Credit Exploratory Data Analysis

By:
Vignesh Kumar

Problem Statement

- To identify patterns which indicate if a client has difficulty paying their instalments.
- To understand the driving factors (or driver variables) behind loan default, the variables which are strong indicators of default.
- To Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)
- To Identify if there are outliers in the dataset and to mention why it is an outlier.
- To Identify if there is data imbalance in the data. Find the ratio of data imbalance.
- To provide univariate and bivariate analysis w.r.t to ‘Target variable’ in the dataset (clients with payment difficulties and all other cases).
- To brief the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.
- To Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable)

Steps Performed part of EDA

- DATA CLEANING
  - Missing value Imputation/ Deletion
  - Data Standardization
- DATA ANALYSIS
  - Feature Engineering
  - Dimensionality reductions
- DATA VISUALIZATION
  - Plots / Graphs using python libraries
  - Univariate / Bivariate / Correlations
- OBSERVATIONS AND FEEDBACK
  - Reasoning based on the Insights

Data Cleaning Highlights

Columns Analysis

- The application data.csv contains many (~40+) columns that have null percentage greater than or equal to 50%.
- Therefore we drop all columns with null values greater than 50%

```
1  check_cols_null_pct(curr_appl_data)
✓ 0.1s

COMMONAREA_MEDI                 69.872
COMMONAREA_AVG                  69.872
COMMONAREA_MODE                 69.872
NONLIVINGAPARTMENTS_MODE        69.433
NONLIVINGAPARTMENTS_AVG         69.433
NONLIVINGAPARTMENTS_MEDI        69.433
```

Missing Values imputation

Those columns that has null values below 50%, are imputed based on the missing type (MAR, MNAR,MCAR)

for most categorical/discrete columns, considered using Mode if the frequency of the topmost value is above 50%. (* refer Notebook for details)

for most numerical columns – considered either Mean, Median, based on distribution and skewness.

```
curr_appl_data1["EMERGENCYSTATE_MODE"].fillna(
    curr_appl_data1["EMERGENCYSTATE_MODE"].mode()[0])
```

Data Cleaning Highlights

DATA STANDARDISATION

- DAYS_COLS - Is a list of days type columns, represented in days count,
- There were inconsistencies in these columns, therefore made all values absolute, then converted it to years format.

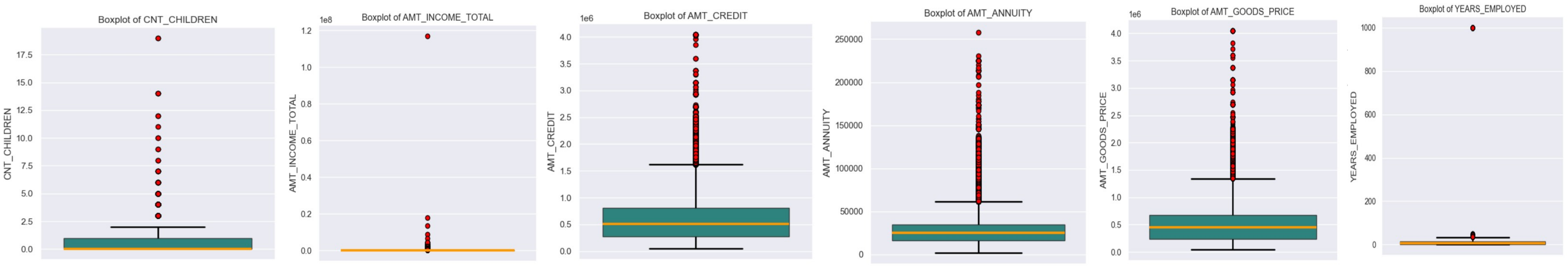
```
curr_appl_data1[DAYS_COLS] = (abs(curr_appl_data1[DAYS_COLS]) / 365).astype(int)
```

- Some categorical column values are replaced with more apt and common terms as part of standardization

```
curr_appl_data1["FLAG_OWN_CAR"] = curr_appl_data1["FLAG_OWN_CAR"].replace(to_replace=["Y", "N"], value=["Yes", "No"])
curr_appl_data1["FLAG_OWN_REALTY"] = curr_appl_data1["FLAG_OWN_REALTY"].replace(to_replace=["Y", "N"], value=["Yes", "No"])
curr_appl_data1["NAME_HOUSING_TYPE"] = curr_appl_data1["NAME_HOUSING_TYPE"].replace(to_replace="House / apartment", value="House")
```

Outlier Analysis

There are many columns that contains <u>outliers</u>:
AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_INCOME_TOTAL, CNT_CHILDREN, CNT_FAM_MEMBERS, CNT_PAYMENT, DAYS _TERMINATION, DAYS_LAST_DUE, YEARS_EMPLOYED

