Credit EDA case study

By – Arjun Mehtani

Problem Statement

- 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.

Steps Required – application.csv dataframe

- Importing required libraries
- Data Sourcing/loading
- Understanding the data by reading data dictionary. – (columns_description.csv)
- Data Cleaning – (handling null values , finding outliers , checking incorrect data types , standardising values)
- Univariate analysis on the entire data frame
- Dividing the data into 2 categories – (TARGET0 and TARGET1)
- Univariate analysis on each category and analysing the trends.
- Bivariate/multivariate analysis

Steps Required – previous_application.csv dataframe

- Importing required libraries
- Data Sourcing/loading
- Understanding the data by reading data dictionary. – (columns_description.csv)
- Data Cleaning – (handling null values , finding outliers , checking incorrect data types , standardising values)
- Univariate analysis on the entire data frame
- Merging the TARGET column from application.csv dataframe using SK_ID_CURR column.
- Dividing the data into 2 categories – (TARGET0 and TARGET1)
- Univariate analysis on each category and analysing the trends.
- Bivariate/multivariate analysis

Data Cleaning

In the data cleaning phase the following activities were performed-

- Checking percentage of nulls in all columns and removing all the columns with more than 40% null values.
- Handling all the columns with nulls less than 40% -(appropriate measure like replacing null values with mean() , median() or mode was taken to accomplish this task.)
- Handling all the incorrect values/datatypes

  1) Finding and handling negative values – (values were replaced with absolute values)

  2) Finding and handling values like XAP , XNA.- (values were replaced with NaN)

(note – For our analysis we have assumed that XNA and XAP represents null values)

- We also removed all the unnecessary columns which were not useful for our analysis after analysing the data dictionary.

Univariate analysis on Numerical variables for application_data.csv

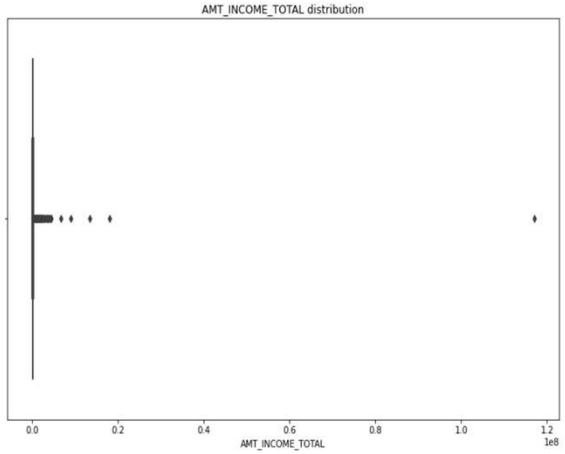


Fig1

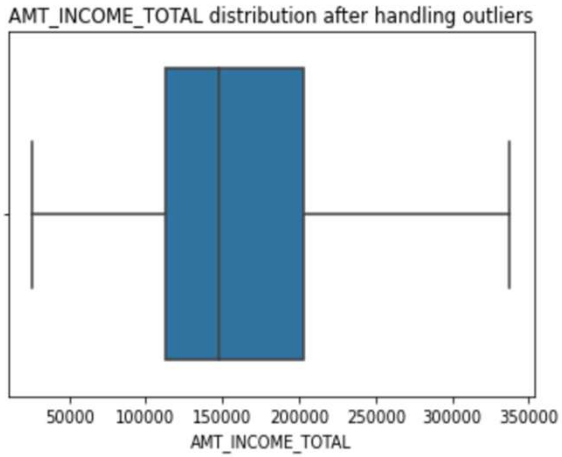


Fig2

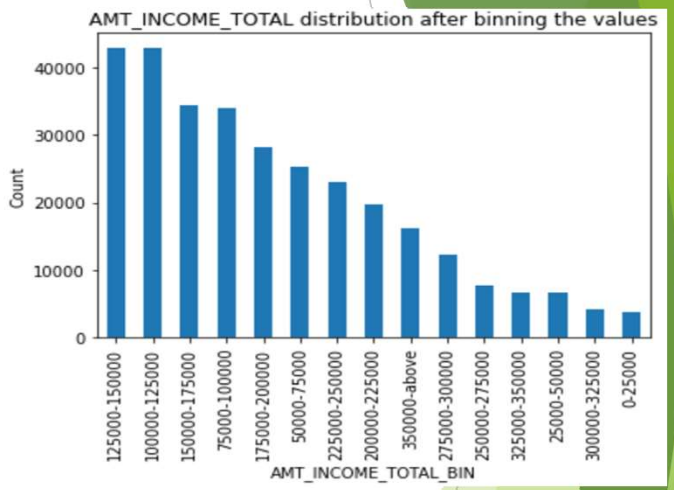


Fig3

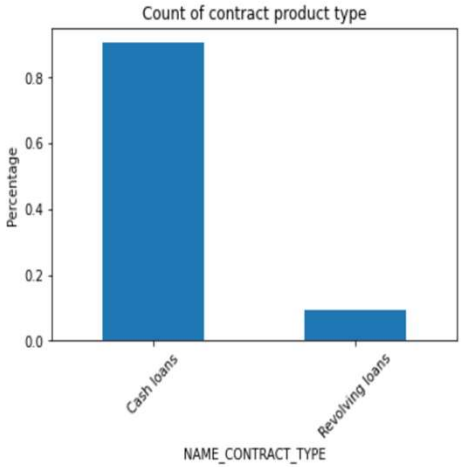
Above 3 figures show the Numerical analysis on AMT_INCOME_TOTAL column.

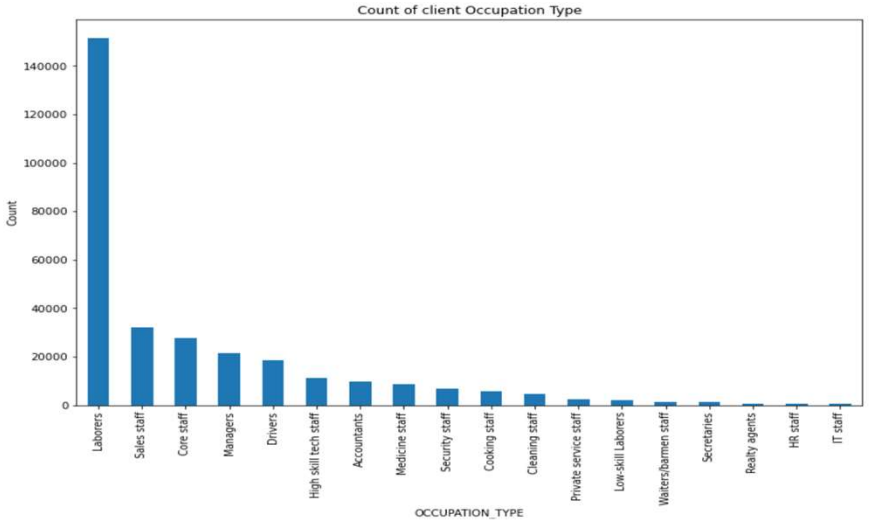
We can clearly see that most client have income in the range 100000 to 150000.

Note -

- (1st figure shows the presence of outliers in the data , 2nd figure is plotted after removing the outliers and the 3rd figure was plotted after binning the values to get a deeper insight of the data)
- We have applied the same procedure for all other numerical columns as well.

