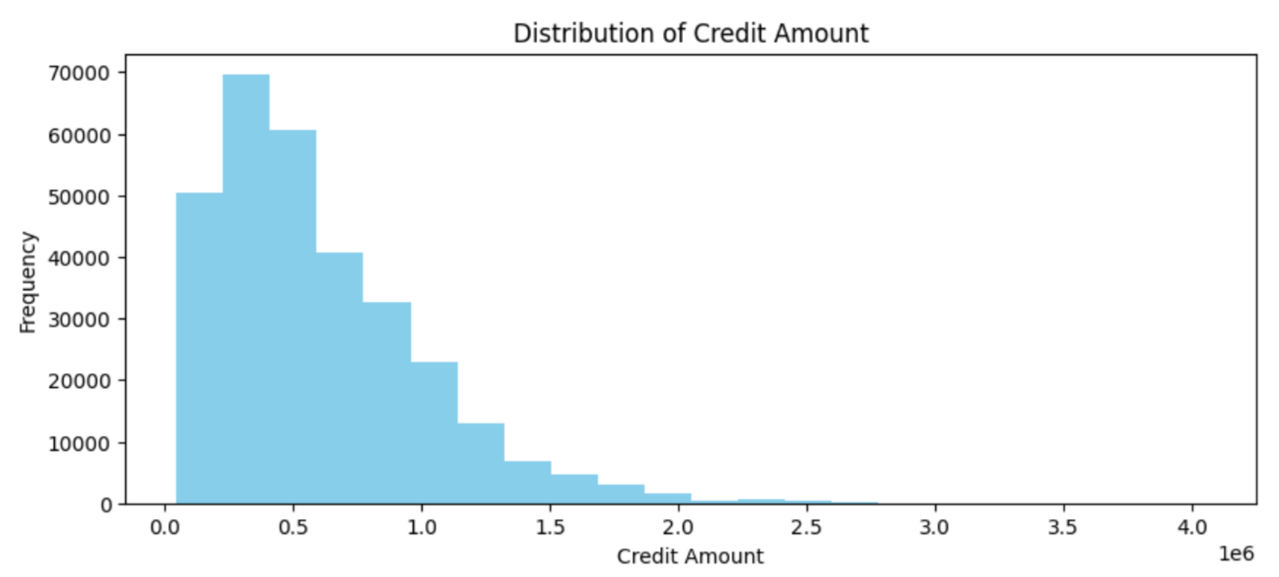
Credit EDA Case Study

Analysis of Application Dataset

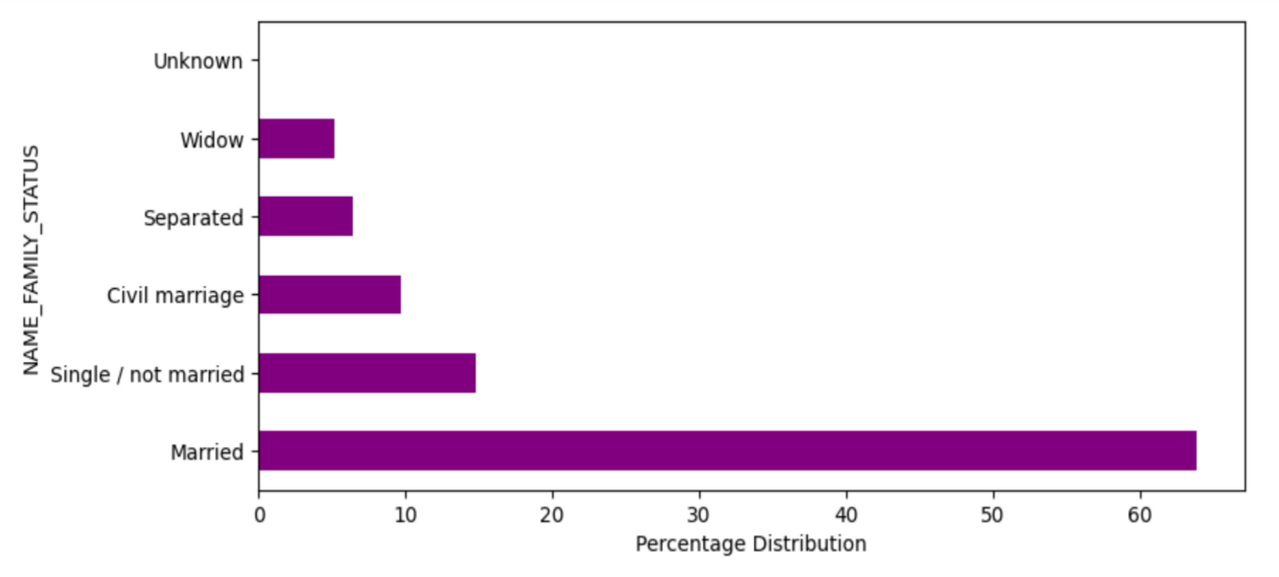
Analysis of frequency of credit amount



***Insights**:*

- Credit loan amount given to most clients is ranging from 2.5 lakhs to 6 lakhs, very few clients have been given credit amount from 15 lakhs onwards.

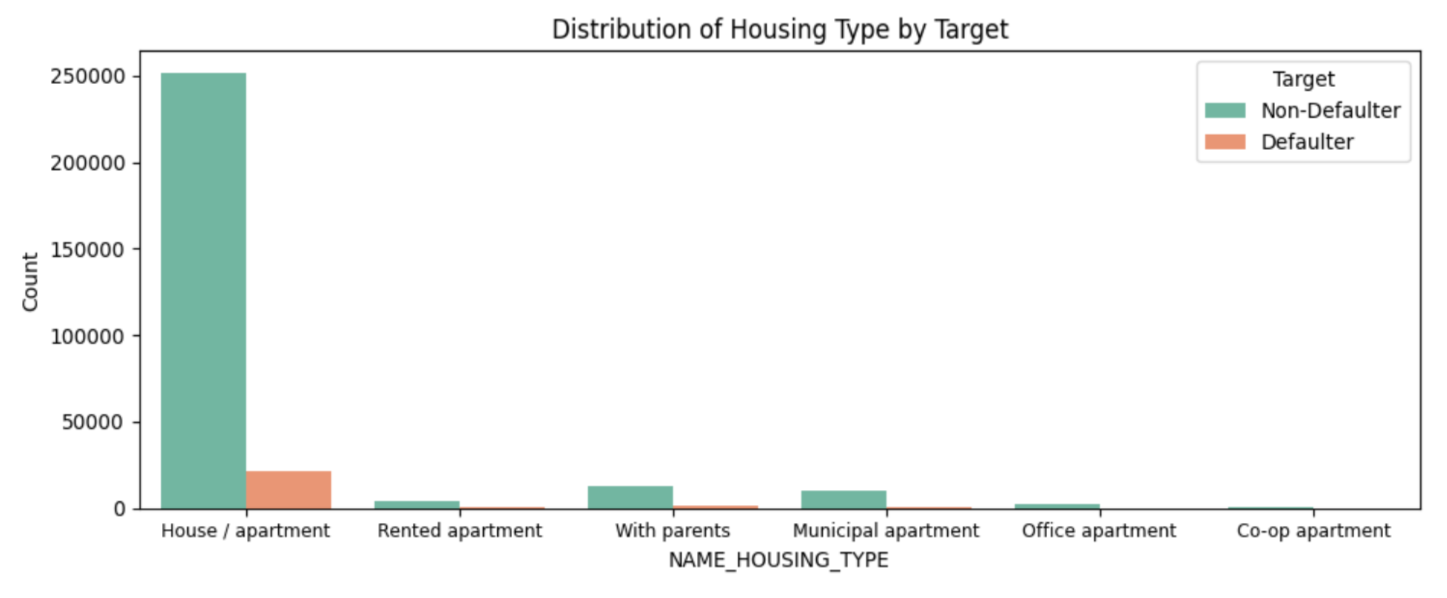
Analysis of percentage distribution of family status



Insights

- Above 60% clients in our dataset are married, around 15% are single, 10% are having civil marriage and then below that are separated and widow.

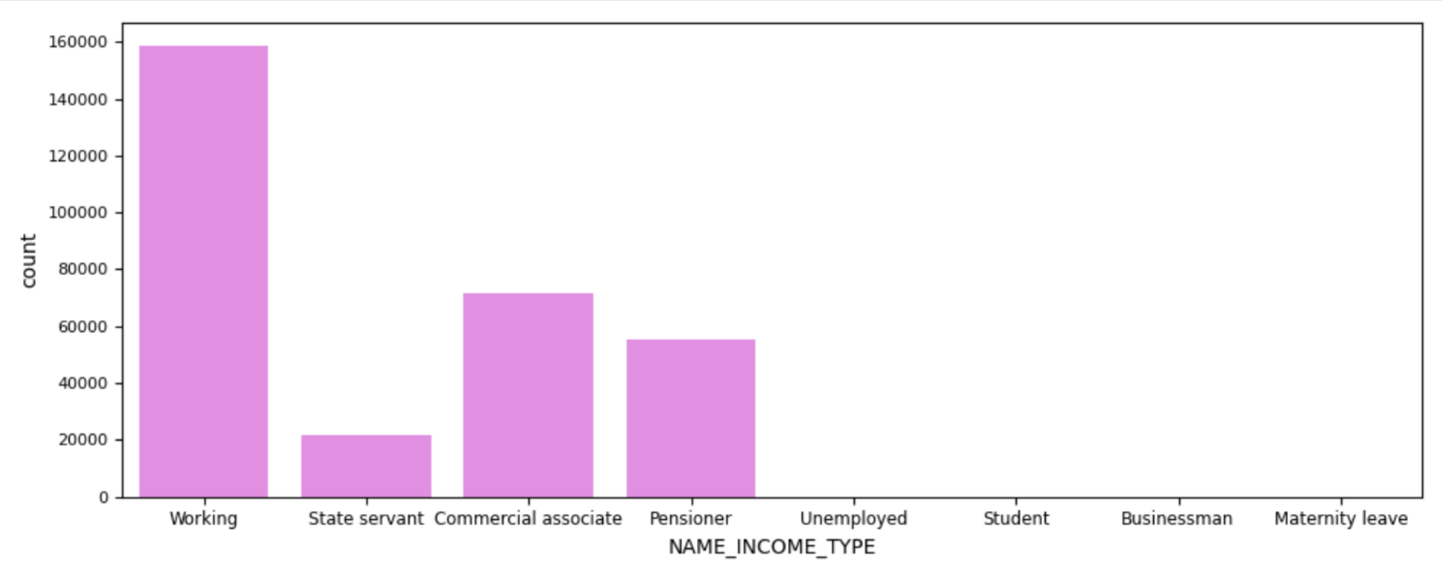
Analysis of Housing types across Target segment



Insights

- Most non-defaulters are the ones with own house/apartment, which is also the case for defaulters, compared to other housing situations.
- Other than House/apartment type, living with parents or in a municipal apartment is more common among non-defaulters compared to defaulters.
- Overall for each housing situation, clients with payment difficulties are less compared to one who pay on time.

Analysis of clients income types



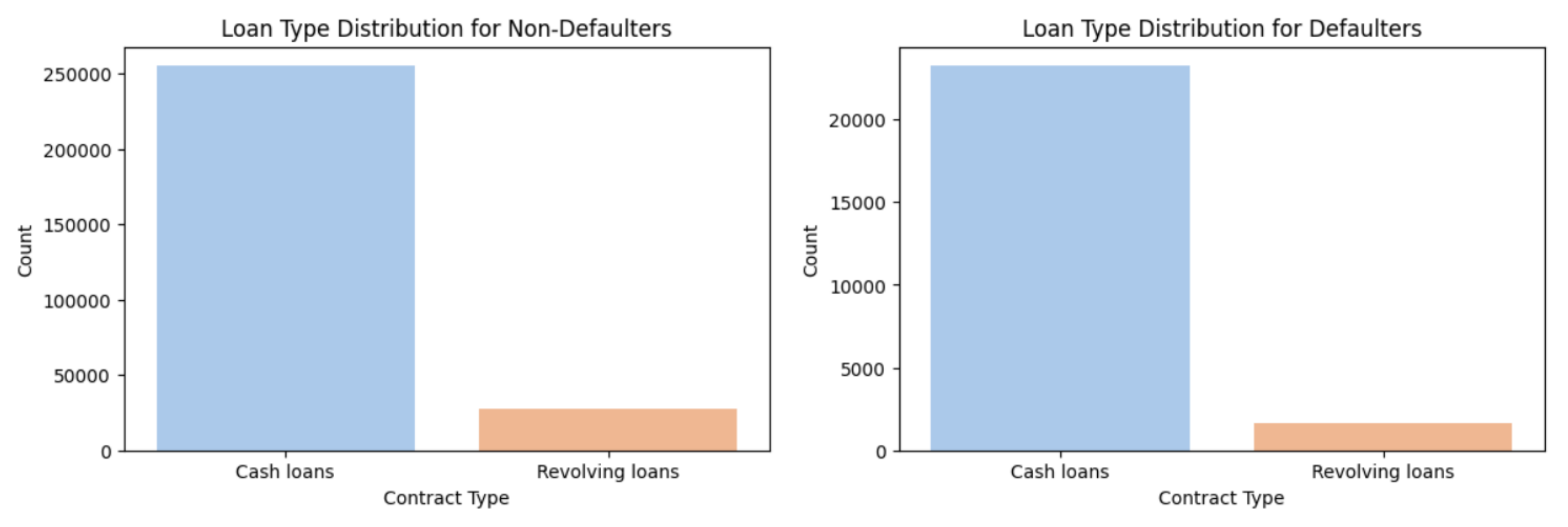
Insights

- Around 1.6L clients have 'Working' background, above 70 thousand have 'Commercial associate' background, 60K clients in our dataset are 'Pensioners', there are also our 20K from 'State servant' background.

