Risk analysis of loan applicant- case study

Business understanding --

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history.

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:

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

2. If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Problem statement --

This case study aims to 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 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.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Understanding the data --

- *'application_data.csv'* contains all the information of the client at the time of application.
  The data is about whether a **client has payment difficulties.**
- *'previous_application.csv'* contains information about the client’s previous loan data. It contains the data on whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**
- *'columns_description.csv'* is data dictionary which describes the meaning of the variables.

Analysis approach --

1. Data cleaning –
   1. Handling missing values.
   2. Outlier treatment.

2. Univariate analysis of variables.

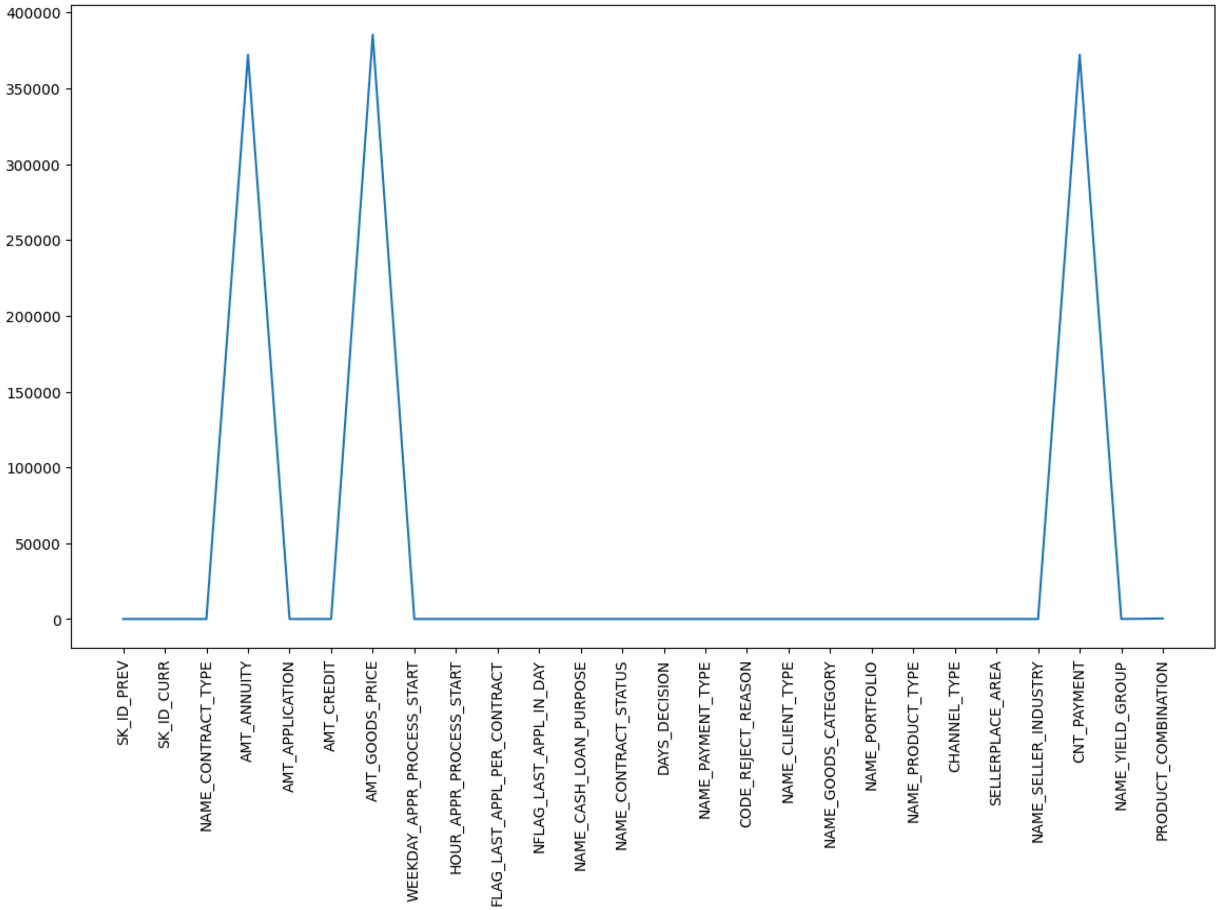
3. Exploration of relationships between variables using correlation matrices and visualization.

4. Merging two data frames into one for further analysis.

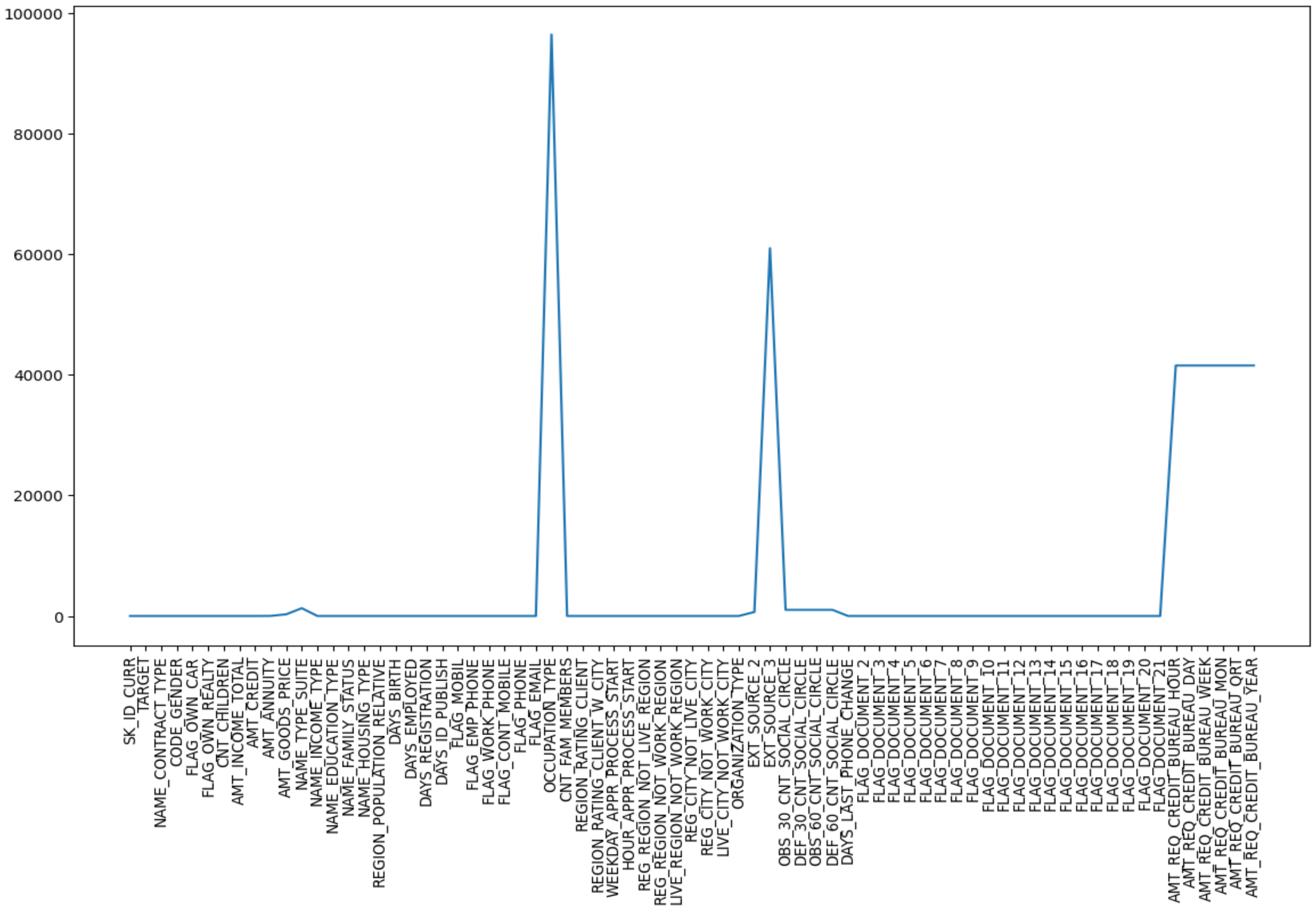
5. Bivariate and multivariate analysis –
   1. Finding the insights about data imbalance.
   2. Finding the top correlation by segmenting the data frame with respect to the target variable.

