EDA CASE STUDY

Submitted by:
Jaidip Ghosh
Shubham Kaser

Data Exploration

- Dimension of the Application dataset is **(307511, 122)**
- Dimension of the Previous Application dataset is **(1670214, 37)**
- Out of **122** columns, **67** columns have null values for Application Dataset
- Out of **307511** rows, **298909** rows have null values for Application Dataset
- Out of **37** columns, **16** columns have null values for Previous Application Dataset
- Out of **1670214** rows, **1670143** rows have null values for Previous Application Dataset

Problem Statement: Use EDA to understand how consumer attributes and loan attributes influence the tendency of defaulter.

Data Cleaning and Manipulation

Possible Data inconsistencies:

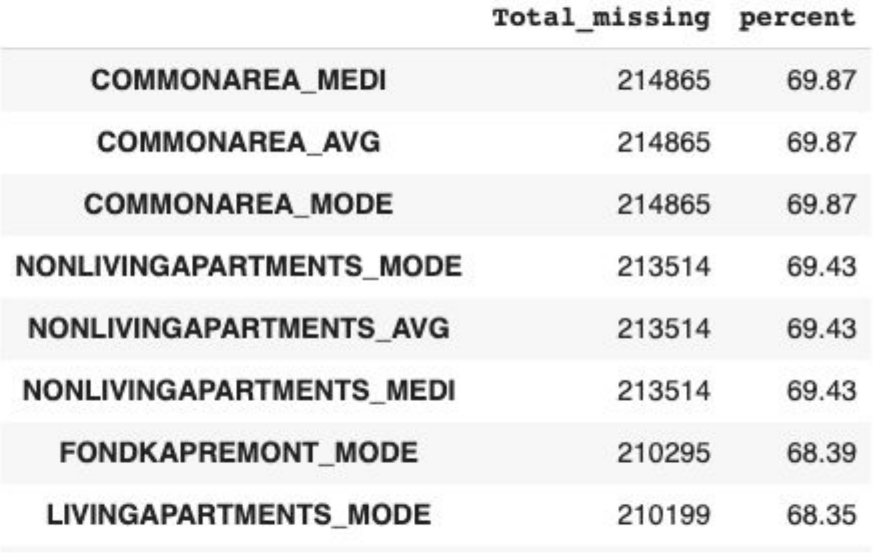
- NaN values in both the datasets
- Duplicate rows of same IDs
- Unacceptable number of outliers

Other Issues:

- Many numerical variables needs to be converted to categorical variables
- Normalization of several variables as the scale was huge due to wide spread of data

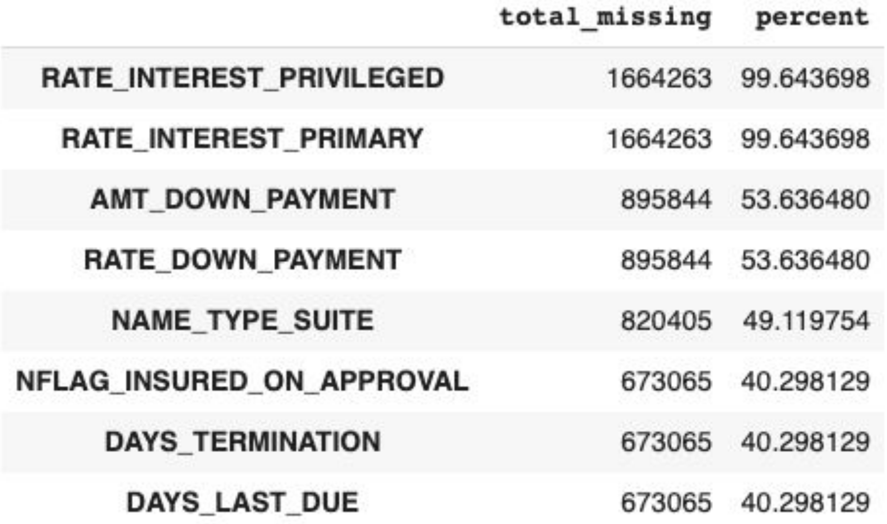
Variables with large fraction of NaN values

Top few Missing values of the dataset
‘application_data’



	Total_missing	percent
COMMONAREA_MEDI	214865	69.87
COMMONAREA_AVG	214865	69.87
COMMONAREA_MODE	214865	69.87
NONLIVINGAPARTMENTS_MODE	213514	69.43
NONLIVINGAPARTMENTS_AVG	213514	69.43
NONLIVINGAPARTMENTS_MEDI	213514	69.43
FONDKAPREMONT_MODE	210295	68.39
LIVINGAPARTMENTS_MODE	210199	68.35

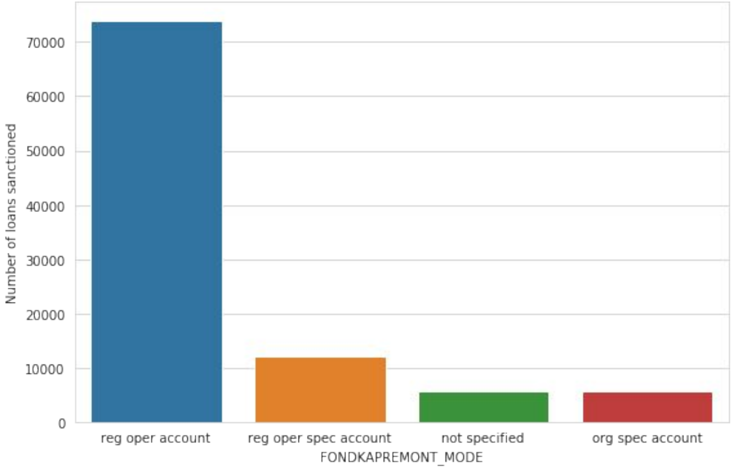
Top few Missing values of the dataset
‘prev_application_data’

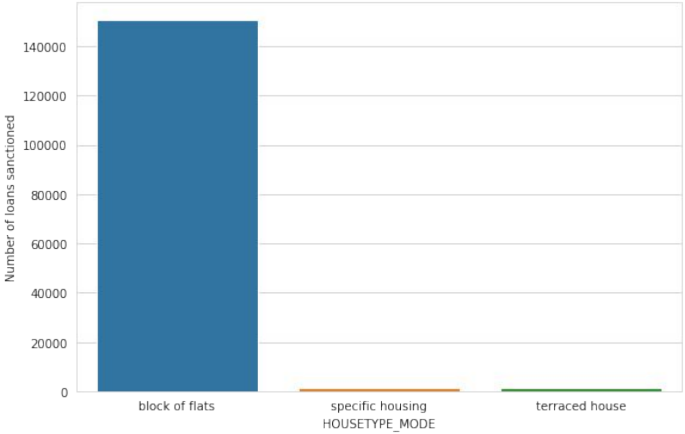


	total_missing	percent
RATE_INTEREST_PRIVILEGED	1664263	99.643698
RATE_INTEREST_PRIMARY	1664263	99.643698
AMT_DOWN_PAYMENT	895844	53.636480
RATE_DOWN_PAYMENT	895844	53.636480
NAME_TYPE_SUITE	820405	49.119754
NFLAG_INSURED_ON_APPROVAL	673065	40.298129
DAYS_TERMINATION	673065	40.298129
DAYS_LAST_DUE	673065	40.298129

Why we dropped several columns?

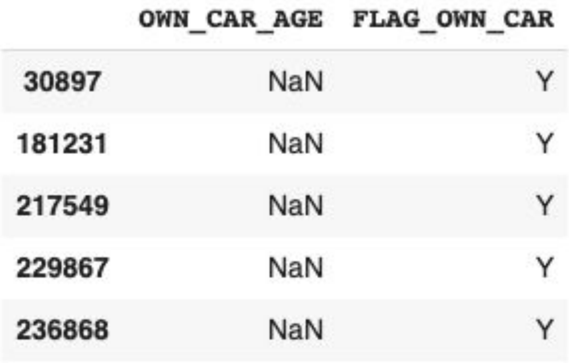
1. Number of variables that had missing values > 50% were found to be 49 whereas variables having missing values > 20% were found to be 50. Number of variables with more than 20% missing values were selected for data imbalance analysis:





The variables with more than 20% missing values found to have data imbalance among the different categories that led the variables to be dropped.

The variable ‘OWN_CAR_AGE’ found to have some interesting error that made it to be dropped. It had NaN for the variable ‘FLAG_OWN_CAR’ with ‘Y’ which implies if a person owns a car, there should have age of the car mentioned in the ‘OWN_CAR_AGE’ variable which was found to be missing instead.



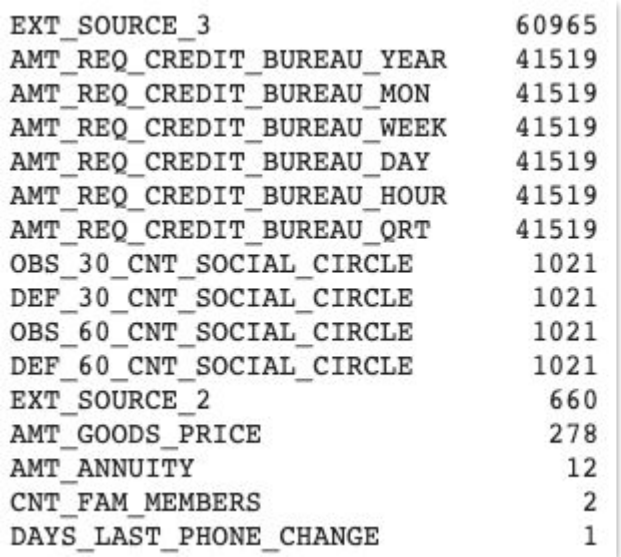
	OWN_CAR_AGE	FLAG_OWN_CAR
30897	NaN	Y
181231	NaN	Y
217549	NaN	Y
229867	NaN	Y
236868	NaN	Y

Missing rows and duplicates

- After dropping columns with missing values, we looked for the rows with more than 20% missing values which resulted to have no such rows existing.
- Then we checked for duplicate values (row wise) in both the datasets and fortunately, no duplicates were found.

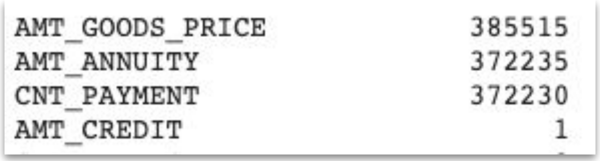
Treating missing values

Count of missing values of variables in ‘application_data’

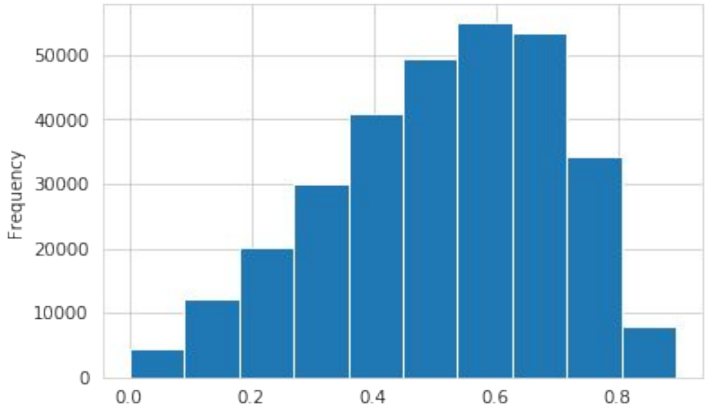


```
EXT_SOURCE_3                  60965
AMT_REQ_CREDIT_BUREAU_YEAR    41519
AMT_REQ_CREDIT_BUREAU_MON     41519
AMT_REQ_CREDIT_BUREAU_WEEK    41519
AMT_REQ_CREDIT_BUREAU_DAY     41519
AMT_REQ_CREDIT_BUREAU_HOUR    41519
AMT_REQ_CREDIT_BUREAU_QRT     41519
OBS_30_CNT_SOCIAL_CIRCLE       1021
DEF_30_CNT_SOCIAL_CIRCLE       1021
OBS_60_CNT_SOCIAL_CIRCLE       1021
DEF_60_CNT_SOCIAL_CIRCLE       1021
EXT_SOURCE_2                    660
AMT_GOODS_PRICE                 278
AMT_ANNUITY                      12
CNT_FAM_MEMBERS                   2
DAYS_LAST_PHONE_CHANGE            1
```

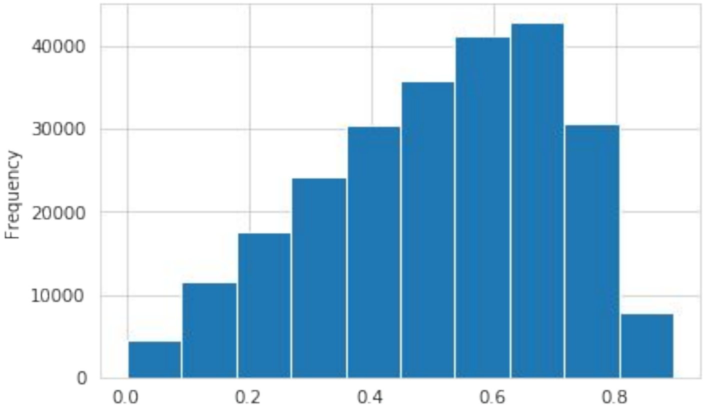
Count of missing values of variables in ‘prev_application_data’

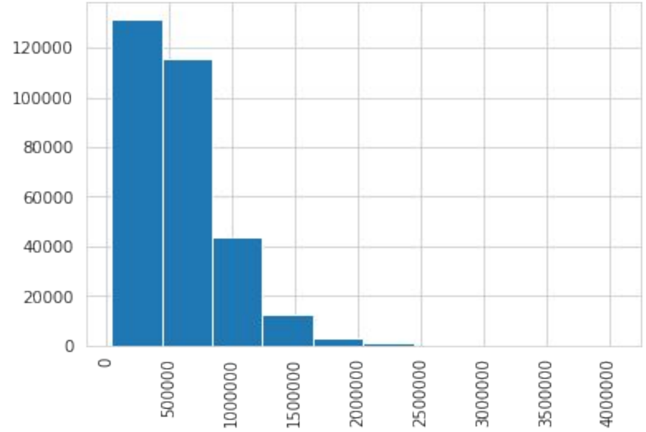


```
AMT_GOODS_PRICE              385515
AMT_ANNUITY                  372235
CNT_PAYMENT                  372230
AMT_CREDIT                        1
```

