Credit EDA Case Study
DS C17

- **Application Data**
- **Previous Application**

Problem statement

• To study and analyse the patterns of given data set of a finance company which specializes in lending various types of loans to urban customers and help to minimize the risk of losing company money while lending to customers.

Solution Approach.

- Analyse the application Data set.
- Identify the key metrics/variable
- Perform univariate analysis on each of the identified key metrics.
- Perform Bivariate analysis to understand patterns
- Perform above analysis on target variable (defaulters or repayers)
- Analyse the Previous application data set
- Identify the key metrics/variable
- Divide the previous application set based on status(Approved/rejected)
- Merge each of these sets with the each of the target data set
- Again perform Uni/Bi variate analysis on the merged data.
- Find correlation of key variables

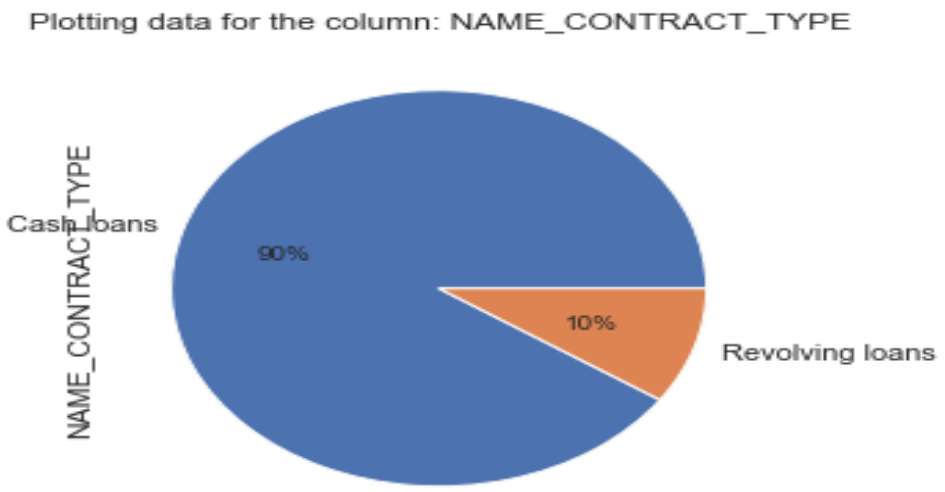
Application data

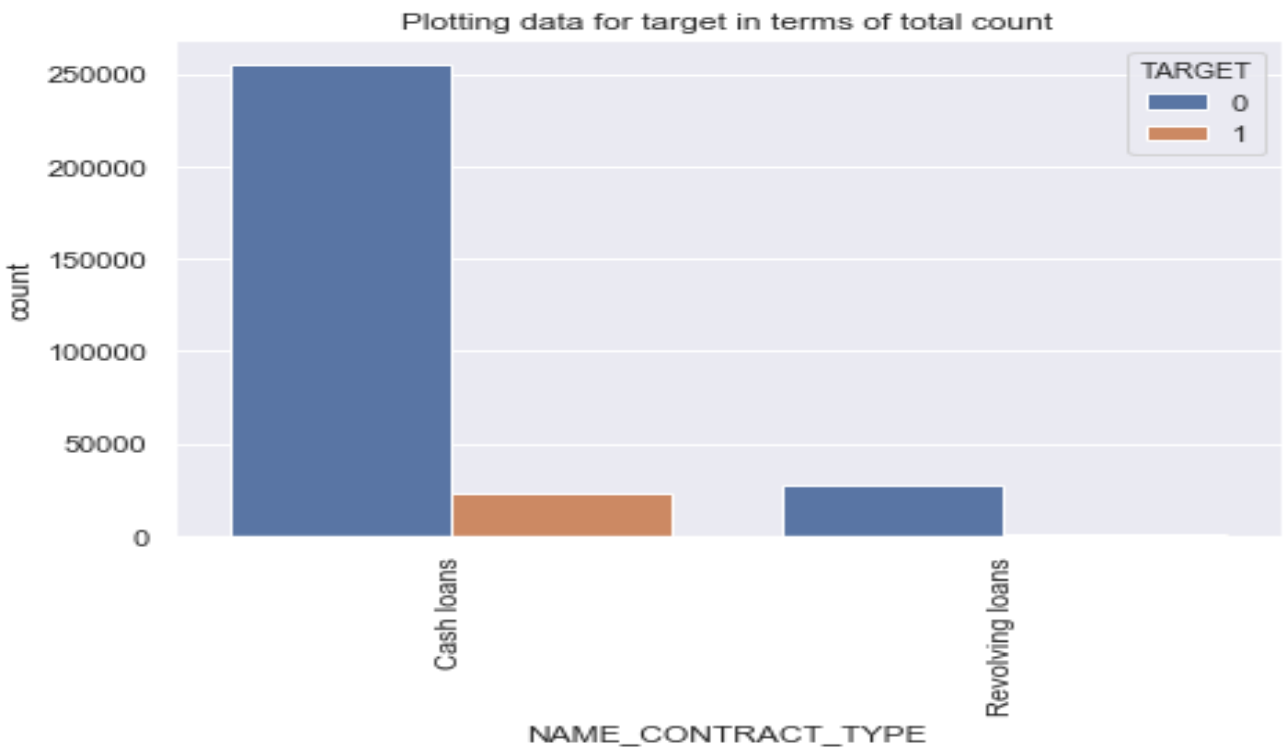
- Key Metrics
- Categorical

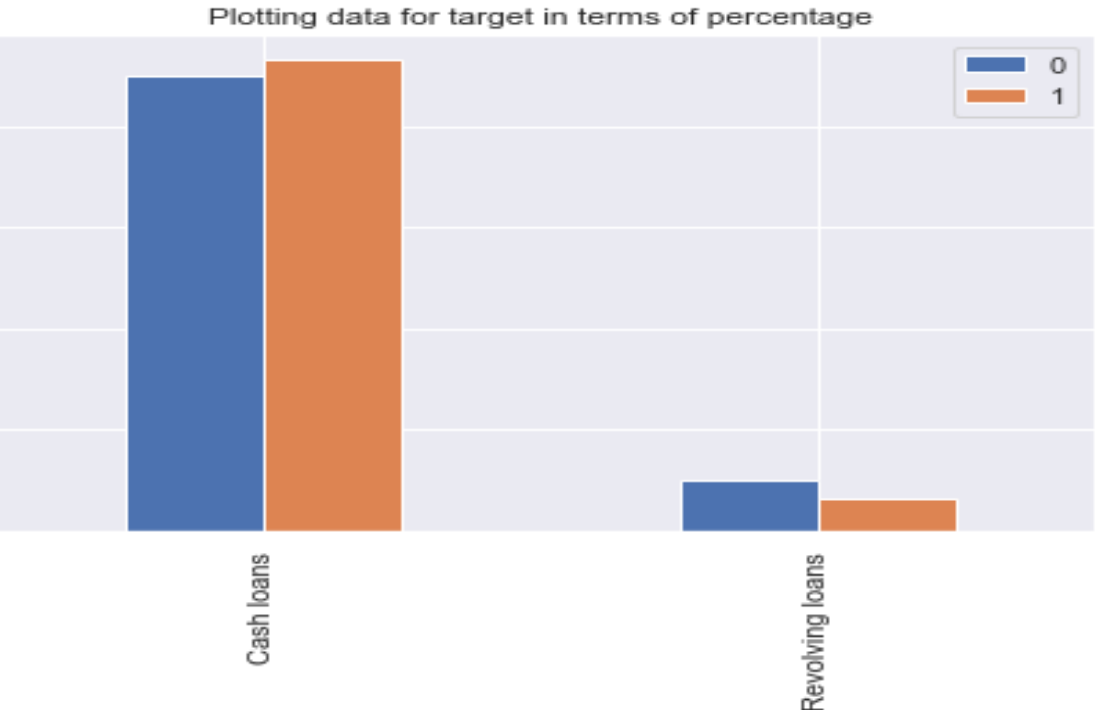
- Gender, Occupation type, family status , Income type, Loan type

- Quantitative

- Annual income, loan amount, Annuity amt, Age, Work exp.

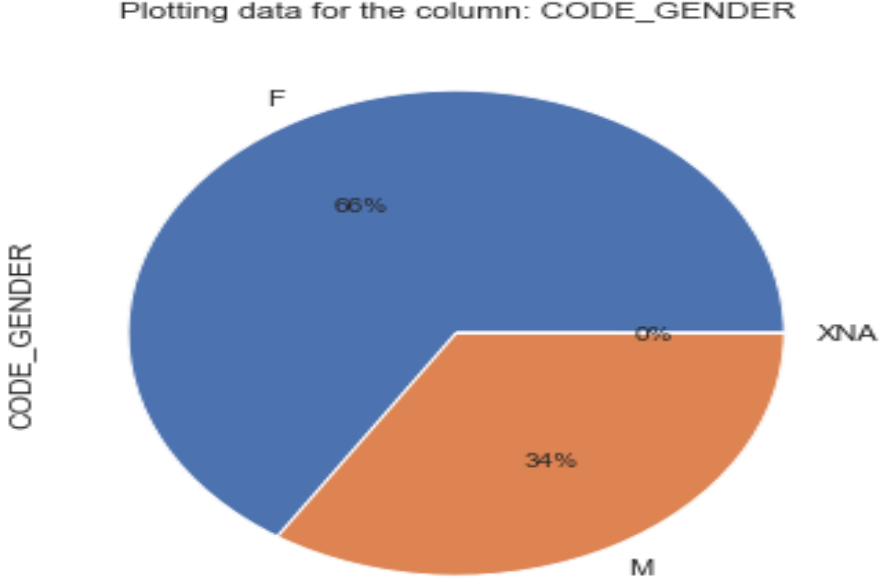


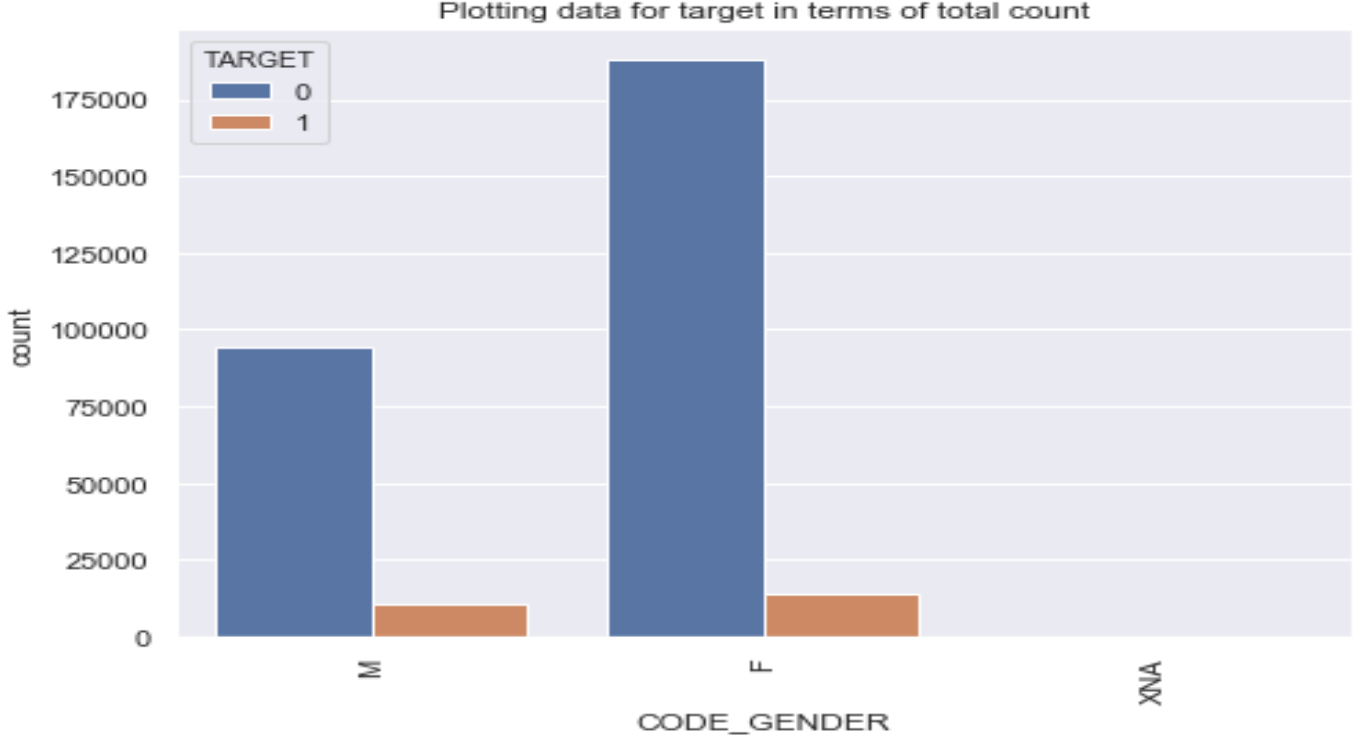




NAME_CONTRACT_TYPE

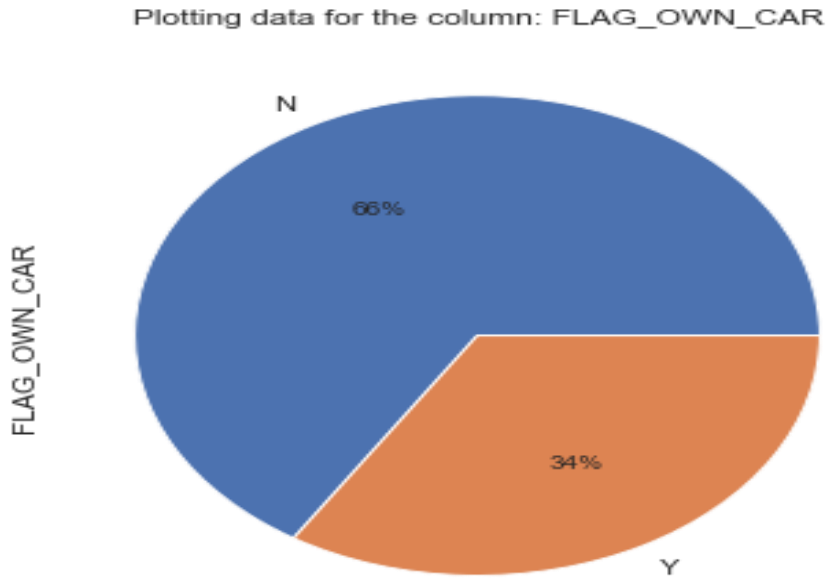
- Identification if loan is cash or revolving- As identified from the plots that Cash Loan default rate is less

