CREDIT EDA CASE STUDY

PROBLEM STATEMENT

When the company receives a loan application, the company must 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.

STEPS TO PERFORM EDA ANALYSIS

- Data Understanding
  - Data reading and data types: To read 'application_data.csv' and 'previous_application.csv' and understand its data types
  - BFS domain terms for better data understanding

- Data Cleaning and Manipulation
  - Handling missing values
  - Standardizing values
  - Handling "null" values

- Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms

- Find the top 10 correlation for the Client with payment difficulties

- Visualizations to explain the numerical/categorical variables

VIEW ON THE DATA SET

```
pre_app.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1670214 entries, 0 to 1670213
Data columns (total 37 columns):
 #   Column                       Non-Null
```

```
app_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Columns: 122 entries, SK_ID_CURR to AMT_REQ_CREDIT_BUREAU_YEAR
dtypes: float64(65), int64(41), object(16)
memory usage: 286.2+ MB
```

- 'application_data.csv' or 'app_data' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties
- 'previous_application.csv' or 'pre_app' 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

HANDLING MISSING VALUES

- Sorted 'application_data' column in terms of maximum count of 'null' values
- For the total 300k rows of data few columns have more than 50% of the rows empty
- Its best practice to remove them before getting started with our analysis

```
app_data.isnull().sum().sort_values(ascending=False)
```

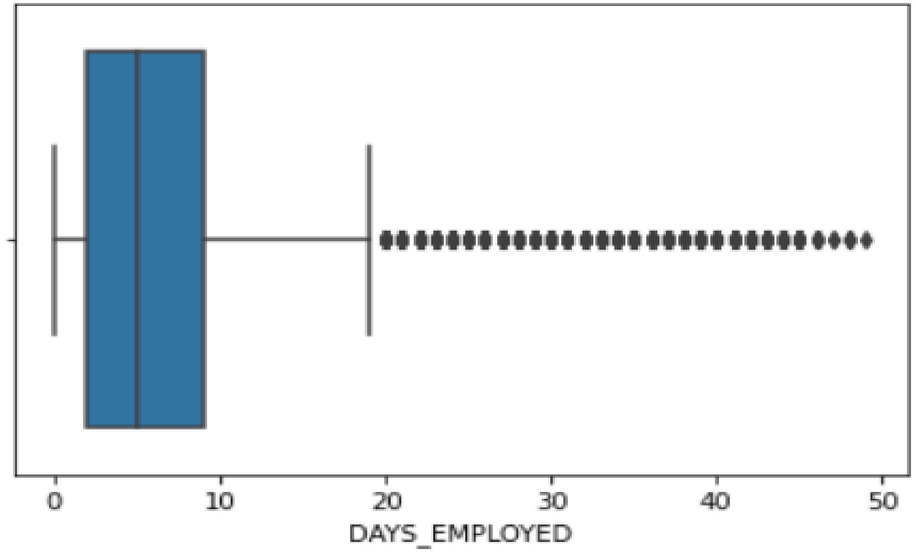
```
EXT_SOURCE_3                  60965
AMT_REQ_CREDIT_BUREAU_YEAR    41519
AMT_REQ_CREDIT_BUREAU_QRT     41519
AMT_REQ_CREDIT_BUREAU_MON     41519
AMT_REQ_CREDIT_BUREAU_WEEK    41519
AMT_REQ_CREDIT_BUREAU_DAY     41519
AMT_REQ_CREDIT_BUREAU_HOUR    41519
NAME_TYPE_SUITE                1292
DEF_60_CNT_SOCIAL_CIRCLE       1021
OBS_30_CNT_SOCIAL_CIRCLE       1021
DEF_30_CNT_SOCIAL_CIRCLE       1021
OBS_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
FLAG_DOCUMENT_18                  0
```

STANDARDIZING VALUES

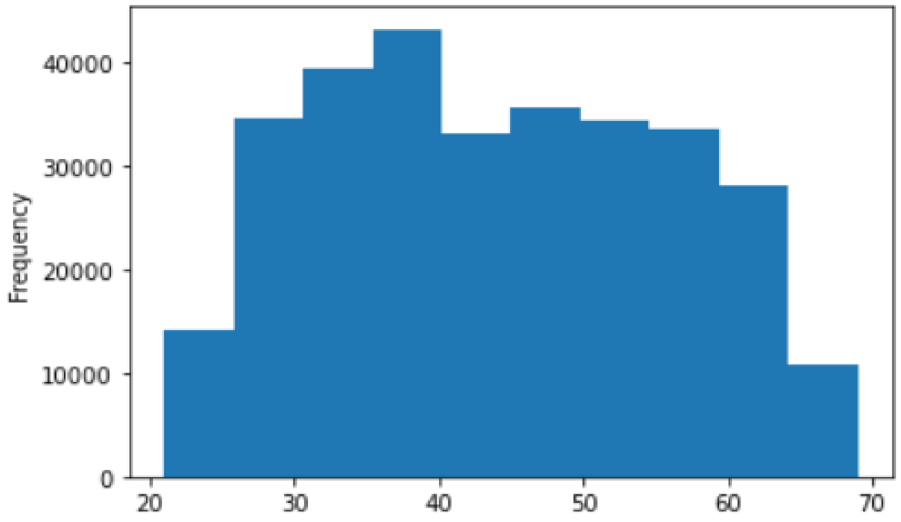
- Converting column values for DAYS_BIRTH and DAYS_EMPLOYED from days to years

- Plotting them into histogram and box plot for a quick analysis on its data distribution

```
sns.boxplot(app_data2.DAYS_EMPLOYED)
plt.show()
```



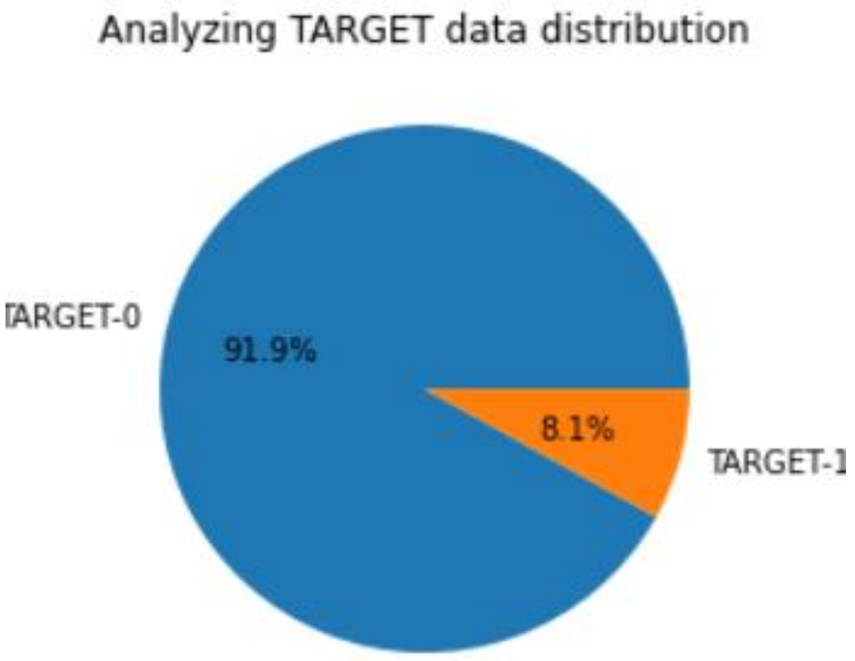

```
app_data2.DAYS_BIRTH.plot.hist()
plt.show()
```



