CREDIT EDA ASSIGNMENT

Kunal M

ACP DS, IIITB

UNDERSTANDING AND APPROACH

UNDERSTANDING - PROBLEM STATEMENT

- Loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history.
- **Risks Associated:**
  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
  - If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.
- **Scenarios for each loan:**
  - The client with payment difficulties
  - All other cases
- **Decisions for any client's loan application:**
  - Approved
  - Cancelled
  - Refused
  - Unused offer

UNDERSTANDING - OBJECTIVES

- **Business Objectives**
  - To identify patterns which indicate if a client has difficulty paying their instalments
    - Action based on patterns such as
      - Denying the loan
      - Reducing the amount of loan
      - Lending (to risky applicants) at a higher interest rate, etc.
  - Ensure that the consumers capable of repaying the loan are not rejected
- **EDA study objectives**
  - Identification of applicants that are capable of repaying using EDA is the aim of this case study
  - Understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.
    - The company can utilise this knowledge for its portfolio and risk assessment.
    - To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

OVERALL APPROACH OF THE ANALYSIS

- Understanding Problem Statement
- Missing Value Handling
  - General Identification
  - Identification and Handling (application_data)
  - Identification and Handling (previous_application)
  - XAP, XNA & 365243 Values
- Categorization for study (Categorical and Numerical Variables Identification)
- Data Imbalance – Analysis and Implication
- Correlation Matrix (To identify variables for bivariant and multivariant analysis)
  - Top 10 correlation for the Client with payment difficulties and other cases
- Univariant Analysis
  - Categorical Variables
  - Outliers Identification / Handling and Univariant Analysis of Numerical Variables

OVERALL APPROACH OF THE ANALYSIS (CONT.)

- Bivariant and Multivariant Analysis
  - Categorical vs Categorical Variables
  - Numerical vs Numerical Variables
  - Combination of Categorical and Numerical Variables
- Understanding the relation between two datasets and merging both
- Univariant Analysis (For Merged Dataset)
  - Categorical Variables
  - Outliers Identification / Handling and Univariant Analysis of Numerical Variables
- Correlation Matrix (For Merged Dataset)
- Bivariant and Multivariant Analysis (For Merged Dataset)
  - Categorical vs Categorical Variables
  - Numerical vs Numerical Variables
  - Combination of Categorical and Numerical Variables
- Conclusion - Portfolio and Risk Assessment

MISSING VALUES HANDLING

MISSING VALUES - IDENTIFICATION

Column Name	%	Column Name	%	Column Name	%	Column Name	%
COMMONAREA_MEDI	69.87	LANDAREA_AVG	59.38	ENTRANCES_MODE	50.35	AMT_REQ_CREDIT_BUREAU_WEEK	13.50
COMMONAREA_AVG	69.87	LANDAREA_MEDI	59.38	ENTRANCES_AVG	50.35	AMT_REQ_CREDIT_BUREAU_DAY	13.50
COMMONAREA_MODE	69.87	LANDAREA_MODE	59.38	ENTRANCES_MEDI	50.35	AMT_REQ_CREDIT_BUREAU_MON	13.50
NONLIVINGAPARTMENTS_MEDI	69.43	BASEMENTAREA_MEDI	58.52	LIVINGAREA_MEDI	50.19	AMT_REQ_CREDIT_BUREAU_QRT	13.50
NONLIVINGAPARTMENTS_MODE	69.43	BASEMENTAREA_AVG	58.52	LIVINGAREA_MODE	50.19	AMT_REQ_CREDIT_BUREAU_HOUR	13.50
NONLIVINGAPARTMENTS_AVG	69.43	BASEMENTAREA_MODE	58.52	LIVINGAREA_AVG	50.19	AMT_REQ_CREDIT_BUREAU_YEAR	13.50
FONDKAPREMONT_MODE	68.39	EXT_SOURCE_1	56.38	HOUSETYPE_MODE	50.18	NAME_TYPE_SUITE	0.42
LIVINGAPARTMENTS_MODE	68.35	NONLIVINGAREA_MEDI	55.18	FLOORSMAX_MEDI	49.76	DEF_30_CNT_SOCIAL_CIRCLE	0.33
LIVINGAPARTMENTS_MEDI	68.35	NONLIVINGAREA_MODE	55.18	FLOORSMAX_AVG	49.76	OBS_60_CNT_SOCIAL_CIRCLE	0.33
LIVINGAPARTMENTS_AVG	68.35	NONLIVINGAREA_AVG	55.18	FLOORSMAX_MODE	49.76	DEF_60_CNT_SOCIAL_CIRCLE	0.33
FLOORSMIN_MODE	67.85	ELEVATORS_MEDI	53.30	YEARS_BEGINEXPLUATATION_AVG	48.78	OBS_30_CNT_SOCIAL_CIRCLE	0.33
FLOORSMIN_MEDI	67.85	ELEVATORS_MODE	53.30	YEARS_BEGINEXPLUATATION_MEDI	48.78	EXT_SOURCE_2	0.21
FLOORSMIN_AVG	67.85	ELEVATORS_AVG	53.30	YEARS_BEGINEXPLUATATION_MODE	48.78	AMT_GOODS_PRICE	0.09
YEARS_BUILD_MODE	66.50	WALLSMATERIAL_MODE	50.84	TOTALAREA_MODE	48.27	AMT_ANNUITY	0.0039
YEARS_BUILD_MEDI	66.50	APARTMENTS_MODE	50.75	EMERGENCYSTATE_MODE	47.40	CNT_FAM_MEMBERS	0.0007
YEARS_BUILD_AVG	66.50	APARTMENTS_MEDI	50.75	OCCUPATION_TYPE	31.35	DAYS_LAST_PHONE_CHANGE	0.0003
OWN_CAR_AGE	65.99	APARTMENTS_AVG	50.75	EXT_SOURCE_3	19.83		

High Missing Values | Mid Missing Values | Low Missing Values

MISSING VALUES - HANDLING

Columns	%age Missing	Insight / Reason	Action Taken
Columns with High Missing Values (>30%), other than OWN_CAR_AGE, EXT_SOURCE_1 and OCCUPATION_TYPE	Between 47.39% to 69.87%	Missing values are to high to be imputed, only columns with causation (OWN_CAR_AGE, EXT_SOURCE_1, OCCUPATION_TYPE) are kept for further analysis before taking action.	Columns not being used for study.
OWN_CAR_AGE	65.99	• Most of missing values for OWN_CAR_AGE are for those, where client/customer doesn't have any car. All these records are valid. • Imputed 5 records where user own a car, and no value for car age, with median value.	**Imputed 5 records with median value, rest kept NaN only.**
NAME_TYPE_SUITE	0.42	Distribution shows that NAME_TYPE_SUITE has 'Unaccompanied' as value for most of its rows (more than 80%).	**Imputated with mode value for missing values.**
AMT_GOODS_PRICE	0.09	• All records where information is missing, are case of Revolving loans. • Mean/Median of AMT_GOODS_PRICE in all is much higher than Mean/Median for Revolving loans.	**Imputated with median value, where NAME_CONTRACT_TYPE is 'Revolving loans'.**
AMT_ANNUITY	0.003902	Only 12 records (< 0.004%) where AMT_ANNUITY is missing.	**Imputed with median value.**
CNT_FAM_MEMBERS	0.00065	Only 2 records (< 0.001%) where CNT_FAM_MEMBERS is missing.	**Imputed with median value.**

MISSING VALUES – HANDLING (CONT.)

Columns	%age Missing	Insight / Reason	Action Taken
DAYS_LAST_PHONE_CHANGE	0.000325	Only 1 records (< 0.001%) where DAYS_LAST_PHONE_CHANGE is missing.	**Imputed with median value.**
EXT_SOURCE_1 EXT_SOURCE_2 EXT_SOURCE_3	Between 0.21% to 56.38%	All 3 stores, normalized score from external data source. If any one is present, it will provide sufficient insight.	**Imputed 172 records with median, where all three are null.**
OCCUPATION_TYPE	31.34	This is important feature and should need more analysis. Please refer table for 'Missing values in OCCUPATION_TYPE'.	**Added three more categories for OCCUPATION_TYPE**
AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY AMT_REQ_CREDIT_BUREAU_WEEK AMT_REQ_CREDIT_BUREAU_MON AMT_REQ_CREDIT_BUREAU_QRT AMT_REQ_CREDIT_BUREAU_YEAR	13.5	Number of enquiries to Credit Bureau about the client before application, within particular duration. These missing values are not expected to harm our analysis. Also, Imputating these values with mean/median is not recommanded. So, keeping them as it is.	No Action
OBS_30_CNT_SOCIAL_CIRCLE DEF_30_CNT_SOCIAL_CIRCLE OBS_60_CNT_SOCIAL_CIRCLE DEF_60_CNT_SOCIAL_CIRCLE	0.33	observation of client's social surroundings with observable for particular DPD (days past due) default. These missing values are not expected to harm our analysis. Also, Imputating these values with mean/median is not recommanded. So, keeping them as it is.	No Action

MISSING VALUES – HANDLING (OCCUPATION_TYPE)

- **Insight:** Occupation type of client seams good attribute for banking system. It is important to study this variable, even though it have high missing values. For this purpose, it is proposed to estimate the occupation type based on Organisation type, to reduce missing values.

Insight	Count	%age	Added new Category
NAME_INCOME_TYPE == 'Pensioner'	55325	17.99	Pensioner
ORGANIZATION_TYPE startswith 'Business Entity'	17893	5.82	Business Entity
ORGANIZATION_TYPE startswith 'Industry'	2316	0.75	XNA
Others	20797	6.76	XNA
Total Records	**96331**	**31.33**	

MISSING HANDLING – PREVIOUS_APPLICATION

Columns	Missing %	Insight / Reason	Action Taken
RATE_INTEREST_PRIMARY	99.64	Very high missing value.	Column not used
RATE_INTEREST_PRIVILEGED	99.64	Very high missing value.	Column not used
AMT_DOWN_PAYMENT	53.64	Very high missing value.	Column not used
RATE_DOWN_PAYMENT	53.64	Very high missing value.	Column not used
NAME_TYPE_SUITE	49.12	Very high missing value.	Column not used
DAYS_FIRST_DRAWING | DAYS_FIRST_DUE DAYS_LAST_DUE_1ST_VERSION DAYS_LAST_DUE | DAYS_TERMINATION	40.30	Although there are large missing values. But when further examined 'Approved' applications, the missing value is only 3.82%	No action
NFLAG_INSURED_ON_APPROVAL	40.30	Very high missing value.	Column not used
AMT_GOODS_PRICE	23.08	Although there are large missing values. But when examined agained 'Approved' applications with NAME_CONTRACT_TYPE other than 'Revolving loans', there is no missing value.	No action required
AMT_ANNUITY	22.29	Although there are some missing values. But when examined agained 'Approved' applications, the missing value is only 0.0008%. Imputating these missing values with median value.	**Imputation of median value**
CNT_PAYMENT	22.29	Although there are some missing values. But when examined agained 'Approved' applications, missing value is only 0.0003%.	**Imputation of median value for 4 records**
PRODUCT_COMBINATION	0.02	346 Rows with missing value.	**Imputation of mode value.**
AMT_CREDIT	0.00	1 record with missing value.	**Imputation of median.**

XAP, XNA & 365243 VALUES

- In addition to missing values in columns, both datasets includes values that represent null or missing value. These are:
  - **XAP represents 'Not Applicable'**
    - Present in 'CODE_REJECT_REASON' and 'NAME_CASH_LOAN_PURPOSE' of previous_application.
    - A valid CODE_REJECT_REASON is available for NAME_CONTRACT_STATUS as 'Refused' and 'Unused offer', for all other status it is XAP.
    - A valid NAME_CASH_LOAN_PURPOSE is available for NAME_CONTRACT_TYPE as 'Cash loans', for all other status it is XAP, which make sence as 'Cash Loan Purpose' should only be valid for 'Cash Loans' only.
  - **XNA represents 'Not Available'**
    - Present in 'ORGANIZATION_TYPE' and 'CODE_GENDER' for application_data and in 'NAME_PRODUCT_TYPE', 'NAME_GOODS_CATEGORY', 'NAME_SELLER_INDUSTRY', 'NAME_CASH_LOAN_PURPOSE', 'NAME_PAYMENT_TYPE', 'NAME_YIELD_GROUP', 'NAME_PORTFOLIO', 'CODE_REJECT_REASON', 'NAME_CLIENT_TYPE', 'NAME_CONTRACT_TYPE' for previous_application.
  - **365243 represents 'Distant high value to indicate missing'**
    - Present in 'DAYS_EMPLOYED' for application_data and in 'DAYS_FIRST_DRAWING', 'DAYS_TERMINATION', 'DAYS_LAST_DUE', 'DAYS_LAST_DUE_1ST_VERSION', 'DAYS_FIRST_DUE' for previous_application.

- **Observation:**
  - All there values, **'XAP', 'XNA' and 365243 represent Missing Not At Random (MNAR)**.
  - All the columns containing 'XAP', 'XNA' and 365243 doesn't have null value, so these values are used to represent the absent of value, and may be due to not available / not applicable / not required or error (human or system).
  - All these values are replaced by null / NaN. Reason for handling these records after handling missing values is that these records are not need to filled by mean / median / mode, but any value is only meaningful where it is present.
  - XNA in CODE_GENDER doesn't make scense. Replacing it with mode value.

