Final Submission of

:Credit Exploratory Data Analysis: Case Study – Assignment March - 2022

SUBMITTED BY:

SOURABH S HUBBALLI

DATED:28/03/2022

Business Understanding

- 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.

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 applicant's using EDA is the aim of this case study.

What are the Datasets provided for Analysis?

- *There are Two major Datasets provided which are mentioned below:*
    1. *Application Data.*
    2. *Previous Application Data.*

- *These file provided in "Comma Separated Values" Format. (.csv).*
- *Another file provided with Column Descriptions for defining and understanding each columns contribution.*

- *Prerequisites:*
    1. *Programming Language: Python.*
    2. *Platform: Jupyter Notebook.*
    3. *Libraries: Pandas, Numpy, Matplotlib, Seaborn, itertools and some Warnings*

Assigning Variables:

Description	***Assigned Variable***
Data Set - 1: "Application_data.csv"	*ap_dt*
Data Set - 2: "previous_application.csv"	*pr_ap_dt*
For Null values defined as	*nulls*
To store Null Total Values	*mis_val*
Null Values in ap_dt> 50%	*nul_50*
Null Values in ap_dt > 15%	*nul_15*
Storing Relevant Values	*nrel*
For Columns Flag	*Col_flag*
To store all flag columns and Target columns	*dt_flg*

Note: There are many other Variables used in data analysis process those variables are mentioned in Jupyter Notebook

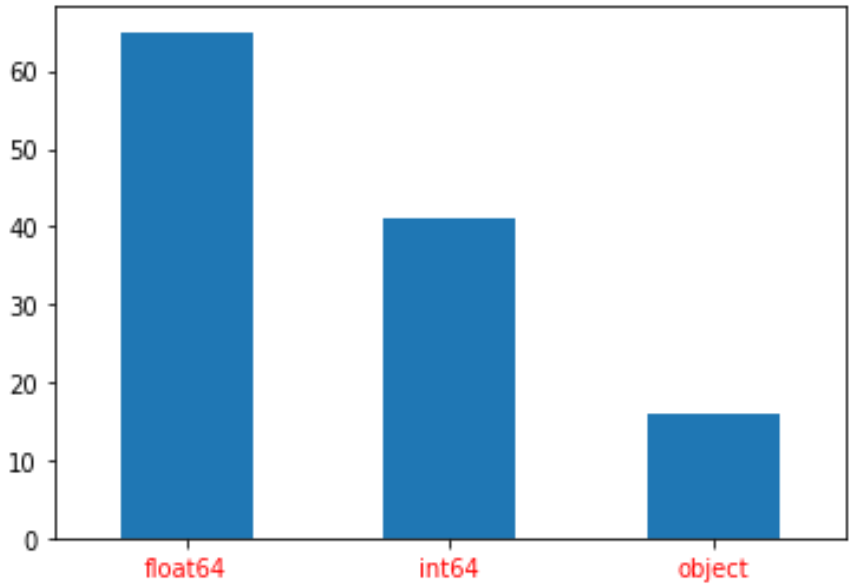
Data Understanding:

Required Details mentioned below:

1. Application_data.csv [ap_dt]

- *Number of Columns: 122*
- *Number of Rows: 307511*
- *Data Types: Integer, Float, Strings*
- *Descriptive view of Data file: There were anomalies like negative numbers, Null values, Days and Years were not in proper Format*

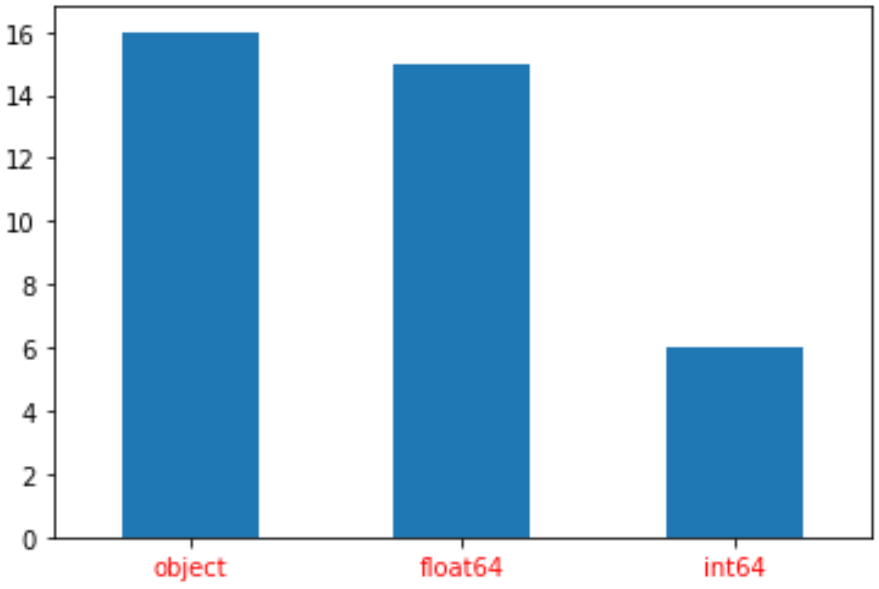
- *Float64: 65*
- *Int64: 41*
- *Object: 16*



2. Previous_Application_data.csv [pr_ap_dt]

1. *Number of Columns: 37*
2. *Number of Rows: 1670214*
3. *Data Types: Integer, Float, Strings*
4. *Descriptive view of Data file: There were anomalies like negative numbers, Null values, Days and Years were not in proper Format*

Float64: 15
Int64: 06
Object: 16



Data Cleaning & Manipulation for Application Data:

How we did that?

- *Rectify the null values.*
- *Filtering unwanted data columns.*
- *Filling the missing values.*
- *Sorting the data.*
- *Fixing the datatype*

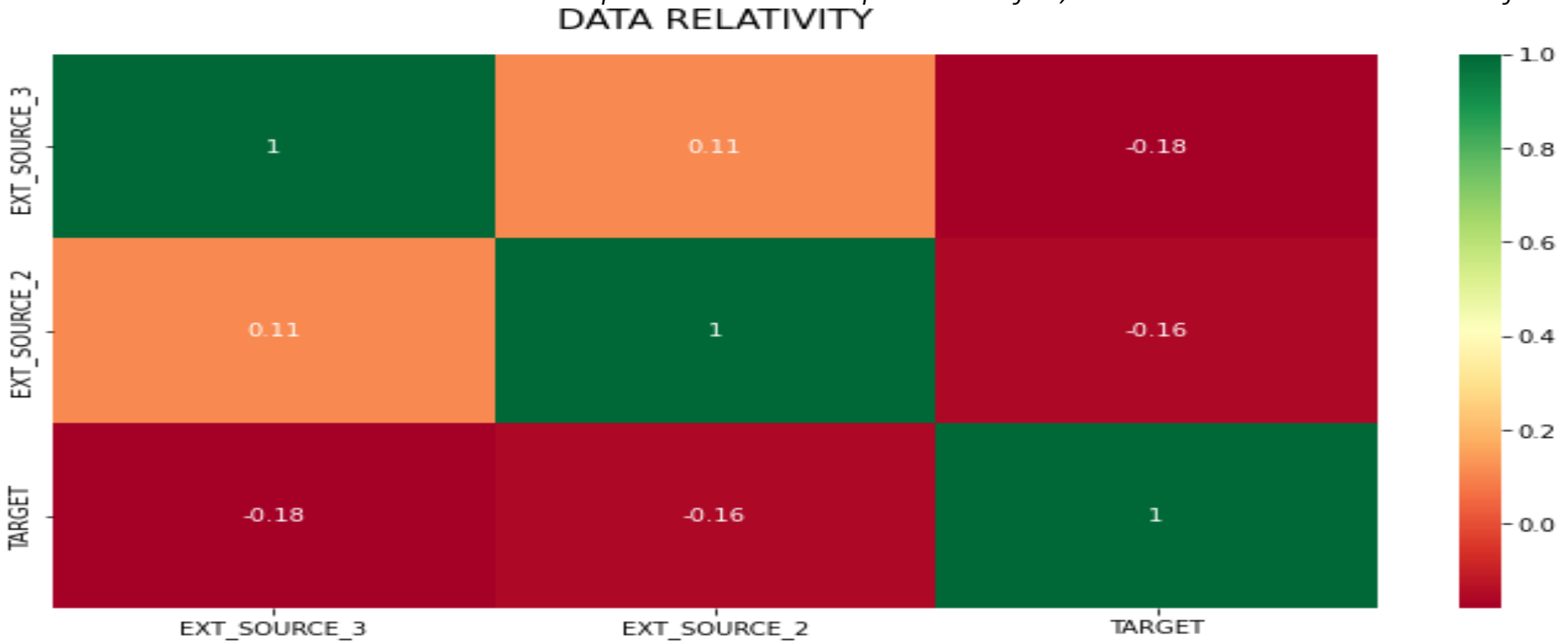
To Remove unwanted or irrelevant columns,

- *First, I have calculated null values "**nulls(ap_dt)"***
- *Then **calculated the values in term of percentage**.*
- *Found that there were above 41 columns which consists more than **50% Null values**.*
- *By comparing the columns contribution with given csv file (Columns_Description.csv), I have removed the irrelevant columns.*
- *Similarly, After removing the 50% data there were 10 Columns which are more than **15% Null Values**.*

Data Cleaning & Manipulation for Application Data:

Correlation & Causation:

- *After double checking the 15% Null Values, There was out-sourced data columns which are provided by externally.*
- *Source Columns:* **EXT_SOURCE_2** & **EXT_SOURCE_2.**
- *What is the relation between these 2 values? As per column description datafile, These are normalized values from external data source.*



Data Cleaning & Manipulation for Application Data:

Analysing EXT_SOURCE_2 & EXT_SOURCE_2 Columns, Flag Columns & Target Columns

- *By Above mentioned Correlation Heatmap, We found that There is no relation and much contribution*
- *These data doesn't cause causation.*
- *So, on this base I have Removed the* **EXT_SOURCE_2** & **EXT_SOURCE_2** columns.

- *After Removing All these columns, we left with* ***71 Columns****.*

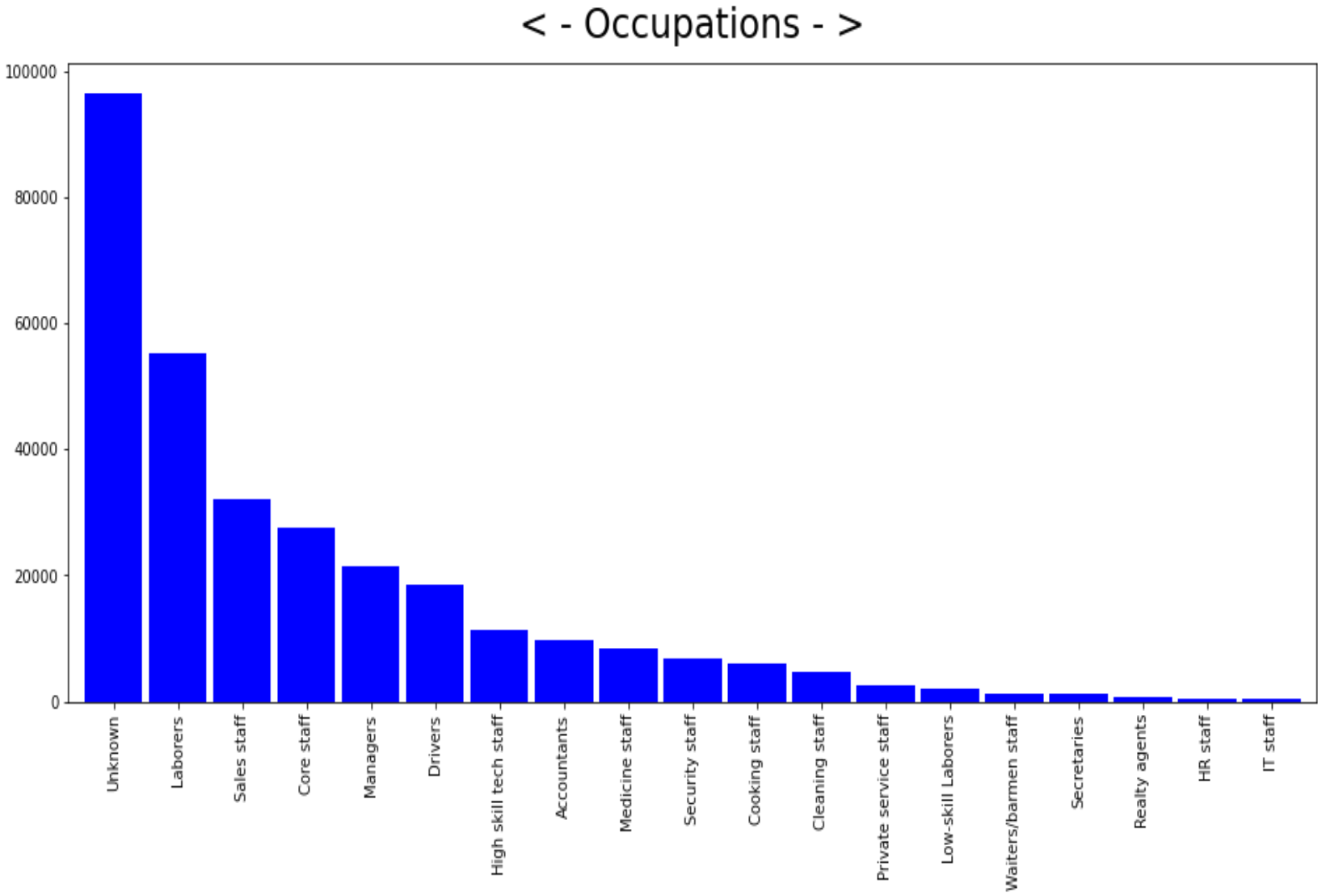
- *This 71 Columns includes* ***28 Flag*** *Columns:*
  - *In which there are Email, phone, Car, work and other important data were stored.*
  - *To analyse the Flag data, I have combined all the flag columns in one variable* ***"col_flag"*** *.*
  - ***Includes "Target"*** *Variable, Which has Explains (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample, 0 - all other cases)*
  - *For analysis we need to find Payers & Defaulters, for that I have changed data from* ***1's & 0's*** *to* ***"Defaulter"*** *and* ***"Repayer"*** *.*

