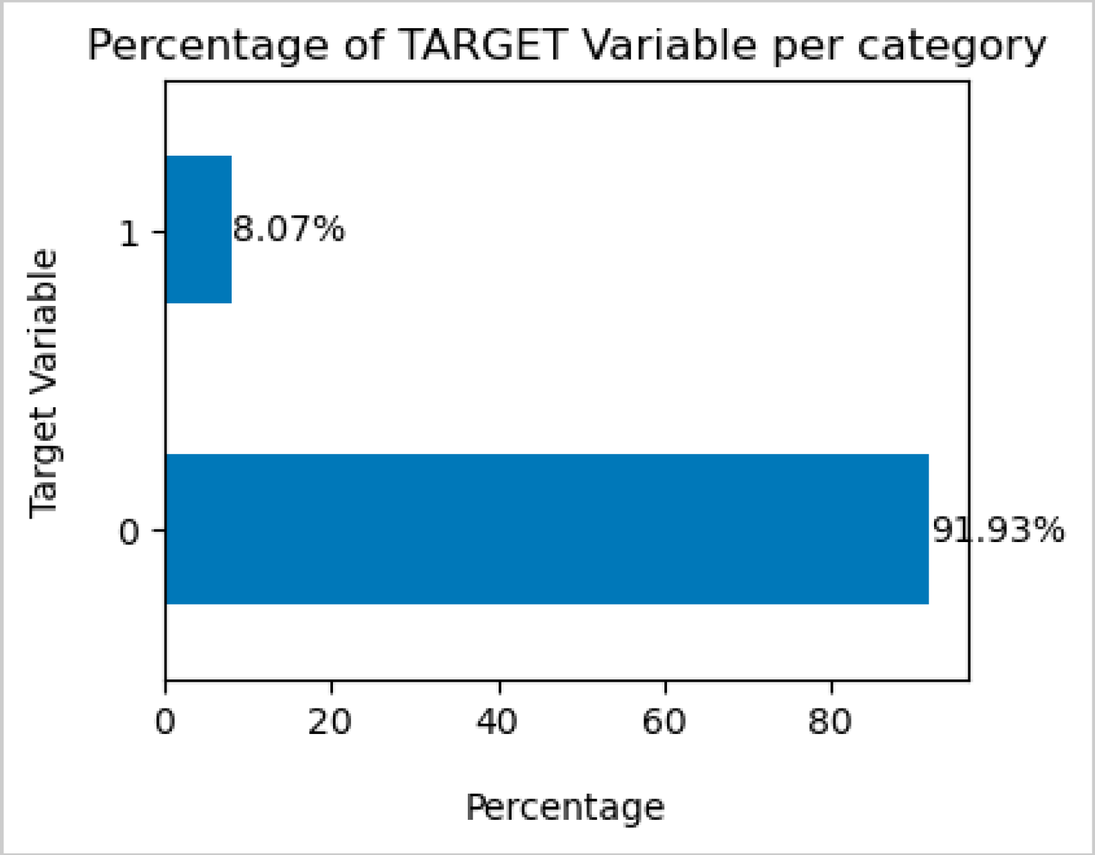
APPLICATION DATA

TARGET VARIABLE

Findings

- The target variable indicates if the client has any payment difficulties. (0=no difficulties, 1= payment difficulties)
- The data is highly imbalanced as 92% of client has no payment difficulties while only 8% of client has payment difficulties. (92:8)
- This indicates that most of the loan is able to be paid back on time

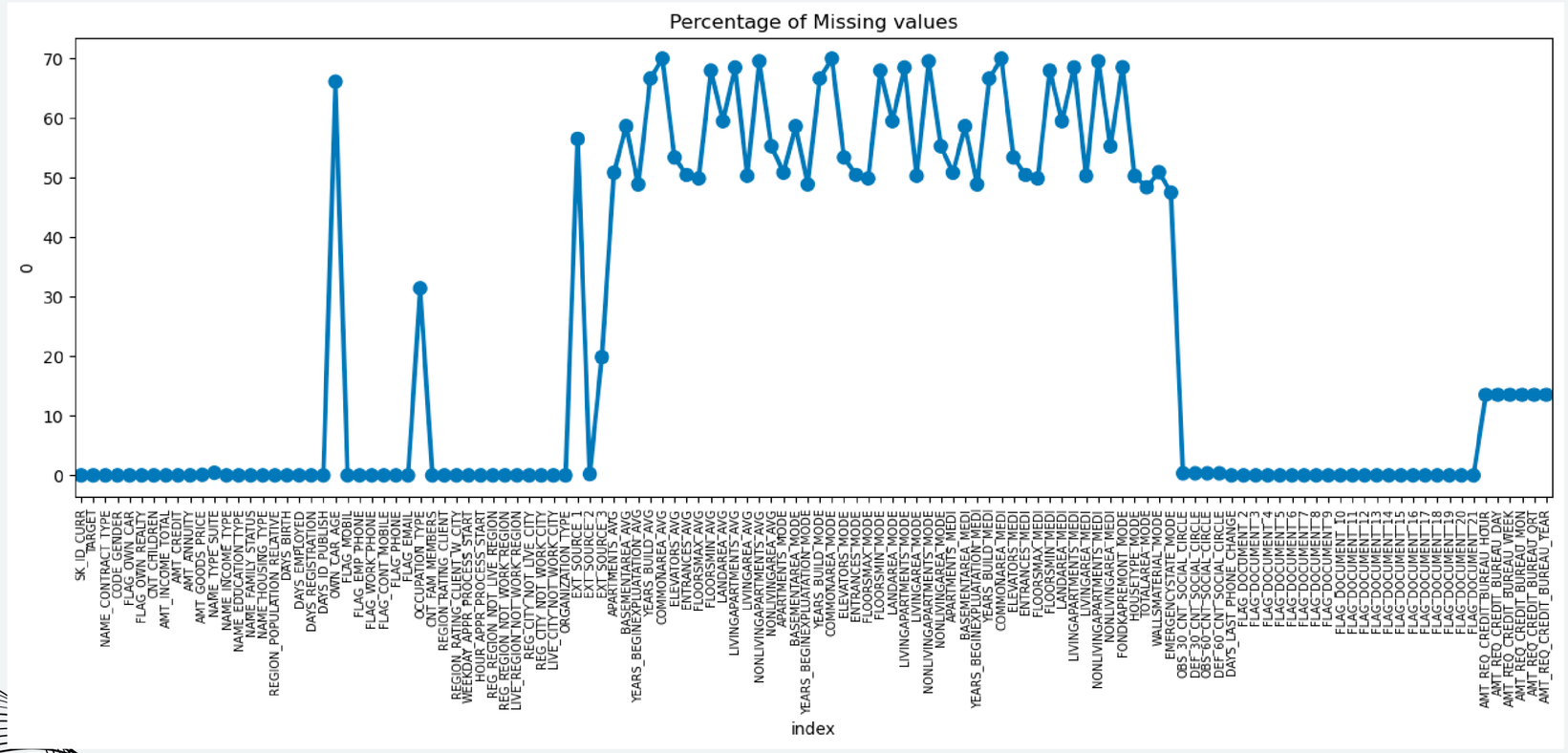


MISSING DATA

Findings

- The percentage of null values identified in the columns ranges from 13.5% to 69.4%.

- Nearly half of the columns contains null values in them.



DATA CLEANING

Findings

We noted that the following columns (**89 columns**) are not relevant to our EDA process, hence we have further removed them in our data cleaning process. Those categories include:

(i) ID

(ii) Day and hour the client applied for the loan

(iii) Client's permanent/ contact address does not match work address

(iv) Normalized score from external data source

(v) Normalized information about building where the client lives

(vi) How many days before application did client change

(vii) Document provided by the client

(viii)Number of enquiries to Credit Bureau

(ix) Observations of client's social surroundings

TREATMENT FOR MISSING VALUES

More than 40%

Drop the whole column (**OWN_CAR_AGE**) as columns with large amount of missing data wouldn't provide useful insights on the analysis.

Between 20%-40%

Since there's only 1 column (***OCCUPATION_TYPE***) that falls within this range, we have replaced the null values using the **mode** as this column contains categorical data.

Below 20%

As the remaining columns with null values fall below 1%, hence suggest to leave for further processing.

