CASE STUDY FOR BANK LOAN DEFAULTERS

ARJUN AMLA

Methodology

Ask questions and define the problem.

Prepare data by collecting and storing the information.

Process data by cleaning and checking the information.

Analyze data to find patterns, relationships, and trends.

Share data with your audience.

Act on the data and use the analysis results.

Methodology Description

Problem Statement

Given a consumer finance company specializes in lending loan (different types) to it's customers.

When a customer applies for a loan, the company needs to make a decision for approval on the basis of the applicant's profile details.

Two major considerations :

1. If the applicant ends up not repaying the loan amount (defaulter), the bank will have to bear a financial loss

2. If the applicant is likely to pay back the loans, then rejecting their application would result as a loss of business to the company

ASK

Business Objectives

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments 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 utilize this knowledge for its portfolio and risk assessment.

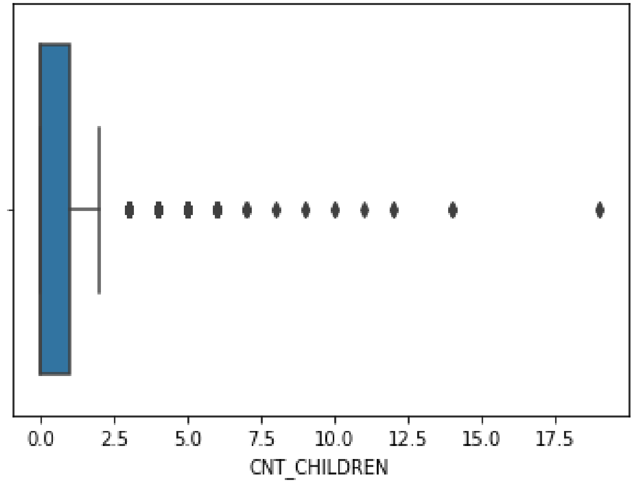
PREPARE

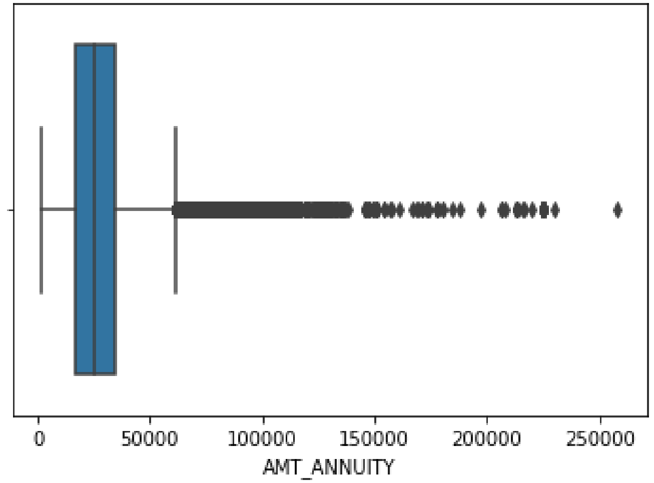
1. In our case study the data is gathered with the organisation which make it's the Internal as well as First- Party data.
2. As this data comes within the organisation its very credible
3. I can also see that the data-is as unbiases as possible

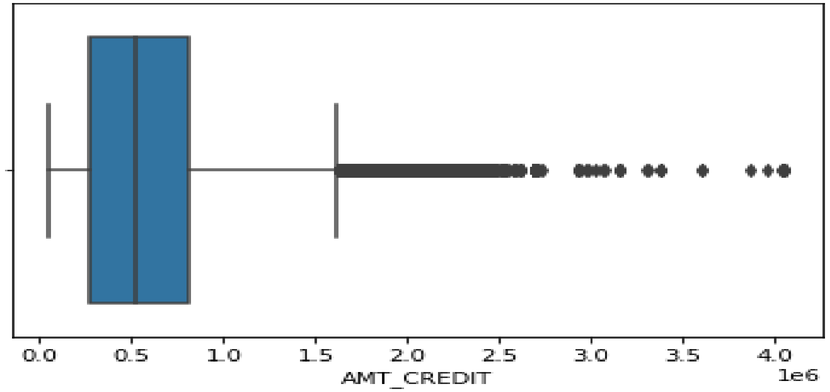
PROCESS

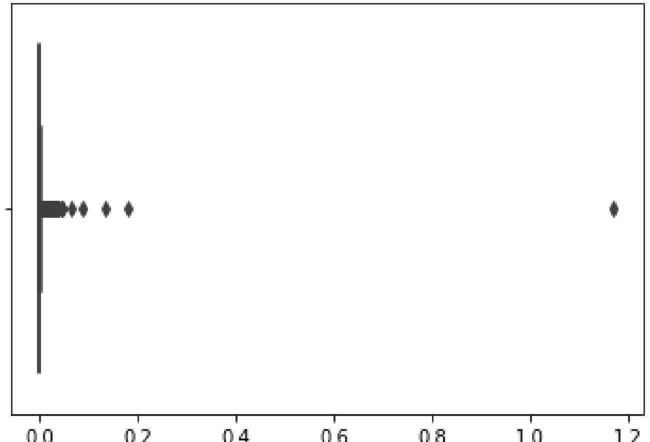
OUTLIERS

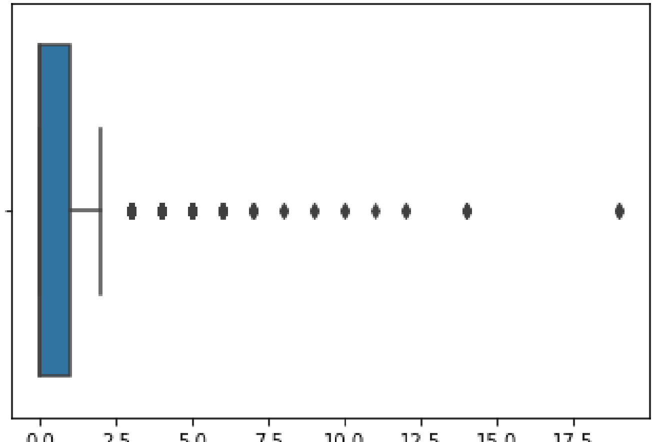
It can be seen that in current application data **AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN** have some number of outliers. **AMT_INCOME_TOTAL** has huge number of outliers which indicate that few of the loan applicants have high income when compared to the others.











MISSING VALUES

First all the columns with missing values>40% are removed from data frame

Then remaining columns are imputed by the median as most of our data contains outliers

OWN_CAR_AGE	65.990810
EXT_SOURCE_1	56.381073
APARTMENTS_AVG	50.749729
BASEMENTAREA_AVG	58.515956
YEARS_BEGINEXPLUATATION_AVG	48.781019
YEARS_BUILD_AVG	66.497784
COMMONAREA_AVG	69.872297
ELEVATORS_AVG	53.295980
ENTRANCES_AVG	50.348768
FLOORSMAX_AVG	49.760822
FLOORSMIN_AVG	67.848630
LANDAREA_AVG	59.376738
LIVINGAPARTMENTS_AVG	68.354953
LIVINGAREA_AVG	50.193326
NONLIVINGAPARTMENTS_AVG	69.432963
NONLIVINGAREA_AVG	55.179164
APARTMENTS_MODE	50.749729
BASEMENTAREA_MODE	58.515956
YEARS_BEGINEXPLUATATION_MODE	48.781019
YEARS_BUILD_MODE	66.497784

SK_ID_CURR	0.000000
TARGET	0.000000
NAME_CONTRACT_TYPE	0.000000
CODE_GENDER	0.000000
FLAG_OWN_CAR	0.000000
FLAG_OWN_REALTY	0.000000
CNT_CHILDREN	0.000000
AMT_INCOME_TOTAL	0.000000
AMT_CREDIT	0.000000
AMT_ANNUITY	0.003902
AMT_GOODS_PRICE	0.090403
NAME_TYPE_SUITE	0.420148
NAME_INCOME_TYPE	0.000000
NAME_EDUCATION_TYPE	0.000000
NAME_FAMILY_STATUS	0.000000
NAME_HOUSING_TYPE	0.000000
REGION_POPULATION_RELATIVE	0.000000
DAYS_BIRTH	0.000000
DAYS_EMPLOYED	0.000000

In [52]:

```
app_df.AMT_ANNUITY.isnull().sum()
```

Out[52]:

12

Filling the missing values with the median

In [53]:

```
app_df.AMT_ANNUITY=app_df.AMT_ANNUITY.fillna(app_df.AMT_ANNUI
TY.median())
```

In [54]:

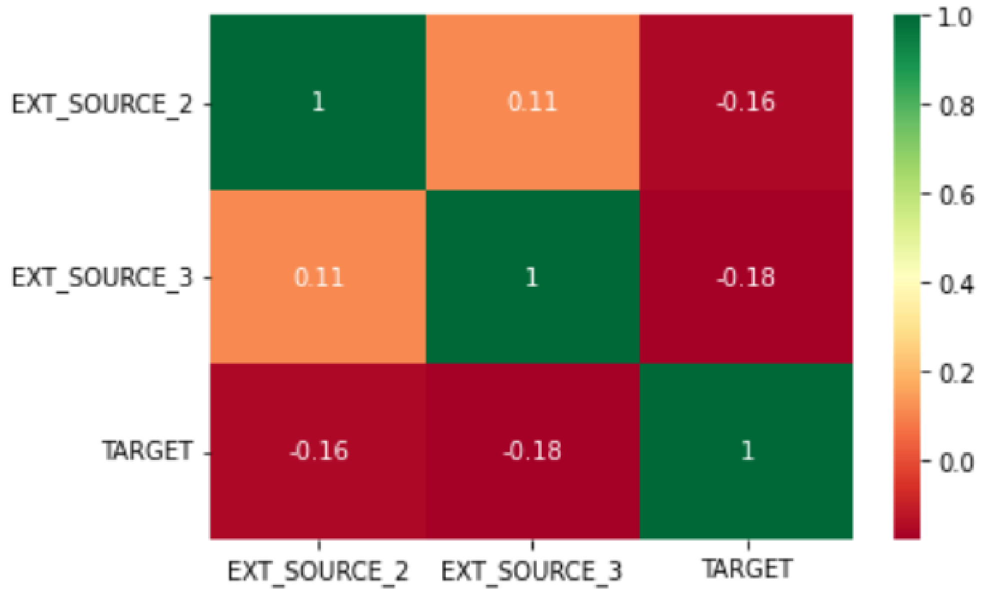
```
app_df.AMT_ANNUITY.isnull().sum()
```

Out[54]:

0

COLUMN REMOVAL

Bases on their correlation with the target variable some of the column are removed



