Credit EDA Case Study

JAI AHUJA
(DDS2090125)

AMYA GUPTA
(DDS2090035)

INDEX

- Introduction
- Business Understanding
- Business Objectives
- Data Imbalance
- Univariate Analysis
  - Categorical Analysis
  - Numerical Analysis
- Bivariate Analysis
- Correlation
- Merged Data Frame
- Top 5 Driver Variables for Default
- Conclusion and Recommendations

INTRODUCTION

This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, 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 as 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 must 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 must 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 contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

1. **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,
2. **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):

1. **Approved :** The Company has approved loan Application
2. **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.
3. **Refused :** The company had rejected the loan (because the client does not meet their requirements etc.).
4. **Unused offer:** Loan has been cancelled by the client but on different stages of the process.

In this case study, We will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

BUSINESS OBJECTIVES

This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants 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 knowledge for its portfolio and risk assessment.

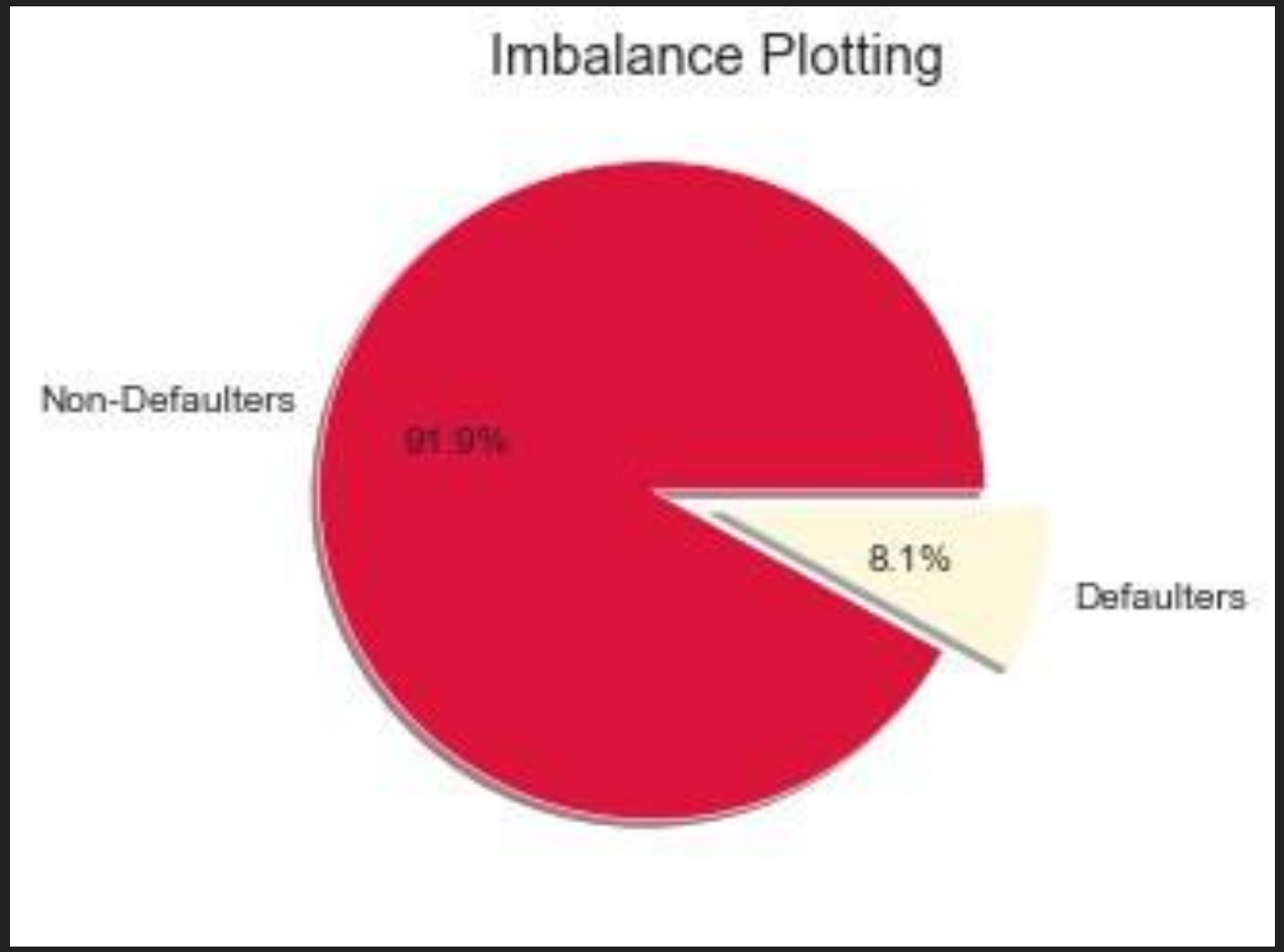
To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

UNDERSTANDING DATA

The dataset used has 3 files as explained below:

1. *'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.**
2. *'previous_application.csv'* contains information about the client's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**
3. *'columns_description.csv'* is data dictionary which describes the meaning of the variables.

DATA IMBALANCE



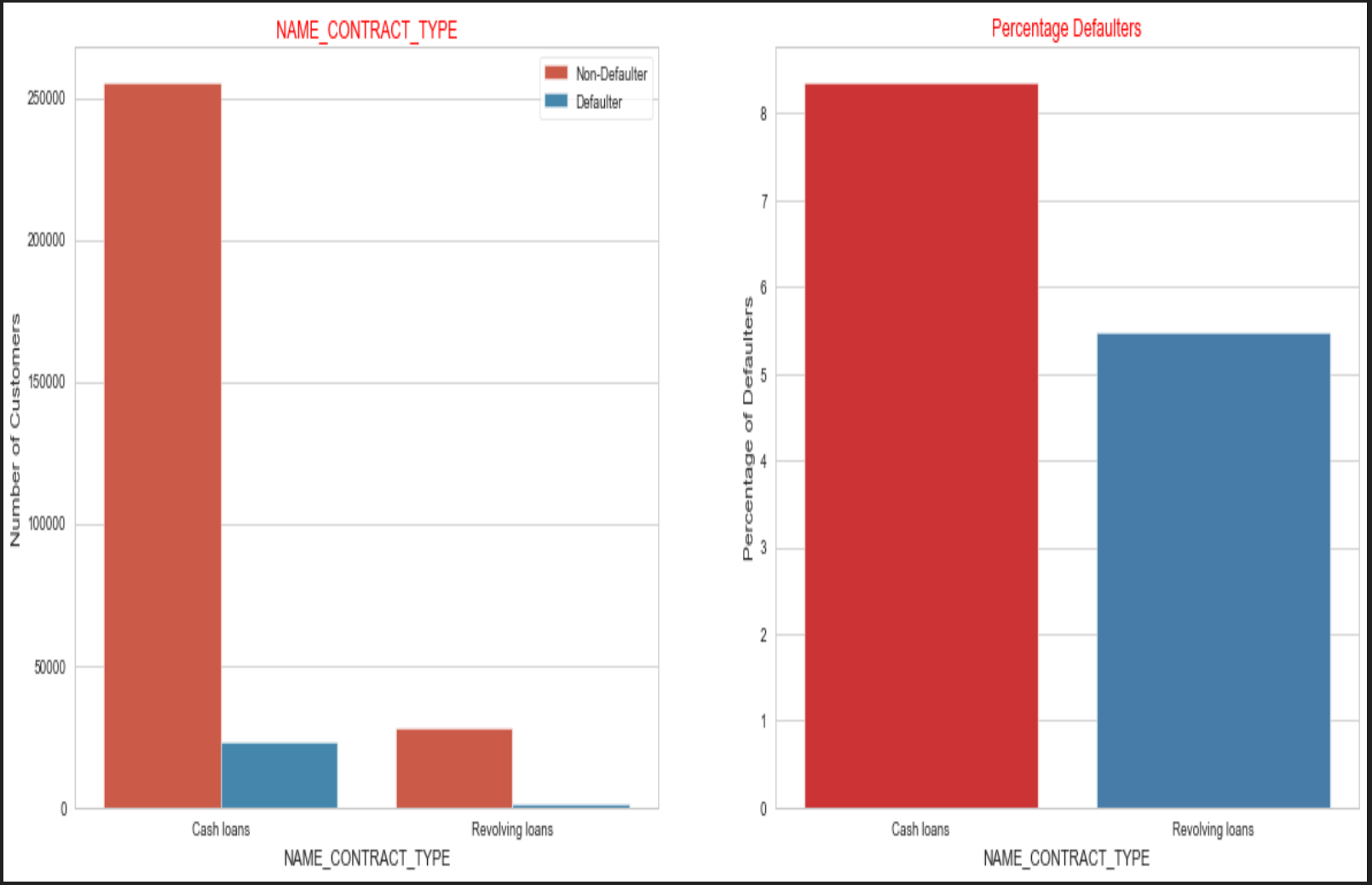
The Data is Imbalanced with Imbalance Ratio of Non-Defaulter to Defaulter as –

