iiit-b International Institute of Information Technology Bangalore

ज्ञानमुत्तमम्

UpGrad

Credit EDA Case Study

Submitted By :

ABHISHEK DAS

PROBLEM STATEMENT

Two types of risks are associated with the bank's decision:

1. TARGET(0)-If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
2. TARGET(1)-If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

ANALYSIS OF DATA IS DONE IN PYTHON ON A JUPITER NOTEBOOK

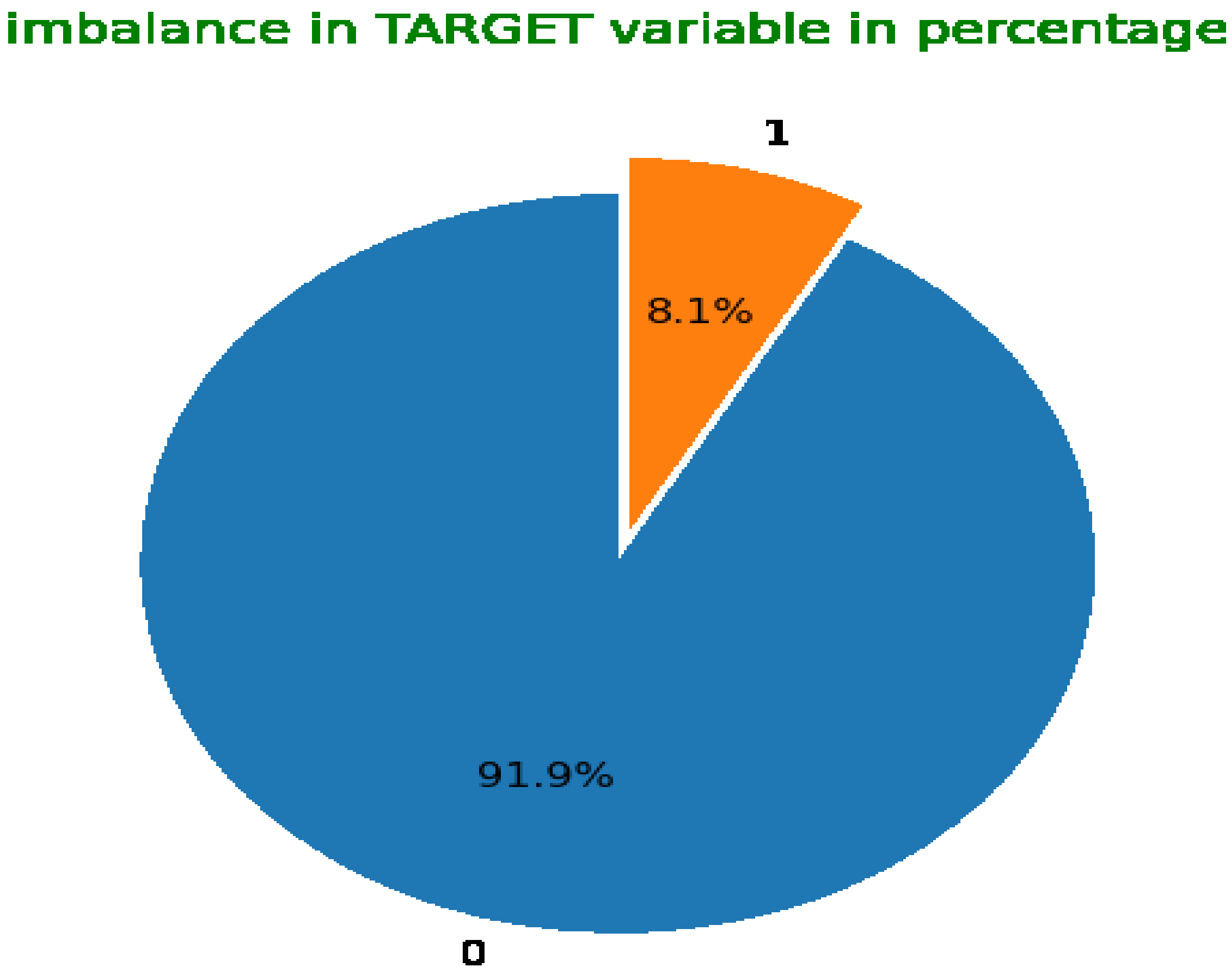
DATA PRE PROCESSING

STEPS:

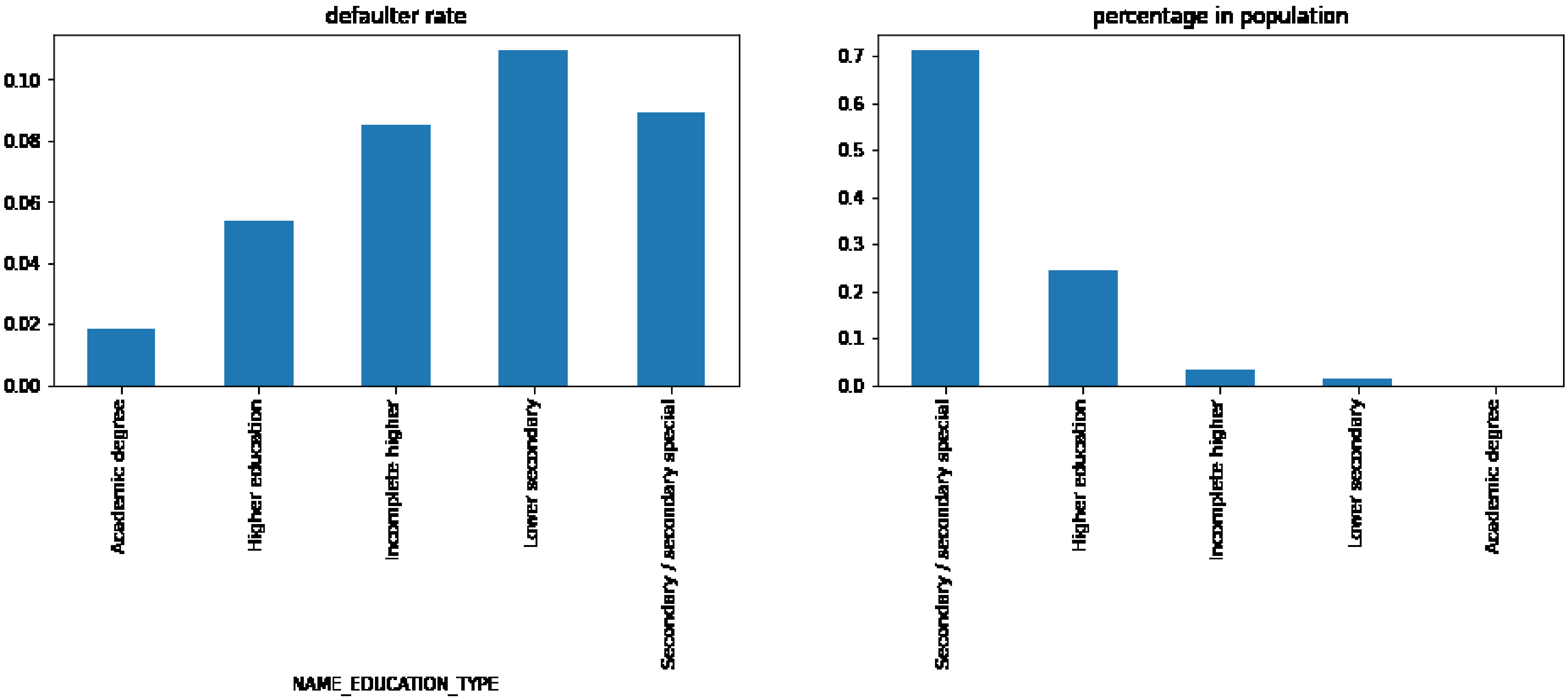
1. **Check the structure of data(shape,size,info,describe)**
2. **Check missing value columns(which to drop,which to handle and how to handle and which column to keep as it is).**
3. **Drop any column which is irrelevant to TARGET variable**
4. **Check data type of each column,fix column with unappropriate data type and binary categorical column(into 0 and 1)**
5. **Check outliers and data imbalance .**
6. **Binning of continuous variables as required during analysis**
7. **Top 10 correlation for defaulter and non defaulter**

Defaulter-8.1%

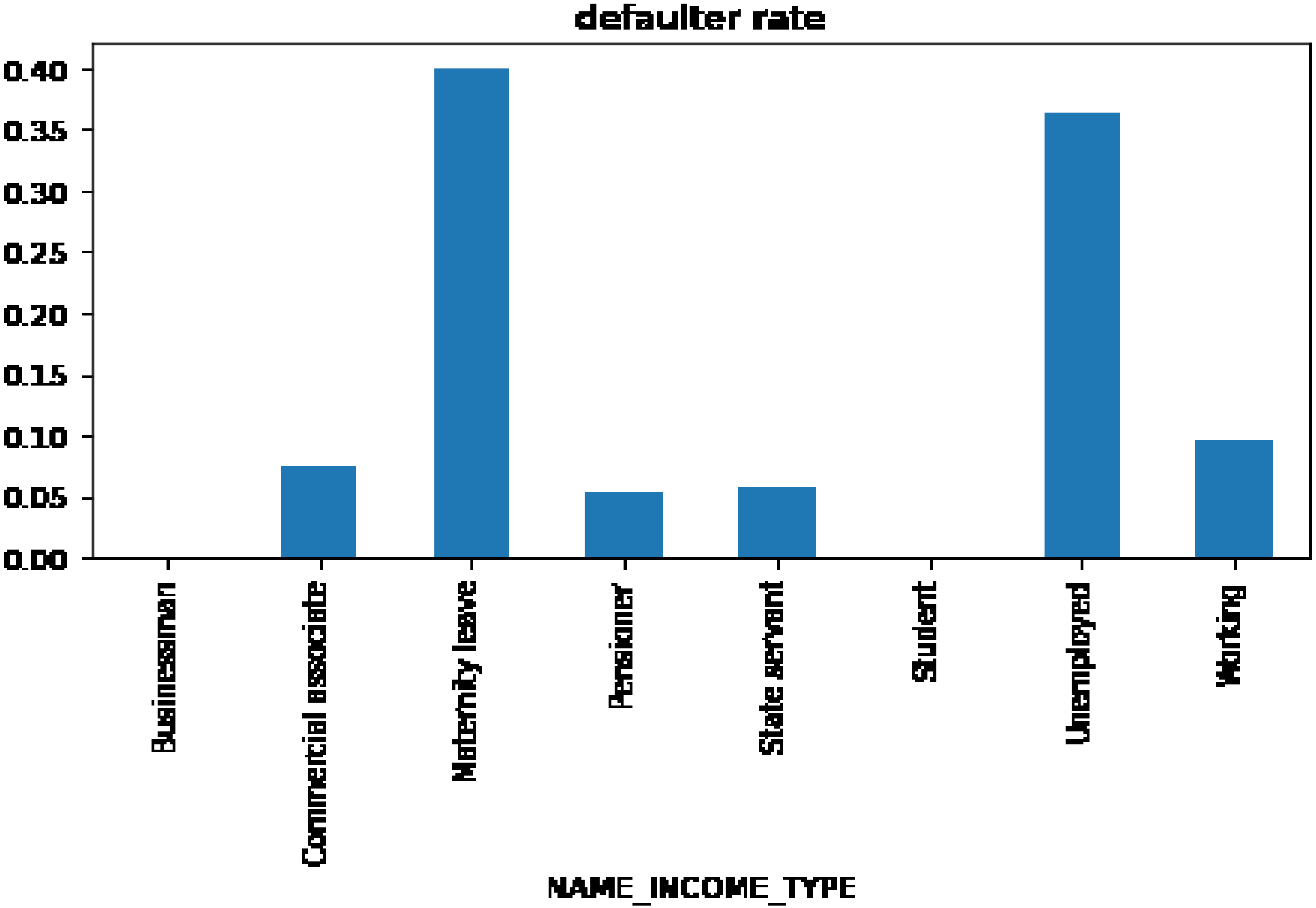
Non-defaulter-91.9%



Higher default rate in lower secondary and least default rate in academic degree(default rate falls with education)

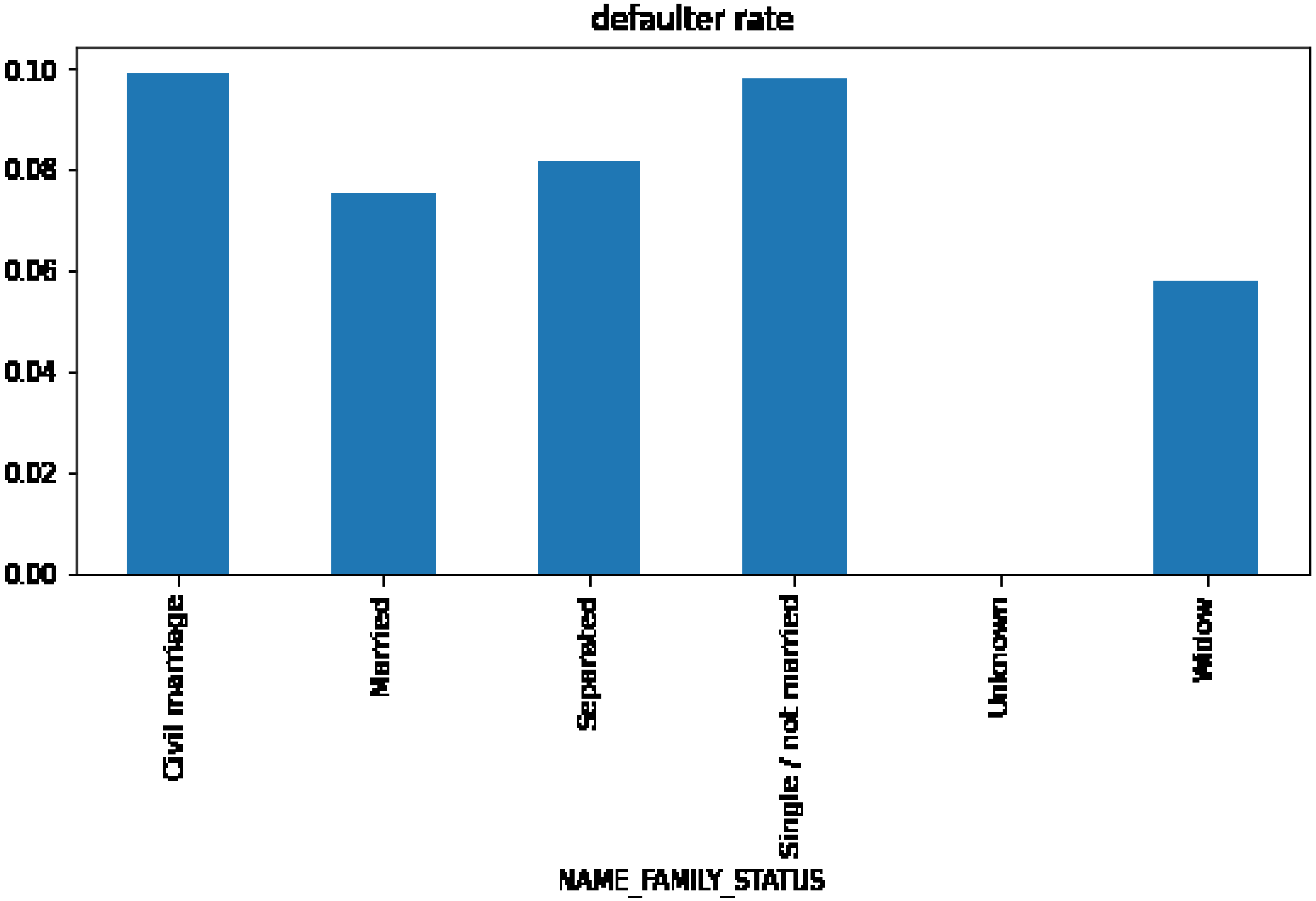


1)Student and buissnessman are safest.
2)Women in maternity leave and Unemployeds had highest default rate



