Case Study - Credit EDA

Submitted by: Janarthanan Balasubramanian & Siva Prakash

Business Understanding

You are working for a consumer finance company, which specialises in lending various types of loans to urban customers. When a loan application is received, the company has to decide for loan approval based on the applicant's profile. There are two associated risks.

(1) Not approving the loan for an applicant who is likely to repay the loan.
(2) Approving the loan for an applicant who is not likely to repay the loan.

With the available dataset, identify patterns that 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 at a higher interest rate, etc.

	Company approves the loan	Company does not approve the loan
Applicant likely to repay loan	✓	Loss of Business
Applicant not likely to repay loan	Financial Loss	✓

The Challenge

to identify the clients who are capable of repaying their loans

so that their loan applications are not rejected.

Objective of Risk Assessment

To understand how consumer attributes and loan attributes influence the tendency of default, i.e. the variables which are strong indicators of default.

The following data sets are available:

application_data.csv	previous_application.csv	columns_description.csv
Contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.	Contains information about the client's previous loan data: whether the previous application had been Approved, Cancelled, Refused or Unused Offer.	Data Dictionary. Describes the meaning of the variables.

Deliverables

Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly.

Task 1: Identify the missing data and use the appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)

Task 2: Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.

Task 3: Identify if there is data imbalance in the data. Find the ratio of data imbalance.

You can plot more than one type of plot to analyse the different aspects due to data imbalance. For example, you can choose your own scale for the graphs, i.e. one can plot in terms of percentage or absolute value. Do this analysis for the 'Target variable' in the dataset (clients with payment difficulties and all other cases).

Task 4: Use a mix of univariate and bivariate analysis etc. Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.

Task 5: Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable, and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5, Target. If you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable, which is increasing or decreasing.

Task 6: Include visualisations and summarise the most important results in the presentation. You are free to choose the graphs, which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the clients with payment difficulties with all other cases.

Rubric for Case Study

*Use a mix of univariate analysis, segmented univariate analysis and bivariate analysis.

Our Approach for Loan Analytics

Credit EDA Case Study - Part 1

Step 1: Load the data sets into Python Data Frames.

Step 2: Explore the data sets. Identify the variables that need to be used for further analysis.

Step 3: Identify data quality issues. How to address null values and missing data? Finalize the approach.

Step 4: Univariate Analysis of Application Data. For categorical variables, we have presented the bar charts showing frequency of values and for numeric variables, we have used histograms and box plots.

Step 5: Univariate Analysis of Previous Applications.

Step 6: Studying the data imbalance and computing the imbalance ratio.

Credit EDA Case Study - Part 2

Step 7: Data Preparation. We have computed derived variables from previous application data file which can be used for further analysis. The two data sets are merged based on SK_ID_CURR.

Step 8: Segmented Univariate Analysis using the merged data.

Step 9: Bivariate Analysis using the merged data.

Step 10: Identifying the driver variables.

Step 11: Studying the correlation between the driver variables using joint plots and pair plots.

Step 12: Drawing insights and providing actionable recommendations.

Note: The Python 3 Notebook submitted with this presentation is organized accordingly to these steps.

1. Data Dictionary

For each field in the dataset

(which contains the chosen categorical and quantitative variables from application_data.csv for analysis),

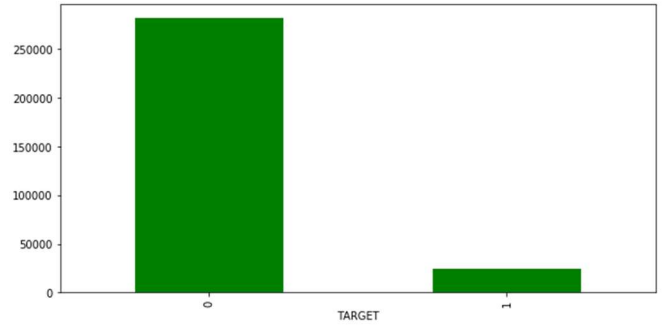
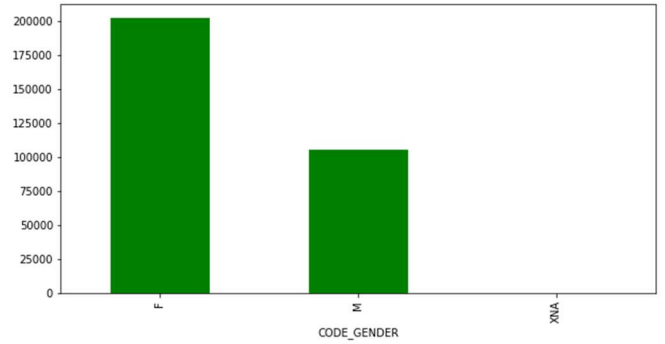
we have presented the following:

- A brief description
- Unique values and value counts
- Identification of **data quality issues**, if any
    a. Whether there are any null values or missing values
    b. How to handle the null or missing values
- **Univariate Analysis:**
    a. Bar chart plotting the frequency of the values (in case of categorical variables)
    b. Histogram for univariate analysis of the quantitative variables
    c. Box Plot to visualize the distribution of the quantitative variables
- **Identification of the Outliers** and insights regarding any imbalance in the data.
- **Studying the Imbalance in the Data.**

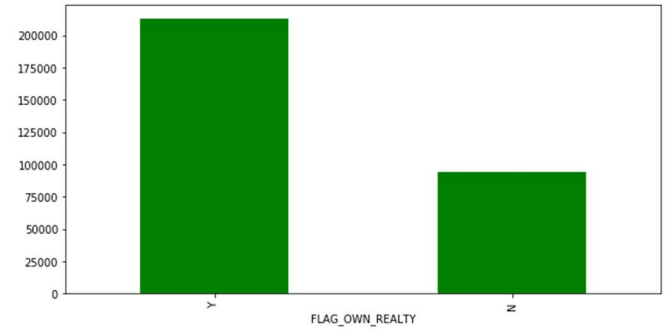
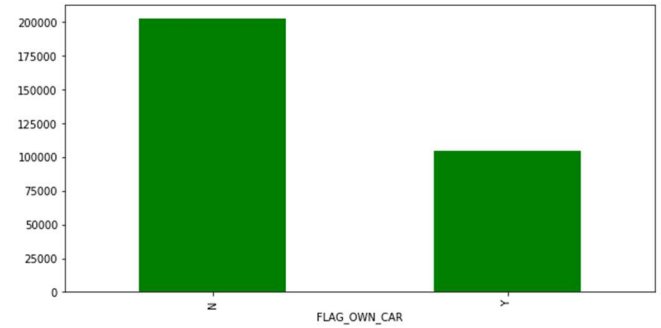
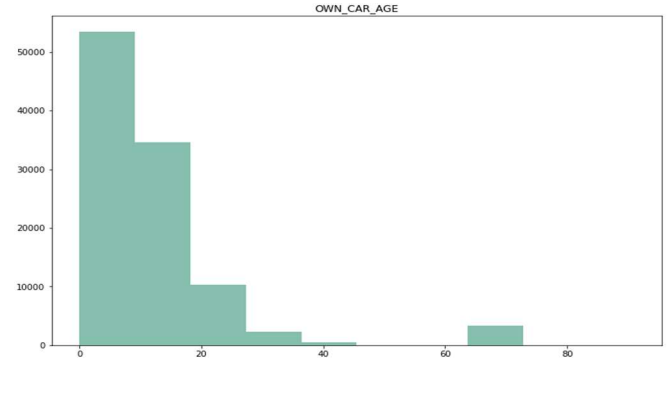
This section of the presentation covers the part 1 of the Credit EDA case study.

1.1 APPLICATION DATA

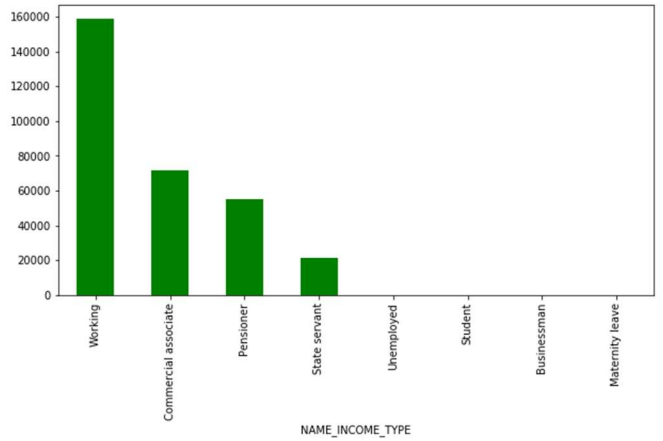
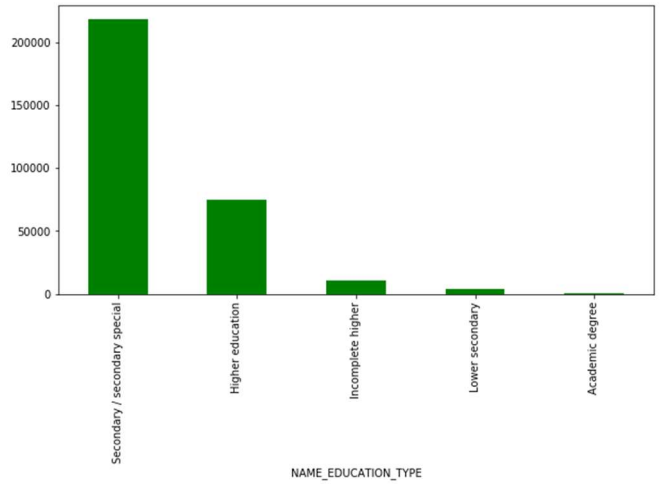
Applicant Profile - Categorical Variables

Column Name	Explanation	Null	Plot
SK_ID_CURR	ID of loan in our sample	None	Unique ID Value. To study the previous applications of the client link with this id.
TARGET	Target variable: 1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample, (24825 records) 0 - all other cases. (282686 records). From the frequency plot we see that the sample selected has very few clients with payment difficulties.	None	
CODE_GENDER	Gender of the client (M / F / XNA) F: 202448 M: 105059 XNA: 4 **Insight:** The number of female applicants is more than the male applicants. 66% of the applicants are female and only 34% are male.	None	

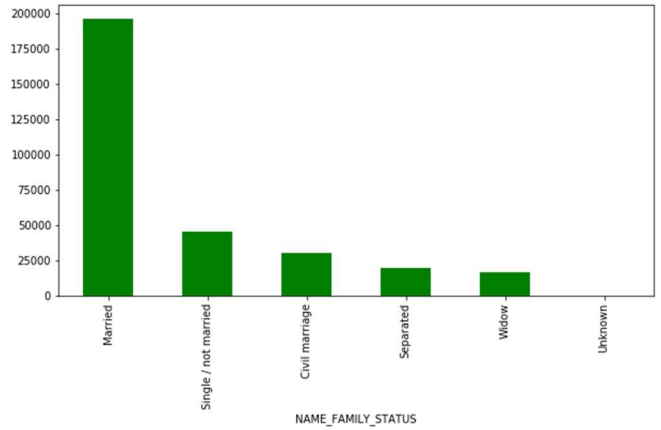
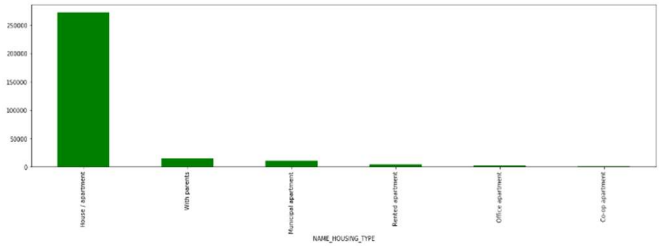
Applicant Profile - Categorical Variables

Column Name	Explanation	Null	Plot
FLAG_OWN_REALTY	Flag if client owns a house or flat Y: 213312 N: 94199 **Insight:** 69% of the applicants owned a house or a flat. 31% did not own a house or a flat.	None	
FLAG_OWN_CAR	Flag if the client owns a car N: 202924 Y: 104587 **Insight:** 66% of the applicants did not own a car.	None	
OWN_CAR_AGE	Age of client's car. The null values correspond to those who do not own a car. For 5 applicants the car age is null in spite of the person owning the car. Set FLAG_OWN_CAR = 'N' for these records. Q1: 5 Q2: 9 Q3: 15 IQR - Q3 - Q1 = 10. Outliers are values beyond the range (-10, 30). There are 4932 such records.	202929	

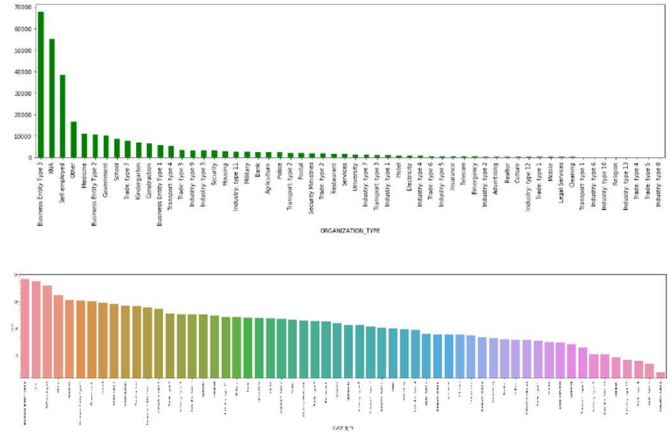
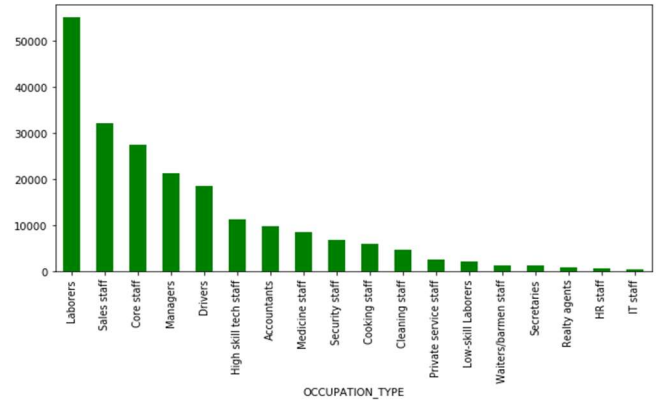
Applicant Profile - Categorical Variables

Column Name	Explanation	Null	Plot
NAME_INCOME_TYPE	Clients income type. Working: 158774 (51.6%) Commercial associate: 71617 (23.3%) Pensioner: 55362 (18.0%) State servant: 21703 (7.0%) Unemployed: 22 Student: 18 Businessman: 10 Maternity leave: 5	None	
NAME_EDUCATION_TYPE	Level of highest education the client achieved Secondary / secondary special: 218391 (71%) Higher education: 74863 (24.3%) Incomplete higher: 10277 (3.3%) Lower secondary: 3816 (1.2%) Academic degree: 164 (negligible)	None	

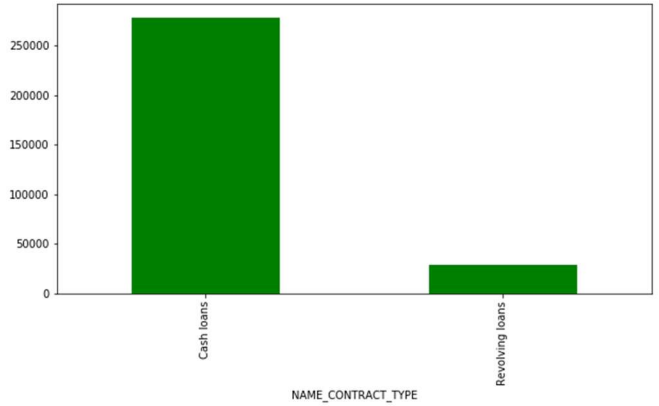
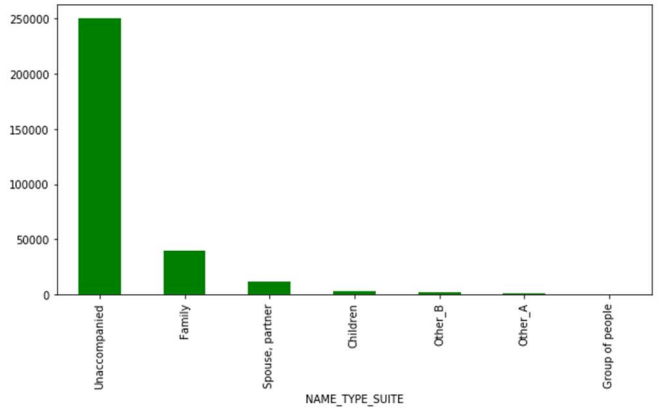
Applicant Profile - Categorical Variables

Column Name	Explanation	Null	Plot
NAME_FAMILY_STATUS	Family status of the client Married: 196432 (63.8%) Single / not married: 45444 (14.7%) Civil marriage: 29775 (9.7%) Separated: 19770 (6.4%) Widow: 16088 (5.2%) Unknown: 2 **Insight:** 73.5% of the applicants are married (or civil marriage). Only 15% are single / unmarried.	None	
NAME_HOUSING_TYPE	What is the housing situation of the client House / apartment: 272868 With parents: 14840 Municipal apartment: 11183 Rented apartment: 4881 Office apartment: 2617 Co-op apartment: 1122	None	 Note: The data is not proper. Most of the applicants have selected their housing situation as house / apartment without further details.

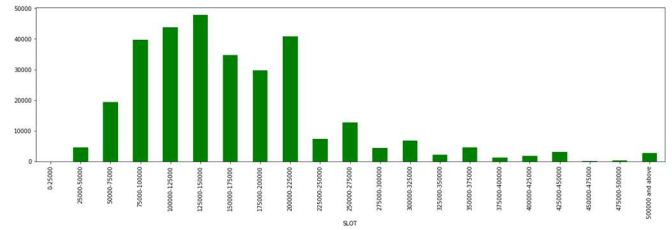
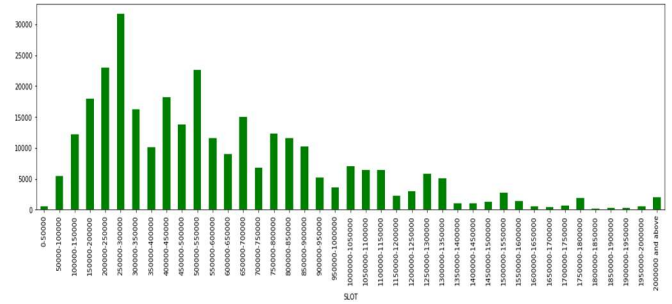
Applicant Profile - Categorical Variables

Column Name	Explanation	Null	Plot
ORGANIZATION_TYPE	Type of organization where client works XNA: 55374 (can be updates as Other) Other: 16681 **Insights:** Most of the applicants are from Business Entity Type 3 (67989). But we can see a long tail, where many small sectors contribute to significant number of applicants. Hence, a plot in logarithmic scale is also shown.	None	
OCCUPATION_TYPE	What kind of occupation does the client have. Laborers: 55186 Sales staff: 32102 Core staff: 27570 Managers: 21371 Drivers: 18603 High skill tech staff: 11380 Accountants: 9813 Medicine staff: 8537 Security staff: 6721 Cooking staff: 5946 Cleaning staff: 4653 Private service staff: 2652 Low-skill Laborers: 2093 Waiters/barmen: 1348 Secretaries: 1305 Realty agents: 751 HR staff: 563 IT staff: 526	96391	

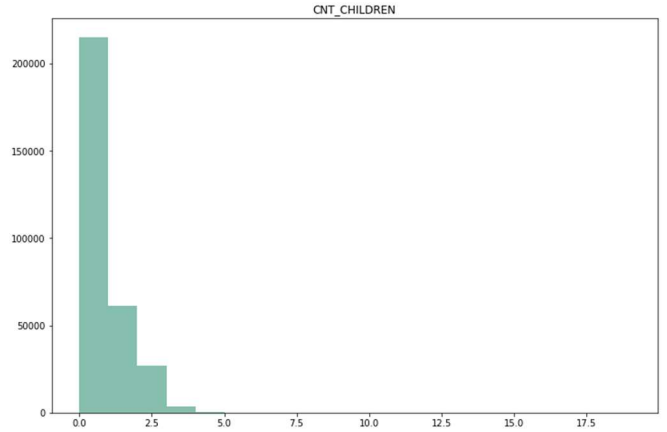
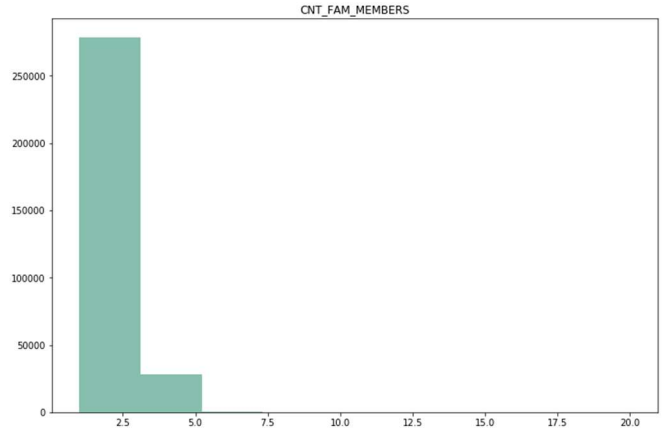
Applicant Profile - Categorical Variables

Column Name	Explanation	Null	Plot
NAME_CONTRACT_TYPE	Identification if loan is cash or revolving. Cash loans: 278232 Revolving loans: 29279 **Insights:** The sample given does not contain any consumer loans. It has only cash loans (90%) and revolving loans (10%).	None	
Note: A loan is typically repaid through fixed monthly payments. Each monthly payment includes both principal and interest. Revolving loan is a type of loan that does not have a fixed number of payments. It is an arrangement which allows for the loan amount to be withdrawn, repaid, and redrawn again in any manner and any number of times, until the arrangement expires.			
NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan. For the missing values, we can update it as Unaccompanied (mode). Unaccompanied: 248526 (80%) Family: 40149 Spouse, partner: 11370 Children: 3267 Other_B: 1770 Other_A: 866 Group of people: 271	1292	

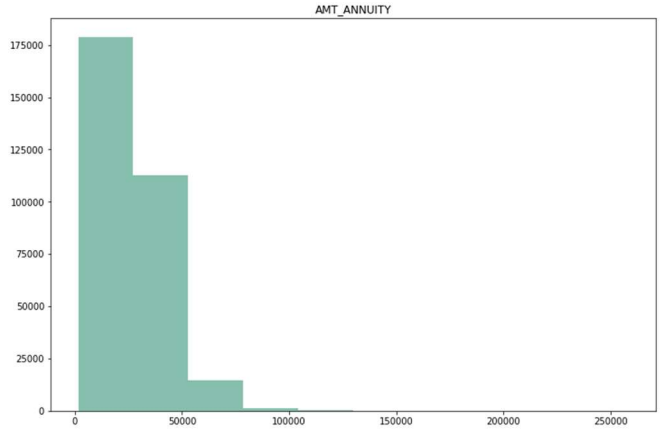
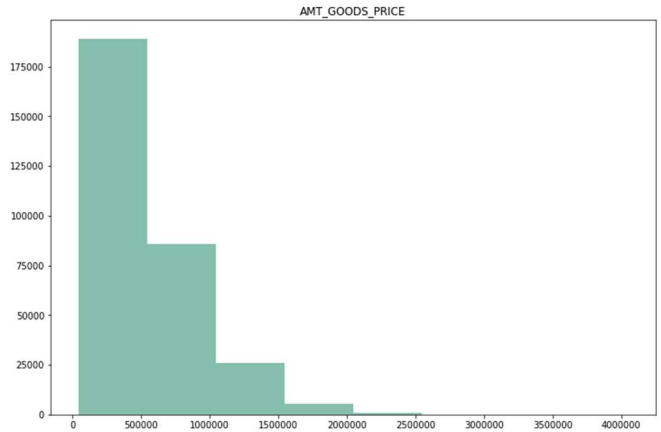
Applicant Profile - Binning of the Quantitative Variables