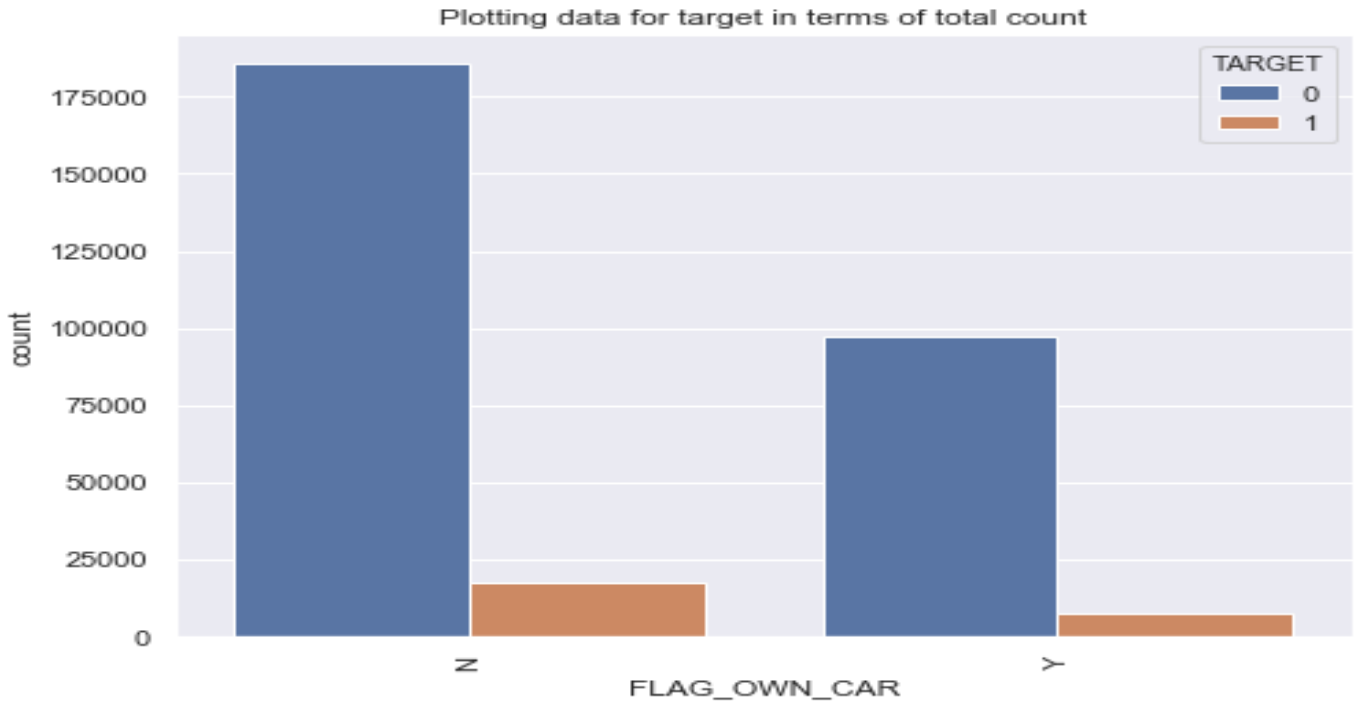


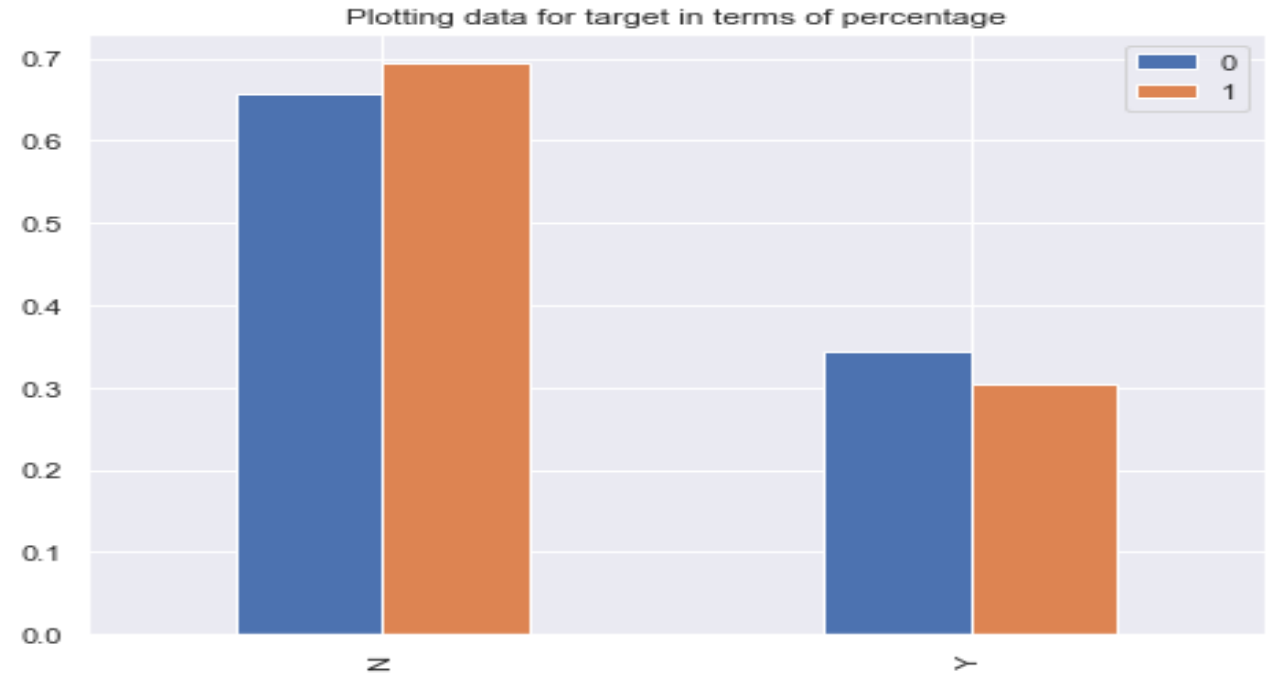


CODE_GENDER

- Gender of the client- Less number of males(hist plot) take loan but the defaulters are higher in case of males(dist plot).

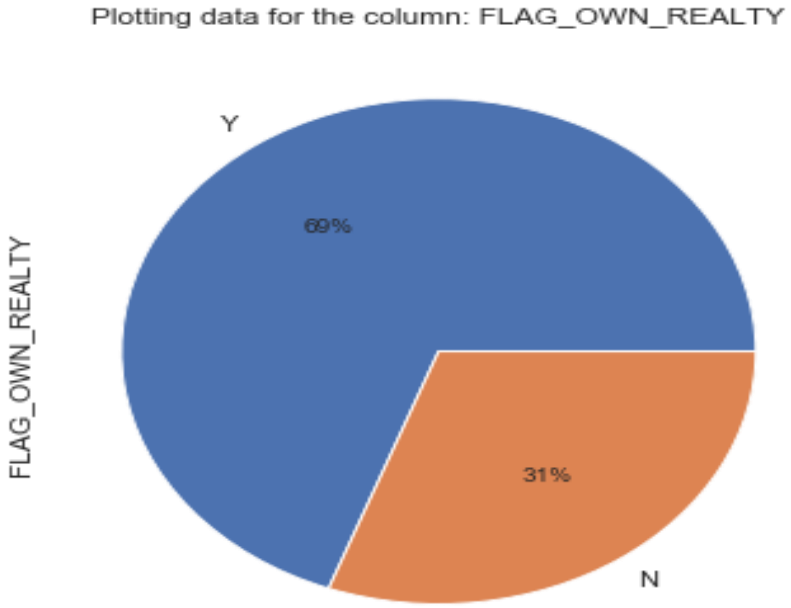


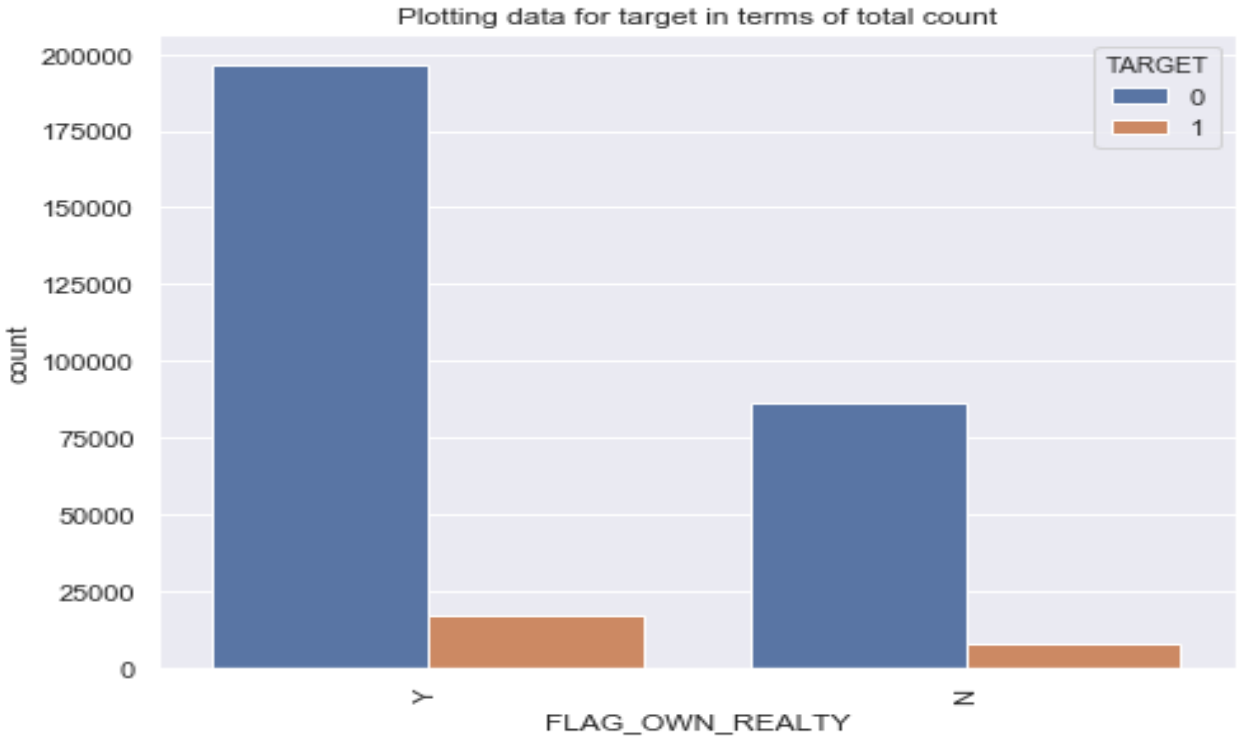




FLAG_OWN_CAR

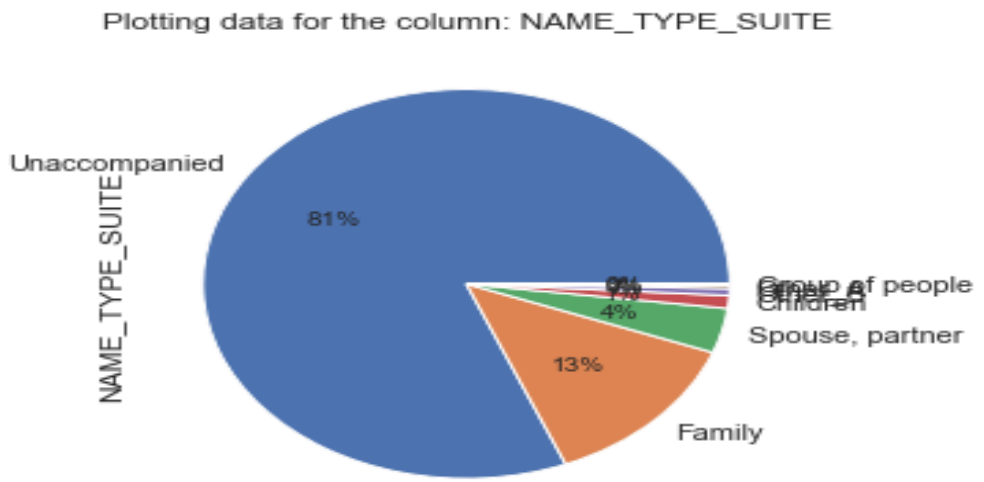
- Flag if the client owns a car- As identified from the plots that the car owners defaulter rate higher then who are not a car owner.

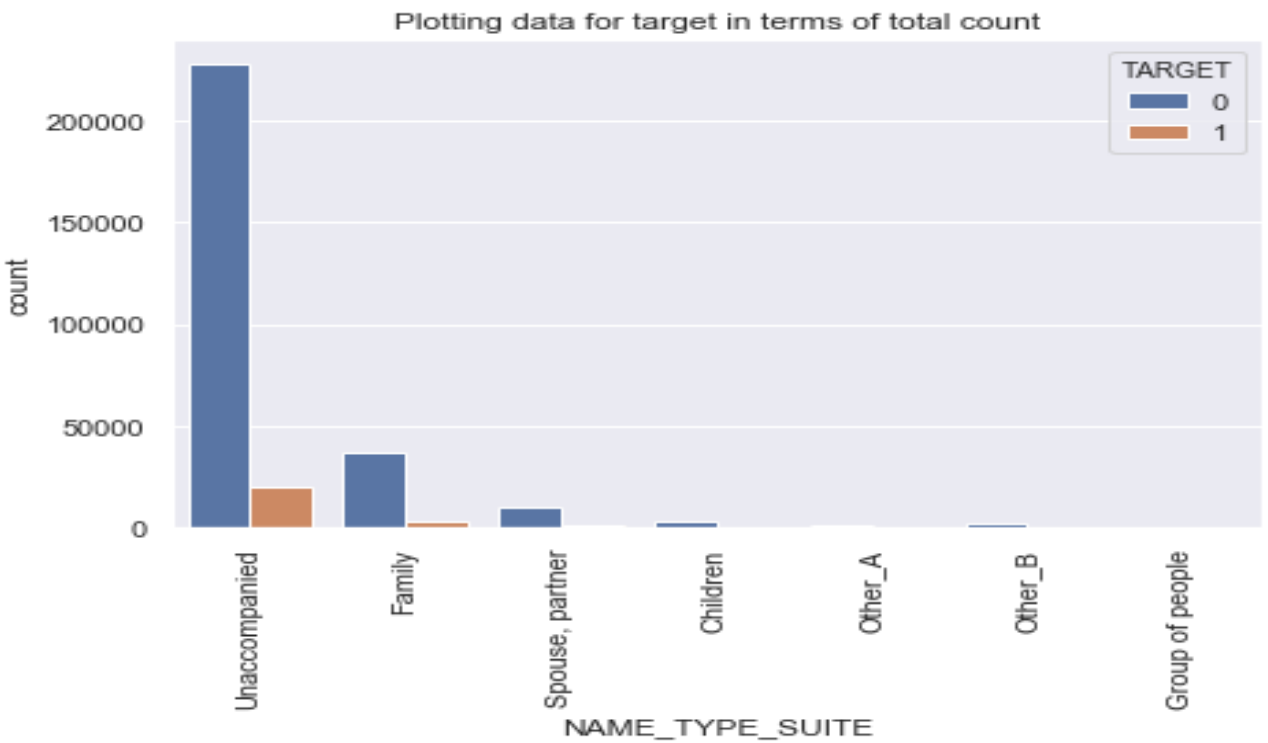


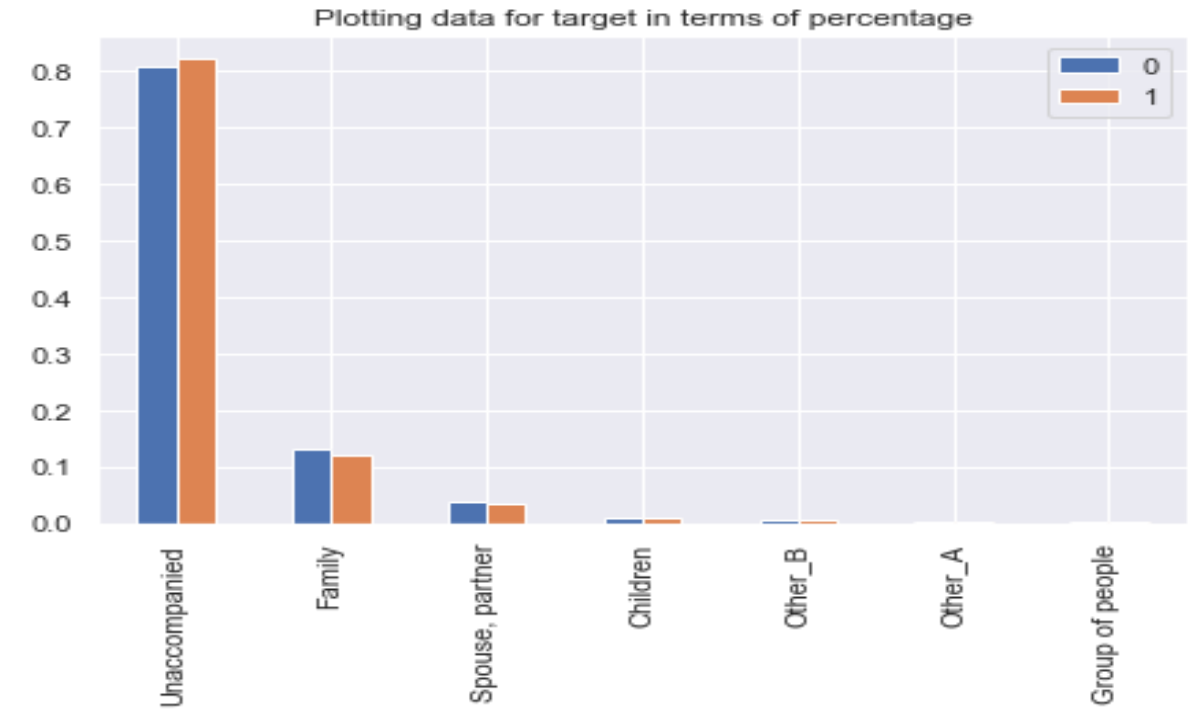


FLAG_OWN_REALTY

- Flag if client owns a house or flat- As identified from the plots that the flat owners defaulter rate higher then who are not a flat owner.

