Credit EDA Case Study

Parita Patel
Suneedhi Sneha

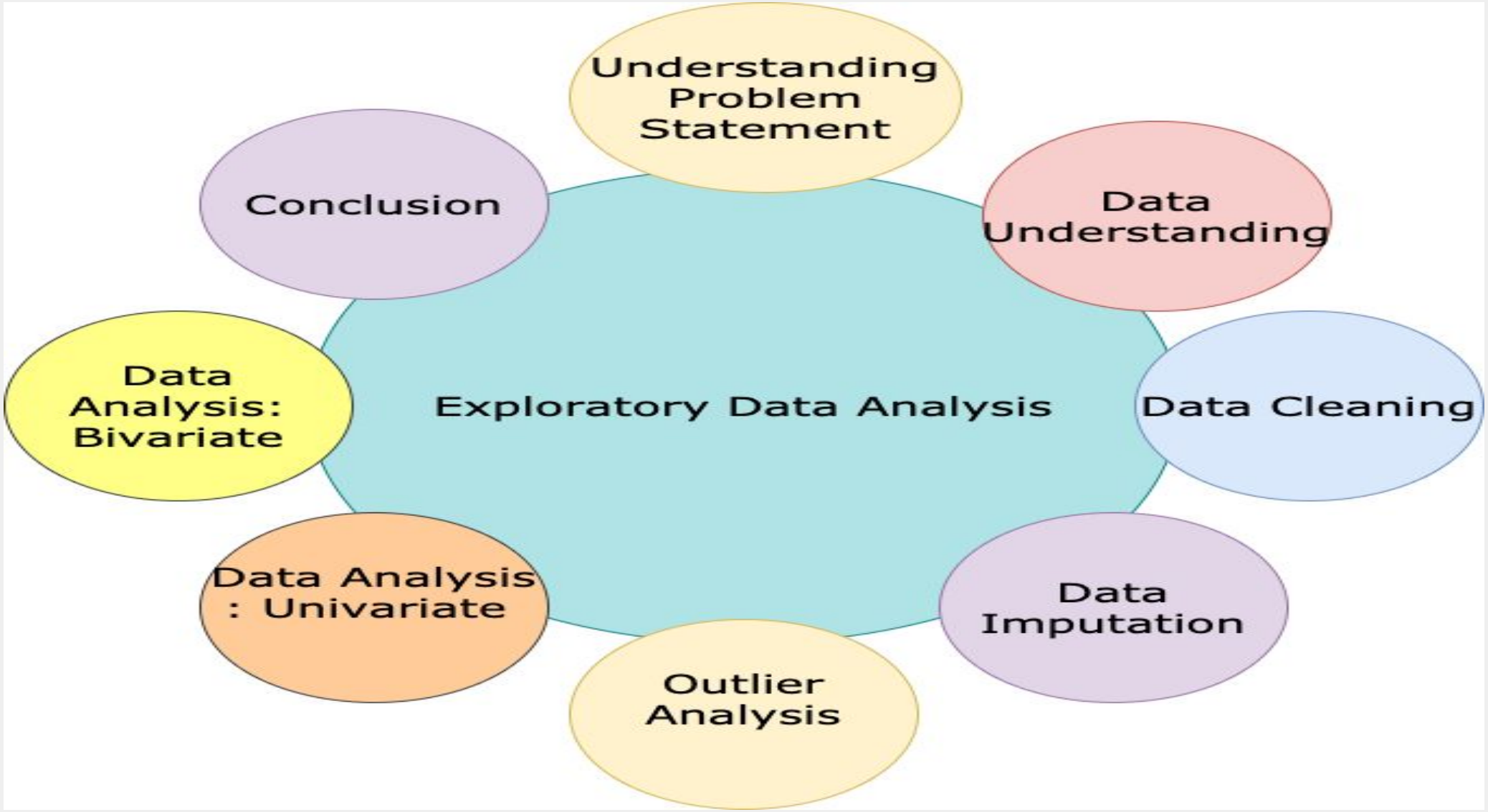
Business Purpose

Loans have always been essential in people's lives in this fast pace of life. Everyone has different reasons to take a loan. Moreover, the reason could be to have a dream car or a home, set up a business, or buy expensive dream products. Even rich people favour taking loans, overspending their cash to get tax benefits, and keeping the cash available for future unexpected and unconventional expenses.

Purpose to identify the defaulter

- Loans are also as crucial for both Lenders as well as for the Borrowers
- Almost all Money Banks make most of their revenues from the interests generated through loans.
- The caveat here is that the lenders make a profit only if the loan gets repaid.Therefore, the Lending Organizations have faced the challenge of analyzing the risk associated with each client.
- Therefore, it is essential to identify the risky behaviours of clients and make educated decisions.

EDA Analysis Approach



Imported Library

```
# Importing the required libraries

import numpy as np # linear algebra
import pandas as pd # data processing

 # visualization
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

 #statistics functions
from scipy import stats
from scipy.stats import norm
```

Reading Data files/Data Understanding

Filename: application_data.csv

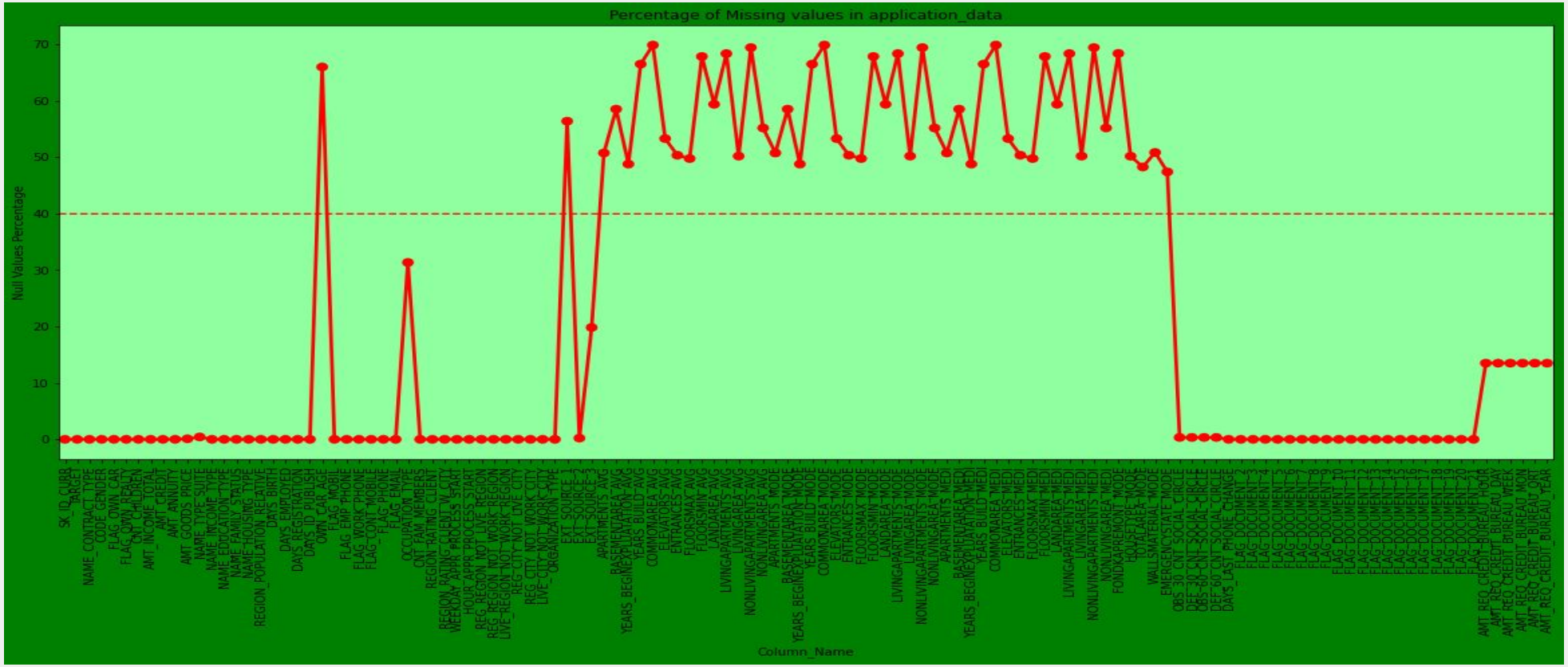
- Number of Rows and Columns in application_data file: (307511, 122)
- dtypes: float64(65), int64(41), object(16)

Filename: previous_application.csv

- Number of Rows and Columns in previous_application file: (1670214, 37)
- dtypes: float64(15), int64(6), object(16)

Data Cleaning & Manipulation

Data Observation: There are 49 columns in application_data dataframe high missing percentage.We will plan to drop the ones with missing value percentage as >=40%

