ITBA
Instituto Tecnológicc
de Buenos Aires

EDA - Credit Fraud

Agustin Lara

Analítica Predictiva

29/03/2023

AGENDA

01 Limpieza de Datos

02 Exploración

03 Corrplot

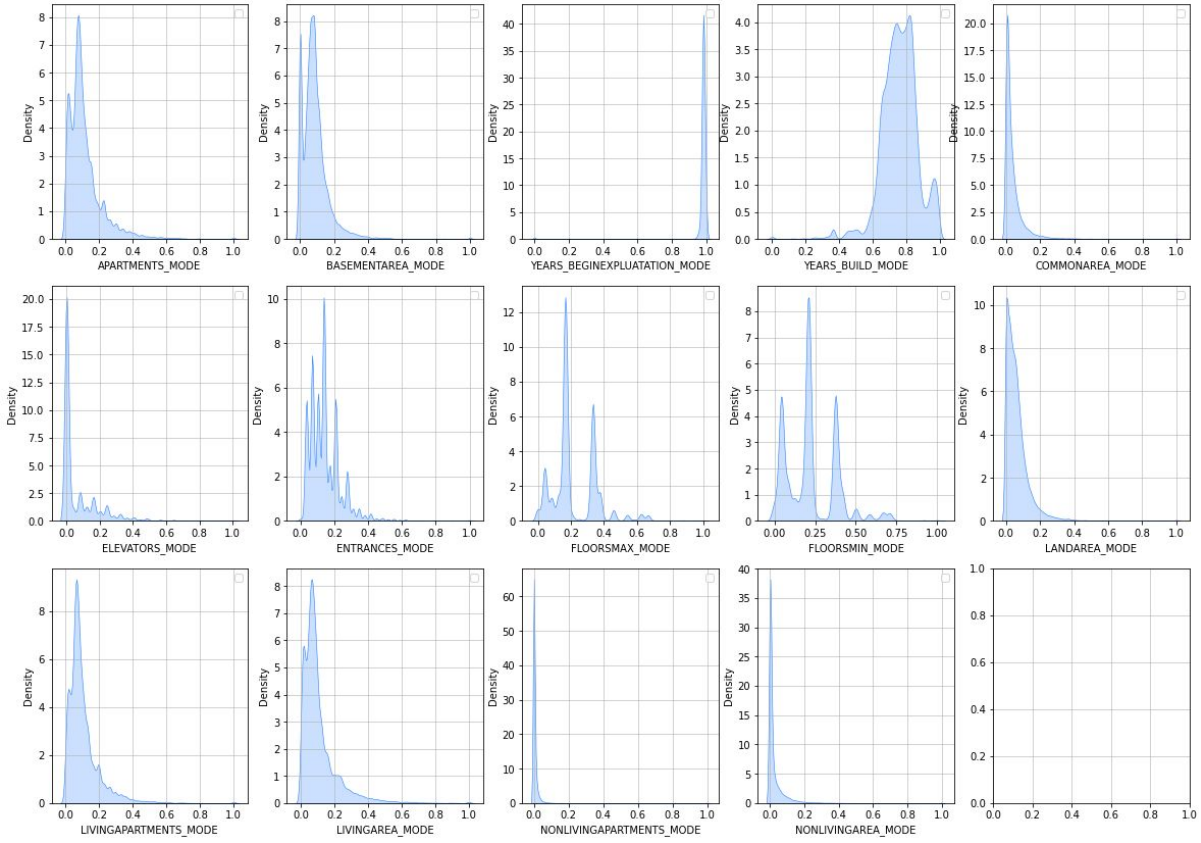
04 Conclusiones

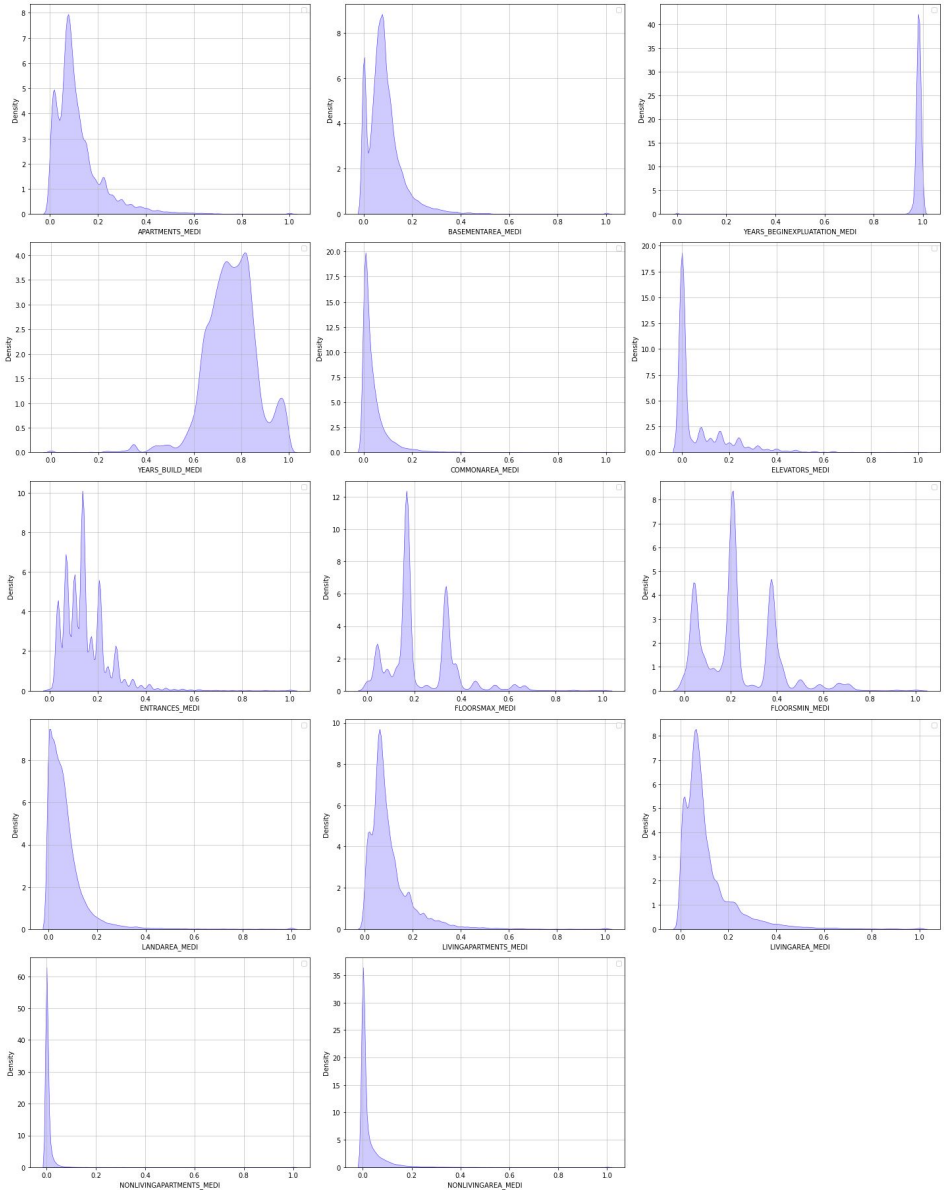
Base: “Credit Card Fraud Detection”

- 307511 registros
- 122 columnas
- dtypes: float64(65), int64(41), object(16)