ANALYZING TARGET DISTRIBUTION

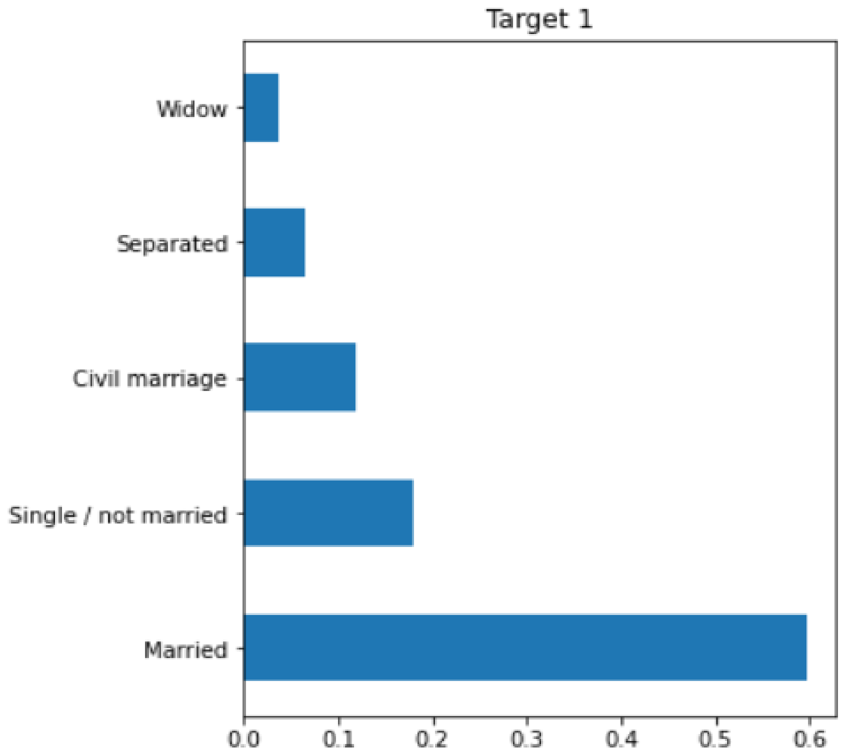
- As per the data plotted in this pie chart using ‘application_data’ dataframe, we can infer that we have in total of 8.1% defaulters

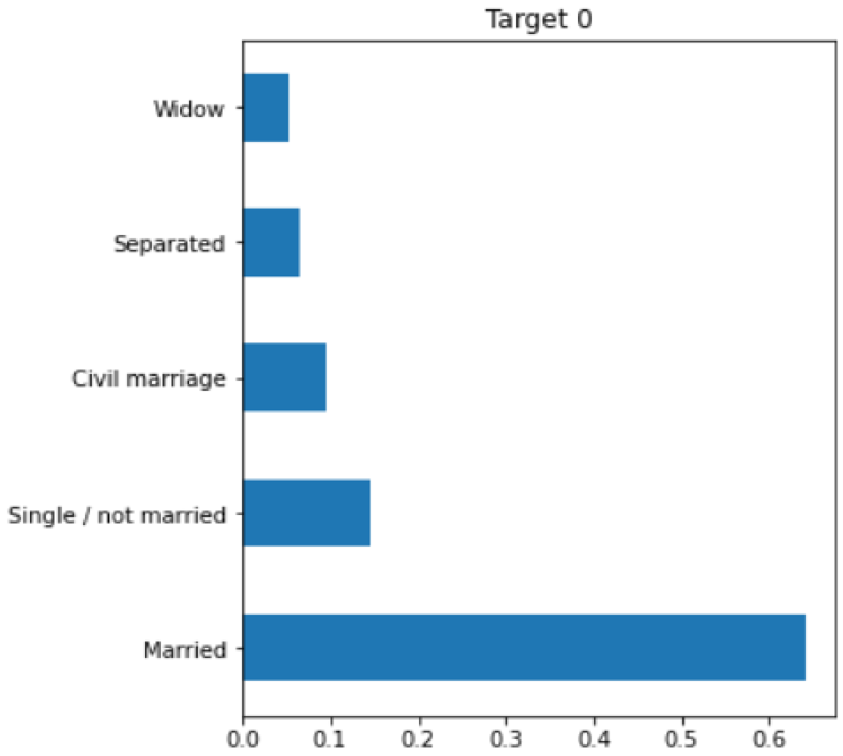
- Further provide insights on why the variables are important for differentiating the clients with payment difficulties with all other cases



UNIVARIATE ANALYSIS ON APP_DATA

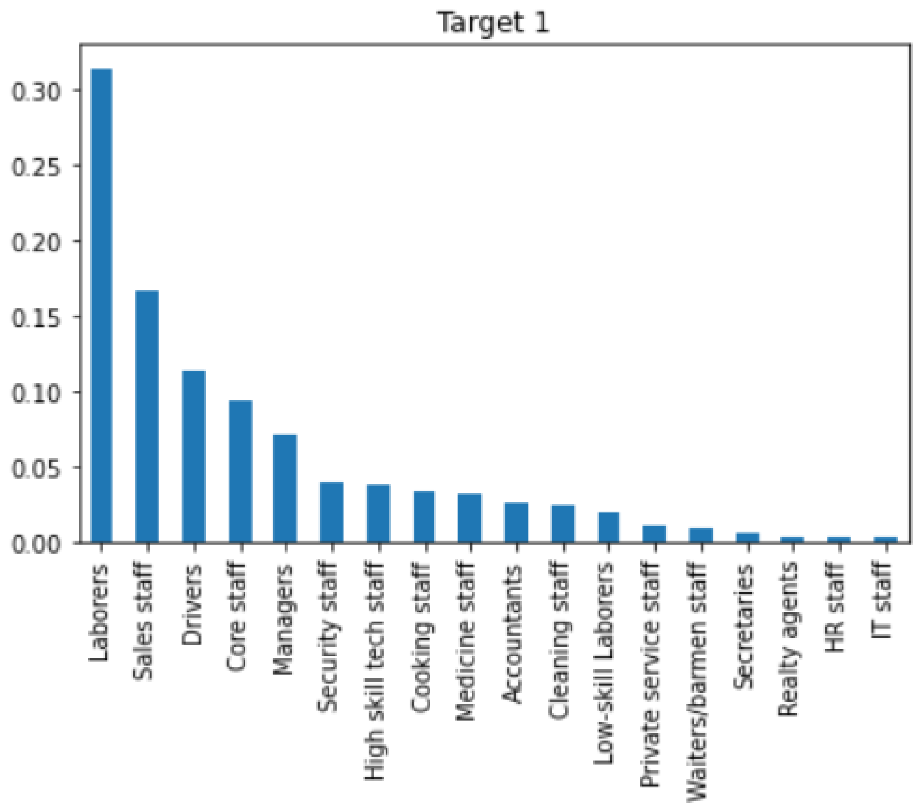
NAME_FAMILY_STATUS distribution across Target 1 and Target 0