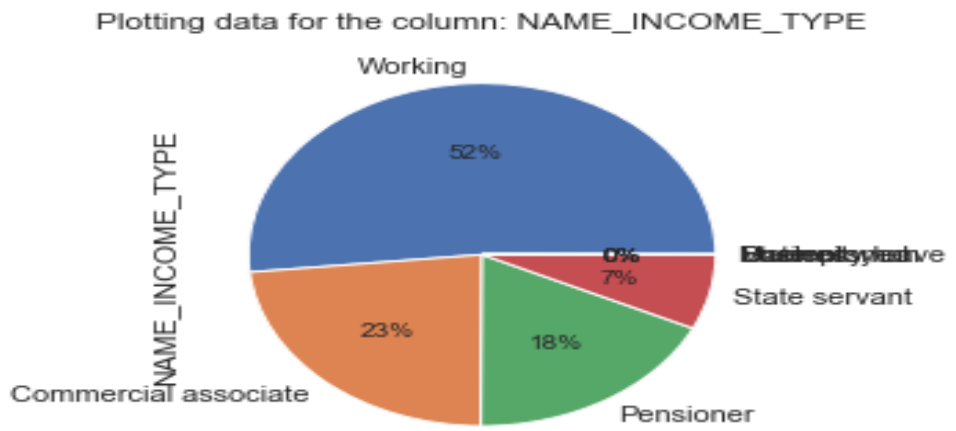


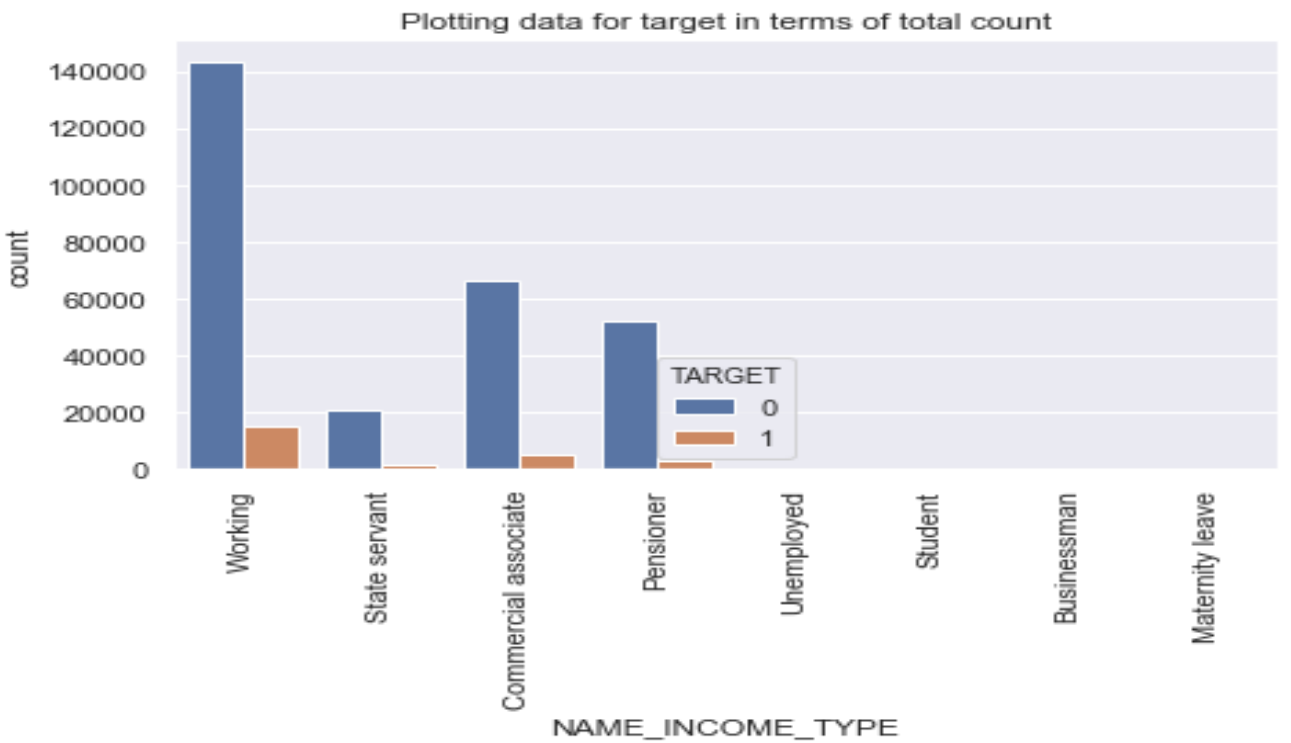


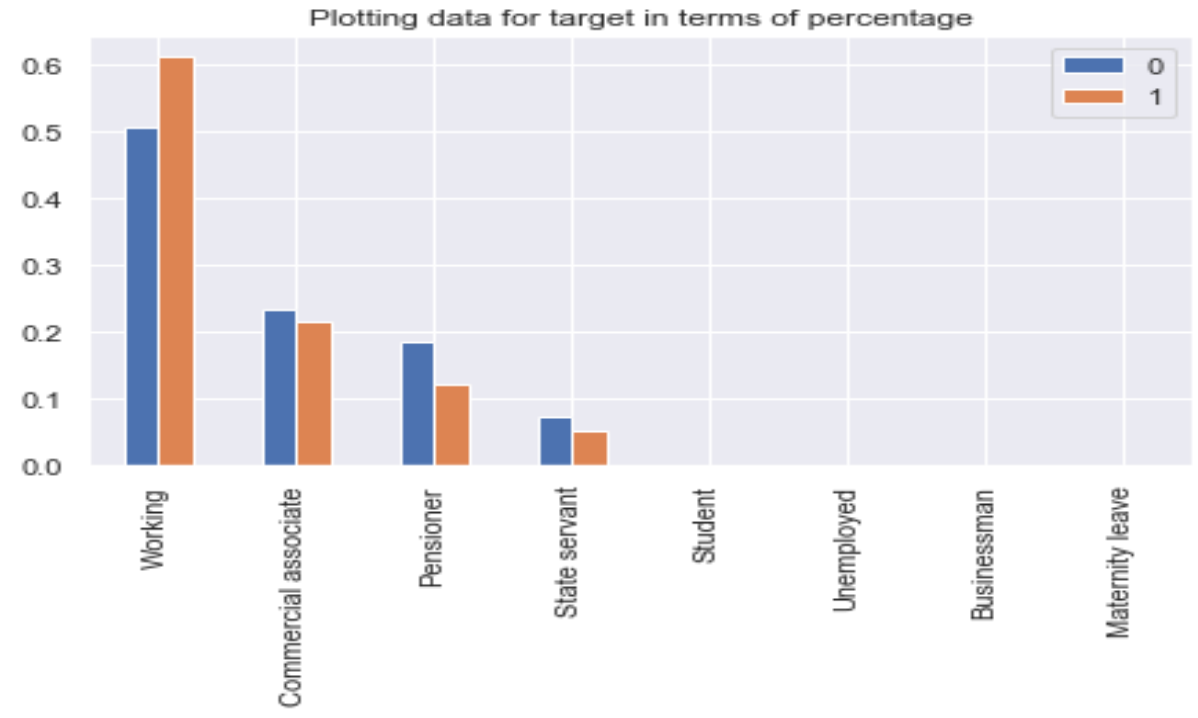


NAME_TYPE_SUITE

- Who was accompanying client when he was applying for the loan- As identified from the plots that the those who unaccompanied has higher defaulter rate.

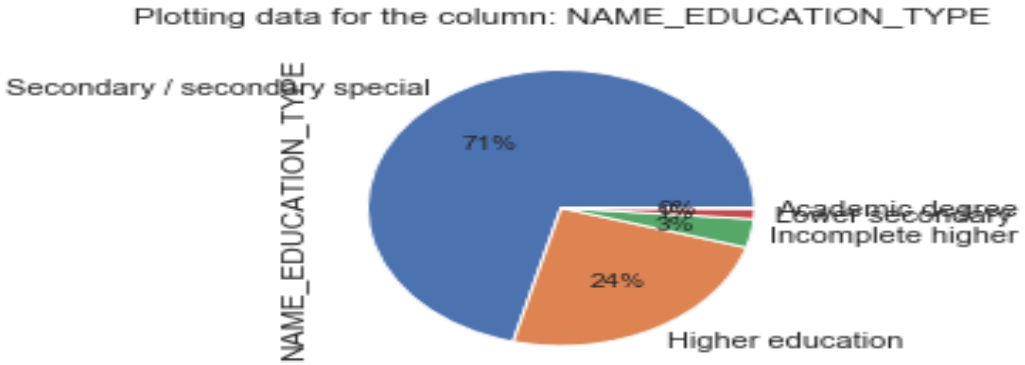


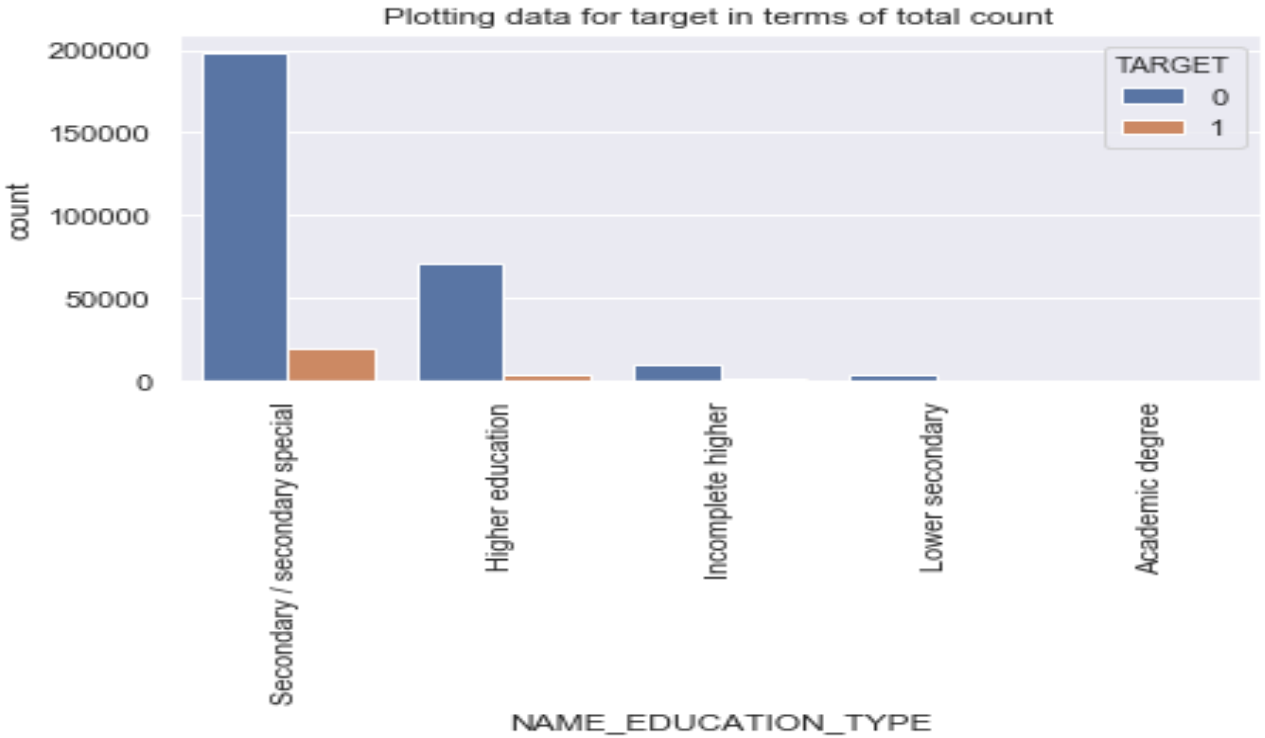


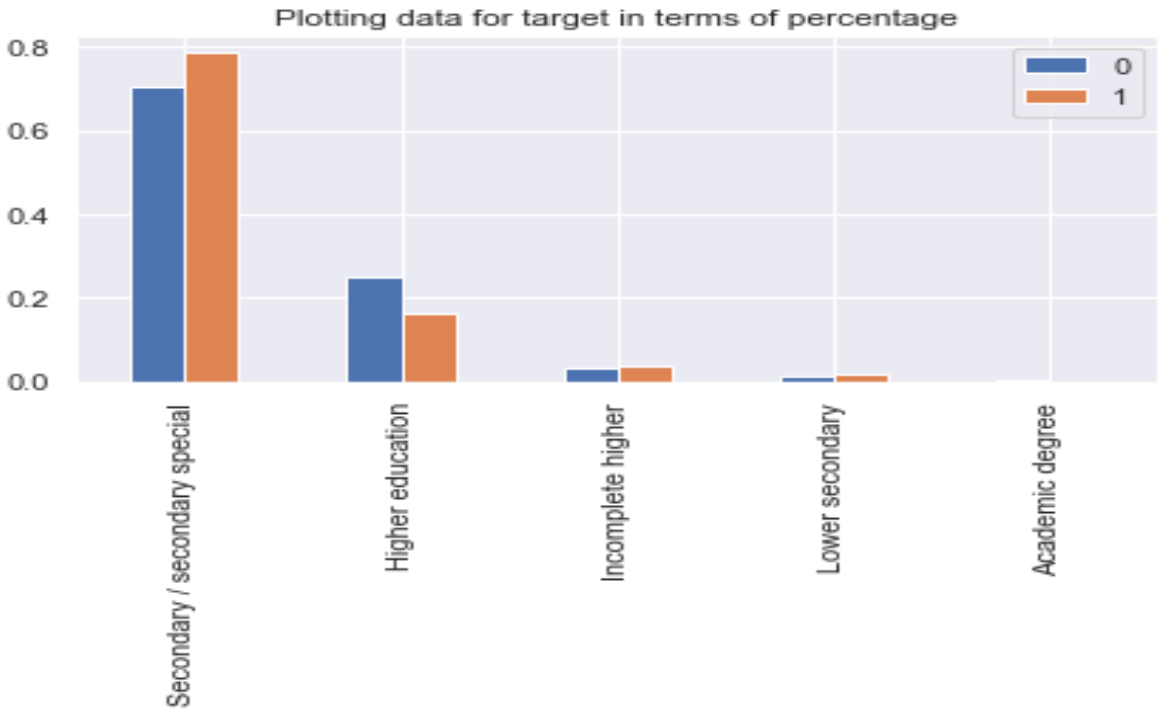


NAME_INCOME_TYPE

- Clients income type (businessman, working, maternity leave,…)- As identified from the plots that the those who are working has higher defaulter rate.
- Pensioner defaulter is lower than non-defaulter.

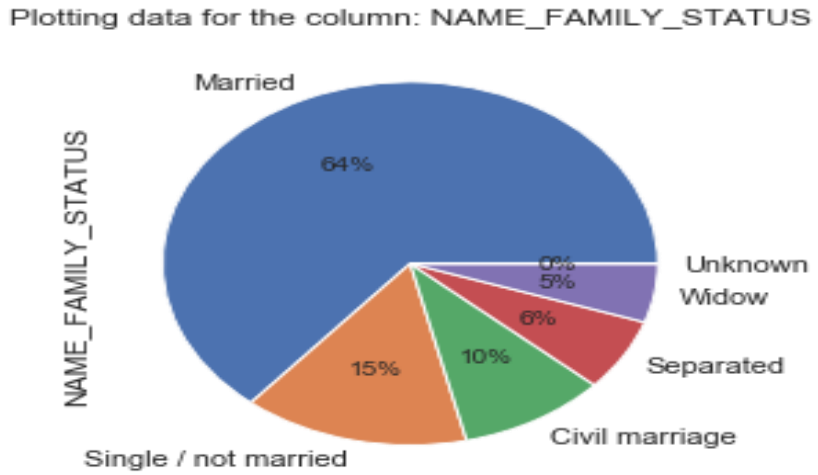


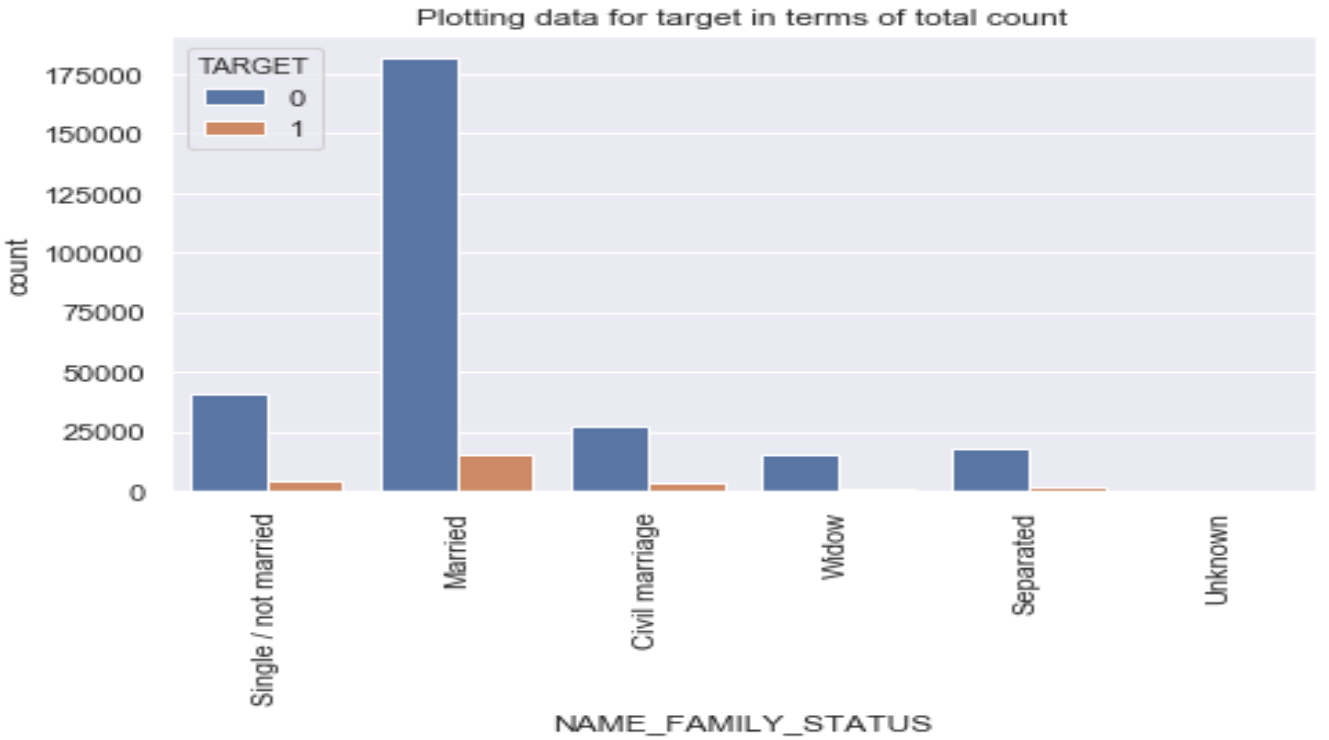


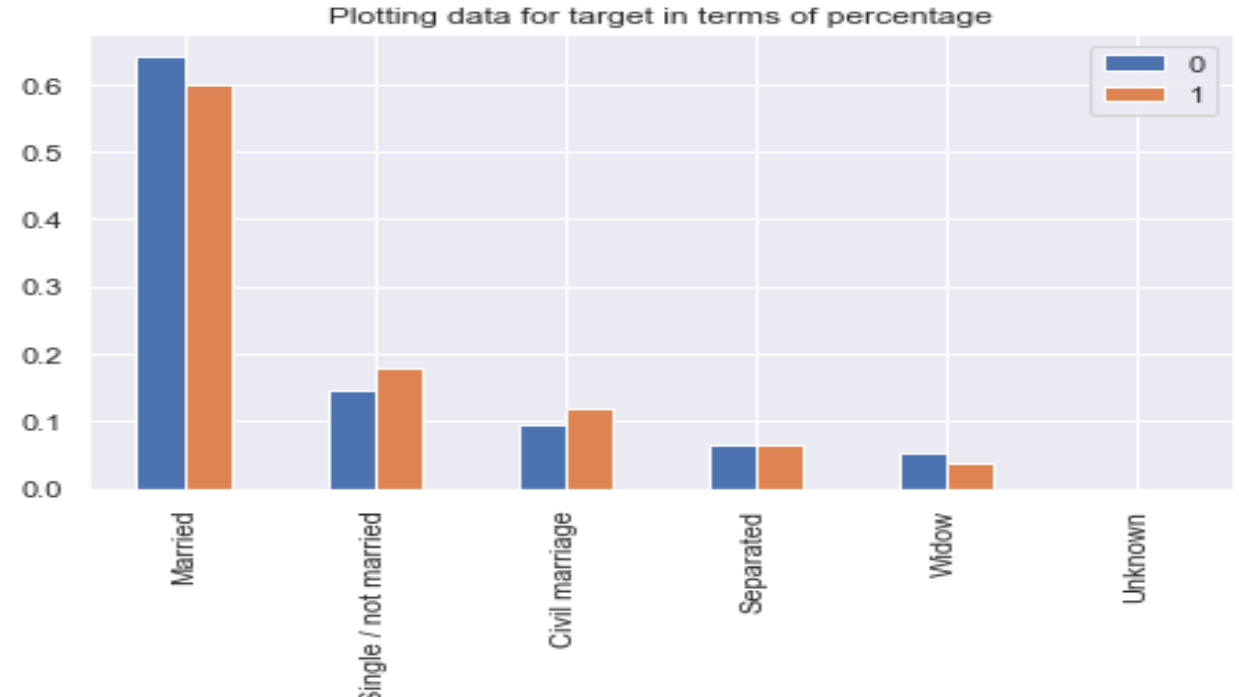


NAME_EDUCATION_TYPE

- Level of highest education the client achieved- As identified from the plots that Most client take loan for secondary education followed by higher education. But the default rate in secondary education is much high and for higher education is much low.