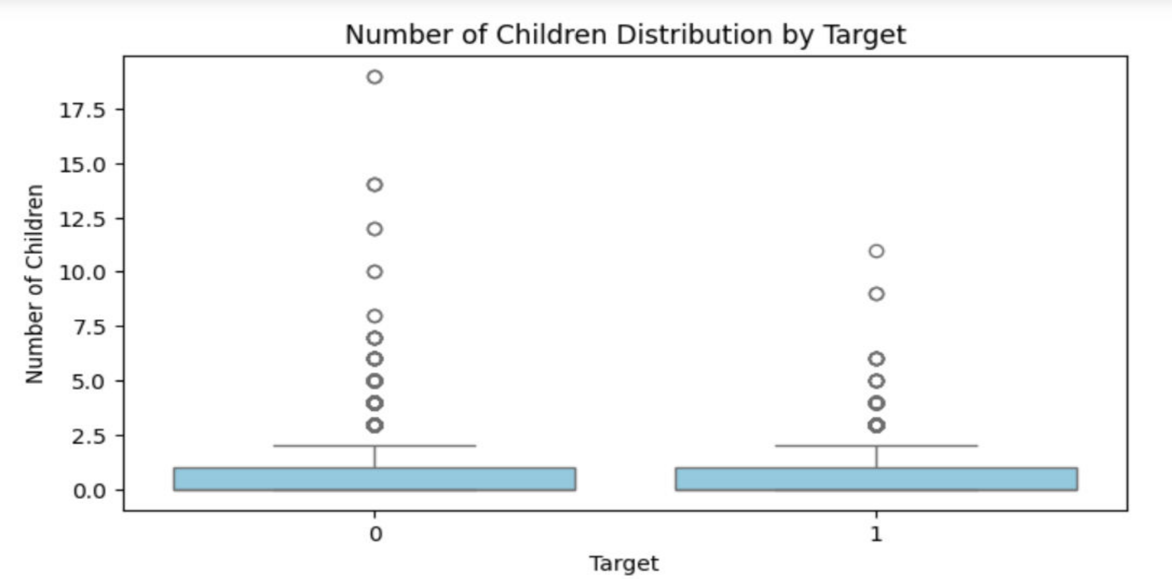
Analysis of types of loan across target segments

Insights

- The majority of loans are Cash loans in both categories (TARGET = 0 and TARGET = 1).The proportion of Cash loans is higher in the category with TARGET = 0, i.e. non-defaulter, compared to the category with TARGET = 1(clients with payment difficulties).
- Both defaulters and non-defaulters show a preference for cash loans over revolving loans.



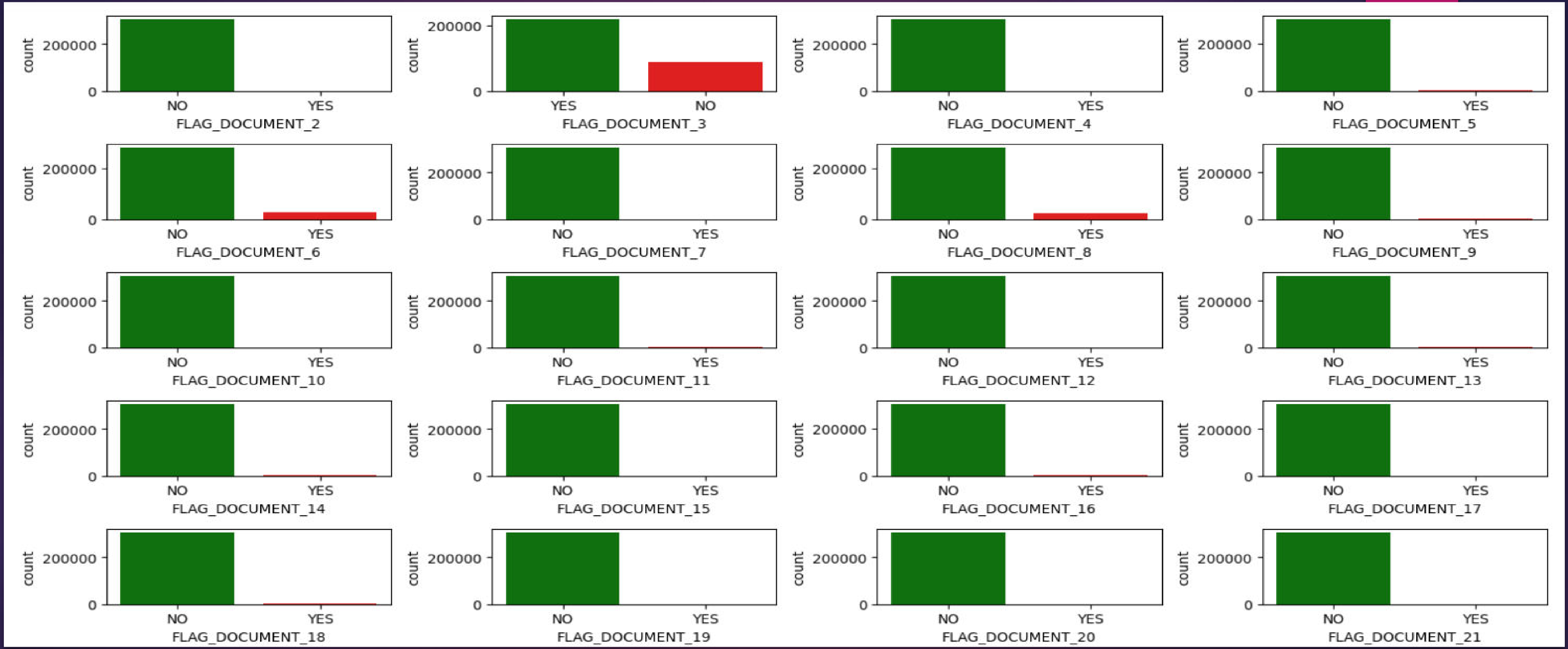
Segment-Wise Child Distribution Analysis



Insights

- For both types of clients, i.e., defaulters(Target==1) and non-defaulters(TARGET==0), we can observe that 75% of clients of our dataset are having at most 1 child. However, there are some outliers with larger numbers of children, reaching a maximum of 19 children for non-defaulters and 11 children for defaulters.

Analysis of Documents provided by the clients



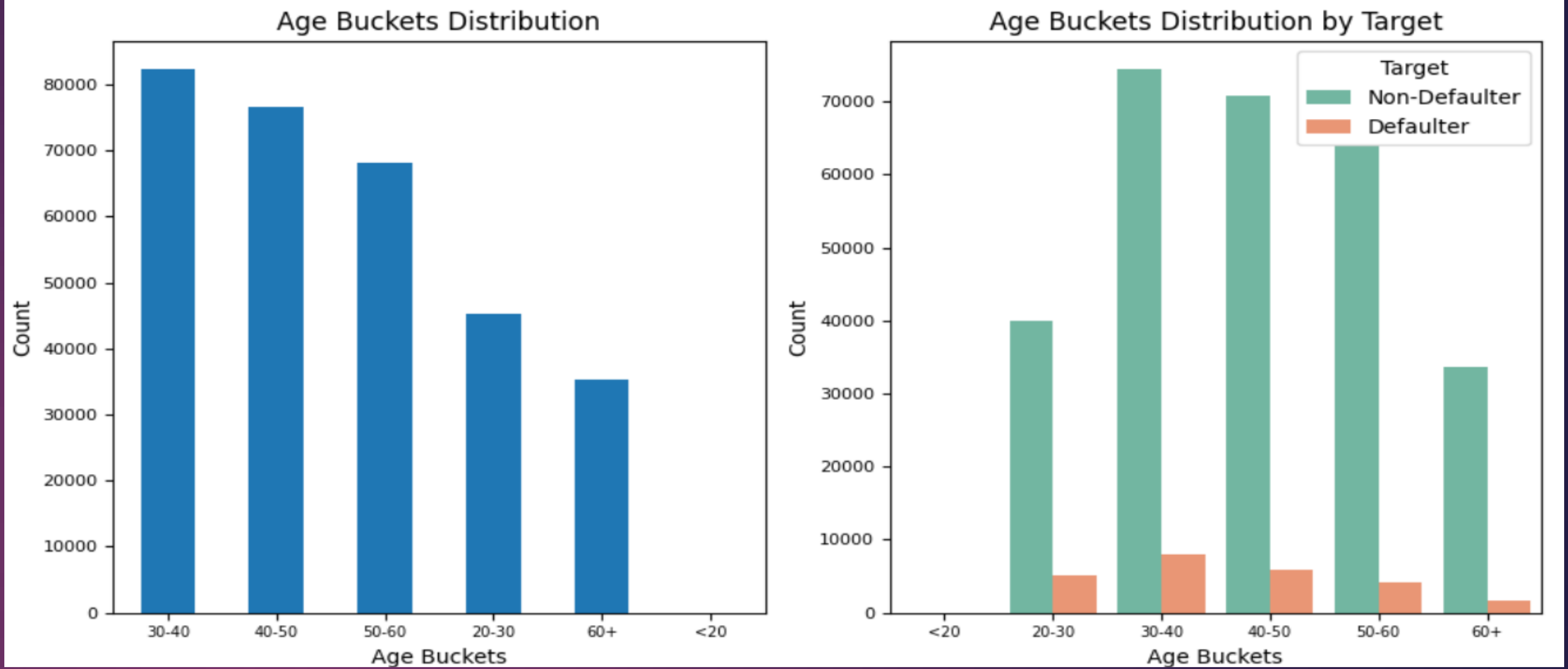
Insights:

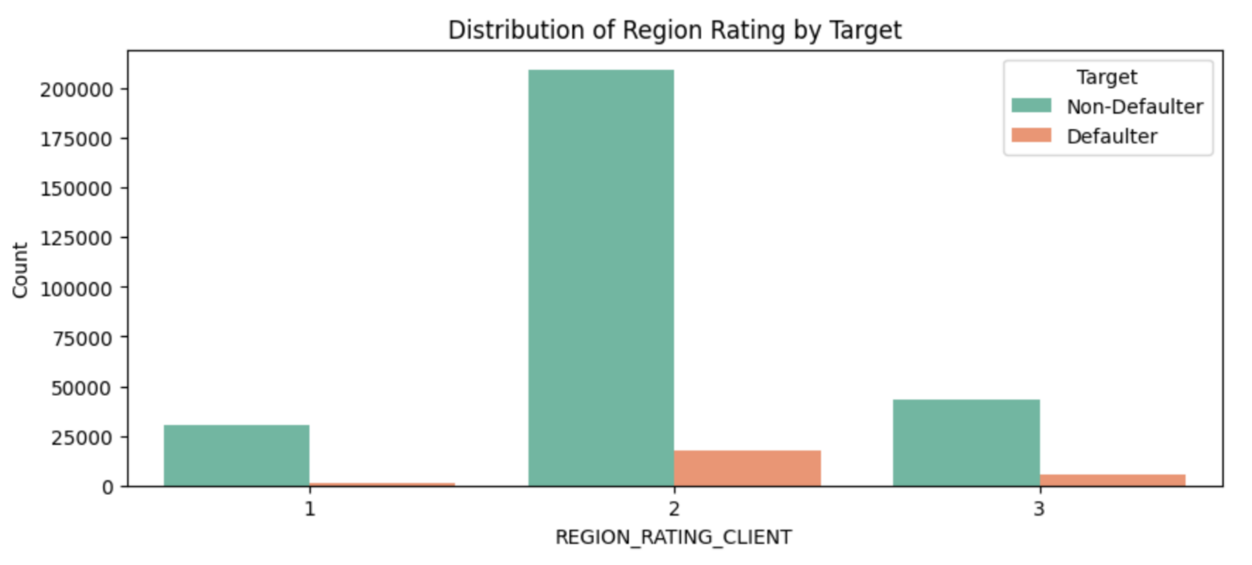
- Out of the total 21 documents, maximum clients have only provided document 3, 6 and 8. Very few of them have provided document 5, 9, 11, 13, 14, 16 and 18.