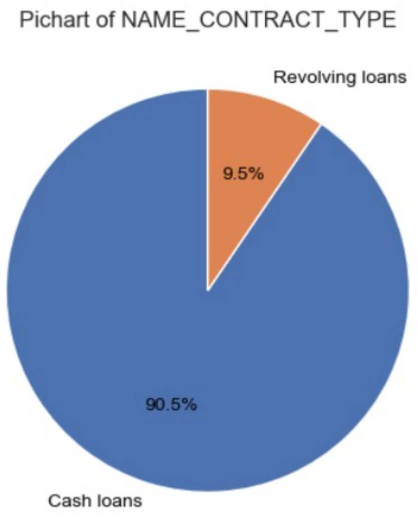


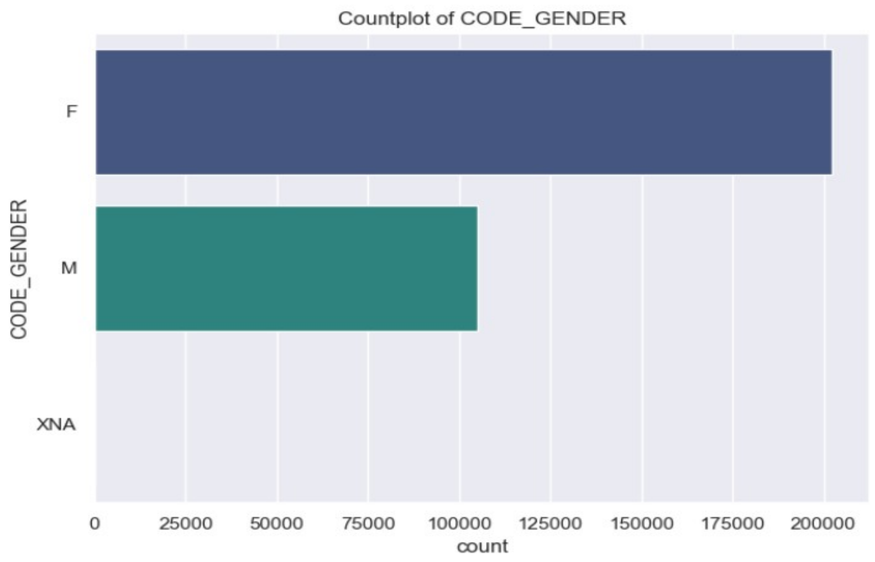
<u>Data Binning</u>
In order to minimize the outliers , we used binning approach for some of the columns:

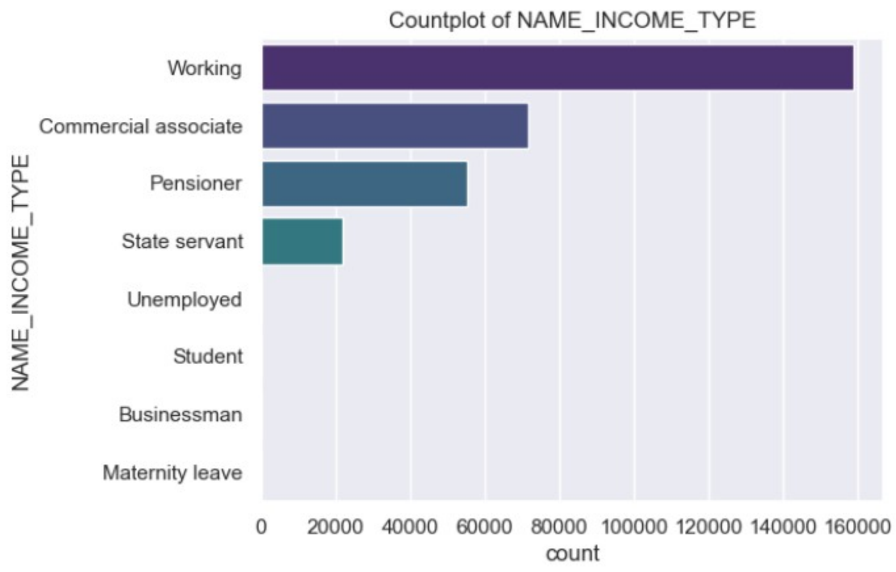
```
curr_appl_data1["AMT_CREDIT_BINS"] = pd.cut(curr_appl_data1['AMT_CREDIT'], bins=[0,200000,400000,600000,800000,1000000,10000000], labels=["0-200K","200-400k"
curr_appl_data1["YEARS_EMPLOYED_BINS"] = pd.cut(curr_appl_data1['YEARS_EMPLOYED'], bins=[-100,10,20,30,40,50,60,1000], labels=["0-10","10-20","20-30","30-40"
curr_appl_data1['AGE_Category'] = pd.cut(curr_appl_data1['YEARS_BIRTH'], [0,30,40,50,60,200], labels=["<30","30-40","40-50","50-60","60+"])
curr_appl_data1
```