Presentation of null values after dropping columns with more than 40% null values present in it --

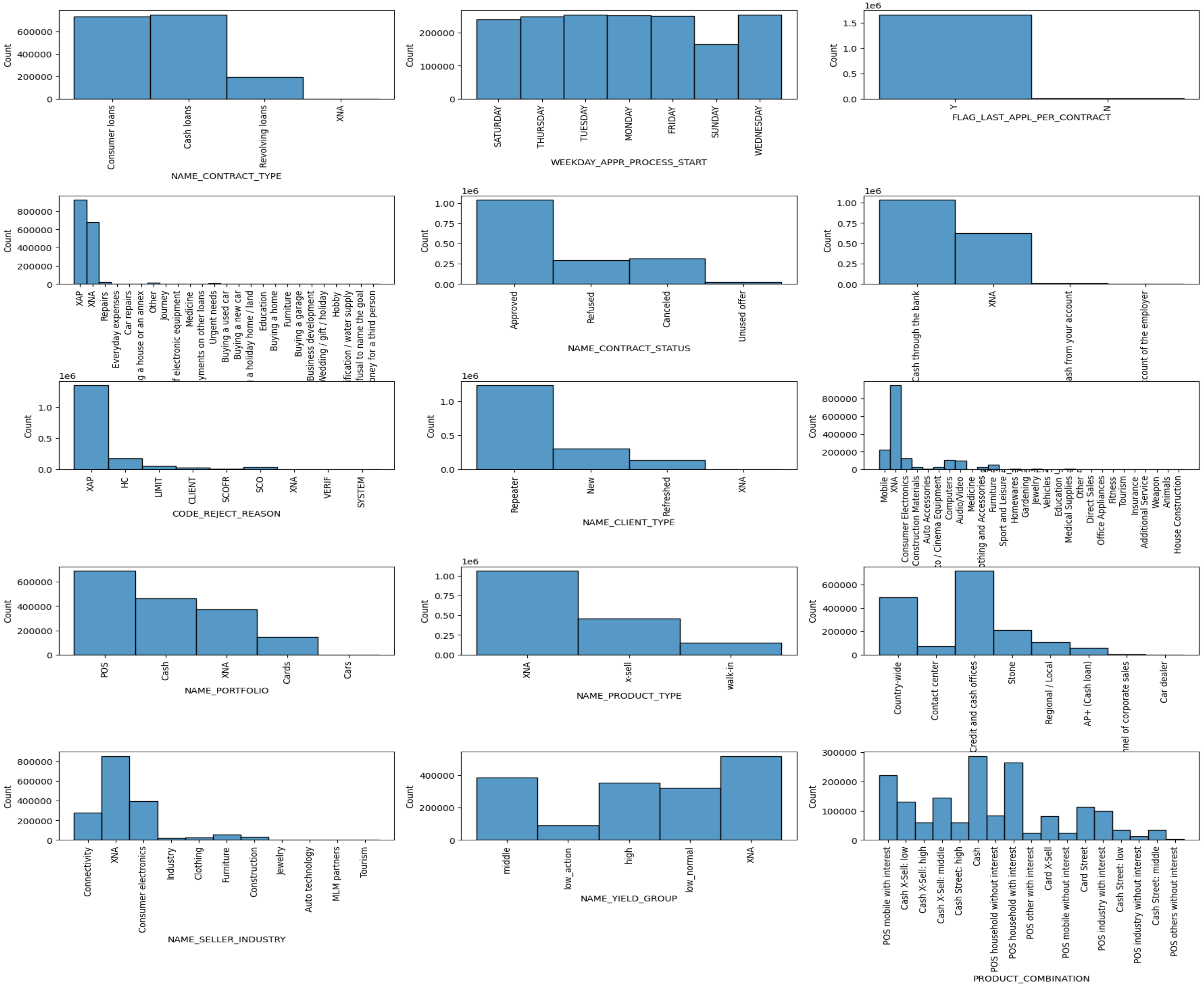
Previous application data



Application data



Univariate analysis of categorical columns of previous application data --



Observations--

1. Cash loans are the highest(747504) with little margin from consumer loans(729145).

2. The client applied for previous application on every day of the week with approximately same frequency except Sunday.

3. In 1661684 cases it was last application for the previous contract.

4. 'XAP' is the highest category of cash loan purpose followed by 'XNA' & 'Repairs'.

5. Previous contract was approved for 1036774 applicants and canceled for 316312 applicants. In 290637 cases it was refused.

6. The client chose 'cash through bank' option in most cases(1033499) and went for cashless in least cases(1084).

7. When applying for previous application the client was repeater in 1231212 cases and new in 301358 cases.

8. The client applied for 'XNA' goods in 950757 cases followed by Mobile goods in 224708 cases in the previous application. Animal and House Construction are the least number of goods here.

9. Previous application was for 'POS' in 691009 cases, 'Cash' in 461514 cases and 'Cars' in only 422 cases.

10. Previous application was 'x-sell' in 456242 cases which is three times as in 'walk-in' cases.

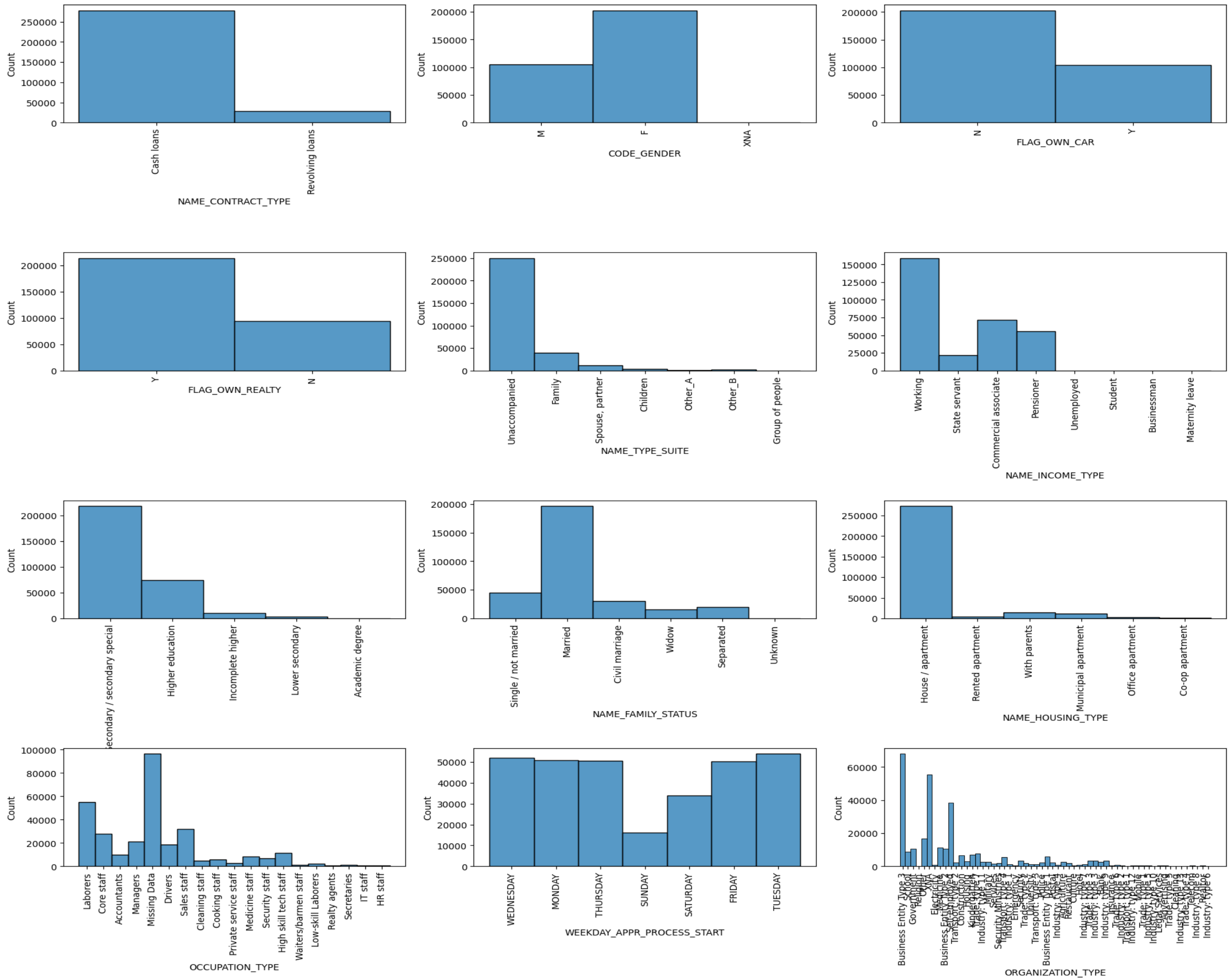
11. We acquired the client on the previous application through 'Credit and cash offices' in highest 719919 cases and least through 'Car dealer' in 449 cases.

12. The industry of seller is 'Consumer electronics' in 398264 cases and 'Tourism' in least 513 cases.

13. Grouped interest rate was in middle category in 385523 cases, high in 353331, low normal in 322066 and low action in least 92025 cases.

14. Product combination in Cash category is highest in 286336 cases followed by POS household with interest category in 263621 cases. It is least in POS others without interest category in 2555 cases.

