Credit EDA Case Study

Ramya D & Rahul Raj, DSC 27

April 2021

Contents

Key Objectives

Identify various patterns among different groups of applicants

Identify patterns which indicate if a client is has payment difficulties.

Understand how customer attributes and past loan attributes influence the tendency of default among the applicants

These insights can be used for taking appropriate actions such as denying loan, increasing interest rate etc. Which will ensure that the applicants capable of repaying the loan are not rejected.

Inspecting Dataset

Given dataset contains information of applicants at both the time of application as well as their previous loans data

- The first dataset 'application_data.csv' contains all the information of the client at the time of application. This data talks about whether a client has payment difficulties.

- A total of 122 attribute for 3,07,511 applicants with their repayment capability is given in this csv file

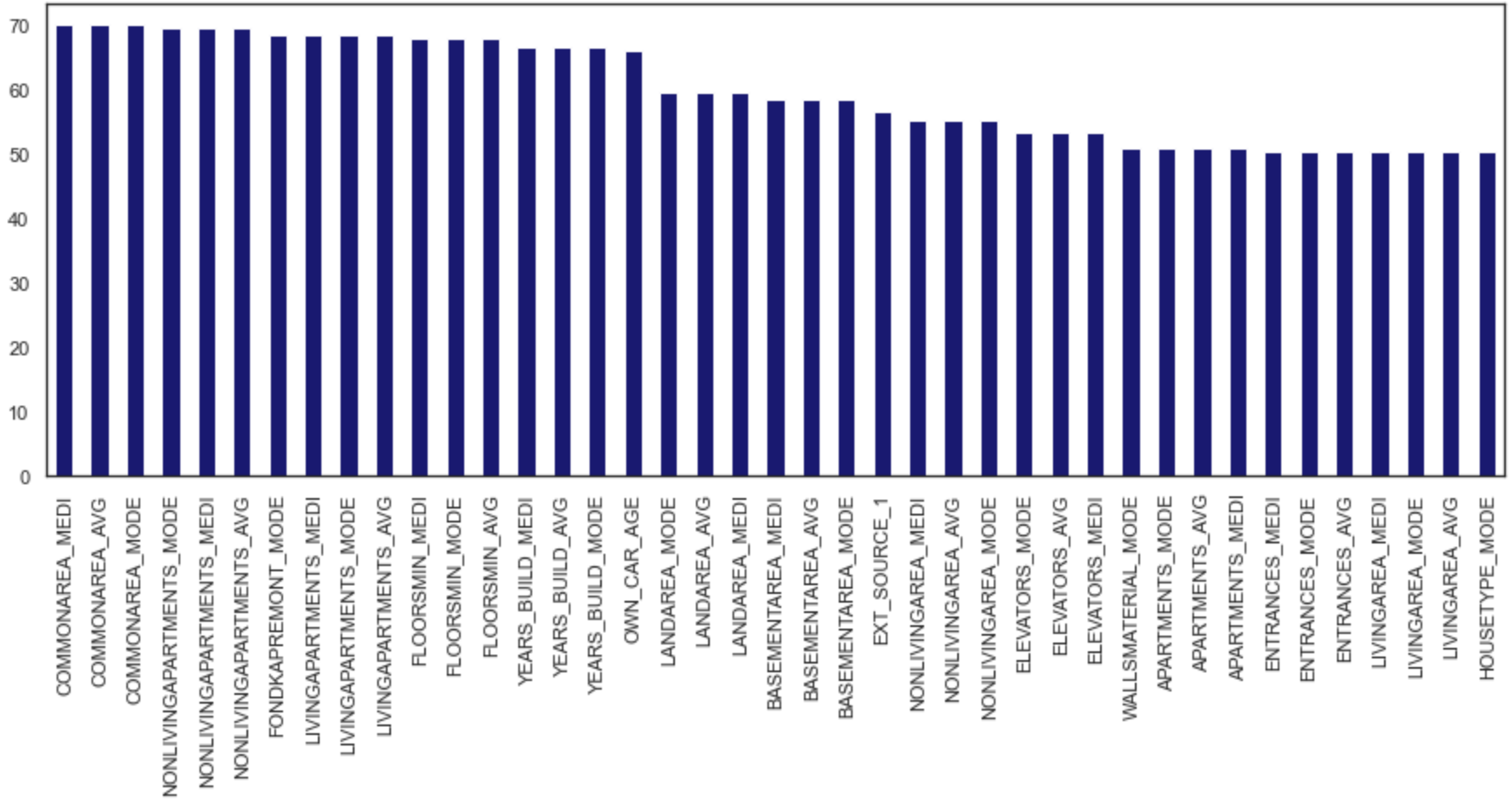
- All the applicants have a unique loan id called “SK_CURR_ID” which is the id given by the bank to that applicant’s loan. The column “TARGET” is an indicator variable (has values 0/1) which gives us the information of payment repaying capabilities of the applicant.

- A quick run of .info() method on this data frame shows that there are 65 float64 columns, 41 int columns and 16 object type columns.

Missing value analysis

Out of 122 columns in the application dataset, there are 67 columns with at least 1 missing value and 41 columns with at least 50% missing values.

- It is clear from the data dictionary that there are no Missing Column Names or Inconsistent column names. The columns with missing value percentage greater than 50 and dropped are shown below.



Missing value analysis (contd.)

There are no duplicate or blank rows in this dataset. After dropping the columns with missing % > 50 and converting “SK_CURR_ID” to index, we are left with 80 columns

- A quick analysis of missing value on row wise observations showed that, the maximum missing observations in a row was 20 (out of 80).
- As this is well within our threshold, we didn’t remove any row wise observations to deal with missing values.
- We also noticed that there are few values in the columns like CODE_GENDER and ORGANIZATION_TYPE which should be treated as missing (These values are given as XNA).
- We converted these values into NaN, so that we can impute them with appropriate value in the next step.
- Like the above approach, In the column DAYS_EMPLOYED, there is an extreme value 365243 used to indicate the missing value. We have replaced this value with NaN as well.

Imputation Approach

Different approaches are used for different columns to not include any bias in our analysis

- The column OCCUPATION_TYPE has 31.34% of its values missing and as this is a categorical variable, instead of imputing them with the most repeated observation, we decided to create a new OCCUPATION_TYPE called “Missing” for these missing observations.
- For the column ORGANIZATION_TYPE with 18.00% of missing values, we observed that this column follows “Missing not at Random (MNAR)” pattern.
- We found a strong relationship between the missing observations in ORGANIZATION_TYPE and “Pensioners” in OCCUPATION_TYPE i.e, If an applicant is Pensioner, we found his ORGANIZATION_TYPE to be missing
- Hence, we decided to impute these missing values with “Pensioner” in ORGANIZATION_TYPE
- For the other columns with missing values, we’ve imputed them with Median/ Mode depending on the type of the column

Data Cleaning

- The columns DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, DAYS_LAST_PHONE_CHANGE has values in the days with reference to the days in which the applicant has made his application.
- Hence, by definition all these columns have negative values. We've converted them to positive and changed their units from Days to Years/Months to extract more information from this column.
- Based on the DAYS_BIRTH column, we've created a new categorical column called AGE_GROUP to map the applicants into the age groups "18_to_25", "25_to_35", "35_to_60", and "60 above"
- In the next step, we converted the Yes/No values in FLAG_OWN_CAR, FLAG_OWN_REALTY to 1/0 so that appropriate numerical analysis can be converted on them.
- Also based on the numeric columns of applicant's income and applicant's loan value, we created a couple of categoric columns to indicate the quantile in which an applicant falls in the dataset.

Data Cleaning (contd.)

- Also, we have created 3 columns called FLAG_OWN_CAR_REALTY, TOTAL_PHN_NOS, TOTAL_DOCUMENTS to indicate the ownership status of both car and apartment of applicant, total number of phone numbers submitted by the applicant, and the total number of documents submitted by the applicant, respectively.

- After correcting the data types of the columns, at the end of this data cleaning step, our data frame has 83 columns (13 category, 25 float, 46 int) for 3,07,511 loan applications.

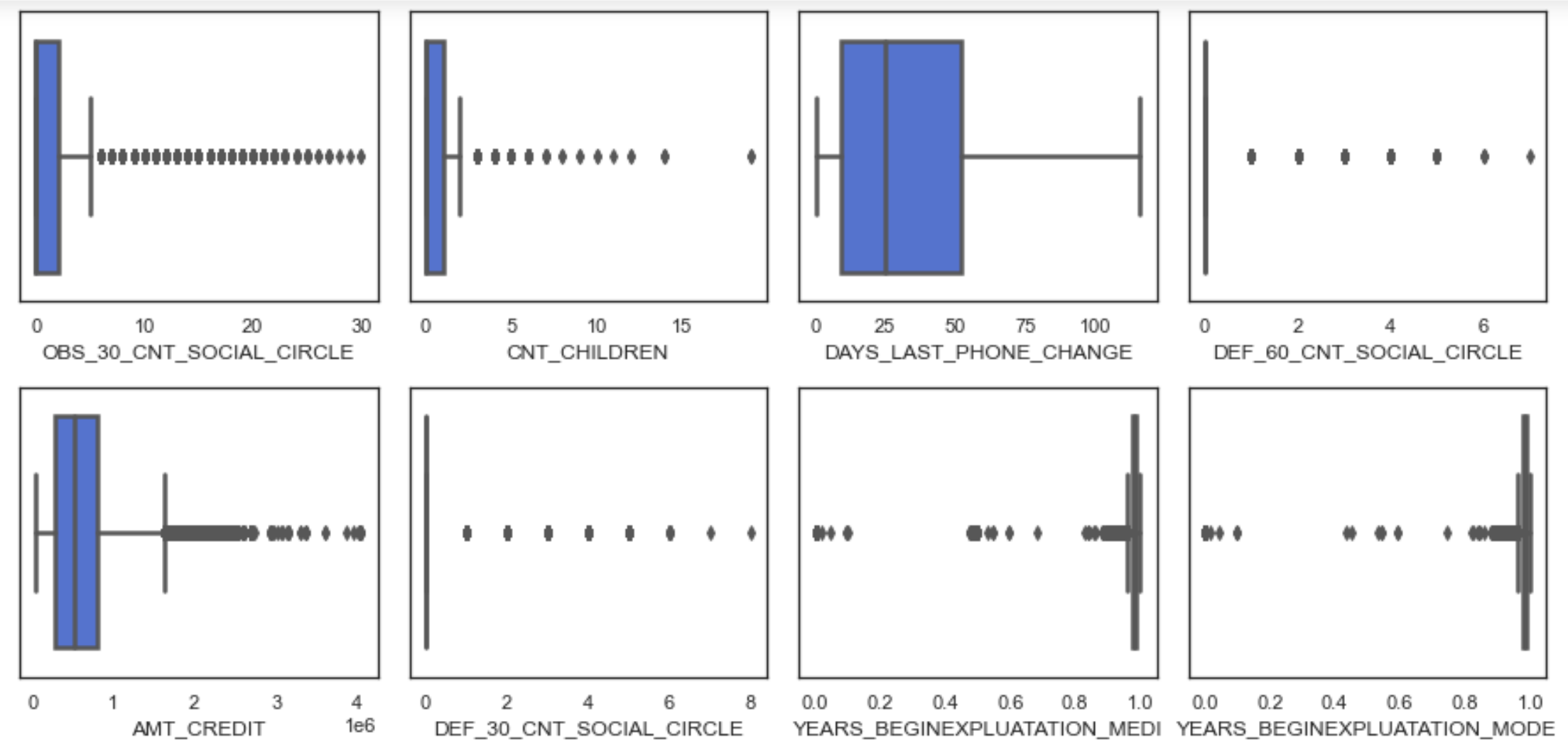
- Careful observation of the columns showed that 31 columns in the dataset are not relevant/ provides too little value to our analysis. We've dropped these columns before proceeding with the final set of analysis on the cleaned dataset

Outlier Analysis

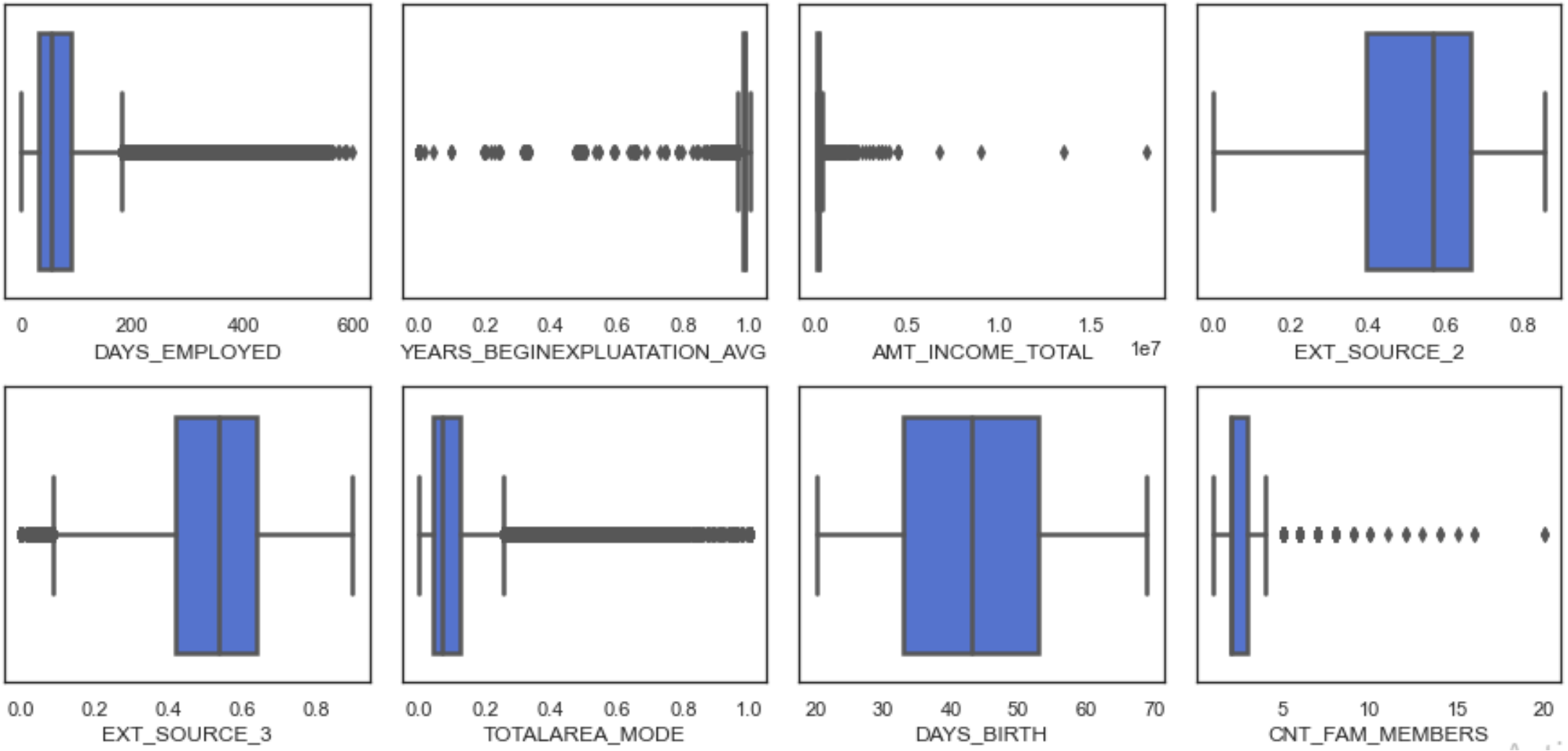
Out of all the columns in our dataset, only 26 were relevant for outlier analysis.

Variable Name	Outlier Count	Lower whisker	Upper whisker
EXT_SOURCE_2	0	-0.012698	1.069094
DAYS_BIRTH	0	5.5	81.5
DAYS_ID_PUBLISH	0	-72	272
TOTAL_PHN_NOS	29	1.5	5.5
DAYS_LAST_PHONE_CHANGE	449	-55.5	116.5
DAYS_REGISTRATION	654	-206	522
CNT_FAM_MEMBERS	4007	0.5	4.5
CNT_CHILDREN	4272	-1.5	2.5
EXT_SOURCE_3	4313	0.088185	0.9652909
YEARS_BEGINEXPLUATATION_MEDI	4762	0.96185	1.00145
YEARS_BEGINEXPLUATATION_AVG	4784	0.96185	1.00145
YEARS_BEGINEXPLUATATION_MODE	5074	0.96185	1.00145
FLOORSMAX_MODE	5104	-0.0832	0.5832
FLOORSMAX_AVG	5215	-0.0832	0.5832
FLOORSMAX_MEDI	5360	-0.0832	0.5832
AMT_CREDIT	6562	-537975	1616625
AMT_ANNUITY	7504	-10584	61704
TOTALAREA_MODE	12091	-0.0884	0.2572
AMT_INCOME_TOTAL	14035	-22500	337500
AMT_GOODS_PRICE	14728	-423000	1341000
OBS_60_CNT_SOCIAL_CIRCLE	19564	-3	5
OBS_30_CNT_SOCIAL_CIRCLE	19971	-3	5
DAYS_EMPLOYED	22960	-60.5	183.5
DEF_60_CNT_SOCIAL_CIRCLE	25769	0	0
DEF_30_CNT_SOCIAL_CIRCLE	35166	0	0
TOTAL_DOCUMENTS	37455	1	1

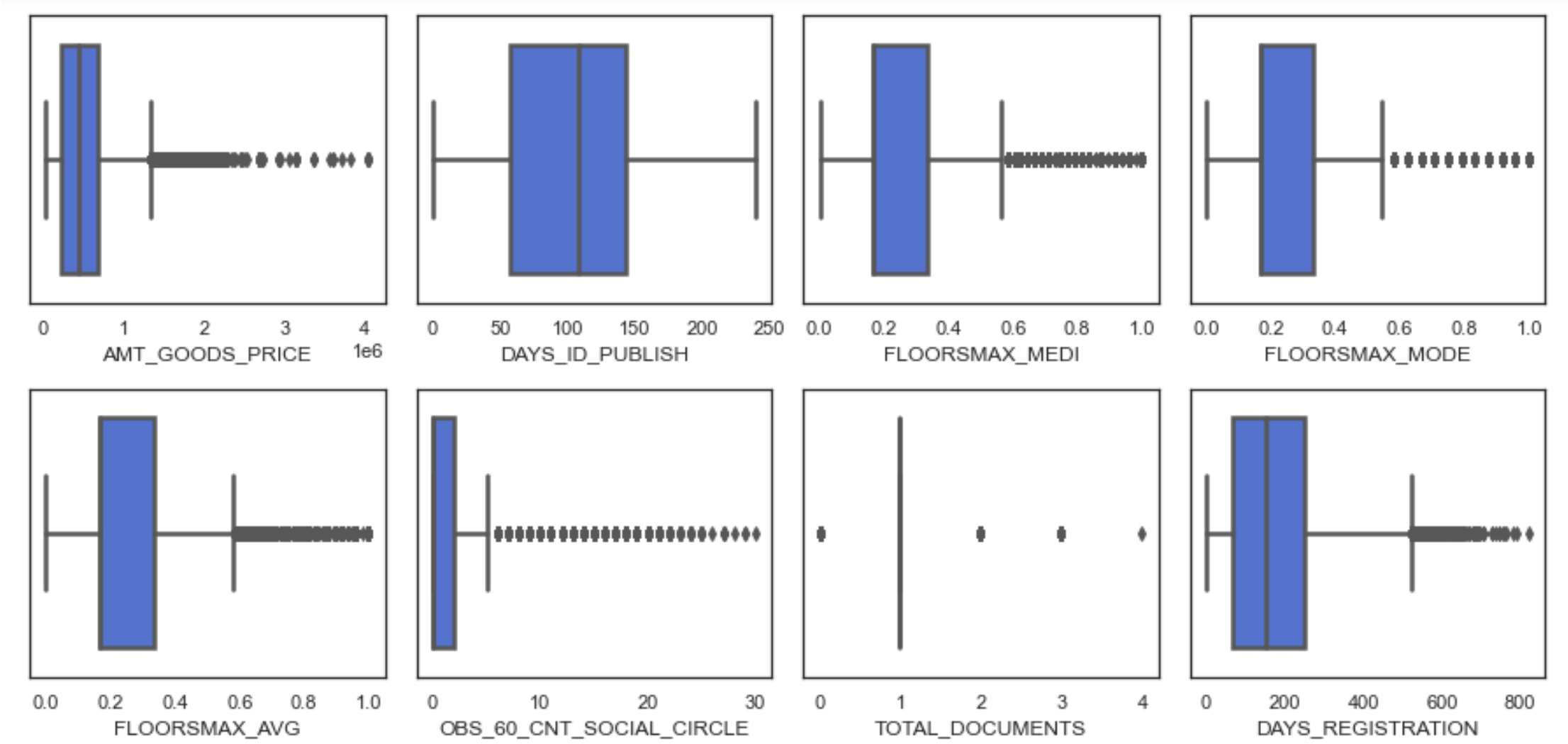
Outlier Analysis



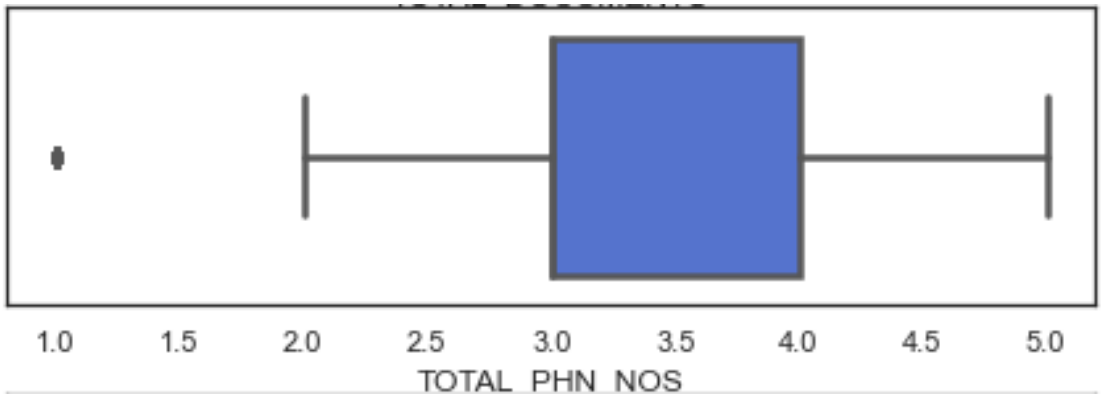
Outlier Analysis (contd.)

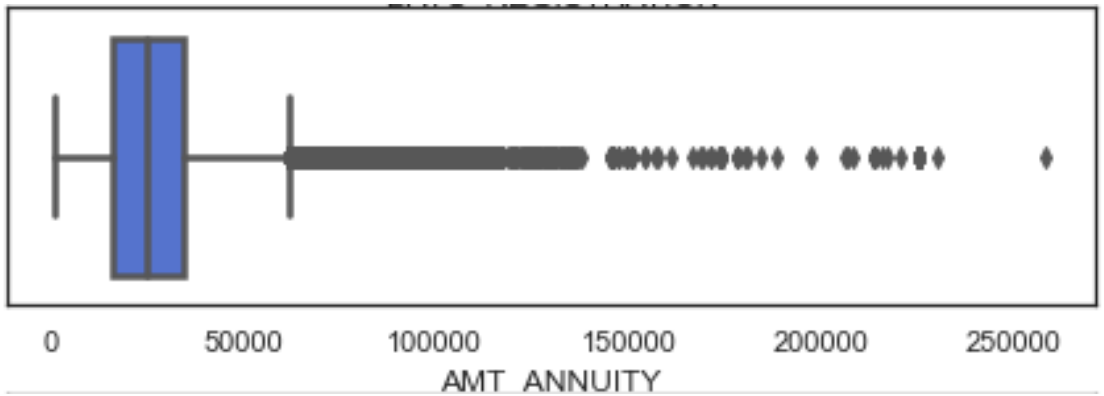


Outlier Analysis (contd.)



Outlier Analysis (contd.)





- Majority of the people submitted just 1 document during loan applications. We found a few applications where the total documents submitted is 0
- We have few applications in which the applicant has changed his registration at least 20 years before the date of applications, these values must be treated as outliers for this column

Outlier Analysis (contd.)

- There appears to be no outliers in the variable DAYS_OF_BIRTH. As all our applications lie between 20 and 80 years
- Applicants who have at least 5 children are treated as outliers in our dataset
- DAYS_EMPLOYED column shows that there are some applicants who have employed for more than 200 months (~17 years) and the highest value we observed here Is around 50 years of experience. We treat all such observations as outliers in this column.
- All the applicants who have changed their phone numbers 10 or more years ago are treated as outliers in our dataset
- From the box plot of the variable OBS_60_CNT_SOCIAL_CIRCLE, we can observe that most of this variable has a value of 1 and an extreme value of 350, which needs to be removed. A similar trend is observed in the other variables which talks about the DPD cases in the social circle of the applicant
- Most of the applicants pay an annuity of around 60,000. We even have some applicants who pay almost 2,50,000 as annuity.

Outlier Analysis (contd.)

- The goods price for which the applicants have taken a loan suggests that the needs of applicants are generally less than 10,00,000 but there are cases when the price of good is almost 40,00,000. (must be home loans/mortgages)

- AMT_CREDIT with values greater than 15,00,000 are observed as outliers in the box plot, but they are not necessarily outliers, and we will keep them as it is for our analysis.

