EDA CREDIT CASE STUDY

Presented By: Malvika Chauhan

Rishav Kumar

Objective

- This case study aims to identify patterns which indicate if a client has difficulty paying their installments 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. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

Data Cleaning and Preparation

Below are the steps which are performed :-

- Data Inspection: It is the act of viewing data for verification and debugging purposes, before, during, or after a translation. With reference to dataset-app.info(), app.describe(), app.shape are some method in inspect the data.

- Identify percentage of null values in columns: We found there are 67 columns in the dataset containing null values in application dataset and 16 in previous application dataset.

- Drop columns with high null values: We followed the strategy of dropping columns having null percentage greater than 50 . Application dataset was having 41 and previous dataset was having 4 columns having null percentage greater than 50 . So we dropped them including some other columns which are not relevant for our analysis.

- Impute other columns with fewer null values:
- **AMT_ANNUITY-** There is a high range of outliers at left side of the distribution and hence imputing with mean would not be the right approach so imputing with median.
- **AMT_GOODS_PRICE-** Since both Median and mode are same and might lead to incorrect analysis . Hence imputing it with mean.
- **OCCUPATION_TYPE-** For OCCUPATION_TYPE filling null values with not Specified as might be customer did not wanted to specify there occupation to all or it might get miss to add in the form due to technical problem while filling the form, so we can't normally predict and add any random occupation to the customer to fill null.
- **NAME_TYPE_SUITE -** Fill null values with mode, we might predict that customer did not accompanied with anyone and came alone.
- **CNT_FAM_MEMBERS-** Replacing null with median as this value can not be fraction.

Outliers:

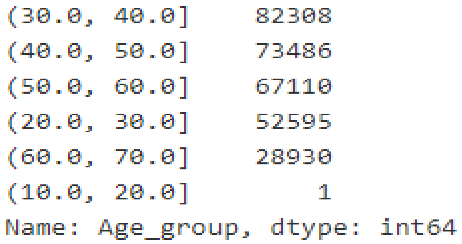
- Handling outlier in **AMT_ANNUITY**, **AMT_GOODS_PRICE** by excluding values outside 99%ile.
- Since **AMT_INCOME_TOTAL** outlier is 337500 and 99th percentile is 450000 and the number of rows with these values are 12866 and 2750 respectively. Replacing the values higher than 450000.0 with 450000.0+10000.0 to make it continuous and distinctly identifiable and not 337500.0 since 99th percentile value is higher.

Data quality issues such as negative value in age column:

- **CNT_FAM_MEMBERS** cannot be float. Converting to integer.
- **DAYS_EMPLOYED**, **DAY_REGISTRATION**, **DAYS_ID_PUBLISH** should be a positive value. Converting into absolute value.
- **DAYS_BIRTH** column is age of the person at the time of loan application. This could be converted to age in years by dividing by 365.25. Also it is with a negative sign, hence needs to be treated.

Binning continuous variables:

- Min age is 20 and max age is 68- Dividing into 7 intervals and creating a new column **Age_group** having range as



```
(30.0, 40.0]    82308
(40.0, 50.0]    73486
(50.0, 60.0]    67110
(20.0, 30.0]    52595
(60.0, 70.0]    28930
(10.0, 20.0]        1
Name: Age_group, dtype: int64
```

- Binning **AMT_INCOME_TOTAL** into income range categories for ease of analysis as 'Low','Average','High','Very High'

Data Analysis

- **Divide data based on target 0 and 1:**

We have a **Imbalance Ratio**: 11.32

Creating two data frames for Target = 0 and Target = 1 as below-

app_T0 = app[app.TARGET == 0]

app_T1 = app[app.TARGET == 1]

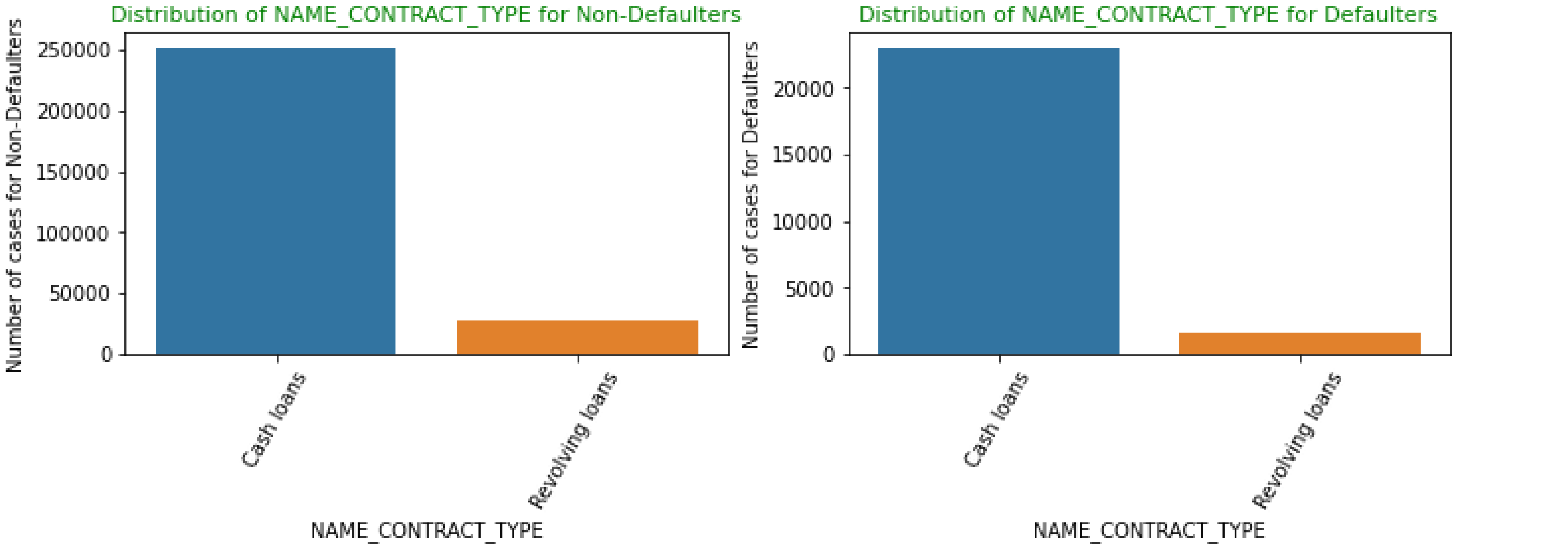
On dividing dataset we analysis that data only consist of 8.11% cases where payment was not made on time.

Also Imbalance Ratio is 11.32.

