Topic : EDA ASSIGNMENT

Steps: Importing Libraries

- First step is importing libraries which are required

Reading Data

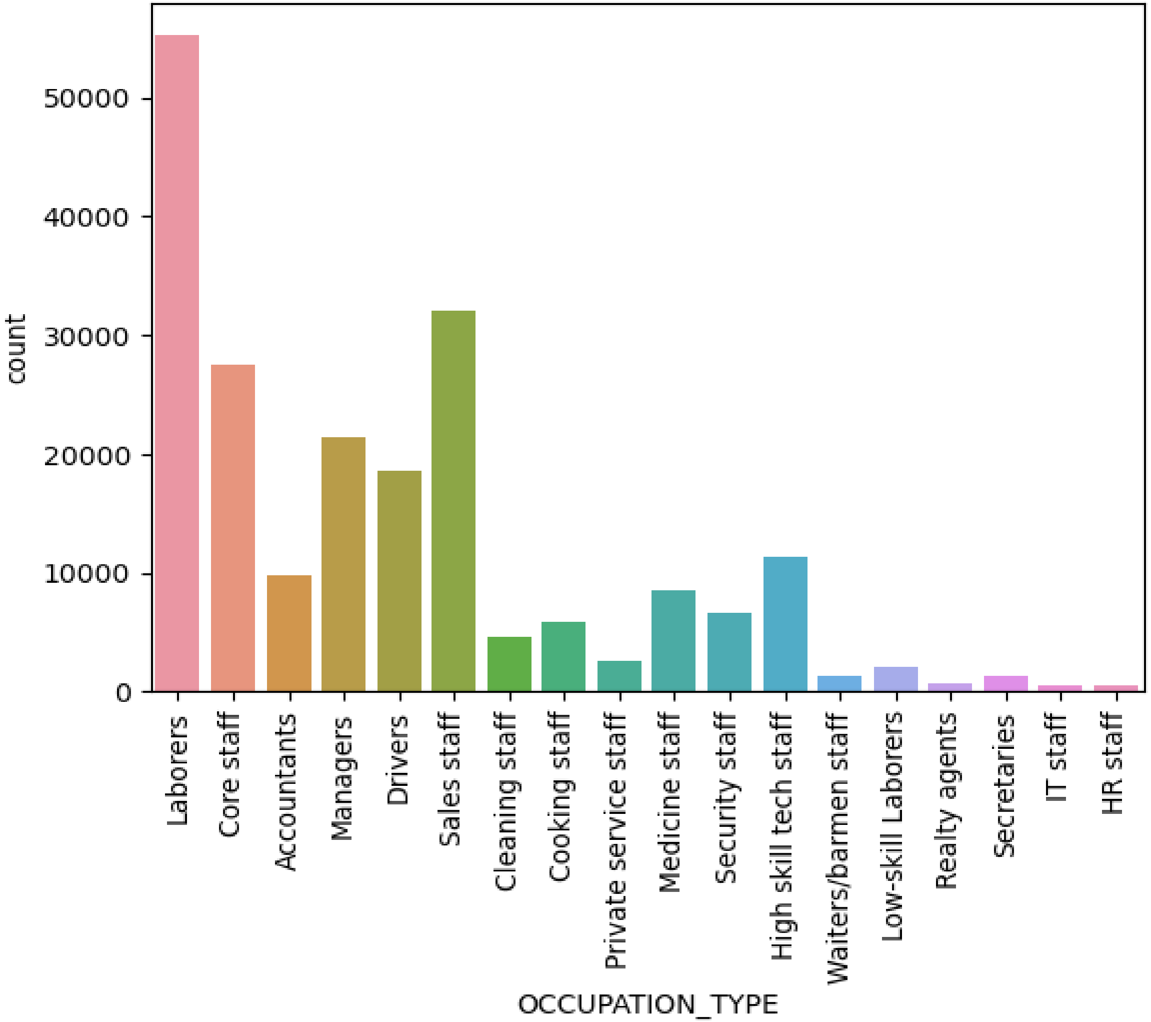
- Application Data and Previous Application is loaded

Checking Data shape and values

- Checking shape and values of Application data.csv file.

