Exploratory Data Analysis (EDA) on Loan Applications

A case study in Risk Analytics in Banking and Financial Services

BY NEHA MOHD SHAHID

PROBLEM STATMENT

When a company issues a loan, there are two main risks:

- 1. Approving a Loan for a Risky Applicant: If the applicant is likely to default, the company could incur financial losses.
- 2. Rejecting a Loan for a Safe Applicant: If the applicant is likely to repay but the loan is denied, the company misses out on potential business.

By conducting Exploratory Data Analysis (EDA) on the company's data, we can analyze patterns and help mitigating these risks.

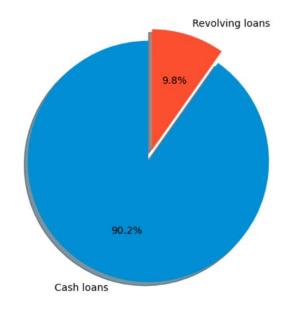
Univariate Analysis

Analysis of Loan Types in Relation to Default Risk

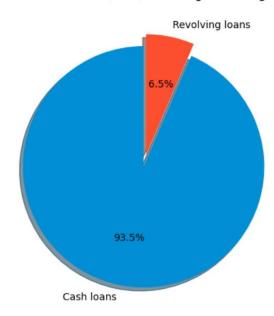
- Cash loans are the majority in both categories, making up 90.2% of non-defaulted loans and 93.5% of defaulted loans.
- The higher percentage of revolving loans in the non-defaulted category suggests that these loans are less likely to default compared to cash loans.
- The decrease in the proportion of revolving loans in the defaulted category (from 9.8% to 6.5%) may indicate that revolving loans are associated with better repayment behavior or lower default risk.

Loan identification analysis over target values (NAME CONTRACT TYPE)





Loan identification (Cash/Revolving) over Target (1)

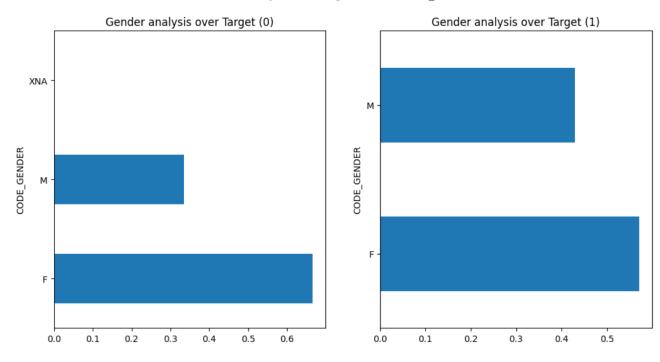


Gender Analysis in Relation to Loan Default Risk

Female borrowers tend to default less frequently compared to their male counterparts as their proportion decreases from 65% in the non-defaulted category to 55% in the defaulted category.

The data suggests that gender may be a relevant factor in assessing the risk of loan default. Female borrowers appear to be slightly lower risk, while male borrowers show a higher tendency toward default.

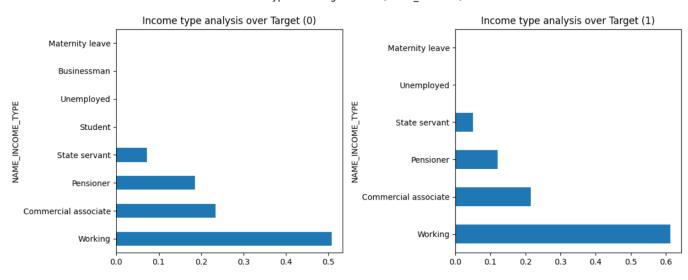
Gender analysis over target values (CODE_GENDER)



Income Type in Relation to Loan Default Risk

- Working individuals proportion is slightly higher among defaulted loans (around 60%) compared to non-defaulted loans (around 50%). It indicates they are more likely to default.
- Pensioners tend to default less frequently compared to other target values as their slightly proportion decreases from 22% in the non-defaulted category to 18% in the defaulted category.
- Pensioners are more likely to pay the loans

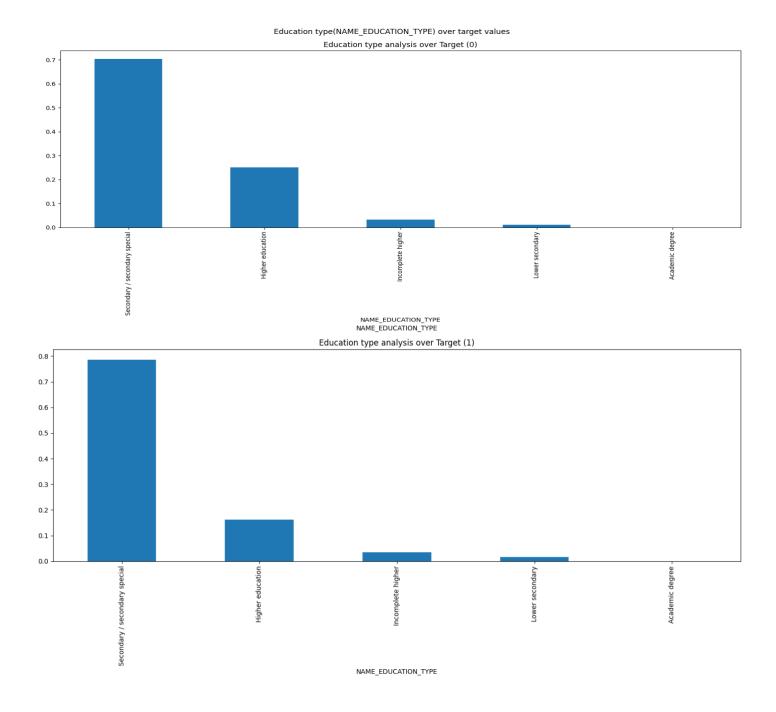
Income type over target values (CODE_GENDER)



