EDA ASSIGNMENT

AGENDA

Problem Statement

Assumption

Approach &Methodology

Graphs & insights

Recommendation or Conclusion

PROBLEM STATEMENT

Introduction

This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Understanding

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
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.

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
All other cases: All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):
Approved: The Company has approved loan Application
Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want.
Refused: The company had rejected the loan (because the client does not meet their requirements etc.).
Unused offer: Loan has been cancelled by the client but at different stages of the process.
In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency to default.

ASSUMPTIONS & DESCRIPTION:

- According to my assumptions and according to column description in my dataset Application data there is lots of flag columns and all are not important and useful for my analysis so I am taking only email, contact columns. I am not dropping those columns but I am taking only important columns.

- Generally when you taking loan or giving loan so there is some important things we consider mostly like what is income, what type of job he is doing, student or working, what is family status, how many member, he is able to pay his previous loan or not, financially how much he is stable, background.

- So according to take care of above things and with the help of columns description I choose my columns for analysis which is on my jupyter notebook and some graphs are in ppt.

- After that I gave some conclusion.

APPROACH & METHODOLOGY

Application data:

- there are 307511 rows and 122 columns they have datatype like object , int, float.

- there columns having negative, positive values which includes days so I am using absolute function there
- There 49 columns which have missing value 40% and After dropping we have left with 73 columns
- Now I have 61 numerical and 12 categorical column.
- Now I have 18 columns which have null values more than zero

Filling missing values:

- AMT_REQ_CREDIT_BUREAU_YEAR","AMT_REQ_CREDIT_BUREAU_QRT","AMT_REQ_CREDIT_BUREAU_MON", "AMT_REQ_CREDIT_BUREAU_WEEK","AMT_REQ_CREDIT_BUREAU_DAY", "AMT_REQ_CREDIT_BUREAU_HOUR“
- These above columns represent discrete and not continuous and for numerical column we use mean or median to fill null value and it is not normally distributed so I used median.
- OBS_30_CNT_SOCIAL_CIRCLE ,DEF_30_CNT_SOCIAL_CIRCLE ,OBS_60_CNT_SOCIAL_CIRCLE ,DEF_60_CNT_SOCIAL_CIRCLE DAYS_LAST_PHONE_CHANGE ,CNT_FAM_MEMBERS ,AMT_ANNUITY ,AMT_GOODS_PRICE ,NAME_TYPE_SUITE
- These above columns have very less missing value so need to impute
- EXT_SOURCE_3 have missing value which is filling by mean() I plot histogram and I got it normally distributed so I used mean()

- In plot you can see second most occupation type is laborer , since there 31% value missing in the column filling those 31% by laborer is not make any sense. so just replace "Nan" value with "Unknown“

Identifying outliers:

from describe we could find all the columns those wo have high difference between max and 75 percentile and important columns which are useful for analysis

- AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.
- AMT_INCOME_TOTAL have large number of outliers it means that income of anyone is very high.
- DAYS_BIRTH has no outliers
- DAYS_EMPLOYED has outlier values around 350000 days which is impossible.

Previous Dataset:

- There are 1670214 rows and 37 columns which int ,float, and object datatype.
- I have 11 columns which have missing value above 40% and I drop all those columns and left with 26 columns now.
- There is 5 columns which have missing values but 2 columns have very less missing so I impute only three columns

- There is 11 numerical and 15 categorical column

Filling Missing values:

- AMT_ANNUITY I used describe() and check my min , max value on the basis of I check my mean median value but I was not clear on that method so I plot histogram and then I saw it is not normally distributed and skewed from right side so mean() is not good for filling missing values there may chances of outliers also so I used median().

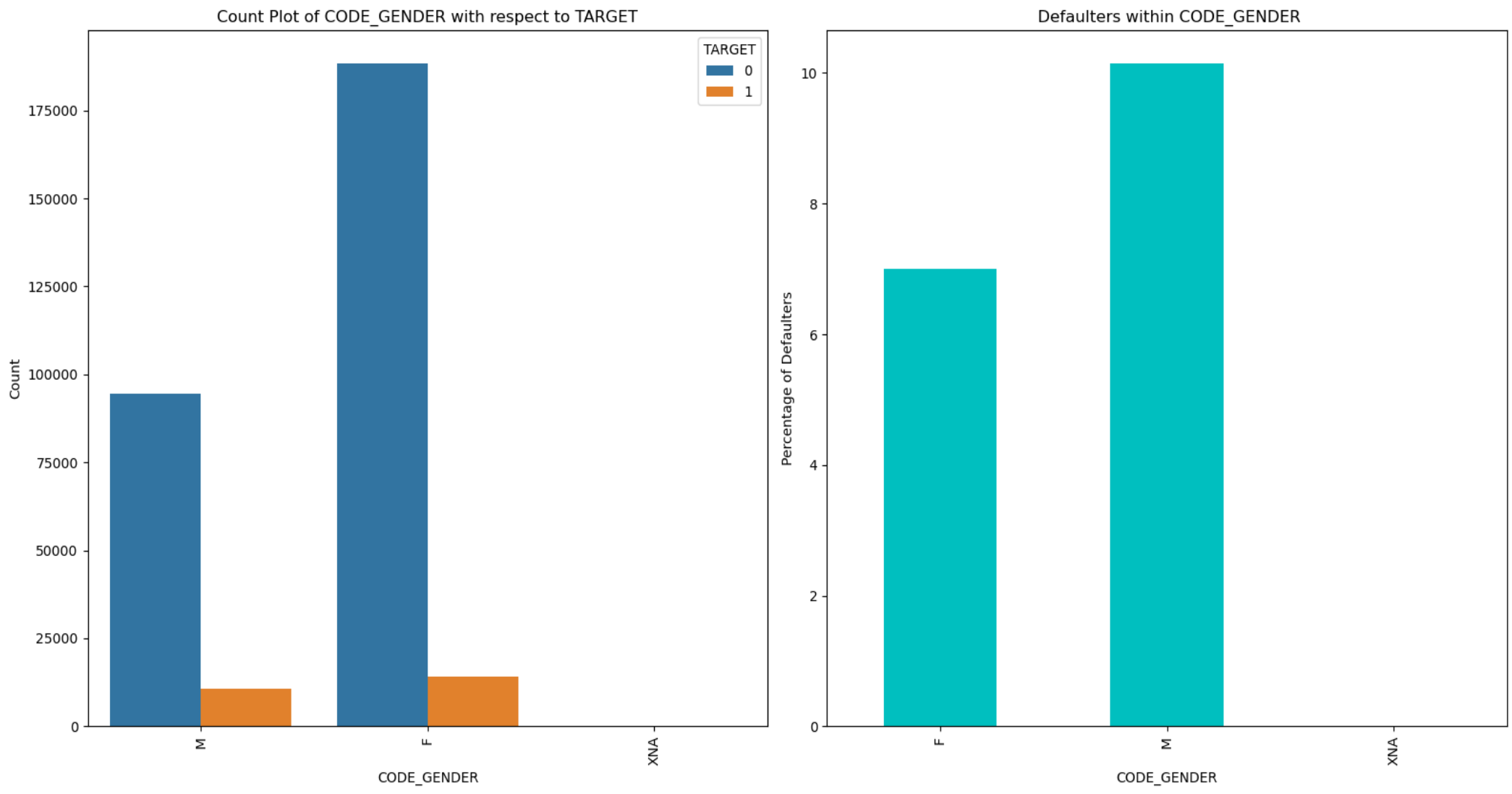
- AMT_GOODS_PRICE same for this column also.