Column Name	Explanation	Null	Plot
AMT_INCOME_TOTAL	Income of the client. Since the distribution was skewed, we have created a categorical variable AMT_INCOME_RANGE and created the bar chart. Minimum: 25650 Maximum: 117000000 Mean: 168797.20 Median: 146997.00 **Outliers:** Q1: 112500 Q2: 146997 Q3: 202500 IQR: Q3 - Q1 = 90000 Outliers are values beyond the range (-22500.00, 337500.00) - there are 14035 such records.	None	
AMT_CREDIT	Credit amount of the loan. Since the distribution was skewed, we have created a categorical variable AMT_CREDIT_RANGE and created the bar chart. Minimum: 45000 Maximum: 4050000 Mean: 599027 Median: 513531 **Outliers:** Q1: 270000 Q2: 513531 Q3: 808650 IQR: Q3 - Q1 = 538650 Outliers are values beyond the range (-537975, 1616625). There are 6562 such records.	None	

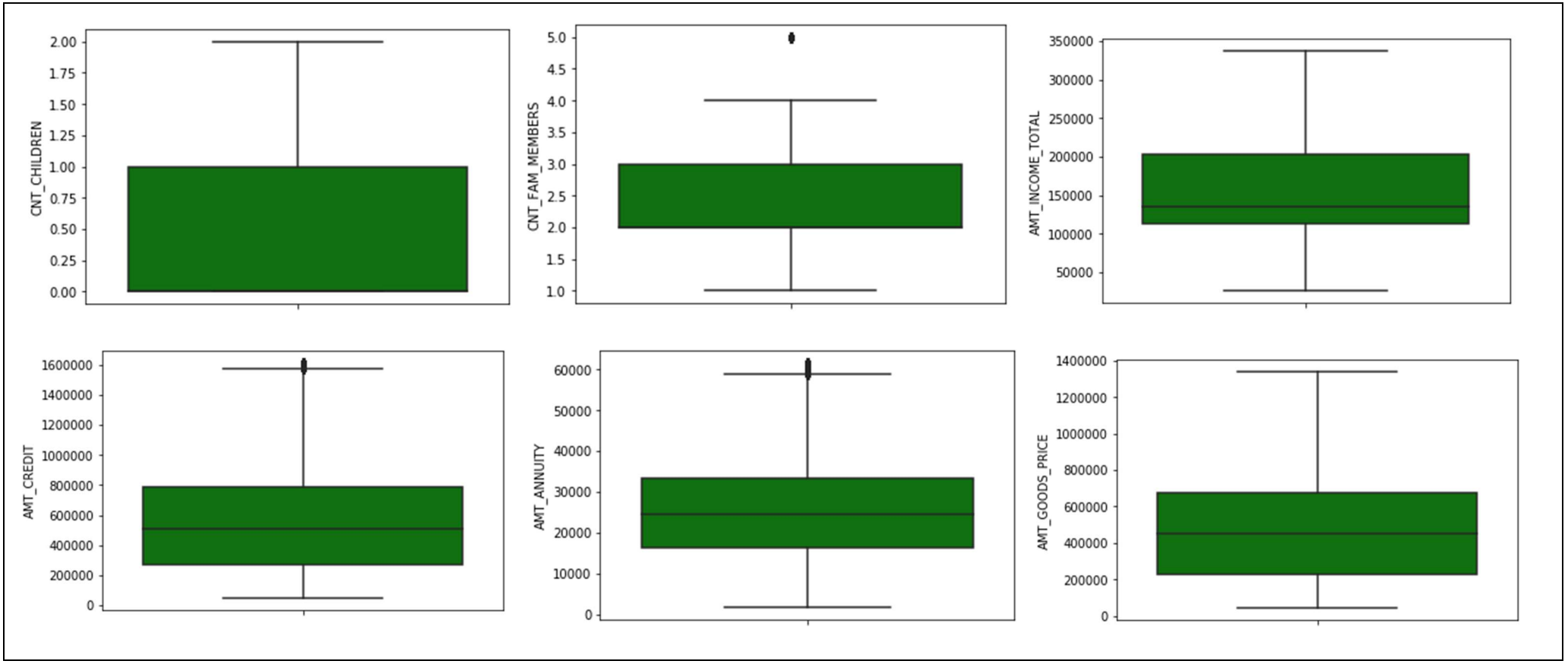
Applicant Profile - Quantitative Variables

Column Name	Explanation	Null	Plot
CNT_CHILDREN	Number of children the client has Minimum: 0 Maximum: 1 Mean: 0.417 Median: 0 Mode: 0 **Outliers:** Q1: 0 Q2: 0 Q3: 1 IQR: Q3 - Q1 = 1 Outliers are values beyond the range (-0.5, 1.5) There are 4272 such records.	None	
CNT_FAM_MEMBERS	How many family members does client have Minimum: 1 Maximum: 20 Mean: 2.15 Median: 2 Mode: 2 **Outliers:** Q1: 2 Q2: 2 Q3: 3 IQR: Q3 - Q1 = 1 Outliers are values beyond the range (0.5, 4.5) There are 529 such records.	2	

Applicant Profile - Quantitative Variables

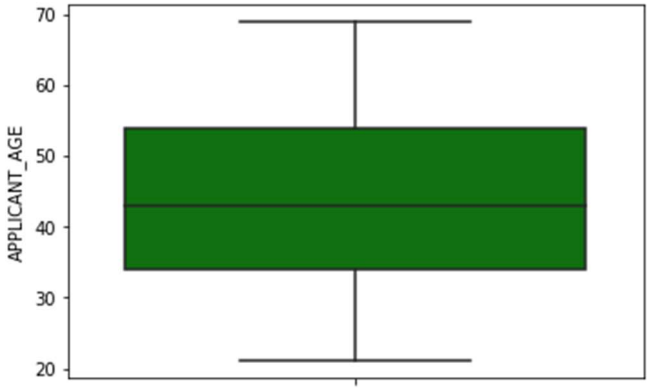
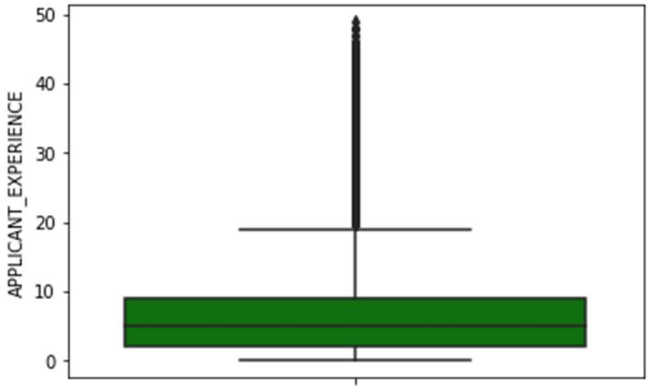
Column Name	Explanation	Null	Plot
AMT_ANNUITY	Loan annuity. In case of Revolving Loans, annuity is 1/20 of the AMT_GOODS_PRICE (in most cases). But the null values are corresponding to cash loans where there is no such relationship. It is better to drop these 12 rows than it affecting our analysis. Minimum: 1615.5 Maximum: 258025.5 Mean: 27108.57 Median: 24903 **Outliers:** Q1: 16524 Q2: 24903 Q3: 34596 IQR: Q3 - Q1 = 18072. Outliers are values beyond the range (-10584, 61704). There are 7504 such records.	12	
AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given. All the null values are corresponding to the Revolving Loans. Assume that AMT_GOODS_PRICE = AMT_CREDIT in these cases. Minimum: 40500 Maximum: 4050000 Mean: 538163 Median: 450000 **Outliers:** Q1: 238500 Q2: 450000 Q3: 679500 IQR: Q3 - Q1 = 441000. Outliers are values beyond the range (-423000, 1341000). There are 14730 such records.	278	

Applicant Profile - - Quantitative Variables (Box Plots - after removal of Outliers)



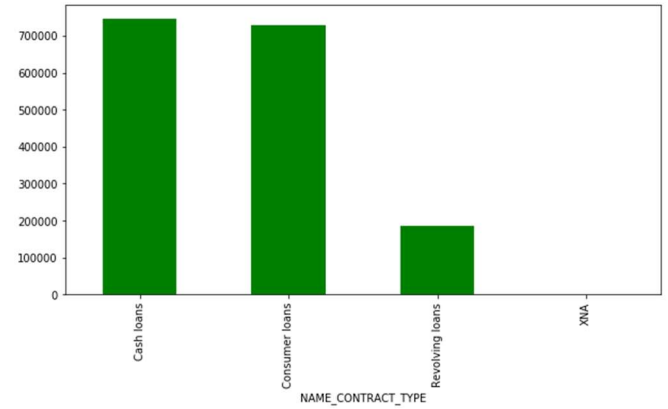
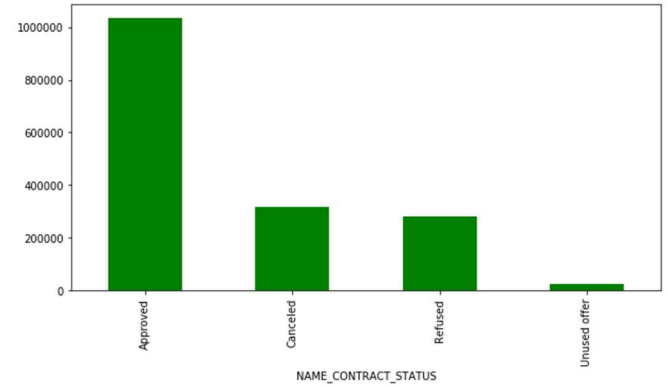
Insights: Due to removal of outliers, the median shifts to a more natural value. For CNT_CHILDREN, the values are distributed between 0 and 1. Whereas, most of the values for CNT_FAM_MEMBERS are in the range of 2-3. We can also see that the total income is distributed in the range 120000 to 200000; but the distribution is skewed and most of the people fall in the bracket of 140000 to 200000. The credit amount is always in the range of 200000 to 800000 and is evenly distributed. We see such even distribution in AMT_ANNUITY and AMT_GOODS_PRICE also.

Applicant Profile - Quantitative Variables

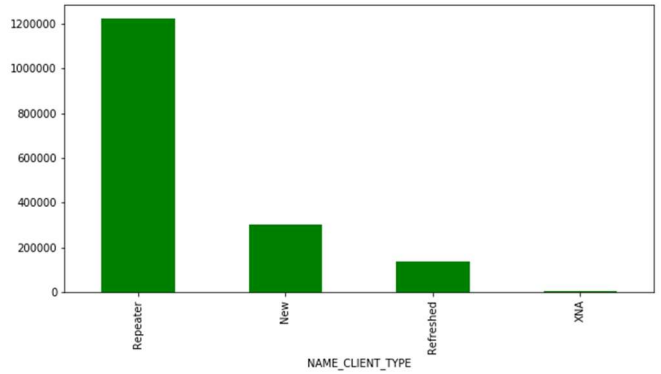
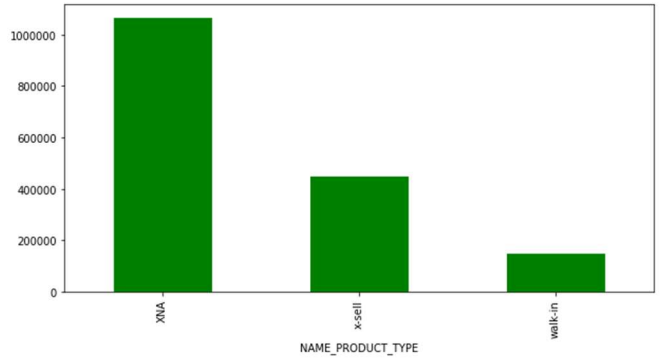
Column Name	Explanation	Null	Plot
DAYS_BIRTH	Client's age in days at the time of application. A derived field APPLICANT_AGE is created to visualize the distribution of the client's age. Discussed in detail in derived fields section.	None	
DAYS_EMPLOYED	How many days before the application the person started current employment. A derived field APPLICANT_EXPERIENCE is created to visualize the distribution of the client's experience in their current job. Discussed in detail in derived fields section.	None	

1.2 Previous Applications

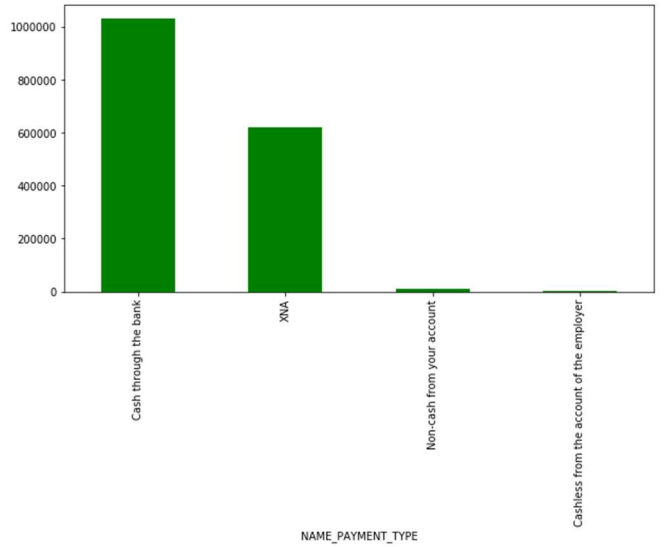
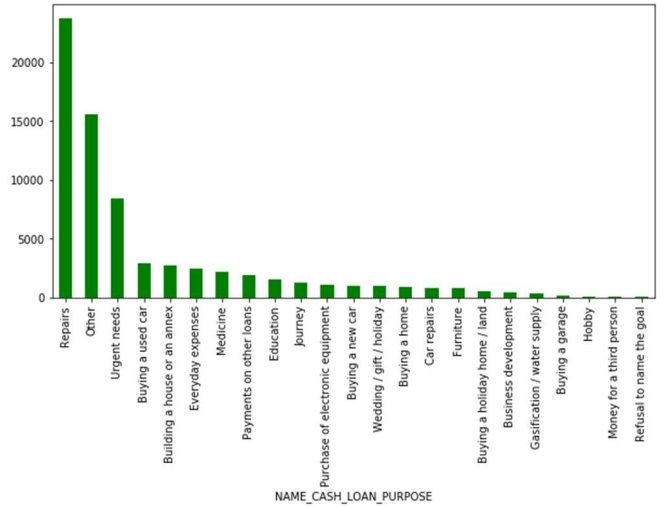
Previous Application - Essential Fields - Categorical Variables

Column Name	Explanation	Null	Plot
NAME_CONTRACT_TYPE	Contract type can be a cash loan, consumer loan or revolving loan. Cash loans: 747147 Consumer loans: 728870 Revolving loans: 184590 XNA: 346	None	
Note that the application data does not contain consumer loans at all. The sample data has only cash loan and revolving loan.			
NAME_CONTRACT_STATUS	The number of loans that were approved, refused, cancelled or unused in the past. Approved: 1036044 Refused: 282169 Cancelled: 316317 Unused Offer: 26423 **Approval Rate:** 77% of the loans were approved.	None	

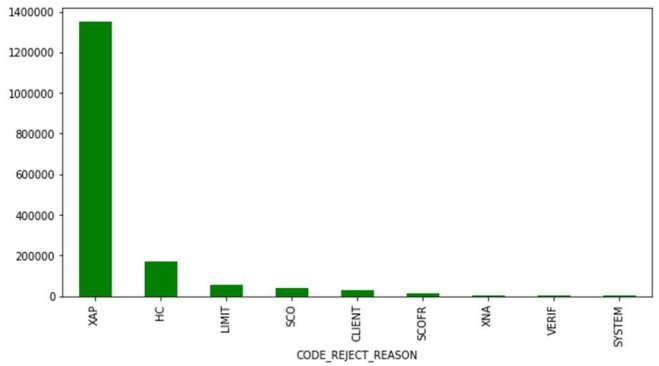
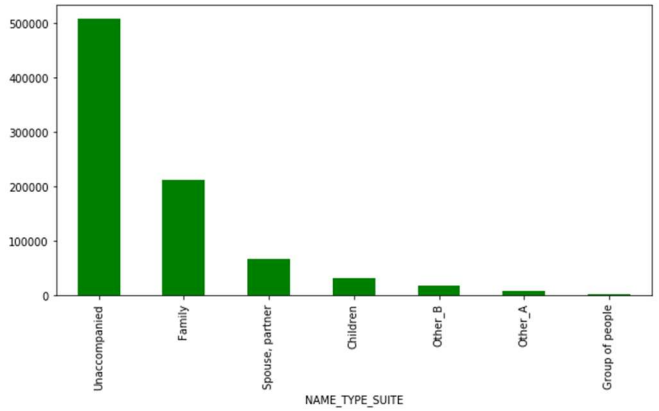
Previous Application - Essential Fields - Categorical Variables

Column Name	Explanation	Null	Plot
NAME_CLIENT_TYPE	Whether the client is new or repeated. Acquisition and retention. Repeater: 1222935 New: 301199 Refreshed: 134882 XNA: 1937 (see whether there are any historical data for these clients. If yes, mark as Repeater, else mark as a New client)	None	
NAME_PRODUCT_TYPE	The product type is having XNA for 1063381 records. Though there are no null values, out of 1660953 records 1063381 are XNA, which is 64%. It is better to drop the column from further analysis.	None	

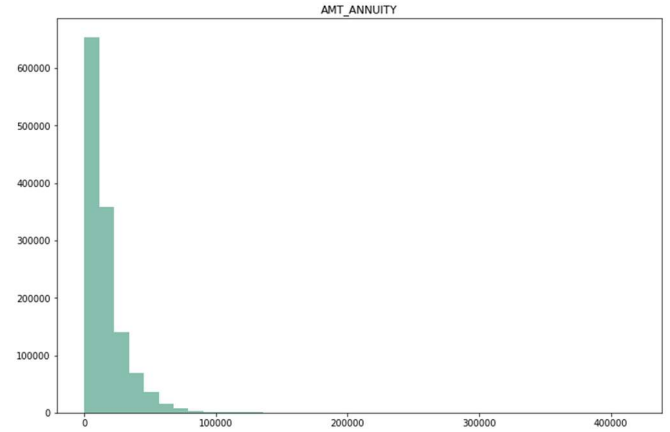
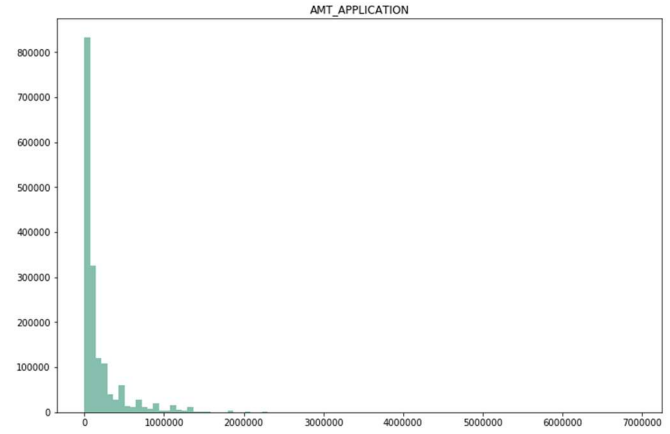
Previous Application - Essential Fields - Categorical Variables

Column Name	Explanation	Null	Plot
NAME_PAYMENT_TYPE	Payment type for repayment of loans. Cash through the bank: 1032927 XNA: 618752 Non-cash from your account: 8189 Cashless from the account of the employer: 1085 Though there are no null values 37% of the records are having the value XNA (not applicable). This may be corresponding to the loans that are not approved or unused offer or cancelled.	None	
NAME_CASH_LOAN_PURPOSE	Purpose for which the client applied for the loan. XAP: 913806 and XNA: 677576 need to be removed to see the distribution. We see that most of the loans were applied by the clients for the purpose or repairs or unforeseen urgent needs. But a significant portion of the clients did not share the real purpose.	None	

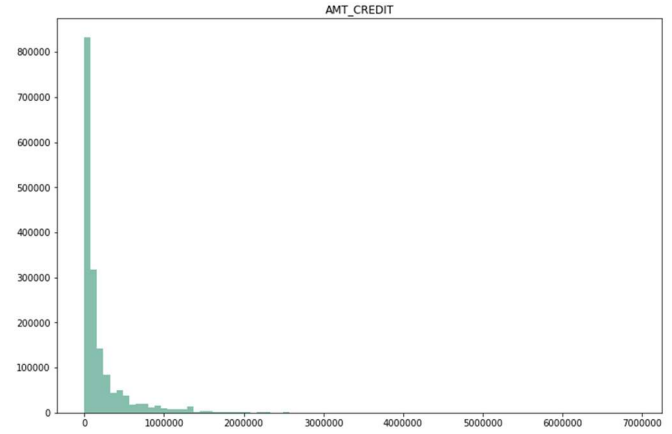
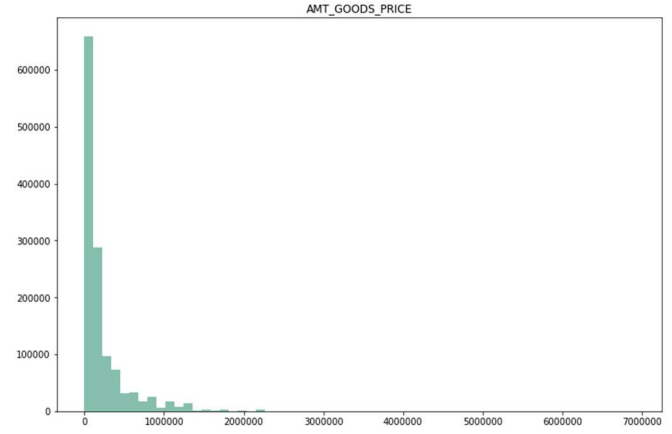
Previous Application - Essential Fields - Categorical Variables

Column Name	Explanation	Null	Plot
CODE_REJECT_REASON	Reason for refusing the loan. For approved or cancelled loans these fields have XAP or XNA (1352353 records). Among the records where the loan was refused, we have 4110 records that has the value XNA.		
NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan. For the missing values, we can update it as Unaccompanied (mode). Unaccompanied: 507349 Family: 212987 Spouse, partner: 66992 Children: 31537 Other_B: 17608 Other_A: 9071 Group of people: 2237		

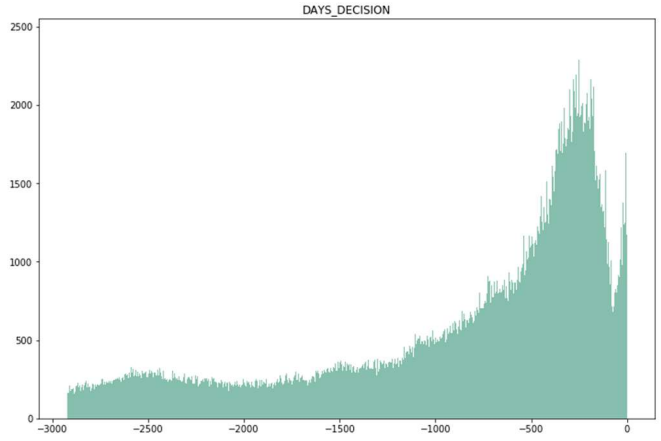
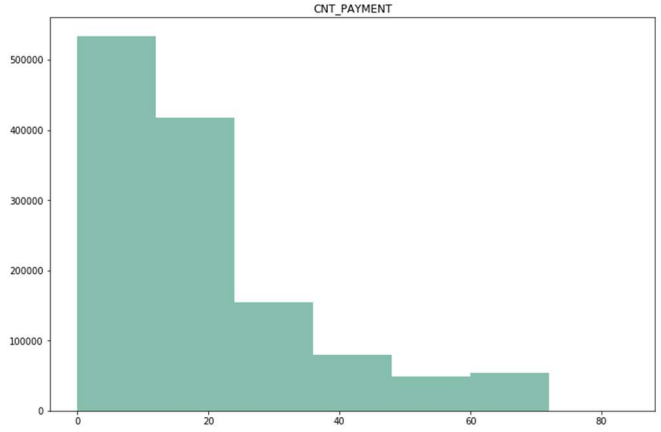
Previous Application - Essential Fields - Quantitative Variables