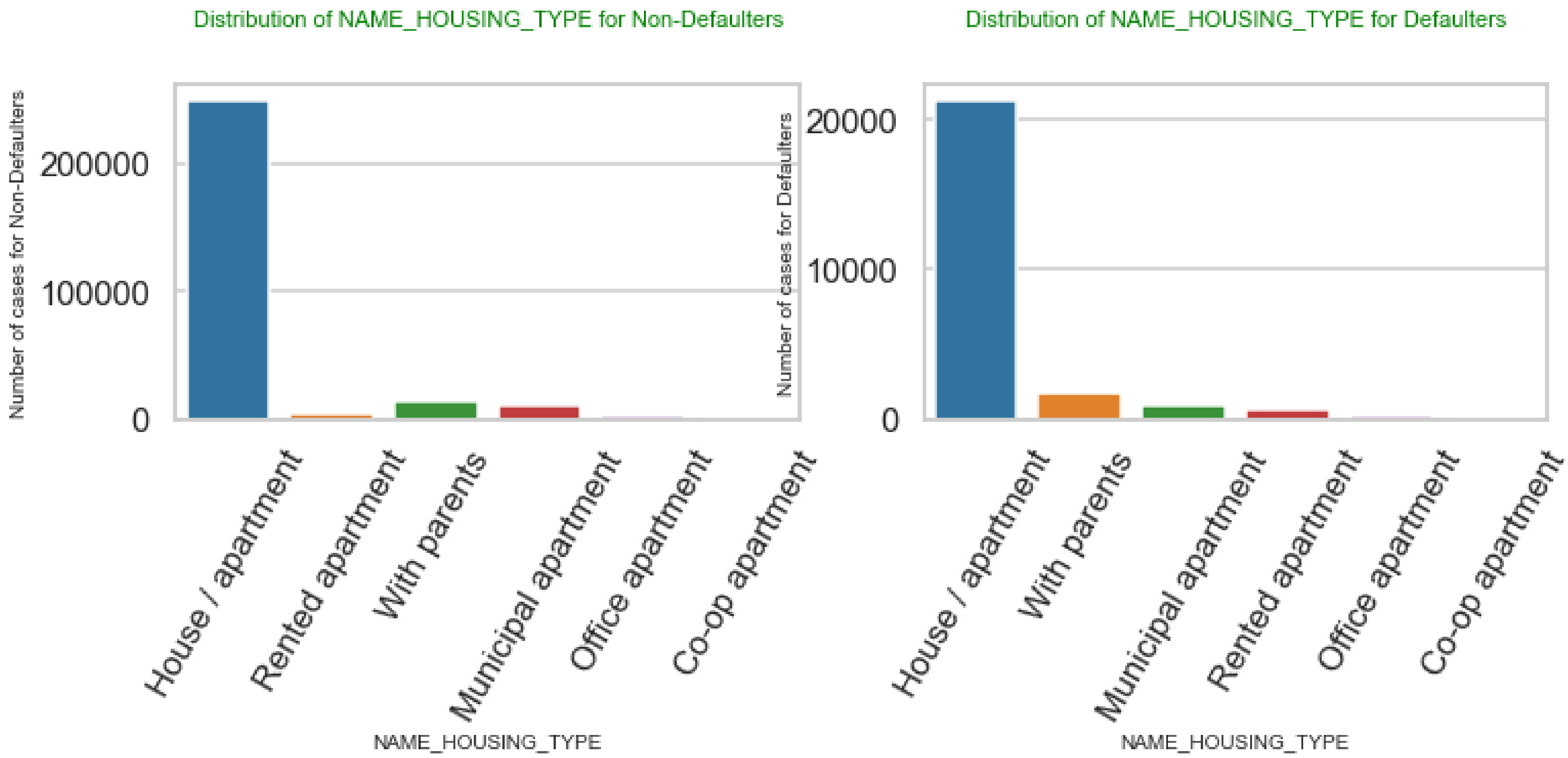
Univariate analysis

- Univariate analysis explores each variable in a data set, separately. It looks at the range of values, as well as the central tendency of the values. It describes the pattern of response to the variable.

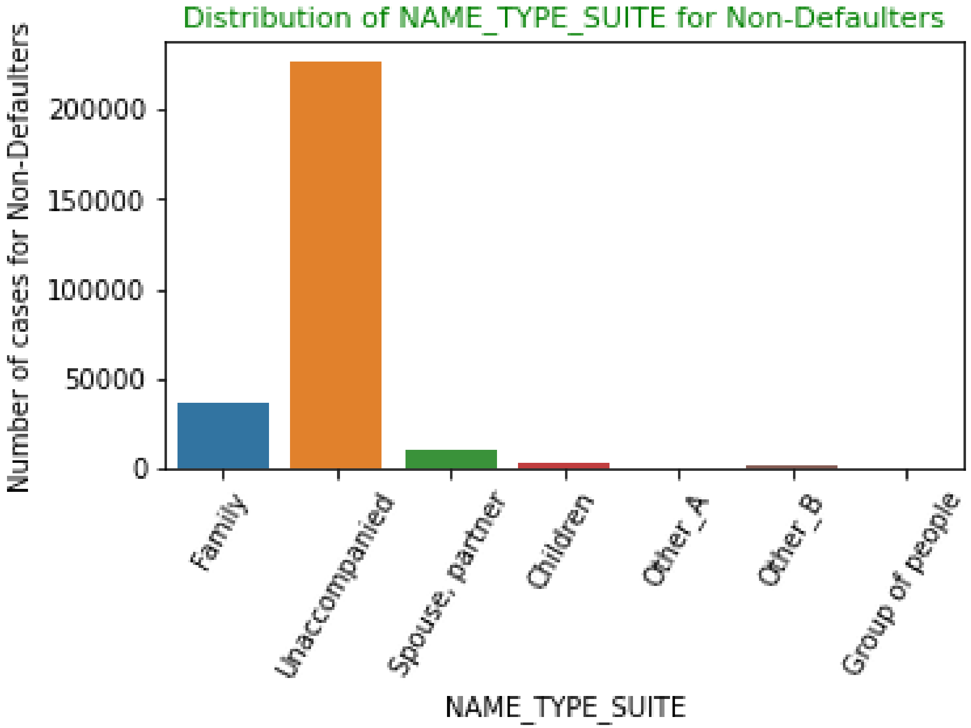
From below we can notice that revolving loans are lesser in the defaulted population. Hence we can infer that revolving loans are comparatively safer. There are around 250000 people has applied for cash loan who are not defaulters while more than 20000 people in defaulters list have applied for loan.

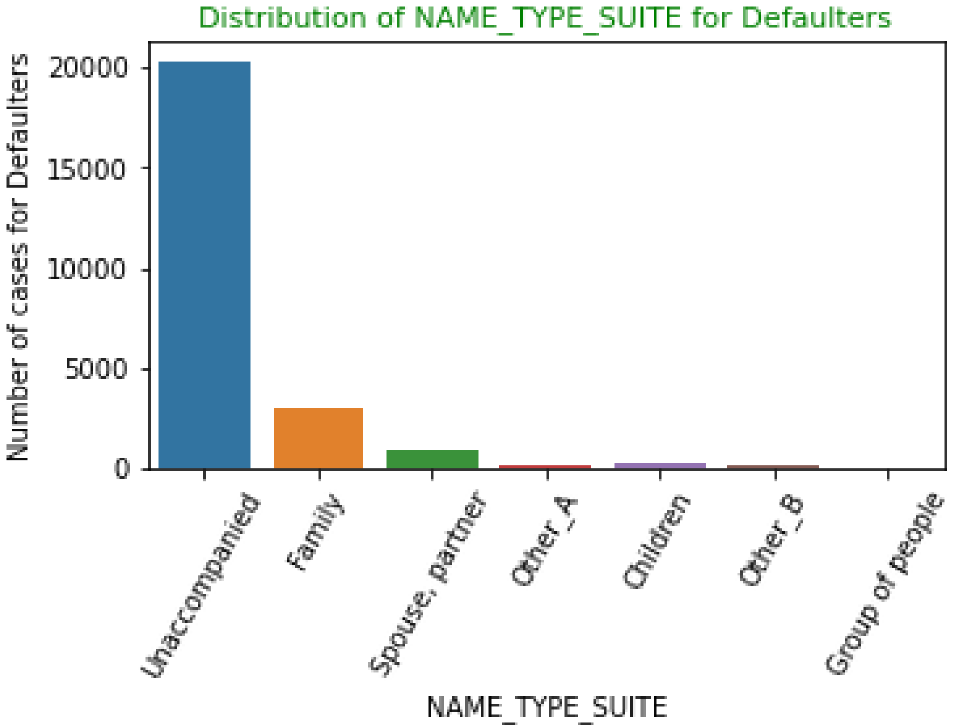


- Population having there own apartments/house have high rate of defaults.