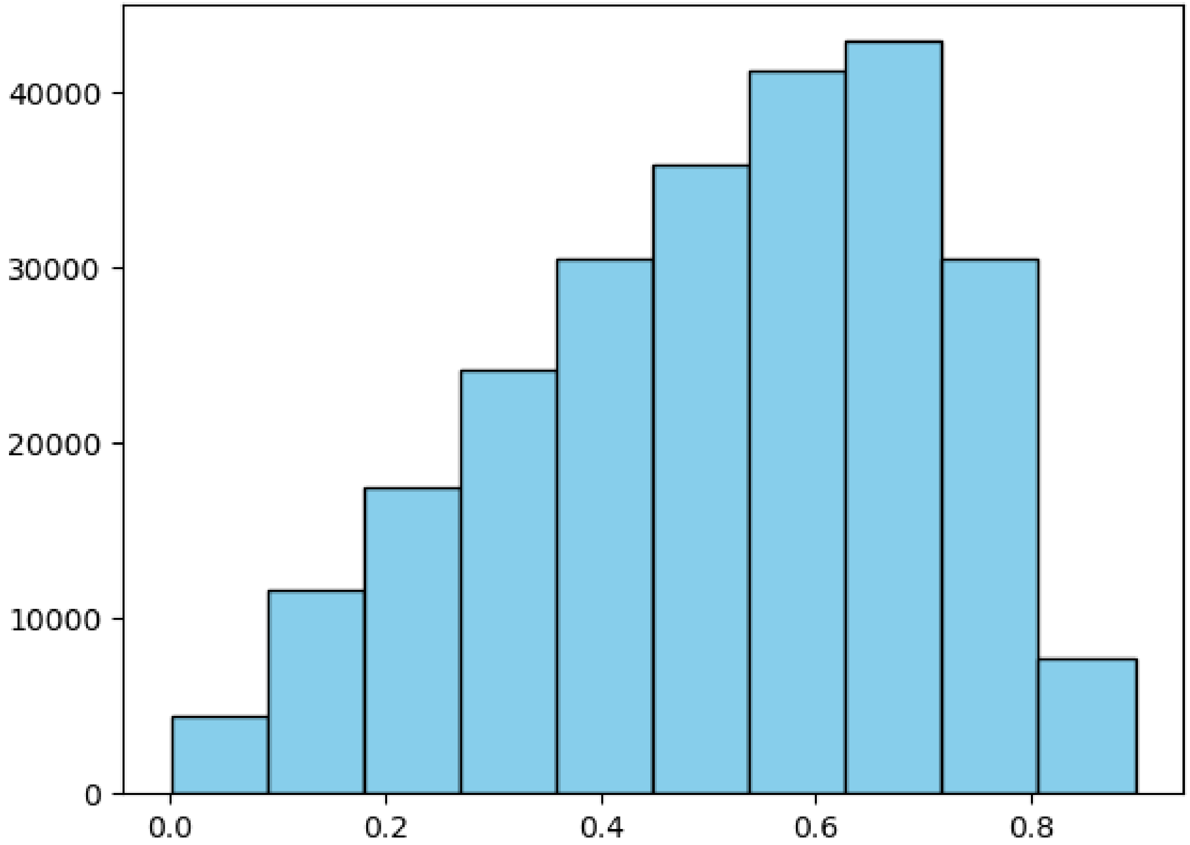
Treatment of missing values

- Columns with missing values less than 45% are taken for treatment.
- Occupation_type has the highest number of missing values under 45% which is 96391.
- Mode of Occupation Type is ‘Labourers’. Missing values cant be replace by mode as same occupation of other customers cant be considered.



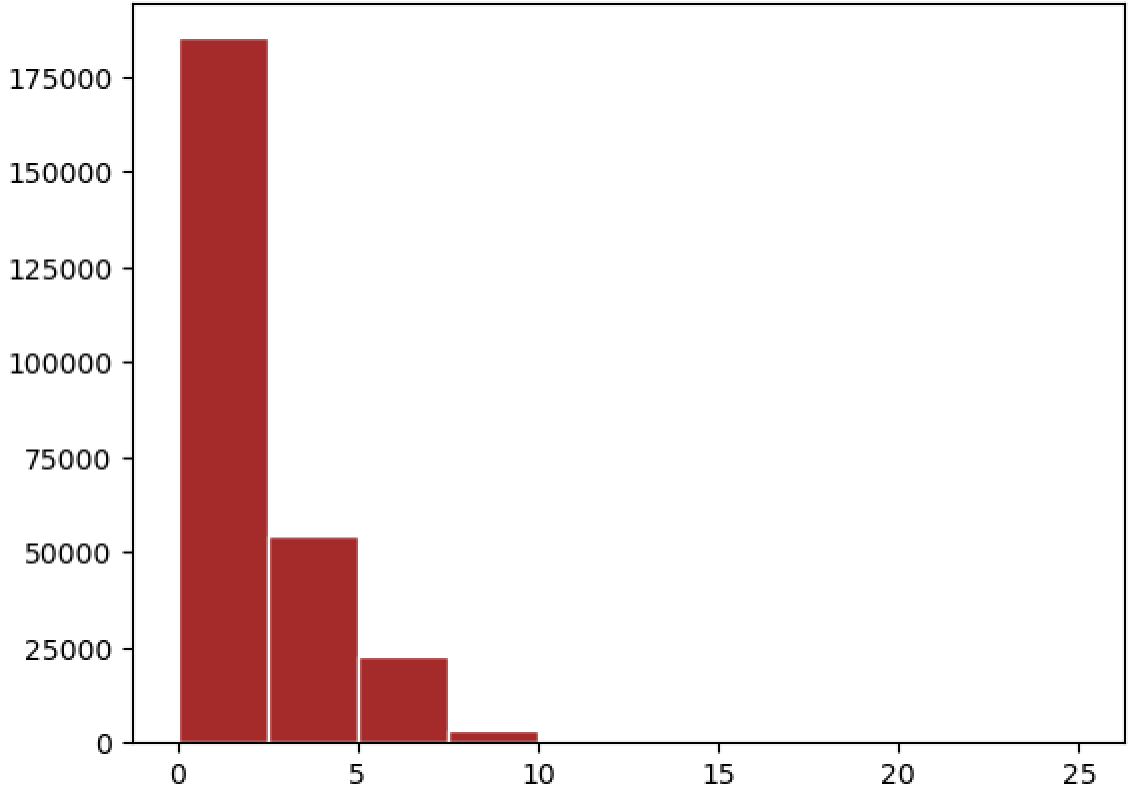
Column EXT_SOURCE_3

- It can be seen from the histogram that in column EXT_SOURCE_3 maximum number of values are between 0.5 to 0.75.
- Replacing values with median as values are skewed left.



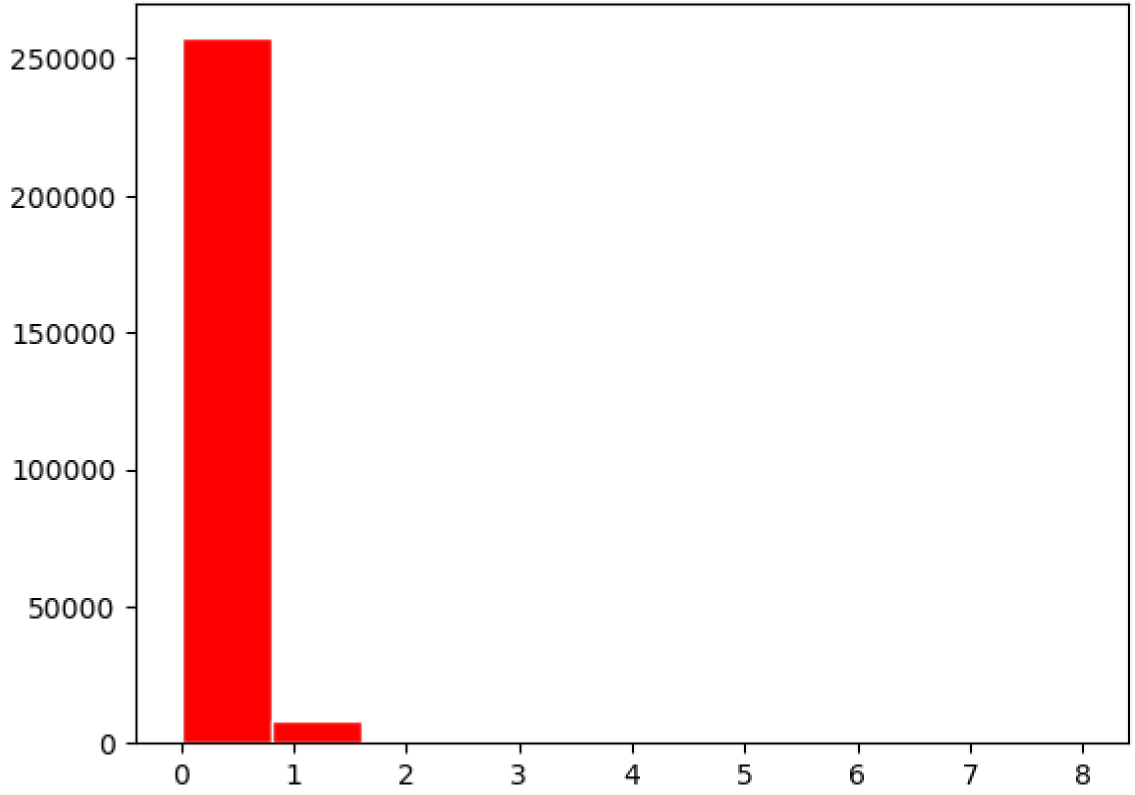
Column 'AMT_REQ_CREDIT_BUREAU_YEAR'