Univariate analysis on Categorical variables for application_data.csv





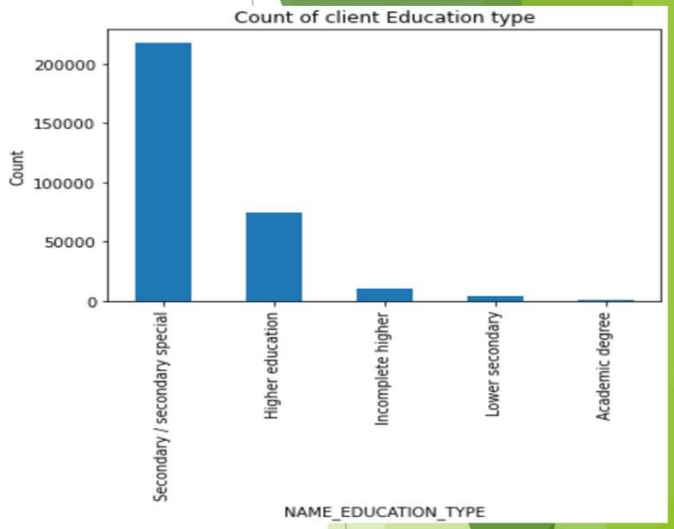


Fig1 Fig2 Fig3

Above 3 figures show the univariate analysis(categorical) on NAME_CONTRACT_TYPE , OCCUPATION_TYPE and NAME_EDUCATION_TYPE column.

We can clearly see that most client opted of cash loans instead of revolving loans.

Most of the clients who opted for loans were Laborers.

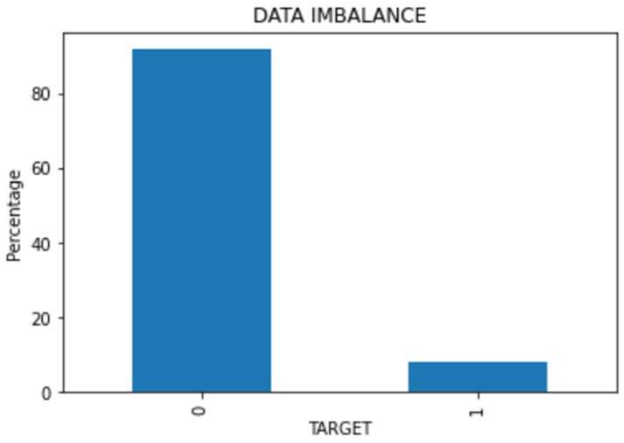
Most of the clients had Secondary/Secondary special Education type.

Note -

We have applied the same procedure for all other categorical columns as well.

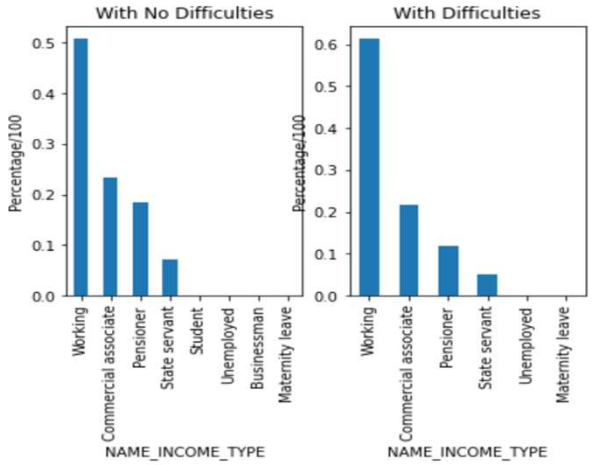
Univariate analysis on the basis of Target Variable for application_data.csv

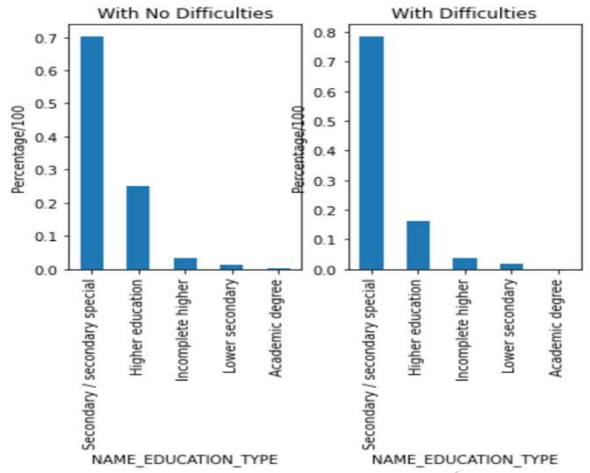
- Data imbalance



We can see that there is a data imbalance-

Percentage of Target 0 - 91.927118
Percentage of Target 1 - 8.072882





Observation -

(NAME_INCOME_TYPE)

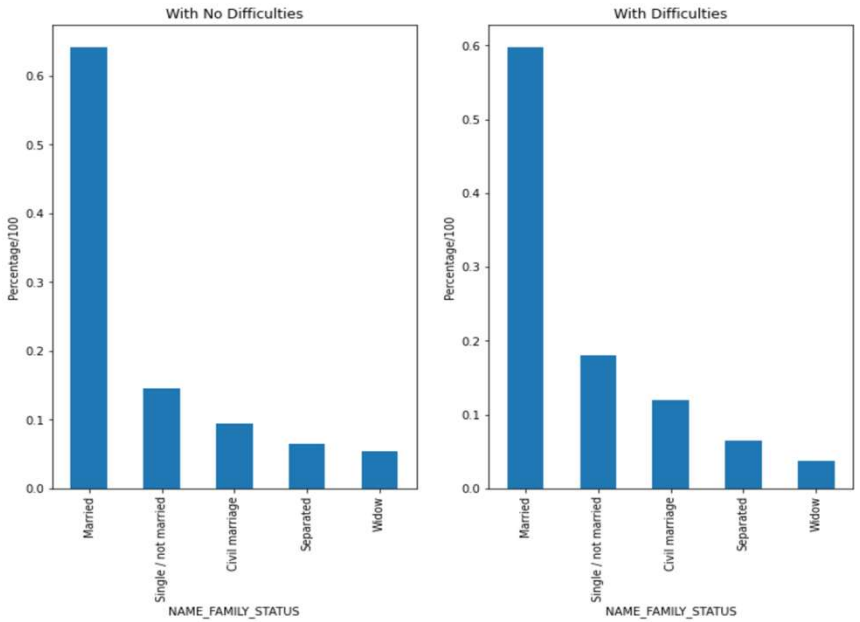
1) We see that the percentage of Working people increases with difficulties in payment
2) We also see that percentage of pensioner decreases with difficulties in payment

(NAME_EDUCATION_TYPE)

1) We see that percentage of Secondary/secondary special increases with difficulties in payment.
2) We also see that percentage of higher education decreases with difficulties in payment.

Univariate analysis on the basis of Target Variable for application_data.csv

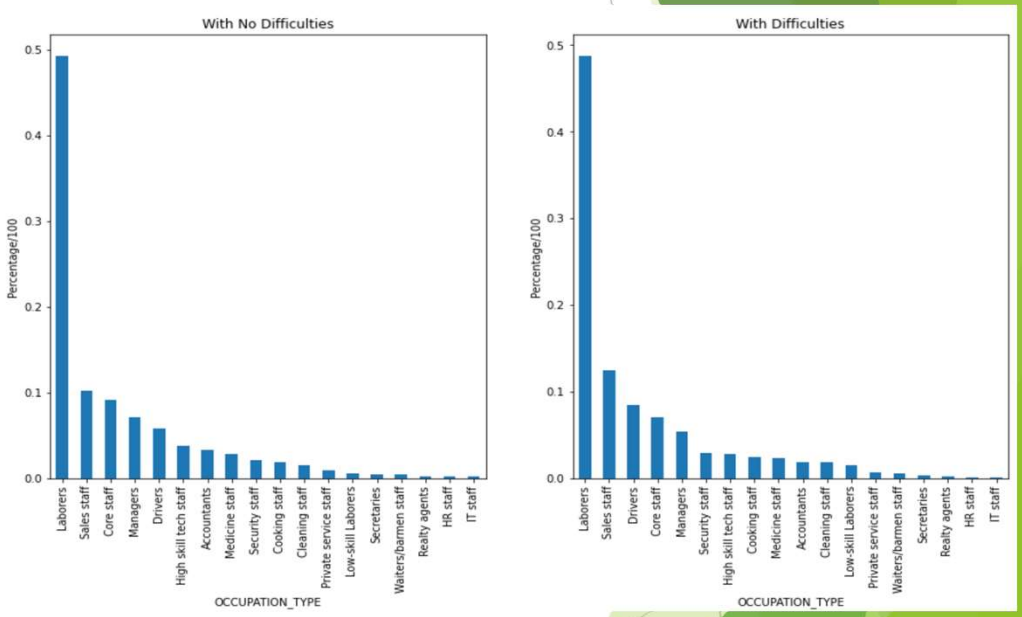
NAME_FAMILY_STATUS



Observation -

1) We see that percentage of married decreases with difficulties in payment
2) We also see that percentage of Single/not married and Civil marriage increase with difficulties in payment