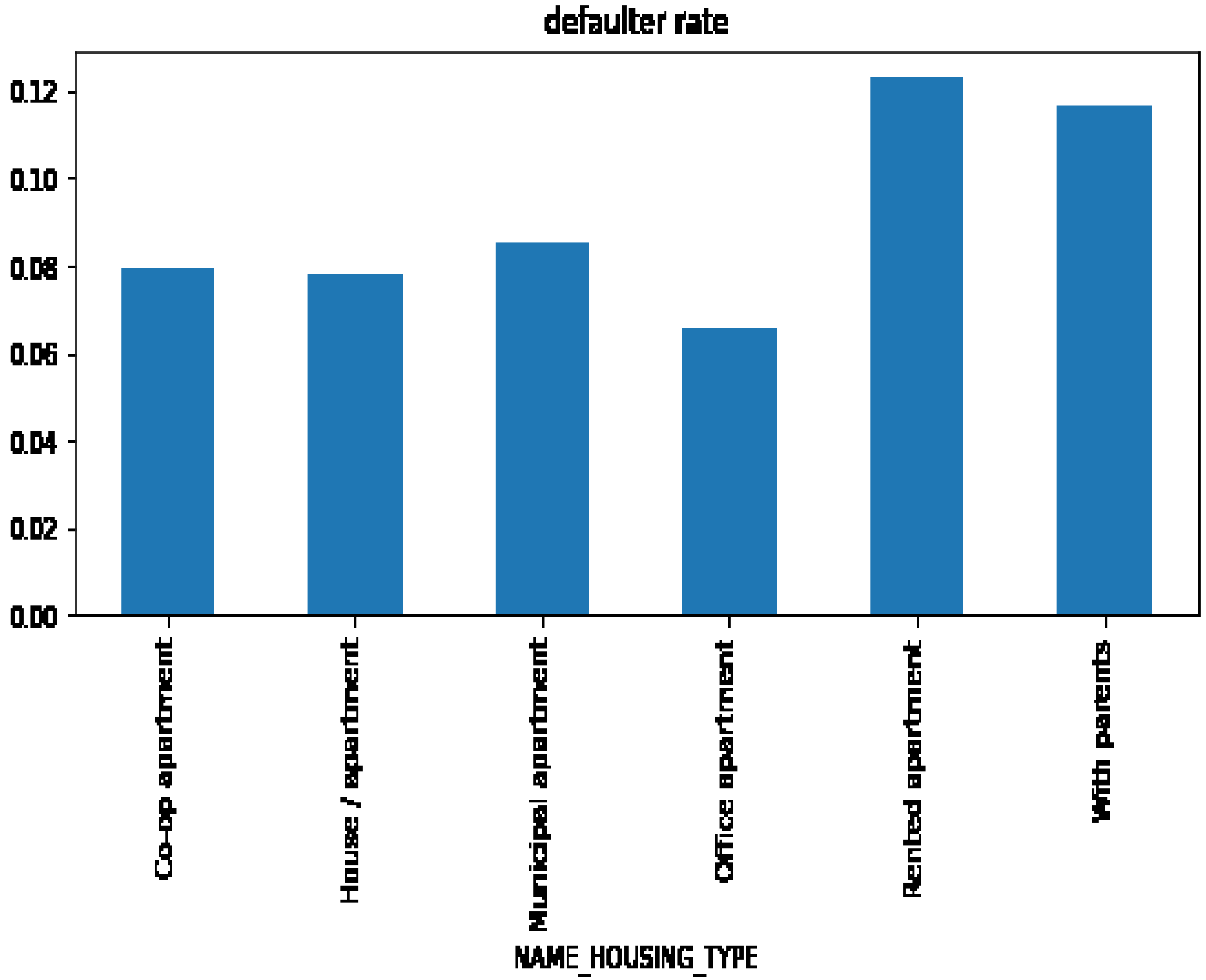
1)Civil married and singles are risky

2)Widows are safest



1)People living at rented apartment and with Parents are highly risky

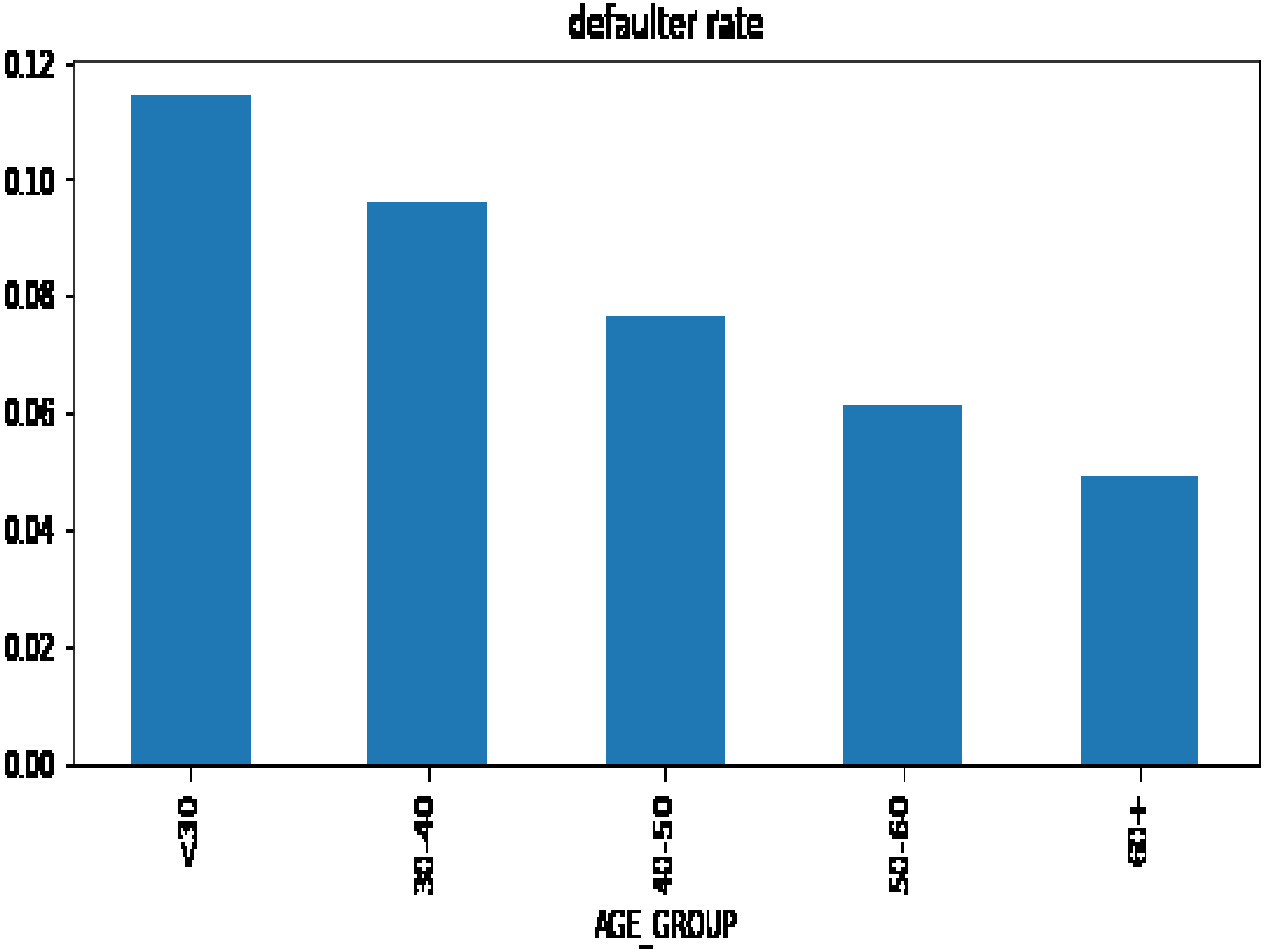
2)People living at office apartment are least at risk



1)Default rate fall with age.

2)People(<30) had highest default rate

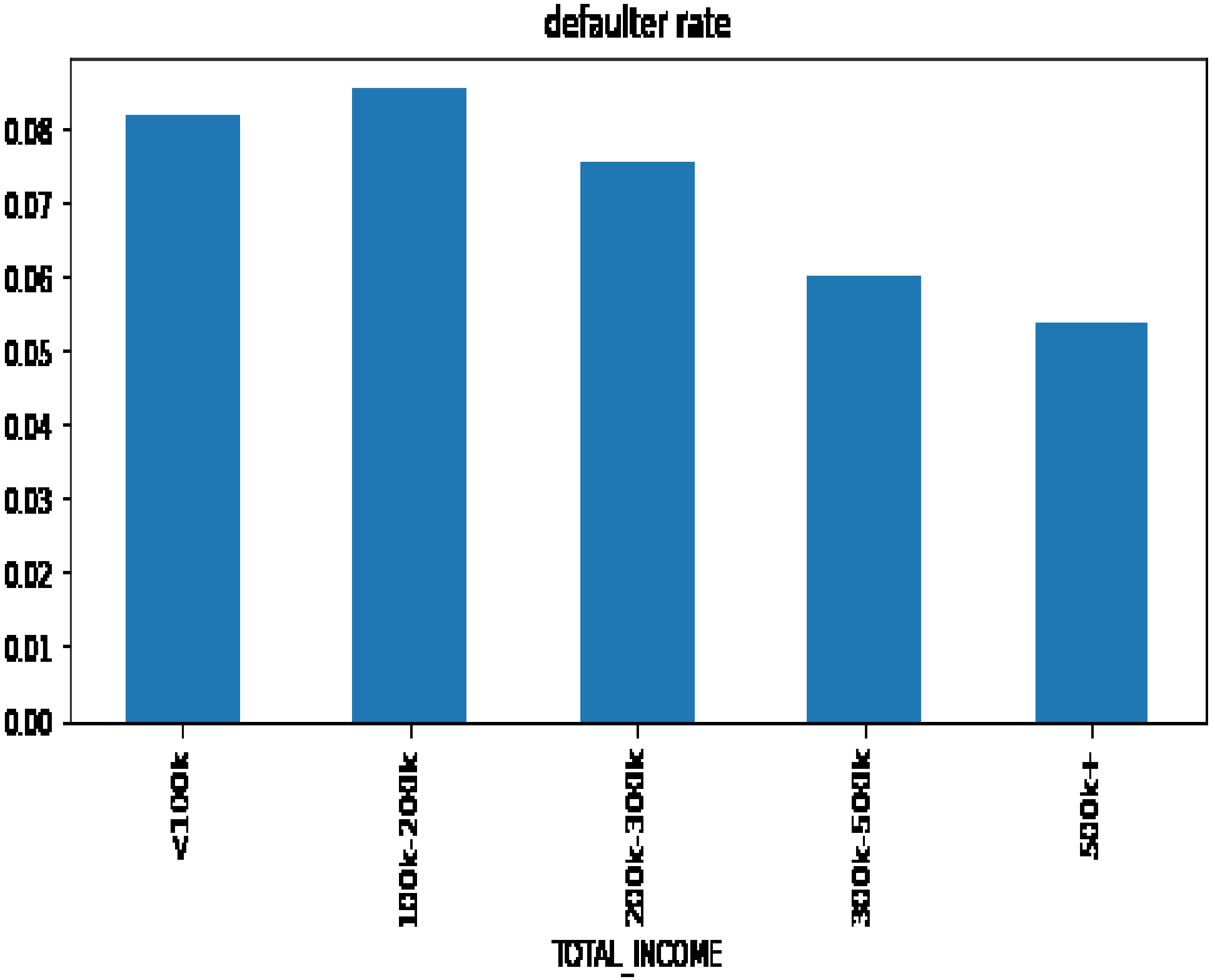
3)people(>60) had least default rate



1)Default rate fall with income.

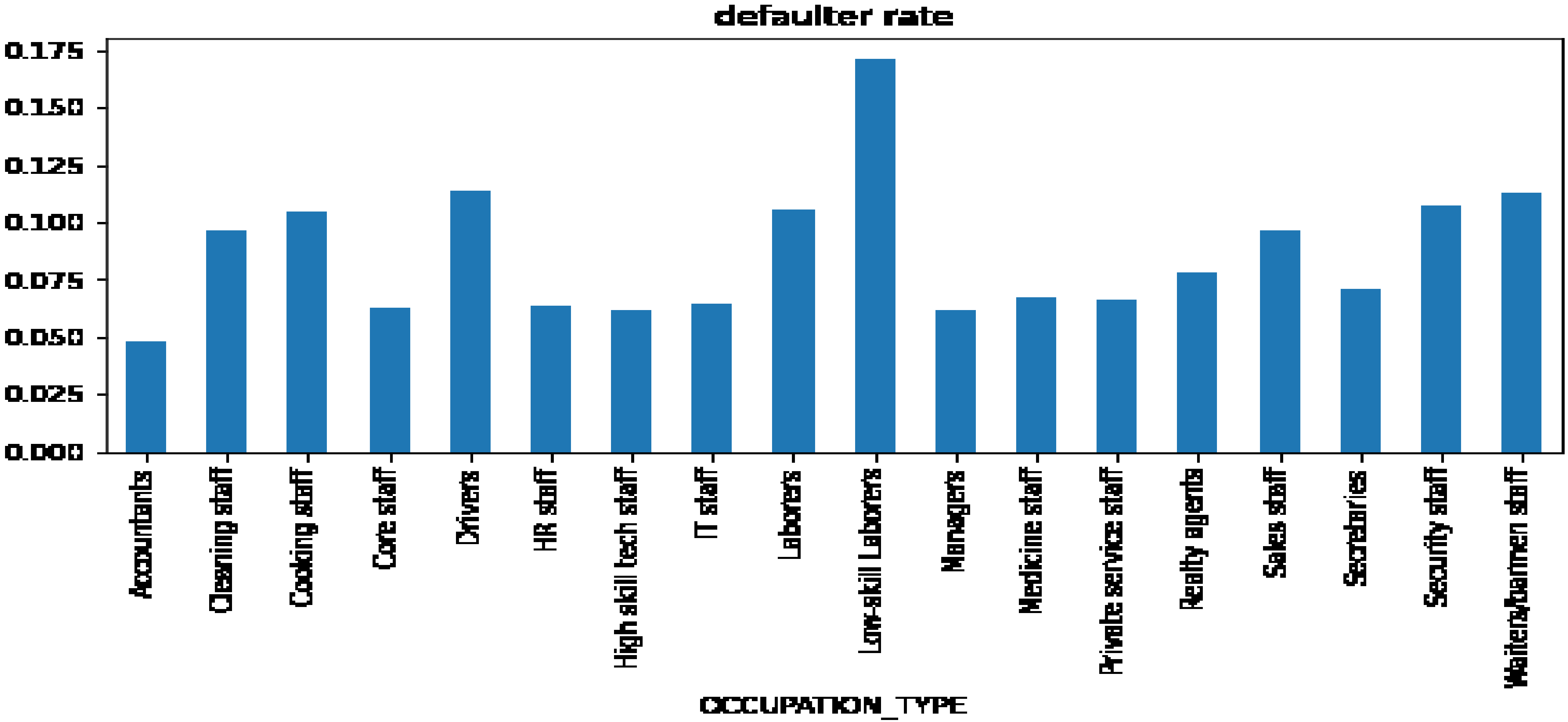
2)People having income <200k are highest risky

