- First we loaded the application_data.csv file and checked its structure.
- Then we found the percentage of missing values of all the columns and removed columns with higher percentage(more then 50%) of null values .

```
df = df.loc[:,round(100*(df.isnull().sum()/Len(df.index)),2)<50]
```

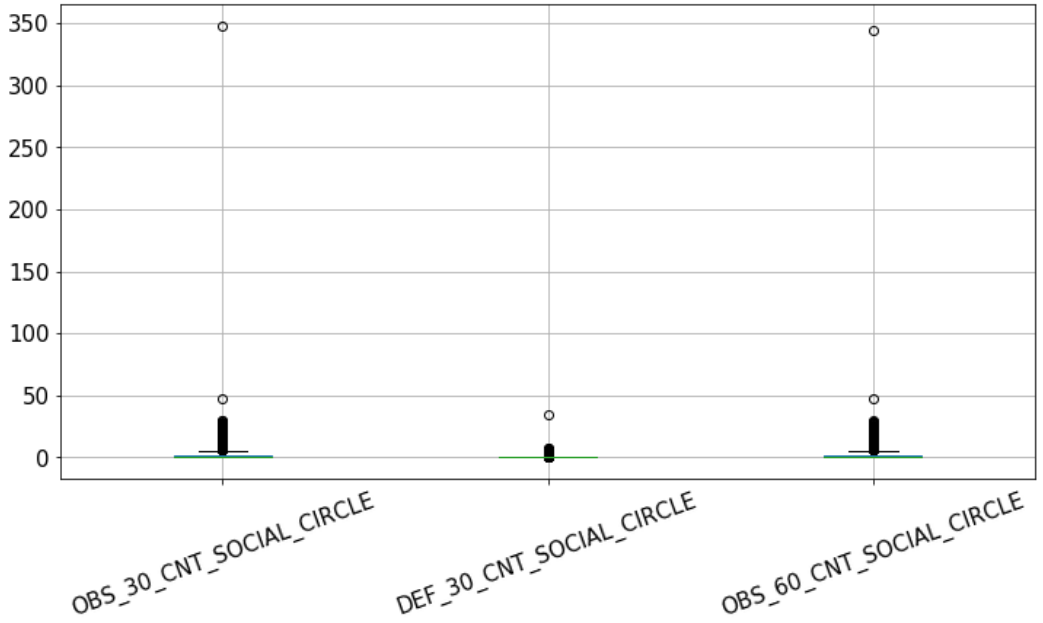
- Selected columns in variable which have low percentage of missing values, then checked its description using describe. After that imputed the null values in these columns using best metric(Mean/ Median).

```
Input_cols = ['AMT_GOODS_PRICE', 'EXT_SOURCE_2', 'EXT_SOURCE_3',
              'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_60_CNT_SOCIAL_CIRCLE', 'CNT_FAM_MEMBERS']
```

AMT_GOODS_PRICE = will impute this columns null value with mean as it don't have any outliers.
EXT_SOURCE_2 = will impute this columns null value with mean as it don't have any outliers.
EXT_SOURCE_3 = will impute this columns null value with mean as it don't have any outliers.
OBS_30_CNT_SOCIAL_CIRCLE = will impute this columns null value with median as this column has outliers which affect the mean.
DEF_60_CNT_SOCIAL_CIRCLE = will impute this columns null value with median as this column has outliers which affect the mean.
CNT_FAM_MEMBERS = will impute this columns null value with median as this column has outliers which affect the mean.

- Imbalance Percentage for Variable Target is below:

  Target = 1 then = 8.07
  Target = 0 then 92.93
  Imbalance % = 8.07/92.93



As we can see in Boxplot of 3 variables that there are some values present in this which are out of the box so they are clearly outliers.
So the Variables 'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE', 'OBS_60_CNT_SOCIAL_CIRCLE' have outliers in their values.

So to handle Outliers, we can remove them from the datasets or in some cases we can replace them also. In this we have replaced them by median.

	REGION_POPULATION_RELATIVE	Poppulation_level	CNT_FAM_MEMBERS	Family Size
0	0.018801	Medium Poppulated	1.0	Small Family
1	0.003541	Less Poppulated	2.0	Small Family
2	0.010032	Medium Poppulated	1.0	Small Family
3	0.008019	Less Poppulated	2.0	Small Family
4	0.028663	Medium Poppulated	1.0	Small Family
...	...	...	...	...
307506	0.032561	Medium Poppulated	1.0	Small Family
307507	0.025164	Medium Poppulated	1.0	Small Family
307508	0.005002	Less Poppulated	1.0	Small Family
307509	0.005313	Less Poppulated	2.0	Small Family
307510	0.046220	Medium Poppulated	2.0	Small Family

307511 rows × 4 columns

For Binning, we have selected 2 variables (REGION_POPULATION_RELATIVE, CNT_FAM_MEMBERS).

We have assigned the binning Values as Below:
'REGION_POPULATION_RELATIVE:

REGION_POPULATION RANGE	Bin
0-0.01	Less Poppulated
0.01-0.1	Medium Poppulated
0.1 - 1.0	Highly Poppulated