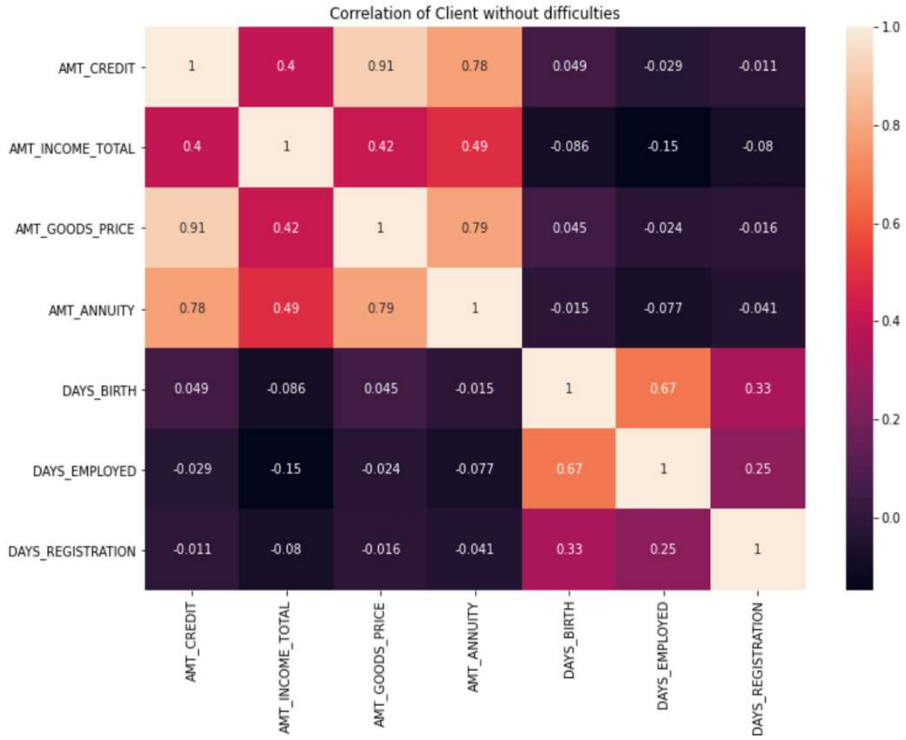
OCCUPATION_TYPE

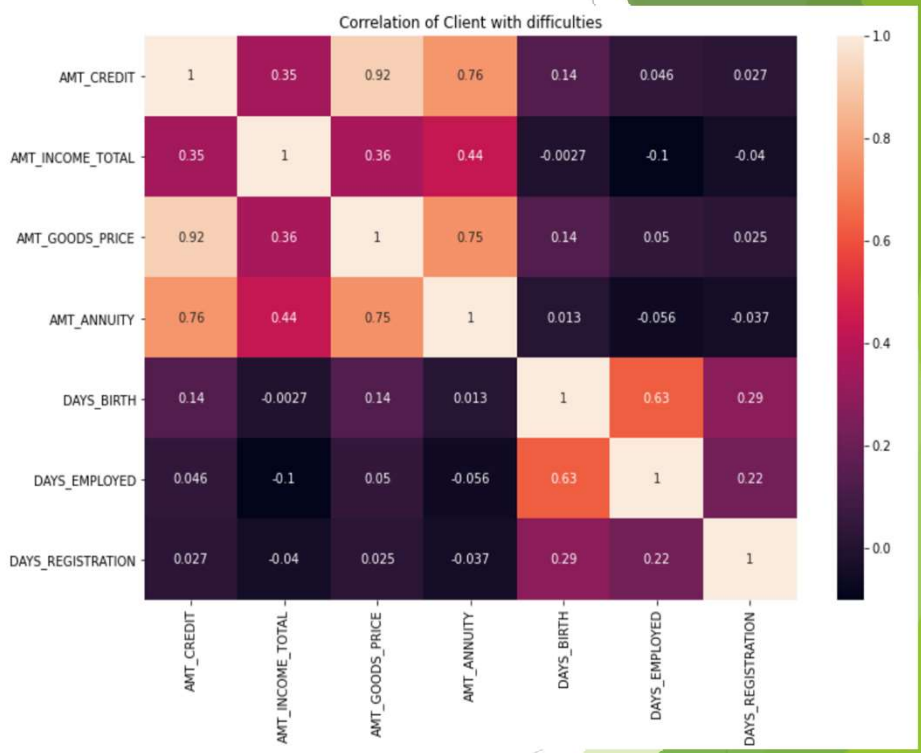


Observation -

1) We see that percentage of sales staff , driver increase slightly with difficulties in payment.
2) We see that percentage of core staff , manager decrease slightly with difficulties in payment.

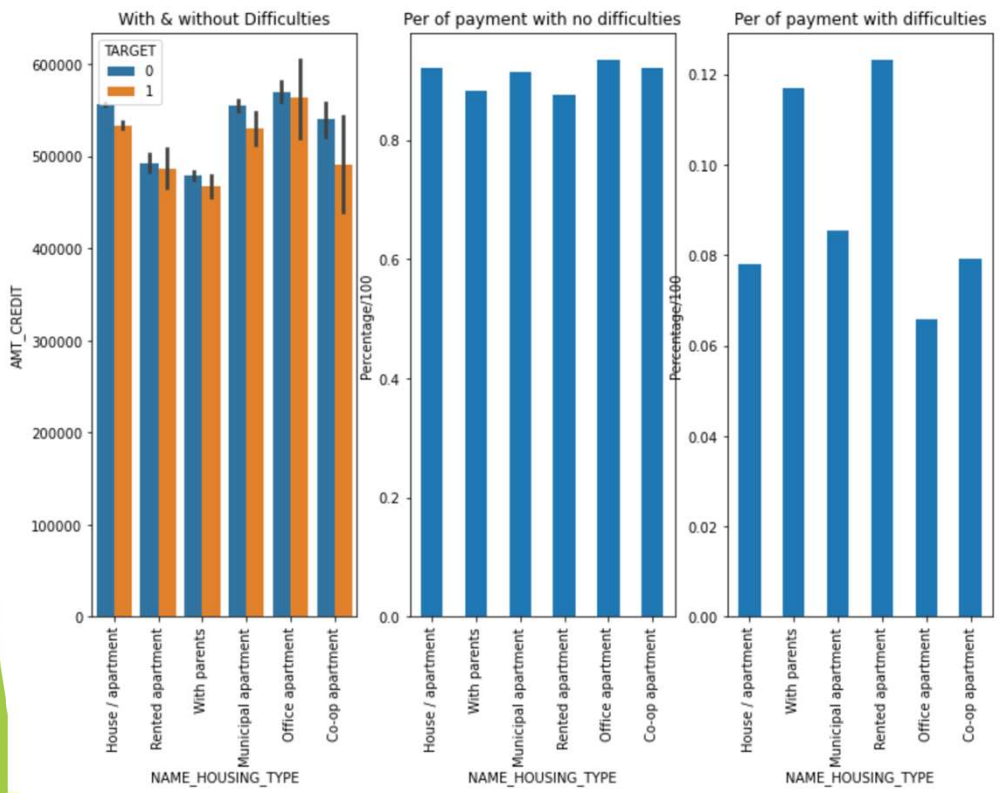
Bivariate analysis/Multivariate analysis for application_data.csv





