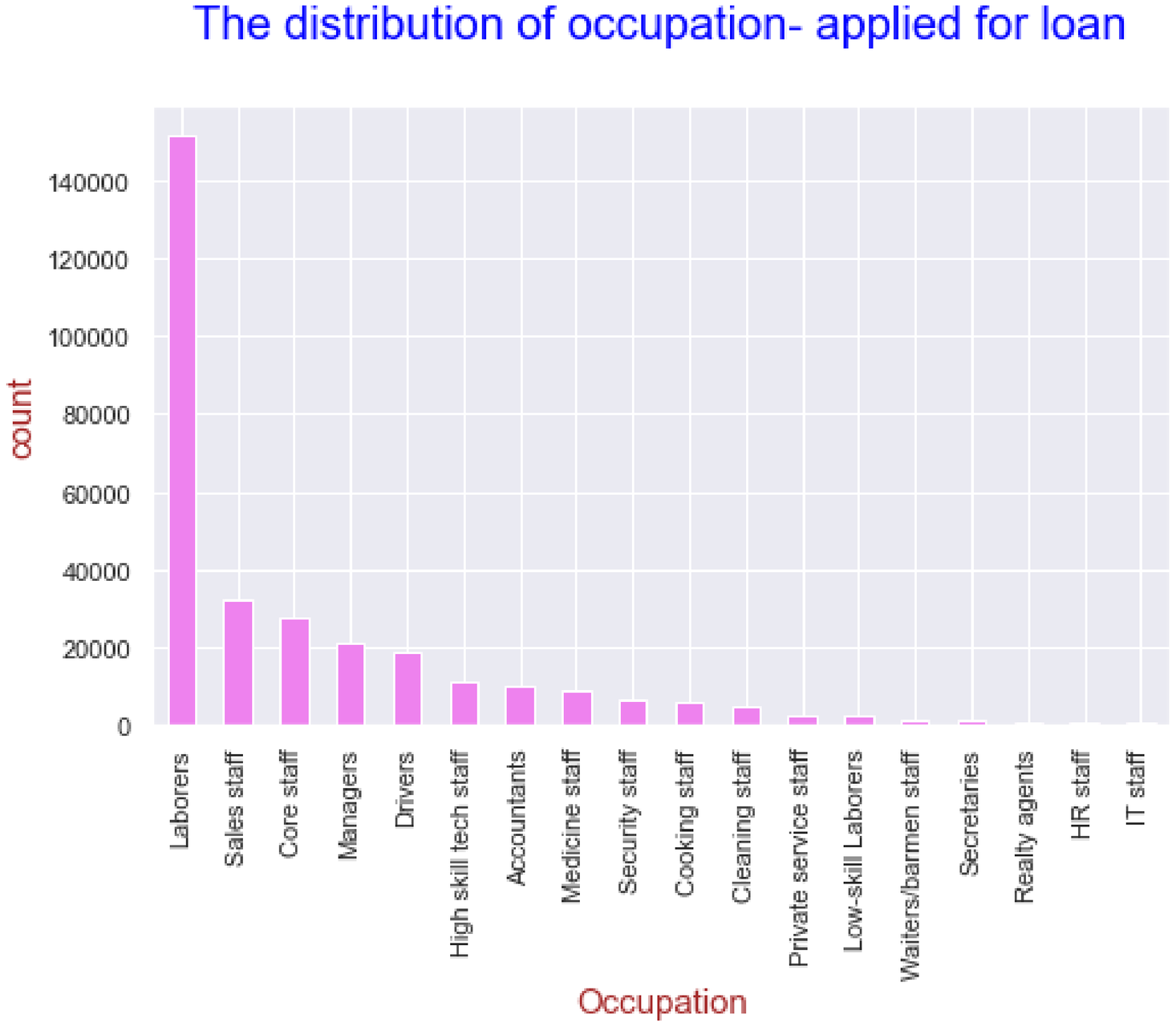


1) Some outliers are noticed in annuity amount.
2) The first quartile is bigger than third quartile for annuity amount which means most of the annuity clients are from first quartile.
3)The value above 258000 is an outlier here.



1)For consumer loans it is the price of the goods for which the loan is given.

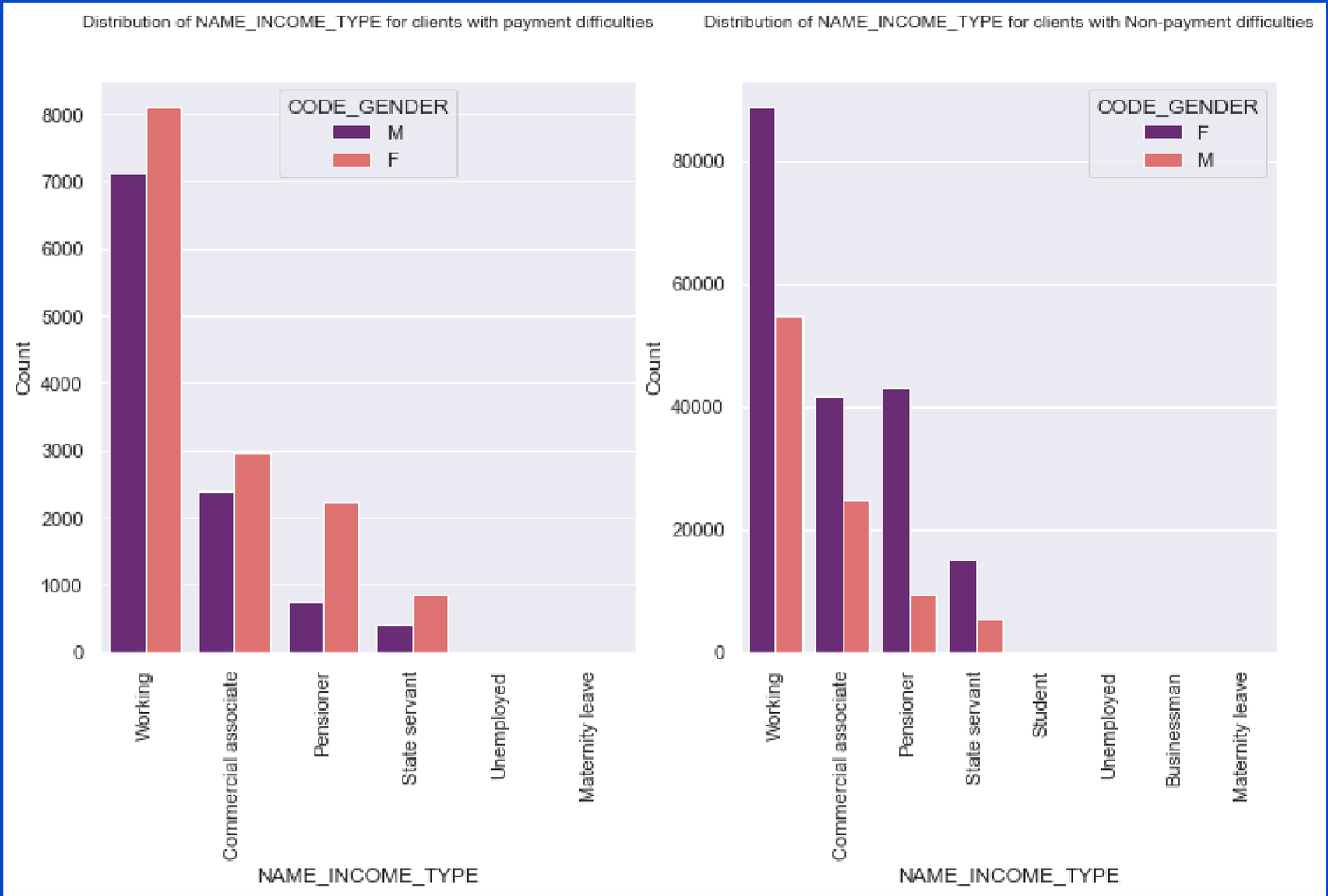
2)Some outliers are noticed in Goods price.
3) The first quartile is bigger than third quartile for goods amount.



Inference from the chart

1) 3 categories, 'Laborers', 'Sales staff', 'Core staffs' shows the major count, who applied
for loan.
2) IT Staffs are the least applied for the loan

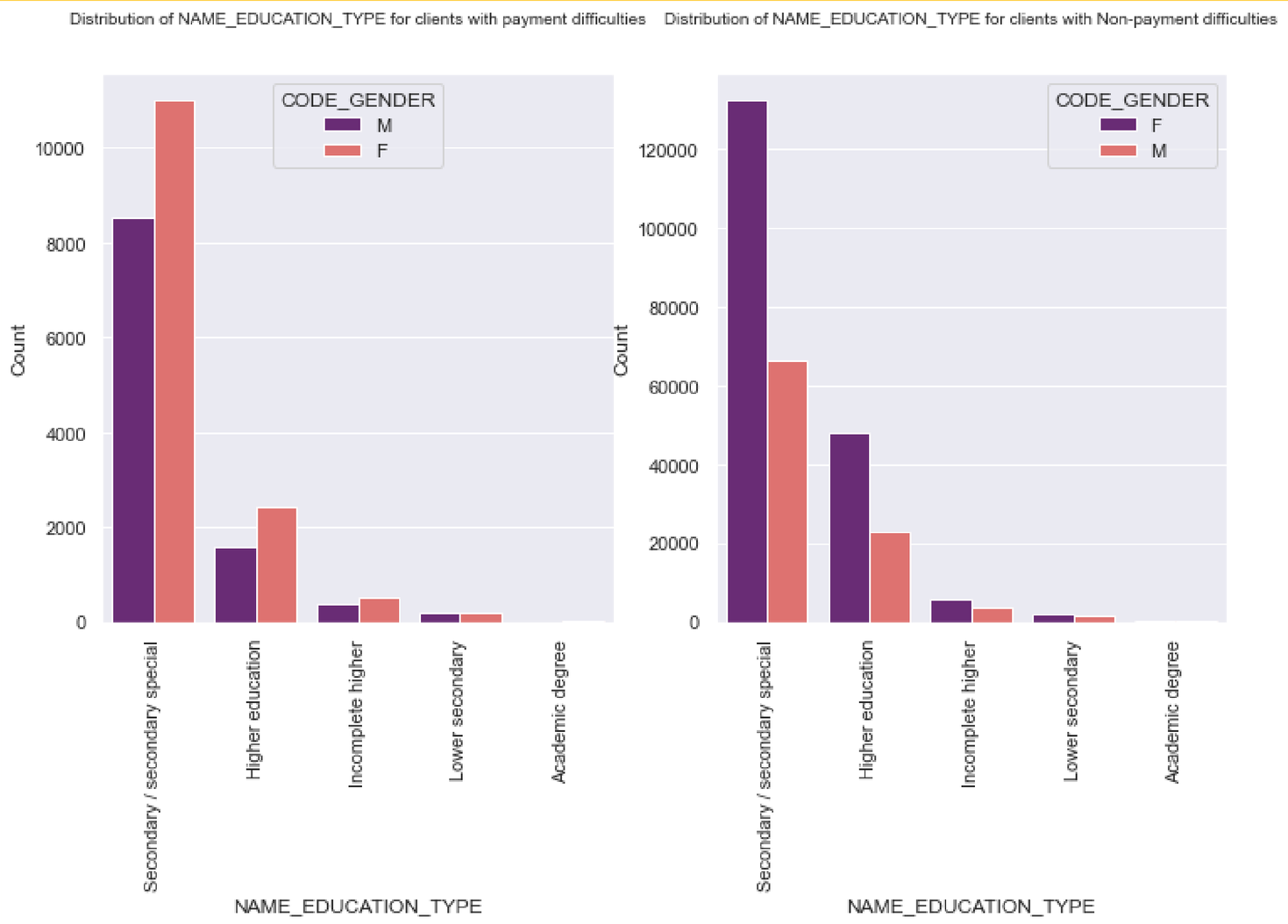
Univariate Analysis



1) Clients who are either at Maternity leave OR Unemployed have payment difficulties

2) For this, Females are having more number of credits than male.

3) Student and Businessmen have no defaults

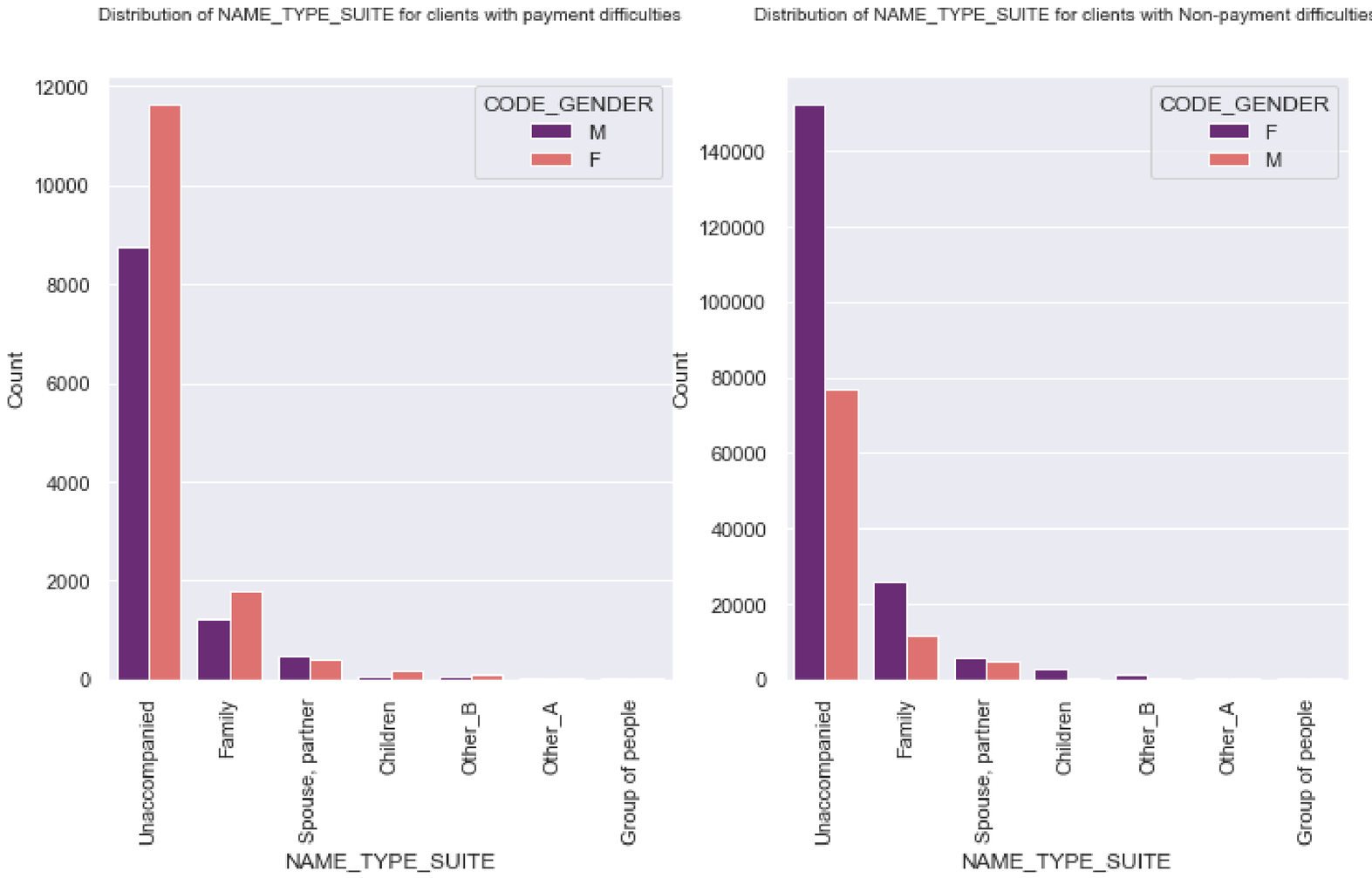


1) The count of Loan Payment Difficulties whose educational qualifications secondary/secondary special is higher compared to higher education, Incomplete higher.

2) And for those who has completed/studying Academic degree has no payments difficulties.

3) The Females are having more number of credits than male.

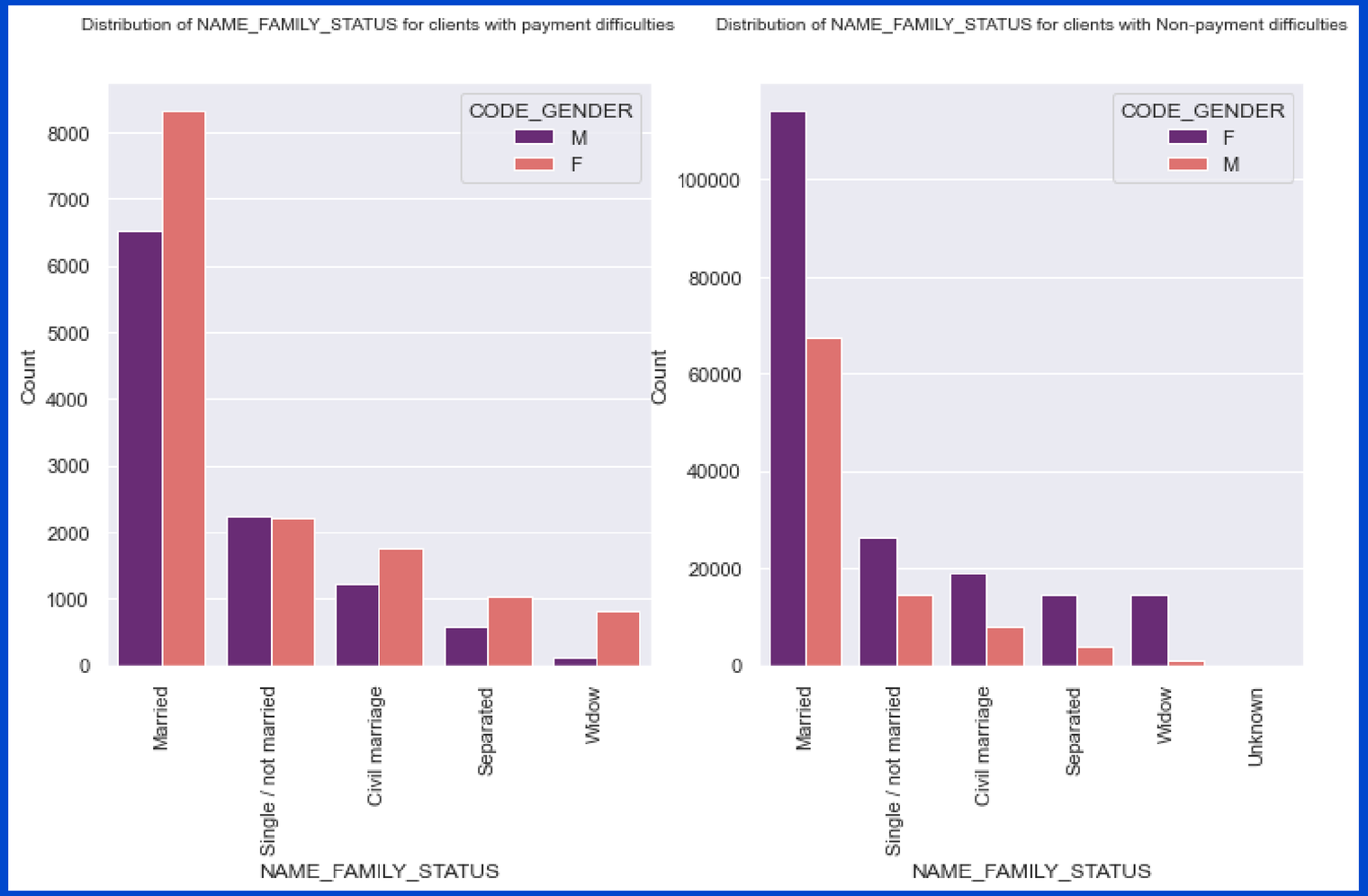
Ordinal variables



1) The count of Loan Payment Difficulties is higher for 'Unaccompanied' than rest other cases.

