CREDIT EXPLORATORY DATA ANALYSIS ASSINGMENT

ASHOK SINGH BANGARI

INTRODUCTION

- This assignment is all about applying EDA on 3 dataset and getting insights from the data which can help the Business
- The dataset are:
- application data: contains all information of the client at the time of application
- previous application: contains information about the client's previous loan data.
- Columns descriptions: contains dictionary, describing the variable

BUSINESS UNDERSTANDING

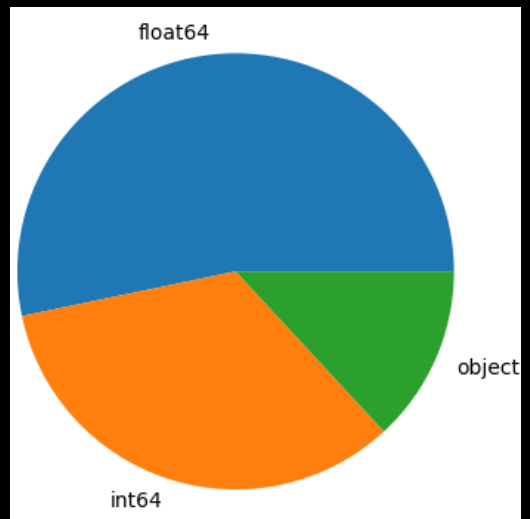
- A Loan lending company has data of different types of client and they need to know more about the clients by just using their data.
- The main aim is to know about the clients category ,who can repay the loan money and know about the clients who are defaulter.
- The company can get in lose either by giving money to defaulter or not giving money to the client who are capable of repaying the money.
- So, the company need every information about the client's category using the data by applying EDA on it.

FILES OF DATASET

Application data(ap_data)

- Data Shape: 307511 rows ,122 columns
- Data types: Float, Integer, Object
- There are null values, some negative values in columns related to days ,days and year are not in proper format

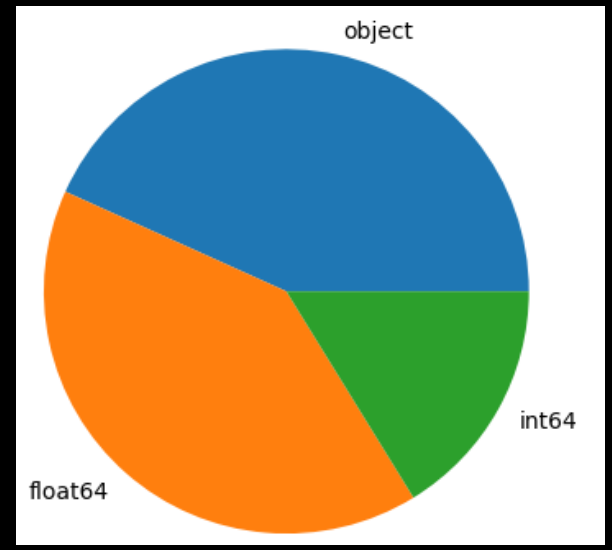
- Float-65%
- Int- 41% ,
- Object-16%



Previous application(ap_prev)

- Data Shape: 1670214 rows , 37columns
- Data types: Integer, Float, Object
- There are null values, some negative values in columns related to days ,days and year are not in proper format.

- Object-43%
- Float64 – 40.5 % ,
- Int64-16.2%

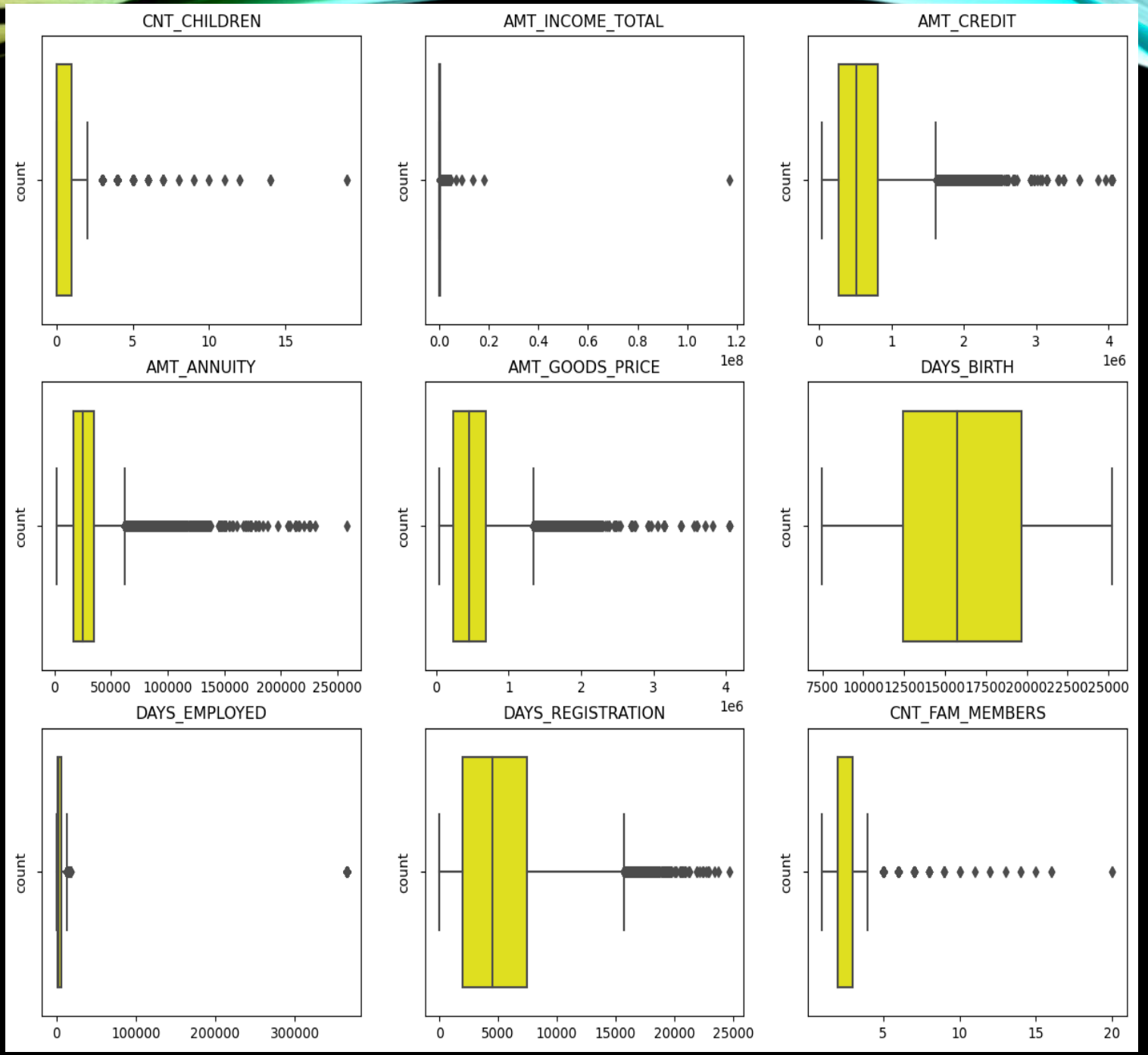


ASSUMPTIONS,APPROACH & METHODOLOGY

- We will try to show only the results, ex-Outliers-showing column's fig which has outliers instead of all fig.
- Dropping null value greater than 40% and 50% depending on the importance of columns.
- Replacing the null value of numerical variable with mode, median etc. and of categorical variables with the other group ex-('other', 'unknown') depends on data
- Binning the values of amount variable in relevant group as they are higher in no.
- Converting negative days values and binning them in a group
- Categorising the columns in the different variable ex- numerical ,categorical & other variables so that analysing process can be easy.