11.39 : 1

UNIVARIATE ANALYSIS

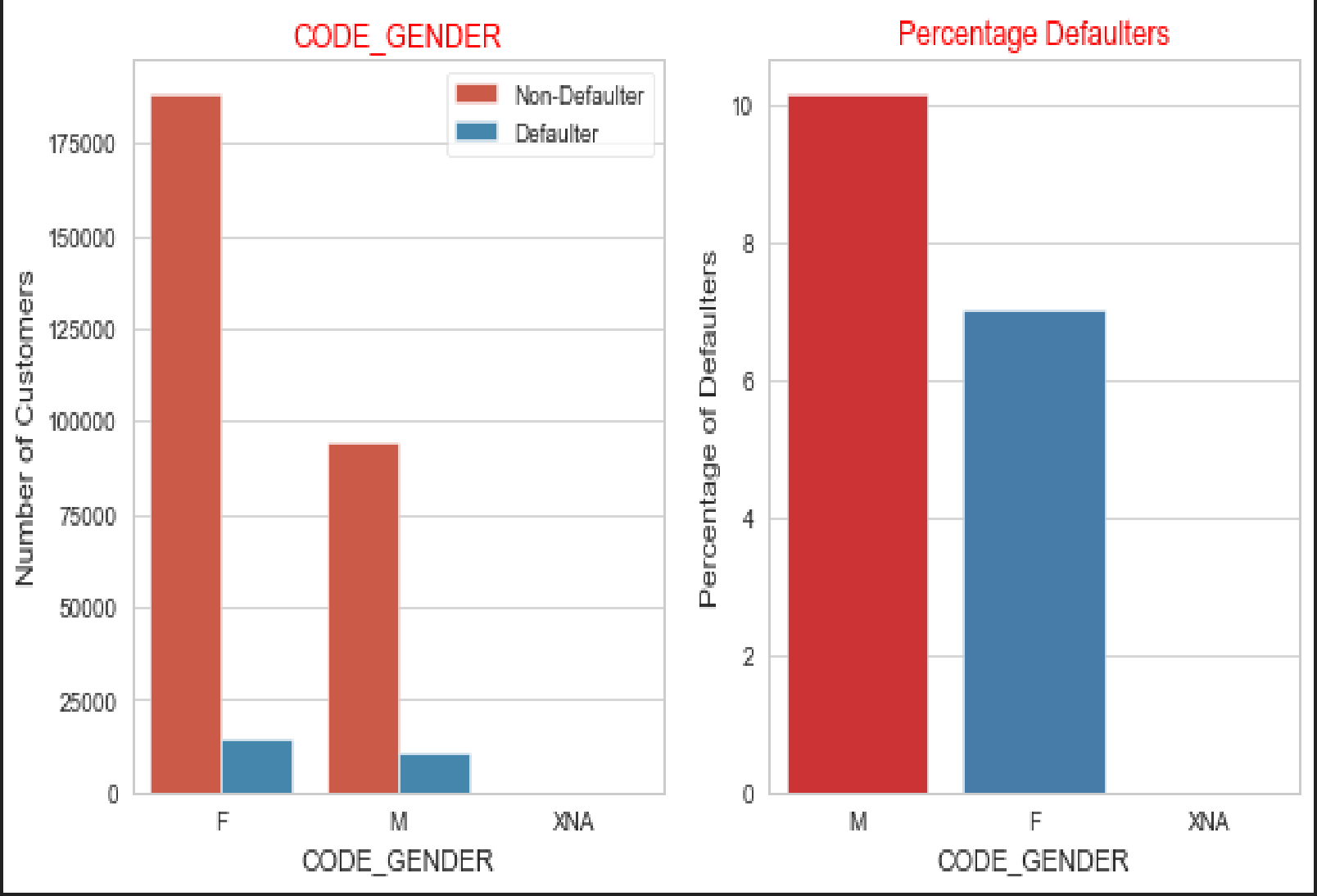
Categorical Variables

CONTRACT TYPE



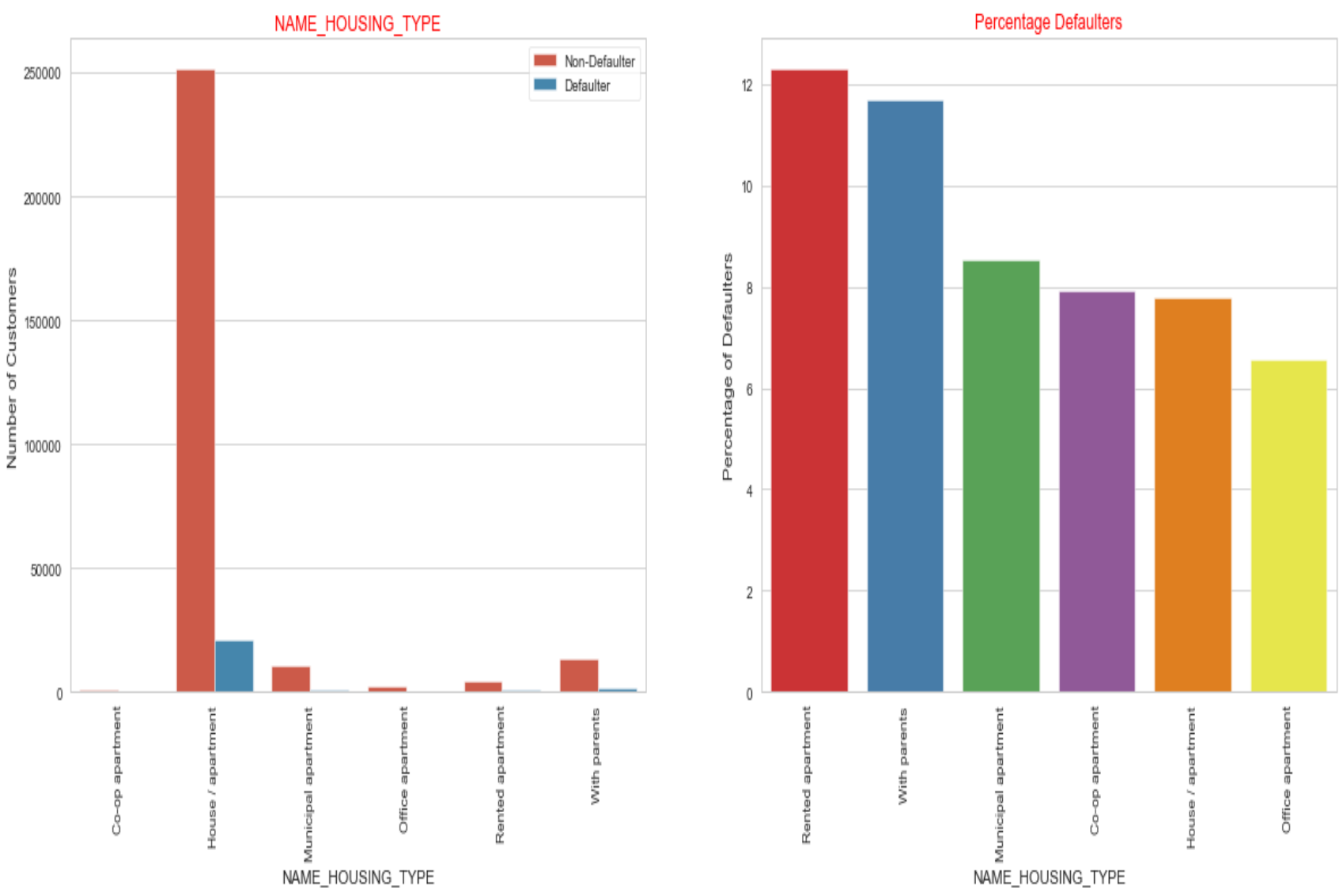
- There are very less customers with revolving loans and 5% of them have not repaid the loan.
- Approximately 8% of people with cash loans have not repaid the loan.

GENDER



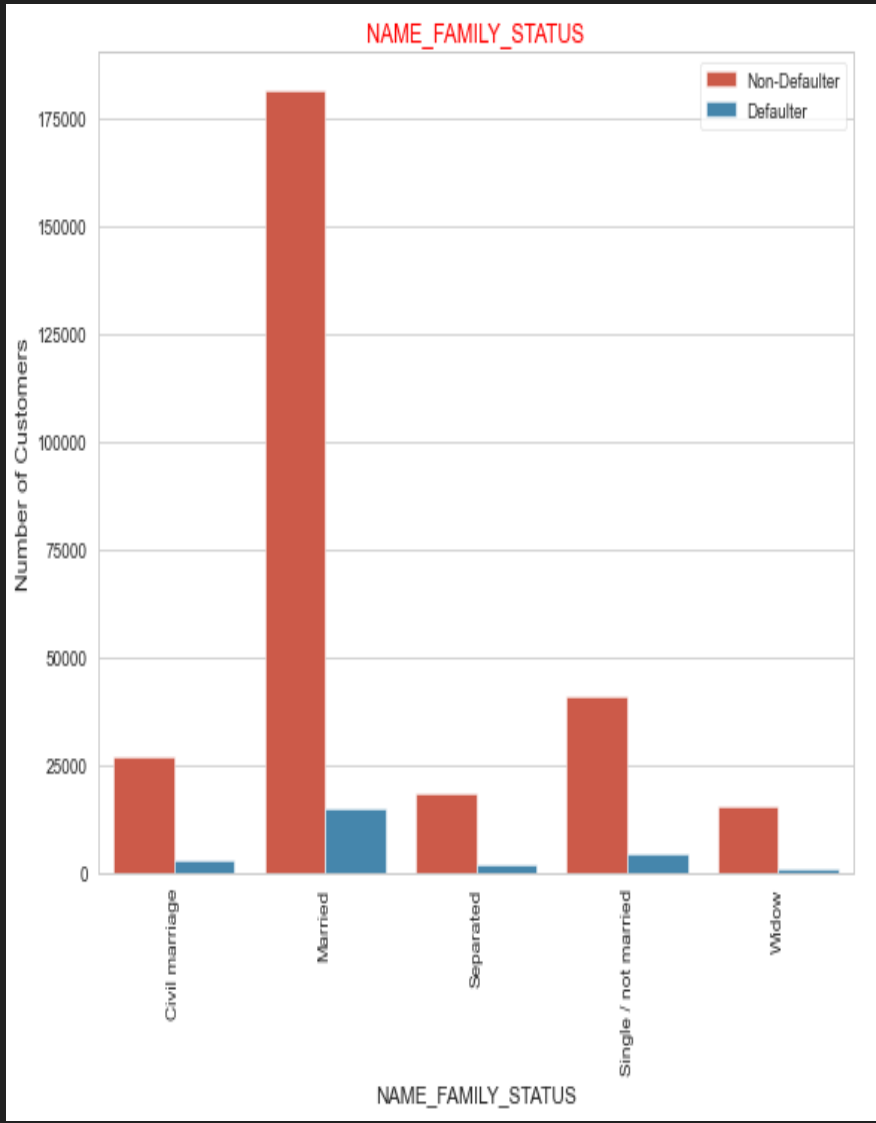
- Female customers have taken more loan, but male customers have higher number of defaulters.
- Female customers pay loan amount on time and banks can target more female customers for lending loan.

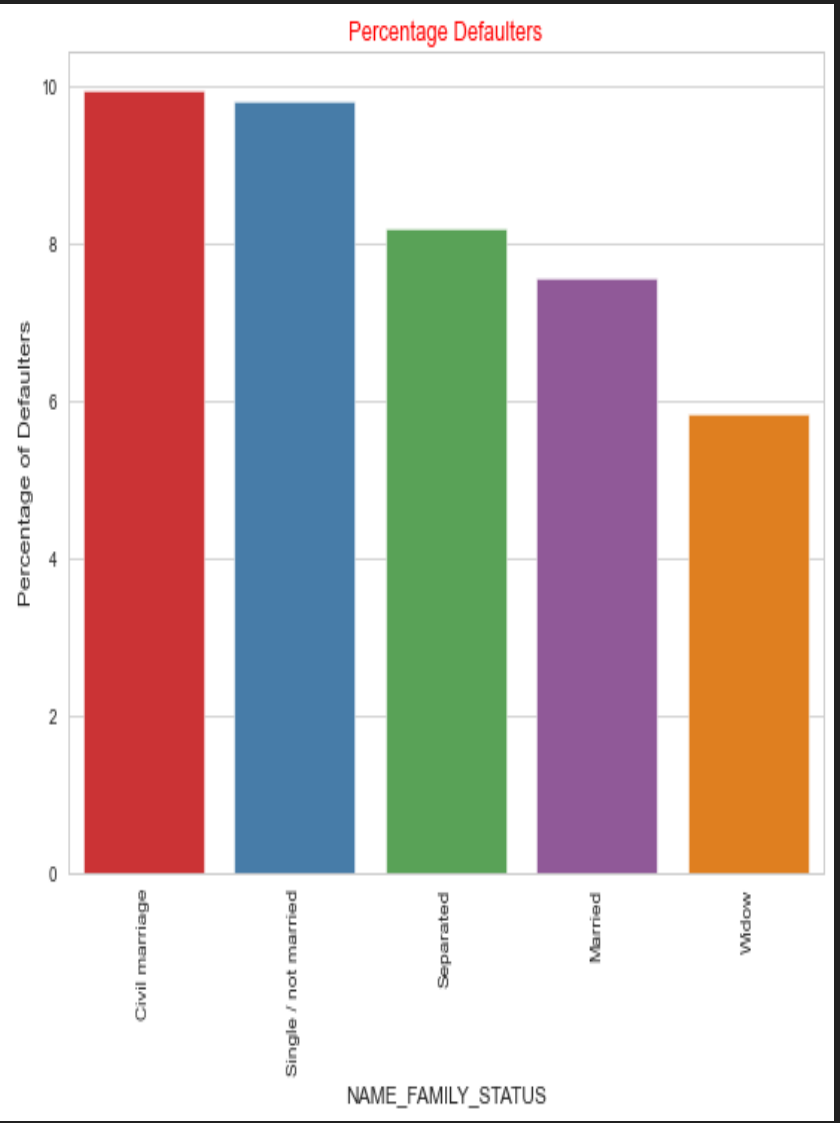
HOUSING TYPE



- People with House/Apartment have taken the greatest number of loans.
- Higher percentage of people who have not repaid the loans live in either Rented Apartments or with their parents.
- Customers owning Office apartment are most likely to make payments on time.