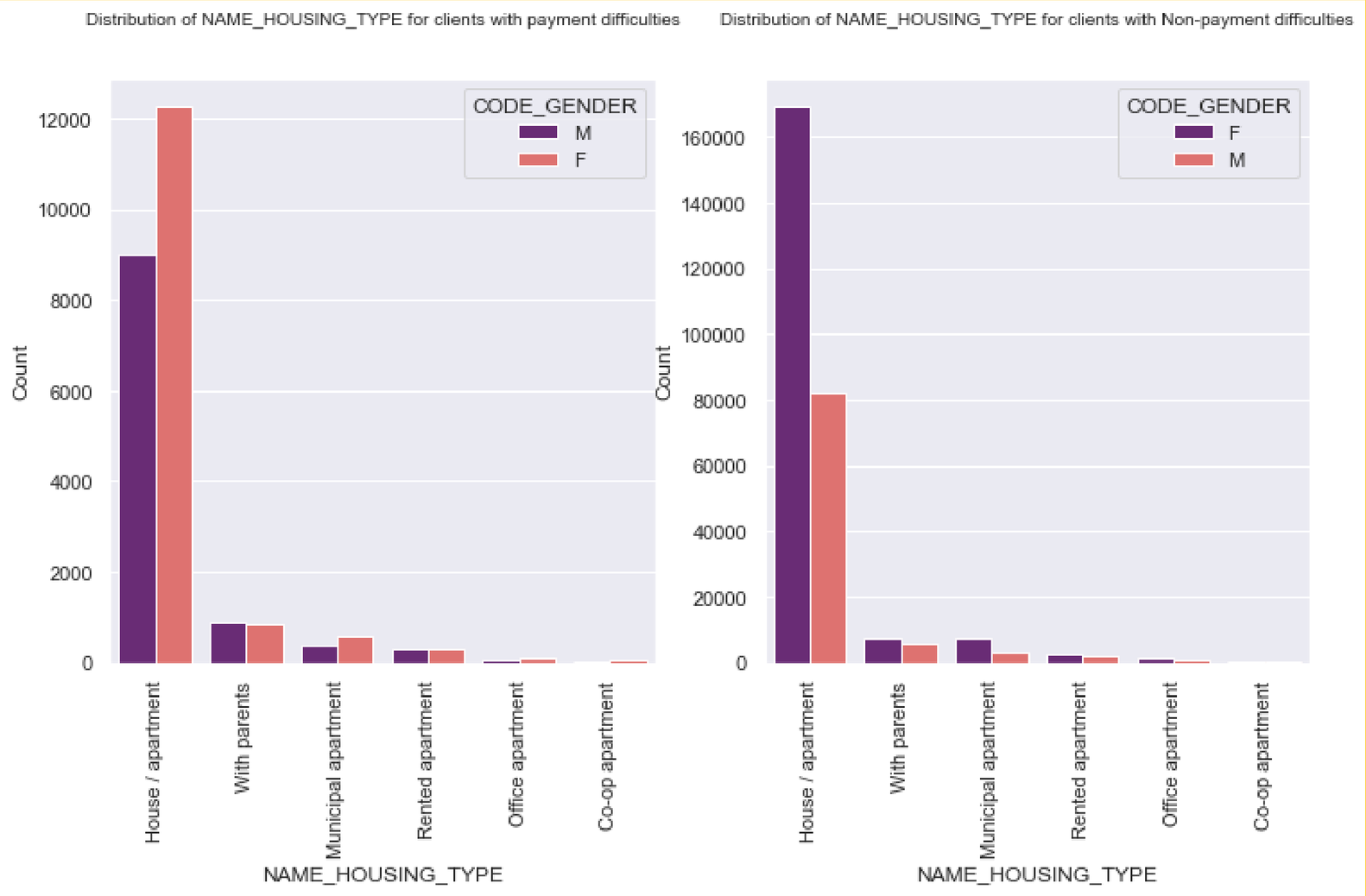
2) For the categories like Other_A and Group of prople has no payments difficulties.

3) The Females are having more number of credits than male



1) And decrease in count for separated and widow with Loan Payment Difficulties when comapred with the percentages from both the charts.

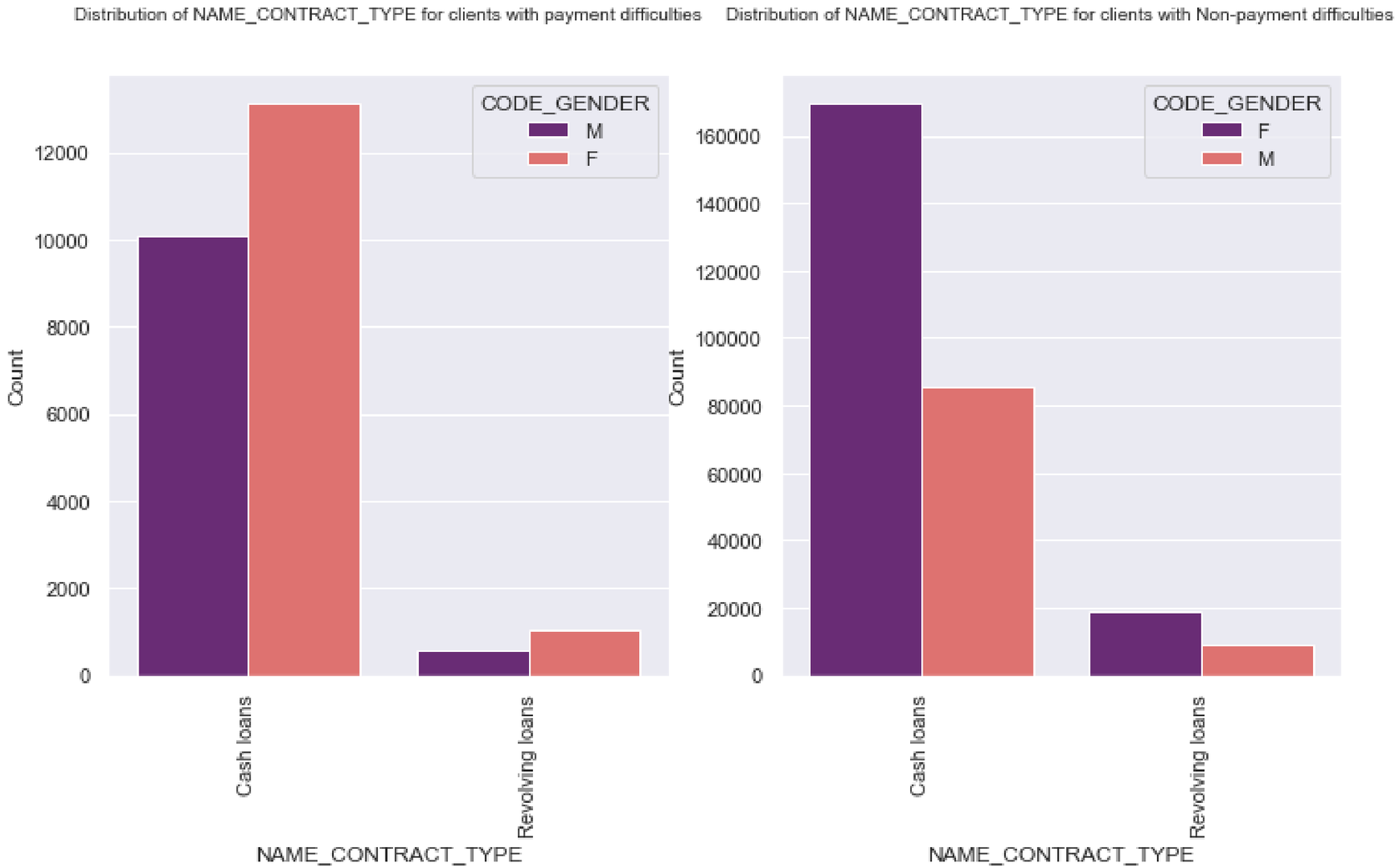
2) clients who have civil marriage or who are single default a lot.



1) We can conclude from the chart, that the Most people live in a House/Apartment

2) Ratio of People who live With Parents is more for defaulter than non-defaulters.

3) We infer that the applicants who live with parents have a higher chance of having payment difficulties.

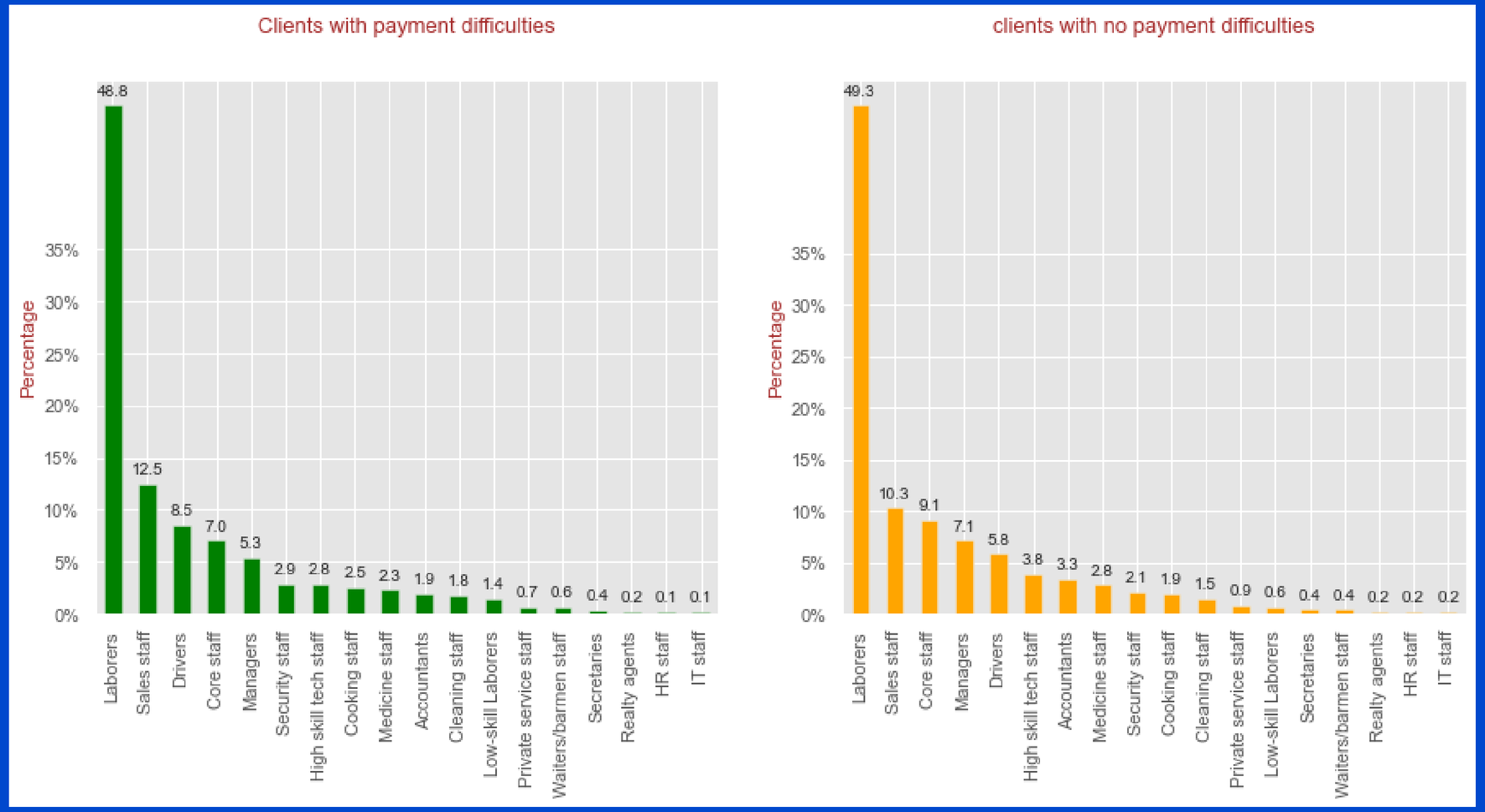


1) For contract type ‘cash loans’ is having higher number of credits than ‘Revolving loans’ contract type.

2) For this also Female is leading for applying credits.

3) For type 1 : there is only Female Revolving loans.

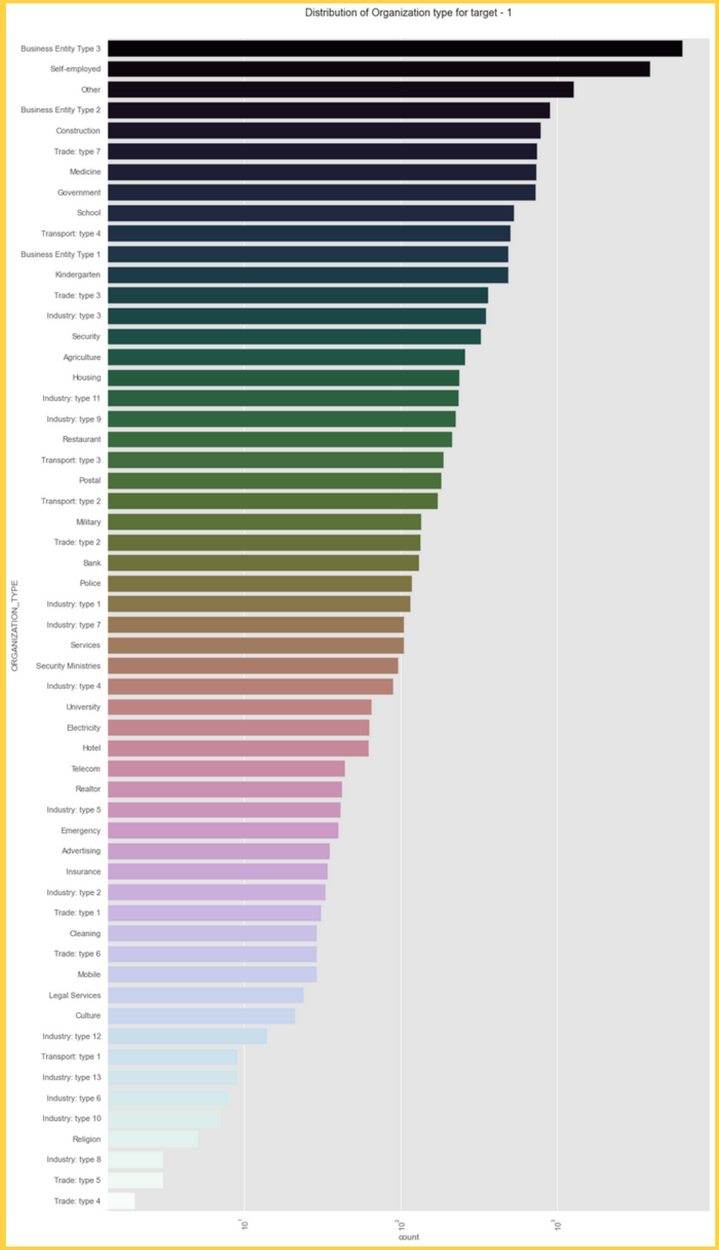
plot to see the distrubution of "OCCUPATION_TYPE" with payments difficulties vs Non-payments difficulties

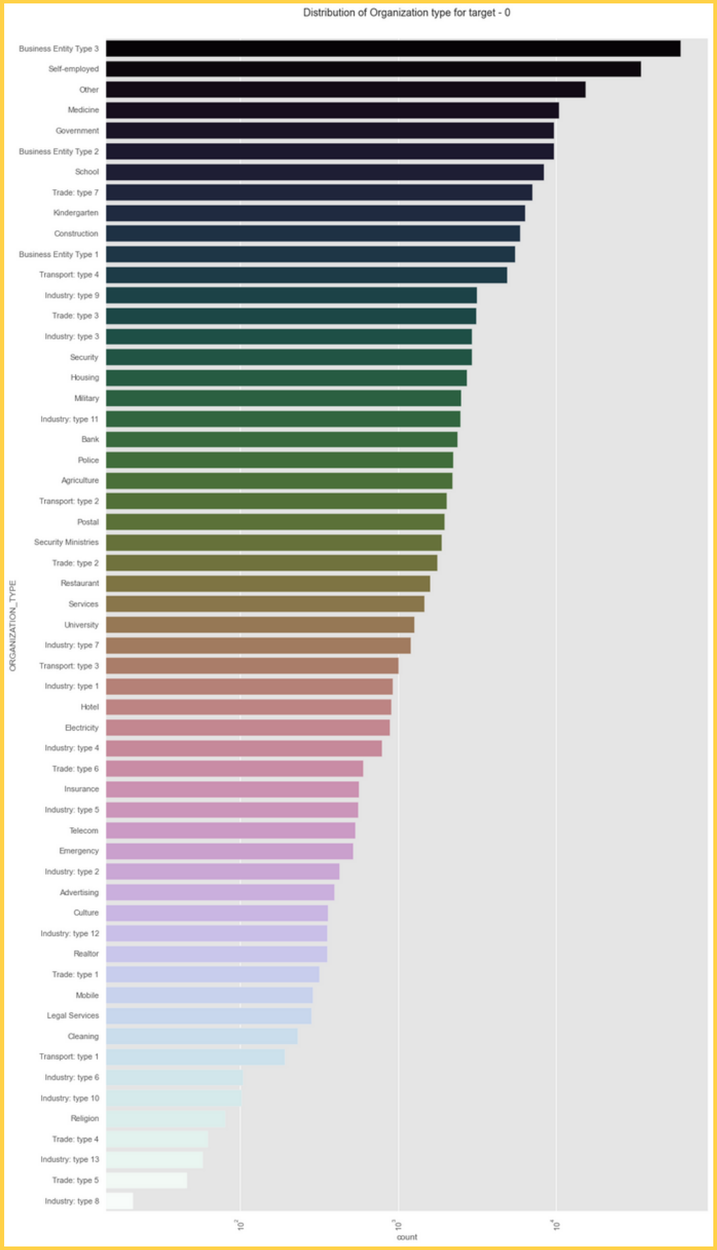


1) 3 categories, 'Laborers', 'Sales staff', 'Core staffs' shows the major count, who applied for loan are having the more percentage of payment difficulties.

2) IT Staffs are the least applied for the loan and non-defaulters in both the cases.

Plotting for Organization type in logarithmic scale



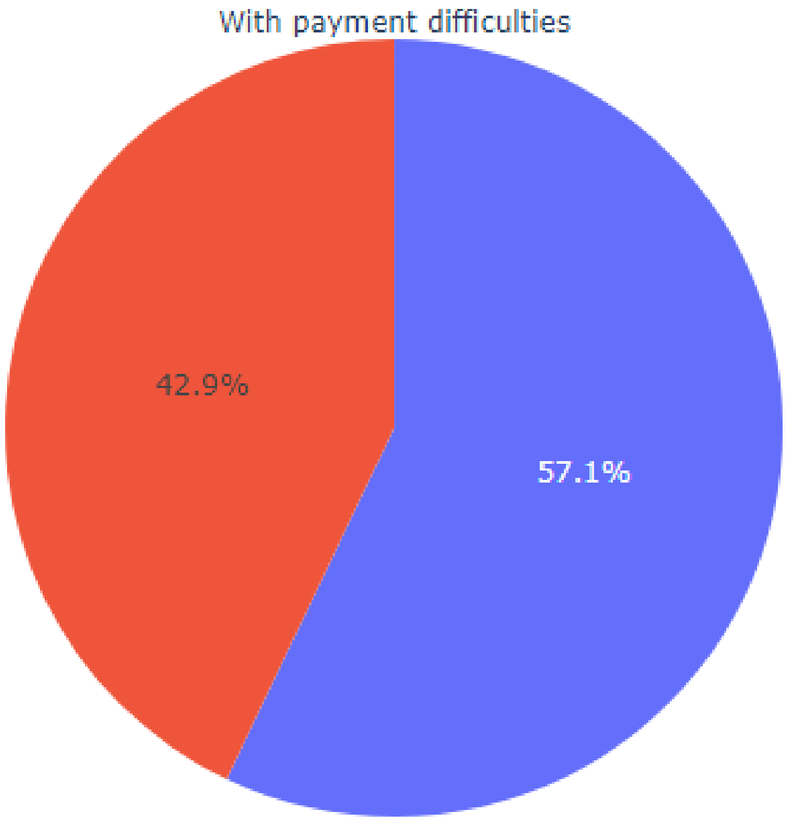


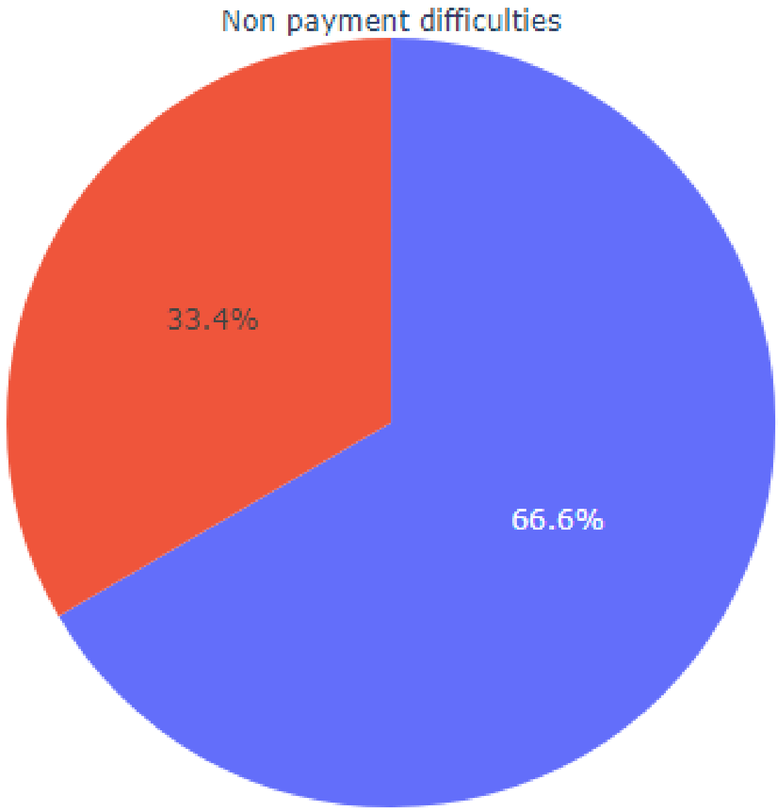
1) Clients which have applied for credits are from most of the organization type ‘Business entity Type 3’ , ‘Self employed’ , ‘Other’ , ‘Medicine’ and ‘Government’.

2) Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4.

3) Same as type 0 in distribution of organization type.

