Credit EDA Case Study

Author: Ashutosh Kumar, Prabhakaran Chandraseakaran
Email : ashutoshind2017@outlook.com, prabhakaran.i@prabs.in

Problem Statement

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. In other words, **the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.**

Data source details

Data source: https://cdn.upgrad.com/UpGrad/temp/afe3690a-9d30-4f1d-b9ee-2d5d17881422/Dataset.txt

This dataset has 3 files as explained below:

1. 'application_data.csv' contains all the information of the client at the time of application.

The data is about whether a client has payment difficulties.

2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

3. 'columns_description.csv' is data dictionary which describes the meaning of the variables.

Missing Data Analysis and Treatment

This is handled through **custom function** created “missing_data” to identify **missing data** across various datasets:

```
# Missing data information:
missing_data(df_application_data).head(20)
```

	Total	Percent
COMMONAREA_MEDI	214865	69.872297
COMMONAREA_AVG	214865	69.872297
COMMONAREA_MODE	214865	69.872297
NONLIVINGAPARTMENTS_MODE	213514	69.432963
NONLIVINGAPARTMENTS_MEDI	213514	69.432963
NONLIVINGAPARTMENTS_AVG	213514	69.432963
FONDKAPREMONT_MODE	210295	68.386172
LIVINGAPARTMENTS_MEDI	210199	68.354953
LIVINGAPARTMENTS_MODE	210199	68.354953
LIVINGAPARTMENTS_AVG	210199	68.354953
FLOORSMIN_MEDI	208642	67.848630
FLOORSMIN_MODE	208642	67.848630
FLOORSMIN_AVG	208642	67.848630
YEARS_BUILD_MEDI	204488	66.497784
YEARS_BUILD_AVG	204488	66.497784
YEARS_BUILD_MODE	204488	66.497784
OWN_CAR_AGE	202929	65.990810
LANDAREA_MODE	182590	59.376738
LANDAREA_AVG	182590	59.376738
LANDAREA_MEDI	182590	59.376738

```
# Finding missing data using our function created in this ipynb
missing_data(df_previous_app_data).head(20)
```

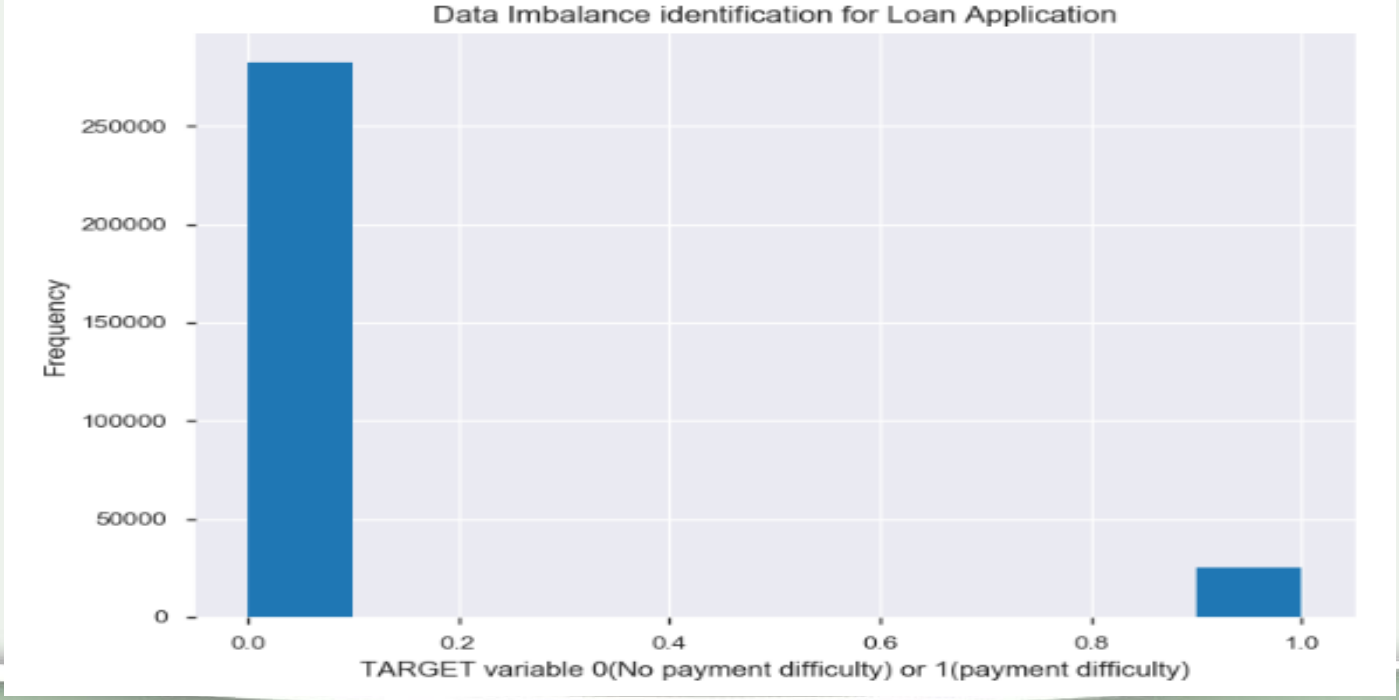
	Total	Percent
RATE_INTEREST_PRIVILEGED	1664263	99.643698
RATE_INTEREST_PRIMARY	1664263	99.643698
RATE_DOWN_PAYMENT	895844	53.636480
AMT_DOWN_PAYMENT	895844	53.636480
NAME_TYPE_SUITE	820405	49.119754
DAYS_TERMINATION	673065	40.298129
NFLAG_INSURED_ON_APPROVAL	673065	40.298129
DAYS_FIRST_DRAWING	673065	40.298129
DAYS_FIRST_DUE	673065	40.298129
DAYS_LAST_DUE_1ST_VERSION	673065	40.298129
DAYS_LAST_DUE	673065	40.298129
AMT_GOODS_PRICE	385515	23.081773
AMT_ANNUITY	372235	22.286665
CNT_PAYMENT	372230	22.286366
PRODUCT_COMBINATION	346	0.020716
AMT_CREDIT	1	0.000060
SK_ID_CURR	0	0.000000
NAME_CONTRACT_TYPE	0	0.000000
WEEKDAY_APPR_PROCESS_START	0	0.000000
HOUR_APPR_PROCESS_START	0	0.000000

Data Imbalance Identification for Loan Application

Imbalance problem is the problem in machine learning where the total number of class of data (positive) is far less than the total number of another class of data (negative).
The **target variable** is what we are asked to predict: either a 0 for the loan was repaid on time, or a 1 indicating the client had payment difficulties. We can first examine the number of loans falling into each category.

```
0     282686
1      24825
Name: TARGET, dtype: int64
```

Plotting the information for **visualisation of imbalance** with respect to the target variable in absolute terms:



Data Imbalance Identification for Loan Application

Yes there is imbalance problem here in the data as we can see from plot in last slide and there are far more loans that were repaid on time than loans that were not repaid on time.

The **ratio of imbalance** is closely is 1139 approximately:

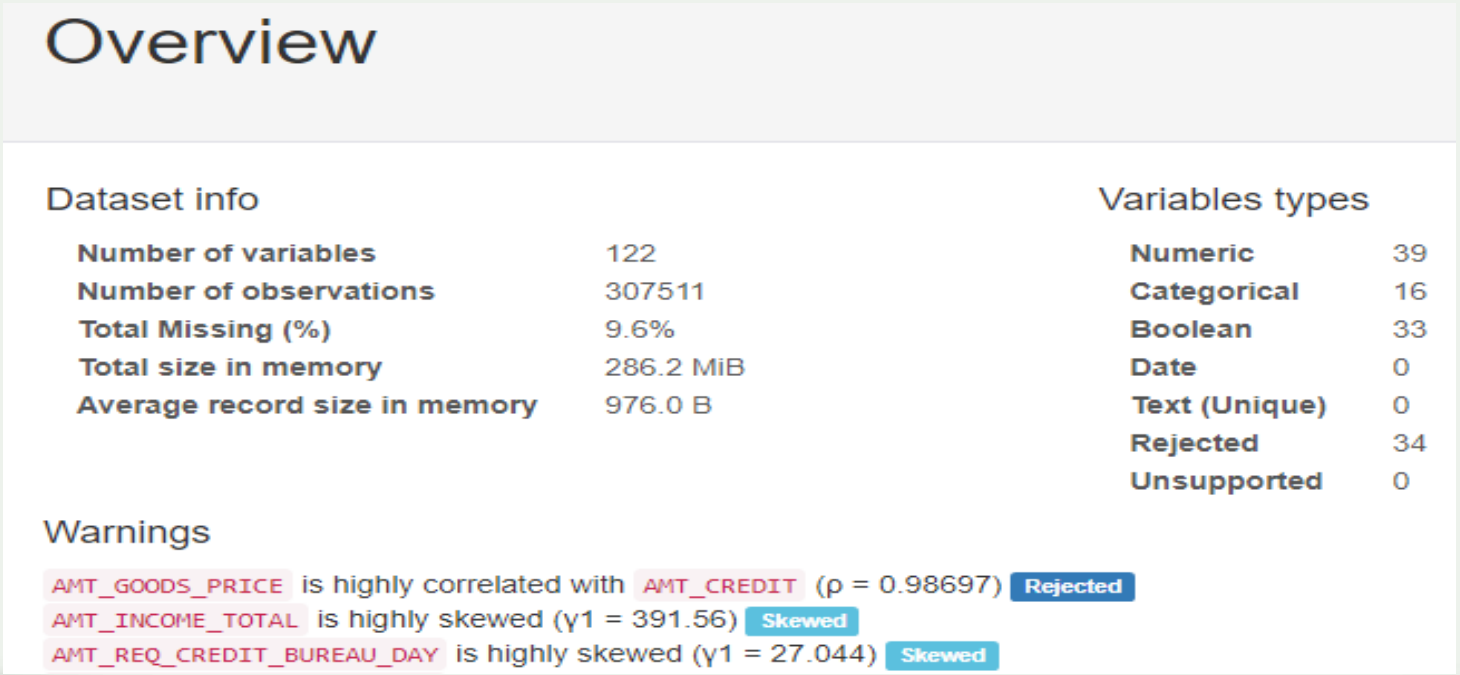
```
The percentage with no payment difficulty is  91.93 % of the dataset
The percentage with payment difficulty is  8.07 % of the dataset
The ratio of imbalance is 1138.72
```

EDA Using Pandas Profiling

Using **built-in pandas profiling library** for building inline and html report.

For each column it gives the important statistics.
For the column type – data are presented in an interactive HTML report: Essentials: type, unique values, missing values, Quantile statistics, Descriptive statistics, Most frequent values, Histogram, Correlations highlighting etc. We will use this for our EDA.

