EDA CASE STUDY

Credit Risk Analysis

By Nupur Agrawal

Introduction

This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Problem Statement

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to 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.

Business Objective

The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments 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 Understanding

There are dataset has 3 files as explained below:

1. 'application_data.csv' contains all the information of the client at the time of application.

The data is about whether a **client has payment difficulties.**

2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data on whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

3. 'columns_description.csv' is data dictionary which describes the meaning of the variables.

Steps Involved

1. Data understanding

2. Data Cleaning and Manipulation

3. Data analysis

4.Risks and Recommendations

Data Cleaning

1.It is done for preparing data for better data analysis

```
# check for null value
application.isnull().sum()
```

```
In [13]: ##Finding the percentage of missing values in all columns
         round(application.isnull().mean()*100,2).sort_values(ascending = False)
```

```
In [15]: #shape of application.shape
         application.shape

Out[15]: (307511, 81)
```

2. After dropping the columns, it got reduced to 81 columns

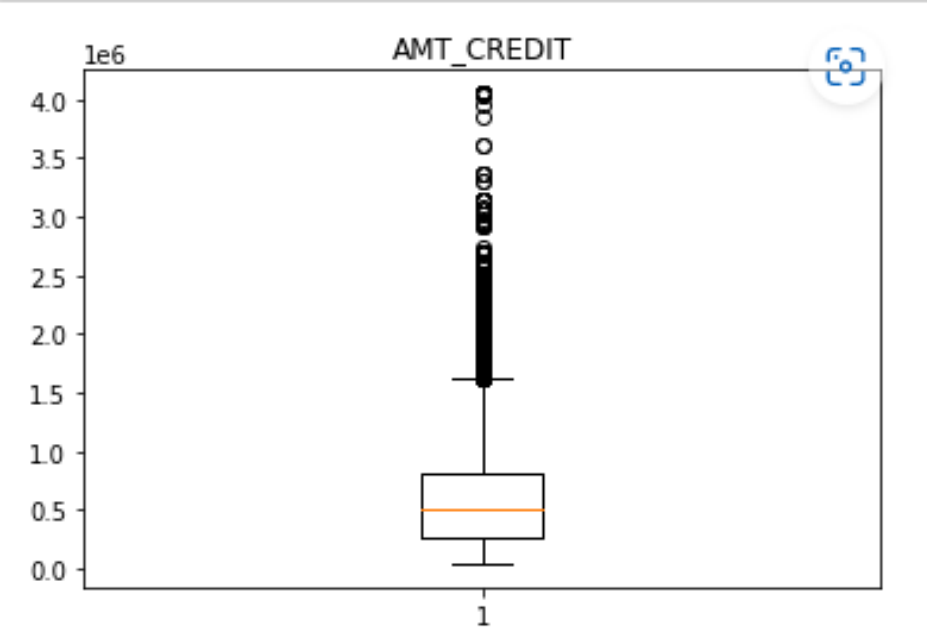
Distribution of AMT_Credit

OBSERVATION:

Here , This columns tell us the credit amount of the client.

The value is greater than 4.0 which is an outlier.

Calculated IQR and found- 1616625.0

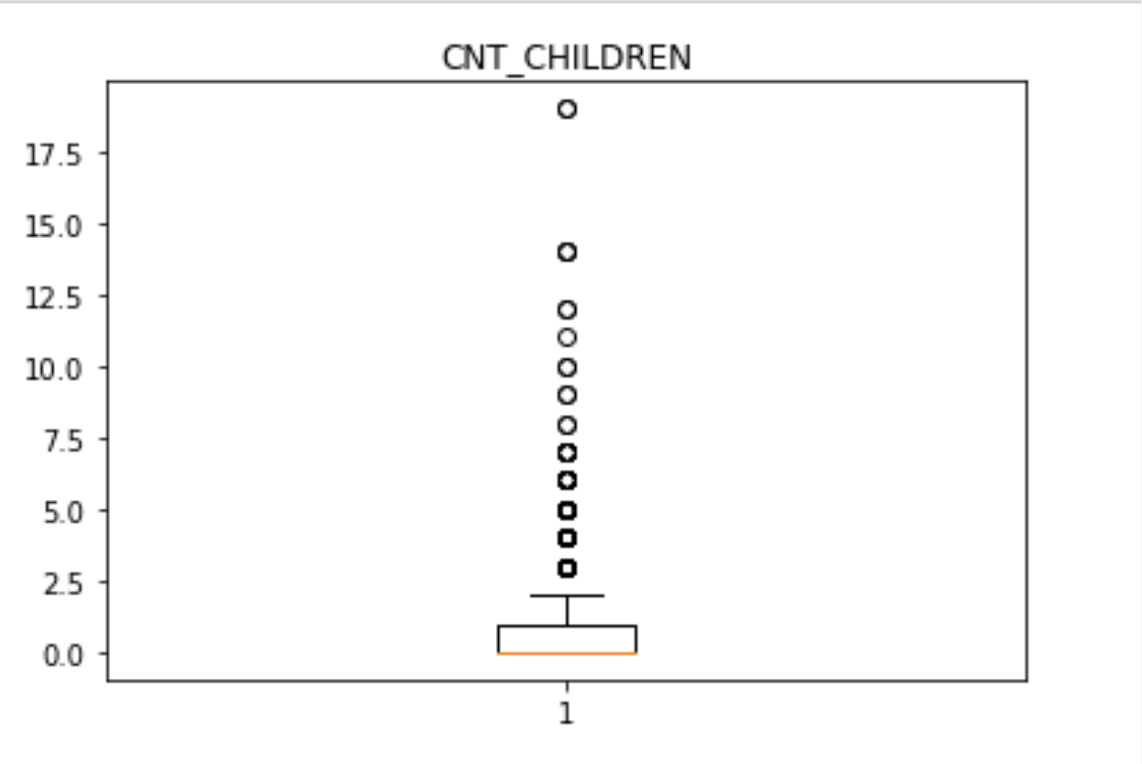


Distribution of CNT_CHILDREN

OBSERVATION:

There is one value 19 as per humans cannot have so many children hence we will also consider this is an outlier:

Here the graph shows that the values greater than 2.5 is considered as outliers ,since count of children cant be in decimal form so 3 is outlier here

