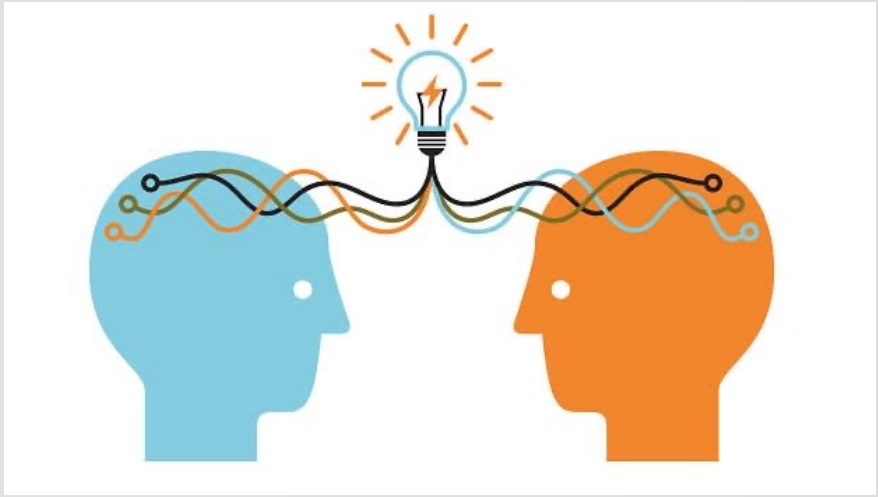
CREDIT EDA ANALYSIS

Presented by-Vikas Bhartiya
Batch no- DS 43
Email-vikas6050@gmail.com

Problem statement

- 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 minimise the risk of losing money while lending to customers.

Business Understanding



- Business Understanding is very very important before approaching the problem.
- Bank is running a good business and they have hired you as Data scientist so that you have to give the meaningful insights.
- Since there are many points for Business understanding so instead of all depicting here we will keep on elaborating time to time when we will solve the problem.

STEPS TO APPROACH

- Data understanding
- Cleaning the Data
- Missing value Treatment
- Changing column in Suitable Format.
- Outliers Treatment
- Preparing data for Analysis
- Univariate Analysis
- Bivariate Analysis
- Multivariate analysis
- Combine both the data and analysis.

AIM of problem

0,1

- In the data we have many features about 121.
- We have column Target and we can call it as dependent variable
- All we have to analysis against the target variable.
- Target-0 means non Defaulter and Target-1 means defaulter.
- Our conclusion should be profitable in terms or business
- Means wrong predicting 0 means bank loosing interest.
- Wrong predicting 1 more dangerous means loosing interest as well as principle amount.

DATA UNDERSTANDING

307511-ROWS,122-COULMN, 9152465-MISSING VALUES

Printing the intial information of the data

```
print ("No. of Rows     : " , bank.shape[0])
print ("No. of Column  : " , bank.shape[1])
print('*'*50)
print ("\nColumn's name: \n" ,bank.columns.tolist())
print('*'*50)
print ("\n Total Missing values :  ", bank.isnull().sum().values.sum())
print('*'*50)
print ("\nUnique values :  \n", bank.nunique())
```

```
No. of Rows     :  307511
No. of Column  :  122
```

41-COLUMN >50% MISSING VALUE SINCE WE HAVE MANY FEATURES BETTER DROP IT INSTEAD OF PREDICTING

total missing value column: 41

	index	missing values	missing percent
48	COMMONAREA_AVG	214865	69.872297
76	COMMONAREA_MEDI	214865	69.872297
62	COMMONAREA_MODE	214865	69.872297
84	NONLIVINGAPARTMENTS_MEDI	213514	69.432963
56	NONLIVINGAPARTMENTS_AVG	213514	69.432963
70	NONLIVINGAPARTMENTS_MODE	213514	69.432963
86	FONDKAPREMONT_MODE	210295	68.386172
82	LIVINGAPARTMENTS_MEDI	210199	68.354953
54	LIVINGAPARTMENTS_AVG	210199	68.354953
68	LIVINGAPARTMENTS_MODE	210199	68.354953
66	FLOORSMIN_MODE	208642	67.848630
80	FLOORSMIN_MEDI	208642	67.848630
52	FLOORSMIN_AVG	208642	67.848630
61	YEARS_BUILD_MODE	204488	66.497784
47	YEARS_BUILD_AVG	204488	66.497784
75	YEARS_BUILD_MEDI	204488	66.497784
21	OWN_CAR_AGE	202929	65.990810
53	LANDAREA_AVG	182590	59.376738
81	LANDAREA_MEDI	182590	59.376738

FILL UP THE REMAINING MISSING VALUE WITH MOST SUITABLE METHOD

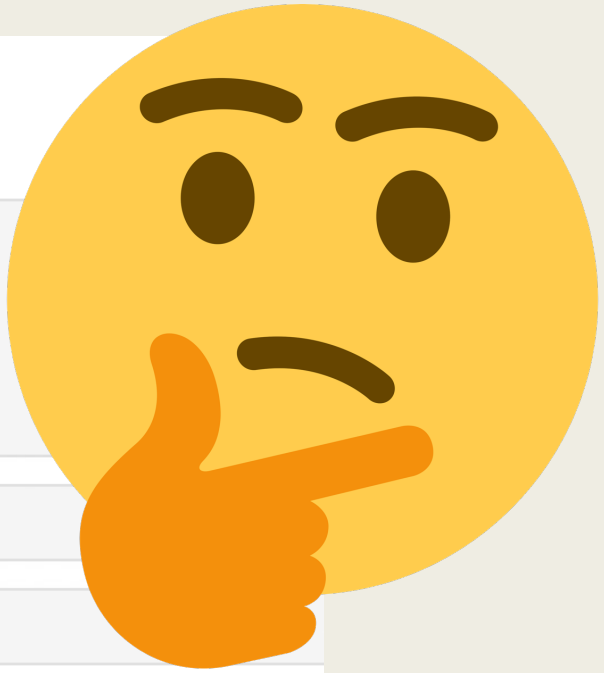
4-Filling missing value with mode and median

```
def fill_na(data, col1=num_col, col2=cat_col):
        for i in col1:
            data[i].fillna(data[i].median(),inplace=True)
        for j in col2:
            data[j].fillna(data[j].mode()[0],inplace=True)
```

```
fill_na(bank_notnull)
```

```
bank_notnull['EMERGENCYSTATE_MODE'].unique()
```

```
array(['No', 'Yes'], dtype=object)
```



DATA CLEANING

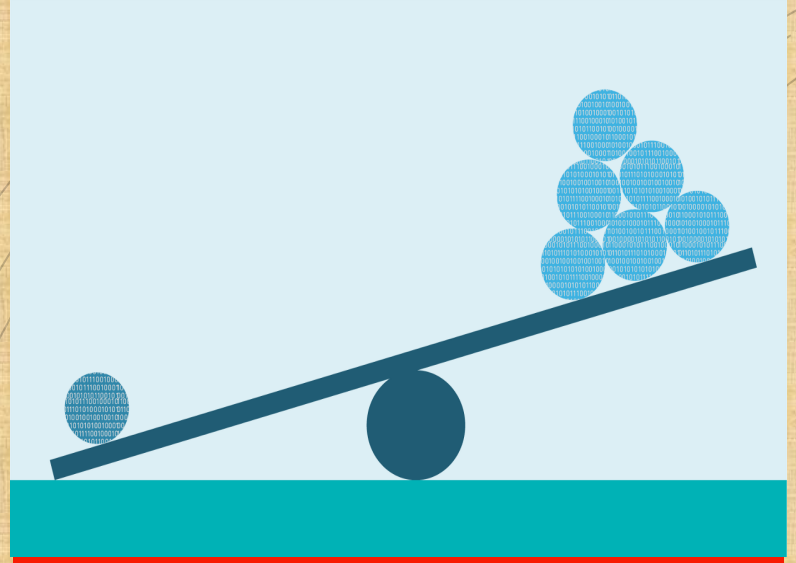
- Data cleaning is most crucial step when preparing for analysis.
- Visit every column in depth.
- XNA, XNP present in data
- Negative value in the YEAR and BIRTH
- Incorrect Data Type
- And many more
- More clean the data More beautiful insight.

SANITY CHECK-CORRECT DATA TYPE

Observation- from above we come to know that days columns has the negative value we will extract this column and convert in positive value

```
#list of days column having negative value
incorret_cols=['DAYS_BIRTH','DAYS_EMPLOYED','DAYS_REGISTRATION','DAYS_ID_PUBLISH','DAYS_LAST_PHONE_CHANGE']
```

```
#creating new column days to year
def clean_col(data,cols):
    for i in cols:
        data[i.replace('DAYS','YEAR')]=data[i].apply(lambda x: round((abs(x)/365),1))
```



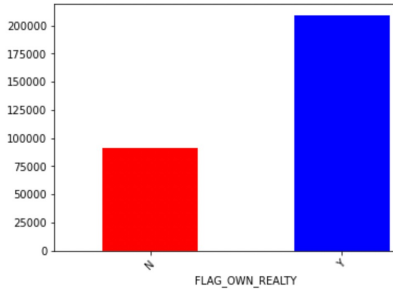
HIGHLY IMBALANCE DATASET

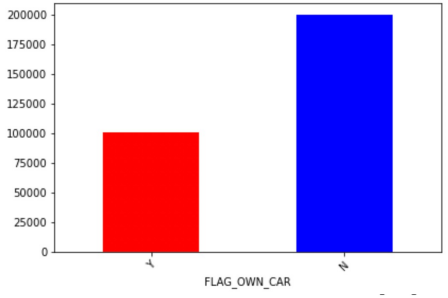
- Dataset is highly imbalance 92% for target 0 and 8% for target-1
- So its very clear to improve business we have to increase 92%
- And reduce 8%

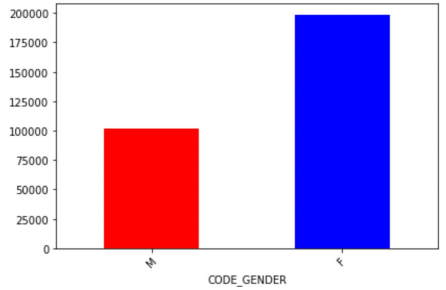
```
bank_cleaned.TARGET.value_counts(normalize=True)*100
```

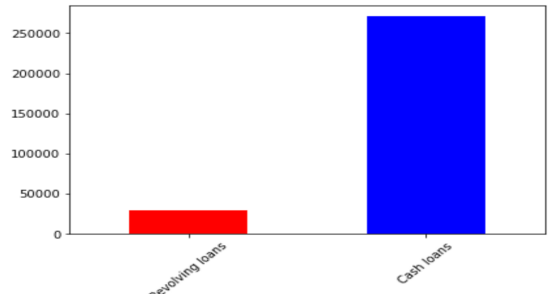
```
0     91.927118
1      8.072882
Name: TARGET, dtype: float64
```

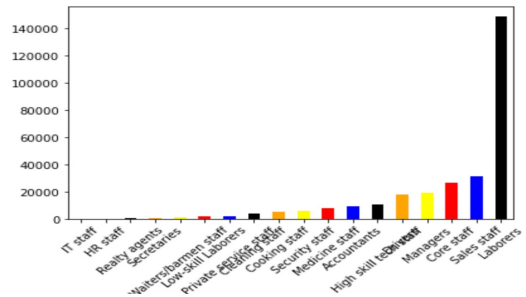
INTIAL DATA INSPECTION
categorical column//Graphical

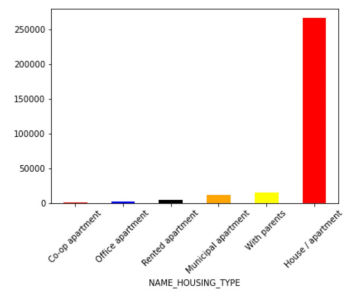


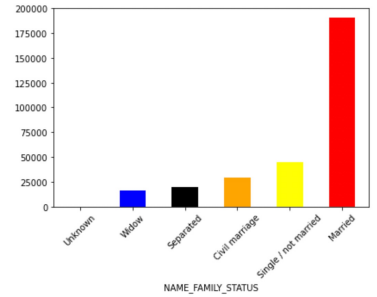


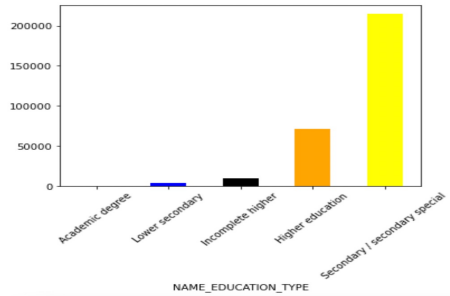


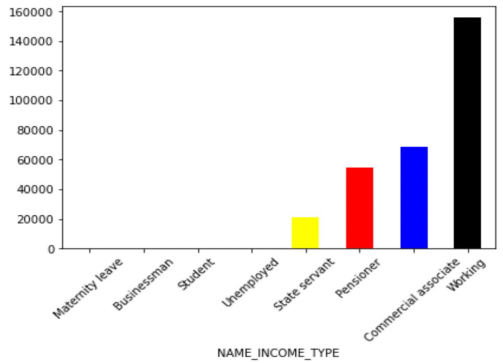


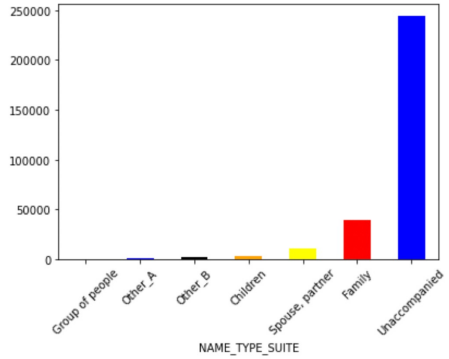










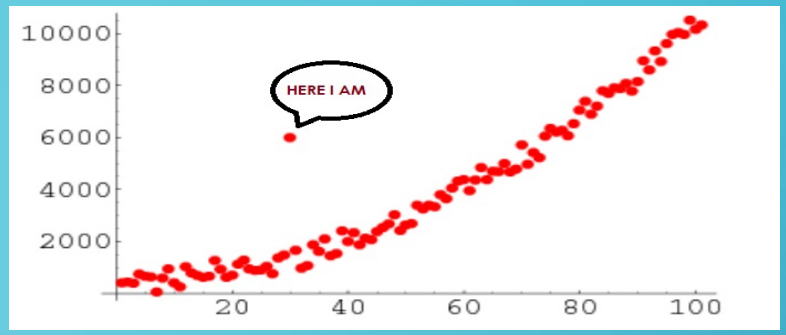


SEE NEXT SLIDE FOR DESCRIPTION

INTIAL DATA INSPECTION CATEGORICAL COLUMN //INSIGHT

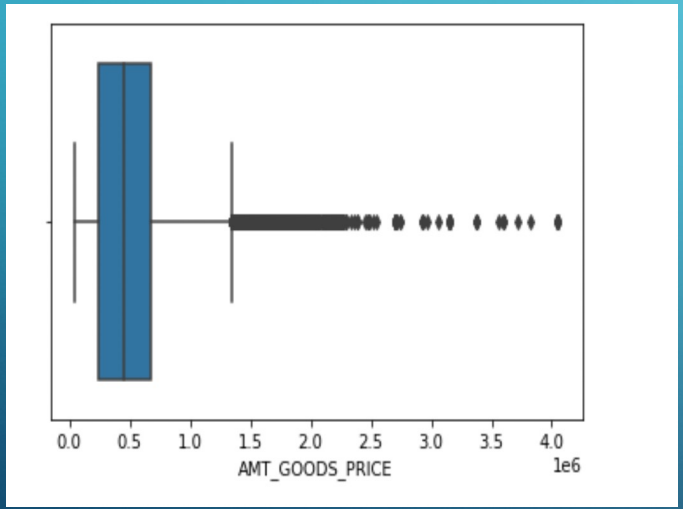
- 👉👉 Cash loan is in high demand compare to the revolving loan
- 👉👉In the data number of females are more than the number of male.
- 👉👉Most of the people applying for the loan does not own the car.
- 👉👉Most of the people has their own house applying for the loan.
- 👉👉A large number of people coming unaccompanied and after that second highest coming with the family.
- 👉👉Most of the working people applying for the loans.
- 👉👉Secondary Education and Higher Education applying are the large number applying for the loans.
- 👉👉People applying for the loans mostly are married.
- 👉👉Labourers and Business entity are the highest applying for the loan.