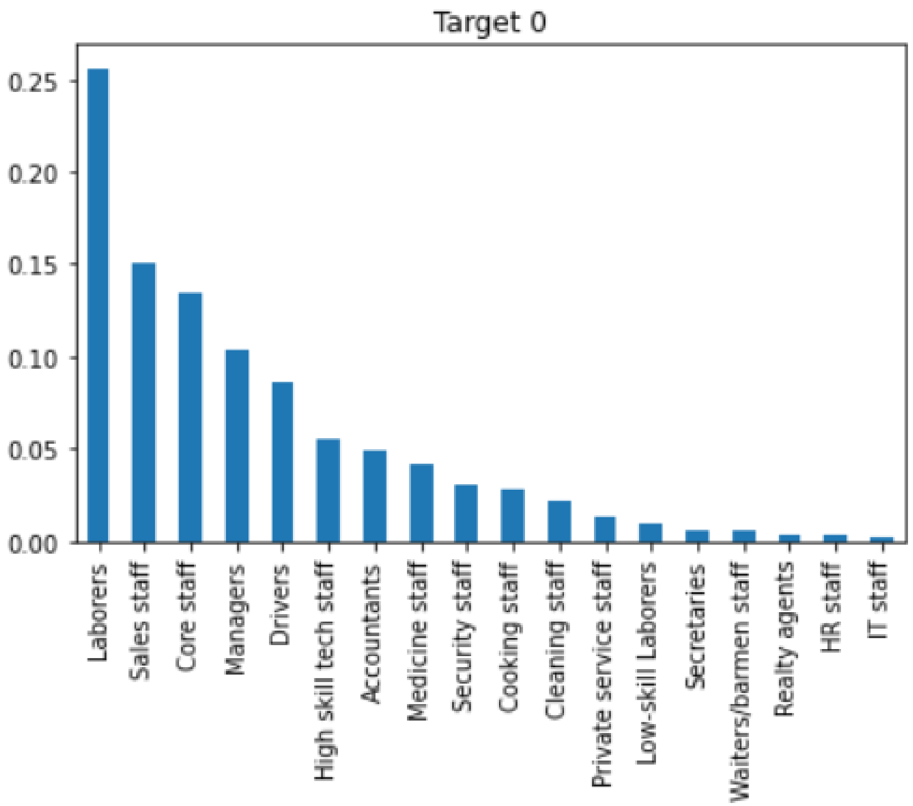




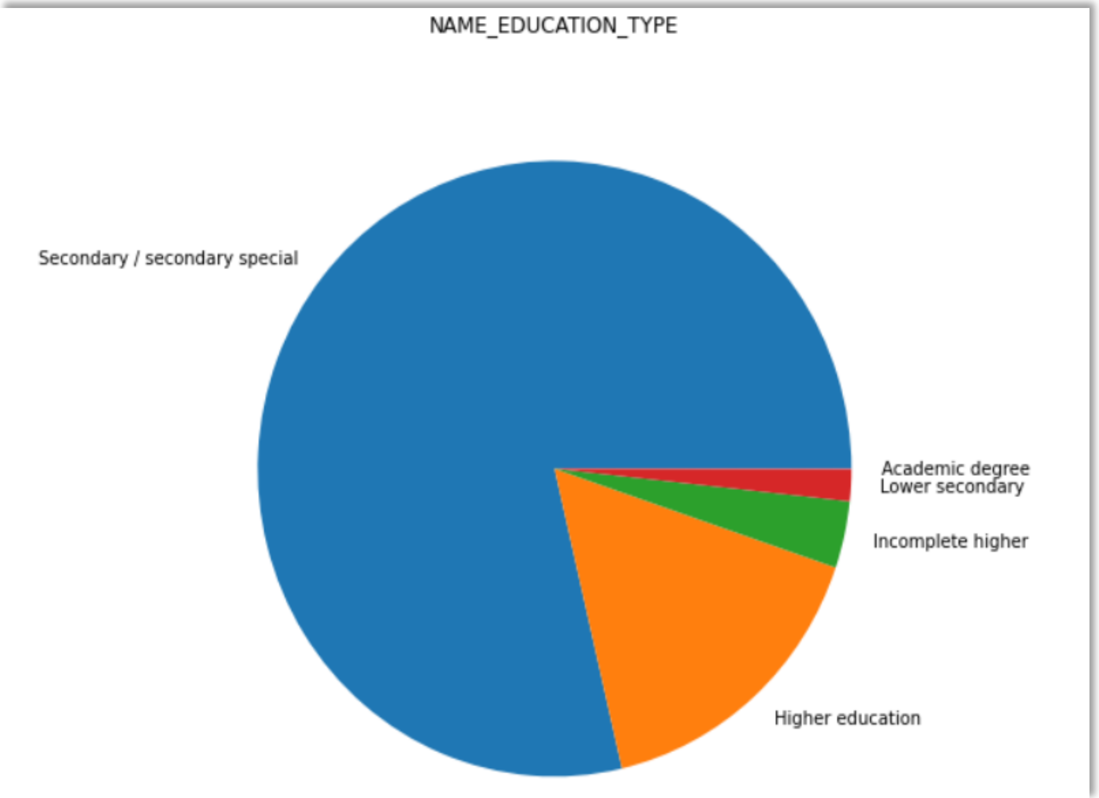
UNIVARIATE ANALYSIS ON APP_DATA

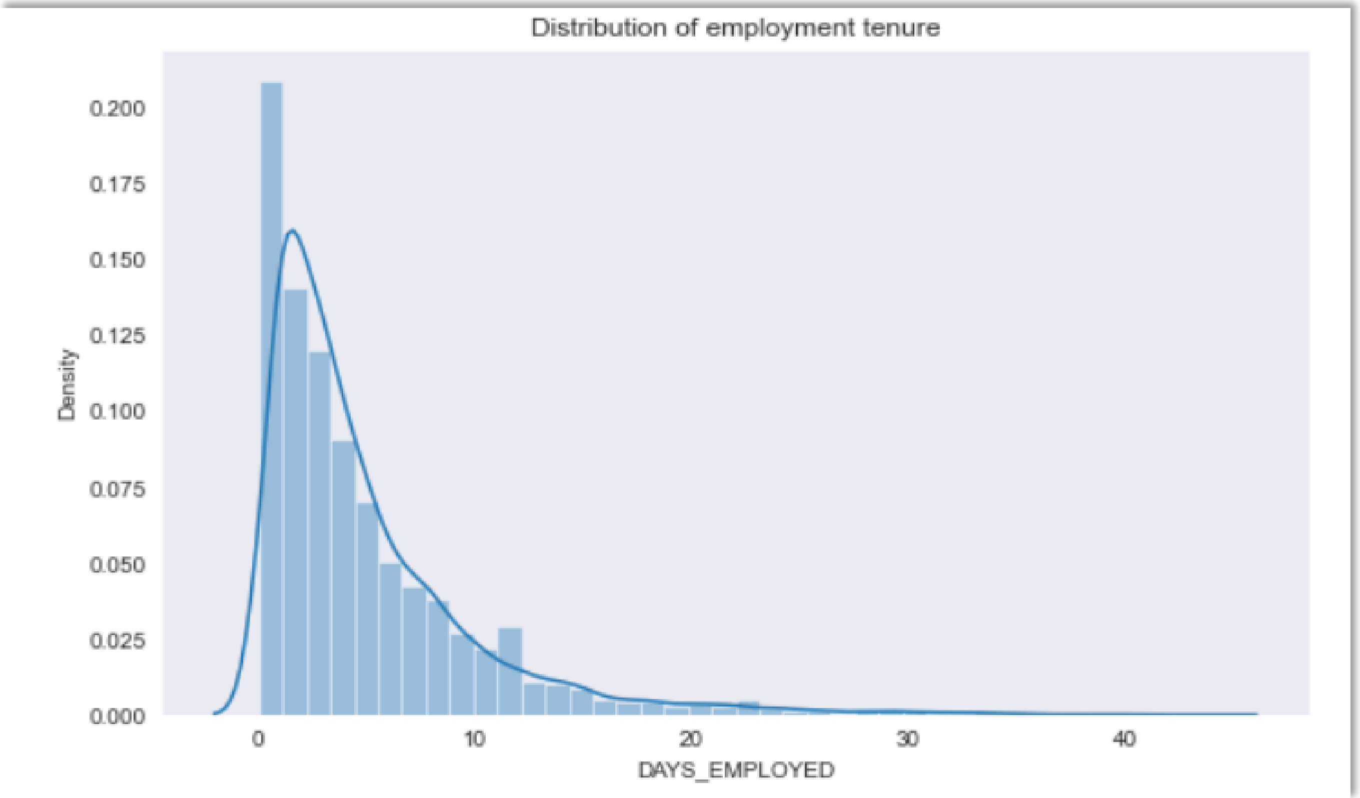
OCCUPATION_TYPE distribution across Target 1 and Target 0