Identifying Outliers:

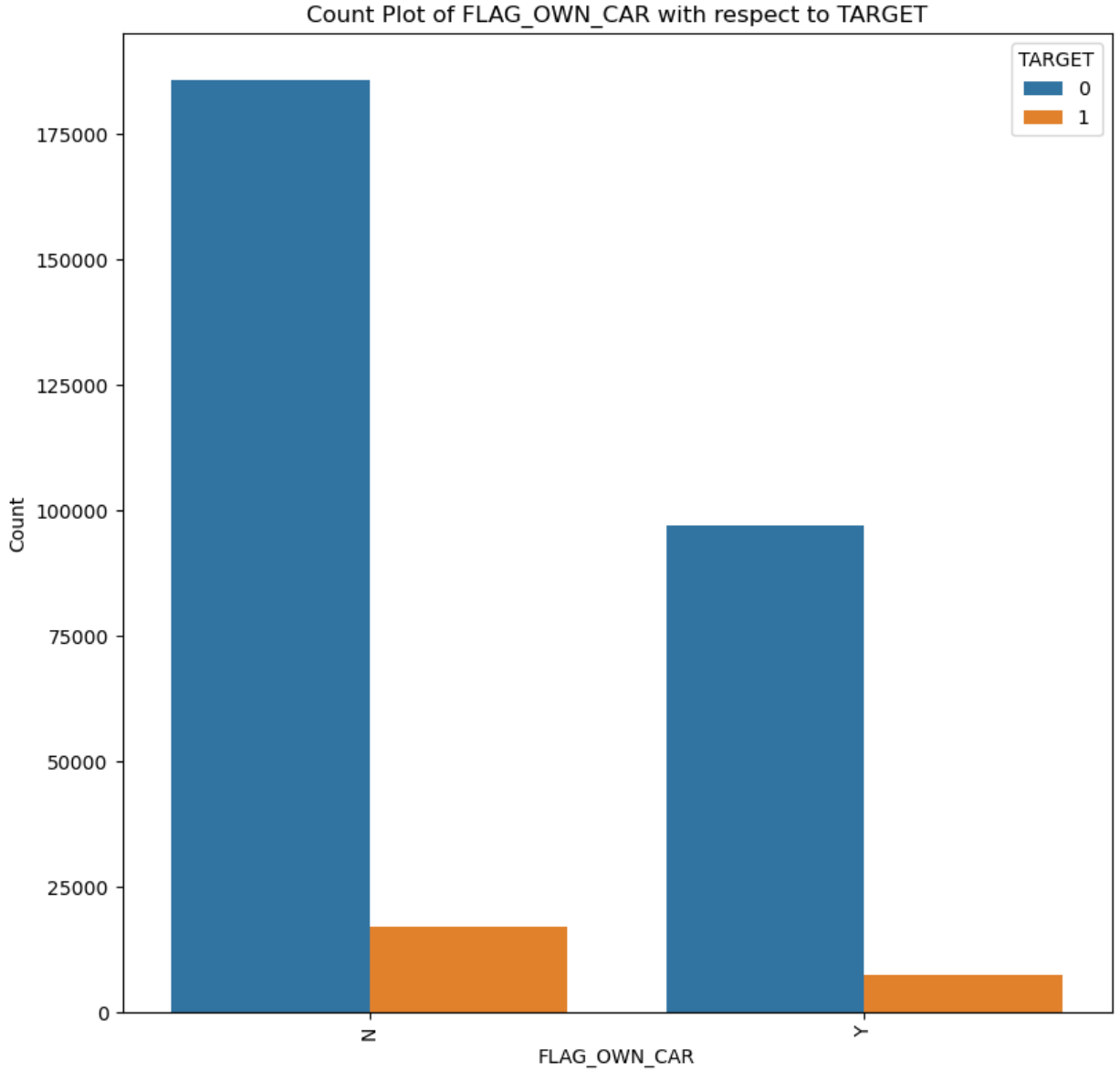
- we could find all the columns those wo have high difference between max and 75 percentile and important column which is used for analysis.