OUTLIERS IN DATASET

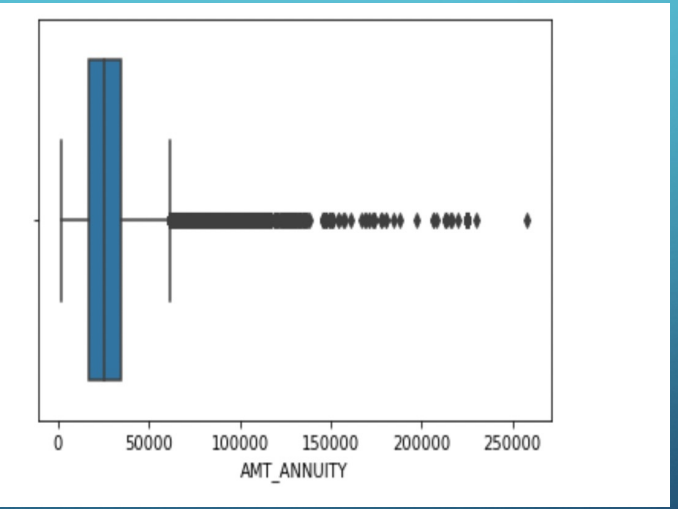


OUTLIERS ARE PRESENT IN THE DATASET I HAVE REMOVED VISUALLY WHERE THE CONTINUITY IS BREAKING.

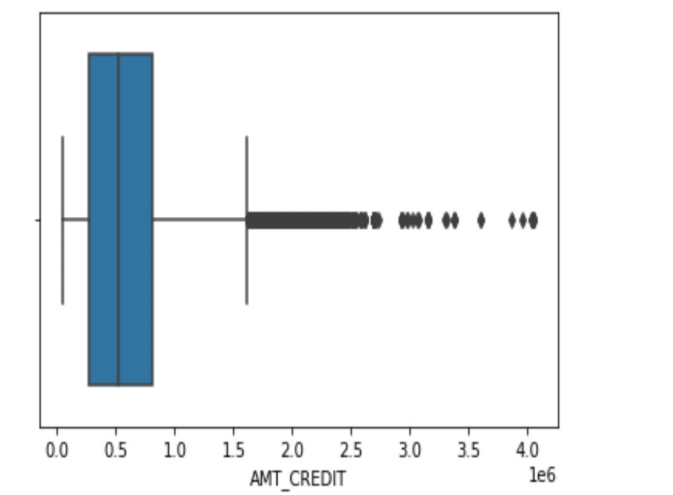
AMT_GOOD_PRICE



AMT_ANNUITY



AMT_CREEDIT



FEATURE SELECTION

```
#Feature selection in 4 parts
nominal_features = ['NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR',
                    'FLAG_OWN_REALTY','NAME_TYPE_SUITE', 'NAME_INCOME_TYPE',
                    'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE',
                    'OCCUPATION_TYPE','ORGANIZATION_TYPE', 'EMERGENCYSTATE_MODE']

ordinal_features = ['NAME_EDUCATION_TYPE','TOTALINCOME_BIN','BIRTH_BIN']

continuous_features = ['AMT_INCOME_TOTAL','AMT_CREDIT', 'AMT_ANNUITY',
                       'AMT_GOODS_PRICE','REGION_POPULATION_RELATIVE','FLOORSMAX_MEDI',
                       'YEAR_BIRTH','YEAR_EMPLOYED','YEAR_REGISTRATION',
                       'YEAR_ID_PUBLISH','YEAR_LAST_PHONE_CHANGE']


discrete_features = [ 'TARGET', 'CNT_CHILDREN','CNT_FAM_MEMBERS',
                      'REGION_RATING_CLIENT','REGION_RATING_CLIENT_W_CITY',
                      'Flag_doc_total', 'Commun_Total',
                       'Total_add_match', 'CODE_GENDER_NUM']
```

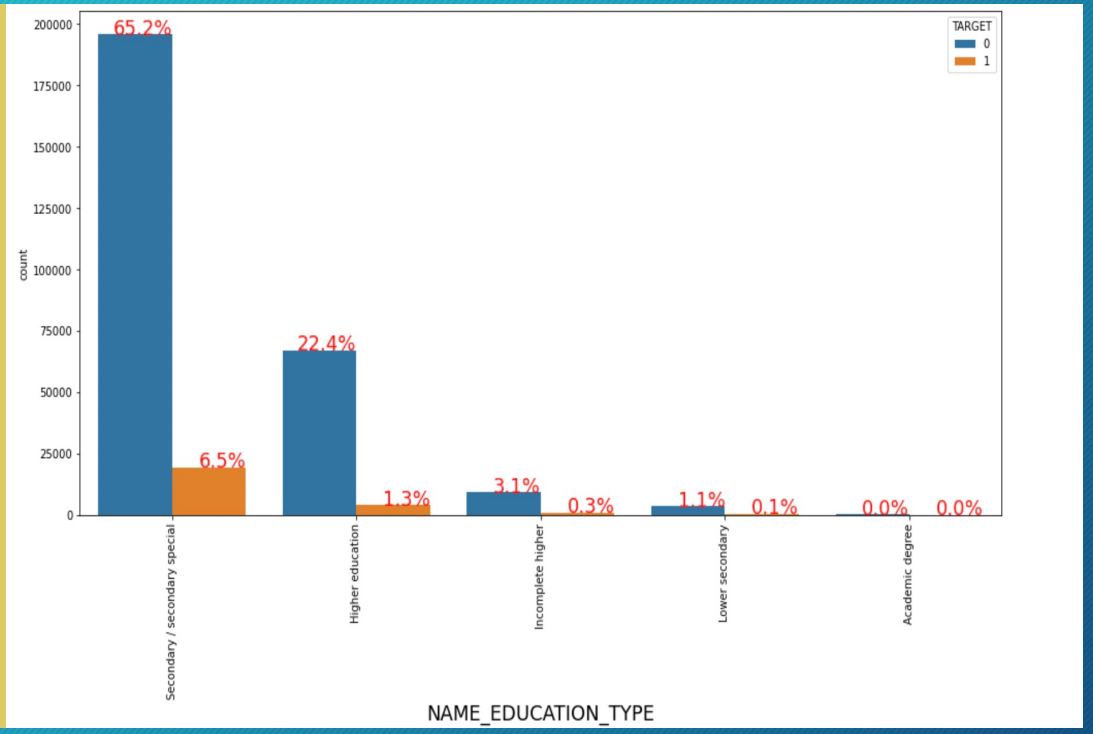
I have divided features in 4 part it will help selecting type of graph we require and how conclude story from graph.

Dependent variable

- All analysis we have to do against target variable
- Target0(92%) and Target1(8%)
- Target0-NonDefaulter
- Target1-Defaulter
- Greater >>>>92% Good customer
- Greater>>>>8% Poor customers
- So start with Univariate then Bivariate and Multivariate analysis.

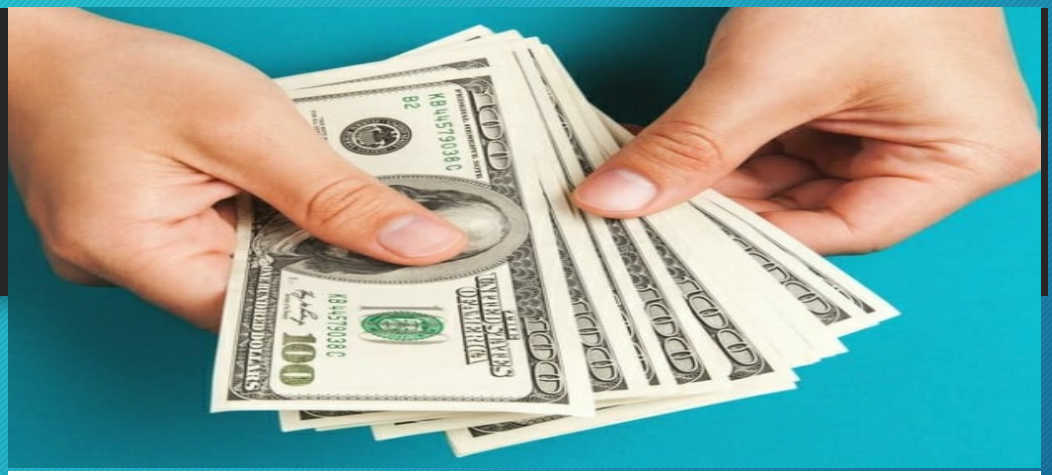
NAME_EDUCATION_TYPE

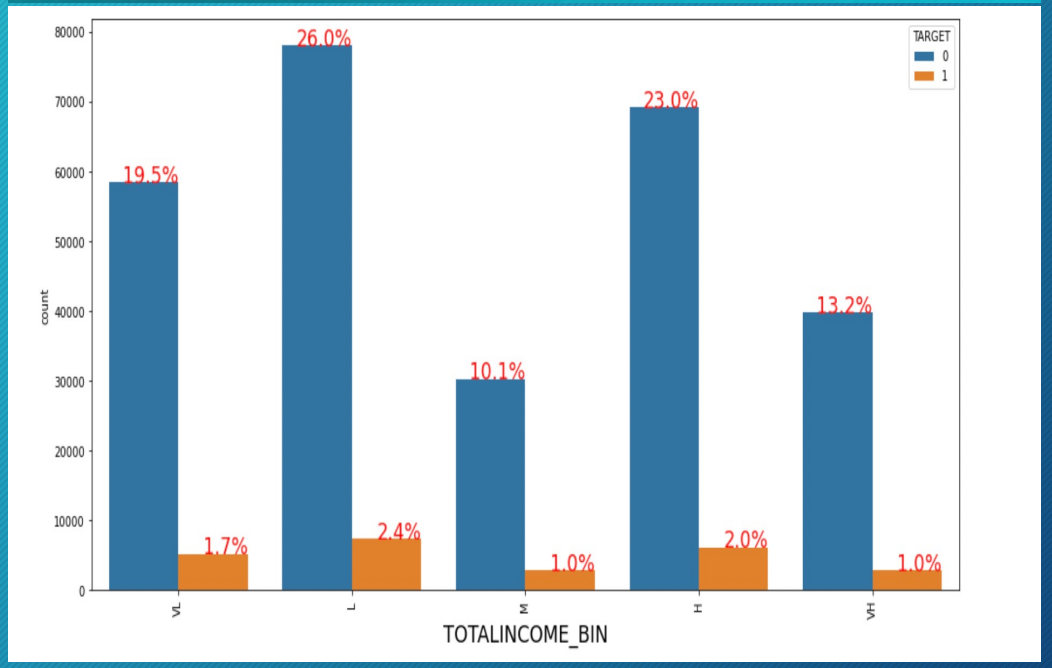
Secondary/special education and Higher Education are contributing almost 94% of the complete data.
In Secondary/special 0-65.2% and 1-6.5%
Higher Education 0-22.4% and 1-1.3
So we can say that Higher education is more better because Non_defaulters are 94.5 percent and defaulter 5.7percent.



TOTAL INCOME

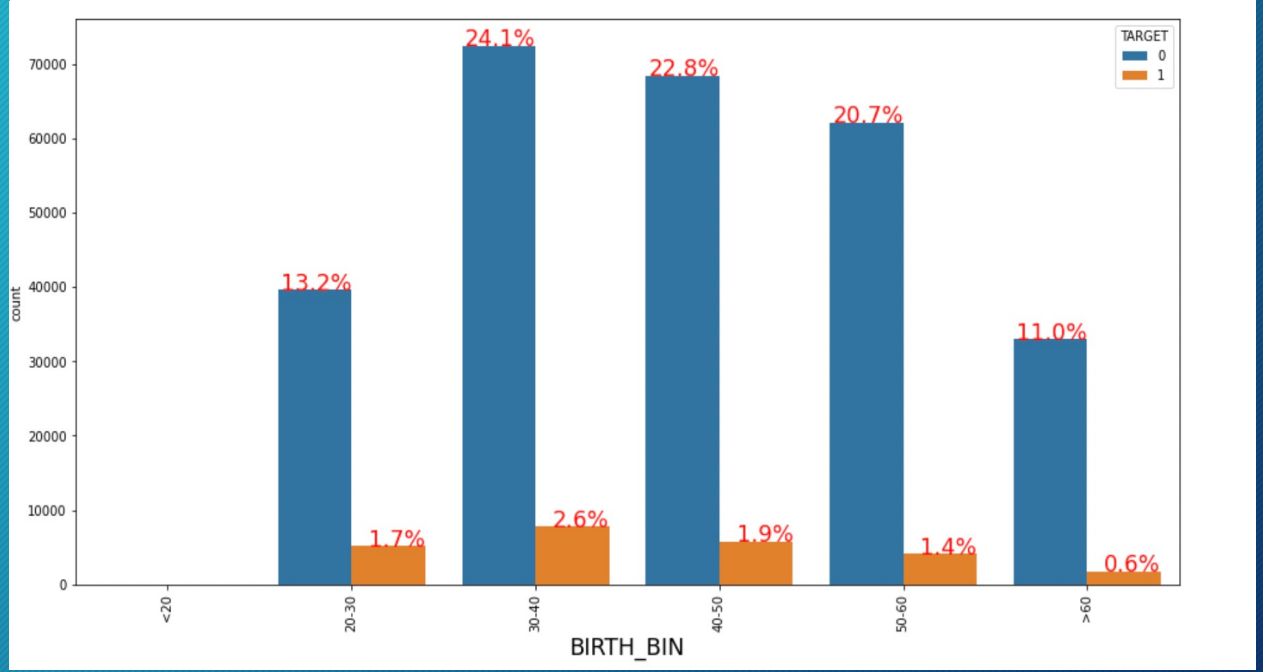
- Those who are having VH income and M(Medium) income are applying for the loan almost half of the other income type.
- Also people in the bin of high income are more BETTER comparing to others.





