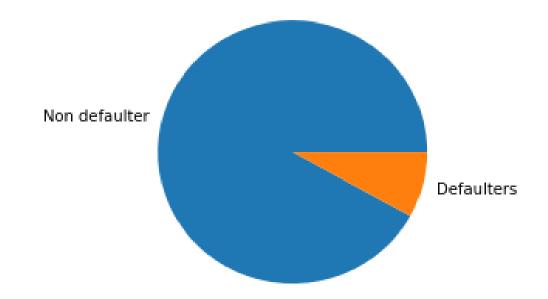
### CREDIT EDA CASE STUDY

BY:: Rohit Lal and Monica Fatwani

# Application Data Frame

## **Target**

Distribution of TARGET (defaulter/non-defaulter)

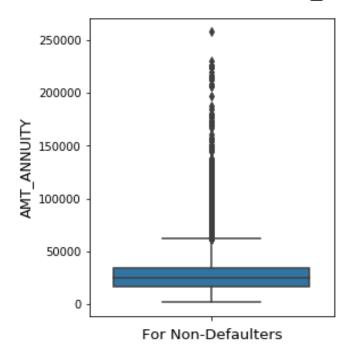


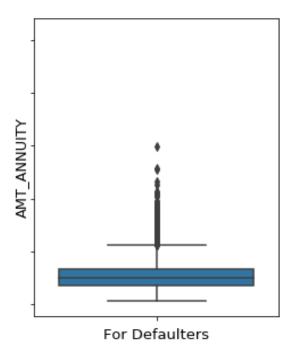
# Univariate Analysis - For Continuous Data Types:

### **AMT\_ANNUITY**

- 1) No strong comparison found in AMT\_ANNUITY with the TARGET variable
- 2) Non-defaulters have in generally loans with very high Annuity amounts (see, outliers in the above plot. Non-defaulters have very high magnitude of Outliers.)
- 3) This can be due to the High Annuity loans are provided only to trusted people which are Non-defaulters

#### Distribution of AMT\_ANNUITY across TARGET variable

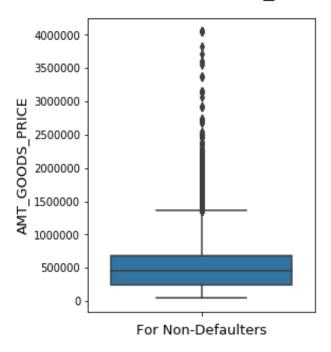


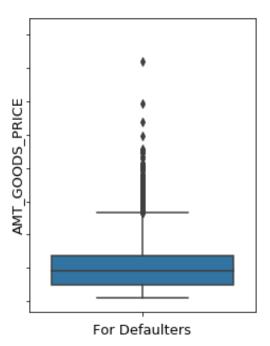


#### AMT\_GOODS\_PRICE

- 1) AMT\_GOODS\_PRICE column has quite high numbers of outliers.
- 2) Again, mean would not be an appropriate measure. We can use Median to estimate the missing values

#### Distribution of AMT\_GOODS\_PRICE across TARGET variable

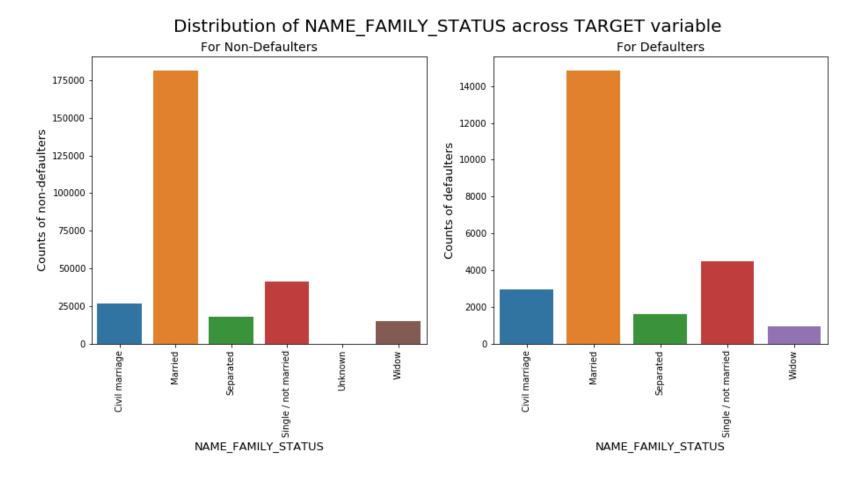




Univariate Analysis -For Categorical Data types(Unordered):

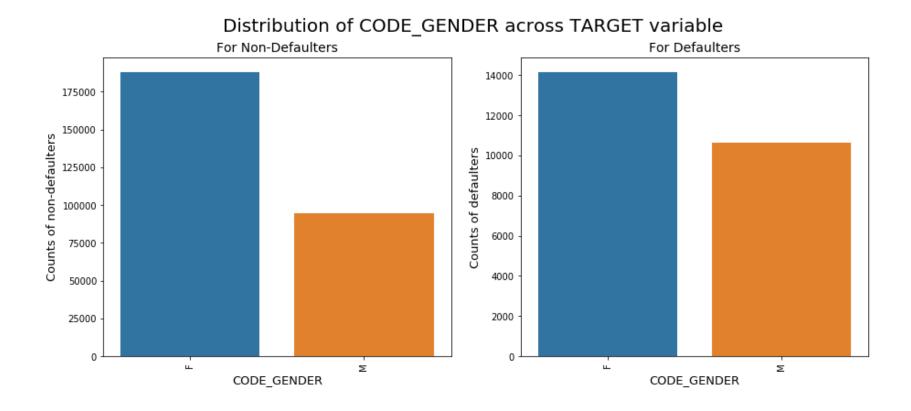
#### NAME\_FAMILY\_STATUS

- 1) Married people are at top in both defaulters and non-defaulters
- 2) Not-married ratio in defaulters is comparatively higher that that of non defaulted population.



