

GROUP 10 DATA PREPARATION AND
VISUALIZATION PROJECT

EXPLORATORY DATA ANALYSIS - HOME CREDIT DEFAULT RISK



application_{train|test}.csv

- This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET).
- Static data for all applications. One row represents one loan in our data sample.

bureau.csv

- All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample).
- For every loan in our sample, there are as many rows as number of credits the client had in Credit Bureau before the application date.

bureau_balance.csv

- This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET).
- Static data for all applications. One row represents one loan in our data sample.

credit_card_balance.csv

Monthly balance snapshots of previous credit cards that the applicant has with Home Credit. This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample

Brief Description of Each Table

Brief Description of Each Table

POS_CASH_balance.csv

- Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit.
- This table has one row for each month of history of every previous credit in Home Credit related to loans in our sample

previous_application.csv

- All previous applications for Home Credit loans of clients who have loans in our sample.
- There is one row for each previous application related to loans in our data sample.

installments_payments.csv

- Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample.

HomeCredit_columns_description.csv

- This file contains descriptions for the columns in the various data files.

application_{train|test}.csv

The application_train.csv table consists of static data relating to the Borrowers with labels. Each row represents one loan application.

The application_test.csv contains the testing dataset, and is similar to application_train.csv, except that the TARGET column has been omitted, which has to be predicted with the help of Statistical and Machine Learning Predictive Models.

	id	NAME_CONTRACT	STATUS
	100001	Cash loans	
	100005	Cash loans	
4	100013	Cash loans	M
5	100028	Cash loans	F
6	100038	Cash loans	M
7	100042	Cash loans	F
8	100057	Cash loans	M
9	100065	Cash loans	M
10	100066	Cash loans	F
11	100067	Cash loans	F
12	100074	Cash loans	F
13	100090	Cash loans	F
	100091	Cash loans	
	100092	Cash loans	

Application

app_train.head(10)

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT
0	100002	1	Cash loans	M	N	Y	0	202500.0	406597.0
1	100003	0	Cash loans	F	N	N	0	270000.0	1293502.0
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000.0
3	100006	0	Cash loans	F	N	Y	0	135000.0	312682.0
4	100007	0	Cash loans	M	N	Y	0	121500.0	513000.0
5	100008	0	Cash loans	M	N	Y	0	99000.0	490495.0
6	100009	0	Cash loans	F	Y	Y	1	171000.0	1560726.0
7	100010	0	Cash loans	M	Y	Y	0	360000.0	1530000.0
8	100011	0	Cash loans	F	N	Y	0	112500.0	1019610.0
9	100012	0	Revolving loans	M	N	Y	0	135000.0	405000.0

```
1 # Overview of training set
2 print(f"Size of training set: {app_train.shape} \n")
```

Size of training set: (307511, 122)

app_test.head(10)

	SK_ID_CURR	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_DOWN_PAYMENT
0	100001	Cash loans	F	N	Y	0	135000.0	568800.0	0.0
1	100005	Cash loans	M	N	Y	0	99000.0	222768.0	0.0
2	100013	Cash loans	M	Y	Y	0	202500.0	663264.0	0.0
3	100028	Cash loans	F	N	Y	2	315000.0	1575000.0	0.0
4	100038	Cash loans	M	Y	N	1	180000.0	625500.0	0.0
5	100042	Cash loans	F	Y	Y	0	270000.0	959688.0	0.0
6	100057	Cash loans	M	Y	Y	2	180000.0	499221.0	0.0
7	100065	Cash loans	M	N	Y	0	166500.0	180000.0	0.0
8	100066	Cash loans	F	N	Y	0	315000.0	364896.0	0.0
9	100067	Cash loans	F	Y	Y	1	162000.0	45000.0	0.0

```
1 # Overview of test set
2 print(f"Size of test set: {app_test.shape} \n")
```

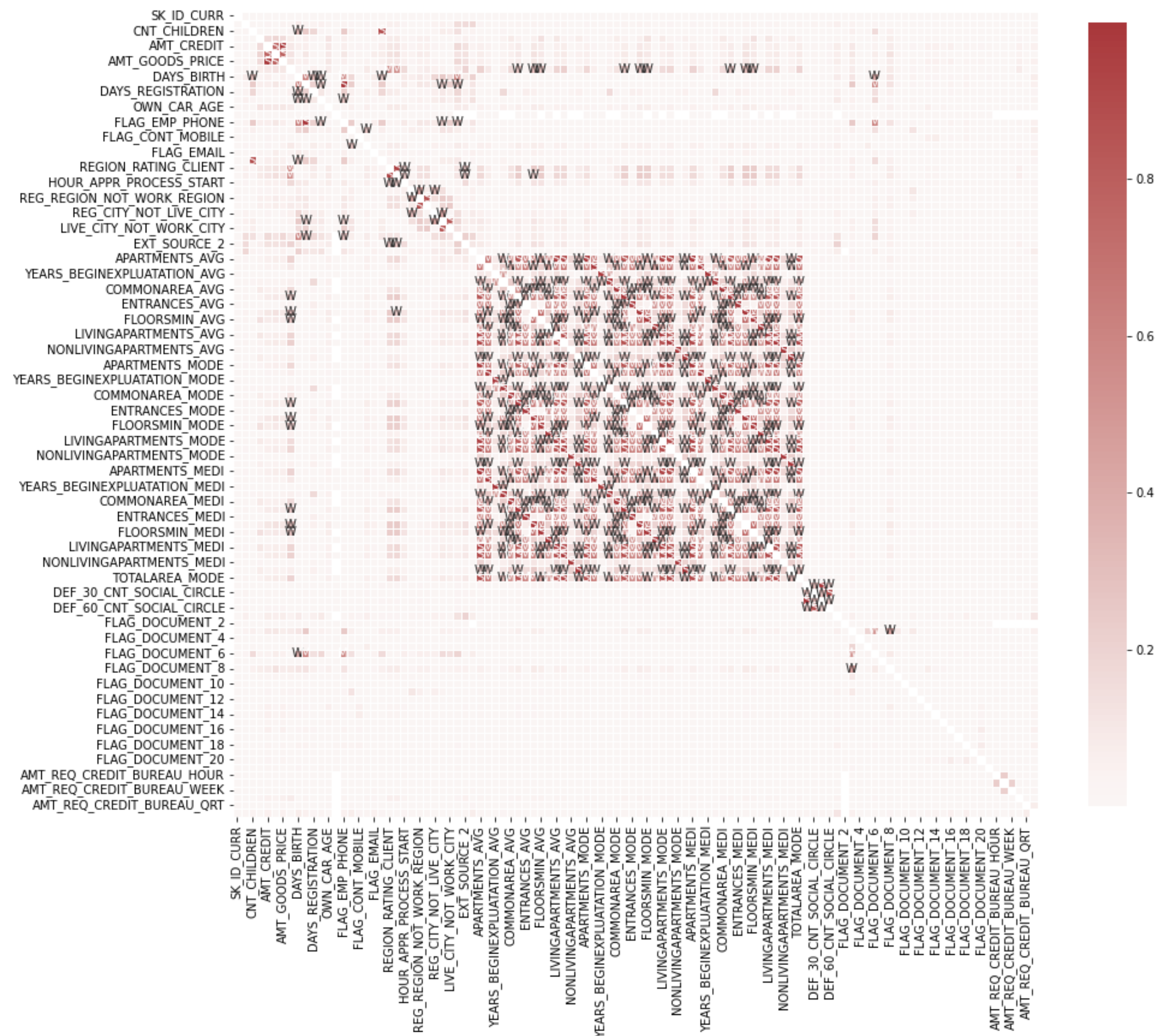
Size of test set: (48744, 121)

Structure Investigation

1	app_train.describe().T								
		count	mean	std	min	25%	50%	75%	max
	SK_ID_CURR	307511.0	278180.518577	102790.175348	100002.0	189145.5	278202.0	367142.5	456255.0
	TARGET	307511.0	0.080729	0.272419	0.0	0.0	0.0	0.0	1.0
	CNT_CHILDREN	307511.0	0.417052	0.722121	0.0	0.0	0.0	1.0	19.0
	AMT_INCOME_TOTAL	307511.0	168797.921875	237123.140625	25650.0	112500.0	147150.0	202500.0	117000000.0
	AMT_CREDIT	307511.0	599025.937500	402490.781250	45000.0	270000.0	513531.0	808650.0	4050000.0

	AMT_REQ_CREDIT_BUREAU_DAY	265992.0	0.000000	0.000000	0.0	0.0	0.0	0.0	9.0
	AMT_REQ_CREDIT_BUREAU_WEEK	265992.0	0.000000	0.000000	0.0	0.0	0.0	0.0	8.0
	AMT_REQ_CREDIT_BUREAU_MON	265992.0	NaN	0.000000	0.0	0.0	0.0	0.0	27.0
	AMT_REQ_CREDIT_BUREAU_QRT	265992.0	NaN	NaN	0.0	0.0	0.0	0.0	261.0
	AMT_REQ_CREDIT_BUREAU_YEAR	265992.0	NaN	0.000000	0.0	0.0	1.0	3.0	25.0

Structure Investigation



	Total	Percent
COMMONAREA_AVG	214865	69.9
NONLIVINGAPARTMENTS_AVG	213514	69.4
FONDKAPREMONT_MODE	210295	68.4
FLOORSMIN_AVG	208642	67.8
YEARS_BUILD_AVG	204488	66.5
OWN_CAR_AGE	202929	66.0
LANDAREA_AVG	182590	59.4
BASEMENTAREA_AVG	179943	58.5
EXT_SOURCE_1	173378	56.4
NONLIVINGAREA_AVG	169682	55.2
WALLSMATERIAL_MODE	156341	50.8
APARTMENTS_AVG	156061	50.7
ENTRANCES_AVG	154828	50.3
HOUSETYPE_MODE	154297	50.2
FLOORSMAX_AVG	153020	49.8

	Total	Percent
YEARS_BEGINEXPLUATATION_AVG	150007	48.8
EMERGENCYSTATE_MODE	145755	47.4
OCCUPATION_TYPE	96391	31.3
EXT_SOURCE_3	60965	19.8
AMT_REQ_CREDIT_BUREAU_DAY	41519	13.5
AMT_REQ_CREDIT_BUREAU_QRT	41519	13.5
AMT_REQ_CREDIT_BUREAU_HOUR	41519	13.5
AMT_REQ_CREDIT_BUREAU_WEEK	41519	13.5
AMT_REQ_CREDIT_BUREAU_MON	41519	13.5
AMT_REQ_CREDIT_BUREAU_YEAR	41519	13.5
NAME_TYPE_SUITE	1292	0.4
DEF_30_CNT_SOCIAL_CIRCLE	1021	0.3
OBS_30_CNT_SOCIAL_CIRCLE	1021	0.3
EXT_SOURCE_2	660	0.2

Variables interpretation

Column	Description
SK_ID_CURR	ID of loan in our sample
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 installments of the loan in our sample, 0 - all other cases)
NAME_CONTRACT_TYPE	Identification if loan is cash or revolving
CODE_GENDER	Gender of the client
FLAG_OWN_CAR	Flag if the client owns a car
FLAG_OWN_REALTY	Flag if client owns a house or flat
CNT_CHILDREN	Number of children the client has
AMT_CREDIT	Credit amount of the loan

Variables interpretation

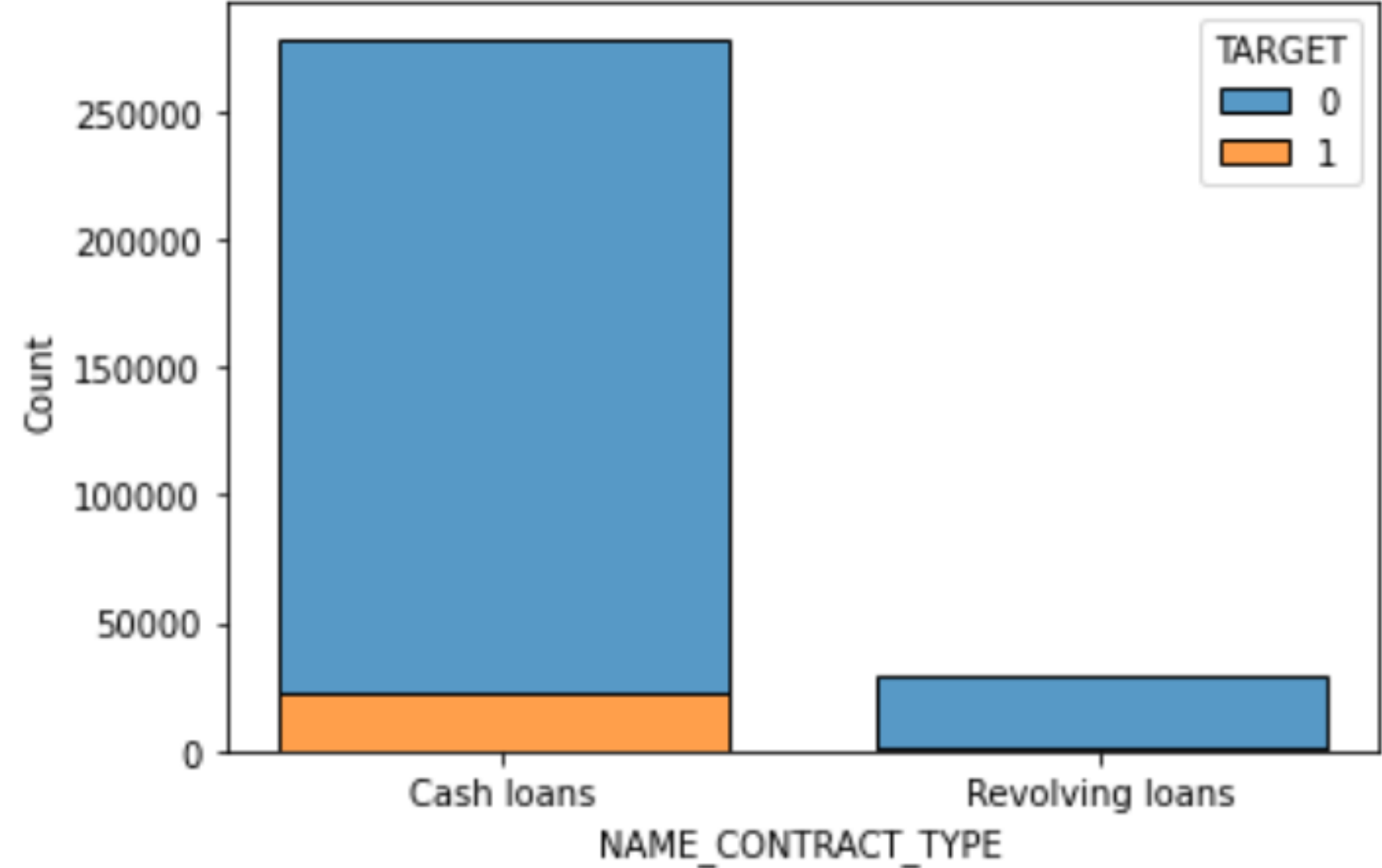
Column	Description
NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan
NAME_INCOME_TYPE	Clients income type (businessman, working, maternity leave,...)
NAME_FAMILY_STATUS	Family status of the client
NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents, ...)
DAYS_BIRTH	Client's age in days at the time of application
OCCUPATION_TYPE	What kind of occupation does the client have
EXT_SOURCE_1	Normalized score from external data source
EXT_SOURCE_2	Normalized score from external data source
EXT_SOURCE_3	Normalized score from external data source

Univariate Analysis

Name_Contract_Type

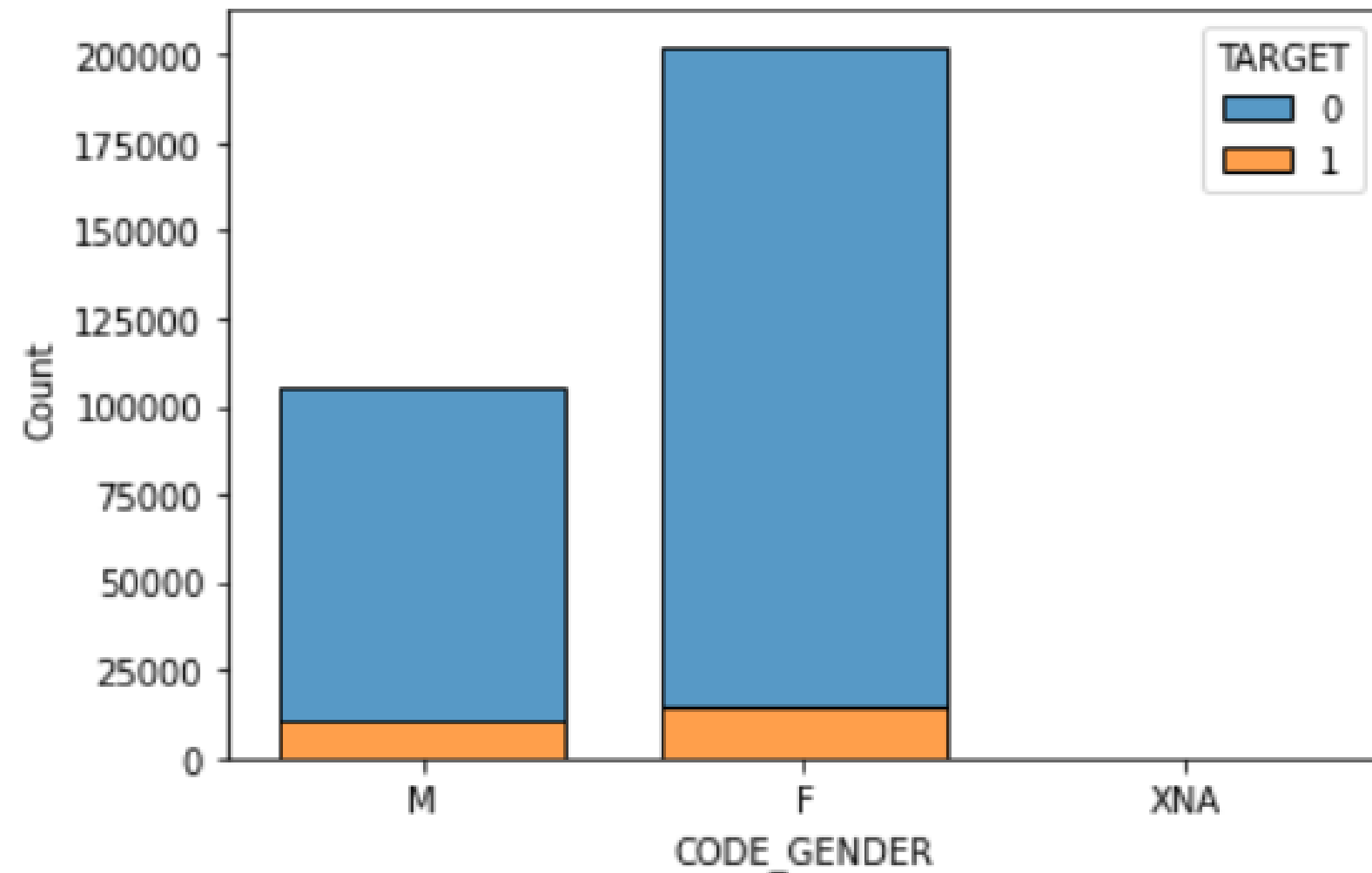
Most of the people are taking loans in the form of cash loans instead of revolving loans such as credit cards. The percentage of repaying the loan of 2 types are nearly the same

	NAME_CONTRACT_TYPE	TARGET	Total	Avg
0	Cash loans	23221	278232	0.083459
1	Revolving loans	1604	29279	0.054783



Univariate Analysis

	CODE_GENDER	TARGET	Total	Avg
0	F	14170	202448	0.069993
1	M	10655	105059	0.101419
2	XNA	0	4	0.000000



Code_Gender

Women took much more number of loans as compared to Men : Whereas Women took a total of 202K+ loans, Men only took 105K+ loans. However, at the same time, Men are slightly more capable of repaying the loan as compared to Women. Whereas Men are able to repay their loans in 10% of the cases, Women are only able to repay in 7% of the cases. There are 4 entries where Gender='XNA'. Since this is not providing us with much information, we can remove these entries later on.

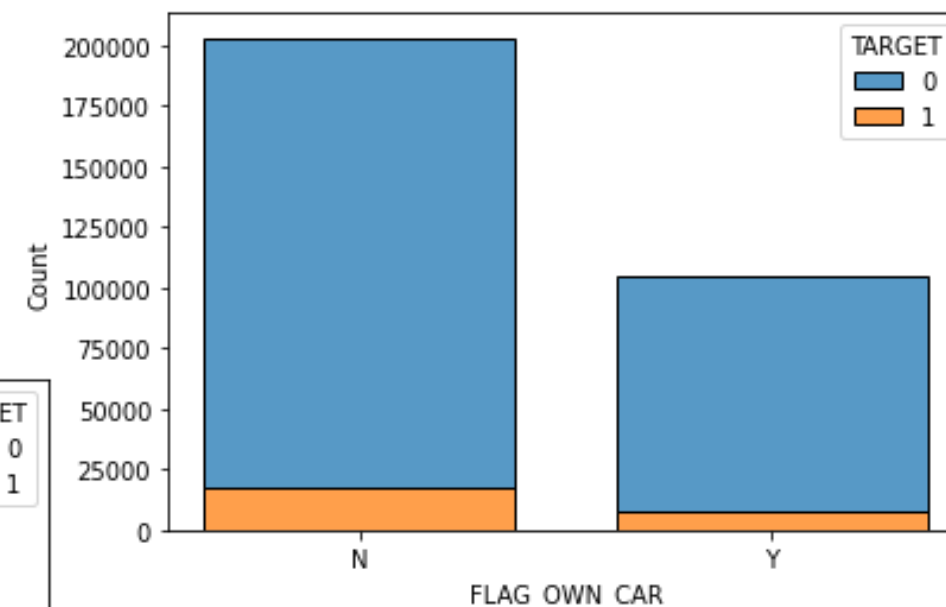
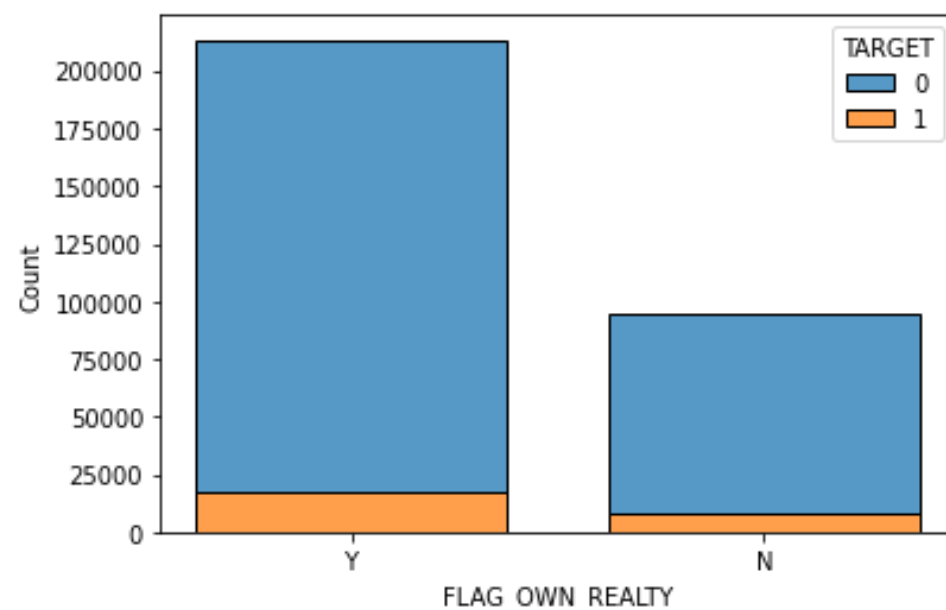
Univariate Analysis

	FLAG_OWN_CAR	TARGET	Total	Avg
0	N	17249	202924	0.085002
1	Y	7576	104587	0.072437

	FLAG_OWN_REALTY	TARGET	Total	Avg
0	N	7842	94199	0.083249
1	Y	16983	213312	0.079616

Flag_Own_Car and Flag_Own_Realty

Most of the applicants for loans own a flat/house, which is a little surprising. However, again, there is not much difference in the loan repayment status for the customer based on this information (7.9% and 8.3% respectively). We can conclude that this feature is not very useful.

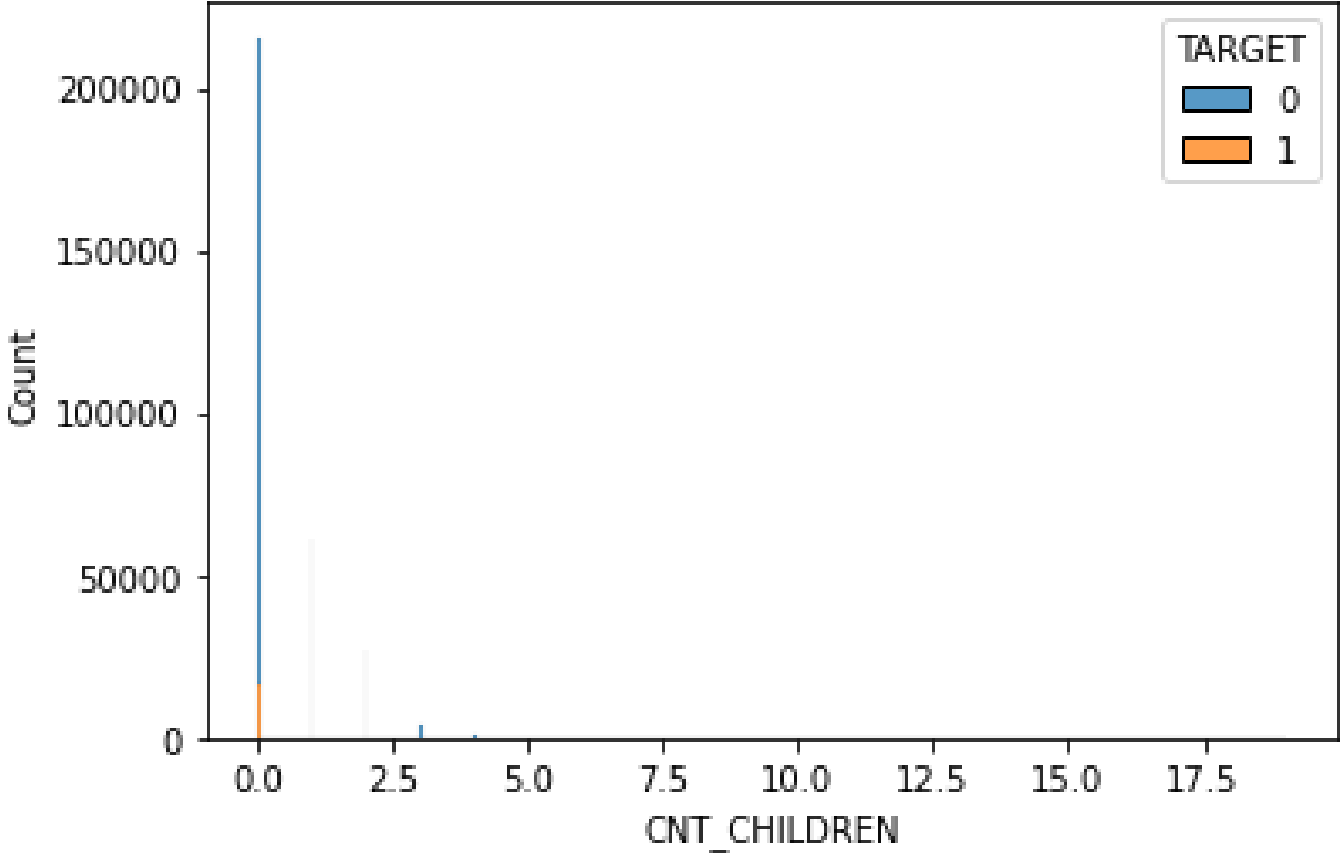


Univariate Analysis

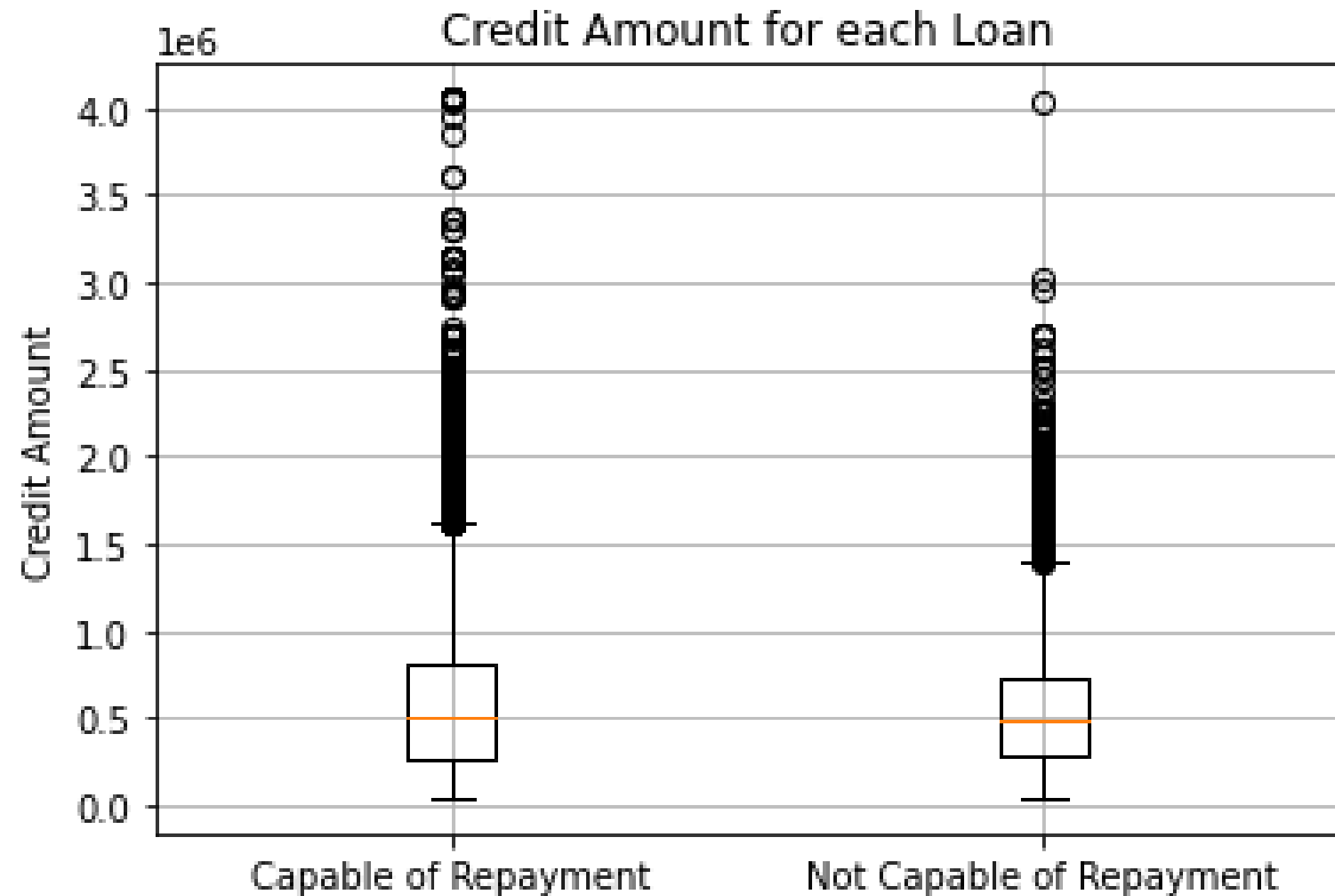
Cnt_Children

The applicants having no children take considerably higher number of loans. However, again, there is not much difference in the loan repayment status for the customer based on this information. We can conclude that this feature is not very useful.