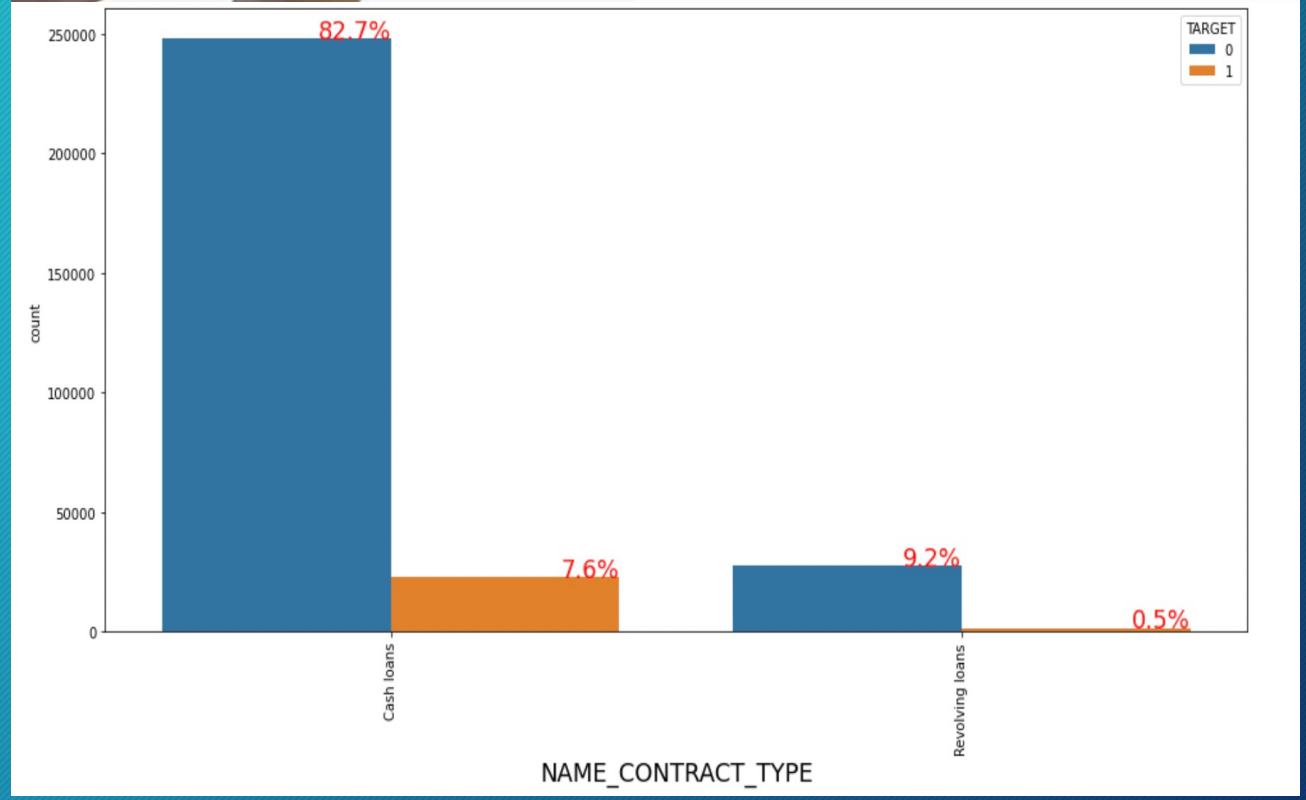
BIRTH_BIN(AGE GROUP)

- We can see from the graph that People in age bin of 50-60 are more reliable. Also people in bin 20-30 and 30-40 has highest number of defaulter compare to other.



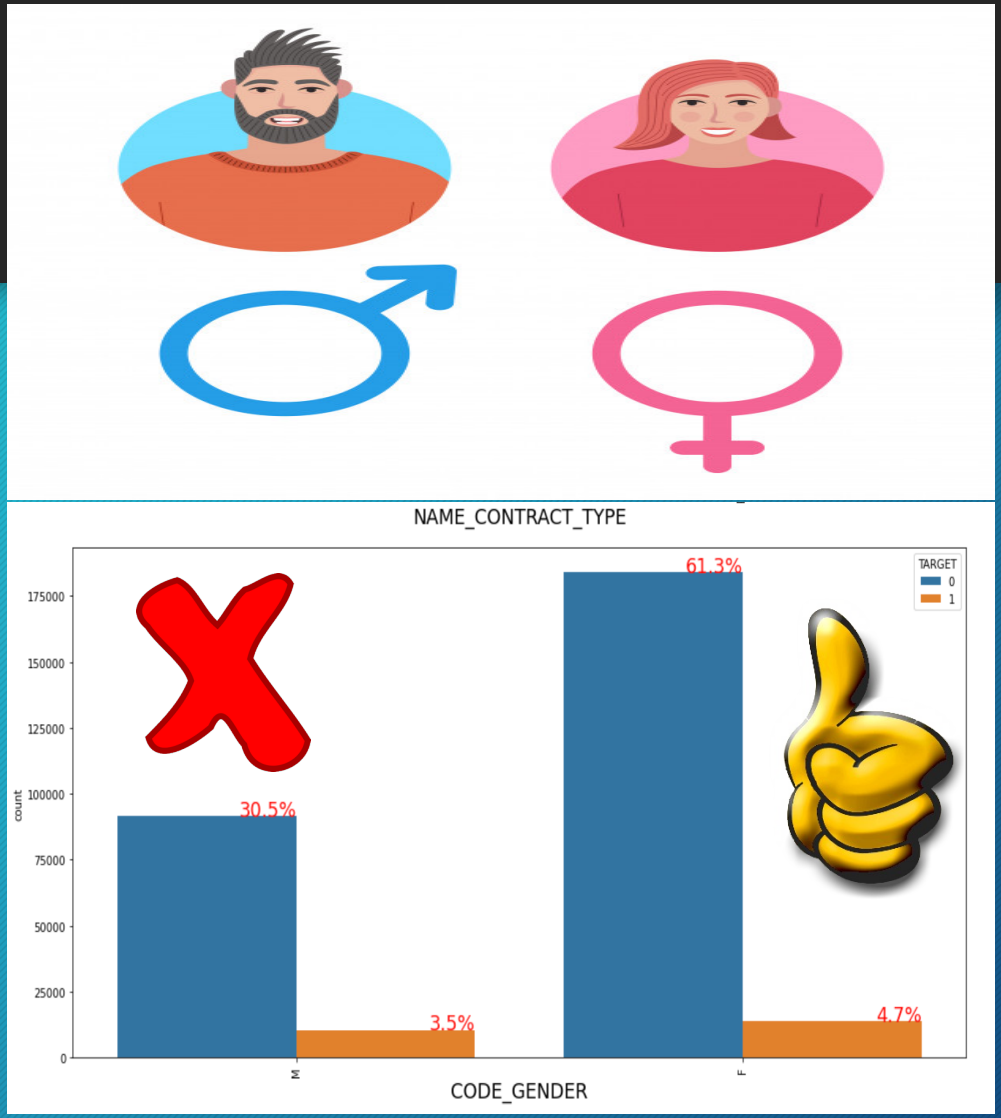
NAME_CONTRACT_TYPE

- Though we can see that 90% of the people applying for the cash loans and only 10% are in the revolving loan and we compare both number of defaulter are less in the revolving loan comparing to the Cash loan.



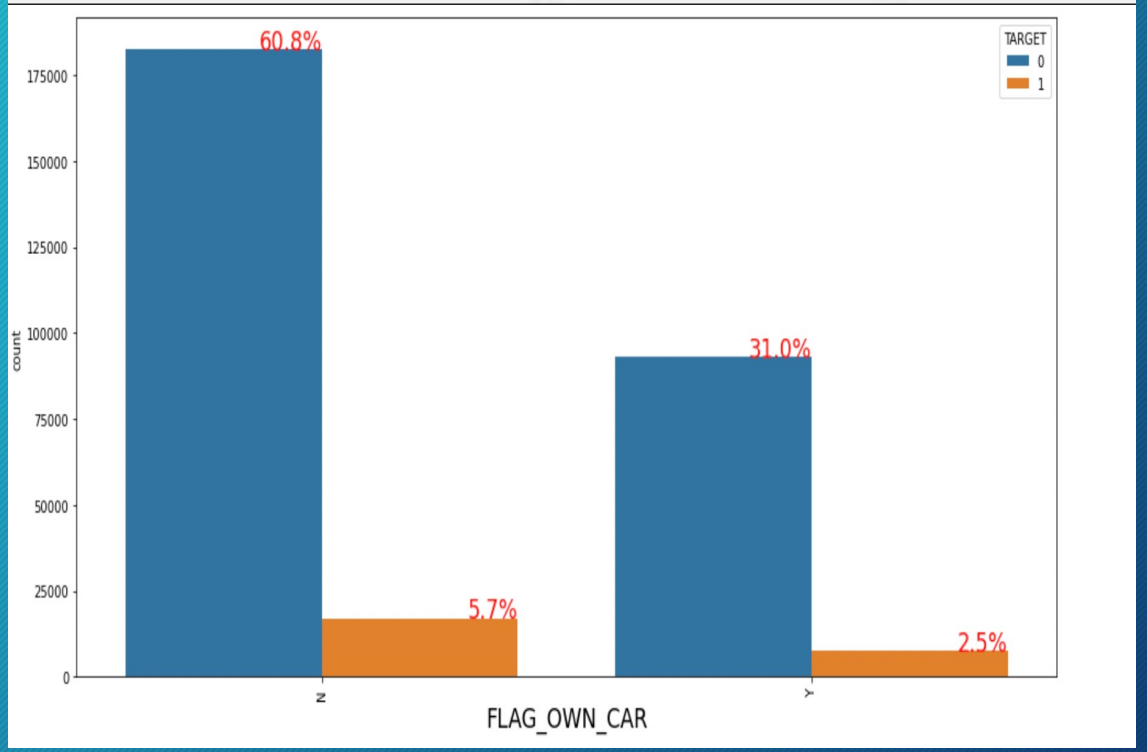
CODE_GENDER

- In application males are 34% and Females are 66%. Also we can see Male defaulter 10.3% and Non-defaulter 89.7%. For female defaulter 7.1% and non-Defaulter is 92.9%.
- So Female are good customer in terms of paying loans and also good in number for applying loans.



FLAG_OWN_CAR

- We can see those who are NOT having the car applying for the loan more than the people having the car.
- There is very little difference but we can say people having car are more reliable than who does not own the car.



FLAG_OWN_REALITY

- Those who have their own house are comparatively twice the people not having their own house and apartment
- but if we talk about the target0 and target 1 both are equal.
- So for business purpose bank should focus on the people who has their own house as these are higher Customers.

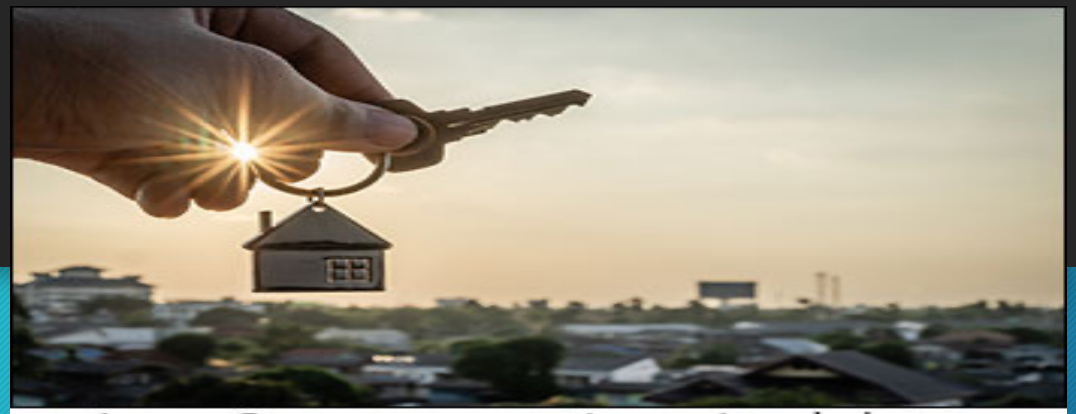
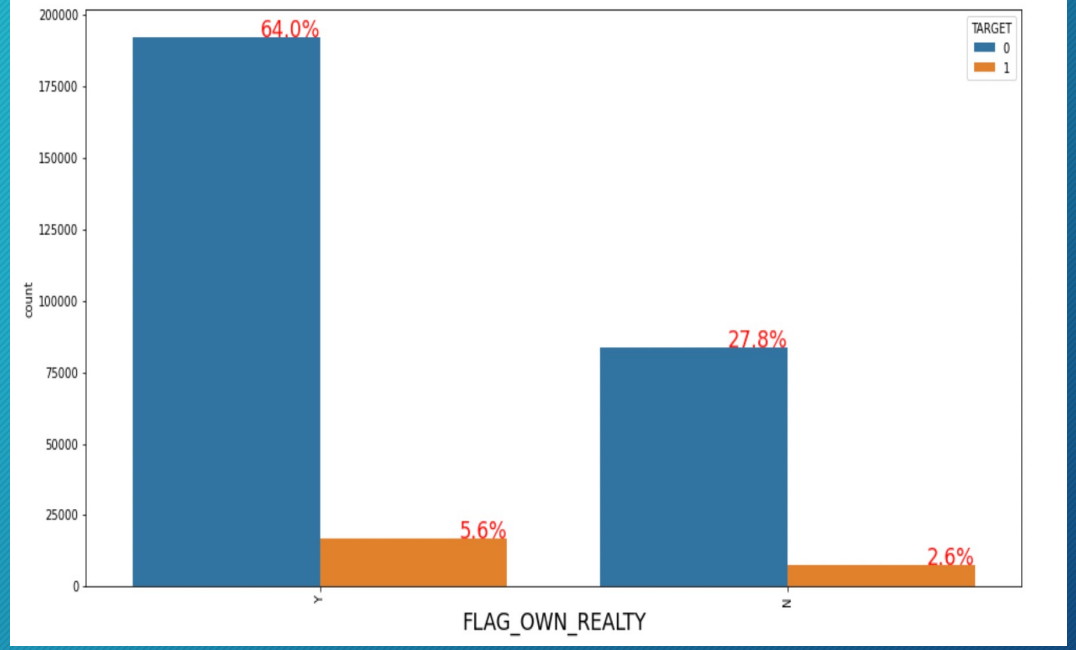
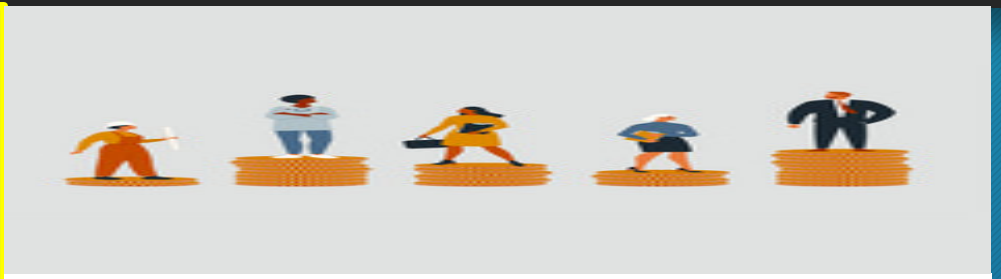


