CREDIT EDA CASE STUDY

SUBMITTED BY:

KARTHIK K S

THE PROBLEM STATEMENT

COMPANY	CONTEXT	PROBLEM STATEMENT
Consumer Finance company specialises in lending various type of loans to urban customers.	Company wants to understand the driving factors or variables behind loan default, i.e. variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.	Working for Consumer Finance company analyse the dataset containing information about loan applicants using EDA to understand how consumer attributes and loan attributes influence the tendency of default.

DATA EXPLORATION and ANALYSIS APPROACH

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.**

Target Variable /Dependent Variable (DV)
The target outcome is 1, in the target variable 'TARGET', in the application_train.csv file.

Description:
1 - client with payment difficulties
0 - all other cases

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.**

Columns_description.csv :
is data dictionary which describes the meaning of the variables.

Data Sourcing → Data Cleaning → Univariate Analysis → Segmented Univariate Analysis → Bivariate Analysis → Summarize results

DATA CLEANING AND MANIPULATION

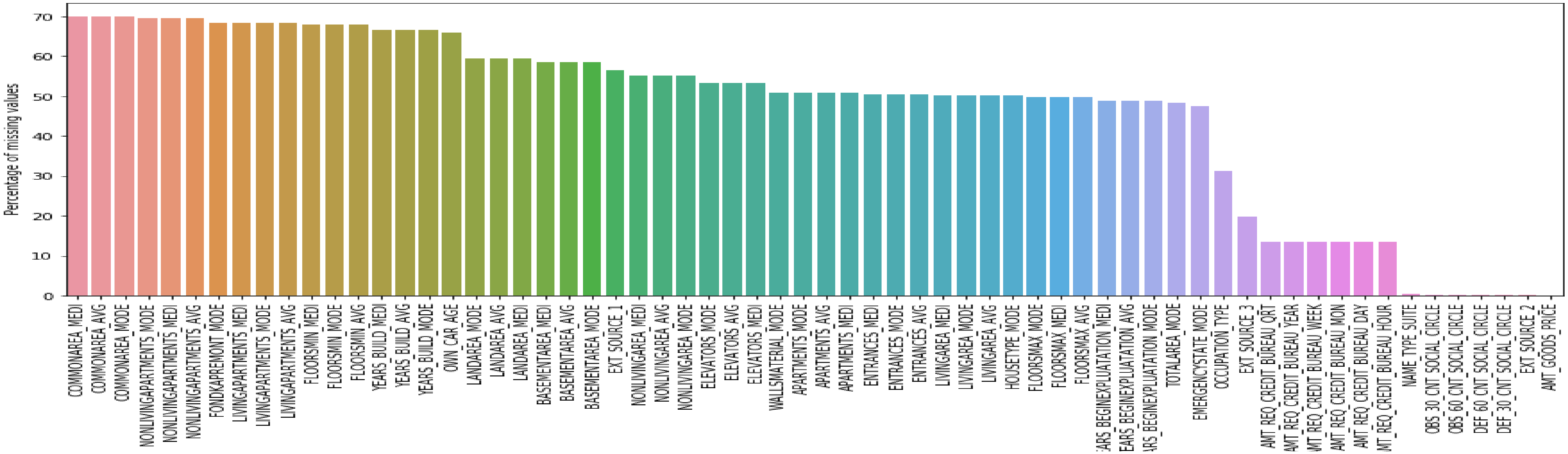
Data Inconsistencies:

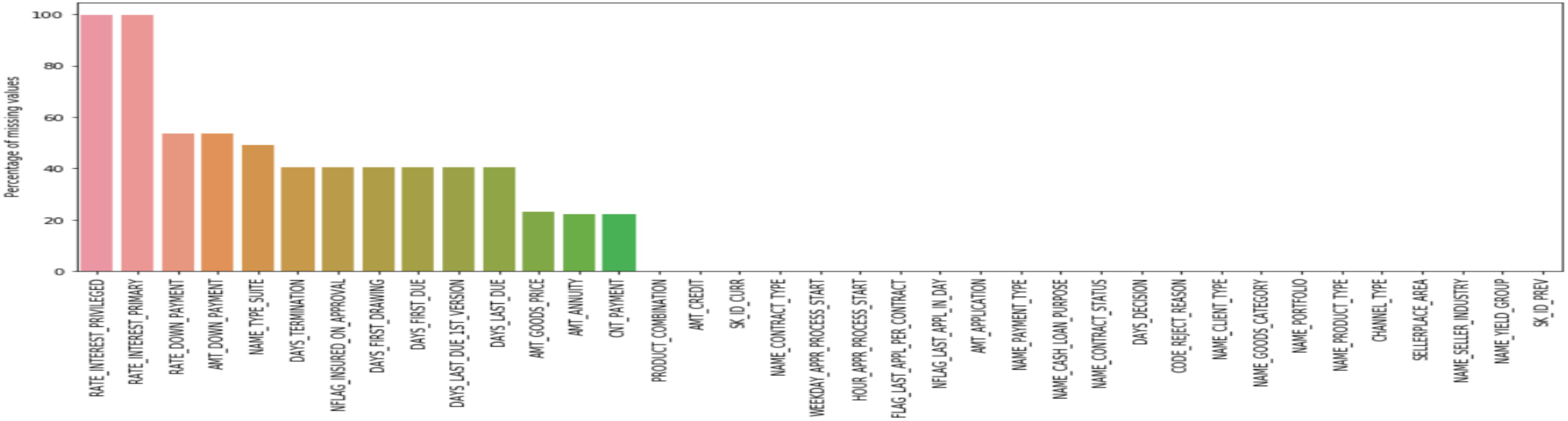
- NA values in the columns (to remove columns having more than 45% null values)
- XNA - CODE_GENDER- Application Data set

Other Issues:

- Few columns (DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH) have negative values
- Data need to be appropriately transformed, if necessary, for easy analysis and plotting the data
- Outliers present in the data set need to be removed

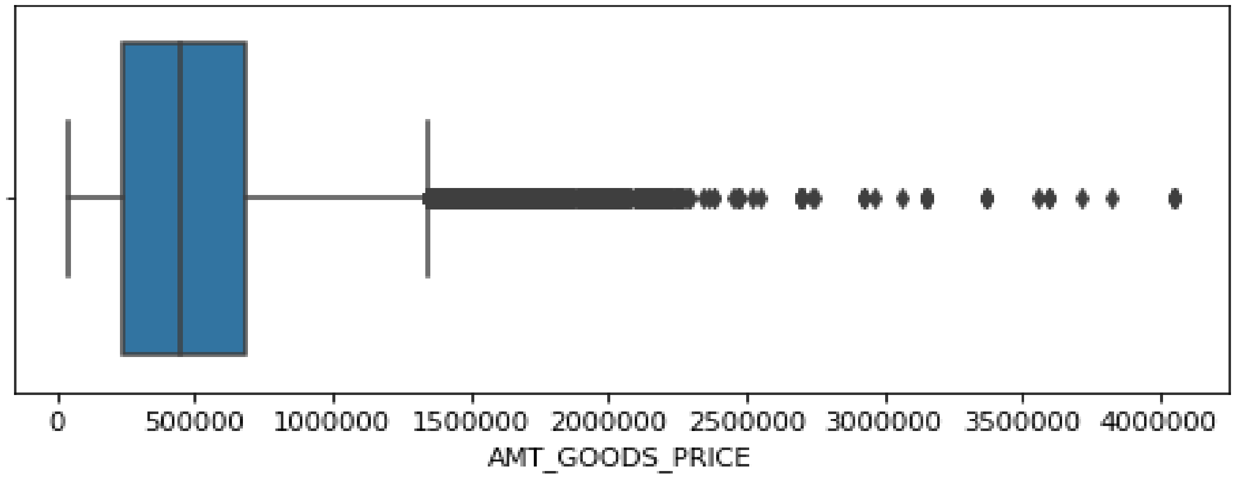
Percentage of missing values for all the columns:



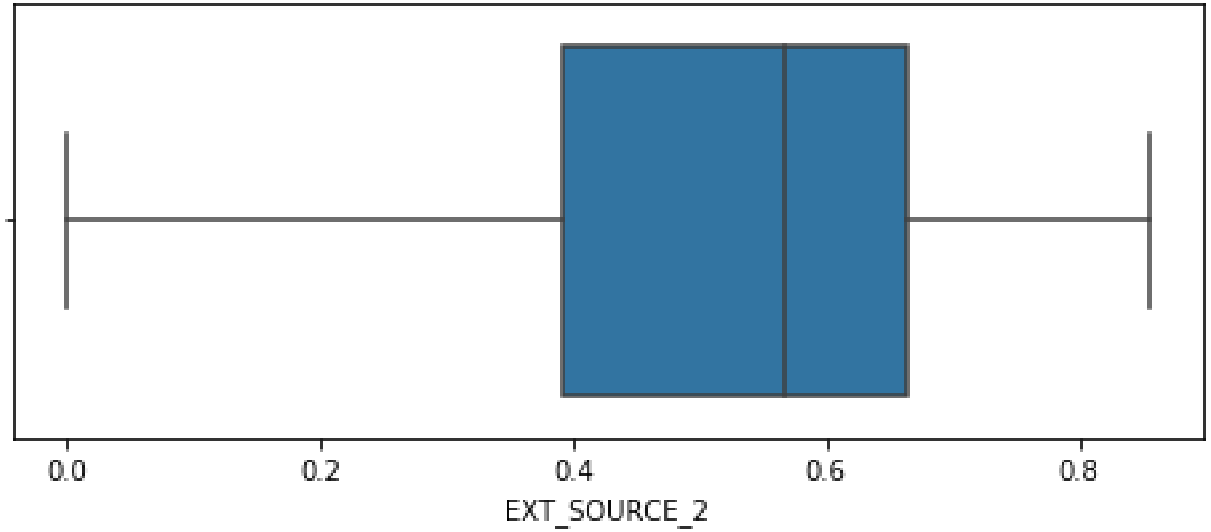


Approach for Missing Value Treatment

1) AMT_GOODS_PRICE:
Since, there are outliers present in the AMT_GOODS_PRICE column, we are imputing it with Median value

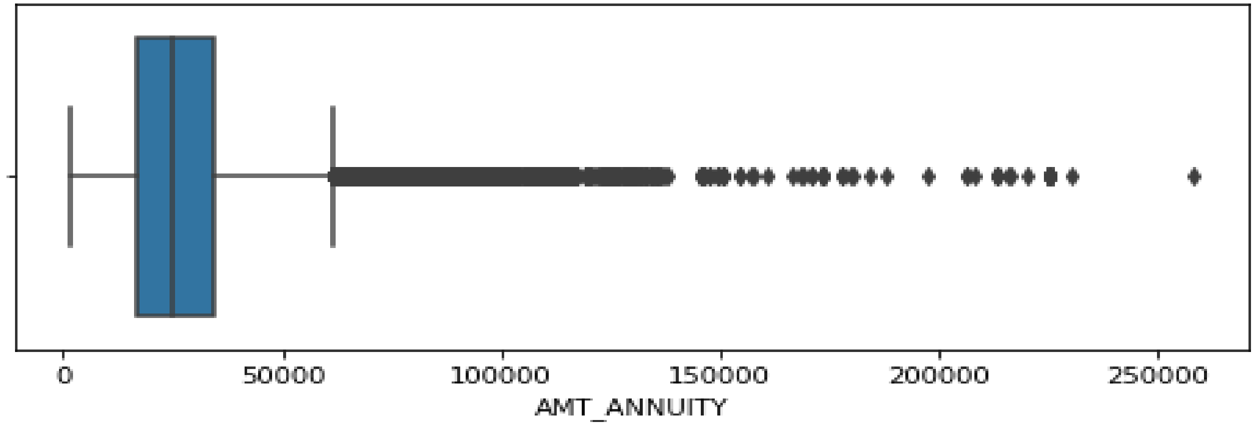


2)EXT_SOURCE_2:
Imputing the missing data with "Mean" as standard deviation is less and there are no outliers present in the data



.

3) AMT_ANNUITY:
Imputing the missing value with "Median" value,
as there are outliers and standard
deviation is high.

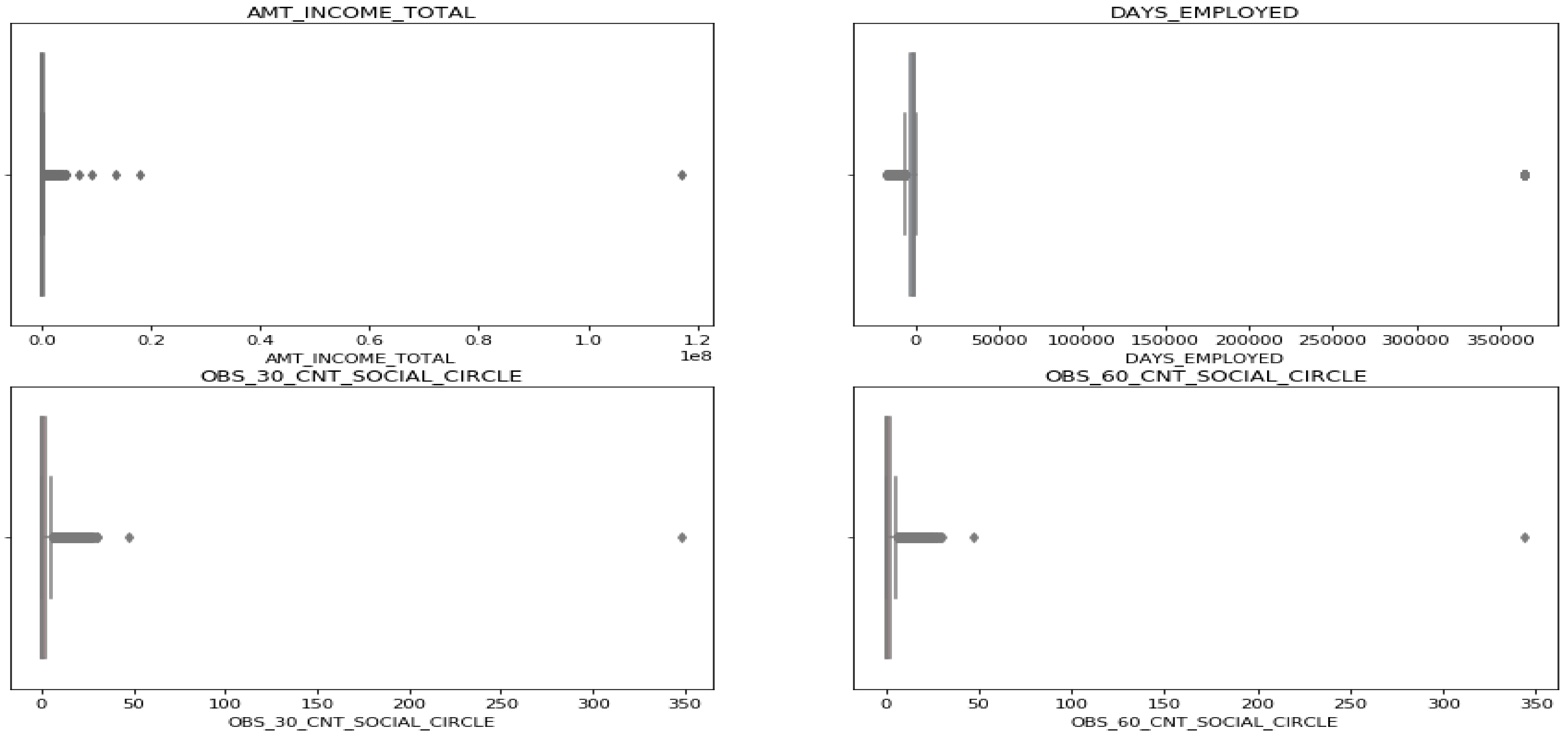


4)SOCIAL_CIRCLE variables:Since, all the columns has a greater number of count 0, we are imputing it with the mode and since it is Zero, it may not affect the analysis.

