Project Name: **Credit Risk Analysis**

Project Objective

identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of the loan, lending (to risky applicants) at a higher interest rate, etc.

Approach/ Steps to be followed

1. Upload
2. remove heavy null columns (>50%)
3. Go through columns_description and develop data understanding.
4. Start individual column analysis and Cleaning if necessary
5. Merge the dataframes
6. Identify Object and Numeric data of dataframe (Created 2 different dataframes to clearly identify)
7. Identify categorical columns (having <=10 unique values) and their cleaning in the merged dataframe.
8. Segmenting the Merged dataframe based on Target group (0 or 1)
9. Running Univariate / Bivariate / multivariate analysis to figure out correlation
10. Conclusion : Summarise the outcome with Correlations identified.

Note : to facilitate the ease of usage, common functions are created before step 3. Calling it **step-Functions**

Step 1 : Usual upload

Step 2 : There are many columns with too much null values.. we have 41 dropped cols with null values from current Applications, while such 4 columns from Previous applications data were dropped. Threshold considered for deletion : 50%

Step 3 : Study of Columns Description :

This step was the most time consuming as data understanding was built from here.

Step 4 : Individual columns analysis (Univariate) and cleaning if necessary:

Approach:

Parameters to check :

- How many null values ?
- does column have any unexpected entry ? Definition of unexpected would be different for each column
- what inference could be drawn from column ?
- Finally - what could be done with this column ?

Outcome:

- Null handling: For Numerical columns, nulls are mostly replaced with 0 or mean value depending on column nature.
- Outliers handling: For Numerical columns, Outliers were handled in some cases by replacing with Upper limit based on IQR.
- Categorical columns were cleaned separately after merge.

Step 5 : Merge:

The two data frames were merged after both data frames were individually cleaned.

Step 6 : Identify Object and Numeric data of dataframe

While keeping the first 2 columns as common, numeric and object columns are separated.

Step 7 : Identify categorical columns

A total of 23 columns were filtered out first. Their nulls were searched and found that 21 out of them have less than 2% nulls, which were replaced by Mode treatment. Other 2 columns were heavily loaded on nulls (~49%) which were dropped.

Step 8 : Segmenting the Merged dataframe based on Target group (0 or 1)

2 separate data frames created from merged and cleaned dataframe. This would help us run bivariate analysis as now – we have a totally cleaned data.

In this step – we have taken care of cleaning required in categorical columns as well.

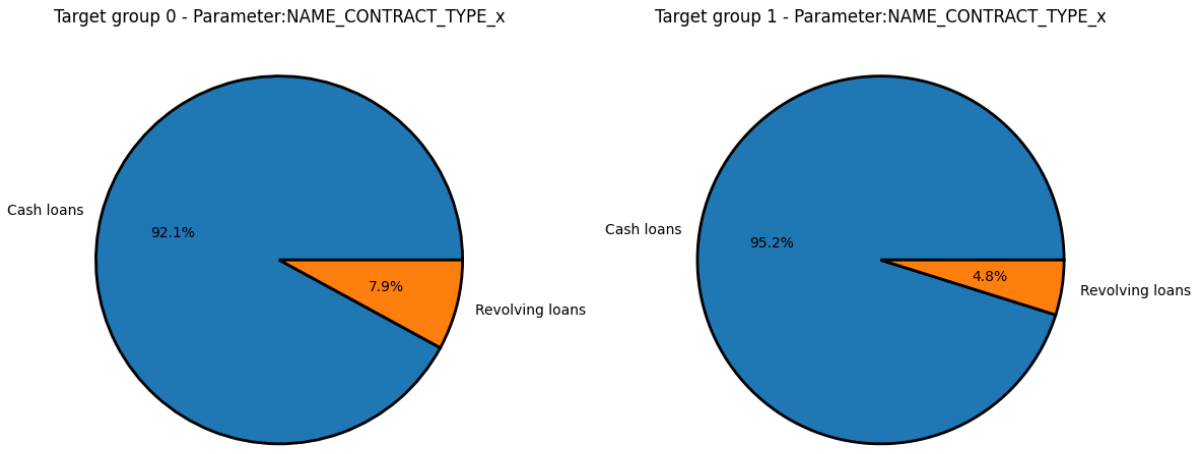
Step 9 : Running Uni/Bivariate / multivariate analysis to figure out correlation

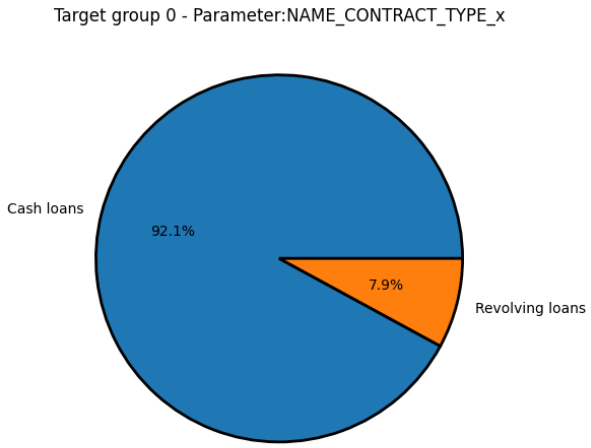
Following type of tools were used to identify the correlations :

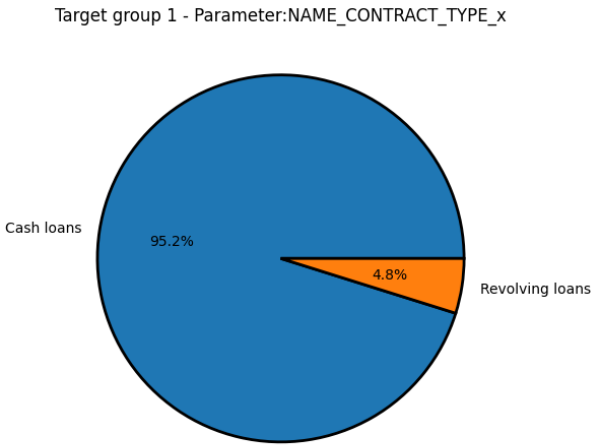
- Pie plots – Target group wise
- Heat map – On numerical data columns
- Scatter Plots – On numerical data columns
- Bar plots - Combination of Numerical and categorical data : Target group wise

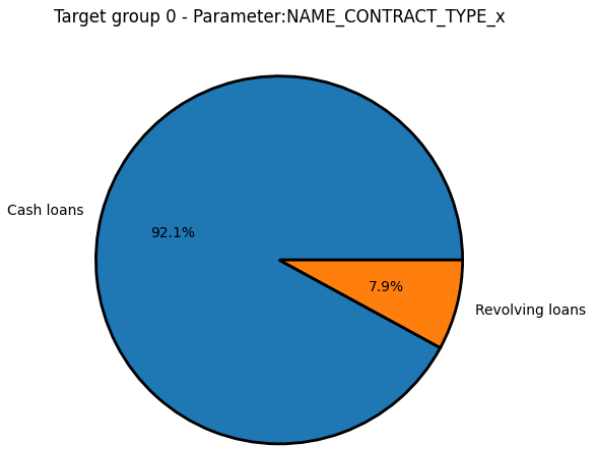
Here are some of the snippets observed during the project study:

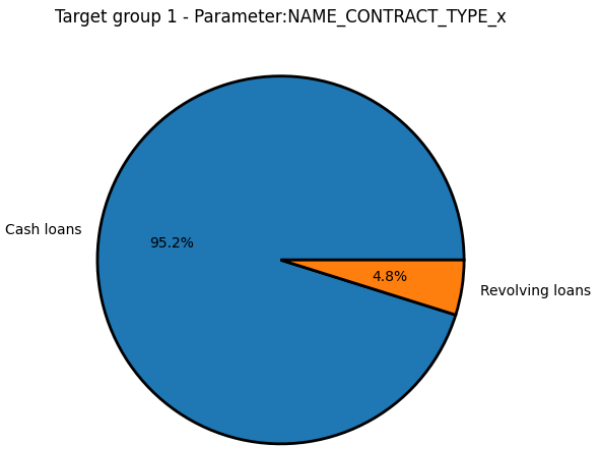
Pie Plots:

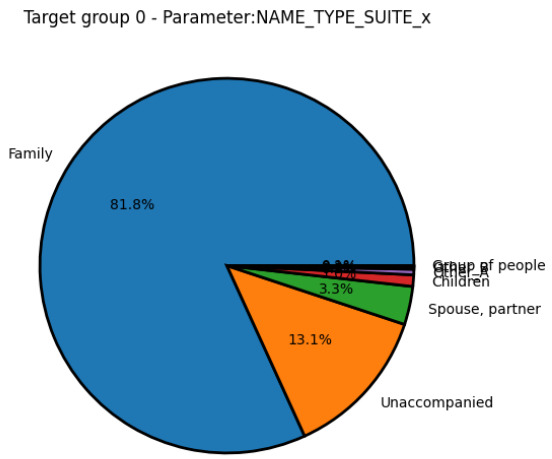


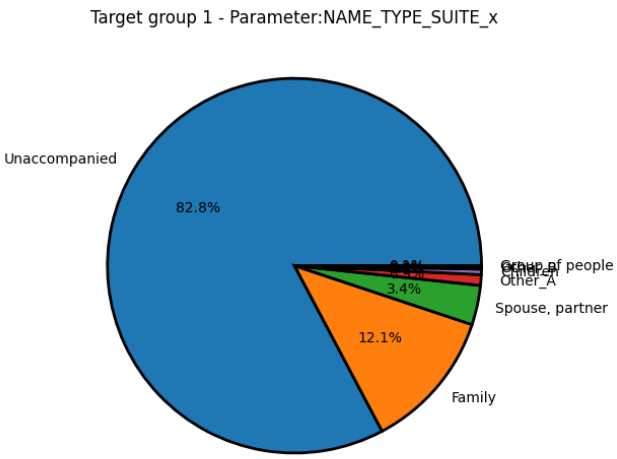

Target group 0 - Parameter:NAME_CONTRACT_TYPE_x
Cash loans
92.1%
7.9%
Revolving loans