- There is an extreme observation who has an income of around 12 crore which need to be removed before doing further analysis

- Since EXT_SCORE_3 is already a normalized value, the box plot confirms that all the observations lie between 0 and 1 and a very small proportion of it lies below 0.1 indicating that these people have a very low credit score

- Box plot of our derived variable TOTAL_PHNS show that, all the applicants have given at least 1 phone number

Outlier Analysis (contd.)

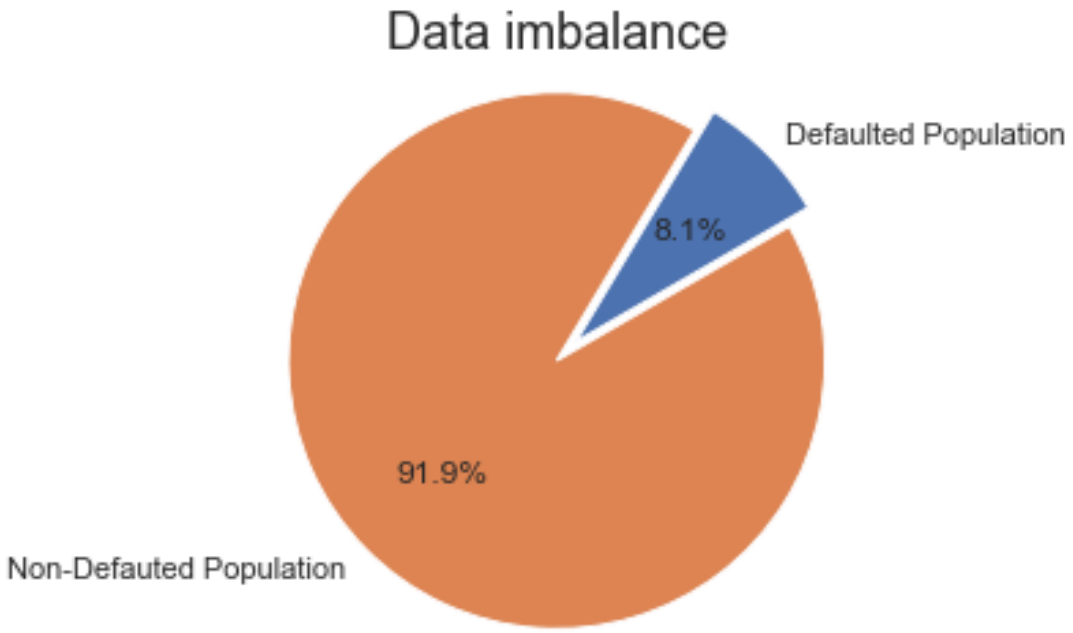
- Most of the applicants come from a family of size 5 members, but we have a very few observations whose family size is greater than 5. We have observed the largest size of the family is around 20

- EXT_SCORE_2 box plot indicates that all the values lie between 0 and 0.8 indicating that we do not have a population who has a very high credit score according to the source EXT_SCORE_2

- Clearly DAYS_ID_PUBLISH shows that there are no outliers for this variable

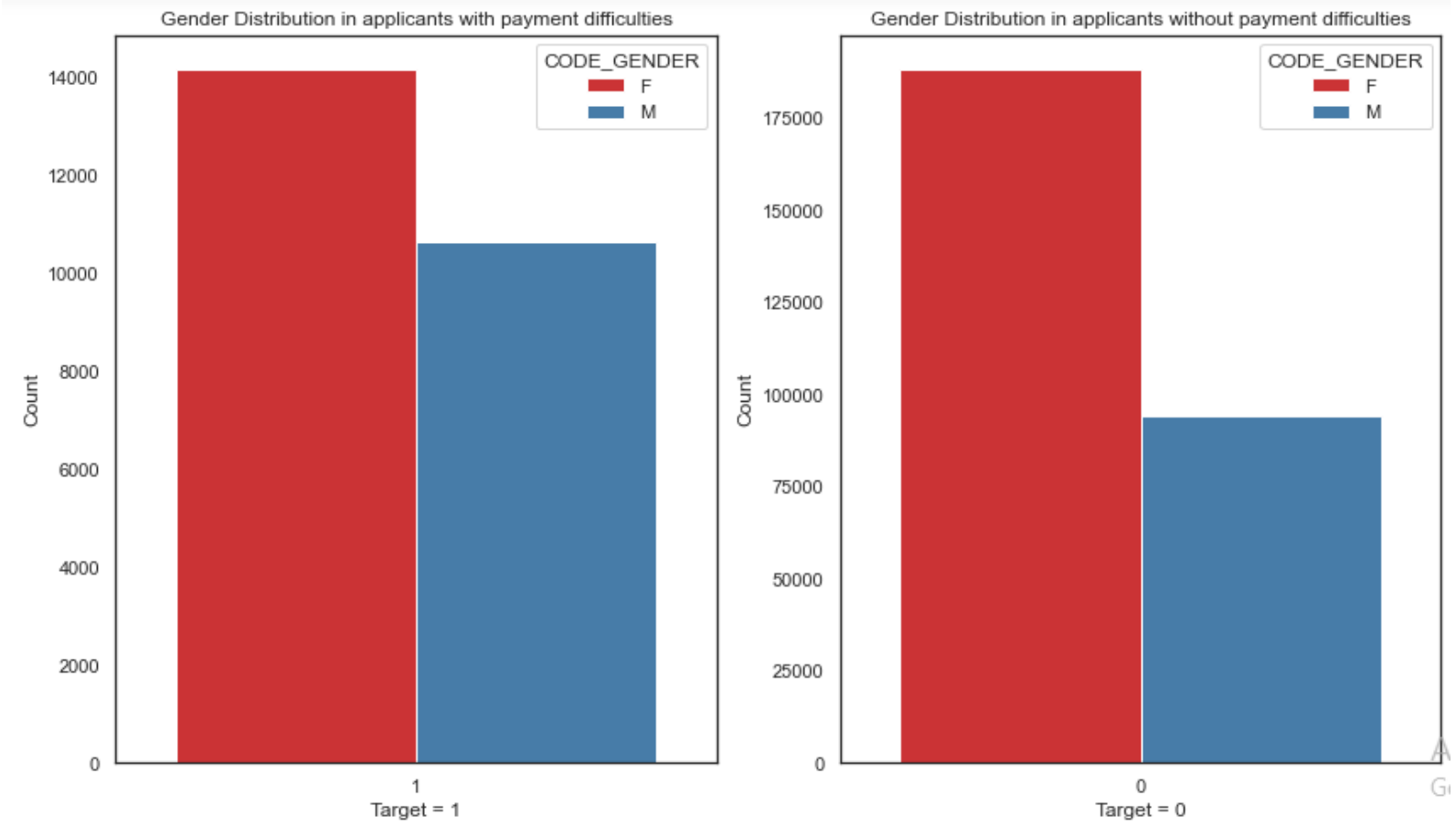
Univariate Analysis – TARGET variable

- A simple value_counts() on the TARGET variable showed that around 8.08% of the applicants in our dataset do not had payment difficulties i.e., and the remaining 91.92% of the applicants completed their payments on time giving us a data imbalance ratio of 11.38 in our dataset



Univariate Analysis – Categorical Variables

Distribution of Gender variable in both populations of with/without payment difficulties.

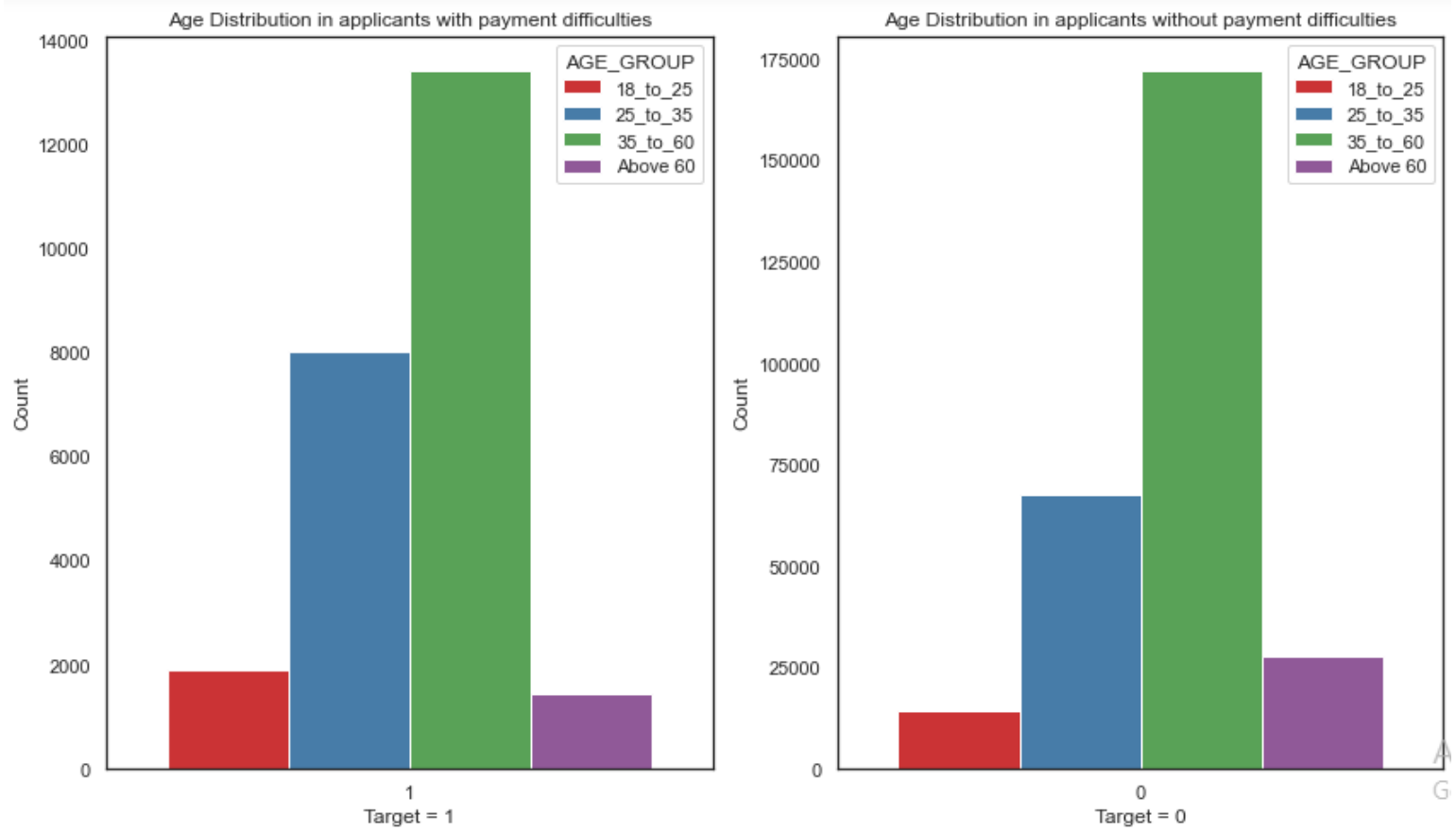


- From the dataset it appears that Female clients applied higher than Male clients for loan
- 66.62% Female clients are non-defaulters while 33.37% male clients are non-defaulters.
- 57.07% Female clients are defaulters while 42.92% male clients are defaulters.

Univariate Analysis – Categorical Variables (contd.)

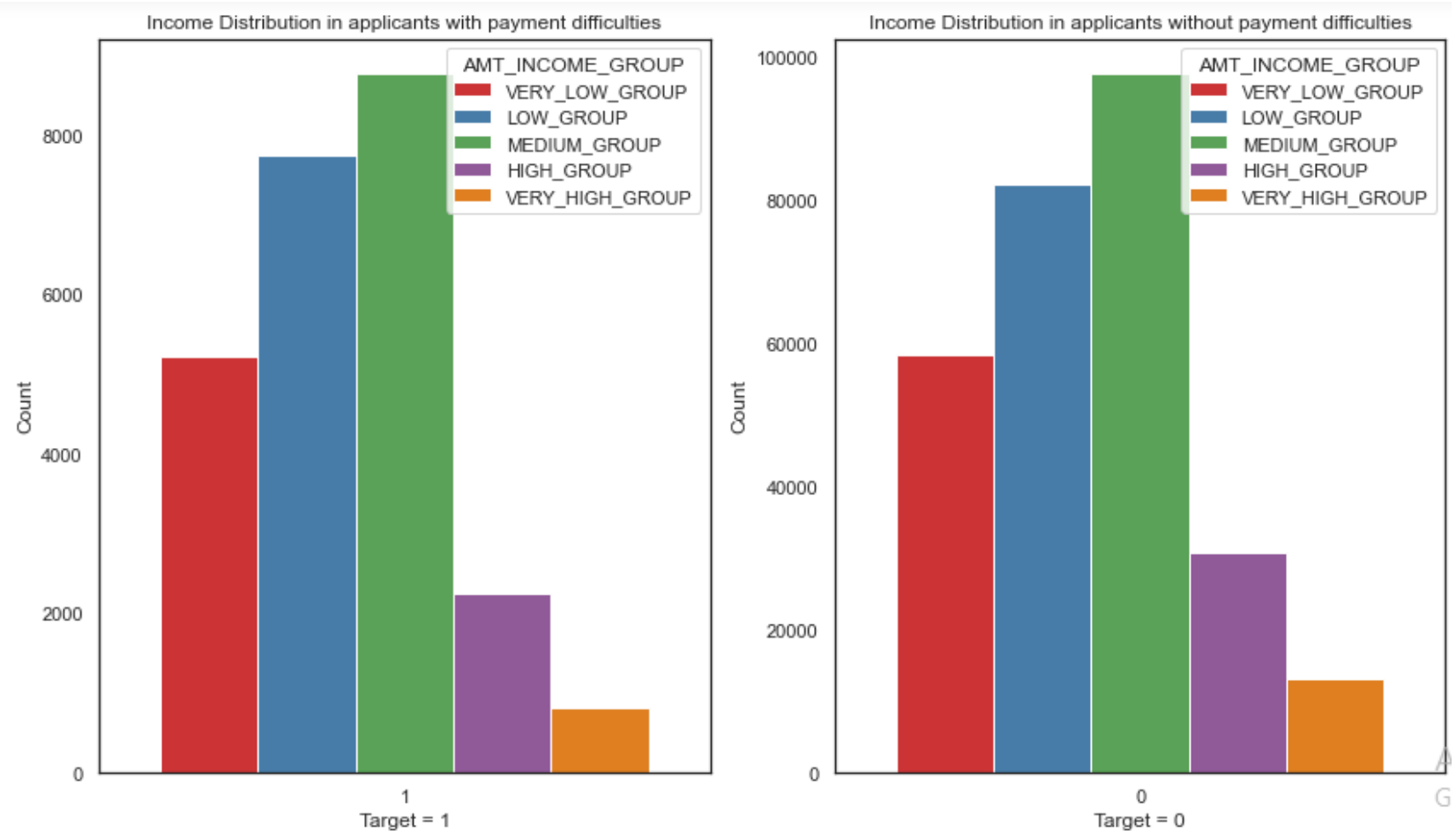
Distribution of Age variable in both populations of with/without payment difficulties.

- 35-60 group seems to applied higher than any other age group for loan in case of Defaulters as well as Non-defaulters.
- Also , 35-60 group facing paying difficulties the most.
- While >60 and 18-25 age group facing paying difficulties less as compared to other age groups.



Univariate Analysis – Categorical Variables (contd.)

Distribution of Income variable in both populations of with/without payment difficulties.

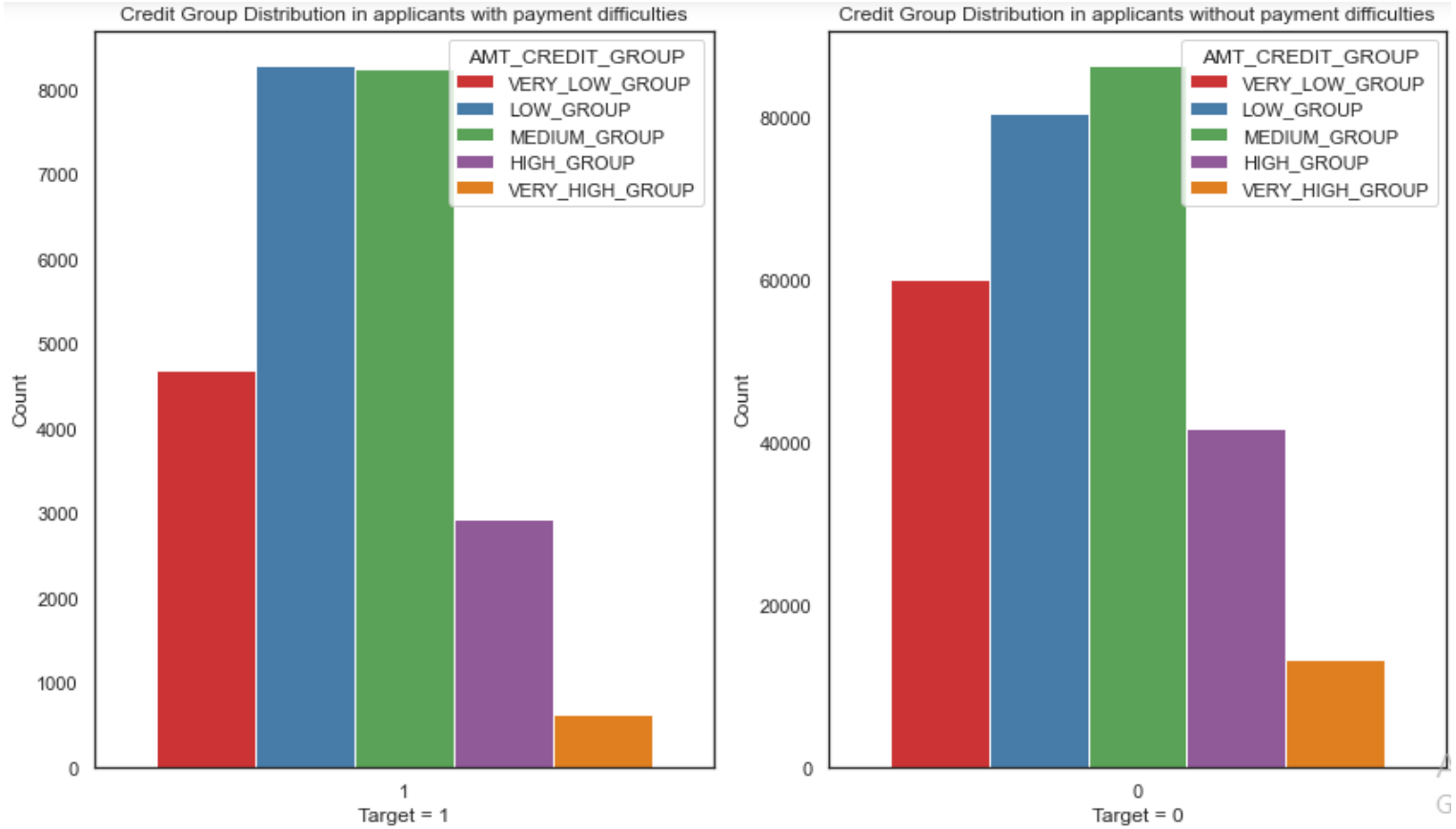


- Applicants with income in the MEDIUM group seems to applied higher than any other income group for loan in case of Defaulters as well as Non-defaulters.
- Also, applicants with income in the MEDIUM group are facing paying difficulties the most when compared to other groups.
- As expected, the applicants with VERY HIGH income, face low paying difficulties when compared to other income groups.

Univariate Analysis – Categorical Variables (contd.)

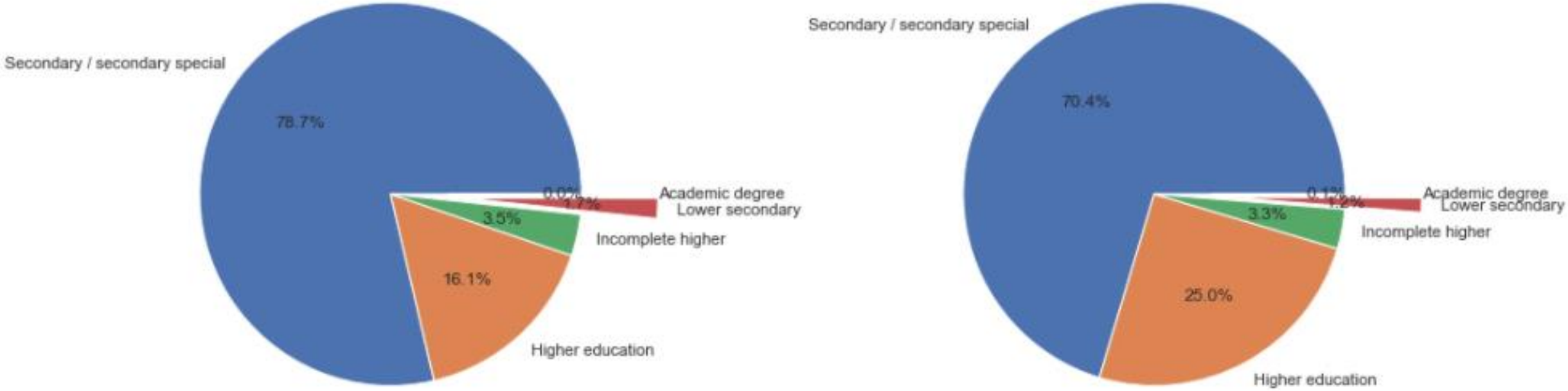
Distribution of Credit Group in both populations of with/without payment difficulties.

- Applicants who have applied for MEDIUM amount of loan are high compared to any other loan group in case of Defaulters as well as Non-defaulters.
- Also , applicants who have applied for a MEDIUM loan are facing paying difficulties the most when compared to other groups.
- Applicants who have applied for VERY HIGH loan, face low paying difficulties when compared to other loan groups possibly because of higher repaying capability from their high income.



Univariate Analysis – Categorical Variables (contd.)

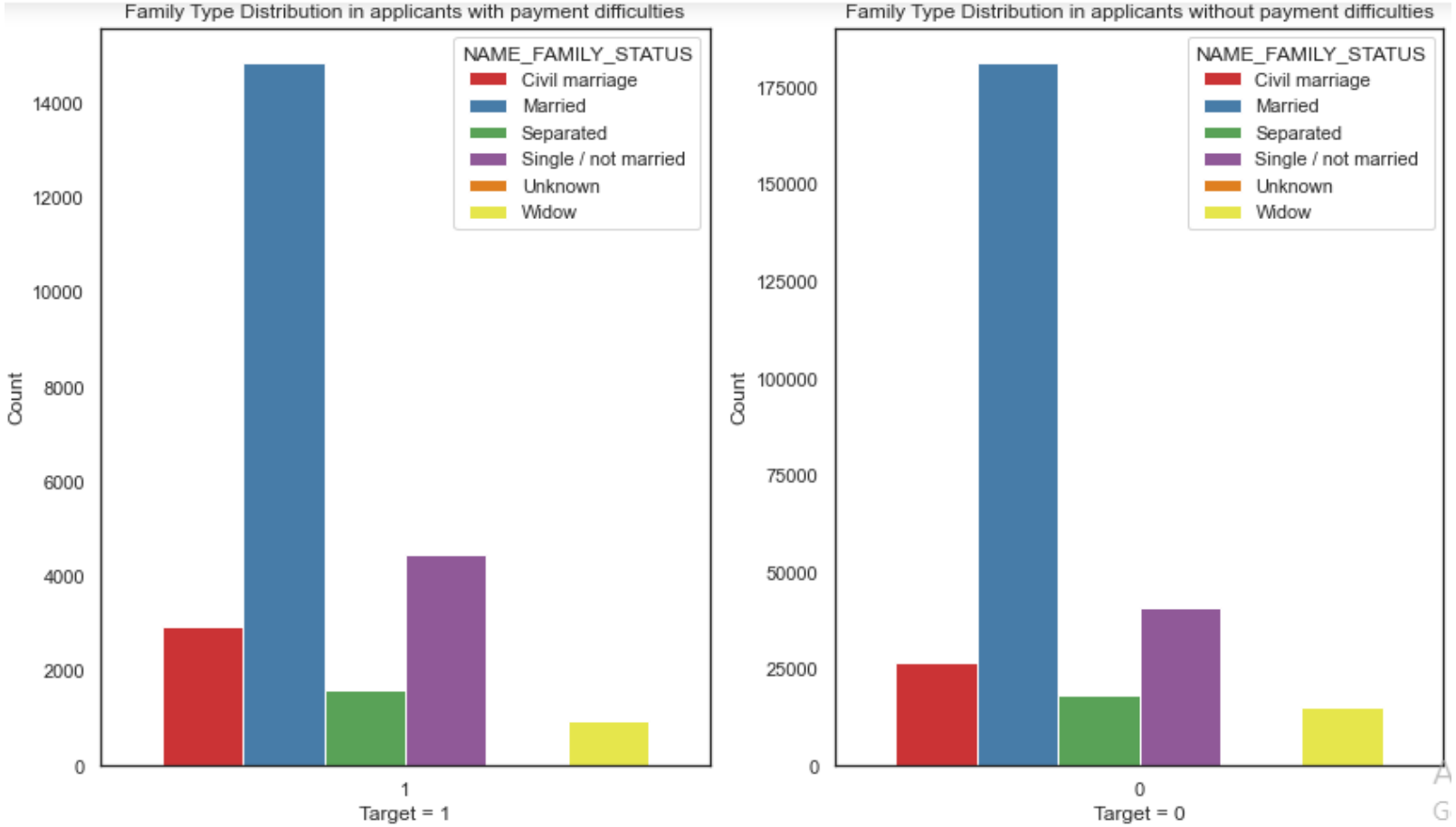
Distribution of Education variable in both populations of with/without payment difficulties.