5)CODE_GENDER: There are 4 null values in this column. Since male percentage is 0.34 and female percentage is 0.66 and it is in the ratio 1:3, so randomly imputing 1 male and 3 female to the four missing values.

Outlier Detection

There are points that lies outside the 95th ,which could be identified fom the below plots, hence IQR method is used to identify the upper and lower quartile range and outliers are removed from the dataset as a part of data cleaning activity



Outlier Treatment

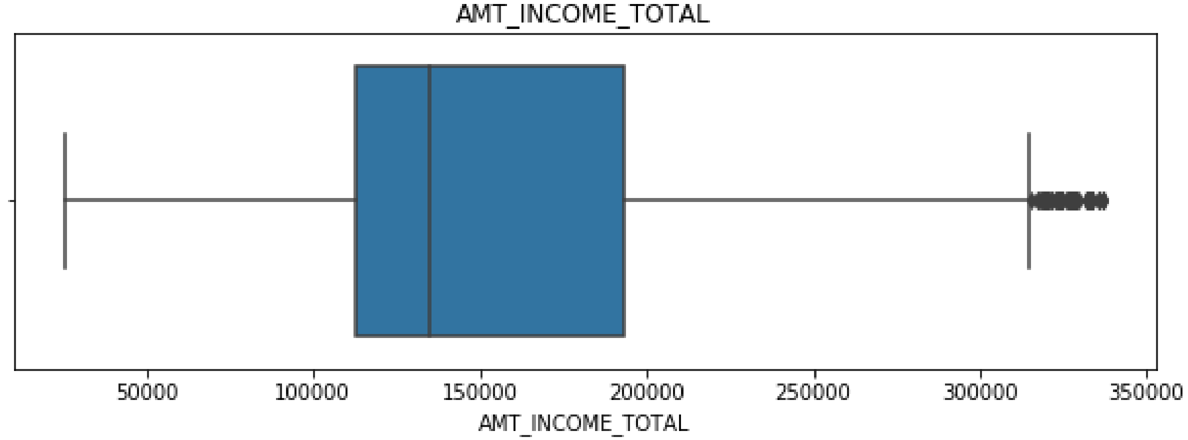
AMT_INCOME_TOTAL:

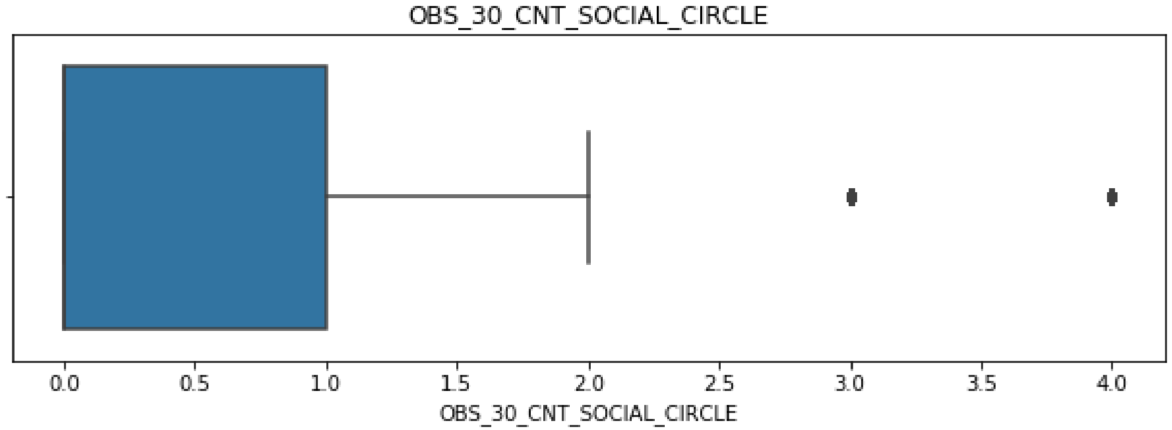
Total number of outliers 15825
Total number of non-outliers 291686
Lower quartile -22500.0
Upper quartile 337500.0

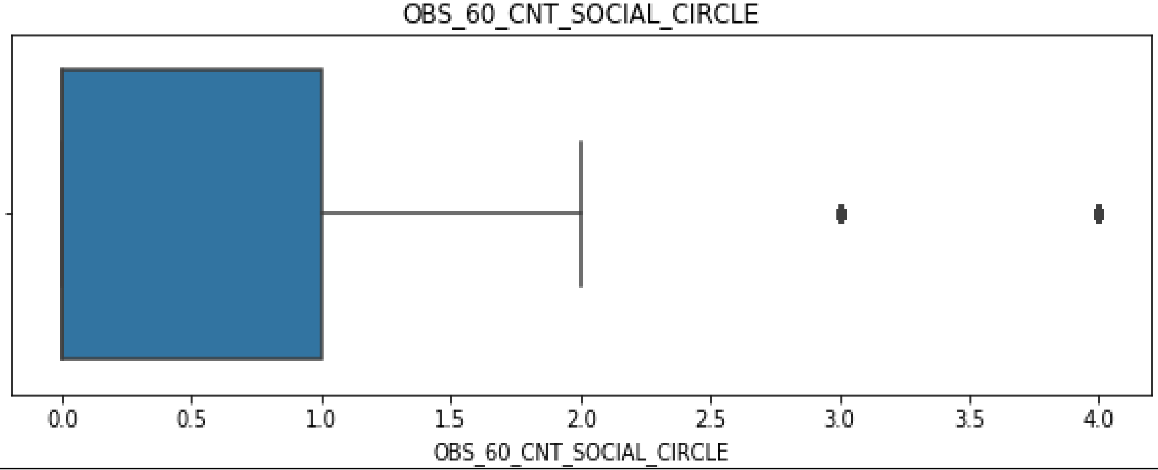
OBS_30_CNT_SOCIAL_CIRCLE:

Total number of outliers 29524
Total number of non-outliers 277987
Lower quartile -3.0
Upper quartile 5.0

OBS_60_CNT_SOCIAL_CIRCLE:
Total number of outliers 29027
Total number of non-outliers 278484
Lower quartile -3.0
Upper quartile 5.0

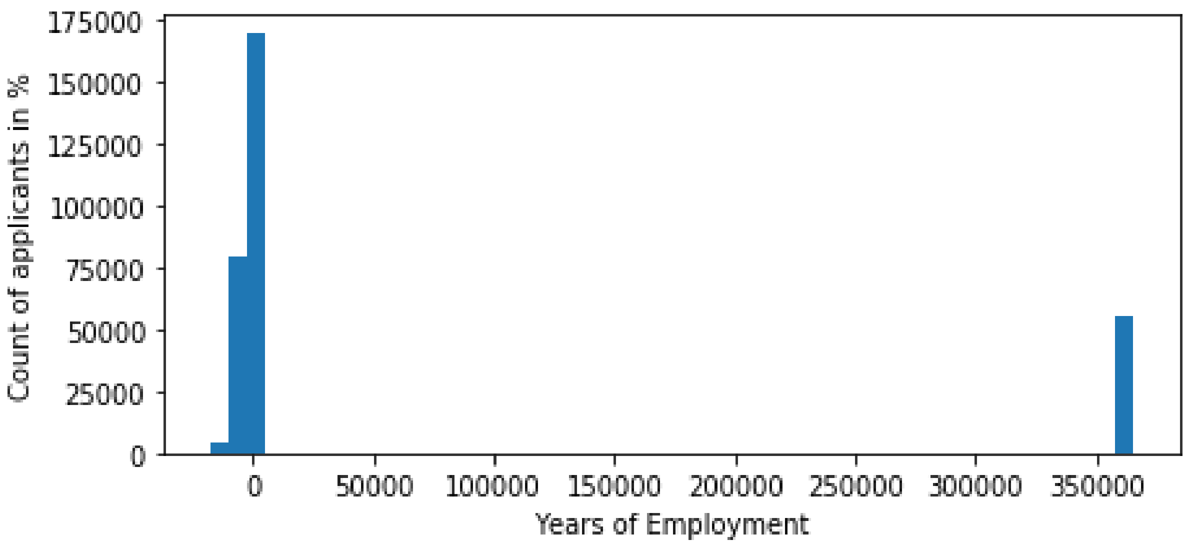




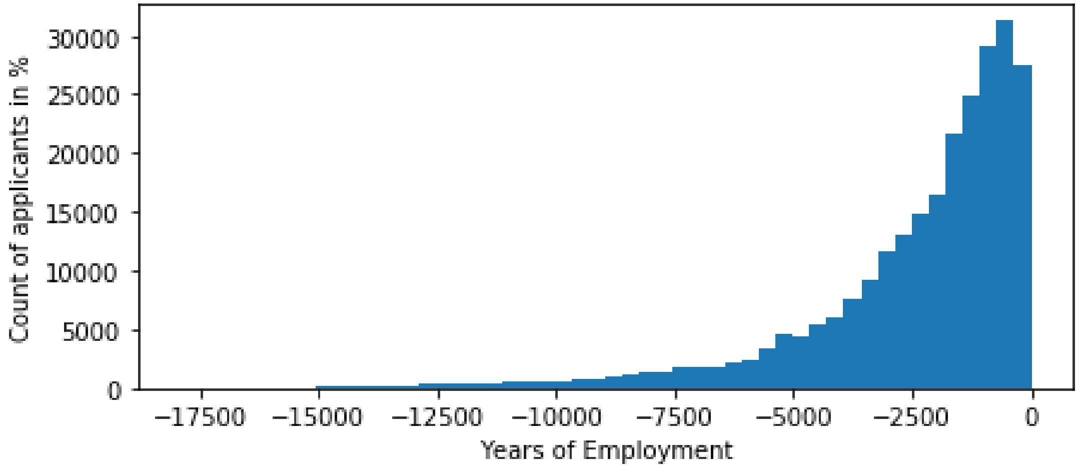


Years of Employment:

- The data looks strange as we have -365243 days of employment which is impossible



- looks like there is data entry error which is clearly an outlier and we are replacing it as null values



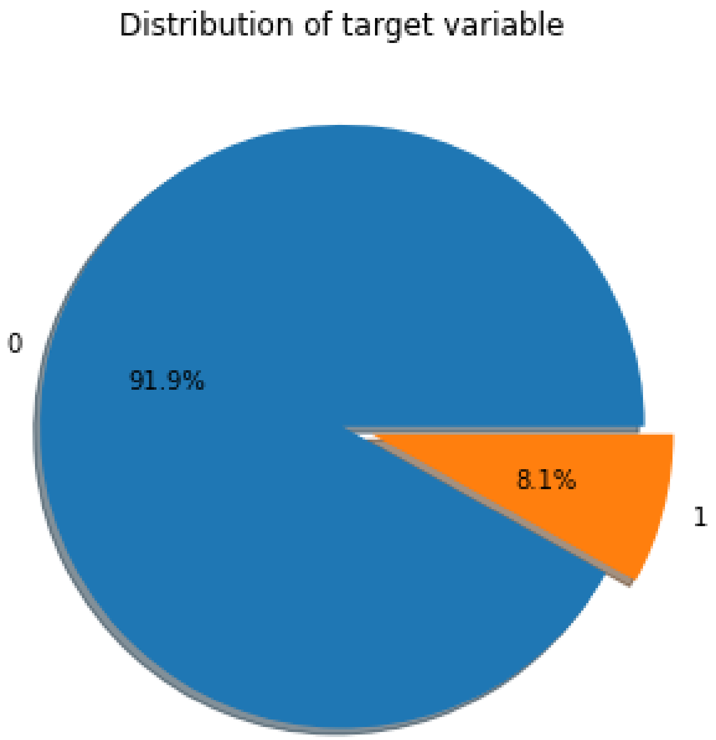
UNIVARIATE ANALYSIS ON CONTINUOUS VARIABLES

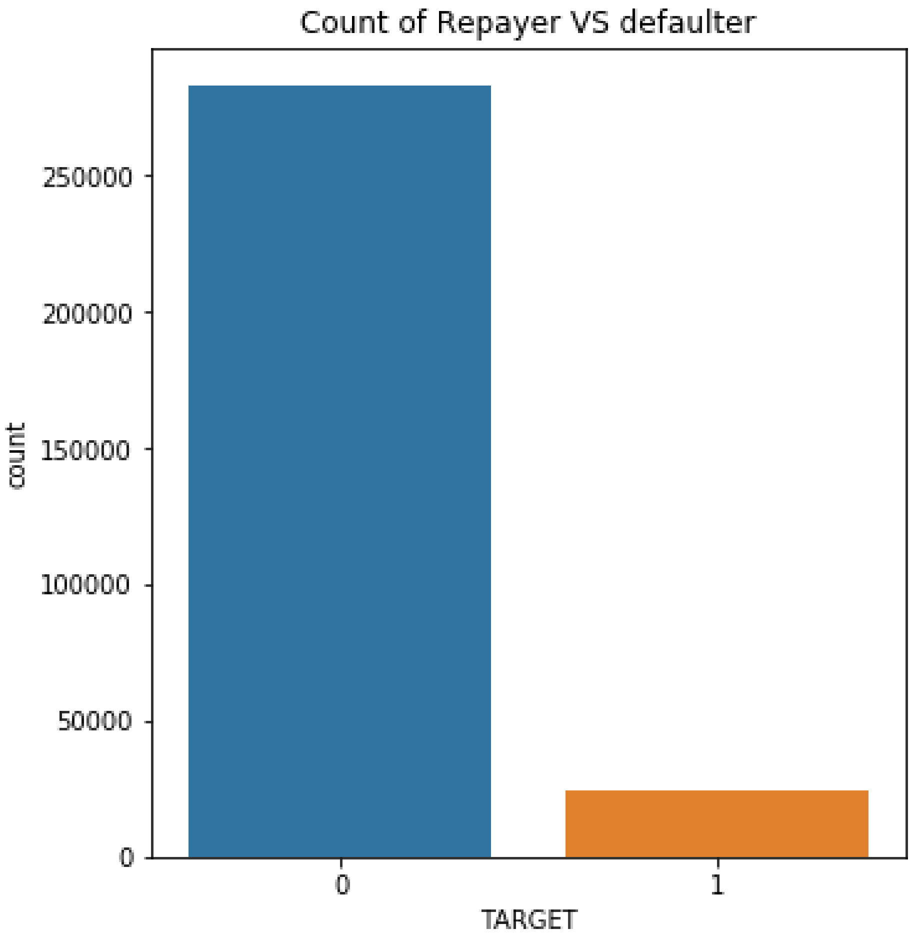
DATA IMBALANCE

RATIO OF DATA IMBALANCE = 8.1 %

Total percentage of clients who paid loans = 91.9

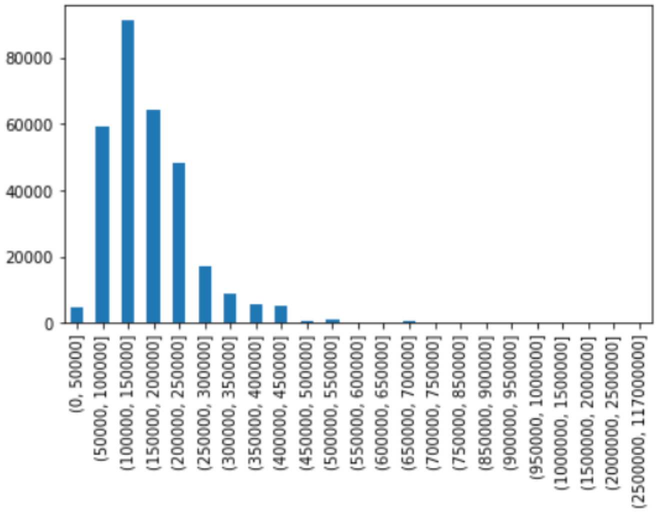
Total percentage of clients who has payment difficulties =8.1

