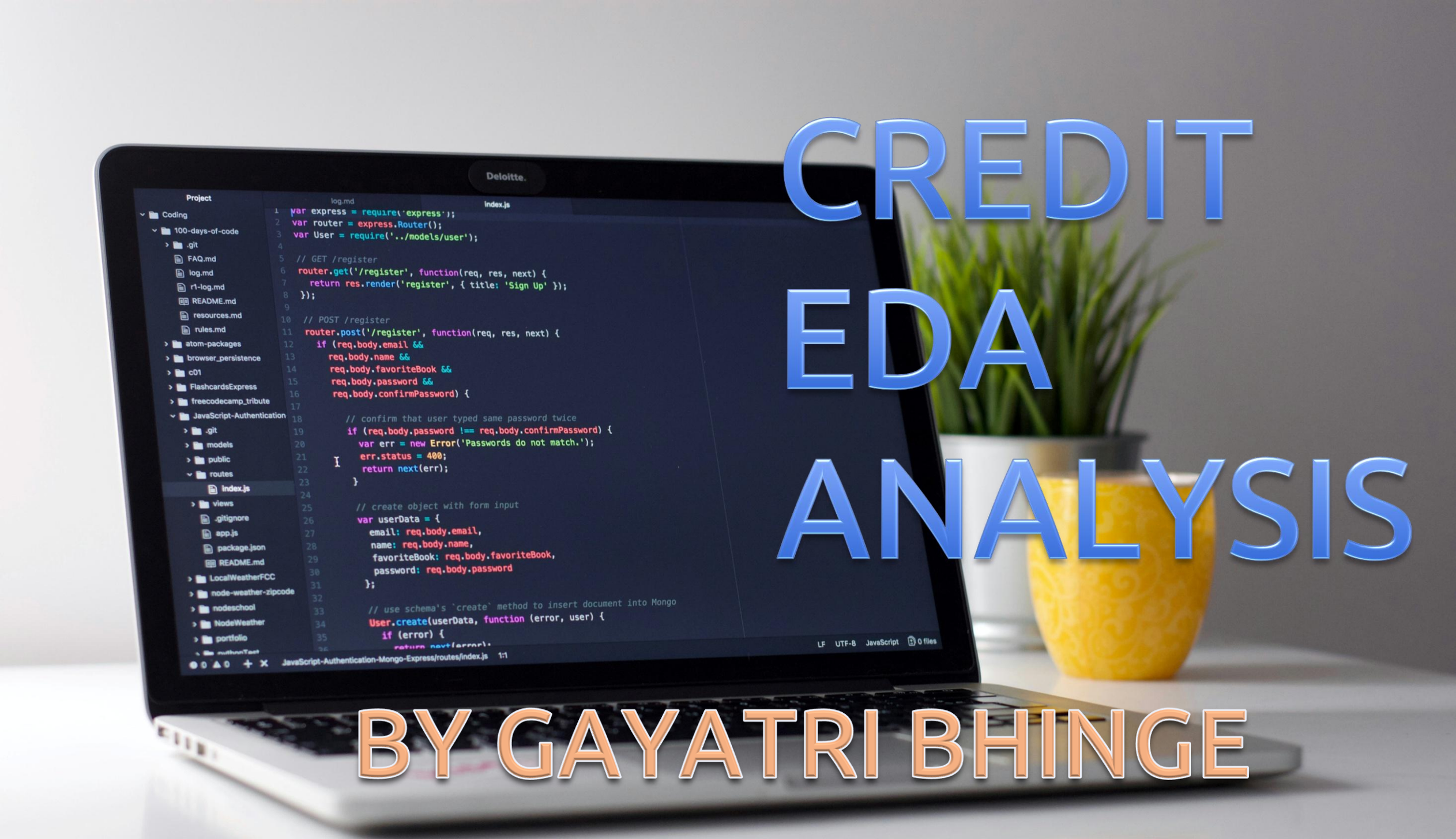
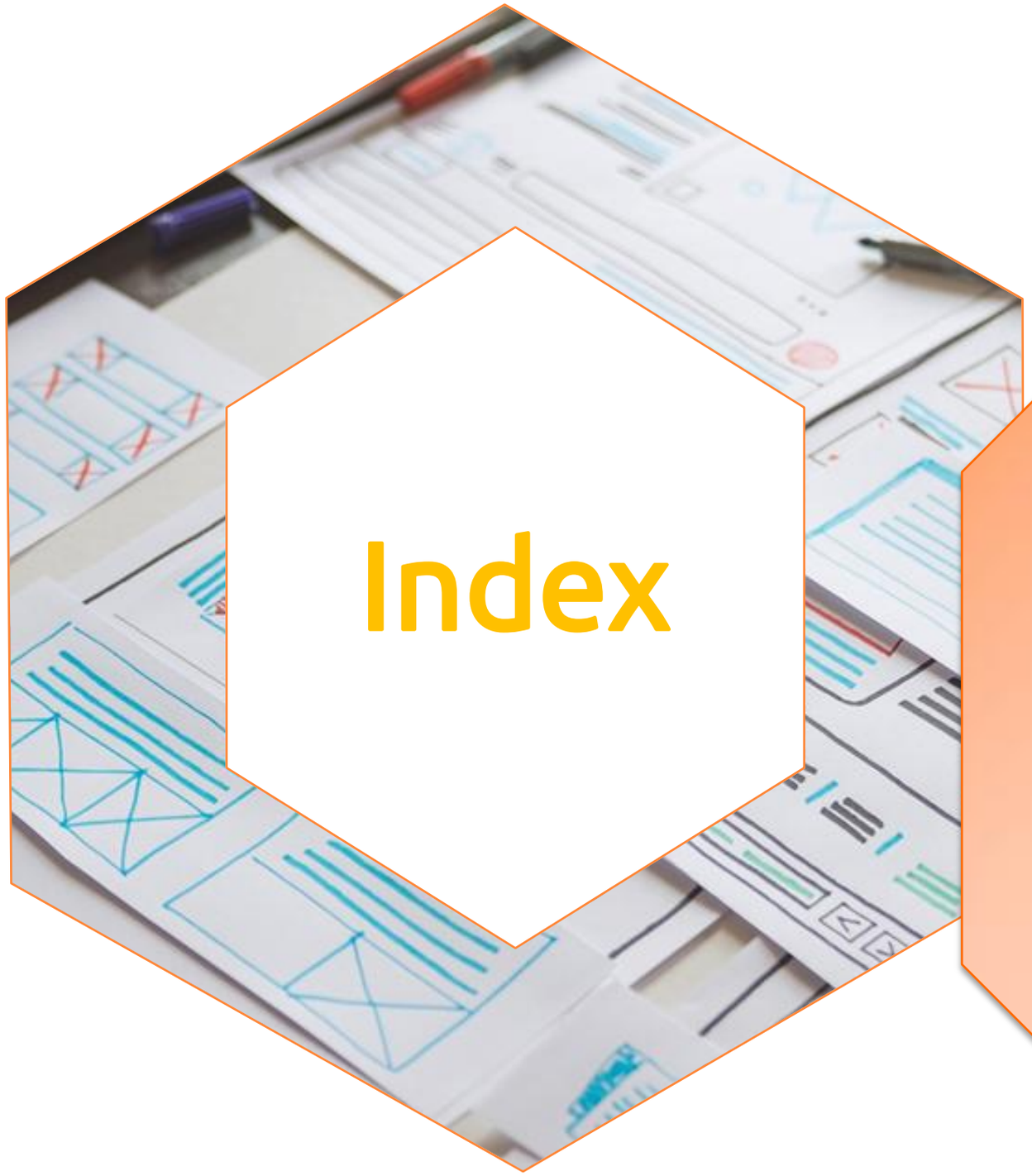

CREDIT
EDA
ANALYSIS
BY GAYATRI BHINGE



Index

1. Introduction
2. Problem Statement
3. Outliers
4. Imbalance in Data
5. Univariate Analysis
6. Bivariate Analysis
7. Conclusion

Introduction

This case study-assignment is based on Exploratory Data Analysis. The aim of this assignment is to apply EDA on real time business case. In this case study we will learn how to develop understanding of risk analytics in banking and financial services and as this dataset is applications of loan by client, so by analyzing this data we understand how data is used to minimize the risk of losing money while approving the loan to applicants.

Problem Statement

To take decision of loan application is very complicated task for any finance company, and it is big risk to approve loan without analysing applicant's profile. It is hard to find insufficient or non-existent credit history of any applicant and due to this applicant may be take advantage of becoming defaulter. As a data analyst for finance company we have to apply EDA on given data to analyses the patterns present in that datasets. When the company receives a loan application, we must ensure that applicant capable to repaying loan amount without fault, and loan is approved after doing analysis of applicant's profile.

There are 2 types of risks depend on decision of loan approval are as below:

1] Rejecting the loan application to capable applicant leads to business loss.

2] Accepting application of defaulter applicant, and if not able to replay loan leads to financial loss. In given dataset there are 2 scenarios, 'Target': 1 and 0.

As task performing EDA on this problem statement we have to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this outcome before approving loan to any applicant.

Data Understanding

This dataset has 3 files as explained below:

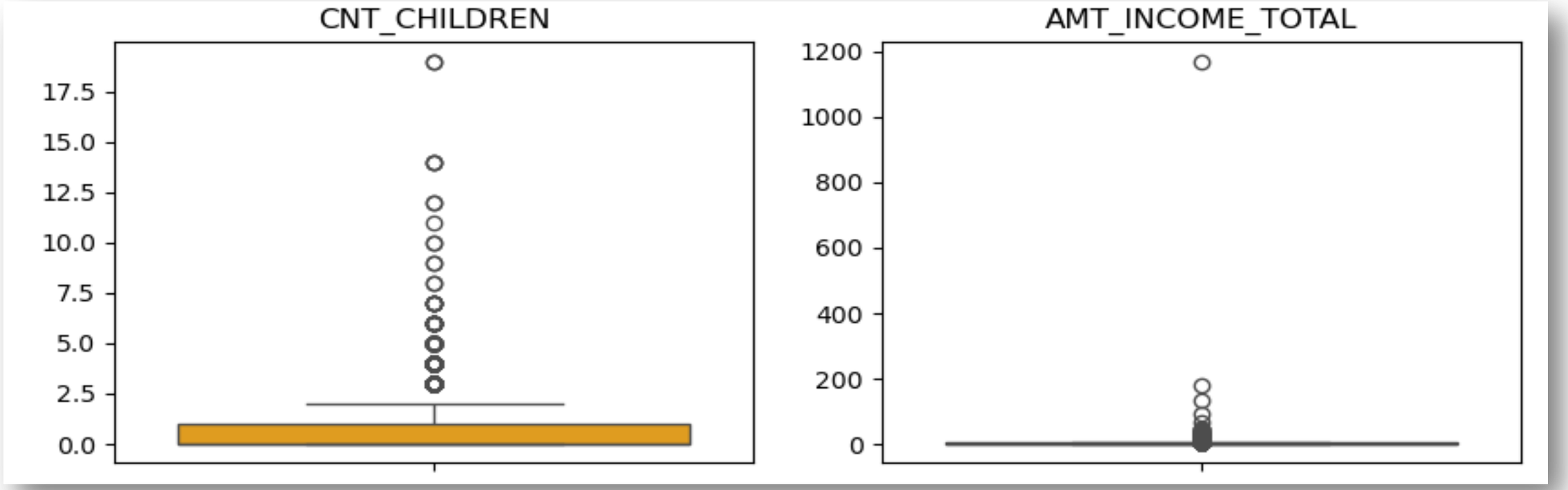
❖'application_data.csv' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.

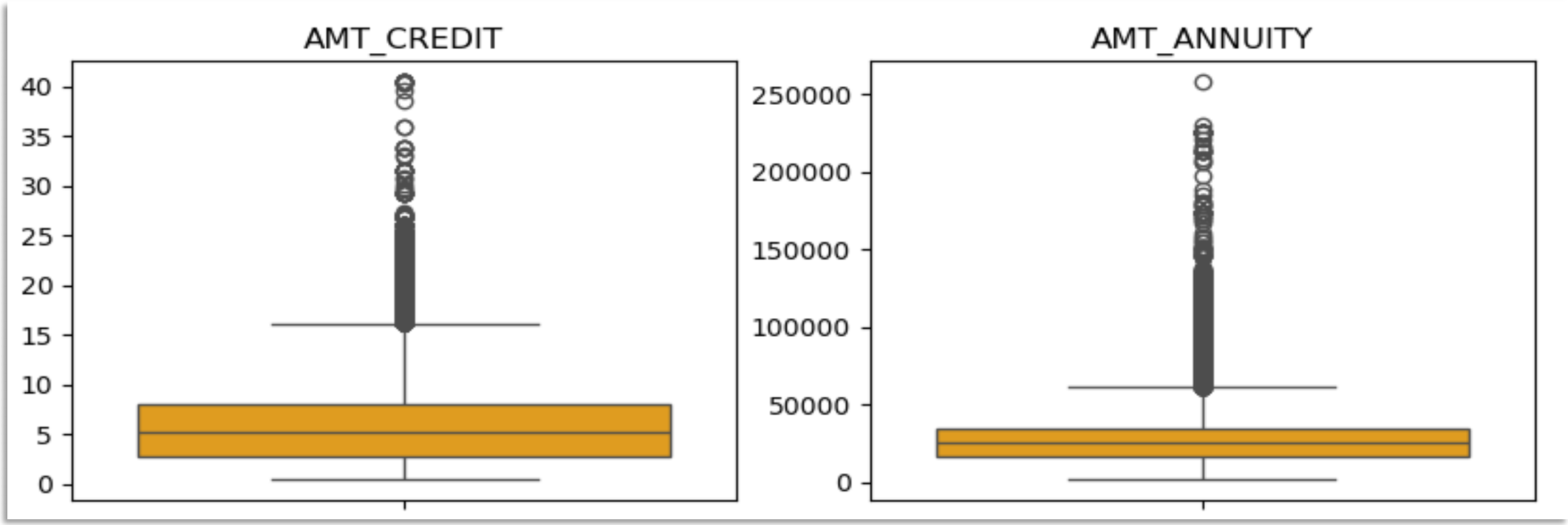
❖'previous_application.csv' contains information about the client's previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.

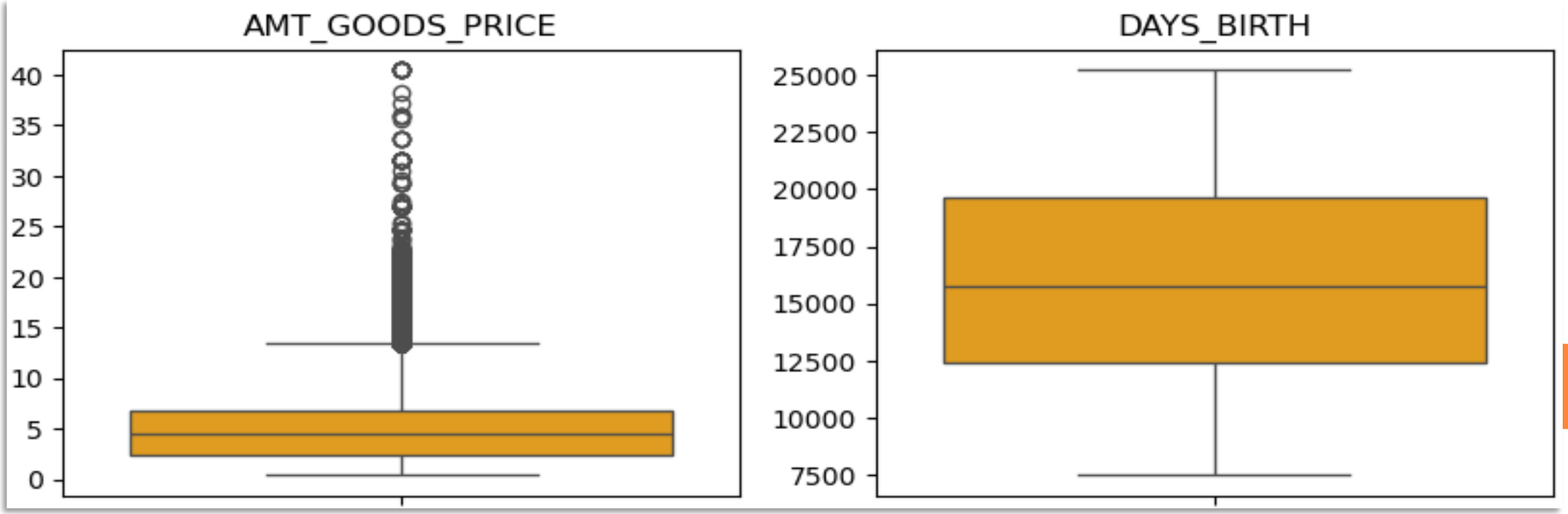
❖'columns_description.csv' is data dictionary which describes the meaning of the variables.

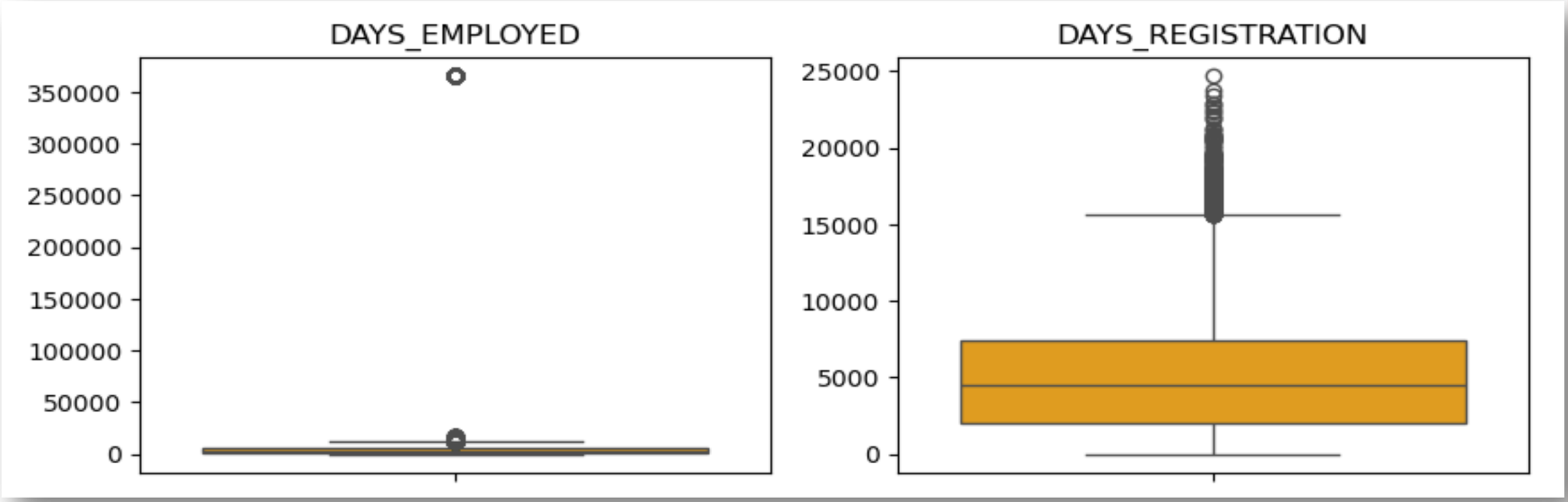
Outliers

Application Data -




AMT_CREDIT
40
35
30
25
20
15
10
5
0
AMT_ANNUITY
250000
200000
150000
100000
50000
0


AMT_GOODS_PRICE
40
35
30
25
20
15
10
5
0
DAYS_BIRTH
25000
22500
20000
17500
15000
12500
10000
7500

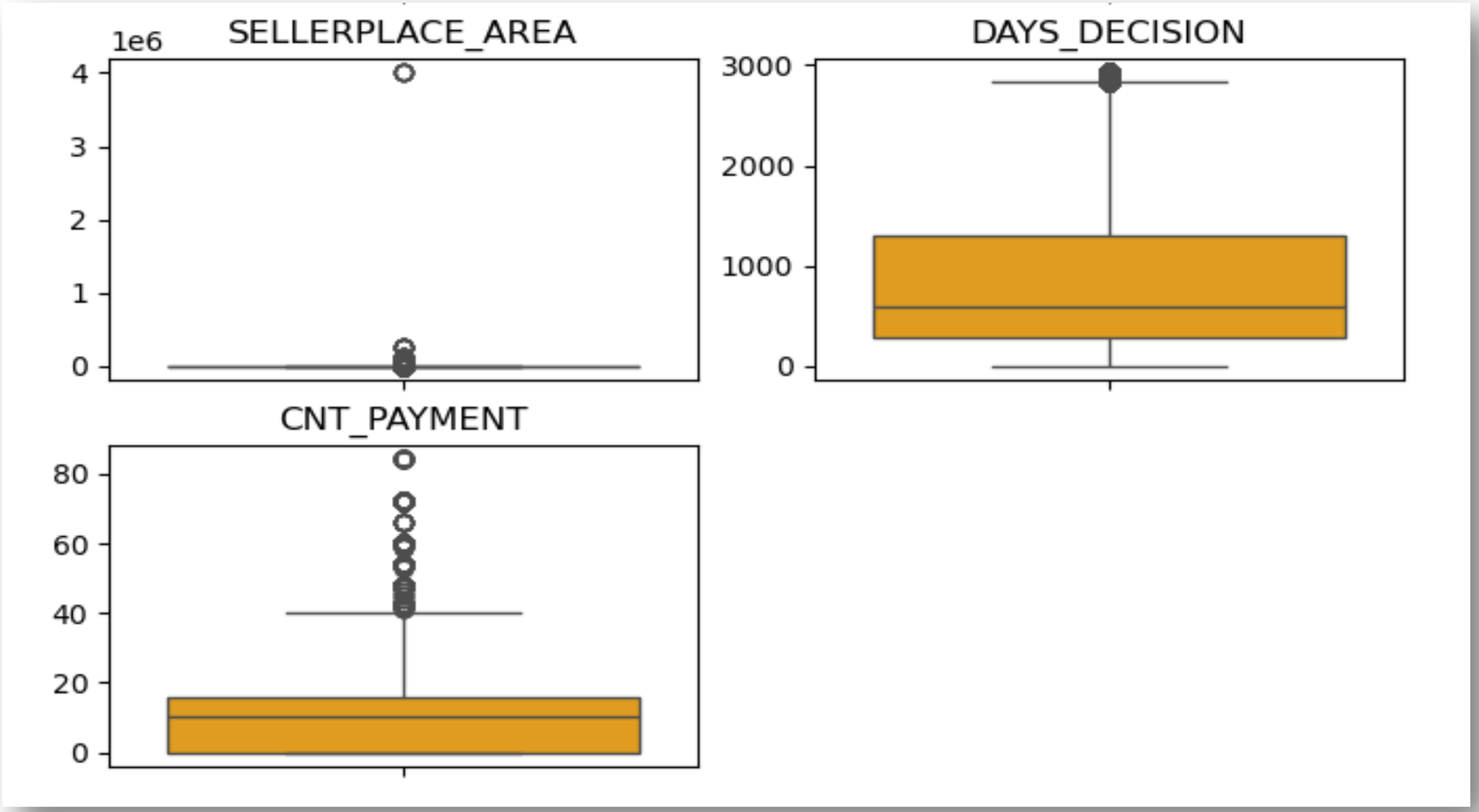

DAYS_EMPLOYED
350000
300000
250000
200000
150000
100000
50000
0
DAYS_REGISTRATION
25000
20000
15000
10000
5000
0

Inferences

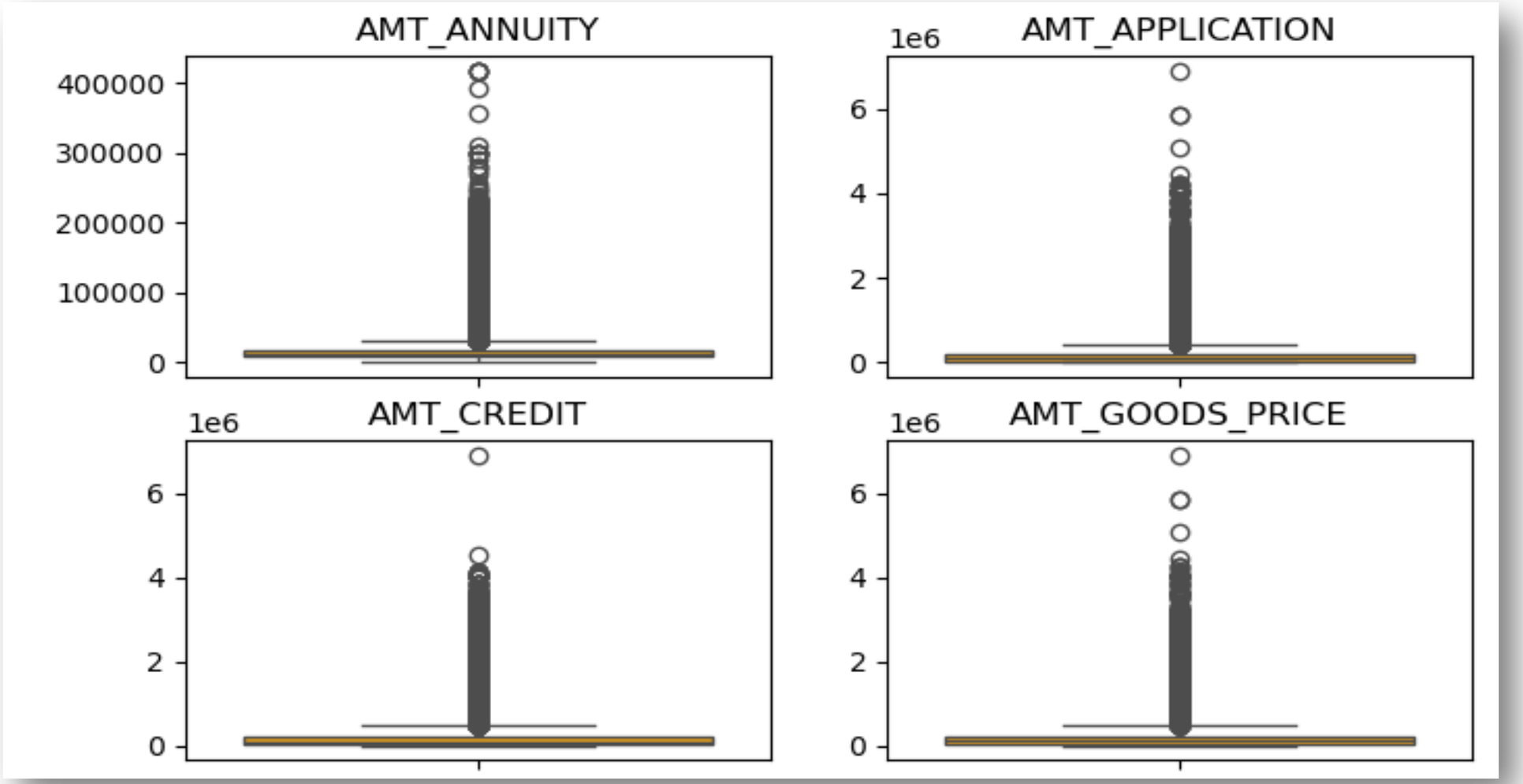
Above boxplots analysis –

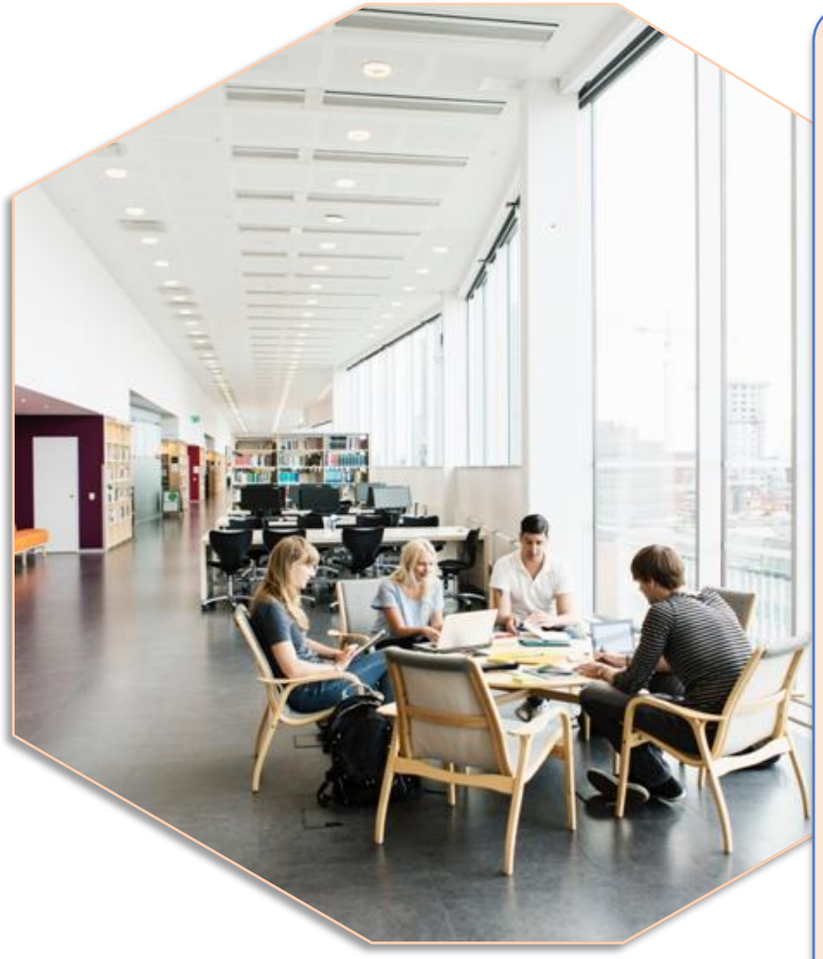
- AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.
- Few of the loan applicants have high income when compared to the others as data in AMT_INCOME_TOTAL has huge number of outliers.
- DAYS_BIRTH has no outliers which means the data available is reliable.
- DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and hence this has to be incorrect entry.
- Outcome for CNT_CHILDREN looks like not correct as number of children's showing outliers.