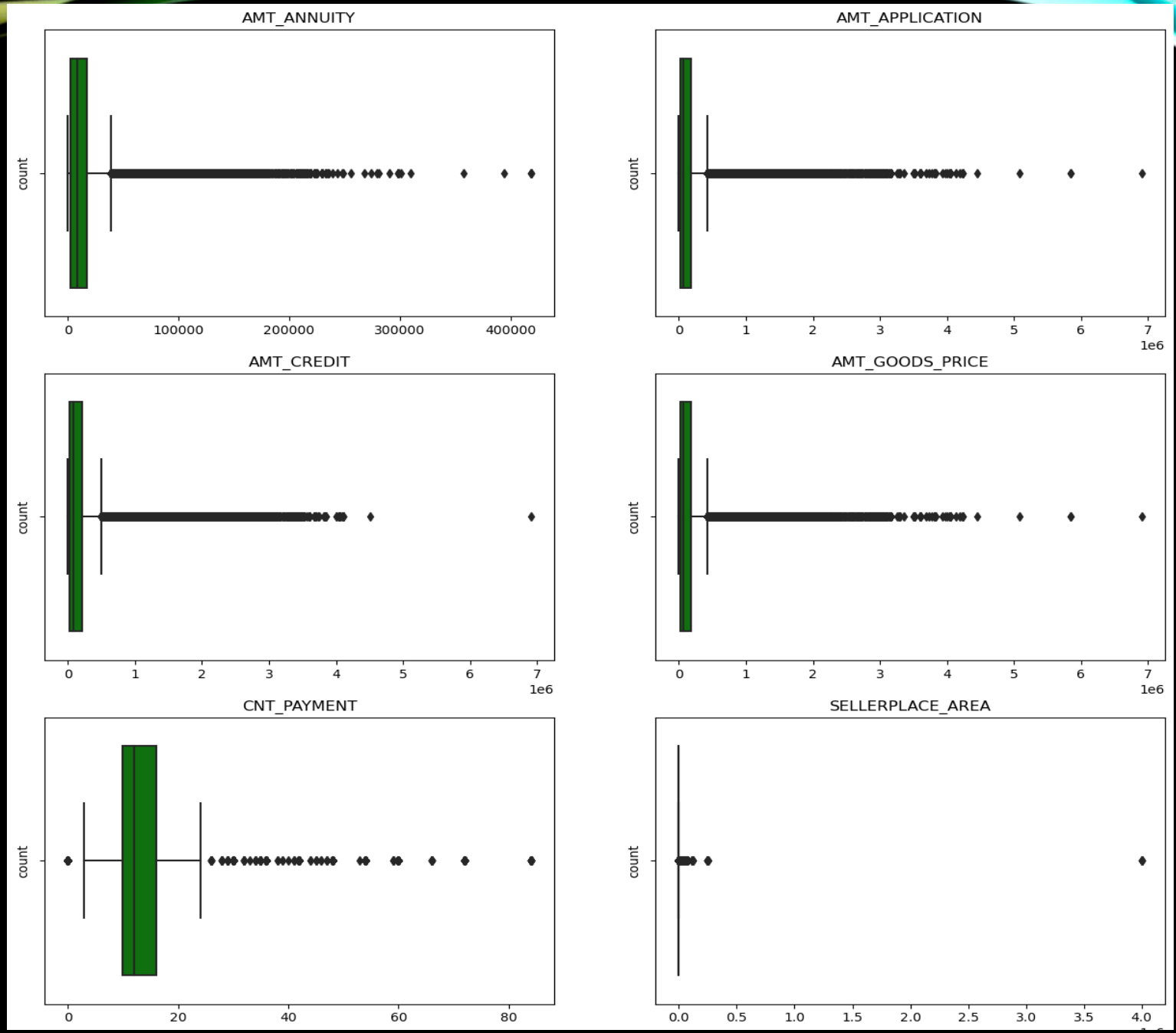
OUTLIERS

- Data set = ap_data
- 'DAYS_EMPLOYED', 'AMT_INCOME_TOTAL' has large amount of outliers, as 'DAYS_EMPLOYED' has highest value - 365243 days which is approx 1000 year (not genuine value) whereas, 'AMT_INCOME_TOTAL ' has 75% value - 20 lakhs and highest value - 11 crores ,The difference is too much between the 75% and the highest value.
- ('AMT_ANNUITY', 'AMT_GOODS_PRICE', ' CMT _FAM_MEMBERS', 'DAY_REGISTRATION', 'AMT_CREDIT','CNT_CHILDREN')has some outliers



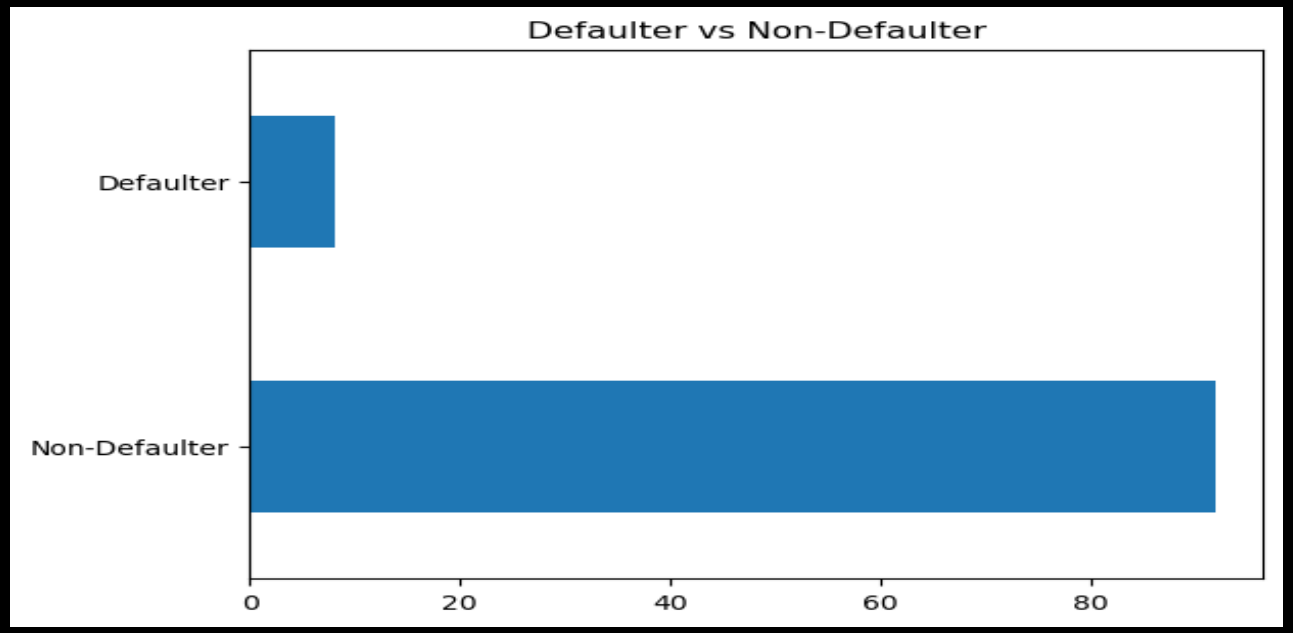
OUTLIERS

- Data set= ap_prev
- ('AMT_CREDIT','AMT_GOODS_PRICE','SELLERPLACE_AREA') has large amount of outliers
- ('CNT_PAYMENT','AMT_ANNUITY','AMT_APPLICATION') has less outliers