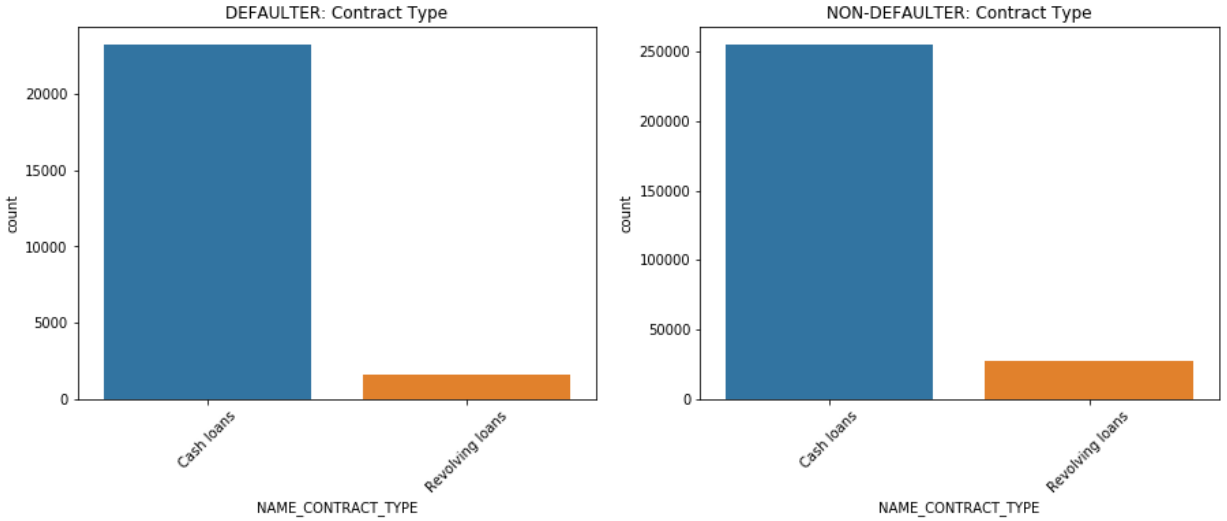
CNT_FAM_MEMBERS

CNT_FAM_MEMBERS RANGE	Bin
0-4	Small Family
4-10	Medium Family
10-20	Large Family

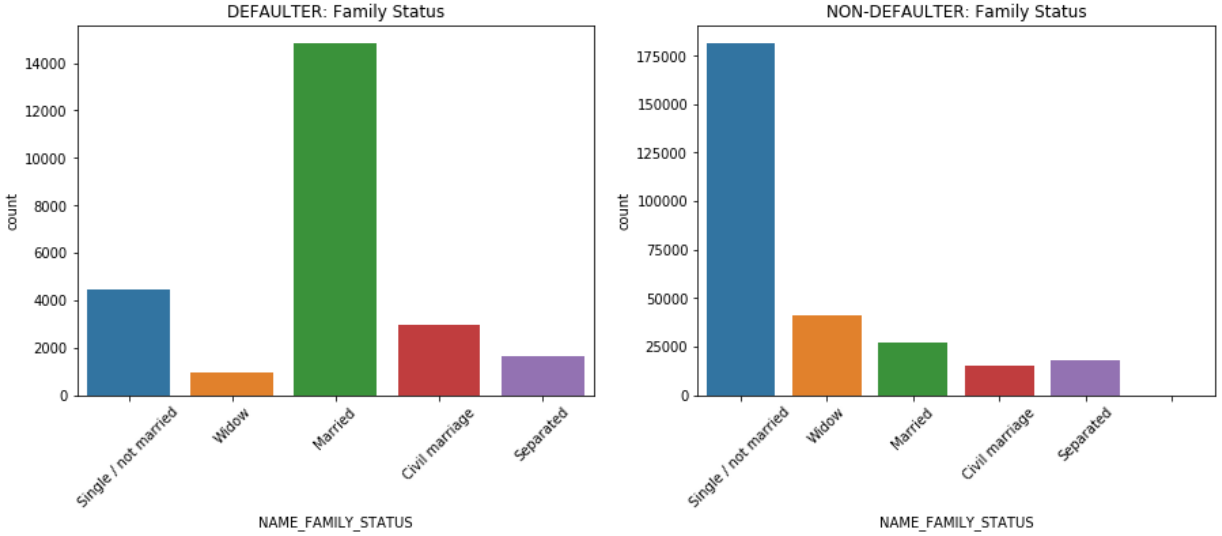
Univariate analysis for categorical variables for both Target 0 and 1.

Below are the Comparison of the target variable across categories of categorical variables. Left plot is for defaulter in all plots.

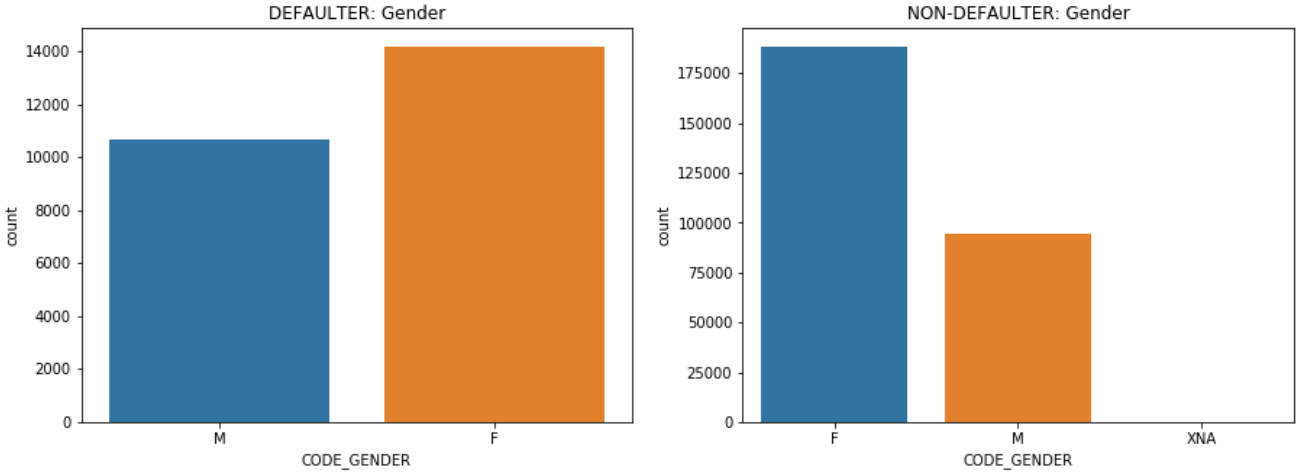
- NAME_CONTRACT_TYPE: Loan type for both Defaulter and Non Defaulter have as cash loans.