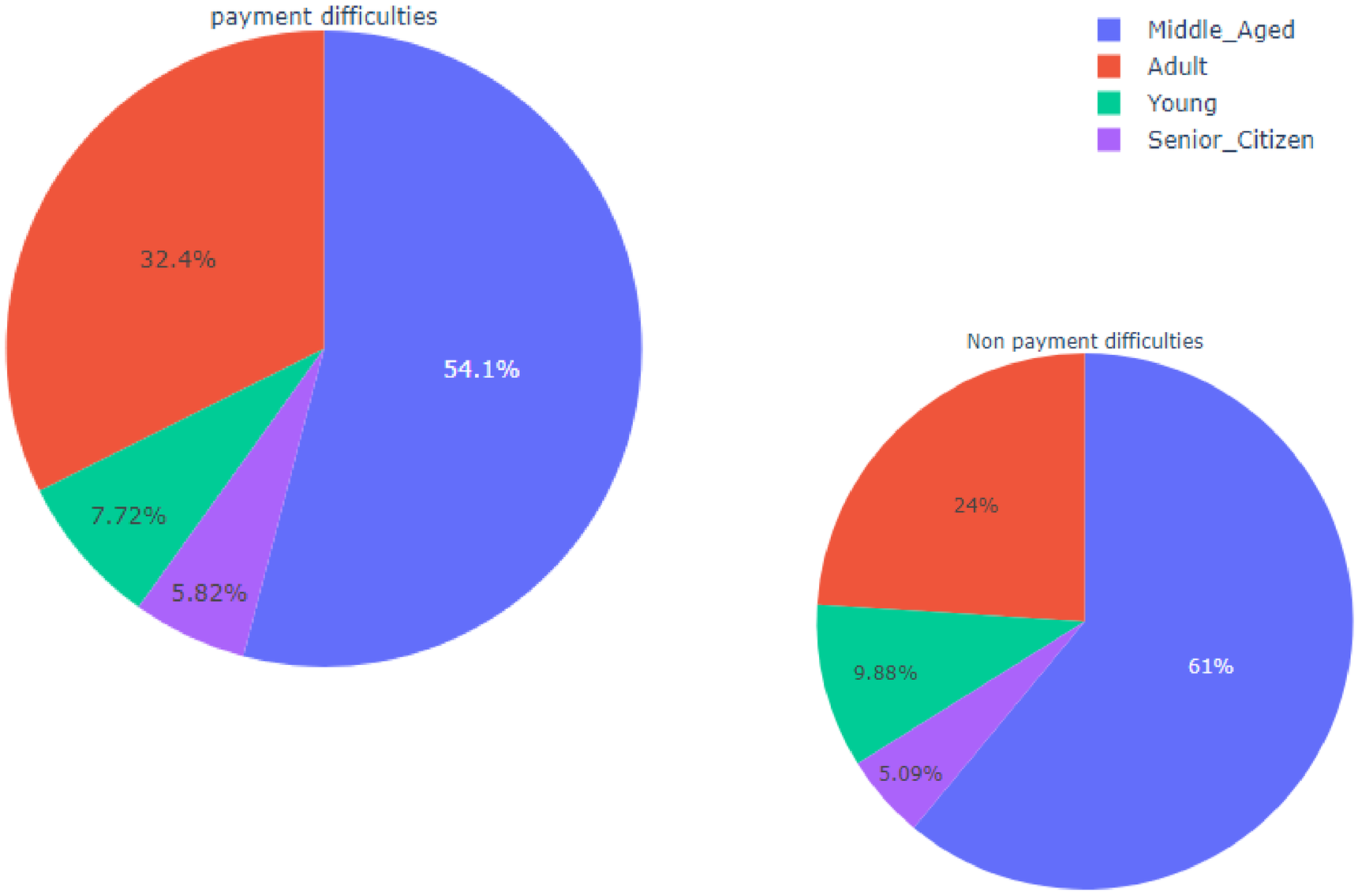
pie chart to see the % of "OCCUPATION_TYPE" with payments difficulties vs Non-payments difficulties





We can make an inference from the above chart, that the number of Females taking loans is much higher than the number of Males for both the cases.

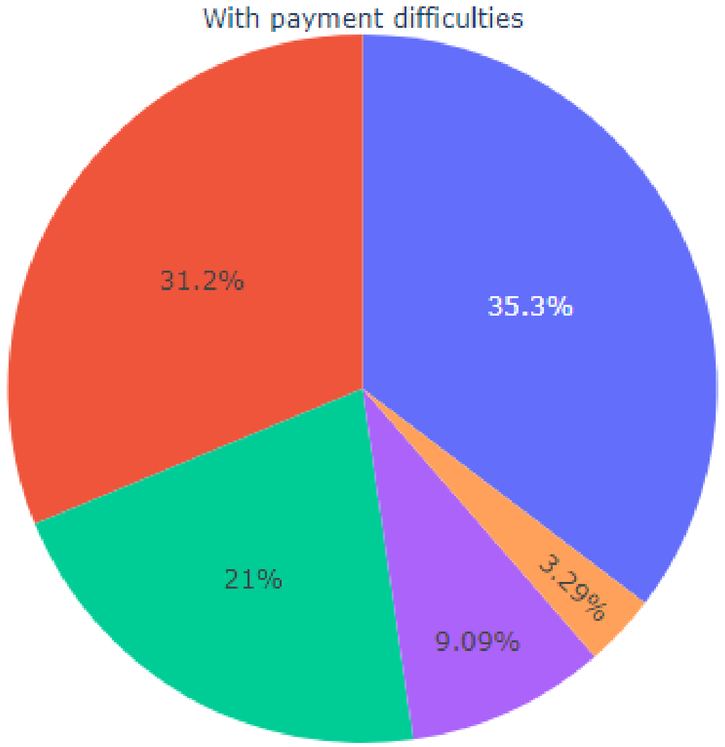
Discrete variable - Binned variables

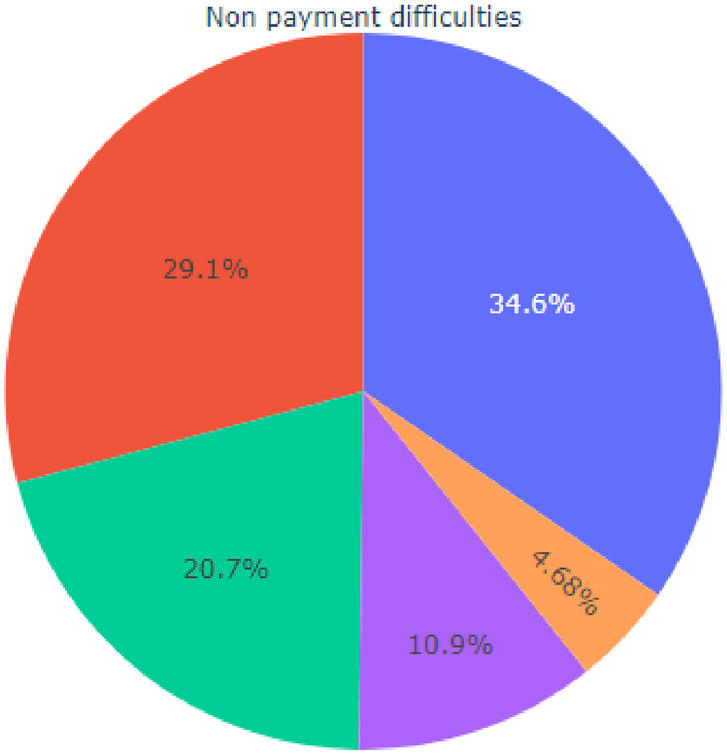


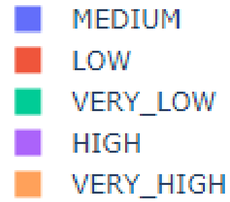
1) We can see from the graph, that there is an increase in the percentage of Loan Payment Difficulties who are young in age when compared to the percentages of Payment Difficulties and Loan-Non Payment Difficulties from 5% to 7 %

2) The same is applicabe to Adults also.

pie chart to see the % of "AMT_INCOME_TOTAL_RANGE" with payments difficulties vs Non-payments difficulties

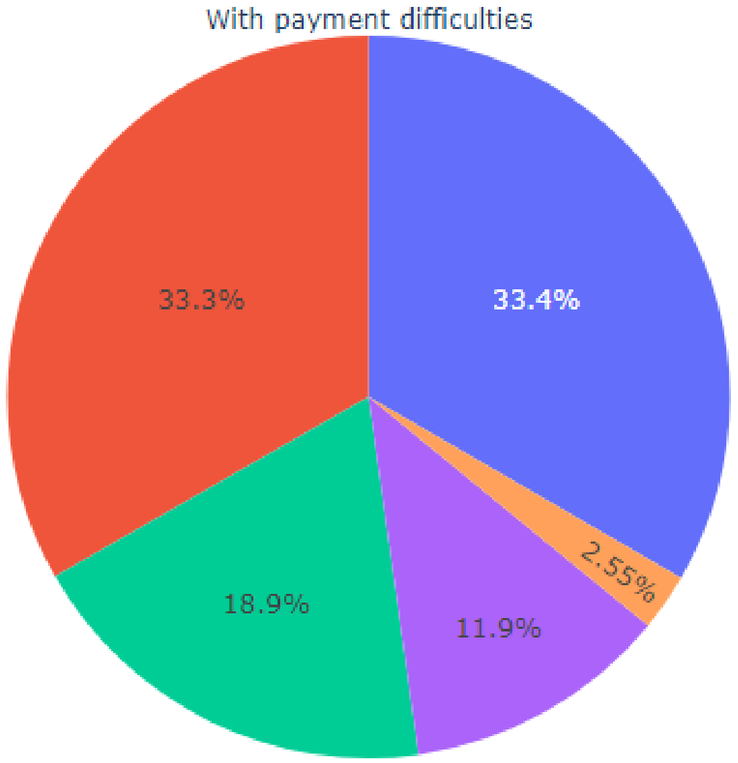


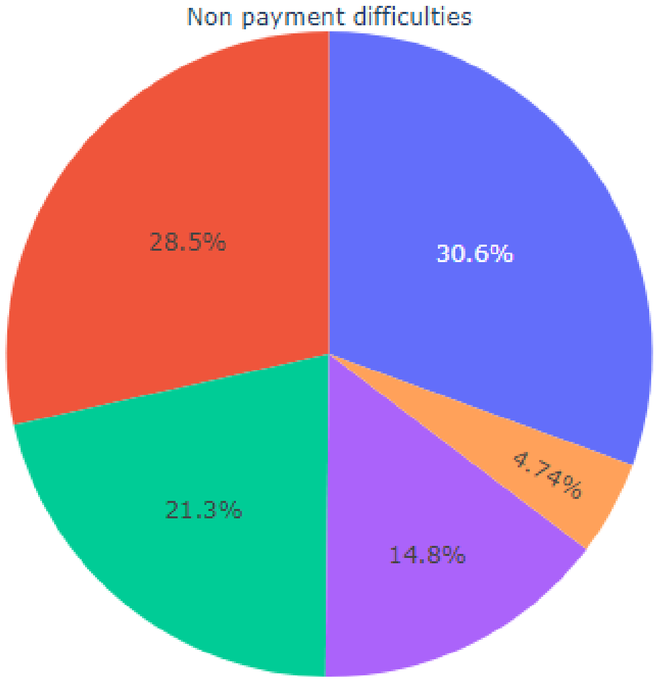


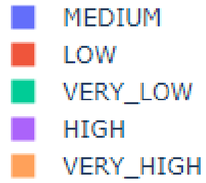


We can see from the graph, that there is an increase in the percentage of Loan Payment Difficulties for 'Medium' and 'Low' Income when compared to the other case.

pie chart to see the % of "AMT_CREDIT_RANGE" with payments difficulties vs Non-payments difficulties

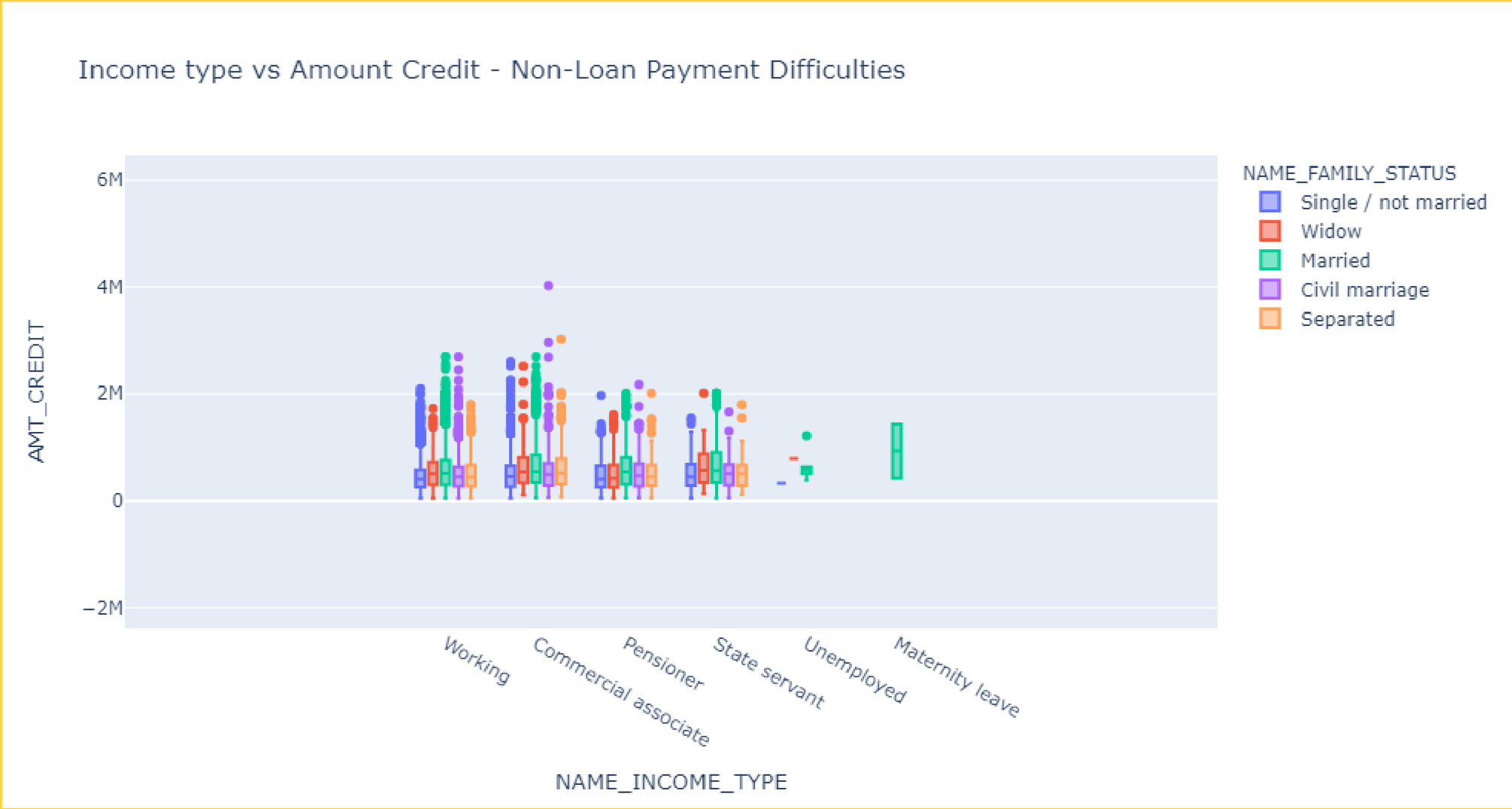




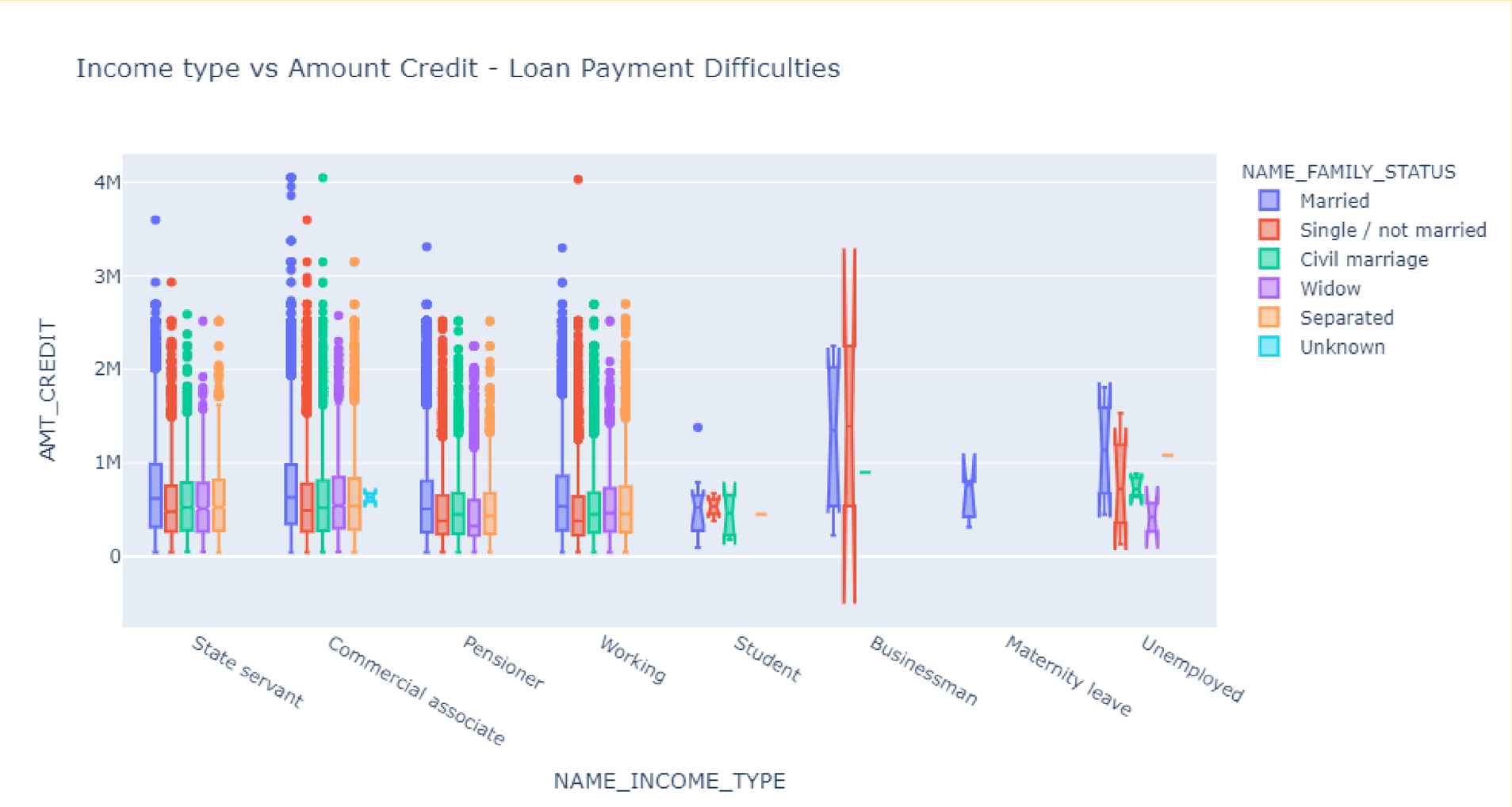


We can see from the graph, that there is an increase in the percentage of Loan Payment Difficulties for 'Medium' and 'Low' credit when compared to the other case.