Snapshot from the report:

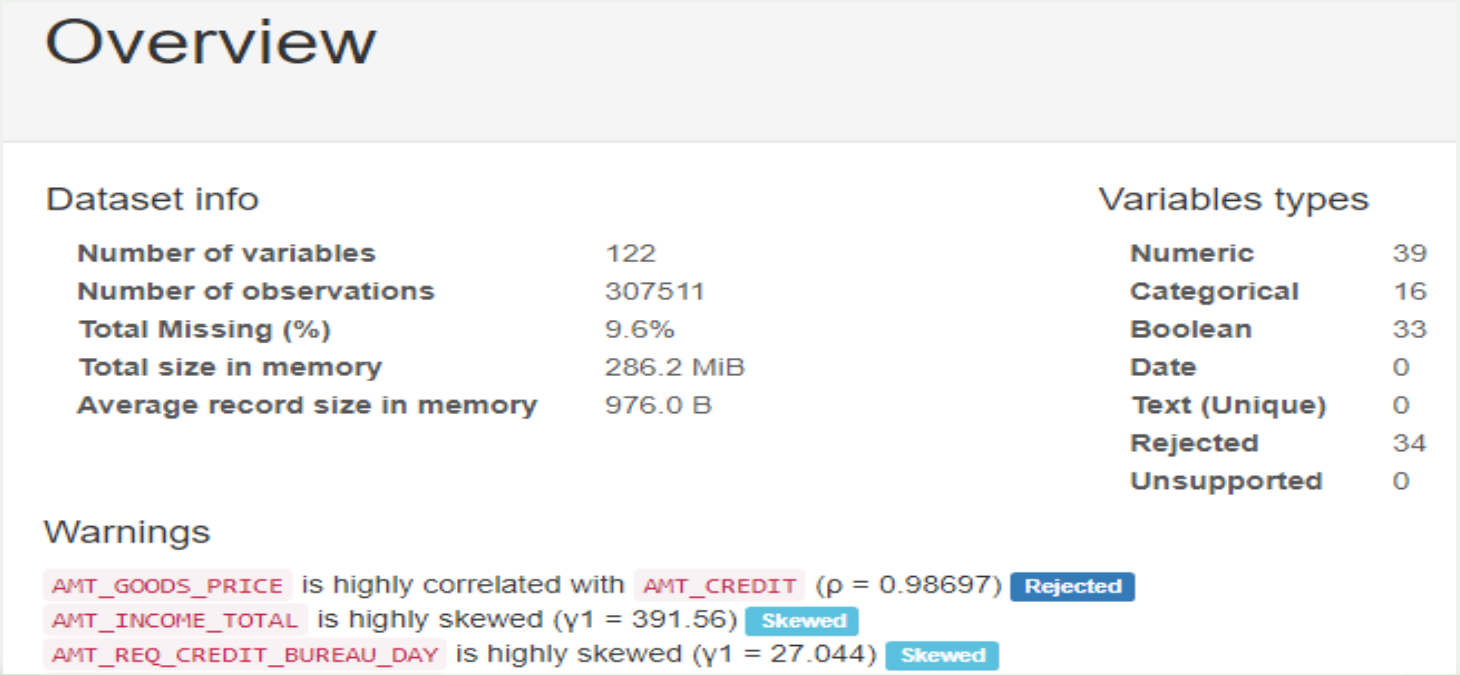


EDA Using Pandas Profiling

Using **built-in pandas profiling library** for building inline and html report.

For each column it gives the important statistics.
For the column type – data are presented in an interactive HTML report: Essentials: type, unique values, missing values, Quantile statistics, Descriptive statistics, Most frequent values, Histogram, Correlations highlighting etc. We will use this for our EDA.

Snapshot from the report:

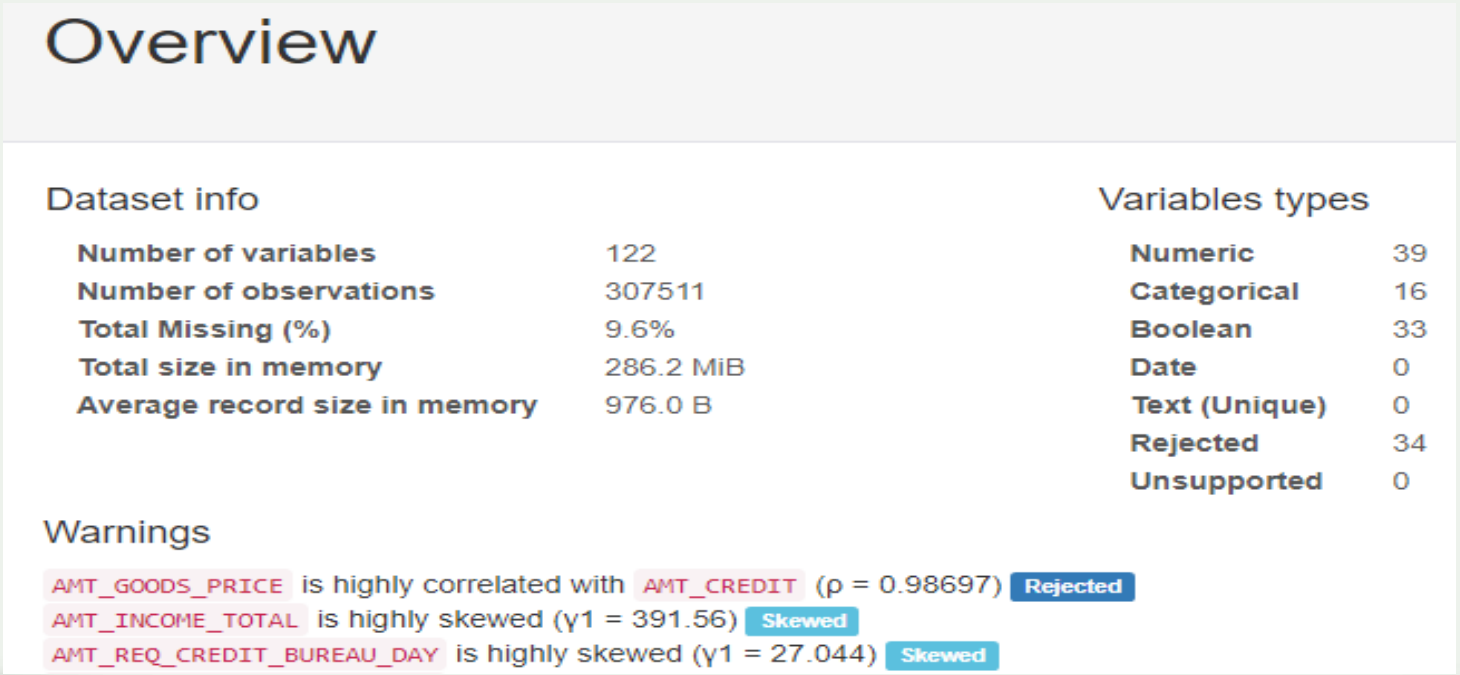


EDA Using Pandas Profiling

Using **built-in pandas profiling library** for building inline and html report.

For each column it gives the important statistics.
For the column type – data are presented in an interactive HTML report: Essentials: type, unique values, missing values, Quantile statistics, Descriptive statistics, Most frequent values, Histogram, Correlations highlighting etc. We will use this for our EDA.

Snapshot from the report:



Further EDA on dataset

- **Missing data treatment** was done with methods like imputations, correlations comparison of variables, removing the columns with more than x% of missing values etc.
- **Feature engineering** was done to create more new columns for analysis and creating derived columns.
- **Datatype conversion** was done whenever required for the given dataset for analysis and reducing the memory usage.
- **Data cleansing** was done for the incorrect values and outliers in the dataset.
- Please refer the ipynb file for more details and process.

Identification of outliers and its treatment

- An **outlier** is an observation that lies outside the overall pattern of a distribution.
- Among the numerical values we can find clear outliers for **CNT_CHILDREN** and **DAYS_EMPLOYED.**
- Using **describe** to determine any extreme values in the dataframe

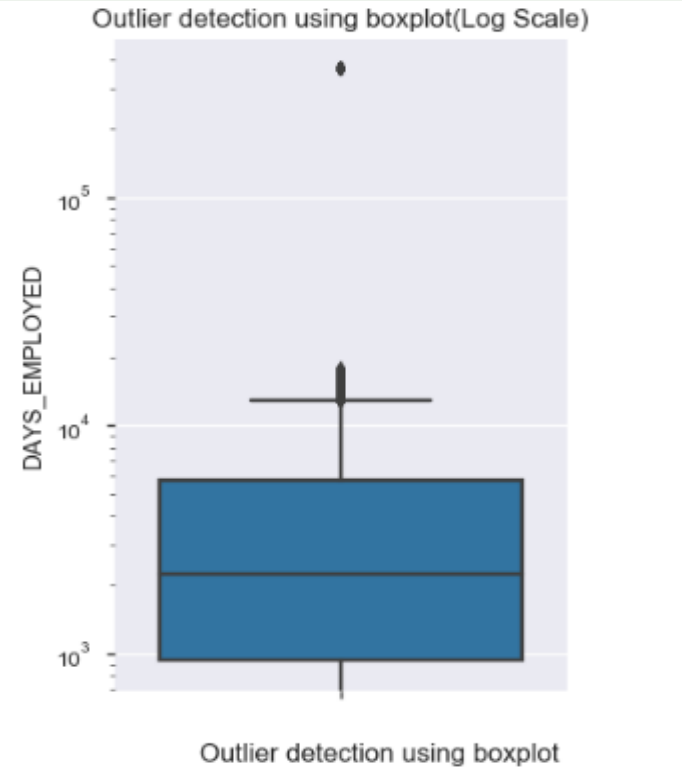
```
df_application_data['DAYS_EMPLOYED'].describe()
```

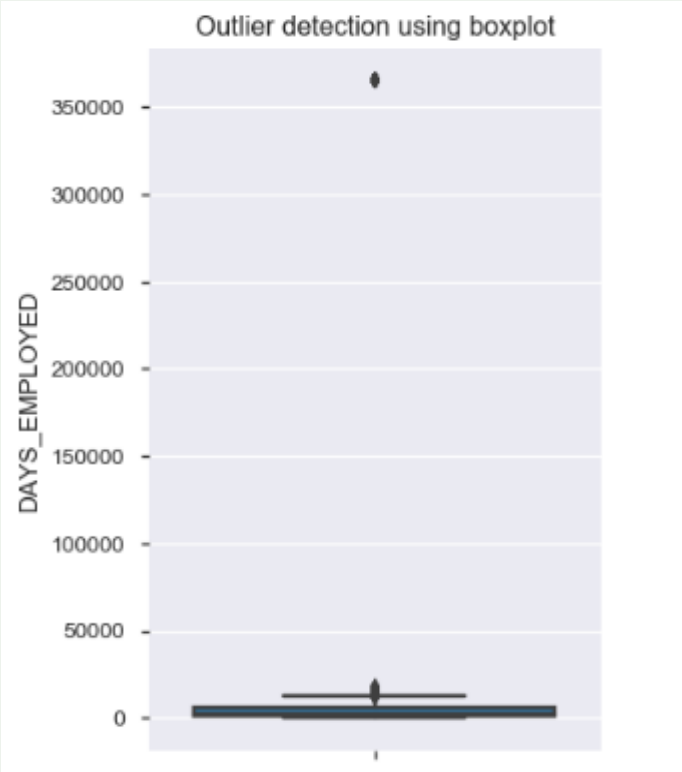
```
count     307507.000000
mean       67725.569893
std       139444.469301
min            0.000000
25%          933.000000
50%         2219.000000
75%         5707.000000
max       365243.000000
Name: DAYS_EMPLOYED, dtype: float64
```

- The maximum value is about 1000 years(365243 days) which does seems not right for days of employment (**DAYS_EMPLOYED**).
- We still have **CNT_CHILDREN** max value as 19 , but we can ignore the same here it might be a rare valid case where a family has these many children !

Identification of outliers and its treatment

Using **boxplot** to identify the outliers in numerical data:





We clearly have outliers present in the data set which needs to be treated or it will lead to skewness in data with impacting significantly on data's mean and other parameters.

Identification of outliers and its treatment

- For the treatment of the outliers we will use **IQR method** (Inter-Quartile Range Method).
- IQR tells how spread the middle values are. It can be used to tell when a value is too far from the middle.
- An **outlier** is a point which falls more than 1.5 times the interquartile range above the third quartile or below the first quartile.
- We have created a **custom python function** remove_outlier to handle all the outliers value in all required datasets based on the IQR method.

Data blending

- Next we have **merged** the two datasets **application_data.csv** and **previous_application.csv** representing the current application and historical loan data.