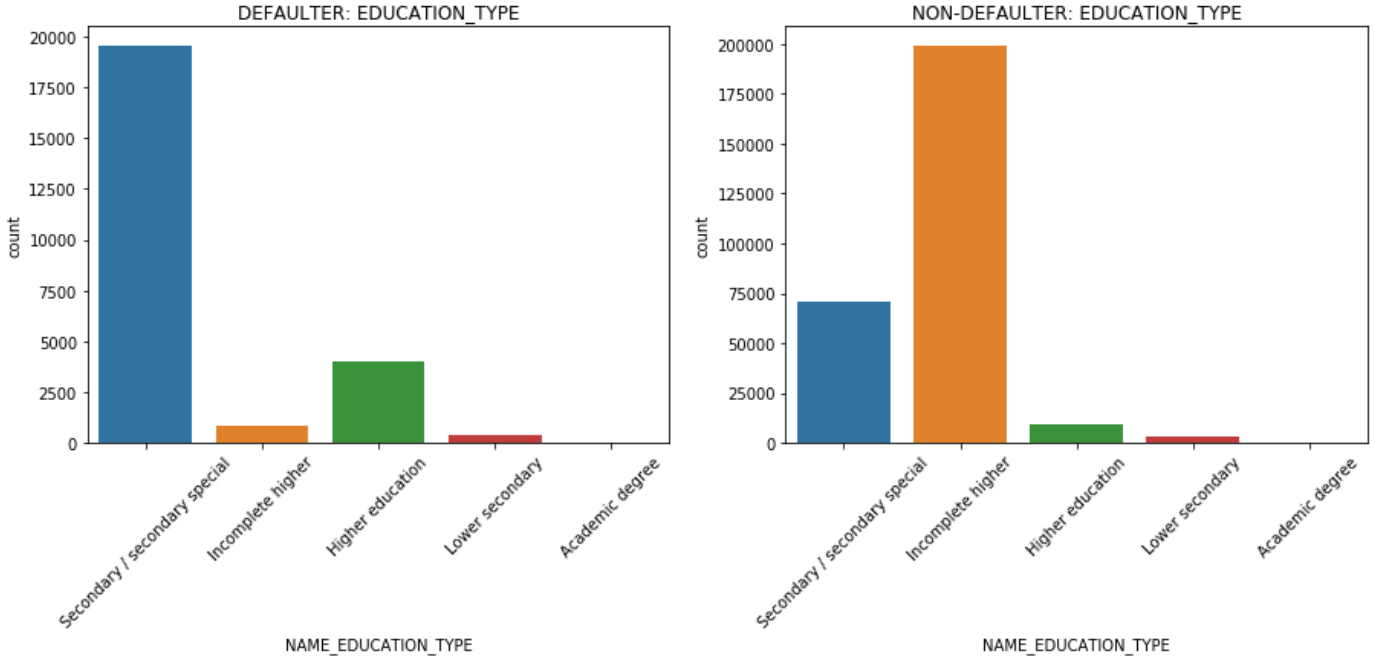
- NAME_FAMILY_STATUS: Married people are more likely to be Defaulter and single/not married are more likely to be Non Defaulter.



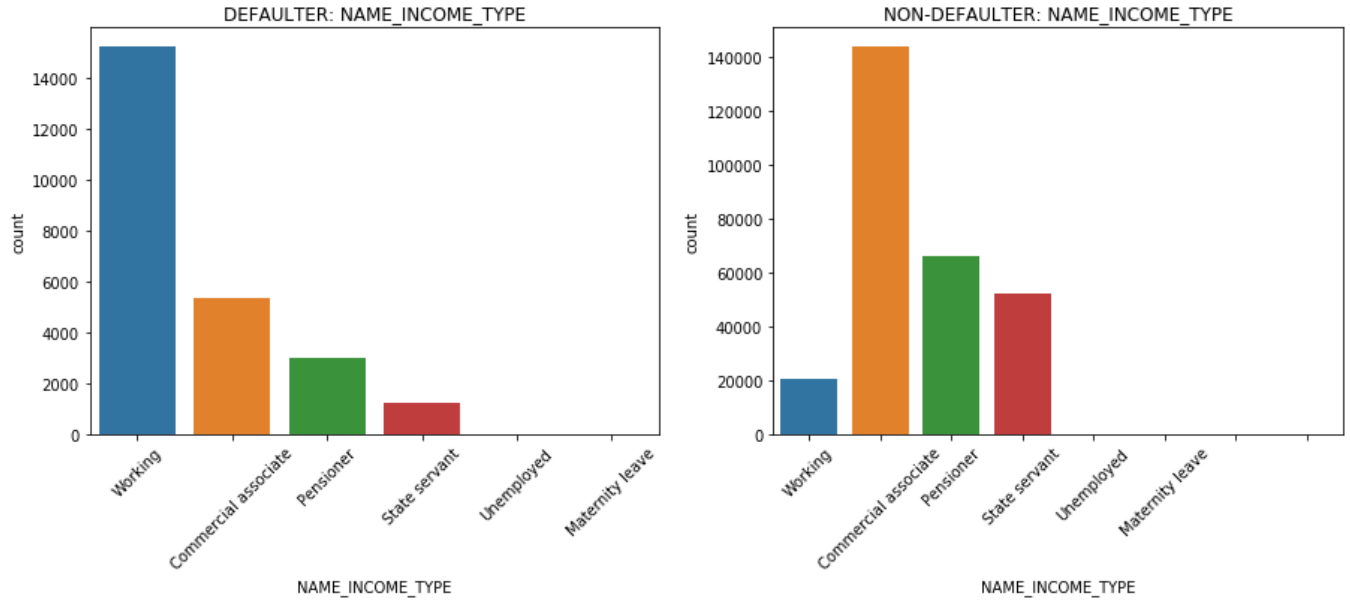
- CODE_GENDER: Most Females are Defaulter and most males are Non Defaulter.



- NAME_EDUCATION_TYPE: Secondary specials educated are mostly like to be Defaulter and Incomplete Higher Education are mostly Non Defaulter.



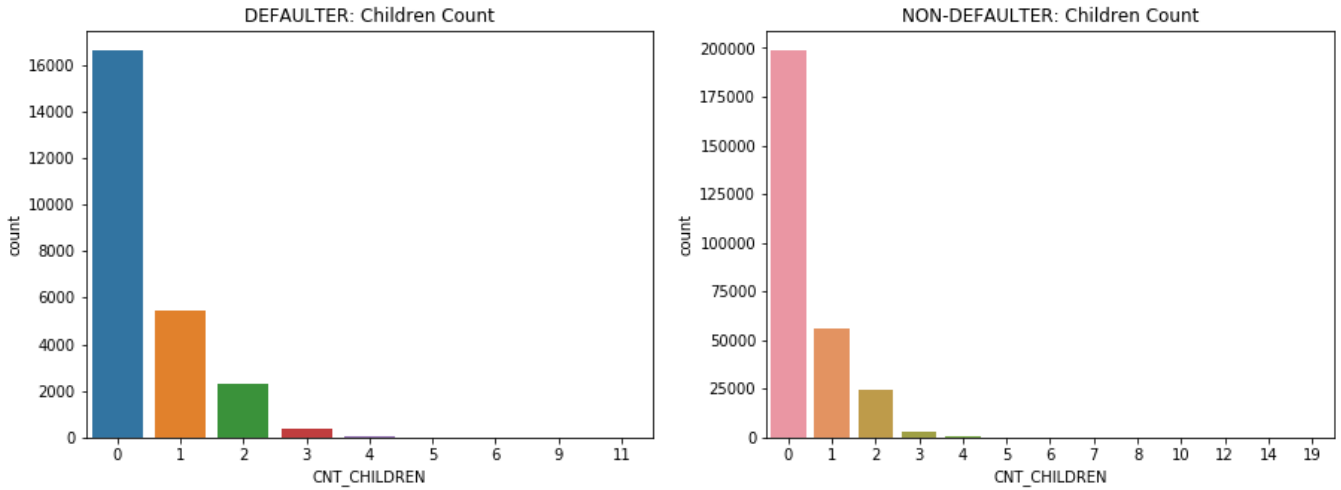
- NAME_INCOME_TYPE: Working Income type people are mostly defaulter and Commercial associate are mostly Non Defaulter.