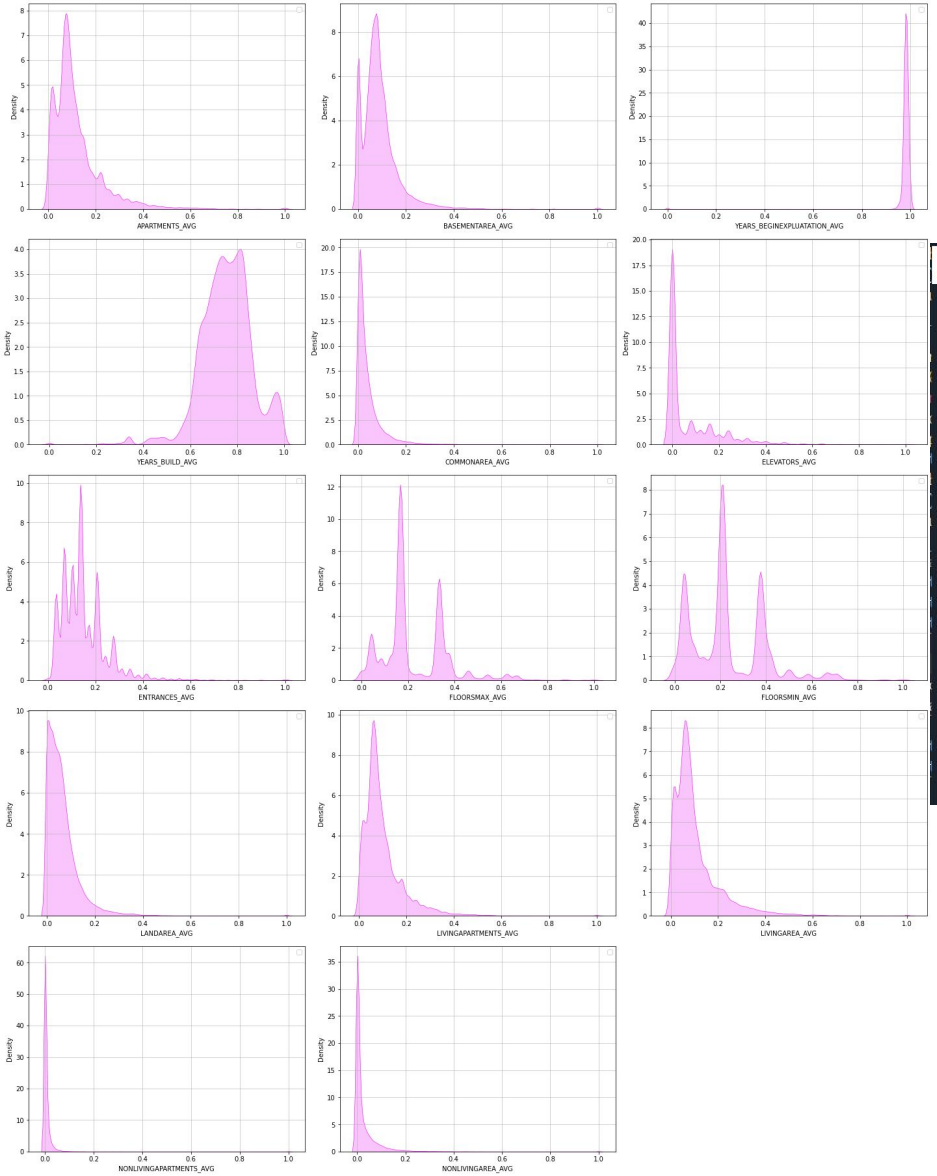
Se busca poder identificar a partir de variables categóricas y numéricas si un cliente se atrasa con los pagos de sus cuotas a la hora de tomar un préstamo

1. Variables normalizadas





2. Cambio de valores



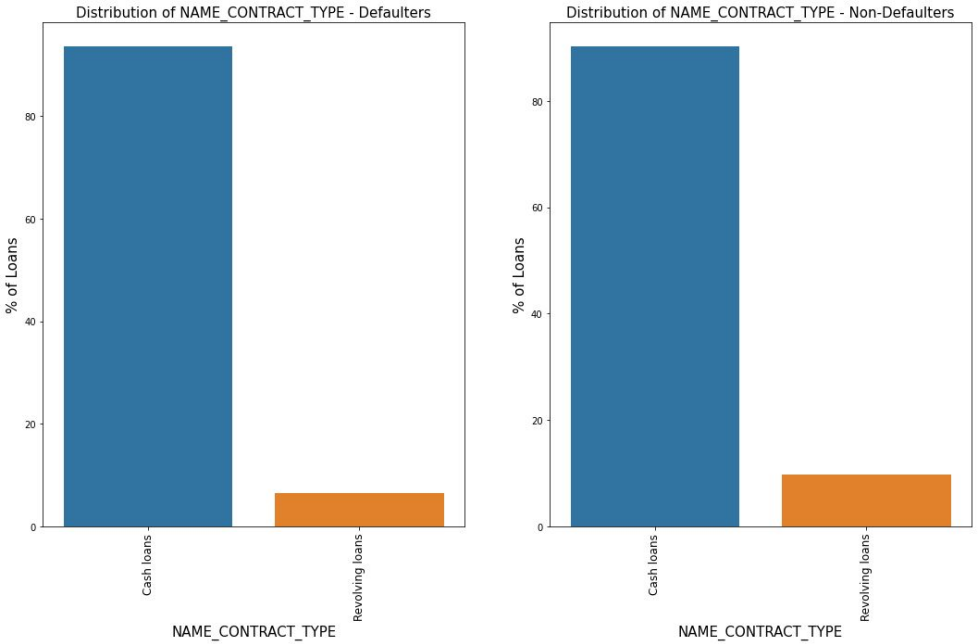
pivot - DataFrame

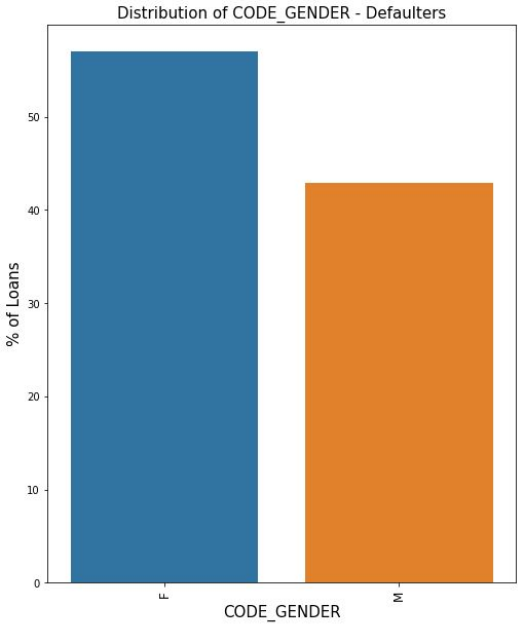
	FLAG_OWN_CAR	FLAG_OWN_REALTY	FLAG_MOBIL	FLAG_EMP_PHONE	WORK_PH	CONT_MC	FLAG_PHONE	FLAG_EMAIL	DOCUMEI	DOCUMEI	DOCUMEI	DOCUMEI	DOCUMEI	DOCUMEI	DOCUMEI
0	N	N	1	0	0	0	0	0	0	0	0	0	0	0	0
1	N	N	1	0	0	0	0	0	0	1	0	0	0	0	0
2	N	N	1	0	0	0	1	0	0	1	0	0	0	0	0
3	N	N	1	0	0	1	0	0	0	0	0	0	0	0	0
4	N	N	1	0	0	1	0	0	0	0	0	0	0	0	0
5	N	N	1	0	0	1	0	0	0	0	0	0	0	0	1
6	N	N	1	0	0	1	0	0	0	0	0	0	1	0	0
7	N	N	1	0	0	1	0	0	0	0	0	1	0	0	0
8	N	N	1	0	0	1	0	0	0	0	0	1	0	0	0
9	N	N	1	0	0	1	0	0	0	0	0	1	0	0	0
10	N	N	1	0	0	1	0	0	0	0	0	1	0	0	0