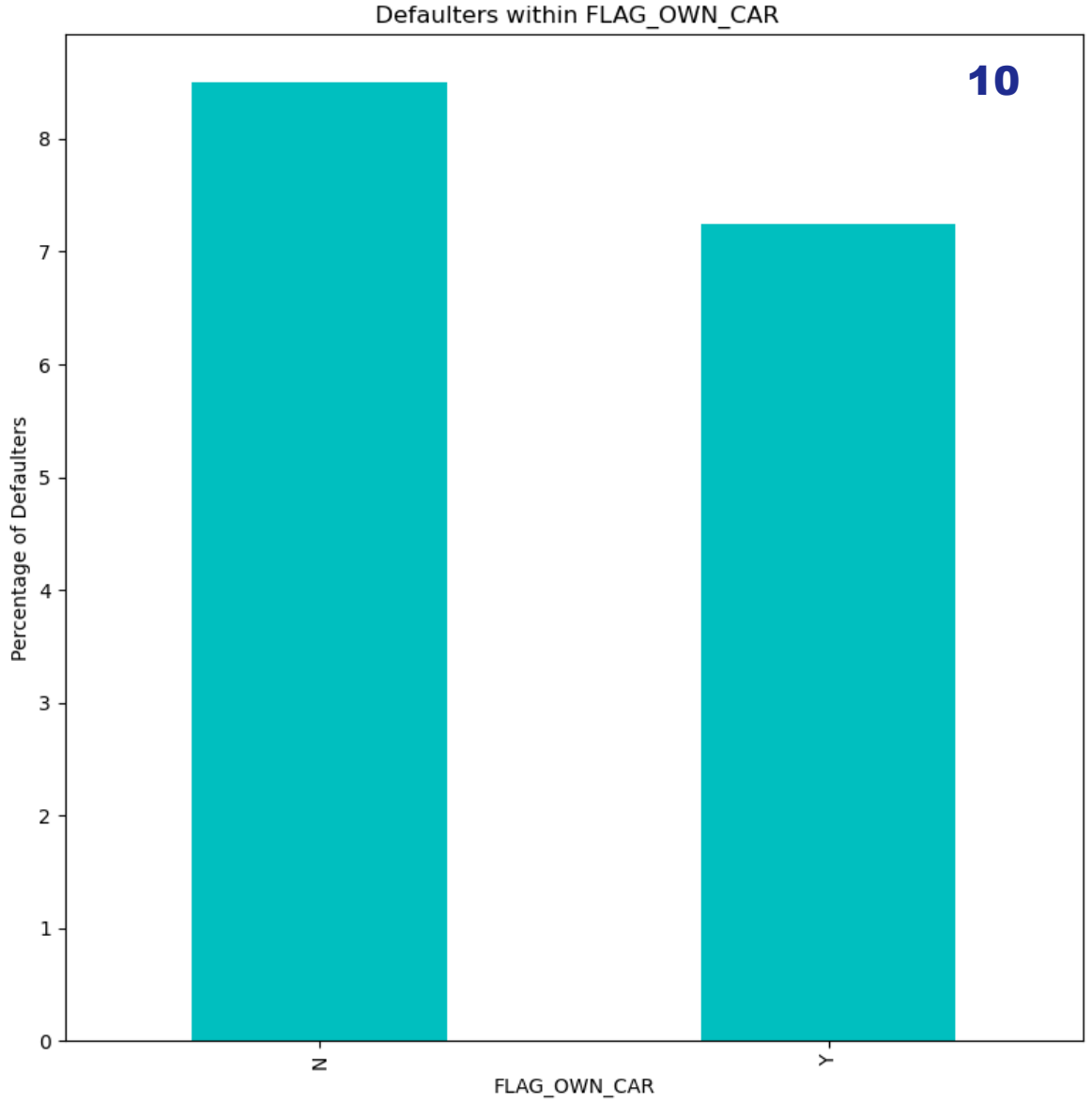
  AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have large number of outliers.
- CNT_PAYMENT has few outlier values.
- DAYS_DECISION has little number of outliers

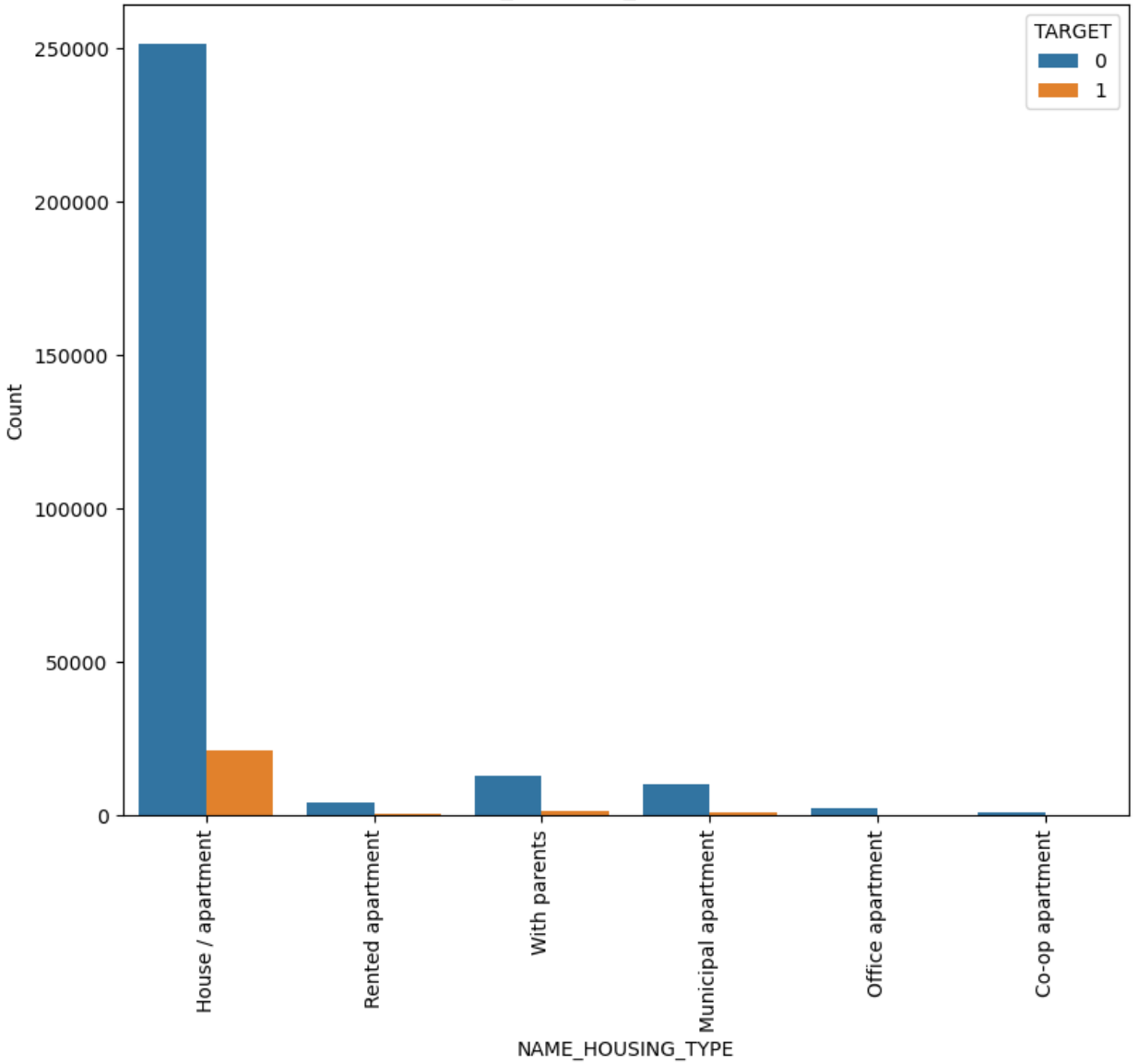
Univariate analysis : Categorical column

