Credit EDA Case Study

BY PRIYANKA AKAVARAM & SAI LALITH SISTLA,

EPDS PROGRAM FROM IIIT BANGALORE AT UPGRAD

Problem statement

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 as 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. 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.

Data and Assumptions

Data source:

- application_data.csv
- previous_application.csv

Assumptions made in application data dataframe

- As looking into the data above , it is observed that there are no data type mismatch
- After checking for missing values and outliers in the columns, we filled the missing values with mean/median/mode and constants appropriately as given below.

Data and Assumptions

Column name	Imputed with/Outlier by	Reason
DAYS_EMPLOYES	Outlier treatment	Outliers causing poor data readability
AMT_ANNUITY	Outlier treatment	
AMT_INCOME_PRICE	Outlier treatment	
EXT_SOURCE_2	Imputed with mean	As there are no outliers
AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_YEAR, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_HOUR	Imputed with mode	Mode
OCCUPATION_TYPE, NAME_TYPE_SUITE, ORGANIZATION_TYPE	Filling missing values with unknown	Huge amount of missing values, hence binning into unknown category instead of imputing it.
CNT_FAM_MEMBERS, DAYS_LAST_PHONE_CHANGE	Imputed with mode	
DEF_60_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, OBS_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE	Imputed with median	As data is continuous and outliers exist imputing it with median instead of mean

Observations for application data

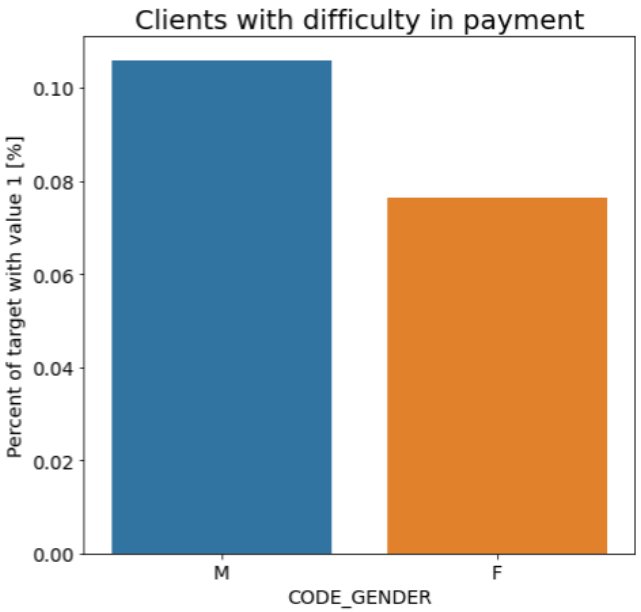
UNIVARIATE ANALYSIS

Observations of CODE_GENDER

- Number of male clients is half the number of female clients

- Males clients have more difficulties compared to female clients while repaying the loan

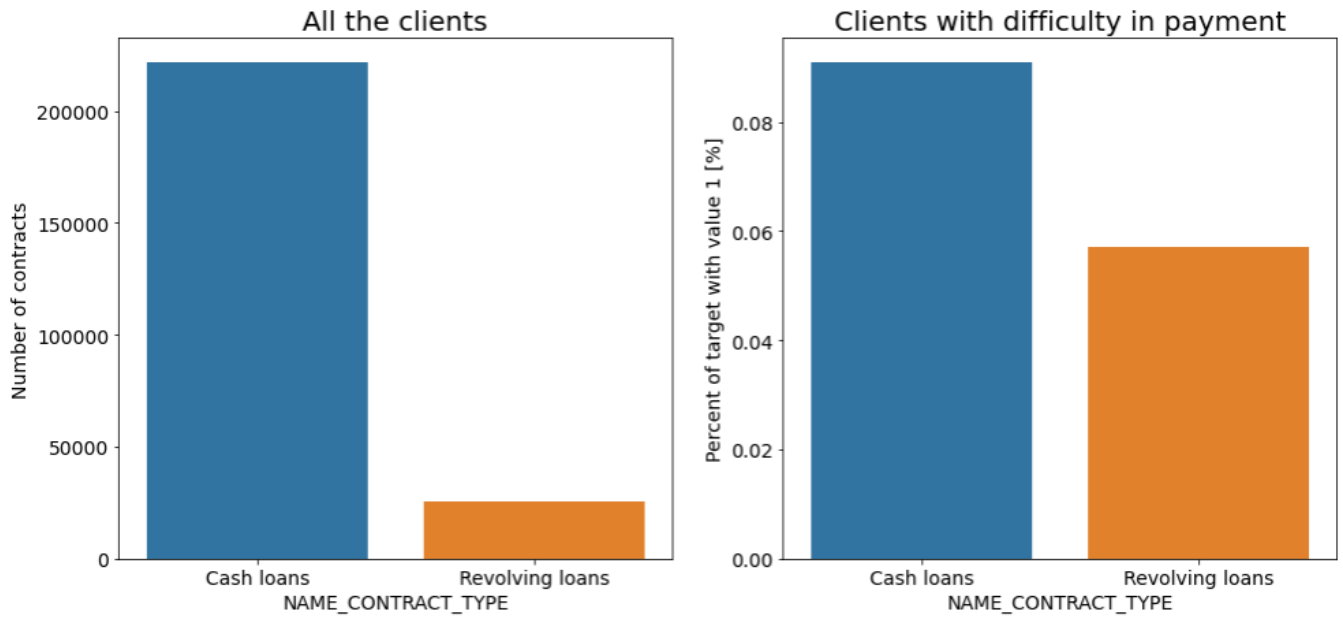
- 10% of male clients, 7% of female clients have difficulty in repaying their loan



Observations of NAME_COTRACT_TYPE

- Revolving loans have low contracts when compared to cash loans

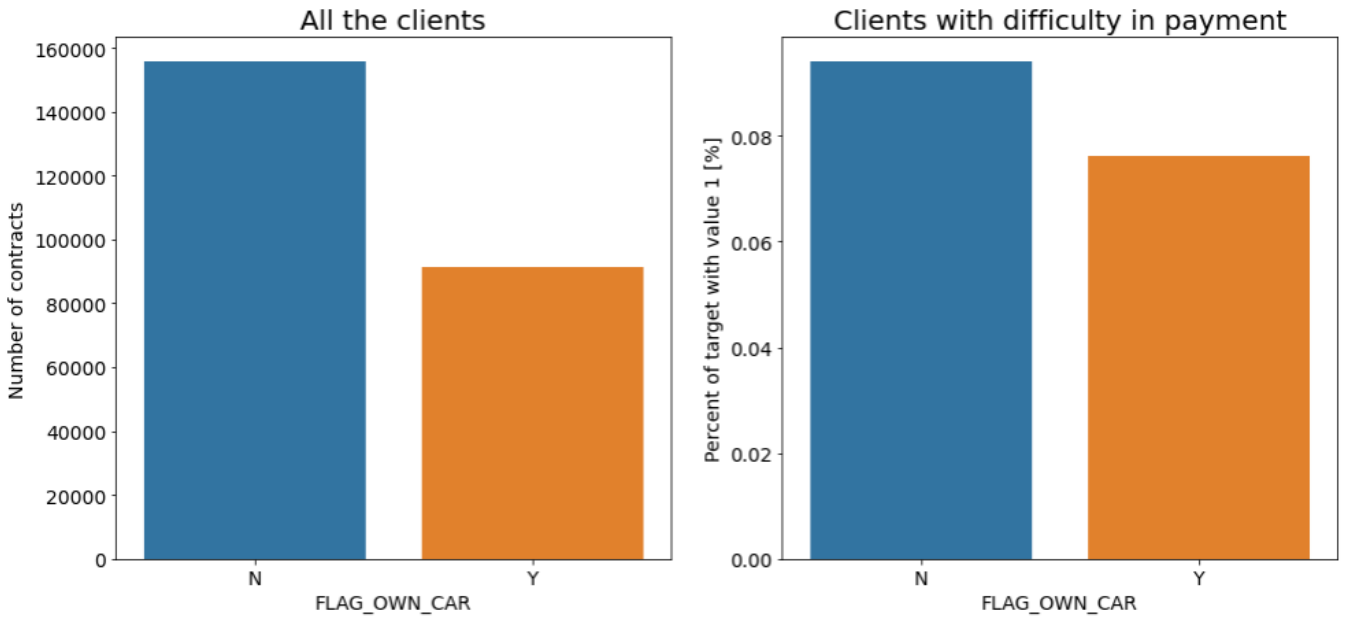
- More number of Revolving loans are not repaid compared to cash loans (compared in frequencies)



Observations for FLAG_OWN_CAR

- Clients who own a car are almost half of those who doesn't own a car.

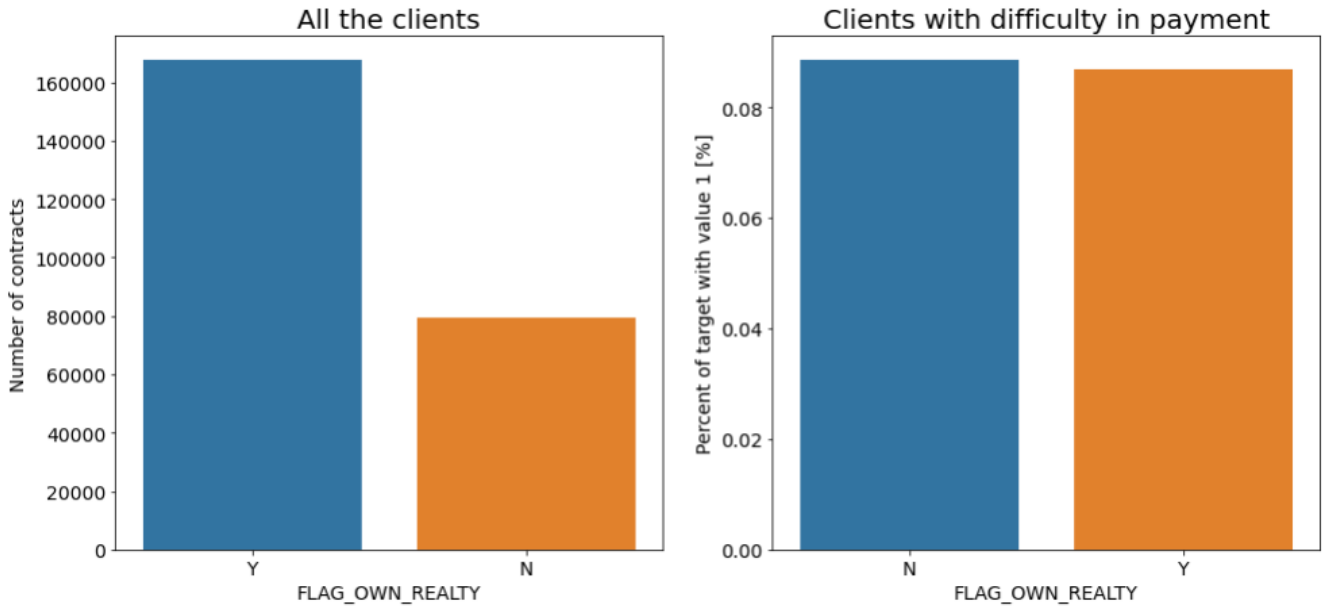
- The clients that owns a car are less likely to not repay a car that the ones that own.



Observations for FLAG_OWN_REALITY

- The clients that owns real estate are more than double of the ones that don't own.

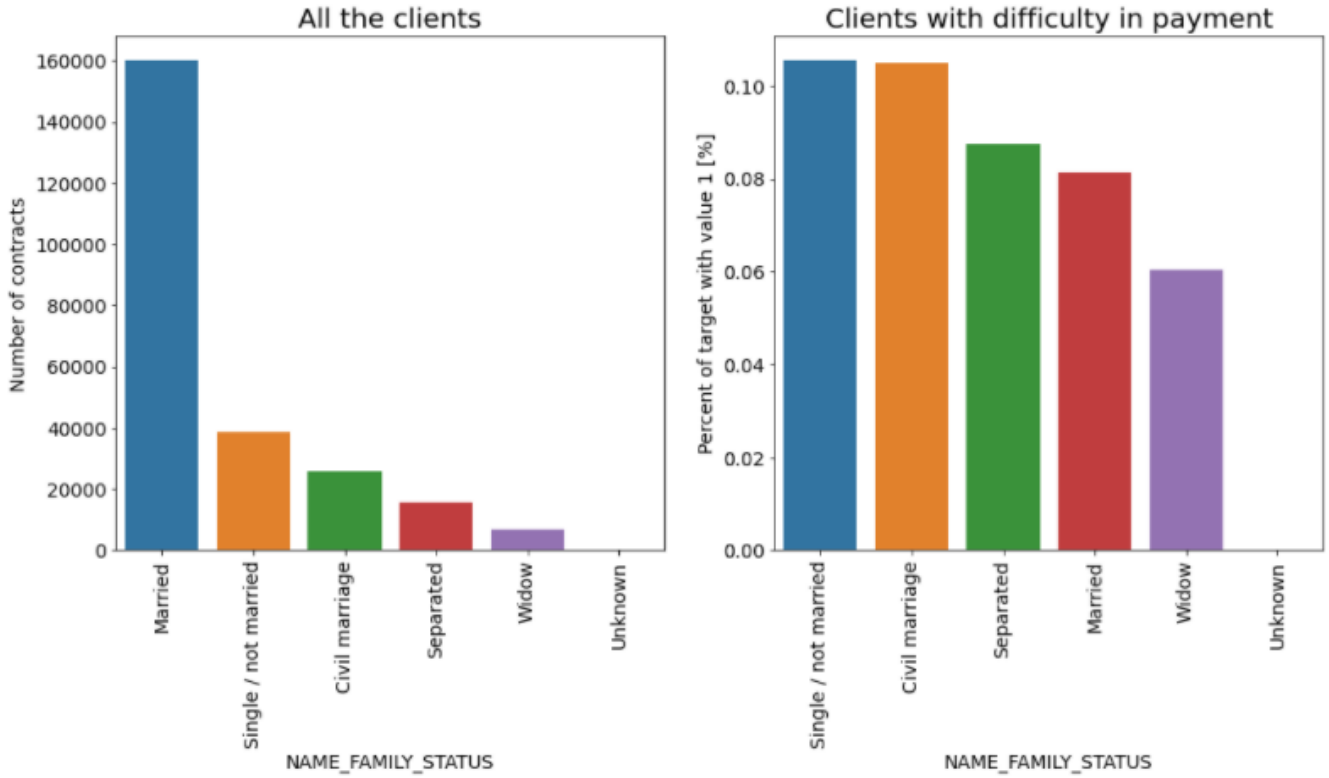
- Both the clients (owning real estate or not owning) have similar difficulties in repayment of their loan.



Observations for NAME_FAMILY_STATUS

- Most clients are married followed by single/not married, civil, separated, widow and unknown.