- From the plot above, we can conclude that secondary/special educated people applying for loans are high in number in both categories and Academic degree educated are least.

Univariate Analysis – Categorical Variables (contd.)

Distribution of Family variable in both populations of with/without payment difficulties.

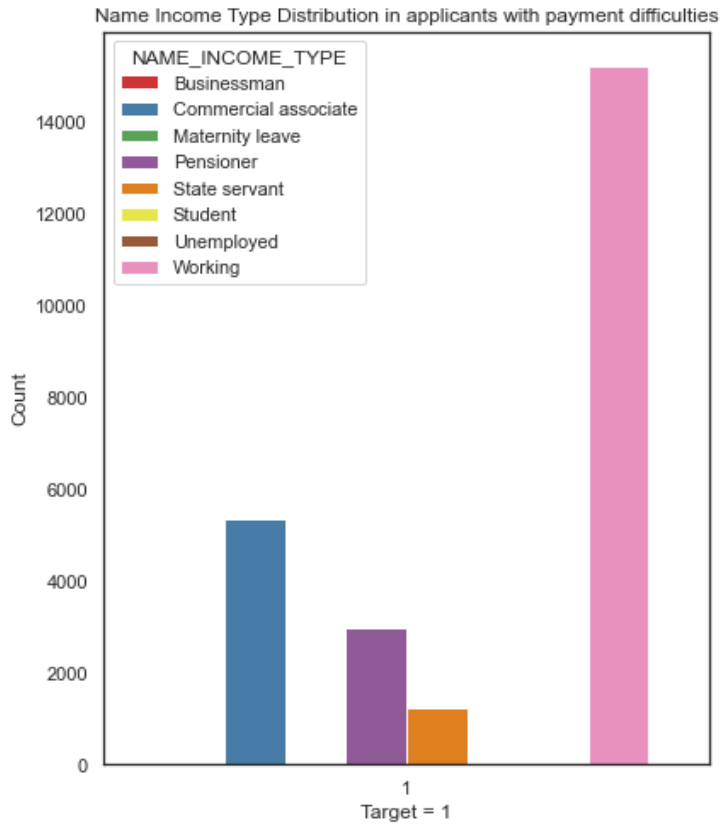


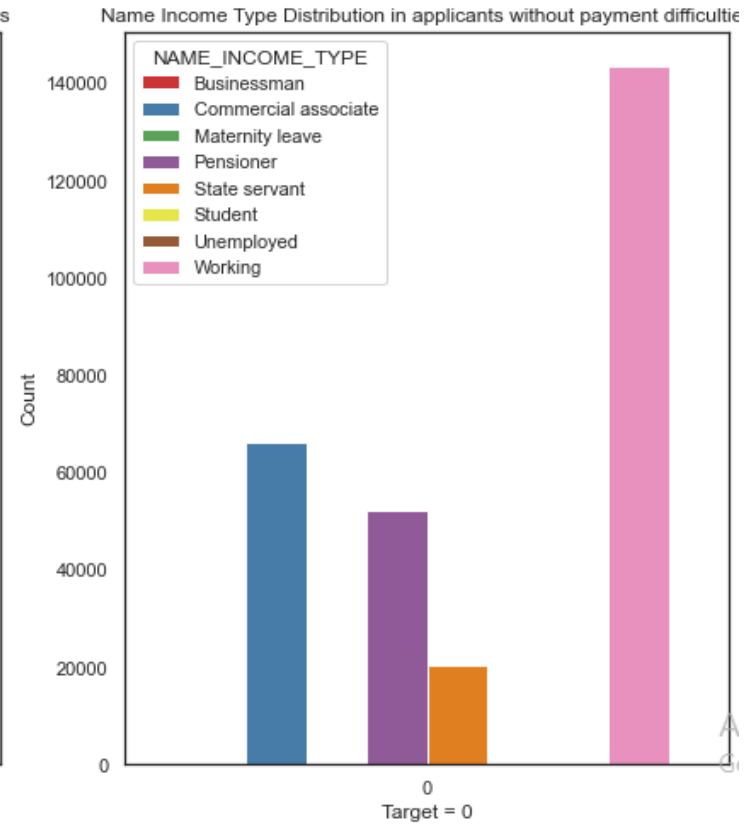
- The order of both default and not default customers is same.
- It also shows that there exists few(1 or 2) unknown values in not default client family status.
- We can say more married people tend to take more Loan as compared to other categories and being married is not impacting default and not defaulting

Univariate Analysis – Categorical Variables (contd.)

Distribution of Type of Income in both populations of with/without payment difficulties.

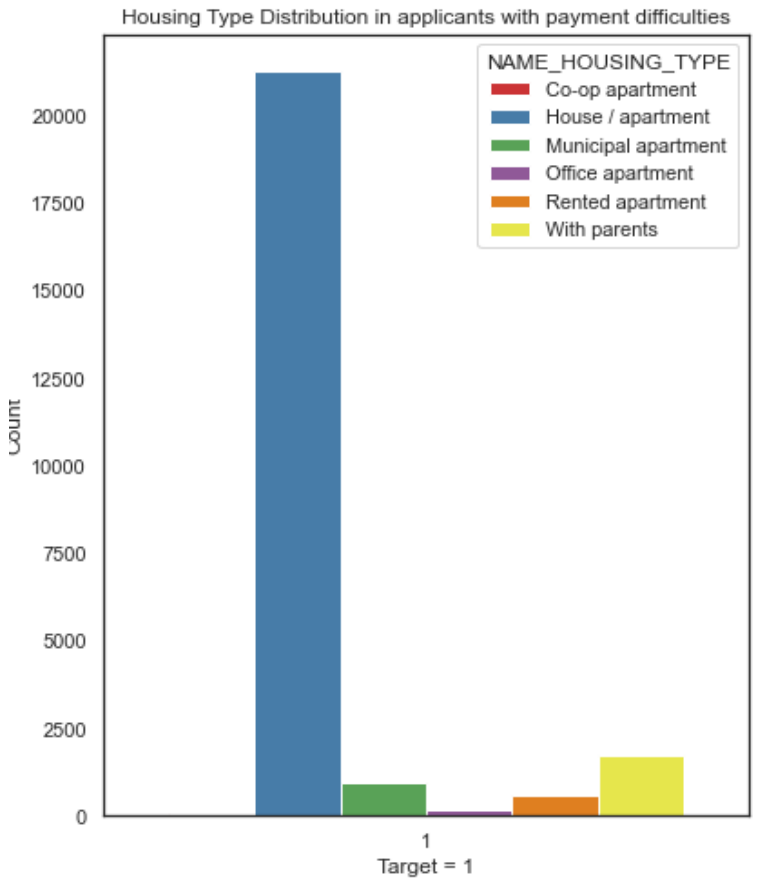
- From the graphs below, we can conclude that Pensioner of not default case are high in number compared to Pensioner of default case.
- It seems there exists both loss and profit due to Pension people to the Bank.
- It also shows that majority of defaulters income type is working and at the same time there is good income to bank from working people..

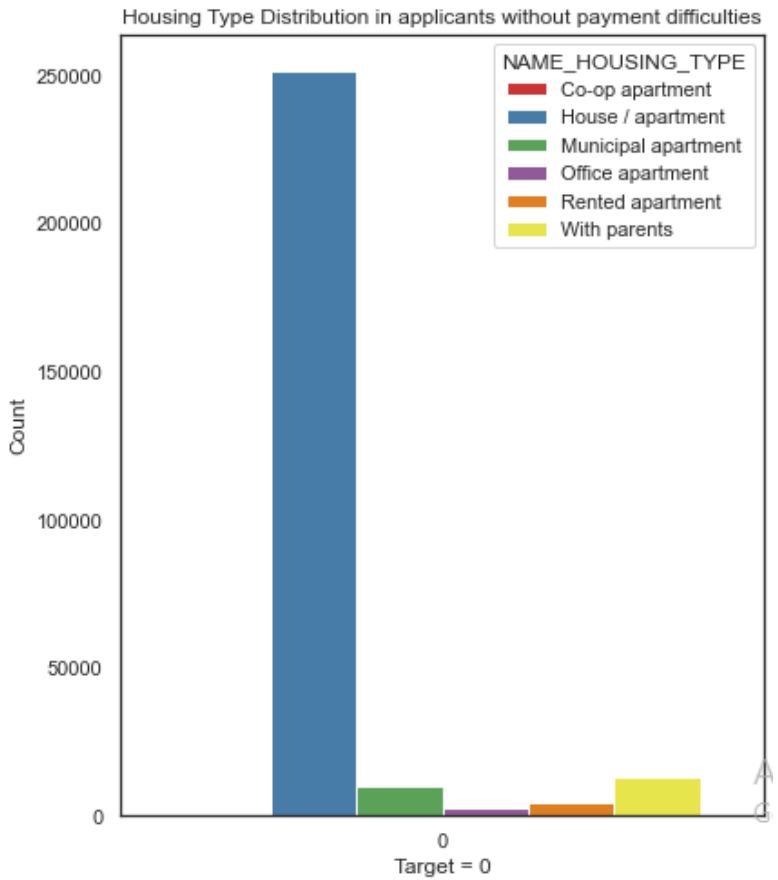




Univariate Analysis – Categorical Variables (contd.)

Distribution of Type of Housing in both populations of with/without payment difficulties.



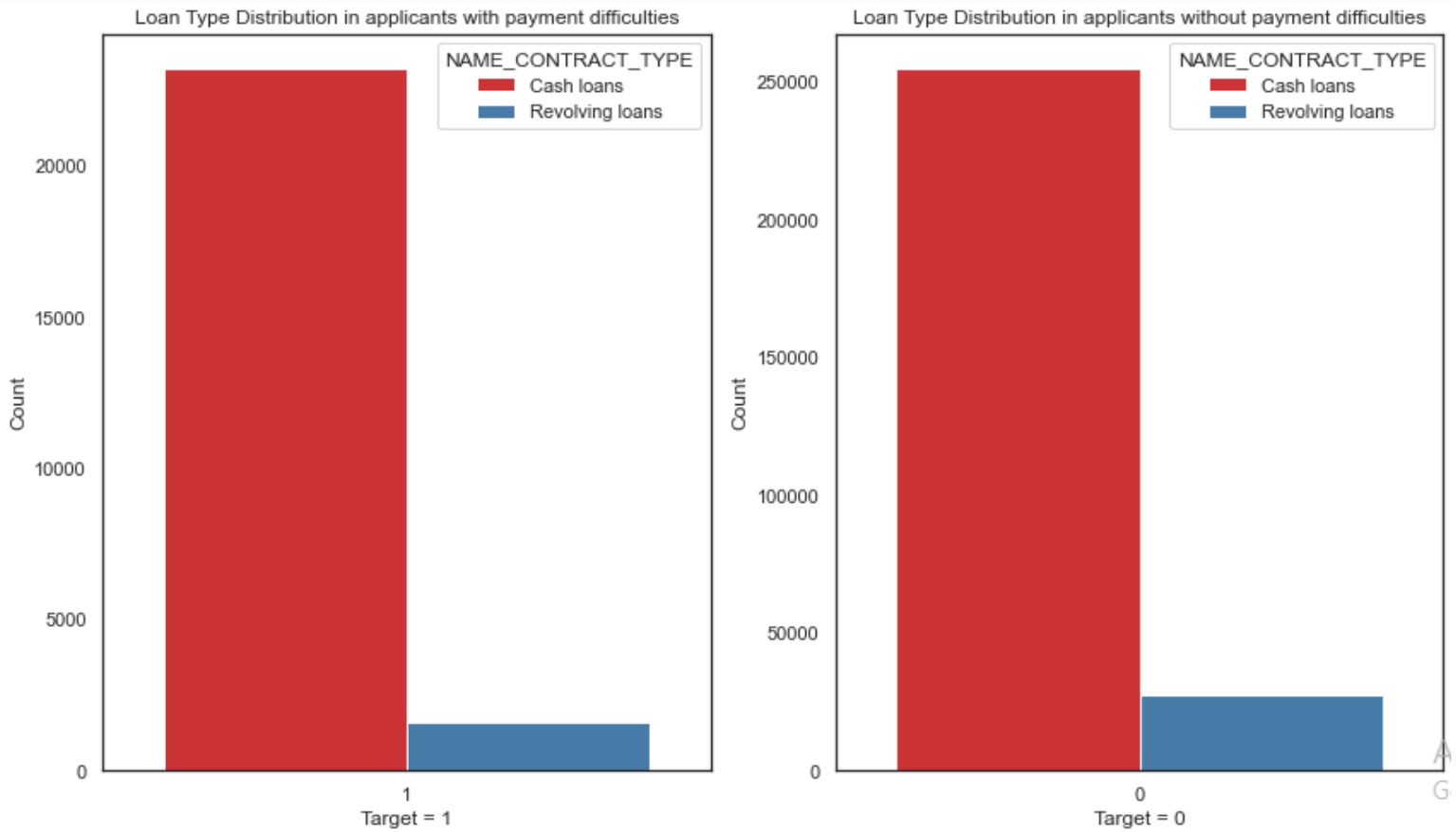


- From the graphs below, we can conclude that Pensioner of not default case are high in number compared to Pensioner of default case.
- It seems there exists both loss and profit due to Pension people to the Bank.
- It also shows that majority of defaulters income type is working and at the same time there is good income to bank from working people..

Univariate Analysis – Categorical Variables (contd.)

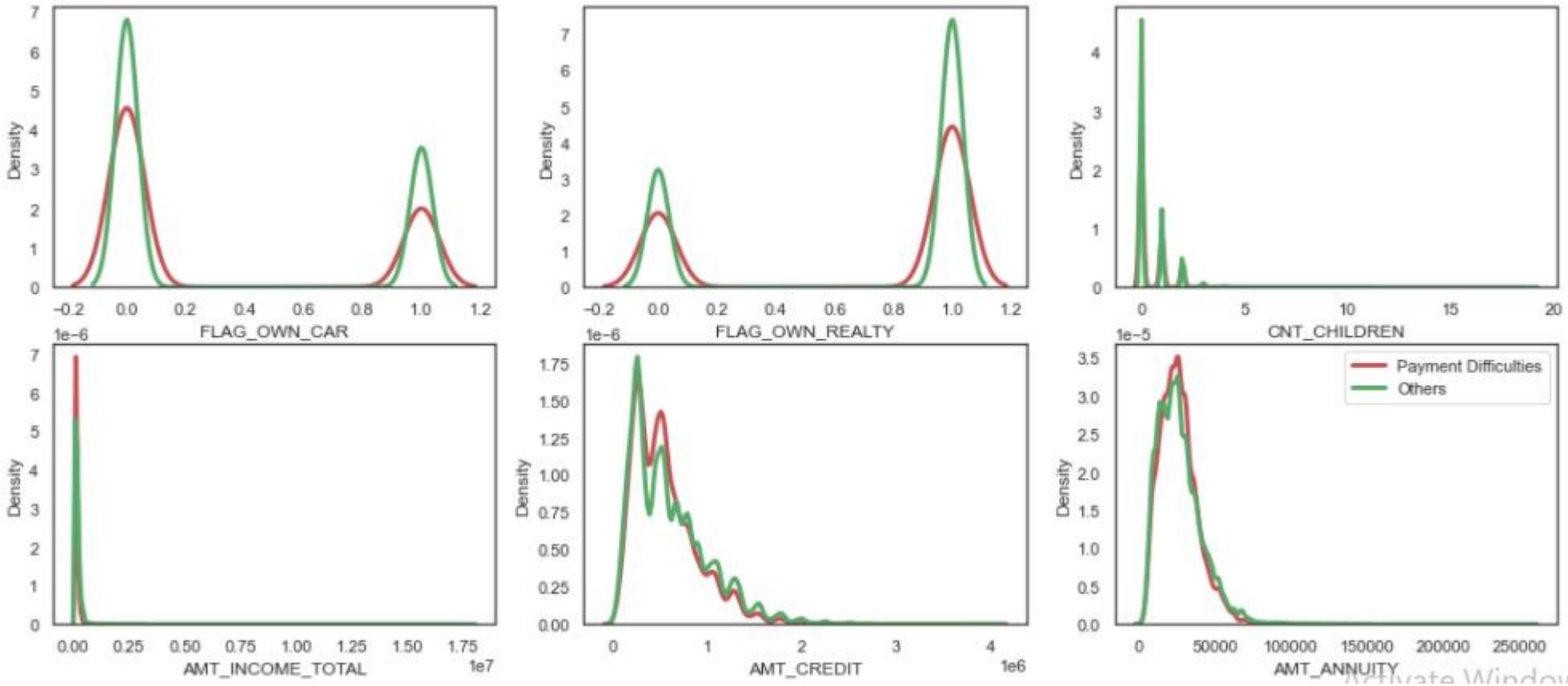
Distribution of Type of Loan in both populations of with/without payment difficulties.

- People tend to take more cash loans, and default percentage of revolving loans is low compared to cash loans



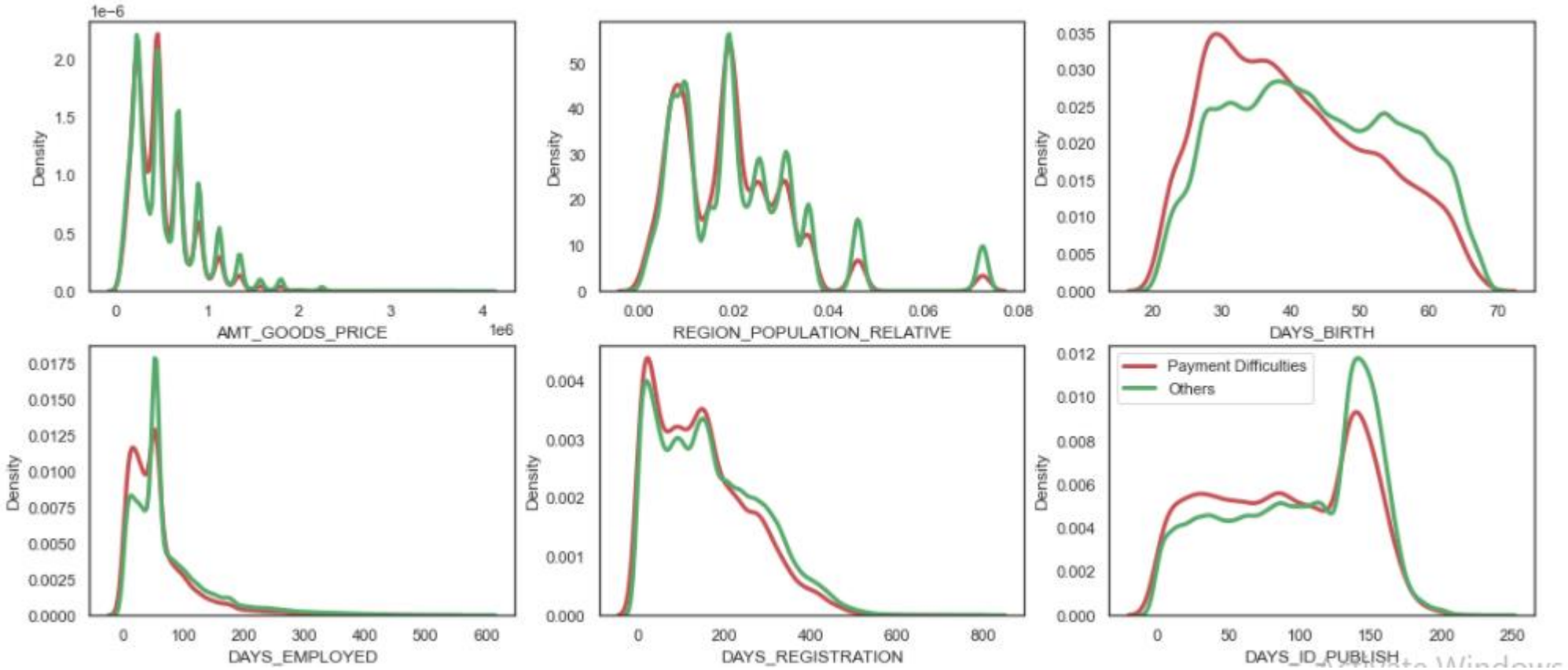
Univariate Analysis – Numerical Variables

KDE Plots of Numerical columns in both populations of with/without payment difficulties



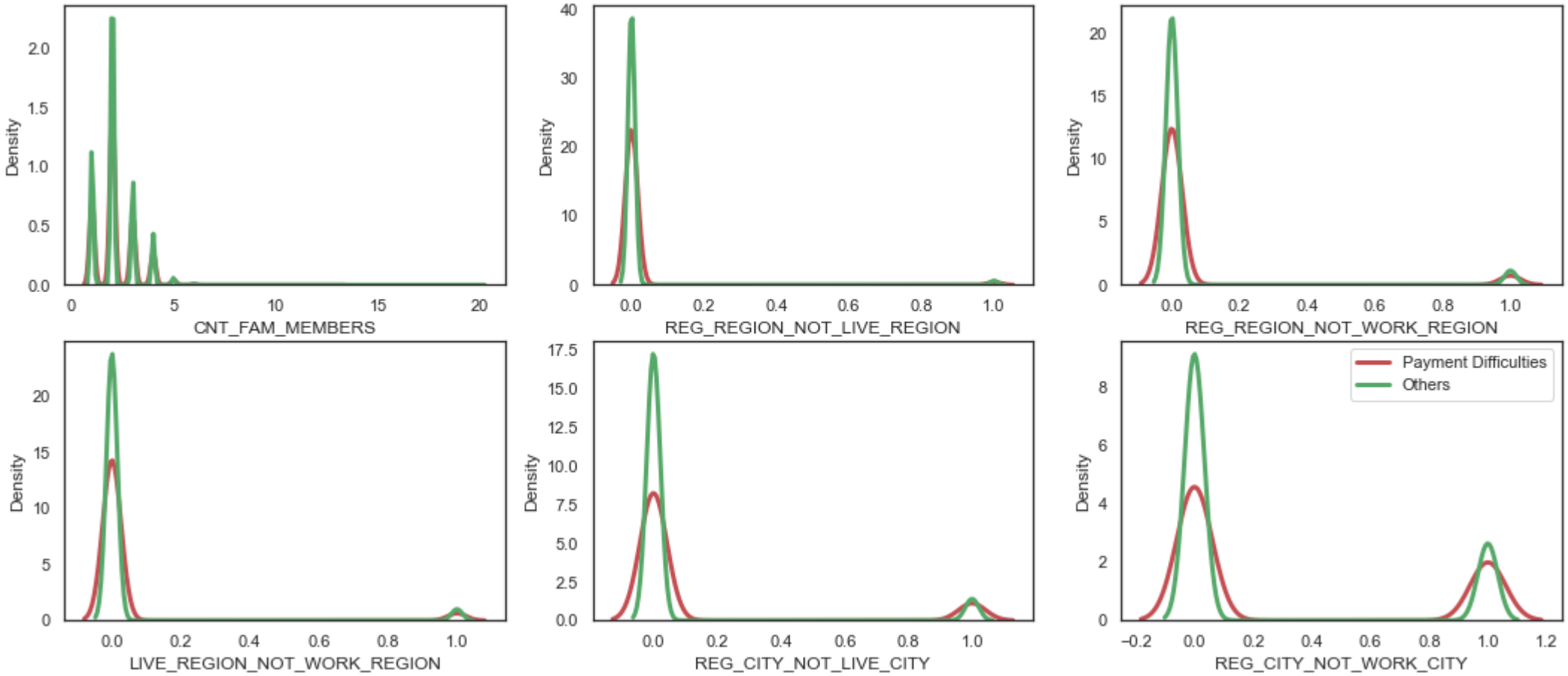
Univariate Analysis – Numerical Variables (contd.)