Target group 1 - Parameter:NAME_CONTRACT_TYPE_x
Cash loans
95.2%
4.8%
Revolving loans

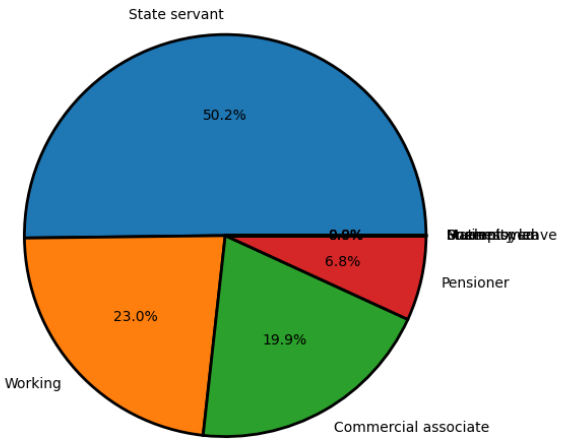

Target group 0 - Parameter:NAME_CONTRACT_TYPE_x
Cash loans
92.1%
7.9%
Revolving loans


Target group 1 - Parameter:NAME_CONTRACT_TYPE_x
Cash loans
95.2%
4.8%
Revolving loans

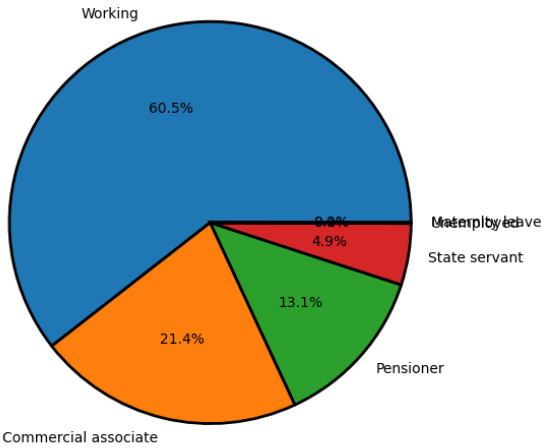

Target group 0 - Parameter:NAME_TYPE_SUITE_x
Family
81.8%
Group of people
Other_A
Children
3.3%
Spouse, partner
13.1%
Unaccompanied


Target group 1 - Parameter:NAME_TYPE_SUITE_x
Unaccompanied
82.8%
Group of people
Other_A
3.4%
Spouse, partner
12.1%
Family

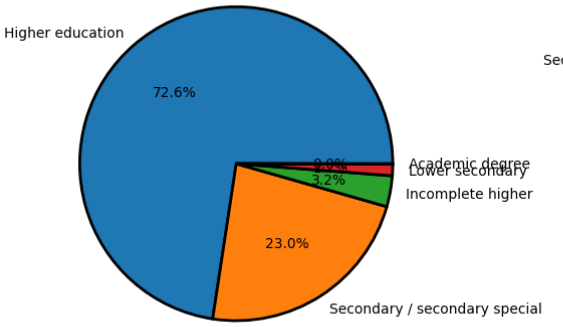
Target group 0 - Parameter:NAME_INCOME_TYPE



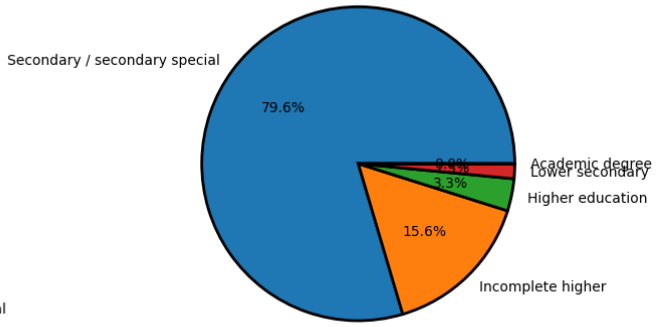
Target group 1 - Parameter:NAME_INCOME_TYPE



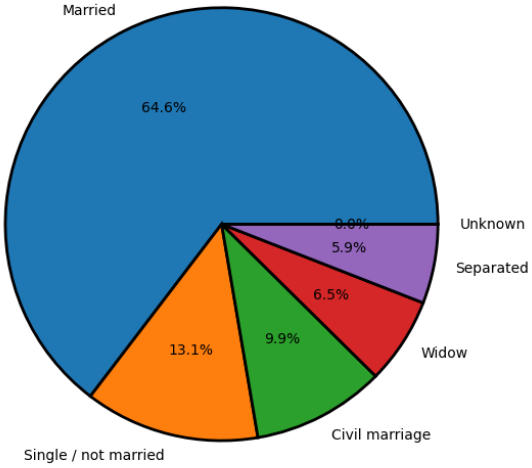
Target group 0 - Parameter:NAME_EDUCATION_TYPE



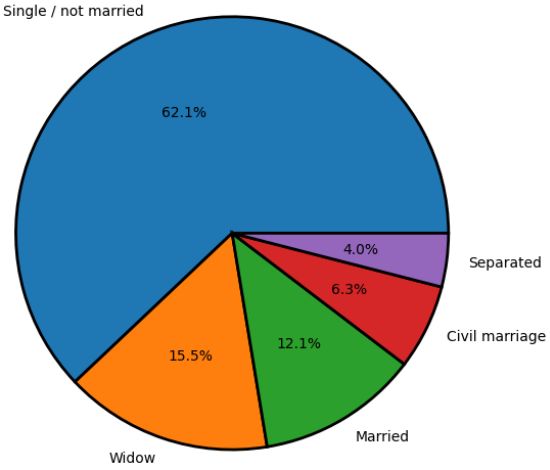
Target group 1 - Parameter:NAME_EDUCATION_TYPE

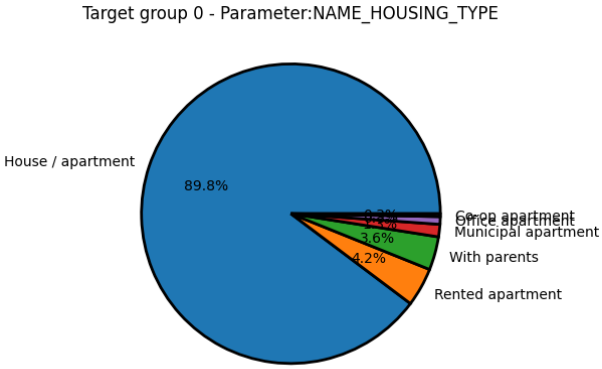


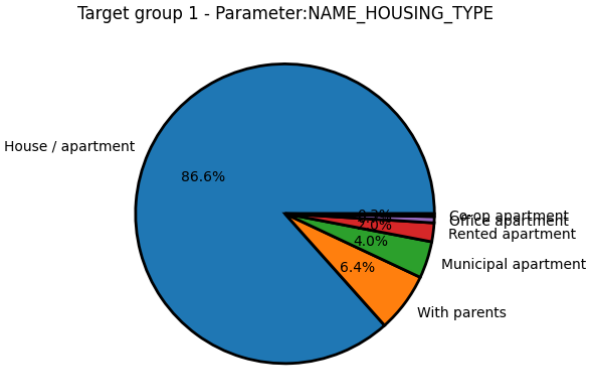
Target group 0 - Parameter:NAME_FAMILY_STATUS

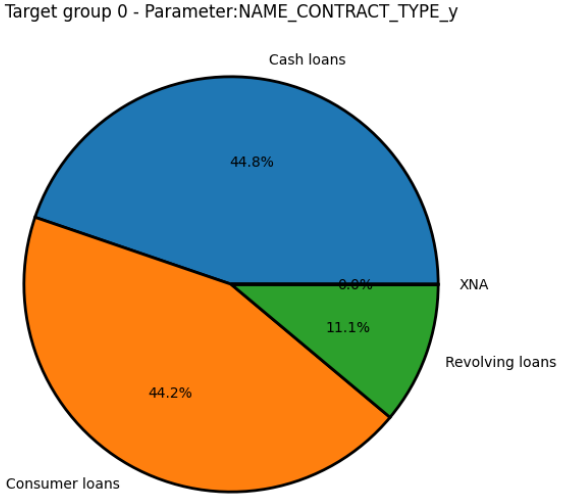


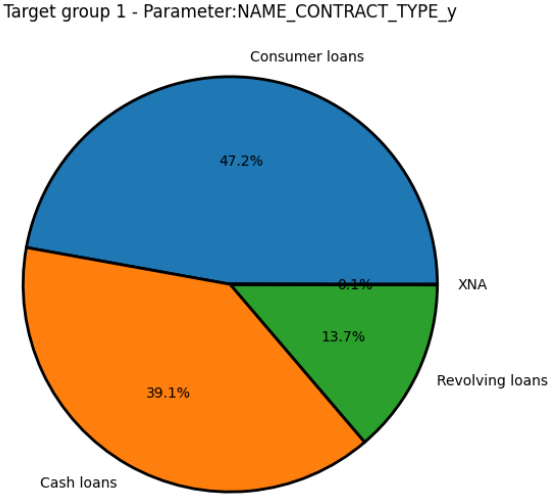
Target group 1 - Parameter:NAME_FAMILY_STATUS