Univariate analysis of categorical columns of application data --



Observations--

1. Cash loans are the highest in 90% cases with nearly ten times lower Revolving loans in 9% cases.

2. 202441 females applied for loan are nearly twice than male applicants with 105053 loan applications.

3. Clients who owns car are 34% which is approximately half than who doesn't own the car(66% cases).

4. 213303 clients own a house or flat and 94195 clients doesn't own it.

5. The client was unaccompanied in highest 81% cases and was with group of people in least 0.08% cases.

6. Nearly half of the applicants are working(51%).

7. Most of the clients(71%) are having highest education of Secondary/secondary special and only 0.05% clients have academic degree.

8. 63% clients are married and 14% are single followed by Civil marriage, Separated, Widow.

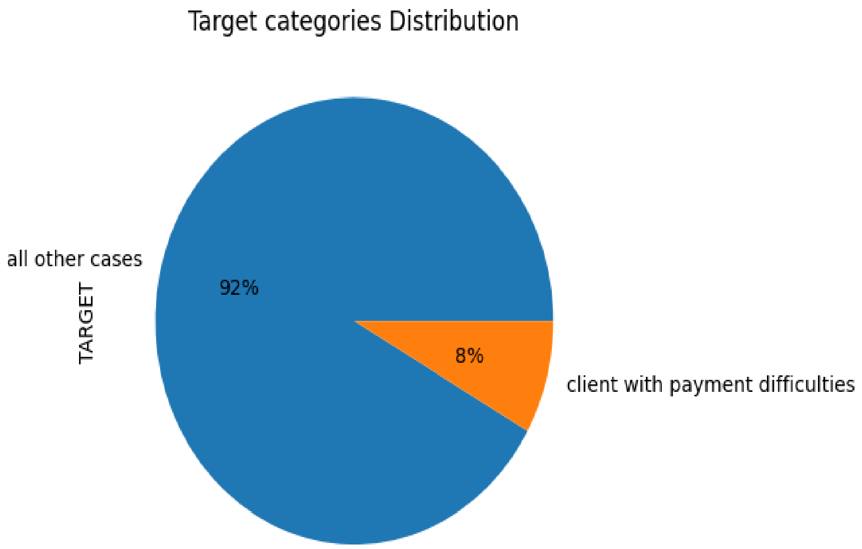
9. 88% clients have houses/apartments and 4% clients are living with parents followed by municipal,rented,office & co-op apartment in least cases.

10. 31% data is missing regarding the occupation of client.Next highest clients are laborers(17%) and least are IT staff in 0.1%.

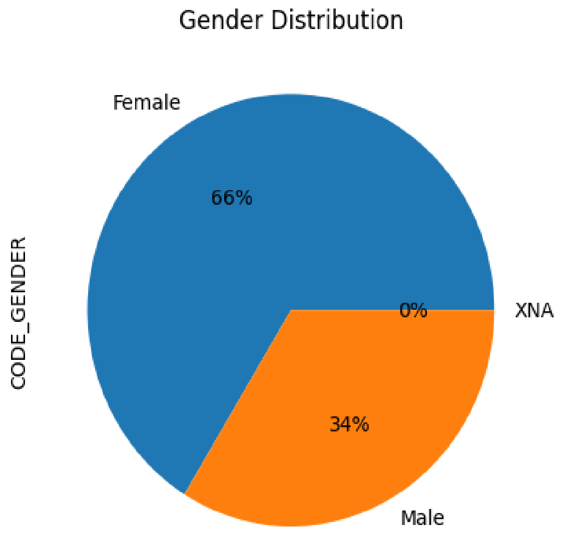
11. The client applied for loan on every day of the week with approximately same frequency except Saturday & Sunday.

12. 22% clients work in Business entity type 3 organization followed by XNA and 12% clients are self-employed.

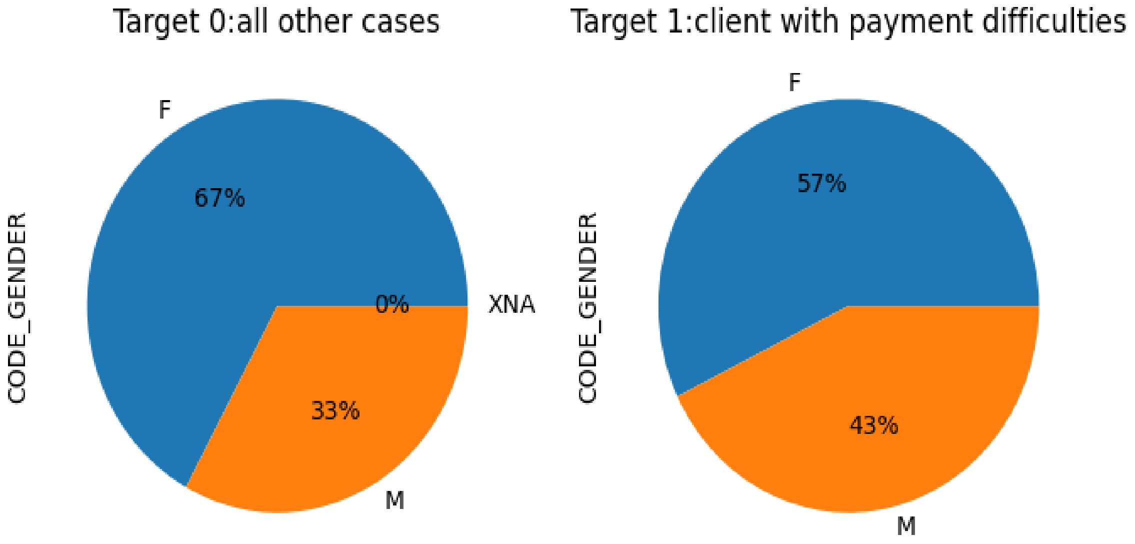
Data imbalance --



Clients with payment difficulties are 8% out of total.

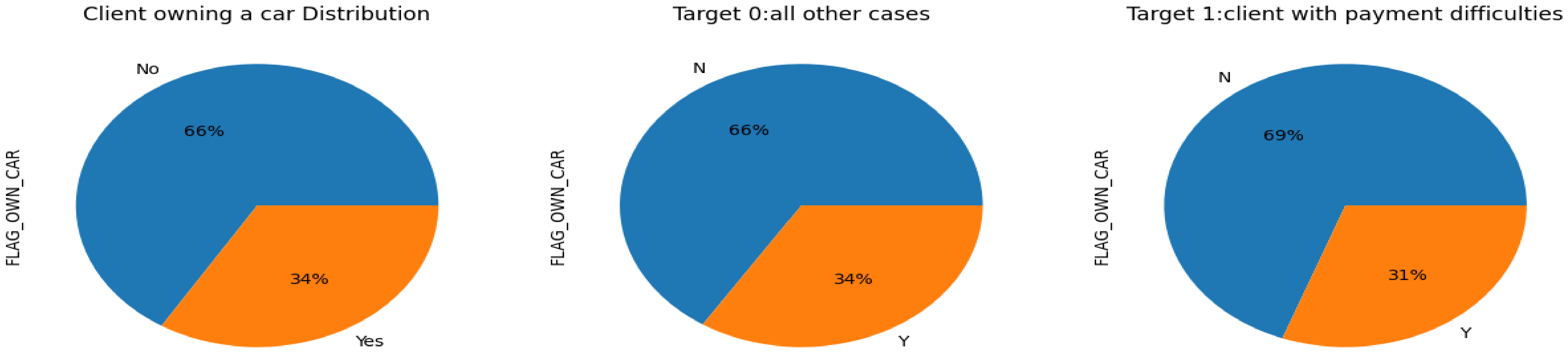


66% females applied for loan and 34% males applied for the loan.



Subplot 1 - 67% females & 33% males are from Target '0' category i.e. all other cases.

Subplot 2 - 57% females & 43% males are from Target '1' category i.e. client with payment difficulties.