DATA TYPE CORRECTION

Some of the object and numerical columns are convert into category column

In [194]:

```
#Conversion of Object and Numerical columns to Categorical Colu
mns
categorical_columns = ['NAME_CONTRACT_TYPE','CODE_GENDER','NA
ME_INCOME_TYPE','NAME_EDUCATION_TYPE',
                       'NAME_FAMILY_STATUS','NAME_HOUSING_TYP
E','WEEKDAY_APPR_PROCESS_START',
                       'ORGANIZATION_TYPE','FLAG_OWN_CAR','FL
AG_OWN_REALTY','LIVE_CITY_NOT_WORK_CITY',
                       'REG_CITY_NOT_LIVE_CITY','REG_CITY_NOT
_WORK_CITY','REG_REGION_NOT_WORK_REGION',
                       'LIVE_REGION_NOT_WORK_REGION','REGION_
RATING_CLIENT','WEEKDAY_APPR_PROCESS_START',
                       'REGION_RATING_CLIENT_W_CITY'
                      ]
for col in categorical_columns:
    app_df[col] =pd.Categorical(app_df[col])
```

```
 28   HOUR_APPR_PROCESS_START       306477 non-null  int64
 29   REG_REGION_NOT_LIVE_REGION    306477 non-null  int64
 30   REG_REGION_NOT_WORK_REGION    306477 non-null  categor
y
 31   LIVE_REGION_NOT_WORK_REGION   306477 non-null  categor
y
 32   REG_CITY_NOT_LIVE_CITY        306477 non-null  categor
y
 33   REG_CITY_NOT_WORK_CITY        306477 non-null  categor
y
 34   LIVE_CITY_NOT_WORK_CITY       306477 non-null  categor
y
 35   ORGANIZATION_TYPE             306477 non-null  categor
y
 36   OBS_30_CNT_SOCIAL_CIRCLE      306477 non-null  float64
 37   DEF_30_CNT_SOCIAL_CIRCLE      306477 non-null  float64
 38   OBS_60_CNT_SOCIAL_CIRCLE      306477 non-null  float64
 39   DEF_60_CNT_SOCIAL_CIRCLE      306477 non-null  float64
```

BINNING

Binning is applied on some numerical column is order to have a better understanding of the data

```
# Creating bins for Credit amount
app_df['AMT_CREDIT']=app_df['AMT_CREDIT']/100000

bins = [0,1,2,3,4,5,6,7,8,9,10,100]
slots = ['0-100K','100K-200K', '200k-300k','300k-400k','400k-
500k','500k-600k','600k-700k','700k-800k',
       '800k-900k','900k-1M', '1M Above']

app_df['AMT_CREDIT_RANGE']=pd.cut(app_df['AMT_CREDIT'],bins=b
ins,labels=slots)
```

```
app_df['AMT_CREDIT_RANGE'].value_counts()
```

```
200k-300k    54501
1M Above     49877
500k-600k    34174
400k-500k    31928
100K-200K    29921
300k-400k    26245
600k-700k    23998
800k-900k    21733
700k-800k    19169
900k-1M       8927
0-100K        6004
Name: AMT_CREDIT_RANGE, dtype: int64
```

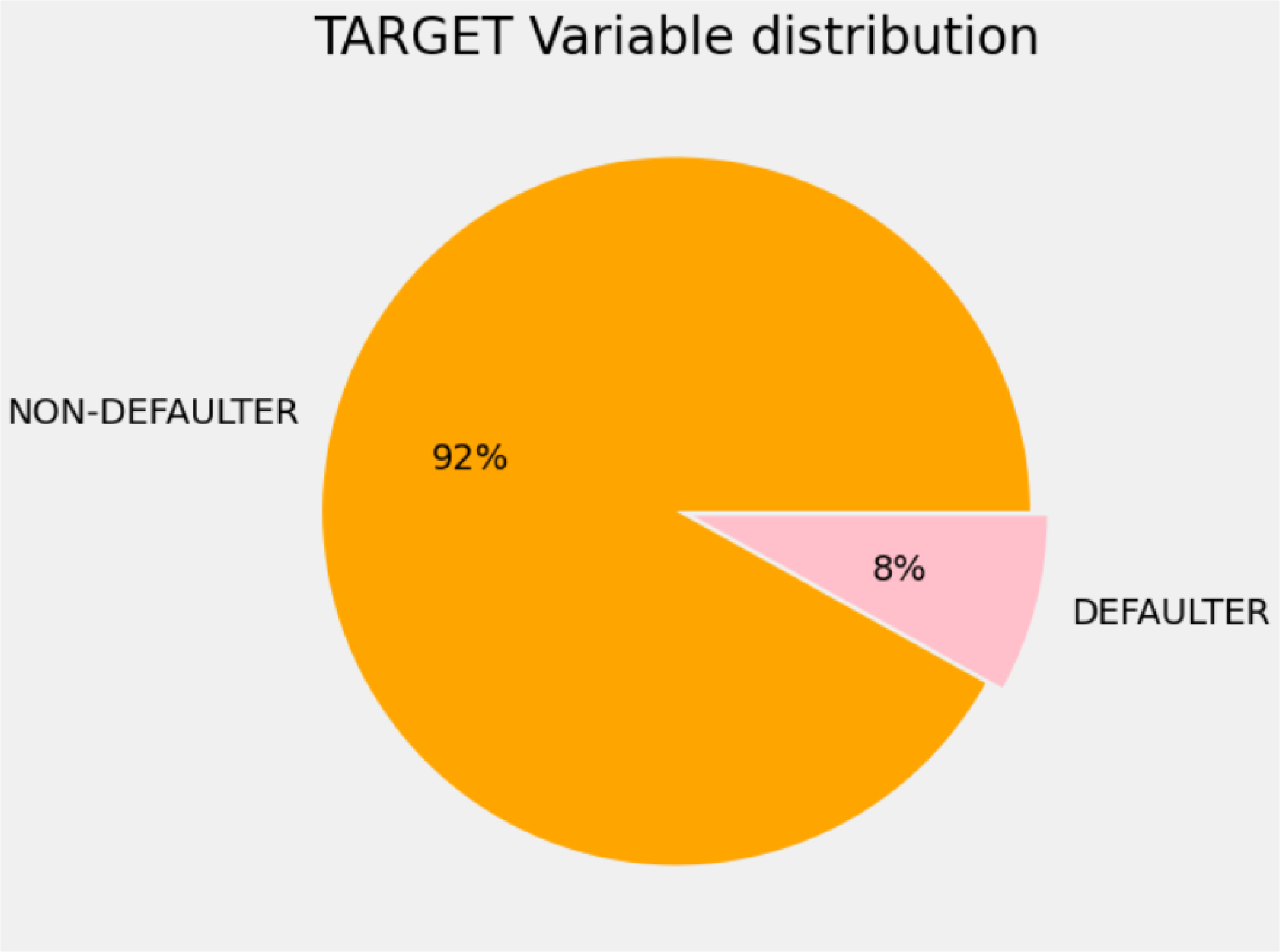
ANALYZE

All the analysis is done on the notebook attached with the presentation

SHARE

IMBALANCE IN DATA

Data is highly imbalanced because among all the applicants, we have 92% of the Applicants who were able to pay back the loan and 8% who defaulted.

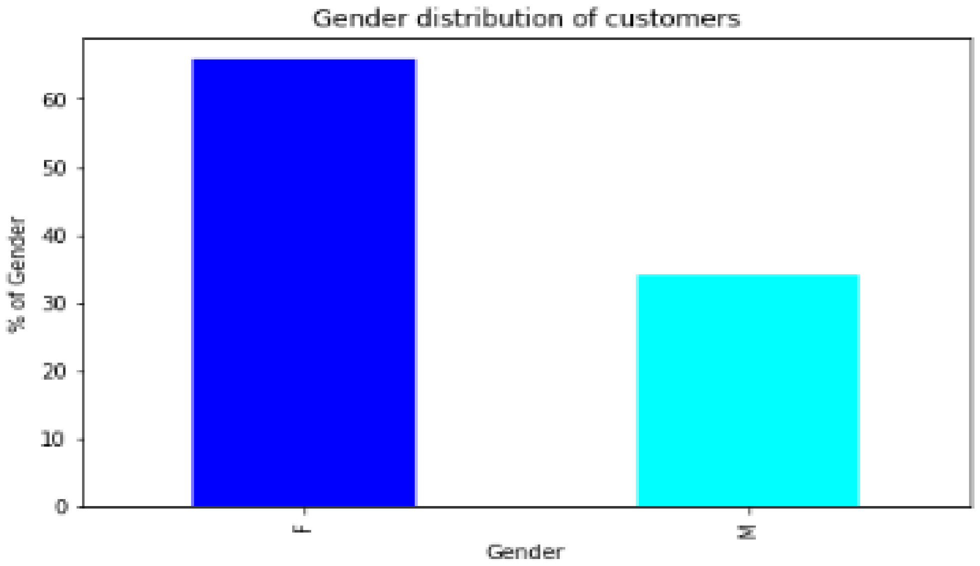


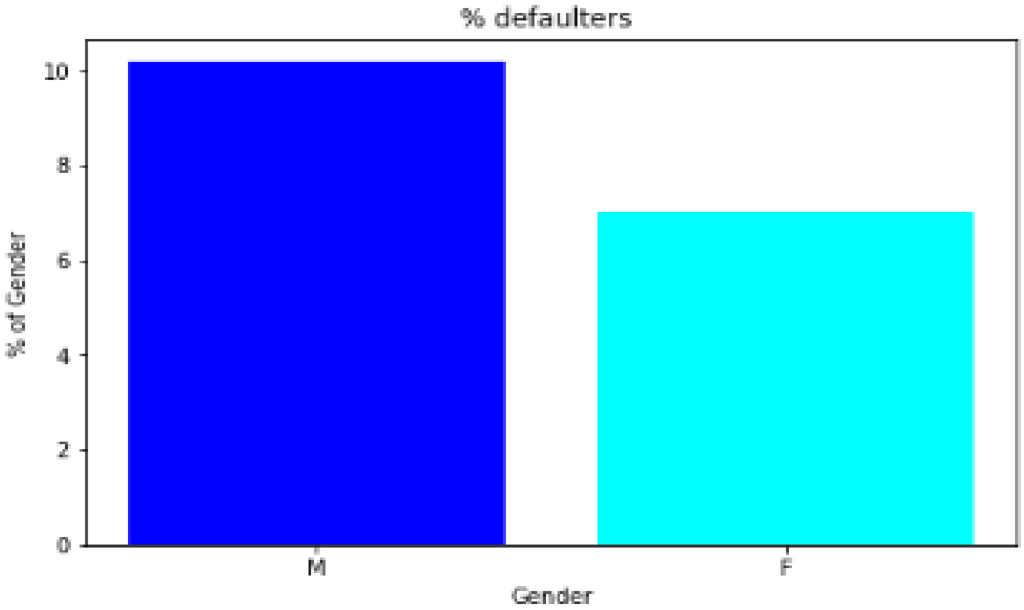
GENDER VS DEFAULTERS

There are more female than male

More no of female are have defaulted the payments due to there grater no

Men are tend to become more defaulters than women even with their small representation