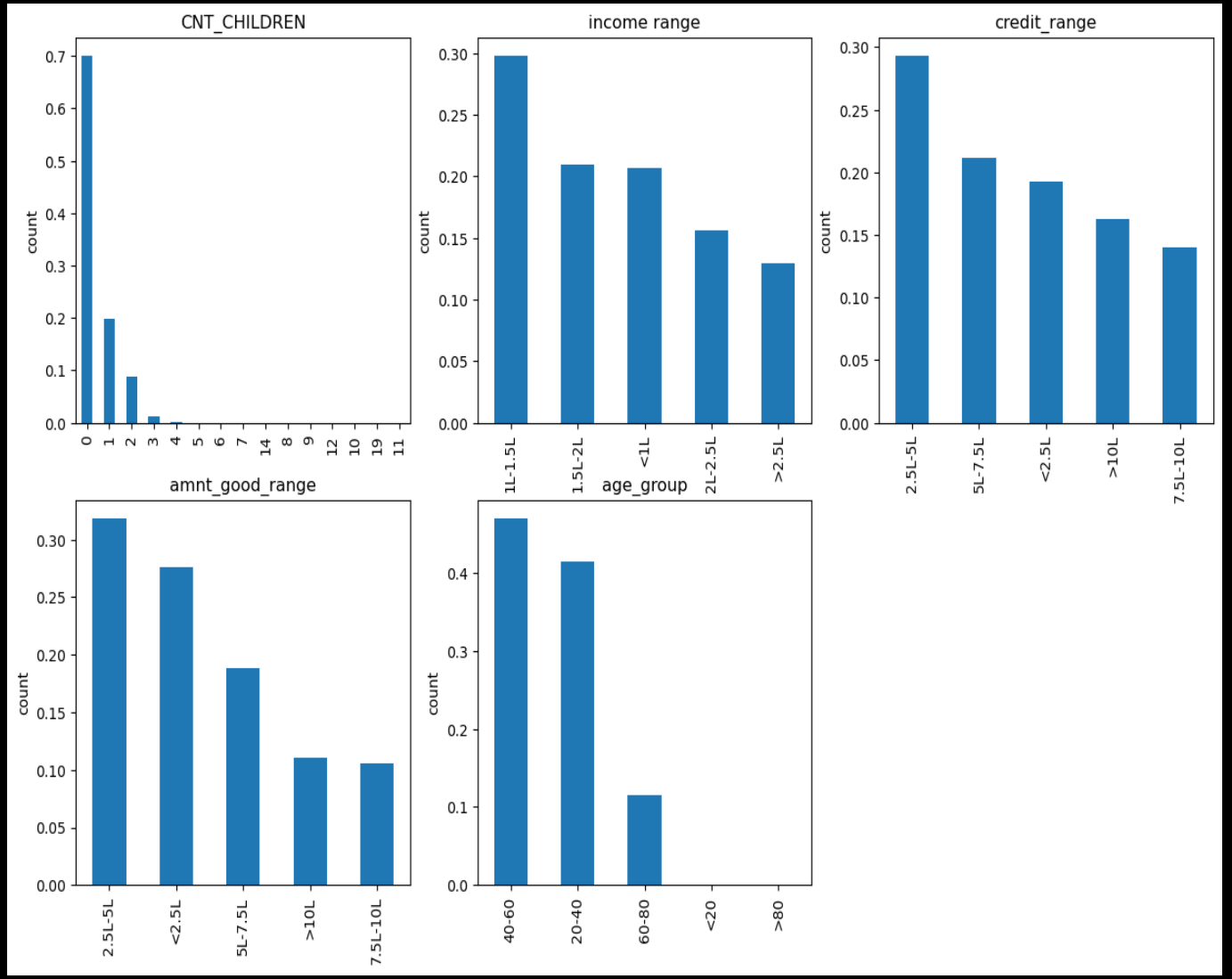
ANALYSIS OF DATA

- **Data Imbalance** : TARGET variable data imbalance
- Defaulters = client with late payment, approx. 8% are there.
- Non-Defaulter = who has no late payment, approx. 91% are there
- The ratio of the data imbalance is approx. 11.38%.



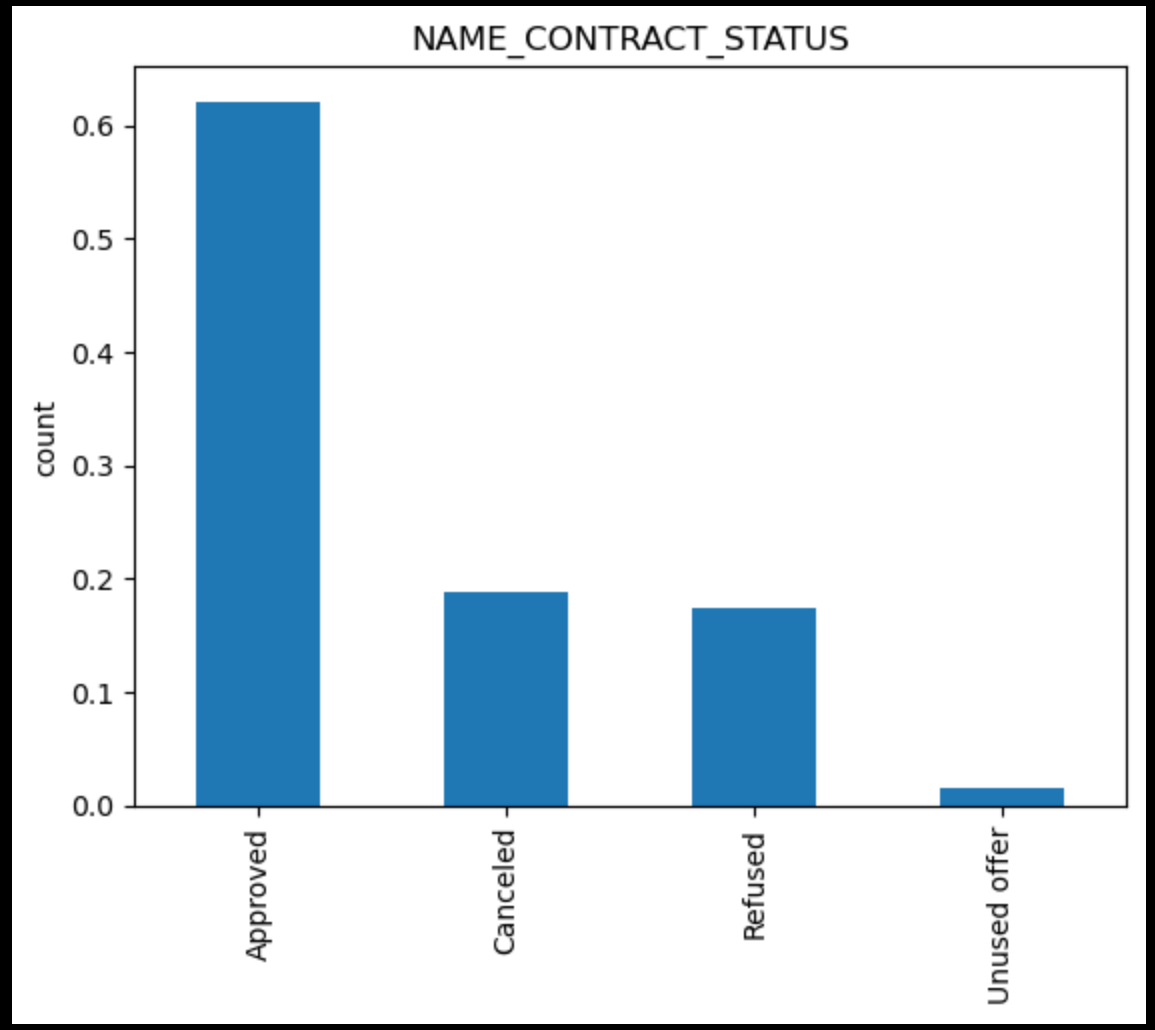
➢ UNIVARIATE ANALYSIS:

- No of children maximum no of customer have are 0 then 1.
- Max income of the customer comes in the range of 1L-1.5L then 1.5L-2L.
- maximum credit range comes in the range of 2.5L-5L then 5L-7.5L.
- Maximum amount of goods coming in range of 2.5l-5L.
- Maximum customer of age group 40-60, then 20-40 were into the application.



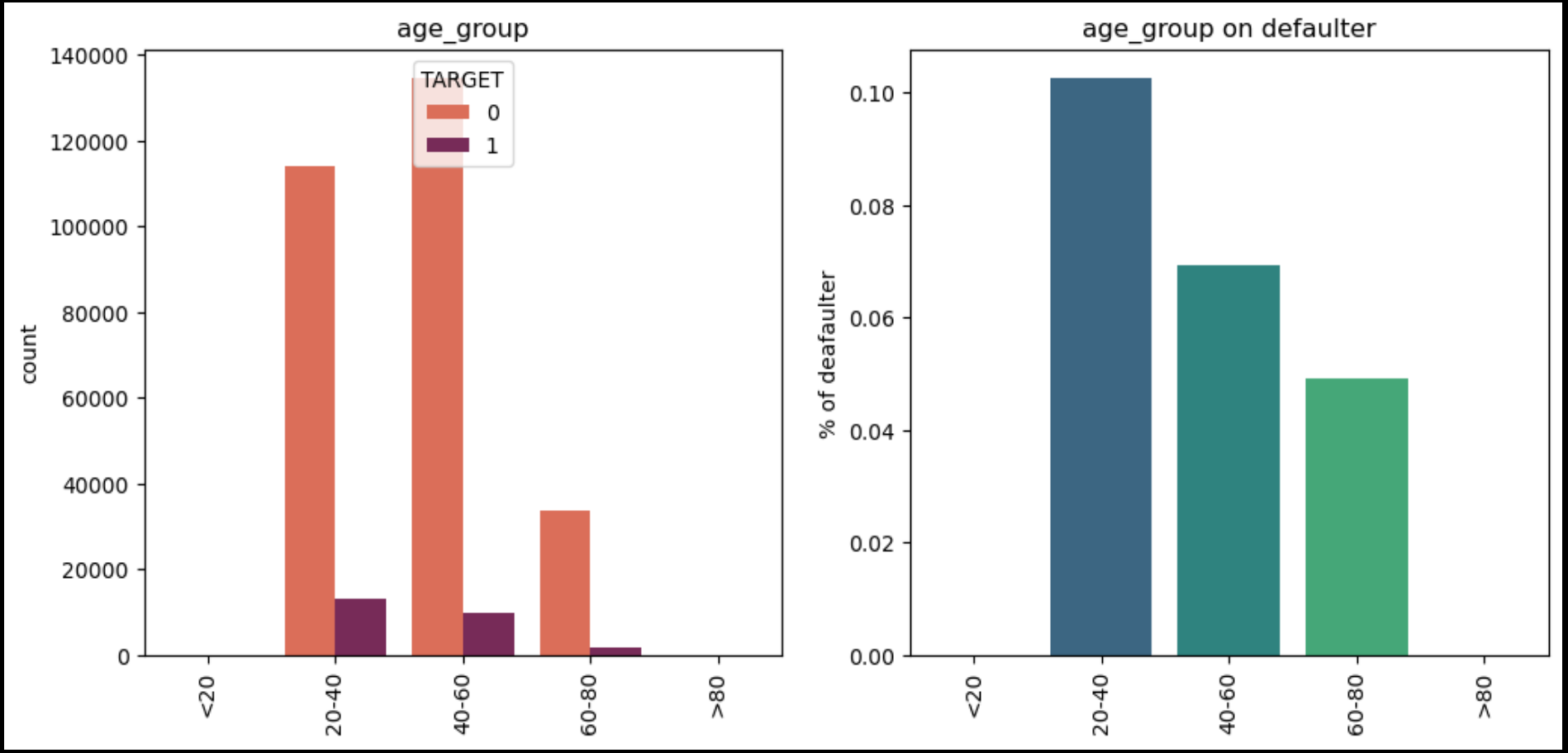
➢ UNIVARIATE ANALYSIS(NAME_CONTRACT_STATUS)