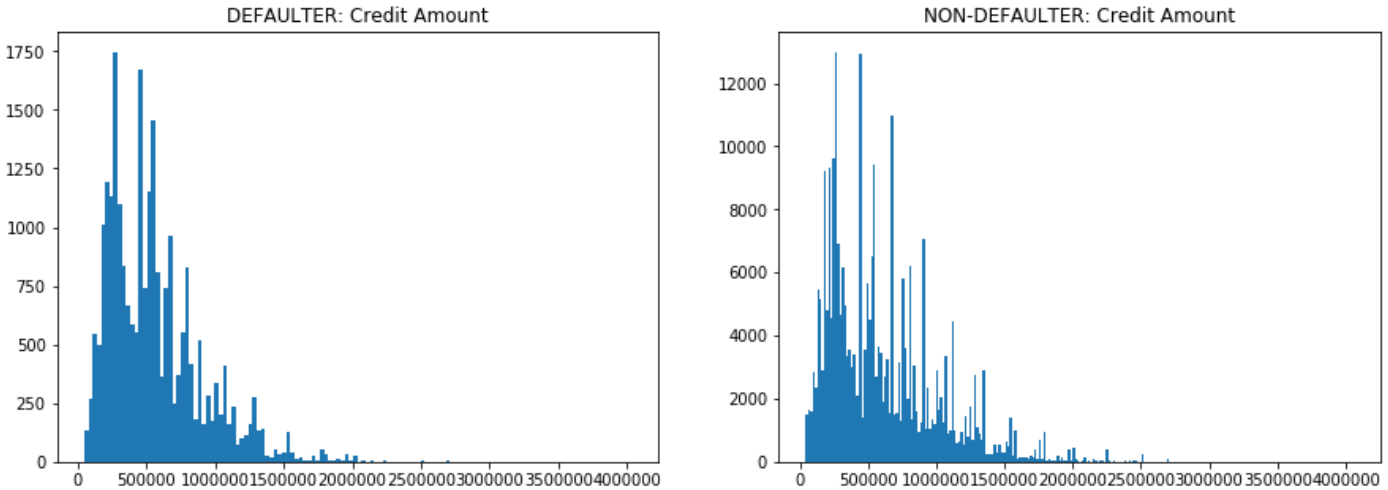
Univariate analysis for numerical variables for both Target 0 and 1.

Below are the Comparison of the target variable across categories of Continues variables. Left plot is for defaulter in all plots.

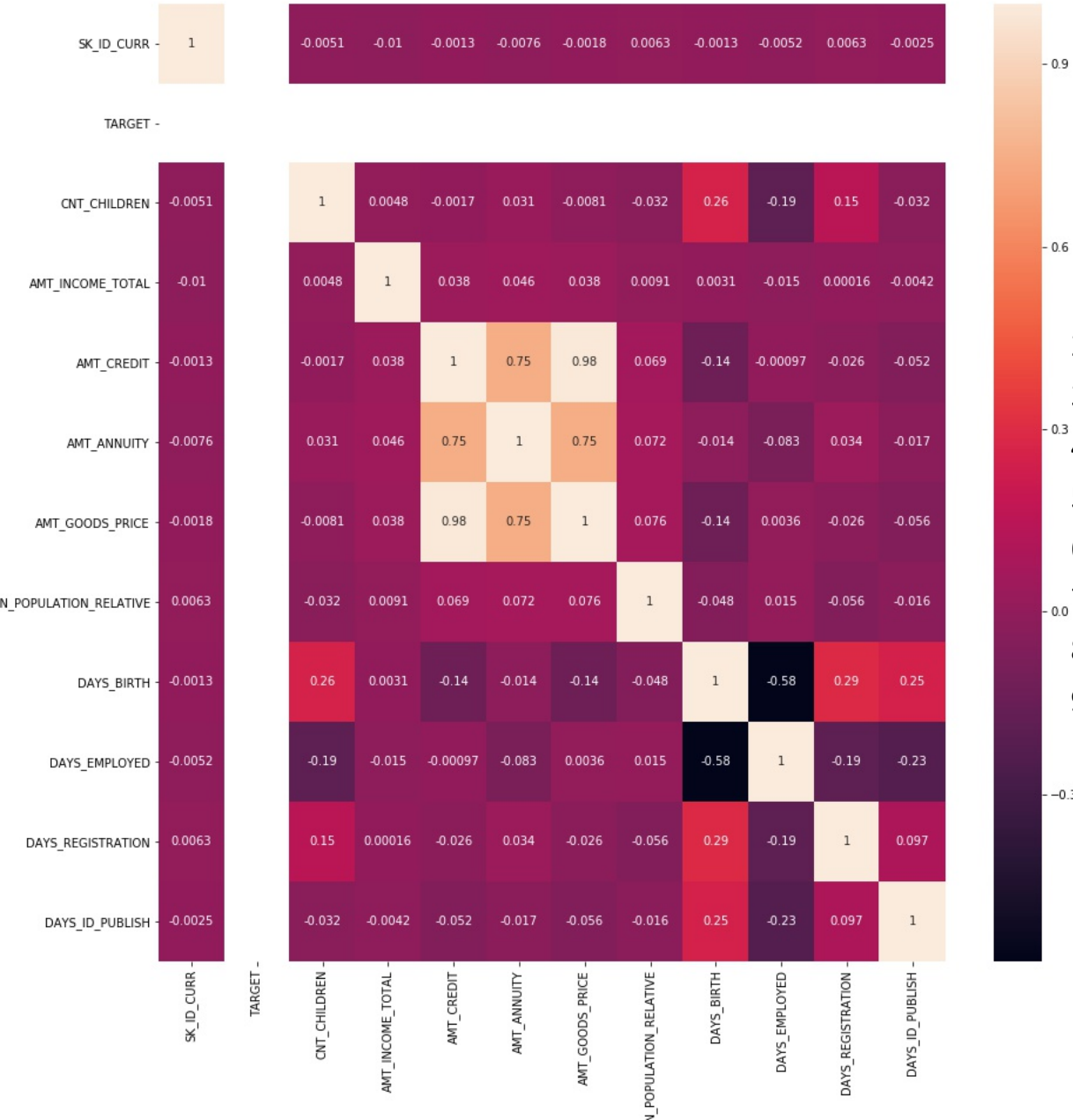
- CNT_CHILDREN: Mostly Defaulter and Non-Defaulter have same children's count i.e. 0.