Univariate Analysis

Key insights from application_data.csv







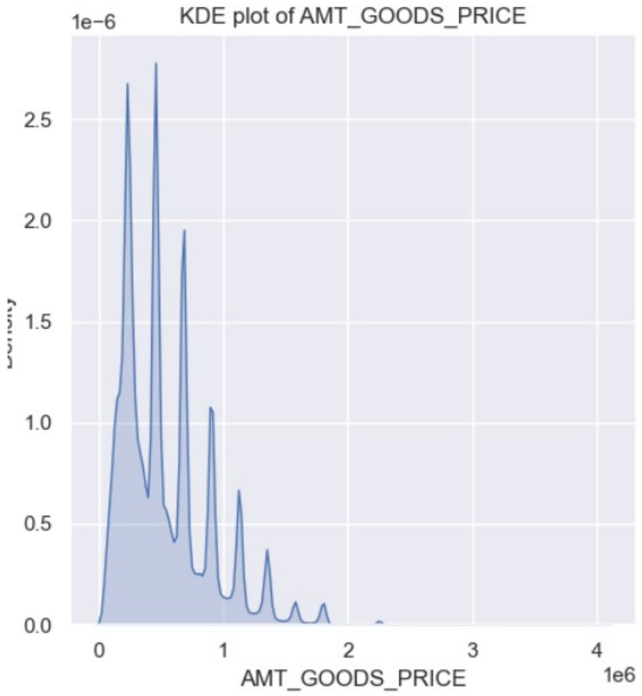
Name_Contract_Type:
Almost 90% of the loan applicants applied for Cash Loans

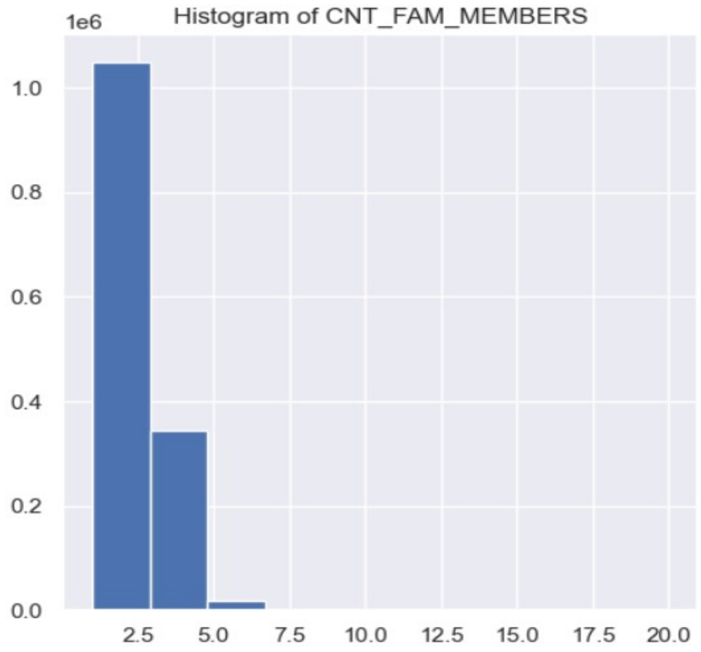
Code Gender:
67.6% of the Loan Applicants are female

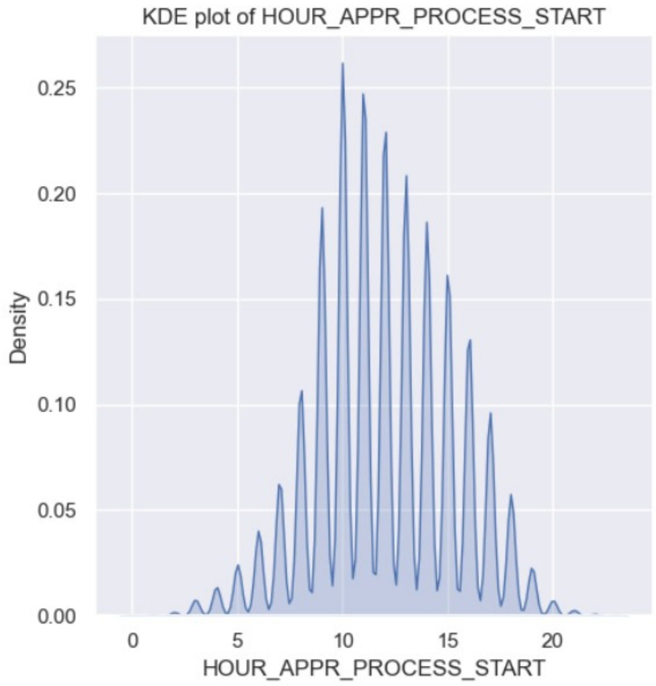
Name_Income_Type:
51% of the applicants are working citizens.