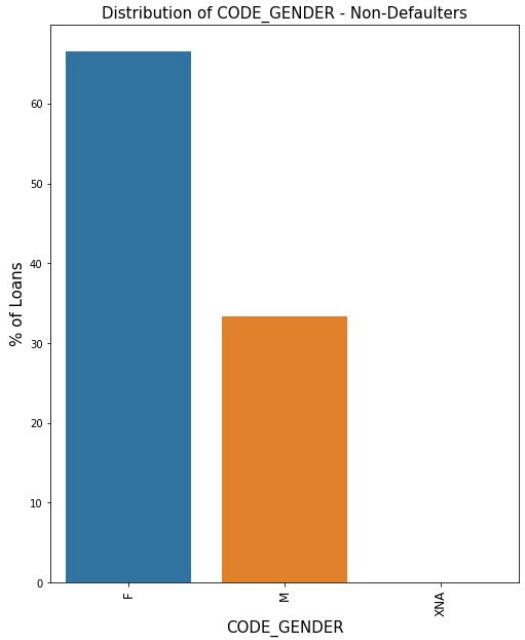
TARGET	FLAG_CONT_MOBILE	FLAG_DOCUMENT_10	FLAG_DOCUMENT_11
N	0.998128666	2.47625E-05	0.0039902
Y	0.998187311	0	0.0030211
DIFF	-5.86455E-05	2.47625E-05	0.0009691

FLAG_DOCUMENT_12	FLAG_DOCUMENT_13	FLAG_DOCUMENT_14	FLAG_DOCUMENT_15	FLAG_DOCUMENT_16
7.07499E-06	0.00	0.30882321727995 00	0.127703529711411 00	0.001026934
0	0.001208459	0.120845921450151 00	0.443101711983887 0	0.006042296
7.07499E-06	0.002520059	0.187977296	-0.315398182	-0.005015362

Index	YP	IN POPULATION REI	DAYS BIRTH	DAYS EMPLOYED	DAYS REGISTRATION	DAYS ID PUBLISH	OWN CAR AGE	OCCUP
0		0.018801	-9461	-637	-3648	-2120	nan	Labore
1		0.003541	-16765	-1188	-1186	-291	nan	Core s
2		0.010032	-19046	-225	-4260	-2531	26	Labore
3		0.008019	-19005	-3039	-9833	-2437	nan	Labore
4		0.028663	-19932	-3038	-4311	-3458	nan	Core s
5		0.035792	-16941	-1588	-4970	-477	nan	Labore
6		0.035792	-13778	-3130	-1213	-619	17	Accoun
7		0.003122	-18850	-449	-4597	-2379	8	Manage
8		0.018634	-20099	365243	-7427	-3514	nan	nan
9		0.019689	-14469	-2019	-14437	-3992	nan	Labore
10		0.0228	-10197	-679	-4427	-738	nan	Core s
11		0.015221	-20417	365243	-5246	-2512	nan	nan