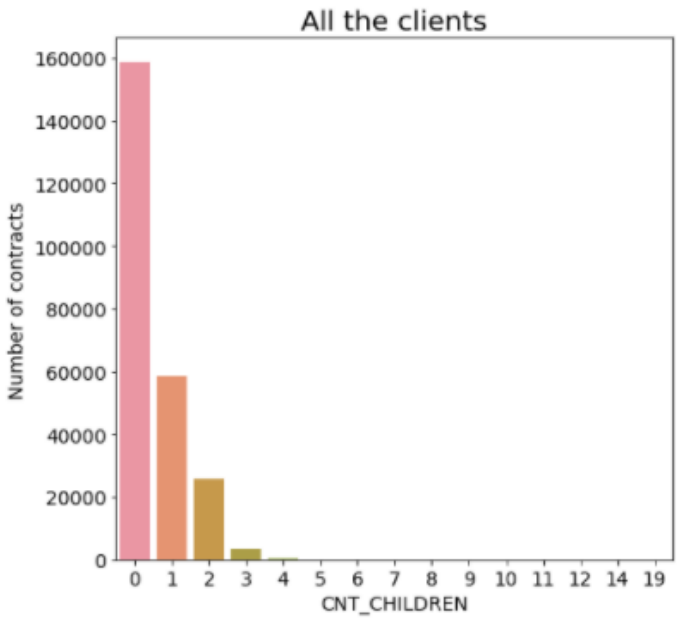
- Civil marriage clients have most percentage of loan non repayment while widows followed by unknown are the lowest

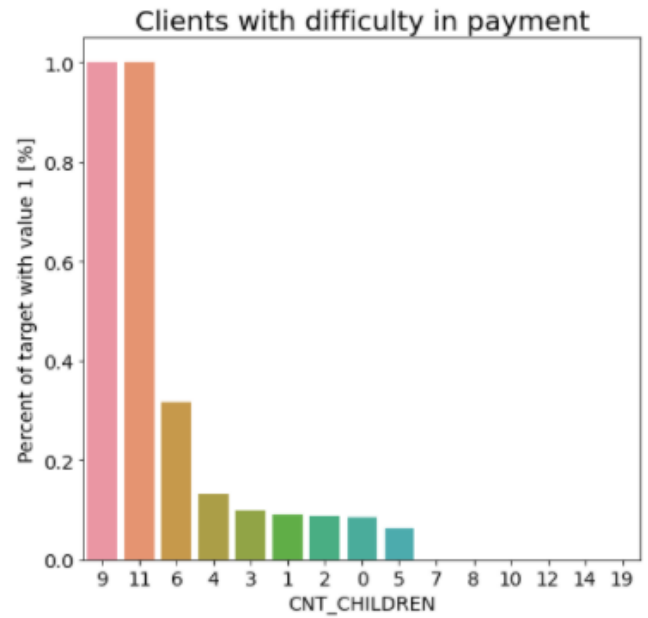


Observations for CNT_CHILDREN

- Most clients with loans have no children, followed by 1 child, 2 child while clients with more than 5 children are the least with loans.

- Clients with higher number of children have difficulties in loan repayment, whilst clients with 9 and 11 children are most likely not repay the loan.

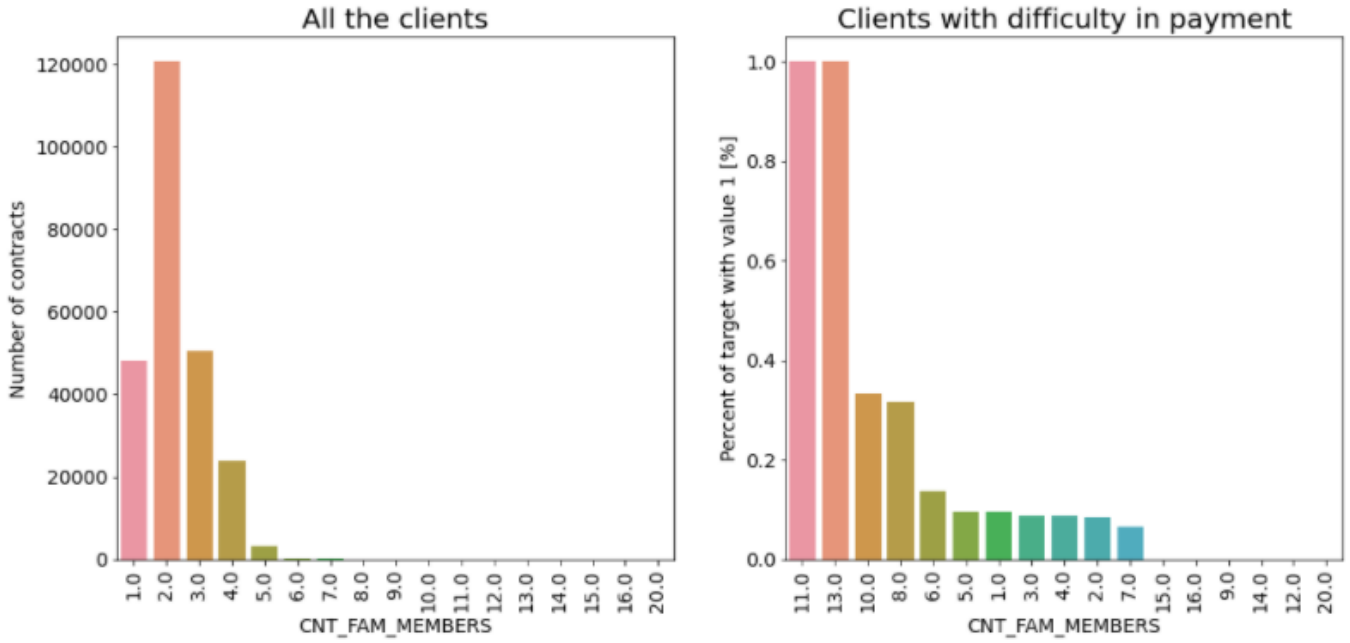




Observations for CNT_FAM_MEMBERS

- Clients with 2 family members are the most with loans, followed by single people an family of 3 etc.,.

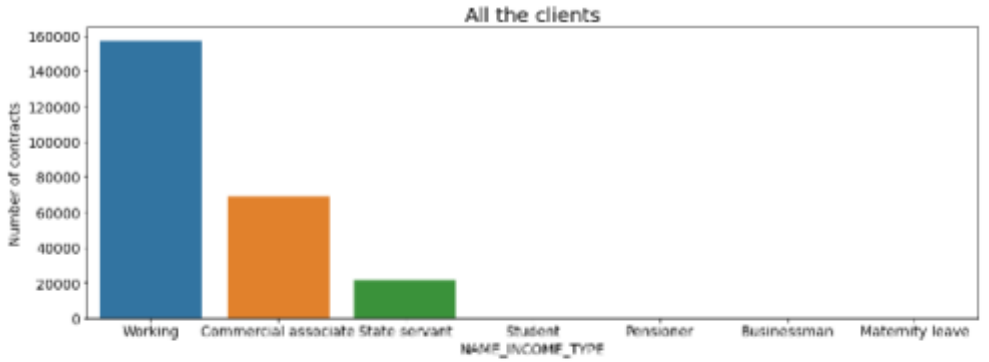
- In terms of repayment, family with 11,13 have 100% non repayment while family with 6,8 members is over 30% non repayment.

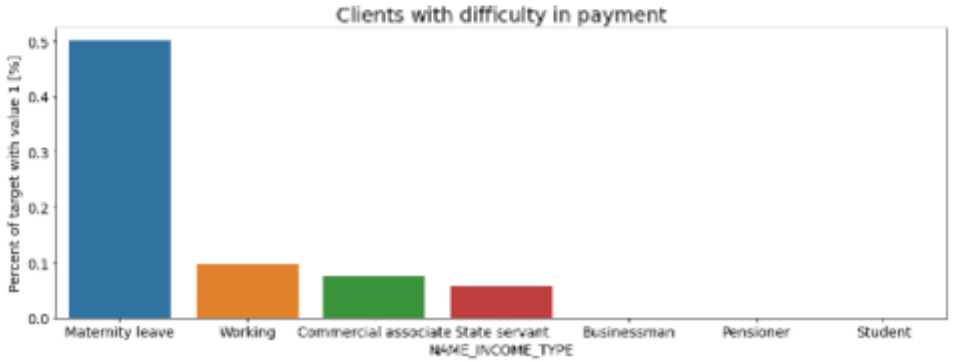


Observations for NAME_INCOME_TYPE

- Most clients are working, followed by commercial associate, Pensioner, stateservent whilst maternity leave clients are the least.

- Maternity leave clients followed by unemployed clients have highest percentage almost at 40% of non repayments of their loans.

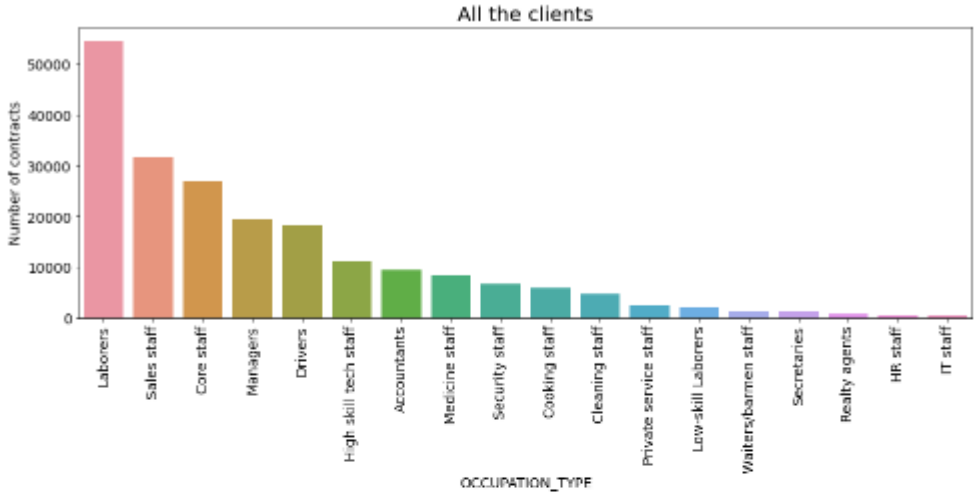


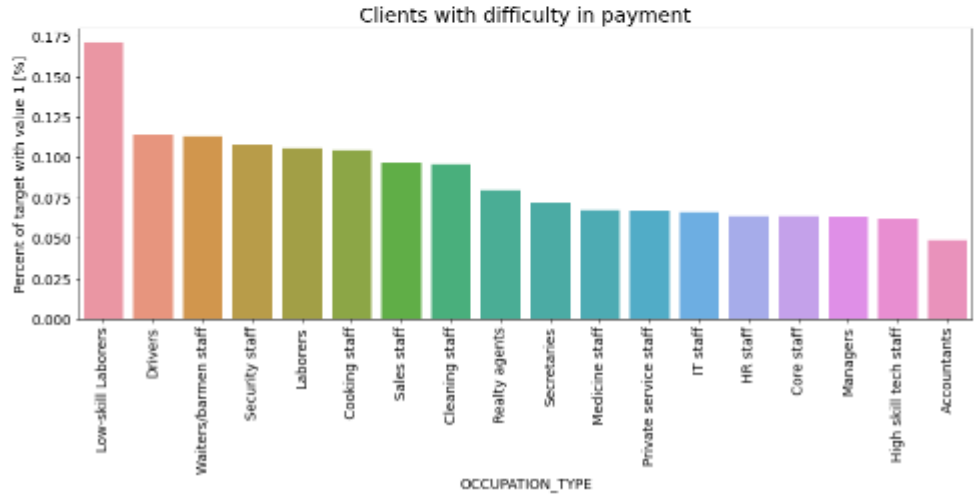


Observations for OCCUPATION_TYPE

- Laborers are the clients with highest loans taken while clients from IT staff are the least

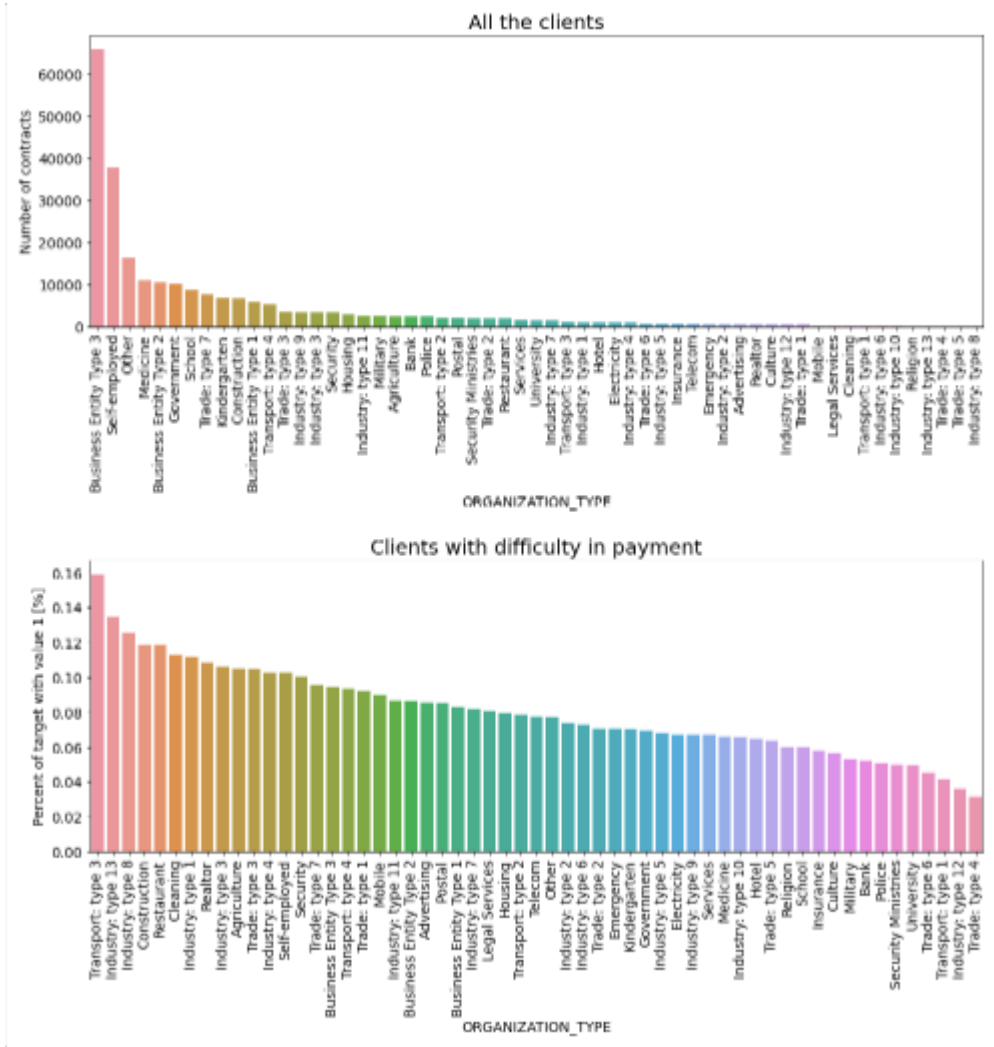
- In terms of non repayment of the loan, low-skill laborers are the highest at 17.5% followed by Drivers, Waiters, security staff and the least being accountants at around 50%