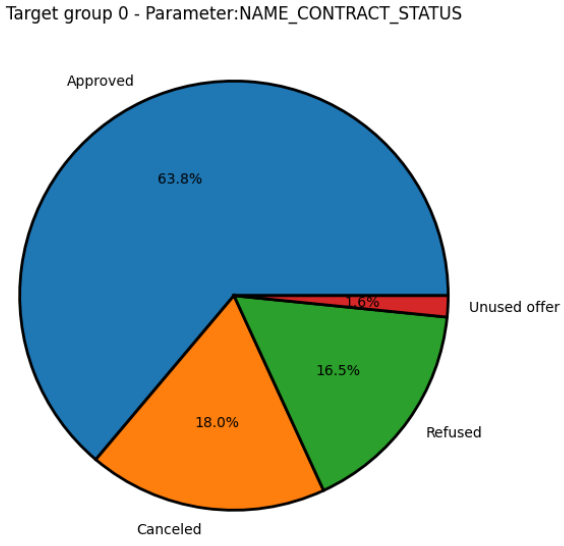


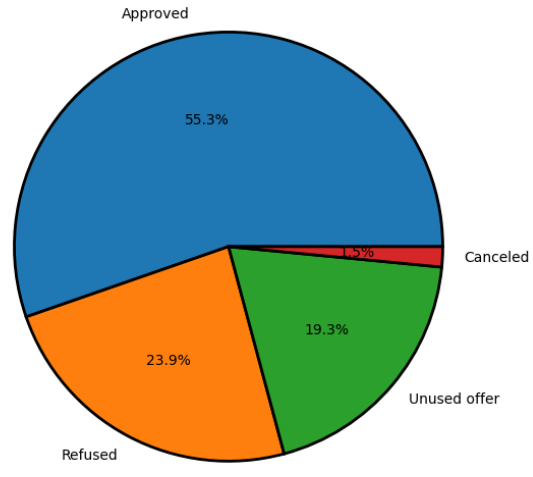

Target group 0 - Parameter:NAME_HOUSING_TYPE
House / apartment
89.8%
Co-op apartment
Office apartment
Municipal apartment
3.6%
With parents
4.2%
Rented apartment

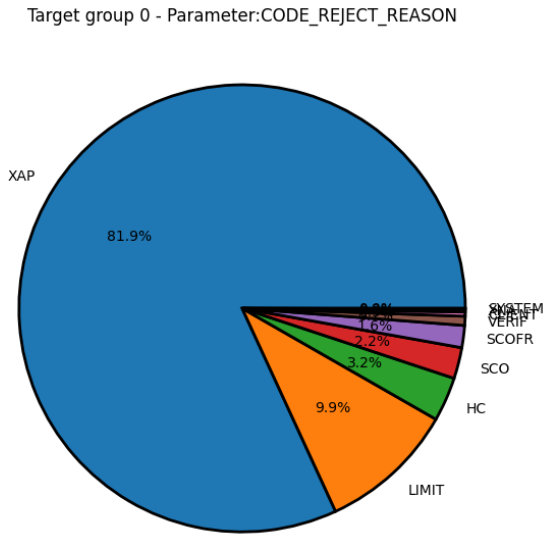

Target group 1 - Parameter:NAME_HOUSING_TYPE
House / apartment
86.6%
Co-op apartment
Office apartment
Rented apartment
4.0%
Municipal apartment
6.4%
With parents

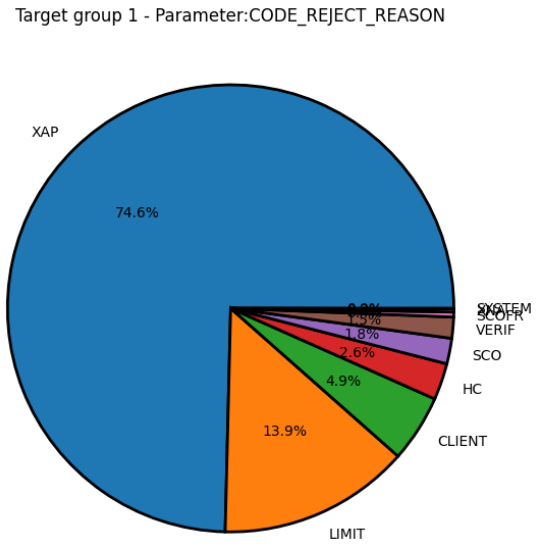

Target group 0 - Parameter:NAME_CONTRACT_TYPE_y
Cash loans
44.8%
XNA
11.1%
Revolving loans
44.2%
Consumer loans

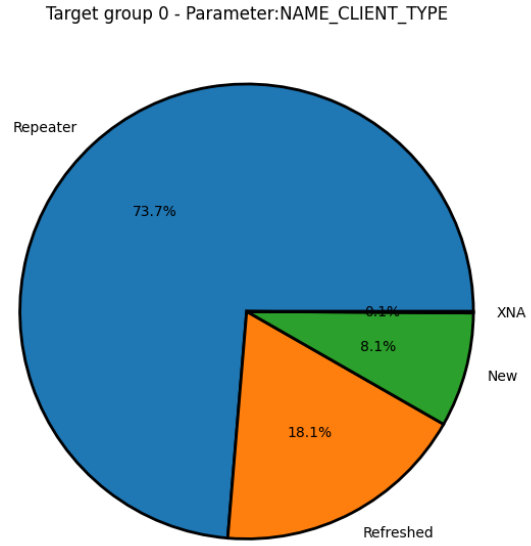

Target group 1 - Parameter:NAME_CONTRACT_TYPE_y
Consumer loans
47.2%
XNA
13.7%
Revolving loans
39.1%
Cash loans

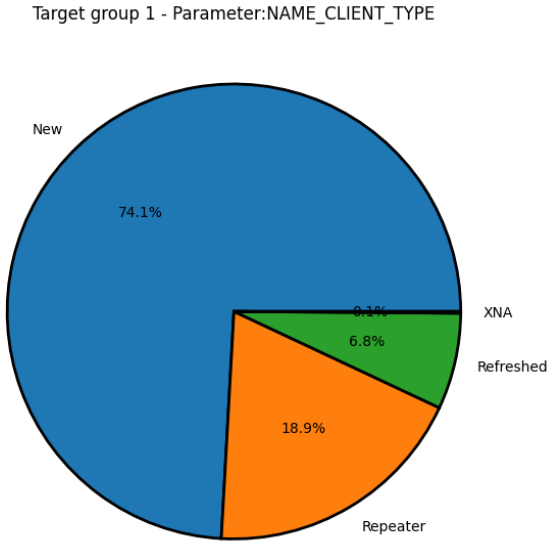

Target group 0 - Parameter:NAME_CONTRACT_STATUS
Approved
63.8%
Unused offer
16.5%
Refused
18.0%
Canceled


Target group 1 - Parameter:NAME_CONTRACT_STATUS
Approved
55.3%
Canceled
19.3%
Unused offer
23.9%
Refused









Inferences from pie chart

1. Loan type

Revolving loans have lesser payment problems compared to Cash loans : 4.8% / 7.9%

2. ****Gender****

Females are better at paying loans compared to males.

3. Car ownership

those who own a car have slightly better chances of being able to pay loans - nearly 3% better chances

4. Real estate ownership

Those who own a real estate : averages seem same for both target groups

5. **Name_suite (accompanied by)**

Those who were accompanied by Family are much better at paying loans compared to others. Hence, "NAME_TYPE_SUITE" is an important parameter to check while deciding if loan should be awarded.

6. **Income type :**

State servernts are pretty good at payments followed by Working class. Working class people TOP in the category of those who experience problem in payments.

7. **Education :**

Even though we saw, Secondary educated applicants were the most, But their biggest chunk falls in the category of those who have had problems in repayment.

Applicants with Higher education are the best at loan re-payments. hence, the most suitable candidates for awarding loan.

8. **Family Status:**

Married people are best at loan repayment while Singles experience difficulties while paying back in most cases. Very important.

9. Housing type:

same in both cases.

10. Weekday:

Same in both cases

11. Contract status of previous application:

Those who were approved loan in previous application are good at paying back in current loan also.