- AMT_CREDIT: Credit amount is mostly distributed between 0-500000 for defaulter and for Non-Defaulter it is distributed equally with high spike between.



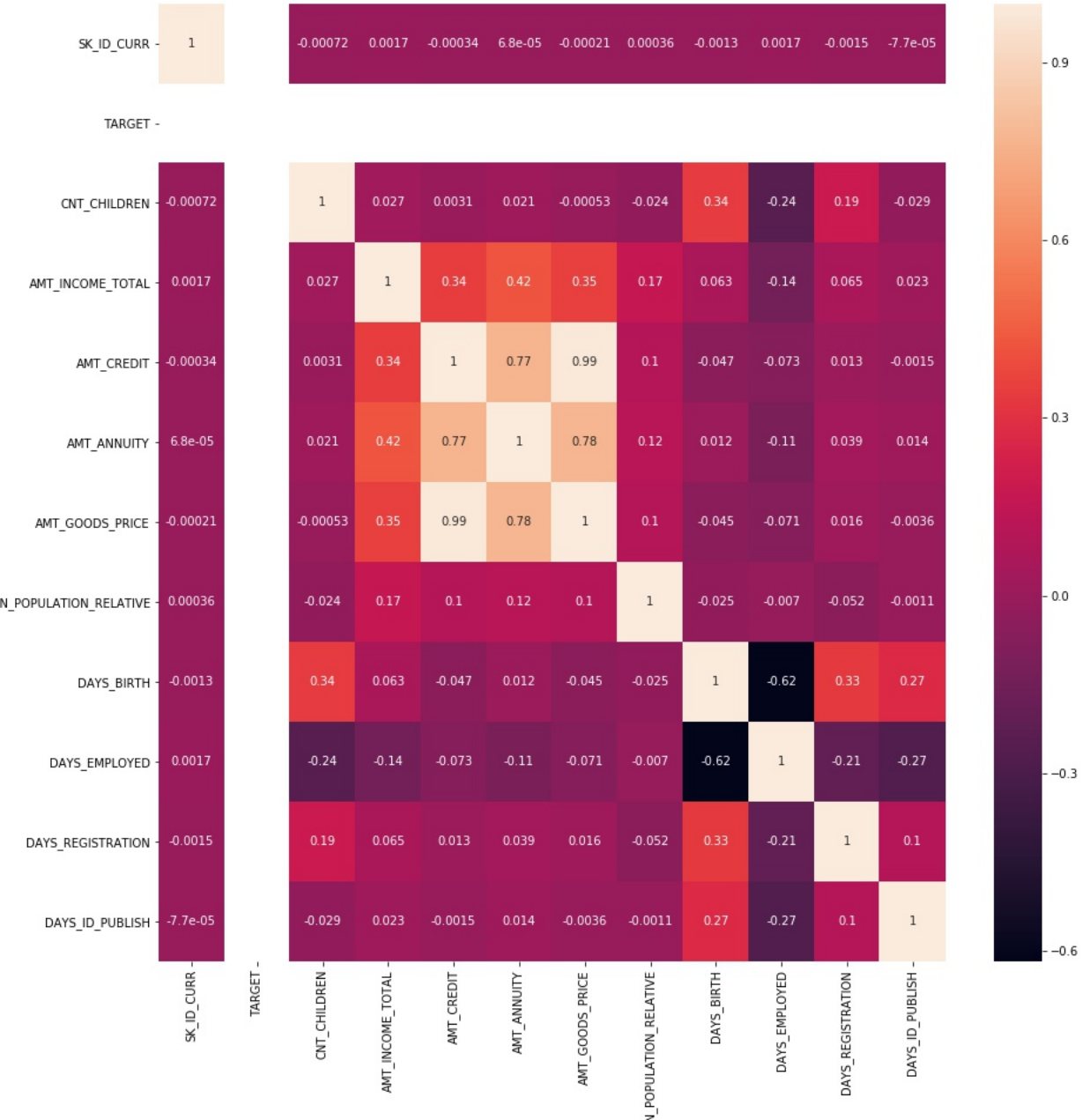
- **HEATMAP FOR DEFAULTER CATEGORY:**



- As per Defaulter Category Heatmap the top 10 correlation for the **Client with payment difficulties** are below:

1. AMT_GOODS_PRICE and AMT_CREDIT
2. AMT_GOODS_PRICE and AMT_AMMUNITY
3. AMT_AMMUNITY and AMT_CREDIT
4. DAYS_BIRTH and DAYS_REGISTRATION
5. CNT_CHILDREN and DAYS_BIRTH
6. DAYS_BIRTH and DAYS_ID_PUBLISH
7. CNT_CHILDREN and DAYS_REGISTRATION
8. DAYS_REGISTRATION and DAYS_ID_PUBLISH
9. REGION_POPULATION_RELATIVE and AMT_GOODS_PRICE
10. REGION_POPULATION_RELATIVE and AMT_AMMUNITY

- **HEATMAP FOR NON-DEFAULTER CATEGORY:**



- As per Non-Defaulter Category Heatmap, the top 10 correlated variables for the **Client without any payment difficulties** are below:

1. AMT_GOODS_PRICE and AMT_CREDIT
2. AMT_GOODS_PRICE and AMT_AMMUNITY
3. AMT_AMMUNITY and AMT_CREDIT
4. AMT_AMMUNITY and AMT_INCOME_TOTAL
5. AMT_GOODS_PRICE and AMT_INCOME_TOTAL
6. AMT_INCOME_TOTAL and AMT_CREDIT
7. CNT_CHILDREN and DAYS_BIRTH
8. DAYS_BIRTH and DAYS_REGISTRATION
9. DAYS_BIRTH and DAYS_ID_PUBLISH
10. CNT_CHILDREN and DAYS_REGISTRATION