- It can be seen that maximum number of values in column 'AMT_REQ_CREDIT_BUREAU_YEAR' are between 0.0 and 3.0.
- Replacing values with **mode** as values are **discrete.**



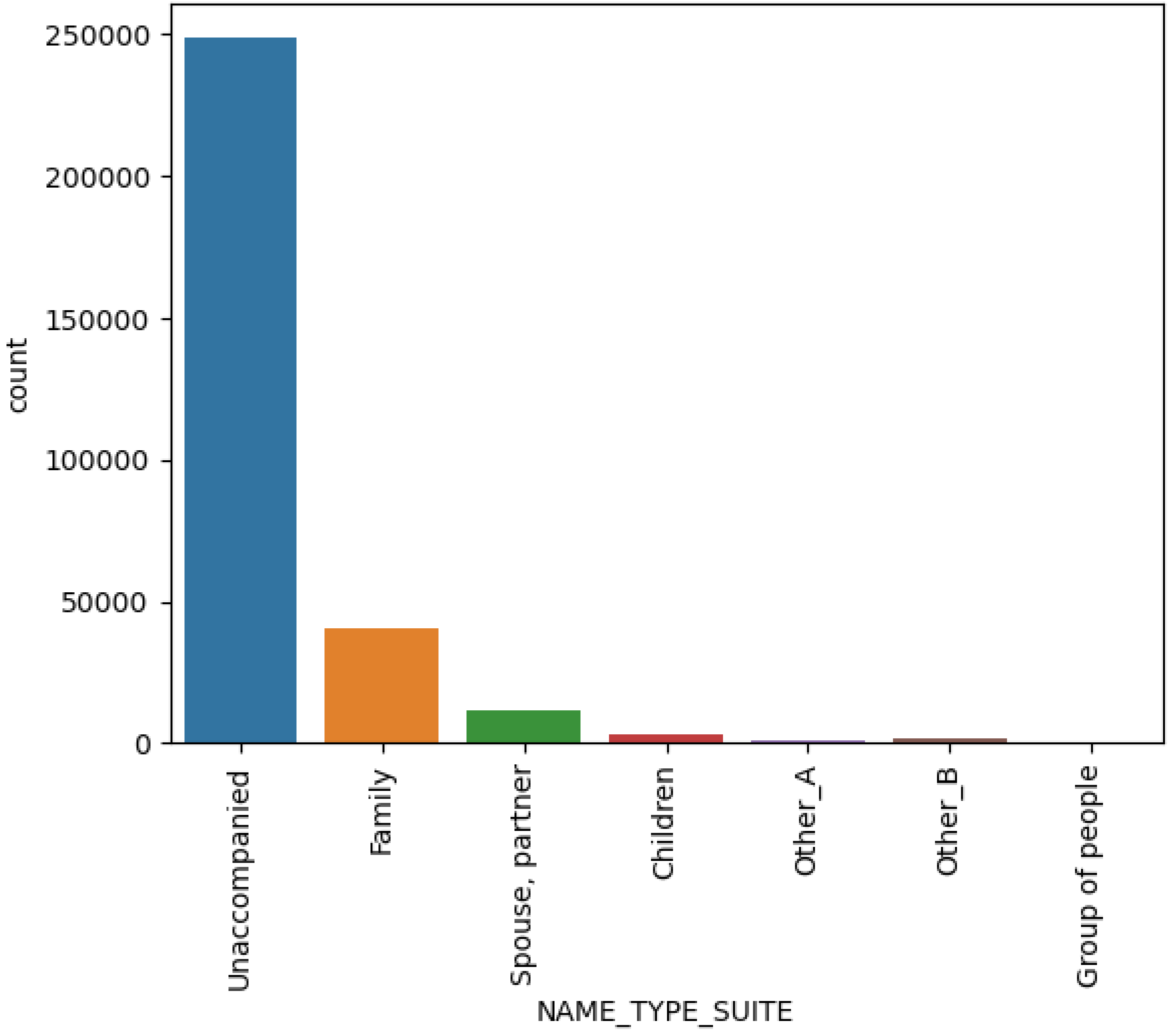
Column 'AMT_REQ_CREDIT_BUREAU_WEEK'

- As Column has only almost single value and its 0.0, so it cant give us variable insights. Its better to drop this column.



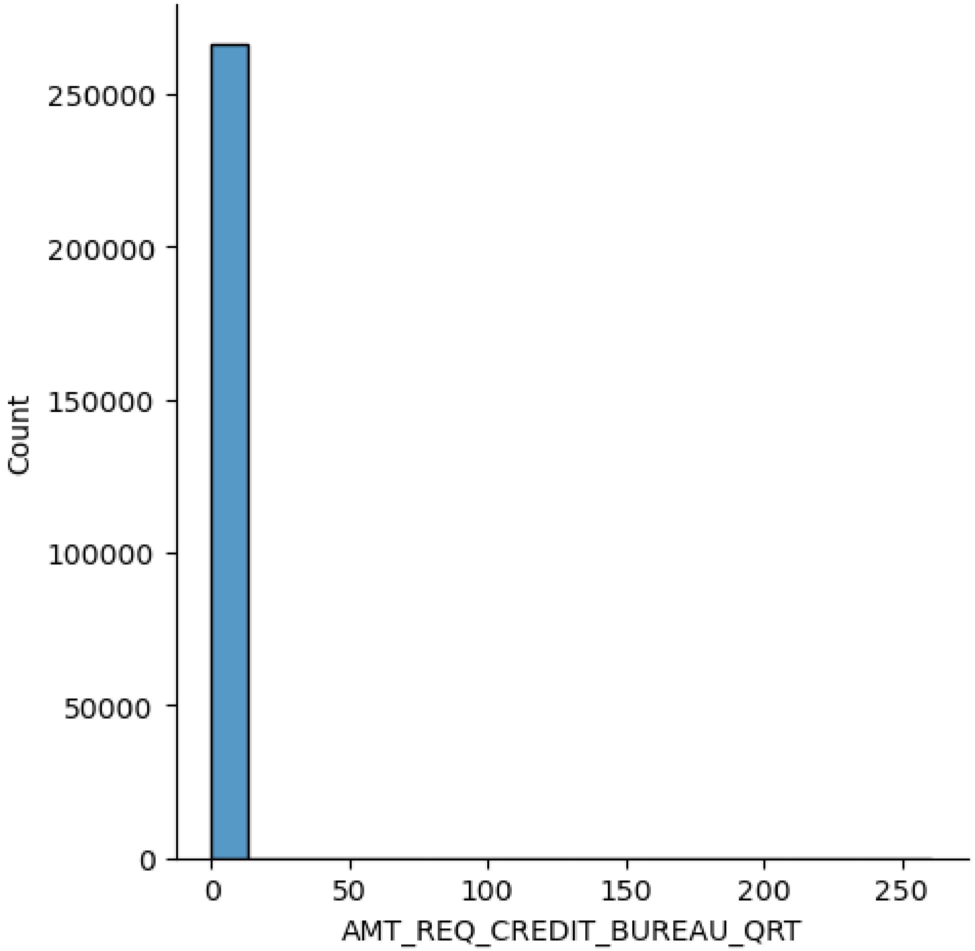
Column 'NAME_TYPE_SUITE'