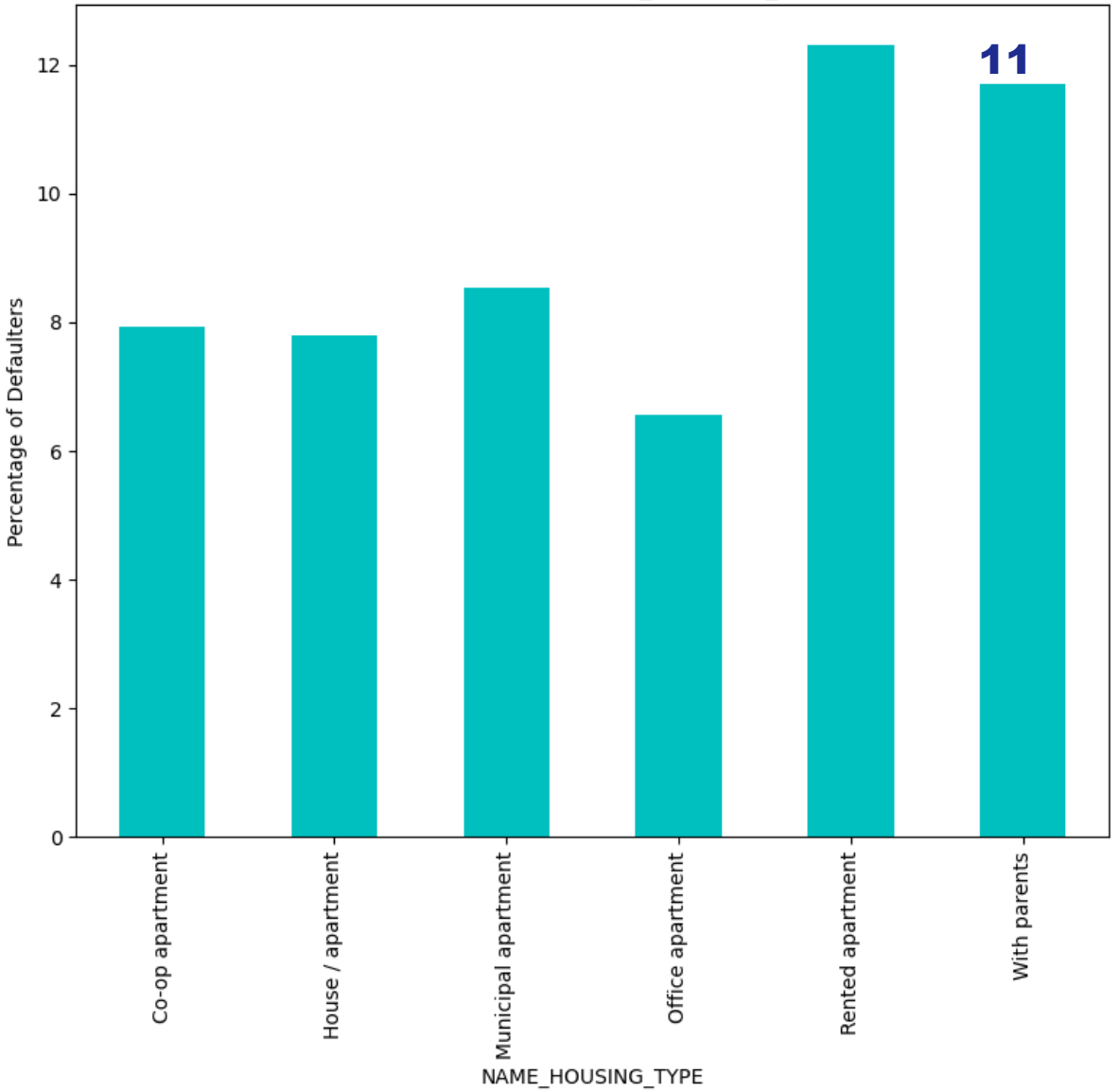


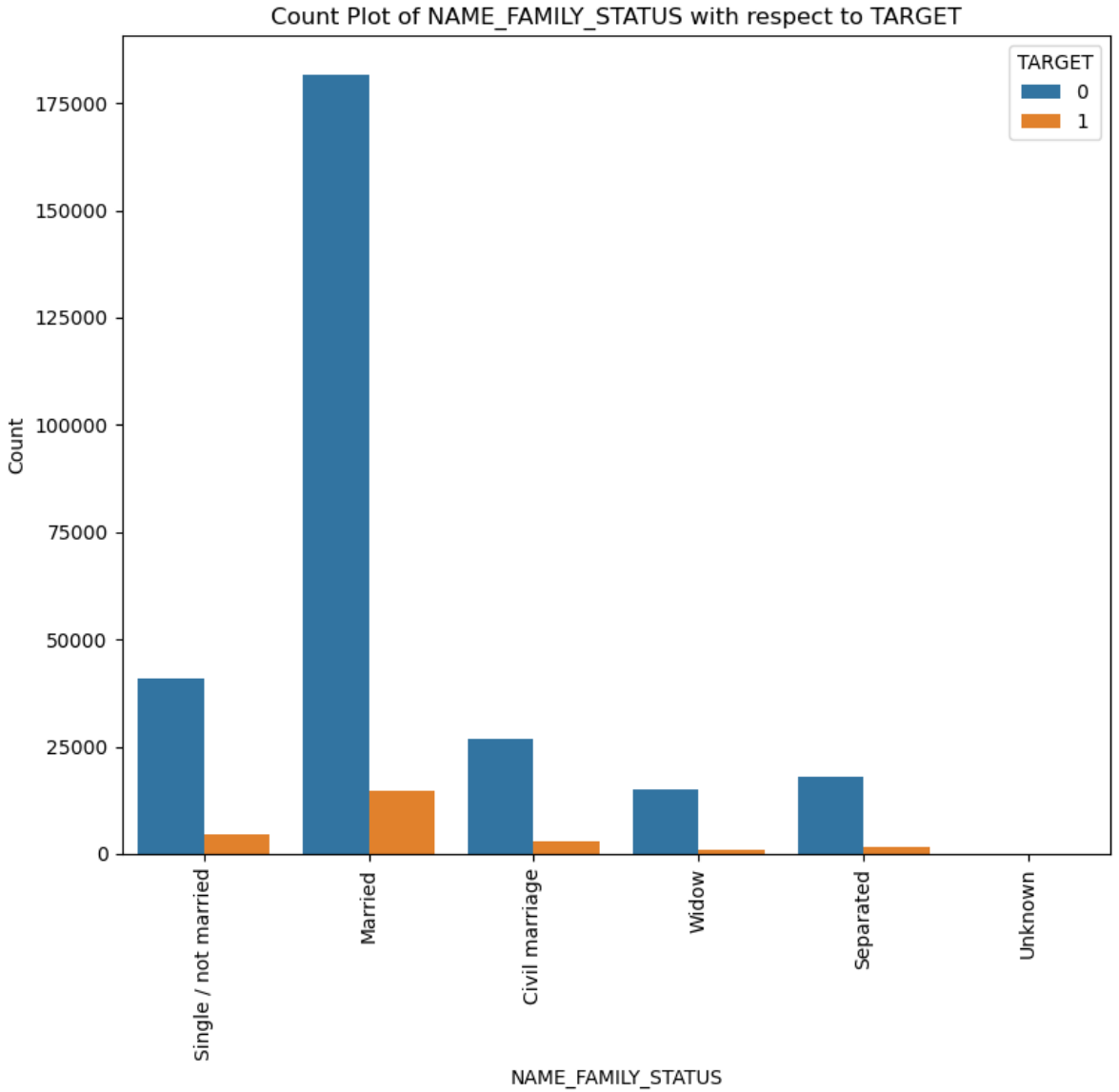
CODE_GENDER: According to my graph as compare to male default rate is higher than female