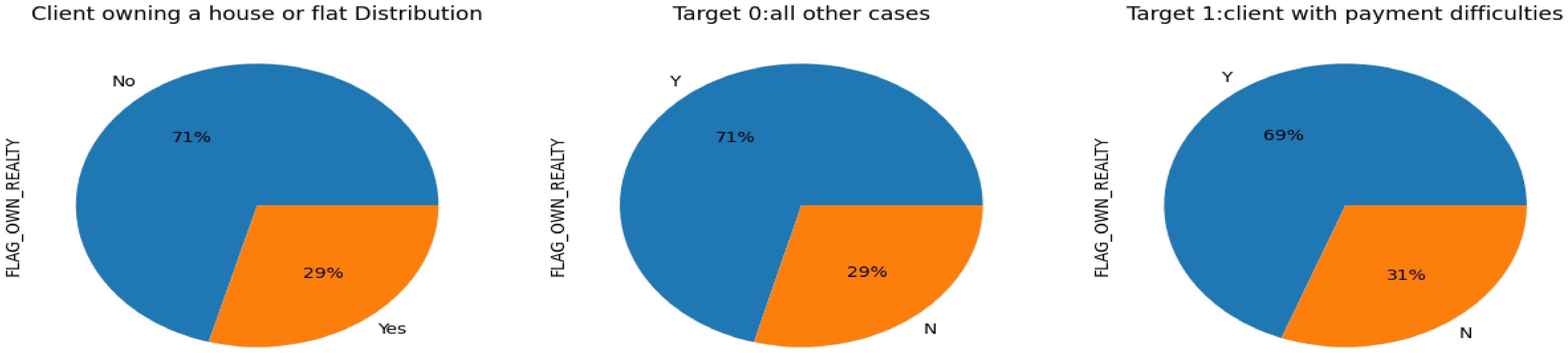


Client owning a car Distribution

Target 0:all other cases

Target 1:client with payment difficulties

Subplot 1 - 34% clients own a car and 66% doesn't out of total clients.
Subplot 2 - In target '0' category, 66% clients doesn't own a car & 34% owns it.
Subplot 3 - In target '1' category, 69% clients doesn't own a car & 31% owns it.