CATEGORISATION FOR STUDY

UNDERSTANDING DATA – CHANGE/SKIP

Column	Dtype	Type of data	Subtype	Distribution Anomaly / Relevance	Action
REGION_POPULATION_RELATIVE	float64	Numerical	Continuous	Relative Field	Skip
DAYS_BIRTH	int64	Numerical	Continuous	YEARS_BIRTH	Change
DAYS_EMPLOYED	int64	Numerical	Continuous	MONTHS_EMPLOYED	Change
FLAG_MOBIL	int64	Categorical	Nominal	All 1	Drop
FLAG_EMP_PHONE	int64	Categorical	Nominal	1 for more than 81%	Skip
FLAG_WORK_PHONE	int64	Categorical	Nominal	0 for more than 80%	Skip
FLAG_CONT_MOBILE	int64	Categorical	Nominal	1 for more than 99%	Skip
FLAG_PHONE	int64	Categorical	Nominal	0 for more than 71%	Skip
FLAG_EMAIL	int64	Categorical	Nominal	0 for more than 94%	Skip
REGION_RATING_CLIENT	int64	Numerical	Discrete		Skip
REGION_RATING_CLIENT_W_CITY	int64	Numerical	Discrete		Skip
REG_REGION_NOT_LIVE_REGION	int64	Categorical	Nominal	0 for more than 98%	Skip
REG_REGION_NOT_WORK_REGION	int64	Categorical	Nominal	0 for more than 94%	Skip
LIVE_REGION_NOT_WORK_REGION	int64	Categorical	Nominal	0 for more than 95%	Skip
REG_CITY_NOT_LIVE_CITY	int64	Categorical	Nominal	0 for more than 92%	Skip
REG_CITY_NOT_WORK_CITY	int64	Categorical	Nominal	0 for more than 76%	Skip
LIVE_CITY_NOT_WORK_CITY	int64	Categorical	Nominal	0 for more than 82%	Skip
EXT_SOURCE_1	float64	Numerical	Continuous	As each of the columns EXT_SOURCE_1, EXT_SOURCE_2 and EXT_SOURCE_3 have missing value, so a new column with mean of all 3 external source rating is created.	Change
EXT_SOURCE_2	float64	Numerical	Continuous		
EXT_SOURCE_3	float64	Numerical	Continuous		
OBS_30_CNT_SOCIAL_CIRCLE	float64	Numerical	Discrete		Skip
DEF_30_CNT_SOCIAL_CIRCLE	float64	Numerical	Discrete		Skip
OBS_60_CNT_SOCIAL_CIRCLE	float64	Numerical	Discrete		Skip
DEF_60_CNT_SOCIAL_CIRCLE	float64	Numerical	Discrete		Skip
FLAG_DOCUMENT_2	int64	Categorical	Nominal	0 for more than 99%	Skip

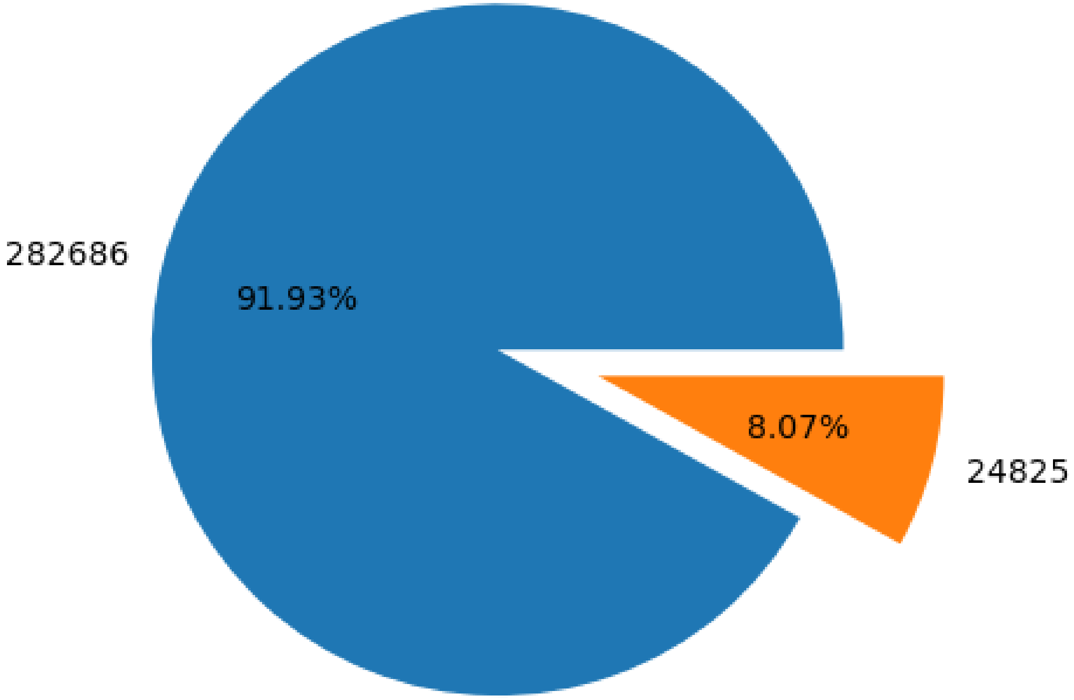
UNDERSTANDING DATA – CHANGE/SKIP

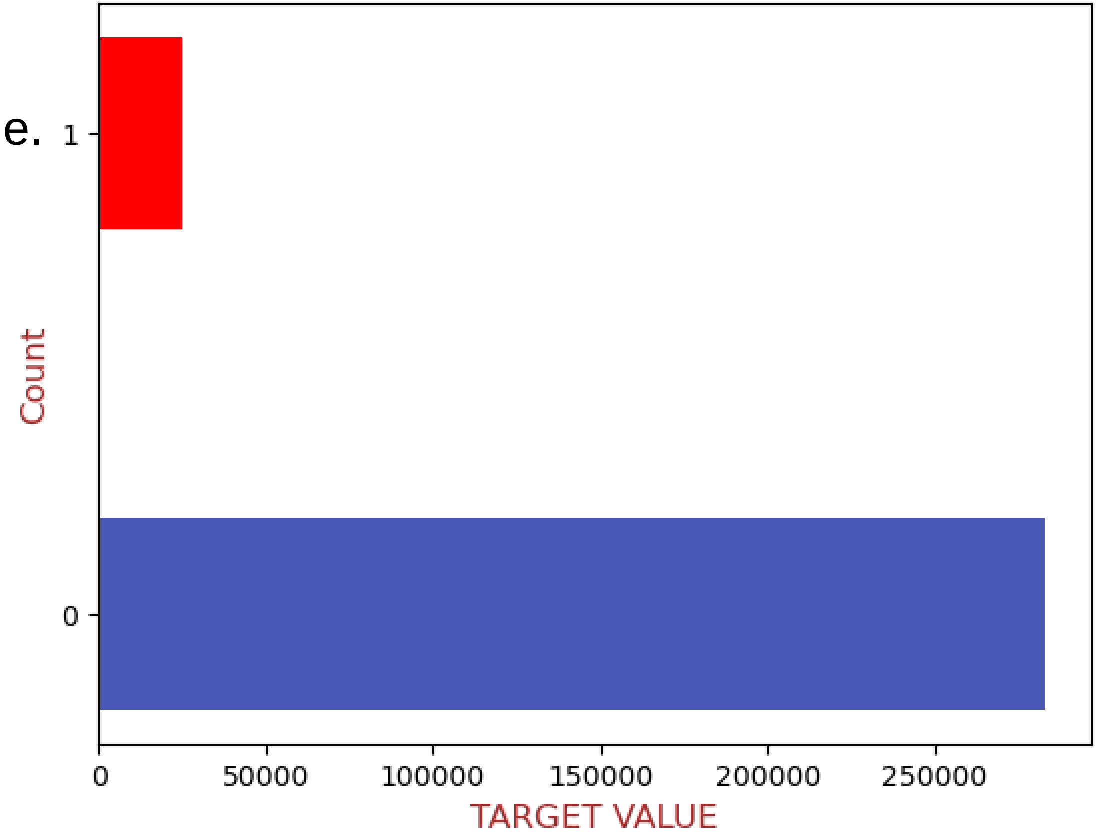
Column	Dtype	Type of data	Subtype	Distribution Anomaly / Relevance	Action
FLAG_DOCUMENT_3	int64	Categorical	Nominal	1 for more than 71%	Skip
FLAG_DOCUMENT_4	int64	Categorical	Nominal	0 for more than 99%	Skip
FLAG_DOCUMENT_5	int64	Categorical	Nominal	0 for more than 98%	Skip
FLAG_DOCUMENT_6	int64	Categorical	Nominal	0 for more than 91%	Skip
FLAG_DOCUMENT_7	int64	Categorical	Nominal	0 for more than 99%	Skip
FLAG_DOCUMENT_8	int64	Categorical	Nominal	0 for more than 91%	Skip
FLAG_DOCUMENT_9	int64	Categorical	Nominal	0 for more than 99%	Skip
FLAG_DOCUMENT_10	int64	Categorical	Nominal	0 for more than 99%	Skip
FLAG_DOCUMENT_11	int64	Categorical	Nominal	0 for more than 99%	Skip
FLAG_DOCUMENT_12	int64	Categorical	Nominal	0 for more than 99%	Skip
FLAG_DOCUMENT_13	int64	Categorical	Nominal	0 for more than 99%	Skip
FLAG_DOCUMENT_14	int64	Categorical	Nominal	0 for more than 99%	Skip
FLAG_DOCUMENT_15	int64	Categorical	Nominal	0 for more than 99%	Skip
FLAG_DOCUMENT_16	int64	Categorical	Nominal	0 for more than 99%	Skip
FLAG_DOCUMENT_17	int64	Categorical	Nominal	0 for more than 99%	Skip
FLAG_DOCUMENT_18	int64	Categorical	Nominal	0 for more than 99%	Skip
FLAG_DOCUMENT_19	int64	Categorical	Nominal	0 for more than 99%	Skip
FLAG_DOCUMENT_20	int64	Categorical	Nominal	0 for more than 99%	Skip
FLAG_DOCUMENT_21	int64	Categorical	Nominal	0 for more than 99%	Skip
AMT_REQ_CREDIT_BUREAU_HOUR	float64	Numerical	Continuous	These columns contains, number of enquiries to Credit Bureau about the client during particular duration before application. Understanding the dataset, number of enquiries upto 1 month back can be analyzed. As values are exclusive, suming upto month to find new column.	Change
AMT_REQ_CREDIT_BUREAU_DAY	float64	Numerical	Continuous		
AMT_REQ_CREDIT_BUREAU_WEEK	float64	Numerical	Continuous		
AMT_REQ_CREDIT_BUREAU_MON	float64	Numerical	Continuous		
AMT_REQ_CREDIT_BUREAU_QRT	float64	Numerical	Continuous		
AMT_REQ_CREDIT_BUREAU_YEAR	float64	Numerical	Continuous		

DATA IMBALANCE

TARGET VARIABLE

- TARGET attribute contains the flag about whether a client has difficulty in paying loan or not. It contains two types of scenarios:
- **The client with payment difficulties (TARGET = 1)**
  - he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
- **All other cases (TARGET = 0)**
  - All other cases when the payment is paid on time.





DATA IMBALANCE IMPLICATION

- **Ratio of data imbalance**
  - Ratio of data imbalance = Number of cases with TARGET as 1 : Number of cases with TARGET as 0
  - Ratio of data imbalance = 8.07 : 91.93 = **<u>1 : 11.4</u>**
- **Implication due to Data Imbalance**
  - TARGET attribute plays a important role in identification of client's capability to pay back loan. As there is a huge difference in two possible options for TARGET column with approximate ratio of 1:11.4, approach of ***<u>'Segmented Analysis based on TARGET variable'</u>*** to be followed, all the data columns should be analyzed separately for two different option of TARGET variable.