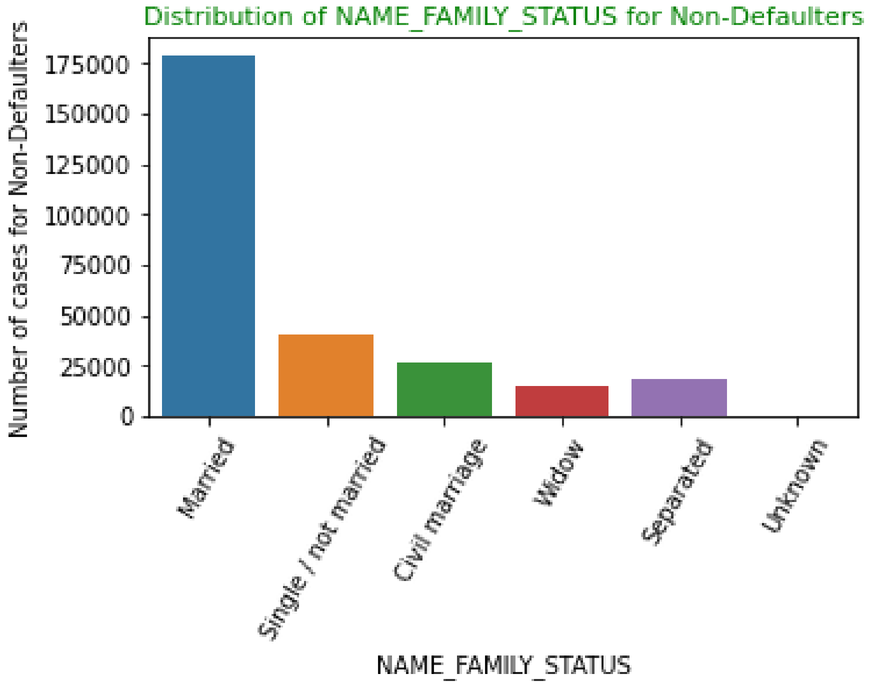


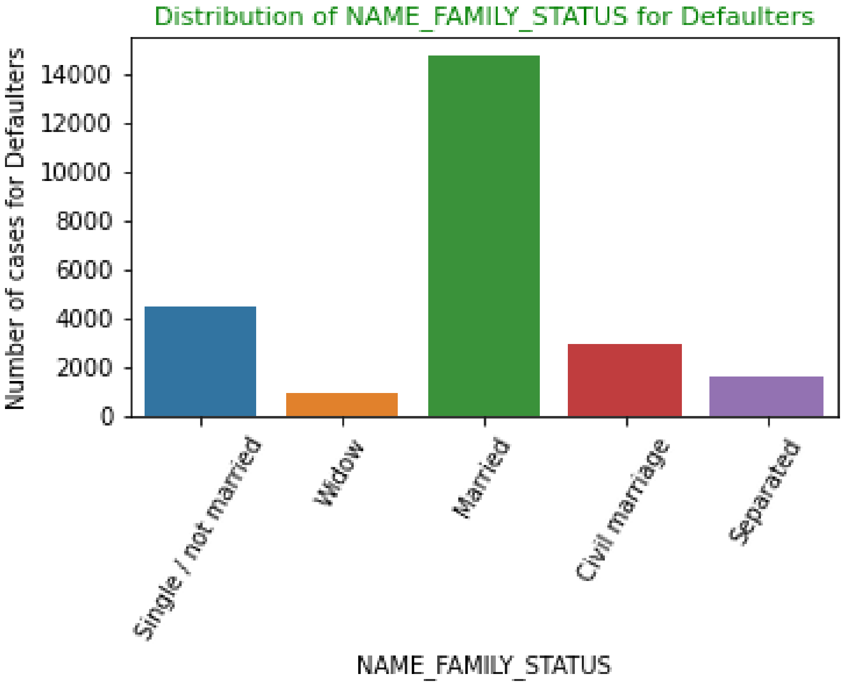
- People applying for loan are mostly “Unaccompanied”.


Distribution of NAME_TYPE_SUITE for Non-Defaulters

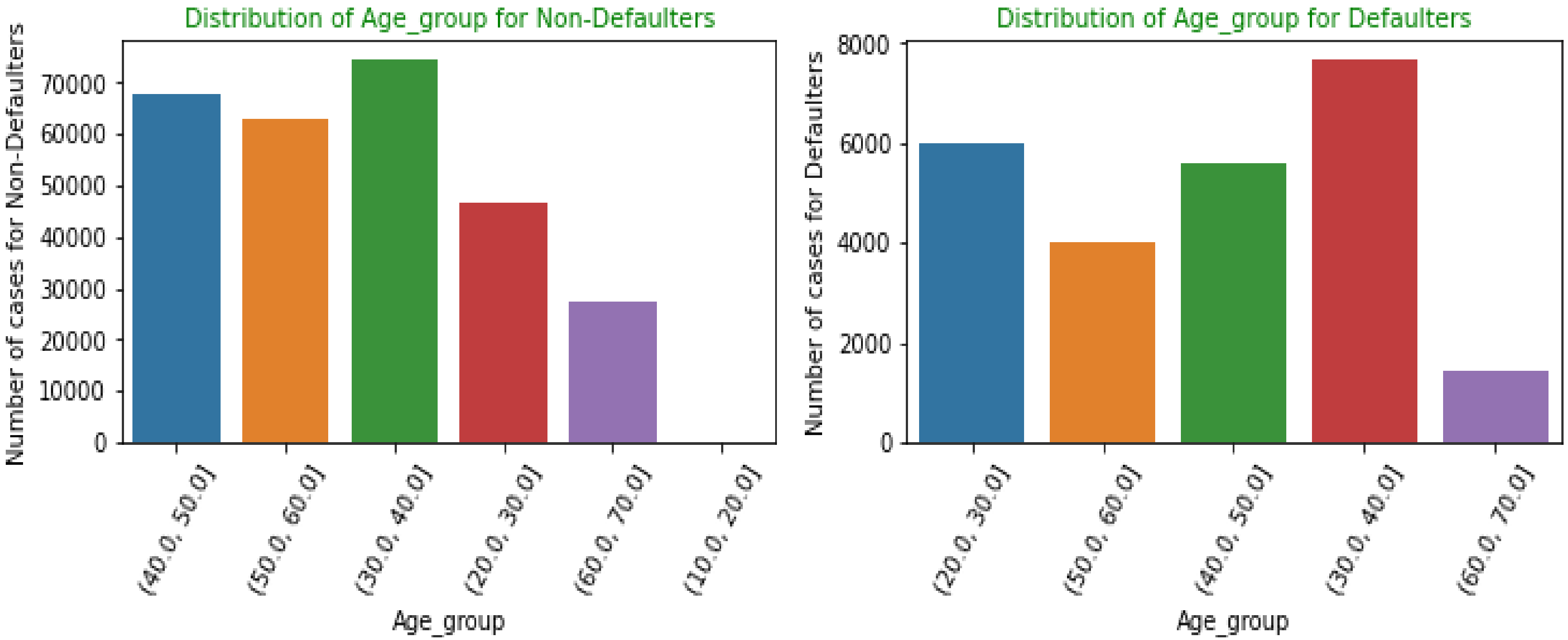

Distribution of NAME_TYPE_SUITE for Defaulters

- Mostly married people apply for loan followed by Single.