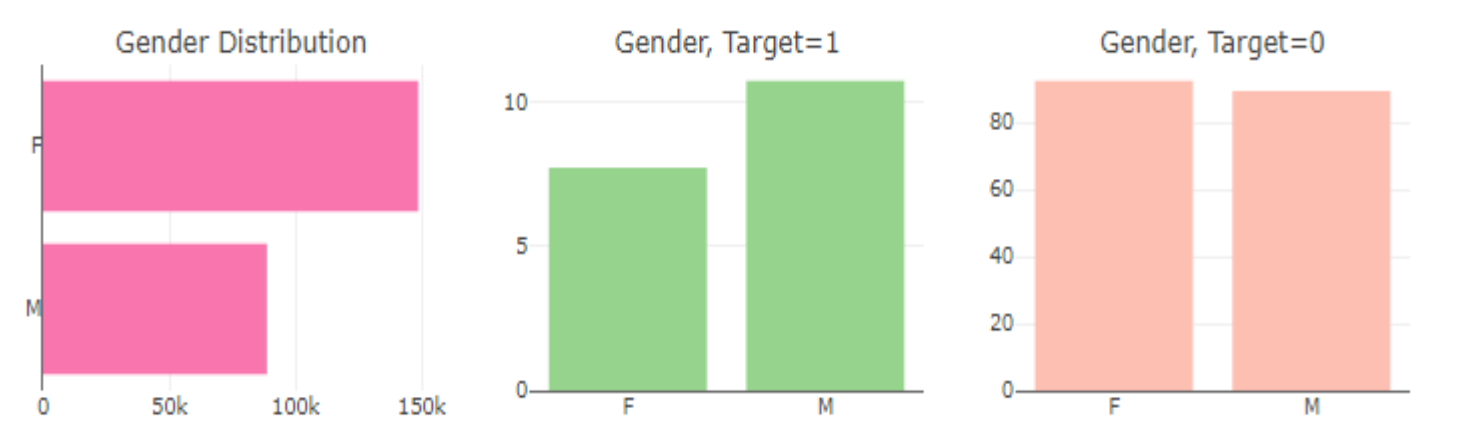
- The merge is done to determine the overall impact of all the variables in both data-set on the Target variable "**TARGET**" which will eventually help us on the decision of credit issue for our clients.

- We are merging the current application data with the clients previous application data **only for the clients which have currently applied for loan**, so using "**inner**" **join** here on the common numerical column **SK_ID_CURR.**

- Post this we have performed basic EDA on the combined data set.

Univariate, Bivariate Analysis and Data Visualization

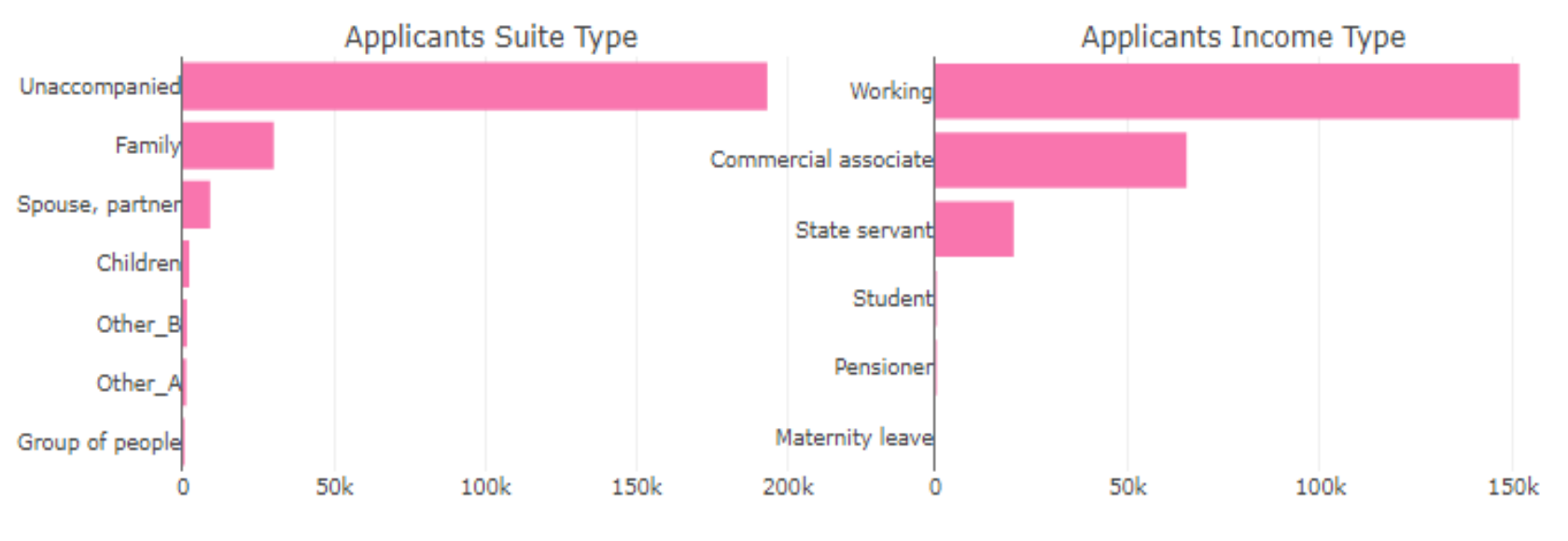
- Here we are beginning to perform data exploration, analysis and visualization.
- **Analysis of gender type of applicant:**



In the applicant's data women have applied for a larger number of loans which is almost the double as the men. In total, there are about 148,758 loan applications filed by females in contrast to about 88,781 applications filed by males.
However, a larger percentage (about 10% of the total) of men had the problems in paying the loan or making instalments within time as compared to women applicants (about 7%).

Univariate, Bivariate Analysis and Data Visualization

- **Identifying income type or suit type:**

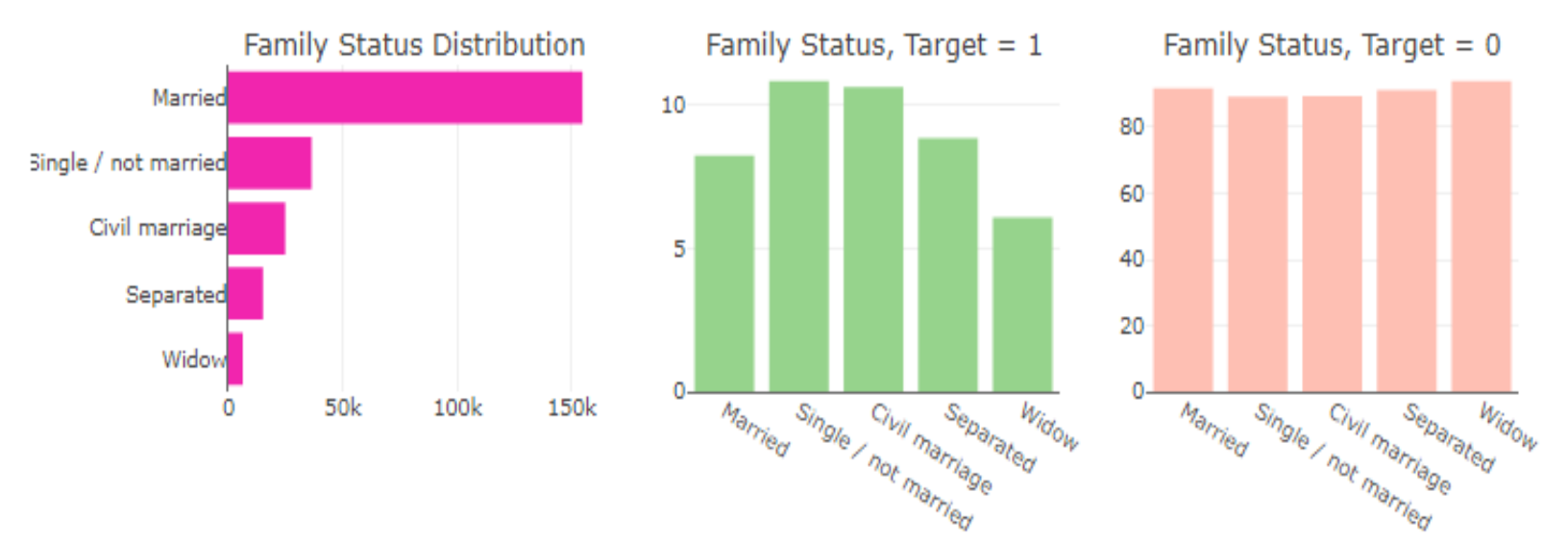


- Top 3 Type Suites which applies for loan are the houses which are: Unaccompanied, Family and Spouse(Partner).

- The income type of people who applies for loan include about 8 categories, top 3 are : Working, Commercial Associate and State servant.

Univariate, Bivariate Analysis and Data Visualization

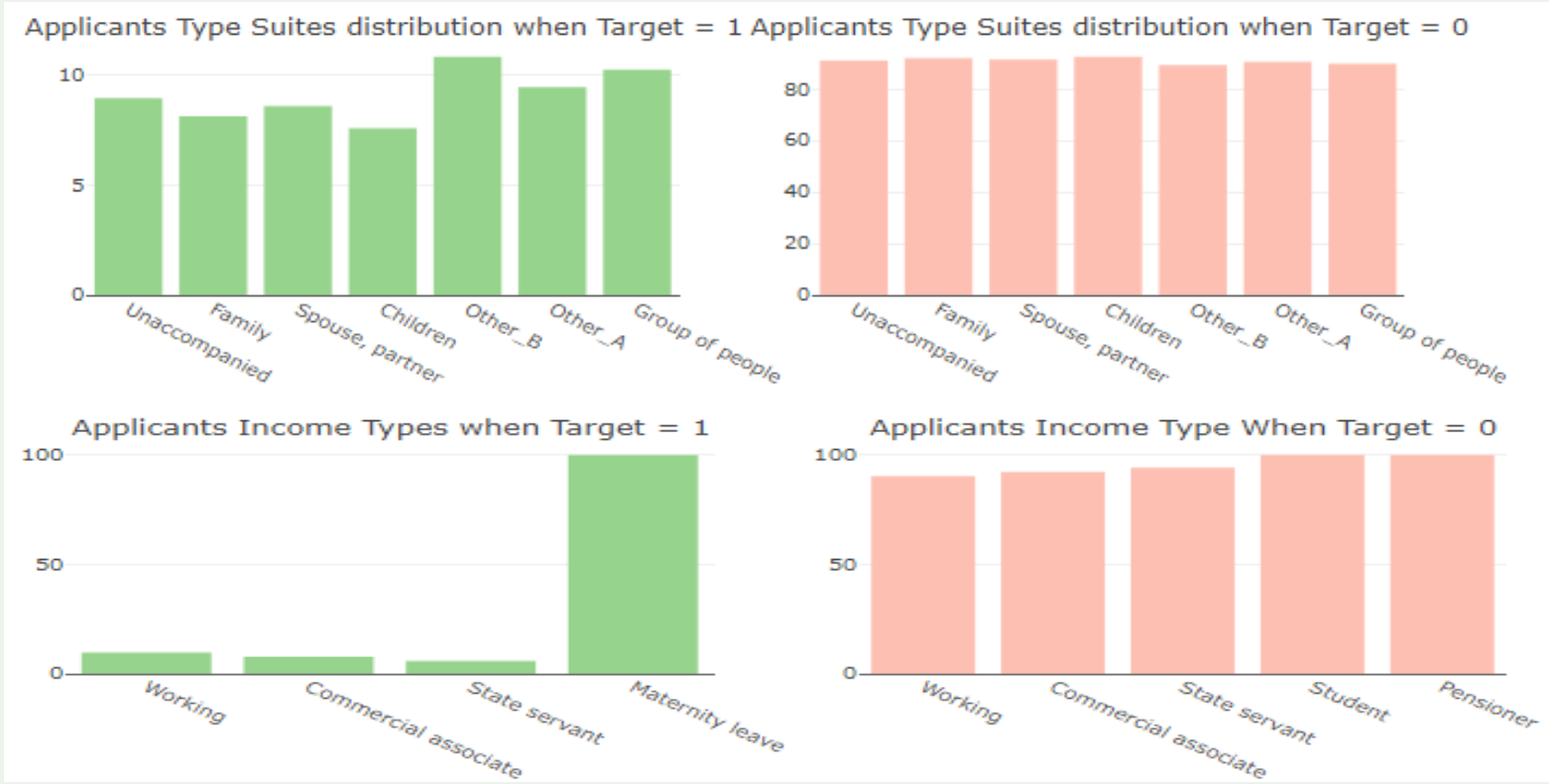
- **Analysis of Family status of applicant:**



- Married people have applied for a larger number of loan applications about 155K, However, people having Civil Marriage and people who are single/not married have the highest percentage (about 10% each) of loan problems and challenges.