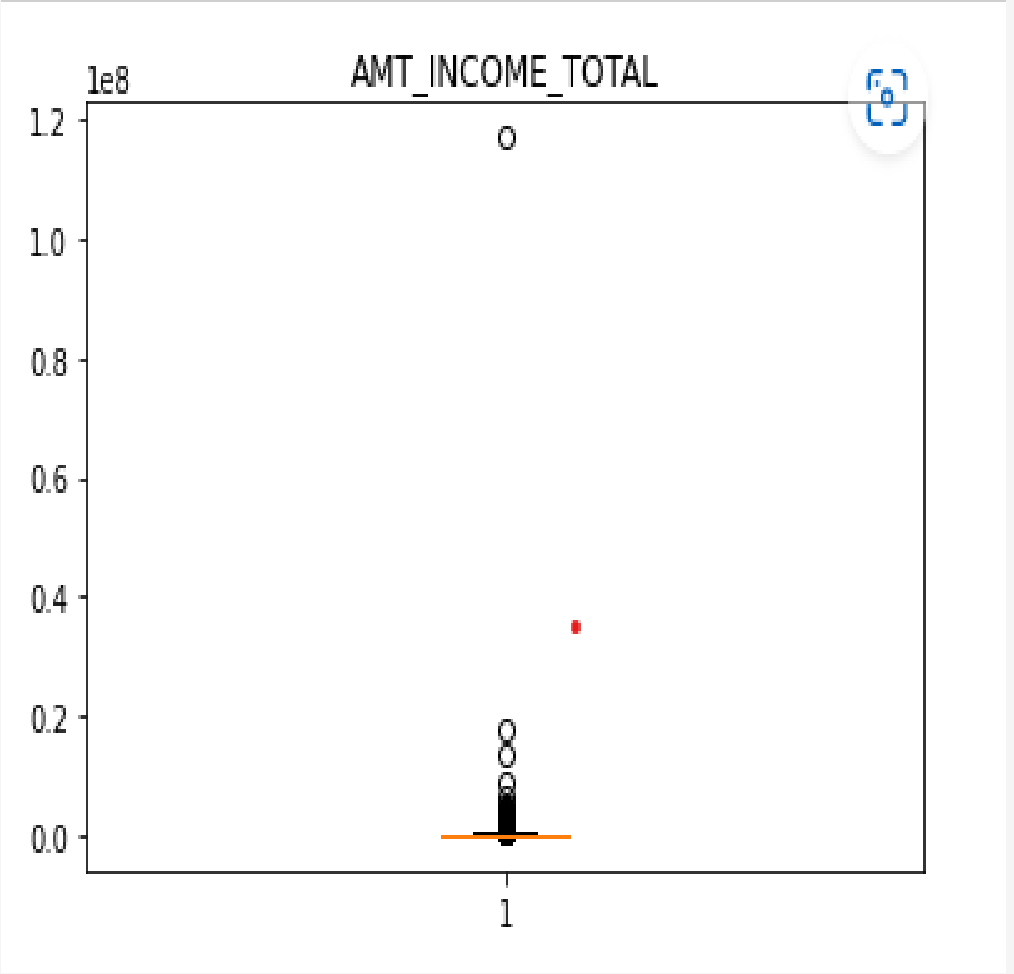


Distribution of AMT_INCOME_TOTAL

OBSERVATION:

Here , This columns tell us the income of the client.

We observe that the MAX amount is very larger the statistical mean which is mean (25,50,75) percentile



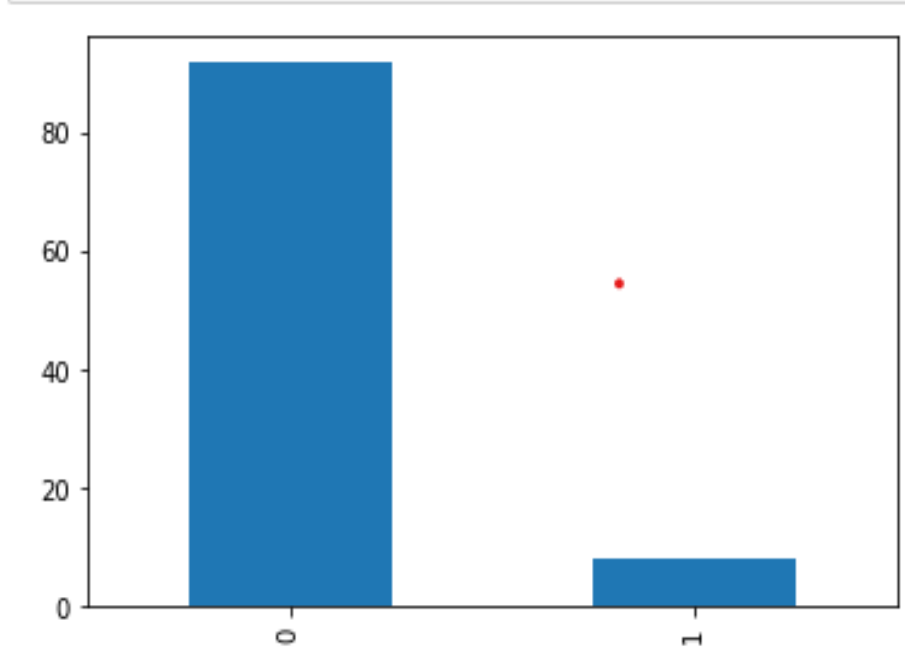
Checking for Imbalance for Target

OBSERVATION:

The figure shows the imbalance Target variable

when Target value 1 then it represents the defaulter with payment difficulties or late payment .This is 8.07% of t he data

Target value s 0 then it represents the non defaulter ,all the other cases apart from target 1 which is 91.93 of the data