Education Type and Payment Difficulties

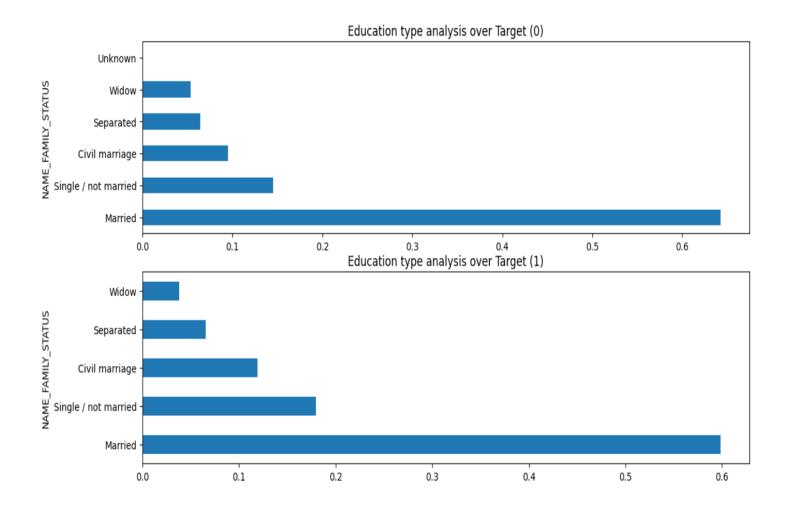
- Clients with "Secondary / Secondary Special" education make up the majority in both target groups.

 However, this proportion is slightly higher in clients with payment difficulties (Target 1).
- Higher education clients represent a smaller proportion in both groups, with a more noticeable presence in clients without payment difficulties (Target 0).
- Clients with higher education are more likely to repay loans



Family Status and Payment Difficulties

- Married clients are the most common in both groups (Target 0 and Target 1), but slightly more prevalent among those without payment difficulties (Target 0).
- Single/not married clients are the second-largest group in both targets, with a higher proportion in clients with payment difficulties (Target 1).
- Widowed clients are small minority, but are most likely to repay loans



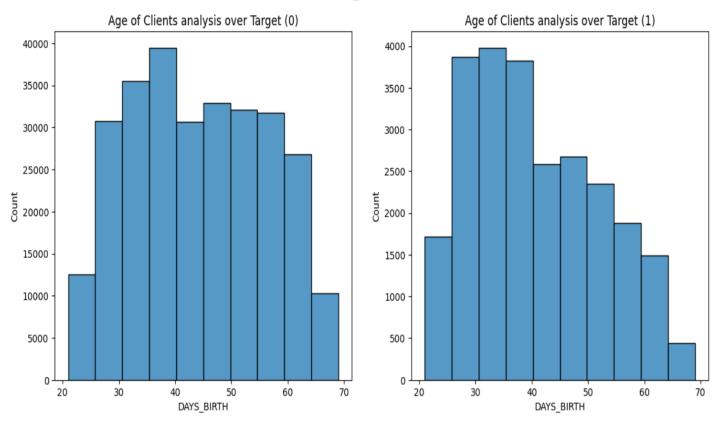
Age of Clients and Payment Difficulties

As the age of clients increases from 40 years, the proportion decreases in the non-defaulted category to the defaulted category.

It indicates that

- Older clients (greater than 40 yrs of age) are most likely to repay loans.
- Younger clients are most likely to default on loan payments

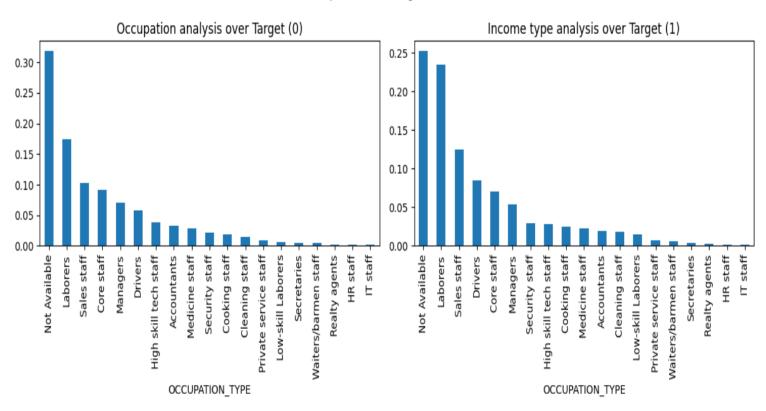
Age of clients (DAYS_BIRTH) over target values



