CREDIT EDA CASE STUDY

By- Vishal D Mehta

&

Kajal Jaiswal

Problem statement :The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. On the basis of the applicant's profile, the company should be able to tell if the applicant is likely to repay the loan or not. This could be made possible by establishing key predictor variables within the given dataset which can in turn be used to aid decision making by the company on granting the loans to its clients. This will help the company to grant loans to clients who are likely not to default and deny those who are likely to default

With the end goal of this case study two data sets were given, namely:

- Application Dataset
- Previous Application Dataset

First, we did analysis for Application Dataset. It has 307511 rows and 122 columns.

Data cleaning was done before the analysis. Steps followed are below:

- First we found the % of missing values in each column in order to figure out which value to delete.

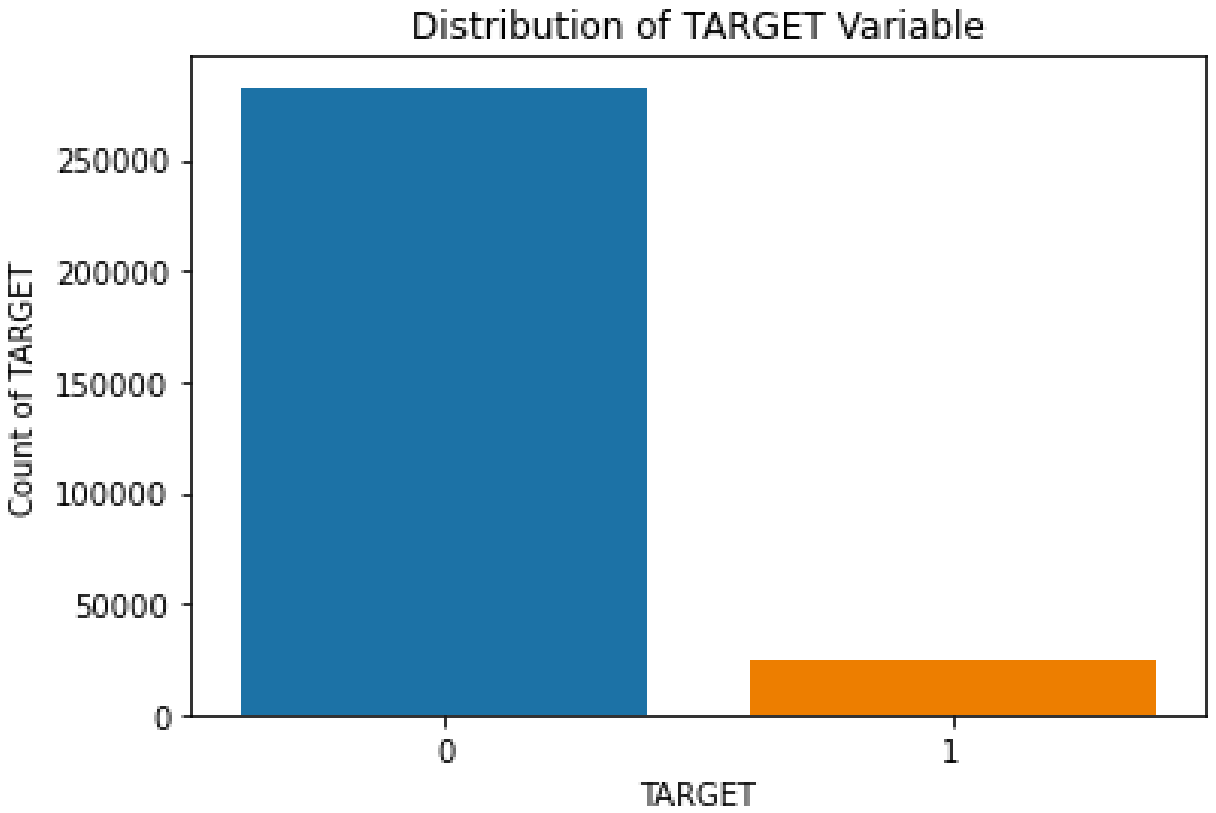
```
empty=app_data.isnull().sum()
null_percent = (app_data.isnull().sum()/app_data.isnull().count()*100)
unique_val = app_data.nunique()
datatypes = app_data.dtypes

pd.concat([empty, null_percent, unique_val, datatypes], axis=1, keys=['Null_count', 'Null_Perce
```

- Removed columns having null percent greater than 57

```
required_index=list(app_null_percent[app_null_percent.Null_Percent <57].index)
app_data=app_data[required_index]
app_data.describe()
```

CHECKING DISTRIBUTION OF TARGET VARIABLE

- We can see from below boxplot that dataset is majorly imbalanced . Loan repaid on time percentage is on higher side(92%,more than 25000 paid on time) and ess than 5000 loans were not repaid(8.07).
- Data imbalance ratio can also measured which comes around 11.39%.

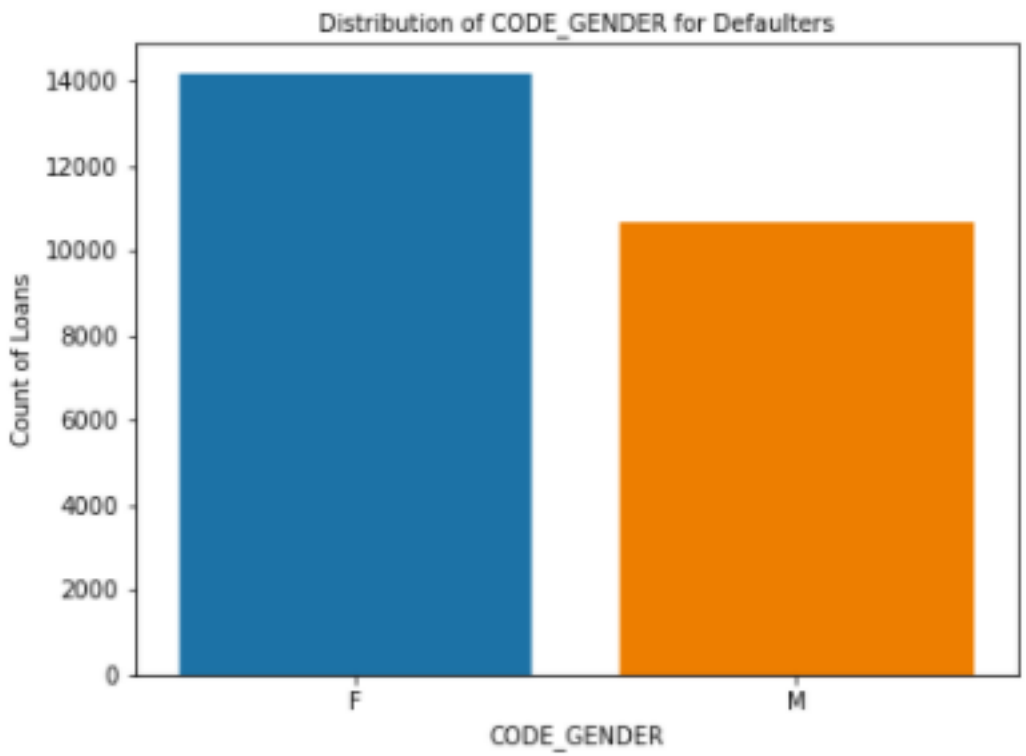


UNIVARIATE ANALYSIS ON CATEGORICAL COLUMNS FOR TARGET 0 AND TARGET 1

Distribution of CODE_GENDER:

For Defaulters means Target=0

- Here, We see that the quantity of Females taking credits is a lot higher than the quantity of Males
- The number of females taking credit loan is nearly about 140000 and that of male is 10500



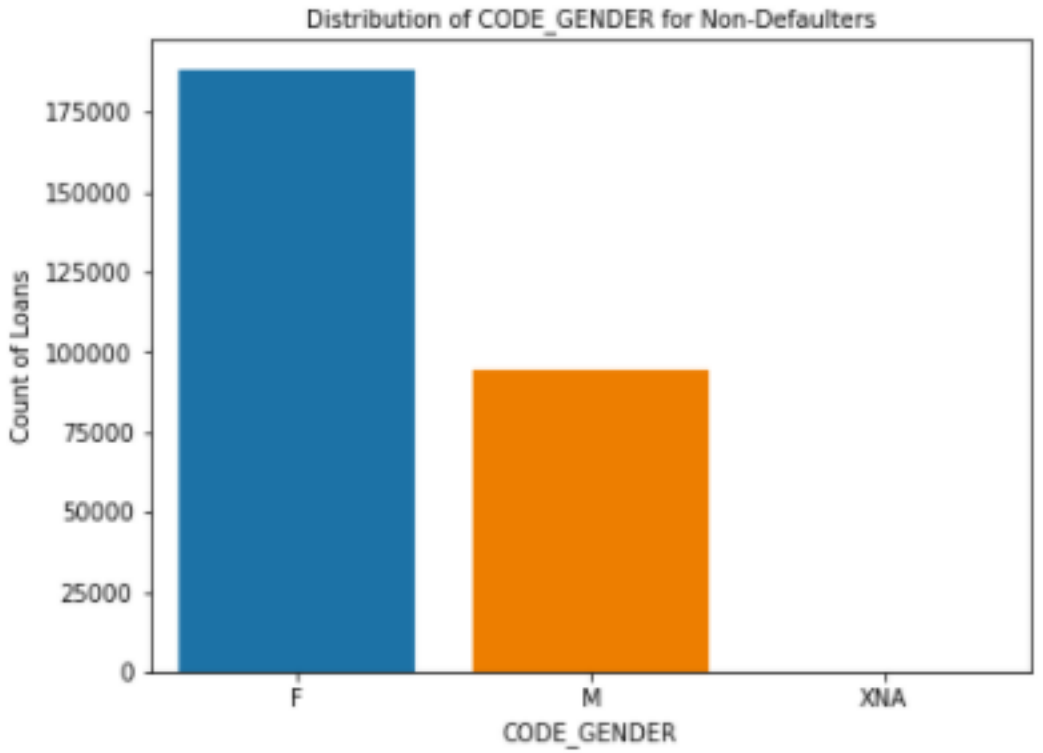
Distribution of CODE_GENDER:

For Non-Defaulters means Target=1

- Here, We see that the quantity of Females taking credits is a lot higher than the quantity of Males.
- The number of females taking credit loan is nearly about 185000 and that of male is nearly around 95000.

For both the case of Target Female are taking more credit than male

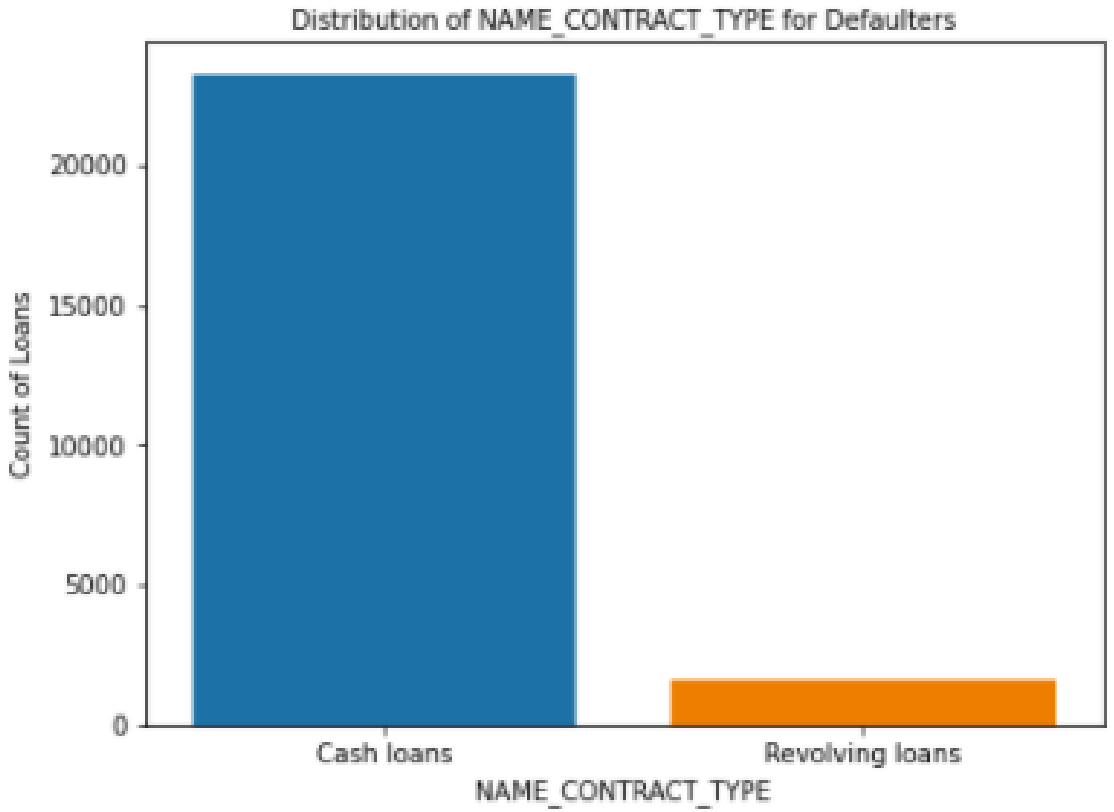
- The total count of female is 202452 (including Xna).
- Total count for male is 105059.



Distribution of NAME_CONTRACT_TYPE:

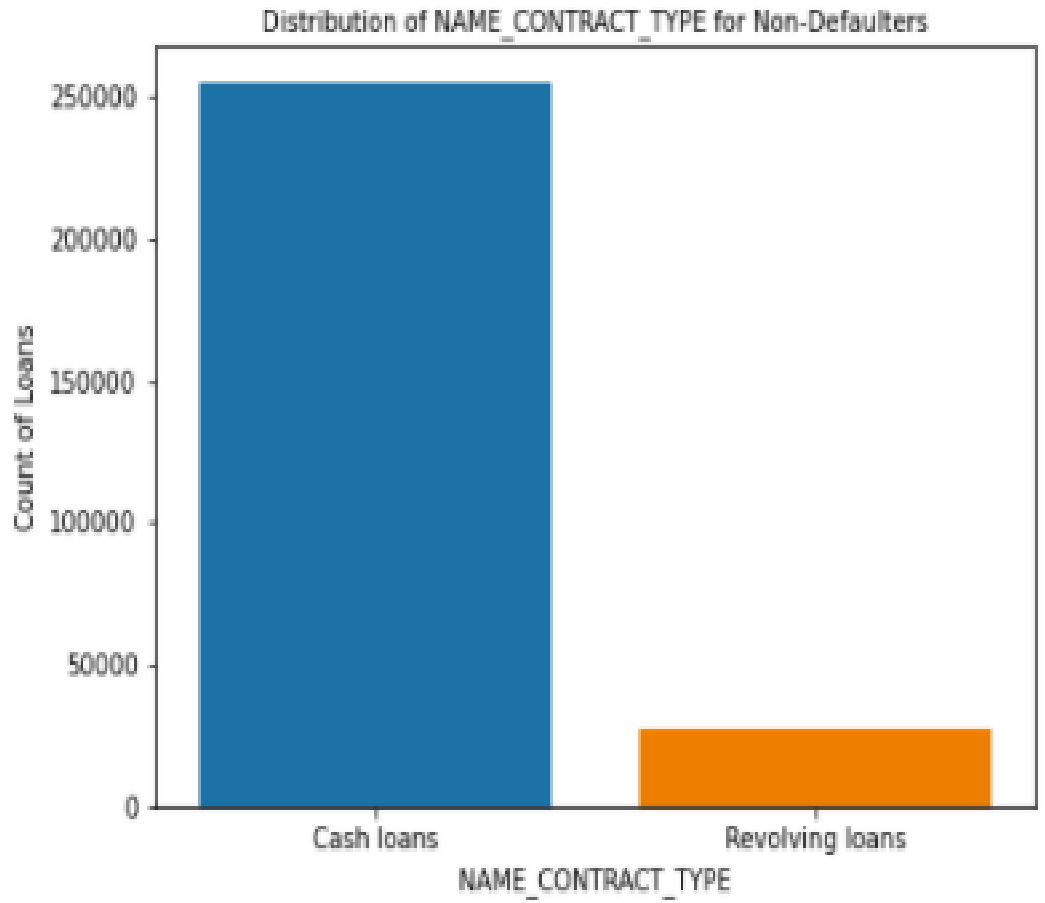
Focuses to be finished up from the chart on the right.

- For contract type 'Cash loans' is having higher number of credits than 'Revolving loans' contract type.
- The count for cash loans is near about 226224 and for revolving loans is 25000.



