Analysis of EDA Case Study (Identification of Loan Defaulters)

By ABHAS RAMTEKE

#Routine Checkings for Application_Data (ds1).

Imported all the necessary libraries in Jupyter notebook(named EDA Assignment).

Read dataset (downloaded in folder by coping as path)

All the necessary things checked first, before making any changes like information, shape & null values to get the clarity of dataset.

Checked null value count & data type of columns.

#Data Cleaning

Identified most of the columns of dataset (ds1) has null values.

In figured out data, we have applied code to all the columns of dataset to remove null values >50%, as these columns has no use of analyse.

After applying code, we left with 81 columns.

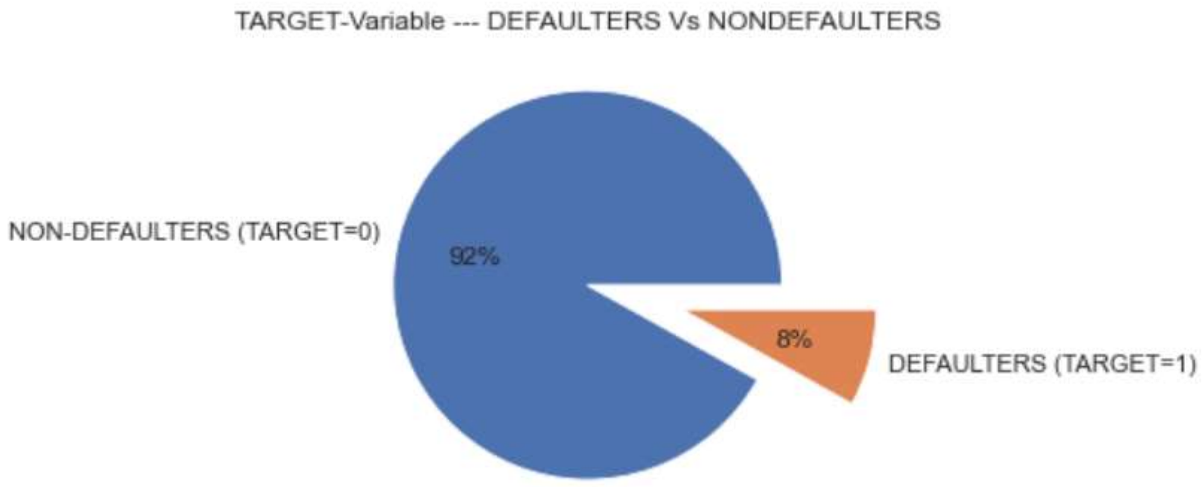
Checked with null percentage, AMT_ANNUITY column has outliers hence, values changed with median.

DATA IMBALANCE

Now,we have checked for imbalance in target variable of dataset.

Applied analysis on 'Target' column, Found difference between Defaulters & Non-Defaulters as shown in pie chart.

92% people comes under Non-Defaulters in Application_dataset while Defaulters are 8% .