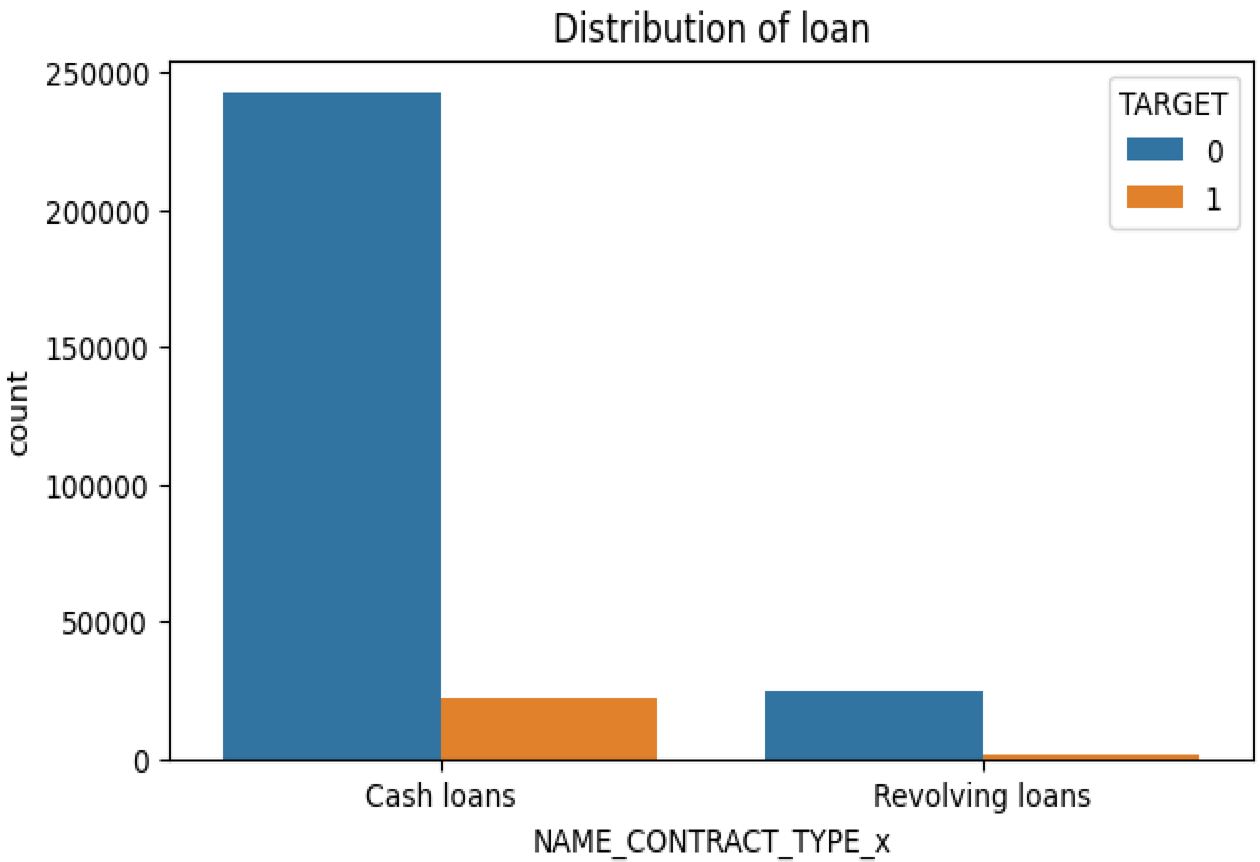


Client owning a house or flat Distribution

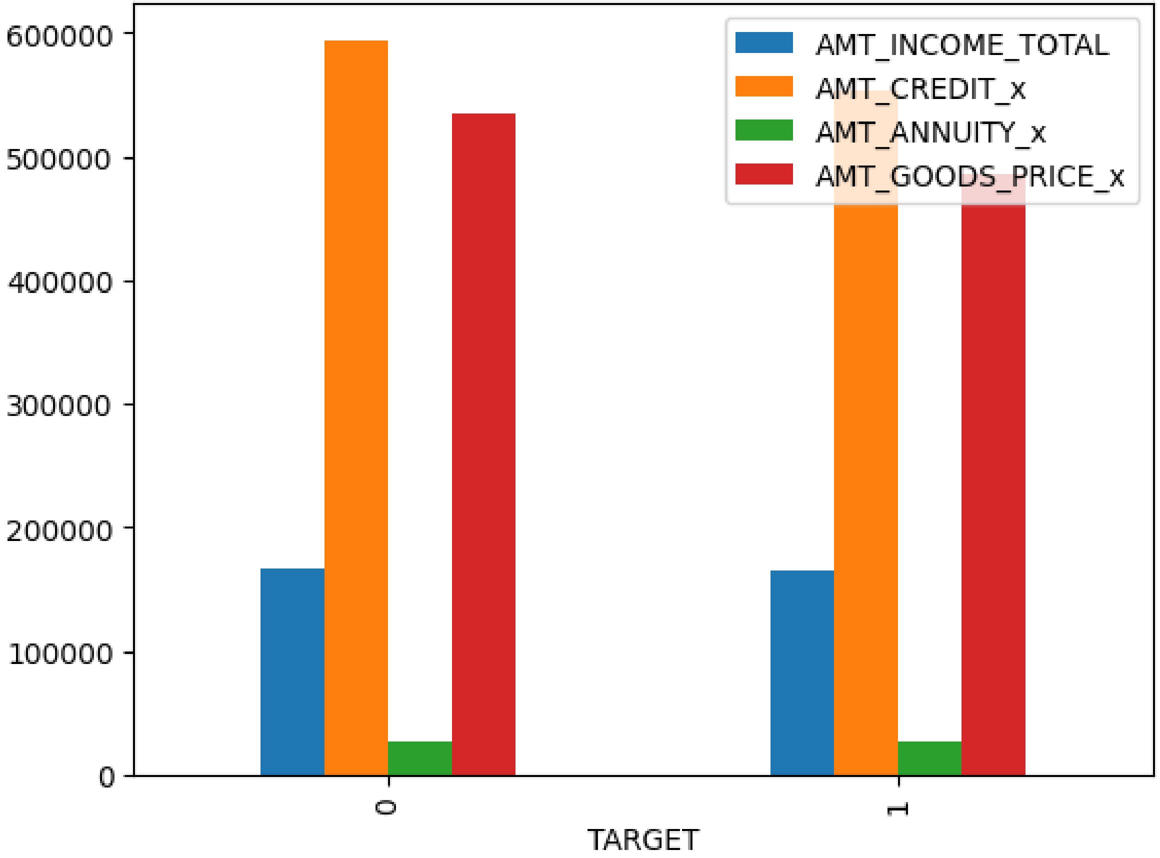
Target 0:all other cases

Target 1:client with payment difficulties

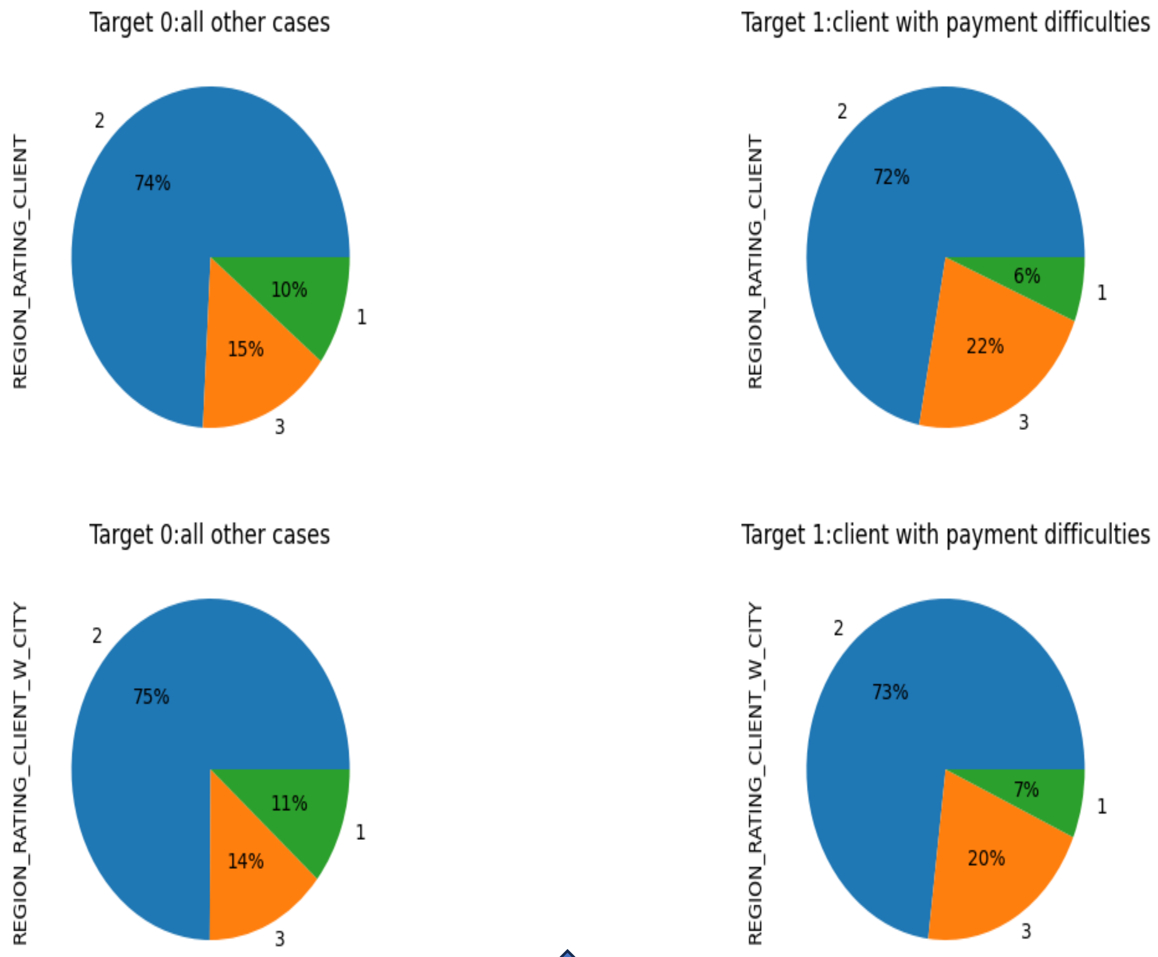
Subplot 1 - 29% clients own a house or flat and 71% doesn't out of total clients.
Subplot 2 - In target '0' category, 29% clients doesn't own a house or flat & 71% owns it.
Subplot 3 - In target '1' category, 31% clients doesn't own a house or flat & 69% owns it.



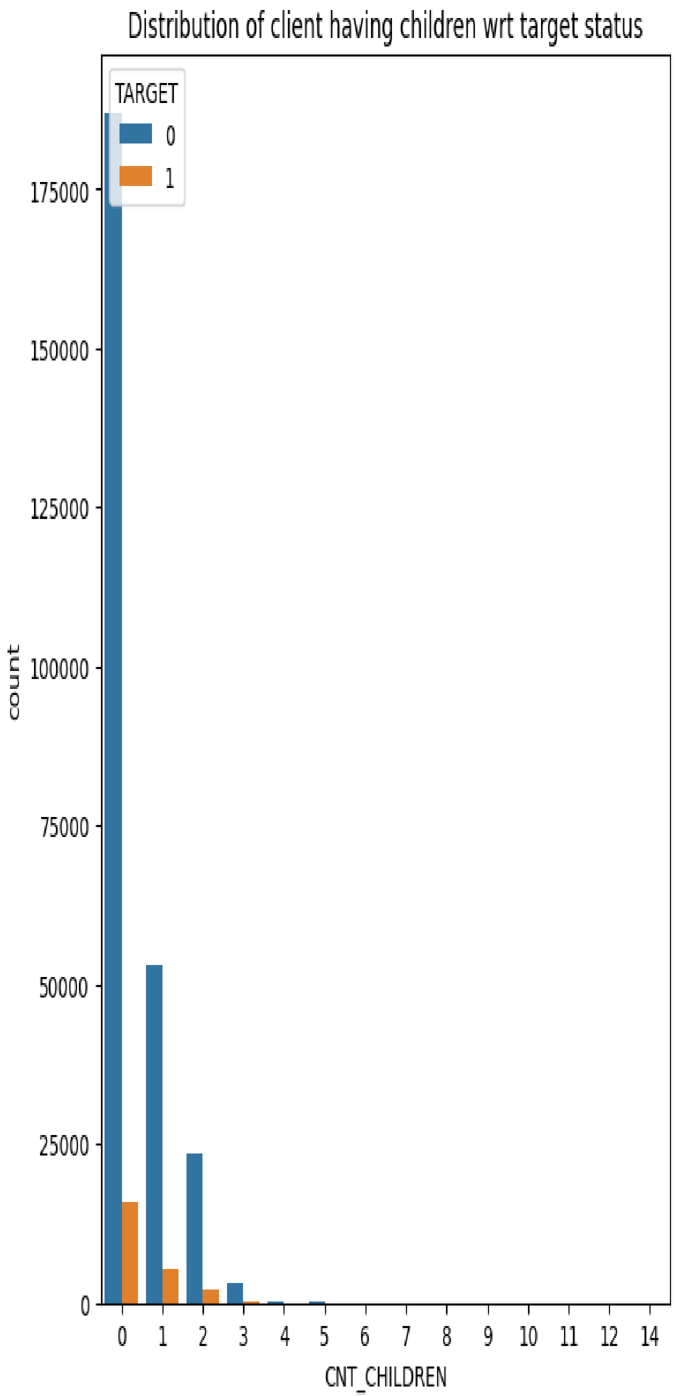
In both the Target categories 'cash' loans are way more than 'revolving' loans.

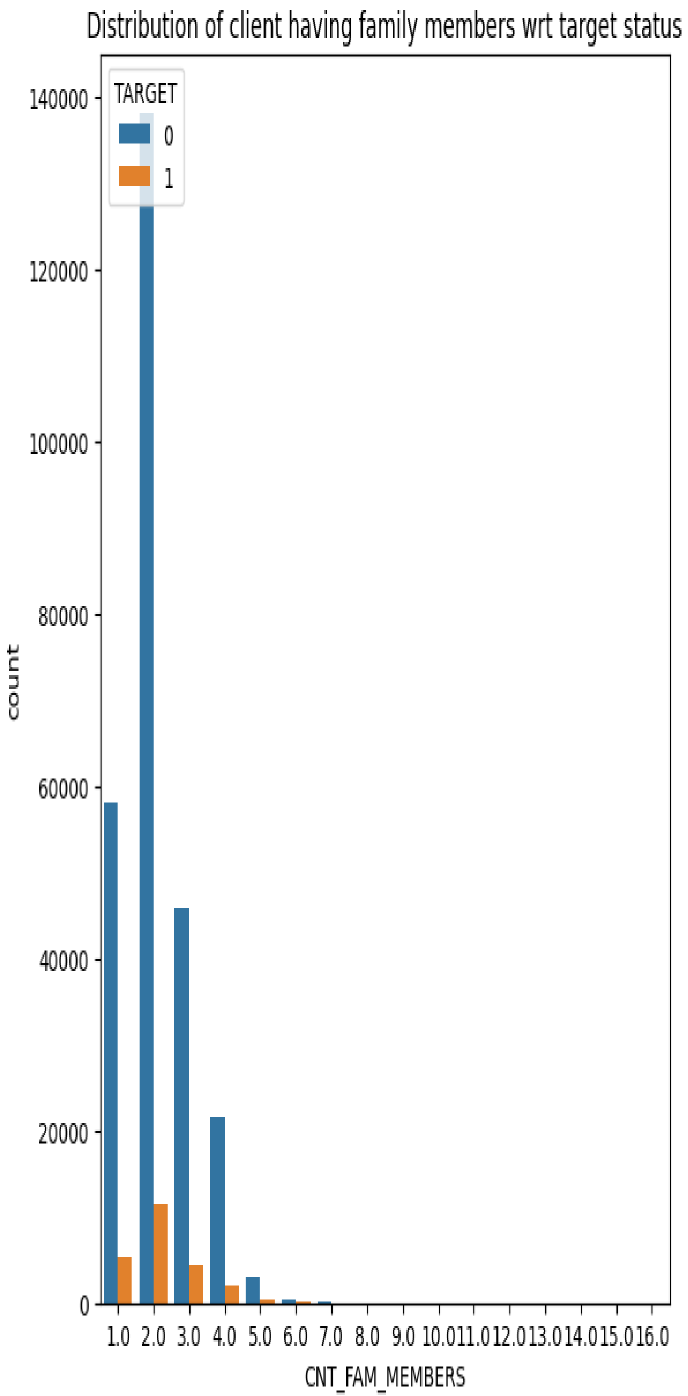


1. Income of client in both the target categories is almost similar.
2. Credit amount is higher in target '0' category.



1. Percentage of clients with payment difficulties are less in rating 1 region(6% < 10%).
2. Percentage of clients with payment difficulties are more in rating 3 region(20% > 14%).