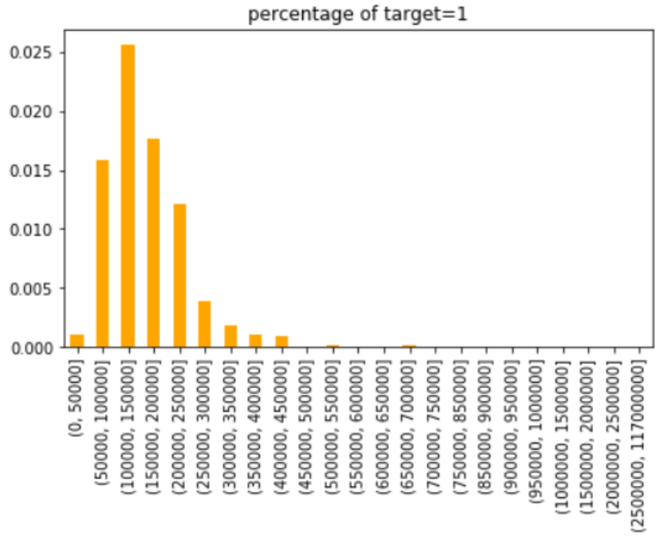




BINNING OF CONTINUOUS VARIABLES

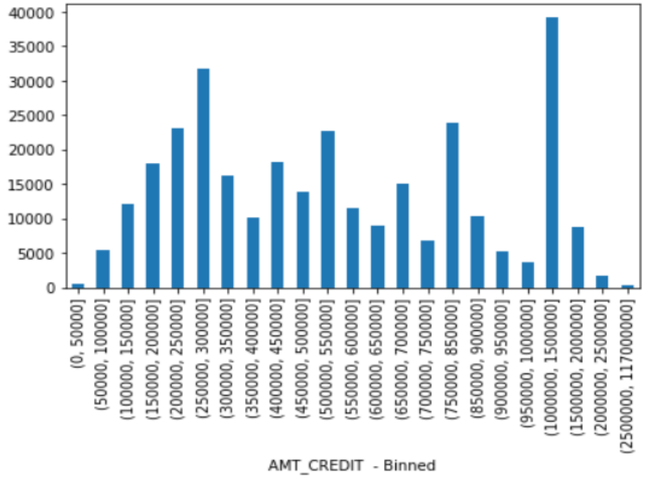
Analysis on Loan Variables:

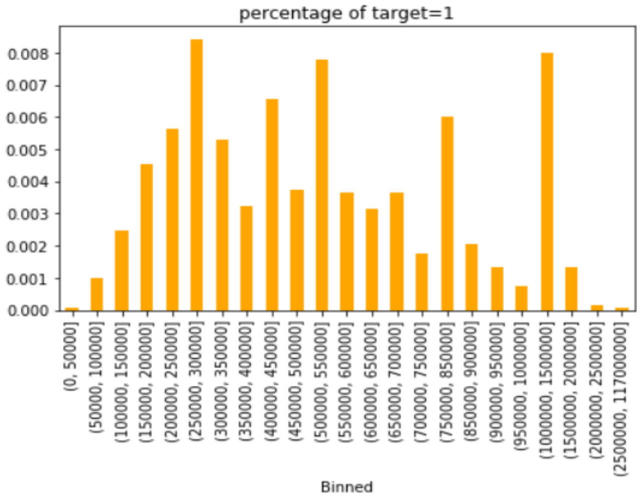




AMT_INCOME_TOTAL :

- If we see the above given distrbution plot it clear say major distrbuiton for amount of anual income is from 0 to 4,50,000 and people with total annual income lying between 100000 and 150000 avail more loans.
- **higher chance of default**: people with total income between **100000 and 150000**

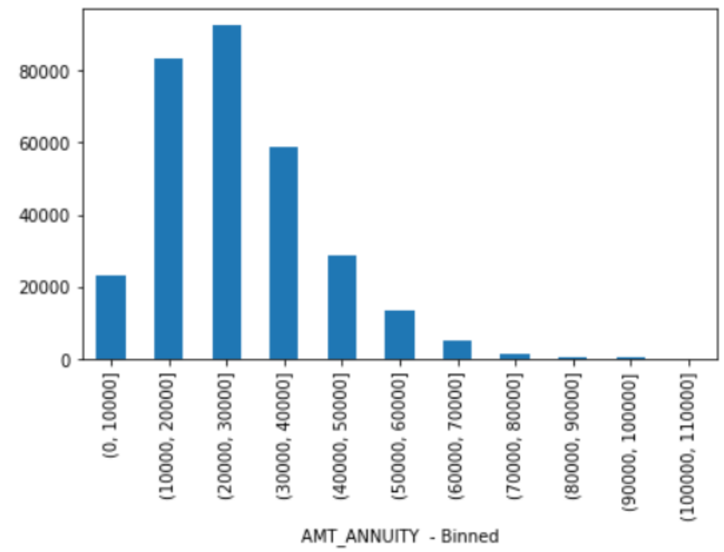


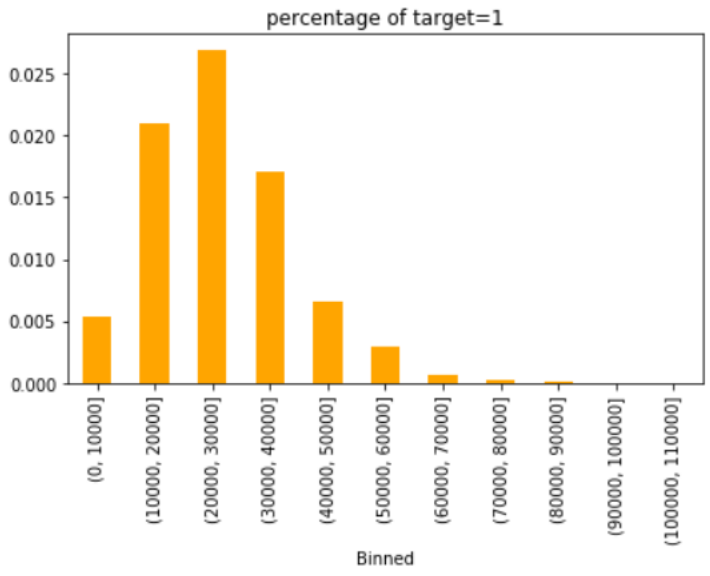


.**AMT_CREDIT**:

- Similarly, people has more credit on range 10,00,000 to 15,00,000 for applying loan

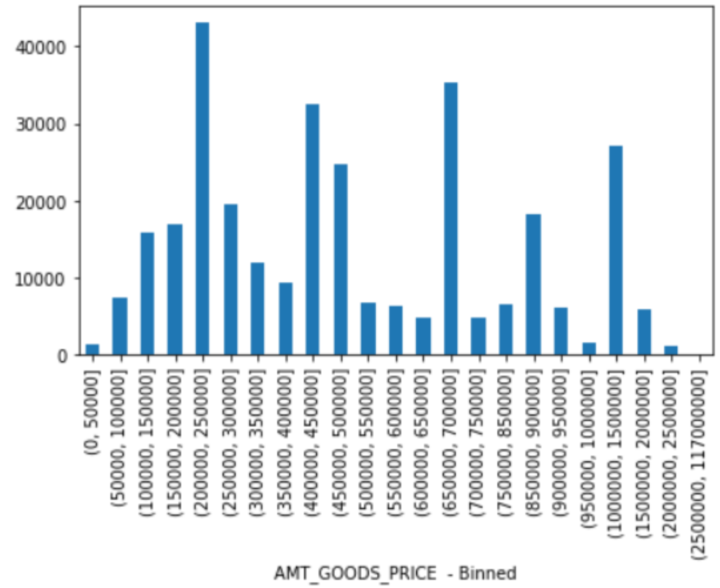
- **higher chance of default** : people with more credit on range **2,50,000 to 3,00,000**

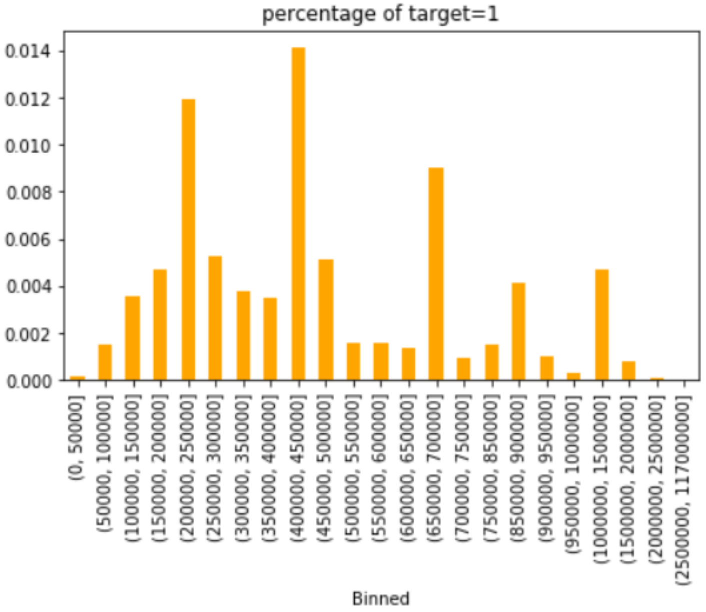




AMT_ANNUITY:

- People have more annuity on range 20,000 to 30,000
- **higher chance of default** : people with more annuity range **20,000 to 30,000**

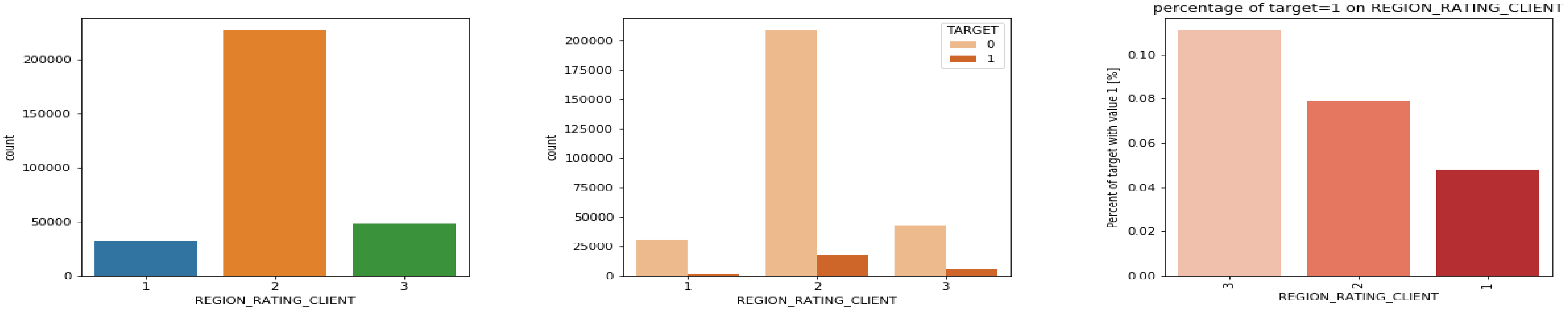




AMT_GOODS_PRICE :

- Maximum loans were given for goods with price range 2,00,000 to 2,50,000
- **higher chance of default** :people with good price range between **4,00,000 to 5,00,000**

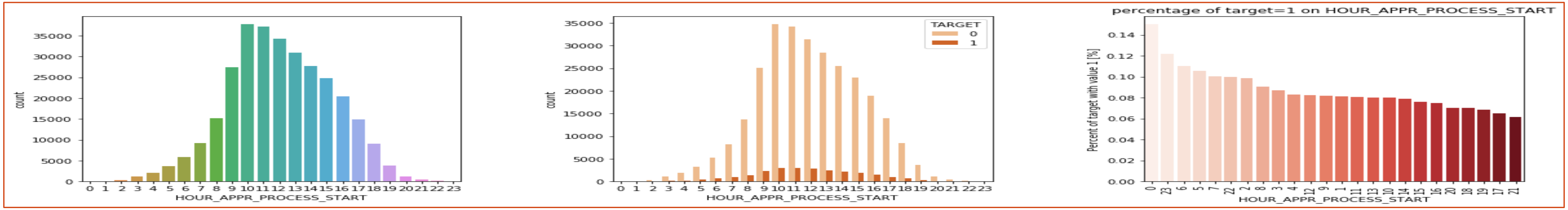
Uni-Variate analysis on Category variable with respect to Target

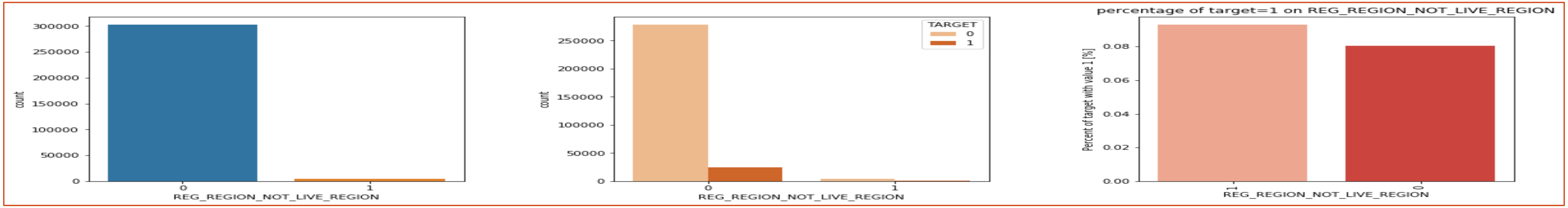


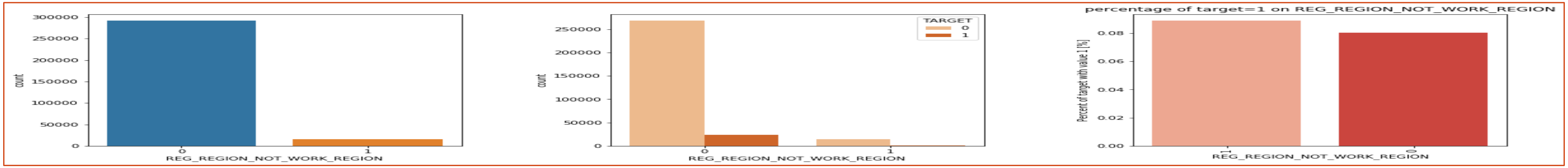
From above plot, it can be seen that

- Clients with Region_rating of 3 are the top loan defaulters than the client with Region_rating of 2 and 1.

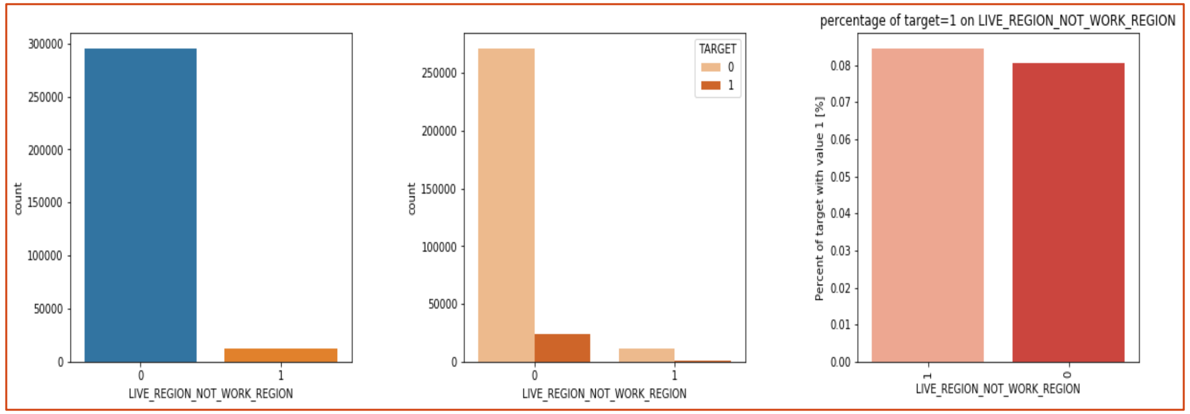
- Clients with Region_rating of 3 after taking city into account are again the top loan defaulters than the client with Region_rating of 2 and 1.