Previous Applications Data



Previous Applications Data





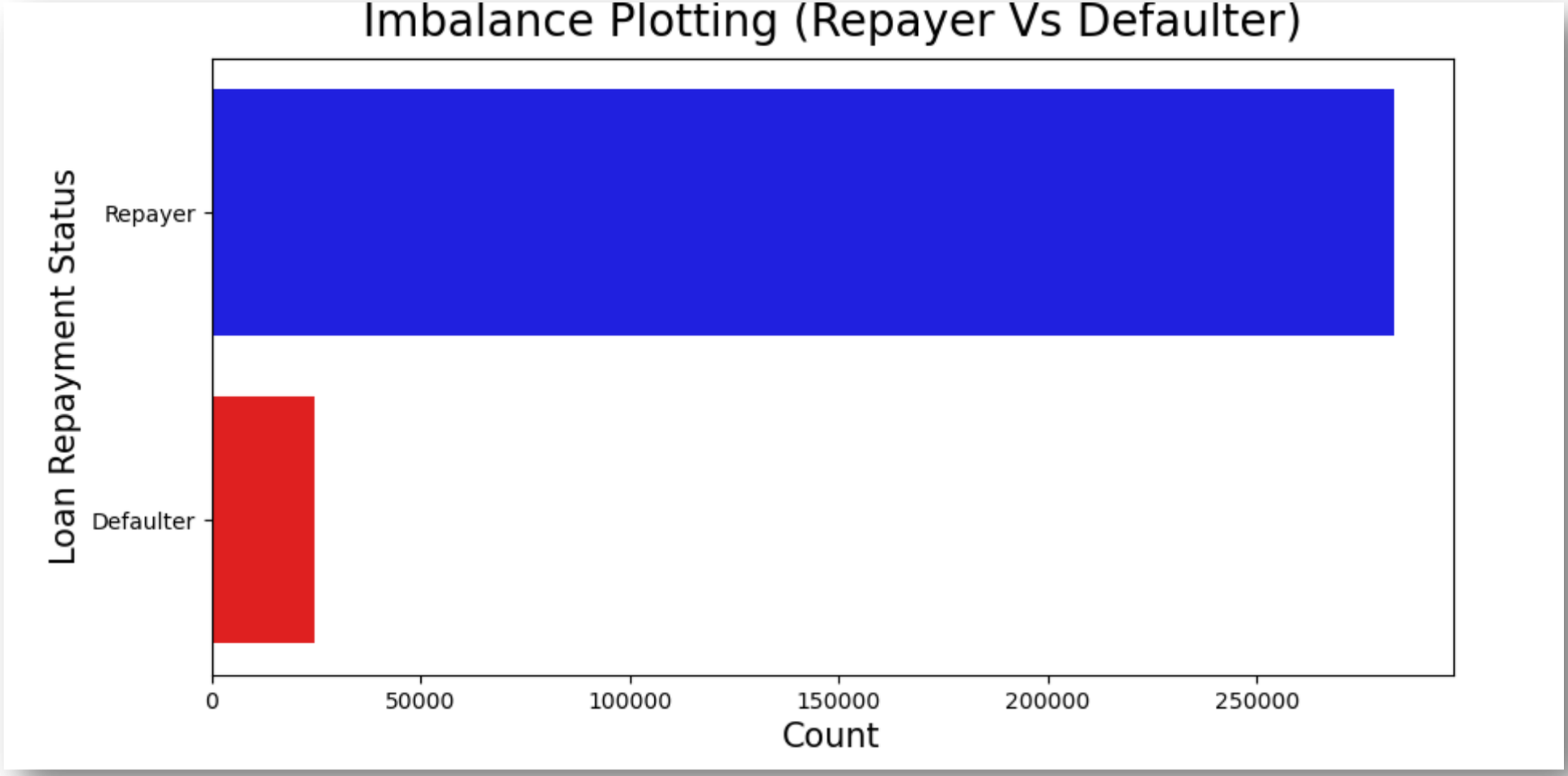
Inferences

Above boxplots analysis –

- DAYS_DECISION has little number of outliers indicating that these previous applications decisions were taken long back.
- CNT_PAYMENT has few outlier values.
- AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.

Imbalance in Data

Application Data



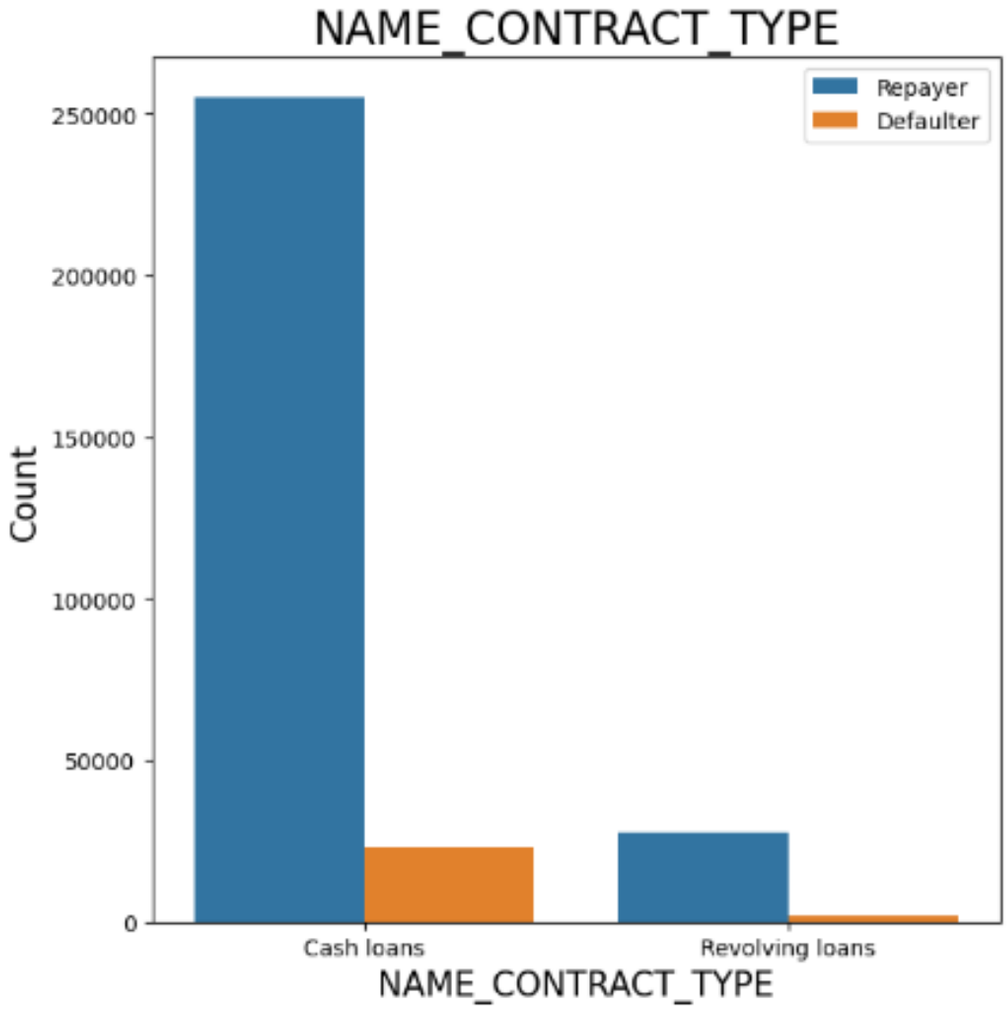
Inferences

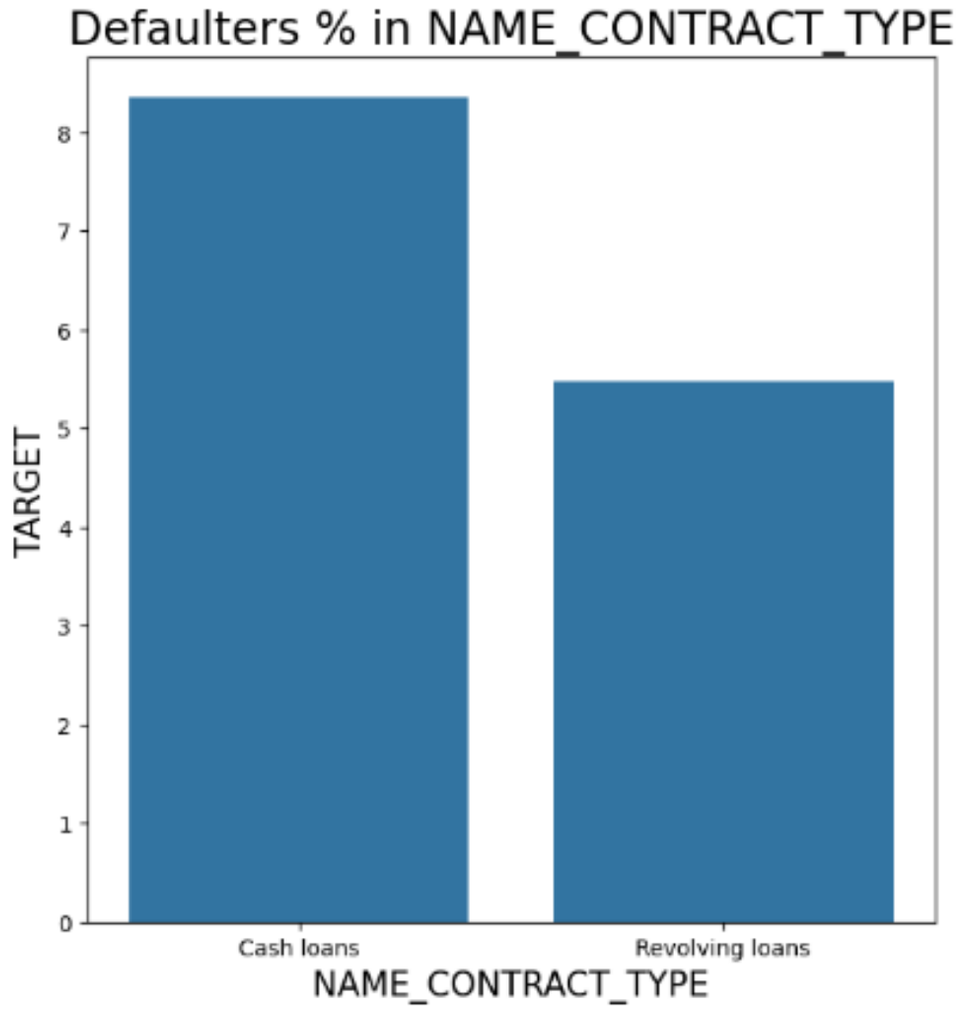
As we see in this graph, there is significant difference between 'Repayer' and 'Defaulter'. We will check ratio as below –

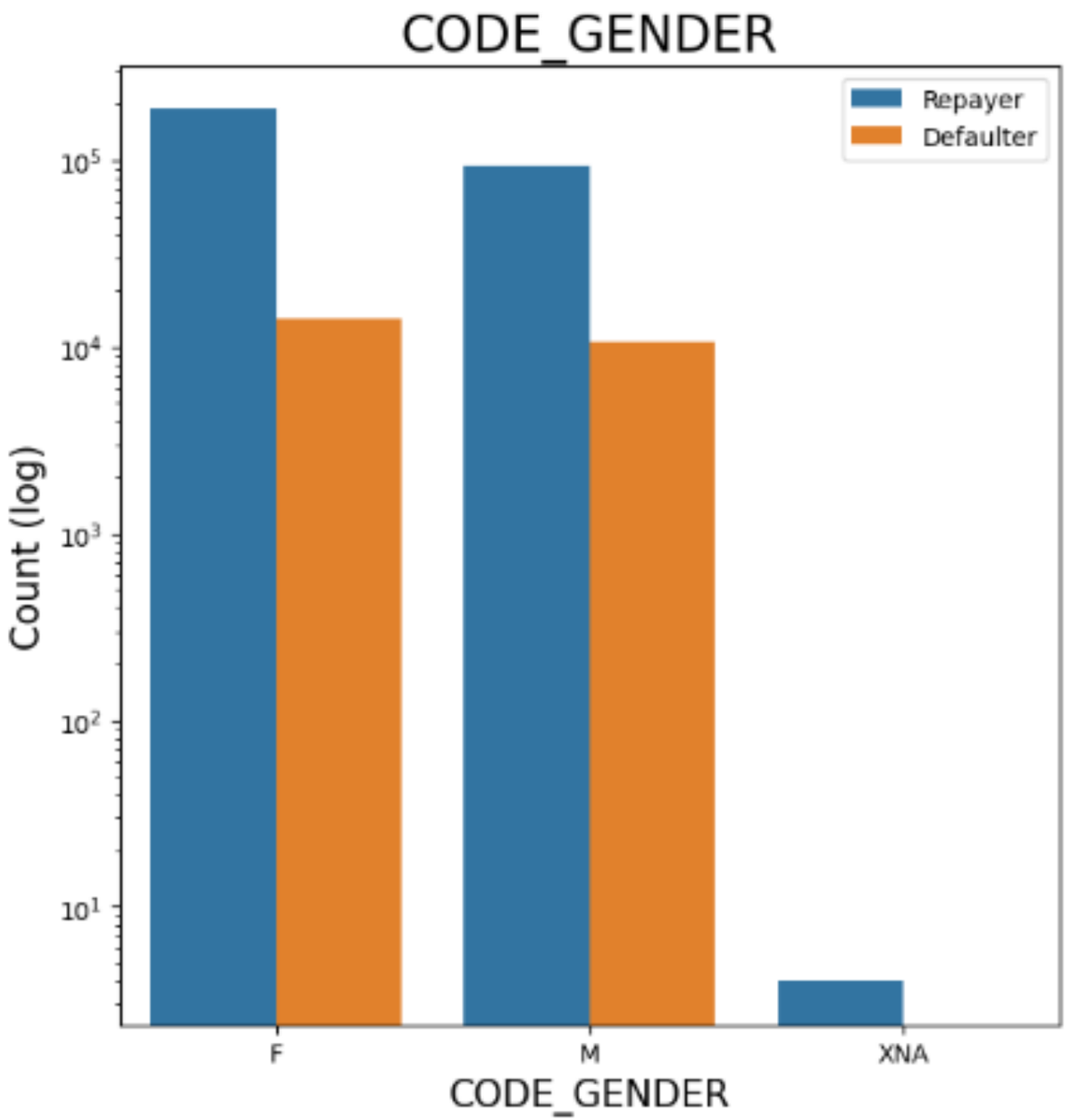
- Repayer Percentage is 91.93%
- Defaulter Percentage is 8.07%
- Imbalance Ratio with respect to Repayer and Defaulter is given: 11.39/1 (approx)