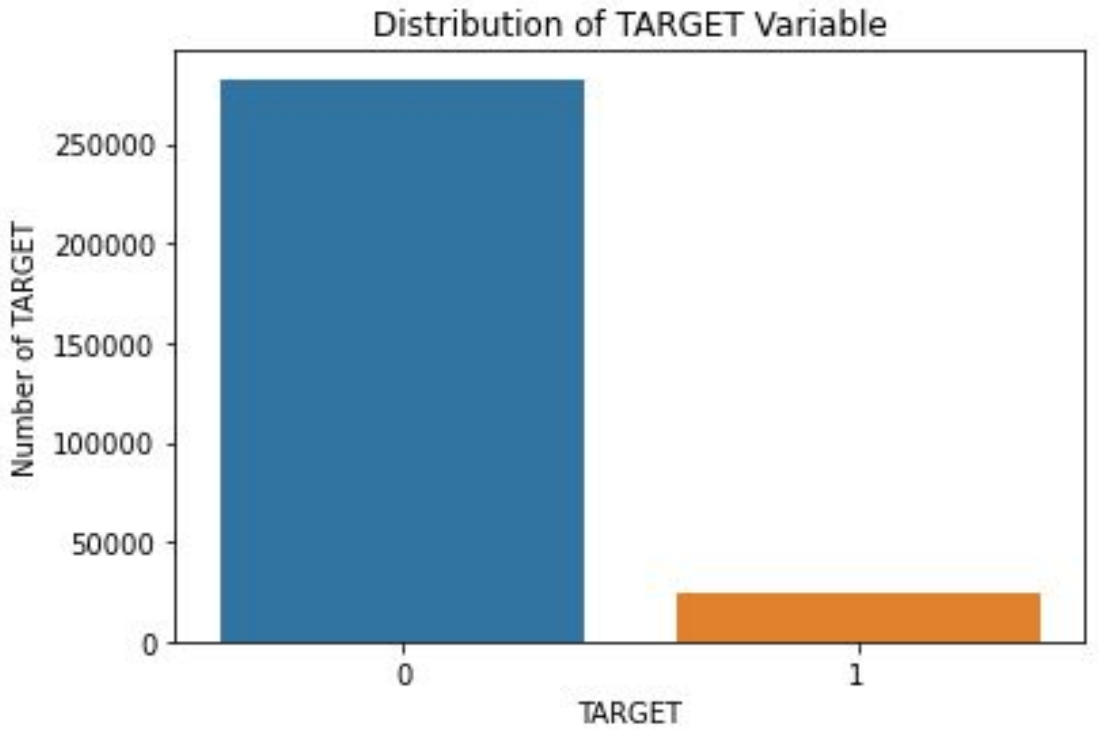


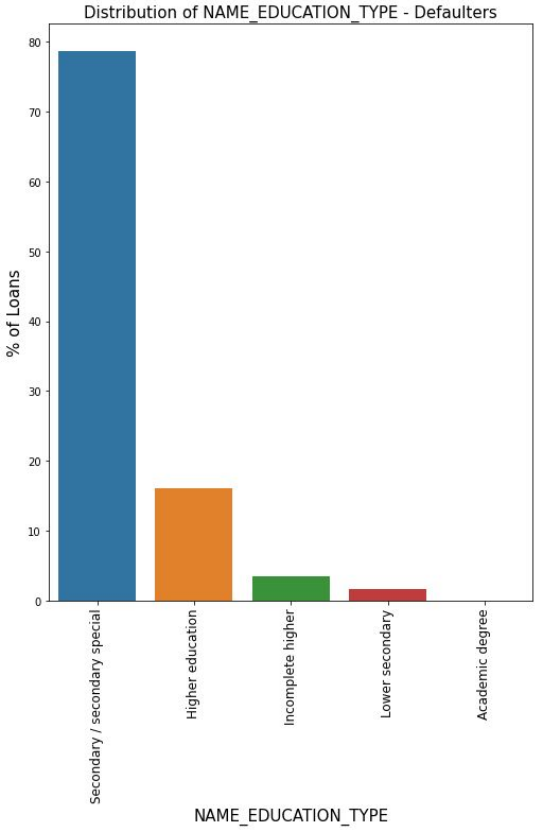


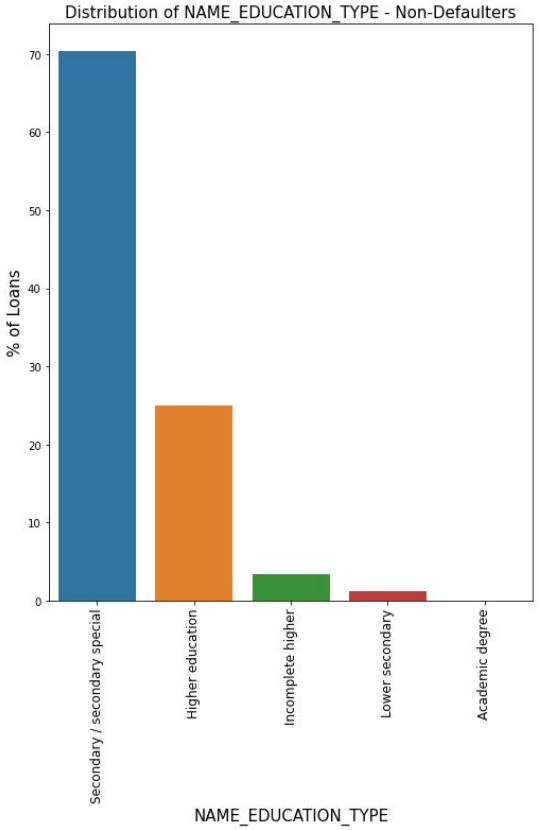


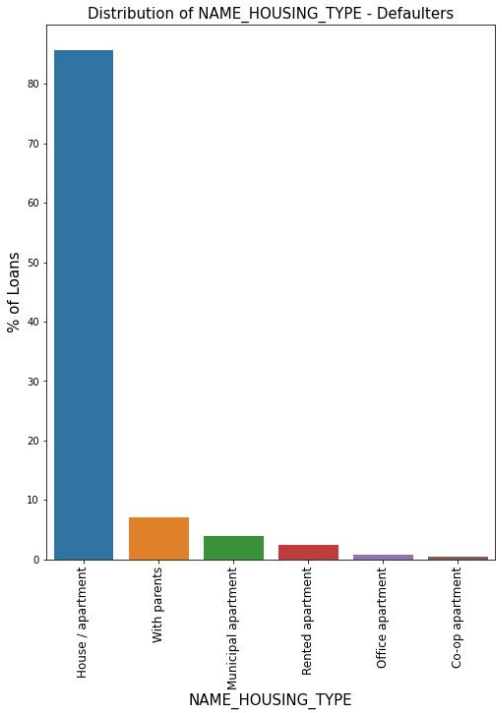
3. Comparación: Defaulters v non Defaulters

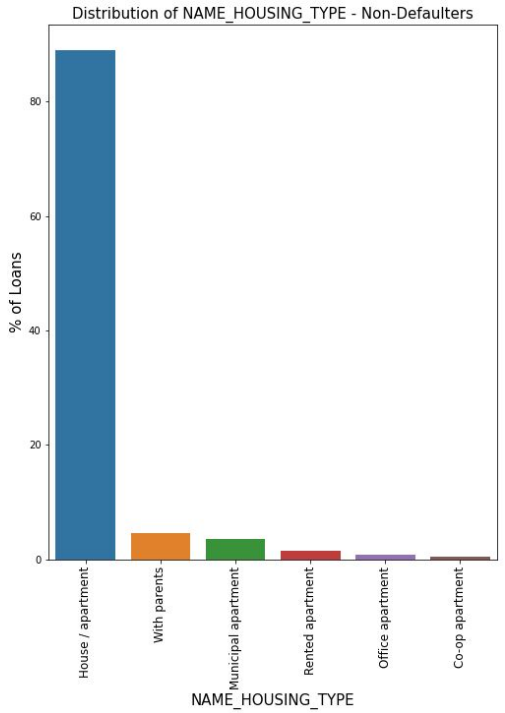
%8.07 Presenta problemas de pago

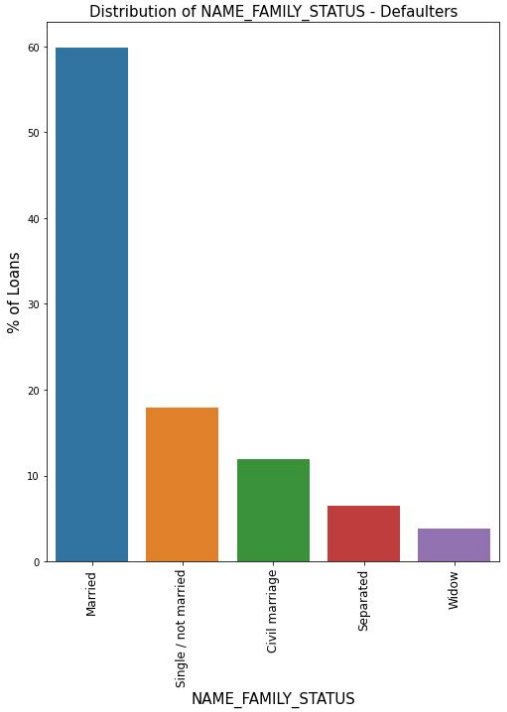


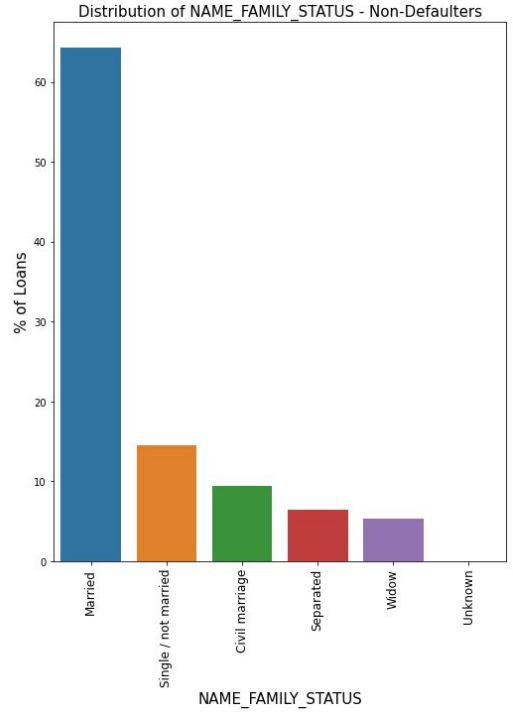

Distribution of NAME_EDUCATION_TYPE - Defaulters
% of Loans
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE

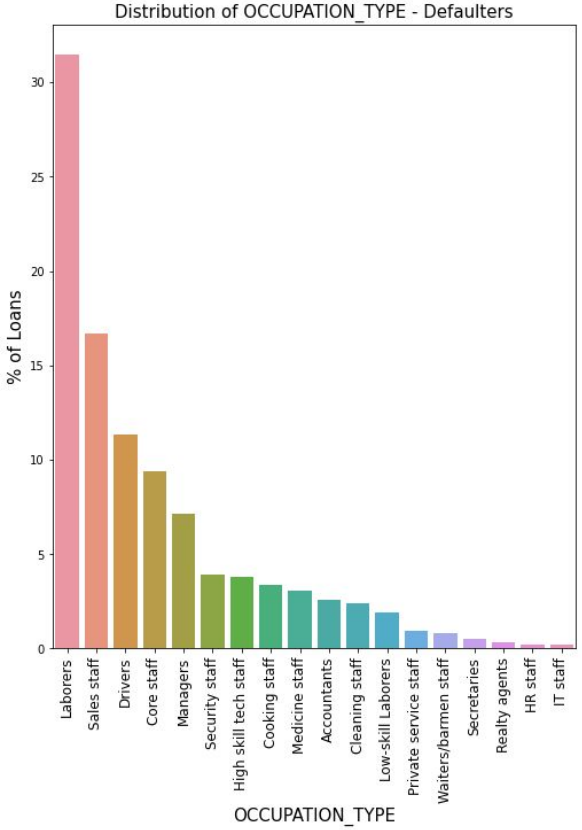

Distribution of NAME_EDUCATION_TYPE - Non-Defaulters
% of Loans
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE

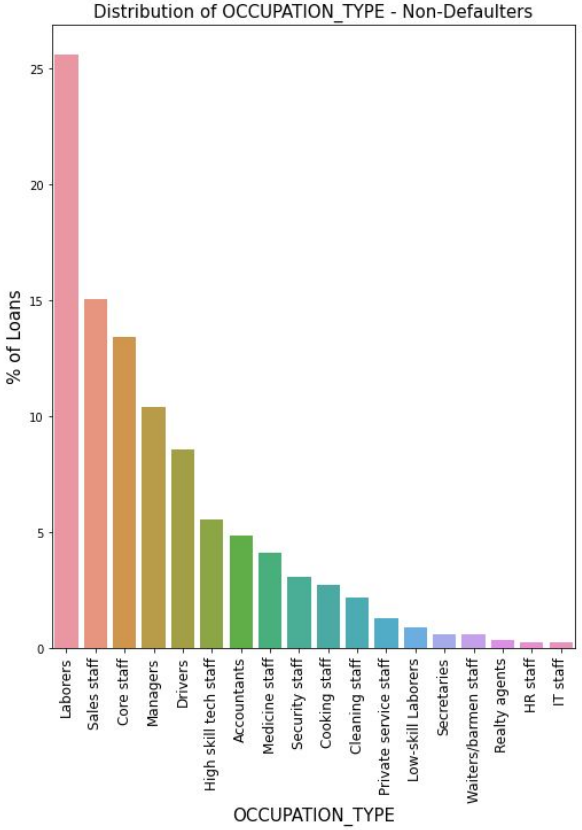

Distribution of NAME_HOUSING_TYPE - Defaulters
% of Loans
House / apartment
With parents
Municipal apartment
Rented apartment
Office apartment
Co-op apartment
NAME_HOUSING_TYPE

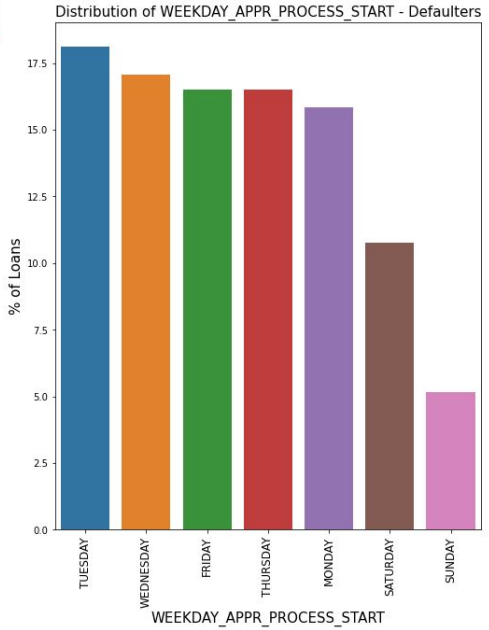

Distribution of NAME_HOUSING_TYPE - Non-Defaulters
% of Loans
House / apartment
With parents
Municipal apartment
Rented apartment
Office apartment
Co-op apartment
NAME_HOUSING_TYPE

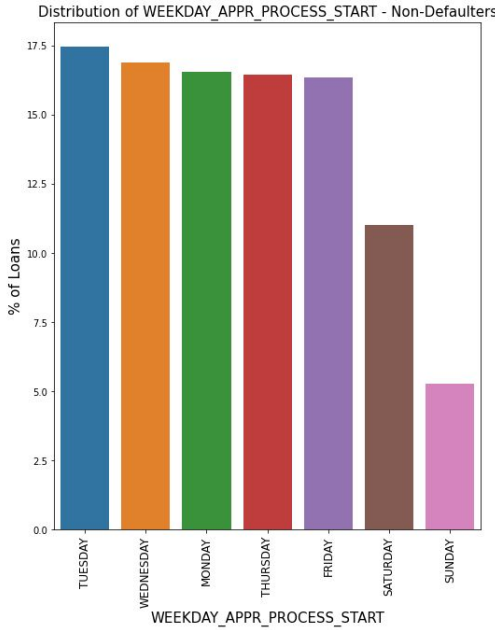

Distribution of NAME_FAMILY_STATUS - Defaulters
% of Loans
Married
Single / not married
Civil marriage
Separated
Widow
NAME_FAMILY_STATUS

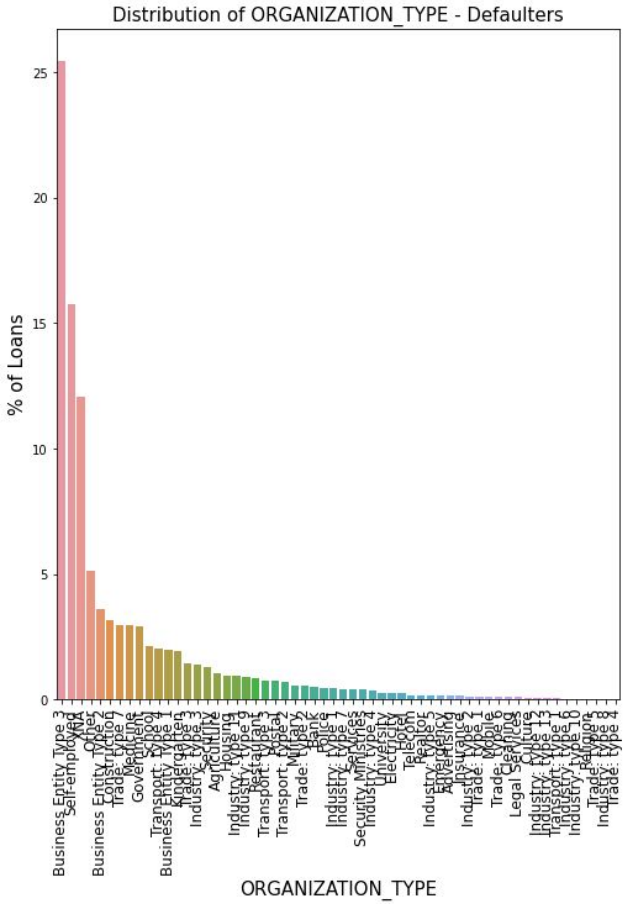

Distribution of NAME_FAMILY_STATUS - Non-Defaulters
% of Loans
Married
Single / not married
Civil marriage
Separated
Widow
Unknown
NAME_FAMILY_STATUS

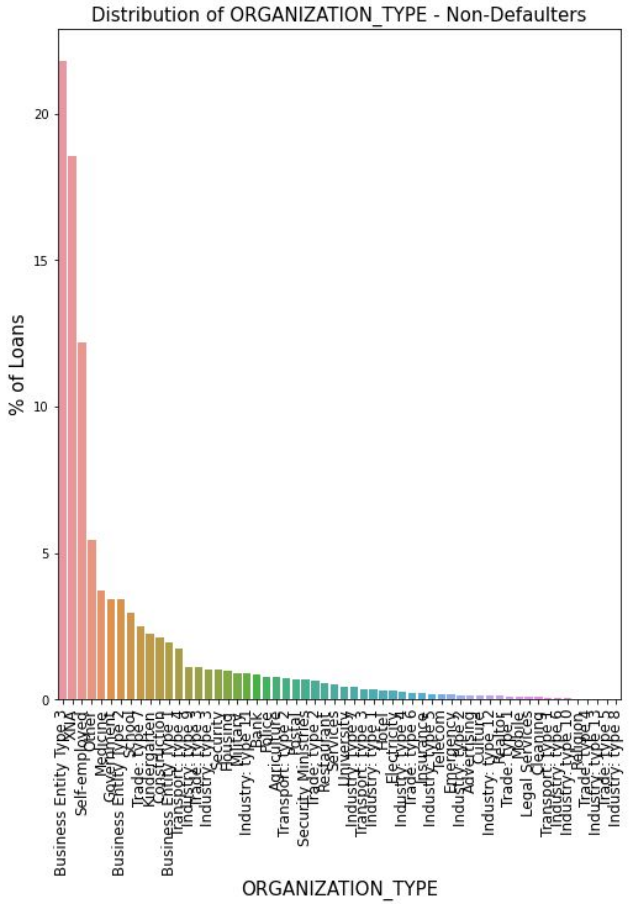

Distribution of OCCUPATION_TYPE - Defaulters
% of Loans
0
5
10
15
20
25
30
Laborers
Sales staff
Drivers
Core staff
Managers
Security staff
High skill tech staff
Cooking staff
Medicine staff
Accountants
Cleaning staff
Low-skill Laborers
Private service staff
Waiters/barmen staff
Secretaries
Realty agents
HR staff
IT staff
OCCUPATION_TYPE