- As Column has maximum number of **unaccompanied** values, we can not show the correlation of this column, as it does not show any insights.

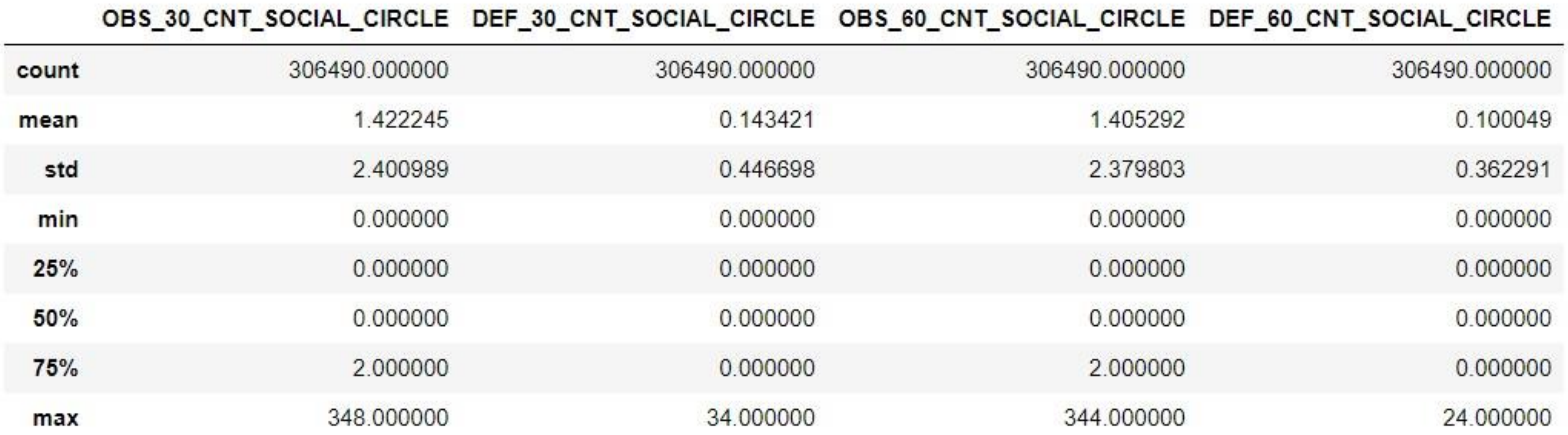


Column 'AMT_REQ_CREDIT_BUREAU_QRT'

- As Column has only almost single value and its 0.0, so it cant give us variable insights. Its better to drop this column
- Similarly with the columns 'AMT_REQ_CREDIT_BUREAU_MON', 'AMT_REQ_CREDIT_BUREAU_DAY','AMT_REQ_CREDIT_BUREAU_HOUR'



Column 'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE’, 'OBS_60_CNT_SOCIAL_CIRCLE’, 'DEF_60_CNT_SOCIAL_CIRCLE'