3)People with income >500k are safest



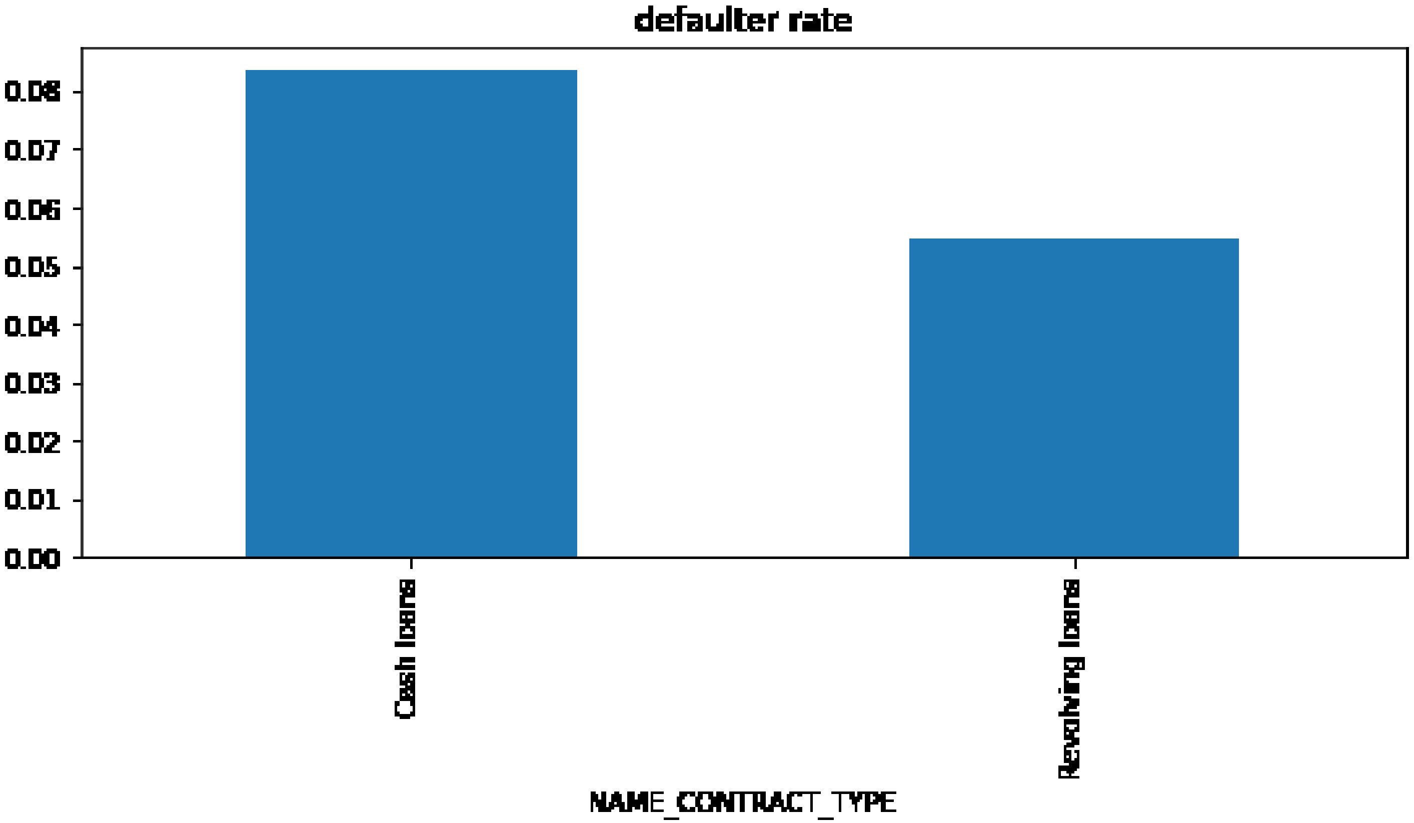
1) low skill labours(highest),waiters,drivers,labourers defaulter rate is high

2) accountant,core staff,hr staff,IT staffs are safest

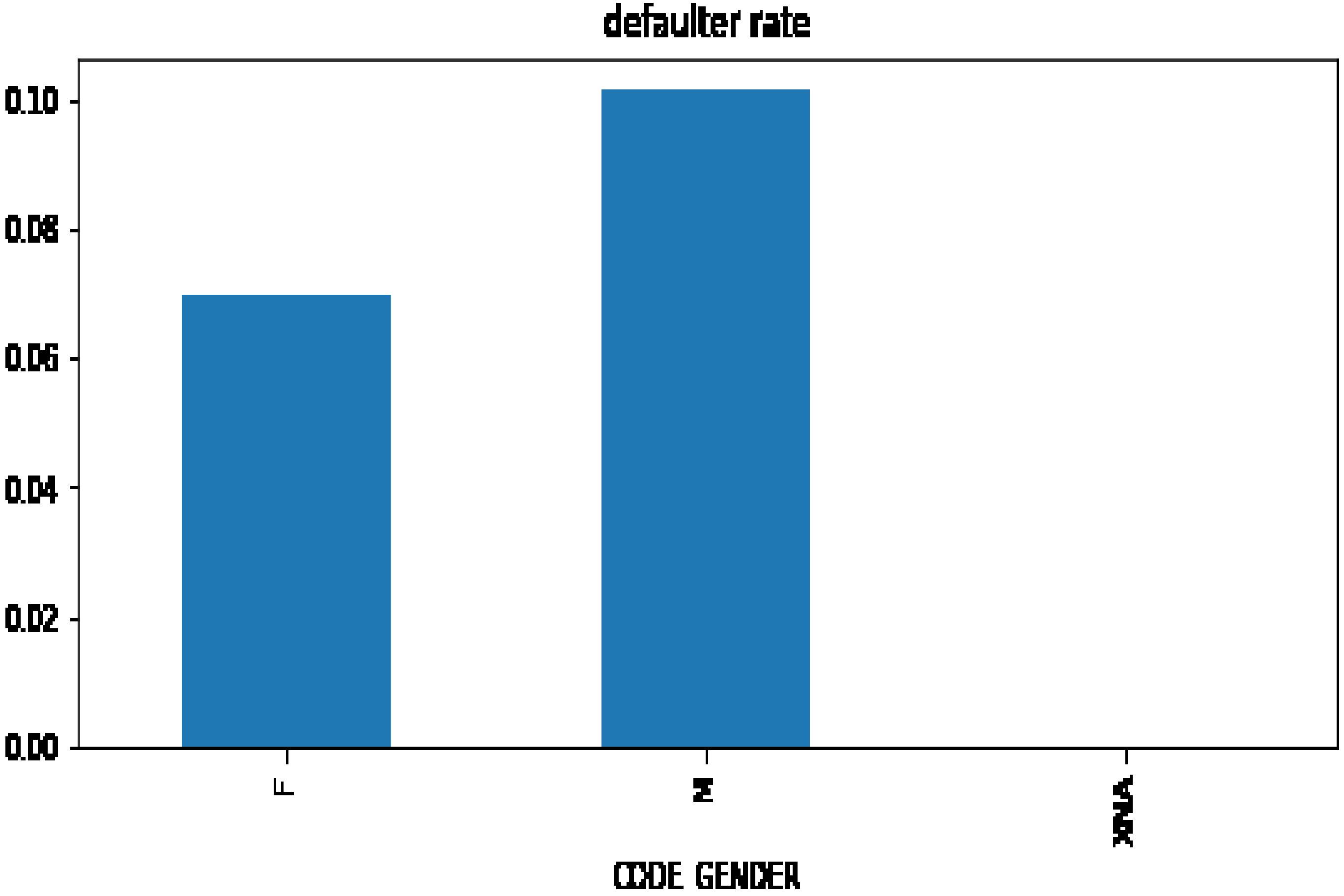


1)Defaulter seem to apply for cash loan

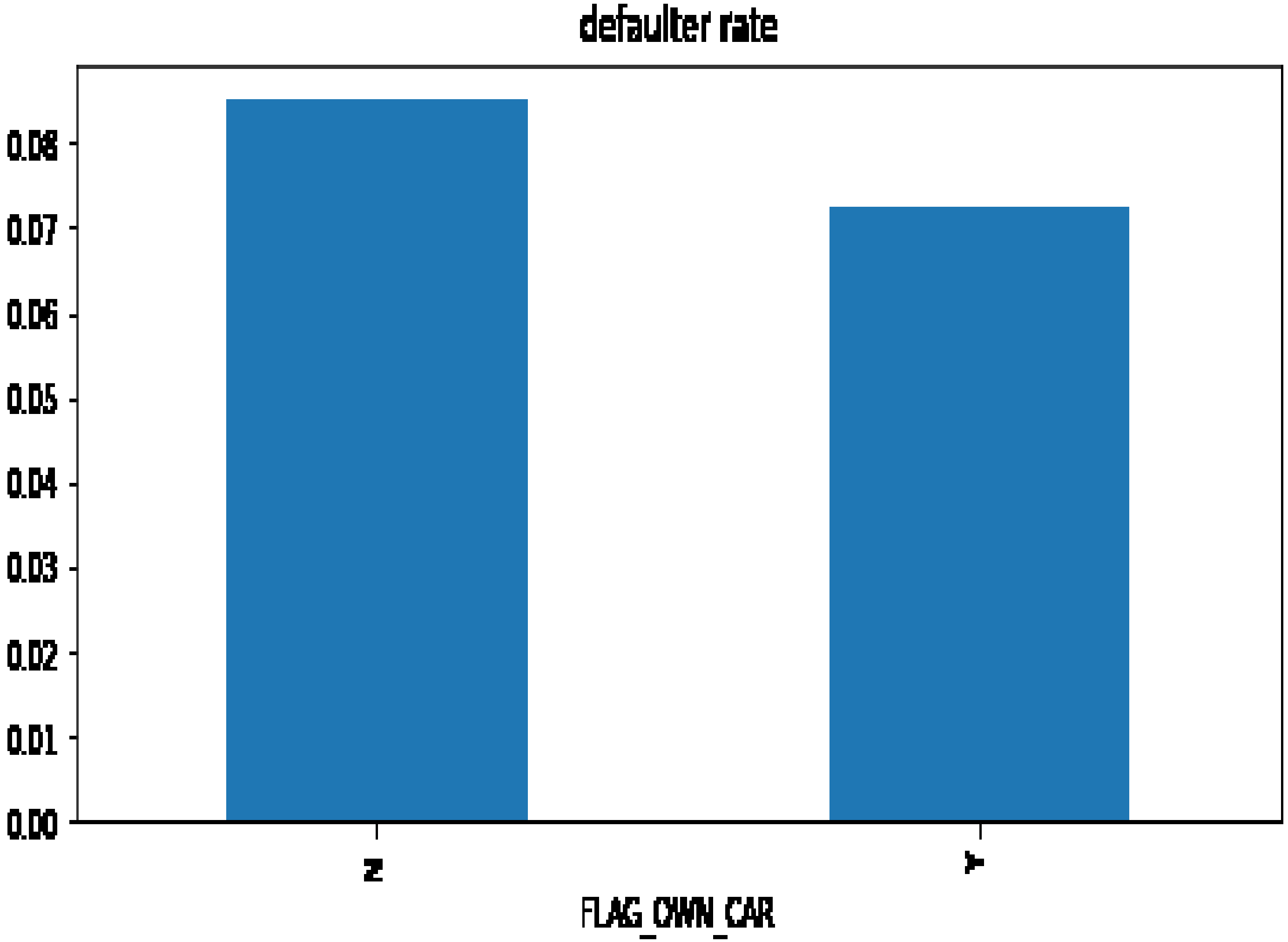
2)non-defaulter applies for revolving loans



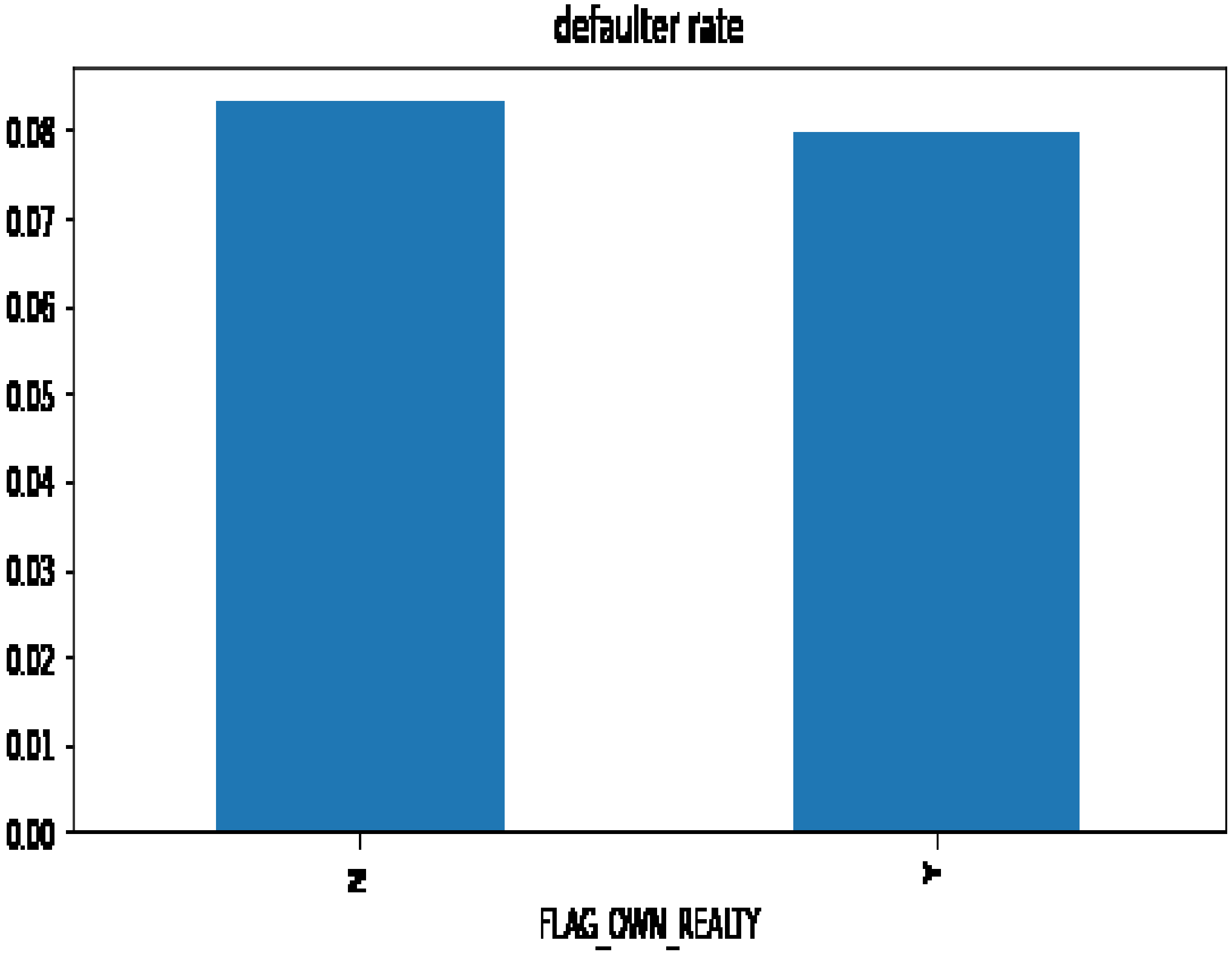
Males seem to be more prone to default



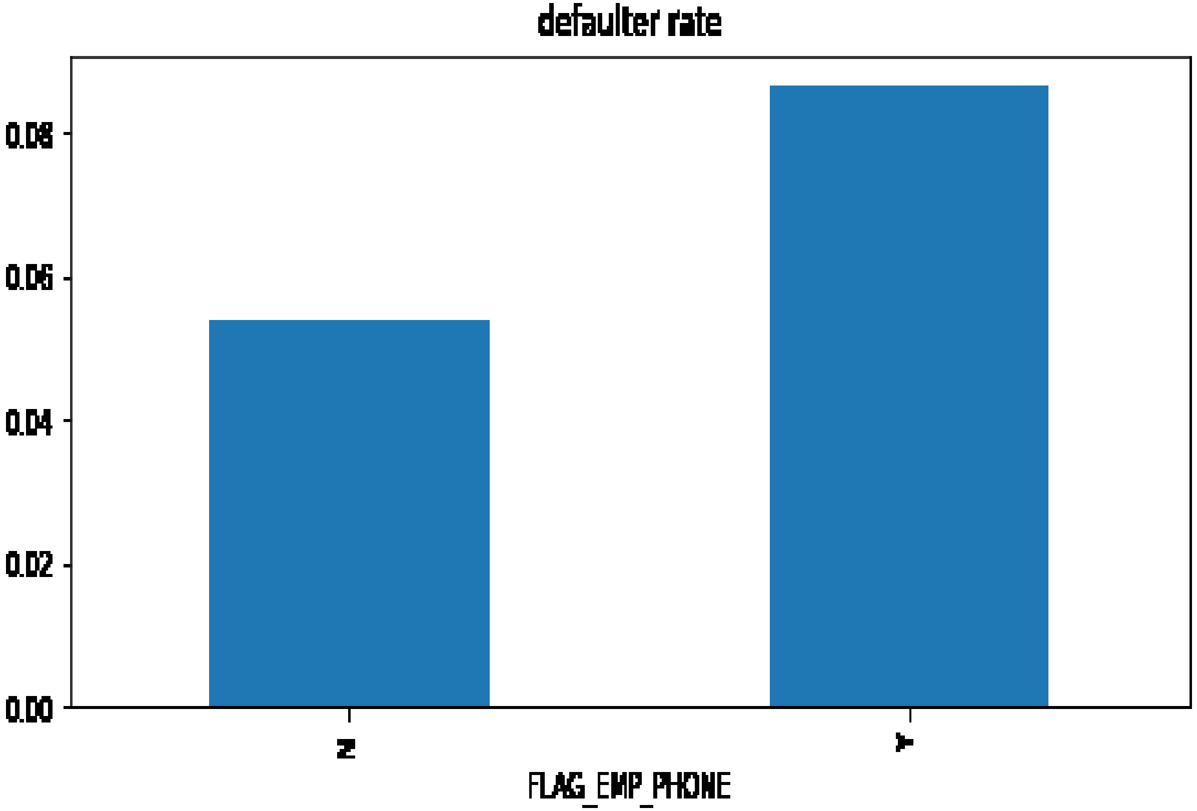
1)Those who don't have car are more prone To default