Column Name	Explanation	Null	Plot
AMT_ANNUITY	Annuity of previous application. Most of the null values correspond to the refused, cancelled or unused offer loans. count: 1036040 min: 0 max: 393868.7 mean: 14715.91 std: 13798.64 Q1: 5939.044 Q2: 10377.18 Q3: 18325.52 IQR: 12386.476 Outliers: Values greater than 36905.234	372214 Only 4 on approved loans	
AMT_APPLICATION	How much credit did client ask. count: 1036044 min: 0 max: 5850000 mean: 180501.6 std: 253533 Q1: 45000 Q2: 90000 Q3: 196130.7 IQR: 151130.7 Outliers: Values greater than 422826.75	0	

Previous Application - Essential Fields - Quantitative Variables

Column Name	Explanation	Null	Plot
AMT_CREDIT	Final credit amount on the previous application. count: 1036043 min: 0 max: 4509688 mean: 202465 std: 275196.1 Q1: 47970 Q2: 102145.5 Q3: 225000 IQR: 177030 Outliers: Values greater than 490545	1	
AMT_GOODS_PRICE	Goods price of good that client asked for (if applicable) on the previous application. count: 993275 min: 0 max: 5850000 mean: 188273.8 std: 256092.6 Q1: 47011.05 Q2: 96705 Q3: 206550 IQR: 159538.95 Outliers: Values greater than 445858.425	380429 42769 on approved loans - mostly revolving loans	

Previous Application - Essential Fields - Quantitative Variables

Column Name	Explanation	Null	Plot
DAYS_DECISION	Relative to current application when was the decision about previous application made. We see that most of the clients apply for repeated loans within 500 days of their previous loan application. Based on the next field CNT_PAYMENT we see that most of the loans are cleared within 12 payments.	None	
CNT_PAYMENT	Term of previous credit at application count: 1036040 min: 0 max: 84 mean: 14.11922 std: 11.9608 Q1: 6 Q2: 12 Q3: 18 IQR: 12 Outliers: Values greater than 36 The median is 12 i.e. most of the loans given are repaid within 12 payments.	372213 Only 4 for approved loans	

1.3 Data Quality Issues

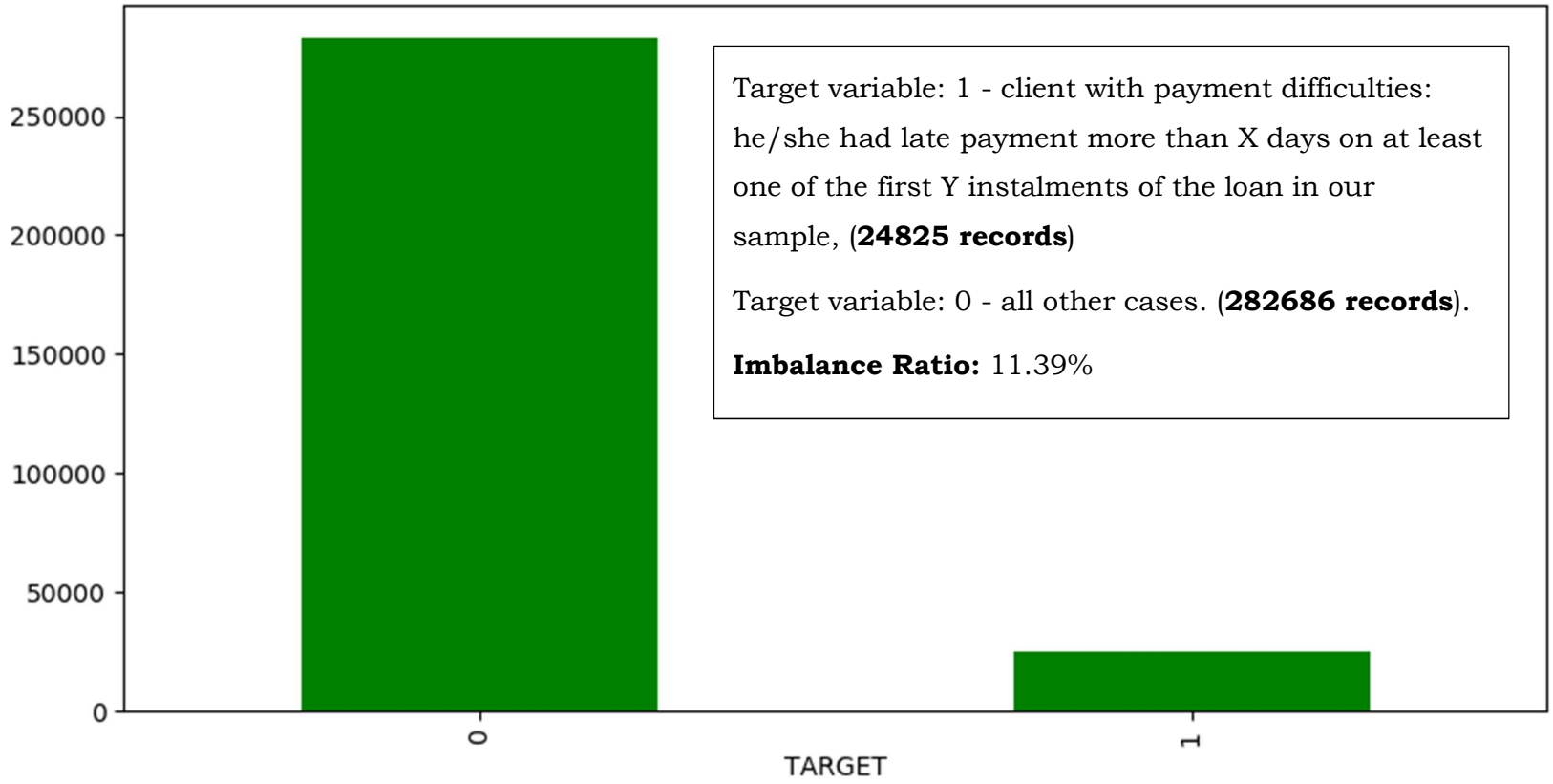
In the data dictionary presented above, we have identified the null values, missing values and any outliers in the features. How to treat the null values or missing values are also explained. There are fields that have no null values, but has the value XNA or XAP for missing values.

APPLICATION DATA	PREVIOUS APPLICATION
CODE_GENDER: 4 XNA values. We will update it with the mode (F). **ORGANIZATION_TYPE:** No null values. But there are fields with value XNA. Can be replaced with 'Other'. **OCCUPATION_TYPE:** 96391 null values. We will update it as 'Unknown'. **CNT_FAM_MEMBERS:** 2 null values. We will update it with median value 2. **NAME_TYPE_SUITE:** 1292 null values. We will update it as Unaccompanied. **AMT_GOODS_PRICE:** 278 null values. The null values are corresponding to the Revolving Loans. We assume that AMT_GOODS_PRICE = AMT_CREDIT. **AMT_ANNUITY:** 12 null values. Approximately (in most of the records) this is 1/20 of the AMT_GOODS_PRICE for Revolving Loans. But the null values are corresponding to cash loans where there is no such relationship. It is better to drop these 12 rows than it affecting our analysis. **OWN_CAR_AGE:** Those who do not own car, for them this value is null. If this value is NULL we will mark FLAG_OWN_CAR = N (to ensure consistency in the dataset).	**AMT_DOWN_PAYMENT:** Down payment on the previous application. 886863 null values. This cannot be treated without knowing whether it is allowed to not make any down payment to get the loan. **CODE_REJECT_REASON:** Why was the previous application rejected. For approved or cancelled loans these fields have XAP or XNA (1352361 records). Among the records where the loan was refused, we have 4110 records that has the value XNA. **CNT_PAYMENT:** Term of the previous credit. 372213 null values. For cancelled, refused or unused offer, this field need not have any value. Only 4 Approved loans are having null values. These records can be dropped. **DAYS_LAST_DUE:** Relative to application date of current application when was the last due date of the previous application. **664513** null values. 211047 records has the value 365243 (invalid value). **DAYS_TERMINATION:** When was the expected termination of the previous application. **664513** null values. 39604 approved loans have null values in DAYS_LAST_DUE, DAYS_TERMINATION. These records can be dropped. 225745 records have the value 365243 (invalid value). **DAYS_FIRST_DRAWING** has 664513 null values and 933777 records with invalid value 365243.

1.4 DATA IMBALANCE

The sample data provided has imbalance with respect to target variable.

From the frequency plot we see that the sample selected has very few clients with payment difficulties (8% only).



Note: We will segregate these data in two data frames for further analysis.

2. Data Preparation

Merging Datasets &

Creating Derived Variables

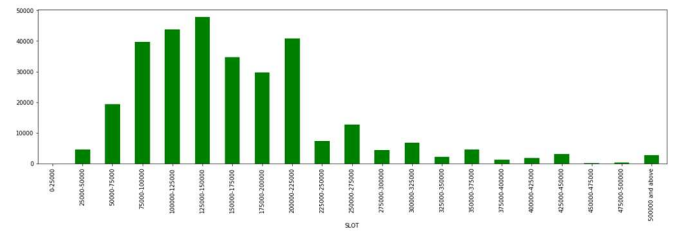
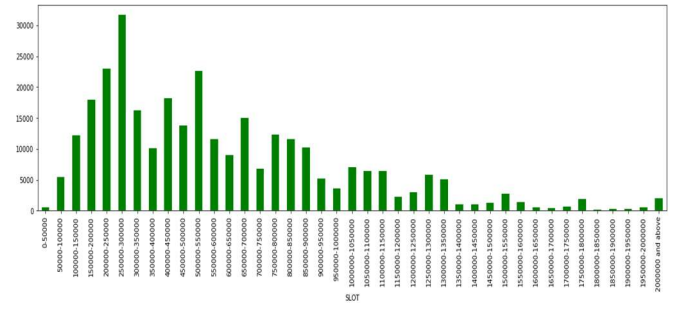
MERGING THE DATASETS

We have two data sets: application_data.csv and previous_applications.csv.

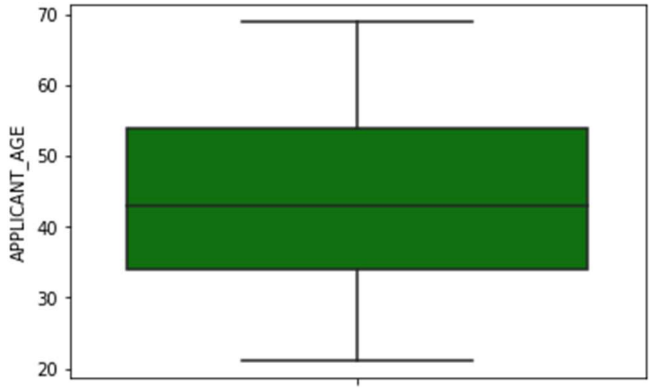
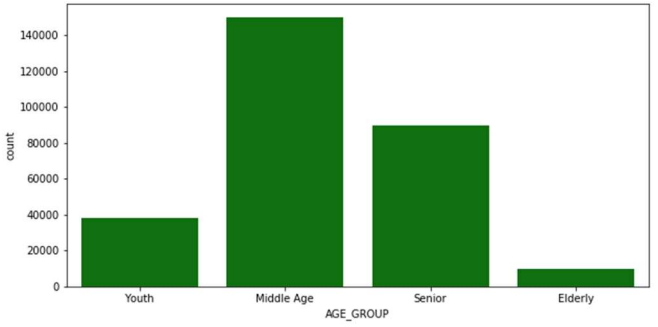
We have selected the fields for our analysis and loaded them into data frames df_applicant_profile and df_prev_app_essential.

Now it is time to merge these datasets and do the further analysis.

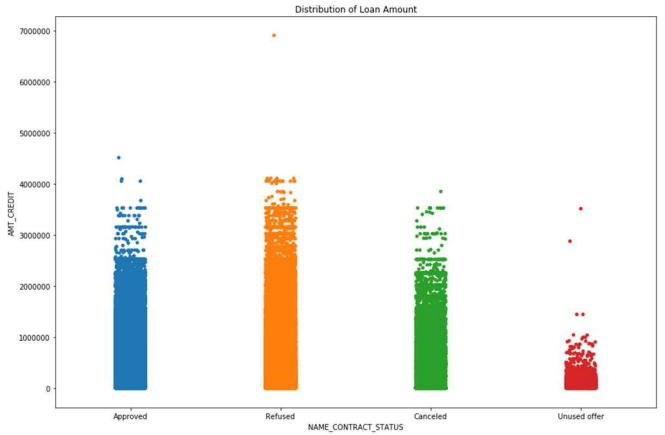
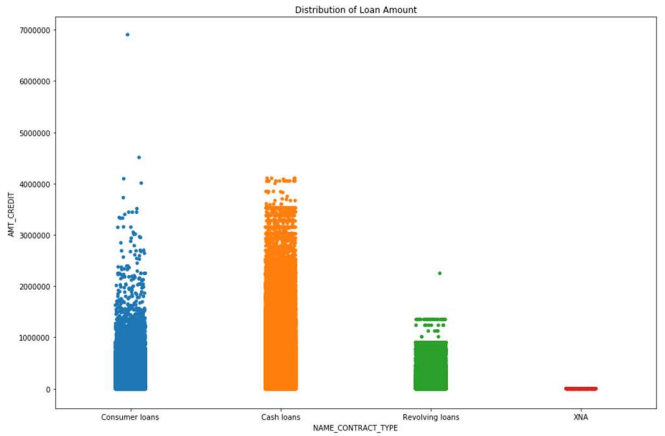
Derived Fields - Merged Dataset

Column Name	Explanation	Null	Plot
AMT_INCOME_RANGE	Created from APPLICANT_PROFILE Bins to categorize the income of the client.	None	
AMT_CREDIT_RANGE	Created from APPLICANT_PROFILE Bins to categorize the loan amount credited.	None	

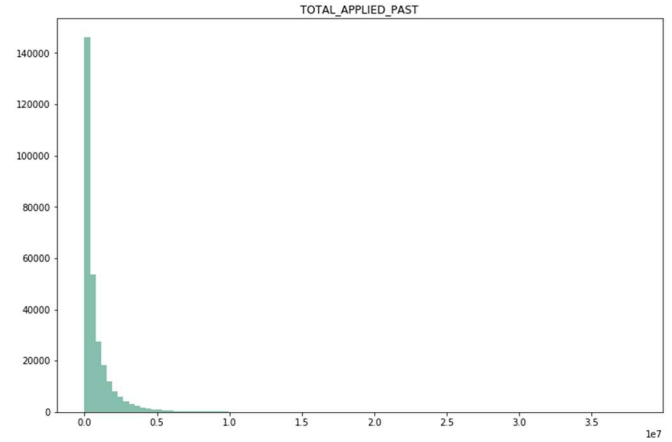
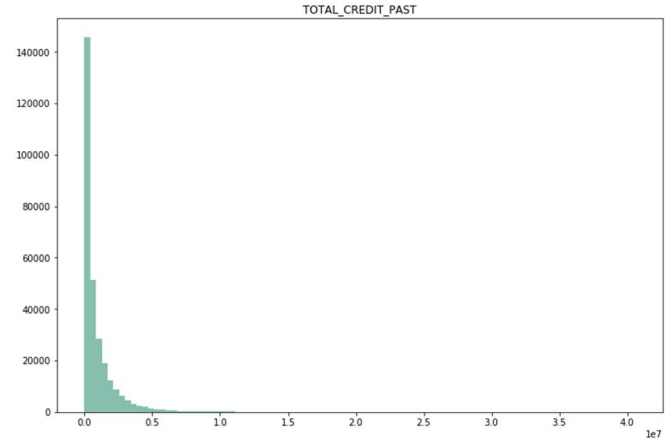
Derived Fields - Merged Dataset

Column Name	Explanation	Null	Plot
APPLICANT_AGE	Created from APPLICANT_PROFILE.DAYS_BIRTH The age of client in years. Count: 291015 records. Minimum: 21 Maximum: 69 Mean: 43.94 Q1: 34 Q2: 43 Q3: 54 IQR = 20 Outliers are out of the range (4, 84). But in this case there are no outliers.	None	
AGE_GROUP	Classified the clients based on their age as Young (18-29) Middle Age (30-49) Senior (50-64) Elderly (> 64)	None	

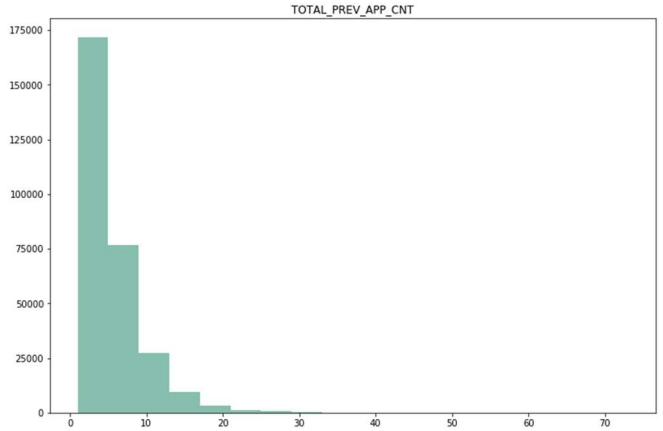
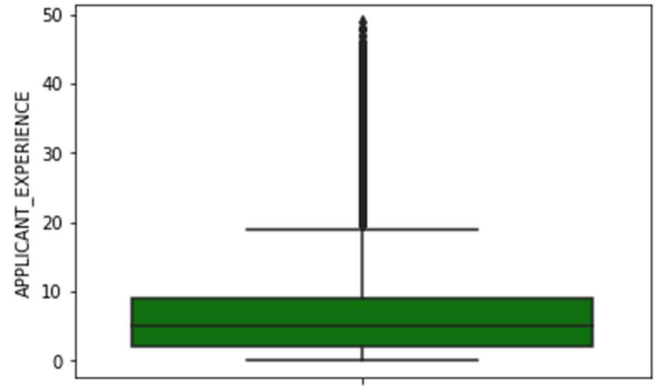
Derived Fields - Merged Dataset

Column Name	Explanation	Null	Plot
PAST_APPROVED_LOANS	How many of the past applications were approved? Vary from 0 - 23 with mean and median at 3.	None	
PAST_REFUSED_LOANS	How many of the past applications were refused? The mean and median are 0, which indicates that most of the loans are getting approved.	None	
PAST_CANCELED_LOANS	How many of the past applications were cancelled? The mean and median are 0 in this case also.	None	
PAST_UNUSED_LOANS	How many of the past applications were unused? The mean, median, Q1, Q2, Q3 are all 0.	None	
PAST_CASH_LOAN	How many cash loans the client has applied? Minimum: 0 Maximum: 60 Mean: 2.15 Q1: 0 Q2: 1 Q3: 3 IQR: 3 Outliers > 7.5	None	
PAST_CONSUMER_LOAN	How many consumer loans the client has applied? Minimum: 0 Maximum: 45 Mean: 2.14 Q1: 1 Q2: 2 Q3: 3 IQR: 2 Outliers > 6	None	
PAST_REVOLVING_LOAN	How many revolving loans the client has applied? Minimum: 0 Maximum: 31 Mean: 0.5 Q1: 0 Q2: 0 Q3: 1 IQR: 1 Outliers > 2.5	None	

Derived Fields - Merged Dataset

Column Name	Explanation	Null	Plot
TOTAL_APPLIED_PAST	How much amount the client has applied for in the past. Min: 0 Max: 37973710 Mean: 847098 Q1: 147438 Q2: 380430 Q3: 1003141	None	
TOTAL_CREDIT_PAST	How much amount was credited to the client in the past. Min: 0 Max: 40561130 Mean: 944952 Q1: 156776 Q2: 424886 Q3: 1143871	None	

Derived Fields - Merged Dataset

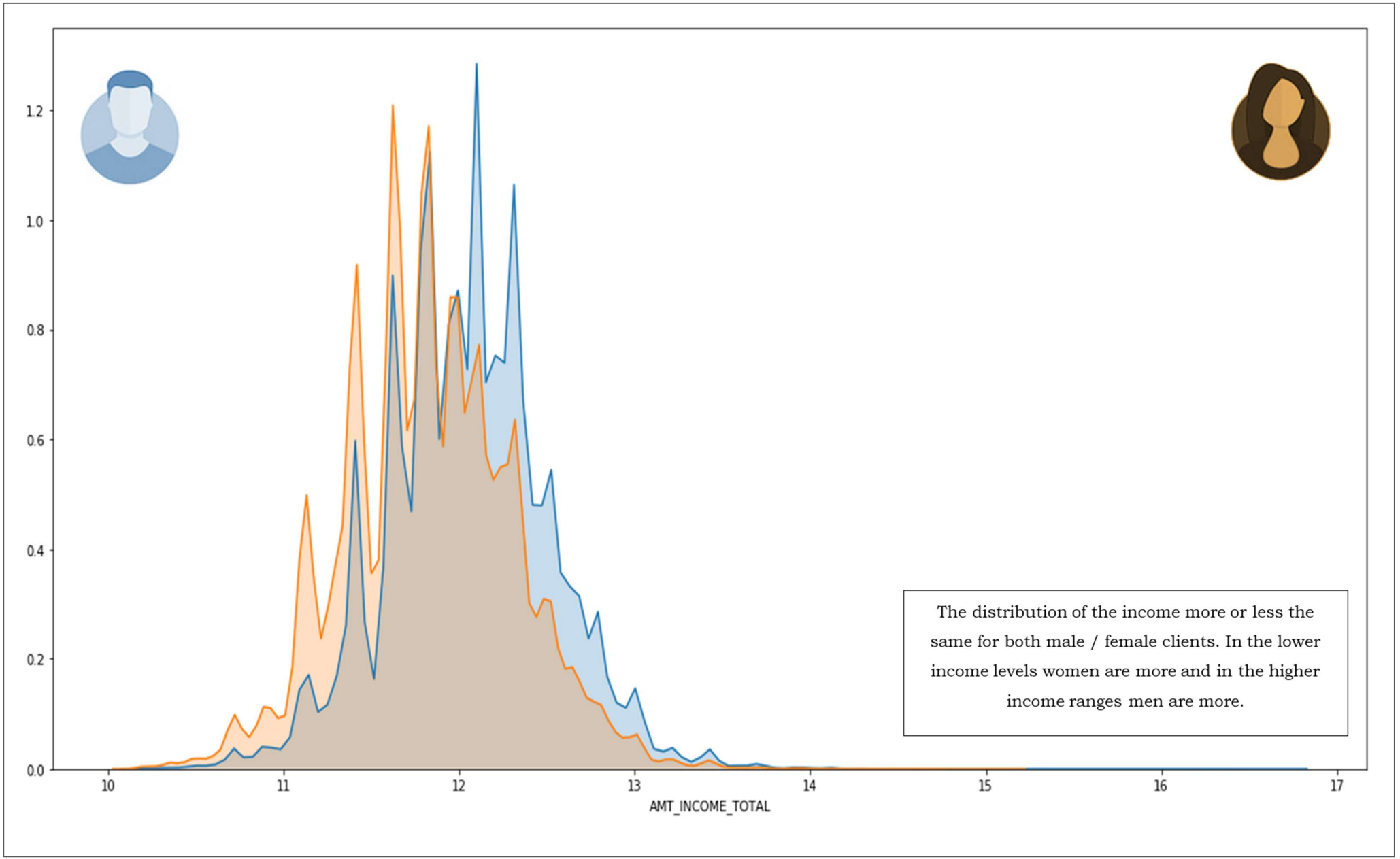
Column Name	Explanation	Null	Plot
TOTAL_PREV_APP_CNT	How many times the client has applied in the past? Minimum: 1 Maximum: 73 Mean: 4.83 Median: 4 Q1: 2 Q2: 4 Q3: 6 IQR: 4 Outliers are those applied more than 12 times.	None	
APPLICANT_EXPERIENCE	From APPLICANT_PROFILE.DAYS_EMPLOYED The experience of client in current job. Minimum: 0 Maximum: 49 Mean: 6.56 Q1: 2 Q2: 5 Q3: 9 IQR = 7 Outliers are greater than 19.5 years. But in this case there are no outliers.	49743	