Analyzing Flag columns & Target column:

Bar Graph Analysis:

- *By Observing the Graph:*
- *defaulters:*
  - *(FLAG_OWN_REALTY,*
  - *FLAG_MOBIL ,*
  - *FLAG_EMP_PHONE,*
  - *FLAG_CONT_MOBILE,*
  - *FLAG_DOCUMENT_3*
- *These columns make relativity thus we can include these below columns:*
  - *FLAG_DOCUMENT_3,*
  - *FLAG_OWN_REALTY,*
  - *FLAG_MOBIL*
- *We can remove all other FLAG columns..*

Imputing Values:



- *In that 10 columns there was a column **"OCCUPATION_TYPE"** , Which describes the **user occupation** was having 31% of Null Values.*
- *I have used **"Unknown"** variable to fill those 31% null values.*
- *First Highest percentage is: **"Unknown"***
- *Second Highest percentage is: **"Laborers"***

Standardize the Values:

Very high value data columns:

AMT_INCOME_TOTAL,
AMT_CREDIT,
AMT_GOODS_PRICE

Converting these numerical columns in categorical columns for better understanding.

Negative values Data columns: -

DAYS_BIRTH,
DAYS_EMPLOYED,
DAYS_REGISTRATION,
DAYS_ID_PUBLISH,
DAYS_LAST_PHONE_CHANGE.

Need to Make it correct those values convert DAYS_BIRTH to AGE in years , DAYS_EMPLOYED to YEARS EMPLOYED.

Standardize the Values:

Standardizing AMT_INCOME_TOTAL, AMT_CREDIT, AMT_GOODS_PRICE columns:

- *It has pricing from 0 to lakhs. so, lets make category and divide the pricing.*
- *Make **Income Range** range from 0 to 10 Lakhs.*

  bins = [0,1,2,3,4,5,6,7,8,9,10,11]

  slot = ['0-1L','1L-2L', '2L-3L','3L-4L','4L-5L','5L-6L','6L-7L','7L-8L','8L-9L','9L-10L','10L Above']
- *Make **Credit Range** range from 0 to 10 Lakhs.*

  bins = [0,1,2,3,4,5,6,7,8,9,10,100]

  slots = ['0-1L','1L-2L', '2L-3L','3L-4L','4L-5L','5L-6L','6L-7L','7L-8L','8L-9L','9L-10L','10L Above']

- *Make **Price of Goods** range from 0 to 10 Lakhs.*

  bins = [0,1,2,3,4,5,6,7,8,9,10,100]

  slots = ['0-1L','1L-2L', '2L-3L','3L-4L','4L-5L','5L-6L','6L-7L','7L-8L','8L-9L','9L-10L','10L Above']

Standardize the Values:

Standardizing – Negative Columns

As Mentioned below,

Negative values Data columns:

- *DAYS_BIRTH,*
- *DAYS_EMPLOYED,*
- *DAYS_REGISTRATION,*
- *DAYS_ID_PUBLISH,*
- *DAYS_LAST_PHONE_CHANGE.*

	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	DAYS_LAST_PHONE_CHANGE
count	307511.000000	307511.000000	307511.000000	307511.000000	307510.000000
mean	-16036.995067	63815.045904	-4986.120328	-2994.202373	-962.858788
std	4363.988632	141275.766519	3522.886321	1509.450419	826.808487
min	-25229.000000	-17912.000000	-24672.000000	-7197.000000	-4292.000000
25%	-19682.000000	-2760.000000	-7479.500000	-4299.000000	-1570.000000
50%	-15750.000000	-1213.000000	-4504.000000	-3254.000000	-757.000000
75%	-12413.000000	-289.000000	-2010.000000	-1720.000000	-274.000000
max	-7489.000000	365243.000000	0.000000	0.000000	0.000000

Before: - ve Values

Using Absolute function converting Negative Values to Positive Values

Standardize the Values:

Standardizing – Negative Columns

As Mentioned below,

Positive values Data columns:

- *DAYS_BIRTH,*
- *DAYS_EMPLOYED,*
- *DAYS_REGISTRATION,*
- *DAYS_ID_PUBLISH,*
- *DAYS_LAST_PHONE_CHANGE.*

	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	DAYS_LAST_PHONE_CHANGE
count	307511.000000	307511.000000	307511.000000	307511.000000	307510.000000
mean	16036.995067	67724.742149	4986.120328	2994.202373	962.858788
std	4363.988632	139443.751806	3522.886321	1509.450419	826.808487
min	7489.000000	0.000000	0.000000	0.000000	0.000000
25%	12413.000000	933.000000	2010.000000	1720.000000	274.000000
50%	15750.000000	2219.000000	4504.000000	3254.000000	757.000000
75%	19682.000000	5707.000000	7479.500000	4299.000000	1570.000000
max	25229.000000	365243.000000	24672.000000	7197.000000	4292.000000

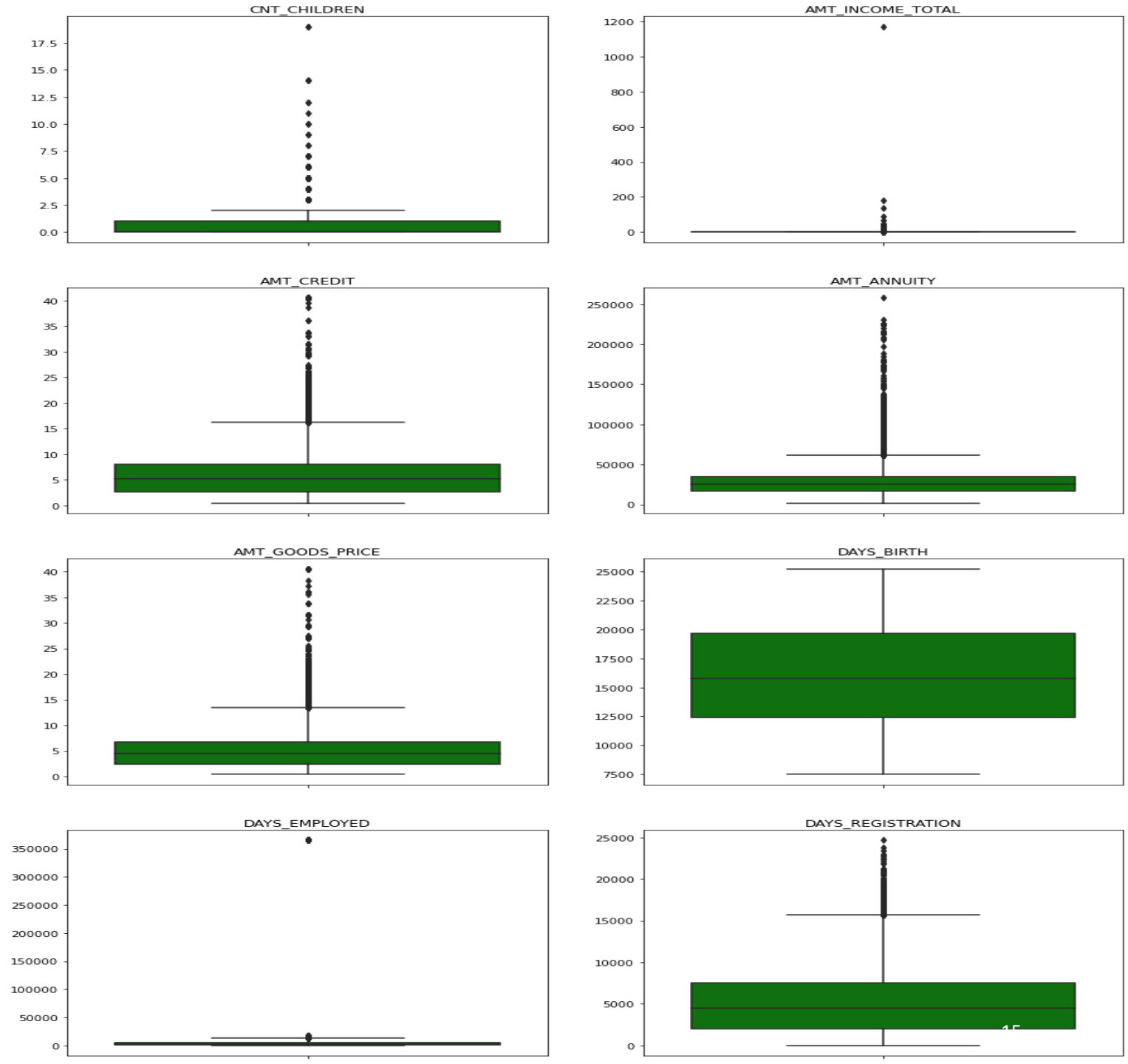
After: + ve Values

By Using Absolute function converting Negative Values to Positive Values

Standardize the Values:

Find the Outliners

- Max Outliners: AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN
- Min Outliners: AMT_INCOME_TOTAL
- No Outliners: DAYS_BIRTH



Summary on Datasets: Application_Data.csv

- *States that: Application_Data.csv:*
- *There are: **3,07,511** Rows & **53** Columns.*
- *Types of datatypes available:*
  - *Integers,*
  - *Float values,*
  - *Strings.*
- *Found the Null values, Filled them with "Unknown" variable.*
- *Removed unwanted columns and other columns.*
- *We have worked on the negative values and converted them into positive values in some of columns.*
- *We have converted Values in proper format.*
- *Now file is neat and clean for further process.*