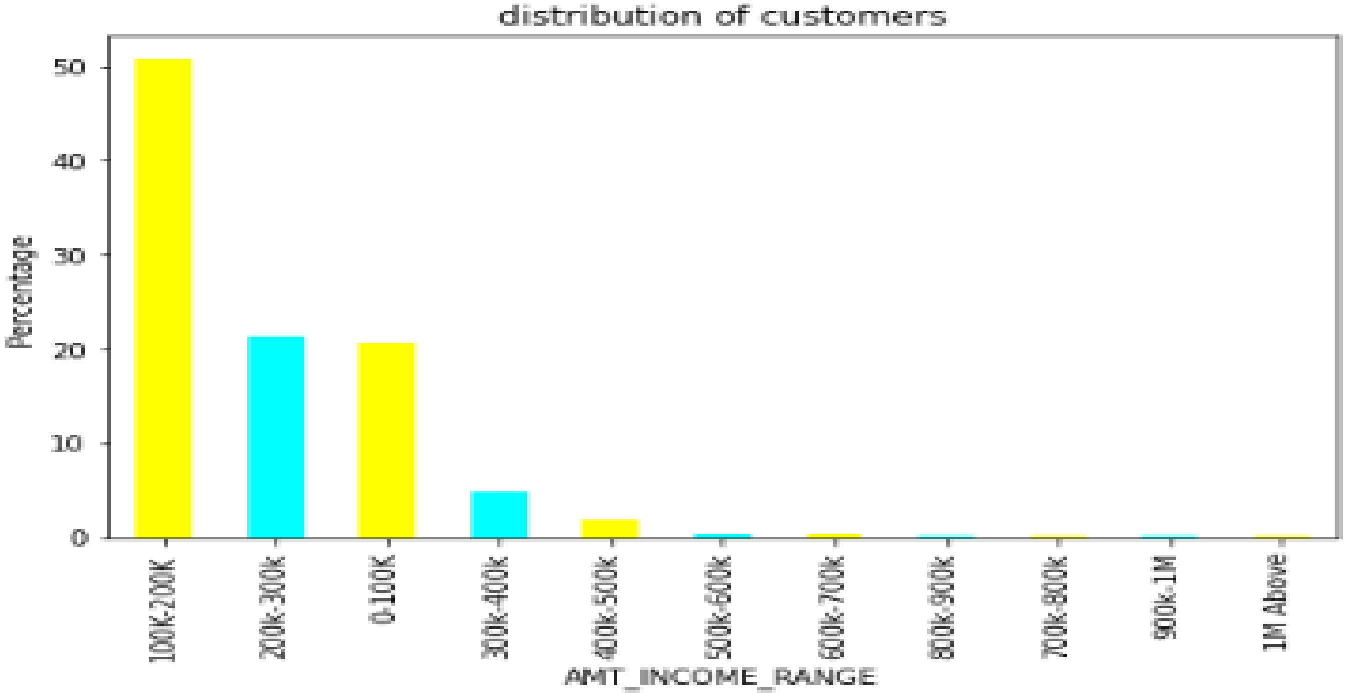


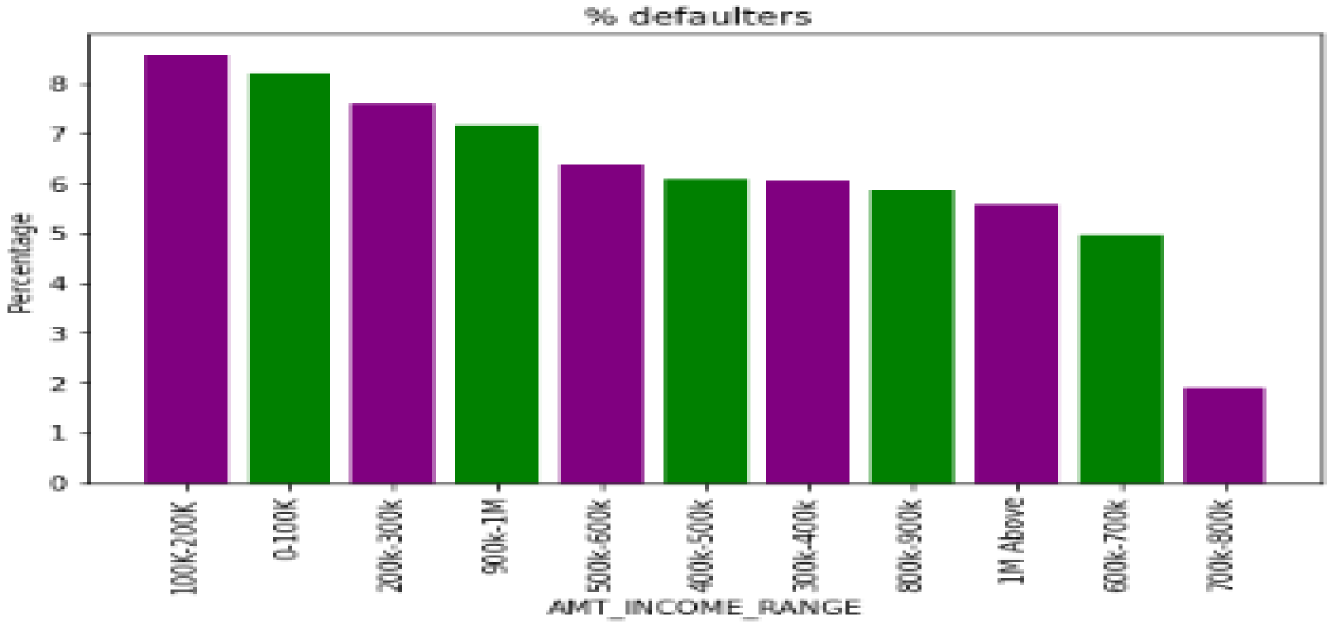


INCOME VS DEFAULTERS

Most of the loans were given to the people with salaries less than 300k

If the salary is less than 500k more chances of defaulting chances for defaulting

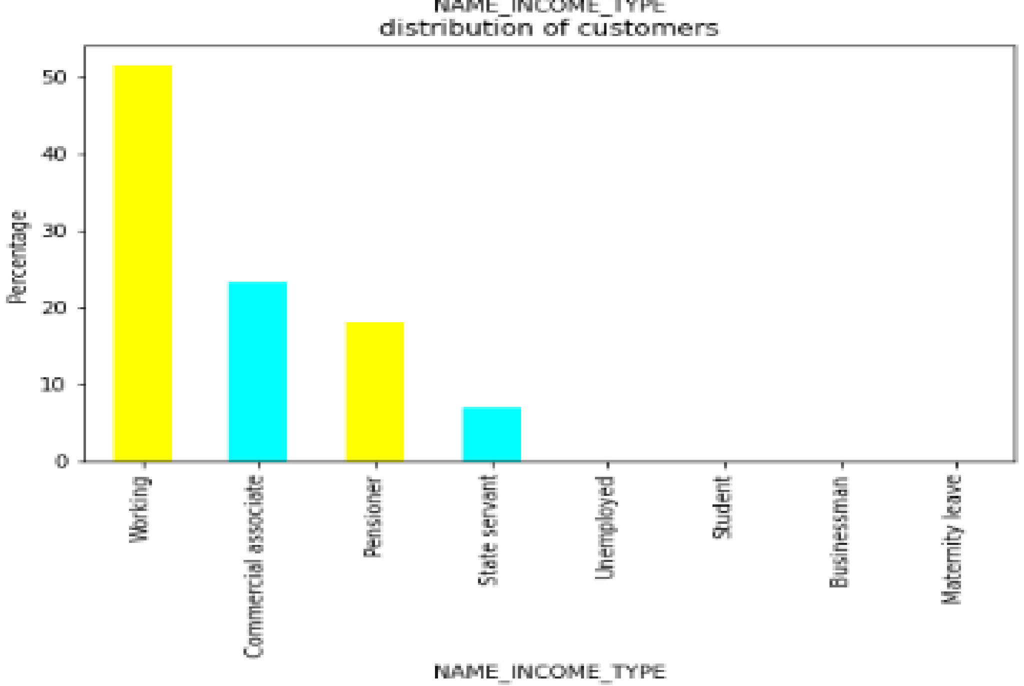


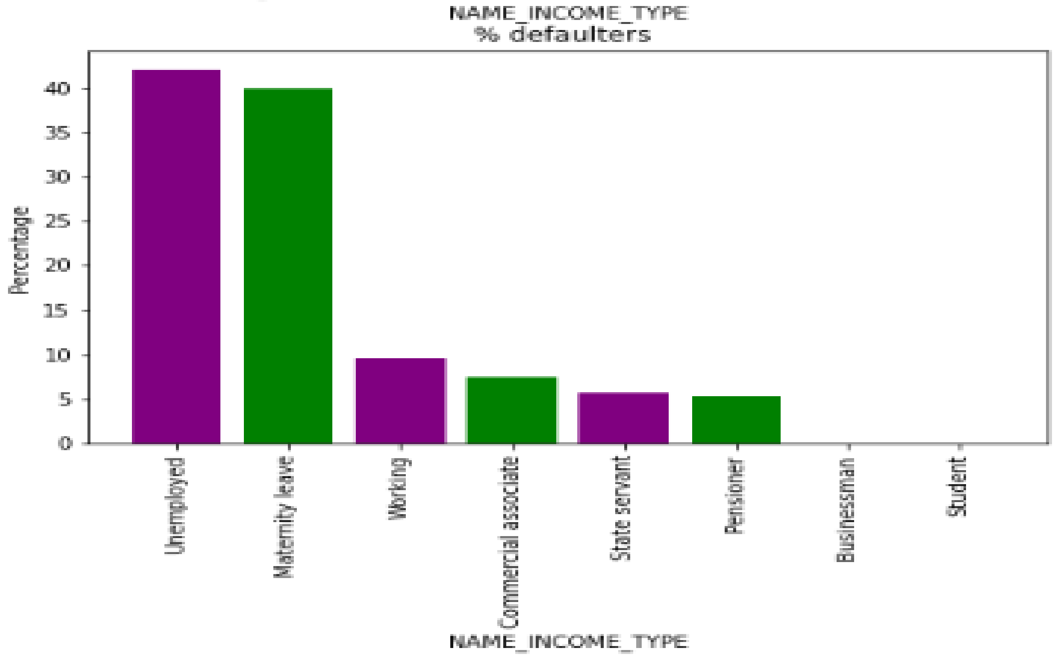


JOB VS DEFAULTERS

Most of the loan are given to working and commercial associates

People who are unemployed or are on maternity leave have high chances to default the payments