3. Data Analysis

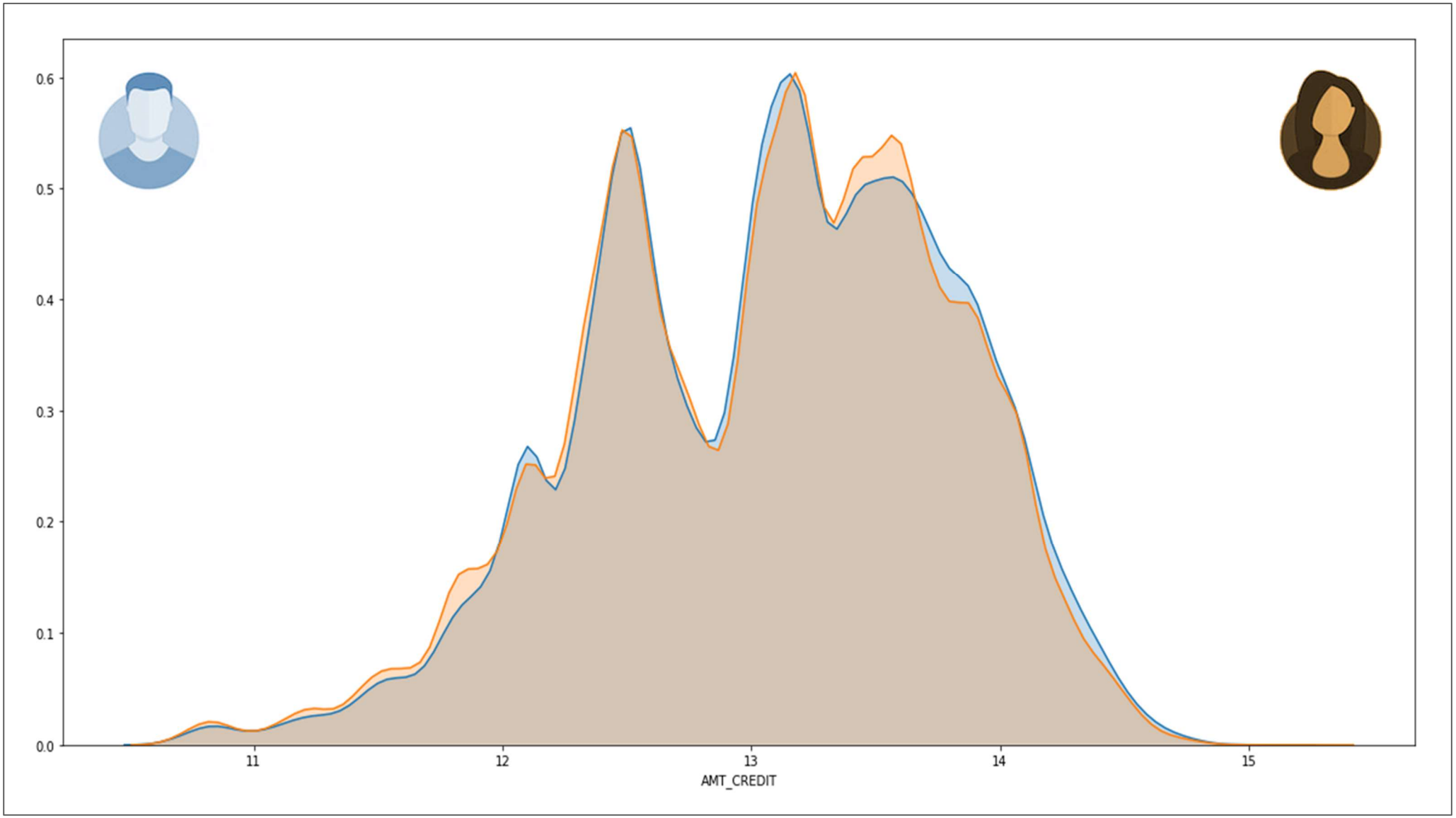
With Merged Data

3.1 Segmented Univariate Analysis

Income of the Clients - Segmented on Gender

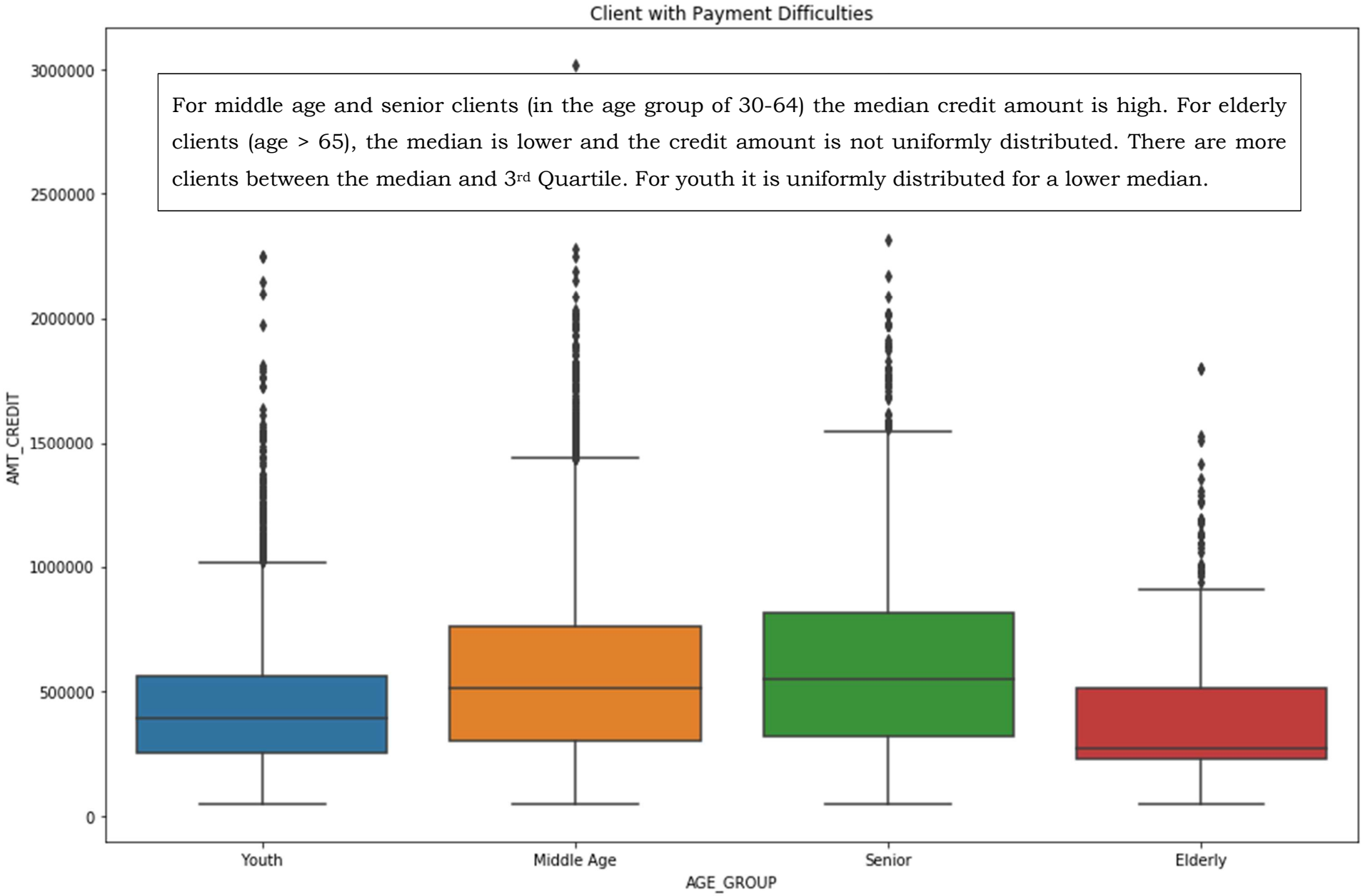


Credit to Clients - Segmented on Gender

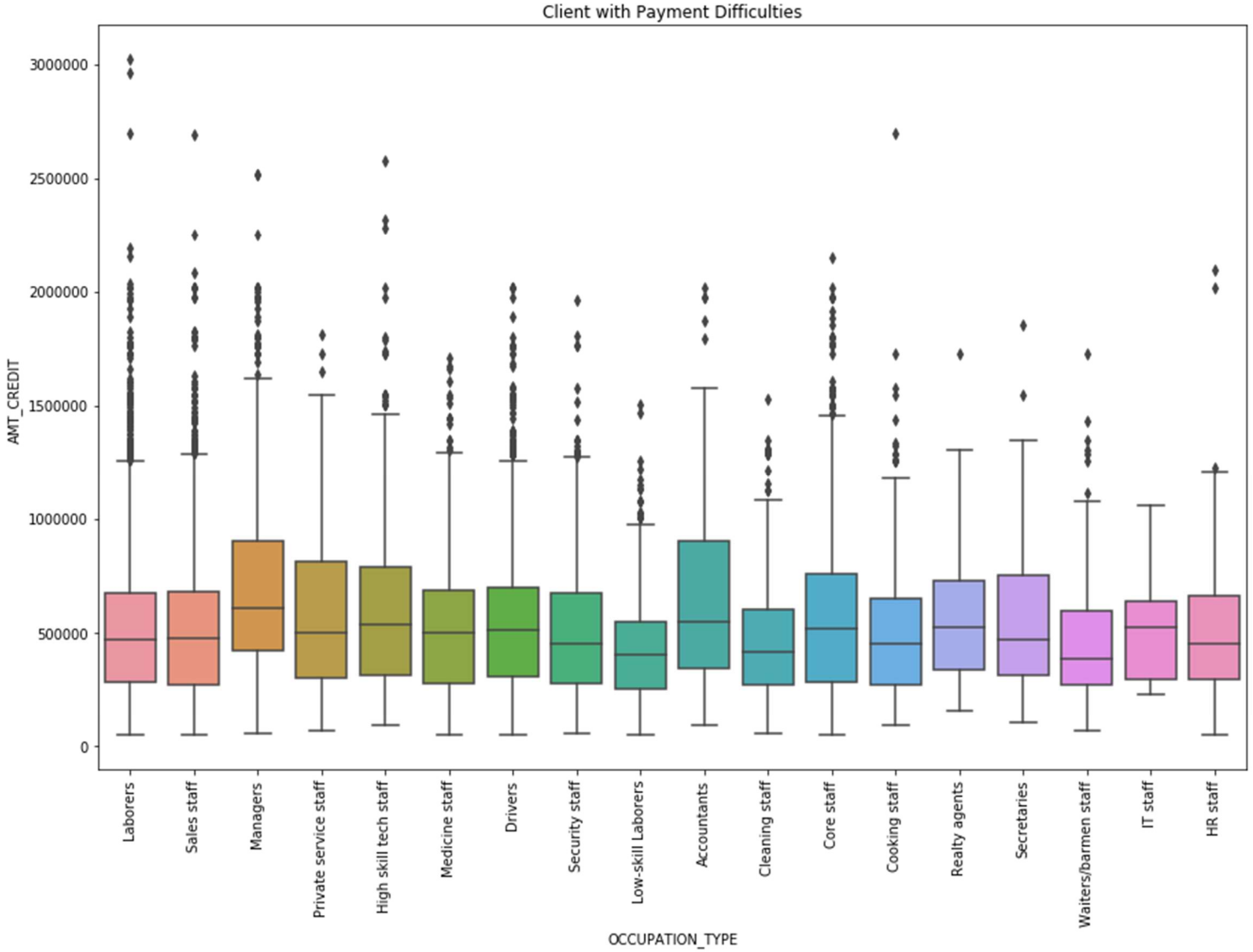


This is the distribution of the credit given to male and female clients. We do not see any discrimination as such.

Credit segmented on Age Group (for Clients with Payment Difficulties)

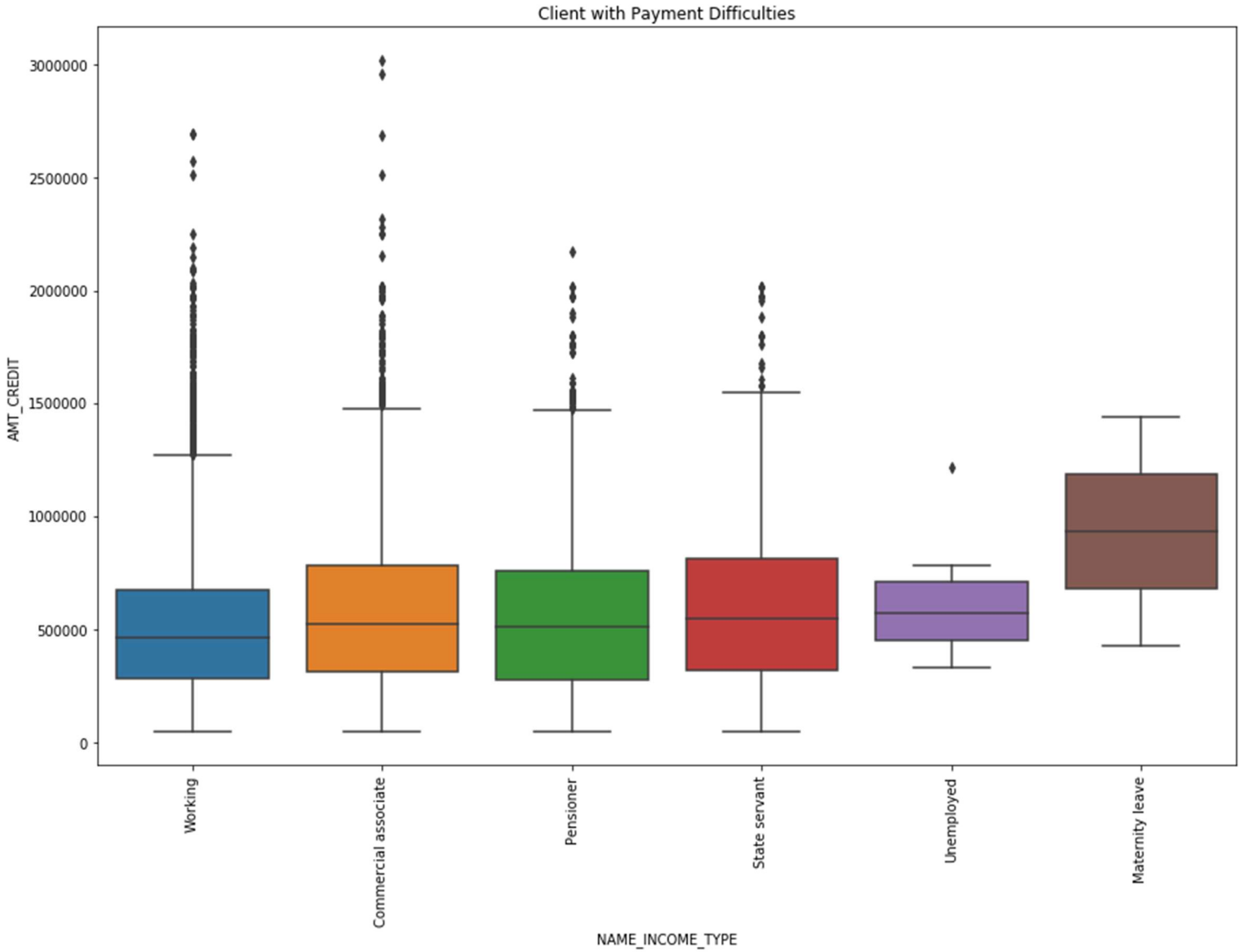


Credit segmented based on Occupation Type for clients with payment difficulties



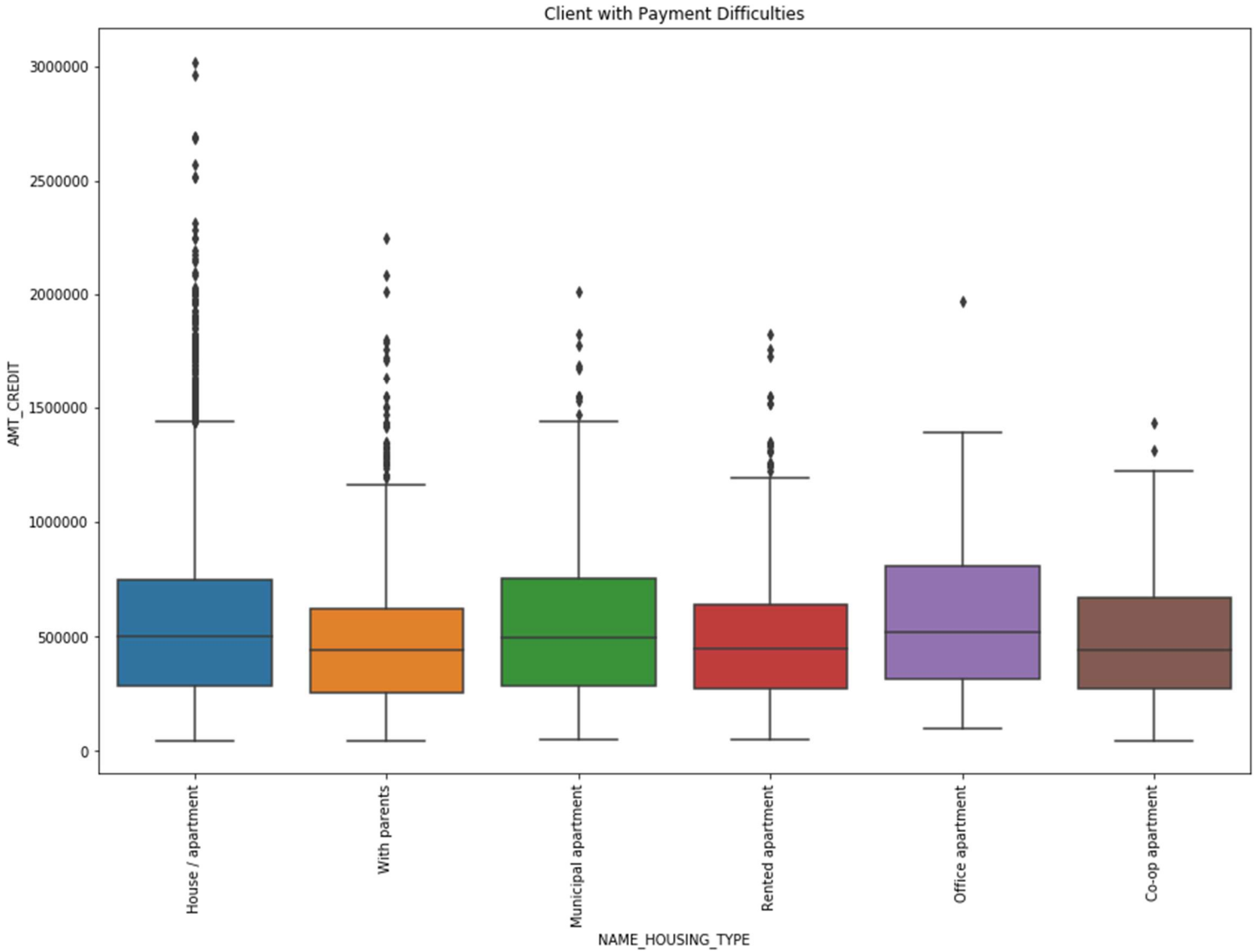
Managers have a higher credit limit followed by **realty agents** & **accountants**.

Credit segmented based on Income Type for clients with payment difficulties



The income of maternity leave is higher in the client with payment difficulties.

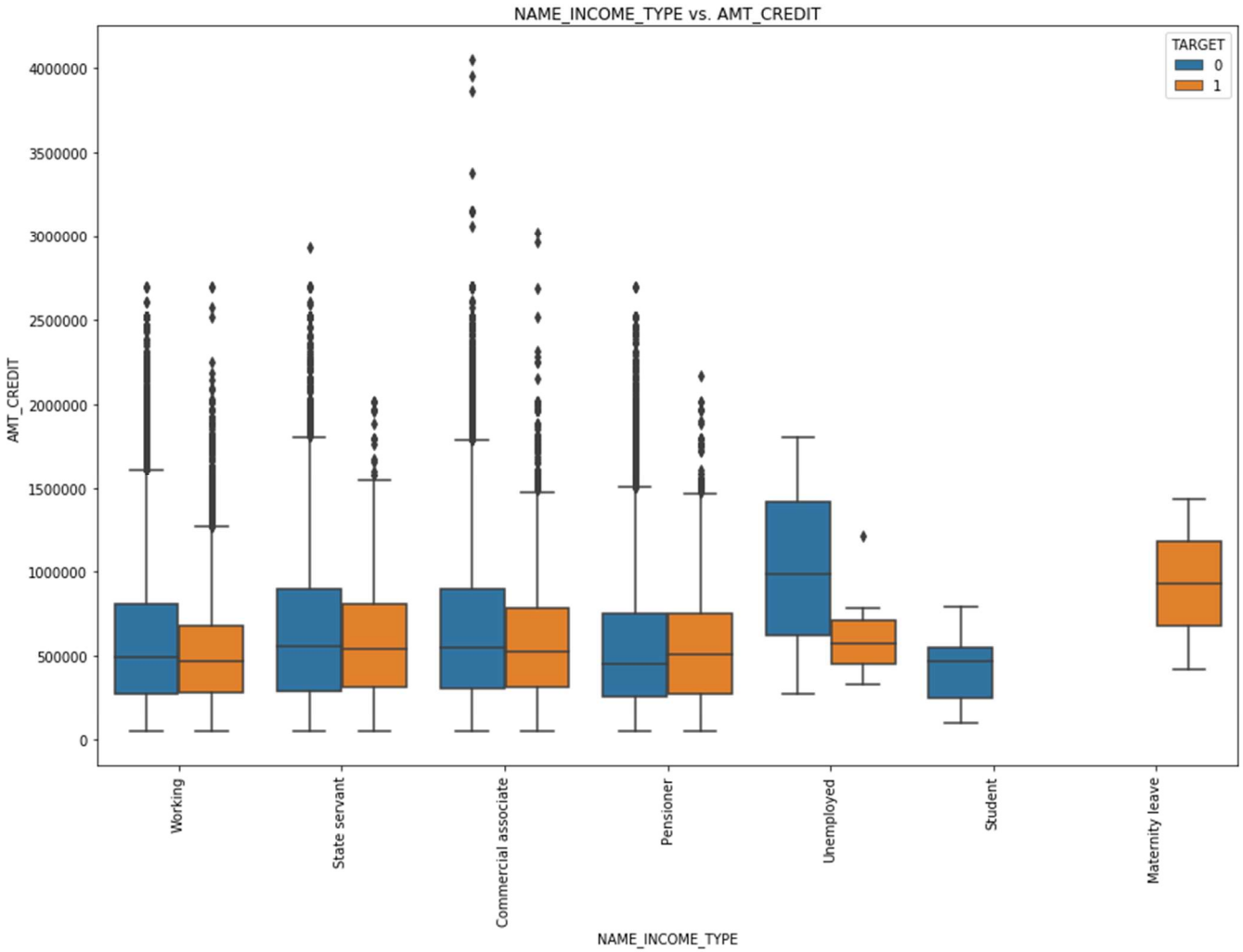
Credit segmented based on Housing Type for clients with payment difficulties



The median is almost in the same range for almost all different housing type. We need to study is further with bivariate analysis.