Summary on Datasets: Previous_Application_Data.csv

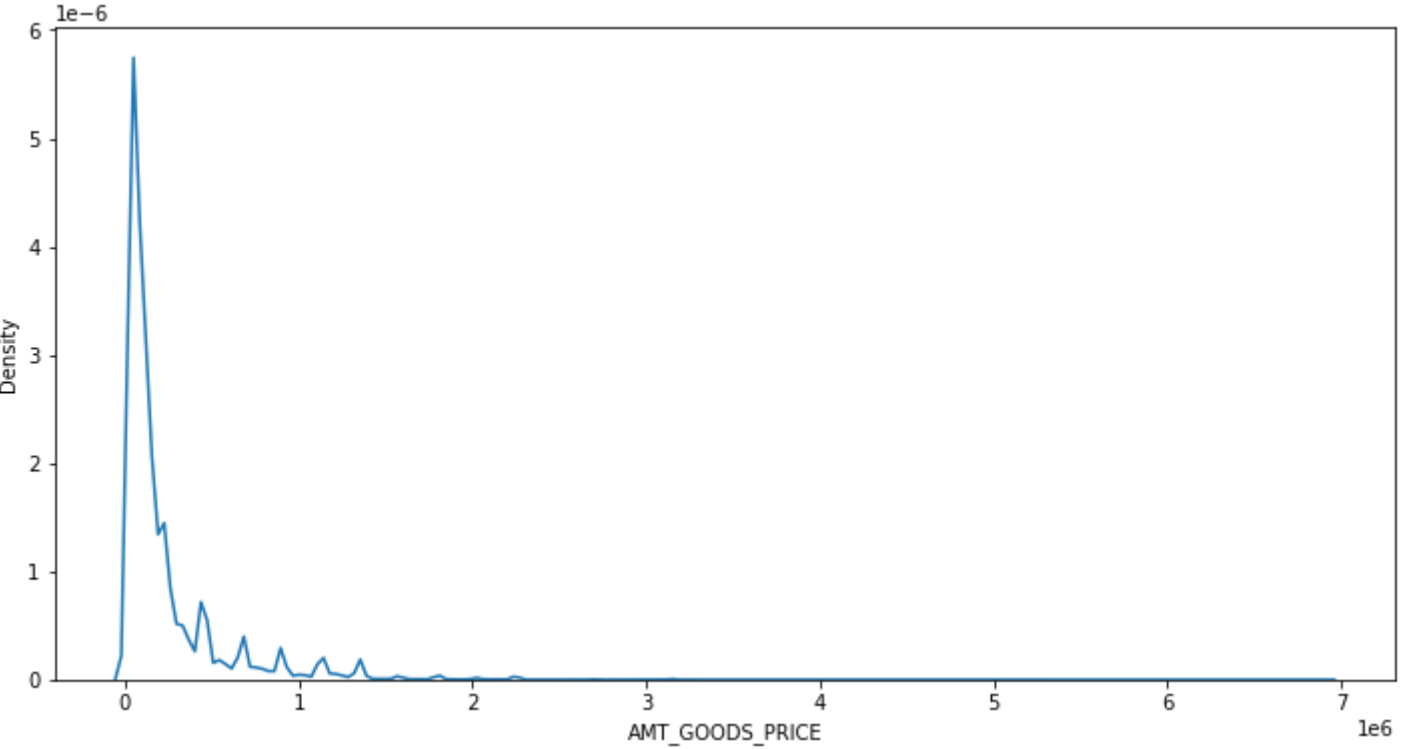
- *States that: Application_Data.csv:*
- *There are: **1670214** Rows & **37** Columns.*
- *Types of datatypes available:*
  - *Integers,*
  - *Float values,*
  - *Strings.*
- *Found the Null values, Filled them with "Unknown" variable.*
- *Removed unwanted columns and other columns.*
- *We have worked on the negative values and converted them into positive values in some of columns.*
- *We have converted Values in proper format.*
- *Now file is neat and clean for further process.*

Data Set Analyzing using Graphical Representation.

Analyzing The Data using Kdeplot:

---

1. Plotting kde plot for "AMT_GOODS_PRICE" to understand the distribution



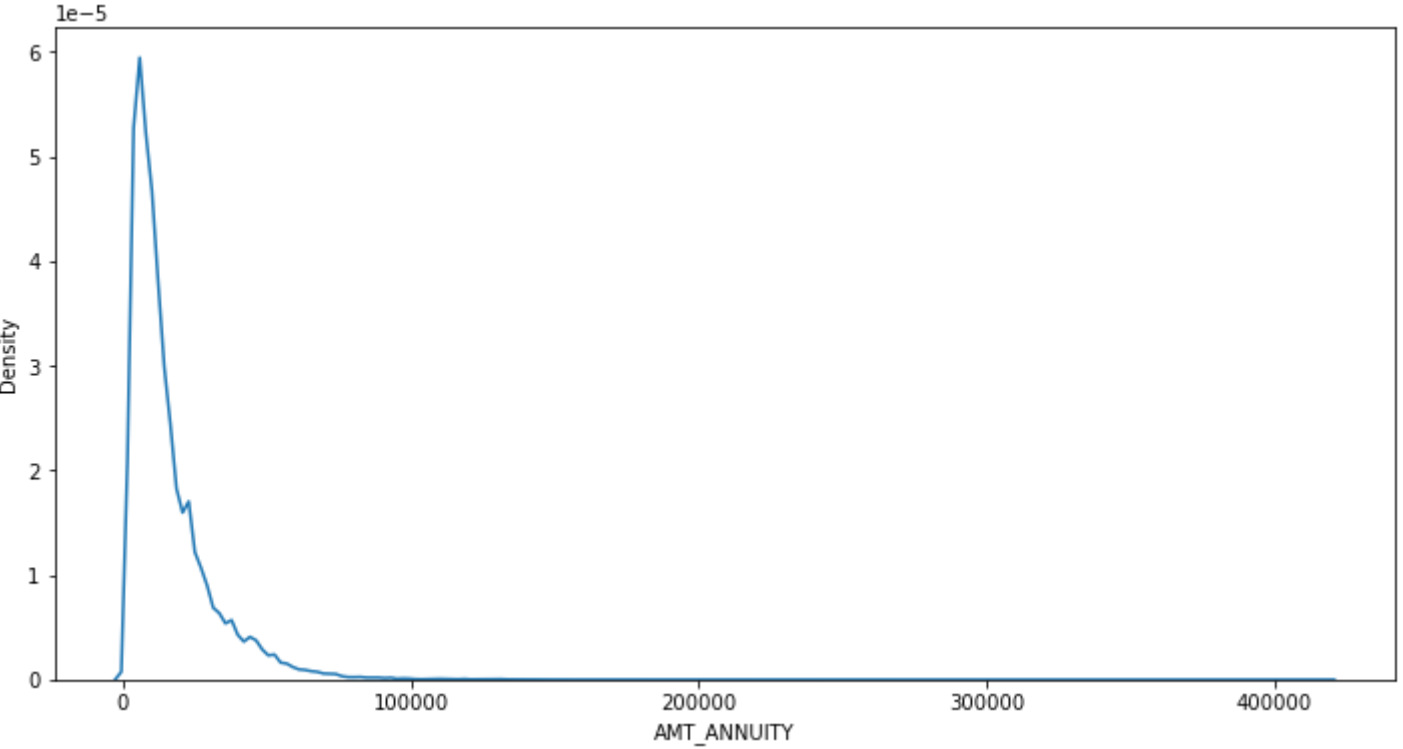
- *There are several peaks along the distribution. Let's impute using the mode, mean and median and see if the distribution is still about the same.*

Data Set Analyzing using Graphical Representation.

Analyzing The Data using Kdeplot:

---

2. plotting a kdeplot to understand distribution of "AMT_ANNUITY"



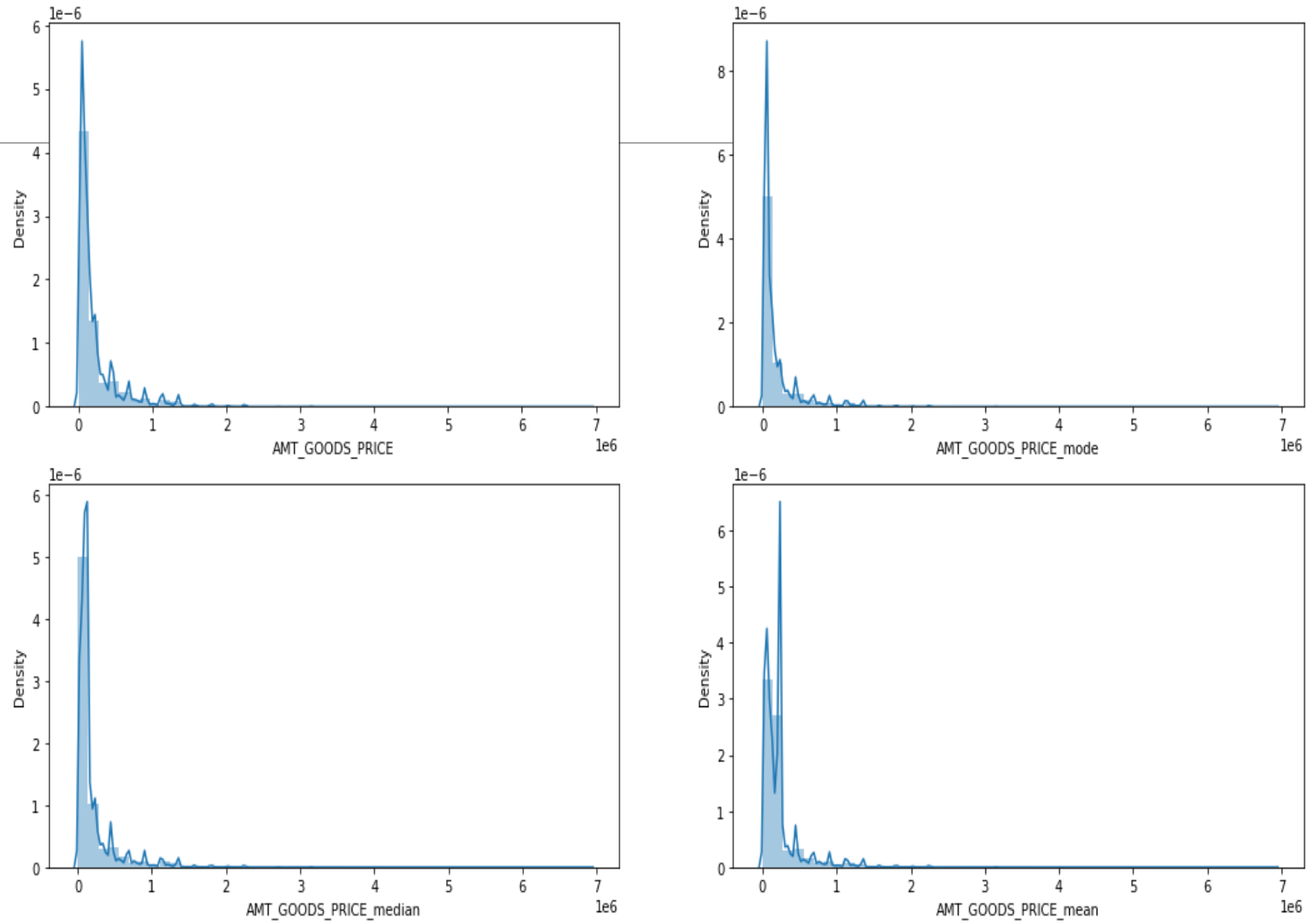
** There is a single peak at the left side of the distribution, and it indicates the presence of outliers and hence imputing with mean would not be the right approach and hence imputing with median.*

Data Set Analyzing using Graphical Representation.