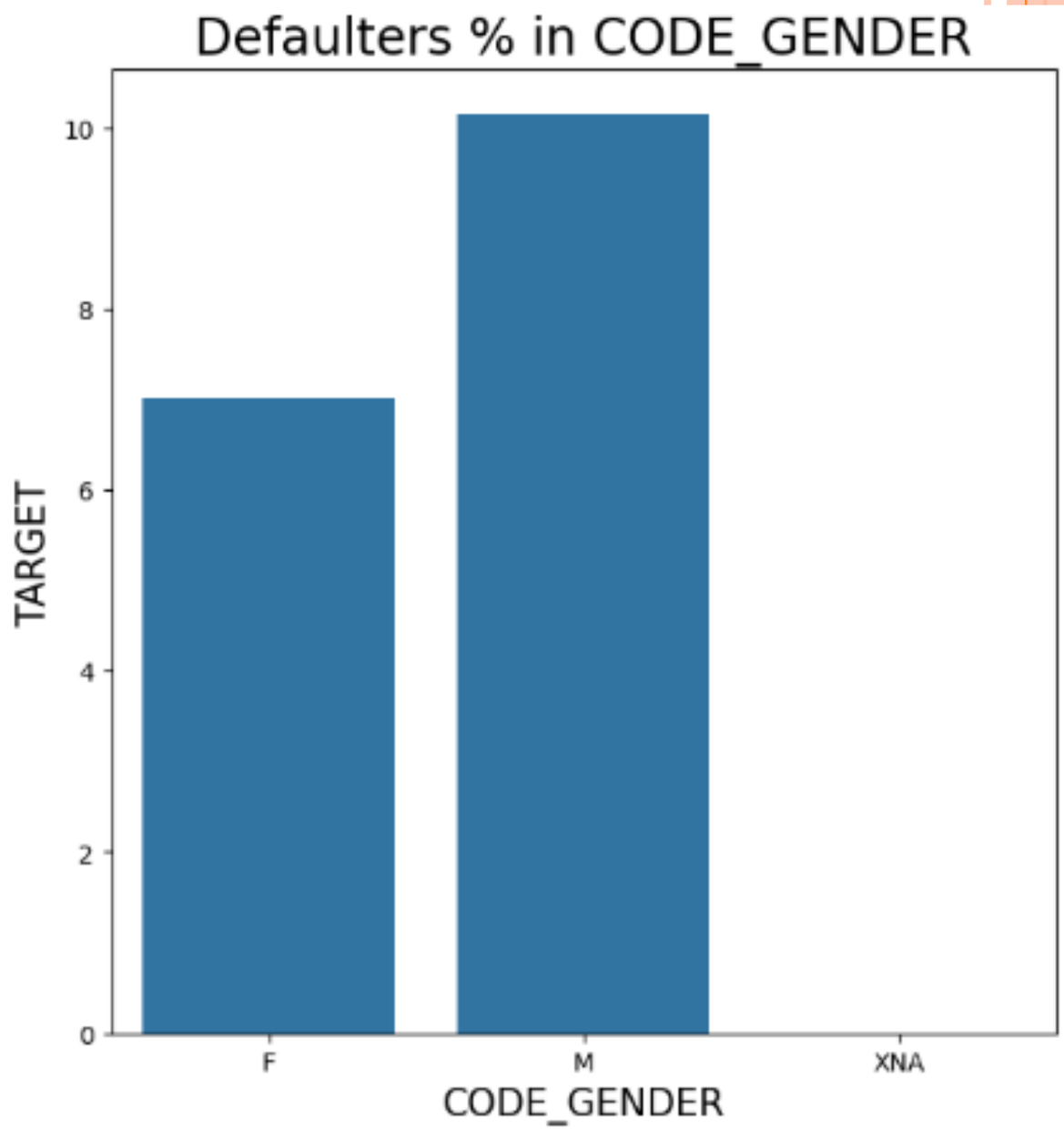
Univariate Analysis

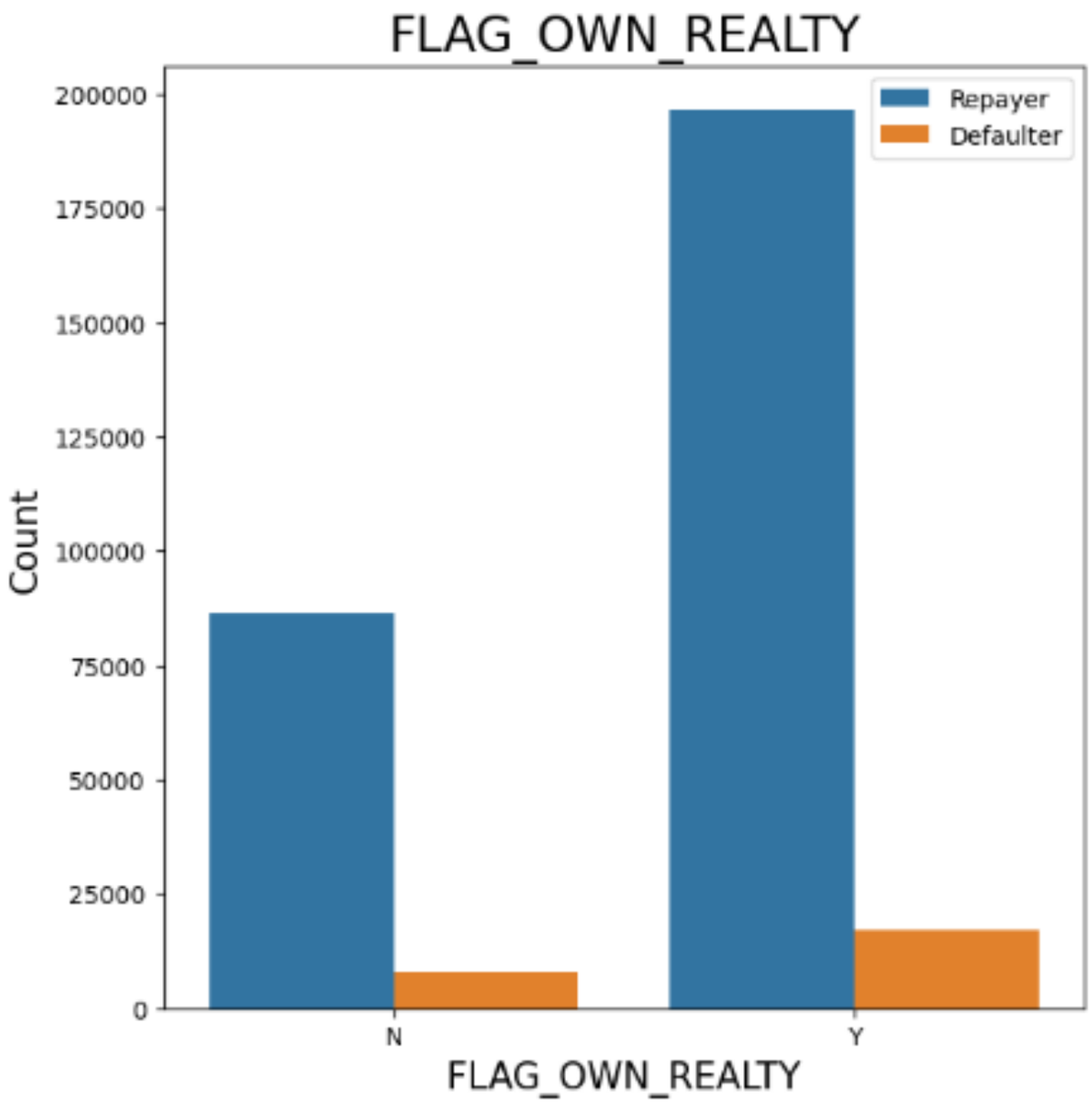
Application Data

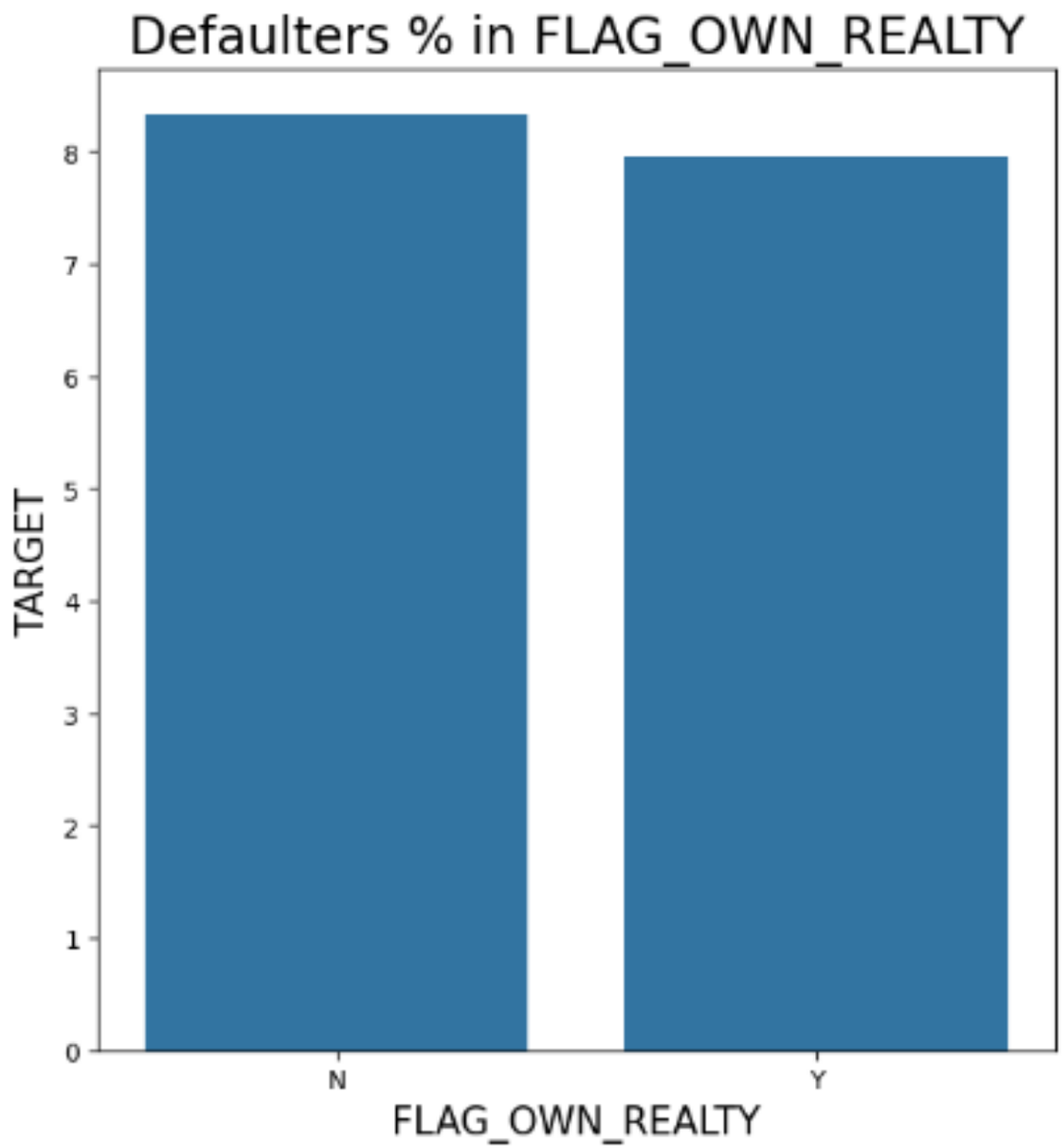


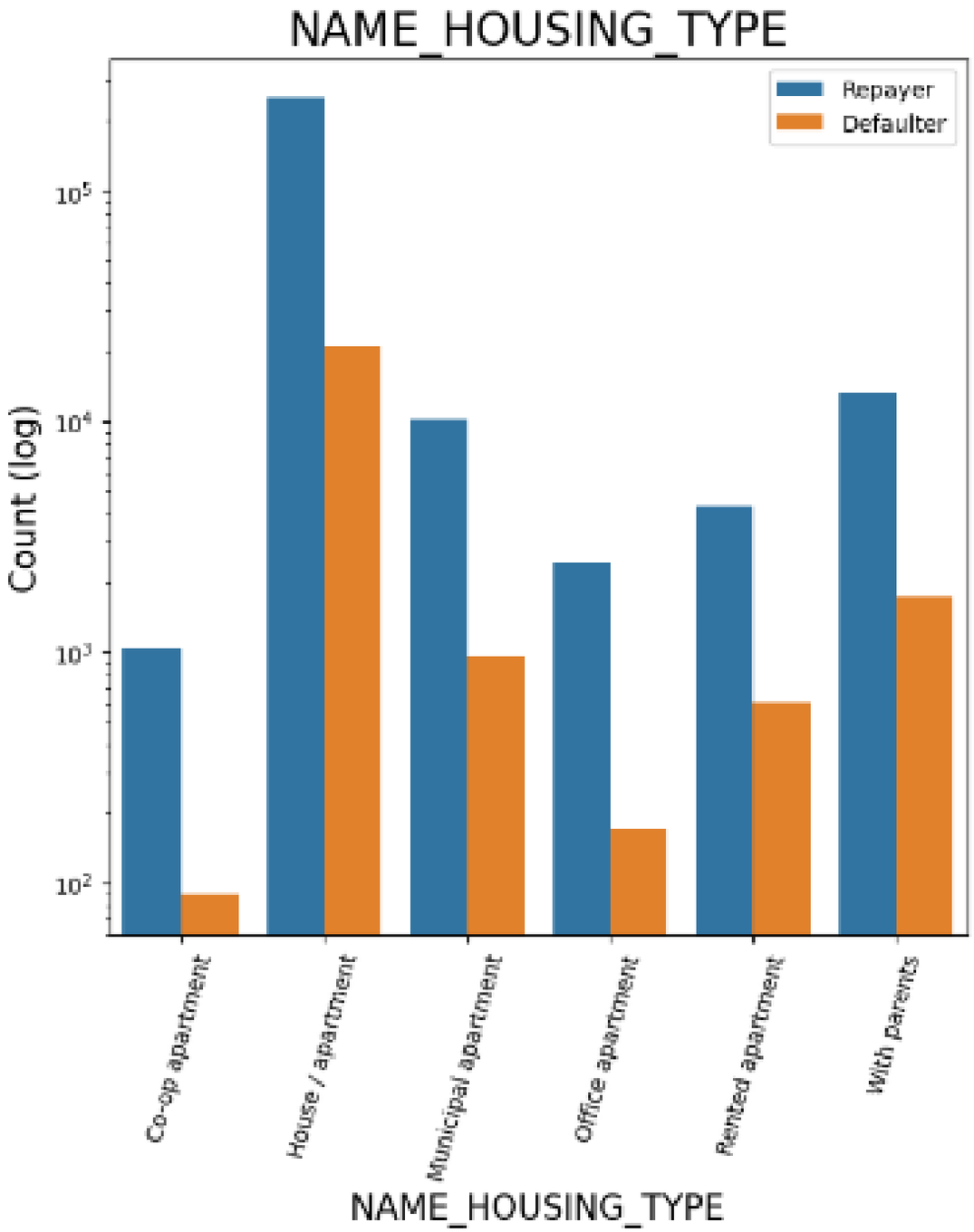


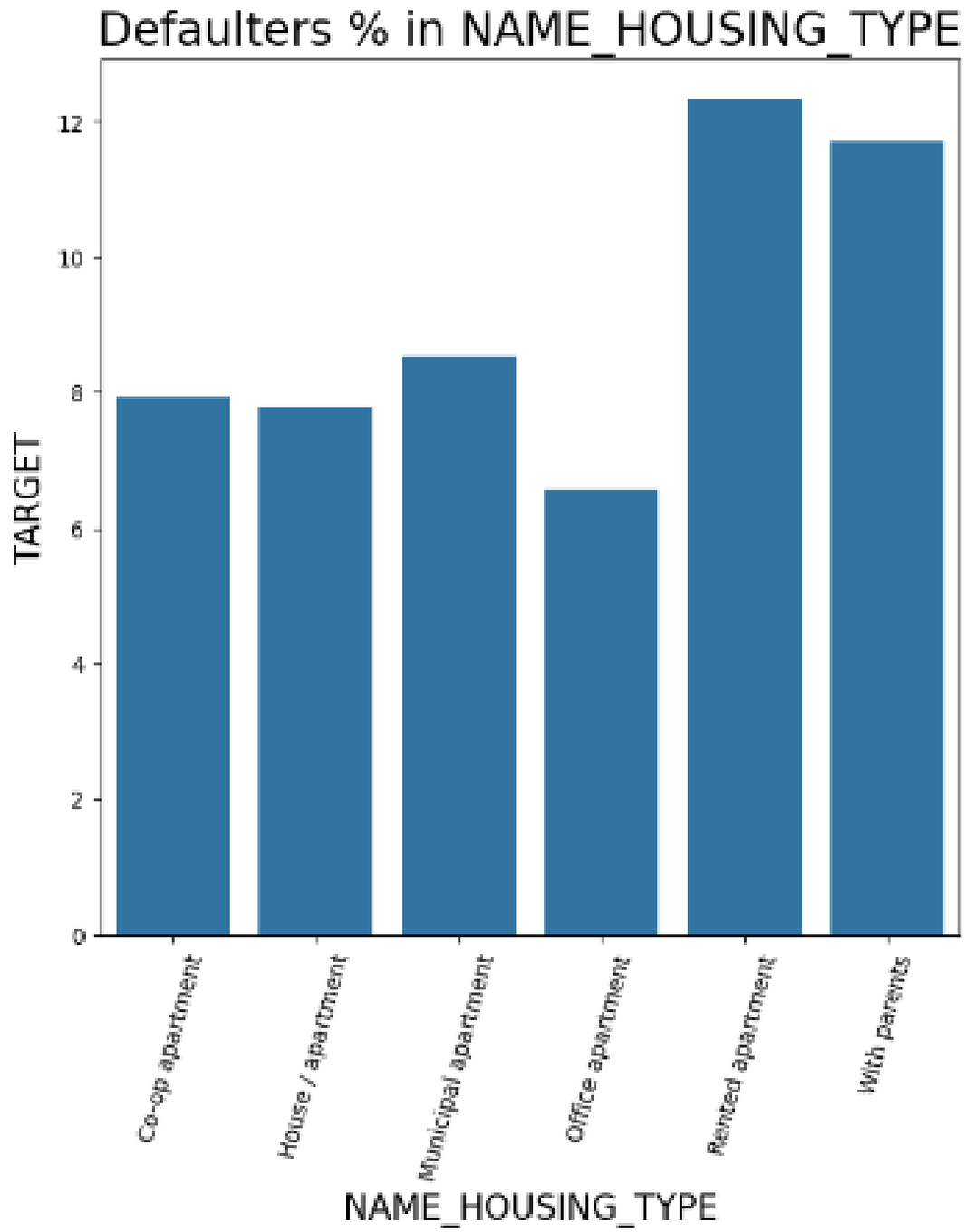

CODE_GENDER
Repayer
Defaulter
Count (log)
10^5
10^4
10^3
10^2
10^1
F
M
XNA
CODE_GENDER


Defaulters % in CODE_GENDER
TARGET
10
8
6
4
2
0
F
M
XNA
CODE_GENDER


FLAG_OWN_REALTY
Repayer
Defaulter
200000
175000
150000
125000
100000
75000
50000
25000
0
Count
N
Y
FLAG_OWN_REALTY

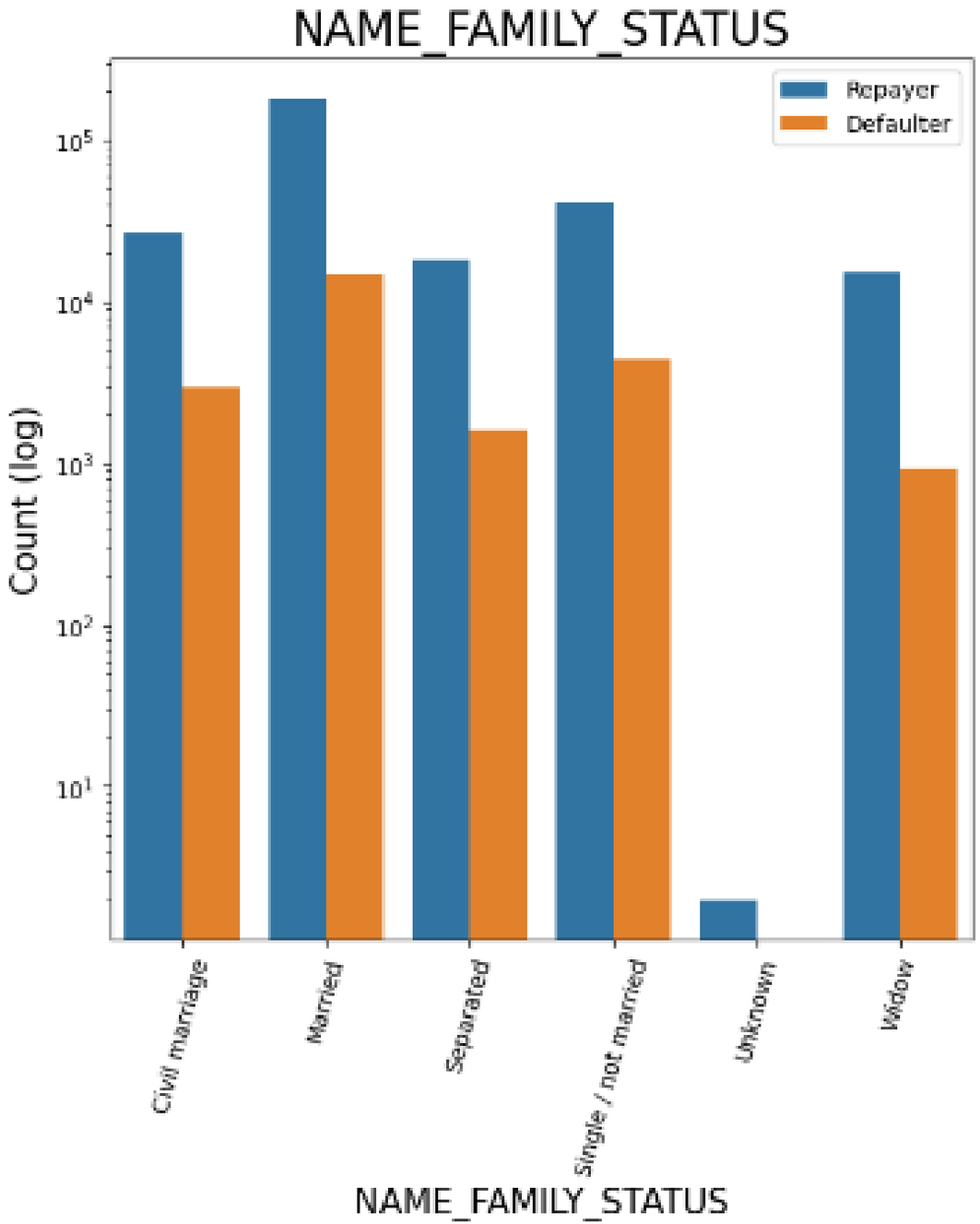

Defaulters % in FLAG_OWN_REALTY
8
7
6
5
4
3
2
1
0
TARGET
N
Y
FLAG_OWN_REALTY

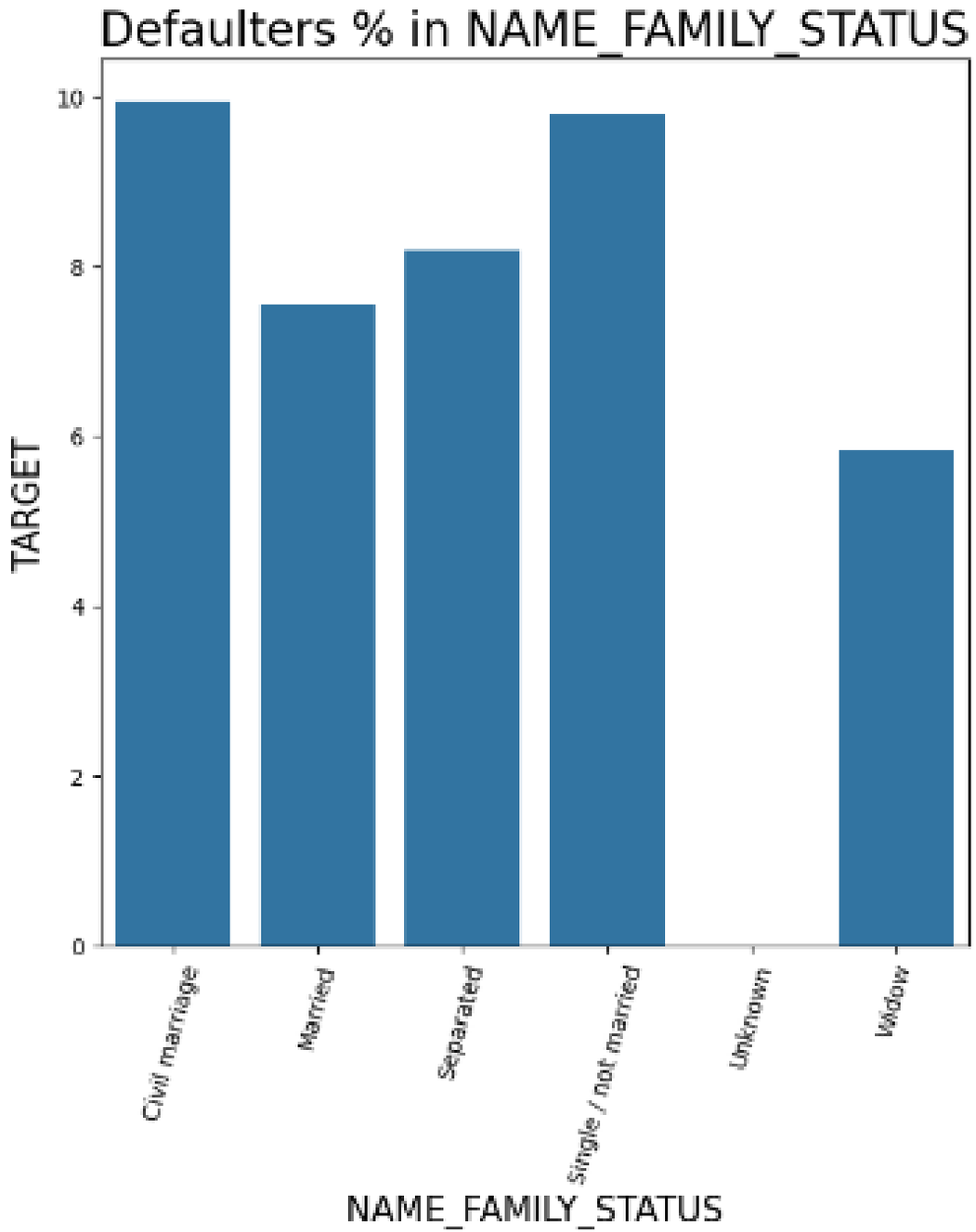

NAME_HOUSING_TYPE
Repayer
Defaulter
Count (log)
10^5
10^4
10^3
10^2
Co-op apartment
House / apartment
Municipal apartment
Office apartment
Rented apartment
With parents
NAME_HOUSING_TYPE

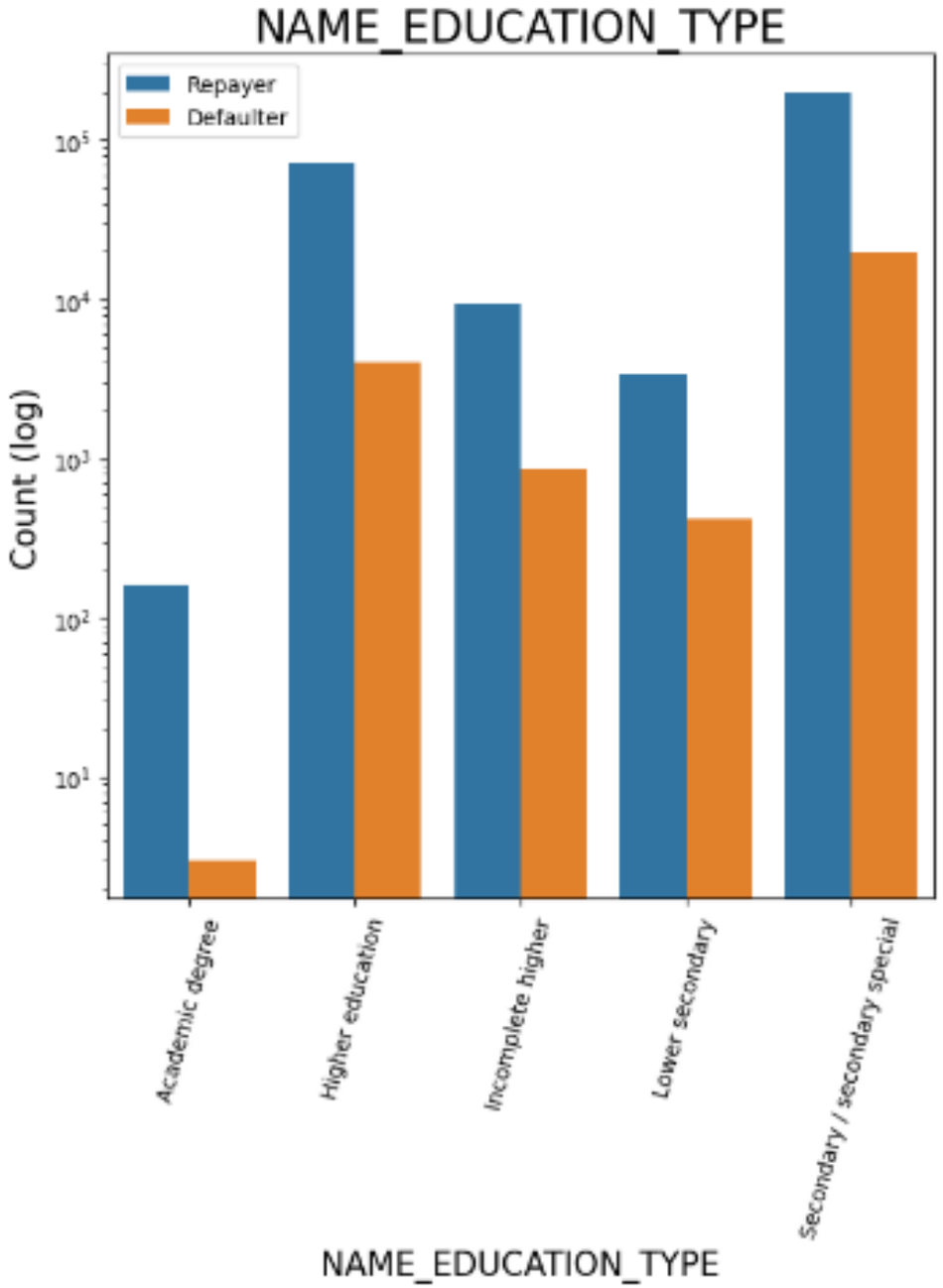

Defaulters % in NAME_HOUSING_TYPE
TARGET
12
10
8
6
4
2
0
Co-op apartment
House / apartment
Municipal apartment
Office apartment
Rented apartment
With parents
NAME_HOUSING_TYPE

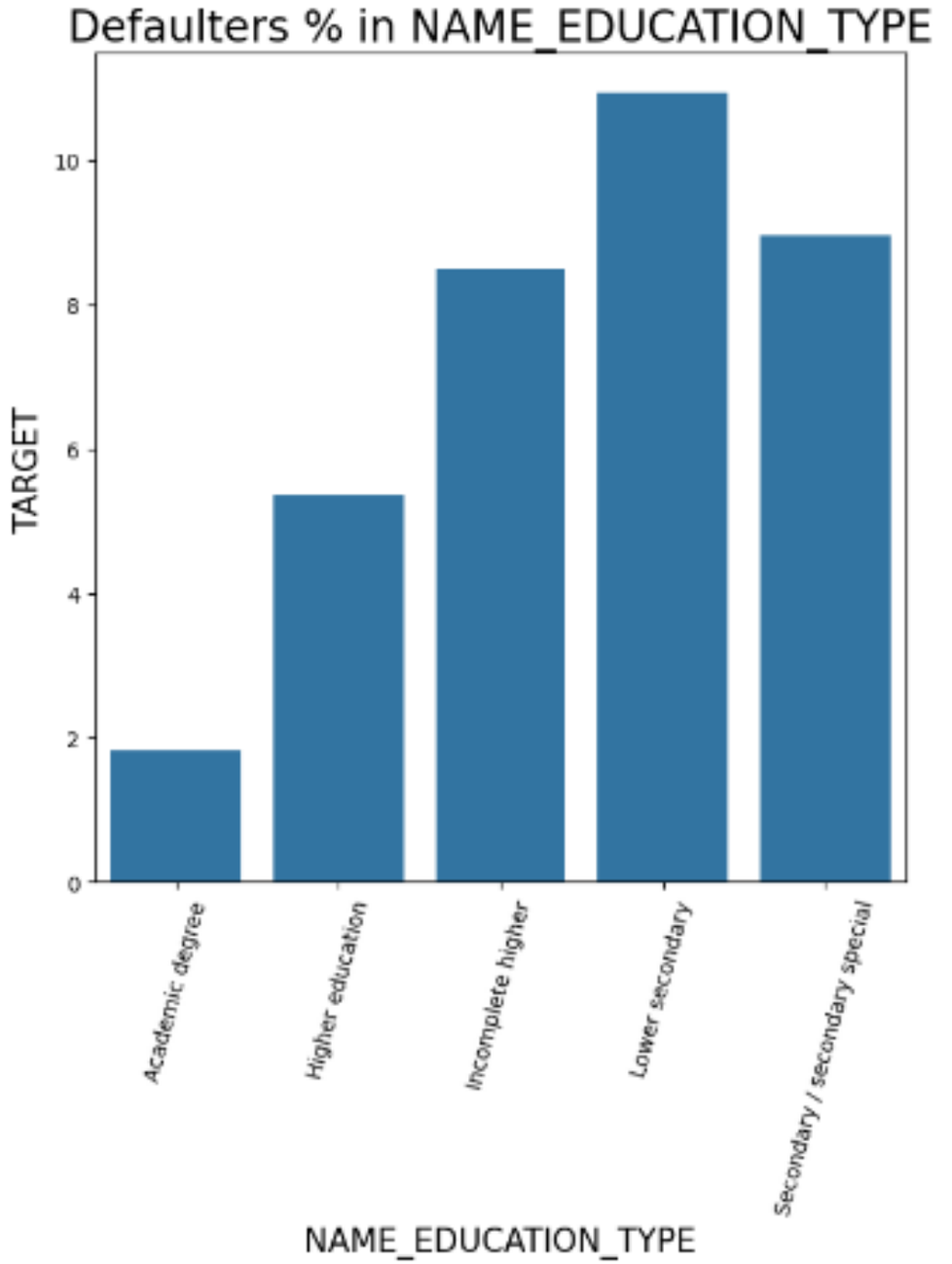
Inferences

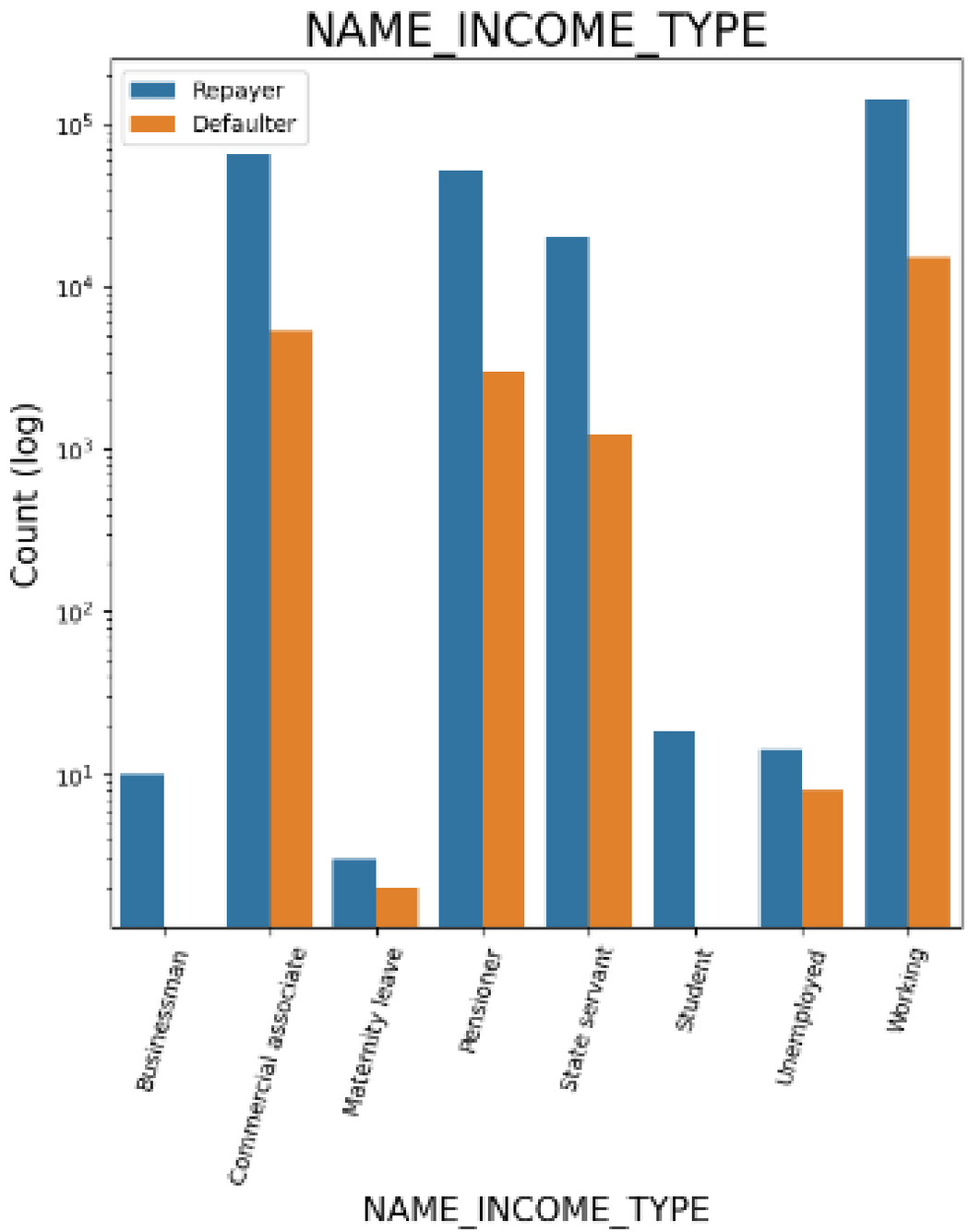
- Revolving loans, means loan for buying any goods, are just a small fraction (10%) from the total number of loans.
- Defaulters are in 8-9% Cash loan applicants and 5-6% Revolving loan applicant.
- The number of female loan applicants is almost double the number of male applicants.
- Based on the percentage of defaulted credits, chance of males' defaulters 10% is higher than female 7%.
- There are property owner are more applicants than don't own property.
- The defaulting rate is same for both, own and don't own property (8%), which means correlation between owning a property and defaulting the loan.
- Majority of applicants are live in House/apartment and applicants living in office apartments have lowest default rate
- The applicants who are living with parents (11.5%) and living in rented apartments(>12%) have higher probability of defaulting