Occupation in Relation to Loan Default Risk

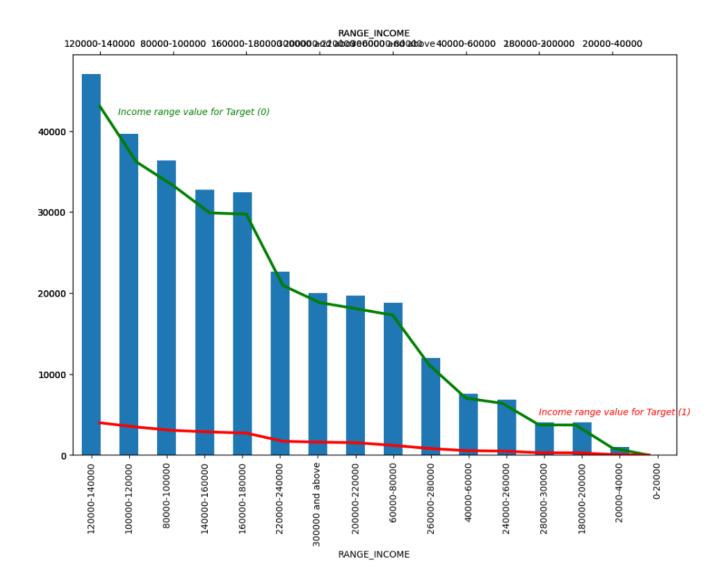
- Clients with occupation as laborers are most likely to default on loan payments.
- Sales staff are also most likely to default on payments.
- HR and IT staff are less likely to apply for loan

Occupation over target values



Range Income in Relation to Loan Default Risk

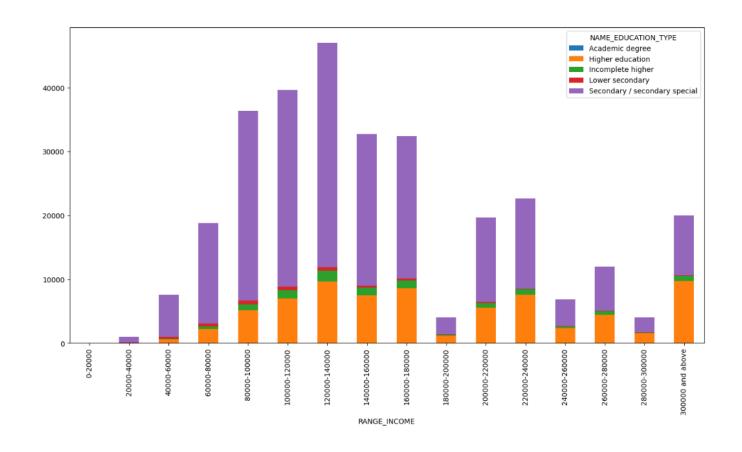
- Clients with higher income range brackets greater than 240000 are most likely to repay the loans.
- Clients with moderate income range bracket i.e., b/w 80000 to 140000 are more likely to default loans.
- Clients with low-income ranges are less likely to apply for loans.



Bivariate Analysis

Education Type VS Income Range

- Clients with low income ranges i.e., less than 80000 have very high percent of clients with secondary level education.
- Clients with mid income range have a balance of education with secondary and higher education.
- Clients with high income range have relatively higher percent of higher education holders.



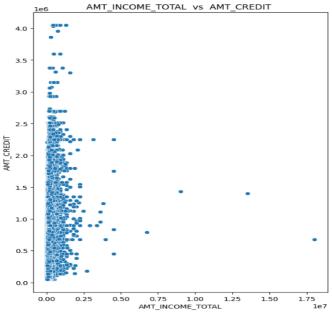
Income vs Credit, Goods price vs Credit

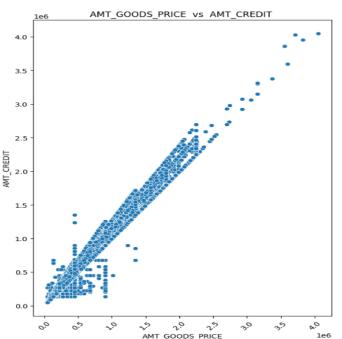
Analysis inference of income and credit amount.

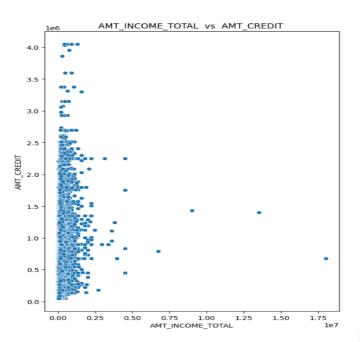
Those who have paid the loan amount on/within time are more likely to get higher credits than those who didn't pay/did late payments.

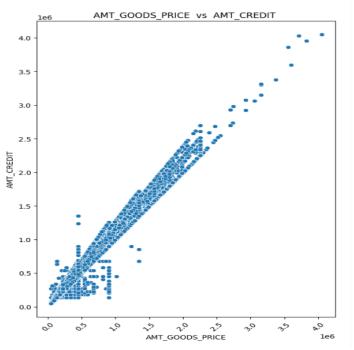
Analysis inference of goods price and credit amount.

People who have higher goods price and have made payments on time have higher credits than those with higher goods price but didn't pay loan.