OUTLIERS HANDLING

For application data and previous application data numerical variables

OUTLIERS HANDLING

- **Insight:** Outliers are observed. These outliers may be actual number like very high credit amount, goods price or number of children /members in family. As business requirement is not to drop any of data point, these outliers are capped to 99 to 99.9 percentile as mentioned below:

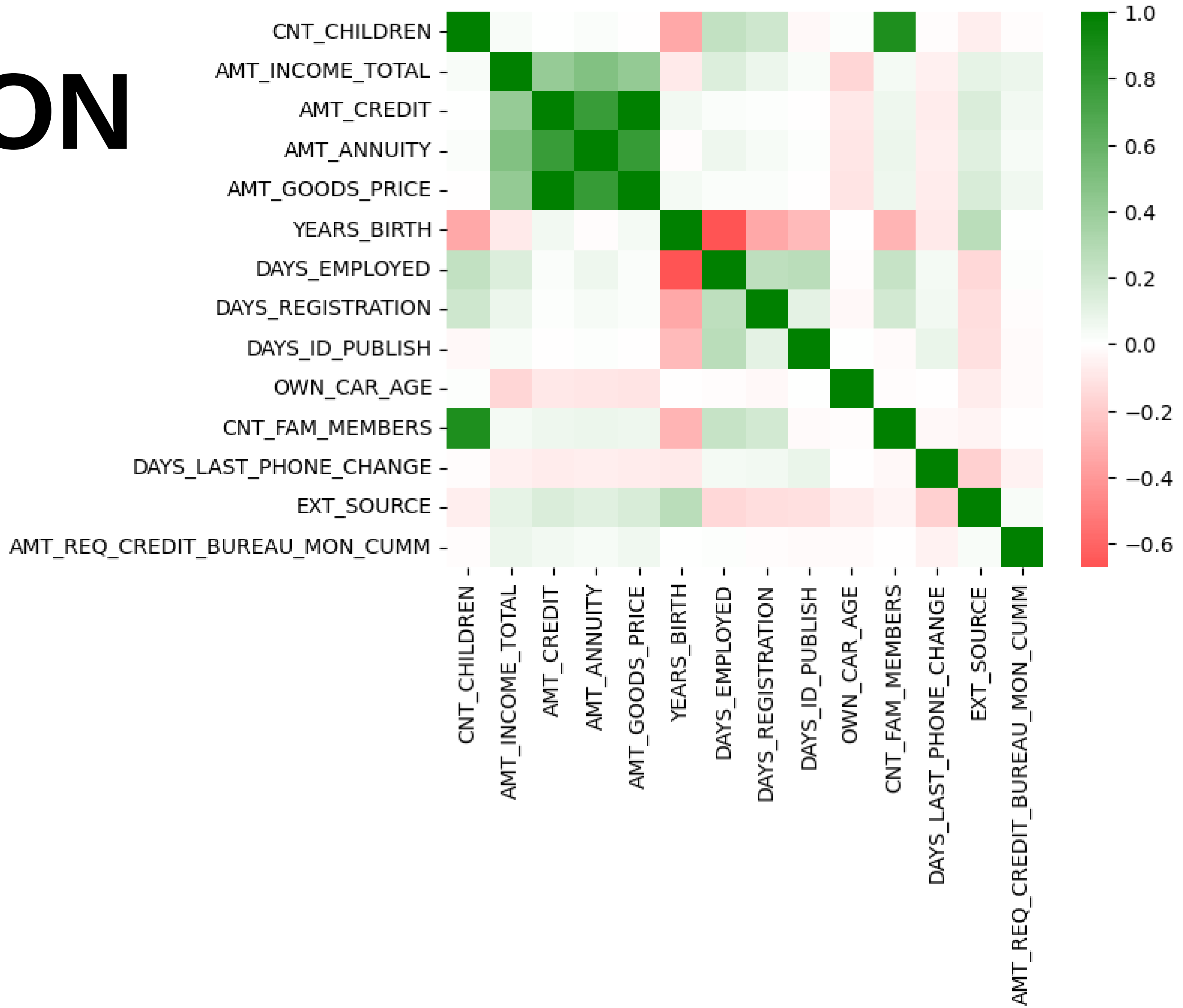
Insight	Count	%age
Application Dataset	CNT_CHILDREN	All outliers greater than 99.9 percentile are capped.
	AMT_INCOME_TOTAL	All outliers greater than 99 percentile are capped.
	AMT_CREDIT	All outliers greater than 99.9 percentile are capped.
	AMT_ANNUITY	All outliers greater than 99.9 percentile are capped.
	AMT_GOODS_PRICE	All outliers greater than 99.9 percentile are capped.
	DAYS_EMPLOYED	All outliers greater than 99 percentile are capped.
	DAYS_REGISTRATION	All outliers greater than 99 percentile are capped.
	CNT_FAM_MEMBERS	All outliers greater than 99.9 percentile are capped.
Previous Application Dataset	AMT_ANNUITY_PREV	All outliers greater than 99.9 percentile are capped.
	AMT_APPLICATION	All outliers greater than 99.9 percentile are capped.
	AMT_CREDIT_PREV	All outliers greater than 99.9 percentile are capped.
	AMT_GOODS_PRICE_PREV	All outliers greater than 99.9 percentile are capped.

CORRELATION MATRIX

Among Numerical Variables

CORRELATION

For clients with NO Payment difficulty (TARGET = 0)



CORRELATION

For clients with Payment difficulty (TARGET = 1)

CORRELATION

Correlation between numerical variables after merging both datasets

COR-RELATION

TARGET = 0 (TOP 10)

Variable 1	Variable 2	Corr
AMT_CREDIT	AMT_GOODS_PRICE	0.987030
CNT_CHILDREN	CNT_FAM_MEMBERS	0.877133
AMT_ANNUITY	AMT_GOODS_PRICE	0.781747
AMT_CREDIT	AMT_ANNUITY	0.777337
YEARS_BIRTH	DAYS_EMPLOYED	-0.670070
AMT_INCOME_TOTAL	AMT_ANNUITY	0.489758
AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.418490
AMT_INCOME_TOTAL	AMT_CREDIT	0.412714
CNT_CHILDREN	YEARS_BIRTH	-0.339760
YEARS_BIRTH	DAYS_REGISTRATION	-0.332980

TARGET = 1 (TOP 10)

Variable 1	Variable 2	Corr
AMT_CREDIT	AMT_GOODS_PRICE	0.982739
CNT_CHILDREN	CNT_FAM_MEMBERS	0.883554
AMT_CREDIT	AMT_ANNUITY	0.753014
AMT_ANNUITY	AMT_GOODS_PRICE	0.752931
YEARS_BIRTH	DAYS_EMPLOYED	-0.623850
AMT_INCOME_TOTAL	AMT_ANNUITY	0.428473
AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.352929
AMT_INCOME_TOTAL	AMT_CREDIT	0.351573
YEARS_BIRTH	DAYS_REGISTRATION	-0.289020
CNT_CHILDREN	YEARS_BIRTH	-0.262880

Top 10 correlated variables are same for clients with payment difficulties with all other cases. Difference exists in order and correlation coefficients only.

Variable 1	Variable 2	Corr (TARGET=0)	Corr (TARGET=1)
AMT_CREDIT	AMT_GOODS_PRICE	0.987030	0.982739
CNT_CHILDREN	CNT_FAM_MEMBERS	0.877133	0.883554
AMT_ANNUITY	AMT_GOODS_PRICE	0.781747	0.753014
AMT_CREDIT	AMT_ANNUITY	0.777337	0.752931
YEARS_BIRTH	DAYS_EMPLOYED	-0.670070	-0.623850
AMT_INCOME_TOTAL	AMT_ANNUITY	0.489758	0.428473
AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.418490	0.352929
AMT_INCOME_TOTAL	AMT_CREDIT	0.412714	0.351573
CNT_CHILDREN	YEARS_BIRTH	-0.339760	-0.289020
YEARS_BIRTH	DAYS_REGISTRATION	-0.332980	-0.262880

Top 10 Order is made on absolute value, as good negative correlation is also good relation among variables.

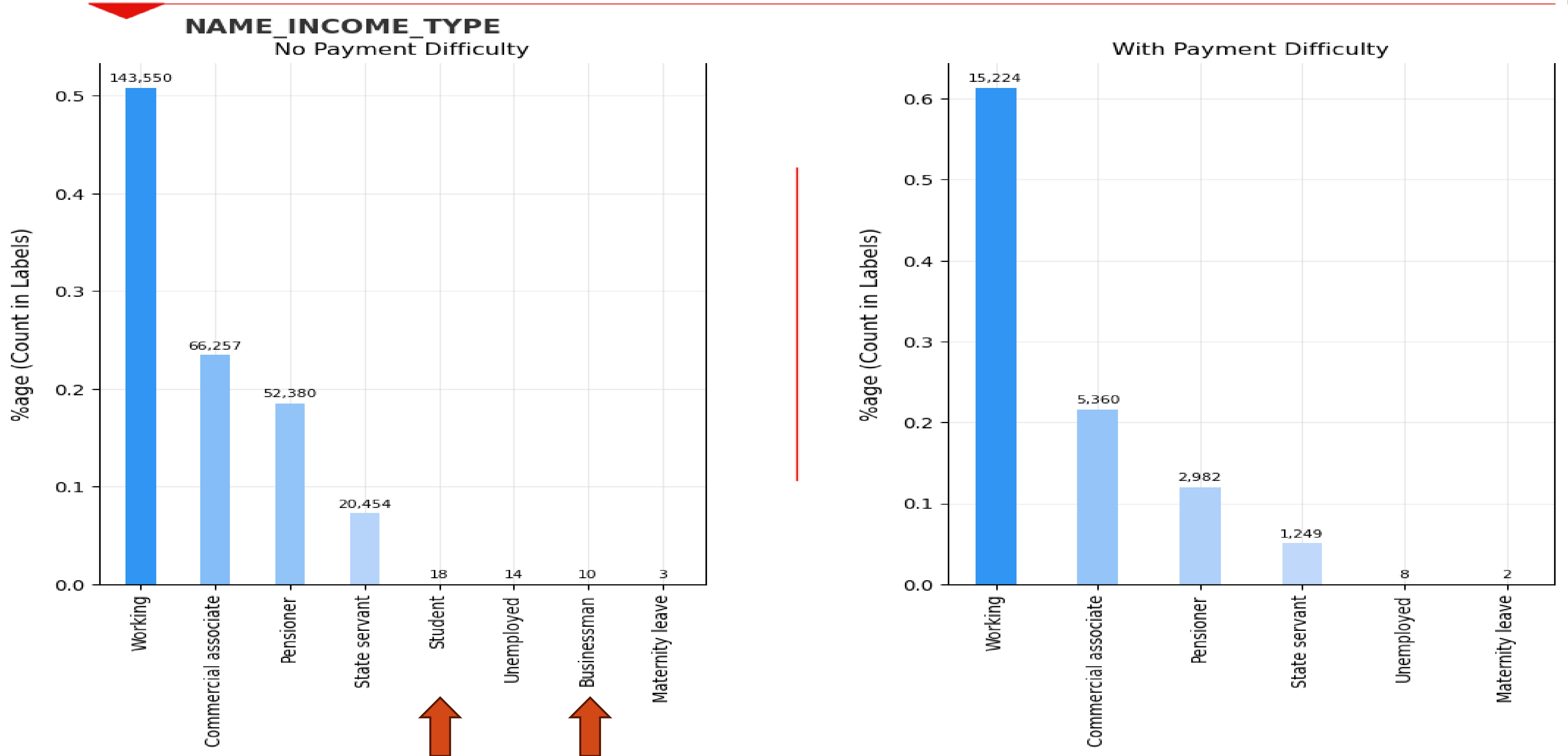
TARGET variable will be studied in addition as differentiator.

IMPORTANT FACTS FROM ANALYSIS

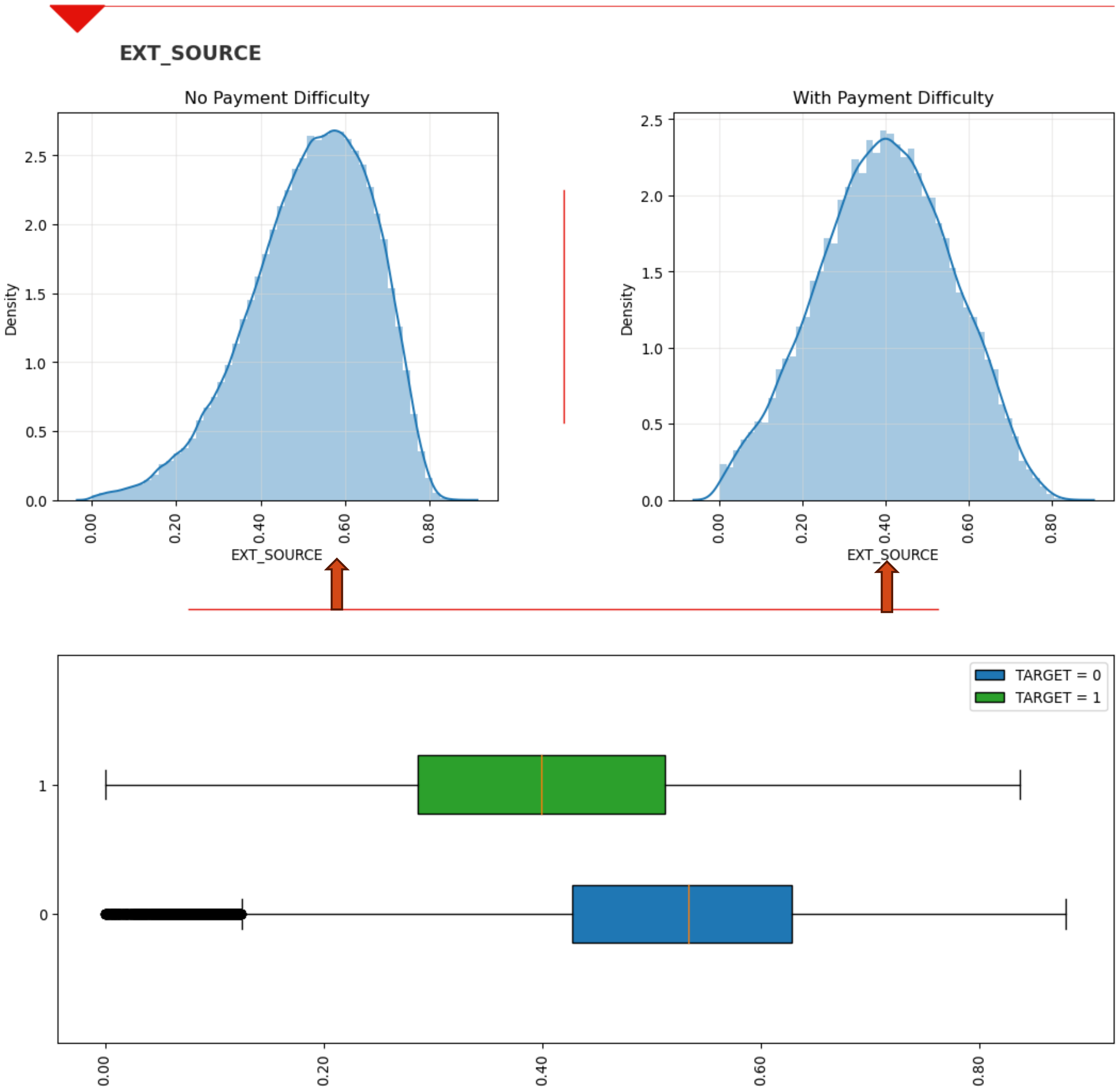
Various analysis technique being used for study (Custom Functions for Visualization): -

- Univariant Analysis
  - Categorical Variables
  - Numerical Variables
- Bivariant Analysis
  - Categorical vs Categorical
  - Continuous vs Continuous
- Multivariant Analysis

- **NAME_INCOME_TYPE**
  - Students and Businessman have 'No Payment Difficulties'. These NAME_INCOME_TYPE can be less risky while repayment.
  - For others, the distribution of NAME_INCOME_TYPE for 'clients with Payment Difficulties' and all other cases are similar.