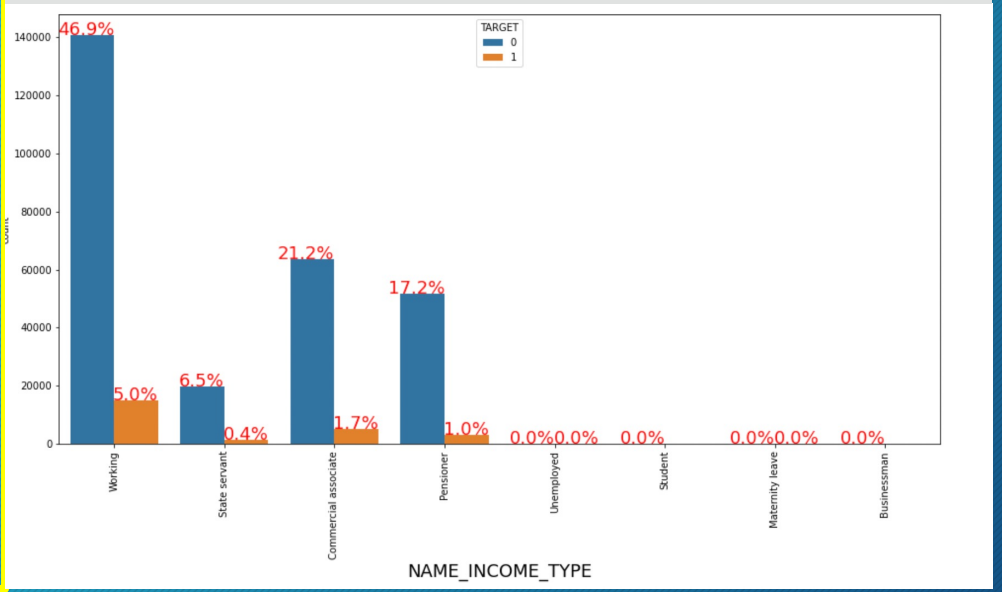
Image Source: sommart\www.istockphoto.com



NAME_INCOME_TYPE

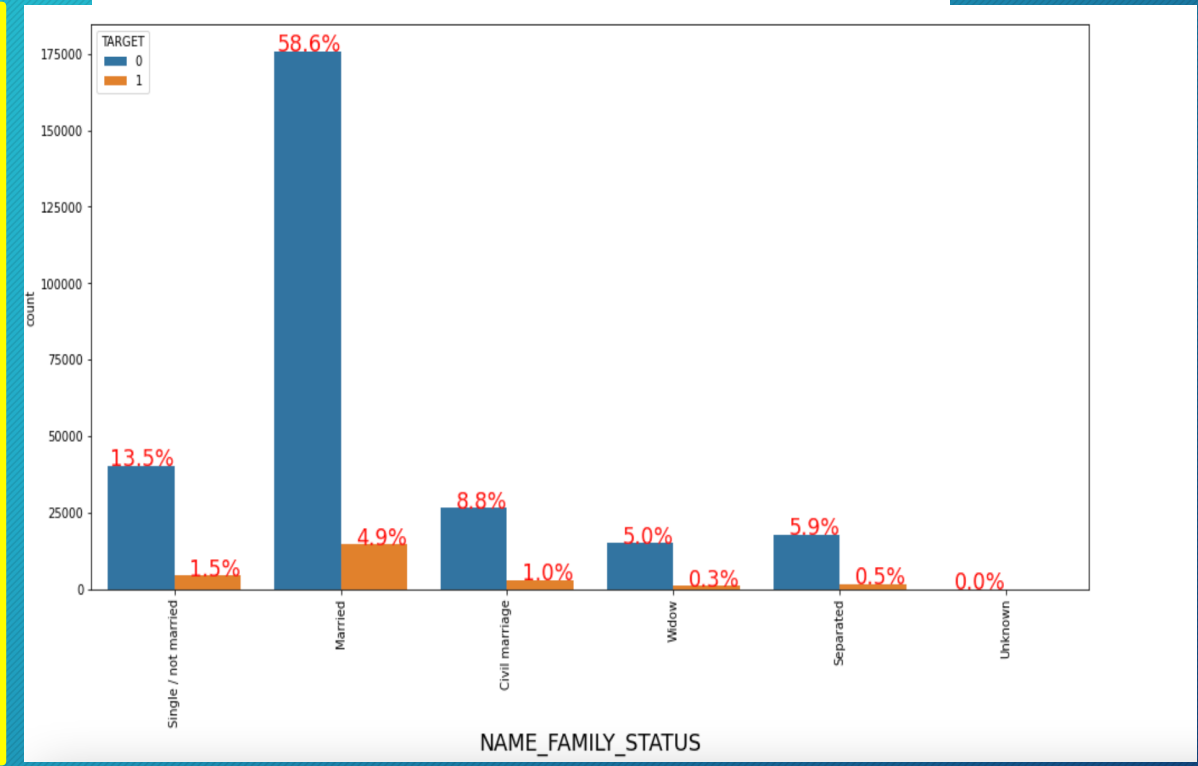
- Working (total-52%), commercial(Total-23%) and Pensioner(18%)
- These 3 are higher in number
- Number of defaulter are very less in Commercials and in Pensioner compare to the Working.
- Pensioners have 94.5% Non-defaulter and only 5.5% of the defaulter.



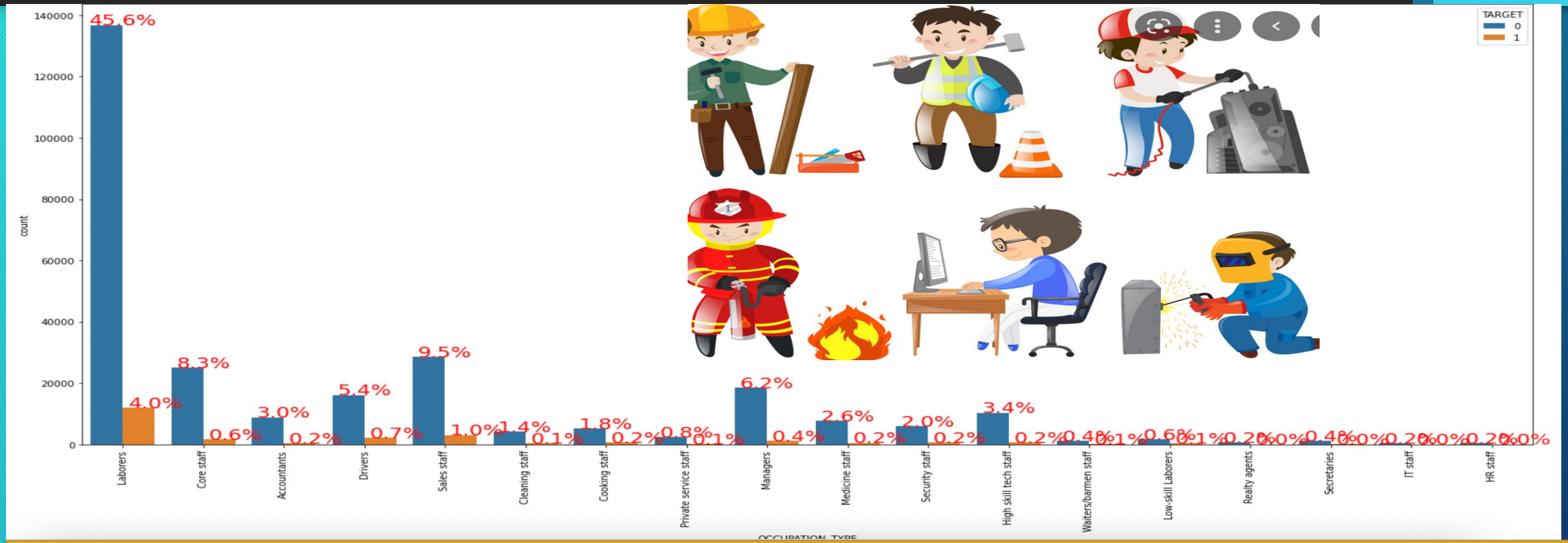


NAME_FAMILY_STATUS

- Total Married(63.5%),Single/Not Married(15%),Civil marriage(9.8%), Separated(6.4%) and Widow(5.3%).
- If we check for defaulter then Civil marriage and single/not married have highest number of defaulter about 10% .
- Widow has very less number of defaulter only 5.66%

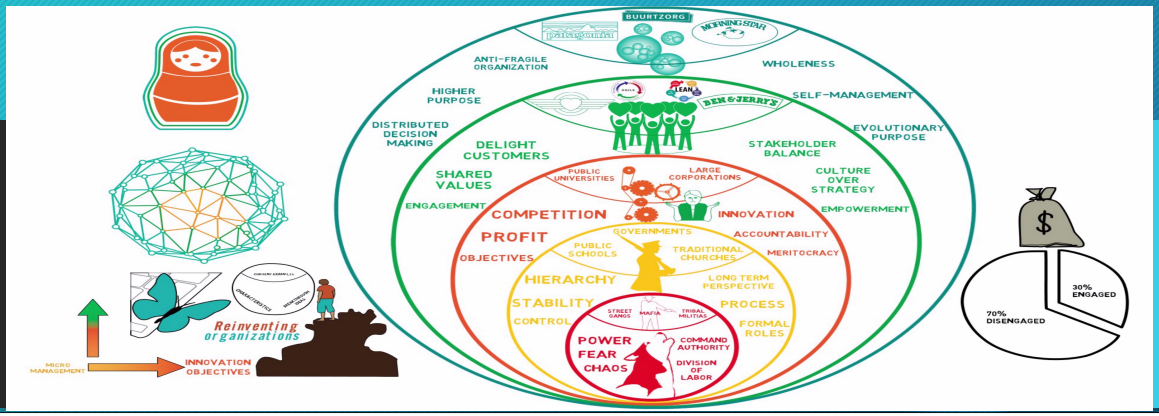


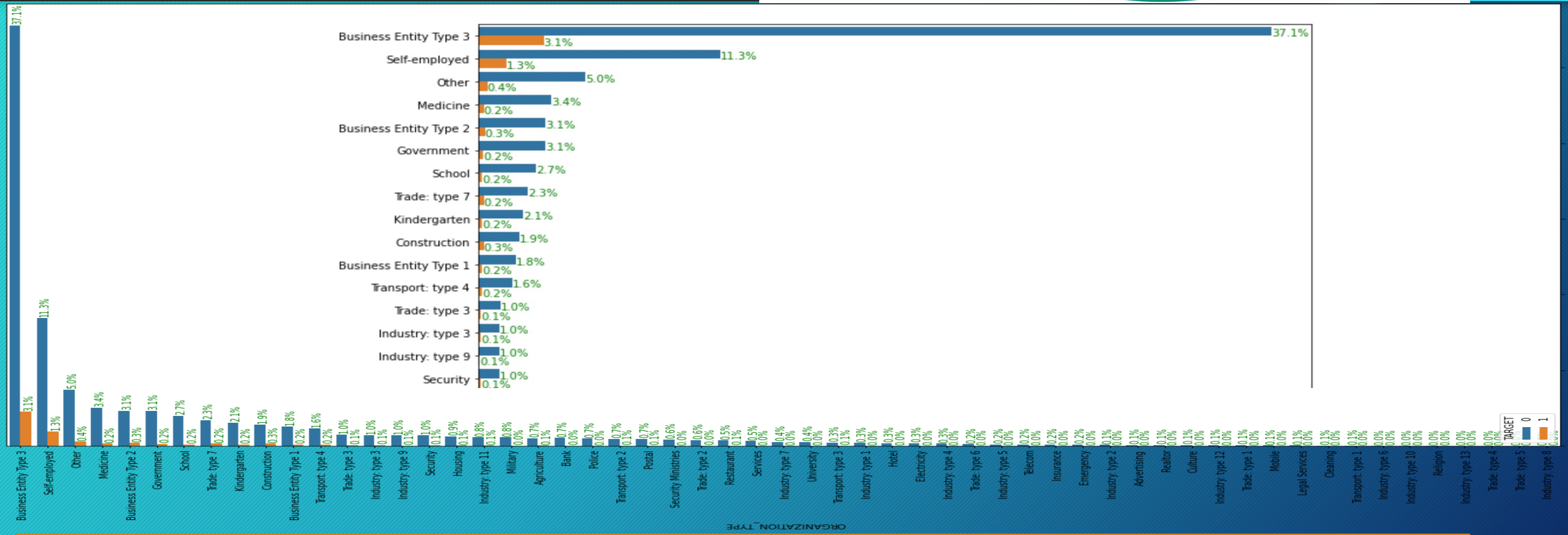
OCCUPATION_TYPE



Laboure's, core staff, managers and Sales staff is four highest category Laboure's are the highest Applicants almost 50% of the total. Numbers of defaulter are highest in Sales Staff on comparing these four. Also core staff and managers have very less number of defaulter.