*Refer Notebook for details

Univariate Analysis

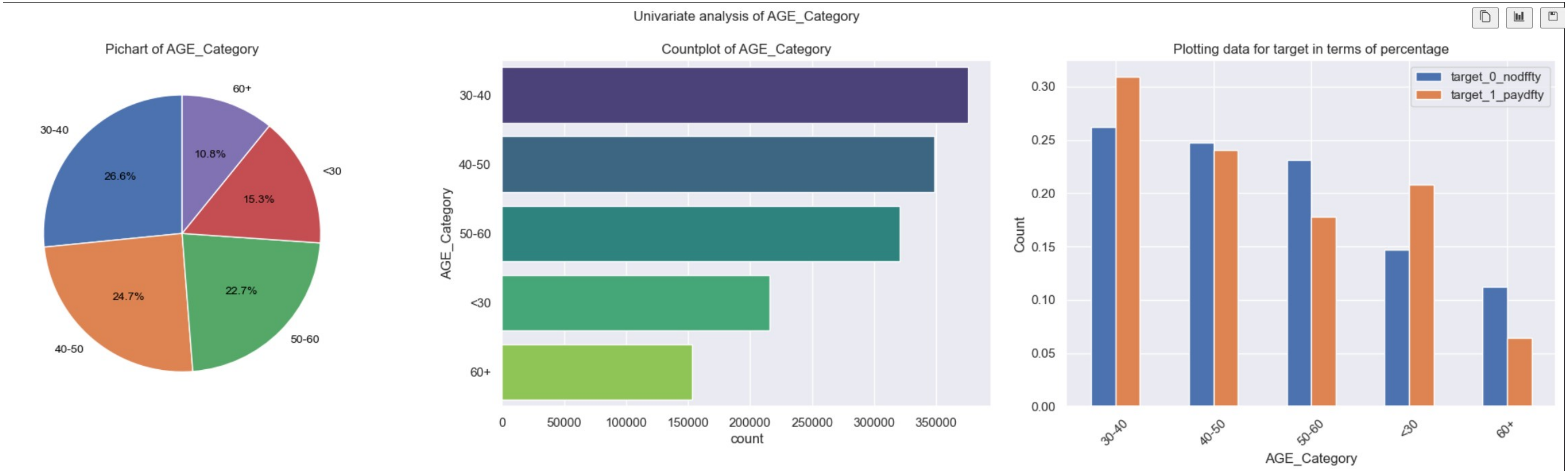






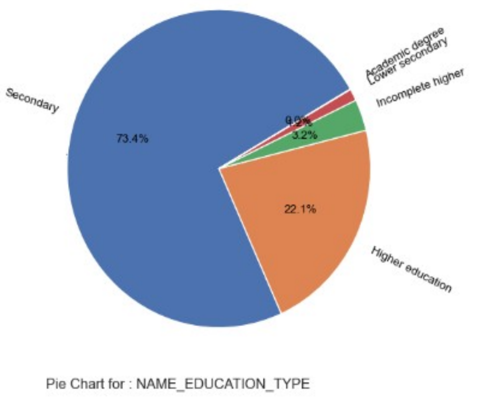
- AMT_GOODS_PRICE, CNT_FAM_MEMBERS columns have outliers and the distribution is slightly skewed towards right.
- whereas the HOUR_APPR_PROCESS_START is normally distributed