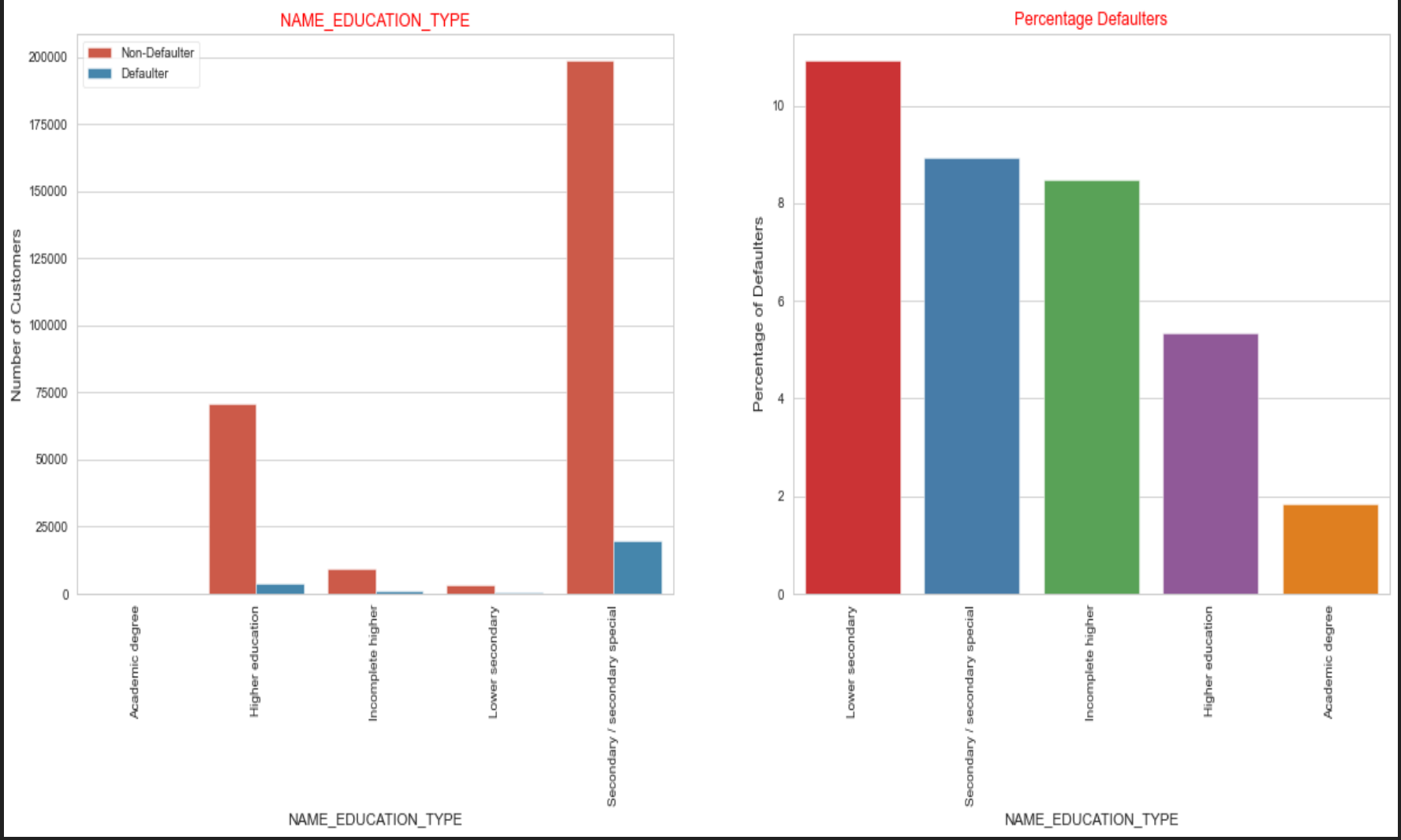
FAMILY STATUS





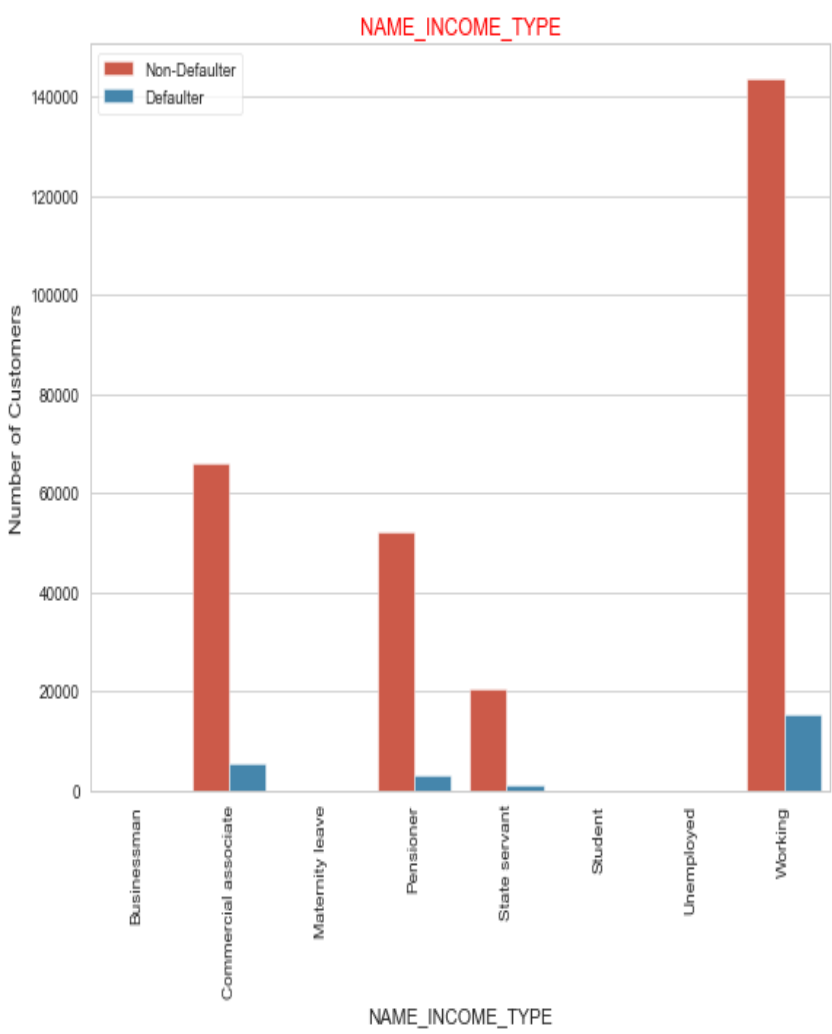
- Married people have taken the greatest number of loans.
- Civil Marriage folks and Single/ Unmarried are the ones who defaulted on the greatest number of loans.
- Widows have defaulted on least number of loans.

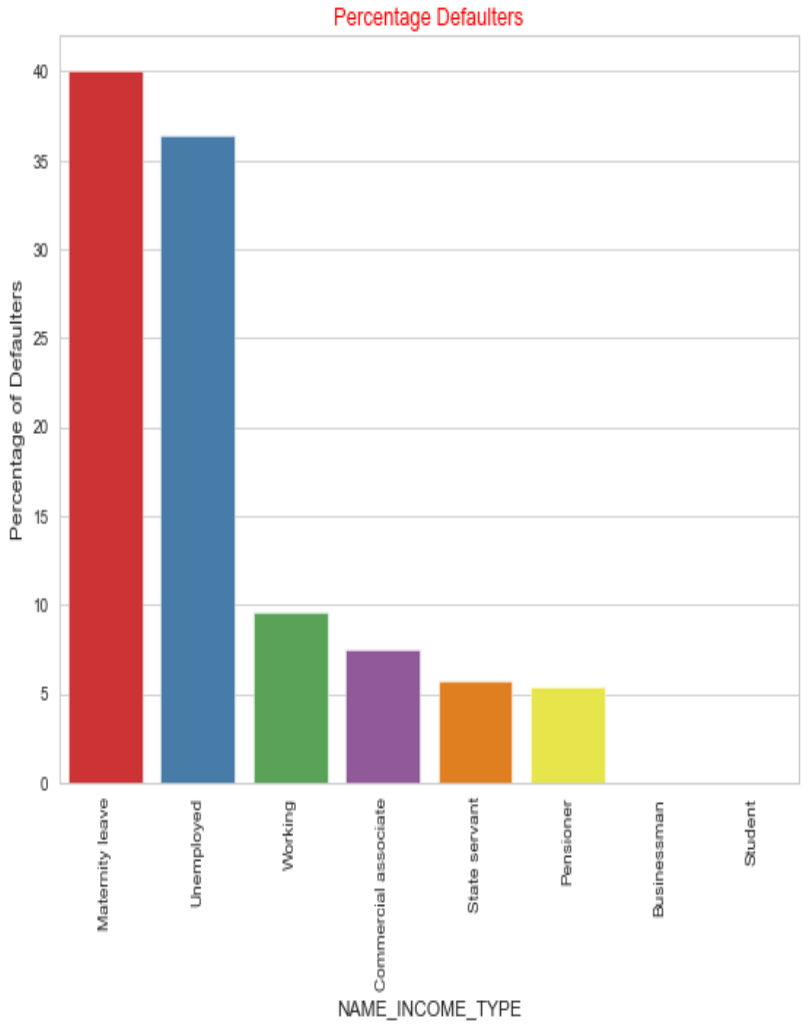
EDUCATION TYPE



- People who have done Secondary/ Secondary Special education have taken higher number of loans.
- People with Lower Secondary although have taken a very few number of loans but have the highest default percentage amongst them.
- People with Academic degree have less than 2% of defaulting rate.

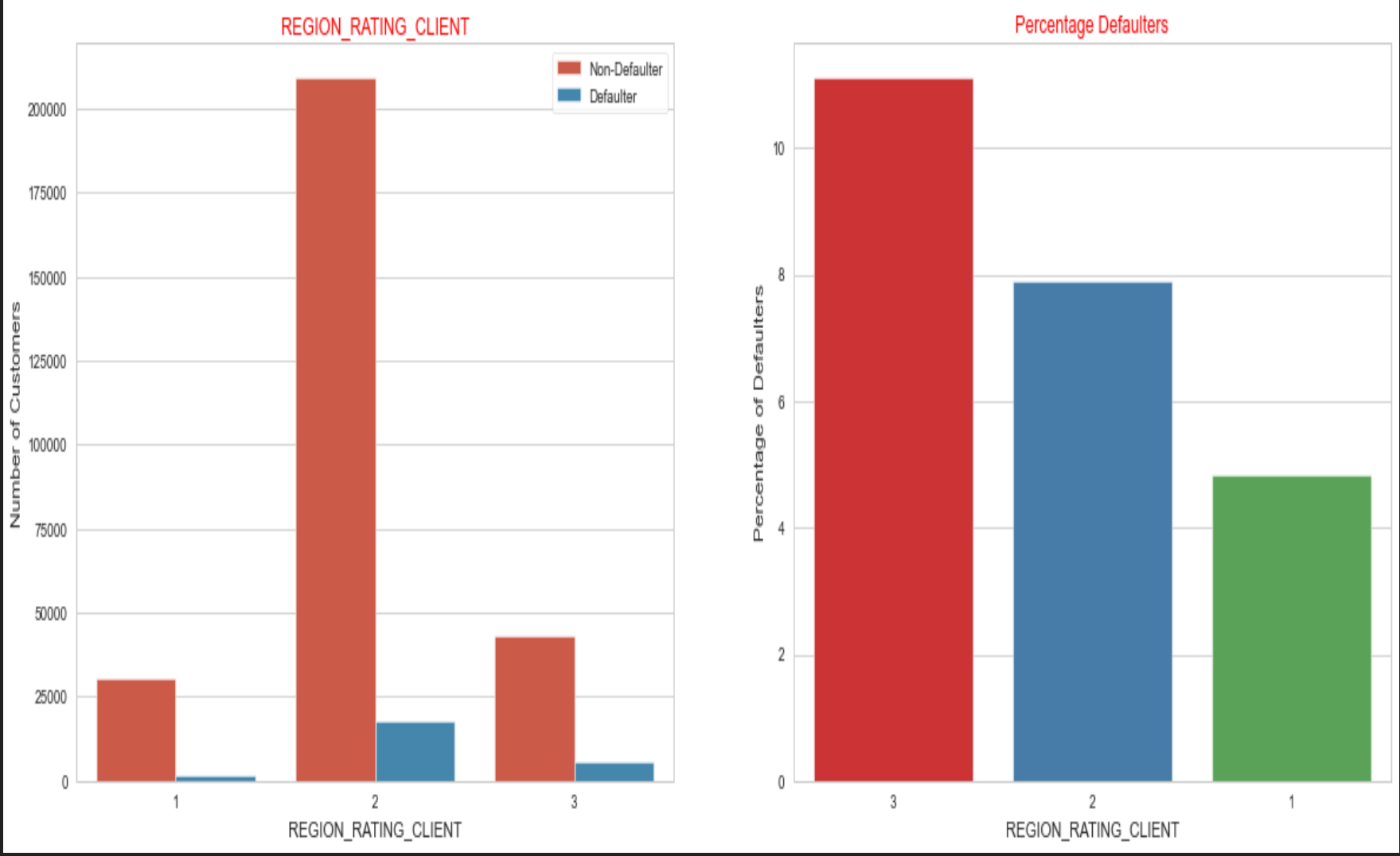
INCOME TYPE





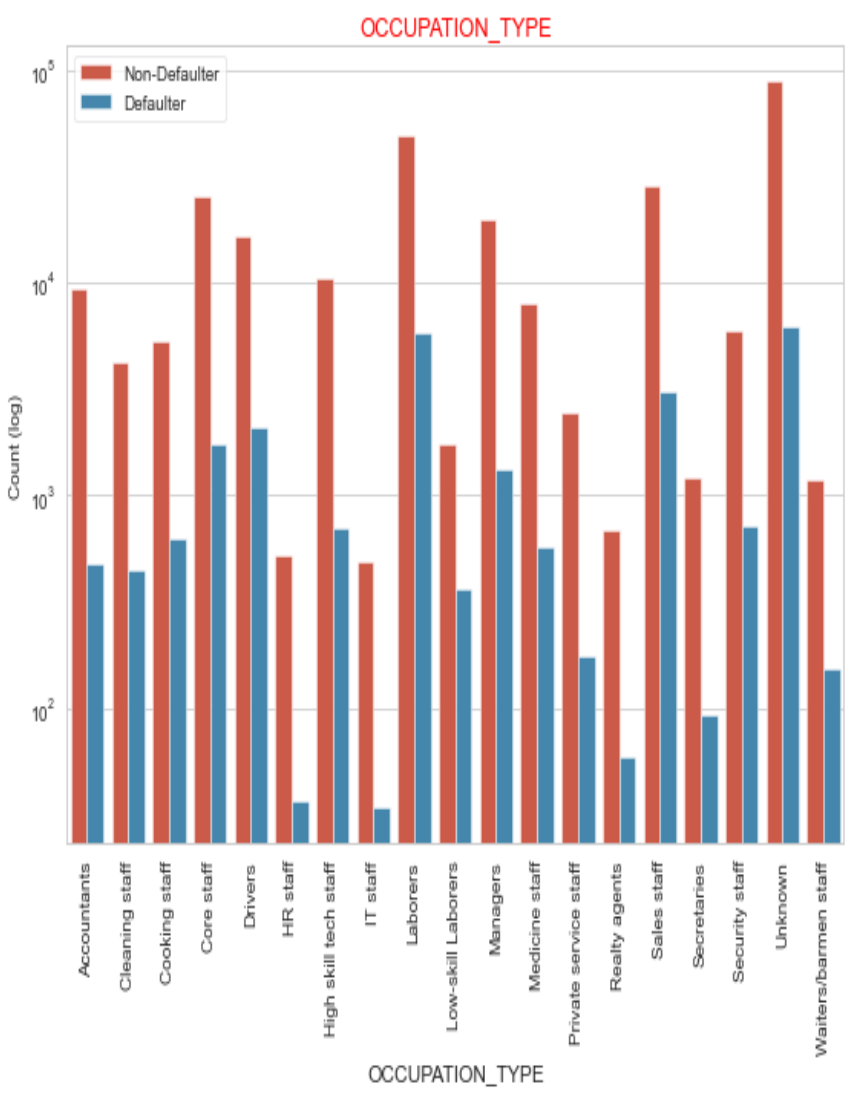
- People who are working have highest number of loans.
- Although females on maternity leaves have taken significantly lower number of loans but have approximately 40% default rate amongst them which is the highest in any category.
- People who are unemployed have a default rate of more than 35%.

