EDA Credit Loan

Assignment
Submitted By:- Harshraj Chavda
Email:- harshchavda439@gmail.com

Objective

We are given two files (.csv) containing relevant information about clients:-

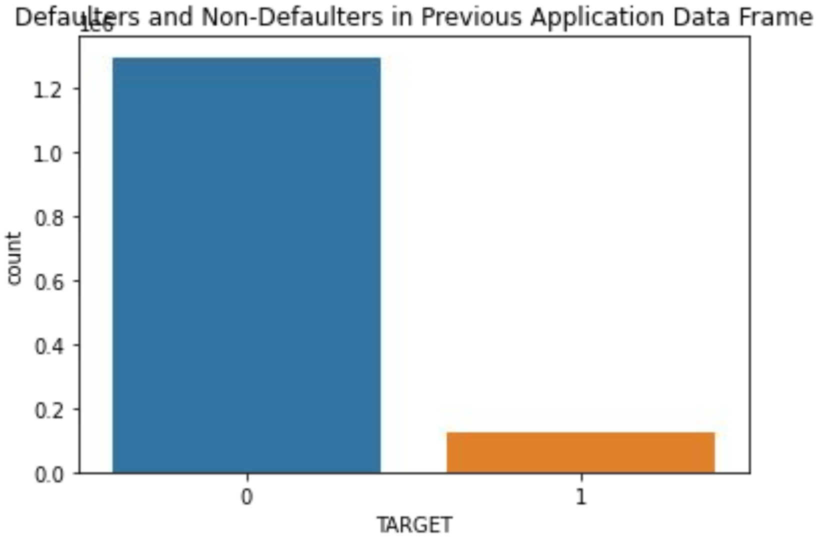
- Application.csv and Previous Application.csv.

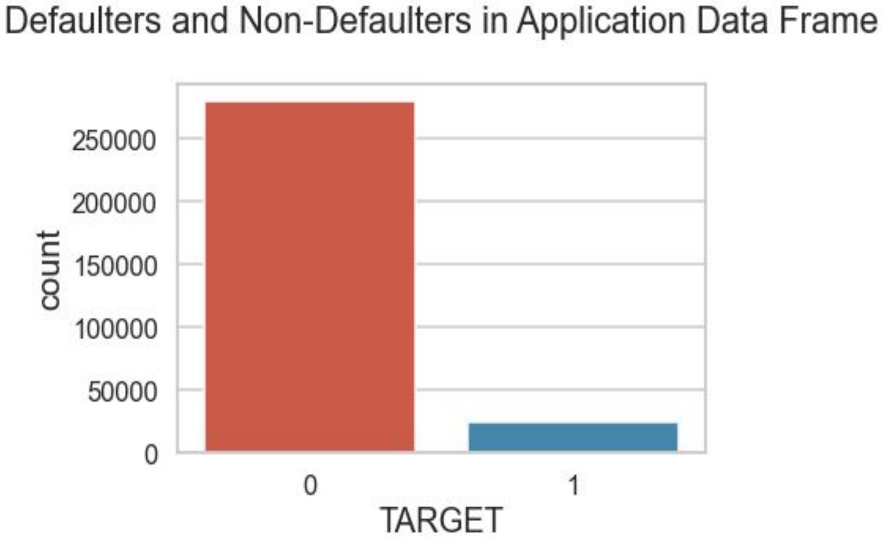
- For Application.csv file we had TARGET variable.
- We need to conduct Exploratory Data Analysis (EDA) to Identify attributes which relate to Defaulters (i.e. Target=1) and Non-Defaulters(i.e. Target=0)

Approach

1. The data set files (.csv) were imported on Jupyter NB for Analysis.
2. First the data was observed for it's general features, i.e. Meta Data check. Like no. of Rows, columns etc.
3. Then we proceed with Data Cleaning, which basically included identifying missing values, imputing them with relevant values i.e. Mode, Mean or Median etc.
4. We checked for Data Imbalance (TARGET col) in Data Frame, i.e. Previous Application and Current Application and reported it.
5. We checked for Outliers in relevant columns. I.e. Numerical Columns.
6. Then we proceeded with Univariate(Segmented) and Bi/Multivariate analysis of all the columns within both Data frame to find distinct trend among Defaulters and Non-Defaulters

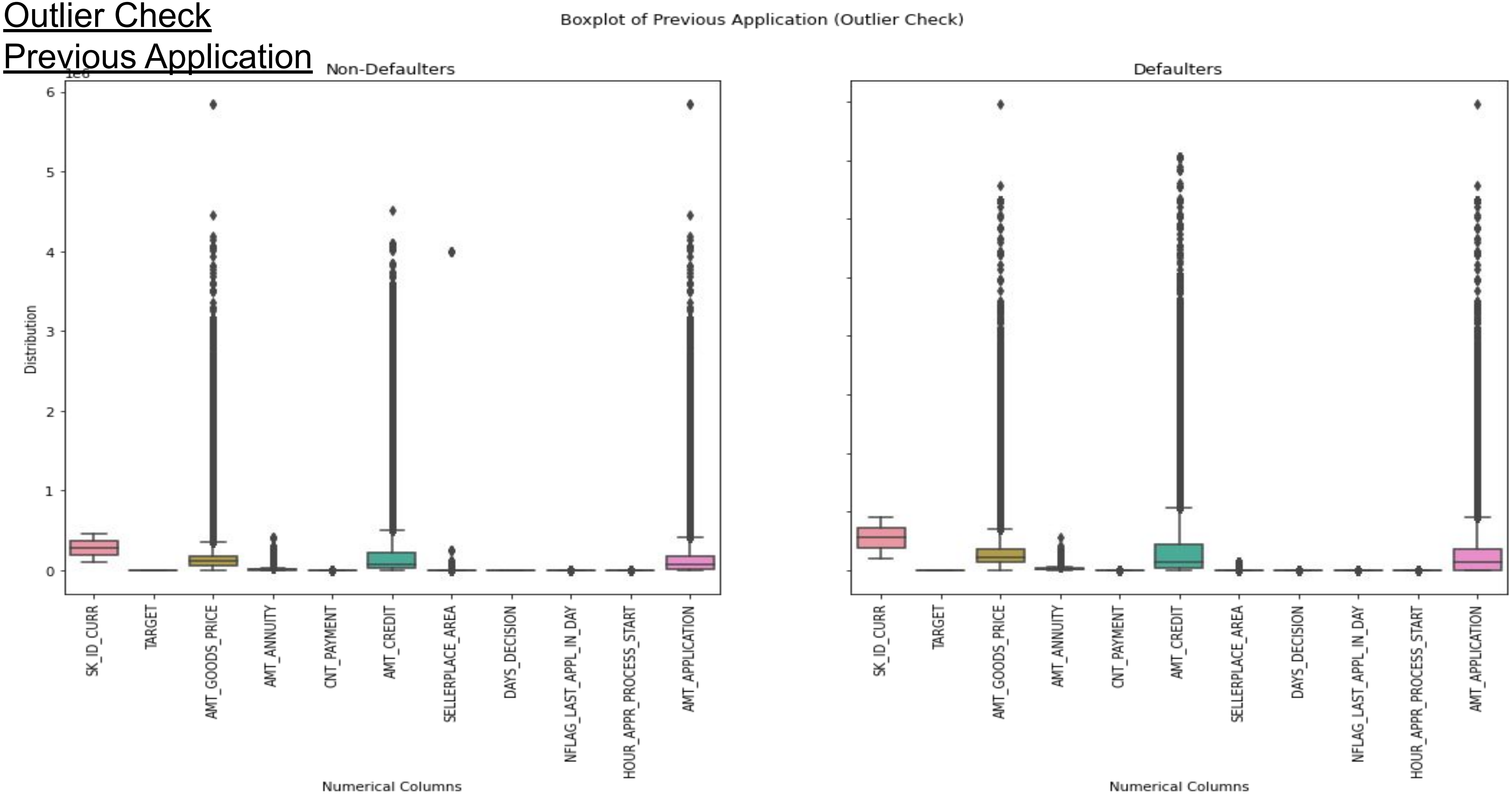
Results
Data Imbalance for both Data Frames