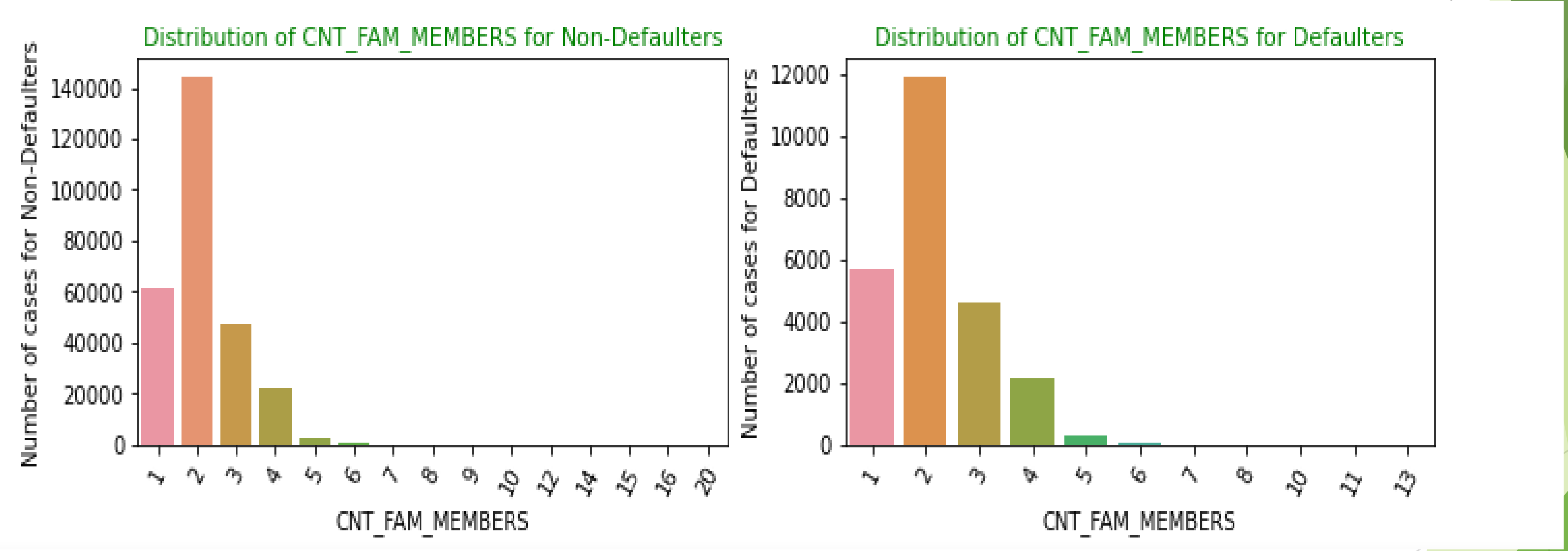




- People of age group 30-40 are mostly likely to apply for loan and also same range people do not able to repay their loan on time.

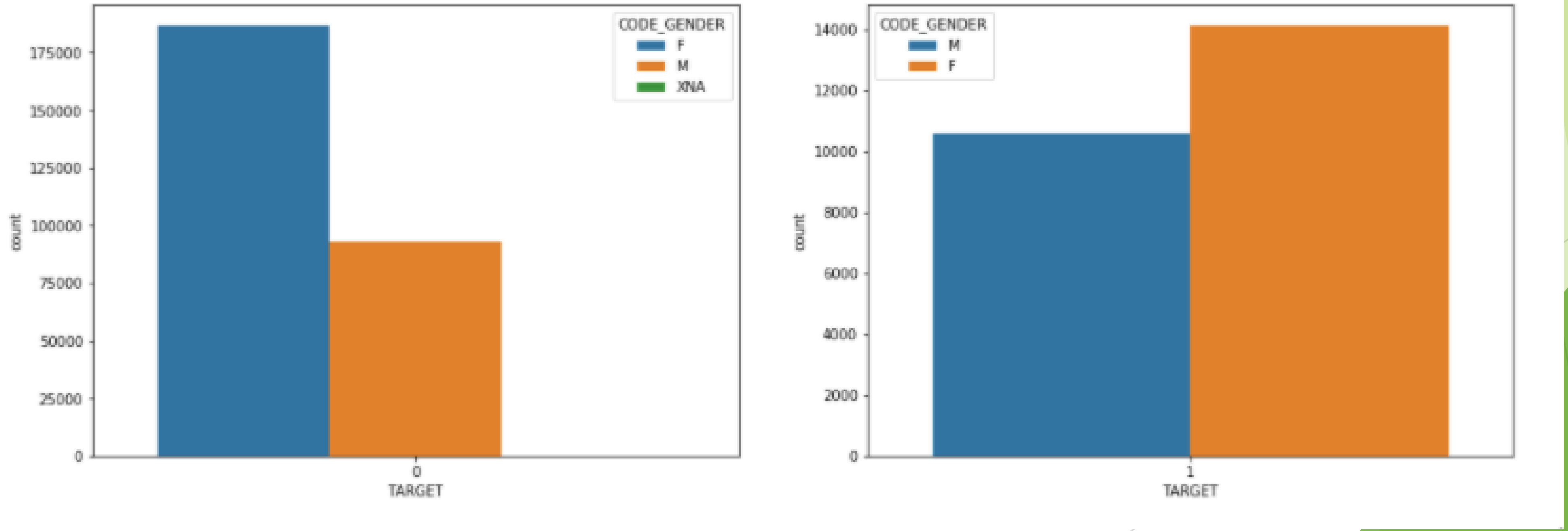


- Family of 2 people most likely to take loan as compare to others and also comes under defaulter list.