### CODE\_GENDER

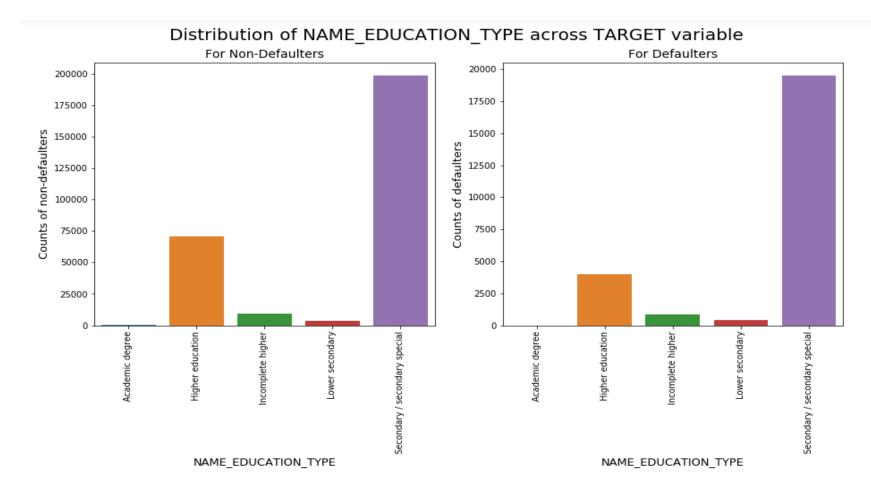
1) Ratio of Female to Male in case of defaulters is comparatively less than that in case of Non-defaulters.



 Univariate Analysis -For Categorical Data types(ordered):

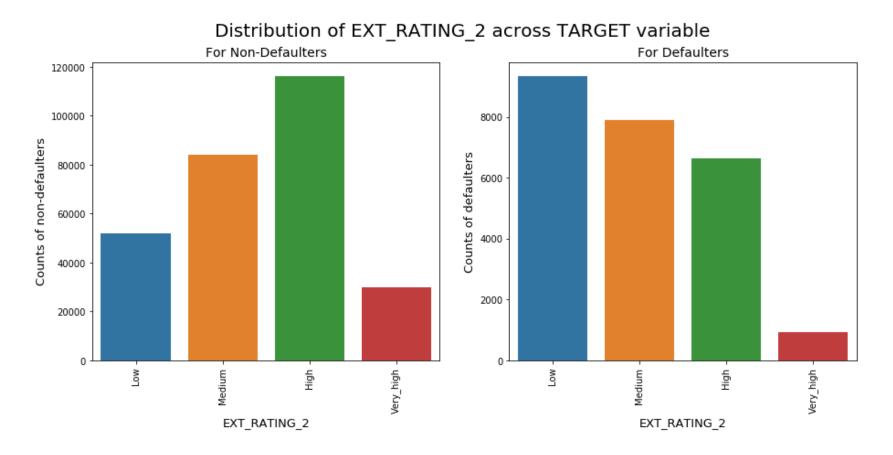
#### NAME\_EDUCATION\_TYPE

- 1) There is no strong relation between NAME\_EDUCATION\_TYPE & TARGET variable.
- 2) In general, Education Type 'Secondary/Secondary Special' is highest and Education Type 'Academic degree' is least for both TARGET categories.



#### EXT\_RATING\_2

- 1) Low Ratings ratio in the Defaulters population is clearly higher as compared to the Non-defaulter population.
- 2) EXT\_RATING\_2 shows strong relationship with our target variable (i.e., defaulters/non-defaulters)