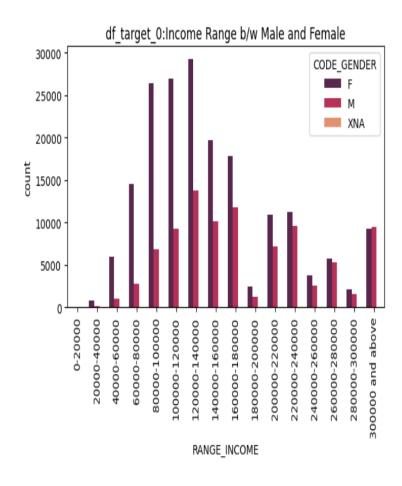


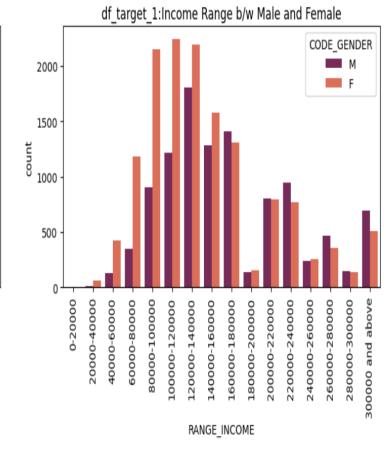




Income Range vs Gender

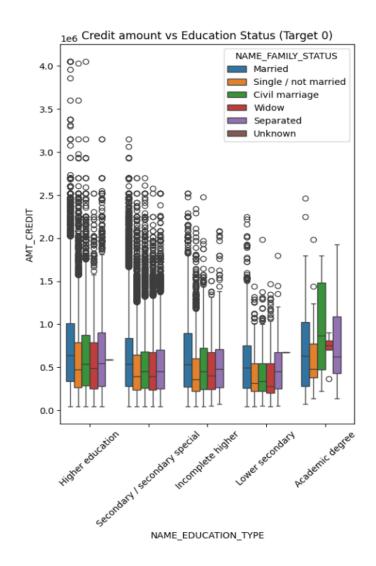
- In both graph, females dominate the income ranges up to 200,000, suggesting that they are the primary earners in these brackets.
- Females, especially in the lower to mid-income ranges, appear more likely to repay loans, possibly due to stable incomes and fewer high-risk loans.
- Males are less represented in lower and mid-income ranges but show a slight increase in the highest income bracket (300,000 and above), particularly in 'df_target_1'.
- Males, particularly in higher income brackets, have a strong repayment capacity due to their selective loan-taking behavior, often linked to higher incomes and strategic borrowing.

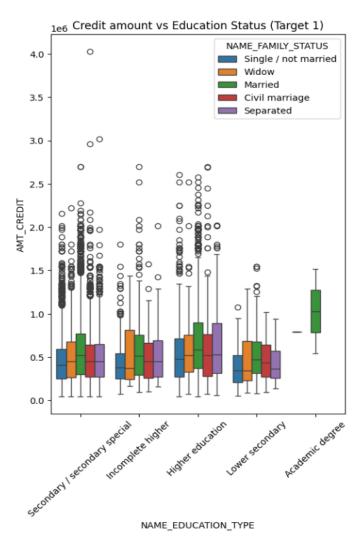




Credit amount vs Education Status

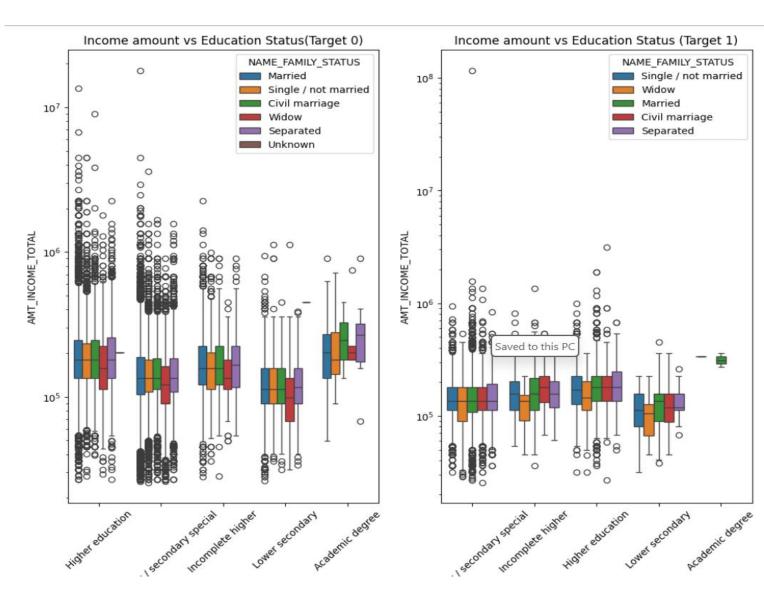
- Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others.
- Civil marriage for Academic degree is having most of the credits in the third quartile.
- Higher education of family status of 'marriage', 'single' and 'civil marriage' are having more outliers.
- Some of the highly educated, married person are having credits higher than those who have done lower secondary education.
- Those with higher education have higher credits and are more likely to make payments on time





Income amount vs Education Status

- For Education type 'Higher education' the income amount mean is mostly equal with family status. It does contain many outliers.
- Less outlier are having for Academic degree but they are having the income amount is little higher that Higher education
- We can see that People with higher education have higher income and don't have difficulties in making loan payment.
- People with higher education who have lesser income are unable to pay the loan. Hence we can conclude that, people with Higher income are most likely to make payments.