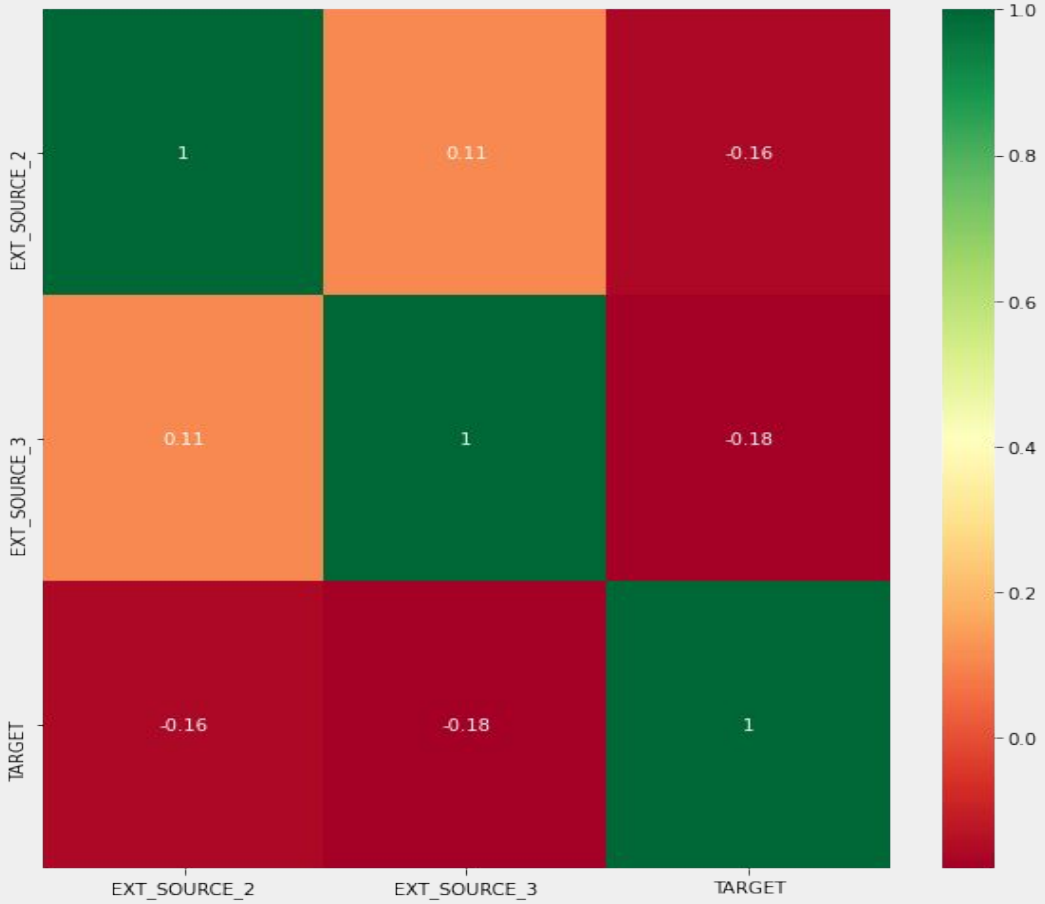


Data Cleaning & Manipulation

Dropped unnecessary columns from application_data:

- Number of Columns dropped : 49
- Before Deletion rows,columns (307511, 122)
- After Deletion rows,columns (307511, 73)

Correlation of EXT_SOURCE_X columns vs TARGET column



Observation:

As per Heatmap, there is almost no correlation between EXT_SOURCE_X columns vs TARGET column.

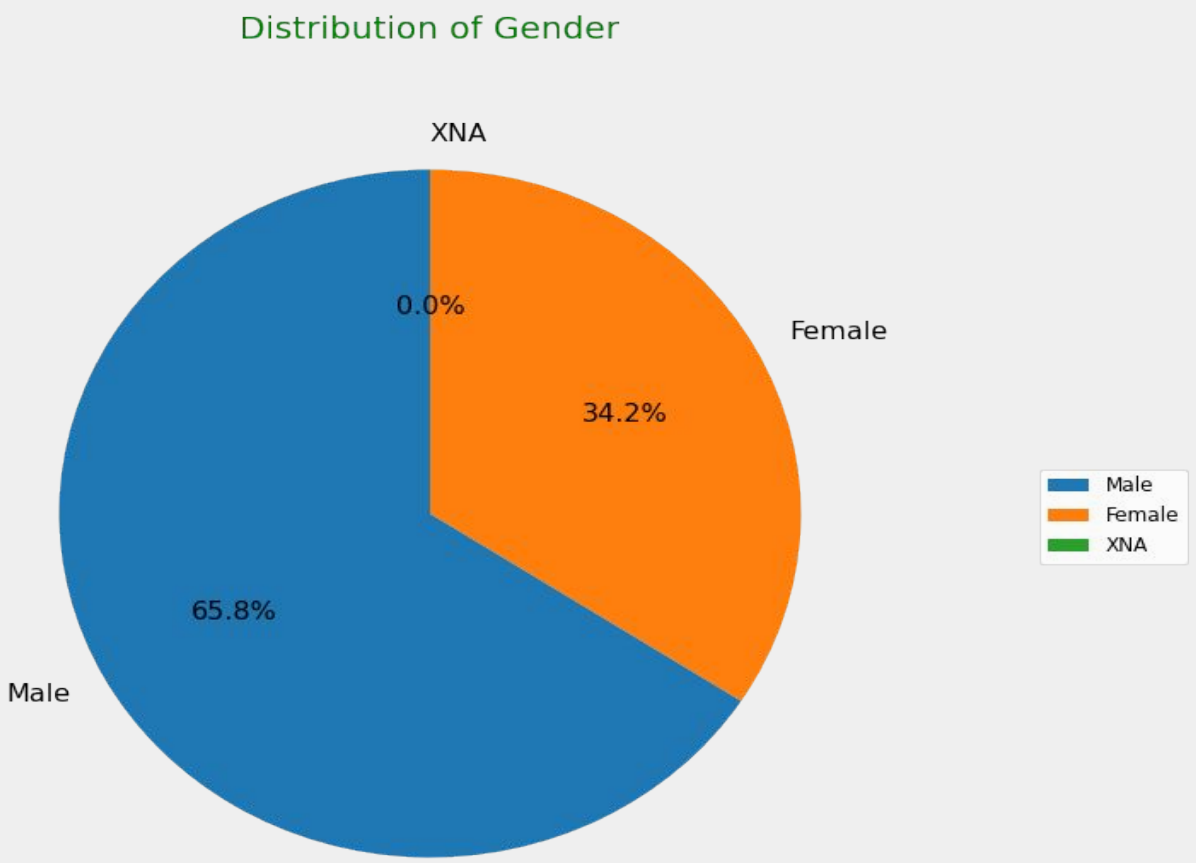
Correlation of AMT_REQ_CREDIT_XXXXX columns vs TARGET column

Observation : Based on the above Heatmap There is almost no correlation between AMT_REQ_CREDIT_XXXXX columns vs TARGET column

Univariate

Categorical Univariate Analysis

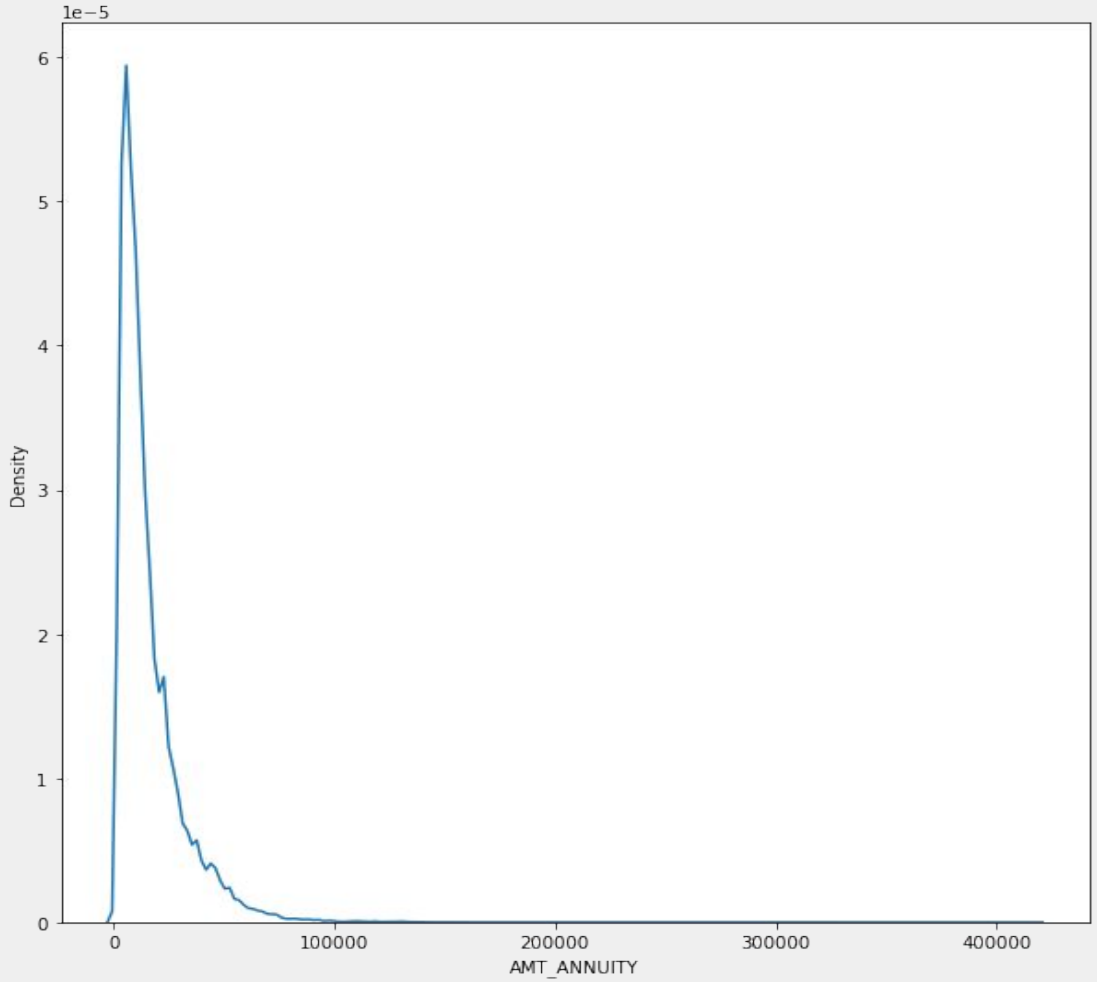
Gender Distribution



Observation : Based on the above analysis there are 4 'XNA' values in Gender_Code Column, which can be imputed by Female i.e. mode of the Gender_Code because its value is around 65% of the total records adding 4 more records won't impact our analysis