Univariate, Bivariate Analysis and Data Visualization

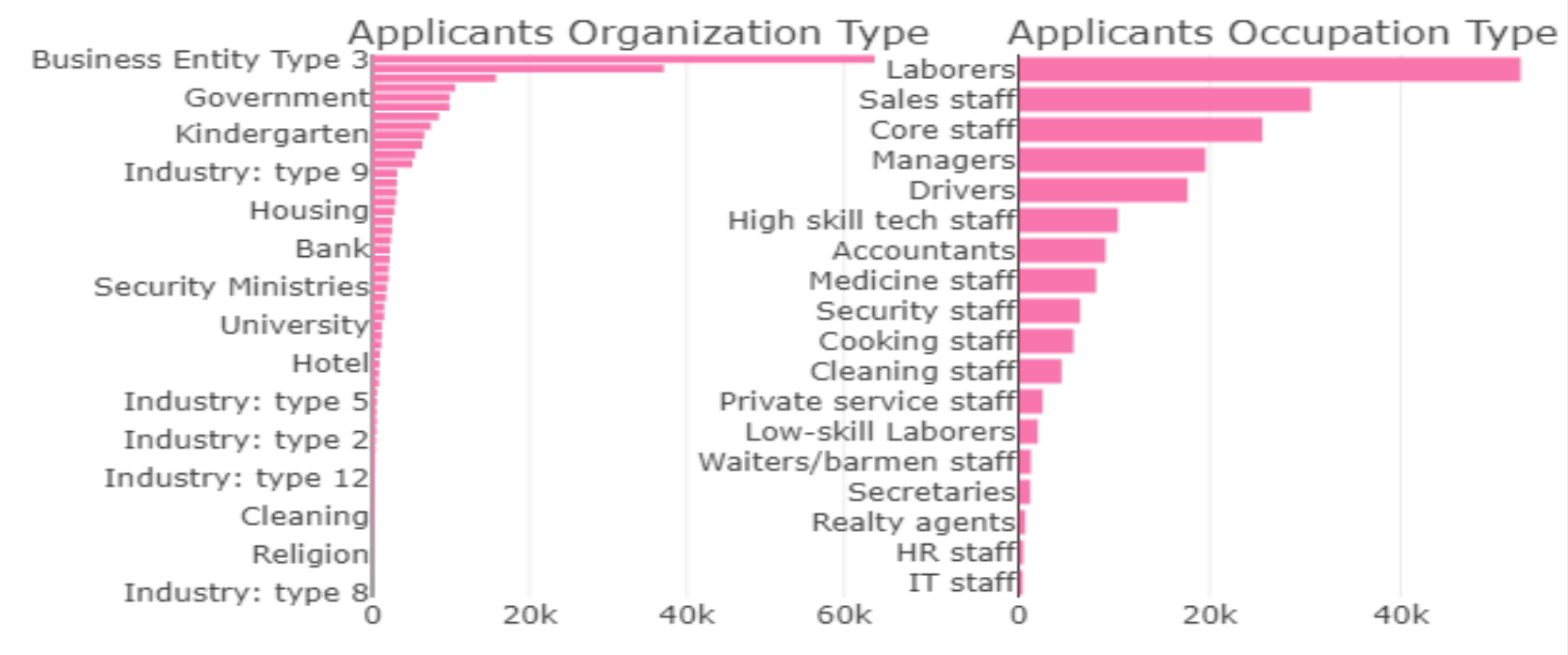
- **Variation of target variable with suit and income type:**



We see that Applicants having Income Types : Maternity Leaves has the highest percentage (about 40%) of Target = 1 i.e.. having more payment problems, while Stste Servants have the least (about 5.8%).

Univariate, Bivariate Analysis and Data Visualization

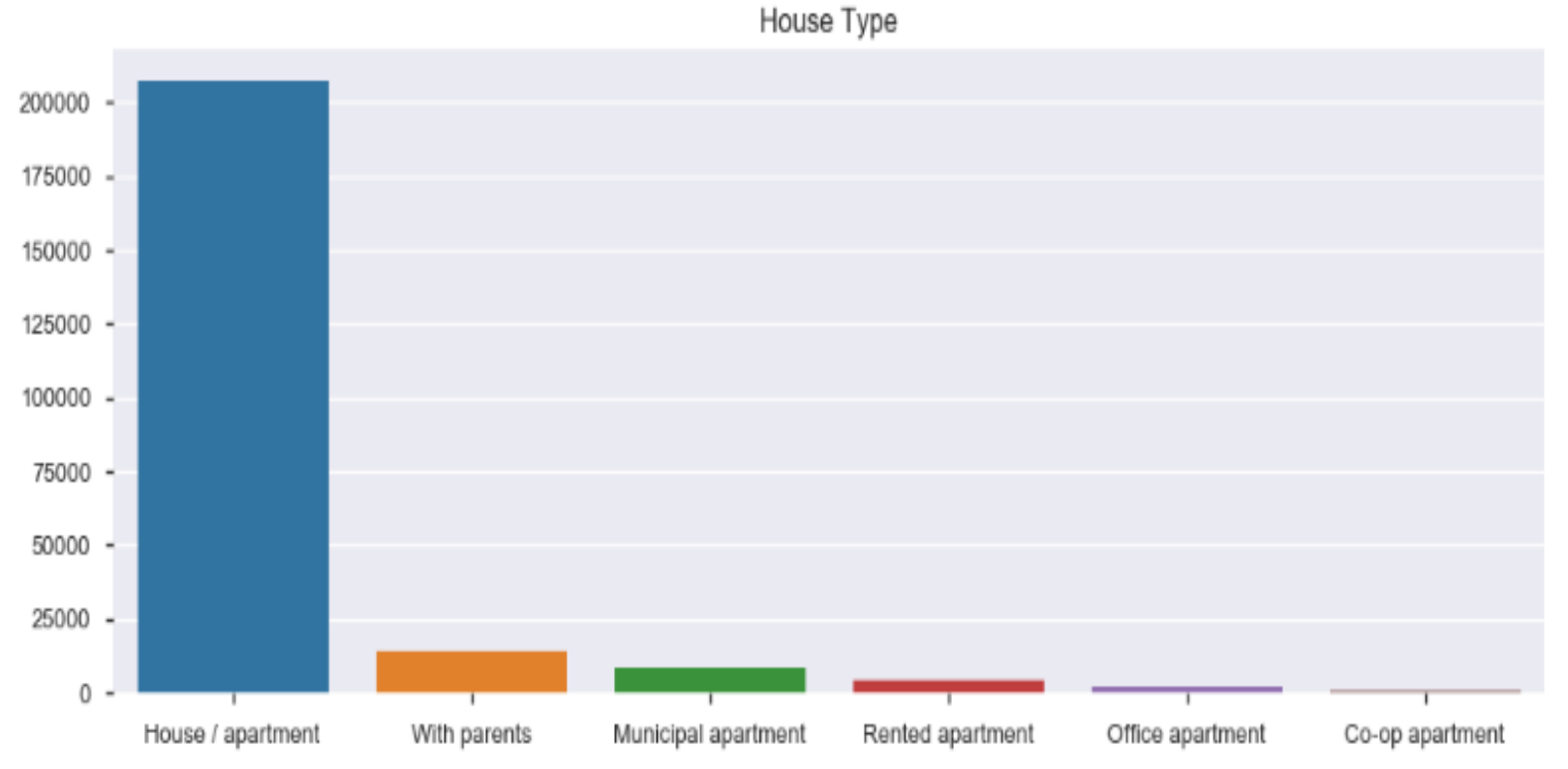
- **Variation with education and occupation type:**



Top Applicant's who applied for loan : Laborers - Approx. 52 K, Sales Staff - Approx. 31 K, Core staff - Approx. 25 K. Entity Type 3 type organizations have filed maximum number of loans equal to approx. 64K

Univariate, Bivariate Analysis and Data Visualization

- **Analysis the target variable with Housing situation of the applicant:**

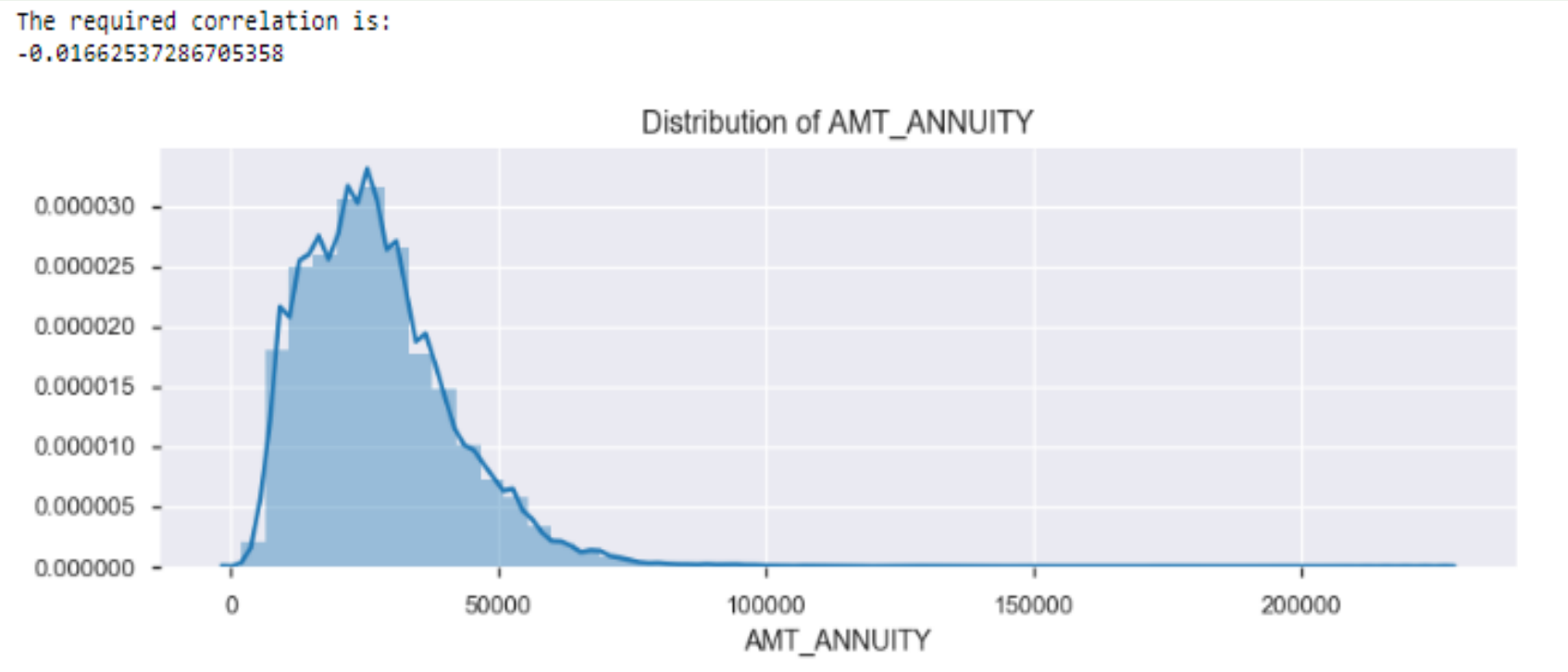


Clients with their own house/apartment(approx. 200k) are mostly applying for loan followed by applicants staying with parents.