- This category has maximum no of approved clients approx. 60%
- Approx. 20% is cancelled and refused



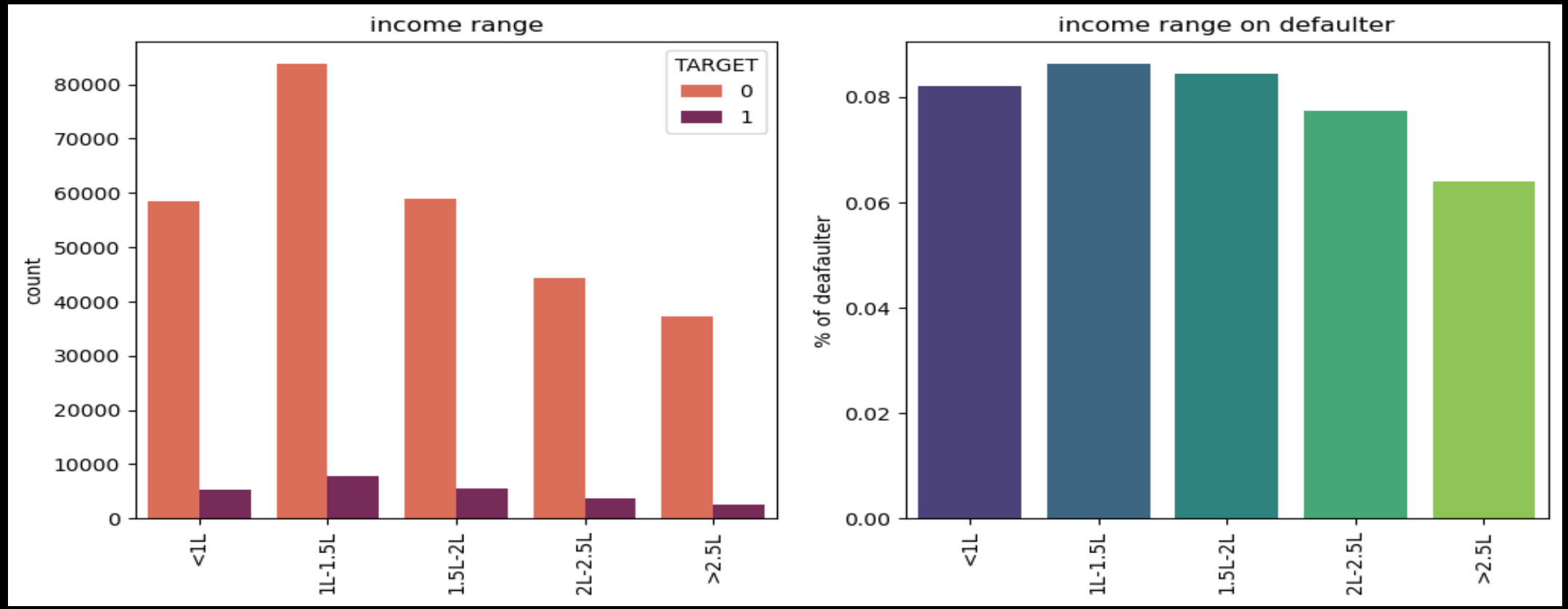
❖ age_group vs TARGET

- The age group of 40-60 are maximum in data who are non defaulter.
- The age group 20-40 has maximum defaulter approx. 85%



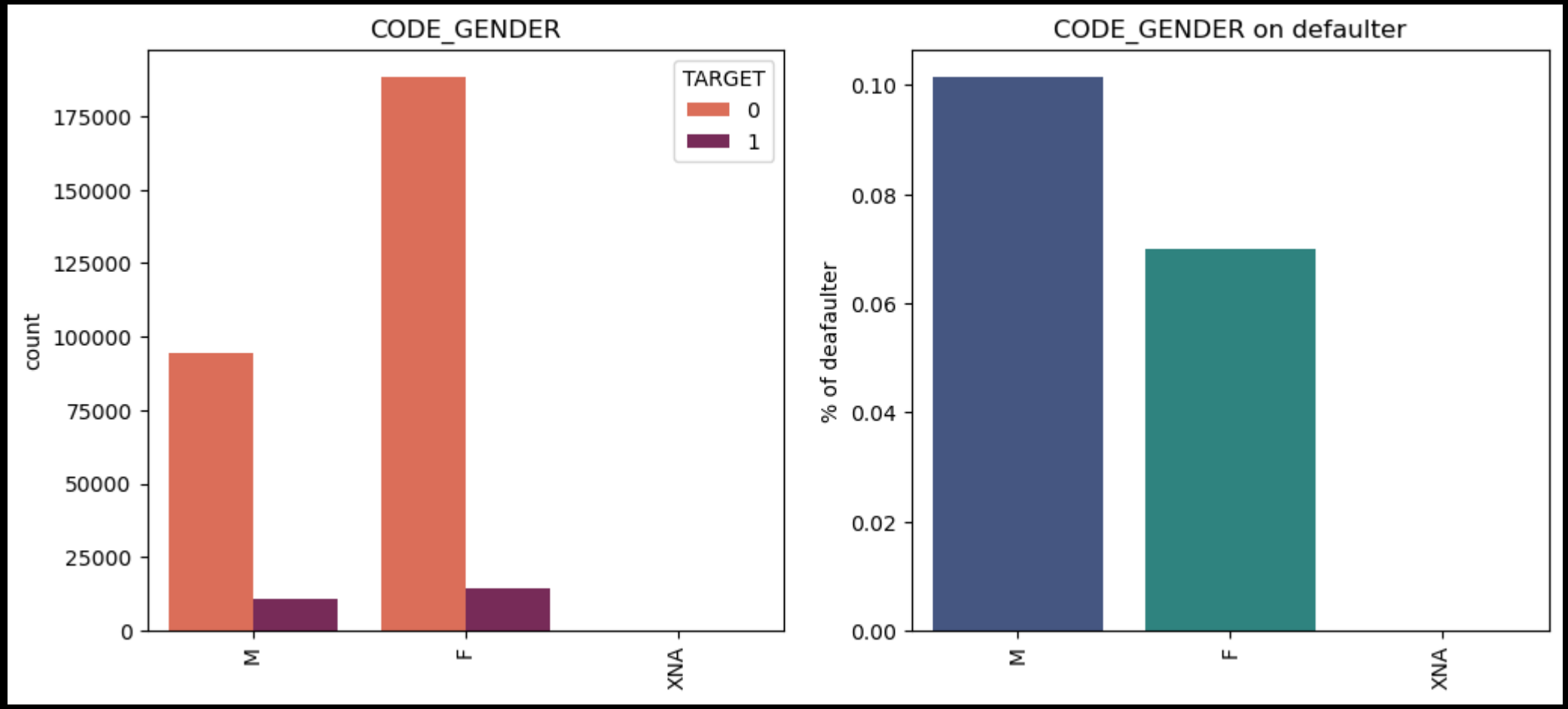
❖ INCOME RANGE VS TARGET

- The income range of 1L -1.5L has maximum non defaulter
- The income range of >2.5 has least no of defaulter
- It somewhere shows that greater the income range lesser the no of defaulters.