Target Variable—Defaulters Vs Non-Defaulters

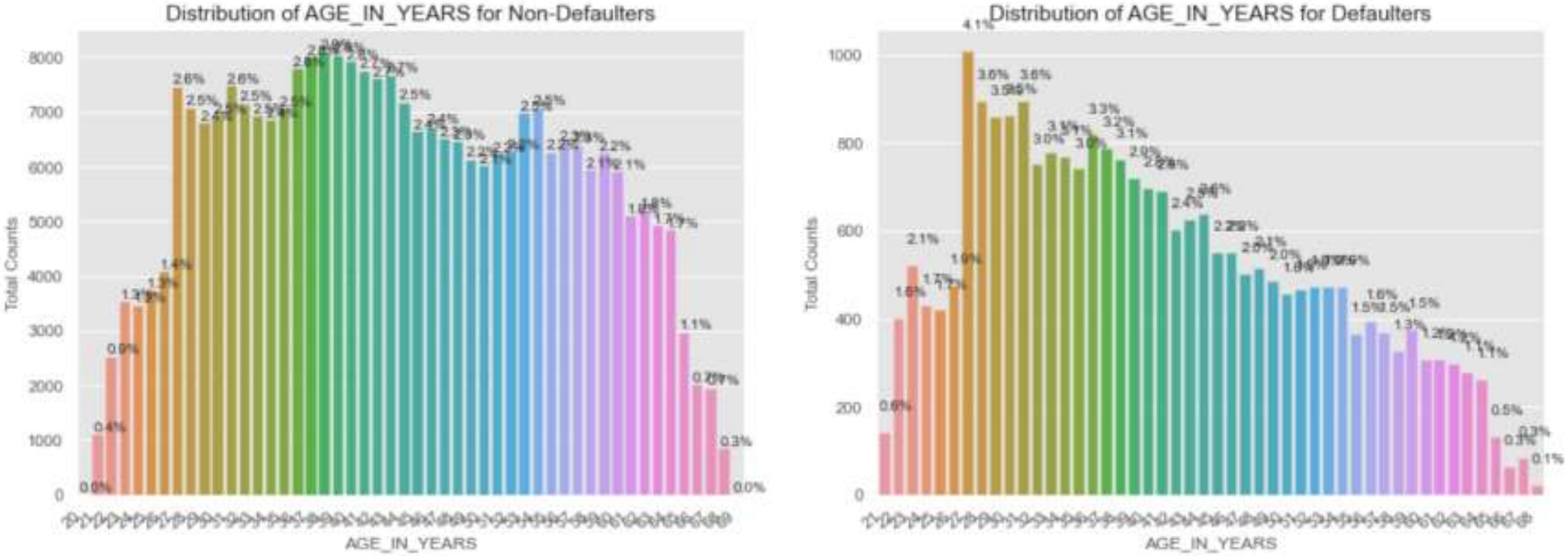
#Univariate Analysis

1. Distribution of AGE_IN_YEARS

Analysis, as per column of AGE_IN_YEARS, the weightage of non defaulters is much more b/w 28-60yrs.

It is observed that, risk is gradually decreasing as the increasing in age.

we can provide loan to people fall under 28 to 50, for ease of clearing loan.



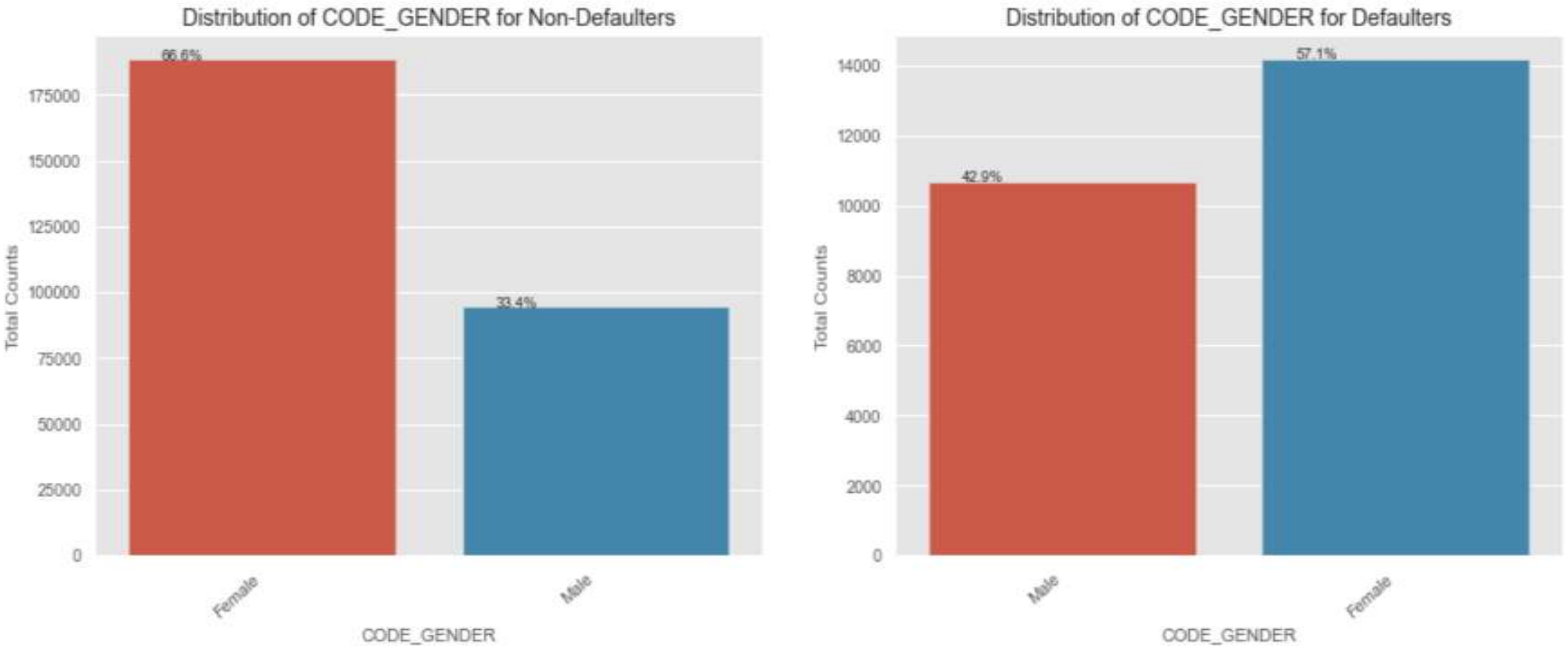
1. Distribution of AGE_IN_YEARS for Defaulters & Non-Defaulters

2. Distribution of CODE_GENDER for Defaulters & Non-Defaulters

Female are more contributing in Non-Defaulters that is 67% to the total count.

As checked with Male, 43% in male category are defaulters as they also seen in less percentage of non-defaulters.