Observations for ORGANIZATION_TYPE

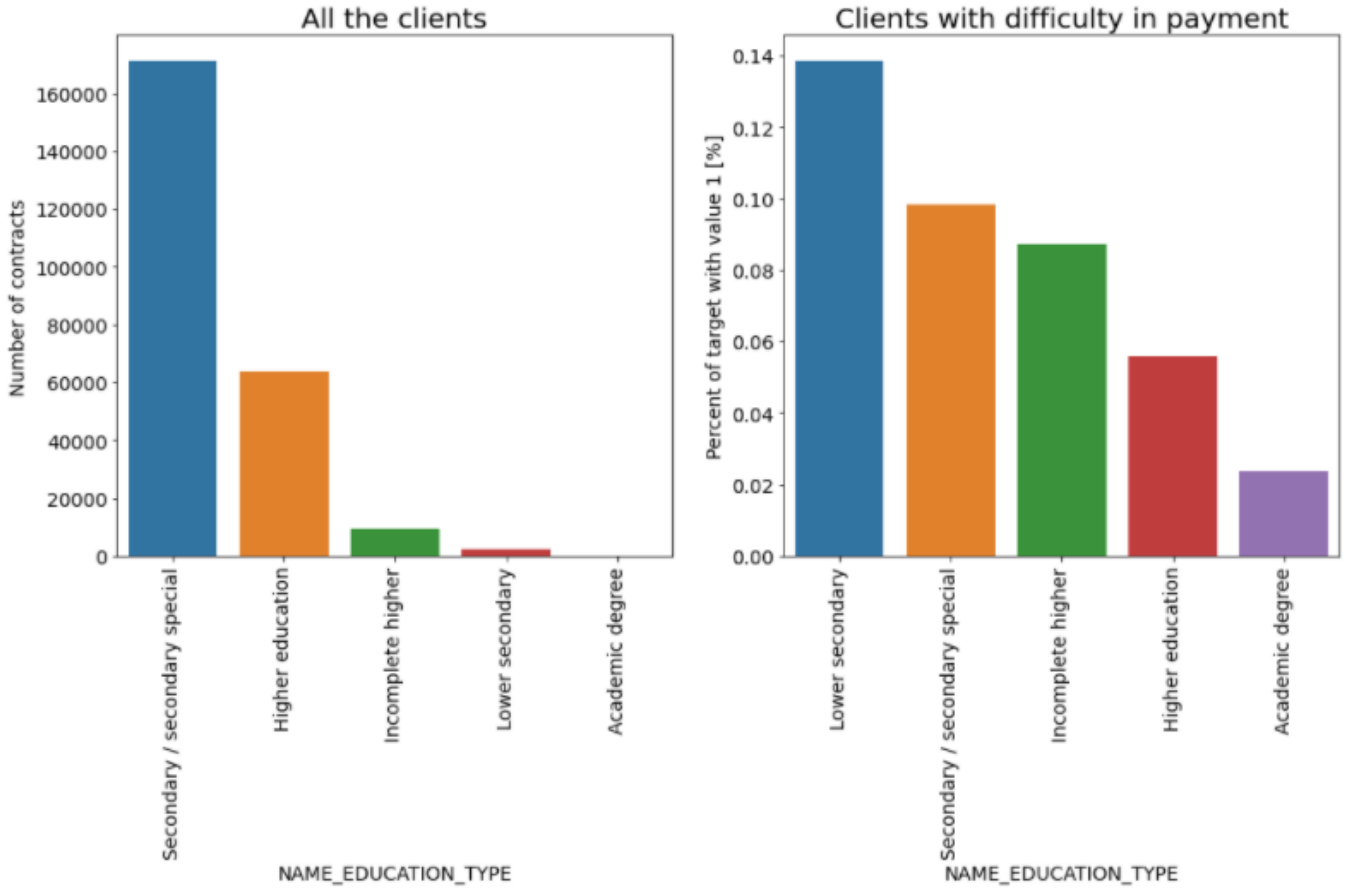
- Transport type3 followed bt industry type1, industry type8 have the highest percentage of loan non repayment while trade type 4 is the least percentage of loan non repayment around 0.4%



Observations for NAME_EDUCATION_TYPE

- Most clients with loan are secondary education followed by higher education while academic degree is the least clients.

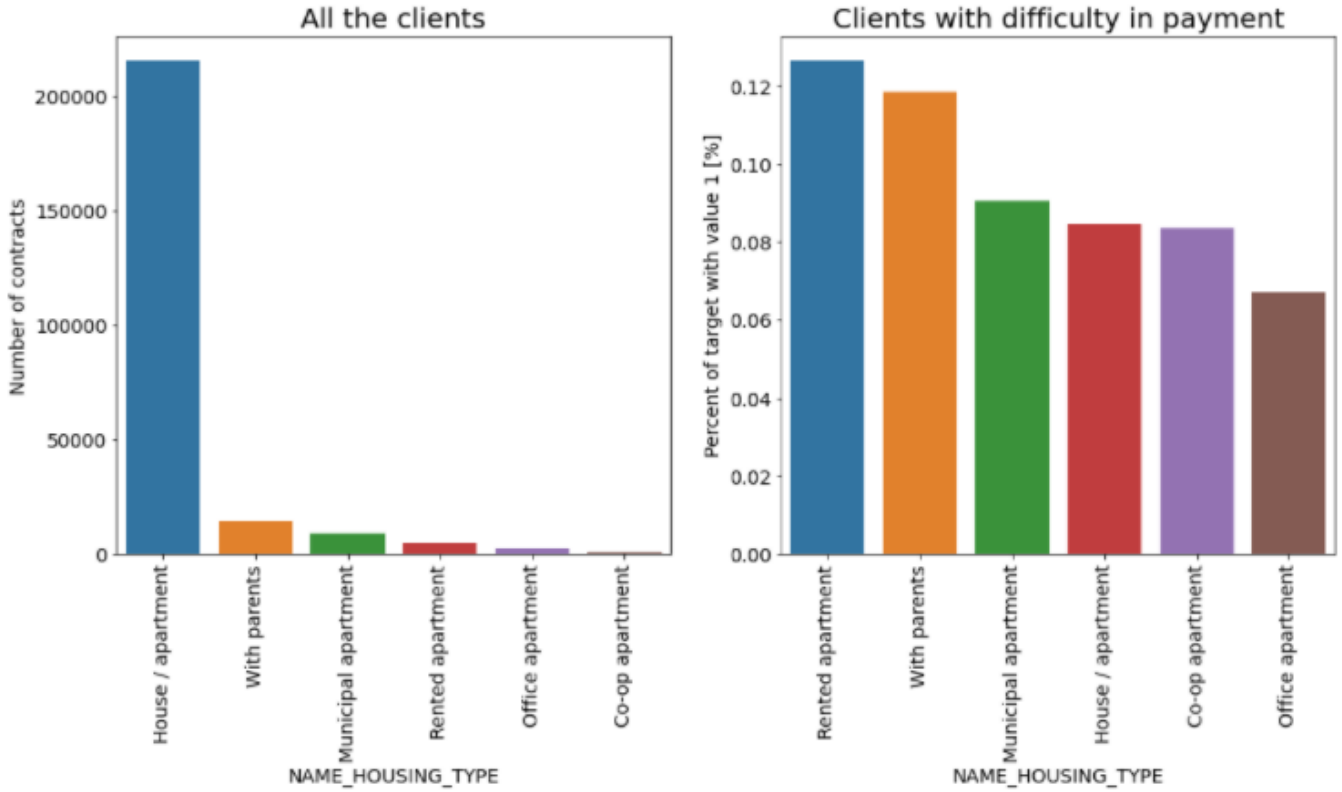
- Academic degree clients have lowest percentage of loan non repayment around 2% while lower secondary clients have the highest percentage around 11%



Observations for NAME_HOUSING_TYPE

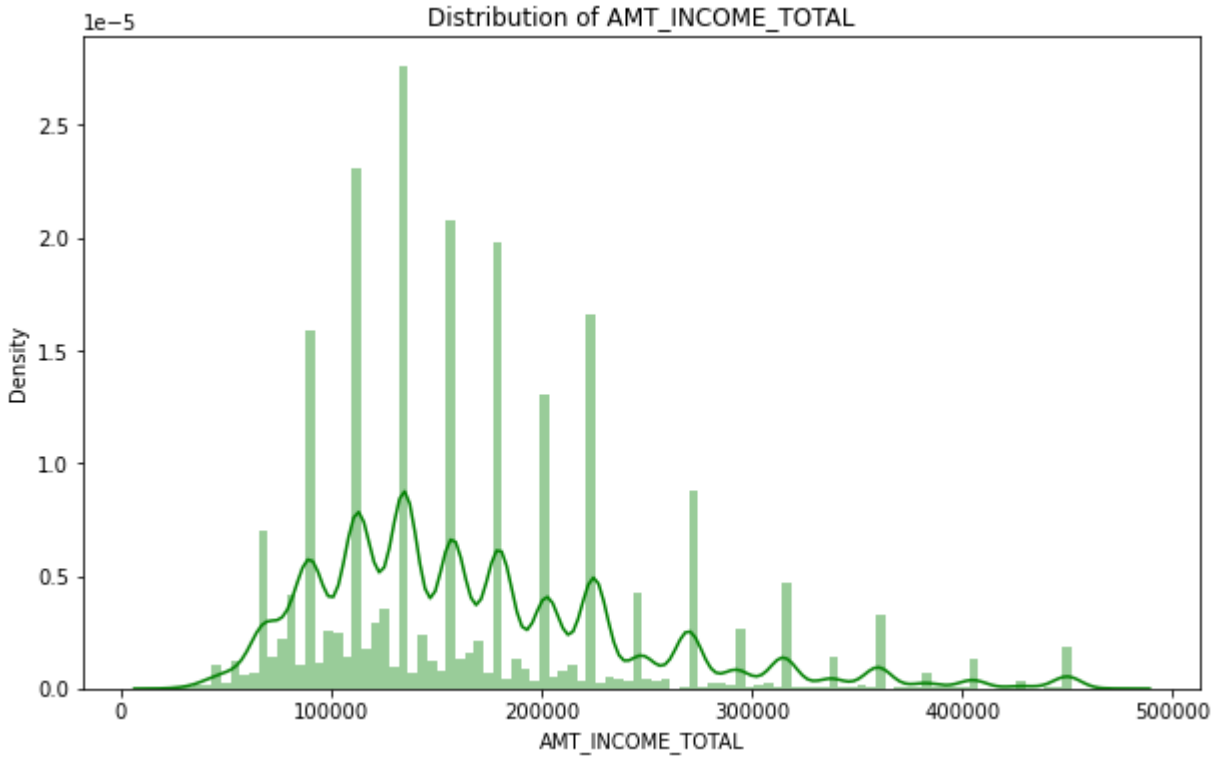
- Apartments are the highest housing type loans taken by the clients.

- Clients with rented apartments have the most difficulty in loan repayments at 12%



Observations of AMT_INCOME_TOTAL

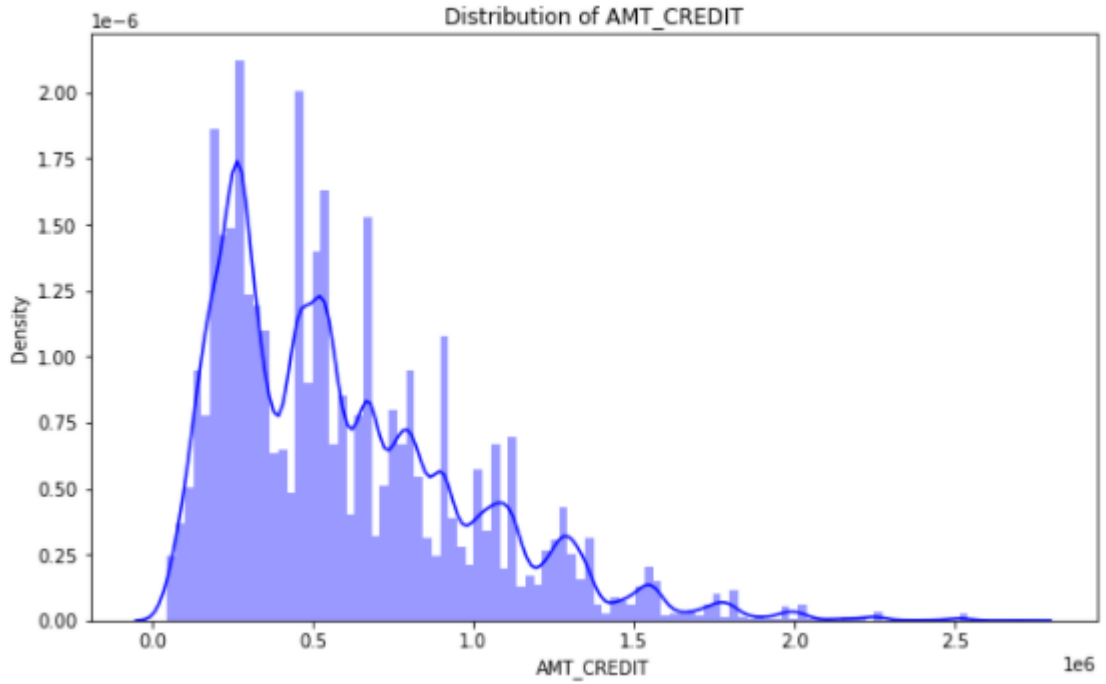
- AMT_INCOME_TOTAL is a right skewed distribution
- More clients have total income between 80000 and 250000.



