HRISHIKESH TONGE
230340325015

EDA

Analyzing Loan Application Data

BUSINESS UNDERSTANDING

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:

- **TYPE 1 ERROR:** IF THE APPLICANT IS LIKELY TO REPAY THE LOAN, THEN NOT APPROVING THE LOAN RESULTS IN A LOSS OF BUSINESS TO THE COMPANY

- **TYPE 2 ERROR:** 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.

BUSINESS UNDERSTANDING

WHEN A CLIENT APPLIES FOR A LOAN, THERE ARE FOUR TYPES OF DECISIONS THAT COULD BE TAKEN BY THE CLIENT/COMPANY

- Approved
- Refused
- Cancelled
- Unused offer

UNDERSTANDING DATA

- **'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.

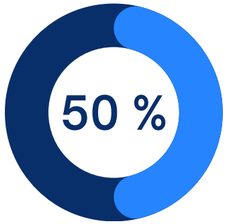
- **'PREVIOUS_APPLICATION.CSV'** CONTAINS INFORMATION ABOUT THE CLIENT'S PREVIOUS LOAN DATA. IT CONTAINS THE DATA WHETHER THE PREVIOUS APPLICATION HAD BEEN APPROVED, CANCELED, REFUSED OR UNUSED OFFER

- **'COLUMNS_DESCRIPTION.CSV'** IS A DATA DICTIONARY WHICH DESCRIBES THE MEANING OF THE VARIABLES.

DATA CLEANING

APPLICATION DATA

1. DEALING WITH NULL VALUES



It was observed and concluded that, columns with null values more than 50 % should be removed. Hence they were dropped at the beginning itself

On further closely looking in the other attributes there were few fields where the definition was not very clear and were not sure if the fields would add value to the analysis.

DAYS_BIRTH and DAYS_EMPLOYED were converted to years

Dropping columns with negative correlation factor

DATA CLEANING

APPLICATION DATA

THE FOLLOWING COLUMNS WERE SELECTED FOR ANALYSIS

We have selected the columns based on

- the null percentage
- the correlation factor
- logical significance of the column.

```
new_df = app_data_filtered[['SK_ID_CURR','TARGET','NAME_CONTRACT_TYPE','CODE_GENDER','FLAG_OWN_CAR','FLAG_OWN_REALTY',
                            'CNT_CHILDREN','AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY','AMT_GOODS_PRICE',
                            'NAME_TYPE_SUITE','NAME_INCOME_TYPE','NAME_EDUCATION_TYPE','NAME_FAMILY_STATUS',
                            'NAME_HOUSING_TYPE','REGION_POPULATION_RELATIVE','DAYS_BIRTH','DAYS_EMPLOYED',
                            'OCCUPATION_TYPE','CNT_FAM_MEMBERS','REGION_RATING_CLIENT_W_CITY',
                            'ORGANIZATION_TYPE','EXT_SOURCE_2','OBS_30_CNT_SOCIAL_CIRCLE','DEF_30_CNT_SOCIAL_CIRCLE',
                            'OBS_60_CNT_SOCIAL_CIRCLE','DEF_60_CNT_SOCIAL_CIRCLE','FLAG_DOCUMENT_2','FLAG_DOCUMENT_3',
                            'FLAG_DOCUMENT_20','FLAG_DOCUMENT_21','AMT_REQ_CREDIT_BUREAU_YEAR']]
```

7] ✓ 0.0s

DATA CLEANING

APPLICATION DATA

2.FILLING NULL VALUES

It was observed that the following two columns still had significant null percentage

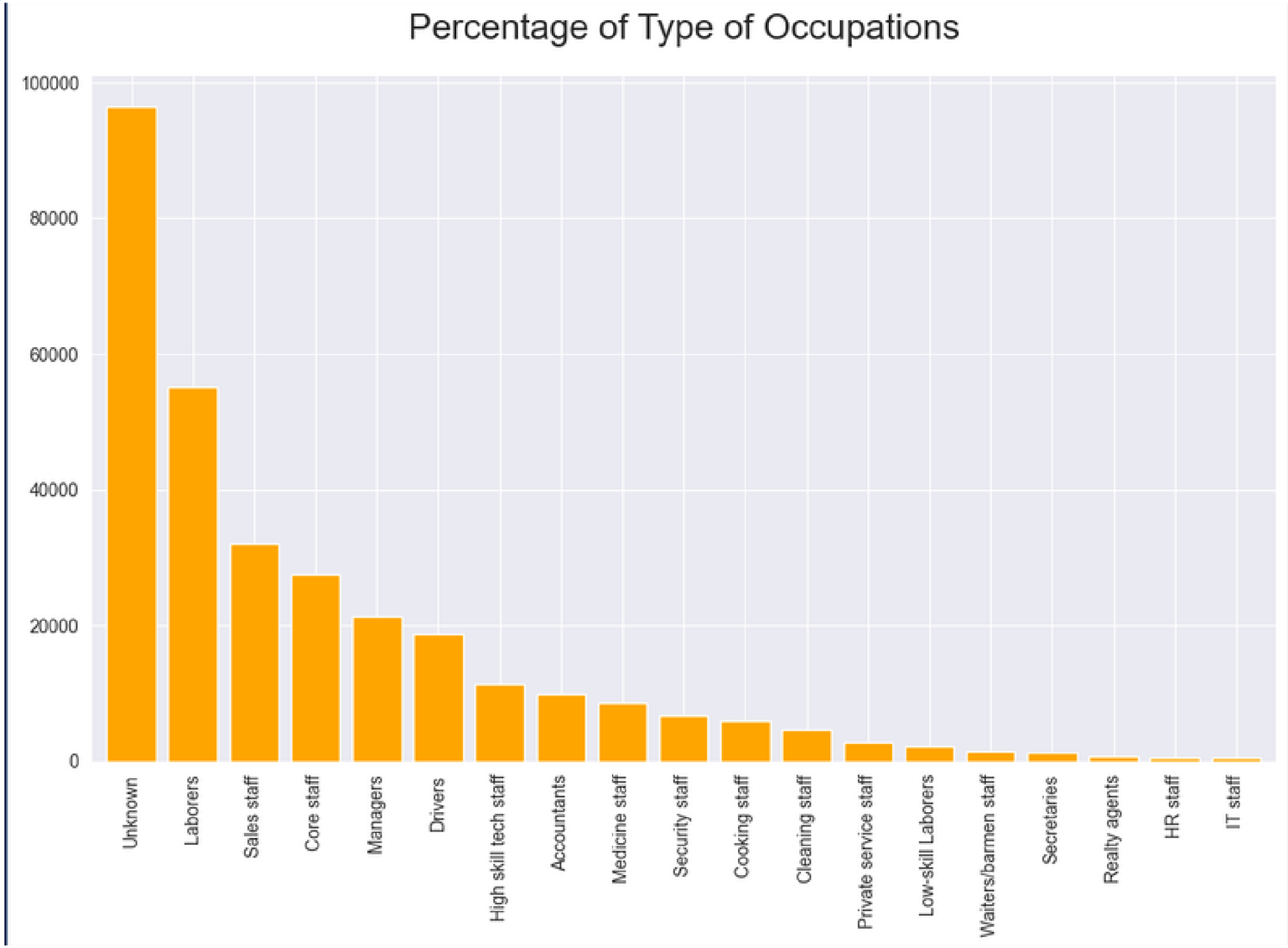
OCCUPATION_TYPE

- Occupation_type had the most significance and consisted of more than 55k records which had null values.
- Thus this was filled by 'Unknown'.
- As replacing it with any other value would create a bias.
- Deleting the records was not a viable option.

AMT_REQ_CREDIT_BUREAU_YEAR

- We checked for mean and median to fill in the data but found it would affect the analysis if mean/median was filled.
- We then filled the data using forward fill.