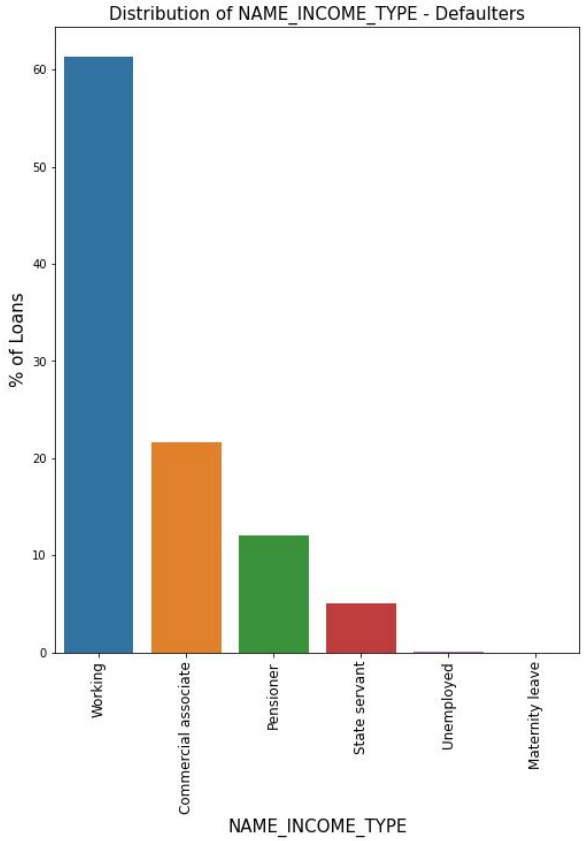

Distribution of OCCUPATION_TYPE - Non-Defaulters
% of Loans
0
5
10
15
20
25
Laborers
Sales staff
Core staff
Managers
Drivers
High skill tech staff
Accountants
Medicine staff
Security staff
Cooking staff
Cleaning staff
Private service staff
Low-skill Laborers
Secretaries
Waiters/barmen staff
Realty agents
HR staff
IT staff
OCCUPATION_TYPE

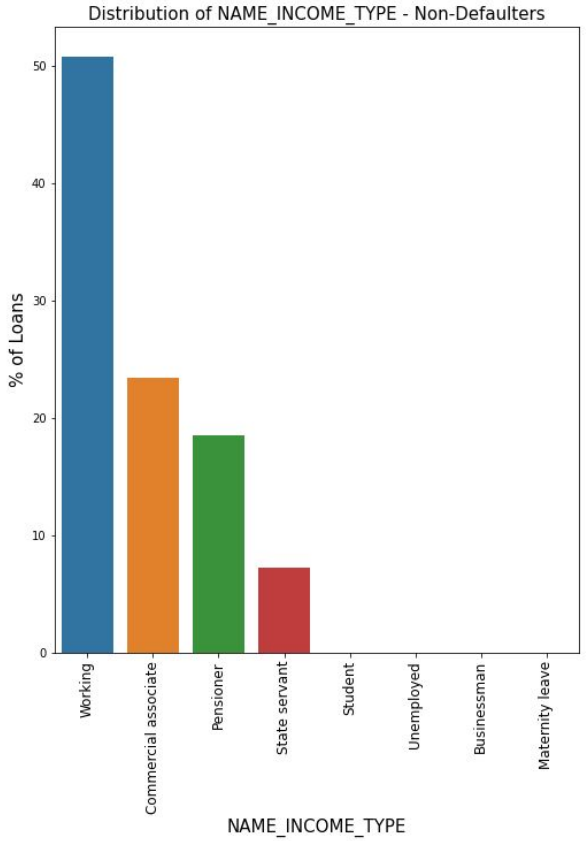

Distribution of WEEKDAY_APPR_PROCESS_START - Defaulters
% of Loans
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
TUESDAY
WEDNESDAY
FRIDAY
THURSDAY
MONDAY
SATURDAY
SUNDAY
WEEKDAY_APPR_PROCESS_START

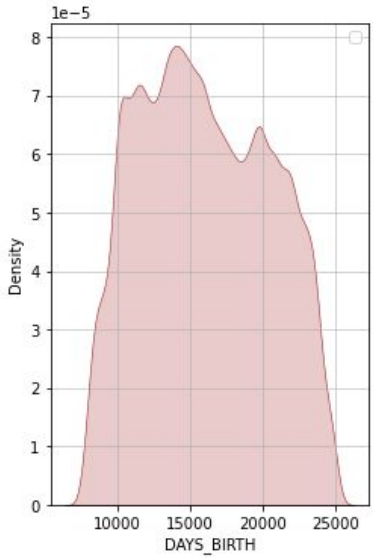

Distribution of WEEKDAY_APPR_PROCESS_START - Non-Defaulters
% of Loans
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
TUESDAY
WEDNESDAY
MONDAY
THURSDAY
FRIDAY
SATURDAY
SUNDAY
WEEKDAY_APPR_PROCESS_START

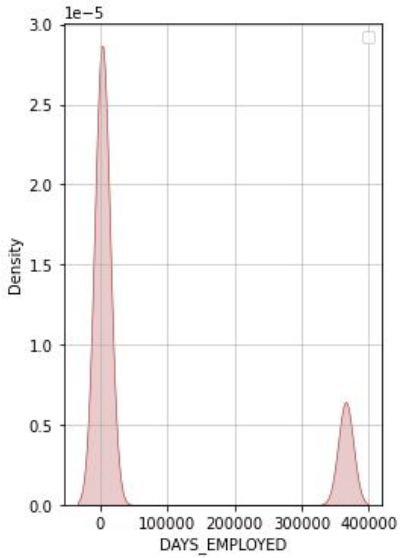

Distribution of ORGANIZATION_TYPE - Defaulters
% of Loans
ORGANIZATION_TYPE

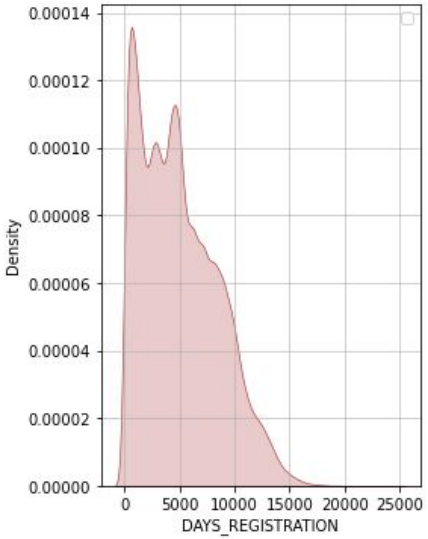

Distribution of ORGANIZATION_TYPE - Non-Defaulters
% of Loans
ORGANIZATION_TYPE

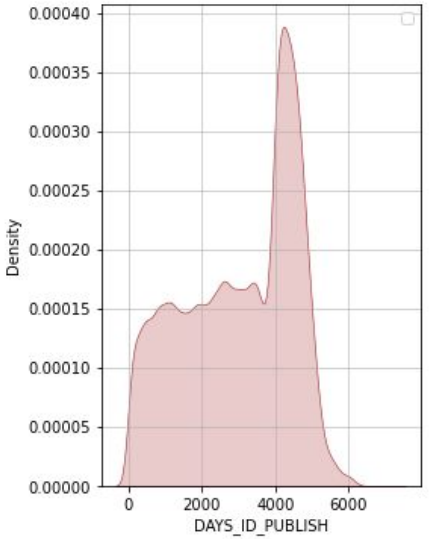

Distribution of NAME_INCOME_TYPE - Defaulters
% of Loans
Working
Commercial associate
Pensioner
State servant
Unemployed
Maternity leave
NAME_INCOME_TYPE