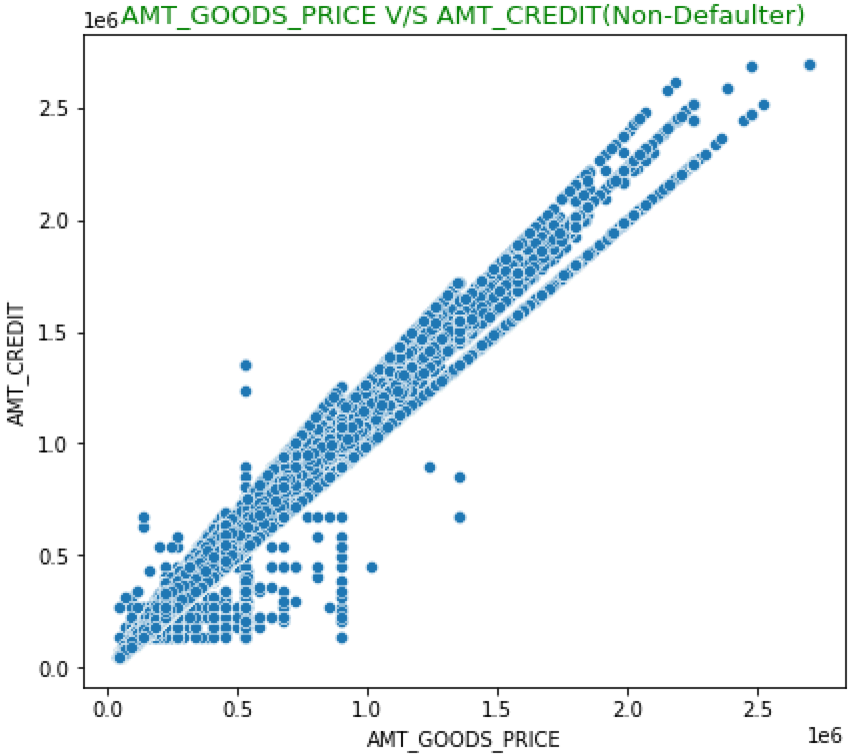
Segmental Univariate Analysis

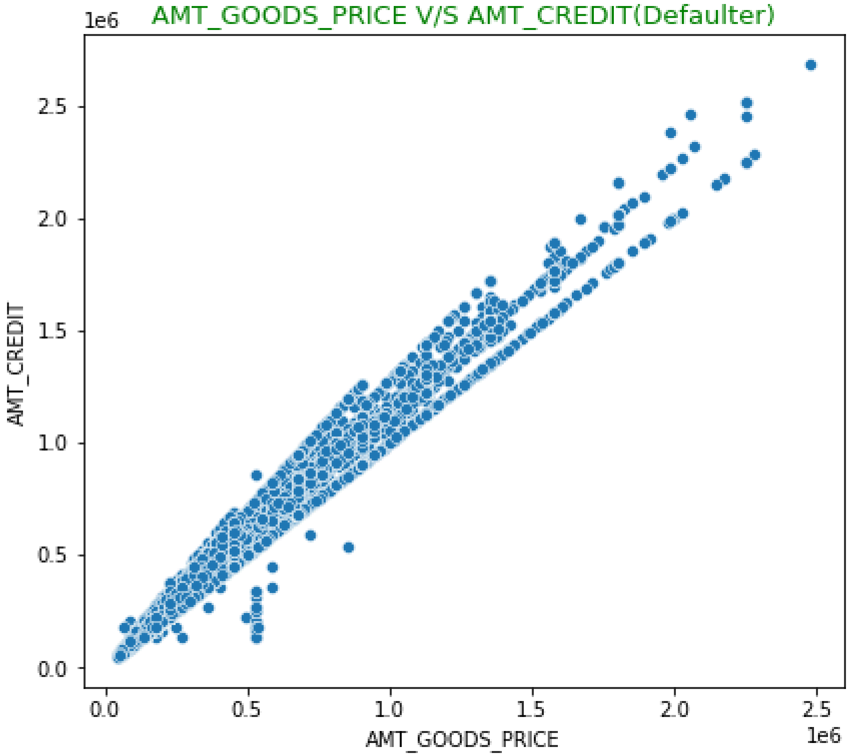
- Segmented univariate analysis allows you to compare subsets of data, which is a powerful technique because it helps you understand how a relevant metric varies across different segments. The way we approach this is by to figure out how to segment/group the variable into smaller buckets that we can compare.
- Here we tried to segment CODE_GENDER in target dataset and found the proportion of males applying for loans and having difficulties in payment is much more than females.



Bivariate analysis

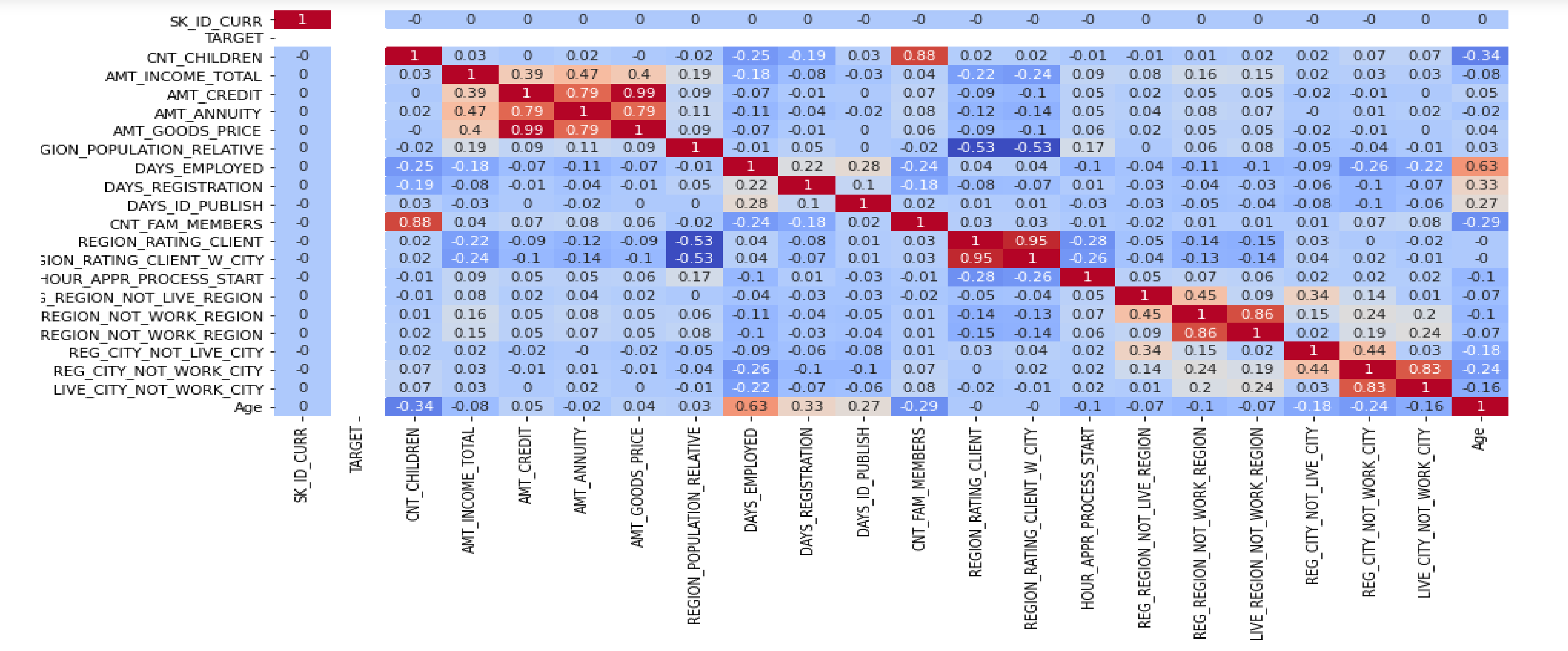
- Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them.
- We can conclude from below graph it becomes difficult for applicant to pay if the 'AMT_CREDIT and AMT_GOODS_PRICE' rises together.


AMT_GOODS_PRICE V/S AMT_CREDIT(Non-Defaulter)


AMT_GOODS_PRICE V/S AMT_CREDIT(Defaulter)

Correlation matrix

- A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables.
- We can use heatmap to represent correlation matrix for dataset Target = 0 and Target = 1
- For target =0