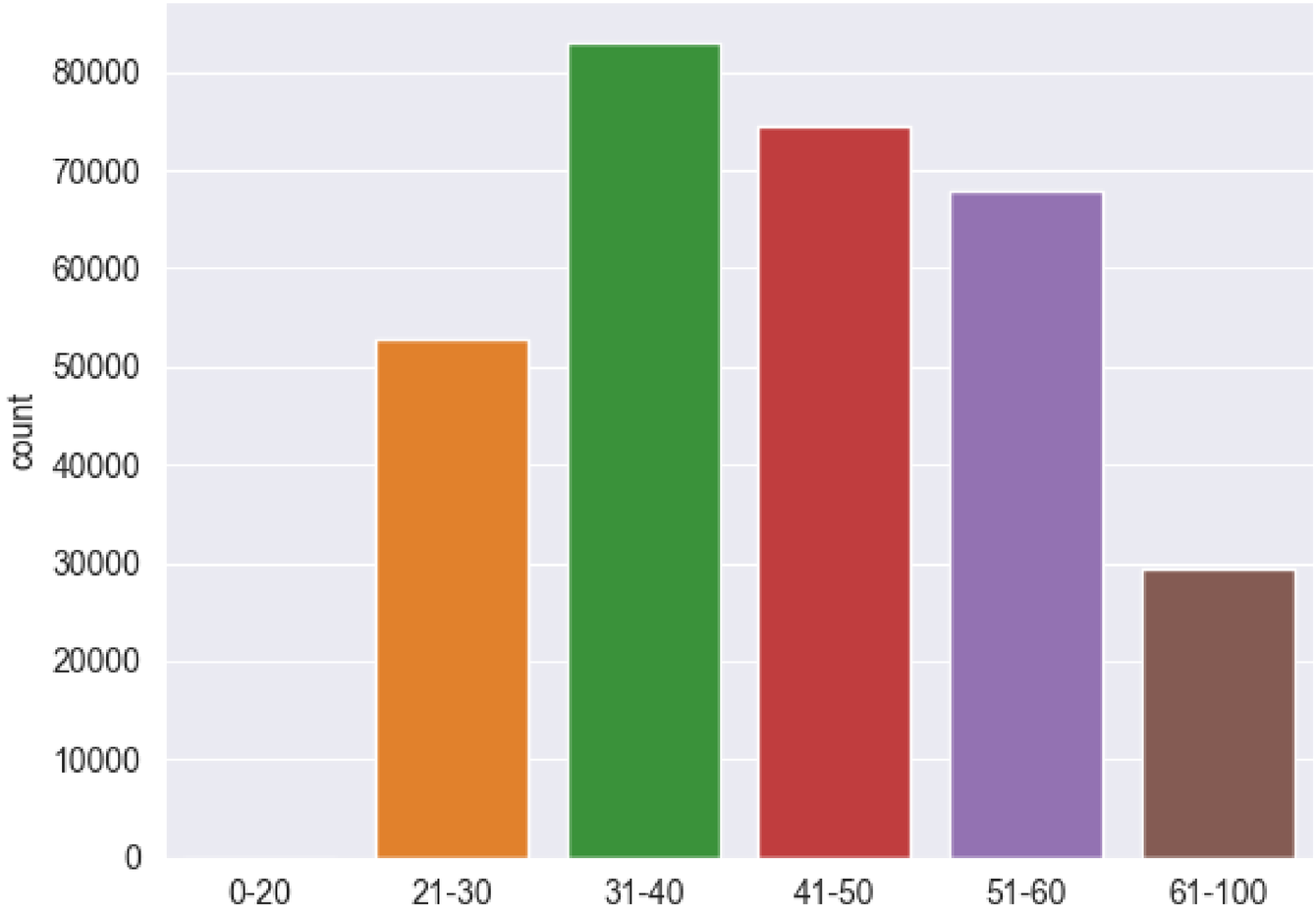
OCCUPATION TYPES



LABOURERS HAVE THE HIGHEST RATE

APPLICATION DATA

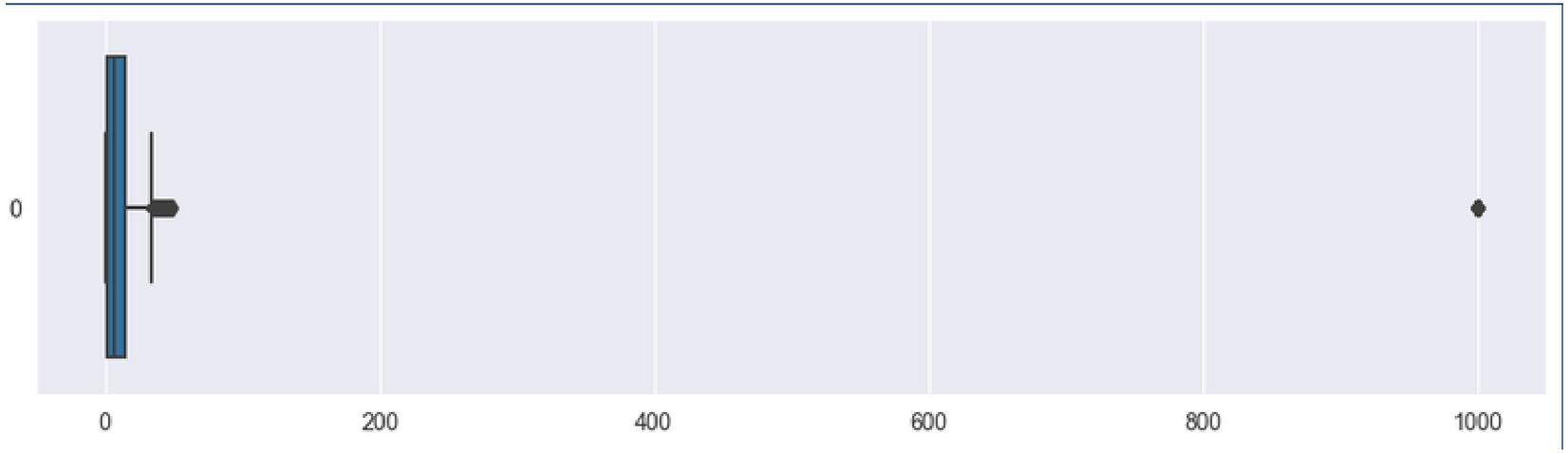
AGE GROUP



AS OBSERVED WE CAN CONCLUDE THAT THE AGE GROUPS 31-40 HAVE THE MAXIMUM APPLICATIONS

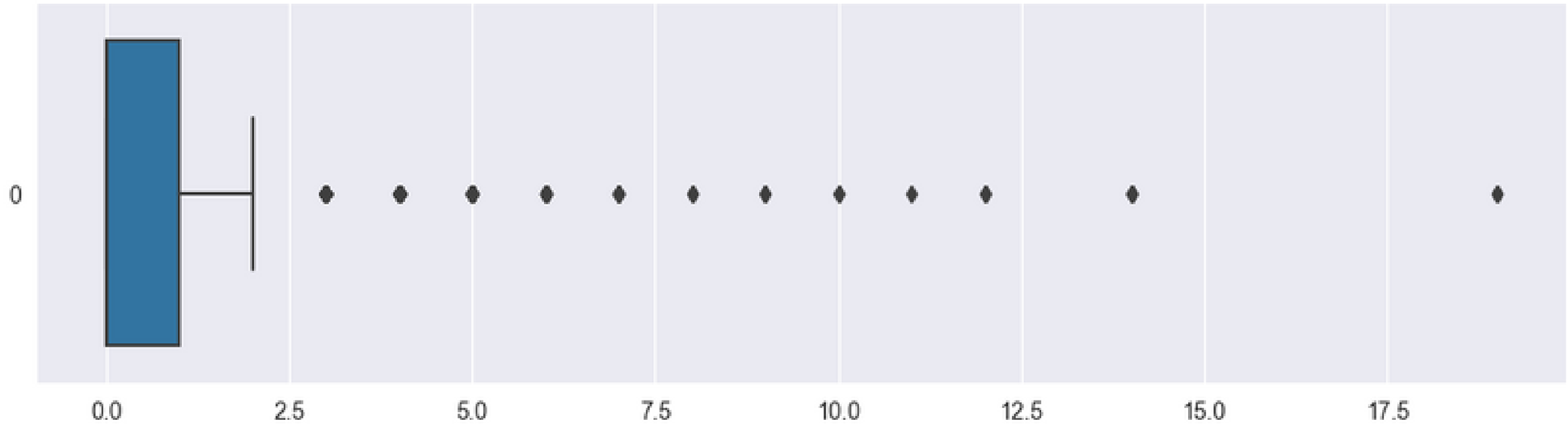
APPLICATION DATA

YEARS OF EXPERIENCE



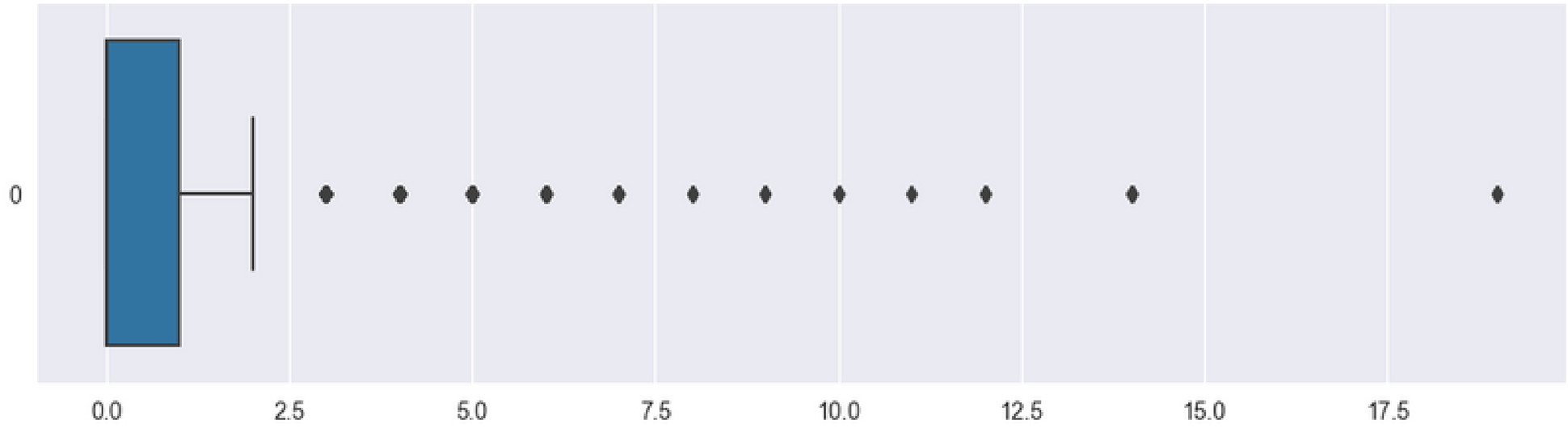
- IF WE SEE THE BOX PLOT OF THE EXPERIENCE IN YEARS, WE CAN SEE THERE ARE AROUND ~55K RECORDS WHERE THE EXPERIENCE IS IN 1000 YEARS, WHICH IS SIGNIFICANTLY.
- THIS DATA IS OF EITHER UNEMPLOYED INDIVIDUALS OR PENSIONERS
- WE HAVE 22 RECORDS OF UNEMPLOYED INDIVIDUALS. THUS THE REST WE CONCLUDE AS PENSIONERS. WE DO NOT DELETE/REPLACE ANYTHING.

CHILDREN COUNT