- **EXT_SOURCE**
  - Very few outliers observed that too only in case of applicants having difficulty in payments.
- Density distribution and box plots shows that ‘applicants with payment difficulty’ have peak at 0.40, whereas ‘applicants with no payment difficulty’ have peak at 0.60.
- Further investigation is done.

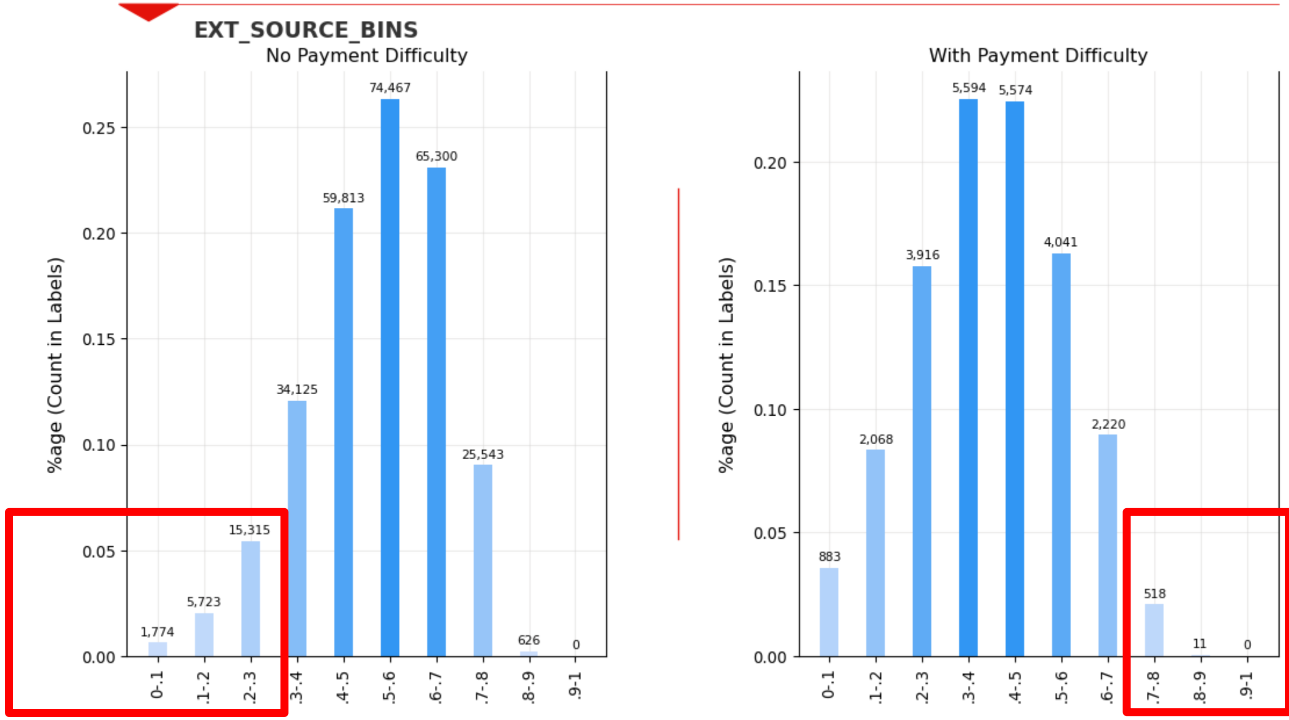


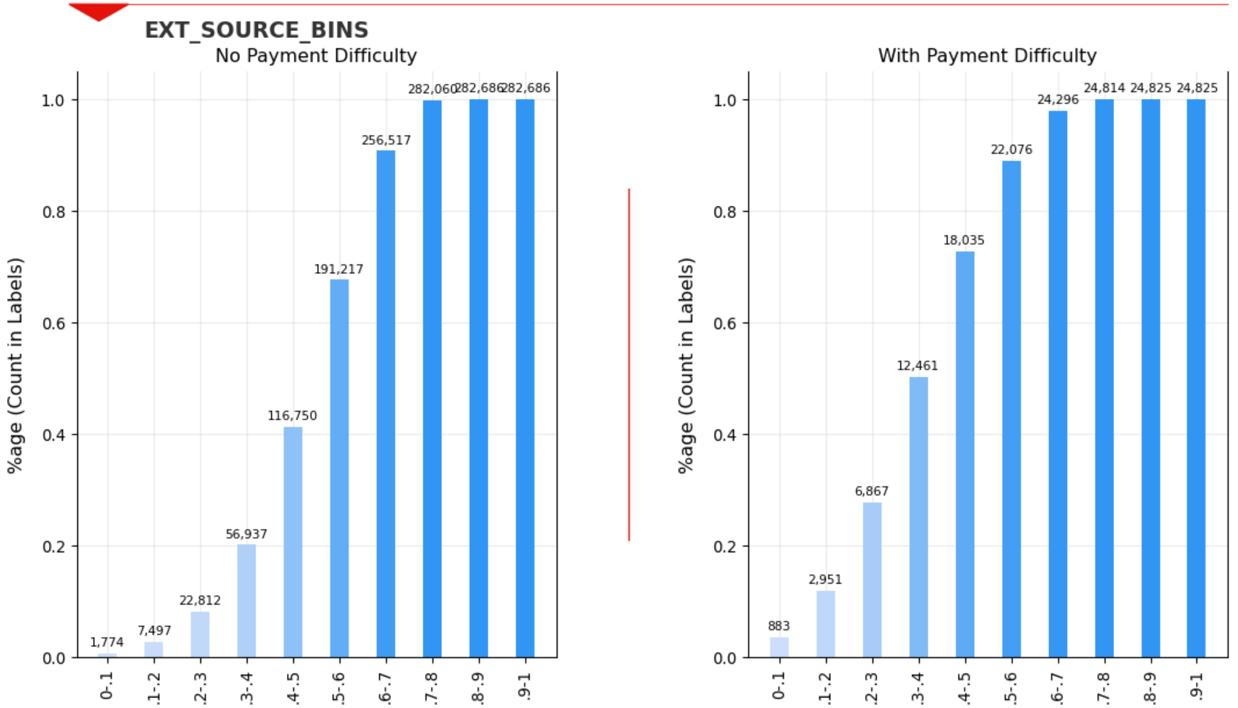
Cont. (next slide)

- **EXT_SOURCE**
  - EXT_SOURCE bins are created with interval of 0.1. Bar Plot for counts and cumulative counts for ‘No Payment Difficulty’ and ‘With Payment Difficulty’ are drawn.

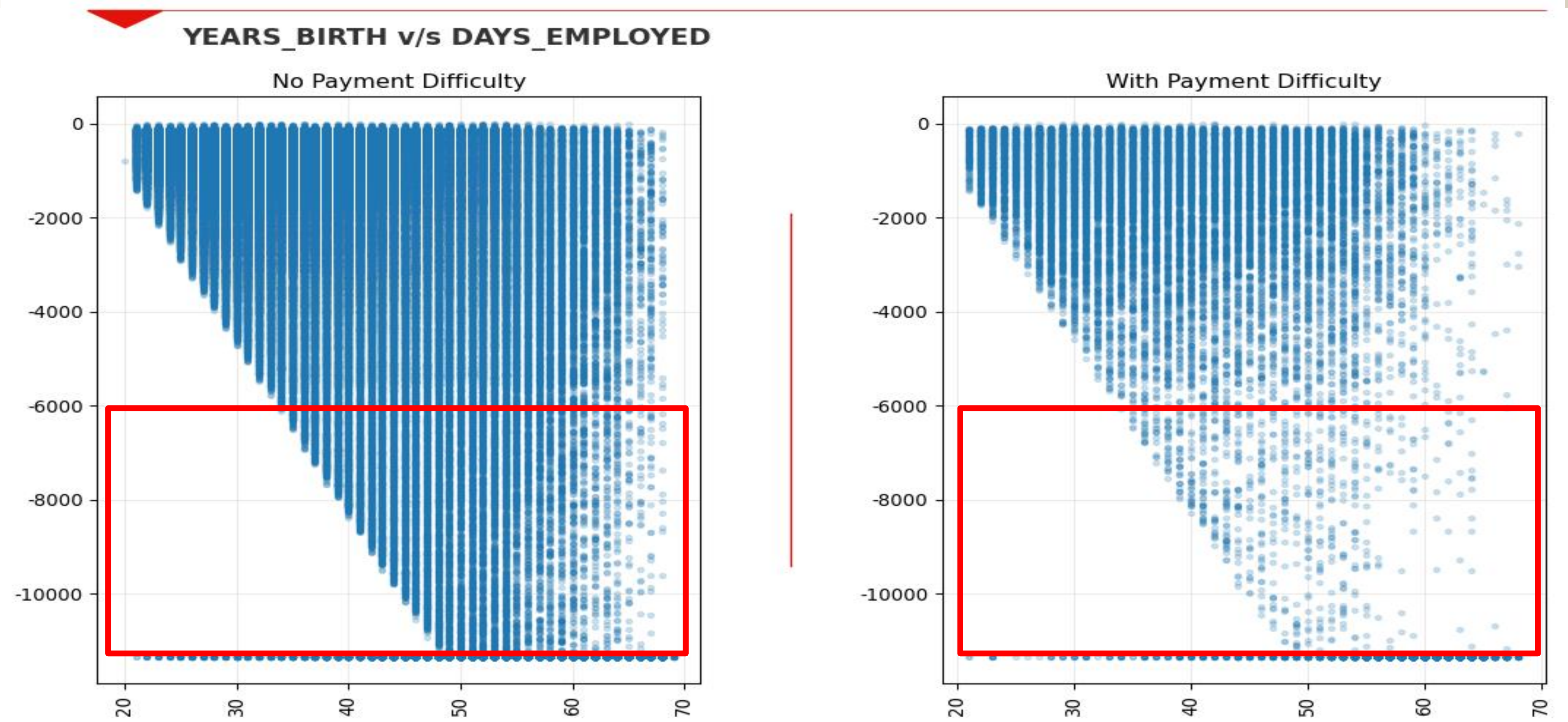
Clients %age having NO payment difficulty with EXT_SOURCE above 0.7.	98.02%
Clients %age having payment difficulty with EXT_SOURCE below 0.3.	28.24% [As there are only 8% clients in all with payment difficulty, this number is 3.5 times higher, which makes it important factor]

- It is concluded that
  - Clients with EXT_SOURCE > 0.7 have high chances of NO difficulty in payments and
  - Clients with EXT_SOURCE < 0.3 have higher chance of difficulty in payments, and should be provided loan with great scrutiny.



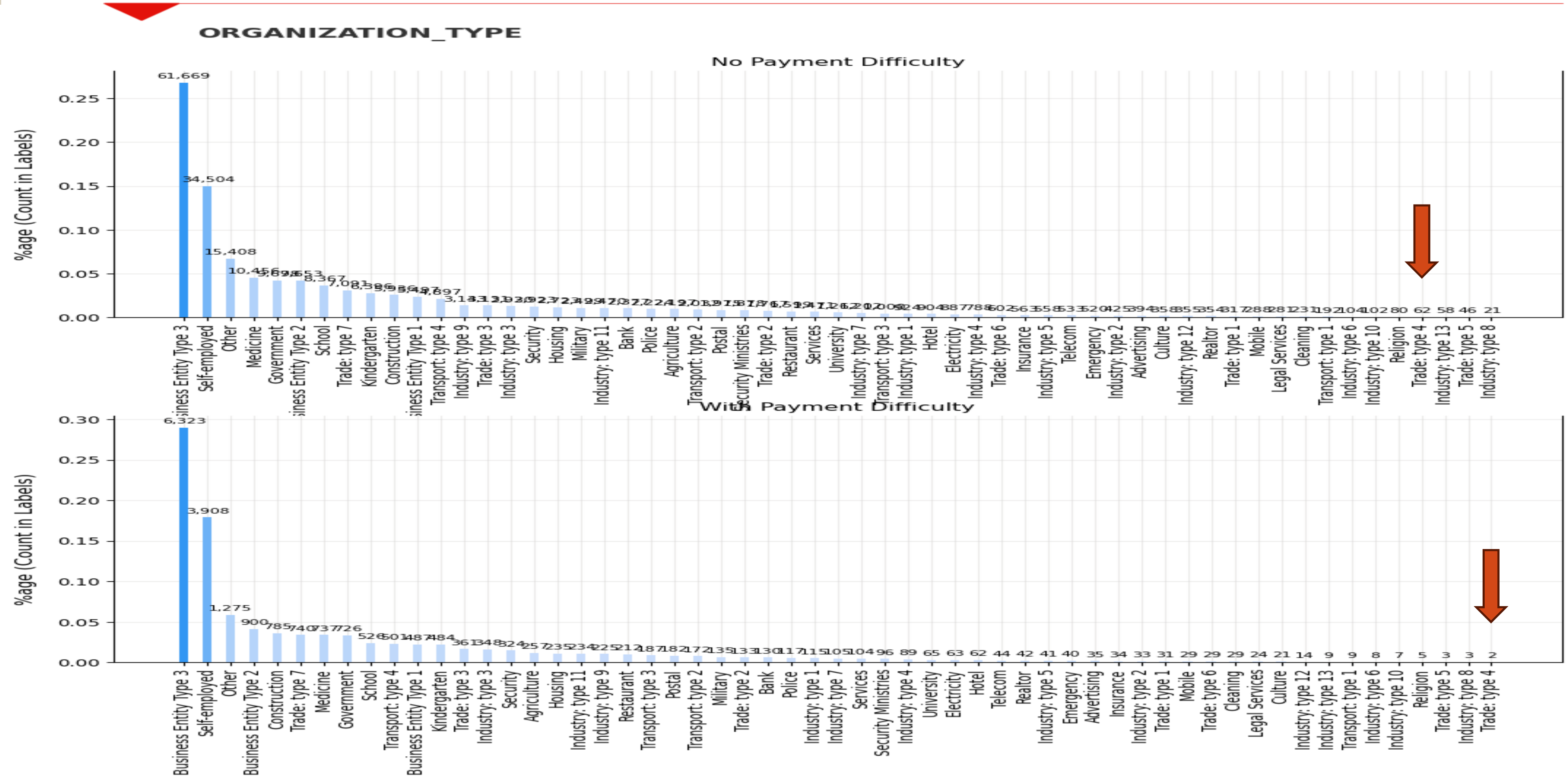


- **YEARS_BIRTH v/s DAYS_EMPLOYED**
  - As data points drastically reduce for chart 'With Payment Difficulty' below 6000. This shows that Clients who are employed for more than 6000 days have very less payment difficulties.