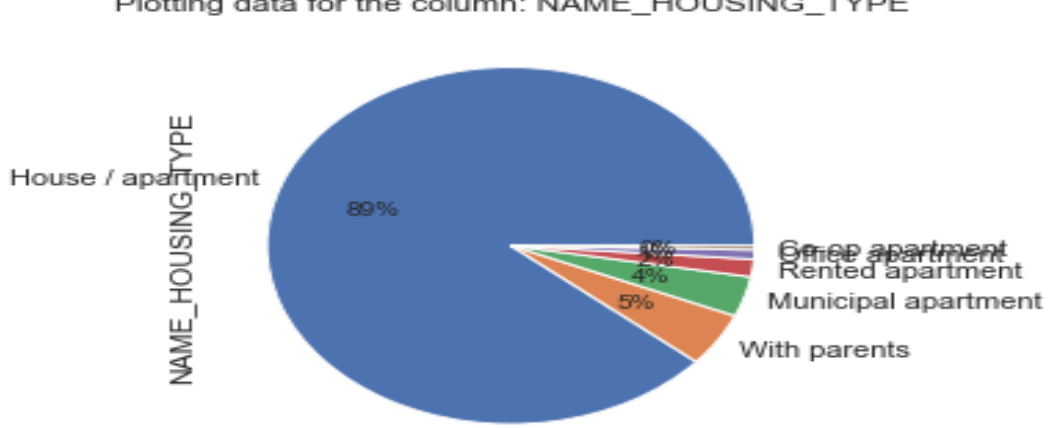


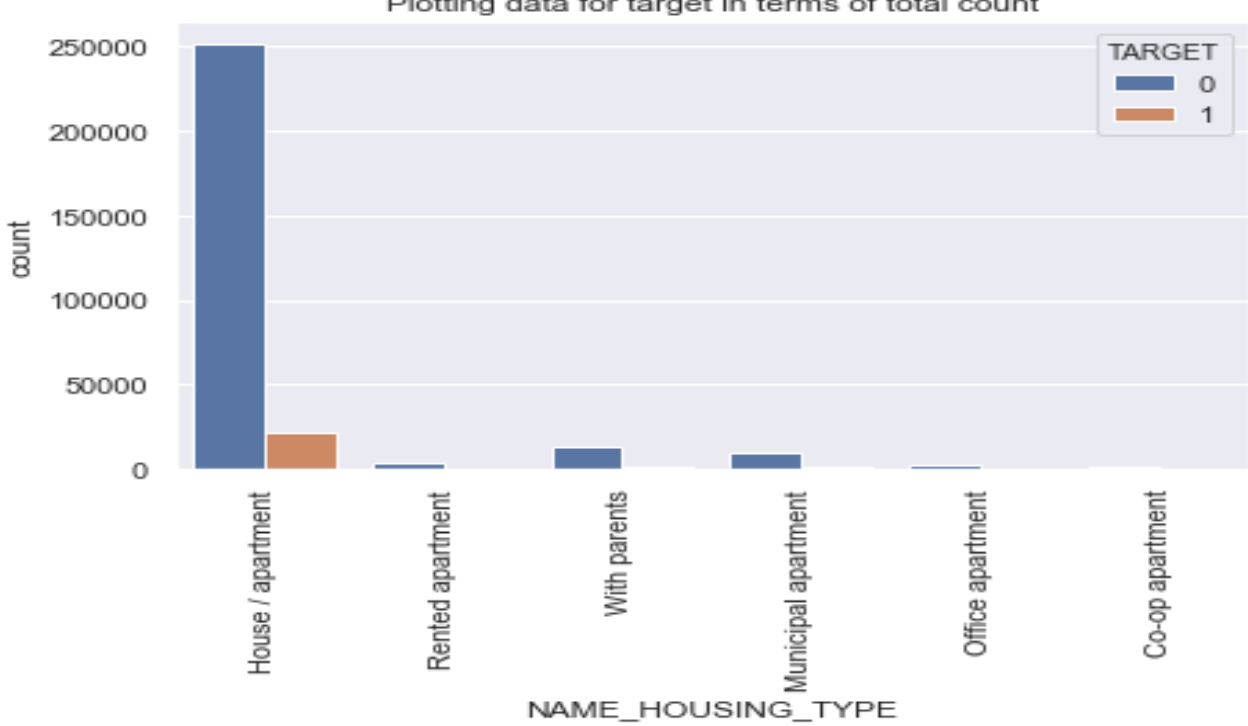


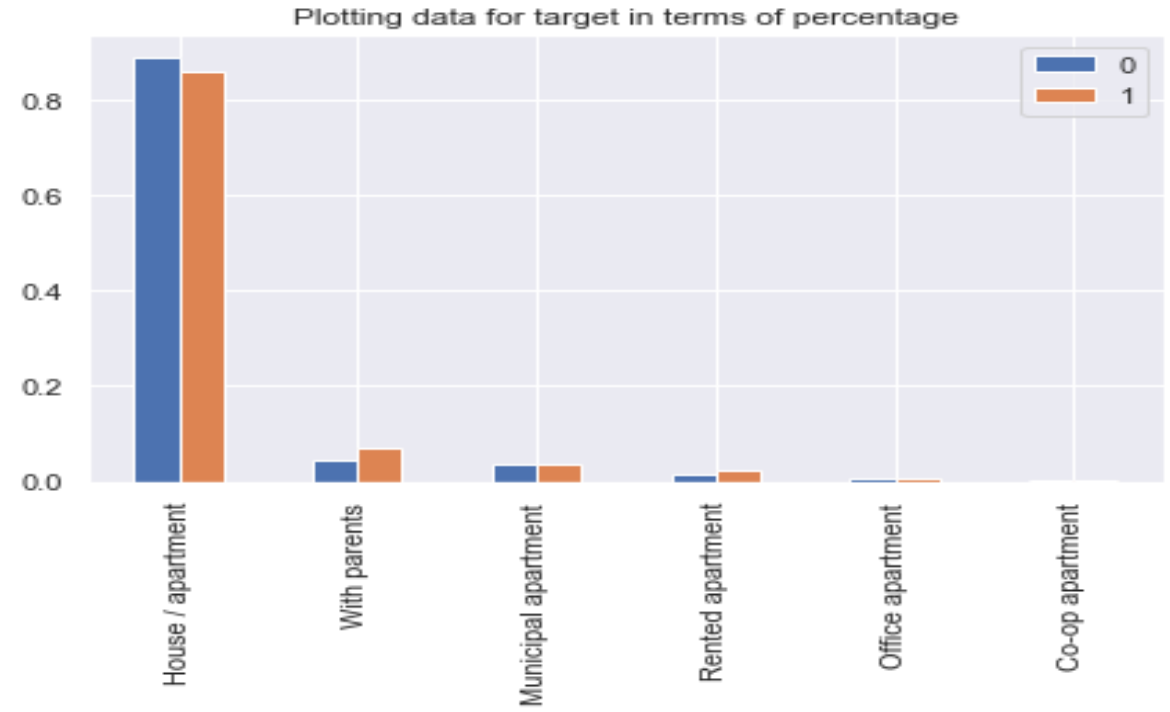


NAME_FAMILY_STATUS

- **Family status of the client**- As identified from the plots that Most married people apply for loan, and mostly they are not defaulters. Single and civil marriage turns out to be more defaulter.

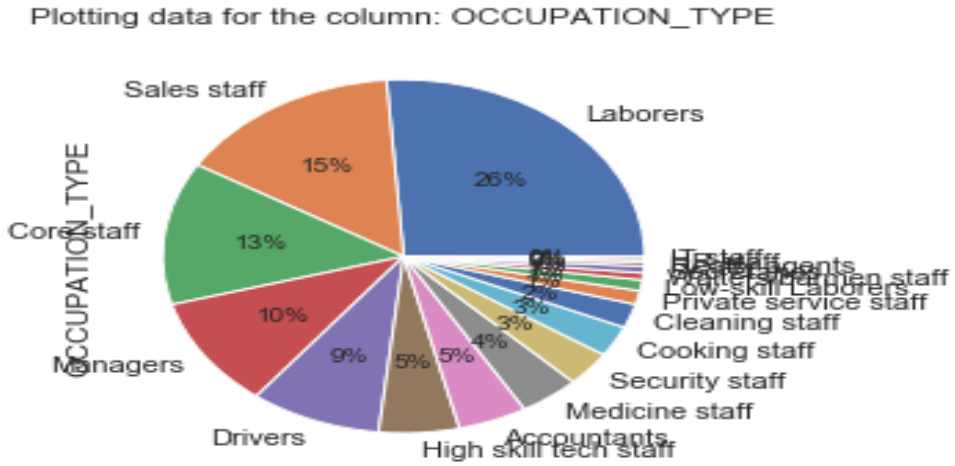


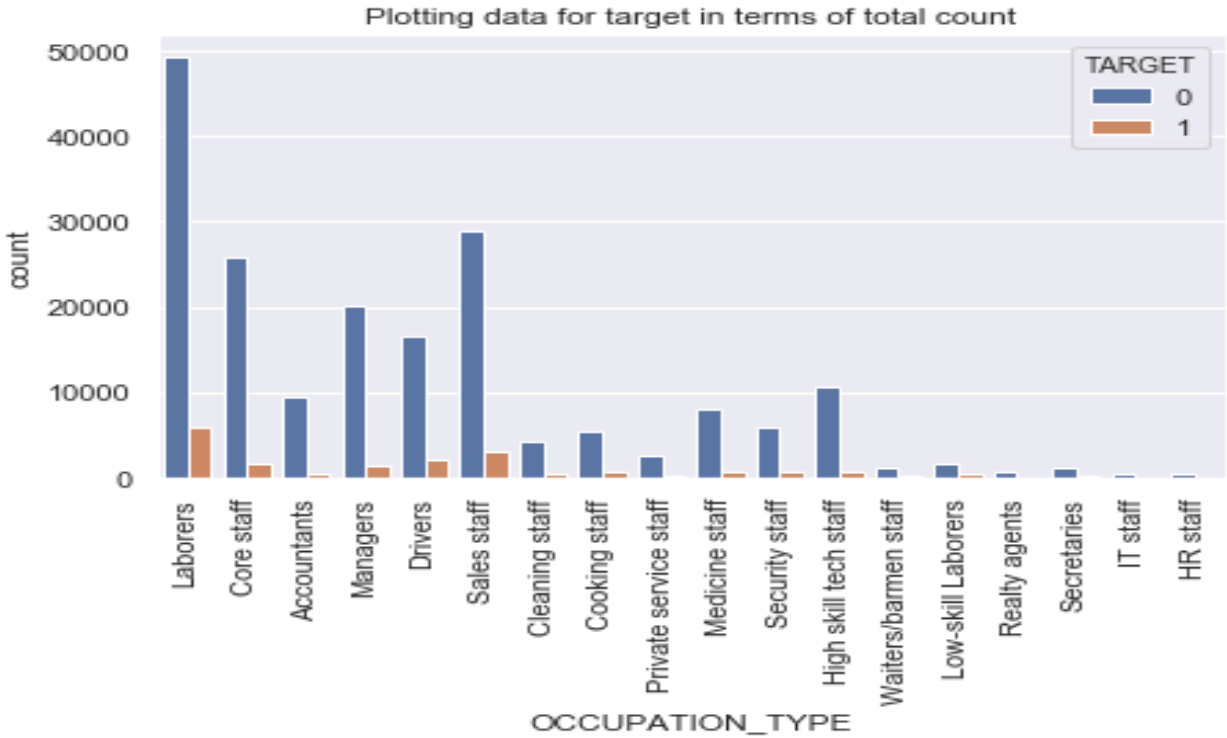


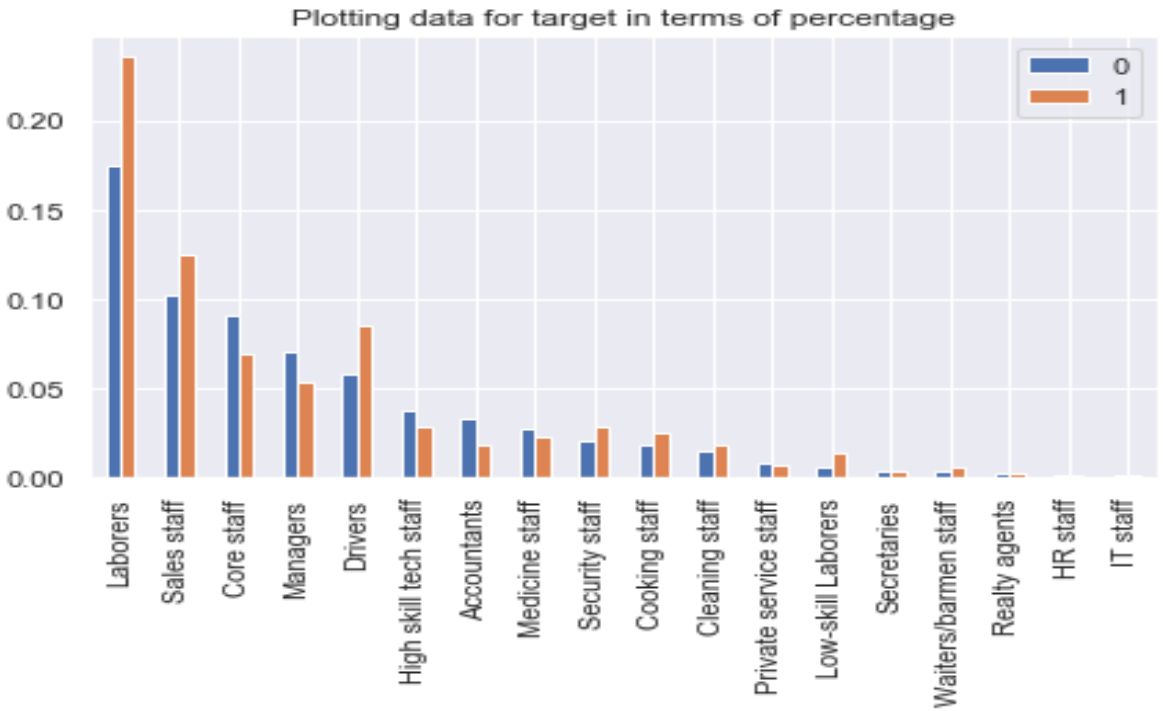


NAME_HOUSING_TYPE

- **What is the housing situation of the client (renting, living with parents, ...)**- As identified from the plots that those who live in House or Apartment has higher default rate.