UNIVARIATE ANALYSIS ON APP_DATA



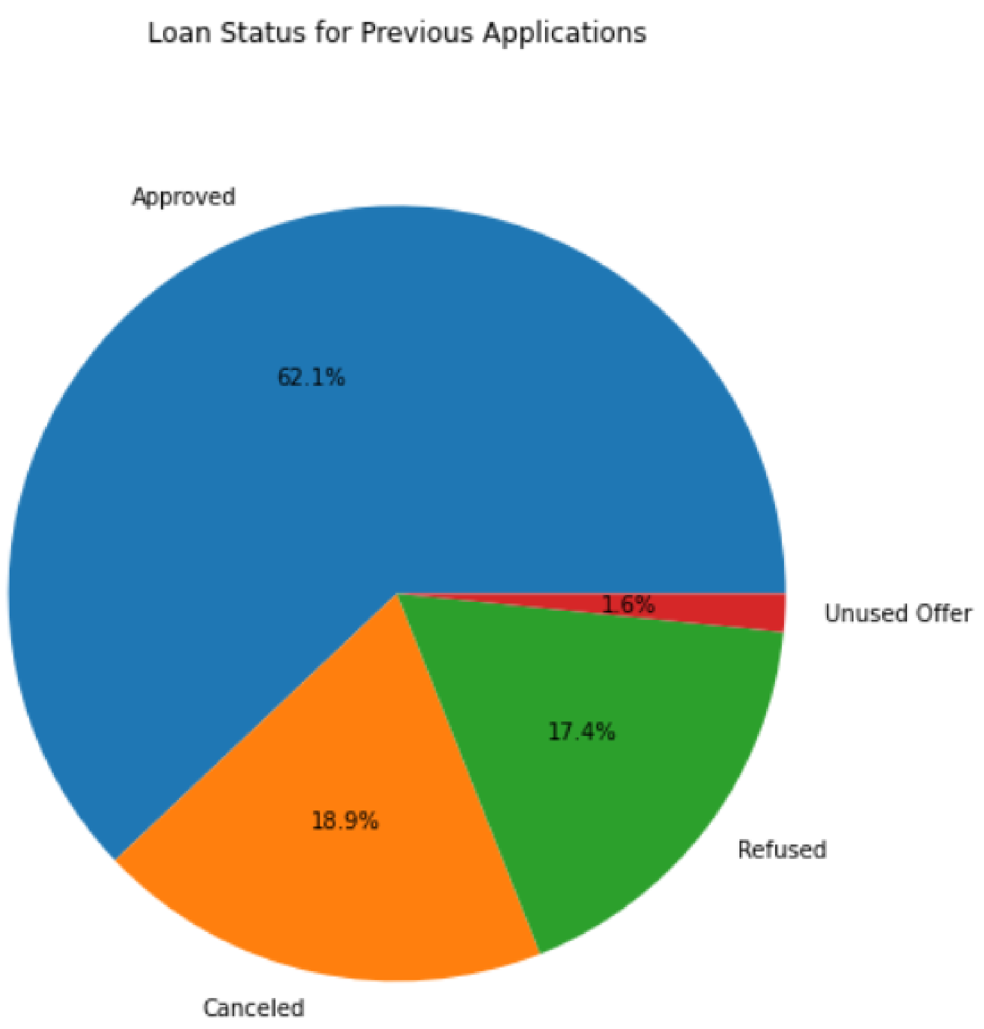


UNIVARIATE ANALYSIS ON PREV_DATA

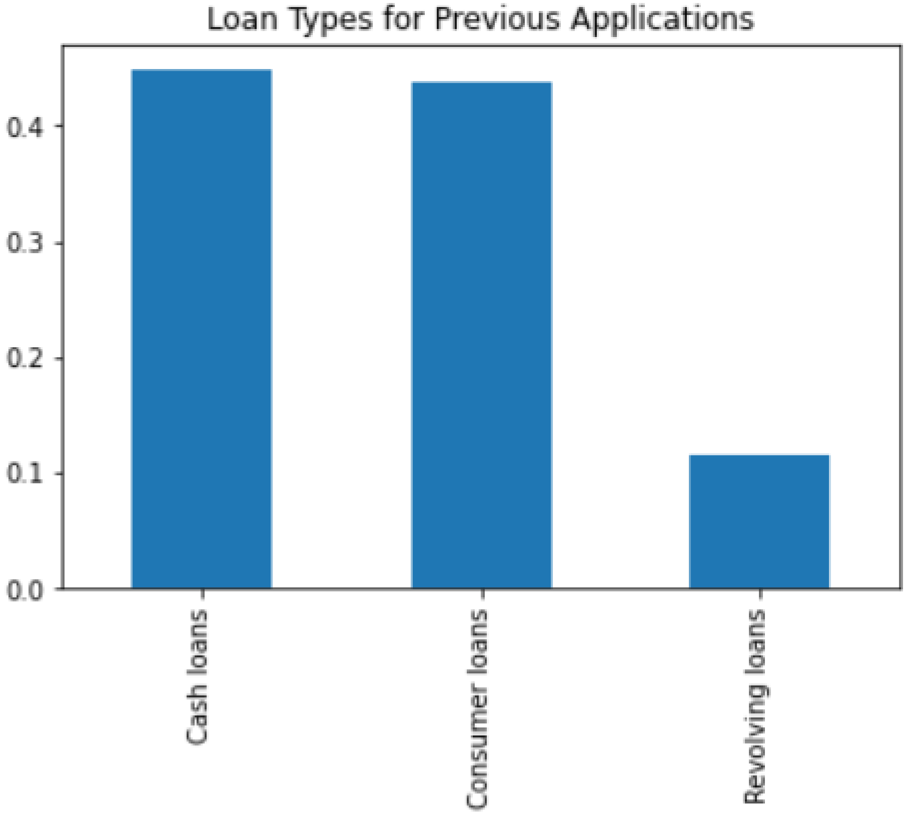
Loan Status for previous application

We can see that most of the applications have been approved

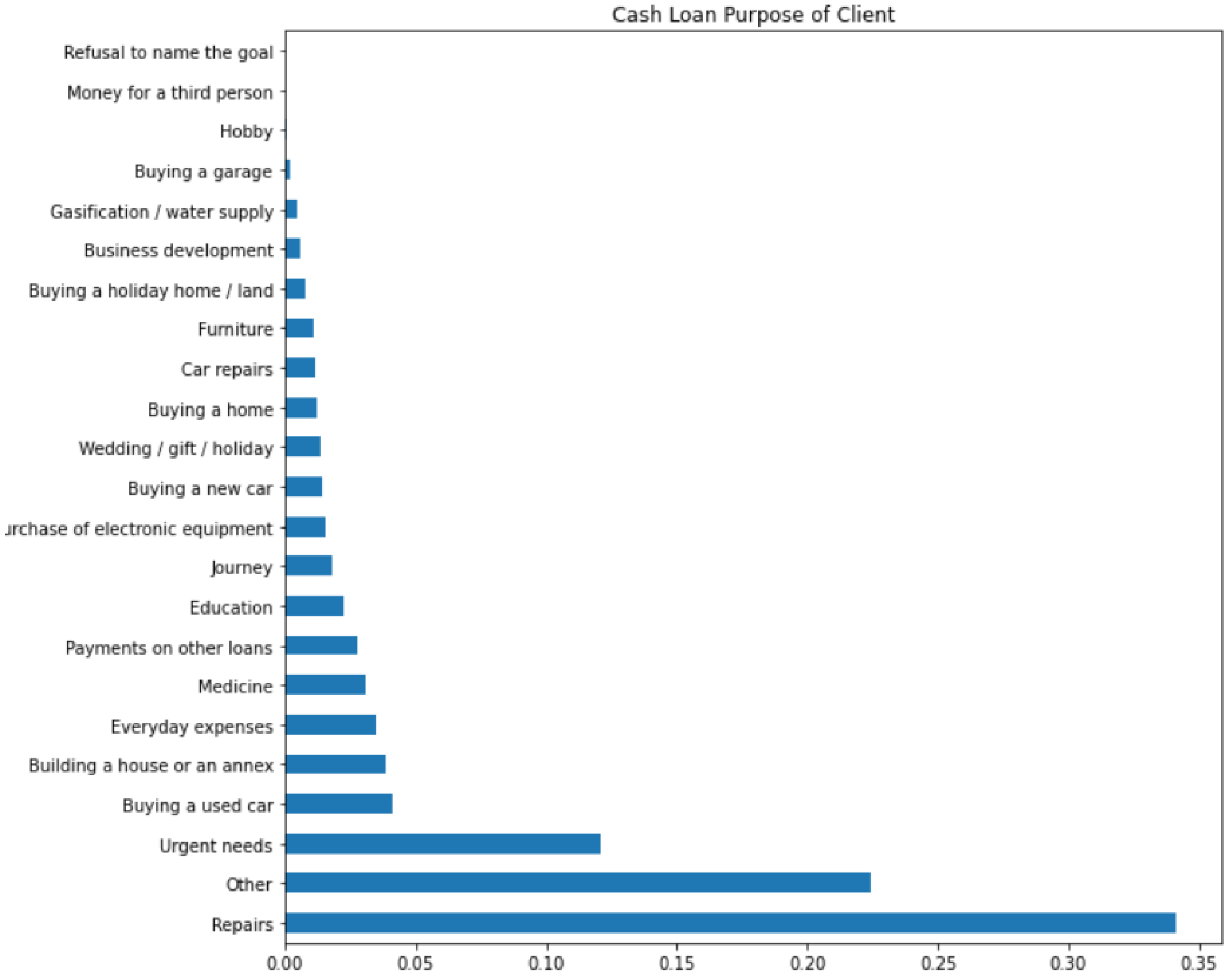
Next we can see they are either cancelled or refused