KDE Plots of Numerical columns in both populations of with/without payment difficulties



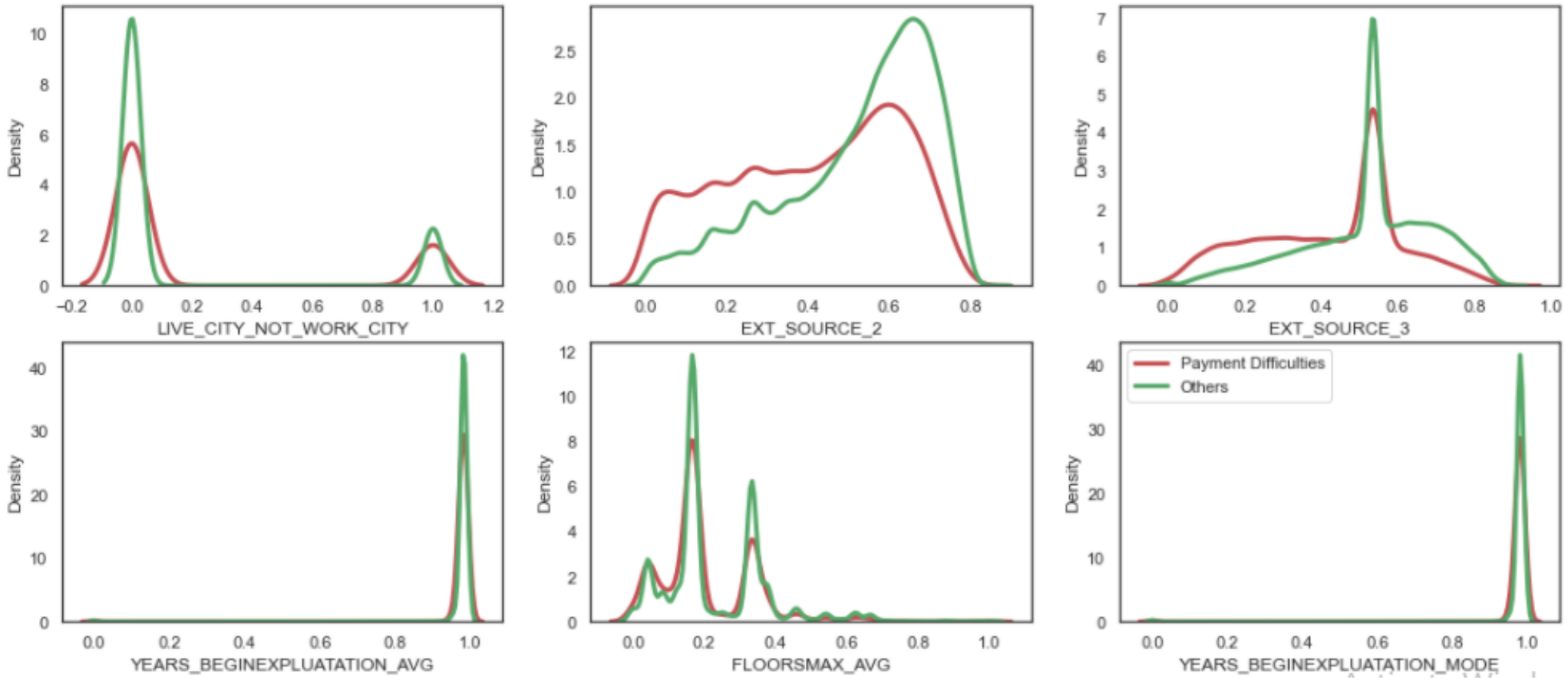
Univariate Analysis – Numerical Variables (contd.)

KDE Plots of Numerical columns in both populations of with/without payment difficulties



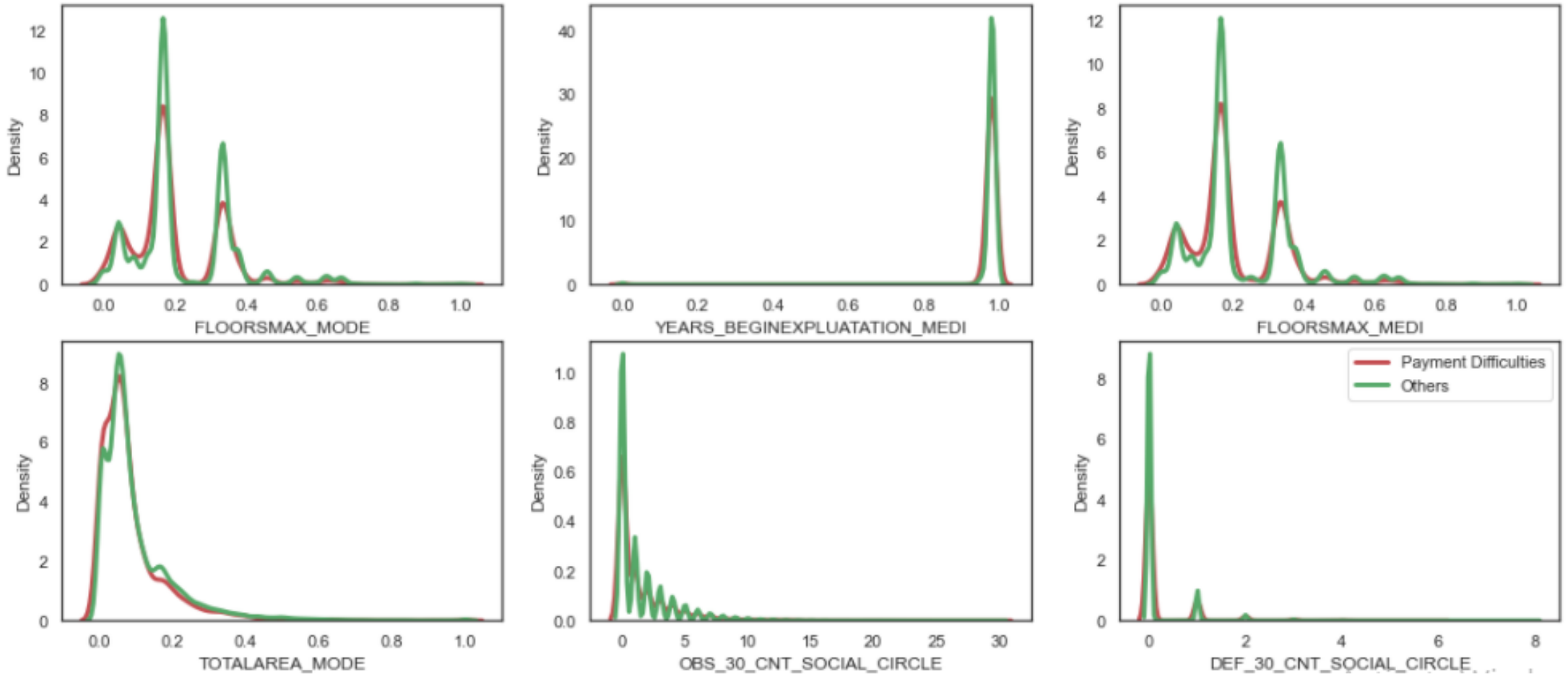
Univariate Analysis – Numerical Variables (contd.)

KDE Plots of Numerical columns in both populations of with/without payment difficulties



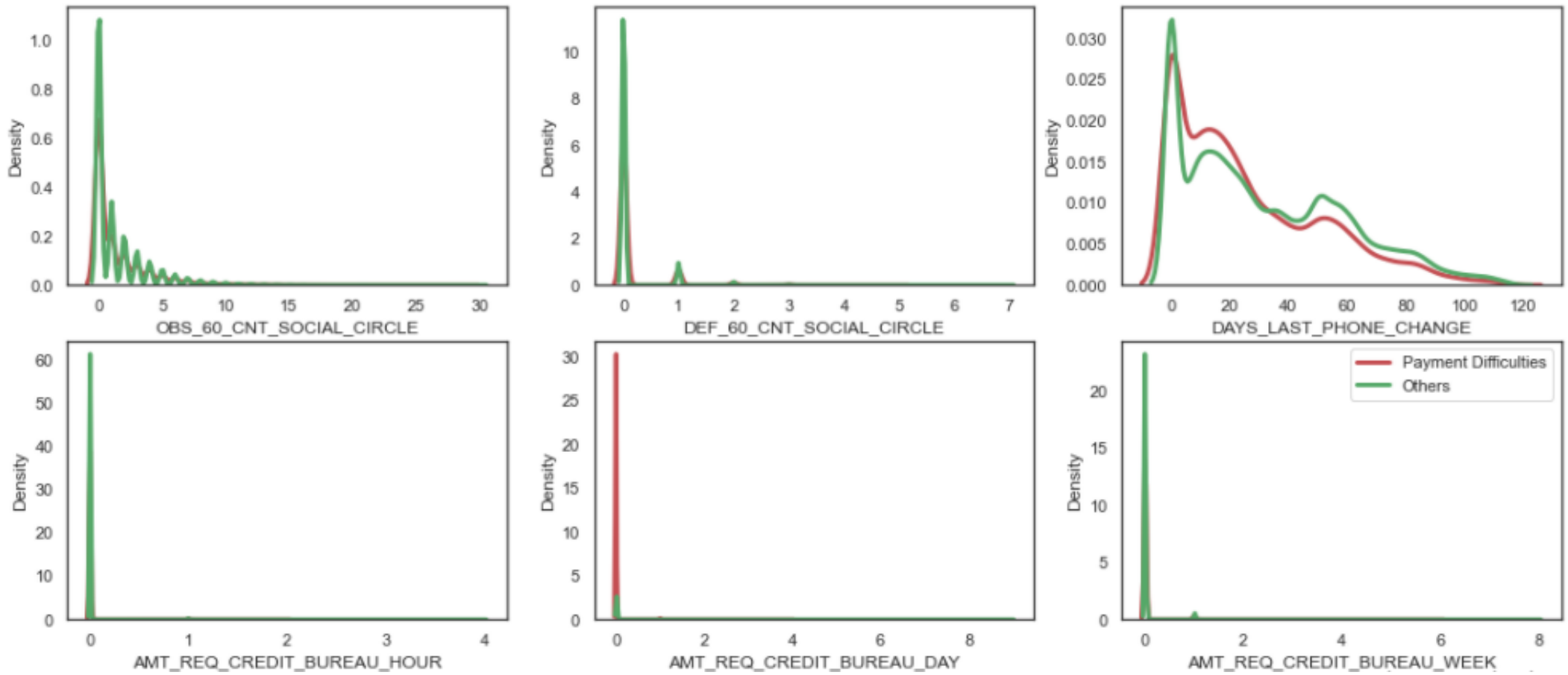
Univariate Analysis – Numerical Variables (contd.)

KDE Plots of Numerical columns in both populations of with/without payment difficulties



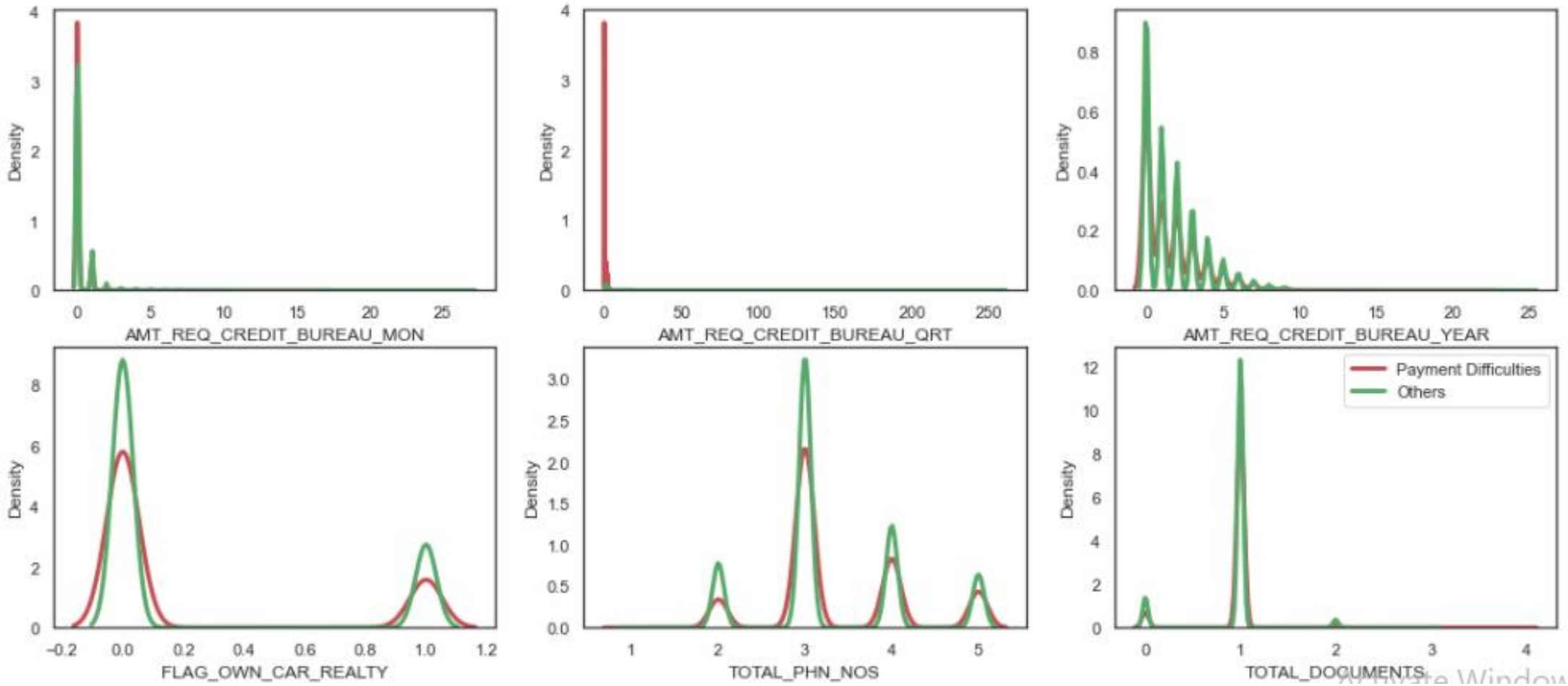
Univariate Analysis – Numerical Variables (contd.)

KDE Plots of Numerical columns in both populations of with/without payment difficulties



Univariate Analysis – Numerical Variables (contd.)

KDE Plots of Numerical columns in both populations of with/without payment difficulties



Univariate Analysis – Numerical Variables (contd.)

Key Insights

- Applicants who don't have a car are more in both the categories of with/without payment difficulties.
- This case is reverse in terms of ownership of a house/apartment. That is, applicants who have their own house makes a higher proportion in the case of both defaulters and non-defaulters.
- Most defaulters are typically in their early 30’s.
- Applicants who have work experience between 40-50 months typically experience paying difficulties compared to other applicants.
- From the density curve of CNT_CHILDREN we observe that the count of the applicant's children do not have any influence on the repaying capabilities of the applicant.

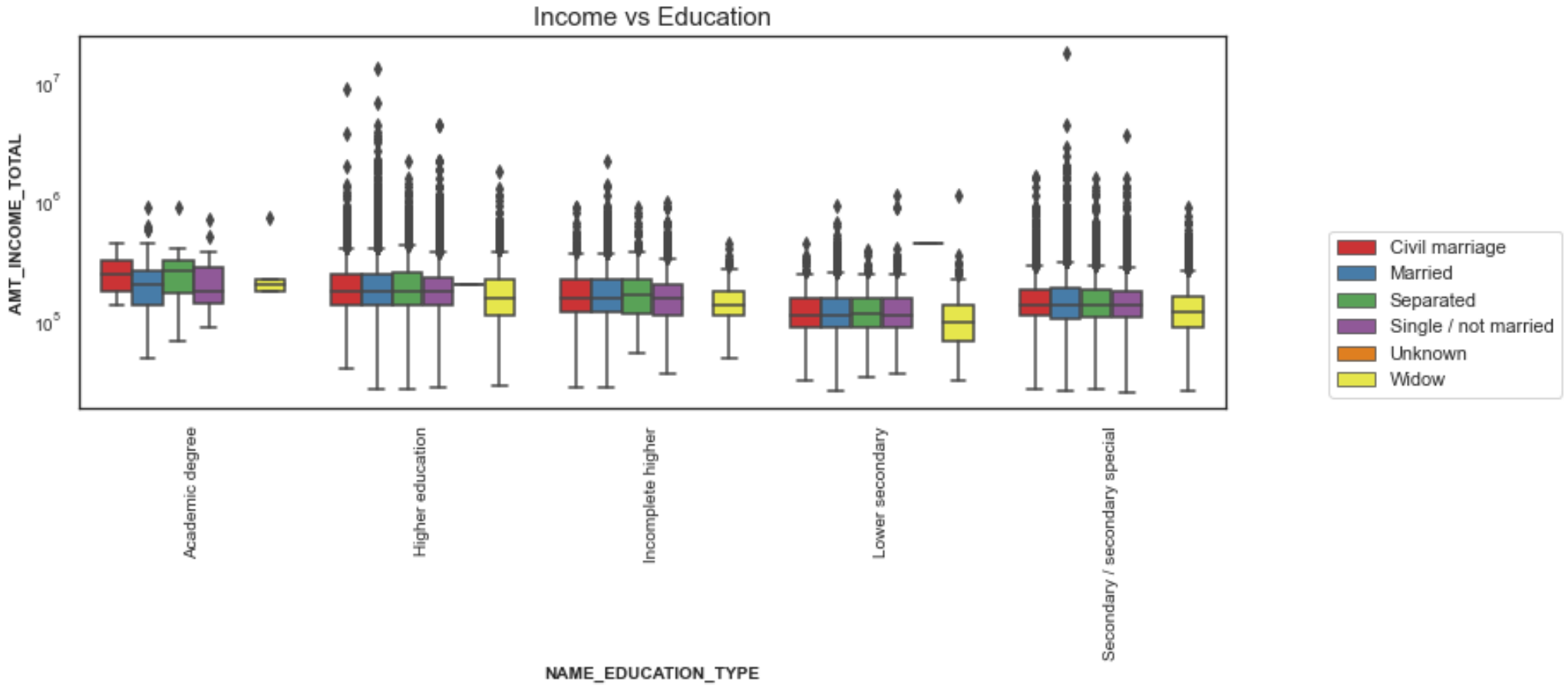
Univariate Analysis – Numerical Variables (contd.)

Key Insights

- Applicants with lower total income are more likely to default
- Applicants who just got employed tends to take more loans
- Applicants who are retired tends to take more loans
- Applicants whose id(s) got published between 130 and 150 months ago tend to take more loans
- Applicants with normalized score between 0.4 and 0.6 is high in case of score pooled from EXT_SOURCE_3. Also, there is no representation of applicants with normalized score above 0.8 in the case of EXT_SOURCE_2.

Bivariate Analysis – Applicants with no payment difficulties

Effect of Income , Education and Family Status



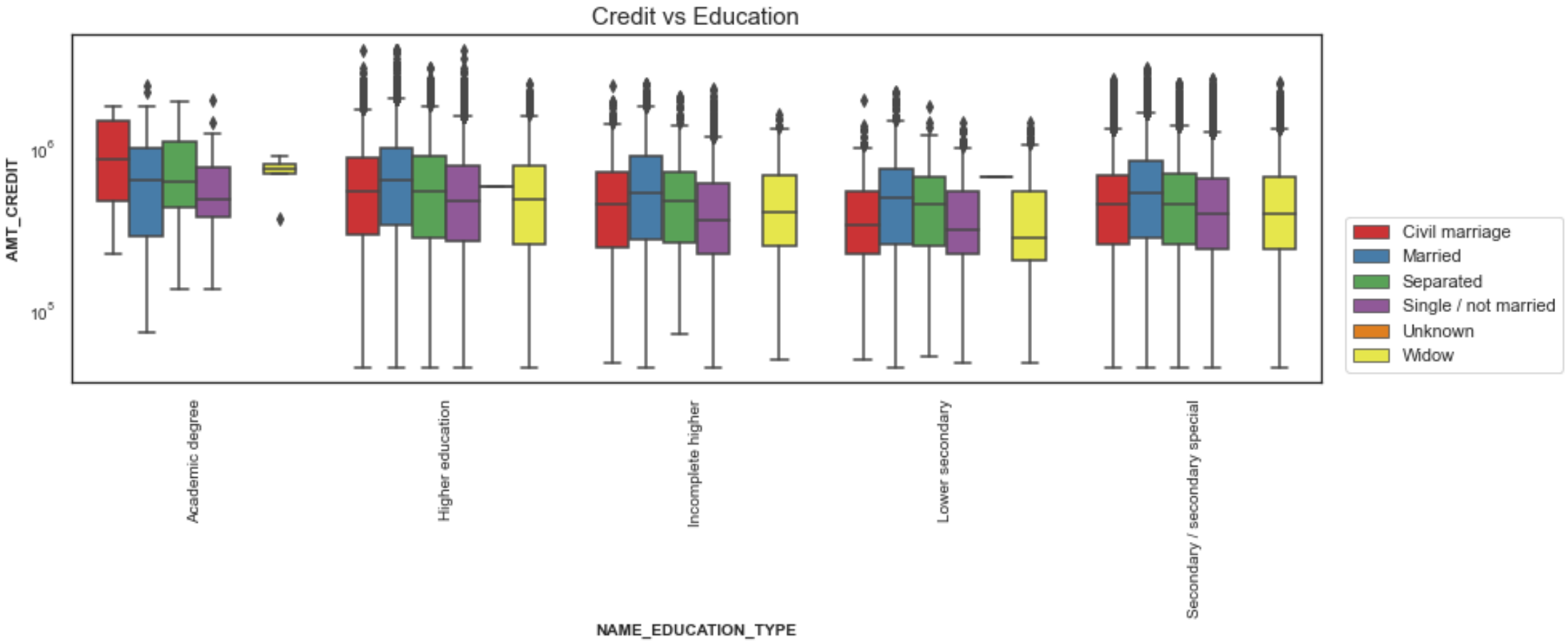
Bivariate Analysis – Applicants with no payment difficulties (contd.)

Effect of Income , Education and Family Status

- Applicants with Academic degree have a very few outliers as compared to other types of education.
- Applicants who have Higher Education, Incomplete Higher Education, Lower Secondary Education and Secondary/Secondary Special have a higher number of outliers.
- As expected, from the above figure it is evident that, applicants who have Higher Education tend to have highest Income compared to others. Also, some of the applicants who haven't completed their Higher Education tend to have higher Income.

Bivariate Analysis – Applicants with no payment difficulties (contd.)

Effect of Credit Amount, Education status and Family status



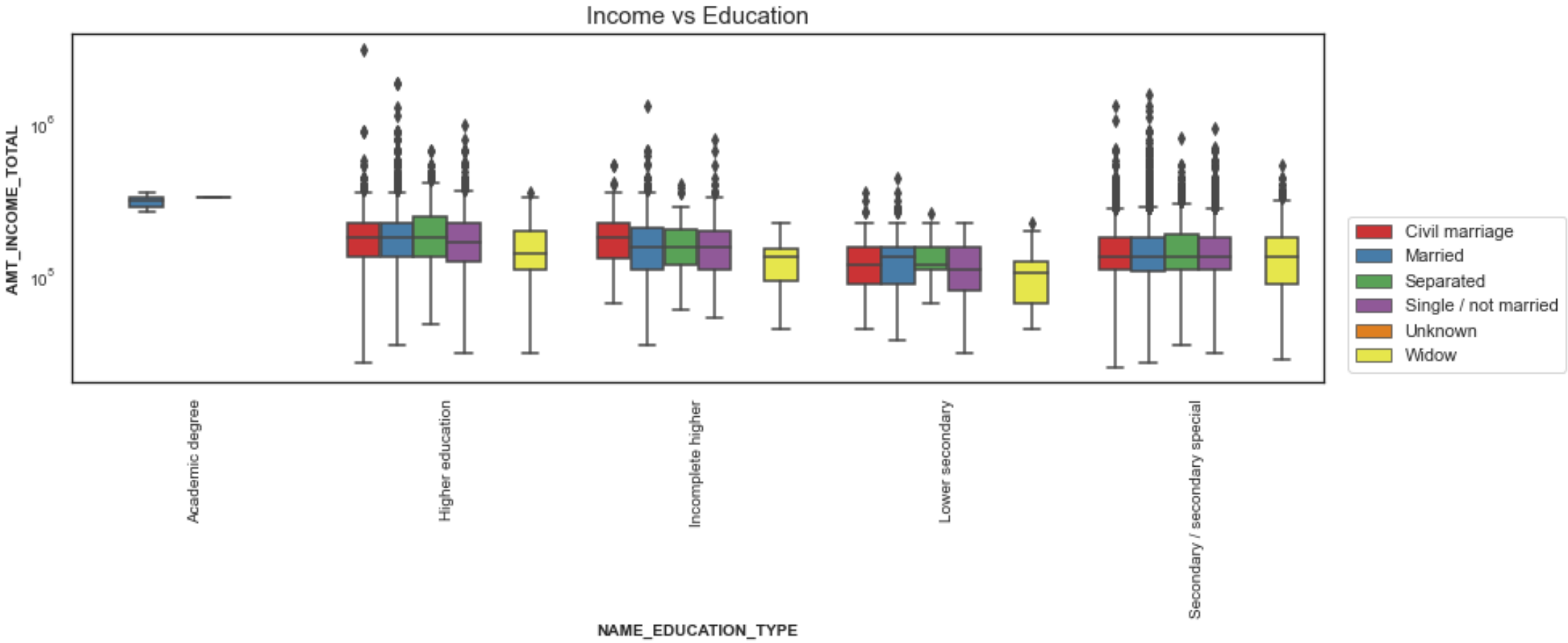
Bivariate Analysis – Applicants with no payment difficulties (contd.)

Effect of Credit Amount, Education status and Family status

- Applicants with all Education types except Academic degree have large number of outliers.
- Applicants with Academic degree and who did Civil Marriage tend to take higher credit loan.
- A very few applicants with Higher Education, Incomplete Higher Education, Lower Secondary Education and Secondary/Secondary Special Education are more likely to take high amount of credit loan(outliers).