Corelation Analysis

DAYS BIRTH -0.0013

DAYS_EMPLOYED -0.001.0

DAYS_ID_PUBLISH -7 7=-15

DAYS REGISTRATION -0.0015

HOUR APPR PROCESS START 4 000011

REG CITY NOT WORK CITY 0.0027

LIVE CITY NOT WORK CITY 4 HIRES

OBS 30 CNT SOCIAL CHICLE 0.0007

DEF 30 CNT SOCIAL CIRCLE CHINISM

OBS 60 CNT SOCIAL CIRCLE 6:88074

DEF_60_CNT_SOCIAL_CIRCLE -0.0024

ANT REQ CREDIT BUREAU HOUR -0.0019

ANT REQ CREDIT BUREAU WEEK -0.0025

AMT REQ CREDIT BUREAU MON -S AL IS

AMT_REQ_CREDIT_BUREAU_ORT -0.0811 AMT REQ CREDIT BUREAU YEAR -0.0043

AMT_REQ_CREDIT_BUREAU_DAY -0.001.)

FXT_SOURCE_2 0.0016

EXT SOURCE 3 4 IIIII III

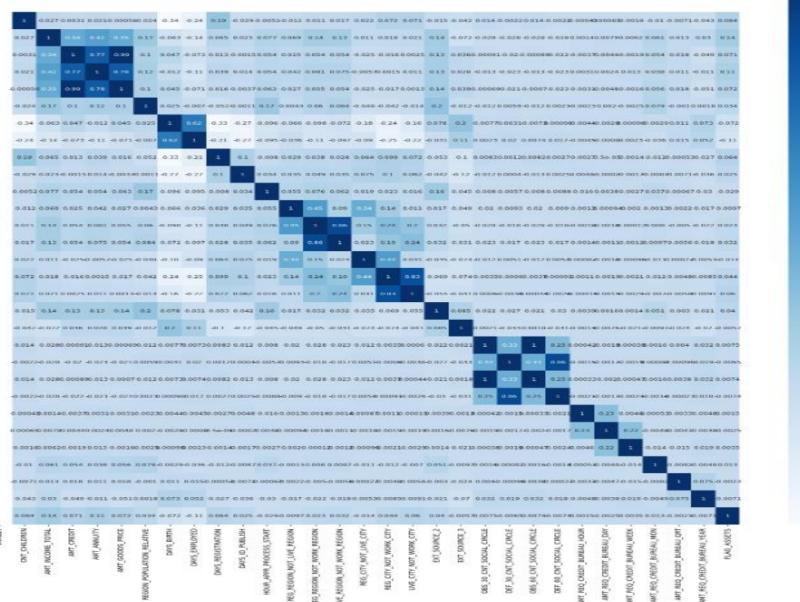
FLAG ASSETS -0.0004

SKJO,CURR

REG_REGION_NOT_LIVE_REGION 0.0001

REG REGION NOT WORK REGION 4 HHITE

LIVE_REGION_NOT_WORK_REGION 0.0028 HEG CHY NOT LIVE CHY -0 0079



0.8

0.6

0.4

- 0.2

-0.0

-0.2

Correlation For Target 0

- There is an inverse relationship between credit amount and age (DAYS_BIRTH), indicating younger clients are more likely to have higher credit amounts.
- Credit amounts tend to be higher for clients with fewer children, suggesting a negative correlation.
- There's an inverse relationship between income amount and the number of children, implying higher incomes for clients with fewer children.
- Clients in densely populated areas tend to have fewer children, as indicated by the negative correlation between REGION_POPULATION_RELATIVE and the number of children.
- The correlation shows that credit amounts are generally higher in densely populated areas, suggesting more credit activity in these regions.
- Income tends to be higher in densely populated areas, aligning with economic opportunities often available in such locations.