Bivariate Analysis of Categorical vs Numerical Variables



Non-Payment difficulties - NAME_INCOME_TYPE vs AMT_CREDIT



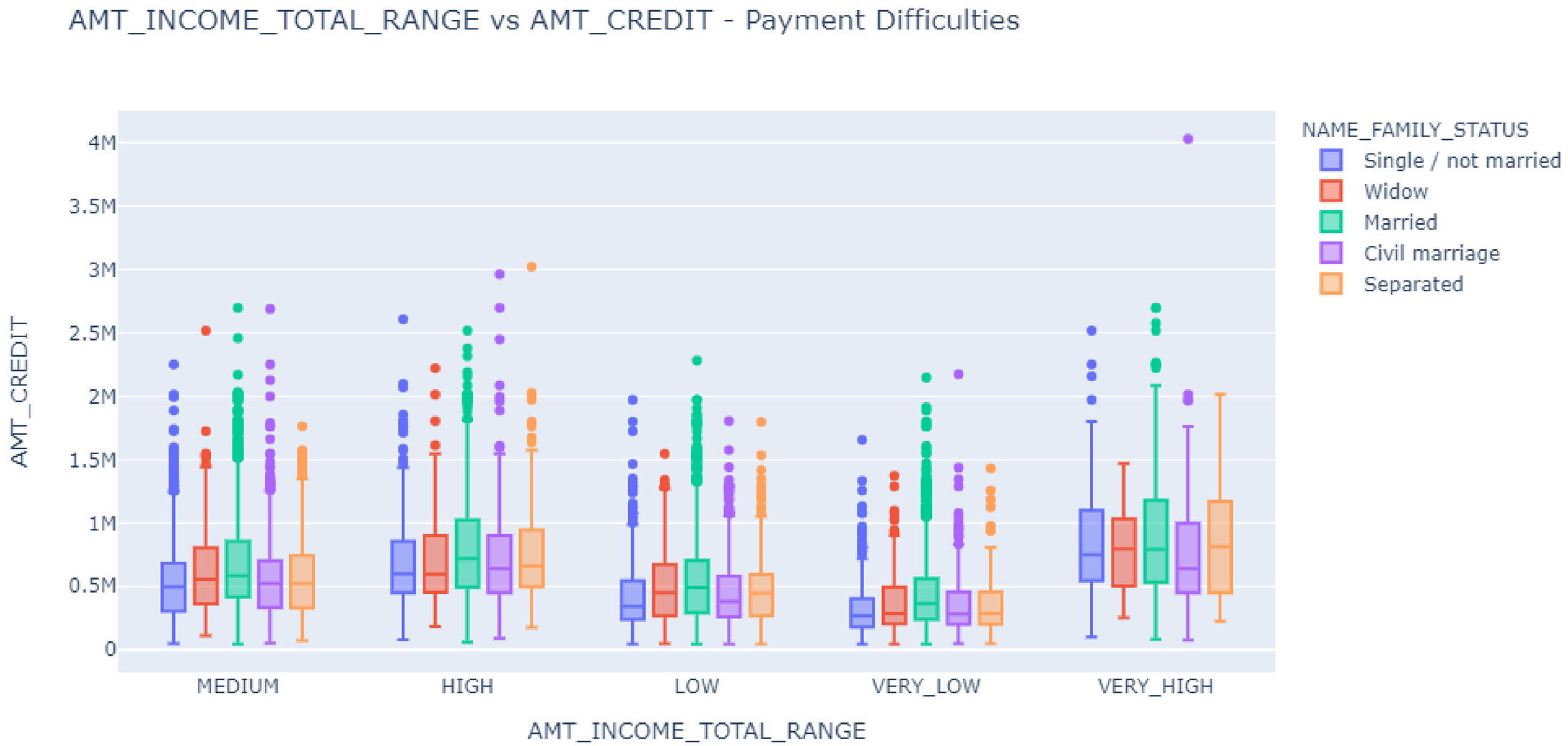
Inference from the above plot

1) The plot for Loan Payment/non-Payment looks almost similar

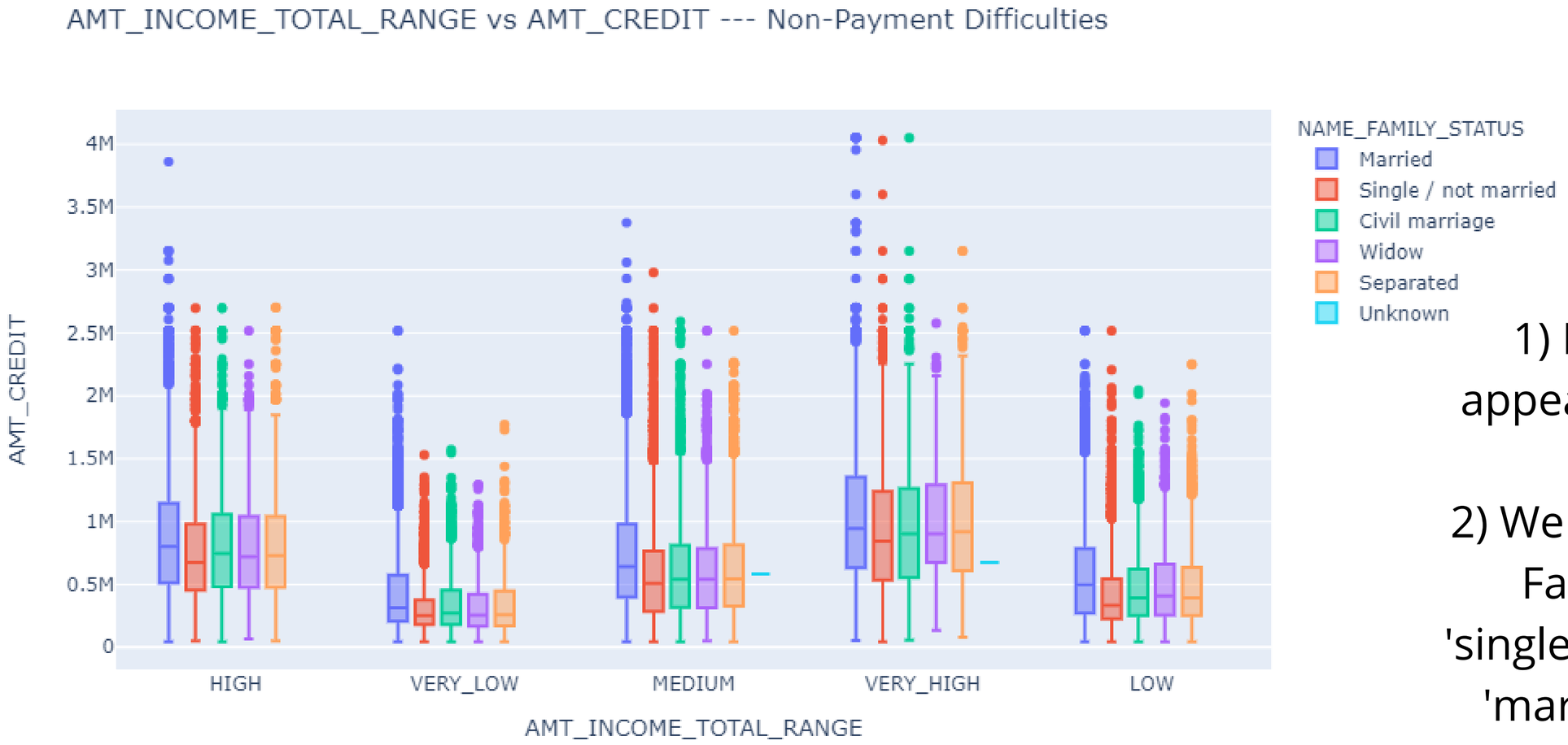
2) The categories like 'Pensioner' and 'State Service' have credits decrease, which means they have low payment difficulties.

3) We can also notice there are outliers present and for the commercial Associate the outliers value is very huge.

Payment difficulties - AMT_INCOME_TOTAL_RANGE vs AMT_CREDIT



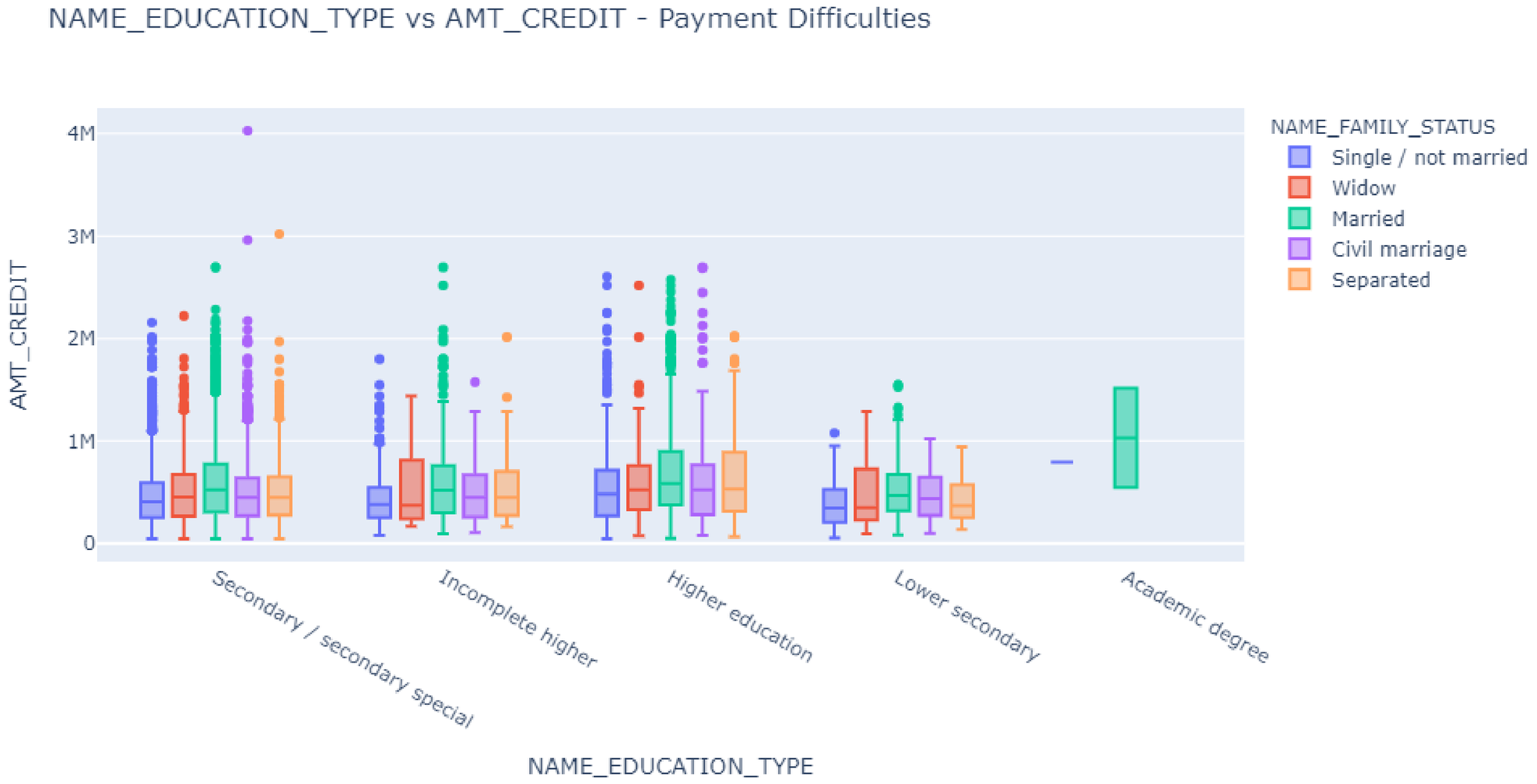
Non- Payment difficulties - AMT_INCOME_TOTAL_RANGE vs AMT_CREDIT



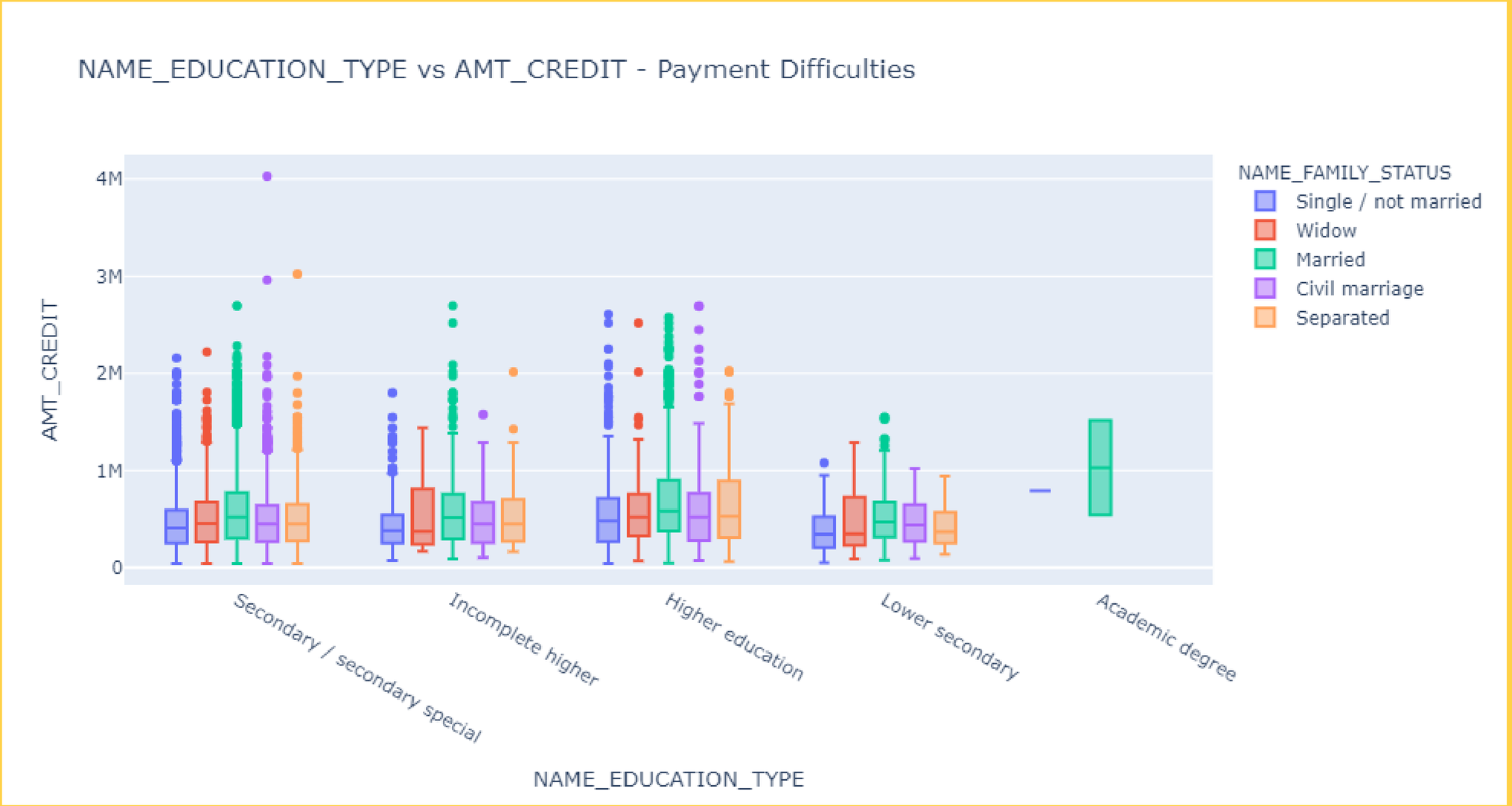
1) Both the plots appears to be similar.

2) We can see that, the Family status of 'single', 'seperated' and 'married' of income range very-high are having higher number of credits than others.

Payment difficulties - NAME_EDUCATION_TYPE vs AMT_CREDIT



Non- Payment difficulties - NAME_EDUCATION_TYPE vs AMT_CREDIT



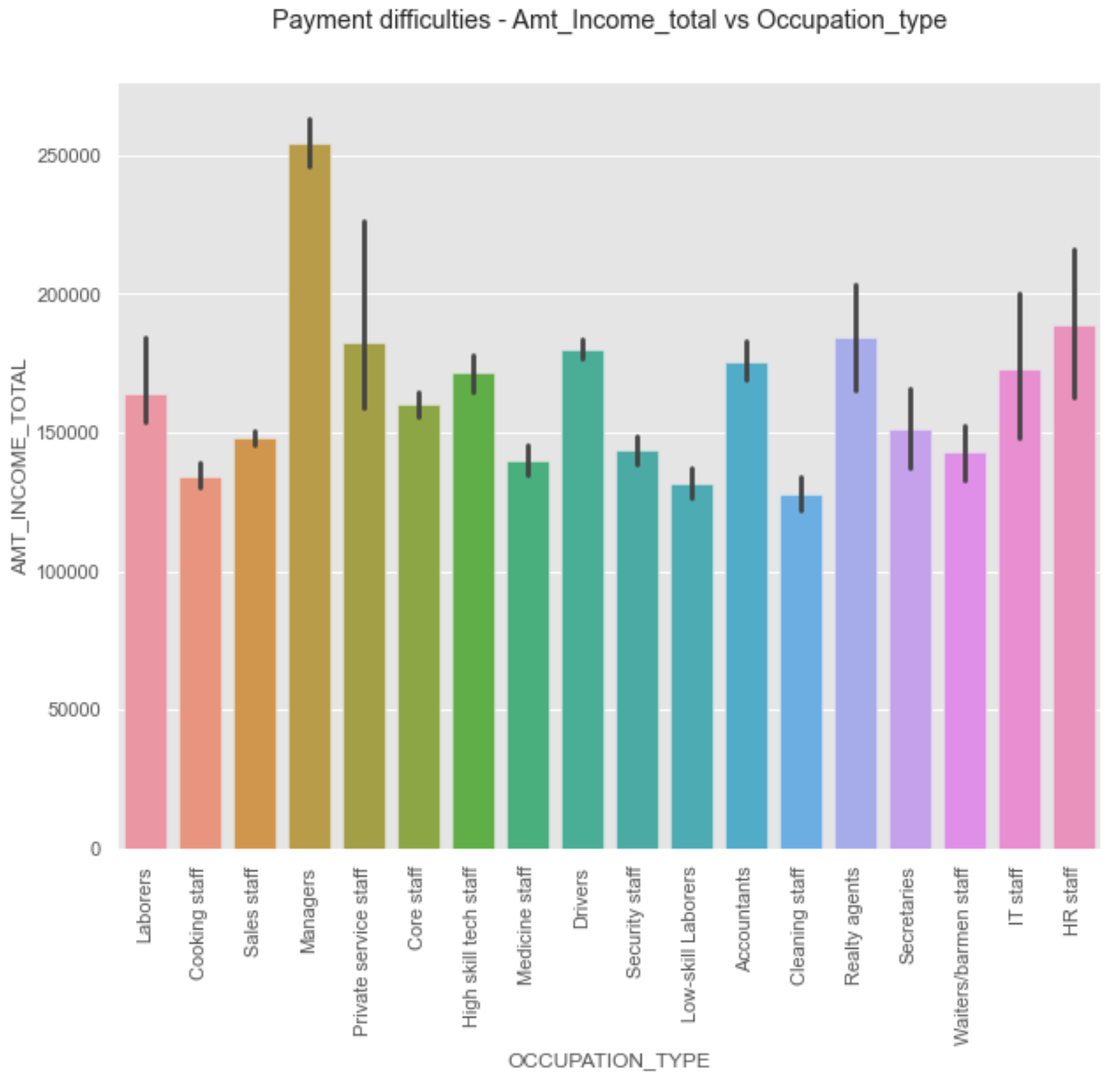
Inference from the above chart

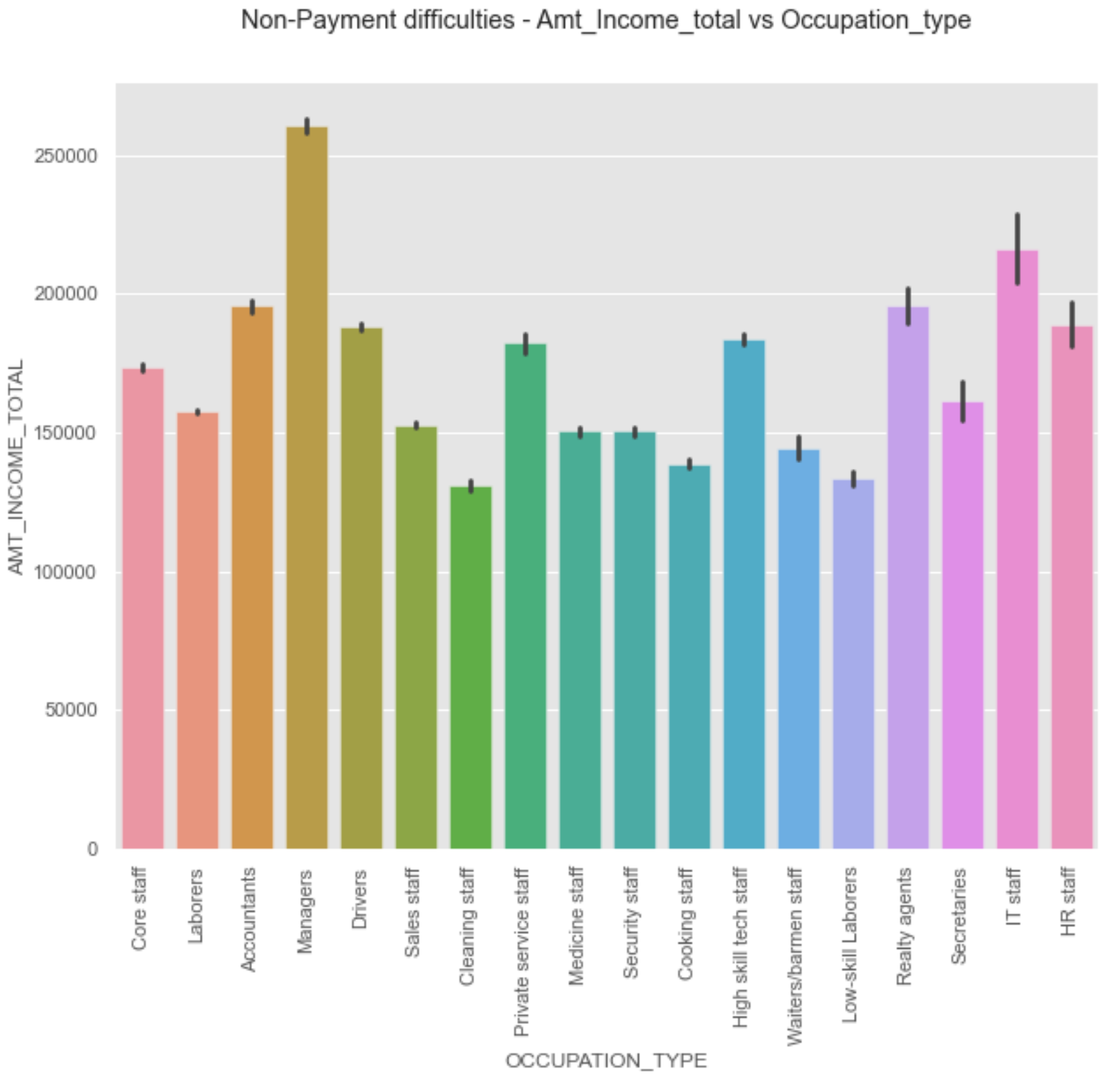
1) Quite similar with Target 0 From the above box plot we can say that Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others.

2) Most of the outliers are from Education type 'Higher education' and 'Secondary'. Civil marriage for Academic degree is having most of the credits in the third quartile.

3) Females are having high payment difficulties than male

Plot to see the Payment difficulties - AMT_INCOME_TOTAL vs OCCUPATION_TYPE





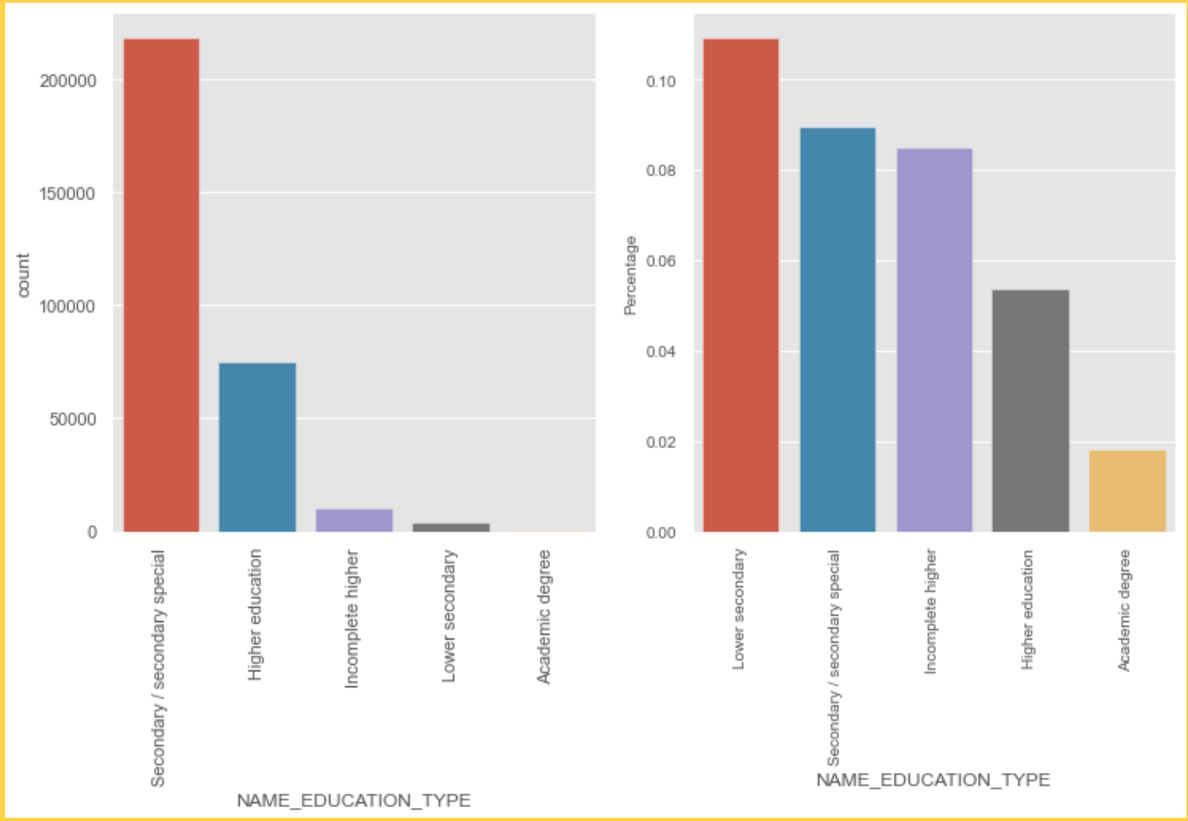
Inferences to be made from the above chart

1) Here we see that both the graphs looks almost similar.