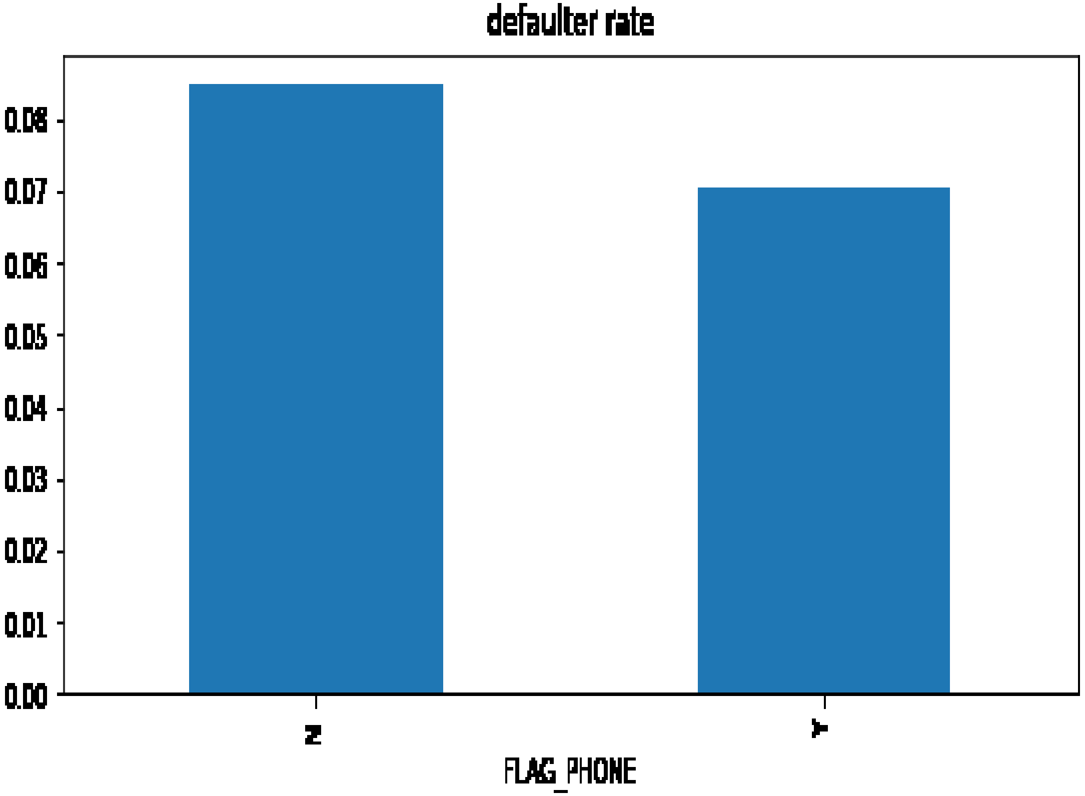


1)Those who don't own flat/house had high Chance of default

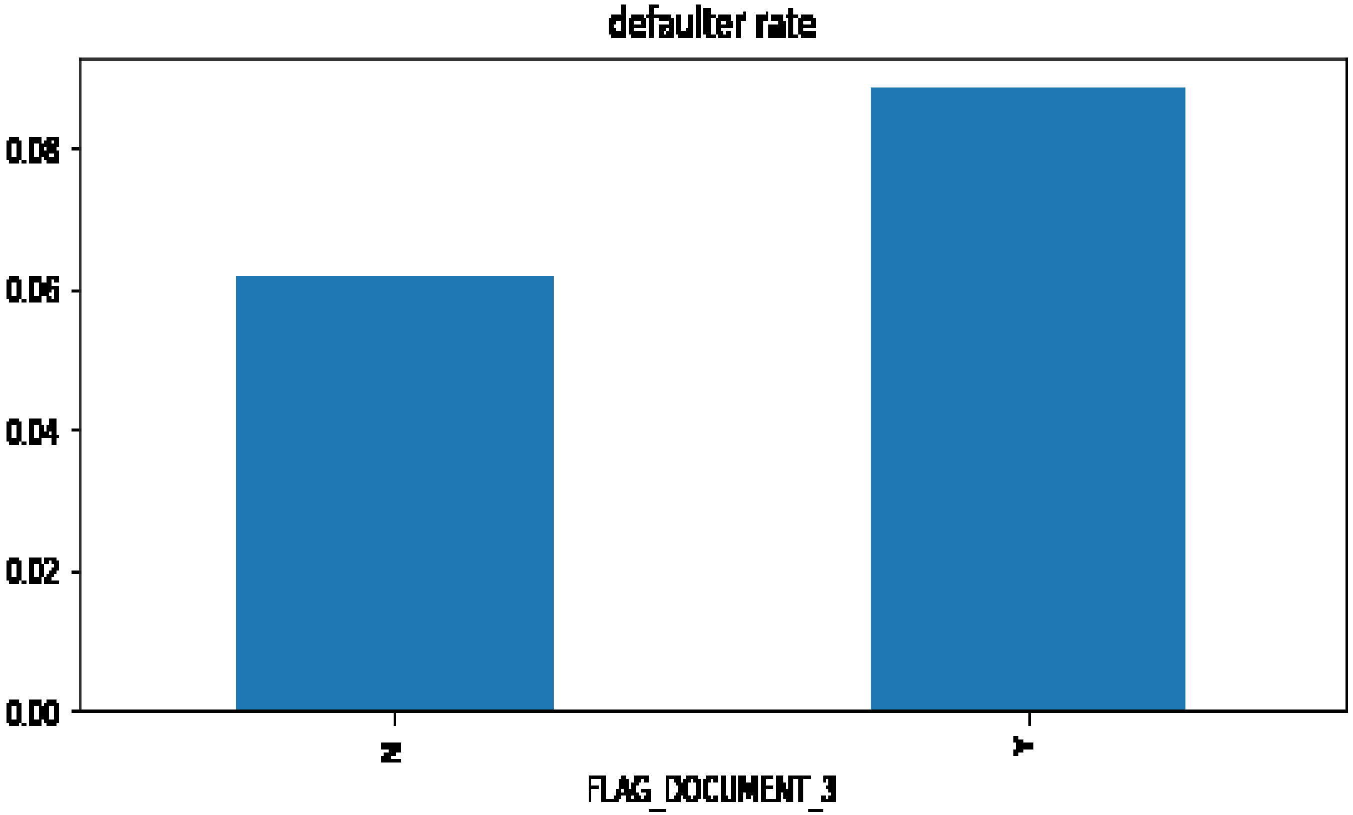


Defaulter had a tendency of giving work phone but not home phone in application

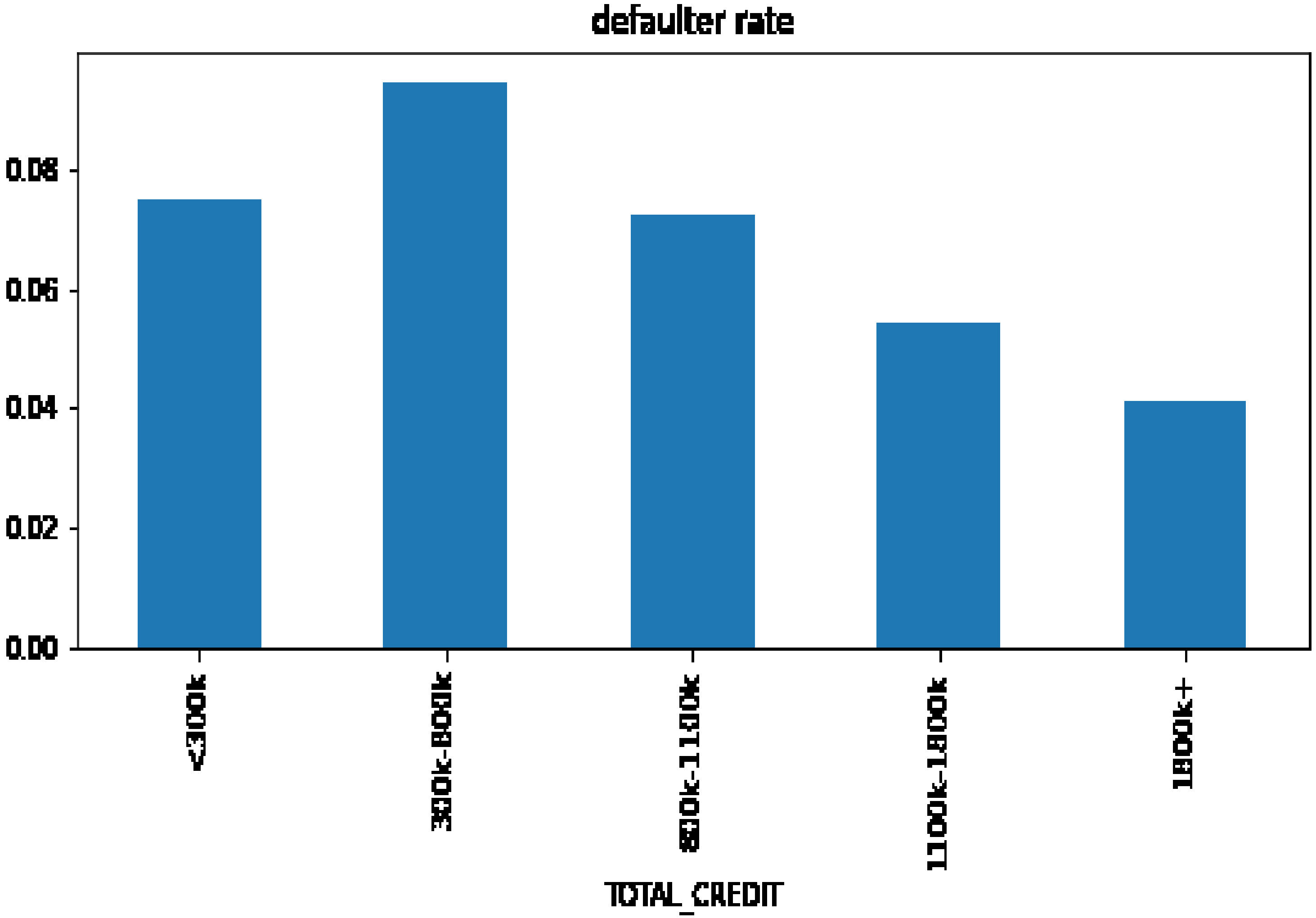




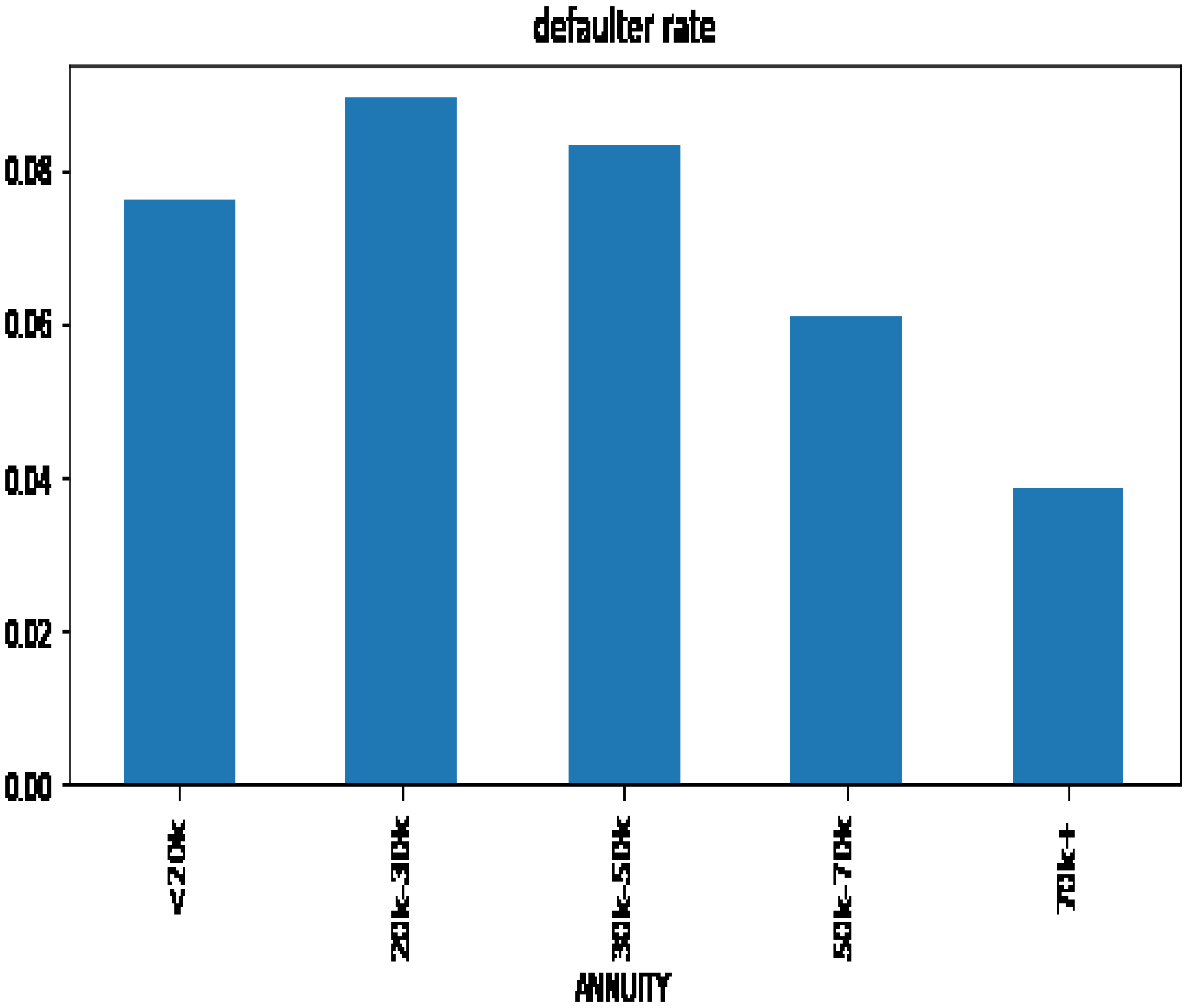
1)People who submitted document 3 had high Default rate



1)Default rate fall with credit amount.
2)Highest default rate(300k-800k)
3)Lowest default rate(>1800k)