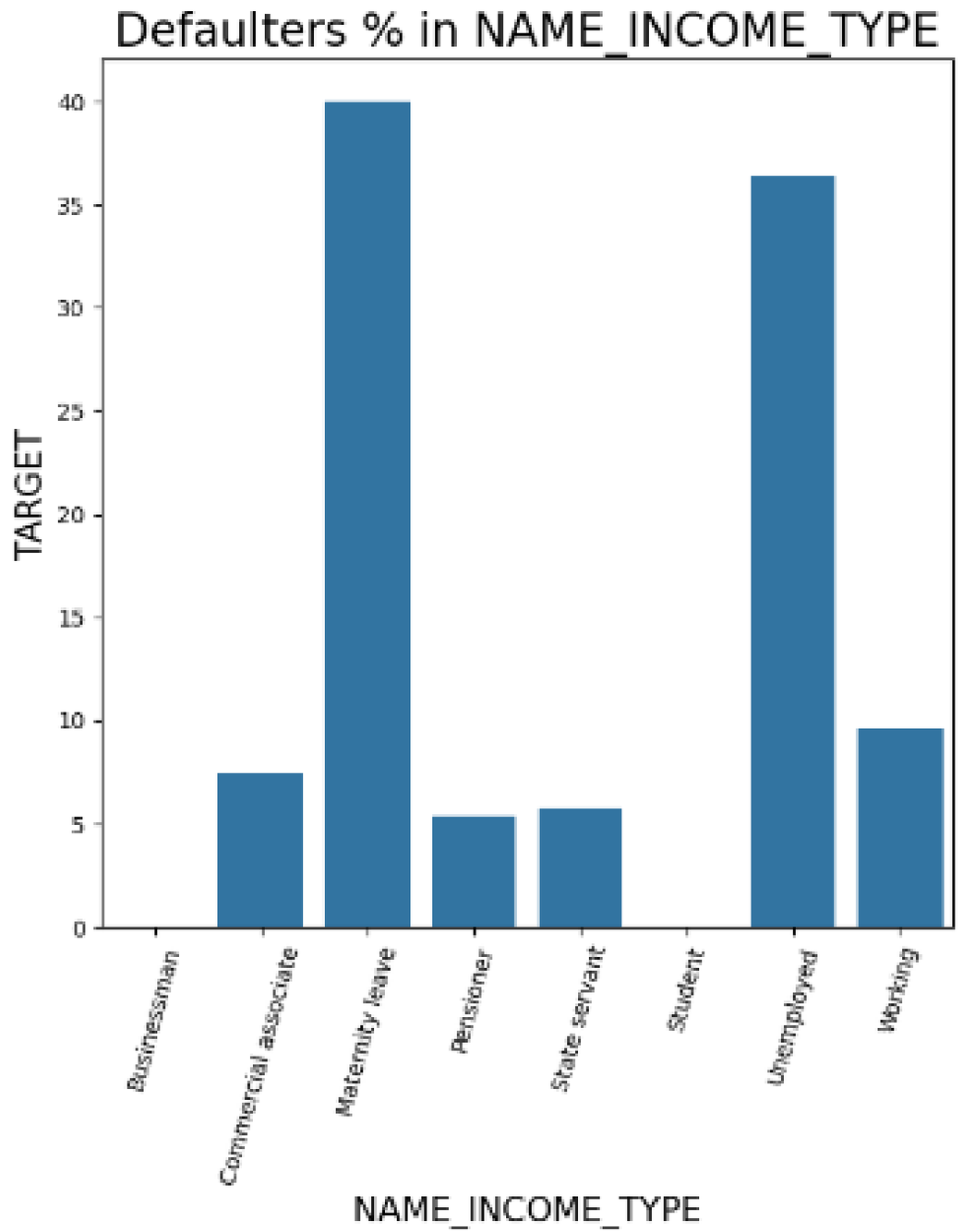

NAME_FAMILY_STATUS
Repayer
Defaulter
Count (log)
10^5
10^4
10^3
10^2
10^1
Civil marriage
Married
Separated
Single / not married
Unknown
Widow
NAME_FAMILY_STATUS

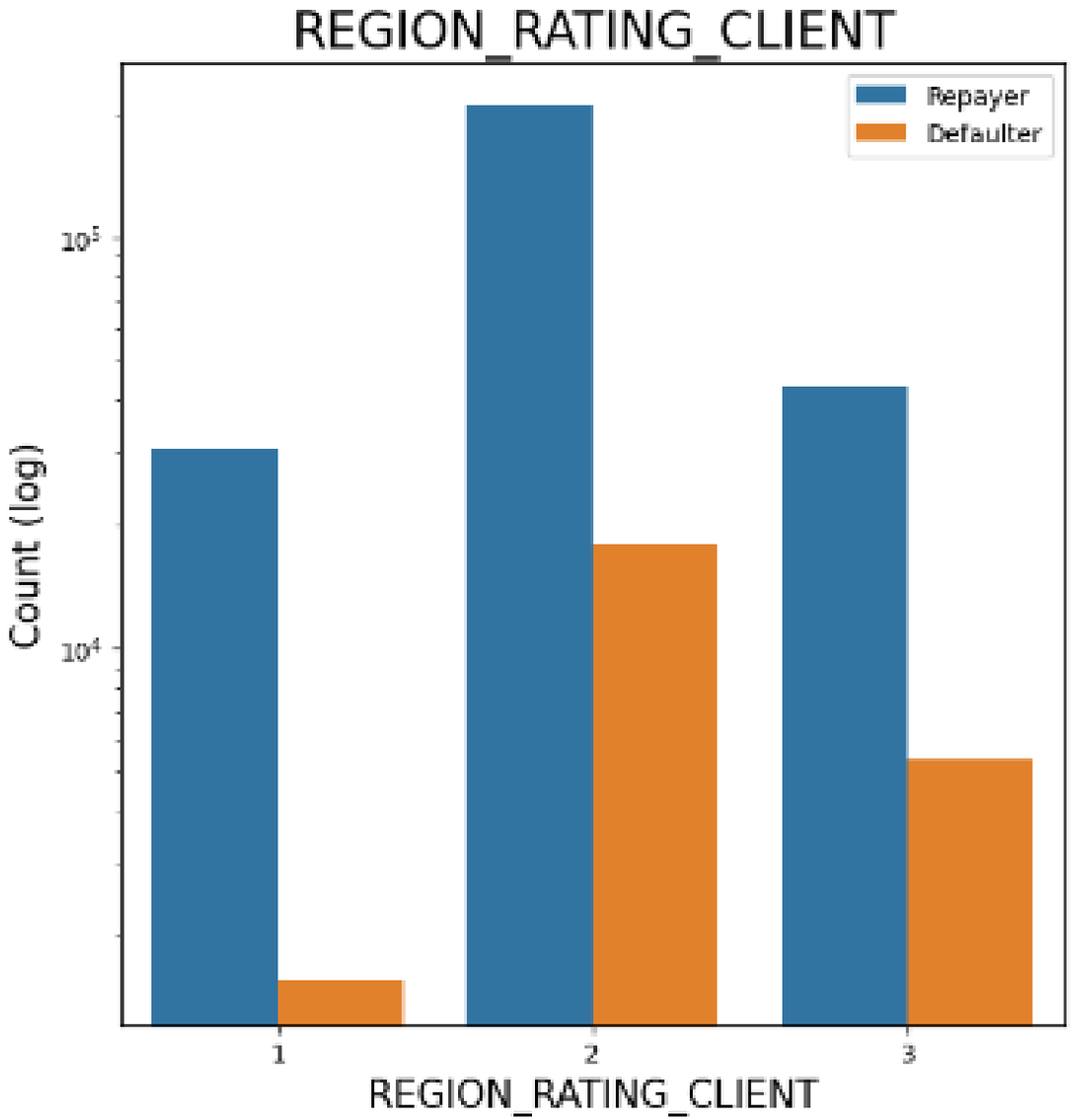

Defaulters % in NAME_FAMILY_STATUS
TARGET
10
8
6
4
2
0
Civil marriage
Married
Separated
Single / not married
Unknown
Widow
NAME_FAMILY_STATUS

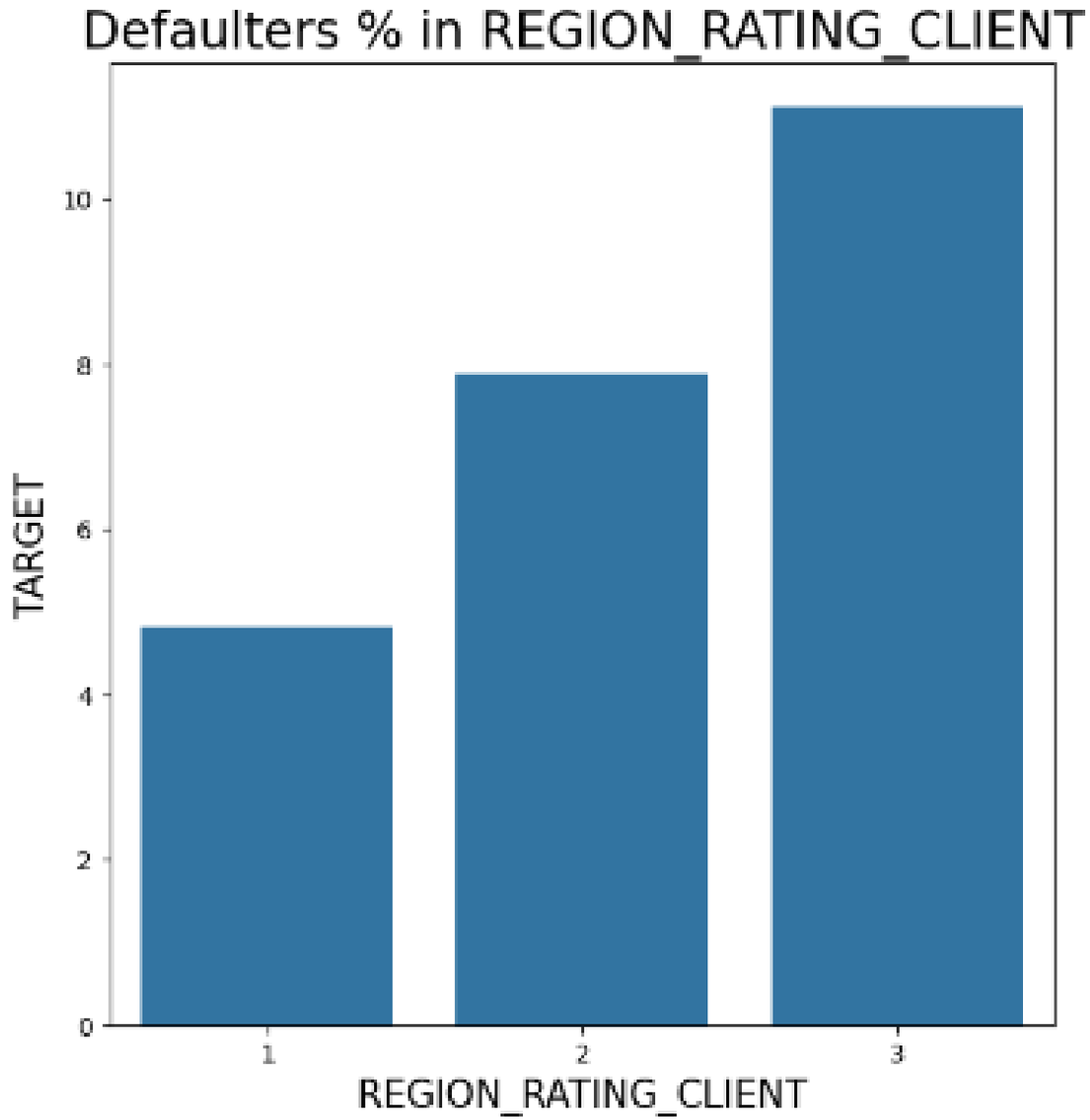

NAME_EDUCATION_TYPE
Repayer
Defaulter
Count (log)
10^5
10^4
10^3
10^2
10^1
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
NAME_EDUCATION_TYPE


Defaulters % in NAME_EDUCATION_TYPE
TARGET
10
8
6
4
2
0
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
NAME_EDUCATION_TYPE


NAME_INCOME_TYPE
Repayer
Defaulter
Count (log)
10^5
10^4
10^3
10^2
10^1
Businessman
Commercial associate
Maternity leave
Pensioner
State servant
Student
Unemployed
Working
NAME_INCOME_TYPE

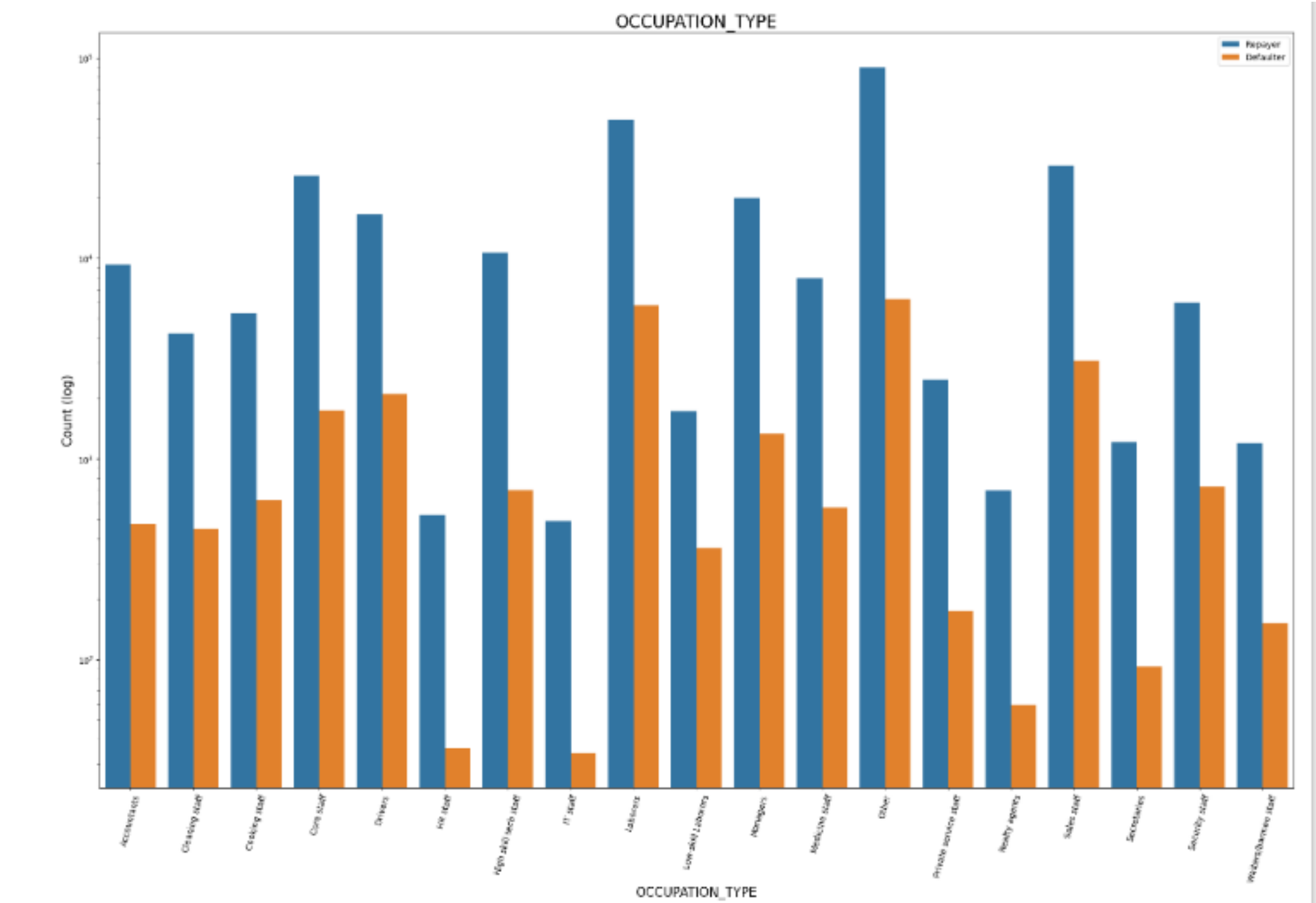

Defaulters % in NAME_INCOME_TYPE
TARGET
40
35
30
25
20
15
10
5
0
Businessman
Commercial associate
Maternity leave
Pensioner
State servant
Student
Unemployed
Working
NAME_INCOME_TYPE

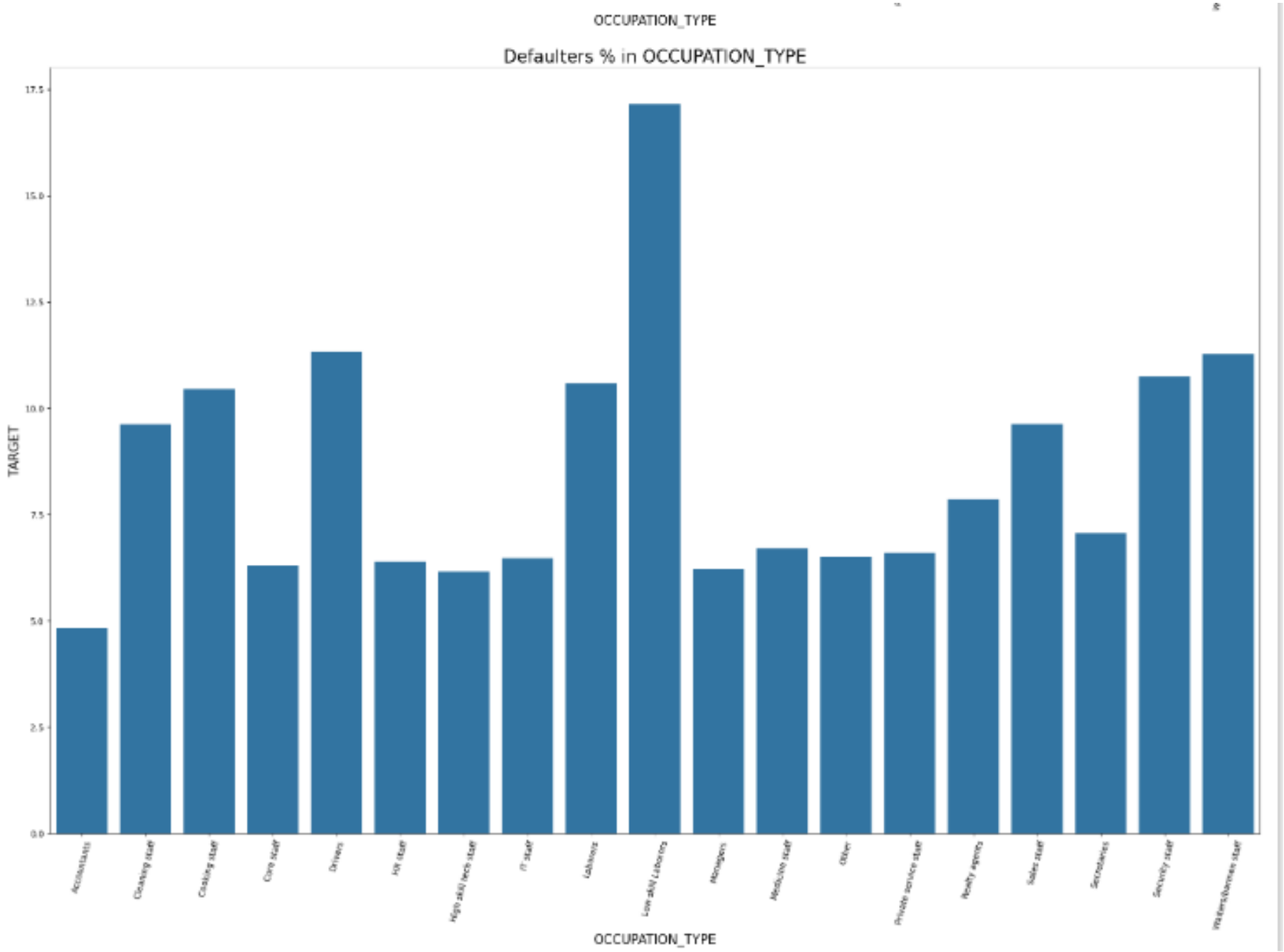

REGION_RATING_CLIENT
Repayer
Defaulter
Count (log)
10^5
10^4
1
2
3
REGION_RATING_CLIENT

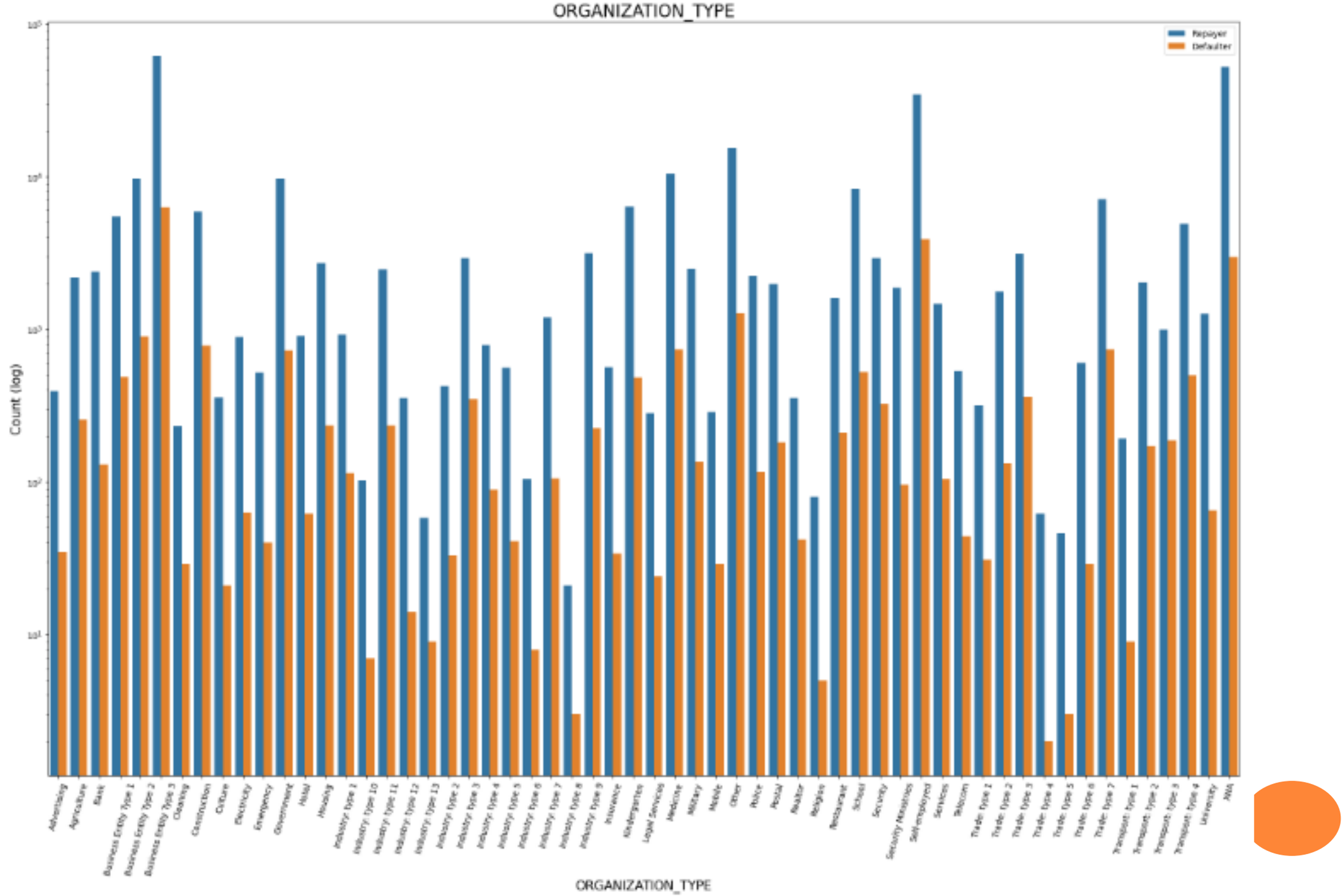

Defaulters % in REGION_RATING_CLIENT
TARGET
10
8
6
4
2
0
1
2
3
REGION_RATING_CLIENT

Inferences

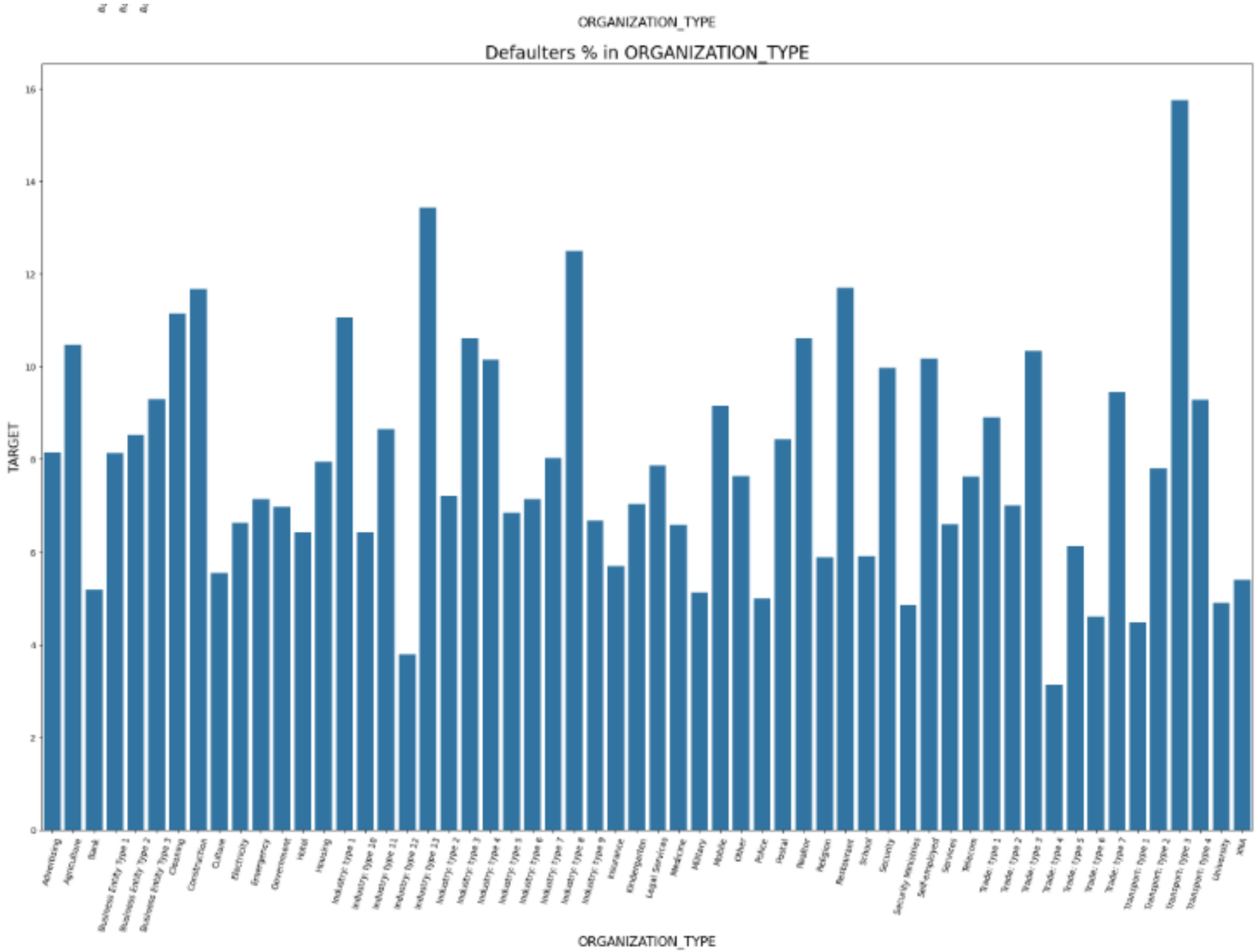
- Married have applied more then followed by Single/not married and civil marriage, applicants.
- In Percentage of defaulters, Civil marriage has the highest percent around (10%) and widow has the lowest around 6% (exception being Unknown).
- Majority of applicants have Secondary/secondary special education, followed by applicants with Higher education.
- Lower secondary category have highest rate of defaulting around 11%. People with Academic degree are least likely to default and very few applicants have an academic degree.
- Most of the applicants are living in Region with Rating 2 place. Region Rating 3 has the highest default rate (11%)
- Applicant living in Region_Rating 1 are safer for approving loanshas the lowest probability of defaulting.