Observations for AMT_CREDIT

- This plot is a right skewed distribution

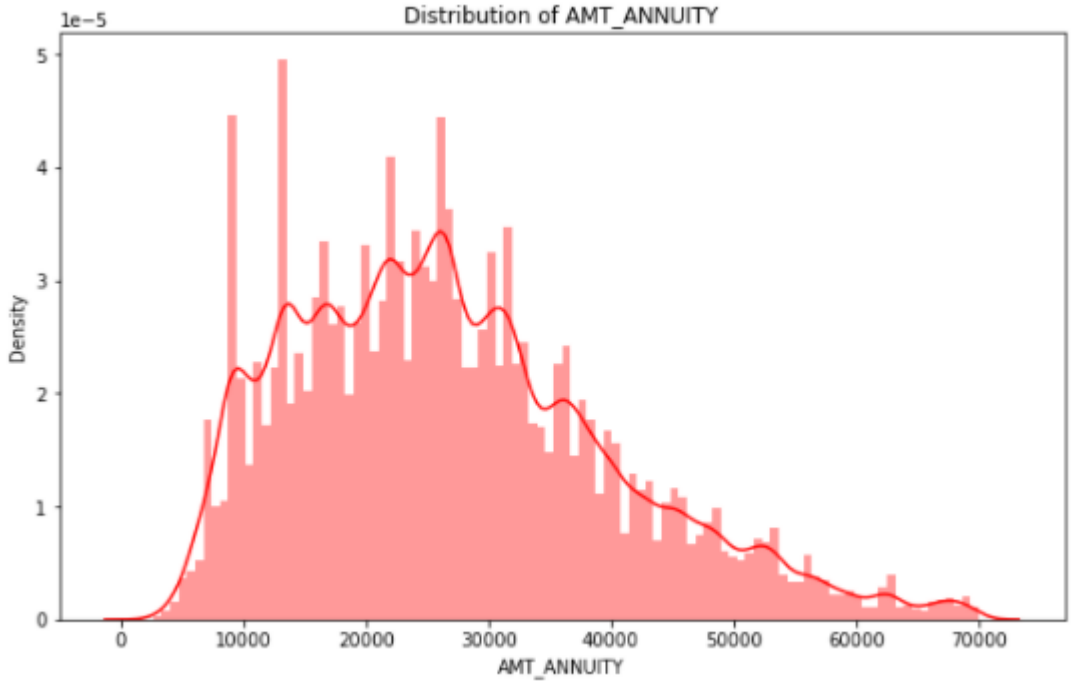
- More clients have high credit between 0 to 1e6



Observations for AMT_ANNUITY

- This plot is a right skewed distribution

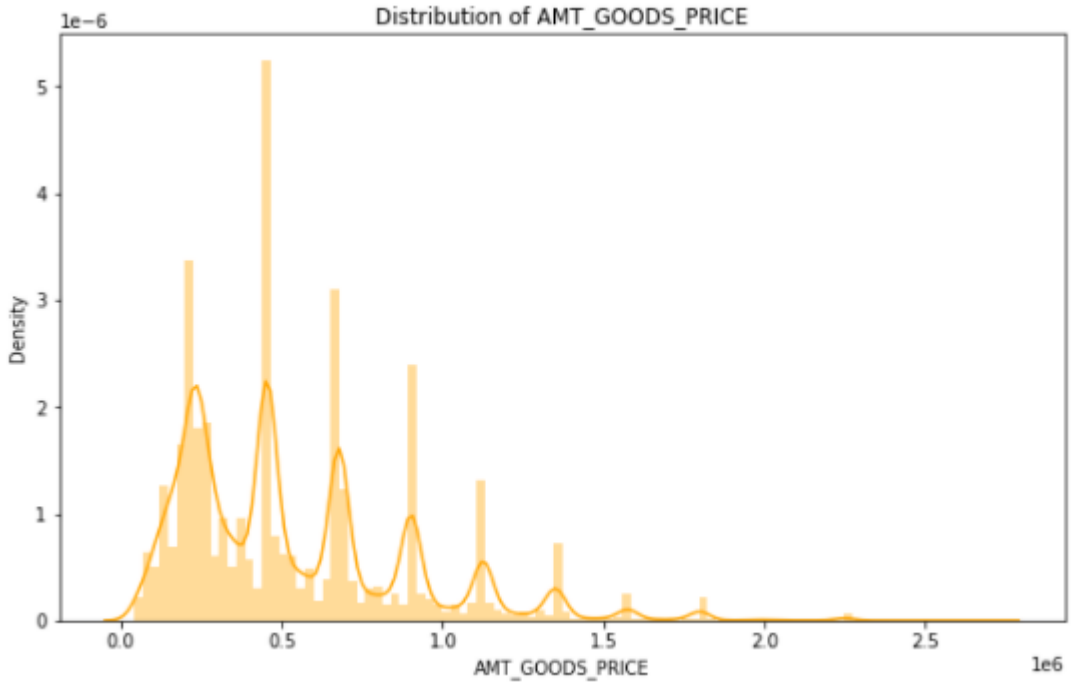
- More clients have high credit between 0 to 50000



Observations for AMT_GOODS_PRICE

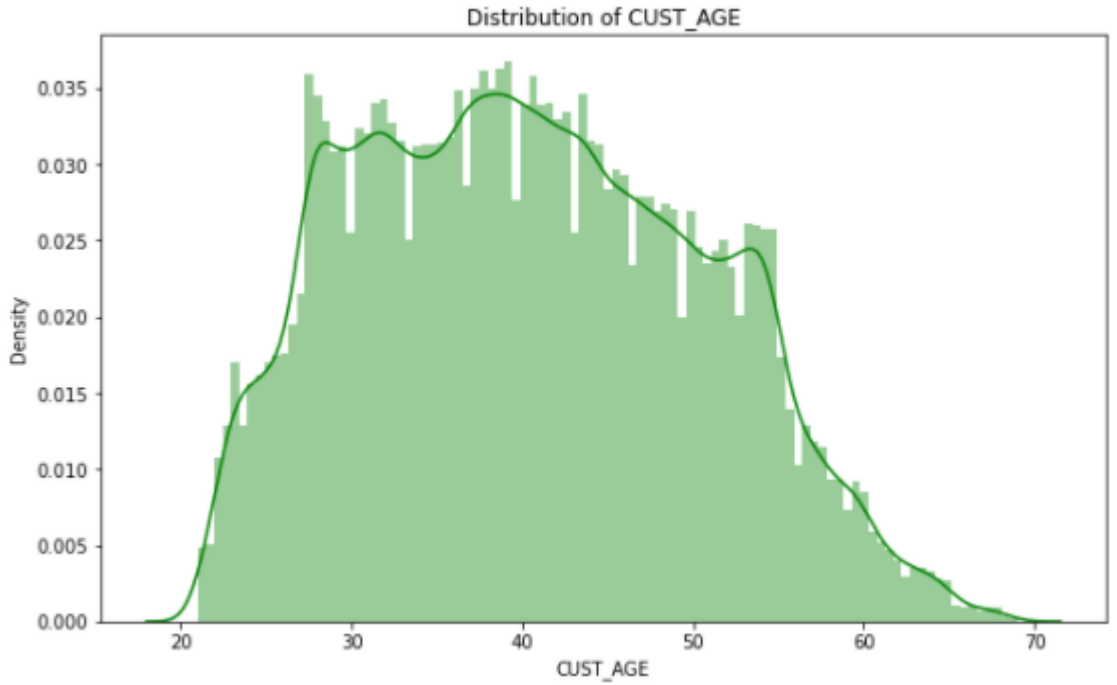
- Also a right skewed distribution

- We can observe that with increase in goods price the density decreases



Observations for CUST_AGE

- We see a sharp increase in clients from ages 20-30 and an exponential drop from ages 55-70.

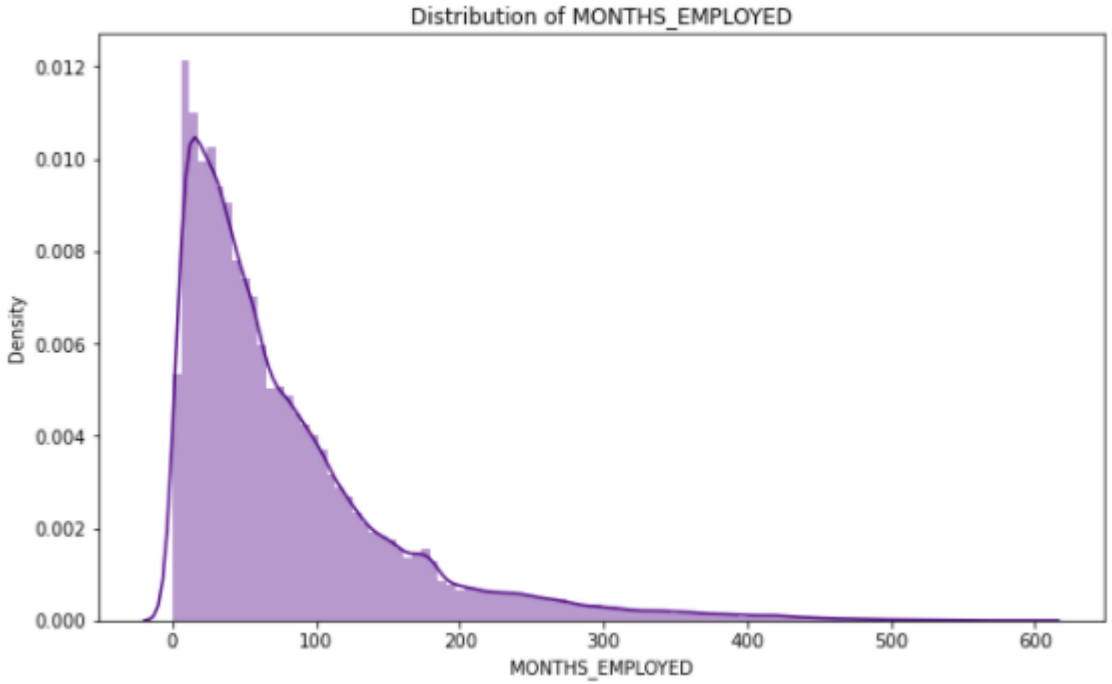


Observations for MONTHS_EMPLOYED

- The plot is right skewed

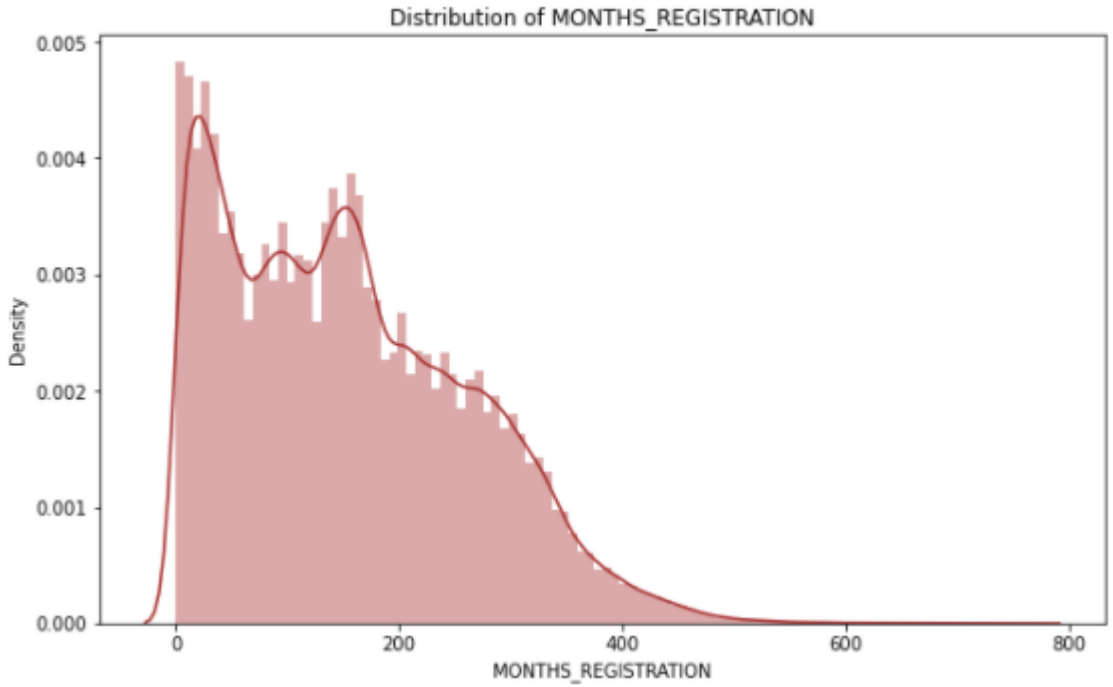
- The peak values in the beginning indicates credit being given for newly employed clients.

- It also suggests that the more months the client has been employed, the less likely they have a loan.



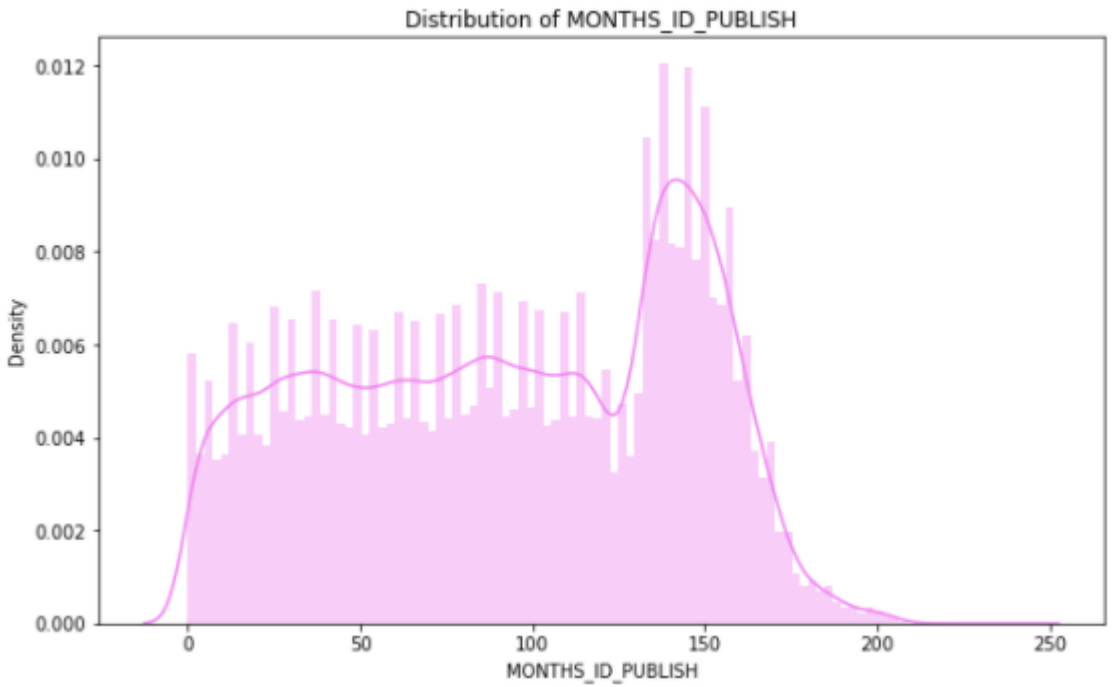
Observations for MONTHS_REGESTRATION

- There are more number of clients that had changed their regestration recently than the ones that have not



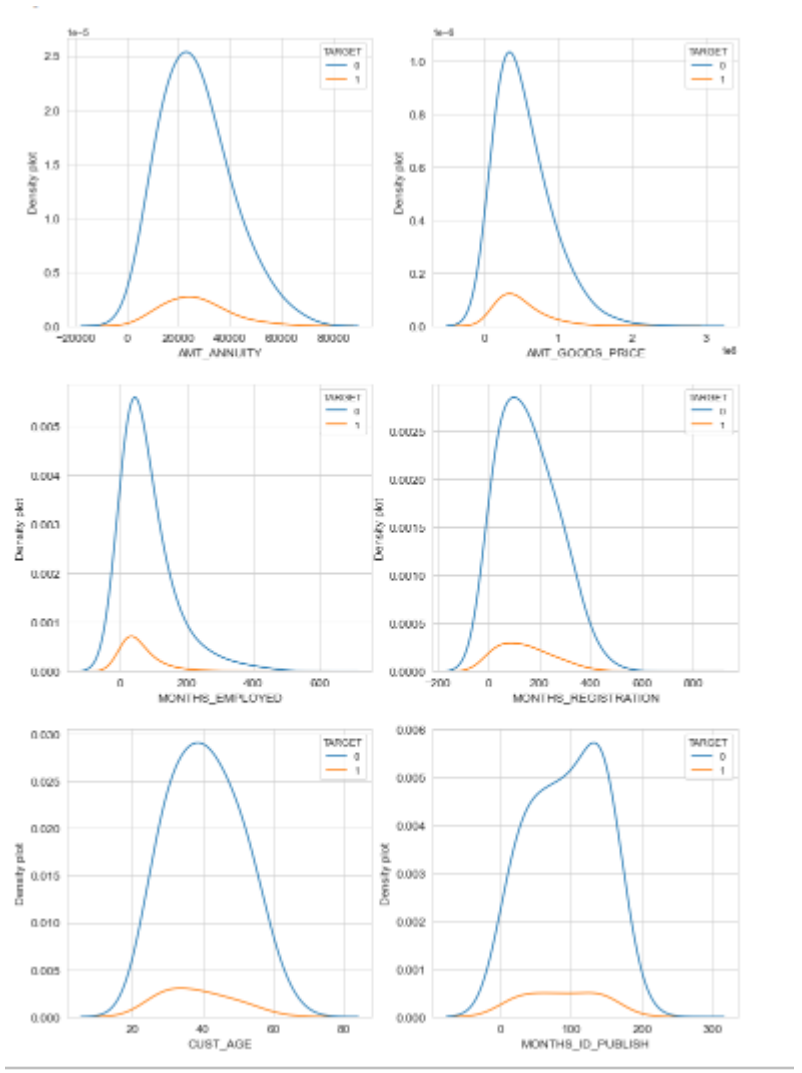
Observations for MONTHS_ID_PUBLISH

- We see a peak at 150 months indicating there is a change in identity published.