For business purpose, continuity of payment in terms of female is more.



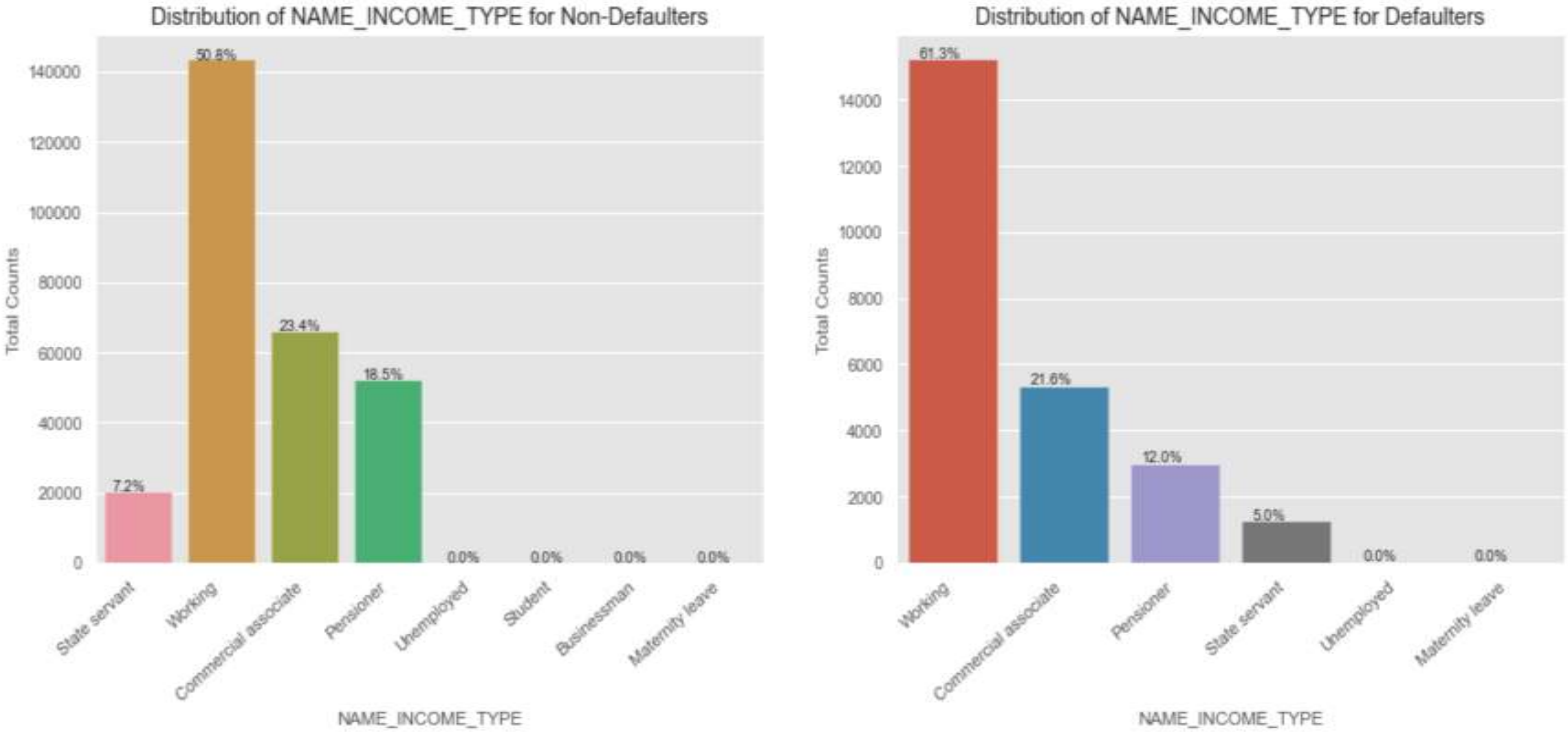
2. Distribution of CODE_GENDER for Defaulters & Non-Defaulters

# Distribution of NAME_INCOME_TYPE for Defaulters & Non-Defaulters

Insights from Name_Income_Type column, Working people category has more percentage of non-defaulters.

Students & Businessman falls under safe category of non-defaulters.

In business term & for companies revenue advantage, working people & commercial associate can help more.