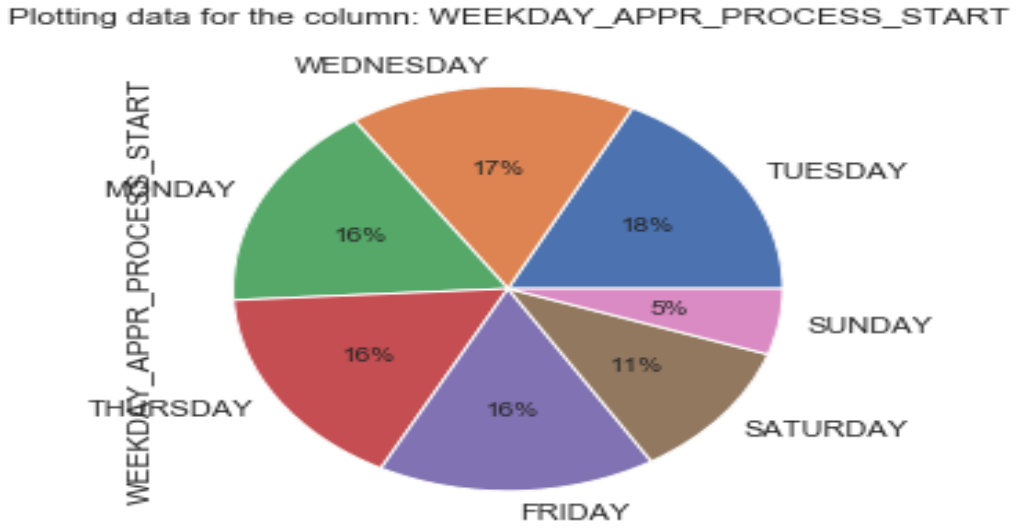


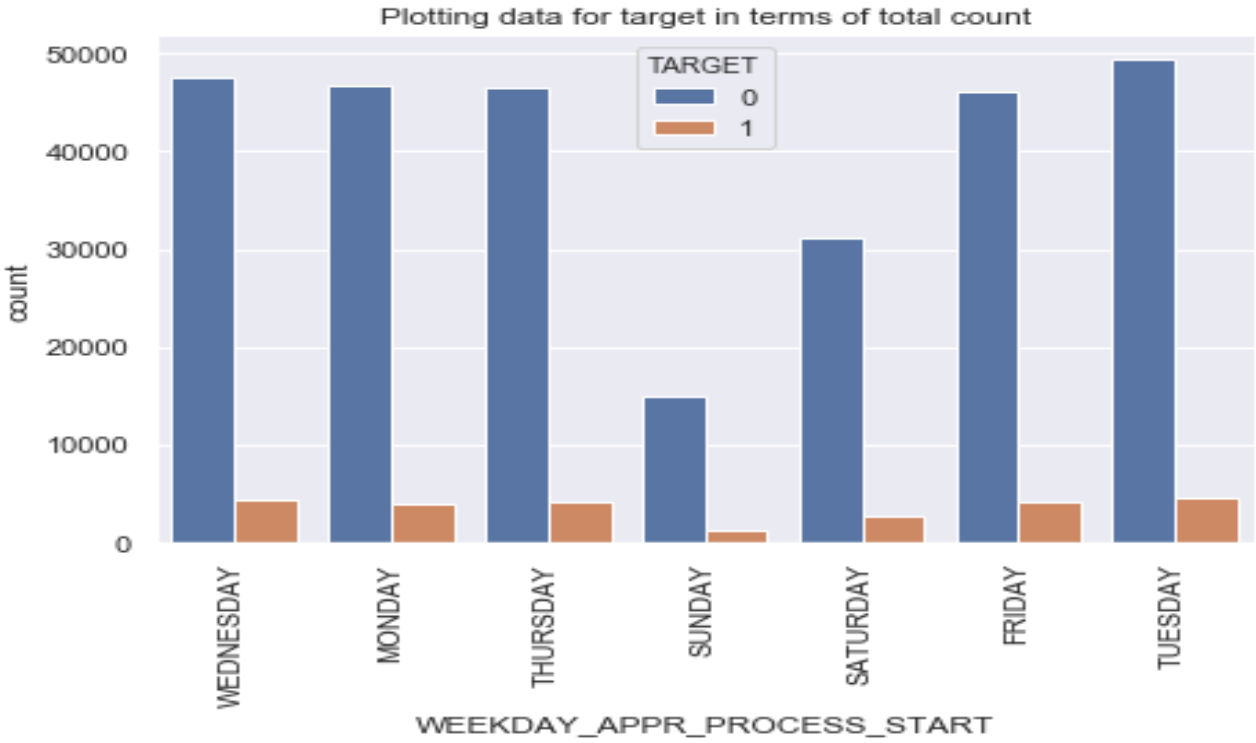


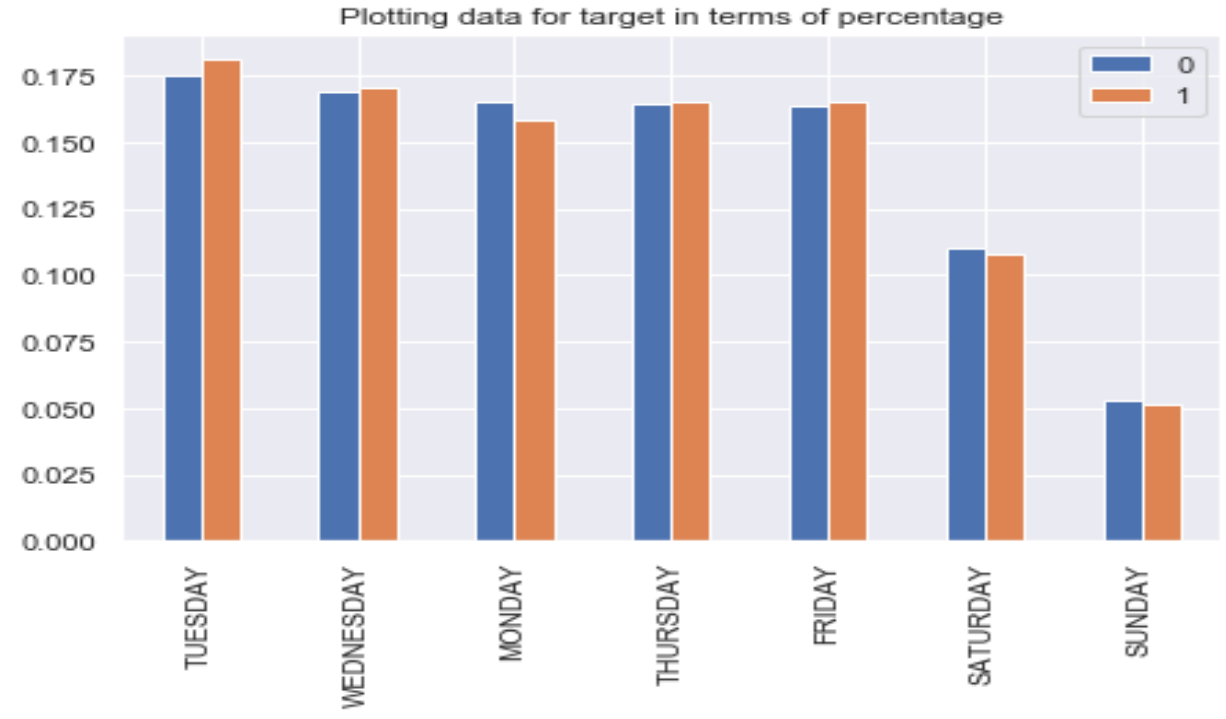


OCCUPATION_TYPE

- **What kind of occupation does the client have-** As identified from the plots that Laborers and different categories of staffs mostly take the loan, but the managers and the high skilled tech staffs are most reliable

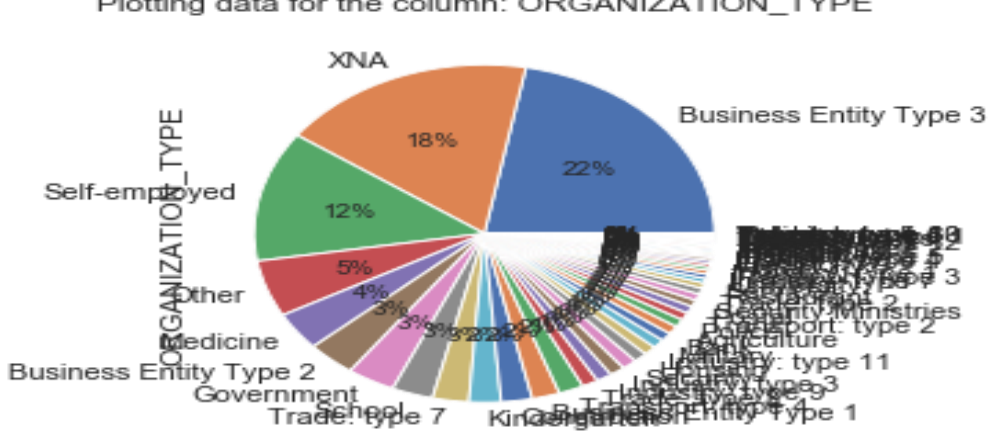


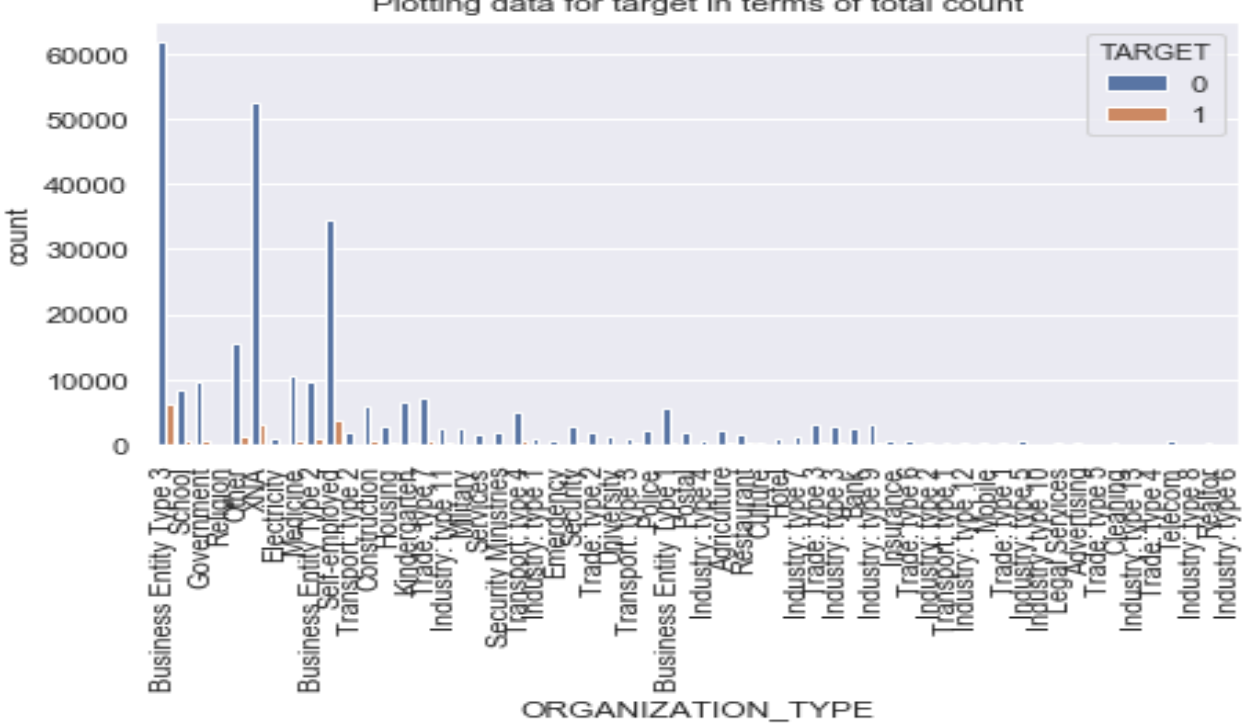


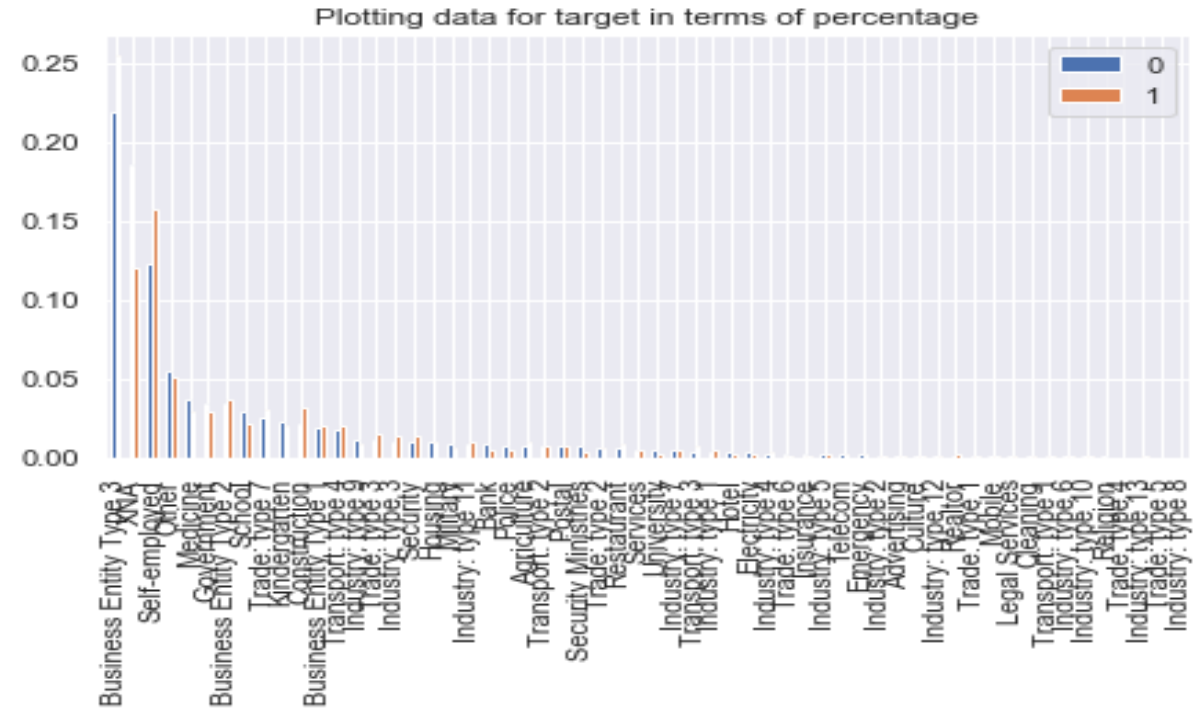


WEEKDAY_APPR_PROCESS_START

- **On which day of the week did the client apply for the loan**- As identified from the plots that Interestingly the Tuesday applied loan has highest default percentage.

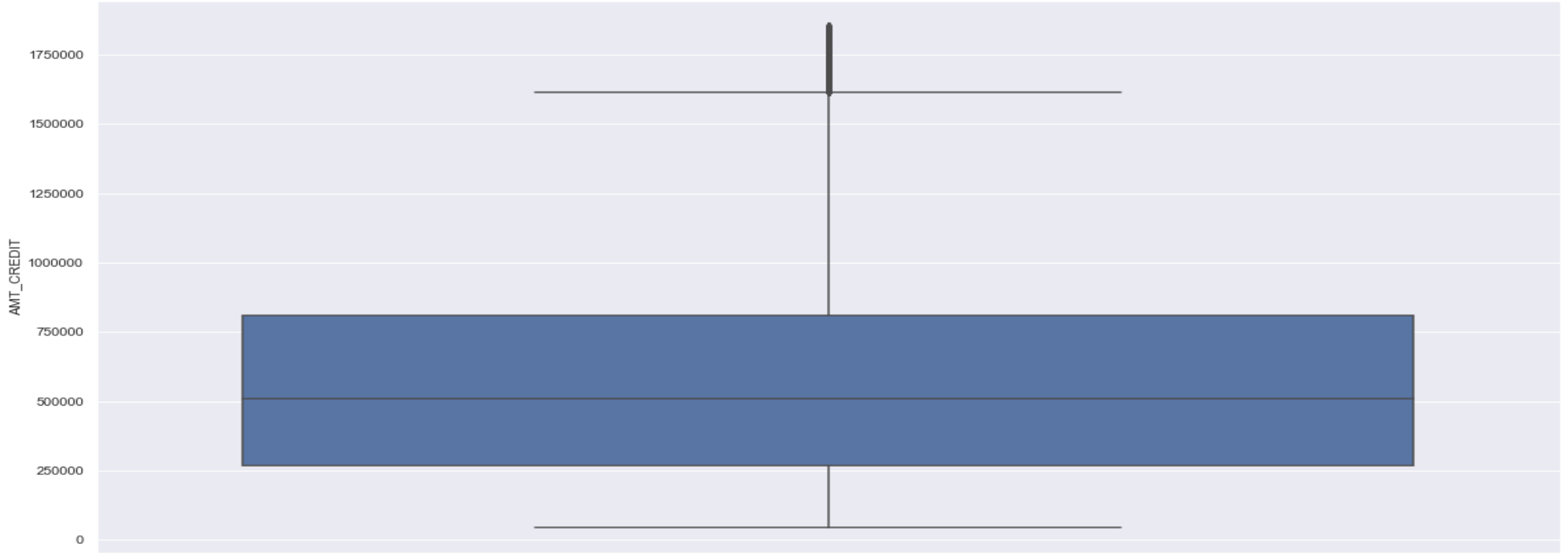






ORGANIZATION_TYPE

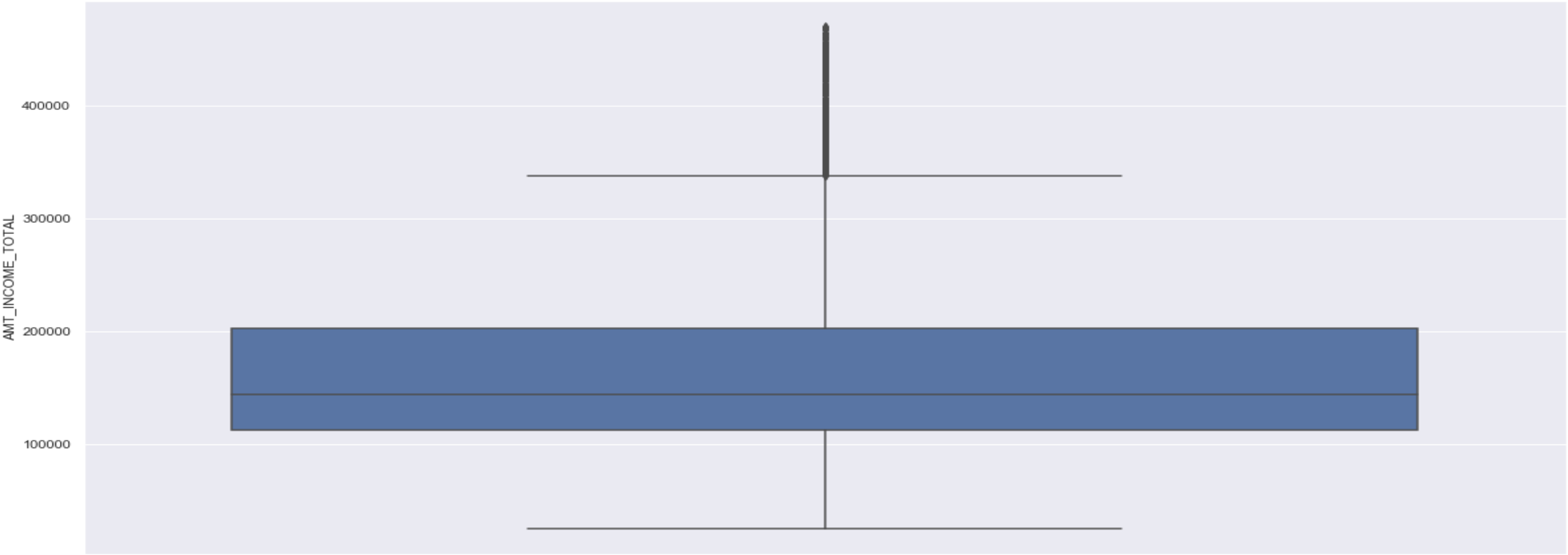
- **Type of organization where client works**- As identified from the plots that the self-employed people has highest default percentage.




```
count 3.044340e+05
mean 5.834403e+05
std 3.723969e+05
min 4.500000e+04
25% 2.700000e+05
50% 5.084955e+05
75% 8.086500e+05
max 1.852808e+06
```