2) The income for managers is very high compared to other categories, who don't defaults.

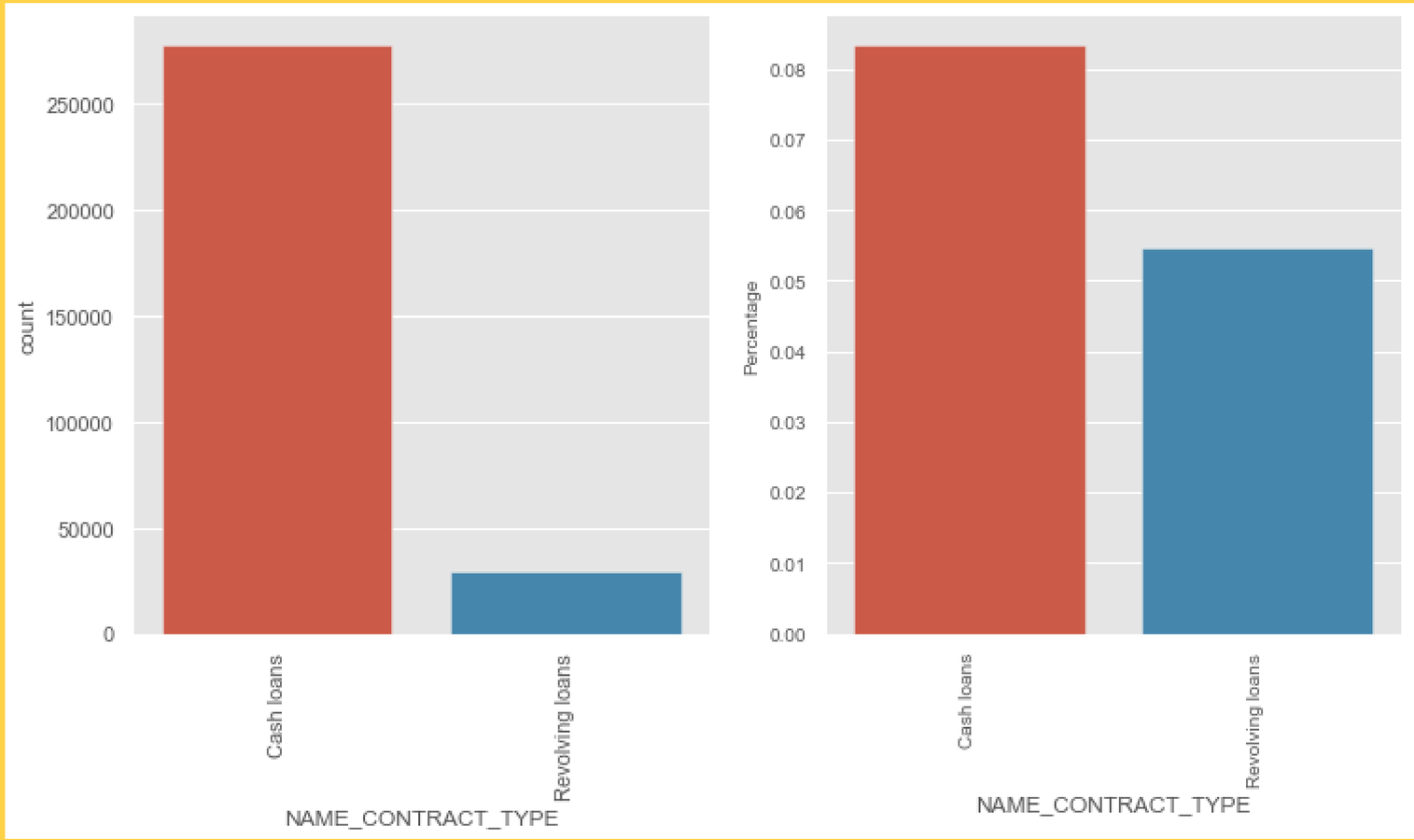
3) The probability of payments difficulties is high for Laborers.

Bivariate Analysis of Categorical vs Categorical Variables



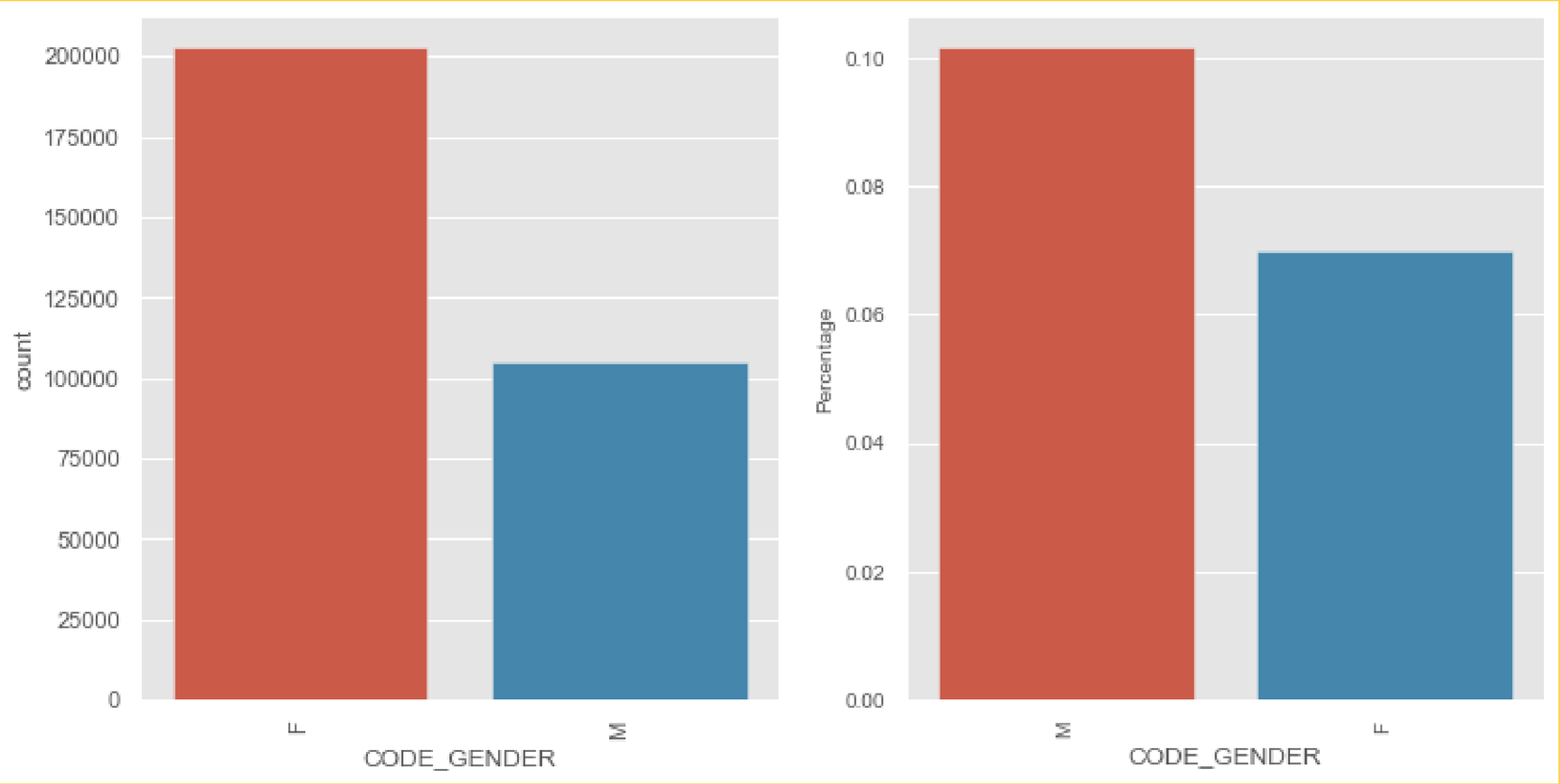
From the above plot, we can infer that, clients with 'Lower secondary' education type have maximum percentage of Loan-Payment Difficulties.

NAME_CONTRACT_TYPE with maximum Loan-Payment Difficulties



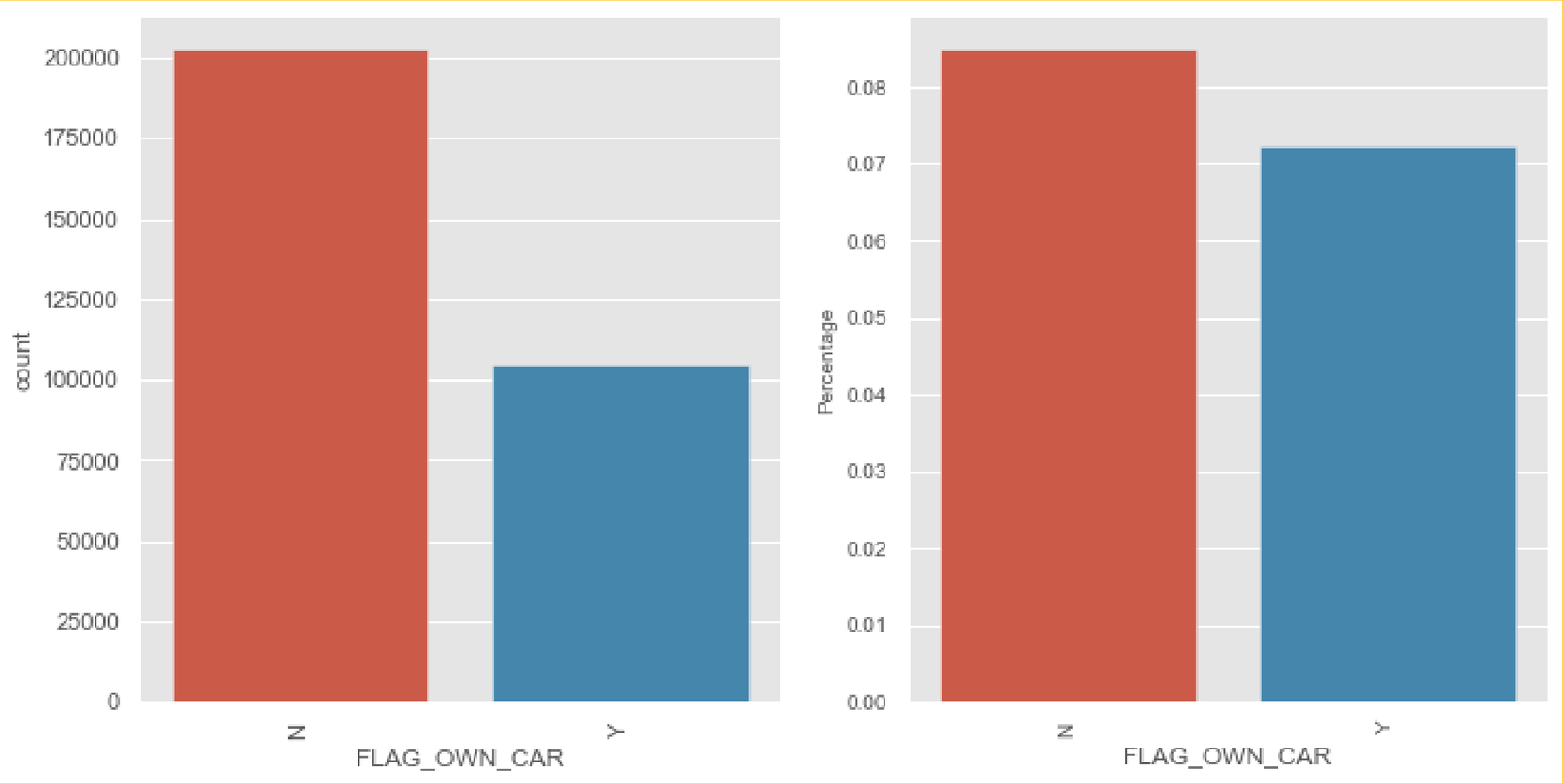
The above plot says that, the clients with 'Cash loans' contract type have maximum percentage of Loan Payemnt Difficulties.

CODE_GENDER with maximum Loan-Payment Difficulties



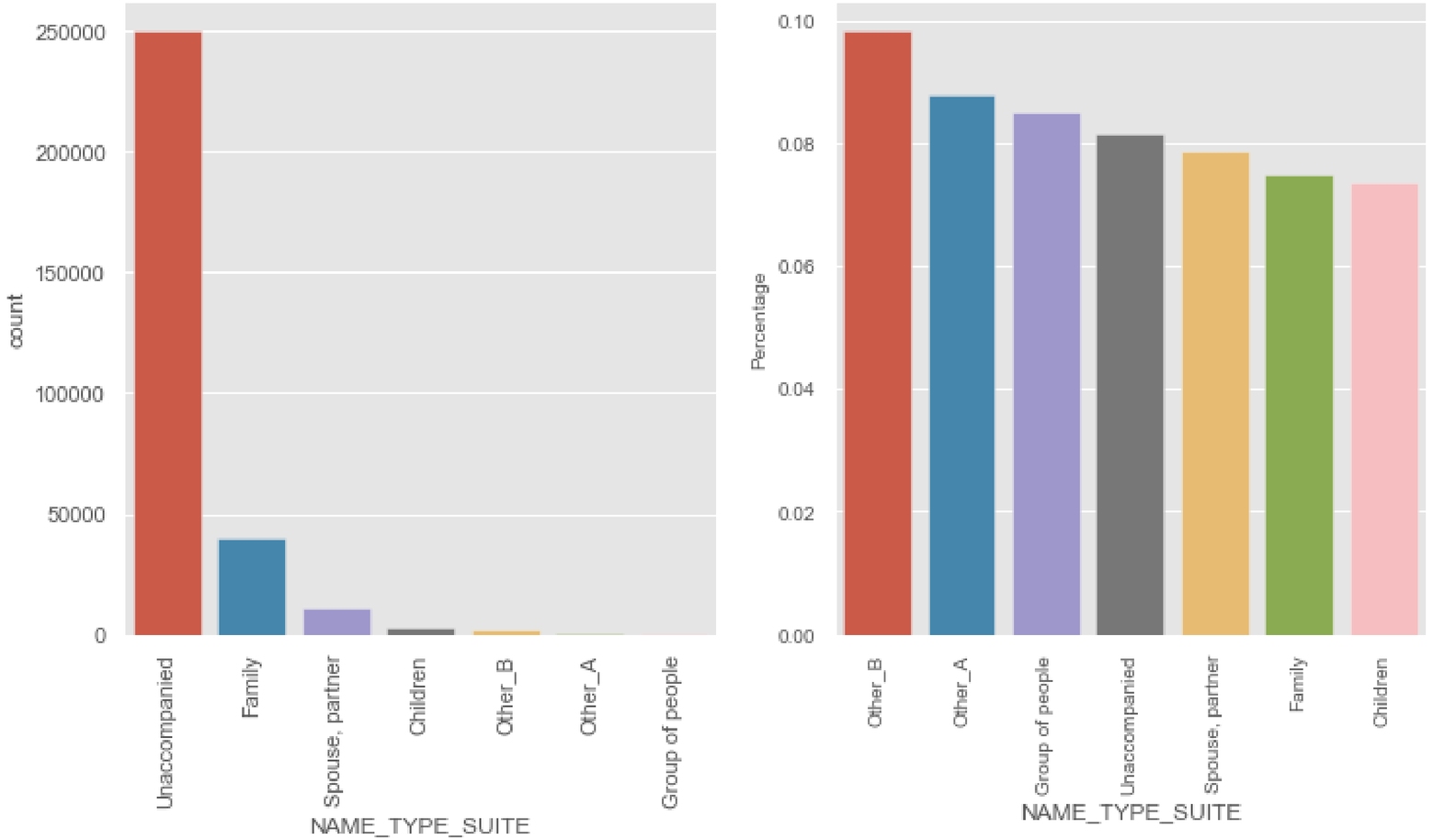
The above plot says that, the clients with 'Males' have maximum percentage of Loan Payemnt Difficulties.

FLAG_OWN_CAR with maximum Loan-Payment Difficulties



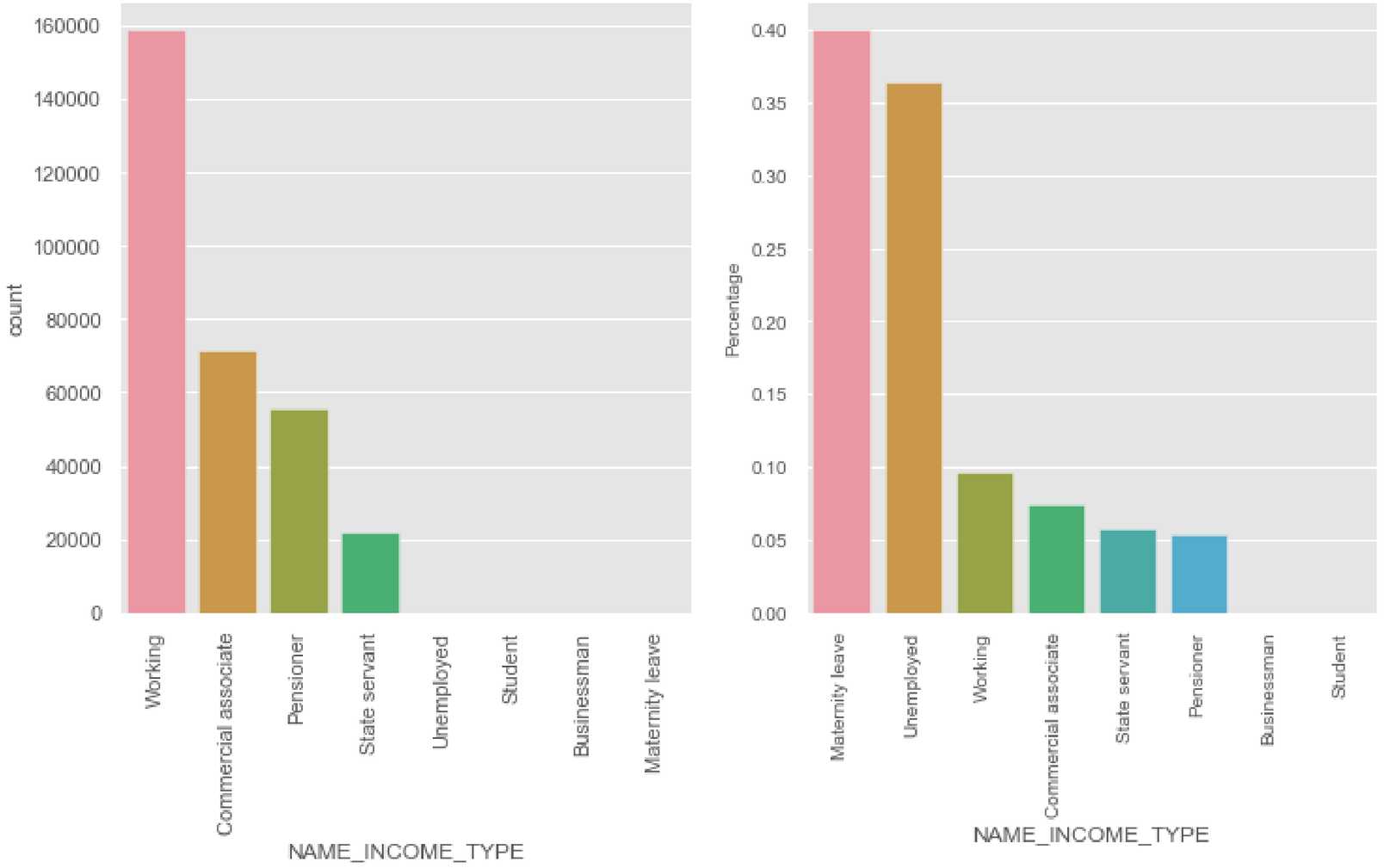
The above plot says that, the clients with 'car' have less percentage of Loan Payemnt Difficulties than the clients with no cars.