CNT_CHILDREN	TARGET	Total	Avg	
0	0	16609	215371	0.077118
1	1	5454	61119	0.089236
2	2	2333	26749	0.087218
3	3	358	3717	0.096314
4	4	55	429	0.128205
5	5	7	84	0.083333
6	6	6	21	0.285714
7	7	0	7	0.000000
8	8	0	2	0.000000
9	9	2	2	1.000000
10	10	0	2	0.000000
11	11	1	1	1.000000
12	12	0	2	0.000000
13	14	0	3	0.000000
14	19	0	2	0.000000



Univariate Analysis



Amt_Credit

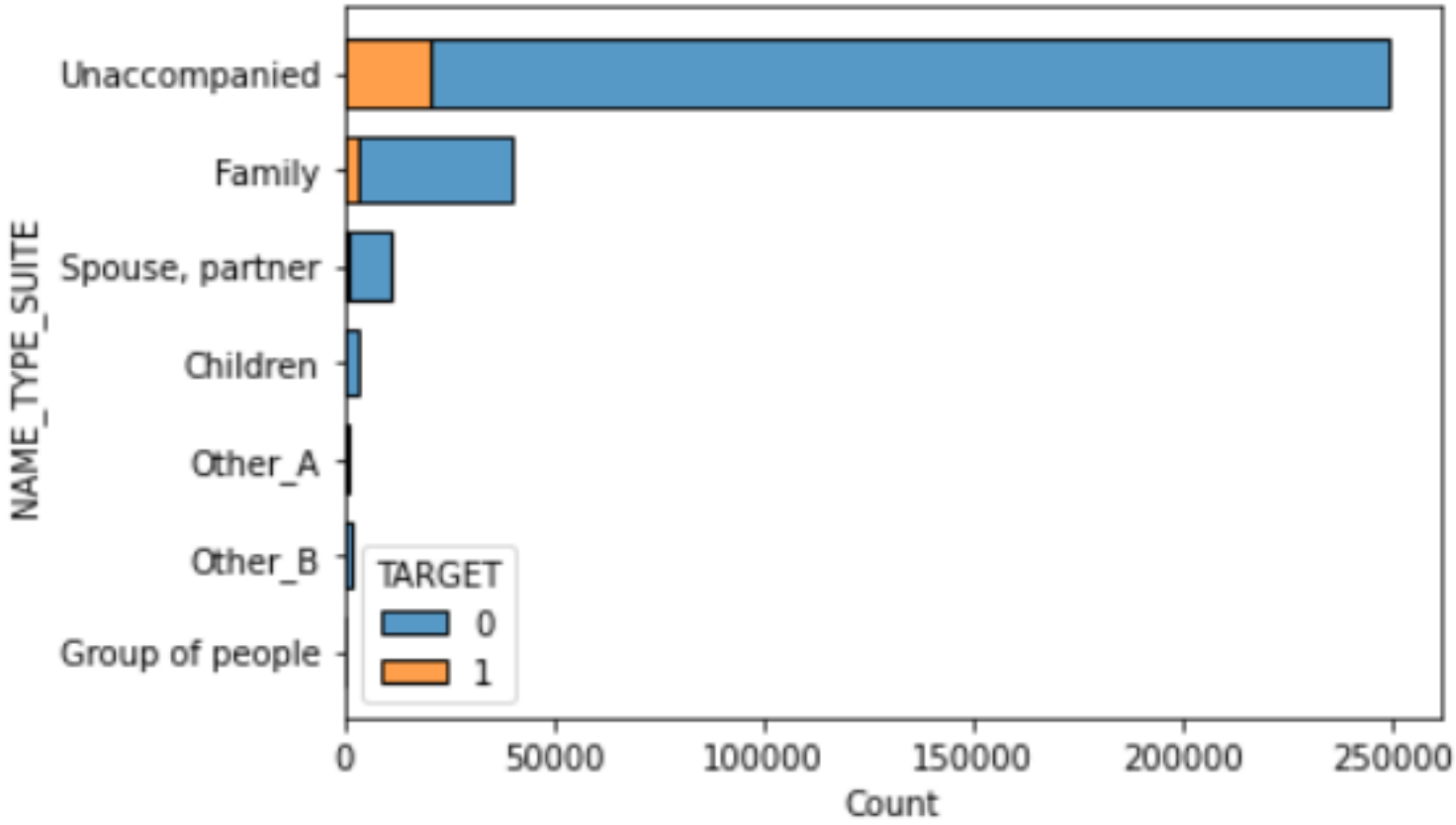
We can see from the Boxplot that the Median Value of the Credit Amount of the Customers who are capable of loan repayment is slightly larger than the Median Value of Customers who are not capable of repayment. This basically means that the customers with higher credit amount have a slightly higher chances of being capable of loan repayment than customers with lower credit amount.

Name_Type_Suite

For the various types of people accompanying the client for loan, the client comes unaccompanied to the bank in the most number of cases, out of which approx. 92% of the time, the bank finds the client to be capable of loan repayment whereas the remaining 8% of the time, the client is not capable of the same. Both in capability and non capability, 'Unaccompanied' as a class is the majority class in this case. The curve over here falls very sharply, which means that there is a lot of variability.

Univariate Analysis

	NAME_TYPE_SUITE	TARGET	Total	Avg
0	Children	241	3267	0.073768
1	Family	3009	40149	0.074946
2	Group of people	23	271	0.084871
3	Other_A	76	866	0.087760
4	Other_B	174	1770	0.098305
5	Spouse, partner	895	11370	0.078716
6	Unaccompanied	20337	248526	0.081830

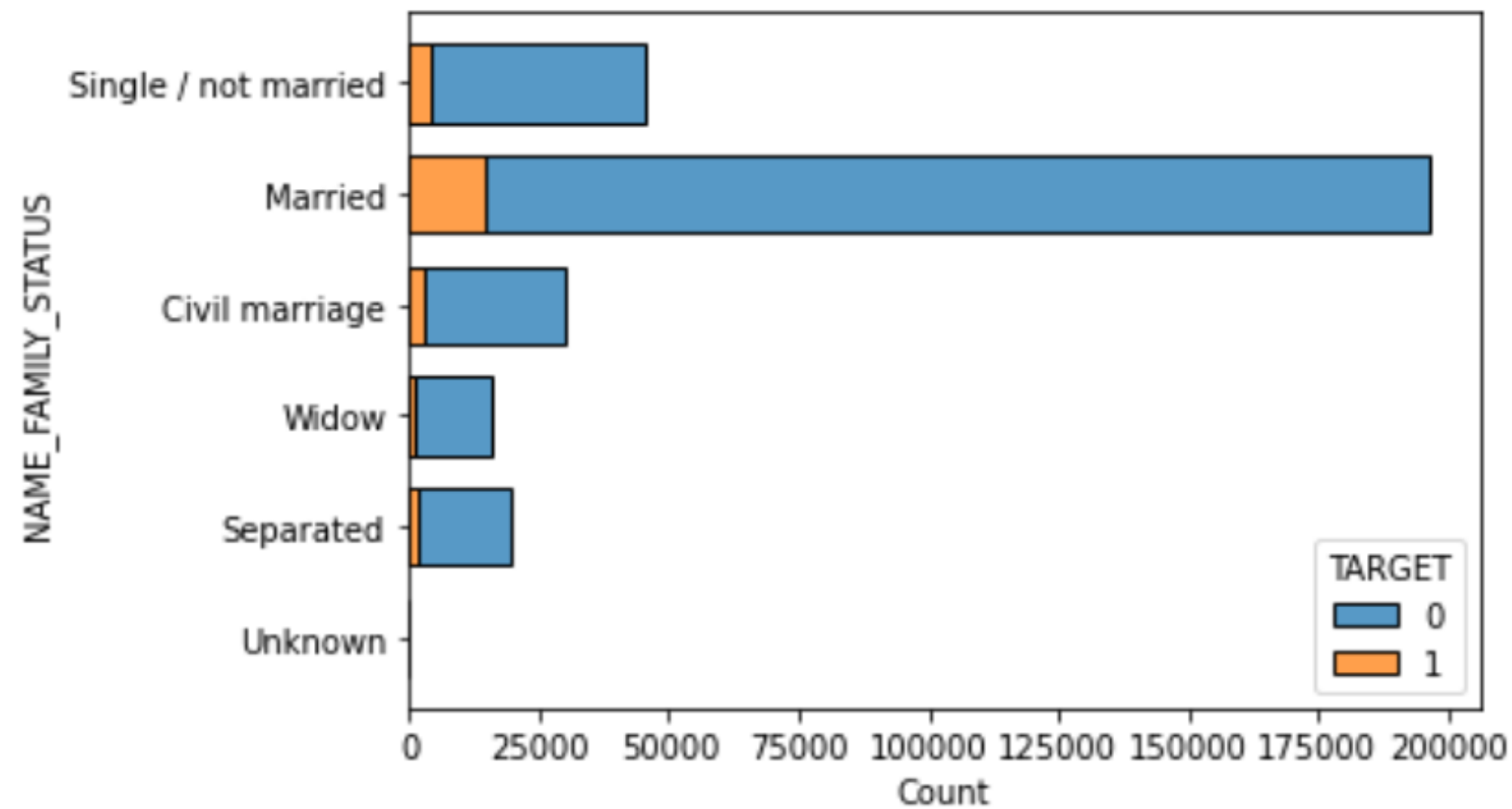


Univariate Analysis

Name_Family_Status

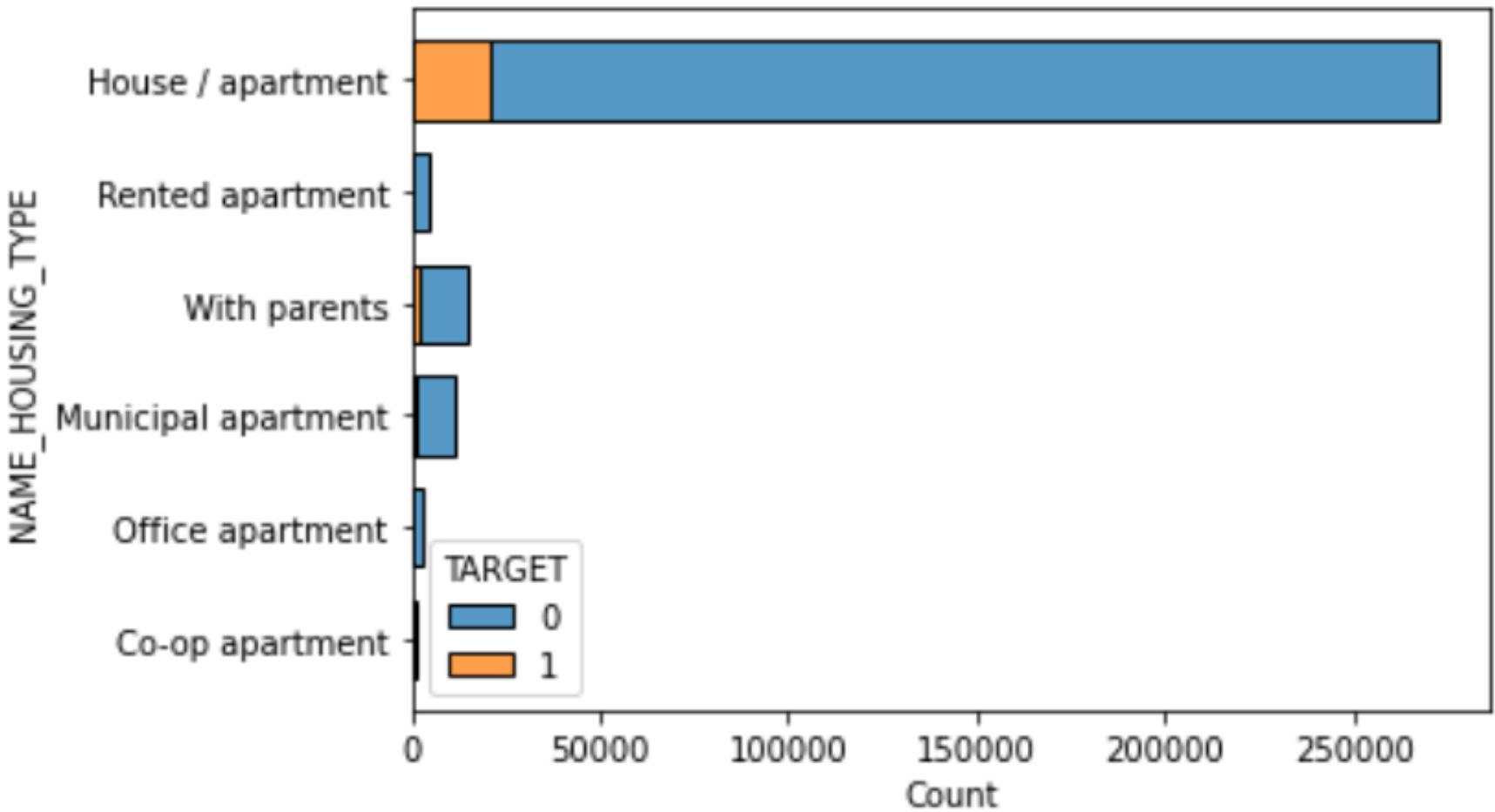
There is variability among the Family Status of the applicants but there is not much variability if the majority class (Married) is ignored. Married people apply for the most number of loans and the number of people deemed incapable of repayment is also the highest.

	NAME_FAMILY_STATUS	TARGET	Total	Avg
0	Civil marriage	2961	29775	0.099446
1	Married	14850	196432	0.075599
2	Separated	1620	19770	0.081942
3	Single / not married	4457	45444	0.098077
4	Unknown	0	2	0.000000
5	Widow	937	16088	0.058242



Univariate Analysis

	NAME_HOUSING_TYPE	TARGET	Total	Avg
0	Co-op apartment	89	1122	0.079323
1	House / apartment	21272	272868	0.077957
2	Municipal apartment	955	11183	0.085397
3	Office apartment	172	2617	0.065724
4	Rented apartment	601	4881	0.123131
5	With parents	1736	14840	0.116981

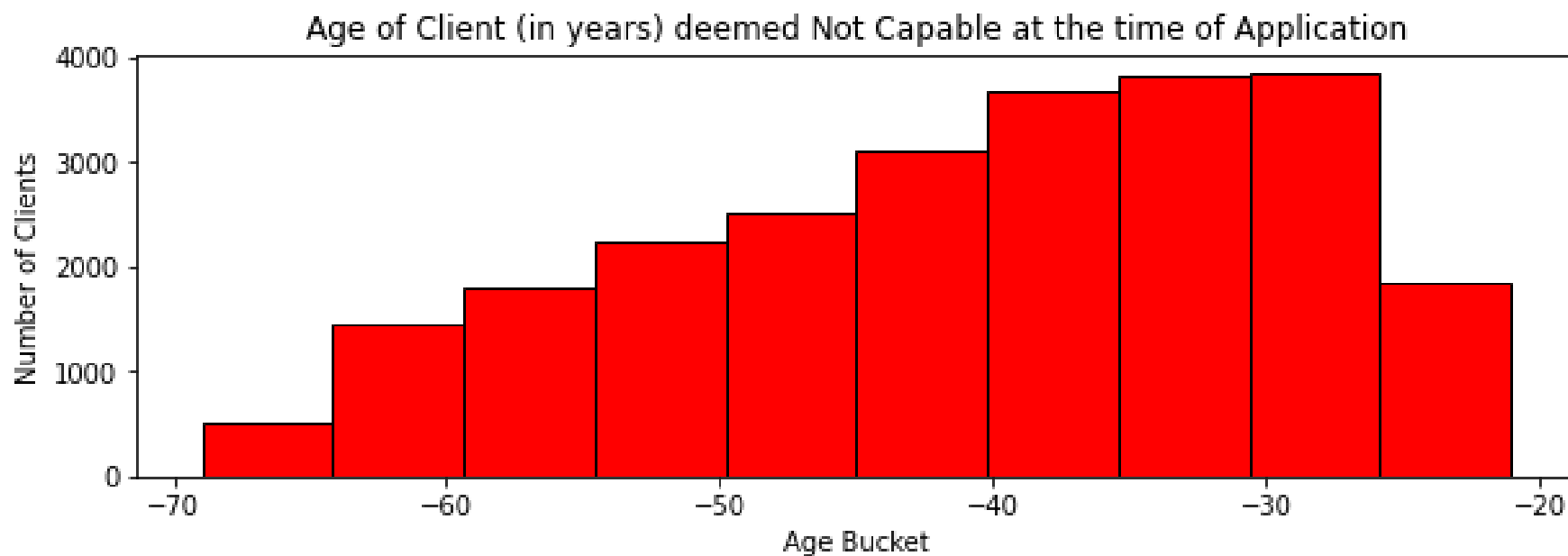
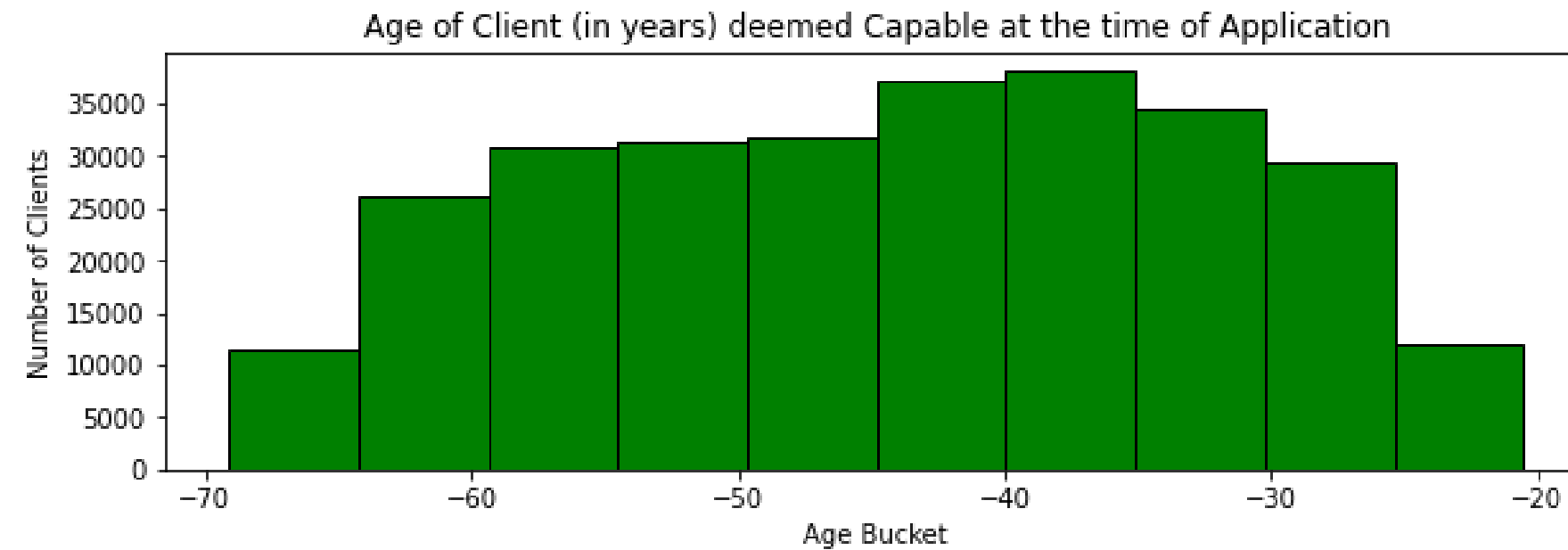
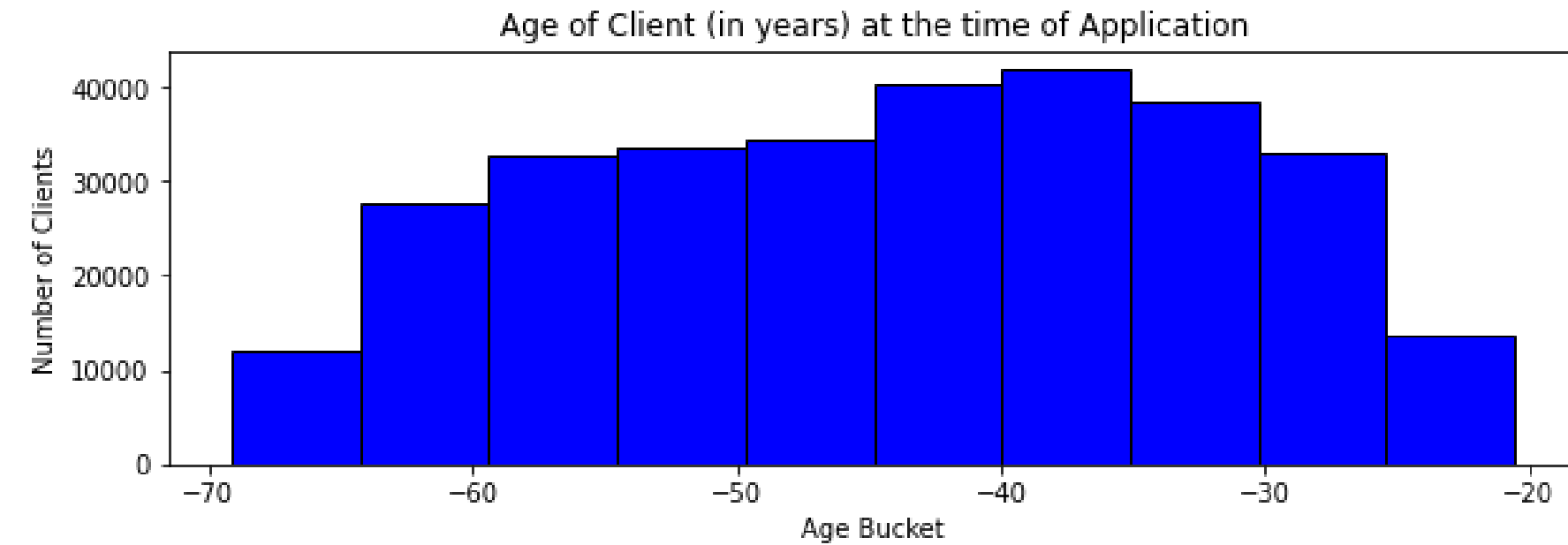


Name_Housing_Type

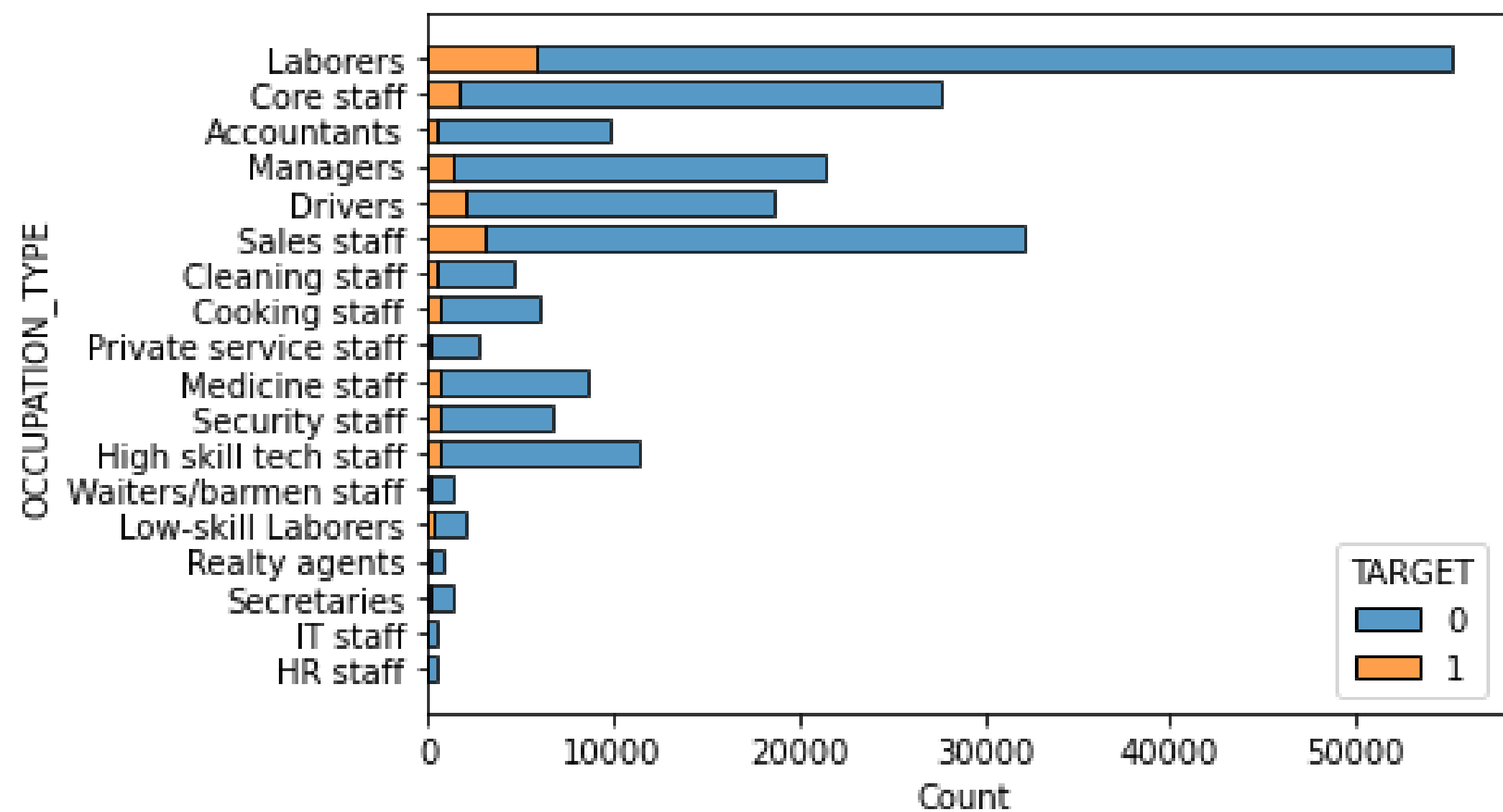
People living in a House/Apartment apply for the most number of loans and the number of people deemed incapable of repayment in this case is also the highest, whereas if percentages are looked at, people living in rented apartment have the highest chance of default.