12. **Client Type**

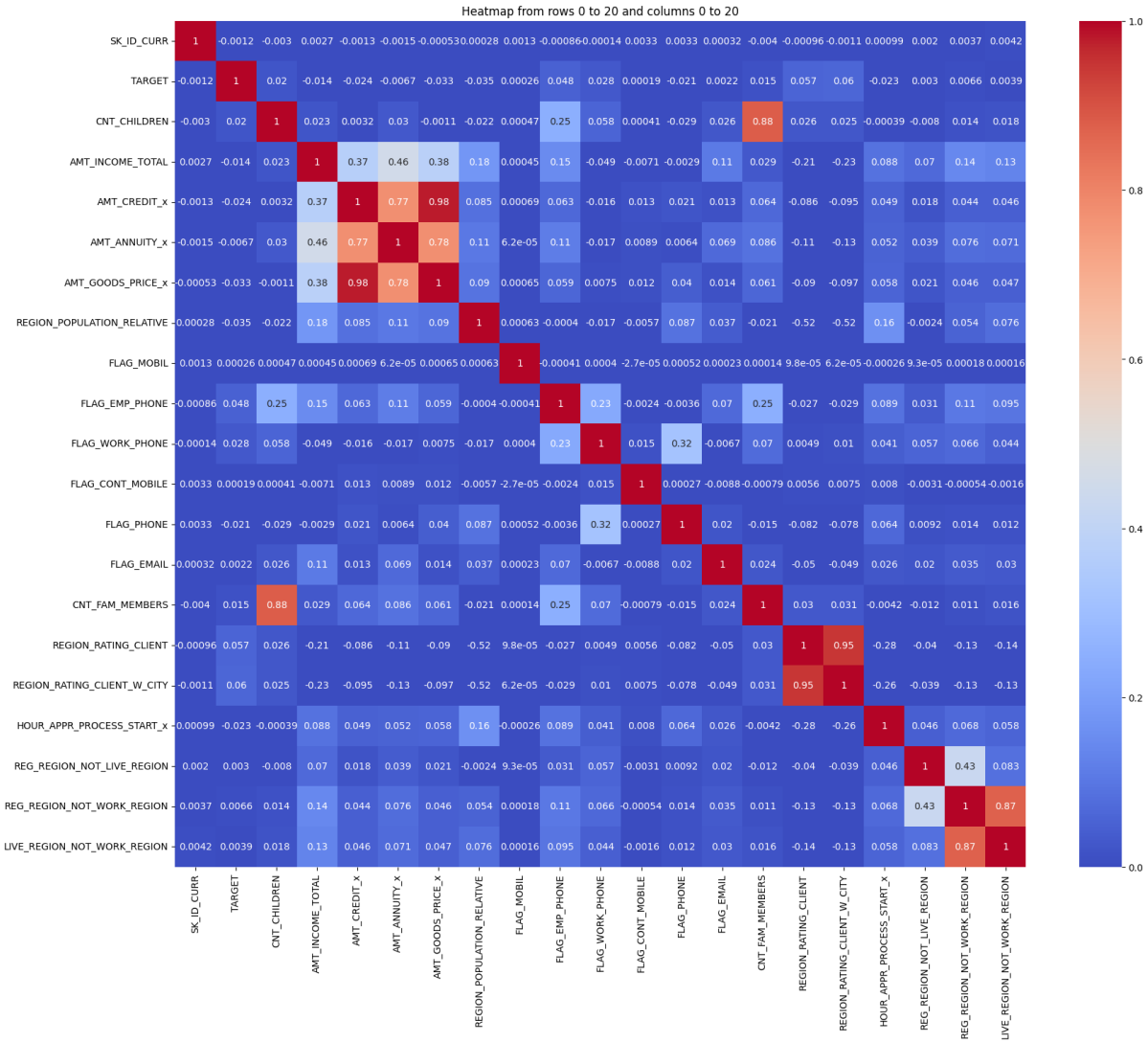
Repeaters are the best at paying loans while New applicants are the riskiest as they experience the maximum cases of payment difficulties.

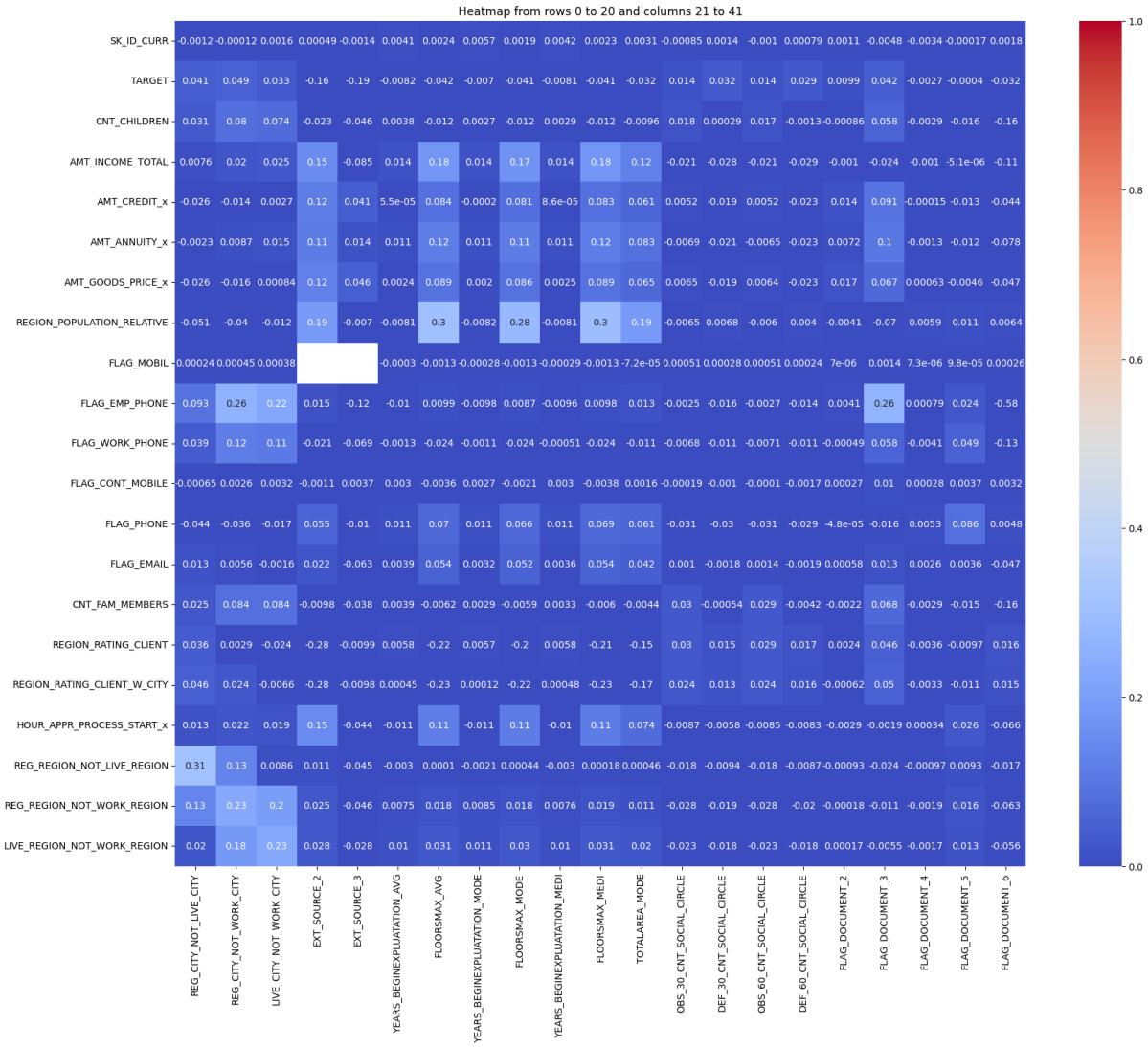
13. Product type

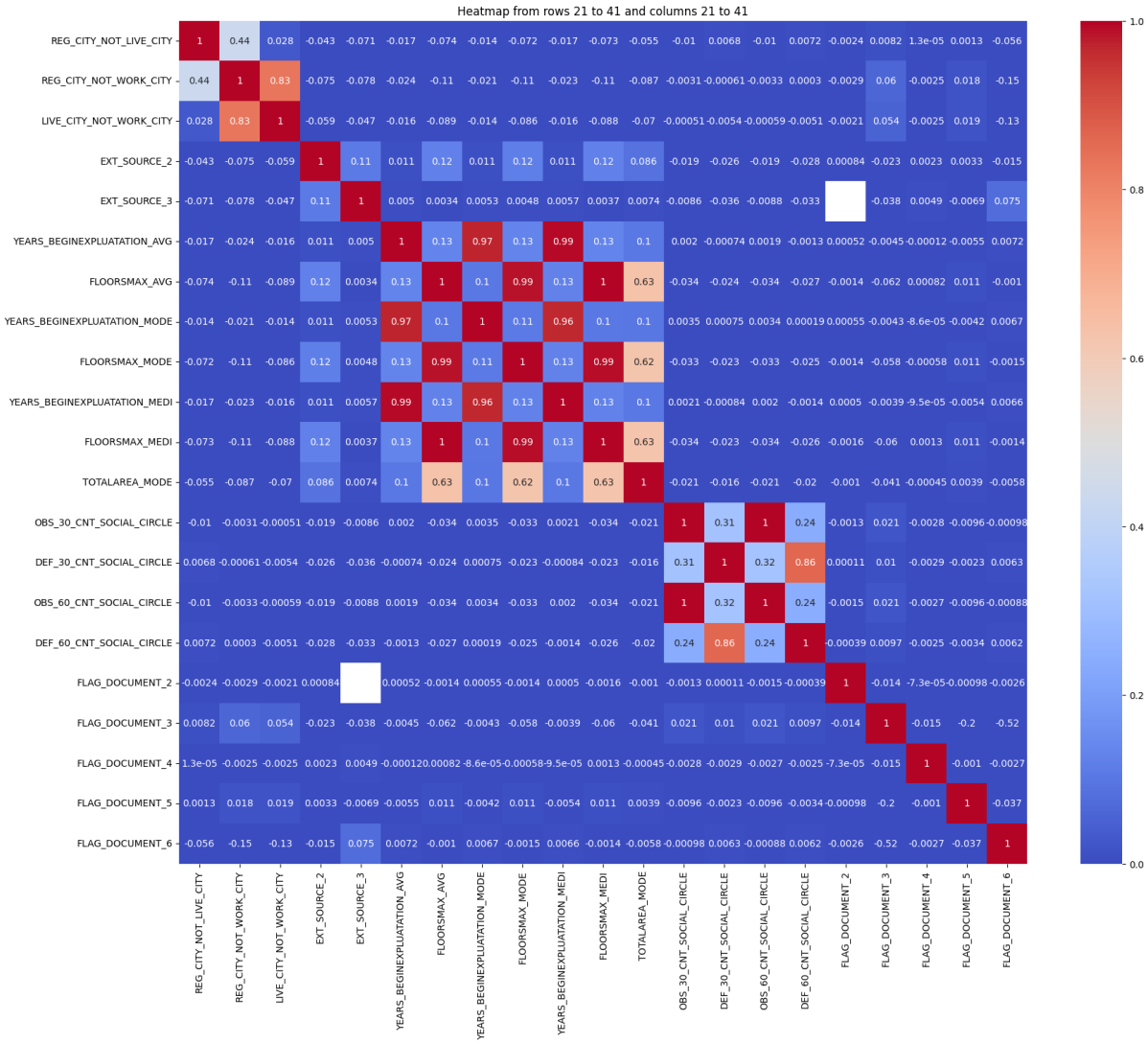
those who were using Product type x-sell in previous applications have good payment record. those with XNA are experiencing most payment problems