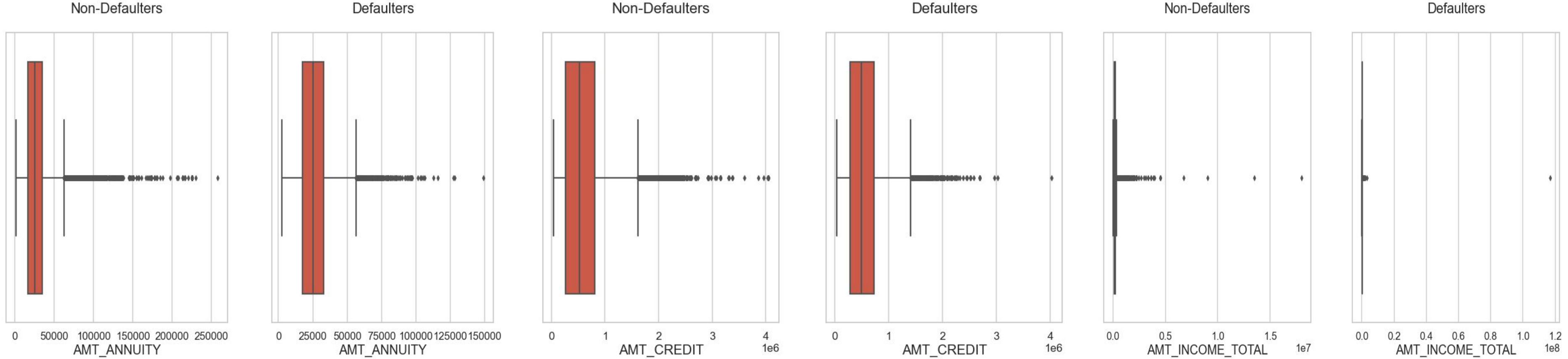
Inference:- For both Data frames the Data Imbalance ratio is 9:1

Outlier Check
Previous Application

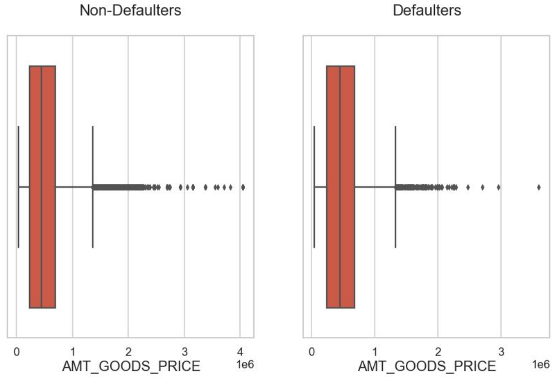


Inference:- ***'AMT_GOODS_PRICE' , 'AMT_CREDIT', 'AMT_APPLICATION' have higher distribution value for Defaulters based on their interquartile size range.***
All these numerical variables have Outliers.

Outlier Check Application



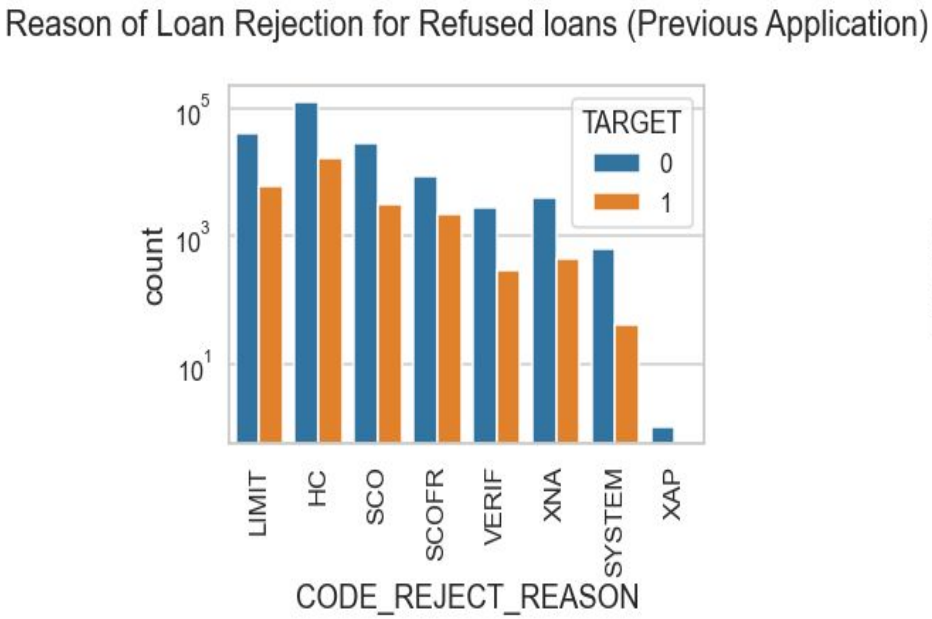
- We didn’t floor or cap any Outliers in our EDA analysis