Provided RAW Data & Imputed Data [Mode, Median, Mean Values]

Analyzing The Data using Kdeplot:

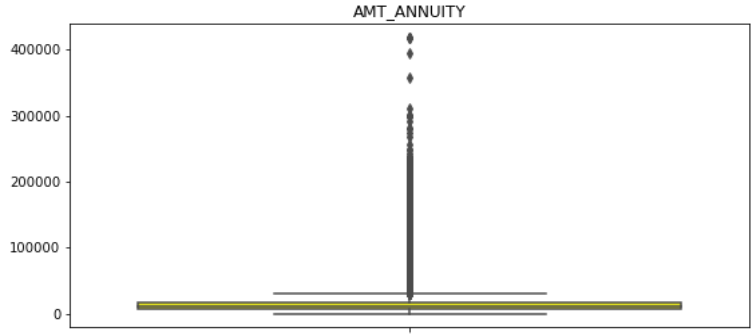
The original distribution is closer with the distribution of data imputed with mode in this case, thus will impute mode for missing values

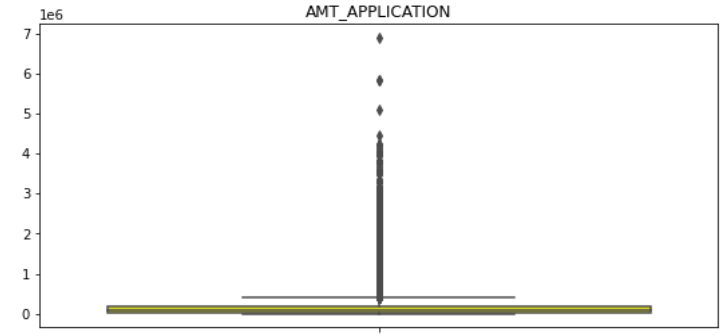


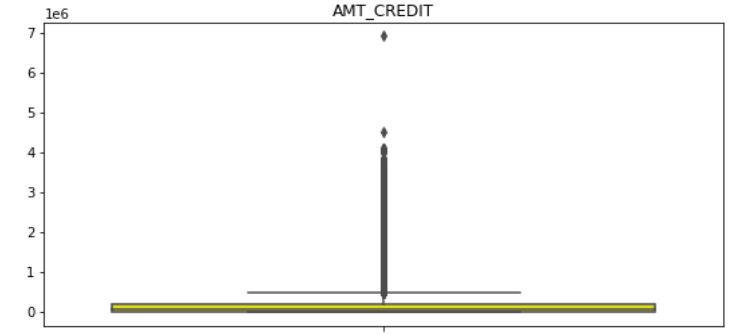
Finding outliers in:

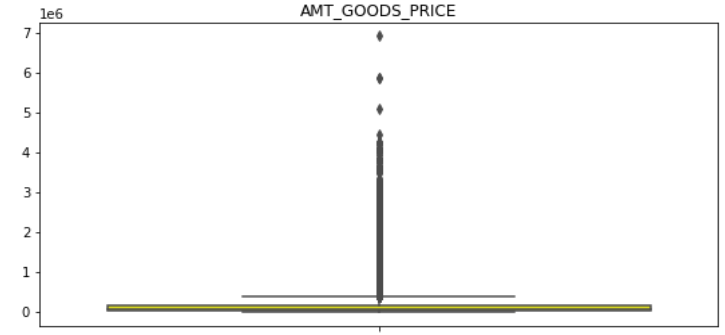
['amt_annuity','amt_application','amt_credit','amt_goods_price','sellerplace_area','days_decision','cnt_payment']

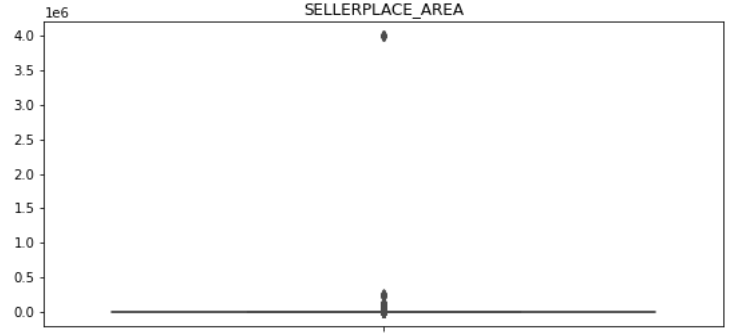
- *Summary It can be seen that in previous application data*
- *AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA consist max. number of outliers.*
- *CNT_PAYMENT consist less outlier values.*
- *DAYS_DECISION has little number of outliers indicating that these previous applications decisions.*

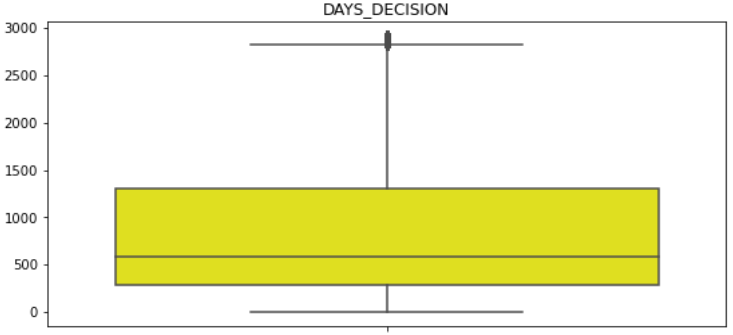


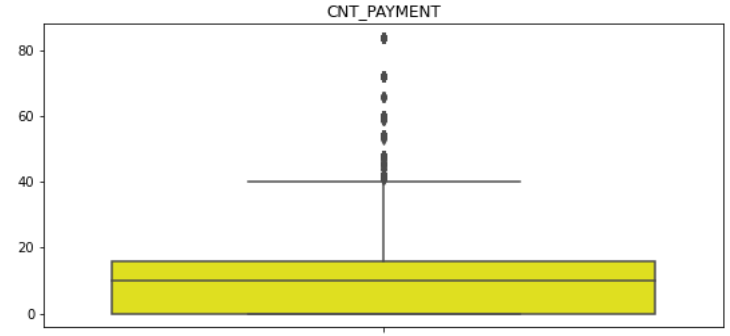








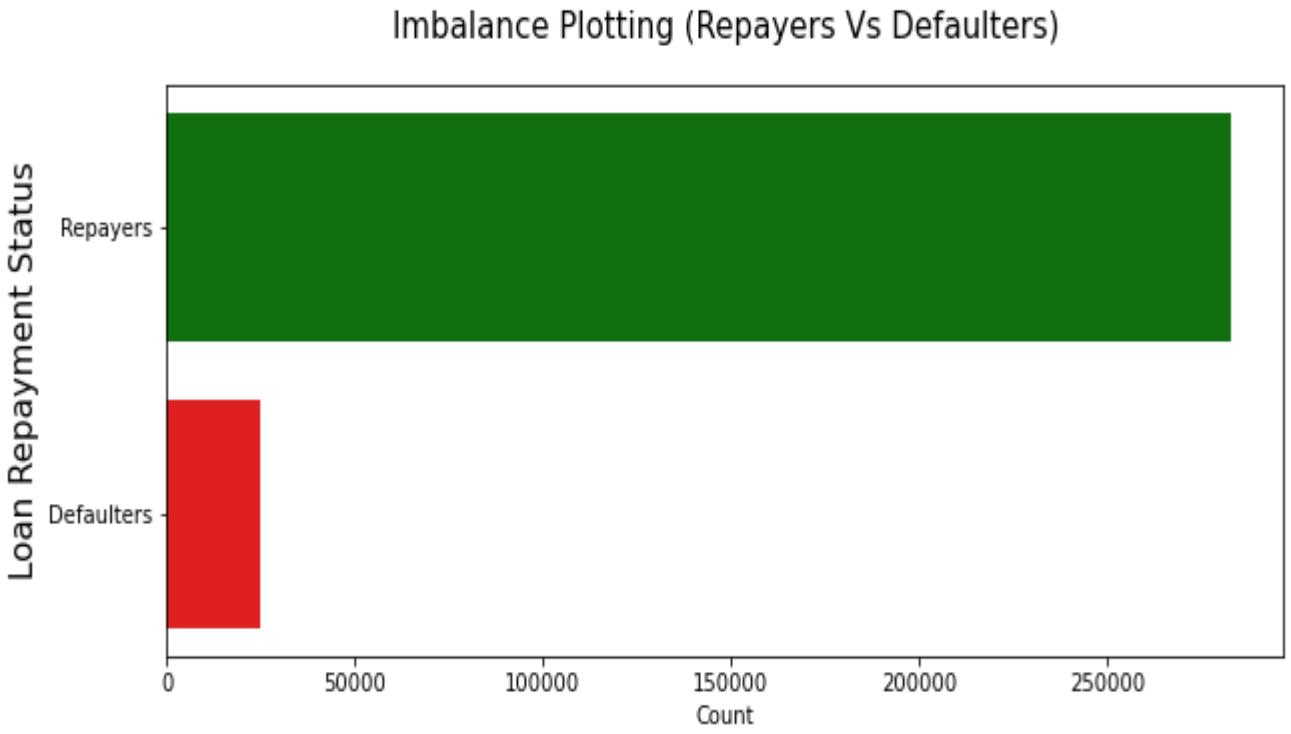




Data Set Analyzing using Graphical Representation.

Repayers & Defaulters

- *Repayer Percentage is 91.93%*
- *Defaulter Percentage is 8.07%*
- *Imbalance Ratio with respect to*
- *Repayer and Defaulter is given: 11.39/1 (approx)*



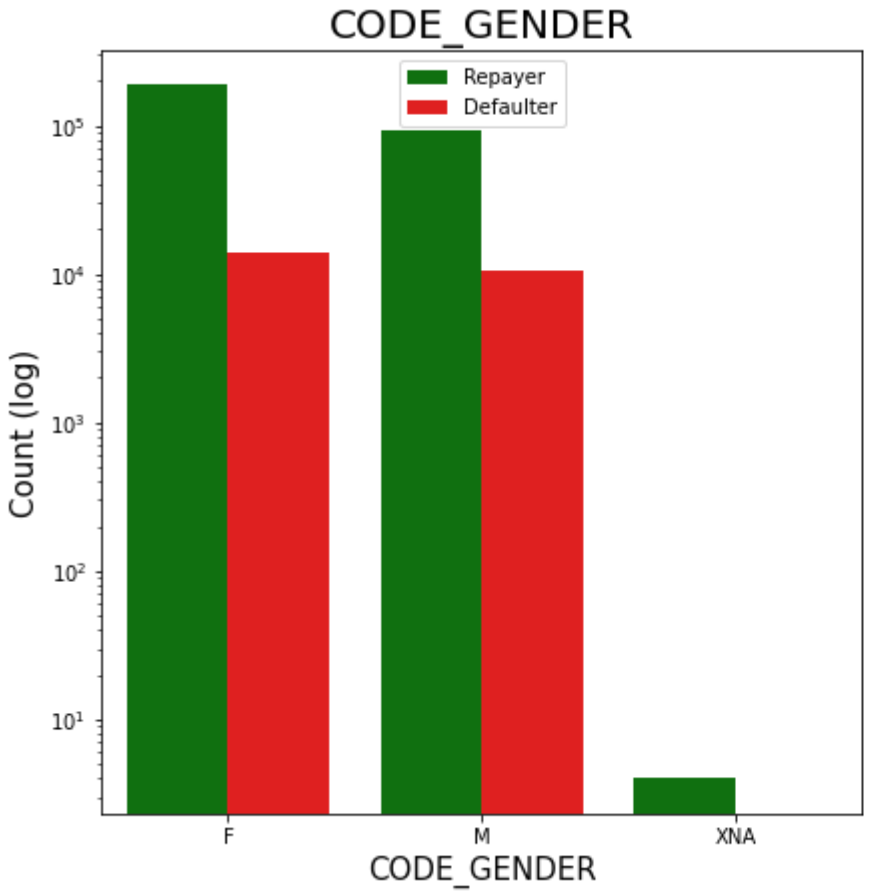
Data Set Analyzing using Graphical Representation.

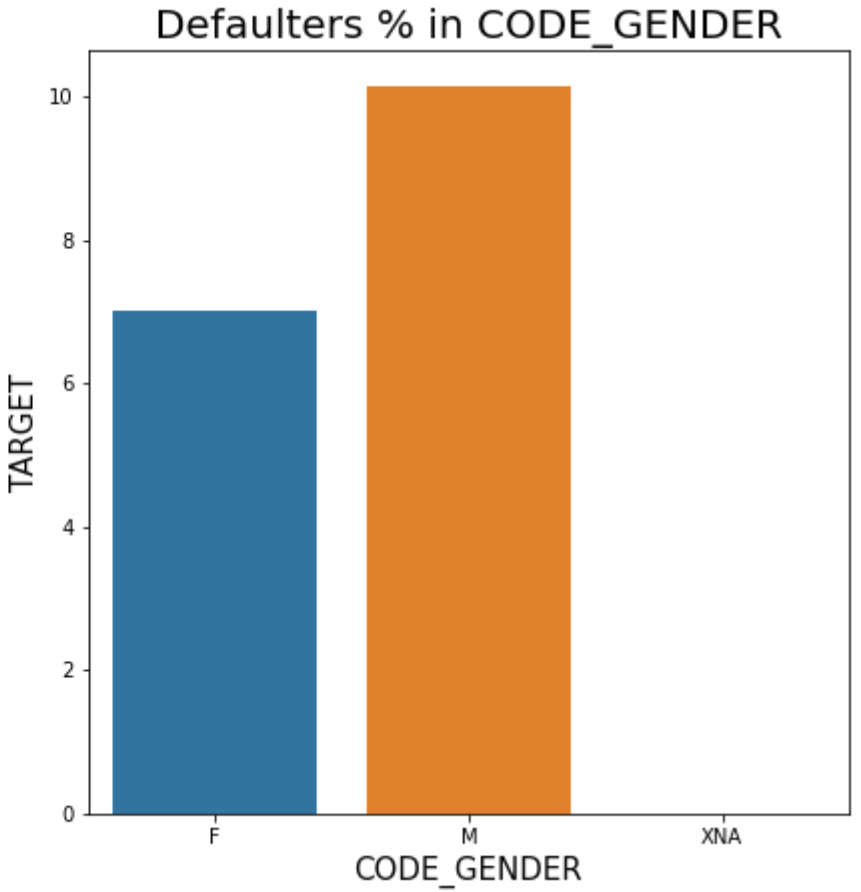
Analyzing Univariate, Bivariate, Multivariate :

Categorical Univariate Variables Analysis

Gender wise Analysis

Based on the percentage of default credits, males have a higher chance of not returning their loans, comparing with women.