- WE HAVE OUTLIERS IN THE CNT_CHILDREN COLUMN 1ST QUARTILE IS MISSING FOR CNT_CHILDREN WHICH MEANS MOST OF THE DATA ARE PRESENT IN THE 1ST QUARTILE.

CHILDREN COUNT



- WE HAVE OUTLIERS IN THE CNT_CHILDREN COLUMN 1ST QUARTILE IS MISSING FOR CNT_CHILDREN WHICH MEANS MOST OF THE DATA ARE PRESENT IN THE 1ST QUARTILE.

IMBALANCE RATIO

THE RATIO OF THE SAMPLE SIZE OF THE LARGEST MAJORITY CLASS AND THAT OF THE SMALLEST MINORITY CLASS. THUS THE LARGER THE VALUE OF IR, THE LARGER THE IMBALANCE EXTENT.

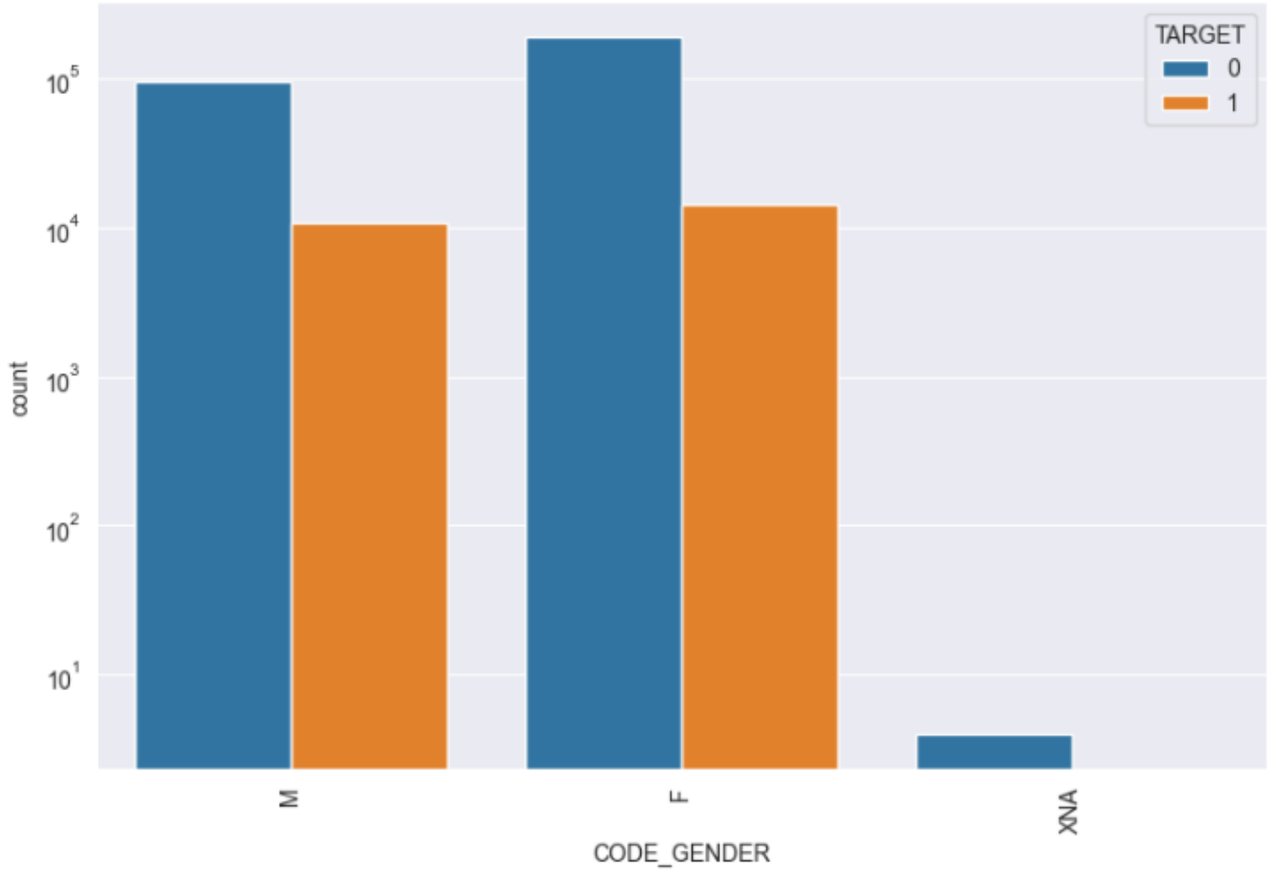
DIVIDING THE DATASET INTO TWO DATASET OF TARGET=1(CLIENT WITH PAYMENT DIFFICULTIES) AND TARGET=0(ALL OTHER)

IMBALANCE RATIO

11.39

THUS OUR IR IS LOW, IMBALANCE EXTENT IS LESS.