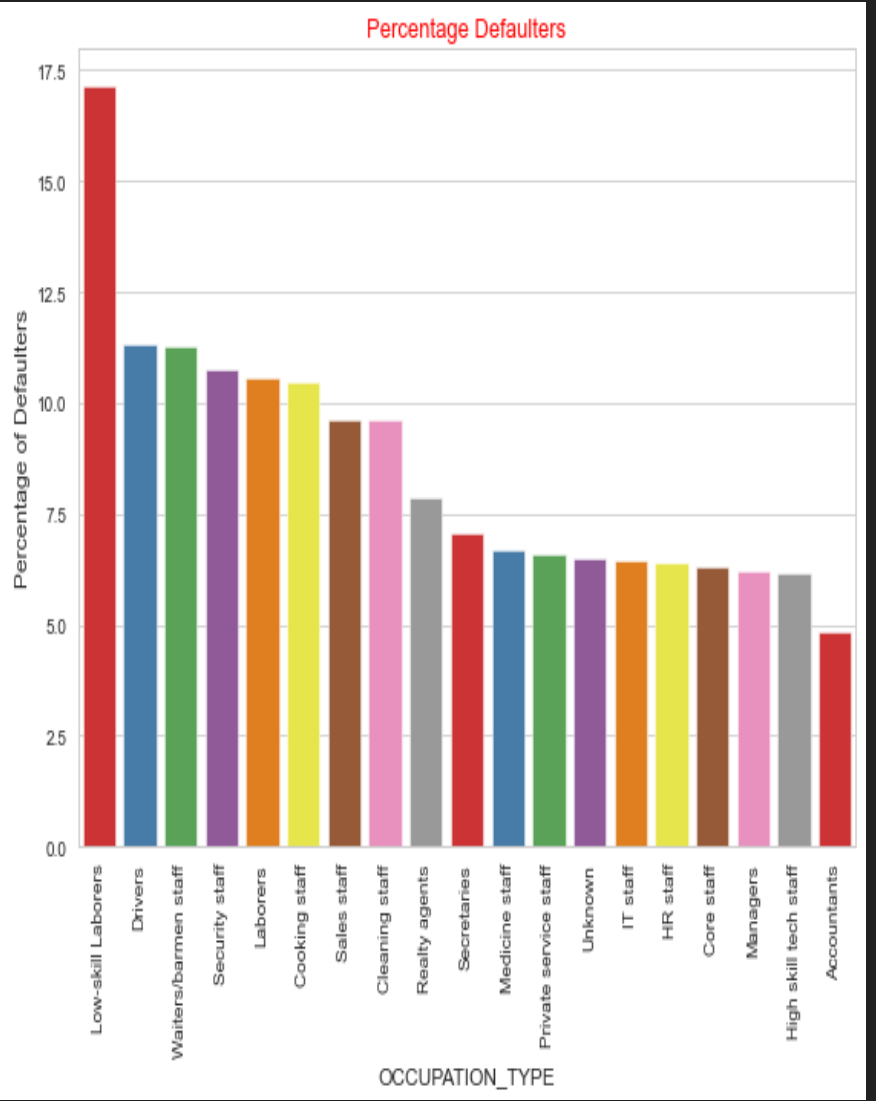
REGION RATING CLIENT



- Most of the people who have applied for loans are living in REGION_RATING_CLIENT 2.
- Applicants living in REGION_RATING_1 have defaulted least no. of loans whereas applicants living in REGION_RATING_3 have defaulted the greatest number of loans.

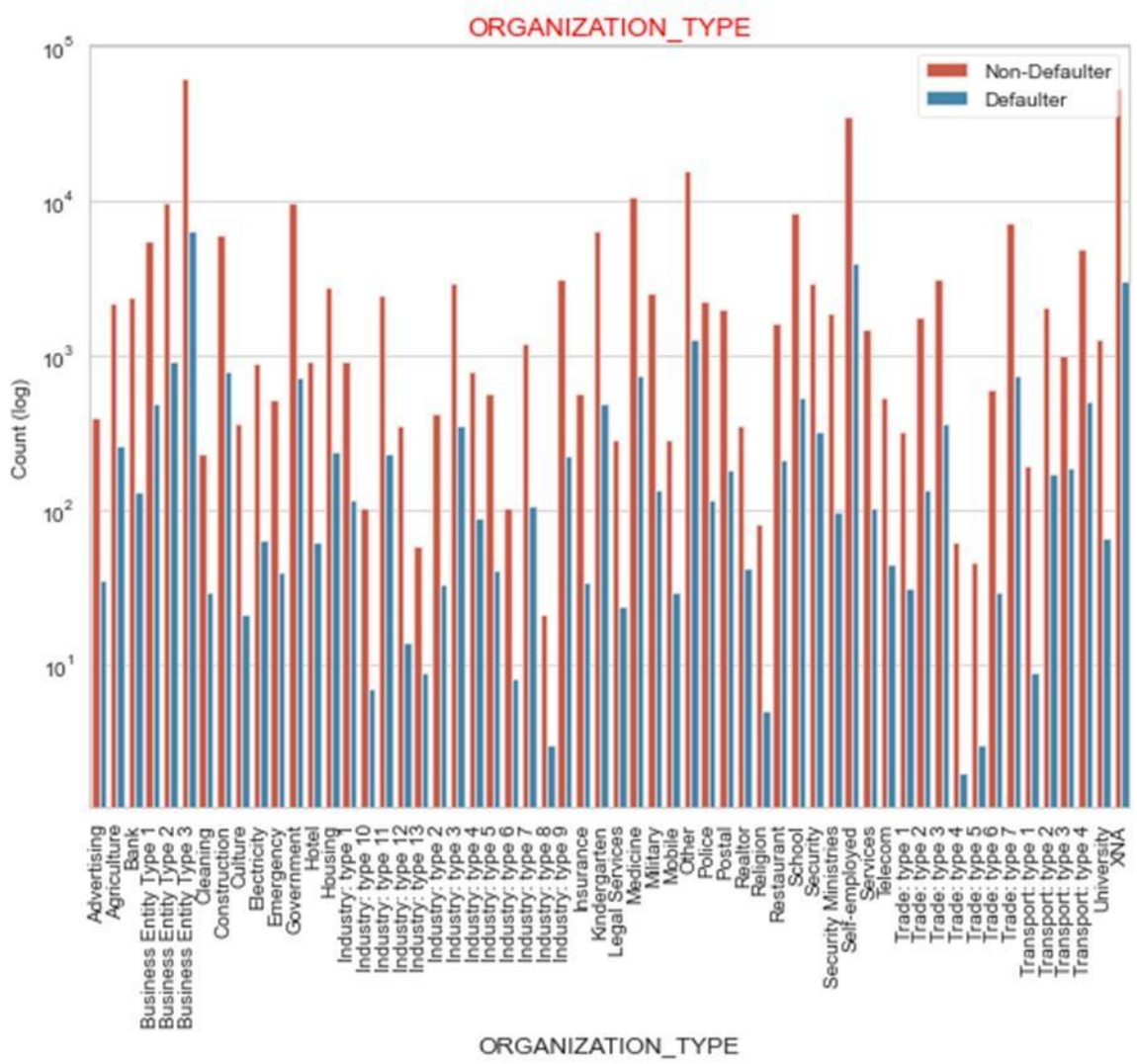
OCCUPATION TYPE

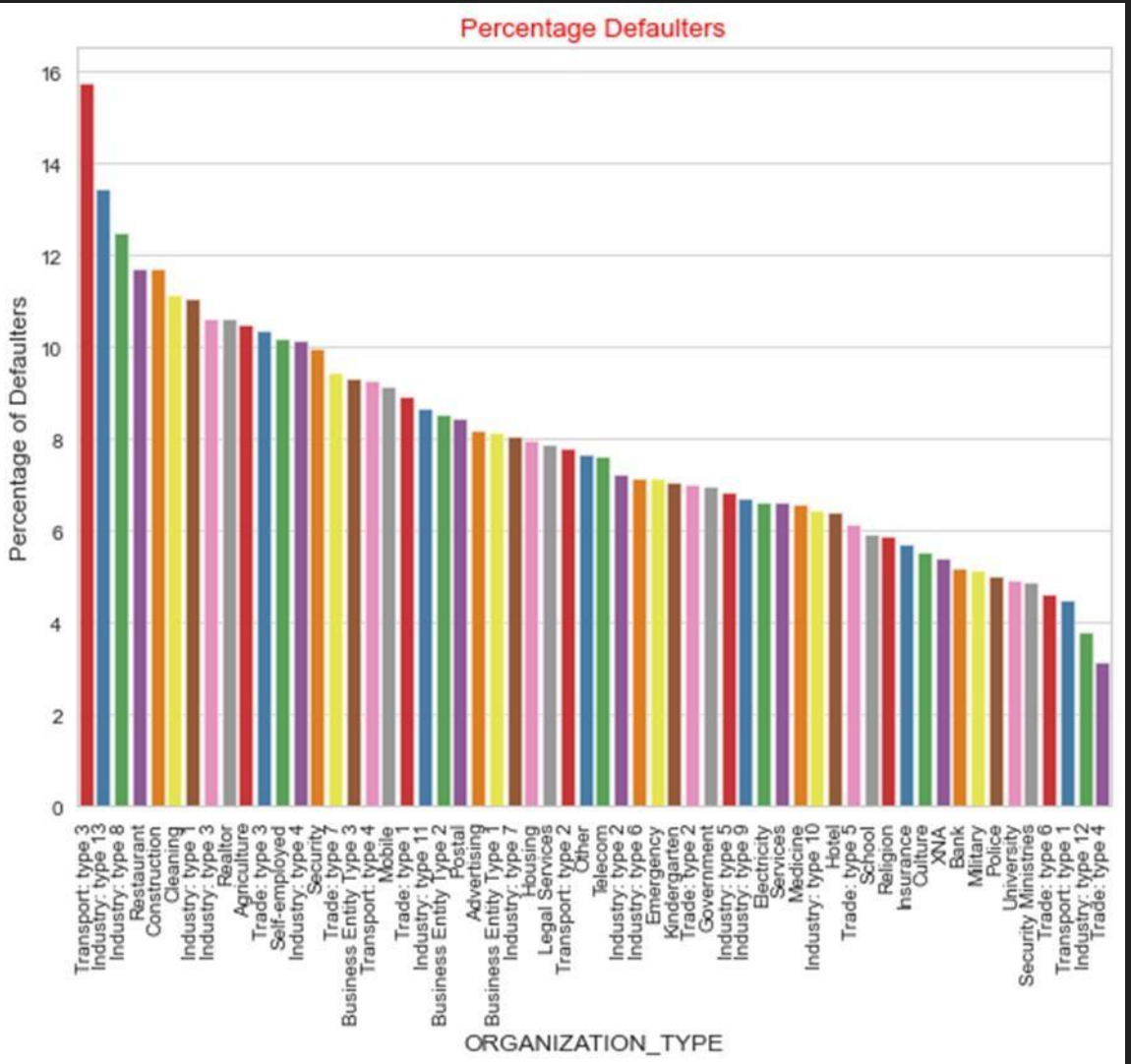




- Laborers have taken most no. of loans followed by Sales staff, core staff, Managers.
- Low-skill Laborer's have defaulted most no. of loans(approx. 17%) followed by Waiters/Barmen staff, Drivers,Secondary staff.
- For a very high number of applications occupation type information is unknown.