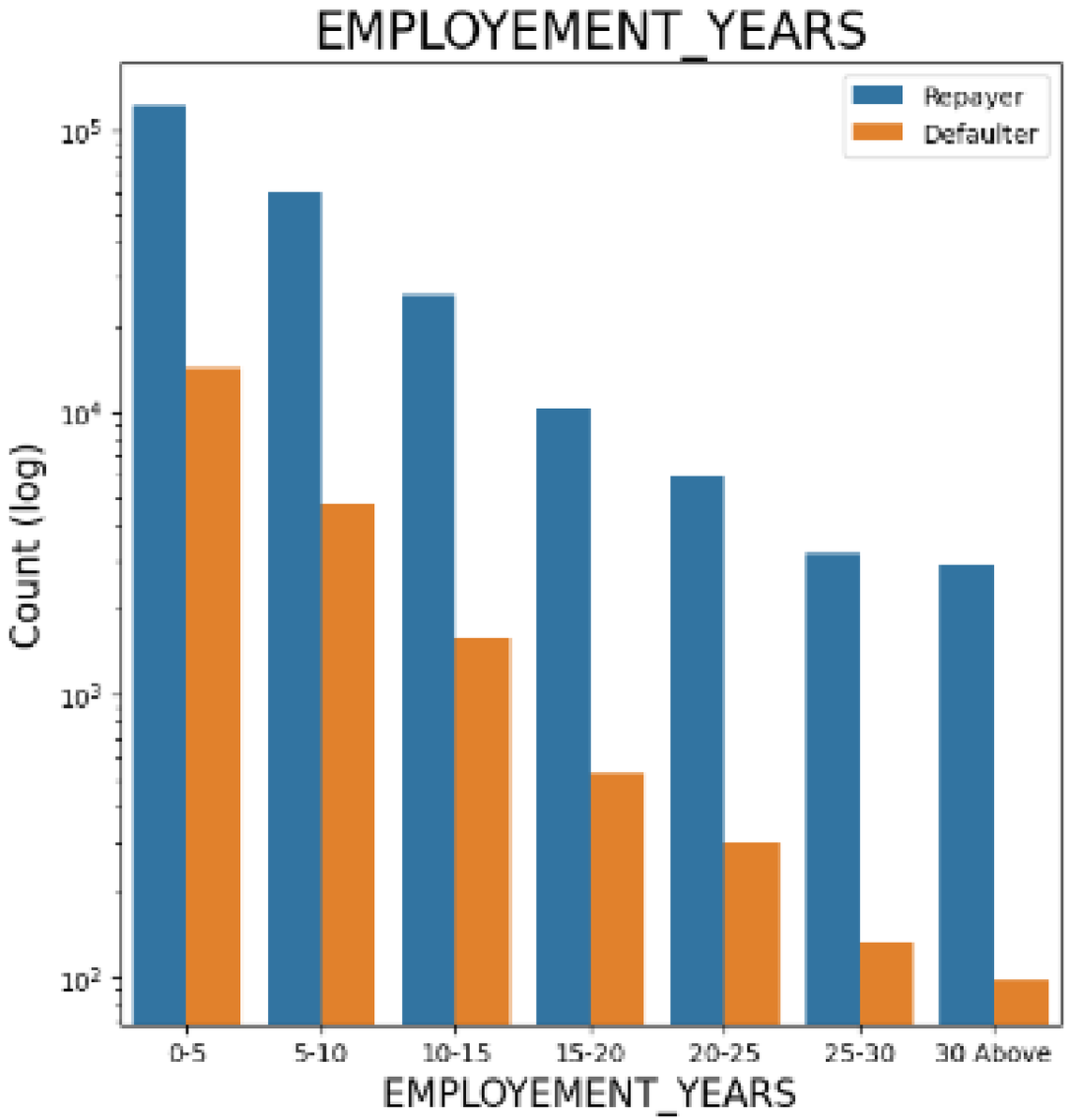

OCCUPATION_TYPE
Repayer
Defaulter
Count (log)
OCCUPATION_TYPE
Defaulters % in OCCUPATION_TYPE

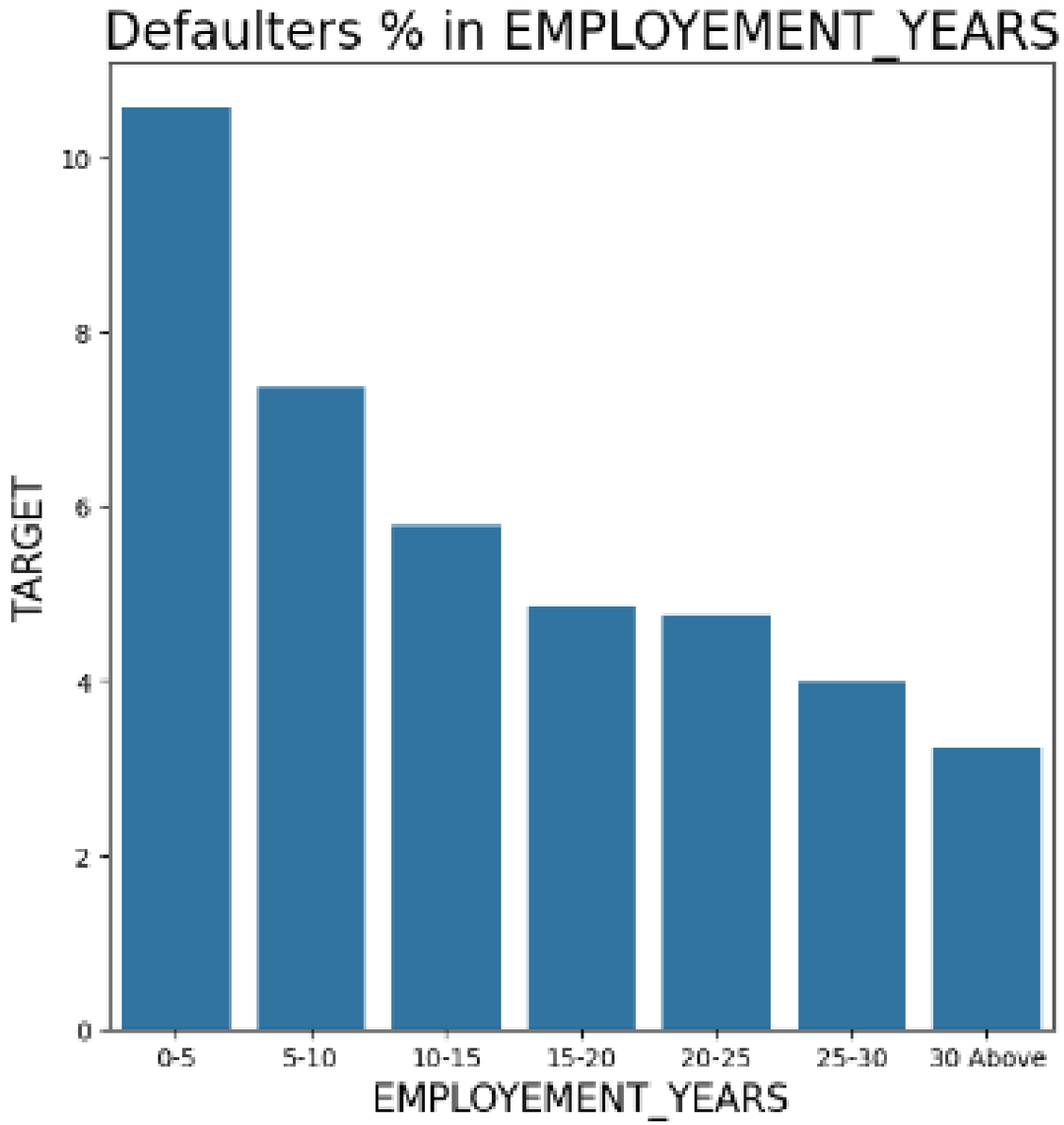

OCCUPATION_TYPE
Defaulters % in OCCUPATION_TYPE
TARGET
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Accountants
Cleaning staff
Cooking staff
Core staff
Drivers
HR staff
High skill tech staff
IT staff
Laborers
Low-skill Laborers
Managers
Medicine staff
Other
Private service staff
Realty agents
Sales staff
Secretaries
Security staff
Waiters/barmen staff
OCCUPATION_TYPE

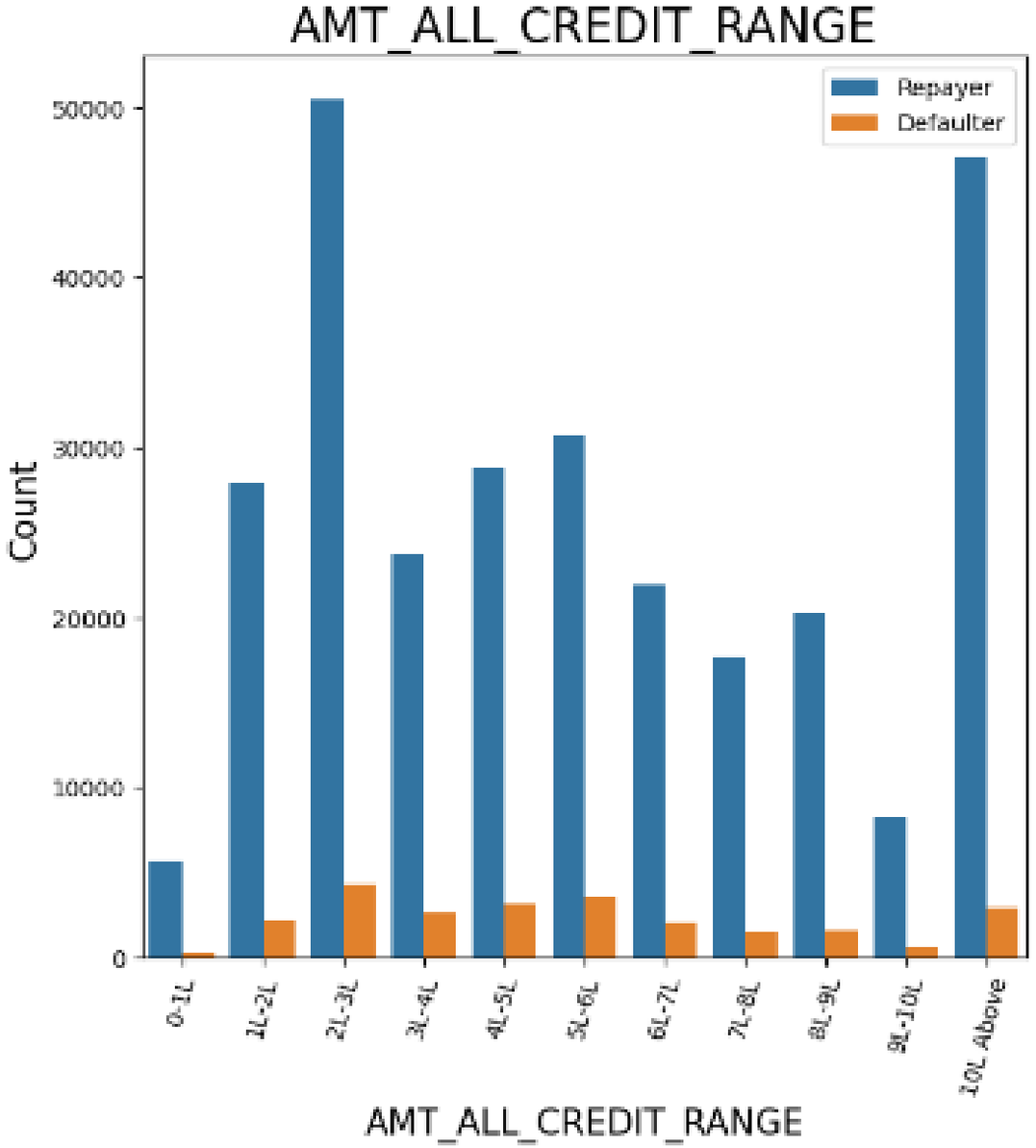


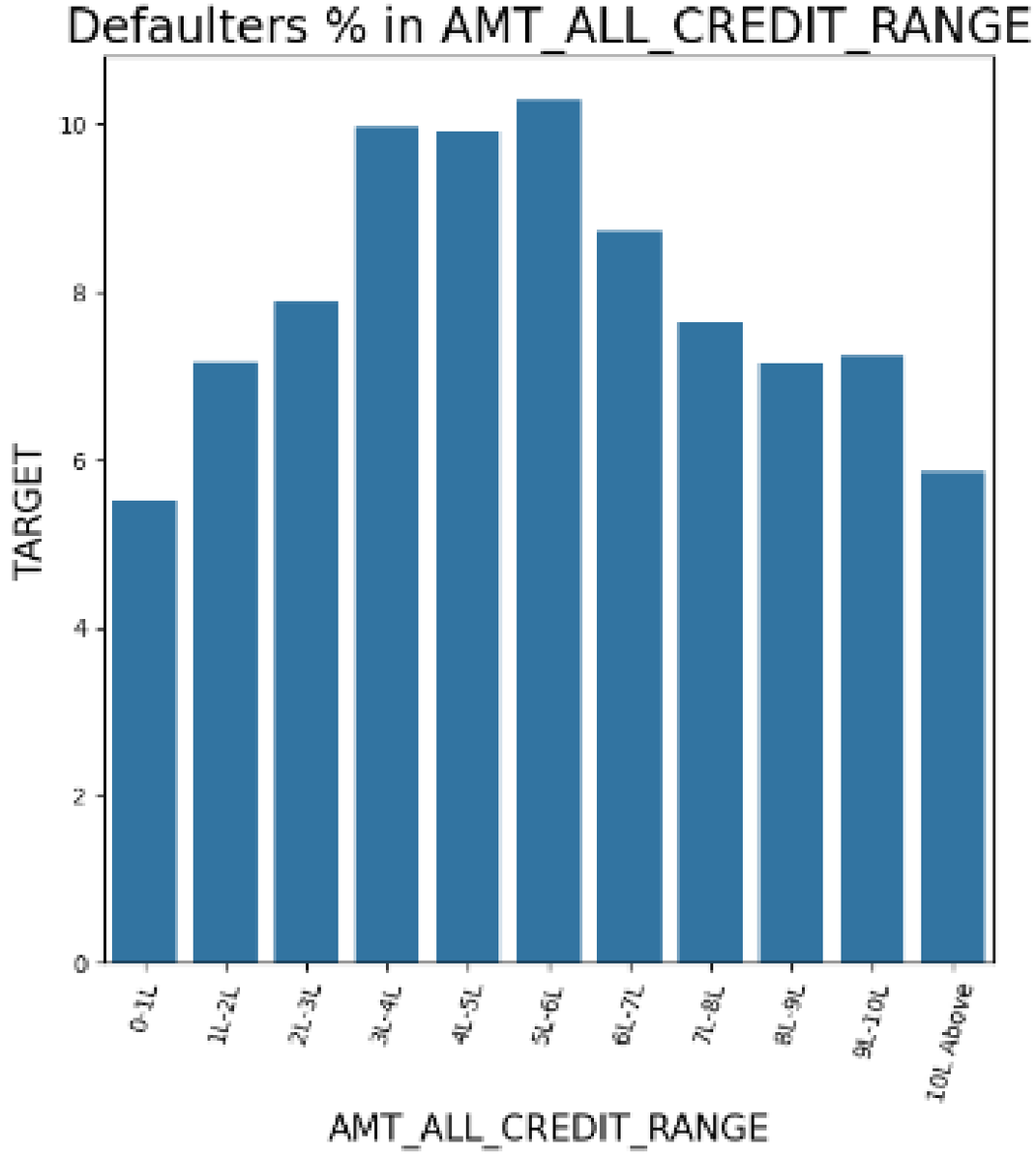
Defaulters % in ORGANIZATION_TYPE


ORGANIZATION_TYPE
Defaulters % in ORGANIZATION_TYPE
TARGET
0
2
4
6
8
10
12
14
16
Advertising
Agriculture
Bank
Business Entity Type 1
Business Entity Type 2
Business Entity Type 3
Cleaning
Construction
Culture
Electricity
Emergency
Government
Hotel
Housing
Industry: type 1
Industry: type 10
Industry: type 11
Industry: type 12
Industry: type 13
Industry: type 2
Industry: type 3
Industry: type 4
Industry: type 5
Industry: type 6
Industry: type 7
Industry: type 8
Industry: type 9
Insurance
Kindergarten
Legal Services
Medicine
Military
Mobile
Other
Police
Postal
Realtor
Religion
Restaurant
School
Security
Security Ministries
Self-employed
Services
Telecom
Trade: type 1
Trade: type 2
Trade: type 3
Trade: type 4
Trade: type 5
Trade: type 6
Trade: type 7
Transport: type 1
Transport: type 2
Transport: type 3
Transport: type 4
University
XNA
ORGANIZATION_TYPE


EMPLOYEMENT_YEARS
Repayer
Defaulter
Count (log)
10^5
10^4
10^3
10^2
0-5
5-10
10-15
15-20
20-25
25-30
30 Above
EMPLOYEMENT_YEARS

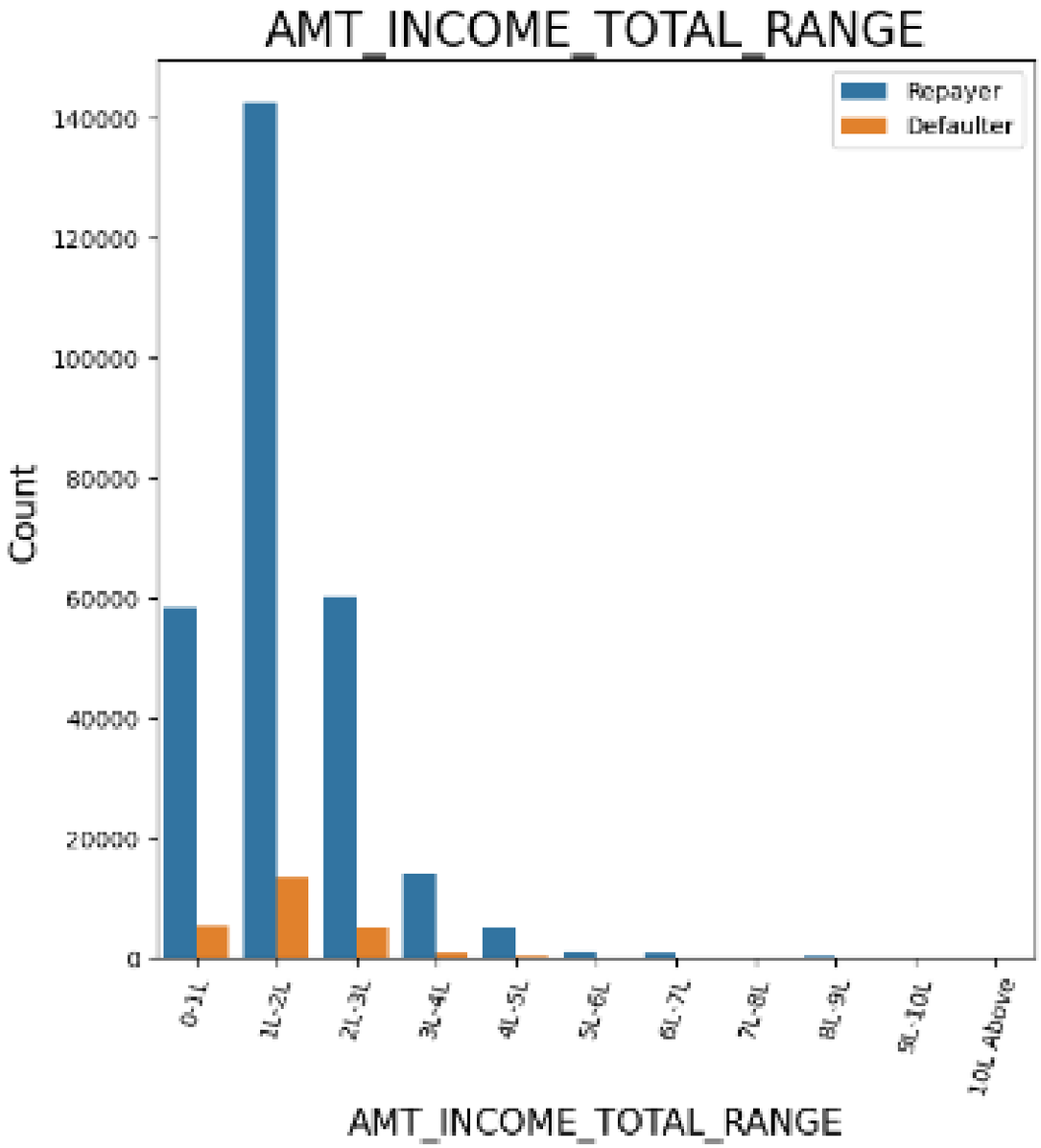

Defaulters % in EMPLOYEMENT_YEARS
TARGET
10
8
6
4
2
0
0-5
5-10
10-15
15-20
20-25
25-30
30 Above
EMPLOYEMENT_YEARS

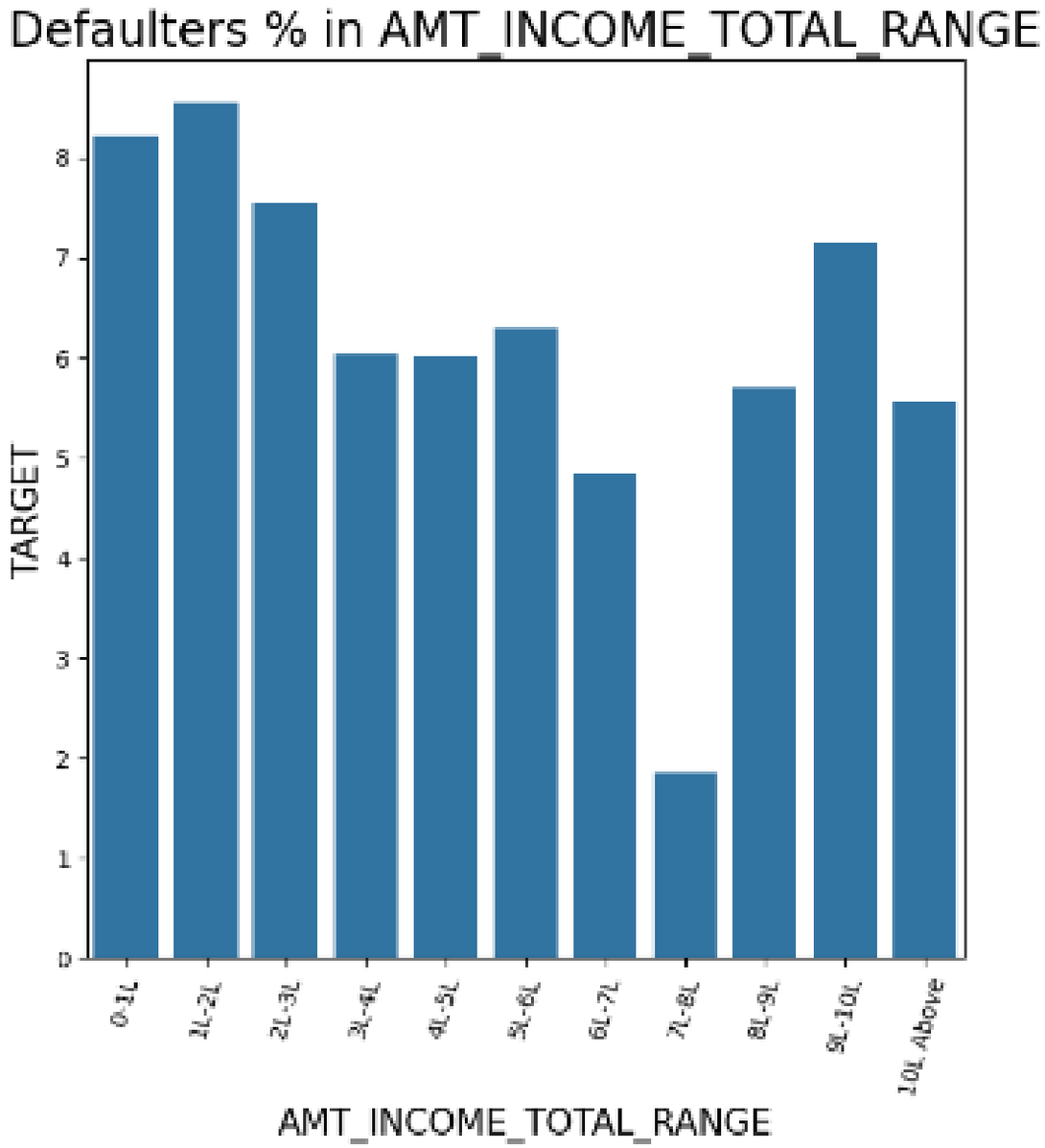

AMT_ALL_CREDIT_RANGE
Repayer
Defaulter
Count
50000
40000
30000
20000
10000
0
0-1L
1L-2L
2L-3L
3L-4L
4L-5L
5L-6L
6L-7L
7L-8L
8L-9L
9L-10L
10L Above
AMT_ALL_CREDIT_RANGE