1) AMT_GOODS_PRICE and AMT_CREDIT have a high correlation
2) AMT_GOODS_PRICE and AMT_ANNUITY have high correlation
3) AMT_ANNUITY and AMT_CREDIT have high correlation

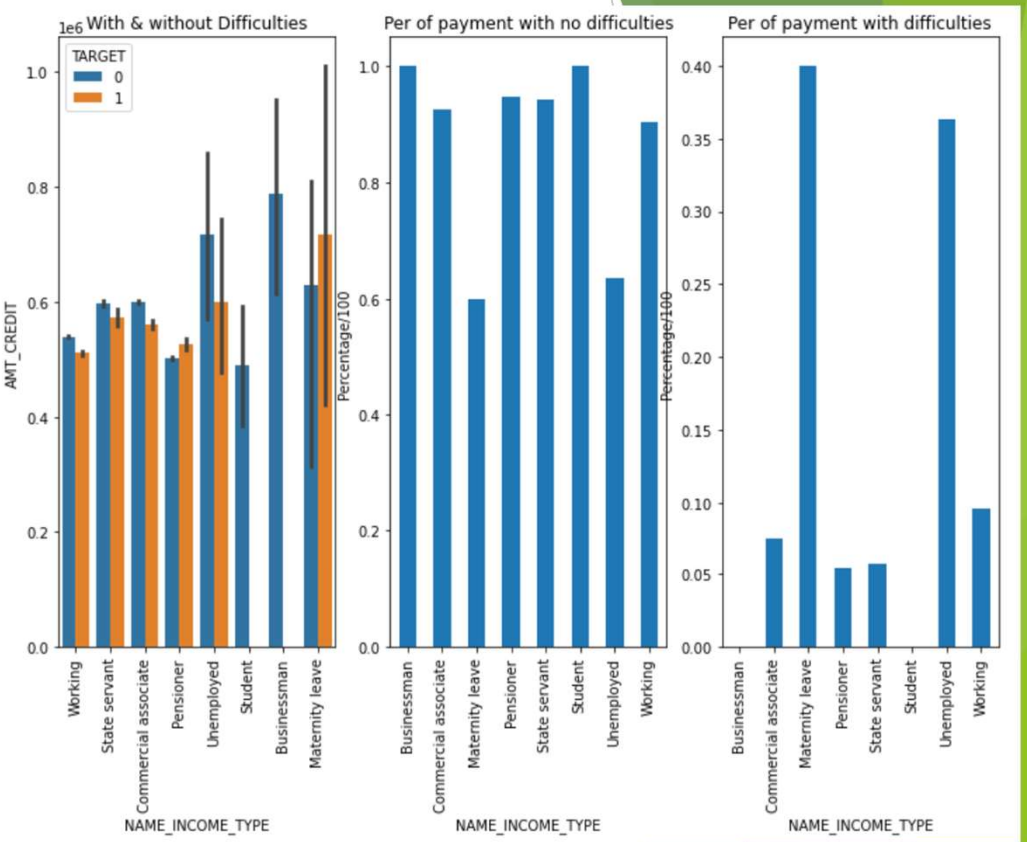
NAME_HOUSING_TYPE and AMT_CREDIT



Observation –

1) We can see that people who live in office apartment have highest credit amount of loan with/without difficulties.
2) We can also see that people who live with parents have the lowest credit amount of loan with/without difficulties.
3) We can also see that people who live in rented apartment or with parents have high percentage of payment with difficulties

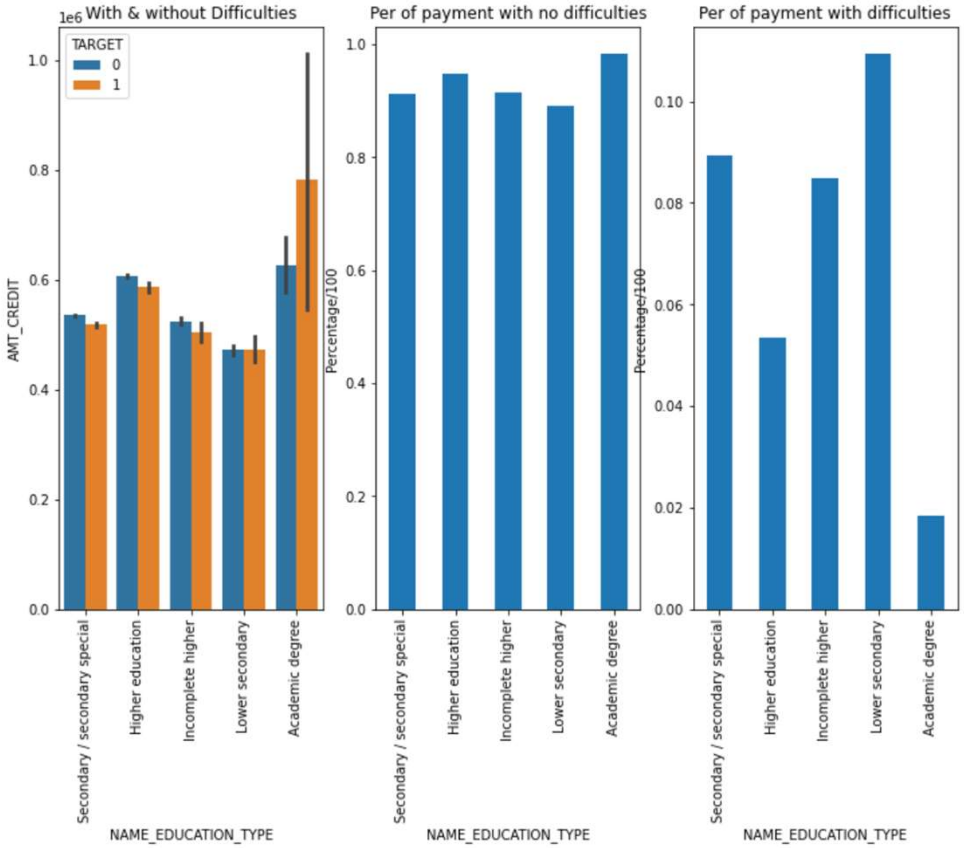
NAME_INCOME_TYPE and AMT_CREDIT



Observation -

1) Businessman have highest credit amount for payment without difficulties.
2) Maternity leave and unemployed have highest percentage of payment with difficulties.