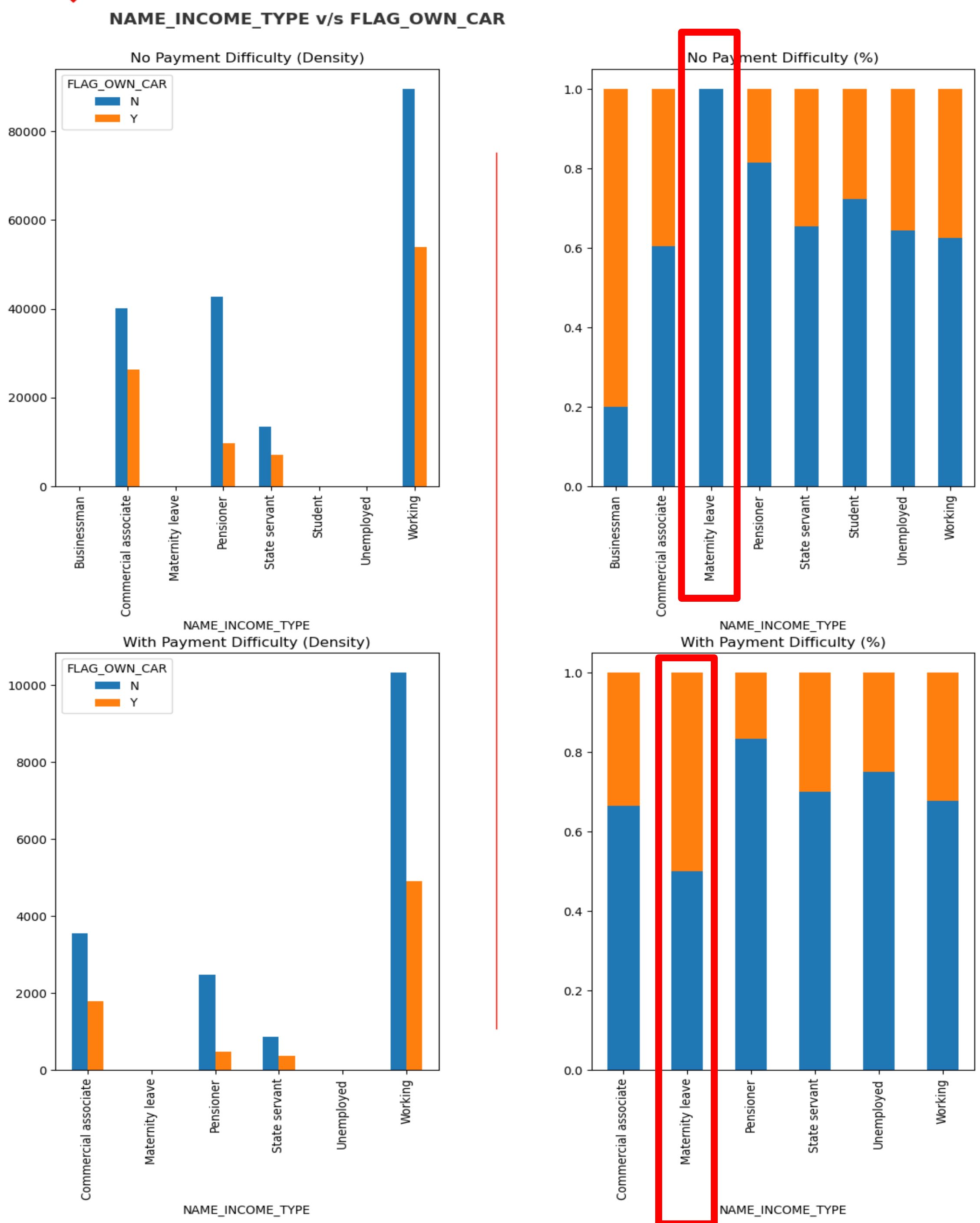


ORGANIZATION_TYPE

- The distribution of ORGANIZATION_TYPE for ‘clients with Payment Difficulties’ and all other cases are similar.
- However, further analysis show that clients in ‘Trade: type 4’ have very less chances of ‘Payment Difficulties’ (High ‘No Payment Difficulty’ compare to others).

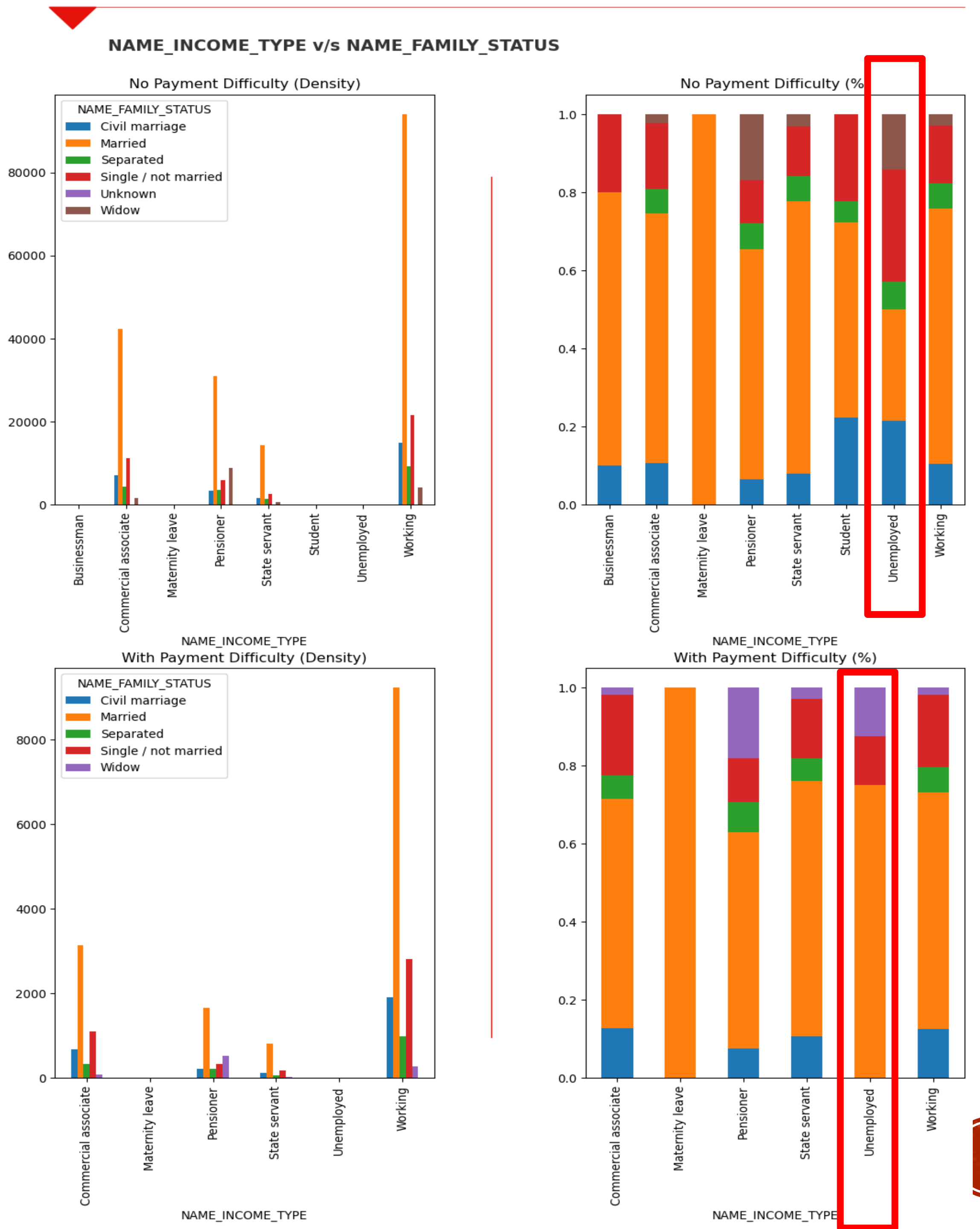


- **NAME_INCOME_TYPE v/s FLAG_OWN_CAR**
  - Clients with ‘Maternity Leave’ income and own car, have payment difficulty.
  - There is a very less correlation between NAME_INCOME_TYPE and FLAG_OWN_CAR.



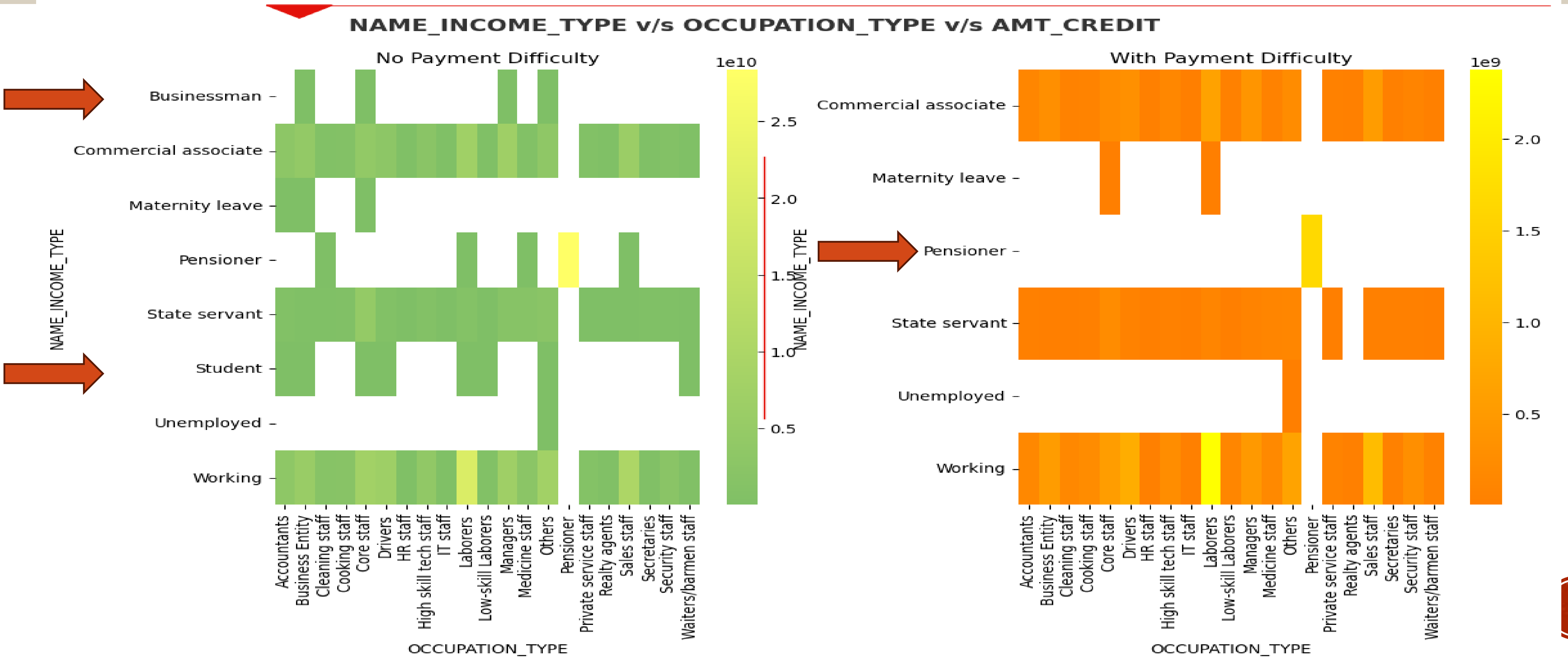
NAME_INCOME_TYPE v/s NAME_FAMILY_STATUS

- Unemployed clients with family status of ‘Civil marriage’ or ‘Separated’ have NO payment difficulty.



