HOME CREDIT DEFAULT RISK

Milestone: Data Exploration and Visualization

Group 7

Raghavi Dube

Sai Shanmukha Bharadwaj Palakodeti

+1 (908)-723-7137 (Raghavi Dube)

+1 (608)- 381- 6259 (Bharadwaj Palakodeti)

Dube.ra@northeastern.edu

Palakodeti.s@northeastern.edu

Percentage of effort contributed by student 1 – 50%

Percentage of effort contributed by student 2 – 50%

Signature of student 1 – RAGHAVI DUBE

Signature of student 2 – BHARADWAJ PALAKODETI

Submission Date - 03.05.2023

- We have split the entire data into training and validation data. The training data consists of 122 columns and 307511 rows.

```
app_train.head()
```

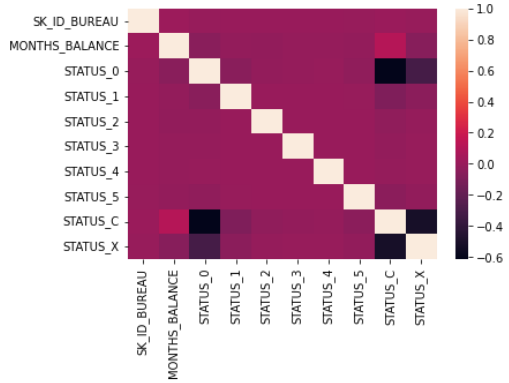
```
Training data shape:  (307511, 122)
```

[3]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDI
0	100002	1	Cash loans	M	N	Y	0	202500.0	406597.
1	100003	0	Cash loans	F	N	N	0	270000.0	1293502.
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000.
3	100006	0	Cash loans	F	N	Y	0	135000.0	312682.
4	100007	0	Cash loans	M	N	Y	0	121500.0	513000.

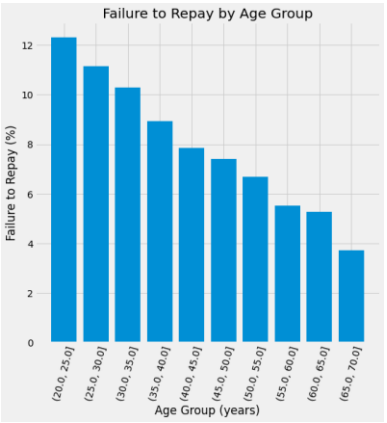
5 rows × 122 columns

- To understand the relationship and dependency amongst various variables in the dataset, we plot a heatmap.



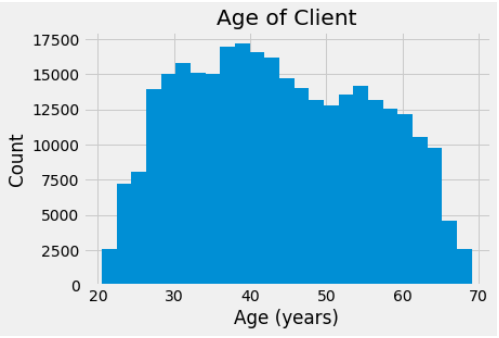
Using this heatmap, we can clearly analyse variable dependency of each variable.

- To understand the relation between failed repayments and age variables, we plotted a bar graph.



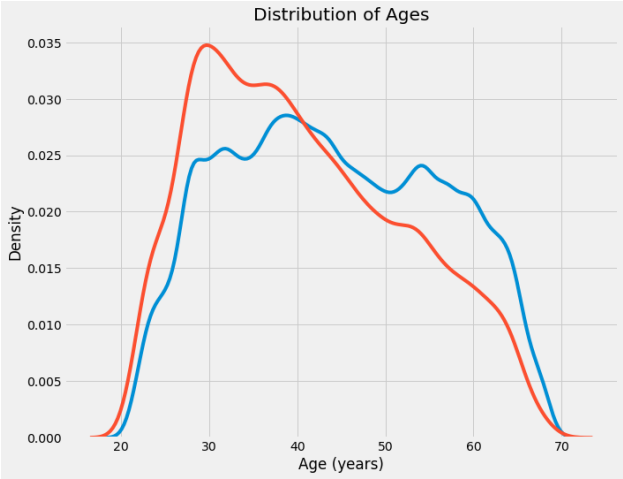
From this plot, we can conclude that the age range from 20 -25 years has the most number of home loan defaulters while the most repayments are made by the age range of 65-70 years.