Bivariate Analysis – Applicants with payment difficulties

Effect of Income , Education and Family status



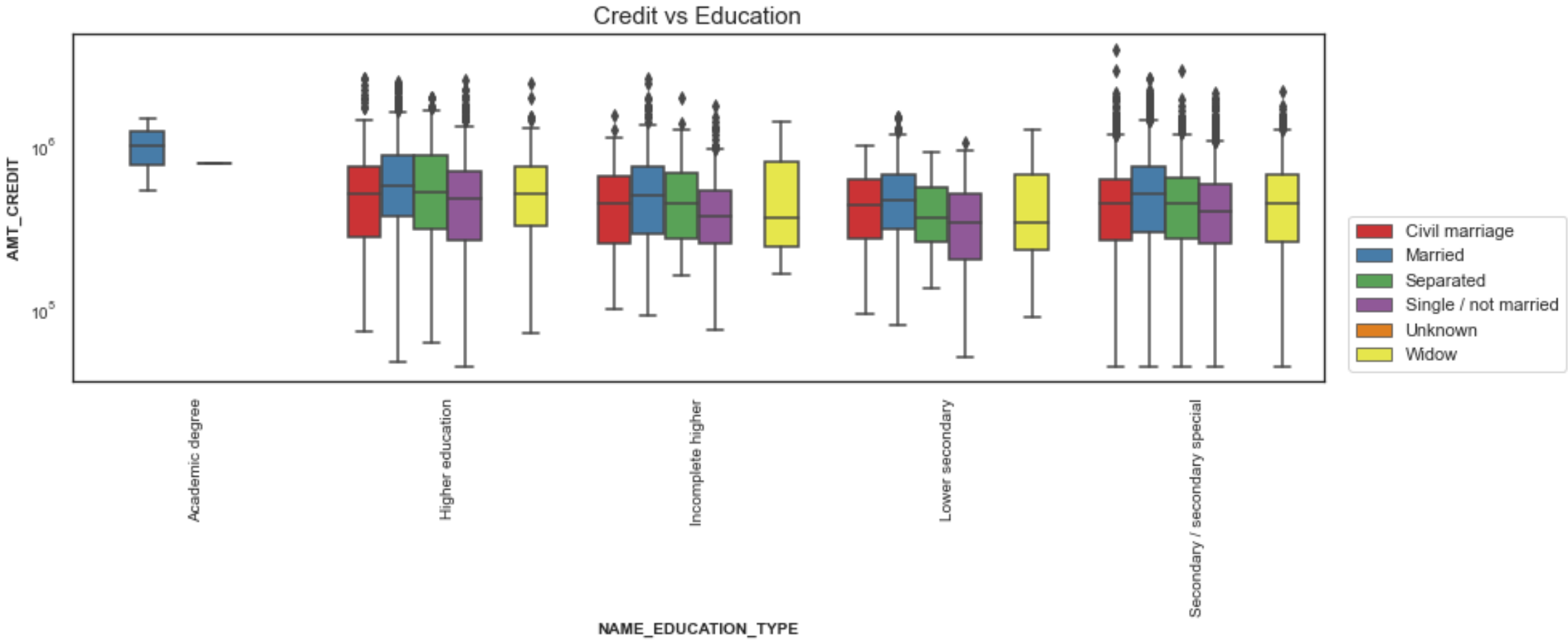
Bivariate Analysis – Applicants with payment difficulties (contd.)

Effect of Income , Education and Family status

- There are very few applicants with Academic degree and payment difficulties. These population has a higher income compared to other education degrees.
- Applicants who have Higher Education, Incomplete Higher Education, Lower Secondary Education and Secondary/Secondary Special have a higher number of outliers.
- Applicants who have payment difficulties and have Secondary/Secondary education have higher income among other education groups

Bivariate Analysis – Applicants with payment difficulties (contd.)

Effect of Credit Amount, Education status and Family status

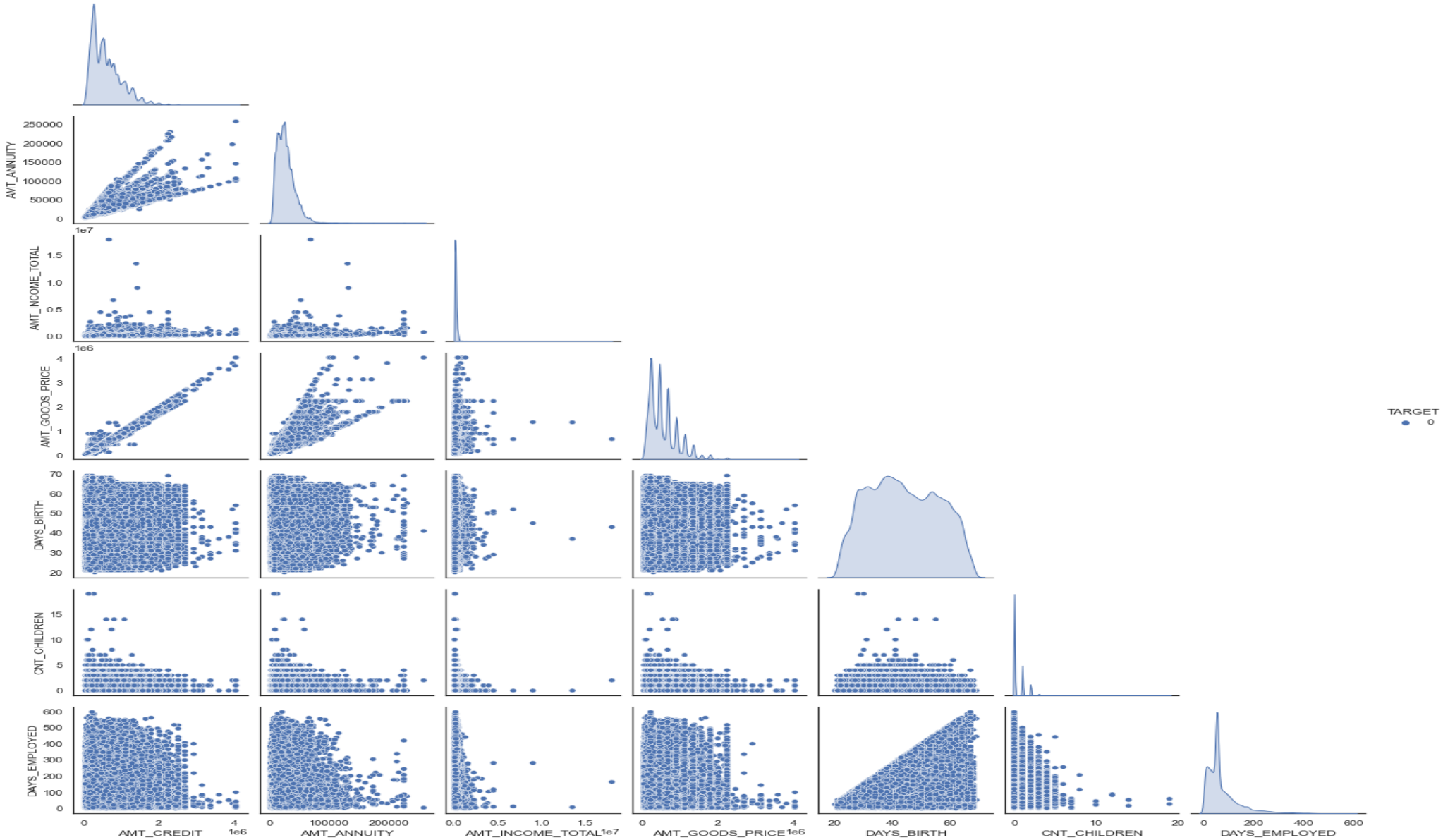


Bivariate Analysis – Applicants with payment difficulties (contd.)

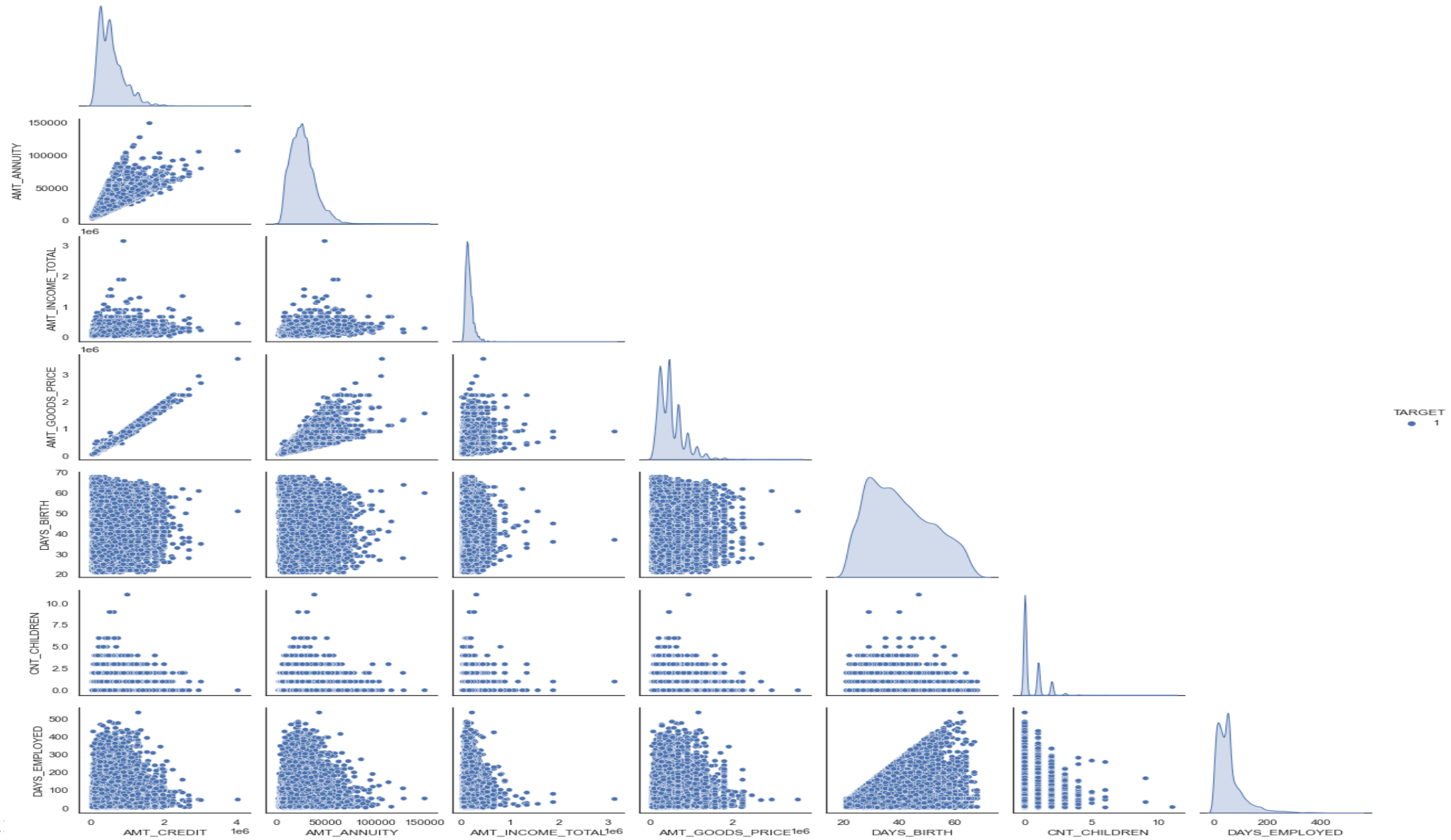
Effect of Credit Amount, Education status and Family status

- There are very few applicants who have an academic degree and went for a higher loan facing payment difficulties
- Applicants who have Higher Education, Incomplete Higher Education, Lower Secondary Education and Secondary/Secondary Special have a higher number of outliers.
- Applicants who have payment difficulties and have Secondary/Secondary education went for a higher loan among other education groups

Pair plots- Applicants with no payment difficulties



Pair plots- Applicants with payment difficulties



Correlations

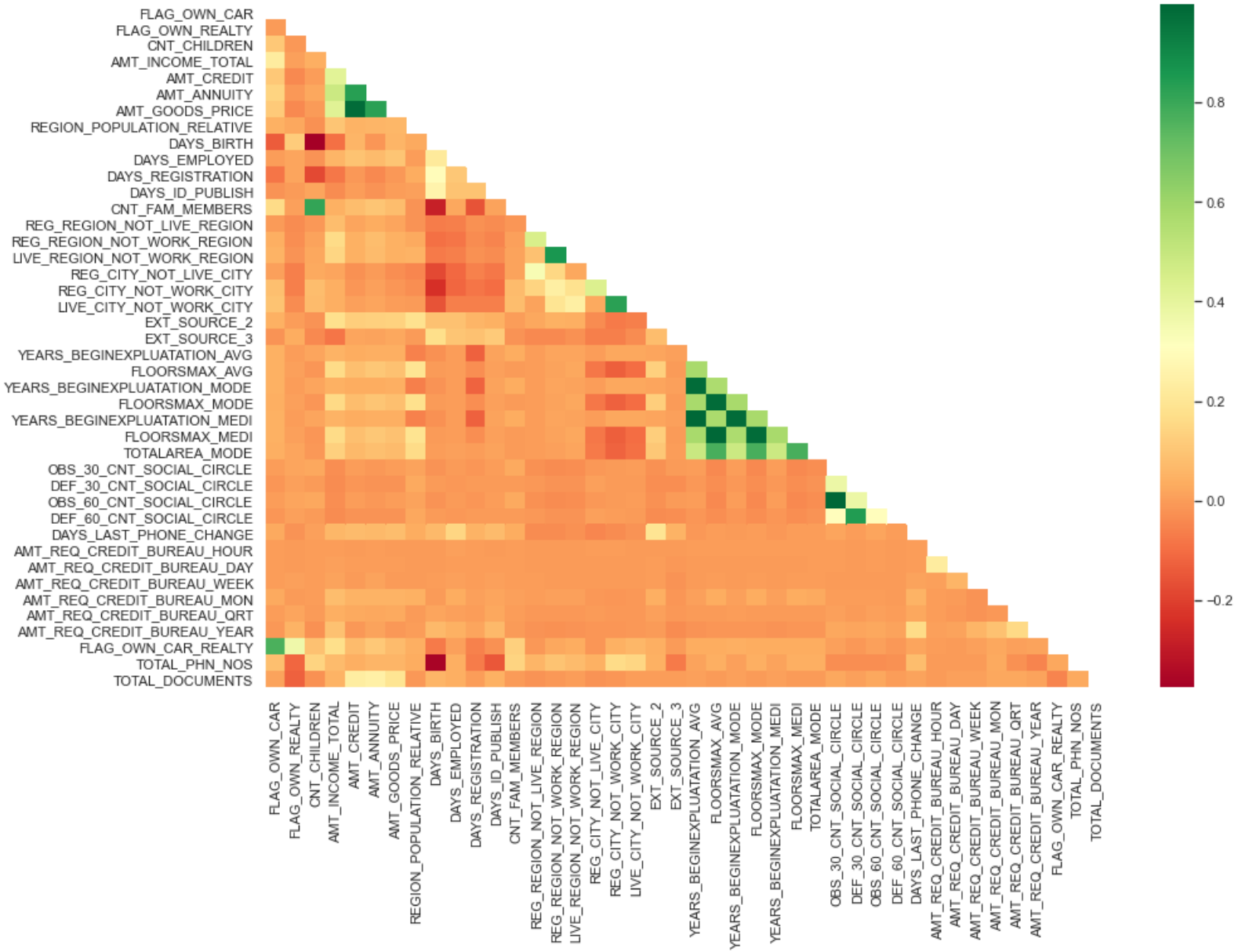
Correlations between numerical variables Using "Pair Plots "

- AMT_CREDIT and AMT_GOODS_PRICE are highly correlated variables for both defaulters and non - defaulters. So as the home price increases the loan amount also increases

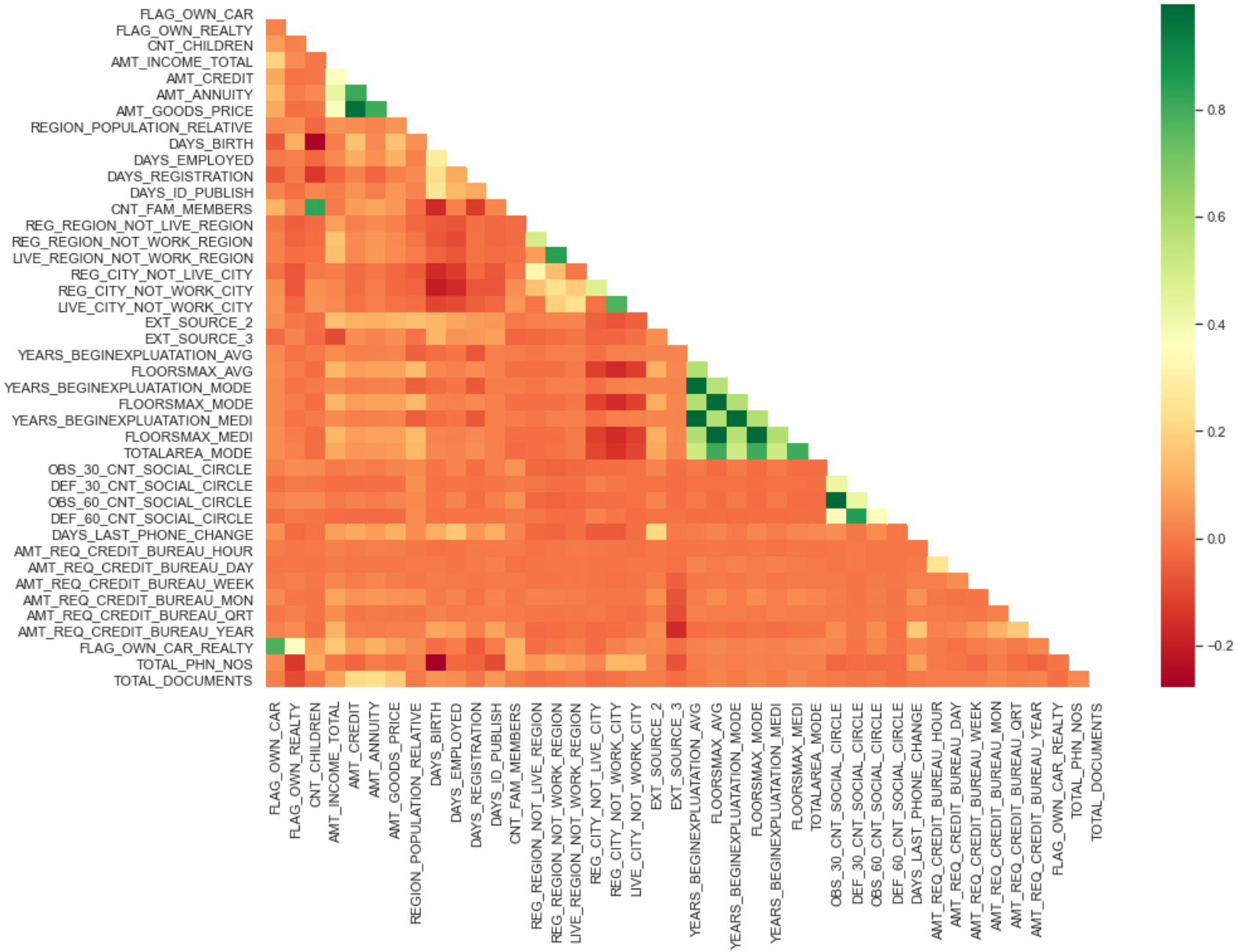
- AMT_CREDIT and AMT_ANNUITY (EMI) are highly correlated variables for both defaulters and non - defaulters . So as the home price increases the EMI amount also increases which is logical

- All three variables AMT_CREDIT, AMT_GOODS_PRICE and AMT_ANNUITY are highly correlated for both defaulters and non-defaulters, which might not give a good indicator for defaulter detection

Correlation Heatmap - Applicants with no payment difficulties



Correlation Heatmap - Applicants with payment difficulties



Correlations

Key Insights which are common to both populations with/without payment difficulties

- AMT_CREDIT is inversely proportional to the DAYS_BIRTH , i.e., applicants who belong to low-age group take high Credit amount and vice-versa
- AMT_CREDIT is inversely proportional to the CNT_CHILDREN, i.e., Credit amount is higher for applicants with few children vice-versa.
- AMT_INCOME_TOTAL is inversely proportional to the CNT_CHILDREN, i.e., more income for applicants with few children vice-versa.
- Applicants with few children live in densely populated area.
- AMT_CREDIT is higher for applicants in densely populated area.
- AMT_INCOME_TOTAL is also higher in for applicants living in densely populated area.

Correlations

Top 10 Correlations in the dataset of Applicants without payment difficulties

Var1	Var2	Correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1
FLOORSMAX_MEDI	FLOORSMAX_AVG	1
YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	1
YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.99
FLOORSMAX_MEDI	FLOORSMAX_MODE	0.99
YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.99
FLOORSMAX_MODE	FLOORSMAX_AVG	0.98
AMT_GOODS_PRICE	AMT_CREDIT	0.98
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.85

Correlations

Top 10 Correlations in the dataset of Applicants with payment difficulties

Var1	Var2	Correlation
YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	1
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1
FLOORSMAX_MEDI	FLOORSMAX_AVG	0.99
YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.99
FLOORSMAX_MEDI	FLOORSMAX_MODE	0.99
YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.99
AMT_GOODS_PRICE	AMT_CREDIT	0.98
FLOORSMAX_MODE	FLOORSMAX_AVG	0.98
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.86
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.84

Inspecting Dataset

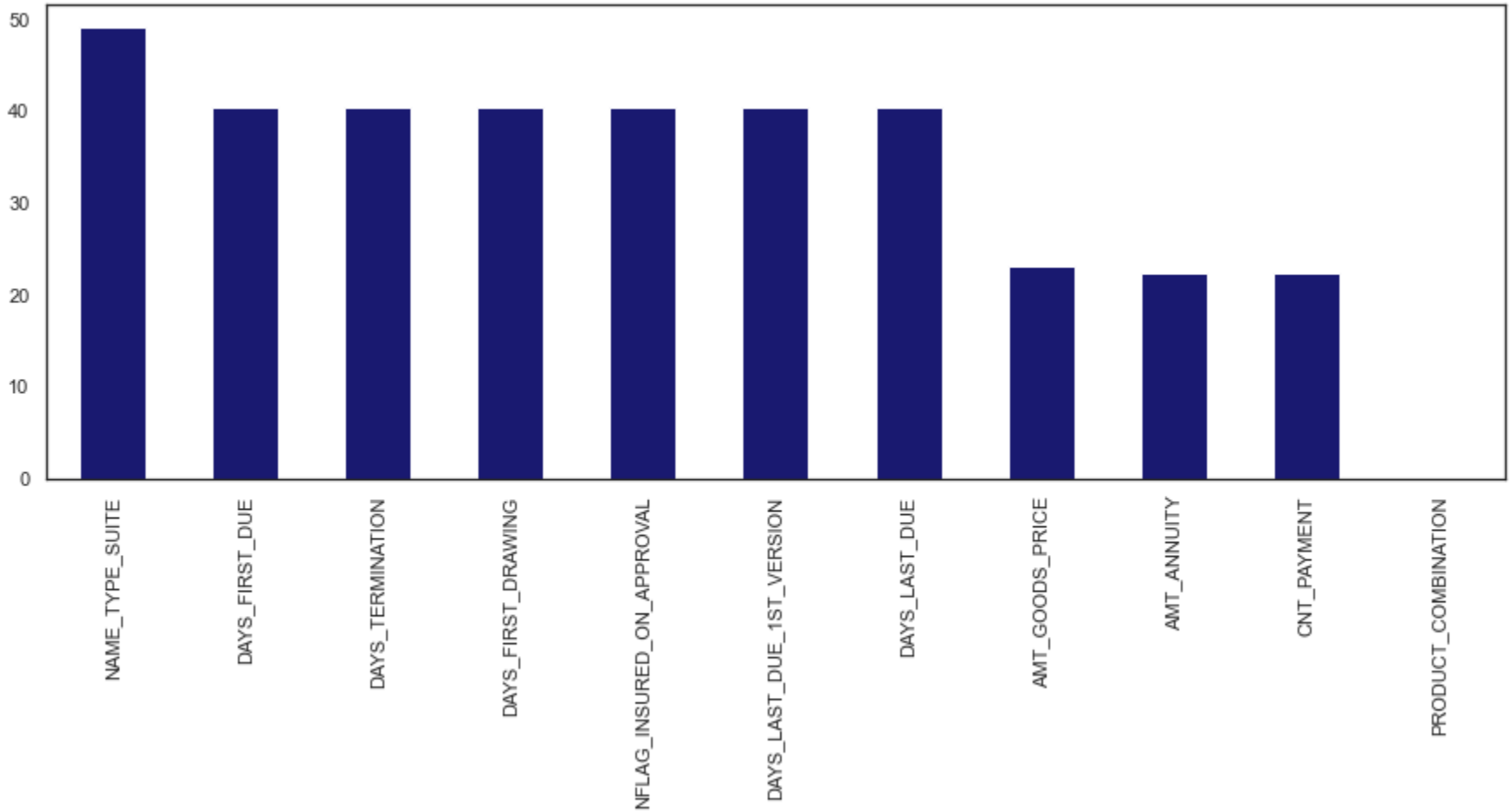
Given previous applications dataset contains information of applicants from their previous loans data

- The first dataset ‘previous_application.csv’ contains information about the client’s previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.
- A total of 37 attributes for 16,70,214 applicants with their previous application status is given in this csv file
- All the observations in this data set have a unique loan current id called “SK_CURR_ID” which is the id given by the bank to that applicant’s loan as well as the unique load ids of the applicant’s previous loans. The column “NAME_CONTRACT_STATUS” is the variable (has values Approved, Cancelled, Refused and Unused offer) which gives us the information of applicant’s previous loan
- A quick run of .info() method on this data frame shows that there are 15 float columns, 6 int columns and 16 object type columns.

Missing value analysis

Out of 37 columns in the application dataset, there are 15 columns with at least 1 missing value and 4 columns with at least 50% missing values.

- It is clear from the data dictionary that there are no Missing Column Names or Inconsistent column names. The columns with missing value percentage greater than 50 and dropped.