3.2. Bivariate Analysis

Income Type vs. Credit



Observations from the Plot:

The credit amount for the clients with different income type are plotted here. The clients with payment difficulty are highlighted in orange.

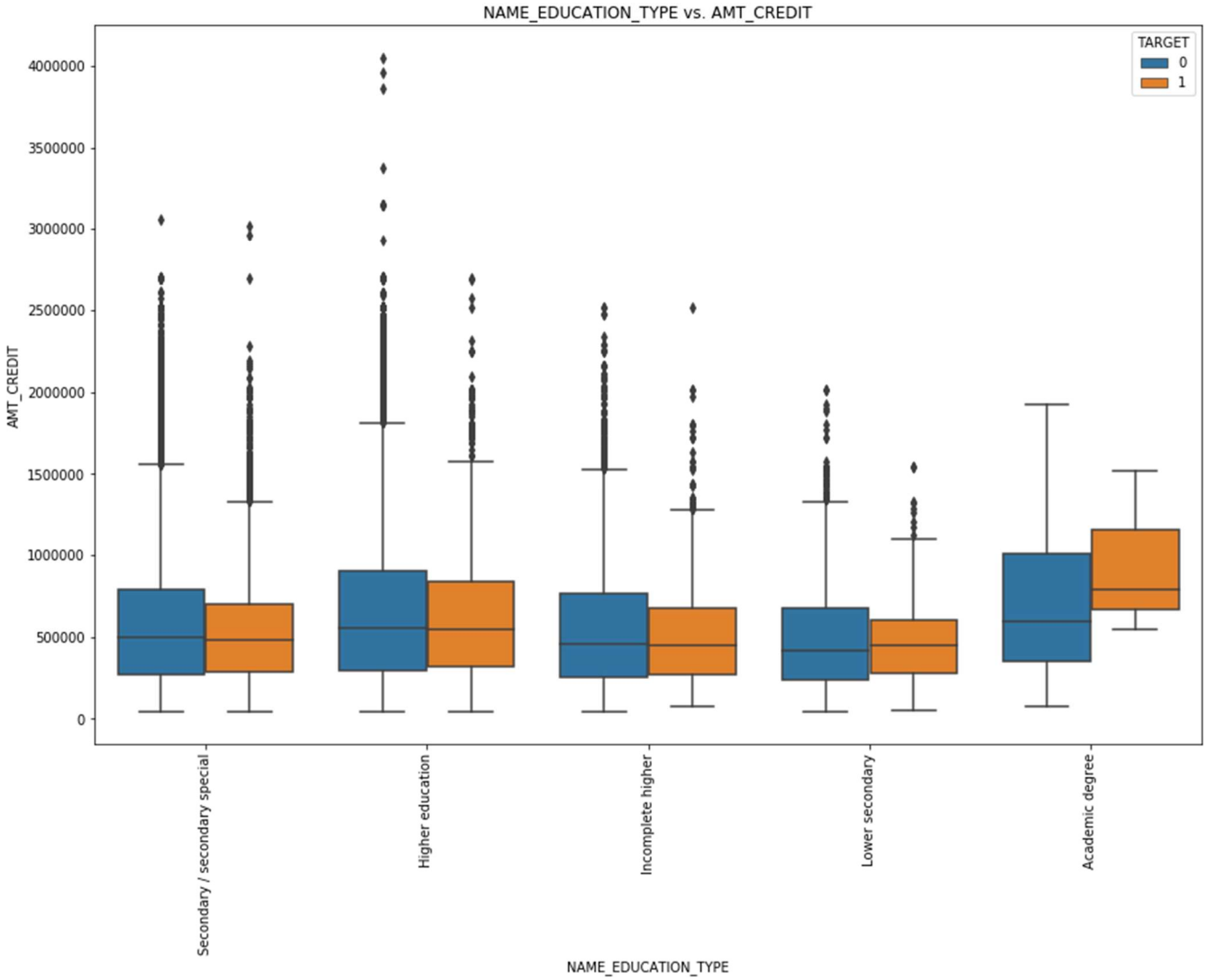
Surprisingly, the unemployed category is having a higher credit compared to others (and no outliers for this category).

In case of students there are no cases reported to have payment difficulty.

Median value for credit amount is in the range of 500000 (except for unemployed)

Note: Though there is imbalance in target variable, the box plot gives the median and the quartiles (percentiles) so the observations do not get skewed as with the case of the count of records.

Education Type vs. Credit



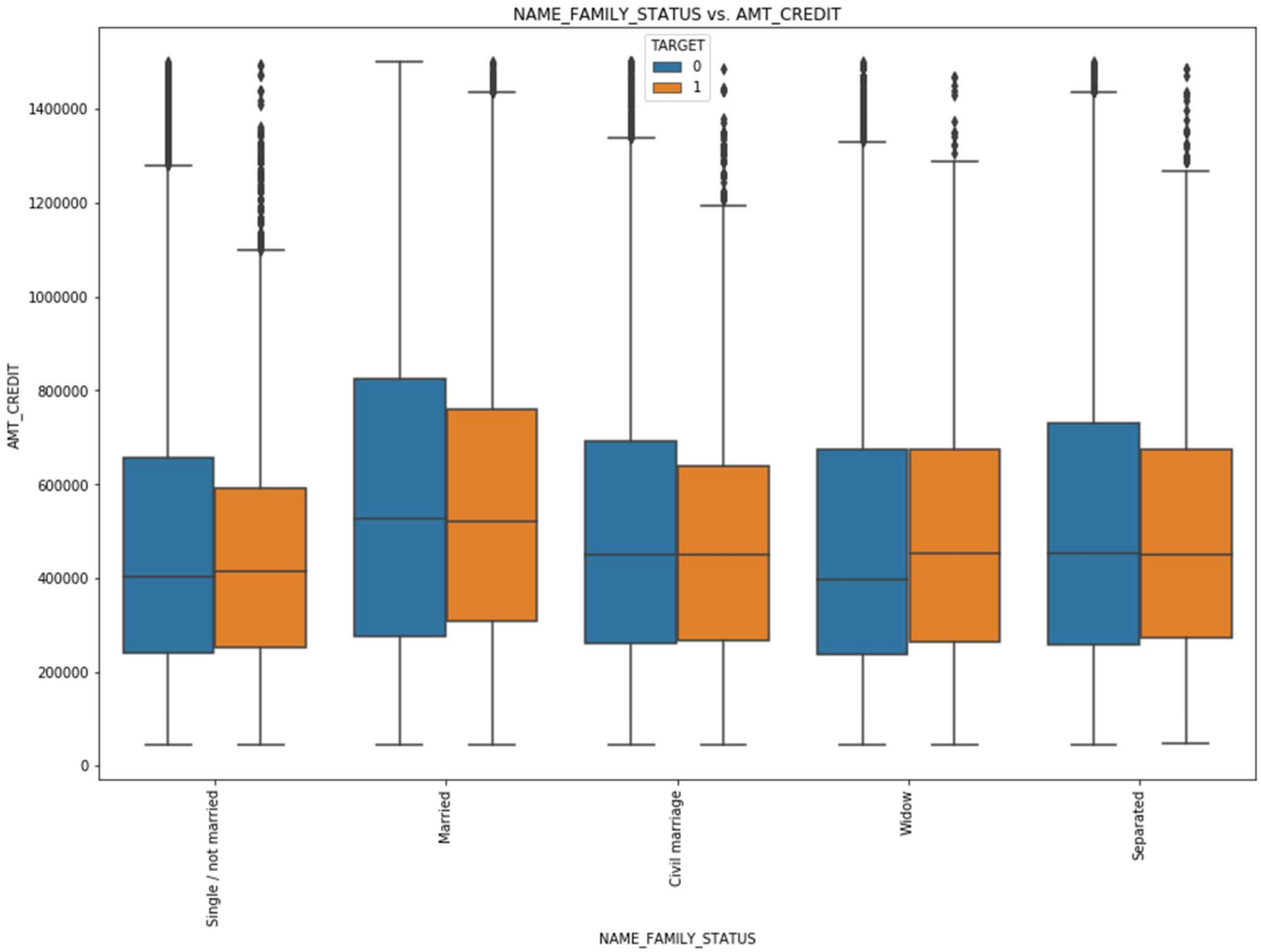
Observations from the Plot:

The number of clients with academic degree is very less (164 records) in the dataset. 71% of the applicants have only secondary education.

However, when we compute the percentile, we see that with whatever sample data we have the median value is higher for the clients with academic degree.

The credit amount progressively increases from lower secondary, to secondary, to incomplete higher, to higher education and to academic degree. This variable has direct impact on how much credit the client is eligible for.

Family Status vs. Credit

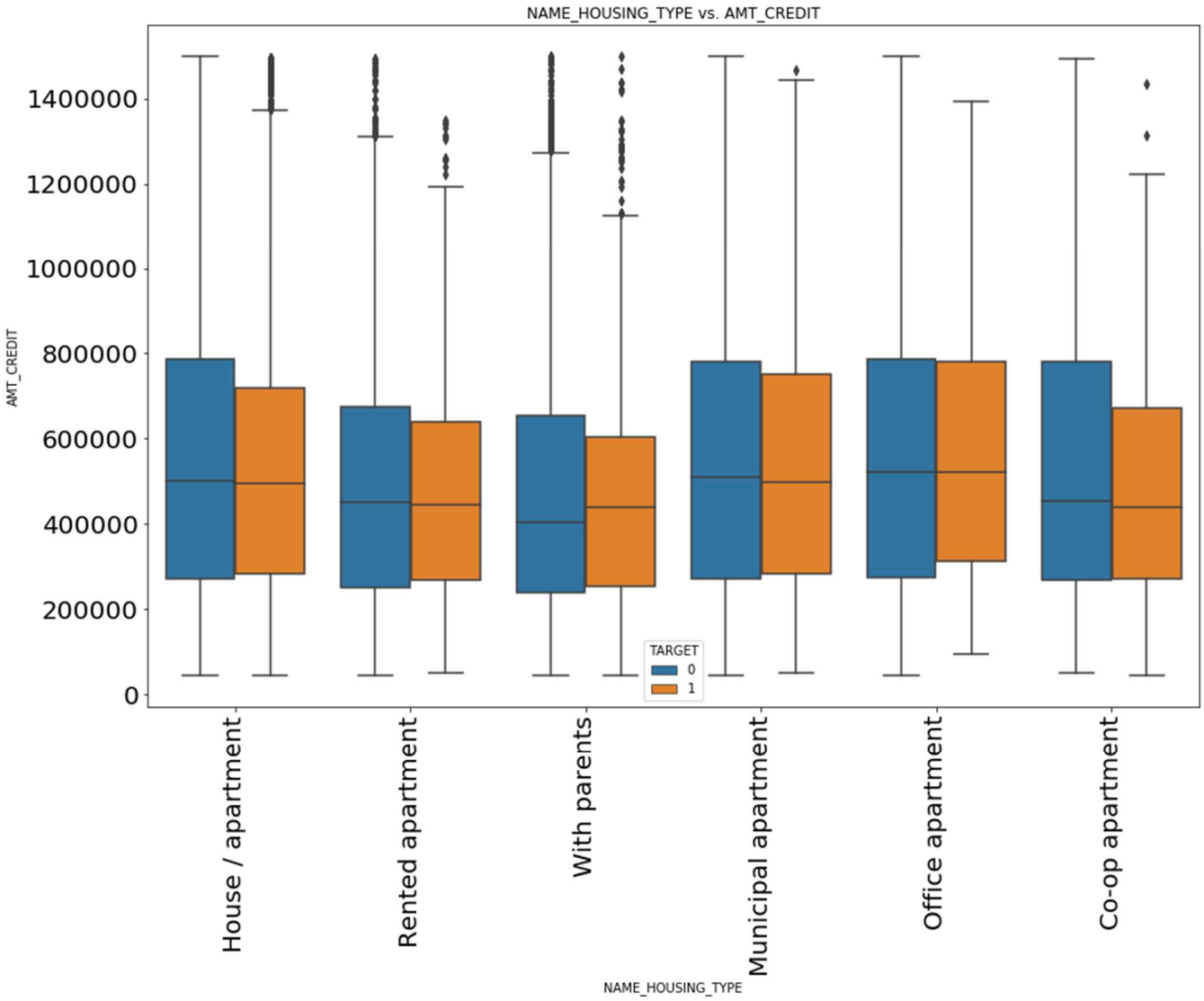


Observations from the Plot:

The credit median is comparatively higher for the married people and lesser for widows.

Note: This plot is presented after removing certain outliers. However, the outliers does not affect the median and quartiles.

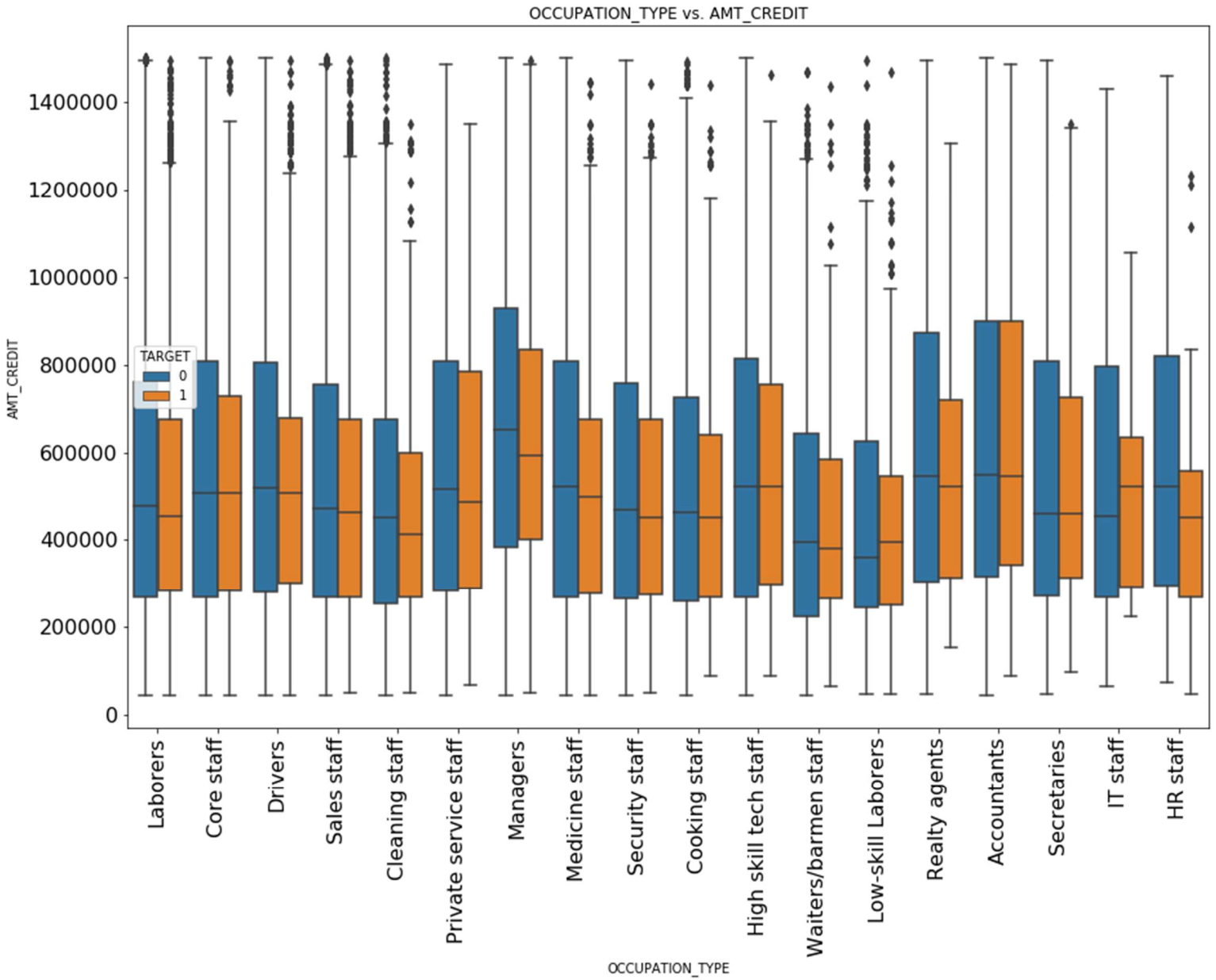
Housing Type vs. Credit



Observations from the Plot:

The plot is showing lower credit limit for those in Rented Apartments or staying with parents.

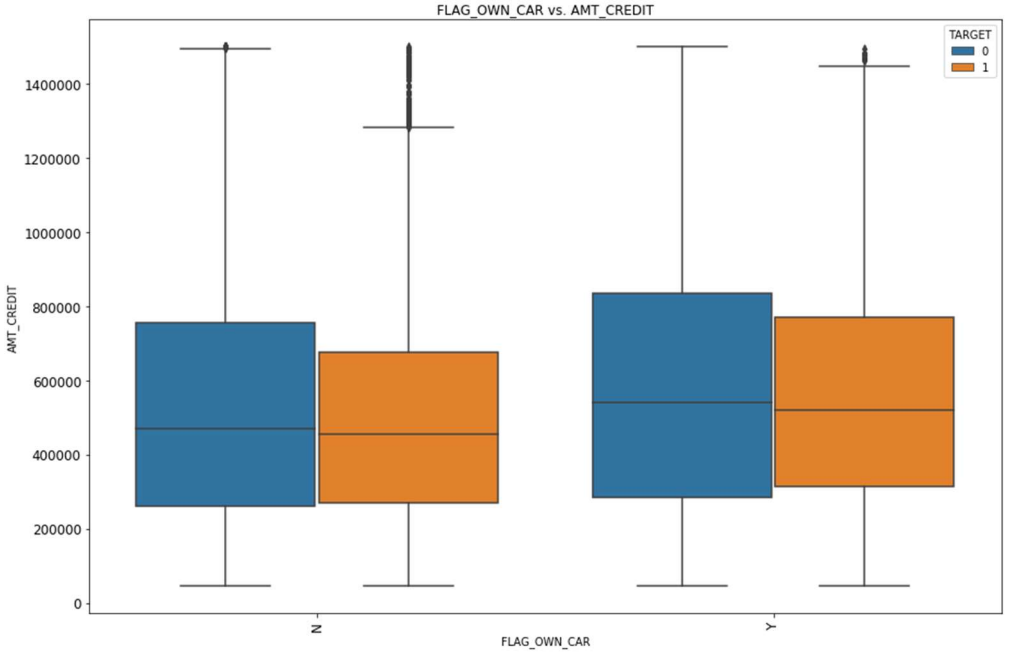
Occupation Type vs. Credit

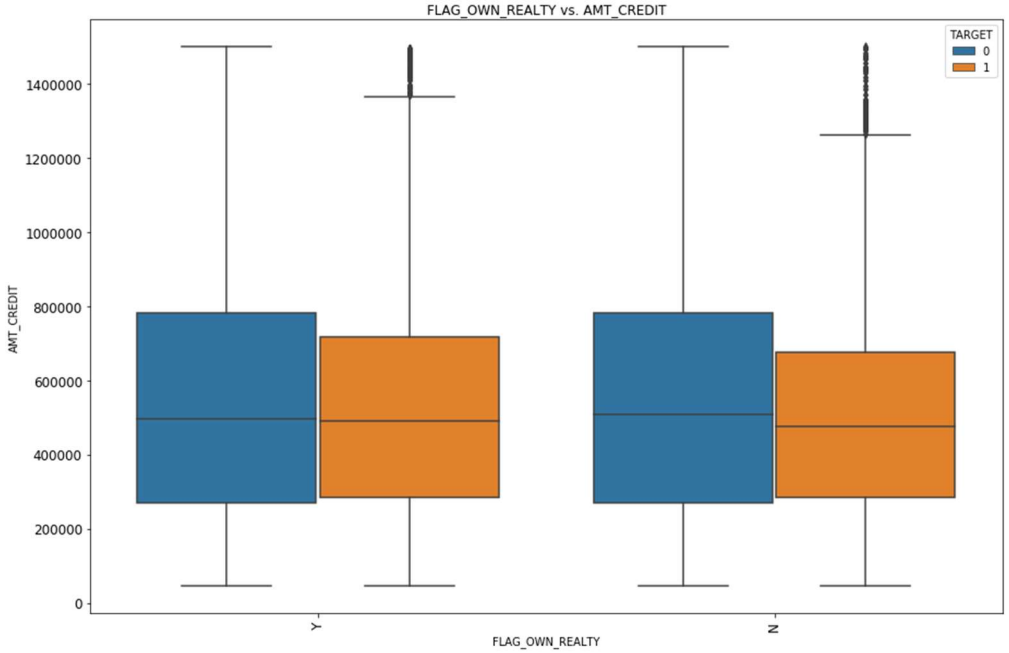


Managers have a higher credit limit followed by **realty agents** & **accountants.**

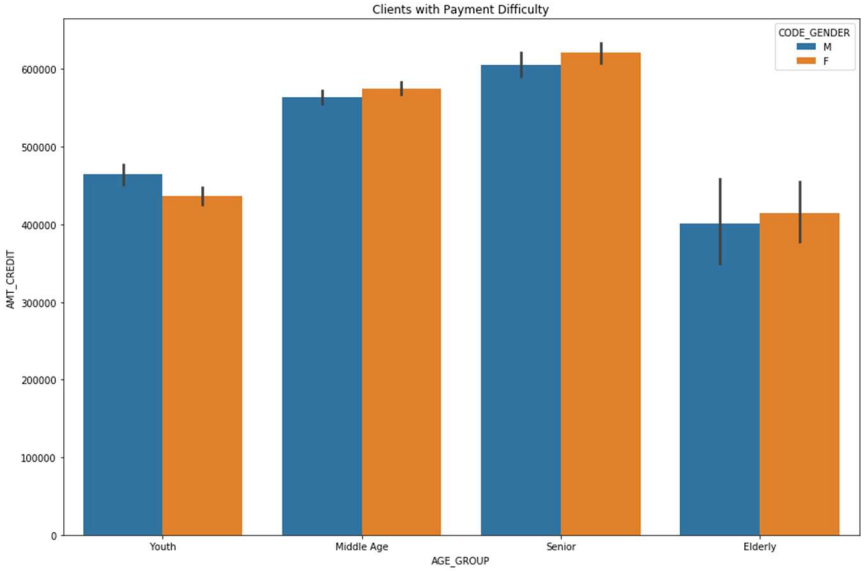
Credit for HR is better than that for IT.