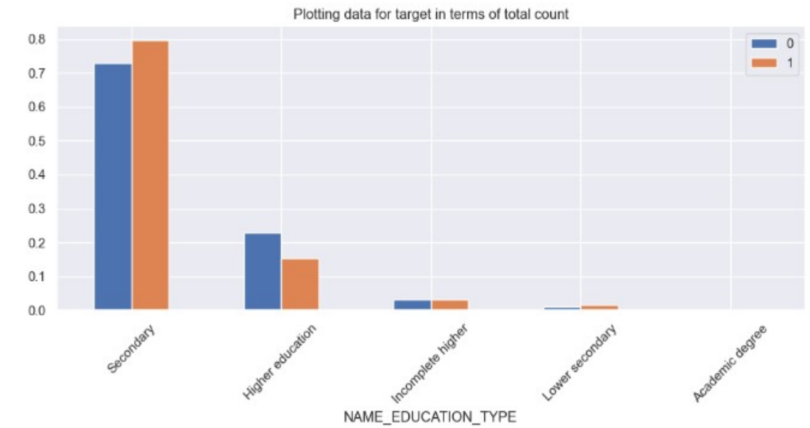
Univariate, Segmented Univariate

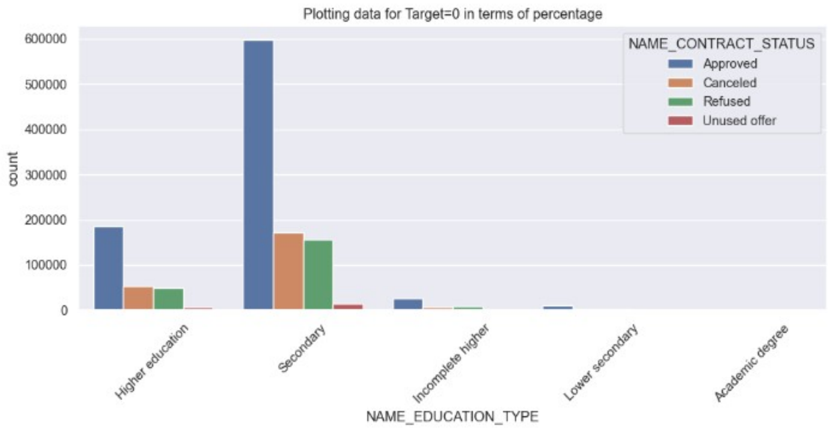


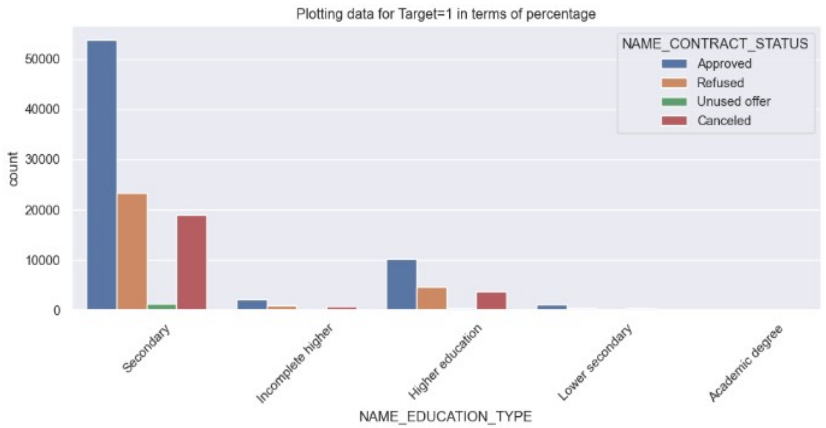
Majority of the loan applicants are between the 30-40 age group, they also have comparatively higher difficulty in loan repayment