Imputing Null Values

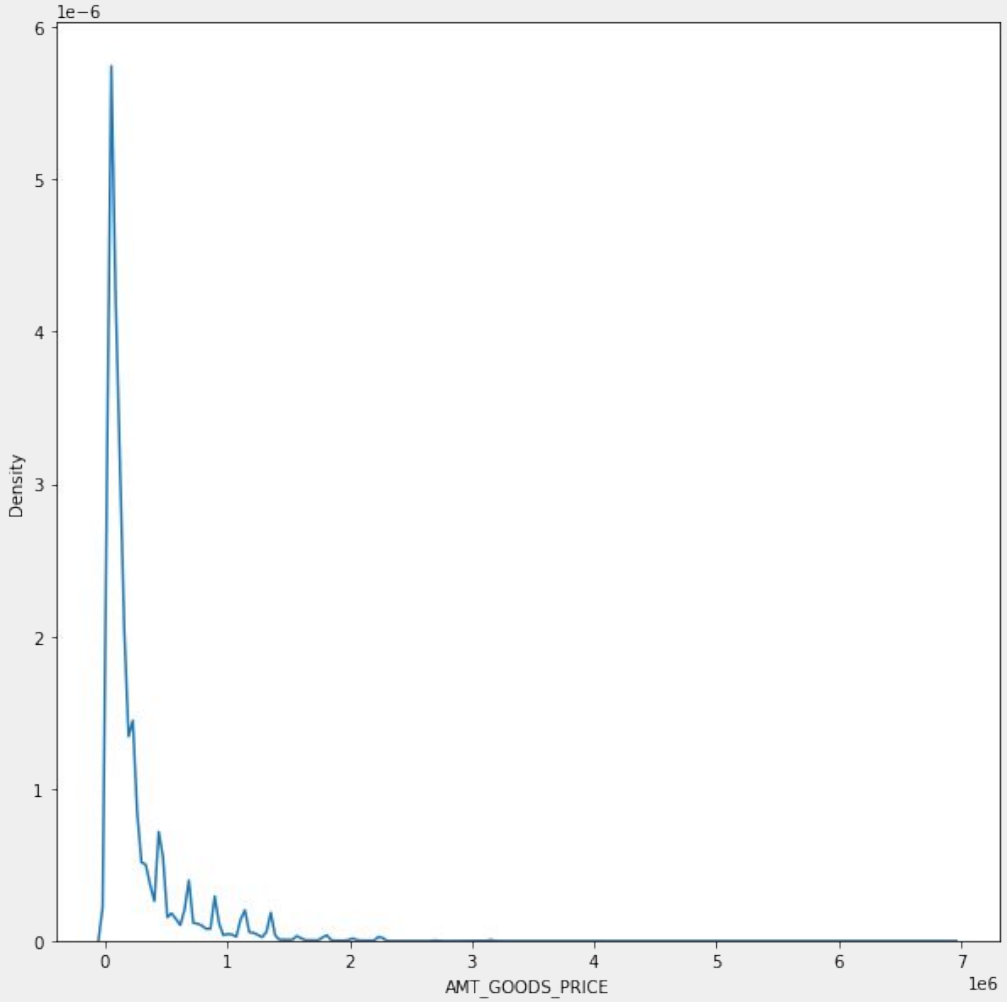
AMT_ANNUITY



Observation: 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.

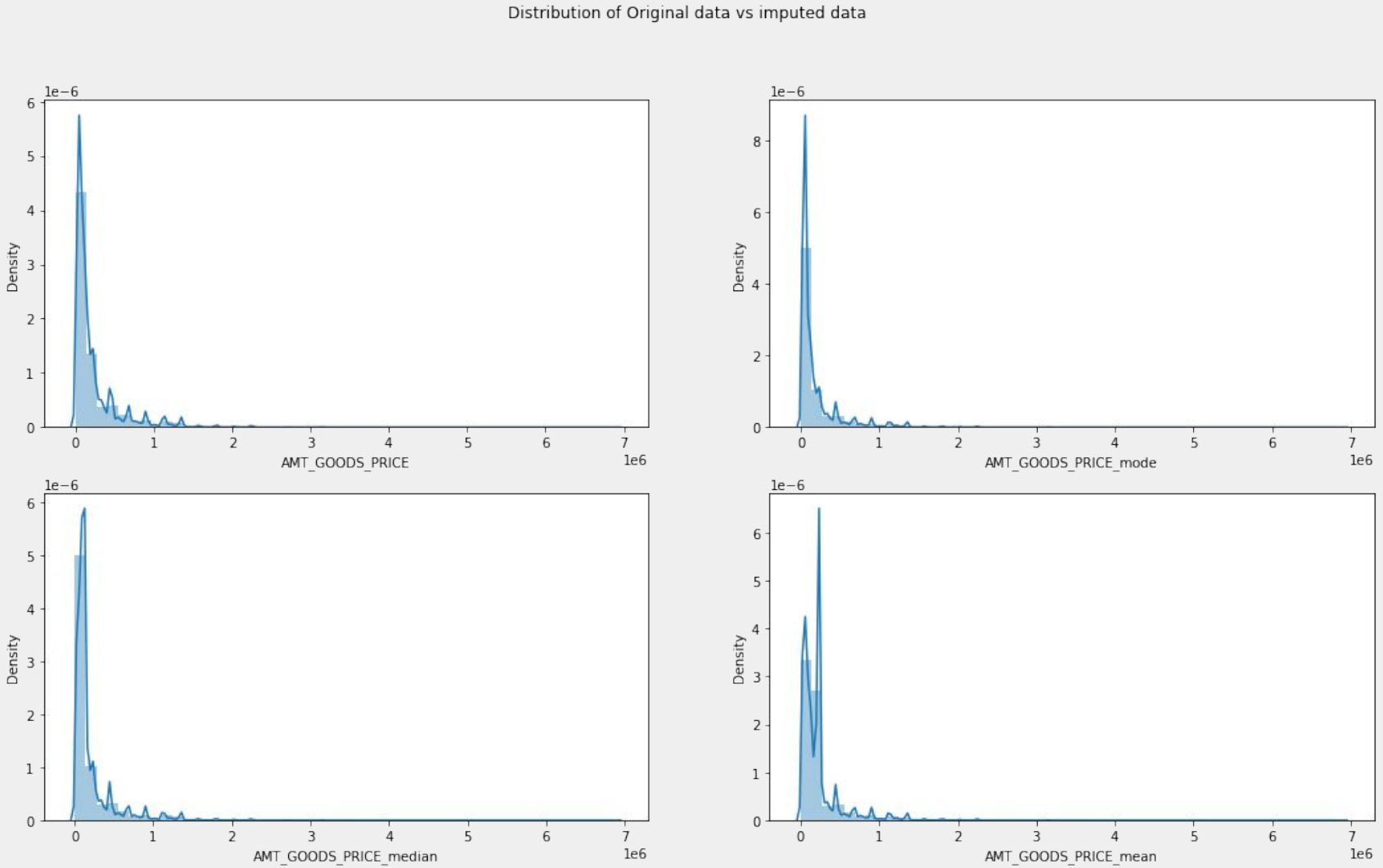
Imputing Null Values

AMT_GOODS



Observation: 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.

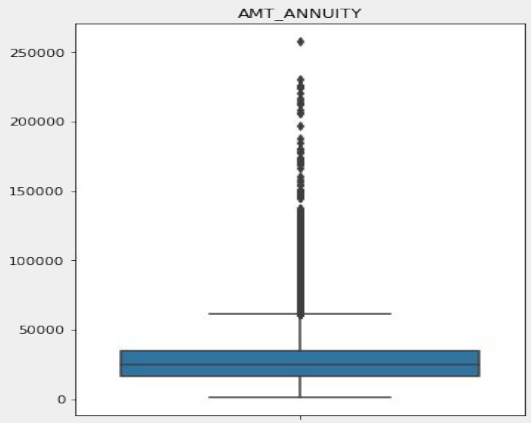
Distribution of Original Data VS Imputed data

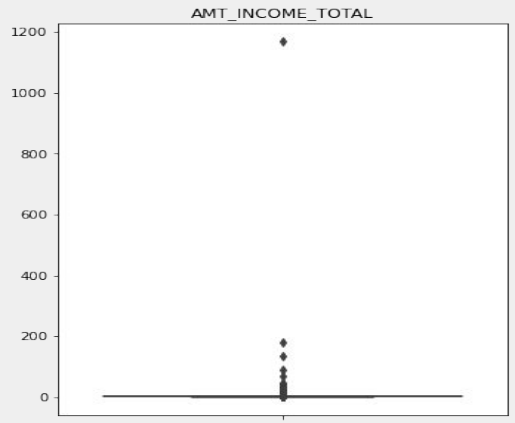


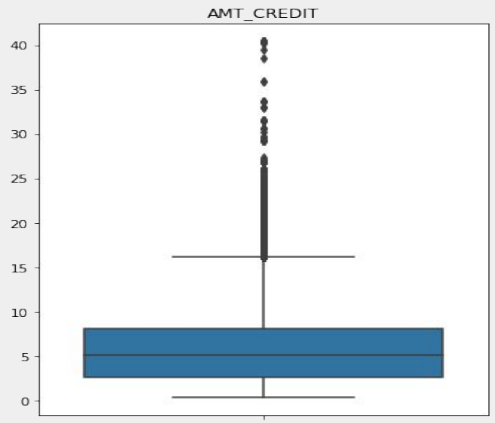
Observation:

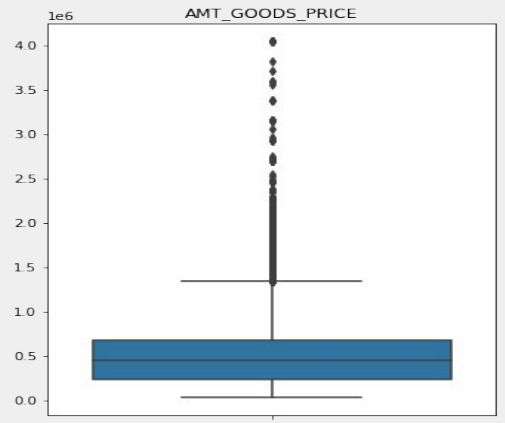
The original distribution is closer with the distribution of data imputed with mode in this case

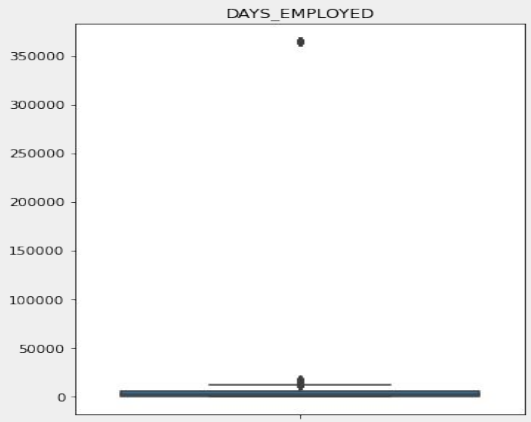
OUTLIERS ANALYSIS

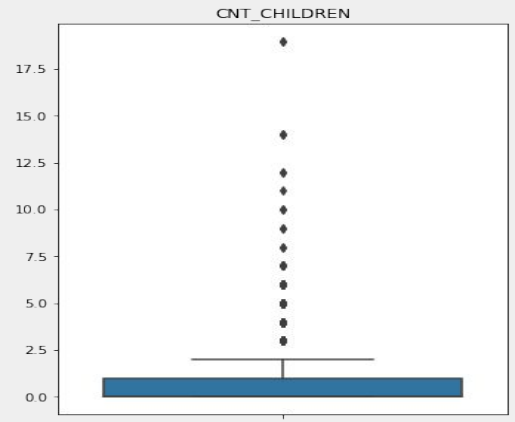


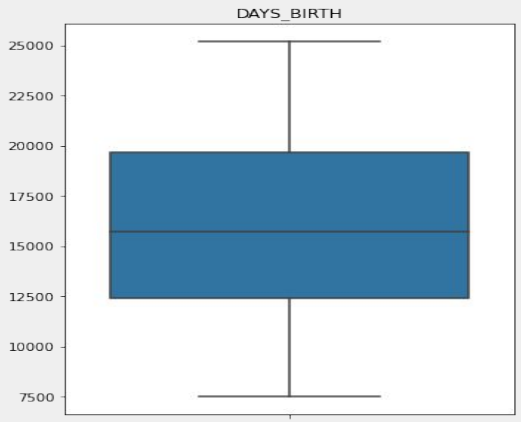








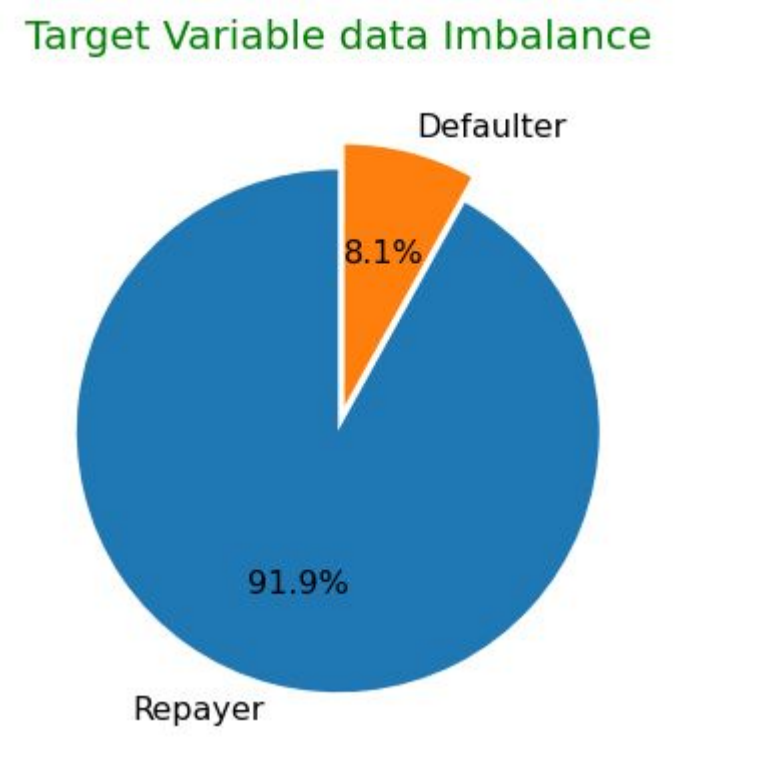


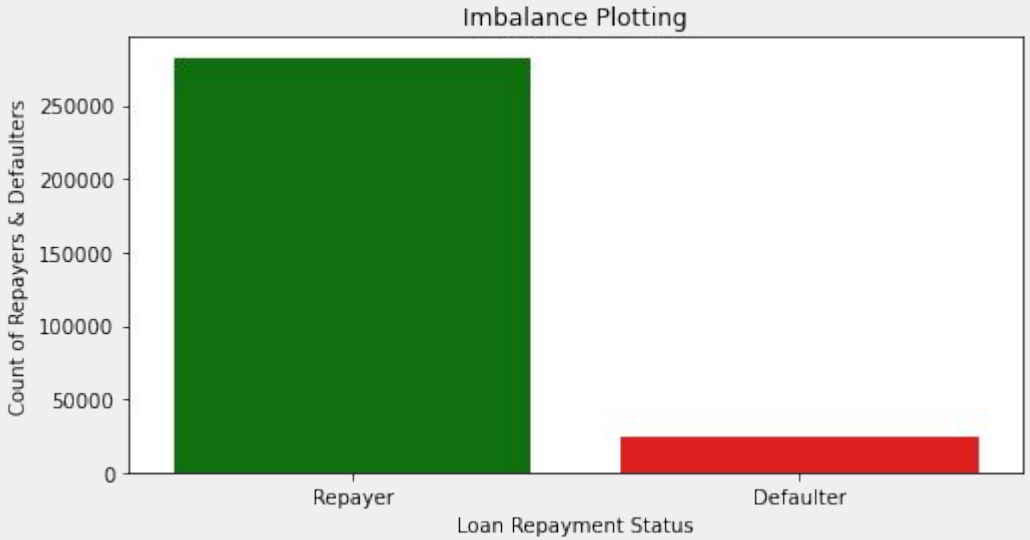


INFERENCE:-

- AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.
- AMT_INCOME_TOTAL has huge number of outliers which indicate that few of the loan applicants have high income when compared to the others.
- DAYS_BIRTH has no outliers which means the data available is reliable.
- DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and hence this has to be incorrect entry.
- We can see the stats for these columns below as well.