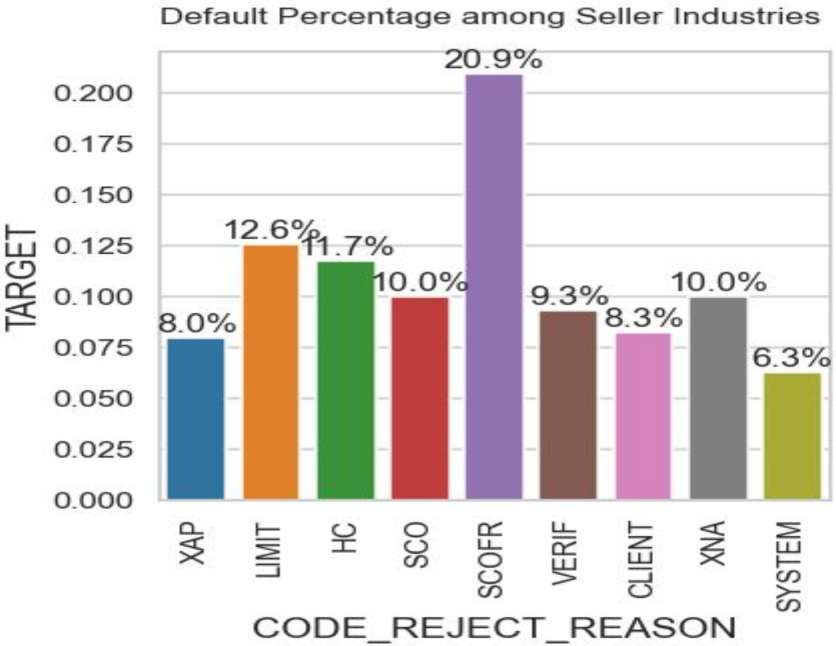


Inference:- ***'AMT_GOODS_PRICE' , 'AMT_CREDIT', 'AMT_APPLICATION' have higher distribution value for Defaulters based on their interquartile size range.***Generally box-plot of distribution of Annuity, Goods Price, credit and Income total reveal that Outliers in Non-Defaulters have higher values compared to Non-Defaulters**.**

Segmented Univariate Analysis
Previous Application

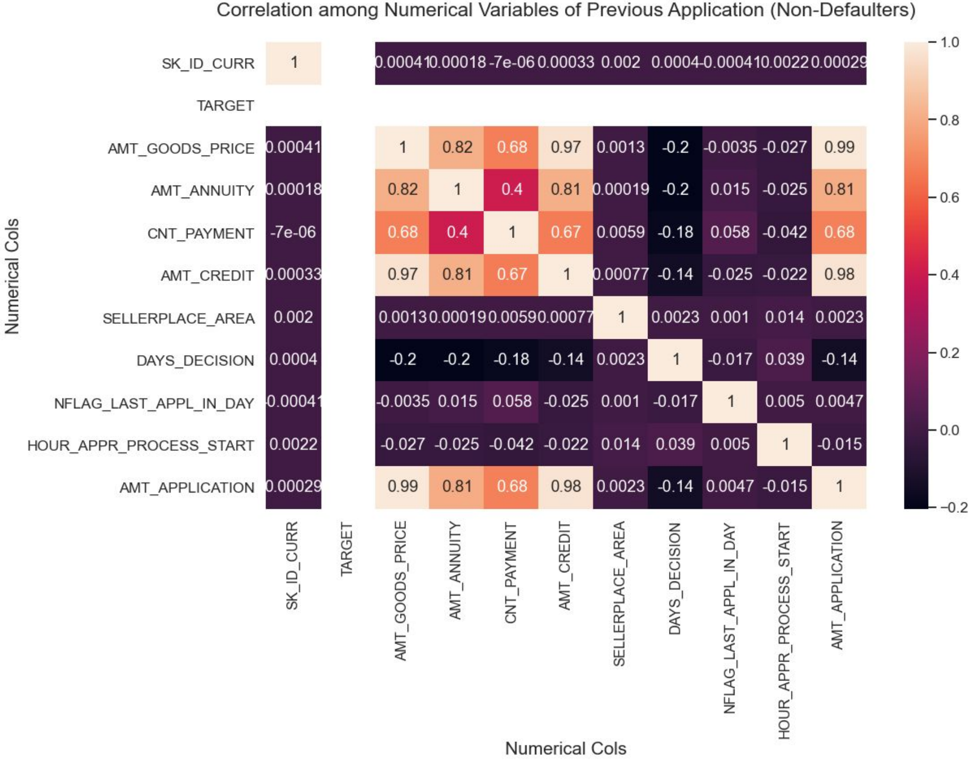
- Such segmented Univariate analysis were conducted for Other Columns as well within Previous application data frame.