ORGANIZATION TYPE

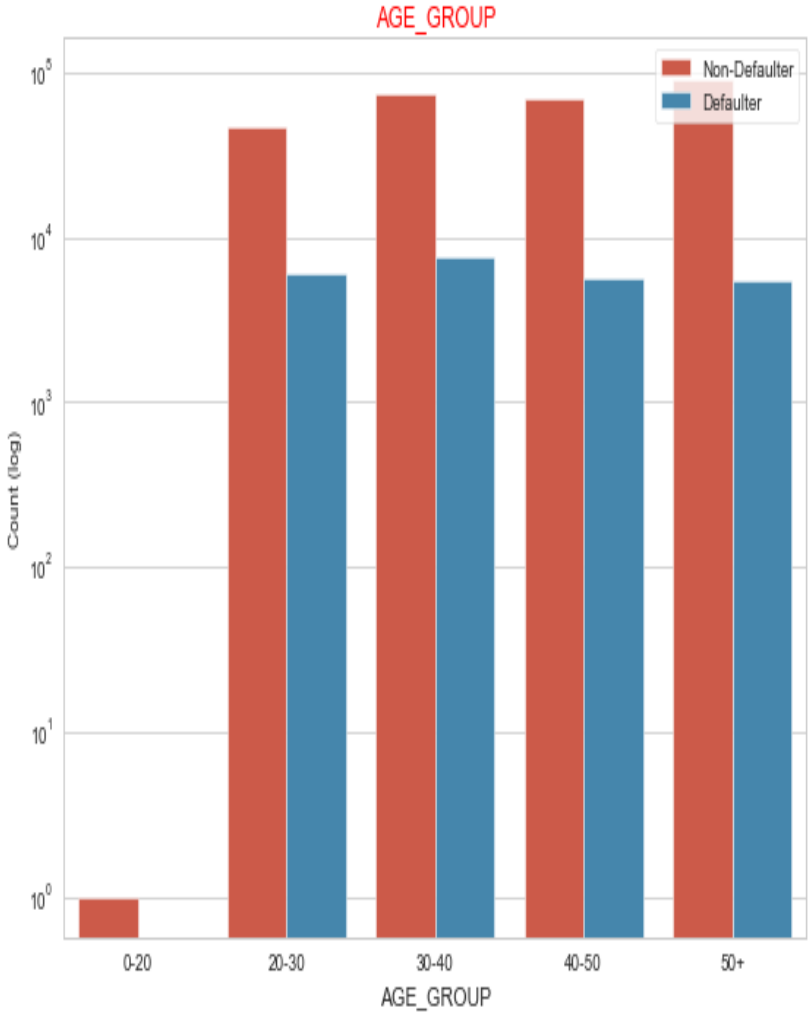


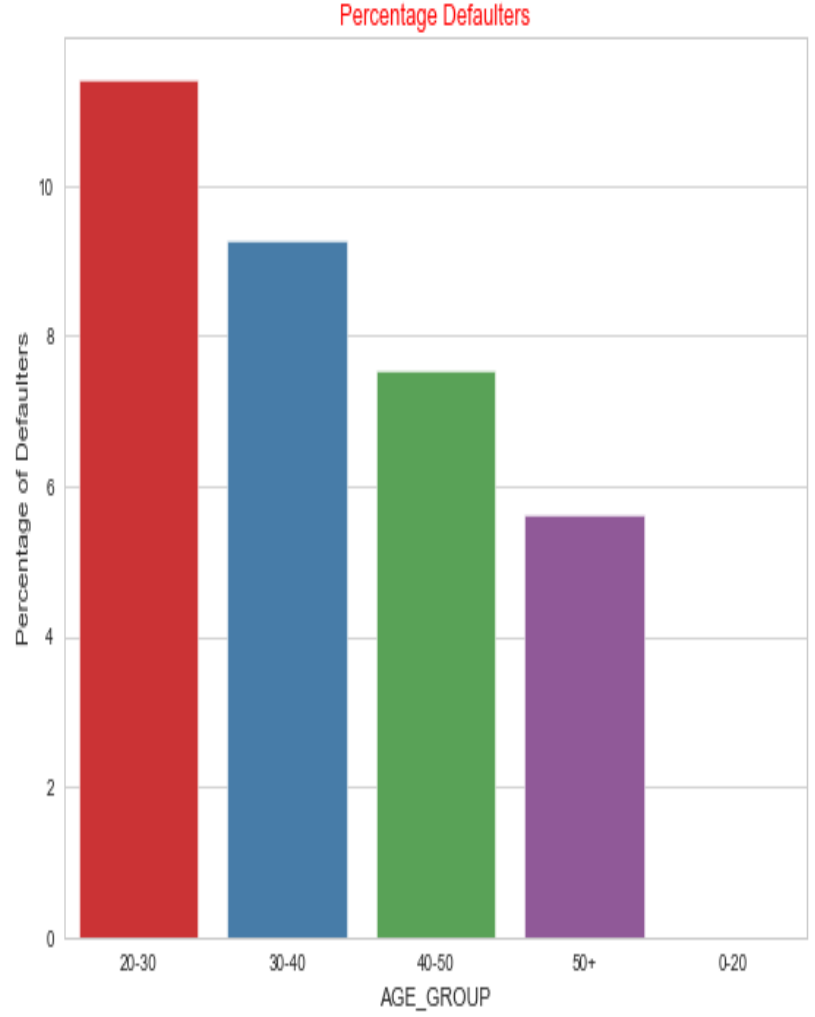


ORGANIZATION TYPE

- Most of the people who applied for loan are from Business Entity Type 3.
- For a very high number of applications, Organization type information is missing(XNA).
- Industry Type 12, Trade type 4 has less defaulters(less than 4%), therefore the applicants from these organizations can be trusted.
- Transport type 3 have more than 15% of defaulters, making it the category with highest defaulters.

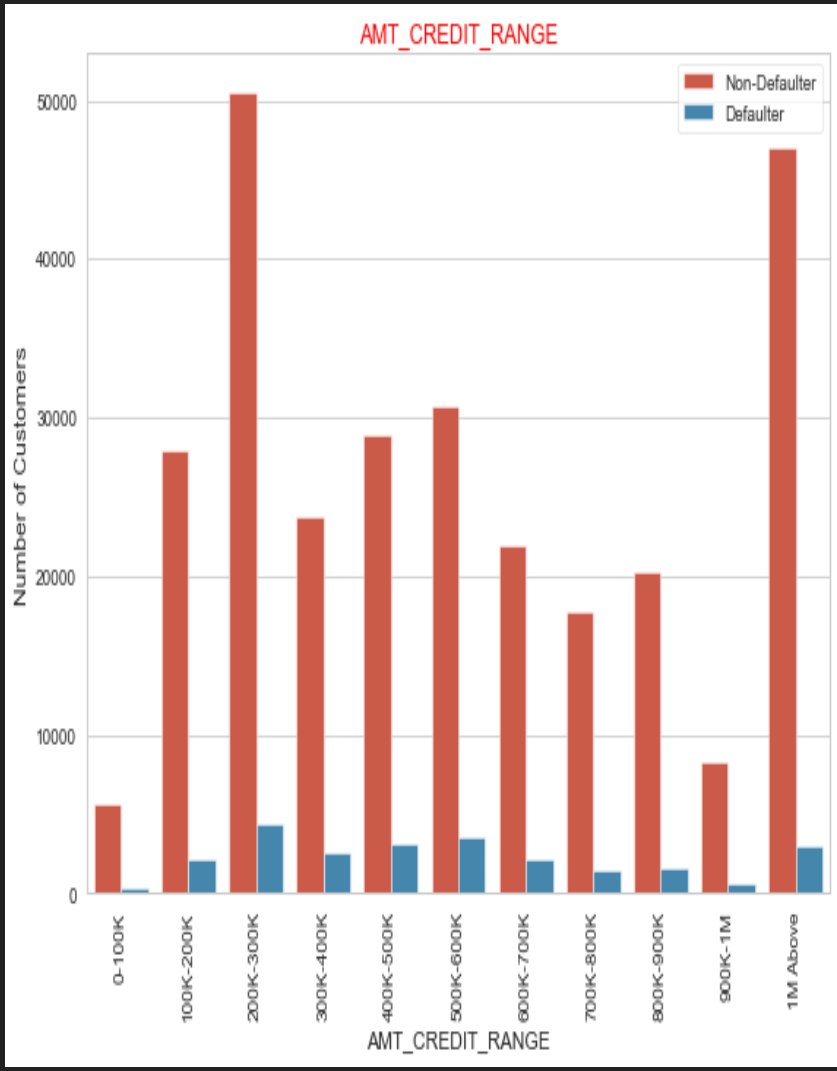
AGE GROUP

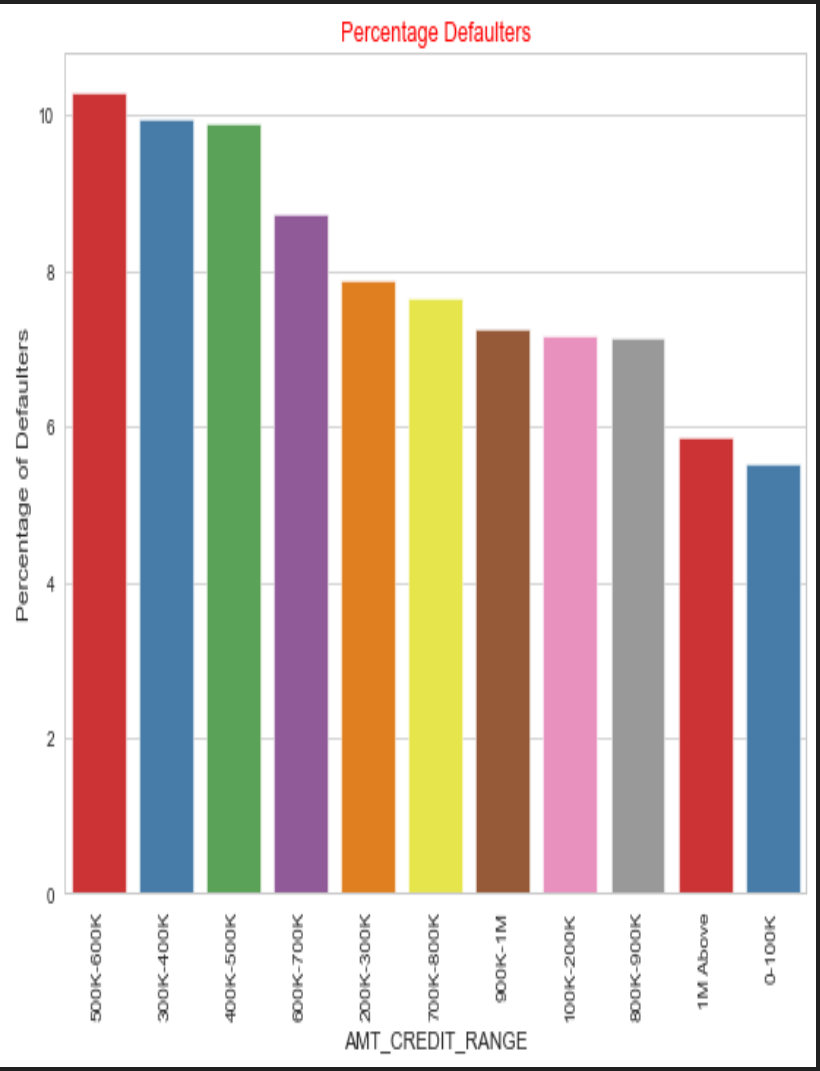




- All age group from 20 to 50+ have applied for loans and most defaulters are in 20-40 age group.

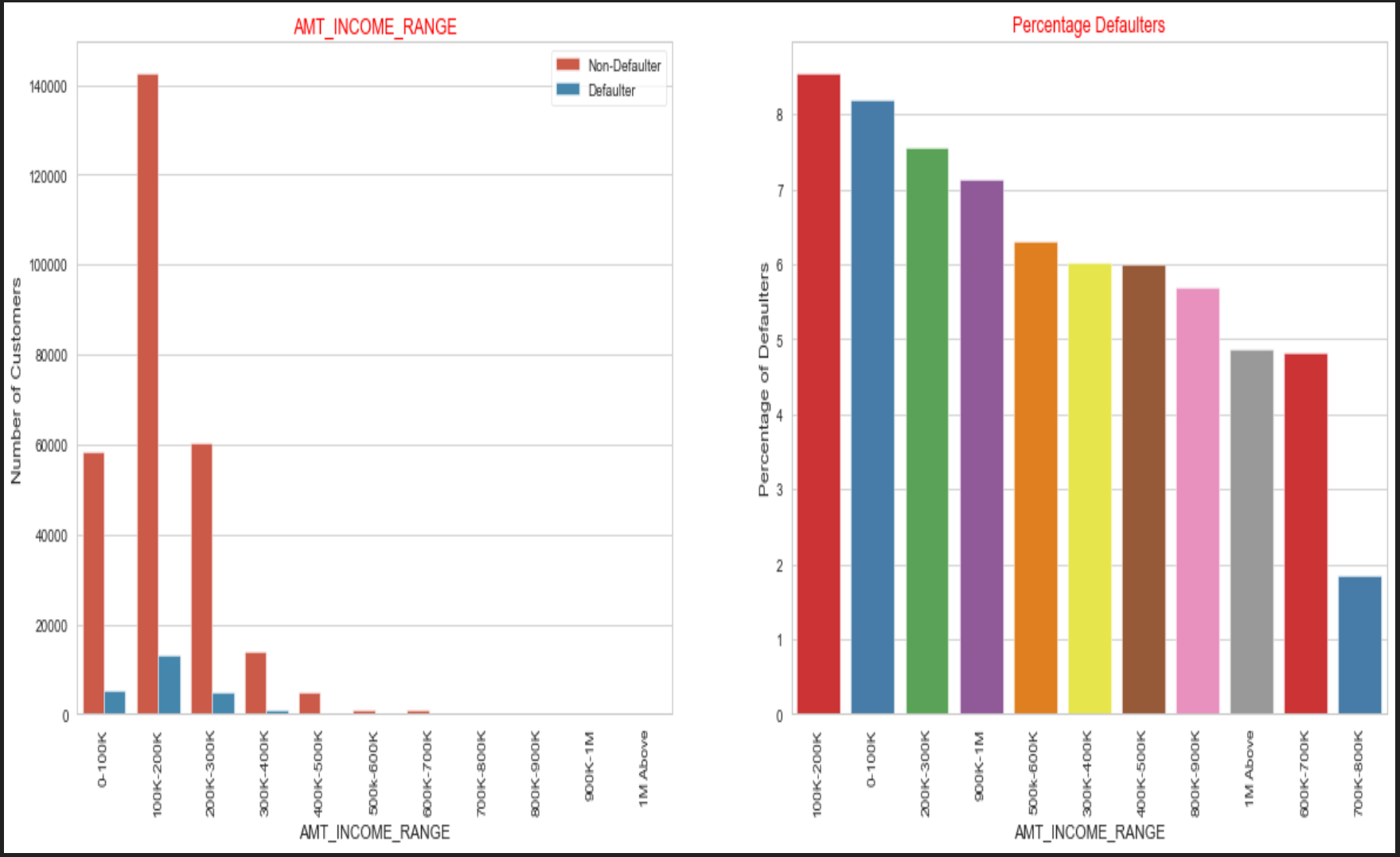
CREDIT RANGE





- People with credit amount between 400K-600K tend to default more than others.