LOAN TYPES FOR PREVIOUS APPLICATION



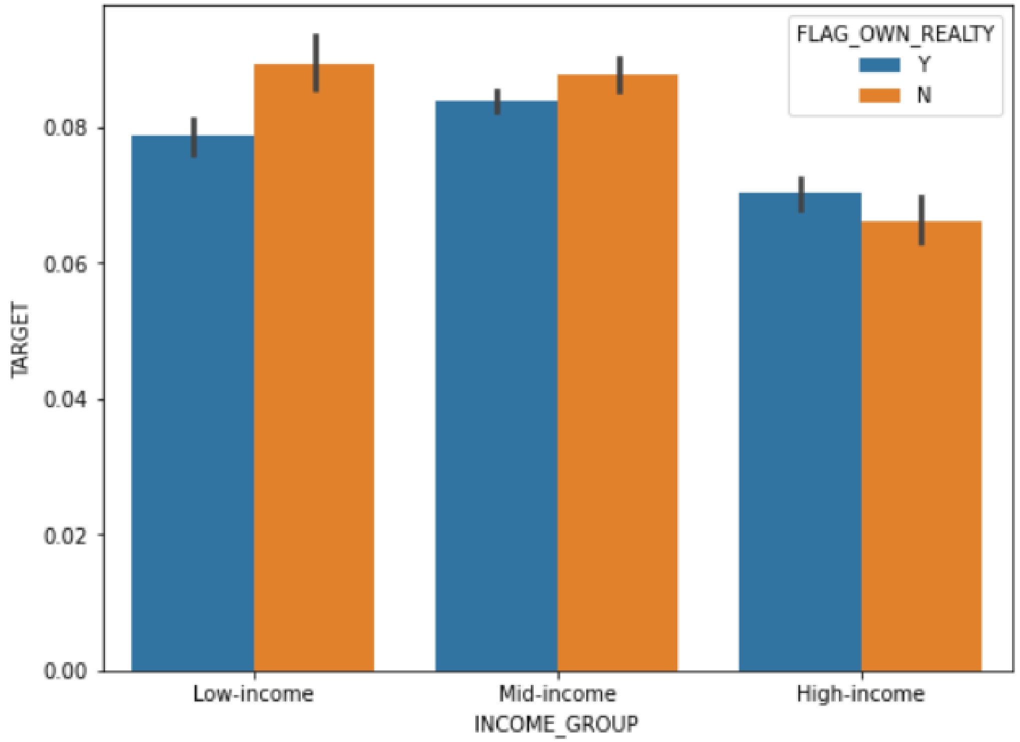
UNIVARIATE ANALYSIS ON CASH LOAN PURPOSE



As we can see from above bar chart most of the cash loans are taken for Repairs

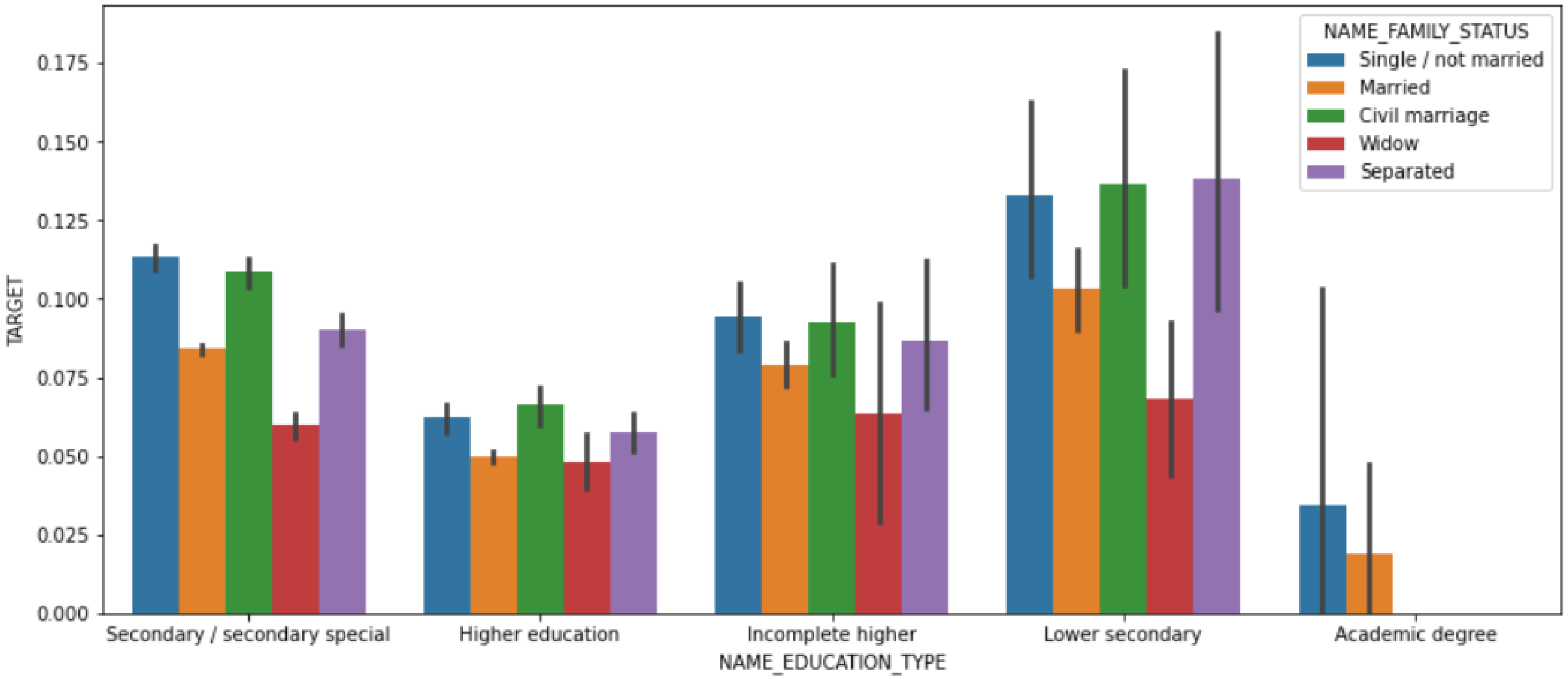
BIVARIATE ANALYSIS ON APP_DATA

- From above bar plot we can see that those from low- and middle-income groups who don't own a realty are more likely to default with their loan payments