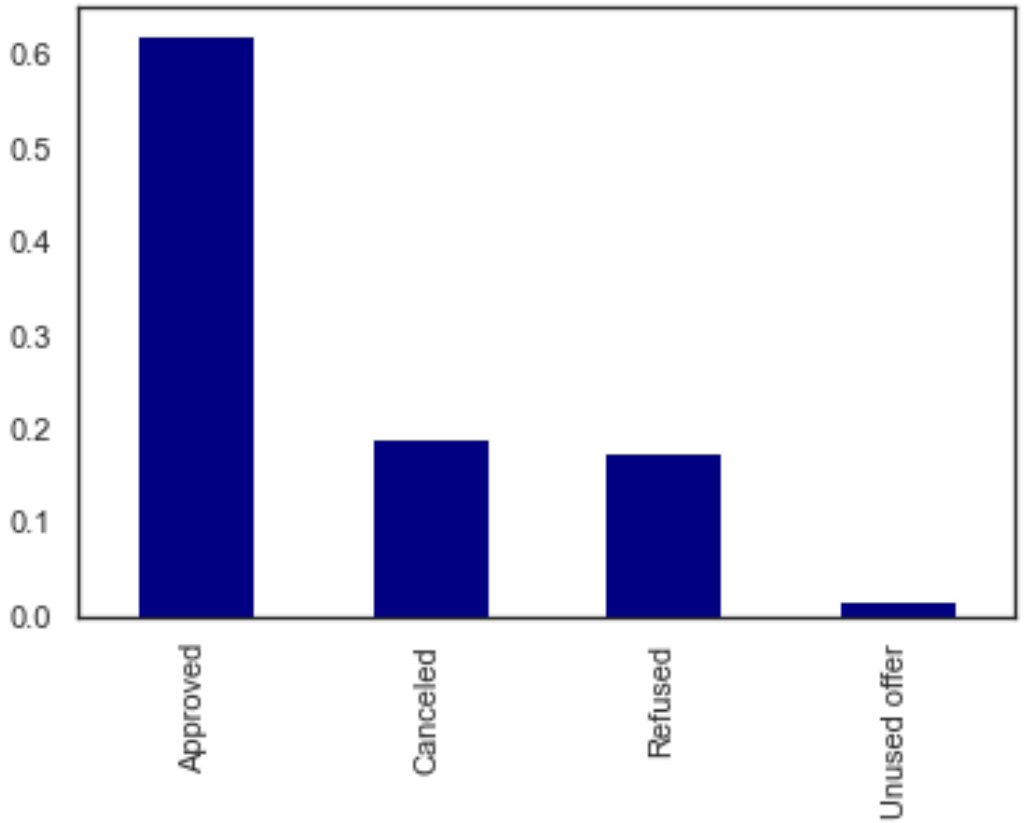
Data Cleaning

- The columns DAYS_DECISION, DAYS_FIRST_DUE, DAYS_LAST_DUE_1ST_VERSION, DAYS_LAST_DUE, DAYS_TERMINATION, DAYS_FIRST_DRAWING have values in the days with reference to the days in which the applicant has made his application.
- Hence, by definition all these columns have negative values. We've converted them to positive and changed their units.
- Relatively, the previous applications data is much cleaner when compared to the applications data, hence there was no need to any extensive cleaning for this dataset.

Analysis of NAME_CONTRACT_STATUS column

Majority of the loans (~60%) in this dataset are related to accepted loans.

- It is clear from the graph below that, the proportion of Cancelled and Refused loans is almost similar and Unused loans account for <5% in our dataset.

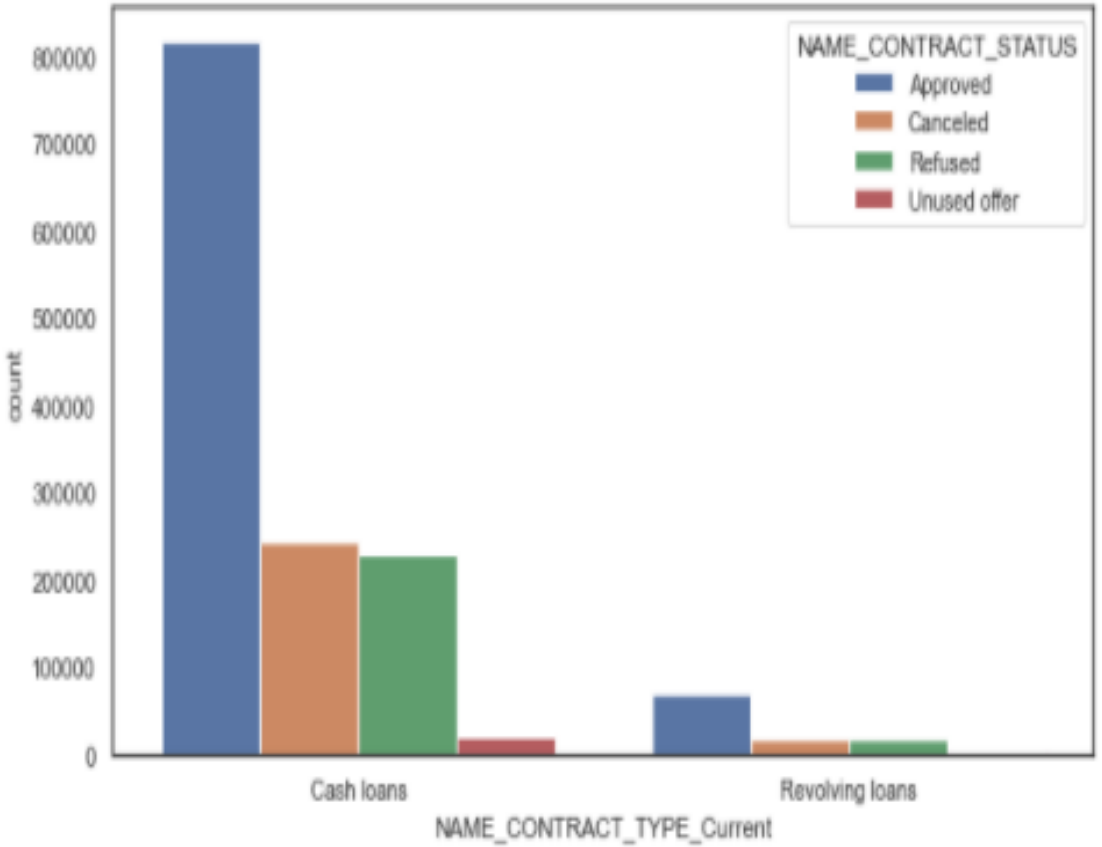


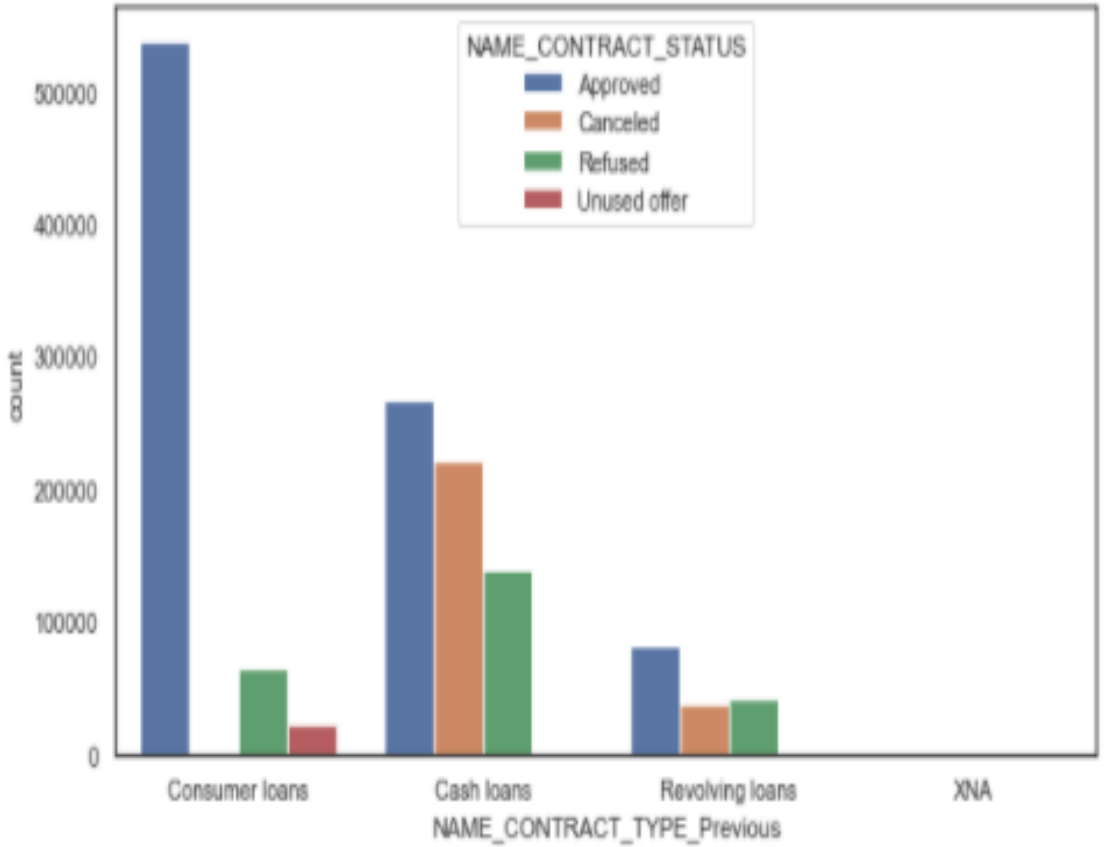
Merging current application dataset with previous applications dataset

- In order to bring previous loan attributes into our applicant attributes dataset, we did a left join on these both datasets
- As “SK_CURR_ID” is common to both the datasets, we used this variable as key to perform the left join
- While we are doing left join with, If we have the multiple copies of a same key in the merging data set then our final merged dataset contains an equal number of copies of same key.
- To distinguish the column names in our datasets, we use suffix values of “_Current”/ “_Previous” in the merging and base datasets respectively.
- After merging, our final dataset has a total of 14,28,946 applications along with 89 columns.

Univariate Analysis – Categorical columns

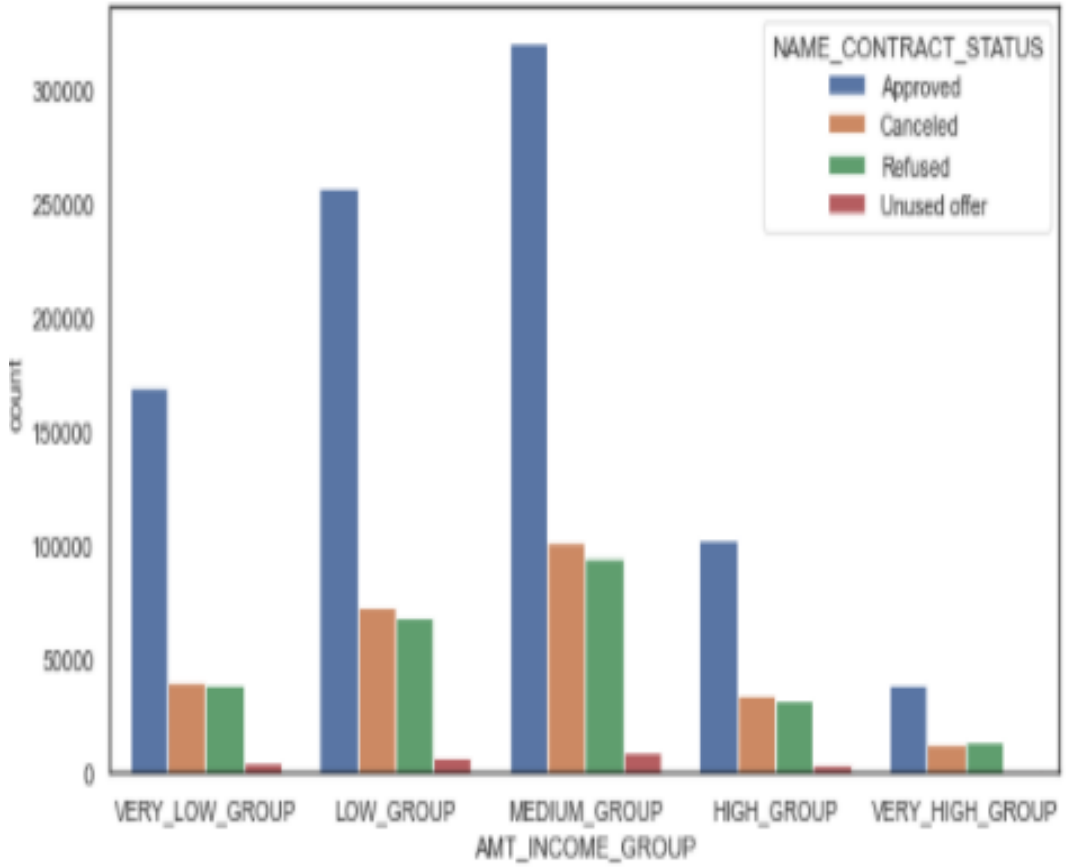
We look at the distribution of various categorical columns w.r.t the NAME_CONTRACT_STATUS

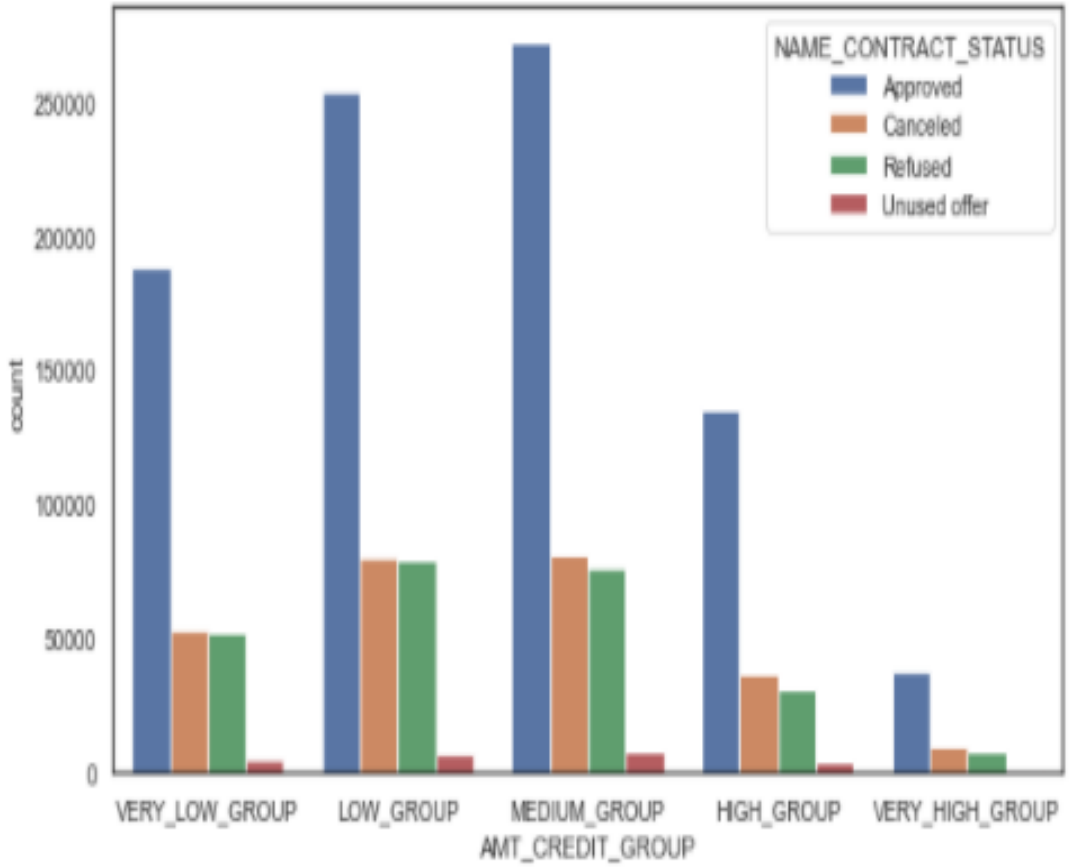




Univariate Analysis – Categorical columns (contd.)

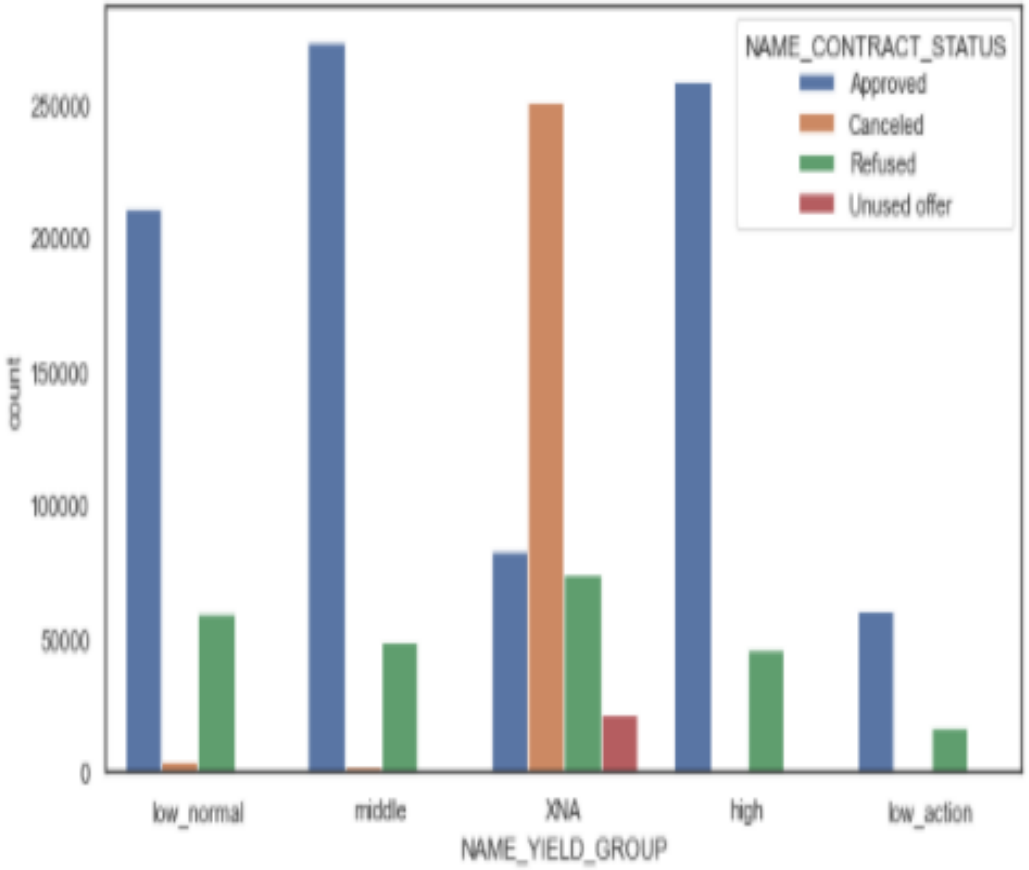
We look at the distribution of various categorical columns w.r.t the NAME_CONTRACT_STATUS

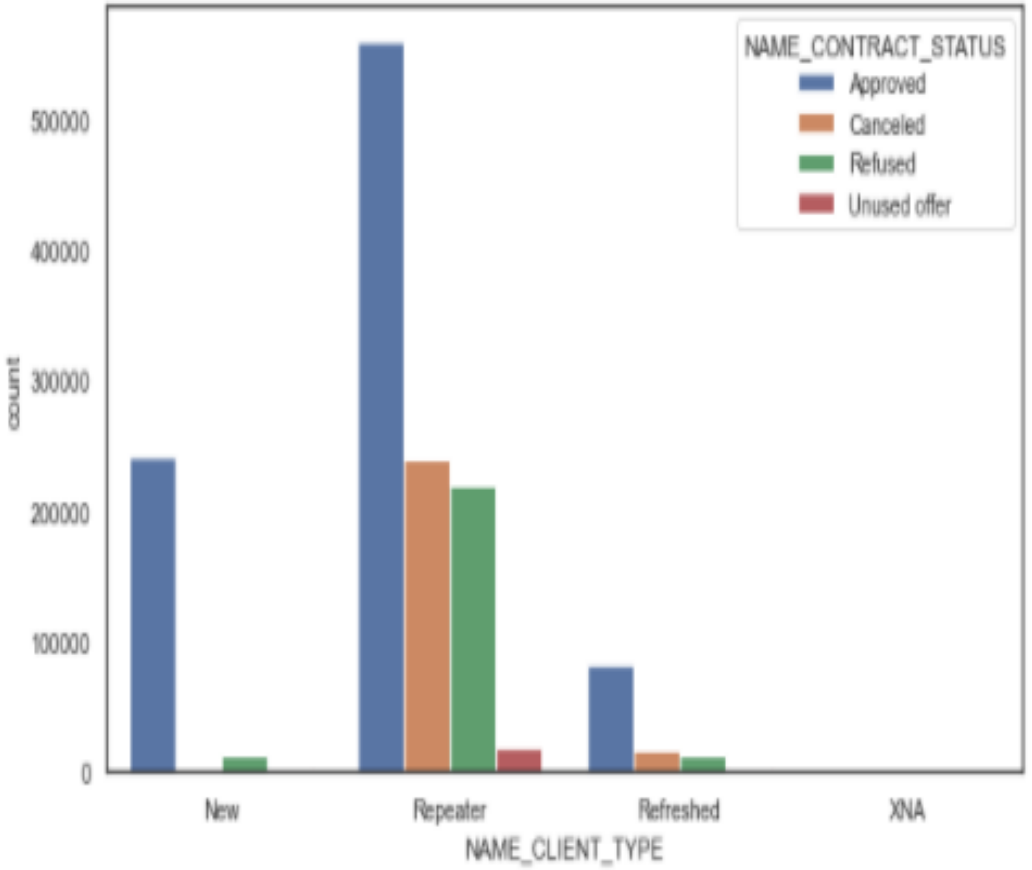




Univariate Analysis – Categorical columns (contd.)

We look at the distribution of various categorical columns w.r.t the NAME_CONTRACT_STATUS





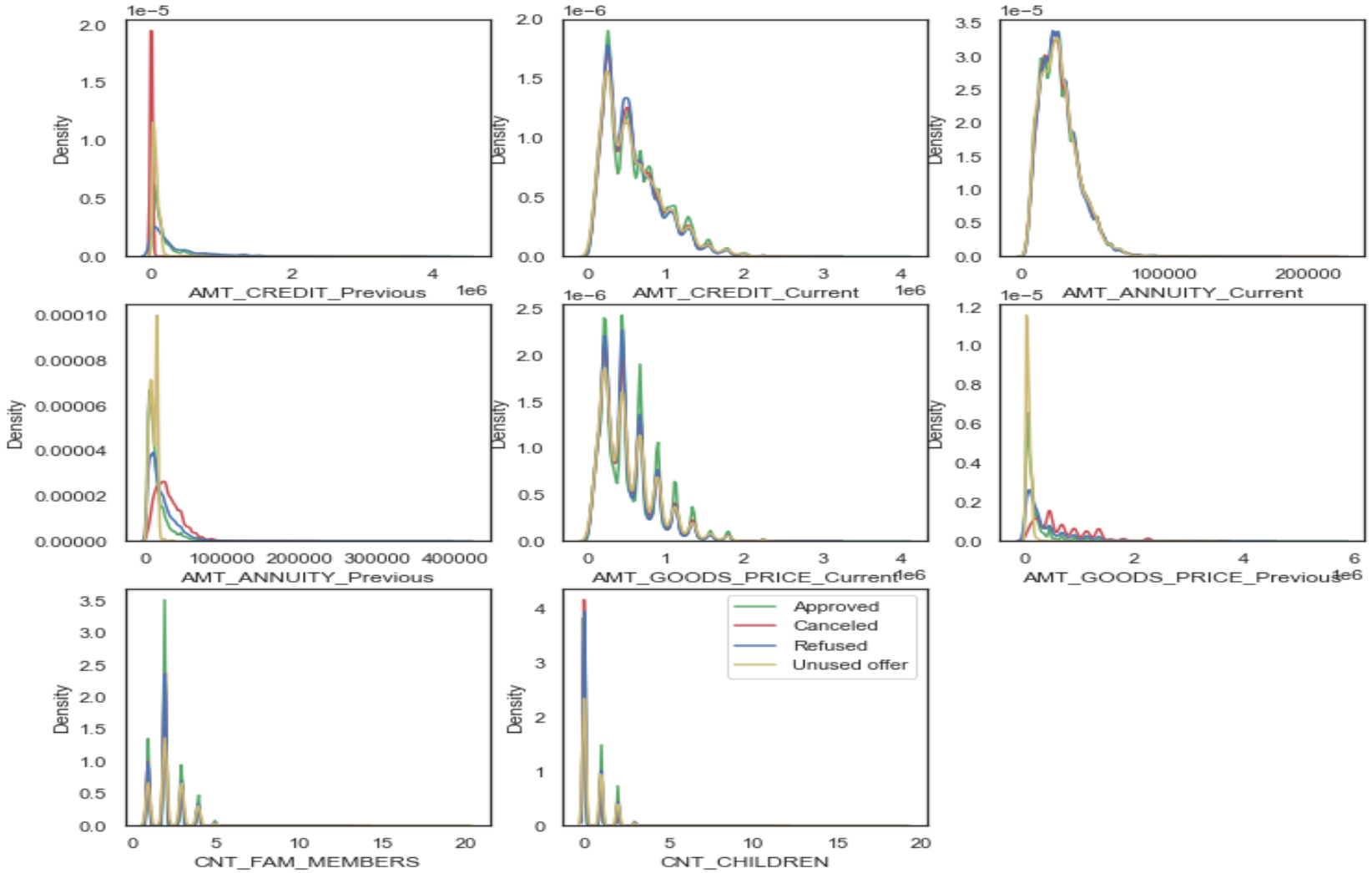
Univariate Analysis – Categorical columns

Key Insights

- Repeater has highest number of approved loans as well as cancelled loans.
- Middle NAME_YIELD_GROUP has highest approval, followed by high and followed by low_normal.
- Value of AMT_CREDIT_GROUP does very low affect on loan approvals.
- For Medium_group AMT_INCOME_GROUP the approval is highest, followed by LOW_GROUP.
- Both in NAME_TYPE_SUITE_Previous and NAME_TYPE_SUITE_Current unaccompanied has the highest proportion when compared with other groups.
- Currently bank is only giving two types of loans - Cash and Revolving Loans, but previously bank was providing Cash, Revolving and Consumer loans.
- Consumer loans were highest previously and now Cash loans are higher..