Correlation for Non-Defaulters

### Correlation for Non-Defaultors

| SK_ID_CURR                 | - 1          | -0.00027       | 0.0022             | 0.00054      | 0.0011        | 0.00091           | 0.00042                      | 0.0014       | 0.001           | -0.0026           | 9.2e-05                   | 9.9e-05                      | 0.00094        | 0.00028        | 0.0011                   |
|----------------------------|--------------|----------------|--------------------|--------------|---------------|-------------------|------------------------------|--------------|-----------------|-------------------|---------------------------|------------------------------|----------------|----------------|--------------------------|
| CNT_CHILDREN               | ~0.00027     | 1              | 0.038              | 0.0023       | 0.025         | -0.0013           | -0.026                       | -0.38        | -0.16           | 0.81              | 0.0017                    | -0.011                       | -0.017         | -0.046         | 0.0097                   |
| AMT_INCOME_TOTAL           | - 0.0022     | 0.038          | 1                  | 0.42         | 0.49          | 0.42              | 0.098                        | -0.094       | -0.17           | 0.045             | 0.1                       | 0.078                        | 0.18           | -0.095         | 0.063                    |
| AMT_CREDIT                 | -0.00054     | 0.0023         | 0.42               | 1            | 0.83          | 0.99              | 0.054                        | 0.057        | -0.12           | 0.079             | 0.051                     | 0.019                        | 0.13           | 0.019          | 0.077                    |
| AMT_ANNUITY                | - 0.0011     | 0.025          | 0.49               | 0.83         | 1             | 0.83              | 0.059                        | -0.012       | -0.13           | 0.096             | 0.054                     | 0.036                        | 0.13           | 0.015          | 0.068                    |
| AMT_GOODS_PRICE            | -0.00091     | -0.0013        | 0.42               | 0.99         | 0.83          | 1                 | 0.063                        | 0.058        | -0.12           | 0.079             | 0.059                     | 0.02                         | 0.14           | 0.024          | 0.082                    |
| REGION_POPULATION_RELATIVE | -0.00042     | -0.026         | 0.098              | 0.054        | 0.059         | 0.063             | 1                            | 0.031        | -0.0026         | -0.018            | 0.13                      | -0.02                        | 0.19           | -0.0059        | 0.036                    |
| DAYS_BIRTH                 | - 0.0014     | -0.38          | -0.094             | 0.057        | -0.012        | 0.058             | 0.031                        | 1            | 0.23            | -0.28             | -0.098                    | -0.066                       | 0.089          | 0.2            | 0.068                    |
| DAYS_EMPLOYED              | - 0.001      | -0.16          | -0.17              | -0.12        | -0.13         | -0.12             | -0.0026                      | 0.23         | 1               | -0.18             | -0.049                    | 0.022                        | -0.082         | -0.0095        | -0.13                    |
| CNT_FAM_MEMBERS            | 0.0026       | 0.81           | 0.045              | 0.079        | 0.096         | 0.079             | -0.018                       | -0.28        | -0.18           | 1                 | -0.0085                   | -0.016                       | 0.00034        | -0.026         | 0.034                    |
| HOUR_APPR_PROCESS_START    | - 9.2e-05    | 0.0017         | 0.1                | 0.051        | 0.054         | 0.059             | 0.13                         | -0.098       | -0.049          | -0.0085           | 1                         | 0.052                        | 0.16           | -0.043         | 0.0063                   |
| REG_REGION_NOT_LIVE_REGION | - 9.9e-05    | -0.011         | 0.078              | 0.019        | 0.036         | 0.02              | -0.02                        | -0.066       | 0.022           | -0.016            | 0.052                     | 1                            | 0.024          | -0.048         | -0.035                   |
| EXT_SOURCE_2               | -0.00094     | -0.017         | 0.18               | 0.13         | 0.13          | 0.14              | 0.19                         | 0.089        | -0.082          | 0.00034           | 0.16                      | 0.024                        | 1              | 0.085          | 0.2                      |
| EXT_SOURCE_3               | -0.00028     | -0.046         | -0.095             | 0.019        | 0.015         | 0.024             | -0.0059                      | 0.2          | -0.0095         | -0.026            | -0.043                    | -0.048                       | 0.085          | 1              | 0.053                    |
| DAYS_LAST_PHONE_CHANGE     | - 0.0011     | 0.0097         | 0.063              | 0.077        | 0.068         | 0.082             | 0.036                        | 0.068        | -0.13           | 0.034             | 0.0063                    | -0.035                       | 0.2            | 0.053          | 1                        |
|                            | SK_ID_CURR - | CNT_CHILDREN - | AMT_INCOME_TOTAL - | AMT_CREDIT - | AMT_ANNUITY - | AMT_GOODS_PRICE - | REGION_POPULATION_RELATIVE - | DAYS_BIRTH - | DAYS_EMPLOYED - | CNT_FAM_MEMBERS - | HOUR_APPR_PROCESS_START - | REG_REGION_NOT_LIVE_REGION - | EXT_SOURCE_2 - | EXT_SOURCE_3 - | DAYS_LAST_PHONE_CHANGE - |

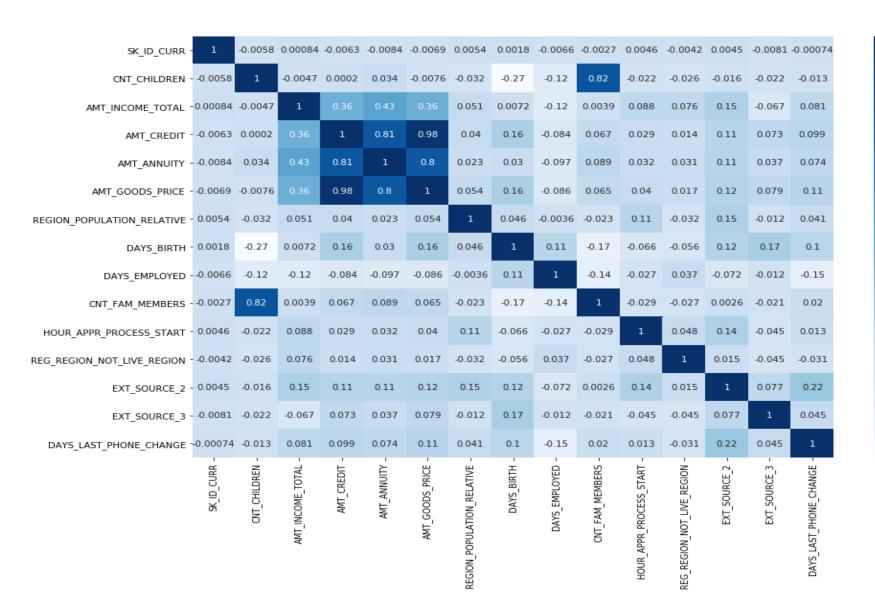
- 0.6 - 0.4 - 0.2 - 0.0 - -0.2

### Correlation for Non-Defaulters(Type 0):

- Total income and Credit amount are strongly related. It means wealthy people take generally higher credit amounts.
- Credit Amount and Annuity are strong related. It is due the behaviour of business - For high credit amount loans, annuity will always be higher.
- Children count and Family member count shows strong relation. Again, it is due the behaviour of the variable. More children constitutes to more family members.
- Children and population density are also inversely related, means in densely populated region, clients have less children. This may due to space related constraints or some government body guidelines in dense regions.
- Credit amount and population density are positively related. It means denser areas constitute good credit amount.

Correlation for Defaulters

#### Correlation for Defaultors



1.0 - 0.8 - 0.6 - 0.4 - 0.2 - 0.0 - -0.2

### Correlation for Non-Defaulters(Type 1):

- For Defaulters also, the prominent correlated variables are same as non-defaulters.
- Total income and Credit amount are strongly related. It means wealthy people take generally higher credit amounts.
- Credit Amount and Annuity are strong related. It is due the behaviour of business - For high credit amount loans, annuity will always be higher.
- Children count and Family member count shows strong relation. Again, it is due the behaviour of the variable. More children constitutes to more family members.
- Children and population density are also inversely related, means in densely populated region, clients have less children. This may due to space related constraints or some government body guidelines in dense regions.
- Credit amount and population density are positively related. It means denser areas constitute good credit amount.