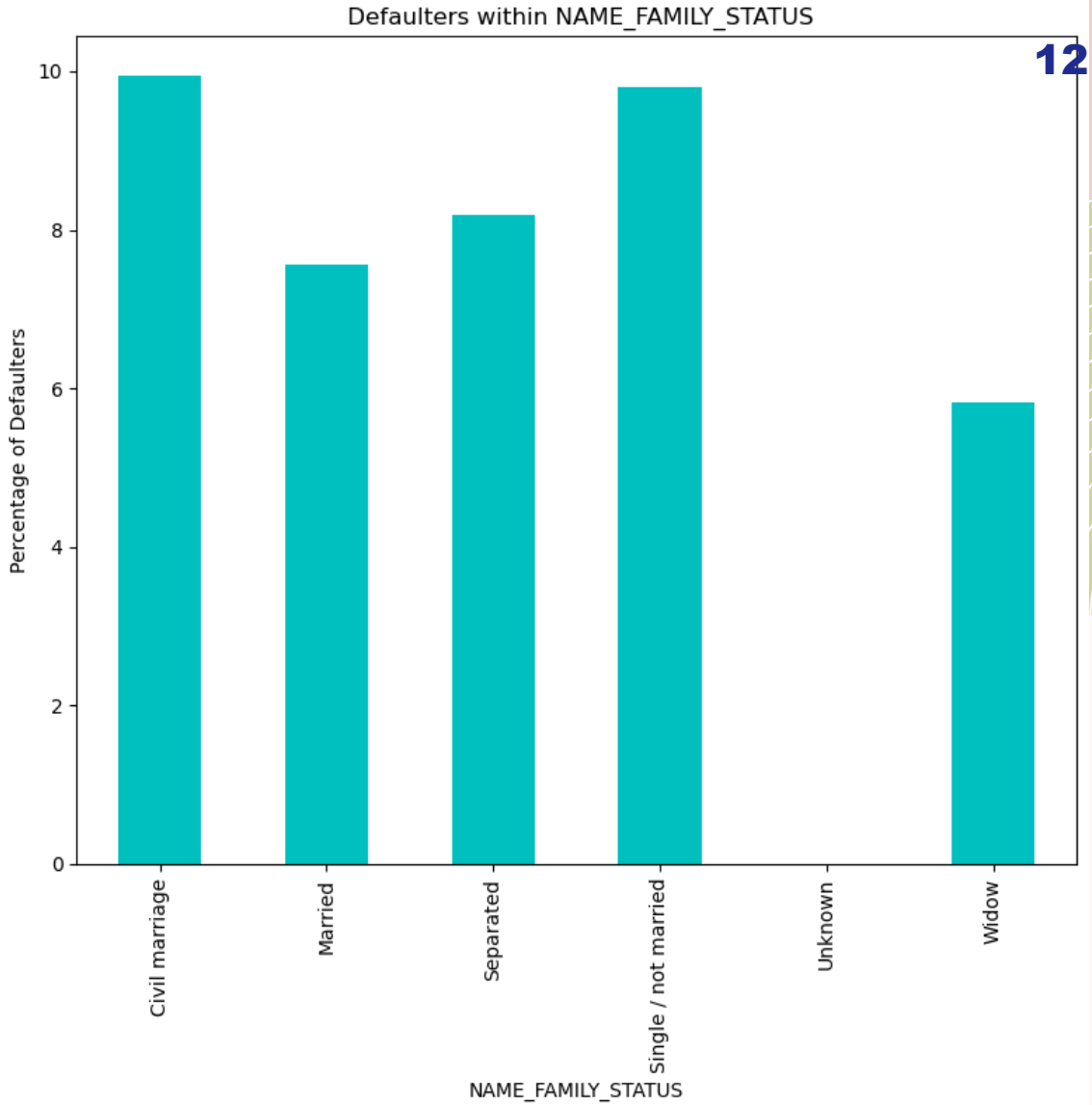

Count Plot of FLAG_OWN_CAR with respect to TARGET
TARGET
0
1
Count
175000
150000
125000
100000
75000
50000
25000
0
N
Y
FLAG_OWN_CAR