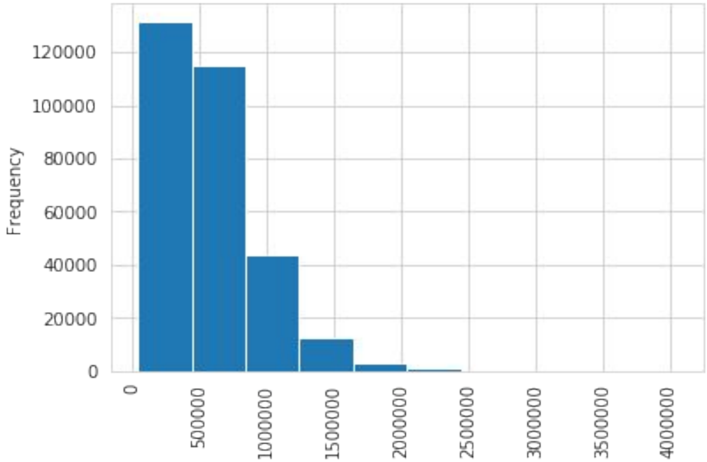


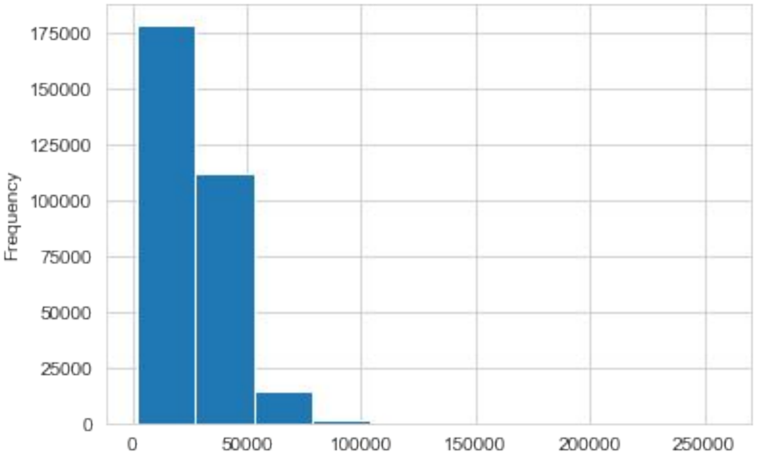
Interpolate



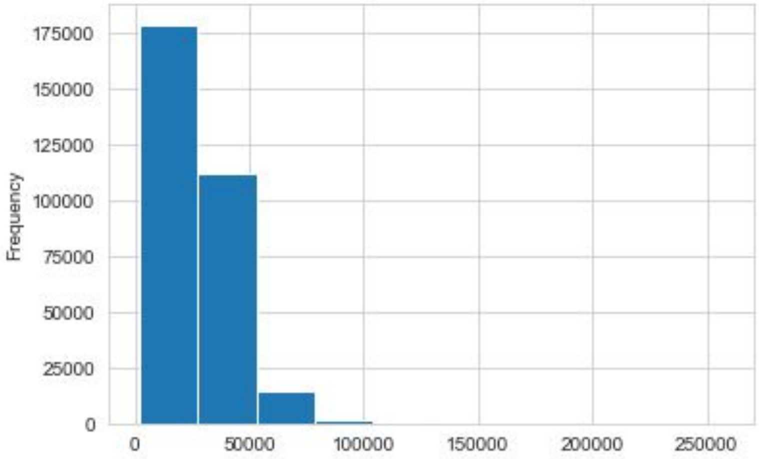


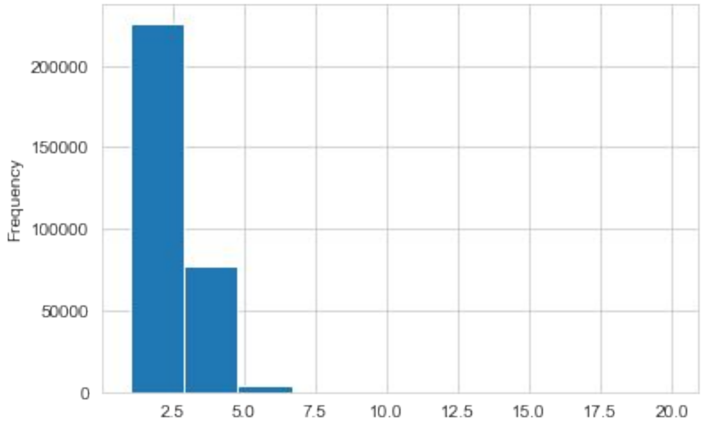
Imputing mean



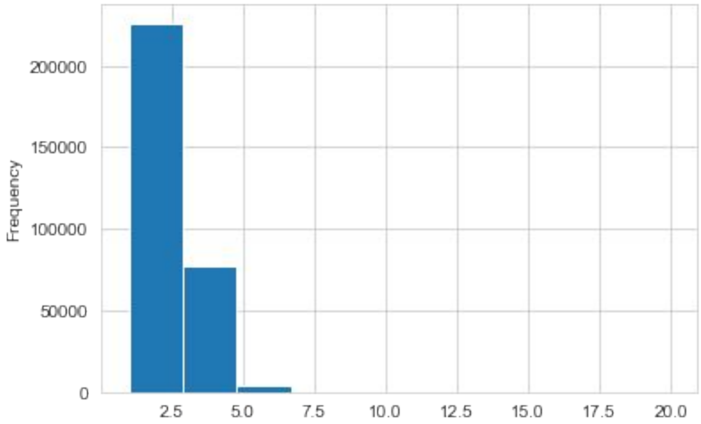

Frequency
175000
150000
125000
100000
75000
50000
25000
0
0
50000
100000
150000
200000
250000

Median


Frequency
175000
150000
125000
100000
75000
50000
25000
0
0
50000
100000
150000
200000
250000


Frequency
200000
150000
100000
50000
0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

Mode

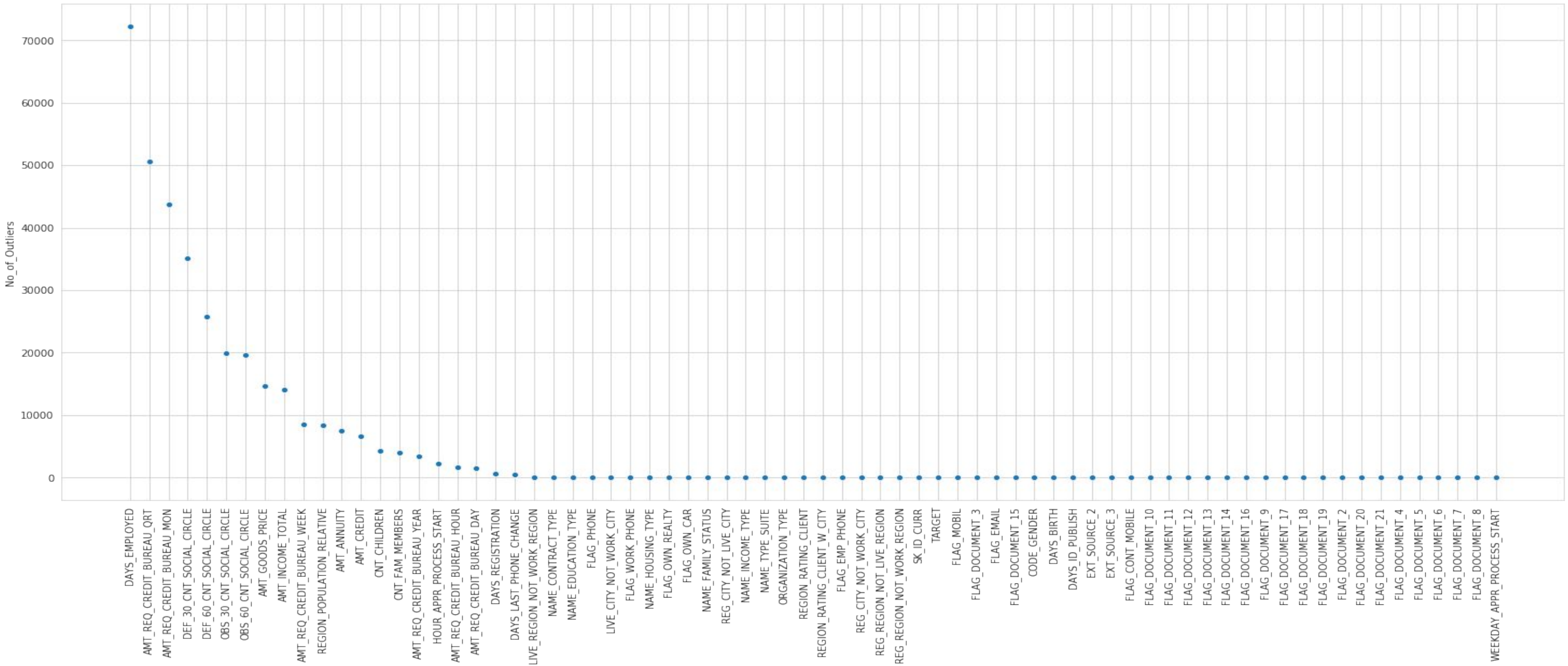

Frequency
200000
150000
100000
50000
0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

Data type conversion

- We converted some of numerical data type into categorical data type.
- *How did we choose specific variables to convert?*

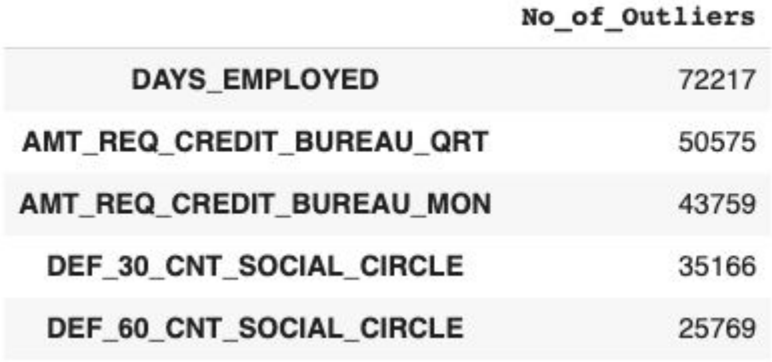
  → We looked for uniques values for each of the numerical variables and the possible categorical variables would be the one with unique values less than 3 (mostly). We filtered those out and converted into categorical variables. Then remaining of the numerical variables still had the possibility to be categorical variable e.g., ‘CNT_CHILDREN’, we analysed them separately and converted them.

Number of Outliers



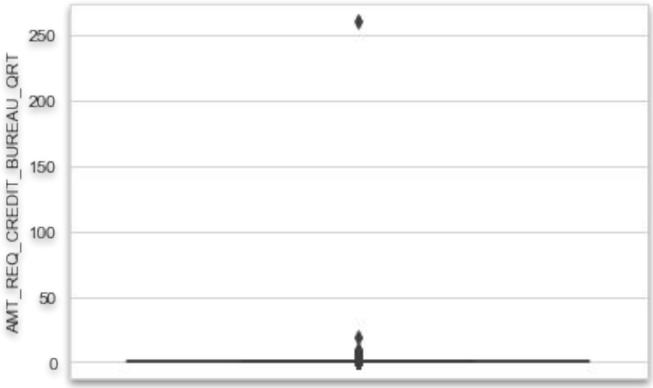
How was the scatter plot for outliers plotted?

- We found Q1, Q4 and IQR for every variables and If any of the point turned out to be beyond the Q1 and Q4 i.e., < Q1 and >Q4, we treated it as outlier and got the total number of outliers for each variables
- The top 5 variables to have the most number of outliers were:

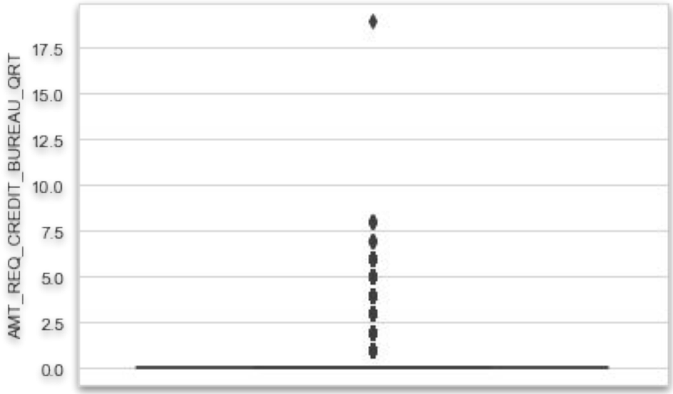


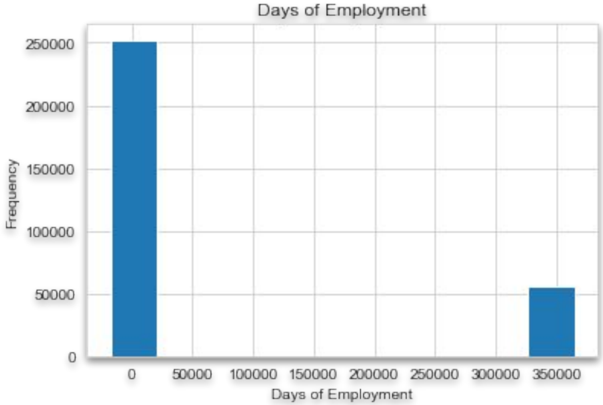
	No_of_Outliers
DAYS_EMPLOYED	72217
AMT_REQ_CREDIT_BUREAU_QRT	50575
AMT_REQ_CREDIT_BUREAU_MON	43759
DEF_30_CNT_SOCIAL_CIRCLE	35166
DEF_60_CNT_SOCIAL_CIRCLE	25769