Subplot 1 - In both the Target categories 'cash' loans are way more than 'revolving' loans.

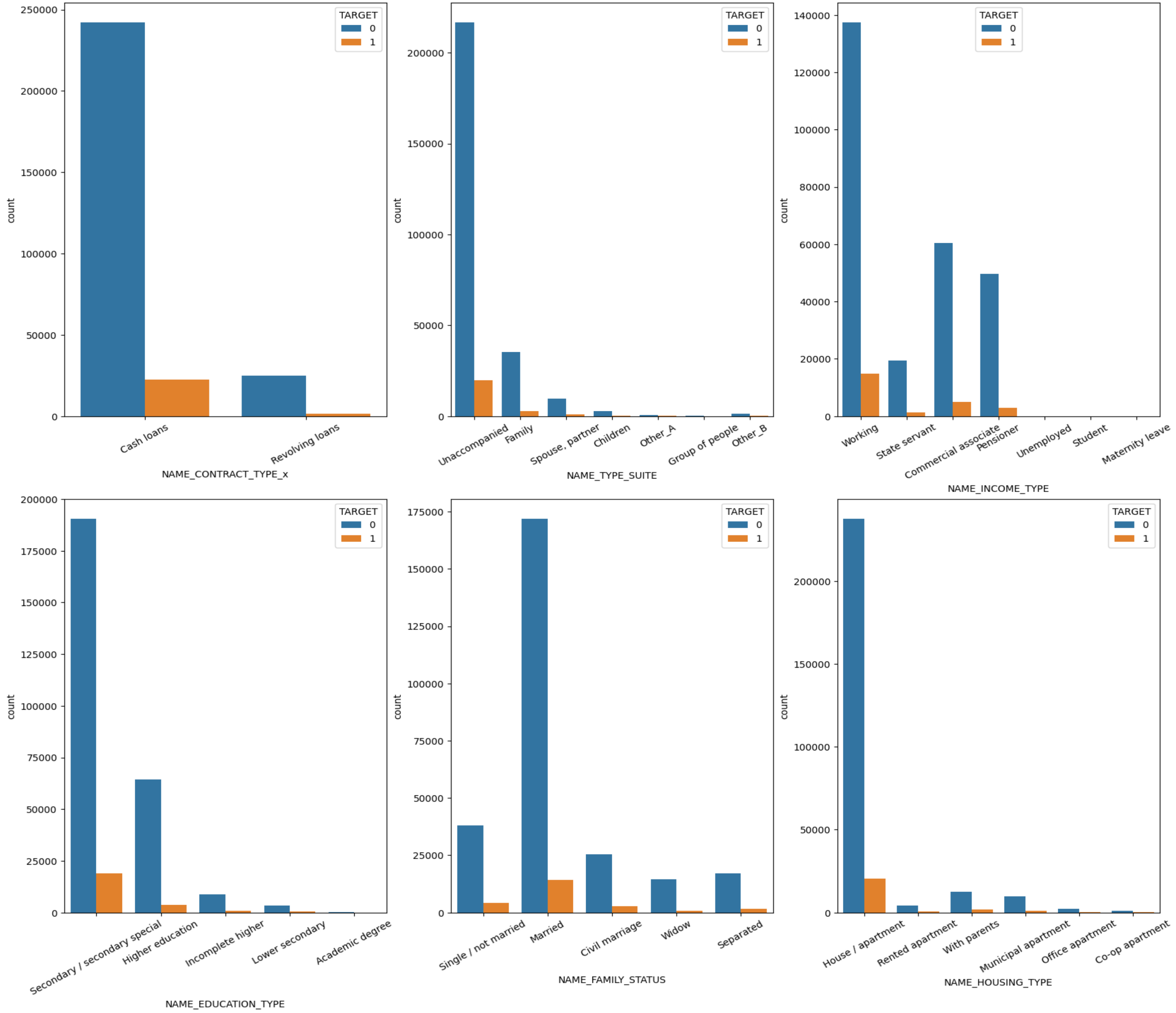
Subplot 2 - In both the Target categories clients were 'Unaccompanied' in highest number of cases.

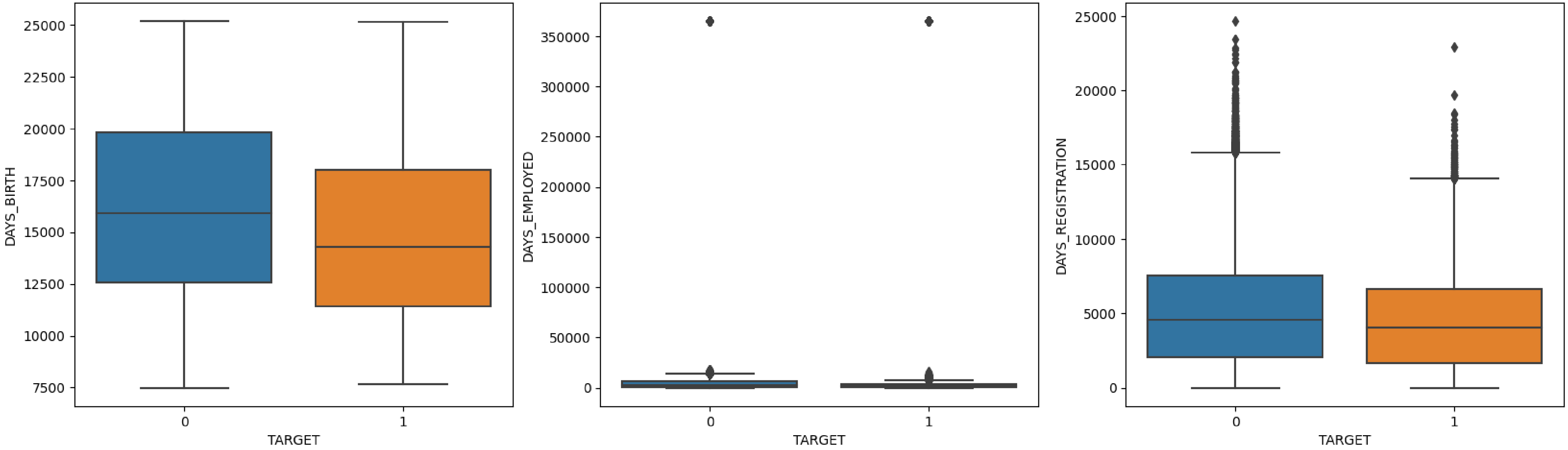
Subplot 3 - In both the Target categories 'working' is the income type of highest number of clients.