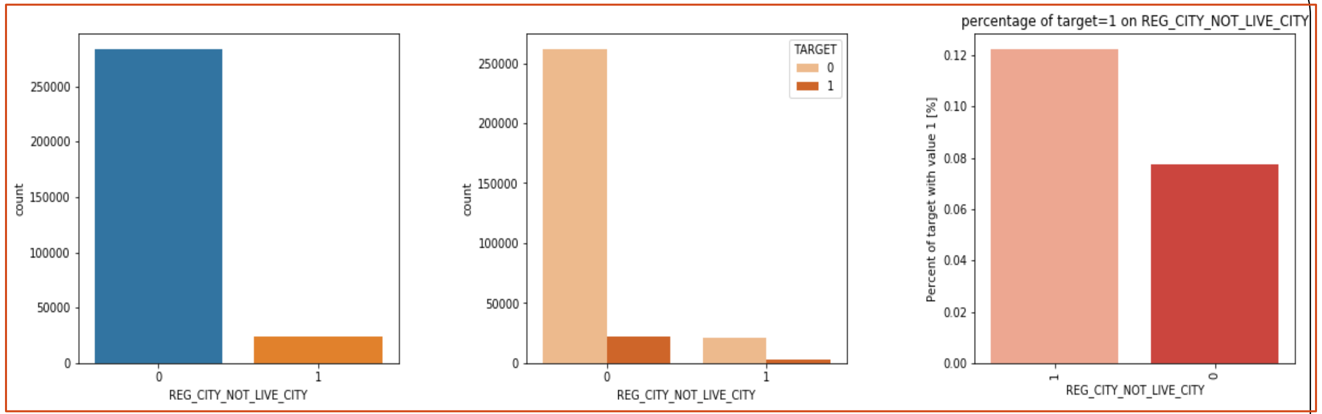


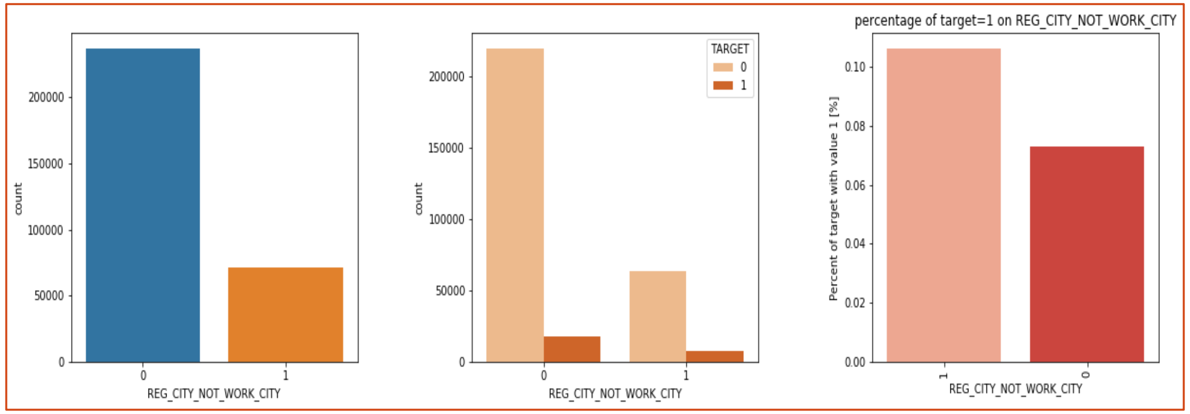


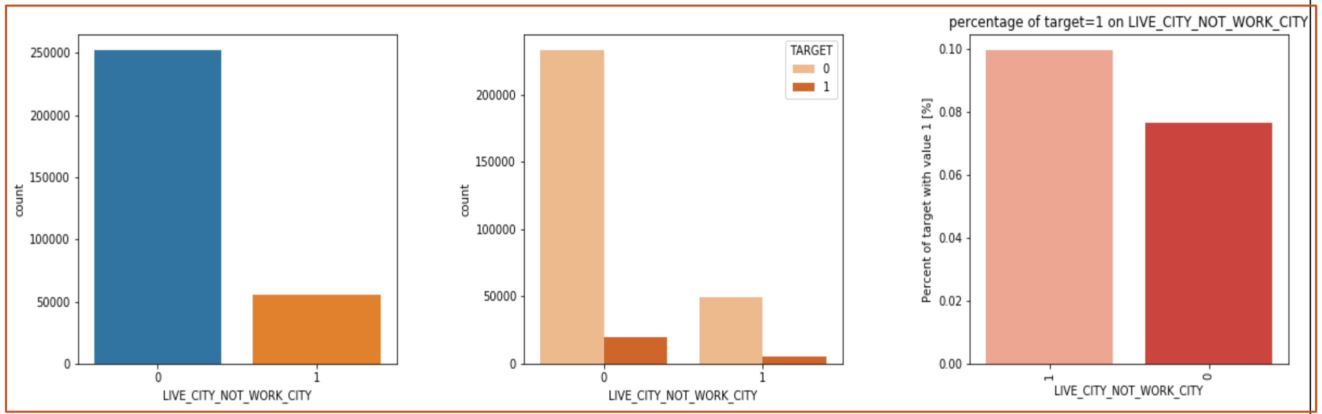


- More number of clients applying loan during 10 to 12am have less chance of loan defaulting than the people who apply late in the evening.
- Clients with different permanent and contact address at region level, have high chances of being loan defaulters than the clients with same address.
- Clients with different permanent and work address at region level, have high chances of being loan defaulters than the clients with same address.

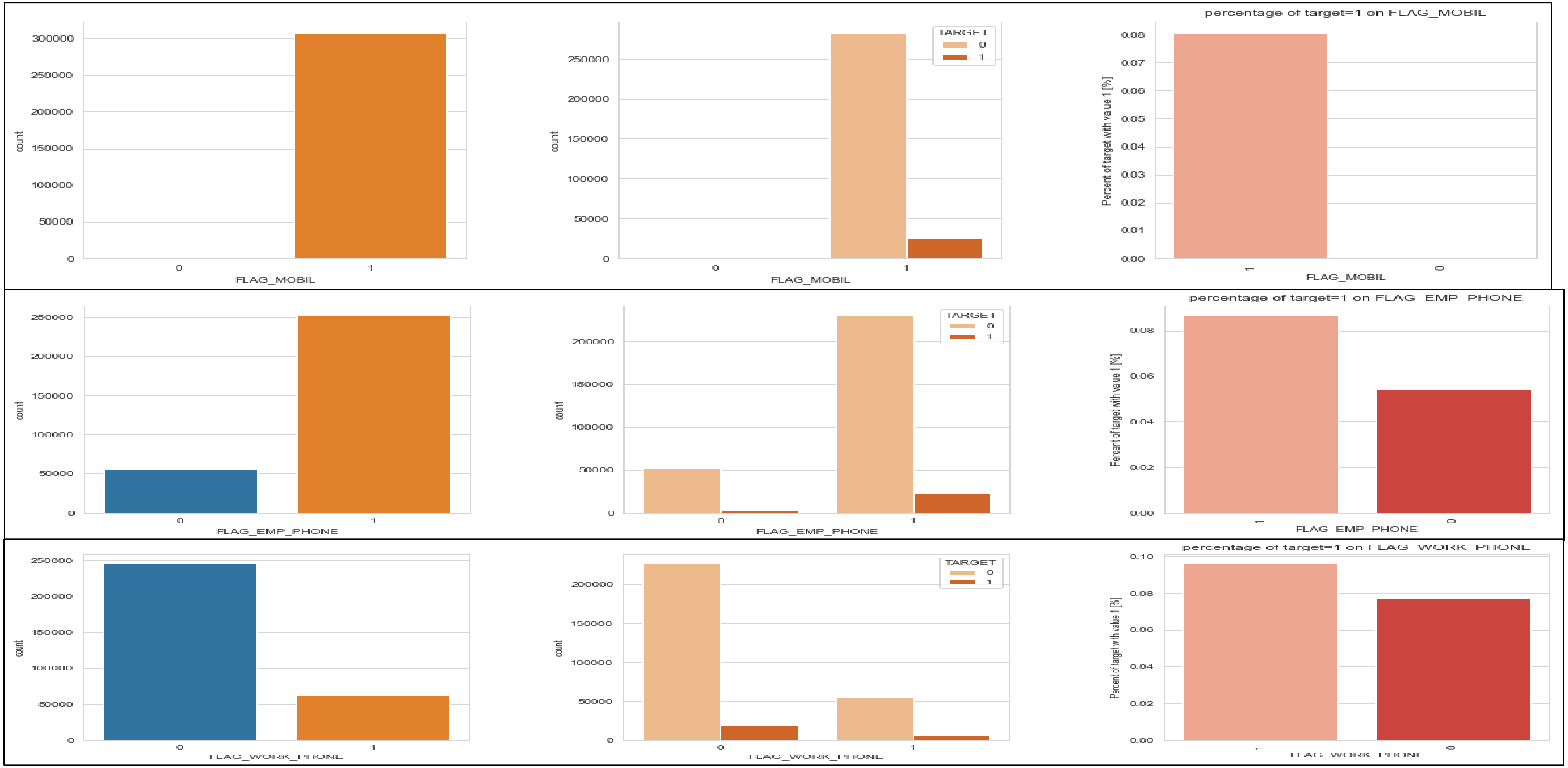




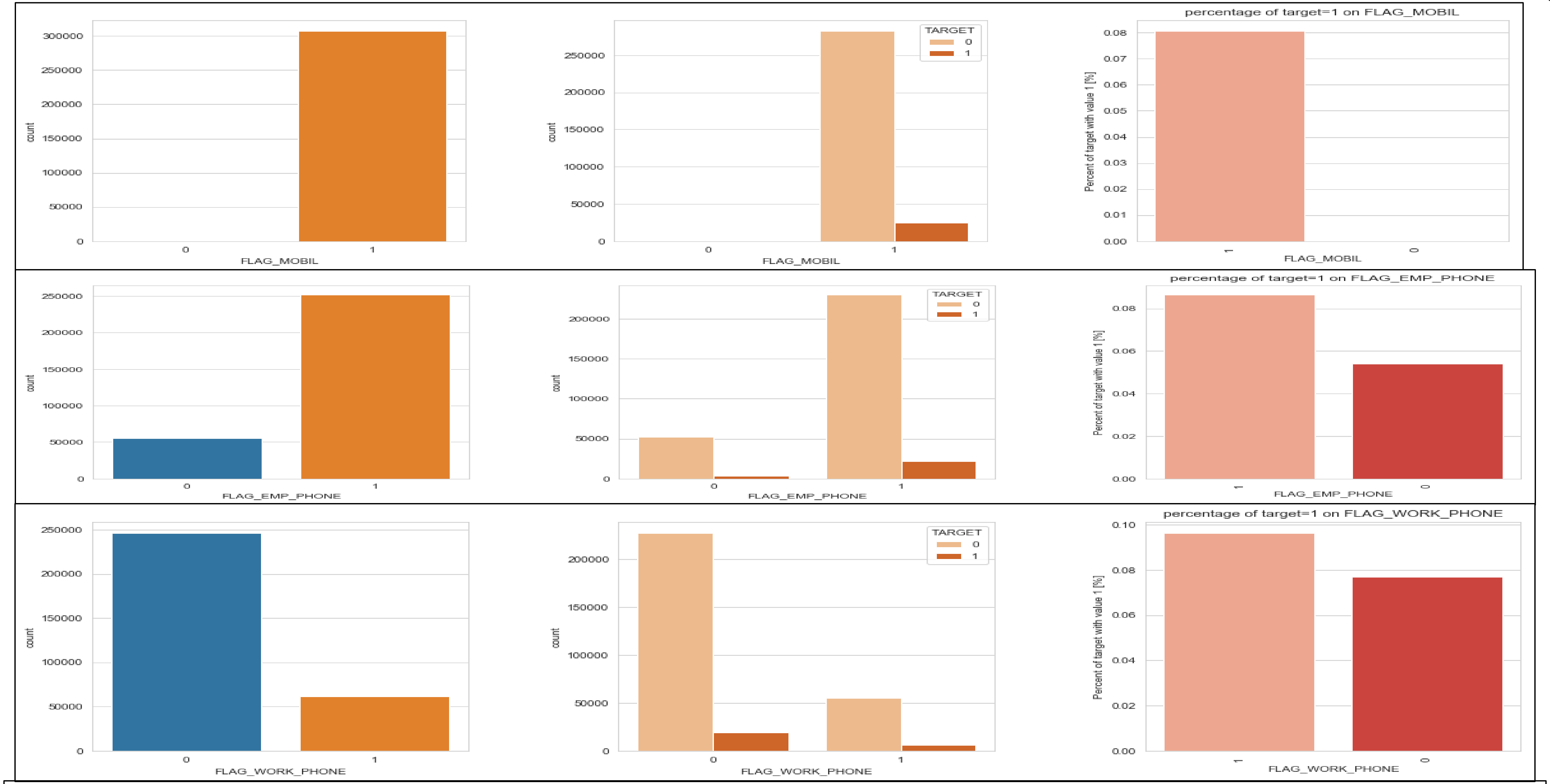




- From above plots it can be seen that the clients with different address(contact, work, permanent) at both region & city level, have high chances of being loan defaulters than the clients with same address.