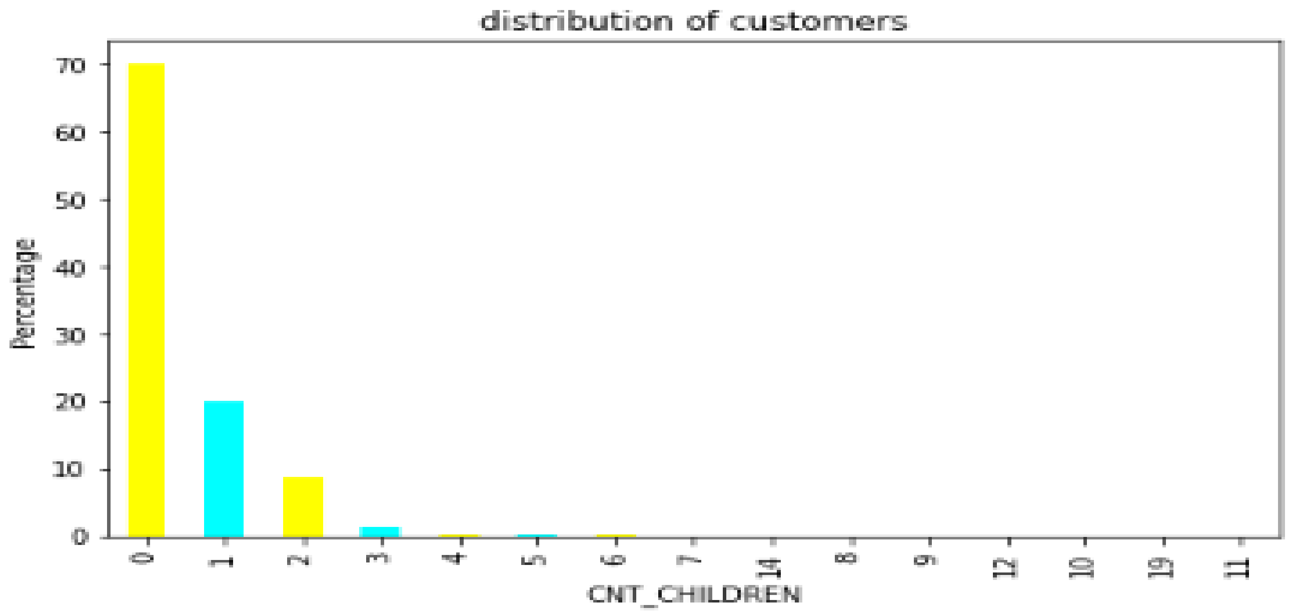


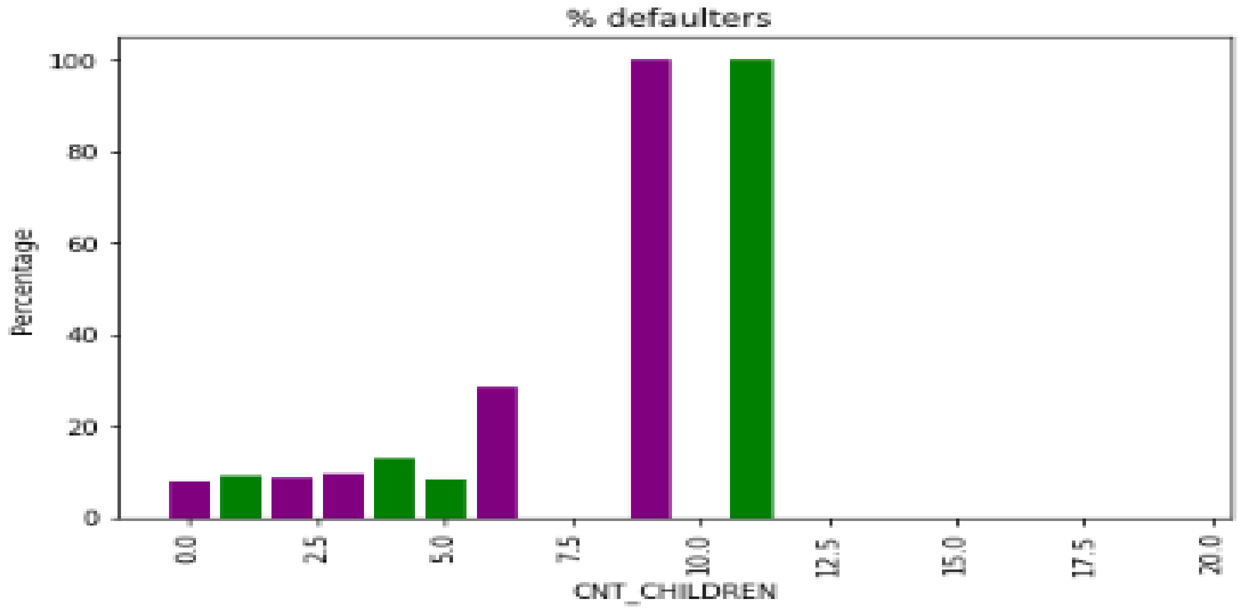


NUMBER OF CHILDERN VS DEFAULTERS

Mostly customers are having zero children

Parents with 6 kids and more are the highest number of defaulters but its worth noting that there number is very less

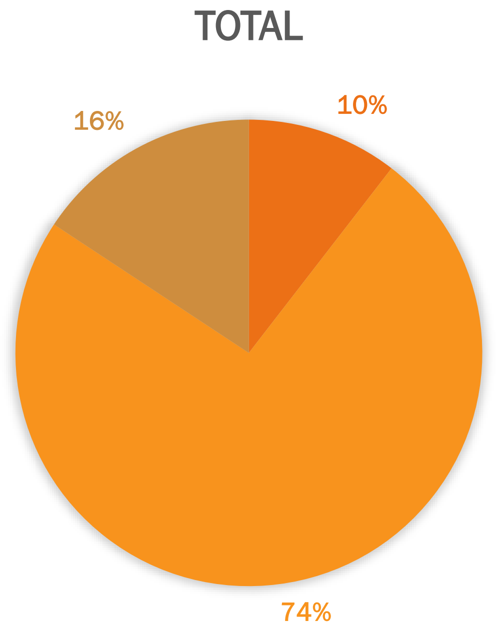


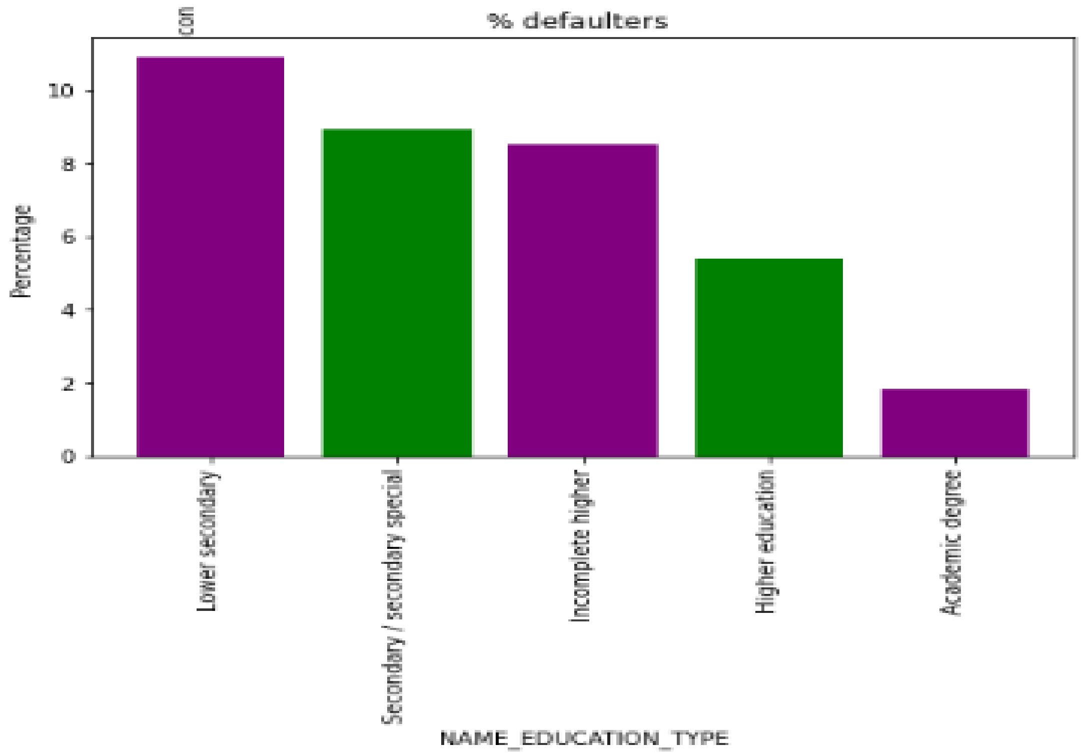


EDUCATION VS DEFAULTERS

Most of the loan are given to people with secondary education and higher education