Observations from Comparision of target0 ant target1 with interval values

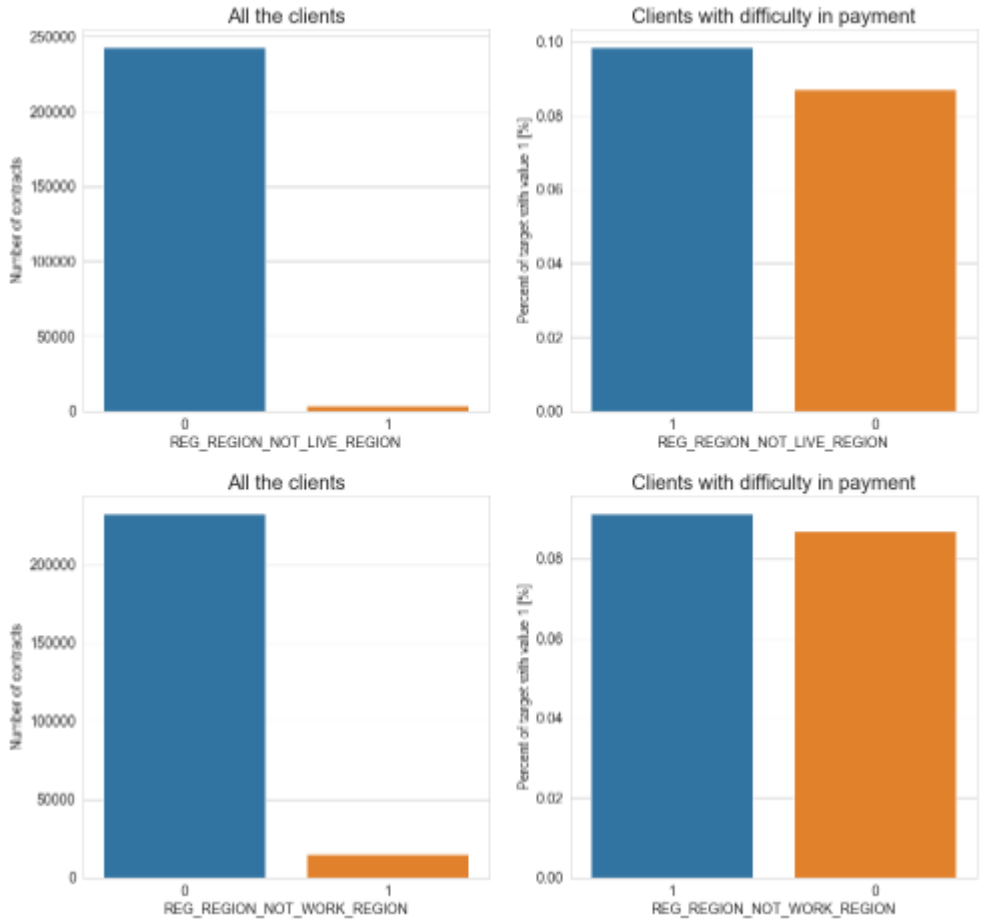
- Density plots of target1 follows the skewness of density plot of target0 for all the above columns



Observations for REG_REGION_NOT_LIVE_REGION, REG_REGION_NOT_WORK_REGION

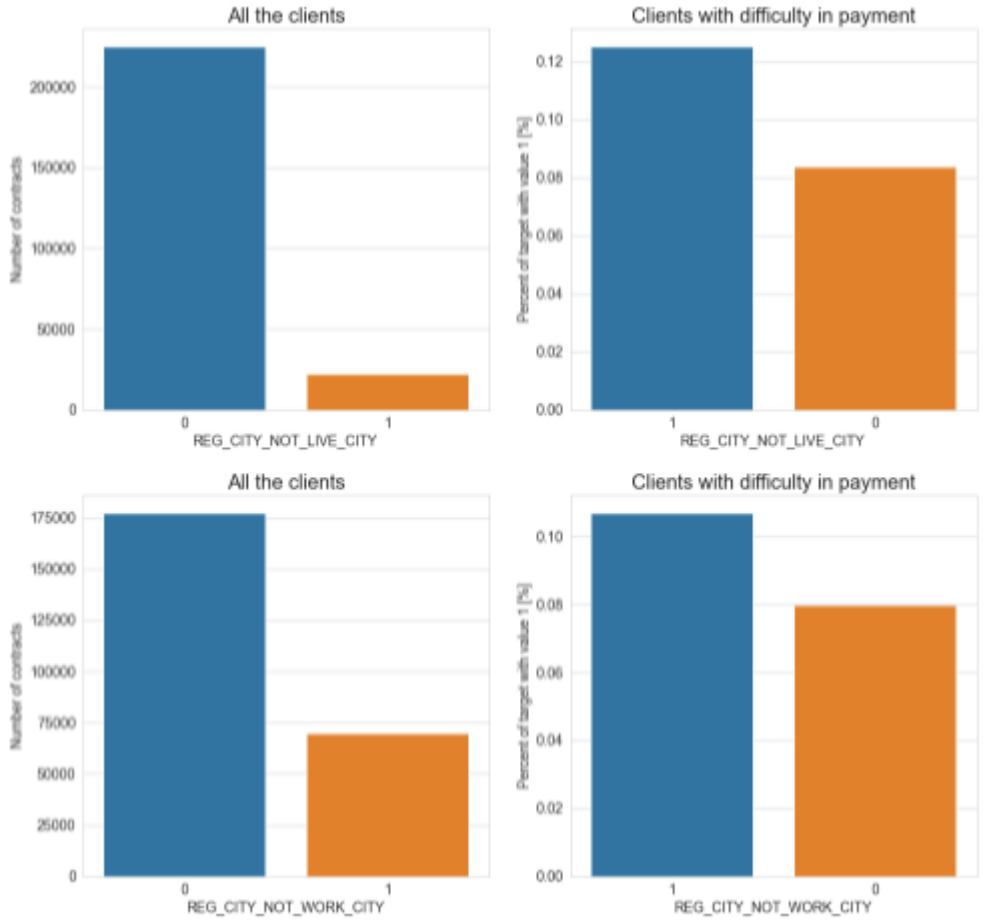
- Both in not live and not work region, very few clients are registered

- The percentage of clients not repaying the loan is slightly higher than the percentage of oan repayment



Observations for REG_CITY_NOT_LIVE_CITY, REG_CITY_NOT_WORK_CITY

- The clients registered in a different city than the working city or living city have higher percentage of loan non repayment



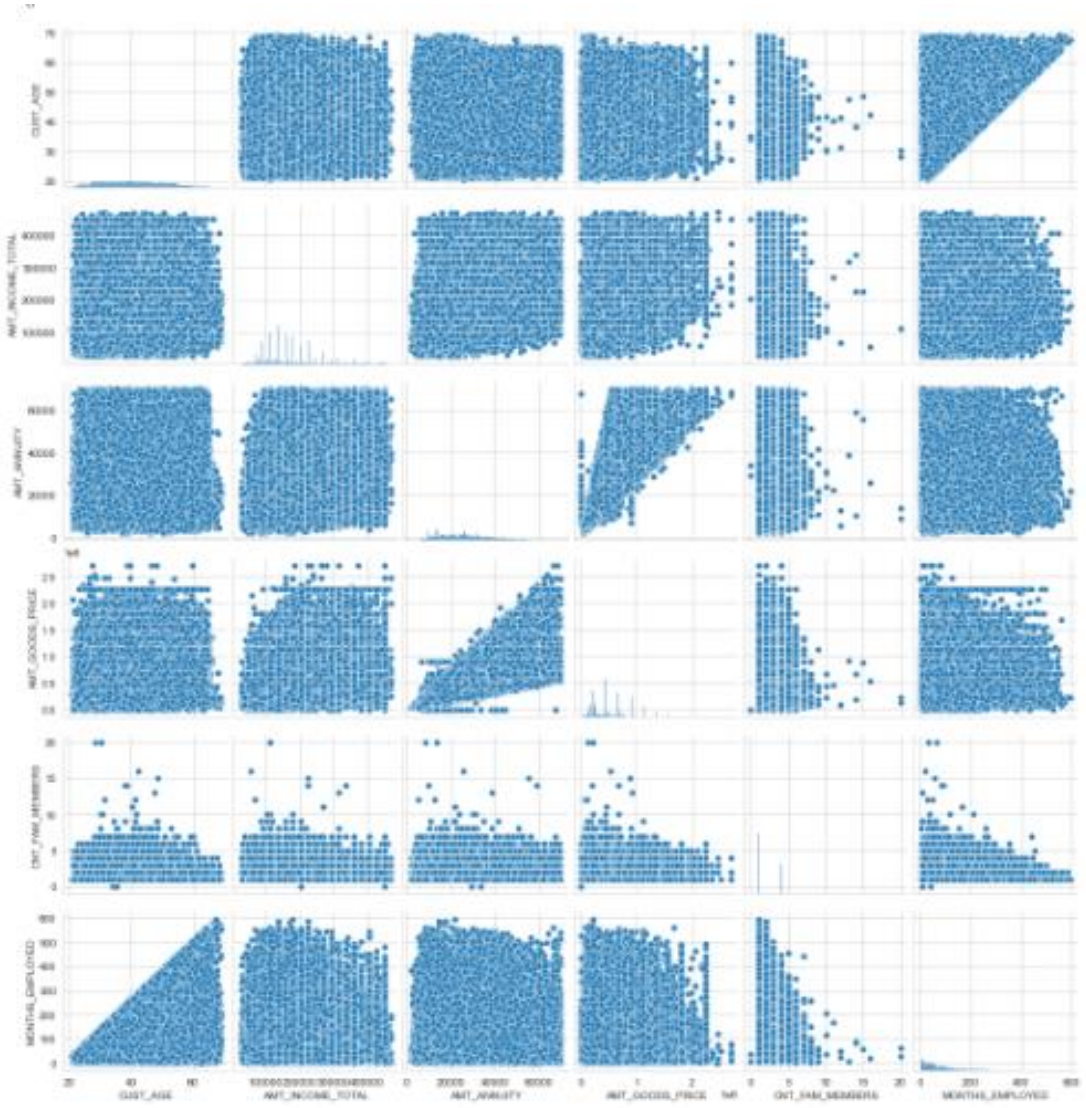
Observations for application data

BIVARIATE ANALYSIS

Observations for pair plot

- There is a high correlation between amount annuity and good price.

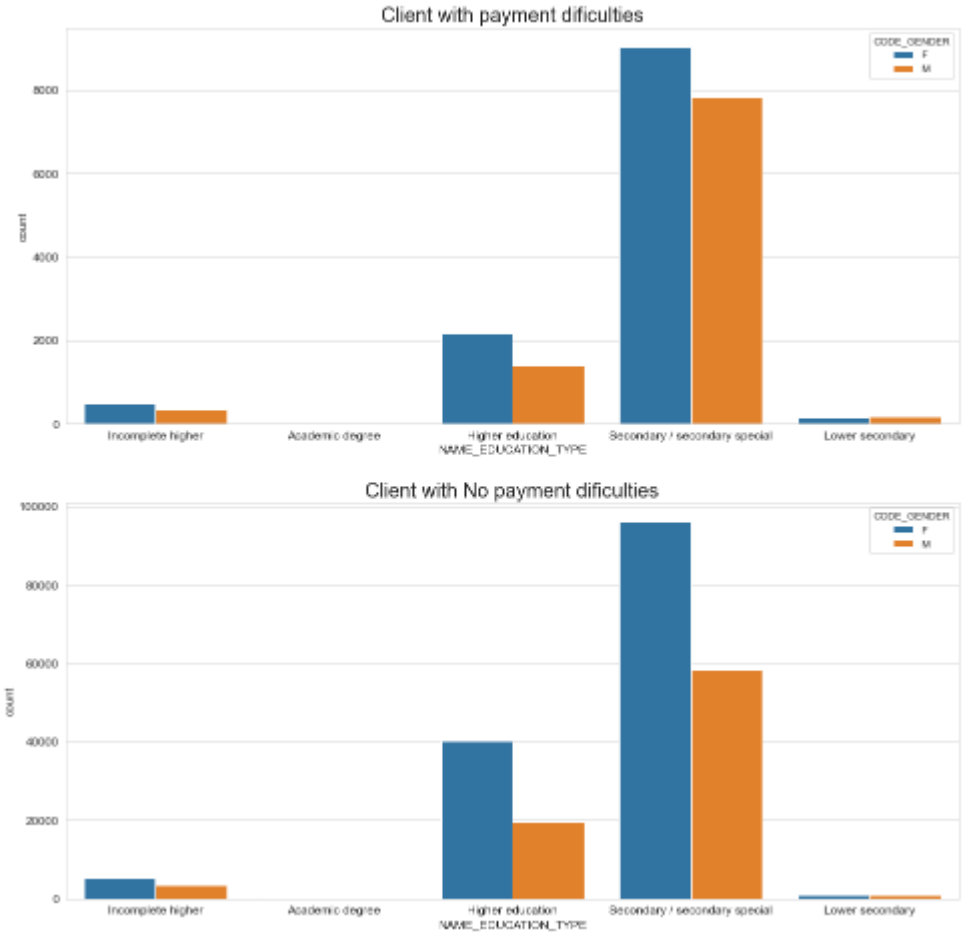
- There is also a correlation between months employed and customer age.