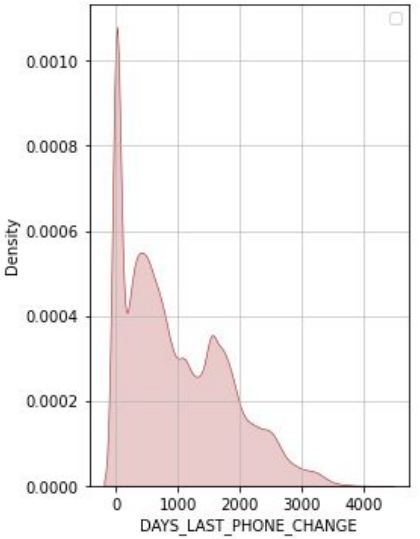

Distribution of NAME_INCOME_TYPE - Non-Defaulters
% of Loans
Working
Commercial associate
Pensioner
State servant
Student
Unemployed
Businessman
Maternity leave
NAME_INCOME_TYPE

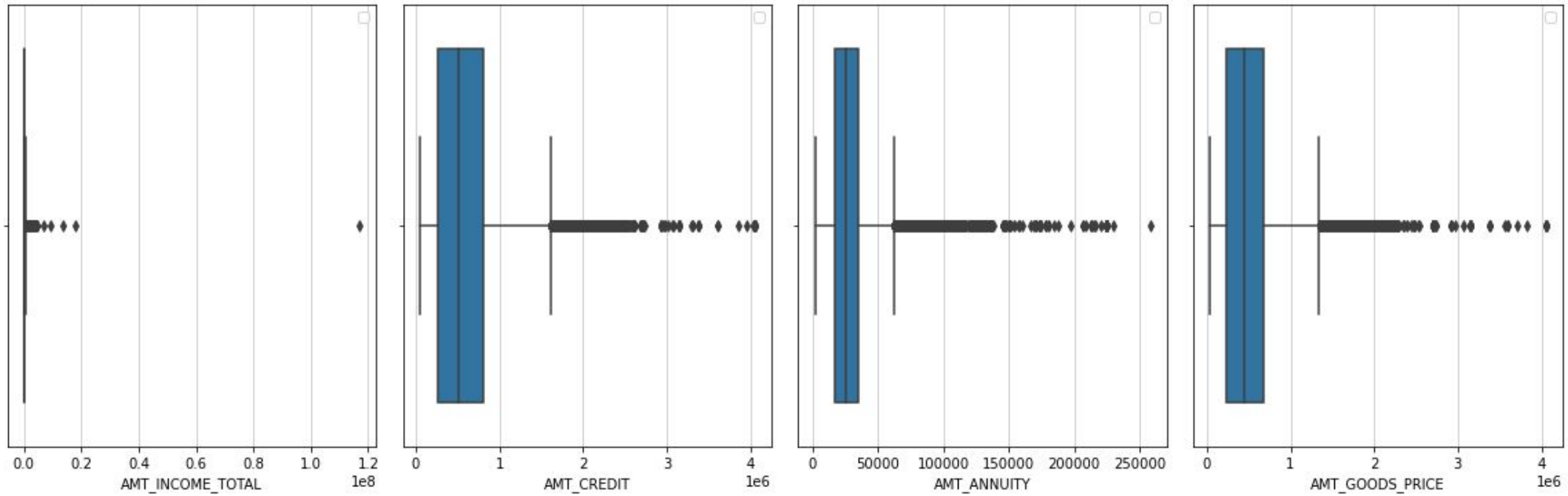

1e−5
Density
DAYS_BIRTH


1e−5
Density
DAYS_EMPLOYED

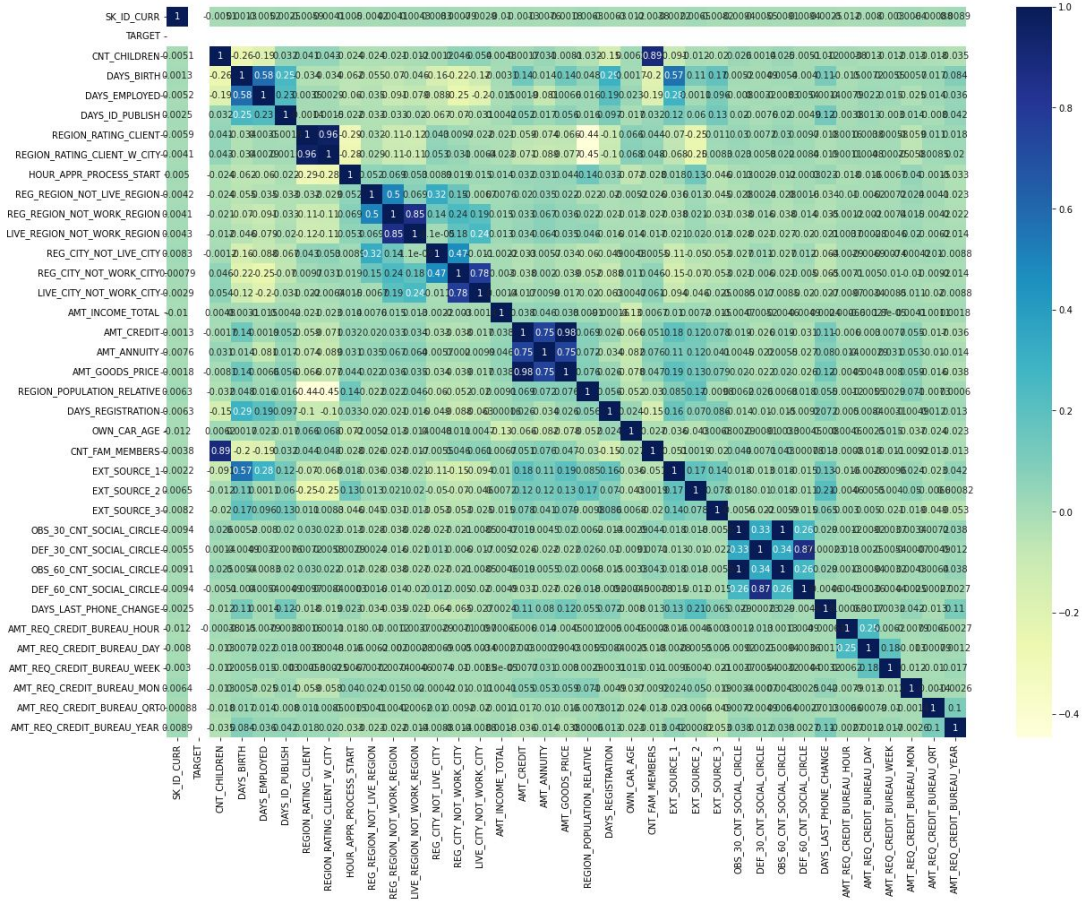

Density
DAYS_REGISTRATION


Density
DAYS_ID_PUBLISH

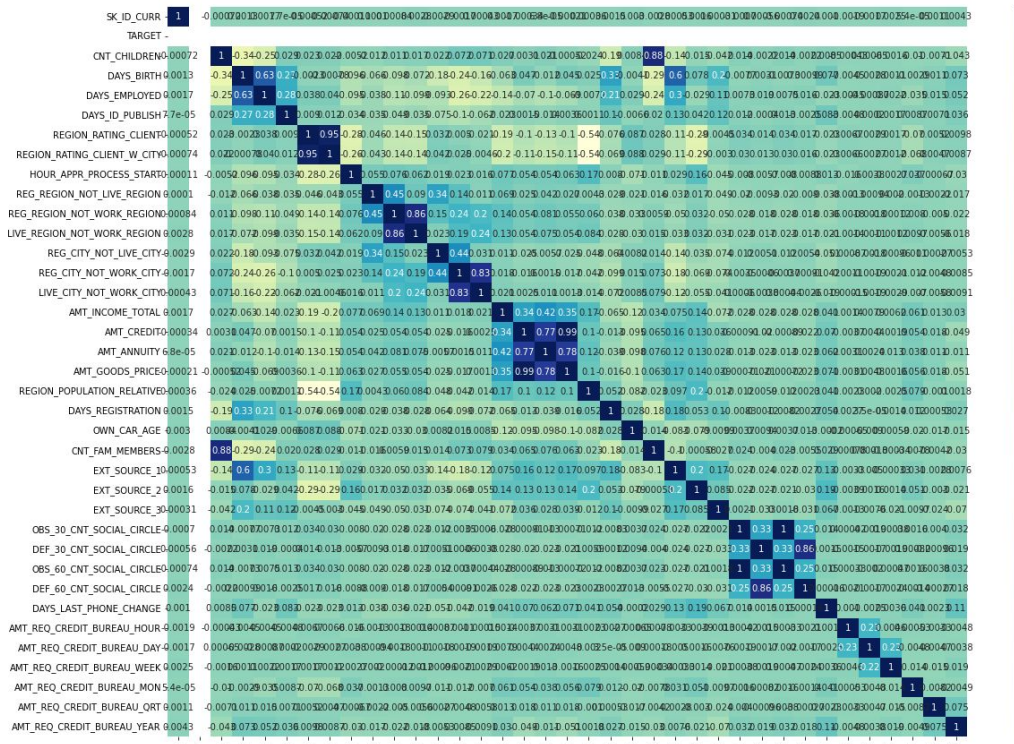

Density
DAYS_LAST_PHONE_CHANGE


0.0
0.2
0.4
0.6
0.8
1.0
1.2
AMT_INCOME_TOTAL