UNIVARIATE ANALYSIS

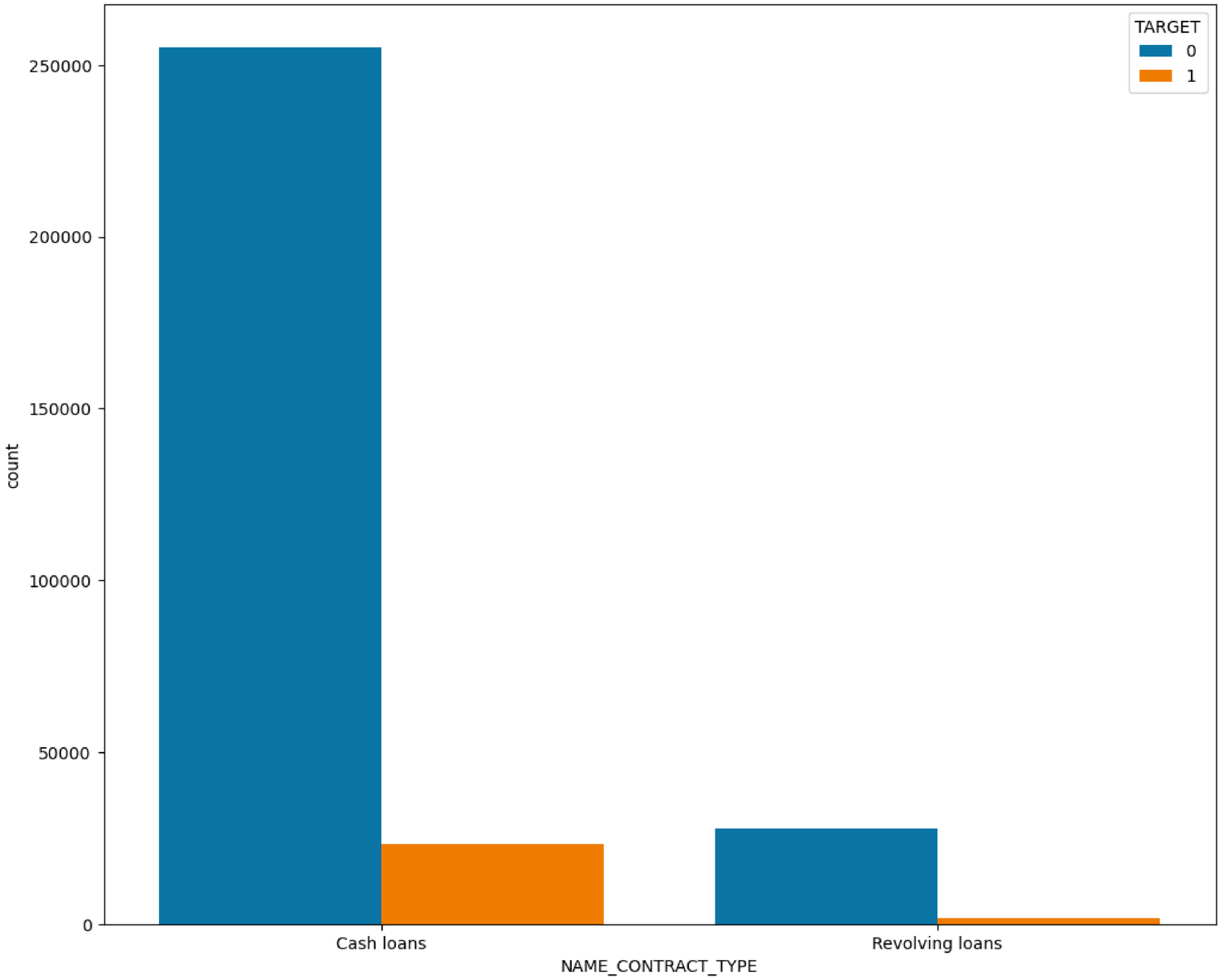
Findings

- We used univariate analysis to analyse categorical data. We identified the type of data in the columns by running '***.info()***' and '***.nunique()***'.

- By analysing single variables, the information that we can obtain is quite limited. However, we did get some very obvious insights for certain columns.

- Since there are so many columns being analysed, we have chosen the top 5 most informative ones to present here. Please refer to the following slides for further information.

Contract type vs target variable



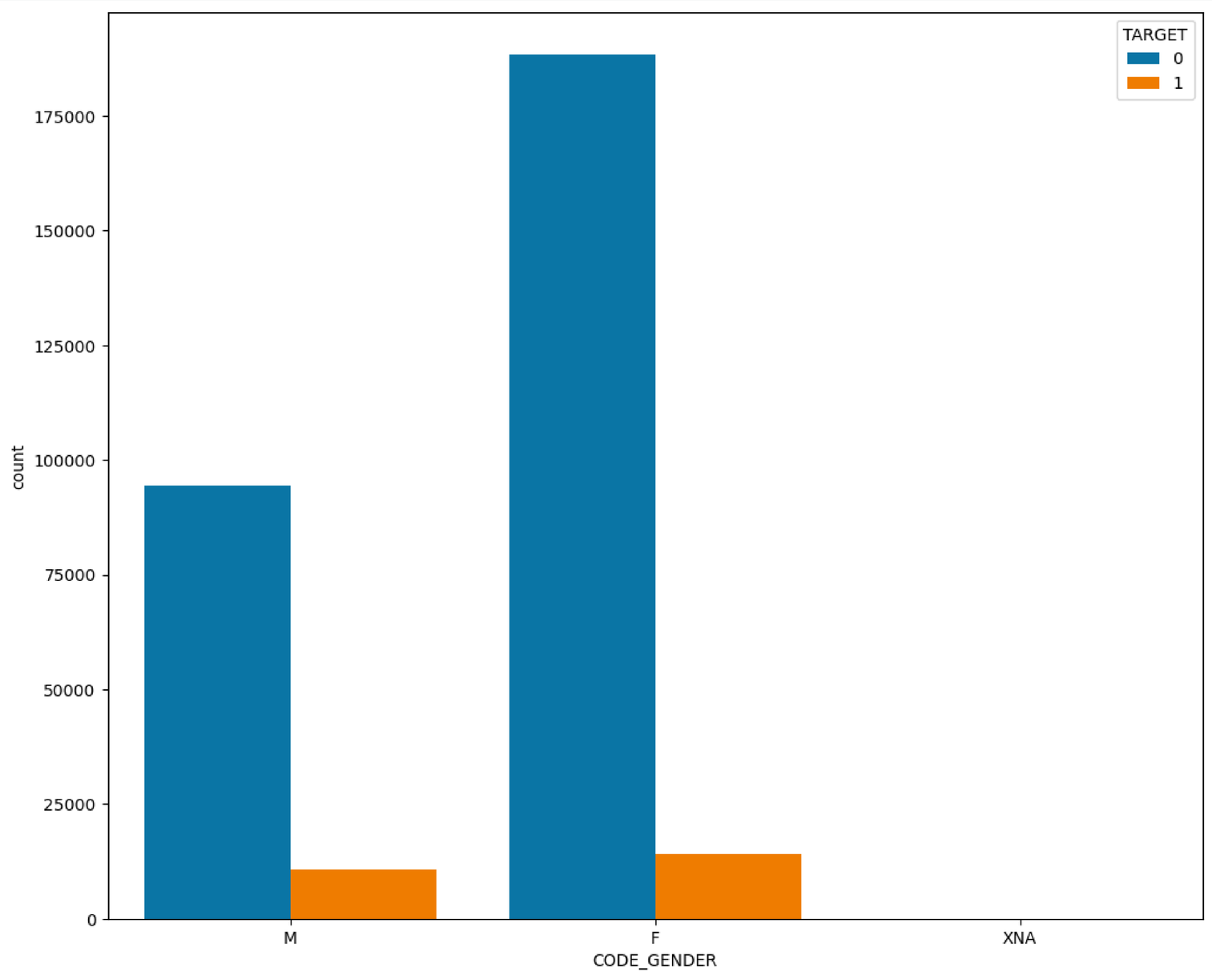
Findings

- Most people applied to cash loans as compared to revolving loans.
- A larger number of clients with payment difficulties can be seen for those who applied for cash loans.

Note:

0 = no payment difficulties
1= client with payment difficultites

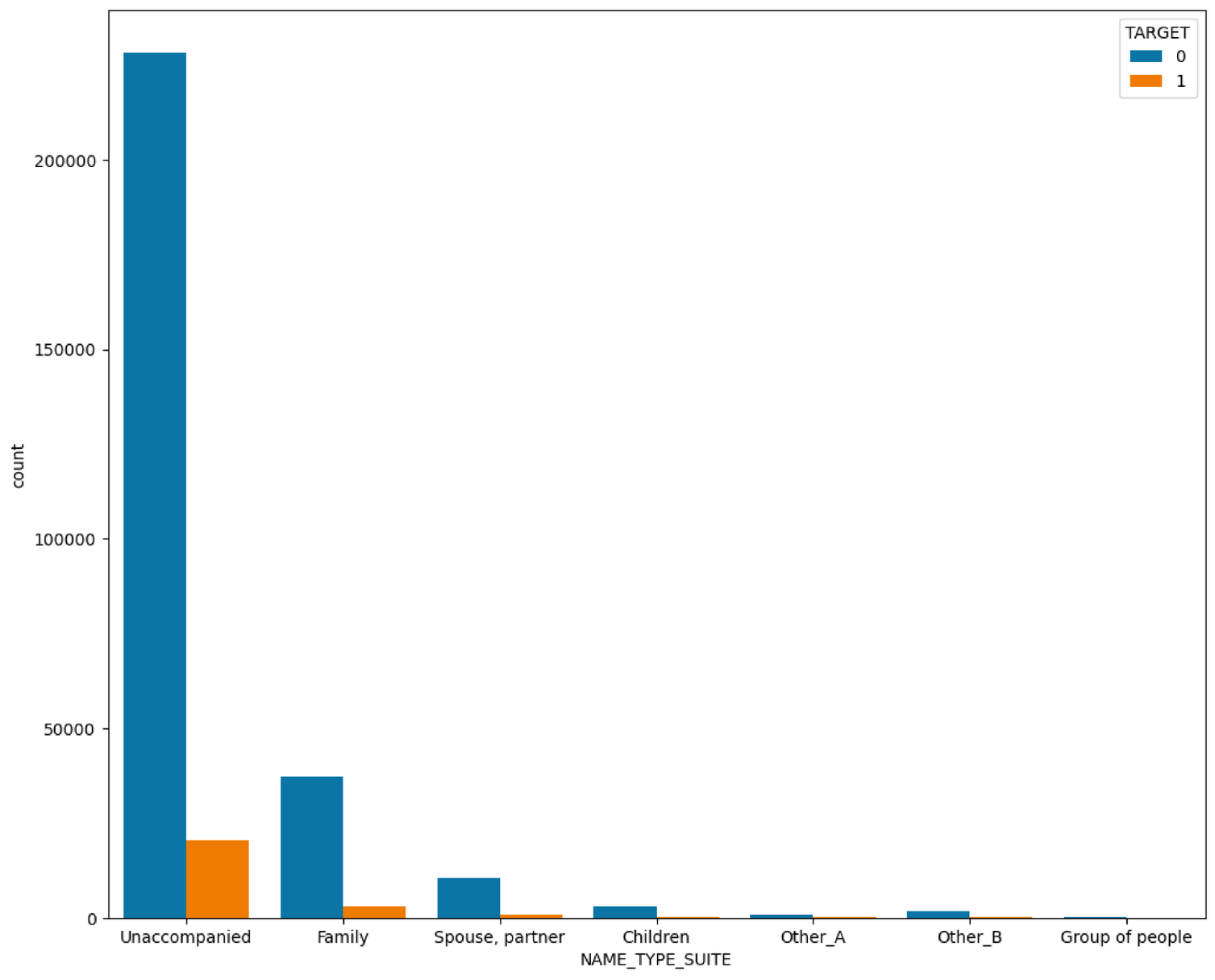
Gender vs target variable



Findings

- More females applied to the loan as compared to male
- More than 175k of females have no payment difficulties, while approx 95k males were in the same position.
- When comparing the ratio of both genders, males has a higher ratio of having payment difficulties compared to females.

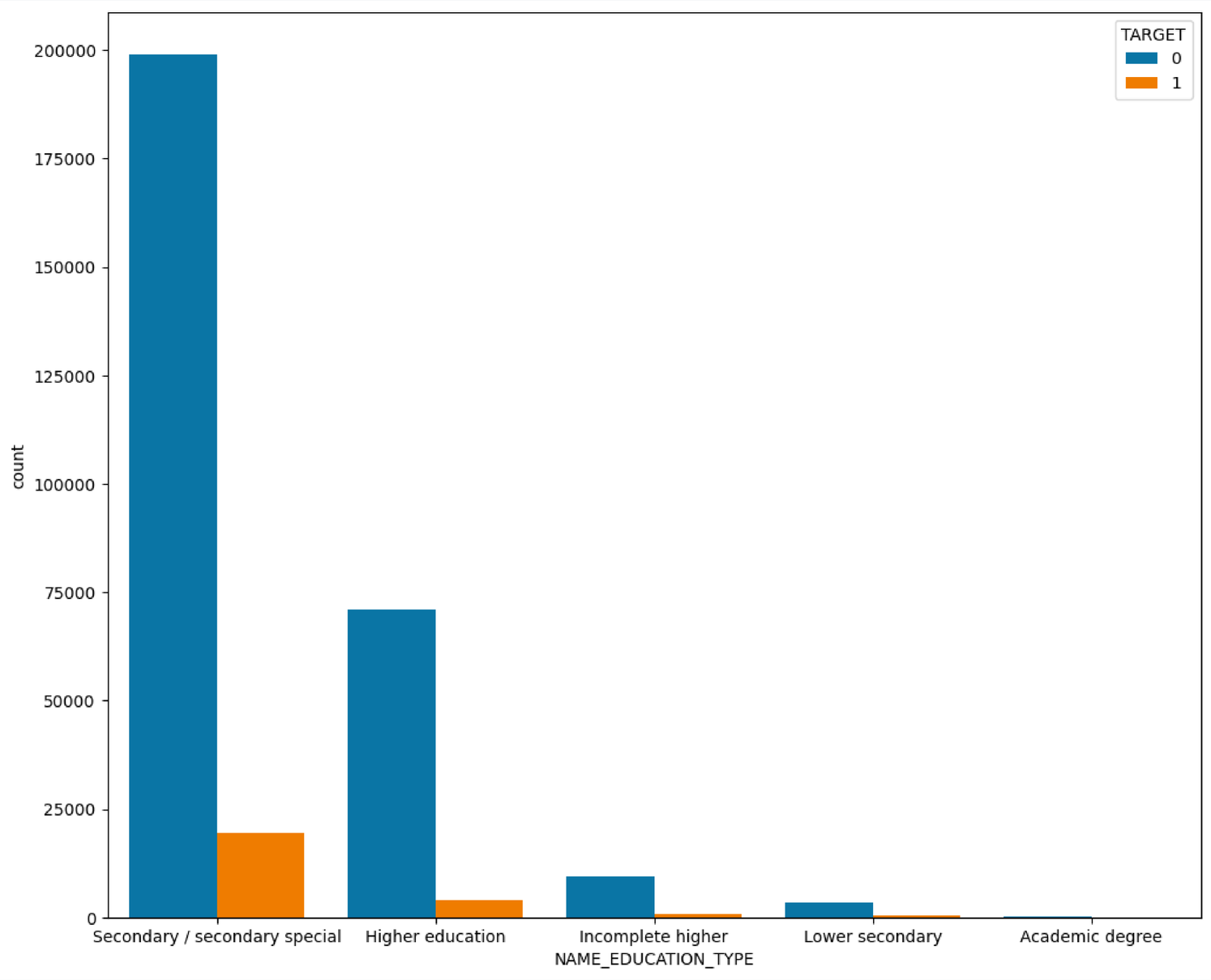
Accompanied person vs target variable



Findings

- Most people went to apply for the loan without anyone accompanying them
- The second highest category is clients accompanied by their family.
- Since the number of unaccompanied is the highest, the number of client with payment difficulties is also the highest under this category.