Age groups count and their analysis with each target segments

Insights

- Despite the smaller representation of clients aged 20 to 30 in our dataset compared to those aged 50 to 60, the former exhibit a higher rate of defaulters. This suggests that individuals in the 20-30 age group are more likely to default on loans compared to those in the 50-60 age group, despite their lower numerical presence.



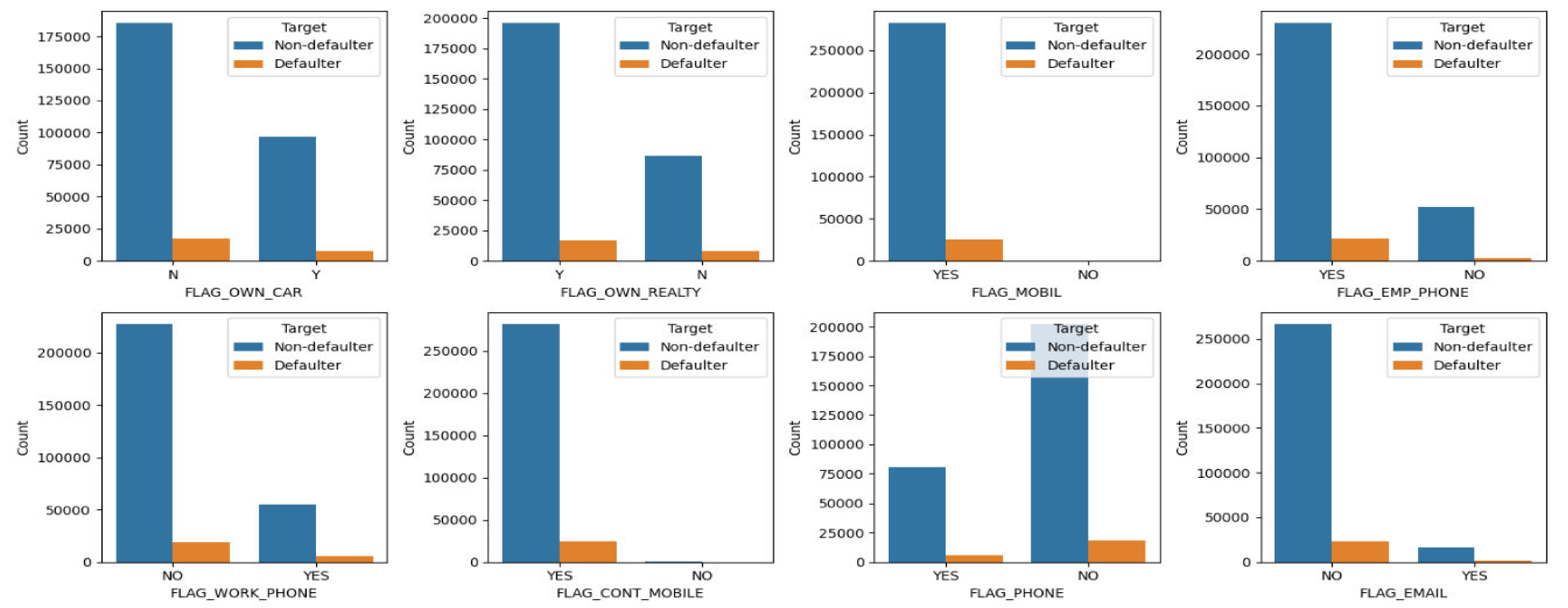


Analysis of the Rated regions

Insights

- The majority of entries in the dataset are from Region, which has a rating of 2 for both defaulters and non-defaulters.
- Despite having fewer data entries, Region, with a rating of 1, demonstrates a more positive non-default behavior, same with region with rating 3.

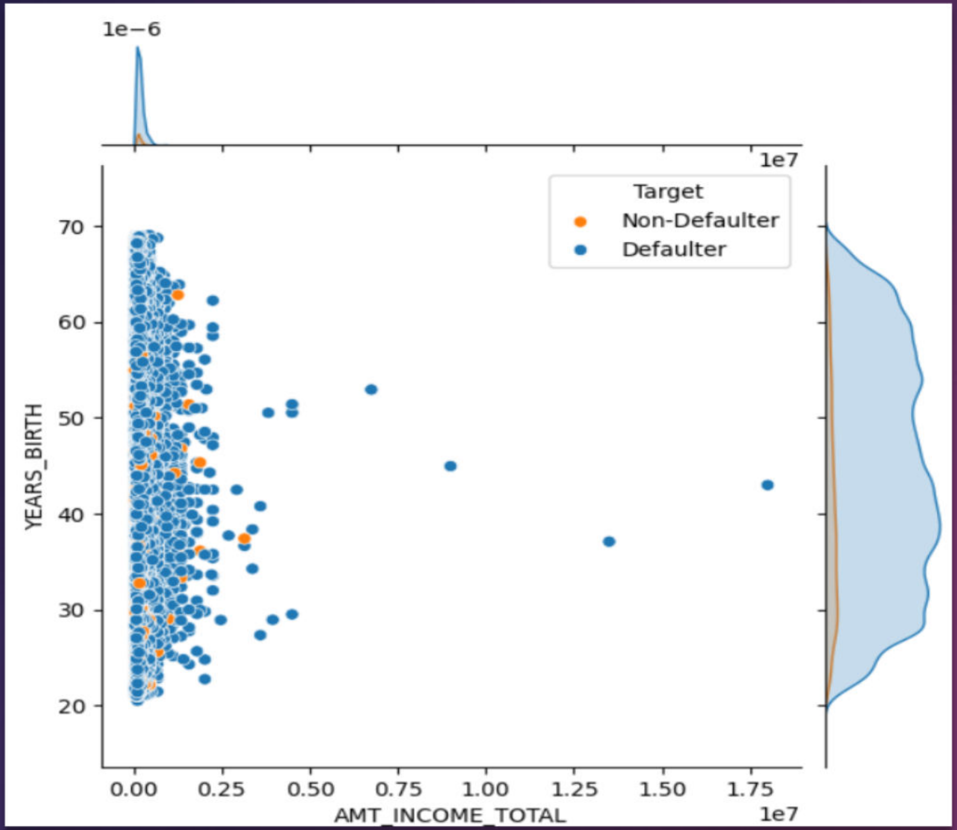
Analyzing personal information provided by the clients w.r.t Target variable



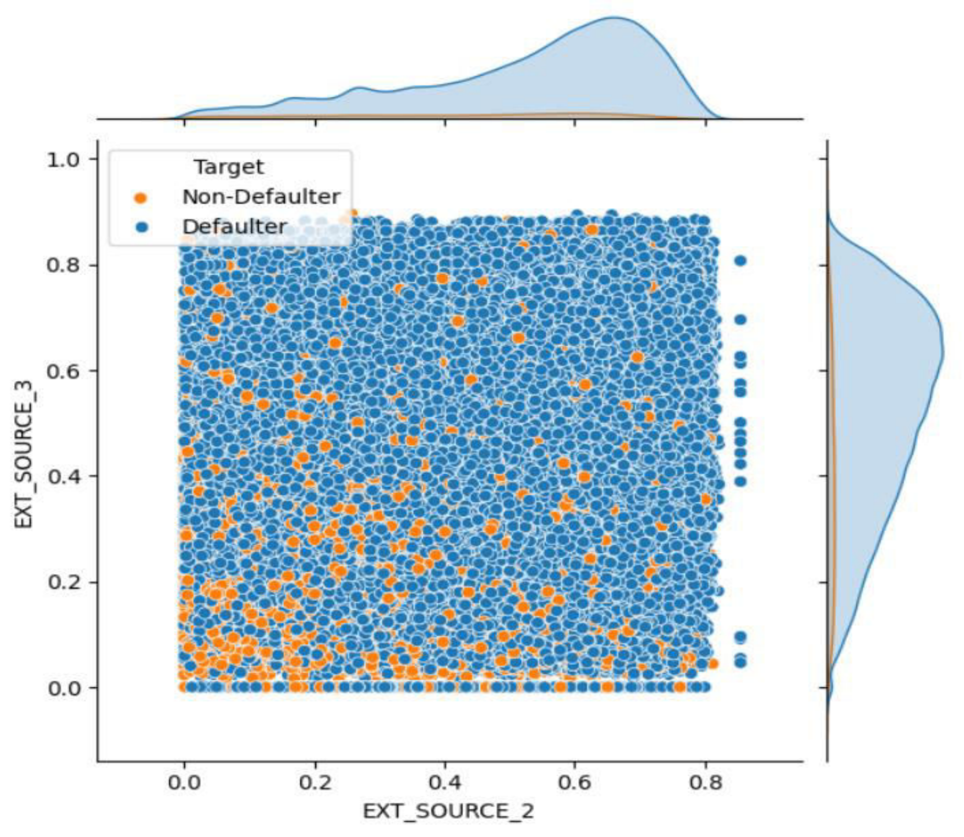
Insights

- Be it defaulter or non-defaulter, all have provided mobile phone(FLAG_MOBIL) and it was reachable(FLAG_CONT_MOBILE)
- Defaulter count seems higher for those who own realy, than those who don't. Whereas it is opposite for the ones who own a car, defaulters counts is greater for those who dont own a car than the ones who do.
- Very few of the clients have provided email and home phone.

Income vs age relation



External sources credit scores relation



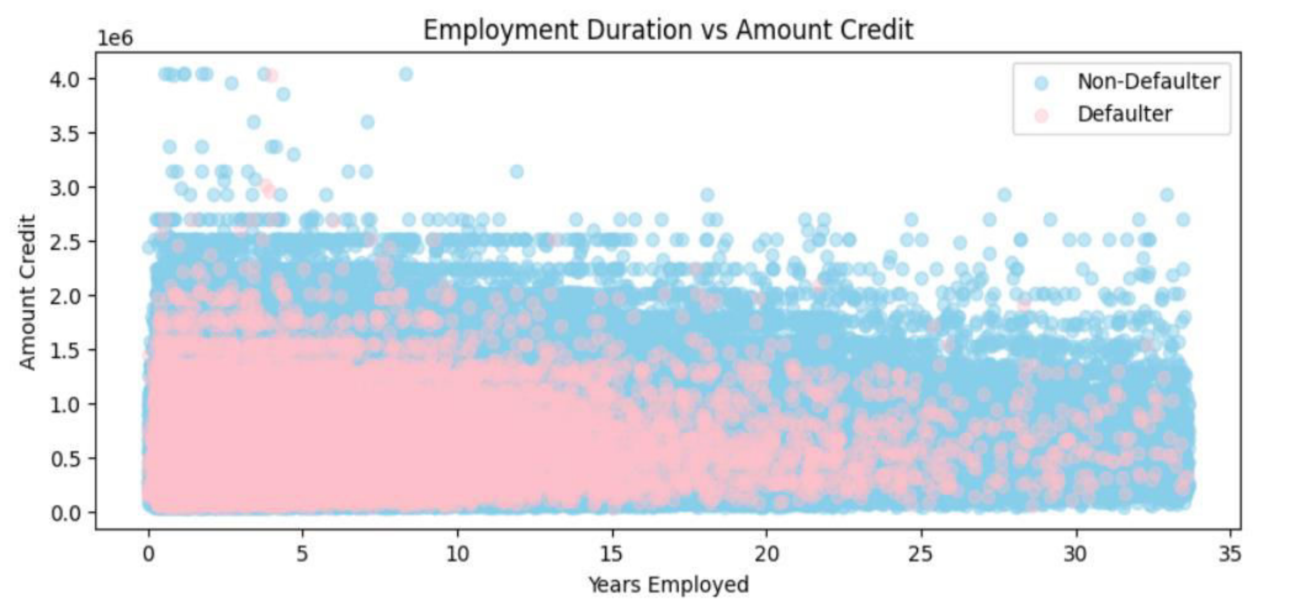
Insights

- Defaulter rate looks just a little bit greater for clients aged between 25 to 40, compared to the ones who are above 55.
- Most clients without payment difficulties (TARGET=0) are clustered at the lower end of the income scale. Those with payment difficulties (TARGET=1) are scattered across the plot, but generally have lower incomes as well.

Insights

- Defaulter rate looks just a little bit greater for clients aged between 25 to 40, compared to the ones who are above 55.
- Most clients without payment difficulties (TARGET=0) are clustered at the lower end of the income scale. Those with payment difficulties (TARGET=1) are scattered across the plot, but generally have lower incomes as well.