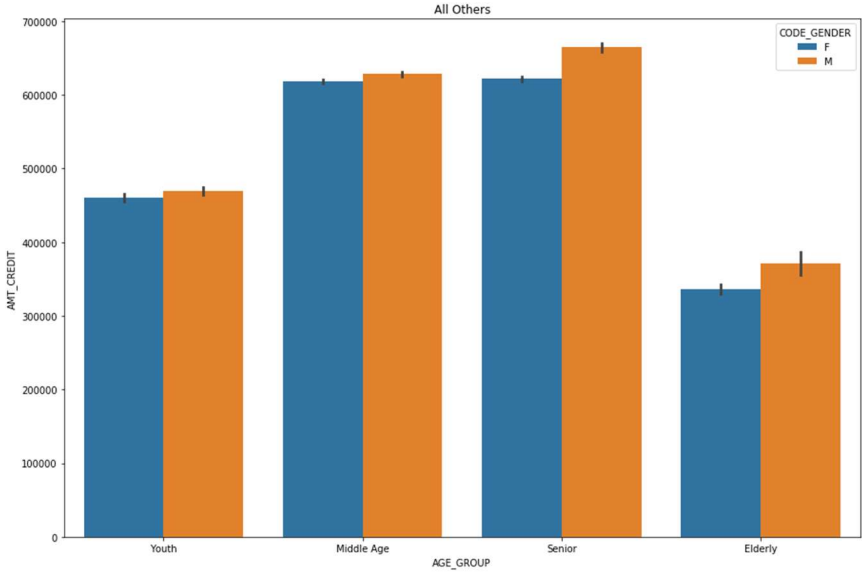
Own Car & Own Realty vs. Credit

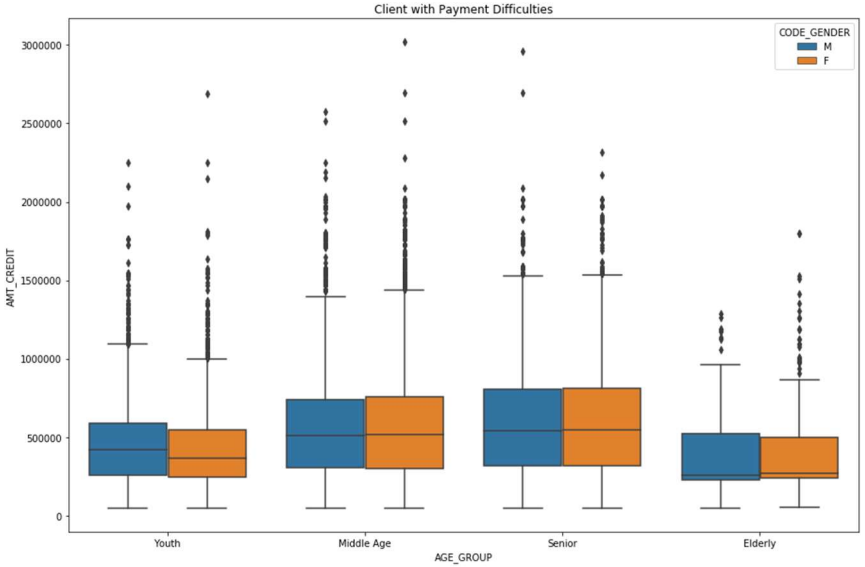


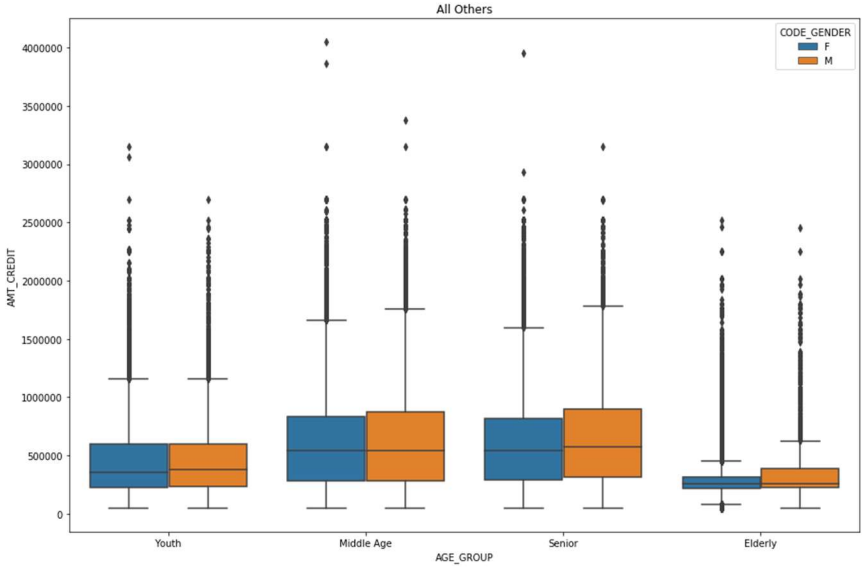


Impact of Age Group and Gender on the Target Variable and the Credit Amount



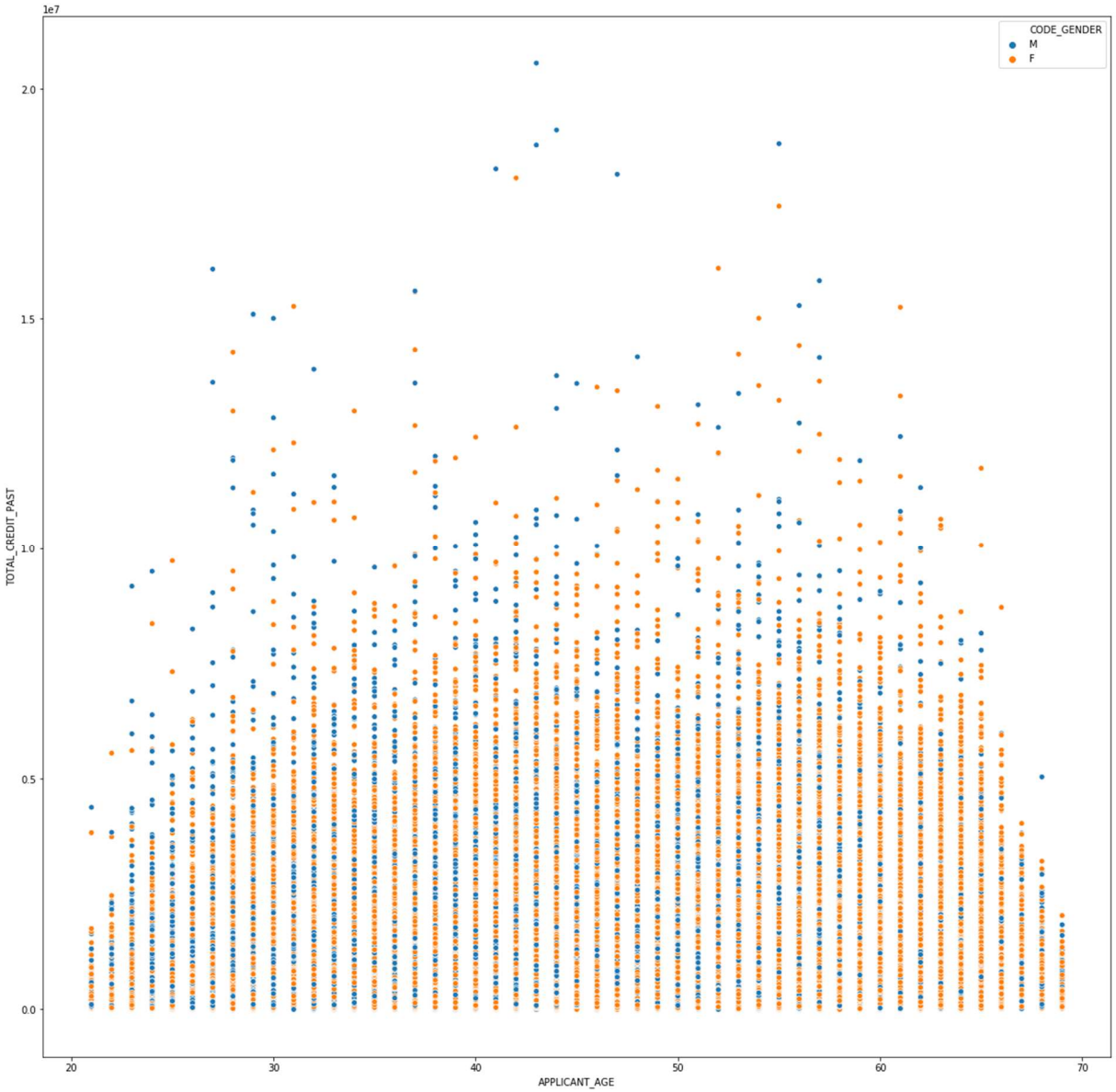






We have segregated the clients into four groups based on their age. Youth (18-29), Middle Age (30-49), Senior (50-64), Elderly (> 64). The following plots show the amount that is given as loan to the clients of different age groups, further showing the bifurcation based on gender. The only insight we get is that most of the loans are sanctioned to middle age or senior men. Gender does not have any influence.

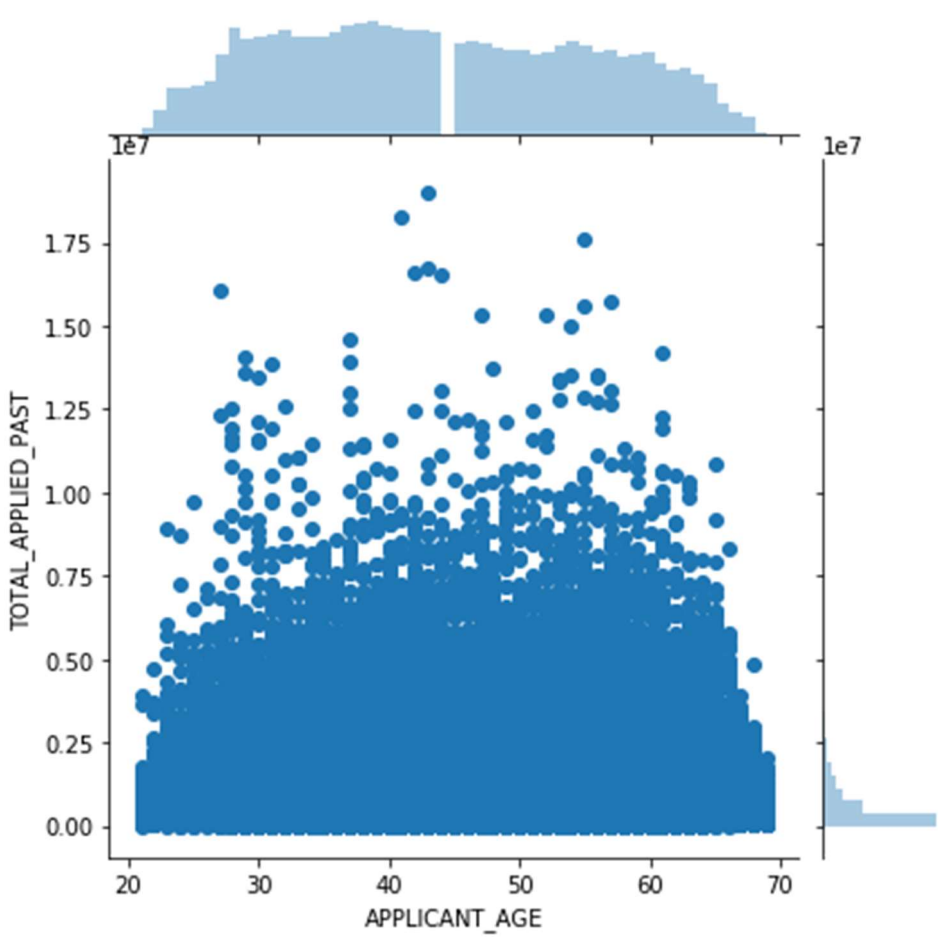
Impact of Age Group and Gender on the Target Variable and the Credit Amount

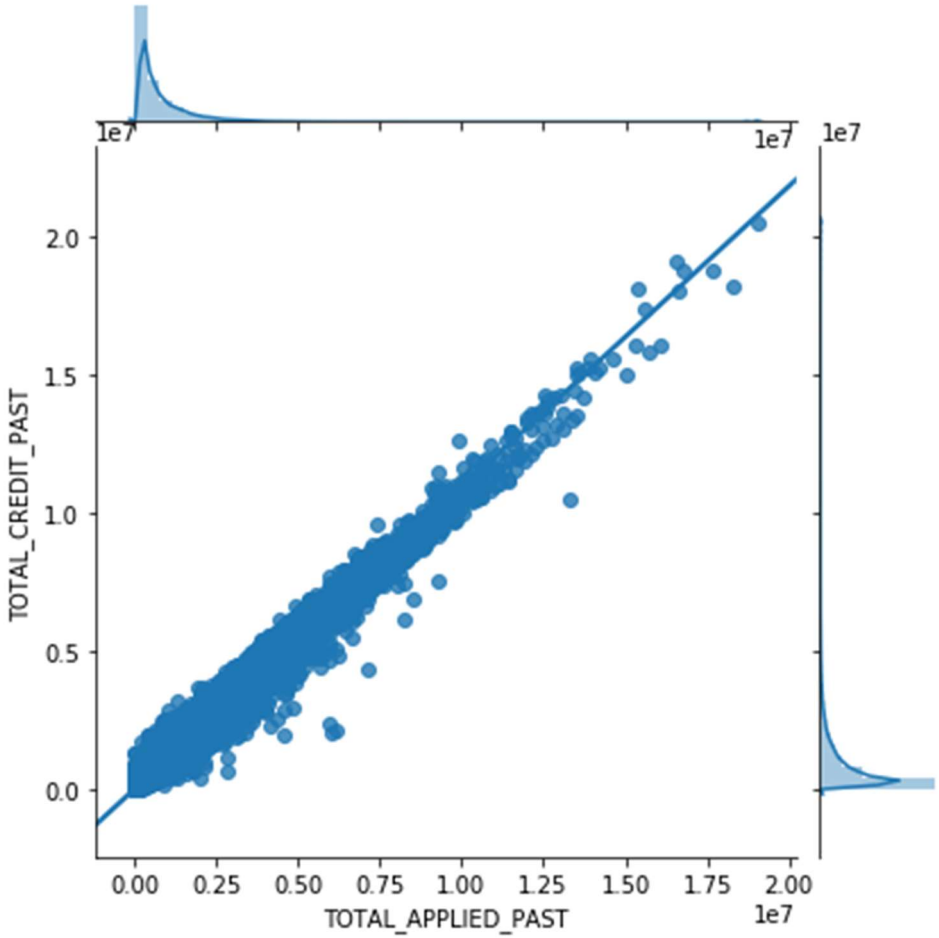


This is another way to visualize the same thing - the relationship between appicant age and the gender on the credit amount. But, here it looks like gender do have an influence with respect to the outliers. We see more of blue dots that indicate male clients. The orange dots (representing female clients are dense at lower credit levels and at higher credit levels the blue dots (male clients) are more significant. This may also be due to the imbalance in the sample data with respect to the percentage of records pertaining to male clients vs. those pertaining to female clients.

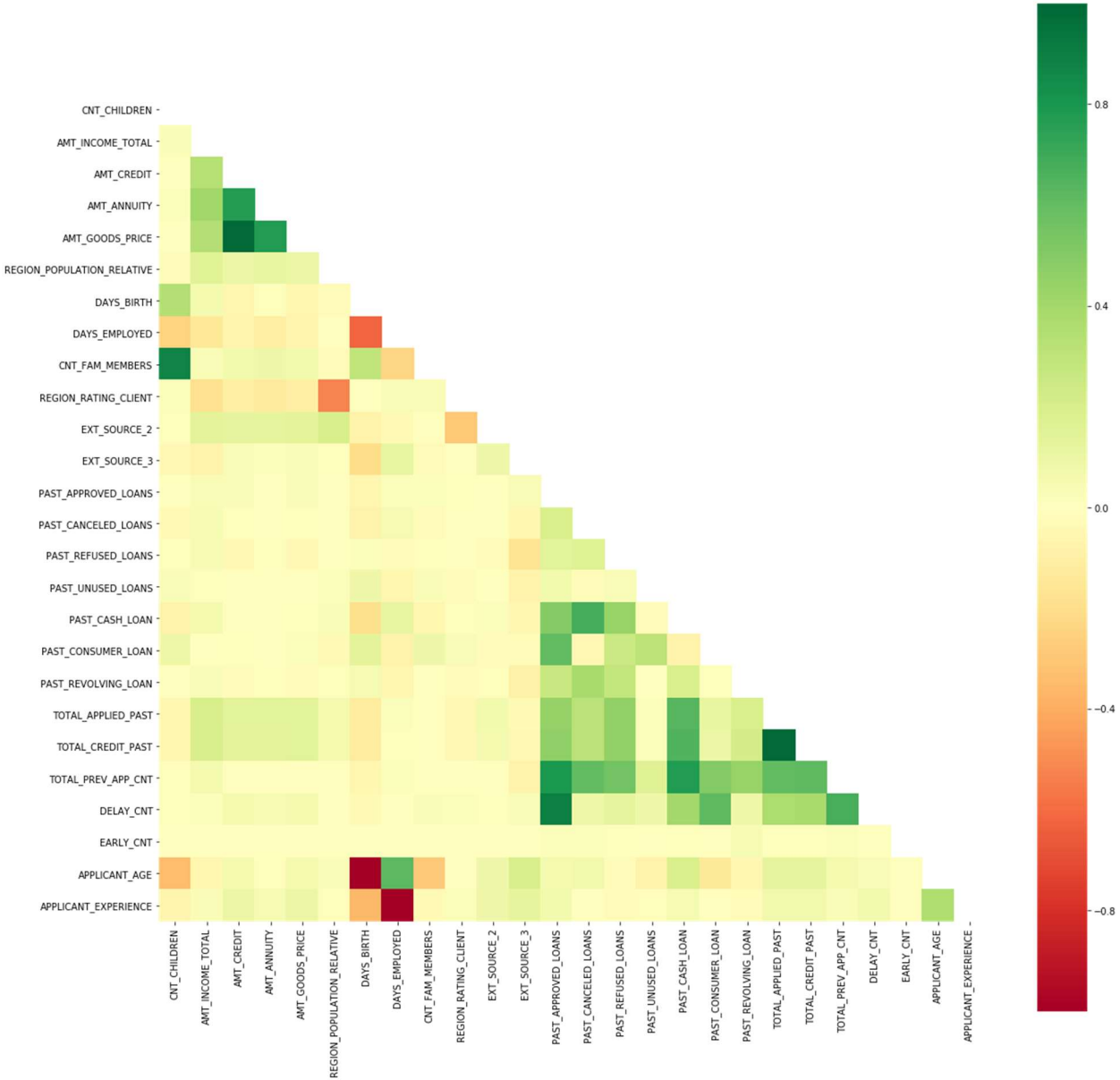
Joint Plot showing the relationship with past applied amount

The scatter plot density is high in the range of 30-60 (middle age and senior). The next plot shows a linear relationship between the amount applied in the past and the amount credited in the past. There are slight variations in the amount applied and amount credited - in most of the cases it is a lesser amount (below the straight line) and in certain cases the amount credited is more than the amount applied (above the line).





Studying the Correlation in Quantitative Variables



We do not see any positive or negative correlation except for few obvious relationships like

AMT_GOODS_PRICE & AMT_CREDIT

TOTAL_APPLIED_PAST & TOTAL_CREDIT_PAST

TOTAL_PREV_APP_CNT & PAST_APPROVED_LOANS

CNT_CHILDREN & CNT_FAM_MEMBERS

Most of these correlations are connected to the derived fields. This makes it further more difficult to identify the driver variables.

4. Driver Variables

For Target

4.1. Identifying Drivers

Categorical Variables

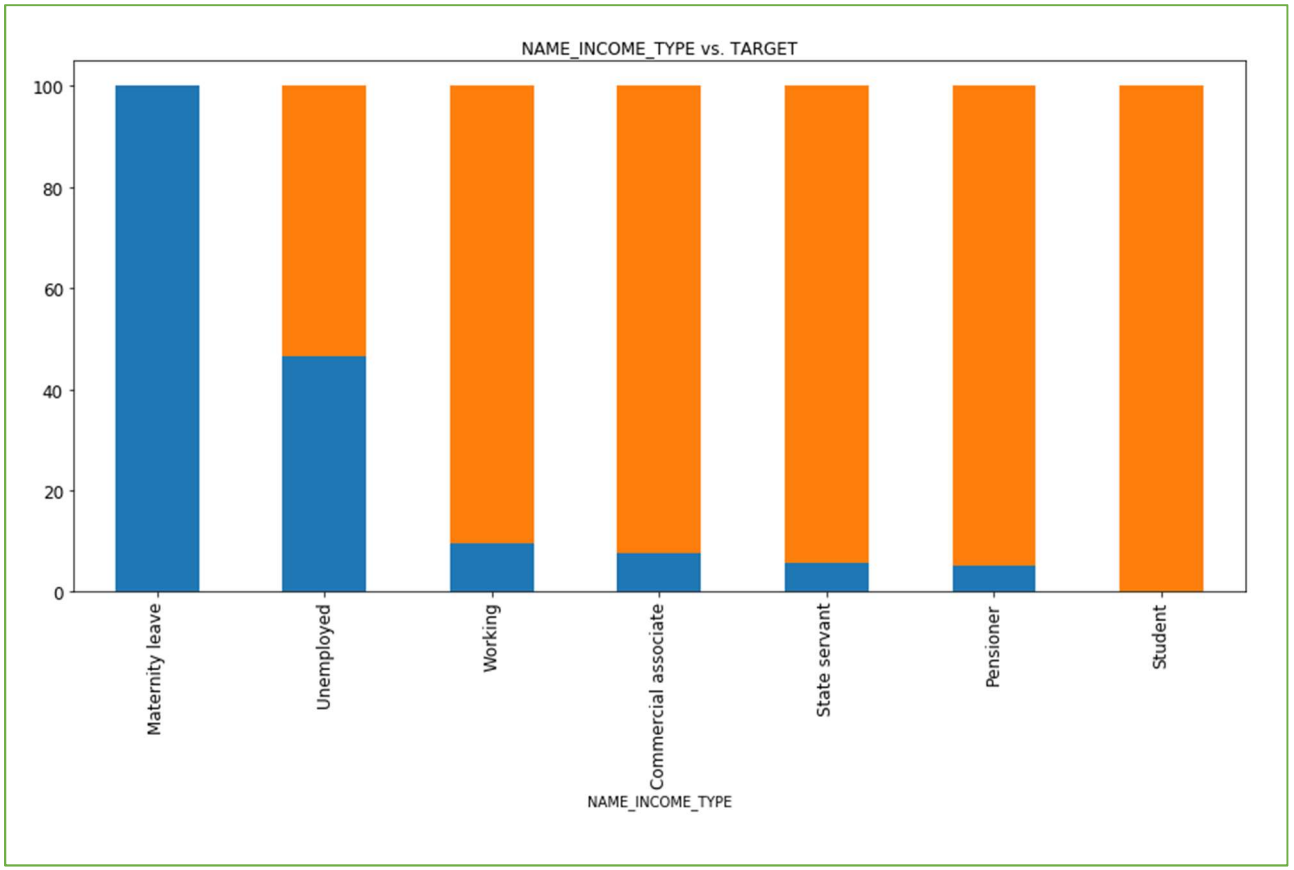
Studying Target Variable for different Categorical Variables

We studied the target variable against different categorical variables using segmented univariate analysis. For each category we plotted the percentage of records with target = 1 (client with payment difficulties) and percentage of records with target = 0 (all others) in a 100% stacked bar plot. This clearly shows in which segment the number of defaulters is high.

Based on the study we identified the following driver variables:

a. Income type is a strong indicator for unemployed or maternity leave categories.
b. Lower the Education higher are the possibilities of payment difficulties.
c. 5-10% is the rate of defaulters irrespective of family status. Not a Predictor.
d. 7-8% irrespective of whether the client owns a car / realty. Not a Predictor.
e. The Middle Age and Youth have more difficulties compared to the Senior & Elderly.
f. Payment difficulty is more probable with those who stay in rented apartment or with parents.
g. Low-skill Laborers, Drivers, Waiters / Barmen Staff, Security Staff, Laborers, Cooking Staff have more than 10% defaulters.