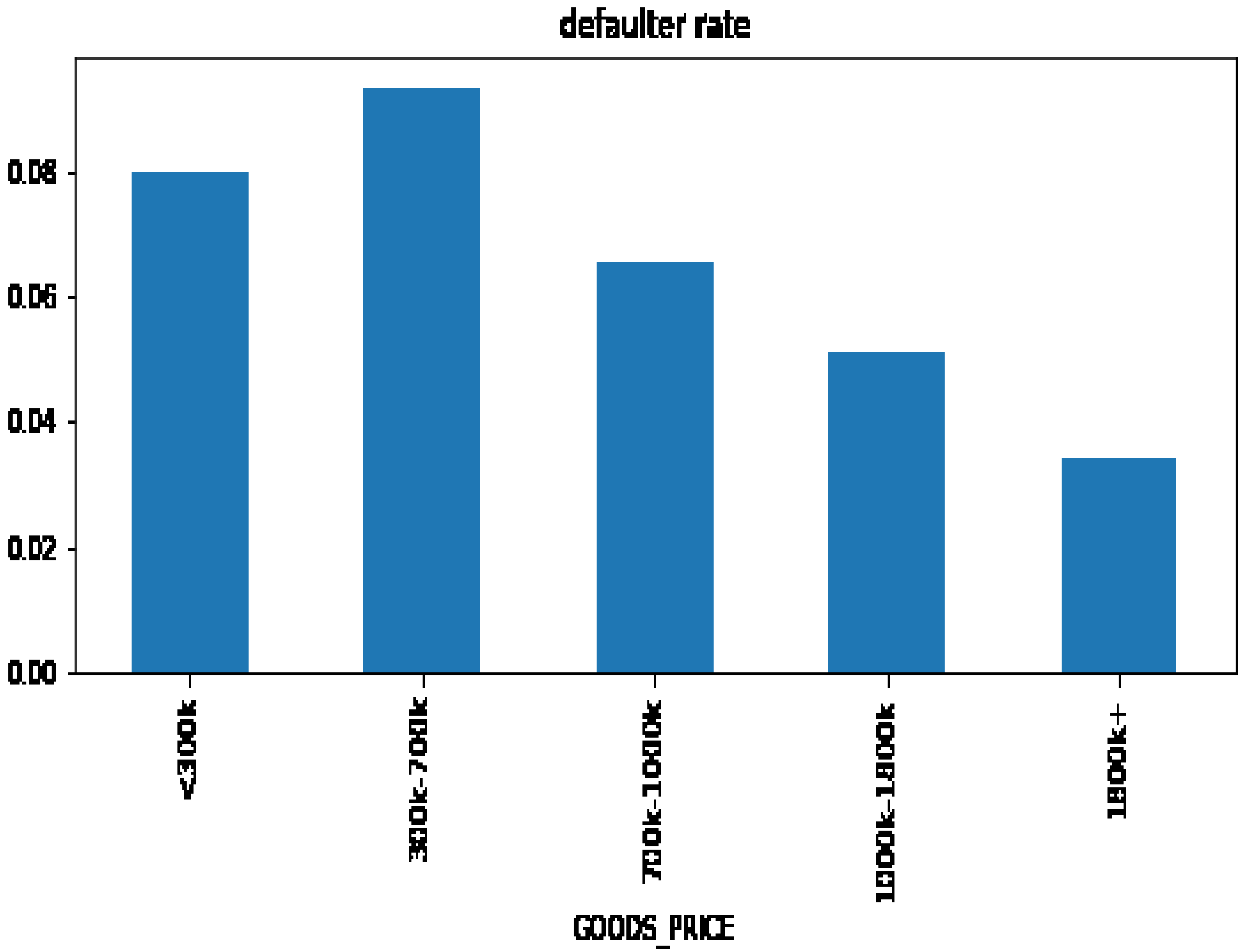


1)Default rate fall with annuity amount.
2)Highest default rate(20k-30k)
3)Lowest default rate(>70k)



1)Default rate fall with goods_price.
2)Highest default rate(300k-700k)
3)Lowest default rate(>1800k)

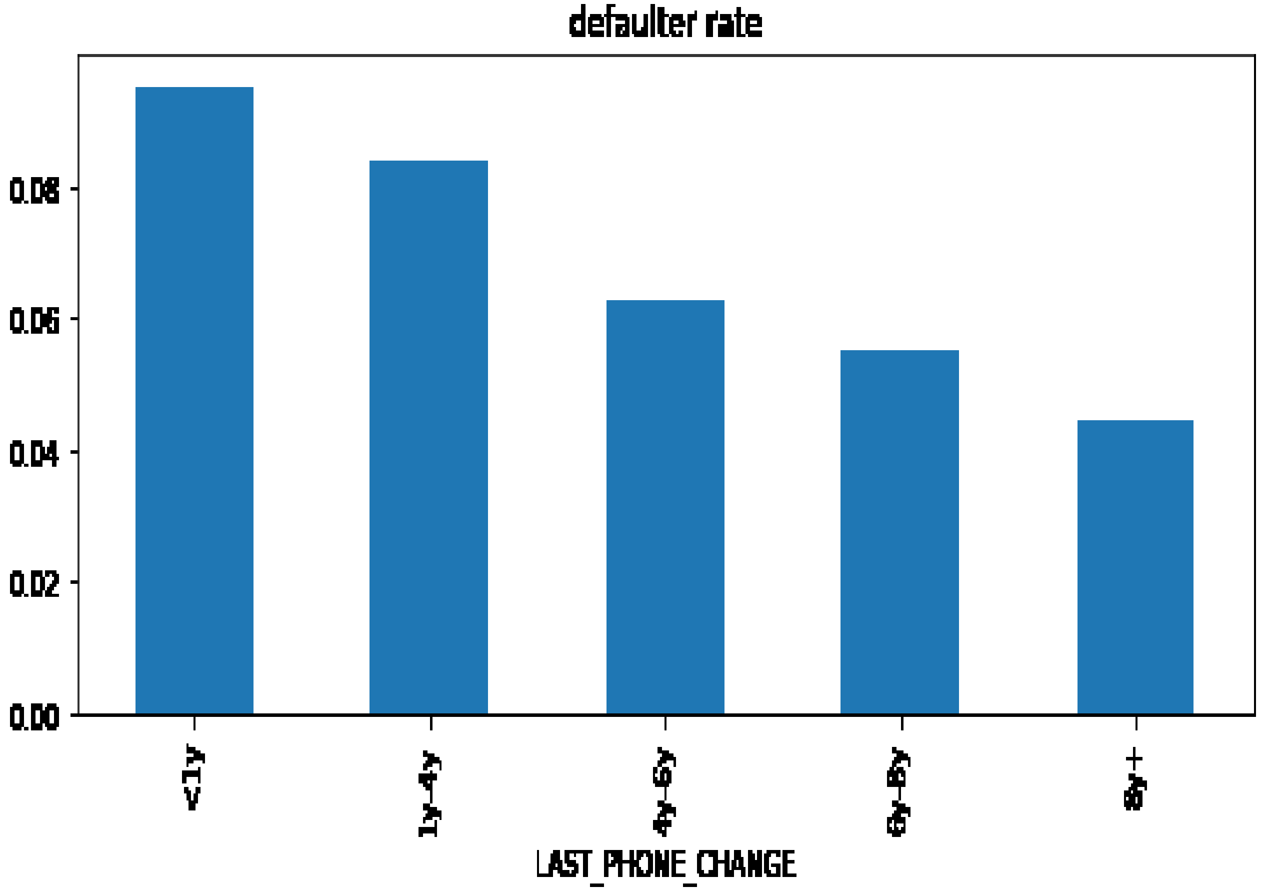
So credit,annuity and good_price showing very High correlation

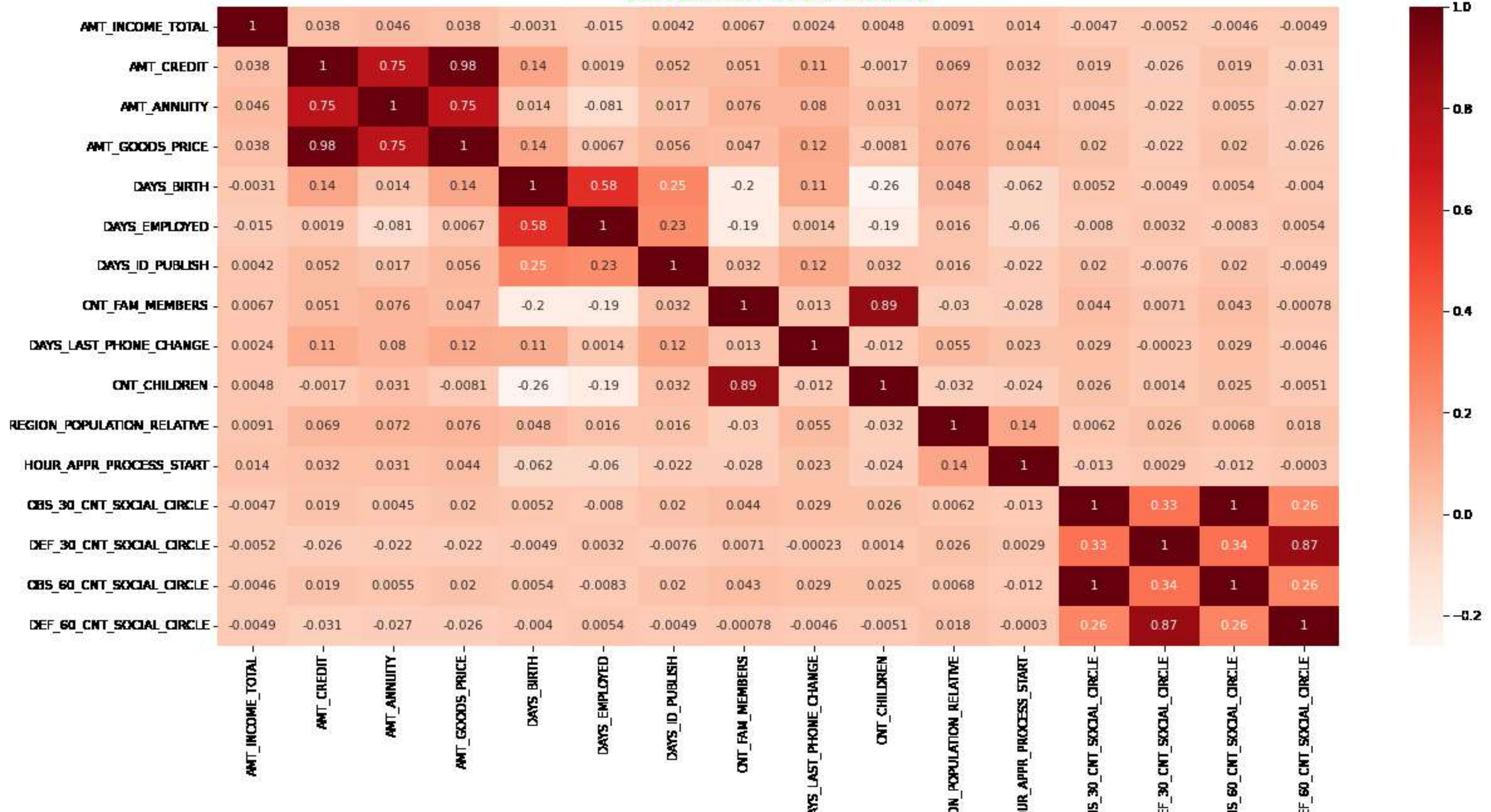


1)Defaulter seem to change their phone before Applying

2)People who changed(<1y)

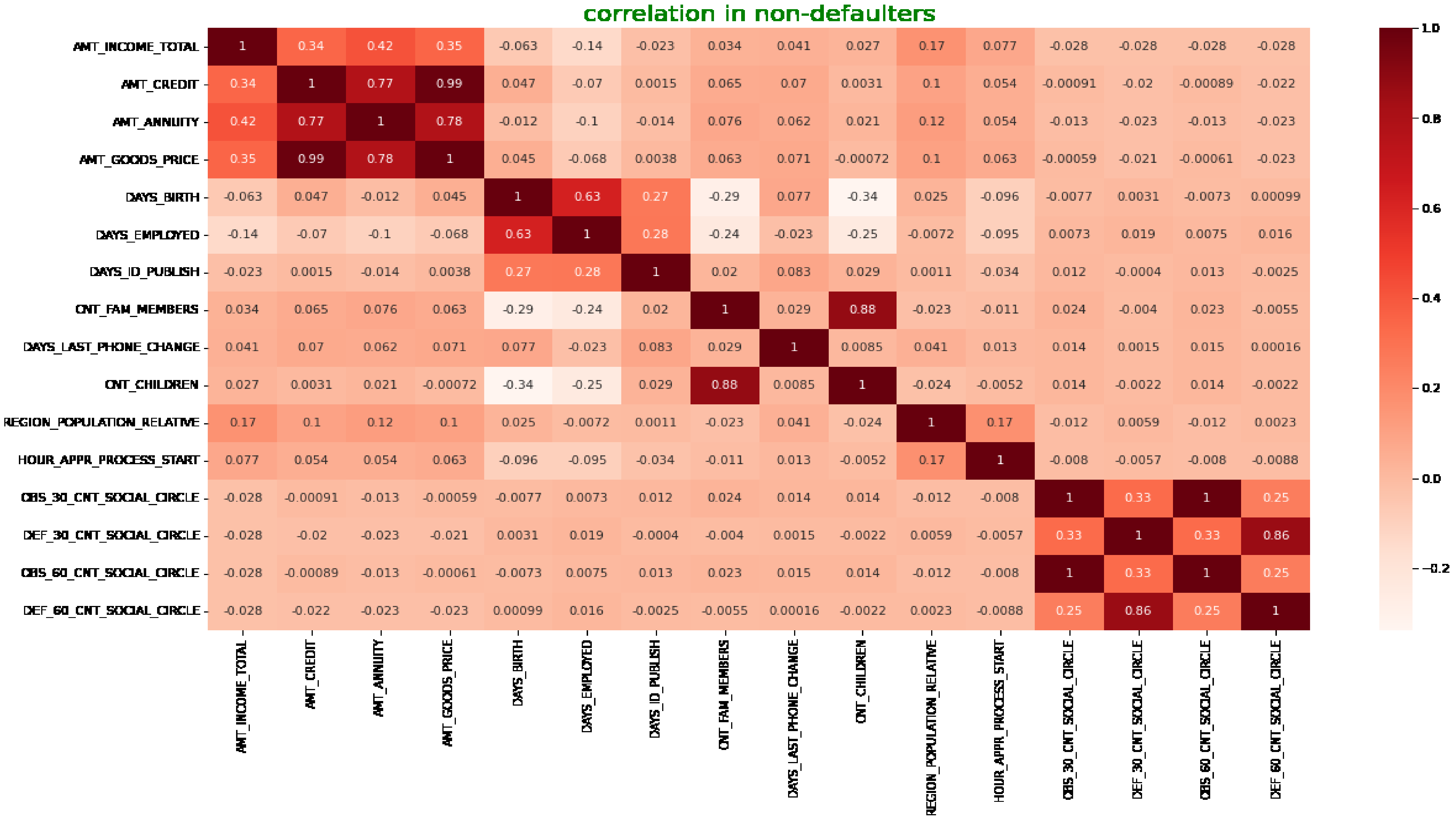
3)People who changed(>8y) are least defaulter