Data Set Analyzing using Graphical Representation.

Analyzing Univariate, Bivariate, Multivariate :

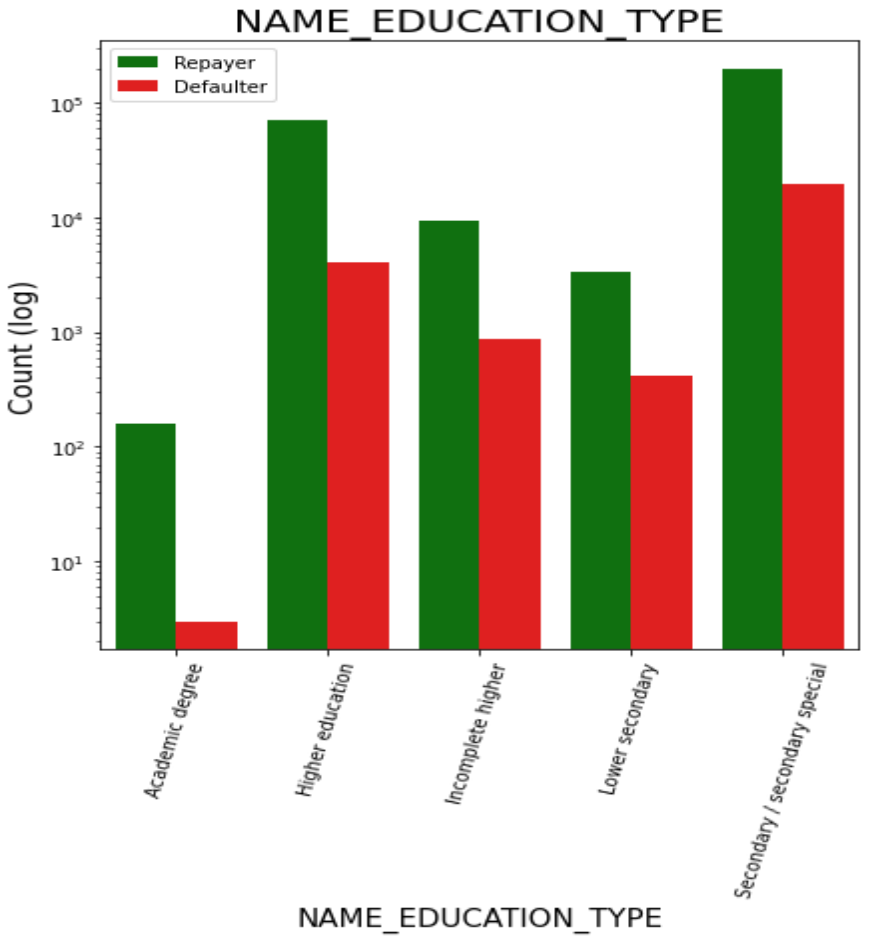
Categorical Univariate Variables Analysis

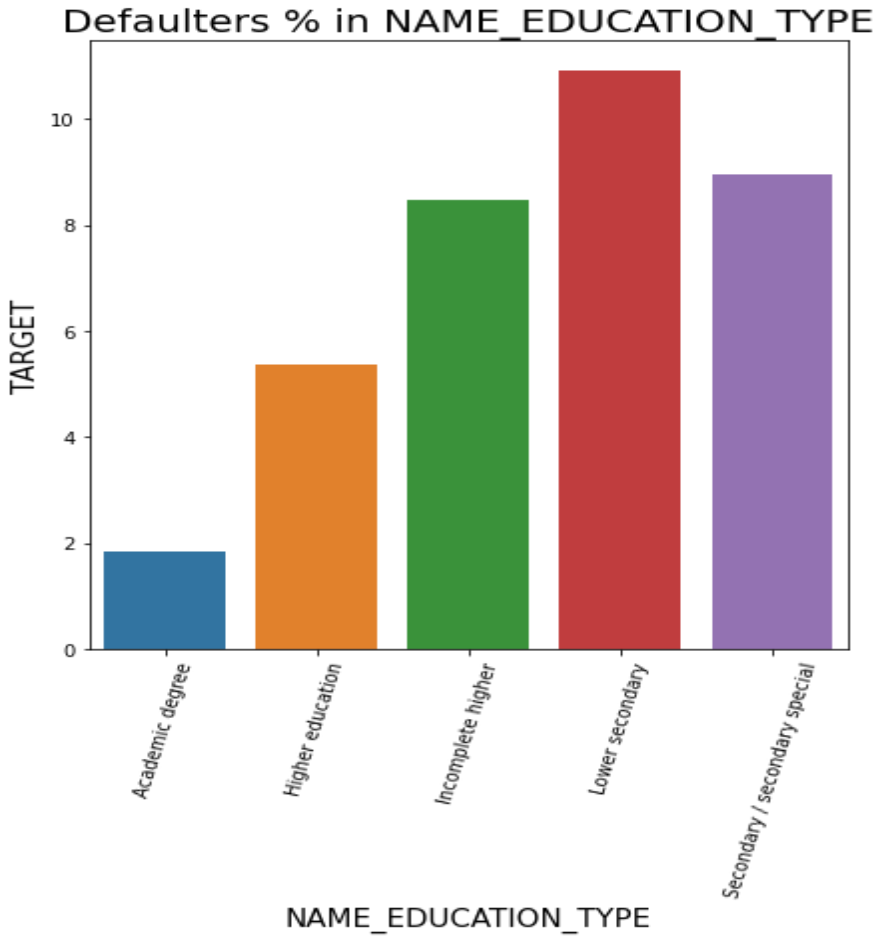
Education wise Analysis

Majority of clients have Secondary/secondary special education, followed by clients with Higher education.

Very few clients have an academic degree Lower secondary category have highest rate of defaulter.

People with Academic degree are least likely to default.





Data Set Analyzing using Graphical Representation.

Analyzing Univariate, Bivariate, Multivariate :

Categorical Univariate Variables Analysis

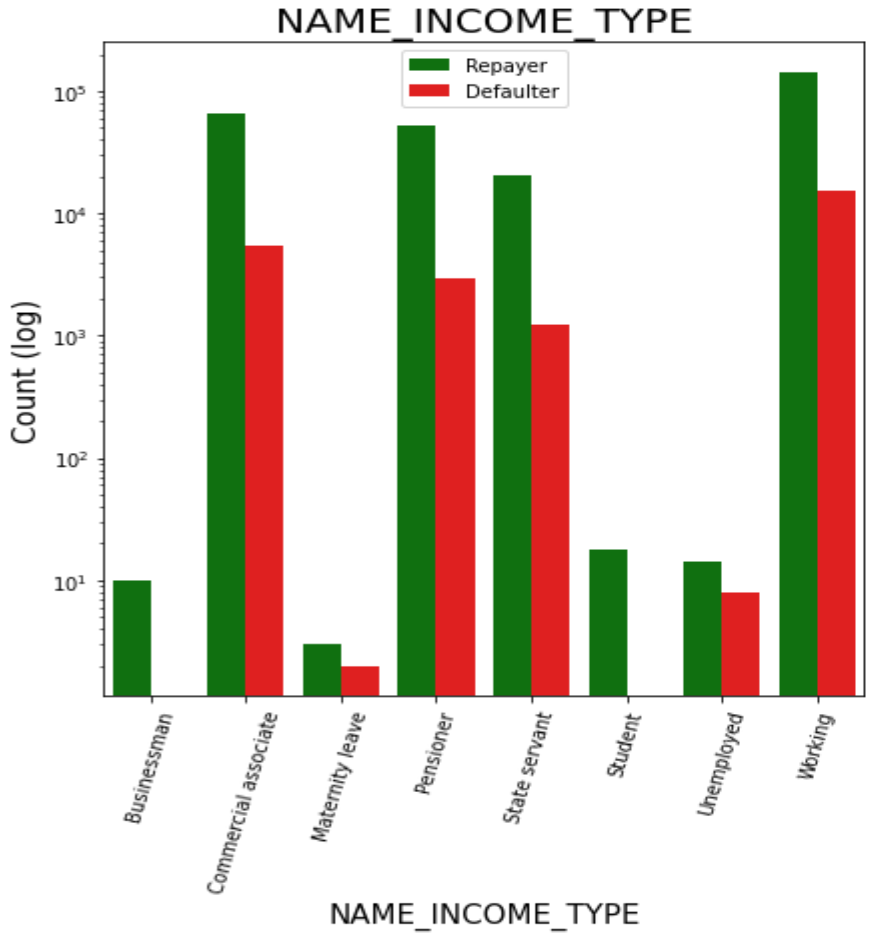
Income wise Analysis

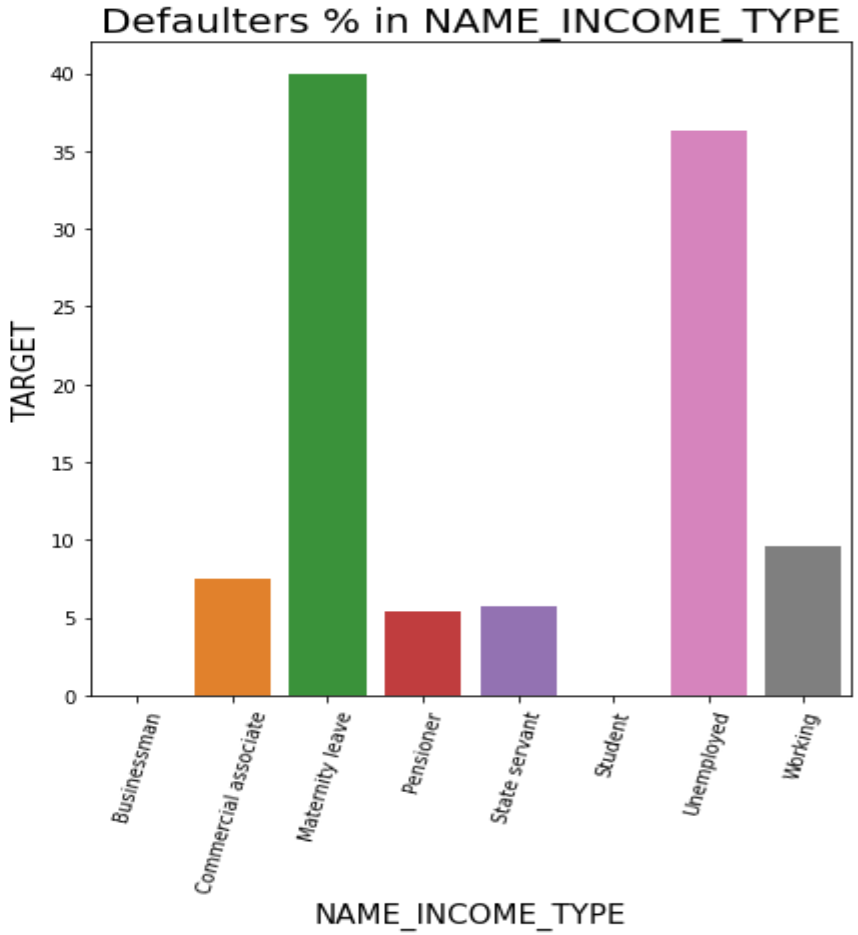
Most of applicants for loans income type is Working, followed by Commercial associate, Pensioner and State servant.

The applicants who are on Maternity leave have defaulting percentage of 40% which is the highest, followed by Unemployed (37%).