Data Imbalance



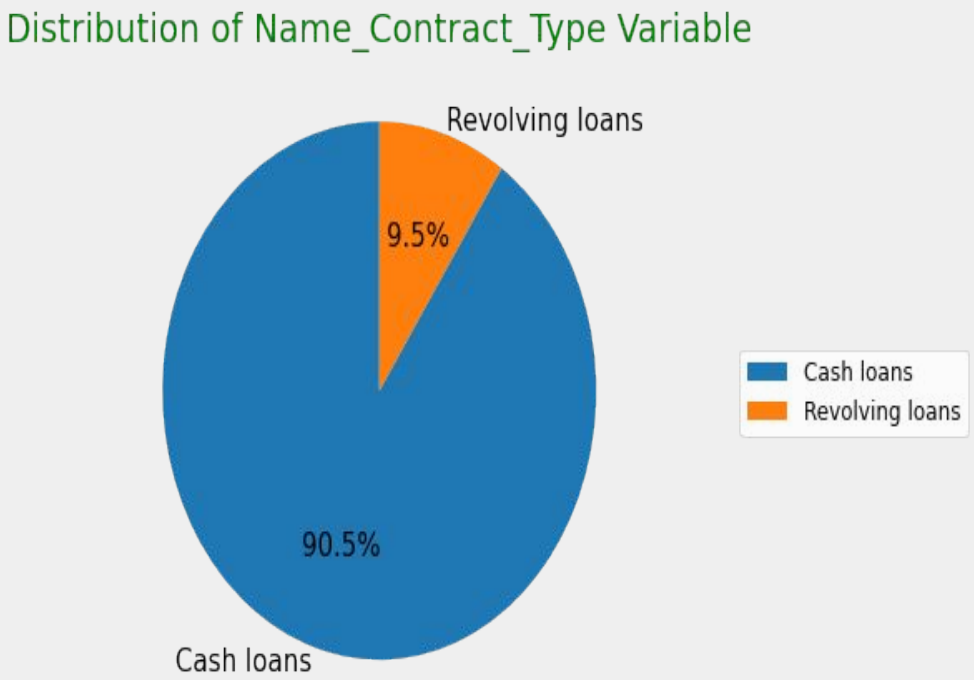


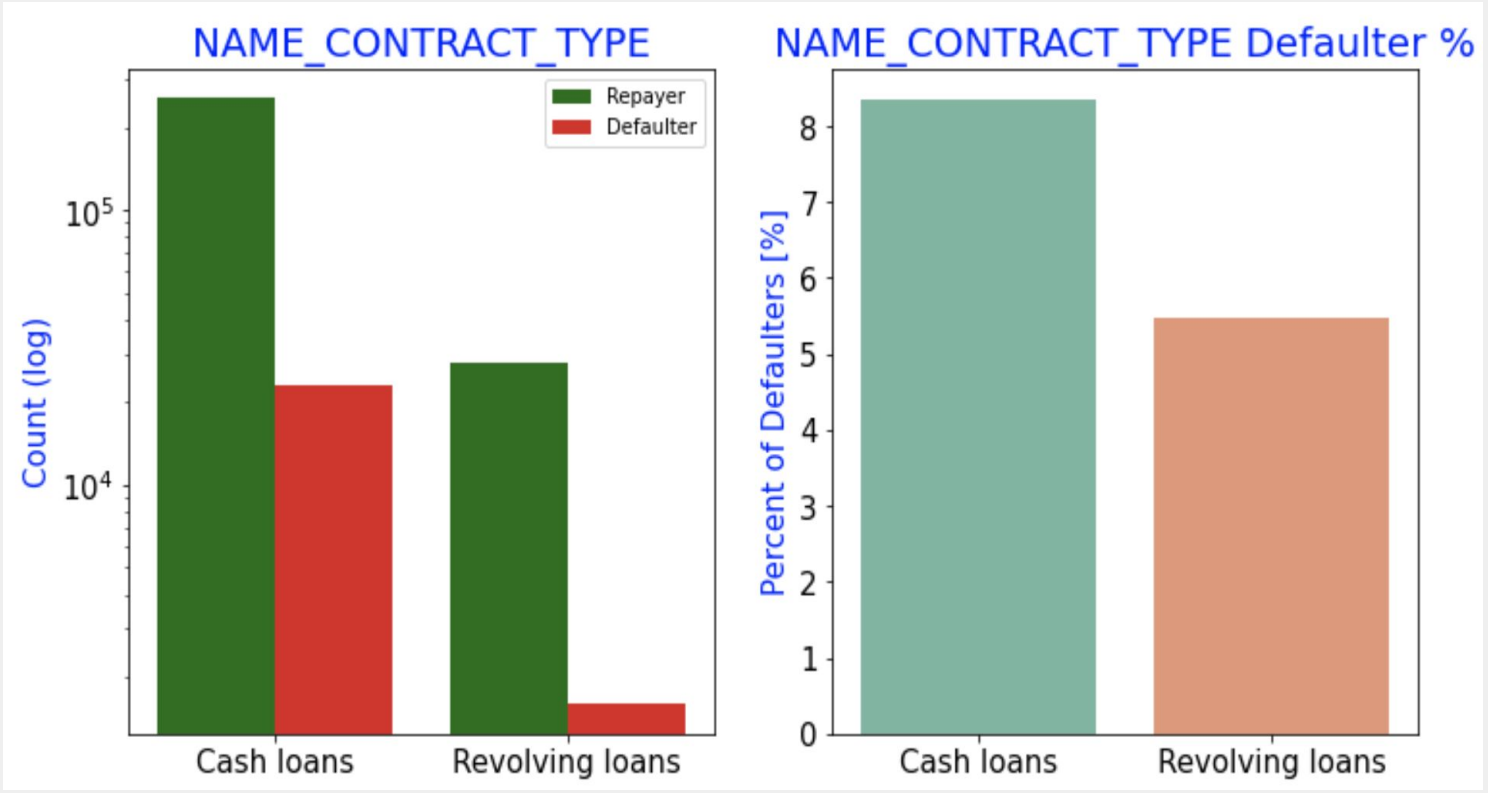
Ratios of imbalance for Repayer and Defaulter in Percentage is: 91.93 and 8.07

Ratios of imbalance for Repayer Vs Defaulter is: 11.39 :1 approx

Categorical Variables Analysis: Segmented Univariate Analysis

NAME_CONTRACT_TYPE

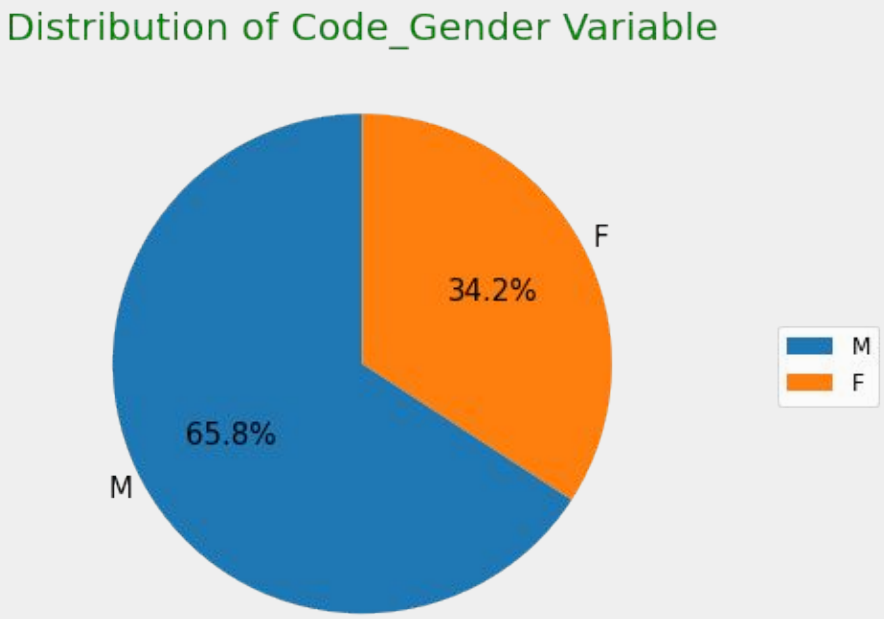


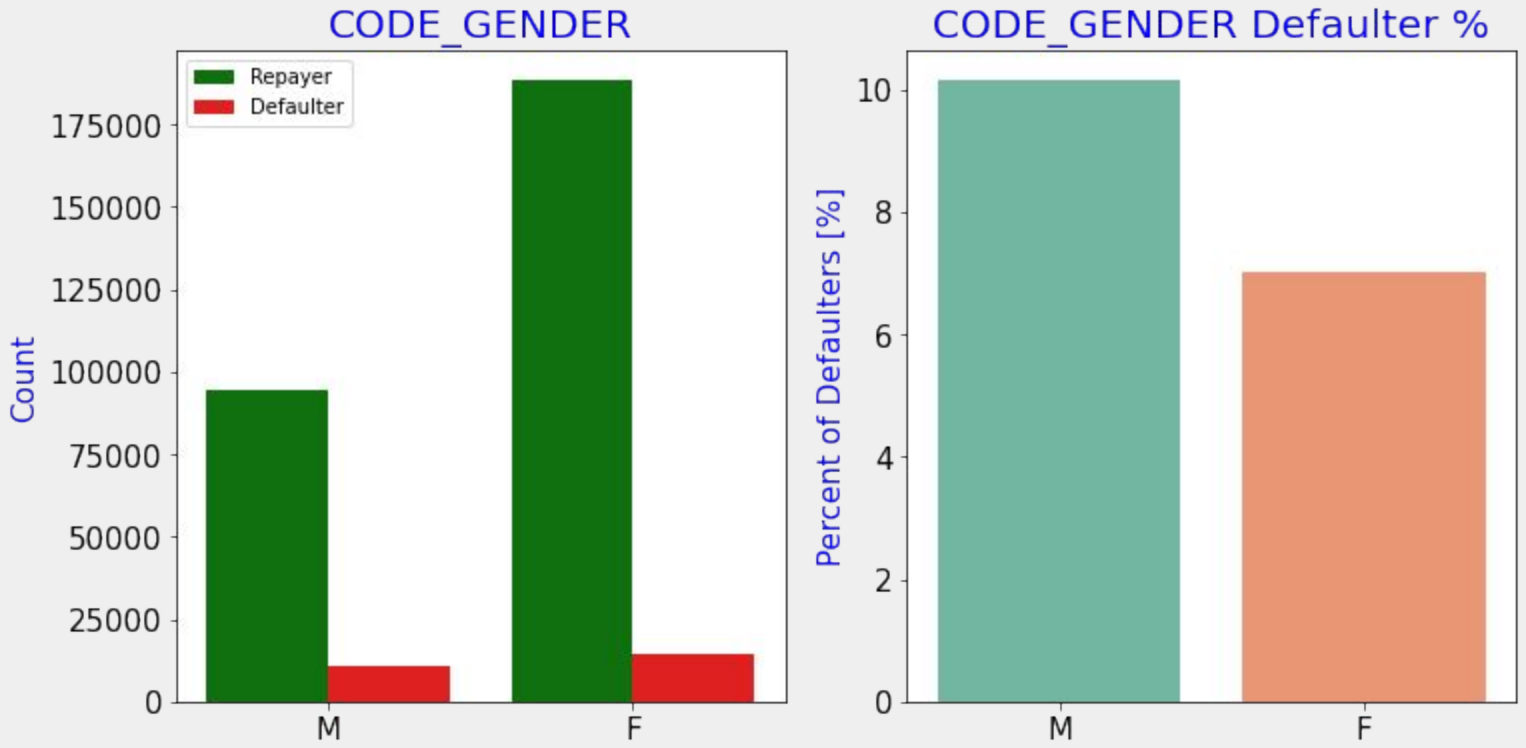


Inferences: Contract type: Revolving loans are just a small fraction (10%) from the total number of loans; Also majority of the Revolving loans have been defaulted.