Univariate, Segmented Univariate and Bivariate Analysis



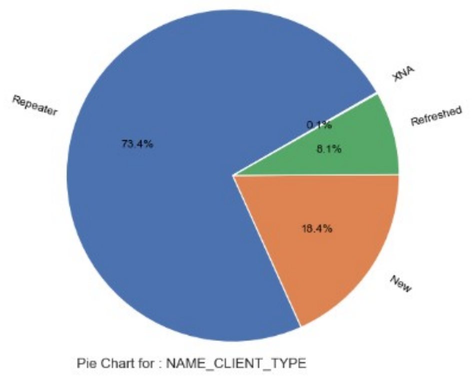


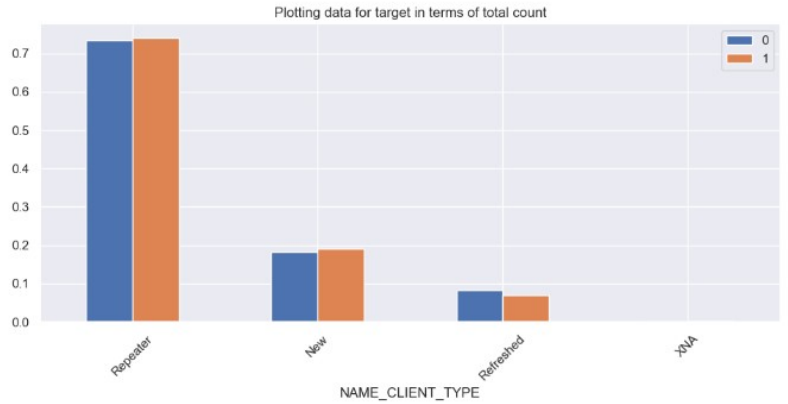


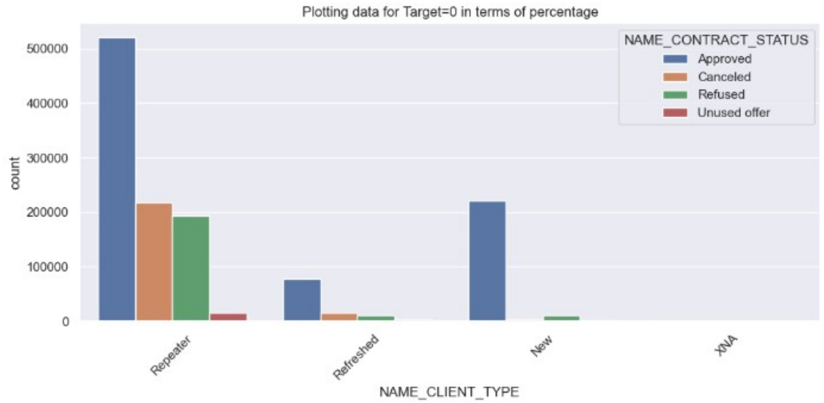


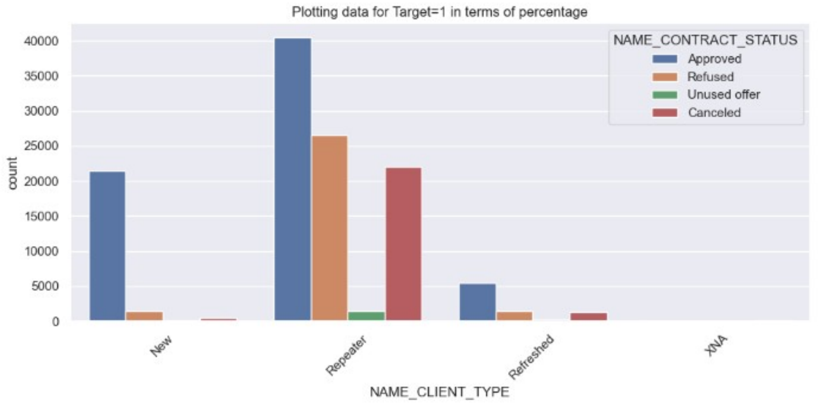
- Client who are Secondary educated have mostly applied for the loan,
- They have the greater chance of paying their installments on time without difficulties among the given categories.
- They also have higher risk of default

Univariate, Segmented Univariate and Bivariate Analysis



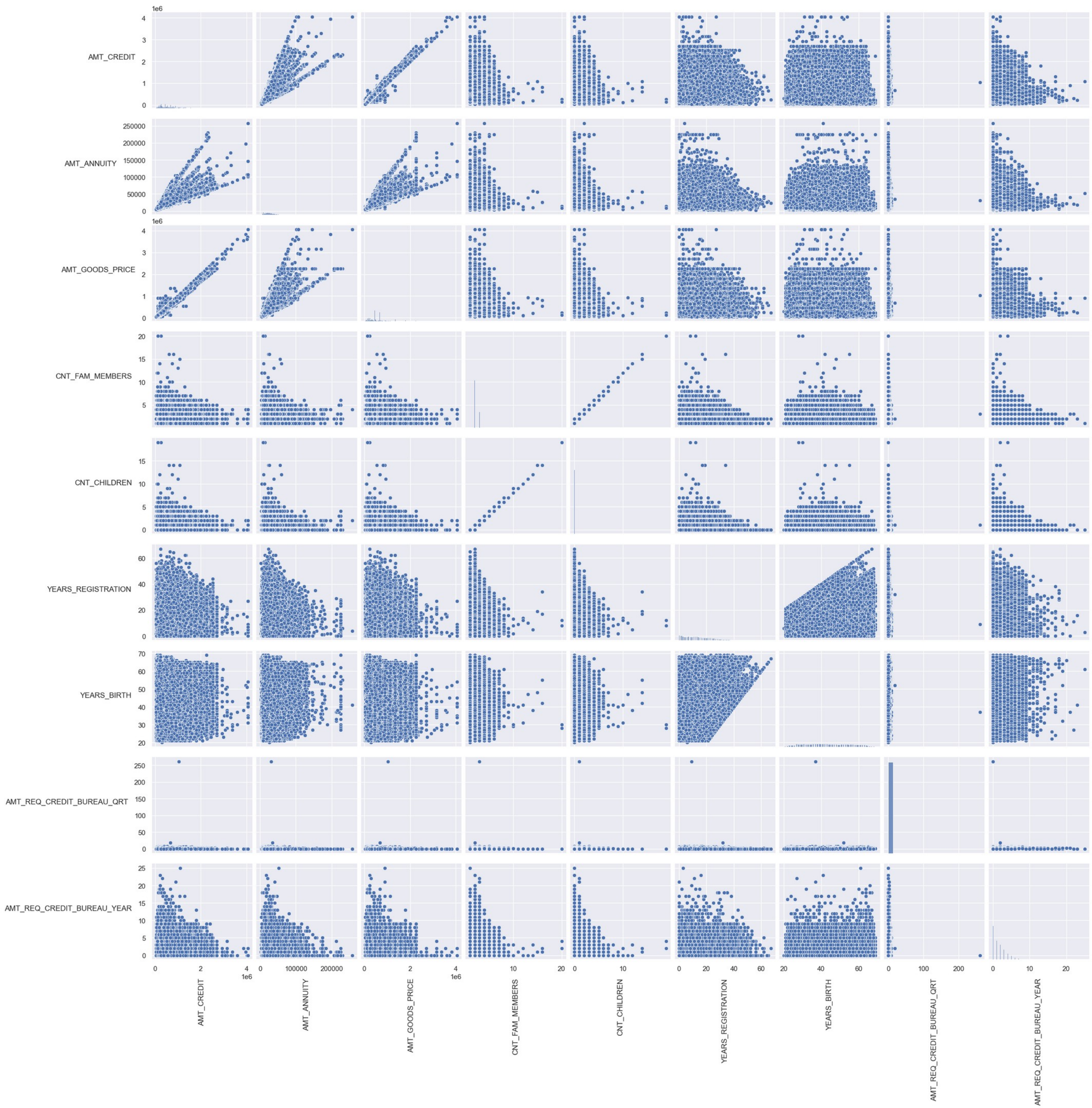






- Most loan applicants are Repeaters

- Also Most loan applicants who do not have difficulty in paying their installments are Repeaters, among the given categories.