Univariate, Bivariate Analysis and Data Visualization

- **Distribution of AMT_ANNUITY: Loan annuity**
- It represent the series of payments to be made (instalments) towards the loan amount.

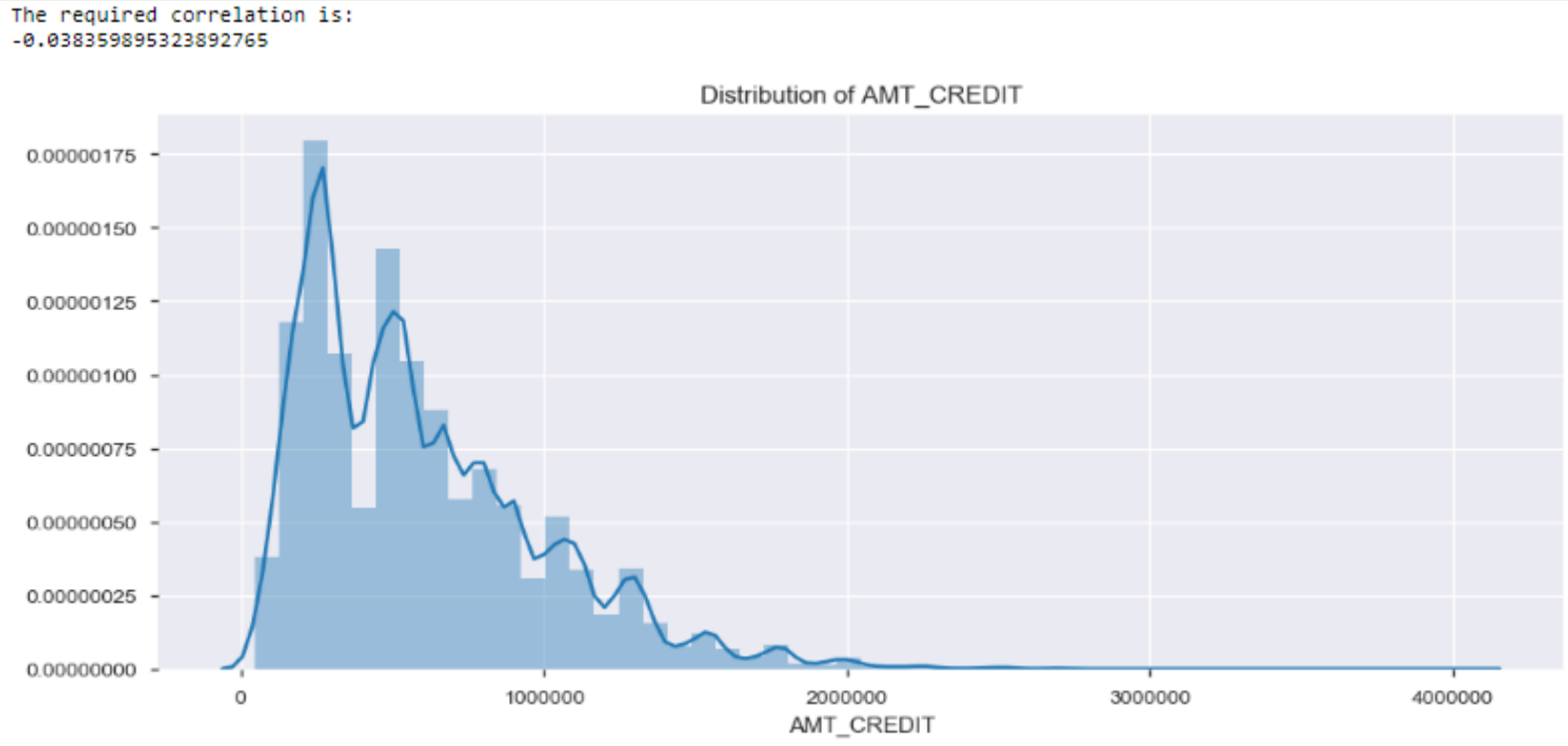
```
The required correlation is:
-0.01662537286705358
```



There is weak negative correlation (-0.0166) with the target variable of AMT_ANNUITY.

Univariate, Bivariate Analysis and Data Visualization

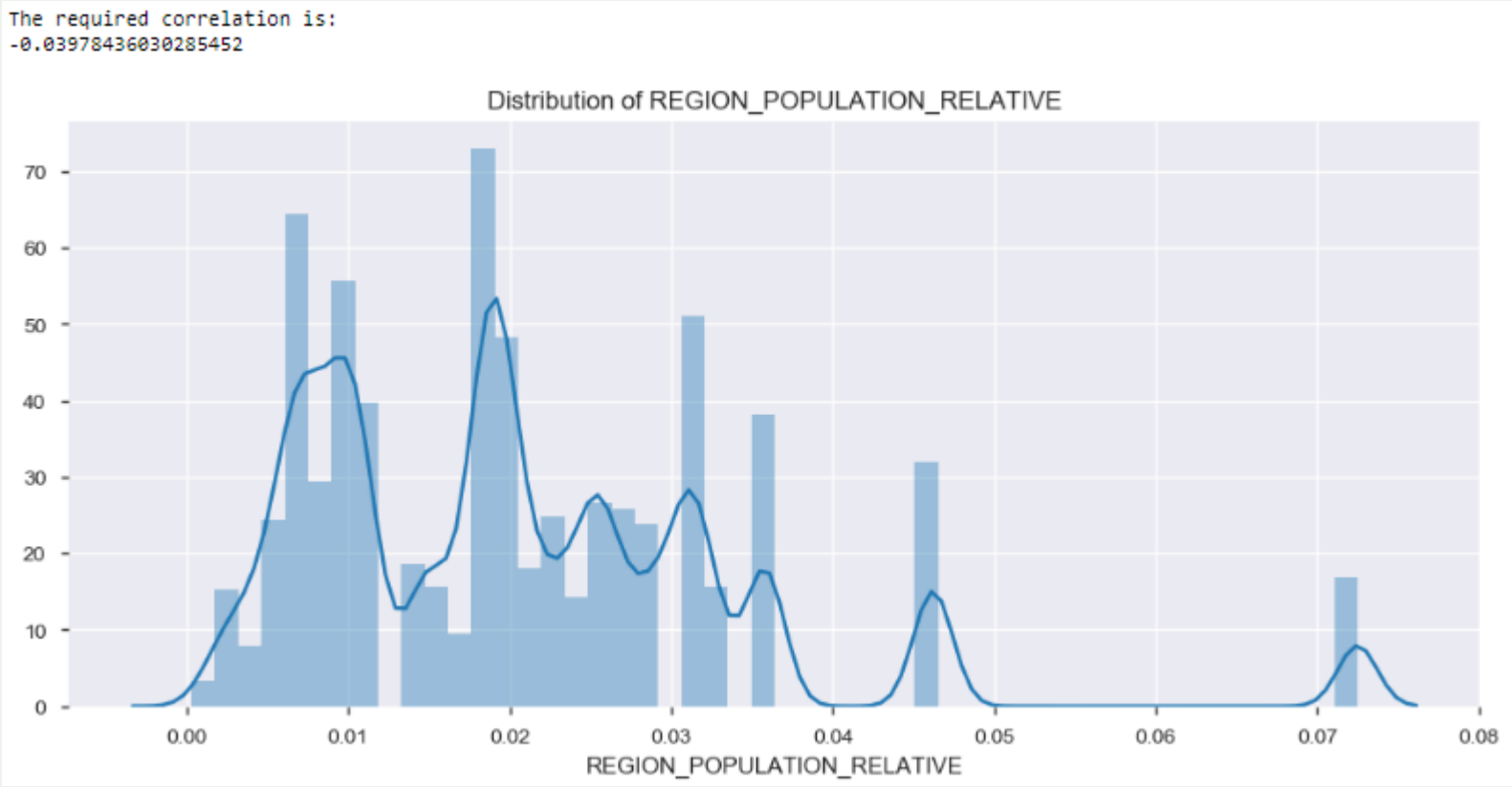
- **Distribution of AMT_CREDIT: Credit amount of the loan**
- It represent TOTAL loan amount.



There is weak negative correlation (-0.0383) with the target variable of **AMT_CREDIT**.

Univariate, Bivariate Analysis and Data Visualization

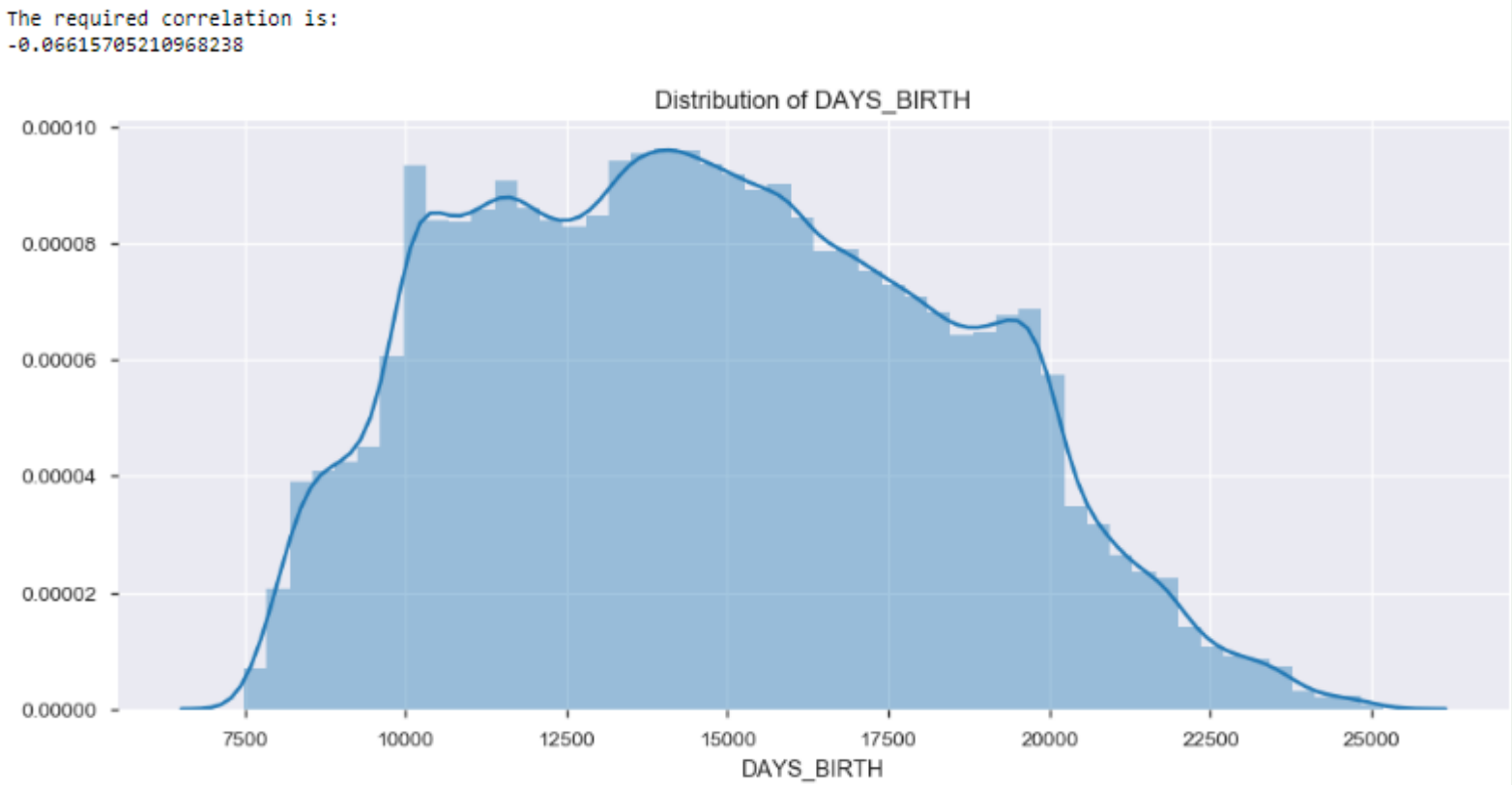
- **Distribution of REGION_POPULATION_RELATIVE :**
- It represent Normalized population of region where client lives (higher number means the client lives in more populated region)



There is weak negative correlation (-0.0397) with the target variable of **REGION_POPULATION_RELATIVE** .