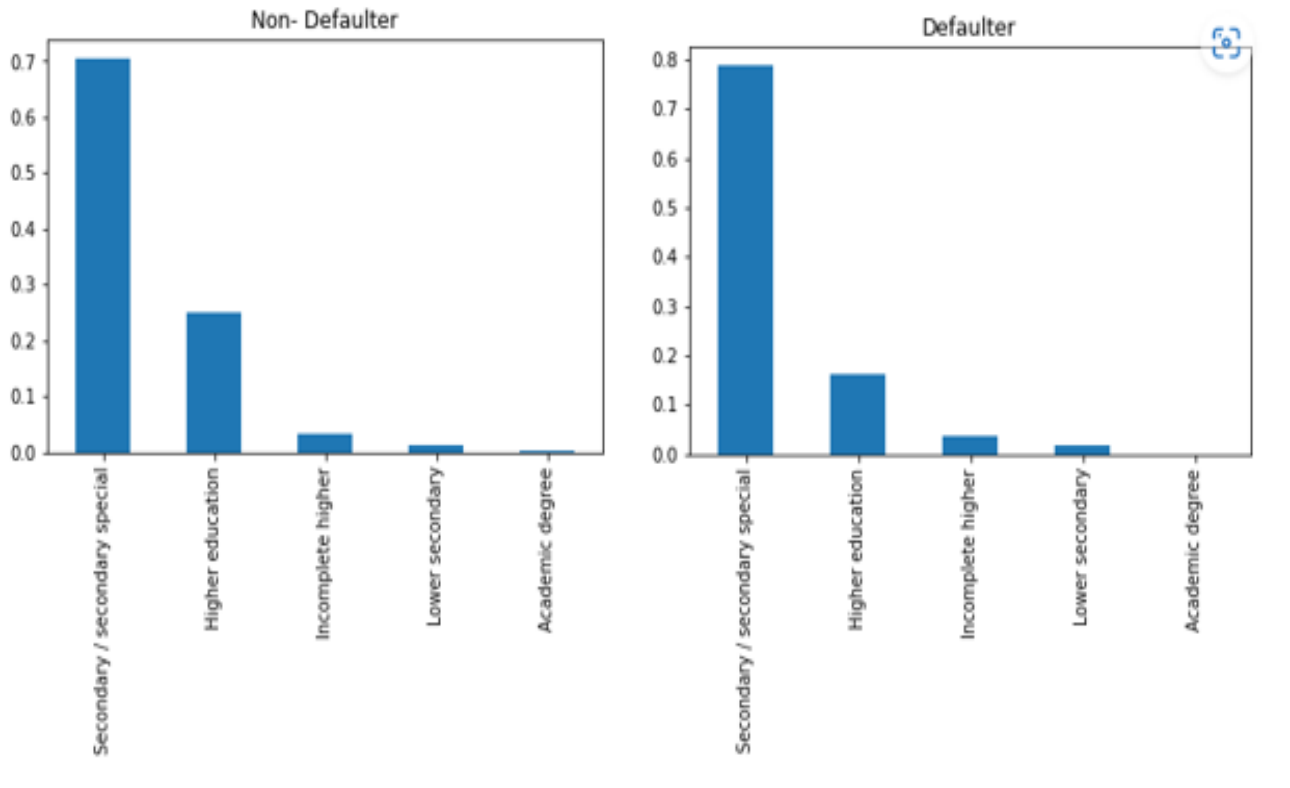



```
0    282686
1     24825
Name: TARGET, dtype: int64
```

UNIVARIATE ANALYSIS

Univariate Analysis for NAME_EDUCATION_TYPE

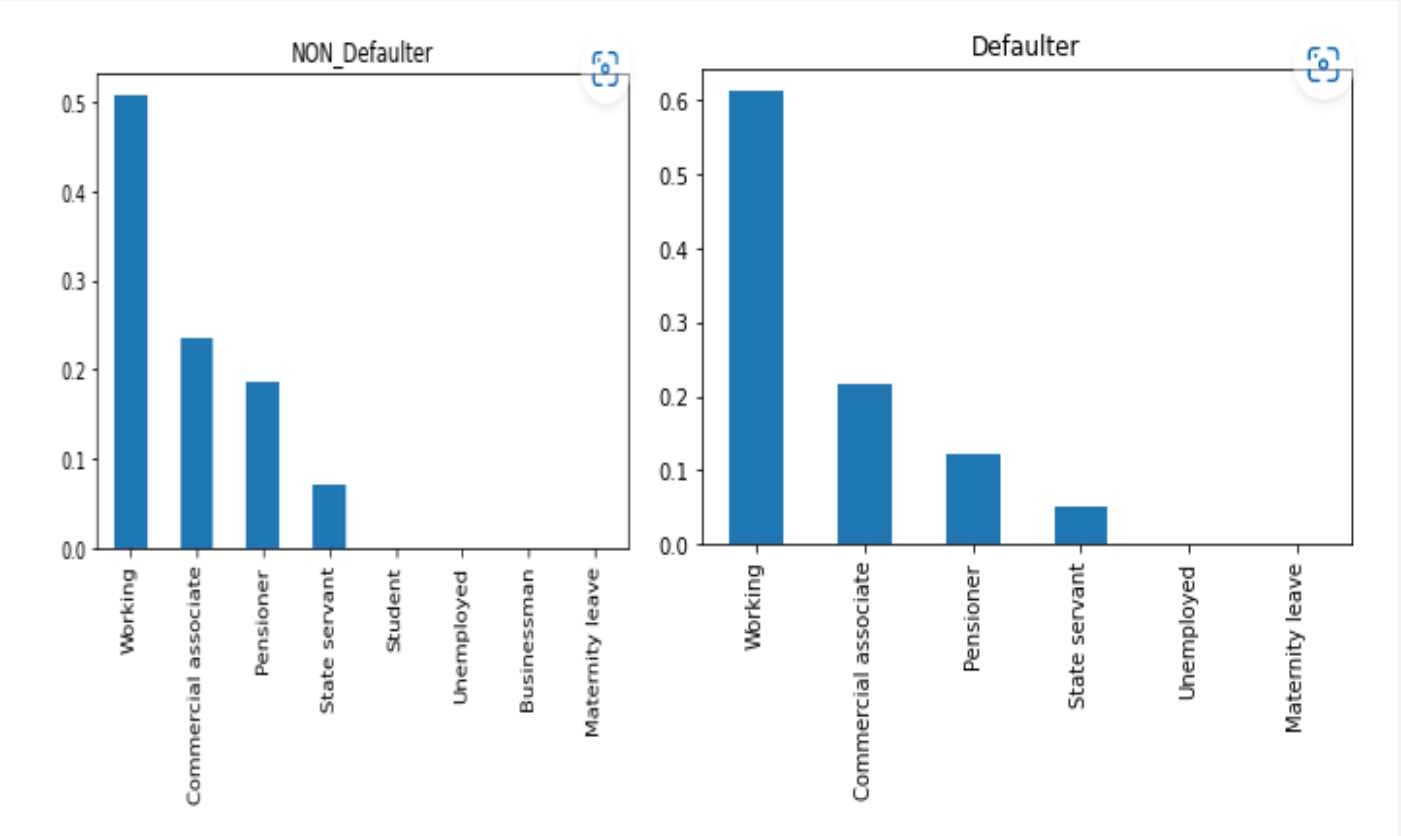
Observation : Here we can see that Academic degree people are applyingfor loan in min number for both target 0 and 1 whereas secondary/ secondary special educated people are applying loans high in number.