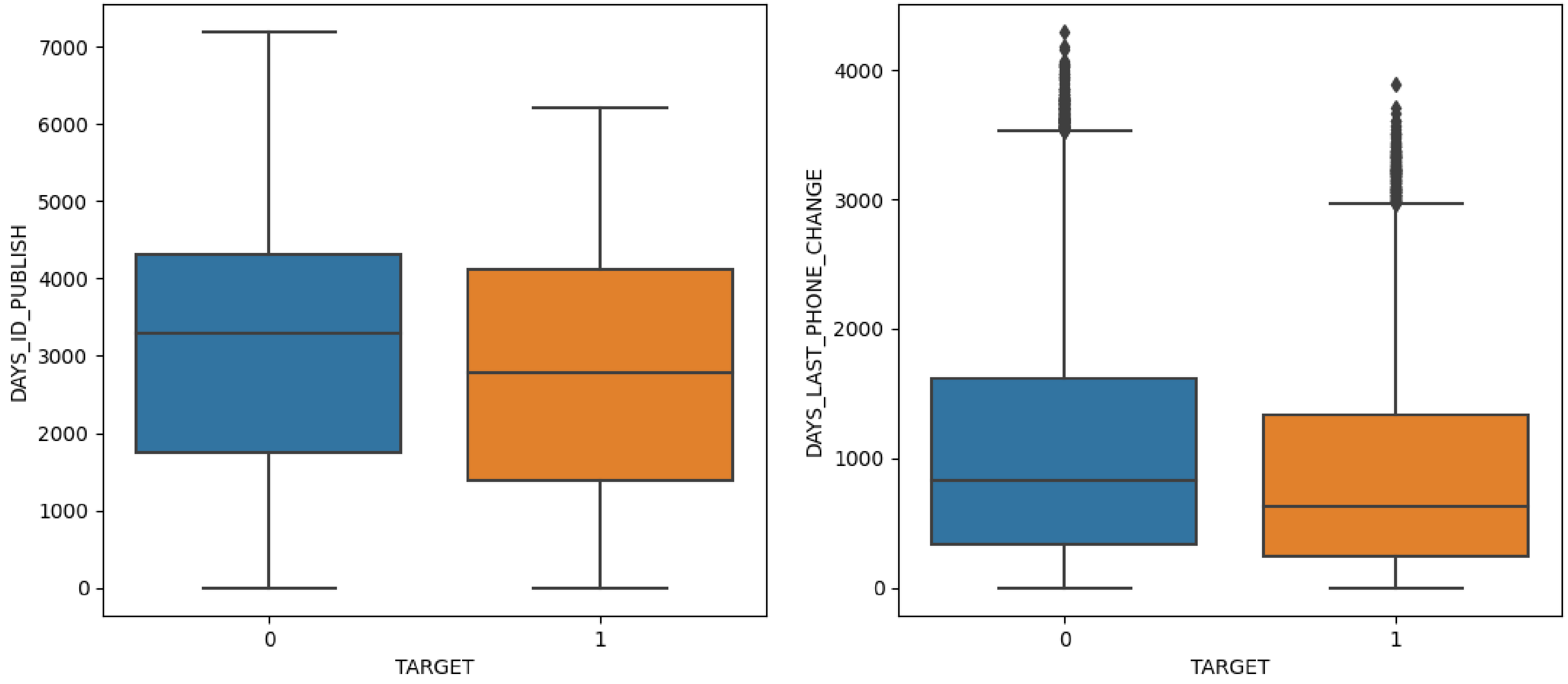
Subplot 4 - Most of the clients from both target categories are having highest education of Secondary/secondary special.

Subplot 5 - Highest number of clients are married in both target categories.

Subplot 6 - Clients have house/apartment in most cases in both target categories.







Subplot 1 - Age distribution of clients in both target categories.

Subplot 2 - How many days before the application the person started current employment.

Subplot 3 - How many days before the application did client change his registration.

Subplot 4 - How many days before the application did client change the identity document with which he applied for the loan.

Subplot 5 - How many days before application did client change phone.

Subplot 1 - Did client provide mobile phone (1=YES, 0=NO).

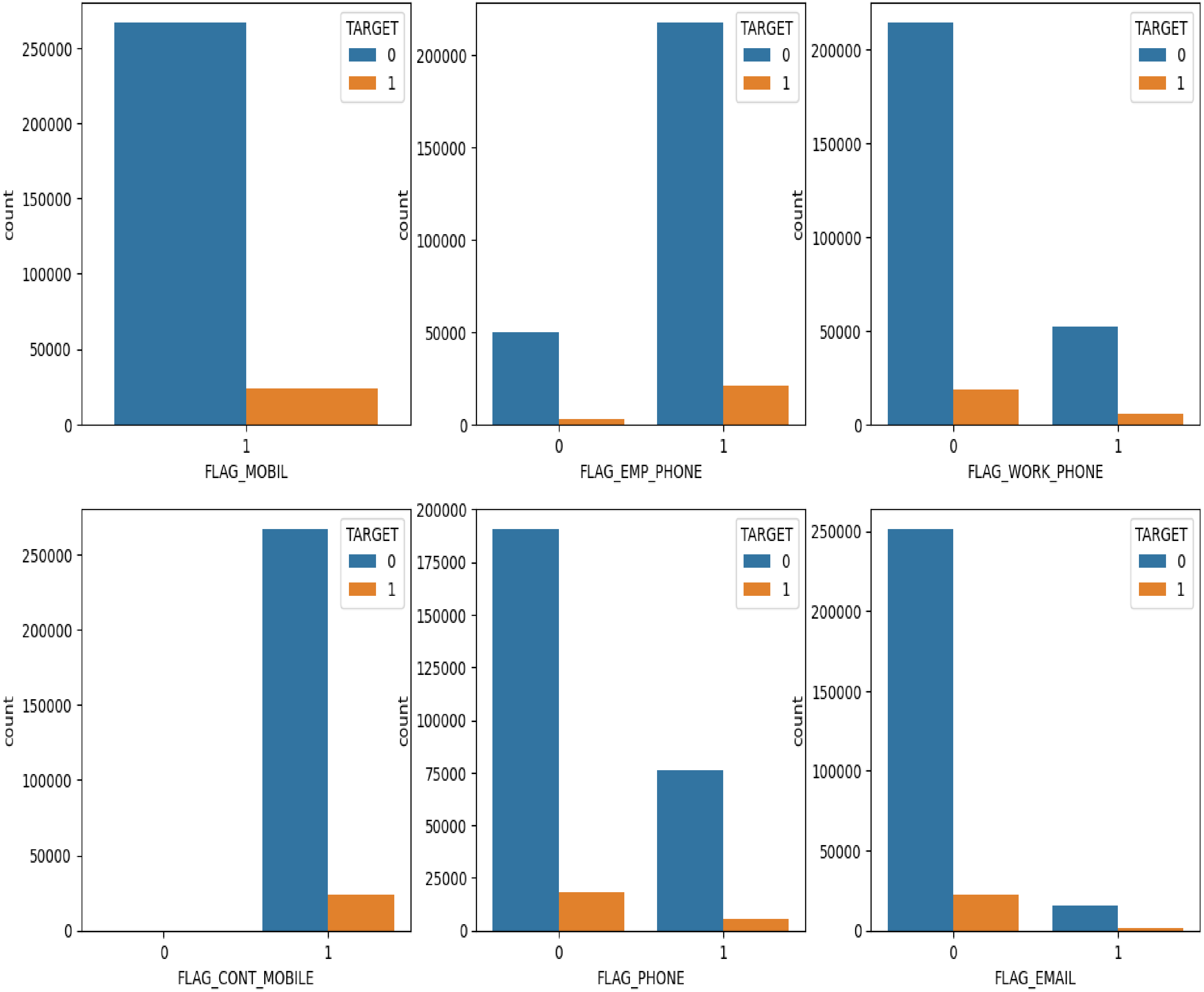
Subplot 2 - Did client provide work phone (1=YES, 0=NO).

Subplot 3 - Did client provide home phone (1=YES, 0=NO).

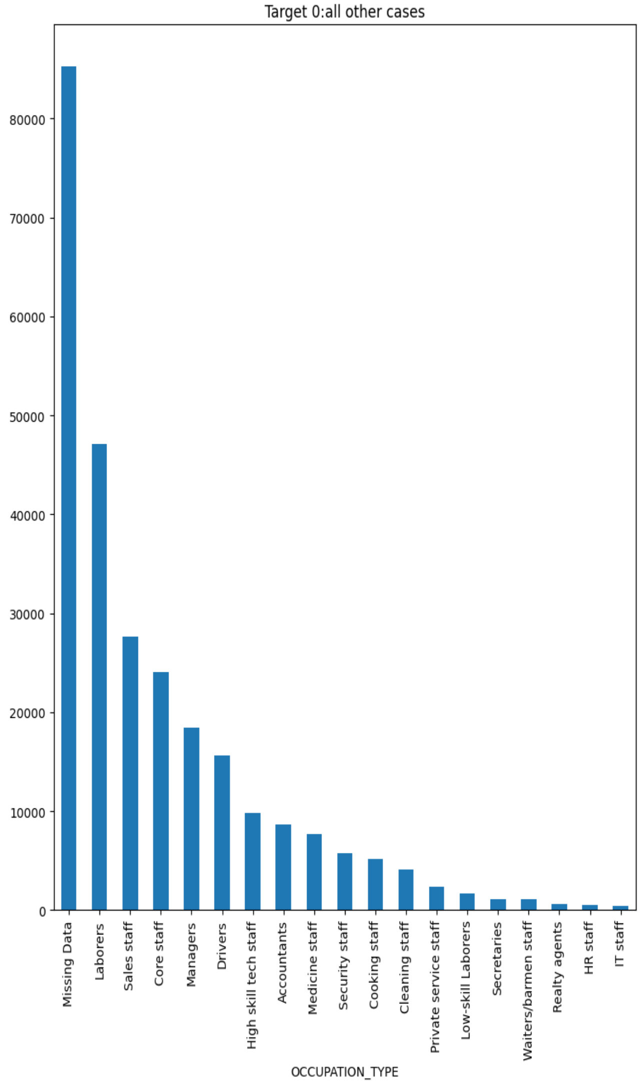
Subplot 4 - Was mobile phone reachable (1=YES, 0=NO).