Univariate Analysis – Numerical columns

We look at the distribution of various numerical columns w.r.t the NAME_CONTRACT_STATUS



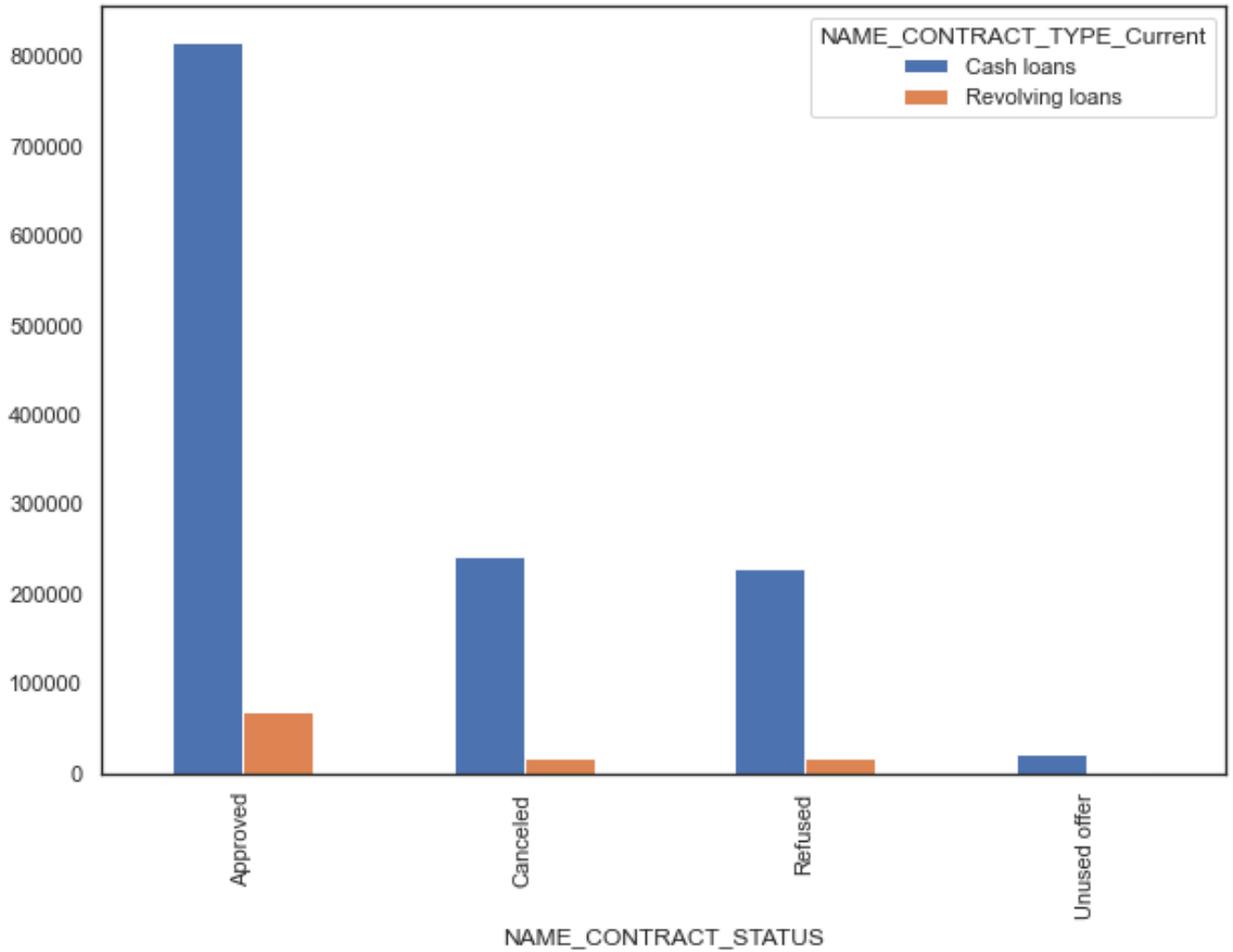
Univariate Analysis – Continuous/Numerical columns (contd.)

Key Insights

- Families who have two members in them tends to take more loans.
- Previously bank had applicants with high unused offers but currently applicants with Approved is high w.r.t AMT_GOODS_PRICE.
- Previously bank had applicants with high unused offers but currently all the groups are similar w.r.t AMT_ANNUITY.
- Previously bank had applicants with high Rejected offers but currently high number of Approved offers w.r.t AMT_CREDIT.

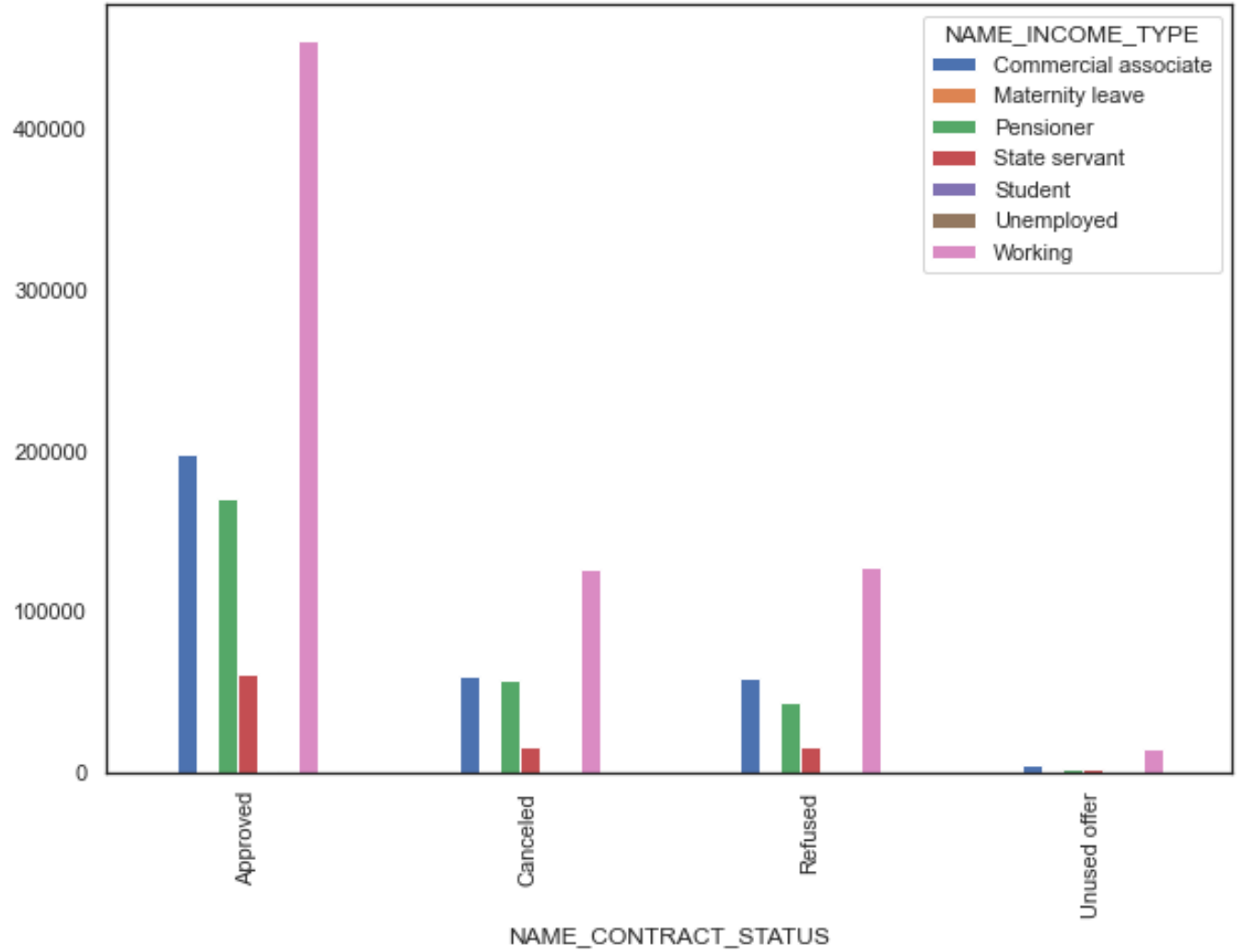
Bivariate Analysis – Categorical columns

Distribution of NAME_CONTRACT_TYPE_Current w.r.t. NAME_CONTRACT_STATUS: Cash loans have the highest count of Approved loans



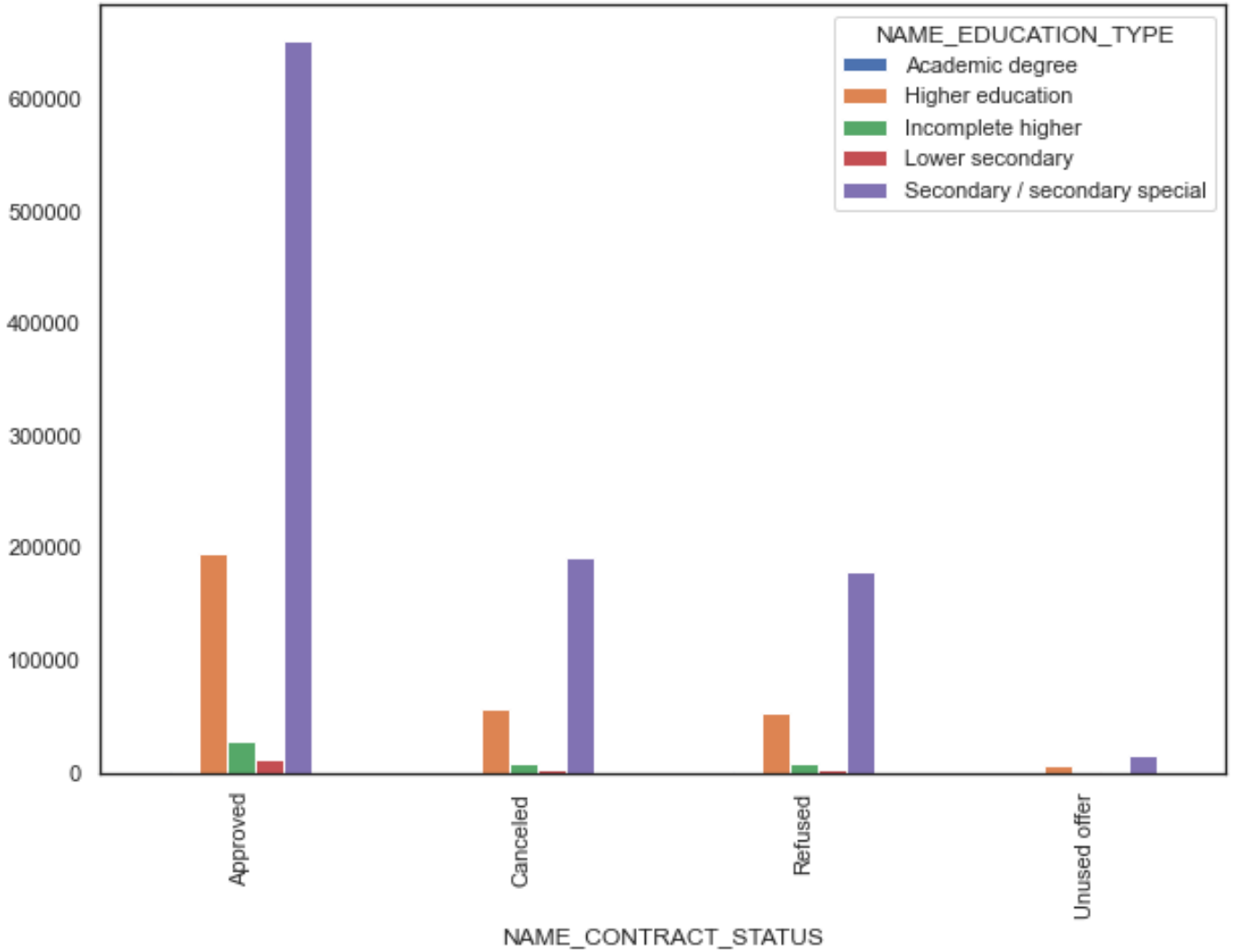
Bivariate Analysis – Categorical columns (contd.)

Distribution of NAME_INCOME_TYPE w.r.t. NAME_CONTRACT_STATUS: Highest number of approvals for working applicant



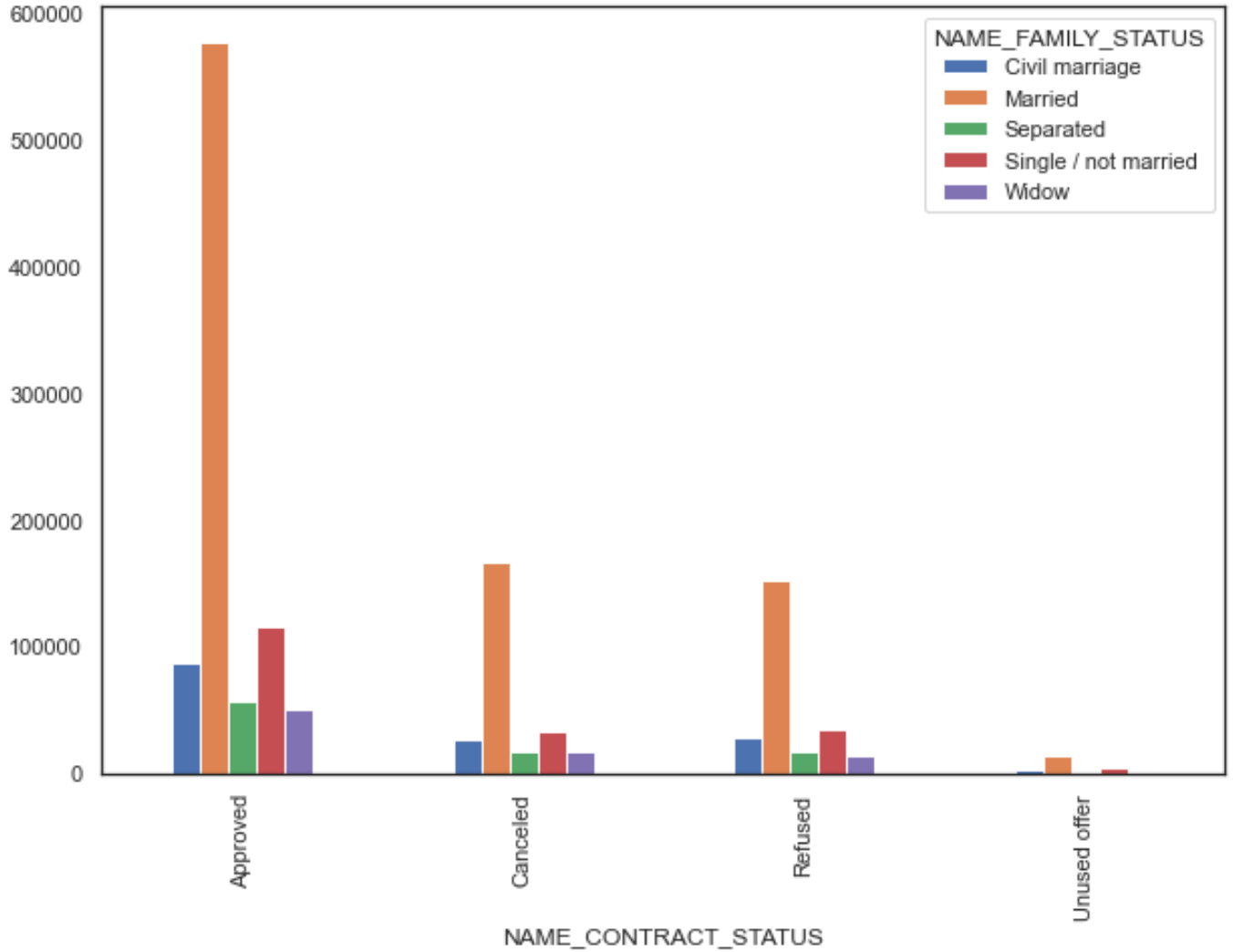
Bivariate Analysis – Categorical columns (contd.)

Distribution of NAME_EDUCATION_TYPE w.r.t. NAME_CONTRACT_STATUS: Highest number of approvals for Secondary/secondary special educated applicant



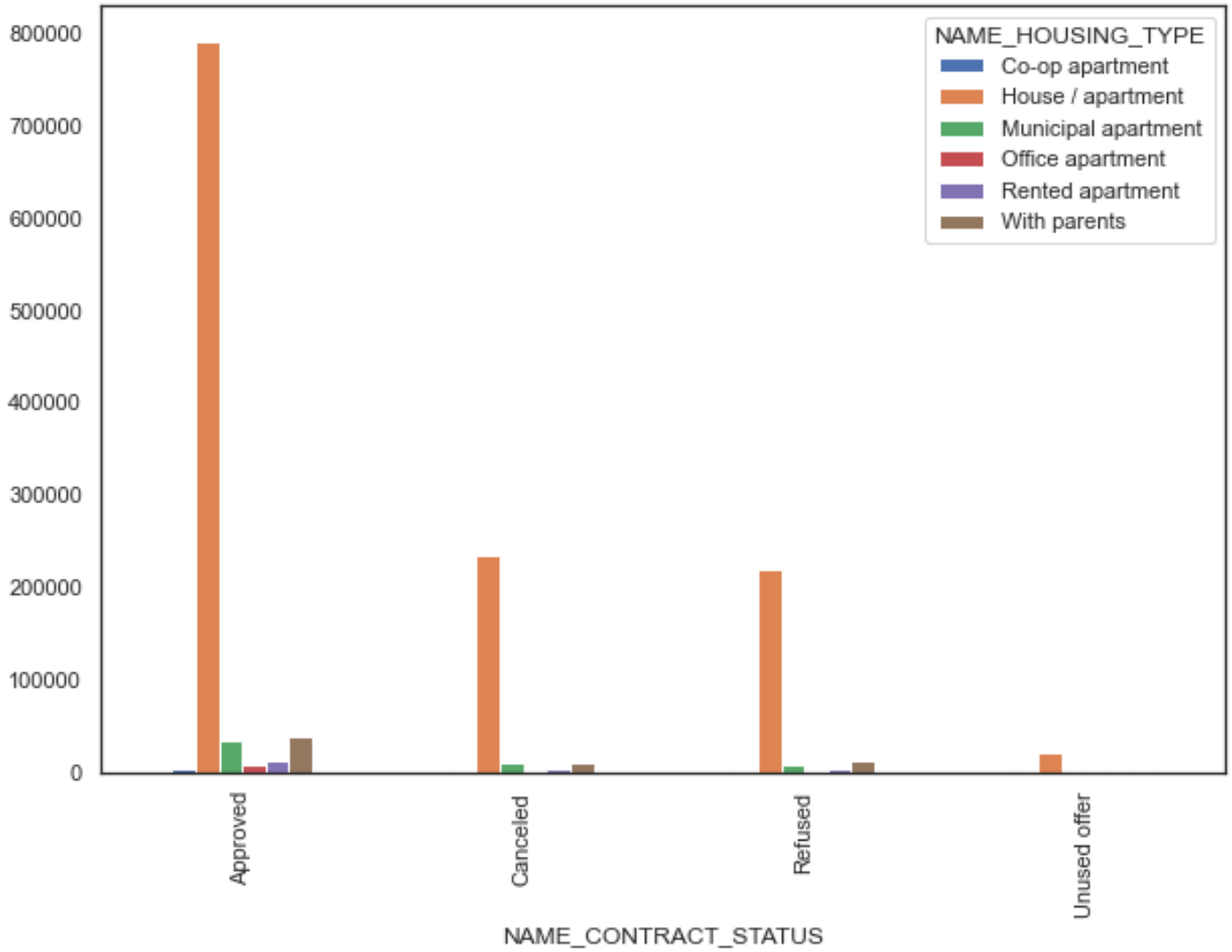
Bivariate Analysis – Categorical columns (contd.)

Distribution of NAME_FAMILY_STATUS w.r.t. NAME_CONTRACT_STATUS: Highest number of approvals for Married applicant



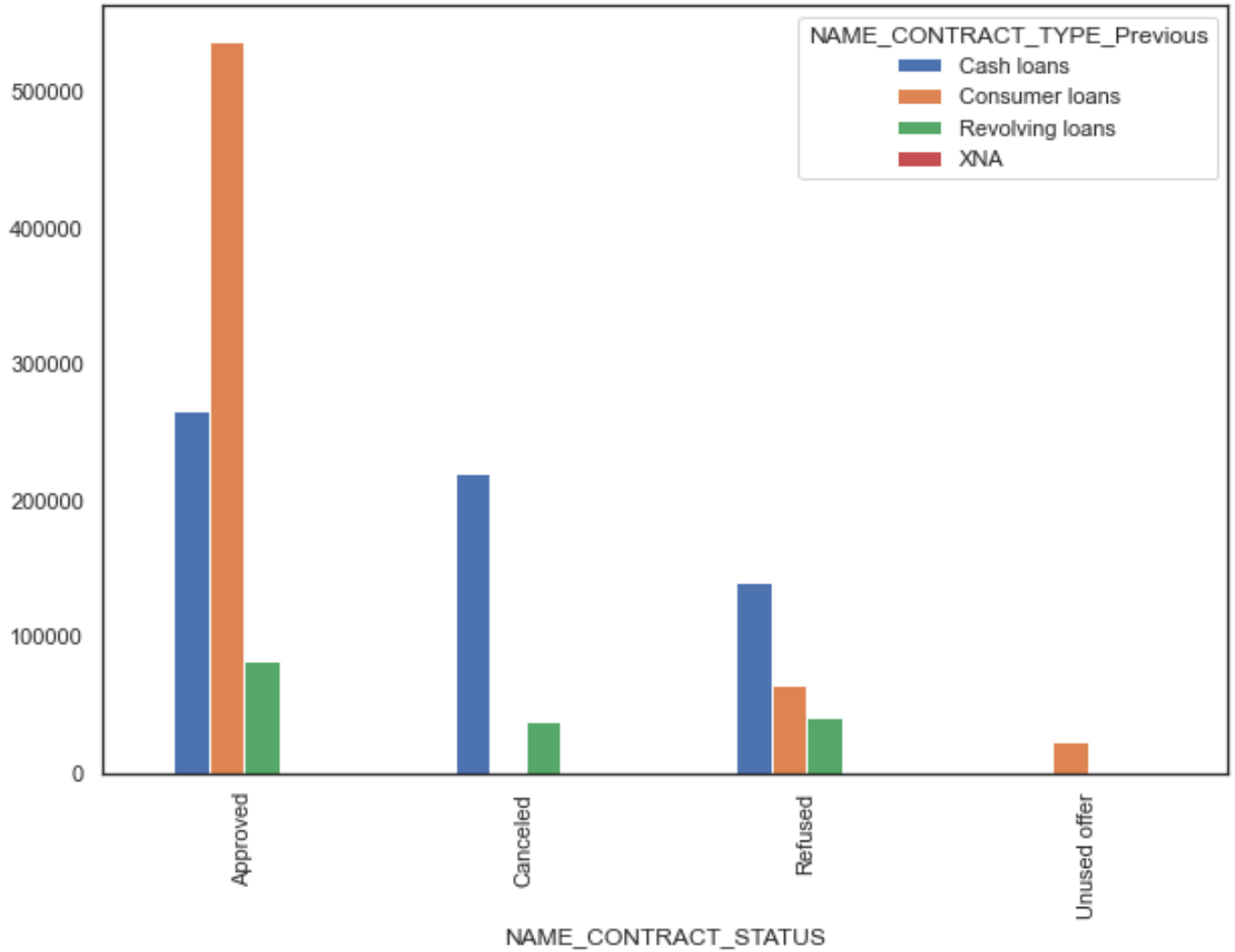
Bivariate Analysis – Categorical columns (contd.)

Distribution of NAME_HOUSING_TYPE w.r.t. NAME_CONTRACT_STATUS: Highest number of approvals for House/apartment owner.