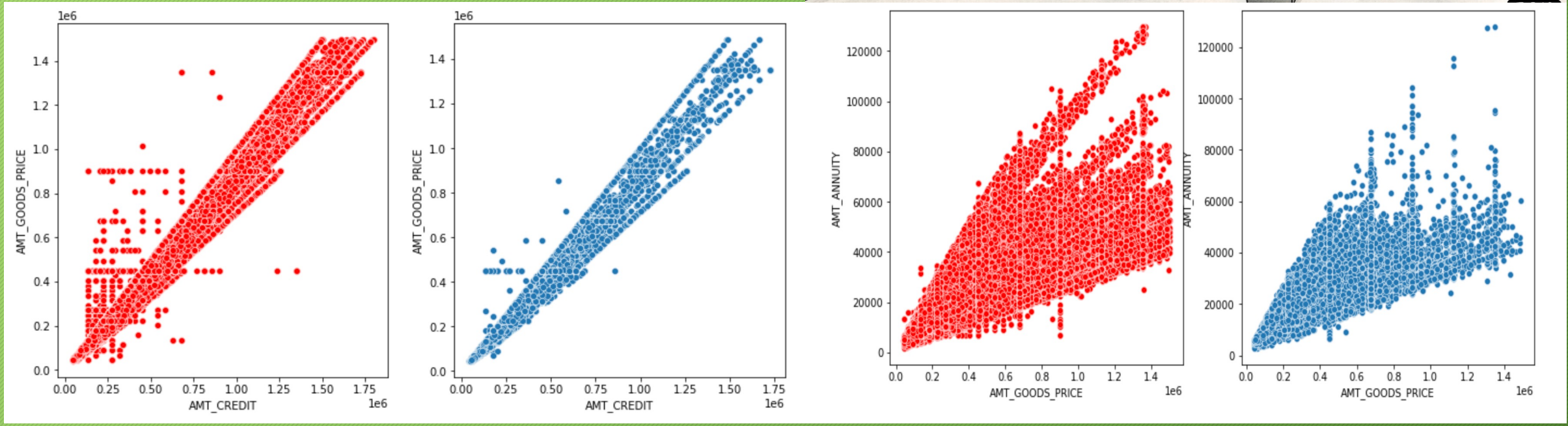
ORGANIZATION_TYPE





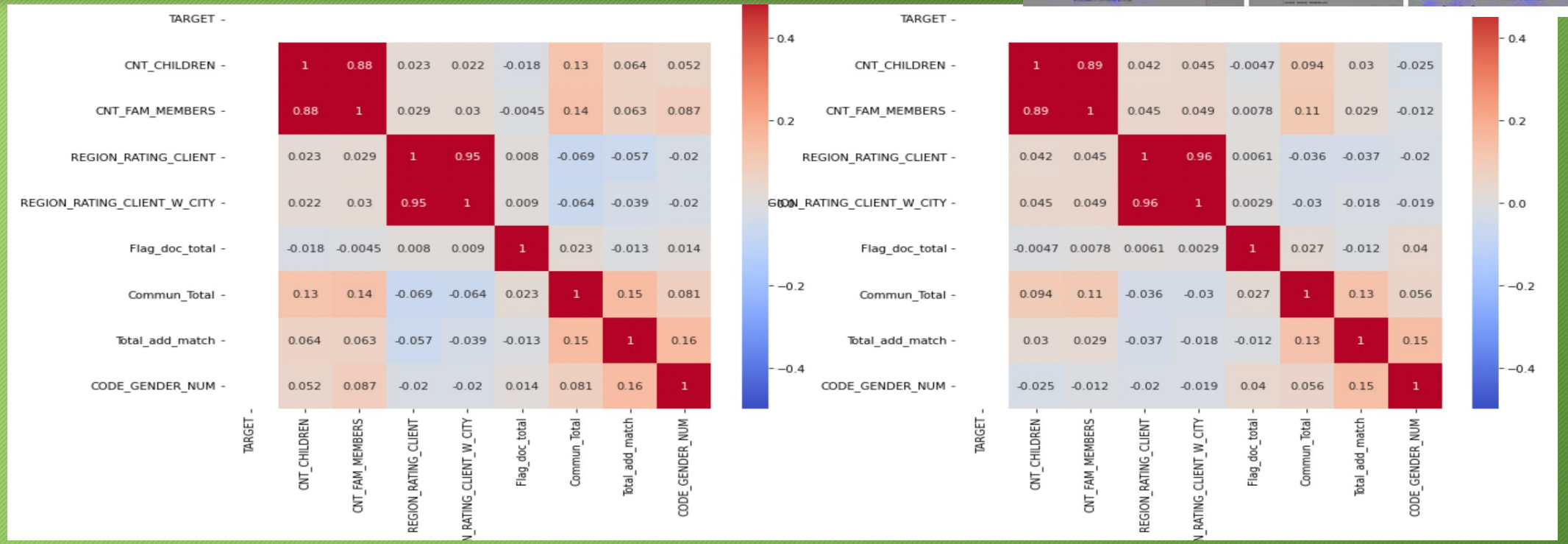
Organization type-Business Entity 3 and Self employed are highest in demanding loan. For defaulter and no defaulter its given in detail in upcoming slides.

BIVARIATE-MULTIVARIATE



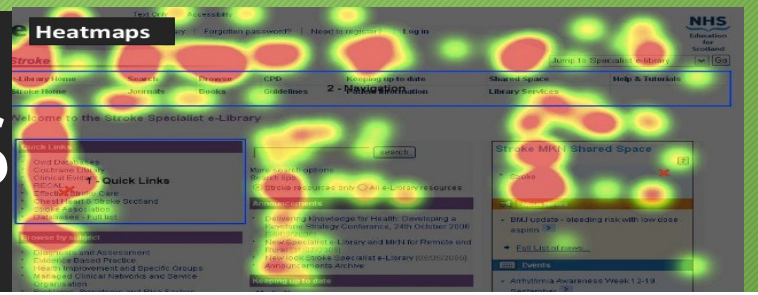
STRONG CO RELATION EXIST BETWEEN THESE FEATURES

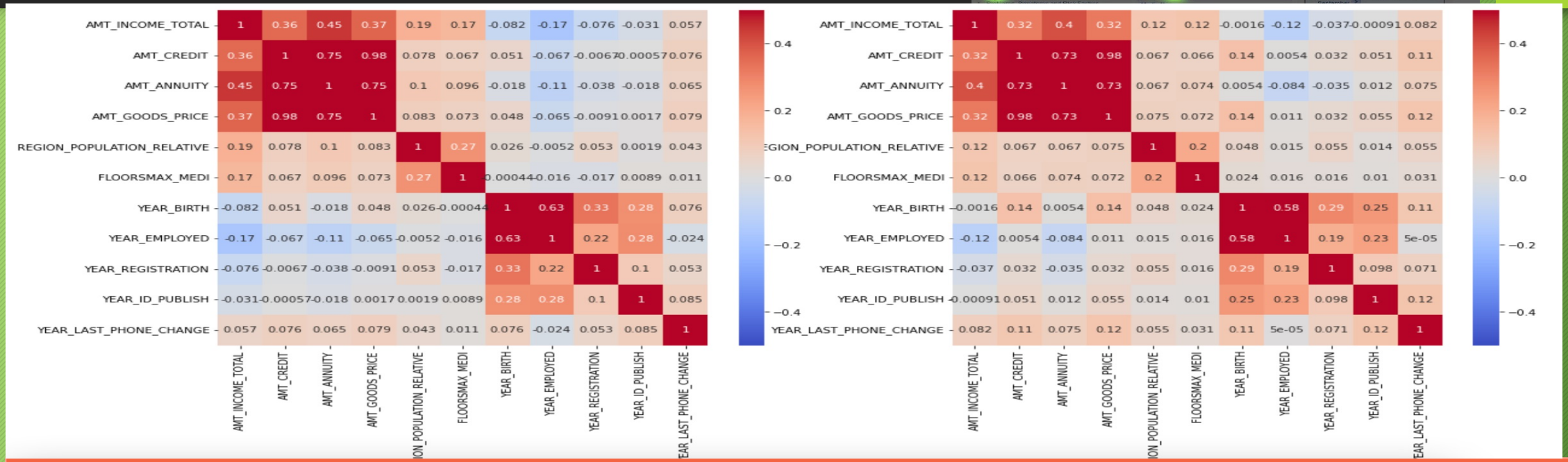
HEAT MAP DISCRETE FEATURES



DISCRETE FEATURES-we can see cnt_childern and cnt_family member has strong co relation and region rating client and region rating client w city and strong co relation.

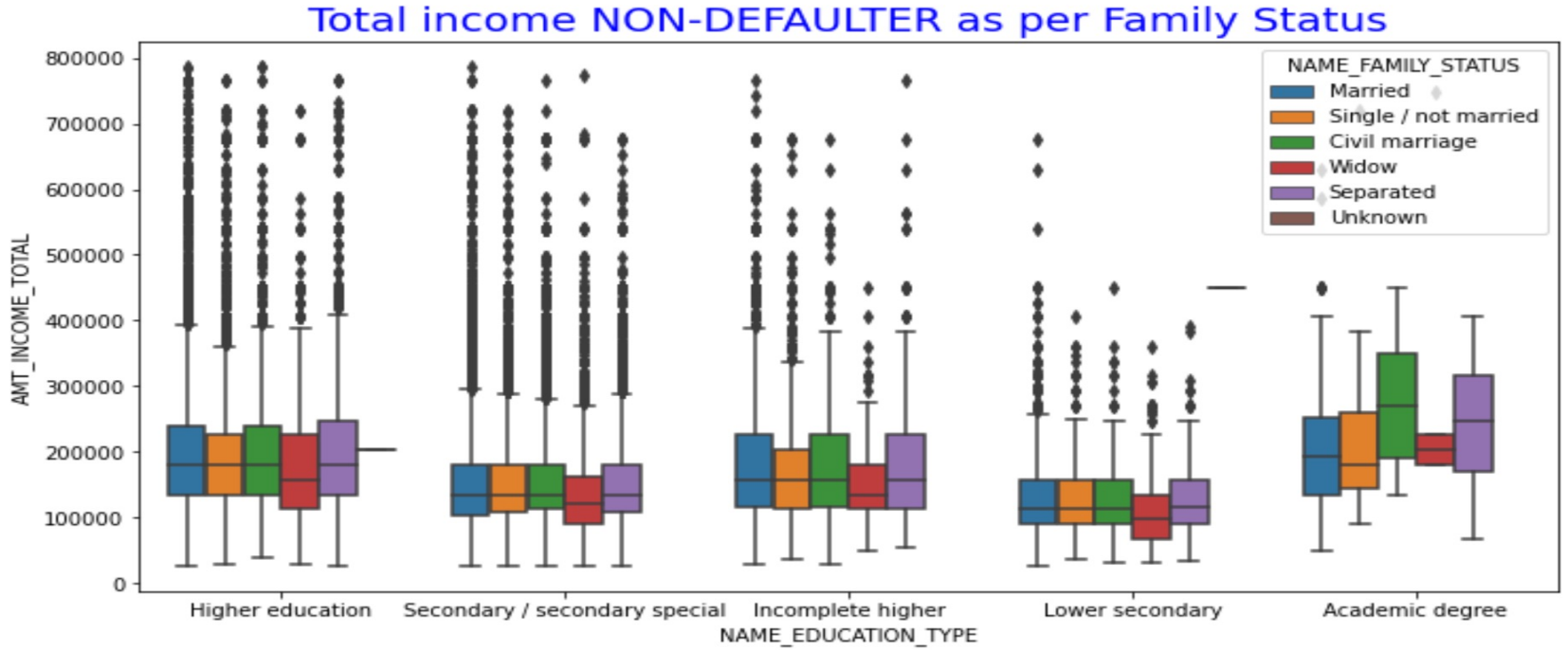
HEAT MAP CONTINIOUS FEATURES





CONTINIOUS FEATURES- WE can see two patches one at the left hand upper corner and one at the right hand lower corner. these both have 4 columns and have have strong correlation among themselves. 1st patch has 4 column namely-'AMT_INCOME_TOTAL','AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE' and other group have YEAR_BIRTH', 'YEAR_EMPLOYED', 'YEAR_REGISTRATION', 'YEAR_ID_PUBLISH','YEAR_LAST_PHONE_CHANGE'

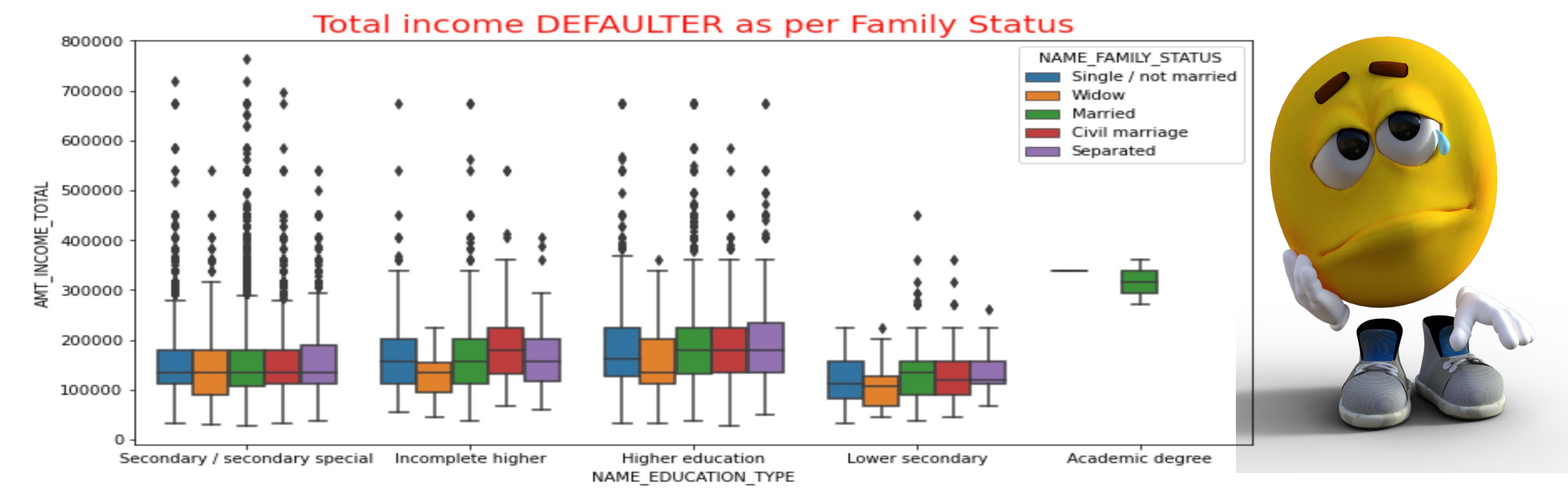
INCOME/EDUCATION/FAMILY STATUS/NON DEFAULTER



Family having the academic degree have higher income compare to other. Also we can see in academic degree those who did civil marriage have higher income among all.

Income range of the lower secondary is very low comparing to all.

INCOME/EDUCATION/FAMILY STATUS/ DEFAULTER

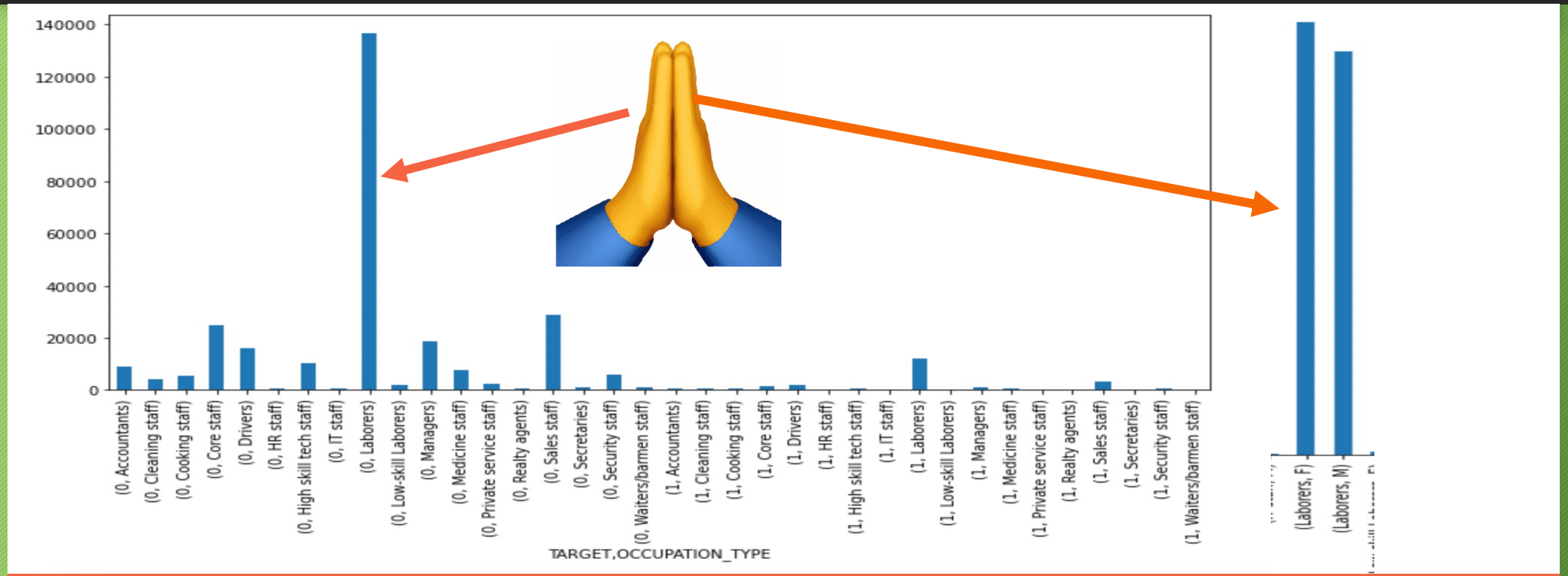


Income range is low of all the defaulter compare to the non defaulter

Income range of widow is less compare to other in same education type.