Treating outliers

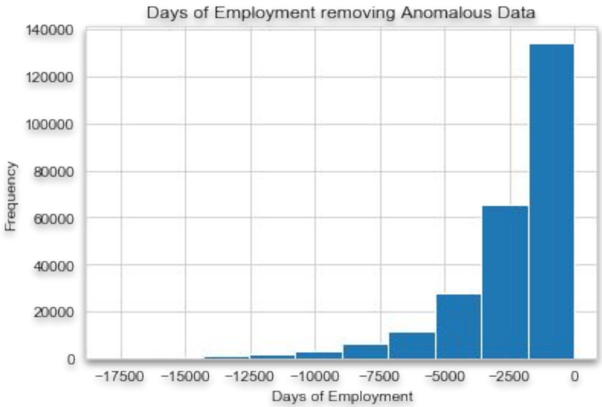


Removing extreme outliers



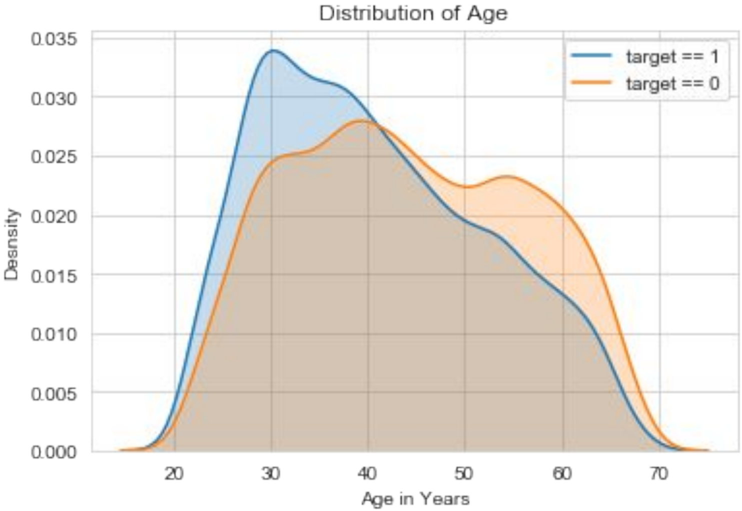


Separating the anomalies in different columns

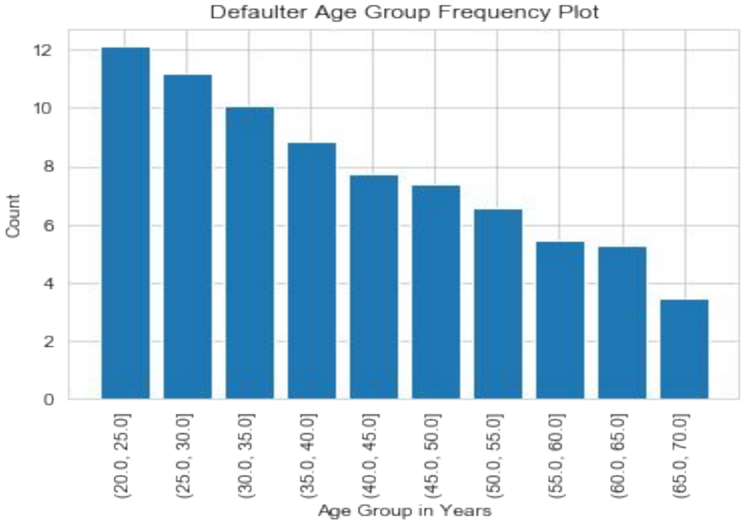


Binning appropriate variables

For the variable ‘DAYS_BIRTH’ , No significant conclusion could be derived from the normal histogram. To visualise effect on the target variable, binning was the appropriate option.



As we see, we can’t really talk about particular age group and their percentage on default.

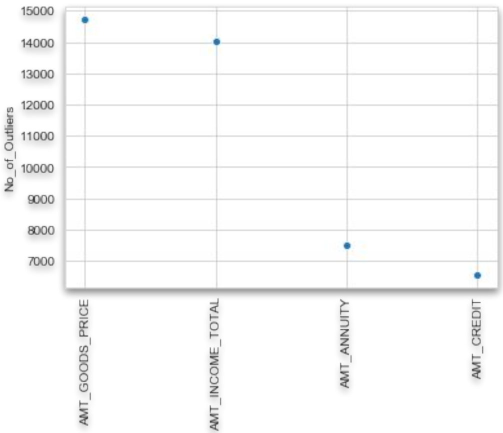


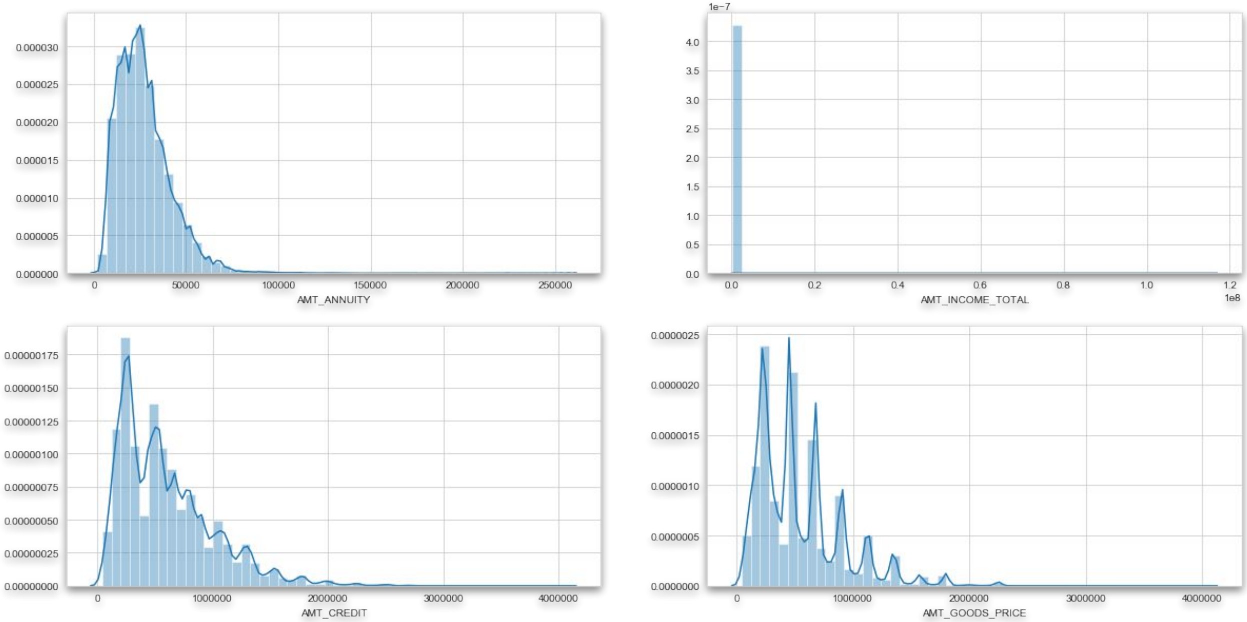
Here, we can see the contribution of particular age group on the list of default

Scaling specific variable

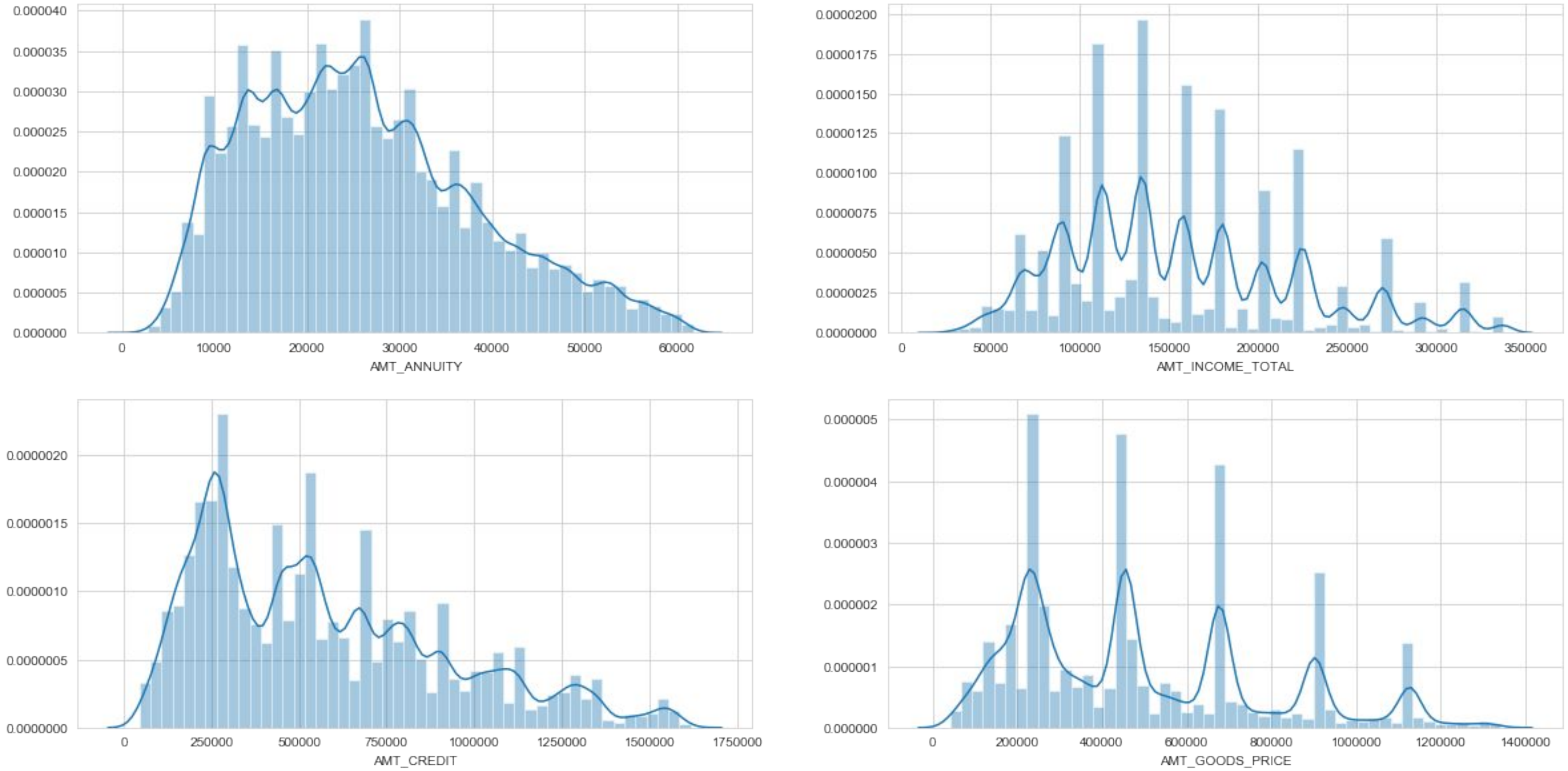
As we see here, the scaling of variables are extremes that has limited the visualisation of normal scale.

Approach here was to find the outliers and simply remove them. Number of outliers for these particular variables are shown below:

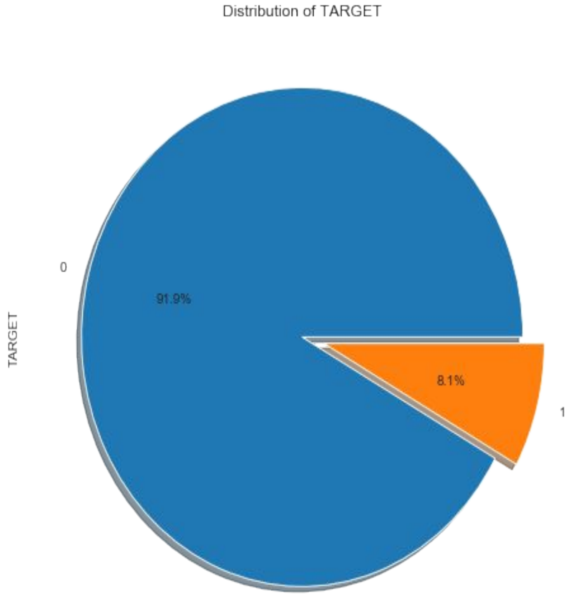


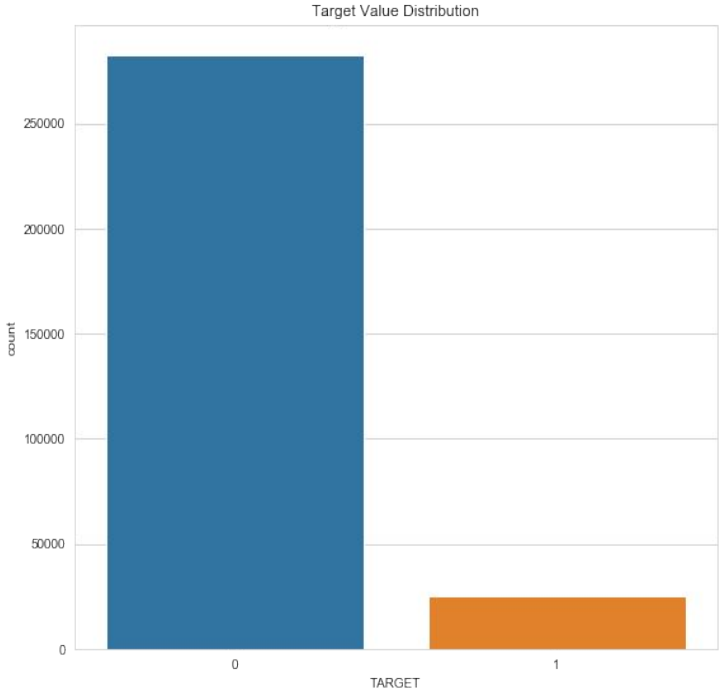


Change in scales after removing outliers can be visualised below:



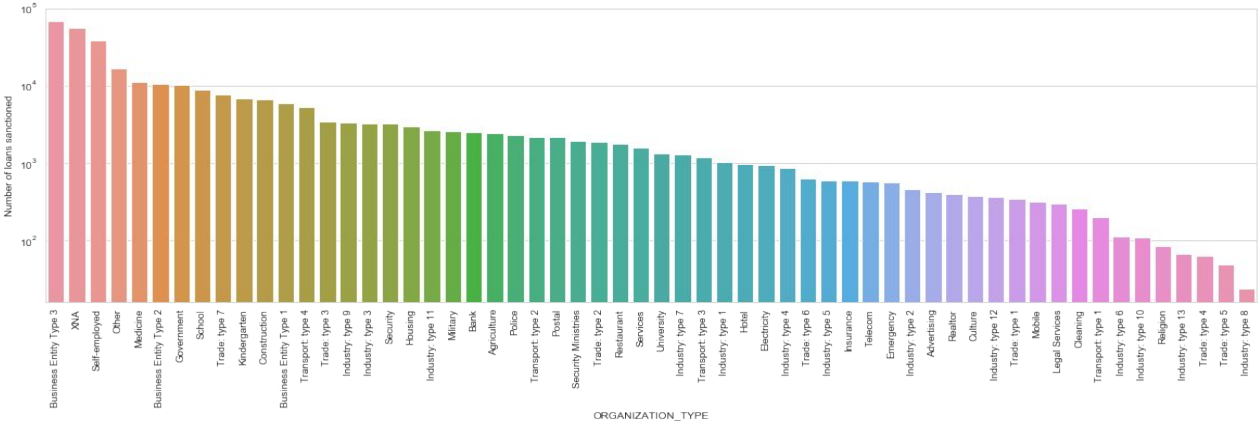
Visualising Data Imbalance of Target Variable

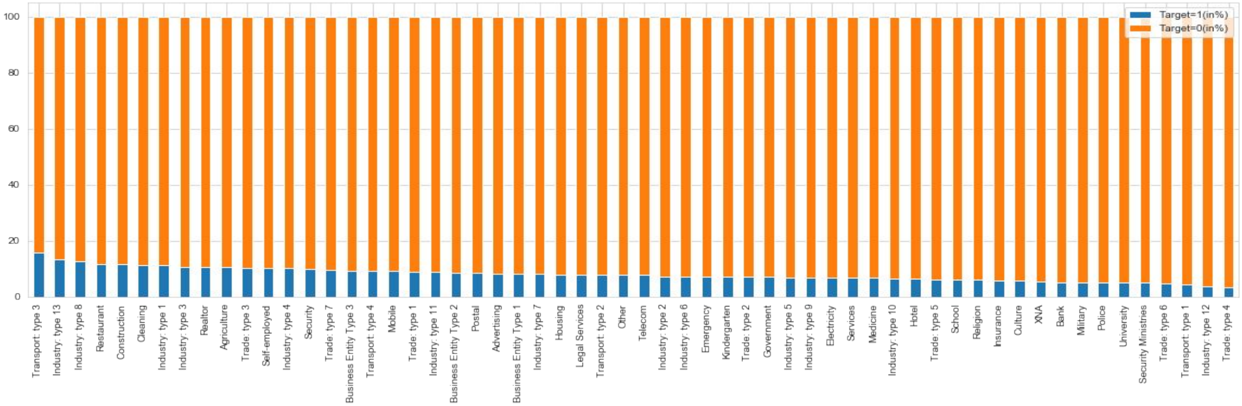




So from the above graph we can clearly view that the data is **imbalanced.** We can say loans that were repaid on time are way more than loans that were not repaid on time. The reason we say the data is imbalanced is that for **0** we have 91% of data and for **1** we have 8% so the data is highly imbalanced.