Observations for NAME_EDUCATION_TYPE, CODE_GENDER

- Both male and female clients with secondary education type have the highest loan non repayment count

- In loan repayments, male clients count is almost half of female clients count



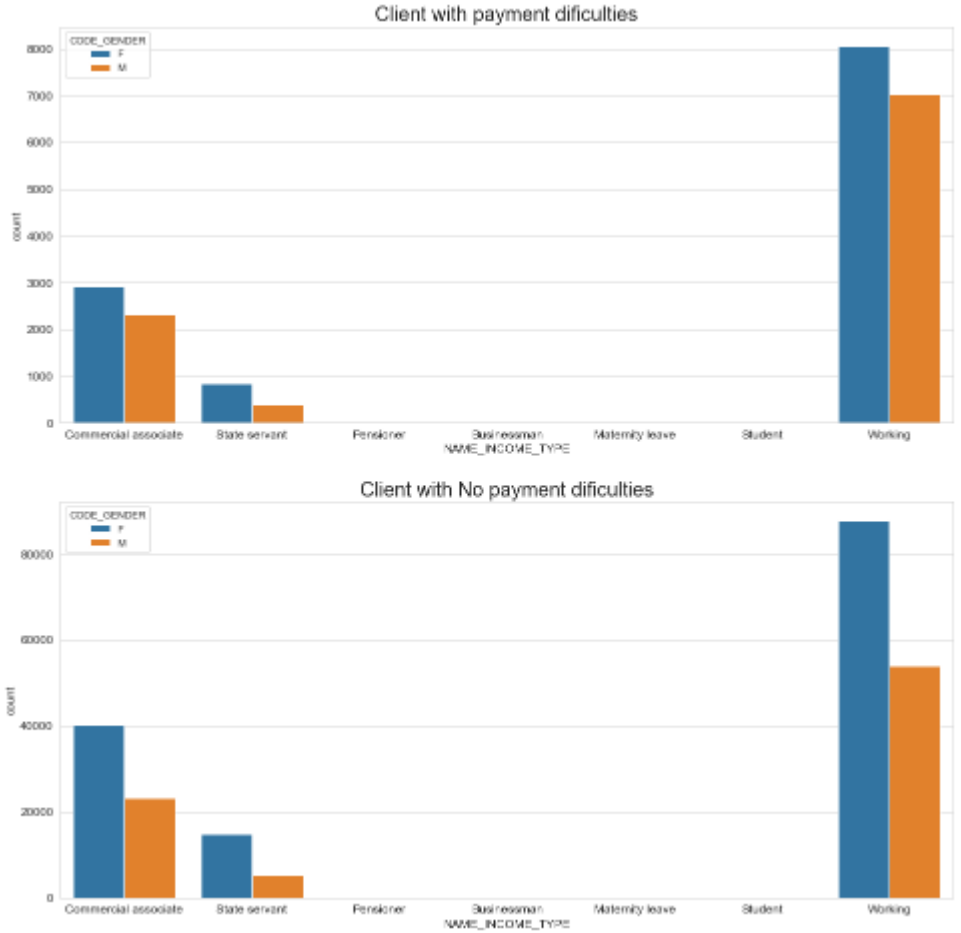
Observations for NAME_INCOME_TYPE, CODE_GENDER

- For income type working, commercial associate, and State Servant the number of credits are higher than other i.e. Maternity leave.

- For this Females are having more number of credits than male.

- Less number of credits for income type Maternity leave.

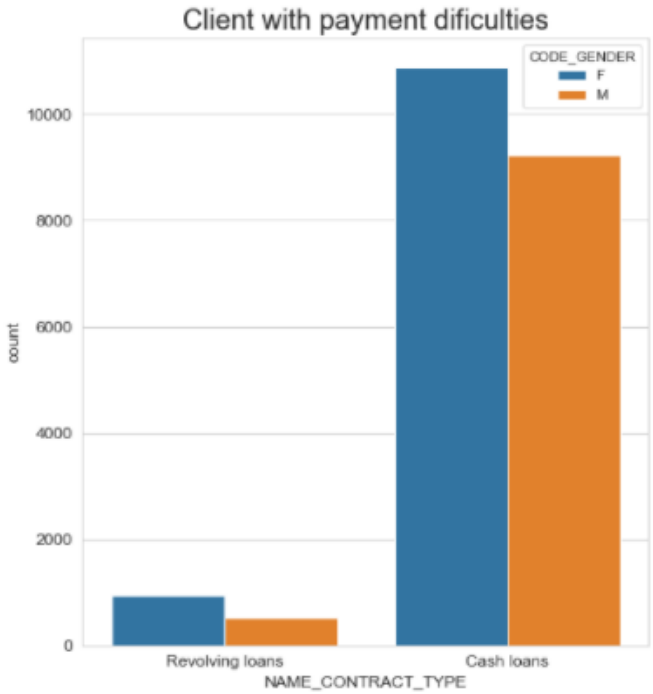
- There is no bar for student , pensioner and Businessman which means they don’t do any late payments.

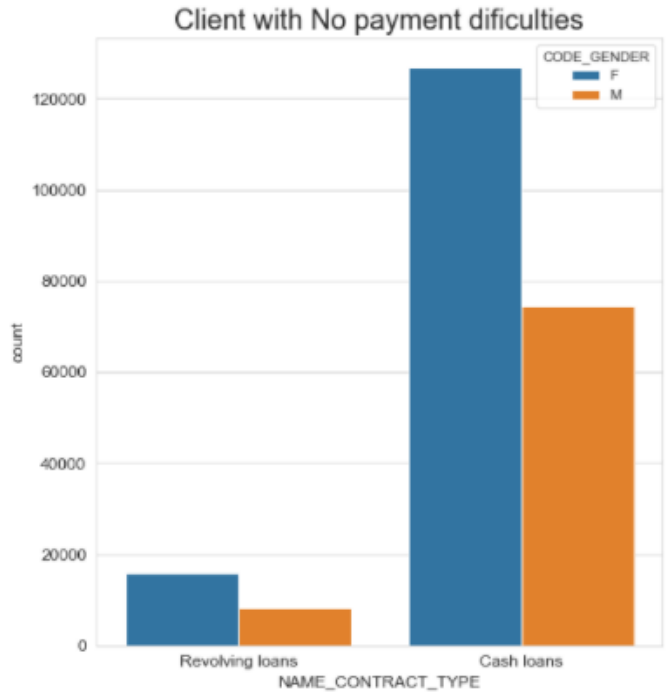


Observations for NAME_CONTRACT_TYPE, CODE_GENDER

- Cash loans of male clients without payment difficulties is half of female clients without payment difficulties.

- Cash loans are the more availed by both clients with and without patment difficulties when compared to revolving loans.





Observations for FLAG_OWN_CAR, FLAG_OWN_REALTY

- There is no observable pattern difference between clients with and without payment difficulties

Observations for NAME_FAMILY_STATUS, CODE_GENDER

- Widow clients have more payment difficulties when compared to other clients.

- Female clients are more than male clients irrespective of their family status and whether they have payment difficulty or not

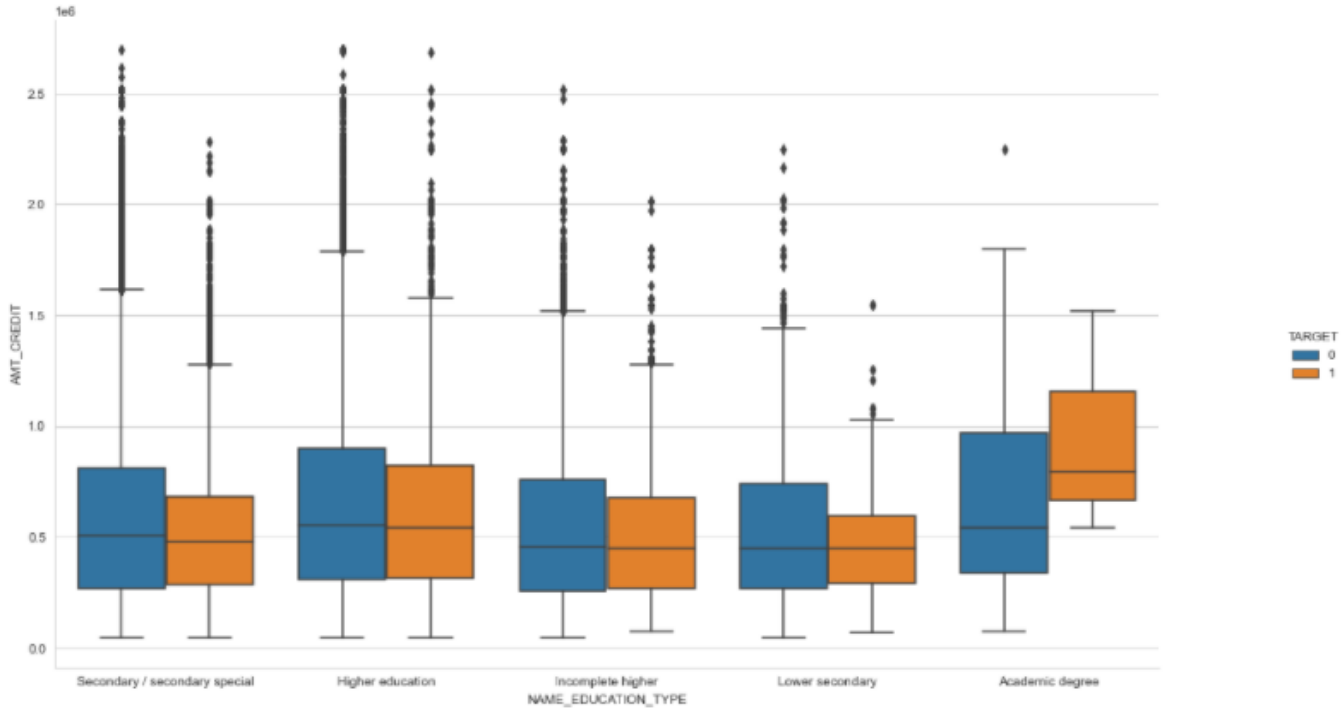
- Married female clients have more payment difficulties than married male clients

Observations for NAME_EDUCATION_TYPE, AMT_CREDIT

- The mean amount credited is almost same among all the education types for clients with and without payment difficulties

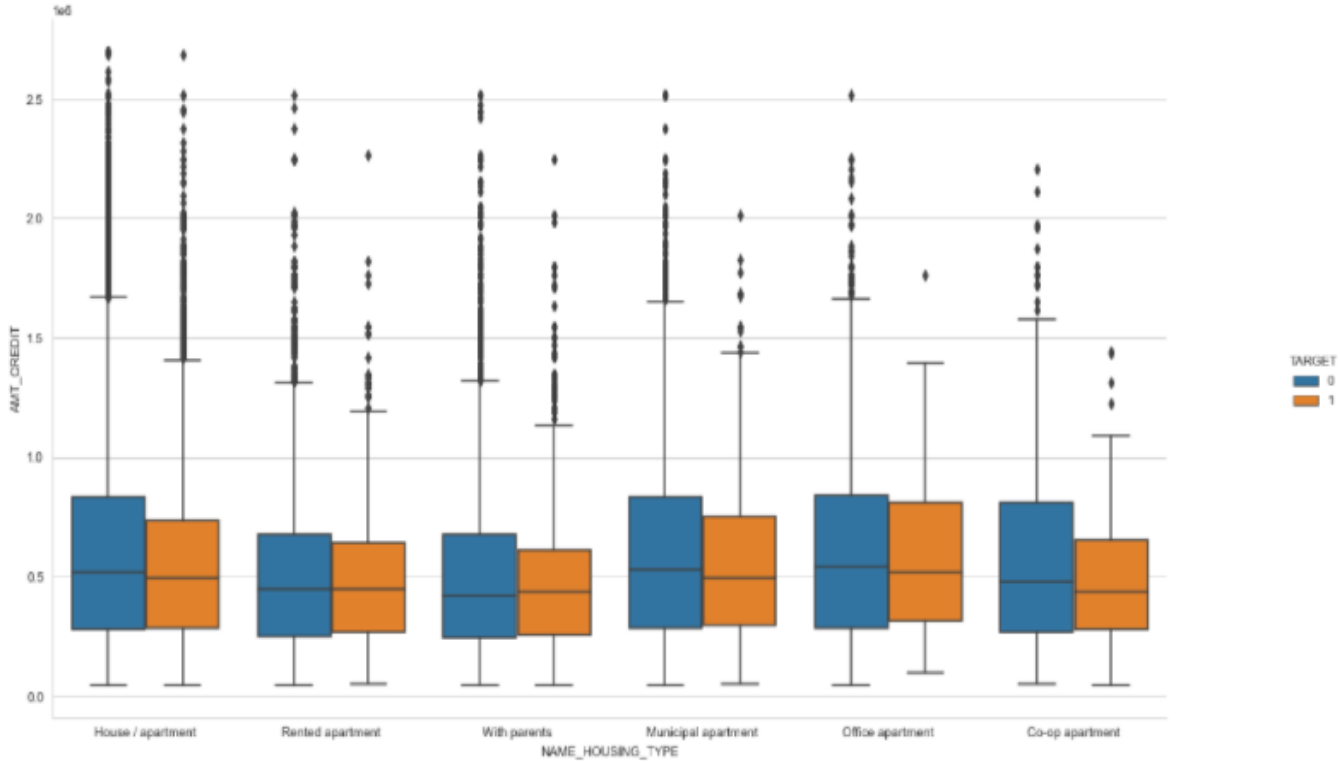
- Clients with higher education seem to have high credit

- Clients with Academic degree seem to have more difficulties for higher credit amount



Observations for NAME_HOUSING_TYPE, AMT_CREDIT

- All the clients have almost the same mean credit per housing type irrespective of the difficulty