Univariate, Bivariate Analysis and Data Visualization

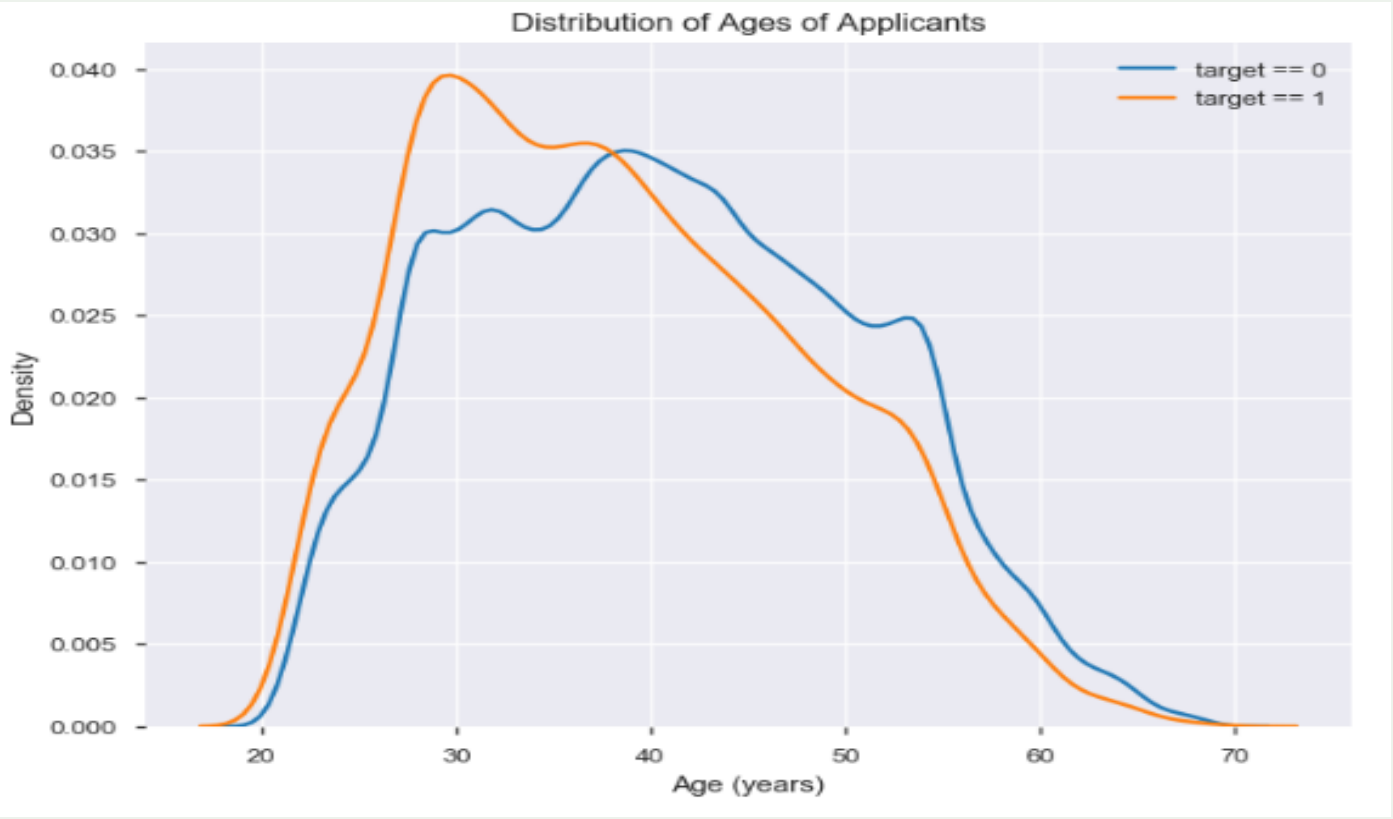
- **Distribution of DAYS_BIRTH(Client's age in days at the time of application):**



There is weak negative correlation (-0.0397) with the target variable of **DAYS_BIRTH** , but this is stronger compared to variables we have tested so far so we will perform more analysis for this important variable.

Univariate, Bivariate Analysis and Data Visualization

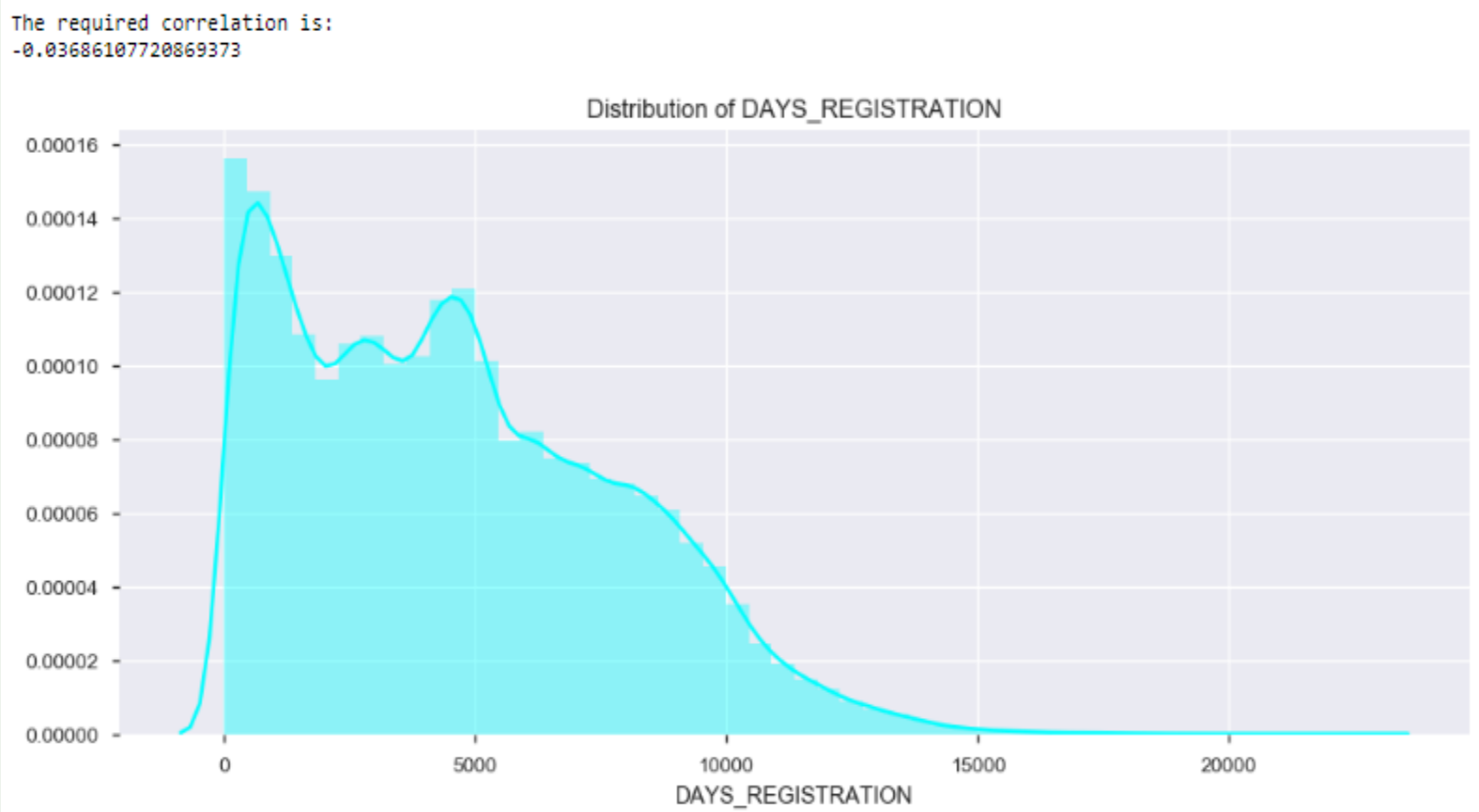
- **Distribution of DAYS_BIRTH(Client's age in days at the time of application):**



The target == 1 curve skews towards the younger end of the range, hence applicants with more age are more likely to repay the loan taken up to some extent as the correlation is weak.

Univariate, Bivariate Analysis and Data Visualization

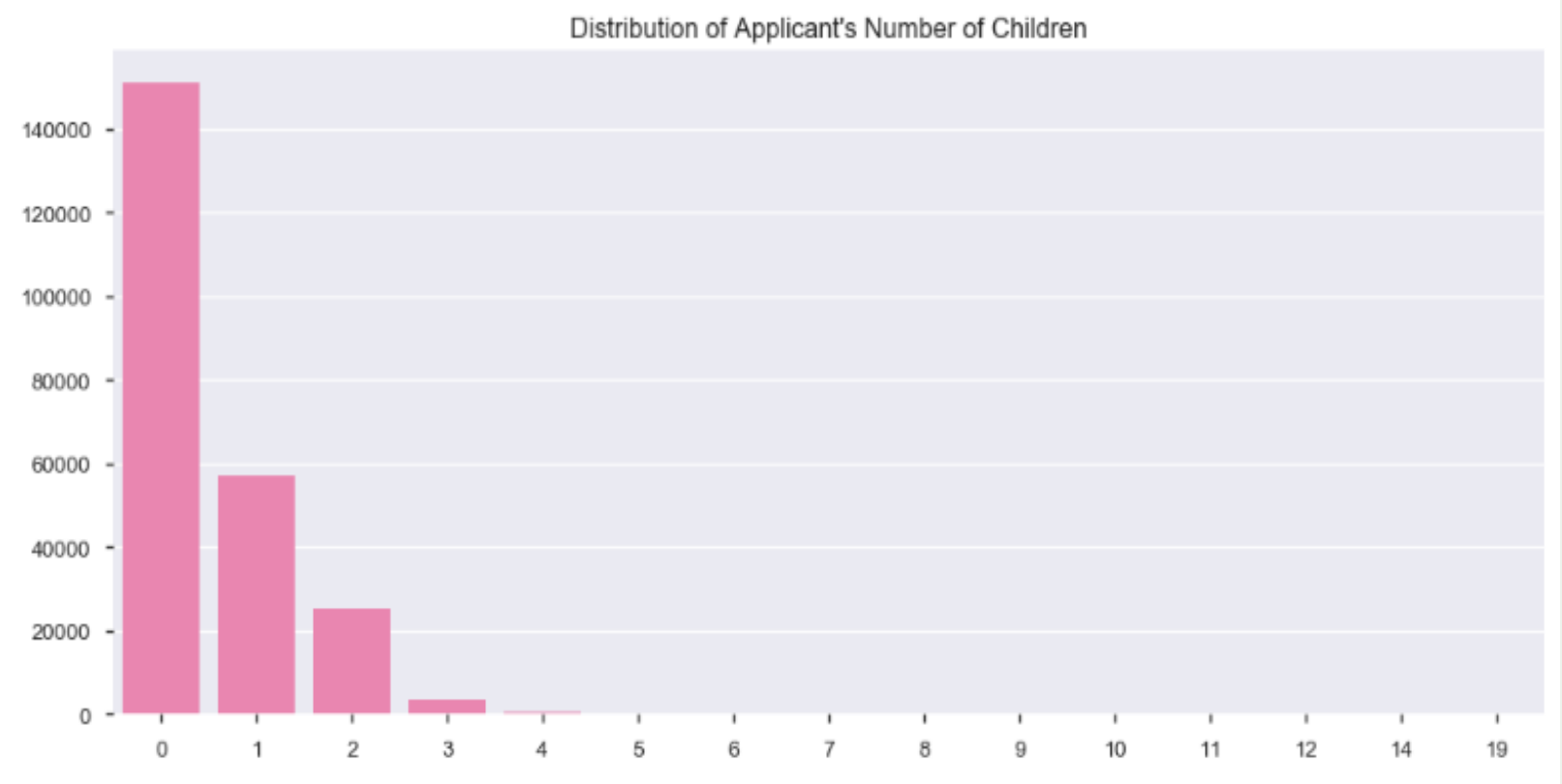
- **Distribution DAYS_REGISTRATION(How many days before the application did client change his registration):**



There is very weak negative correlation (-0.0368) with the target variable of **DAYS_REGISTRATION**.