NAME_TYPE_SUITE with maximum Loan-Payment Difficulties



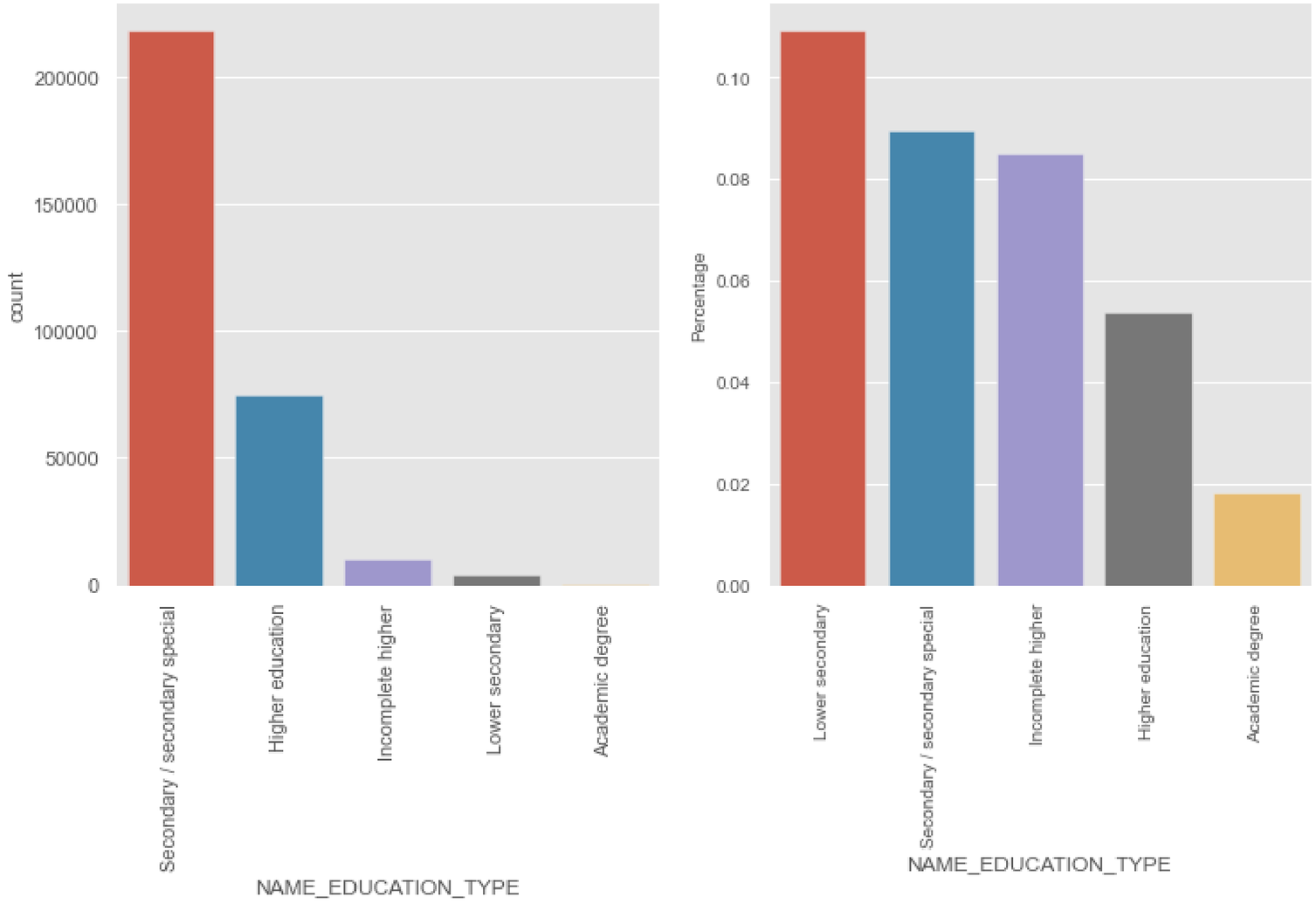
The above plot says that, the clients with 'Other-B' have maximum percentage of Loan Payemnt Difficulties.

NAME_INCOME_TYPE with maximum Loan-Payment Difficulties



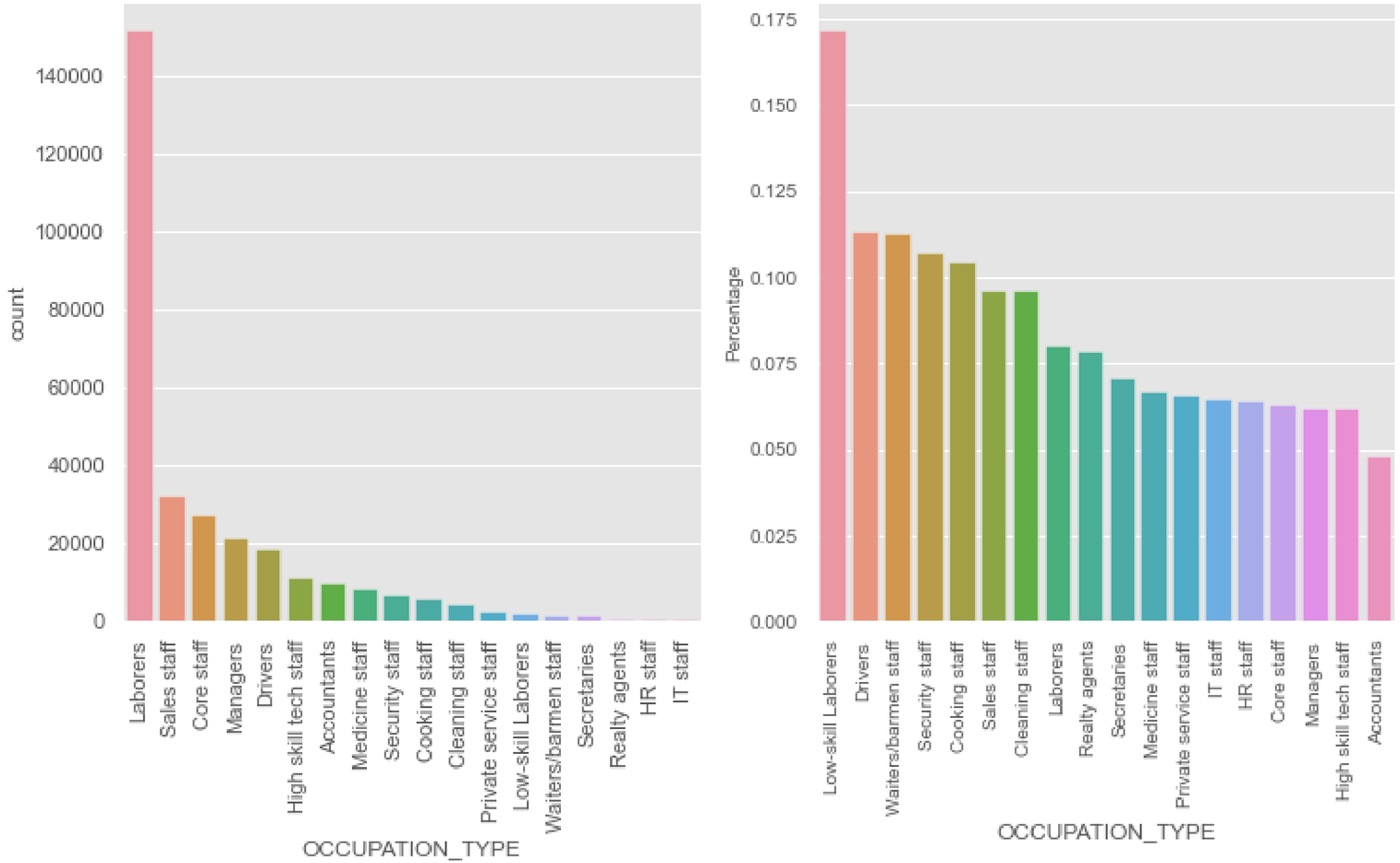
The above plot says that, the clients with 'Maternity leave' category have maximum percentage of Loan Payemnt Difficulties.

NAME_EDUCATION_TYPE with maximum Loan-Payment Difficulties



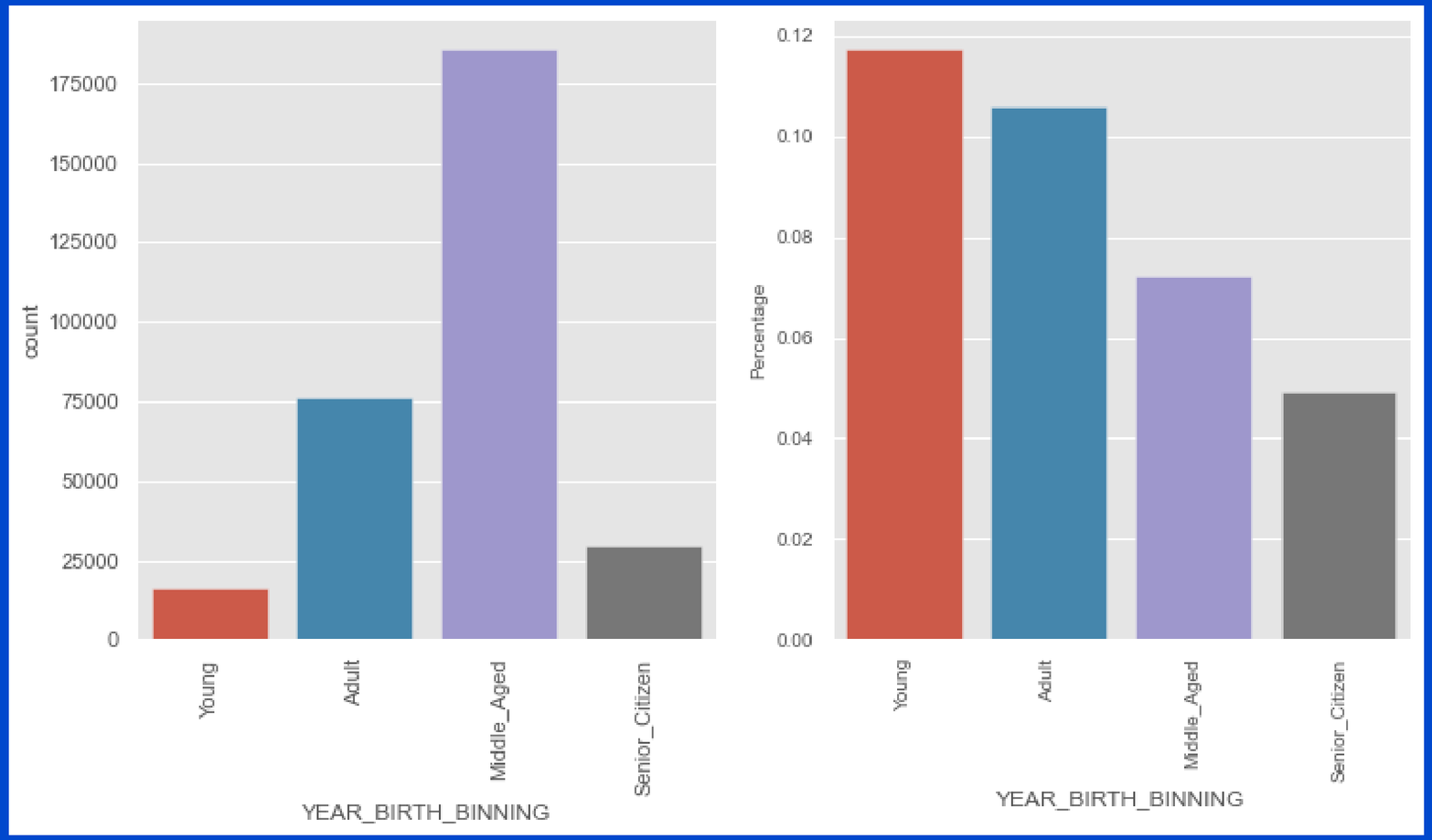
The above plot says that, the clients with 'Lower secondary' type have maximum percentage of Loan Payemnt Difficulties.

OCCUPATION_TYPE with maximum Loan-Payment Difficulties



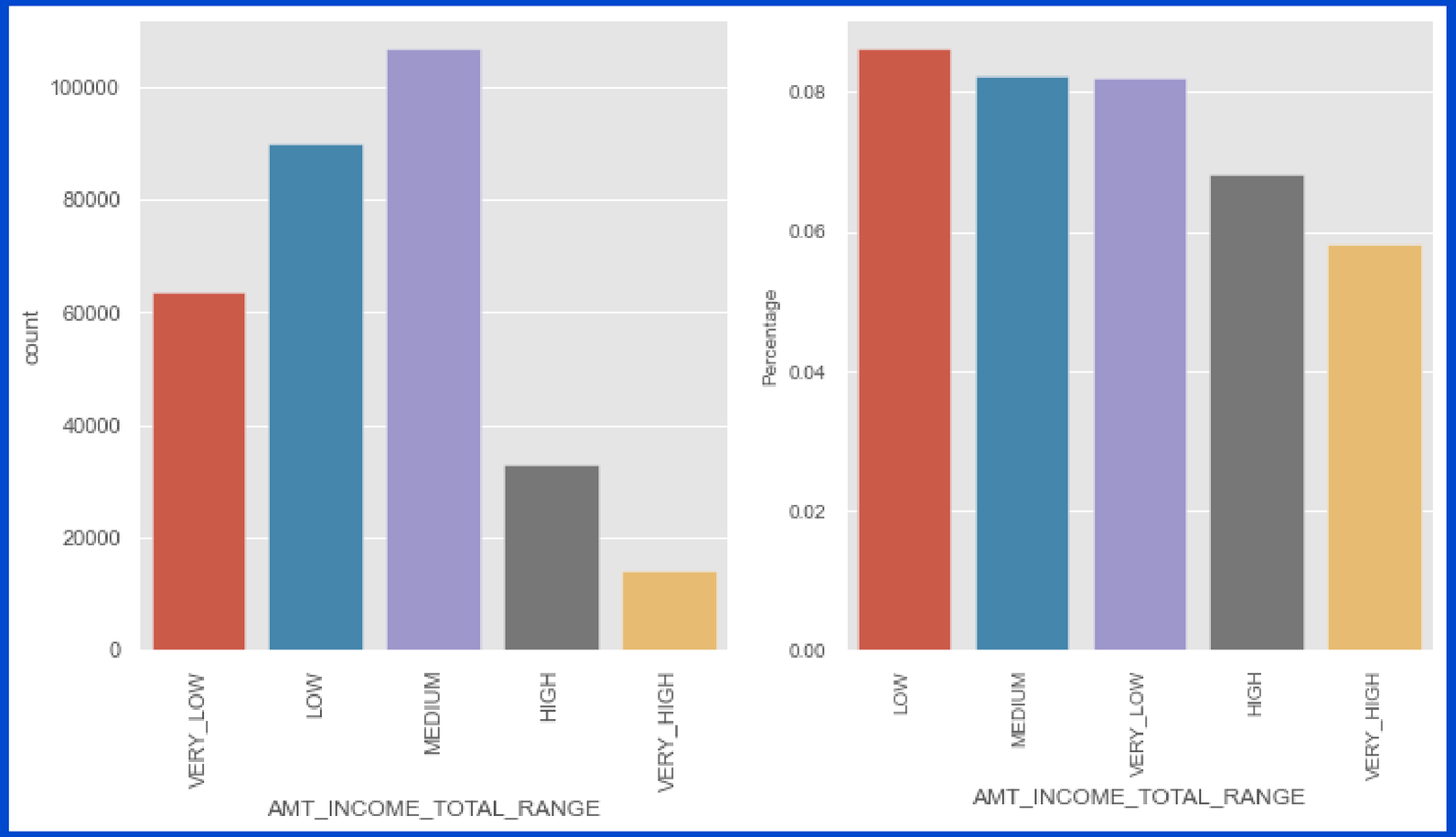
The above plot says that, the clients with 'Low skilled Laborers' category have the maximum percentage of Loan Payemnt Difficulties.

DAYS_BIRTH_BINNING_CATEGORIES with maximum Loan-Payment Difficulties



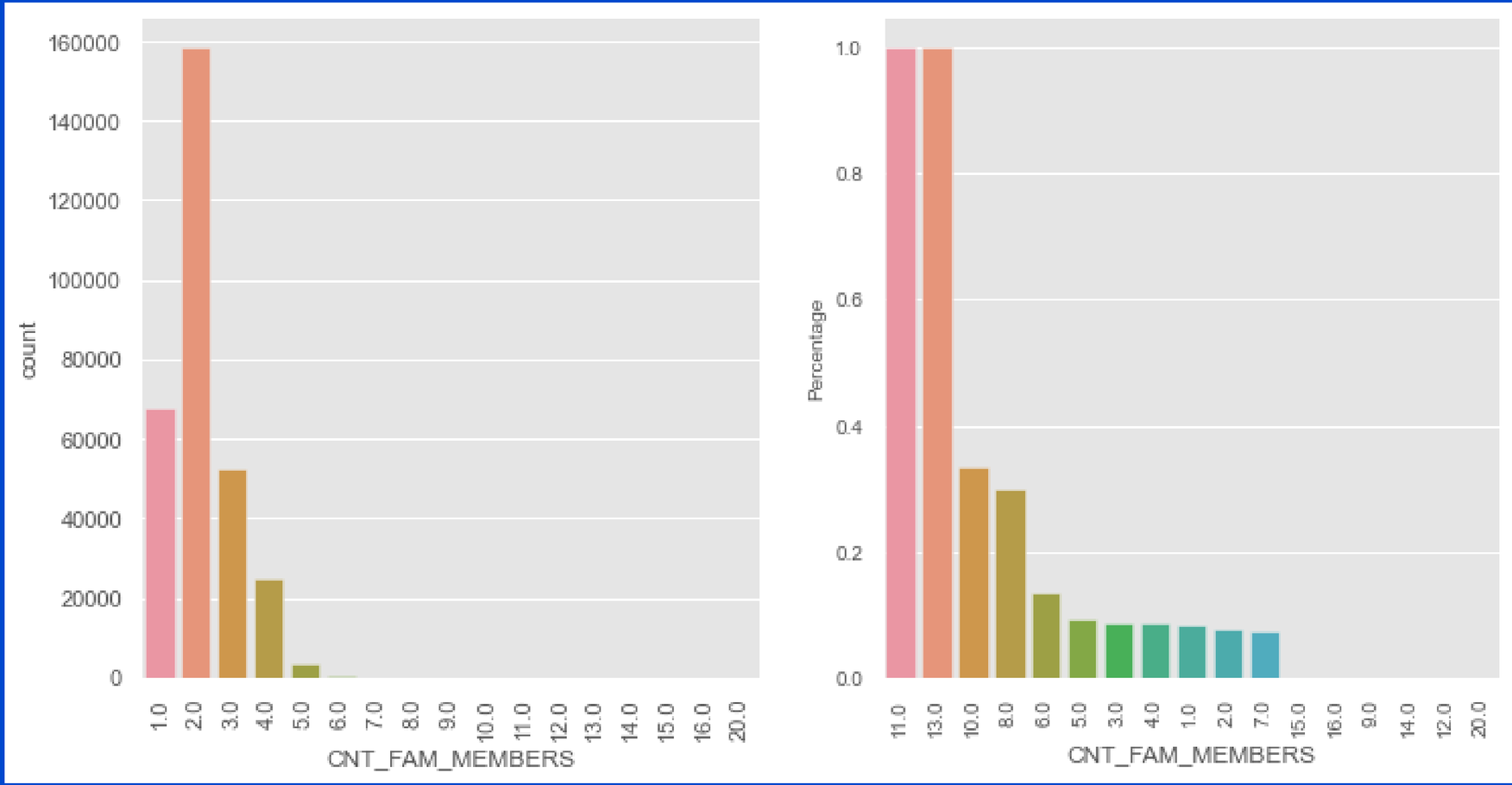
The above plot says that, the clients with 'Young' people have the maximum percentage of Loan Payemnt Difficulties.

AMT_INCOME_TOTAL_RANGE with maximum Loan-Payment Difficulties



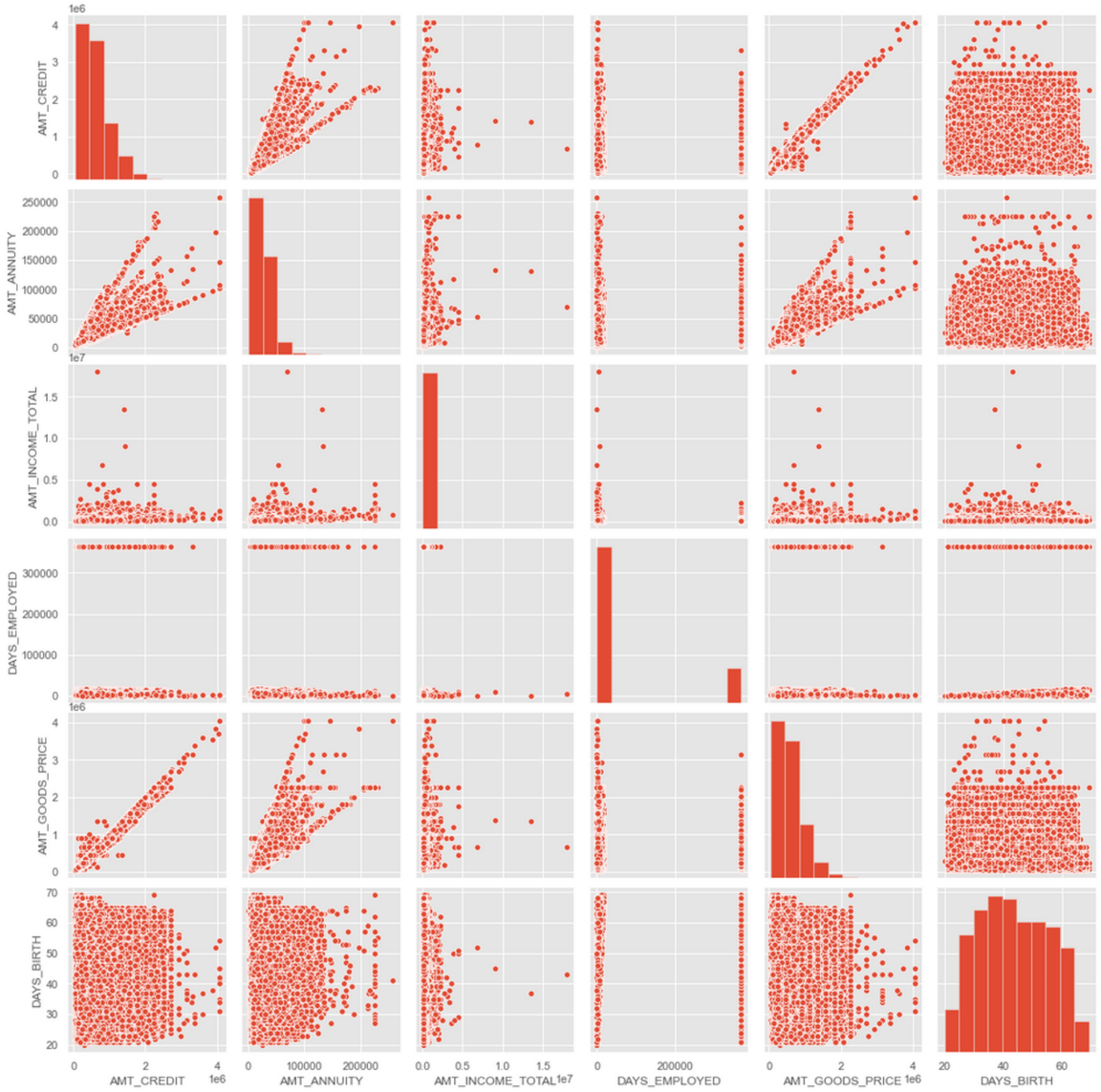
The above plot says that, the clients with 'Low' income have the maximum percentage of Loan Payemnt Difficulties.