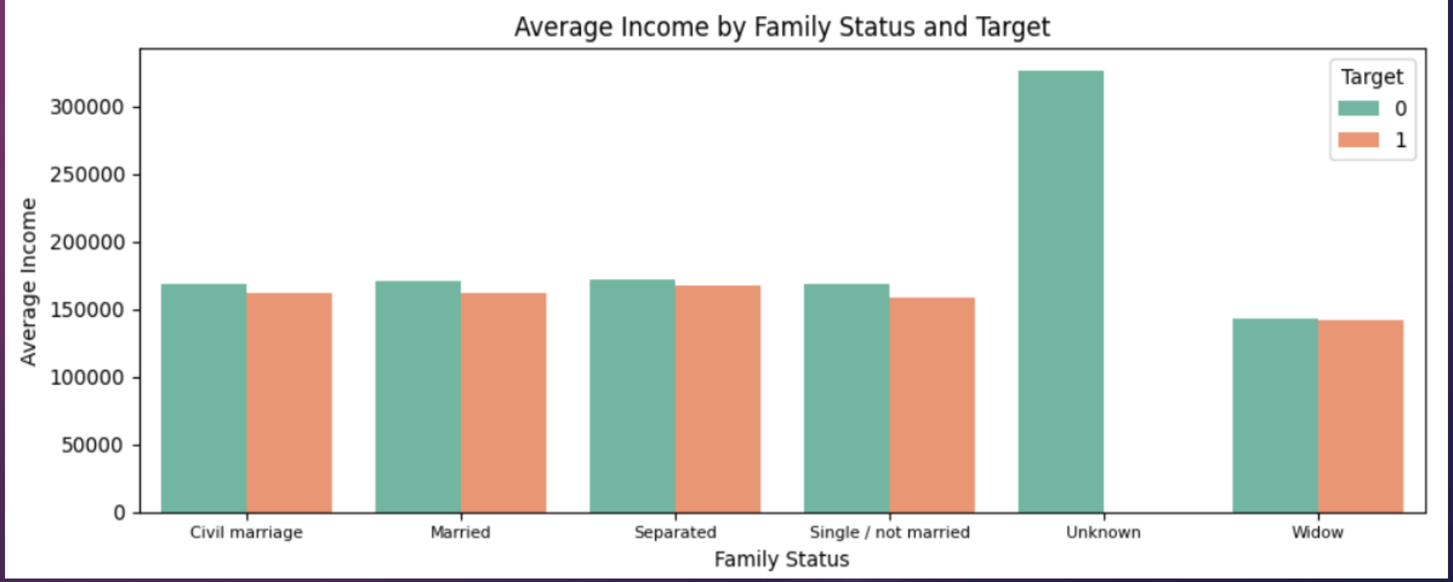
Analysis of employment duration w.r.t credit amount



Insights

- Defaulters count are less compared to non-defaulters for employment years above 20 and even lesser for employment above 30
- Defaulters are also less frequent compared to non-defaulters for early employment years but for higher credit amounts.

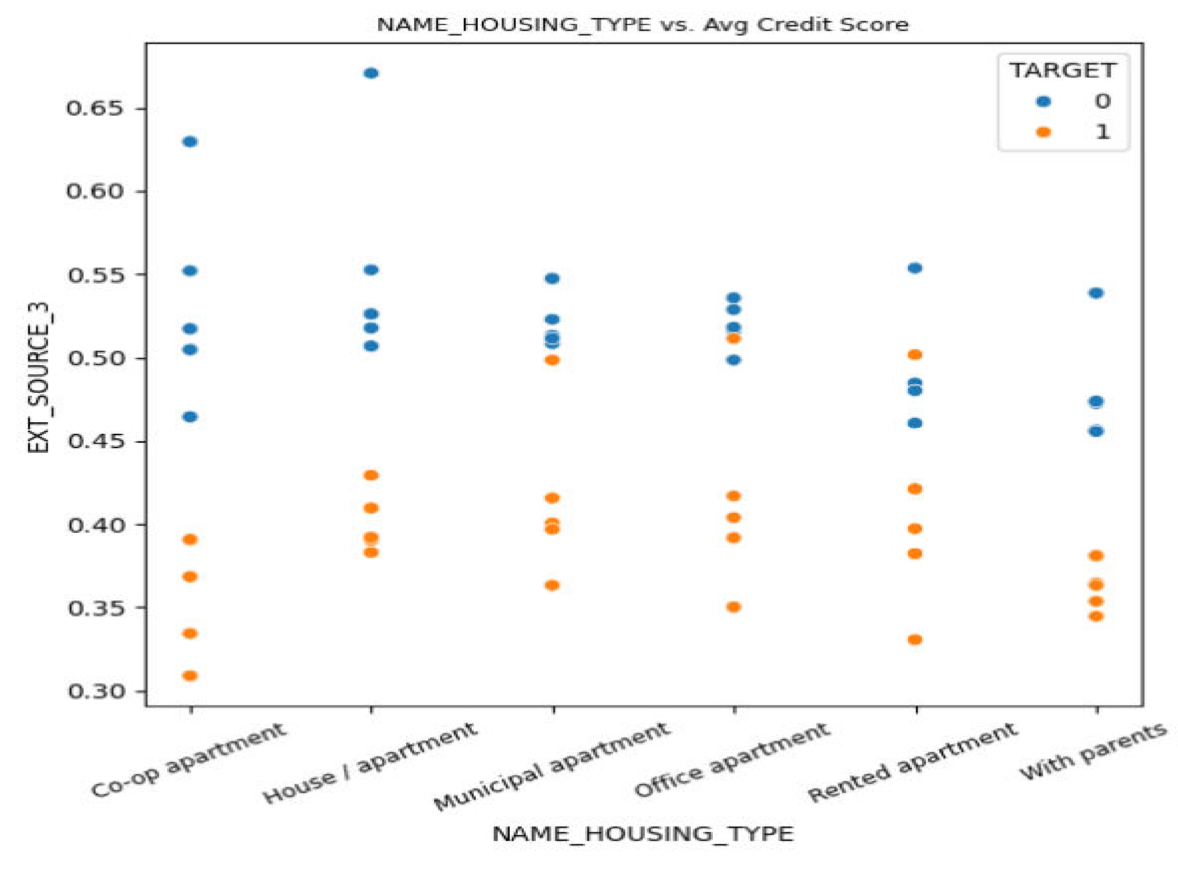
Analyzing average income of different Family status



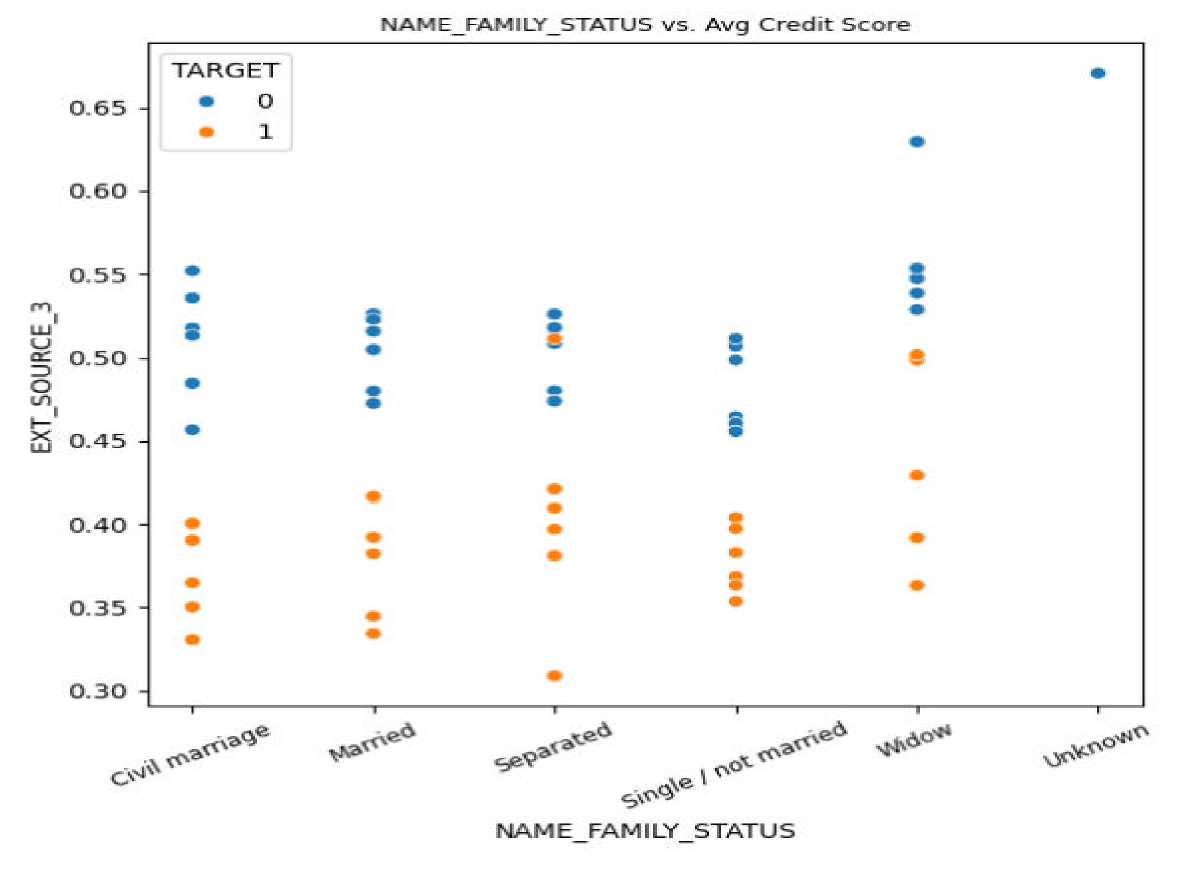
Insights

- Across all family statuses, individuals without payment difficulties(TARGET = 0) tend to have higher average incomes than those with payment difficulties(TARGET = 1), although the difference is very small.
- The difference in average income between those with and without payment difficulties is smallest among Widowed individuals, they also have the lowest average incomes.
- Clients under 'Unknown' category don't seem to have any defaulters and they have highest average income.

Housing Type Influence on Credit Scores



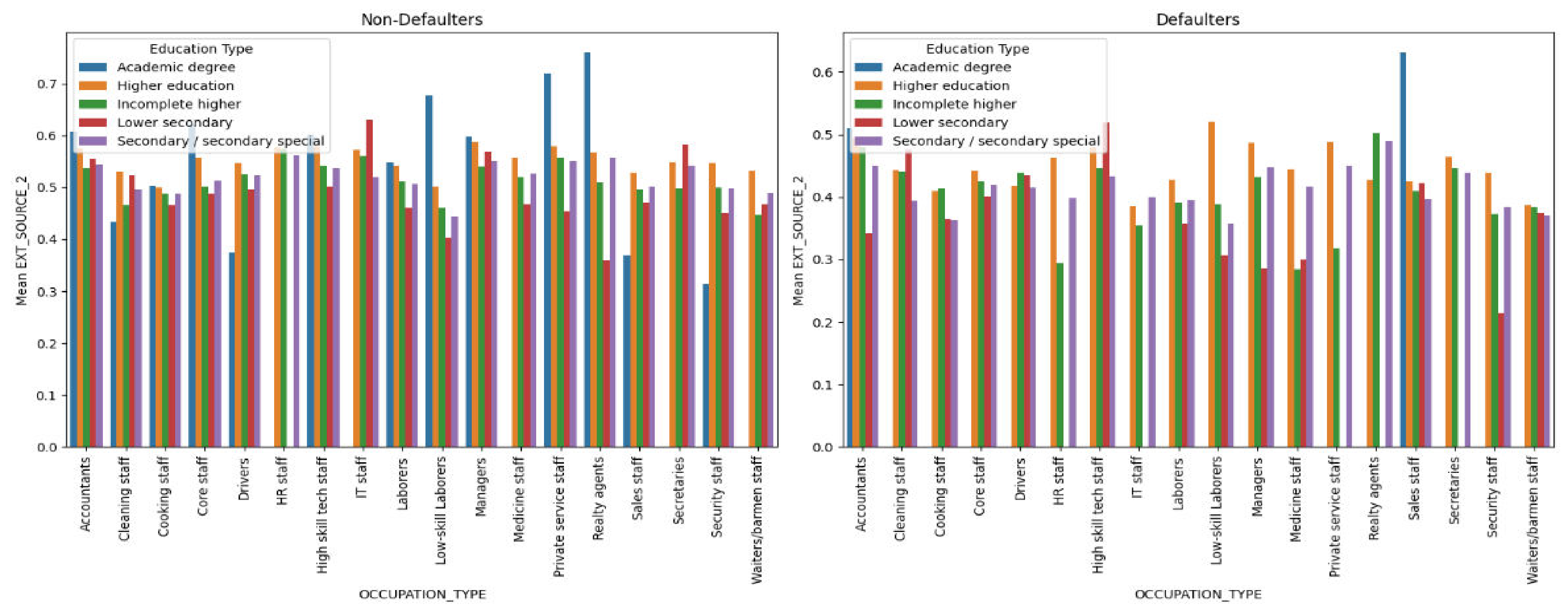
Family Status influence on Credit Scores