Distribution of NAME_INCOME_TYPE for Non-Defaulters

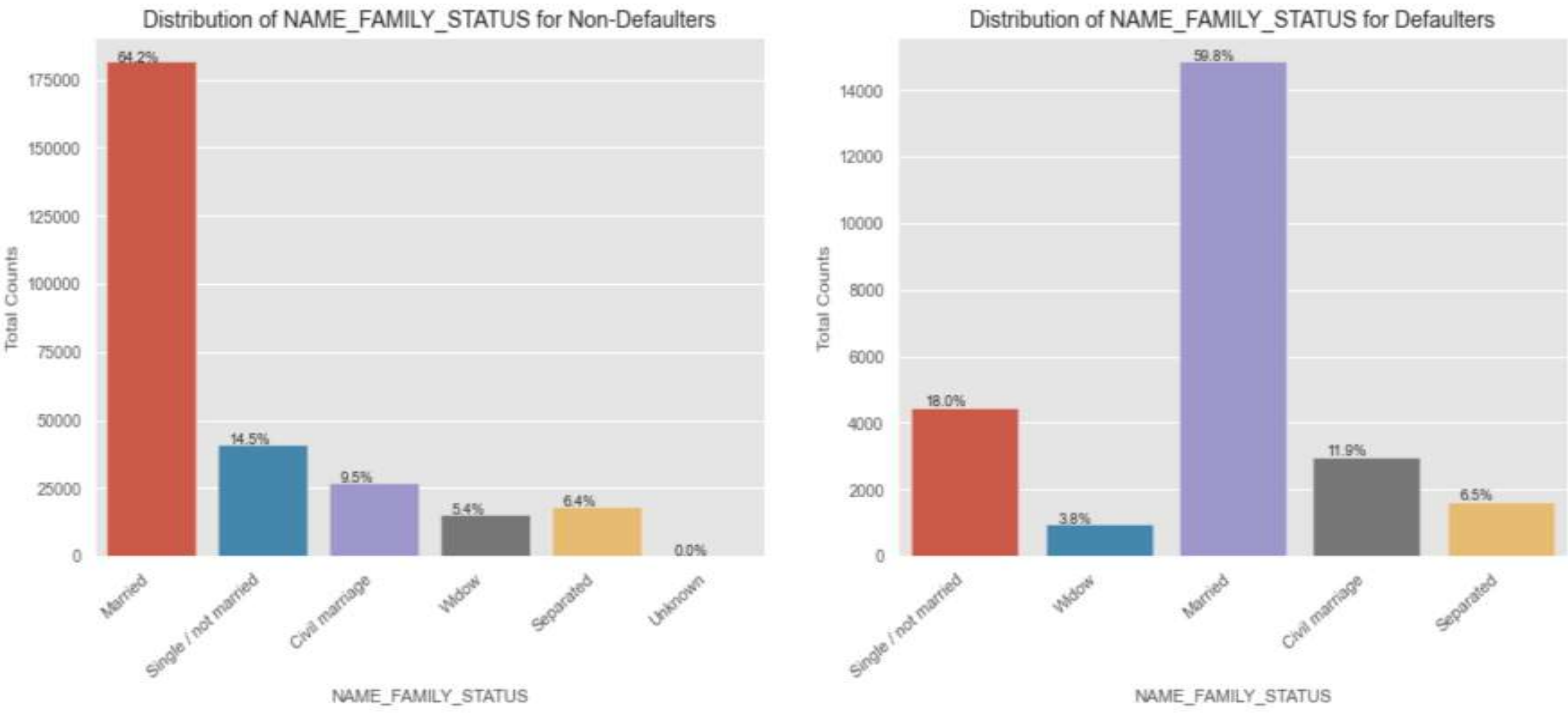
Distribution of NAME_INCOME_TYPE for Defaulters

3. Distribution of NAME_INCOME_TYPE for Defaulters & Non-Defaulters

# Distribution of NAME_FAMILY_STATUS for Defaulters & Non-Defaulters

After analysis over NAME_FAMILY_STATUS column, percentage of married people have more in terms of non-defaulters approx. 64% where in defaulting distribution married people are high.

But comparatively, we see that Single/not Married people contribute 14.5% to Non Defaulters and 18% to the defaulters of the application dataset.