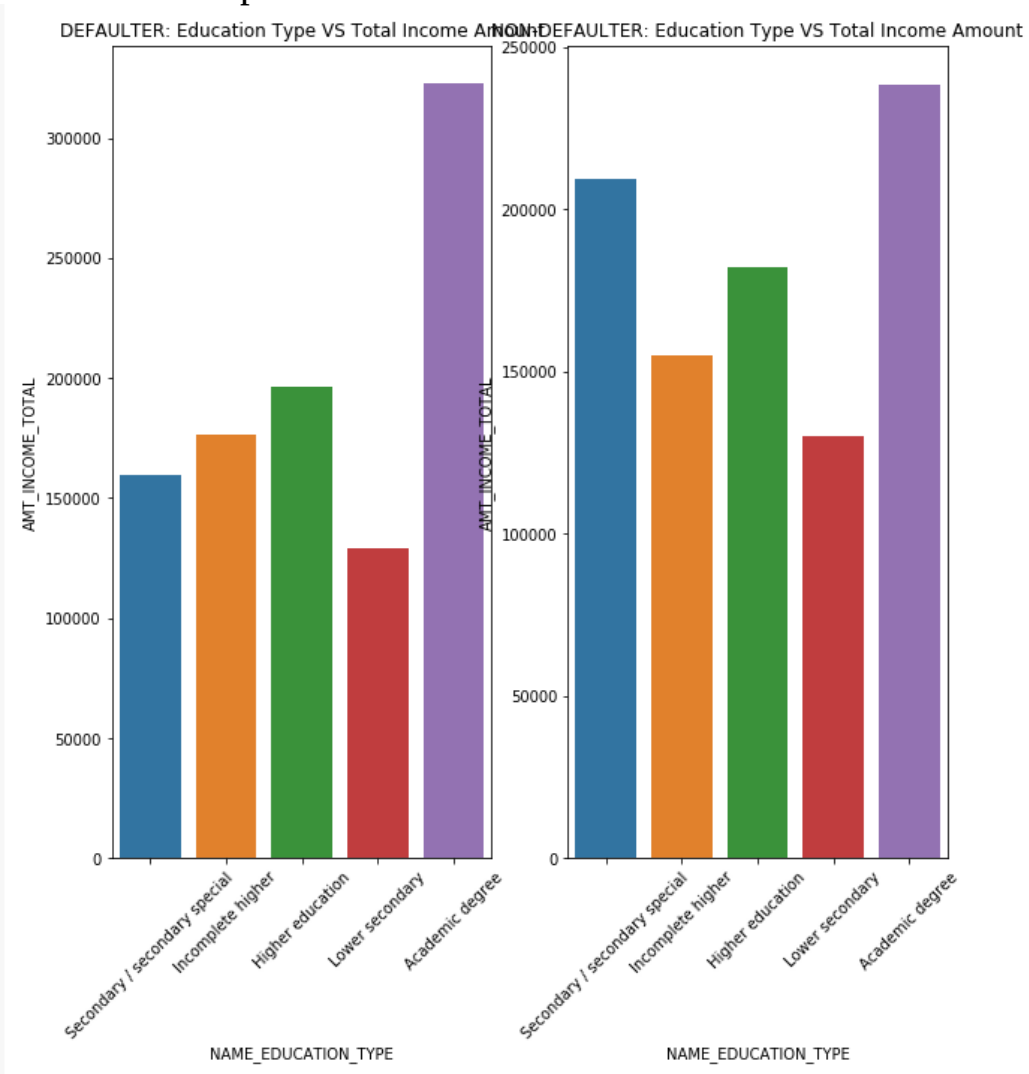
So As per heatmap analysis of both the Categories (Defaulter and Non-defaulter, it is confirmed that TOP-3 Correlated variables are same in both the Categories.

1. AMT_GOODS_PRICE and AMT_CREDIT
2. AMT_GOODS_PRICE and AMT_AMMUNITY
3. AMT_AMMUNITY and AMT_CREDIT

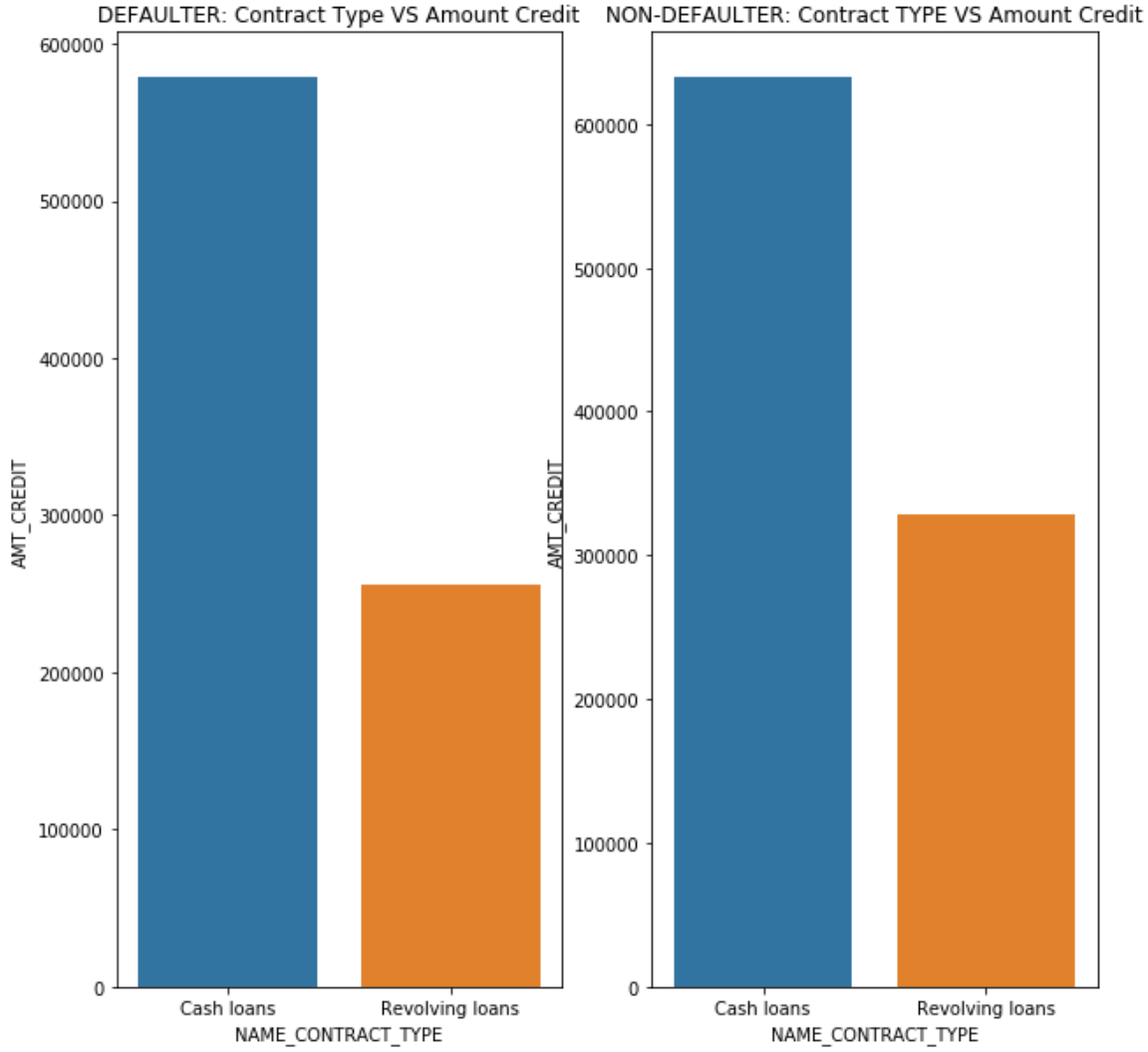
Bivariate analysis for variables for both Target 0 and 1.