Categorical Variables Analysis: Segmented Univariate Analysis

CODE_GENDER

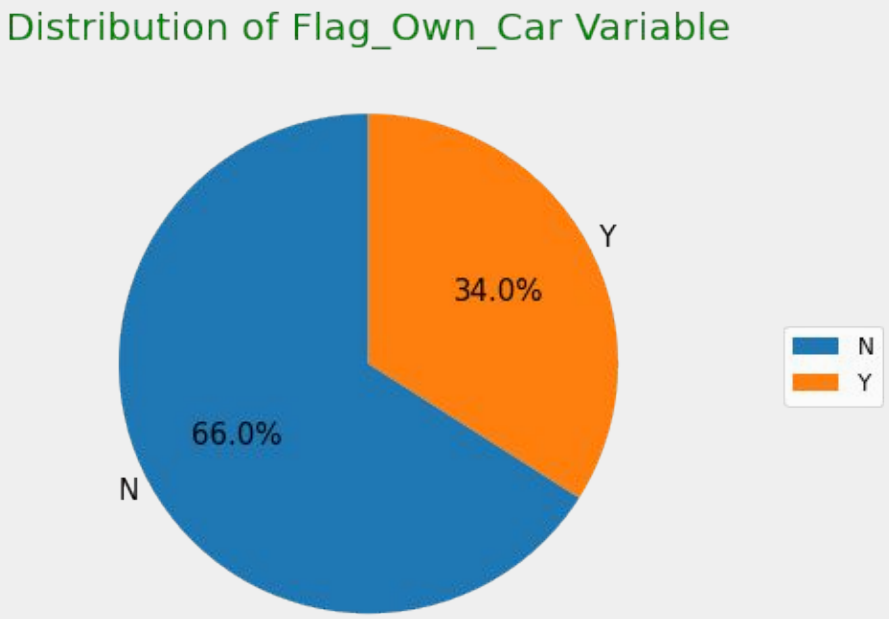


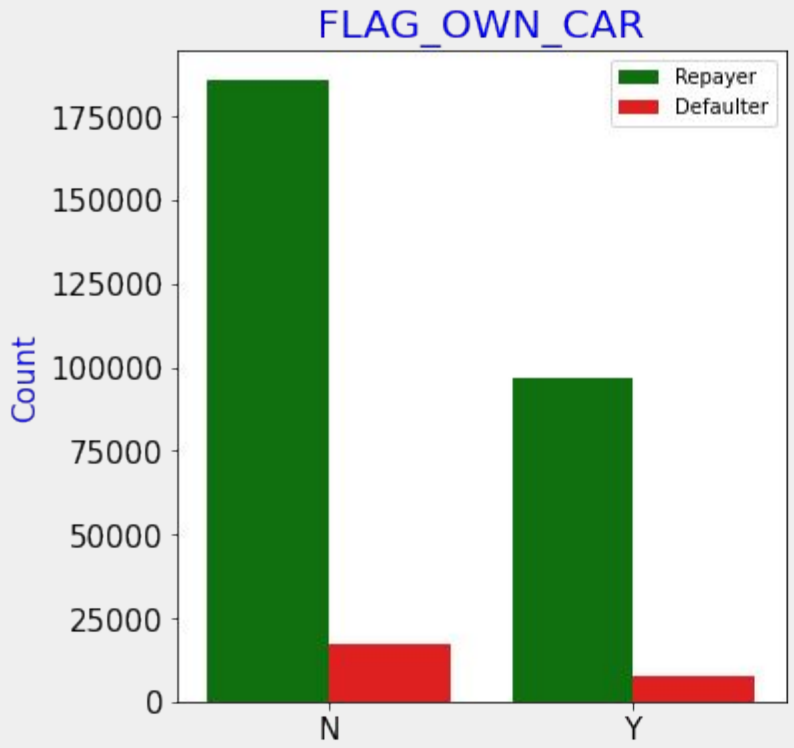


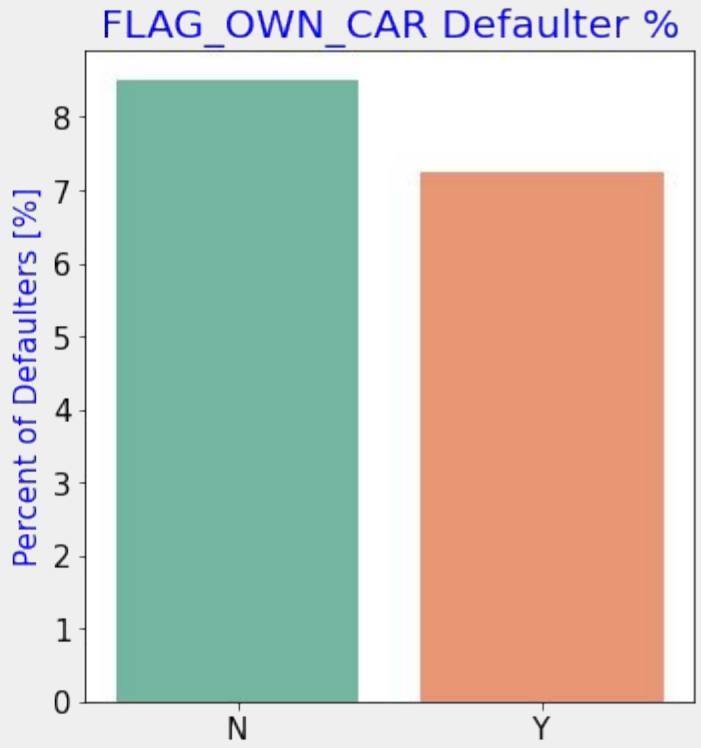
Inferences: The number of female clients is almost double the number of male clients. Based on the percentage of defaulted credits, males have a higher chance of not returning their loans (approx 10%), comparing with women (~7%)