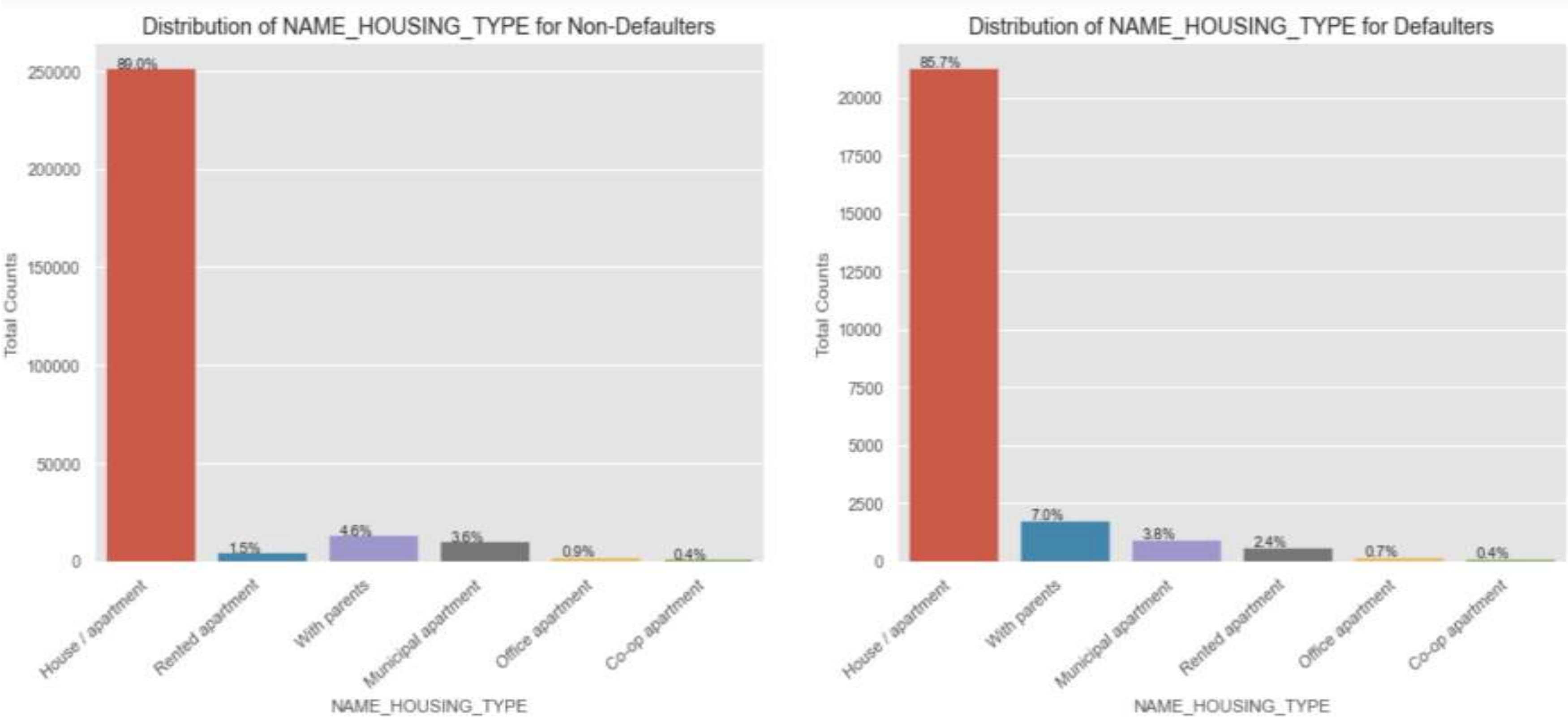
4. Distribution of NAME_FAMILY_STATUS for Defaulters & Non-Defaulters

# Distribution of NAME_HOUSING_TYPE for Defaulters & Non-Defaulters

Analysis over Name_Housing_type column, people who have House/Apartments, applies for more loan ie., 89% people from the dataset.

People living with parents, tries to default more often when comparing with other.

In business term, we can intact more people from owned House/Apartment category.



5. Distribution of NAME_HOUSING_TYPE for Defaulters & Non-Defaulters

#Routine Checkings for Previous_Application_Data (ds2).

Imported all the necessary libraries in Jupyter notebook(named EDA Assignment).

Read dataset (downloaded in folder by coping as path)

All the necessary things checked first, before making any changes like information, shape & null values to get the clarity of dataset.

Checked null value count & data type of columns.