Insights

- *Non-defaulters (TARGET = 0) tend to have a higher average credit score(EXT_SOURCE_3) compared to defaulters (TARGET = 1) across most housing types and family statuses*
- *Above the average credit score of 0.45, be it client's housing type or their family status, most seem to have no payment difficulties.*

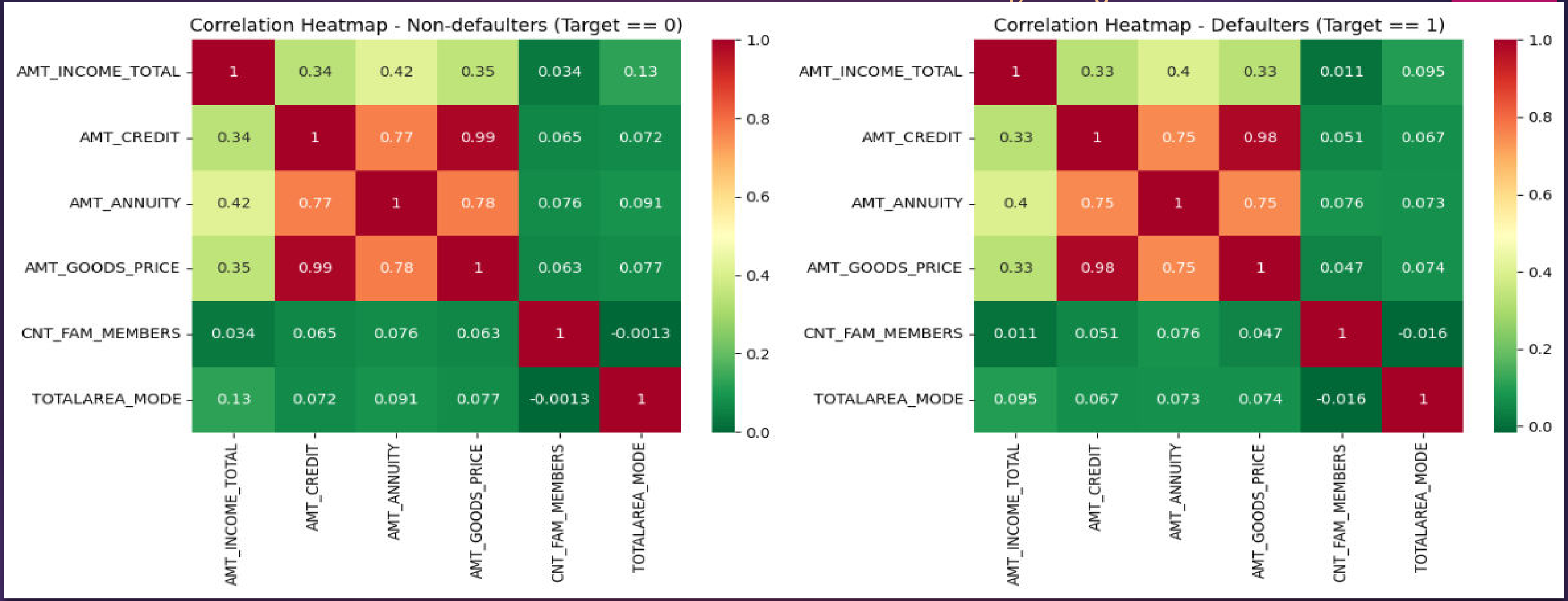
For each target segment, analyzing mean credit scores for different occupations across different education types



Insights:

- Academic Degree seems to be missing from most of the occupations in defaulters plot.
- For the non-defaulters plot, in most occupations, individuals with an Academic Degree tend to have a higher average EXT_SOURCE_2 compared to those with lower education levels.
- Sales staff: In non-defaulters plot, those with an Academic Degree have a lower average EXT_SOURCE_2 compared to other education levels. However, it looks opposite in the defaulters plot.

Correlation between numeric variables across each target segment

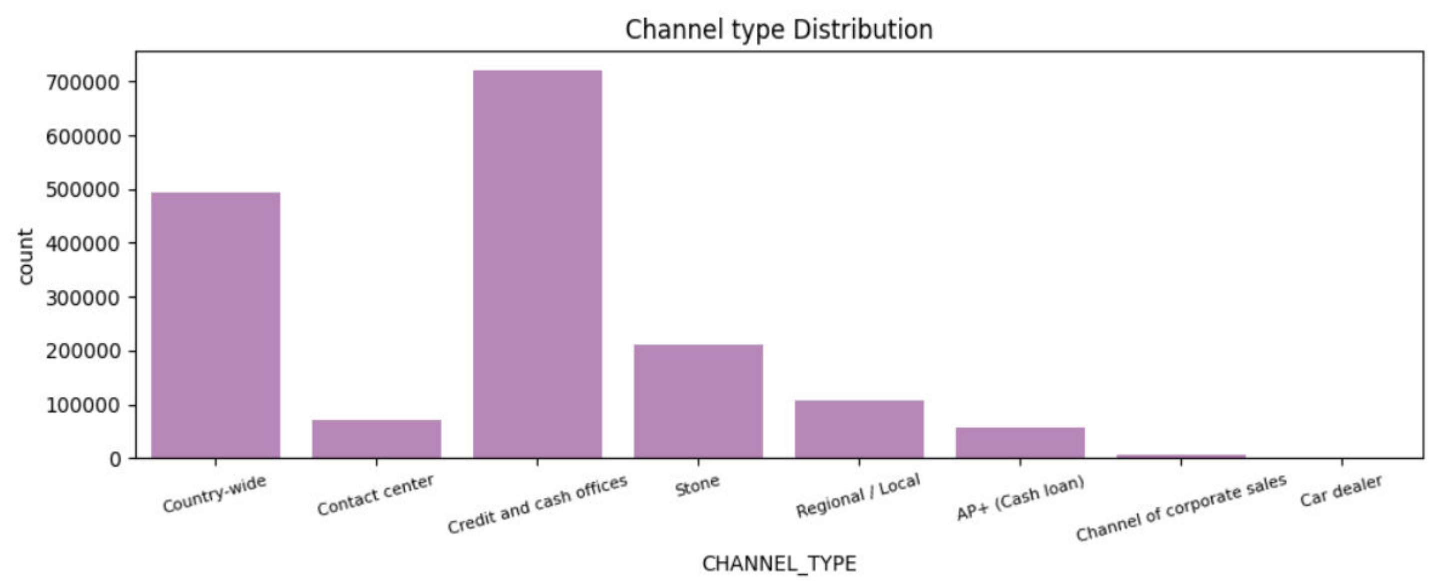


Insights:

- There is a strong positive correlation between AMT_CREDIT and AMT_GOODS_PRICE, as indicated by the dark green color in both heatmaps.
- AMT_ANNUITY also shows a significant positive correlation with both AMT_CREDIT and AMT_GOODS_PRICE
- CNT_FAM_MEMBERS shows a weaker correlation with other variables, implying that number of family members have no relationship with these financial attributes.
- For non-defaulters, there’s a noticeable positive correlation between AMT_INCOME_TOTAL and other monetary attributes like AMT_CREDIT, but this correlation is less pronounced for defaulters.
- Non-defaulters show a very weak positive correlation between income and property size.

Previous Application Dataset Analysis

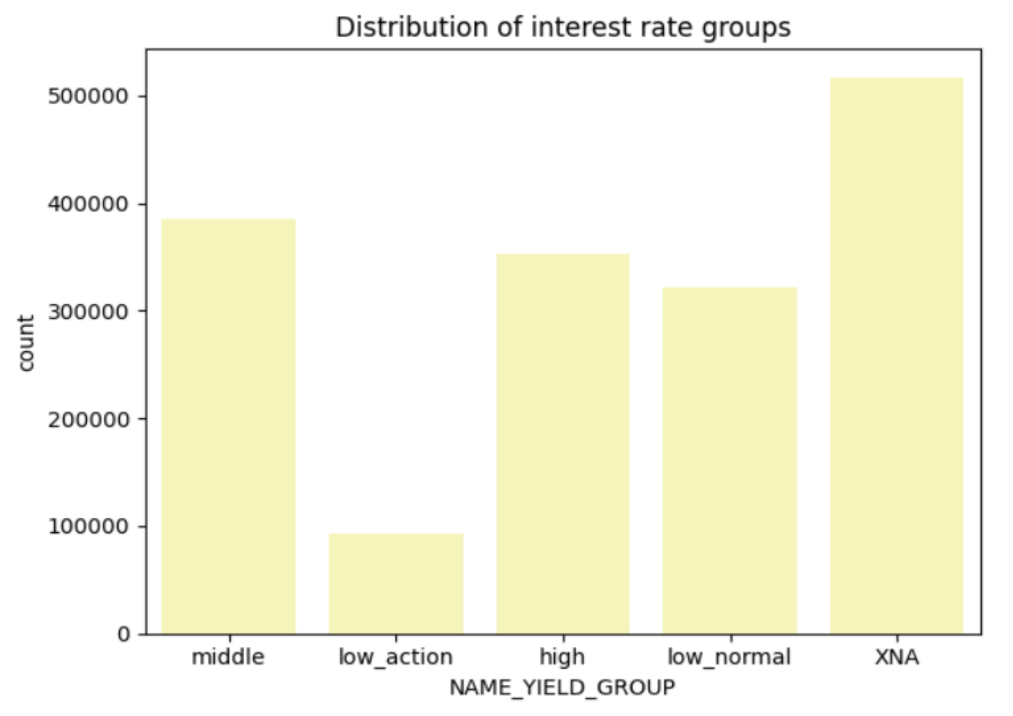
Distribution of different types of channel



Insights

- Maximum number of clients, which is almost 700K customers, applied through "credit and cash offices", and very few, around 450 applied for loan through "car dealer".

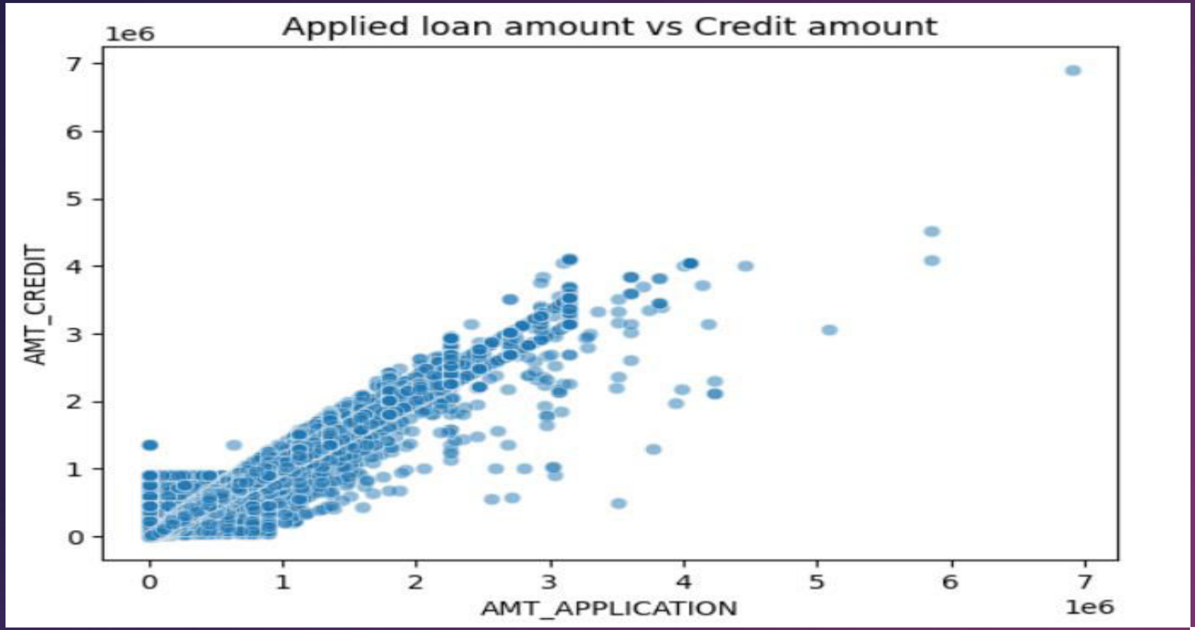
Distribution of interest rate groups



Insights:

- Apart from XNA which probably are representing missing values or unknown data, moderate interest rates (Middle) are the highest and most preferred among clients, depending on the credit amount, and least preferred are the low_action category.