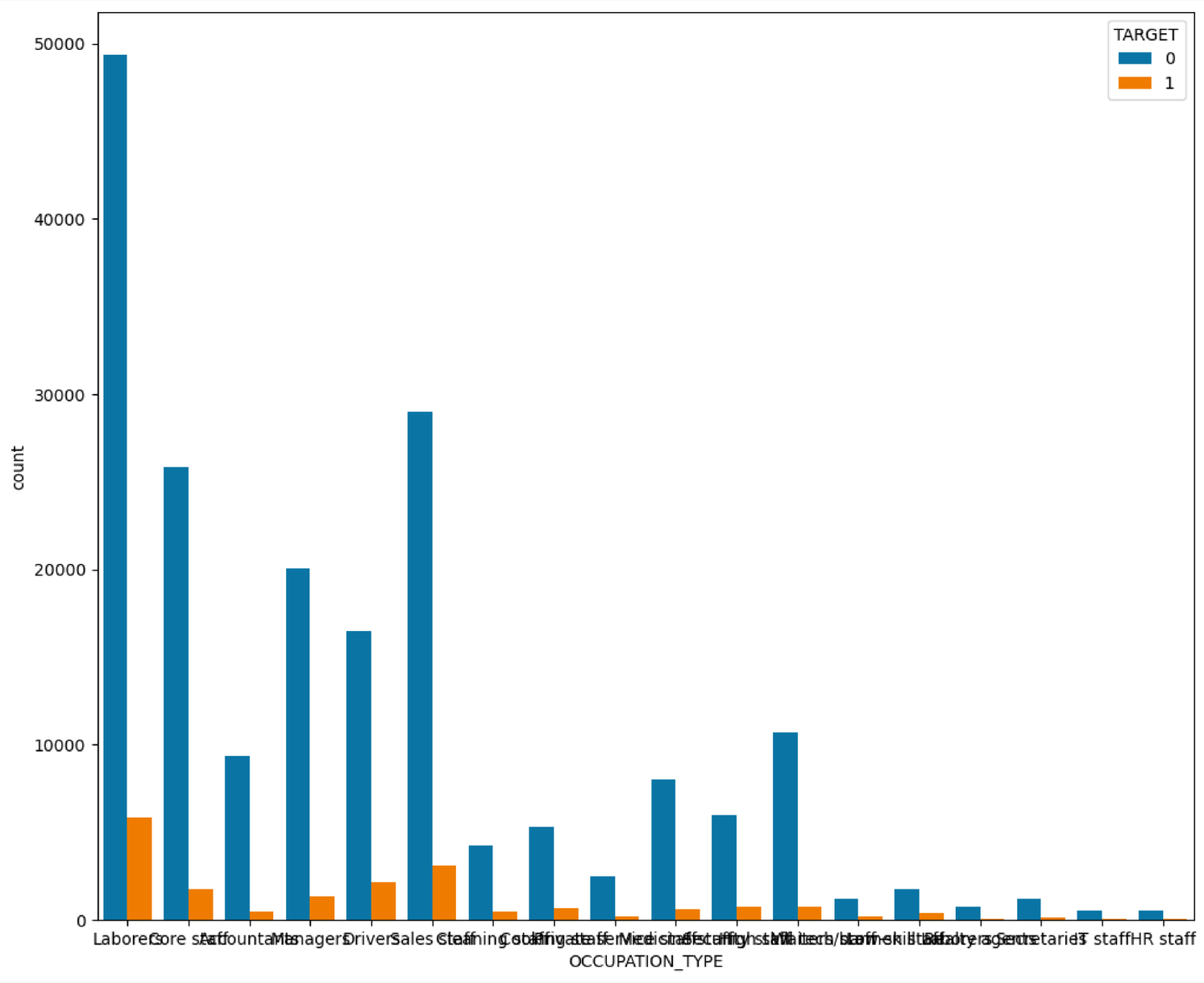
Education level vs target variable



Findings

- Most loan applicants are secondary school graduates, with higher education being the second highest category
- Approx 25k of applicants under secondary/ secondary special have payment difficulties.

Occupation type vs target variable



Findings

- The highest category of applicants are labourers (151k applicants), which also has the highest number of clients with payment difficulties.
- The second and third highest category of applicants are sales staff (32k applicants) and core staff (28k applicants) respectively.

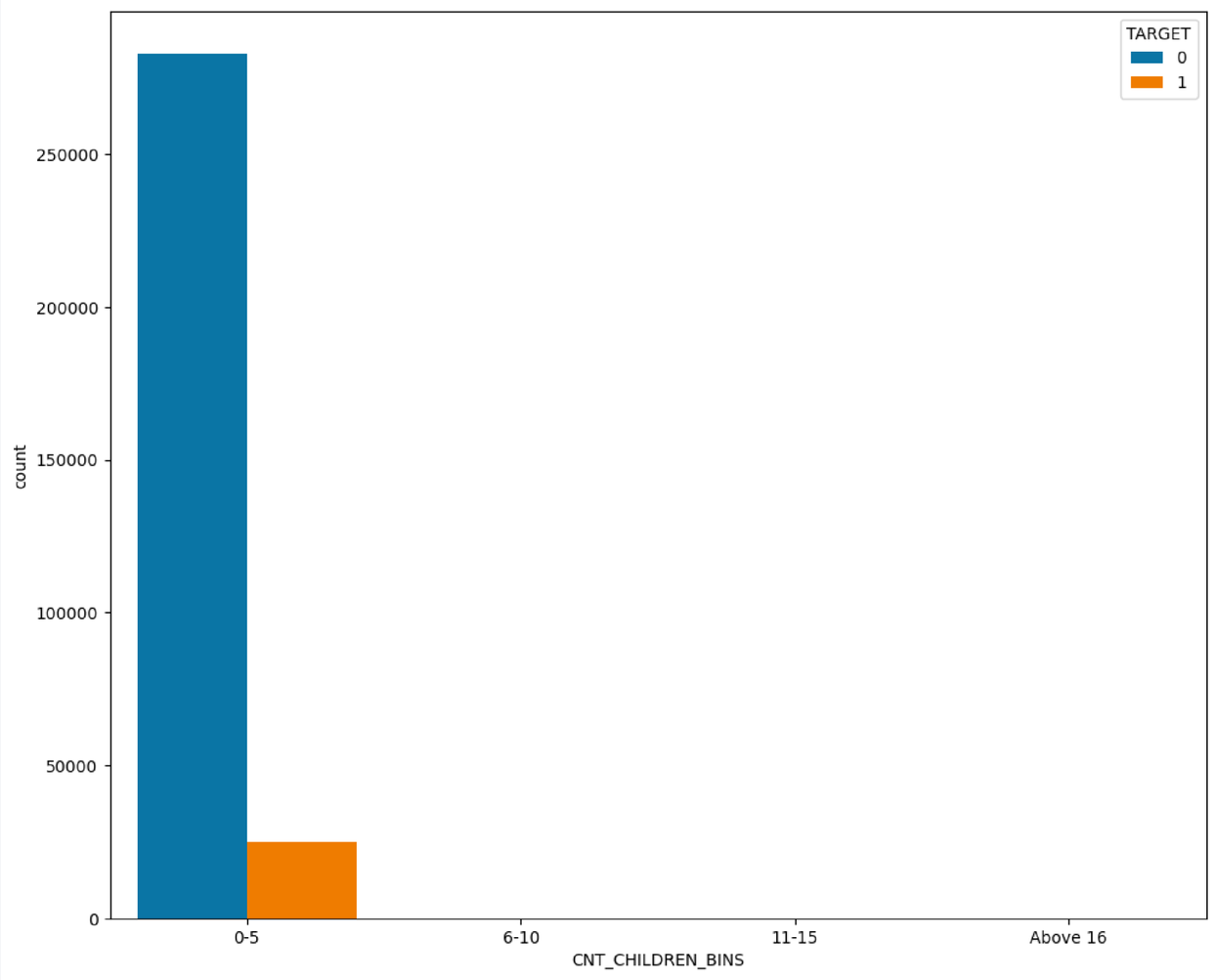
NUMERICAL ANALYSIS

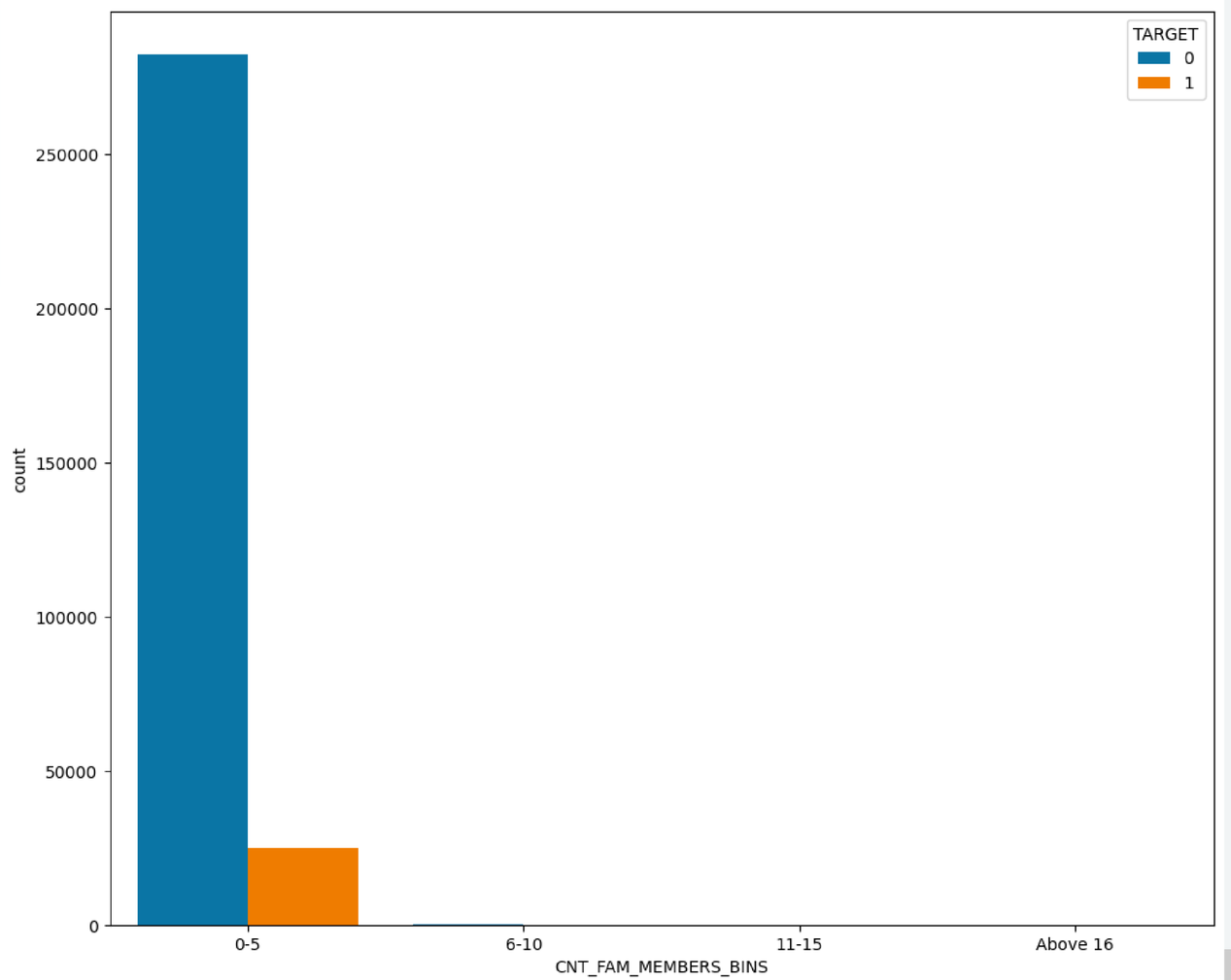
Findings

- We used numerical analysis to analyse numerical data. We identified the type of data in the columns by running '***.info()***' and '***.nunique()***'.