	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	HOUR_APPR_PROCESS_START	REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	REG_CITY_NOT_LIVE_CITY	REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	Age
SK_ID_CURR	1		-0	0	0	0	0	0	0	0	-0	-0	-0	-0	-0	0	0	0	-0	-0	0	0
TARGET																						
CNT_CHILDREN	-0		1	0.03	0	0.02	-0	-0.02	-0.25	-0.19	0.03	0.88	0.02	0.02	-0.01	-0.01	0.01	0.02	0.02	0.07	0.07	-0.34
AMT_INCOME_TOTAL	0		0.03	1	0.39	0.47	0.4	0.19	-0.18	-0.08	-0.03	0.04	-0.22	-0.24	0.09	0.08	0.16	0.15	0.02	0.03	0.03	-0.08
AMT_CREDIT	0		0	0.39	1	0.79	0.99	0.09	-0.07	-0.01	0	0.07	-0.09	-0.1	0.05	0.02	0.05	0.05	-0.02	-0.01	0	0.05
AMT_ANNUITY	0		0.02	0.47	0.79	1	0.79	0.11	-0.11	-0.04	-0.02	0.08	-0.12	-0.14	0.05	0.04	0.08	0.07	-0	0.01	0.02	-0.02
AMT_GOODS_PRICE	0		-0	0.4	0.99	0.79	1	0.09	-0.07	-0.01	0	0.06	-0.09	-0.1	0.06	0.02	0.05	0.05	-0.02	-0.01	0	0.04
GION_POPULATION_RELATIVE	0		-0.02	0.19	0.09	0.11	0.09	1	-0.01	0.05	0	-0.02	-0.53	-0.53	0.17	0	0.06	0.08	-0.05	-0.04	-0.01	0.03
DAYS_EMPLOYED	0		-0.25	-0.18	-0.07	-0.11	-0.07	-0.01	1	0.22	0.28	-0.24	0.04	0.04	-0.1	-0.04	-0.11	-0.1	-0.09	-0.26	-0.22	0.63
DAYS_REGISTRATION	0		-0.19	-0.08	-0.01	-0.04	-0.01	0.05	0.22	1	0.1	-0.18	-0.08	-0.07	0.01	-0.03	-0.04	-0.03	-0.06	-0.1	-0.07	0.33
DAYS_ID_PUBLISH	-0		0.03	-0.03	0	-0.02	0	0	0.28	0.1	1	0.02	0.01	0.01	-0.03	-0.03	-0.05	-0.04	-0.08	-0.1	-0.06	0.27
CNT_FAM_MEMBERS	-0		0.88	0.04	0.07	0.08	0.06	-0.02	-0.24	-0.18	0.02	1	0.03	0.03	-0.01	-0.02	0.01	0.01	0.01	0.07	0.08	-0.29
REGION_RATING_CLIENT	-0		0.02	-0.22	-0.09	-0.12	-0.09	-0.53	0.04	-0.08	0.01	0.03	1	0.95	-0.28	-0.05	-0.14	-0.15	0.03	0	-0.02	-0
GION_RATING_CLIENT_W_CITY	-0		0.02	-0.24	-0.1	-0.14	-0.1	-0.53	0.04	-0.07	0.01	0.03	0.95	1	-0.26	-0.04	-0.13	-0.14	0.04	0.02	-0.01	-0
HOUR_APPR_PROCESS_START	-0		-0.01	0.09	0.05	0.05	0.06	0.17	-0.1	0.01	-0.03	-0.01	-0.28	-0.26	1	0.05	0.07	0.06	0.02	0.02	0.02	-0.1
_REGION_NOT_LIVE_REGION	0		-0.01	0.08	0.02	0.04	0.02	0	-0.04	-0.03	-0.03	-0.02	-0.05	-0.04	0.05	1	0.45	0.09	0.34	0.14	0.01	-0.07
REGION_NOT_WORK_REGION	0		0.01	0.16	0.05	0.08	0.05	0.06	-0.11	-0.04	-0.05	0.01	-0.14	-0.13	0.07	0.45	1	0.86	0.15	0.24	0.2	-0.1
REGION_NOT_WORK_REGION	0		0.02	0.15	0.05	0.07	0.05	0.08	-0.1	-0.03	-0.04	0.01	-0.15	-0.14	0.06	0.09	0.86	1	0.02	0.19	0.24	-0.07
REG_CITY_NOT_LIVE_CITY	-0		0.02	0.02	-0.02	-0	-0.02	-0.05	-0.09	-0.06	-0.08	0.01	0.03	0.04	0.02	0.34	0.15	0.02	1	0.44	0.03	-0.18
REG_CITY_NOT_WORK_CITY	-0		0.07	0.03	-0.01	0.01	-0.01	-0.04	-0.26	-0.1	-0.1	0.07	0	0.02	0.02	0.14	0.24	0.19	0.44	1	0.83	-0.24
LIVE_CITY_NOT_WORK_CITY	0		0.07	0.03	0	0.02	0	-0.01	-0.22	-0.07	-0.06	0.08	-0.02	-0.01	0.02	0.01	0.2	0.24	0.03	0.83	1	-0.16
Age	0		-0.34	-0.08	0.05	-0.02	0.04	0.03	0.63	0.33	0.27	-0.29	-0	-0	-0.1	-0.07	-0.1	-0.07	-0.18	-0.24	-0.16	1

Observation from Top 10 correlated pairs

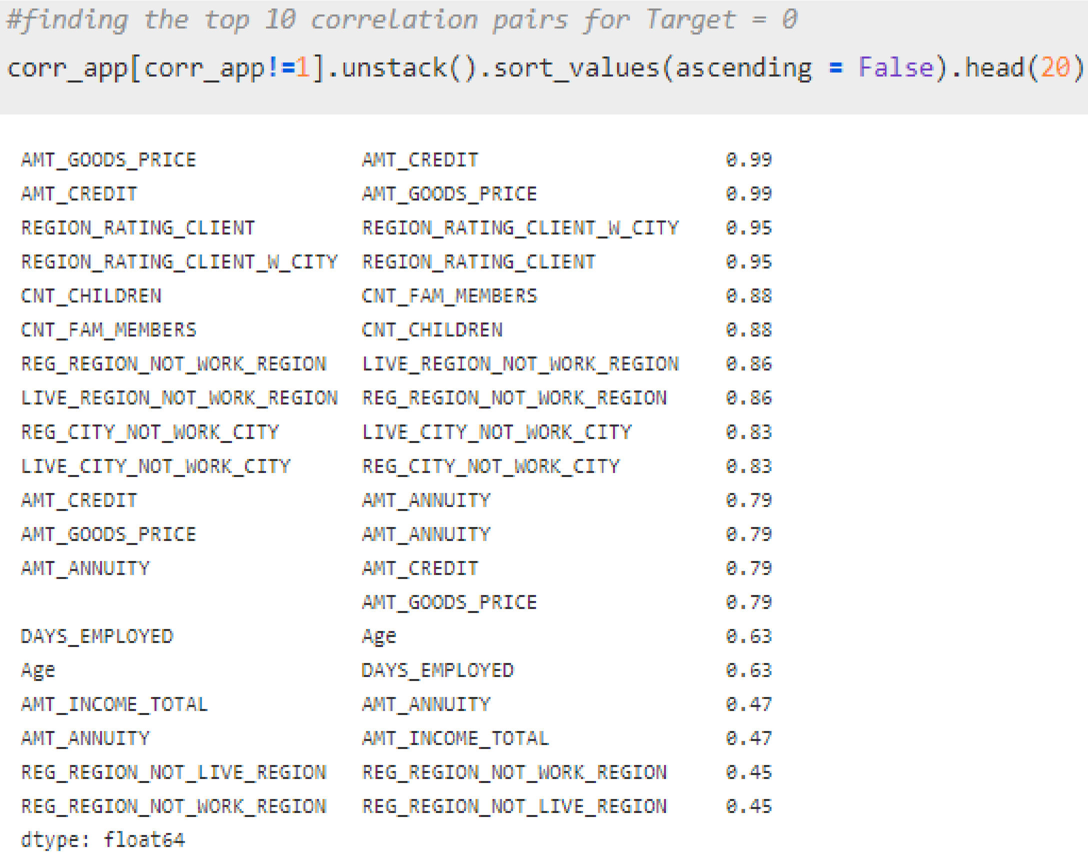
1.Increase in the count of children there is a increase in family member

2.AMT_INCOME_TOTAL increases with the increase in AMT_CREDIT and AMT_ANNUNITY.