People with Lower secondary education have more chances to be a defaulter

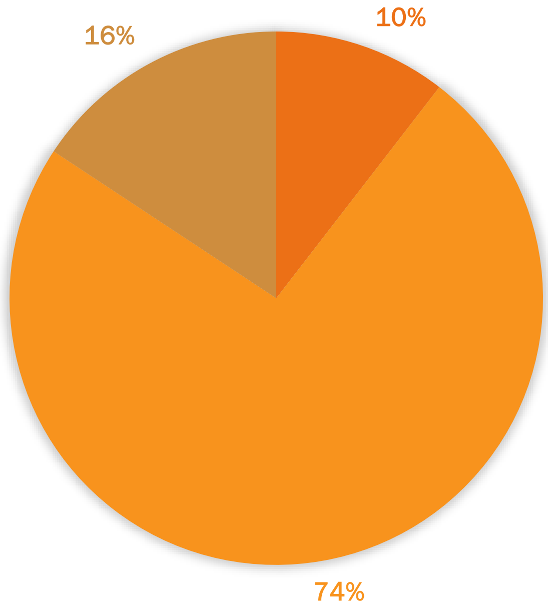


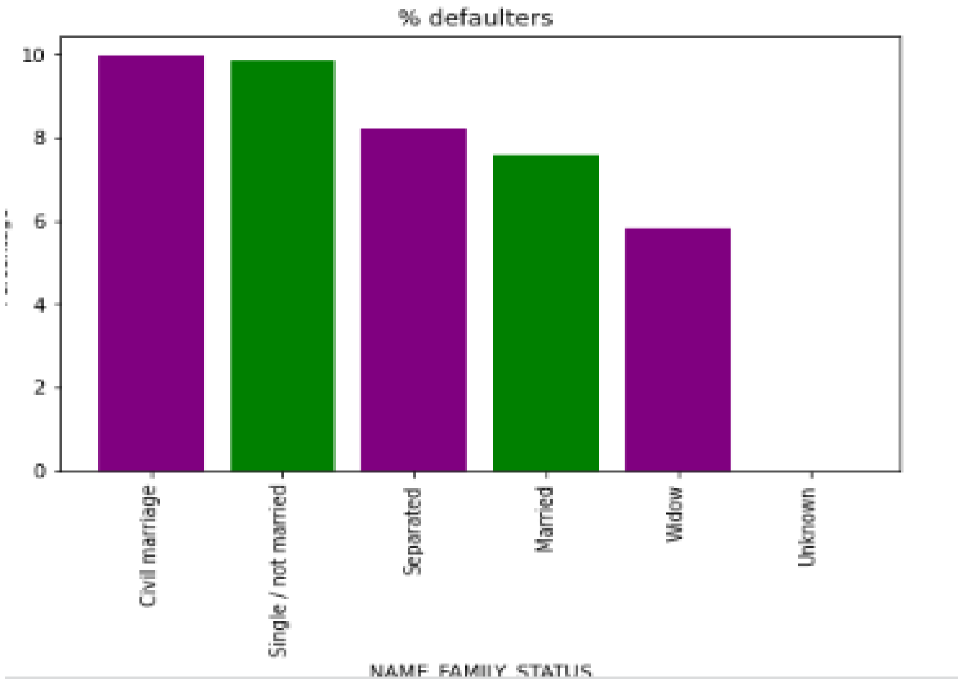


MARITAL STATUS VS DEFAULTERS

Most of the loans are given to the married people

People with civil marriage or single or separated are more likely to default

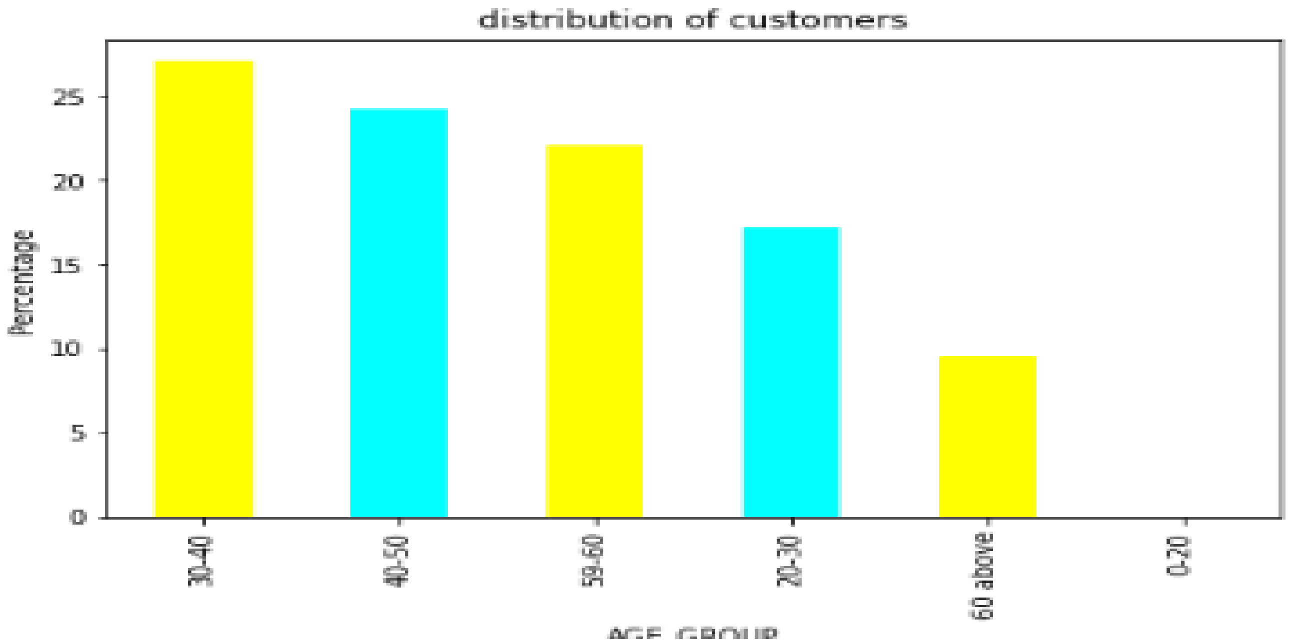


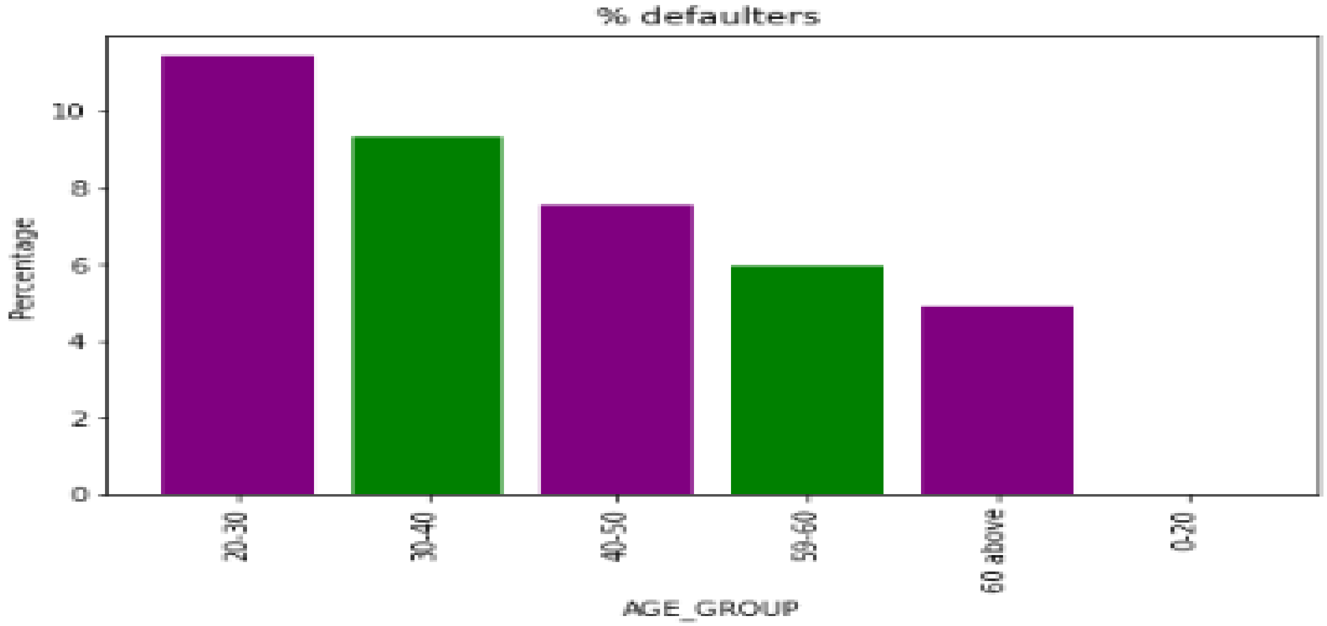


AGE VS DEFAULTERS

Most of the loans are given to people within age 30-40

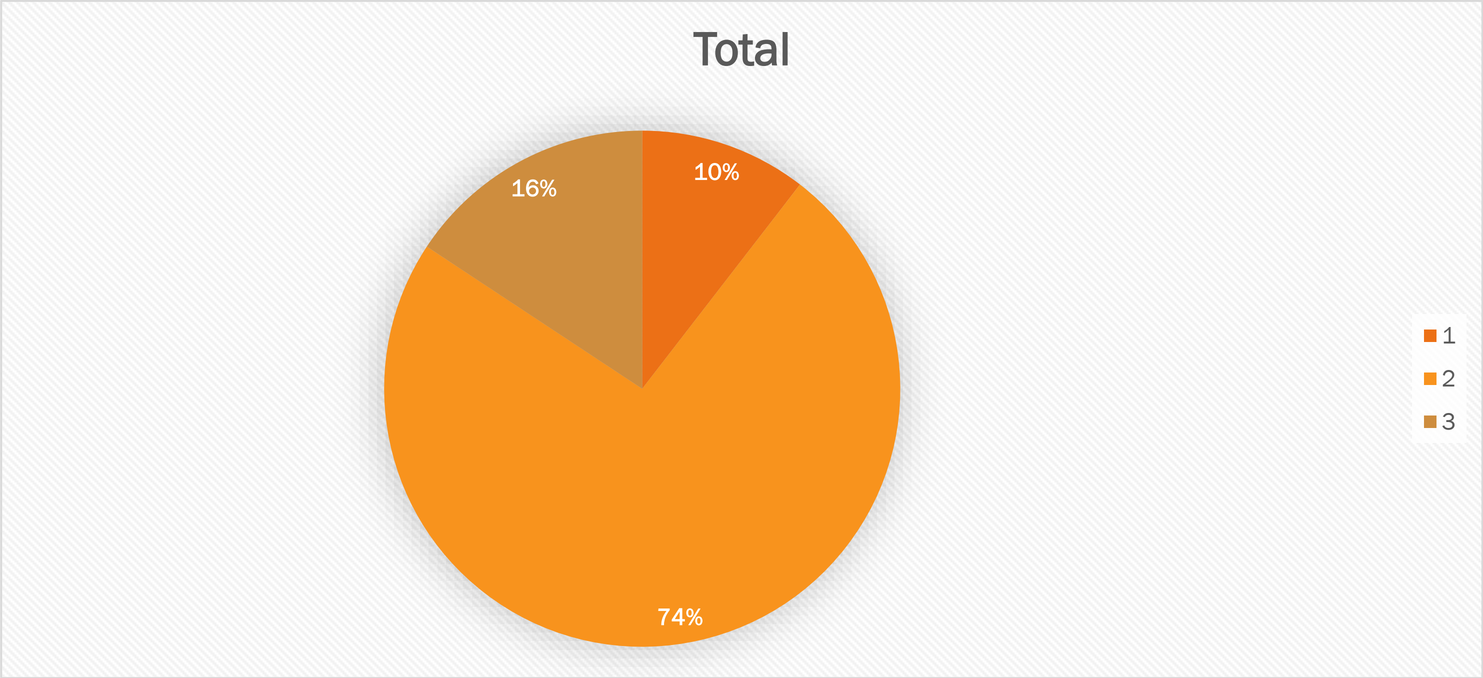
People in the age group range 20-40 have higher probability of defaulting People above age of 50 have low probability of defaulting

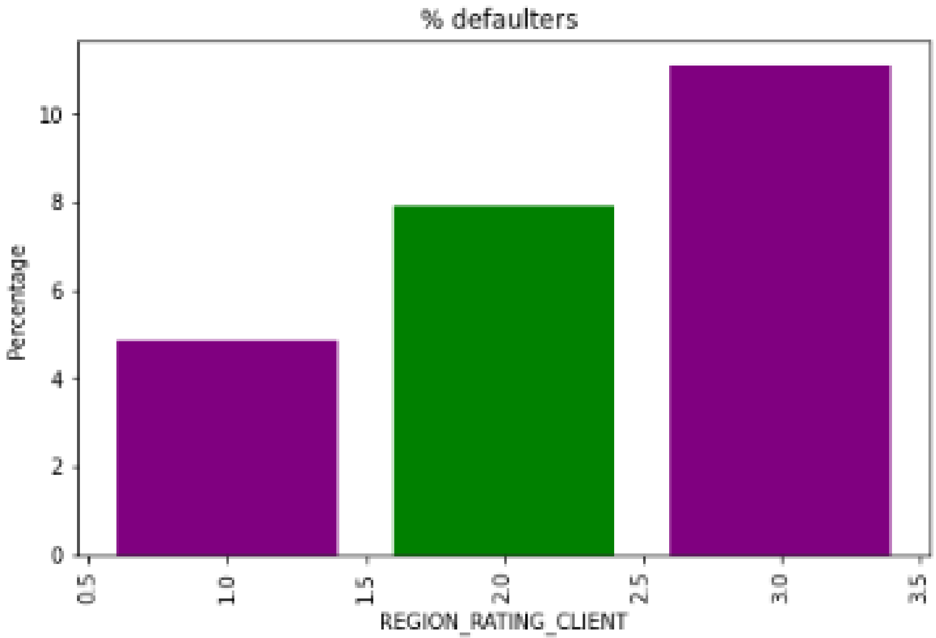




REGION VS DEFAULTERS

Most of the loans were given in region 2 but a great number of people from region 3 are defaulting