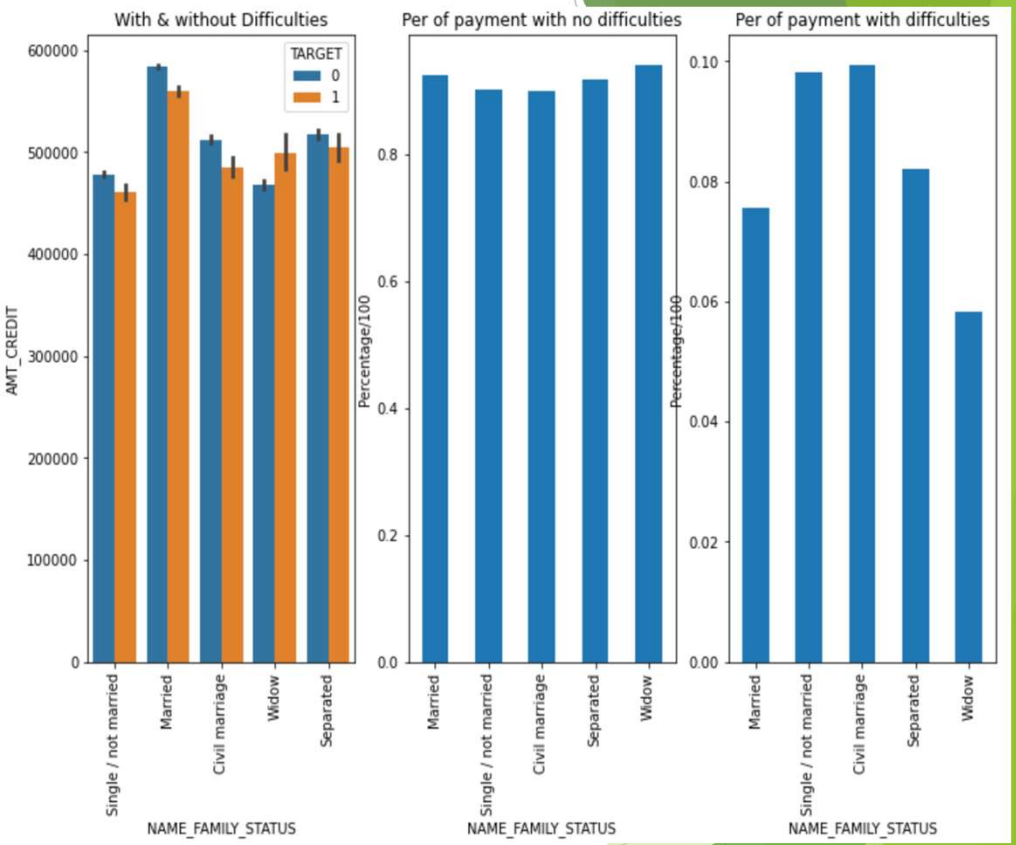
NAME_EDUCATION_TYPE and AMT_CREDIT



Observation -

1) We can see that client with academic degree has the highest credit amount and least percentage in terms of payment with difficulties.
2) We can also see that client with education of lower secondary has the highest percentage of payment with difficulties.

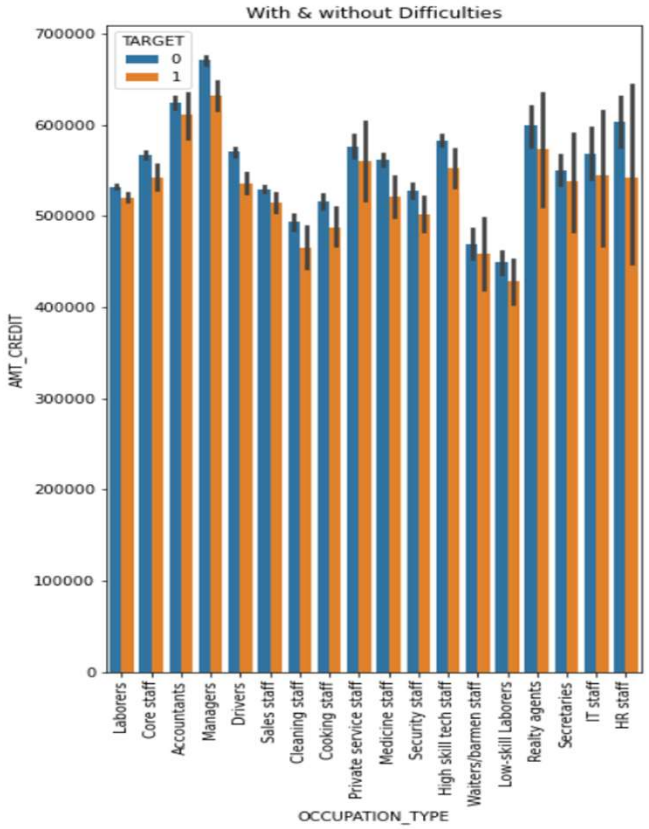
NAME_FAMILY_STATUS and AMT_CREDIT

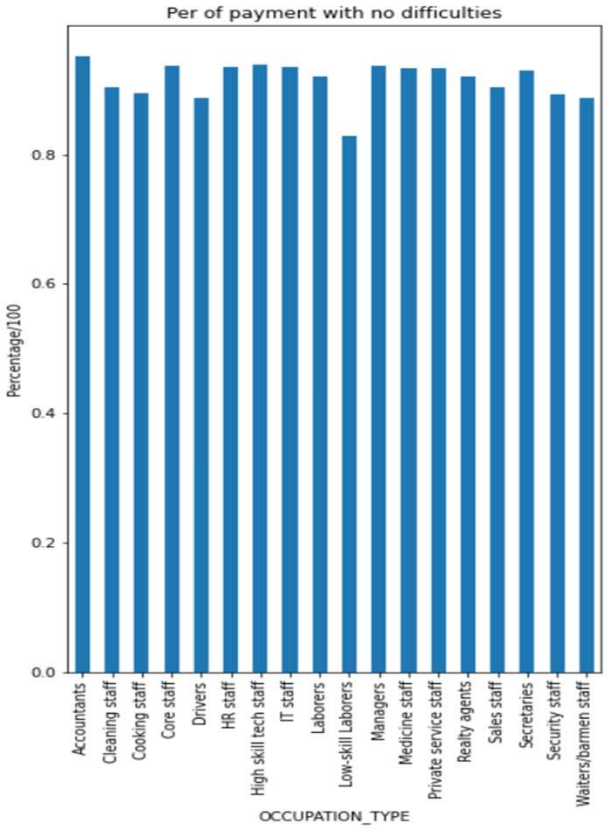


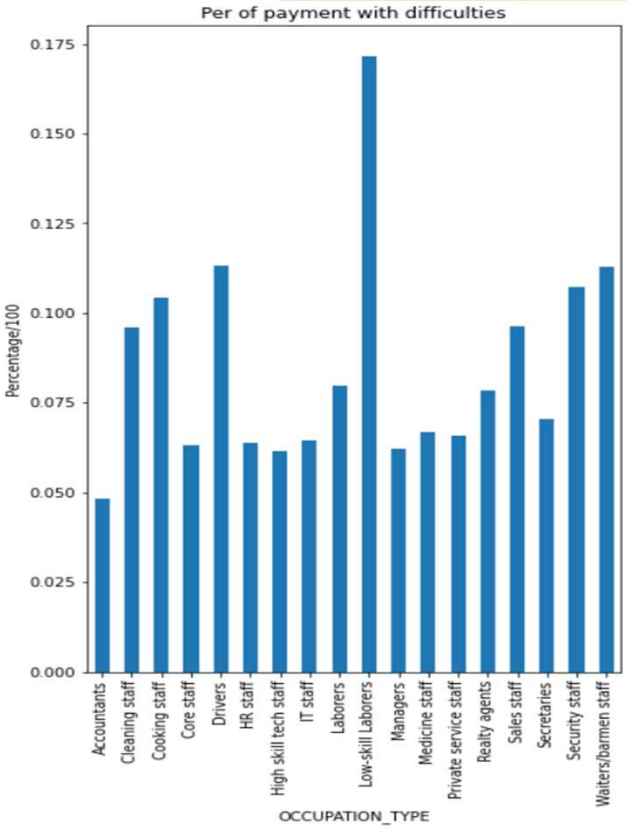
Observation –

1) We can see that married people have the highest no. of loan credit.
2) We can also see that single , civil marriage have more percentage of payment with difficulties.