AMT_CREDIT

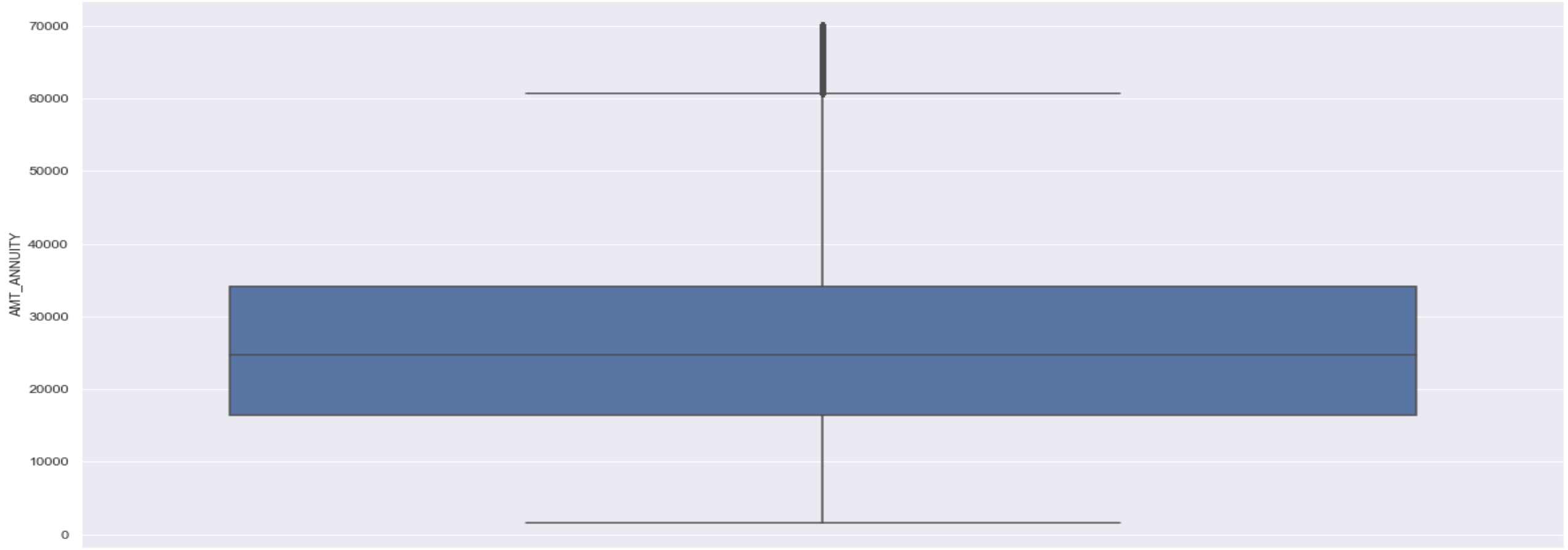
Credit amount of the loan- As identified from the plots that the Loan amount is distributed between 270000 to 800000.




```
count 304417.000000
mean 162911.014841
std 77494.004409
min 25650.000000
25% 112500.000000
50% 144000.000000
75% 202500.000000
max 469800.000000
```

AMT_INCOME_TOTAL

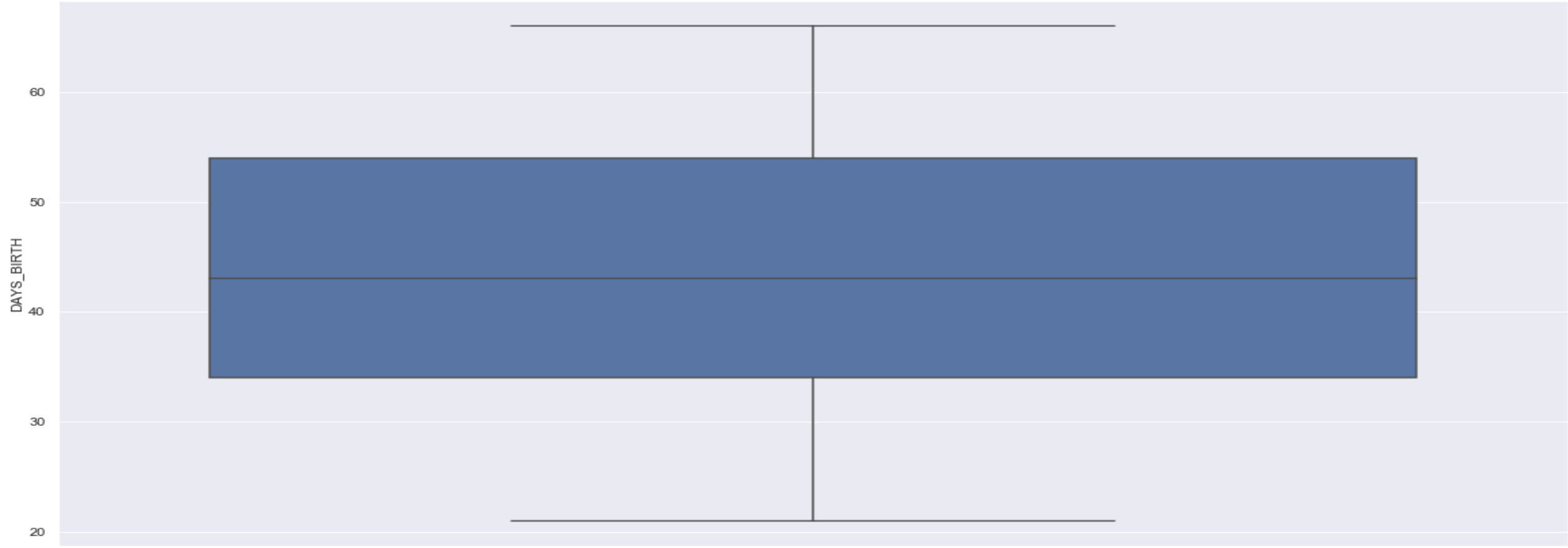
Income of the client- As identified from the plots that the Income amount is distributed between 112500 to 202500.




```
count
304418.000000
mean 26498.619144
std 13032.387753
min 1615.500000
25% 16456.500000
50% 24745.500000
75% 34182.000000
max 69988.500000
```

AMT_ANNUITY

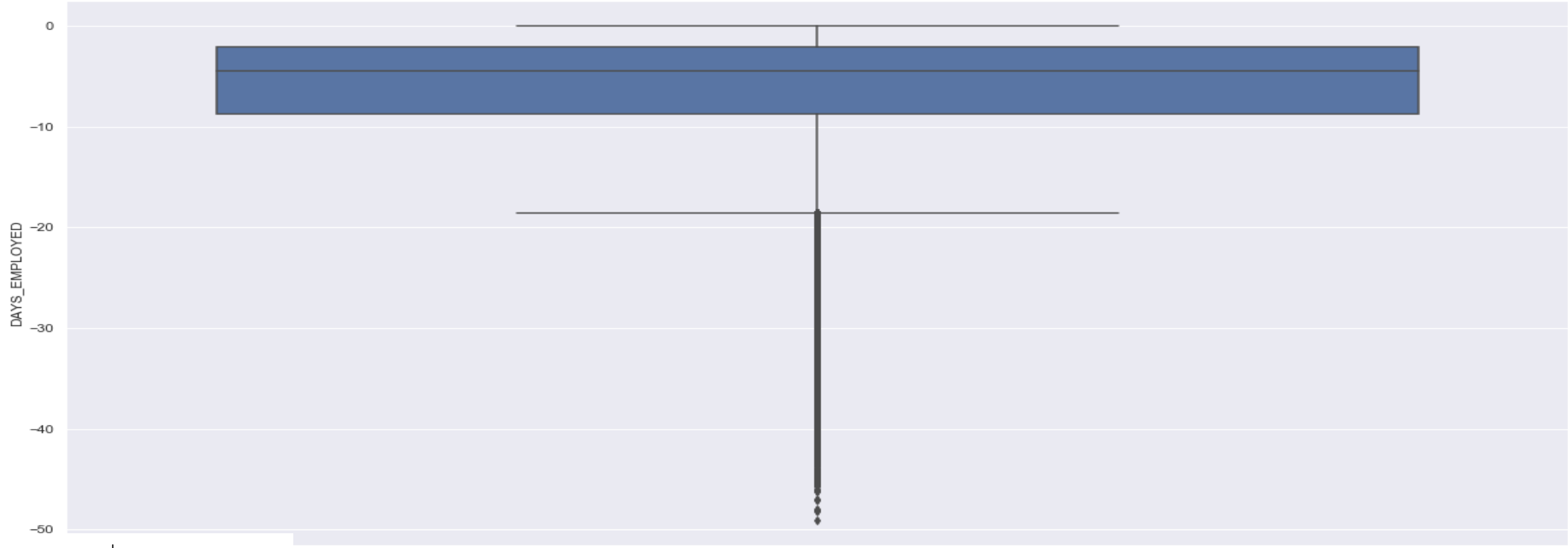
Loan annuity- As identified from the plots that the Annuity amount is distributed between 16000 to 35000.




```
count
303709.000000
mean 43.642437
std 11.740048
min 21.000000
25% 34.000000
50% 43.000000
75% 54.000000
max 66.000000
```

DAYS_BIRTH

Client's age in days at the time of application- As identified from the plots that the Applicants age is distributed b/w 33 to 55 yrs




```
count
252137.000000
mean -6.531971
std 6.406466
min -49.073973
25% -8.698630
50% -4.515068
75% -2.101370
max 0.000000
```

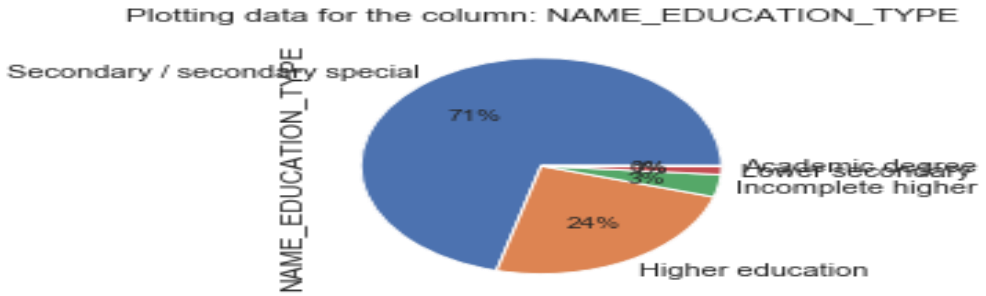
DAYS_EMPLOYED

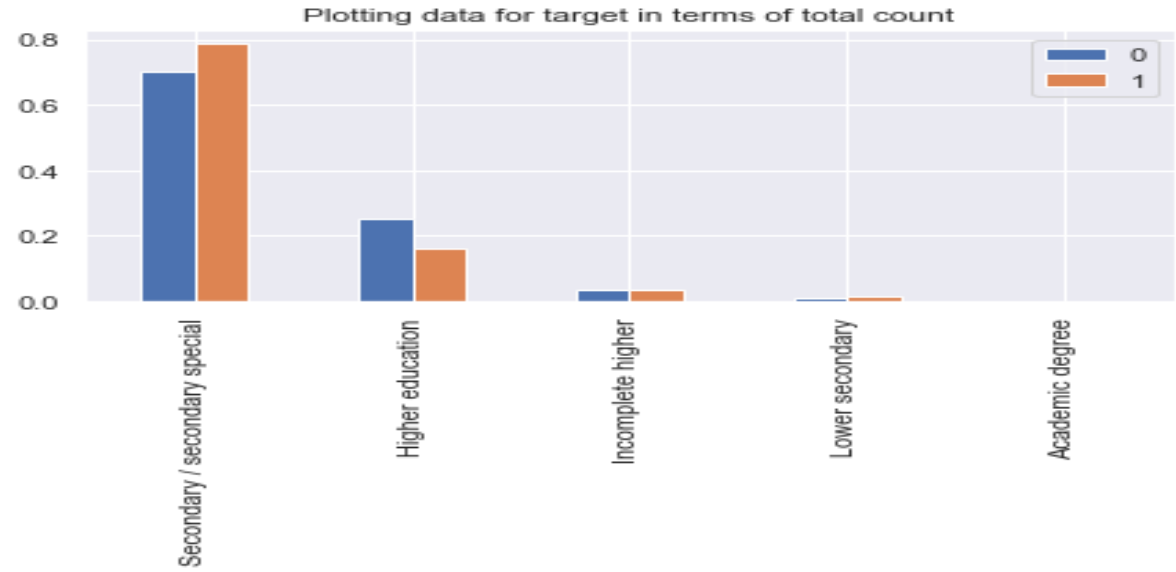
How many days before the application the person started current employment- As identified from the plots that Applicants are at least have 2 to 8yrs work experience before applying loan.

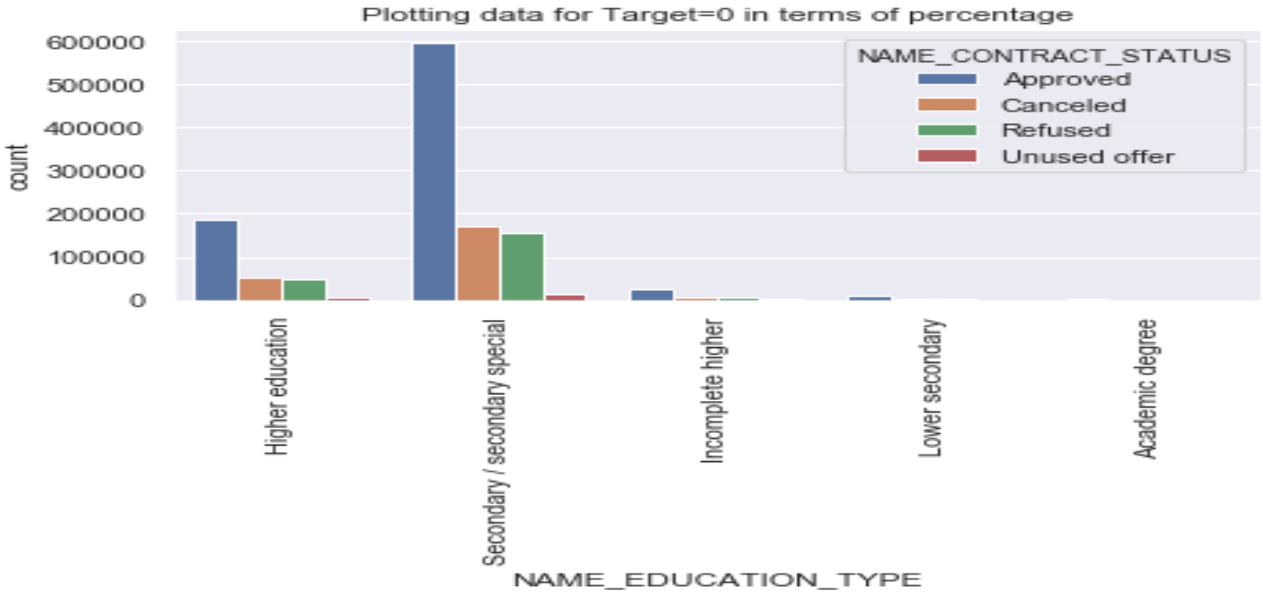
Analysis Using the Previous Application Data