Univariate Analysis for categorical Variables





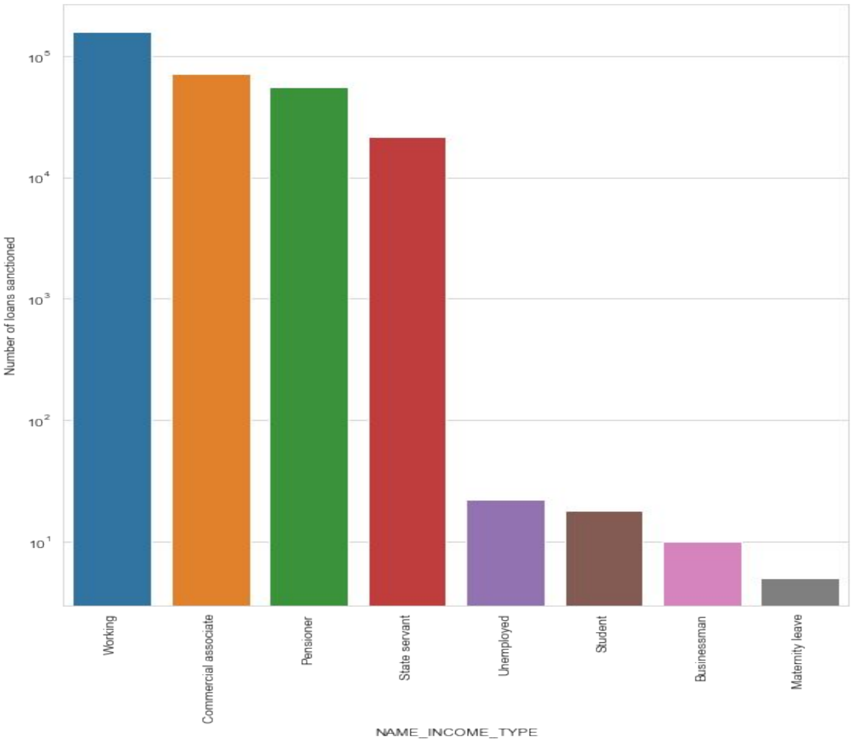
Conclusion:
As per the analysis when we plotted Organization Type with Target Value, we found the top 5 organization type likely to default is

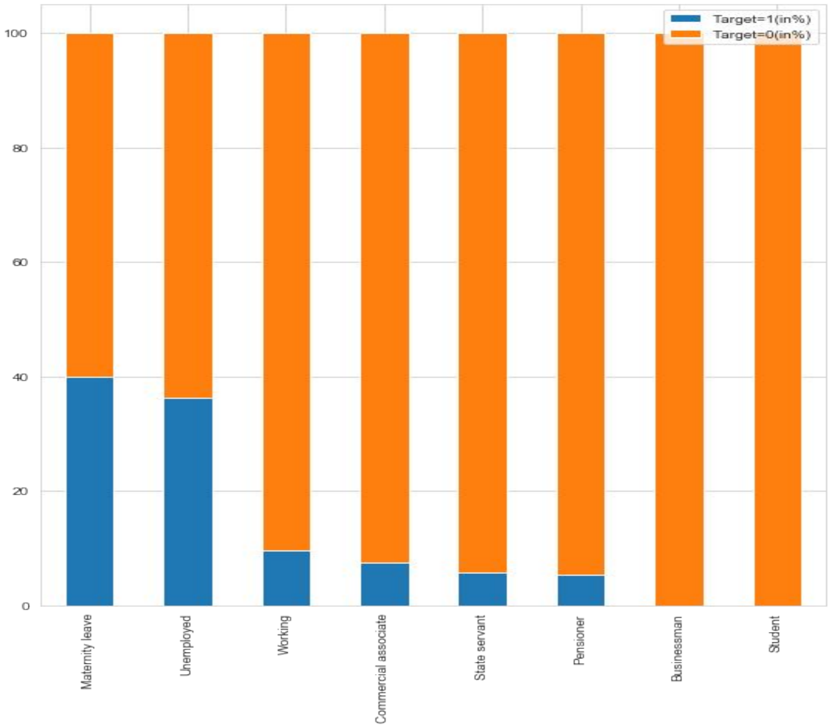
1. Transport: type 3
2. Industry: type 13
3. Industry: type 8
4. Restaurant
5. Construction

And the top 5 organization, where it's least likely to be defaulter

1. Security Ministries
2. Trade: type 6
3. Transport: type 1
4. Industry: type 12
5. Trade: type 4

Analysis of ‘NAME_INCOME_TYPE with target variable

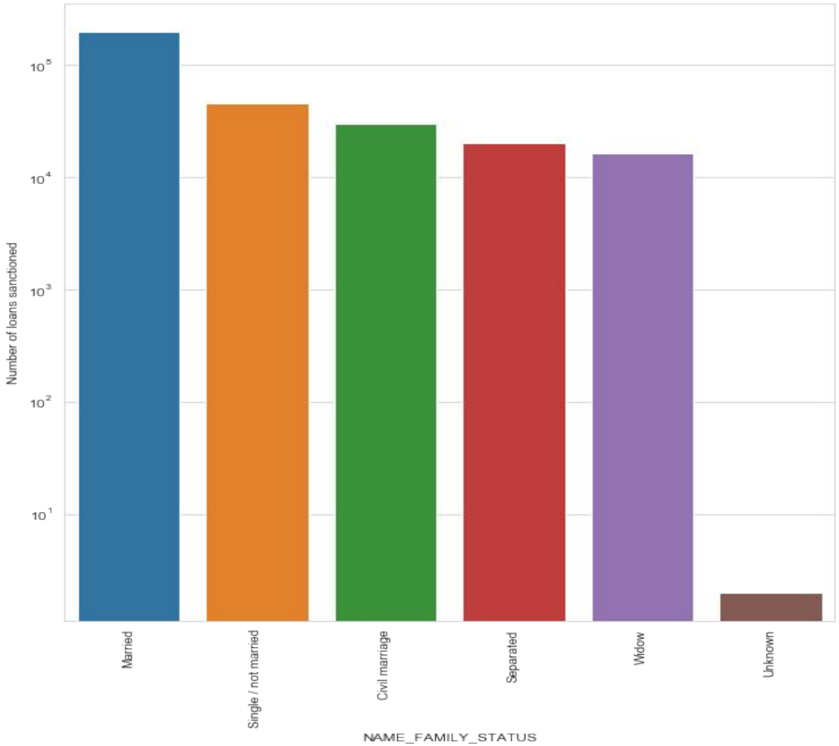


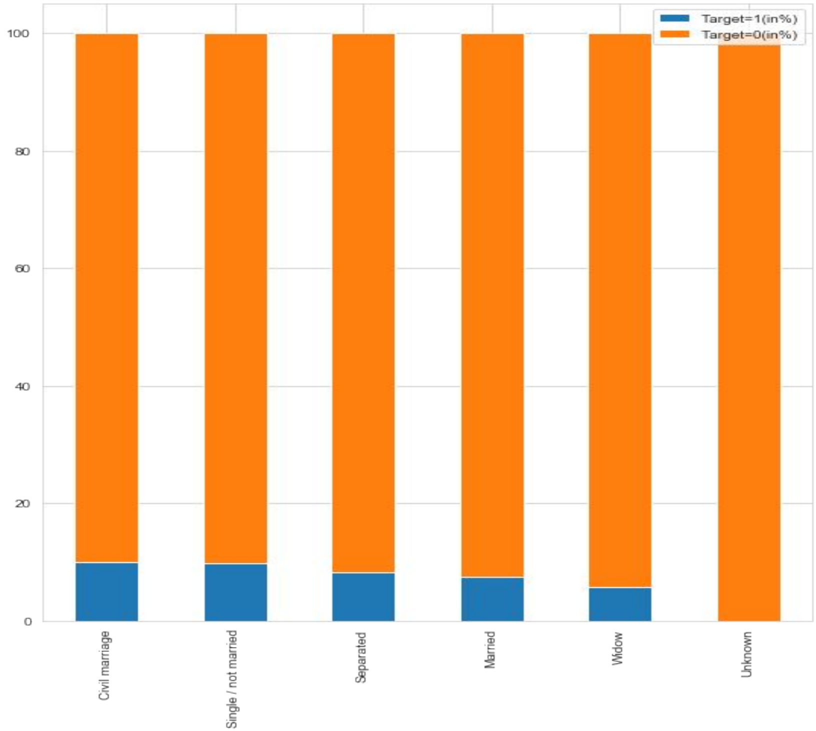


Conclusion: We can clearly see data imbalance here. Although the number of loans sanctioned for **Unemployed** and **Maternity leave** are very less, they are still **more likely to default** if we look into their %age having Target value 1, that are 36% and 40% respectively. While for the **Working** & **Commercial associate** they are less likely to default.

And the most safest to invest in Income Type is **Businessman**, **Student**.

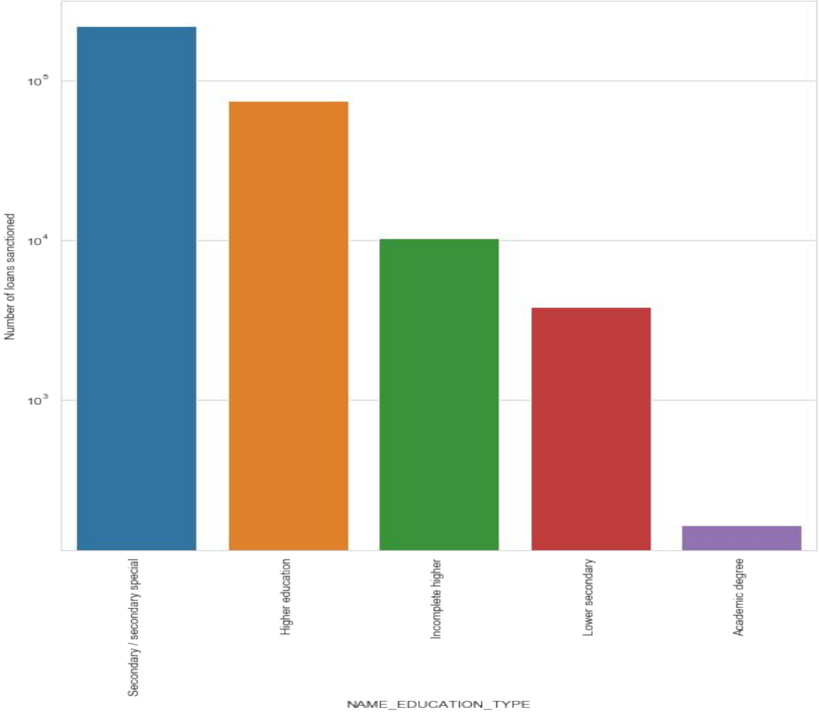
Analysis of ‘NAME_FAMILY_STATUS with target variable

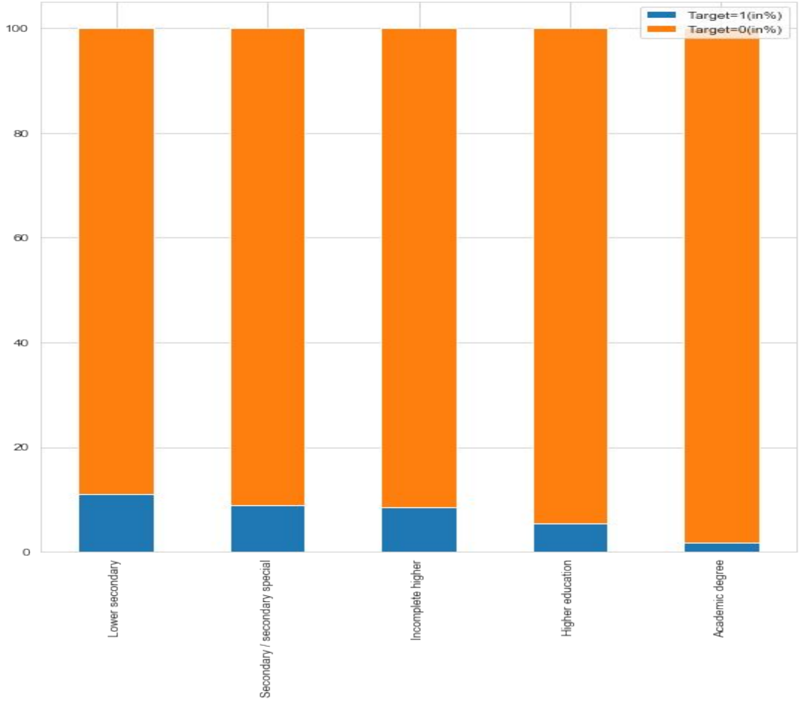




Conclusion: If we skip unknown, data are almost balanced except for Unknown (which is very less), for **Civil Marriage & Single/not married** the chance of **being a defaulter is the most**, ~10%. For married number of the data is so huge, the % of defaulter is equivalent to 7.5% , this leads to **married** people are the **safest to invest for.**