Income Type

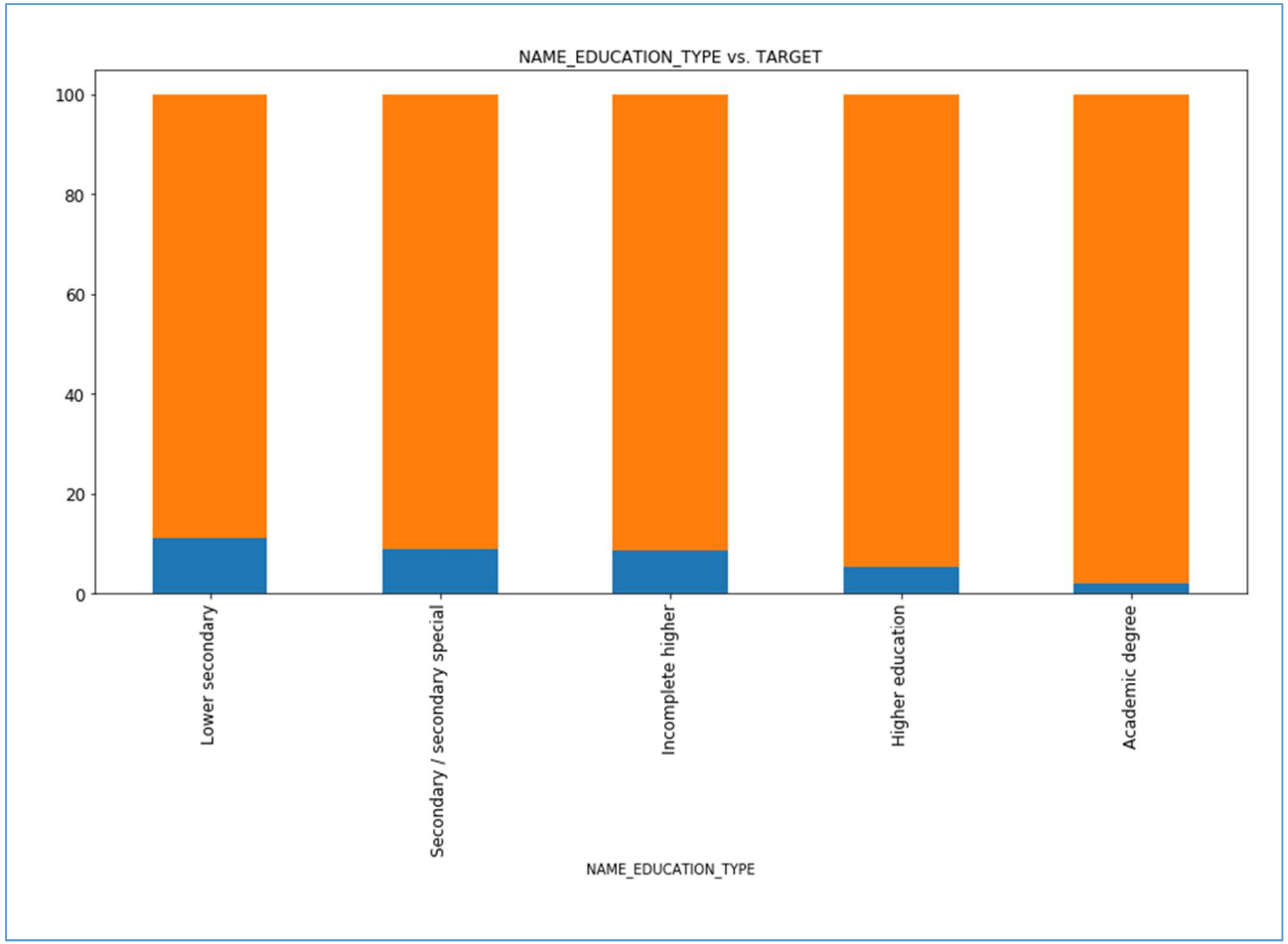


Income Type	CWPD%	OTHERS%
NAME_INCOME_TYPE	%OF1	%OF0
Commercial associate	7.64	92.36
Maternity leave	100	0
Pensioner	5.25	94.75
State servant	5.74	94.26
Student	0	100
Unemployed	46.67	53.33
Working	9.54	90.46

INCOME TYPE = UNEMPLOYED OR MATERNITY LEAVE

is a strong indicator of client with payment difficulties.

Education Type

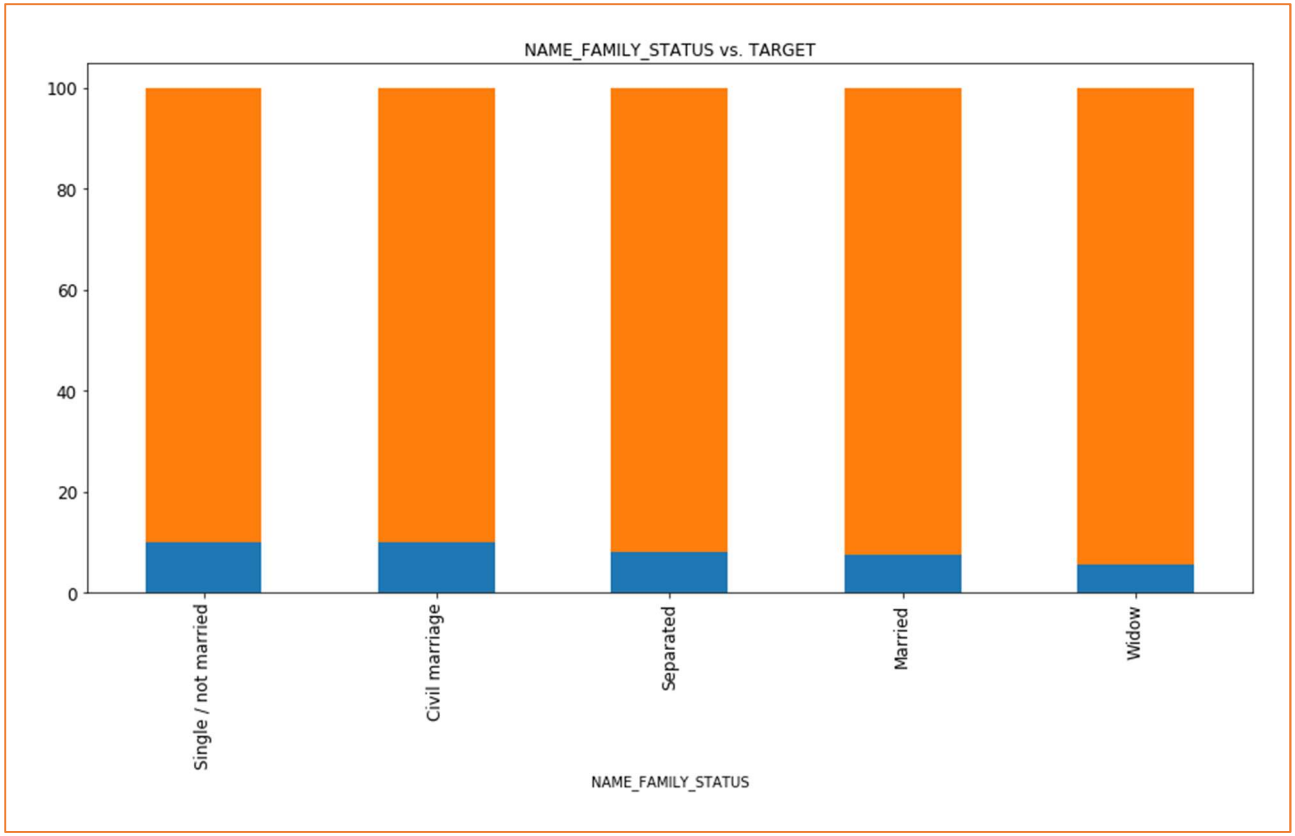


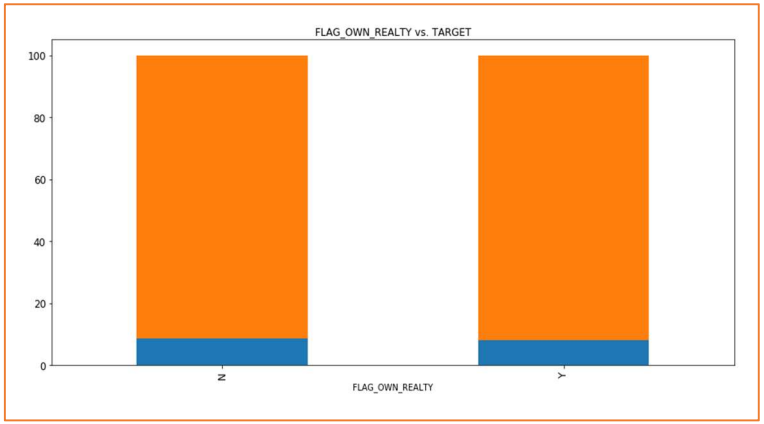
Income Type	CWPD%	OTHERS%
NAME_EDUCATION_TYPE	%OF1	%OF0
Academic degree	1.99	98.01
Higher education	5.42	94.58
Incomplete higher	8.55	91.45
Lower secondary	10.98	89.02
Secondary / secondary spe	8.88	91.12

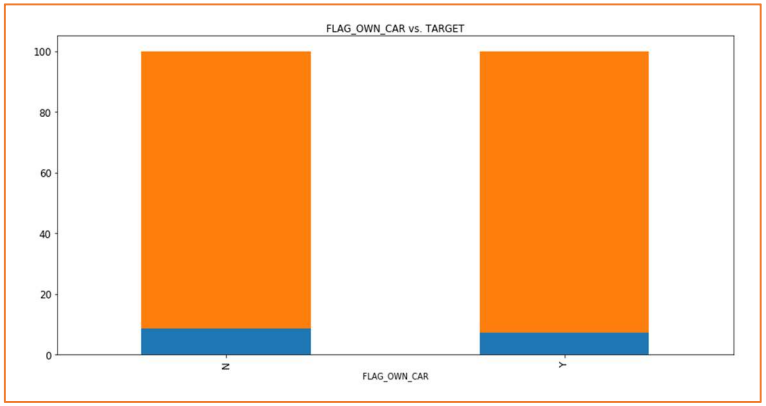
Lower the Education

Higher are the possibilities of payment difficulties.

Family Status, Owning a Car or Realty





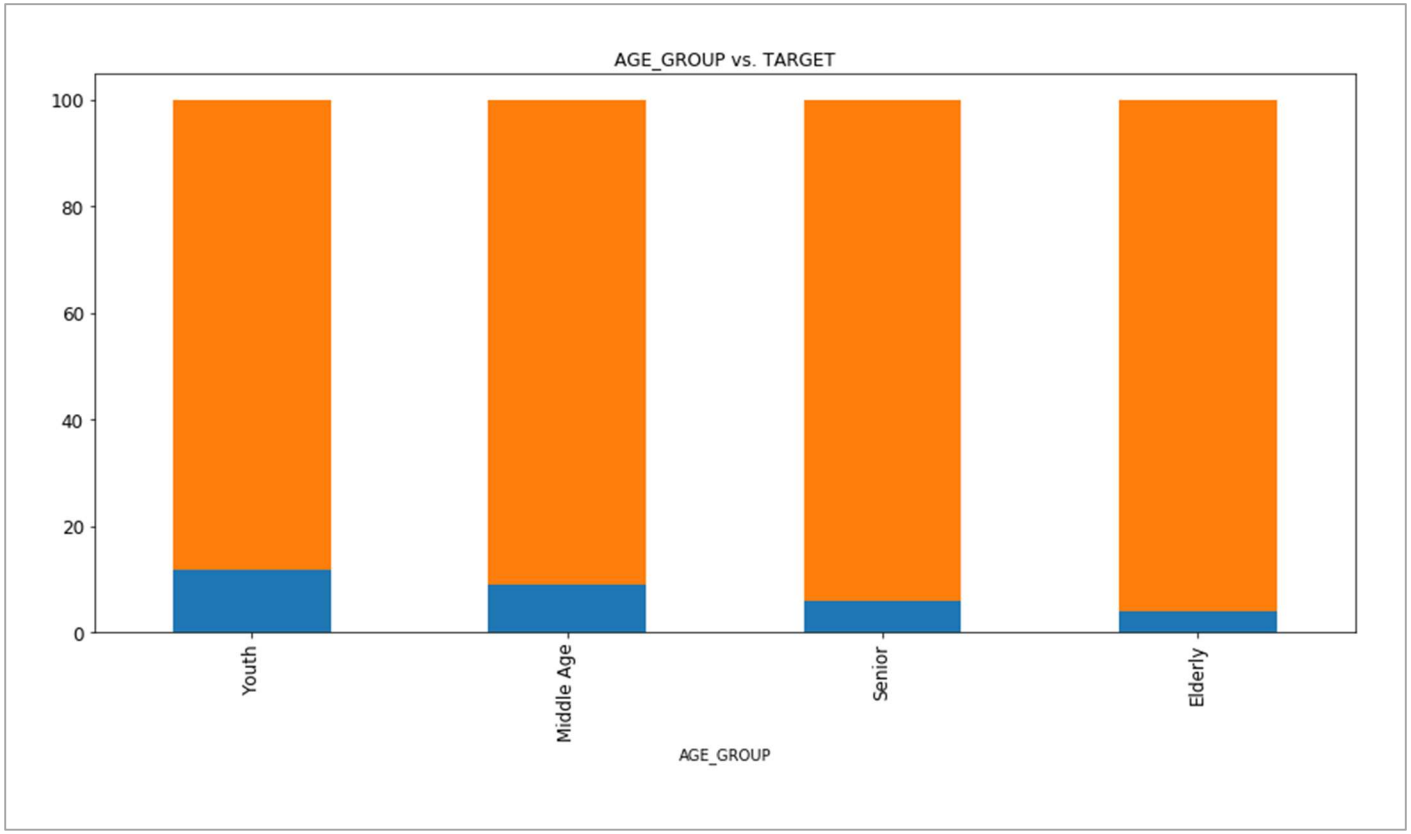


5-10% is the rate of defaulters irrespective of family status.

7-8 in case of clients owning a car / owning a realty.

These three CANNOT be PREDICTOR VARIABLES.

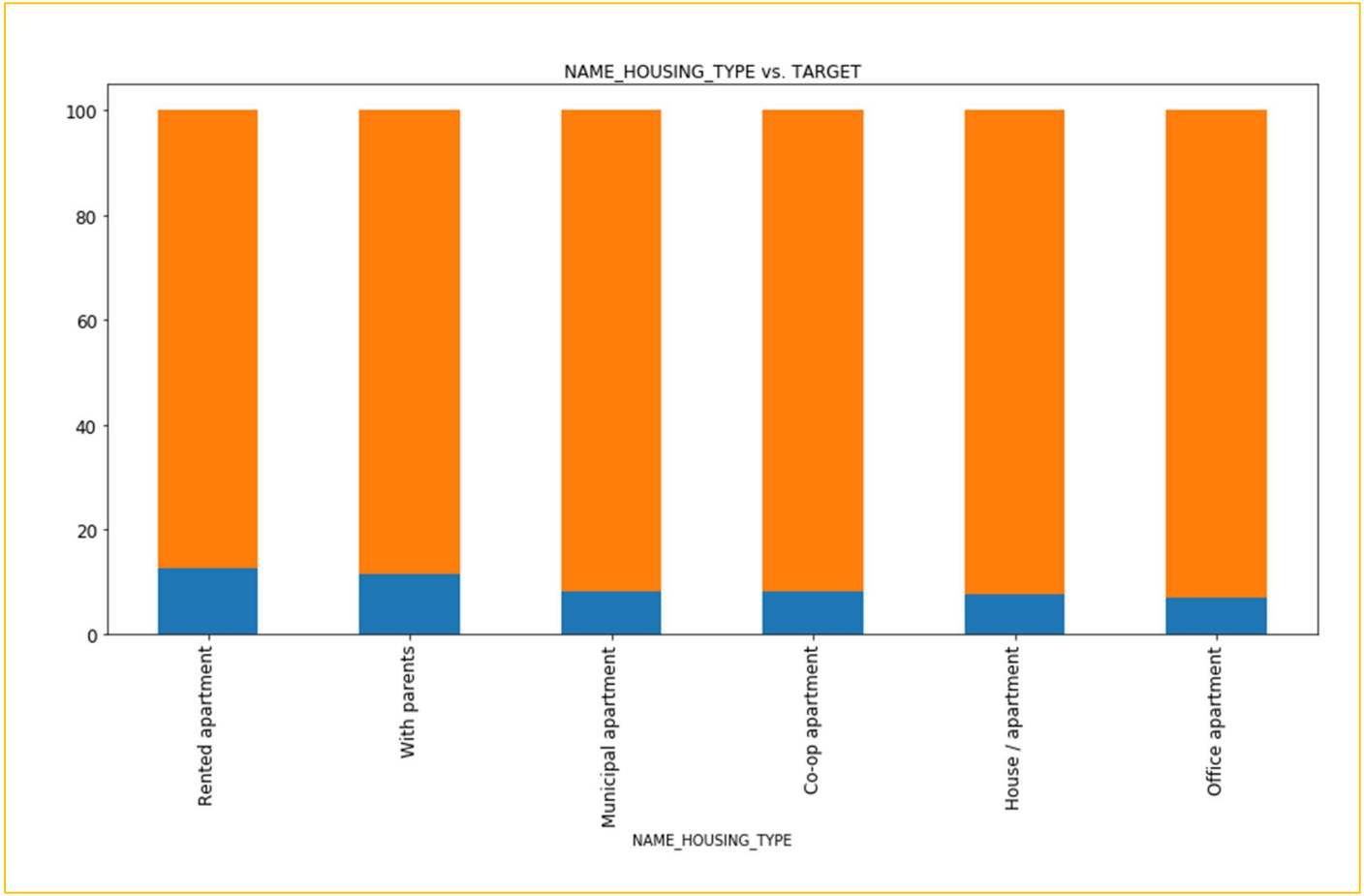
Age Group



Age Group	CWPD%	OTHERS%
AGE_GROUP	%OF1	%OF0
Elderly	4.04	95.96
Middle Age	8.75	91.25
Senior	5.82	94.18
Youth	11.68	88.32

The Middle Age and Youth have more defaulters compared to the Senior & Elderly.

Housing Type



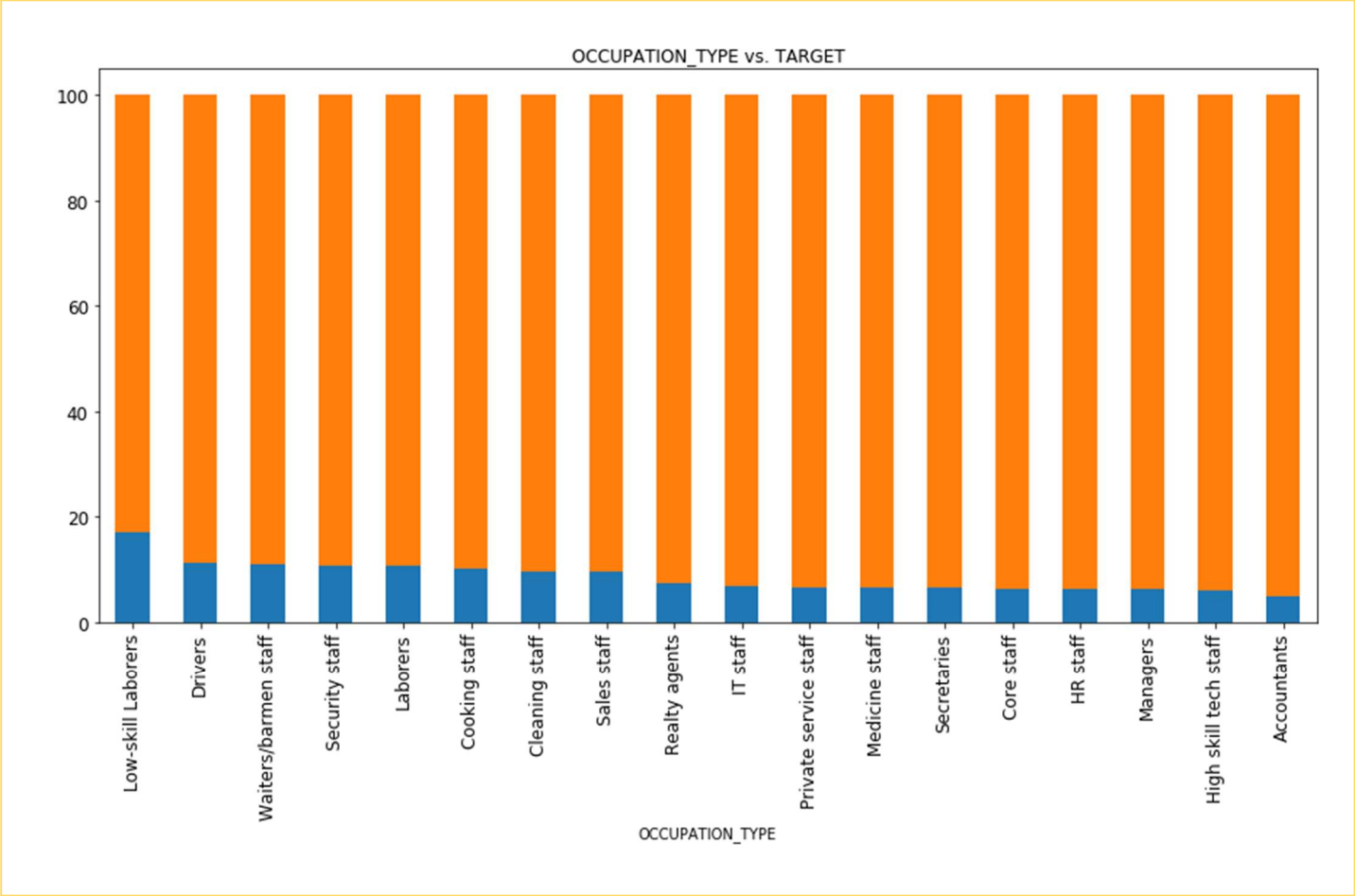
Housing Type	CWPD%	OTHERS%
NAME_HOUSING_TYPE	%OF1	%OF0
Rented apartment	12.58	87.42
With parents	11.68	88.32
Municipal apartment	8.38	91.62
Co-op apartment	8.14	91.86
House / apartment	7.8	92.2
Office apartment	6.75	93.25

There are high number of clients with payment difficulties

who stay in RENTED APARTMENT or WITH PARENTS.

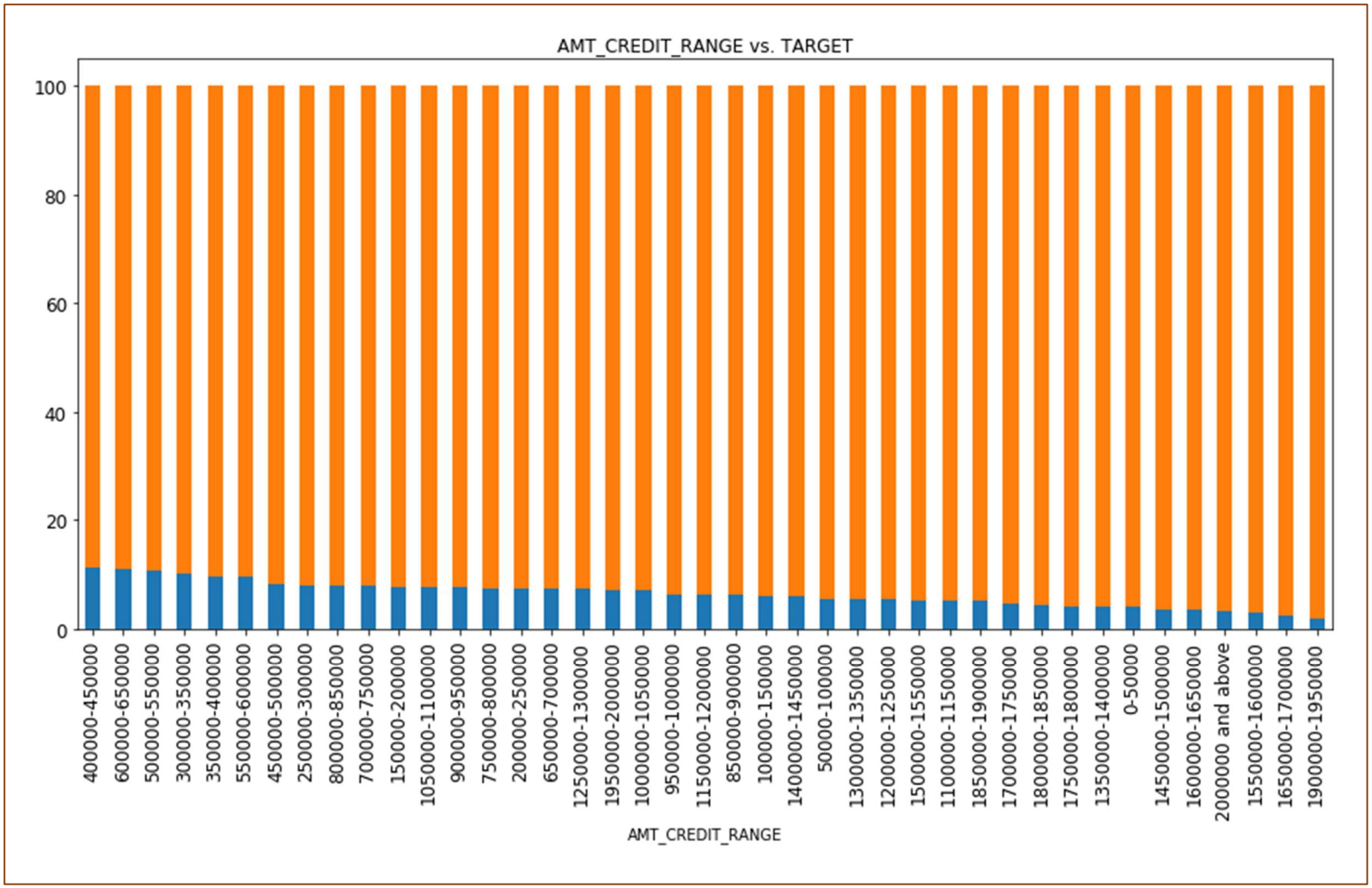
Occupation Type

Low-skill Laborers, Drivers, Waiters / Barmen Staff, Security Staff, Laborers, Cooking Staff have more than 10% defaulters.