- We mainly used KDE plot to analyse the numerical data as it gives a better presentation of the dispersion of the data for each target variable.

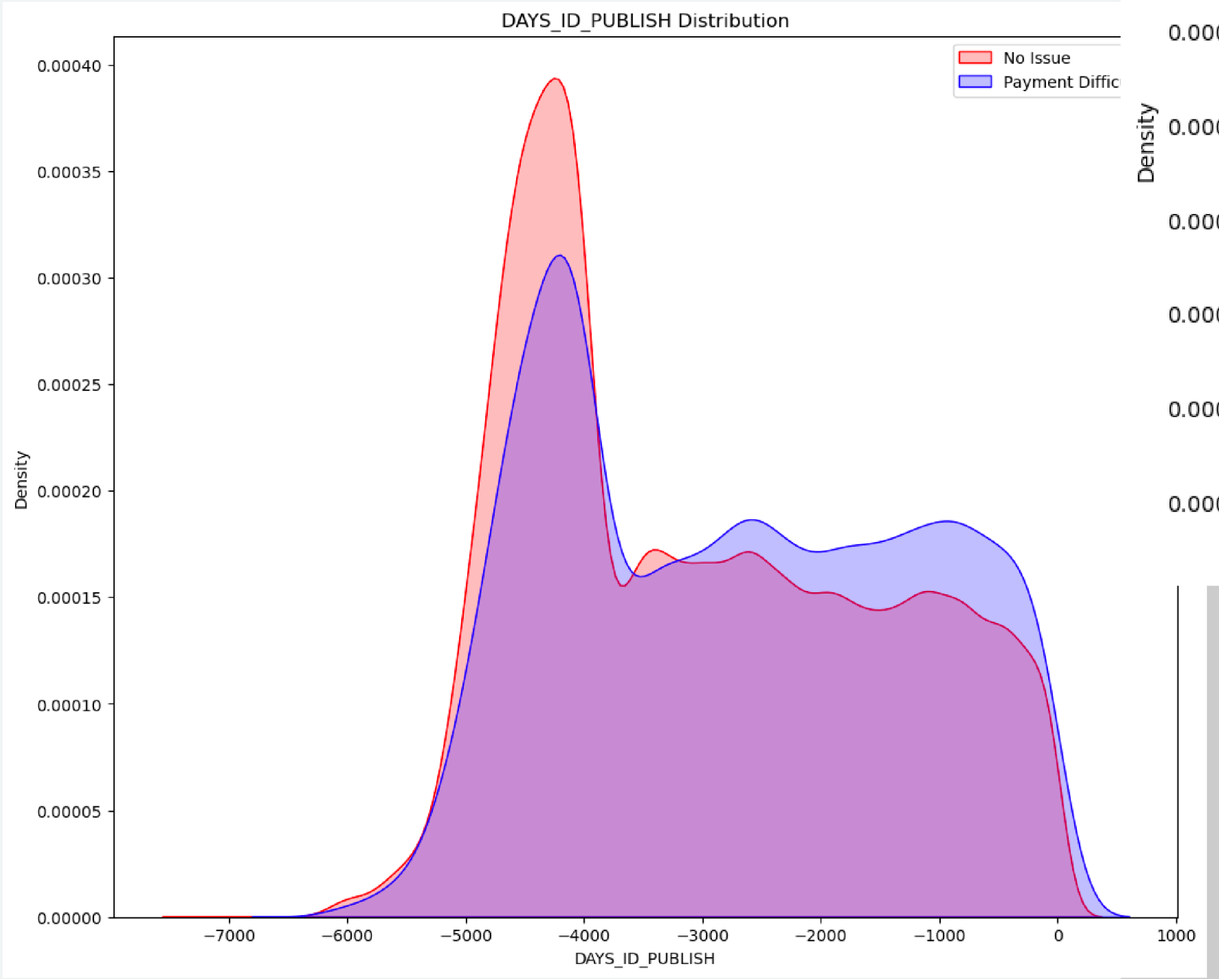
Family features vs target variable

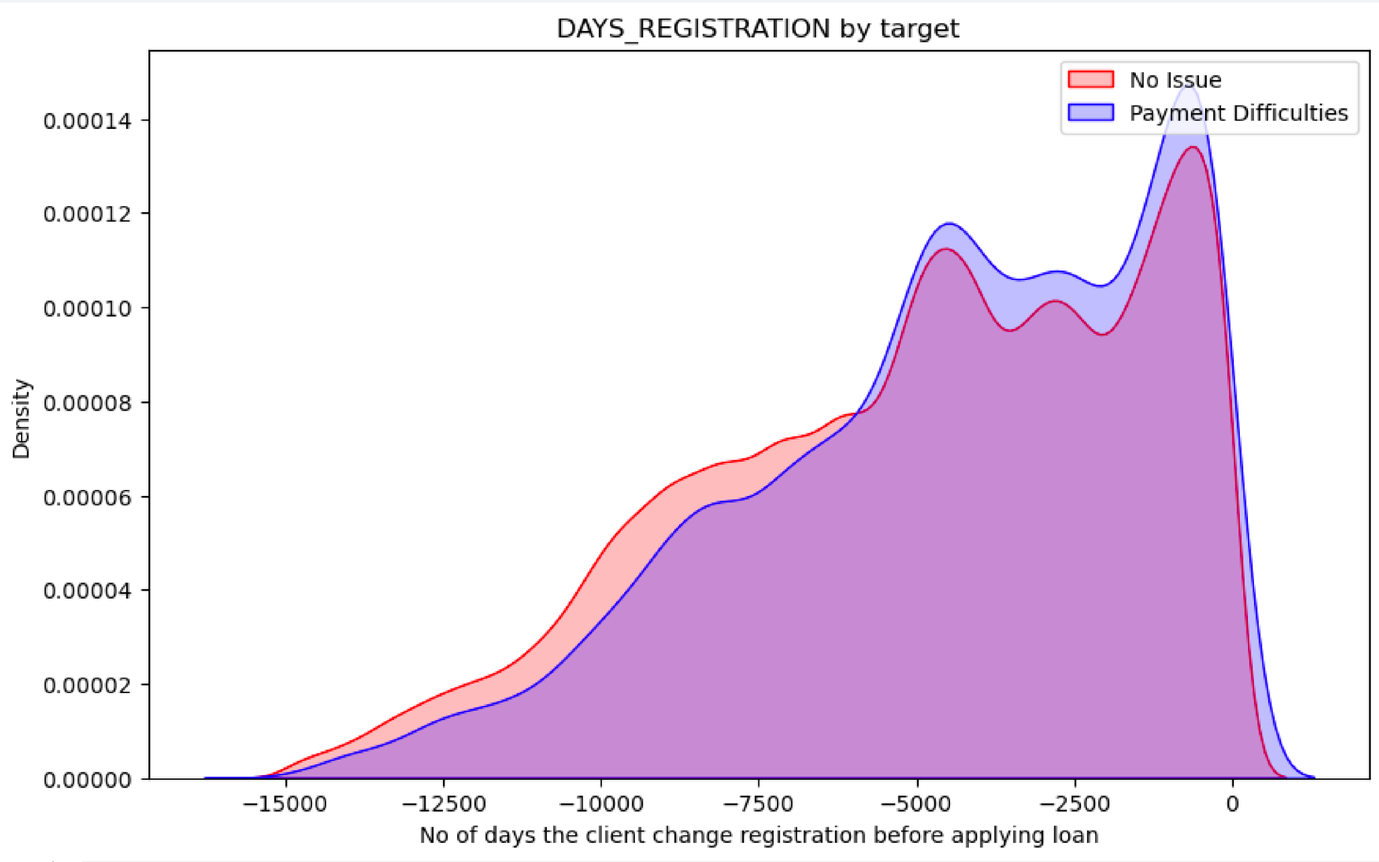




The family of the majority of the applicants are within 0-5 person (including children)