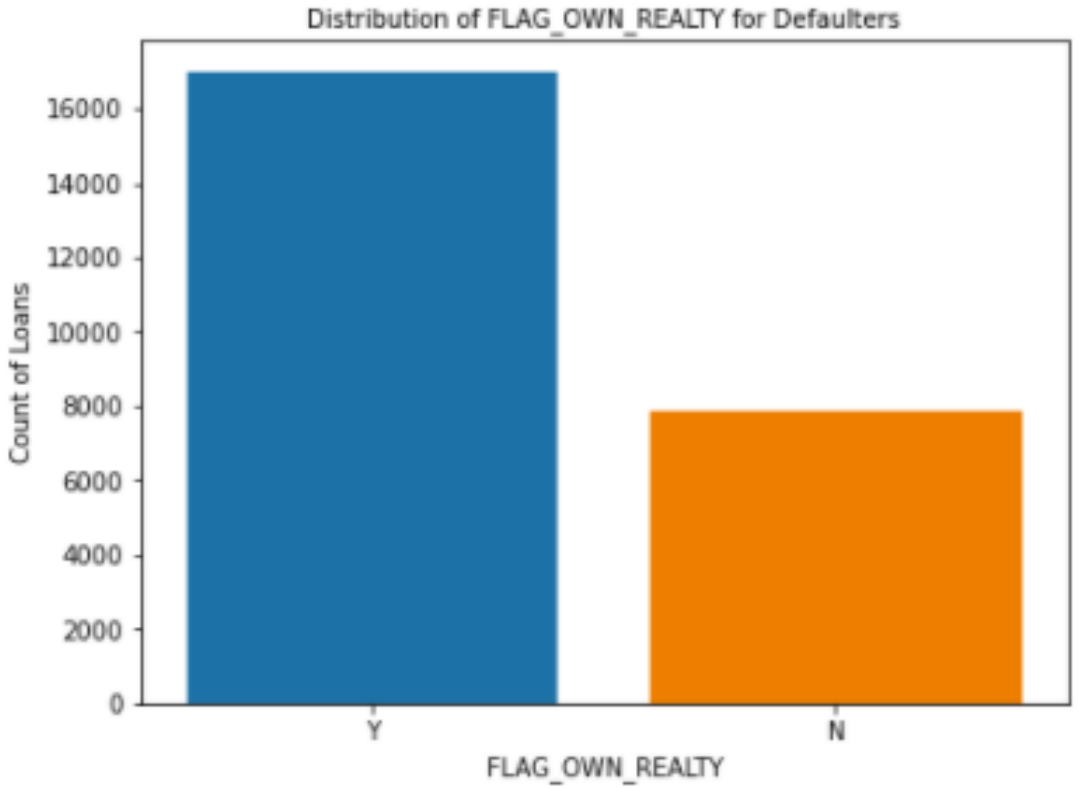
Distribution of NAME_CONTRACT_TYPE:
Focuses to be finished up from the chart on the right.

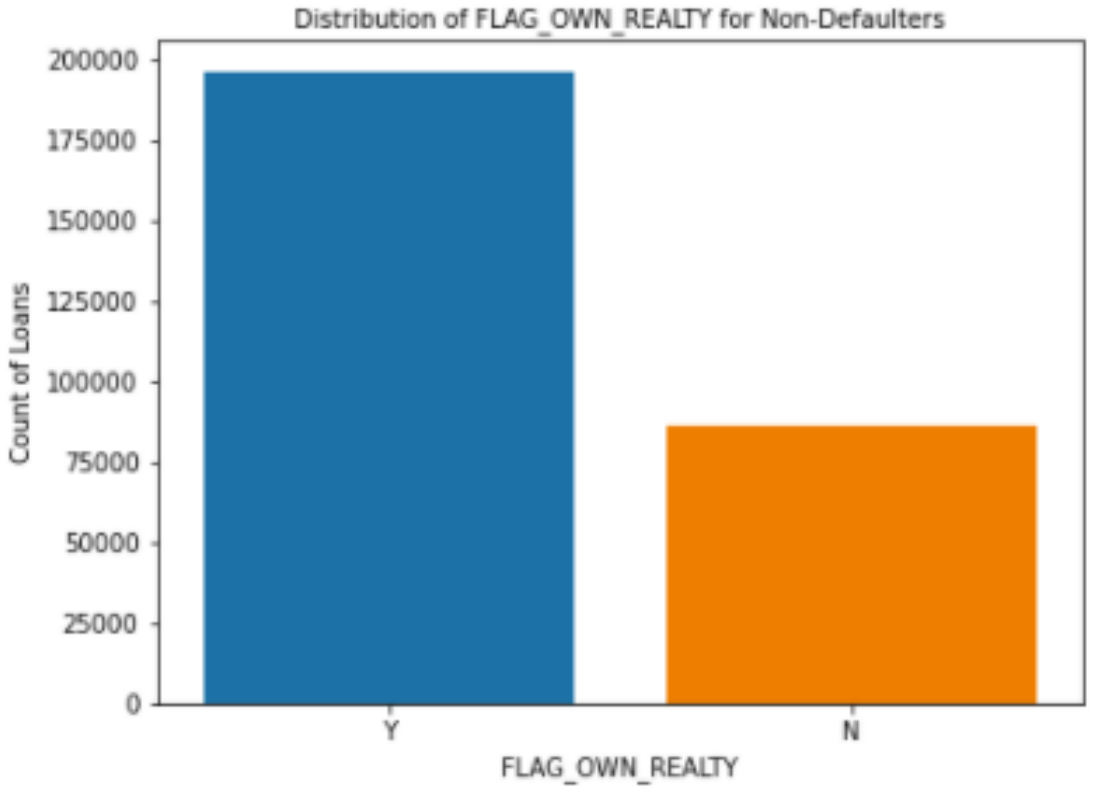
- For contract type 'Cash loans' is having higher number of credits than 'Revolving loans' contract type.
- The count for cash loans is near about 250000 and for revolving loans is 25000



DISTRIBUTION OF FLAG_OWN_REALITY

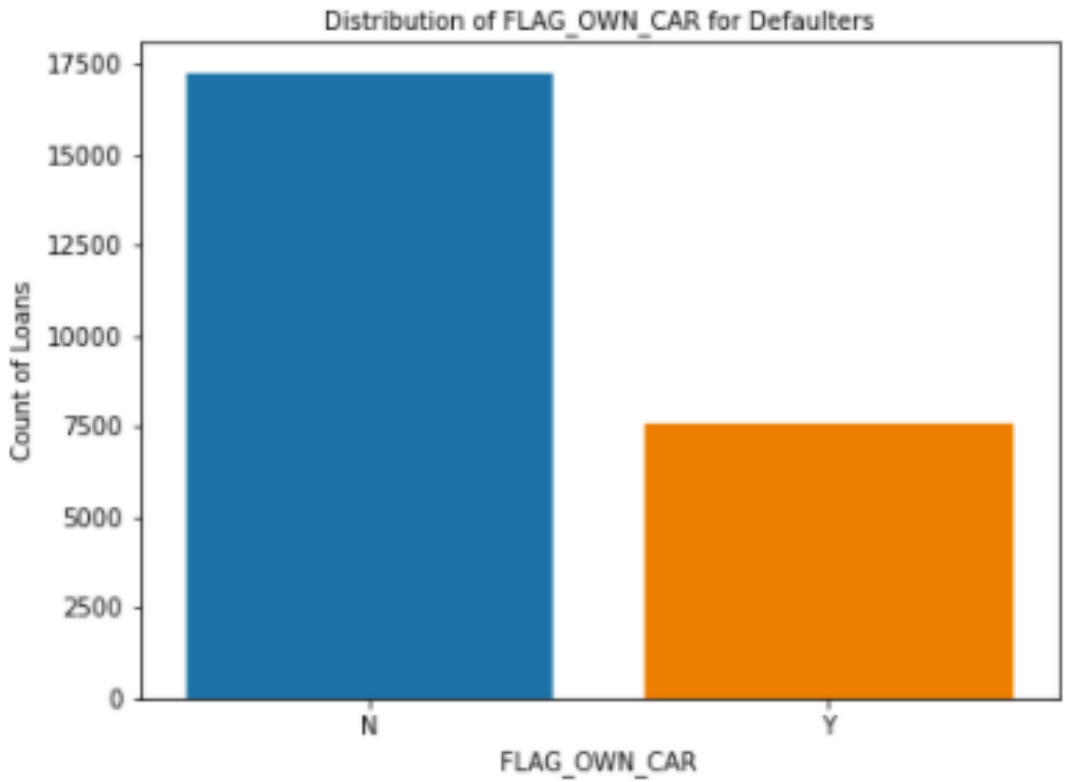
From the chart below, it is clear that, for both the target values 0 & 1, individuals who own reality have higher counts in contrast to those who don't.

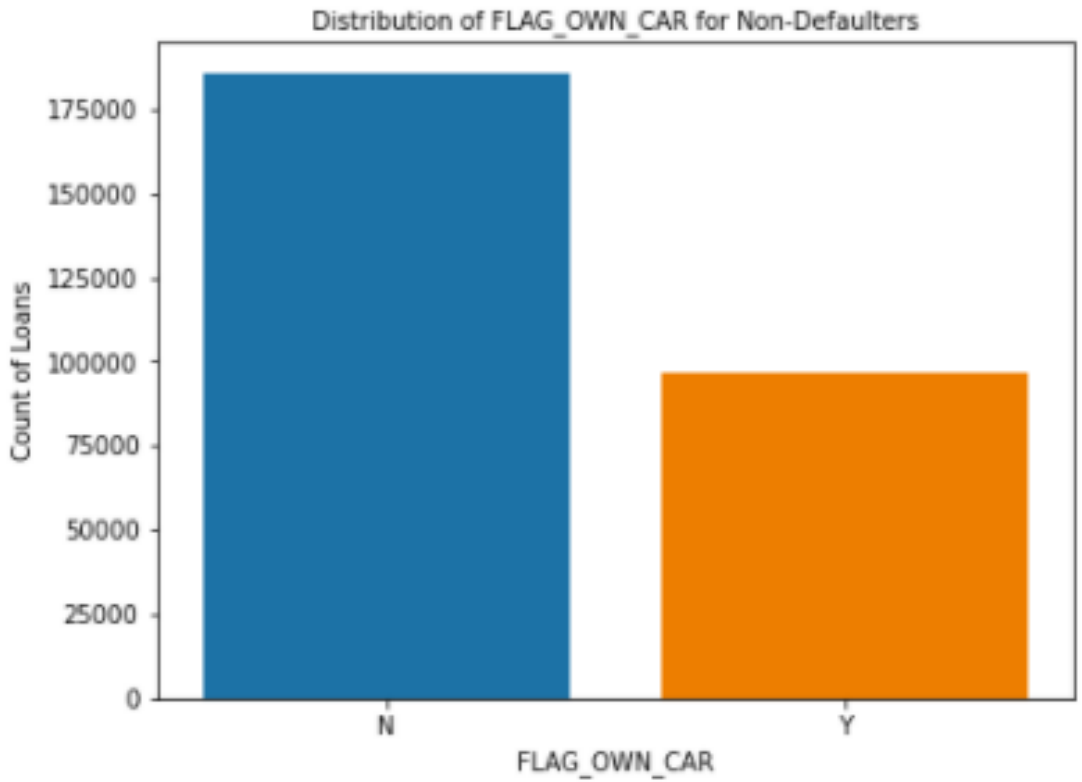




DISTRIBUTION OF FLAG_OWN_CAR

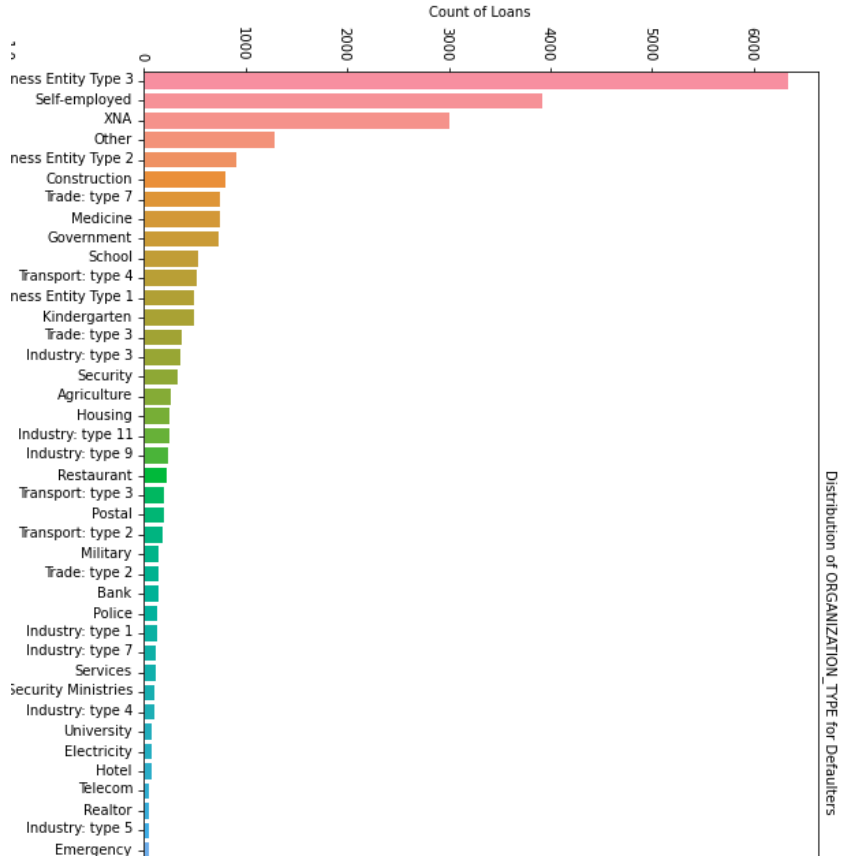
In the case of owning a car we can see that those who dont own have higher count in both the target cases compared to one's who do have





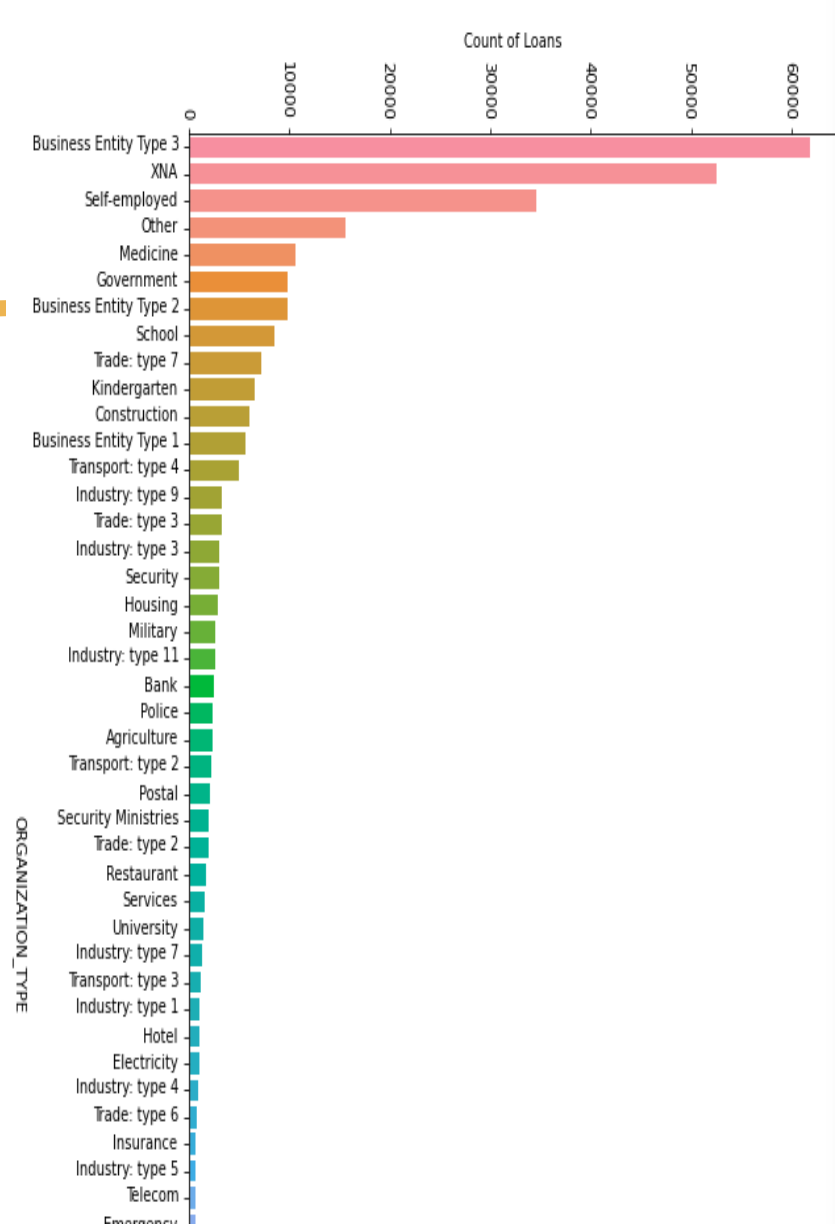
DISTRIBUTION OF ORGANIZATION_TYPE FOR TARGET 0

- Points to be concluded from the graph on the right.
- Clients which have applied for credits are from most of the organization type 'Business entity Type 3' , 'Self employed' , 'Other' , 'Medicine' and 'Government'.
- Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4.