Academic degree Except married people there are no other defaulter present in the graph

OCCUPATION TYPE/CODE_GENDER/TARGET

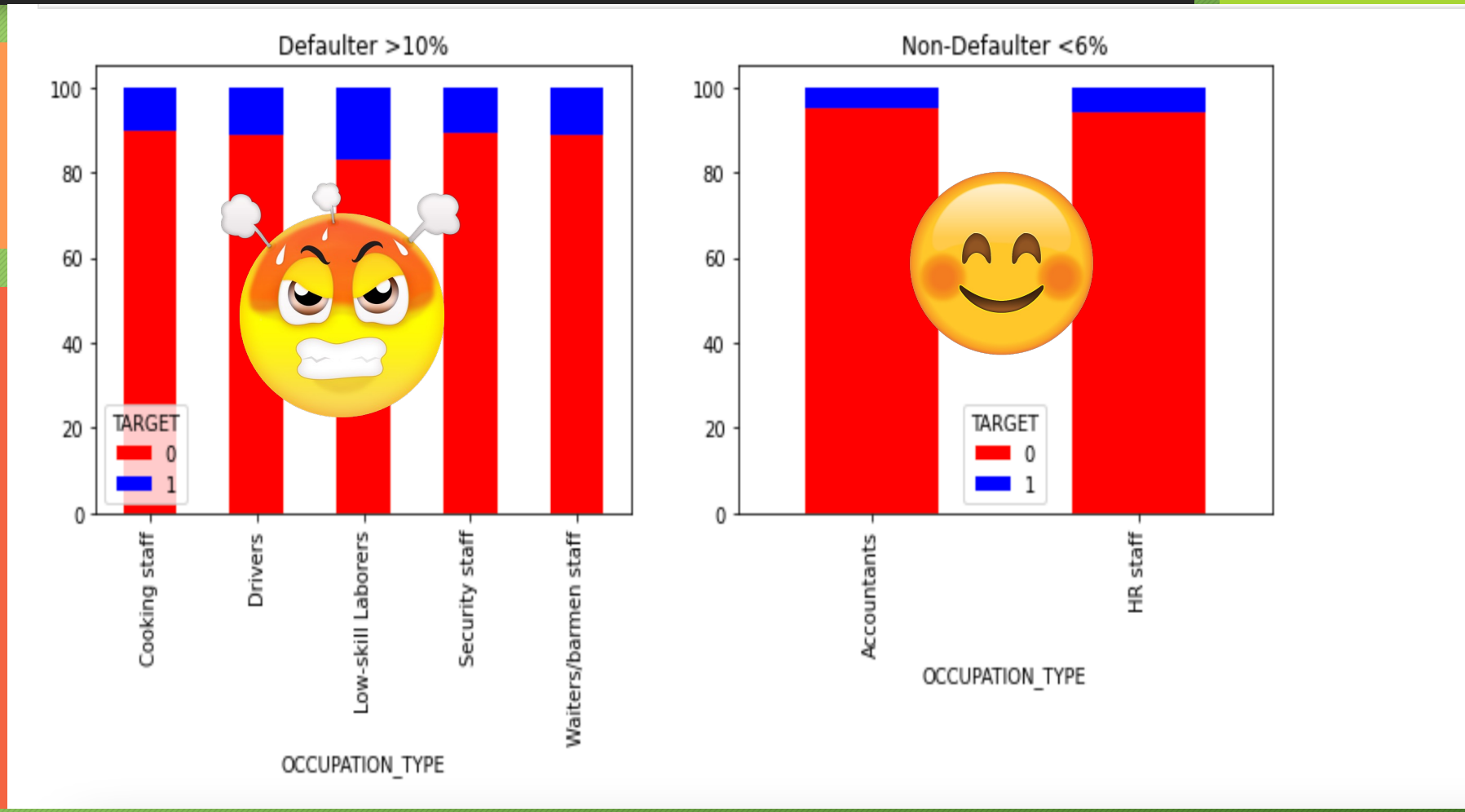


We can see Laboure's (both male and female), Sales staff(female), Drivers(female) and core staff(Female) are mostly applying for the loans. But from here we can not conclude who is the good or bad customers for bank.

OCCUPATION TYPE- DEFAUALTER//NON DEFAULTER

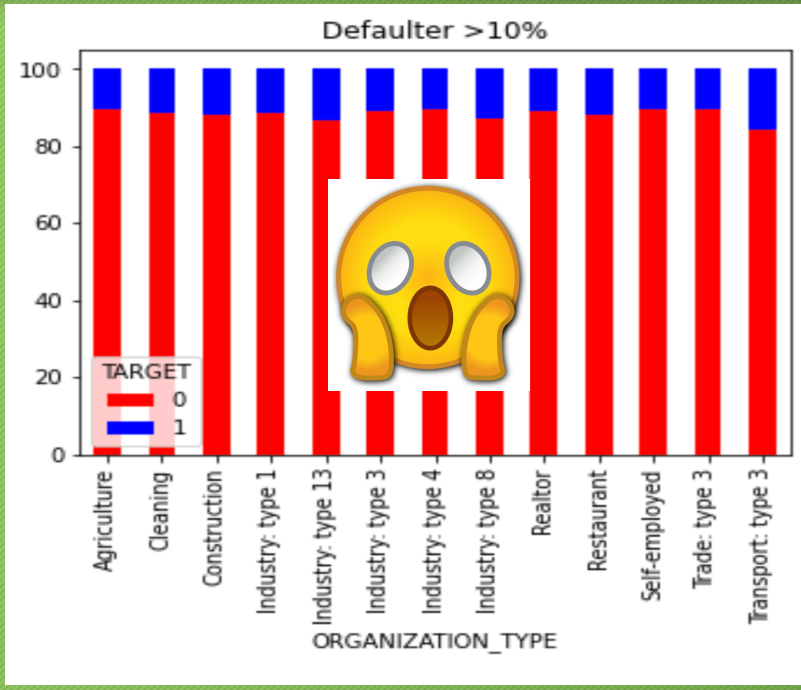
😊😊 Accountants and HR staff are the good as being Non-Defaulter

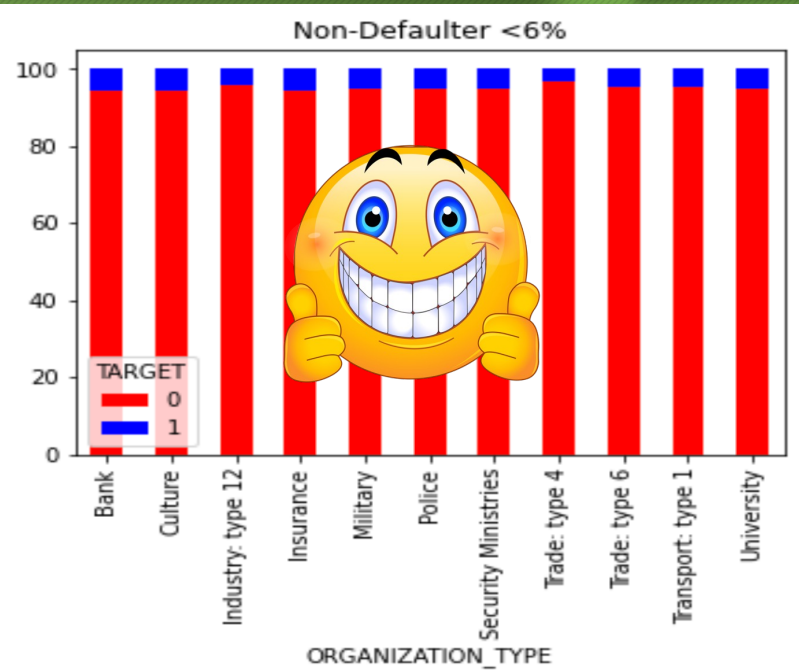
😥😥 Cooking staff, Drivers,low skill laborer's, security staff and waitress/barmen staff are poor performer for replaying the loan amount



ORGANIZATION_TYPE - DEFAUALTER//NON DEFAULTER

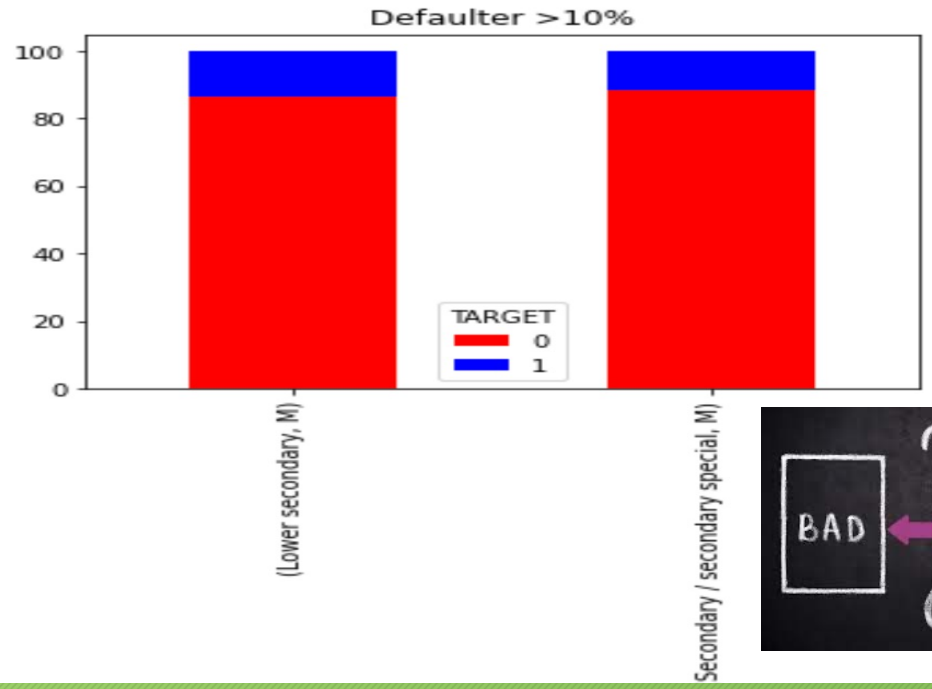
😊😊 Good customer- Bank,Culter,Industry type12,Insurance,Military,Police,Security Ministries, Trade type-4,trade type 6,Transport type-1 and university

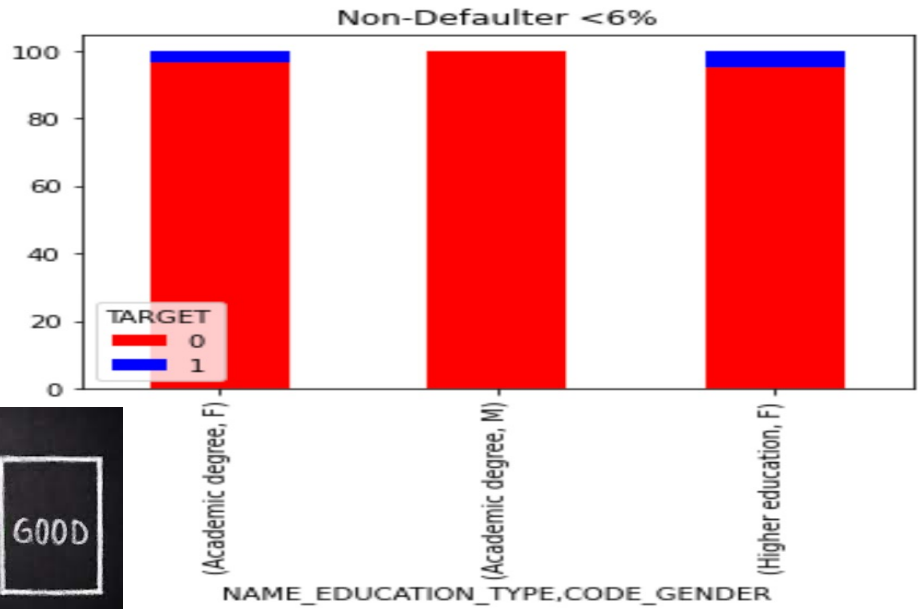




😢😢 Poor customer- Agriculter,cleaning,construtions, industry type-1, industry type13,industry type3,industry type4,industry type8,Realtor,Restaurant,Self Employed,Trade Type-3 and Transport type 3