Univariate Analysis for NAME_INCOME_TYPE

Observation : Here we can see that the student, business man are non default whereas for bank there is good income from working people,

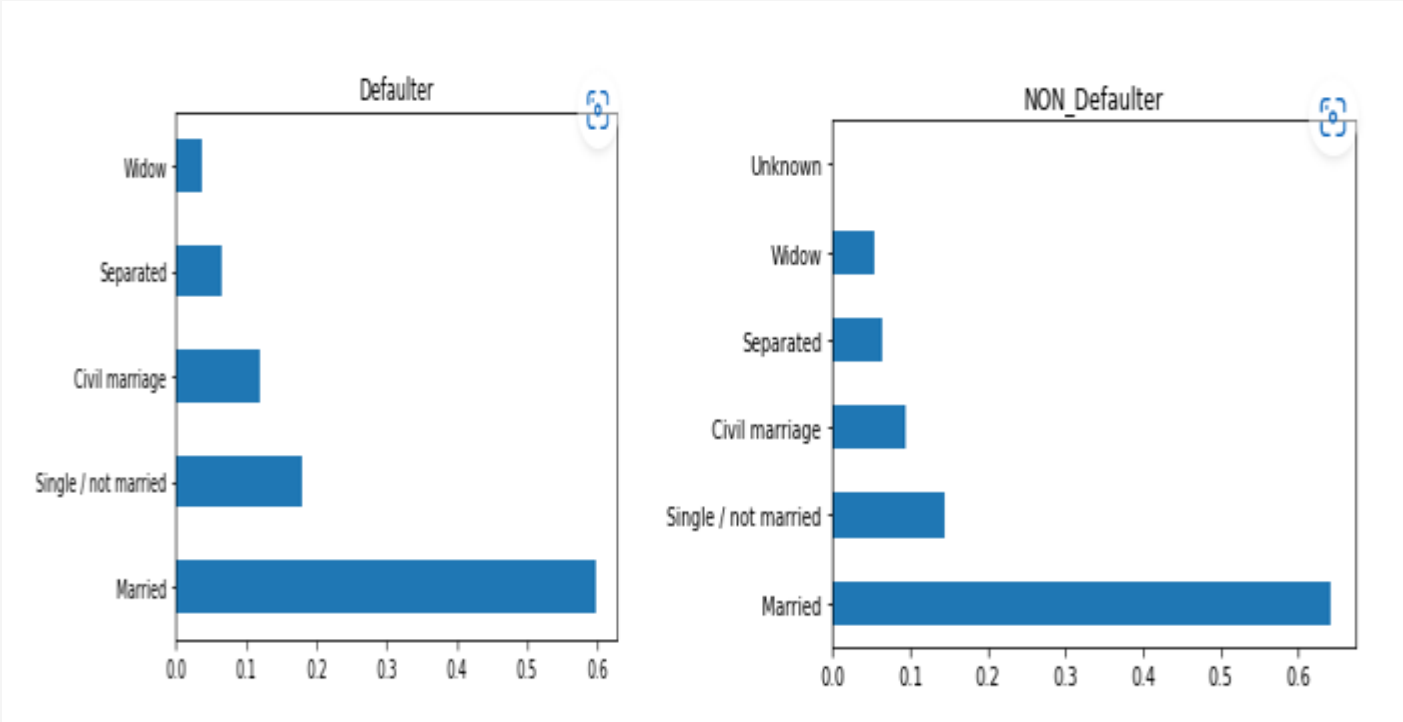
Also Business doesn't have payment Difficulties



Univariate Analysis for NAME_FAMILY_STATUS

Observation : Here we can see that the married people apply for loans mostly as its not affecting defaulter and non- defaulter.

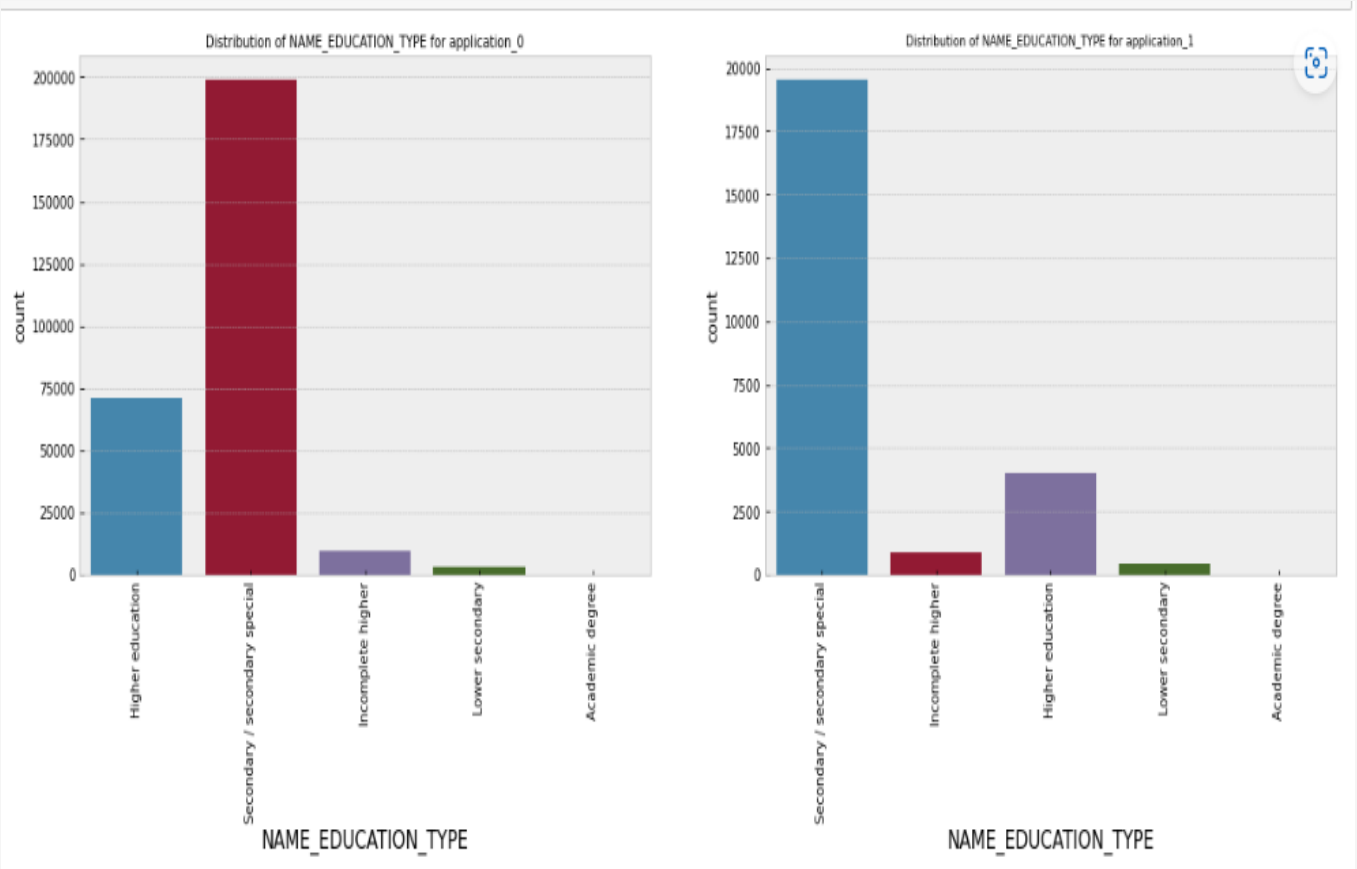
Clients who are single/not married have more difficulties with on time payments.