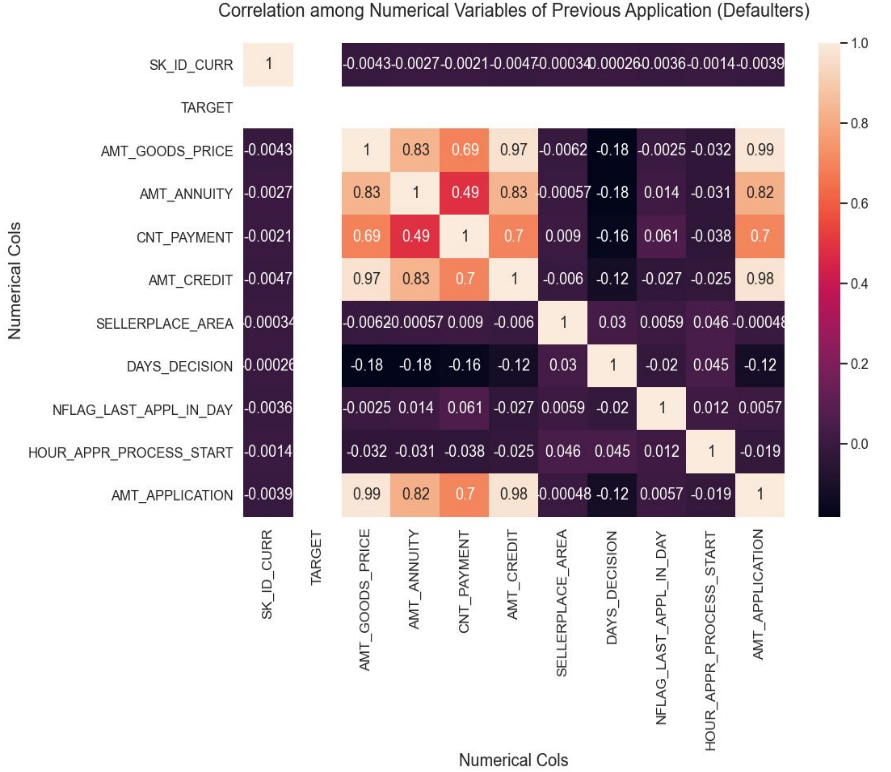


Inference:- 'Among Individuals whose loans were refused, majority of them were of 'HC' and 'LIMIT' rejects.In general, whose loans were rejected, majority of the defaulters (~21%) had 'SCOFR' as the reject code.

Bivariate Analysis- Correlation of Numerical Columns Previous Application



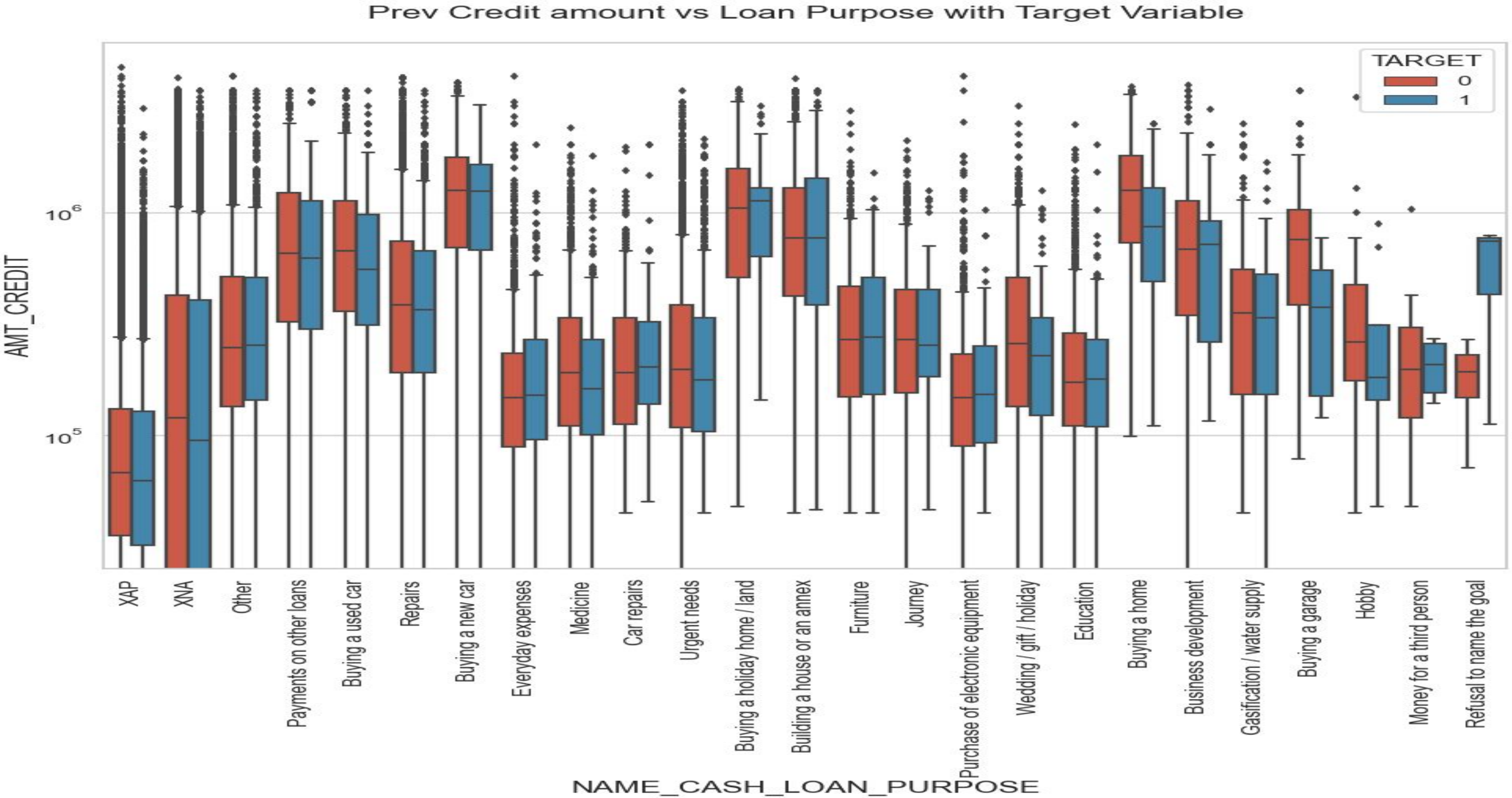
	SK_ID_CURR	TARGET	AMT_GOODS_PRICE	AMT_ANNUITY	CNT_PAYMENT	AMT_CREDIT	SELLERPLACE_AREA	DAYS_DECISION	NFLAG_LAST_APPL_IN_DAY	HOUR_APPR_PROCESS_START	AMT_APPLICATION
SK_ID_CURR	1		0.00041	0.00018	-7e-06	0.00033	0.002	0.0004	-0.00041	0.0022	0.00029
TARGET											
AMT_GOODS_PRICE	0.00041		1	0.82	0.68	0.97	0.0013	-0.2	-0.0035	-0.027	0.99
AMT_ANNUITY	0.00018		0.82	1	0.4	0.81	0.00019	-0.2	0.015	-0.025	0.81
CNT_PAYMENT	-7e-06		0.68	0.4	1	0.67	0.0059	-0.18	0.058	-0.042	0.68
AMT_CREDIT	0.00033		0.97	0.81	0.67	1	0.00077	-0.14	-0.025	-0.022	0.98
SELLERPLACE_AREA	0.002		0.0013	0.00019	0.0059	0.00077	1	0.0023	0.001	0.014	0.0023
DAYS_DECISION	0.0004		-0.2	-0.2	-0.18	-0.14	0.0023	1	-0.017	0.039	-0.14
NFLAG_LAST_APPL_IN_DAY	-0.00041		-0.0035	0.015	0.058	-0.025	0.001	-0.017	1	0.005	0.0047
HOUR_APPR_PROCESS_START	0.0022		-0.027	-0.025	-0.042	-0.022	0.014	0.039	0.005	1	-0.015
AMT_APPLICATION	0.00029		0.99	0.81	0.68	0.98	0.0023	-0.14	0.0047	-0.015	1