Below are the Comparison of the target variable across categories of variables. Left plot is for defaulter in all plots.

- Education Type VS Total Income: For Defaulter and Non Defaulter Academic Degree people are having more total Income compare to other educated person.

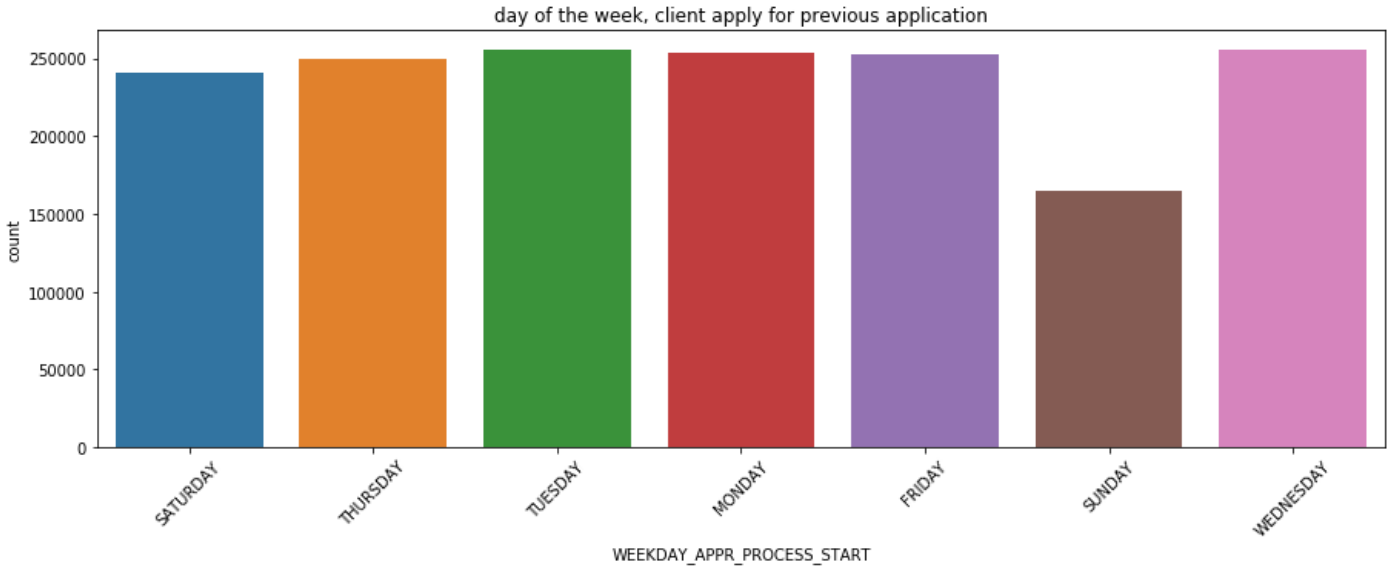


- Contract Type VS Amount Credit: For Cash loan types defaulter have Less Credited amount compare to Non-defaulter.

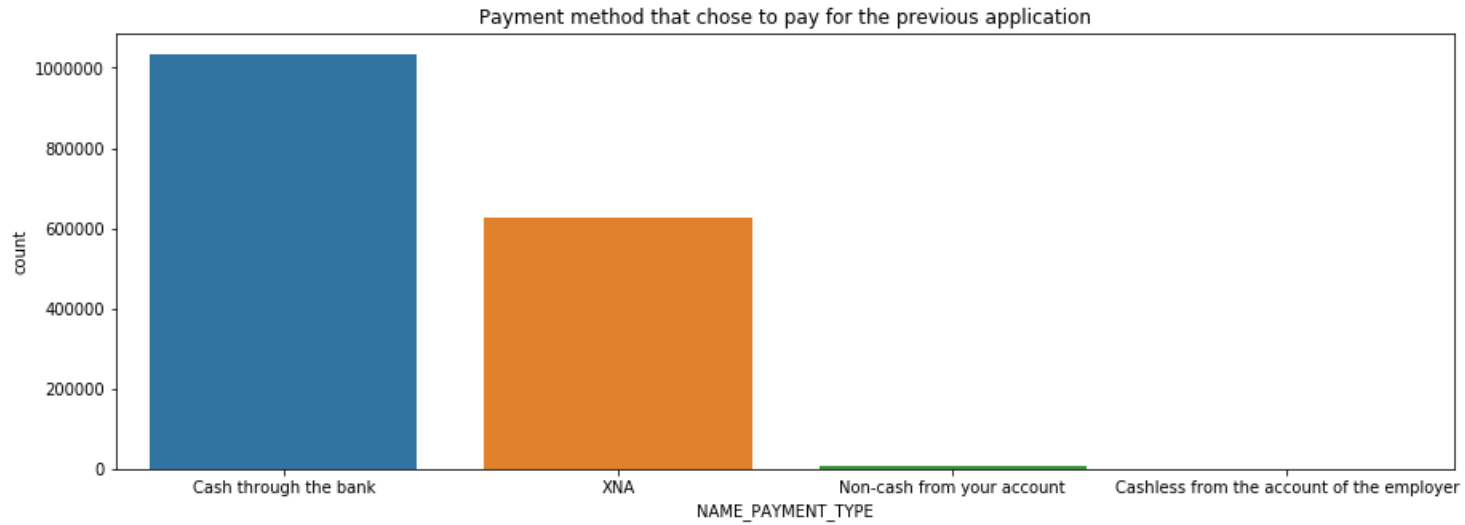


Univariate analysis for variables of Previous Application Data
Below are the analysis of variables for Previous Application Data.

- WEEKDAY_APPR_PROCESS_START: Sunday was the day when less application applied in previous applicaiton.