From above graphs, it can be concluded that the clients who provided all their numbers(mobile, home and work) are the highest loan defaulters then the clients who did not provide any of their numbers



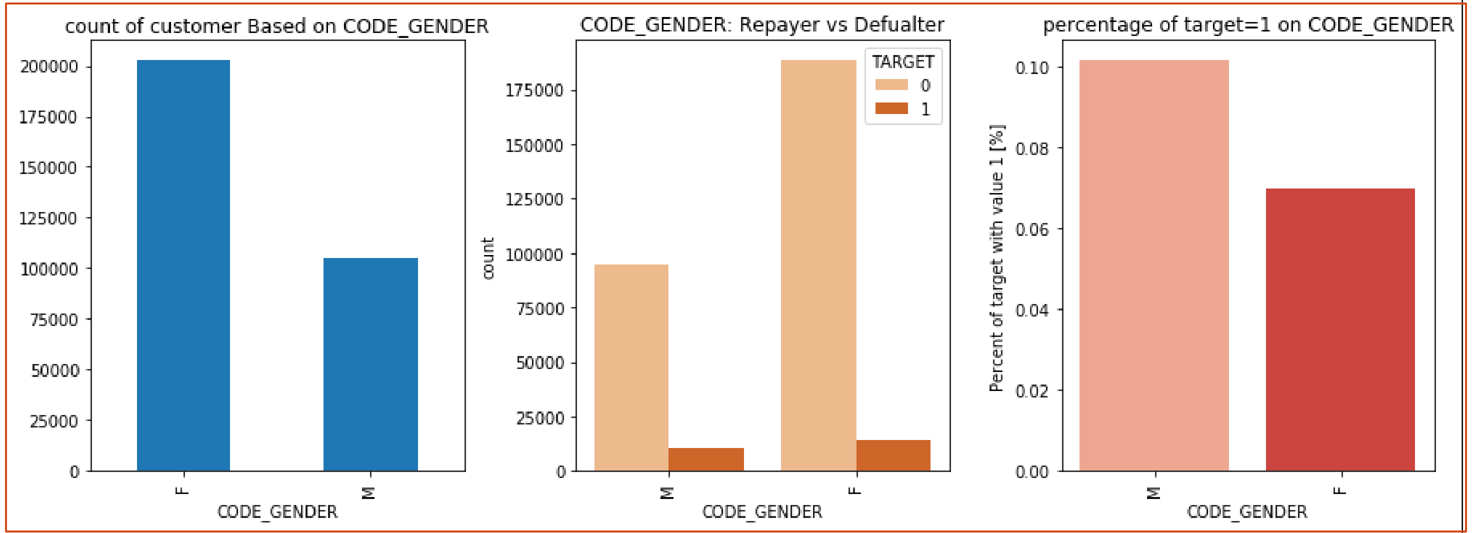
From above graphs, it can be concluded that the clients who provided all their numbers(mobile, home and work) are the highest loan defaulters then the clients who did not provide any of their numbers

UNIVARIATE ANALYSIS ON CATEGORICAL VARIABLES

1)CODE GENDER

	CODE_GENDER	TARGET
1	M	0.101418
0	F	0.069992

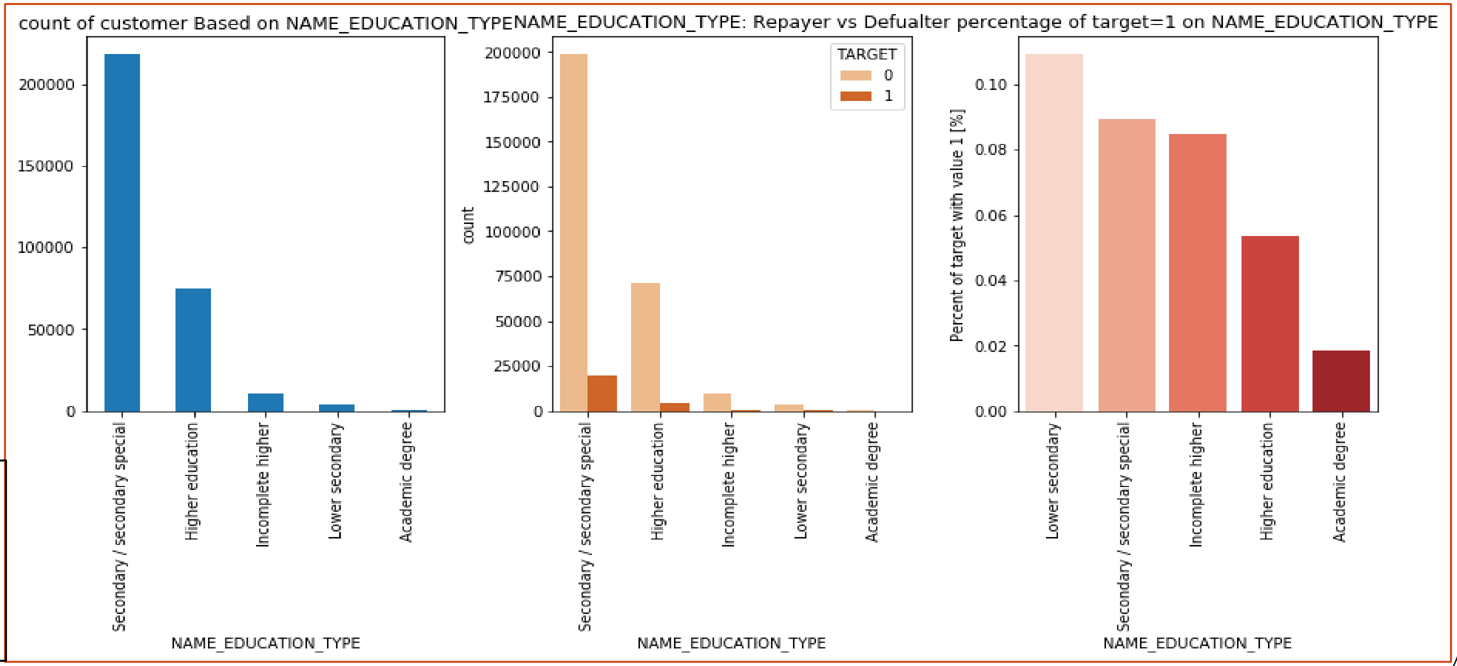
Clearly seen from the graph, Males being less applicant of loans than females, but has a high chance of being a loan defaulter



2)NAME EDUCATION TYPE

3	Lower secondary	0.109277
4	Secondary / secondary special	0.089399
2	Incomplete higher	0.084850
1	Higher education	0.053551
0	Academic degree	0.018293

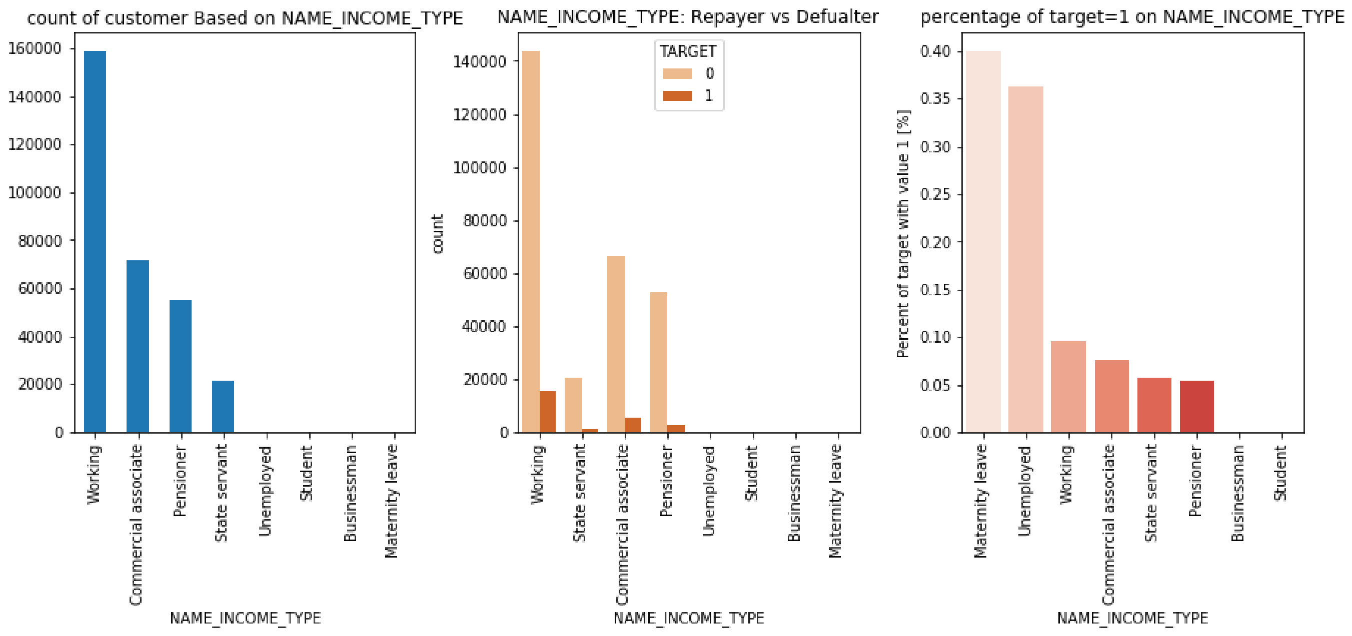
From the given plot it is clearly seen that the customer with "lower secondary education" apply for a smaller number of loans. But they cannot not repay the loan, making them have a high chance of loan defaulter in the list.



3)NAME INCOME TYPE

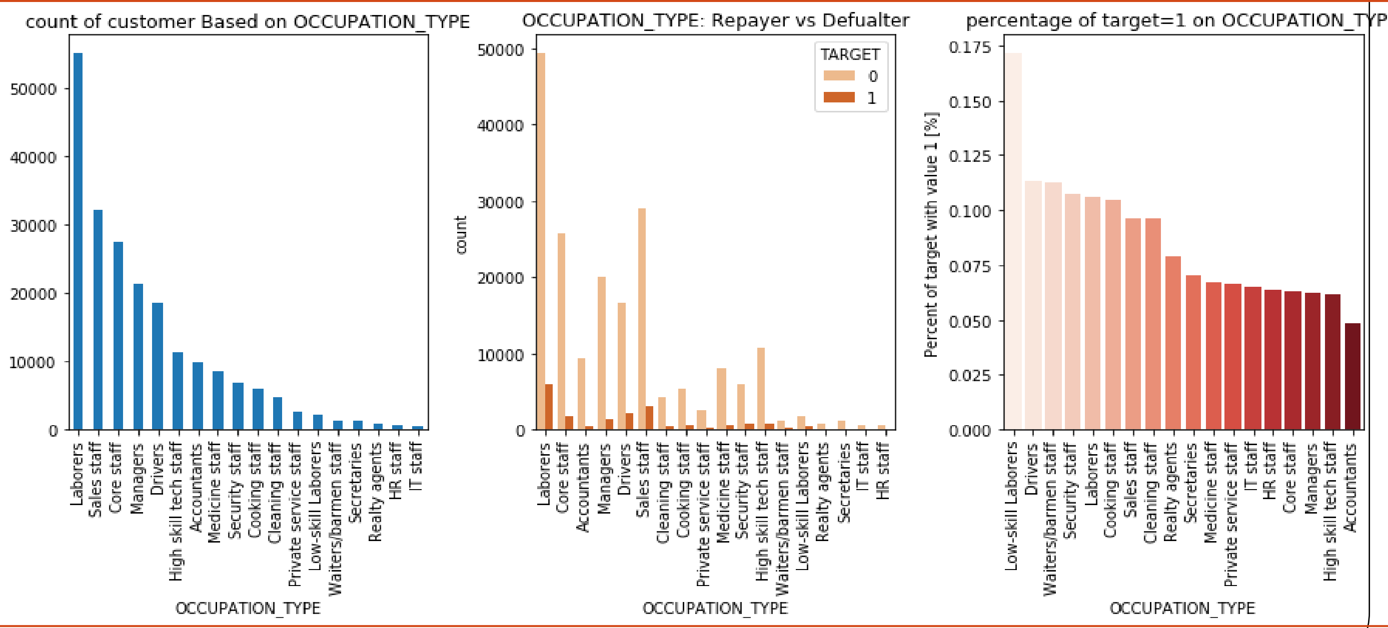
	NAME_INCOME_TYPE	TARGET
2	Maternity leave	0.400000
6	Unemployed	0.363636
7	Working	0.095885
1	Commercial associate	0.074843
4	State servant	0.057550
3	Pensioner	0.053864
0	Businessman	0.000000
5	Student	0.000000

From above graph it is seen that working professionals apply for a greater number of loans but they have less chances of being loan defaulters whereas customers on maternity leave are topmost in loan defaulter list among name_income_type category



4)OCCUPATION TYPE

	OCCUPATION_TYPE	TARGET
9	Low-skill Laborers	0.171524
4	Drivers	0.113261
17	Waiters/barmen staff	0.112760
16	Security staff	0.107424
8	Laborers	0.105788
2	Cooking staff	0.104440
14	Sales staff	0.096318
1	Cleaning staff	0.096067

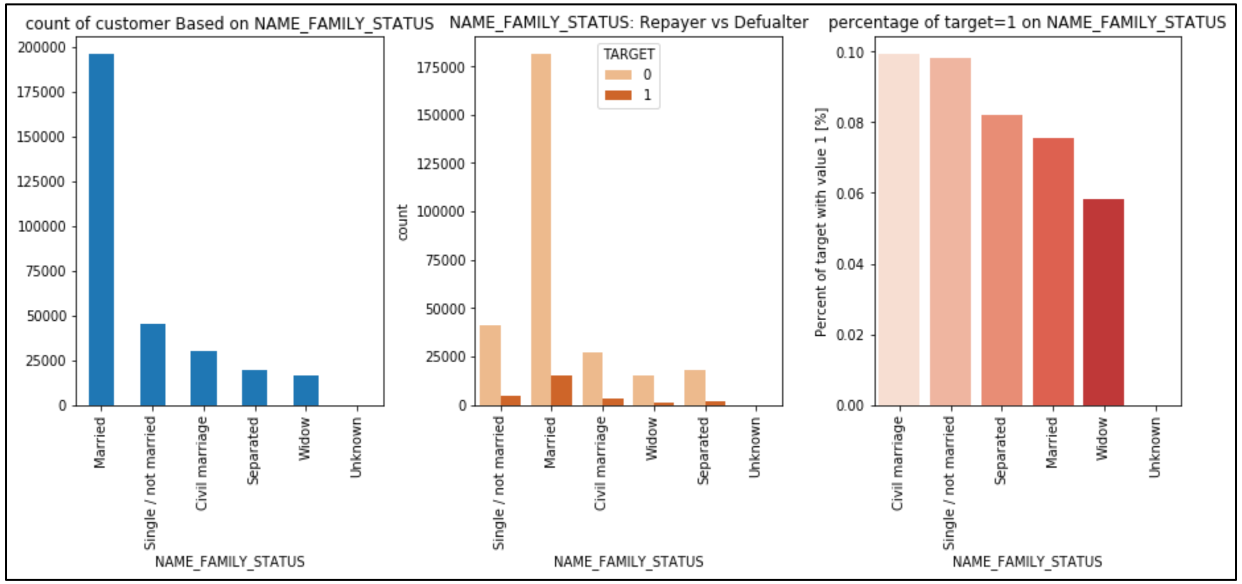


It can be seen from above graph that Low skill laborers apply for a smaller number of loans but are in the highest category of loan defaulter followed by drivers, waiters and barmen staff.

5)NAME FAMILY STATUS

	NAME_FAMILY_STATUS	TARGET
0	Civil marriage	0.099446
3	Single / not married	0.098077
2	Separated	0.081942
1	Married	0.075599
5	Widow	0.058242
4	Unknown	0.000000

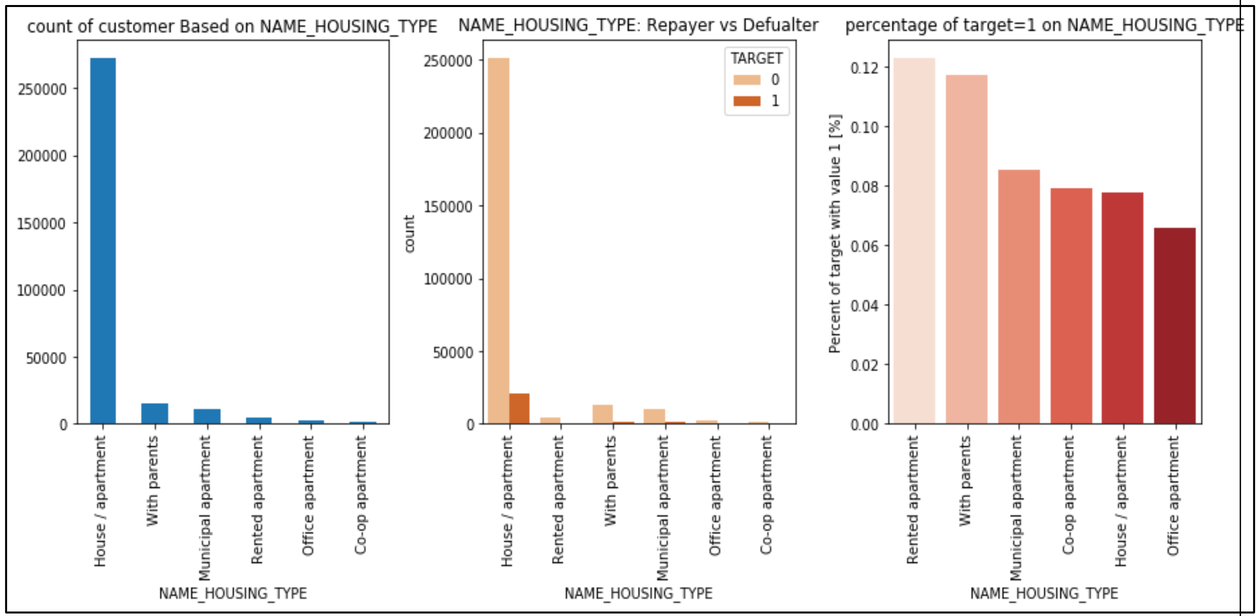
It can be clearly seen from above plot that civil marriage category customers has higher chances of being loan defaulter whereas widow category being the least defaulter