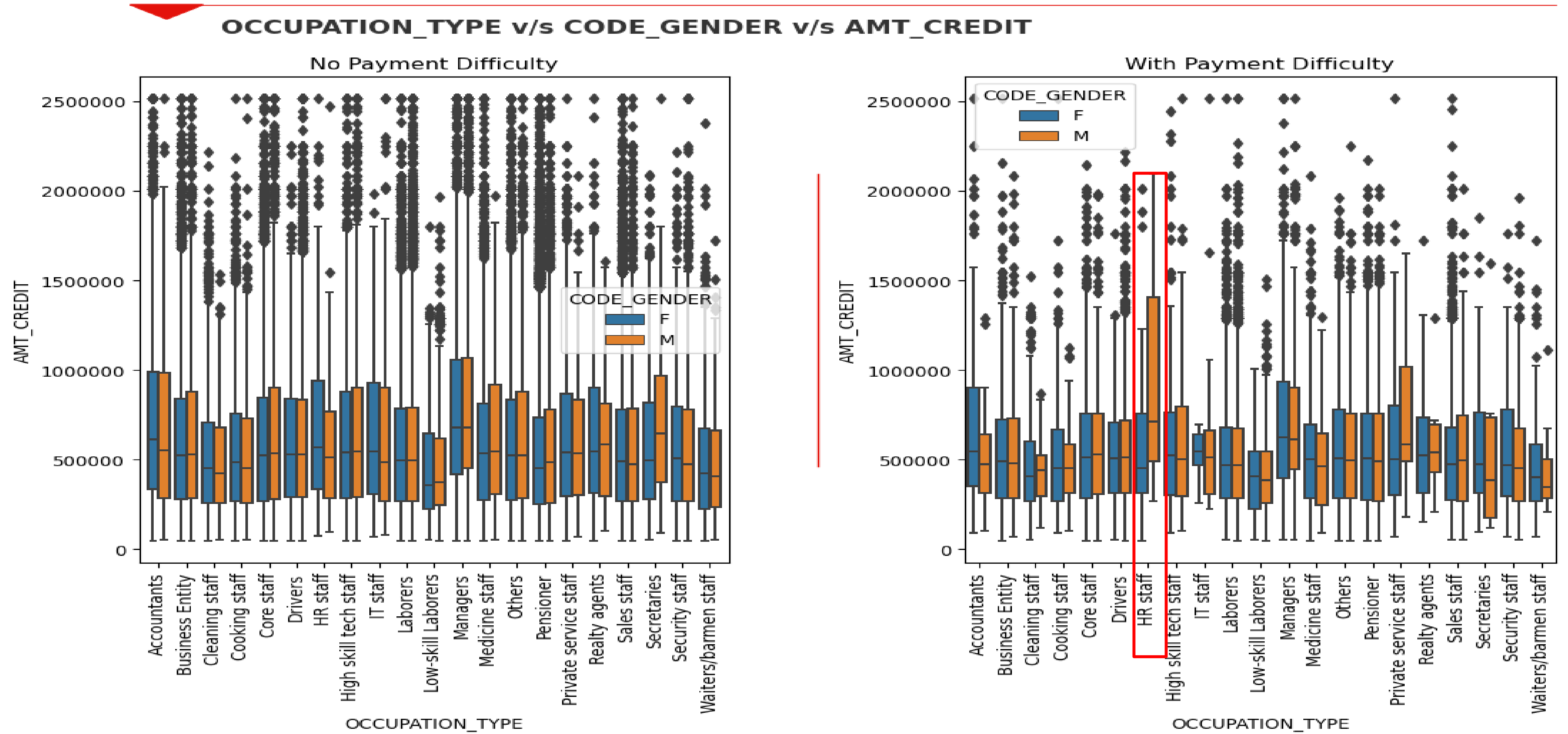
NAME_INCOME_TYPE v/s OCCUPATION_TYPE v/s AMT_CREDIT

- Businessman and Students have no payment difficulty.
- Pensioners with occupation type other than ‘Pensioner’ and <BLANK>, have no payment difficulty.

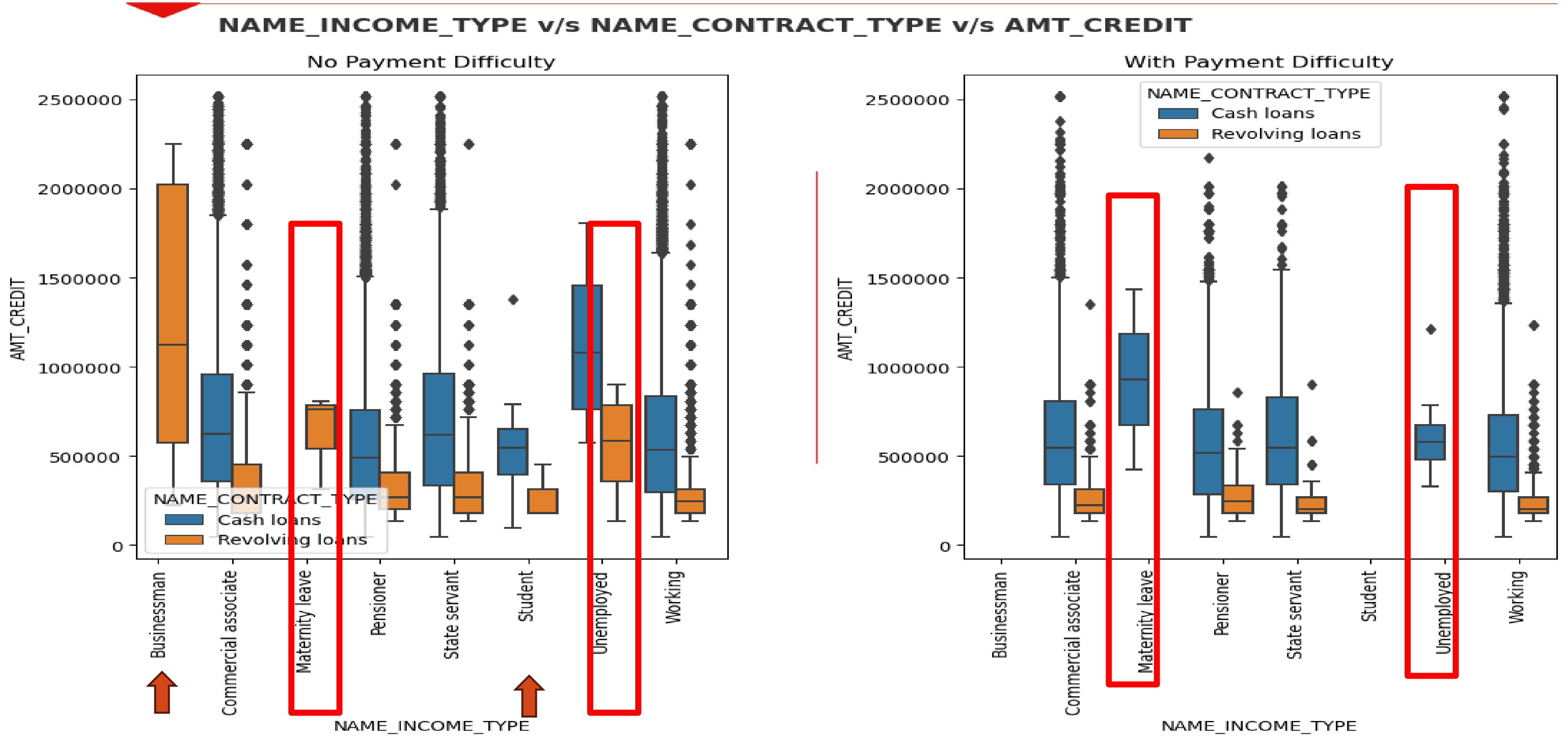


OCCUPATION_TYPE v/s CODE_GENDER v/s AMT_CREDIT

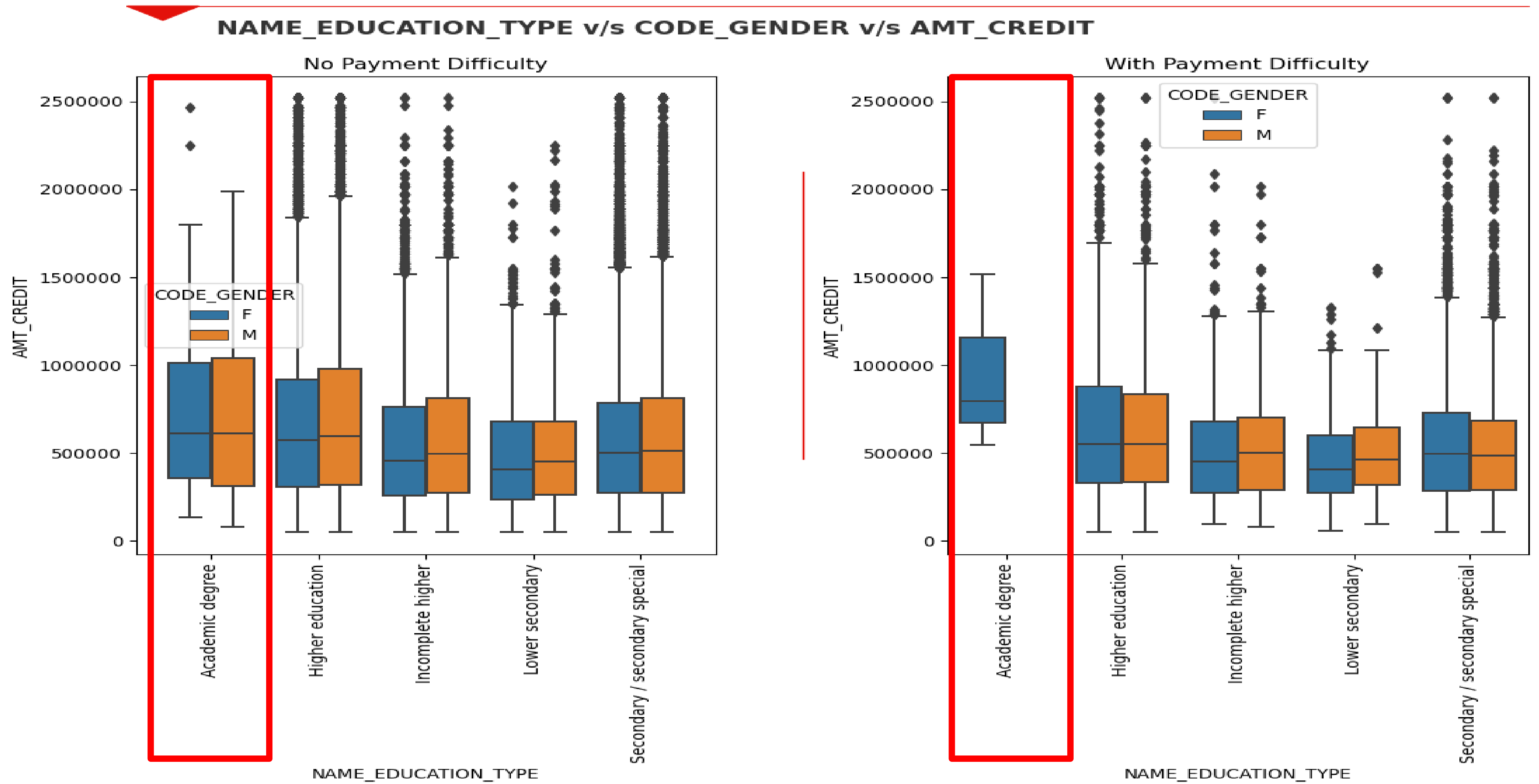
- Male HR staff with higher Credit Amount (above 12 lakh) have higher chances of payment difficulty.
- For rest of income types, boxplots for ‘clients with Payment Difficulties’ and all other cases are similar.