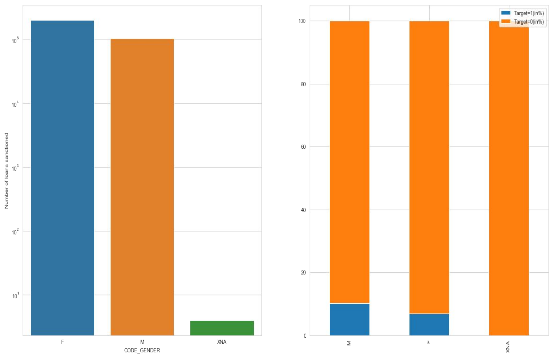
Analysis of ‘NAME_EDUCATION_TYPE with target variable



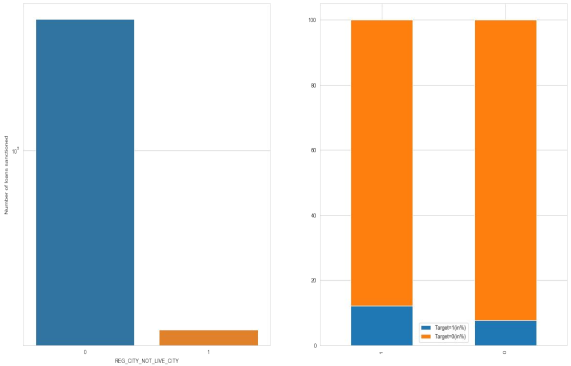


Conclusion: Data is quite imbalanced here. We can say for **higher education**, chances of **default is less** as the number of loan sanctioned is significant with average Target as 5%, but for the high taret average which in this case is **Lower Secondary**, there is a **strong chance** they will be the **defaulter** and for the S**econdary / secondary special** it's not **safe to invest** as it has the highest number of application and it's among top two of the defaulters list

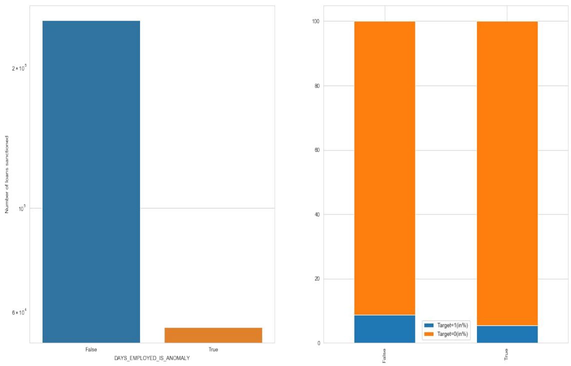
Analysing some more categorical variables with Target variable



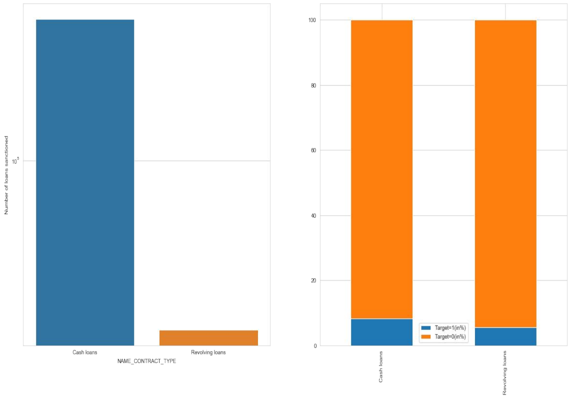
Variable: CODE_GENDER



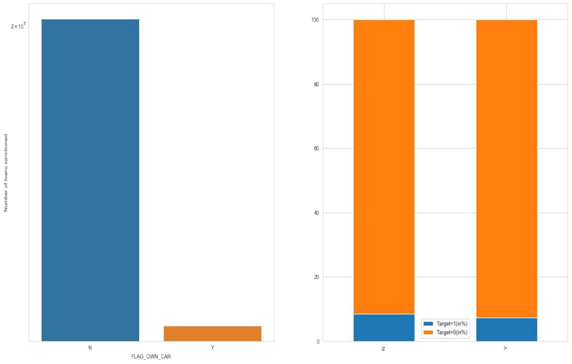
Variable: FLAG_OWN_REALTY



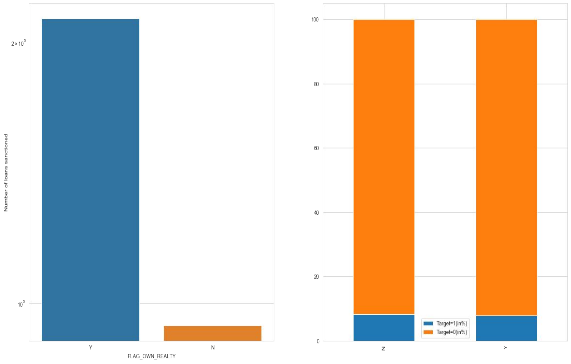
Variable: FALG_OWN_CAR



Variable: NAME_CONTRACT_TYPE

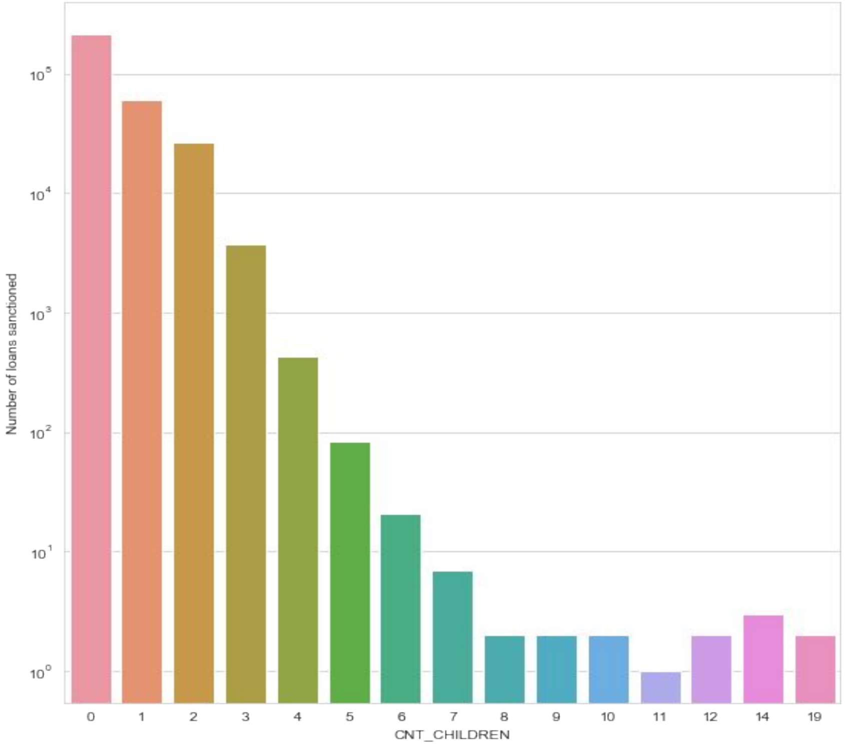


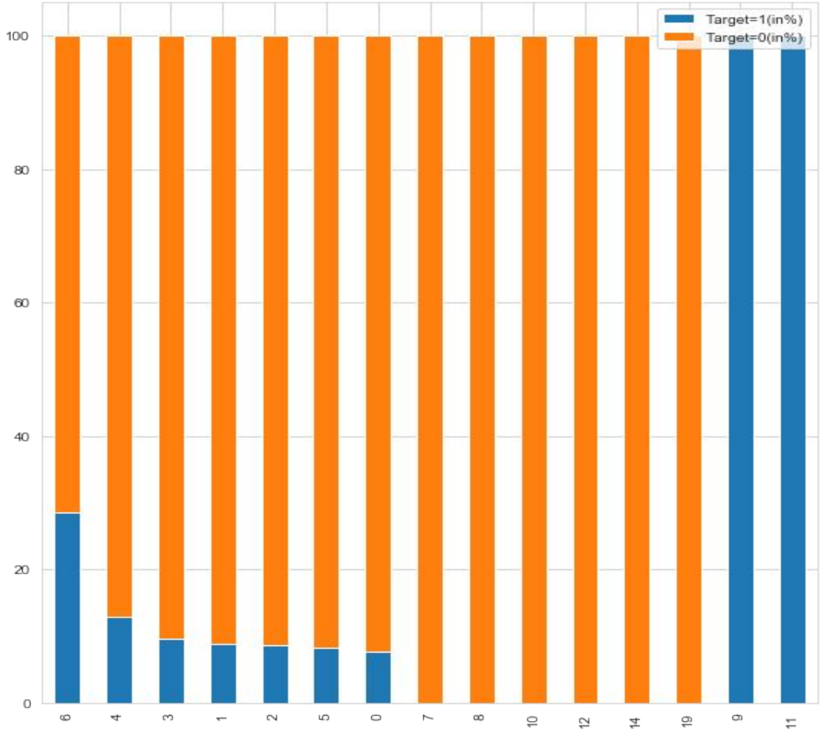
Variable: DAYS_EMPLOYED_IS_ANOMALY



Variable:REG_CITY_NOT_LIVE_CITY

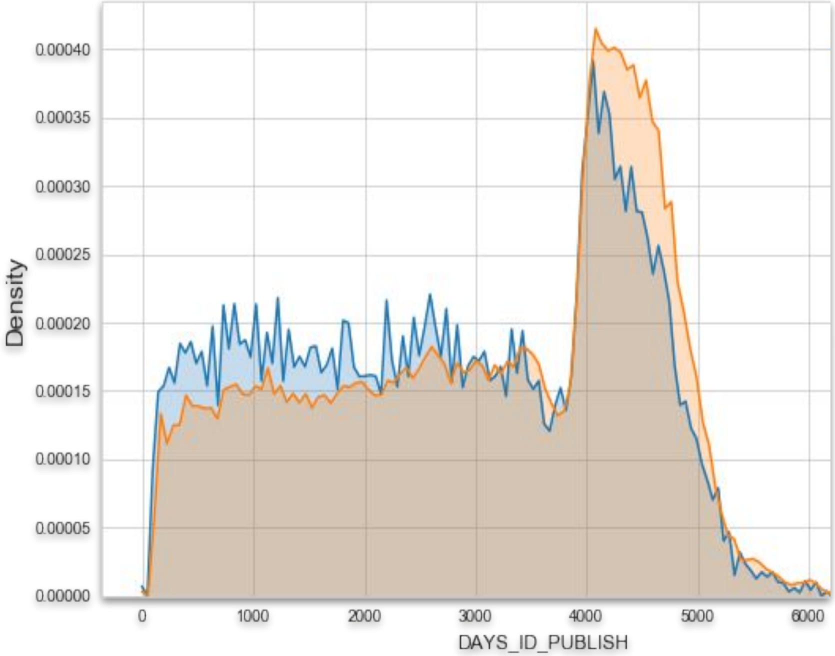
Analysing the ordered categorical variable i.e., numerical variable converted to categorical



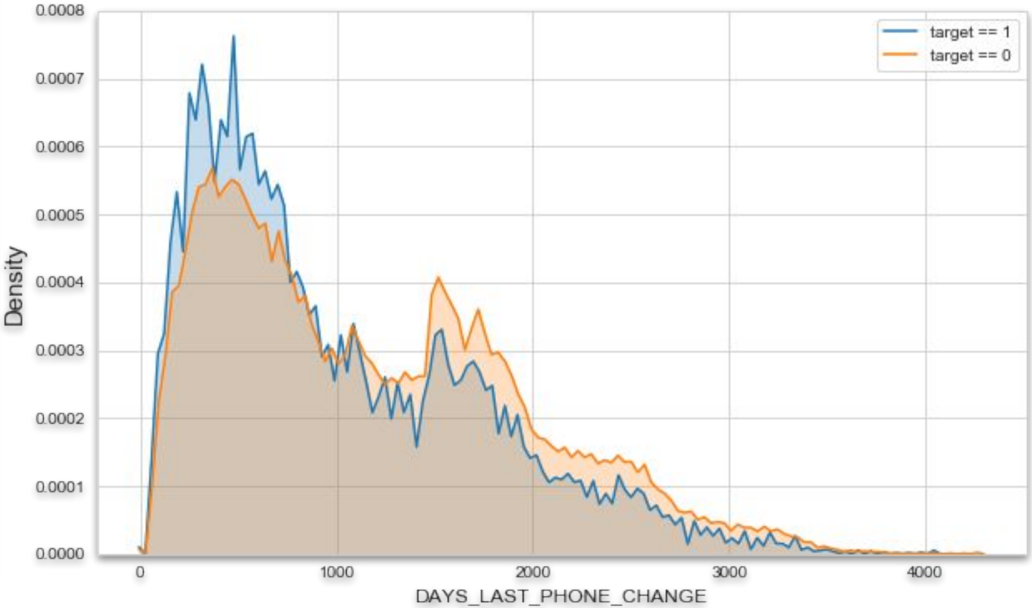


Conclusion: For number of children greater than 6, as there are very less amount of data so we can't conclude whether they will be defaulter or not. Clients having **4 or 6** number of children, they are **most likely to not repay the loan**
Client having **0** children, they are the **safest to give loan to** as they show only ~7% of default rate

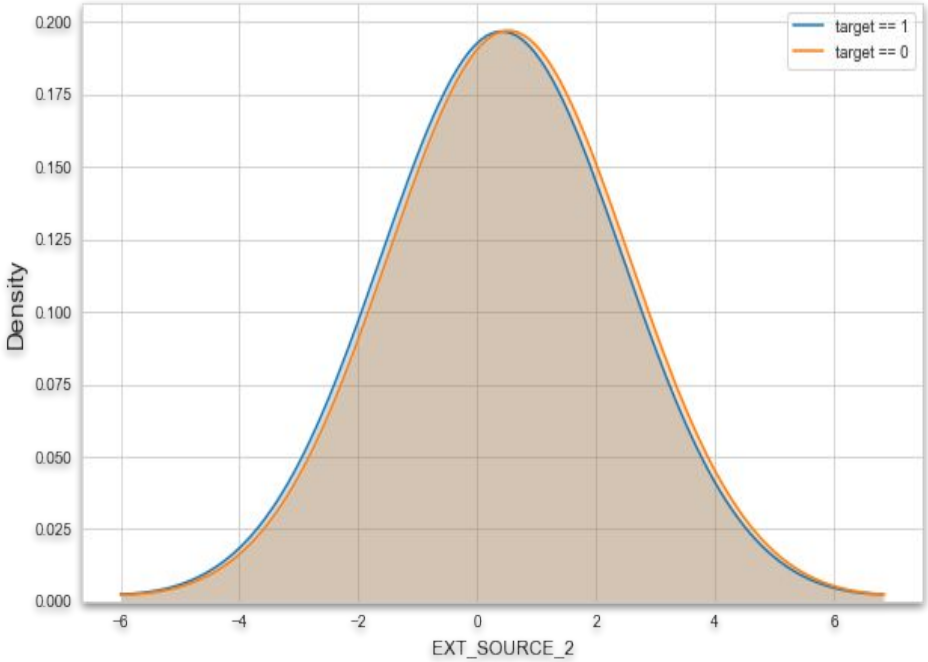
Analysing numerical variables with Target variable