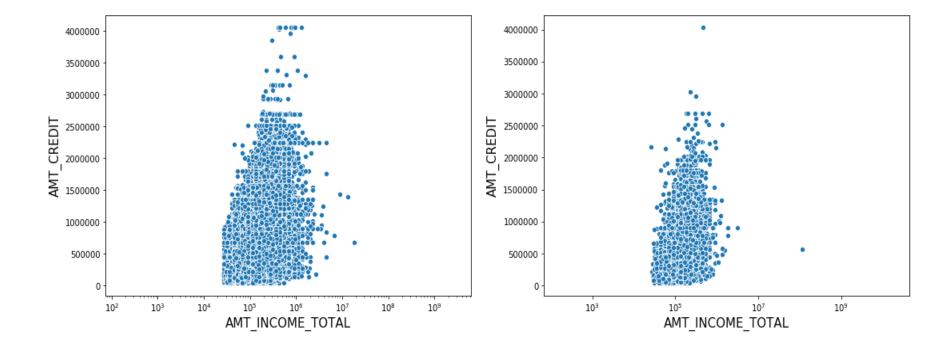
### Bivariate (Continuous-Continuous)

### CREDIT v/s INCOME

- Higher income candidates in general present in Non-defaulter population.
- Log Scale for INCOME\_TOTAL in above plots is chosen to take care of outlier points.

CREDIT vs INCOME for Non-Defaulters

CREDIT vs INCOME for Defaulters

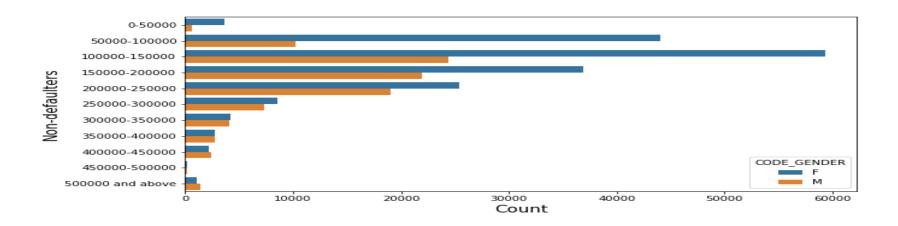


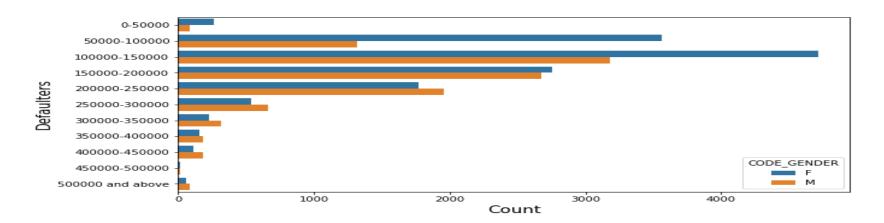
 Bivariate(Continuous-Categorical)

#### AMT\_INCOME\_RANGE v/s CODE\_GENDER

- 1) Female counts are higher than male.
- 2) Income range from 50000 to 250000 is having more number of credits.

 ${\tt Distribution\ of\ AMT\_INCOME\_RANGE\ and\ CODE\_GENDER\ for\ TARGET}$ 

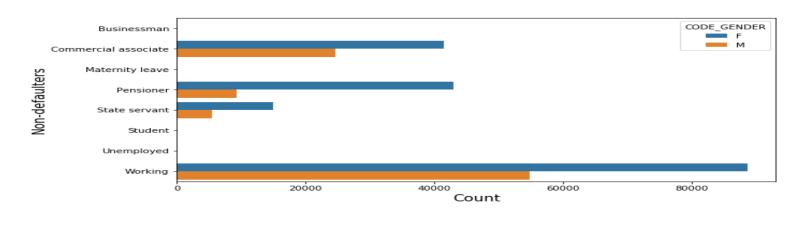


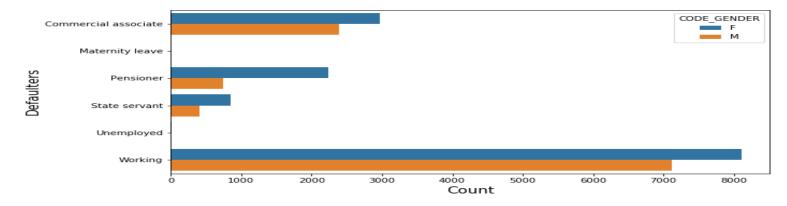


#### NAME\_INCOME\_TYPE v/s CODE\_GENDER

- 1) For income type working, commercial associate, and State Servant, pensioner the number of credits are higher than others.
- 2) For this Females are having more number of credits than male. Less number of credits for income type student, Businessman and Unemployed.

 ${\tt Distribution\ of\ NAME\_INCOME\_TYPE\ and\ CODE\_GENDER\ for\ TARGET}$ 



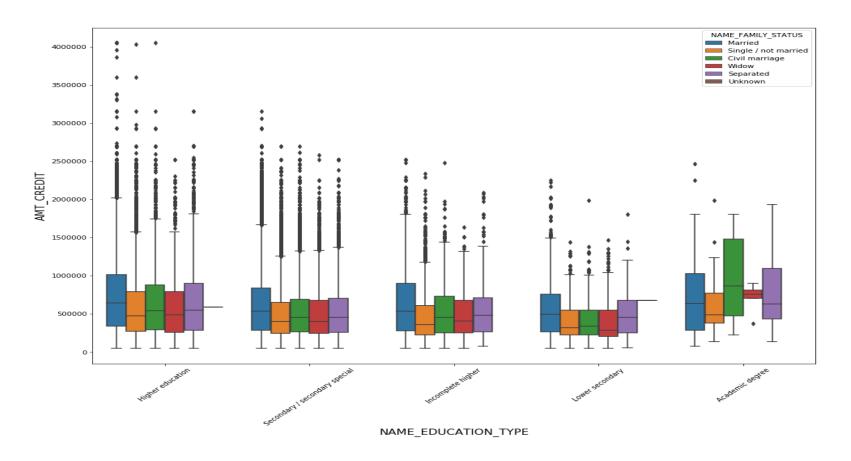


Multivariate Analysis:

### NAME\_EDUCATION\_TYPE v/s AMT\_CREDIT v/s NAME\_FAMILY\_STATUS - Non Defaulters

- 1)In Academic degree category, AMT\_CREDIT counts are much higher for Civil Marriage group.
- 2)In Academic degree category, AMT\_CREDIT counts are minimum for Widow group.