DISTRIBUTION OF ORGANIZATION_TYPE FOR TARGET 1

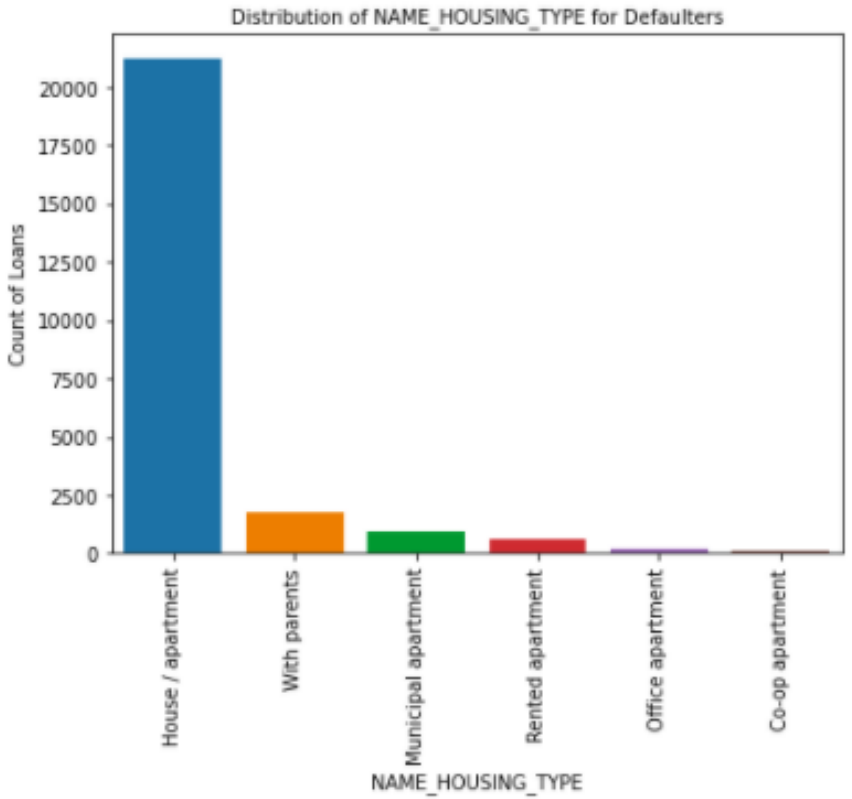
- Points to be concluded from the graph on the right.
- Clients which have applied for credits are from most of the organization type 'Business entity Type 3' , 'Self employed' , 'Other' , 'Medicine' and 'Government'.
- Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4.
- Same as type 0 in distribution of organization type.

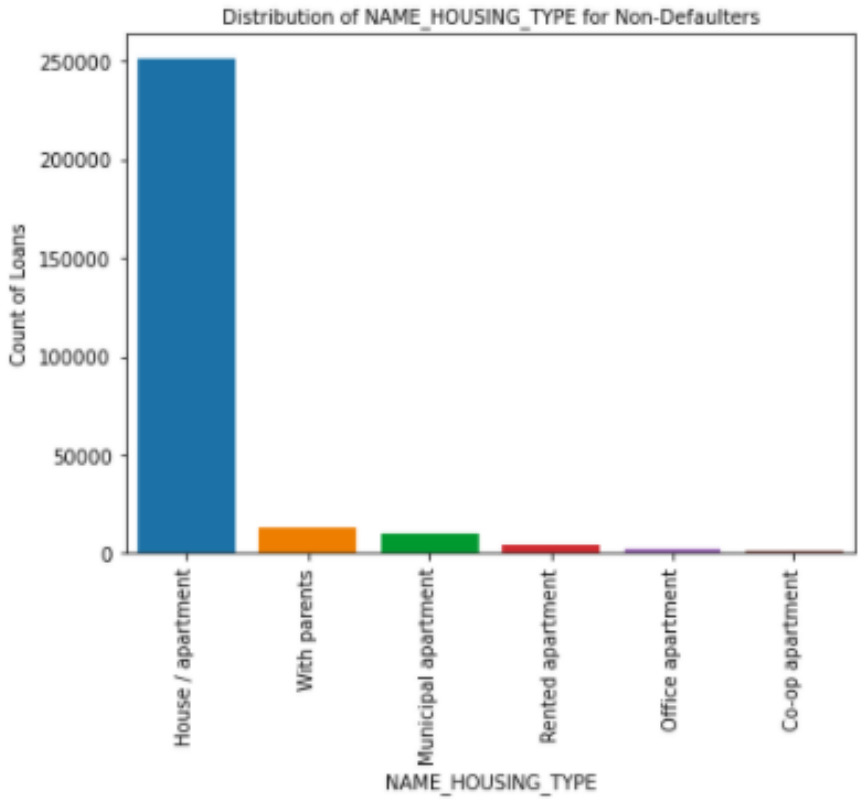


DISTRIBUTION OF NAME_HOUSING_TYPE

Observation for the housing type is :

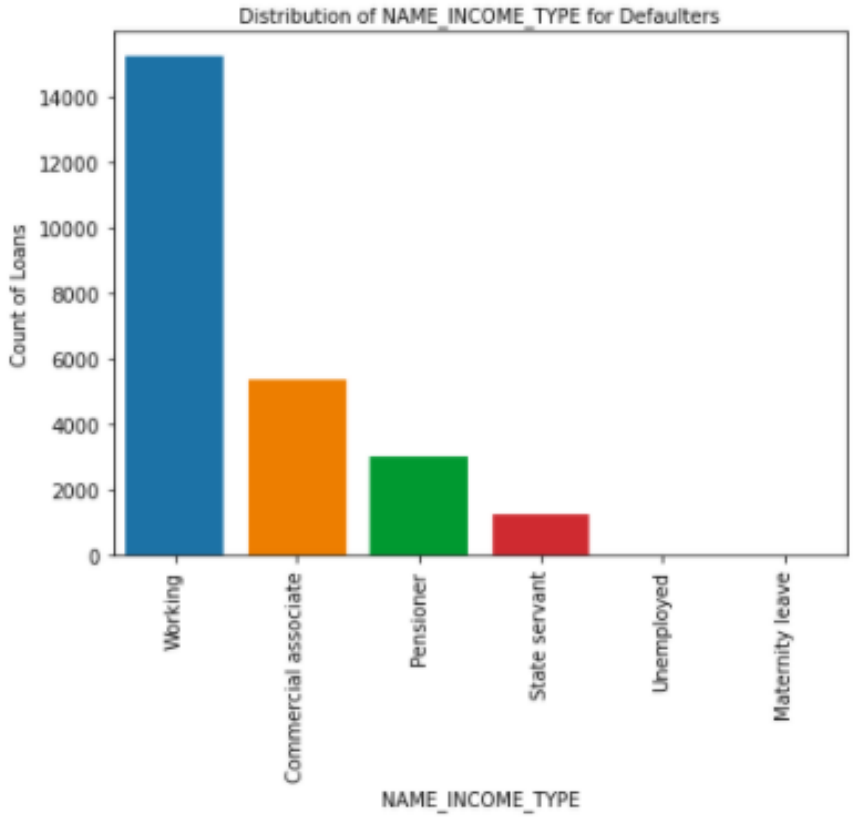
- Most people live in a House/Apartment have higher values for defaulting and non-defaulting ,co-op apartment have the least count.
- Ratio of People who live With Parents is slightly higher for defaulter than non-defaulters. It tells us that applicant who live with parents have a higher chance of having payment difficulties.

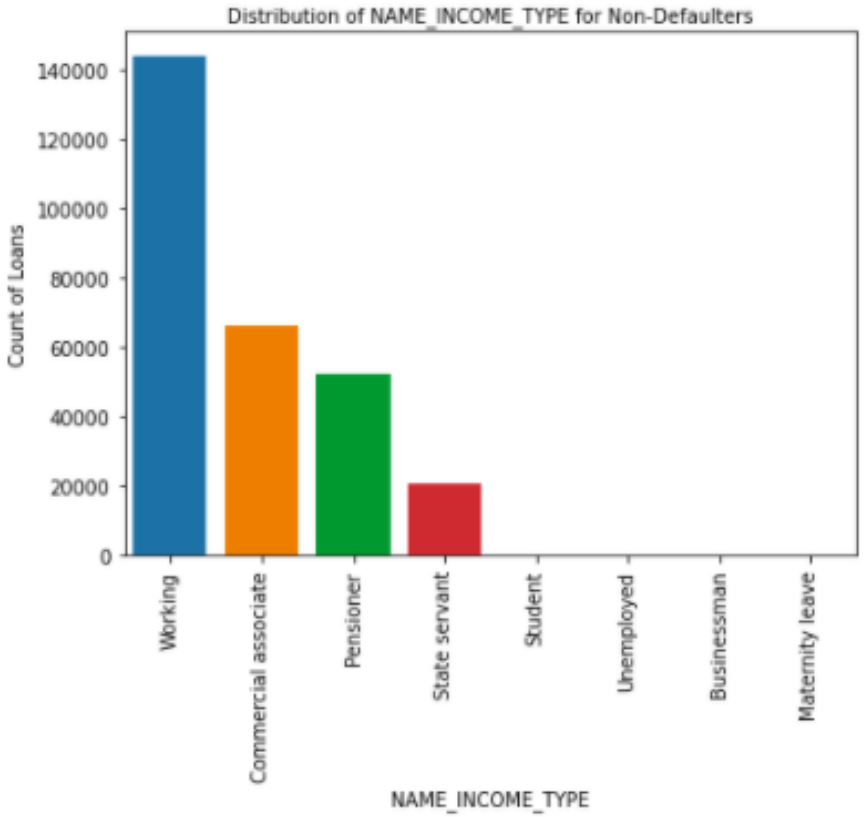




DISTRIBUTION OF NAME_INCOME_TYPE

- Commercial associates, Pensioner, State Servants have a higher ratio to total in non-defaulters.

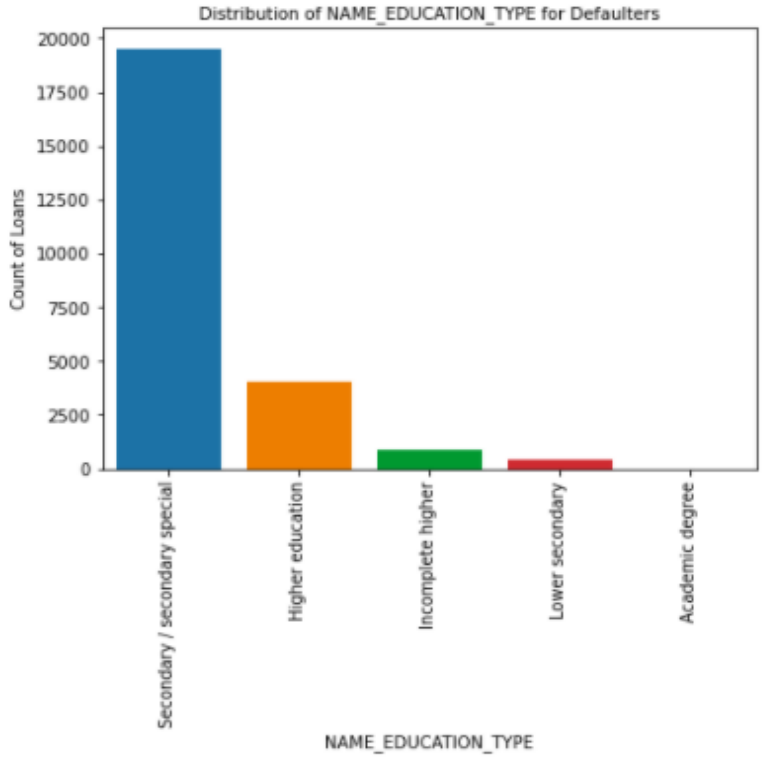


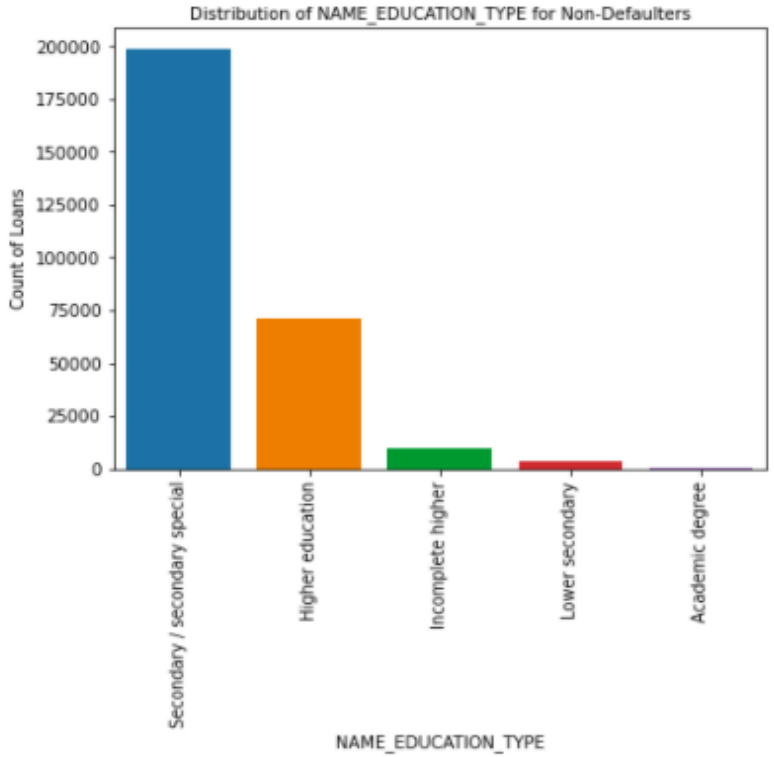


DISTRIBUTION OF NAME_EDUCATION_TYPE

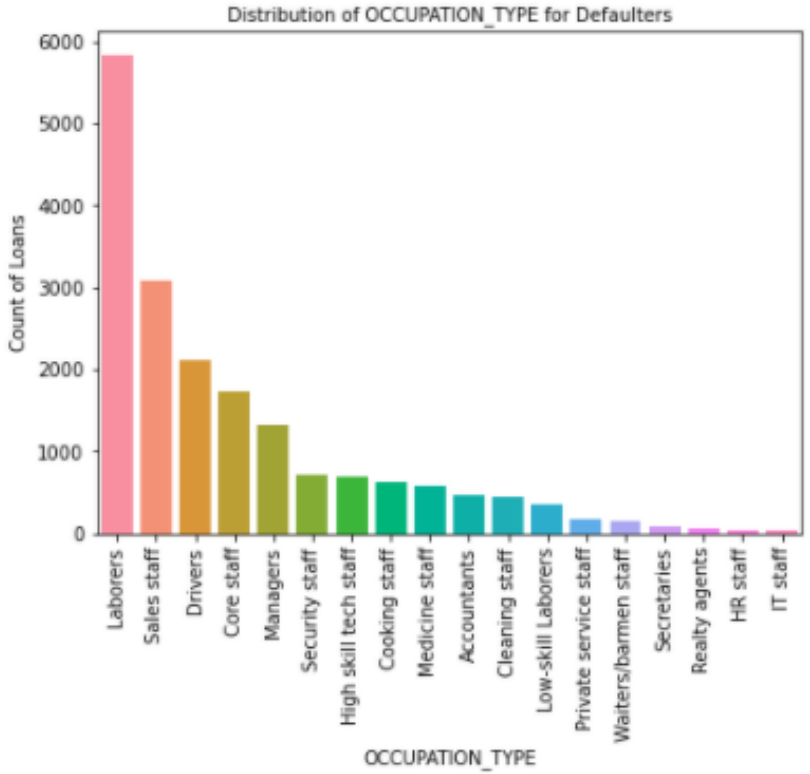
While the category with highest count remains same.(Secondary/secondary special)

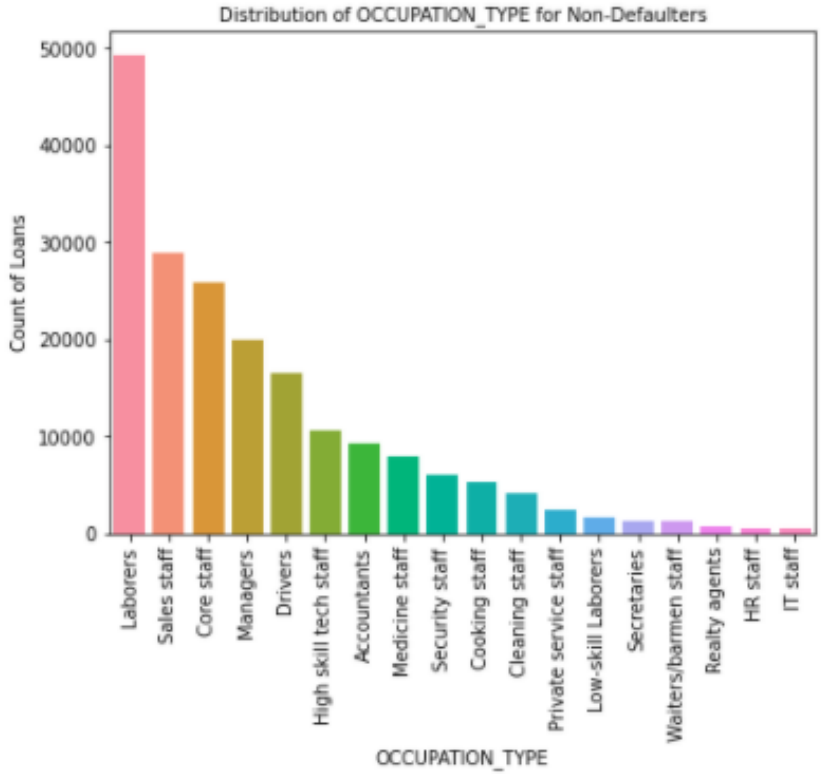
- This chart tells us that people with Academic Degree rarely take loans and are rarely defaulters. So they are potentially good customers.
- People with higher education are less likely to have payment difficulties. The Ratio is higher for non-defaulters than defaulters.