Days_Birth

Most number of people applying for loans are in the range of (35-40) years whereas this is followed by people in the range of (40-45) years whereas the number of applicants in people aged < 25 or aged > 65 is very low.



Again, for the people who are deemed capable of loan repayment, people in the same age buckets of (35-40) years and (40-45) years are deemed to be most capable. People aged in the buckets (25-30) years and (30-35) years have a large chance of being deemed not capable for loan repayment



	OCCUPATION_TYPE	TARGET	Total	Avg
0	Accountants	474	9813	0.048303
1	Cleaning staff	447	4653	0.096067
2	Cooking staff	621	5946	0.104440
3	Core staff	1738	27570	0.063040
4	Drivers	2107	18603	0.113261
5	HR staff	36	563	0.063943
6	High skill tech staff	701	11380	0.061599
7	IT staff	34	526	0.064639
8	Laborers	5838	55186	0.105788
9	Low-skill Laborers	359	2093	0.171524
10	Managers	1328	21371	0.062140
11	Medicine staff	572	8537	0.067002
12	Private service staff	175	2652	0.065988
13	Realty agents	59	751	0.078562
14	Sales staff	3092	32102	0.096318
15	Secretaries	92	1305	0.070498
16	Security staff	722	6721	0.107424
17	Waiters/barmen staff	152	1348	0.112760

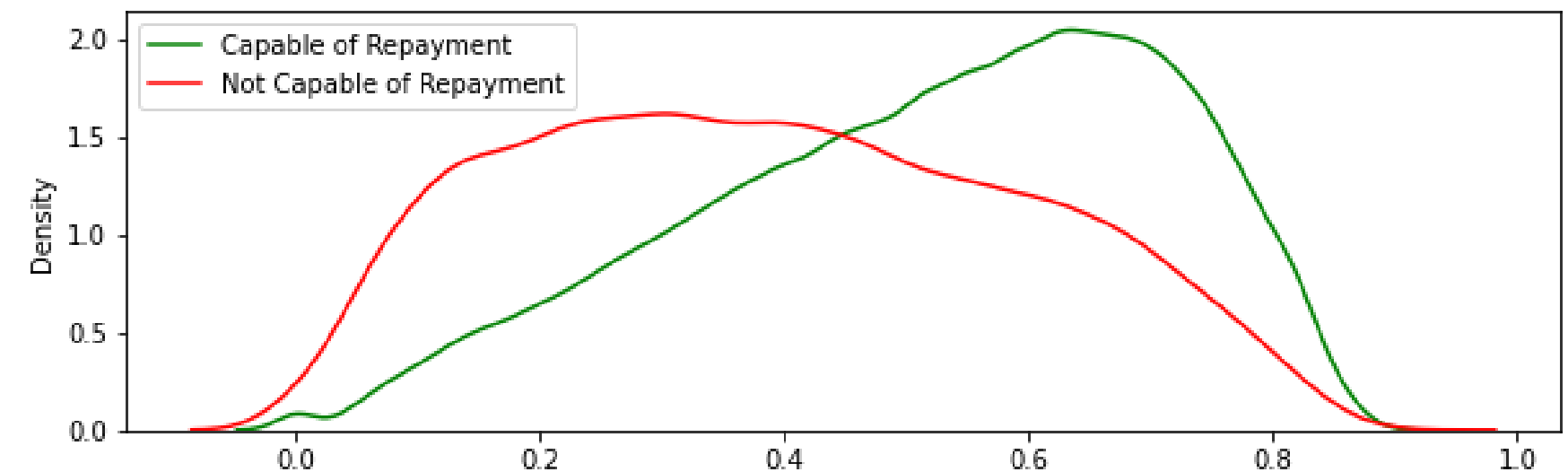
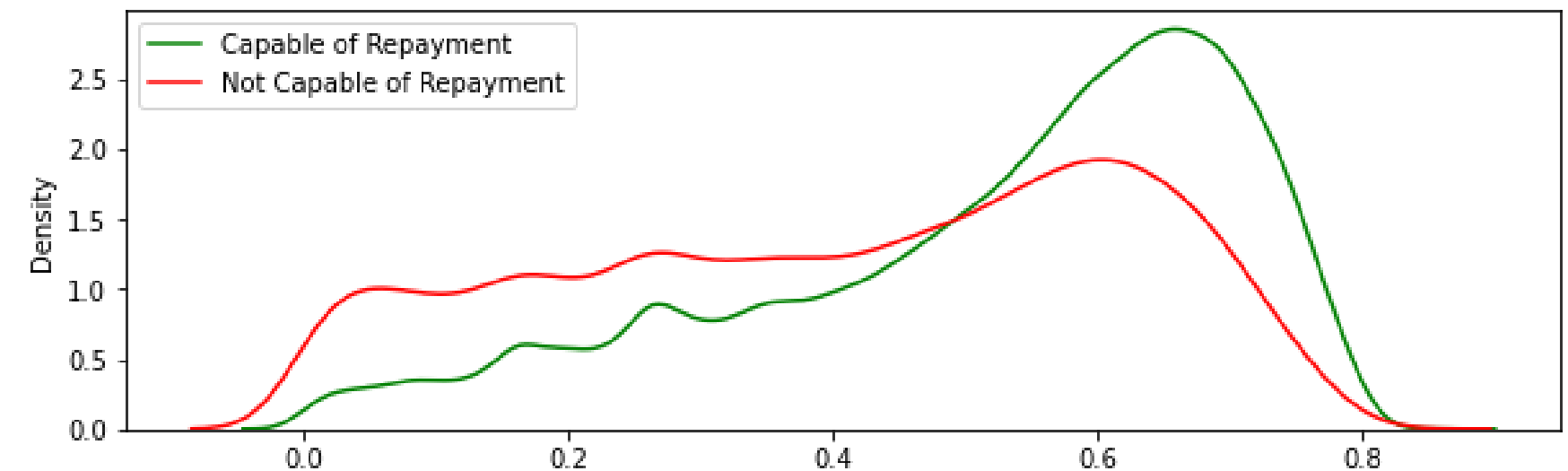
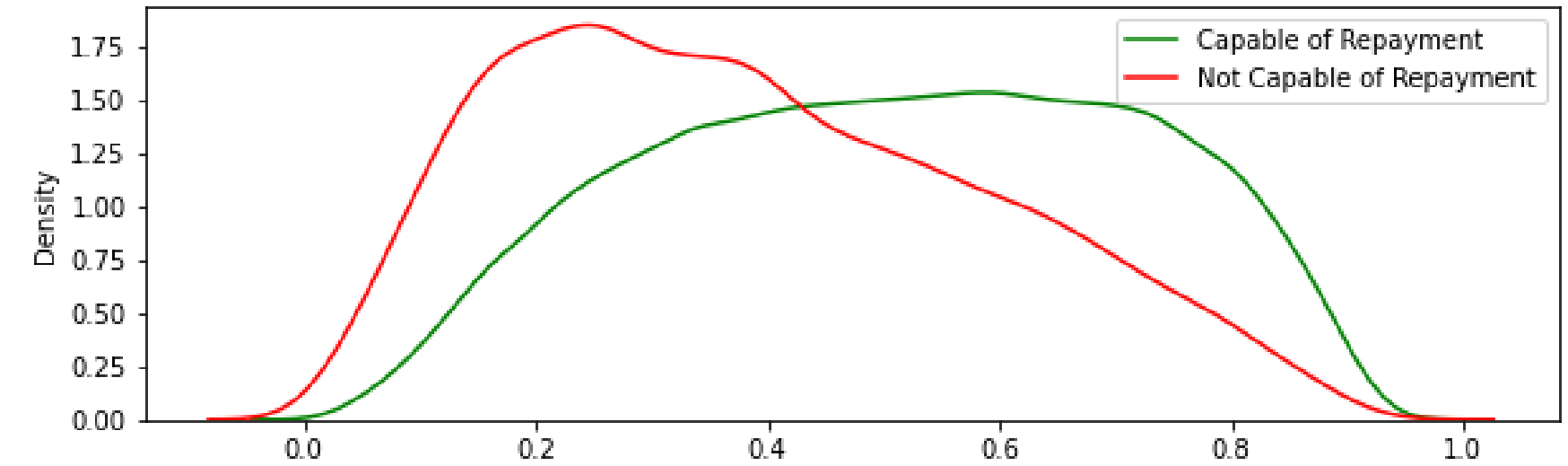
Occupation_Type

Out of all the possible Occupation Types, the majority of applicants have not provided their Occupation Type in the application (approx. 31.39%) which is followed by Laborers (approx. 18%). Out of all the occupations, Waiters/barmen staff are considered to be the least capable of repayment followed by Laborers -> though laborers have considerably higher applications as compared to Waiters/barmen staff.

EXT_SOURCE_1/2/3

There is some considerable difference among the 2 classes, as we can see from the PDF plot. Therefore, 'Ext_Source_1', 'Ext_Source_2' and 'Ext_Source_3' is going to be an important features.

Univariate Analysis



bureau.csv

This table consists of all client's previous credit records with financial institutions other than Home Credit Group which were reported by the the Credit Bureau.

		BUREAU	CREDIT
		5714462	Closed
		5714463	Active
4	5354	5714464	Active
5	5354	5714465	Active
6	5354	5714466	Active
7	5354	5714467	Active
8	5354	5714468	Active
9	2297	5714469	Closed
10	2297	5714470	Closed
11	2297	5714471	Active
12	2297	5714472	Active
13	2297	5714473	Closed
14	2297	5714474	Active
15	2297	5714475	Active

bureau_data.head()

	SK_ID_CURR	SK_ID_BUREAU	CREDIT_ACTIVE	CREDIT_CURRENCY	DAYS_CREDIT	CREDIT_DAY_OVERDUE	DAYS_CREDIT_ENDDATE	DAYS_ENDDATE_FACT
0	215354	5714462	Closed	currency 1	-497	0	-153.0	-153
1	215354	5714463	Active	currency 1	-208	0	1075.0	Na
2	215354	5714464	Active	currency 1	-203	0	528.0	Na
3	215354	5714465	Active	currency 1	-203	0	NaN	Na
4	215354	5714466	Active	currency 1	-629	0	1197.0	Na

bureau_data.shape

(1716428, 17)

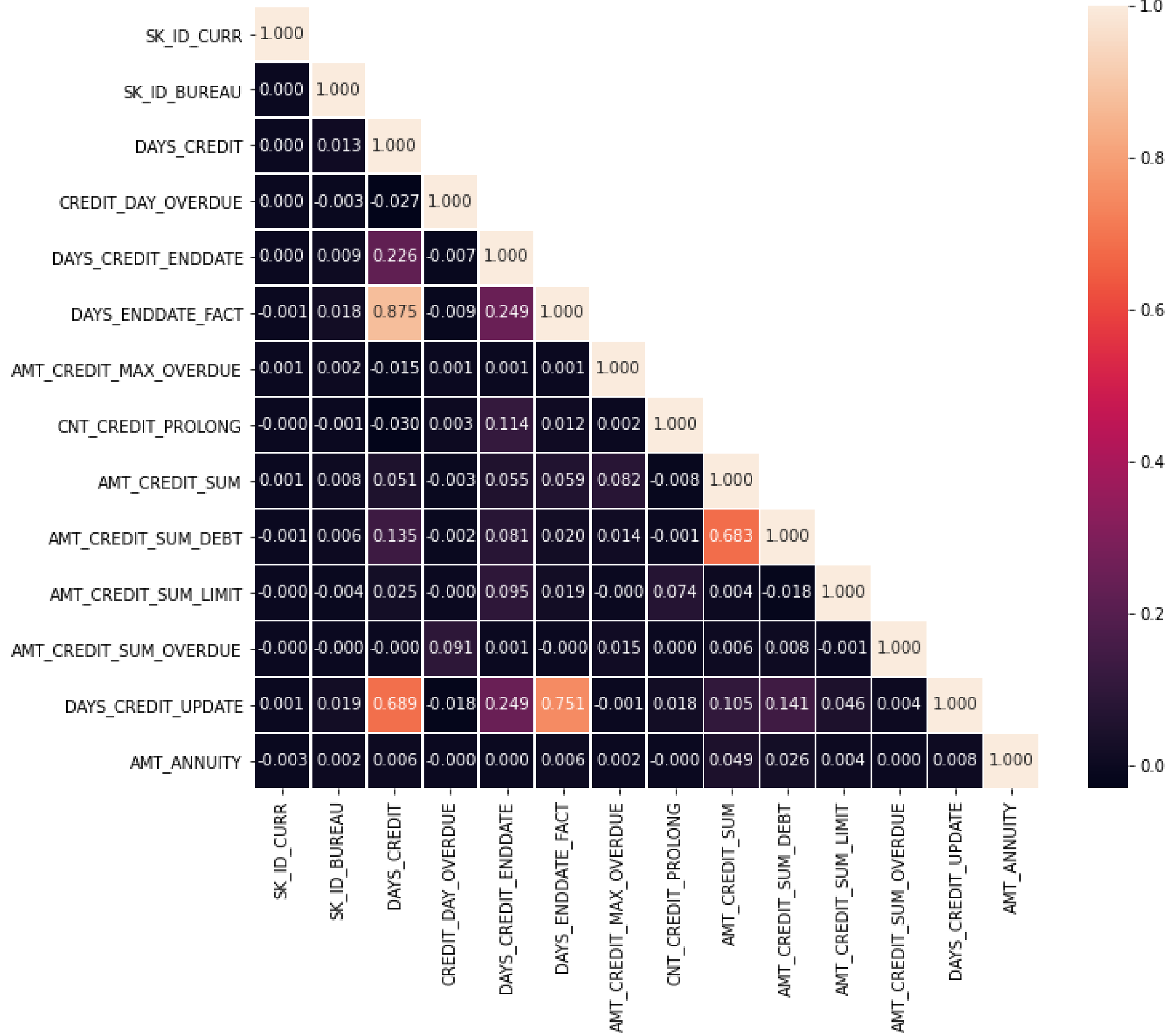
Structure Investigation

```
1 bureau_data.describe().T
```

	count	mean	std	min	25%	50%	75%	max
SK_ID_CURR	1716428.0	2.782149e+05	1.029386e+05	100001.000	188866.75	278055.0	367426.00	4.562550e+05
SK_ID_BUREAU	1716428.0	5.924434e+06	5.322657e+05	5000000.000	5463953.75	5926303.5	6385681.25	6.843457e+06
DAYS_CREDIT	1716428.0	-1.142108e+03	7.951649e+02	-2922.000	-1666.00	-987.0	-474.00	0.000000e+00
CREDIT_DAY_OVERDUE	1716428.0	8.181666e-01	3.654443e+01	0.000	0.00	0.0	0.00	2.792000e+03
DAYS_CREDIT_ENDDATE	1610875.0	5.105174e+02	4.994220e+03	-42060.000	-1138.00	-330.0	474.00	3.119900e+04
DAYS_ENDDATE_FACT	1082775.0	-1.017437e+03	7.140106e+02	-42023.000	-1489.00	-897.0	-425.00	0.000000e+00
AMT_CREDIT_MAX_OVERDUE	591940.0	3.825418e+03	2.060316e+05	0.000	0.00	0.0	0.00	1.159872e+08
CNT_CREDIT_PROLONG	1716428.0	6.410406e-03	9.622391e-02	0.000	0.00	0.0	0.00	9.000000e+00
AMT_CREDIT_SUM	1716415.0	3.549946e+05	1.149811e+06	0.000	51300.00	125518.5	315000.00	5.850000e+08
AMT_CREDIT_SUM_DEBT	1458759.0	1.370851e+05	6.774011e+05	-4705600.320	0.00	0.0	40153.50	1.701000e+08
AMT_CREDIT_SUM_LIMIT	1124648.0	6.229515e+03	4.503203e+04	-586406.115	0.00	0.0	0.00	4.705600e+06
AMT_CREDIT_SUM_OVERDUE	1716428.0	3.791276e+01	5.937650e+03	0.000	0.00	0.0	0.00	3.756681e+06
DAYS_CREDIT_UPDATE	1716428.0	-5.937483e+02	7.207473e+02	-41947.000	-908.00	-395.0	-33.00	3.720000e+02
AMT_ANNUITY	489637.0	1.571276e+04	3.258269e+05	0.000	0.00	0.0	13500.00	1.184534e+08

Structure Investigation

Correlation matrix



Total	Percent	Percent
AMT_ANNUITY	1226791	71.473490
AMT_CREDIT_MAX_OVERDUE	1124488	65.513264
DAYS_ENDDATE_FACT	633653	36.916958
AMT_CREDIT_SUM_LIMIT	591780	34.477415
AMT_CREDIT_SUM_DEBT	257669	15.011932
DAYS_CREDIT_ENDDATE	105553	6.149573
AMT_CREDIT_SUM	13	0.000757
CREDIT_ACTIVE	0	0.000000

CREDIT_CURRENCY	0	0.000000
DAYS_CREDIT	0	0.000000
CREDIT_DAY_OVERDUE	0	0.000000
SK_ID_BUREAU	0	0.000000
CNT_CREDIT_PROLONG	0	0.000000
AMT_CREDIT_SUM_OVERDUE	0	0.000000
CREDIT_TYPE	0	0.000000
DAYS_CREDIT_UPDATE	0	0.000000
SK_ID_CURR	0	0.000000

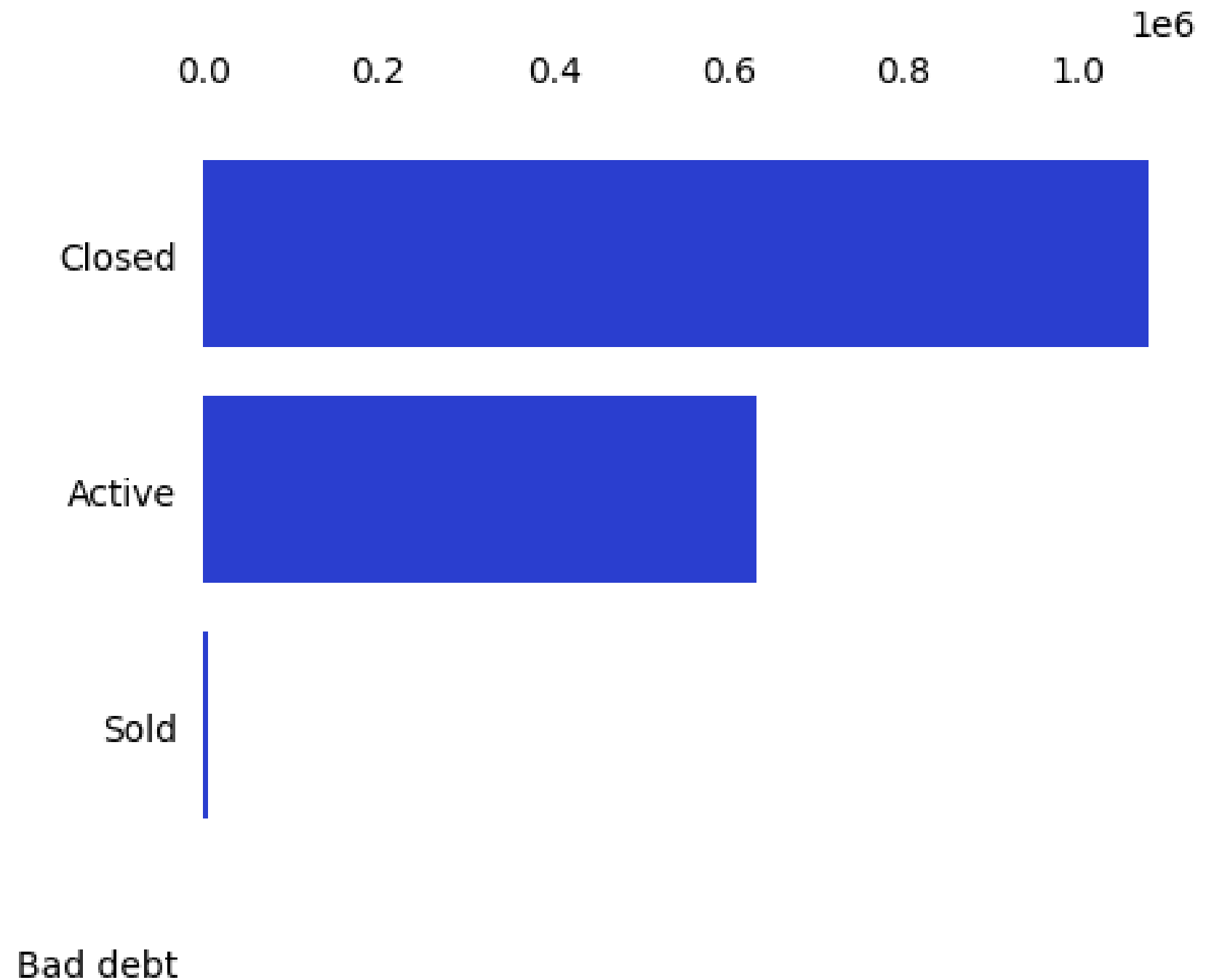
Variables interpretation

Column	Description
SK_BUREAU_ID	Recoded ID of previous Credit Bureau credit related to our loan (unique coding for each loan application)
CREDIT_ACTIVE	Status of the Credit Bureau (CB) reported credits
DAYS_CREDIT	How many days before current application did client apply for Credit Bureau credit
CREDIT_DAY_OVERDUE	Number of days past due on CB credit at the time of application for related loan in our sample
CREDIT_TYPE	Type of Credit Bureau credit (Car, cash,...)

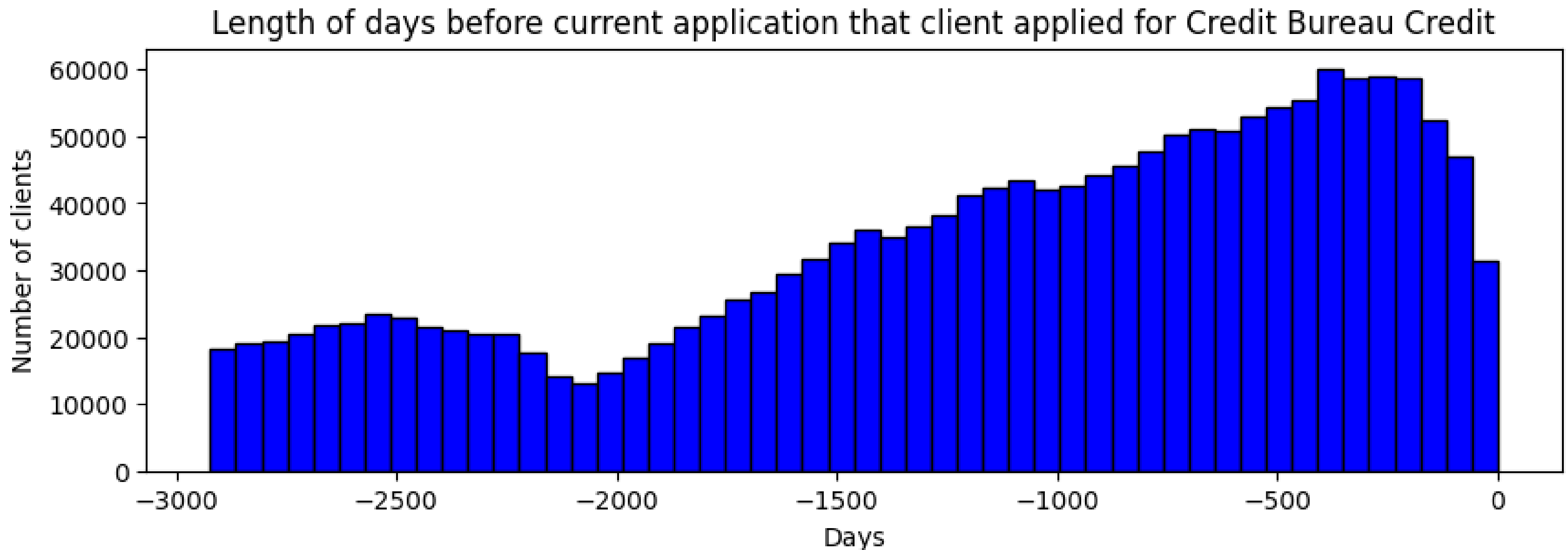
Univariate Analysis

CREDIT_ACTIVE

- Most of the applications in the bureau_data is closed, following by the status of being Active
- There are very few loans that are "Sold" or considered to be "Bad debt"



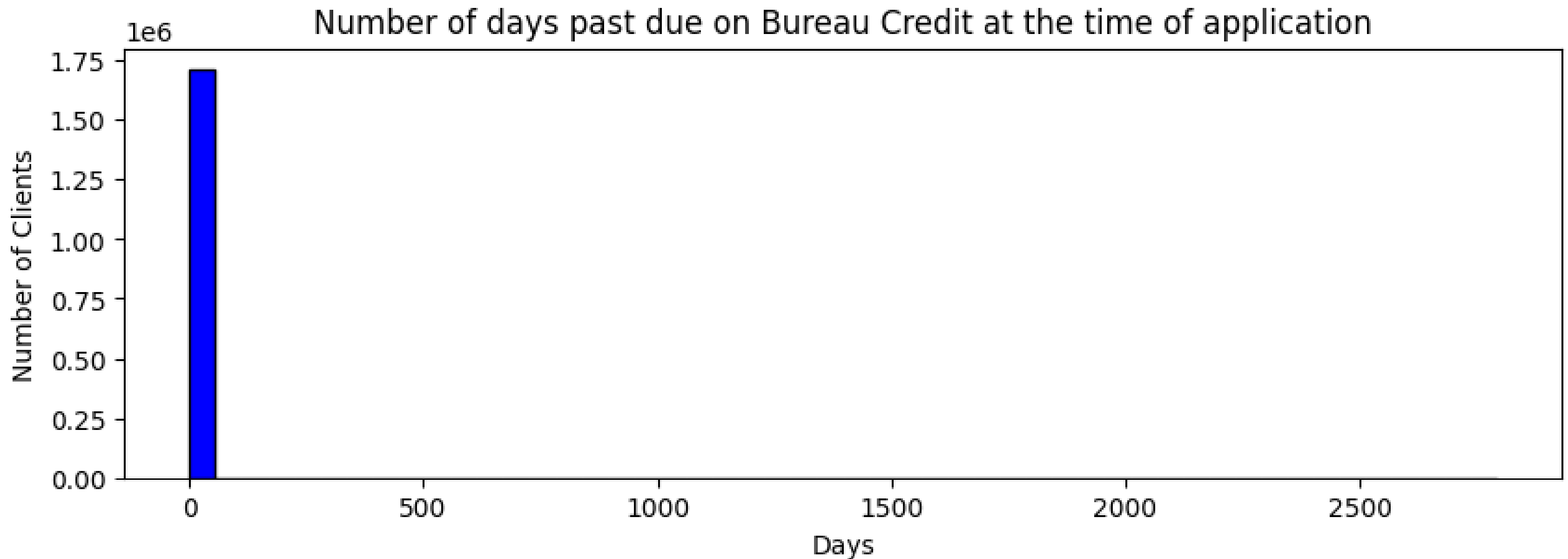
Univariate Analysis



DAYS_CREDIT

Most of clients applied for Bureau Credit is less than 500 days before the data of loan application

Univariate Analysis



CREDIT_DAY_OVERDUE

- This shows that most of the clients have a low "days past due (close to 0)" on Credit Bureau at the time of their application since the histogram is very peaked near 0

Univariate Analysis

CREDIT_DAY_OVERDUE

Looking further:

- there are total of 4217 (out of 1716428) datapoints where the CREDIT_DAY_OVERDUE value is greater than 0

Percentile	Number of days past due on CB Credit
0	0
10	0
20	0
30	0
40	0
50	0
60	0
70	0
80	0
90	0
100	2792

There is 0 day past due till the 90th percentile, while the 100th percentile is a value = 2792 (max =2792)

Zooming into 99th - 100 percentile:

- We can see from here that only the Top 0.3 percentile of values over here are non-zeroes.