Correlations of application data

Observations for correlations of appData

Observation

- The top 10 correlations are

1. OBS_60_CNT_SOCIAL_CIRCLE OBS_30_CNT_SOCIAL_CIRCLE ,

2. AMT_GOODS_PRICE AMT_CREDIT ,

3. REGION_RATING_CLIENT REGION_RATING_CLIENT_W_CITY ,

4. REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION ,

5. CNT_FAM_MEMBERS CNT_CHILDREN ,

6. DEF_30_CNT_SOCIAL_CIRCLE DEF_60_CNT_SOCIAL_CIRCLE,

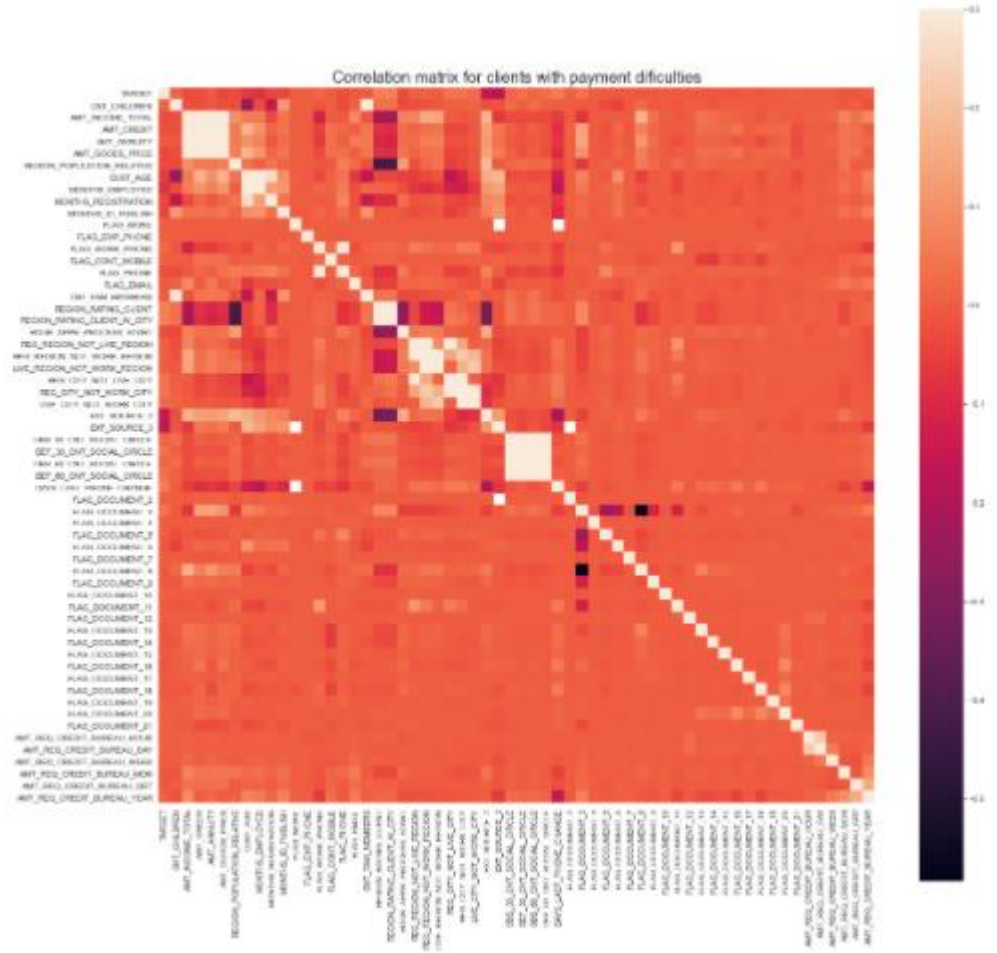
7. AMT_ANNUITY AMT_CREDIT ,

8. AMT_ANNUITY AMT_GOODS_PRICE,

9. LIVE_CITY_NOT_WORK_CITY REG_CITY_NOT_WORK_CITY ,

10. FLAG_DOCUMENT_3 FLAG_DOCUMENT_8.

- From top to bottom the correlation decreases



Top 10 pairs from application data in descending order of their correlation

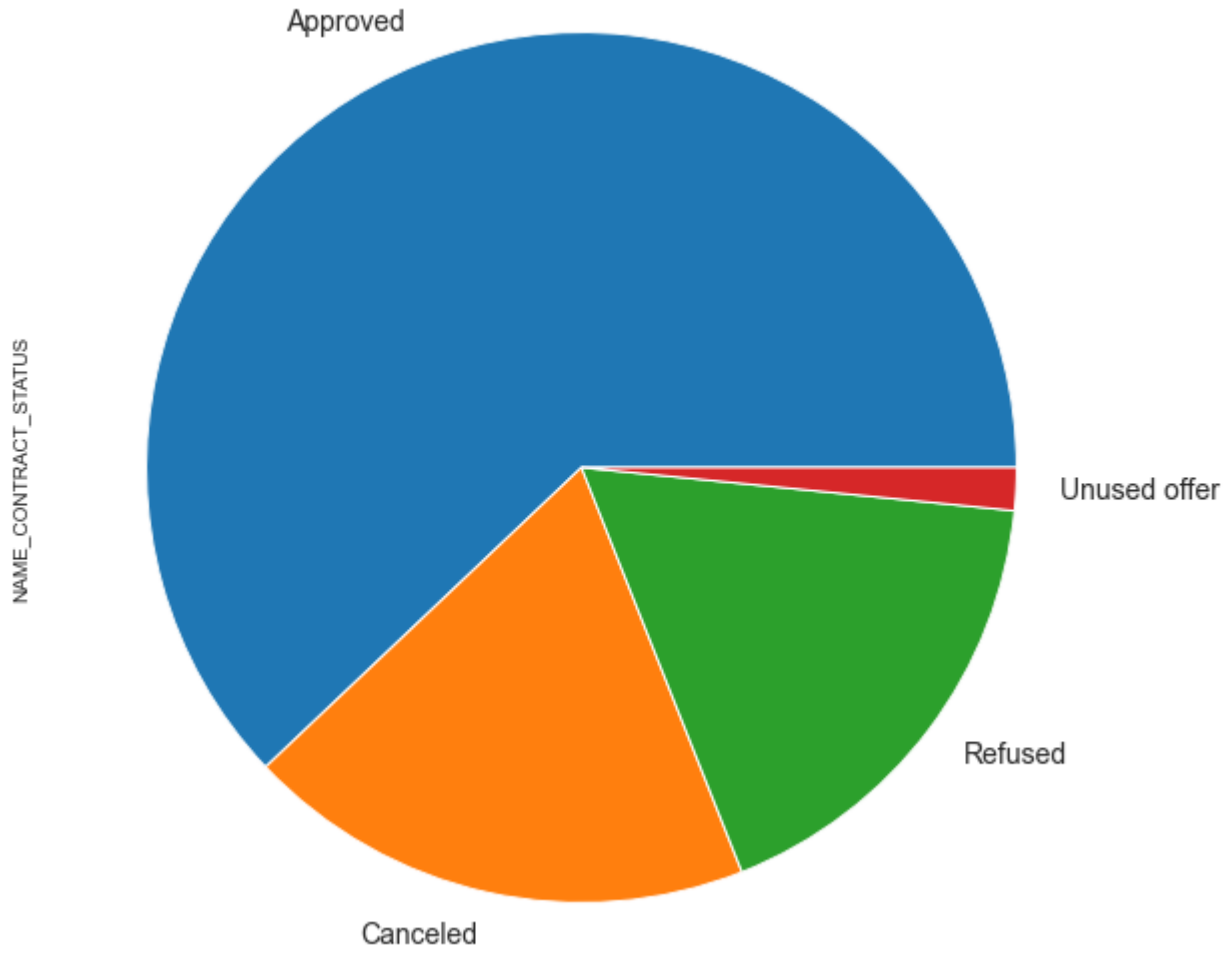
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE
AMT_GOODS_PRICE	AMT_CREDIT
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION
CNT_FAM_MEMBERS	CNT_CHILDREN
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE
AMT_ANNUITY	AMT_CREDIT
AMT_ANNUITY	AMT_GOODS_PRICE
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY
FLAG_DOCUMENT_3	FLAG_DOCUMENT_8.

Observations from Previous application data analysis

UNIVARIATE ANALYSIS

Observations for NAME_CONTRACT_STATUS

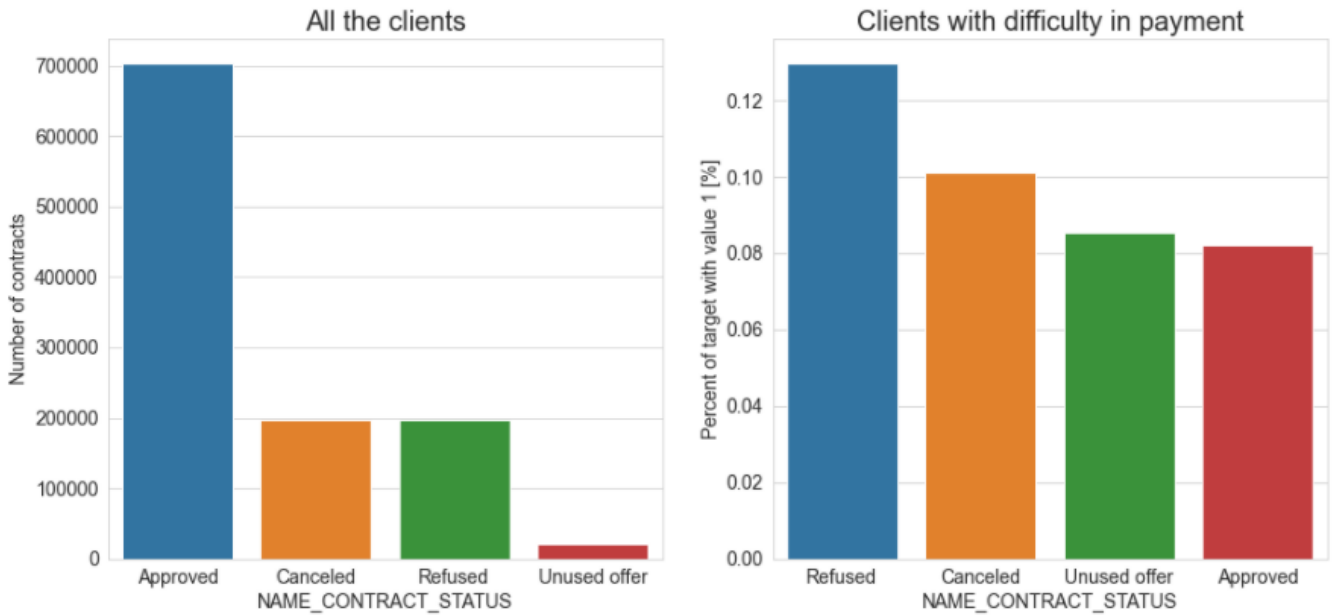
- Most of the clients loans are approved and very less number of clients have not availed their approved loans.



Observations for NAME_CONTRACT_STATUS

- Most of the loans in previous data are approved

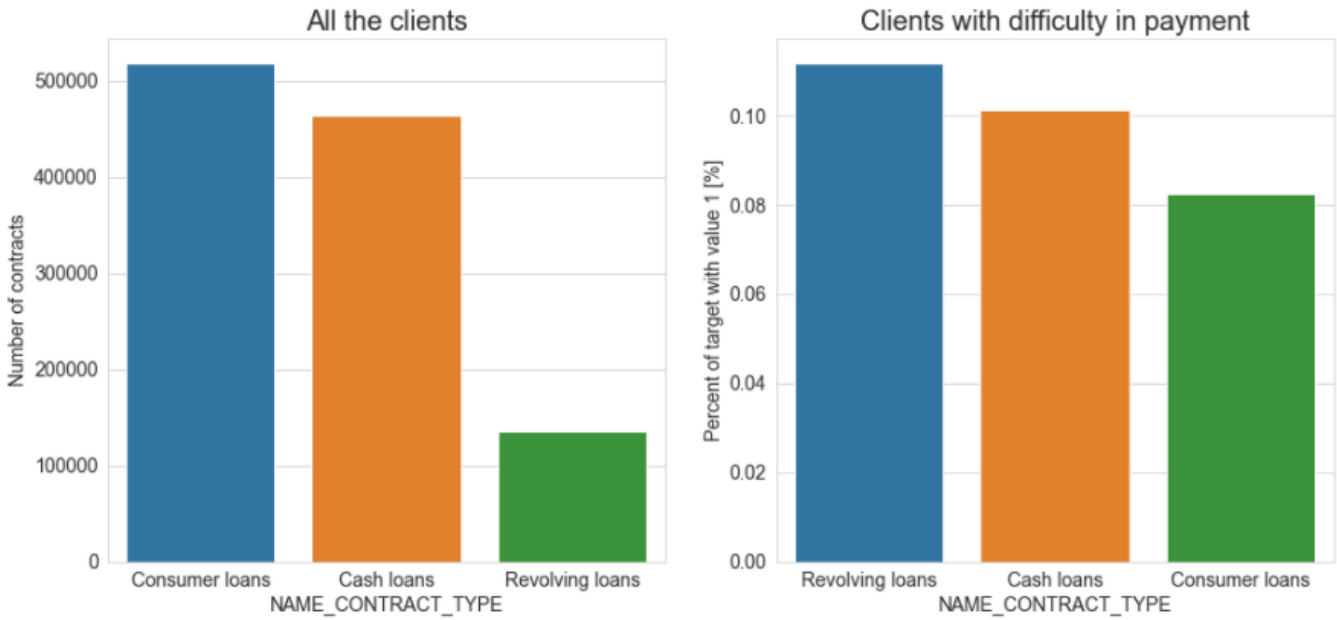
- Clients whose loans were refused previously have high difficulty in loan repayment when compared to clients with approved loan contracts.



Observations for NAME_CONTRACT_TYPE

- Most applications in previous applications data were of cash loans followed by consumer loans while revolving loans were the least.

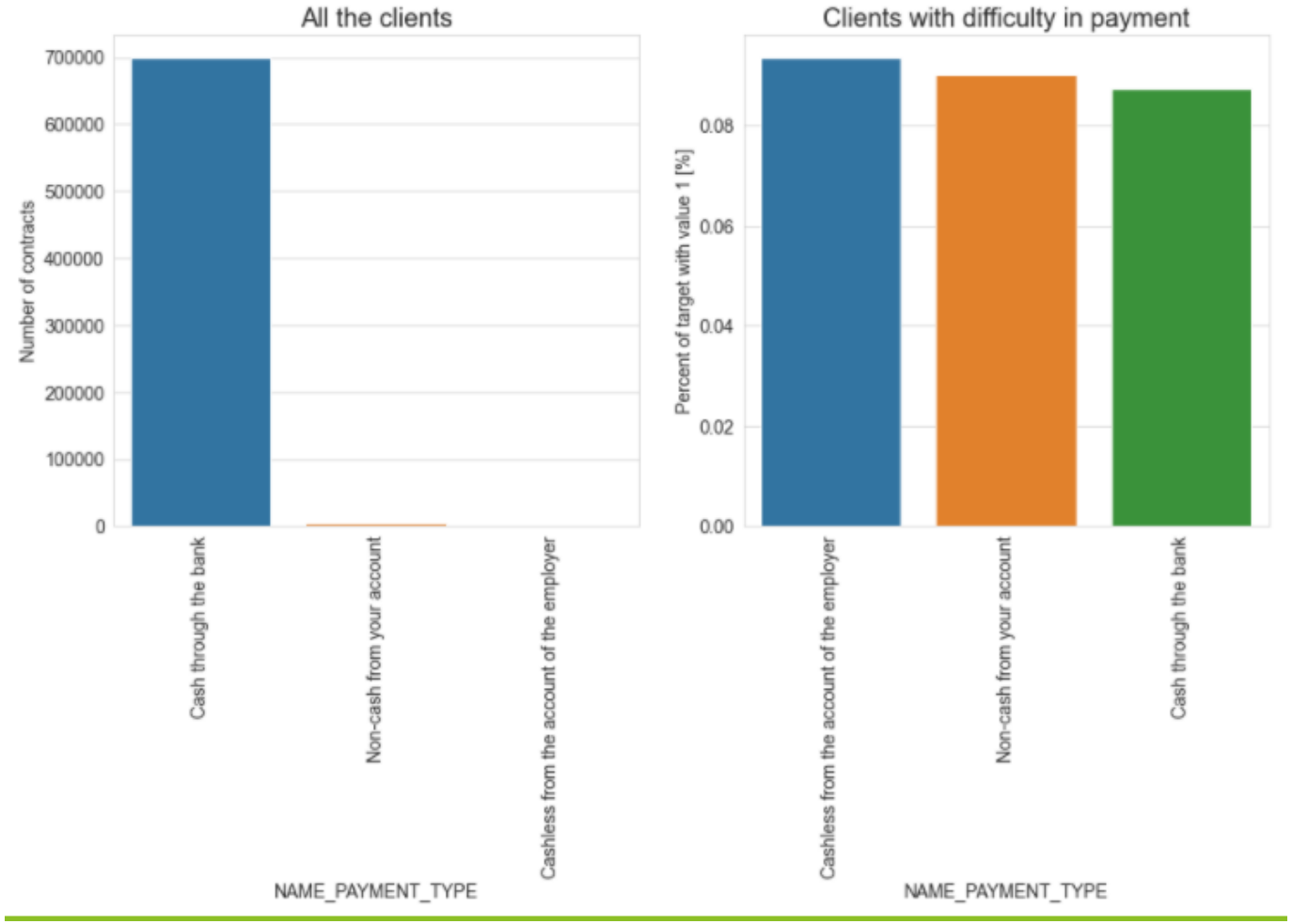
- Clients with revolving loans have high percentage of difficulty in loan repayments.