CNT_FAM_MEMBERS with maximum Loan-Payment Difficulties

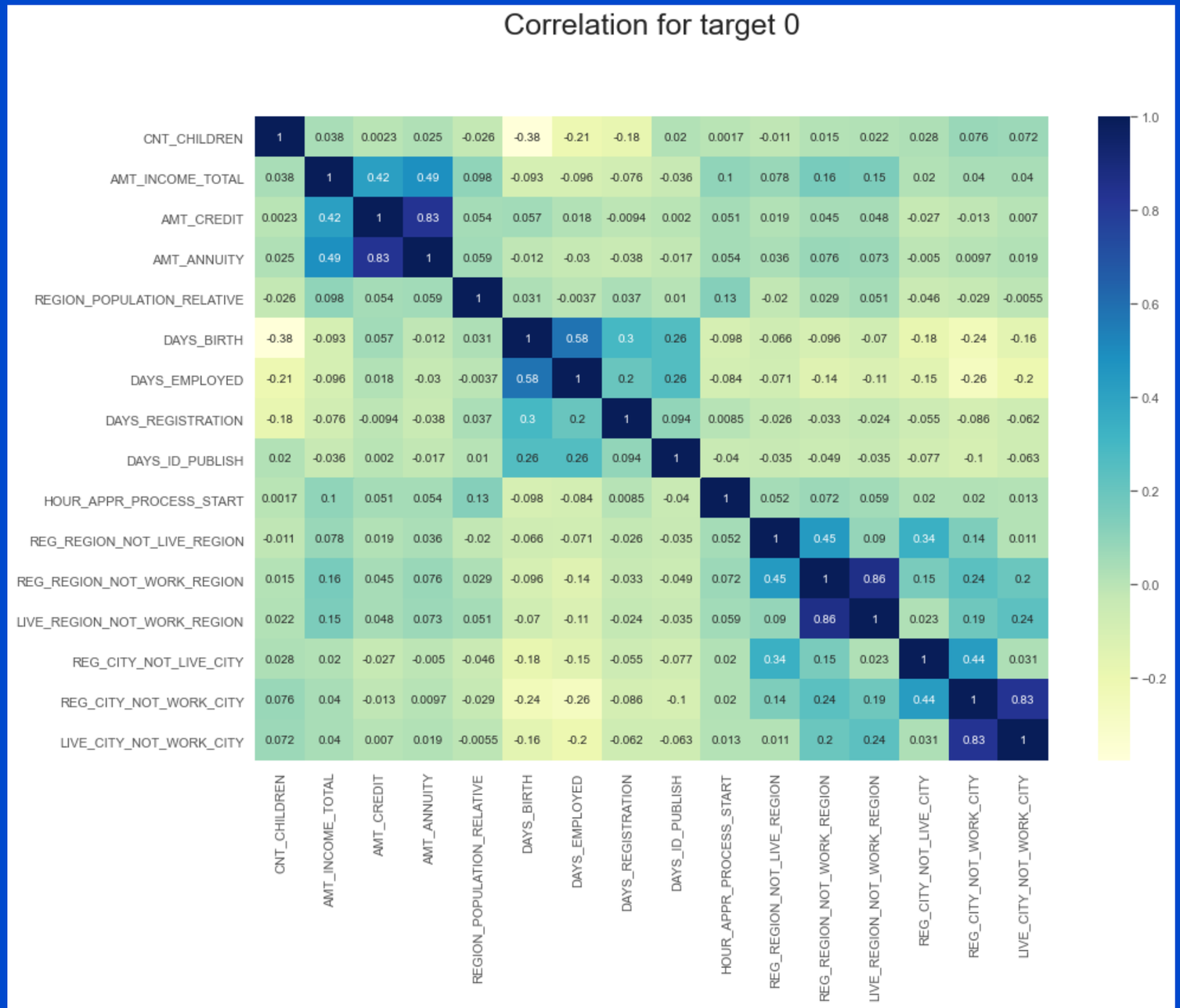


The above plot says that, the clients with '11 family members' category have the maximum percentage of Loan Payemnt Difficulties.

Bivariate Analysis of Numerical vs Numerical Variables

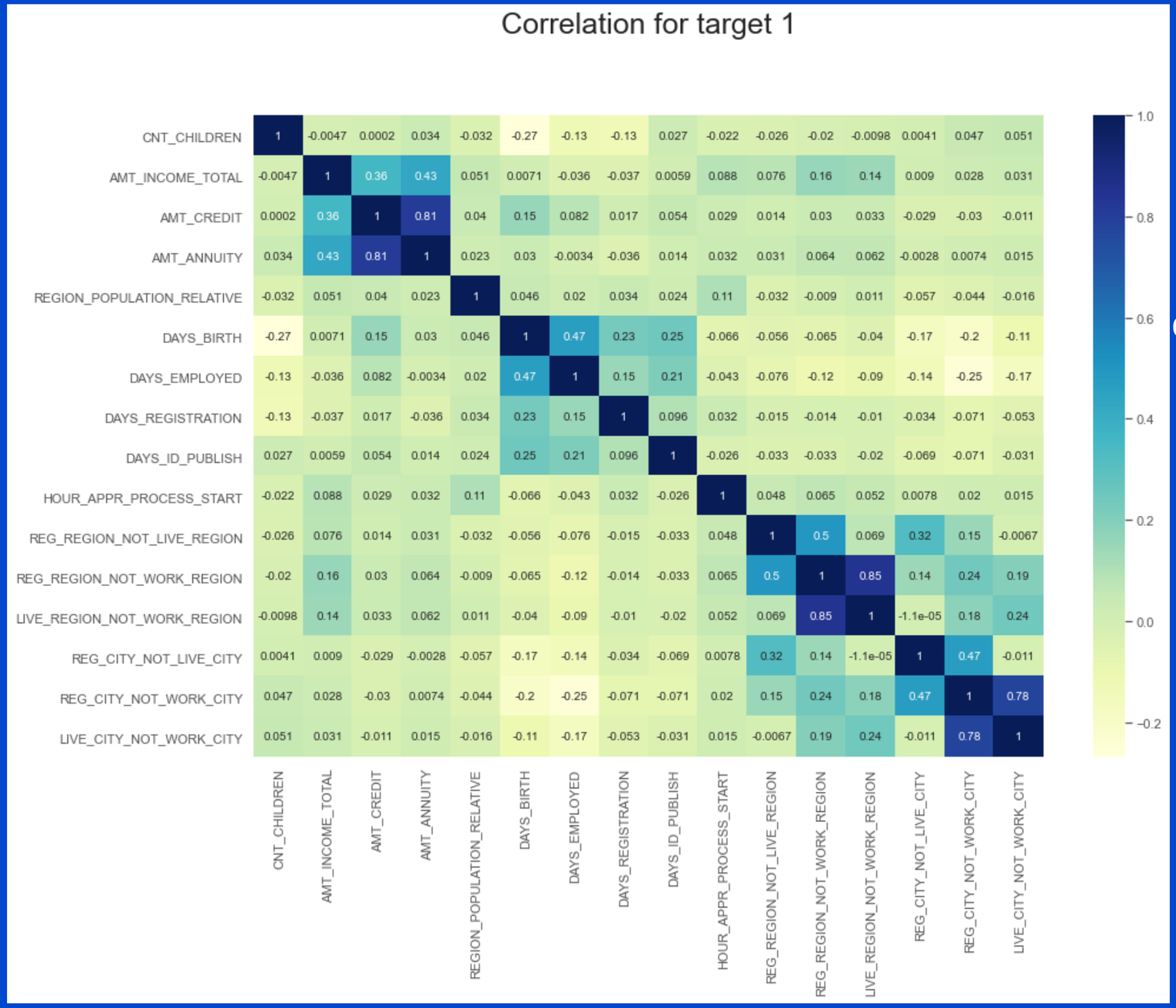


Multivariate Analysis



Inference

1) Credit amount is inversely proportional to the date of birth, which means Credit amount is higher for low age and vice-versa.
2) Credit amount is inversely proportional to the number of children client have, means Credit amount is higher for less children count client have and vice-versa.
3) Income amount is inversely proportional to the number of children client have, means more income for less children client have and vice-versa.
4) less children client have in densely populated area.
5) Credit amount is higher to densely populated area.
6) The income is also higher in densely populated area.

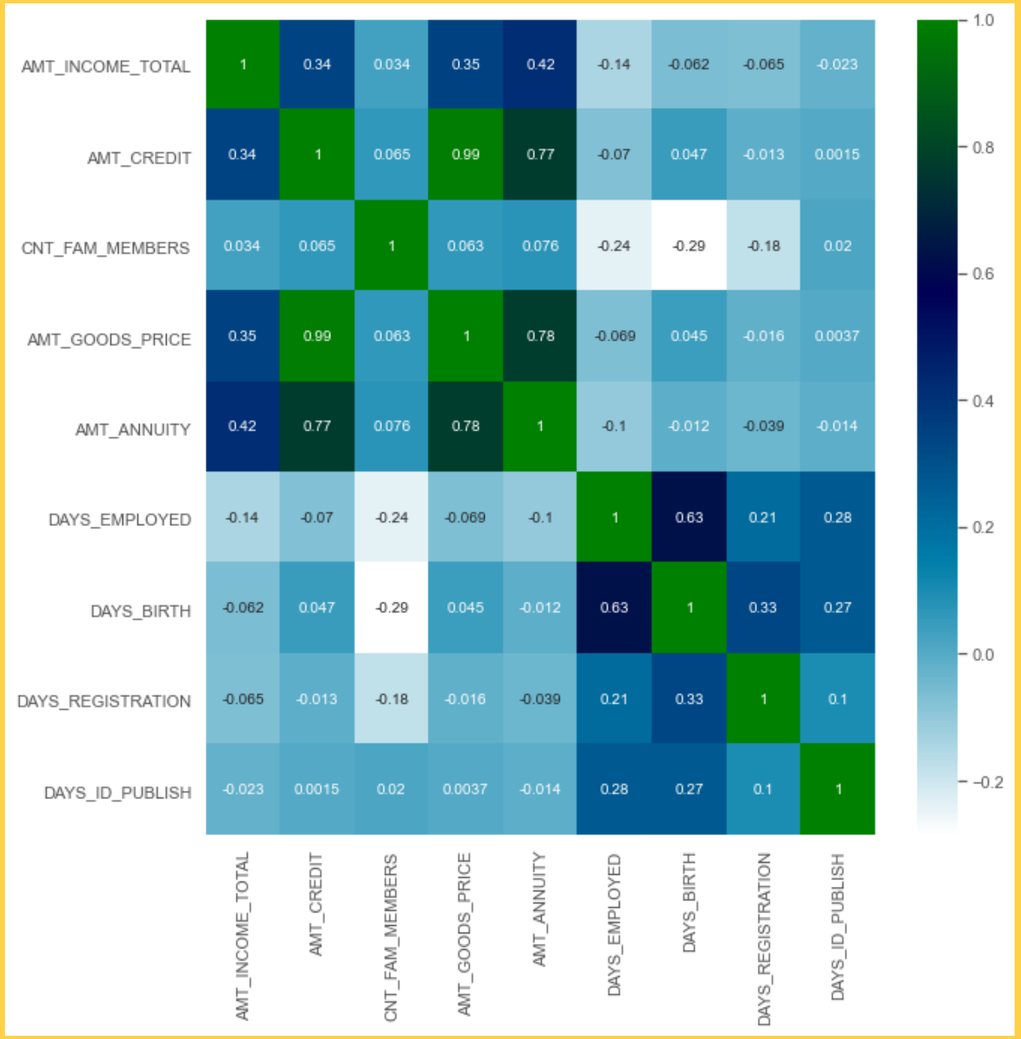


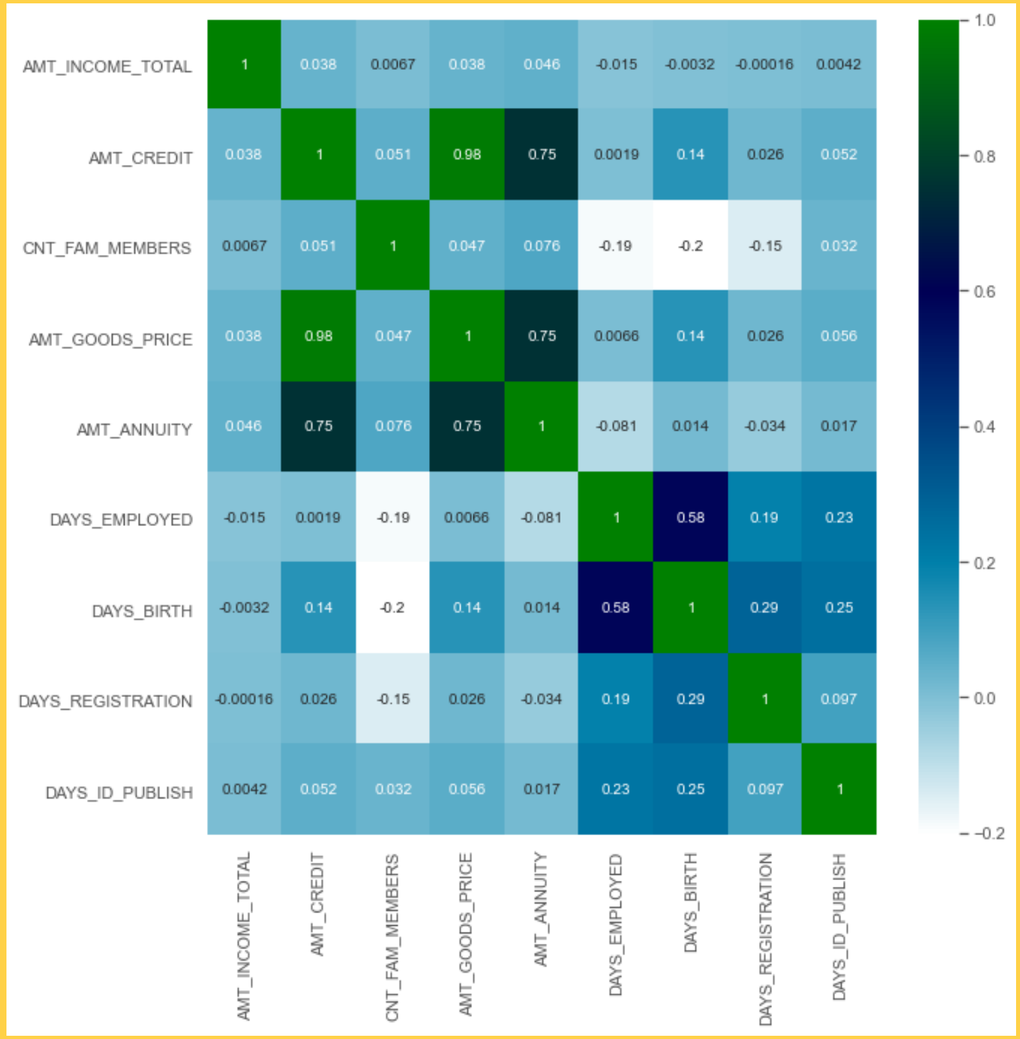
Inference

This heat map for Target 1 is also having quite a same observation just like Target 0. But for few points are different. They are listed below.

1) The client's permanent address does not match contact address are having less children and vice-versa

2) The client's permanent address does not match work address are having less children and vice-versa