Percentile	Number of days past due on CB Credit
99.0	0.00
99.1	0.00
99.2	0.00
99.3	0.00
99.4	0.00
99.5	0.00
99.6	0.00
99.7	0.00
99.8	13.00
99.9	52.57
100.0	2792.00

Univariate Analysis

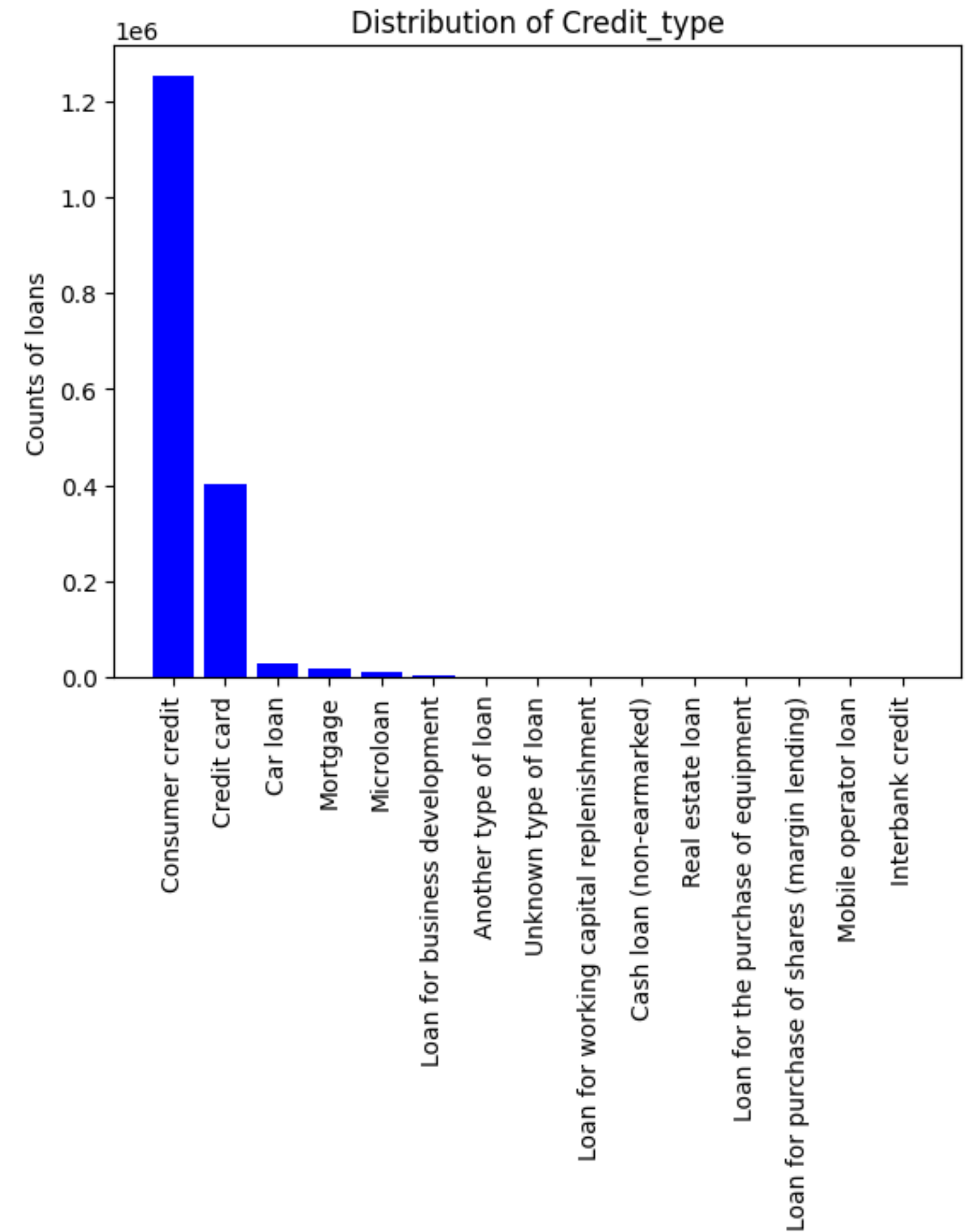
CREDIT_TYPE

- Type of Credit Bureau credit (Car, cash,...)

Comment:

- Consumer Credit and Credit Cards are the mostly registered credit types in the Credit Bureau

=> We may assign the 13 remaining as one - Rare



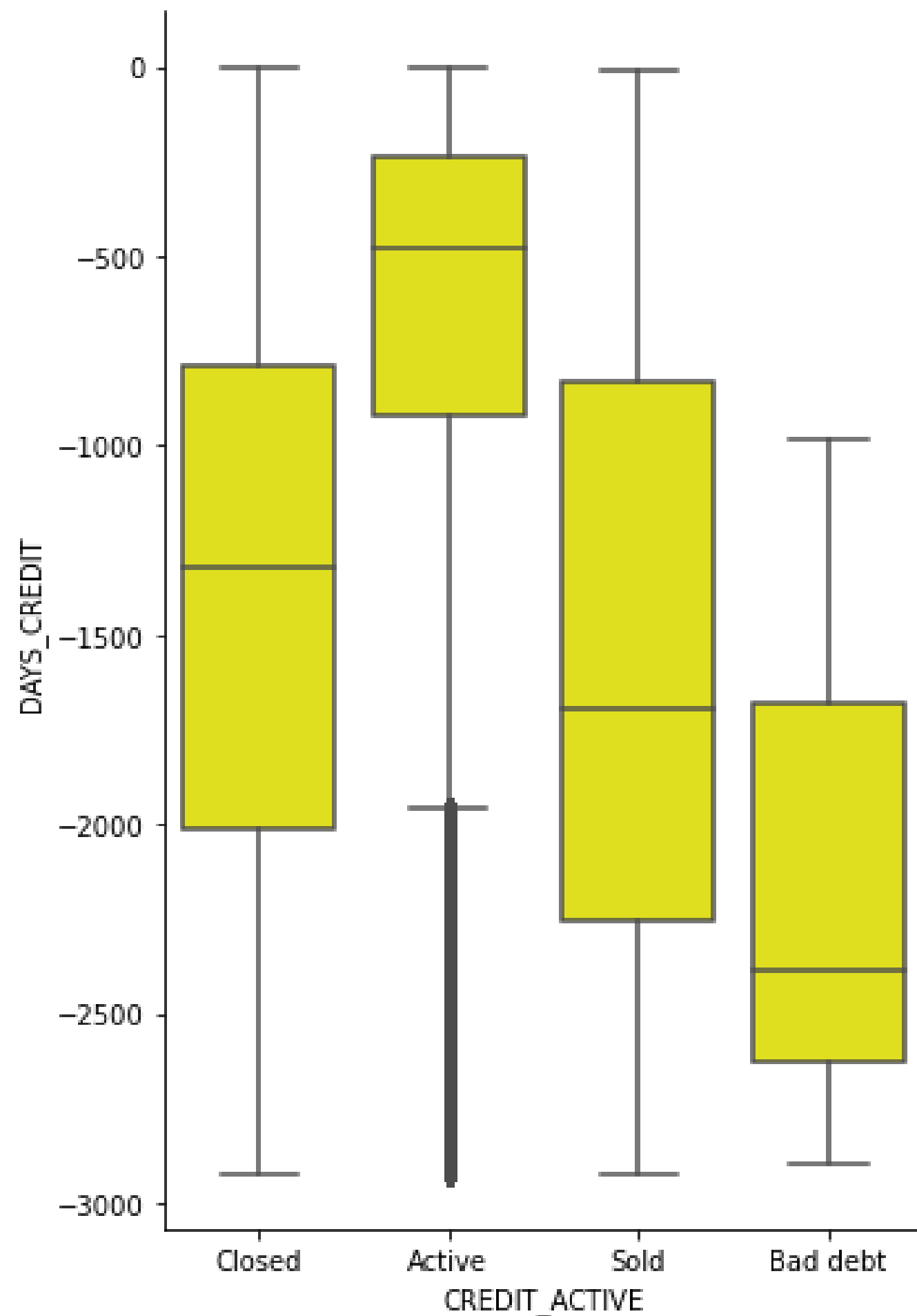
Univariate Analysis

CREDIT_ACTIVE vs DAYS_CREDIT

- CREDIT_ACTIVE: Status of the Credit Bureau (CB) reported credits
- DAYS_CREDIT: How many days before current application did client apply for Credit Bureau credit

Comment:

- When the Credit Status is Active, it means that the corresponding "DAYS_CREDIT" i.e number of days before application, the median value is approximately 500 days



bureau_balance.csv

his table consists of Monthly balance of each credit for each of the previous credit that the client had with financial institutions other than Home Credit.

		MONTHS	STATUS	
	5715448	0	C	
	5715448	-1	C	
4	5715448	-2	C	
5	5715448	-3	C	
6	5715448	-4	C	
7	5715448	-5	C	
8	5715448	-6	C	
9	5715448	-7	C	
10	5715448	-8	C	
11	5715448	-9		0
12	5715448	-10		0
13	5715448	-11	X	
14	5715448	-12	X	
15	5715448	-13	X	

```
bureau_balance.head()
```

	SK_ID_BUREAU	MONTHS_BALANCE	STATUS
0	5715448	0	C
1	5715448	-1	C
2	5715448	-2	C
3	5715448	-3	C
4	5715448	-4	C

```
bureau_balance.shape  
(27299925, 3)
```



Total	Percent	Percent
SK_ID_BUREAU	0	0.0
MONTHS_BALANCE	0	0.0
STATUS	0	0.0

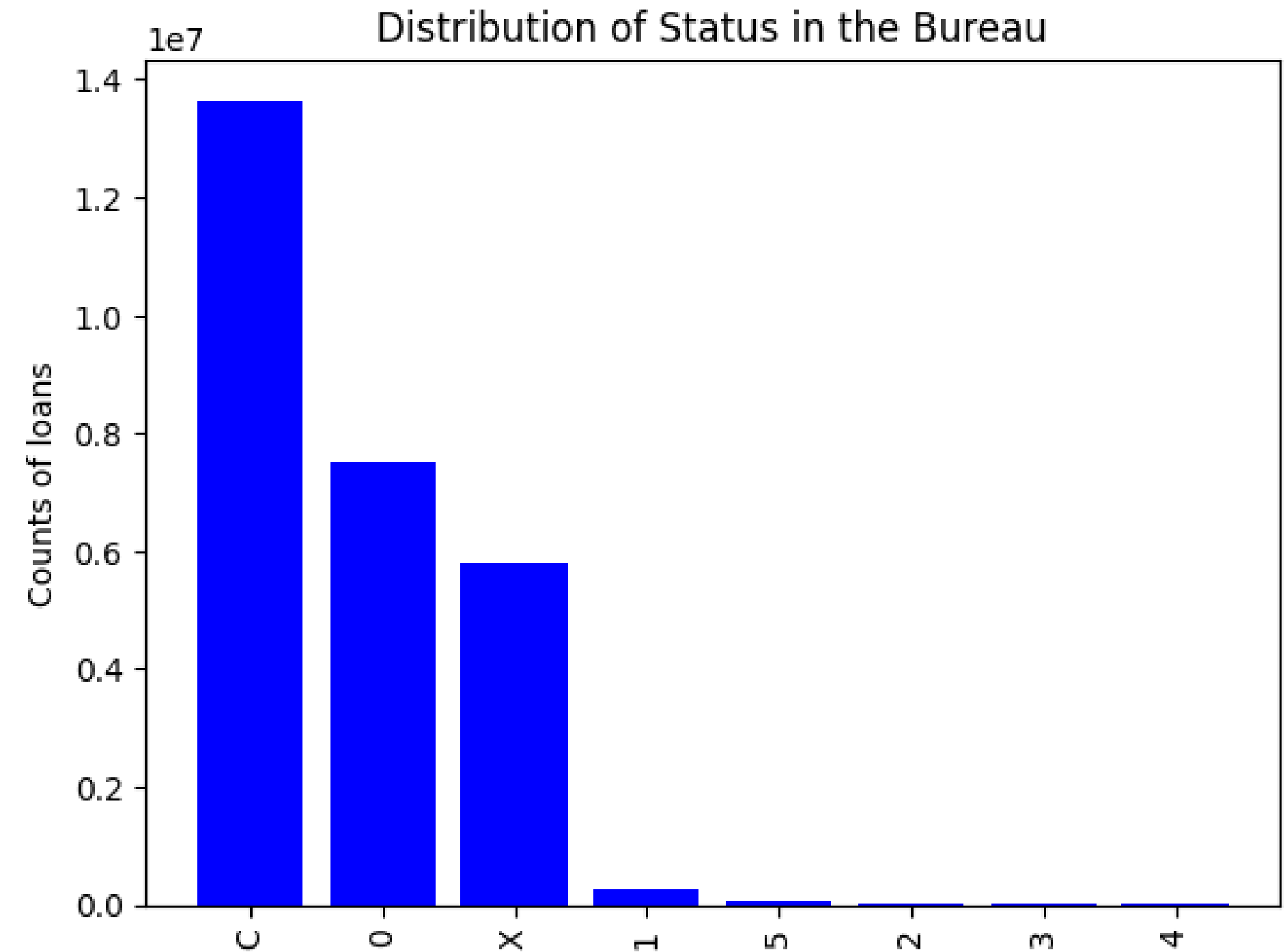
Variables interpretation

Column	Description
SK_BUREAU_ID	Recoded ID of Credit Bureau credit (unique coding for each application)
MONTHS_BALANCE	Month of balance relative to application date (-1 means the freshest balance date)
STATUS	Status of Credit Bureau loan during the month (active, closed, DPD0-30,... [C means closed, X means status unknown, 0 means no DPD, 1 means maximal did during month between 1-30, 2 means DPD 31-60,... 5 means DPD 120+ or sold or written off])

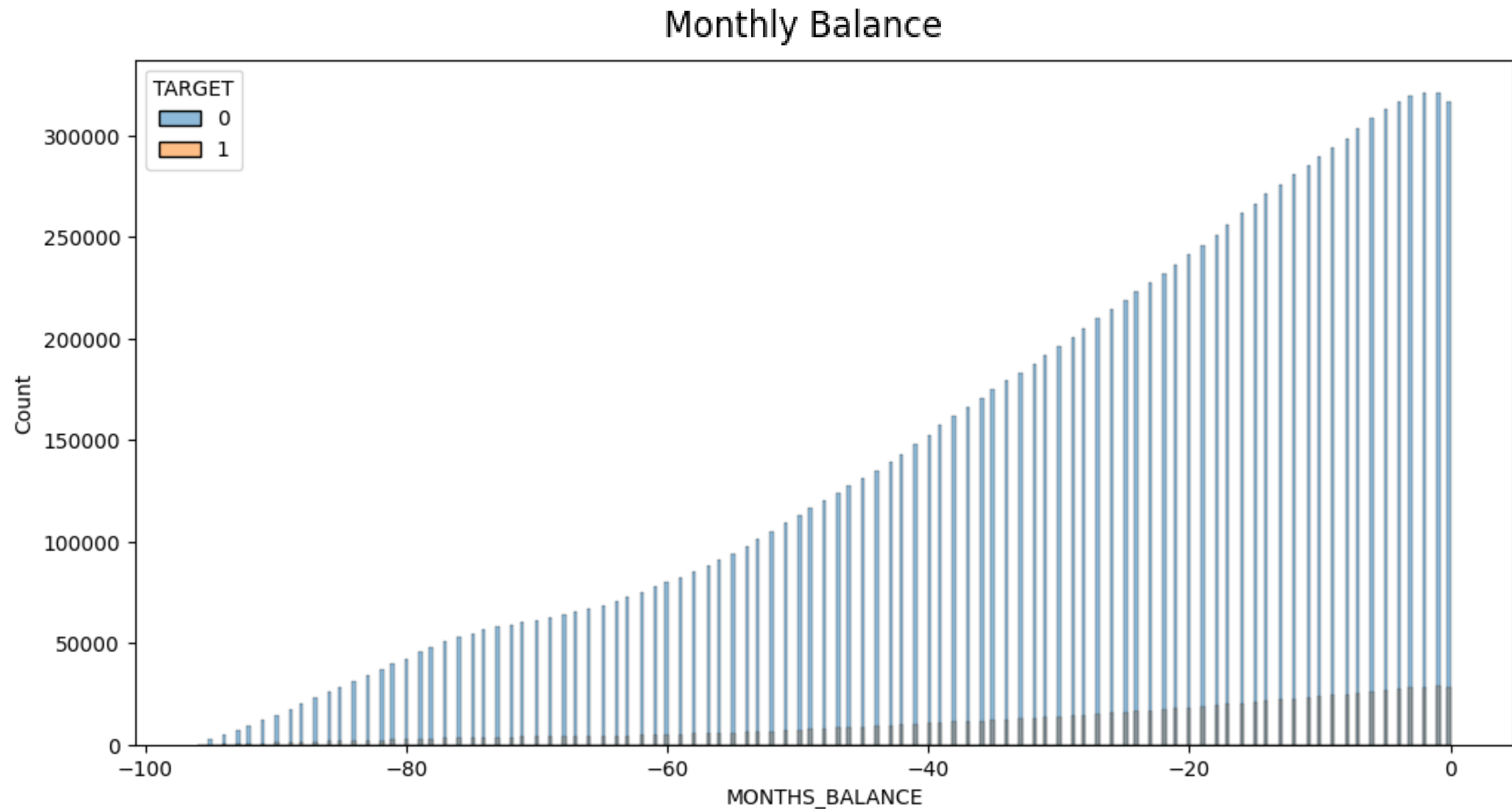
Univariate Analysis

STATUS

- Most of the loans are closed in the Credit Bureau, which is followed by clients with 0 DPD and then by applicants whose status is unknown
- We can conclude that there are very few annuity defaulters in the data

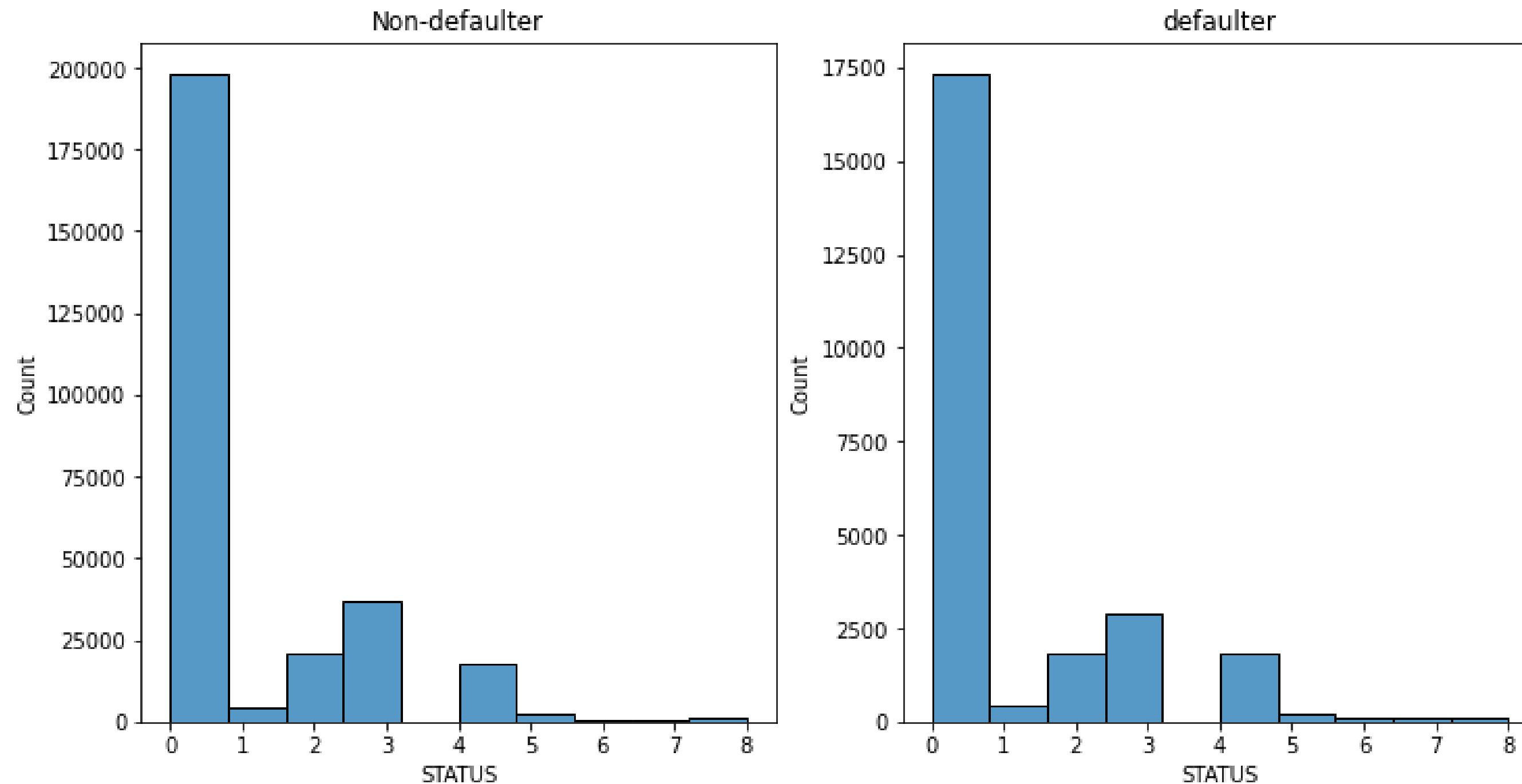


Univariate Analysis



MONTHS_BALANCE

Univariate Analysis



Status of defaulter vs non-defaulter

We see a same trend in status both in defaulter and non defaulter

previous_application.csv

This table contains the static data of the previous loan which the client had with Home Credit.

	SK_ID_CUR	NAME_	
	20495	271877 Consumer	
	2802425	108129 Cash loans	25188.1
4	2523466	122040 Cash loans	15060.74
5	2819243	176158 Cash loans	47041.34
6	1784265	202054 Cash loans	31924.4
7	1383531	199383 Cash loans	23703.93
8	2315218	175704 Cash loans	
9	1656711	296299 Cash loans	
10	2367563	342292 Cash loans	
11	2579447	334349 Cash loans	
12	1715995	447712 Cash loans	11368.62
13	2257824	161140 Cash loans	13832.75
14	230894	258628 Cash loans	122
15	321676 Consumer		

previous

previous_application.head(10)

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAYS
0	2030495	271877	Consumer loans	1730.430	17145.0	17145.0	0.0	17145.0	
1	2802425	108129	Cash loans	25188.615	607500.0	679671.0	NaN	607500.0	
2	2523466	122040	Cash loans	15060.735	112500.0	136444.5	NaN	112500.0	
3	2819243	176158	Cash loans	47041.335	450000.0	470790.0	NaN	450000.0	
4	1784265	202054	Cash loans	31924.395	337500.0	404055.0	NaN	337500.0	
5	1383531	199383	Cash loans	23703.930	315000.0	340573.5	NaN	315000.0	
6	2315218	175704	Cash loans	NaN	0.0	0.0	NaN	NaN	
7	1656711	296299	Cash loans	NaN	0.0	0.0	NaN	NaN	
8	2367563	342292	Cash loans	NaN	0.0	0.0	NaN	NaN	
9	2579447	334349	Cash loans	NaN	0.0	0.0	NaN	NaN	

previous_application.shape
(1670214, 37)

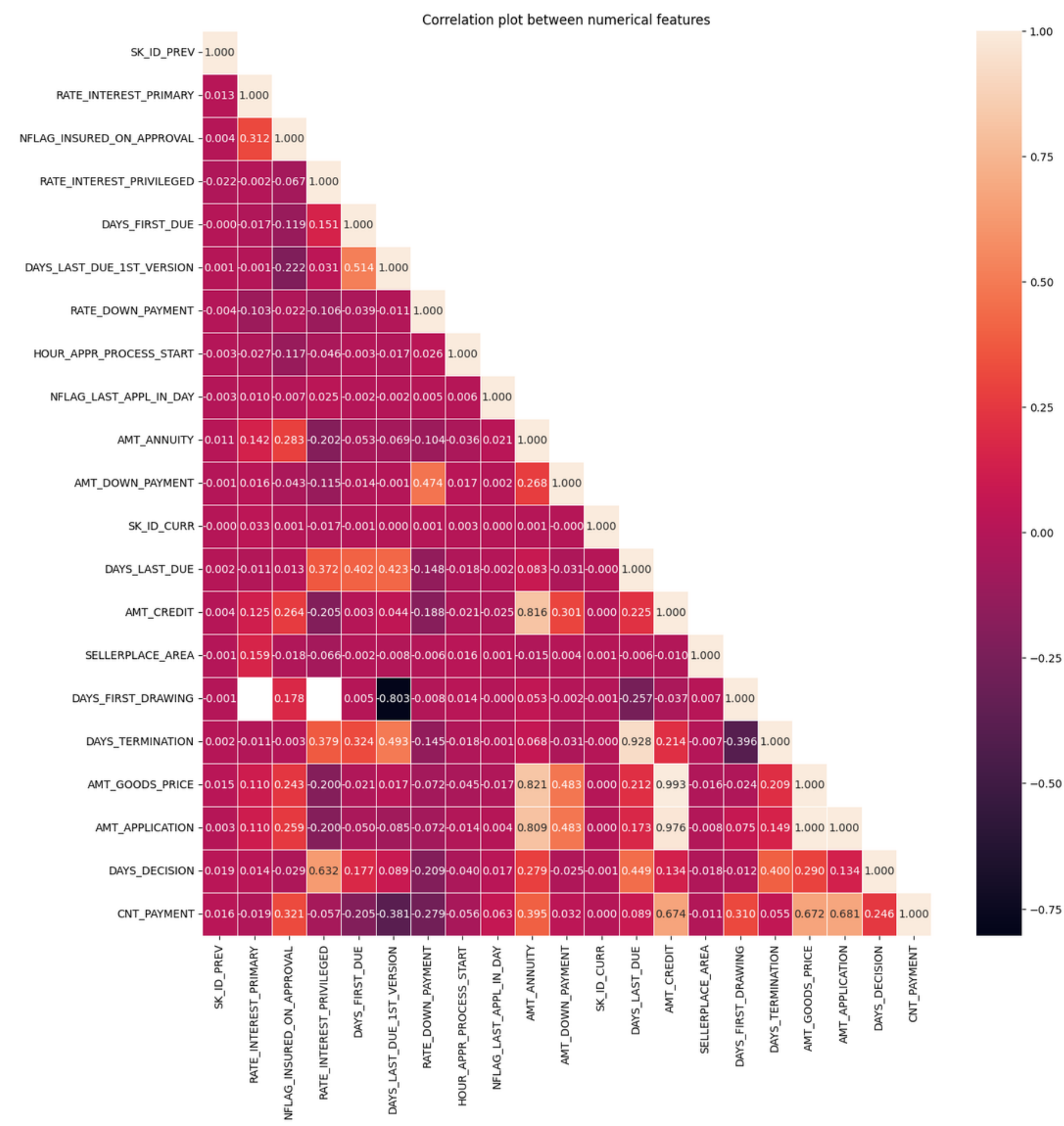
Structure Investigation

```
1 previous_application.describe().T
```

	count	mean	std	min	25%	50%	75%	max
SK_ID_PREV	1670214.0	1.923089e+06	532597.958696	1.000001e+06	1.461857e+06	1.923110e+06	2.384280e+06	2845382.000
SK_ID_CURR	1670214.0	2.783572e+05	102814.823849	1.000010e+05	1.893290e+05	2.787145e+05	3.675140e+05	456255.000
AMT_ANNUITY	1297979.0	1.595512e+04	14782.137335	0.000000e+00	6.321780e+03	1.125000e+04	2.065842e+04	418058.145
AMT_APPLICATION	1670214.0	1.752339e+05	292779.762386	0.000000e+00	1.872000e+04	7.104600e+04	1.803600e+05	6905160.000
AMT_CREDIT	1670213.0	1.961140e+05	318574.616547	0.000000e+00	2.416050e+04	8.054100e+04	2.164185e+05	6905160.000
AMT_DOWN_PAYMENT	774370.0	6.697402e+03	20921.495410	-9.000000e-01	0.000000e+00	1.638000e+03	7.740000e+03	3060045.000
AMT_GOODS_PRICE	1284699.0	2.278473e+05	315396.557937	0.000000e+00	5.084100e+04	1.123200e+05	2.340000e+05	6905160.000
HOUR_APPR_PROCESS_START	1670214.0	1.248418e+01	3.334028	0.000000e+00	1.000000e+01	1.200000e+01	1.500000e+01	23.000
NFLAG_LAST_APPL_IN_DAY	1670214.0	9.964675e-01	0.059330	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000
RATE_DOWN_PAYMENT	774370.0	7.963682e-02	0.107823	-1.497876e-05	0.000000e+00	5.160508e-02	1.089091e-01	1.000
RATE_INTEREST_PRIMARY	5951.0	1.883569e-01	0.087671	3.478125e-02	1.607163e-01	1.891222e-01	1.933299e-01	1.000
RATE_INTEREST_PRIVILEGED	5951.0	7.735025e-01	0.100879	3.731501e-01	7.156448e-01	8.350951e-01	8.525370e-01	1.000
DAYS_DECISION	1670214.0	-8.806797e+02	779.099667	-2.922000e+03	-1.300000e+03	-5.810000e+02	-2.800000e+02	-1.000
SELLERPLACE_AREA	1670214.0	3.139511e+02	7127.443459	-1.000000e+00	-1.000000e+00	3.000000e+00	8.200000e+01	4000000.000
CNT_PAYMENT	1297984.0	1.605408e+01	14.567288	0.000000e+00	6.000000e+00	1.200000e+01	2.400000e+01	84.000
DAYS_FIRST_DRAWING	997149.0	3.422099e+05	88916.115833	-2.922000e+03	3.652430e+05	3.652430e+05	3.652430e+05	365243.000
DAYS_FIRST_DUE	997149.0	1.382627e+04	72444.869708	-2.892000e+03	-1.628000e+03	-8.310000e+02	-4.110000e+02	365243.000
DAYS_LAST_DUE_1ST_VERSION	997149.0	3.376777e+04	106857.034789	-2.801000e+03	-1.242000e+03	-3.610000e+02	1.290000e+02	365243.000
DAYS_LAST_DUE	997149.0	7.658240e+04	149647.415123	-2.889000e+03	-1.314000e+03	-5.370000e+02	-7.400000e+01	365243.000
DAYS_TERMINATION	997149.0	8.199234e+04	153303.516729	-2.874000e+03	-1.270000e+03	-4.990000e+02	-4.400000e+01	365243.000
NFLAG_INSURED_ON_APPROVAL	997149.0	3.325702e-01	0.471134	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000