Defaulters within FLAG_OWN_CAR
Percentage of Defaulters
8
7
6
5
4
3
2
1
0
N
Y
FLAG_OWN_CAR

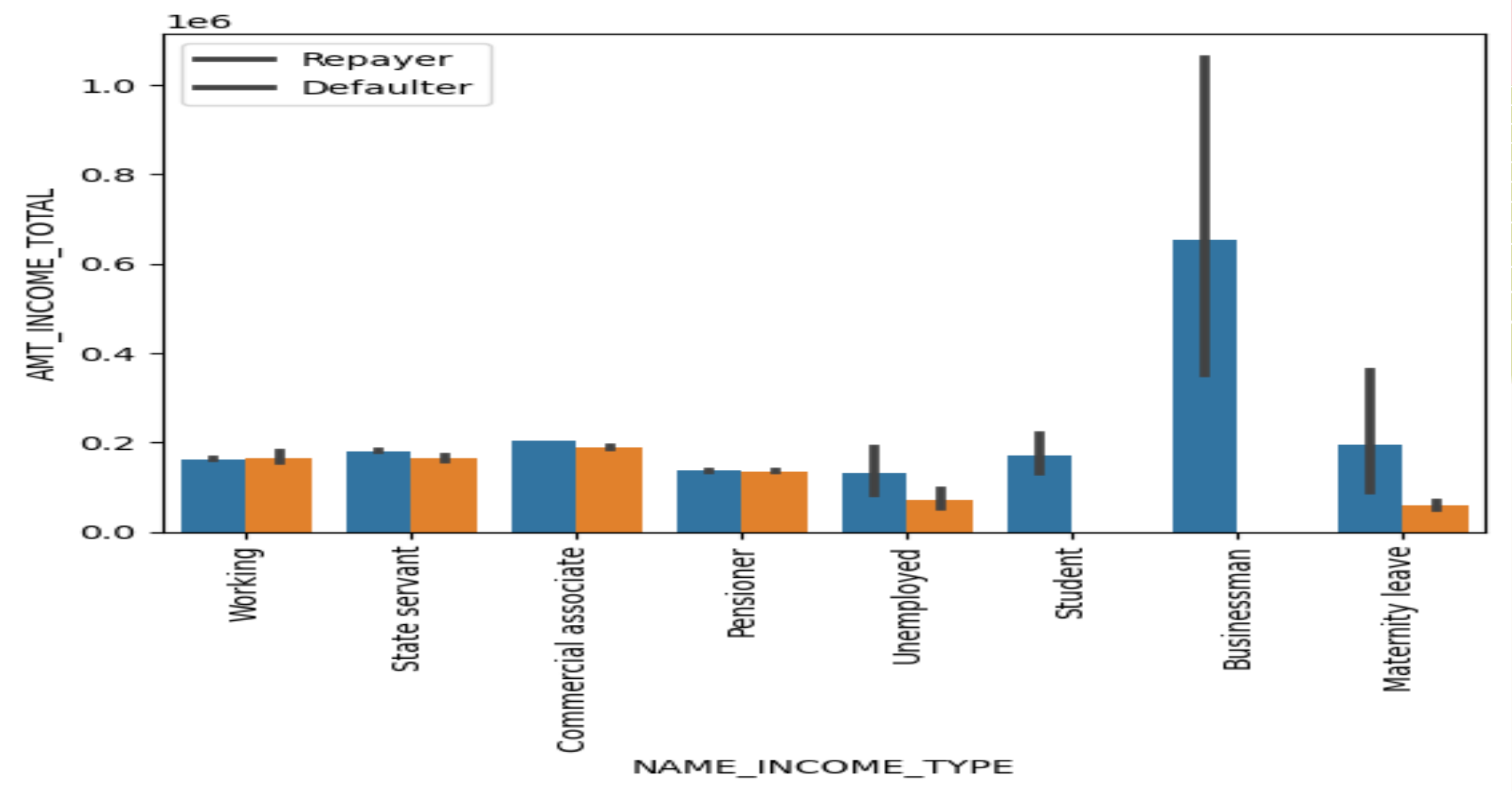

Count Plot of NAME_HOUSING_TYPE with respect to TARGET
TARGET
0
1
Count
250000
200000
150000
100000
50000
0
House / apartment
Rented apartment
With parents
Municipal apartment
Office apartment
Co-op apartment
NAME_HOUSING_TYPE


Defaulters within NAME_HOUSING_TYPE
Percentage of Defaulters
12
10
8
6
4
2
0
11
Co-op apartment
House / apartment
Municipal apartment
Office apartment
Rented apartment
With parents
NAME_HOUSING_TYPE