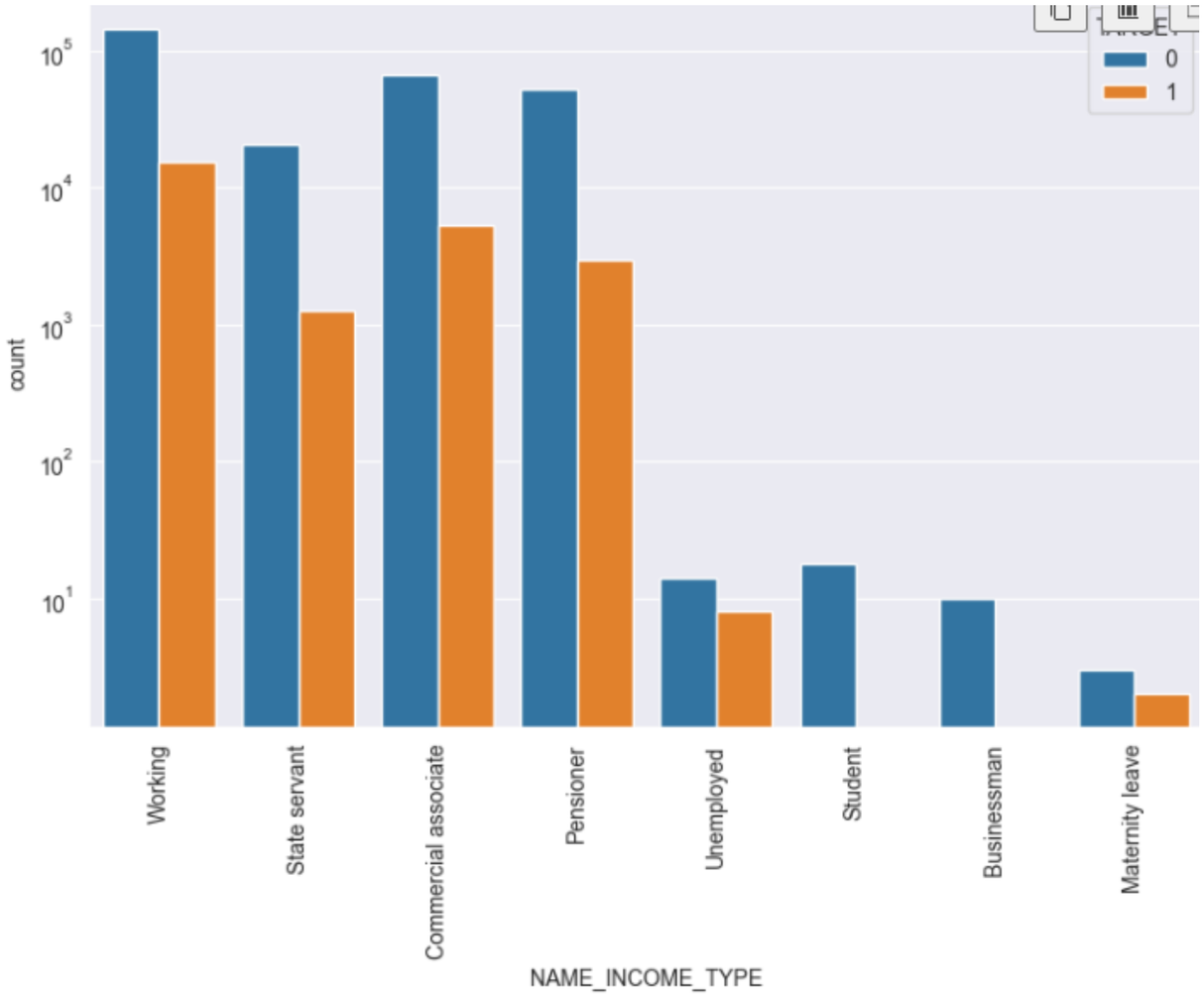
UNIVARIATE



GENDER:

THE % OF DEFAULTERS ARE MORE IN MALE THAN FEMALE

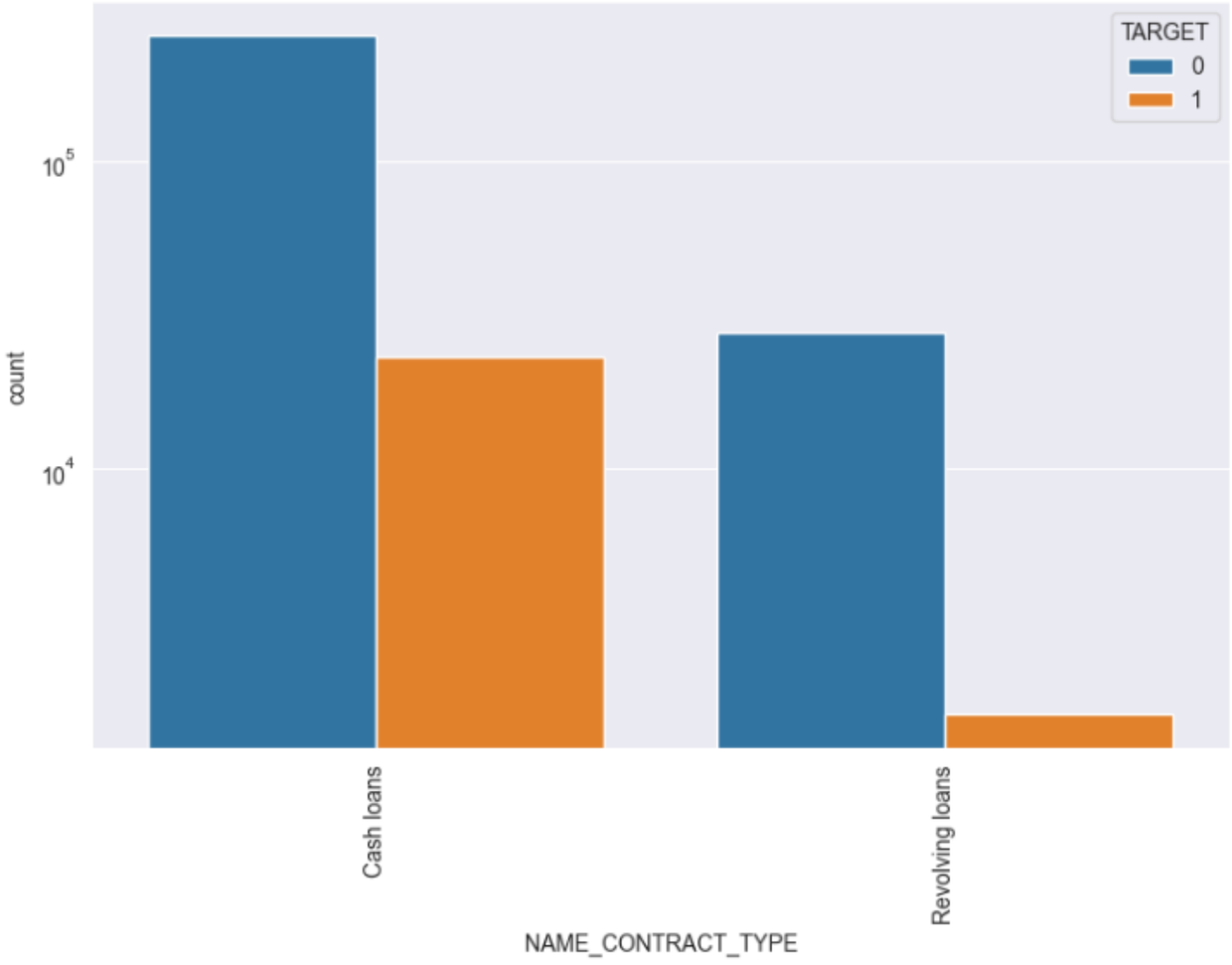
UNIVARIATE



INCOME TYPE:

- STUDENT AND BUSINESS ARE HIGHER IN PERCENTAGE OF LOAN REPAYMENT.
- WORKING, STATE SERVENT AND COMMERCIAL ASSOCIATES ARE HIGHER IN DEFAULT PERCENTAGE.
- MATERNITY CATEGORY IS SIGNIFICANTLY HIGHER PROBLEM IN REPAYMENT.

UNIVARIATE



NAME_CONTRACT_TYPE

- FOR CONTRACT TYPE ‘CASH LOANS’ ARE HIGH IN NUMBER OF CREDITS THAN ‘REVOLVING LOANS’ CONTRACT TYPE.
- BY ABOVE GRAPH 'REVOLVING LOANS' IS SMALL AMOUNT COMPARED TO ‘CASH LOANS’