Univariate, Bivariate Analysis and Data Visualization

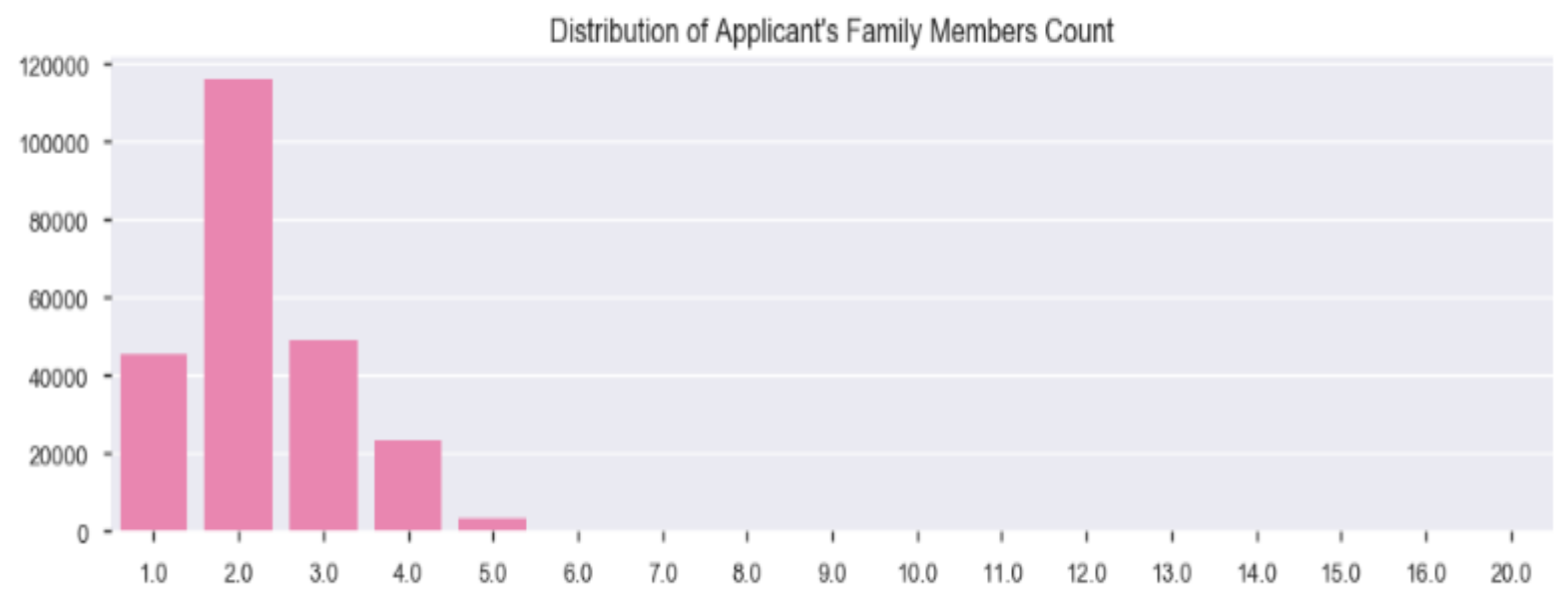
- **Distribution CNT_CHILDREN (Number of children the client has):**



Most of the clients have no children (around 145k), followed by 1 children (approx. 60k)

Univariate, Bivariate Analysis and Data Visualization

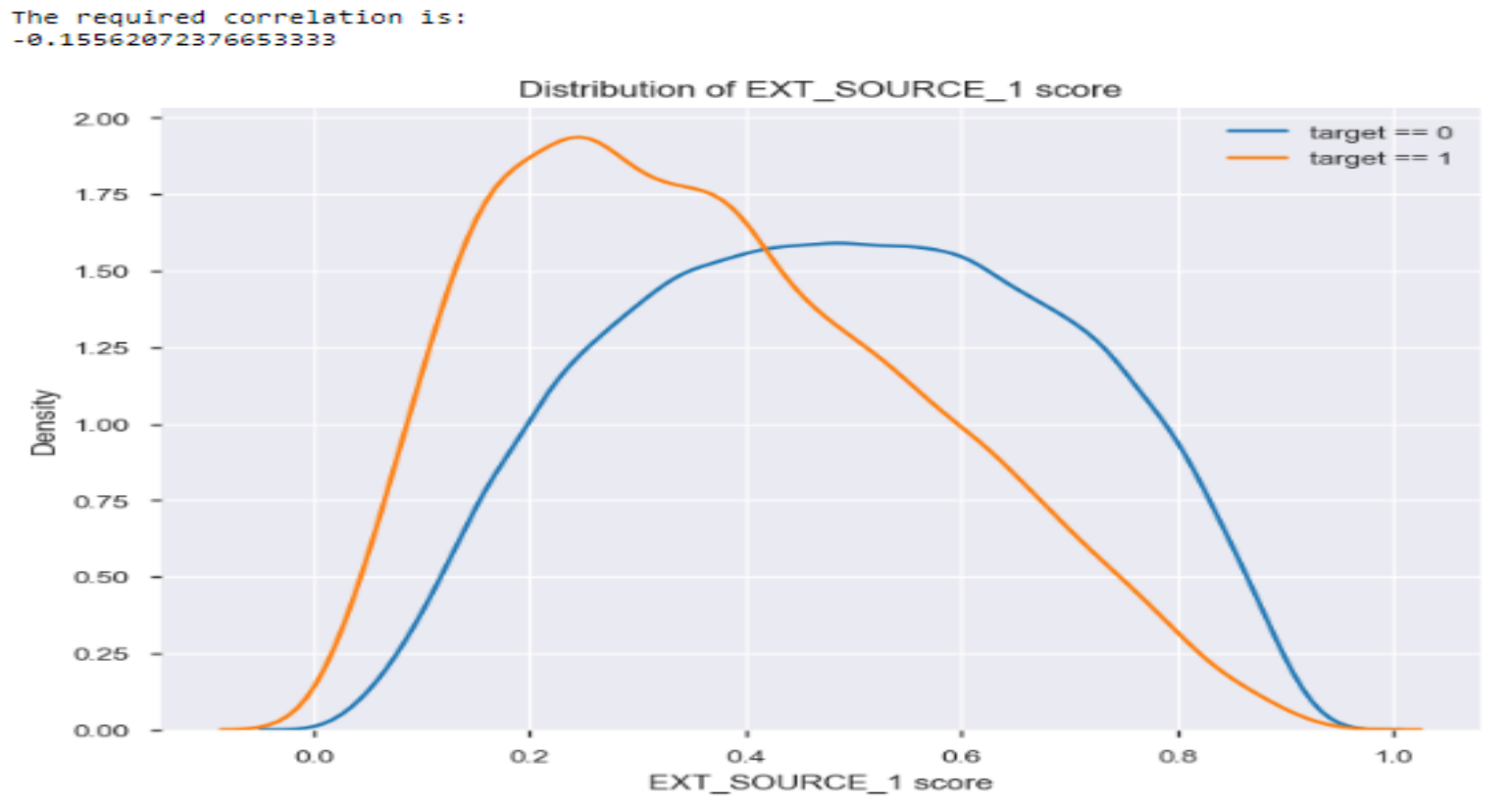
- **Distribution CNT_FAM_MEMBERS (Number of family member of client):**



Most of the clients have 2 members in family (approx. 120k) followed by 3 and 1 members in family.

Univariate, Bivariate Analysis and Data Visualization

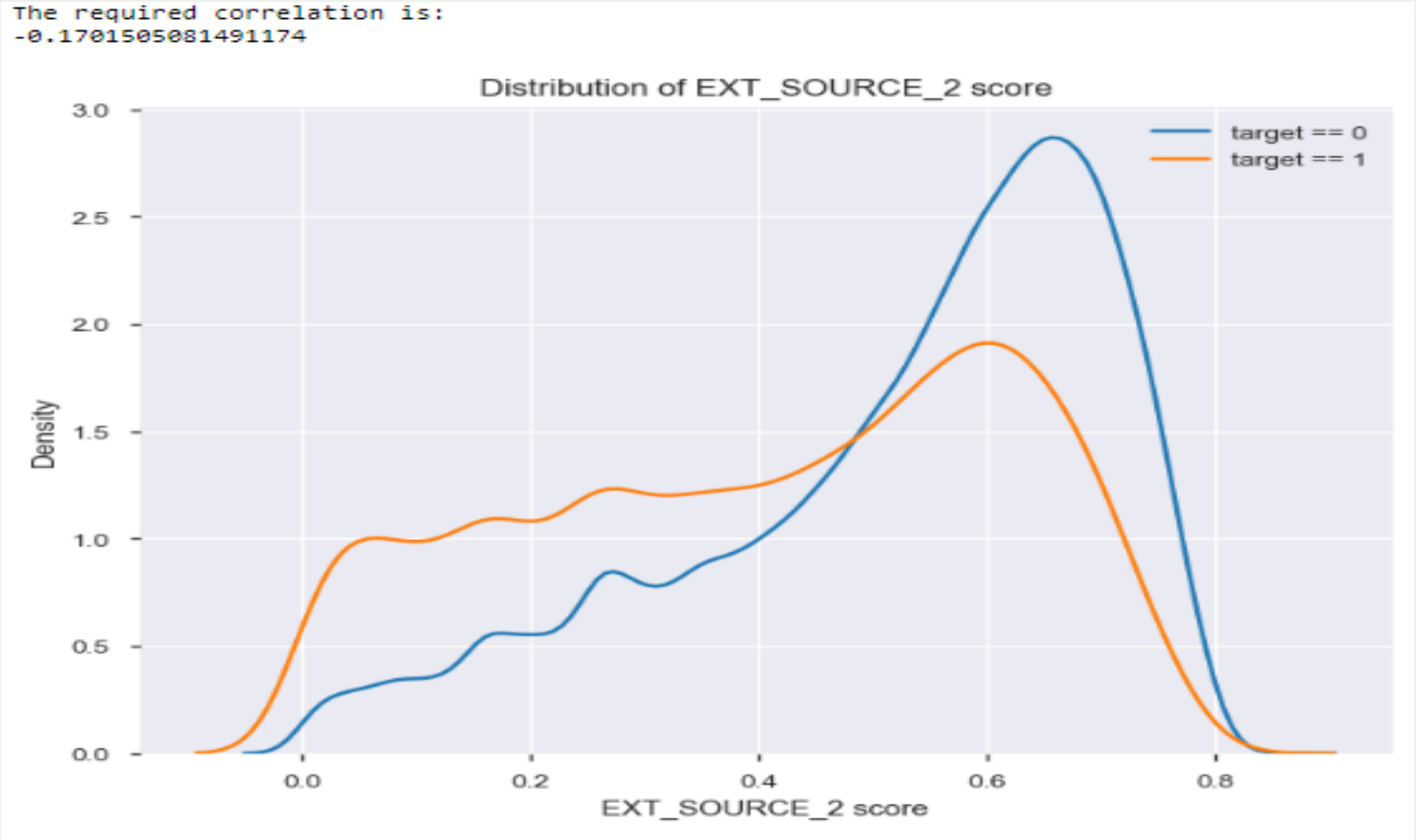
- **Distribution EXT_SOURCE_1 (Normalized score from external data source 1):**



There is weak negative correlation (-0.155) with the target variable of **EXT_SOURCE_1** .