The rest under average around 10% defaulters.

Student and Businessmen though less in numbers, do not have default record. Safest two categories for providing loan..





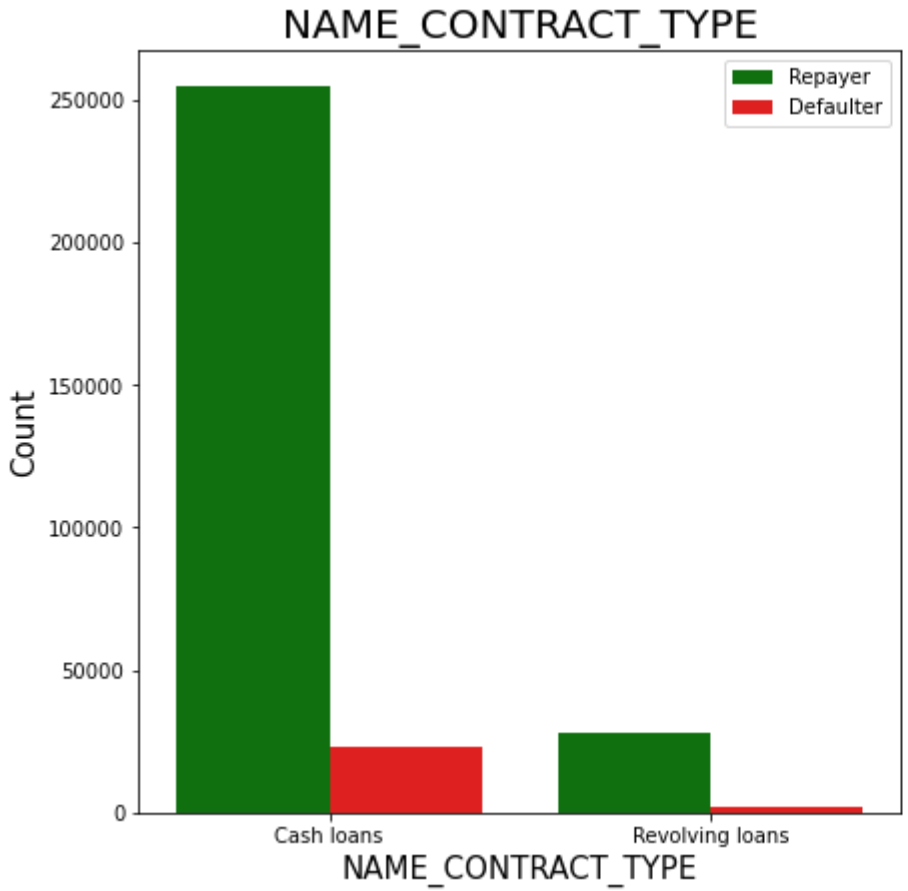
Data Set Analyzing using Graphical Representation.

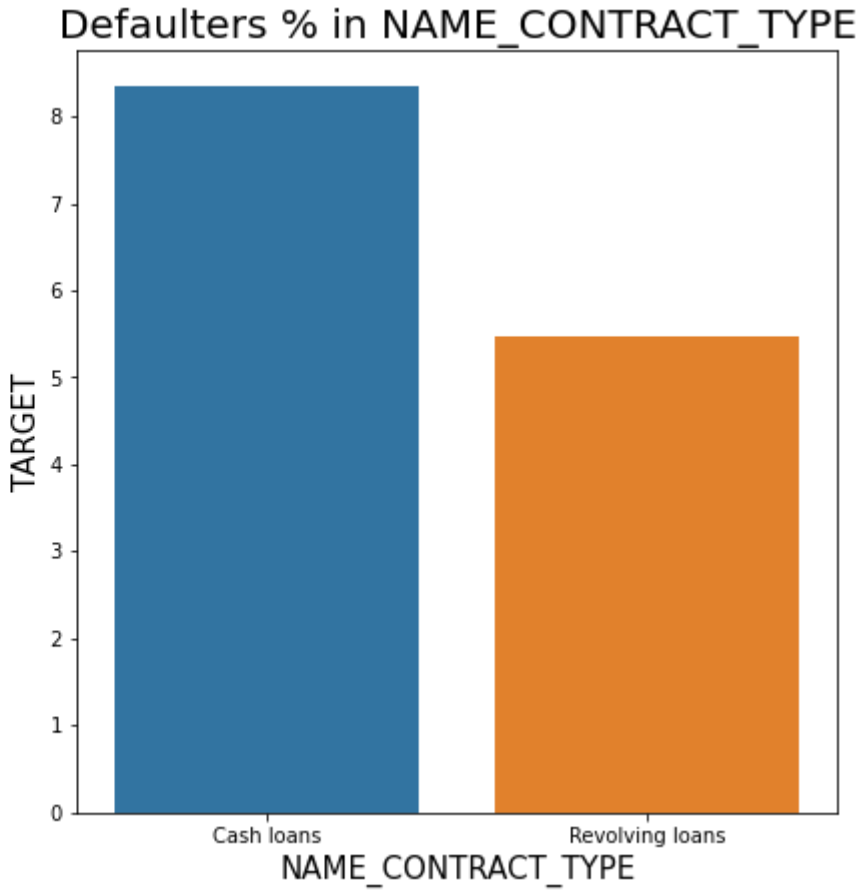
Analyzing Univariate, Bivariate, Multivariate :

Categorical Univariate Variables Analysis

Contract wise Analysis

Contract type: Revolving loans are just a small fraction (10%) from the total number of loans Around 8-9% Cash loan applicants and 5-6% Revolving loan applicant are in defaulters





Data Set Analyzing using Graphical Representation.

Analyzing Univariate, Bivariate, Multivariate :

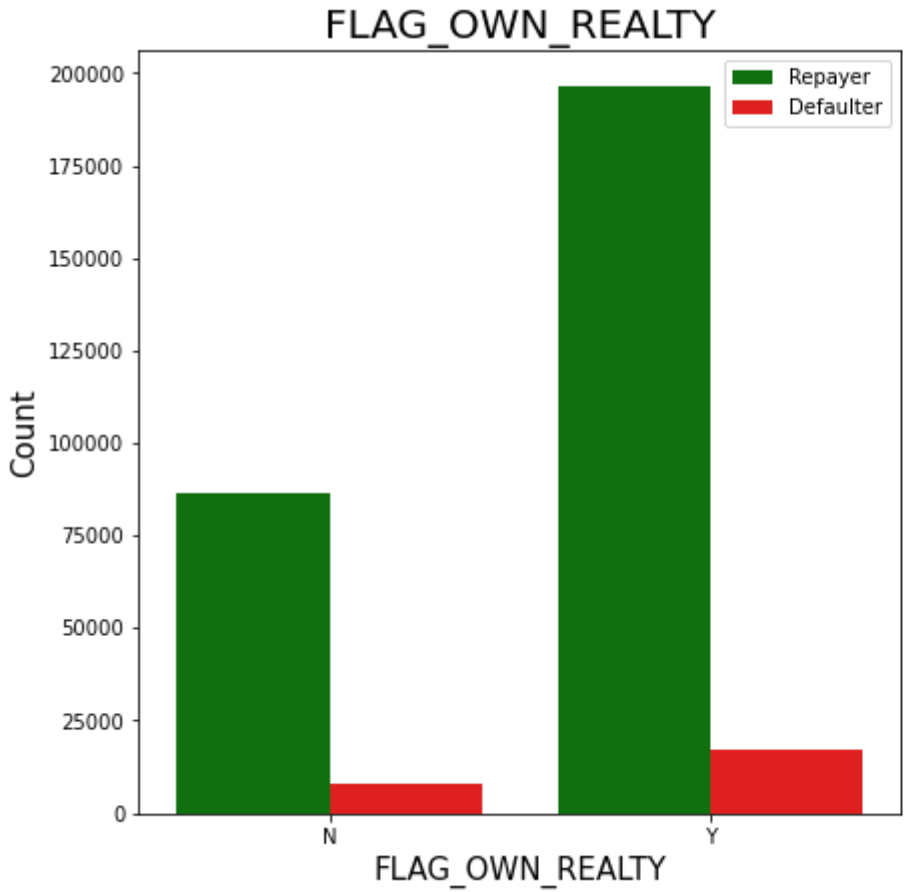
Categorical Univariate Variables Analysis

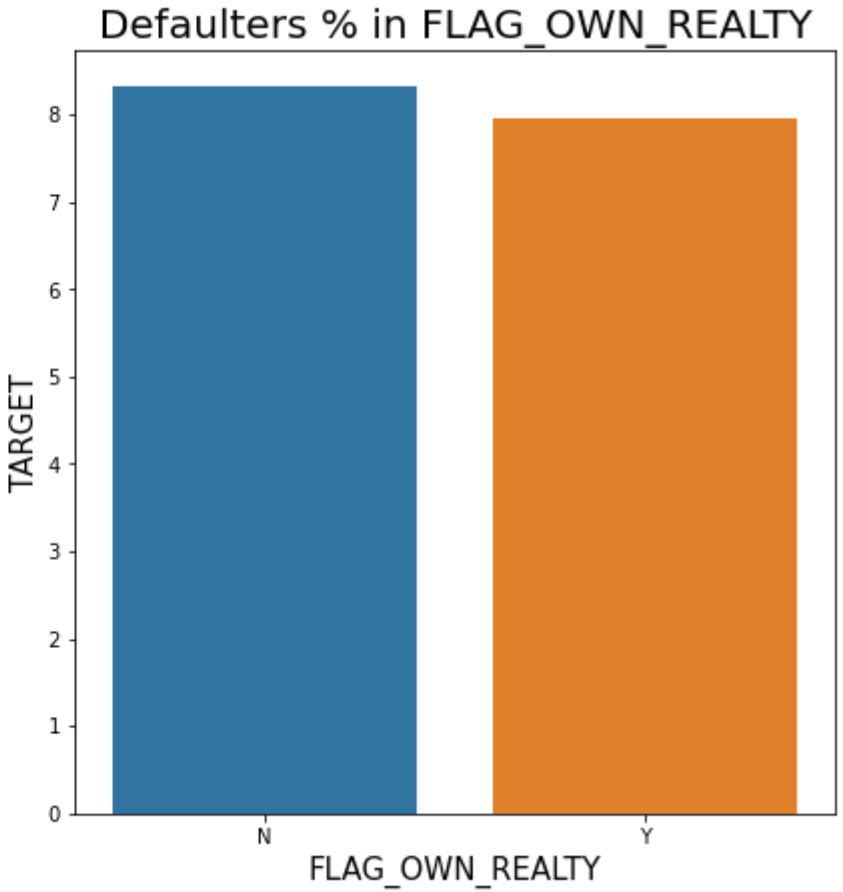
Real Estate Analysis

The clients who own real estate are more than double of the ones that don't own.

The defaulting rate of both categories are around the same (~8%).

Thus we can infer that there is no correlation between owning a reality and defaulting the loan.





Data Set Analyzing using Graphical Representation.