0.4

- 0.2

- 0.0

	*	0.004	ac 0017	n 031	0.0001	0.032	0.26	0.19	0.15	0.032	0.024	0.024	0.021	0.012	0.0012	0.046	0.054	0.012	0.02	0.026	0014	0.025	0002	00001	00.00	0.012	0.013-4	0038 4	.035	10
	E 0048	1	0.030	nosh	0.000	0.00934	0.0029	-0.0330	0001	0.0042	0.014	0.0076	0.015	0.015	0 0022	0.003	6 0014	0.0012	-0.015	0.00474	005.00	00480	0048	00000	00022	De-000	00410	00130	onte o	10
	E 0017	0.038		6.75	0.98	0.000	0.14-0	00001	0.026	0.012	0.032	0.02	0.022	0.034	-0.003	0.038	= 017	0.12	0.078	0.019	8.02E	0.019	= 031	0.006	0.003.0	00771	055-4	0017-0	1036	0.0
	8 631	0.048	0.75		0.75	0.072	0.014	-0.983	0.034	-0.017	0.011	0.025	0.057	0.064	0.0057	0.002	0.0099	0.12	0.041	0.0045	0.0220	.0035	0.027	0.0540	00029		1053		024	12.3
	0.0081	0.030	0.98		4	0.076	0.24	0.0036	0.026	0.056	0.044	0.022	9.030	0.036	0.004	0.039	0.017	012	0.079	8.02	0.022	0.02	0.026-0	0.0040	0042	0.000 1	1,000 4	0.010-0	.036	1.0
	-0.052	0.000	1009	nozz	s oze.	1	H 048	0025	0.050	-0.1118	11.114	0.022	0.022	noan	-0 no	0.052	40.002	BAF-	0 0096	00002	o nen c	monis	E 038-4	0.0018	0.00000	0029	ors-o	00770	ome i	ra
	0.26	0.002	9 0.14	0.014	0.54	0.048	1	0.57	0.29	0.25	0.062	0.055	0.07	0.046	0.16	-0.22	0.12	011	0.17	0.00514	0.004%	100530	0037	0.015	0.06730	00560	0057	0.017 0	084	0:0
	-0.19	41 0 15	w 0009	AD 08:1	0.0036	0815	0.57	a	-0.19	-0.25	-0.06	0.055	-0.00M	-0.077	-0.085	-0.24	-02 -	0 0015	0.096	0.00096	00336	00928	005-86	0.0026	0.027	0015	s.026.6	0.015	035	4014
	0.15	0.0001	60.026	0.034	0.026	0.056	0.29	0.19	1	0.097	0.033	0.02	0.021	0.016	0.049	0.088	0.063	0.07	0.086	0.014	0.01	0.015.0	0092	0.005	0.00840	.00210	0049	0.012 0	0000	1-0
	-0 052	-0 005	240 155 2	-0:01Y	-0.056	-0.00m	-0.25	-0.75	0 1147	1	0.007	0822	×055	8 007	0.087	0:01	0.051	-0.0m	49.119	-0.02 (ooze.	40.007 E	00498	10058	0.019	11 00% 4	014.0	2008-4	0424	101
	0.024	0.034	0.032	0.021	0.044	014	0.062	0.06	0.632	0.022		0.052	2.069	0.053	0.0099	0.010	0.005	012	0.046	0.013	0029	0.013	0003	0.018	0.0160	0067	0.04.0	.0035-0	0.000	0.10
	0.024	0.007	6 0.02	0.035	0.022	0.022	0,055	0.033	0.02	0.033	0.052	1	0.7	0.060	0.32	0.15	0.0067	0.013	0.045	0.028	0024	0.0280	0.0016	0.01	0.00630	.00721	1024 0	2043 0	.022	0:10
	-0.000	0.015	0.044	HOE.	0.016	0077	40.07	-D 23.899	0.021	поэл	0.069	9			014	0.50	11.79	0.071	-0.011	41-0-50	0.014	0.088	0.0104	0.00000	0.0024	-047-074.0	015-0	00102-0	10000	-34
	0.032	0.017	0.034	0.004	0.036	0.046	0.046	0.077	0.016	0.02	0.003	0.009	0.85	1	2.2e 01	0.18	0.24	0.02	0.013	0.028	0.021	0.027	0.02.0	1.00070	00028	0.0046	0.02 0	.0062.0	0.014	10
	er (1811)	ets cietur	2-0 0+n	0.00%	-0.035	-0 tte.	40 18	-0.005	пон	11 000.7	i come	0.47	Hart	T Tend	7	nar	en en 1	e0 115	<0 mm m	45 1507 7	0.011	0.027	00124	00074	0.1300498	1001-040	00042	0 H1 40	contin-	1.00
	2.046	0.000	0.036	0.002	0.039	0.052	0.22	0.24	0.088	0.07	0.010	0.15	0.24	0.58	0.47		0.76	0.07	0.053	0.021	2.006	0.021	0.000	0.0071	200.0	0.01	0.01.0	.0002	014	1.0
	SE CON-SE	en 19111	40.017	II DOWN	40 01 0	-0.117	40.10	-0.0	12. 6196-15	11-0.111	0 O(A-	0.000	11:19	11.75	40.011	0.78	*	-D 110W	40 807%	ep incomes	0 01 /4	00000	40 000 4	n otres	0.1203.446	000054		el 117 40	190840	no
	0.011	0.807	2 0.12	0.12	0.13	0.17	0.11	0.0015	0.07	0.06	0.13	0.012	0.021	8.02	0.05	0.07	0.046	1	0.076	0.016	0.01	0.018	0.011	0.0040	0.0055	0.004	0.00 0	.0006.0	000820	id
	49 1007	40 11114	20 07 0	11 (2541)	10 03%	2 21741	11.20	to cover -	o oue	47.11	45 DIES.	0.045	-0.011	-01 OTLN	40 UN. T	40-10% 4	an anna	onze		01110	0 0774	00004	o min	11001	o mes -	11 10 21 4	2119-4	D 28 (FW 40	10VA-0	-
	0.026	0.004	70.019	0.0045	0.02	0.80621	0051	0.0089	0.014	0.02	0.013	0.028	0.036	0.028	0.027	0.021	0.0085	0.018	0.0056	2	-	÷	0.26	0012	0.00920	.00370	00340	00720	038	1.100
	n oosa	HI 00%	240 026	-0.000	49 600	08784	0.004.3	0.0057	081	11 00 76	S 002/9	0.0024	40.01.6	-0.021	36.00EX	-0 006	4333	-0.01	-0 E22	and the same	18	M MAIL	0.87	notte	100250	1,00540	000.0	00119	013/4	1011
	0.025	0.004	60.019	0.0055	0.02	0.0068	0.0053	0.0092	0.015	0.02	0.012	0.028	0.038	0.027	0.027	0.021	0.0085	0.018	0.0055	1	024	2	0.26	1,0013	0.00940	00020	00430	0064.0	038	100
	-B 0051	60 00e	90.611	-0.0027	-0 02%	OHIHA	0.003.8	00540	111192	0.0069	0.0003	emms.	40 OS W	-0.00	0.015	-0 005	-0.02	-0.011	-0 015	0.26	11.00	0.2E	*	10049-	0.00340	001-040	00290	HORSE SE	HE2 FO	ime
	0.0003	0.0000	60.006	0.014	0.0045	0.0052	0.015	0.0076	0.005	0.0038	0.016	0.01	0.0013	0.0027	0.0029	0.0071	0.0097	0.0046	0.003	0.0012	0.0120	.00130	1.0049	-1	9.25	.00620	00790	00660	00270	31
	-0.014	0.000	20.003	0.0002	B 0043	0.00558	6 0073	H-0073.4	0084	H0 01.4	0.0164	0.0062	-0.0036	1.00031	Ø 0069	0.005	0.0034	0.0055	o nes-	0.00908	100754	1:0094E	0036	B25	э.	R3R -	0.03.30	1100291	00124	0.0
	0.012	Line 0	50.0077	0.035	0.008	0,00201	0.0056	0.015	0031	0.000	0.0067	0.0072	0.0074	0.0044	00.0074	0.01	0.0085	0.004	0.021	0.00370	0.005-40	0.00320	00440	1.00EZ	0.18	*	0.012	0.03.0	017	1.0
	0.013	0.806	1 9 055	0.053	2.059	00711	0057	-0.0260	0049	0.014	804	0.024	oms.	B 02-	0.000	20.05	-0.011	0.05	-0.010	0.00344	000 K	10431	00254	0.0079	G. 013-	0.012	1	901340	00261	10
			1-0-01/2																									_	01 0	ai
	0.035	0.001	9.0.036	0.014	0.038	0.0006	0.004	0.035	0.013	0.042	0.033	0.023	0.022	0.024	0.0000	0.014	0.0008	0.00082	20.053	0.036	mers.	0.038.0	100270	0.0023	0.0012	0.017-0	0026	-1	*	0.0
	0.069	0.034	0.07		0.048	0.020	40.003		0.057	0.0047	0.008	0.021	e neus	000	-0.025	0018	18.004		100000	n 00% 4	DOTA	0.00%-6	00000	e cone	0.012	0.004 8	0.021-0	0071	0.01	1
DAMOST	LDMEN	TOTAL	ANT_CREDIT	ANT_ANNUITY	MIC.	LAME	DAYS BIRTH	DAYS_EMPLOYED	MATION	DAYS ID PUBLISH	START	#000	RCION	HOION	ECI.	E,	E,	EXT_SOUNCE_2	EXT_SOURCE_3	E E	ORG.	CHOLL	CHOLE	E.	W DAY	WEEK	NOW.7	16,01	U,YEAR	AGGFTS
	ONT_CHILDREN	ANT INCOME TOTAL	AN.	A.	ANT GOODS PRICE	M.	DAY	18. BI	DAYS, AEGISTRATION	0,	POCESS	3	WORK	WORK	NOT_U	01,00	04,70	8,	07,30	SOCIE	30CK	SOCIAL	SOCIAL	BUNEAL	E.	BUTEN	BUREA	100	BUREA	RAD
		¥			TW.	POPULA		0	PA S	ă	HOUR, APPROPESS, STANT	ON NO	W. NOT.	100	REGLOTIY MOT LINE, CITY	REG CITY NOT WORK CITY	CITY_NOT_WDRK_CIT			OBS 30 CNT SOCIAL CINCLE	DEF_34_CNT_SOCIAL_CIRCLE	085_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE) BDT	IGEO,	MD	OREDI	9,	CREDIT	
						REGION POPULATION RELATIVE					Đ,	HEG JEGION NOT LIVE JEGION	REG_REGION_NOT_WORK_REGION	UVE_REGION_NOT_WORK_REGION	561	REG	N.			18		9,990	18	AMT, NEQ, CNEDIT, BUNEAU, HOUR	ANT, RED_CREDIT, BUREAU, DAY	AHT, NEQ, CAEDIT, BUNEAU, WEEK	ANT, REO, CREDIT, BUREAU, MON	ANT NEQ CHEDIT BUNEAU QRI	AMT_REQ_CREDIT_BUREAU_YEAR	
						~						*	25	š										₹	4	₹	4	4	毫	