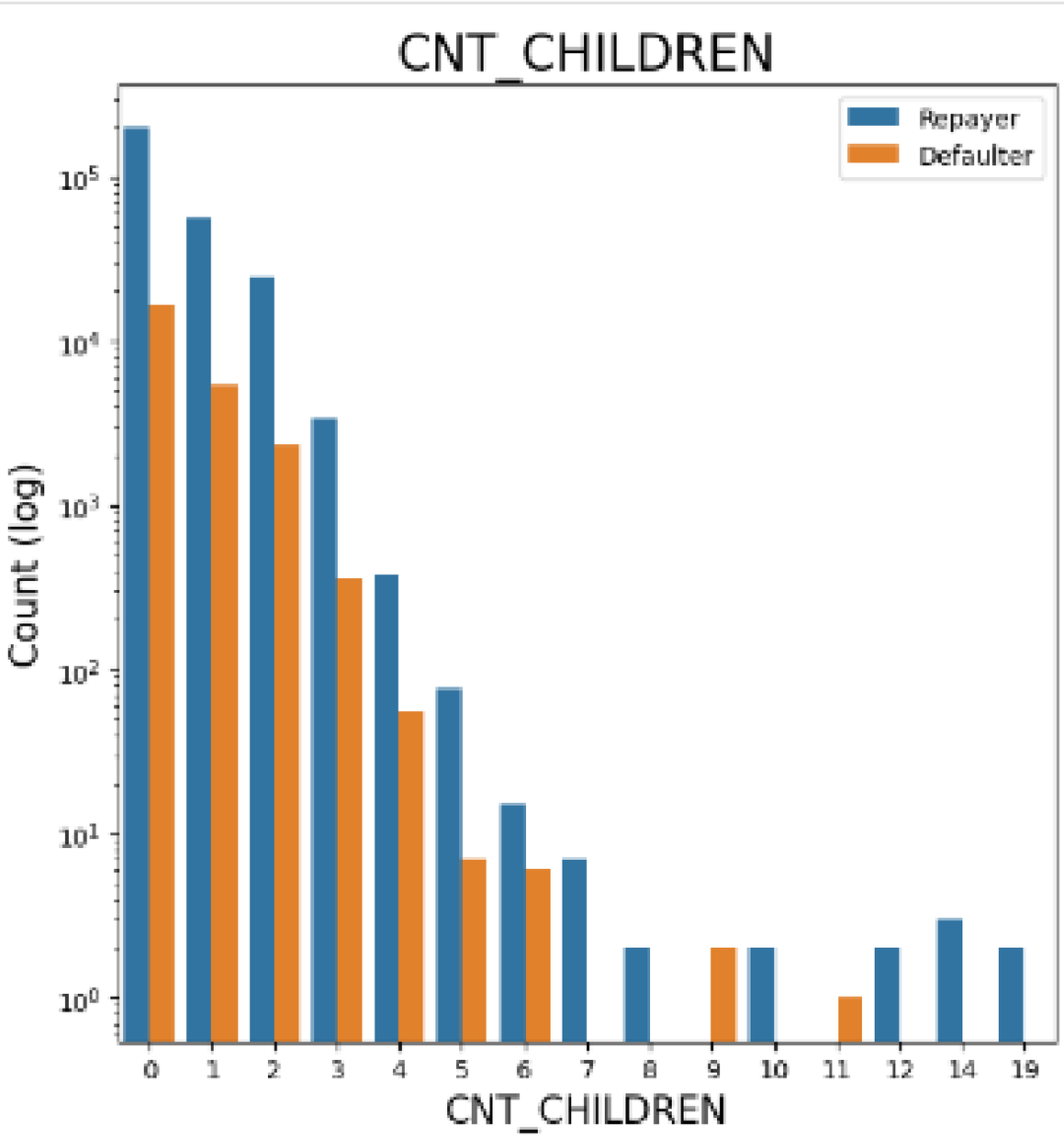

Defaulters % in AMT_ALL_CREDIT_RANGE
TARGET
10
8
6
4
2
0
0-1L
1L-2L
2L-3L
3L-4L
4L-5L
5L-6L
6L-7L
7L-8L
8L-9L
9L-10L
10L Above
AMT_ALL_CREDIT_RANGE

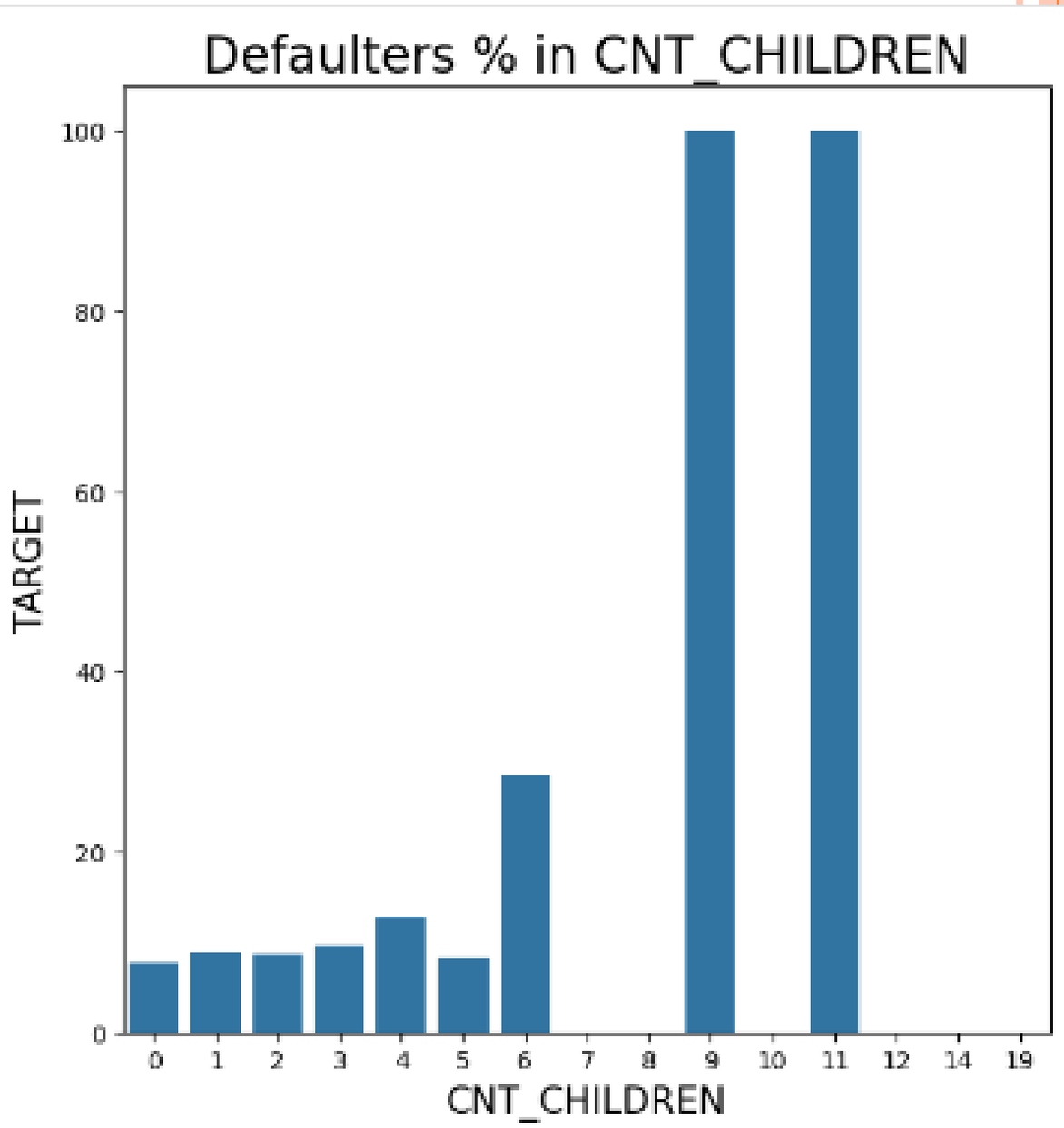
Inferences

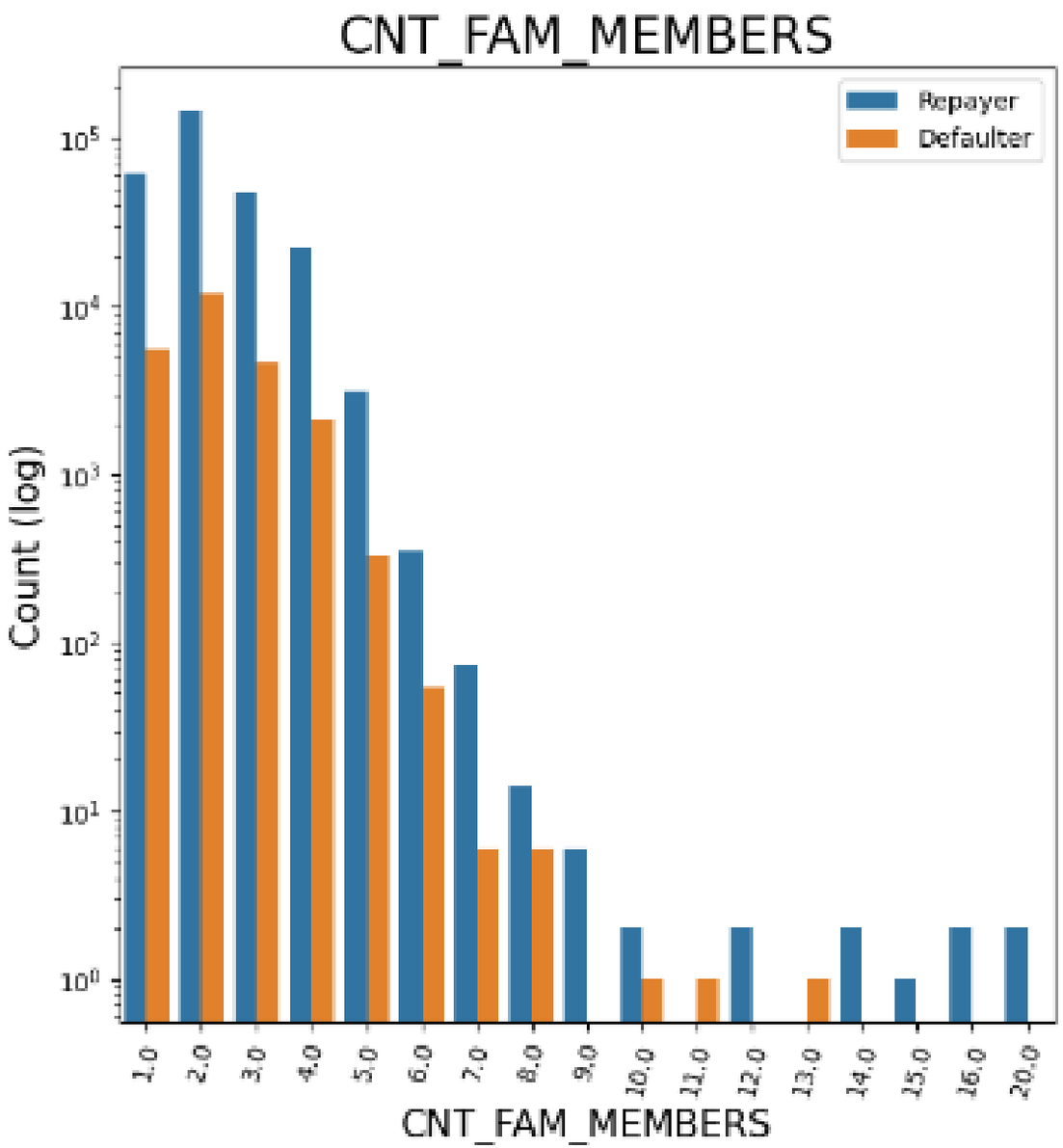
- Category with highest percent of defaulters are Low-skill Laborers (above 17%), followed by Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff.
- IT staff are less likely to apply for Loan, and most of the loans are taken by Laborers, followed by Sales staff.
- Self employed people have relative high defaulting rate, to be safer side loan disbursement should be avoided or provide loan with higher interest rate to mitigate the risk of defaulting.
- Organizations with highest percent of defaulters are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%).
- With people having 40+ year experience have less than 1% default rate. With increase of employment year, defaulting rate is gradually decreasing.
- Majority of the applicants having working experience between 0-5 years are defaulters. The defaulting rating of this group is also the highest which is around 10%.

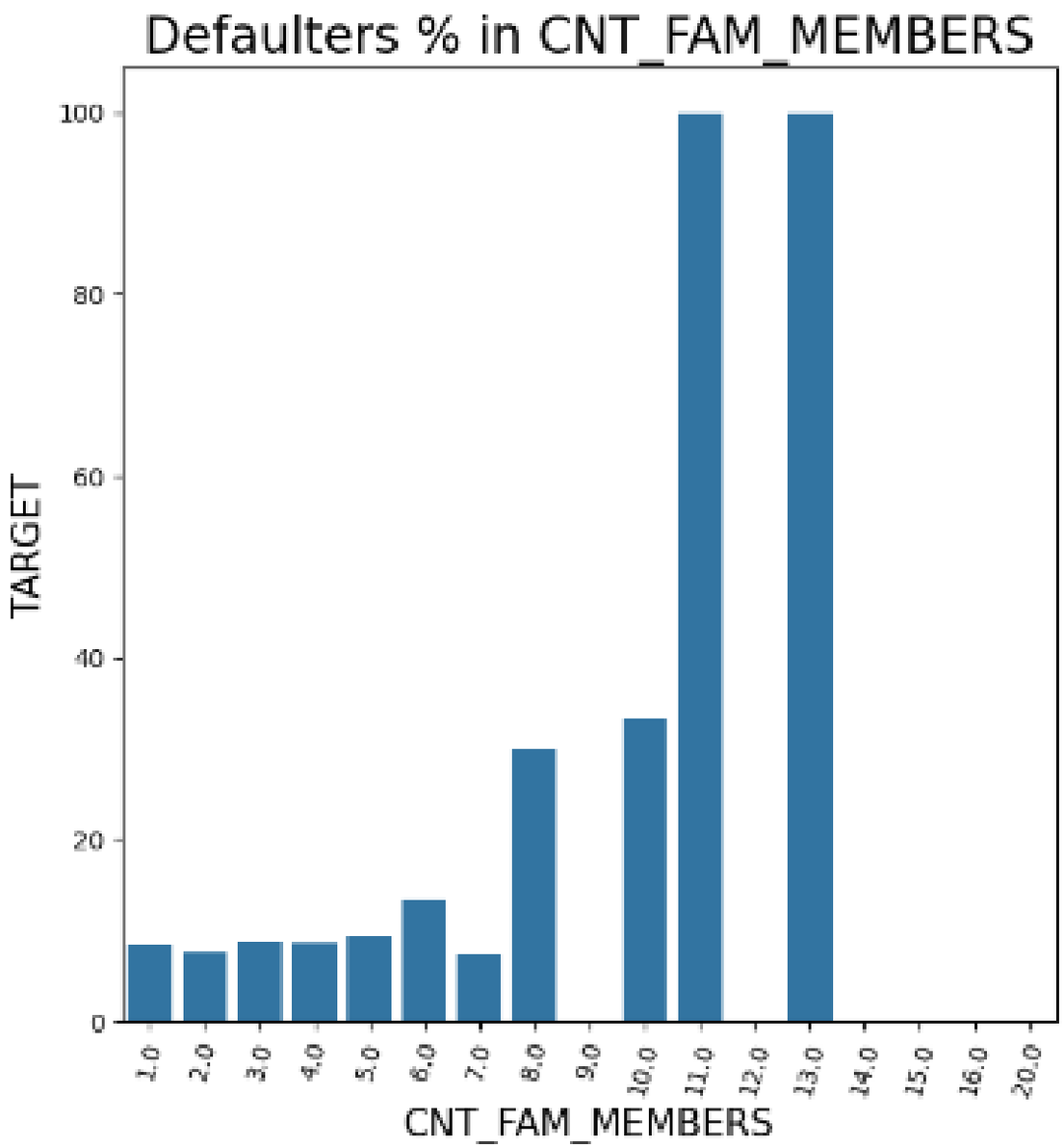

AMT_INCOME_TOTAL_RANGE
Repayer
Defaulter
Count
140000
120000
100000
80000
60000
40000
20000
0
0-1L
1L-2L
2L-3L
3L-4L
4L-5L
5L-6L
6L-7L
7L-8L
8L-9L
9L-10L
10L Above
AMT_INCOME_TOTAL_RANGE


Defaulters % in AMT_INCOME_TOTAL_RANGE
TARGET
8
7
6
5
4
3
2
1
0
0-1L
1L-2L
2L-3L
3L-4L
4L-5L
5L-6L
6L-7L
7L-8L
8L-9L
9L-10L
10L Above
AMT_INCOME_TOTAL_RANGE


CNT_CHILDREN
Repayer
Defaulter
Count (log)
10^5
10^4
10^3
10^2
10^1
10^0
0 1 2 3 4 5 6 7 8 9 10 11 12 14 19
CNT_CHILDREN


Defaulters % in CNT_CHILDREN
TARGET
100
80
60
40
20
0
0 1 2 3 4 5 6 7 8 9 10 11 12 14 19
CNT_CHILDREN

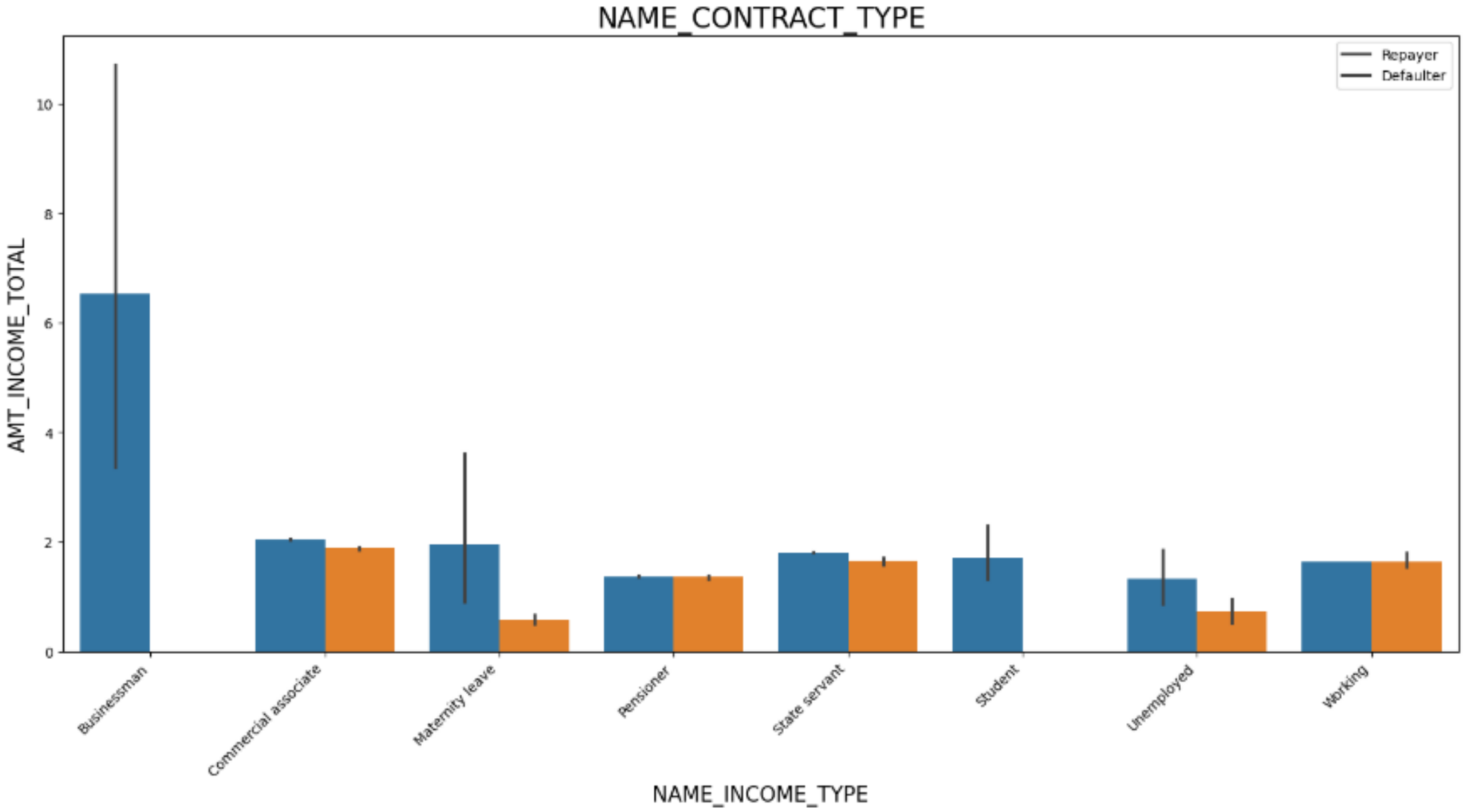

CNT_FAM_MEMBERS
Repayer
Defaulter
Count (log)
10^5
10^4
10^3
10^2
10^1
10^0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
14.0
15.0
16.0
20.0
CNT_FAM_MEMBERS


Defaulters % in CNT_FAM_MEMBERS
TARGET
100
80
60
40
20
0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
14.0
15.0
16.0
20.0
CNT_FAM_MEMBERS

Inferences

- Applicants who get loan for 3L-6L have most number of defaulters than other loan range.
- Applicants with high number have loan in range of 2L-3L.
- Majority of the applications have Income total less than 3L.
- Application with Income less than 3L, has high probability of defaulting, and applicant with Income 7L-8L are less likely to default.
- Applicants who have more than 4 children has a very high default rate with child count 9 and 11 showing 100% default rate
- Most of the applicants do not have children, and very few clients have more than 3 children.
- Family member follows the same trend as children where having more family members increases the risk of defaulting.

Categorical Bivariate / Multivariate Analysis

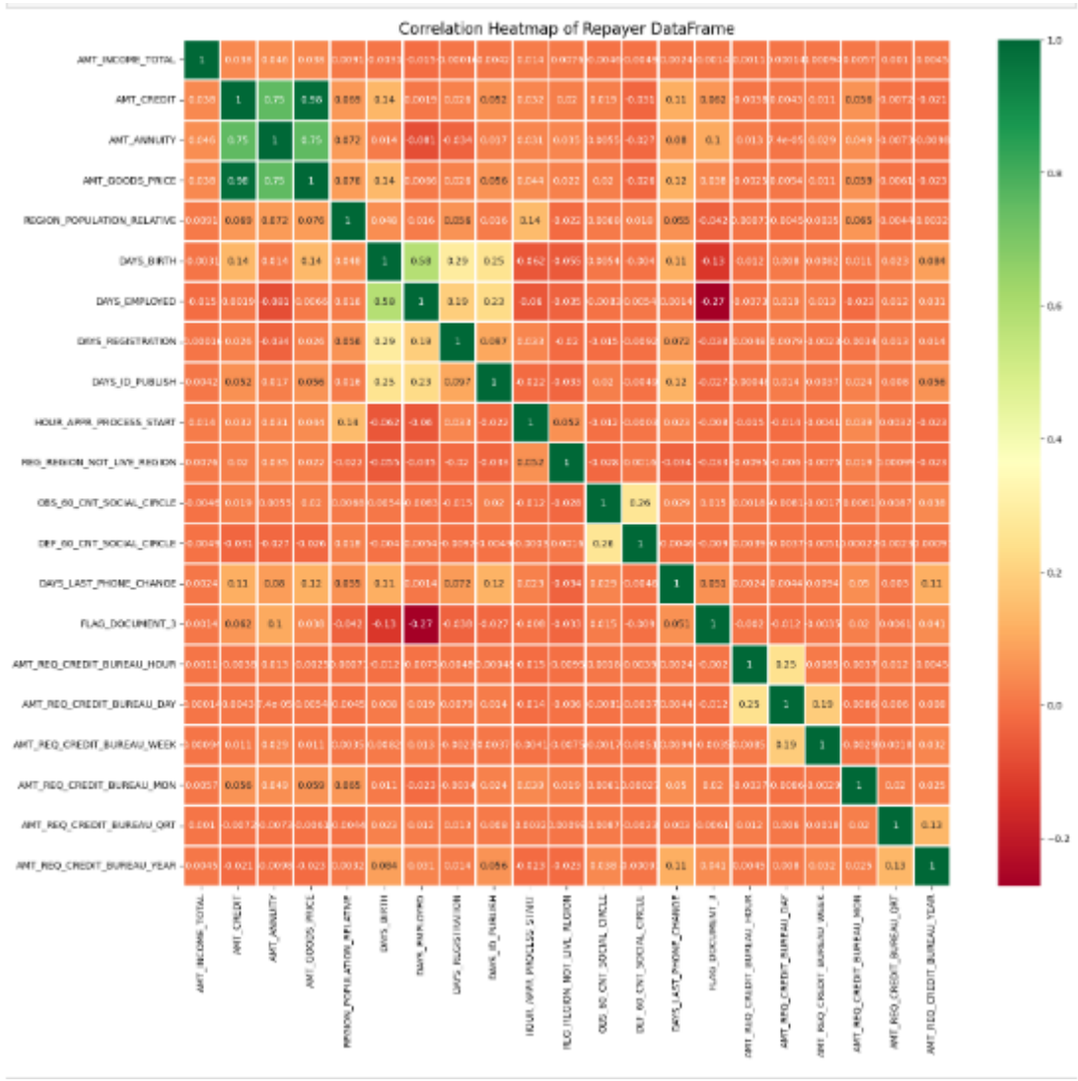


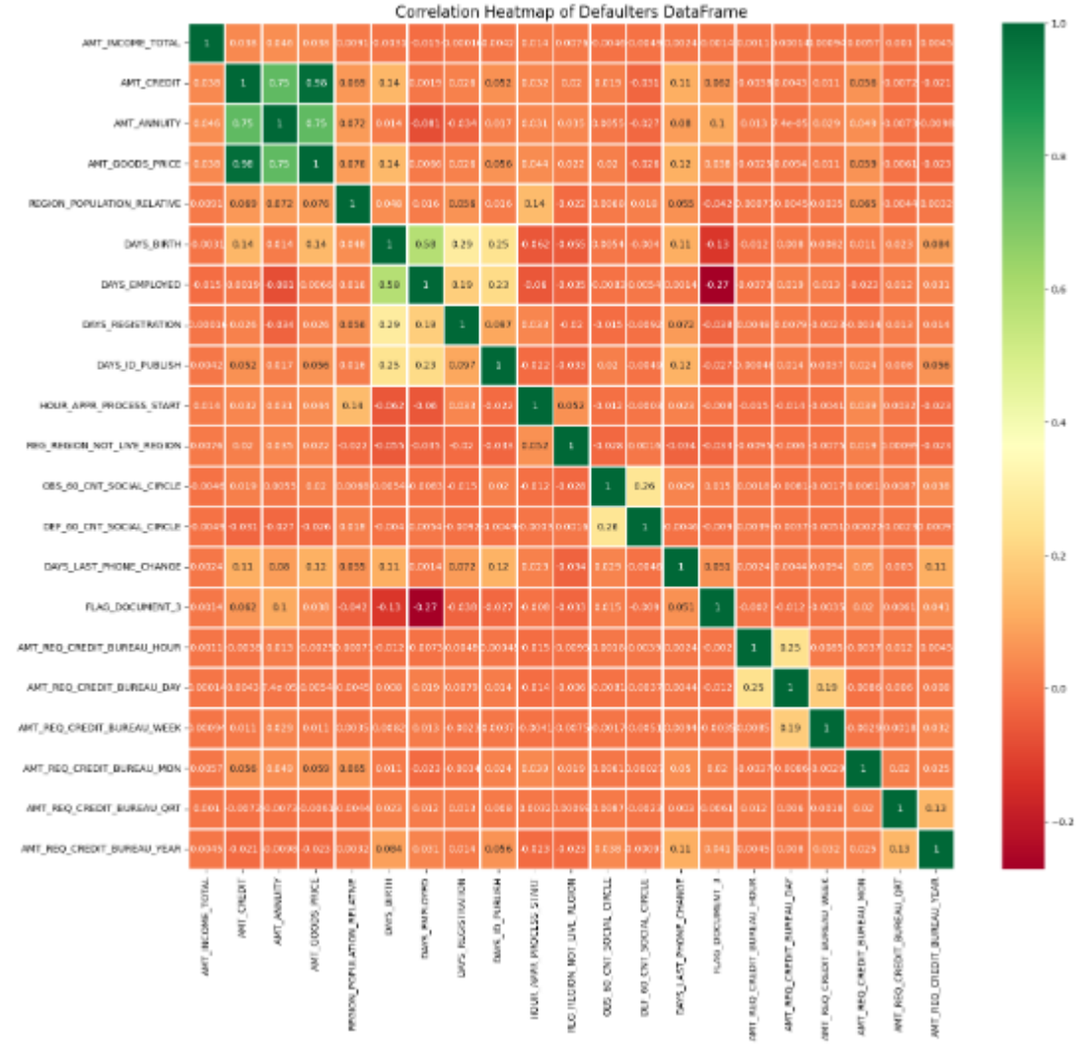
Inferences

- Businessman income is the highest and the estimated range with default 95% confidence level and it seems to indicate that the income of a Businessman could be in the range of slightly close to 4L and slightly more than 10L.