	SK_ID_CURR	TARGET	AMT_GOODS_PRICE	AMT_ANNUITY	CNT_PAYMENT	AMT_CREDIT	SELLERPLACE_AREA	DAYS_DECISION	NFLAG_LAST_APPL_IN_DAY	HOUR_APPR_PROCESS_START	AMT_APPLICATION
SK_ID_CURR	1		-0.0043	-0.0027	-0.0021	-0.0047	-0.00034	0.00026	-0.0036	-0.0014	-0.0039
TARGET											
AMT_GOODS_PRICE	-0.0043		1	0.83	0.69	0.97	-0.0062	-0.18	-0.0025	-0.032	0.99
AMT_ANNUITY	-0.0027		0.83	1	0.49	0.83	-0.00057	-0.18	0.014	-0.031	0.82
CNT_PAYMENT	-0.0021		0.69	0.49	1	0.7	0.009	-0.16	0.061	-0.038	0.7
AMT_CREDIT	-0.0047		0.97	0.83	0.7	1	-0.006	-0.12	-0.027	-0.025	0.98
SELLERPLACE_AREA	-0.00034		-0.0062	-0.00057	0.009	-0.006	1	0.03	0.0059	0.046	-0.00048
DAYS_DECISION	0.00026		-0.18	-0.18	-0.16	-0.12	0.03	1	-0.02	0.045	-0.12
NFLAG_LAST_APPL_IN_DAY	-0.0036		-0.0025	0.014	0.061	-0.027	0.0059	-0.02	1	0.012	0.0057
HOUR_APPR_PROCESS_START	-0.0014		-0.032	-0.031	-0.038	-0.025	0.046	0.045	0.012	1	-0.019
AMT_APPLICATION	-0.0039		0.99	0.82	0.7	0.98	-0.00048	-0.12	0.0057	-0.019	1

Inference:- 'AMT_APPLICATION' is directly correlated with 'AMT_CREDIT', 'AMT_GOODS_PRICE' and highly correlated to 'AMT_ANNUITY', 'CNT_PAYMENT.'CNT_PAYMENT' is equally (moderately) correlated to 'AMT_GOODS_PRICE', 'AMT_CREDIT' and 'AMT_APPLICATION'.Bivariate Analysis of Numerical variables dont show any significant different trends among Non-Defaulter and Defaulters.

Multivariate Analysis- Categorical-Numerical var for TARGET var. Previous Application



Inference:- Highest credit amount was received for 'Buying a new car' and 'Buying New Home' among Cash loans.Lowest credit amount was received for 'Purchase of electronic equipment' and 'Everyday Expenses'.Everyday expenses, building house or annex, purchase of electronic equipment and Furniture show slighlty higher Credit amount received for cash loan purpose for Defaulters compared to non-defaulters.

Conclusions of Previous Application Data Frame

- There were 15 Categorical columns and 9 Numerical columns.
- There were 1420436 clients.
- The data was higjly imbalanced, ~90% were Non-Defaulters and ~10% were Defaulters.
- 'NAME_CONTRACT_STATUS' is an important variable, majority of clients whose previous loan was approved were Non-defaulters, but intrestingly many people who were refused loan in previous application were Non-Defaulters. Bank should also focus on it's policies to avoid Type I error i.e. False Positive- Wrongly considering a cllient as Defaulter.
- Boxplots of Defaulters and Non-defaulters reveal some significant difference across numerical columns:-

'AMT_GOODS_PRICE' , 'AMT_CREDIT', 'AMT_APPLICATION' have higher distribution value for Defaulters based on their interquartile size range. ¶

- Segmented Univariate Analysis of Categorical variables showed most opted sub-category within that variable via count plot and Default % bar plot showed the sub-category with highest default %.

Examples:-(with significant Default %)

- Cash loan Purpose -1) Refusal to give reason, 2) Money for third Person, 3) Hobby; all ~20%, such loans should be given at higher intrest rate or lower credit amount or simply refused.
- Remaining Default % are within range of 7%~12%, for such sub-categories, those with highest Default% should be given loan at higher intrest rate or lower credit amount.

Multivariate Analyses helped us reveal peculiar characteristics of Defaulters.

- Multivariate Analyses revealed that some Numerical columns are highly correlated, i.e.
- 'AMT_APPLICATION' is directly correlated with 'AMT_CREDIT', 'AMT_GOODS_PRICE' and highly correlated to 'AMT_ANNUITY', 'CNT_PAYMENT.
- 'CNT_PAYMENT' is equally (moderately) correlated to 'AMT_GOODS_PRICE', 'AMT_CREDIT' and 'AMT_APPLICATION'.

There are few goods categories where Defaulters received loans for higher price for same goods than Non-Defaulters; i.e. 'Additional Service' and 'Direct Sales'- based on IQR and also 'Education'- based on Median value

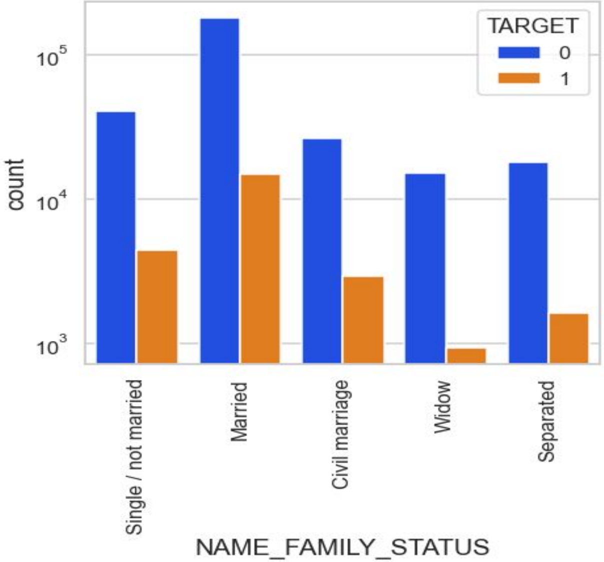
- This gives us an idea on how to treat the client asking for higher credit amount than others, i.e. possible Defaulter.