correlation in defaulters

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	AYS_LAST_PHONE_CHANGE	CNT_CHILDREN	ON_POPULATION_RELATIVE	UR_APPR_PROCESS_START	S_30_CNT_SOCIAL_CIRCLE	EF_30_CNT_SOCIAL_CIRCLE	S_60_CNT_SOCIAL_CIRCLE	EF_60_CNT_SOCIAL_CIRCLE
AMT_INCOME_TOTAL	1	0.038	0.046	0.038	-0.0031	-0.015	0.0042	0.0067	0.0024	0.0048	0.0091	0.014	-0.0047	-0.0052	-0.0046	-0.0049
AMT_CREDIT	0.038	1	0.75	0.98	0.14	0.0019	0.052	0.051	0.11	-0.0017	0.069	0.032	0.019	-0.026	0.019	-0.031
AMT_ANNUITY	0.046	0.75	1	0.75	0.014	-0.081	0.017	0.076	0.08	0.031	0.072	0.031	0.0045	-0.022	0.0055	-0.027
AMT_GOODS_PRICE	0.038	0.98	0.75	1	0.14	0.0067	0.056	0.047	0.12	-0.0081	0.076	0.044	0.02	-0.022	0.02	-0.026
DAYS_BIRTH	-0.0031	0.14	0.014	0.14	1	0.58	0.25	-0.2	0.11	-0.26	0.048	-0.062	0.0052	-0.0049	0.0054	-0.004
DAYS_EMPLOYED	-0.015	0.0019	-0.081	0.0067	0.58	1	0.23	-0.19	0.0014	-0.19	0.016	-0.06	-0.008	0.0032	-0.0083	0.0054
DAYS_ID_PUBLISH	0.0042	0.052	0.017	0.056	0.25	0.23	1	0.032	0.12	0.032	0.016	-0.022	0.02	-0.0076	0.02	-0.0049
CNT_FAM_MEMBERS	0.0067	0.051	0.076	0.047	-0.2	-0.19	0.032	1	0.013	0.89	-0.03	-0.028	0.044	0.0071	0.043	-0.00078
DAYS_LAST_PHONE_CHANGE	0.0024	0.11	0.08	0.12	0.11	0.0014	0.12	0.013	1	-0.012	0.055	0.023	0.029	-0.00023	0.029	-0.0046
CNT_CHILDREN	0.0048	-0.0017	0.031	-0.0081	-0.26	-0.19	0.032	0.89	-0.012	1	-0.032	-0.024	0.026	0.0014	0.025	-0.0051
REGION_POPULATION_RELATIVE	0.0091	0.069	0.072	0.076	0.048	0.016	0.016	-0.03	0.055	-0.032	1	0.14	0.0062	0.026	0.0068	0.018
HOUR_APPR_PROCESS_START	0.014	0.032	0.031	0.044	-0.062	-0.06	-0.022	-0.028	0.023	-0.024	0.14	1	-0.013	0.0029	-0.012	-0.0003
OBS_30_CNT_SOCIAL_CIRCLE	-0.0047	0.019	0.0045	0.02	0.0052	-0.008	0.02	0.044	0.029	0.026	0.0062	-0.013	1	0.33	1	0.26
DEF_30_CNT_SOCIAL_CIRCLE	-0.0052	-0.026	-0.022	-0.022	-0.0049	0.0032	-0.0076	0.0071	-0.00023	0.0014	0.026	0.0029	0.33	1	0.34	0.87
OBS_60_CNT_SOCIAL_CIRCLE	-0.0046	0.019	0.0055	0.02	0.0054	-0.0083	0.02	0.043	0.029	0.025	0.0068	-0.012	1	0.34	1	0.26
DEF_60_CNT_SOCIAL_CIRCLE	-0.0049	-0.031	-0.027	-0.026	-0.004	0.0054	-0.0049	-0.00078	-0.0046	-0.0051	0.018	-0.0003	0.26	0.87	0.26	1



correlation in non-defaulters

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	DAYS_LAST_PHONE_CHANGE	CNT_CHILDREN	REGION_POPULATION_RELATIVE	HOUR_APPR_PROCESS_START	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE
AMT_INCOME_TOTAL	1	0.34	0.42	0.35	-0.063	-0.14	-0.023	0.034	0.041	0.027	0.17	0.077	-0.028	-0.028	-0.028	-0.028
AMT_CREDIT	0.34	1	0.77	0.99	0.047	-0.07	0.0015	0.065	0.07	0.0031	0.1	0.054	-0.00091	-0.02	-0.00089	-0.022
AMT_ANNUITY	0.42	0.77	1	0.78	-0.012	-0.1	-0.014	0.076	0.062	0.021	0.12	0.054	-0.013	-0.023	-0.013	-0.023
AMT_GOODS_PRICE	0.35	0.99	0.78	1	0.045	-0.068	0.0038	0.063	0.071	-0.00072	0.1	0.063	-0.00059	-0.021	-0.00061	-0.023
DAYS_BIRTH	-0.063	0.047	-0.012	0.045	1	0.63	0.27	-0.29	0.077	-0.34	0.025	-0.096	-0.0077	0.0031	-0.0073	0.00099
DAYS_EMPLOYED	-0.14	-0.07	-0.1	-0.068	0.63	1	0.28	-0.24	-0.023	-0.25	-0.0072	-0.095	0.0073	0.019	0.0075	0.016
DAYS_ID_PUBLISH	-0.023	0.0015	-0.014	0.0038	0.27	0.28	1	0.02	0.083	0.029	0.0011	-0.034	0.012	-0.0004	0.013	-0.0025
CNT_FAM_MEMBERS	0.034	0.065	0.076	0.063	-0.29	-0.24	0.02	1	0.029	0.88	-0.023	-0.011	0.024	-0.004	0.023	-0.0055
DAYS_LAST_PHONE_CHANGE	0.041	0.07	0.062	0.071	0.077	-0.023	0.083	0.029	1	0.0085	0.041	0.013	0.014	0.0015	0.015	0.00016
CNT_CHILDREN	0.027	0.0031	0.021	-0.00072	-0.34	-0.25	0.029	0.88	0.0085	1	-0.024	-0.0052	0.014	-0.0022	0.014	-0.0022
REGION_POPULATION_RELATIVE	0.17	0.1	0.12	0.1	0.025	-0.0072	0.0011	-0.023	0.041	-0.024	1	0.17	-0.012	0.0059	-0.012	0.0023
HOUR_APPR_PROCESS_START	0.077	0.054	0.054	0.063	-0.096	-0.095	-0.034	-0.011	0.013	-0.0052	0.17	1	-0.008	-0.0057	-0.008	-0.0088
OBS_30_CNT_SOCIAL_CIRCLE	-0.028	-0.00091	-0.013	-0.00059	-0.0077	0.0073	0.012	0.024	0.014	0.014	-0.012	-0.008	1	0.33	1	0.25
DEF_30_CNT_SOCIAL_CIRCLE	-0.028	-0.02	-0.023	-0.021	0.0031	0.019	-0.0004	-0.004	0.0015	-0.0022	0.0059	-0.0057	0.33	1	0.33	0.86
OBS_60_CNT_SOCIAL_CIRCLE	-0.028	-0.00089	-0.013	-0.00061	-0.0073	0.0075	0.013	0.023	0.015	0.014	-0.012	-0.008	1	0.33	1	0.25
DEF_60_CNT_SOCIAL_CIRCLE	-0.028	-0.022	-0.023	-0.023	0.00099	0.016	-0.0025	-0.0055	0.00016	-0.0022	0.0023	-0.0088	0.25	0.86	0.25	1

top 10 correlation for defaulters

1)amt_credit vs amt_goods_price(corr=0.98)

2)cnt_family_member vs cnt_children(corr=0.89)

3)DEF_30_CNT_SOCIAL_CIRCLE vs DEF_60_CNT_SOCIAL_CIRCLE(corr=0.87)

4)amt_credit vs amt_annuity(corr=0.75)

5)amt_annuity vs amt_goods_price(corr=0.75)

6)days_birth vs days_employed(corr=0.58)

7)DEF_30_CNT_SOCIAL_CIRCLE vs OBS_60_CNT_SOCIAL_CIRCLE(corr=0.34)

8)DEF_30_CNT_SOCIAL_CIRCLE vs OBS_30_CNT_SOCIAL_CIRCLE(corr=0.33)

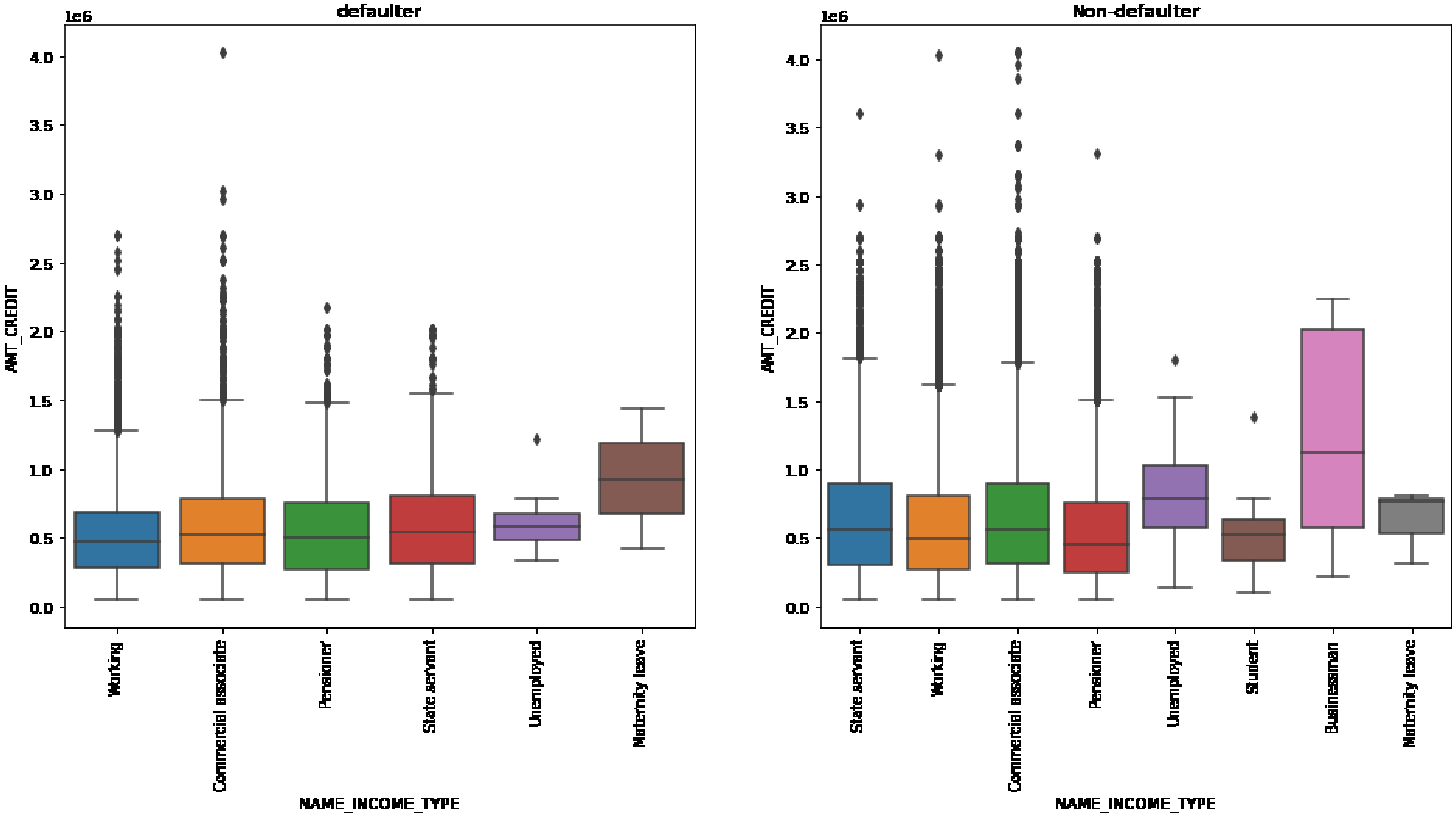
9)DEF_60_CNT_SOCIAL_CIRCLE vs OBS_30_CNT_SOCIAL_CIRCLE(corr=0.26)

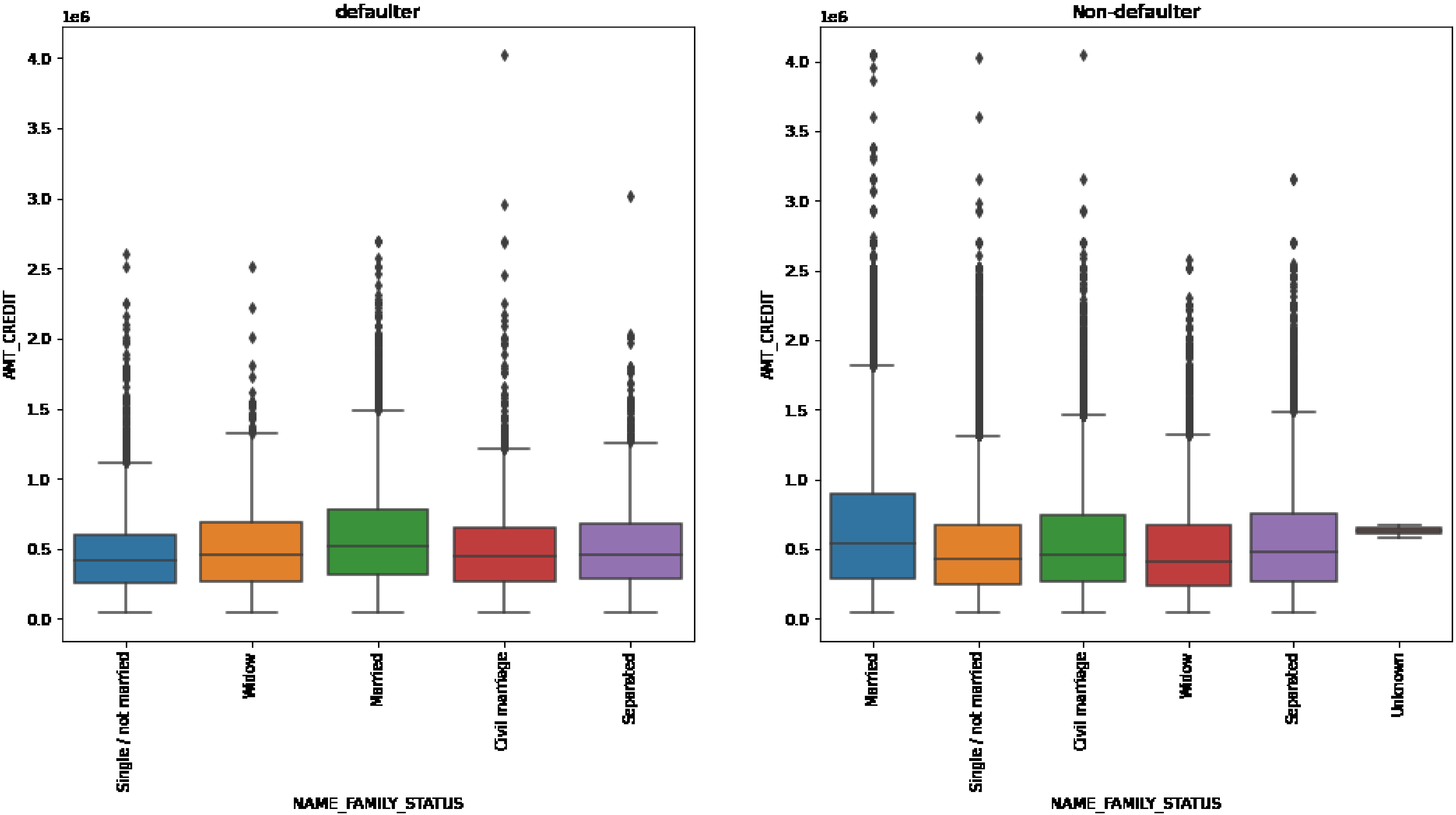
10)days_birth vs days_id_publish(corr=0.25)