OCCUPATION_TYPE vs AMT_CREDIT



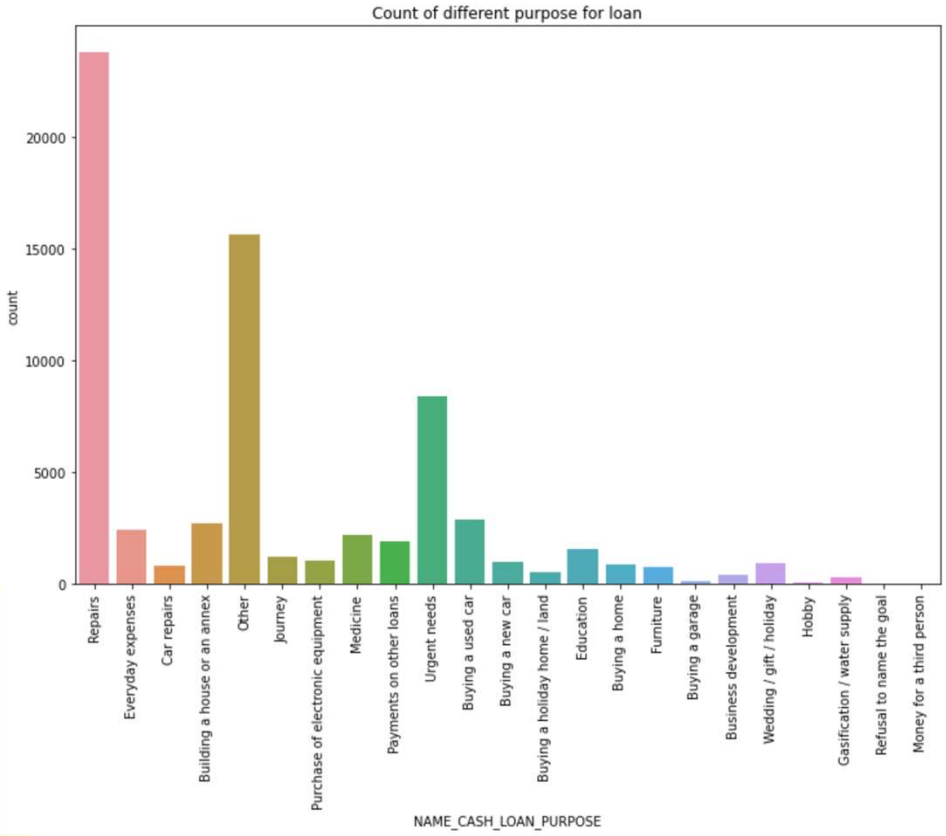




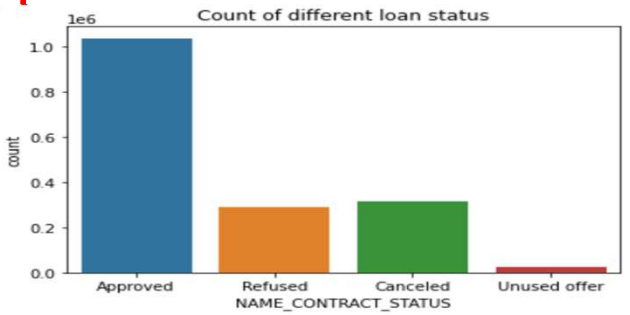
Observation –

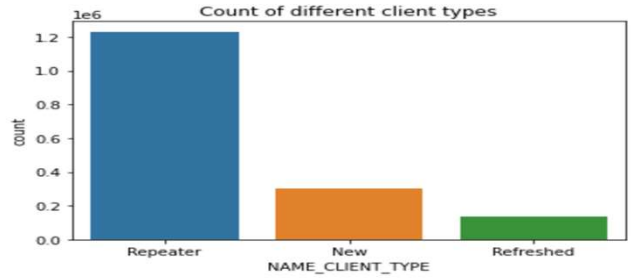
We can clearly see that low skilled laborers have high percentage of difficulties in loan payment

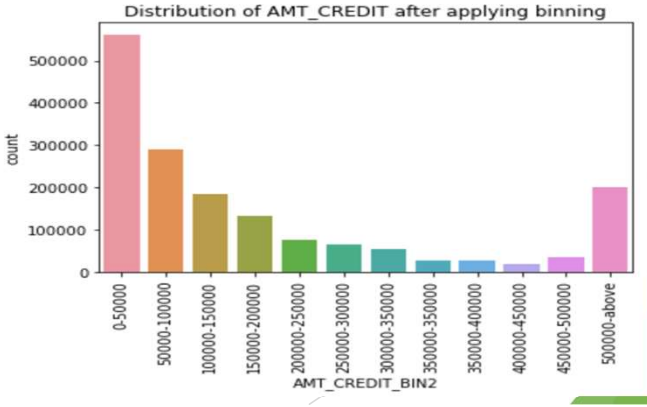
Univariate analysis for previous_application.csv



We can clearly see that Repairs are the major reasons for clients to take cash loans.







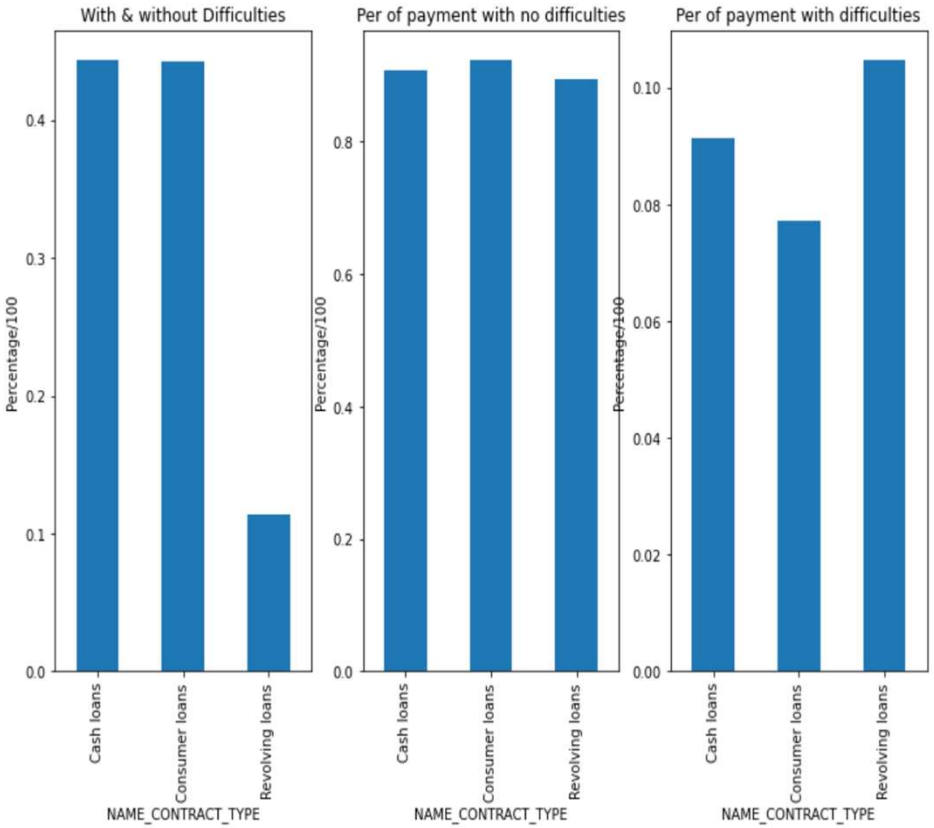
Observation-

We can clearly see that majority of the client loans were Approved while Unused offer was in the minority.