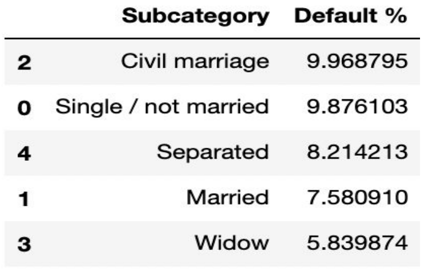
Among Goods categories, clients who received loans for higher goods amounts for 'Direct Sales', 'Additional Service' and 'Education' were Defaulters.

Segmented Univariate Analysis

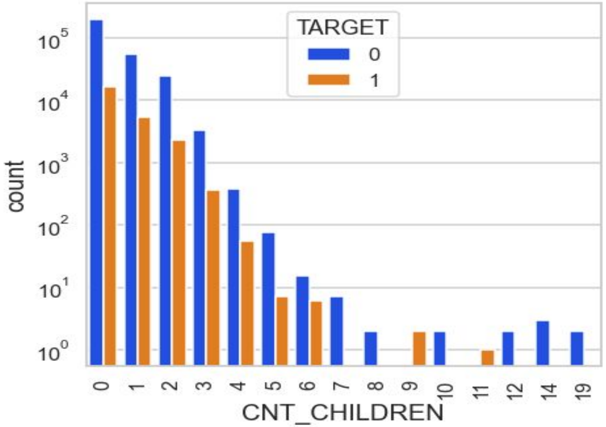
Application

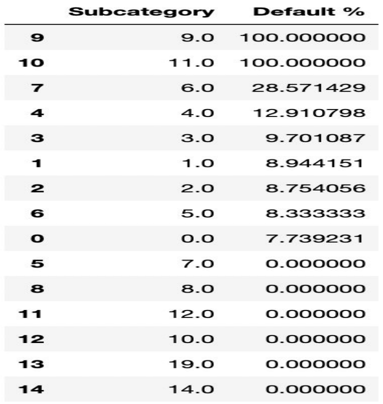
- Such segmented Univariate analysis were conducted for Other Columns as well within Application data frame.



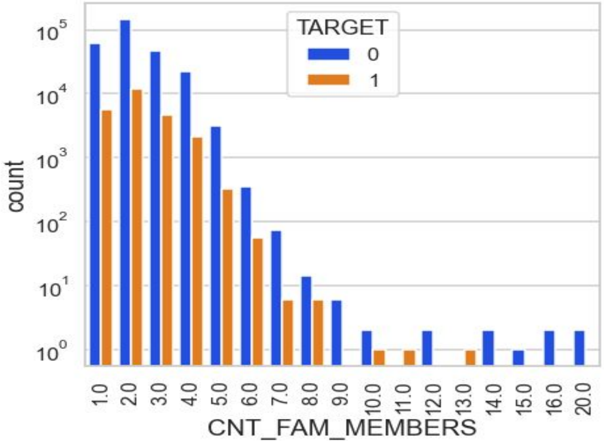


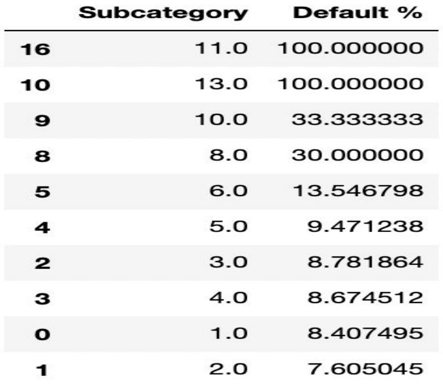
	Subcategory	Default %
2	Civil marriage	9.968795
0	Single / not married	9.876103
4	Separated	8.214213
1	Married	7.580910
3	Widow	5.839874





	Subcategory	Default %
9	9.0	100.000000
10	11.0	100.000000
7	6.0	28.571429
4	4.0	12.910798
3	3.0	9.701087
1	1.0	8.944151
2	2.0	8.754056
6	5.0	8.333333
0	0.0	7.739231
5	7.0	0.000000
8	8.0	0.000000
11	12.0	0.000000
12	10.0	0.000000
13	19.0	0.000000
14	14.0	0.000000