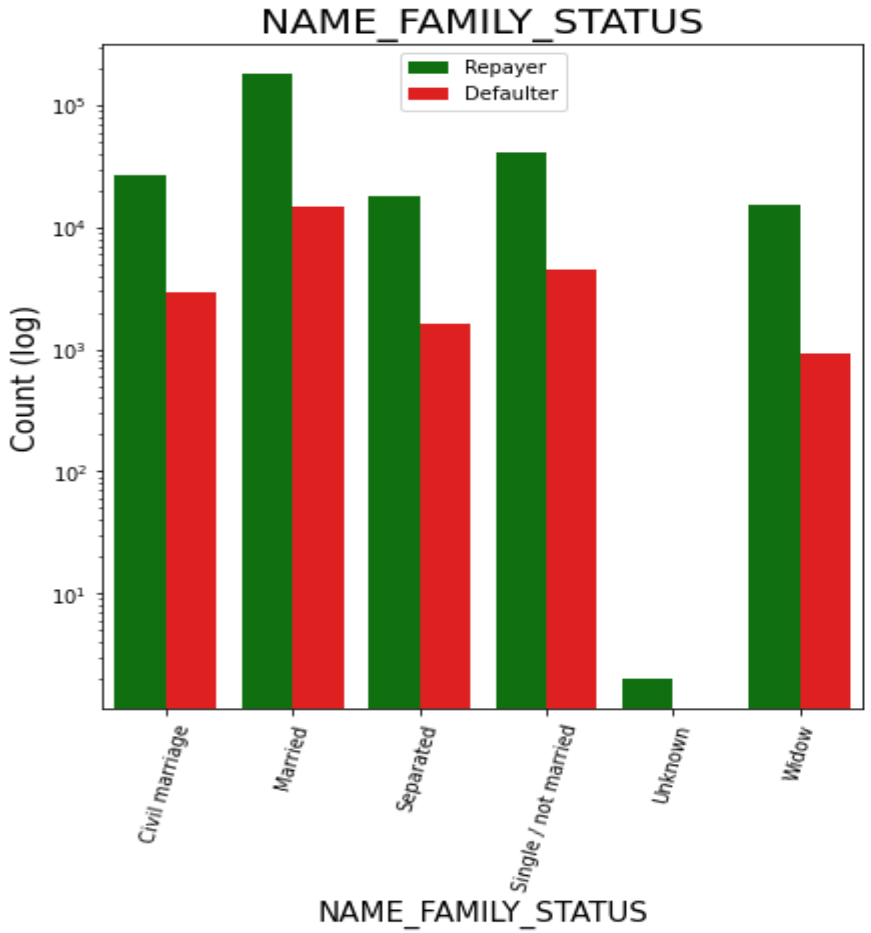
Analyzing Univariate, Bivariate, Multivariate :

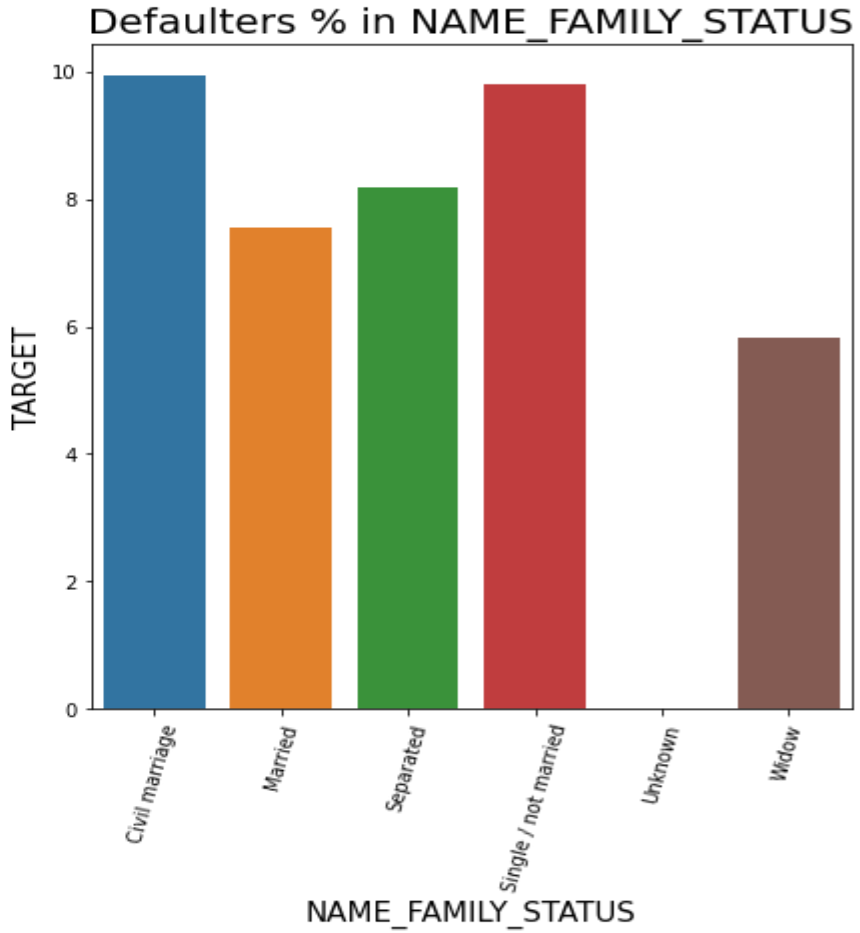
Categorical Univariate Variables Analysis

Family Analysis

Most of the people who have taken loan are married, followed by Single/not married and civil marriage. In Percentage of

defaulters Civil marriage has the highest percent around and widow has the lowest.





Data Set Analyzing using Graphical Representation.

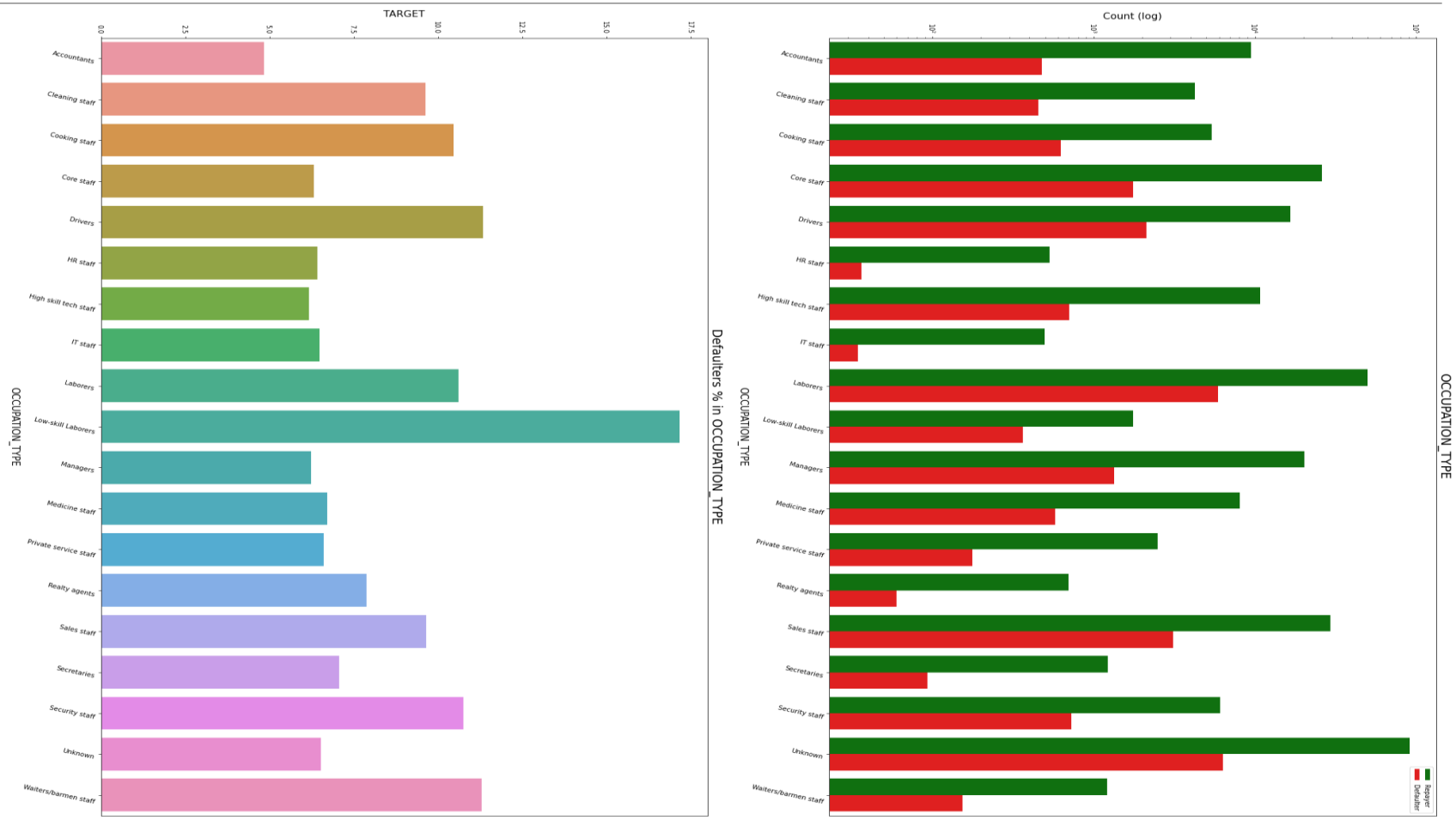
Analyzing Univariate, Bivariate, Multivariate :

Categorical Univariate Variables Analysis

Occupation Analysis

Category with highest percent of defautess are Low-skill Laborers (above 17%), followed by Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff.

IT staff are less likely to apply for Loan.

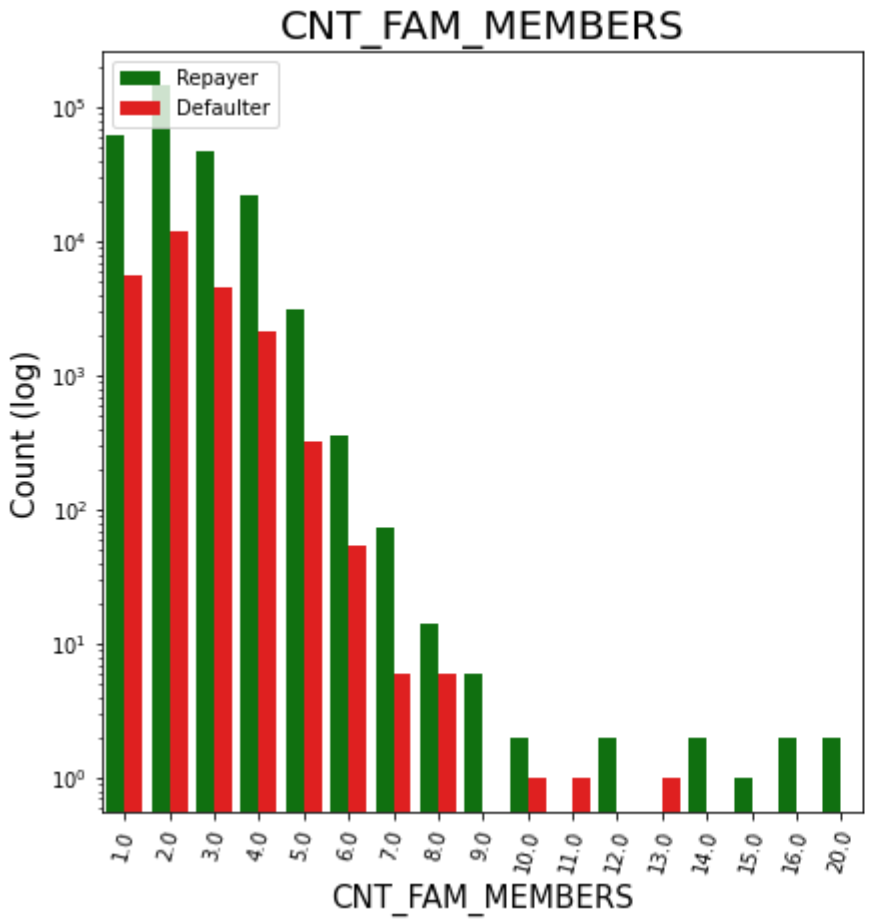


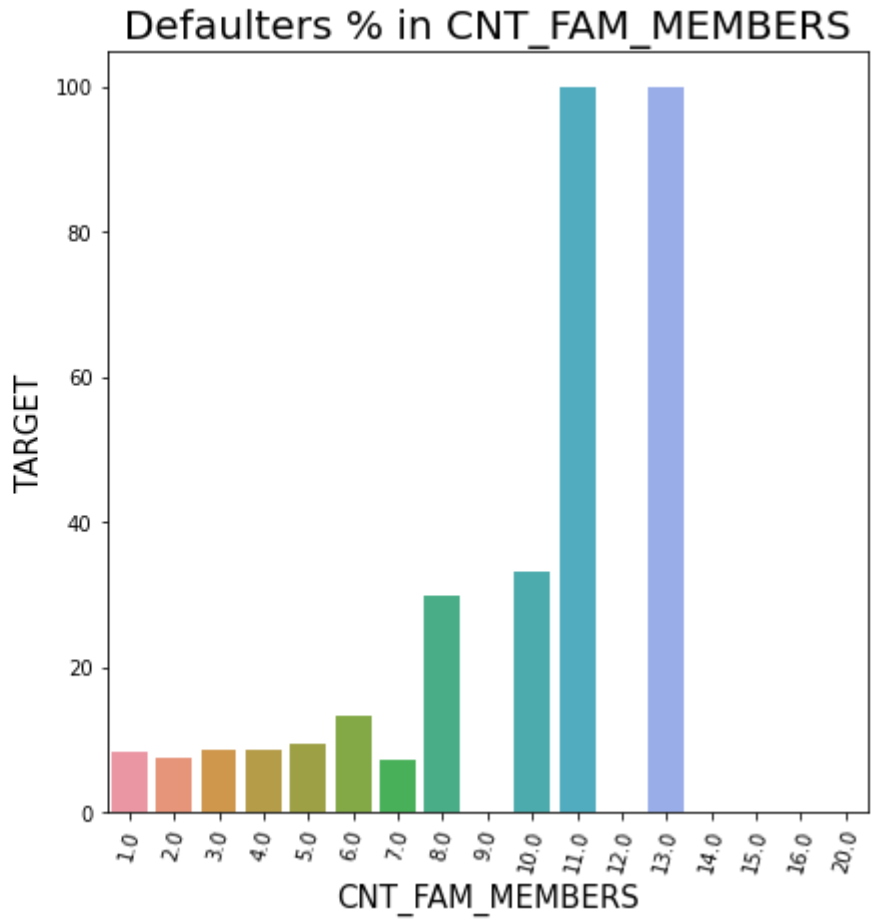
Data Set Analyzing using Graphical Representation.