TOP 10 INFUENCING FACTORS

AMT_CREDIT
AMT_GOODS_PRICE

AMT_ANNUITY
AMT_INCOME_TOTAL

CNT_CHILDREN
CNT_FAM_MEMBERS

AMT_GOODS_PRICE
AMT_INCOME_TOTAL

DEF_60_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE

AMT_INCOME_TOTAL
AMT_CREDIT

AMT_ANNUITY
AMT_GOODS_PRICE

OBS_60_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE

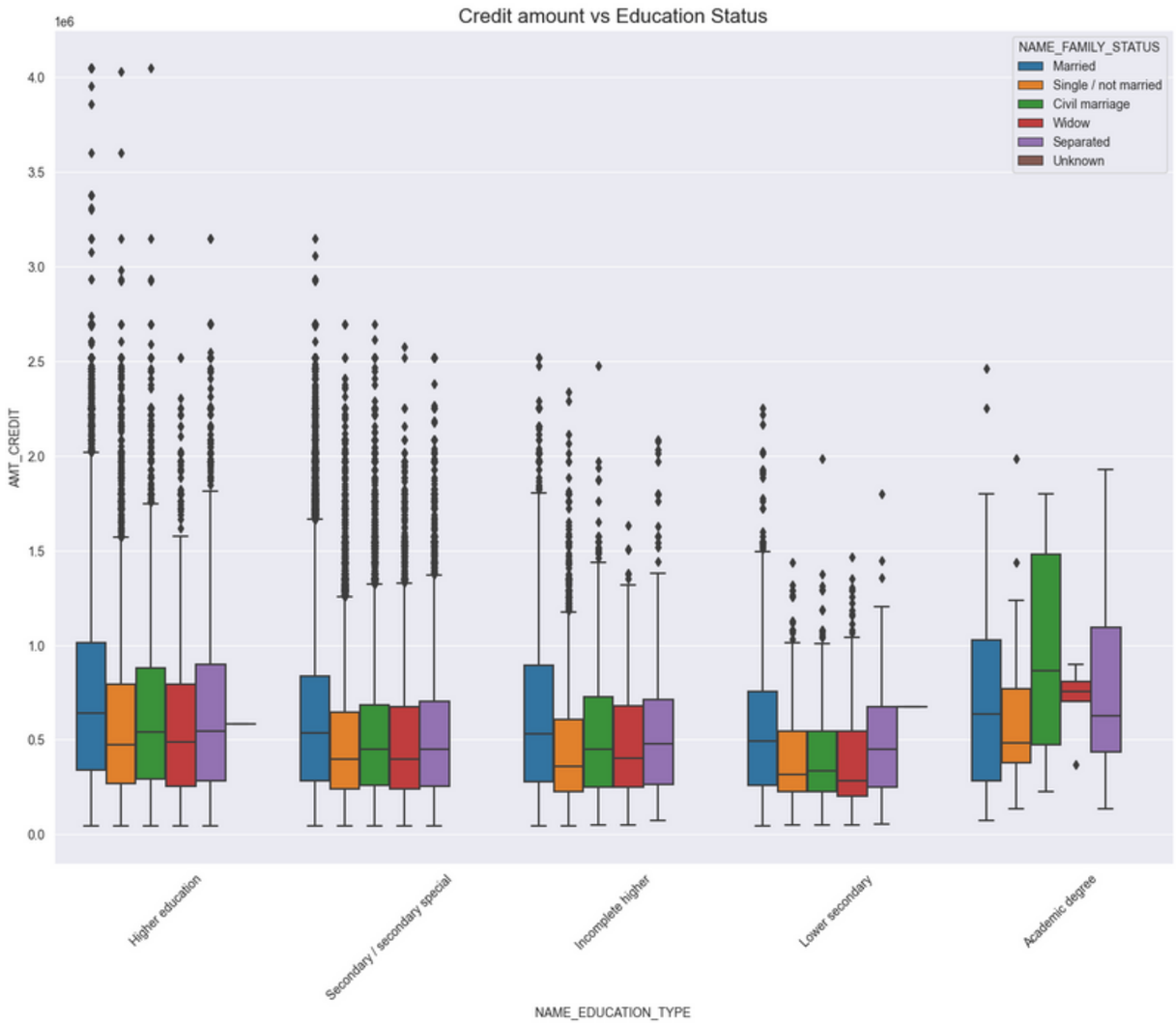
AMT_CREDIT
AMT_ANNUITY

DEF_30_CNT_SOCIAL_CIRCLE
OBS_30_CNT_SOCIAL_CIRCLE

APPLICATION DATA

BIVARIATE

EDUCATION TYPE VS CREDIT AMOUNT (PAYMENT / NON PAYMENT DIFFICULTIES) FOR TARGET0



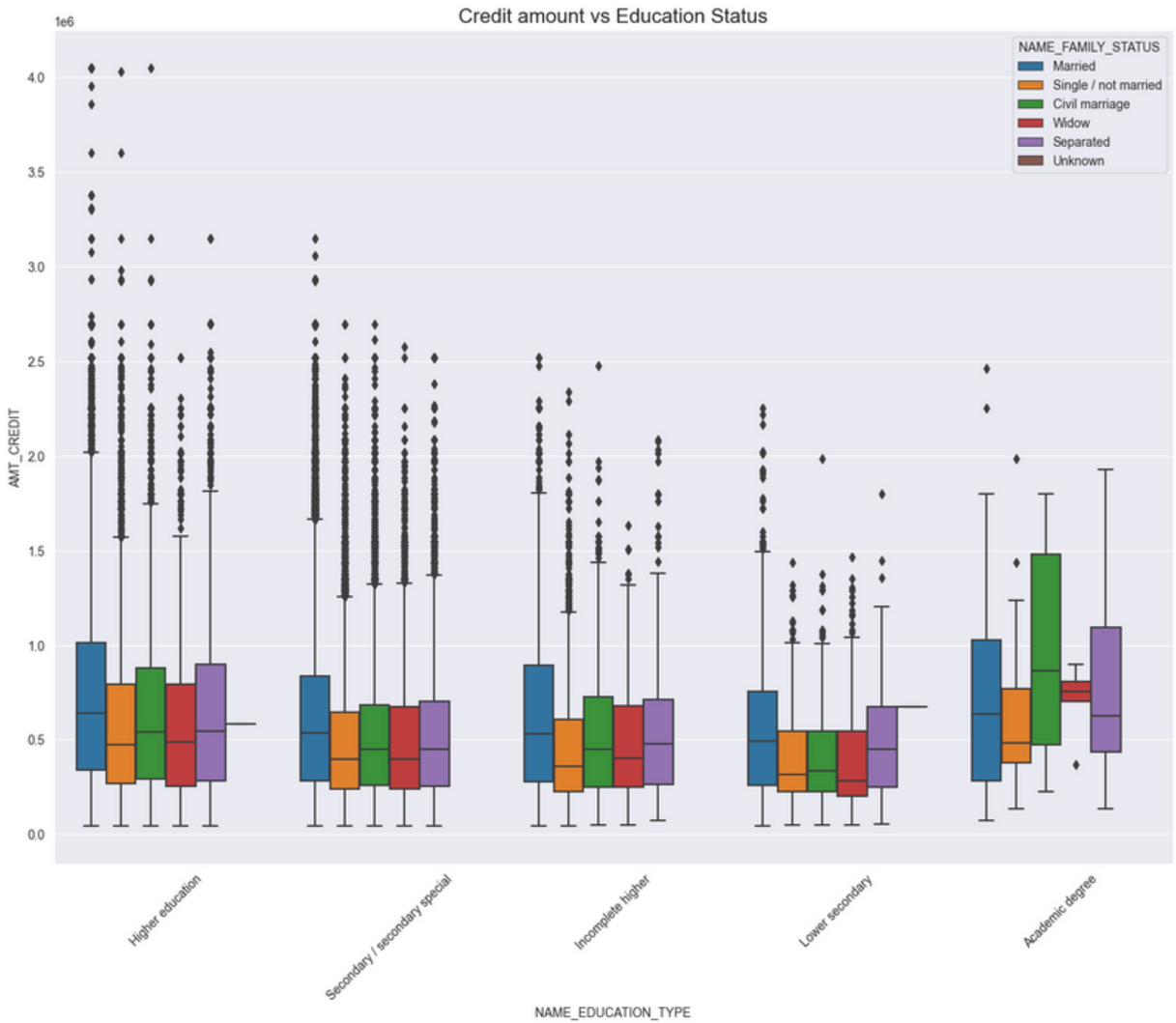
OBSERVATION - TARGET 0

FAMILY STATUS OF '**CIVIL MARRIAGE**', '**MARRIAGE**' AND '**SEPARATED**' OF **ACADEMIC DEGREE EDUCATION** ARE HAVING HIGHER NUMBER OF CREDITS THAN OTHERS.

ALSO, **HIGHER EDUCATION** OF FAMILY STATUS OF '**MARRIAGE**', '**SINGLE**' AND '**CIVIL MARRIAGE**' ARE HAVING MORE OUTLIERS. **CIVIL MARRIAGE** FOR **ACADEMIC DEGREE** IS HAVING MOST OF THE CREDITS IN THE THIRD QUARTILE.