All other charts are not giving much - so ignoring them.

Heatmaps:


Heatmap from rows 0 to 20 and columns 0 to 20
SK_ID_CURR
TARGET
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT_x
AMT_ANNUITY_x
AMT_GOODS_PRICE_x
REGION_POPULATION_RELATIVE
FLAG_MOBIL
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
HOUR_APPR_PROCESS_START_x
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION
1.0
0.8
0.6
0.4
0.2
0.0


Heatmap from rows 0 to 20 and columns 21 to 41
SK_ID_CURR -0.0012 -0.00012 0.0016 0.00049 -0.0014 0.0041 0.0024 0.0057 0.0019 0.0042 0.0023 0.0031 -0.00085 0.0014 -0.001 0.00079 0.0011 -0.0048 -0.0034 -0.00017 0.0018
TARGET 0.041 0.049 0.033 -0.16 -0.19 -0.0082 -0.042 -0.007 -0.041 -0.0081 -0.041 -0.032 0.014 0.032 0.014 0.029 0.0099 0.042 -0.0027 -0.0004 -0.032
CNT_CHILDREN 0.031 0.08 0.074 -0.023 -0.046 0.0038 -0.012 0.0027 -0.012 0.0029 -0.012 -0.0096 0.018 0.00029 0.017 -0.0013 -0.00086 0.058 -0.0029 -0.016 -0.16
AMT_INCOME_TOTAL 0.0076 0.02 0.025 0.15 -0.085 0.014 0.18 0.014 0.17 0.014 0.18 0.12 -0.021 -0.028 -0.021 -0.029 -0.001 -0.024 -0.001 -5.1e-06 -0.11
AMT_CREDIT_x -0.026 -0.014 0.0027 0.12 0.041 5.5e-05 0.084 -0.0002 0.081 8.6e-05 0.083 0.061 0.0052 -0.019 0.0052 -0.023 0.014 0.091 -0.00015 -0.013 -0.044
AMT_ANNUITY_x -0.0023 0.0087 0.015 0.11 0.014 0.011 0.12 0.011 0.11 0.011 0.12 0.083 -0.0069 -0.021 -0.0065 -0.023 0.0072 0.1 -0.0013 -0.012 -0.078
AMT_GOODS_PRICE_x -0.026 -0.016 0.00084 0.12 0.046 0.0024 0.089 0.002 0.086 0.0025 0.089 0.065 0.0065 -0.019 0.0064 -0.023 0.017 0.067 0.00063 -0.0046 -0.047
REGION_POPULATION_RELATIVE -0.051 -0.04 -0.012 0.19 -0.007 -0.0081 0.3 -0.0082 0.28 -0.0081 0.3 0.19 -0.0065 0.0068 -0.006 0.004 -0.0041 -0.07 0.0059 0.011 0.0064
FLAG_MOBIL 0.00024 0.00045 0.00038 -0.0003 -0.0013 -0.00028 -0.0013 -0.00029 -0.0013 -7.2e-05 0.00051 0.00028 0.00051 0.00024 7e-06 0.0014 7.3e-06 9.8e-05 0.00026
FLAG_EMP_PHONE 0.093 0.26 0.22 0.015 -0.12 -0.01 0.0099 -0.0098 0.0087 -0.0096 0.0098 0.013 -0.0025 -0.016 -0.0027 -0.014 0.0041 0.26 0.00079 0.024 -0.58
FLAG_WORK_PHONE 0.039 0.12 0.11 -0.021 -0.069 -0.0013 -0.024 -0.0011 -0.024 -0.00051 -0.024 -0.011 -0.0068 -0.011 -0.0071 -0.011 -0.00049 0.058 -0.0041 0.049 -0.13
FLAG_CONT_MOBILE -0.00065 0.0026 0.0032 -0.0011 0.0037 0.003 -0.0036 0.0027 -0.0021 0.003 -0.0038 0.0016 -0.00019 -0.001 -0.0001 -0.0017 0.00027 0.01 0.00028 0.0037 0.0032
FLAG_PHONE -0.044 -0.036 -0.017 0.055 -0.01 0.011 0.07 0.011 0.066 0.011 0.069 0.061 -0.031 -0.03 -0.031 -0.029 -4.8e-05 -0.016 0.0053 0.086 0.0048
FLAG_EMAIL 0.013 0.0056 -0.0016 0.022 -0.063 0.0039 0.054 0.0032 0.052 0.0036 0.054 0.042 0.001 -0.0018 0.0014 -0.0019 0.00058 0.013 0.0026 0.0036 -0.047
CNT_FAM_MEMBERS 0.025 0.084 0.084 -0.0098 -0.038 0.0039 -0.0062 0.0029 -0.0059 0.0033 -0.006 -0.0044 0.03 -0.00054 0.029 -0.0042 -0.0022 0.068 -0.0029 -0.015 -0.16
REGION_RATING_CLIENT 0.036 0.0029 -0.024 -0.28 -0.0099 0.0058 -0.22 0.0057 -0.2 0.0058 -0.21 -0.15 0.03 0.015 0.029 0.017 0.0024 0.046 -0.0036 -0.0097 0.016
REGION_RATING_CLIENT_W_CITY 0.046 0.024 -0.0066 -0.28 -0.0098 0.00045 -0.23 0.00012 -0.22 0.00048 -0.23 -0.17 0.024 0.013 0.024 0.016 -0.00062 0.05 -0.0033 -0.011 0.015
HOUR_APPR_PROCESS_START_x 0.013 0.022 0.019 0.15 -0.044 -0.011 0.11 -0.011 0.11 -0.01 0.11 0.074 -0.0087 -0.0058 -0.0085 -0.0083 -0.0029 -0.0019 0.00034 0.026 -0.066
REG_REGION_NOT_LIVE_REGION 0.31 0.13 0.0086 0.011 -0.045 -0.003 0.0001 -0.0021 0.00044 -0.003 0.00018 0.00046 -0.018 -0.0094 -0.018 -0.0087 -0.00093 -0.024 -0.00097 0.0093 -0.017
REG_REGION_NOT_WORK_REGION 0.13 0.23 0.2 0.025 -0.046 0.0075 0.018 0.0085 0.018 0.0076 0.019 0.011 -0.028 -0.019 -0.028 -0.02 -0.00018 -0.011 -0.0019 0.016 -0.063
LIVE_REGION_NOT_WORK_REGION 0.02 0.18 0.23 0.028 -0.028 0.01 0.031 0.011 0.03 0.01 0.031 0.02 -0.023 -0.018 -0.023 -0.018 0.00017 -0.0055 -0.0017 0.013 -0.056
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
EXT_SOURCE_2
EXT_SOURCE_3
YEARS_BEGINEXPLUATATION_AVG
FLOORSMAX_AVG
YEARS_BEGINEXPLUATATION_MODE
FLOORSMAX_MODE
YEARS_BEGINEXPLUATATION_MEDI
FLOORSMAX_MEDI
TOTALAREA_MODE
OBS_30_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
FLAG_DOCUMENT_2
FLAG_DOCUMENT_3
FLAG_DOCUMENT_4
FLAG_DOCUMENT_5
FLAG_DOCUMENT_6
1.0
0.8
0.6
0.4
0.2
0.0



Inferences from heatmaps:

Strong correlations found between:

1. AMT_GOODS_PRICE_x & AMT_CREDIT_X

2. AMT_CREDIT_x & AMT_ANNUITY_x

3. AMT_ANNUITY_x & AMT_GOODS_PRICE_x

4. CNT_CHILDREN & CNT_FAM_MEMBERS

5. DAYS_LAST_DUE vs DAYS_TERMINATION

6. REGION_RATING_CLIENT_W_CITY & REGION_RATING_CLIENT

7. REG_REGION_NOT_WORK_REGION & LIVE_REGION_NOT_WORK_REGION

Moderatley effective relations

1. CNT_PAYMENT vs AMT_APPLICATION

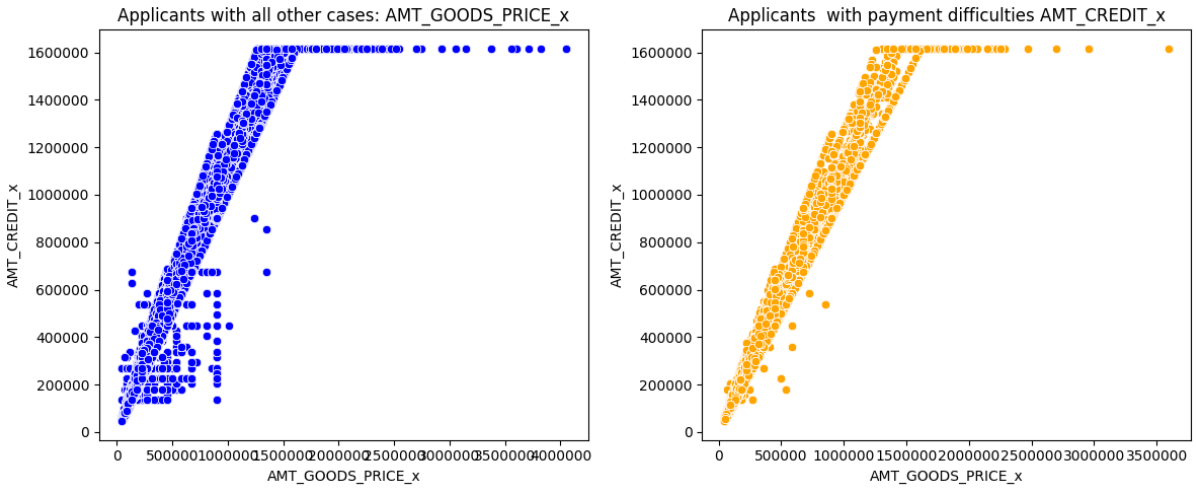
2. AMT_CREDIT_y vs CNT_PAYMENT

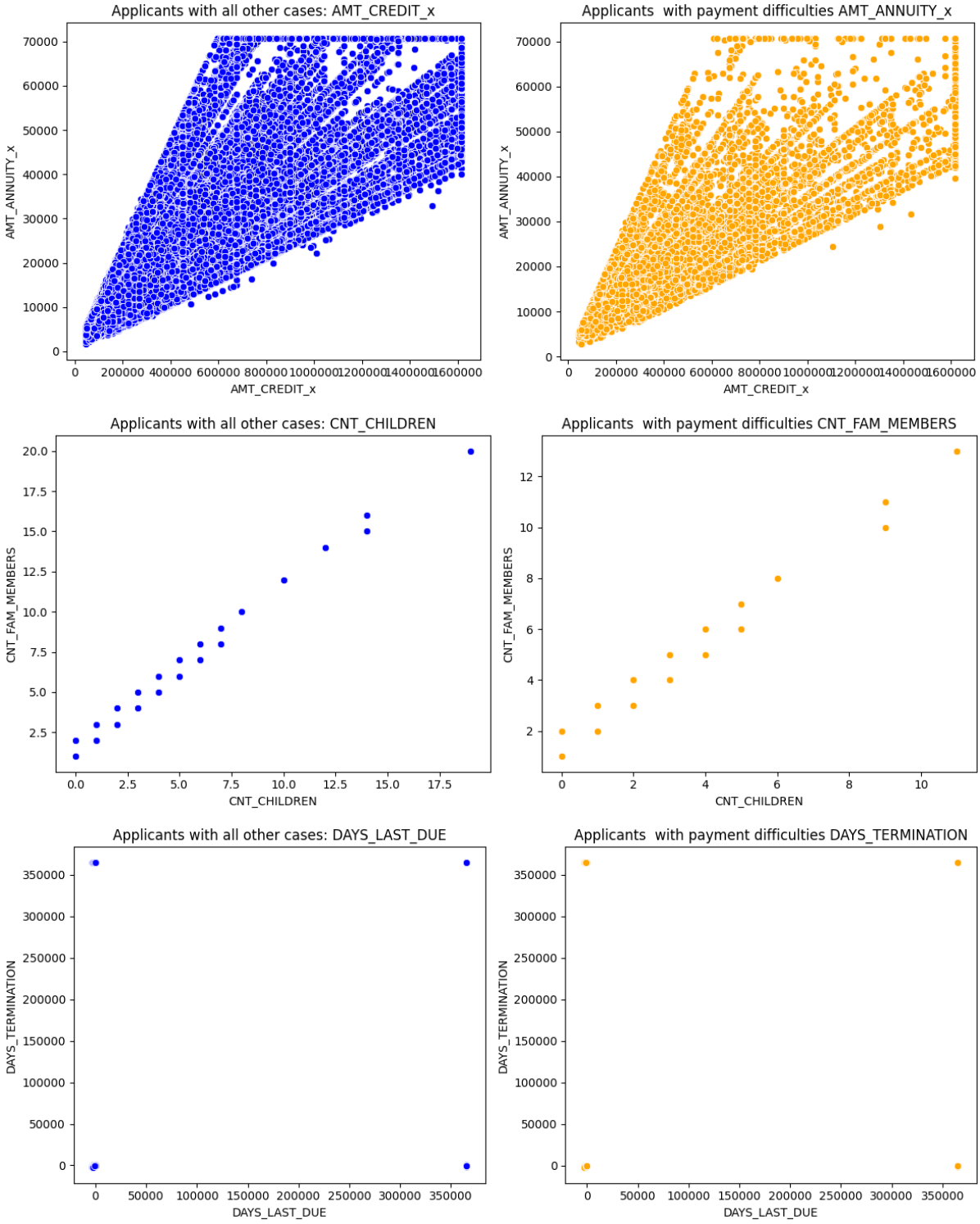
3. AMT_GOODS_PRICE_y =CNT_PAYMENT

******This is only for understanding - as their direct correlation with TARGET GROUP is not seen here******

****Running scatterplot on all with hue of TARGET****

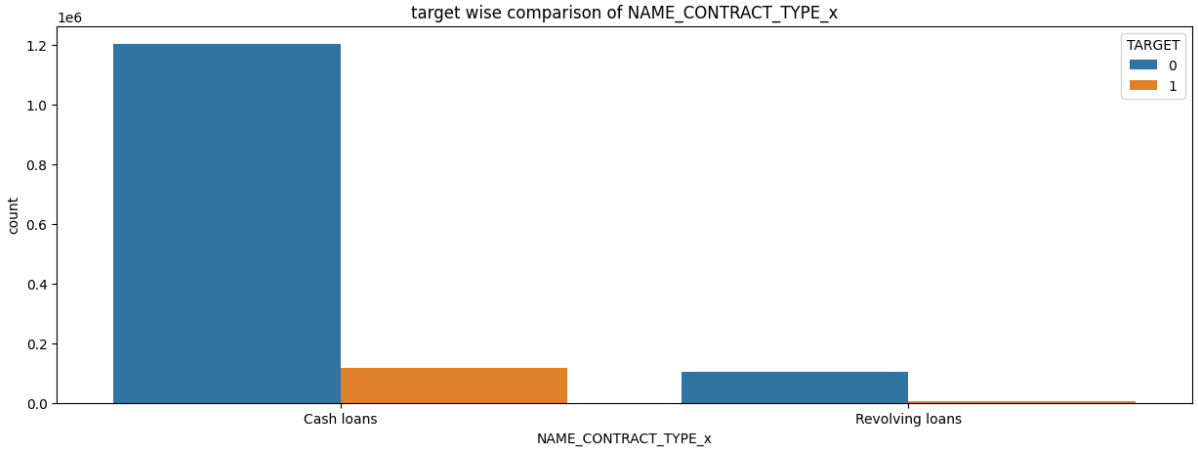
Scatterplots :