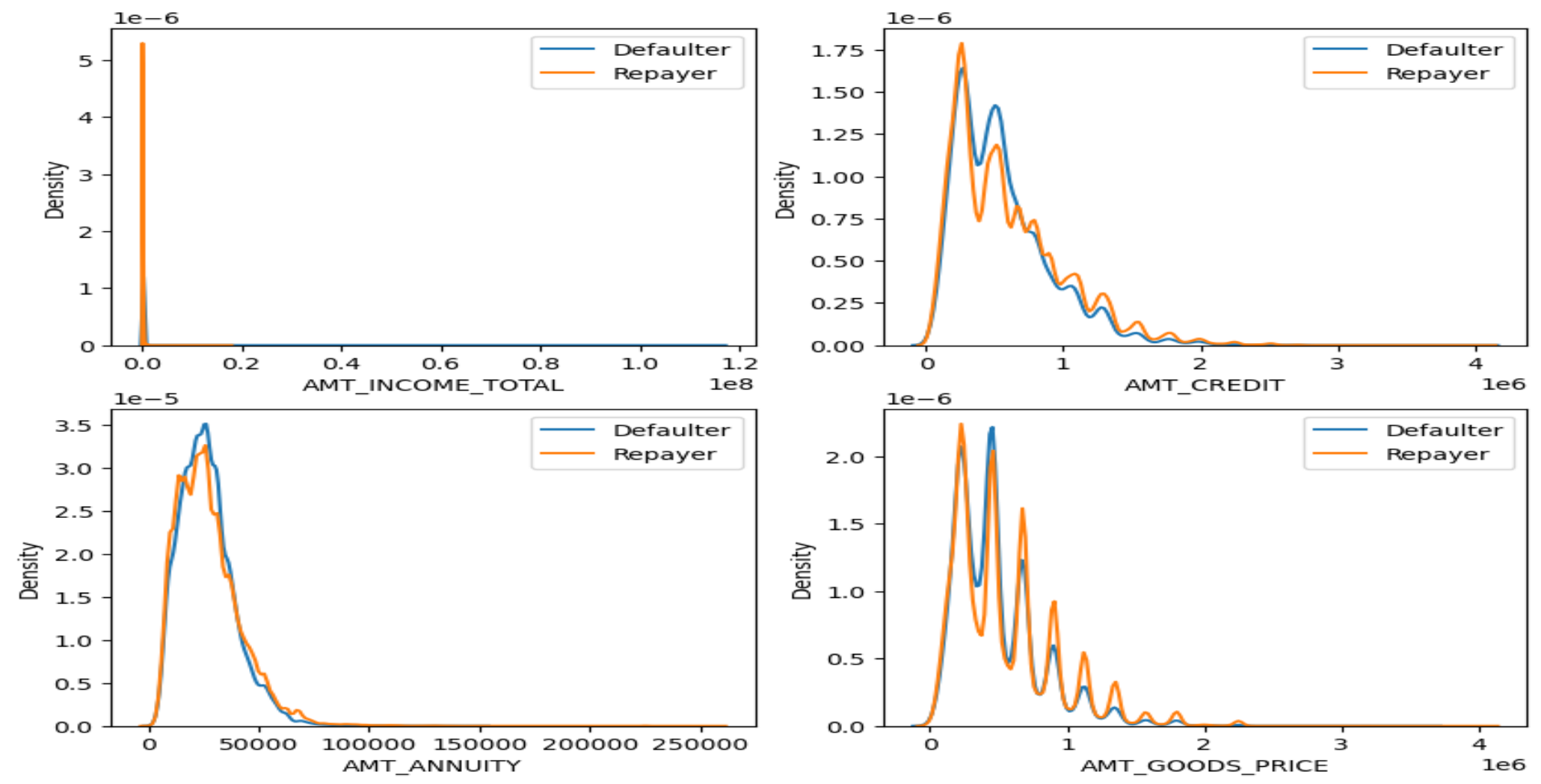

Count Plot of NAME_FAMILY_STATUS with respect to TARGET
TARGET
0
1
Count
175000
150000
125000
100000
75000
50000
25000
0
Single / not married
Married
Civil marriage
Widow
Separated
Unknown
NAME_FAMILY_STATUS


Defaulters within NAME_FAMILY_STATUS
Percentage of Defaulters
10
8
6
4
2
0
Civil marriage
Married
Separated
Single / not married
Unknown
Widow
NAME_FAMILY_STATUS