6)NAME HOUSING TYPE

	NAME_HOUSING_TYPE	TARGET
4	Rented apartment	0.123131
5	With parents	0.116981
2	Municipal apartment	0.085397
0	Co-op apartment	0.079323
1	House / apartment	0.077957
3	Office apartment	0.065724

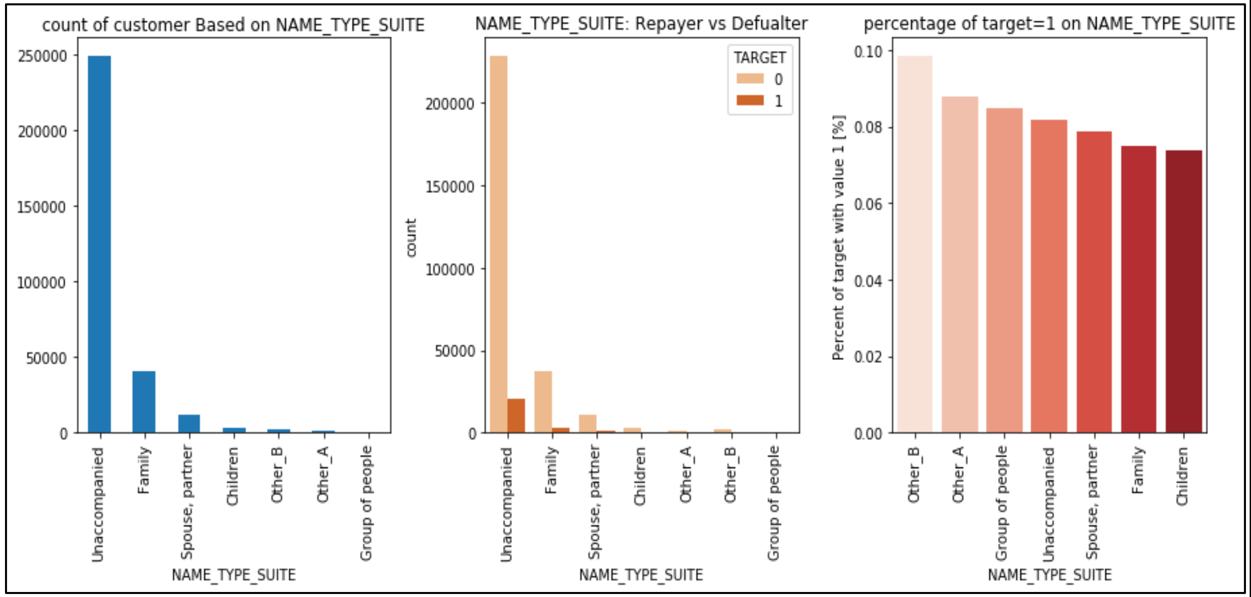
From graph, we can see that people staying in House/apartments apply for a greater number of loans but people staying in rented apartments have the high chances of being loan defaulter among Name_housing_type category



7)NAMETYPE SUITE

	NAME_TYPE_SUITE	TARGET
4	Other_B	0.098305
3	Other_A	0.087760
2	Group of people	0.084871
6	Unaccompanied	0.081830
5	Spouse, partner	0.078716
1	Family	0.074946
0	Children	0.073768

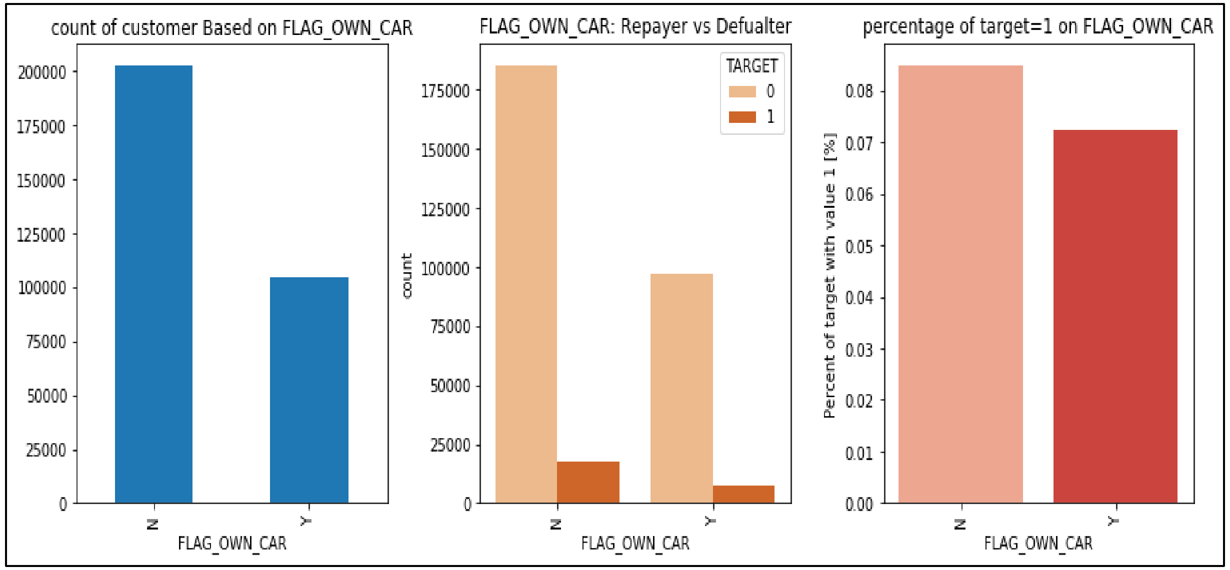
It can be clearly seen from the plot that most of the people apply loan being unaccompanied but customers of other B category have more chance of not being able to repay the loan on time, thus becoming the highest loan defaulter among NameType_Suite category



8)FLAG OWN CAR

	FLAG_OWN_CAR	TARGET
0	N	0.085002
1	Y	0.072437

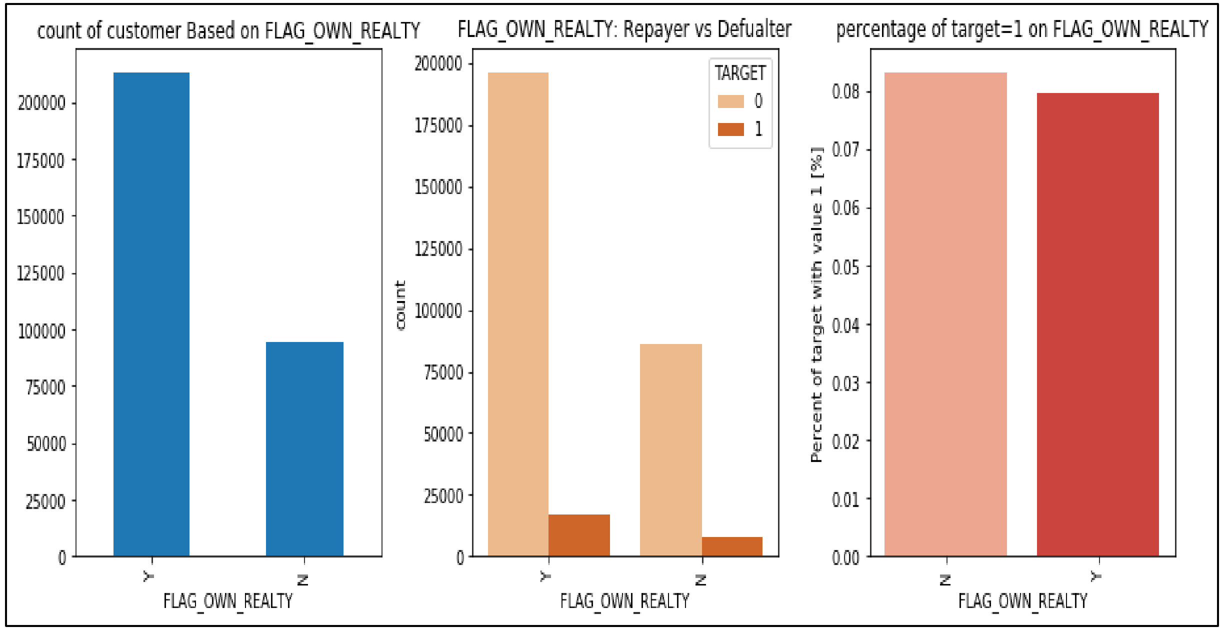
It can be seen from above plot that people who do not own a car,
have the high chance of being loan defaulter than the people who own a car.



9)FLAG OWN REALTY

	FLAG_OWN_REALTY	TARGET
0	N	0.083249
1	Y	0.079616

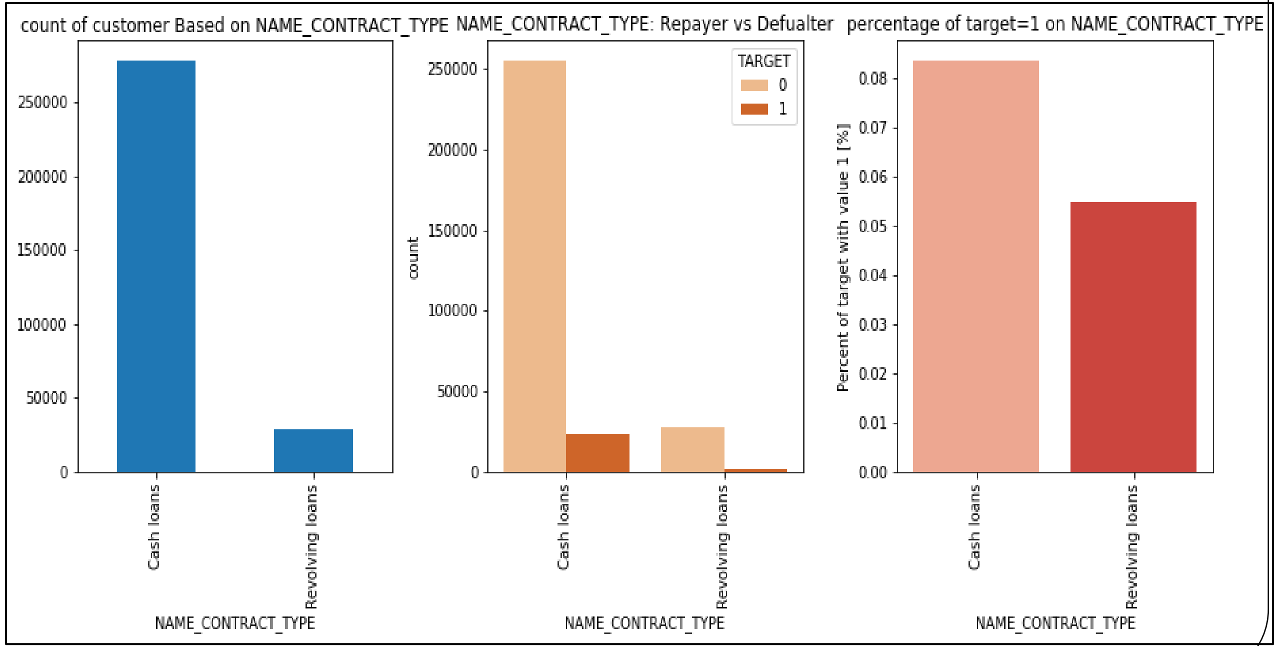
From the given plot, we can see that people who do not have their own house/flat apply for a smaller number of loans but has high chances of being defaulters than the people who have their own house or flat.



10)NAME CONTRACT TYPE

	NAME_CONTRACT_TYPE	TARGET
0	Cash loans	0.083459
1	Revolving loans	0.054783

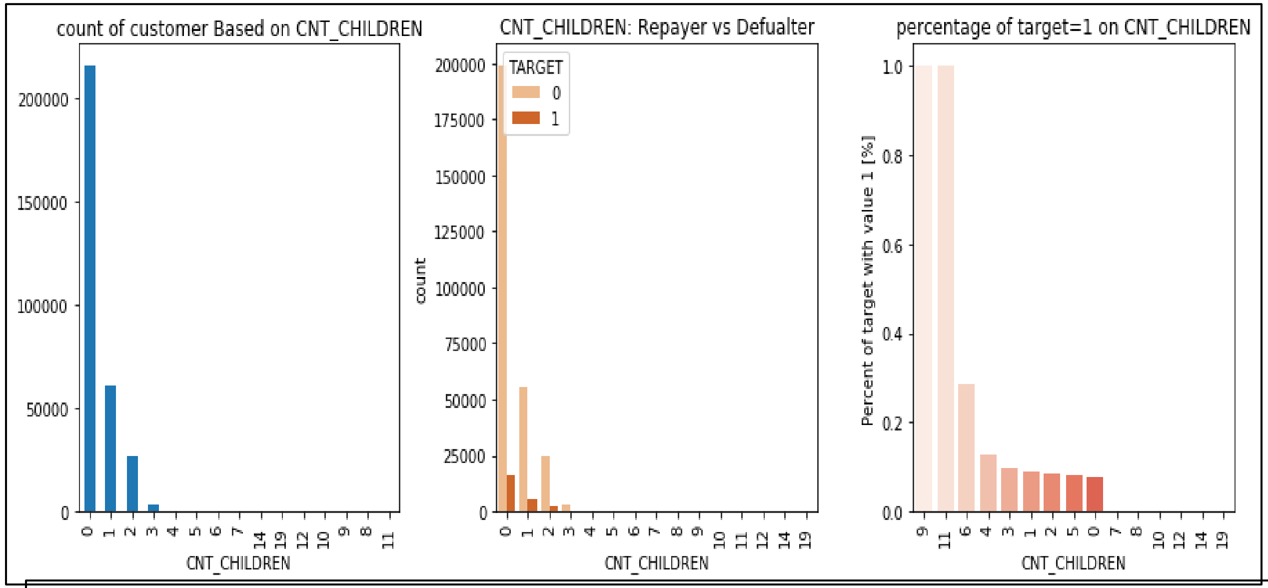
From the plot, it can be seen that people with cash loans have the high chance of being loan defaulter than people with revolving loans.