Analyzing Univariate, Bivariate, Multivariate :

Categorical Univariate Variables Analysis

No. Family Members Analysis





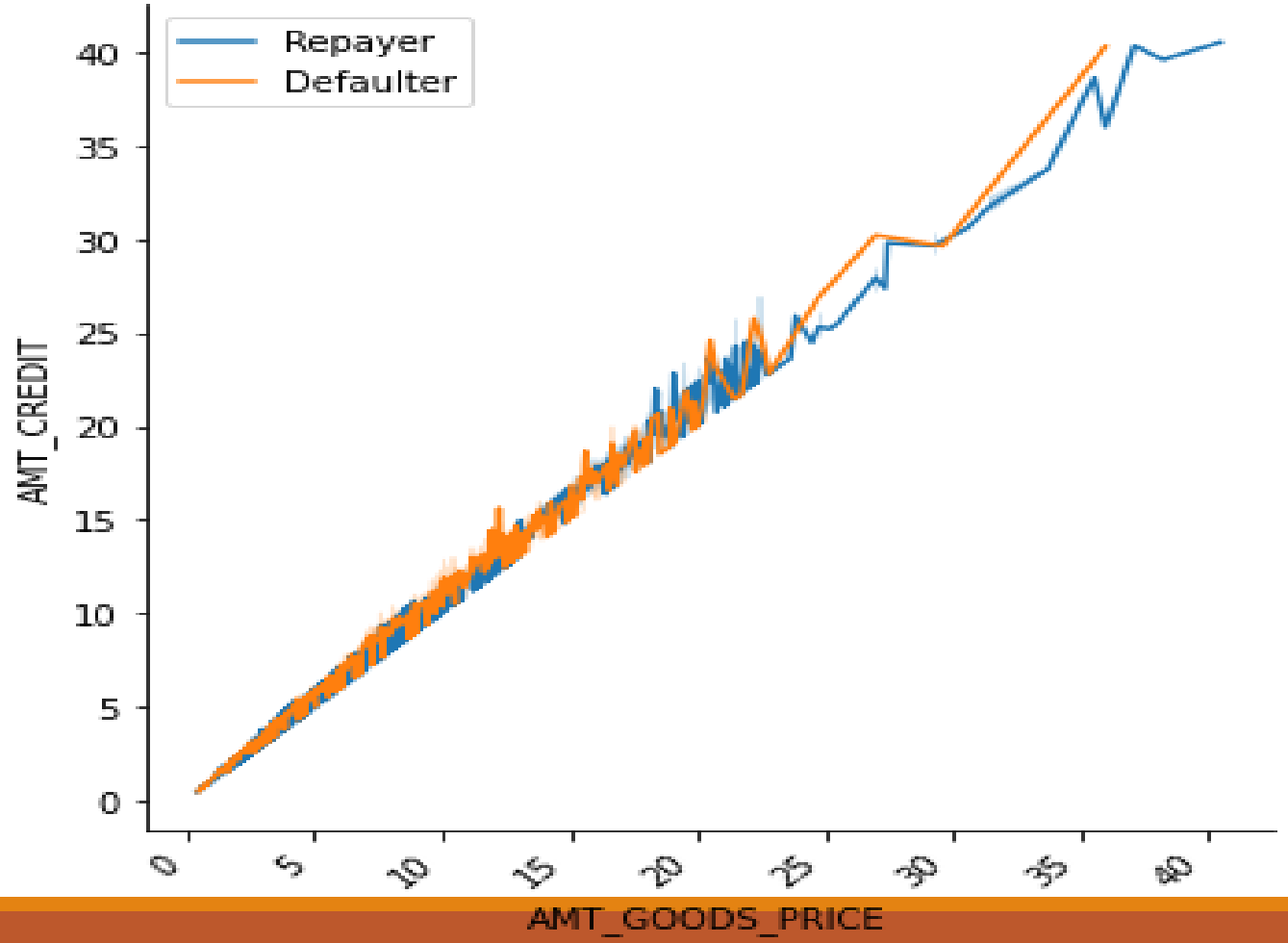
Data Set Analyzing using Graphical Representation.

Analyzing Univariate, Bivariate, Multivariate :

Categorical Univariate Variables Analysis

Numerical Univariate Analysis

When the credit amount goes beyond 30 Lakhs, there is an increase in defaulters.



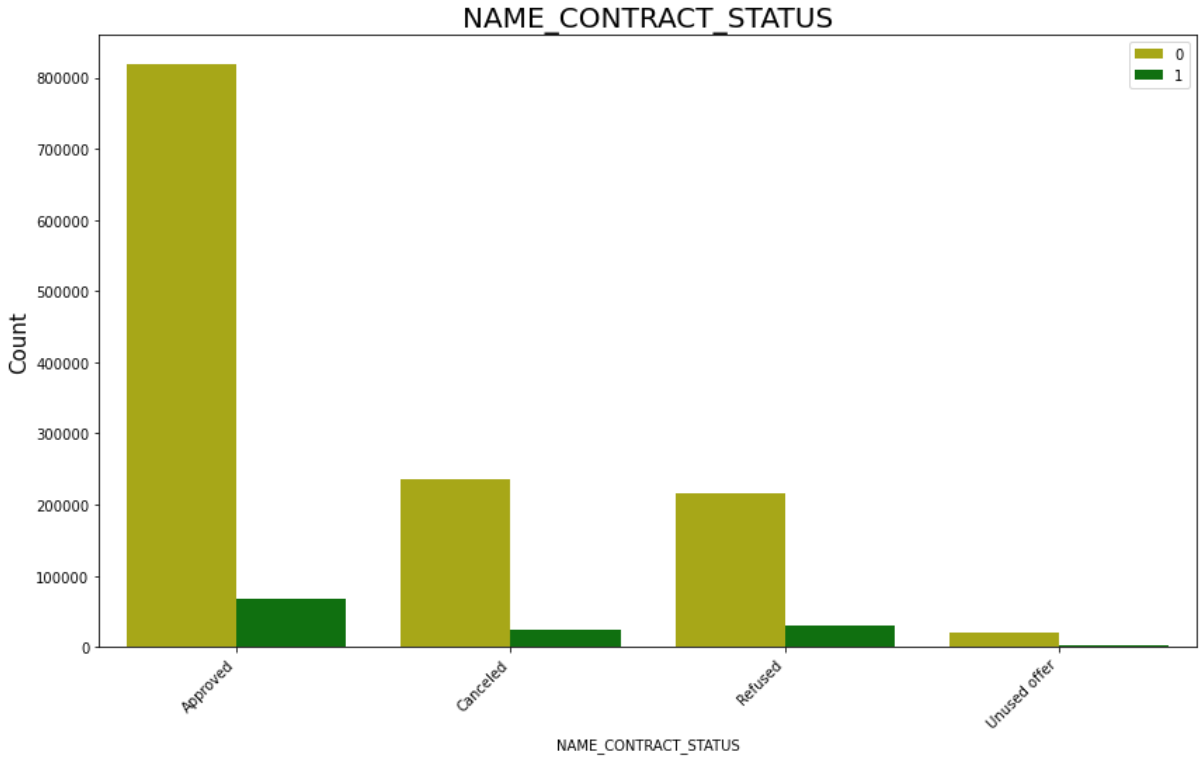
Data Set Analyzing using Graphical Representation.

Analyzing Univariate, Bivariate, Multivariate :

Categorical Univariate Variables Analysis

90% of the previously cancelled client have actually rep the loan. Revising the interest rates would increase business opportunity for these clients88% of the clients who have been previously refused a loan has payer back the loan in current case. Refusal reason should be recorded for further analysis as these clients could turn into potential repaying customer.

		Counts	Percentage
NAME_CONTRACT_STATUS	TARGET		
Approved	0	818856	92.41%
	1	67243	7.59%
Canceled	0	235641	90.83%
	1	23800	9.17%
Refused	0	215952	88.0%
	1	29438	12.0%
Unused offer	0	20892	91.75%
	1	1879	8.25%

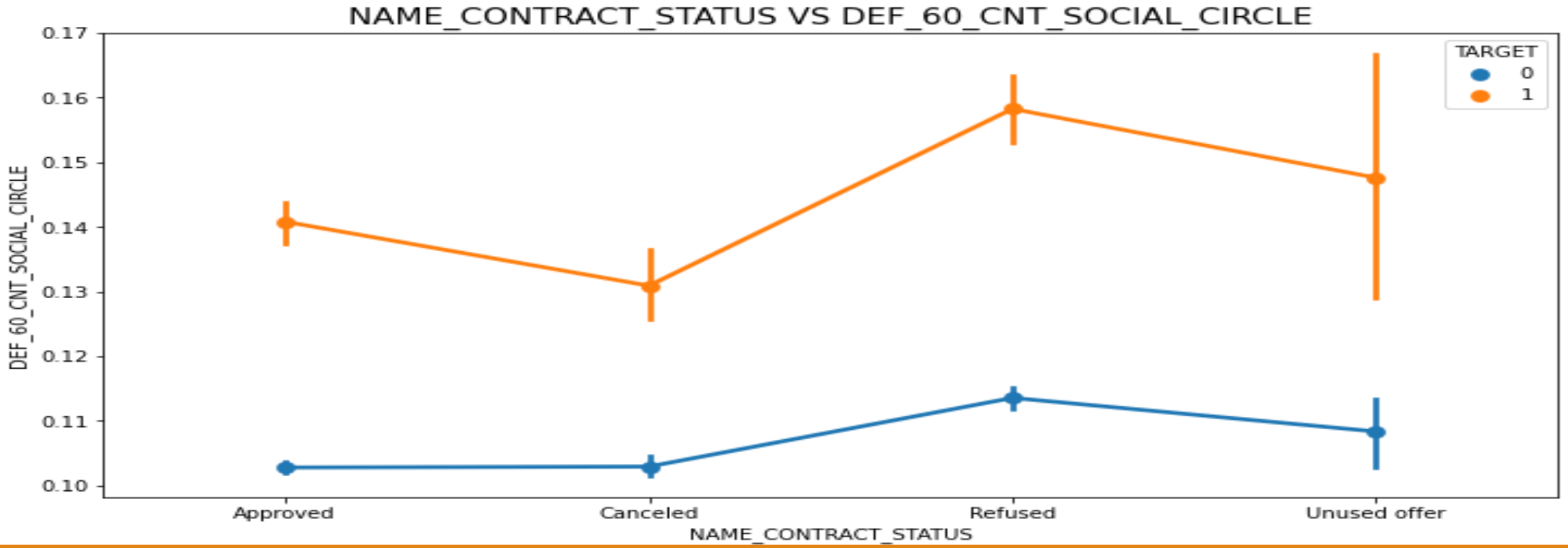


Data Set Analyzing using Graphical Representation.

Analyzing Univariate, Bivariate, Multivariate :

Categorical Univariate Variables Analysis

Clients who have average of 0.13 or higher their DEF_60_CNT_SOCIAL_CIRCLE score tend to default more and thus analysing client's social circle could help in disbursement of the loan.



THANK YOU