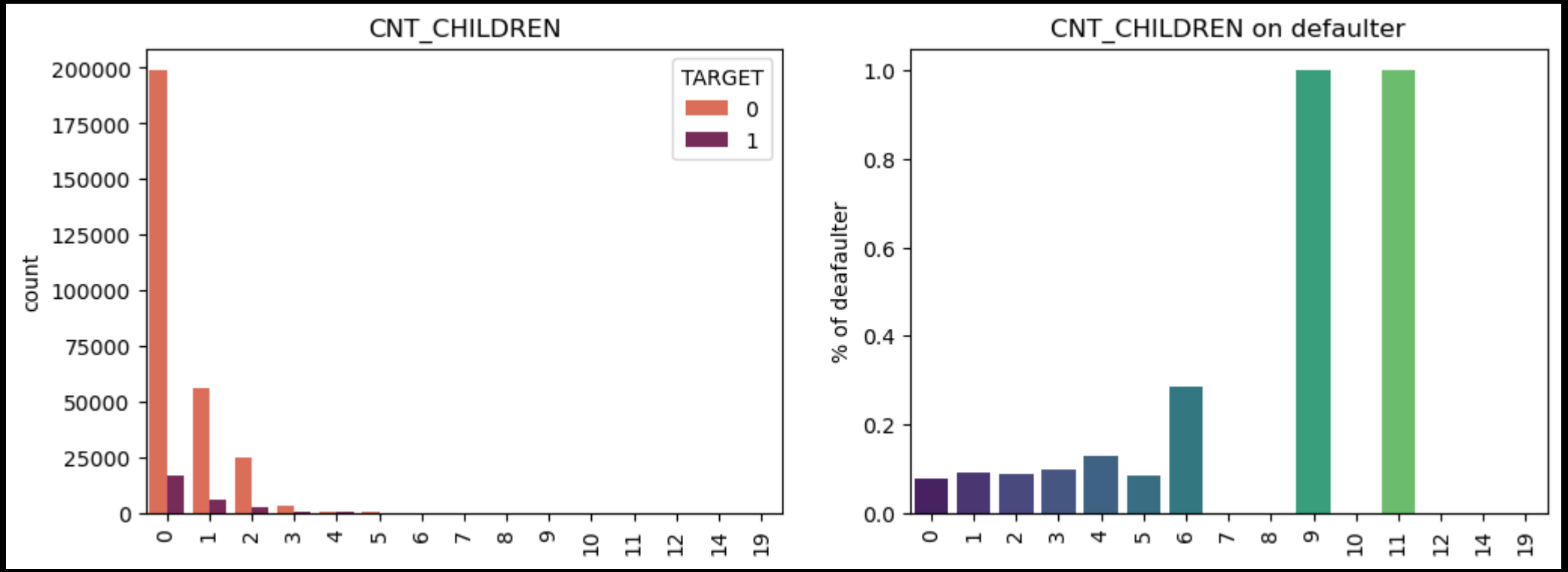
❖ CODE_GENDER VS TARGET

- The defaulters are maximum in Males then Female



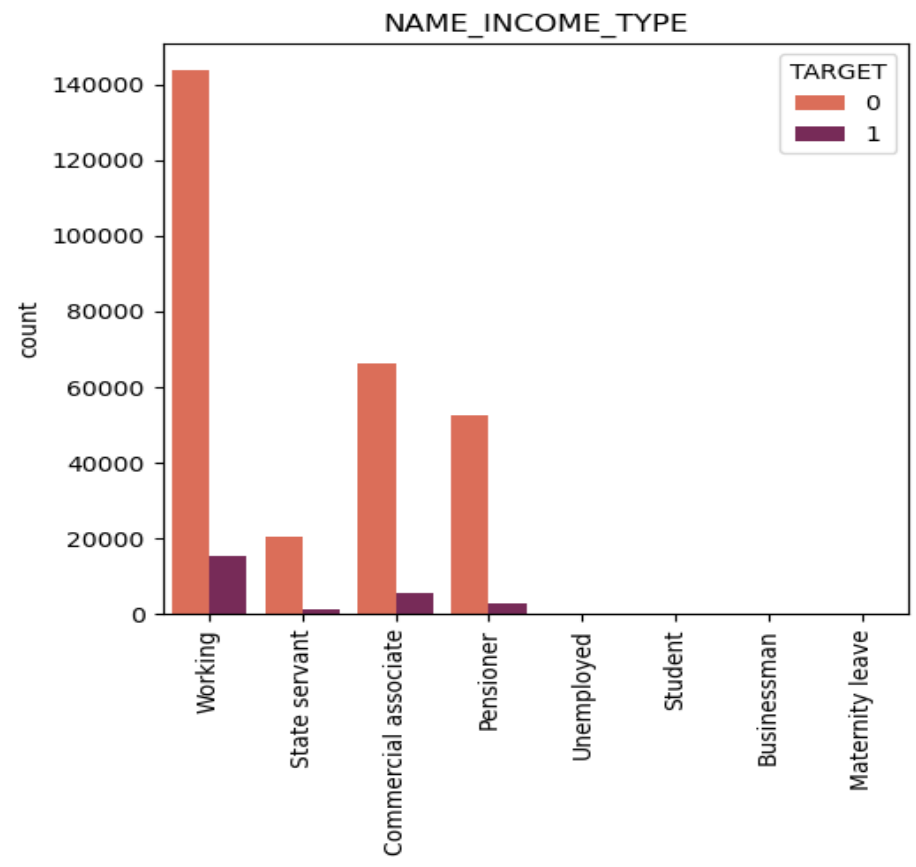
❖ CNT_CHILDREN VS TARGET

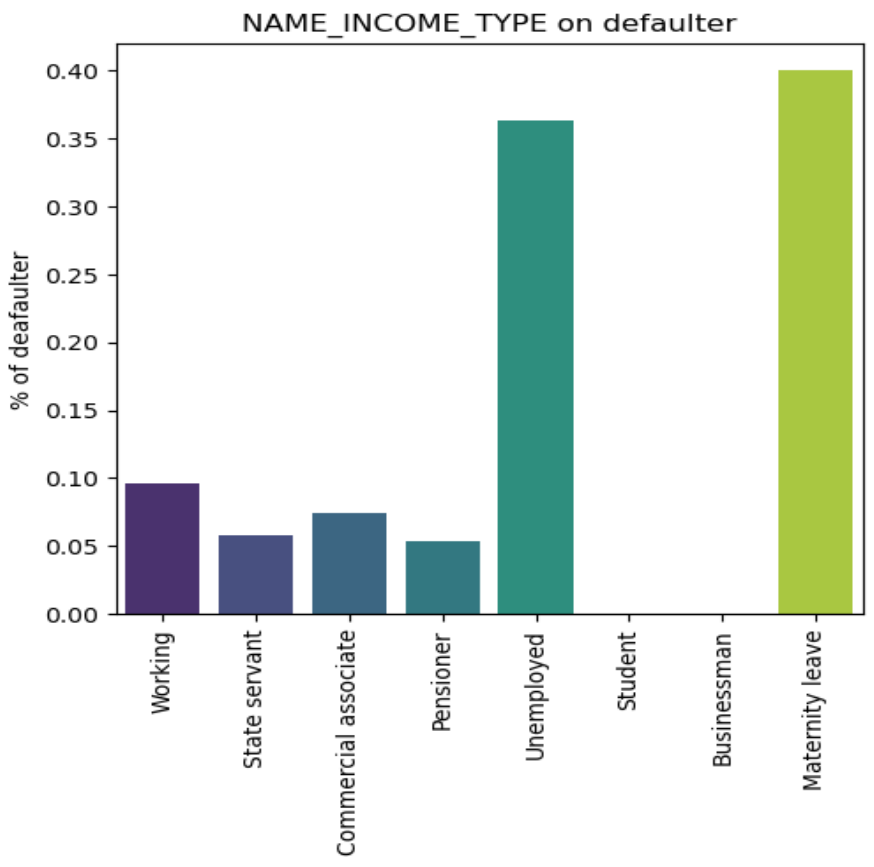
- The client has no child at the time of application.
- Maximum defaulter has no of children for around 9-10