BIVARIATE

EDUCATION TYPE VS CREDIT AMOUNT (PAYMENT / NON PAYMENT DIFFICULTIES) FOR TARGET1



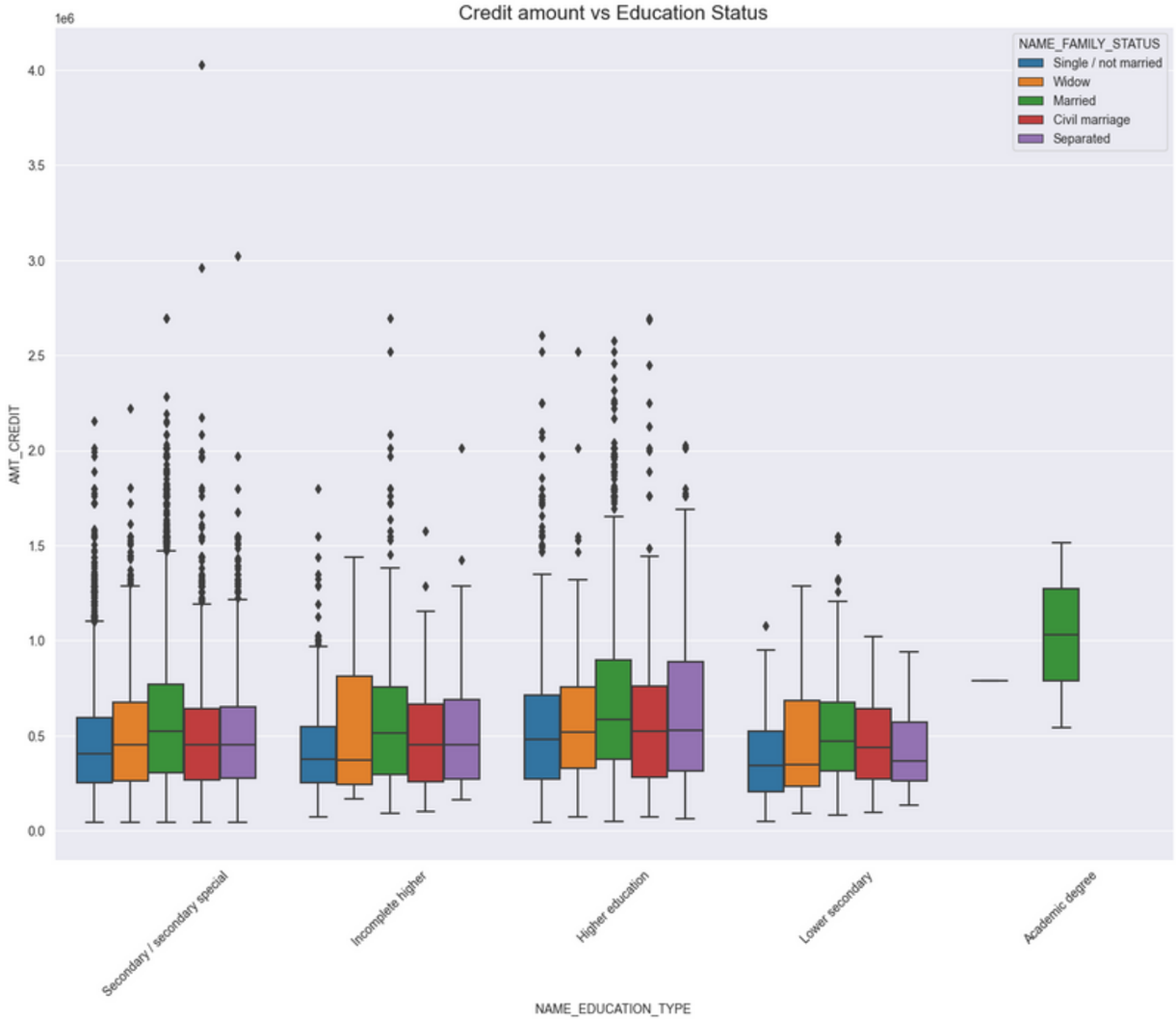
OBSERVATION - TARGET 1

QUITE SIMILAR WITH TARGET 0 FROM THE ABOVE BOX PLOT WE CAN SAY THAT FAMILY STATUS OF '**CIVIL MARRIAGE**', '**MARRIAGE**' AND '**SEPARATED**' OF **ACADEMIC DEGREE** EDUCATION ARE HAVING HIGHER NUMBER OF CREDITS THAN OTHERS.

MOST OF THE OUTLIERS ARE FROM EDUCATION TYPE '**HIGHER EDUCATION**' AND '**SECONDARY**'. **CIVIL MARRIAGE** FOR **ACADEMIC** DEGREE IS HAVING MOST OF THE CREDITS IN THE THIRD QUARTILE

BIVARIATE

INCOME VS CREDIT AMOUNT (PAYMENT / NON PAYMENT DIFFICULTIES) FOR TARGET0



OBSERVATION - TARGET 1

QUITE SIMILAR WITH TARGET 0 FROM THE ABOVE BOX PLOT WE CAN SAY THAT FAMILY STATUS OF '**CIVIL MARRIAGE**', '**MARRIAGE**' AND '**SEPARATED**' OF **ACADEMIC DEGREE** EDUCATION ARE HAVING HIGHER NUMBER OF CREDITS THAN OTHERS.

MOST OF THE OUTLIERS ARE FROM EDUCATION TYPE '**HIGHER EDUCATION**' AND '**SECONDARY**'. **CIVIL MARRIAGE** FOR **ACADEMIC** DEGREE IS HAVING MOST OF THE CREDITS IN THE THIRD QUARTILE

DATA CLEANING

PREVIOUS APPLICATION DATA

1. DEALING WITH NULL VALUES

50 %

It was observed and concluded that, columns with null values more than 50 % should be removed. Hence they were dropped at the beginning itself

Selecting significant columns which will add value to the analysis

```
prev_df = prev_data[['SK_ID_PREV',
'SK_ID_CURR',
'NAME_CONTRACT_TYPE',
'AMT_ANNUITY',
'AMT_APPLICATION',
'AMT_CREDIT',
'AMT_GOODS_PRICE',
'NAME_CASH_LOAN_PURPOSE',
'NAME_CONTRACT_STATUS',
'CODE_REJECT_REASON',
'NAME_CLIENT_TYPE']]
✓ 0.1s
```

MERGING DATA

PREVIOUS APPLICATION DATA
+
APPLICATION DATA

MERGED_DF = PD.MERGE(LEFT = NEW_DF,RIGHT=PREV_DF, ON='SK_ID_CURR',HOW='INNER')