Observations for NAME_PAYMENT_TYPE

- Most popular payment type seems to be cash through bank.

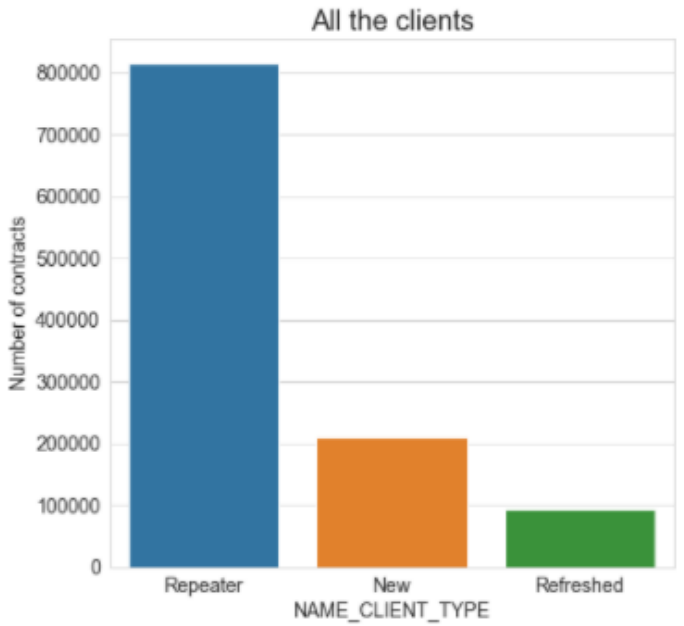
- It can be seen in the second graph that clients with all types of payment have similar percentage of difficulty in loan repayments

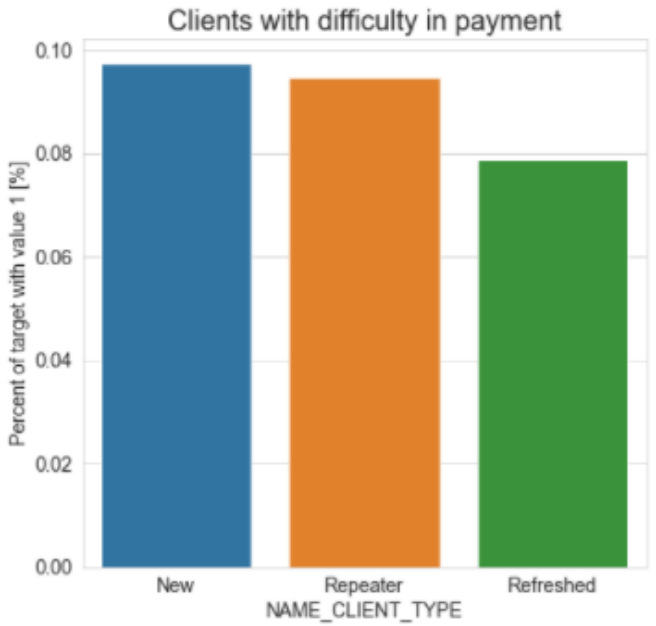


Observations for NAME_CLIENT_TYPE

- Most of the clients are repeaters

- From the second graph, new clients have more difficulty in loan repayments

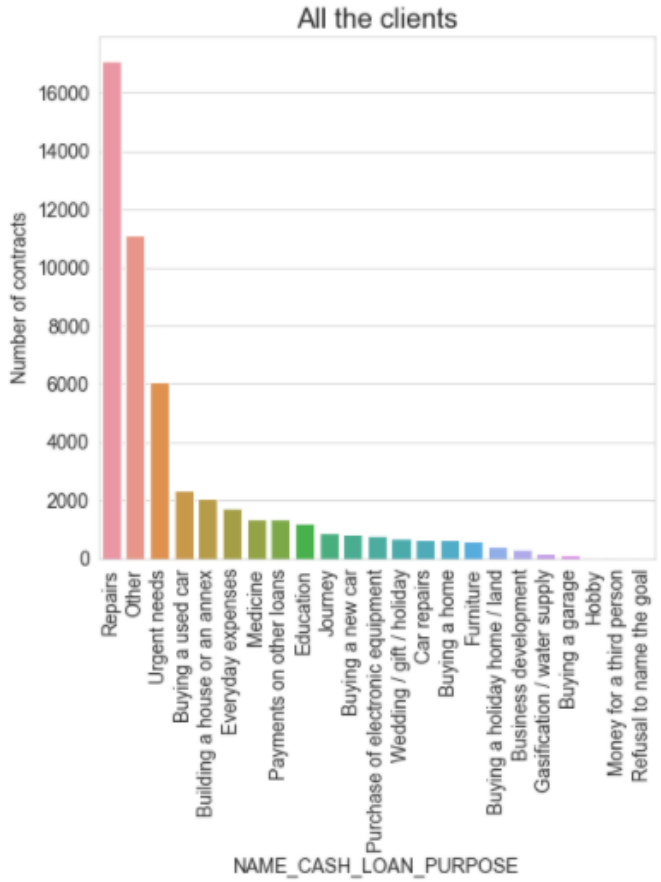


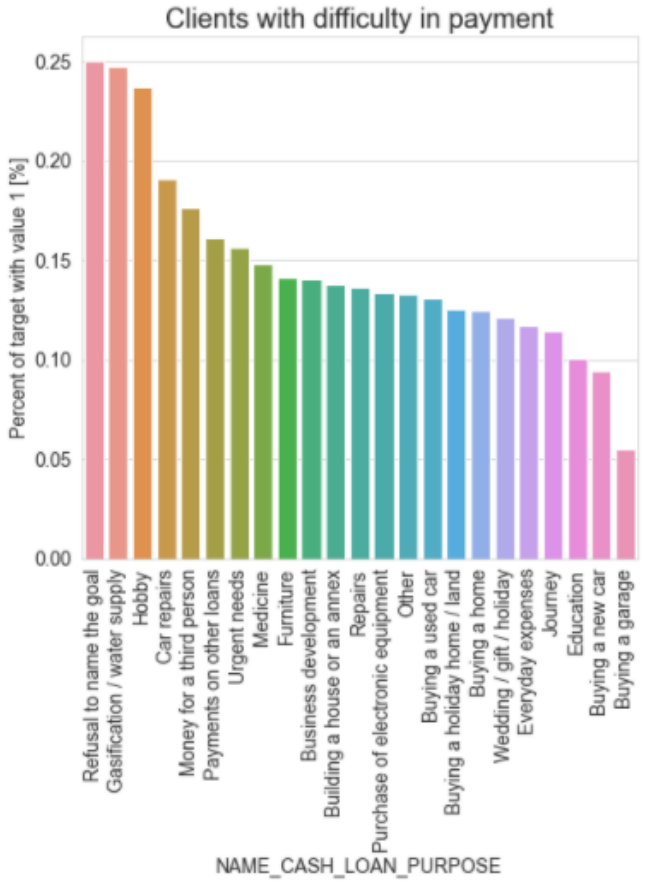


Observations for NAME_CASH_LOAN_PURPOSE, CASH LOAN PURPOSE

- Most contracts taken by clients state their goal is repairs

- Clients who refuse to state their goals have high percentage of difficulty in payment of loans





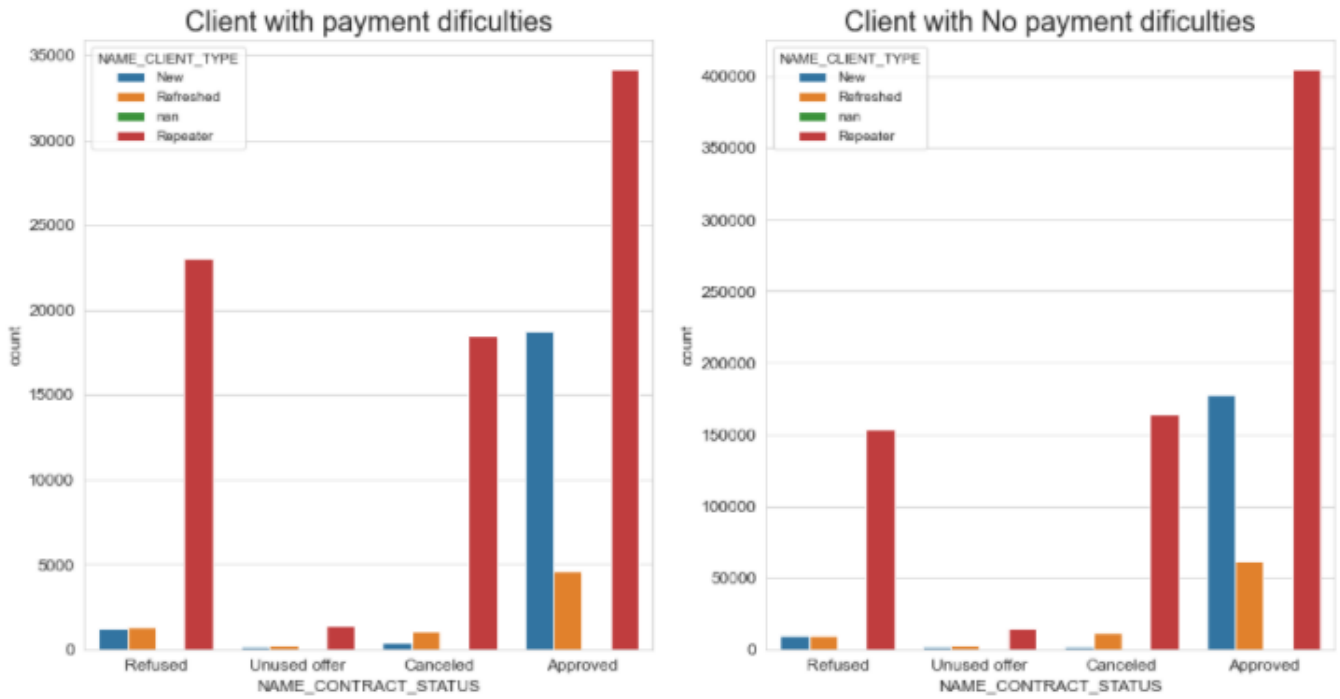
Observations of previous application data

BIVARIATE ANALYSIS

Observations of NAME_CONTRACT_STATUS, NAME_CLIENT_TYPE

- Clients who are repeaters and have been refused before have more chances of difficulty in repayment of loans

- Clients who are new and the previous application getting cancelled have difficulty in repayment of loans

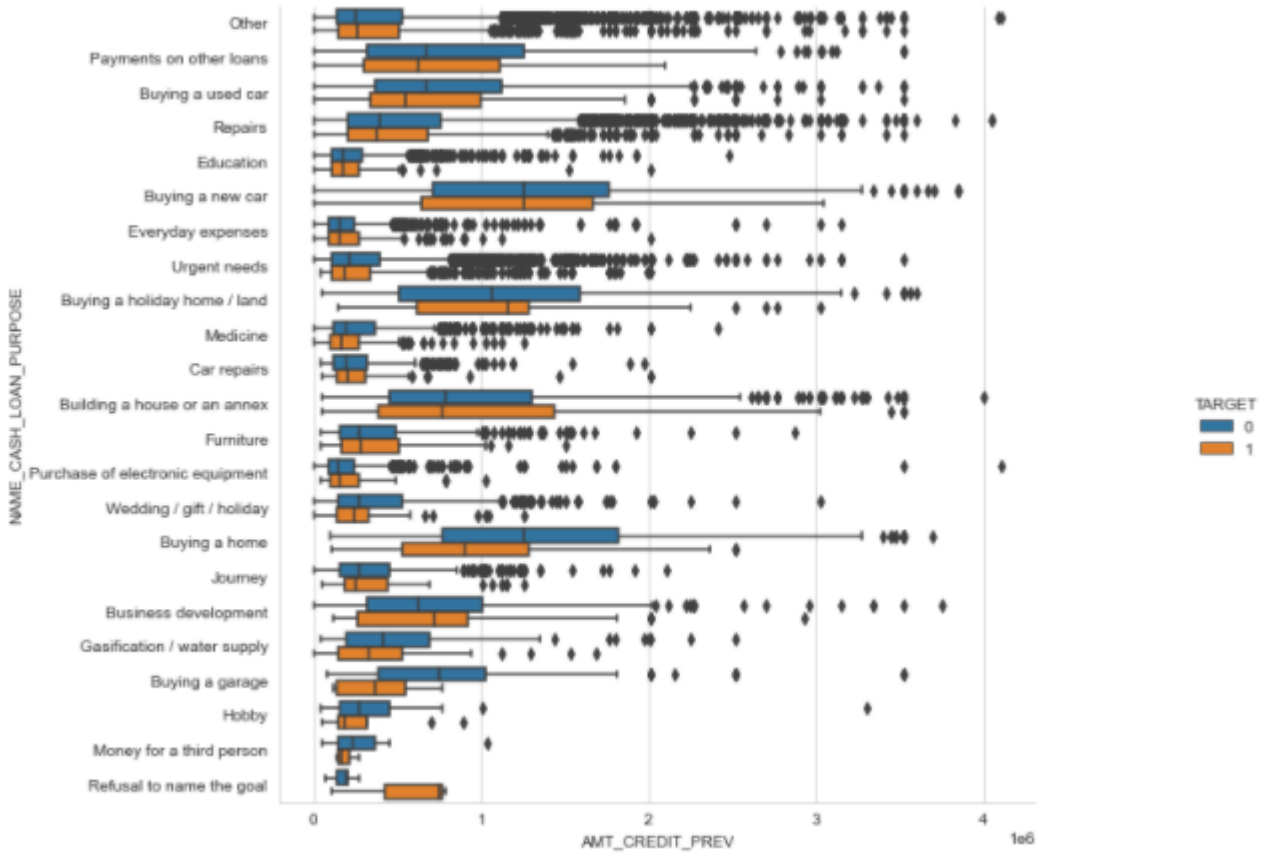


Observations of NAME_CONTRACT_STATUS, NAME_CONTRACT_TYPE

- Clients with cash loans which were refused before have more difficulty in loan repayment

Observations of AMT_CREDIT_PREV, NAME_CASH_LOAN_PURPOSE

- Clients who refuse to name their goal with high amount credit tend to have payment difficulties



Conclusions

Defaulter's demography

Application data

- The clients with maternity leave as their income type seem to have high % of payment difficulties, so they are the driving factors to be avoided
- The clients with low skilled laborers as their occupation type have high% of payment difficulties, so they are the driving factors to be avoided
- The clients with lower secondary as their education type have higher% of payment difficulties, so they are also one of the driving factors to be avoided

Previous Application Data

- The clients with Refusal to name the goal as their loan purpose have higher % of payment difficulties, so they are also one of the driving factors to be avoided
- The clients with Refused loans as their previous contract status also have higher % of payment difficulties, so they are also one of the driving factors to be avoided
- The clients with Revolving Loans as their previous contract type also have higher % of payment difficulties, so they are also one of the driving factors to be avoided

Credible applications

- According to data most of the defaulters are working clients, however this is just a correlation but not a causation. Hence proper scrutiny must be done on other parameters before the refusing the application of loan.
- In the given data female applicants have more difficulties in count when compared to male applicants but when compared in percentages, they do not make much difference. Hence there should not be any bias based on gender.