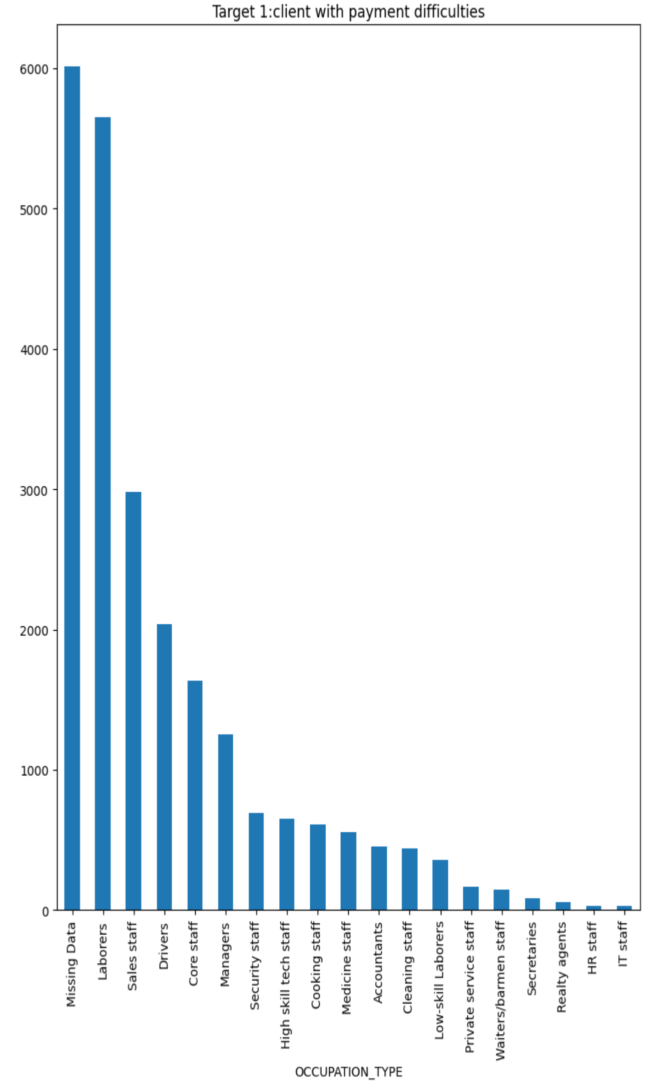
Subplot 5 - Did client provide home phone (1=YES, 0=NO).

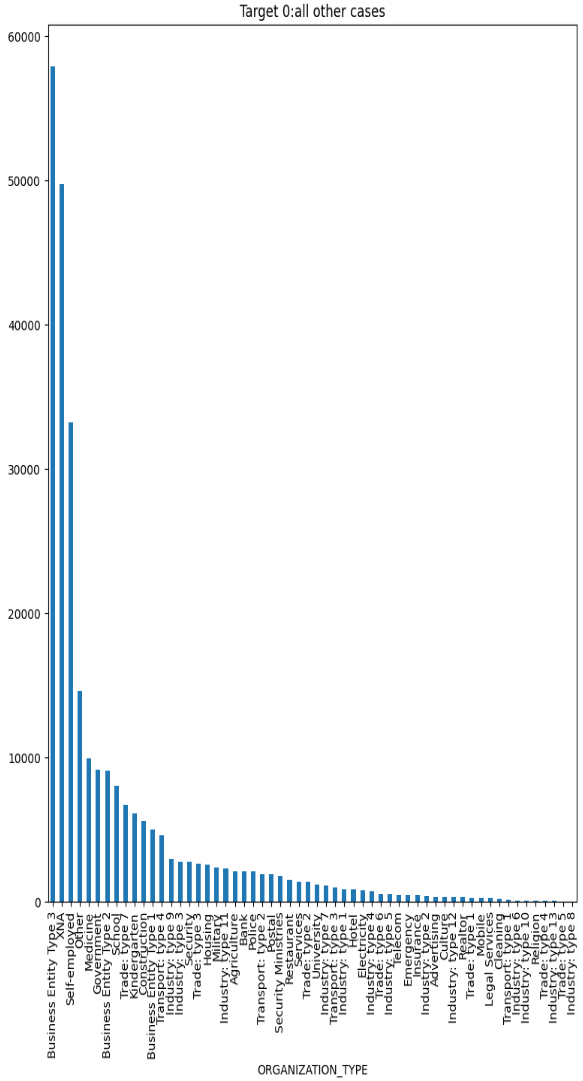
Subplot 6 - Did client provide email (1=YES, 0=NO).

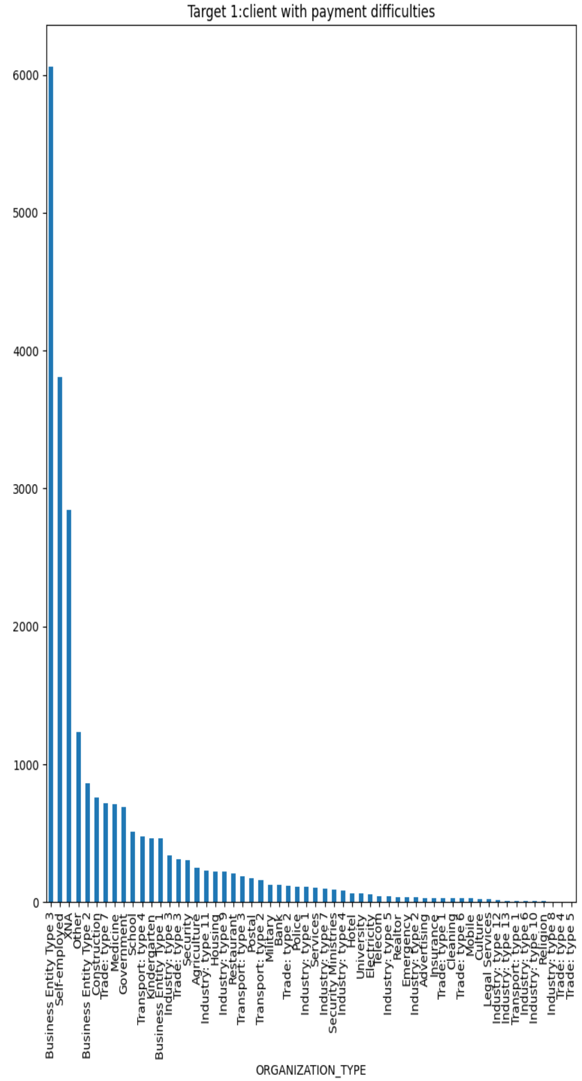


1. Data is missing in both the categories.
2. Highest number of clients belong to 'Laborers' category in both target categories.









1. Data is missing(i.e. 'XNA' category) in both the categories.
2. Highest number of clients belong to 'Business Entity Type 3' category in both target categories followed by 'Self-employed.

Top correlations by segmenting the data frame with respect to the target variable

Target '0' category i.e. all other cases			Target '1' category i.e. client with payment difficulties		
Var1	Var2	Correlation	Var1	Var2	Correlation
FLAG_EMP_PHONE	DAYS_EMPLOYED	0.999774	DAYS_EMPLOYED	FLAG_EMP_PHONE	0.999695
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998399	OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998257
AMT_GOODS_PRICE_x	AMT_CREDIT_x	0.986700	AMT_GOODS_PRICE_y	AMT_APPLICATION	0.983176
AMT_GOODS_PRICE_y	AMT_APPLICATION	0.986603	AMT_CREDIT_x	AMT_GOODS_PRICE_x	0.982312
AMT_APPLICATION	AMT_CREDIT_y	0.970960	AMT_CREDIT_y	AMT_APPLICATION	0.969466
AMT_GOODS_PRICE_y	AMT_CREDIT_y	0.968847		AMT_GOODS_PRICE_y	0.963026
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.949235	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.957446
CNT_CHILDREN	CNT_FAM_MEMBERS	0.879662	CNT_FAM_MEMBERS	CNT_CHILDREN	0.886857
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.867661	DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.869502
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.857442	REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.855025

Conclusion --

After the data cleaning, we performed univariate analysis of variables, bivariate and multivariate analysis of variables with respect to ‘target’ variable.

We derived some insights about the important variables for differentiating the **clients with payment difficulties with all other cases.**

We also derived top correlations of variables by segmenting the data frame with respect to ‘target’ variable. And it was done using the correlation matrices and heatmaps.