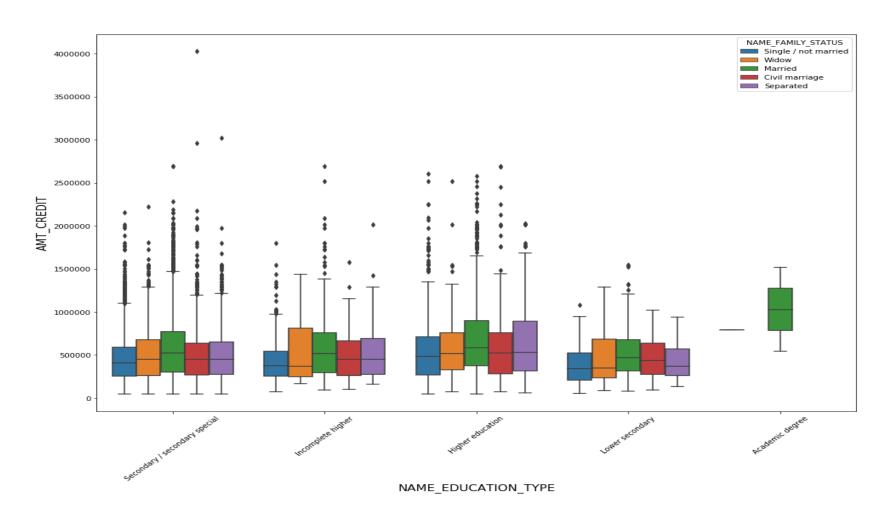
NAME\_EDUCATION\_TYPE vs AMT\_CREDIT for Non-Defaulters with grouping by NAME\_FAMILY\_STATUS



#### NAME\_EDUCATION\_TYPE v/s AMT\_CREDIT v/s NAME\_FAMILY\_STATUS - Defaulters

- 1) In Academics category, only Married group has defaulters.
- 2) So, bank must prefer applicants with Academics degree and Non-married.

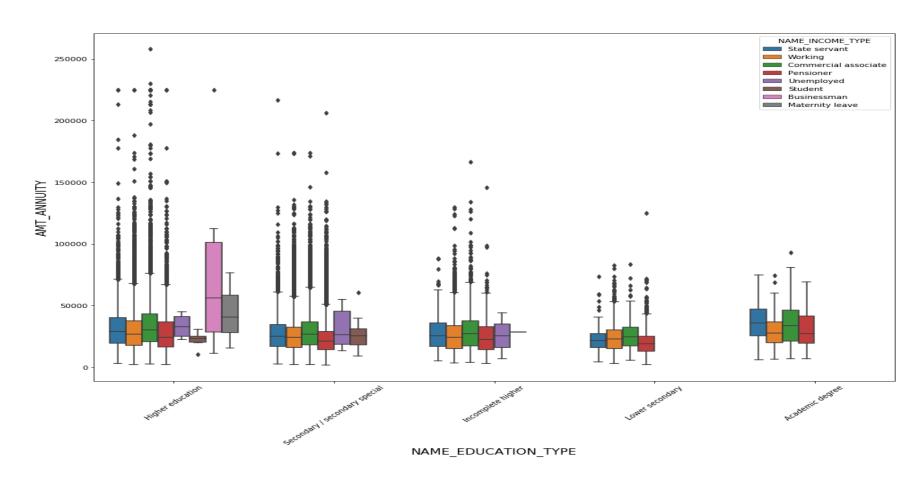
NAME\_EDUCATION\_TYPE vs AMT\_CREDIT for Defaulters with grouping by NAME\_FAMILY\_STATUS



#### NAME\_EDUCATION\_TYPE v/s AMT\_ANNUITY v/s NAME\_INCOME\_TYPE - Non-Defaulters

- 1) There are many applicants from Secondary/secondary special education type whose loan gets approved.
- 2) Higher Education:Businessman, Higher Education:Maternity leave and Higher Education:students should be targeted as they pay their loan on time as per the data

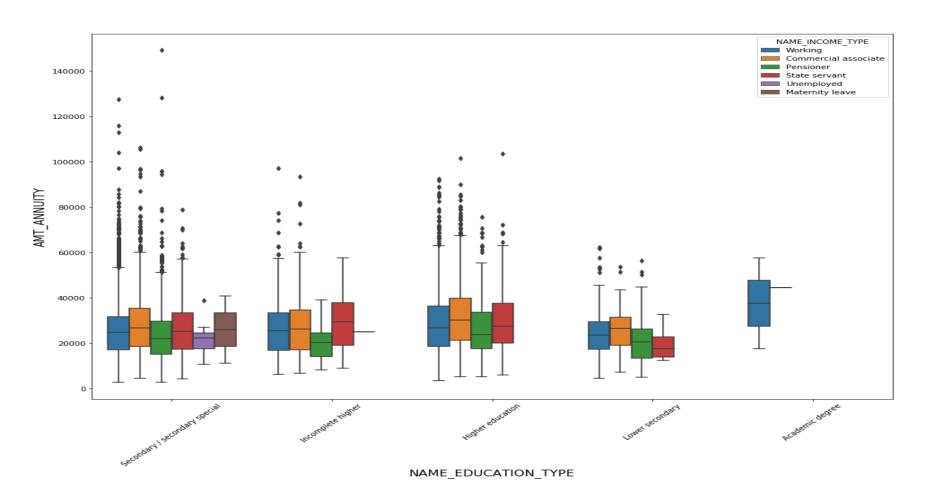
NAME\_EDUCATION\_TYPE vs AMT\_ANNUITY for Non-Defaulters with grouping by NAME\_INCOME\_TYPE



#### NAME\_EDUCATION\_TYPE v/s AMT\_ANNUITY v/s NAME\_INCOME\_TYPE-Defaulters

- Bank must focus on Academic degree Education Type non-working type as they are having higher number of successful payments on time.