INCOME RANGE

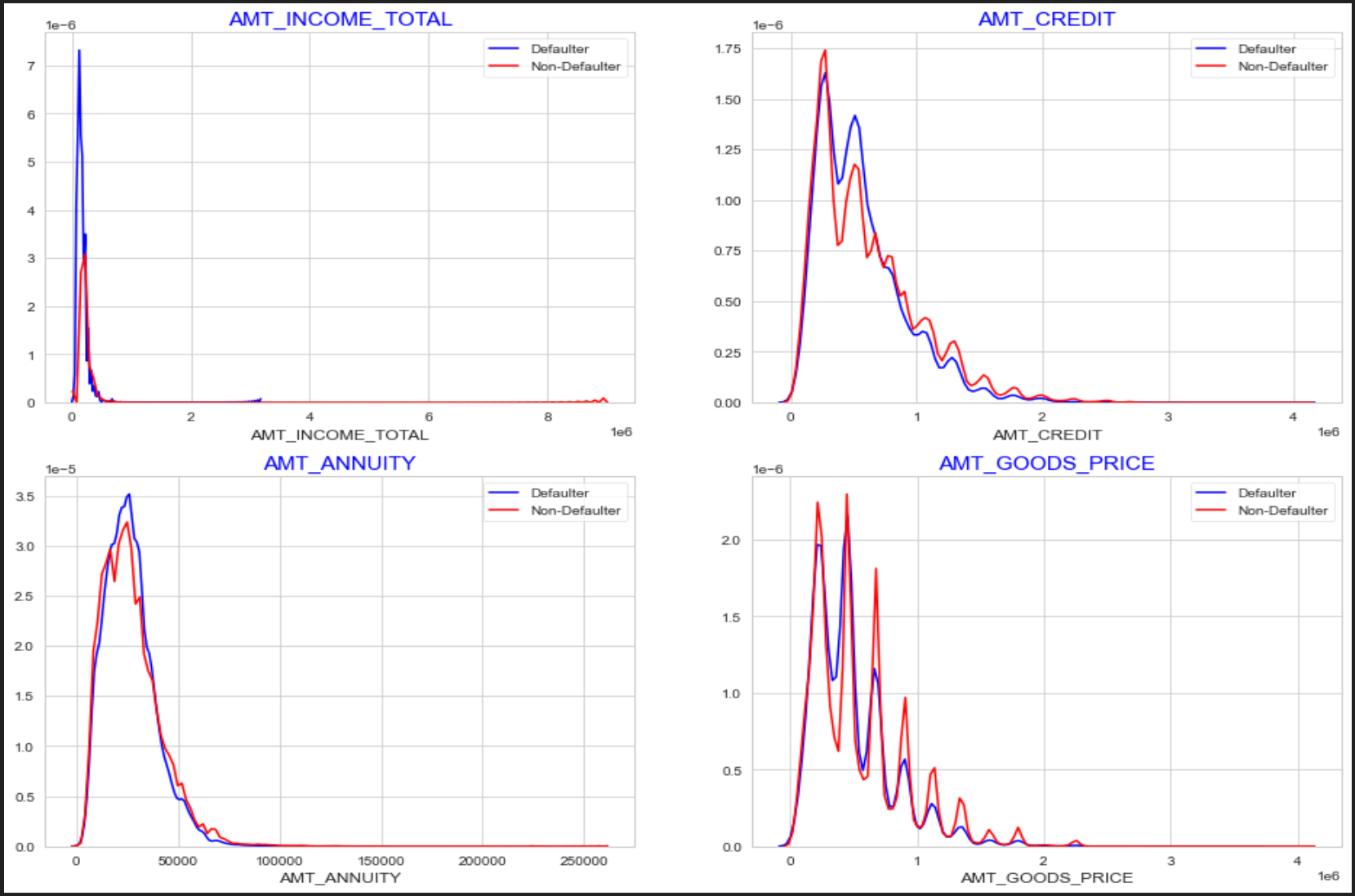


- Most of the applicants have income less than 300K.
- Applicants with income more than 700K are less likely to default.
- Applicants with income less than 300K are more likely to default.

UNIVARIATE ANALYSIS

Numerical Variables

Numerical Variables: INCOME, CREDIT, ANNUITY, GOODS PRICE

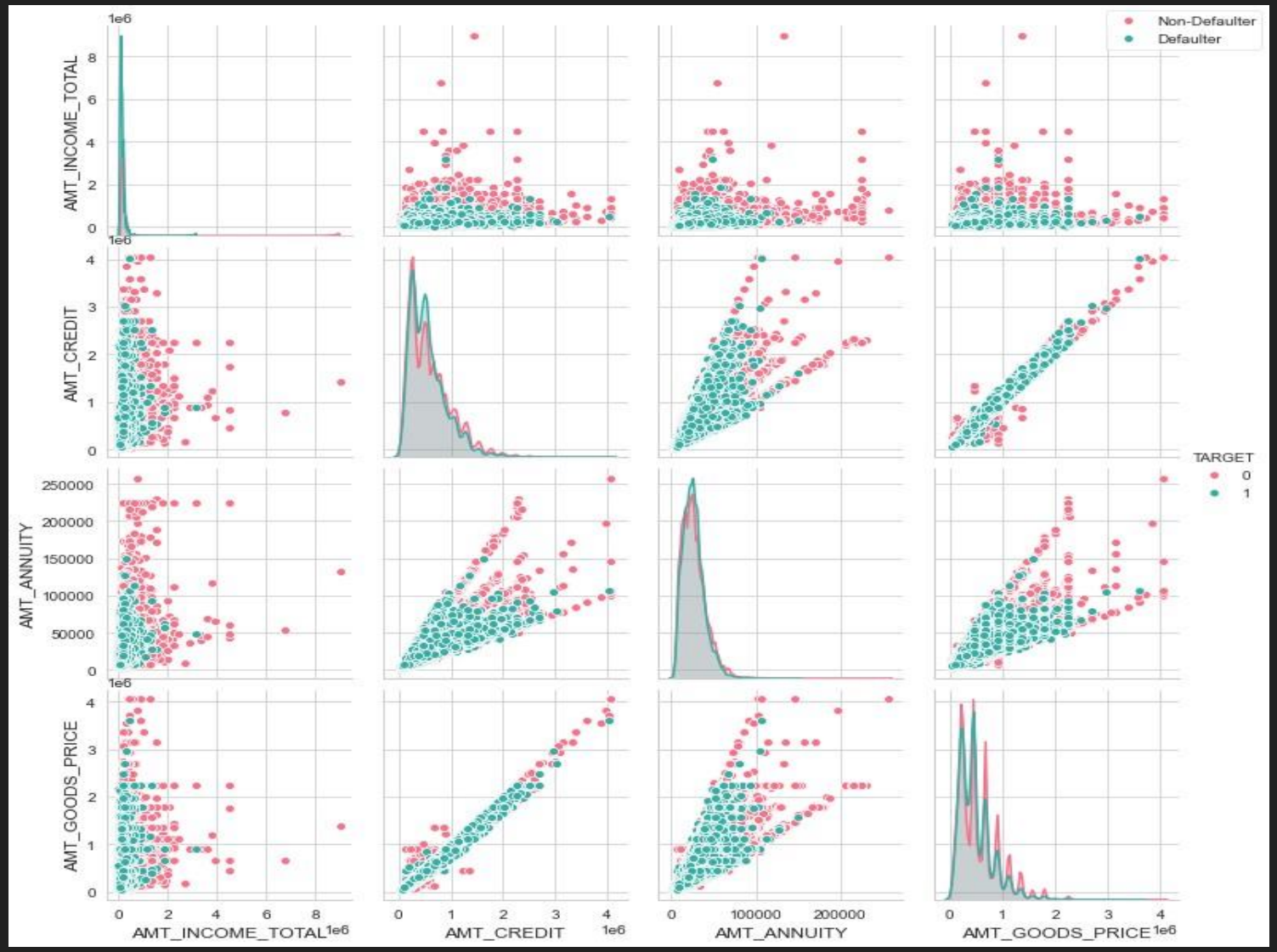


- Most no of loans are given for goods price below 10 lakhs.
- Most people pay annuity below 50000 for the credit loan.
- Credit amount of the loan is mostly
- less then 10 lakhs.
- The Non-defaulters and Defaulters distribution overlap in all the plots and hence we cannot use any of these variables in isolation to make a decision.

BIVARIATE ANALYSIS

Numerical Variables

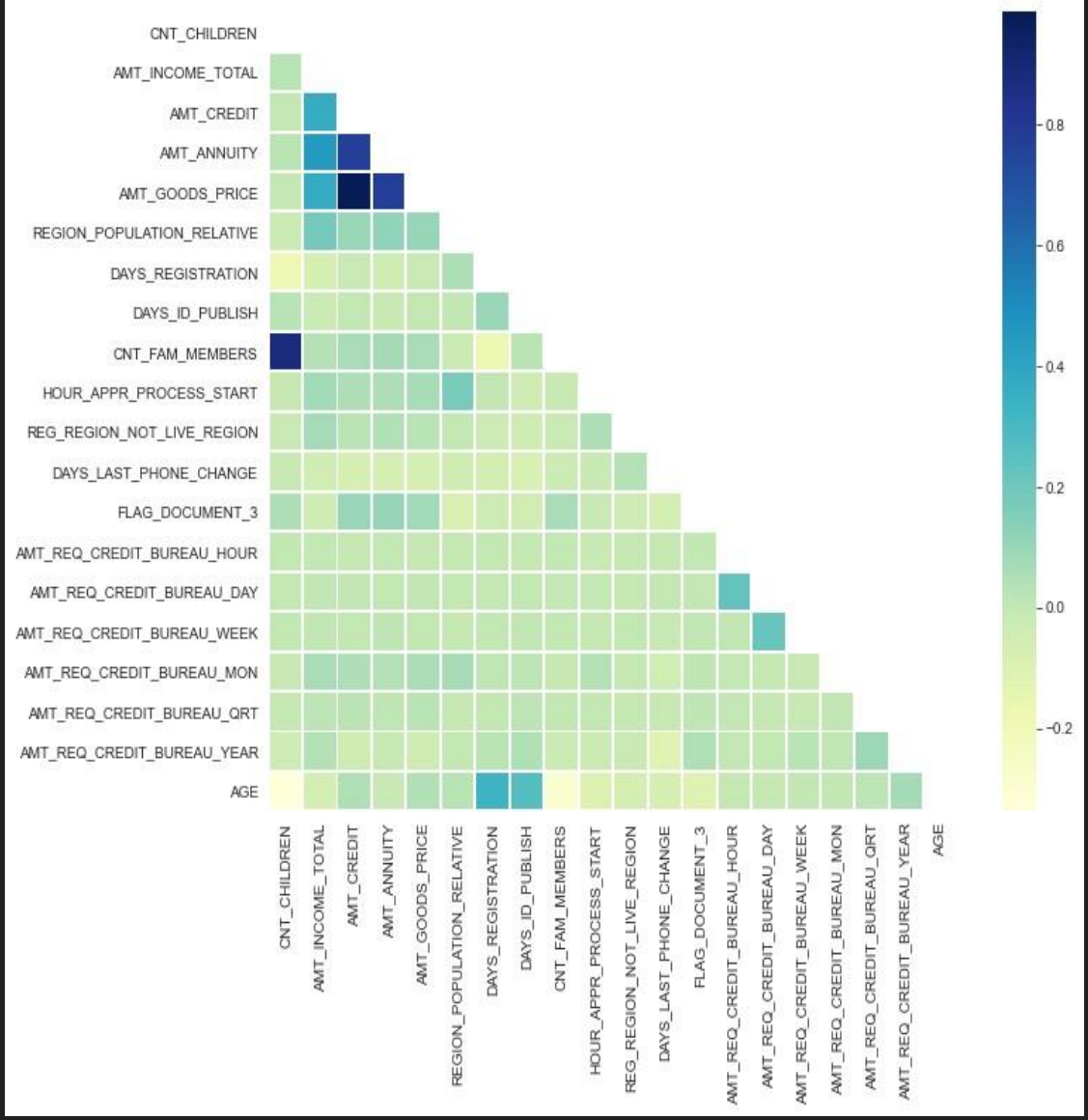
PAIR PLOT OF NUMERICAL VARIABLES



- AMT_CREDIT and AMT_GOODS_PRICE are highly correlated, as the points are forming a straight line, thus showing a linear relationship between the two.
- We can see that as AMT_CREDIT & AMT_GOODS_PRICE exceeds 3M, the proportion of defaulters decreases significantly.
- When AMT_ANNUITY > 150K & AMT_CREDIT > 3M, the percentage of defaulters decreases.