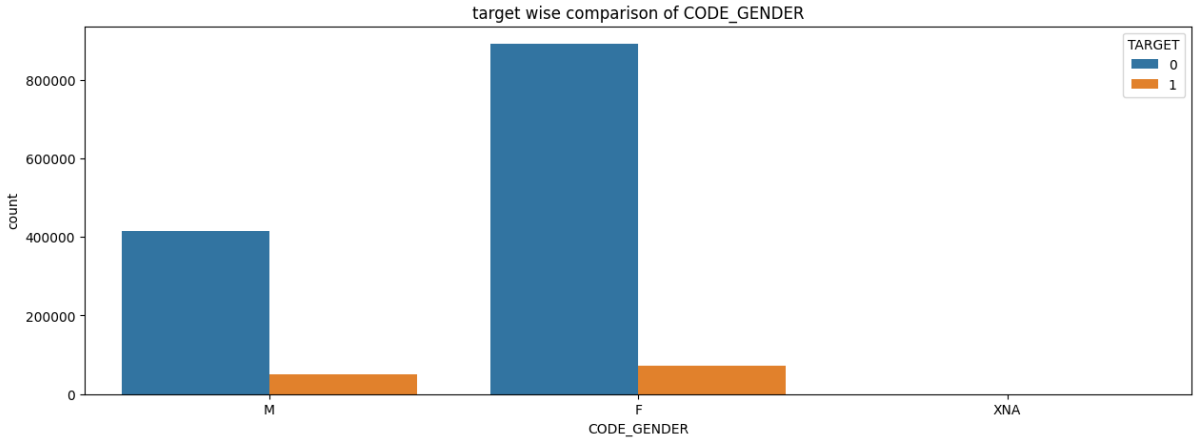


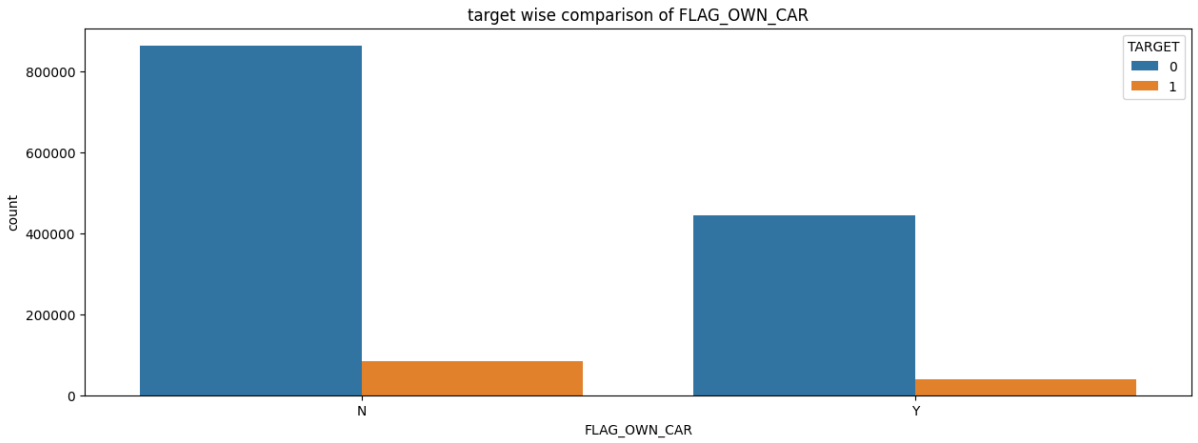


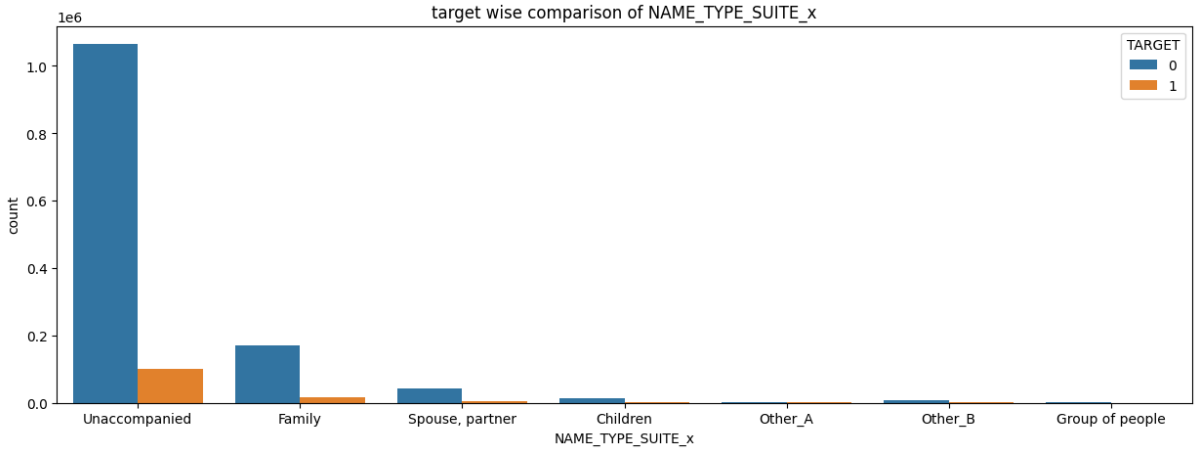
Inference : Although there are parameters that show high correlation and some that do not - yet there is no such combination that shows a different behavior for those with Payment difficulties vs those with all other cases

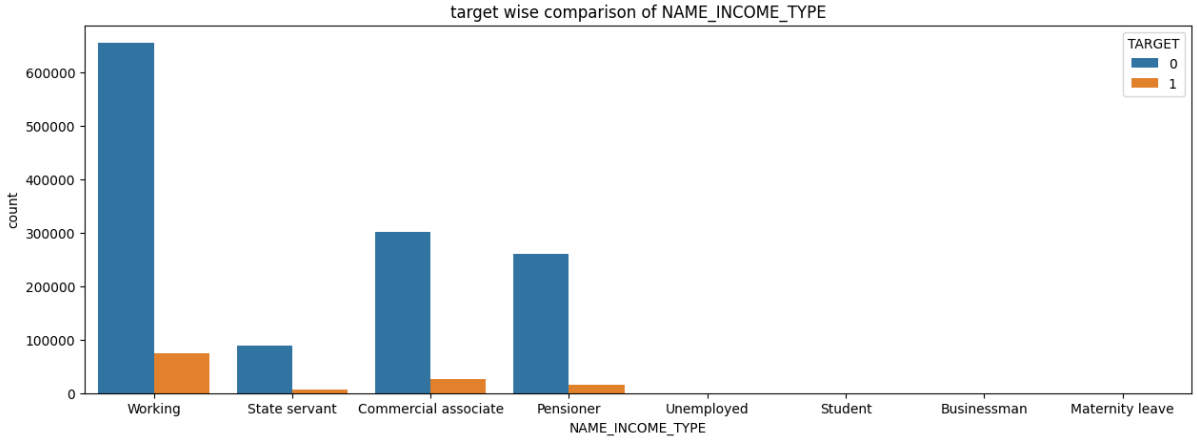
Count plots :

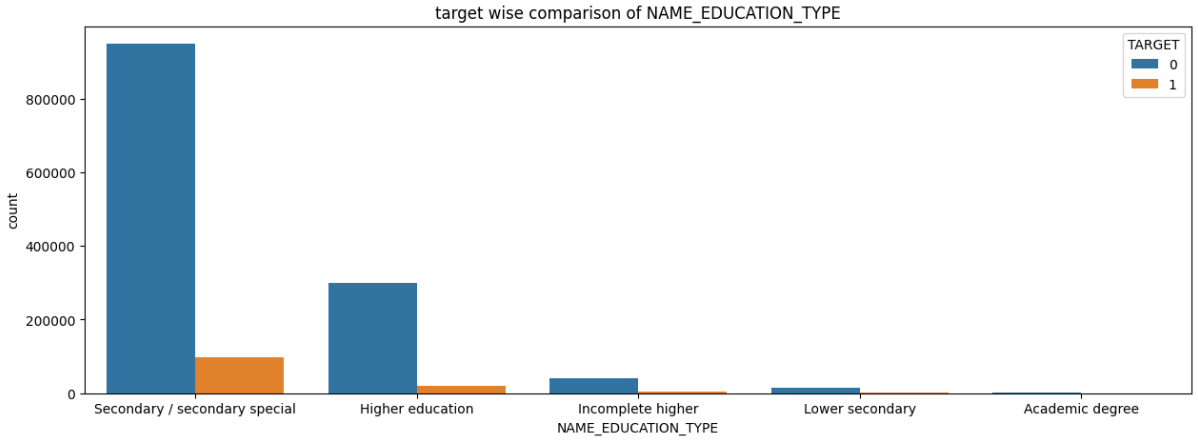

target wise comparison of NAME_CONTRACT_TYPE_x
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
count
TARGET
0
1
Cash loans
Revolving loans
NAME_CONTRACT_TYPE_x

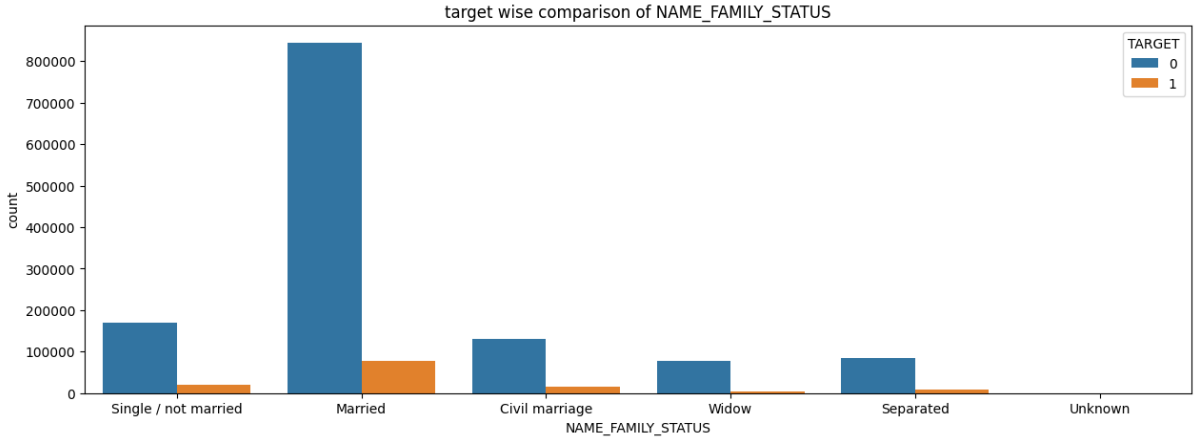

target wise comparison of CODE_GENDER
800000
600000
400000
200000
0
count
TARGET
0
1
M
F
XNA
CODE_GENDER