Structure Investigation

Correlation matrix



	Total	Percent
RATE_INTEREST_PRIVILEGED	1664263	99.643698
RATE_INTEREST_PRIMARY	1664263	99.643698
AMT_DOWN_PAYMENT	895844	53.636480
RATE_DOWN_PAYMENT	895844	53.636480
NAME_TYPE_SUITE	820405	49.119754
NFLAG_INSURED_ON_APPROVAL	673065	40.298129
DAYS_TERMINATION	673065	40.298129
DAYS_LAST_DUE	673065	40.298129
DAYS_LAST_DUE_1ST_VERSION	673065	40.298129
DAYS_FIRST_DUE	673065	40.298129
DAYS_FIRST_DRAWING	673065	40.298129
AMT_GOODS_PRICE	385515	23.081773
AMT_ANNUITY	372235	22.286665
CNT_PAYMENT	372230	22.286366
PRODUCT_COMBINATION	346	0.020716
AMT_CREDIT	1	0.000060
NAME_YIELD_GROUP	0	0.000000
NAME_PORTFOLIO	0	0.000000
NAME_SELLER_INDUSTRY	0	0.000000

	Total	Percent
SELLERPLACE_AREA	0	0.000000
CHANNEL_TYPE	0	0.000000
NAME_PRODUCT_TYPE	0	0.000000
SK_ID_PREV	0	0.000000
NAME_GOODS_CATEGORY	0	0.000000
NAME_CLIENT_TYPE	0	0.000000
CODE_REJECT_REASON	0	0.000000
SK_ID_CURR	0	0.000000
DAYS_DECISION	0	0.000000
NAME_CONTRACT_STATUS	0	0.000000
NAME_CASH_LOAN_PURPOSE	0	0.000000
NFLAG_LAST_APPL_IN_DAY	0	0.000000
FLAG_LAST_APPL_PER_CONTRACT	0	0.000000
HOUR_APPR_PROCESS_START	0	0.000000
WEEKDAY_APPR_PROCESS_START	0	0.000000
AMT_APPLICATION	0	0.000000
NAME_CONTRACT_TYPE	0	0.000000
NAME_PAYMENT_TYPE	0	0.000000

Variables interpretation

Column	Description
SK_ID_CURR	ID of loan in our sample
NAME_CONTRACT_TYPE	Contract product type (Cash loan, consumer loan [POS] ,...) of the previous application
NAME_CASH_LOAN_PURPOSE	Purpose of the cash loan
NAME_CONTRACT_STATUS	Contract status (approved, cancelled, ...) of previous application
NAME_PAYMENT_TYPE	Payment method that client chose to pay for the previous application
CODE_REJECT_REASON	Why was the previous application rejected
NAME_TYPE_SUITE	Who accompanied client when applying for the previous application
NAME_CLIENT_TYPE	Was the client old or new client when applying for the previous application

Variables interpretation

Column	Description
NAME_GOODS_CATEGORY	What kind of goods did the client apply for in the previous application
NAME_PORTFOLIO	Was the previous application for CASH, POS, CAR, ...
NAME_PRODUCT_TYPE	Was the previous application x-sell o walk-in
CHANNEL_TYPE	Through which channel we acquired the client on the previous application
SELLERPLACE_AREA	Selling area of seller place of the previous application
NAME_SELLER_INDUSTRY	The industry of the seller
CNT_PAYMENT	Term of previous credit at application of the previous application
NAME_YIELD_GROUP	Grouped interest rate into small medium and high of the previous application
PRODUCT_COMBINATION	Detailed product combination of the previous application

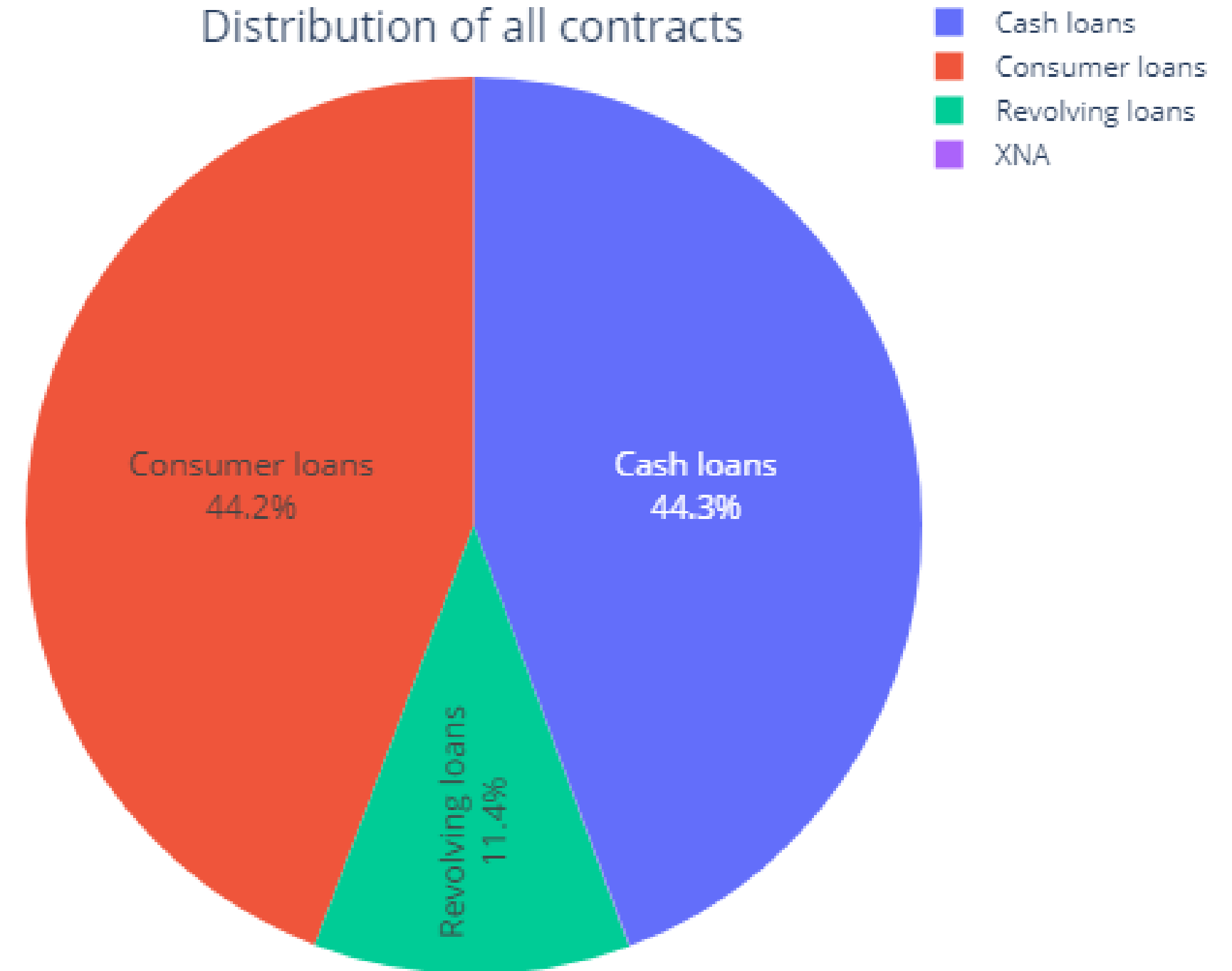
Univariate Analysis

NAME_CONTRACT_TYPE

We see that most of the previous loans have been either Cash Loans or Consumer Loans, which correspond to roughly 44% of loans each. The remaining 11.41% corresponds to Revolving Loans, and there are some loans named XNA whose types are actually not known, but they are very few in numbers.

Contract product type of previous application

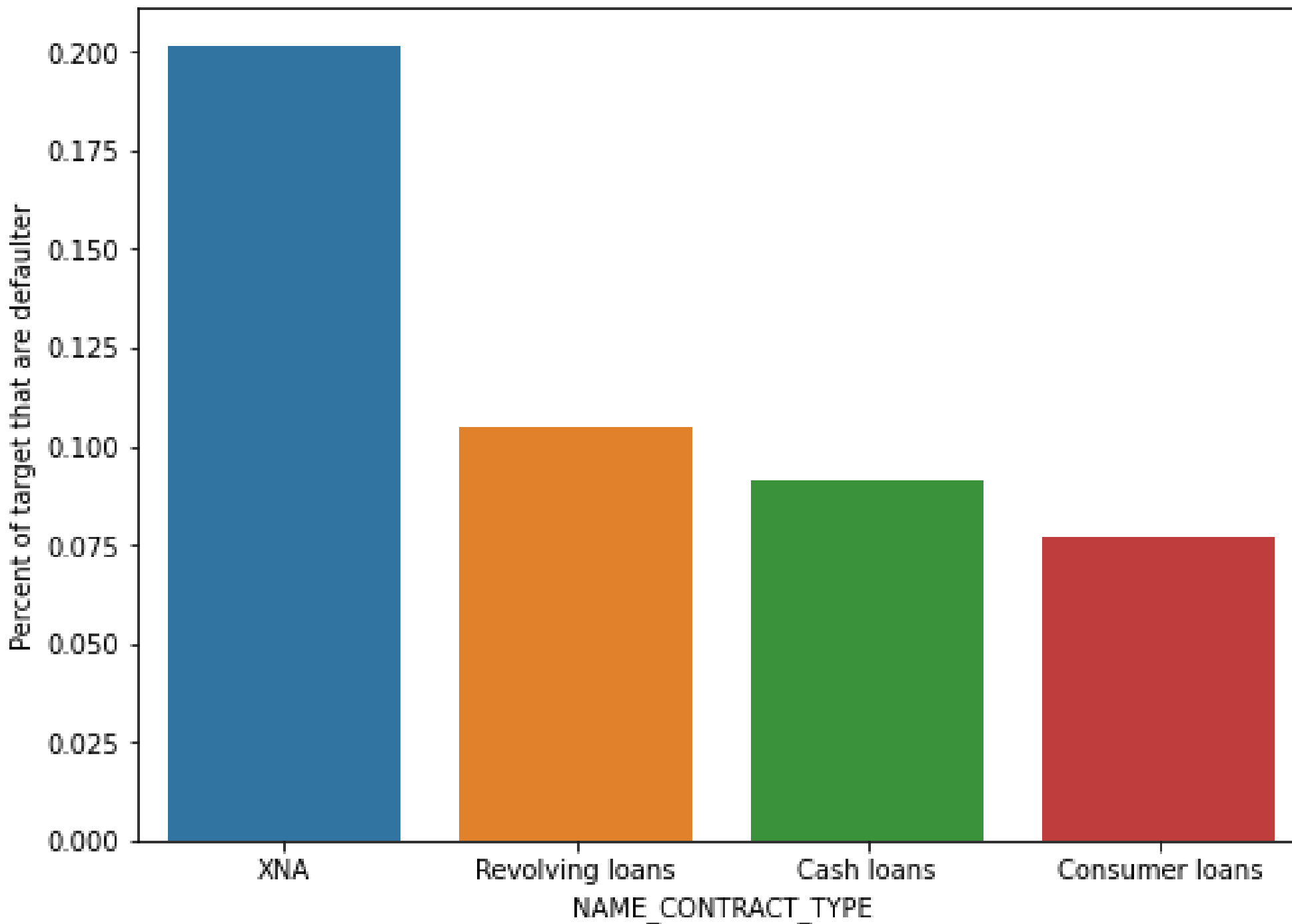
Distribution of all contracts



Univariate Analysis

NAME_CONTRACT_TYPE

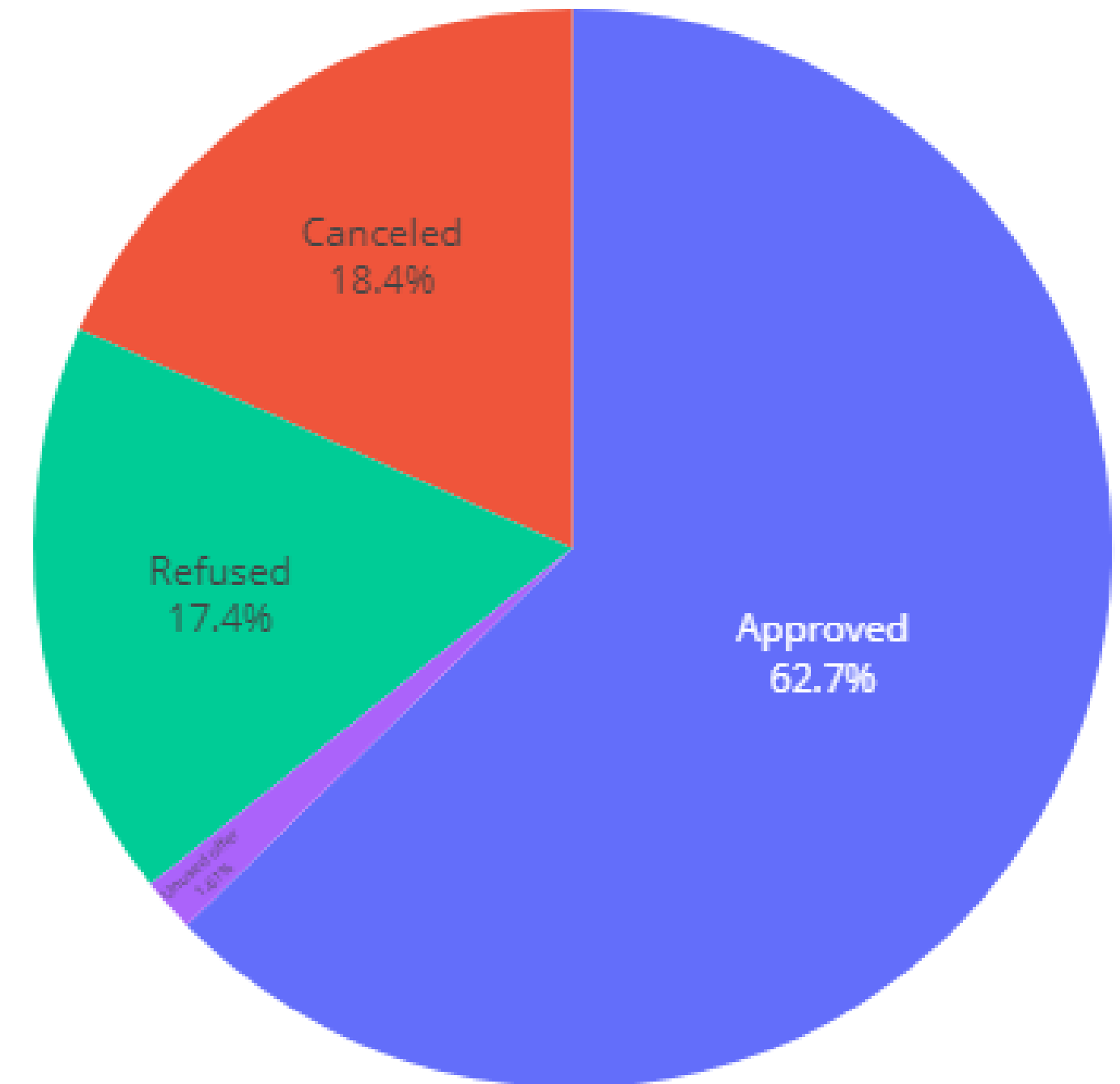
We see that the Percentage of Defaulters for XNA type of loan are the highest, at 20% Default rate. The next highest Default Rate is among Revolving Loans, which is close to 10.5%.



Univariate Analysis

Approval status of application

Distribution of all contracts



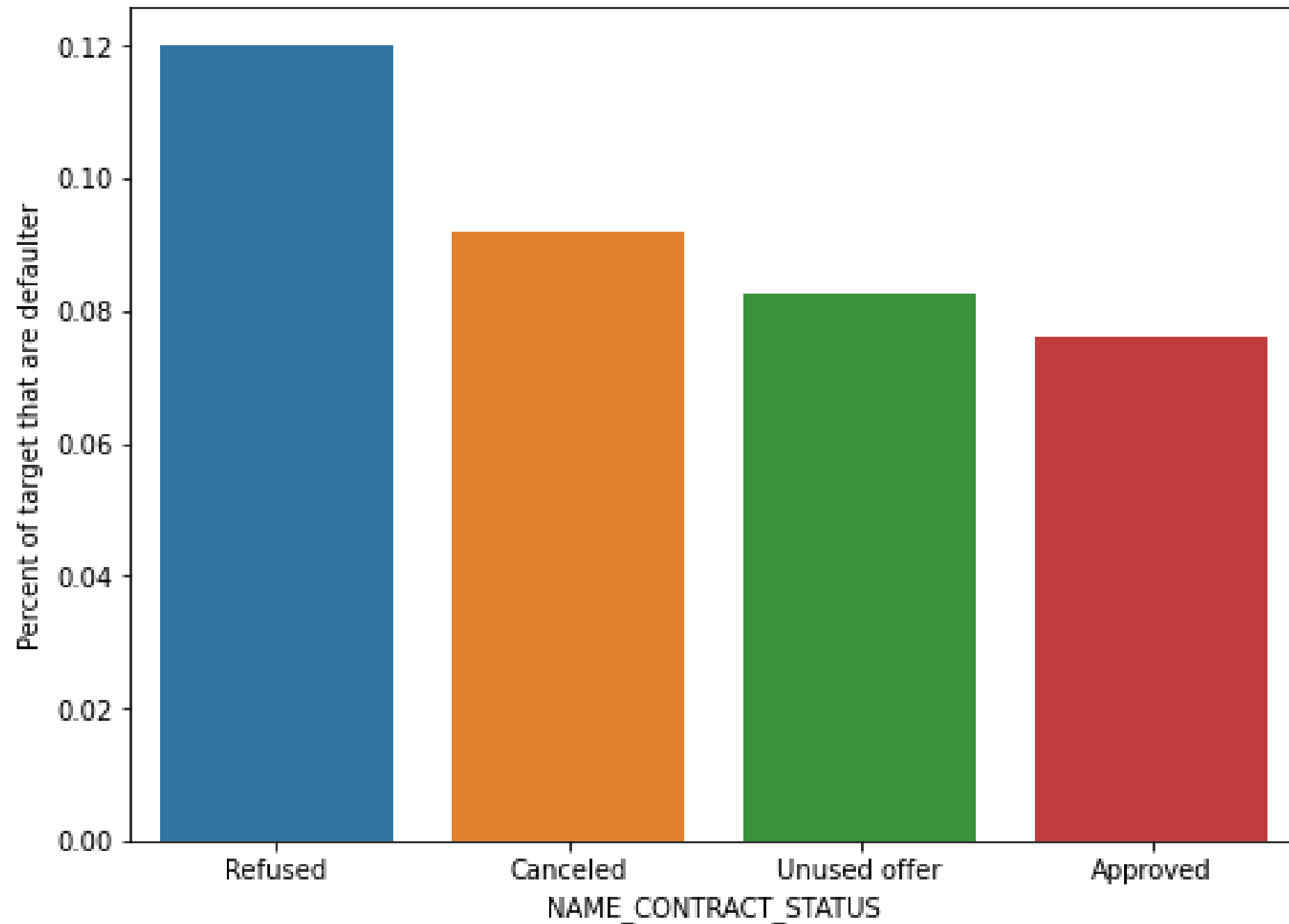
NAME_CONTRACT_STATUS

The most common type of Contract Status is the Approved Status. About 63% of the previous Credits have an Approved Status. The next two common status are Canceled and Refused, which both correspond to about 18% of the loans. This implies that most of the loans get approved and only some fraction of them do not. The least occurring type of contract status is Unused Offer which corresponds to just 1.61% of all the loans.

Univariate Analysis

NAME_CONTRACT_STATUS

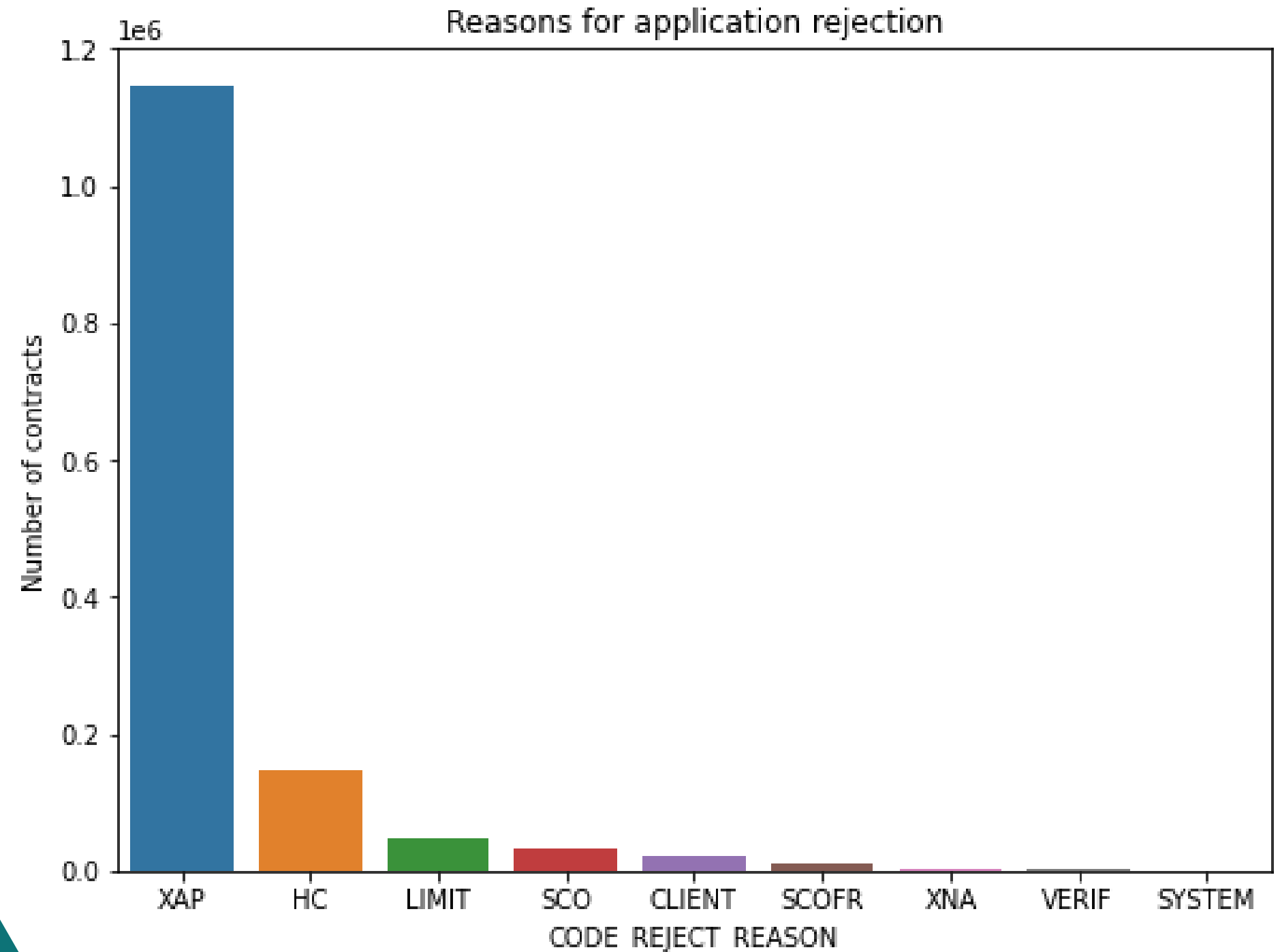
Looking at the subplot for percentage of defaulters, we see that the those loans which previously had Refused Status tend to have defaulted the highest in the current loans. They correspond to about 12% of Defaulters from that category. These are followed by Canceled Status which correspond to close to 9% of Default Rate. This behaviour is quite expected logically, as these people must have been refused due to not having adequate profile. The least default rate is observed for Contract Status of Approved.



Univariate Analysis

CODE_REJECT_REASON

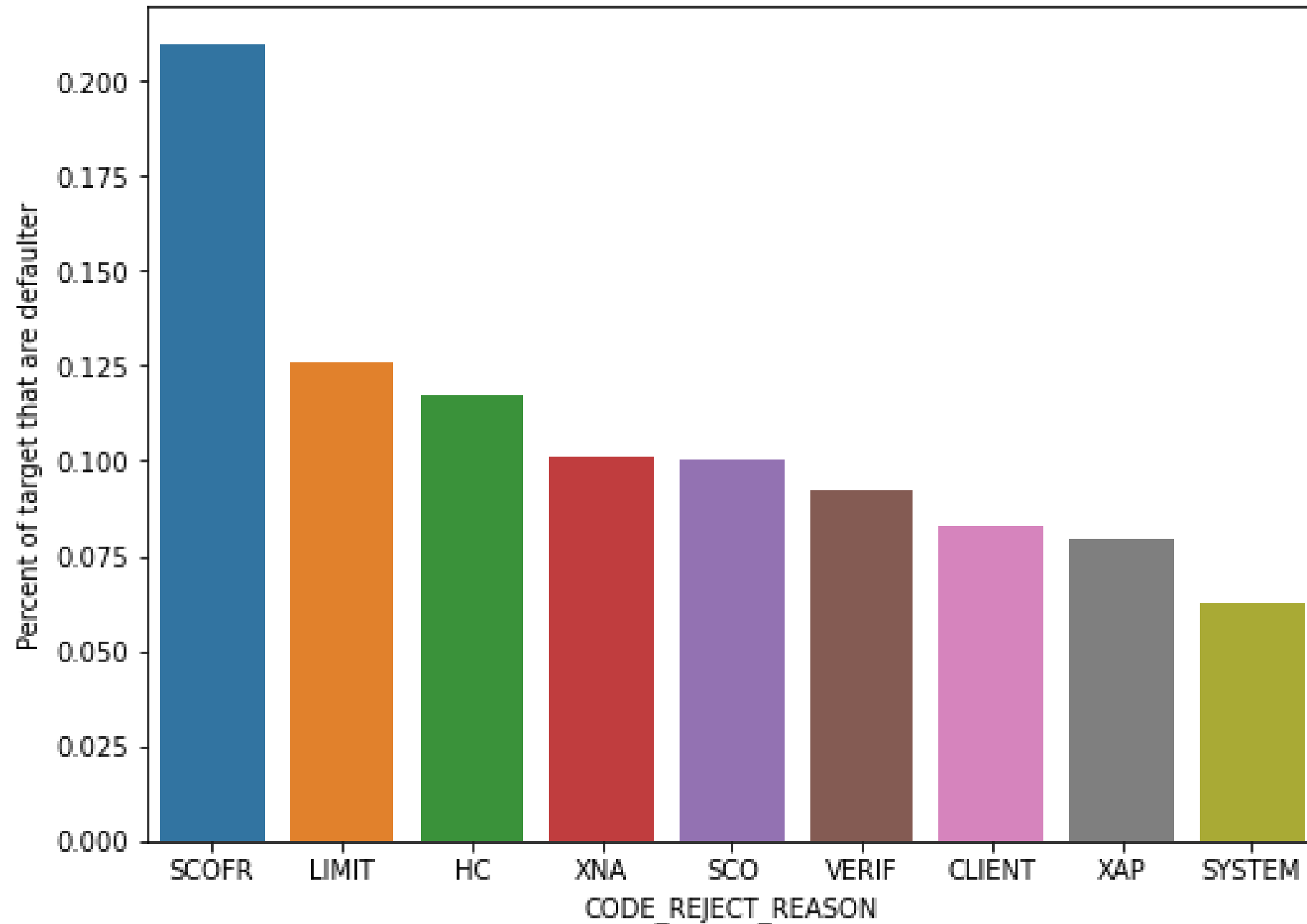
The most common type of reason of rejection is XAP, which is about ~81%. The other reasons form only a small part of the rejection reasons. HC is the second highest rejection reason with just 10.33% of occurrences.