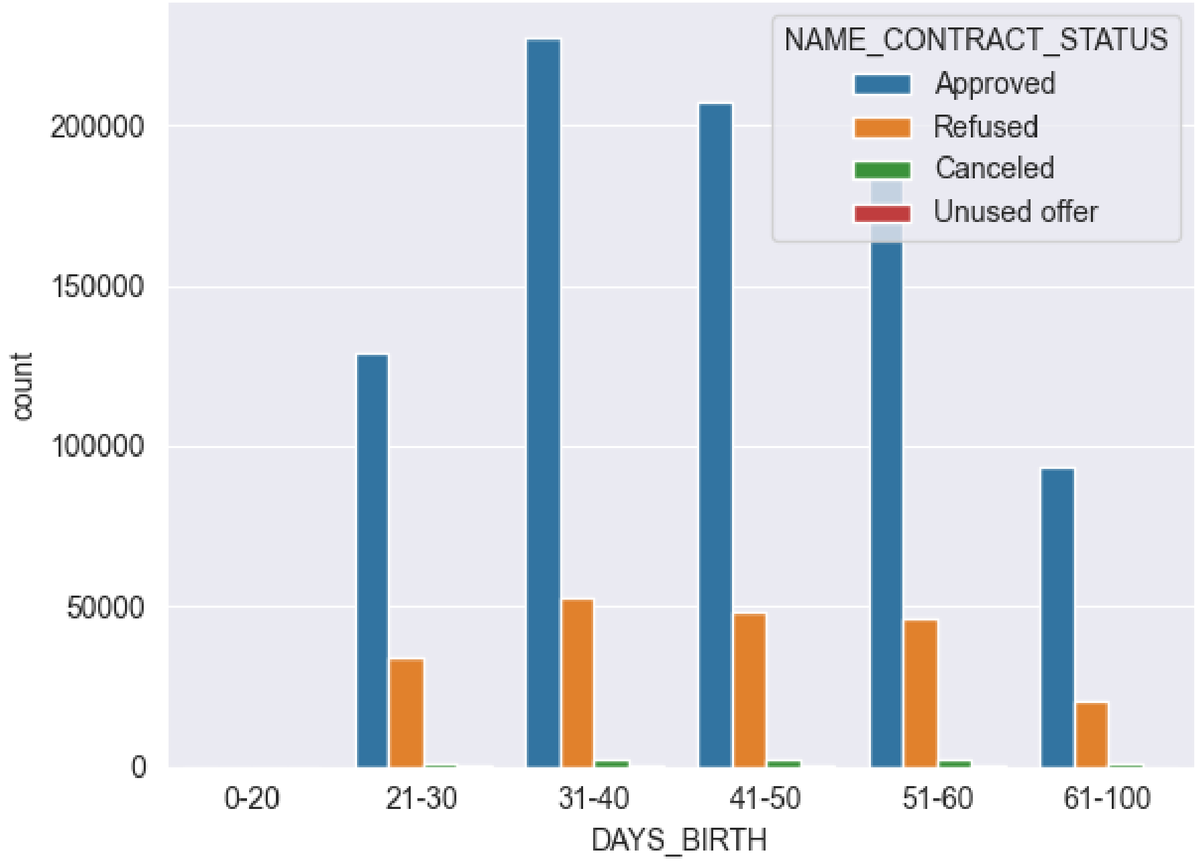
CREATING **MERGED_DF** WHICH IS MERGED ON THE COLUMN SK_ID_CURR.

```
merged_df.columns
✓ 0.0s

Index(['SK_ID_CURR', 'TARGET', 'NAME_CONTRACT_TYPE_x', 'CODE_GENDER',
       'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL',
       'AMT_CREDIT_x', 'AMT_ANNUITY_x', 'AMT_GOODS_PRICE_x', 'NAME_TYPE_SUITE',
       'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS',
       'NAME_HOUSING_TYPE', 'REGION_POPULATION_RELATIVE', 'DAYS_BIRTH',
       'DAYS_EMPLOYED', 'OCCUPATION_TYPE', 'CNT_FAM_MEMBERS',
       'REGION_RATING_CLIENT_W_CITY', 'ORGANIZATION_TYPE', 'EXT_SOURCE_2',
       'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE',
       'OBS_60_CNT_SOCIAL_CIRCLE', 'DEF_60_CNT_SOCIAL_CIRCLE',
       'FLAG_DOCUMENT_2', 'FLAG_DOCUMENT_3', 'FLAG_DOCUMENT_20',
       'FLAG_DOCUMENT_21', 'AMT_REQ_CREDIT_BUREAU_YEAR', 'SK_ID_PREV',
       'NAME_CONTRACT_TYPE_y', 'AMT_ANNUITY_y', 'AMT_APPLICATION',
       'AMT_CREDIT_y', 'AMT_GOODS_PRICE_y', 'NAME_CASH_LOAN_PURPOSE',
       'NAME_CONTRACT_STATUS', 'CODE_REJECT_REASON', 'NAME_CLIENT_TYPE'],
      dtype='object')
```