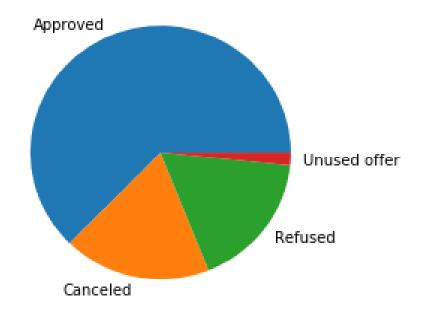
NAME\_EDUCATION\_TYPE vs AMT\_ANNUITY for Defaulters with grouping by NAME\_INCOME\_TYPE



Previous Application dataframe

### NAME\_CONTRACT\_STATUS

### Distribution of Contract Status



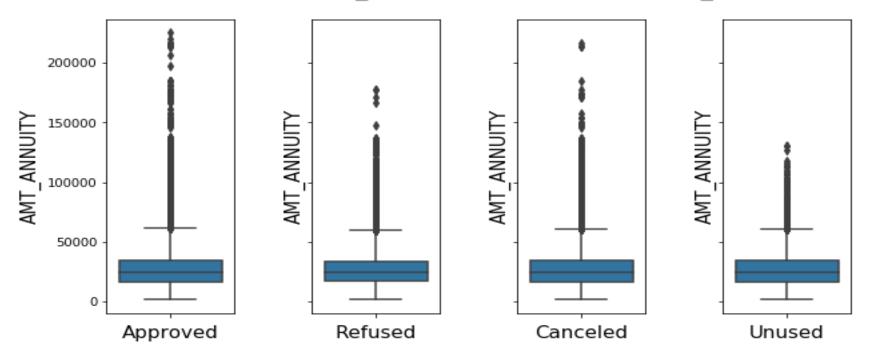
### Univariate Analysis-

(Continuous Variables across Contract Status)

### AMT\_ANNUITY

- No strong comparison visible in AMT\_ANNUITY with the CONTRACT\_STATUS variable
- Refused & Unused have comparatively lesser Outliers count than the Approved & Canceled.

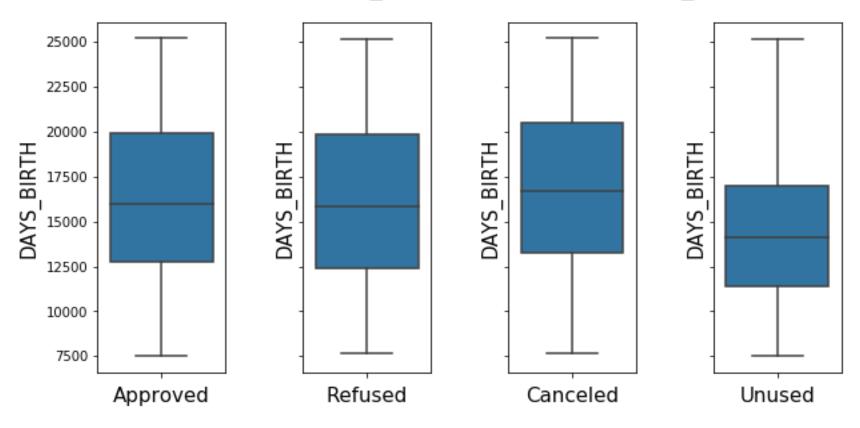
Distribution of AMT\_ANNUITY across CONTRACT\_STATUS



### DAYS\_BIRTH

 Unused loans have in general lesser DAYS\_BIRTH, i.e., younger clients are often constituting Unused loans.

Distribution of DAYS\_BIRTH across CONTRACT\_STATUS



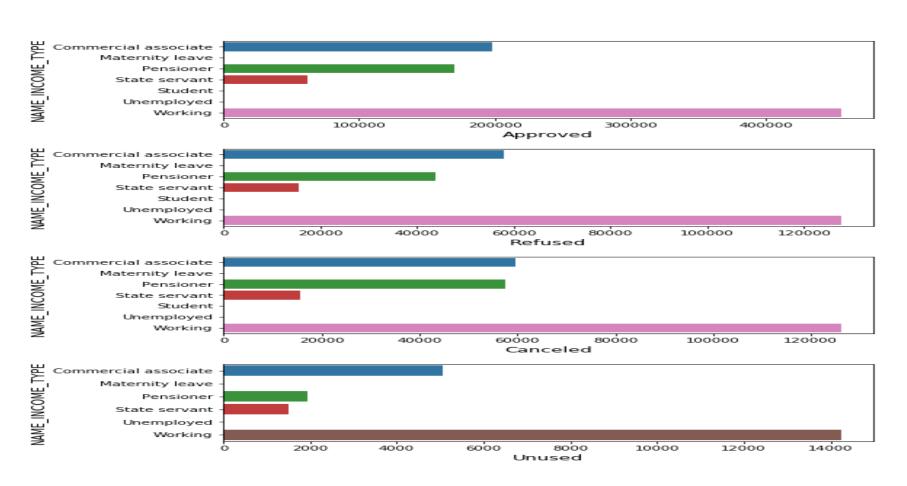
# Univariate Analysis-

(Categorical Variables across Contract Status)

#### NAME\_INCOME\_TYPE

In Canceled category, there is a huge difference in income type Pensioner & State servant