❖ NAME_INCOME_TYPE VS TARGET

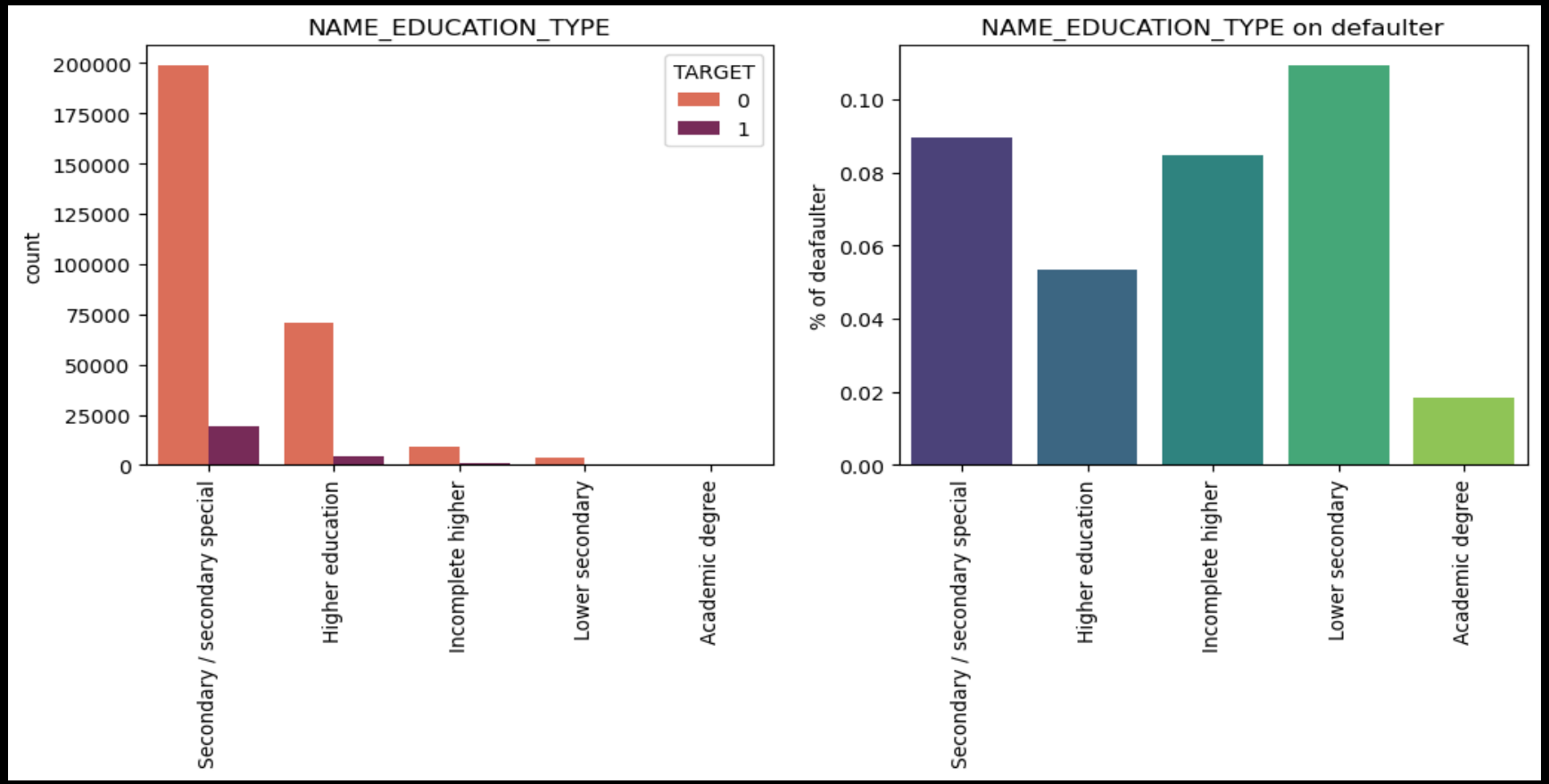
- Working are more interested in the application form as compare to others.
- Unemployed & clients have Maternity leave has the maximum no of defaulters.





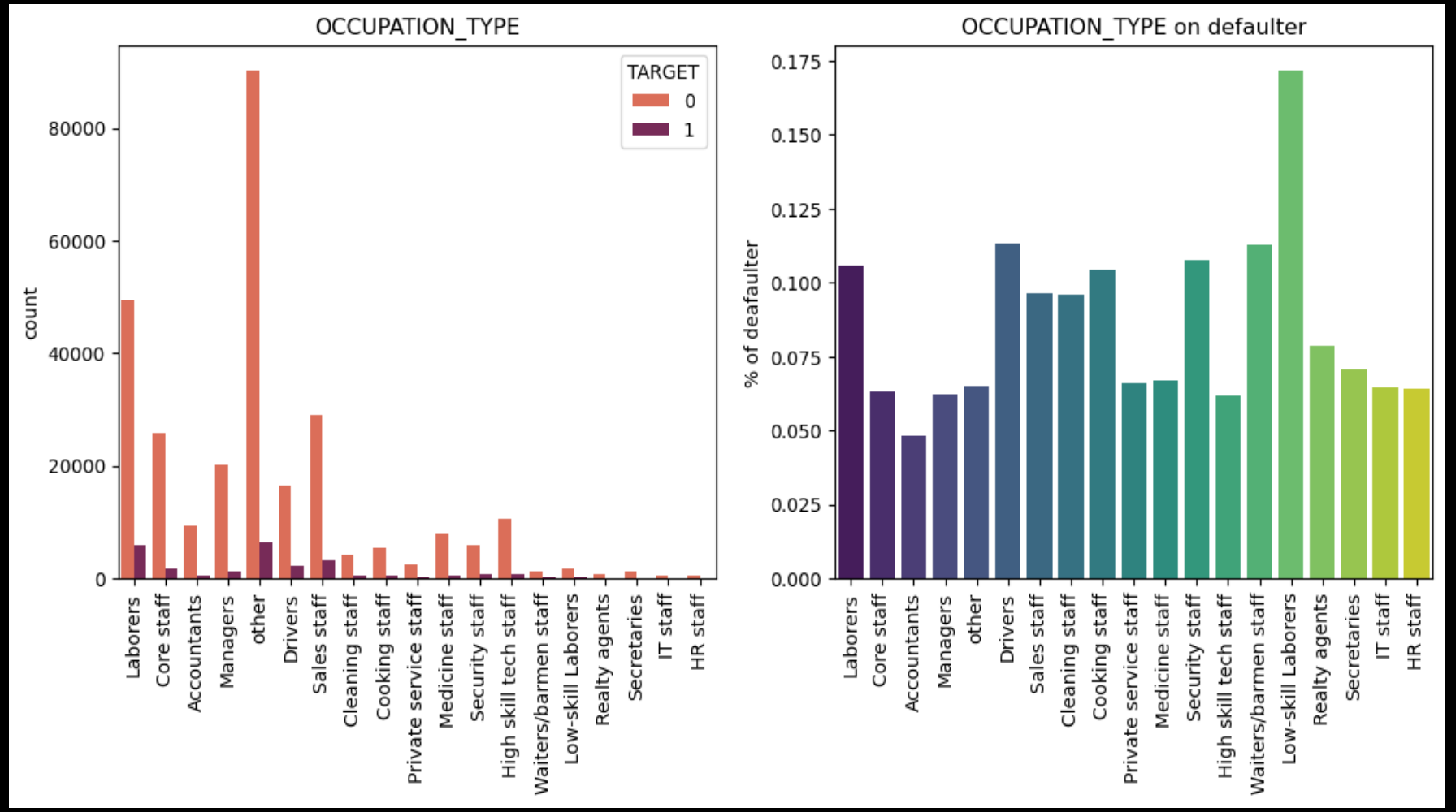
❖ NAME_EDUCATION_TYPE VS TARGET

- Secondary special are maximum at the time of application.
- Lower secondary has the maximum no of defaulters and client having Academic degree are the least no of defaulters.



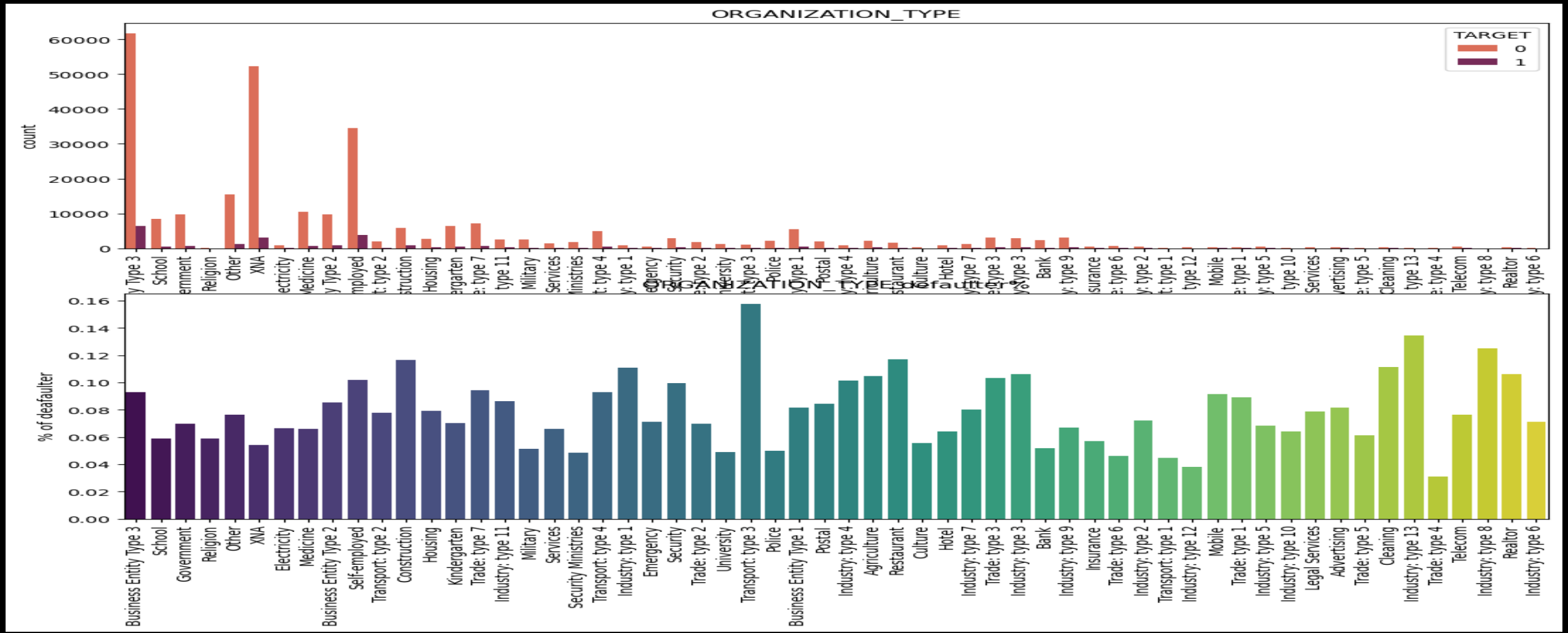
❖ OCCUPATION_TYPE VS TARGET

- Laborers are maximum in non defaulters
- Low skilled laborers has maximum no of defaulters and Accountant has minimum no of defaulters.