BIVARIATE ANALYSIS ON APP_DATA

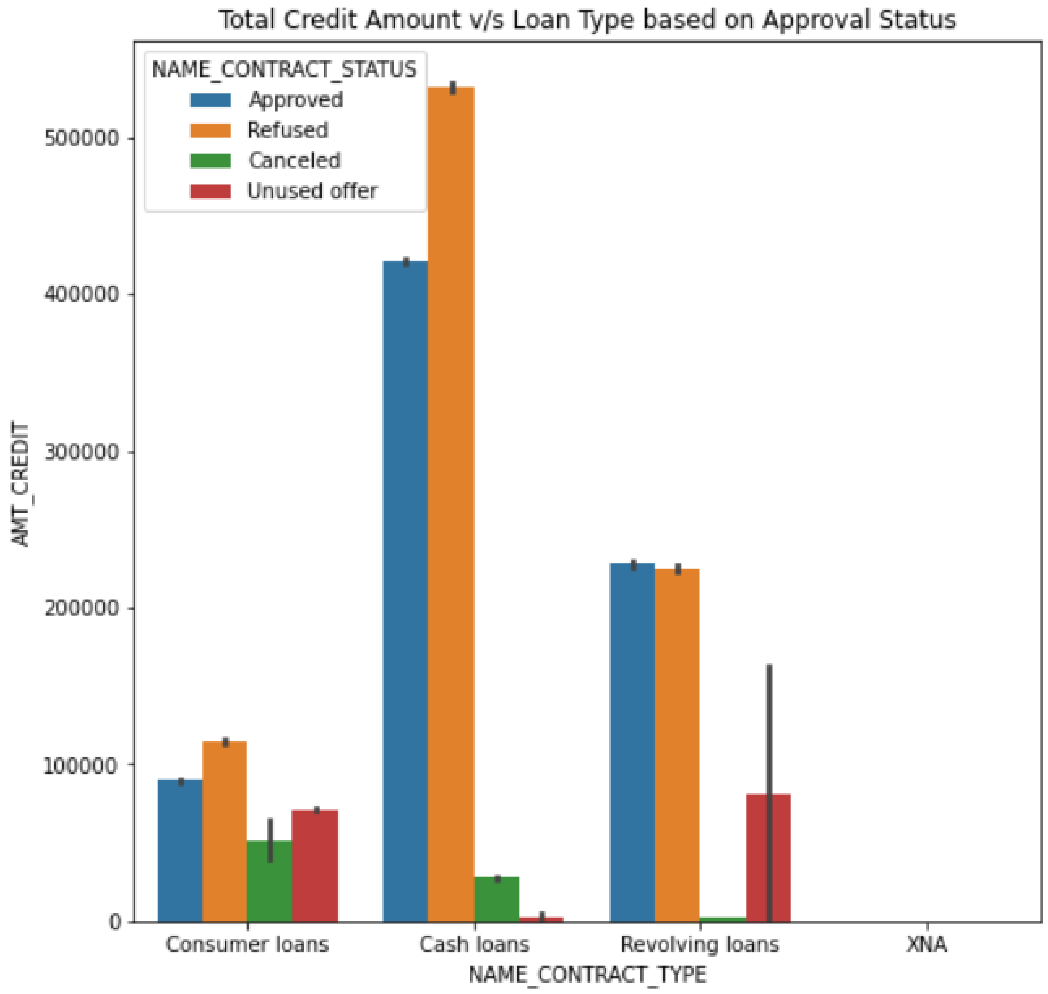
NAME_EDUCATION_TYPE and NAME_FAMILY_STATUS



BIVARIATE ANALYSIS ON PREV_DATA

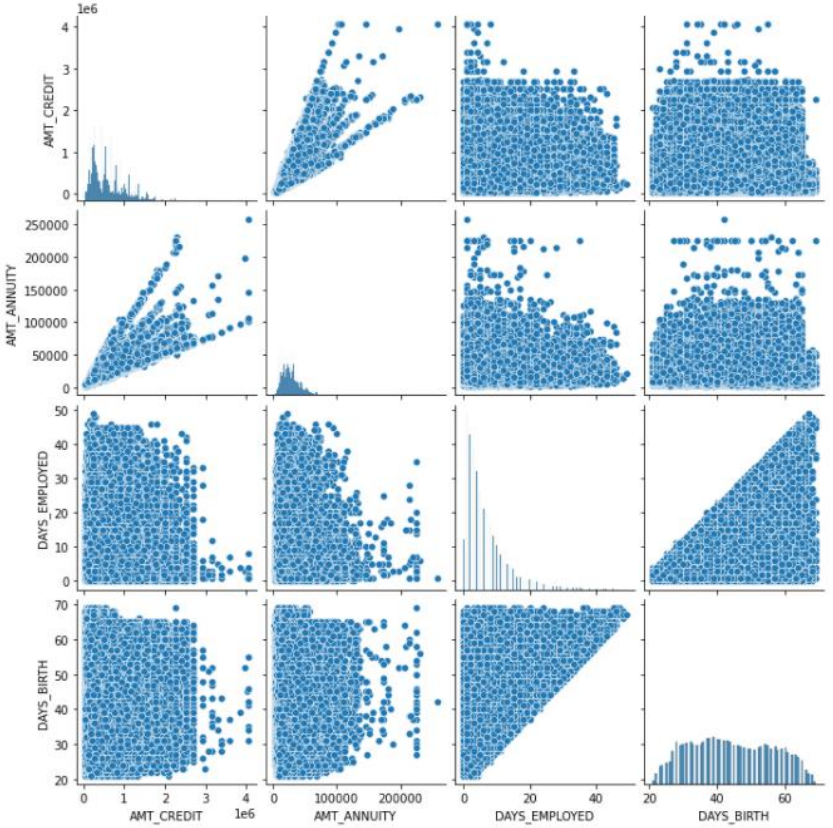
The rejection rate for cash loans above 500k seems higher

Anything up to 400k seems to be getting processed

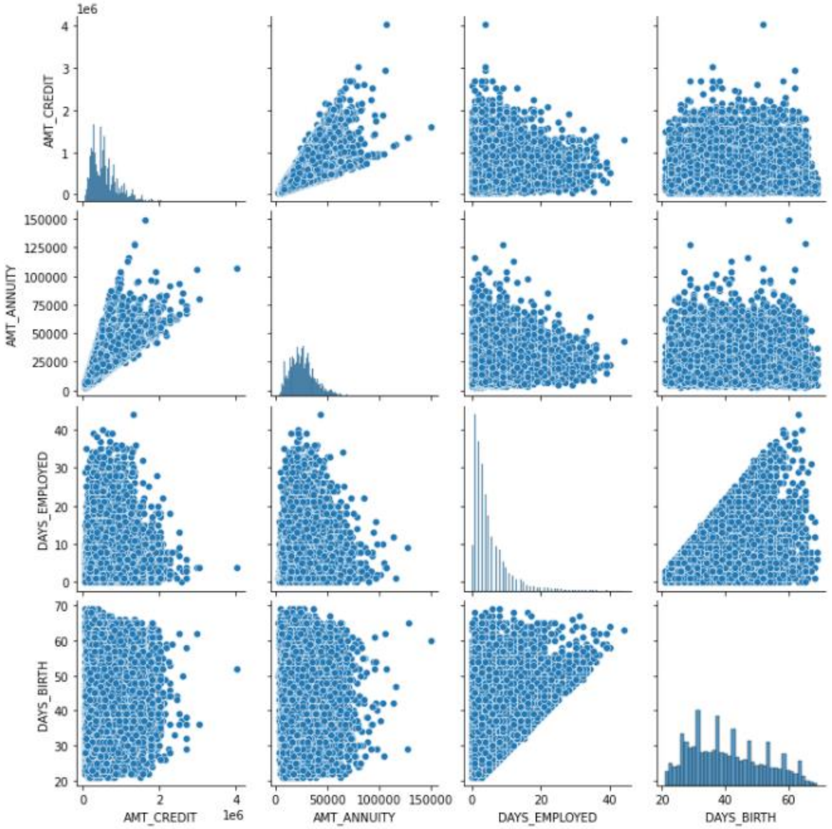


MULTIVARIATE ANALYSIS

AMT_CREDIT & AMT_ANNUITY are directly proportional
The loan applications are distributed across all age groups