11)CNT CHILDREN

	CNT_CHILDREN	TARGET
9	9	1.000000
11	11	1.000000
6	6	0.285714
4	4	0.128205
3	3	0.096314
1	1	0.089236
2	2	0.087218

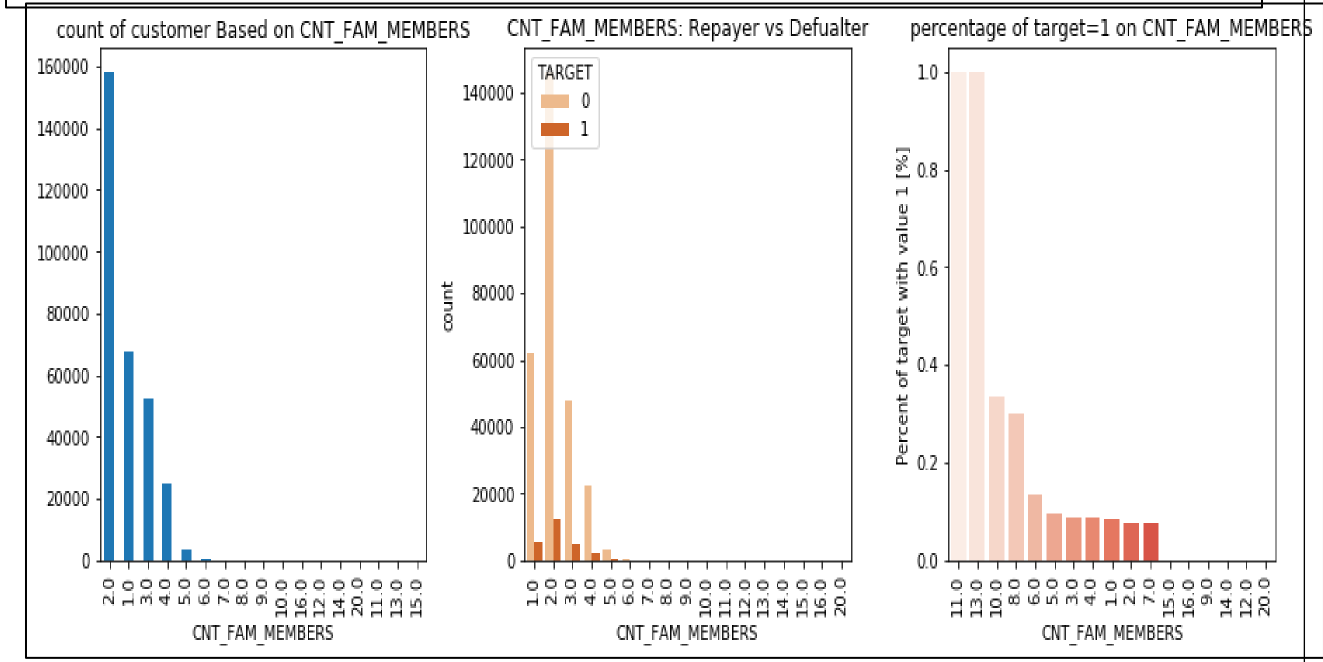
It can be clearly seen from above plot that the customers having a greater number of children, have high chances of not paying the loan than the other customers who have one or two children.



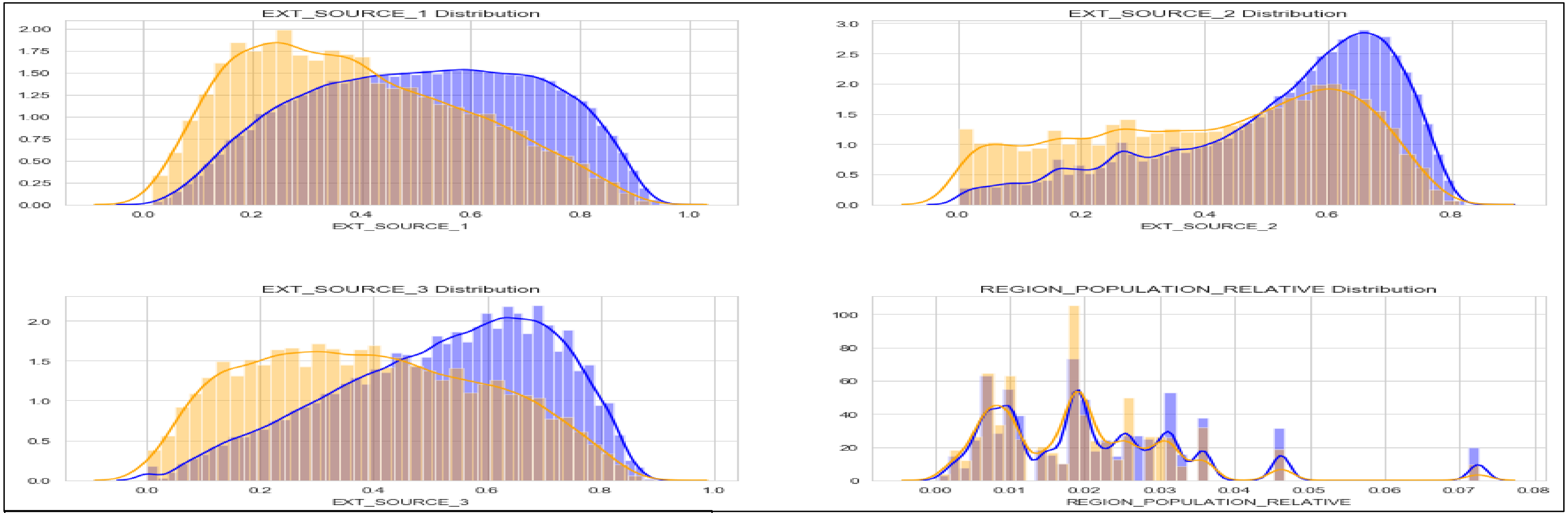
12)CNT FAMILY MEMBERS

	CNT_FAM_MEMBERS	TARGET
10	11	1.000000
12	13	1.000000
9	10	0.333333
7	8	0.300000
5	6	0.134804
4	5	0.094020
2	3	0.087603
3	4	0.086488
0	1	0.083644
1	2	0.075835

It can be seen from above plot that more the number of members in a family, higher the chances of being a loan defaulter.



Analysis on External source and Region population relative

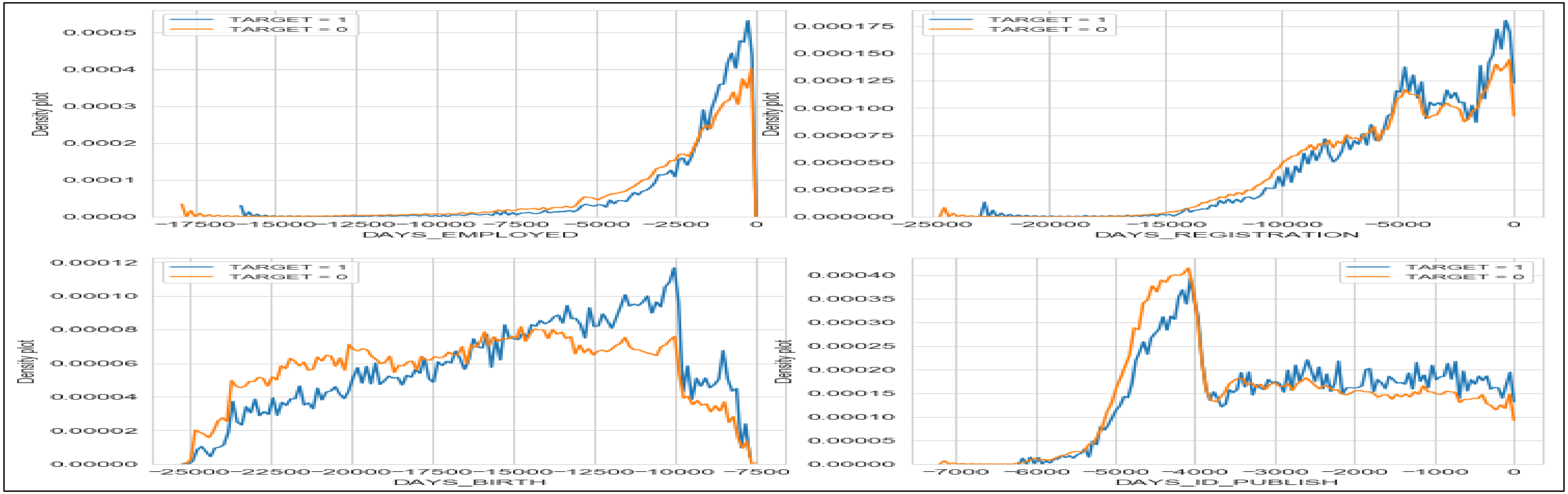


orange = defaulters   blue = Non defaulters

Inference :

1. **External source 1** shows a negative correlation with Target variable. When the Points from External Source gets higher, there is a higher chance of customer being default and shows maximum default at range 0.5 to 0.7
2. **External source 2**- There is a moderation distribution among defaulters in all range while there is maximum range for non defaulters when the ranger is higher at 0.6 and 0.8
3. **External source 3** - There is similar to distribution to distribution 1, When the Points from External Source gets higher, there is a higher chance of customer being default and shows maximum default at range 0.6 to 0.8
4. **Region population relative** - There is maximum default happening at 0.02

Analysis on 'DAYS_EMPLOYED', 'DAYS_REGISTRATION','DAYS_BIRTH','DAYS_ID_PUBLISH'



Inference:

1. From the previous plot, the applicants with less than 5 years of employment (-2500)are less likely to repay the loan.
2. The applicants with less than 5 years (-2500) of Registration_days (How many days before the application did client change his registration) are less likely to repay the loan
3. The applicants who are at thier 30's (1000) are less likely to repay the loan, while it decreases when tha age increases. Also, people less than 20(-8000) repay their loans
4. The applicants who changed the identity document with which he applied for the loan before 10 days (-4000) are less likely to repay the loan

Derived Metrics:

Type Driven Metrics:

Bins have been been created for continuous variables for easy and meaningful visualization

Data driven Metrics :

New columns have been created using the Amount credit and Amount annuity, Also for Total income and Amount credit inorder to visualze if there is any **correlation** between the variables

test_df['ratio__AMT_CREDIT__ANNUITY_RATIO'] = test_df['AMT_CREDIT'] / test_df['AMT_ANNUITY']

test_df['ratio__AMT_INCOME_TOTAL__AMT_CREDIT'] = test_df['AMT_INCOME_TOTAL'] / test_df['AMT_CREDIT']

Business Driven Metrics:

Total income can be classified into High, Medium and Low through a condition and the total income can be classified into three categories to identify how it is related with the Target variable

Driver Variables:

From the Univariate analysis done on various categories with respect to our target variable, following is the list of variables that influence the target variable :

1. **OCCUPATION_TYPE - *OCCUPATION-*** Occupation plays a major role as it is the source of income through which they can repay loan - Low skill laborers apply for a smaller number of loans but are in the highest category of loan defaulter followed by drivers, waiters and barmen staff.
2. **DAYS_BIRTH – *Age*** – Age plays a major role in providing loan as from the analysis the applicants who are at their 30's are less likely to repay the loan, while it decreases when tha age increases. Also, people less than 20 repay their loans
3. **DAYS_EMPLOYED- *WORK EXPERIENCE* -** Higher the work experience, higher will be their income .From the analysis the applicants with less than 5 years of employment (-2500)are less likely to repay the loan.
4. **AMT_CREDIT - *CREDIT HISTORY* –** The credit amount of the people plays a major role too. As per the analysis, the people has more credit on range 10,00,000 to 15,00,000 for more number of loans, higher chance of default : people with more credit on range 2,50,000 to 3,00,000
5. **AMT_INCOME_TOTAL - *INCOME*** – Income variable plays a major role as from analysis if we see the distrbution plot it clear says major distribution for amount of annual income is from 0 to 4,50,000 and people with total annual income lying between 100000 and 150000 avail more loans

CORRELATION ANALYSIS

The following are the top 10 positively correlated fields

+ve corr:

TARGET	1.000000
DAYS_BIRTH	0.078239
DAYS_EMPLOYED	0.074958
REGION_RATING_CLIENT_W_CITY	0.060893
REGION_RATING_CLIENT	0.058899
DAYS_LAST_PHONE_CHANGE	0.055218
DAYS_ID_PUBLISH	0.051457
REG_CITY_NOT_WORK_CITY	0.050994
FLAG_EMP_PHONE	0.045982
REG_CITY_NOT_LIVE_CITY	0.044395

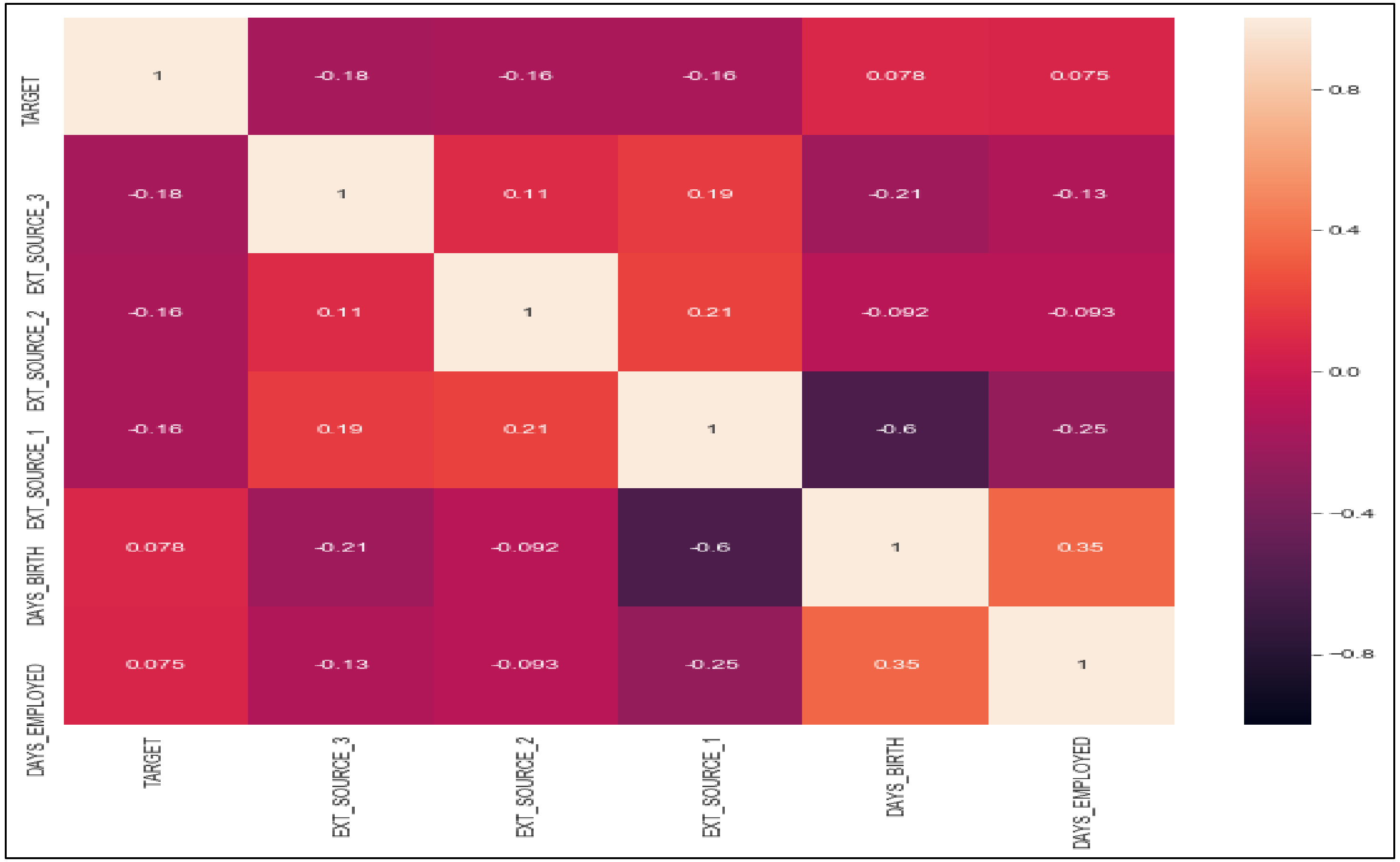
The following are the top 10 negatively correlated fields

-ve corr:

ELEVATORS_MEDI	-0.033863
ELEVATORS_AVG	-0.034199
REGION_POPULATION_RELATIVE	-0.037227
AMT_GOODS_PRICE	-0.039645
FLOORSMAX_MODE	-0.043226
FLOORSMAX_MEDI	-0.043768
FLOORSMAX_AVG	-0.044003
EXT_SOURCE_1	-0.155317
EXT_SOURCE_2	-0.160472
EXT_SOURCE_3	-0.178919

From the previous plot, as we can see EXT_SOURCE_3','EXT_SOURCE_2','EXT_SOURCE_1','DAYS_BIRTH','DAYS_EMPLOYED' are more correlated features, hence examining a little closer through heat map