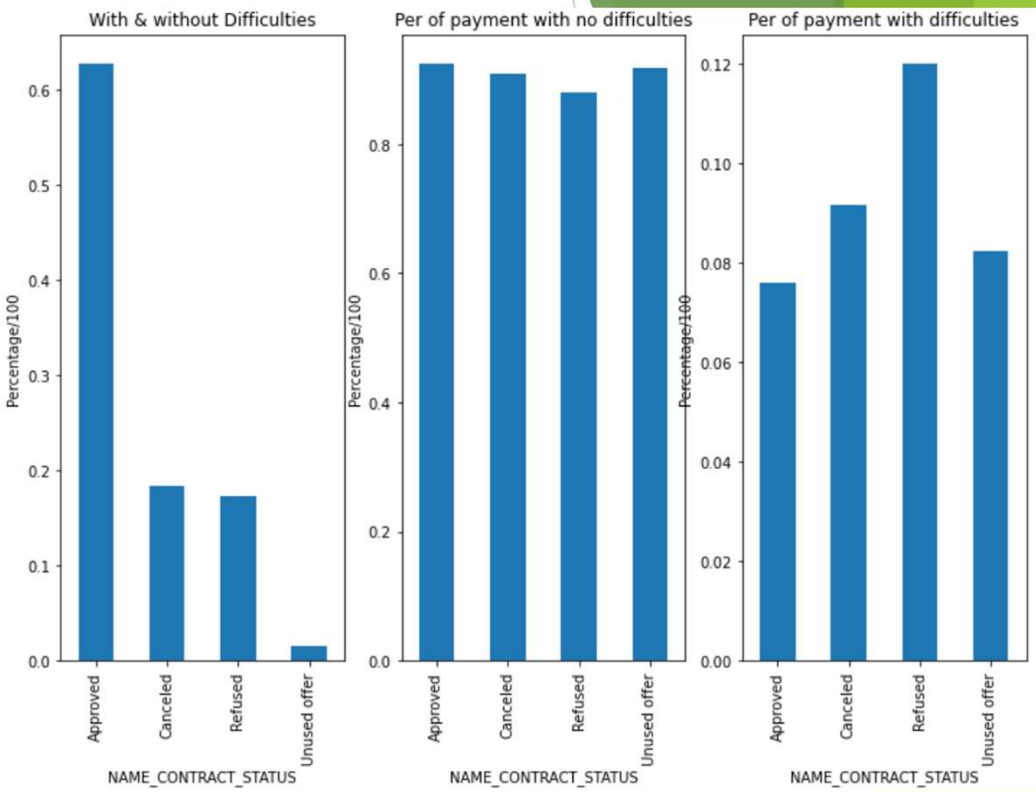
We can clearly see that majority client are repeater.

We can see that majority of credit amount which client received was in range - (0-50000)

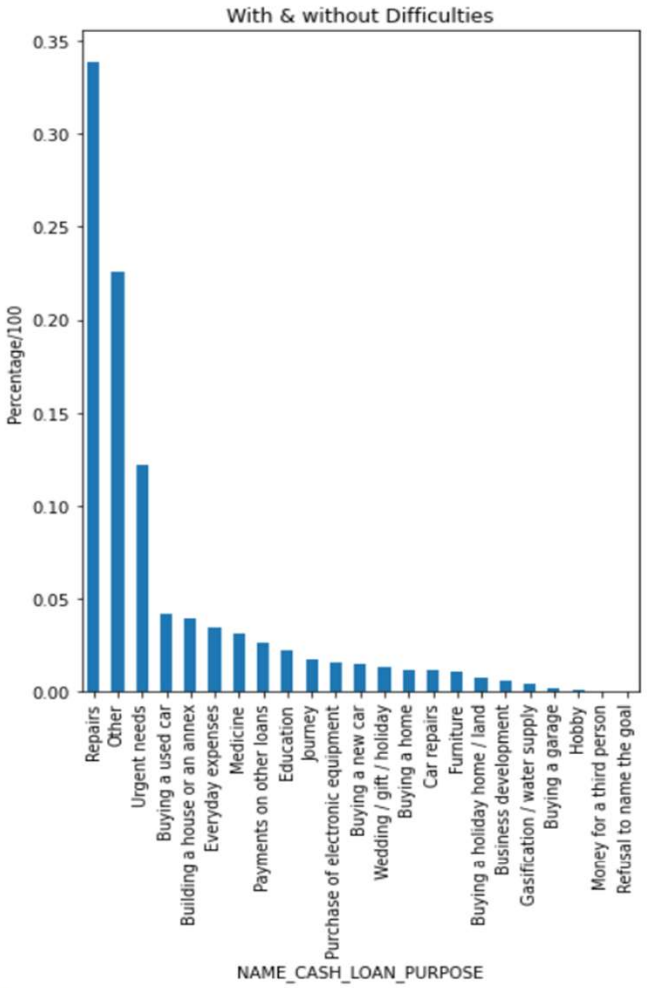
Univariate analysis on the basis of TARGET variable for previous_application.csv

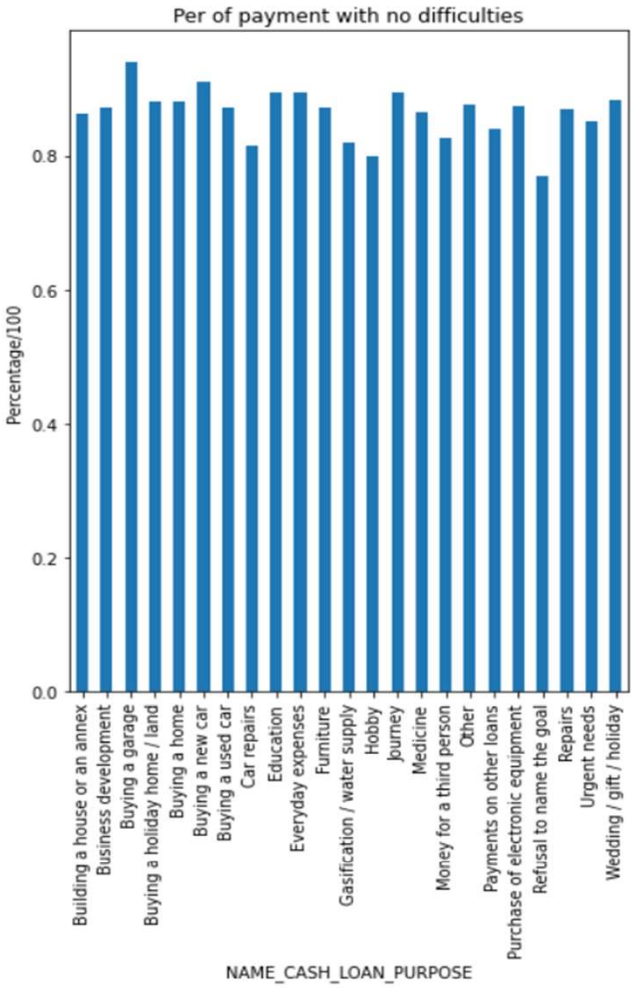


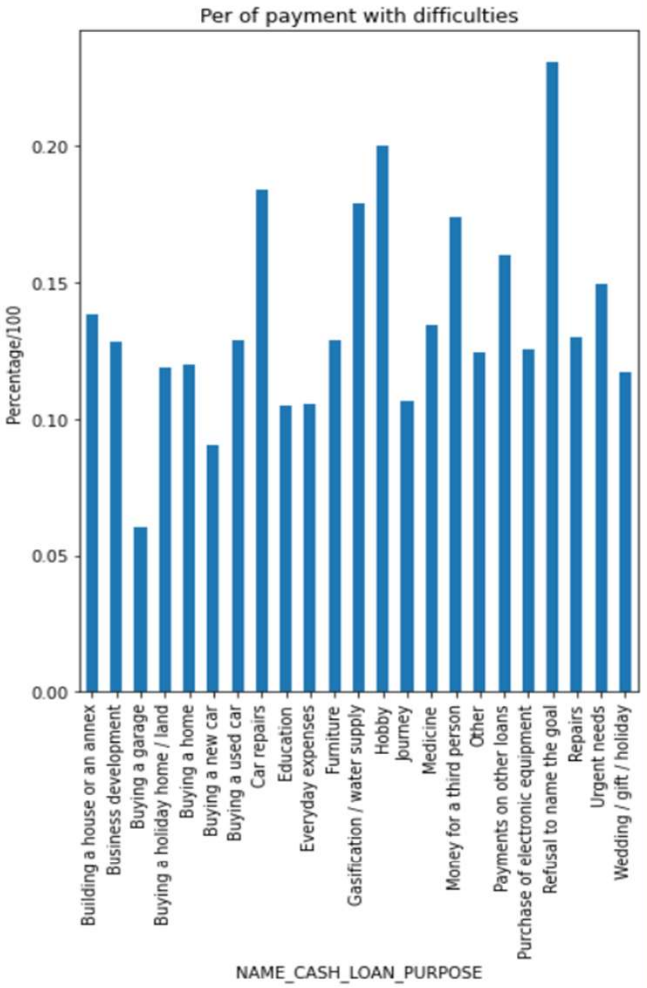
We can clearly see that Revolving loans are majority in case of loan payment with difficulties



We can also see that most of the loans with NAME_CONTRACT_STATUS as Refused had highest percentage of payment difficulties.