AMT_CREDIT & AMT_ANNUITY are directly proportional
The AMT_ANNUITY is inversely proportion or having negative causation with DAYS_EMPLOYED, means customers who are employed for a long time are more likely to repay on-time



TOP 10 CORRELATION FOR THE CLIENT WITH PAYMENT DIFFICULTIES

MT_GOODS_PRICE & AMT_CREDIT show good correlation of up to 0.987251. Thus, for consumer requesting for high loan credit have high goods price

Customers with high CNT_FAM_MEMBERS are likely to have more CNT_CHILDREN, thus showing that they might falls under defaulters because of family commitments

	level_0	level_1	0
84	AMT_INCOME_TOTAL	SK_ID_CURR	0.001739
126	AMT_CREDIT	SK_ID_CURR	0.000364
128	AMT_CREDIT	AMT_INCOME_TOTAL	0.342796
168	AMT_ANNUITY	SK_ID_CURR	0.000070
170	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418954
...	...	...	...
1758	FLAG_DOCUMENT_21	FLAG_DOCUMENT_16	0.000170
1759	FLAG_DOCUMENT_21	FLAG_DOCUMENT_17	0.000299
1760	FLAG_DOCUMENT_21	FLAG_DOCUMENT_18	0.000565
1761	FLAG_DOCUMENT_21	FLAG_DOCUMENT_19	0.000437
1762	FLAG_DOCUMENT_21	FLAG_DOCUMENT_20	0.000399

level_0	level_1	Correlation_TARGET0
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508
AMT_GOODS_PRICE	AMT_CREDIT	0.987251
CNT_FAM_MEMBERS	CNT_CHILDREN	0.878575
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861812
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859331
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830366
AMT_GOODS_PRICE	AMT_ANNUITY	0.776686
AMT_ANNUITY	AMT_CREDIT	0.771309
FLAG_DOCUMENT_6	FLAG_DOCUMENT_3	0.486443
FLAG_DOCUMENT_8	FLAG_DOCUMENT_3	0.461069

MERGE TWO DATAFRAMES

- Merging App_Data and Prev_Data to understand the data for customers who are re-applying for loan

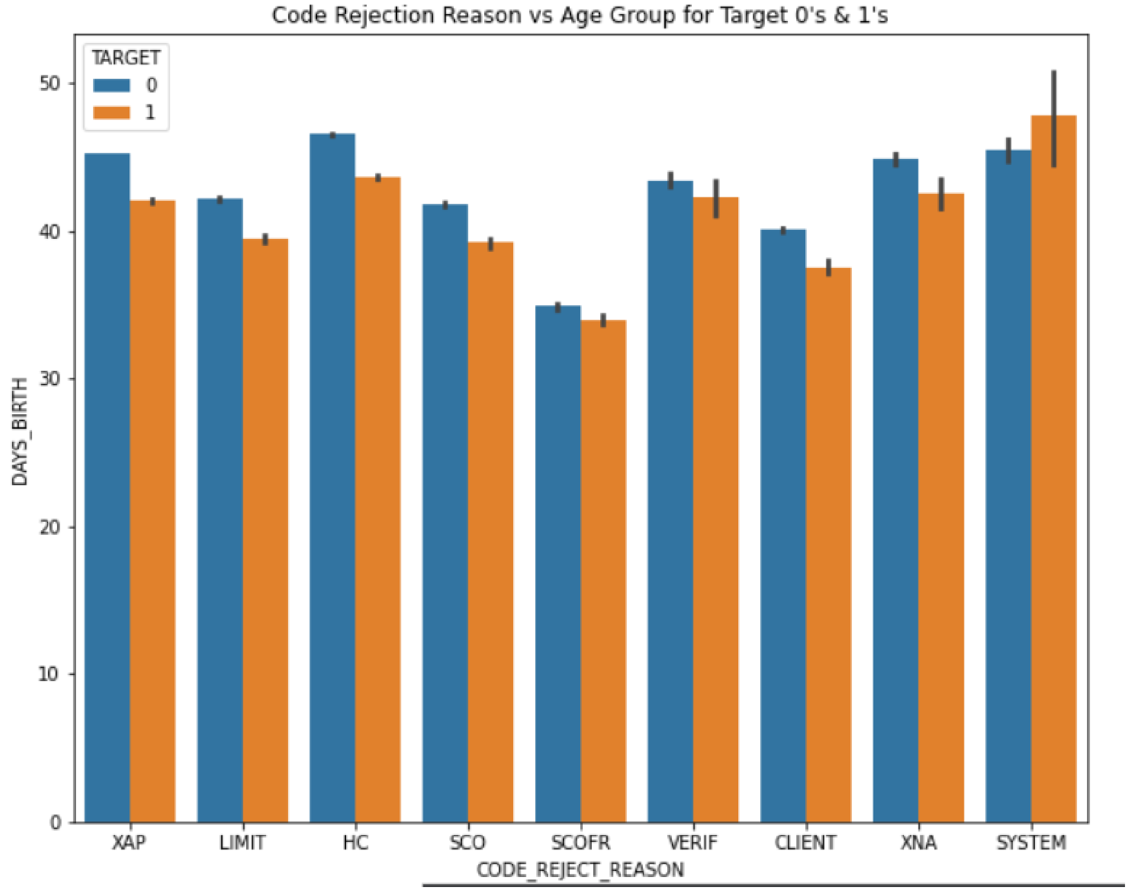
- Infer some analysis based on which their previous loan application was approved or rejected and what might be different in the current application

```
df_merged.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1413608 entries, 0 to 1413607
Data columns (total 18 columns):
 #   Column               Non-Null Count    Dtype
---  ------               --------------    -----
 0   SK_ID_CURR           1413608 non-null  int64
 1   TARGET               1413608 non-null  int64
 2   CODE_GENDER          1413608 non-null  object
 3   AMT_INCOME_TOTAL     1413608 non-null  float64
 4   AMT_CREDIT_x         1413608 non-null  float64
 5   AMT_ANNUITY_x        1413608 non-null  float64
 6   NAME_TYPE_SUITE      1410082 non-null  object
 7   NAME_INCOME_TYPE     1413608 non-null  object
 8   NAME_EDUCATION_TYPE  1413608 non-null  object
 9   NAME_FAMILY_STATUS   1413608 non-null  object
 10  DAYS_BIRTH           1413608 non-null  float64
 11  DAYS_EMPLOYED        1140025 non-null  float64
 12  OCCUPATION_TYPE      1413608 non-null  object
 13  SK_ID_PREV           1413608 non-null  int64
 14  AMT_ANNUITY_y        1106420 non-null  float64
 15  AMT_APPLICATION      1413608 non-null  float64
 16  AMT_CREDIT_y         1413607 non-null  float64
 17  CODE_REJECT_REASON   1413608 non-null  object
dtypes: float64(8), int64(3), object(7)
memory usage: 204.9+ MB
```