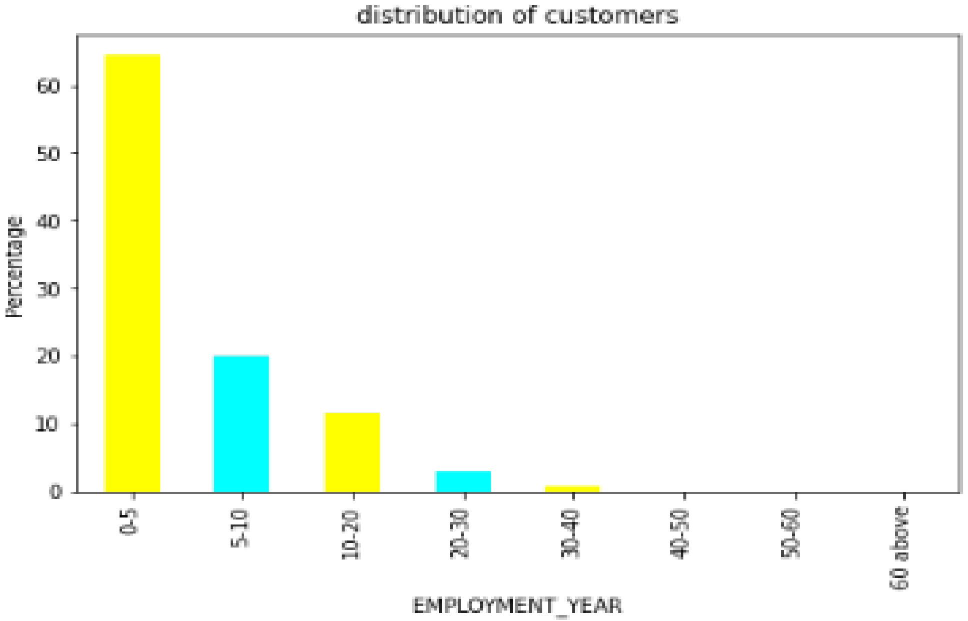


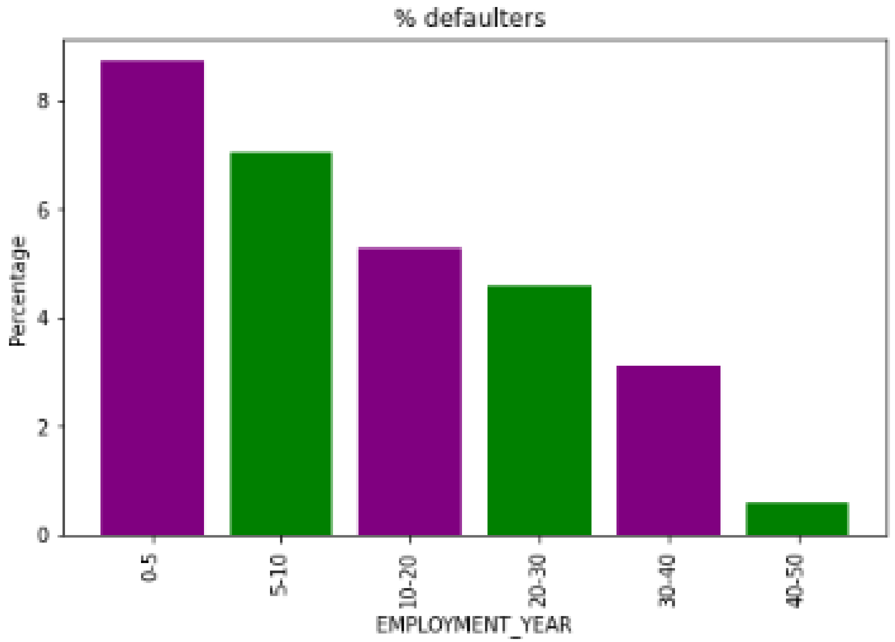


EMPLOYMENT TIME VS DEFAULTERS

Most of the loans are given to people in their early years of employment

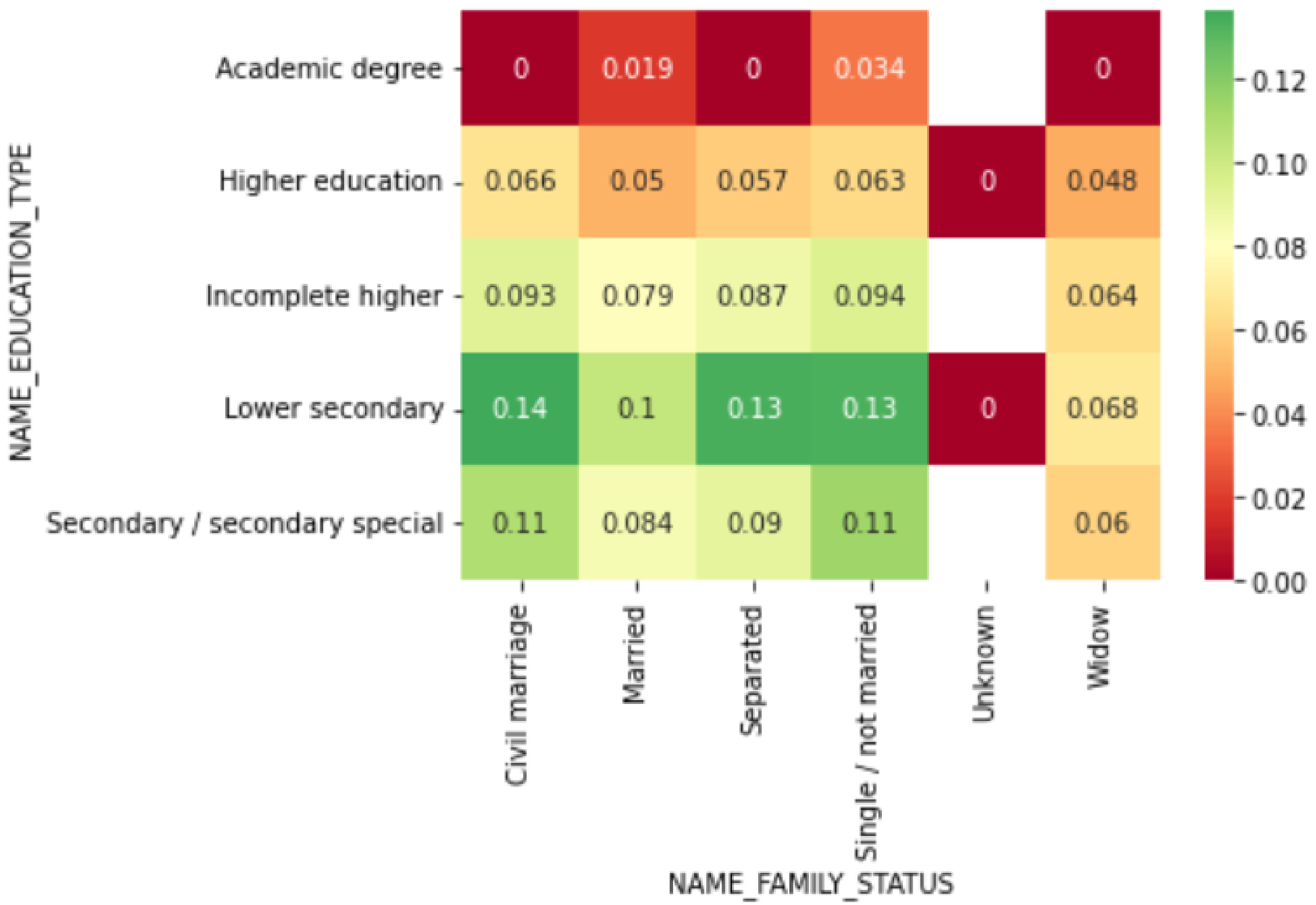
More the experience less the default percentage





EDUCATION VS MARITAL SATATUS

People with civil marriage and lower Sc eductaion likely to default more followed by lower Sc seperated and single people