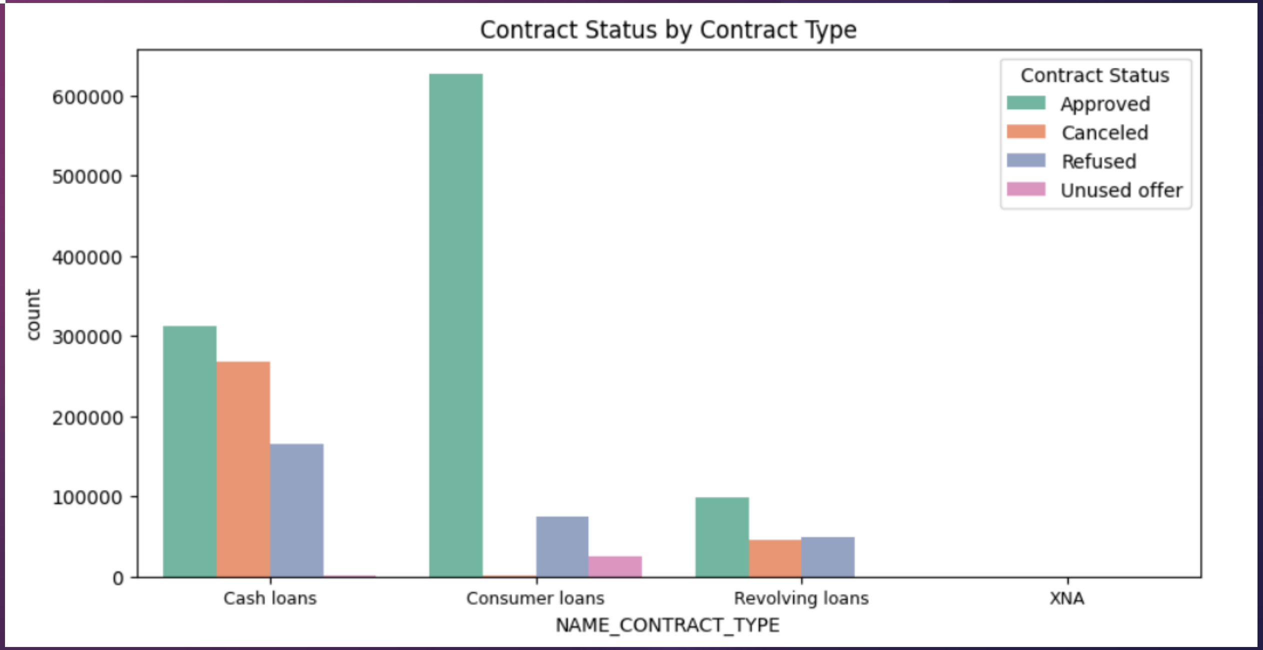
Loan amount applied vs credit amount proceeded



Insights

- Scatter plot shows a positive correlation between both variables, indicating a positive linear relationship, i.e. as the amount requested in the loan application increases, the credit amount approved also tends to increase.
- However, there are some outliers, particularly at the higher end of the “AMT_APPLICATION” range.

Contract type w.r.t contract status

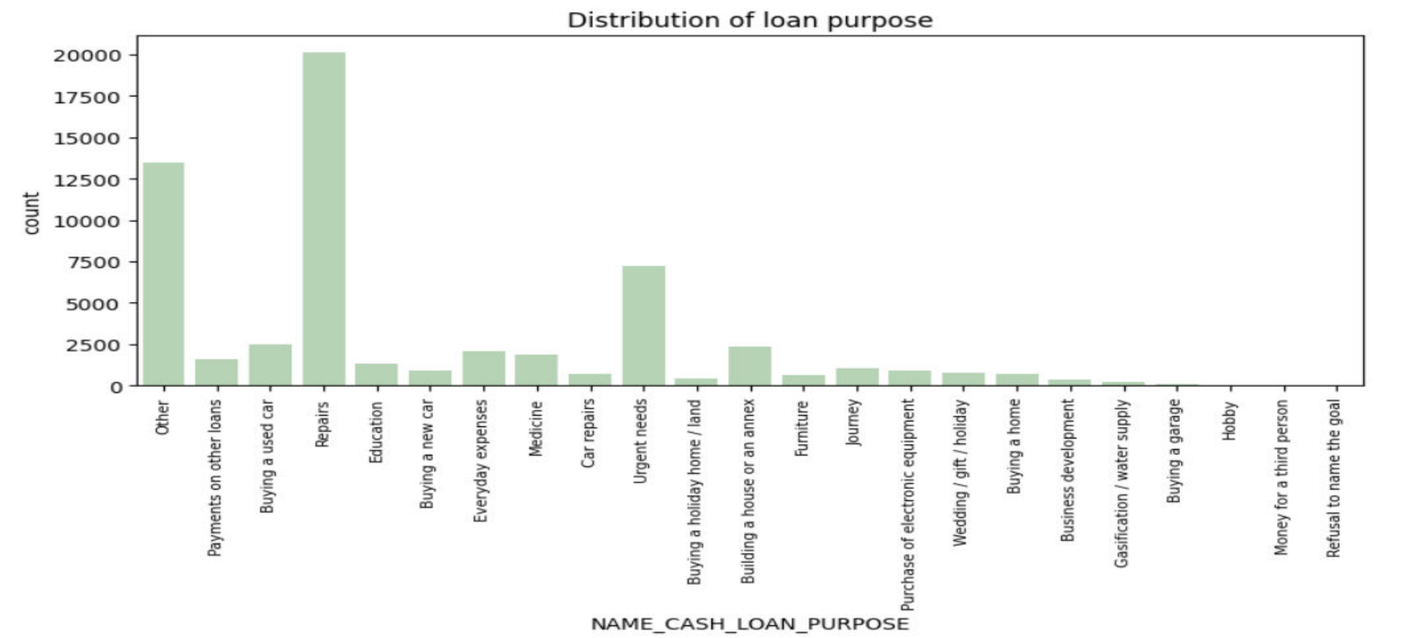


Insights:

- Consumer loans have the highest approval count, whereas Cash loans have a moderate approval rate and Revolving loans have the lowest approval count.
- Cash loans have the highest cancellation count, Revolving loans have a moderate cancellation count and Consumer loans have a very low cancellation count.
- Revolving loans have the highest refusal rate, then are Consumer loans and least are the Revolving loans.

Merged data Analysis

Distribution of various purposes for loan



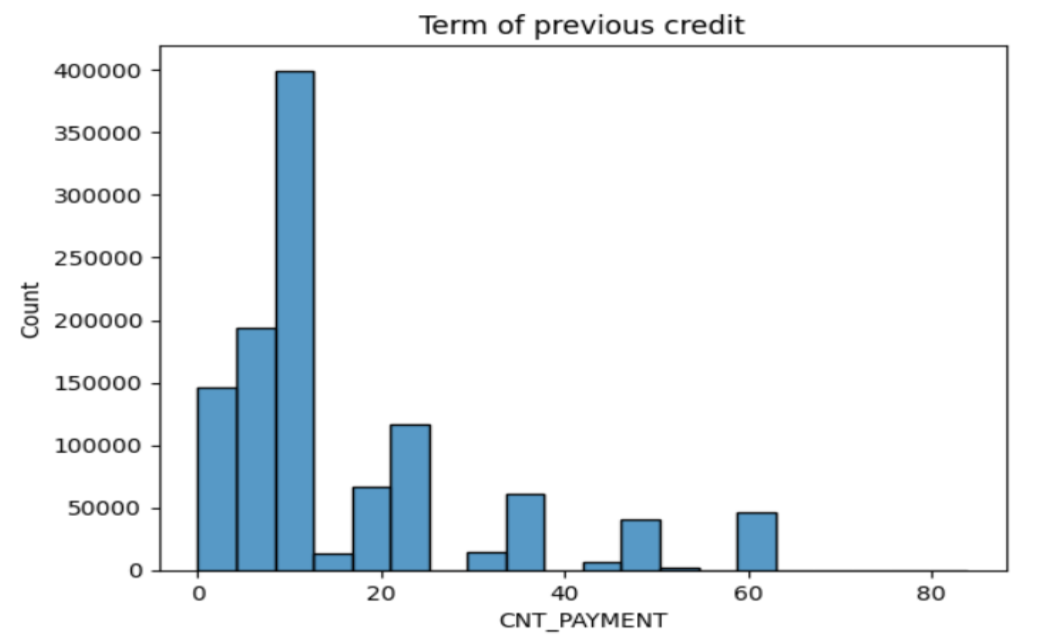
Insights

- Most common purpose for loan is for 'Repairs' which is around 20000, i.e. around 30% of loans, then comes 'Others' and 'Urgent needs'.
- Other common loan purposes include 'Buying a used car', 'Building a house or an annex' and 'Everyday expenses', which are around approximate 2500.

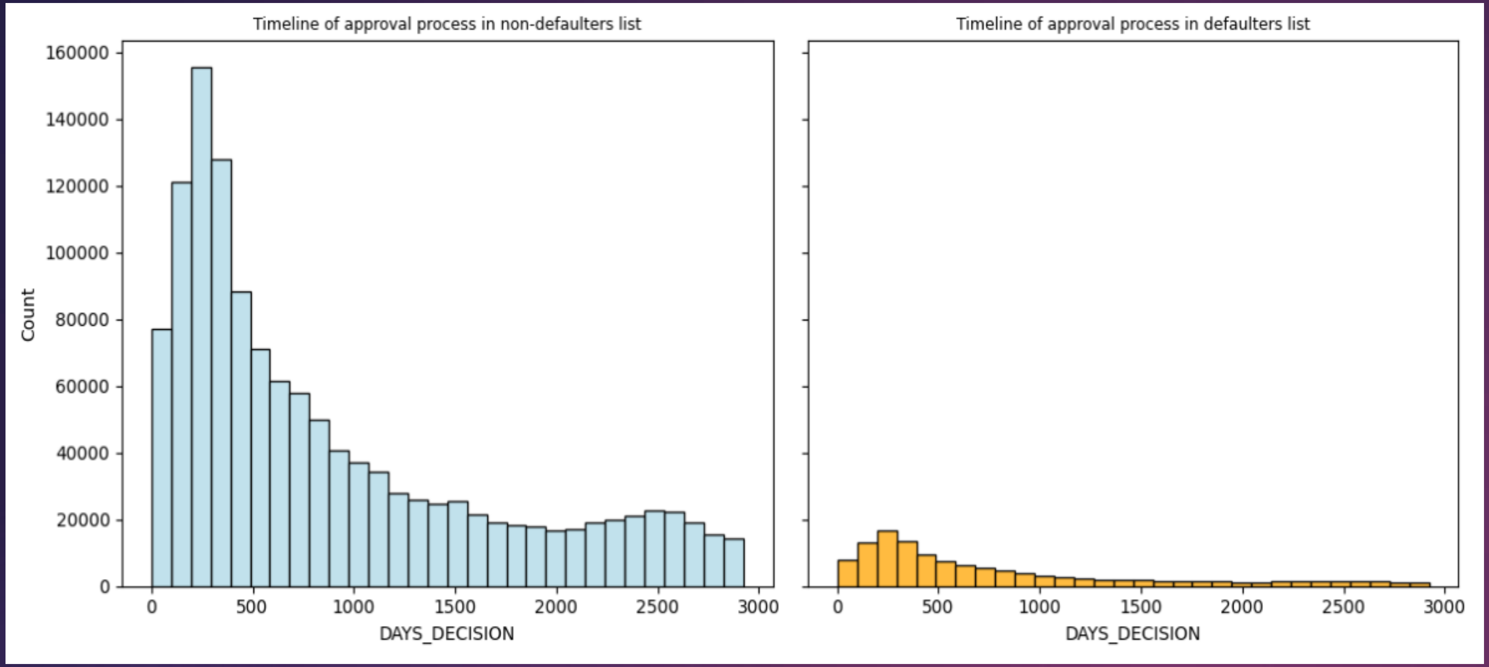
Credit repayment loan duration

Insights

- The most common loan terms/duration fall between 10 and 20 months/years (since it is not specified in the dataset).
- There is a significant decrease in frequency for loan terms longer than 20 months/years.