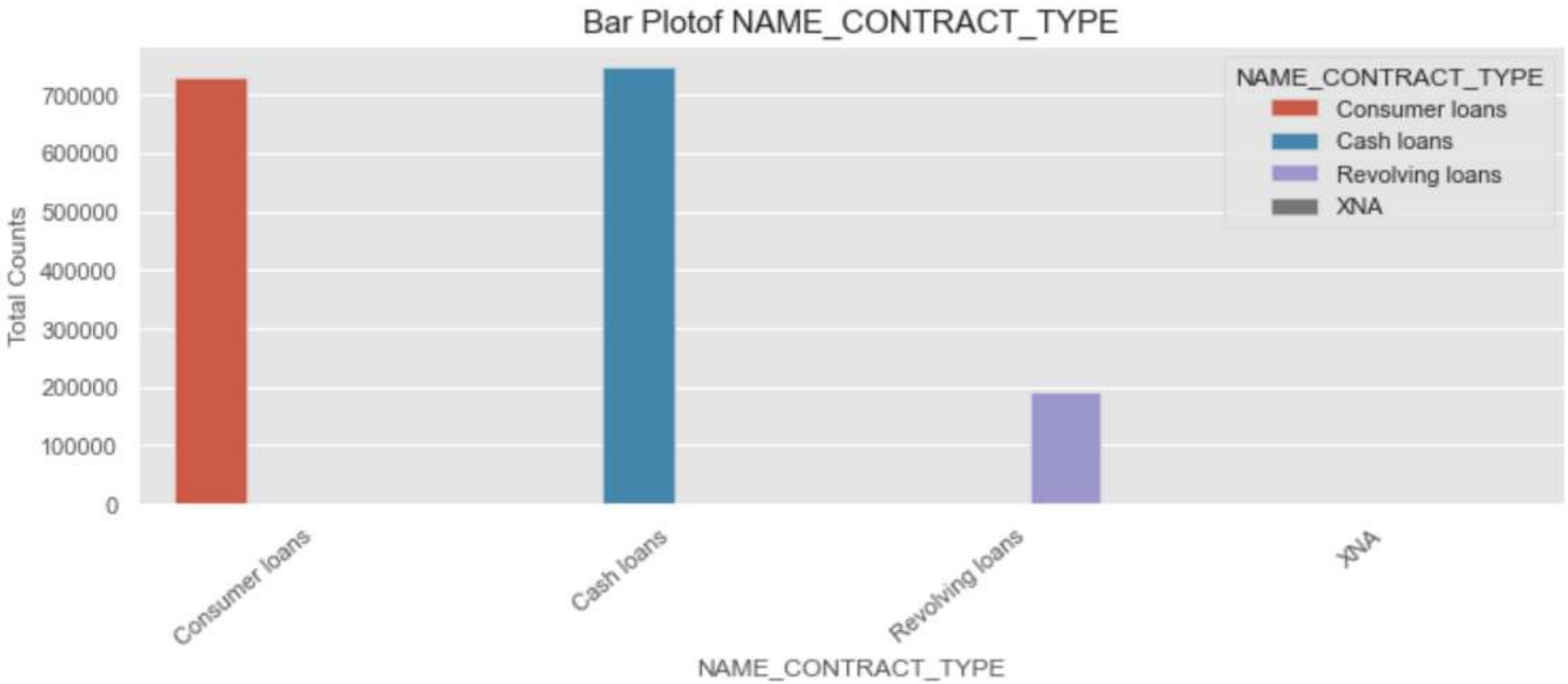
#Univariate Analysis

As per univariate analysis, done on previous_application_dataset, insights are as follows:

Cash Loans & Consumer Loan has more percentage than other as applicants prefer the same.

Revolving Loan percentage to the applications ratio is low.

As business purpose, we can provide more consumer loan & cash loan.