Analysis of NAME_EDUCATION_TYPE

Observation :

Here we can see that the higher the education the more the individual is capable of paying the loan

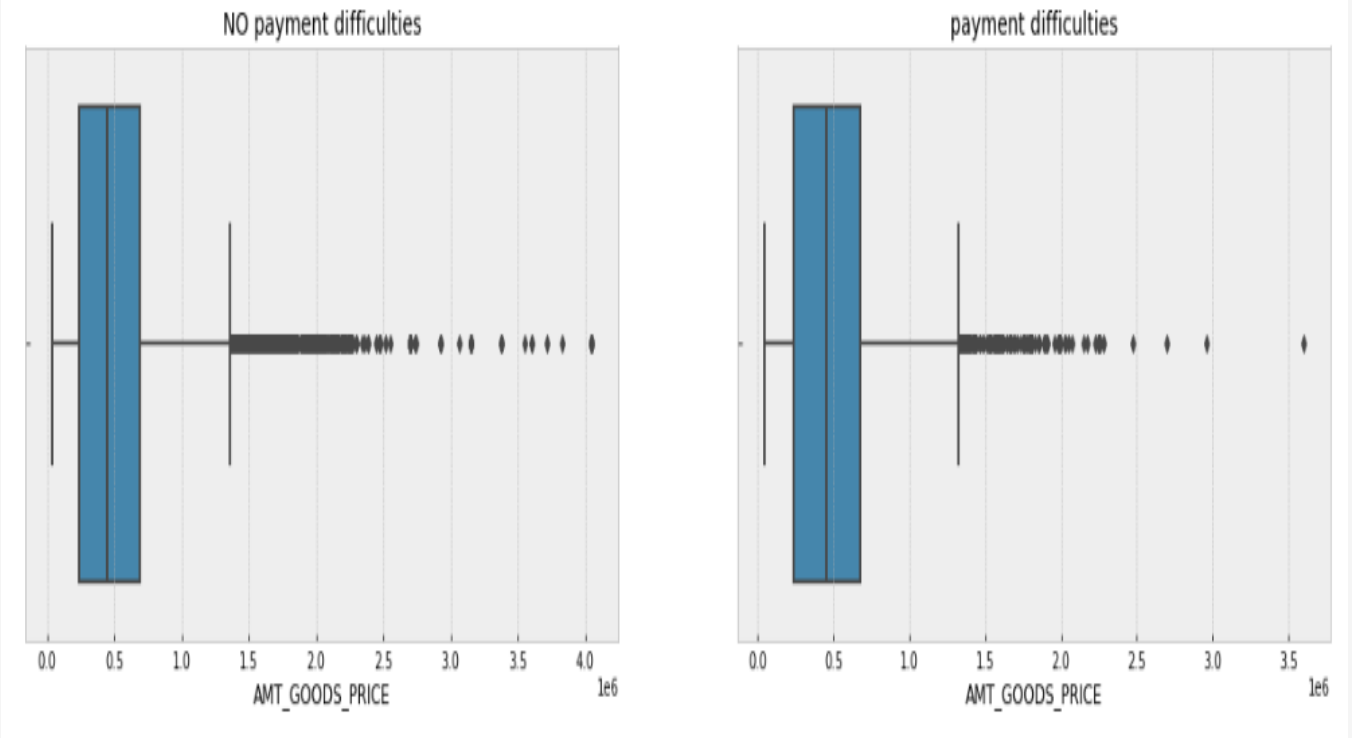


Univariate Analysis for Numerical Variable

Analysis for AMT_GOODS_PRICE

OBSERVARTION :

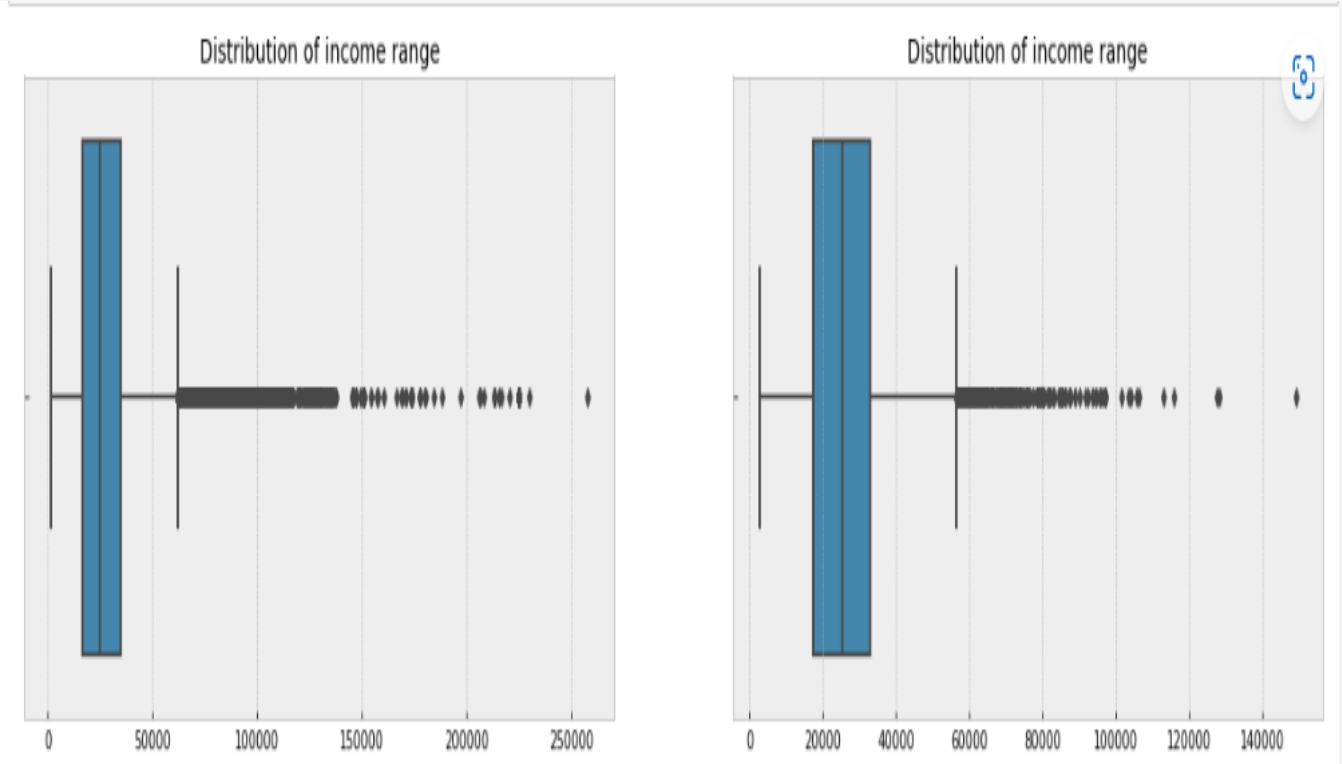
Here we can see that also both are having the mid value about 0.5 and customer with no payment difficulty lies between 0.3 to 0.7.