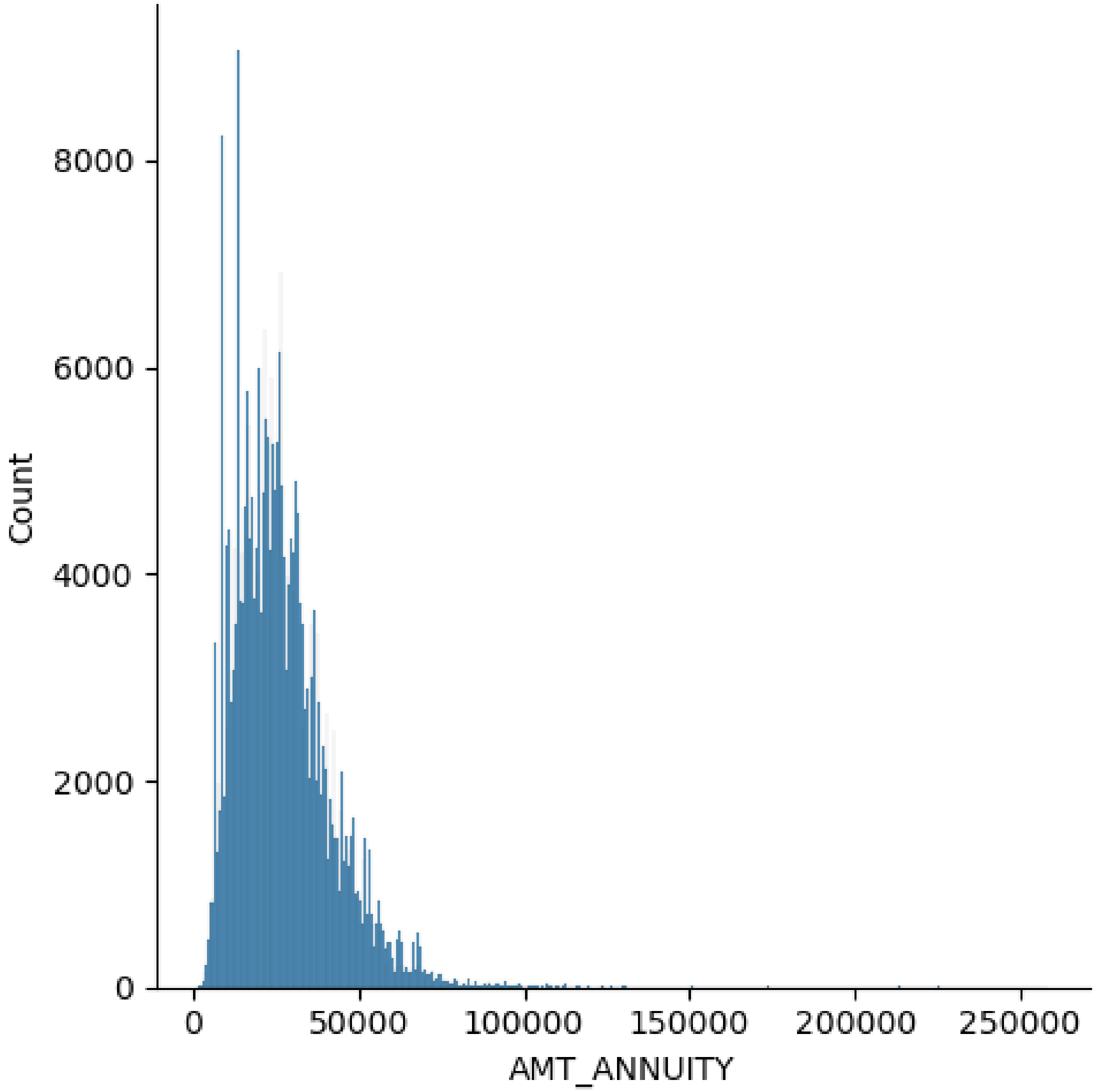


	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE
count	306490.000000	306490.000000	306490.000000	306490.000000
mean	1.422245	0.143421	1.405292	0.100049
std	2.400989	0.446698	2.379803	0.362291
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000
75%	2.000000	0.000000	2.000000	0.000000
max	348.000000	34.000000	344.000000	24.000000

- As all four Column has only almost single value and its 0.0, so it cant give us variable insights. Its better to drop this column

Column 'AMT_ANNUITY'

- Replacing values with median as values are skewed right.



Columns ‘EXT_SOURCE_2’, ‘AMT_GOODS_PRICE’, ‘CNT_FAM_MEMBERS’, ‘DAYS_LAST_PHONE_CHANGE’

- Now the remaining columns are ‘EXT_SOURCE_2’, ‘AMT_GOODS_PRICE’, ‘CNT_FAM_MEMBERS’ and ‘DAYS_LAST_PHONE_CHANGE’
- These columns have less than 1% null values, replacing null values with mean of present values in these columns.

Now we can check that Data Cleaning Process is done and columns don’t have any nulls including Categorical and Numerical columns.

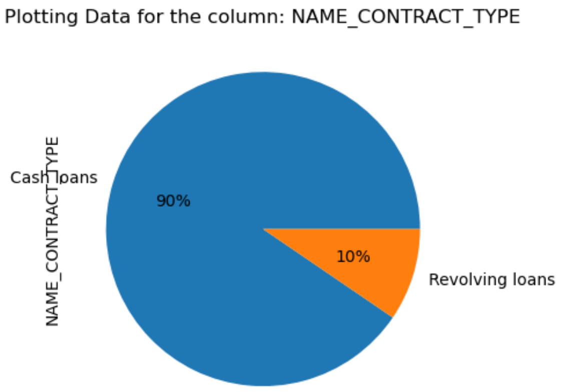
Next step is to check Correlation of all columns with Target Column.

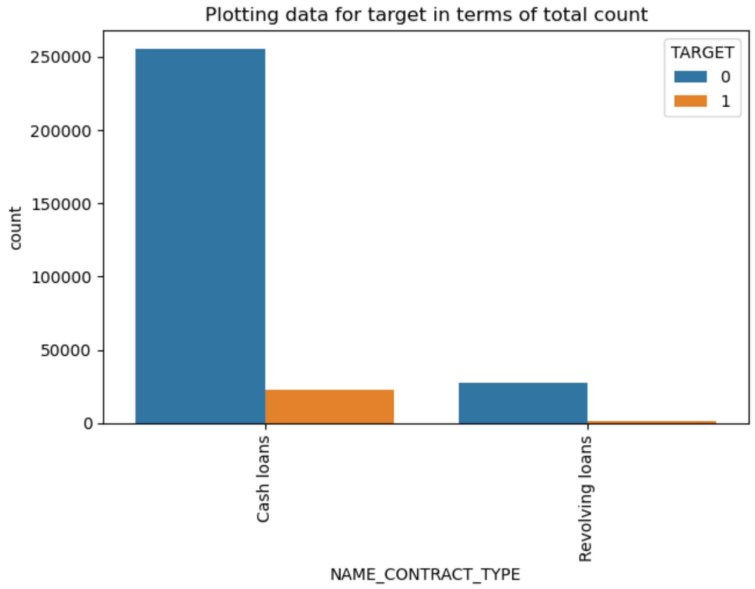
First checking correlation with 12 Categorical Columns.

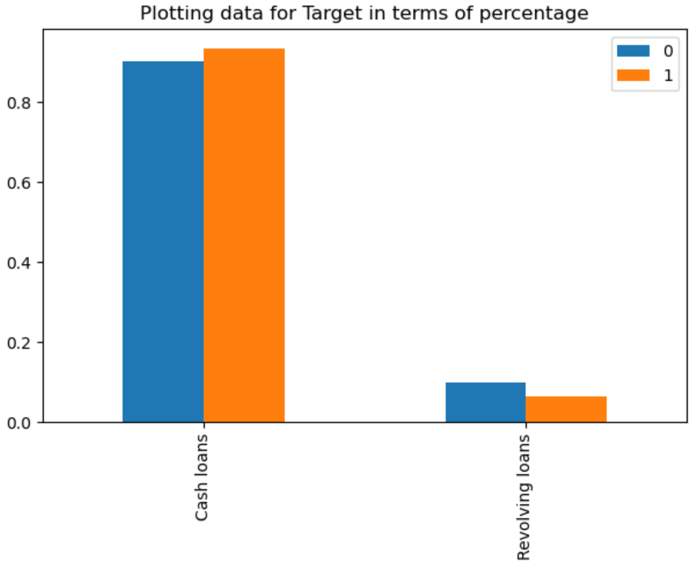
- Plotting ‘Pie chart’, ‘Count plot’ and ‘bar plot in terms of Percentage’ of all 12 Categorical Columns.
- 3 Figures of Each Categorical Column that is 36 figures are plotted.
- Each of 12 Categorical Column plots are given from next page.

Column: NAME_CONTRACT_TYPE

- It can be seen from pie chart that 90% are Cash Loans and only 10% are Revolving Loans.
- It can be seen from the count plot that compare to customers who are NPA (Non-Performing Assets), customers who have paid loan amount on time is very less.