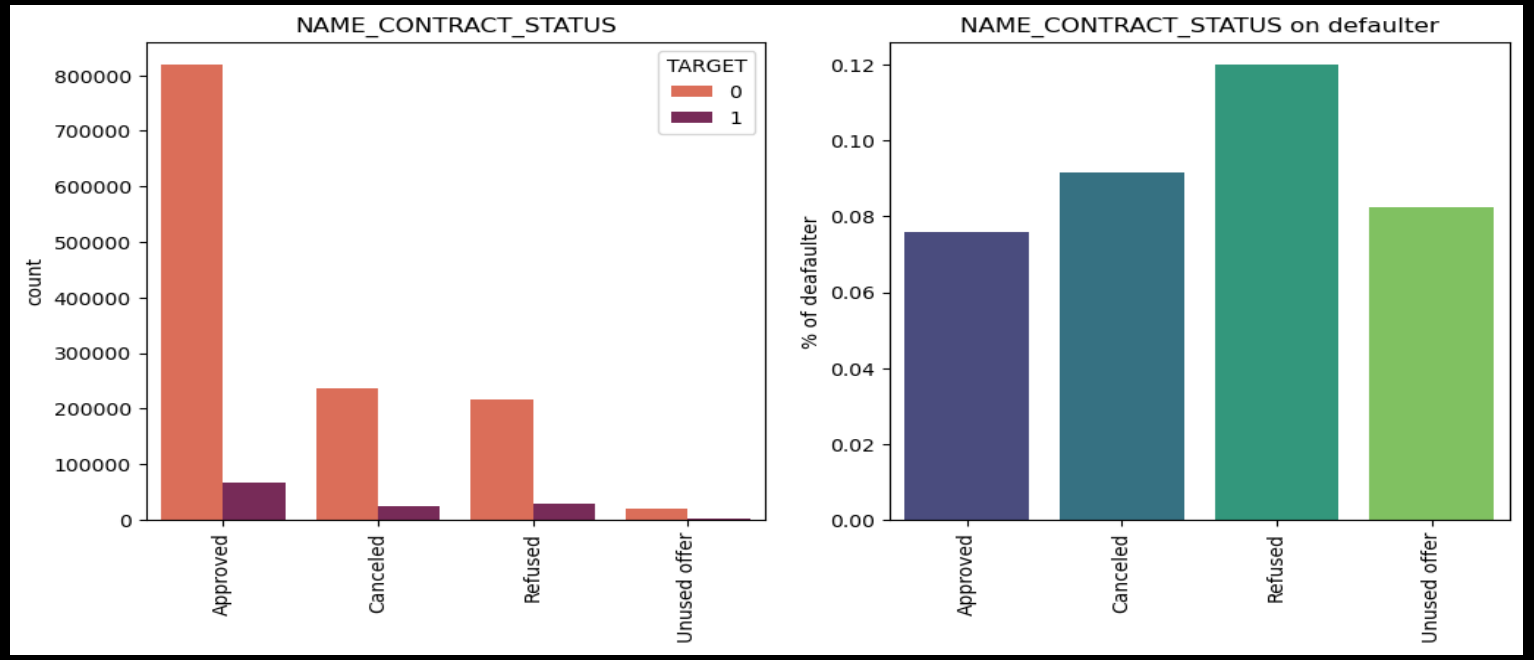
❖ ORGANIZATION_TYPE VS TARGET

- We can say that business entity 3 has maximum no of non defaulter.
- Transport type 3 then industry type 3 then construction then self employed has maximum no of defaulters.



➢ NAME_CONTRACT_STATUS VS TARGET

- The no of approved clients are maximum in the dataset.
- Clients whose application got Refused has maximum no of defaulters
- The Approved clients has approx. 92.4% of non defaulters and approx. 7.5% of defaulters
- The Cancelled clients has approx. 90.8% of non defaulters and approx. 9.1% of defaulters, so before canceling there should analysis of those 90.8% non defaulters
- The Refused clients has approx. 88% of non defaulters and 11.9% of defaulters

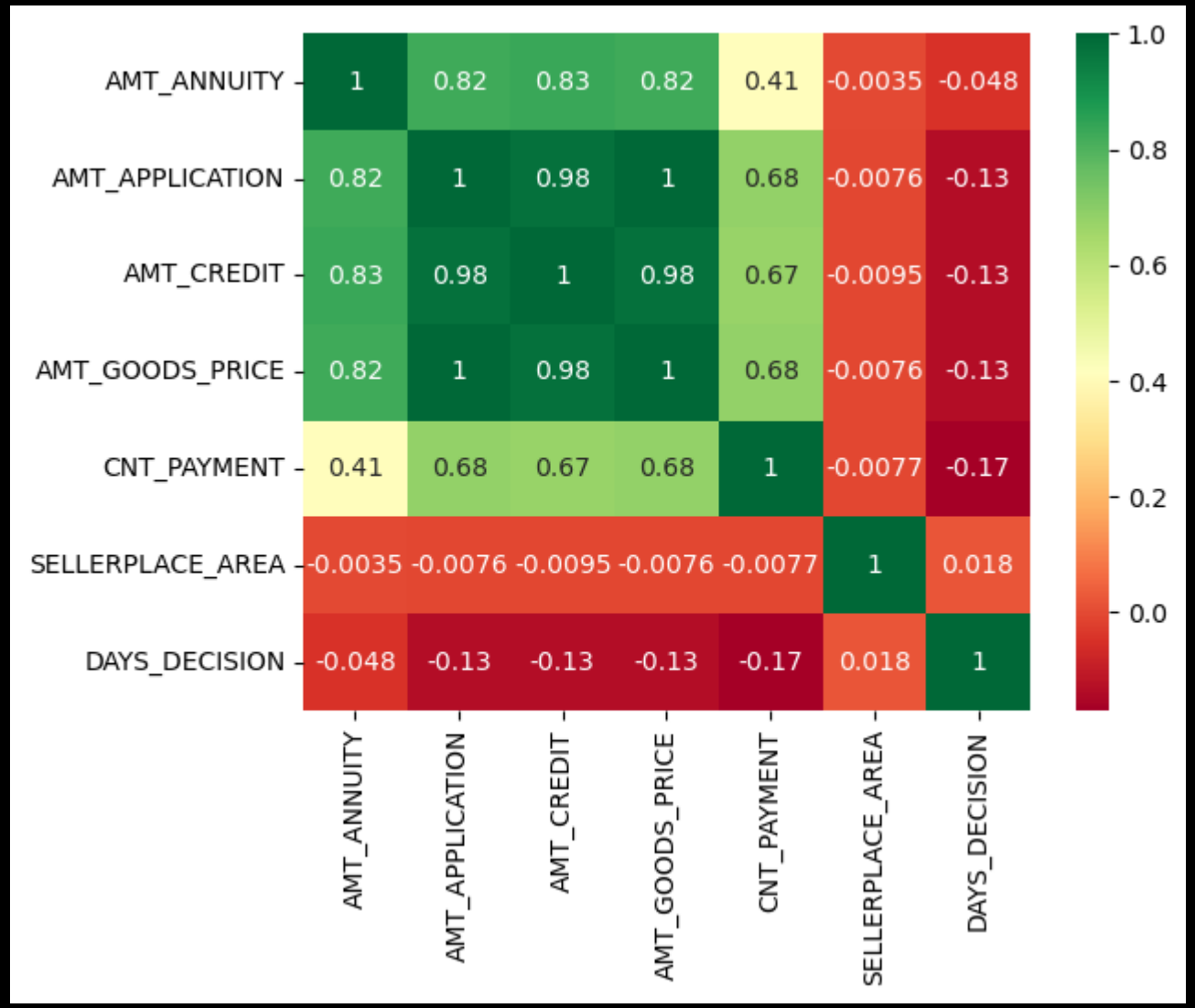


NAME_CONTRACT_STATUS	TARGET	
Approved	0	92.411345
	1	7.588655
Canceled	0	90.826431
	1	9.173569
Refused	0	88.003586
	1	11.996414
Unused offer	0	91.748276
	1	8.251724