Numeric Multivariate Analysis

This is based on Target values 0 and 1 for correlation and other analysis.

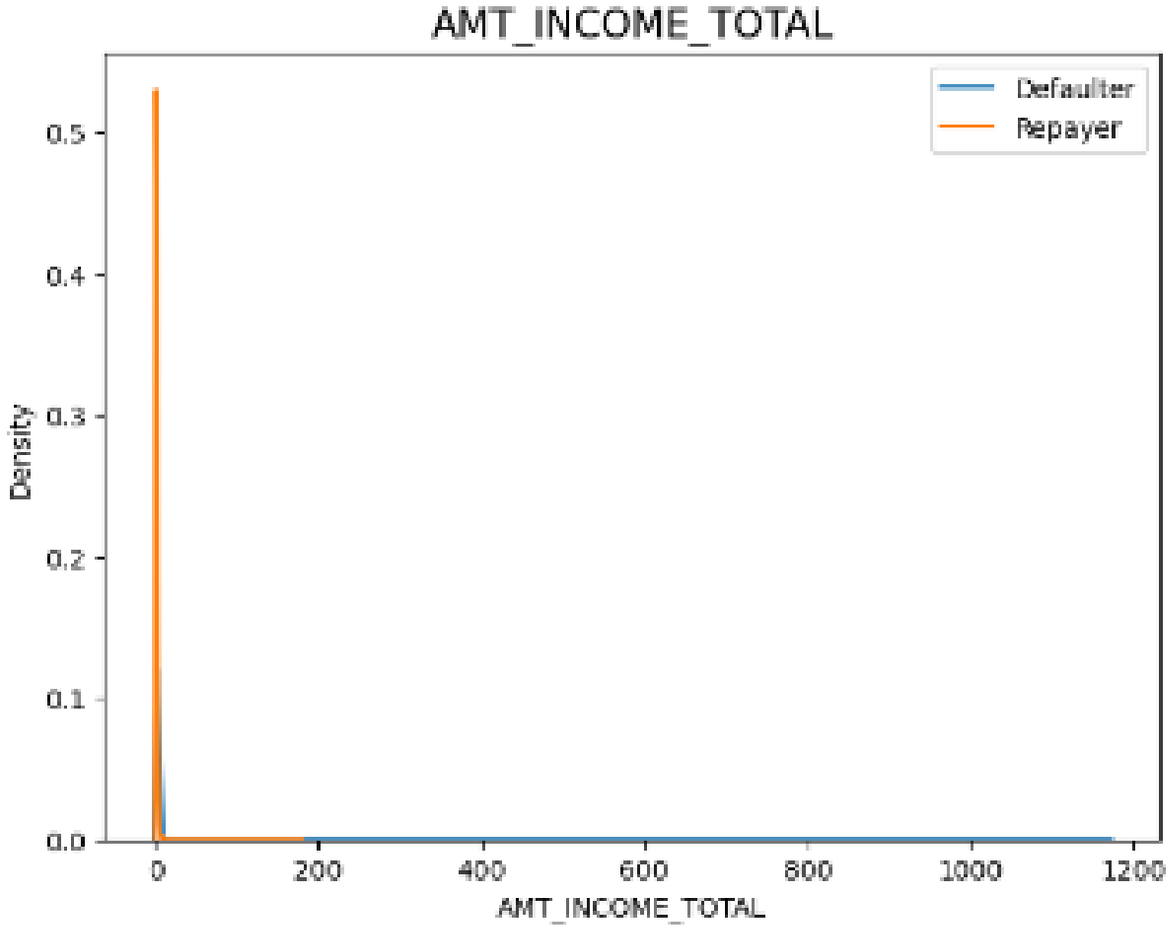


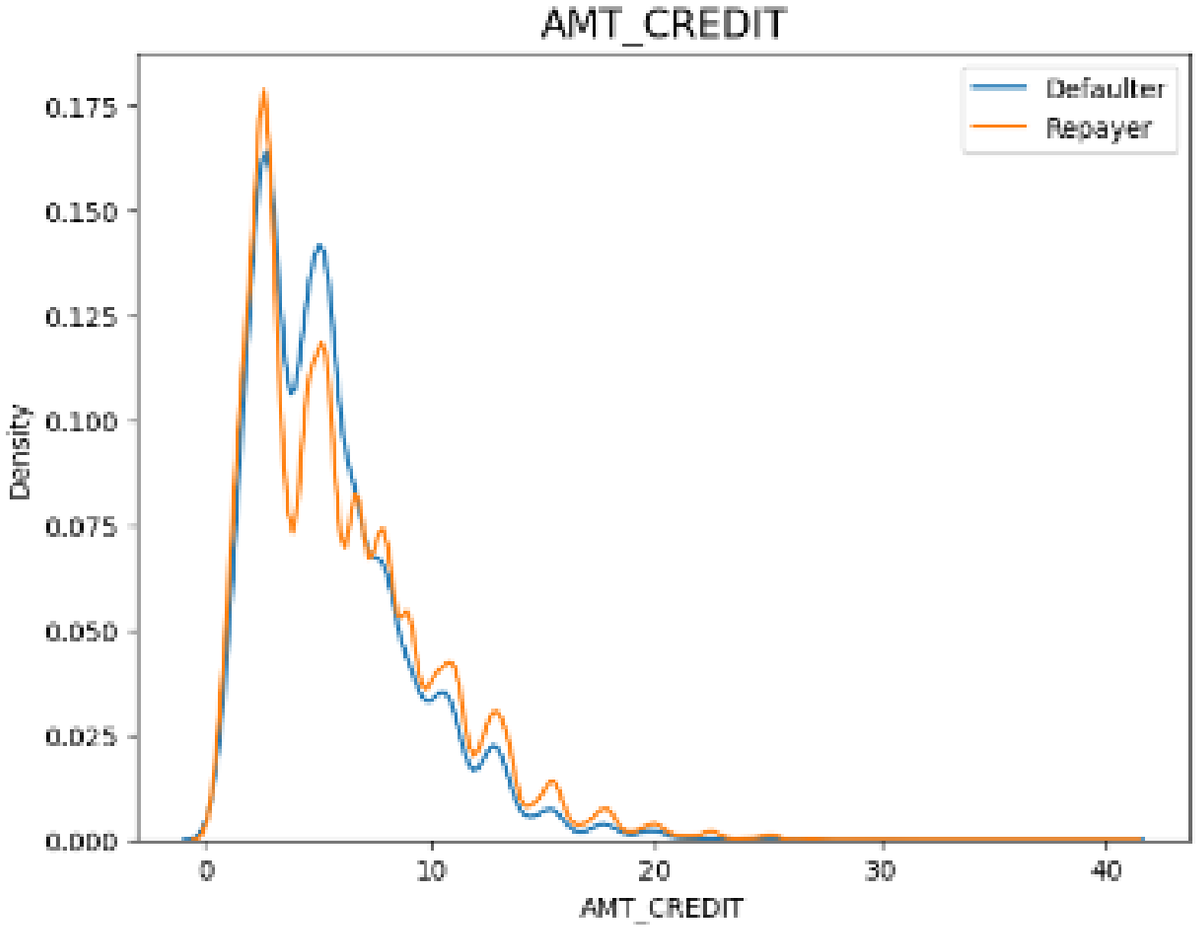


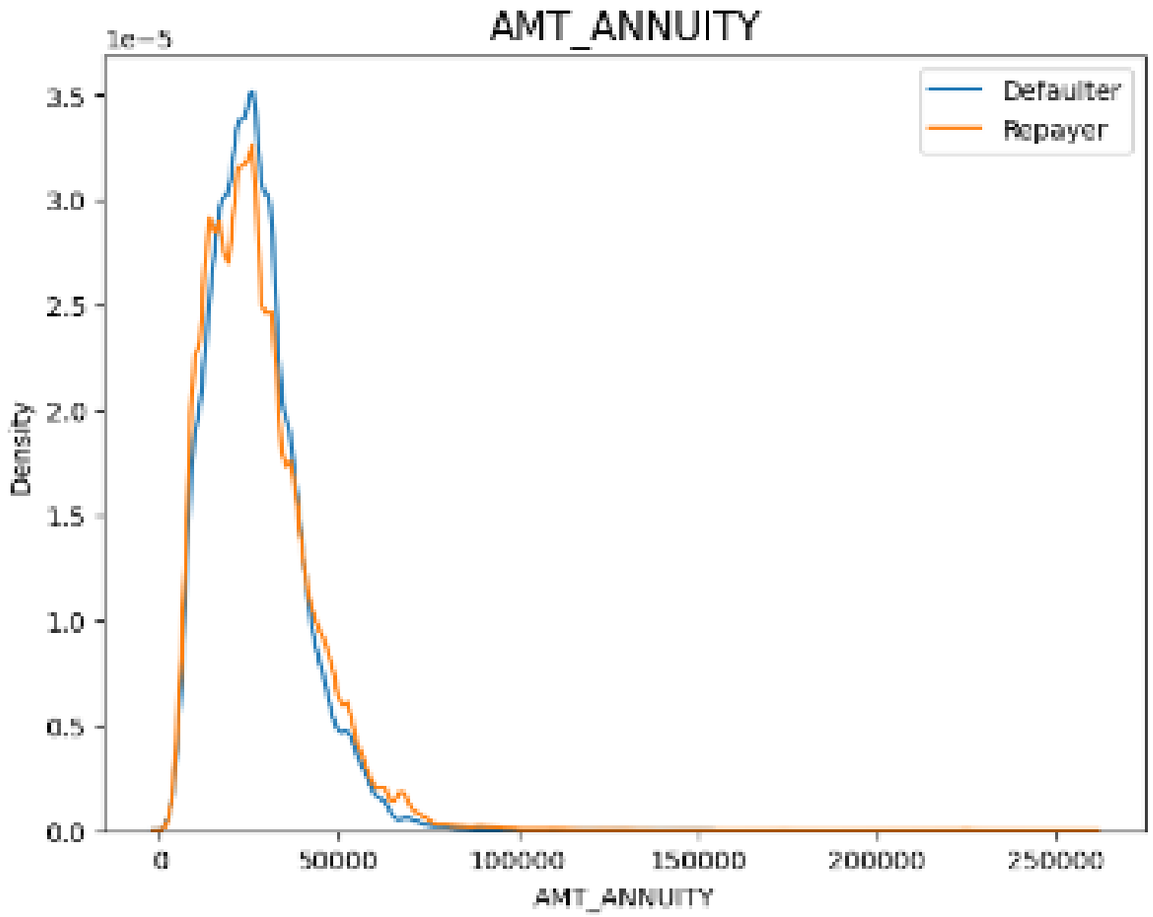
Inferences

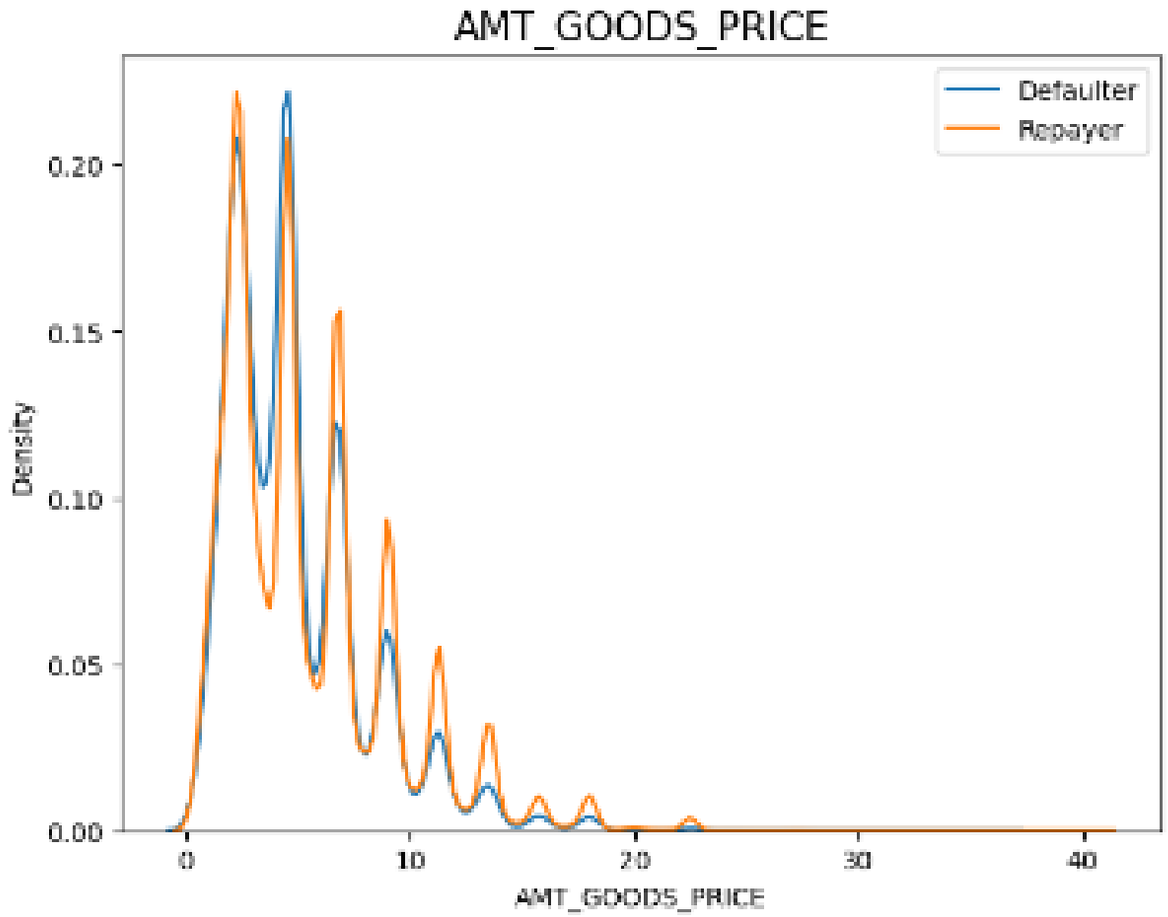
- Credit amount is highly correlated with:
  - Goods Price Amount
  - Loan Annuity
  - Total Income
- We can also see that Repayers have high correlation in number of days employed.
- Credit amount is highly correlated with good price amount which is same as Repayers.
- There is a slight increase in defaulted to observed count in social circle among Defaulters is 0.264) when compared to Repayers is 0.254
- Days_birth and number of children correlation has reduced to 0.259 in Defaulters when compared to 0.337 in Repayers.
- Loan annuity correlation with credit amount has slightly reduced in Defaulters is 0.75 when compared to Repayers is 0.77
- We can also see that Repayers have high correlation in number of days employed is 0.62 when compared to Defaulters is 0.58. 8 .There is a severe drop in the correlation between total income of the client and the credit amount (0.038) amongst Defaulters whereas it is 0.342 among Repayers.

Density distribution amount





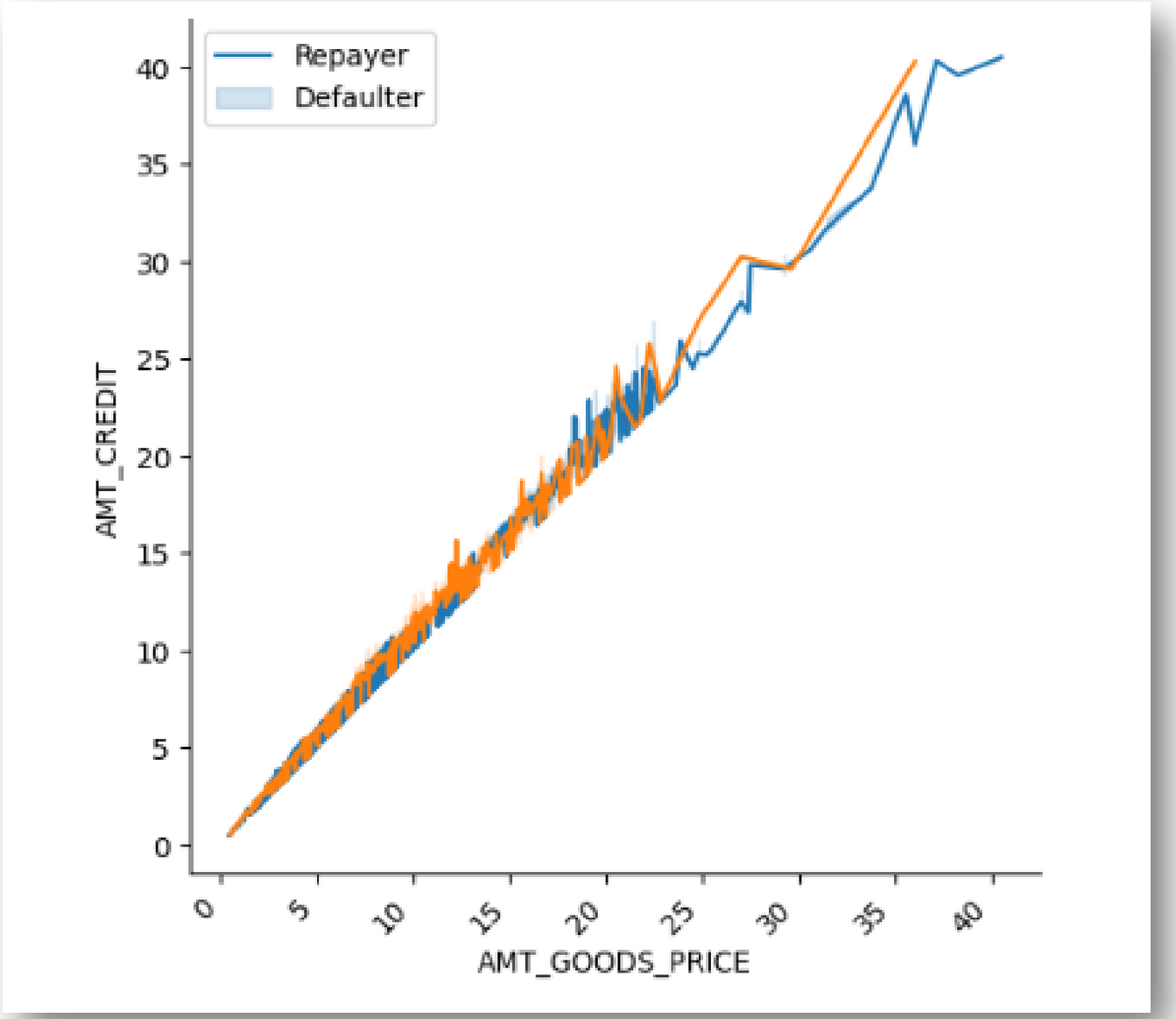

AMT_ANNUITY
1e−5
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Density
0
50000
100000
150000
200000
250000
AMT_ANNUITY
Defaulter
Repayer


AMT_GOODS_PRICE
0.20
0.15
0.10
0.05
0.00
Density
0
10
20
30
40
AMT_GOODS_PRICE
Defaulter
Repayer

Inferences

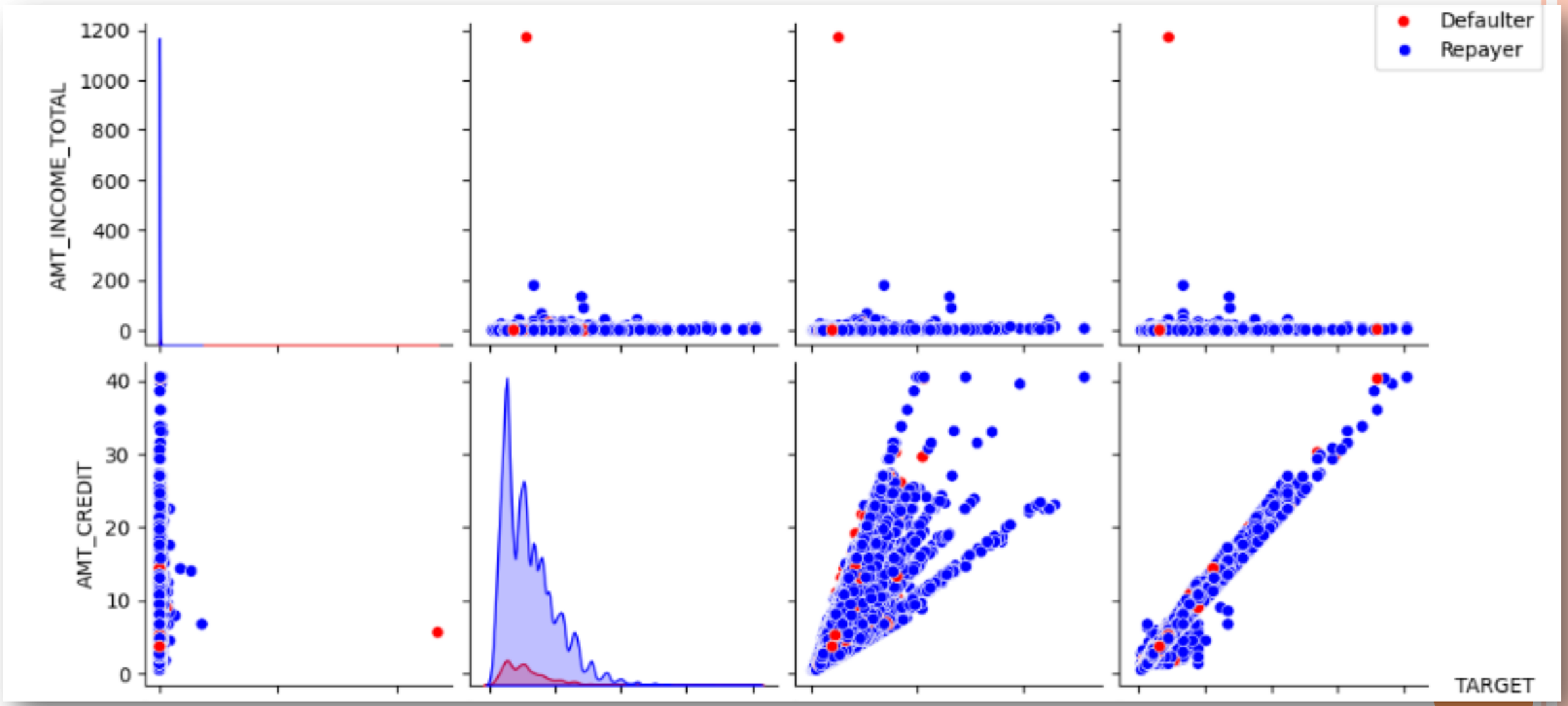
- The Repayers and Defaulters distribution overlap in all the plots and hence we cannot use any of these variables in isolation to make a decision.
- Most no of loans are given for goods price below 10L. Credit amount of the loan is mostly less then 10L
- Most people pay annuity below 50K for the credit loan.