3.AMT_GOOD_PRICE increase with the increase in AMT_ANNUNITY

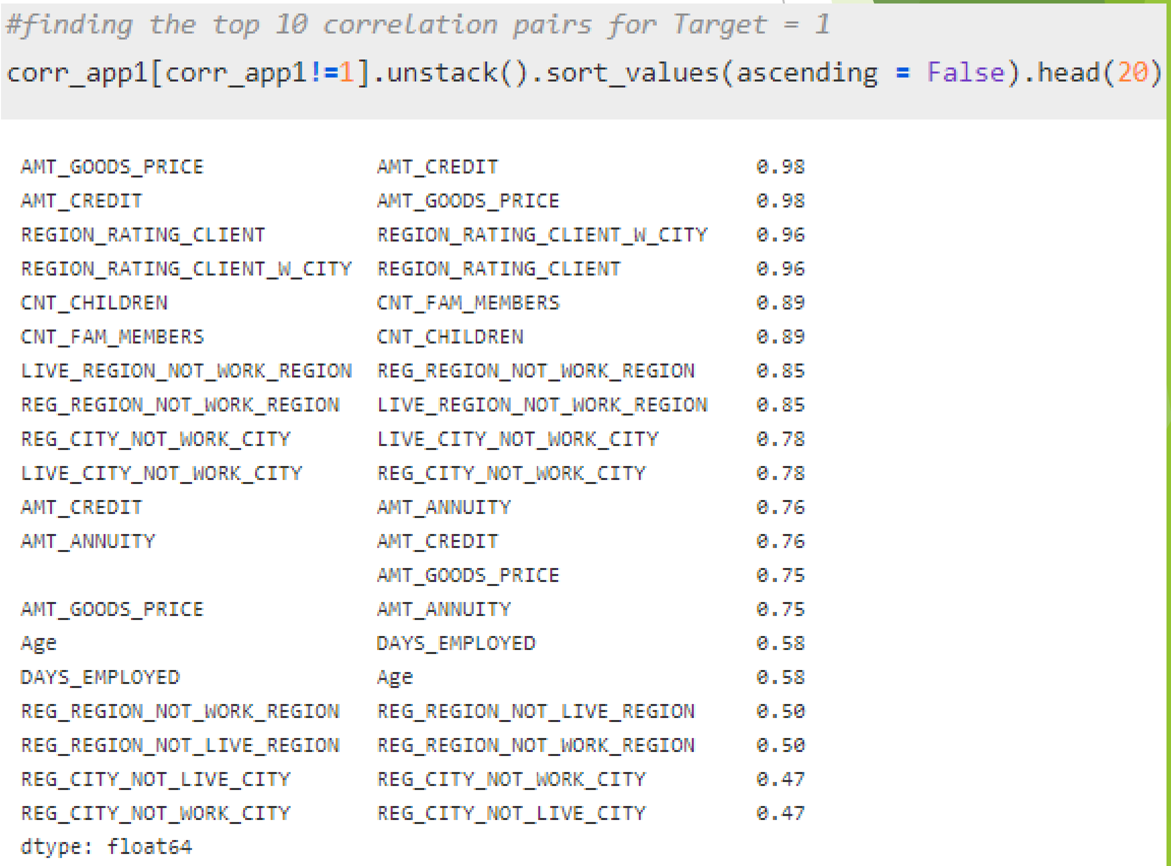
4.AMT_GOOD_PRICE,AMT_CREDIT and AMT_ANNUNITY are somehow related to each other.

5.People work location is different from their permanent address.



```
#finding the top 10 correlation pairs for Target = 0
corr_app[corr_app!=1].unstack().sort_values(ascending = False).head(20)
```

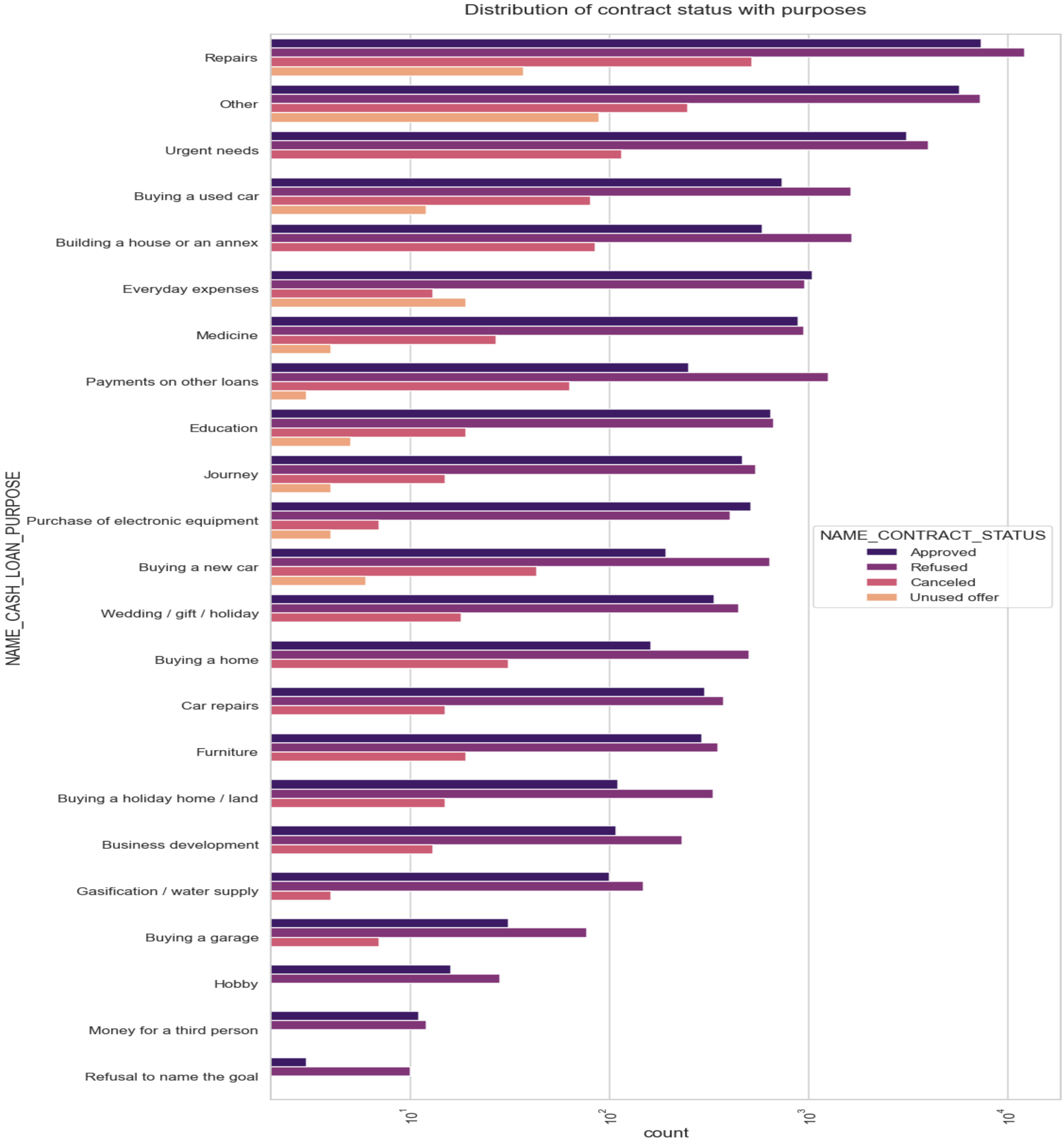
```
AMT_GOODS_PRICE              AMT_CREDIT                     0.99
AMT_CREDIT                   AMT_GOODS_PRICE                0.99
REGION_RATING_CLIENT         REGION_RATING_CLIENT_W_CITY    0.95
REGION_RATING_CLIENT_W_CITY  REGION_RATING_CLIENT           0.95
CNT_CHILDREN                 CNT_FAM_MEMBERS                0.88
CNT_FAM_MEMBERS              CNT_CHILDREN                   0.88
REG_REGION_NOT_WORK_REGION   LIVE_REGION_NOT_WORK_REGION    0.86
LIVE_REGION_NOT_WORK_REGION  REG_REGION_NOT_WORK_REGION     0.86
REG_CITY_NOT_WORK_CITY       LIVE_CITY_NOT_WORK_CITY        0.83
LIVE_CITY_NOT_WORK_CITY      REG_CITY_NOT_WORK_CITY         0.83
AMT_CREDIT                   AMT_ANNUITY                    0.79
AMT_GOODS_PRICE              AMT_ANNUITY                    0.79
AMT_ANNUITY                  AMT_CREDIT                     0.79
                             AMT_GOODS_PRICE                0.79
DAYS_EMPLOYED                Age                            0.63
Age                          DAYS_EMPLOYED                  0.63
AMT_INCOME_TOTAL             AMT_ANNUITY                    0.47
AMT_ANNUITY                  AMT_INCOME_TOTAL               0.47
REG_REGION_NOT_LIVE_REGION   REG_REGION_NOT_WORK_REGION     0.45
REG_REGION_NOT_WORK_REGION   REG_REGION_NOT_LIVE_REGION     0.45
dtype: float64
```