Top 10 correlation in non-defaulter

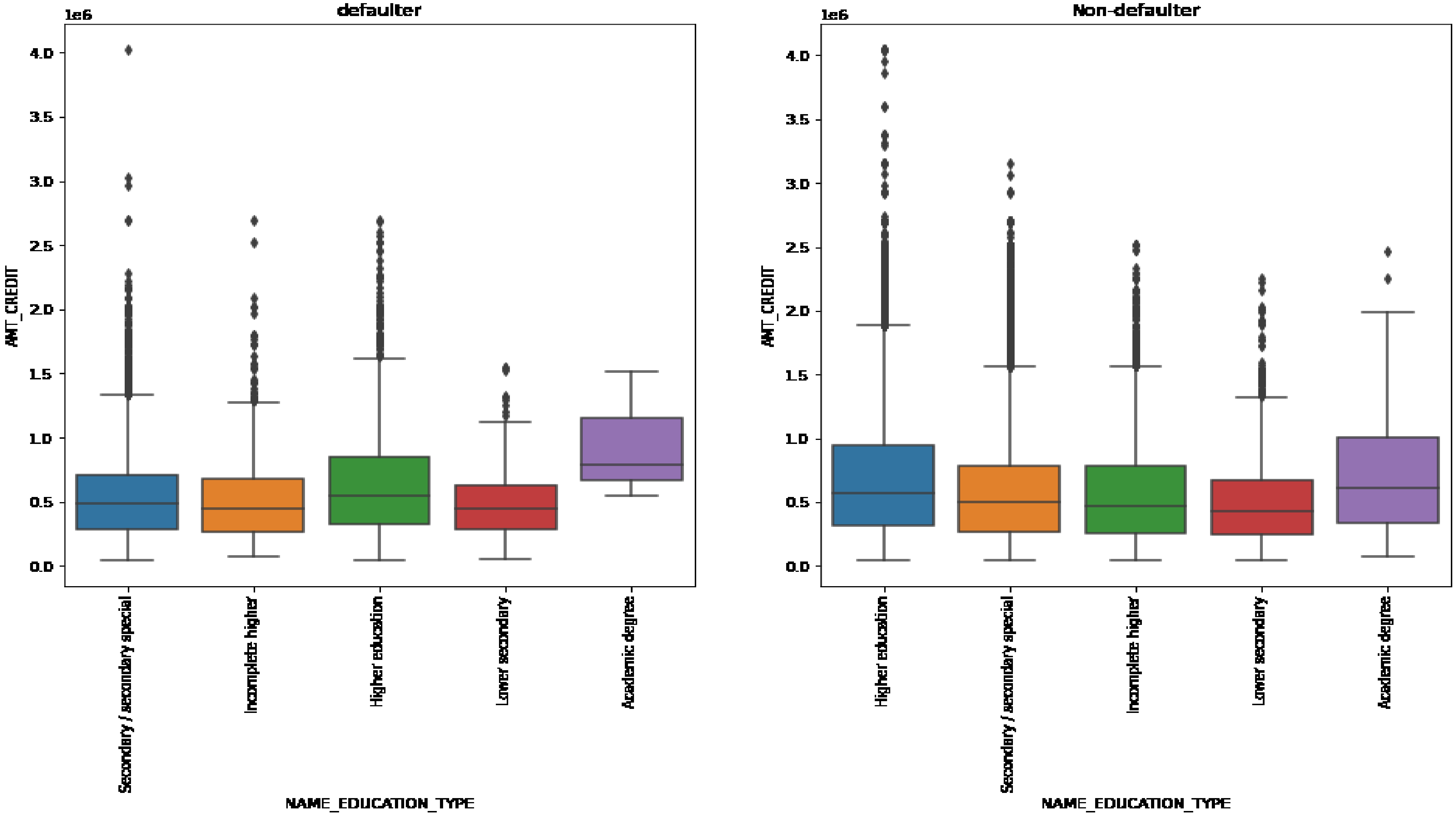
1)amt_credit vs amt_goods_price(corr=0.99)
2)cnt_family_member vs cnt_children(corr=0.88)
3)DEF_30_CNT_SOCIAL_CIRCLE vs DEF_60_CNT_SOCIAL_CIRCLE(corr=0.86)
4)amt_annuity vs amt_goods_price(corr=0.78)
5)amt_credit vs amt_annuity(corr=0.77)
6)days_birth vs days_employed(corr=0.63)
7)DEF_30_CNT_SOCIAL_CIRCLE vs OBS_60_CNT_SOCIAL_CIRCLE(corr=0.33)
8)DEF_30_CNT_SOCIAL_CIRCLE vs OBS_30_CNT_SOCIAL_CIRCLE(corr=0.33)
9)days_employed vs days_id_publish(corr=0.28)
10)DEF_60_CNT_SOCIAL_CIRCLE vs OBS_30_CNT_SOCIAL_CIRCLE(corr=0.25)

observation-sequence of top 10 correlation in defaulter and non-defaulter is almost same


defaulter
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
AMT_CREDIT
Working
Commercial associate
Pensioner
State servant
Unemployed
Maternity leave
NAME_INCOME_TYPE
Non-defaulter
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
AMT_CREDIT
State servant
Working
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
NAME_INCOME_TYPE



defaulter
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
AMT_CREDIT
Single / not married
Widow
Married
Civil marriage
Separated
NAME_FAMILY_STATUS
Non-defaulter
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
AMT_CREDIT
Married
Single / not married
Civil marriage
Widow
Separated
Unknown
NAME_FAMILY_STATUS


defaulter
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
AMT_CREDIT
Secondary / secondary special
Incomplete higher
Higher education
Lower secondary
Academic degree
NAME_EDUCATION_TYPE
Non-defaulter
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
AMT_CREDIT
Higher education
Secondary / secondary special
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE

observations

1. **women in maternity leave who applied for high credit had very high chance of default**

2. **to give small credit loan(below 800k),students are good choice**

3. **unemployed people are comparatively safer to give big credit(above 8ook)**

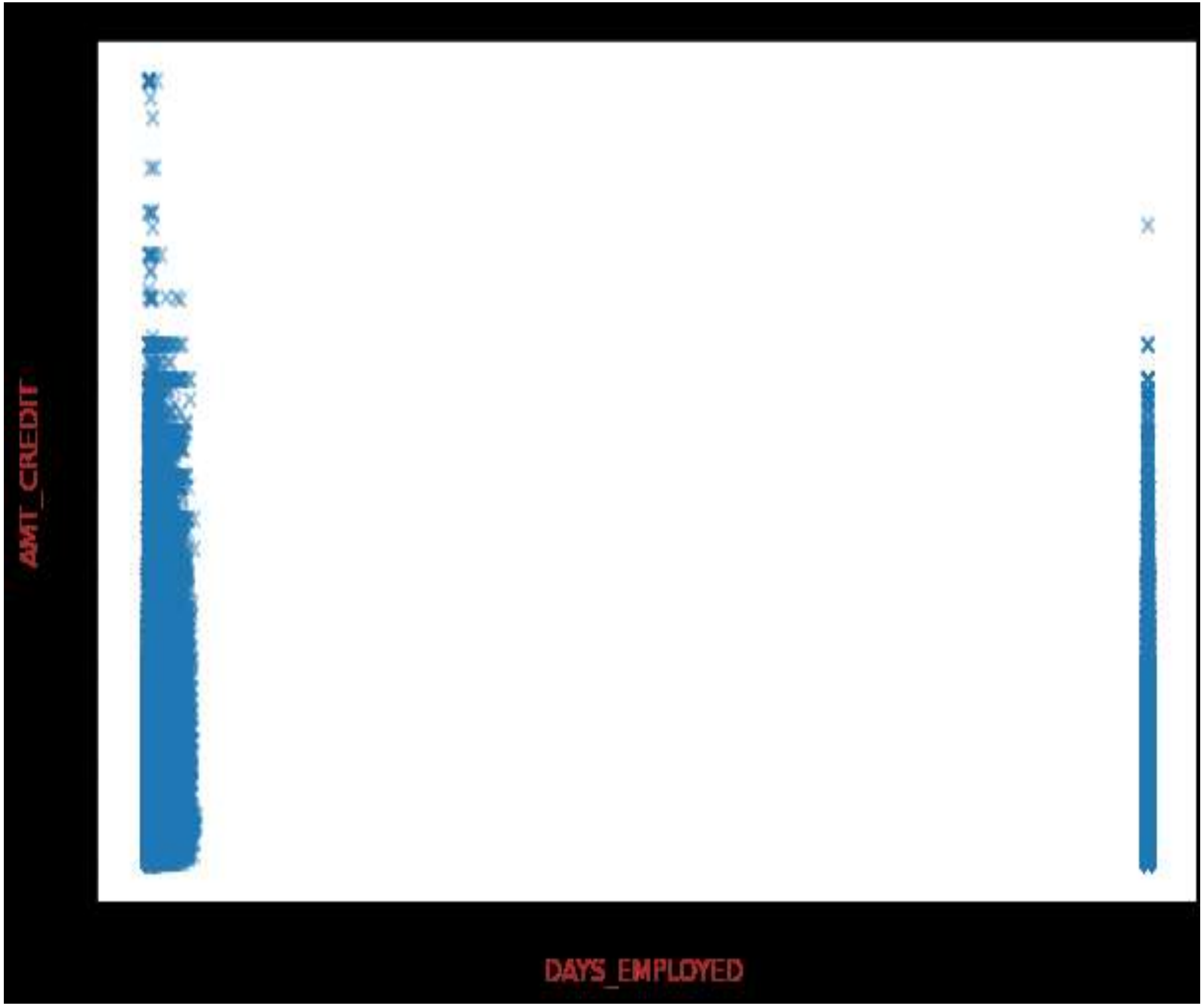
4. **married person are comparative safe to give loan,only when you are giving above 800k credit(just increase credit amount to ensure further more safety)**

5. **below 500k credit,person with academic degree is very safe**

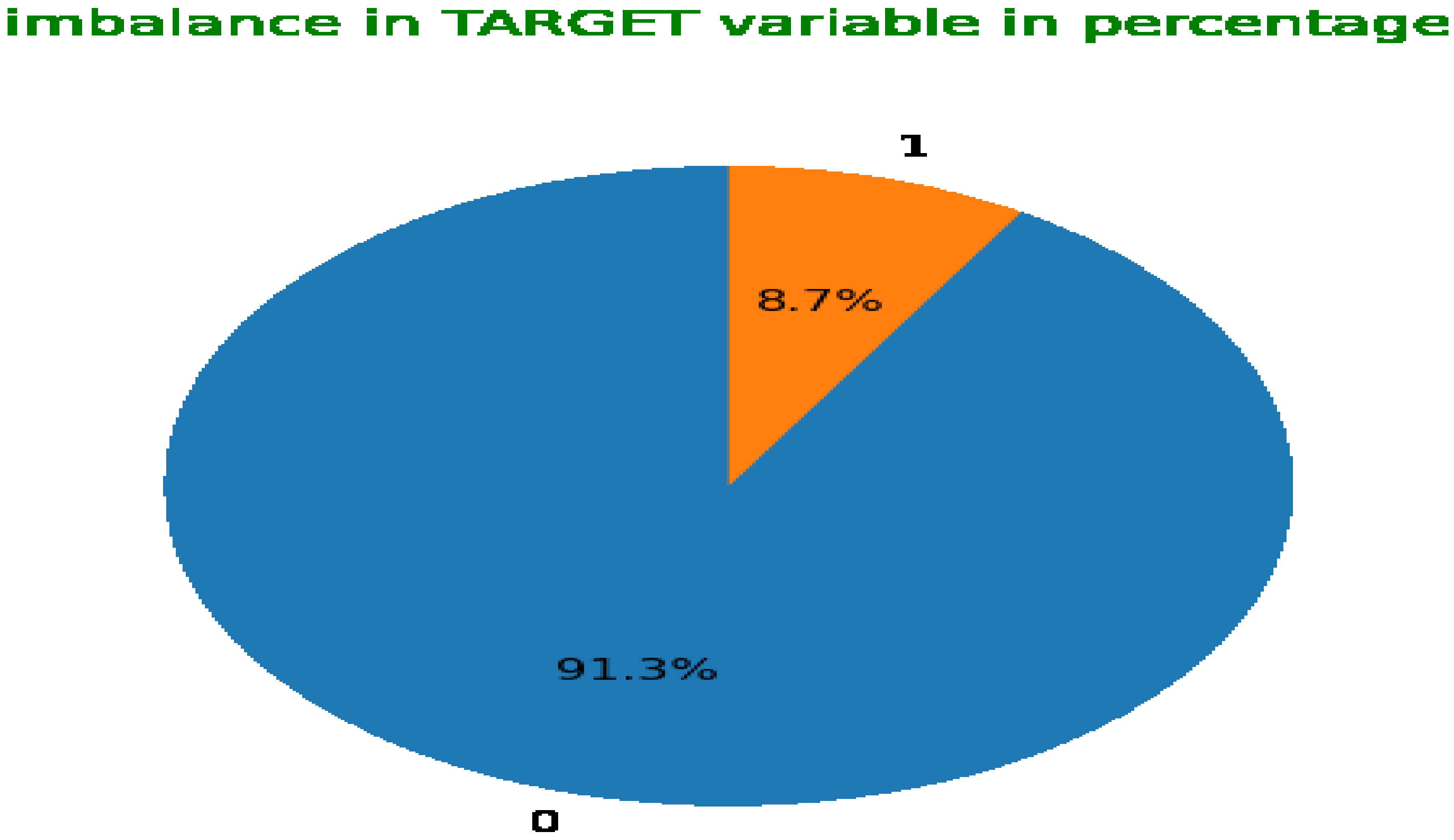
6. **for giving loan to person with secondary education go beyond 1000k(toward high credit)**

7. **person with lower secondary education is highestrisk,choosing those who ask for credit beyond 1500k ensure safety**

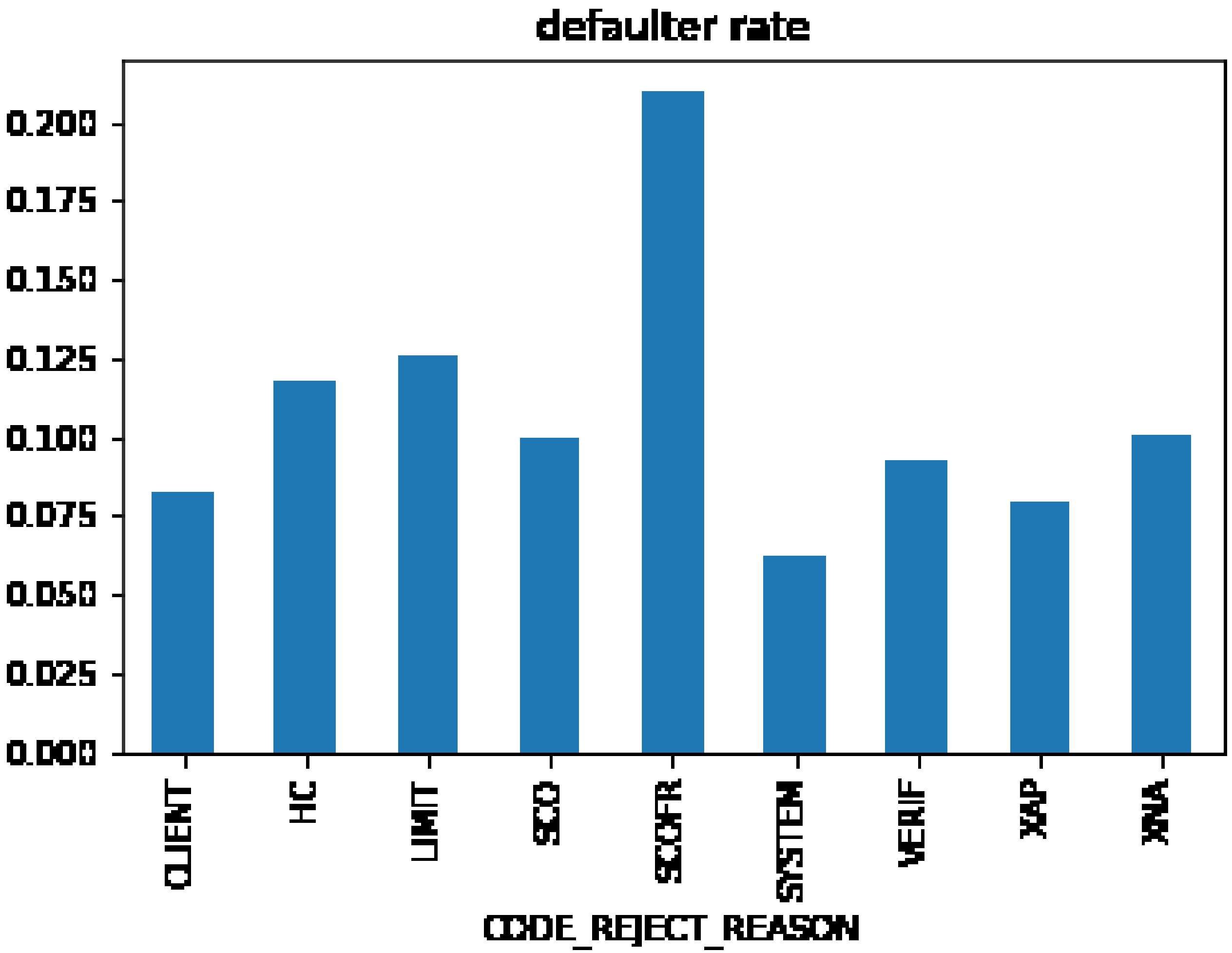
as you move toward higher days of employment,defaulting chance reduces