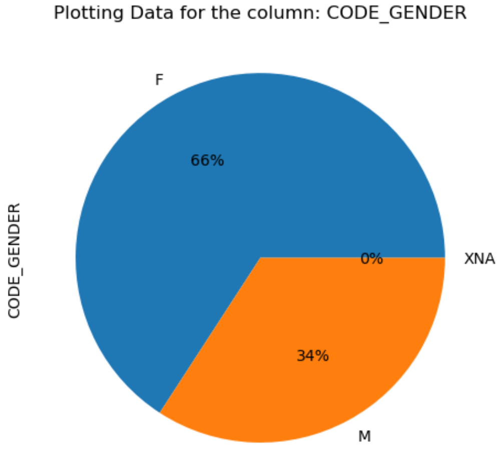


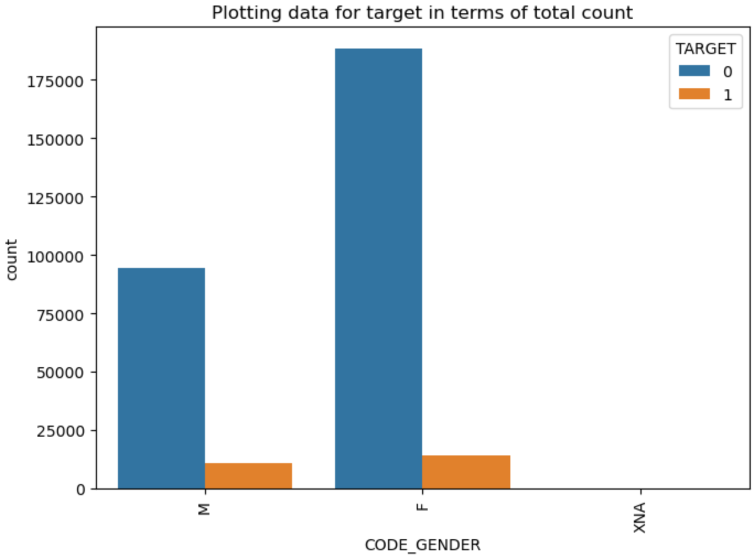


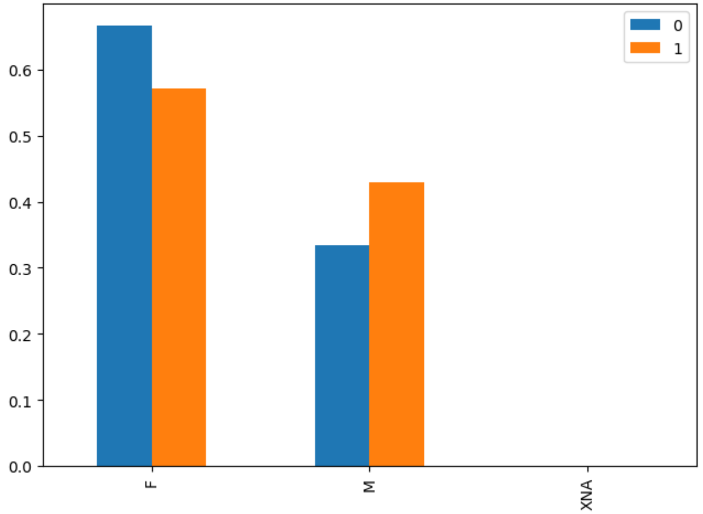


Column: CODE_GENDER

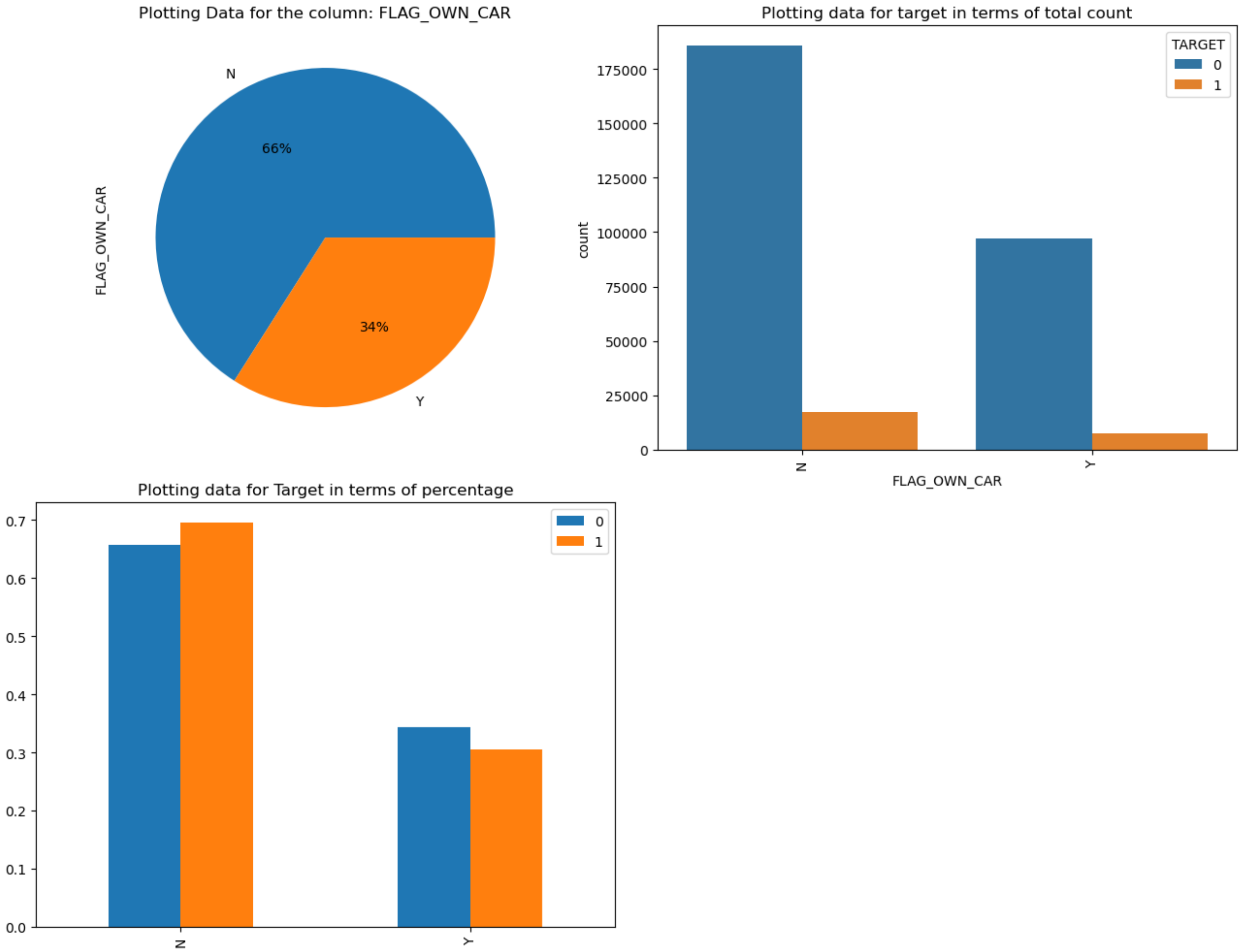
- It can be seen from pie chart that 66% of customers are Female while only 34% are Male (customers who have taken loan).
- It can be seen from the count plot that in both gender of customers NPA (Non-Performing Assets) are very high than customers who have paid loan amount on time.