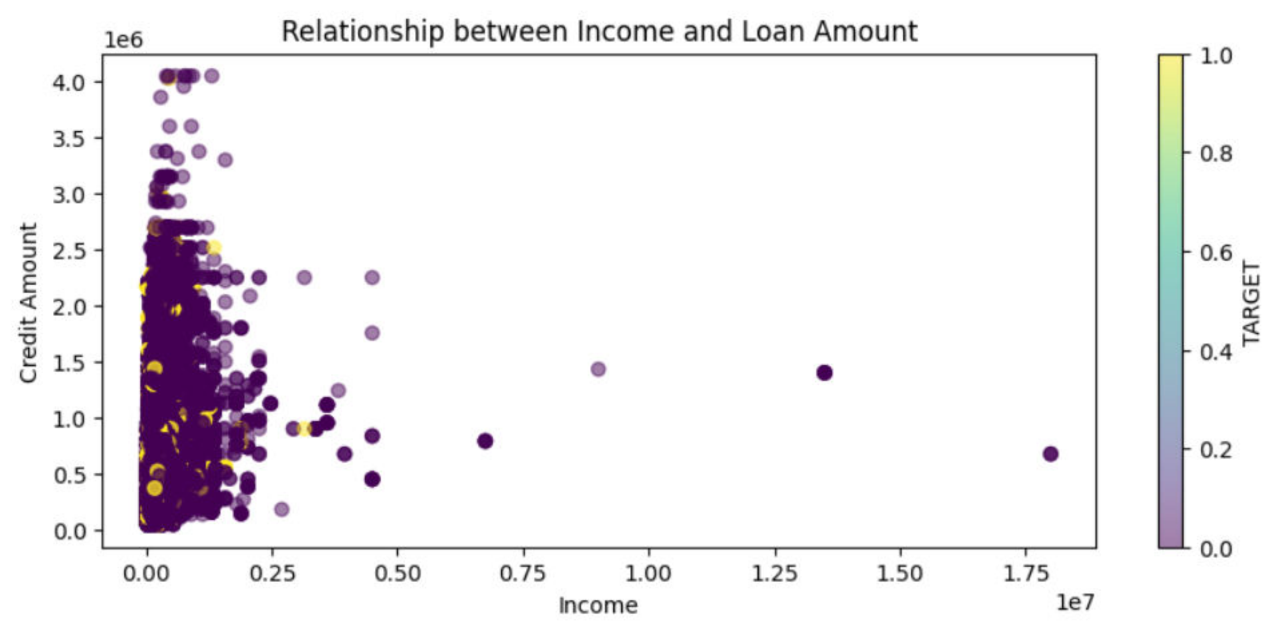
Time taken for loan approval



Insights

- The decision about the previous application for non-defaulters tends to be made relatively quickly, with a peak around 500 days.
- For defaulters, the decisions are more spread out, but generally, they are made in less than 1000 days.

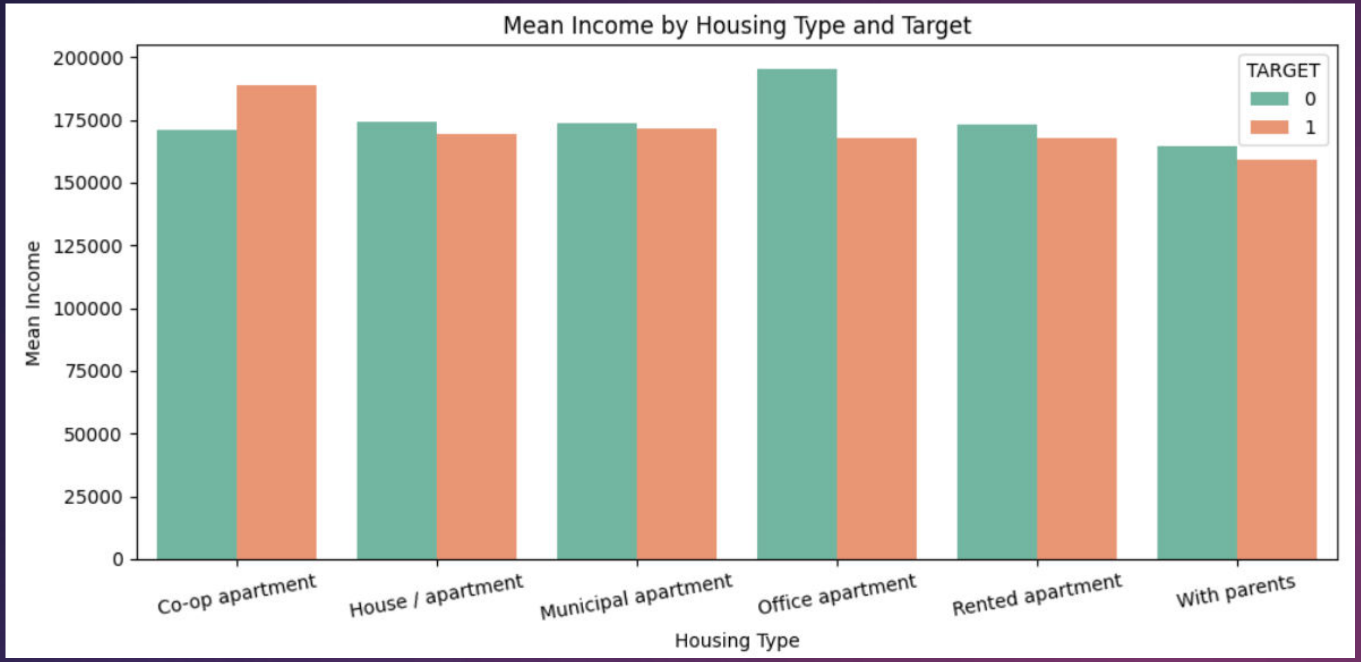
Income vs Credit Relation



Insights

- There is a positive correlation between Income and Loan Amount, as most data points are concentrated at lower values for both variables and follow a very little diagonal pattern towards higher values.
- Purple dominating yellow, indicates that defaulters more prevalent in the dataset, especially at the lower end of income and loan amounts.

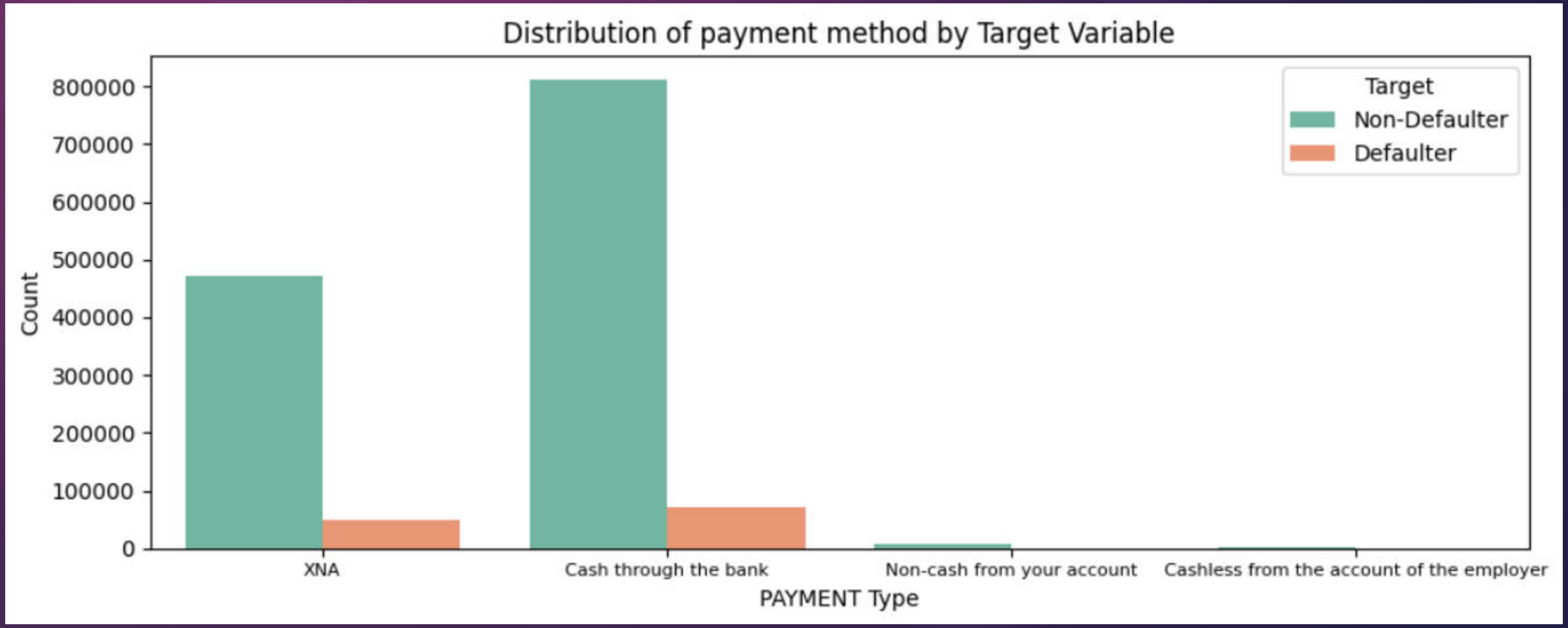
Average income across different housing types



Insights

- For every housing type, the mean income for non-defaulters (TARGET = 0) is slightly higher than for defaulters (TARGET = 1), except for the clients with co-op apartment.
- Mean-income of deaulters of co-op apartment is greater than their non-defaulter.
- Mean-income of non-defaulters is highest with the clients having office apartments.

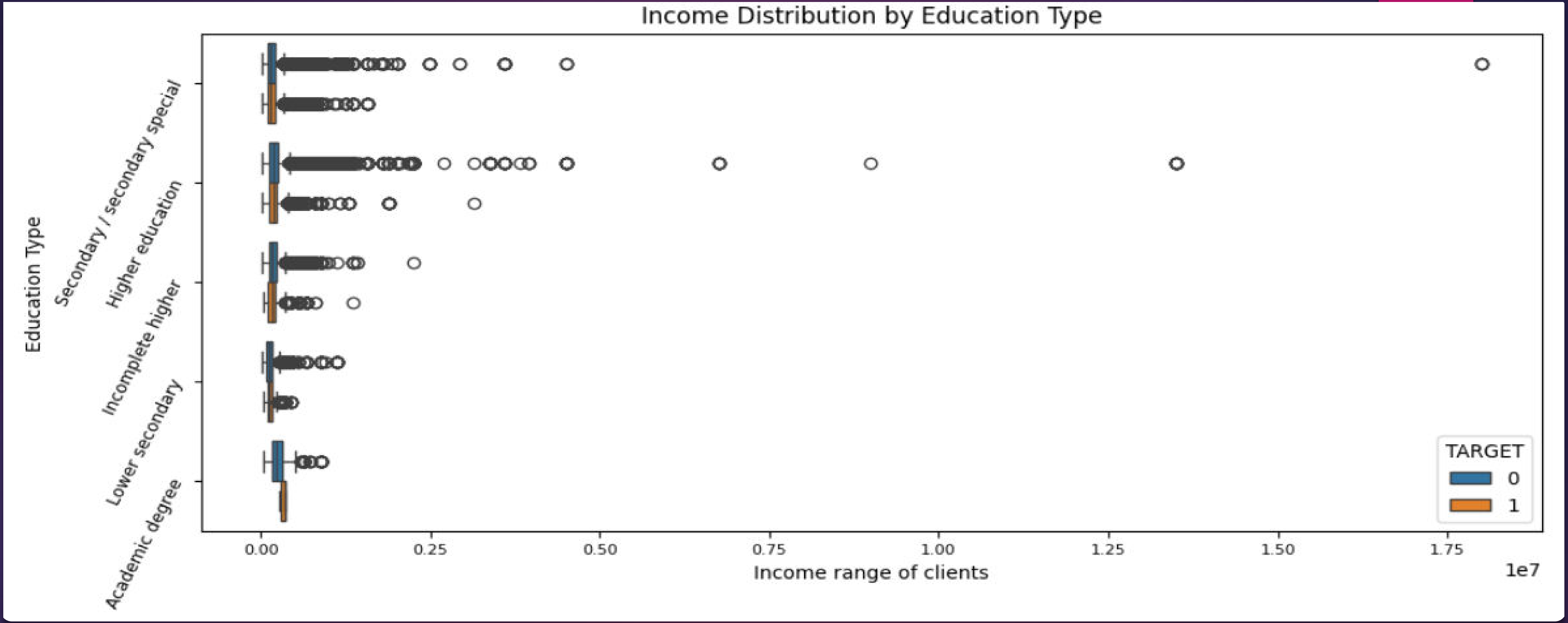
Payment methods clients choose



Insights

- "Cash through the bank" is the most common payment method for both non-defaulters (TARGET = 0) and defaulters (TARGET = 1). A significantly higher number of successful repayments i.e. (TARGET = 0), were made using this method compared to defaulters.