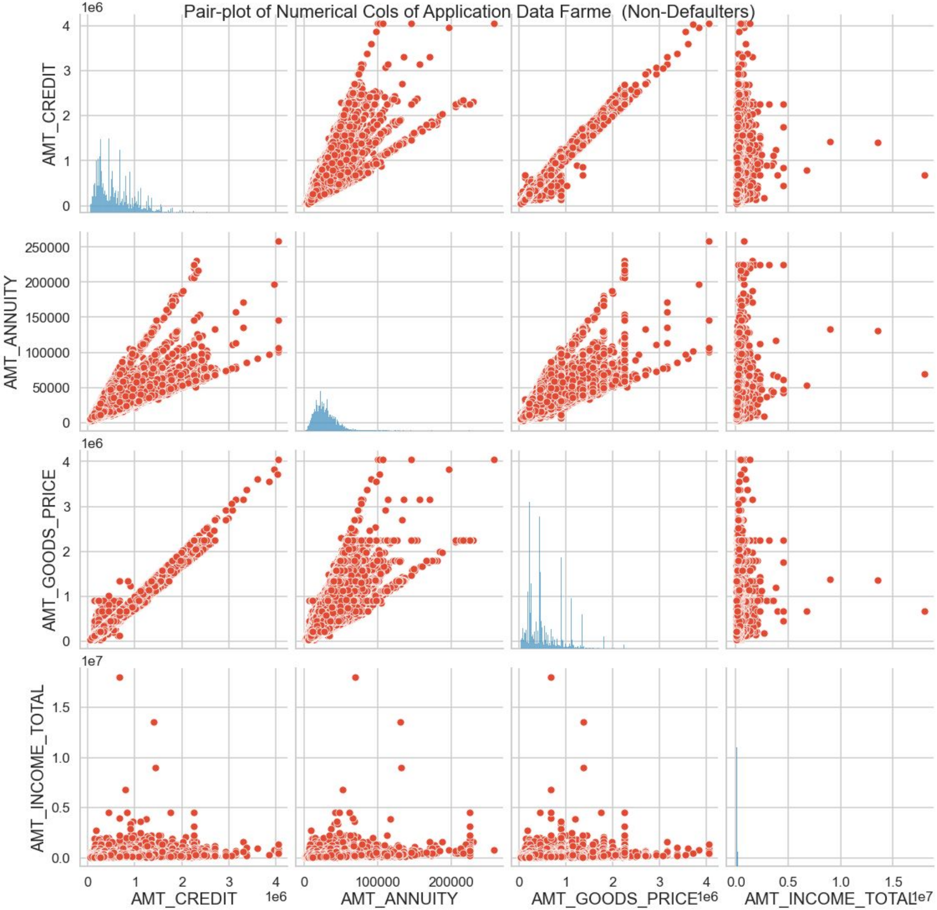




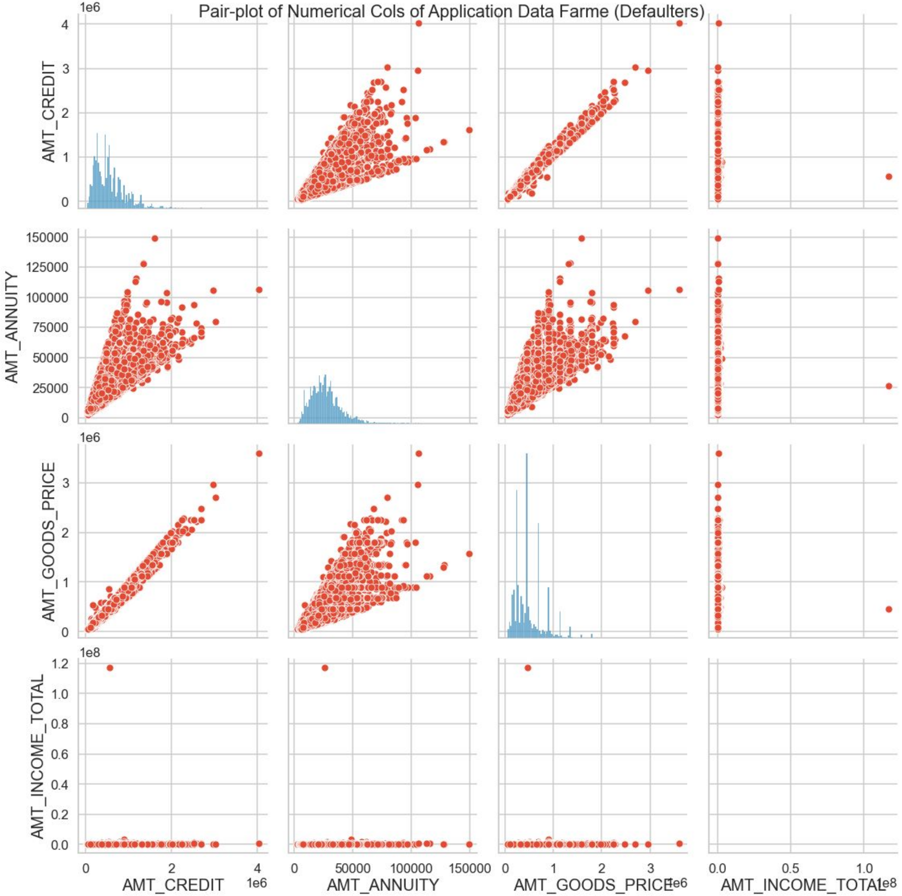
	Subcategory	Default %
16	11.0	100.000000
10	13.0	100.000000
9	10.0	33.333333
8	8.0	30.000000
5	6.0	13.546798
4	5.0	9.471238
2	3.0	8.781864
3	4.0	8.674512
0	1.0	8.407495
1	2.0	7.605045

Inference:-Majority Clients are Married (not civil marriage) or Single, and with low number of Family members and children.Greater the no. of children and Family members greater is the Default %.

Correlation among numnerical cols for Non-Defaulters and Defaulters - Application



Pair-plot of Numerical Cols of Application Data Farme (Non-Defaulters)



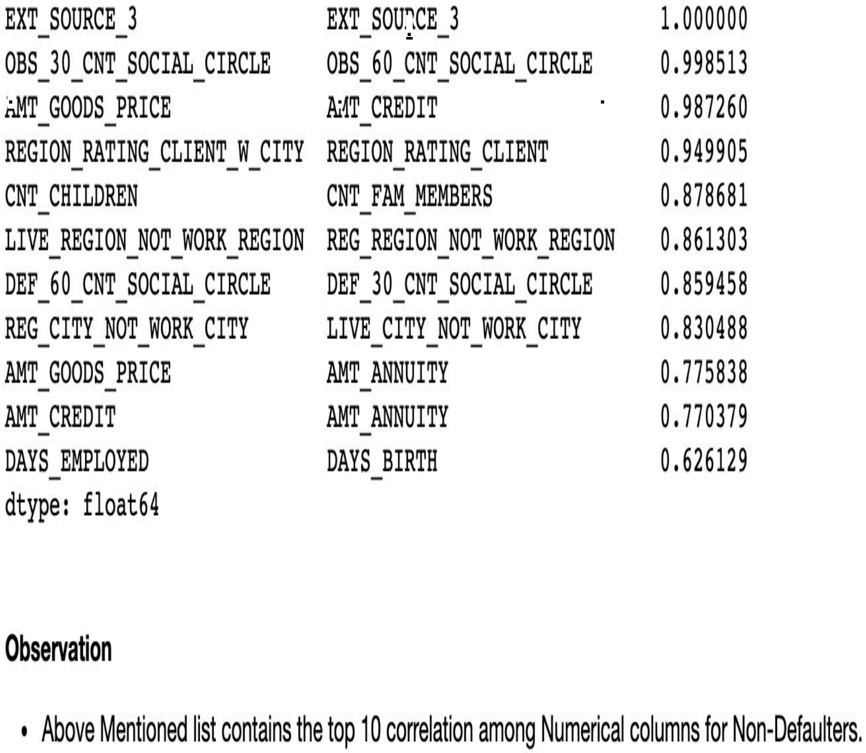
Pair-plot of Numerical Cols of Application Data Farme (Defaulters)

Inference:-'AMT_CREDIT' and 'AMT_GOODS_PRICE' are directly correlated with each other.'AMT_ANNUITY' is highly corelated with 'AMT_CREDIT' and 'AMT_GOODS_PRICE'.'AMT_INCOME_TOTAL doesnt show any significant correlation with either of these:- 'AMT_CREDIT', 'AMT_GOODS_PRICE', 'AMT_ANNUITY'Following pair-plot shows similar trend for the above mentioned Numerical cols for both Non-Defaulter and Defaulters.