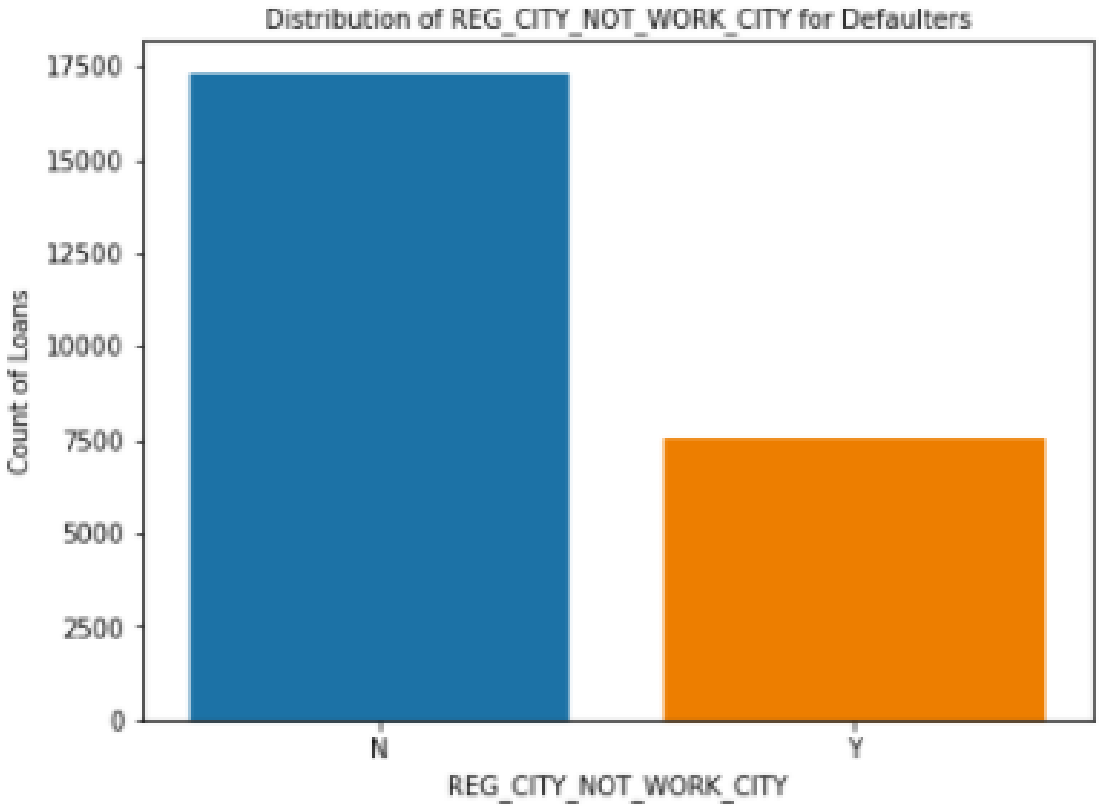
DISTRIBUTION OF OCCUPATION TYPE

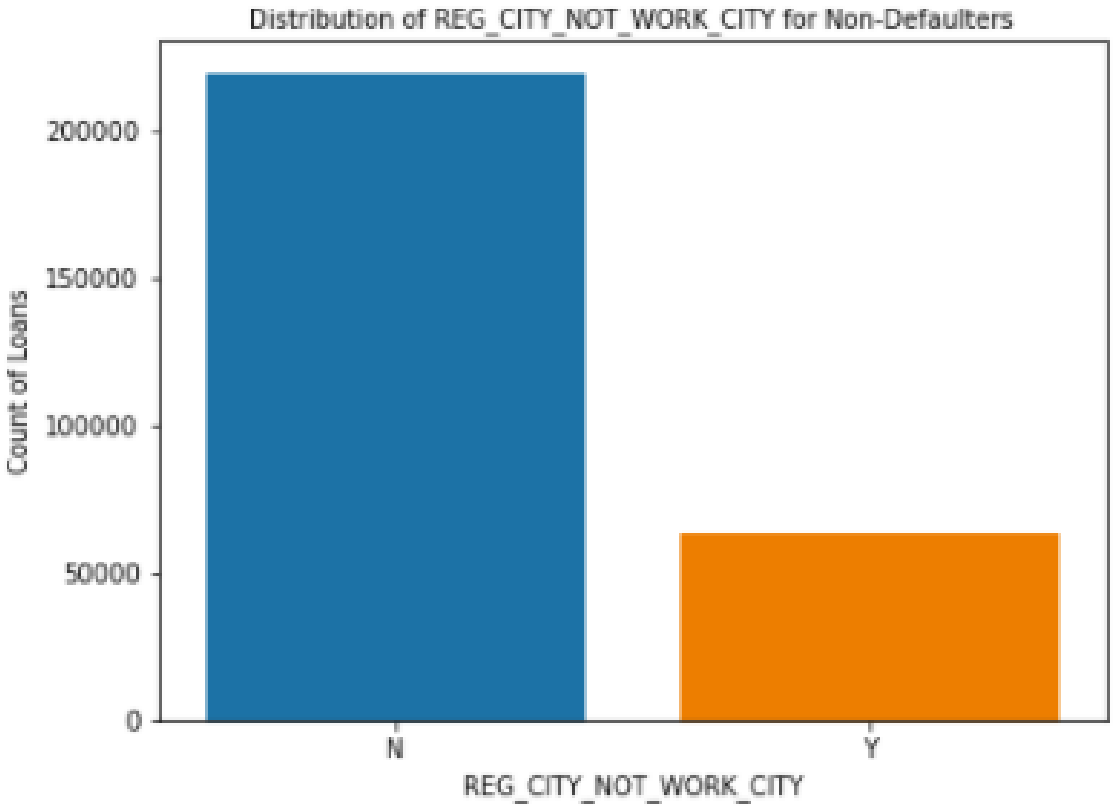




DISTRIBUTION OF REG_CITY_WORK_CITY

We observe that the Ratio of people whose Registration City is not the same as live city or work city is higher in case of defaulters are compared to defaulters. It tells us that people who live or work in a city different than the registration city are more likely to have payment difficulties

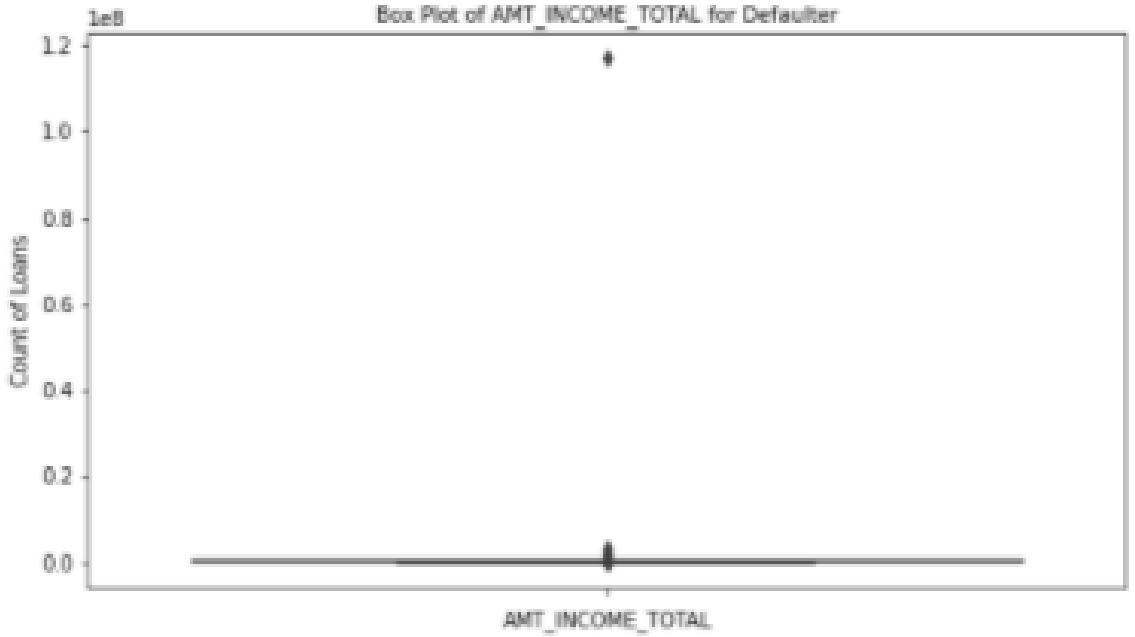


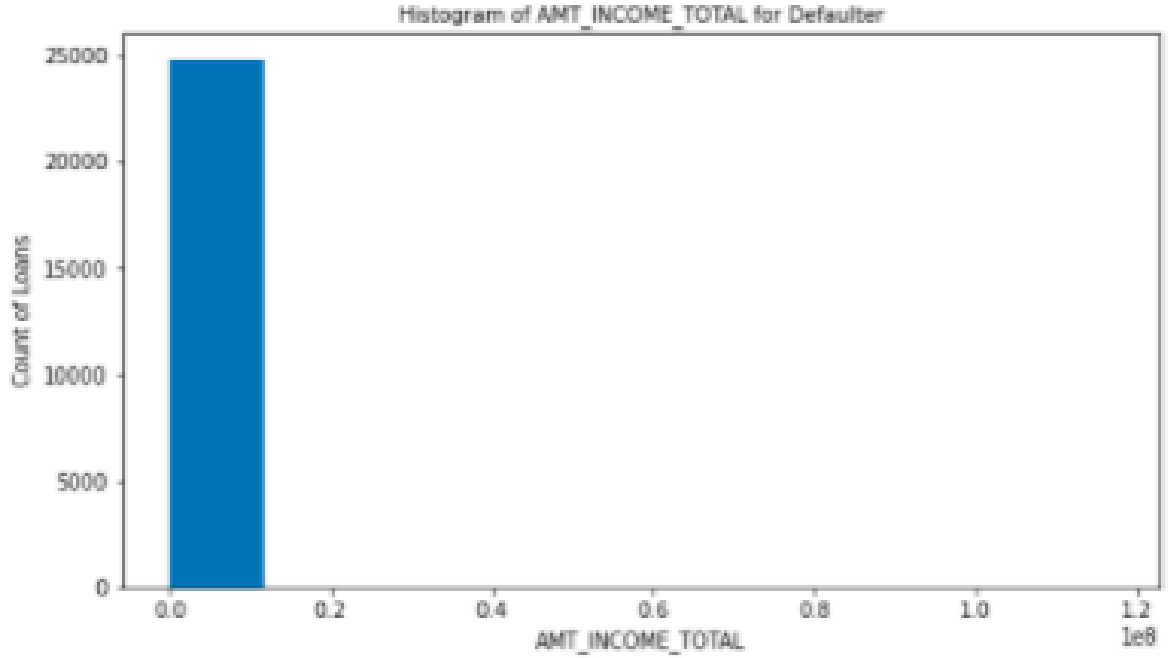


ANALYSIS ON NUMERICAL COLUMN- UNIVARIATE & BIVARIATE

ANALYSIS ON AMT_TOTAL_INCOME

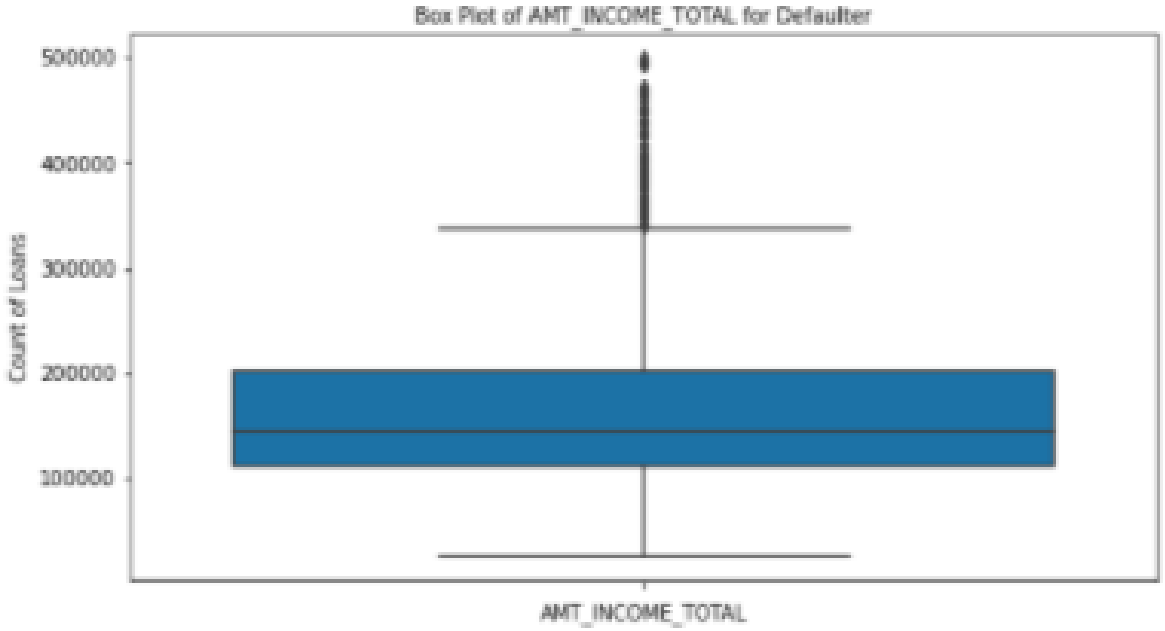
Now, we have filtered the numeric column by selecting datatype "float64" and plotted a box plot and hist plot for AMT_TOTAL_INCOME and we can see that there are lots of outliers for amt_income_total and we need to handle them accordingly

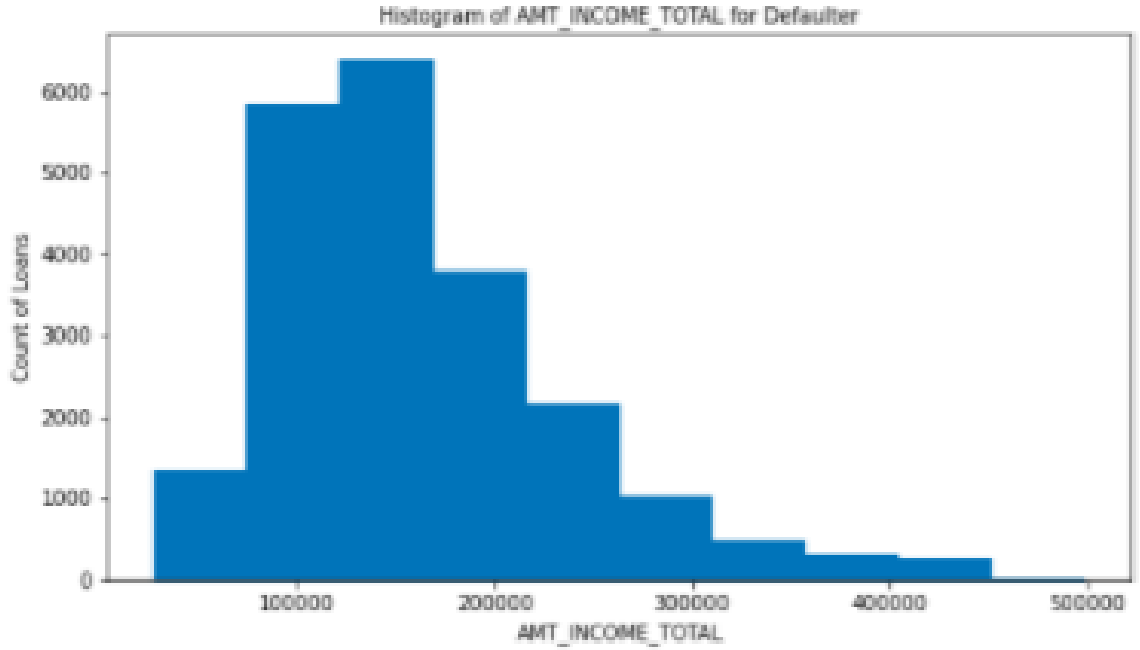




After counting quantile for AMT_INCOME_TOTAL, we found that 99% 99% of the values are within the value of 517500 and hence we can remove all entries above 99 percentile. Hence, plotting the graph there after-