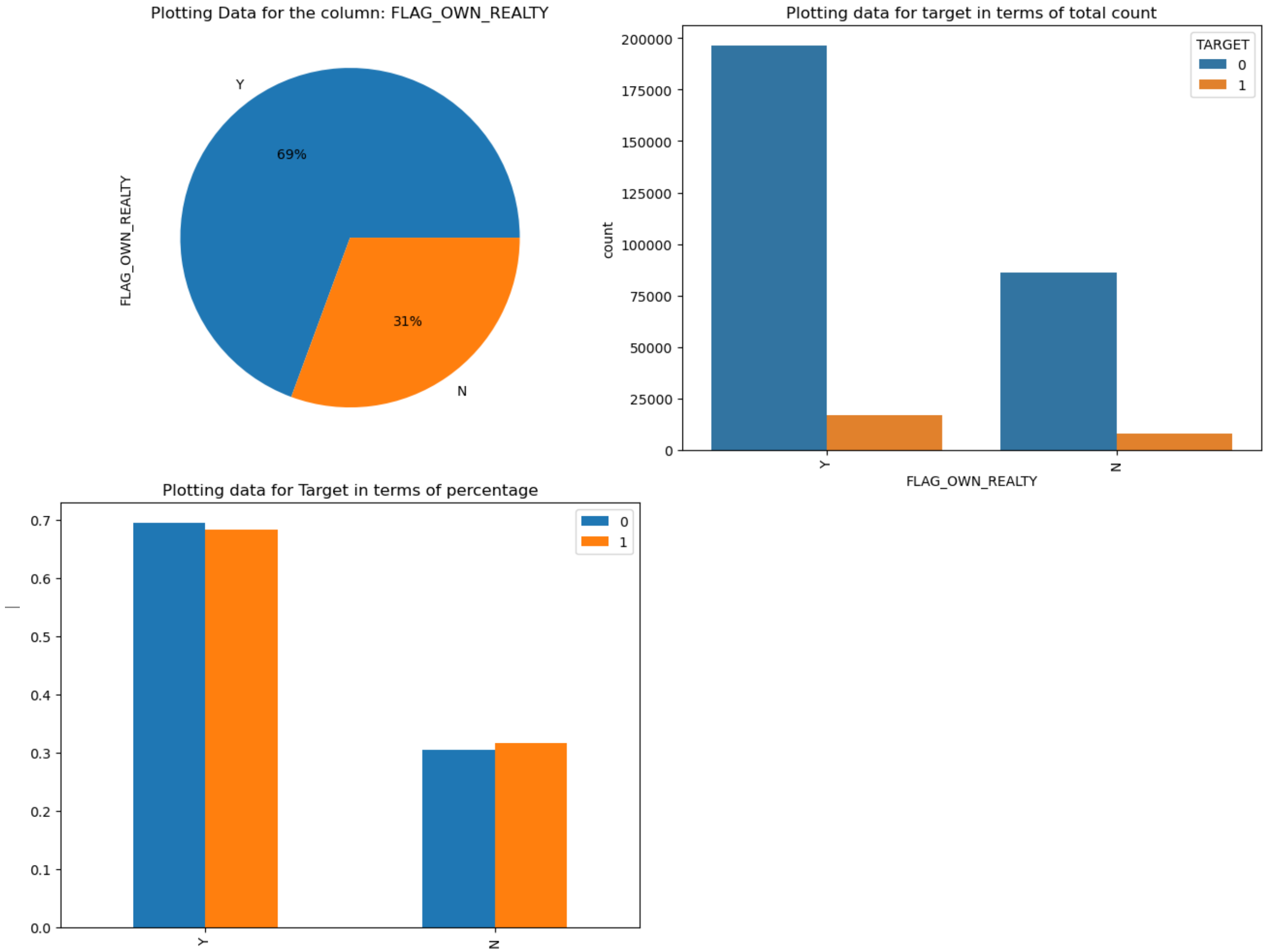

Plotting Data for the column: CODE_GENDER


Plotting data for target in terms of total count


Plotting data for Target in terms of percentage

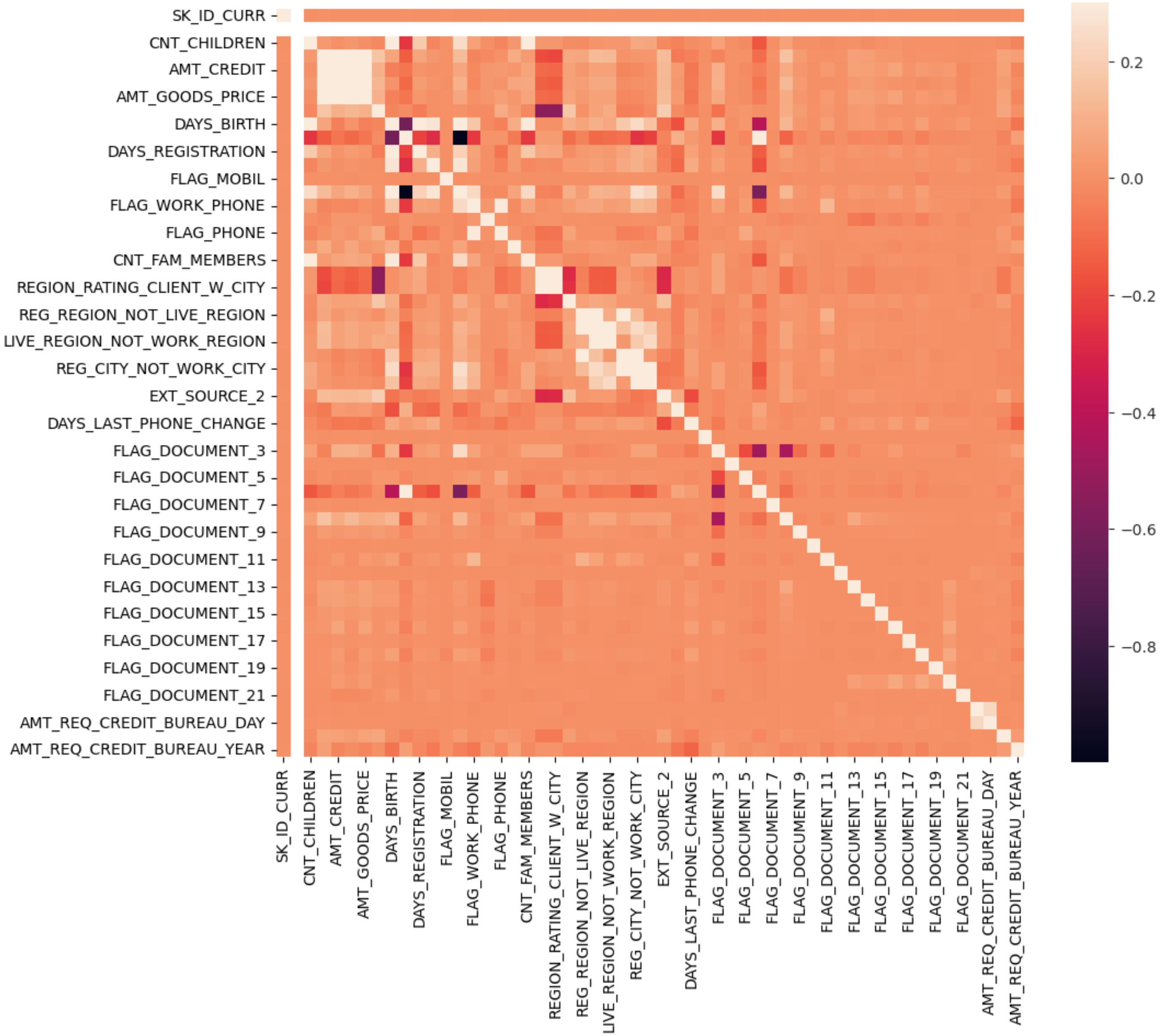
- Similarly all Charts shows the insights of remaining categorical columns.
- Below are FLAG_OWN _CAR and FLAG_OWN_REALTY



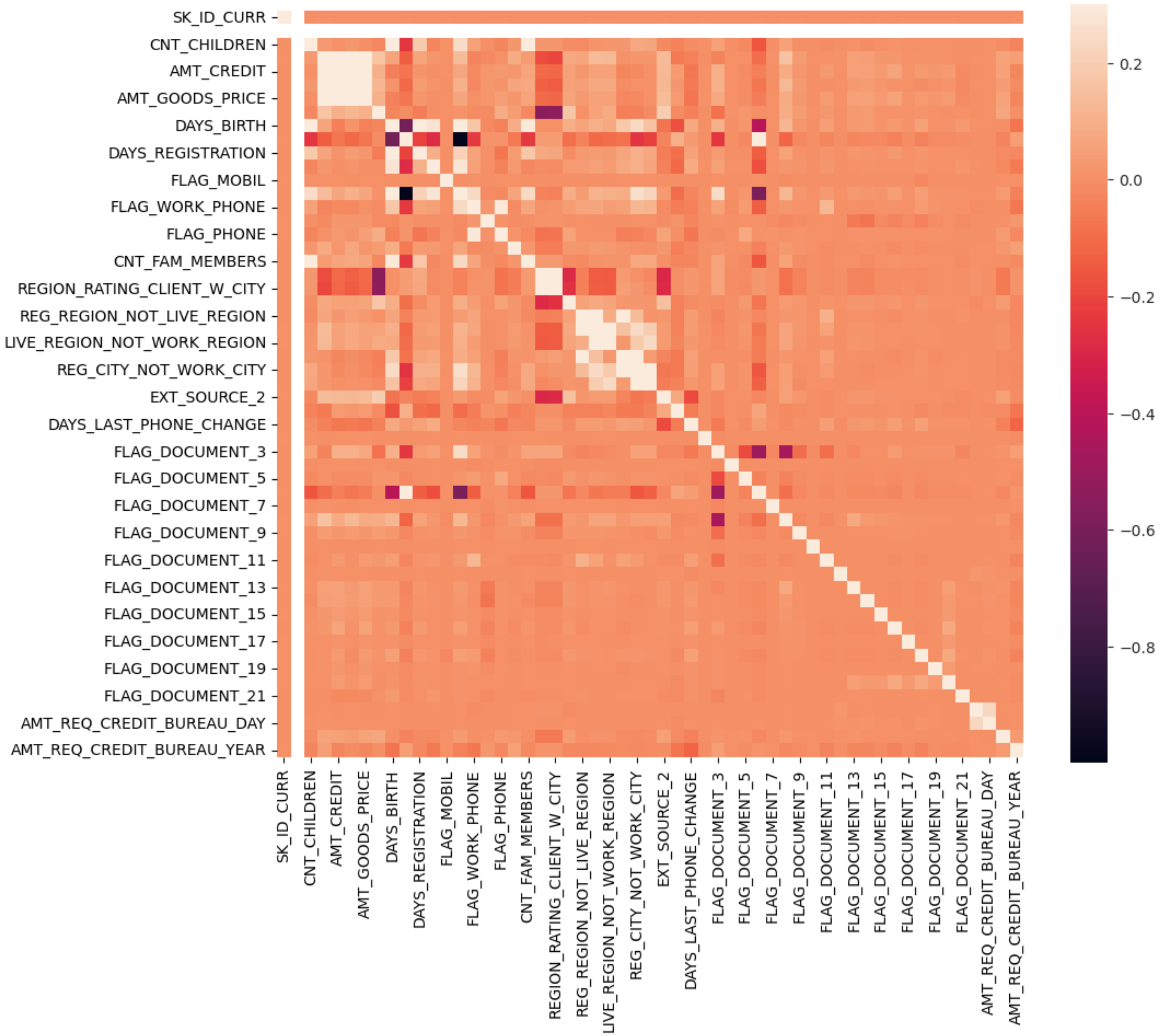


Now we have to work on numerical columns for correlation insights
For this we are using heatmaps so that correlation of all numerical columns can be seen with each other.

- Adjacent figure shows the correlation of all numerical columns with each other.
- In this case we have only taken cases with Target equals to Zero that is Loans With Payment Difficulties or Non-Performing Assets.



- From these we have to find top 10 correlations and removing those correlations which have correlation absolutely equals to 1.00
- The same process of heatmap is done for numerical columns for Target equals to 1 that is customers who have paid loan on time.



- Similar process of EDA is done on previous application file for 10% on more data and then merging it for finding insights.