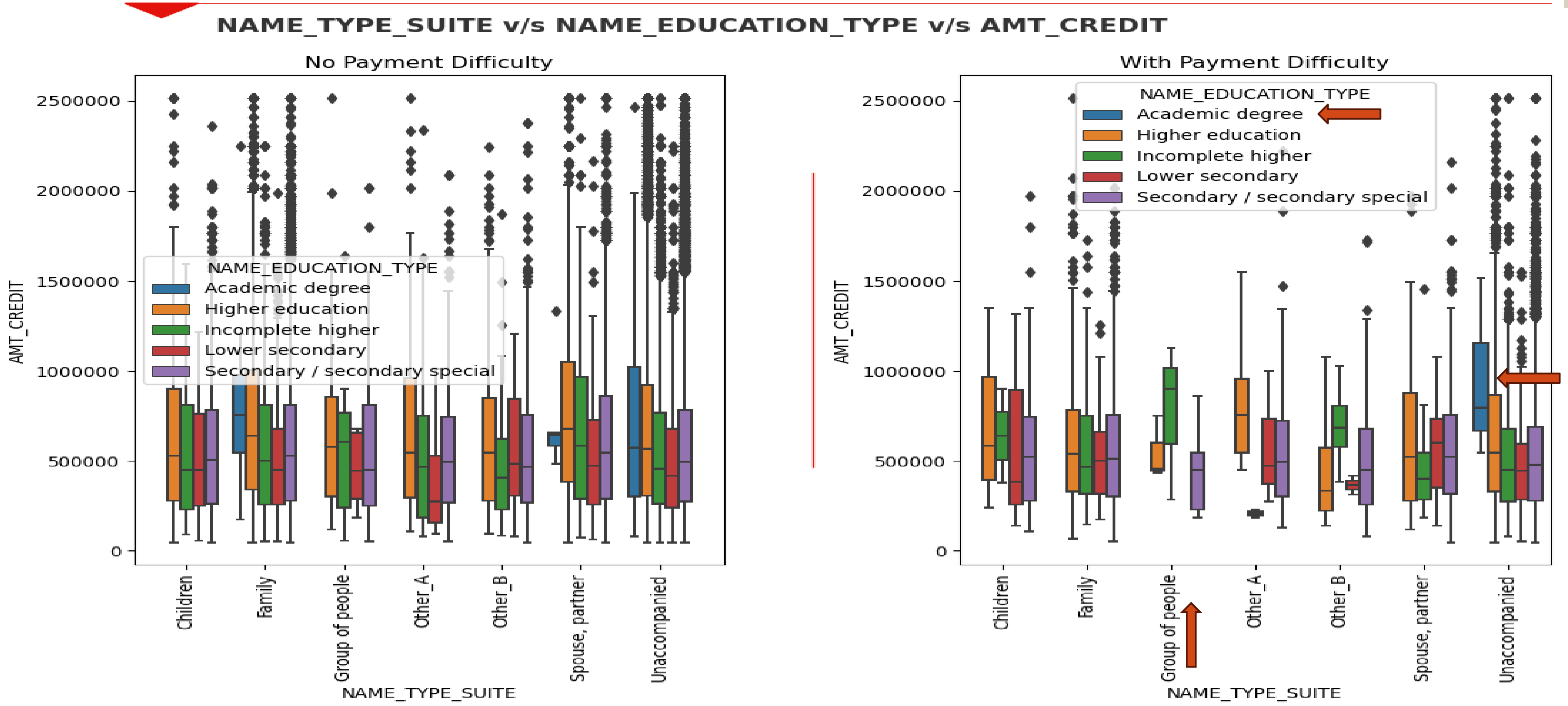
- **NAME_INCOME_TYPE v/s NAME_CONTRACT_TYPE v/s AMT_CREDIT**
  - Unemployed and client with income type 'Maternity leave' have NO payment difficulty in repayment of 'Revolving loans'.
  - Unemployed and client with income type 'Maternity leave' have payment difficulty in repayment of 'Cash loans'.
  - Businessman and Students have no payment difficulty.



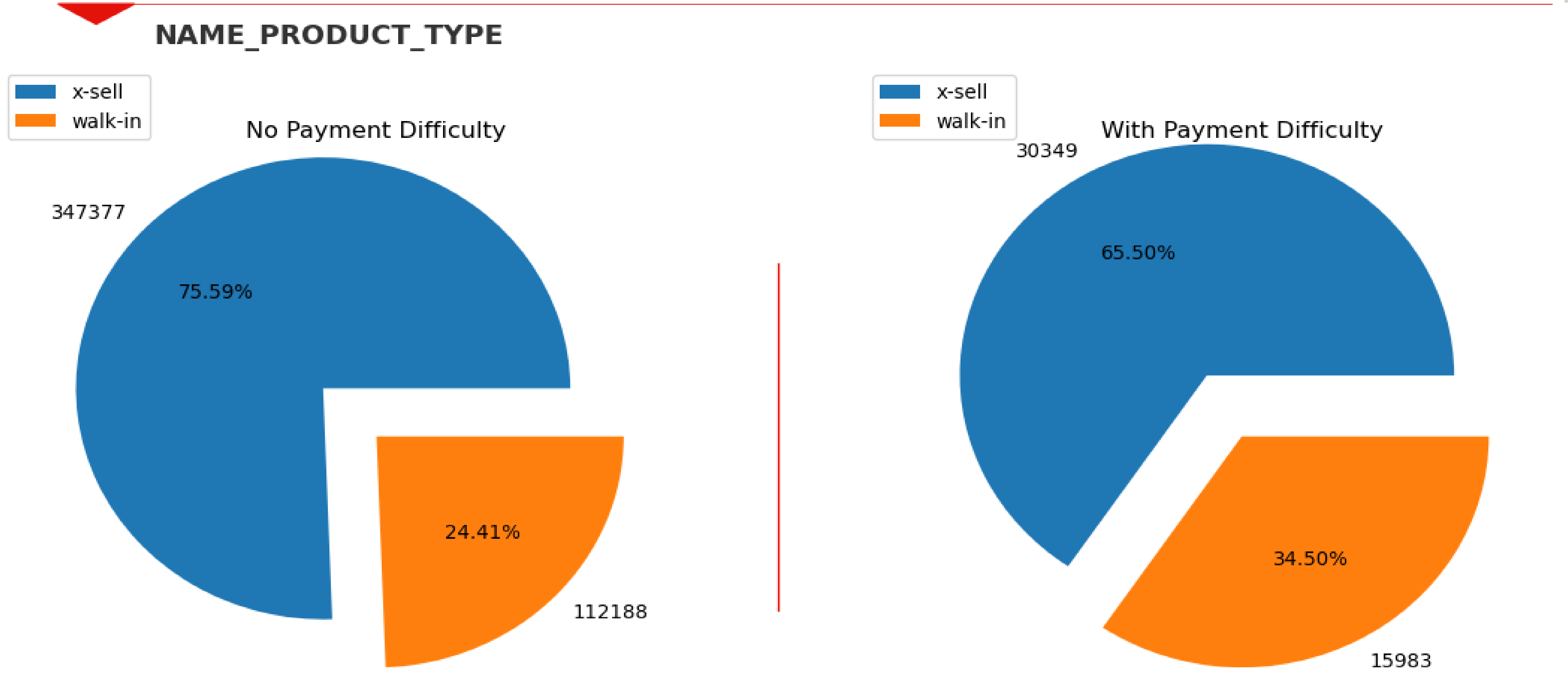
- **NAME_EDUCATION_TYPE v/s CODE_GENDER v/s AMT_CREDIT**
  - Male clients with ‘Academic degree’ have no Payment Difficulties.
  - For other education types and gender, boxplots for ‘clients with Payment Difficulties’ and all other cases are similar.



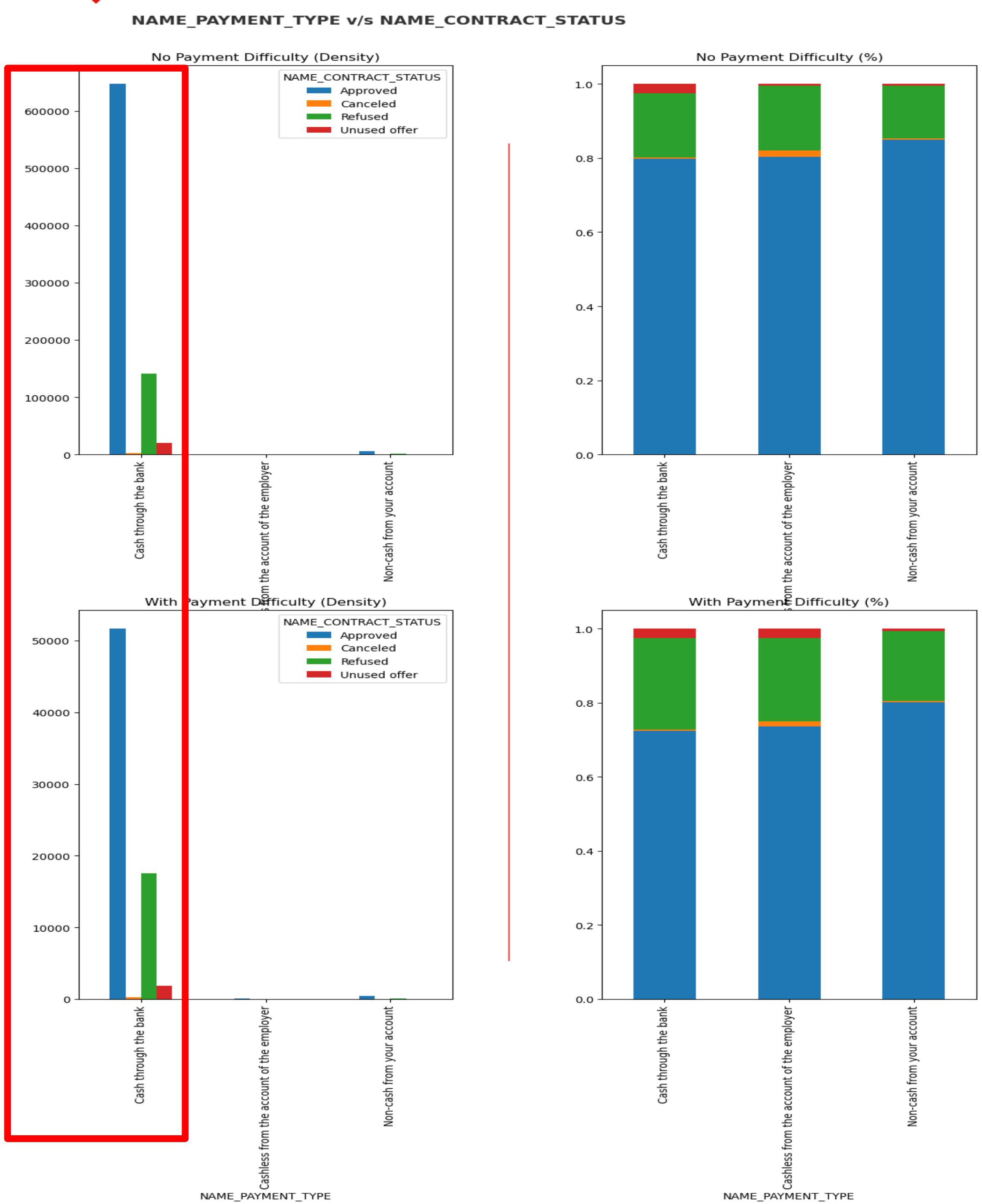
- **NAME_TYPE_SUITE v/s NAME_EDUCATION_TYPE v/s AMT_CREDIT**
  - Clients with 'Academic degree' and Accompanied (not 'Unaccompanied') have no Payment Difficulties.
  - Clients with 'Lower secondary' and categorized part of 'Group of people' have no Payment Difficulties.



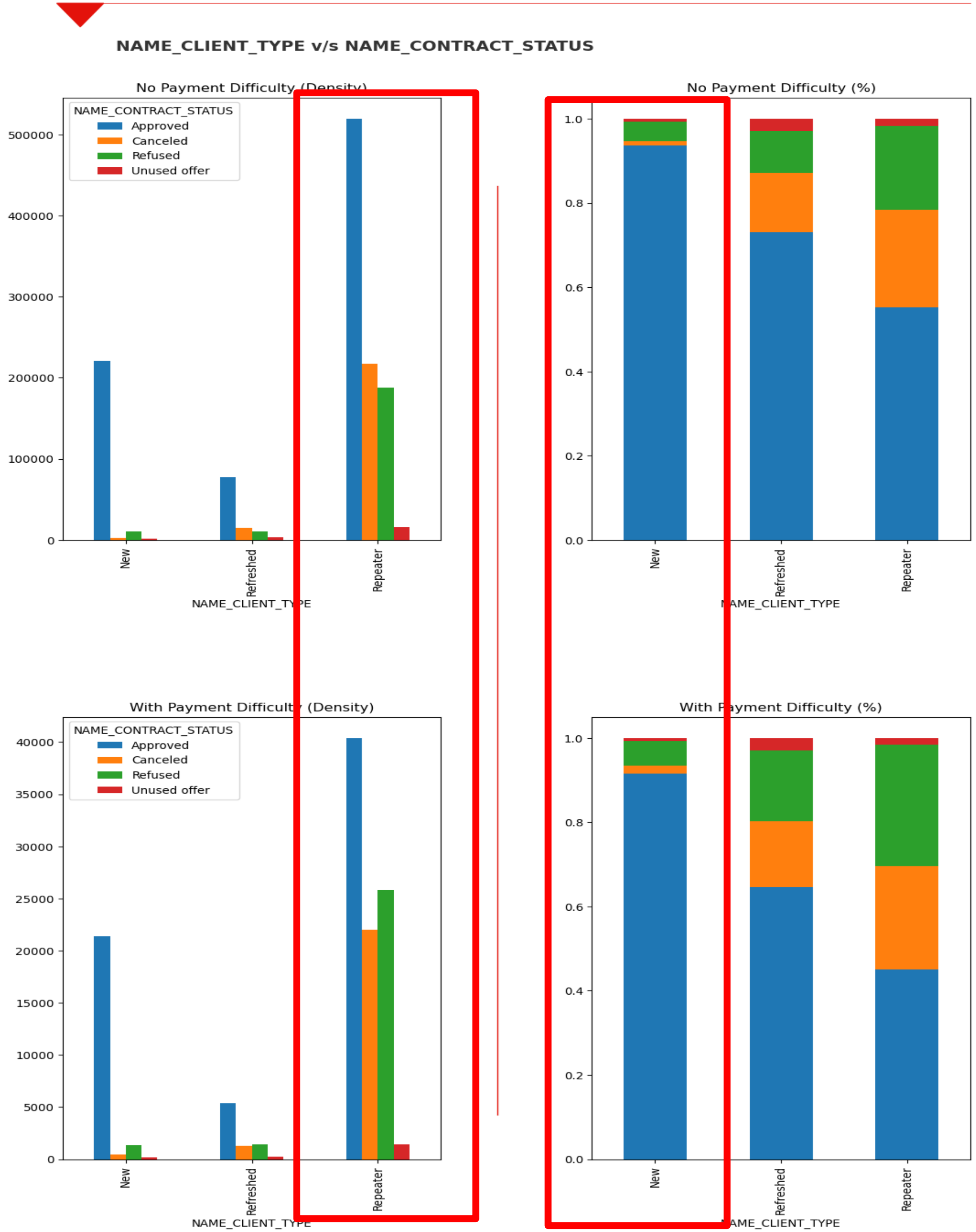
- **NAME_PRODUCT_TYPE**
  - There is higher percentage of walk-in clients with ‘Payment difficulty’.
  - The distribution is biased with XNA (Not Available) mostly, so nothing can be concluded with confidence.