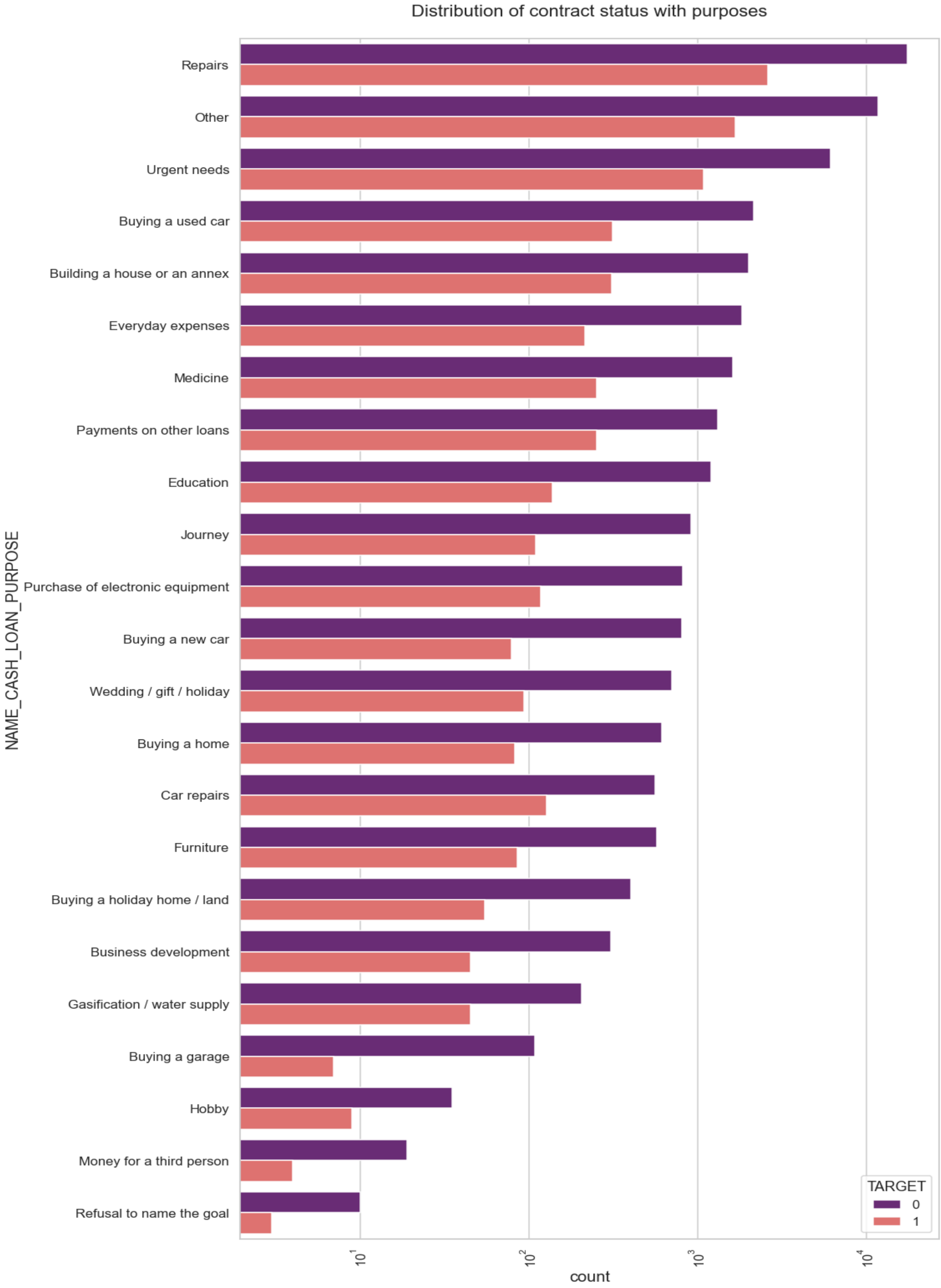
```
#finding the top 10 correlation pairs for Target = 1
corr_app1[corr_app1!=1].unstack().sort_values(ascending = False).head(20)
```

```
AMT_GOODS_PRICE              AMT_CREDIT                     0.98
AMT_CREDIT                   AMT_GOODS_PRICE                0.98
REGION_RATING_CLIENT         REGION_RATING_CLIENT_W_CITY    0.96
REGION_RATING_CLIENT_W_CITY  REGION_RATING_CLIENT           0.96
CNT_CHILDREN                 CNT_FAM_MEMBERS                0.89
CNT_FAM_MEMBERS              CNT_CHILDREN                   0.89
LIVE_REGION_NOT_WORK_REGION  REG_REGION_NOT_WORK_REGION     0.85
REG_REGION_NOT_WORK_REGION   LIVE_REGION_NOT_WORK_REGION    0.85
REG_CITY_NOT_WORK_CITY       LIVE_CITY_NOT_WORK_CITY        0.78
LIVE_CITY_NOT_WORK_CITY      REG_CITY_NOT_WORK_CITY         0.78
AMT_CREDIT                   AMT_ANNUITY                    0.76
AMT_ANNUITY                  AMT_CREDIT                     0.76
                             AMT_GOODS_PRICE                0.75
AMT_GOODS_PRICE              AMT_ANNUITY                    0.75
Age                          DAYS_EMPLOYED                  0.58
DAYS_EMPLOYED                Age                            0.58
REG_REGION_NOT_WORK_REGION   REG_REGION_NOT_LIVE_REGION     0.50
REG_REGION_NOT_LIVE_REGION   REG_REGION_NOT_WORK_REGION     0.50
REG_CITY_NOT_LIVE_CITY       REG_CITY_NOT_WORK_CITY         0.47
REG_CITY_NOT_WORK_CITY       REG_CITY_NOT_LIVE_CITY         0.47
dtype: float64
```

Merging with previous application dataset



Distribution of contract status with purposes



Distribution of contract status with purposes

Merging with previous application dataset

- Observations after merging both dataset:

1. Most rejection of loans came from purpose 'repairs'.

2. For education purposes we have equal number of approves and rejection.

3. Buying a used car is having significant higher rejection than approves.

4. Also when purpose is Building a house there is less scope of cancelling the loan.

5. Loan purposes with 'Repairs' are facing more difficulties in payment on time.

6. There are few places where loan payment is significant higher than facing difficulties. They are 'Buying a garage', 'Business development', 'Buying land', 'Buying a new car' and 'Education‘. We can focus on these purposes for which the client is having for minimal payment difficulties. 'Building a house', 'Everyday house', 'Everyday on the loans' are some purpose in which payment of difficulties is almost same.

Conclusion and Recommendation

- With some following ways bank can be able to handle losses.

1.Revolving loans are comparatively safer than Cash loan.

2.Females are more likely to repay loan inspite of failing to repay on time. Bank can target more to female for profit.

3.Though married people used to take loan more but they are also first in defaulter list, so bank should give

loan to them with caution to avoid loss.

4.Bank can target people whose purpose of loan is education as they are less likely to reject it also they

are less likely to comes under default category.

5.Banks should focus less on income type ‘Working’ as they are having most number of unsuccessful payments.

6.Bank can focus mostly on housing type ‘with parents’ , ‘House\apartment’ and ‘municipal apartment’ for successful payments.

Thank You