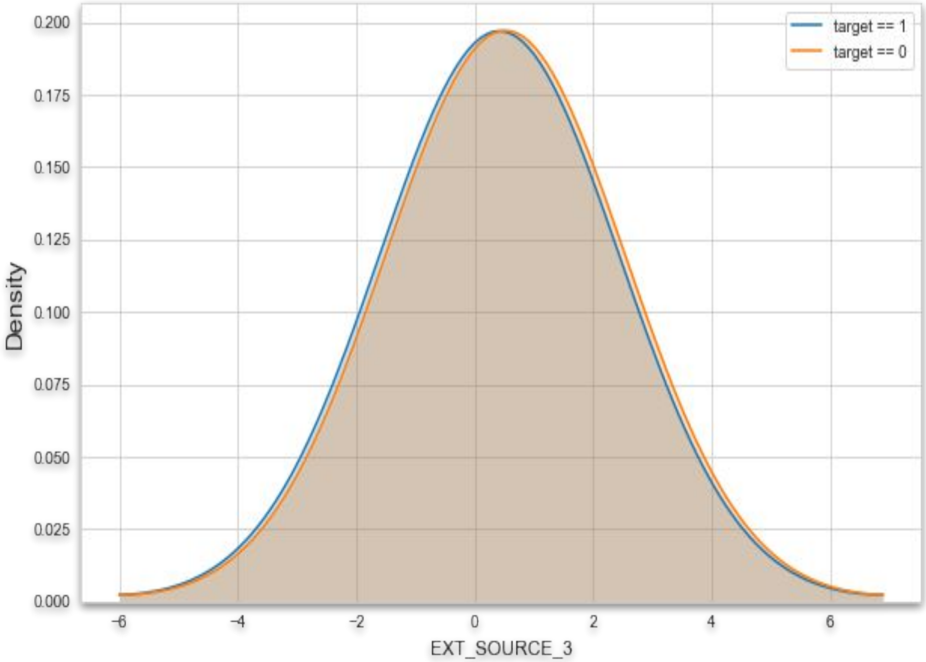


Conclusion: We can see a person who applied for the loan 3000 days before changing the identity document with which he had applied for loan is more likely to default

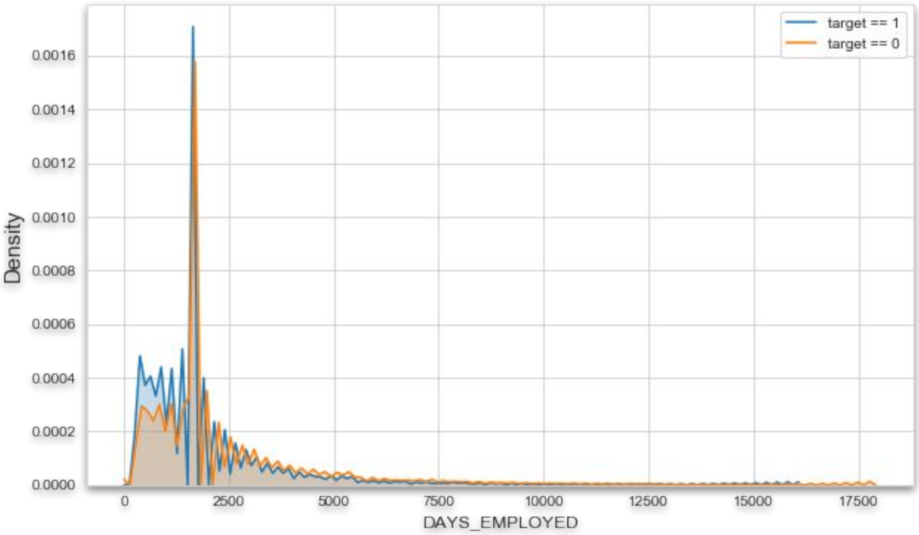


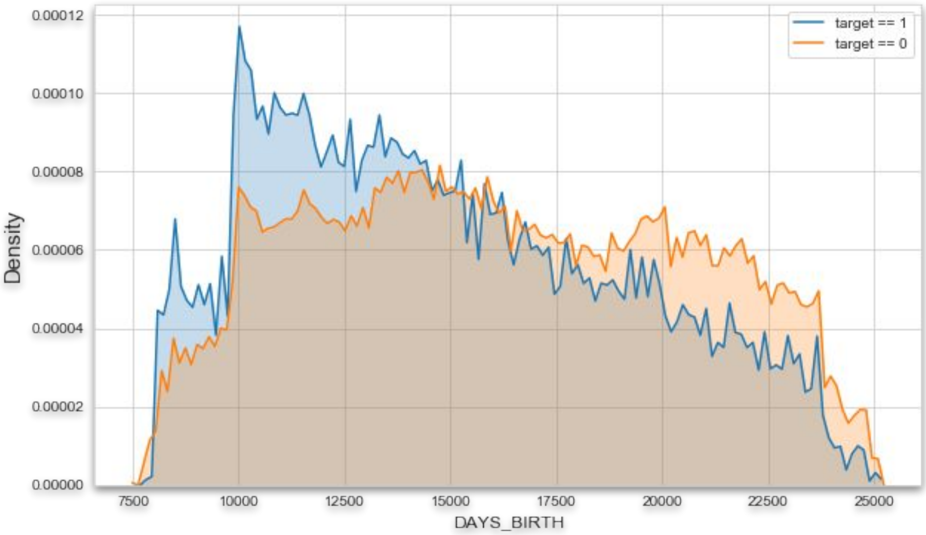
Conclusion: As we can see it clearly, if the applicant has changed the phone number 1000 days before applying for loan, is more likely to default.





Conclusion: Both of the EXT_SOURCE are negatively correlated with the target that means, we can say as the EXT_SOURCE increases, the client is more likely to repay the loan.



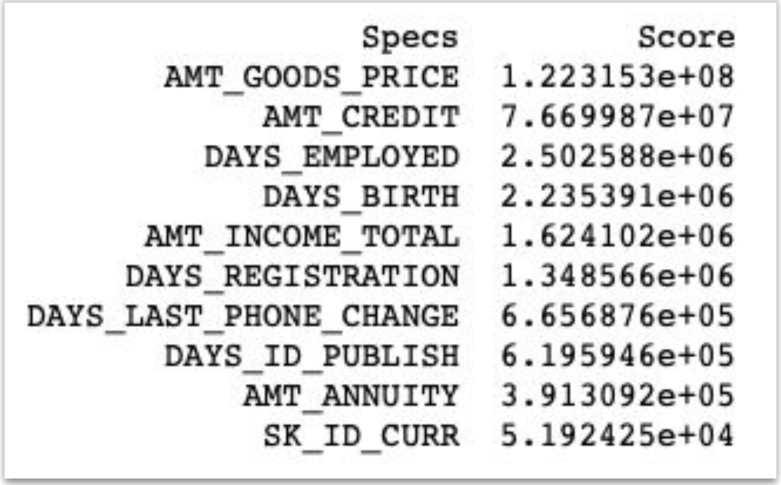


Conclusion: As the plot shows, clients who are employed for less than 2000 days are more likely to default.

Conclusion: As the plot shows, if the days of birth is less than ~14000 days prior to application, then they are most likely to default

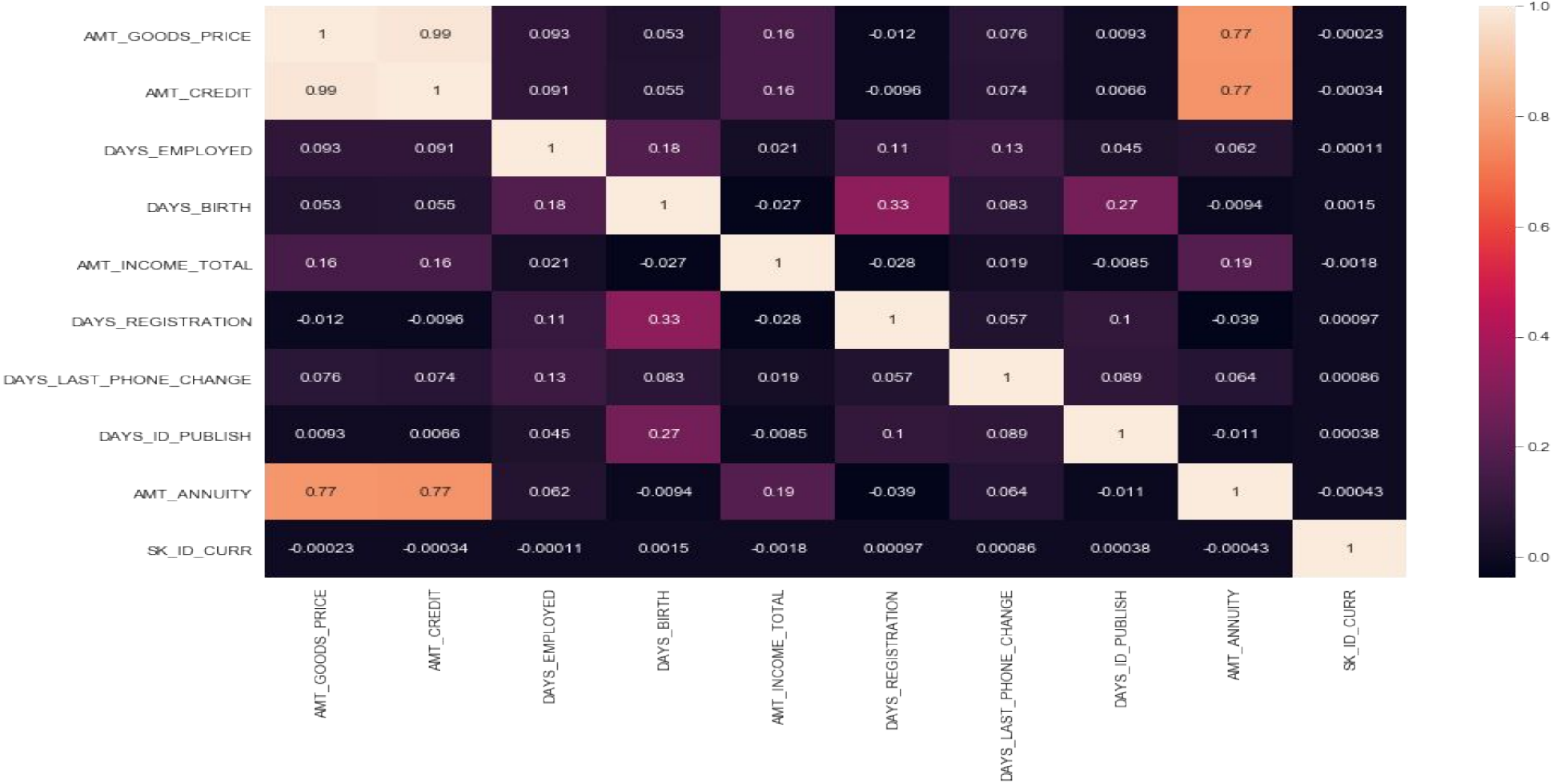
Finding the top correlation

The approach to get the variables that are highly correlated with the Target variable was to find f-score. Based on the f-score, the top 10 variables that turned out to be highly correlated with Target variable were:



```
                 Specs         Score
        AMT_GOODS_PRICE  1.223153e+08
             AMT_CREDIT  7.669987e+07
          DAYS_EMPLOYED  2.502588e+06
             DAYS_BIRTH  2.235391e+06
       AMT_INCOME_TOTAL  1.624102e+06
      DAYS_REGISTRATION  1.348566e+06
 DAYS_LAST_PHONE_CHANGE  6.656876e+05
        DAYS_ID_PUBLISH  6.195946e+05
            AMT_ANNUITY  3.913092e+05
             SK_ID_CURR  5.192425e+04
```

Visualizing these variables on heatmap

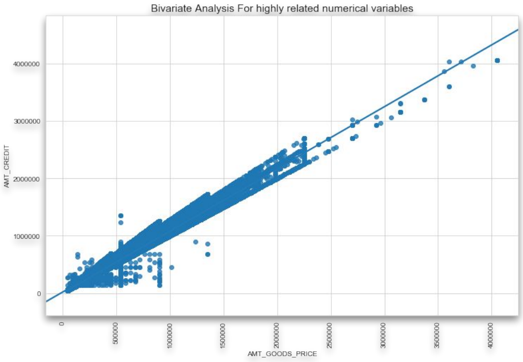


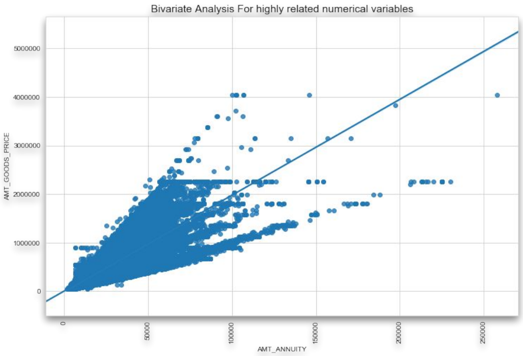
	AMT_GOODS_PRICE	AMT_CREDIT	DAYS_EMPLOYED	DAYS_BIRTH	AMT_INCOME_TOTAL	DAYS_REGISTRATION	DAYS_LAST_PHONE_CHANGE	DAYS_ID_PUBLISH	AMT_ANNUITY	SK_ID_CURR
AMT_GOODS_PRICE	1	0.99	0.093	0.053	0.16	-0.012	0.076	0.0093	0.77	-0.00023
AMT_CREDIT	0.99	1	0.091	0.055	0.16	-0.0096	0.074	0.0066	0.77	-0.00034
DAYS_EMPLOYED	0.093	0.091	1	0.18	0.021	0.11	0.13	0.045	0.062	-0.00011
DAYS_BIRTH	0.053	0.055	0.18	1	-0.027	0.33	0.083	0.27	-0.0094	0.0015
AMT_INCOME_TOTAL	0.16	0.16	0.021	-0.027	1	-0.028	0.019	-0.0085	0.19	-0.0018
DAYS_REGISTRATION	-0.012	-0.0096	0.11	0.33	-0.028	1	0.057	0.1	-0.039	0.00097
DAYS_LAST_PHONE_CHANGE	0.076	0.074	0.13	0.083	0.019	0.057	1	0.089	0.064	0.00086
DAYS_ID_PUBLISH	0.0093	0.0066	0.045	0.27	-0.0085	0.1	0.089	1	-0.011	0.00038
AMT_ANNUITY	0.77	0.77	0.062	-0.0094	0.19	-0.039	0.064	-0.011	1	-0.00043
SK_ID_CURR	-0.00023	-0.00034	-0.00011	0.0015	-0.0018	0.00097	0.00086	0.00038	-0.00043	1

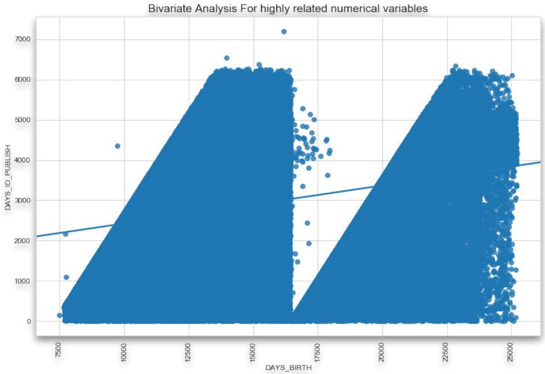
Conclusion: In the above correlation matrix, the variables with top correlation with Target and the pairs which have highest correlation among themselves in that list have been shown.