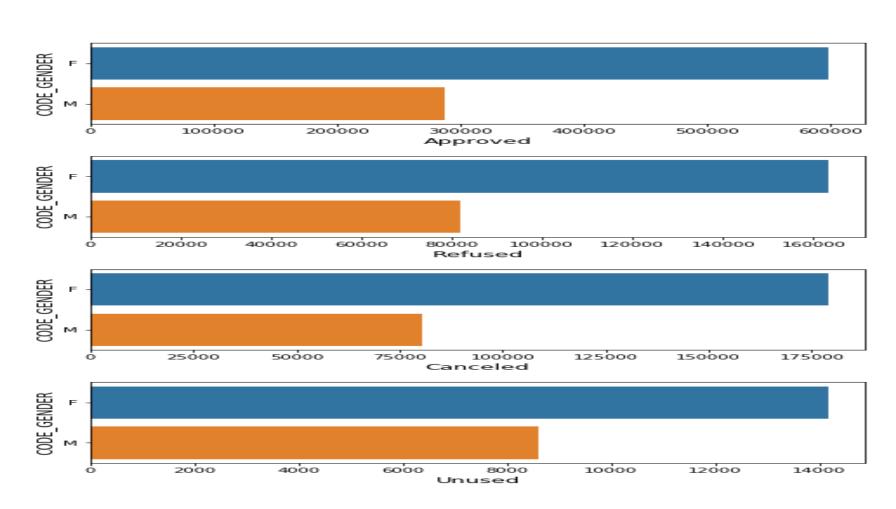
#### Distribution of NAME\_INCOME\_TYPE vs CONTRACT\_STATUS



#### CODE\_GENDER

Males have considerably lesser counts in Canceled category as compared to Females.

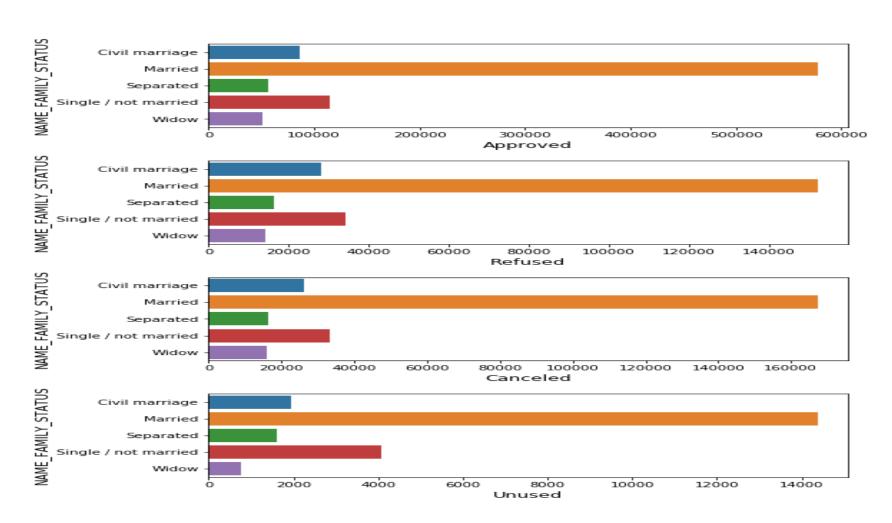
#### Distribution of CODE\_GENDER vs CONTRACT\_STATUS



#### NAME\_FAMILY\_STATUS

- 1) Married people are at top in all CONTRACT\_STATUS categories.
- 2) Single/Not-married ratio is comparatively higher in Unused category.

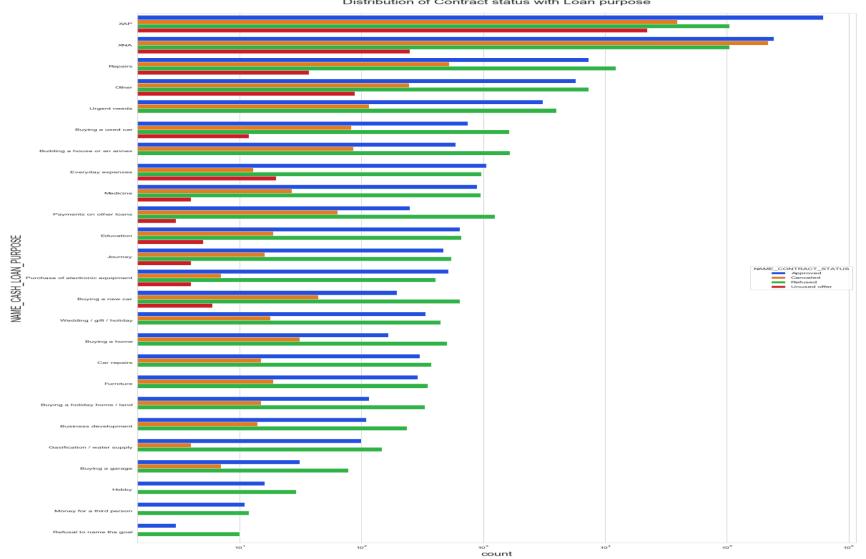
  Distribution of NAME\_FAMILY\_STATUS VS CONTRACT\_STATUS



## Bivariate Analysis

#### NAME\_CASH\_LOAN\_PURPOSE v/s NAME\_CONTRACT\_STATUS

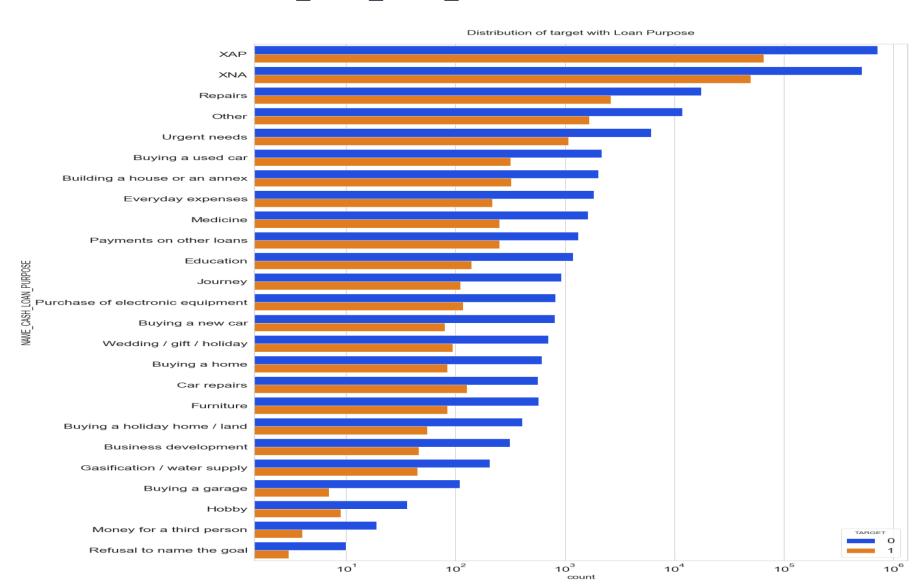




#### NAME\_CASH\_LOAN\_PURPOSE v/s NAME\_CONTRACT\_STATUS

- Repairs have most reject loans.
- In Medicines we have almost equal number of approved & rejected loans.
- In Education we have almost equal number of approved & rejected loans.
- Paying other loans and buying a new car is having significant higher rejection than approves.