No of days client changed their ID / registration prior to loan application

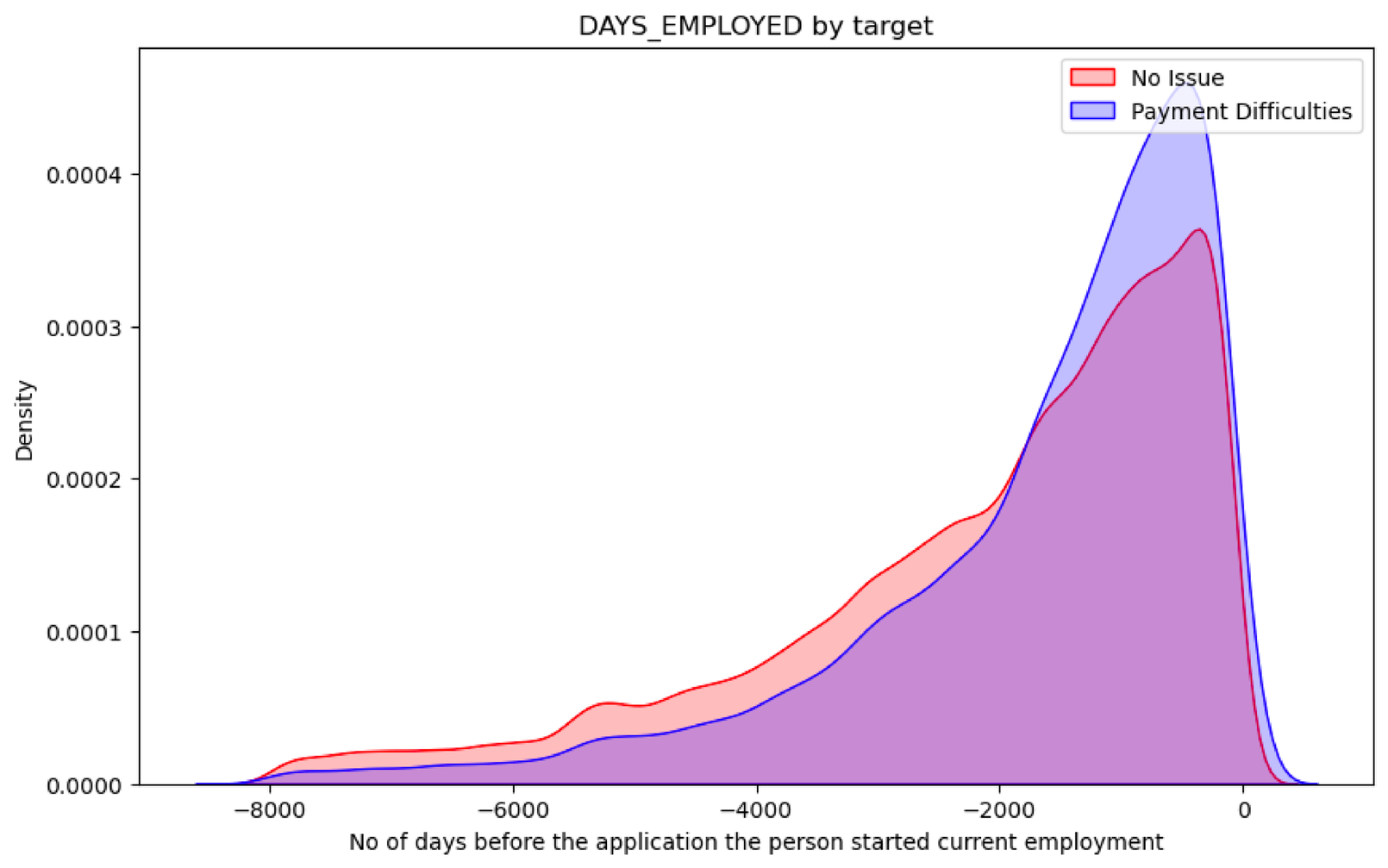




Findings

Larger amount of applicants with payment difficulties were seen for those who changed their ID/ registration fewer days prior to their loan application.

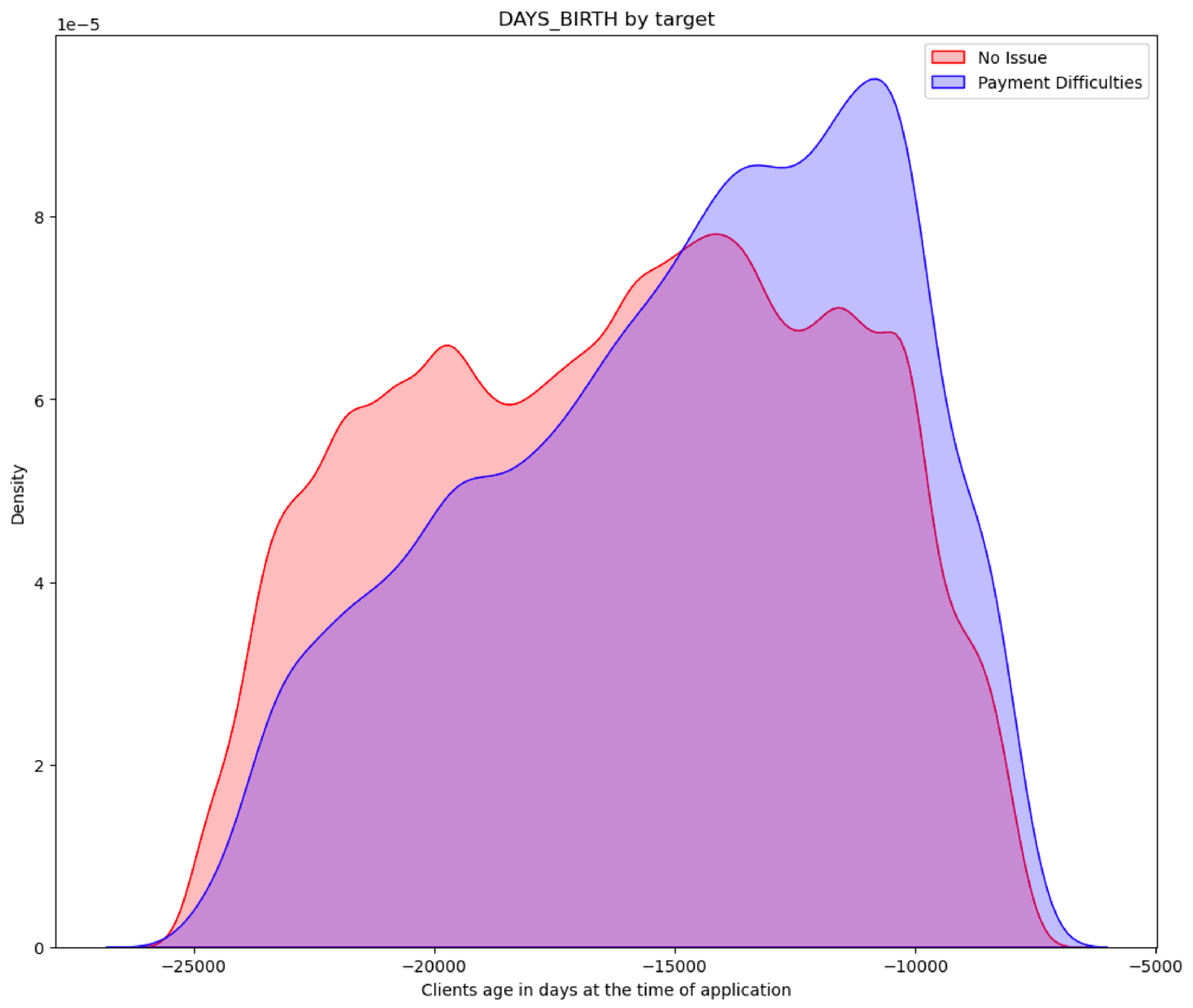
No of days client started their current employment prior to loan application



Findings

Larger amount of applicants with payment difficulties were seen for those who started their current employment fewer days prior to their loan application.

Client's age at the time of loan application



Findings

- Larger amount of applicants with payment difficulties were seen for those who are younger in age at the time of application
- We have seen a larger amount of older applications with no payment difficulties as compared to younger applicants