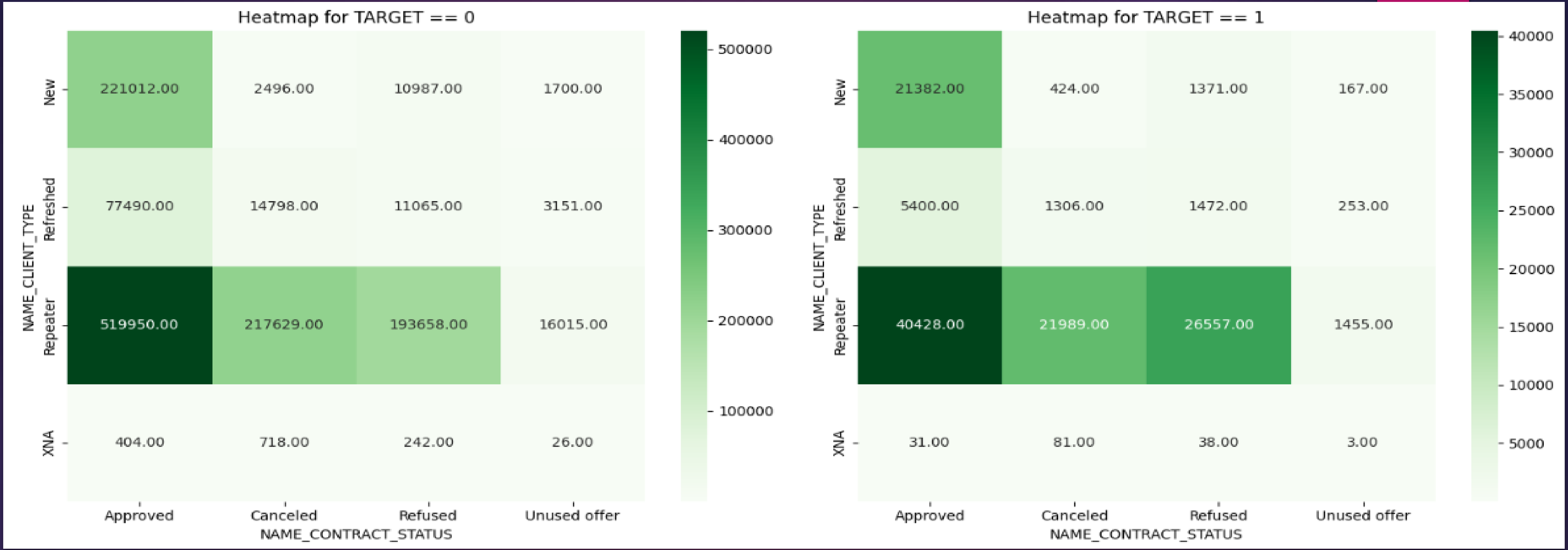
Distribution of income across different education level



Insights

- There is no significant difference in the median income between Non-Defaulters(Target = 0) and Defaulters(Target = 1) across all education levels, except for the presence of outliers, particularly in the ‘Higher Education’ category, indicates that there are individuals with incomes much higher than the median
- Individuals with higher levels of education (Academic degree and Higher education) have higher average incomes.

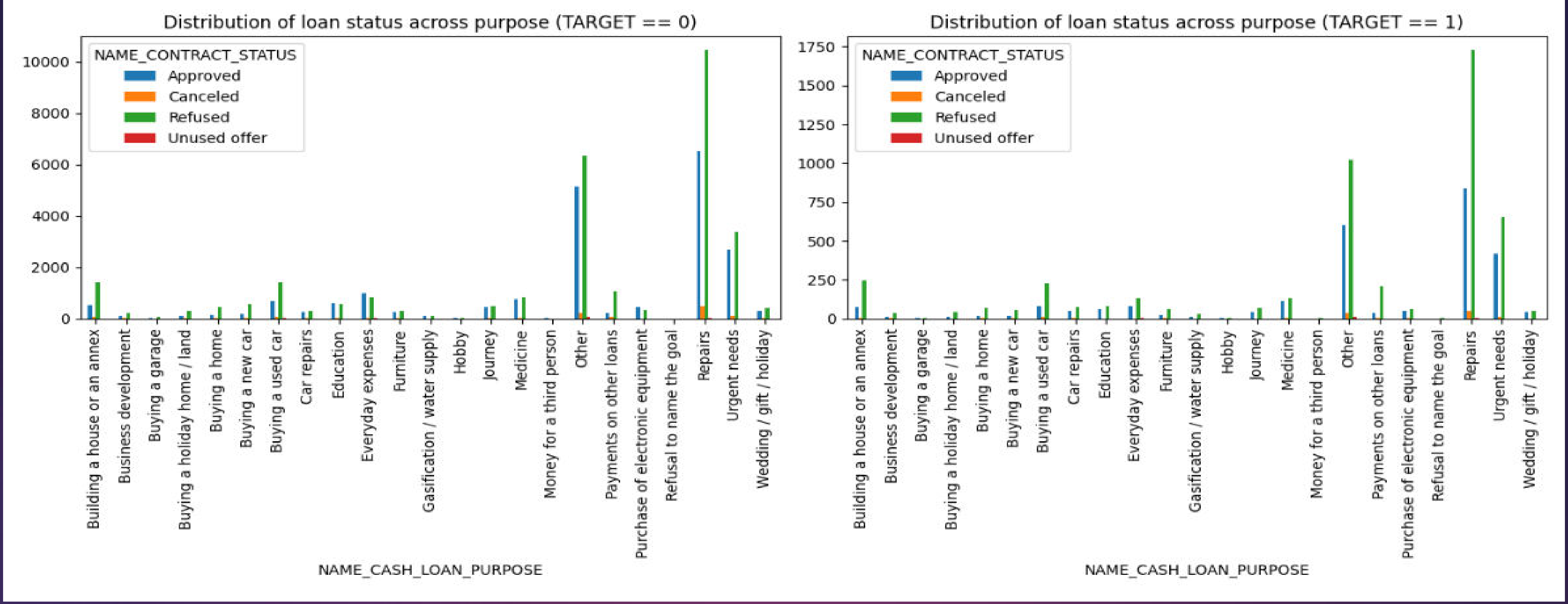
Loan status across client type combination count



Insights

- The heatmaps provide a visual comparison of the counts of different contract statuses across client types, segmented by the target variable (TARGET == 0 for non-defaulters and TARGET == 1 for defaulters).
- ‘Repeaters' are the most common client type with approved contracts For both TARGET == 0 and TARGET == 1
- 'New Clients' category for Non-defaulters have a higher count of approvals compared to the defaulters.
- The count of 'refused' contracts for 'Repeaters' are high in loan default as compared to the non-defaulters.

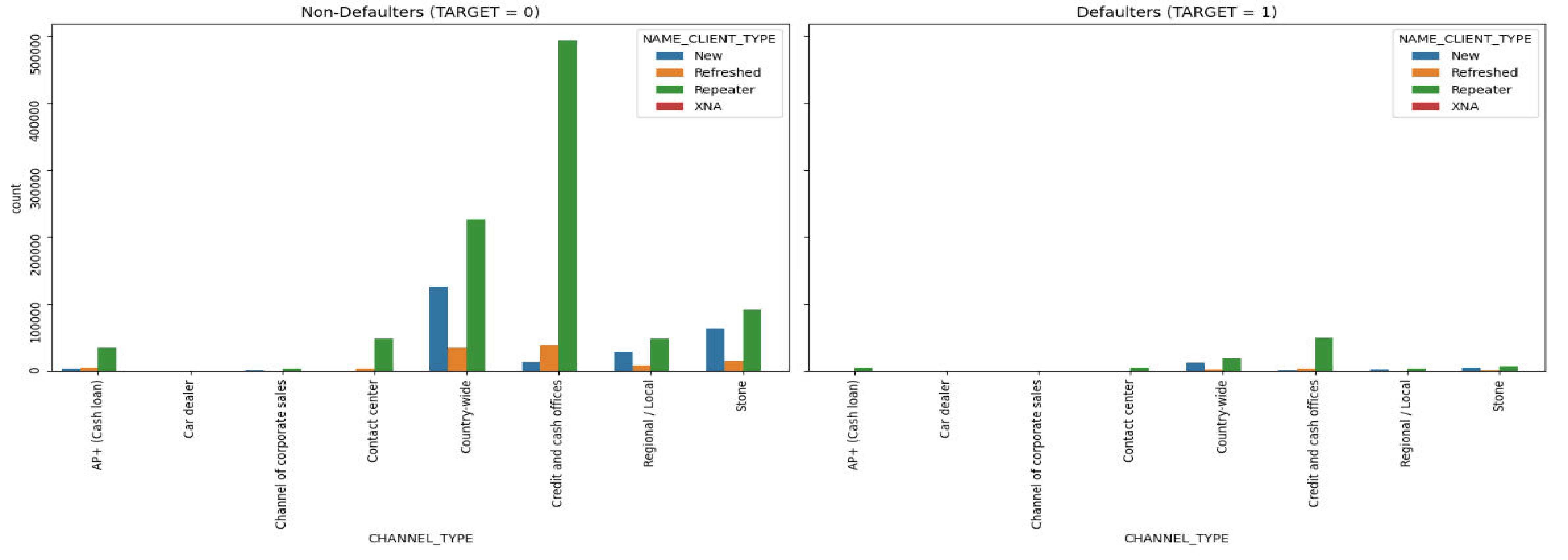
Status of loan for different loan purposes



Insights

- Across both target segment 'Repairs' category has the highest approved and refused loan status, similar with the 'Other' category.
- For non-defaulters (TARGET == 0), the majority of loans across various purposes are refused, same goes for defaulters (Target ==1)

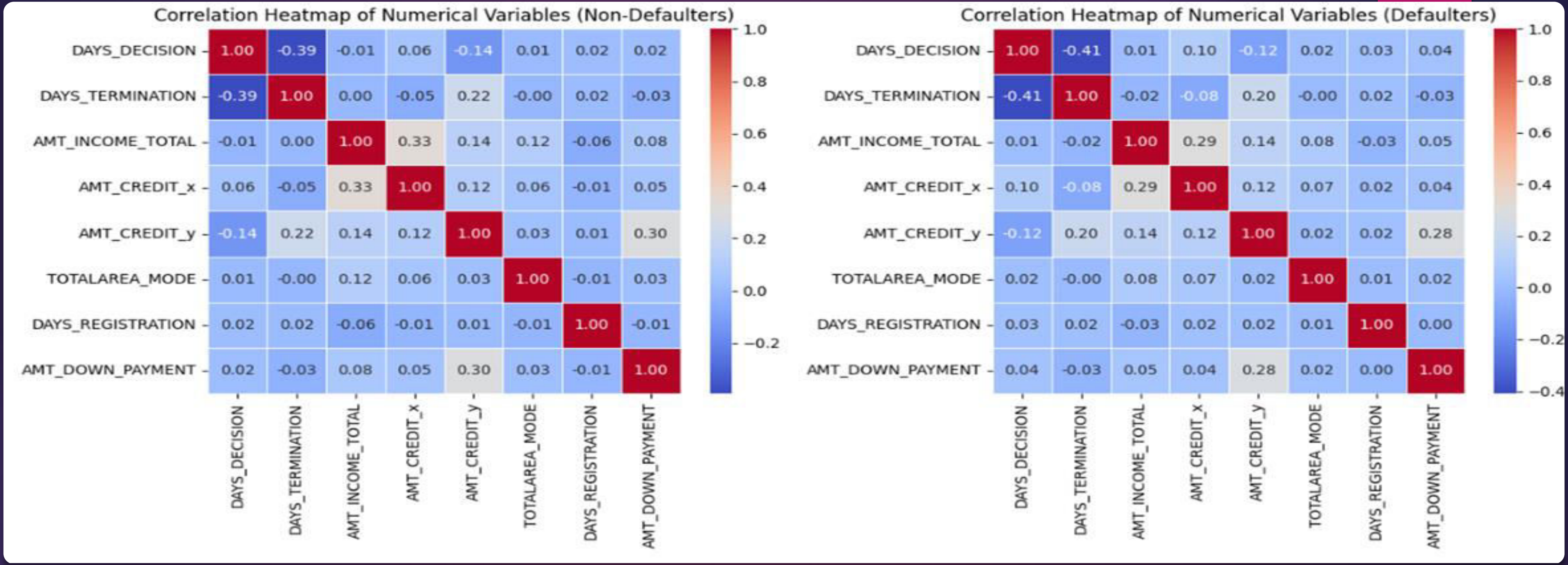
Type of clients applied across different Channels for loan w.r.t Target variable



Insights:

- For the Non-Defaulters (TARGET = 0): The "Credit and cash offices" channel has the highest count of clients, with 'Repeaters' being the most prevalent client type.
- Incase of Non-defaulters, The "Country-wide" channel has a significant number of 'New' clients
- For Defaulters (TARGET = 1): All channels have fewer defaulting clients compared to non-defaulting ones.
- There is increase in refused contracts for 'Repeaters' in the defaulters' category.

Correlation across different types of numerical variables w.r.t Target Variable



Insights:

- For relation between variables in both non-defaulters and defaulters plot:
- DAYS_DECISION has a weak negative correlation with both loan termination time and income. Moderate positive correlation between loan amount (both AMT_CREDIT_x and AMT_CREDIT_y) and income. A weak positive correlation between loan amount (AMT_CREDIT_x/y) and down payment. A weak positive correlation between income and property size (TOTALAREA_MODE).
- The overall correlation strength among variables is weaker for Defaulters than for Non-Defaulters

THE END