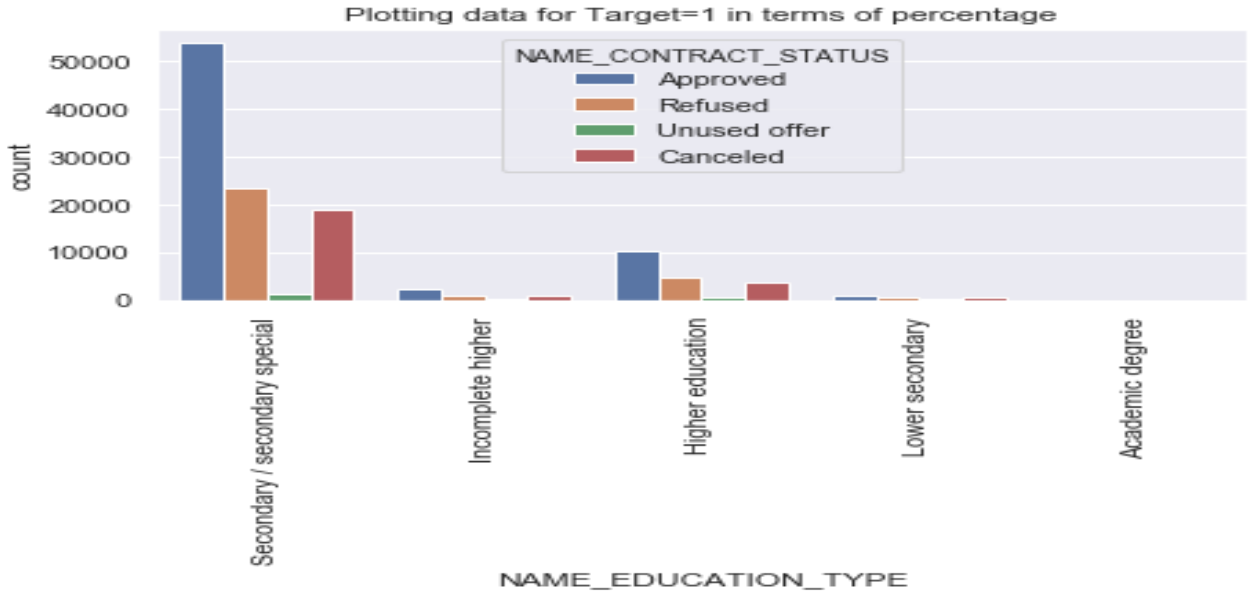
Column Reference- NAME_EDUCATION_TYPE & NAME_CONTRACT_STATUS

- **People tend to make more loan for 'Secondary special' and their loan is also approved.**



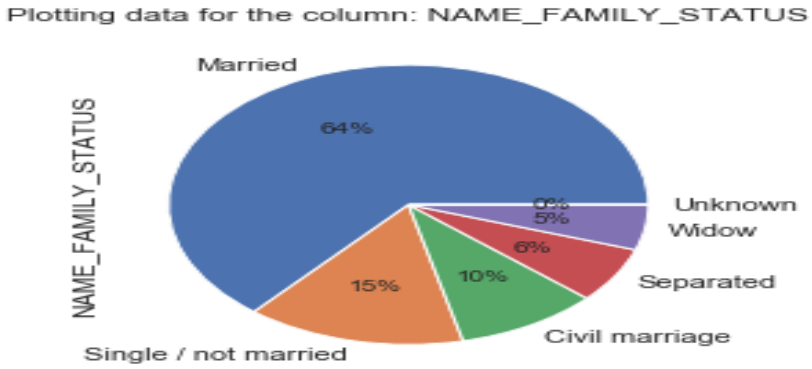


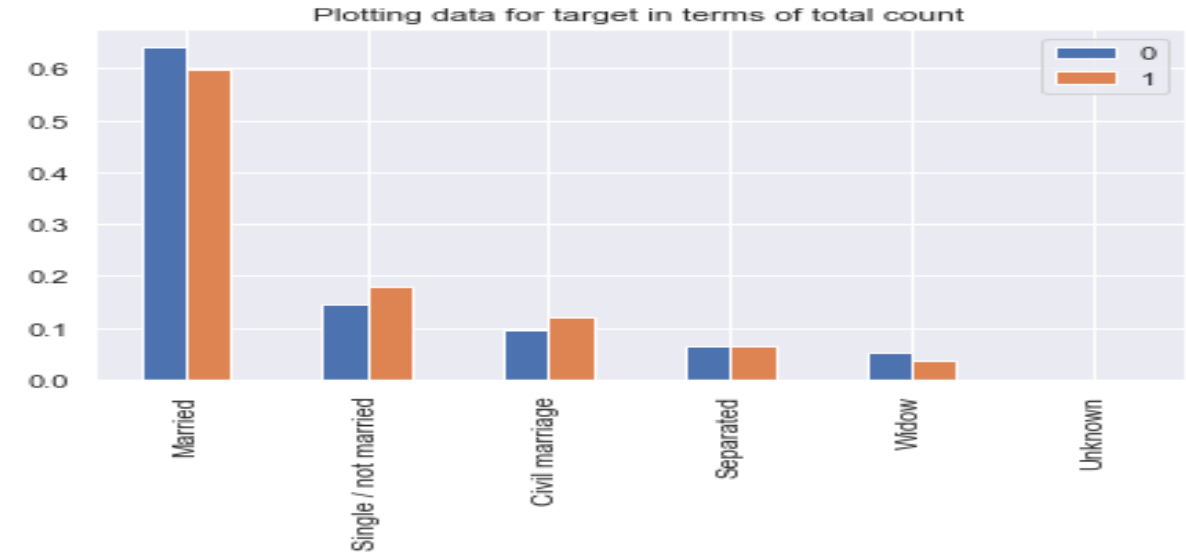


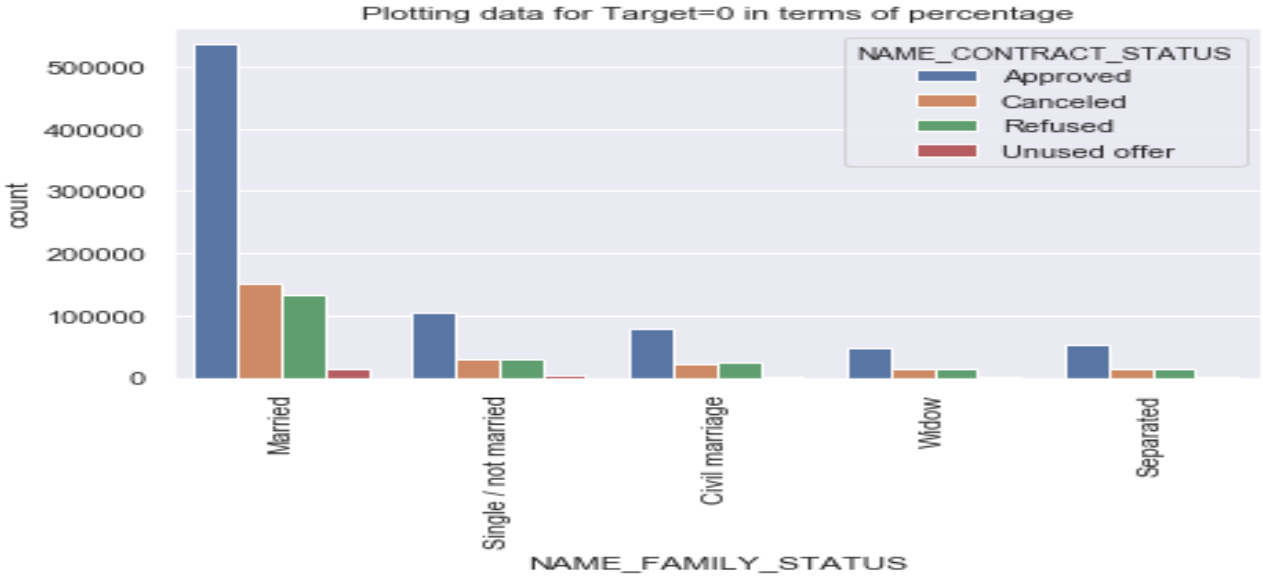


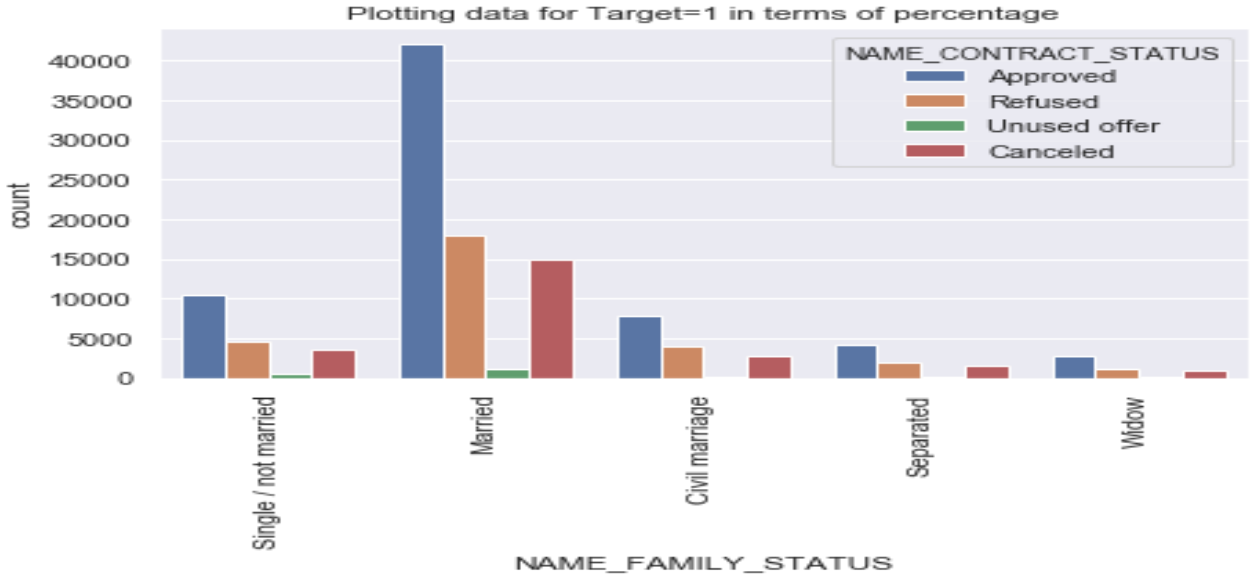
Column Reference- Name_FAMILY_STATUS & NAME_CONTRACT_STATUS

There is a clear difference for the categories for "Approved, Refused, Unused and Cancelled" for the category: Married. Married people tends to pay loan on time than Singles.



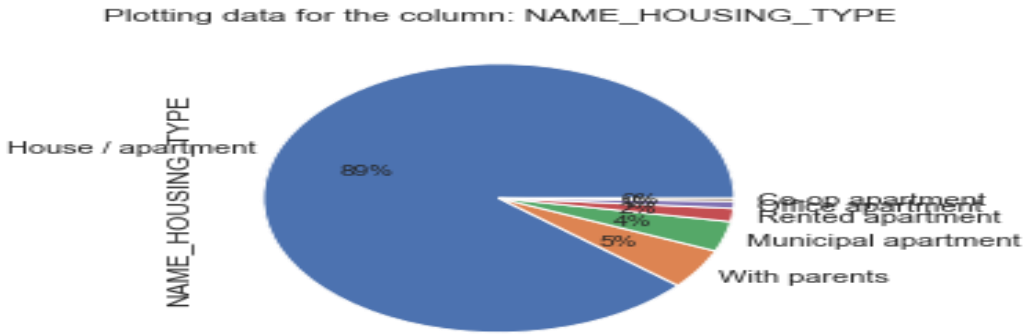


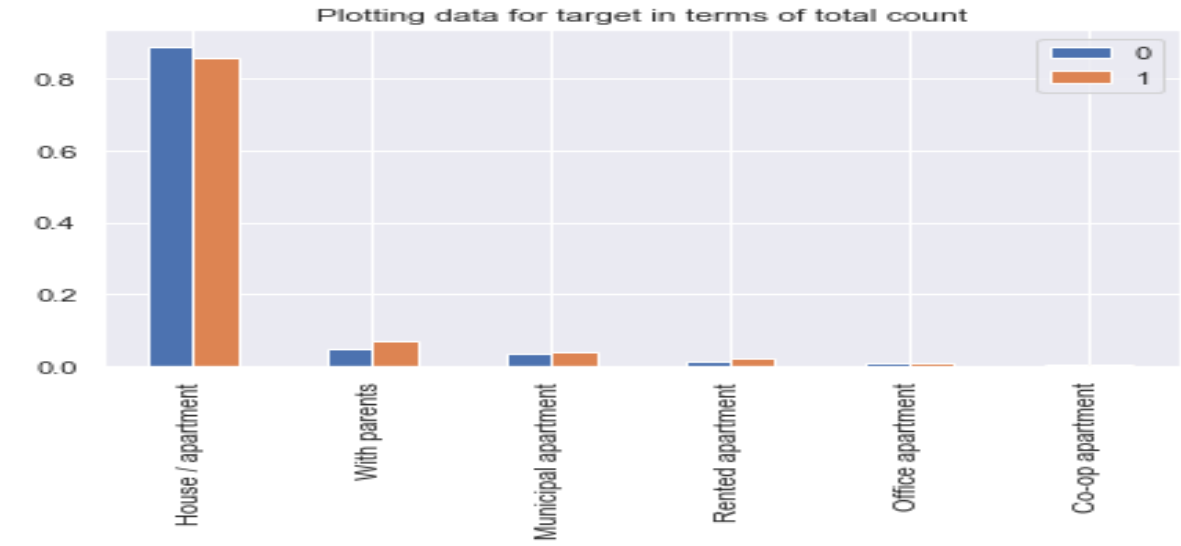


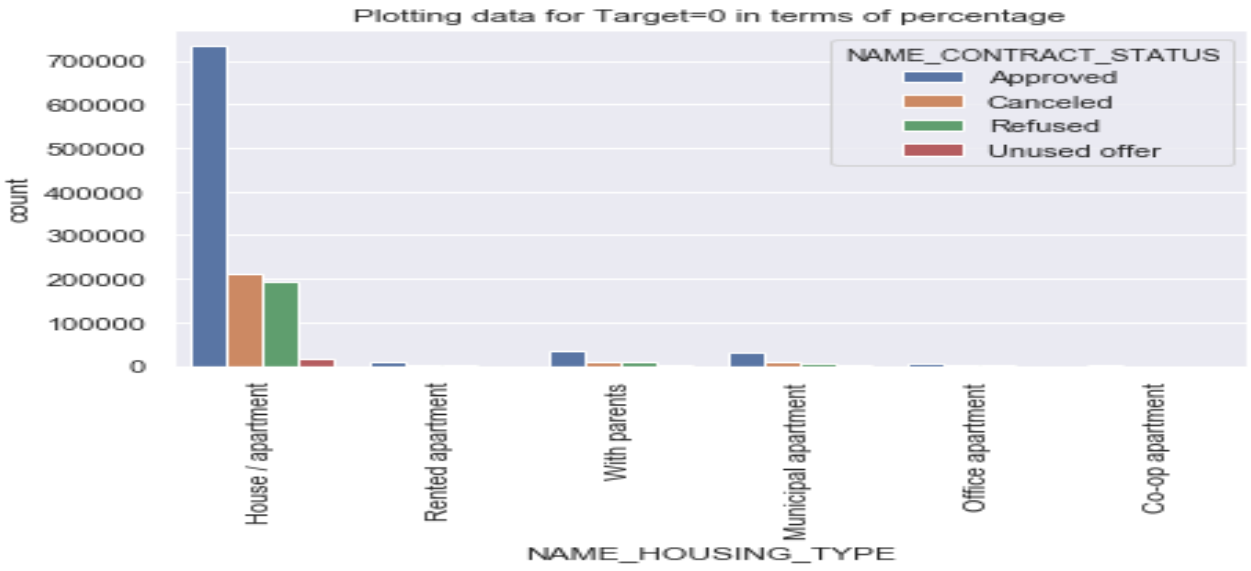


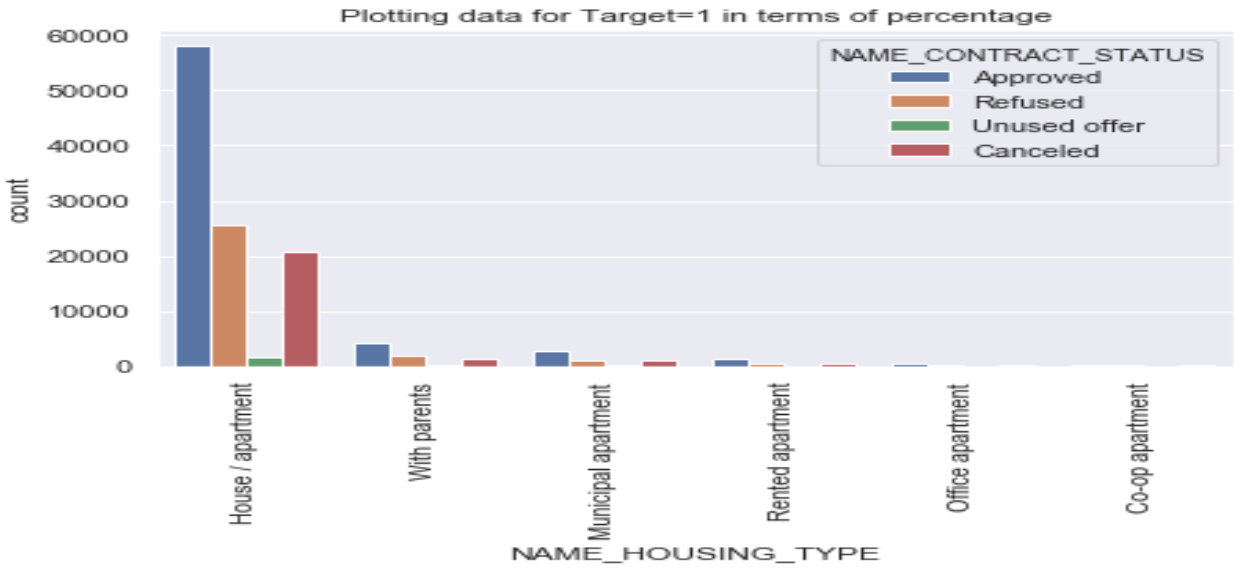
Column Reference- Name_HOUSING_TYPE & NAME_CONTRACT_STATUS

There is a clear difference for the categories for "Approved, Refused, Unused and Cancelled" for the category: House/apartment.