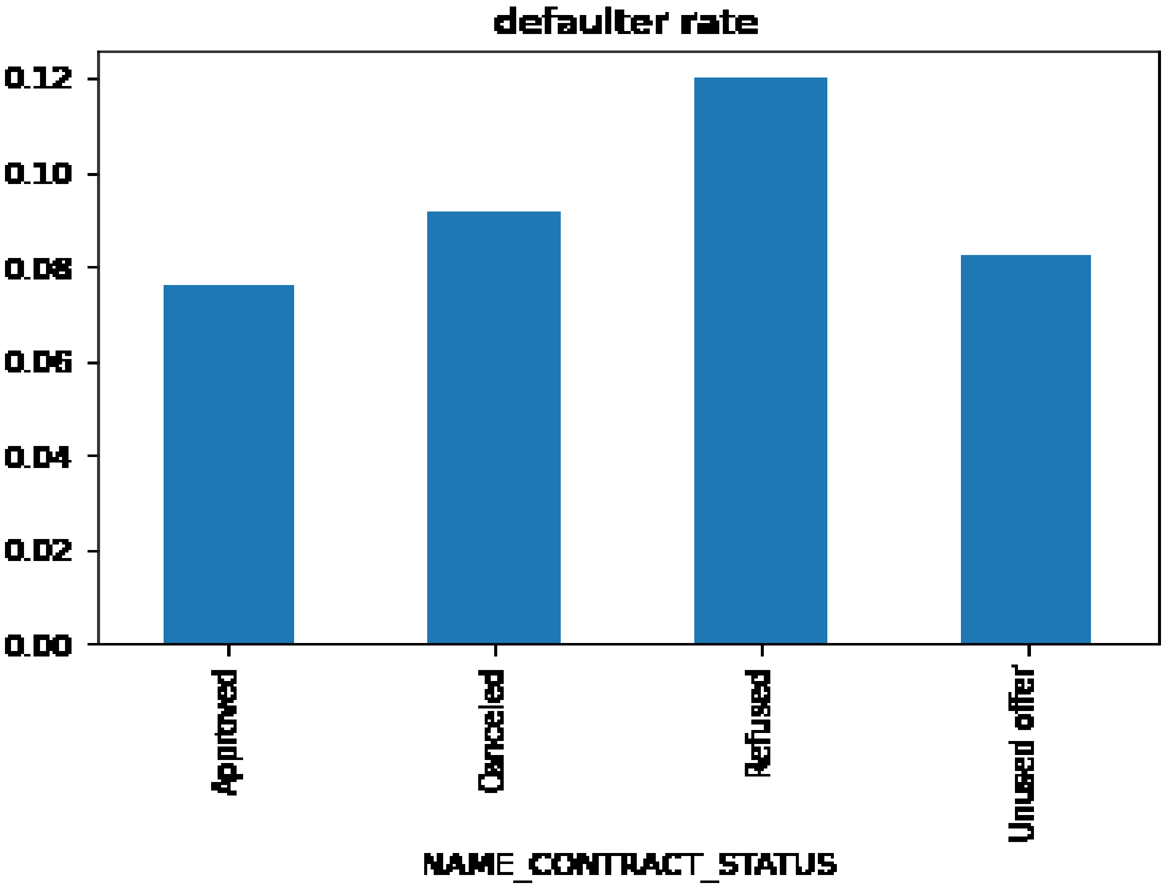


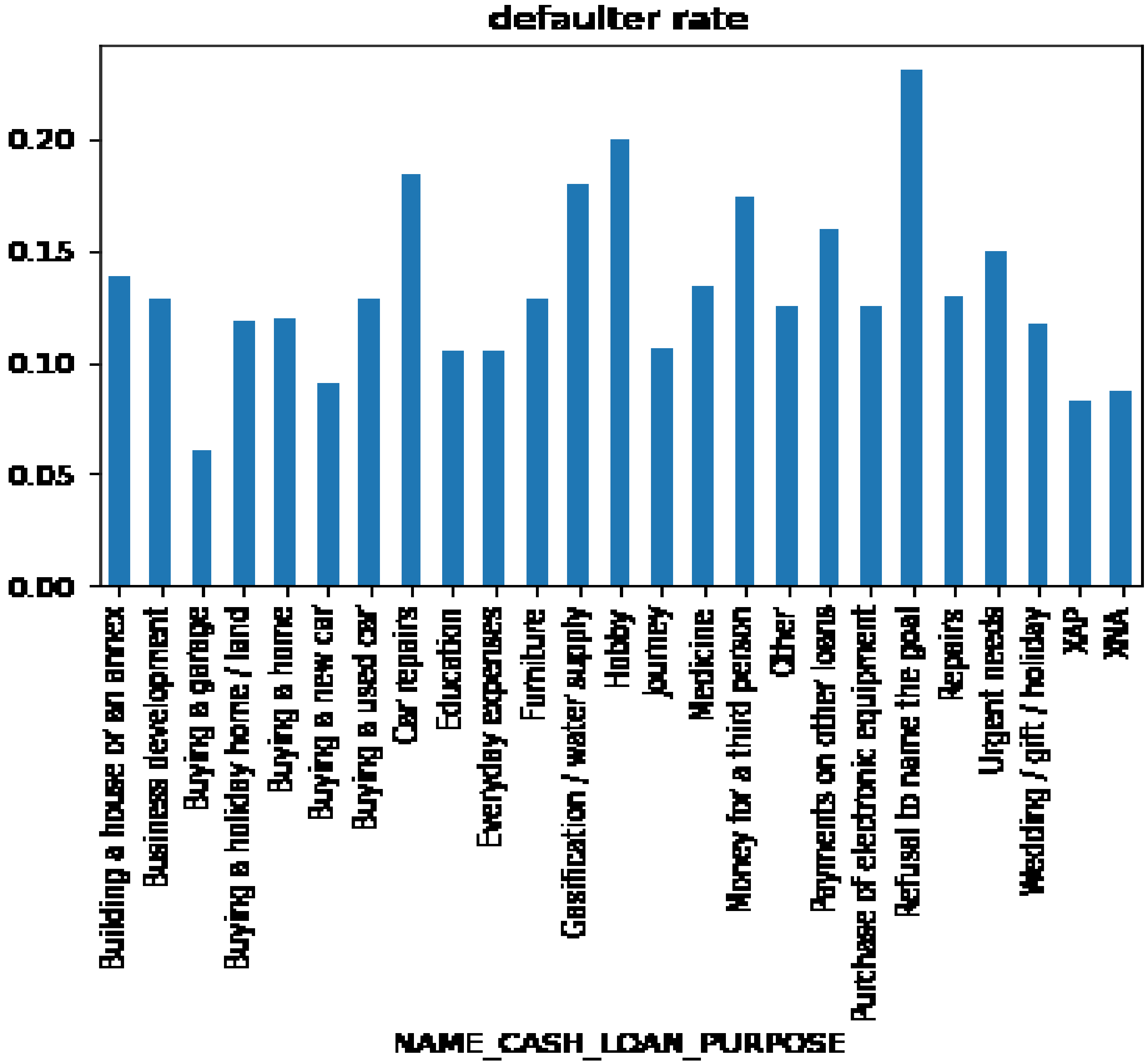
For previous applicants,8.7% are defaults

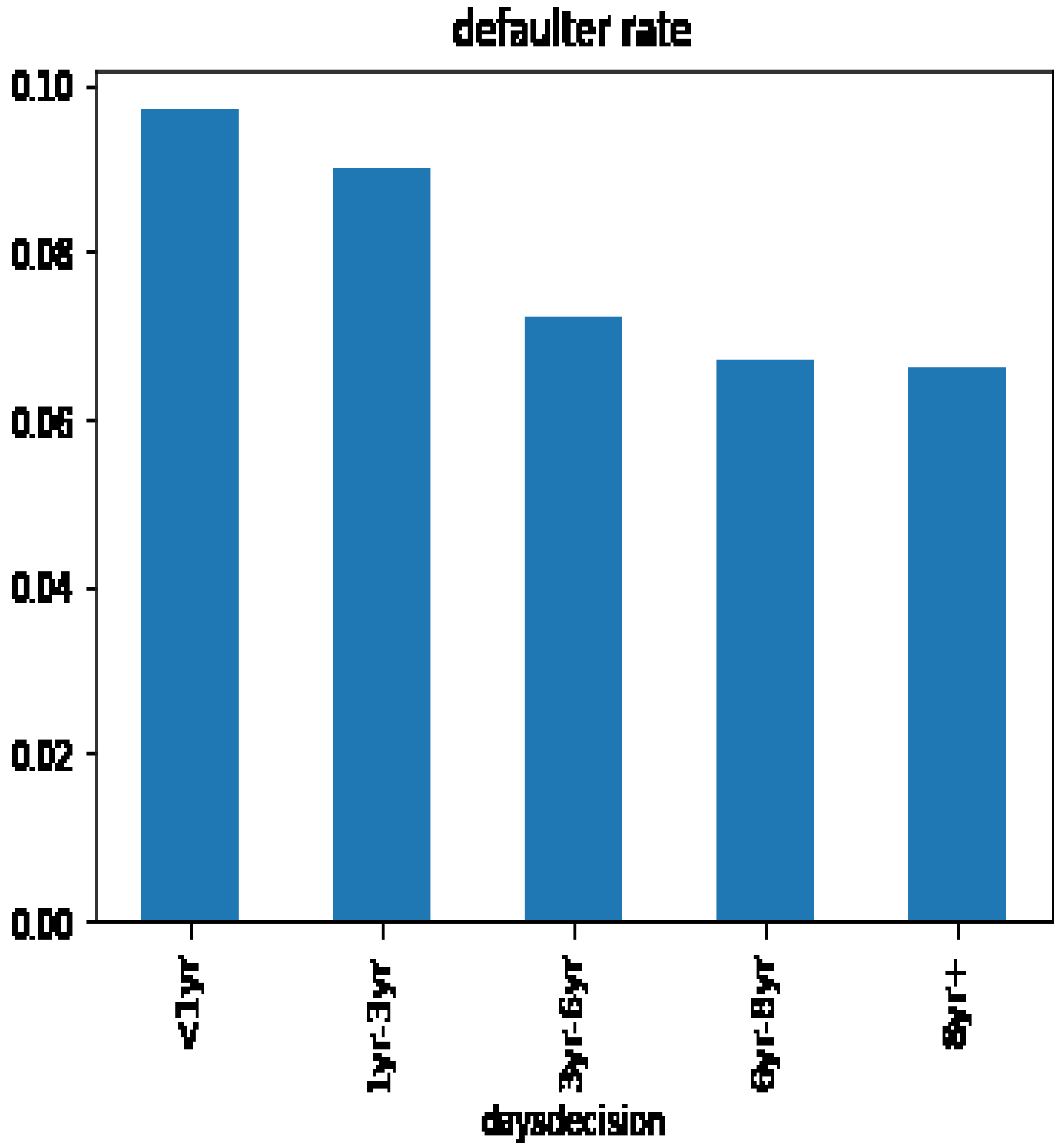


Previous data insights






defaulter rate
0.00
0.05
0.10
0.15
0.20
Building a house or an annex
Business development
Buying a garage
Buying a holiday home / land
Buying a home
Buying a new car
Buying a used car
Car repairs
Education
Everyday expenses
Furniture
Gasification / water supply
Hobby
Journey
Medicine
Money for a third person
Other
Payments on other loans
Purchase of electronic equipment
Refusal to name the goal
Repairs
Urgent needs
Wedding / gift / holiday
XAP
XNA
NAME_CASH_LOAN_PURPOSE


defaulter rate
0.00
0.02
0.04
0.06
0.08
0.10
<1yr
1yr-3yr
3yr-6yr
6yr-8yr
8yr+
daysdecision

CONCLUSIONS ABOUT DEFAULTERS

1)firstly,around 8% of population is found to be defaulter and mens are more likely to be defaulter then females
2)education=defaulter rate increases with decline in education so people with lower secondary education and then secondary education had highest default rate
3)income type=unemployed(second highest defaulter rate) and women in maternity(highest defaulter rate) had highest default rate
4)family status=in defaulter rate,civil married is highest followed by single(second)
5)housing status=people living in rented apartment had highest share of defaulters,followed by those who live with parents
6)age=defaulter rate falls with age so people under 30 and then in 30's had highest default rate
7)income=default rate increase with fall of income. people with total income below 200000 are highest in risk
8)occupation type=low skill labourers had highest default rate.other serving class like waiter,driver,labourers are also at high risk
9)contract type=people applying for cash loan had high default rate
10)own_car=people who dont own car had high chance of defaulting
11)house/flat=people who dont own real state(house/flat) had higher risk of default
12)in defaulter there is a pattern that they provide work phone but not home phone
13)doc_3=people who submitted this document had higher default rate
14)total_credit=people who applied for credit of 3lac to 8lac had highest default rate then default rate fall with credit

15)annuity and goods price-following same trend as credit, people with annuity(20k-30k) and good price(300k-700k) had highest default rate,falling with increase in value
16)changing phone=defaulter had tendency of changing phone right before application.people who change phone(<1yr) are highest in risk.default rate decreases with time
17)women in maternity leave who applied for high credit had very high chance of default
18)employment_days=people recently employed had highest chance of default
19)persons who was rejected in prev app due to reject code SCOFR,turned defaulter in most number
20)defaulter had a tendency to refuse to name the goal for which they applied for cash loan,as observed
21)defaulters seem to apply for next loan application quicker than non-defaulter

CONCLUSION ABOUT NON-DEFAULTERS

1)firstly,around 92% of population is found to be non-defaulter and females are more likely to be non-defaulter then males
2)education=non-defaulter tendency increases with education level so people with academic degree had least chance of default
3)income type=students and buissnessman are most likely to be non-defaulter(minimum risk)
4)family status=widow is least risky
5)housing status=people living at office apartment are least risky
6)age=defaulter rate falls with age so people with age 60+ are least risky
7)income=default rate falls with increase in income. people with total income above 500k are least in risk
8)occupation type=accountant,core staff,hr staff,IT staffs are safest
9)contract type=people applying for revolving loan are safe
10)own_car=people who own car had non-defaulting tendency
11)house/flat=people who own real state(house/flat) had non-default tendency
12)non-defaulter provides home phone
13)total_credit= default rate fall with credit amount,people applied for 1800k credit are safest
14)annuity and goods price-following same trend as credit, people with annuity(>70k) and good price(>1800k) had least default rate

15)changing phone=as high the duration with which person didnt change his phone from app date,that much low is risk.person who didnt change phone from 8y+ had leat risk
16)for small credit(which is risky),students are safest.
17)for unemployed people(who are risky),its safer to give big credit(above 8ook)
18)married person are comparative safe to give loan,only when you are giving above 800k credit(just increase credit amount to ensure further more safety)
19)for giving loan to person with secondary eucation(second high in risk) go beyond 1000k(toward high credit)
20)for person with lower secondary education(highest risk),choosing those who ask for credit beyond 1500k ensure safety
21)as you move toward higher days of employment,defaulting chance reduces
22)person whose early application rejected due to system(code reject reason) is least risky