#### NAME\_CASH\_LOAN\_PURPOSE v/s TARGET

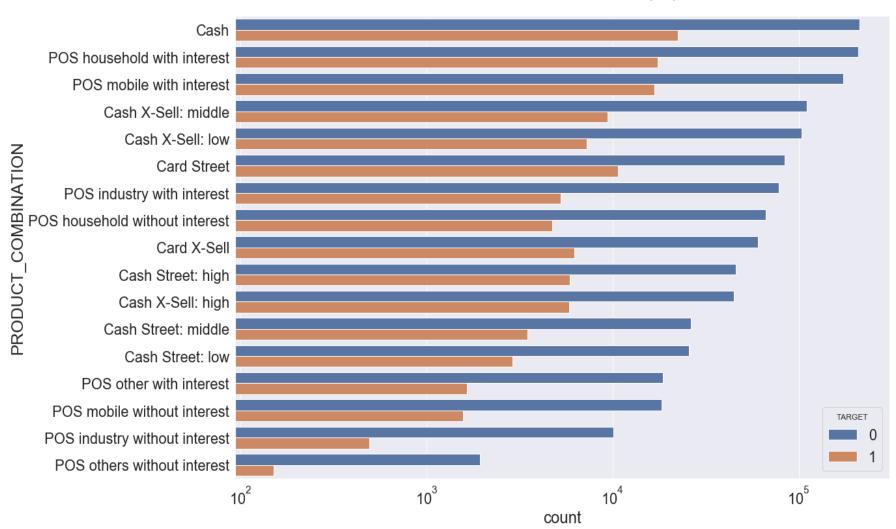


#### NAME\_CASH\_LOAN\_PURPOSE v/s TARGET

- Repairs have considerably high count of defaulters.
- Buying a garage has comparatively lesser defaulters and more non-defaulters.
- Other loan purposes where non-defaulters count are significantly greater than the defaulter count are Business Development, Buying a home, Buying a land, Buying a new car.

#### PRODUCT\_COMBINATION v/s TARGET

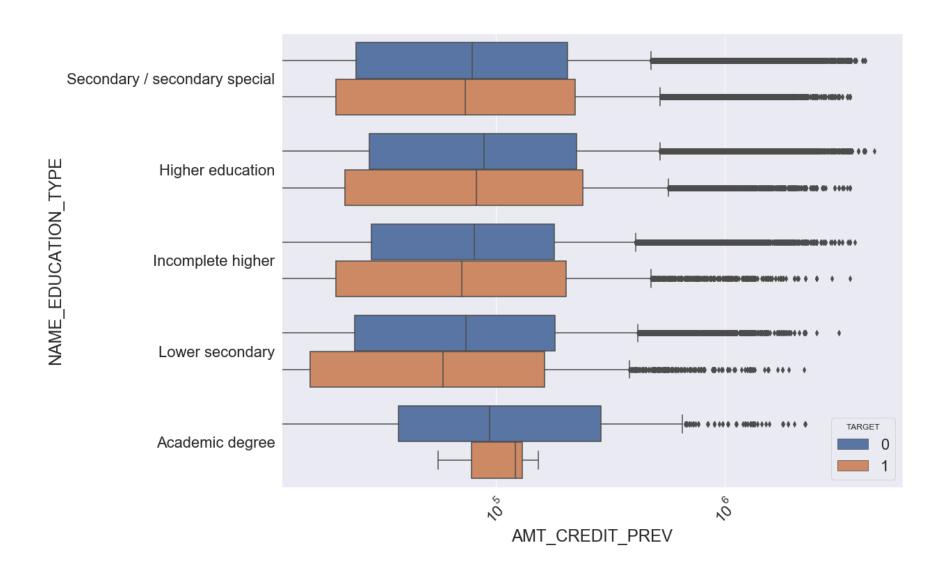
Distribution of contract status with purposes



#### PRODUCT\_COMBINATION v/s TARGET

- POS others without interest, POS industry without interest Product combinations should be targeted as their defaulter percentage is low compared to others.
- Highest defaulter rate is for Product combination cash and highest non-defauters are also from Cash product combination. This may be because Cash combination count is high in our data set.

#### AMT\_CREDIT\_PREV v/s NAME\_HOUSING\_TYPE v/s TARGET



#### AMT\_CREDIT\_PREV v/s NAME\_HOUSING\_TYPE v/s TARGET

- Office apartment, Co-op apartment have higher AMT\_CREDIT for defaulters.
- Municipal apartment have higher AMT\_CREDIT for non-defaulters.
- So, banks can focus on giving loans to House/apartment, Municipal Apartments, as they show positive results being non-defaulters.

### CONCLUSION

- Higher Education:Businessman, Higher Education:Maternity leave and Higher Education:students should be targeted as they pay their loan on time as per the data.
- Bank must focus on Academic degree Education Type non-working type as they are having higher number of successful payments on time.
- Loan purpose Repair is having higher number of unsuccessful payments on time.
- Buying a garage, Business development, Buying land, Buying a new car and Education should be targeted as they are having minimum payment difficulties.
- POS others without interest, POS industry without interest Product combinations should be targeted as their defaulter percentage is low compared to others.
- Banks can focus on giving loans to House/apartment, Municipal Apartments, as they show positive results being non-defaulters.

# Thanks for reading!