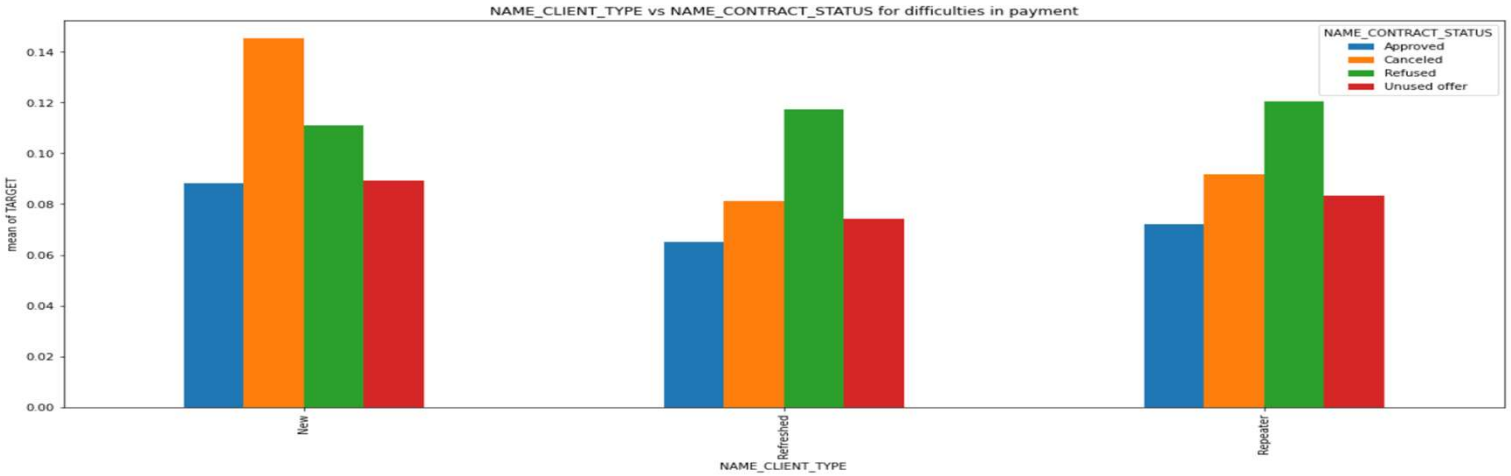


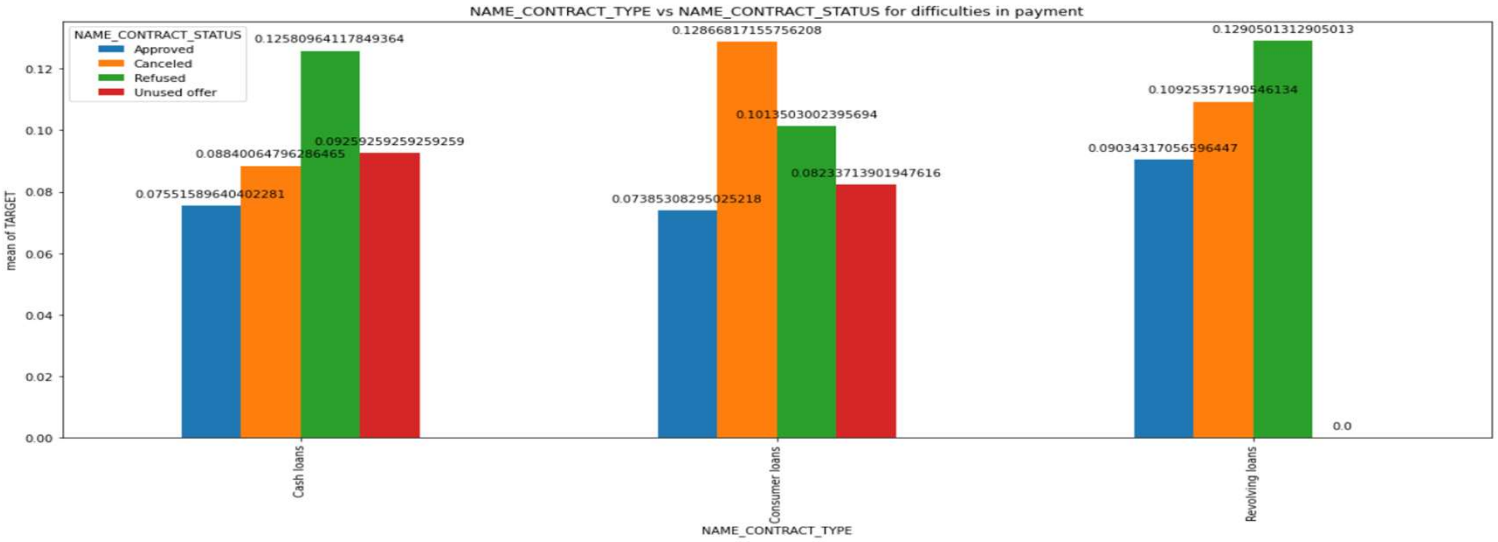




We can clearly see that majority of the purpose of loan is for 'Repairs' while 'Refusal to name the goal' is majority in case of payment with difficulties.

Bivariate analysis/Multivariate analysis for previous_application.csv





Observation -

We can observe that New client whose previous application was cancelled have higher percentage of difficulties in payment.

We can clearly see that clients who opted for revolving loans and refused the previous application have more percentage of difficulties in payment

Conclusion

Application_data -

1) Client with NAME_INCOME_TYPE as Maternity leave and unemployed have highest percentage of payment with difficulties hence more likely to be loan defaulter.
2) Clients with NAME_HOUSING_TYPE as ‘rented apartment’ or ‘with parents’ have high percentage of payment with difficulties hence more likely to be loan defaulter.
3) Clients with NAME_EDUCATION_TYPE as ‘Lower Secondary’ have high percentage of payment with difficulties hence more likely to be loan defaulter.
4) Clients with INCOME more than 350000 have lowest percentage of payment with difficulties hence less likely to be loan defaulter.
5) Clients with OCCUPATION_TYPE as low skilled laborers have high percentage of difficulties in loan payment hence more likely to be loan defaulter.

Previous_application-

1) Clients with NAME_CONTRACT_STATUS as Refused in the previous application had highest percentage of payment difficulties hence more likely to be loan defaulter.
2) Clients who opted for Revolving loans in previous application had highest percentage of payment difficulties hence more likely to be loan defaulter.
3) Clients whose purpose of loan was 'Refusal to name the goal’ in the previous application had highest percentage of payment difficulties hence more likely to be loan defaulter.
4) Clients whose previous application was rejected due to 'SCOFR’ had highest percentage of payment difficulties hence more likely to be loan defaulter.

Thank You