Univariate Analysis

CODE_REJECT_REASON

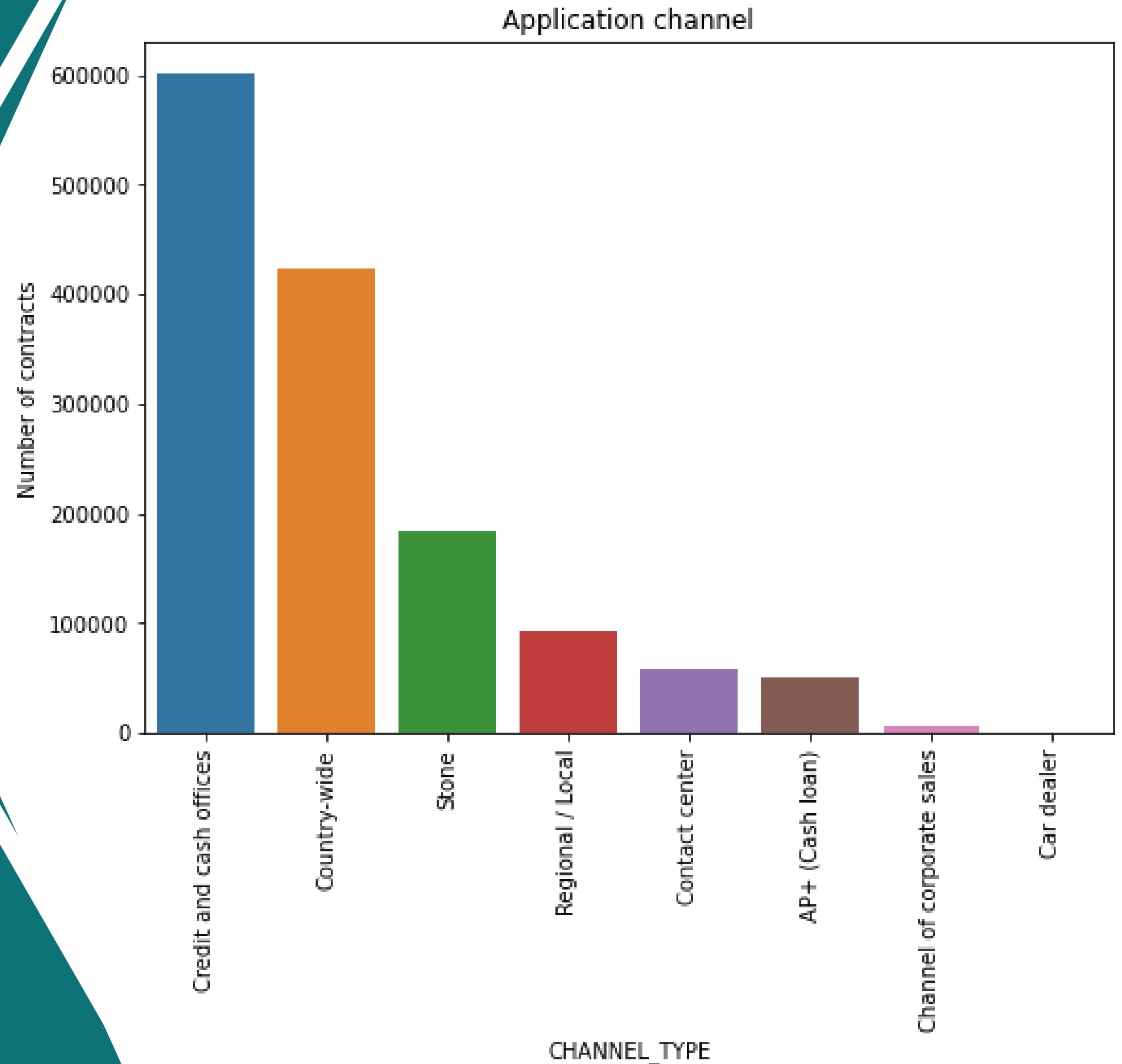
The distribution of percentage of defaulters for each category of CODE_REJECT_REASON is quite interesting. Those applicants who had their previous applications rejected by Code SCOFT have the highest percentage of Defaulters among them (~21%). This is followed by LIMIT and HC which have around 12.5% and 12% of Defaulters.



Univariate Analysis

CHANNEL_TYPE

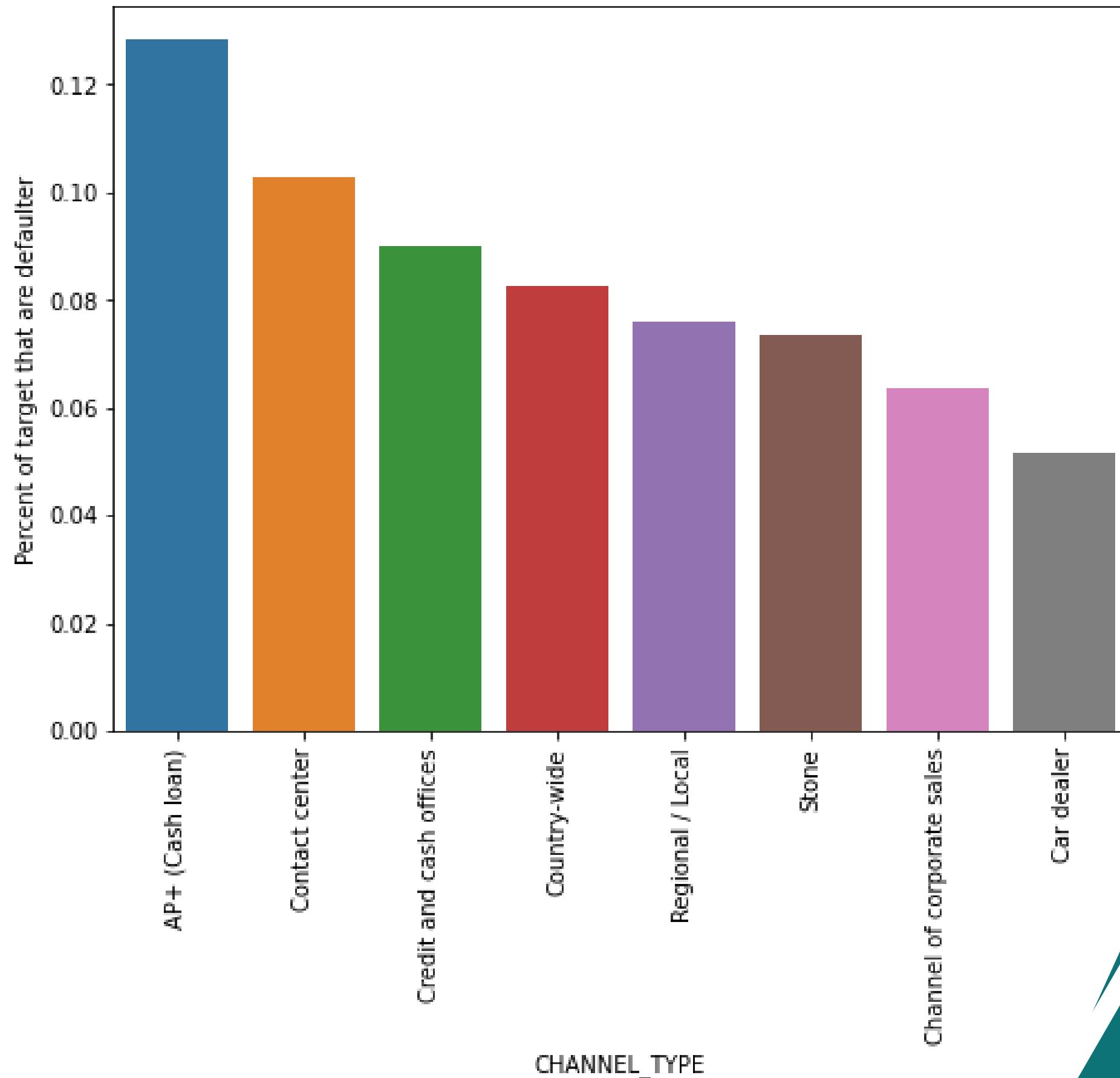
From the first subplot we see that most of the applications were acquired through the Credit and cash offices which were roughly 42.47% applications, which were followed by Country-wide channel corresponding to 29.93% applications. Rest of the channel types corresponded to only a select number of applications.



Univariate Analysis

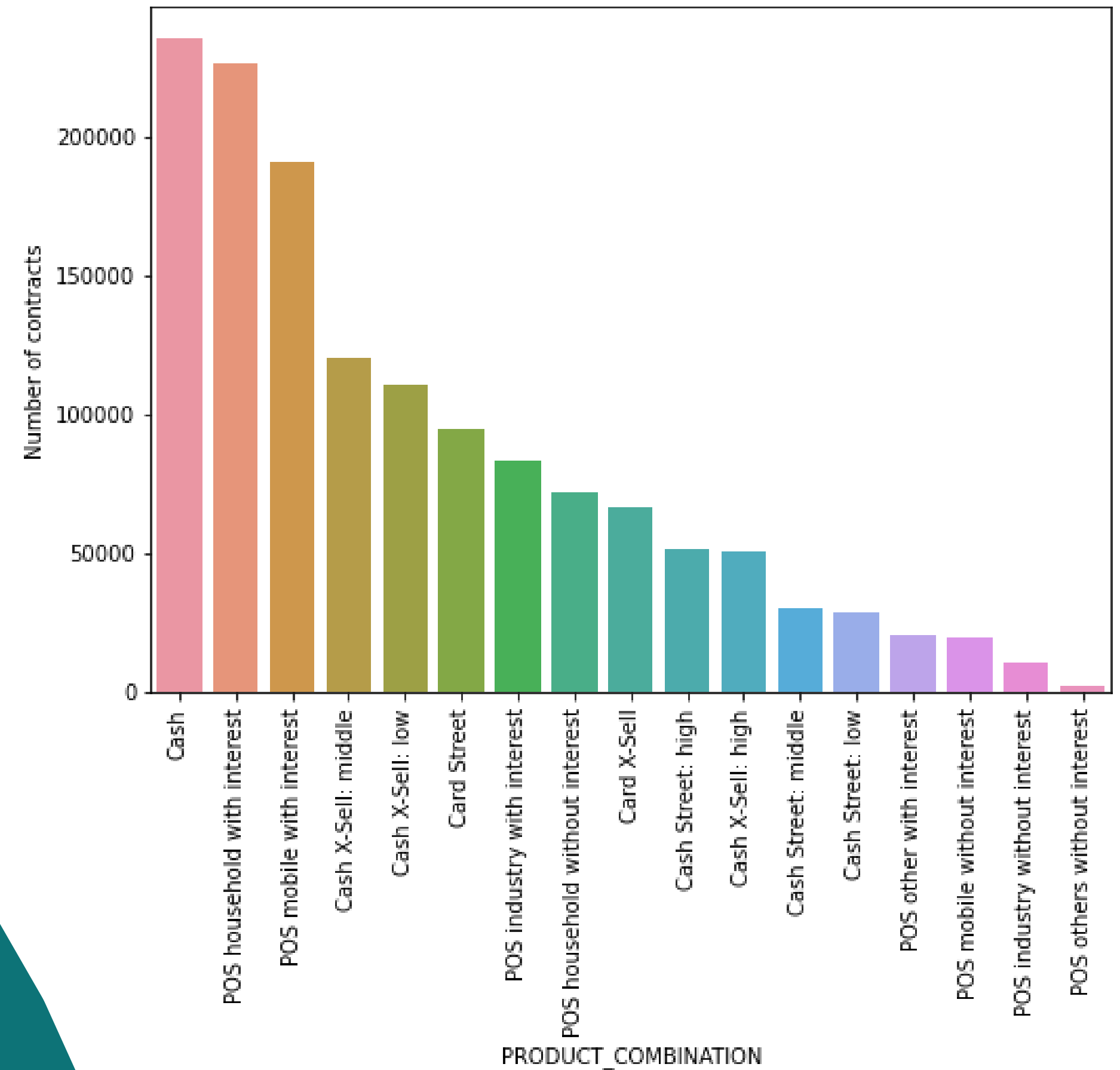
CHANNEL_TYPE

The highest Defaulting Percentage was seen among applications who had a channel type of AP+ (Cash loan) which corresponded to about 13% defaulters in that category. The rest of the channels had lower default percentages than this one. The channel Car Dealer showed a lowest Percentage of Defaulters in that category (only 5%).



Univariate Analysis

Product combination of the application



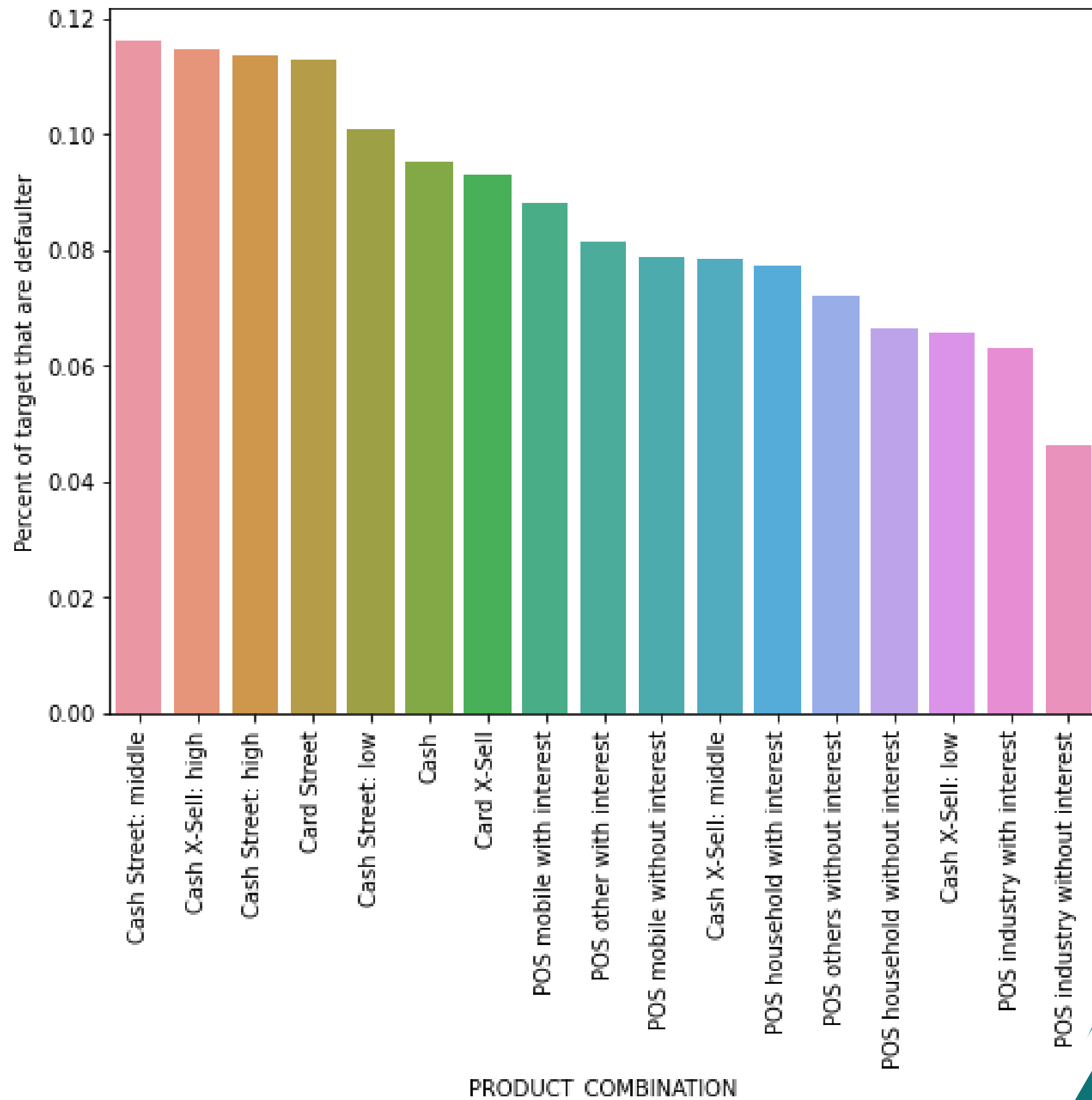
PRODUCT_COMBINATION

The 3 most common types of Product Combination are Cash, POS household with interest and POS mobile with interest. They correspond to roughly 50% of all the applications

Univariate Analysis

PRODUCT_COMBINATION

Looking at the Percentage of Defaulters per category plot, we see a highest defaulting tendency among Cash Street: mobile category, Cash X-sell: high, Cash Street: high and Card Street which all are near about 11-11.5% defaulters per category. The lowest Percentage of Defaulters are in the POS Industry without interest Category, which correspond to about 4.5% Defaulters.



installments_payments.csv

This table lists out the repayment history of each of the loan that the applicant had with Home Credit Group. The table contains features like the amount of instalment, how much did the client pay for each instalments, etc.

ASK_ID_CU NUM_1				
	14186	161674	1	
	1330831	151639	0	
4	2085231	193053	2	1
5	2452527	199697	1	3
6	2714724	167756	1	2
7	1137312	164489	1	12
8	2234264	184693	4	11
9	1818599	111420	2	4
10	2723183	112102	0	14
11	1413990	109741	1	4
12	1782554	106597	1	3
13	2558880	154793	1	5
	570206	147645	0	
	197273			

bureau_data.head(10)

	SK_ID_PREV	SK_ID_CURR	NUM_INSTALMENT_VERSION	NUM_INSTALMENT_NUMBER	DAYS_INSTALMENT	DAYS_ENTRY_PAYMENT	AMT_INSTALMENT	AMT_PAYMENT
0	1054186	161674	1.0	6	-1180.0	-1187.0	6948.360	6948.360
1	1330831	151639	0.0	34	-2156.0	-2156.0	1716.525	1716.525
2	2085231	193053	2.0	1	-63.0	-63.0	25425.000	25425.000
3	2452527	199697	1.0	3	-2418.0	-2426.0	24350.130	24350.130
4	2714724	167756	1.0	2	-1383.0	-1366.0	2165.040	2165.040
5	1137312	164489	1.0	12	-1384.0	-1417.0	5970.375	5970.375
6	2234264	184693	4.0	11	-349.0	-352.0	29432.295	29432.295
7	1818599	111420	2.0	4	-968.0	-994.0	17862.165	17862.165
8	2723183	112102	0.0	14	-197.0	-197.0	70.740	70.740
9	1413990	109741	1.0	4	-570.0	-609.0	14308.470	14308.470

bureau_data.shape
(13605401, 8)

Structure Investigation

```
1 install_payments.describe().T
```

	count	mean	std	min	25%	50%	75%	max
SK_ID_PREV	13605401.0	1.903365e+06	536202.905546	1000001.0	1434191.000	1896520.000	2369094.000	2843499.000
SK_ID_CURR	13605401.0	2.784449e+05	102718.310411	100001.0	189639.000	278685.000	367530.000	456255.000
NUM_INSTALLMENT_VERSION	13605401.0	8.566373e-01	1.035216	0.0	0.000	1.000	1.000	178.000
NUM_INSTALLMENT_NUMBER	13605401.0	1.887090e+01	26.664067	1.0	4.000	8.000	19.000	277.000
DAYS_INSTALLMENT	13605401.0	-1.042270e+03	800.946284	-2922.0	-1654.000	-818.000	-361.000	-1.000
DAYS_ENTRY_PAYMENT	13602496.0	-1.051114e+03	800.585883	-4921.0	-1662.000	-827.000	-370.000	-1.000
AMT_INSTALLMENT	13605401.0	1.705091e+04	50570.254429	0.0	4226.085	8884.080	16710.210	3771487.845
AMT_PAYMENT	13602496.0	1.723822e+04	54735.783981	0.0	3398.265	8125.515	16108.425	3771487.845

Structure Investigation

Correlation matrix



Total	Percent	Percent
DAYS_ENTRY_PAYMENT	2905	0.021352
AMT_PAYMENT	2905	0.021352
SK_ID_PREV	0	0.000000
SK_ID_CURR	0	0.000000
NUM_INSTALLMENT_VERSION	0	0.000000
NUM_INSTALLMENT_NUMBER	0	0.000000
DAYS_INSTALLMENT	0	0.000000
AMT_INSTALLMENT	0	0.000000

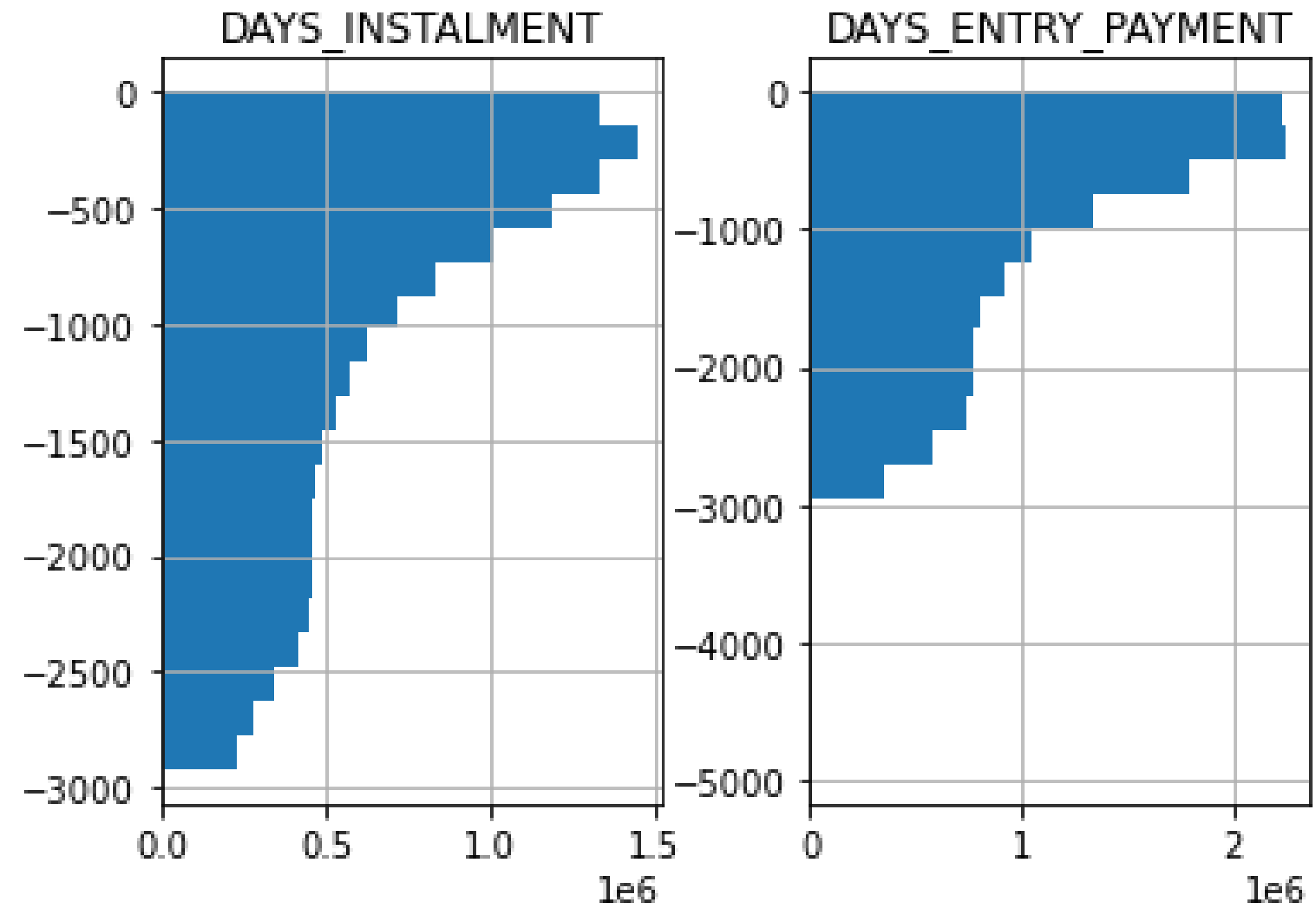
Variables interpretation

Column	Description
SK_ID_CURR	ID of loan in our sample
NUM_INSTALLMENT_VERSION	Version of installment calendar (0 is for credit card) of previous credit. Change of installment version from month to month signifies that some parameter of payment calendar has changed
NUM_INSTALLMENT_NUMBER	On which installment we observe payment
DAYS_INSTALLMENT	When the installment of previous credit was supposed to be paid (relative to application date of current loan)
DAYS_ENTRY_PAYMENT	When was the installments of previous credit paid actually (relative to application date of current loan)
AMT_INSTALLMENT	What was the prescribed installment amount of previous credit on this installment
AMT_PAYMENT	What the client actually paid on previous credit on this installment

Univariate Analysis

DAYS_INSTALMENT and DAYS_ENTRY_PAYMENT

- The histogram of prescribed installment day and actually paid day also suggests late payment of the customer



POS_CASH_balance.csv

This table contains the Monthly Balance Snapshots of previous Point of Sales and Cash Loans that the applicant had with Home Credit Group. The table contains columns like the status of contract, the number of installments left, etc.