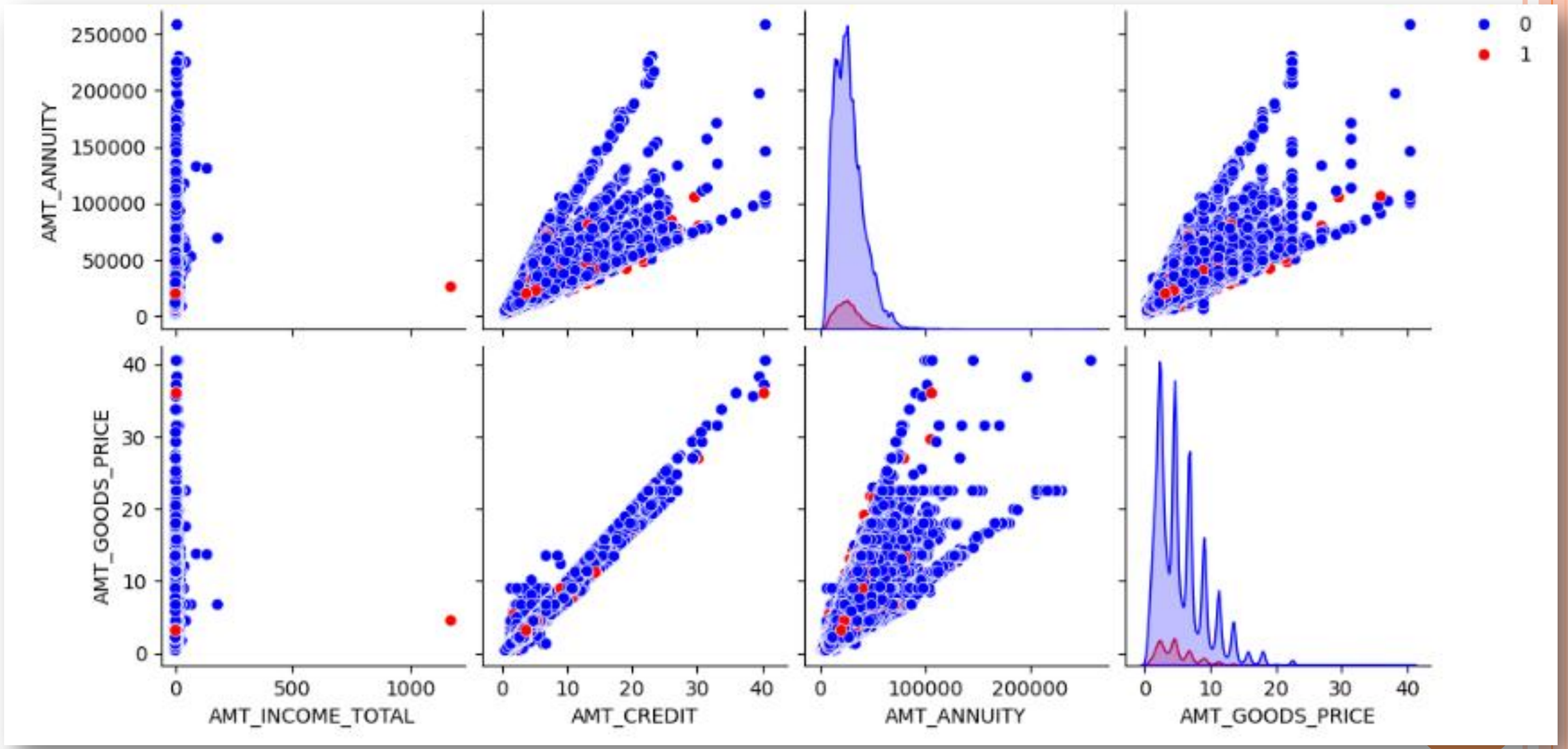
Numerical Bivariate Analysis



Inferences

➢When the credit amount goes beyond 30L, there is an increase in defaulters.


Defaulter
Repayer
AMT_INCOME_TOTAL
1200
1000
800
600
400
200
0
AMT_CREDIT
40
30
20
10
0
TARGET

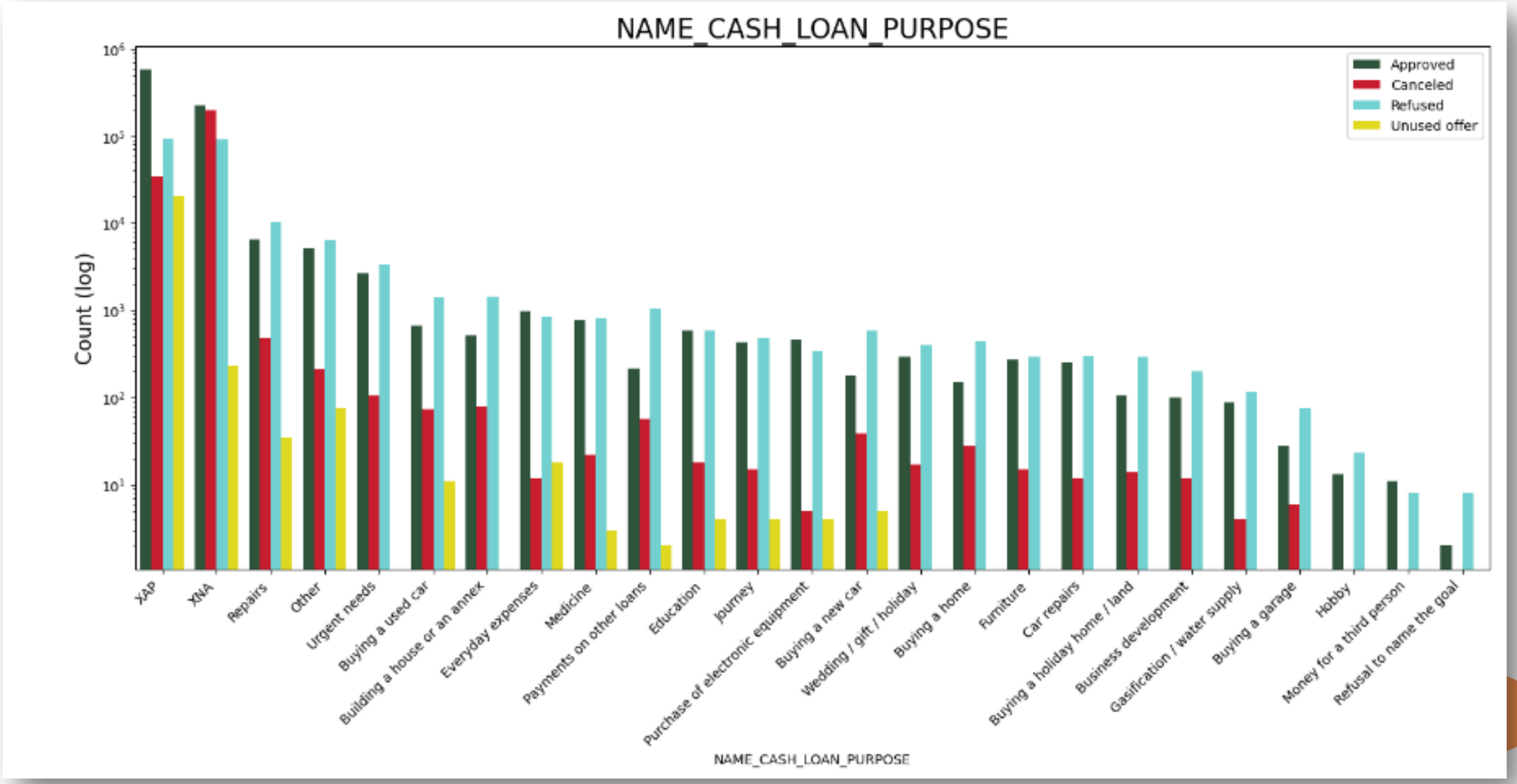

AMT_ANNUITY
AMT_GOODS_PRICE
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
0
1

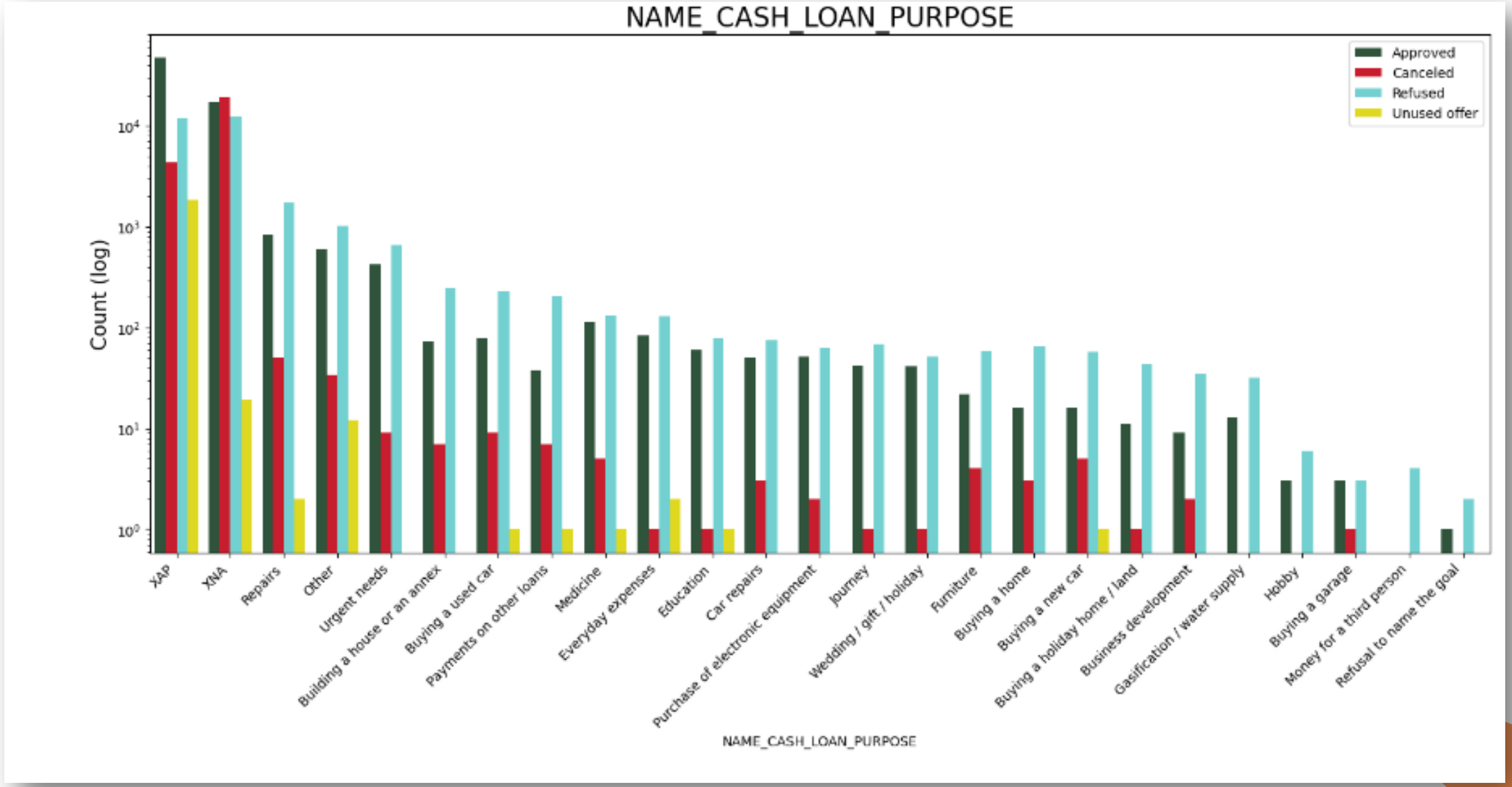
Inferences

- When Annuity Amount > 15K and Good Price Amount > 20L, there is a lesser chance of Defaulters, and there are very less Defaulters for AMT_CREDIT >20L
- Loan Amount(AMT_CREDIT) and Goods price(AMT_GOODS_PRICE) are highly correlated as based on the scatter plot where most of the data are consolidated in form of a line

Merged Data frames Analysis

Bisecting the "loan_data" dataframe based on
Target value 0 and 1 for correlation and other analysis



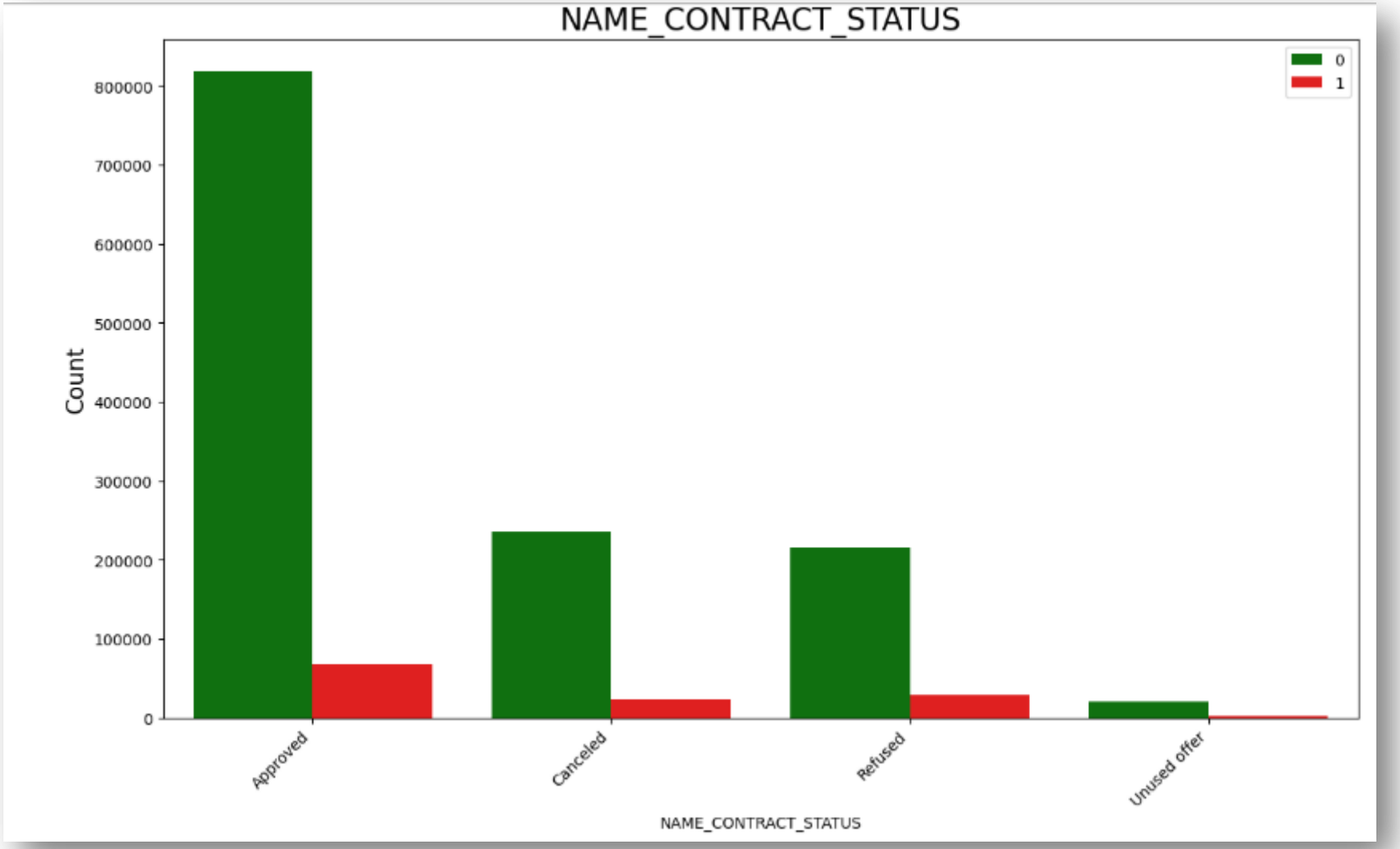

NAME_CASH_LOAN_PURPOSE
Approved
Canceled
Refused
Unused offer
Count (log)
10^4
10^3
10^2
10^1
10^0
XAP
XNA
Repairs
Other
Urgent needs
Building a house or an annex
Buying a used car
Payments on other loans
Medicine
Everyday expenses
Education
Car repairs
Purchase of electronic equipment
Journey
Wedding / gift / holiday
Furniture
Buying a home
Buying a new car
Buying a holiday home / land
Business development
Gasification / water supply
Hobby
Buying a garage
Money for a third person
Refusal to name the goal
NAME_CASH_LOAN_PURPOSE

Inferences

- Loan purpose has high number of unknown values (XAP, XNA) • Loan taken for the purpose of Repairs looks to have highest default rate • Huge number application have been rejected by bank or refused by applicant who is applied for Repair or Other.

- From this we can infer that repair is considered high risk by bank. Also, either they are rejected or bank offers loan on high interest rate which is not feasible by the clients and they refuse the loan.

Univariate Merged Data

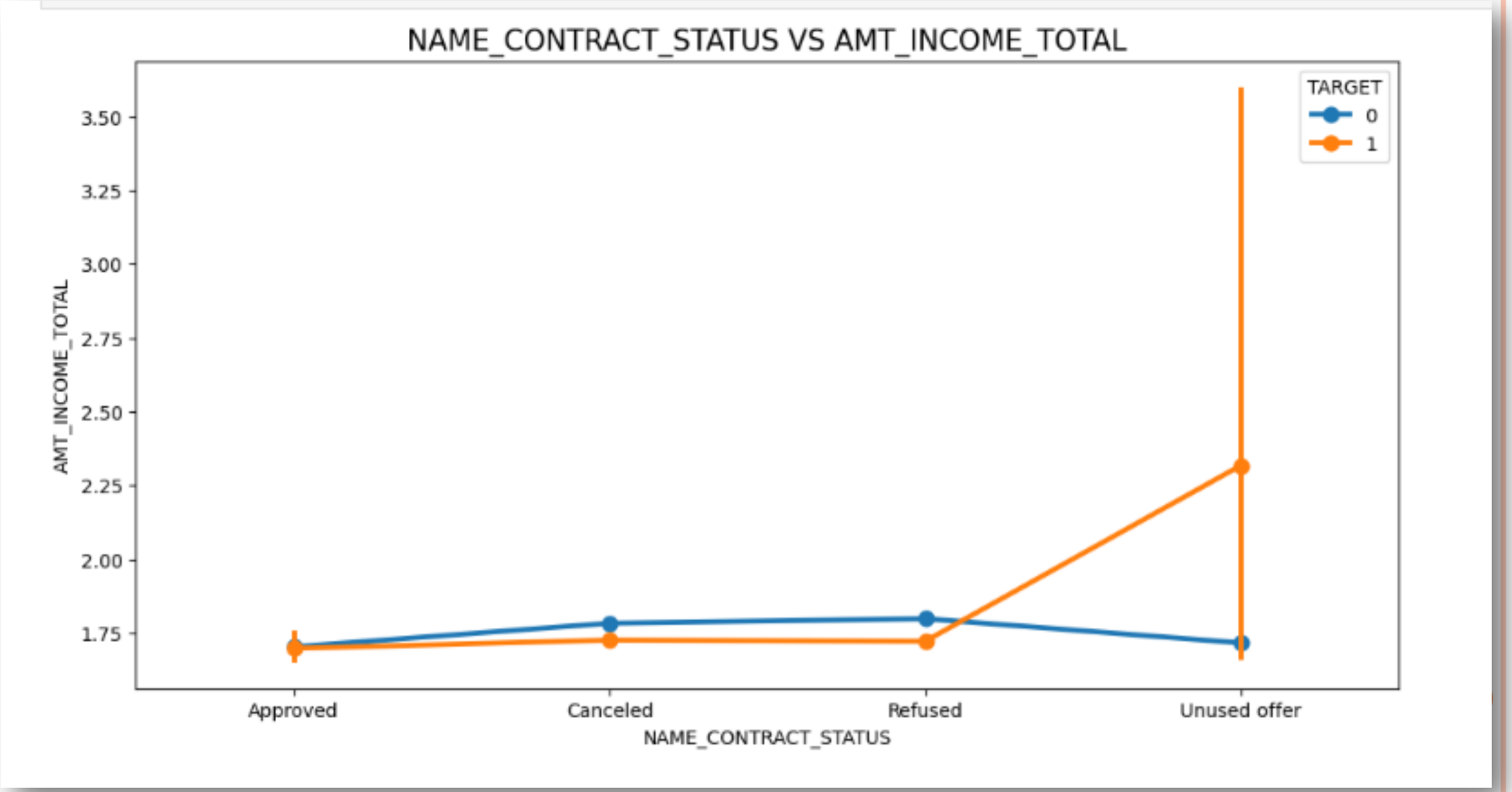


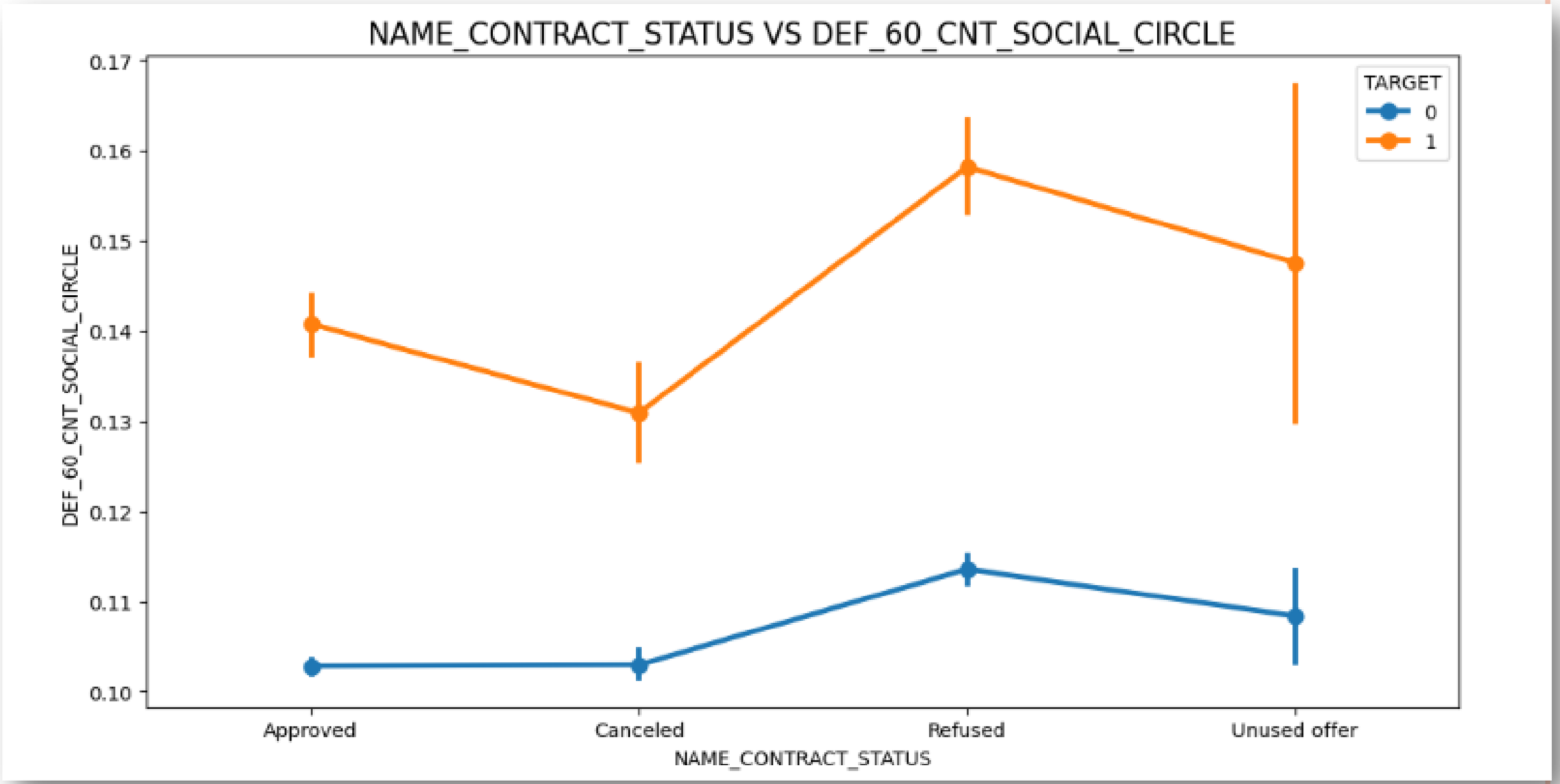
NAME_CONTRACT_STATUS	TARGET	Counts	Percentage
Approved	0	818856	92.41%
	1	67243	7.59%
Canceled	0	235641	90.83%
	1	23800	9.17%
Refused	0	215952	88.0%
	1	29438	12.0%
Unused offer	0	20892	91.75%
	1	1879	8.25%

Inferences

- Refusal reason should be recorded for further analysis as these clients could turn into potential repaying customer.
- 90% of the previously cancelled applicants have actually Repayed the loan. Revising the interest rates would increase business opportunity for these clients
- 88% of the clients who have been previously refused a loan has payed back the loan in current case.

Bivariate Merged Data




NAME_CONTRACT_STATUS VS DEF_60_CNT_SOCIAL_CIRCLE
TARGET
0
1
DEF_60_CNT_SOCIAL_CIRCLE
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.10
Approved
Canceled
Refused
Unused offer
NAME_CONTRACT_STATUS

Inferences

- Applicants who have average of 0.13 or higher their DEF_60_CNT_SOCIAL_CIRCLE score tend to default more and thus analysing client's social circle could help in disbursement of the loan.

- The point plot show that the applicants who have not used offer earlier have defaulted even when there average income is higher than others

Conclusions

Decisive Factor whether an applicant will be Defaulter:

- REGION_RATING_APPLICANTS: Applicants who live in Rating 3 has highest defaults.
- CNT_CHILDREN & CNT_FAM_MEMBERS: Applicants who have children => 9 default 100% and so their applications are to be rejected.
- AMT_GOODS_PRICE: When the credit amount goes beyond 3L, there is an increase in defaulters.
- OCCUPATION_TYPE: Due to default rate is huge avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff.
- CODE_GENDER: Females are good repayer where as men are at relatively higher default rate.
- NAME_FAMILY_STATUS : Civil marriage or who are single applicants who have default a lot.

Decisive Factor whether an applicant will be Defaulter:

- NAME_EDUCATION_TYPE: Lower Secondary & Secondary education applicants are default a lot.
- NAME_INCOME_TYPE: Maternity leave OR Unemployed applicants are default a lot.
- ORGANIZATION_TYPE: Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self-employed applicants have relative high defaulting rate, and thus should be avoided to be approved for loan or provide loan with higher interest rate to might be the risk of defaulting.
- DAYS_BIRTH: Avoid young applicants who are in age group of 20-40 as they have higher probability of defaulting
- DAYS_EMPLOYED: Applicants who have less than 5 years of employment have high default rate.

Decisive Factor whether an applicant will be Repayer:

- CNT_CHILDREN: Applicants with 0 to 2 children tend to repay the loans.
- DAYS_BIRTH: If age above of 50 have low probability of defaulting.
- DAYS_EMPLOYED: 40+ year experience having less than 1% default rate.
- AMT_INCOME_TOTAL: Income more than 700,000 are less likely to default.
- NAME_EDUCATION_TYPE: Academic degree has fewer defaults.

Decisive Factor whether an applicant will be Defaulter:

- NAME_INCOME_TYPE: Student and Businessmen have no defaults.
- REGION_RATING_APPLICANTS: RATING 1 is safer.
- ORGANIZATION_TYPE: Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%
- NAME_CASH_LOAN_PURPOSE: Loans bought for Hobby, Buying garage are being repayed mostly.

Below are few conditions where interest rate is high might getting any default risk leading to business loss:

- CNT_CHILDREN & CNT_FAM_MEMBERS: higher interest should be imposed on their loans., who have 4 to 8 children has a very high default rate.

- AMT_INCOME: Those applicants could be offered loan with higher interest compared to other income category, since 90% of the applications have Income total less than 3 and they have high probability of defaulting.

- NAME_HOUSING_TYPE: Offering the loan would might the loss if any of applicants are from the category of people who live in rented apartments & living with parents.

- AMT_CREDIT: Having higher interest specifically for Applicants who get loan for 3L - 6L credit range would be ideal because they tend to default more than others.

- NAME_CASH_LOAN_PURPOSE: Repairs purpose seems to have highest default rate. Very high number applications have been rejected by bank or refused by client in previous applications as well which has purpose as repair or other. This shows that purpose repair is taken as high risk by bank and either bank offers very high loan interest rate which they are rejected.

Business Advices after this EDA as below:

- 88% of the applicants who were refused by bank for loan earlier have now turned into a repaying applicants. Hence documenting the reason for rejection could moderate the business loss and these clients could be contacted for further loans.

- 90% of the previously cancelled applicants have actually repayed the loan. Record the reason for cancellation which might help the bank to determine and negotiate terms with these repaying customers in future for increase business opportunity.

Thank You!!