Income Type	CWPD%	OTHERS%
NAME_OCCUPATION_TYPE	%OF1	%OF0
Accountants	4.85	95.15
Cleaning staff	9.7	90.3
Cooking staff	10.27	89.73
Core staff	6.32	93.68
Drivers	11.38	88.62
HR staff	6.31	93.69
High skill tech staff	6.09	93.91
IT staff	6.71	93.29
Laborers	10.57	89.43
Low-skill Laborers	17.14	82.86
Managers	6.19	93.81
Medicine staff	6.55	93.45
Private service staff	6.65	93.35
Realty agents	7.3	92.7
Sales staff	9.67	90.33
Secretaries	6.54	93.46
Security staff	10.63	89.37
Waiters/barmen staff	11.01	88.99

Credit Amount



Credit Amount	CWPD%	OTHERS%
AMT_CREDIT_RANGE	%OF1	%OF0
400000-450000	11.32	88.68
600000-650000	10.96	89.04
500000-550000	10.62	89.38
300000-350000	10.13	89.87
350000-400000	9.75	90.25
550000-600000	9.74	90.26
450000-500000	8.31	91.69
250000-300000	8.07	91.93
800000-850000	7.99	92.01
700000-750000	7.86	92.14
150000-200000	7.79	92.21
1050000-1100000	7.65	92.35
900000-950000	7.6	92.4
750000-800000	7.55	92.45
200000-250000	7.38	92.62
650000-700000	7.37	92.63
1250000-1300000	7.29	92.71
1950000-2000000	7.24	92.76
1000000-1050000	7.21	92.79
950000-1000000	6.35	93.65

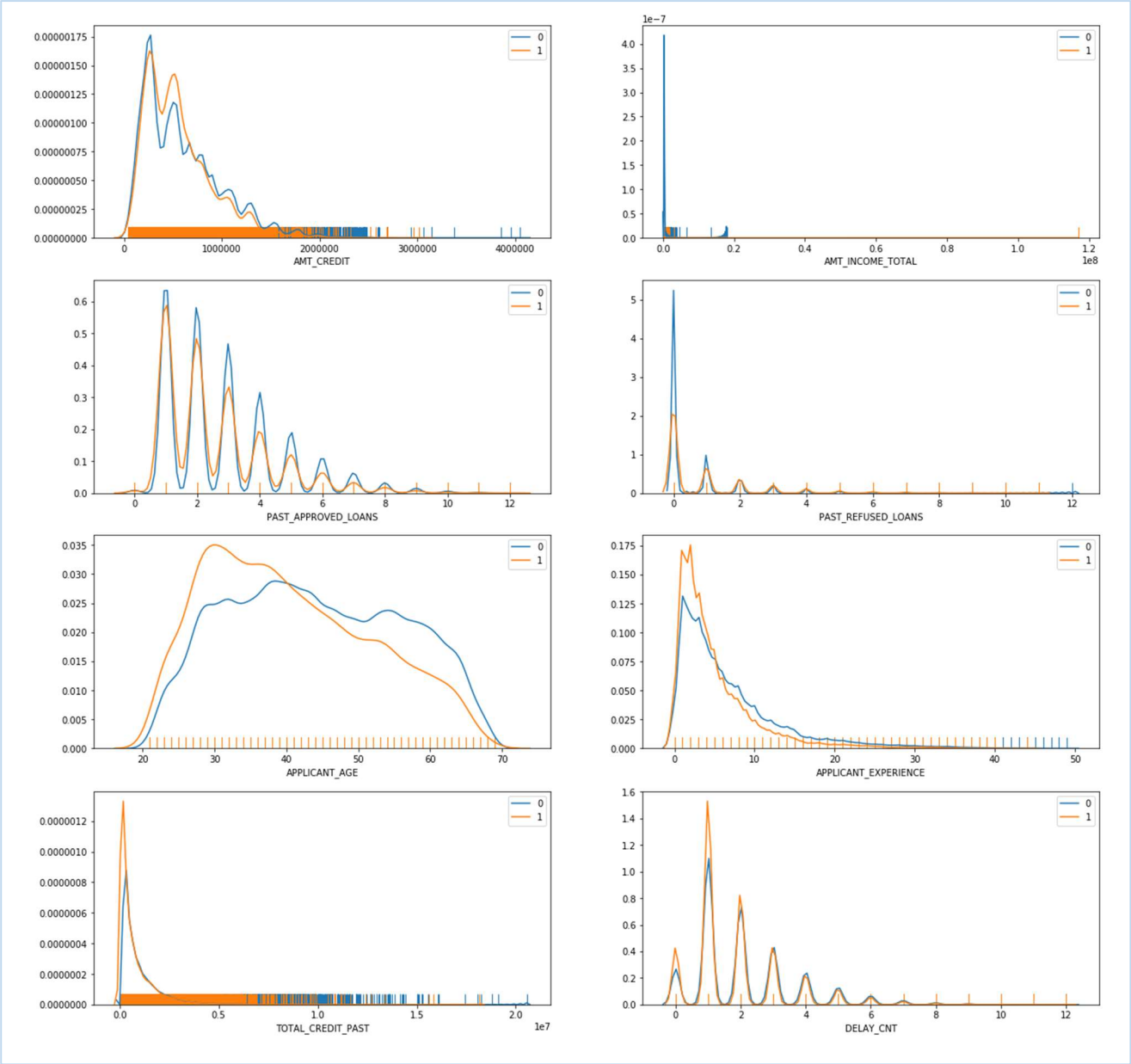
Credit Amount	CWPD%	OTHERS%
AMT_CREDIT_RANGE	%OF1	%OF0
1150000-1200000	6.28	93.72
850000-900000	6.24	93.76
100000-150000	6.06	93.94
1400000-1450000	5.95	94.05
50000-100000	5.62	94.38
1300000-1350000	5.47	94.53
1200000-1250000	5.36	94.64
1500000-1550000	5.28	94.72
1100000-1150000	5.21	94.79
1850000-1900000	5.08	94.92
1700000-1750000	4.67	95.33
1800000-1850000	4.38	95.62
1750000-1800000	4.23	95.77
1350000-1400000	4.19	95.81
0-50000	4.16	95.84
1450000-1500000	3.65	96.35
1600000-1650000	3.48	96.52
2000000 and abo	3.17	96.83
1550000-1600000	3.11	96.89
1650000-1700000	2.45	97.55
1900000-1950000	2	98

We see more clients with payment difficulties in specific ranges.

4.2. Identifying Drivers

Quantitative Variables

Studying Target Variable for different Quantitative Variables



Credit Amount

Income of the client

Past approved loans

Past Refused Loans

Client's Age

Client's Experience

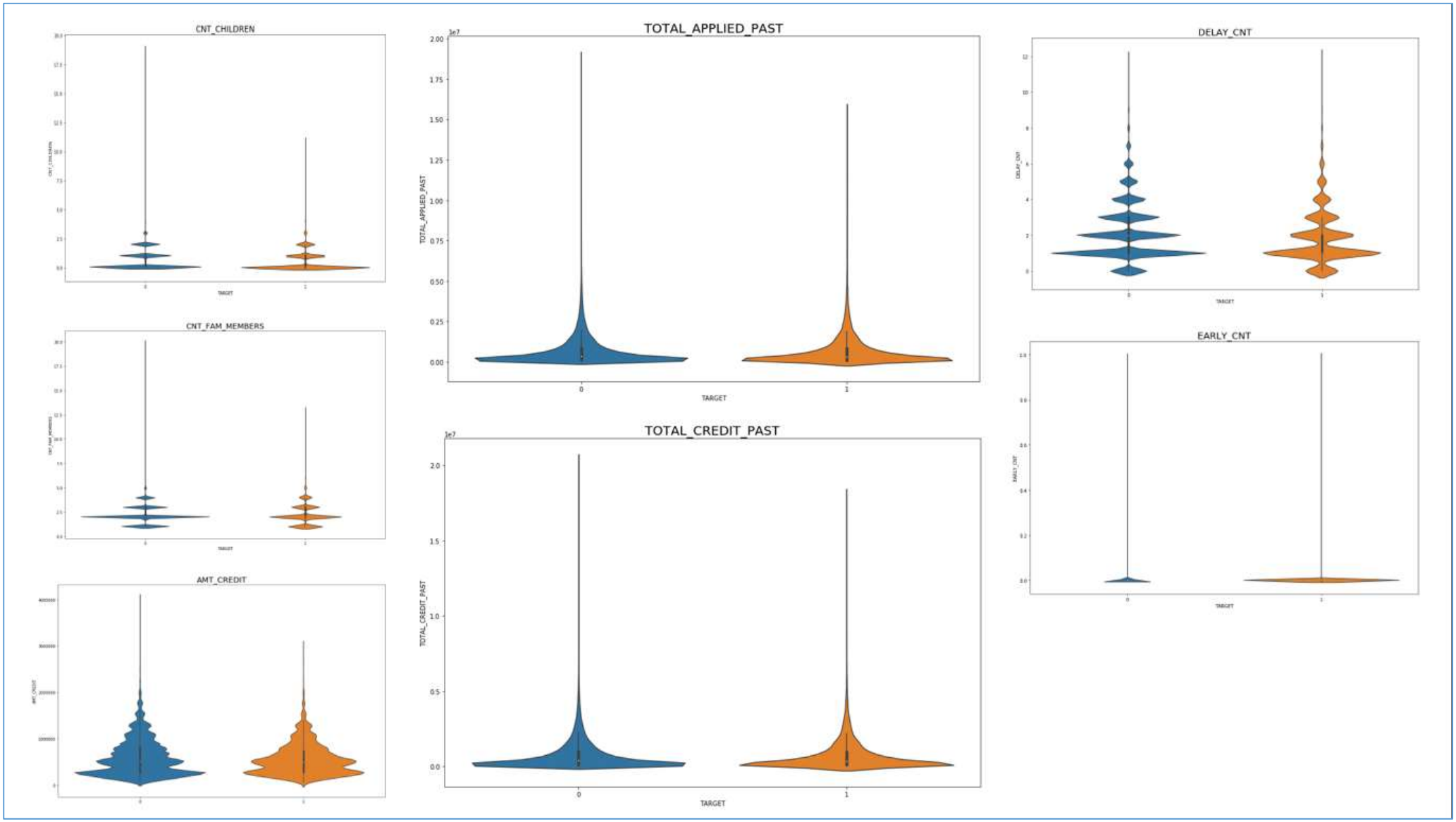
Credits in the Past

Delays in the Past

These parameters are individually plotted here.

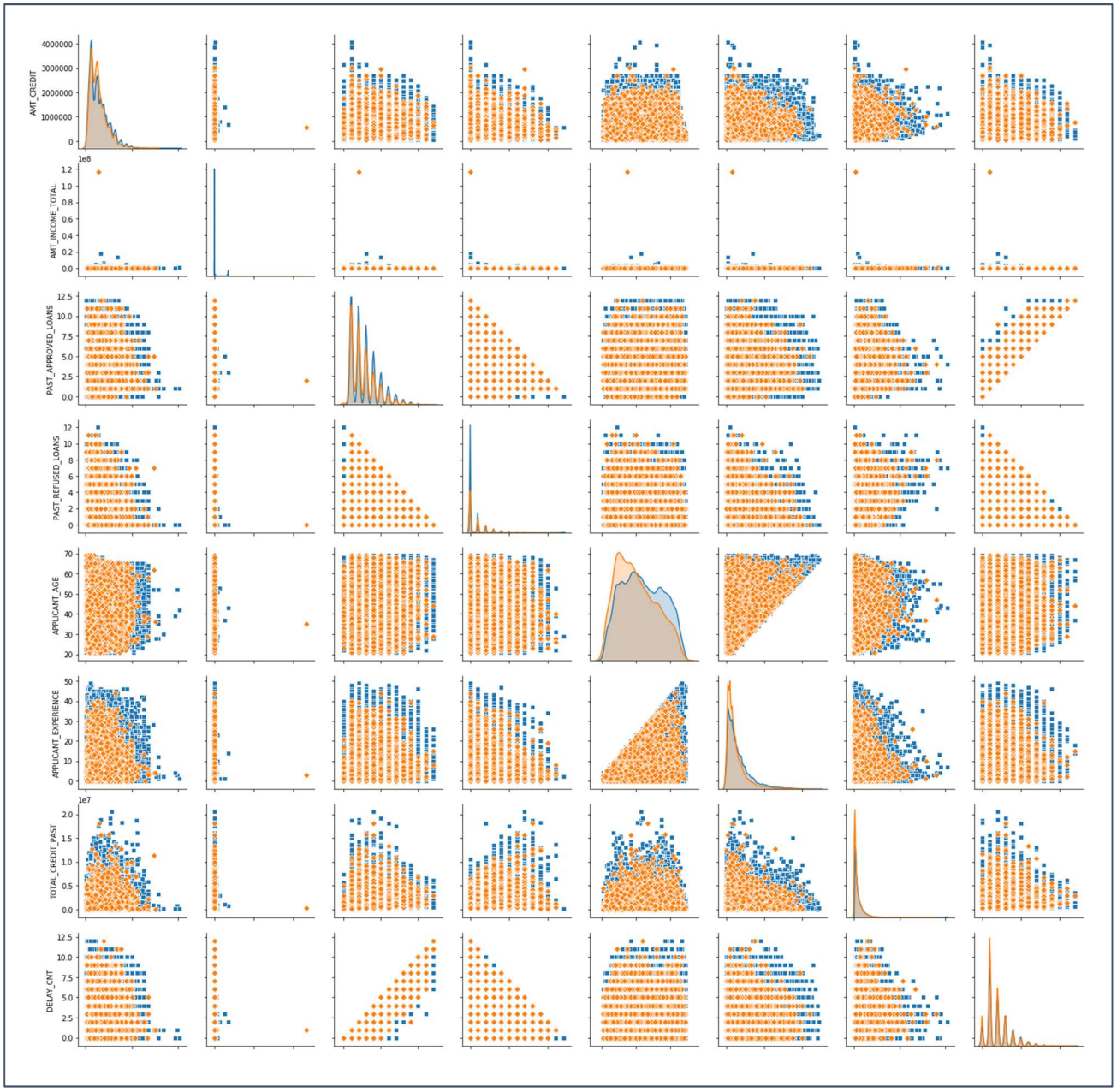
In the next plot we will see how they are correlated with each other.

Plotting the Quantitative Variables - Probability Density with Violin Plot



4.3. Correlation between Driver Variables

Correlation between the Driver Variables

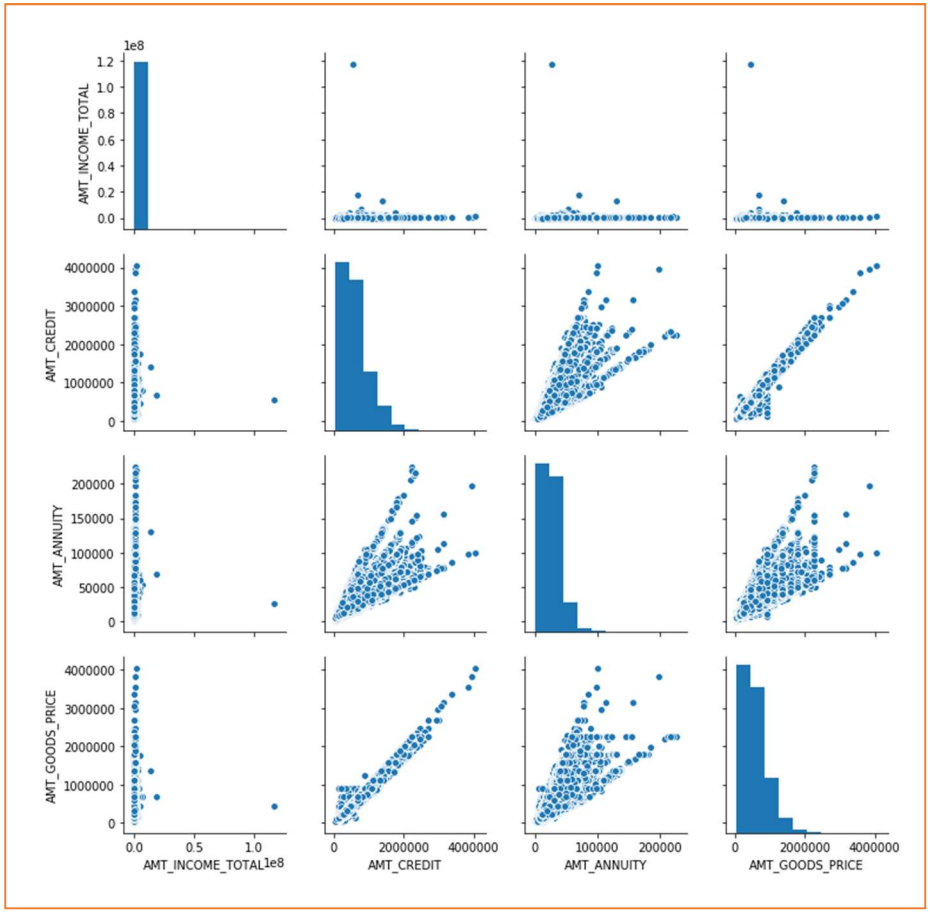


The driver variables considered are:

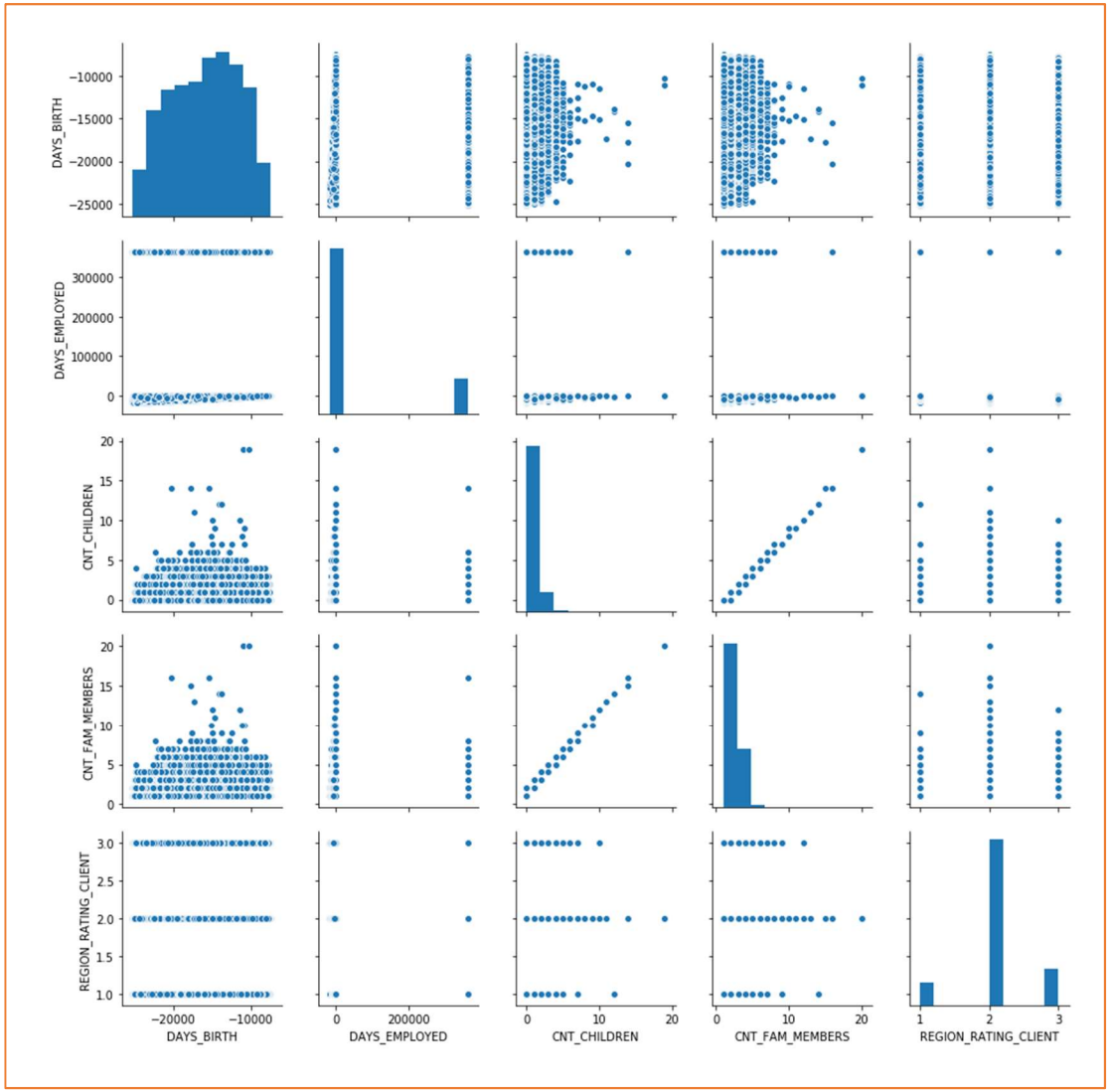
- Credit Amount
- Income of the client
- Past approved loans
- Past Refused Loans
- Client's Age
- Client's Experience
- Credits in the Past
- Delays in the Past

The orange marks represents the clients with payment difficulties. From the density we can see that these variables give a strong indication on the clients with payment difficulties.

Studying the Correlation with Pair Plot for Quantitative Variables



This plot depicts the correlation between the various amount fields in the data set. The price of the goods, annuity and credit amount show a linear relationship with each other because the loan applied and sanctioned depends on the price of the goods and so is the annuity.



Personal attributes of the client (number of children, family members, region, age and experience) are depicted against each other.

CONCLUSION

Based on the study we identified the following driver variables:

a. Income type is a strong indicator for unemployed or maternity leave categories.
b. Lower the Education higher are the possibilities of payment difficulties.
c. 5-10% is the rate of defaulters irrespective of family status. Not a Predictor.
d. 7-8% irrespective of whether the client owns a car / realty. Not a Predictor.
e. The Middle Age and Youth have more difficulties compared to the Senior & Elderly.
f. Payment difficulty is more probable with those who stay in rented apartment or with parents.
g. Low-skill Laborers, Drivers, Waiters / Barmen Staff, Security Staff, Laborers, Cooking Staff have more than 10% defaulters.
h. Applicant age and experience has a greater kernel density distribution for defaulters.
i. Delays in the past payment is another strong driver.