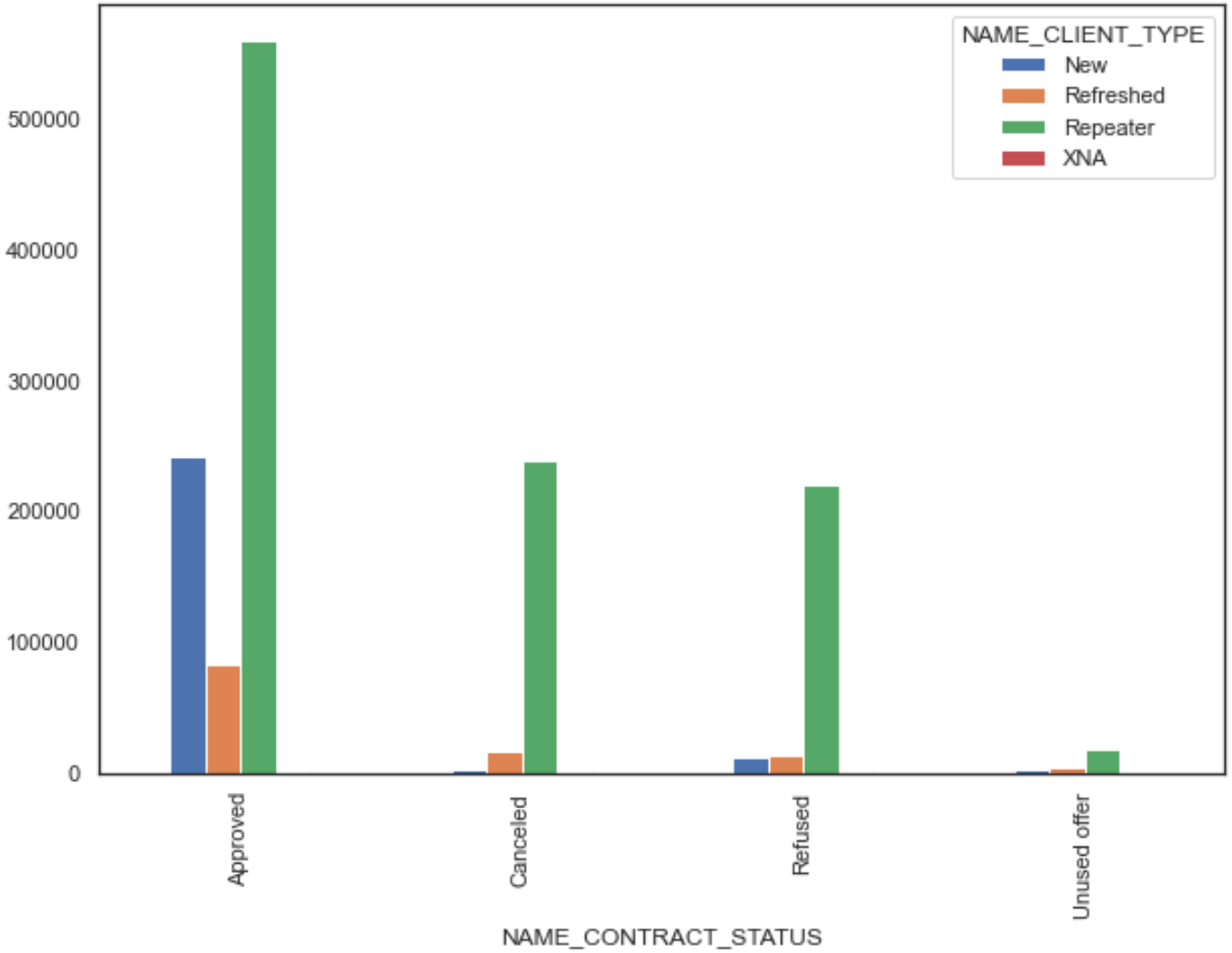
Bivariate Analysis – Categorical columns (contd.)

Distribution of NAME_CONTRACT_TYPE_Previous w.r.t. NAME_CONTRACT_STATUS: Highest number of approvals for Consumer Loans.



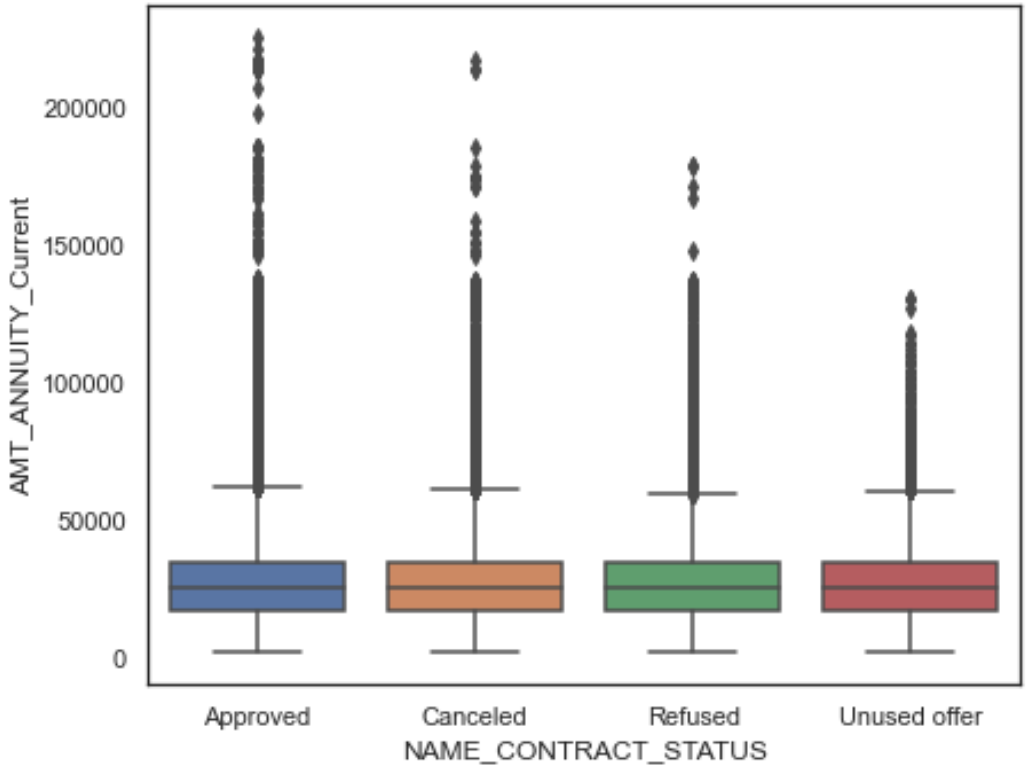
Bivariate Analysis – Categorical columns (contd.)

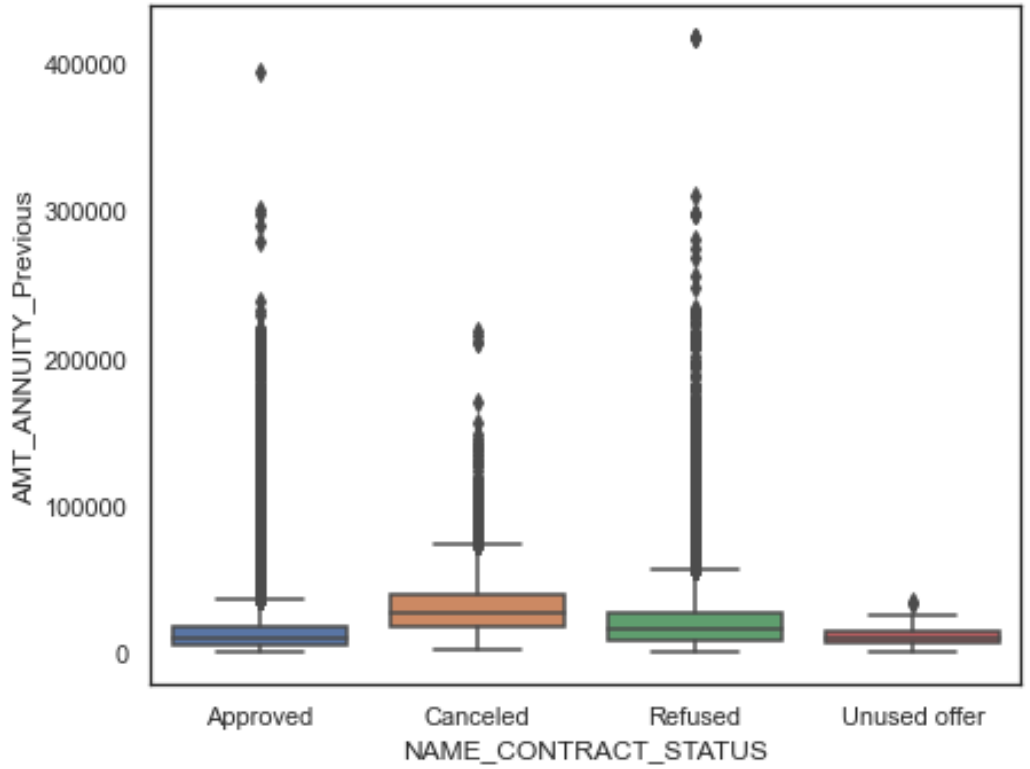
Distribution of NAME_CLIENT_TYPE w.r.t. NAME_CONTRACT_STATUS: Applicants who made Repeated applications got approved greater number of times compared to others



Bivariate Analysis – Numerical columns

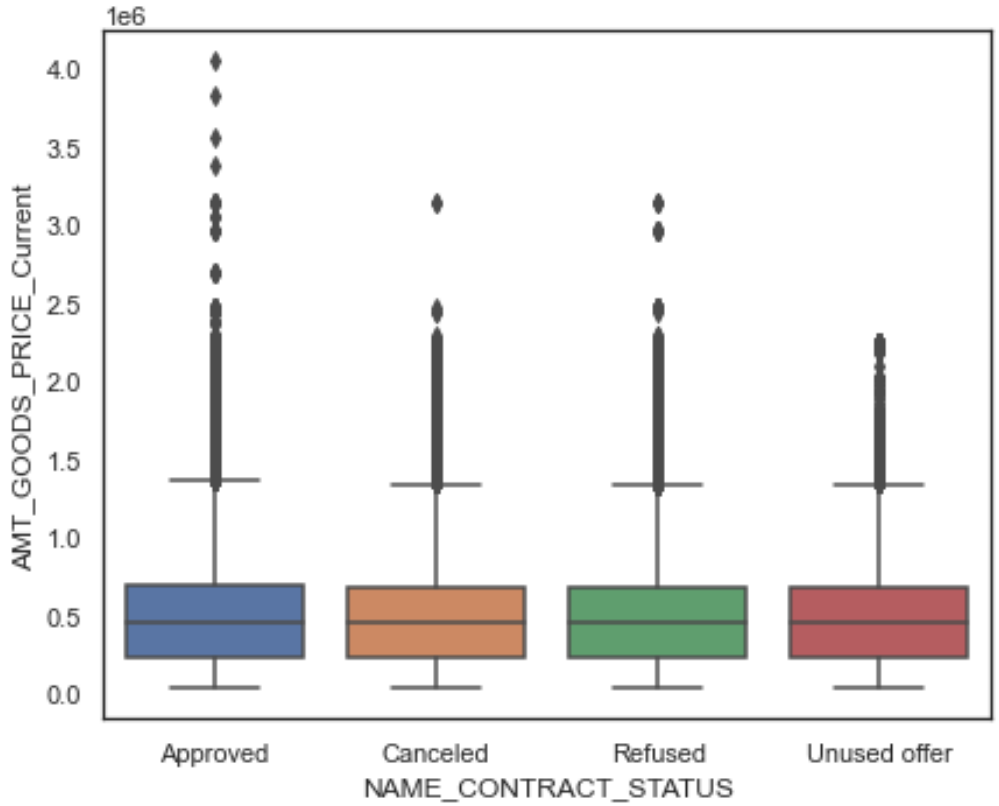
Distribution of various numerical columns in the merged dataset w.r.t. NAME_CONTRACT_STATUS

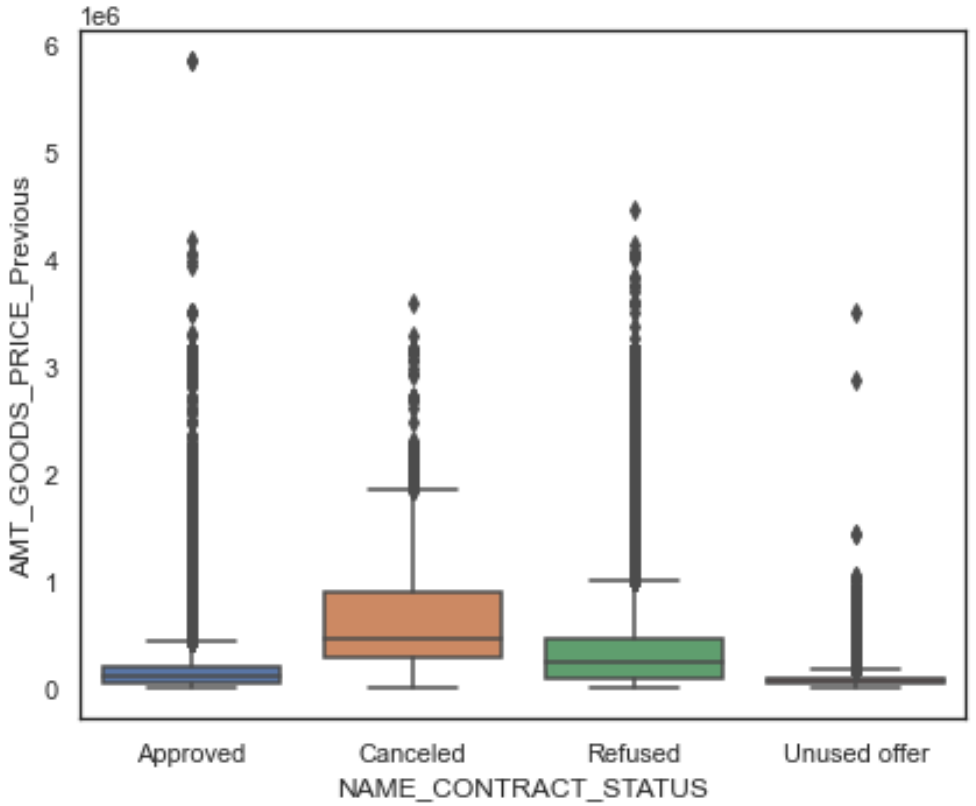




Bivariate Analysis – Numerical columns (contd.)

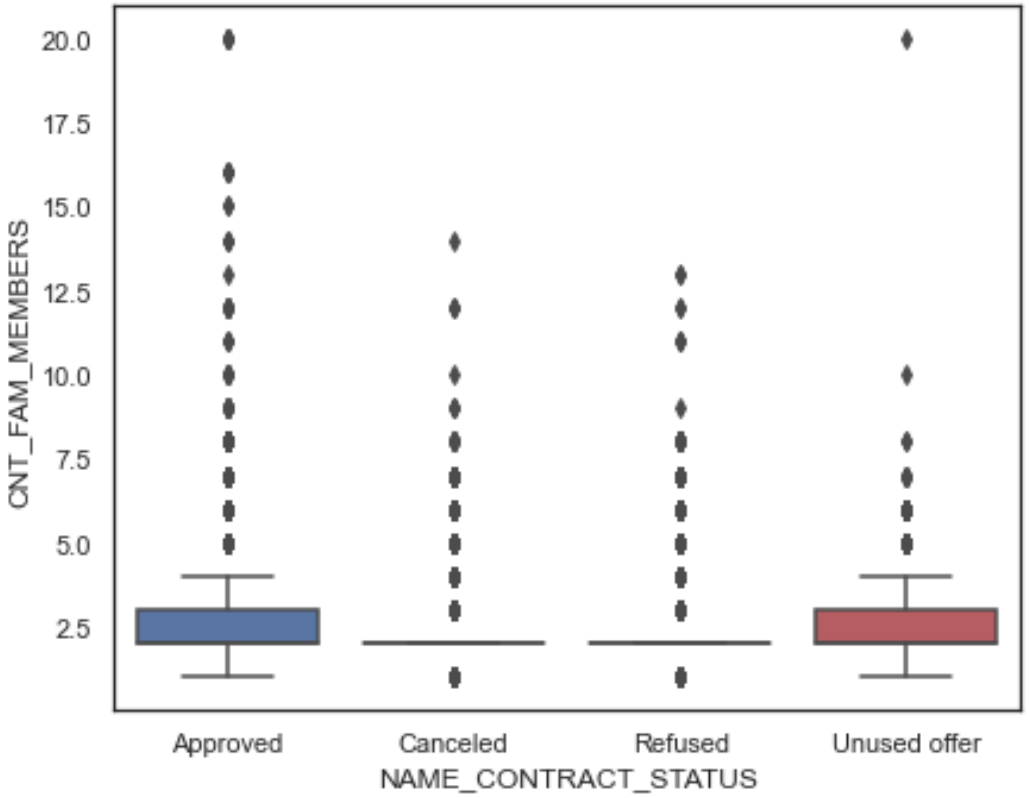
Distribution of various numerical columns in the merged dataset w.r.t. NAME_CONTRACT_STATUS

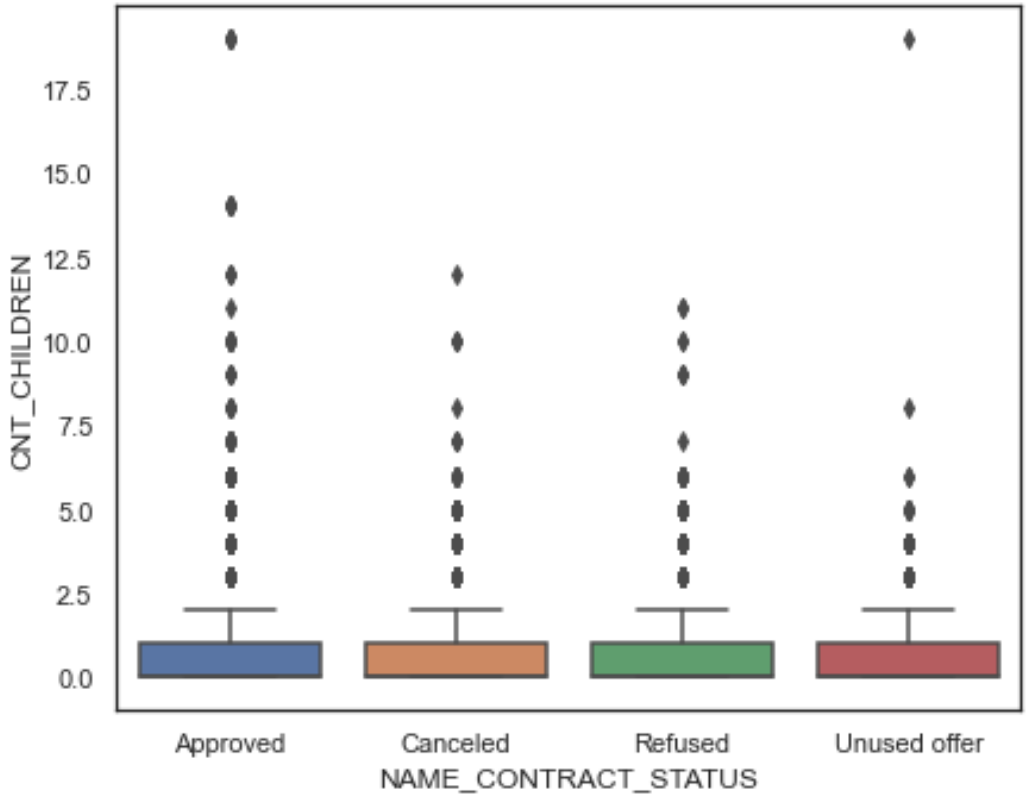




Bivariate Analysis – Numerical columns (contd.)

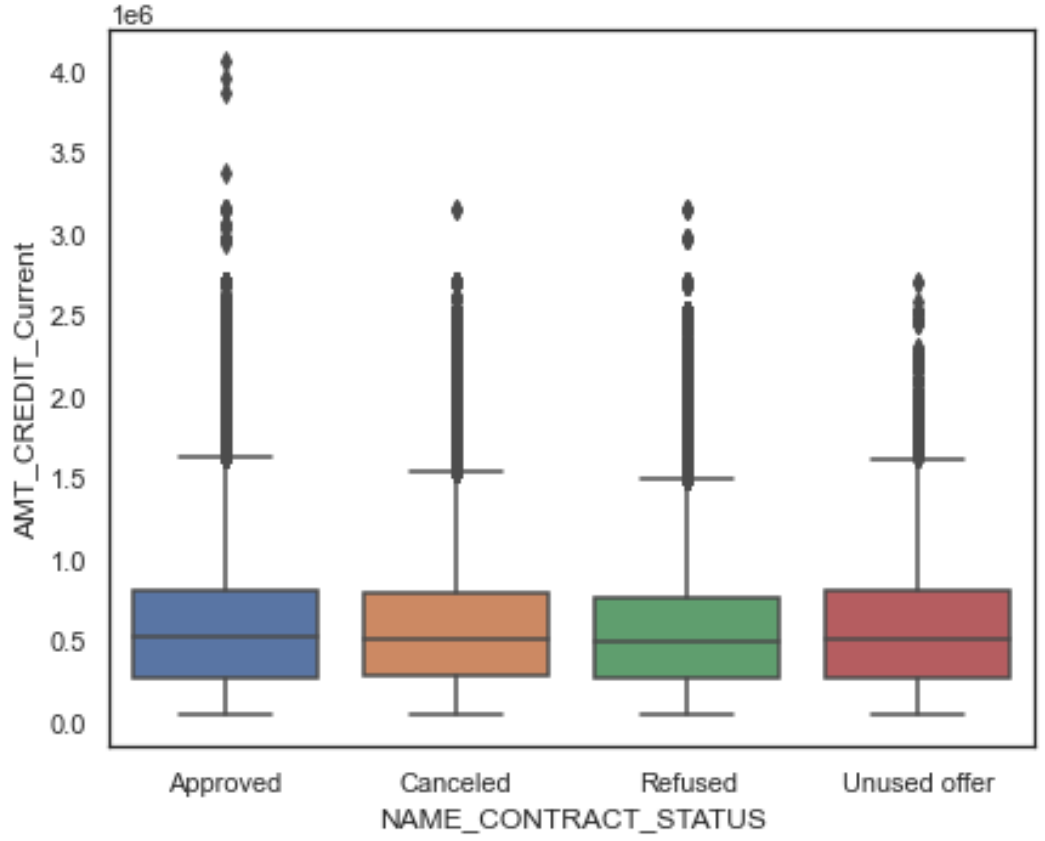
Distribution of various numerical columns in the merged dataset w.r.t. NAME_CONTRACT_STATUS

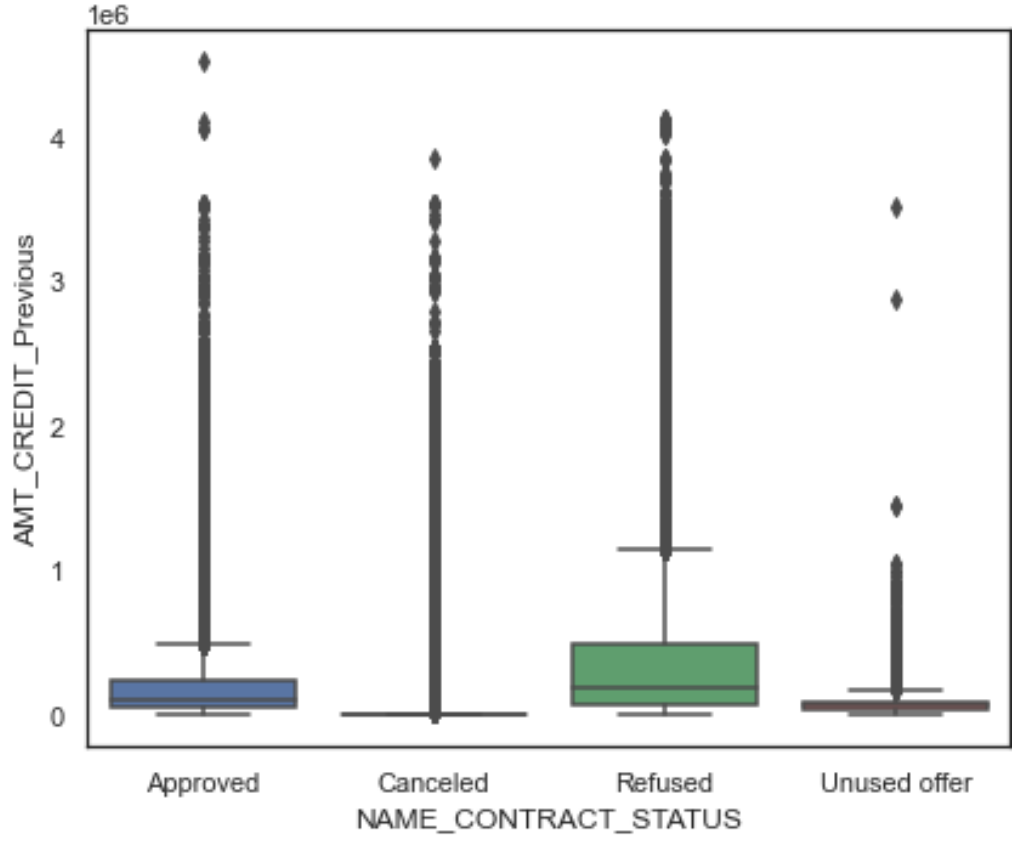




Bivariate Analysis – Numerical columns (contd.)

Distribution of various numerical columns in the merged dataset w.r.t. NAME_CONTRACT_STATUS





Bivariate Analysis – Numerical columns (contd)

Distribution of various numerical columns in the merged dataset w.r.t. NAME_CONTRACT_STATUS

- AMT_CREDIT_Previous has highest refused cases and AMT_CREDIT_Current have a similar distribution for all 4 cases.
- Family with 2-3 people are the majority group in the approved and unused categories.
- Previously most of the applications were cancelled or refused but now Refused/Cancelled/Approved/Unused all have similar distribution w.r.t AMT_GOODS_PRICE and AMT_ANNUITY.