CORRELATION

NON-DEFAULTERS

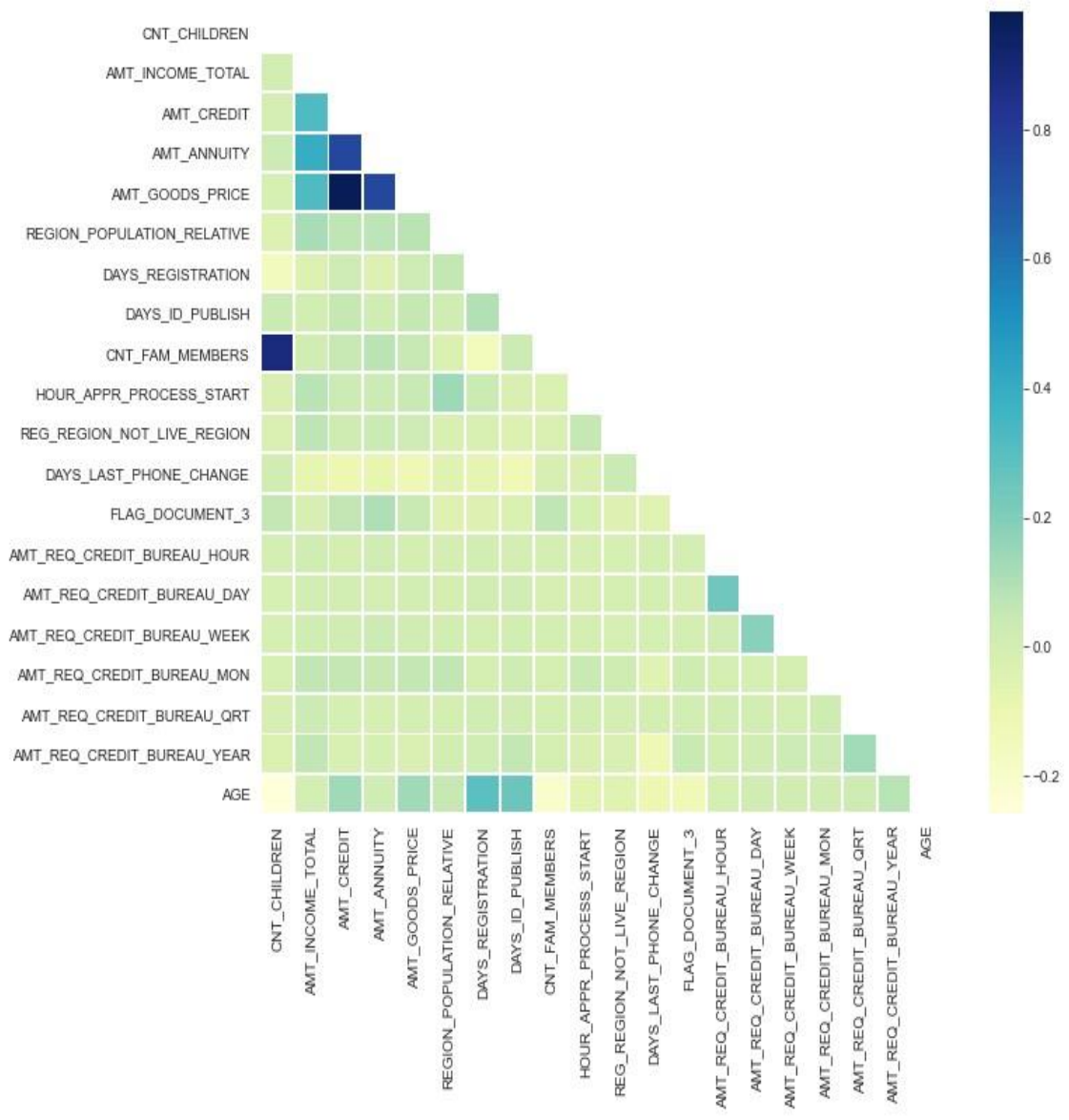


TOP 10 CORRELATION

Column 1	Column 2	Correlation
AMT_CREDIT	AMT_GOODS_PRICE	0.987022
CNT_CHILDREN	CNT_FAM_MEMBERS	0.878573
AMT_GOODS_PRICE	AMT_ANNUITY	0.776462
AMT_CREDIT	AMT_ANNUITY	0.771344
AMT_ANNUITY	AMT_INCOME_TOTAL	0.447743
AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.376369
AMT_CREDIT	AMT_INCOME_TOTAL	0.369424
AGE	CNT_CHILDREN	0.336938
DAYS_REGISTRATION	AGE	0.333033
AGE	CNT_FAM_MEMBERS	0.285818

- These variables are intercorrelated with each other:
  - AMT_CREDIT
  - AMT_INCOME_TOTAL
  - AMT_GOODS_PRICE
  - AMT_ANNUITY
- A high degree correlation can be seen between CNT_CHILDREN and CNT_FAM_MEMBERS.

DEFAULTERS

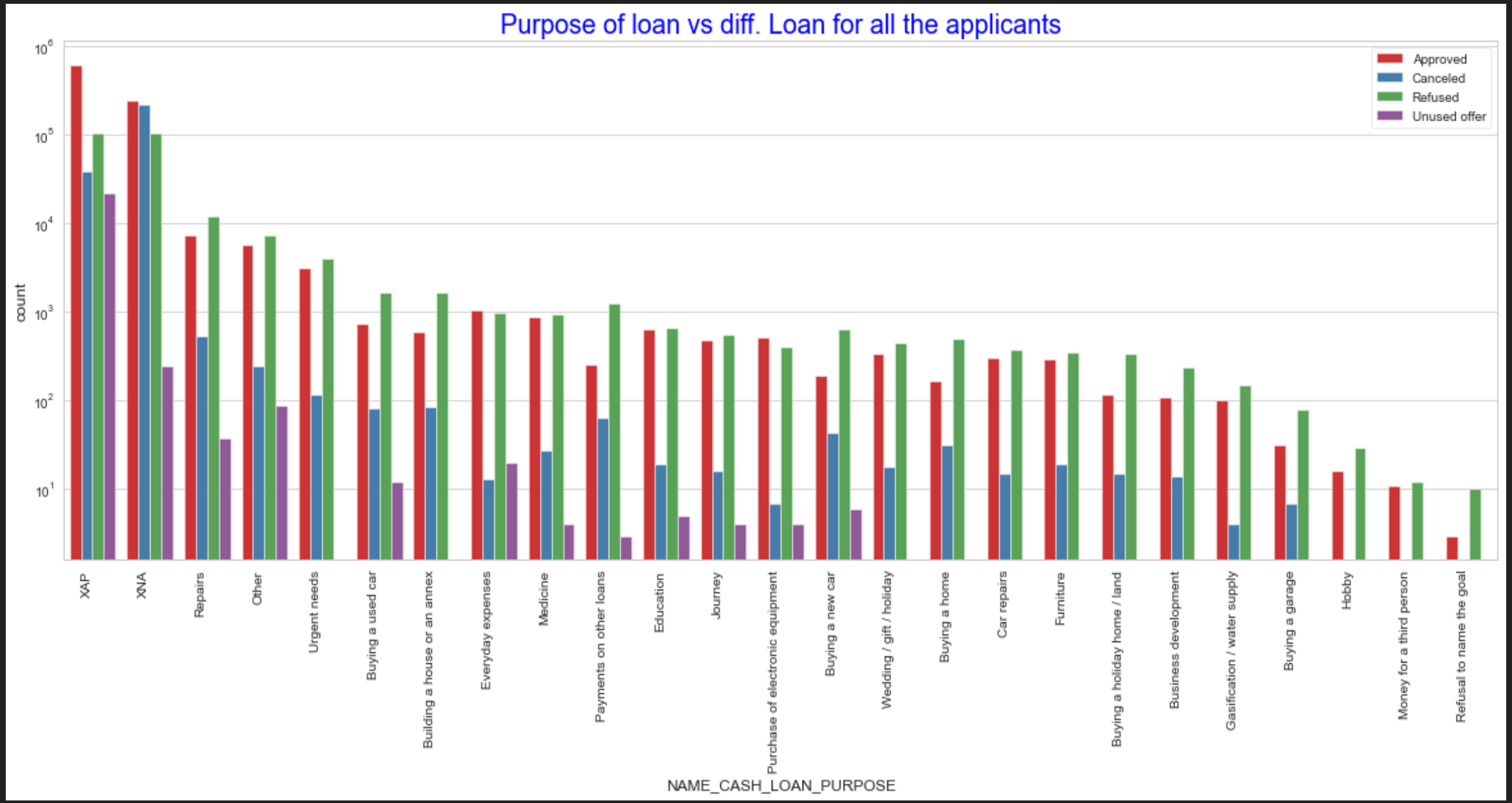


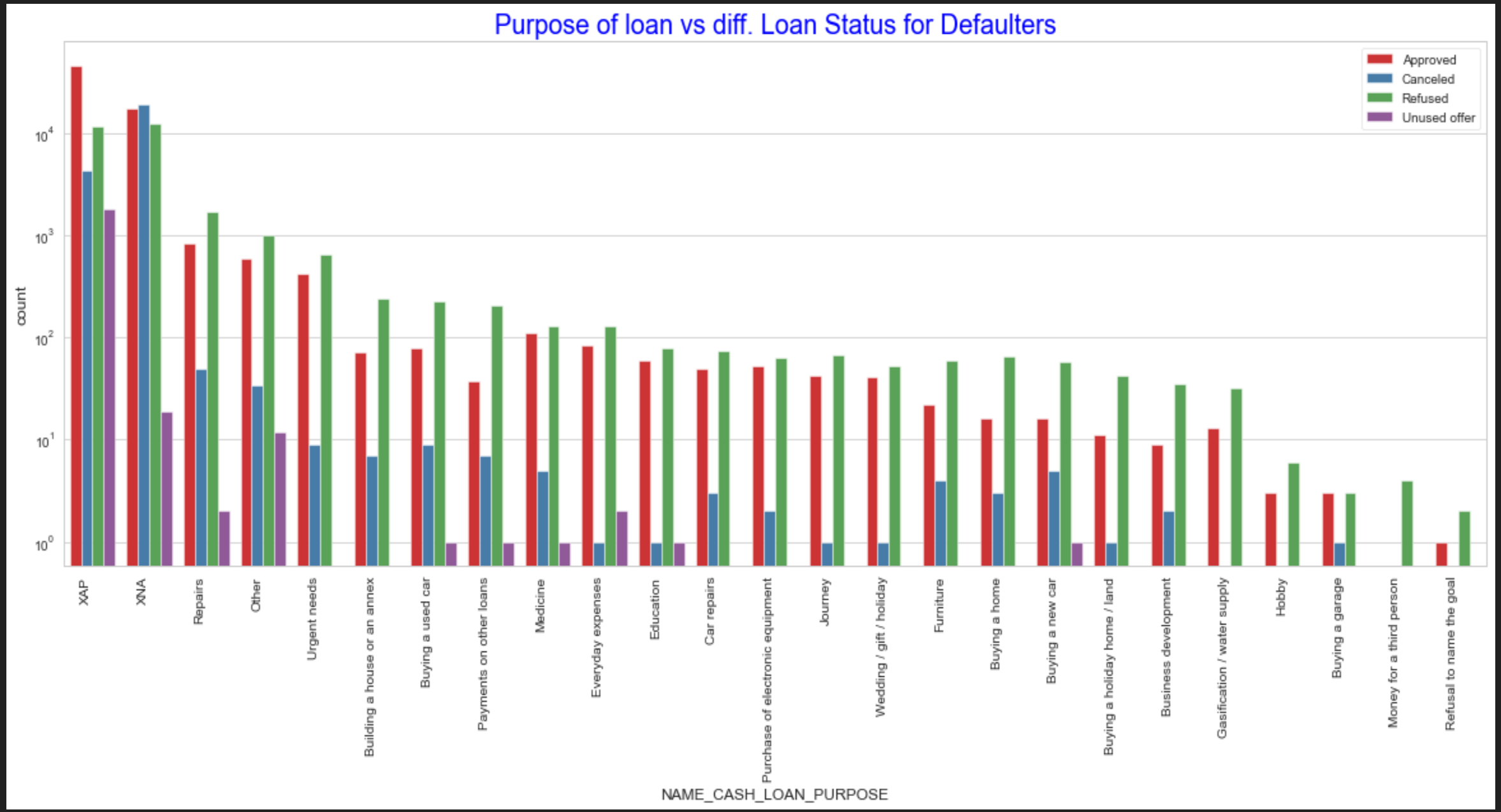
TOP 10 CORREALTION

Column 1	Column 2	Correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998270
AMT_CREDIT	AMT_GOODS_PRICE	0.982784
CNT_FAM_MEMBERS	CNT_CHILDREN	0.885481
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.869015
AMT_ANNUITY	AMT_GOODS_PRICE	0.752295
AMT_CREDIT	AMT_ANNUITY	0.752195
AMT_ANNUITY	AMT_INCOME_TOTAL	0.398260
OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.337383
DEF_30_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.334029
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.327456

- These variables are intercorrelated with each other:
  - AMT_CREDIT
  - AMT_INCOME_TOTAL
  - AMT_GOODS_PRICE
  - AMT_ANNUITY
- A high degree correlation can be seen between CNT_CHILDREN and CNT_FAM_MEMBERS.