BIVARIATE ANALYSIS

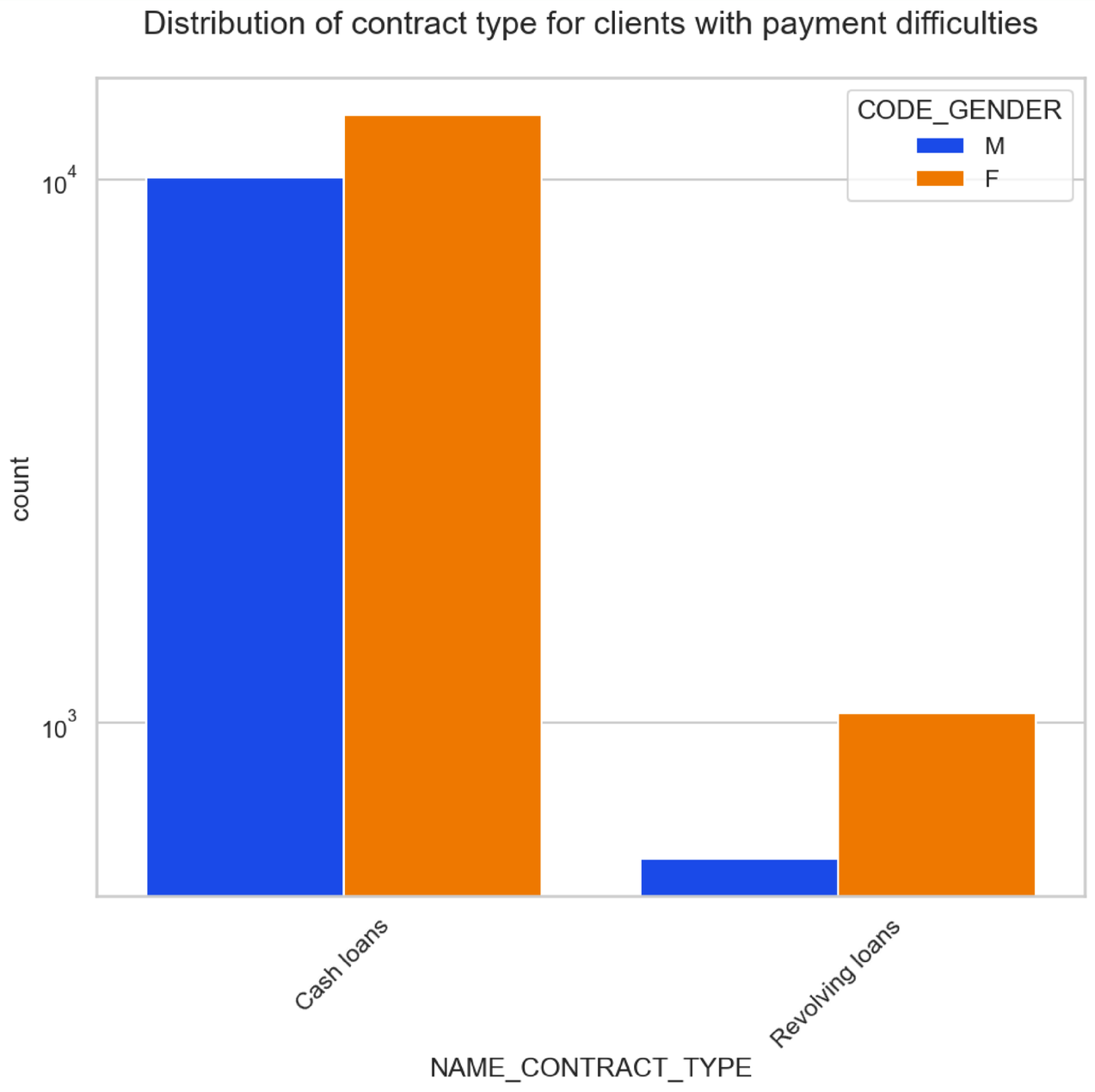
Findings

- Kindly refer to the next few slides for the findings for certain categorical data vs clients with payment difficulties. As per previous section, we only take the top 5 analysis.

- According to our analysis, it has shown that females are mostly the ones with payment difficulties as compared to males from the graphs extracted.

- We focused on analysing clients with payment difficulties as we wanted to understand which groups are the ones that have issues repaying their loan and other characteristics associated to them.

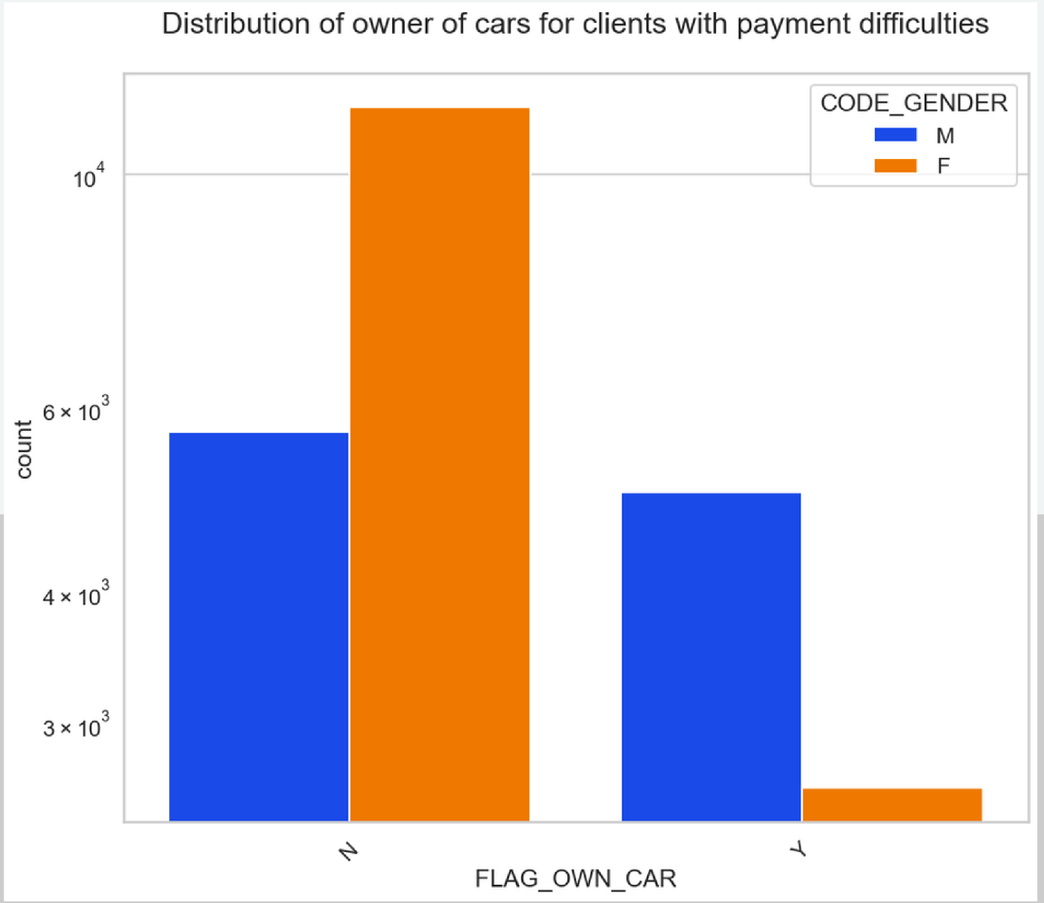
Contract type vs gender (payment difficulties)



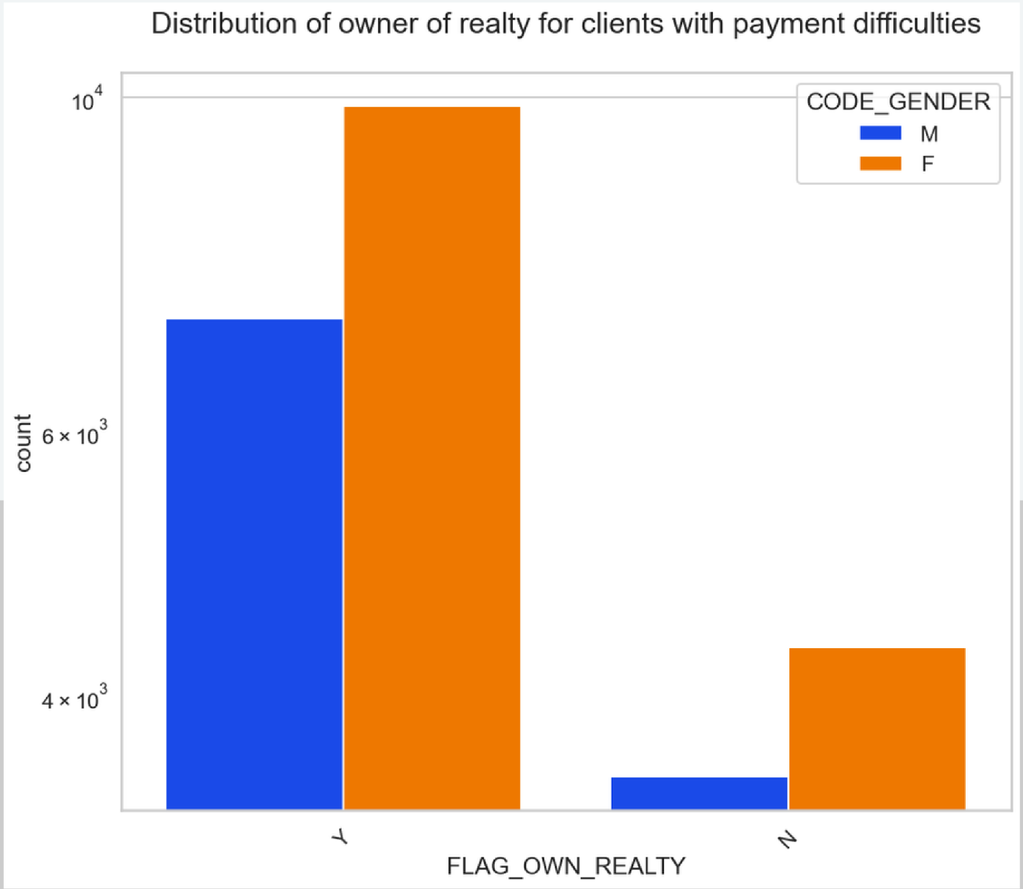
Findings

- Most applicants with payment difficulties applied for cash loans as compared to revolving loans
- As seen from the bar chart on the left, females are more likely to end up having payment difficulties as compared to males regardless of the type of loans they applied to.