	ASK_ID_CUMONTH		
	182943	-31	
	367990	-33	
4	397406	-32	12
5	269225	-35	48
6	334279	-35	36
7	342166	-32	12
8	204376	-38	48
9	153211	-35	36
10	112740	-31	12
11	274851	-32	24
12	287361	-32	12
13	237959	-39	12
	278261	-32	
	146161	-25	

pos_cash.head(10)

	SK_ID_PREV	SK_ID_CURR	MONTHS_BALANCE	CNT_INSTALMENT	CNT_INSTALMENT_FUTURE	NAME_CONTRACT_STATUS	SK_DPD	SK_DPD_DEF
0	1803195	182943	-31	48.0	45.0	Active	0	0
1	1715348	367990	-33	36.0	35.0	Active	0	0
2	1784872	397406	-32	12.0	9.0	Active	0	0
3	1903291	269225	-35	48.0	42.0	Active	0	0
4	2341044	334279	-35	36.0	35.0	Active	0	0
5	2207092	342166	-32	12.0	12.0	Active	0	0
6	1110516	204376	-38	48.0	43.0	Active	0	0
7	1387235	153211	-35	36.0	36.0	Active	0	0
8	1220500	112740	-31	12.0	12.0	Active	0	0
9	2371489	274851	-32	24.0	16.0	Active	0	0

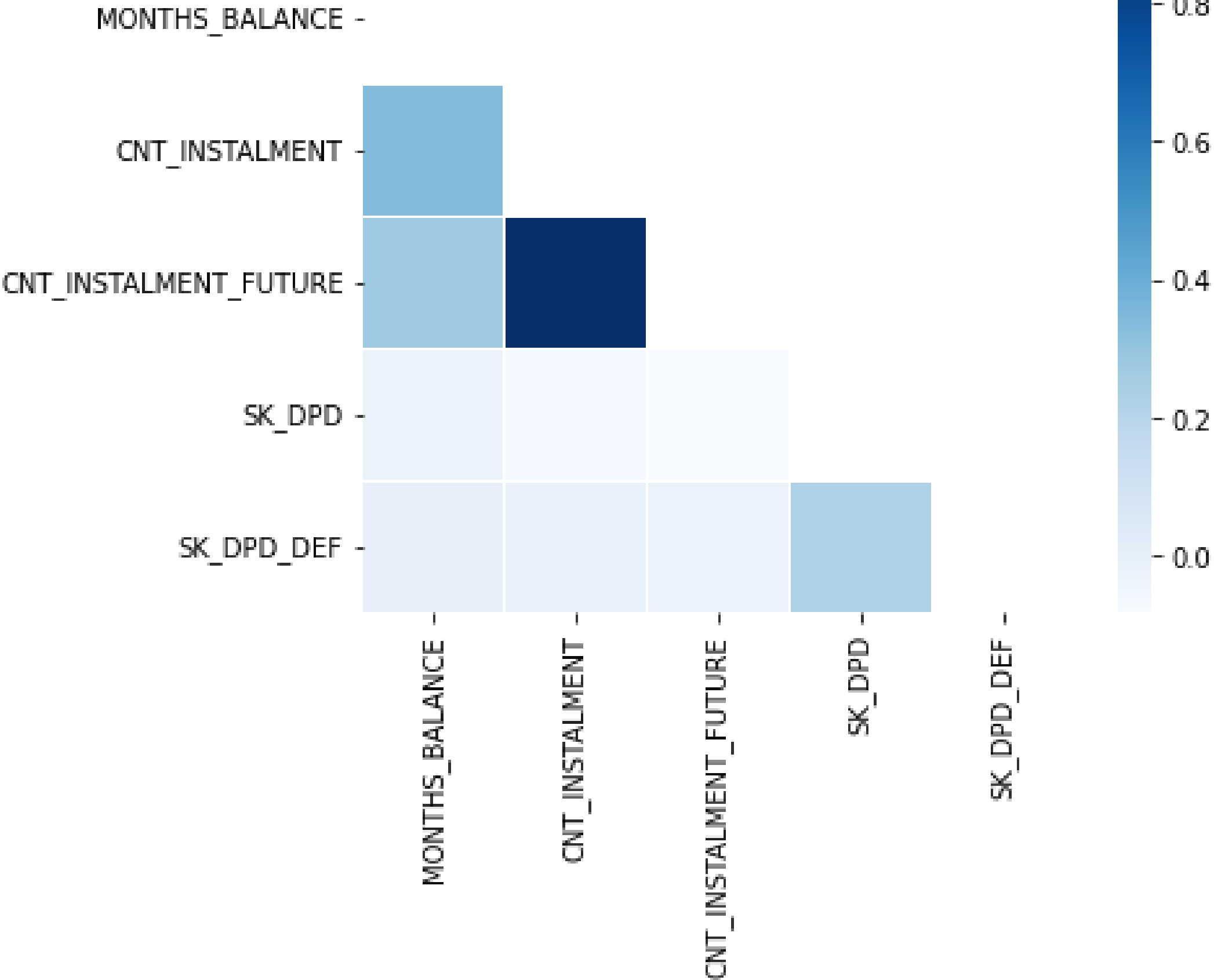
pos_cash.shape
(10001358, 8)

Structure Investigation

1	pos_cash.describe().T								
		count	mean	std	min	25%	50%	75%	max
	SK_ID_PREV	10001358.0	1.903217e+06	535846.530722	1000001.0	1434405.0	1896565.0	2368963.0	2843499.0
	SK_ID_CURR	10001358.0	2.784039e+05	102763.745090	100001.0	189550.0	278654.0	367429.0	456255.0
	MONTHS_BALANCE	10001358.0	-3.501259e+01	26.066570	-96.0	-54.0	-28.0	-13.0	-1.0
	CNT_INSTALMENT	9975287.0	1.708965e+01	11.995056	1.0	10.0	12.0	24.0	92.0
	CNT_INSTALMENT_FUTURE	9975271.0	1.048384e+01	11.109058	0.0	3.0	7.0	14.0	85.0
	SK_DPD	10001358.0	1.160693e+01	132.714043	0.0	0.0	0.0	0.0	4231.0
	SK_DPD_DEF	10001358.0	6.544684e-01	32.762491	0.0	0.0	0.0	0.0	3595.0

Structure Investigation

Correlation Heatmap for Numerical features



**Correlation
matrix**

Total	Percent	Percent
CNT_INSTALMENT_FUTURE	26087	0.260835
CNT_INSTALMENT	26071	0.260675
SK_ID_PREV	0	0.000000
SK_ID_CURR	0	0.000000
MONTHS_BALANCE	0	0.000000
NAME_CONTRACT_STATUS	0	0.000000
SK_DPD	0	0.000000
SK_DPD_DEF	0	0.000000

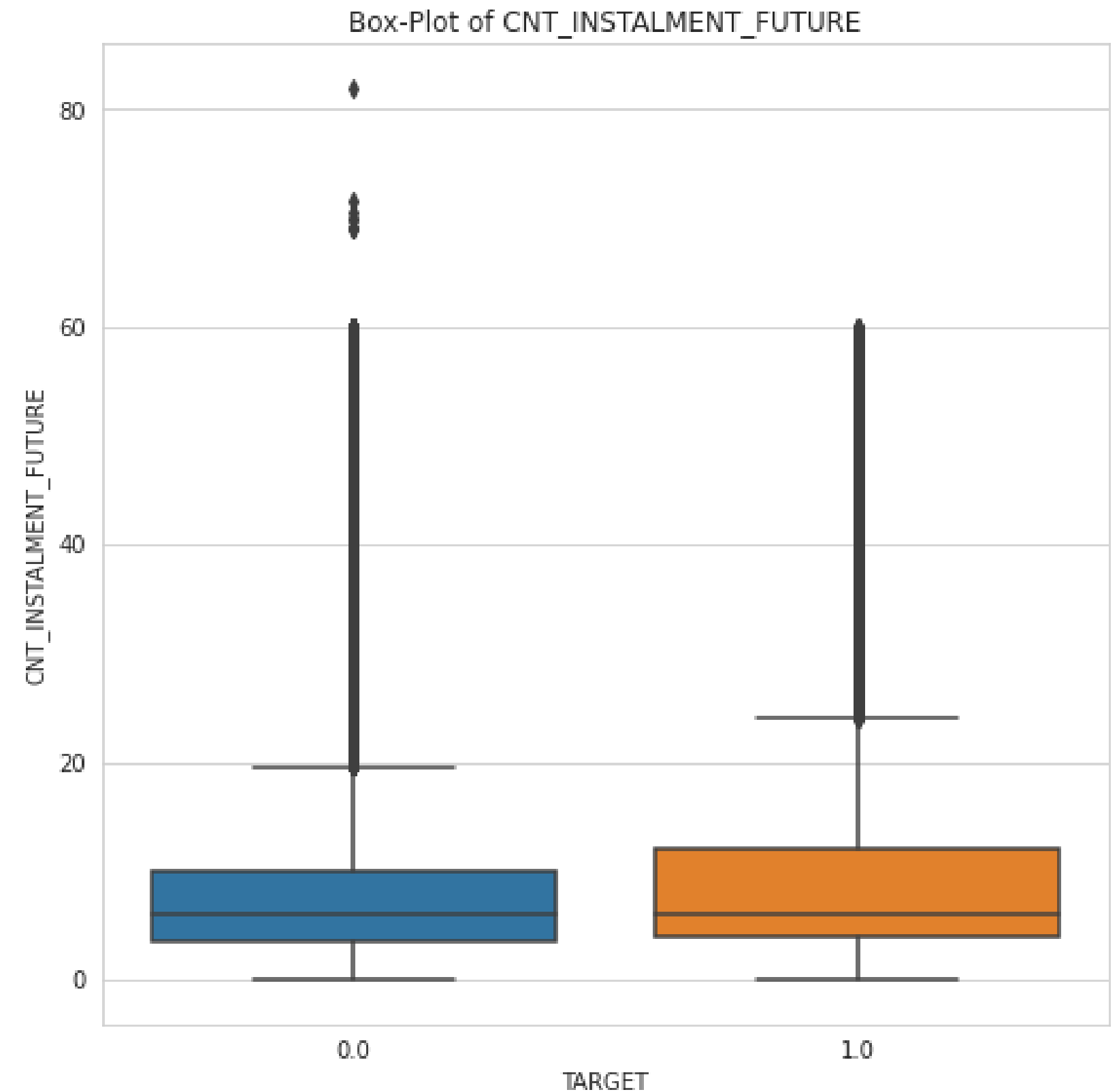
Variables interpretation

Column	Description
SK_ID_CURR	ID of loan in our sample
MONTHS_BALANCE	Month of balance relative to application date (-1 means the information to the freshest monthly snapshot, 0 means the information at application - often it will be the same as -1 as many banks are not updating the information to Credit Bureau regularly)
CNT_INSTALMENT	Term of previous credit (can change over time)
CNT_INSTALMENT_FUTURE	Installments left to pay on the previous credit
NAME_CONTRACT_STATUS	Contract status during the month
SK_DPD	DPD (days past due) during the month of previous credit
SK_DPD_DEF	DPD during the month with tolerance (debts with low loan amounts are ignored) of the previous credit

Univariate Analysis

CNT_INSTALLMENT_FUTURE

- Looking at the above box-plot for CNT_INSTALLMENT_FUTURE, we see that the percentile values >50% for Defaulters are usually higher than those of Non-Defaulters. Even the upper limit whisker for Defaulters is higher than that of Non-Defaulters. This suggests that the Defaulters tend to have more number of Installments remaining on their previous credits as compared to Non-Defaulters.



credit_card_balance.csv

This table consists of the monthly data related to any or multiple Credit Cards that the applicant had with the Home Credit Group. The table contains fields like balance, the credit limit, amount of drawings, etc. for each month of the credit card.

credit_card_balance				
ASK_ID_CUIMONTH				
	22384	378907	-6	
	2582071	363914	-1	63975..
4	1740877	371185	-7	31815.23
5	1389973	337855	-4	236572.1
6	1891521	126868	-1	453919.5
7	2646502	380010	-7	82903.82
8	1079071	171320	-6	353451.6
9	2095912	118650	-7	47962.13
10	2181852	367360	-4	291543.1
11	1235299	203885	-5	201261.2
12	1108284	209660	-7	102076.6
13	2740914	340339	-1	131669.7
14	1085699	302517	-4	147511.1
15	171537			

credit_card_balance.head(10)

	SK_ID_PREV	SK_ID_CURR	MONTHS_BALANCE	AMT_BALANCE	AMT_CREDIT_LIMIT_ACTUAL	AMT_DRAWINGS_ATM_CURRENT	AMT_DRAWINGS_CURRENT
0	2562384	378907	-6	56.970001	135000	0.0	877.500000
1	2582071	363914	-1	63975.554688	45000	2250.0	2250.000000
2	1740877	371185	-7	31815.224609	450000	0.0	0.000000
3	1389973	337855	-4	236572.109375	225000	2250.0	2250.000000
4	1891521	126868	-1	453919.468750	450000	0.0	11547.000000
5	2646502	380010	-7	82903.812500	270000	0.0	0.000000
6	1079071	171320	-6	353451.656250	585000	67500.0	67500.000000
7	2095912	118650	-7	47962.125000	45000	45000.0	45000.000000
8	2181852	367360	-4	291543.062500	292500	90000.0	289339.437500
9	1235299	203885	-5	201261.187500	225000	76500.0	111026.703125

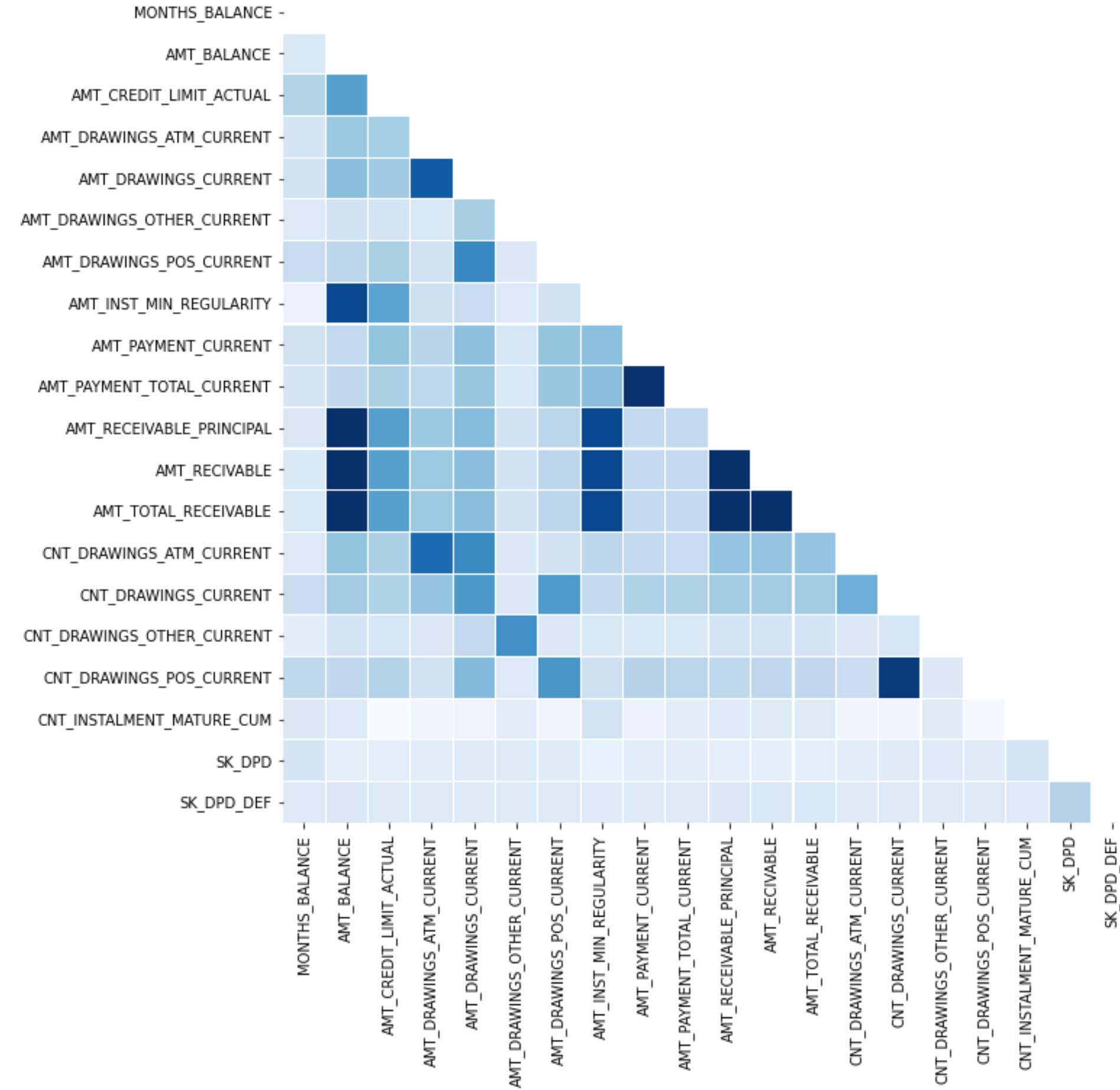
credit_card_balance.shape
(3840312, 12)

Structure Investigation

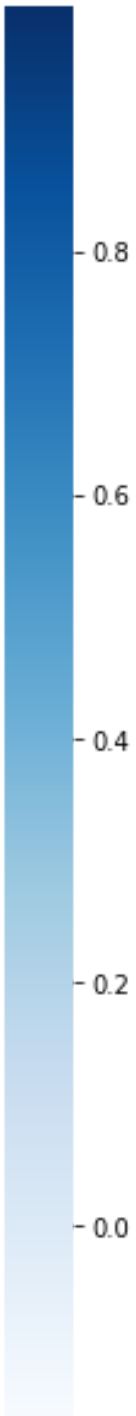
1	pos_cash.describe().T								
		count	mean	std	min	25%	50%	75%	max
	SK_ID_PREV	10001358.0	1.903217e+06	535846.530722	1000001.0	1434405.0	1896565.0	2368963.0	2843499.0
	SK_ID_CURR	10001358.0	2.784039e+05	102763.745090	100001.0	189550.0	278654.0	367429.0	456255.0
	MONTHS_BALANCE	10001358.0	-3.501259e+01	26.066570	-96.0	-54.0	-28.0	-13.0	-1.0
	CNT_INSTALMENT	9975287.0	1.708965e+01	11.995056	1.0	10.0	12.0	24.0	92.0
	CNT_INSTALMENT_FUTURE	9975271.0	1.048384e+01	11.109058	0.0	3.0	7.0	14.0	85.0
	SK_DPD	10001358.0	1.160693e+01	132.714043	0.0	0.0	0.0	0.0	4231.0
	SK_DPD_DEF	10001358.0	6.544684e-01	32.762491	0.0	0.0	0.0	0.0	3595.0

Structure Investigation

Correlation Heatmap for Numerical features



Correlation
matrix



	Total	Percent
AMT_PAYMENT_CURRENT	767988	19.998063
AMT_DRAWINGS_ATM_CURRENT	749816	19.524872
CNT_DRAWINGS_POS_CURRENT	749816	19.524872
AMT_DRAWINGS_OTHER_CURRENT	749816	19.524872
AMT_DRAWINGS_POS_CURRENT	749816	19.524872
CNT_DRAWINGS_OTHER_CURRENT	749816	19.524872
CNT_DRAWINGS_ATM_CURRENT	749816	19.524872
CNT_INSTALMENT_MATURE_CUM	305236	7.948208
AMT_INST_MIN_REGULARITY	305236	7.948208
SK_ID_PREV	0	0.000000
AMT_TOTAL_RECEIVABLE	0	0.000000
SK_DPD	0	0.000000

	Total	Percent
NAME_CONTRACT_STATUS	0	0.000000
CNT_DRAWINGS_CURRENT	0	0.000000
AMT_PAYMENT_TOTAL_CURRENT	0	0.000000
AMT_RECIVABLE	0	0.000000
AMT_RECEIVABLE_PRINCIPAL	0	0.000000
SK_ID_CURR	0	0.000000
AMT_DRAWINGS_CURRENT	0	0.000000
AMT_CREDIT_LIMIT_ACTUAL	0	0.000000
AMT_BALANCE	0	0.000000
MONTHS_BALANCE	0	0.000000
SK_DPD_DEF	0	0.000000

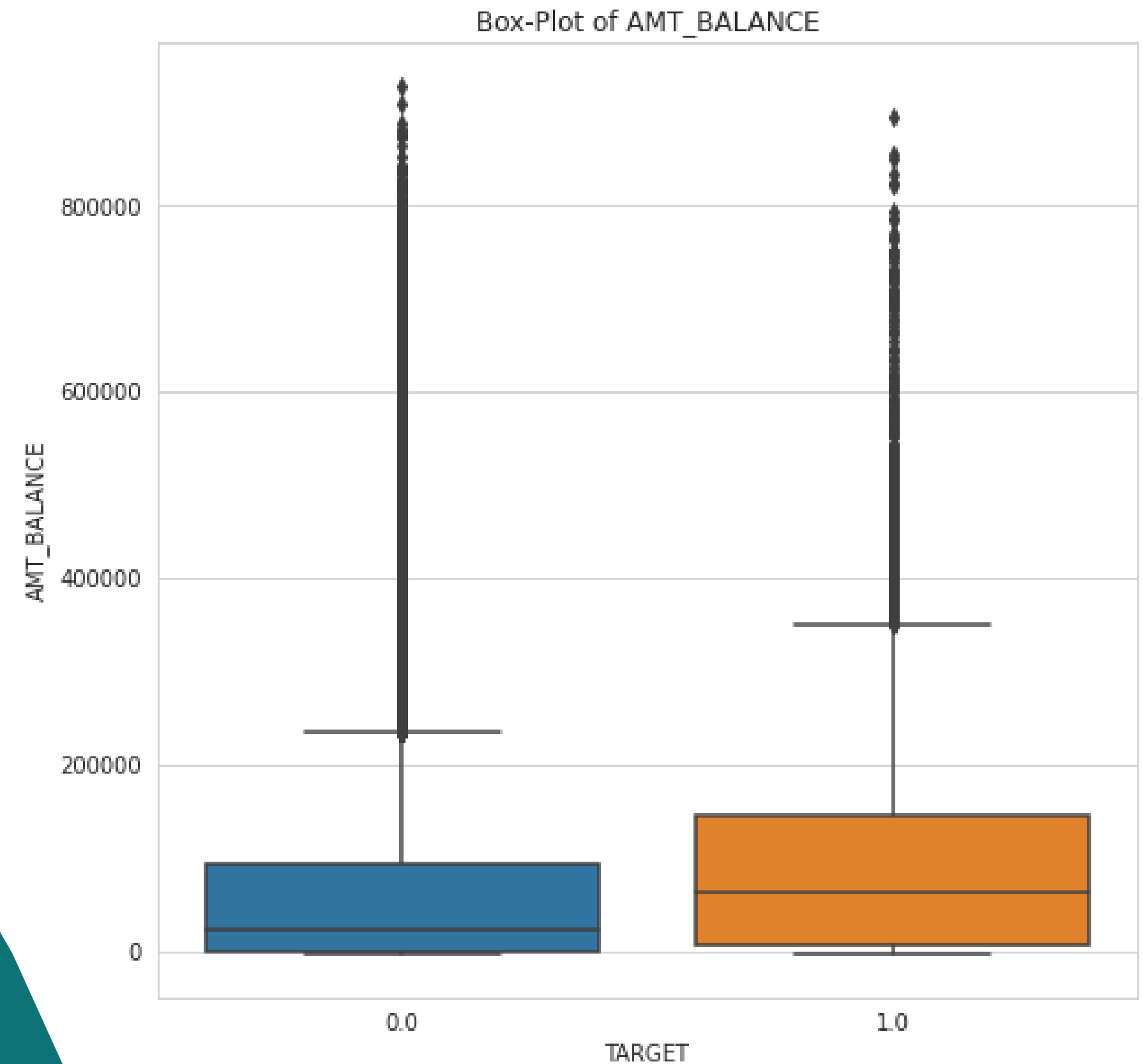
Variables interpretation

Column	Description
SK_ID_CURR	ID of loan in our sample
MONTHS_BALANCE	Month of balance relative to application date (-1 means the freshest balance date)
AMT_BALANCE	Balance during the month of previous credit
AMT_CREDIT_LIMIT_ACTUAL	Credit card limit during the month of the previous credit
AMT_INST_MIN_REGULARITY	Minimal installment for this month of the previous credit
AMT_PAYMENT_CURRENT	How much did the client pay during the month on the previous credit
AMT_PAYMENT_TOTAL_CURRENT	How much did the client pay during the month in total on the previous credit
AMT_RECEIVABLE_PRINCIPAL	Amount receivable for principal on the previous credit
AMT_RECIVABLE	Amount receivable on the previous credit
AMT_TOTAL_RECEIVABLE	Total amount receivable on the previous credit
NAME_CONTRACT_STATUS	Contract status (active signed,...) on the previous credit

Univariate Analysis

AMT_BALANCE

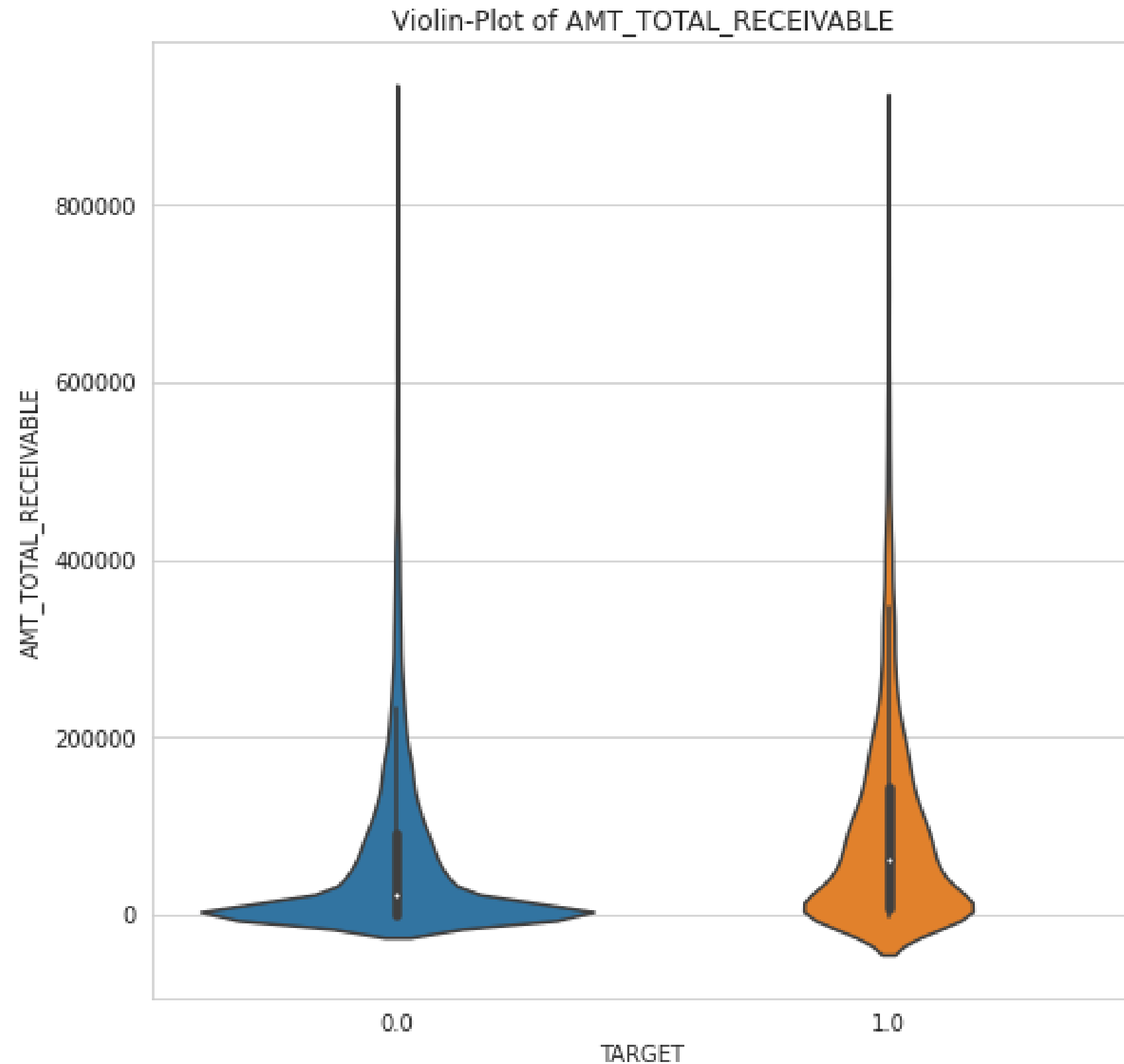
- It can be seen that the Defaulters have a higher value of AMT_BALANCE as compared to Non-Defaulters. They show a higher values of all the quantiles and even the whiskers. This could imply that the Credit amount for Defaulters could also be relatively higher as compared to Non-Defaulters.
- We see that the Defaulters here too appeared to have a higher minimum installment each month as compared to Non-Defaulters. This usually tells about the spending and borrowing habit of the people. The defaulters show a higher spending and borrowing habits as compared to Non-Defaulters.