Name: TARGET, dtype: float64

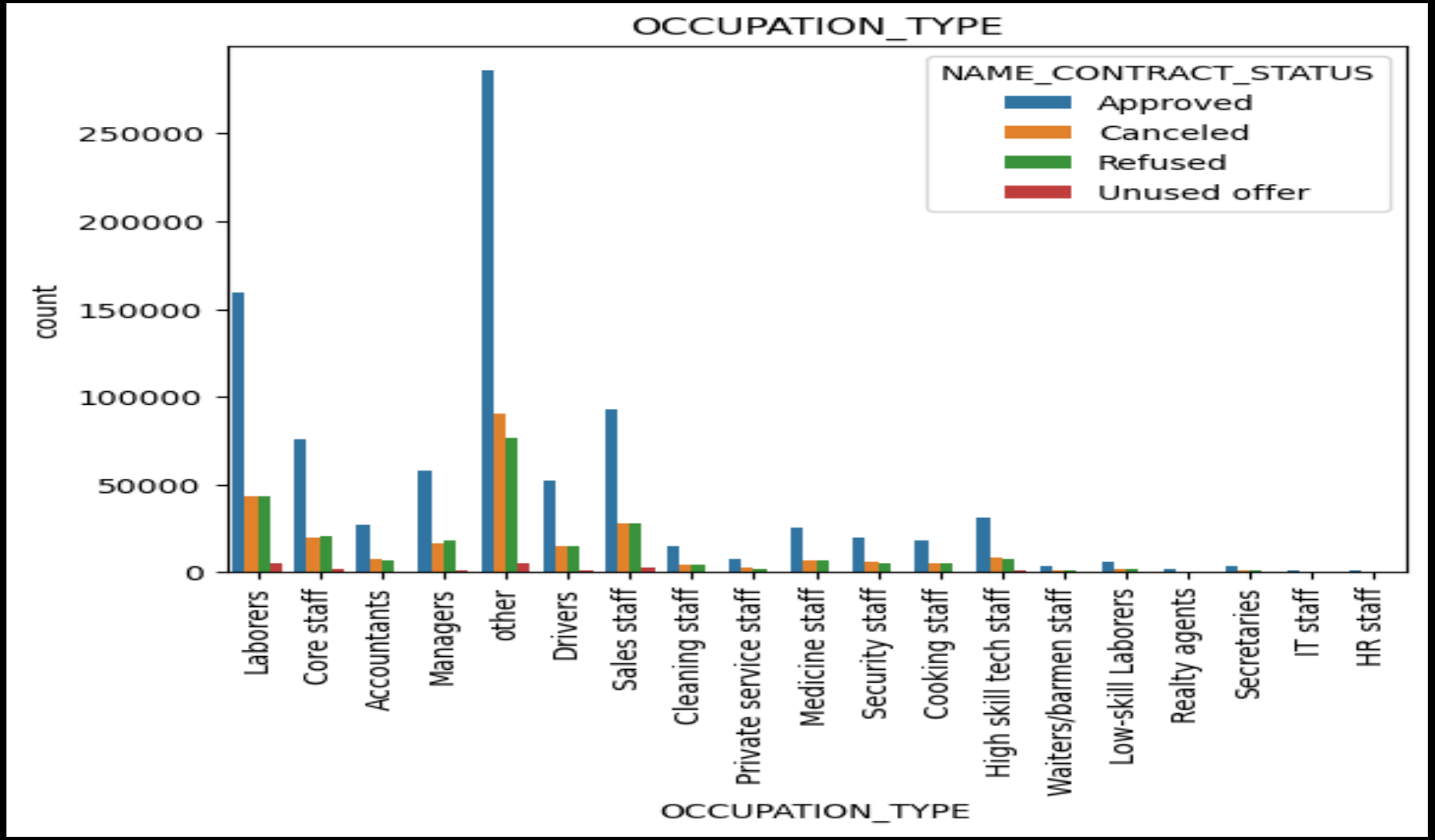
➢ CORRELATION BETWEEN NUMERICAL VARIABLES

- There is very high correlation between 'AMT_CREDIT' and 'AMT_GOODS_PRICE'.
- There is very high correlation between 'AMT_CREDIT' and 'AMT_APPLICATION'.
- Very less correlation between 'CNT_PAYMENT' and 'DAYS_DECISION'



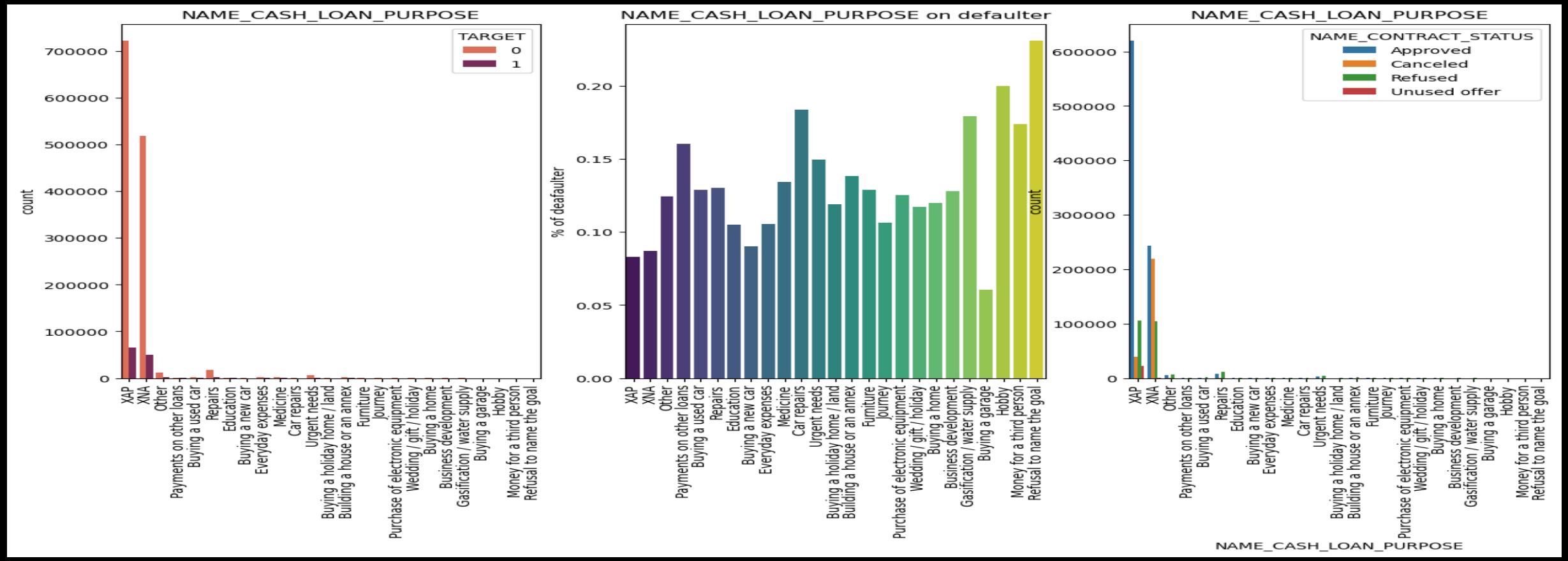
❖ OCCUPATION_TYPE VS NAME_CONTRACT_STATUS

- After the missing values Laborers are the maximum in Approved, Cancelled and Refused category



❖ NAME_CASH_LOAN_PURPOSE VS TARGET

- Most of the values are missing, the remaining highest is Repairs category which has highest non defaulters
- Repairs and Others category has the highest no of application approval.
- Clients having loan purpose of 'Hobby' then 'Car Repairs' then 'water supply work' then 'payments on other loans' has the maximum no of defaulters.



RECOMMENDATIONS & CONCLUSIONS

- Age of clients 40-60 has high no of clients who Repay, Should avoid the 20-40 age group.
- Client's Income range greater than 2.5L can repay the loan, avoid the income range 1 - 1.5L.
- Females are less defaulter comparing with Males
- Clients with 1-2 children can repay the loan, children more than 6 increases the rate of defaulter
- Academic degree Education type has less no of defaulter where as secondary special has high no of defaulters.
- Need to avoid low skill laborers then Waiters/ Barmen staff as they have high no of defaulters.
- Business Entity 3 has high no. of clients that Repay whereas Transport type 3,Industry type 13,Realtor can have defaulters.
- Client taking loan for hobbies can be defaulter whereas client buying garage item has minimum no of defaulters.
- Student can be non defaulter
- Unemployed clients can be defaulter.

THANK YOU