NAME_EDUCATION_TYPE + CODE_GENDER

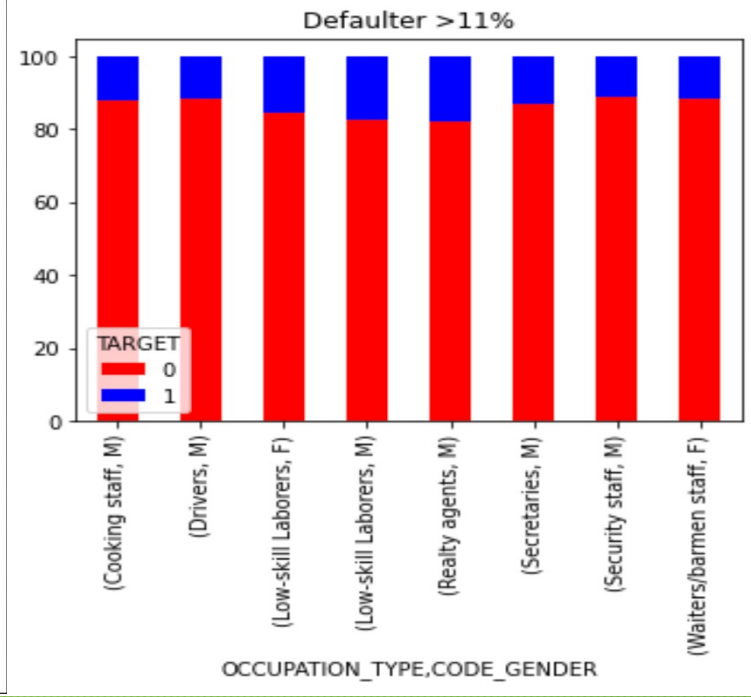


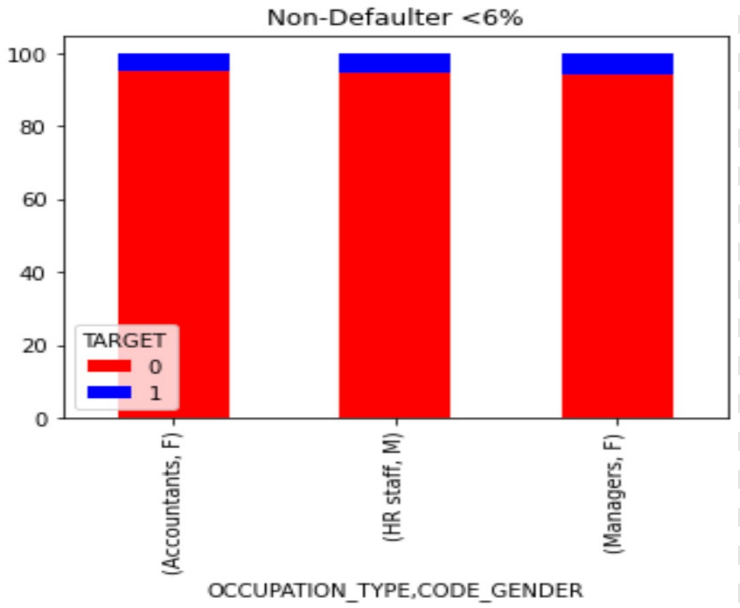


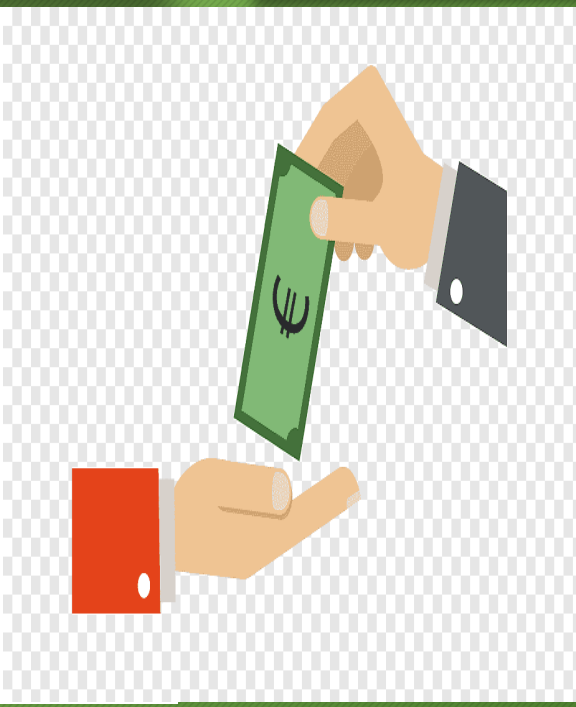
😊😊 Female having higher education and both male and female having academic degree have good record to being as non defaulter

😢😢 Poor customer - Males belongs to Lower secondary education and Secondary/Secondary special are more defaulter.

OCCUPATION_TYPE + CODE_GENDER





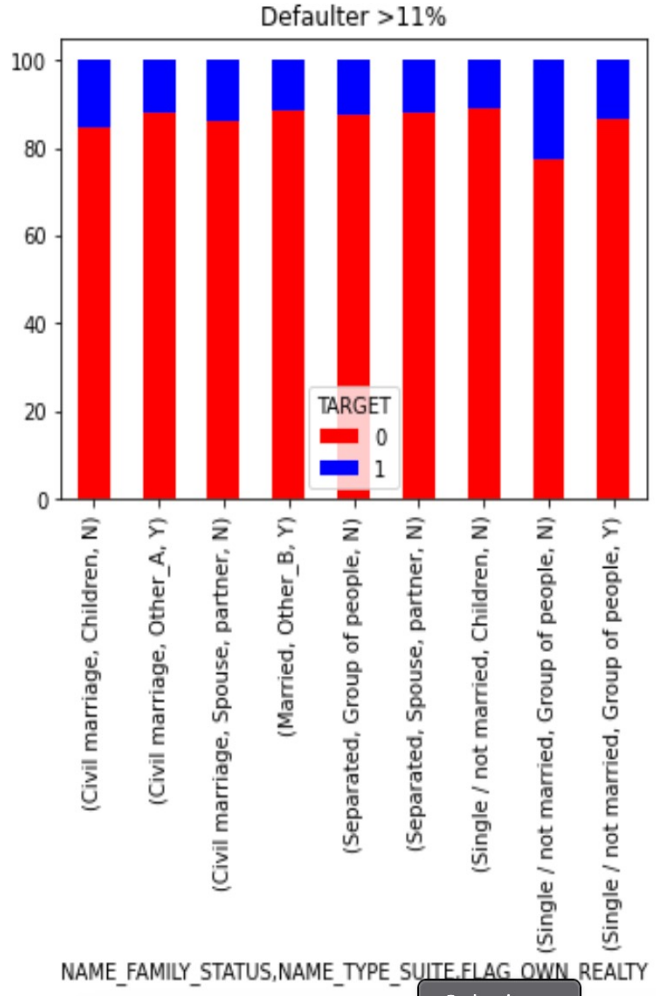


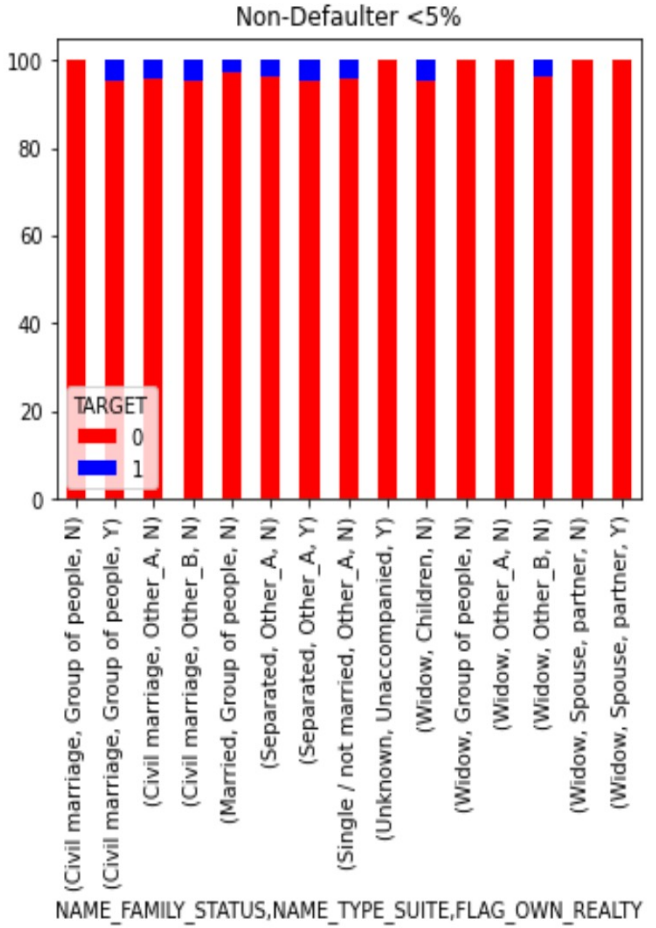
😊😊 Good customer - Female having higher education and both male and female having academic degree have good record to being as non defaulter

😥😥 Poor customer - Males belongs to Lower secondary education and Secondary/Secondary special are more defaulter.

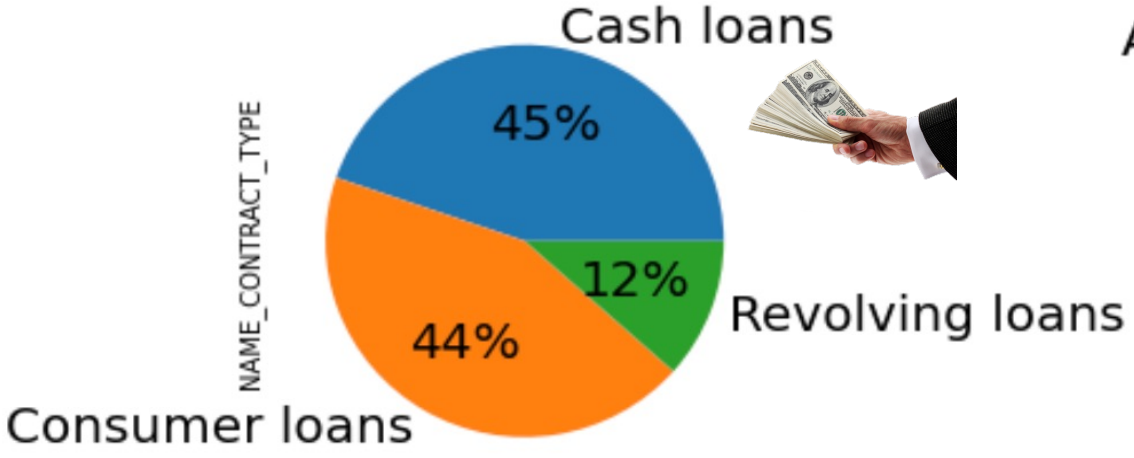
NAME_FAMILY_STATUS+ NAME_TYPE_SUITE + FLAG_OWN_REALTY

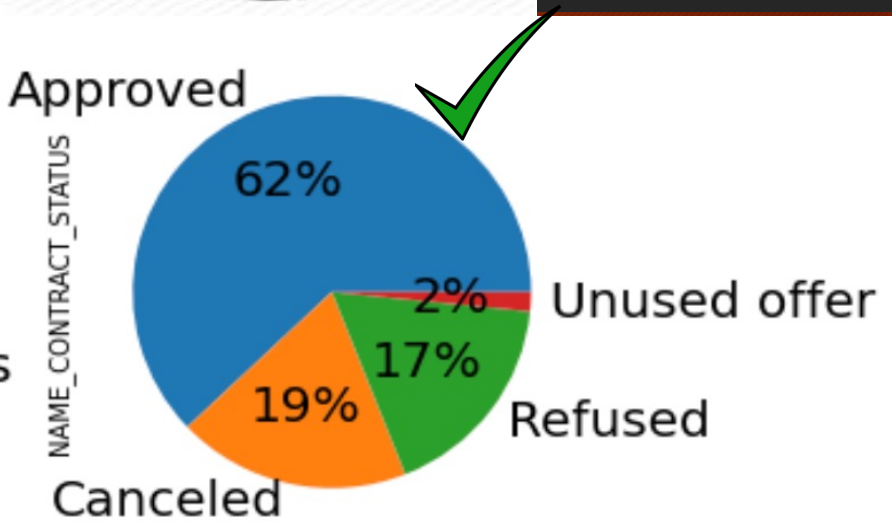
Group of people single/not married does not have house or apartment are 22.7 percent in defaulter category





PREVIOUS APPLICATION

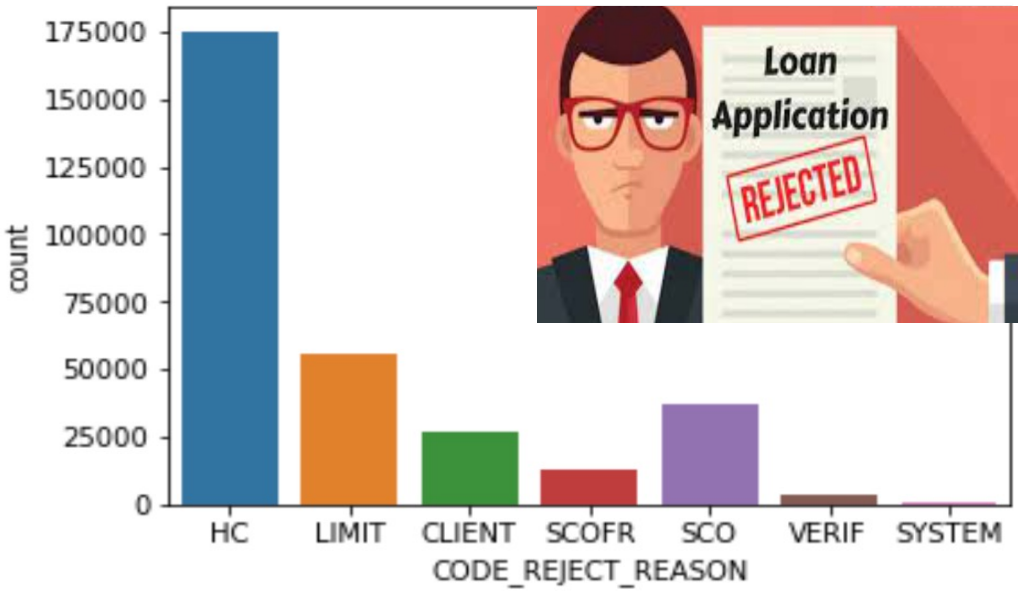


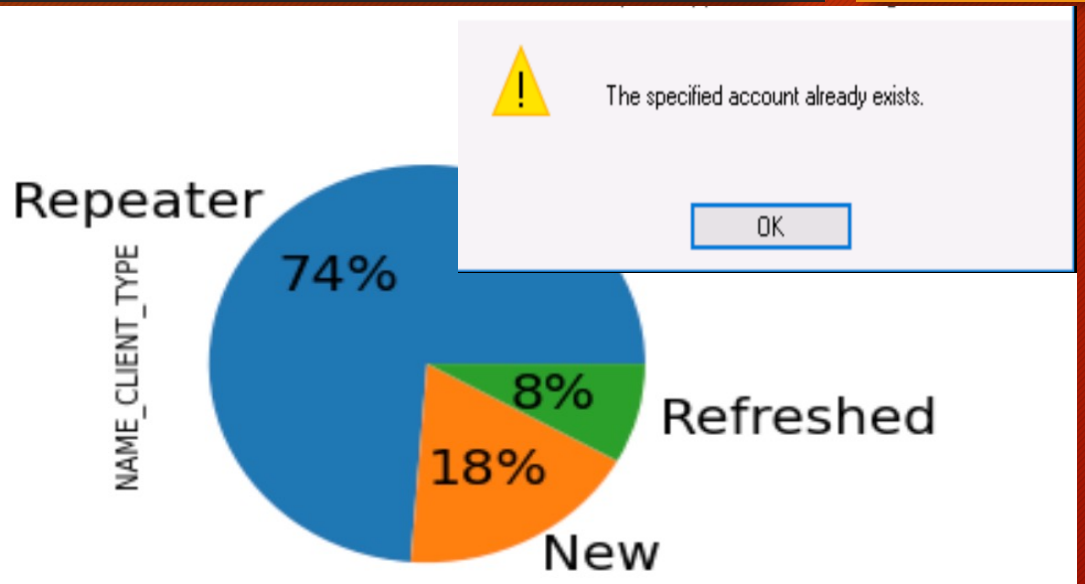


Consumer and Cash loans are almost same and high in number but revolving loans are very few.

Very few loans are the unused offer, approved loans and highest in number,
Refused are slightly less than Cancelled.