Categorical Variables Analysis: Segmented Univariate Analysis

FLAG_OWN_CAR



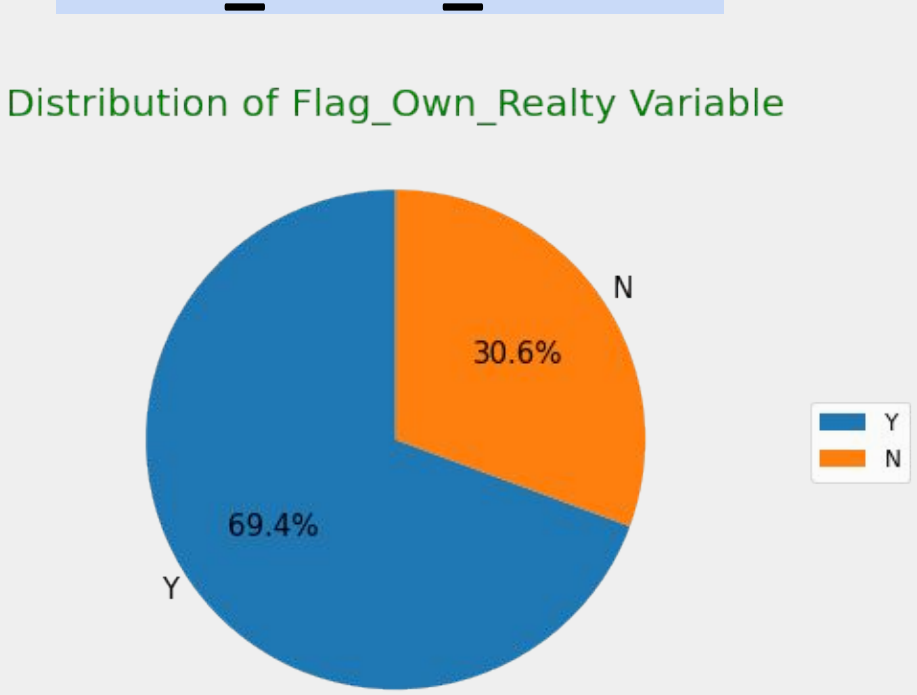


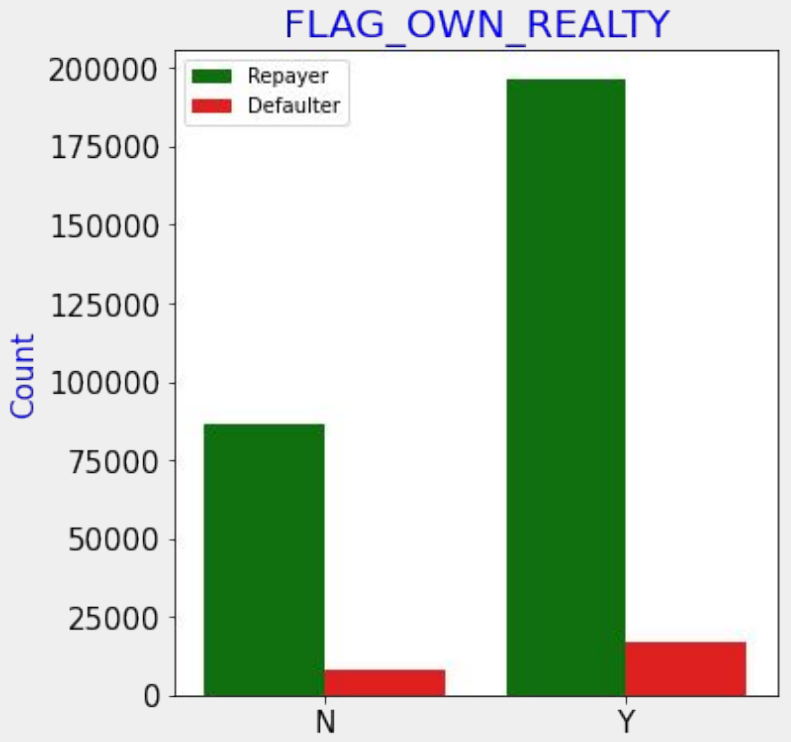


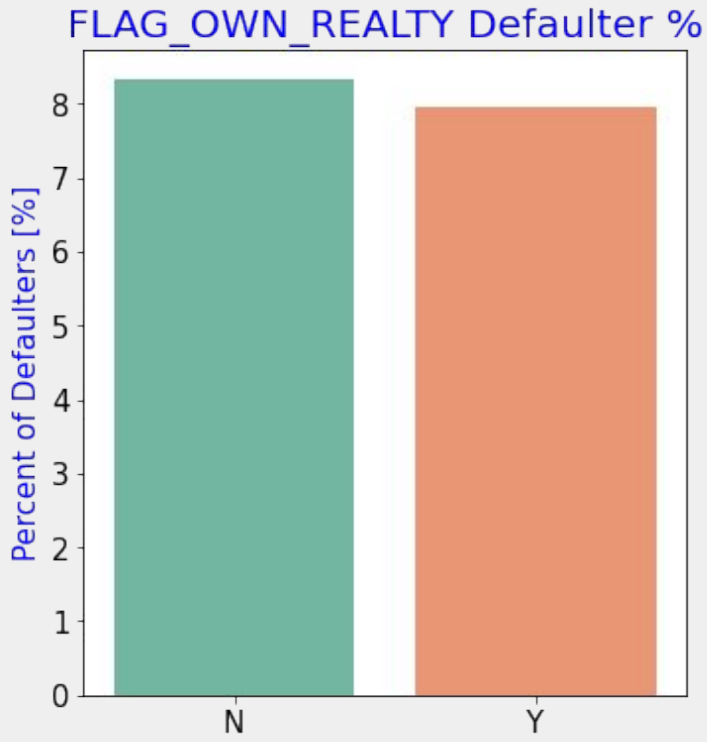
Inferences: The number of female clients is almost double the number of male clients. Based on the percentage of defaulted credits, males have a higher chance of not returning their loans (approx 10%), comparing with women (~7%)

Categorical Variables Analysis: Segmented Univariate Analysis

FLAG_OWN_REALTY

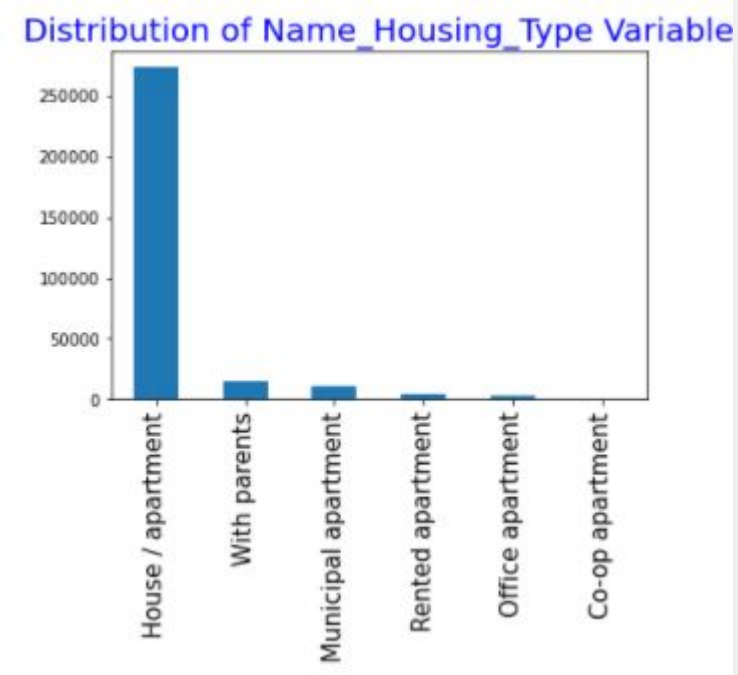




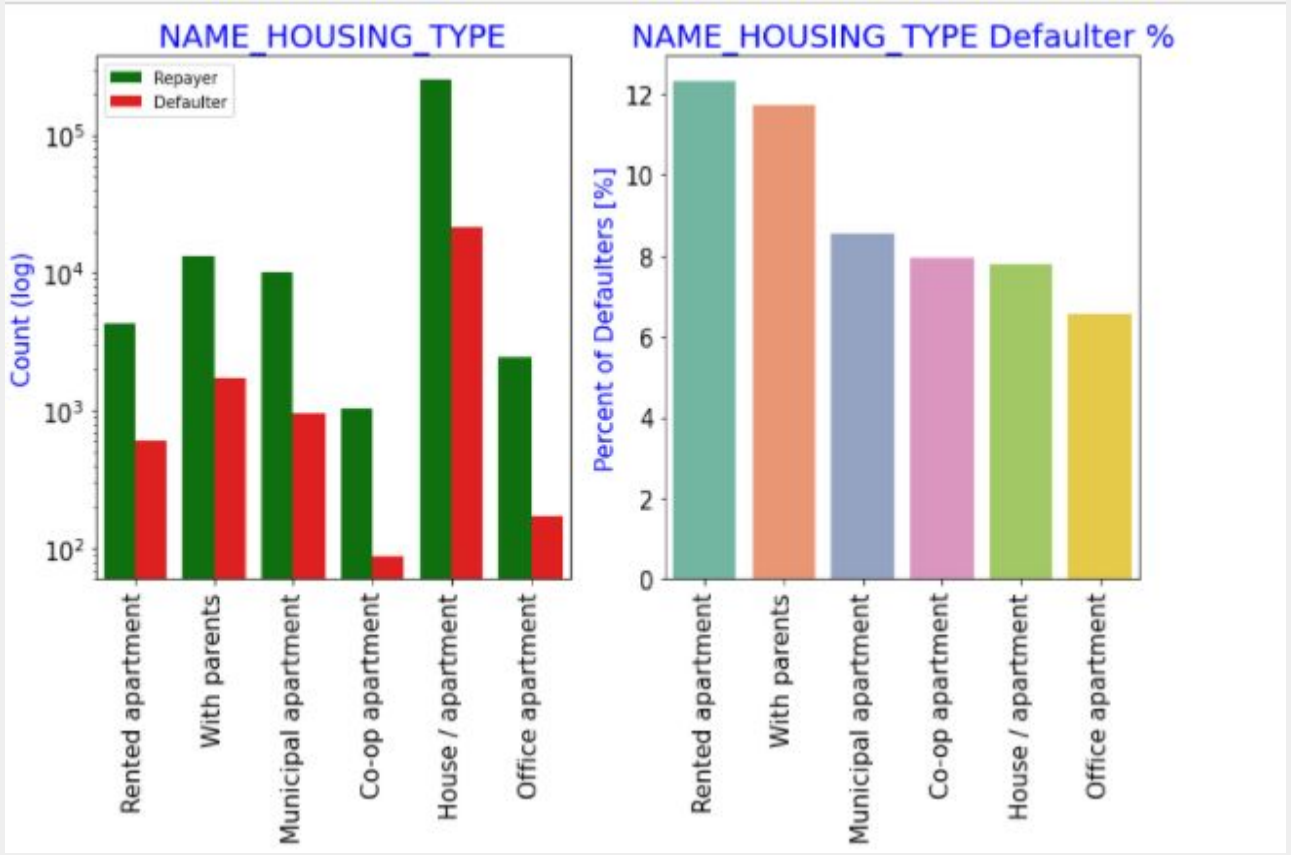


Inferences: The clients who own real estate are more than double of the ones that don't own. But the defaulting rate of both categories are around the same (~8%). Thus there is no correlation between owning a reality and defaulting the loan.