Analysis for AMT_Credit

Observation:

For AMT_CREDIT between 250000 and approximately 650000.There are more clients with payment difficulty

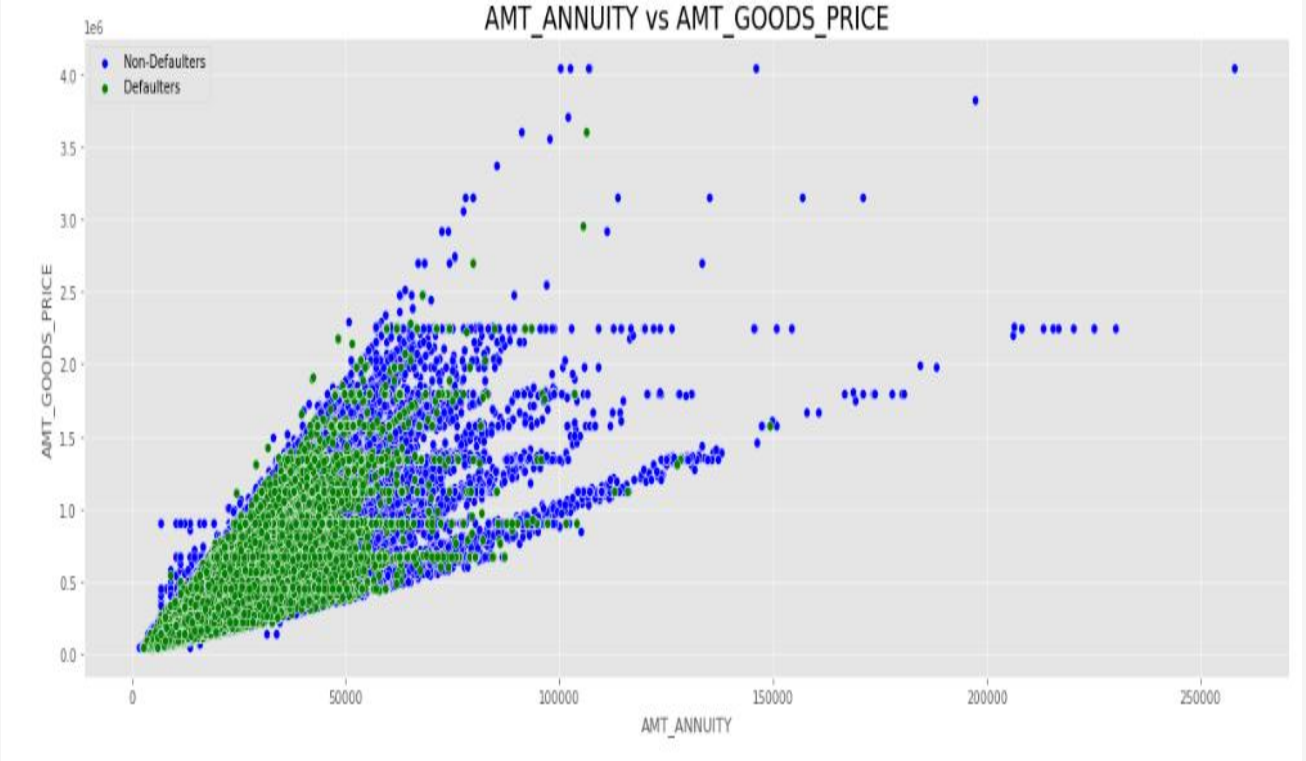


Bivariate Analysis

Analysis for ('AMT_ANNUITY','AMT_GOODS_PRICE')

Observation: here we can see that the strong correlation between AMT_ANNUITY,AMT_GOODS_PRICE- values above 70000 are tend to decrease in defaulters. Also we can see that graph moderates but they are not thoroughly correlated because there are above par values for both the columns.

This means Annuity increases so does the Goods price



Analysis for ('AMT_GOODS_PRICE','AMT_CREDIT')

Observation: the graph shows the Low, Median, high values of defaulters and non defaulter the higher credit value is for those who pay on time and is non defaulter.

They have strong Correlation which means the Goods price increase so does the credit

Correlation

Observation :

For Non-Defaulters Decision is taken by FLOORSMAX_MEDI,YEARS_BEGINEXPLUATATION_MEDI,FLOORSMAX_MODE,TOTALAREA_MODE

Observation Defaulter Decision is taken by-OBS_60_CNT_SOCIAL_CIRCLE,YEARS_BEGINEXPLUATATION_MEDI YEARS_BEGINEXPLUATATION_MEDI,FLOORSMAX_MODE

	col 1	col 2	Correlation	Corr_Abs
970	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00	0.999991
788	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	1.00	0.999988
824	FLOORSMAX_MEDI	FLOORSMAX_AVG	1.00	0.999985
826	FLOORSMAX_MEDI	FLOORSMAX_MODE	1.00	0.999916
754	FLOORSMAX_MODE	FLOORSMAX_AVG	1.00	0.999872
790	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	1.00	0.999587
718	YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	1.00	0.999578
142	AMT_GOODS_PRICE	AMT_CREDIT	1.00	0.999371
385	CNT_FAM_MEMBERS	CNT_CHILDREN	0.99	0.990113
1006	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.99	0.988079

Top 10 Correlations from target_1 : Loan Defaulter dataframe

OBSERVATION

The loan annuity correlation with credit amount and also with goods price has slightly reduced in defaulters(0.748) when compared to repayers(0.777)

Days birth and number of children correlation has reduced to 0.256 in defaulters when compared to 0.336 in repayers.