Some of the high linear relationships observed as below:

- 'AMT_CREDIT','AMT_ANNUITY','AMT_GOODS_PRICE’,
- Therefore More the price of the goods, higher the credit amount

- 'CNT_FAM_MEMBERS','CNT_CHILDREN’,

- 'YEARS_REGISTRATION','YEARS_BIRTH’,

- 'AMT_CREDIT vs 'AMT_REQ_CREDIT_BUREAU_YEAR’,

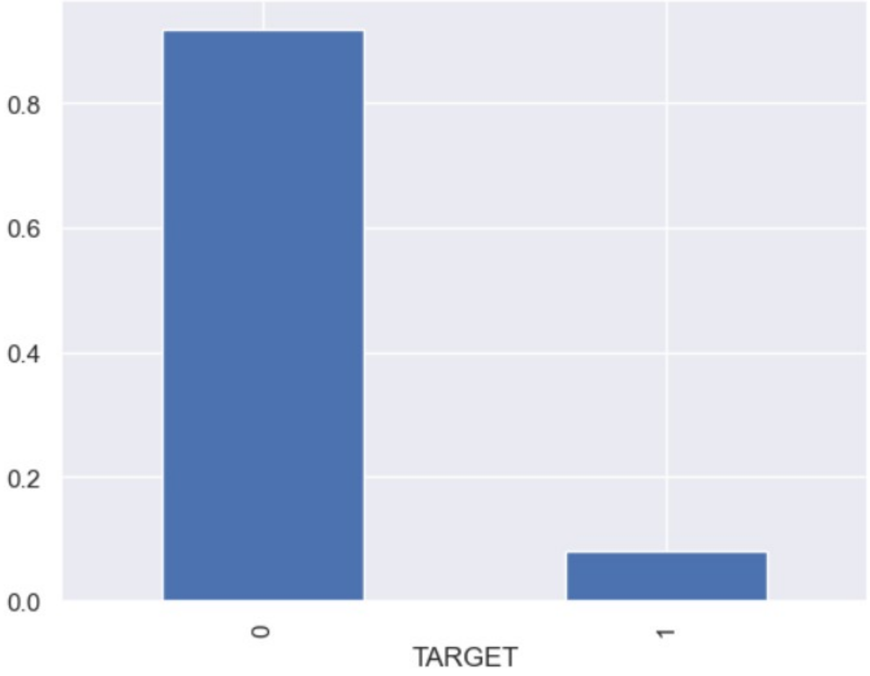
Data Imbalance Ratio

Upon analyzing the Target variable, we can conclude that the Data is **Highly Imbalanced.**
The Imbalance ratio for ‘Client with Payment difficulties’ [Target] ==1 and ‘All other cases’ [Target] == 0 is **11.387.**

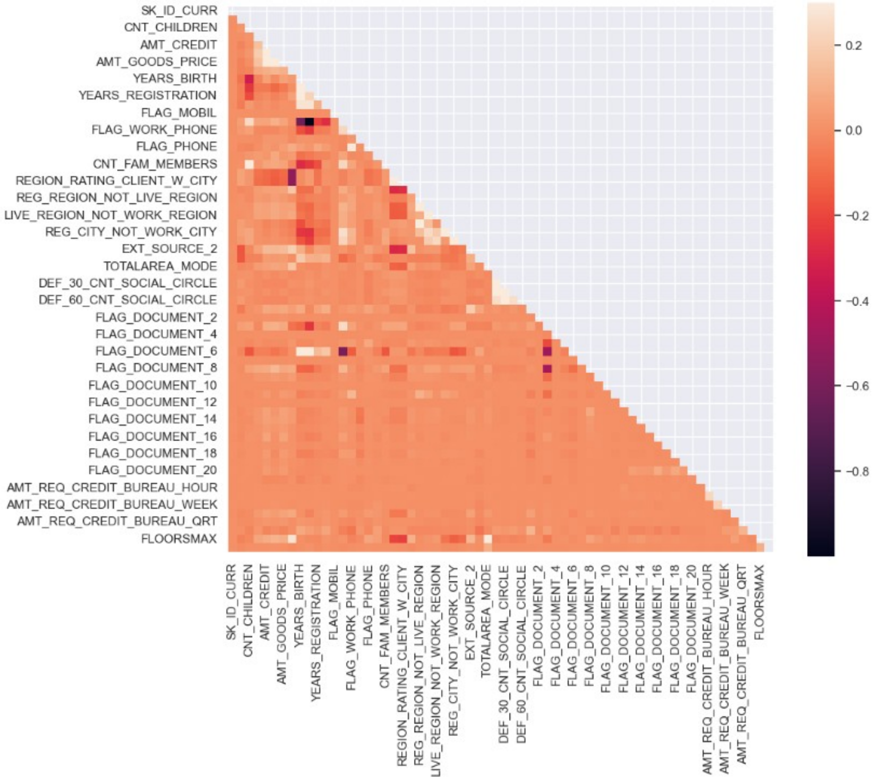
```
# curr_appl_data1['TARGET'].head(10)
target_ratio_one = (curr_appl_data1['TARGET'] == 1).sum()
target_ratio_zero = (curr_appl_data1['TARGET'] == 0).sum()
target_ratio = target_ratio_zero/target_ratio_one
target_ratio
```