SK_ID_CURR -

REGION_POPULATION_RELATIVE -0.11163

HOUR APPR PROCESS START - 0.005

REG_REGION_NOT_UVE_REGION - 0.0042

PLG_REGION_NOT_WORK_REGION - 0.0042

REG CITY NOT LIVE CITY -0 HINTS

REG CITY NOT LIVE CITY -0 HINTS

REG CITY NOT WORK CITY -0 HINTS

LIVE CITY NOT WORK CITY -0 HINTS

TARGET
CNT_CHILDREN -8.0051

AMT_NCOME_TOTAL - 0.91

AMT_CREDIT -8.0013

AMT_ANNUITY -8.0078

ANT_GOODS_PICCE -0.0010

DAYS_EMPLOYED =0.0052

DAYS_EMPLOYED =0.0052

DAYS_REGISTRATION =0.0062

DAYS_ID_PUBLISH =0.0025

FIXT_SOURCE_> 0.0863

EXT SOURCE 3 -0 0010

FLAG_ASSETS & GOODS

OBS 30 CNT SOCIAL CIRCLE =0.0054

DEF_30_CNT_SOCIAL_CIRCLE =0.0051

OBS 60 CNT_SOCIAL_CIRCLE =0.0051

DEF_60_CNT_SOCIAL_CIRCLE =0.0081

ANT_REQ_CREDIT_BUREAU_NOUR -0.012

ANT_REQ_CREDIT_BUREAU_WEEK -0.003

ANT_REQ_CREDIT_BUREAU_MON -0.00666

ANT_REQ_CREDIT_BUREAU_ORT -0.00666

ANT_REQ_CREDIT_BUREAU_YEAR -0.00666

Correlation For Target 1

This heat map for Target 1 is also having quite a same observation just like Target 0. But for few points are different. They are listed below.