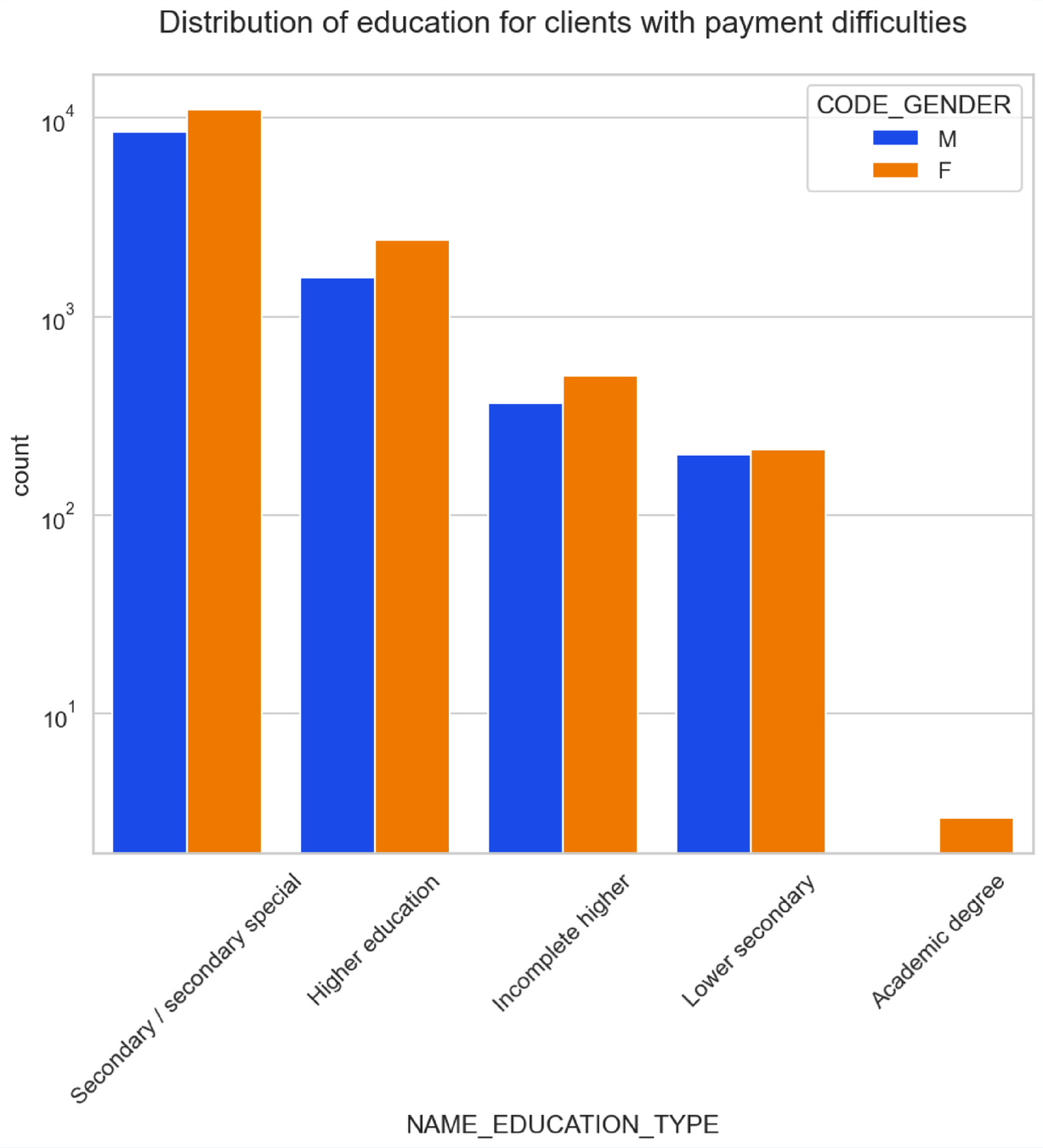
Owner of car/realty vs gender (payment difficulties)



As compared to males, most female applicants that have payment difficulties are non-car owners.



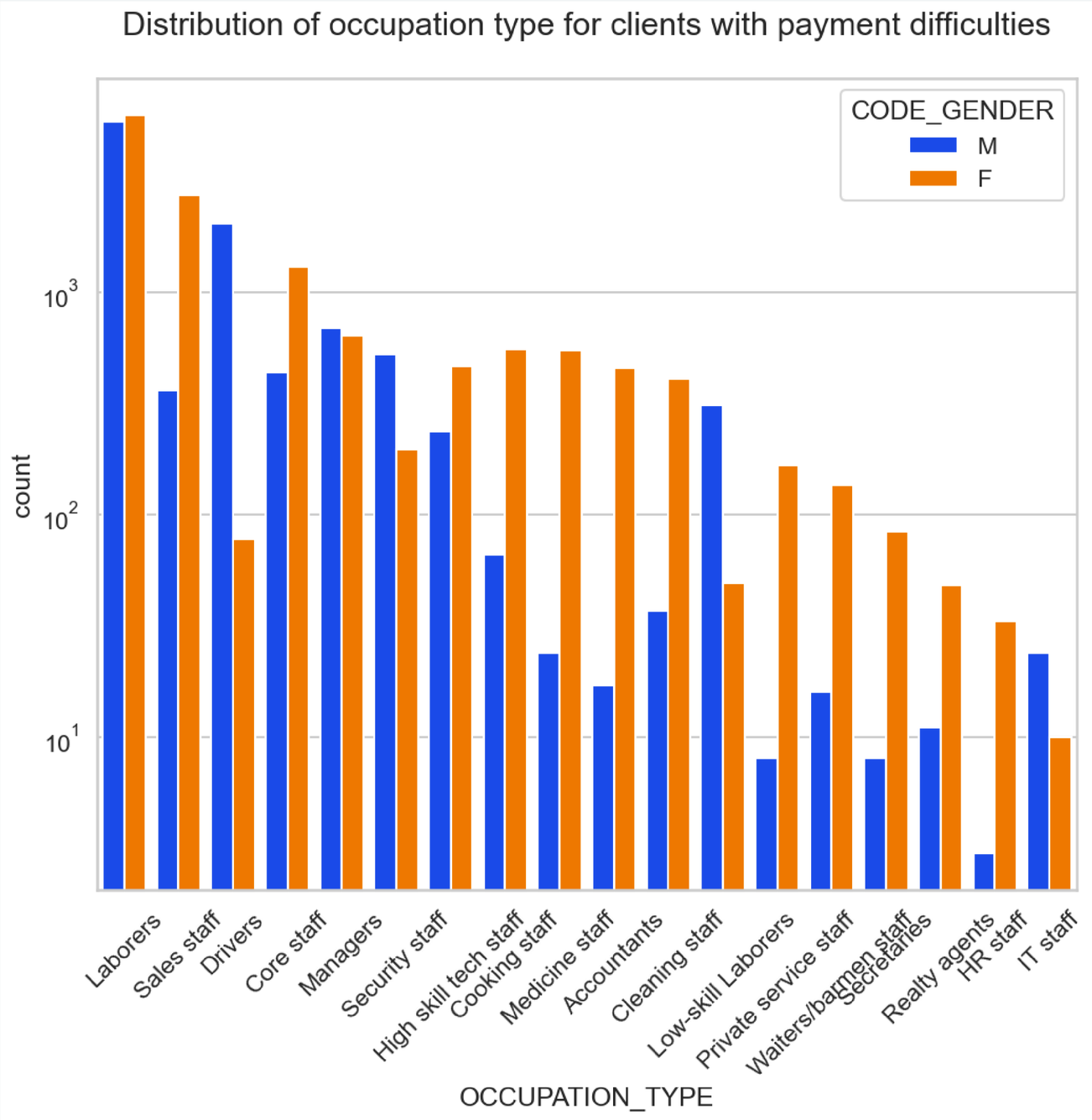
However, for applicants that own a realty, there are more females with payment difficulties as compared to male applicants.



Education type vs gender (payment difficulties)

Findings

- The majority of applicants with payment difficulties completed secondary school as their highest education achieved.
- The number of female applicants with payment difficulties are slightly higher than male applicants for all categories as seen in the bar chart on the left.



Occupation type vs gender (payment difficulties)

Findings

- The majority of applicants with payment difficulties are labourers.
- The number of female applicants with payment difficulties are higher than male applicants for all categories except for drivers, security staff, cleaning staff and IT staff.

SUMMARY APPLICATION DATA

APPLICATION DATA

- More users applied for cash loans as compared to revolving loans.
- More females submitted their loan application than males.
- Most applicants went to apply for loan unaccompanied by any other person, attended secondary school as their highest level education and work as labourers.
- Most applicants come from families of around 5 family members.
- More applicants with payment difficulties were seen for those who changed their ID/ registration, started their current job fewer days prior to their loan applications.
- Younger clients are seen to have payment difficulties as compared to older clients.
- As our our bivariate analysis analysing clients with payment difficulties using gender, females are more likely to be having payment difficulties than males.

THANK YOU