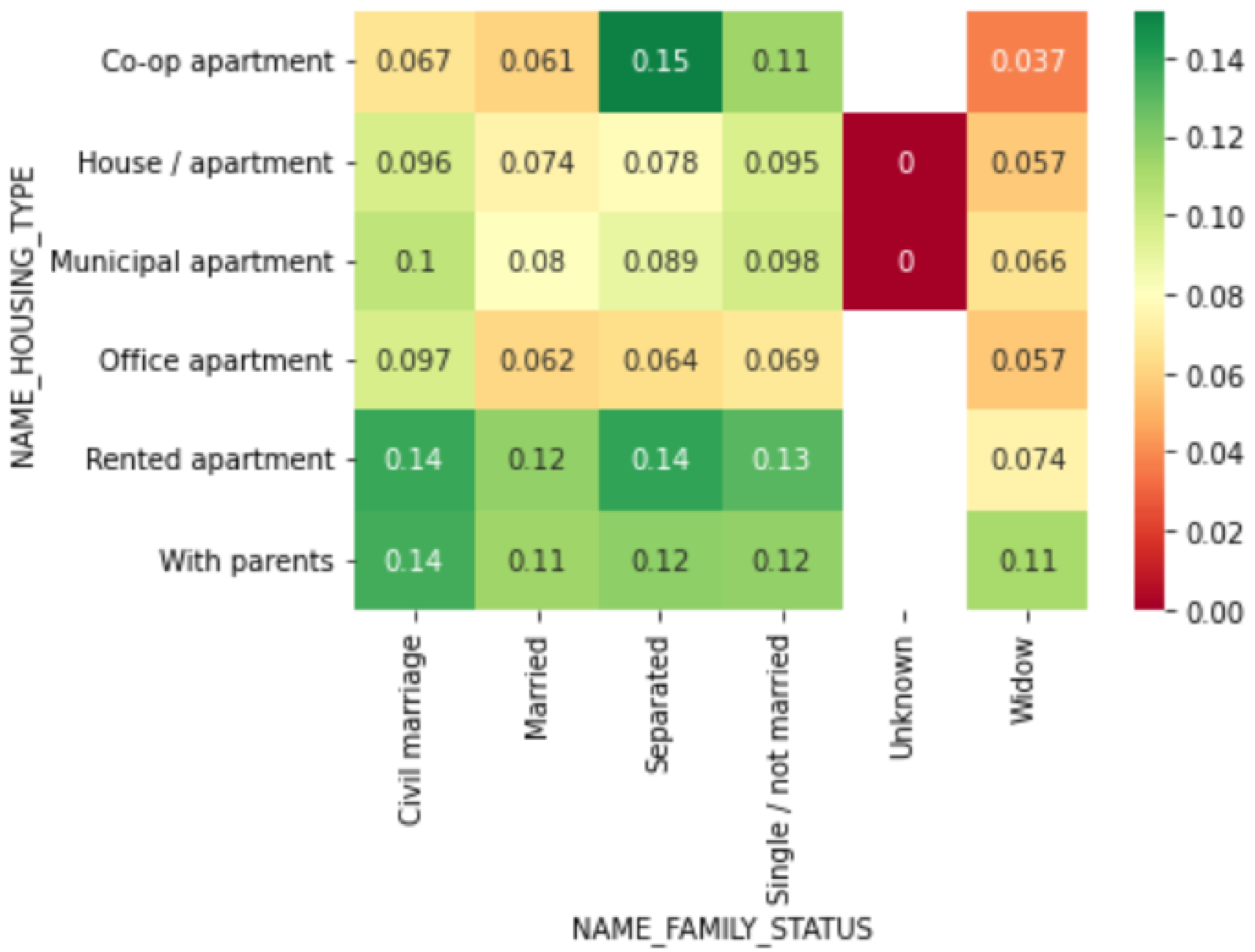


OCCUPATION VS MARITAL STATUS

Single,low skilled labours and widowed HR staff are likely to default More

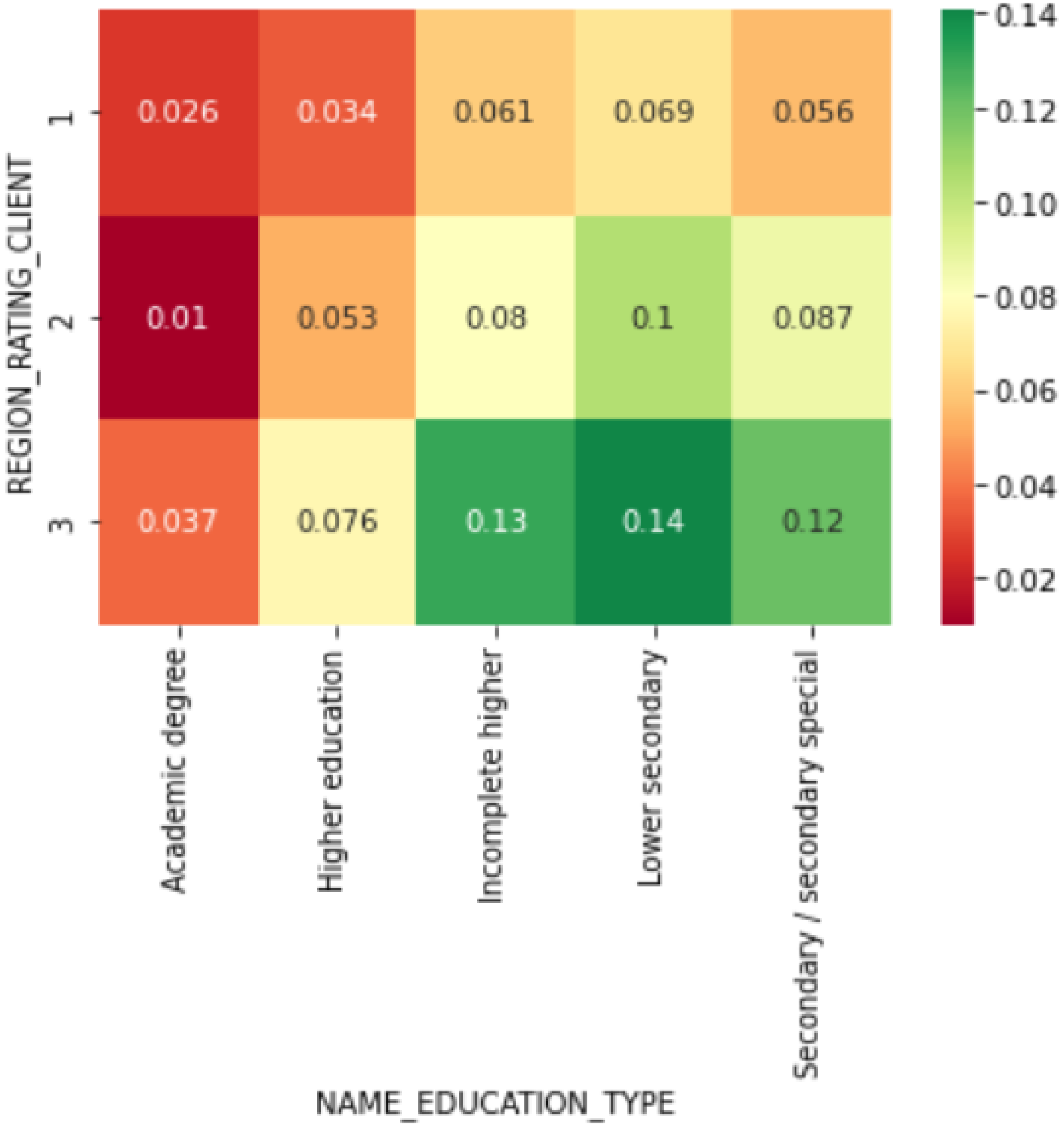
PROPERTY VS MARITAL STATUS

People living in co-op appartements and are seperated are likely to default the loan followed by seperated and civil marriage living in rented appartements



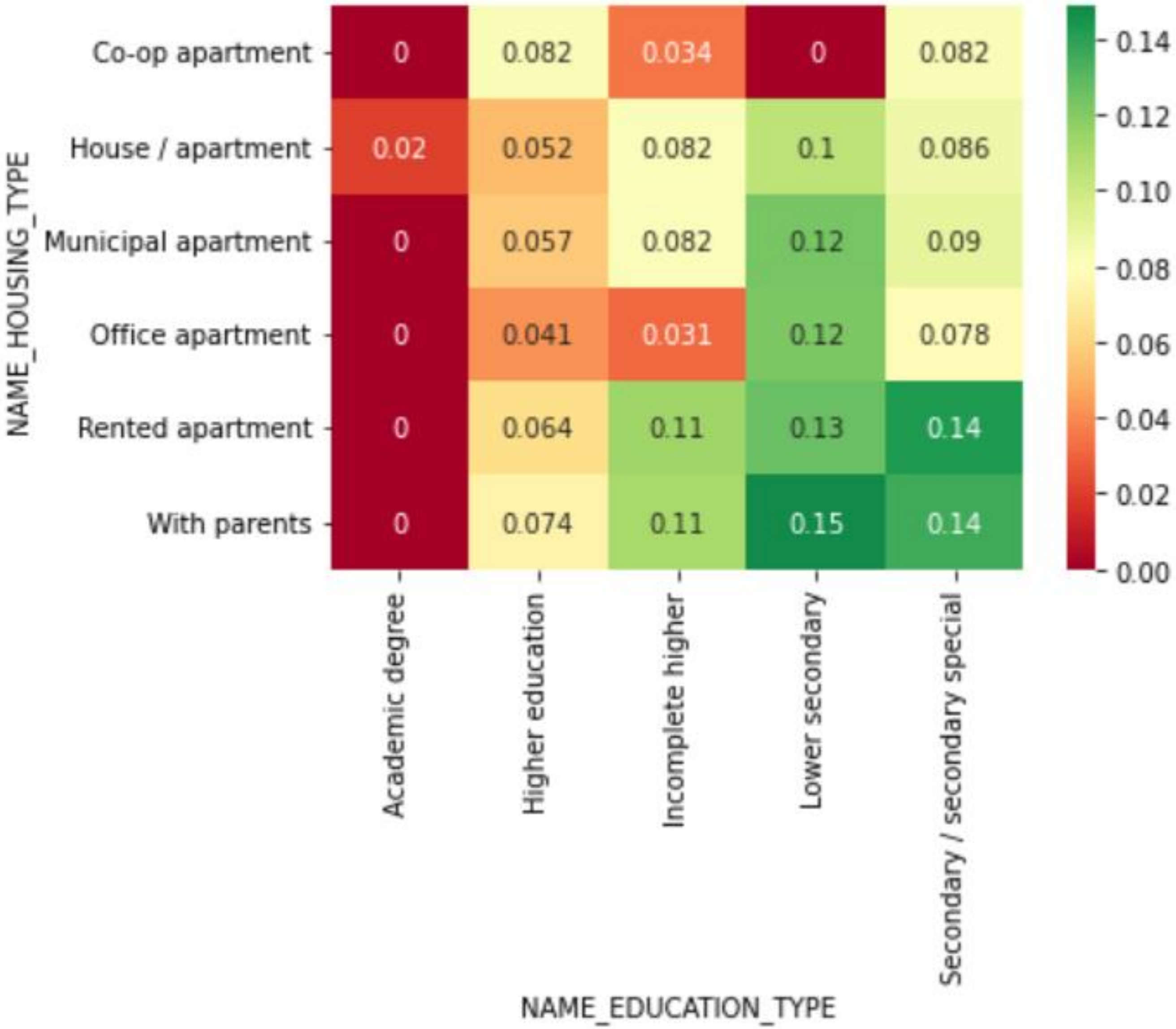
REGION VS EDUCATION

People living in region 3 and having lower Sc education are are likely to default follwed by people in the same region who havn't completed their higer education



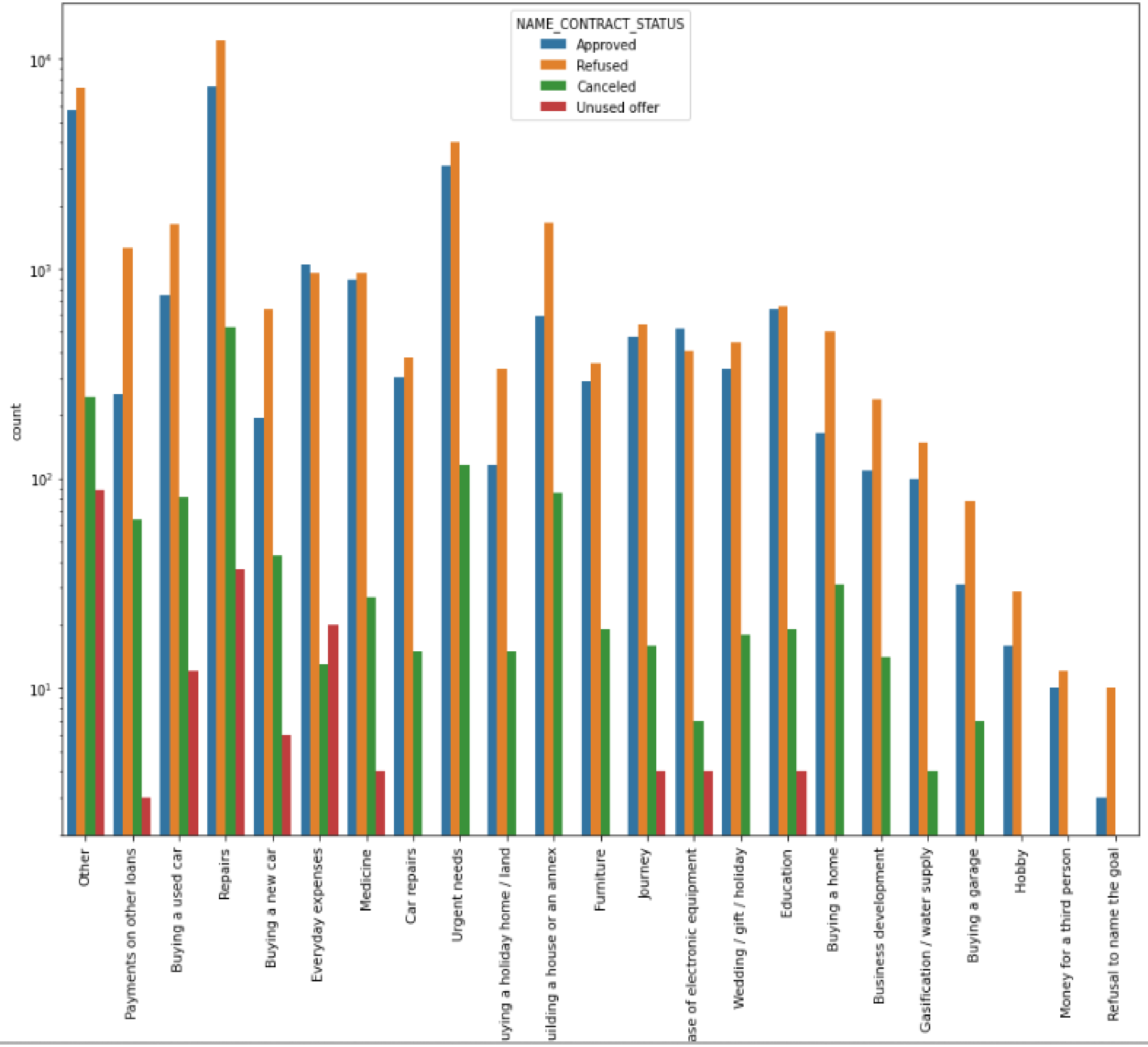
PROPERTY VS EDUCATION

People living with their parents and having lower Sc education are likely to default followed by people haven't completed their higher education