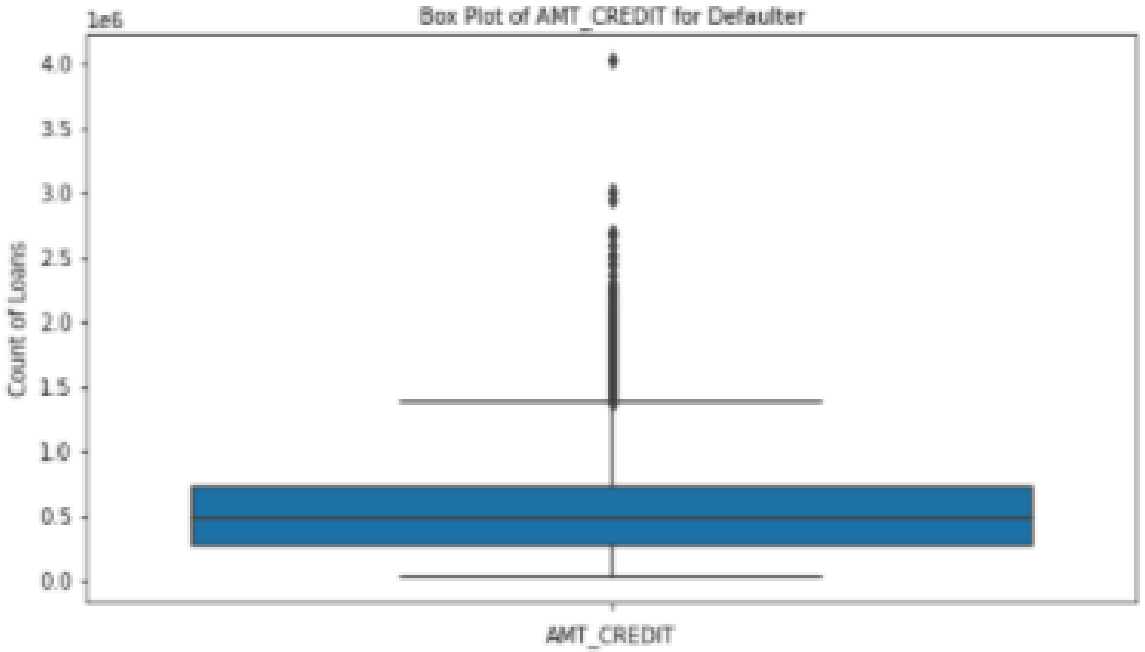
We can now see that defaulters are generally who have salary on the lower side between 100000 and 200000 with some on higher end and some on lower end

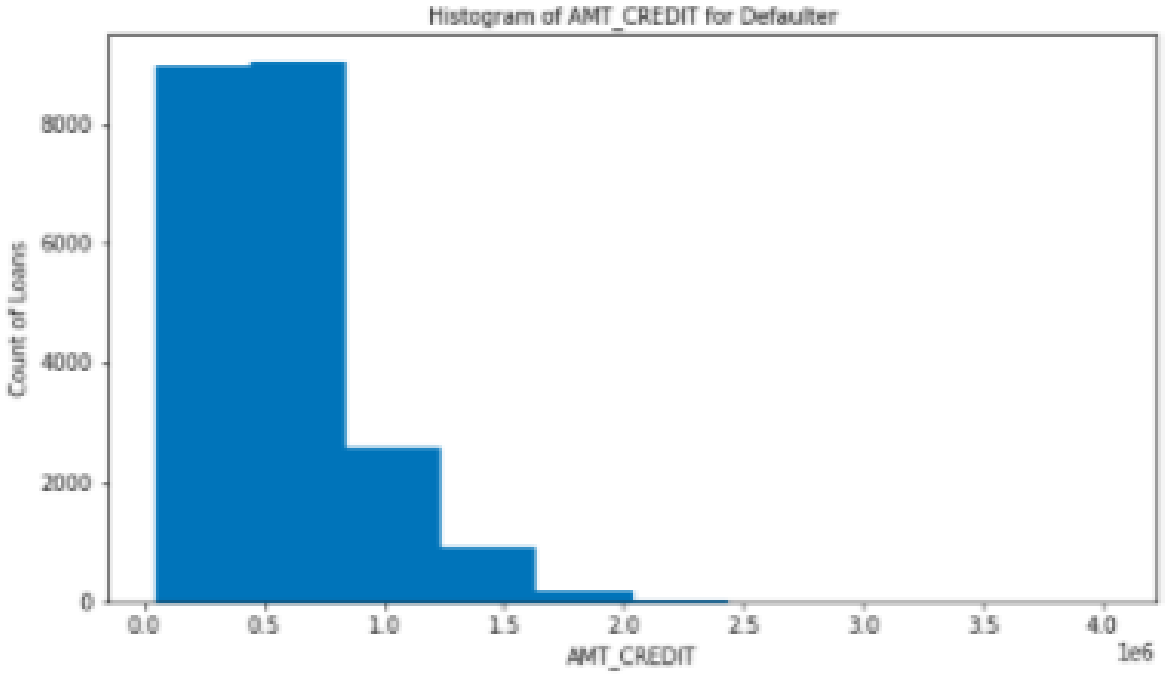




ANALYSIS ON AMT_CREDIT

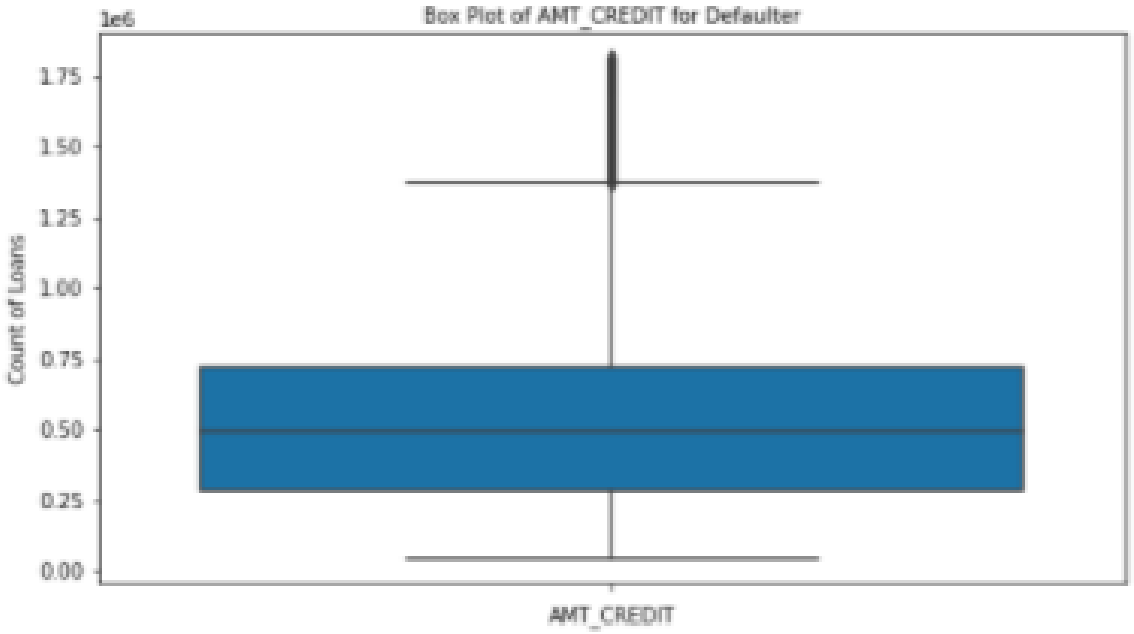
By calculating quantile for AMT_CREDIT, 99 percent values for this is within 1817491 we can remove the value above this considering it as outlier

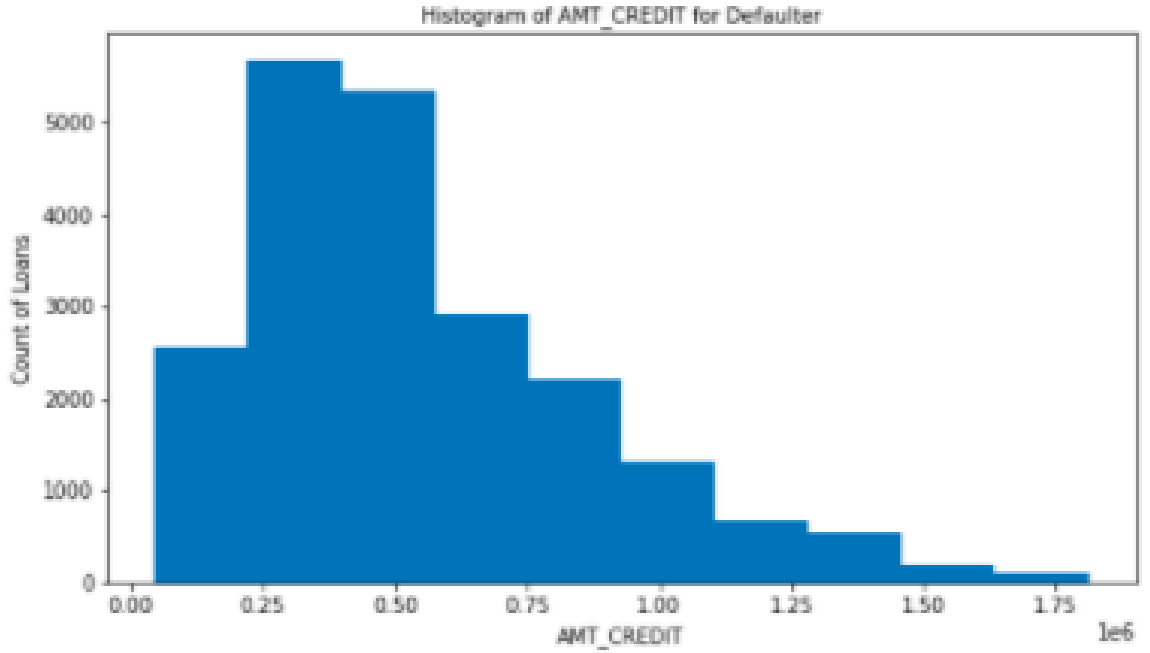




Considering 99% values to be outliers and removing it

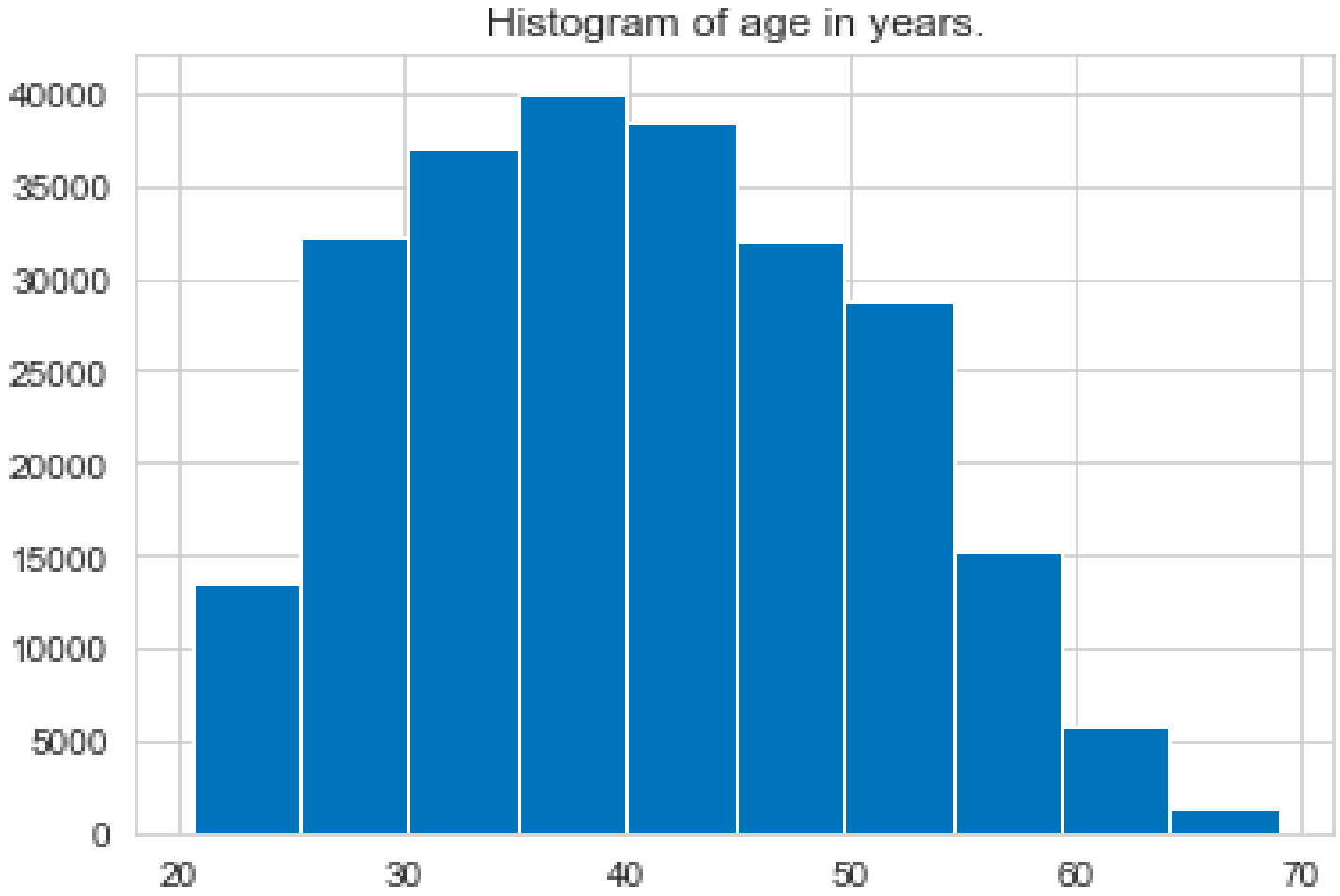
we observe that the max no of defaulters credit amount lies between 250000 to around 500000 for defaulters.





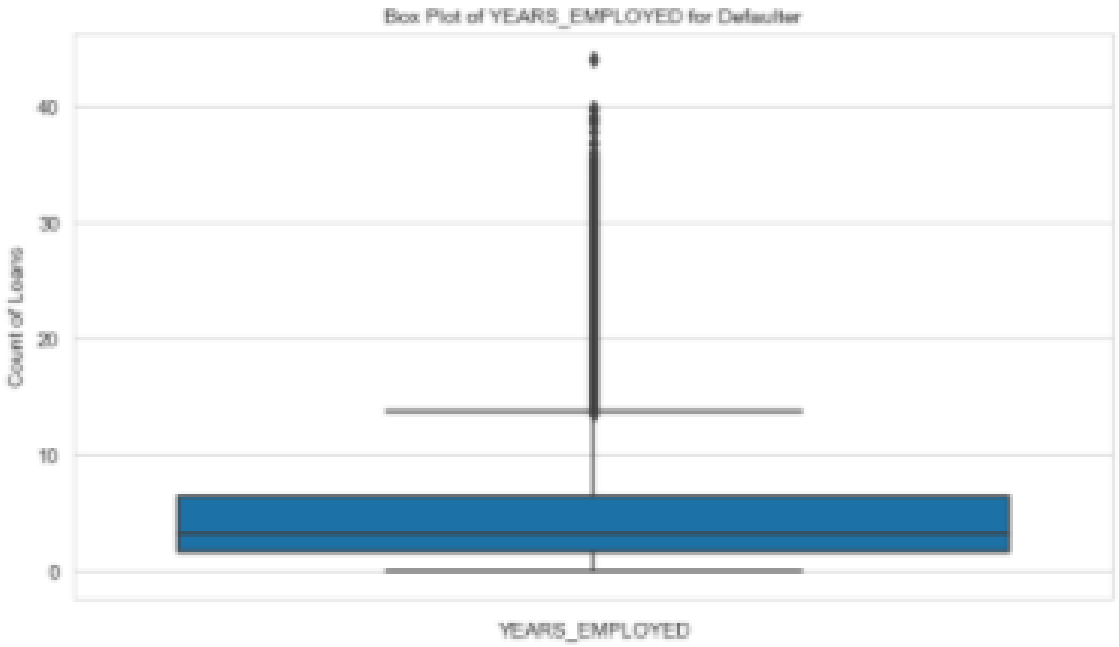
ANALYSIS ON AGE

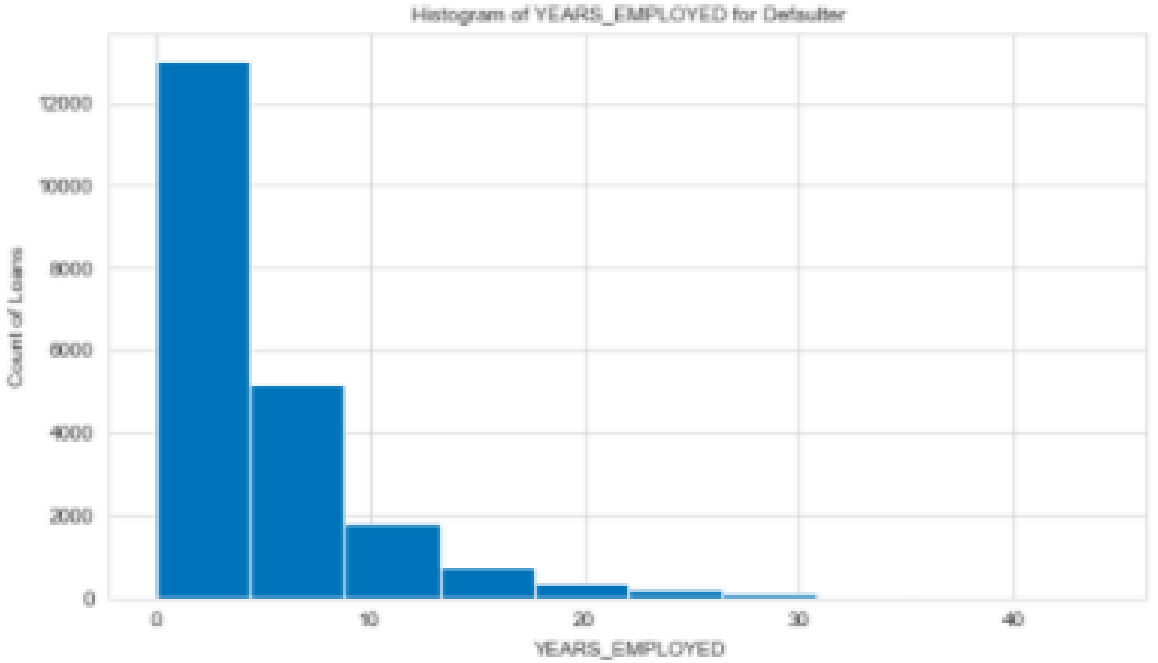
Age is by all accounts genuinely distributed with respect to loan count ranging highest count in range of 35-40 and then sort of dropping down as the age progressess