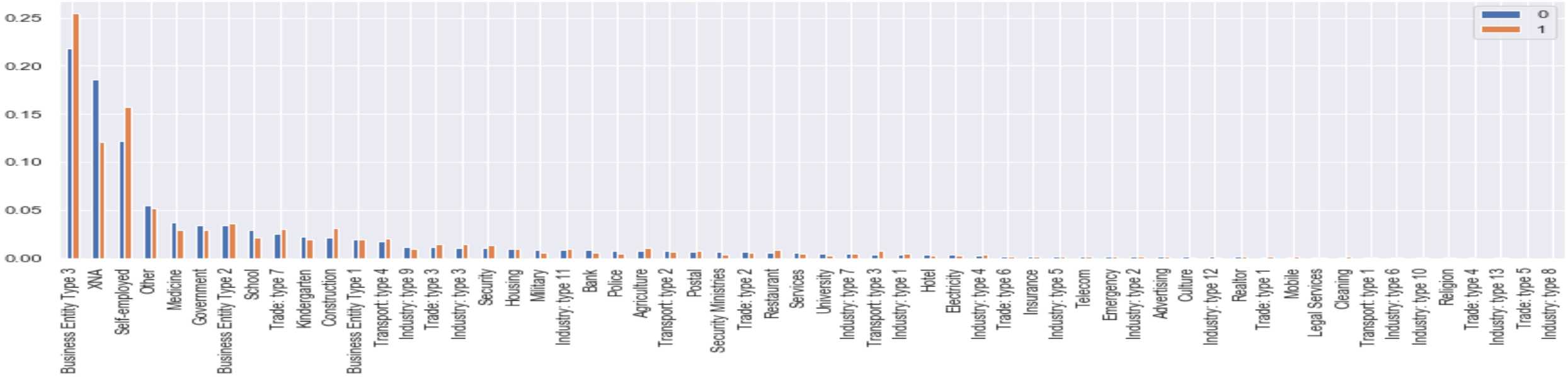


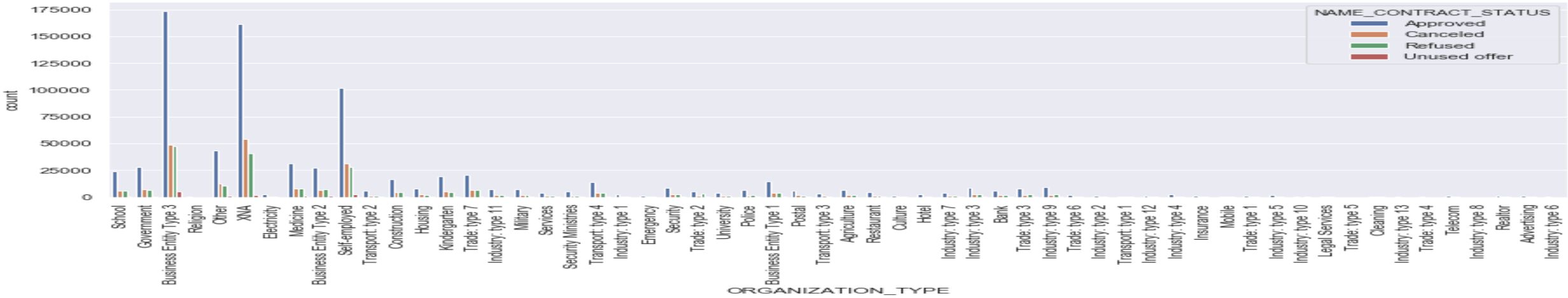




Column Reference- ORGANIZATION_TYPE & NAME_CONTRACT_STATUS

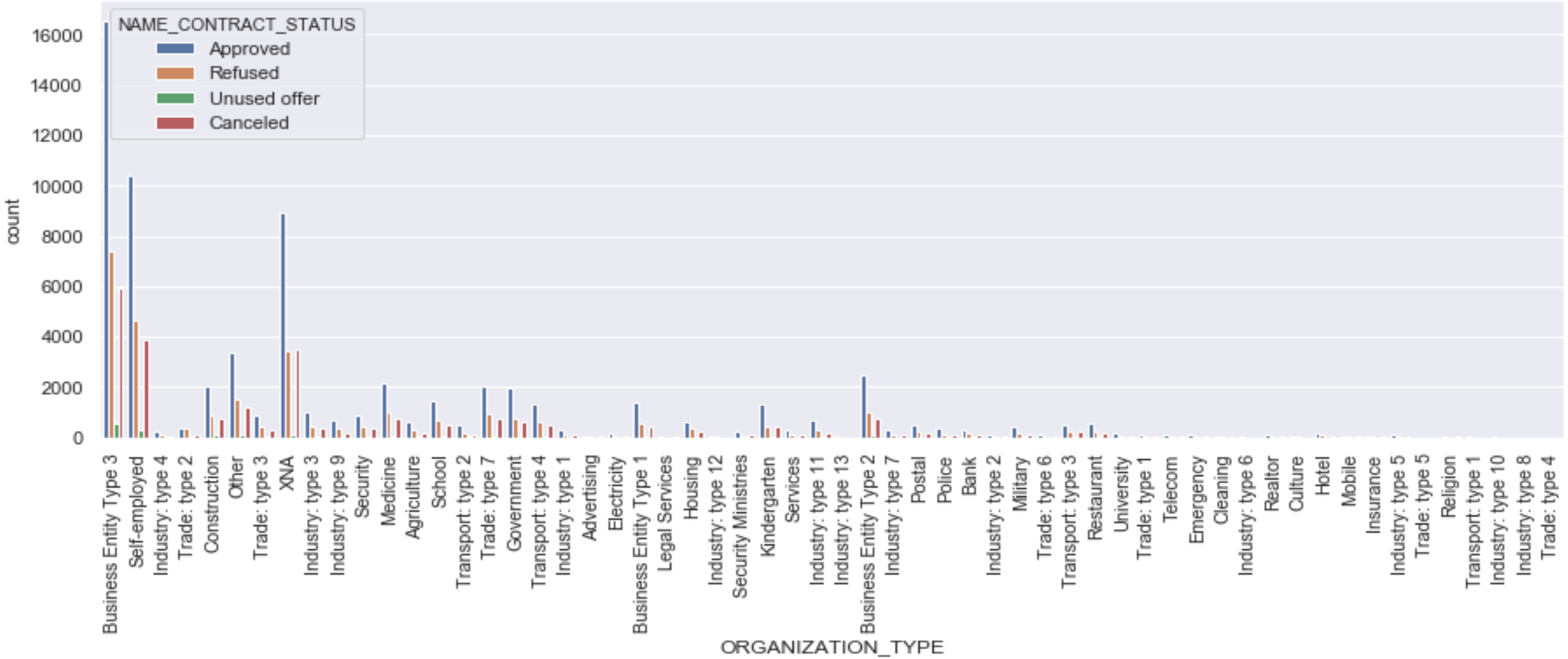
- **Its clear that Business Entry Type-3 has maximum default rate**





- **Business Entry Type 3 has maximum Approve Rate from non Defaulter list**

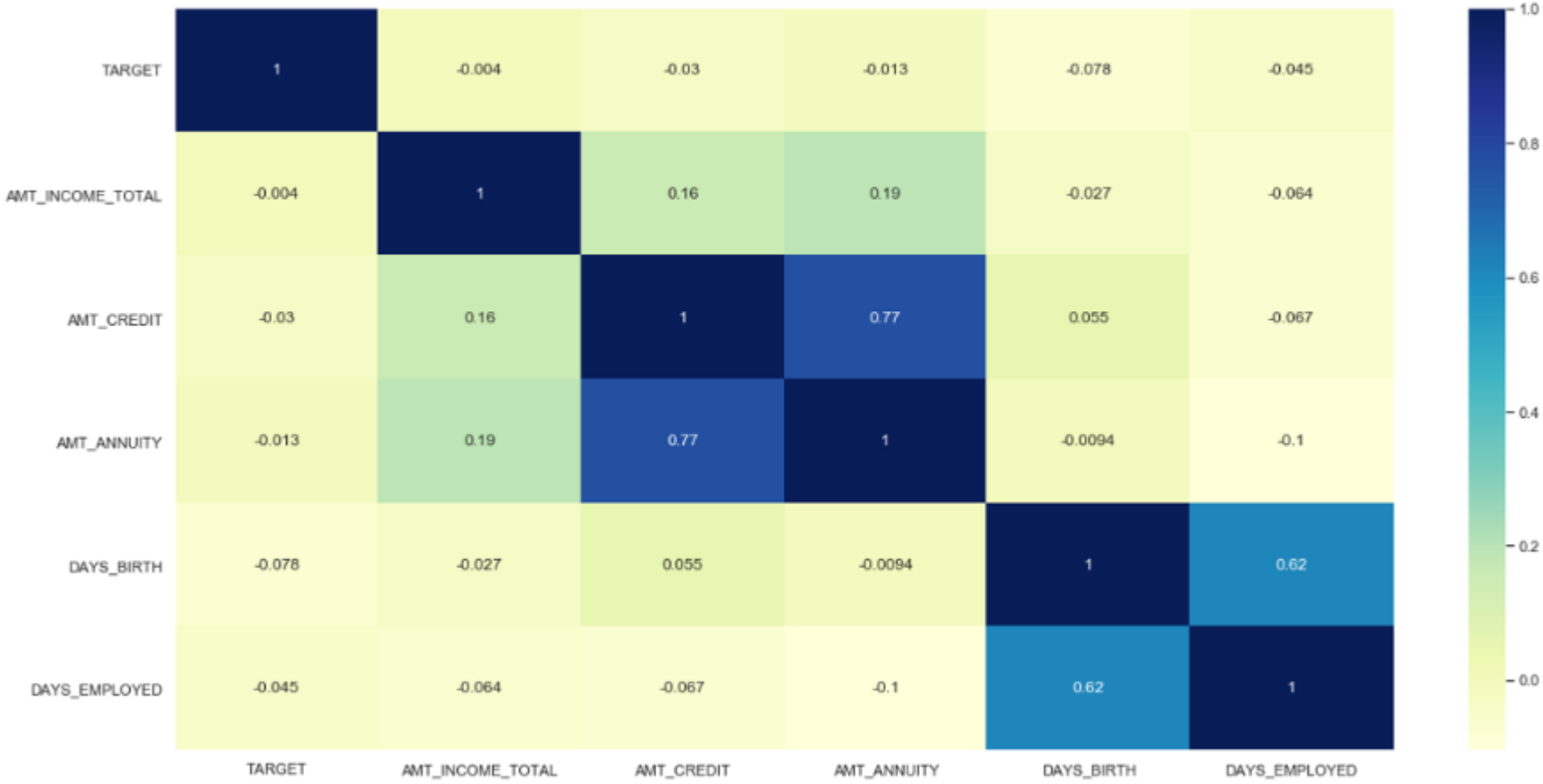
Column Reference- ORGANIZATION_TYPE & NAME_CONTRACT_STATUS



- **Business Entry Type 3 has maximum Approve Rate from Defaulter list**

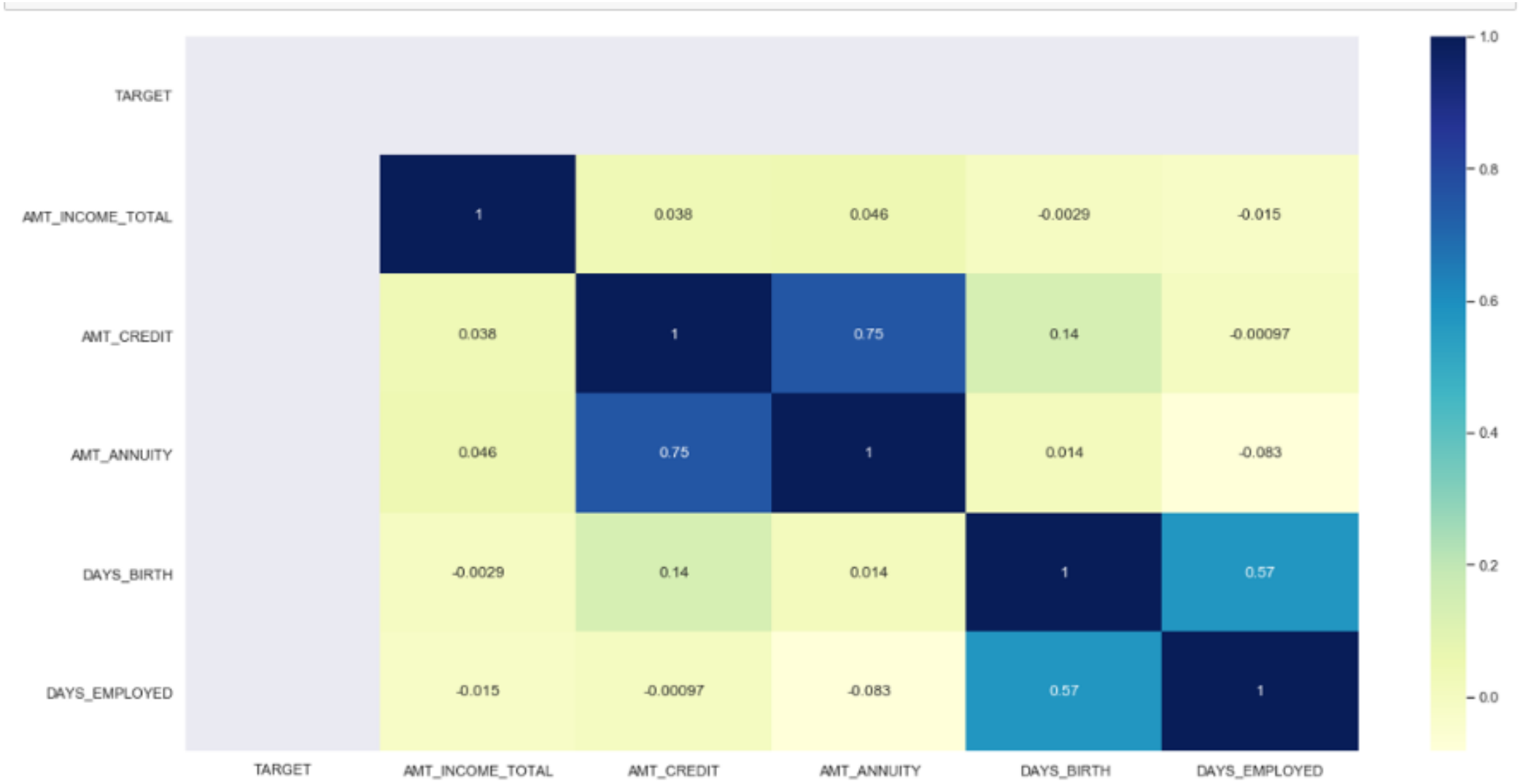
Correlation :

- Correlation for entire set :



- Correlation for data set Target = 0

- Correlation for data set for target =1



For Target 1 correlation matrix

	VAR1	VAR2	Correlation
893	FLAG_EMP_PHONE	DAYS_EMPLOYED	0.999758
2689	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508
2411	FLOORSMAX_MEDI	FLOORSMAX_AVG	0.997018
2342	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.993582
2413	FLOORSMAX_MEDI	FLOORSMAX_MODE	0.988153
412	AMT_GOODS_PRICE	AMT_CREDIT	0.987250
2275	FLOORSMAX_MODE	FLOORSMAX_AVG	0.985603
2206	YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.971032
2344	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.962064
1379	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149

For Target 0 correlation matrix

	VAR1	VAR2	Correlation1
893	FLAG_EMP_PHONE	DAYS_EMPLOYED	0.999702
2689	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998269
2411	FLOORSMAX_MEDI	FLOORSMAX_AVG	0.997187
2342	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.996124
2413	FLOORSMAX_MEDI	FLOORSMAX_MODE	0.989195
2275	FLOORSMAX_MODE	FLOORSMAX_AVG	0.986594
412	AMT_GOODS_PRICE	AMT_CREDIT	0.983103
2206	YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.980466
2344	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.978073
1379	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637

Observations :

1. FLAG_EMP_PHONE & DAYS_EMPLOYED 0.999758,
2. OBS_60_CNT_SOCIAL_CIRCLE & OBS_30_CNT_SOCIAL_CIRCLE 0.998508
3. FLOORSMAX_MEDI & FLOORSMAX_AVG 0.997018
4. YEARS_BEGINEXPLUATATION_MEDI & YEARS_BEGINEXPLUATATION_AVG 0.993582
5. FLOORSMAX_MEDI & FLOORSMAX_MODE 0.988153
6. AMT_GOODS_PRICE & AMT_CREDIT 0.987250
7. FLOORSMAX_MODE & FLOORSMAX_AVG 0.985603
8. YEARS_BEGINEXPLUATATION_MODE & YEARS_BEGINEXPLUATATION_AVG 0.971032
9. YEARS_BEGINEXPLUATATION_MEDI &YEARS_BEGINEXPLUATATION_MODE 0.962064
10. REGION_RATING_CLIENT_W_CITY & REGION_RATING_CLIENT 0.950149

It can be inferred from the top 10 co-related columns that that Target 0 & Target 1 dataset following the same pattern.

Thank You!!!