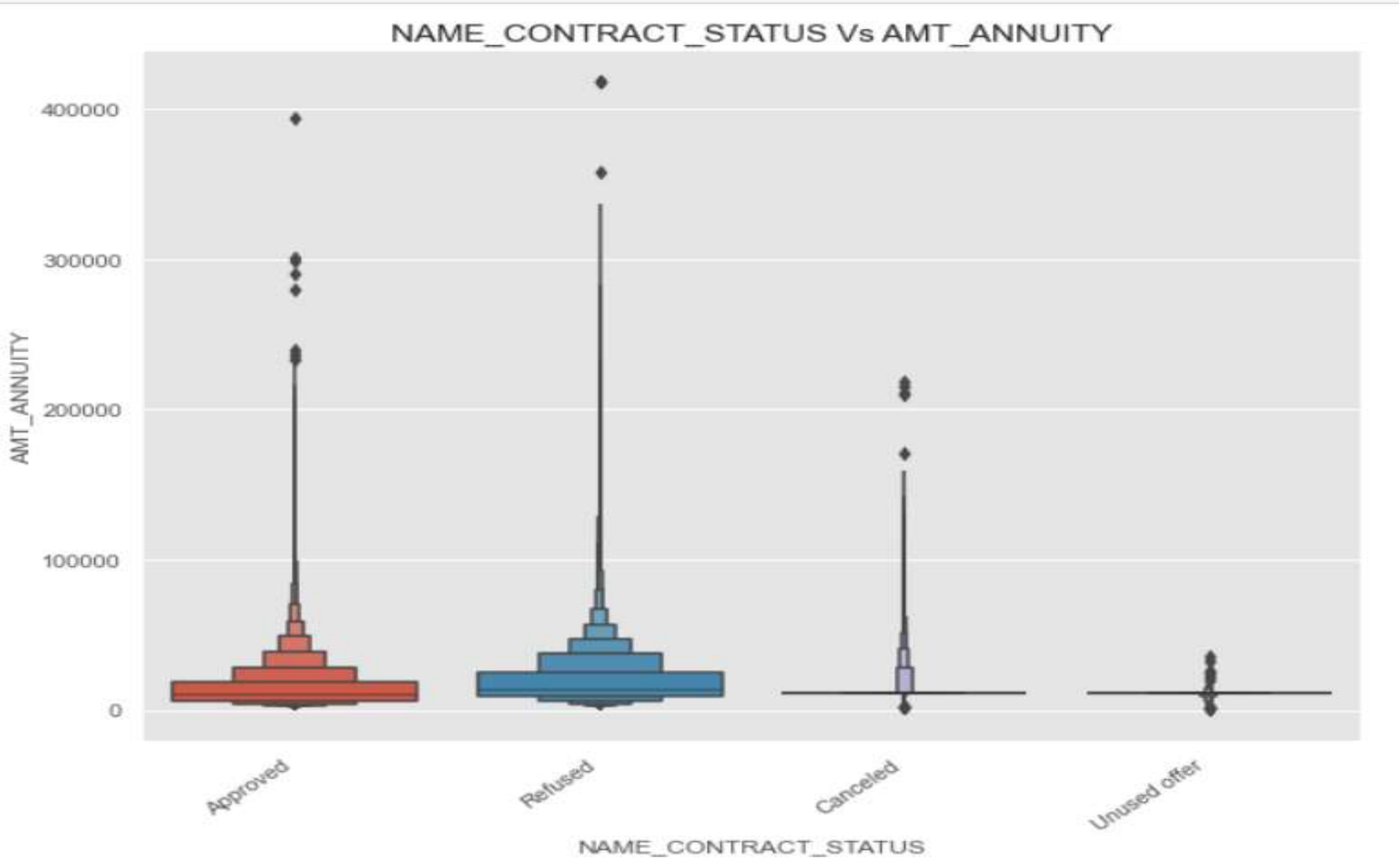
1. Bar Plot of Name_Contract_Type

#Bivariate Analysis

Bivariate analysis is combination of two different columns here taken Name_Contract_Status Vs Amt_annuity as follows:

Refused data is more in number to the AMT_ANNUITY.

Lot of applications are cancelled due to low AMT_ANNUITY.



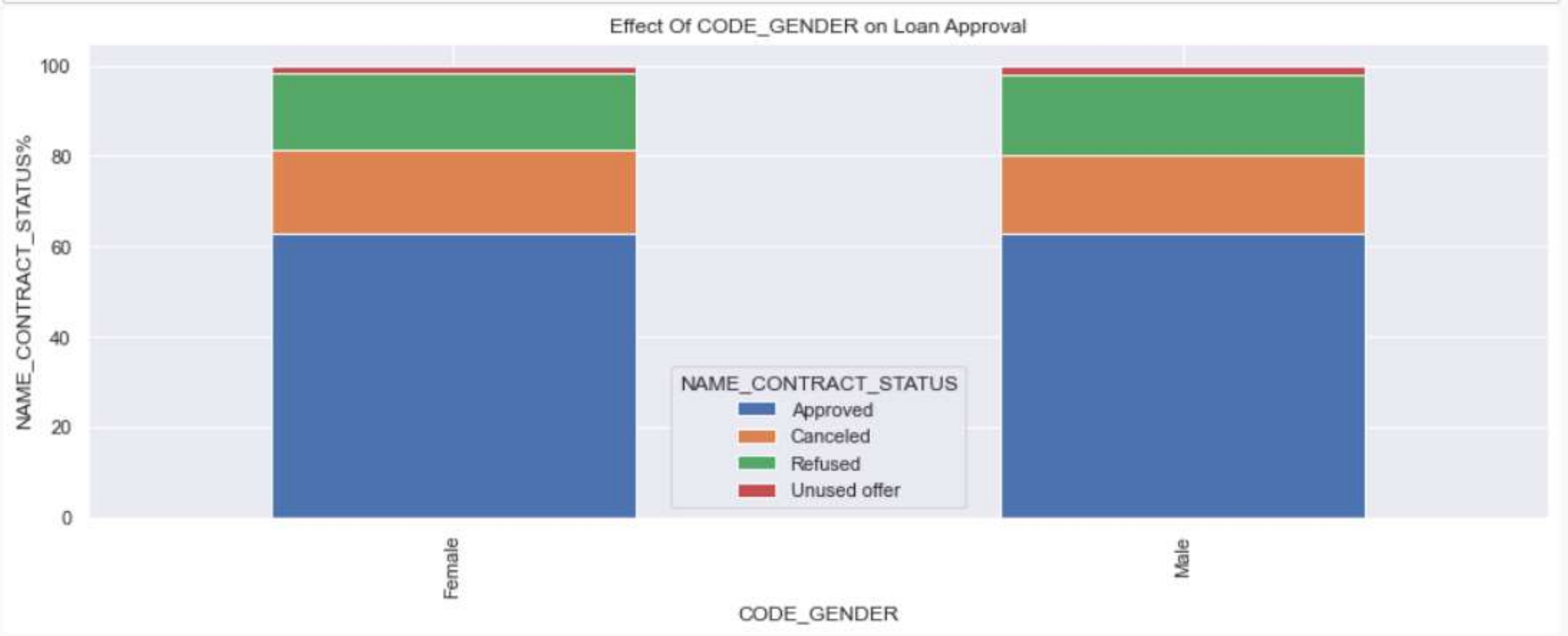
1. Bivariate analysis (Name_Contract_Status Vs AMT_Annuity)

#After Merging of both Dataset.

Insights from the merging dataset, code gender doesn't have any much effect on application rejection.

We saw earlier in univariate analysis, female have lesser chances to default than males

Bank can add more weightage to female for sectioning loan amount.