ANALYSIS OF DAYS EMPLOYED

First we Derived variable "Years Employed" from days employed and plotted graph for the same. We can see that huge no of defaulters of credit are individuals who don't work and individuals who are in beginning phases of there profession that is individuals who have experience going under 10 years





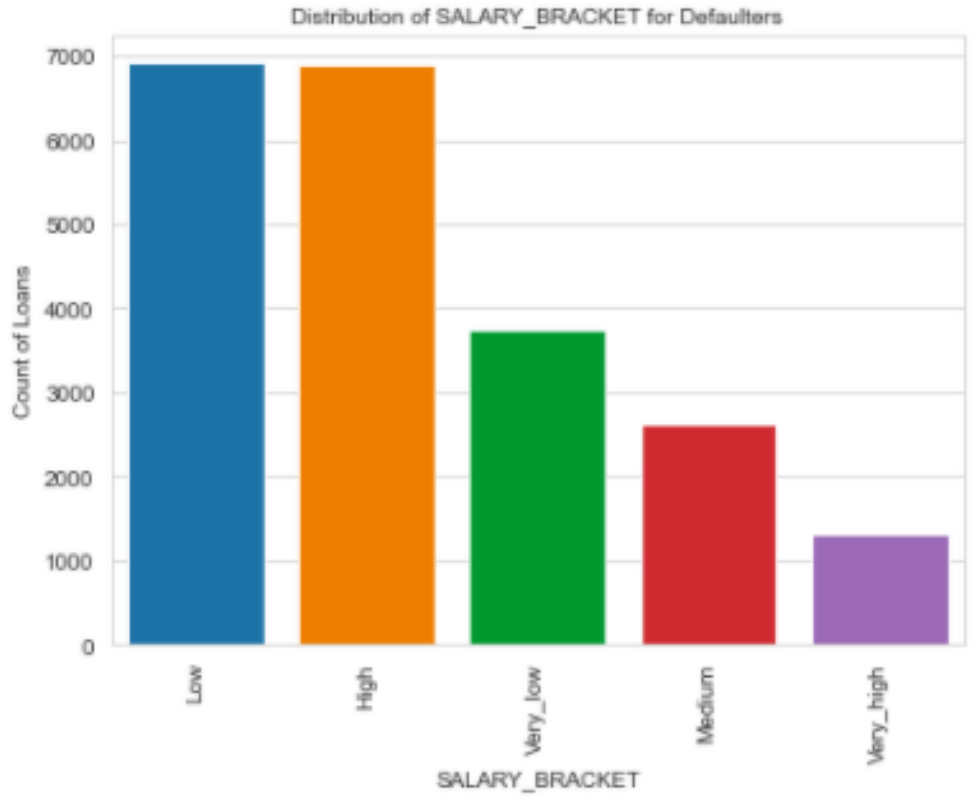
ANALYSIS ON NAME_INCOME_TYPE AND NAME_EDUCATION TYPE

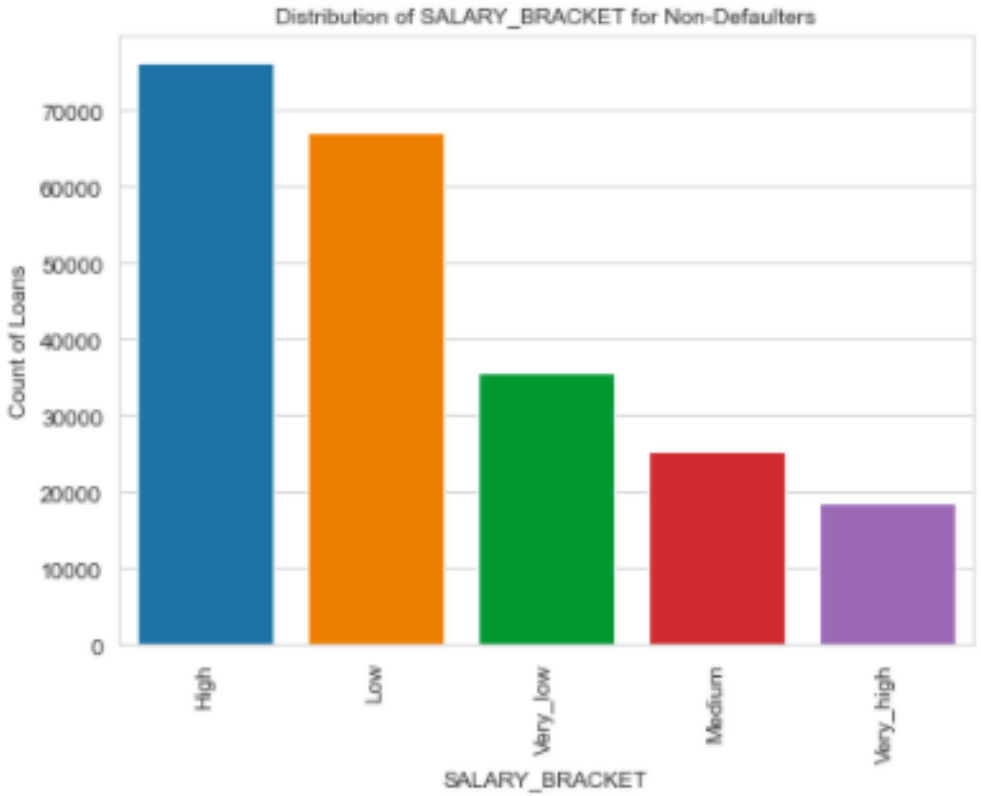
After doing analysis on NAME_INCOME_TYPE and NAME_EDUCATION_TYPE

- We found that non working people are generally student and pensioners having count of 17 and 10 respectively but main customers of availing loan are people who are working with count of 155766
- Most of the loans are taken by working people with secondary education having count of 116167.

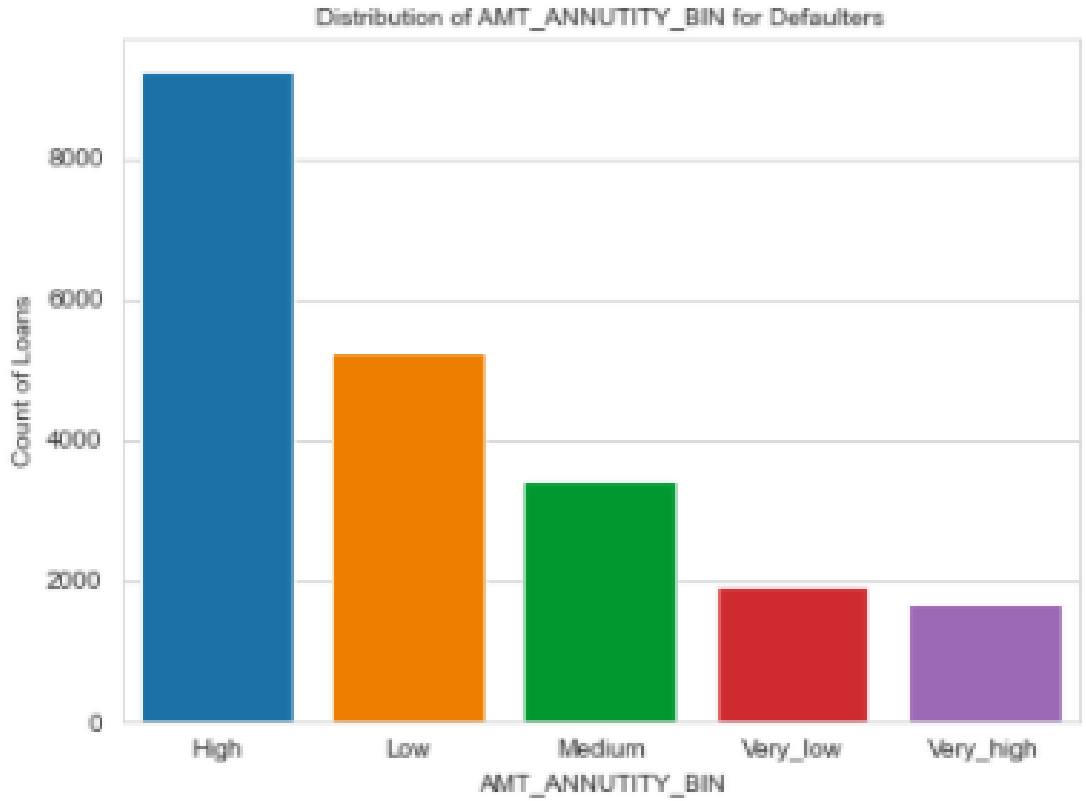
CONTINUOUS VARIABLES ANALYSIS BY BINNING

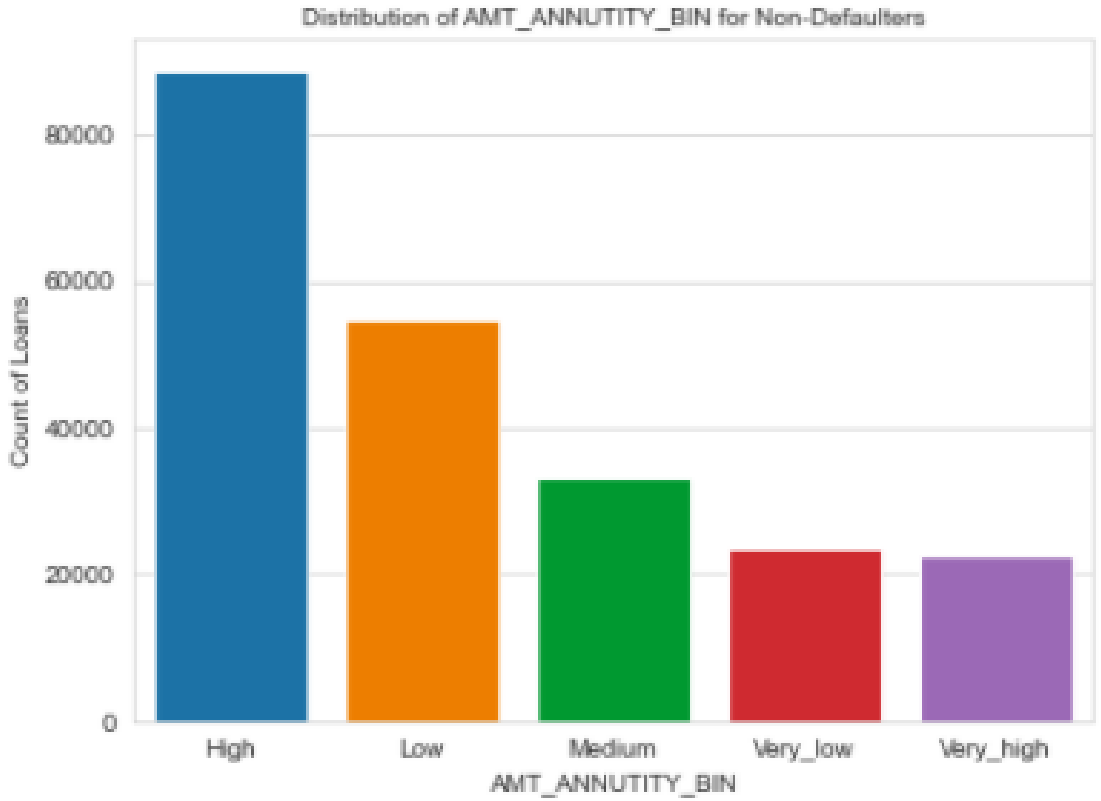
- Here we applied quantile-based discretization function on AMT_INCOME-TOTAL binning the quantile into 0, 10,3 5, 50 & 100 labelling 'Very_low', 'Low', 'Medium',' High', ' Very_high' respectively.
- Naming it "SALARY_BRACKET" and plotting graph of it we found that people having lower salary bracket tend to be defaulters and also are high loan counts and higher salary one are the ones who dont seem to be defaulters .For the rest scenarios are similar in both the TARGET case .





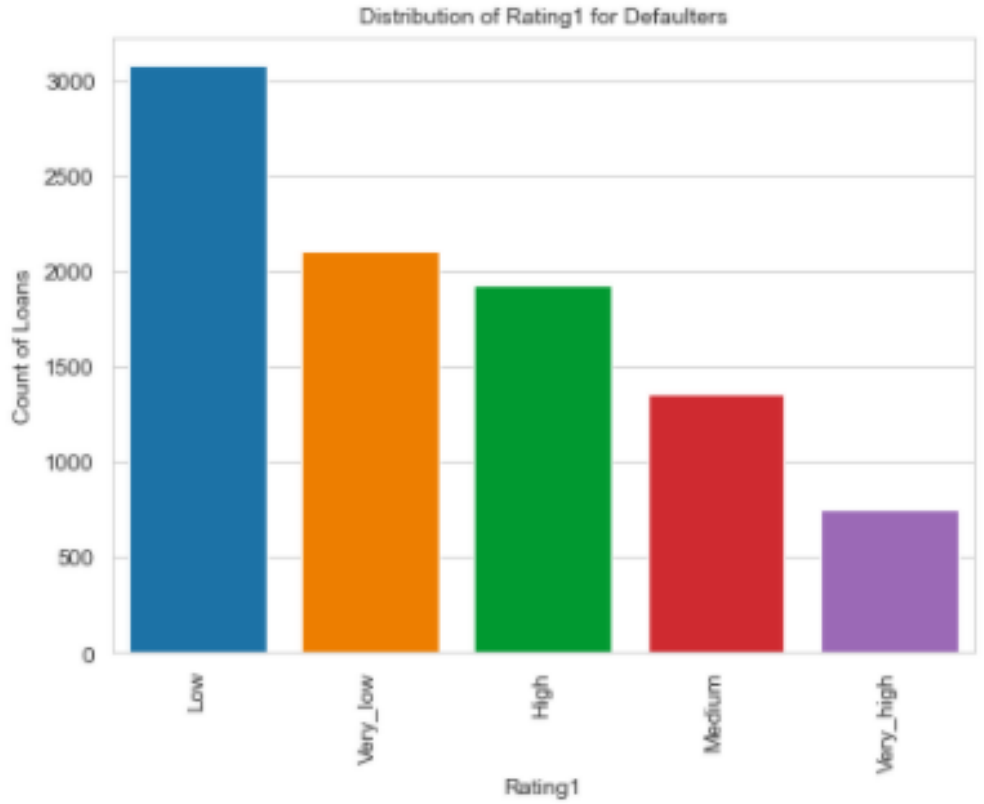
Similarly, applying qcut function in AMT_ANNUTITY_BIN, it is clear that maxinum number of defaulters have Low_annuity Values, while maximum number of non-defaulters have high annuity