✓ 0.0s

```
11.387150050352467
```

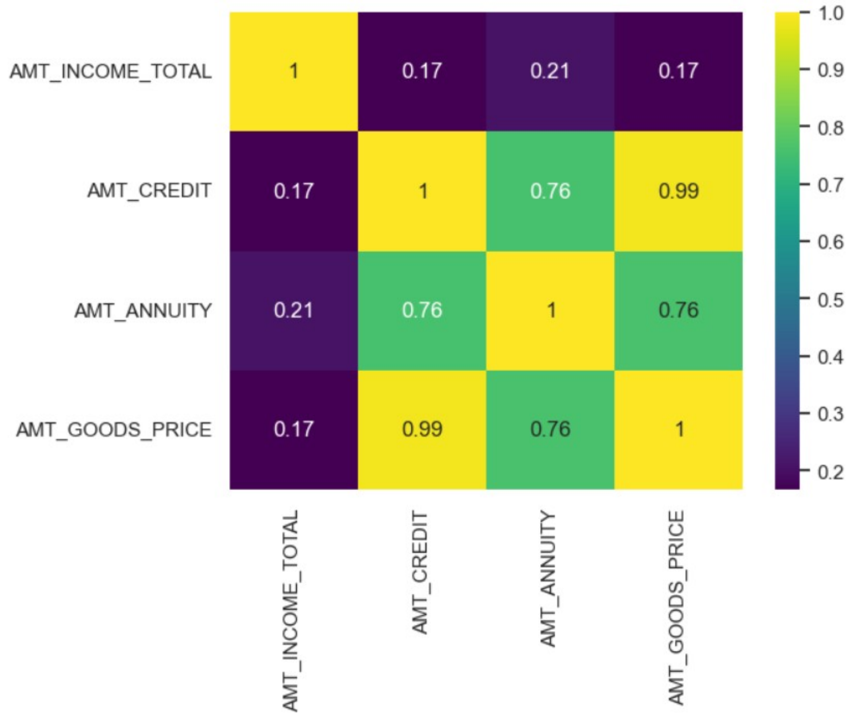


Correlation Matrix



A common Heatmap comprising of all numerical columns
It is more or less similar for both targets classes [0 & 1]

- AMT_CREDIT is high for Youngsters 'YEARS_BIRTH'
- AMT_CREDIT is high for low 'CNT_CHILDREN



A Heatmap comprising of Most critical features

Top 10 Correlations

Top 10 Correlations

```python
corr_0 = curr_target_train_0.corr(numeric_only=True).abs()
corr_0 = corr_0.unstack()
correlation_0 = corr_0.sort_values()
correlation_0 = corr_0.dropna()
# correlation_0

correlation_0 = correlation_0[correlation_0 != 1.0]
correlation_target_zero = correlation_0.reset_index()
correlation_target_zero.sort_values(by=0, ascending=False).head(10)
```

✓ 2.4s

	level_0	level_1	0
507	YEARS_EMPLOYED	FLAG_EMP_PHONE	1.000
752	FLAG_EMP_PHONE	YEARS_EMPLOYED	1.000
2014	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.999
1891	OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.999
313	AMT_GOODS_PRICE	AMT_CREDIT	0.987
190	AMT_CREDIT	AMT_GOODS_PRICE	0.987
1196	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950
1134	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950
78	CNT_CHILDREN	CNT_FAM_MEMBERS	0.879
1055	CNT_FAM_MEMBERS	CNT_CHILDREN	0.879

From this analysis we can infer some of the features:

- There is a high correlation between AMT GOODS_PRICE and AMT_CREDIT
- Therefore More the price of the goods, higher the credit amount

- Correlation between CNT_FAM_MEMBERS and CNT_CHILDREN denotes more the count of family members there is a higher chance that the count of children's will be higher

- There is also a strong correlation between REGION_RATING_CLIENT and REGION_RATING_CLIENT_W_CITY ,
- Which denotes both of the metrics are more or less aligned to each other

Thank You