Univariate Analysis

AMT_TOTAL_RECEIVABLE

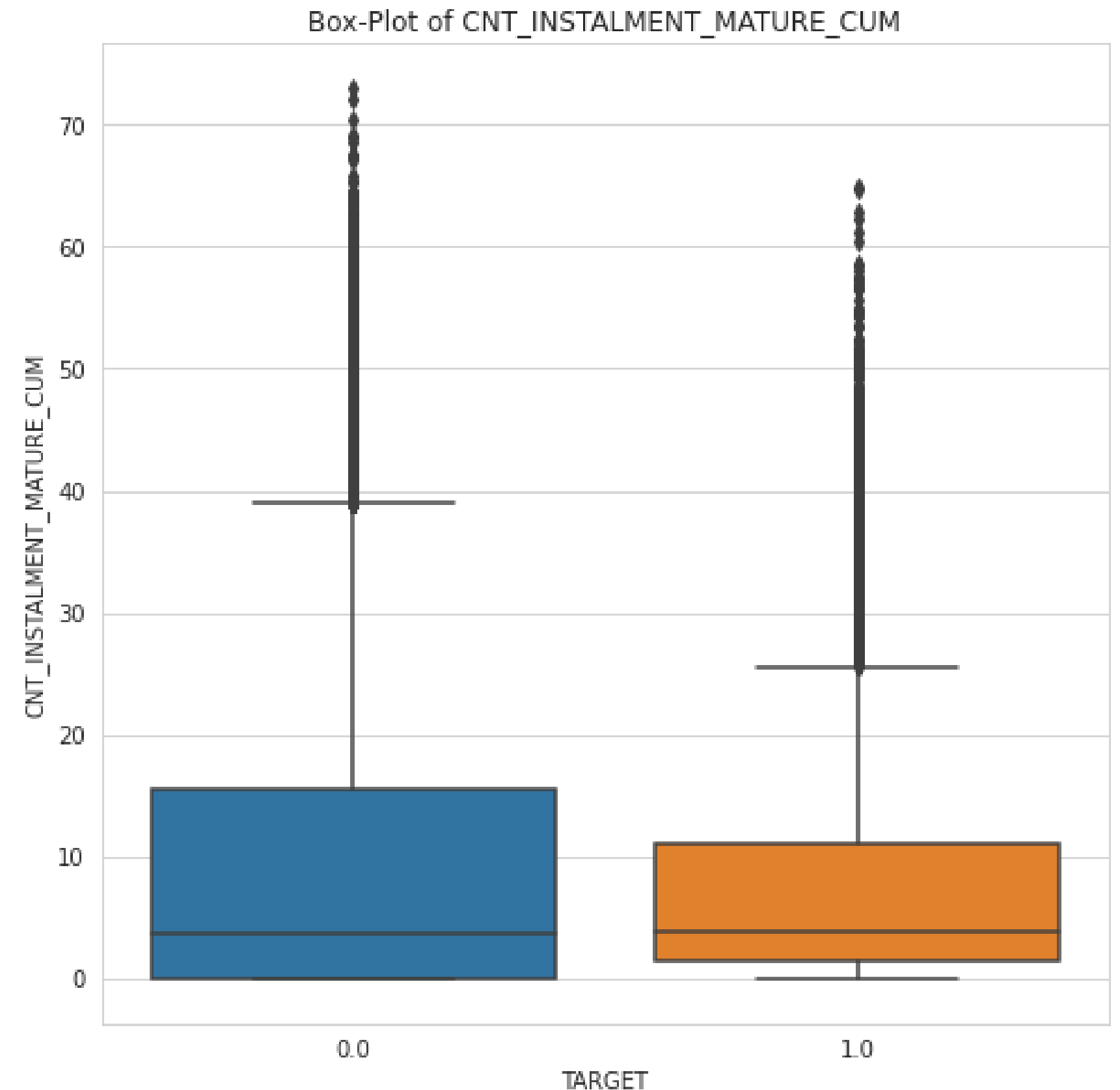
Looking at the box plot of AMT_TOTAL_RECEIVABLE, we see a similar behaviour as seen with other amounts as well, which is that the Defaulter usually had higher Amount Receivable on their previous credit, which may imply the higher amounts of credits that they may have taken. The PDF also shows a very higher peak at lower amounts for Non-Defaulter as compared to Defaulters.



Univariate Analysis

CNT_INSTALMENT_MATURE_CUM

- From the above plot, we see a very interesting behaviour. This plot shows that the Non-Defaulters usually had higher range of values for the number of installments paid as compared to Defaulters. This might show the defaulting behaviour, where in the defaulters usually would pay fewer number of installments on their previous credit.



Merging

application_{train|test}.csv

- Main tables – our train and test samples
- Target (binary)
- Info about loan and loan applicant at application time

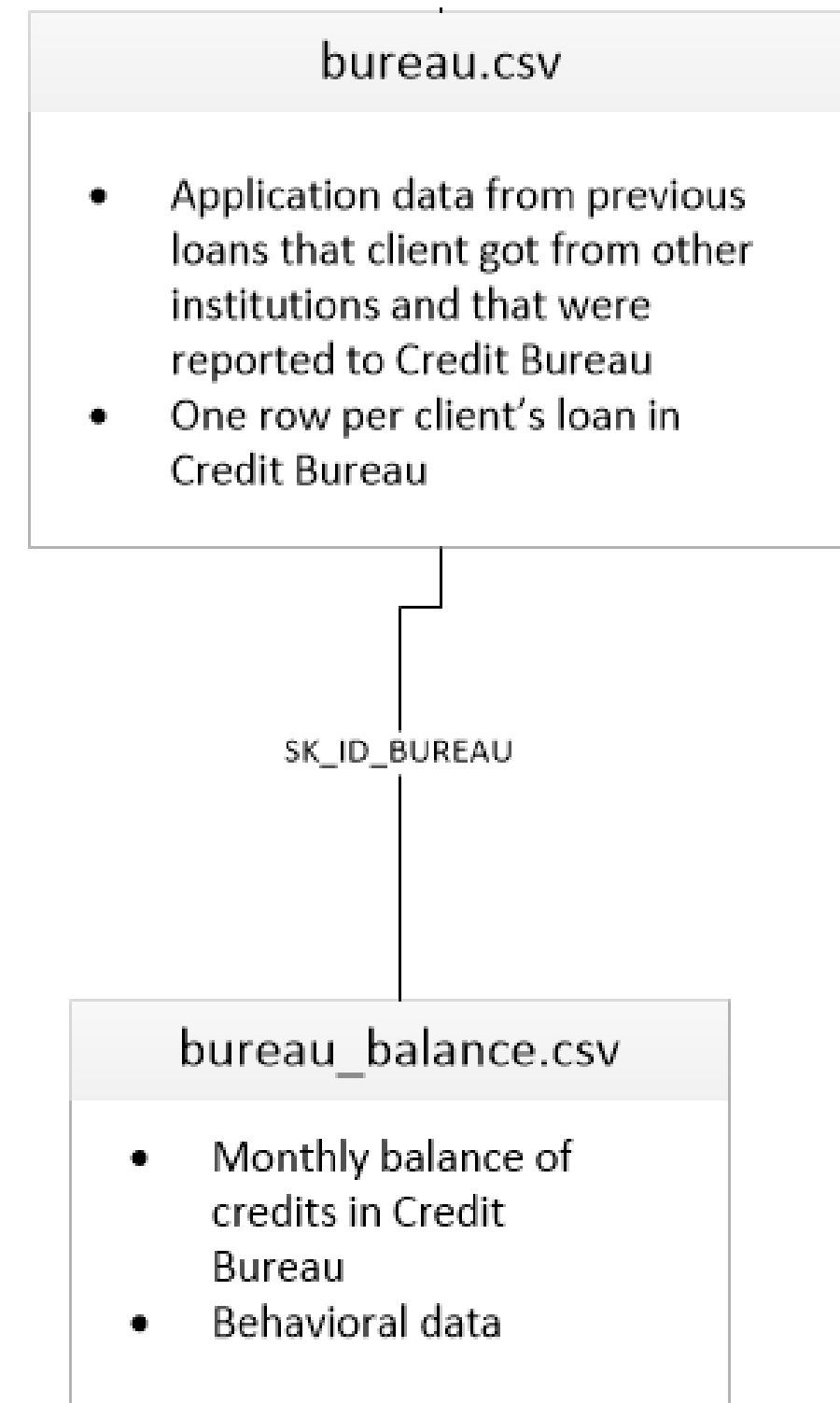
Merging application_train with application_test

Both datasets have exactly the same format with only TARGET column being present in train set as the only difference

Merging bureau.csv with bureau_balance.csv

Before we merge data with bureau, we need to merge bureau dataframe with related information in bureau_balance file

- Collapse bureau_balance dataframe to mean values grouped by SK_ID_BUREAU
- Merge this with bureau dataframe
- Collapse bureau dataframe to mean values grouped by SK_ID_CURR

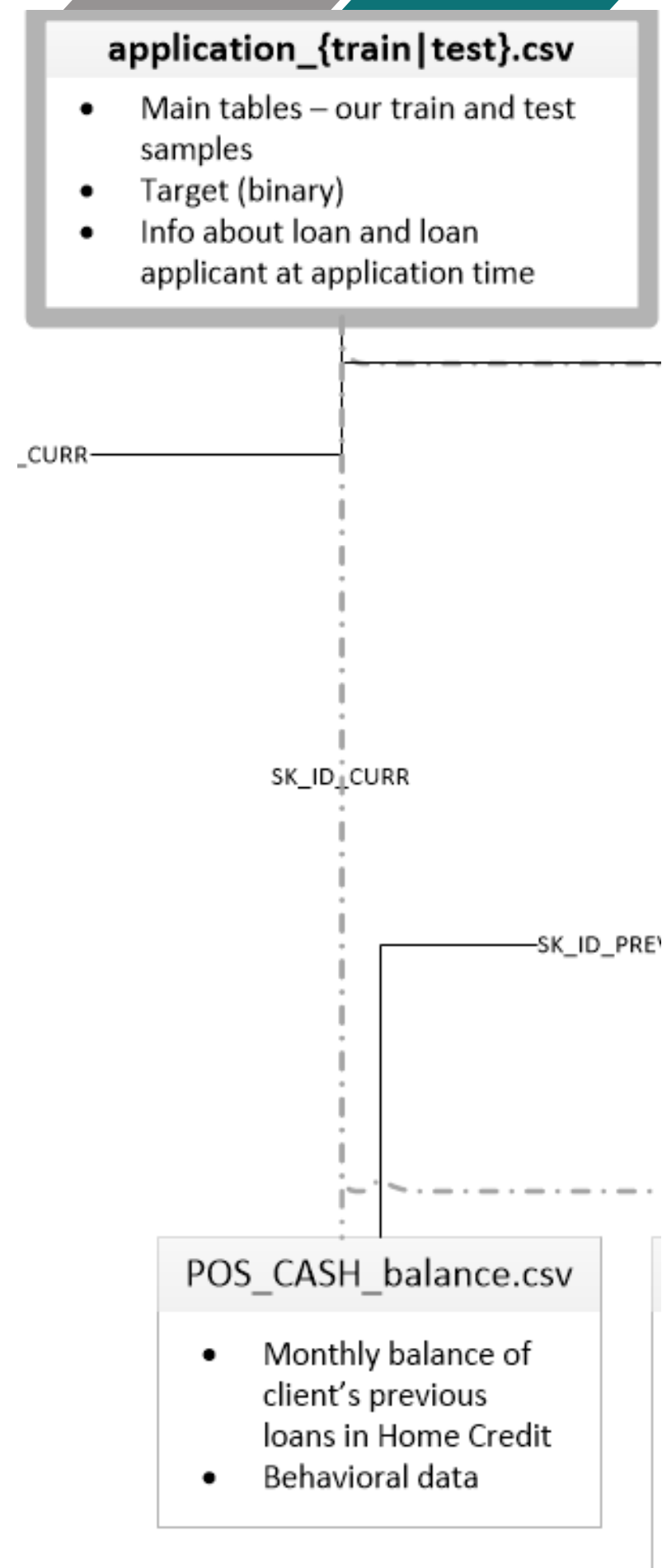


Merging

Merging with POS_CASH_balance.csv

monthly balance snapshots of previous Point Of Sale s and cash loans that the applicant had with Home Credit (one row for each month of history)

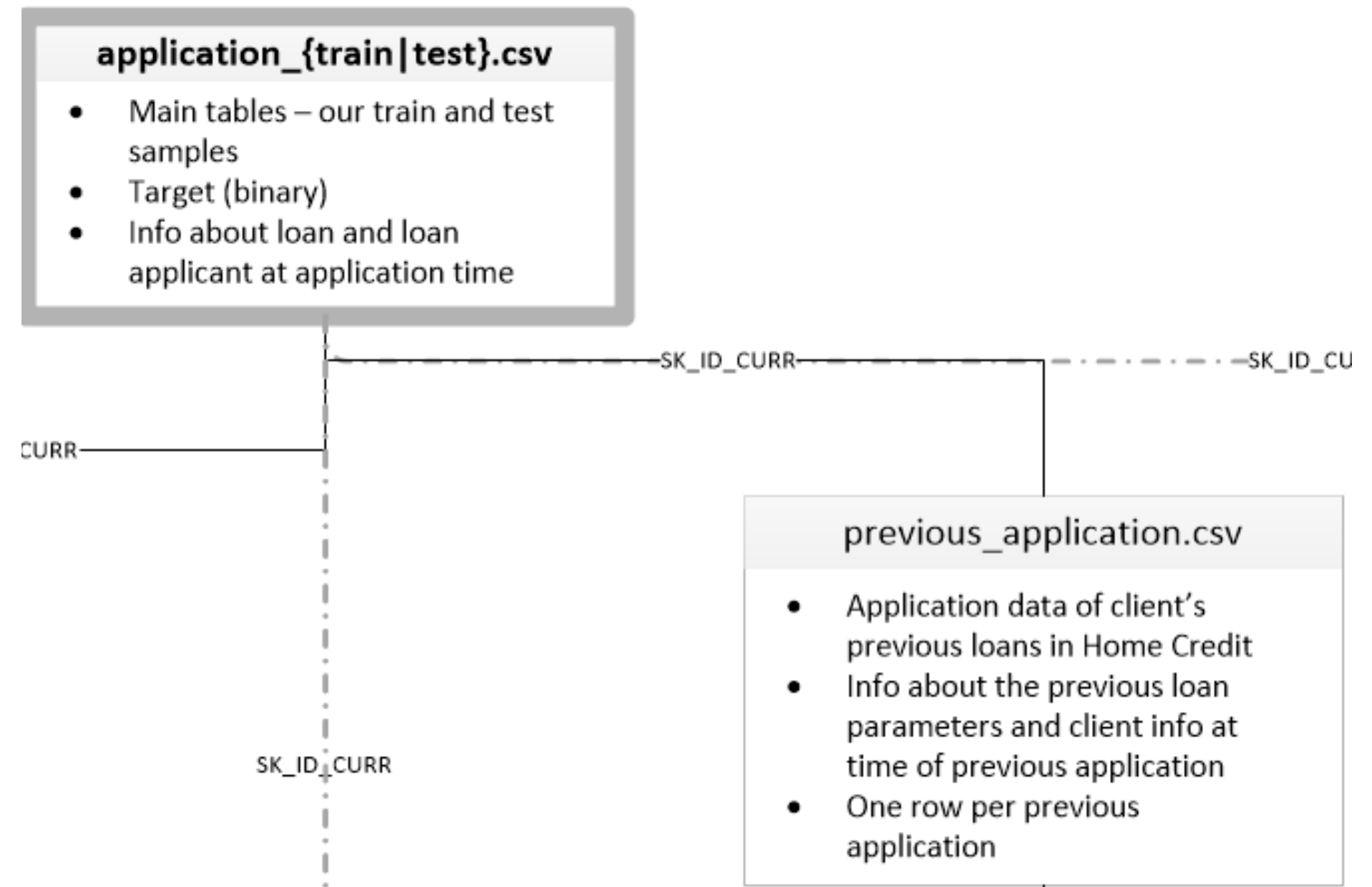
- Use key SK_ID_CURR to map with the main dataframe



Merging with previous_application.csv

previous_application reflects clients' previous applications for loans to Home Credit.

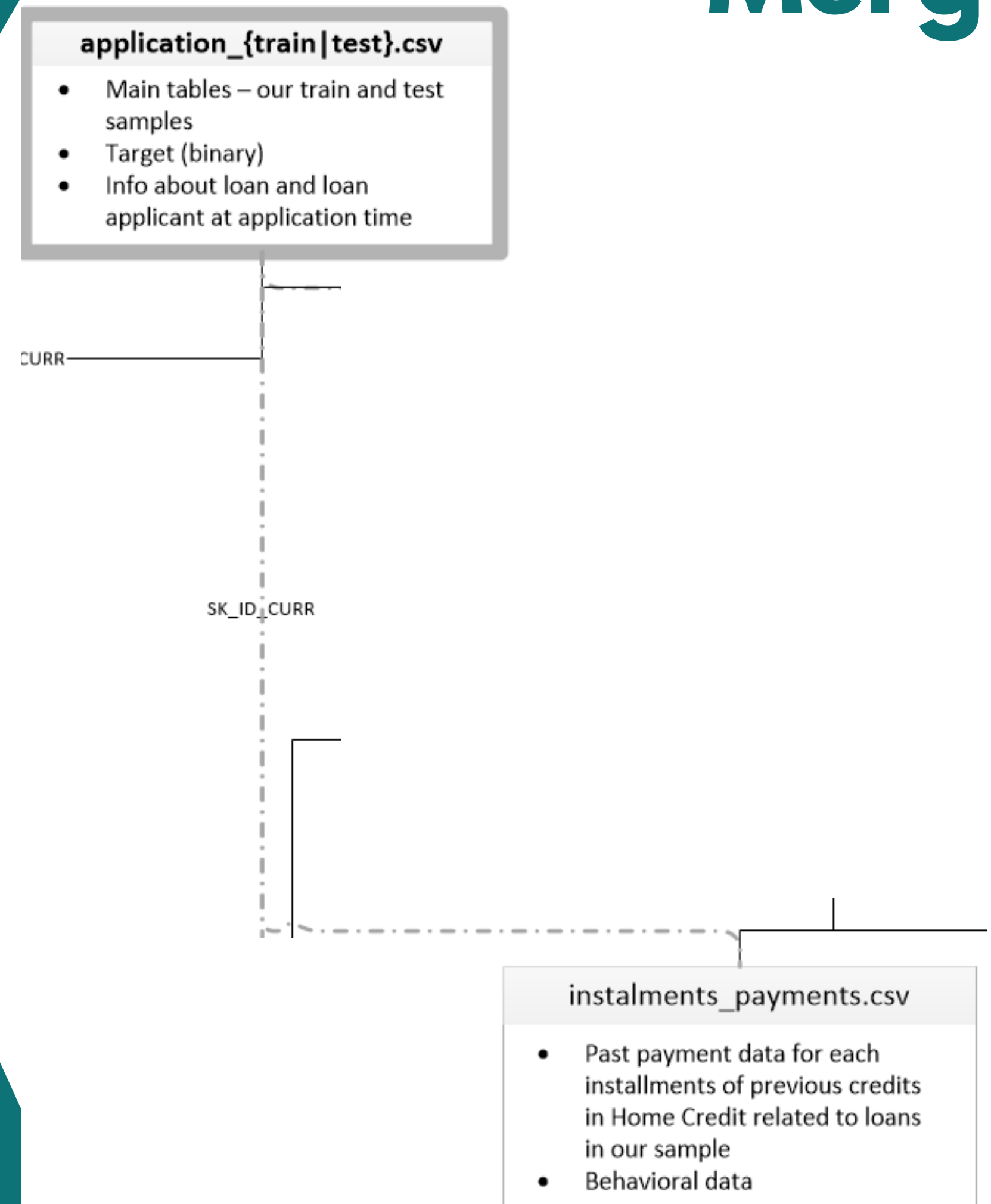
- Use key SK_ID_CURR to map with the main dataframe



Merging with instalments_payments.csv

repayment history for previous credits with Home Credit (one row for each payment)

- Use key SK_ID_CURR to map with the main dataframe



Merging

Merging with credit_card_balance

- Monthly balance snapshots of applicant's credit cards (one row for each month of history) Use key
- SK_ID_CURR to map with the main dataframe

application_{train|test}.csv

- Main tables – our train and test samples
- Target (binary)
- Info about loan and loan applicant at application time

SK_ID_CURR

SK_ID_CURR

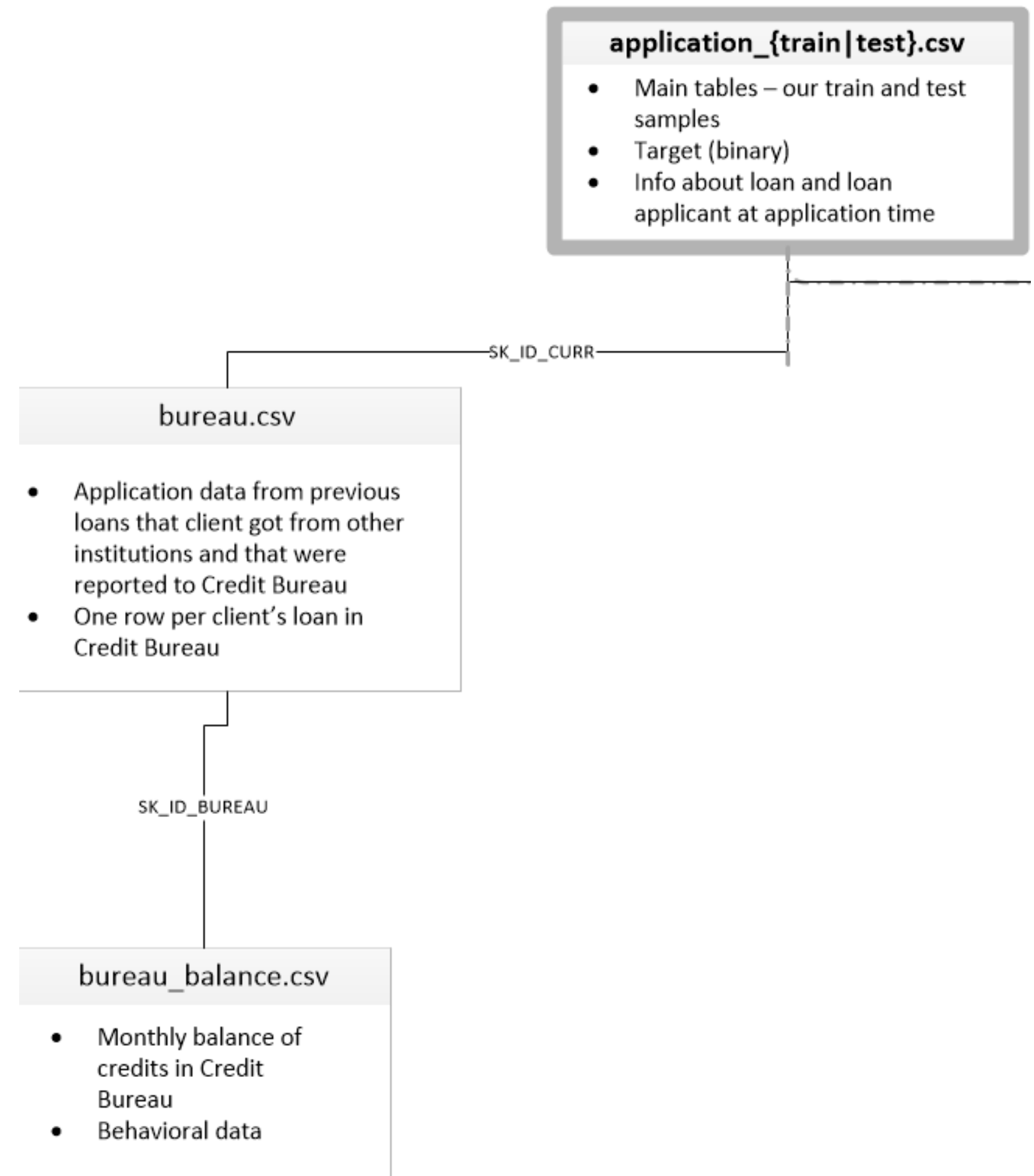
credit_card_balance.csv

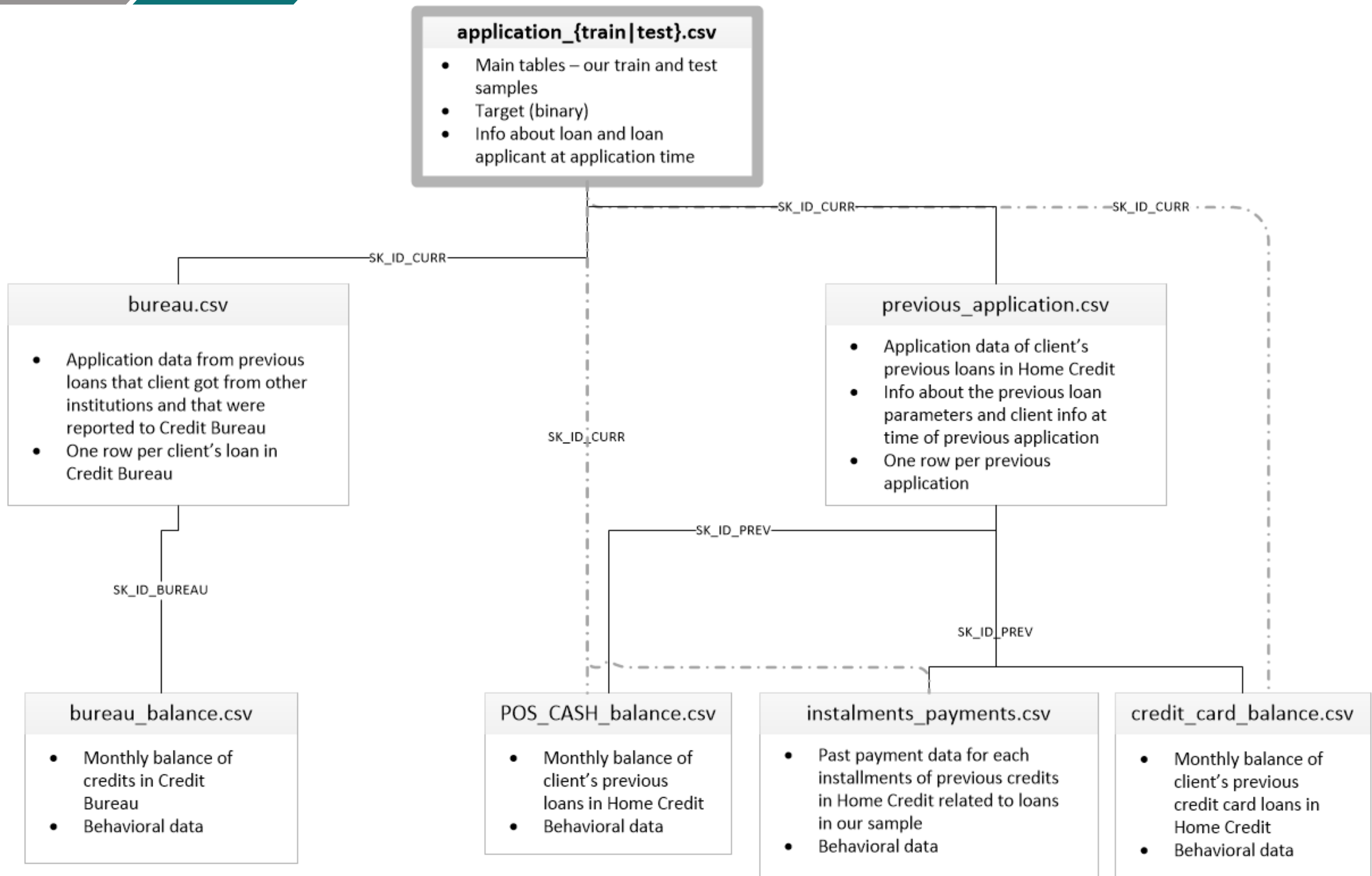
- Monthly balance of client's previous credit card loans in Home Credit
- Behavioral data

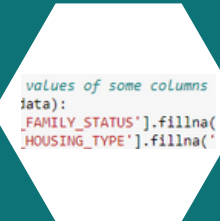
Merging

Merging with the bureau branch

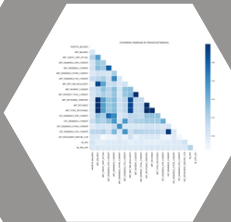
Use key SK_ID_CURR to map with the main dataframe







Removing attributes
having $> 60\%$ missing
values and fillna



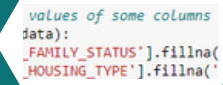
Removing numerical
attribute having high
correlation ($> 90\%$)



Drop columns with
percentage of unique
values $> 80\%$

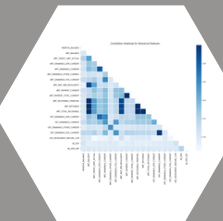
Cleaning the data

Cleaning the data



```
values of some columns  
data):  
    'FAMILY_STATUS'].fillna(  
    'HOUSING_TYPE'].fillna('
```

Fill "0", "Unknown" in
numeric, categorical
columns respectively

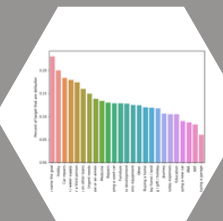


Handling
outliers



```
_DOCUMENT_2', 'FLAG_DOCU  
_DOCUMENT_7', 'FLAG_DOCU  
_DOCUMENT_11', 'FLAG_DOCU  
_DOCUMENT_14', 'FLAG_DOCU  
_DOCUMENT_17', 'FLAG_DOCU  
_DOCUMENT_20', 'FLAG_DOCU  
REQ_CREDIT_BUREAU_DAY',  
REQ_CREDIT_BUREAU_MON',  
CHILDREN']
```

Encoding categorical
variables



Visualize
features