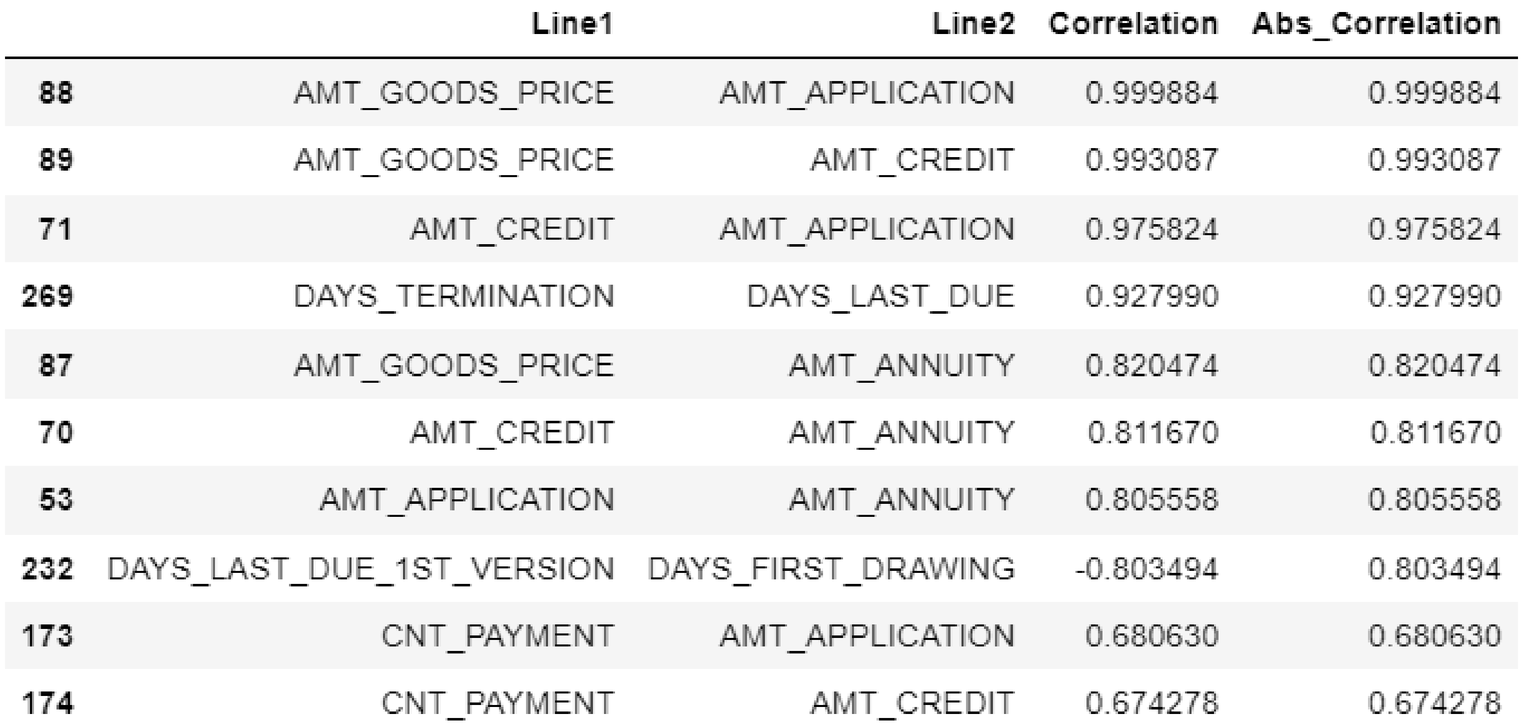
Merger of columns of both dataset (Name_Contract_Status Vs Code_Gender)

#Correlation on Previous Application Data

Cnt_Payment vs Amt_Appication is having more correlation value than Cnt_Payment vs Amt Credit.

Amt_Goods_price having correlation factor 0.9998 is more than factor 0.9930.

AMT_CREDIT in Line1 having correlation 0.975 is more than 0.811.



	Line1	Line2	Correlation	Abs_Correlation
88	AMT_GOODS_PRICE	AMT_APPLICATION	0.999884	0.999884
89	AMT_GOODS_PRICE	AMT_CREDIT	0.993087	0.993087
71	AMT_CREDIT	AMT_APPLICATION	0.975824	0.975824
269	DAYS_TERMINATION	DAYS_LAST_DUE	0.927990	0.927990
87	AMT_GOODS_PRICE	AMT_ANNUITY	0.820474	0.820474
70	AMT_CREDIT	AMT_ANNUITY	0.811670	0.811670
53	AMT_APPLICATION	AMT_ANNUITY	0.805558	0.805558
232	DAYS_LAST_DUE_1ST_VERSION	DAYS_FIRST_DRAWING	-0.803494	0.803494
173	CNT_PAYMENT	AMT_APPLICATION	0.680630	0.680630
174	CNT_PAYMENT	AMT_CREDIT	0.674278	0.674278

Correlation of Previous Application Dataset

THANK YOU.