MERGED DATAFRAME ANALYSIS

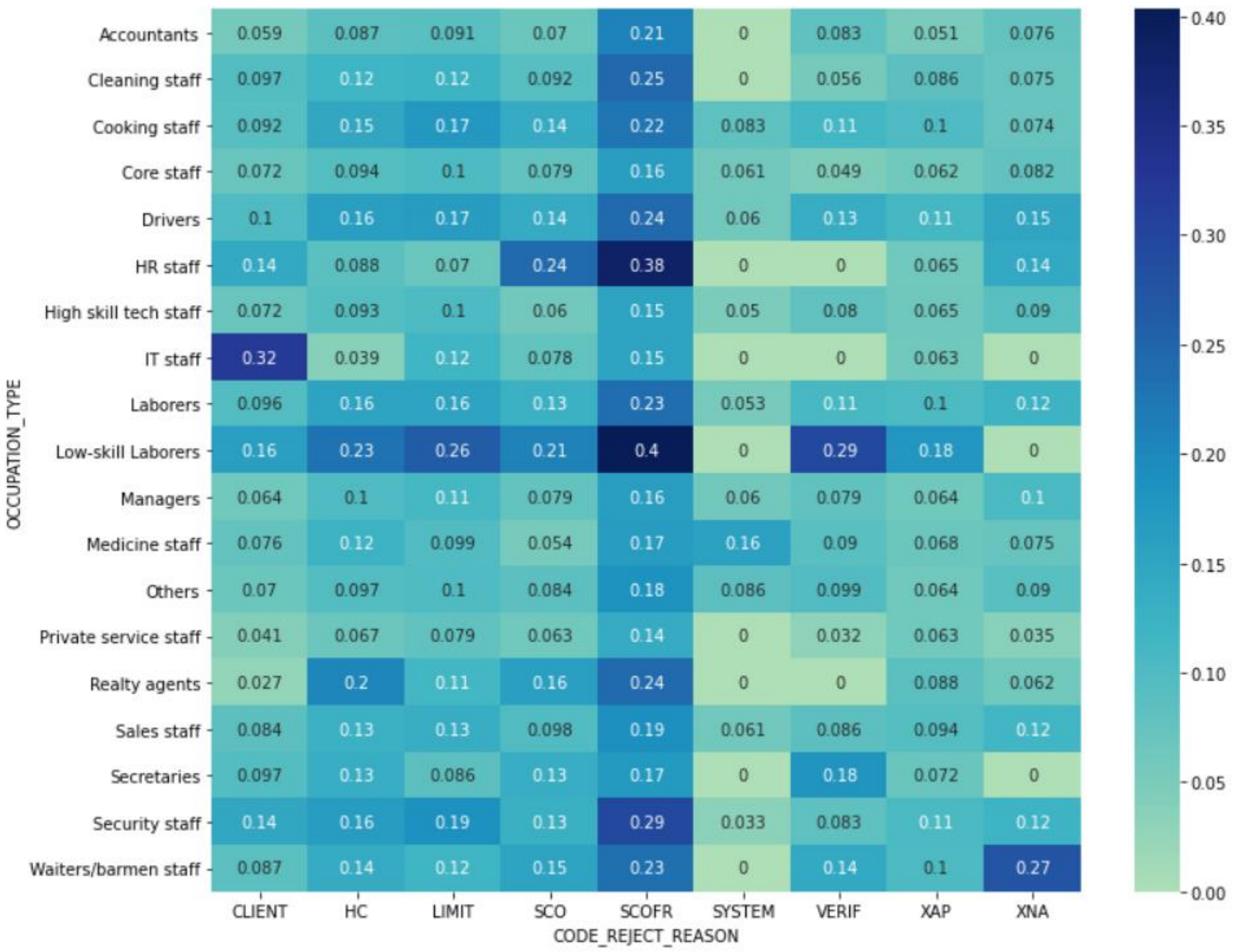
- With this we can infer that "SYSTEM" CODE_REJECT_REASON has the highest count across applicants across age band upto close to 50 years



MERGED DATAFRAME ANALYSIS

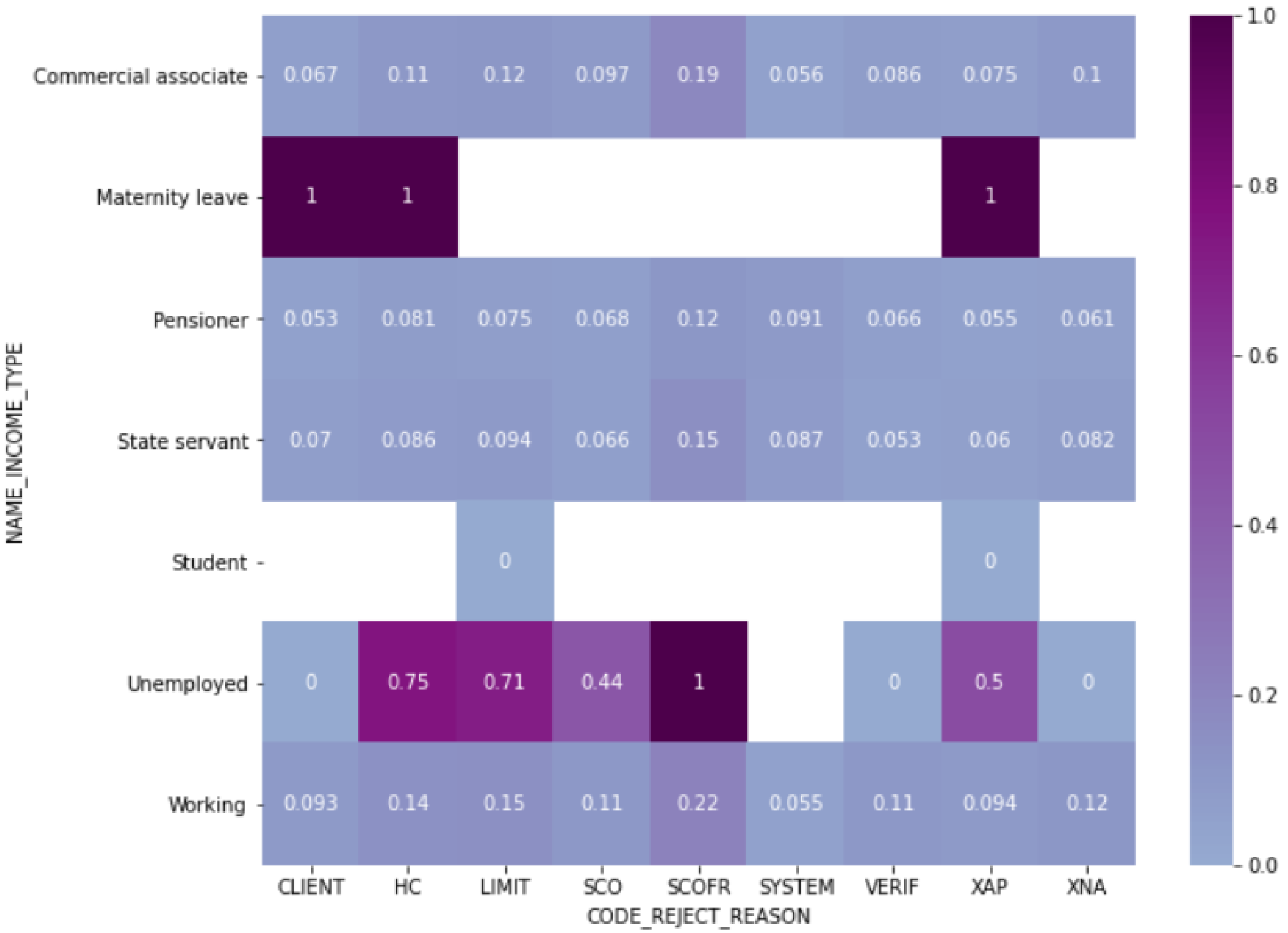
- In this we can infer that "SCOFR" CODE_REJECT_REASON” is the most common rejection reason across all OCCUPATION_TYPE customers

- Low-skill Laborers have the maximum rejections across all rejection's types, and we can infer that as they have daily and inconsistent wages, they are more likely to get their loan applications rejected



MERGED DATAFRAME ANALYSIS

- In this we can infer that Unemployed customers are more likely to default as they dont have an income to repay the loan on time and are risky
- Also, women on Maternity Leave are more likely to become defaulters as they might be on unpaid leave



CONCLUSION

- Defaulter's rate is 8.7%
- Cash loan types tend to have higher rejection rate when compared to Revolving loans
- Low-skill Laborers have the maximum rejections
- Women on Maternity Leave are more likely to become defaulters
- "SCOFR" CODE_REJECT_REASON" is the most common rejection reason across all OCCUPATION_TYPE customers
- Laboure occupation type by far has the highest rate for facing the difficulty with the loan repayments
- MT_GOODS_PRICE & AMT_CREDIT show good correlation of up to 0.987251
- Customers with high CNT_FAM_MEMBERS are likely to have more CNT_CHILDREN
- Customers who are employed for a long time are more likely to repay on-time
- Low and middle-income groups who don't own a house/reality, are more likely to default on their payments