Heat Map Analysis on External Source



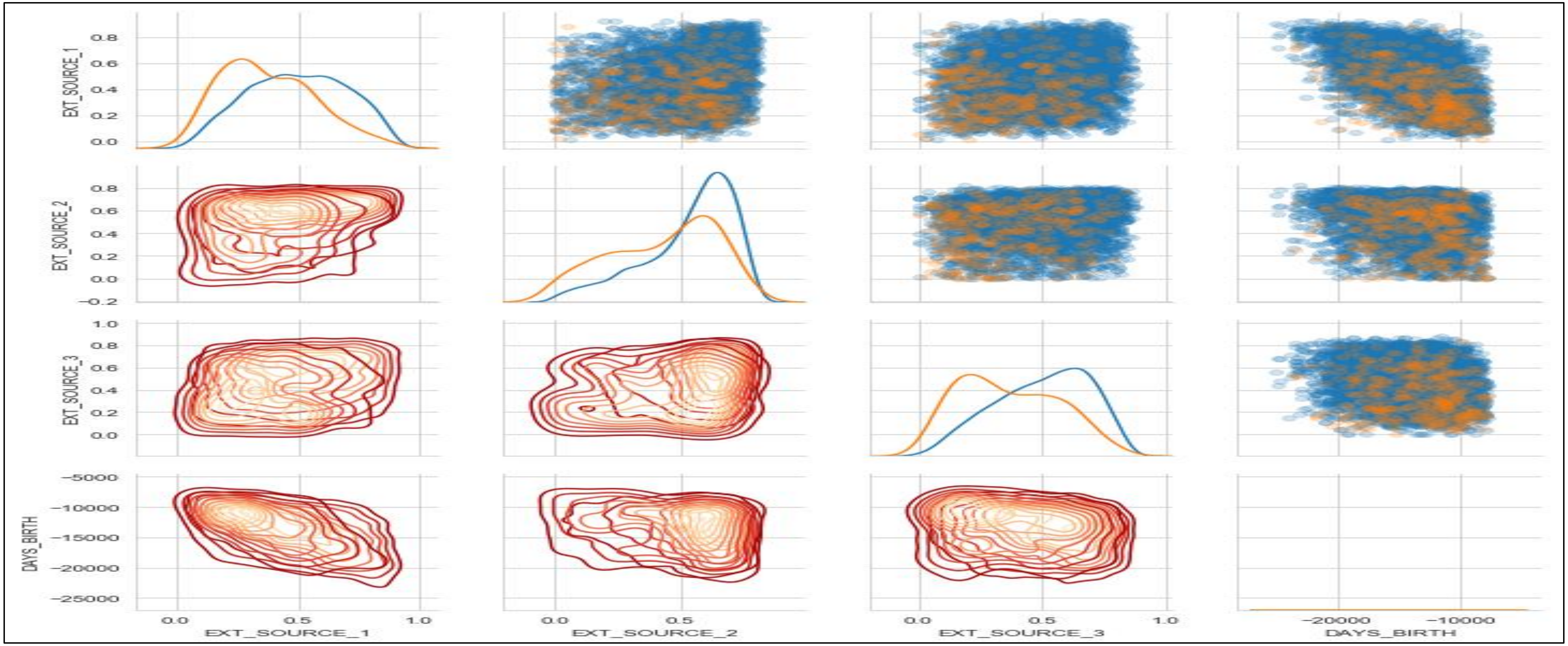
From the plot, highly negatively correlated variables are

1. External source 3 and days birth
2. External source 2 and days birth
3. External source 1 and days birth
4. External source 1 and External source 2
5. External source 1 and External source 3

Inference:

The heat map shows that all the external sources show a negative correlation with the Target

GRAPHICAL ANALYSIS: Pair plot and a Pair grid

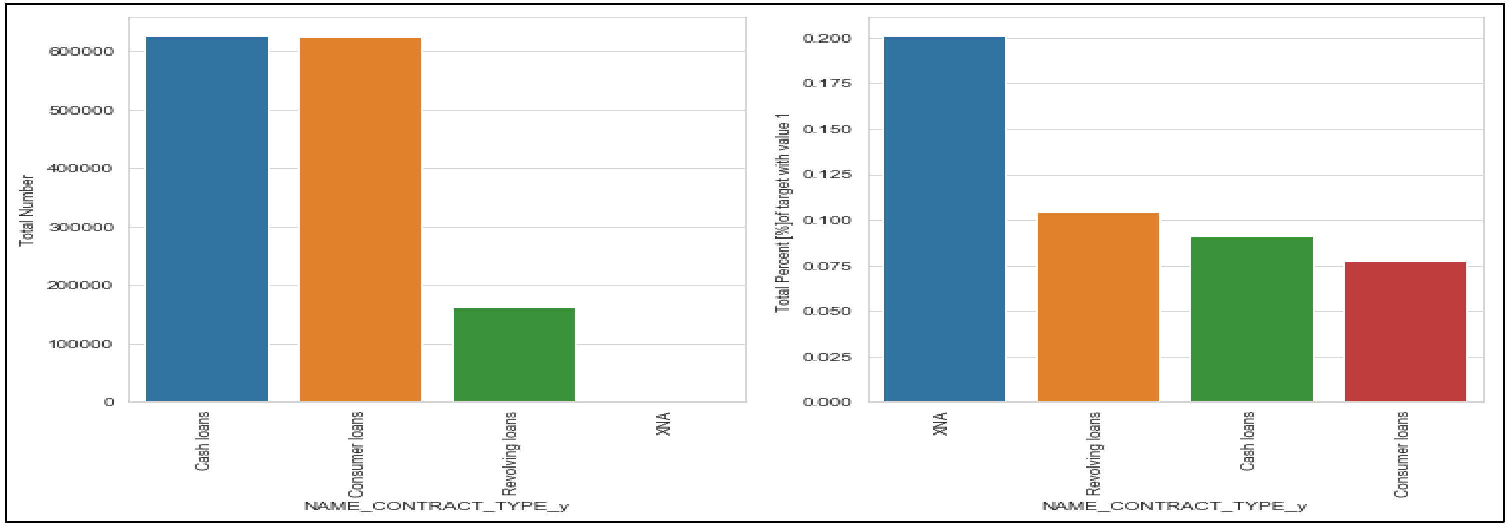


Having a look at previous plot, DAYS_BIRTH and EXT_SOURCE_, we can see that for TARGET=1 (i.e., orange) there is a high negative correlation.

UNIVARIATE ANALYSIS ON COMBINED DATASET

1) NAME_CONTRACT_TYPE_y

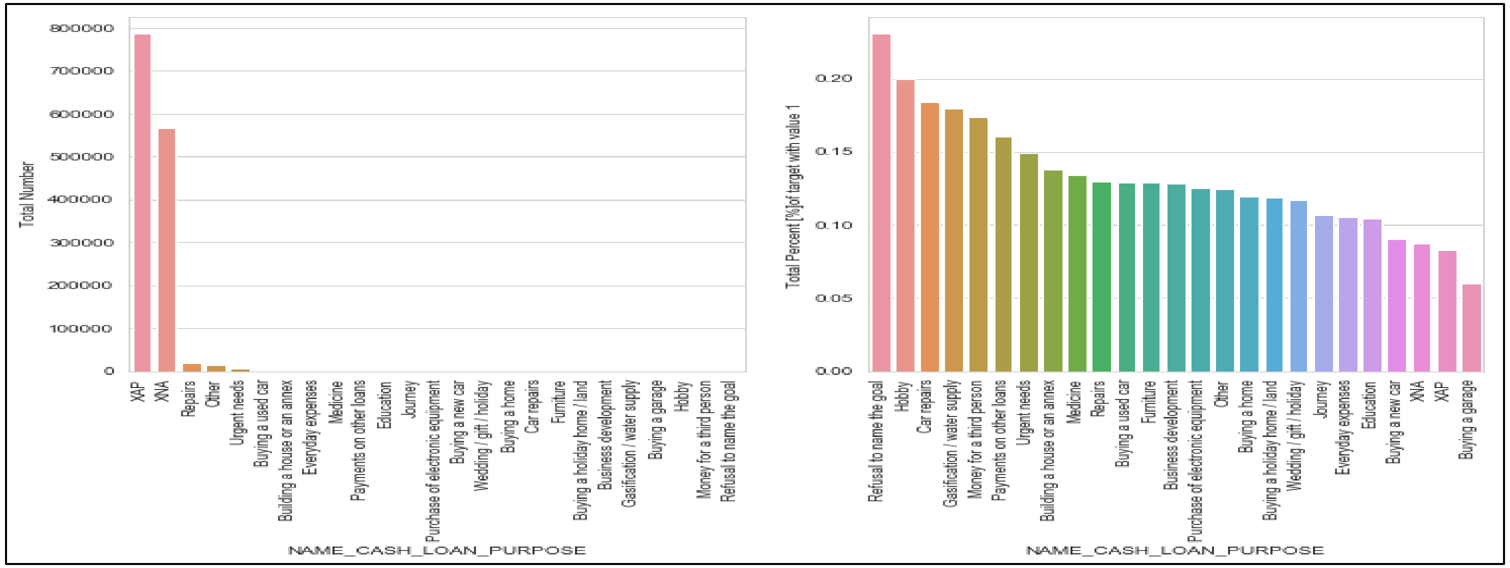
People with Revolving loans has the higher chance of being a default, while in the application data set people with Cash loans had the higher chance of default



2) NAME_CASH_LOAN_PURPOSE

XAP, XNA counts are higher, which are missing data and this needs to be removed. Apart from this - Repairs, Other, Urgent needs, Buying a used car, Building a house or an annex accounts for the largest number of contracts.

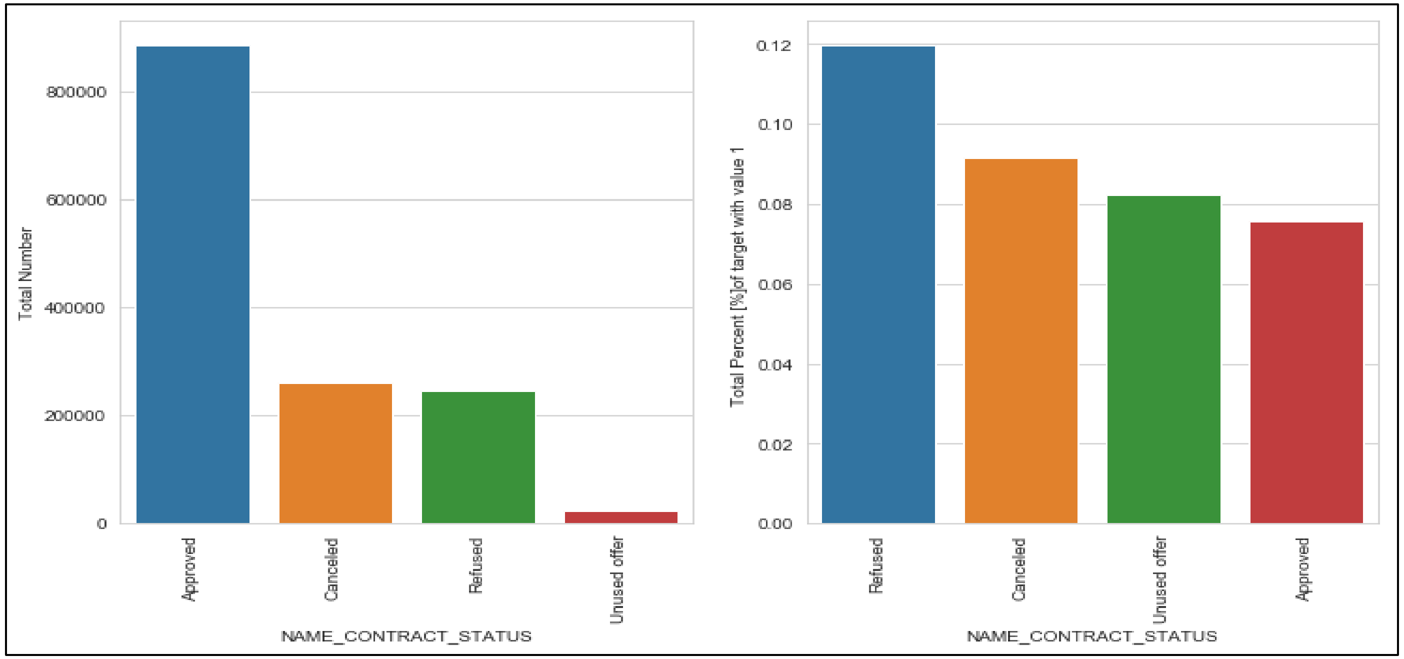
When looking at the percent of defaults for current applications in the sample, Refusal to name the goal - ~23% , Hobby (20%), Car repairs (~18%)



3)NAME_CONTRACT_STATUS

Most previous applications statuses are in Approved (~850K), Canceled and Refused (~240K). There are only ~20K in status Unused offer.

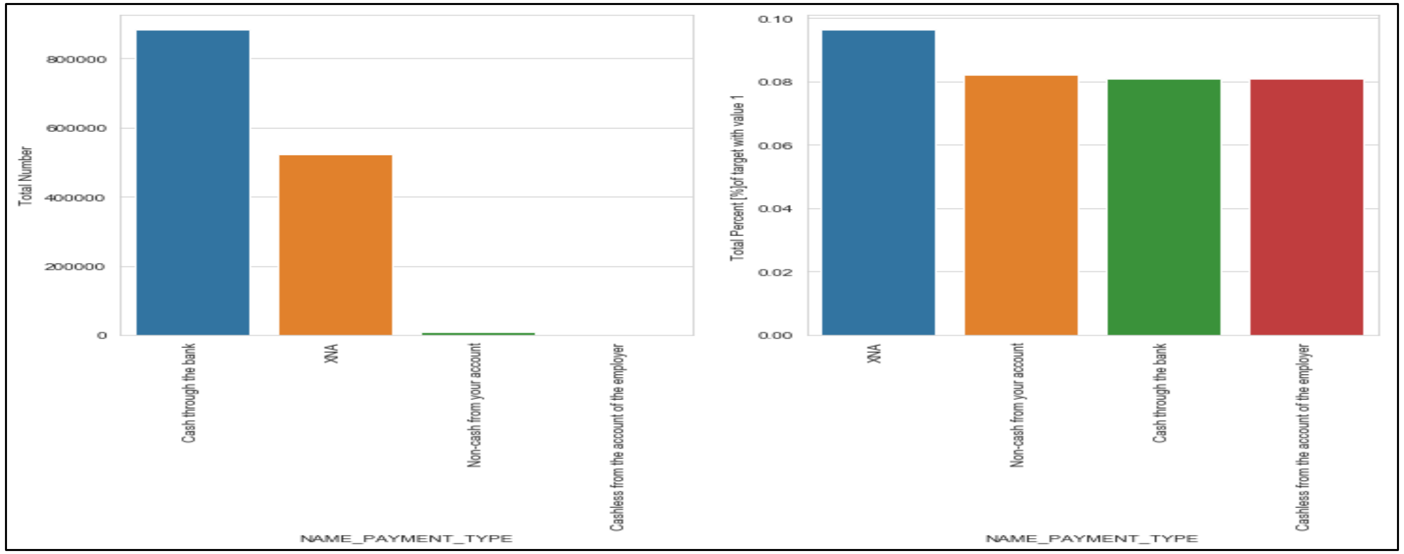
from the above plot the percent of defaults are Refused (12%), followed by Canceled (9%), Unused offer (~8%) and Approved (lowest percent of defaults in current applictions, with less than 8%).



4)NAME_PAYMENT_TYPE

Looking at the plot higher number of previous applications were paid with Cash through the bank (~850K). Payments using Non-cash from your account or Cashless from the account of the employer are very less.

And in the percentage distribution these three types of payments in previous applications results is almost the same



5) NAME_CLIENT_TYPE

Most of the previous applications have client type Repeater (~1M), just over 200K are New and ~100K are Refreshed.

In terms of default percent for current applications of clients with history of previous applications, current clients with previous applications have values of percent of defaults ranging from from 8.5%, 8.25% and 7% corresponding to client types in the past New, Repeater and Refreshed, respectively