1e8
0
1
2
3
4
AMT_CREDIT
1e6
0
50000
100000
150000
200000
250000
AMT_ANNUITY
0
1
2
3
4
AMT_GOODS_PRICE
1e6

Defaulters

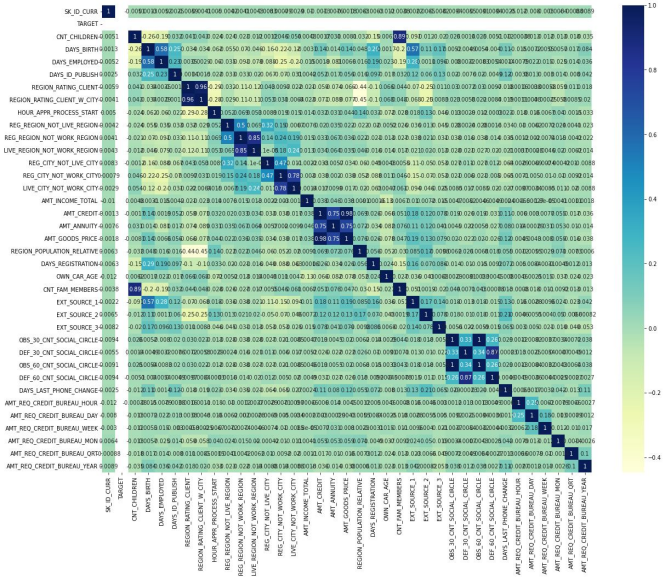


No Defaulters

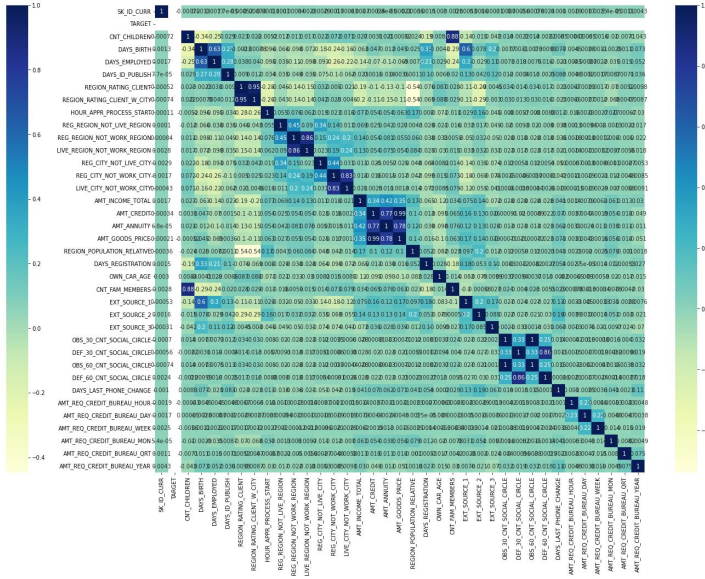


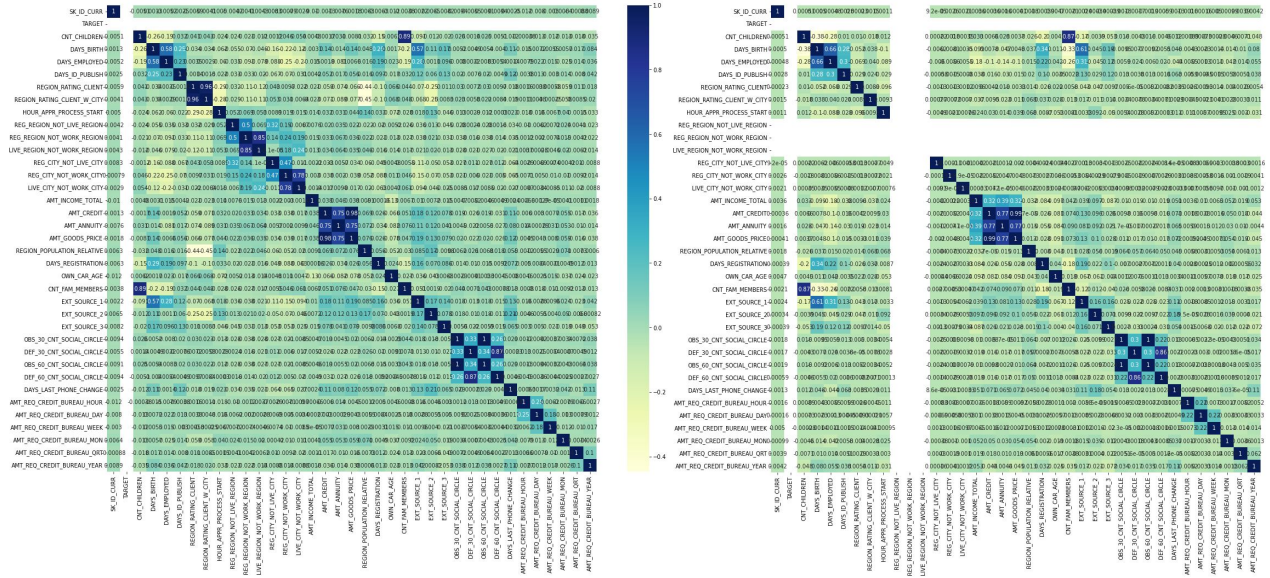
ITBA

Defaulters



No Defaulters





%36.81

Personas con problemas de pago eliminadas

CONCLUSIONES

- Es una base desbalanceada
- Mejorar limpieza de datos y outliers para no perder tantos datos
- Parece ser más sencillo identificar quien NO tendrá problemas con el pago