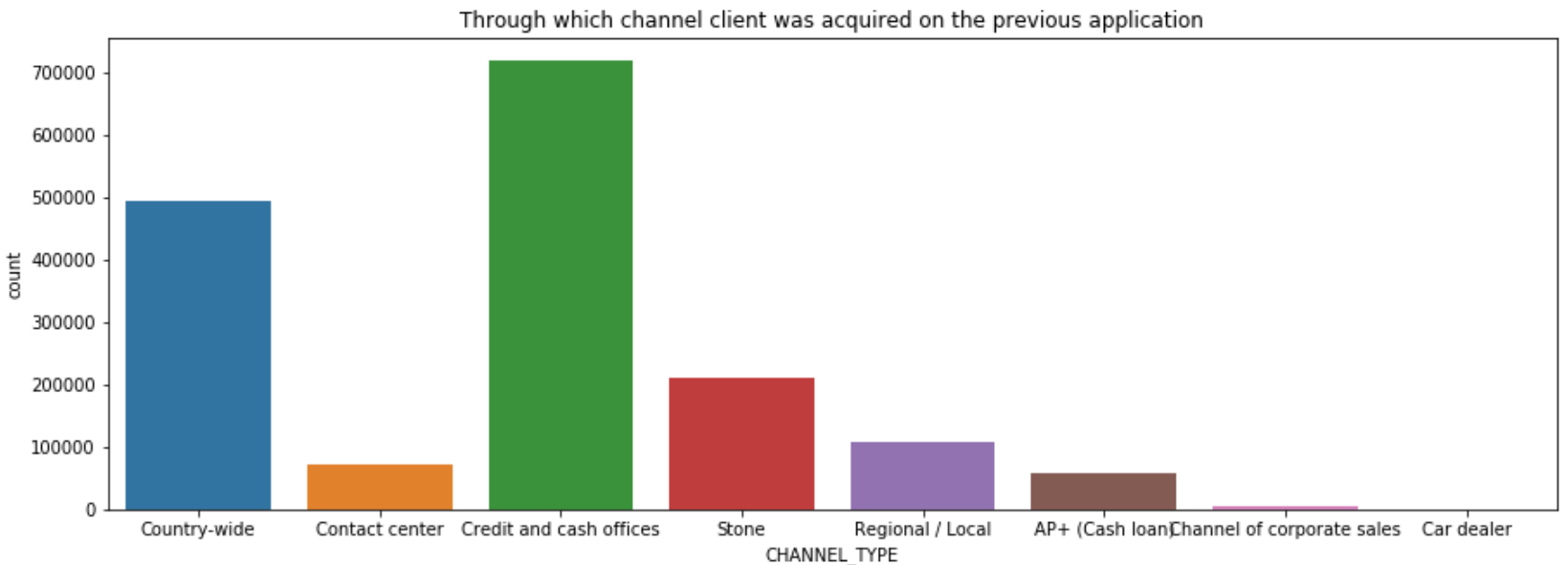


- NAME_PAYMENT_TYPE: As per shown in graph, mostly customers payed their previous application amount in Cash and mostly others didnt declare their payment method.



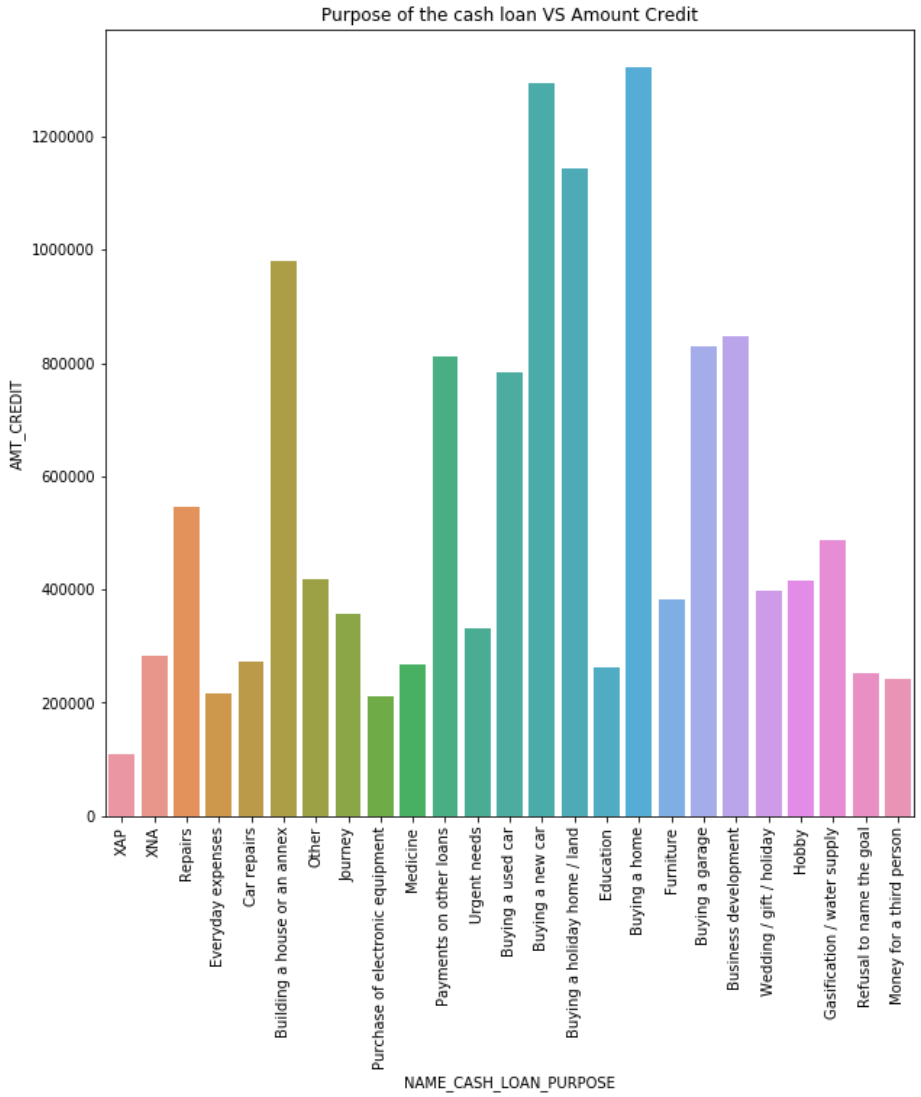
NAME_CONTRACT_STATUS: As per shown in graph, mostly application approved last time. But compare to refused, Cancelled were more.

CHANNEL_TYPE: As per shown in graph, mostly customers aquired on the previous application through Credit and Cash offices



Bivariate analysis for variables of Previous Application Data
Below are the analysis of variables for Previous Application Data.

Purpose of the cash loan VS Amount Credit: Amount credited for Most of the Cash loans for the Buying a home.



Contract Type VS Application Amount: Application amount for refused applications are more high.