PREVIOUS APPLICATION DATA



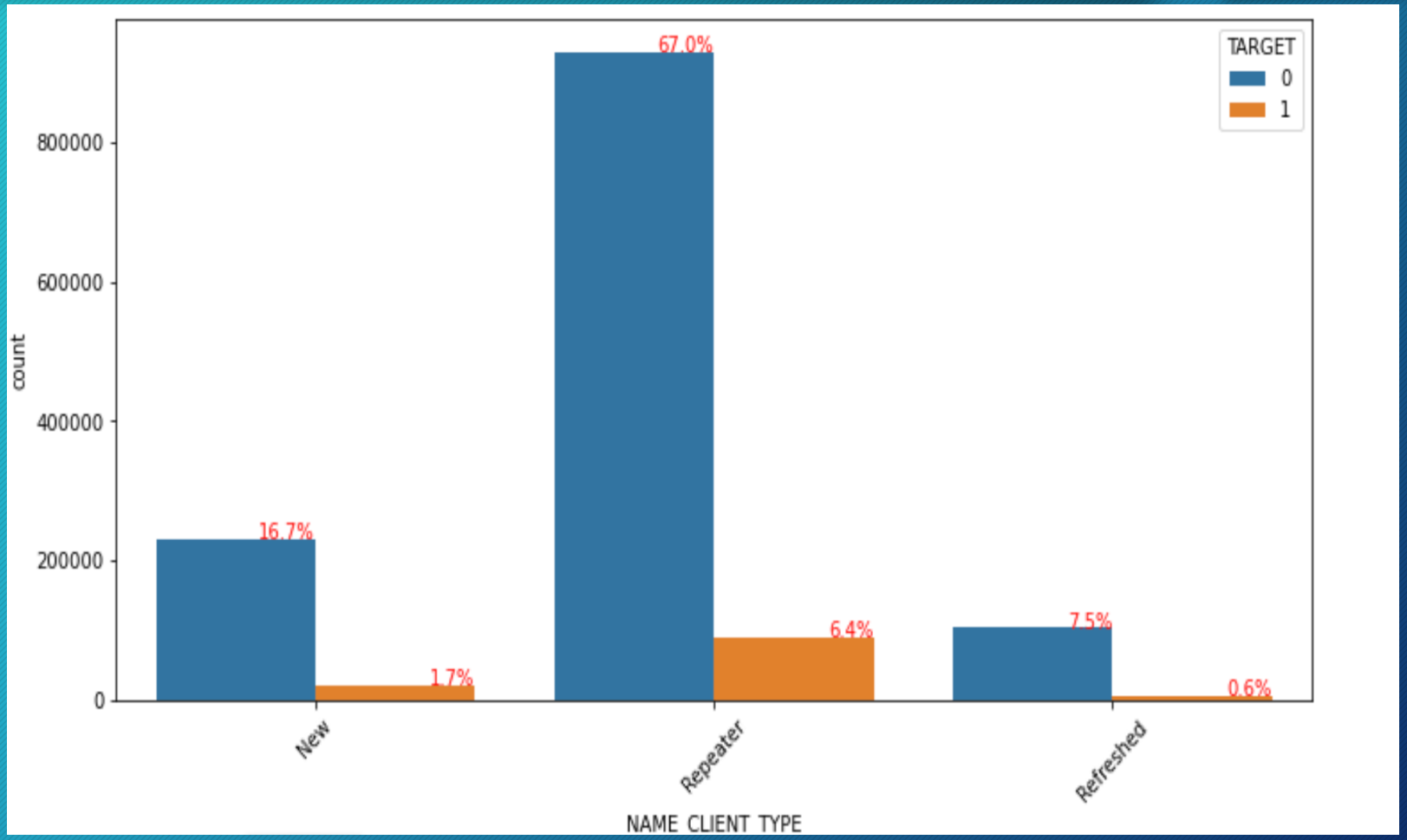


The reason 'High Credit'(HC) is highest for previous application rejection. Second highest reason for rejection is Limit.

Most of applications are repeaters.

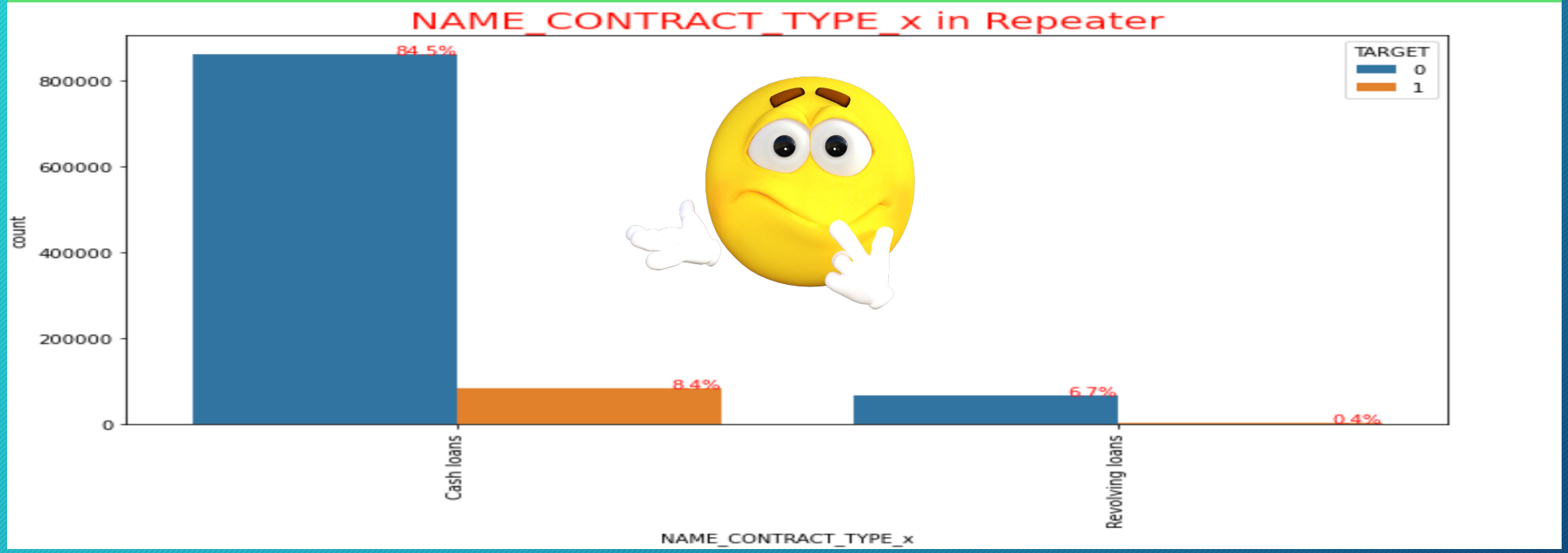
COMBINED DATASET

Repeater are highest in number and if we see defaulter vs non defaulter then there is hardly much difference but still we can say Refreshed is little comparatively better but we will segregate data and view some in-depth analysis.



NAME_CONTRACT_TYPE_x HUE AS TARGET

Repeaters in Combined dataset

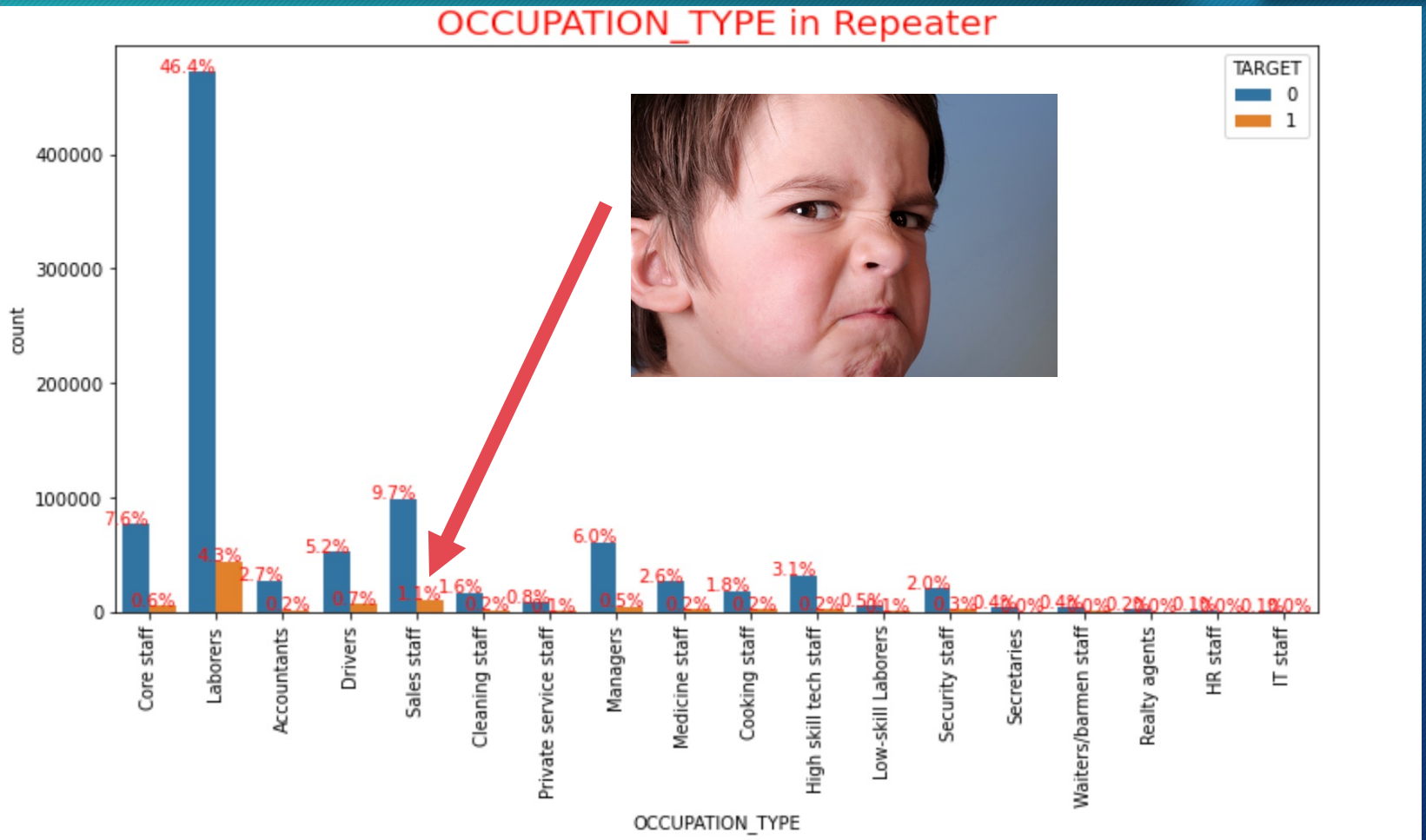


We can see total cash loans are -92.9% and Revolving loans are only 7.1%. But comparing these two we can see that number of defaulters are very less in the Revolving loans

OCCUPATION TYPE IN REPEATER DATASET EXTRACTED FROM COMBINED

Total numbers in occupation type Laboure's(50.7%),Sales Staff(10.8%),core staff(8.2%),Managers(6.5%) ,drivers(5.7%) and Accountants(2.9%).

Out of these Sales Staff have highest number of defaulter.

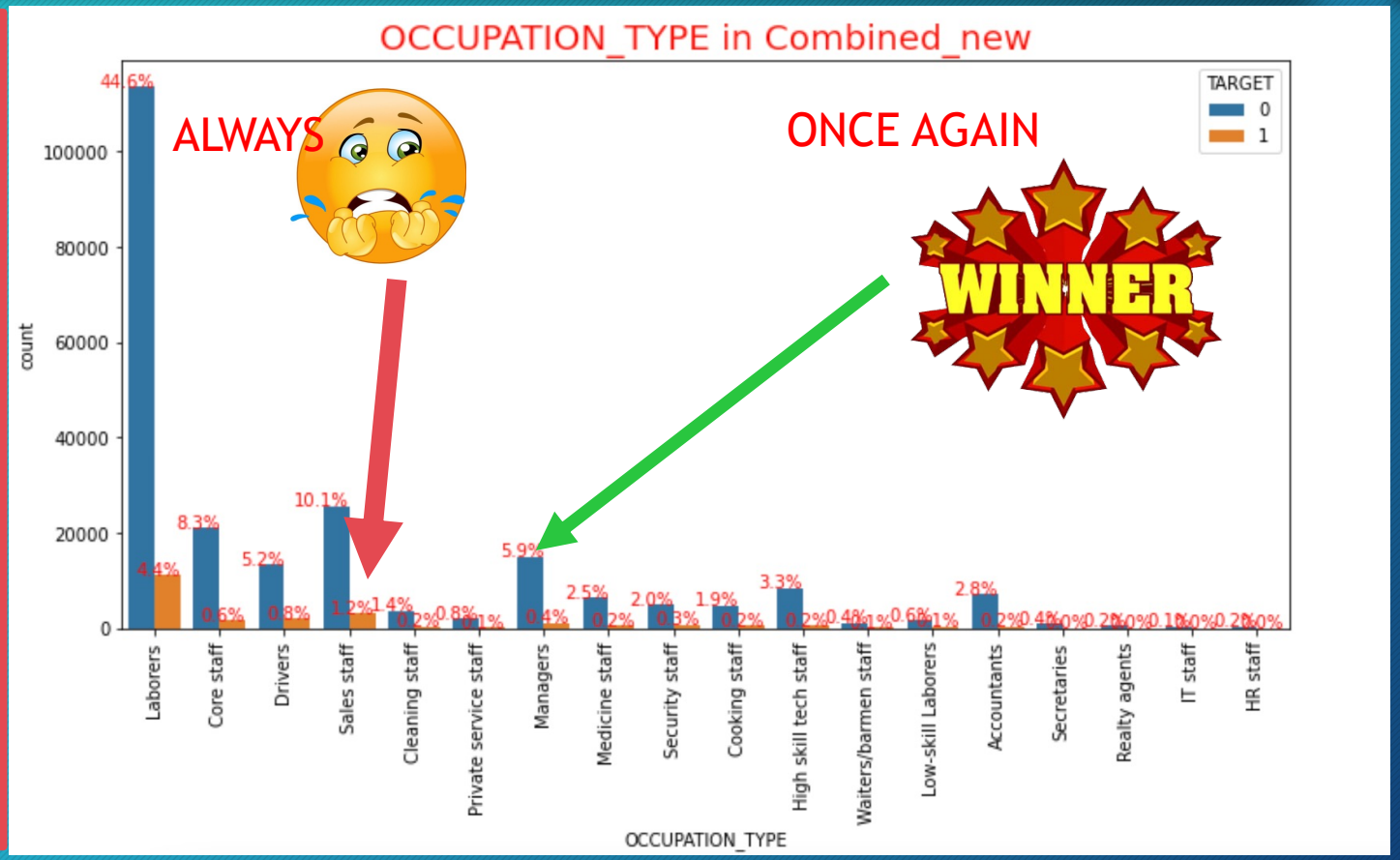


OCCUPATION_TYPE IN NEW EXTRACTED FRON COMBINED

we can observe Total Laborers 49%,Corestaff 8.9% ,sales staff 11.3 and managers 6.3%

also NON-Defaulter laborers 91%, core staff 93.2%, Sales staff- 89.3% and managers 93.6%.

So again managers are better and sale staff is poor in paying loans




THANK
YOU


to be continued