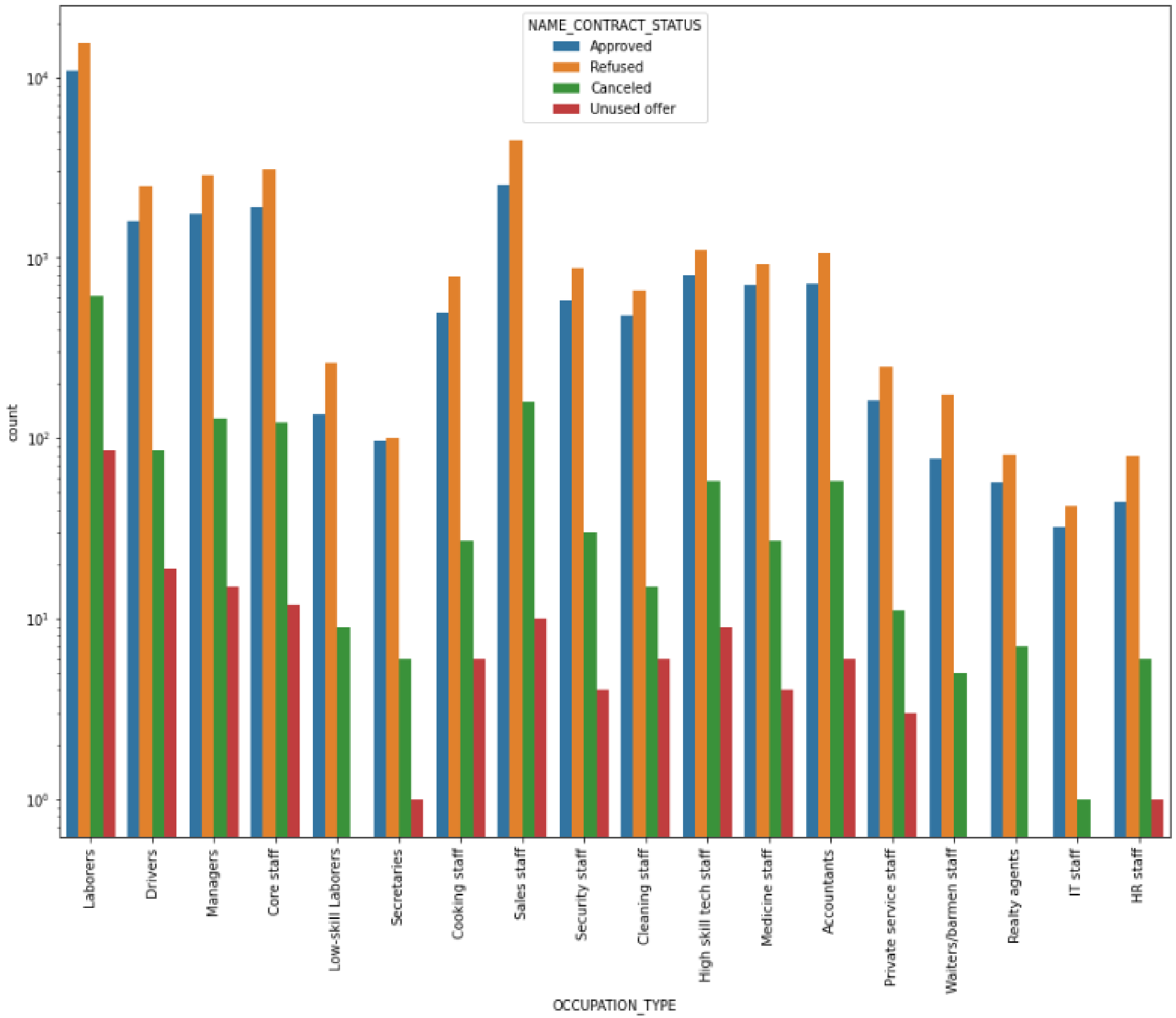
PURPOSE VS OFFER

For repair most of the loans are given and most of the offers are reject



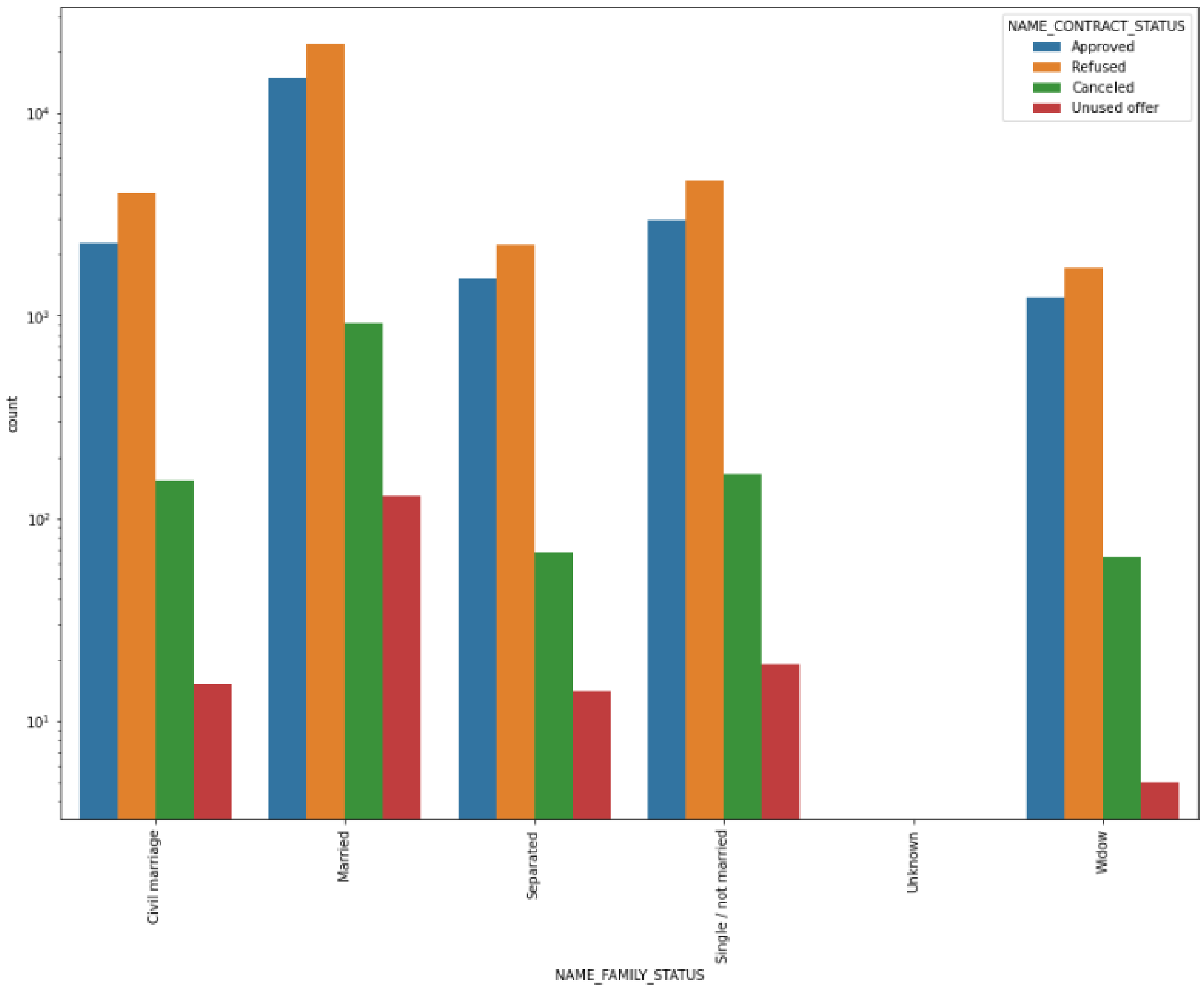
OCCUPATION VS OFFERS

Most of the loan are offered to labours



MARITAL STATUS VS OFFERS

Most of the offers are given to the married people



ACT

OBSERVATIONS

Following are main driving factors for one to default:

- Cash loan takers are more likely to default a payment.
- Men are more defaulters than women.
- Parents with 6 kids and more are more phorone to become a defaulter.
- People who live in rented apartments or are living with their parents likely to default a payment.
- People in the age group range 20-40 have higher probability of defaulting.
- A great number of people from region 3 are defaulting despite of their smaller number.

CONT..

- People given loan between 300k and 700k have high possibility to default.
- People who are unemployed or are on maternity leave have high chances to default the payments.
- People with Lower secondary education have more chances to be a defaulter.
- People with civil marriage or single or separated are more likely to default.
- Low-skill Laborers ,drivers Waiters/barmen staff, Security staff, Laborers and Cooking staff,sales staff are the highest categories to default.
- More the number of family members more one likely to default.

CONT..

- Single people who are accompanied by the group of people in loan application are likely to default more.
- Single low skilled labours and widowed HR staff are likely to default More
- Married unemployed are likely to default most.
- People living in co-op appartements and are separated are likely to default the loan followed by separated and civil marriage living in rented appartements
- People living in region 3 and are single or having civil marriage are likely to default .

CONT…

- People living with their parents and having lower Sc education are likely to default followed by people haven't completed their higher education.
- If the salary is less than 500k more chances of defaulting chances for default.
- Organizations with highest percent of loans not repaid are Transport: type 3 , Industry: type 13 , Industry: type 8 and Restaurant.
- People with civil marriage and lower Sc education likely to default more followed by lower Sc separated and single people .
- People living in region 3 and having lower Sc education are likely to default followed by people in the same region who haven't completed their higher education.

SUGGESTIONS

- A proper education background is important before giving loans
- People with more than 6 kids needs extra attention before giving out loans.
- Its better to avoid giving loans to unemployed married people
- People with employment less 3 years should be given loan at a higher interest rate in order to mitigate the risk
- Application of People on maternity leave and have secondary education need a through scrutiny check before giving loan

CONT..

- People with higher family count can be charged more interest
- Applications of people In region 3 needs to be dealt more carefully
- People living in co-op appartements and are separated are likely to default
- Widowed HR staff need more attention while giving loan
- A good look in the previous history of the customer will benefit bank more
- Whether a customer is contacted before or not need to be checked before giving loan

THANK YOU!