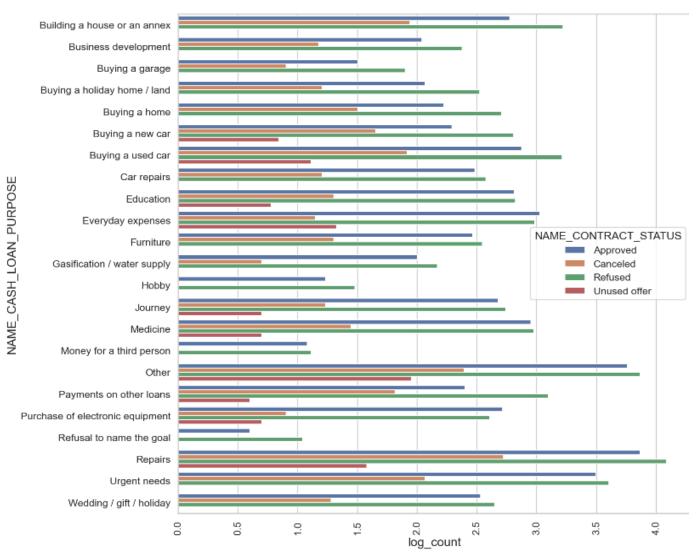
- Clients whose permanent address does not match their contact address tend to have fewer children, highlighting mobility or lifestyle differences.
- There's a similar observation where clients whose permanent address does not match their work address also tend to have fewer children, which could reflect work-related relocations.

Analysis after merging previous data

Distribution of contract status with purposes

- Most rejection of loans came from purpose 'Repairs'.
- For education purposes we can see equal number of approves and rejection.
- Purchase of Electronic equipment has higher approved than rejection.

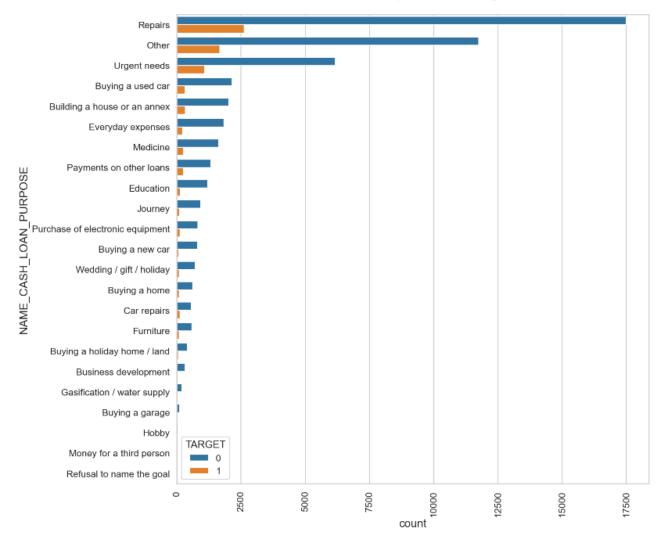




Distribution of purposes with target

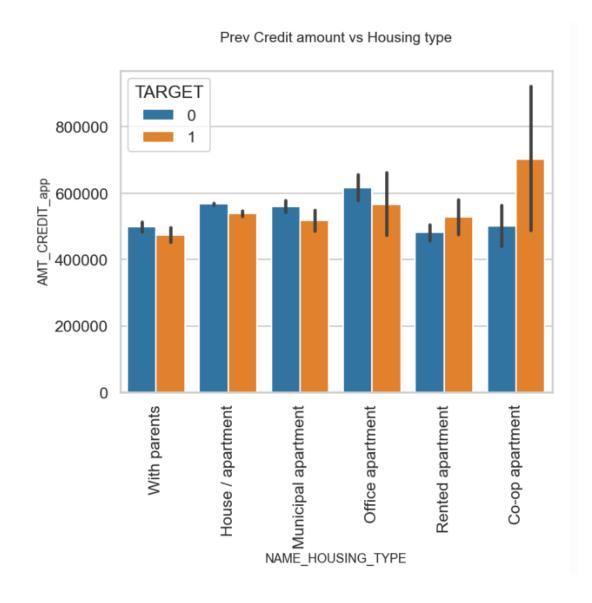
Loan purposes with 'Repairs' are facing more difficulties in payment on time. loan payment is significant higher than facing difficulties.

Distribution of purposes with target



PREV CREDIT AMOUNT VS HOUSING TYPE

- Here for Housing type, office apartment is having higher credit for target = 0 and co-op apartment is having higher credit for target = 1.
- So, we can conclude that bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment.
- Bank can focus mostly on housing type with parents or House/apartment or municipal apartment for successful payments.



CONCLUSION

- The loan purpose 'Repair' is having higher number of unsuccessful payments on time hence bank should avoid giving loan for this.
- > Banks should focus more on contract type 'Student', 'pensioner' and 'Businessman' with housing 'type other than 'Co-op apartment' for successful payments.
- Pensioners & people with Higher income are most likely to make payments.
- People with housing type 'With parents' can be targeted as they are having least number of unsuccessful payments.

Thank You