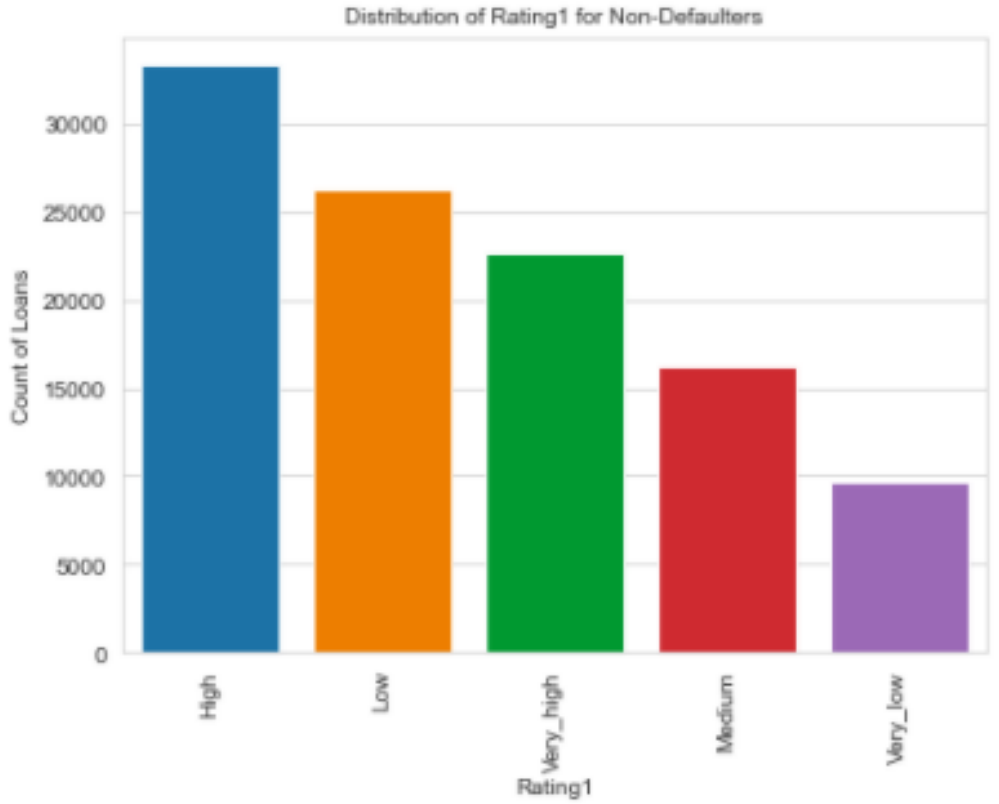




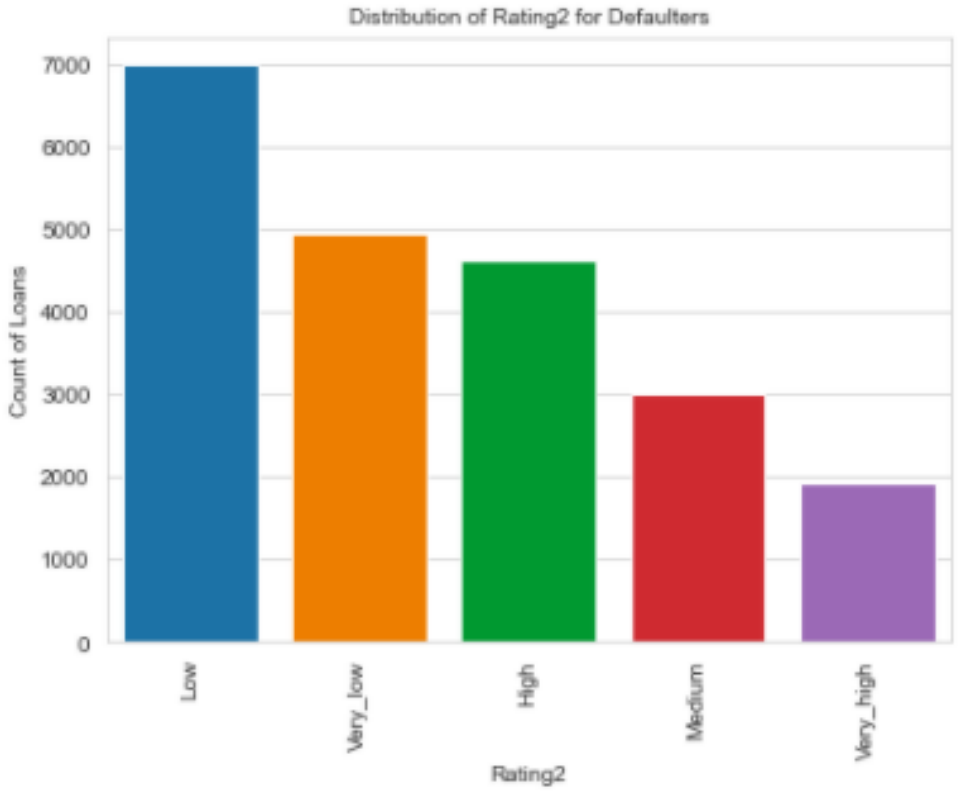
DISTRIBUTION OF EXT_SOURCE_1 & EXT_SOURCE_2

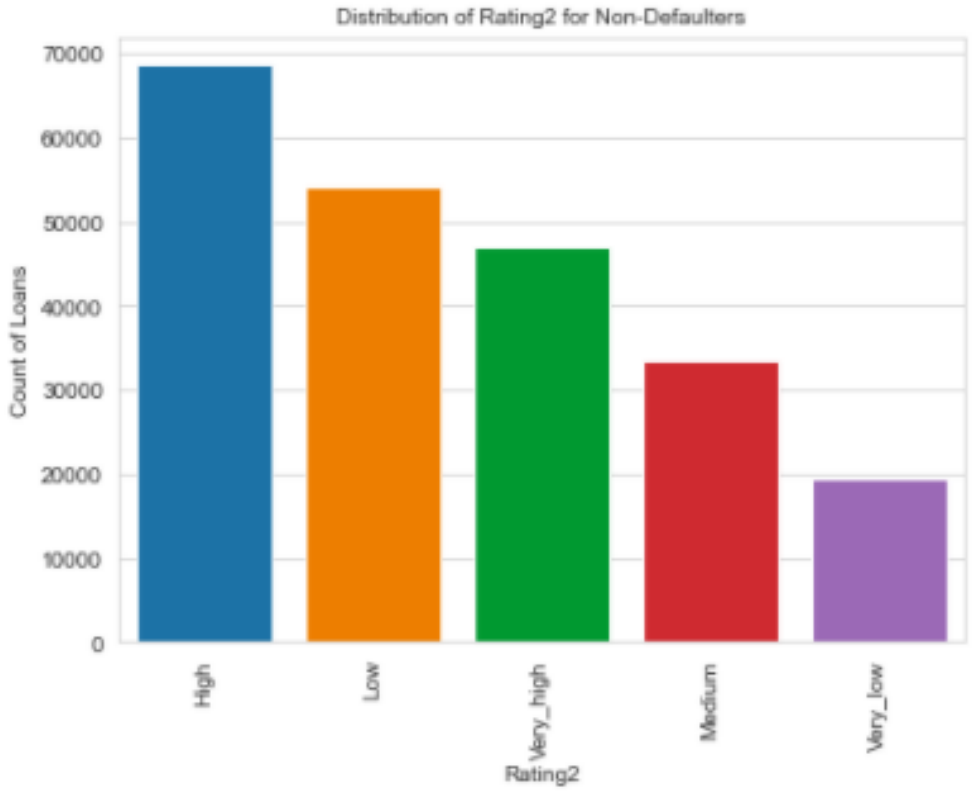
- Here we applied quantile-based discretization function on EXT_SOURCE_1 binning the quantile into 0, 10,3 5, 50 & 100 labelling 'Very_low', 'Low', 'Medium',' High', ' Very_high' respectively.
- Naming it "Rating 1" and plotting graph of it we found that a large number of defaulters have very Low rating, while a large number of non-defaulters have a high rating.





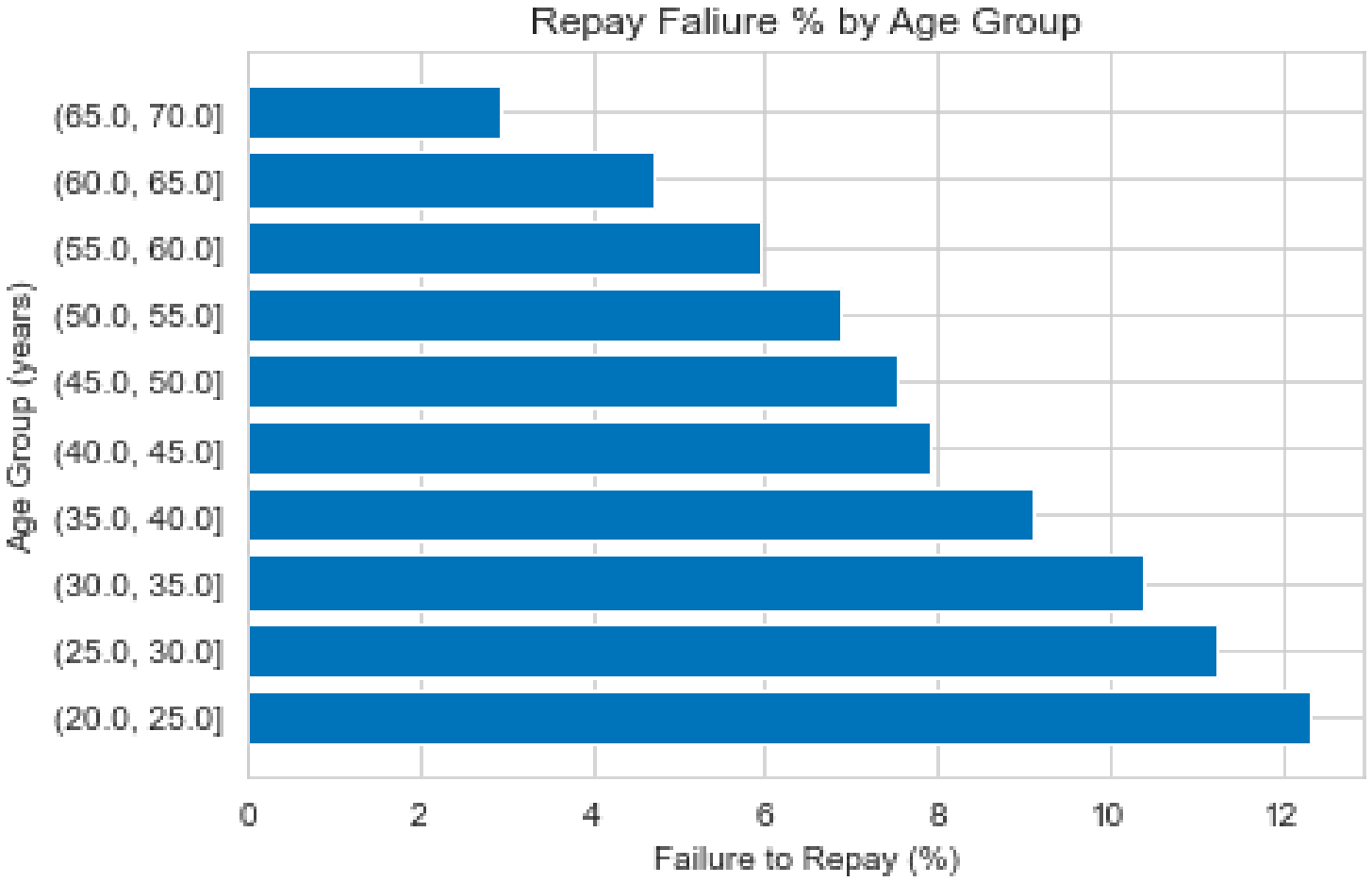
Similarly, applying qcut function in EXT_SOURCE_2 , plotting graph of it shows that large number of defaulters have very Low rating, while a large number of non-defaulters have a high rating.





Analysing what is the impact of age in terms of people taking loans by binning the age. Graph the age bins and the average of the target as a bar plot:

- We can see that maximum failure to repay is from age group 20 to 25 and majority is lying in 20 to 40 years of range



BIVARIATE ANALYSIS

After analysing the columns excel provided to us selecting the below columns for our bivariate Analysis

['EXT_SOURCE_1', 'EXT_SOURCE_3', 'EXT_SOURCE_2','AMT_GOODS_PRICE', 'AMT_ANNUITY', 'CNT_FAM_MEMBERS',' DAYS_LAST_PHONE_CHANGE', 'AMT_CREDIT', 'AMT_INCOME_TOTAL', 'DAYS_REGISTRATION', 'REGION_POPULATION_RELATIVE', 'CNT_CHILDREN', 'HOUR_APPR_PROCESS_START','REGION_RATING_CLIENT_W_CITY', 'REGION_RATING_CLIENT', 'DAYS_ID_PUBLISH', 'DAYS_EMPLOYED', 'DAYS_BIRTH']

---

first finding out the correlation between different attributes for the defaulters ussing corr() function, we get the plot shown below-

CORRELATION BETWEEN DIFFERENT ATTRIBUTES FOR DEFAULTER

We can make the below inferences for correlation between the different variables :

5 most positive correlations

AMT_CREDIT - AMT_GOODS_PRICE

REGION_RATING_CLIENT_W_CITY - REGION_RATING_CLIENT

CNT_CHILDREN - CNT_FAM_MEMBERS

AMT_CREDIT - AMT_ANNUITY

AMT_GOODS_PRICE - AMT_ANNUITY

5 most negative correlations

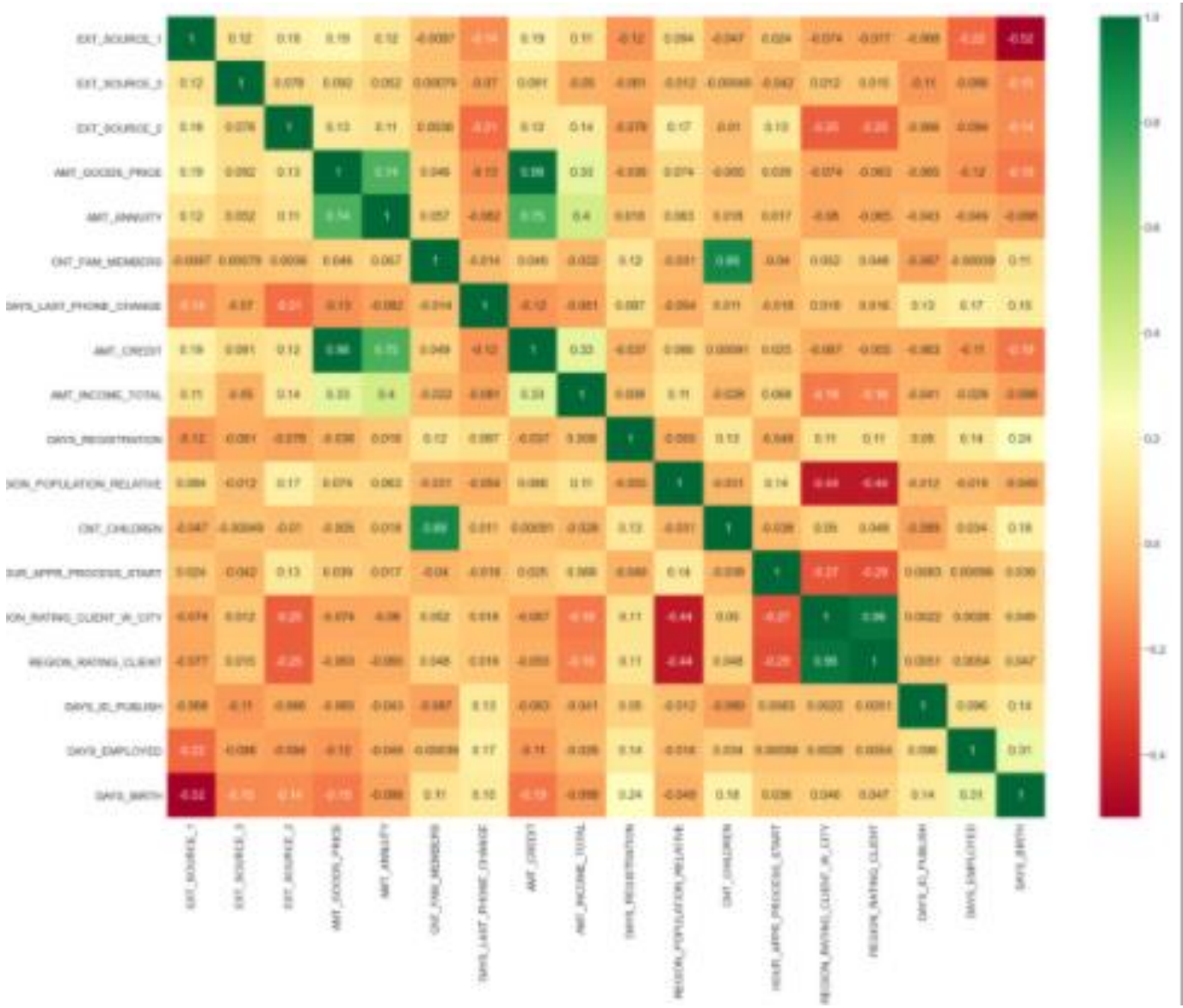
HOUR_APPR_PROCESS_START - REGION_RATING_CLIENT_W_CITY

REGION_RATING_CLIENT - HOUR_APPR_PROCESS_START

REGION_POPULATION_RELATIVE - REGION_RATING_CLIENT

REGION_RATING_CLIENT_W_CITY - REGION_POPULATION_RELATIVE

EXT_SOURCE_1 - DAYS_BIRTH



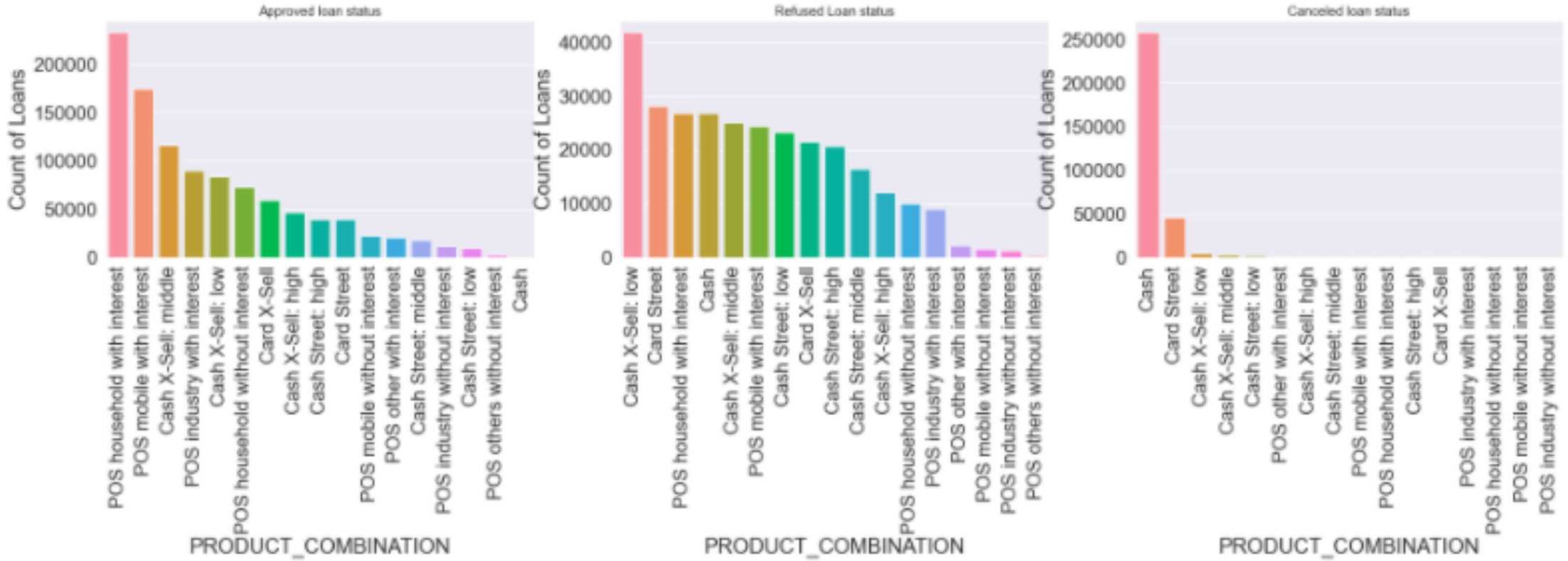
ANALYSIS OF PREVIOS APPLICATION DATASET

- The Previous application dataset has 1670214 rows and 37 columns where 16 object type, 15 are float64 type and 6 are int64 type columns are present.
- Now analysing for null values , we identified that there are columns having high null values ,we need to handle them efficently so that it does not impact our analysis . RATE_INTEREST_PRIVILEGED and RATE_INTEREST_PRIMARY having 99% data as null which is not helpful to us thus we will discard the rows having null above 55% and continue our anaylsis.