ANALYSING AGE AND CONTRACT APPROVAL

PREVIOUS APPLICATION DATA
+
APPLICATION DATA

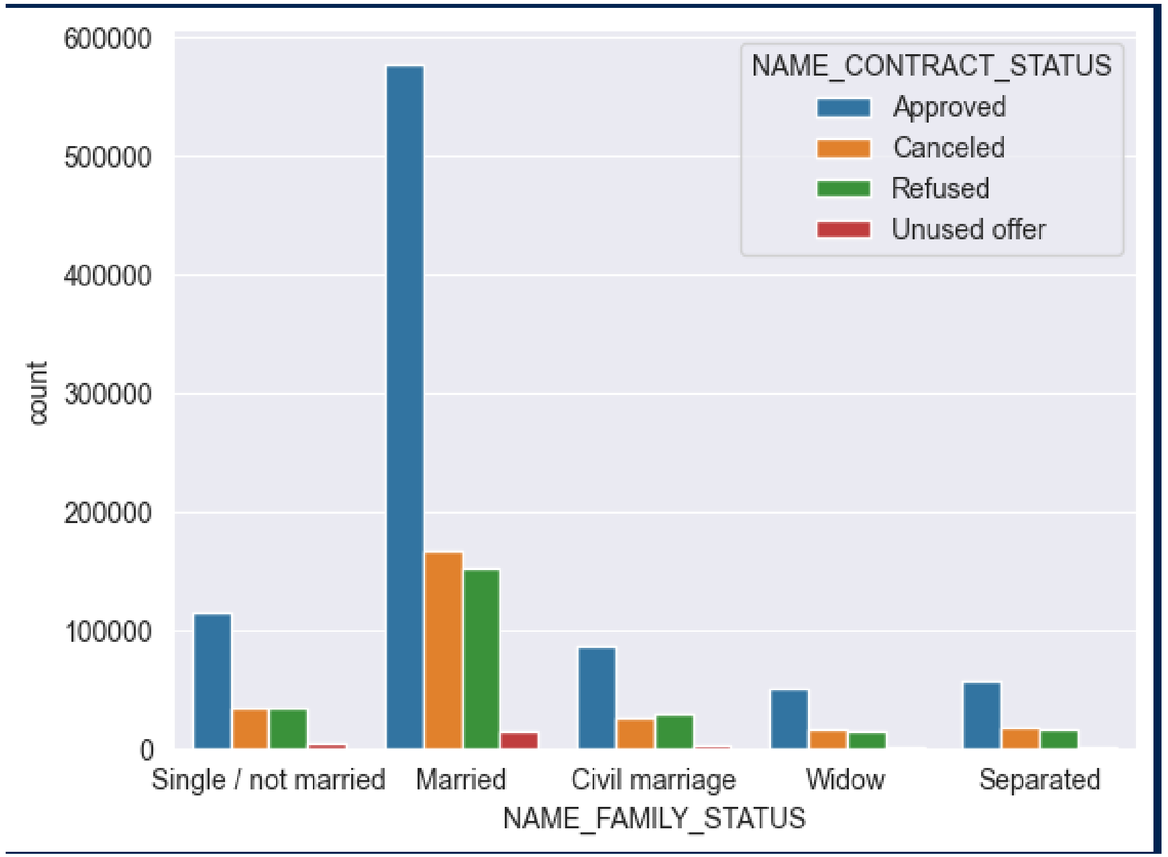


OBSERVATION

- LOAN MOSTLY GET APPROVED FOR THE AGE BAND OF 31 TO 40.
- THE SECOND HIGHEST AGE BAND WHERE LOAN GETS APPROVED IS 41-5

ANALYSING FAMILY STATUS AND CONTRACT APPROVAL

PREVIOUS APPLICATION DATA
+
APPLICATION DATA

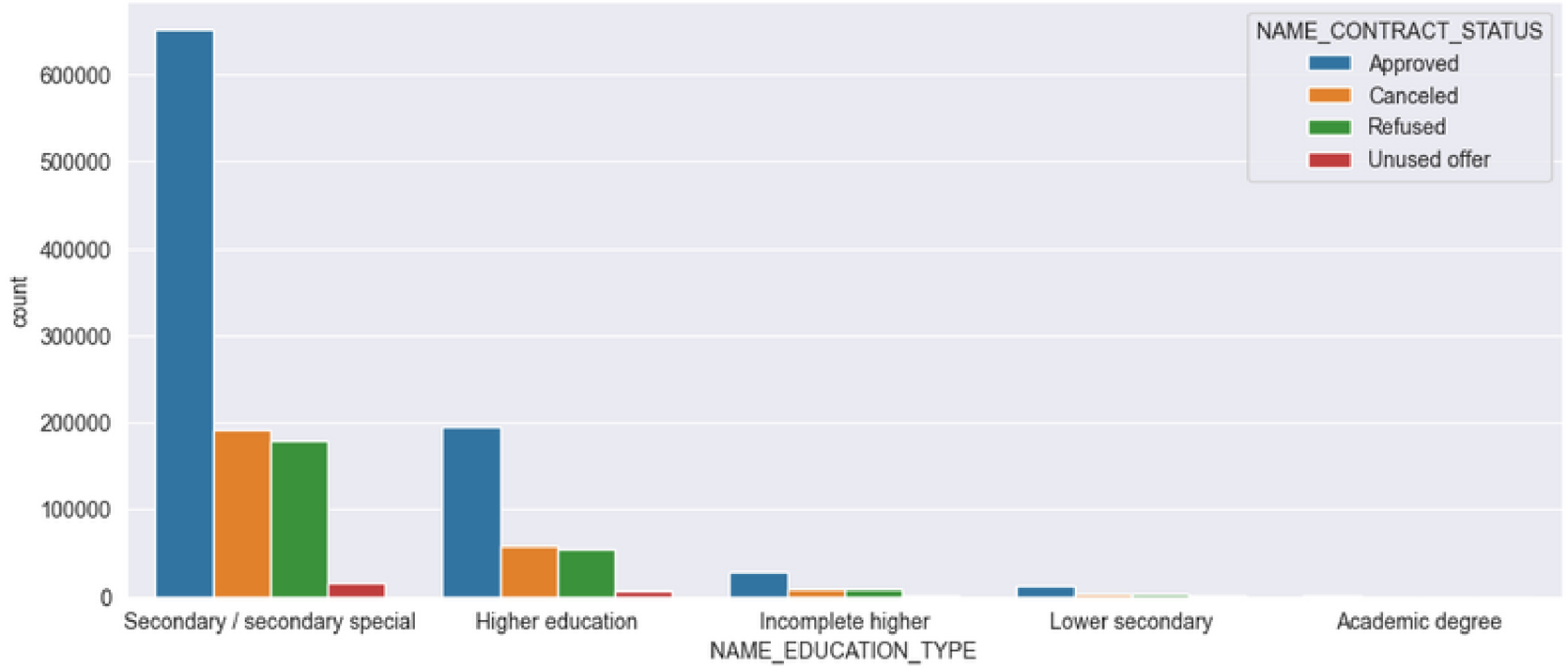


OBSERVATION

- THE LOAN GETS APPROVED MOSTLY FOR MARRIED PEOPLE.

ANALYSING EDUCATION TYPE AND CONSTRACT APPROVAL

PREVIOUS APPLICATION DATA
+
APPLICATION DATA

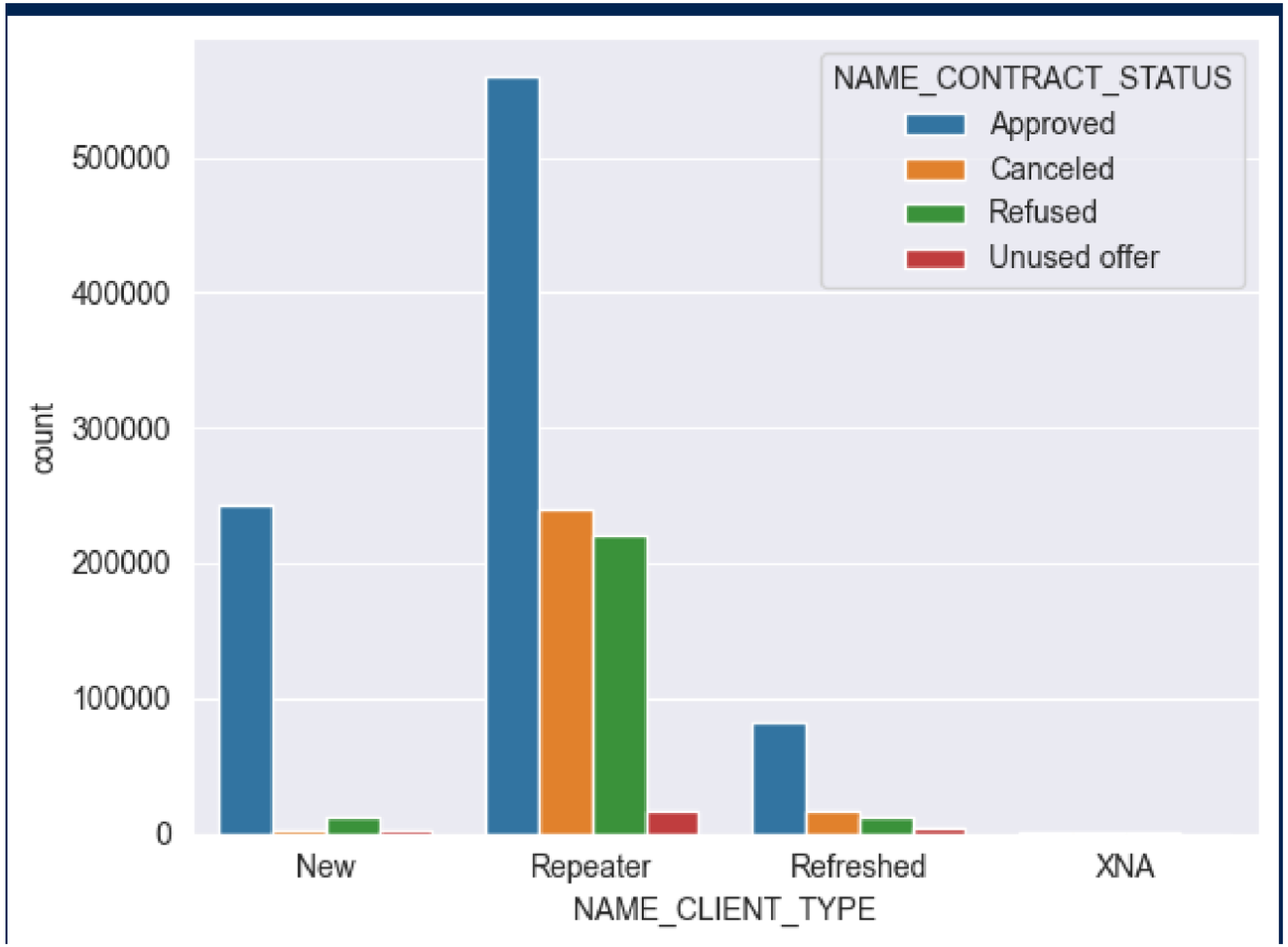


OBSERVATION

- THE PEOPLE WHO ARE EDUCATED TILL SECONDARY / SECONDARY-SPECIAL HAVE THE HIGHEST COUNT OF GETTING THE LOAN APPROVED

REPEATING CASES

PREVIOUS APPLICATION DATA + APPLICATION DATA



OBSERVATION

- MAXIMUM APPLICATIONS WHICH ARE APPROVED ARE OF THE APPLICATIONS WHICH ARE REFRESHED
- WHICH MEANS THE PEOPLE WHO ARE APPLYING FOR THE LOAN MAYBE SECOND/THIRD TIME, THEIR APPLICATIONS GET APPROVED FASTER.

LOGISTIC REGRESSION

PREVIOUS APPLICATION DATA
+
APPLICATION DATA

CONFUSION MATRIX

```
[206000    17  4658     0]
 [  1807     3   419     0]
 [ 44460    33  6073     0]
 [   173     0     0     0]
```

MODEL ACCURACY

80%

LOGISTIC REGRESSION

PREVIOUS APPLICATION DATA
+
APPLICATION DATA

CLASSIFICATION REPORT

- WE HAVE ACHIEVD 80% ACCURACY BY USING LOGISTIC REGRESSION MODEL.
- MODEL HAS GIVEN 82% PRECISION FOR APPROVED LOANS AND HIGHEST AFTER THAT 55% FOR REFUSED LOAN OFFER.
- RECALL VALUE IS HIGHEST FOR APPROVED LOAN OFFER WITH 98%

	precision	recall	f1-score	support
Approved	0.82	0.98	0.89	210675
Canceled	0.06	0.00	0.00	2229
Refused	0.54	0.12	0.20	50566
Unused offer	0.00	0.00	0.00	173
accuracy			0.80	263643
macro avg	0.35	0.27	0.27	263643
weighted avg	0.76	0.80	0.75	263643

THUS WE CAN USE THE LOGISTIC REGRESSION TO EFFECTIVELY PREDICT THE LIKELIHOOD OF PAYMENT DIFFICULTIES FOR LOAN APPLICANTS

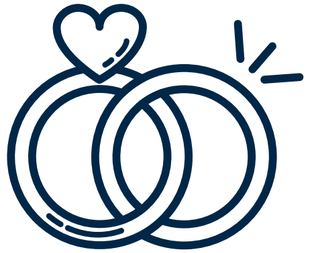
FINAL CONCLUSIONS

TARGET APPLICANTS

STUDENTS

BUSINESS MEN

PENSIONERS



MARRIED INDIVIDUALS

EDUCATION TYPE

SECONDARY/ SECONDARY SPECIAL

TARGET APPLICANTS AGE

31 TO 40 YRS

k You

HRISHIKESH TONGE