- **NAME_PAYMENT_TYPE v/s NAME_CONTRACT_STATUS**
  - Client with NAME_PAYMENT_TYPE as ‘Cash through the bank’ have highest approval among all.

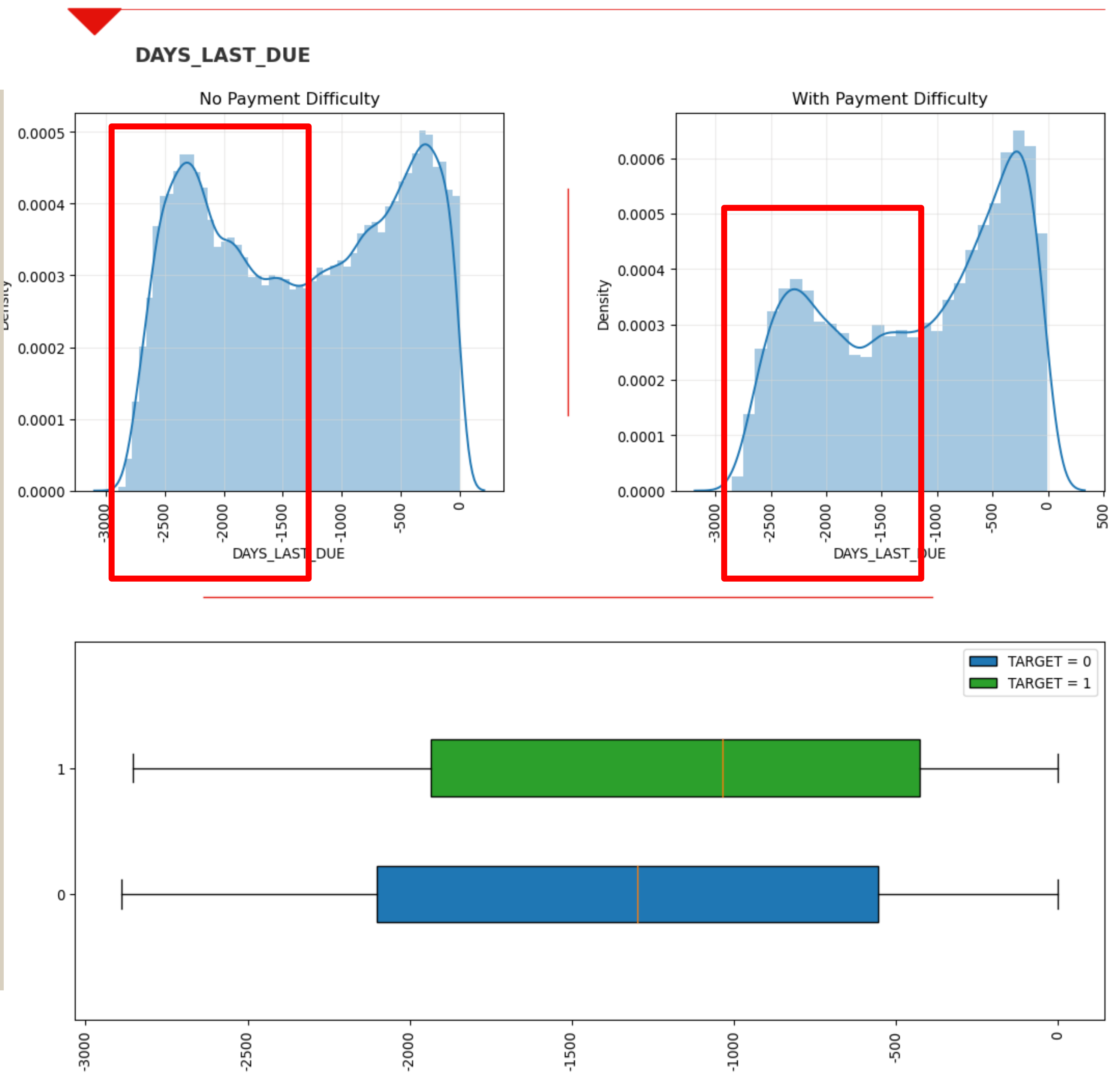


- **NAME_CLIENT_TYPE v/s NAME_CONTRACT_STATUS**
  - ‘New’ clients have highest Approval in terms of percentage in same category.
  - ‘Repeater’ clients have highest Approval in terms of count in all categories.

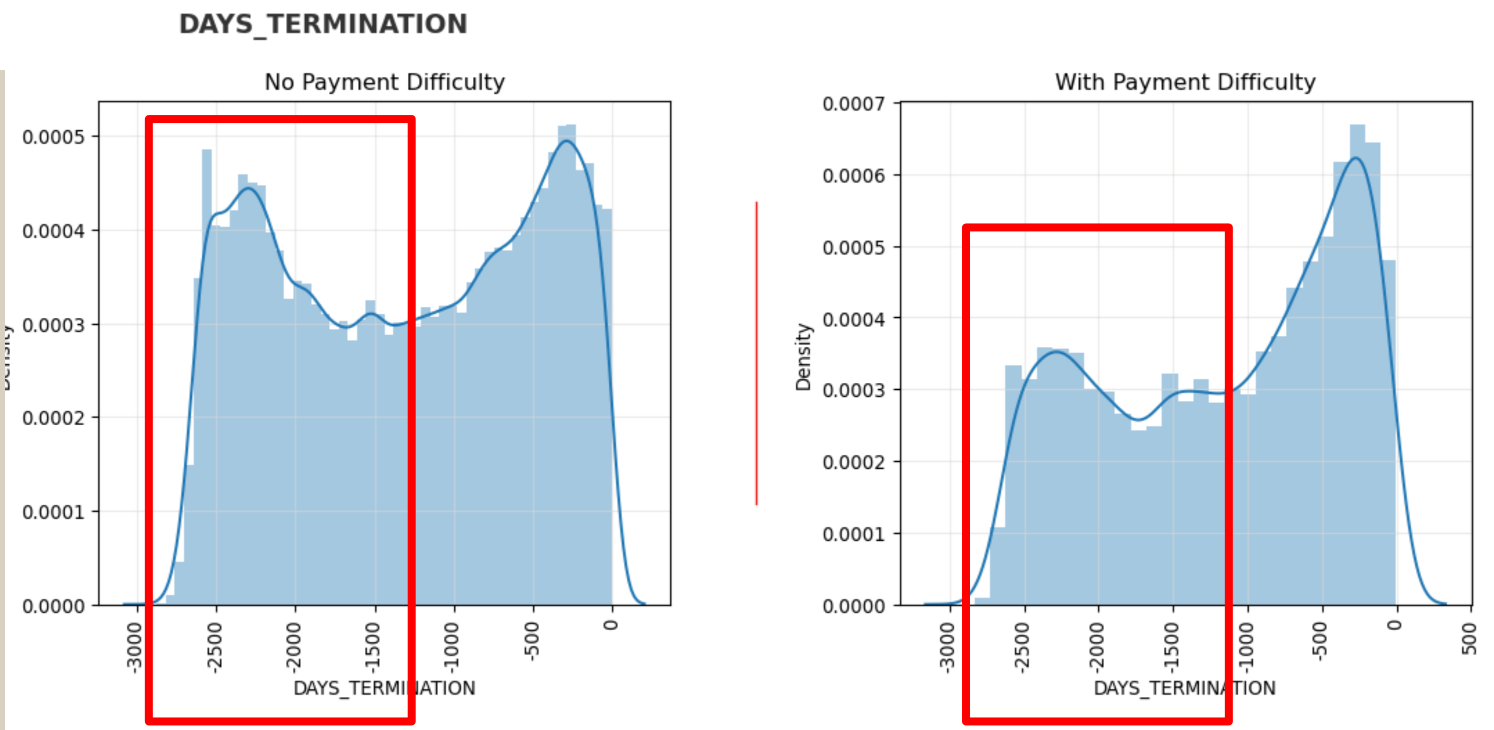


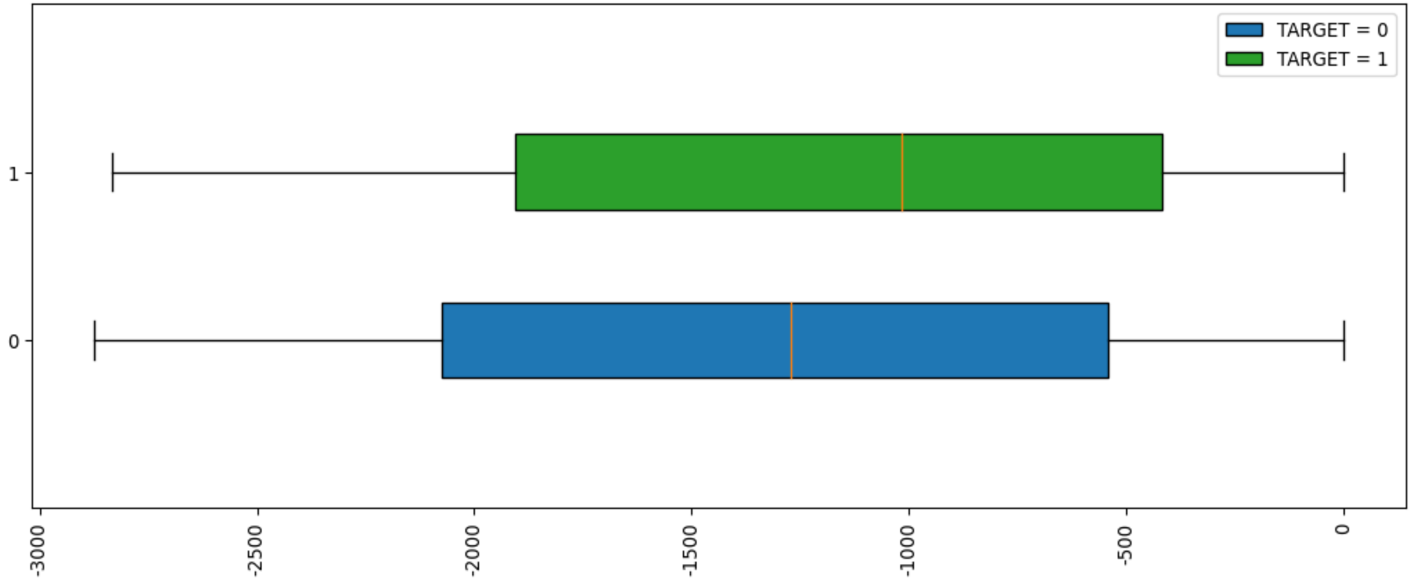
ANALYSIS OF PREVIOUS LAST_DUE

- **DAYS_LAST_DUE**
  - The clients with larger relative DAYS_LAST_DUE have better chances of ‘No Payment Difficulty’.



- **DAYS_TERMINATION**
  - The clients with larger relative DAYS_TERMINATION have better chances of ‘No Payment Difficulty’.





CONCLUSION

PORTFOLIO AND RISK ASSESSMENT

DRIVING FACTORS

Primary Factors

- NAME_INCOME_TYPE
- NAME_CLIENT_TYPE
- NAME_EDUCATION_TYPE
- CODE_GENDER
- EXT_SOURCE

Secondary Factors

- ORGANIZATION_TYPE
- DAYS_EMPLOYED
- NAME_FAMILY_STATUS
- NAME_CONTRACT_TYPE

Mild Influencing Factors

- AMT_CREDIT
- NAME_TYPE_SUITE
- FLAG_OWN_CAR
- DAYS_LAST_DUE
- DAYS_TERMINATION

PORTFOLIO OF CLIENTS THAT ARE LIKELY TO HAVE NO DIFFICULTY WHILE REPAYMENT

- Clients who are ‘Students’ or ‘Businessman’.
- Unemployed clients with family status of ‘Civil marriage’ or ‘Separated’.
- Clients with organization type as ‘Trade: type 4’.
- Clients with average rating of three external sources EXT_SOURCE > 0.7.
- Clients who are employed for more than 6000 days.
- Unemployed and client with income type ‘Maternity leave’.
- Male clients with ‘Academic degree’.
- Clients with ‘Academic degree’ and Accompanied (not ‘Unaccompanied’) have no Payment Difficulties.
- Clients with ‘Lower secondary’ and categorized part of ‘Group of people’ have no Payment Difficulties.
- ‘Repeater’ clients have highest Approval in terms of count in all categories.
- ‘New’ clients have highest Approval in terms of percentage in same category.
- The clients with larger relative last due and termination of previous applications.

PORTFOLIO OF CLIENTS – THAT ARE LIKELY TO DEFAULT

- The clients that fall under these category are risky candidates, are following actions may be taken: -
  - Denying the loan (if client fall under multiple categories as mentioned below)
  - Reducing the amount of loan
  - Lending (to risky applicants) at a higher interest rate, etc

- Clients with average rating of three external sources EXT_SOURCE < 0.3.
- Clients with 'Maternity Leave' income and own car.
- Male HR staff with higher Credit Amount (above 12 lakh).
- Unemployed and client with income type 'Maternity leave' have payment difficulty in repayment of 'Cash loans'.

THANK YOU