- We plot a histogram to understand the client age group of the given data.



This histogram shows us that the most number of clients come from the age range of 35-45 years while least home loans were taken by people in the age range of 20-23 and 67 – 70 years.

- To further understand the age distribution of clients, we plot a distribution plot.



- Handling the missing values by filling the null records by the column mean.

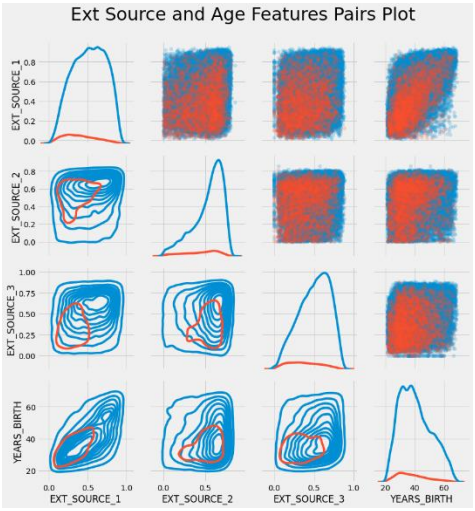
```
col = list(bureau.columns)
for each in col:
    if bureau.dtypes[each]!=np.object:
        bureau[each].fillna((bureau[each].mean()), inplace=True)
```

```
bureau.isnull().sum()
```

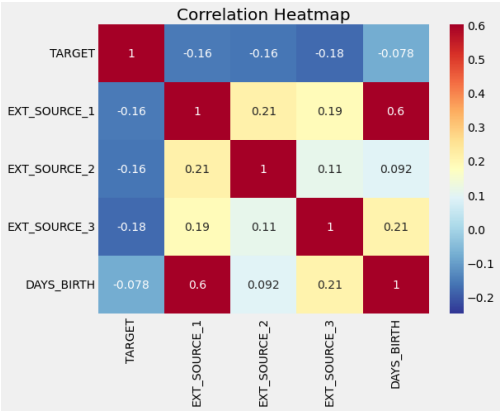
```
SK_ID_CURR                0
SK_ID_BUREAU              0
CREDIT_ACTIVE             0
CREDIT_CURRENCY           0
DAYS_CREDIT               0
CREDIT_DAY_OVERDUE        0
DAYS_CREDIT_ENDDATE       0
CNT_CREDIT_PROLONG        0
AMT_CREDIT_SUM            0
AMT_CREDIT_SUM_DEBT       0
AMT_CREDIT_SUM_OVERDUE    0
CREDIT_TYPE               0
DAYS_CREDIT_UPDATE        0
dtype: int64
```

To develop a model which is accurate and to reduce errors, it is important to handle all the null values in the data.

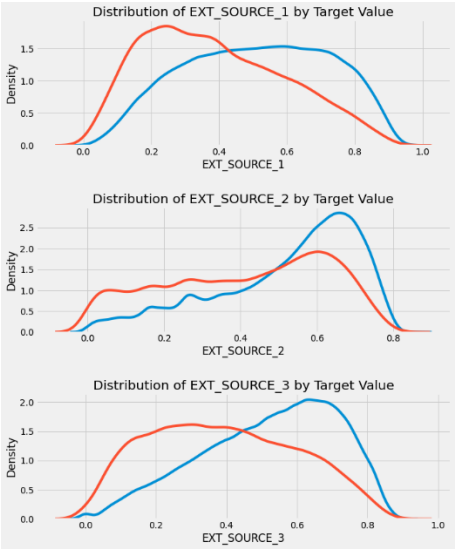
- To understand the relationship between age and ext source, we plot pair plots. This visualization helps us to understand the co-dependency of these variables easily.



- We majorly focus on two variables – TARGET VALUE and EXT SOURCE.
  To explore these variables and their relationship to a greater extent, we use the following visualizations –
  1) HEATMAP – To understand their co-dependency degree.



  2) DISTRIBUTION PLOT –



- We plot a distribution plot of credit income percent by target value.