The correlation is strong between family member and children counts, although the correlation increases for the defaulters.

	Var1	Var2	Correlation
0	AMT_GOODS_PRICE	AMT_CREDIT	0.982
1	CNT_FAM_MEMBERS	CNT_CHILDREN	0.894
2	AMT_CREDIT	AMT_ANNUITY	0.749
3	REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	0.443
4	DAYS_BIRTH	DAYS_EMPLOYED	0.307
5	REGION_RATING_CLIENT	HOUR_APPR_PROCESS_START	0.290
6	DAYS_REGISTRATION	DAYS_BIRTH	0.241
7	AMT_CREDIT	DAYS_BIRTH	0.190
8	AMT_GOODS_PRICE	DAYS_BIRTH	0.185
9	CNT_CHILDREN	DAYS_BIRTH	0.177

Merging Dataset application and prev_application

```
merged = pd.merge(left=application  ,right=prev_application,how='inner',on='SK_ID_CURR',suffixes='_a')
```

```
merged.head()
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE_	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CRE
0	100002	1	Cash loans	M	N	Y	0	202500	4(
1	100003	0	Cash loans	F	N	N	0	270000	12(
2	100003	0	Cash loans	F	N	N	0	270000	12(
3	100003	0	Cash loans	F	N	N	0	270000	12(
4	100004	0	Revolving loans	M	Y	Y	0	67500	1(

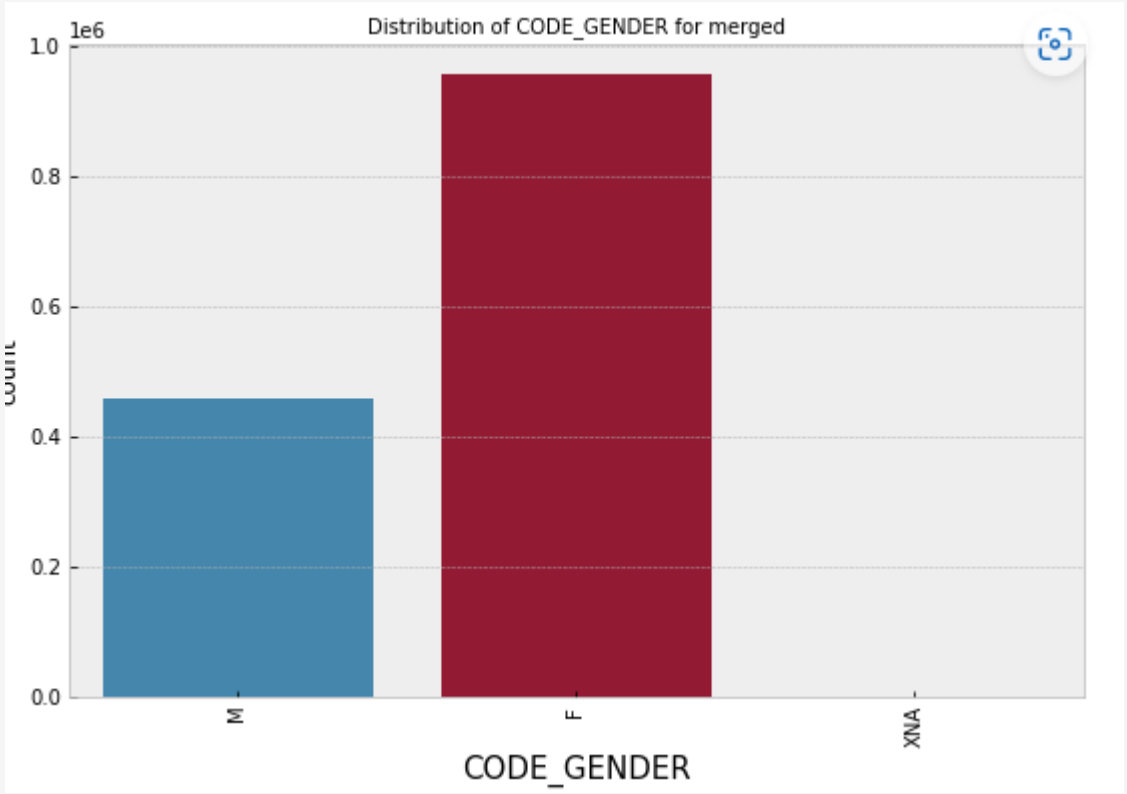
5 rows × 117 columns