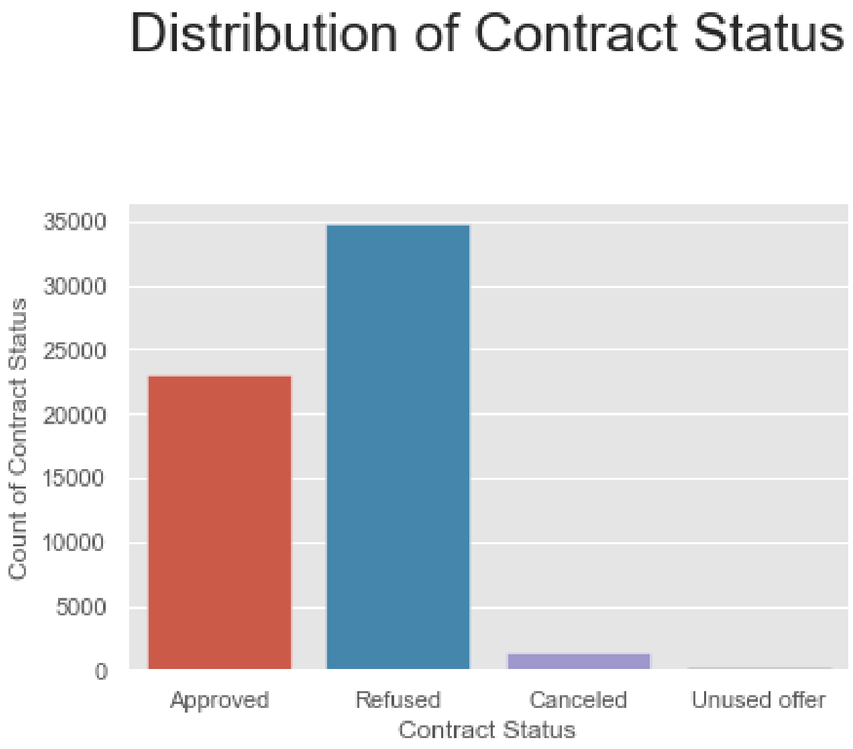


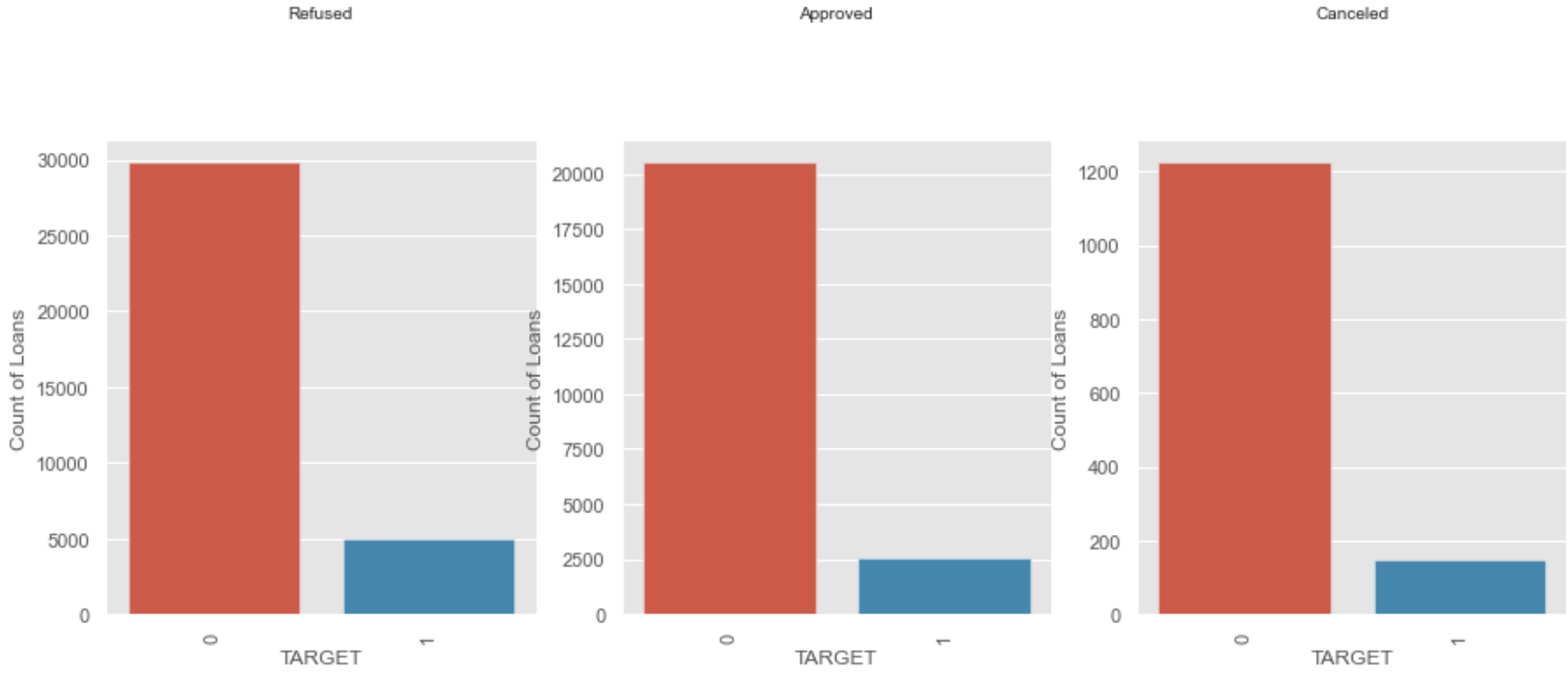


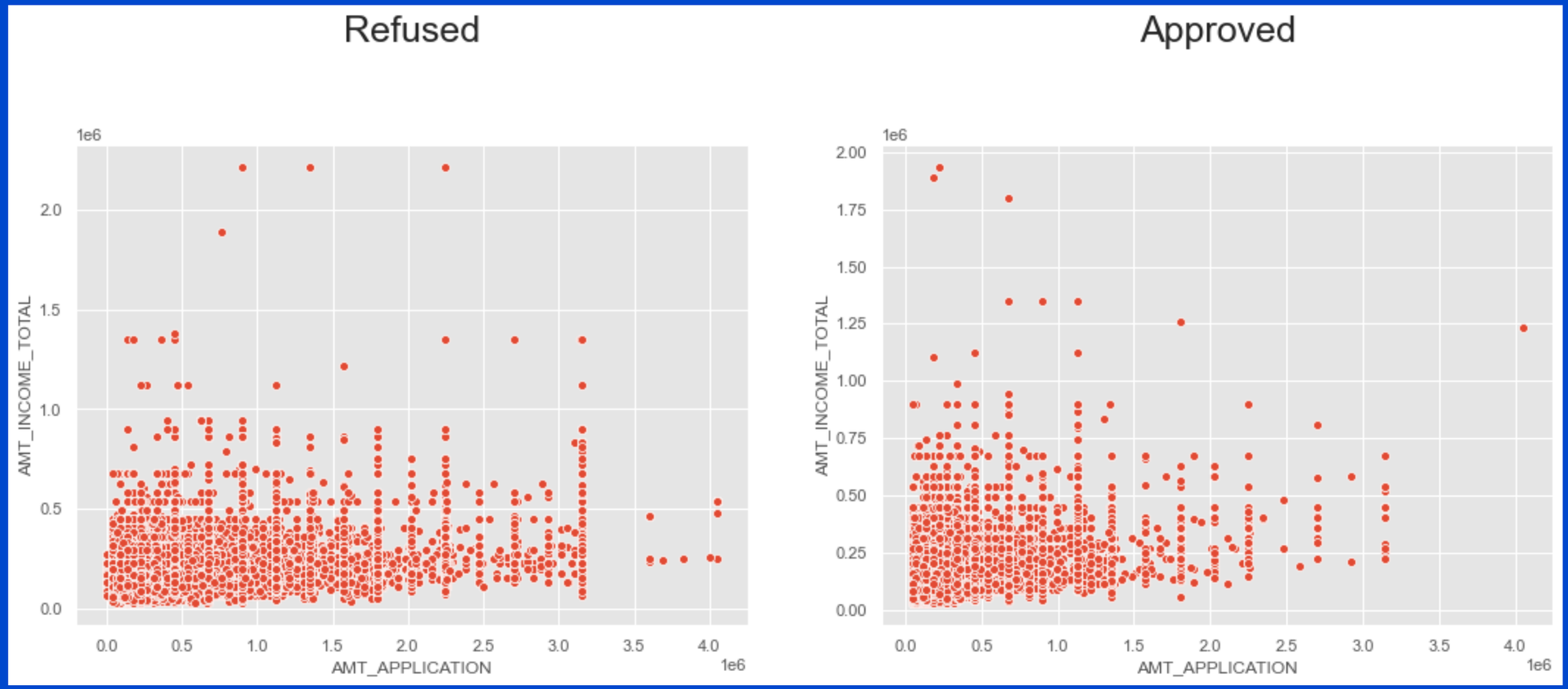
1) We can see that, there is high correlation between credit amount and goods price.

2) There appears to be some deviancies in the correlation of defaulters and non-defaulters such as credit amount comparison income type.

Previous application

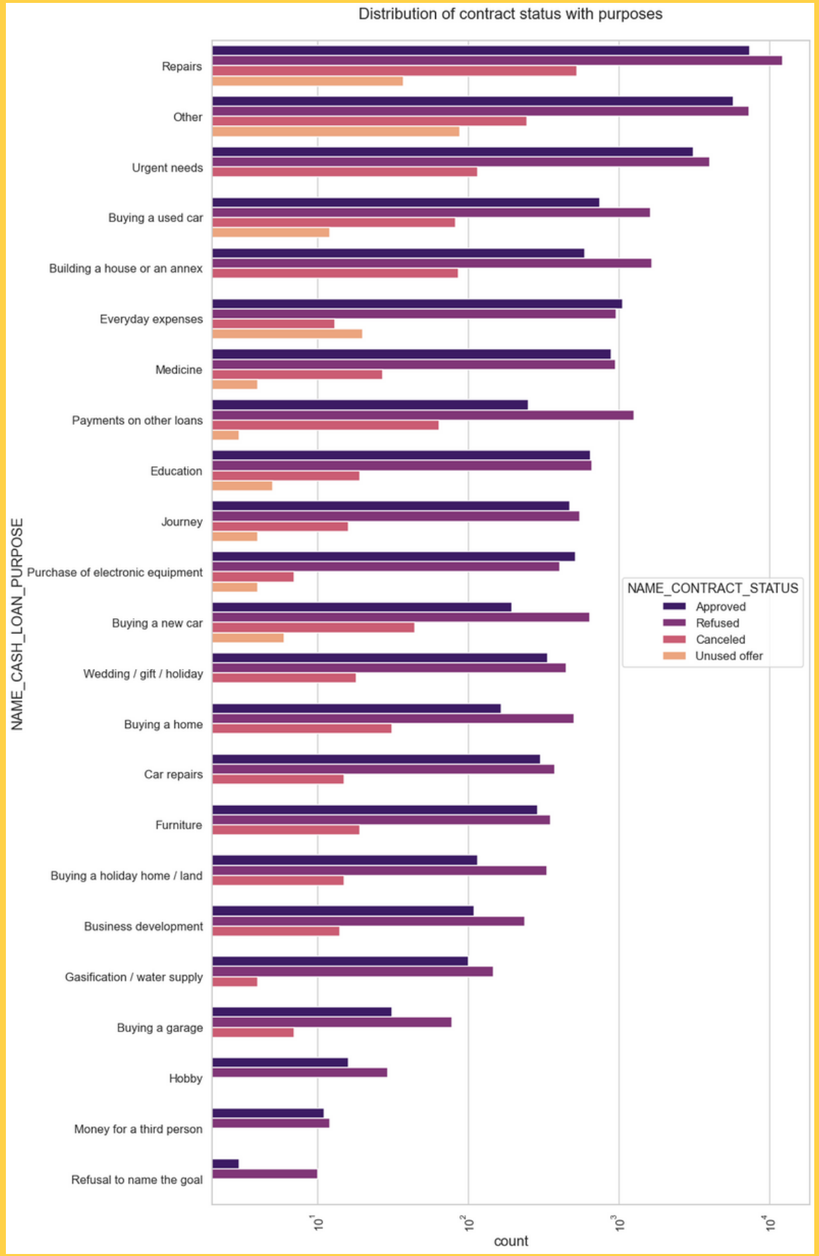






Loan request higher than 200k had a higher rejection rate. Also loan rejection rate was much lower if the income was higher than 500k.

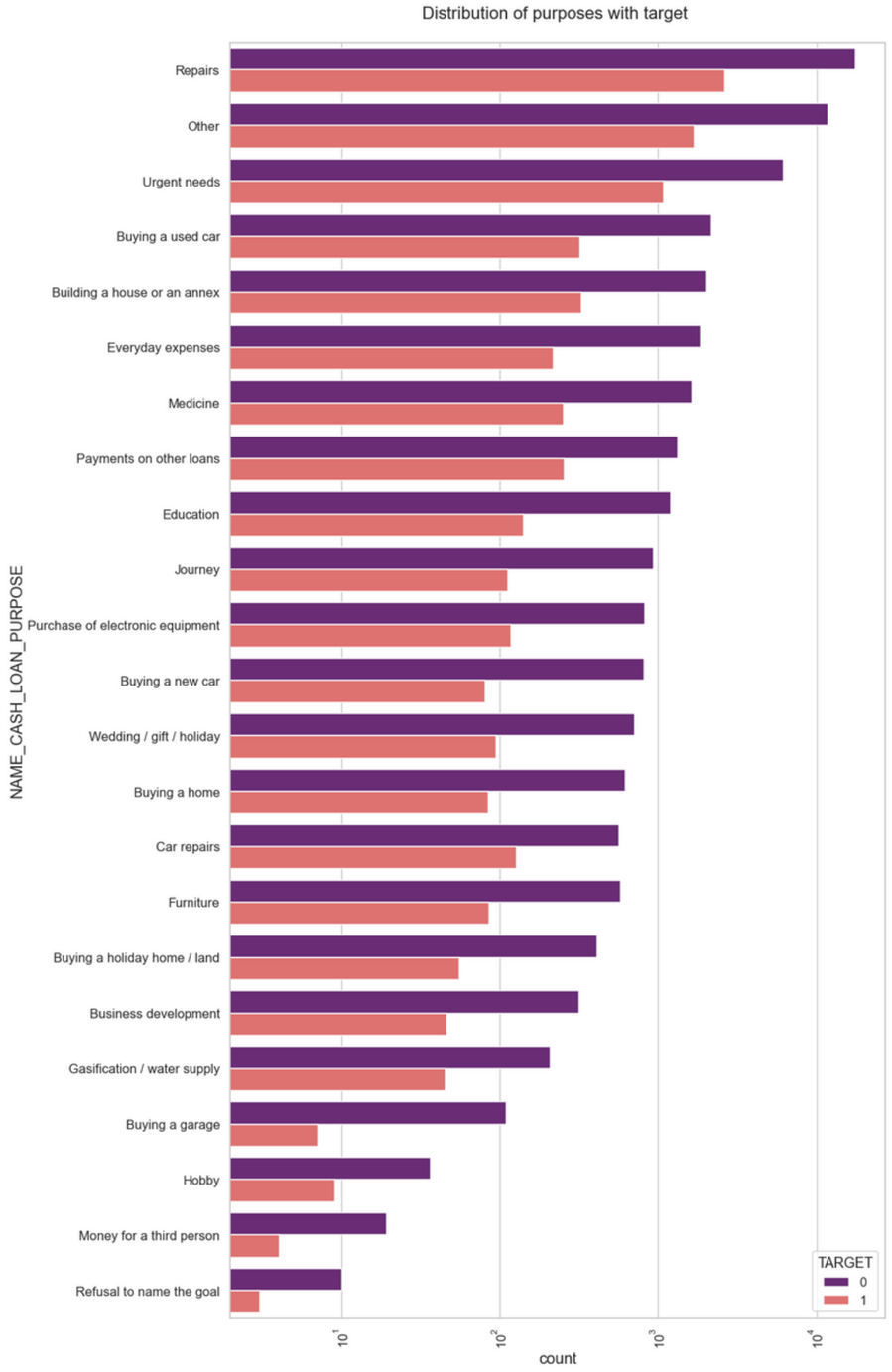
Performing univariate analysis



1) Most rejection of loans came from purpose 'repairs'.

2) For education purposes we have equal number of approves and rejection

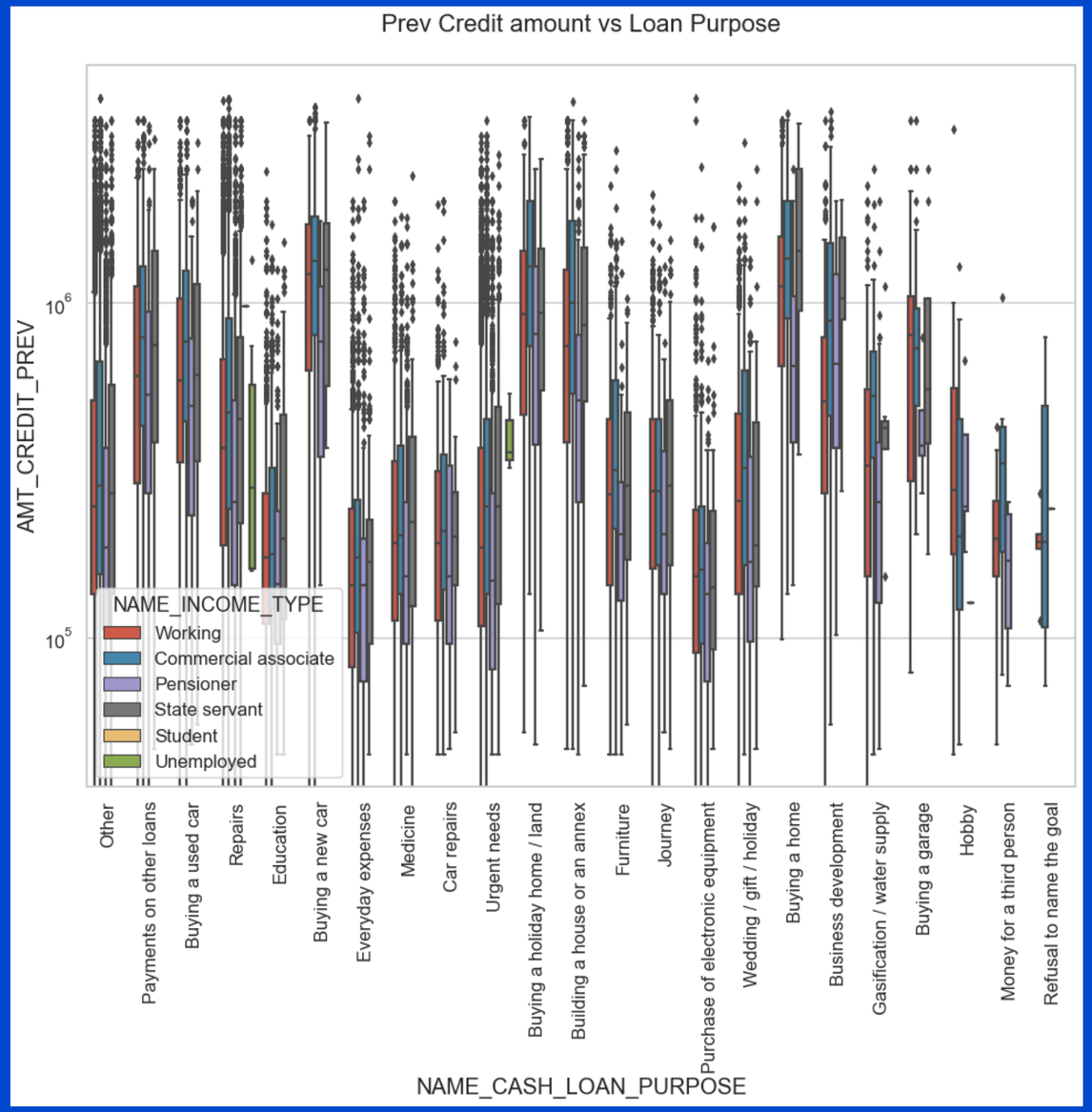
3) Paying other loans and buying a new car is having significant higher rejection than approves.



1) Loan purposes with 'Repairs' are facing more difficulites in payment on time.

2) There are few places where loan payment is significant higher than facing difficulties. They are 'Buying a garage', 'Business developemt', 'Buying land','Buying a new car' and 'Education' Hence we can focus on these purposes for which the client is having for minimal payment difficulties.

Performing bivariate analysis

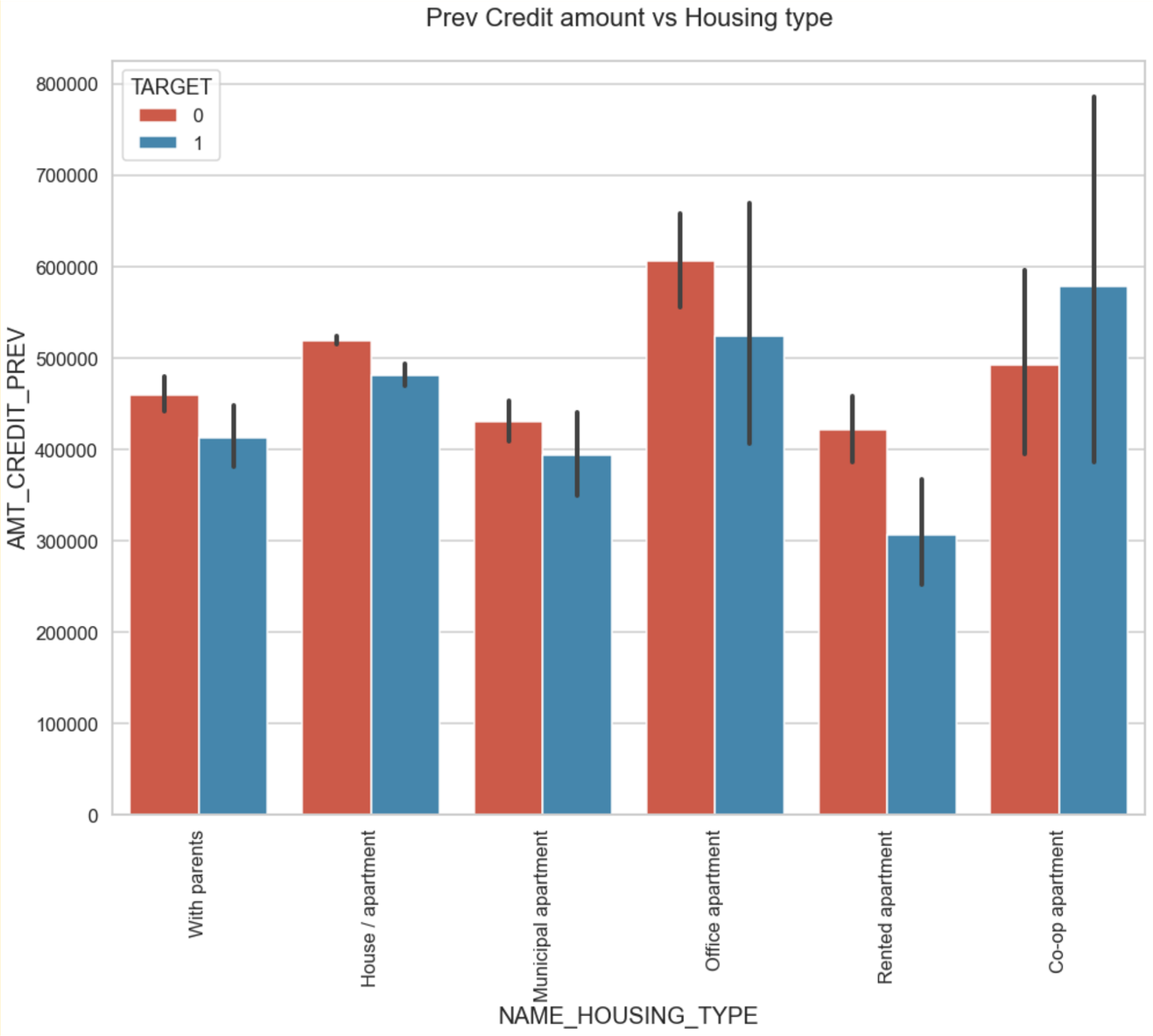


1) The credit amount of Loan purposes like 'Buying a home','Buying a land','Buying a new car' and'Building a house' is higher.

2) Income type of state servants have a significant amount of credit applied

3) Money for third person or a Hobby is having less credits applied for.

Plotting for Credit amount prev vs Housing type in logarithmic scale



Here for Housing type, office appartment is having higher credit of target 0 and co-op apartment is having higher credit of target 1. So, we can conclude that bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment. Bank can focus mostly on housing type with parents or House\appartment or miuncipal appartment for successful payments.

Conclusion

After analysing the application data and previous data, we can conclude from the observations made on various parameters/columns and find out the attributes which can be repayer or defaulters.

Below are the few observed metrix that clients falls under repayer category:

1) DAYS_BIRTH: Clients having age 50 or above are likely to fall under less defaulters category

2) NAME_INCOME_TYPE: We have infered that, the Student and Businessmen category have no defaults.

3) DAYS_EMPLOYED: Clients with more number of experience has the less probability to default.

4) AMT_INCOME_TOTAL: Clients having income more, has very less history of defaulters.

5) NAME_CASH_LOAN_PURPOSE: Loans bought for Hobby, Buying garage are being repayed mostly.

6) CNT_CHILDREN: Clients with less child or less family members shows the pattern of less defaulters

7) NAME_EDUCATION_TYPE: People with Academic degree has less defaults.

8) ORGANIZATION_TYPE: Clients with Trade Type 4 and 5 and Industry type 8 have tend to repay their loans.

Below are the few observed metrix that clients falls under defaulter category:

1) OCCUPATION_TYPE: Low-skill Laborers, Drivers, Security staff, Laborers and Cooking staff, these people are having high payment difficulties.

2) NAME_INCOME_TYPE: We have infered that, Clients who are either at Maternity leave OR Unemployed has high rate of repaying their loans.

3) DAYS_BIRTH: Young and Adult people tend to have high difficulties in repaying the loan.

4) NAME_EDUCATION_TYPE: Lower Secondary and Secondary education categories have the high probabilites of default.

5) DAYS_EMPLOYED: Clients with less employment rate is having high payment difficulties.

6) NAME_FAMILY_STATUS : civil marriage or single categories default a lot. So, we their applications can be rejected.

7) CODE_GENDER: The percentage of loan default is more for Men in Gender category.

8) ORGANIZATION_TYPE: Clients with Trade Type 4 and 5 and Industry type 8 have tend to repay their loans.

Hence we recommend the bank to look into the above metrices before approving the loans to their clients


thank
you