Univariate, Bivariate Analysis and Data Visualization

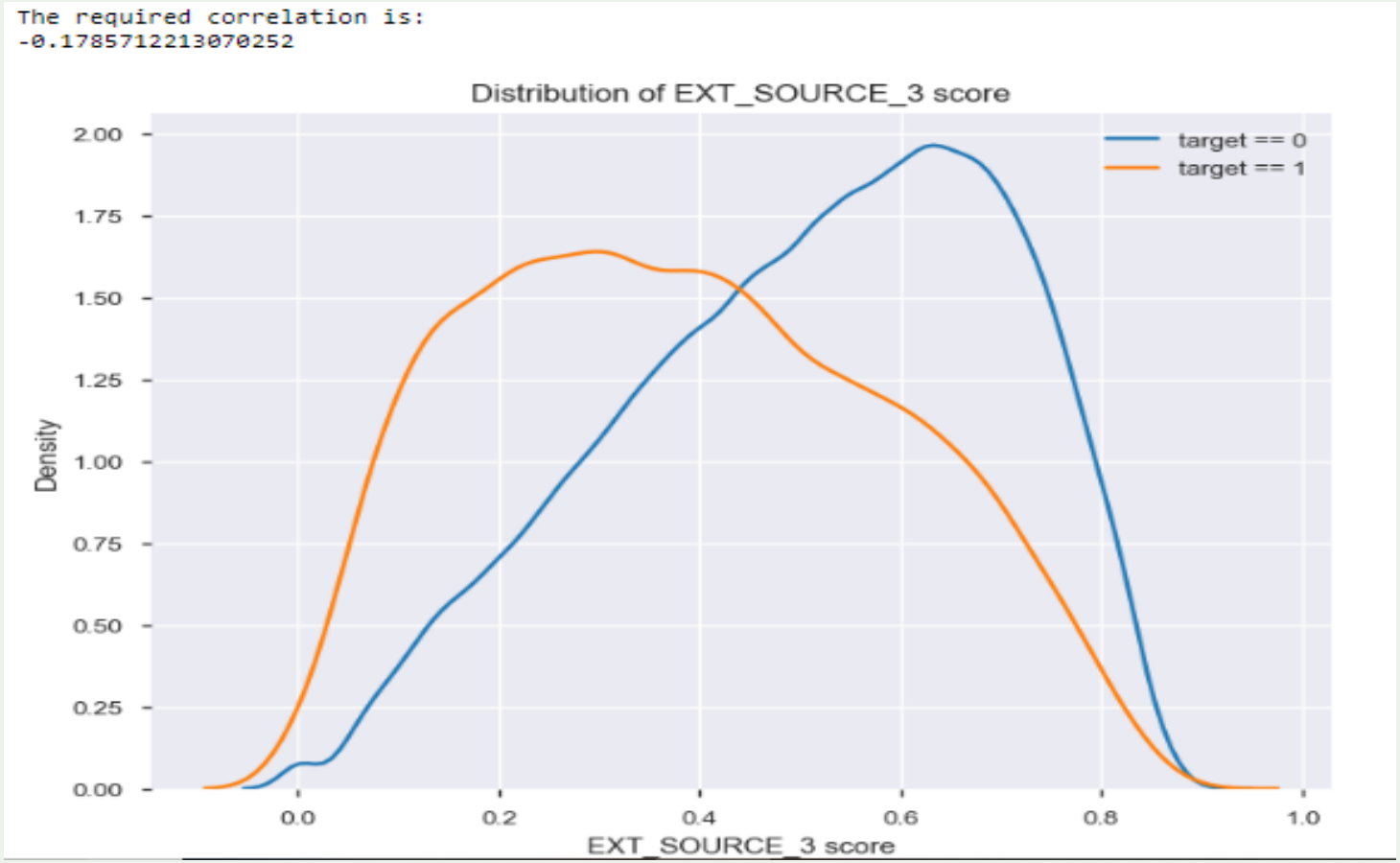
- **Distribution EXT_SOURCE_2 (Normalized score from external data source 2):**



There is weak negative correlation (-0.170) with the target variable of **EXT_SOURCE_2** .

Univariate, Bivariate Analysis and Data Visualization

- **Distribution EXT_SOURCE_3 (Normalized score from external data source 3):**



There is weak negative correlation (-0.178) with the target variable of **EXT_SOURCE_3** .

Univariate, Bivariate Analysis and Data Visualization

- **Checking how the correlation looks like after data cleansing, preparation and engineering**

Univariate, Bivariate Analysis and Data Visualization

- **Checking how the correlation looks like after data cleansing, preparation and engineering**
  - Frankly, there was no clear picture from the heatmap above as there are too many variables present which makes making decision very difficult and visualization looks quite clumsy.
  - We will find the correlation with respect to the target variable “TARGET” and will consider only the variables with relatively stronger correlation for our final analysis.
  - As evident from the above analysis, below variables are having highest and lowest correlation with target variable:

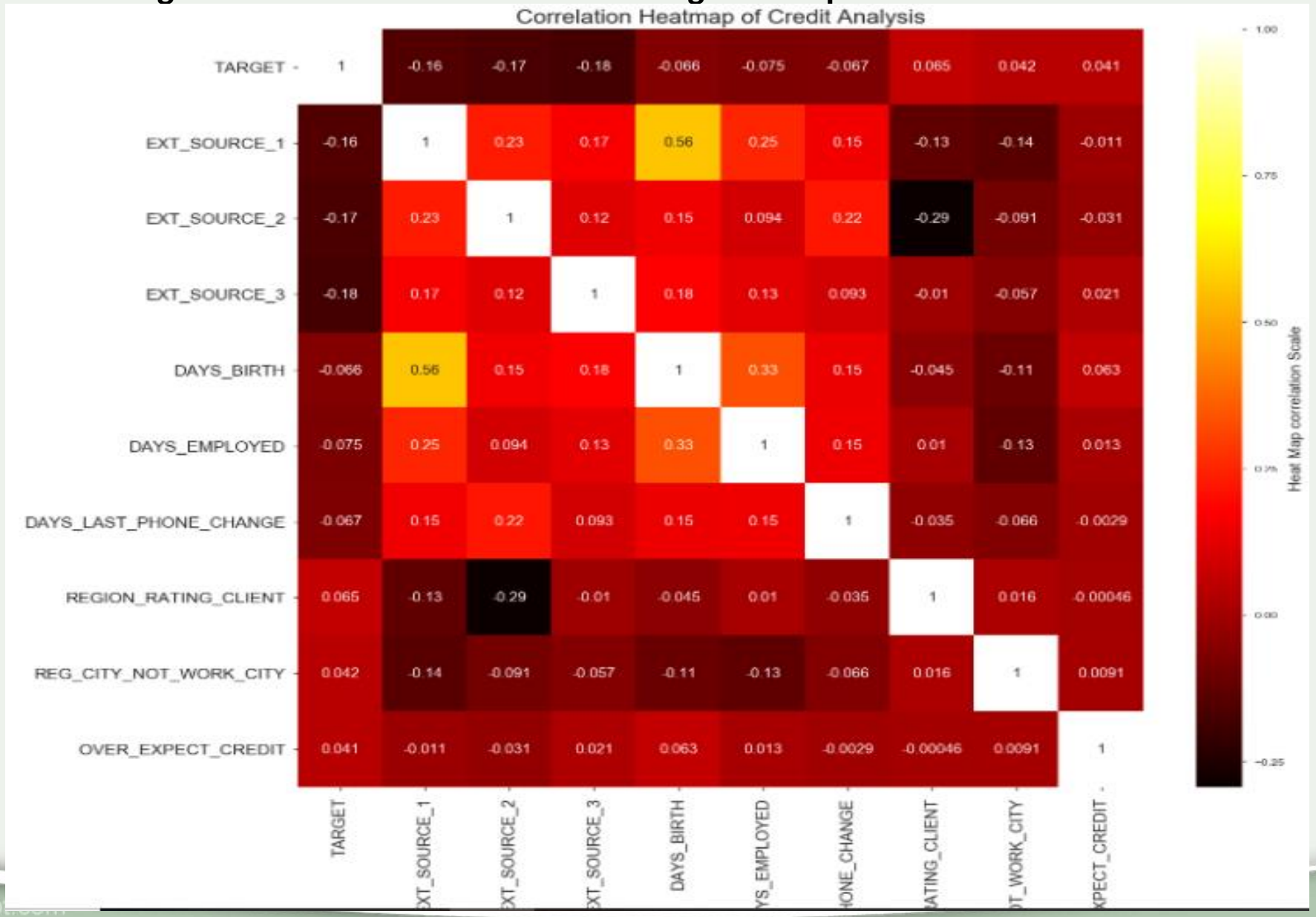
  - **Most Negative Correlations:**
  - #EXT_SOURCE_3 -0.178571
  - #EXT_SOURCE_2 -0.170151
  - #EXT_SOURCE_1 -0.155621
  - #DAYS_EMPLOYED -0.074915
  - #DAYS_LAST_PHONE_CHANGE -0.067110
  - #DAYS_BIRTH -0.066157
  - **Most Positive Correlations:**
  - #REGION_RATING_CLIENT 0.064646
  - #REG_CITY_NOT_WORK_CITY 0.042304
  - #REG_CITY_NOT_LIVE_CITY 0.041903
  - #OVER_EXPECT_CREDIT 0.041403

Univariate, Bivariate Analysis and Data Visualization

- **Checking how the correlation looks like using pair-wise scatter plot for selected variables:**

Univariate, Bivariate Analysis and Data Visualization

- **Checking how the correlation looks like using heatmap for selected variables:**



Conclusion from Case Study

- All **three EXT_SOURCE** features, **DAYS_EMPLOYEED**, **DAYS_BIRTH** of applicants shares the **negative correlations** with the target,
- It indicates that **as the value of these parameters increases, the client is more likely to repay the loan.**
- The information obtained from external data source(**EXT_SOURCE**), the age of applicant (**DAYS_BIRTH**) and employment duration of the applicant (**DAYS_EMPLOYEED**) is important information in determining if the loan should be granted or not and **high values are desirable for the loan approval.**

- On the other hand, there is a **very weak positive correlation** with **REGION_RATING_CLIENT**, **REG_CITY_NOT_WORK_CITY, REG_CITY_NOT_LIVE_CITY** and **derived variable OVER_EXPECT_CREDIT** with the target, **as they increase the client is less likely to pay the loan.**
- Here, **REGION_RATING_CLIENT** = rating of the region where client lives, **REG_CITY_NOT_WORK_CITY** = Flag if client's permanent address does not match work address, **REG_CITY_NOT_LIVE_CITY** = Flag if client's permanent address does not match contact address, **OVER_EXPECT_CREDIT** = actual credit larger than goods price.
- Hence **more risk is involved with increase in any of these parameters**, so credit agency should re-consider when any of these variables values is on the higher side.
- The **lower values of these variables are desirable for loan approval.**

References and Interpretations

- Some general interpretations of the absolute value of the correlation coefficient are:

00-.19 “very weak”
20-.39 “weak”
40-.59 “moderate”
60-.79 “strong”
80-1.0 “very strong”

- Both the positive and negative correlations falls under the category of **very weak correlation with the target variable in our case study**.

THANKS !!