Merged 2 data set .i.e, application and Previous application dataset and joined through inner query

Analysis on Merged dataset.

Analysis on CODE_GENDER

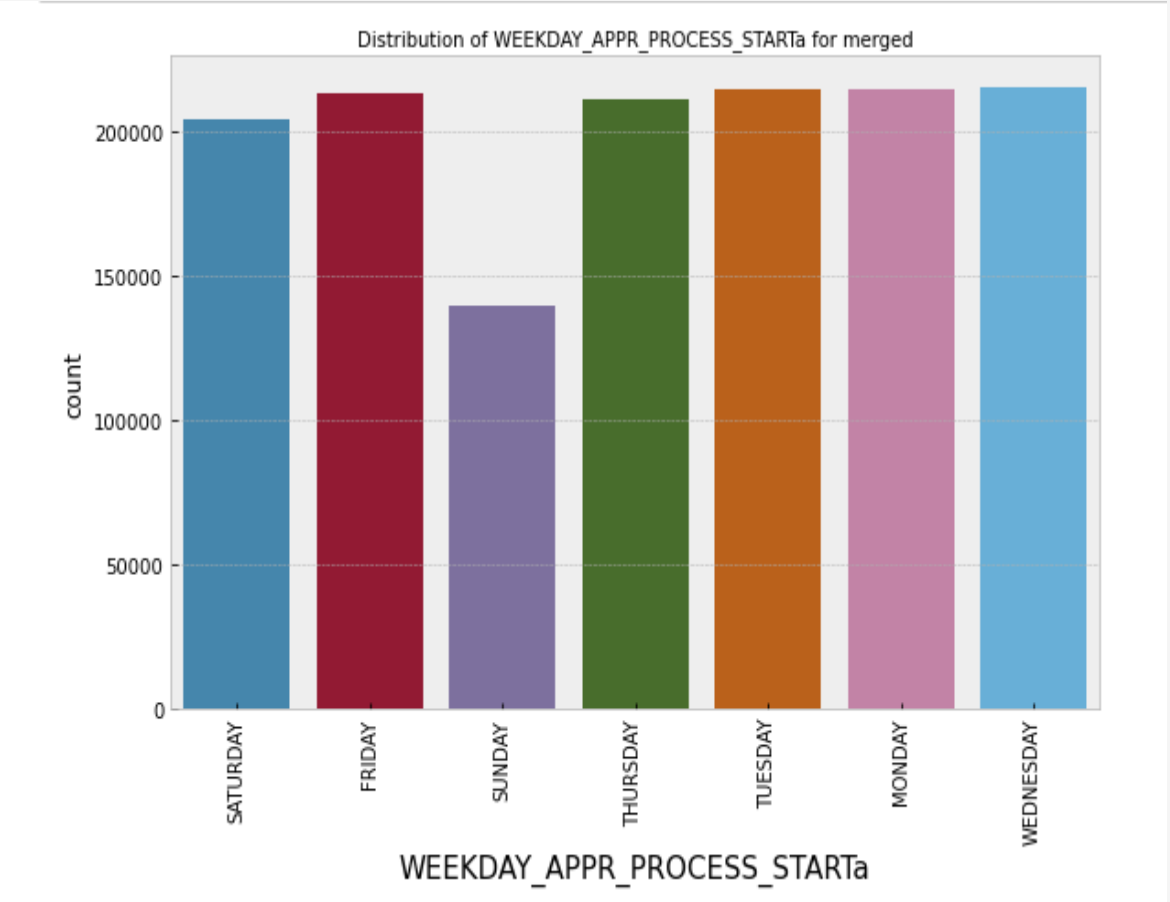
Observation: we can see that female have higher no of approval loans



Analysis on Merged dataset.

Analysis on WEEKDAY_APPR_PROCESS_STARTa

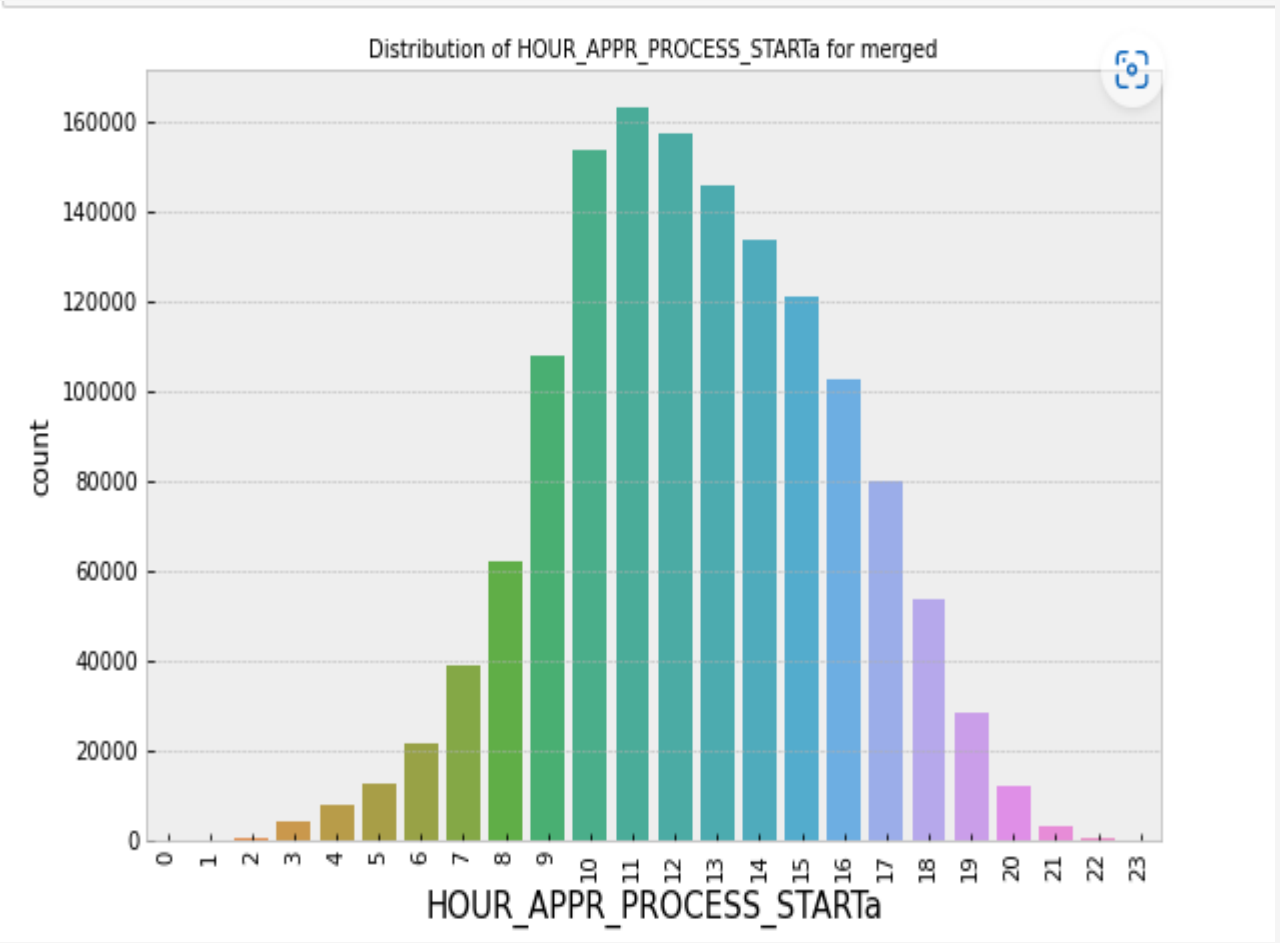
Observation: we can see that female have higher no of approval loans



Analysis on Merged dataset.

Analysis on HOUR_APPR_PROCESS_STARTa

Observation: the maximum hour to process the approval is 11 hour



Conclusion & Recommendations

1. To minimize the risk of loss, variables like- AMT_ANNUITY,DAYS_BIRTH,CODE_GENDER,NAME_HOUSING_TYPE,NAME_EDUCATION_TYPE should be considered

2. 2. Most number of unsuccessful payments are done from loan purpose 'Repair’.

3. Recommended group where loan can be credited-

   - Old group of any income group
   - Female client
   - Clients with high income category

4.Single people default more so it safe to give to married people

5. The people who have House/Apartment, tend to apply for more loan and People living with parents tend to be default