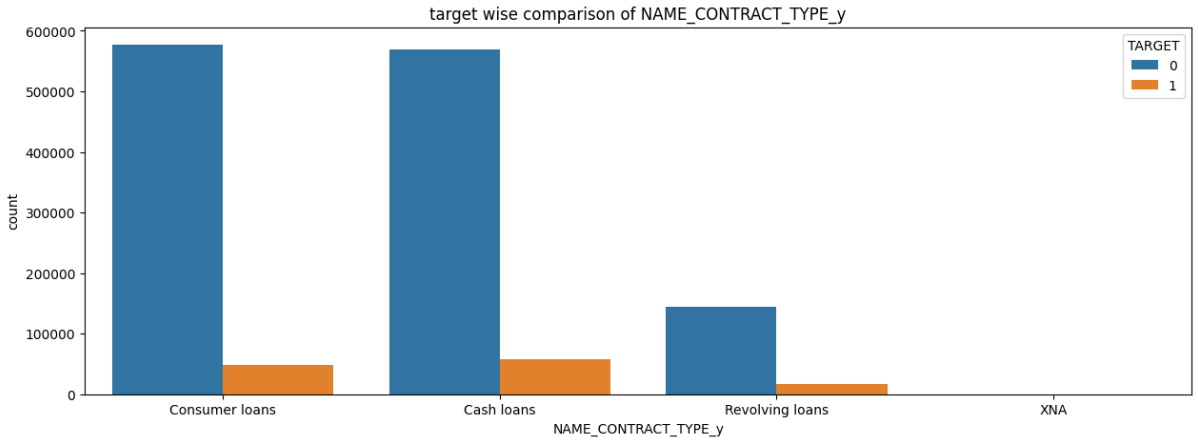

target wise comparison of FLAG_OWN_CAR
800000
600000
400000
200000
0
count
TARGET
0
1
N
Y
FLAG_OWN_CAR

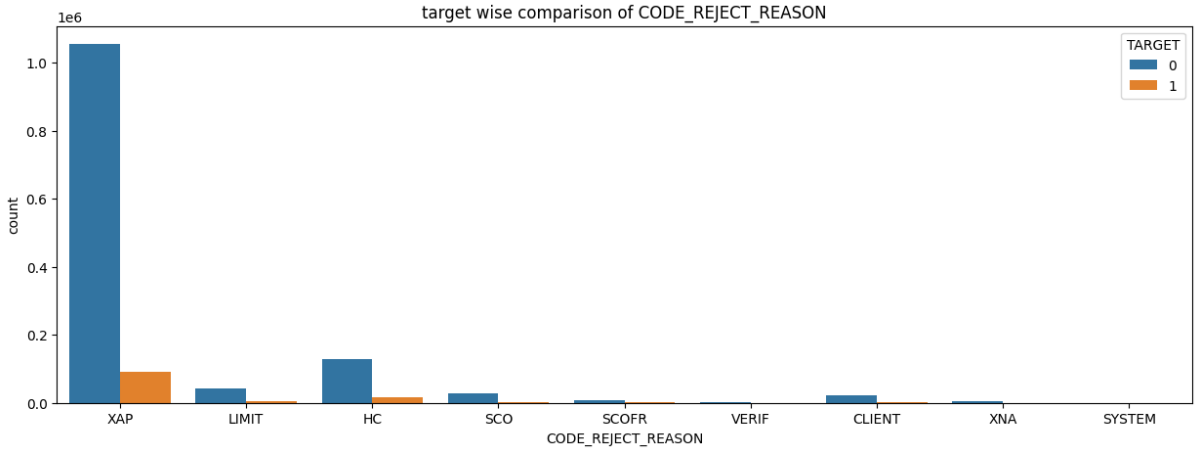

target wise comparison of NAME_TYPE_SUITE_x
1e6
1.0
0.8
0.6
0.4
0.2
0.0
count
TARGET
0
1
Unaccompanied
Family
Spouse, partner
Children
Other_A
Other_B
Group of people
NAME_TYPE_SUITE_x

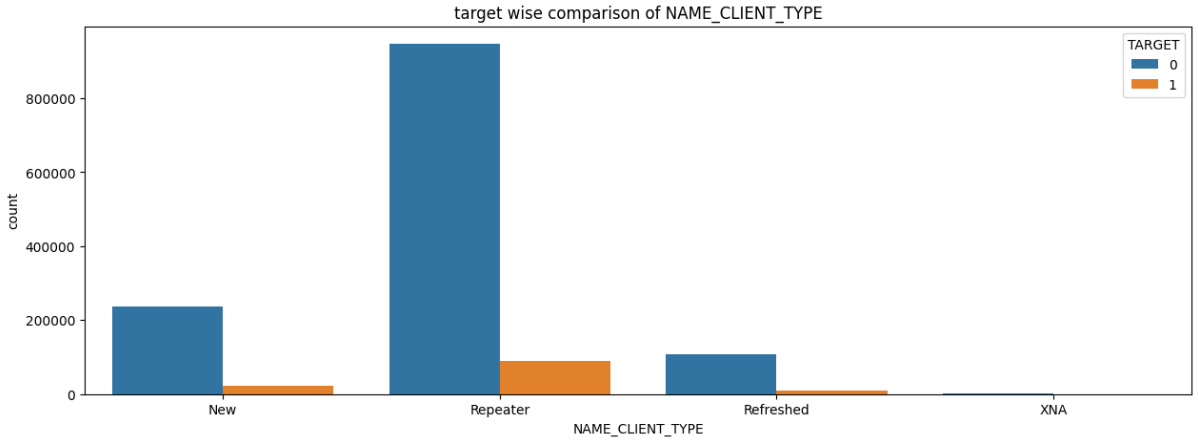

target wise comparison of NAME_INCOME_TYPE
600000
500000
400000
300000
200000
100000
0
count
TARGET
0
1
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
NAME_INCOME_TYPE


target wise comparison of NAME_EDUCATION_TYPE
800000
600000
400000
200000
0
count
TARGET
0
1
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE


target wise comparison of NAME_FAMILY_STATUS
800000
700000
600000
500000
400000
300000
200000
100000
0
count
TARGET
0
1
Single / not married
Married
Civil marriage
Widow
Separated
Unknown
NAME_FAMILY_STATUS







Count plot inferences:

1. Contract type wise :

   very high imbalance on Cash loans.

   ratio of 0 and 1 seems similar for all loan types. hence no distinguishing factor.

2. Gender :

Females apply far more for the loans compared to men. However, male feel find more difficulties paying.. to be proven with ratio.

3. Car Ownership:

very high number of people who apply loans do not have car.

no visible differentiation in behavior between those who own and those who dont own a car.

4. Realty ownership:

there are more applicants with realty who apply for loan than those who don't

same like car ownership, no differentiation.

5. Suite type (Who accompanied customer during prev application) - (OLD and NEW)

People mostly visit unaccompanied, followed by Family.

visibly no differentiation.. ratio to be checked.

6. Income type:

Working people form biggest chunk of applicants.

payment difficulty ratio seems higher for commercial associates and Pensioners. ratio calculation to be done to objectify number

7. Education:

applicants are mostly either secondary educated and then followed by those with "higher education". all others constitute a small part.

No visible differentiating behavior

8. Family status:

Most of applicants are married, followed by Single.

those in civil marriage have higher visible ratio of applicants with payment difficulties compared to others. objectify with ratio calculation.

9. housing type :

Most of applicants have House/Apartment.

those who live with parents have higher ratio of applicants with payment difficulties

10. Application processing day (OLD and Current)

no visible difference. lesser business of credit on weekends.

11. EMERGENCYSTATE_MODE:

this parameter is heavily biased on "NO"

no visible differentiation

12. Contract type (loan type):

Consumer and cash loans are the most in demand, though almost equal to each other. Revolving loans on third number.

Applicants are finding slightly more difficulty on cash loans compared to Consumer loans. THis may be due to amount of loan. can be checked.

13. Contract status (from previous application):

those who were awarded loan last time are the highest with ease of payment in current applications as well.

However, differentiation ratio across types to be checked.. visibly not identifiable.

14. Payment type:

Most of people prefer to pay by "Cash through the bank".

XNA type applicants have higher numbers having issues in paying back.

15. Rejection_reason:

for prev applications, XAP was the biggest reason for loan rejection. however, a very good percentage of those have not faced difficulty in paying back now.

HC (High credit ratio ?) applicants in previous applications, have higher ratio of paying back again

16. Client type:

Repeaters are the biggest chunk of the current applicants as well.

differentiation ratio seems same across all types.

17. Portfolio :

POS is the most common Portfolio type, followed by Cash.

ratio to be calculated . "CASH" category seems to have higher ratio of payment difficulties

18. Channel_type:

Credit and cash offices bring the maximum business for company.

County wide seems to have higher ratio. to be checked

19. yield group (Loan rate grouping):

XNA is the most commmon type of RoI group though not clear about its meaning.

ratio seems similar across categories.

The analysis indicates a clear need for calculating ratio of all cateories to confirm. This ratio or percentage is more easily visible in pie plots. Hence, together with inference of pie plots, final outcome can be concluded.

Step 10 : CONCLUSION

Following are the most important Parameters that affect the outcome as mentioned in the above cell

Gender , Income type, Education , Family Status , Client type

out of these parameters, the most effective correlations seen are :

1. Gender

2. Education

3. Family status

Project Conclusion:

For use case : important insights are derived from the project for the client if any.

for Personal benefit : This project has given me enough confidence to handle big data and deal with multiple various parameters without having deep interactive understanding from the client.