Bivariate Analysis - Application

Non-Defaulters

Numerical var Numerical var Correlation Value



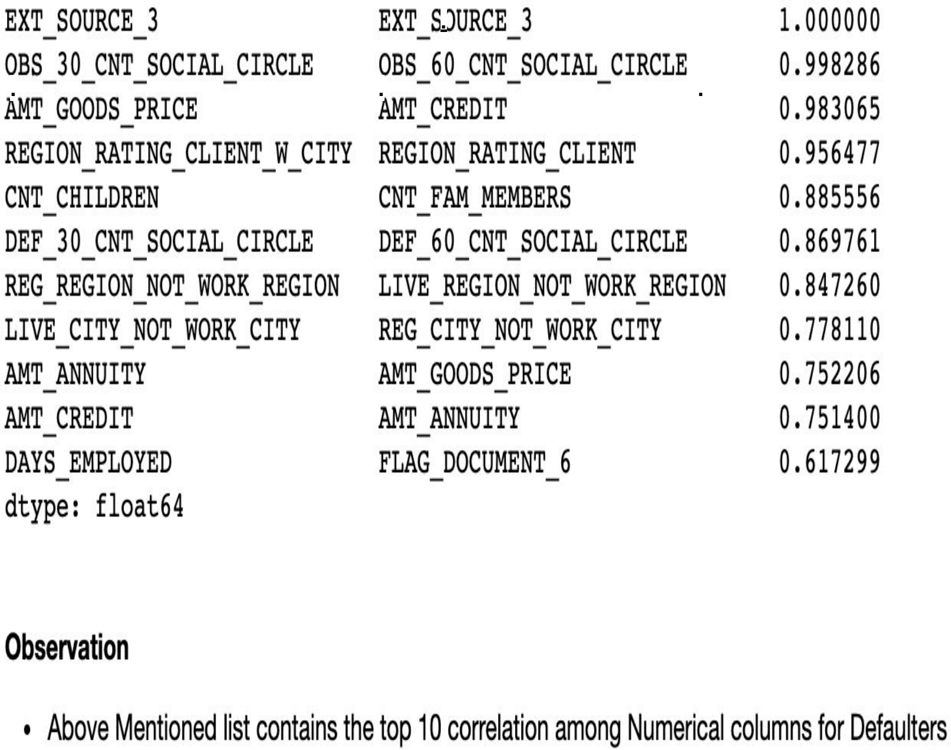
```
EXT_SOURCE_3                 EXT_SOURCE_3                 1.000000
OBS_30_CNT_SOCIAL_CIRCLE     OBS_60_CNT_SOCIAL_CIRCLE     0.998513
AMT_GOODS_PRICE              AMT_CREDIT                   0.987260
REGION_RATING_CLIENT_W_CITY  REGION_RATING_CLIENT         0.949905
CNT_CHILDREN                 CNT_FAM_MEMBERS              0.878681
LIVE_REGION_NOT_WORK_REGION  REG_REGION_NOT_WORK_REGION   0.861303
DEF_60_CNT_SOCIAL_CIRCLE     DEF_30_CNT_SOCIAL_CIRCLE     0.859458
REG_CITY_NOT_WORK_CITY       LIVE_CITY_NOT_WORK_CITY      0.830488
AMT_GOODS_PRICE              AMT_ANNUITY                  0.775838
AMT_CREDIT                   AMT_ANNUITY                  0.770379
DAYS_EMPLOYED                DAYS_BIRTH                   0.626129
dtype: float64
```

Observation

- Above Mentioned list contains the top 10 correlation among Numerical columns for Non-Defaulters.

Defaulters

Numerical var Numerical var Correlation Value



```
EXT_SOURCE_3                 EXT_SOURCE_3                 1.000000
OBS_30_CNT_SOCIAL_CIRCLE     OBS_60_CNT_SOCIAL_CIRCLE     0.998286
AMT_GOODS_PRICE              AMT_CREDIT                   0.983065
REGION_RATING_CLIENT_W_CITY  REGION_RATING_CLIENT         0.956477
CNT_CHILDREN                 CNT_FAM_MEMBERS              0.885556
DEF_30_CNT_SOCIAL_CIRCLE     DEF_60_CNT_SOCIAL_CIRCLE     0.869761
REG_REGION_NOT_WORK_REGION   LIVE_REGION_NOT_WORK_REGION  0.847260
LIVE_CITY_NOT_WORK_CITY      REG_CITY_NOT_WORK_CITY       0.778110
AMT_ANNUITY                  AMT_GOODS_PRICE              0.752206
AMT_CREDIT                   AMT_ANNUITY                  0.751400
DAYS_EMPLOYED                FLAG_DOCUMENT_6              0.617299
dtype: float64
```

Observation

- Above Mentioned list contains the top 10 correlation among Numerical columns for Defaulters

Summary for Application Data Frame

- The Data was highly imabalanced.
- Data Cleaning was very similar to Previous Application Data Frame.

Few columns such as Age, Education Type, Gender, Marrital Status, Family Members, Occupation Type showed remarkable trend for Defaulter % for specific sub-categories and/or overall variable distribution in comparison to Non-defaulters.

- For this Application a list of Default % was utilised to identify sub-categories having highest Default % instead of barplot which was used in Previous Application, because the no. of columns were higher.
- For certain columns/variables which showed showed similar trend for Defaulters and Non-Defaulters, they are ignored as they dont provide any valuable information.
- Default % values for each columns for their Sub-categories have been mentioned at the end of Analysis of those columns.