MERGED DATAFRAME

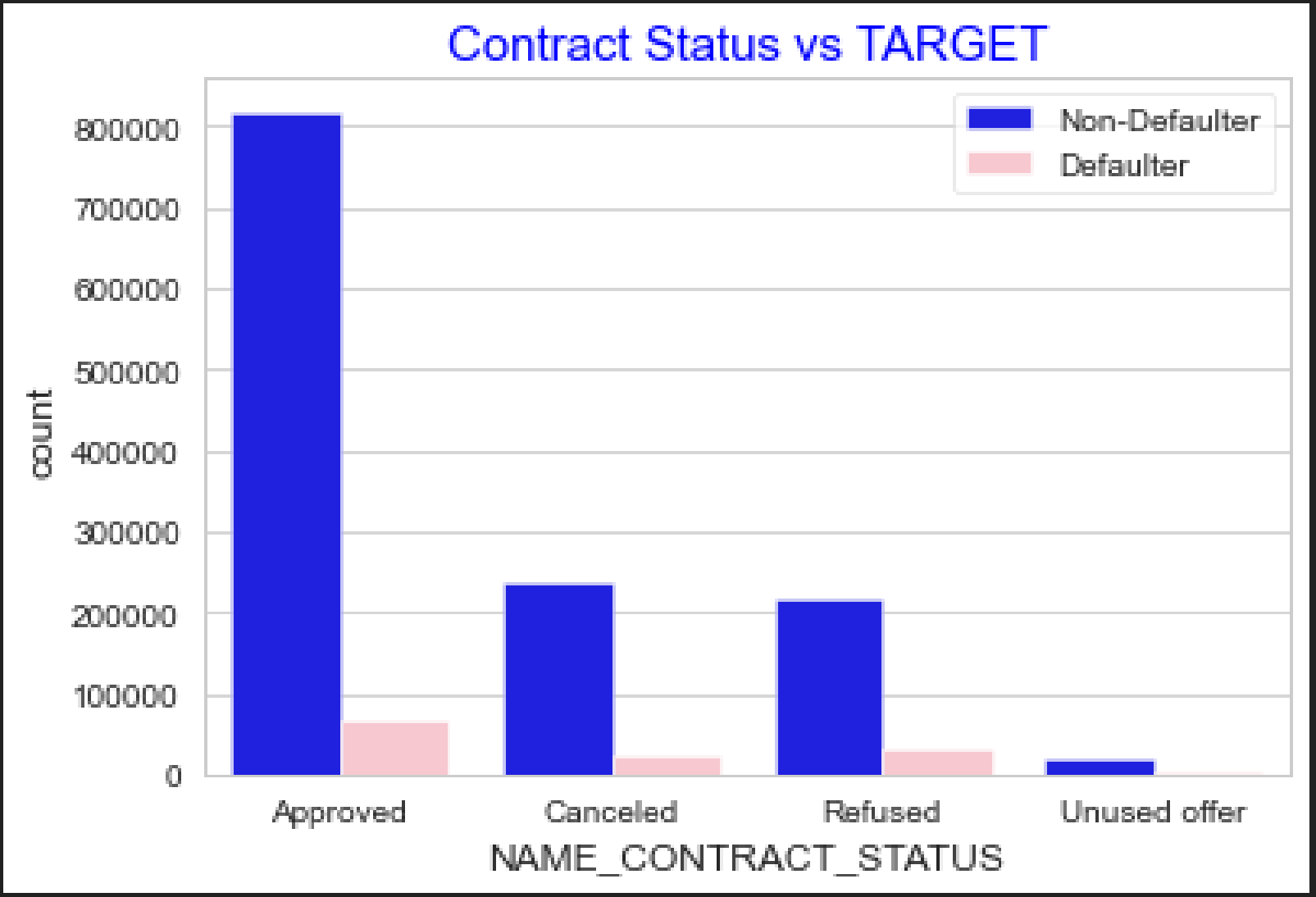

Purpose of loan vs diff. Loan for all the applicants
Approved
Canceled
Refused
Unused offer
count
10^6
10^5
10^4
10^3
10^2
10^1
XAP
XNA
Repairs
Other
Urgent needs
Buying a used car
Building a house or an annex
Everyday expenses
Medicine
Payments on other loans
Education
Journey
Purchase of electronic equipment
Buying a new car
Wedding / gift / holiday
Buying a home
Car repairs
Furniture
Buying a holiday home / land
Business development
Gasification / water supply
Buying a garage
Hobby
Money for a third person
Refusal to name the goal
NAME_CASH_LOAN_PURPOSE


Purpose of loan vs diff. Loan Status for Defaulters
Approved
Canceled
Refused
Unused offer
count
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

OBSERVATIONS FROM PREVIOUS GRAPHS:

- Purpose of loan is unknown for very high number of applicants.
- Banks have rejected high number of applications taken for Repair and Other purposes, also applicants have refused these offers more number of times.
- There are few places where proportion of Non-Defaulters is significantly higher.
- They are-
  - 'Buying a garage'
  - 'Business development'
  - 'Buying land'
  - 'Buying a new car'
  - 'Education'
- Hence we can focus on these purposes for which default percentage is less.

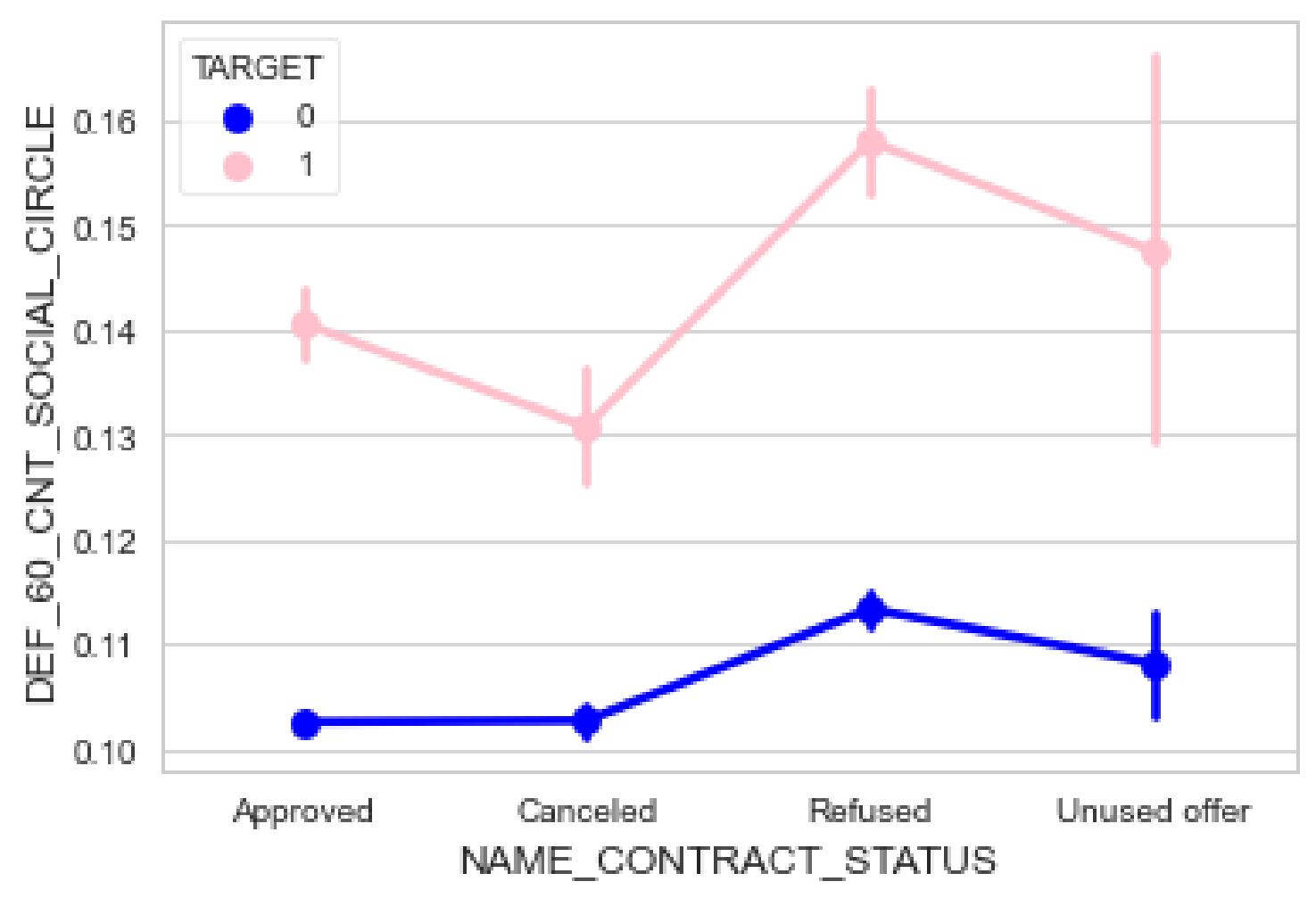
CONTRACT STATUS vs LOAN REPAYMENT STATUS



- 90% of clients who cancelled their loan previously have successfully repaid their current loan, bank should record the reason for cancellation of these clients and bring in some policies accordingly as so they can be potential customers for the bank.

- Major portion of clients who have been previously refused a loan have repaid back the loan in current case. Refusal reason should be recorded for further analysis as these clients would turn into potential repaying customer.

SOCIAL CIRCLE vs TARGET LOAN REPAYMENT STATUS



- Clients who have average of 0.13 or higher DEF_60_CNT_SOCIAL_CIRCLE score tend to default more and hence client's social circle has to be analyzed before providing the loan.

CONCLUSIONS

FACTORS WHICH INDICATE THAT THE PERSON WILL BE A NON-DEFAULTER ARE:

- LOANS_EDUCATION_TYPE: People who have Academic Degrees have less defaults as compared to other people.
- NAME_INCOME_TYPE: Students and Businessmen have no defaults.
- NAME_FAMILY_STATUS: Widows are least likely to default on the loans.
- AMT_INCOME_TOTAL: Customers who have income in the range of 700K and 800K are least likely to default.
- ORGANIZATION_TYPE: Clients with Trade Type:4 & 6, Industry Type:12 Transport Type: 1 are least likely to default.

FACTORS WHICH INDICATE THAT THEPERSON WILL BE A DEFAULTER ARE:

- CODE_GENDER: Male Customers are more likely to default than females.
- NAME_FAMILY_STATUS: People who are single or have done Civil Marriage are more likely to default.
- NAME_INCOME_TYPE: Clients who are on maternity leave or are Unemployed are most likely to default on their payments.
- NAME_HOUSING_TYPE: People who live with in rented apartments or with their parents are more likely to default on loan.
- OCCUPATION_TYPE: Low-Skilled Laborer's are most likely to default on the loan.
- AGE_GROUP: People in the Age Group of 20-40 have are most likely to default on the loan.

The following variables indicate the people from the below categories tend to default on the loan which can be prevented by providing them loans at higher interest rate to cushion any default risk and further, preventing any business loss:

- NAME_HOUSING_TYPE: People living in the Rented Apartments are the ones who take a large number of loans but also have a higher default rate. So, completely shutting them off would be loss for the business.
- AMT_CREDIT: There are a large chunk of people who earn in the range of 100K and 200K and also those people have a higher default rate. So, keeping a higher interest rate would make sense.
- NAME_EDUCATION_TYPE: People with Secondary/Special education applied for most Percentage of loans and thus, keeping a nominal interest rate for those folks would help in the business.
- NAME_CASH_LOAN_PURPOSE: Loans taken for the purpose of Repairs have a higher default rate and hence, the bank charges a higher interest rate for that client which the client cannot bear and hence, cancel the loan in other stages of the application.
- OCCUPATION_TYPE: There are quite a few low-skilled and other labourers who apply for the loans and these people also have a higher default rate. Since, these people also have low incomes so the bank should keep a decent amount of interest rate which would not lead to carry out any defaults for these people.

MORE SUGGESTIONS

- There are a significant number of people who had cancelled the loan application but have now turned into Non-Defaulters. So, the bank could collect the information on what made them cancel the service and improve on those services for the clients.

- Almost 85% of the clients who were refused the loan in the previous application have repayed the loan or have no difficulty in repaying the loan. Thus, refusing these clients any further would be bad for business and hence, bank should recheck the reasons behind the refusals for these customers.

THANK YOU