- Checking the percentage of different loan status for previous application data we found that loan approval percent is around 62% cancelled and refused add up the remaining and unused offer is negligible.

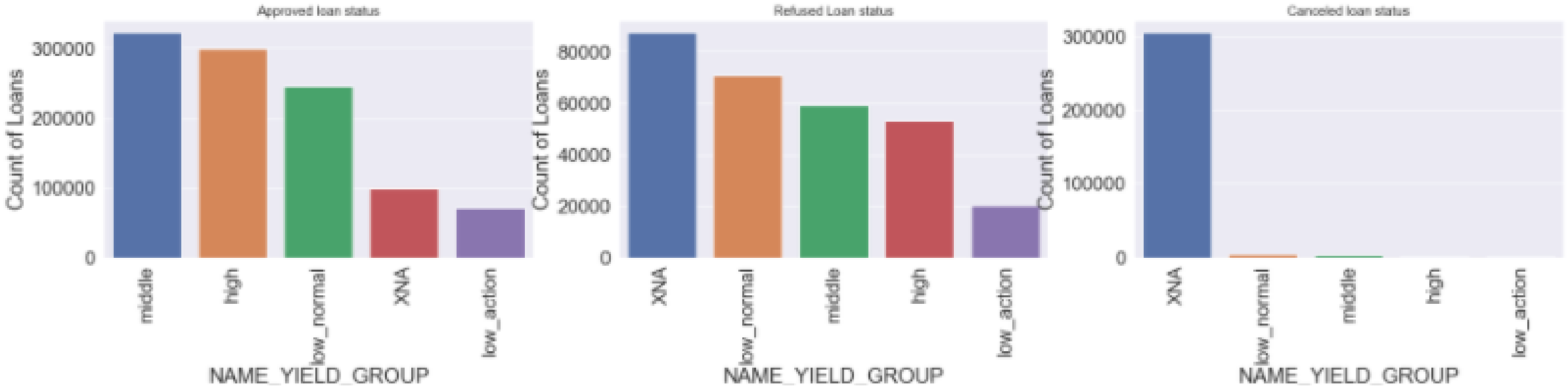
DISTRIBUTION OF PRODUCT_COMBINATION

We can see from above most approved yield group is middle and most refused and cancelled are from XNA or the unidentified group and low-normal yield followed by middle



DISTRIBUTION OF NAME_YIELD_GROUP

We can see from above most approved yield group is middle and most refused and cancelled are from XNA or the unidentified group and low-normal yield followed by middle



BIVARIATE ANALYSIS OF VARIABLES

In this analysis we will find out the correlation of different attributes for approved and refused status loans .

CORRELATION FOR APPROVED

For approved status top 5 attributes having positive correlation are :

1.)Amt_CREDIT with AMT_APPLICATION

2.)AMT_CREDIT with AMT_GOODS_PRICE

3.)AMT_APPLICATION with AMT_ANNUITY

4.)DAYS_TERMINATION with DAYS_LAST_DUE

5.)AMT_DOWN_PAYMENT with RATE_DOWN_PAYMENT

For approved status top 5 attributes having negative correlation are :

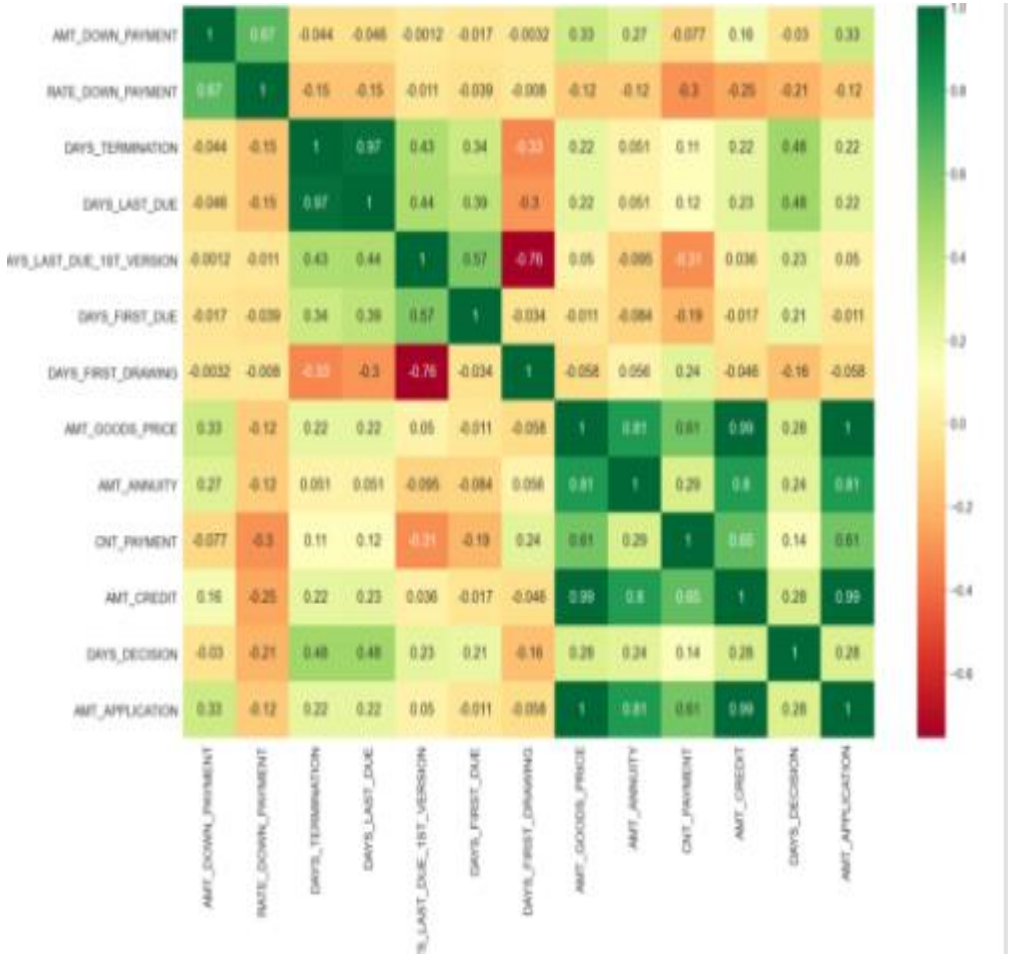
1.) DAYS_FIRST_DRAWING with DAYS_LAST_DUE_1ST_VERSION

2.) DAYS_FIRST_DRAWING with DAYS_TERMINATION

3.) DAYS_FIRST_DRAWING with DAYS_LAST_DUE

4.) CNT_PAYMENT with DAYS_LAST_DUE_1ST_VERSION

5.) CNT_PAYMENT with RATE_DOWN_PAYMENT



CORRELATION FOR REFUSED

For refused status top 3 attributes having positive correlation are :

1.)AMT_APPLICATION with AMT_GOODS_PRICE

2.) AMT_APPLICATION with AMT_CREDIT

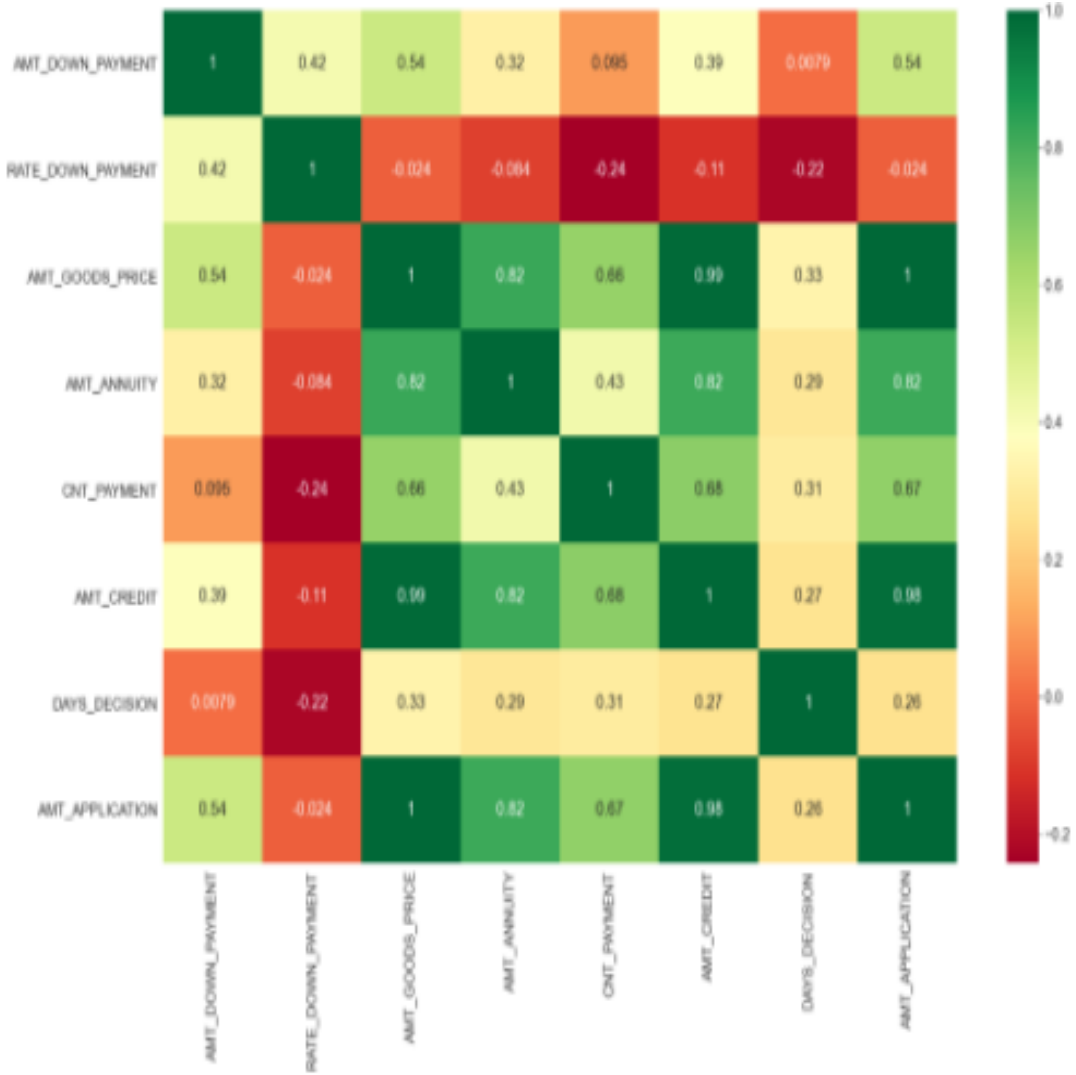
3.)AMT_APPLICATION with AMT_ANNUITY

For refused status top 3 attributes having negative correlation are :

1.) CNT_PAYMENT with RATE_DOWN_PAYMENT

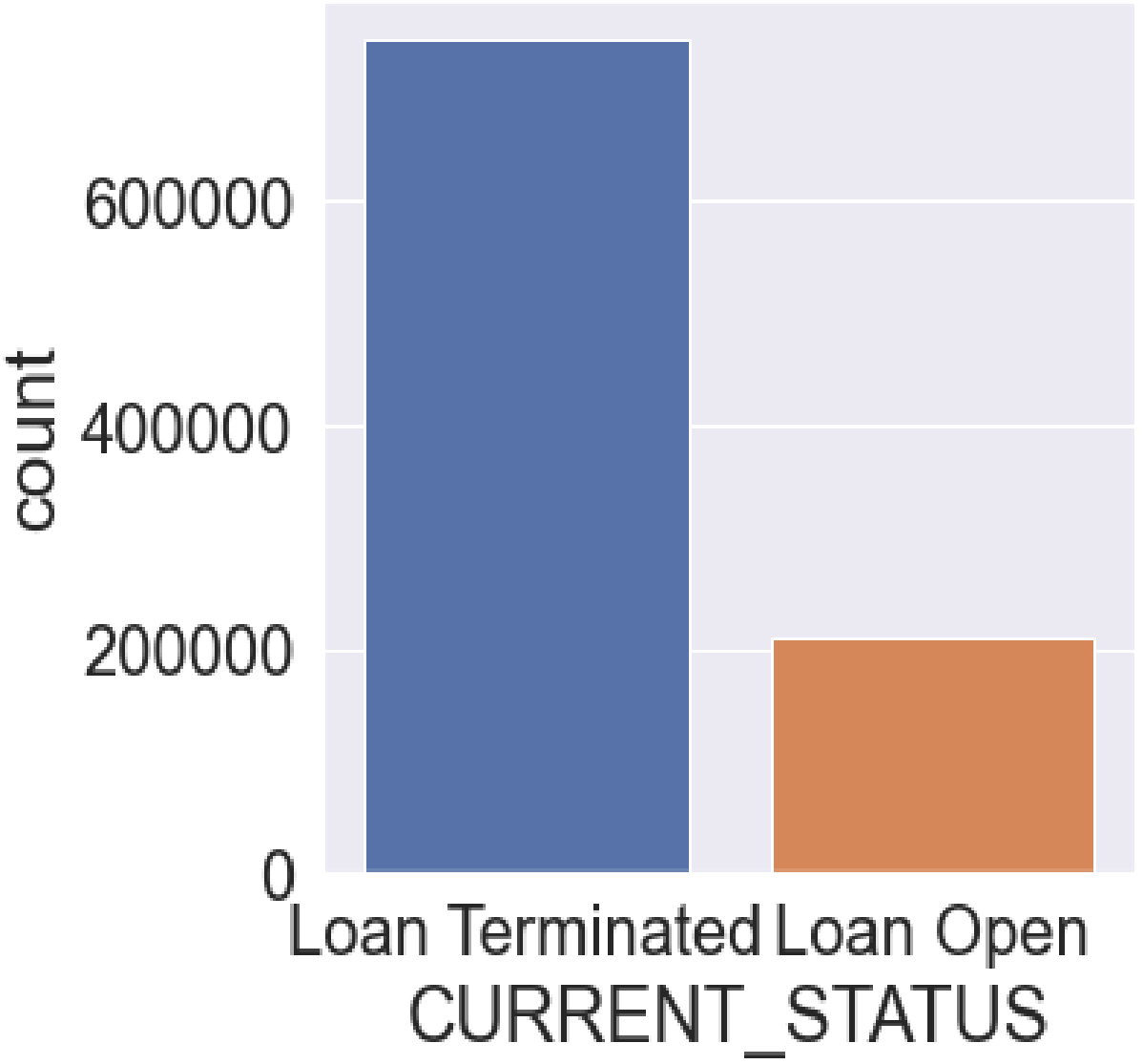
2.) DAYS_DECISION with RATE_DOWN_PAYMENT

3.) AMT_CREDIT with RATE_DOWN_PAYMENT



DISTRIBUTION OF APPROVED DATA BASED ON LOAN TERMINATED AND LOAN OPEN

The plot on right shows that high count for loan terminated compared to the loan which are open



CONCLUSION

- Bank should focus on people having Academic degree. They are potential customers.
- Get as much as clients from housing type 'With parents' as they are having least number of unsuccessful payments.
- Banks should focus less on income type 'Working' as they are having most number of unsuccessful payments.
- Target should focus from age group above 40 as the % failure of repay is lower for them.