Bivariate analysis

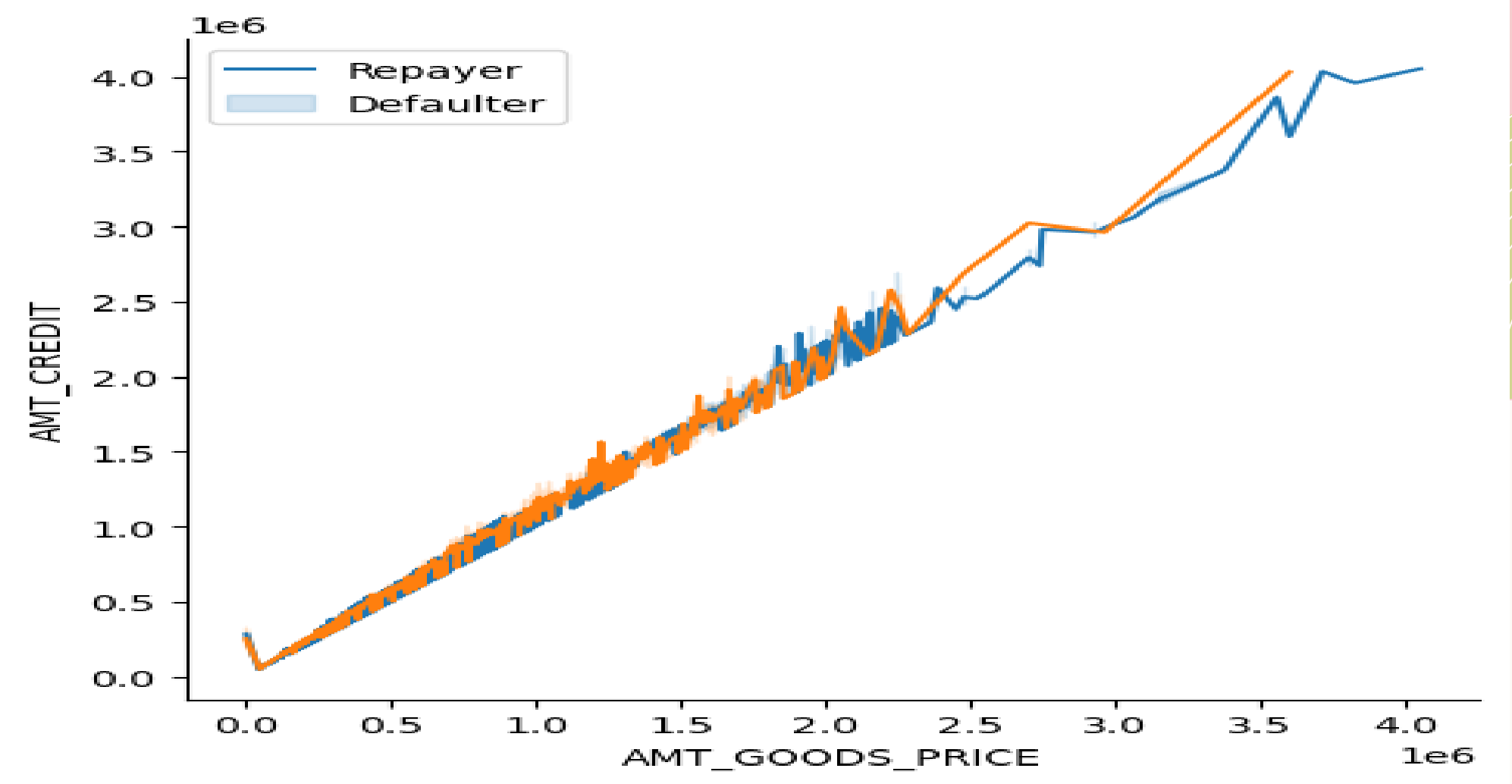


Univariate : Numerical column



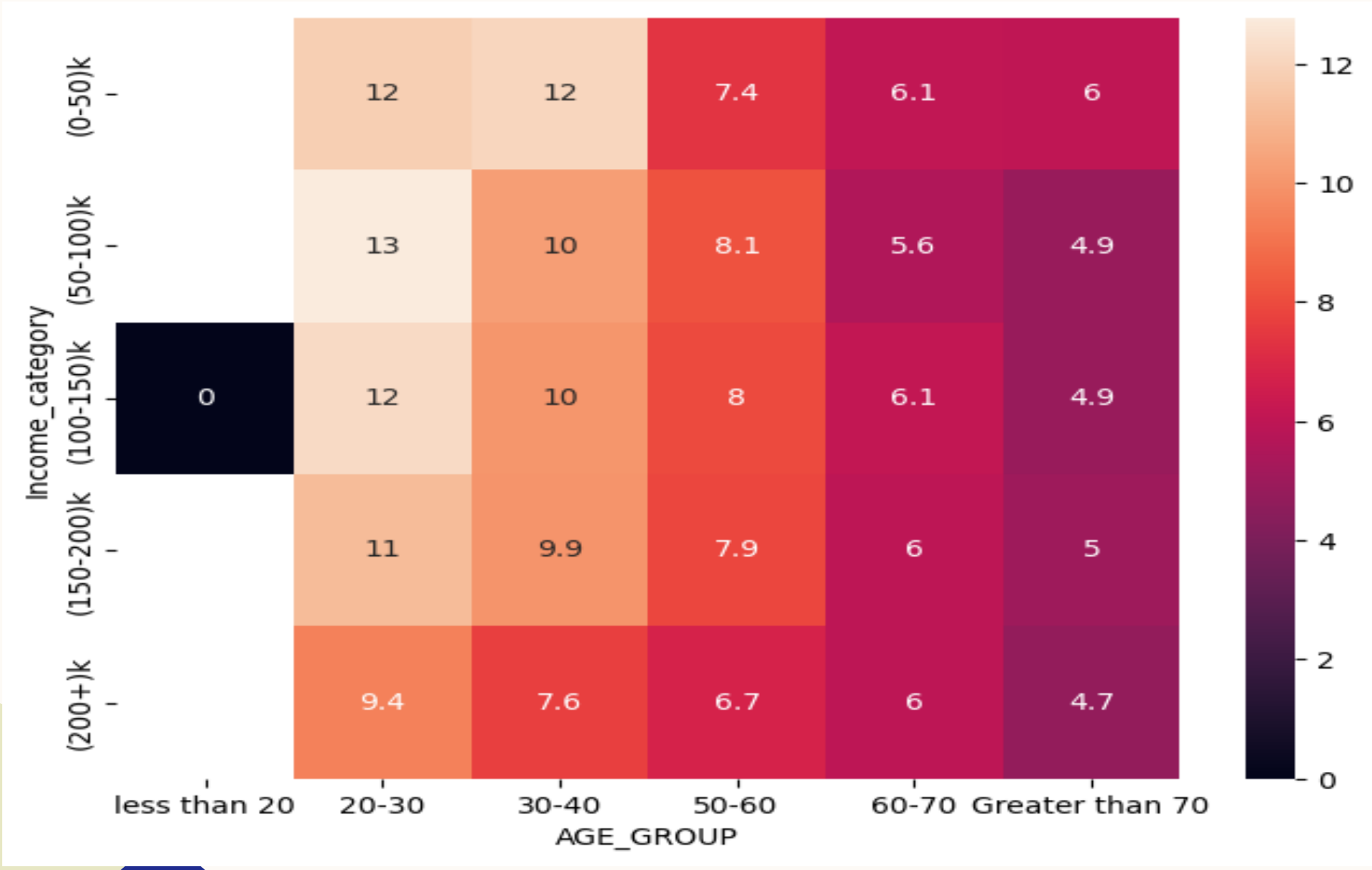
The repayers and defaulters distribution overlap in all the plots so we don’t use this plots for conclusion.

Bivariate : Numerical



According to above I conclude When credit amount goes up to 30-32 lakhs default rate increases.

HEATMAP : RATE OF DEFAULTING W.R.T CATEGORY INCOME AND AGE GROUP



Data Analysis: Conclusion

- I have done my univariate and bivariate analysis for categorical columns and numerical columns.
- NAME_INCOME_TYPE: according to my graph Student and Businessmen have no defaults.
- REGION_RATING_CLIENT: according to my graph RATING 1 is safer.
- ORGANIZATION_TYPE: Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less
- NAME_CASH_LOAN_PURPOSE: Loans bought for Hobby, Buying garage are being repaid mostly.
- CNT_CHILDREN: People with zero to two children tend to repay the loans.
- According to my heatmap analysis Age group greater than 70 whose salary is 50k-200k default rate is low.

Default chances:

- CODE_GENDER: according to my graph as compare to female male are higher default rate
- NAME_FAMILY_STATUS : according to my graph civil marriage and single are more chances to default.
- NAME_EDUCATION_TYPE: according to my graph Lower Secondary & Secondary education people are more chances to default
- NAME_INCOME_TYPE: according to my graph people who are in Maternity leave and Unemployed chances of default are more.
- REGION_RATING_CLIENT: according to my graph client who live in Rating 3 region has high chances of defaults.
- OCCUPATION_TYPE: Low-skill Laborers has high chances of defaults.
- CNT_CHILDREN & CNT_FAM_MEMBERS: people who have children 9 and more than 9 chances of defaults are more.
- When credit amount goes up to 30-32lakhs defaults rate increases.

THANK YOU