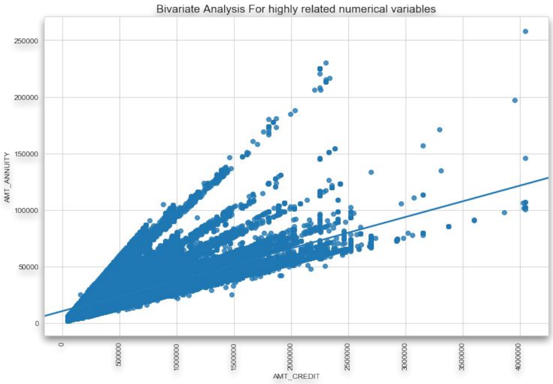
1.	AMT_GOODS_PRICE	AMT_CREDIT	0.986588
2.	AMT_ANNUITY	AMT_GOODS_PRICE	0.774661
3.	AMT_CREDIT	AMT_ANNUITY	0.770127
4.	DAYS_BIRTH	DAYS_REGISTRATION	0.331912
5.	DAYS_BIRTH	DAYS_ID_PUBLISH	0.272691

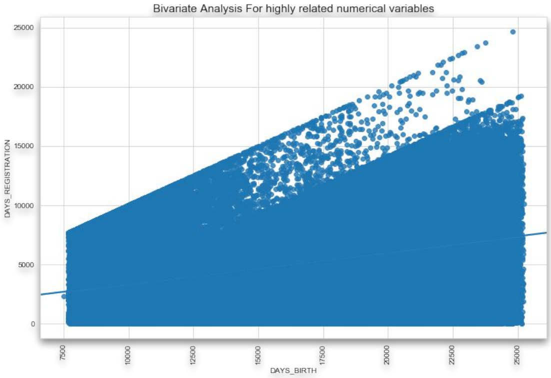
Bivariate Analysis of highly correlated variables











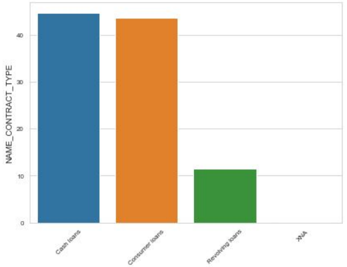
Conclusion: From the above five plots, we can say the top three correlated variables are directly proportional among one another.

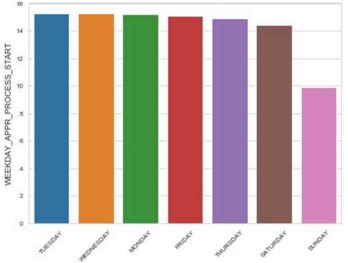
The plot with the variable DAYS_BIRTH | DAYS_REGISTRATION, the correlation varies from 0 to +1 for different data points of DAYS_BIRTH

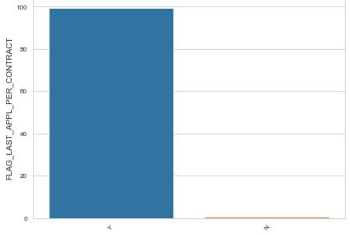
The plot with the variable DAYS_BIRTH | DAYS_ID_PUBLISH, the correlation varies from 0 to +1 till the DAYS_BIRTH 16000, the pattern discontinued due to the fall in values in the variable DAYS_ID_PUBLISHED then continues with the same previous pattern.

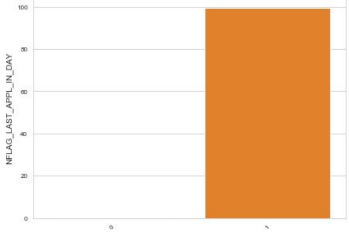
EDA For Previous Application

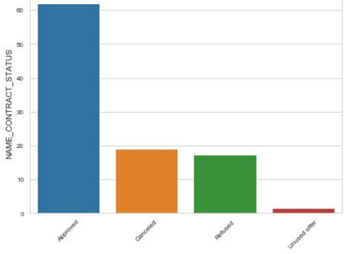
1. **FLAG_LAST_APPL_PER_CONTRACT**: We have only data for last application for the previous contract (100%)
2. **NFLAG_LAST_APPL_IN_DAY:** All the datas are for the application that was the last application per day of the client. No error had been made.
3. **NAME_PRODUCT_TYPE** : 60% of the product type of clients' previous application is not available.
4. **NAME_CONTRACT_TYPE** : The data looks balance except for the revolving loans, number of clients are less.
5. **NAME_CONTRACT_STATUS:** 60% of the client's applications are approved where as ~20% of the applications are cancelled or Refused.
6. **NAME_PAYMENT_TYPE:** Most of the clients chose to go for Cash payment mode and the data looks imbalance.

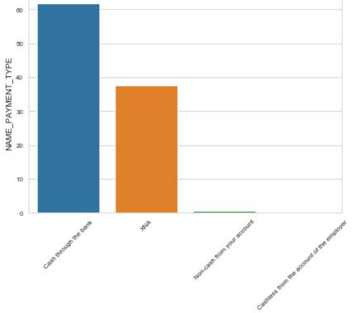






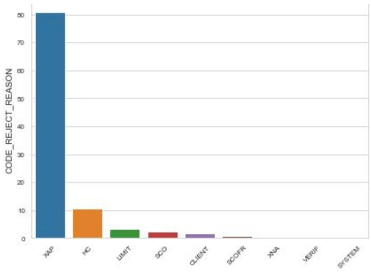


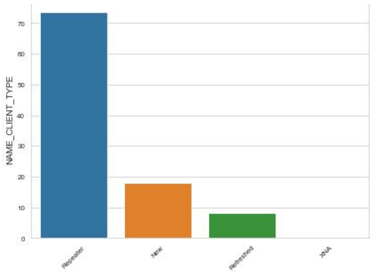


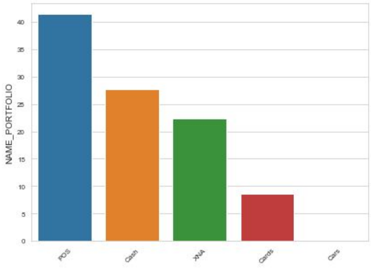


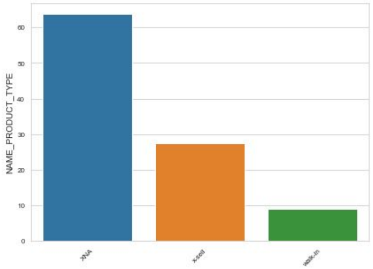
EDA For Previous Application

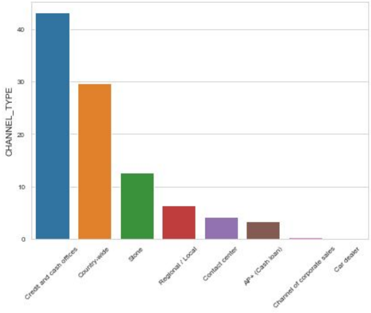
7. **NAME_CLIENT_TYPE:** Existing client i.e., category 'Repeater' applied for loan the most (~75%) following with new clients with ~18%.
8. **NAME_PORTFOLIO:** The clients with the previous application for POS have applied for loan the most i.e., ~40% following the clients with previous application for Cash with ~22%
9. **NAME_YIELD_GROUP**: ~22% of the clien's interest rate of loan are categorised into middle following with ~20% of the clients with high interest rate in their previous application.
10. **WEEKDAY_APPR_PROCESS_START**: We can see every day almost same percentage of applications have been registered and on Sunday just ~10% of applications are registered
11. **CHANNEL_TYPE**: Almost 40% of the clients' applications are aquired from credit and cash offices.
12. **CODE_REJECT_REASON**: 80% of the reason for rejecting the application is XAP i.e., the reason has not been mentioned for most of the rejection. Following with HC that is the reason for rejection of 10% of the application

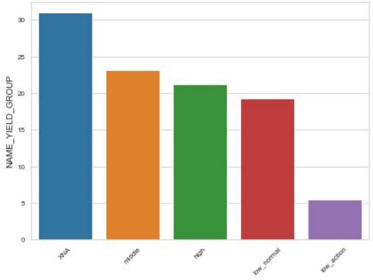




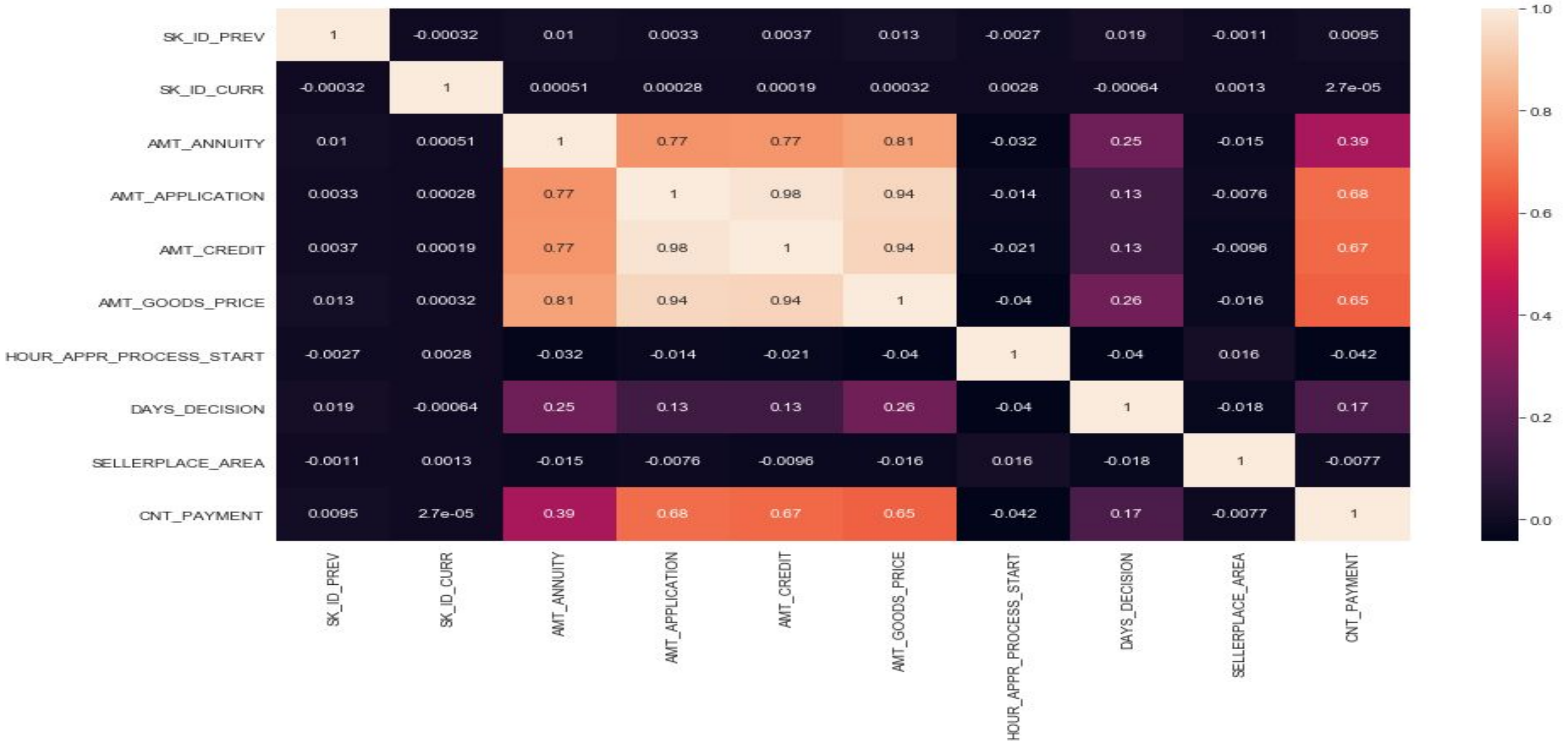








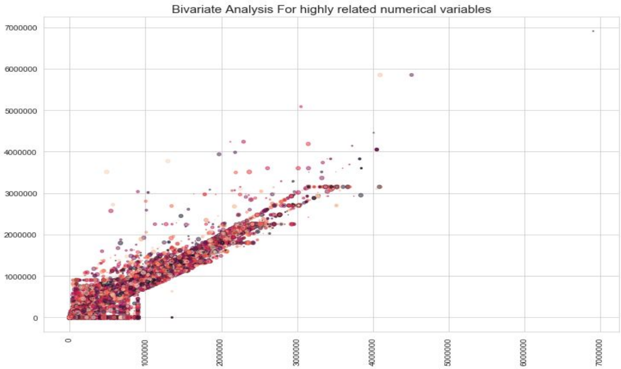
Correlation among different variables Previous Application Data



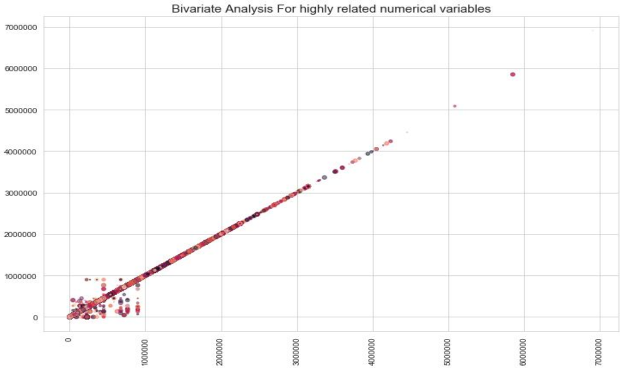
The top 5 correlated variables from the previous application dataset are:

1. AMT_CREDIT	AMT_APPLICATION	0.975824
2. AMT_GOODS_PRICE	AMT_APPLICATION	0.944614
3. AMT_GOODS_PRICE	AMT_CREDIT	0.937753
4. AMT_GOODS_PRICE	AMT_ANNUITY	0.808610
5. AMT_CREDIT	AMT_ANNUITY	0.773140

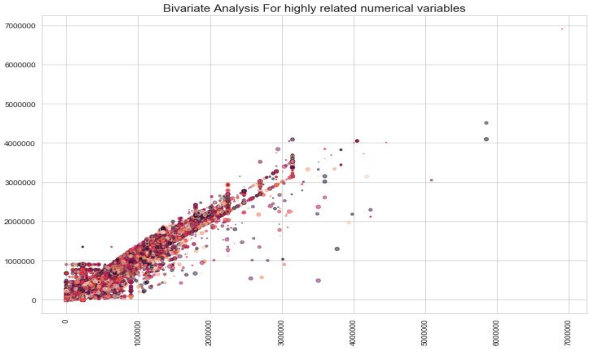
Bivariate Analysis of highly correlated variables



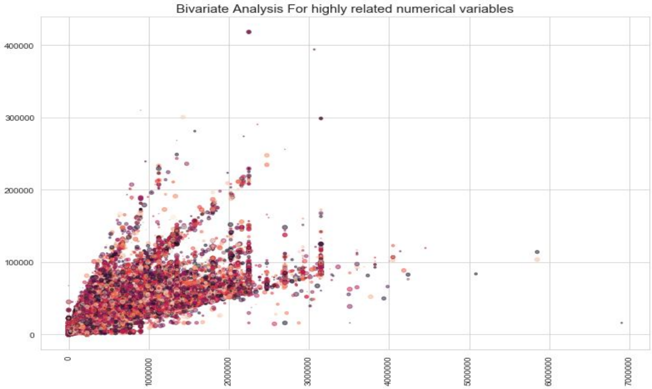
Bivariate Analysis For highly related numerical variables



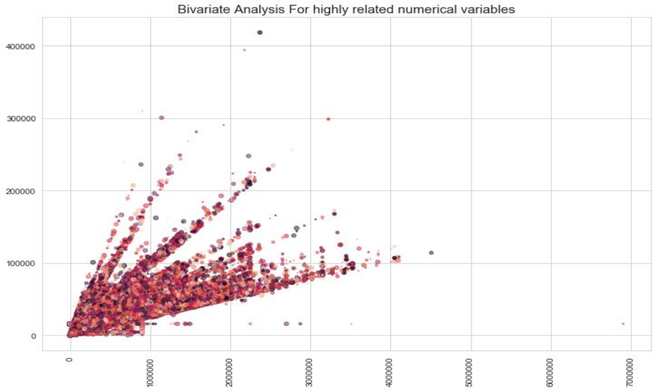
Bivariate Analysis For highly related numerical variables



Bivariate Analysis For highly related numerical variables



Bivariate Analysis For highly related numerical variables



Bivariate Analysis For highly related numerical variables

Conclusion
So from the above five plot, all the variable are having a positive Correlation, or directly proportional with each other.

So for **Previous Application Data**
AMT_CREDIT | AMT_APPLICATION
AMT_GOODS_PRICE | AMT_APPLICATION
AMT_GOODS_PRICE | AMT_CREDIT
AMT_GOODS_PRICE | AMT_ANNUITY
AMT_CREDIT | AMT_ANNUITY

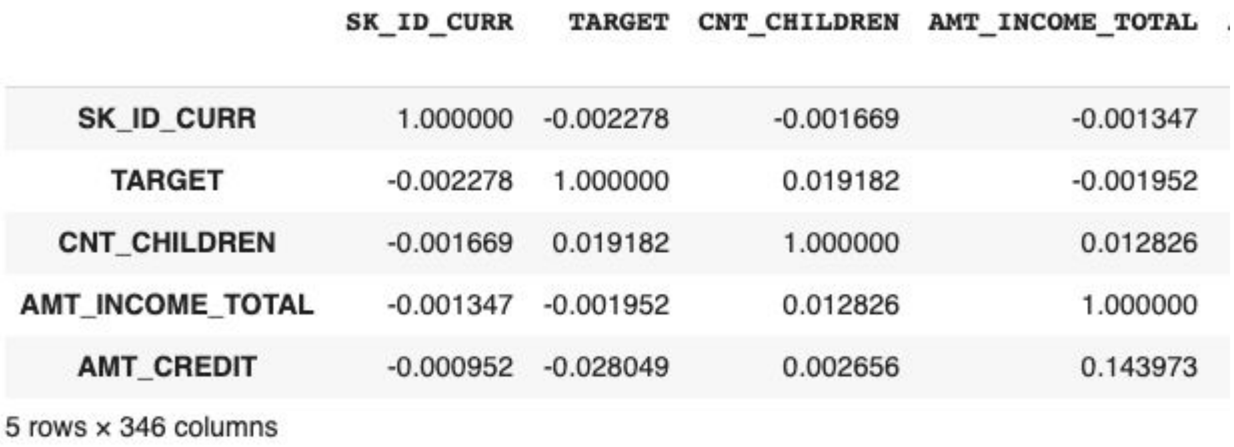
And for **Current Application data**
AMT_GOODS_PRICE | AMT_CREDIT
AMT_ANNUITY | AMT_GOODS_PRICE
AMT_CREDIT | AMT_ANNUITY
DAYS_BIRTH | DAYS_REGISTRATION
DAYS_BIRTH | DAYS_ID_PUBLISH

So the **Common correlated variables** are
1. AMT_GOODS_PRICE | AMT_CREDIT
2. AMT_ANNUITY | AMT_GOODS_PRICE
3. AMT_CREDIT | AMT_ANNUITY

Merging both the dataframes

The approach to merge both the data frame was:

1. Dropped the IDs that were in prev_application but not in application_data.
2. Converted categorical variables into numerical by implementing dummies on the categories.
3. Aggregated numerical and categorical columns with [min, max, mean] and [mean] respectively
4. Finally merged the both the dataset on ‘SK_CURR_ID’
5. Head of the merged Dataframe →



	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL
SK_ID_CURR	1.000000	-0.002278	-0.001669	-0.001347
TARGET	-0.002278	1.000000	0.019182	-0.001952
CNT_CHILDREN	-0.001669	0.019182	1.000000	0.012826
AMT_INCOME_TOTAL	-0.001347	-0.001952	0.012826	1.000000
AMT_CREDIT	-0.000952	-0.028049	0.002656	0.143973

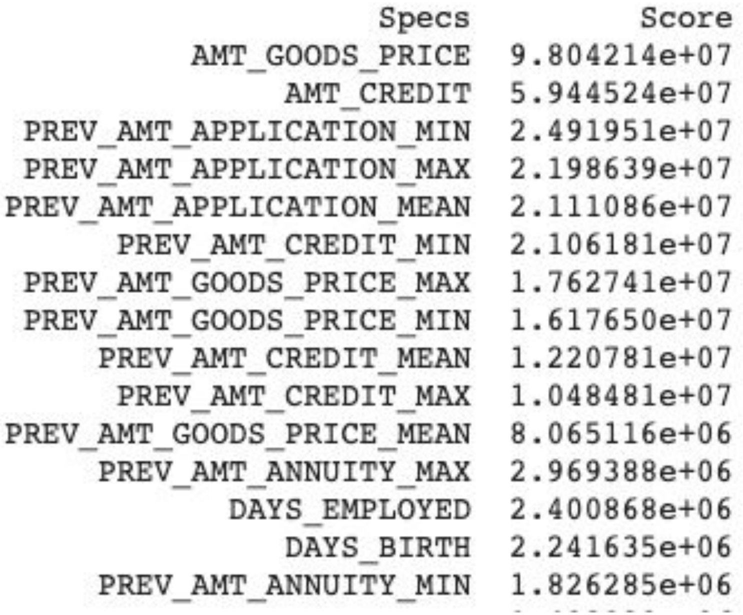
5 rows × 346 columns

Finding correlation of the merged dataframe

As the number of columns were 346, it was difficult to analyze heatmap and conclude the top correlated variables out of it

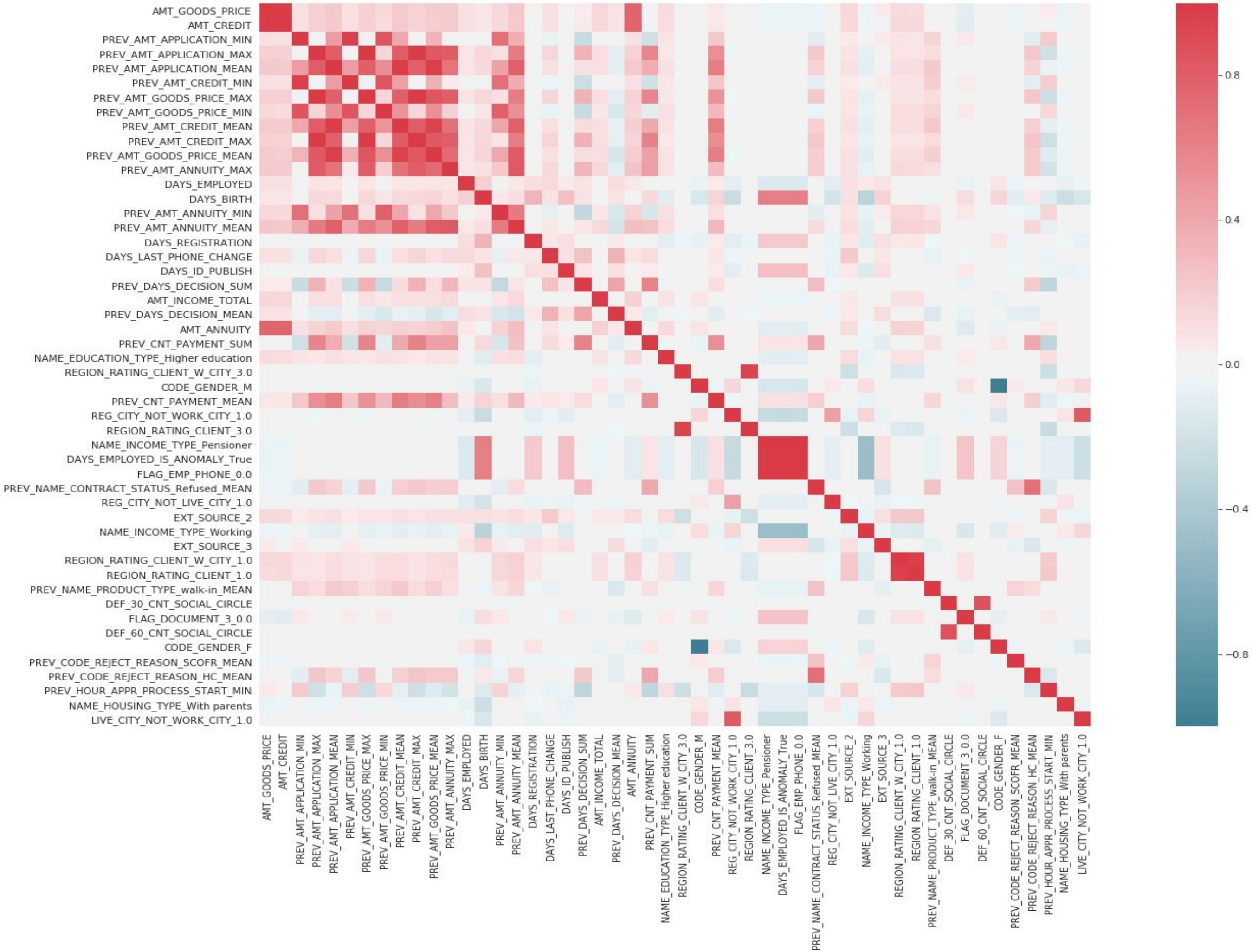
We found out top 50 correlated variables with the target variable by calculating f-score and then plotted heat map for those 50 variables to find out the highly correlated variables among one another.

Top few correlated variables are shown in the image:



```
                     Specs         Score
            AMT_GOODS_PRICE  9.804214e+07
                 AMT_CREDIT  5.944524e+07
   PREV_AMT_APPLICATION_MIN  2.491951e+07
   PREV_AMT_APPLICATION_MAX  2.198639e+07
  PREV_AMT_APPLICATION_MEAN  2.111086e+07
        PREV_AMT_CREDIT_MIN  2.106181e+07
   PREV_AMT_GOODS_PRICE_MAX  1.762741e+07
   PREV_AMT_GOODS_PRICE_MIN  1.617650e+07
       PREV_AMT_CREDIT_MEAN  1.220781e+07
        PREV_AMT_CREDIT_MAX  1.048481e+07
  PREV_AMT_GOODS_PRICE_MEAN  8.065116e+06
       PREV_AMT_ANNUITY_MAX  2.969388e+06
              DAYS_EMPLOYED  2.400868e+06
                 DAYS_BIRTH  2.241635e+06
       PREV_AMT_ANNUITY_MIN  1.826285e+06
```

Top 50 correlations



Some of top correlations

```
FLAG_EMP_PHONE_0.0                DAYS_EMPLOYED_IS_ANOMALY_True       0.999884
DAYS_EMPLOYED_IS_ANOMALY_True     FLAG_EMP_PHONE_0.0                  0.999884
DAYS_EMPLOYED_IS_ANOMALY_True     NAME_INCOME_TYPE_Pensioner          0.999698
NAME_INCOME_TYPE_Pensioner        DAYS_EMPLOYED_IS_ANOMALY_True       0.999698
NAME_INCOME_TYPE_Pensioner        FLAG_EMP_PHONE_0.0                  0.999582
FLAG_EMP_PHONE_0.0                NAME_INCOME_TYPE_Pensioner          0.999582
PREV_AMT_APPLICATION_MAX          PREV_AMT_GOODS_PRICE_MAX            0.994357
PREV_AMT_GOODS_PRICE_MAX          PREV_AMT_APPLICATION_MAX            0.994357
AMT_GOODS_PRICE                   AMT_CREDIT                          0.986229
AMT_CREDIT                        AMT_GOODS_PRICE                     0.986229
PREV_AMT_CREDIT_MAX               PREV_AMT_APPLICATION_MAX            0.985563
PREV_AMT_APPLICATION_MAX          PREV_AMT_CREDIT_MAX                 0.985563
PREV_AMT_CREDIT_MAX               PREV_AMT_GOODS_PRICE_MAX            0.983324
PREV_AMT_GOODS_PRICE_MAX          PREV_AMT_CREDIT_MAX                 0.983324
PREV_AMT_CREDIT_MEAN              PREV_AMT_APPLICATION_MEAN           0.975854
```