Housing Type



Majority of people live in House/apartment

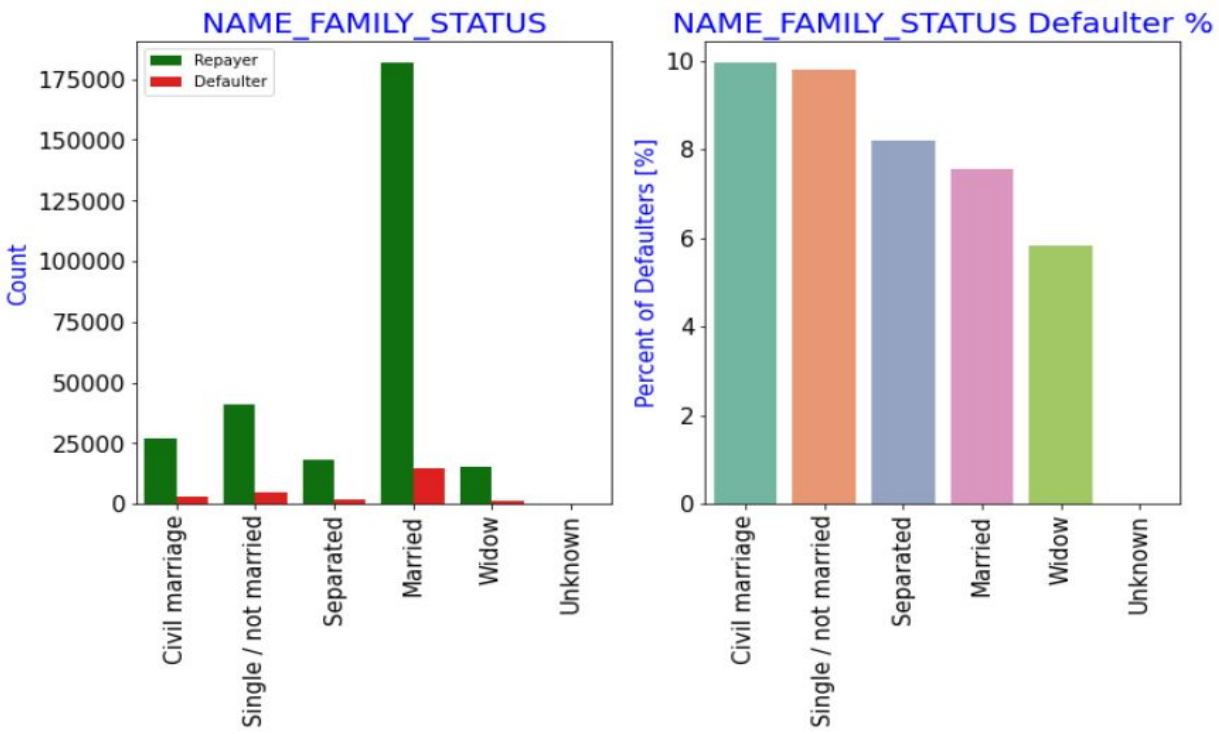


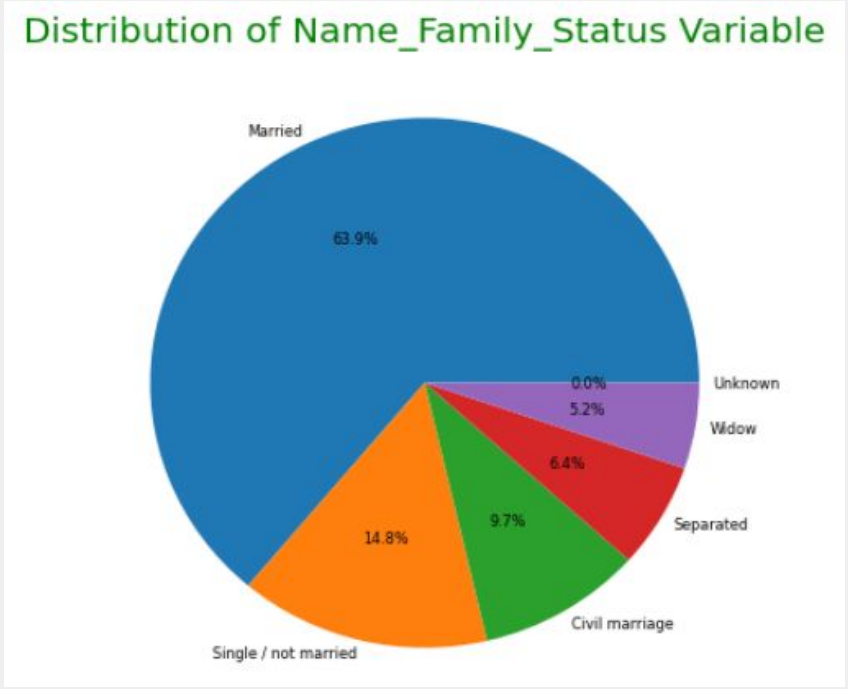
People living in office apartments have lowest default rate
- People living with parents (around 11.5%) and living in rented apartments(> 12%) have higher probability of defaulting

Inferences:

- Majority of people live in House/apartment
- People living in office apartments have lowest default rate
- People living with parents (around 11.5%) and living in rented apartments(> 12%) have higher probability of defaulting

FAMILY/MARITAL STATUSF

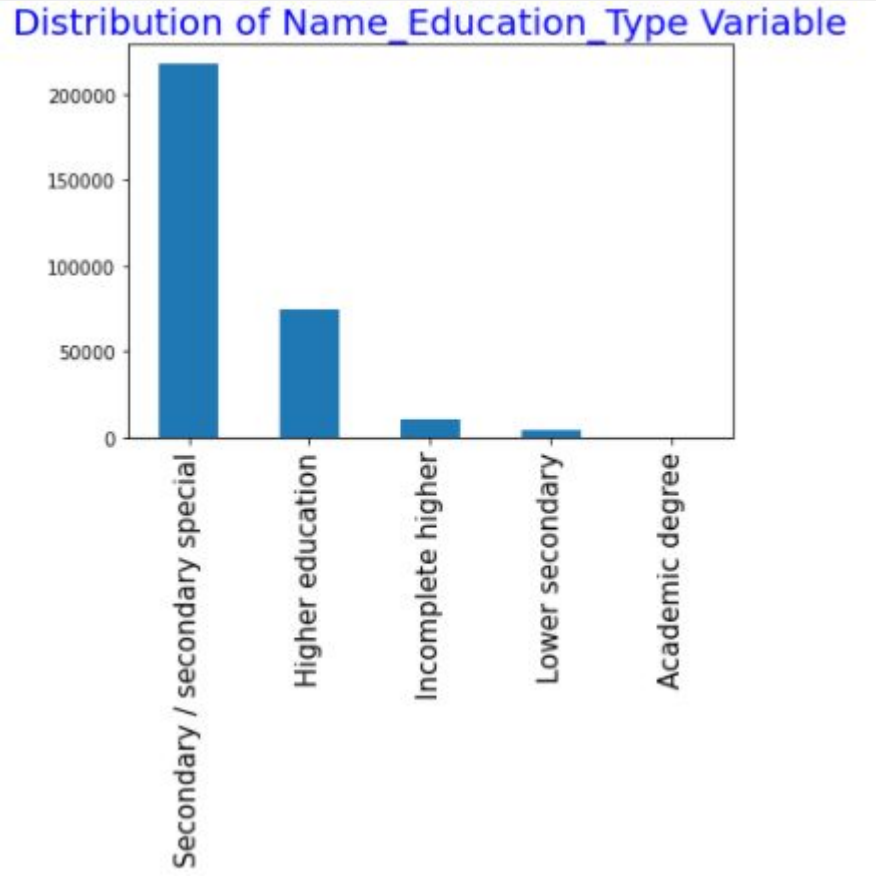


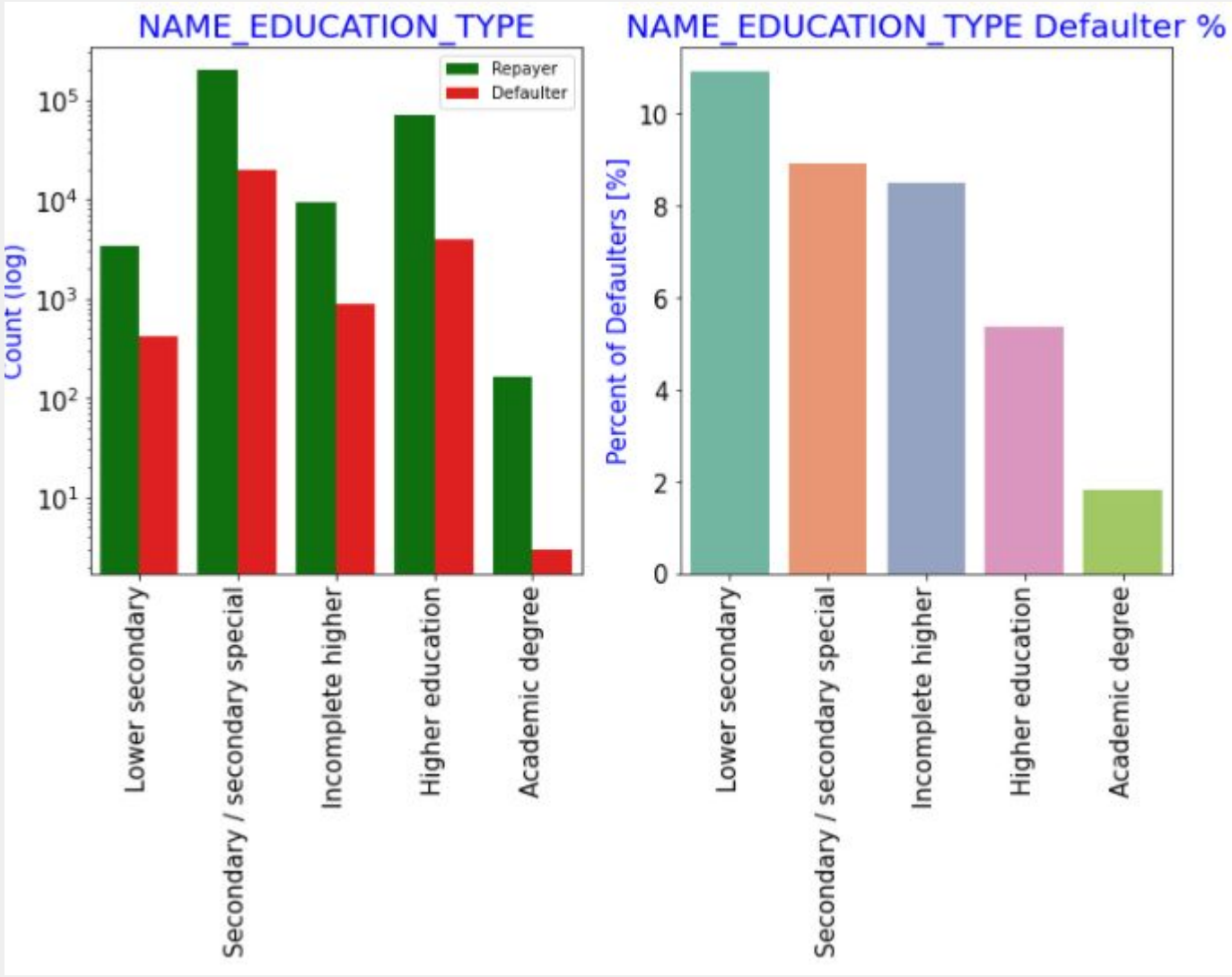


Inferences:

- Most of the people who have taken loan are married, followed by Single/not married and civil marriage
- In terms of percentage of not repayment of loan, Civil marriage has the highest percent of not repayment (10%), with Widow the lowest (exception being Unknown).

Education Type

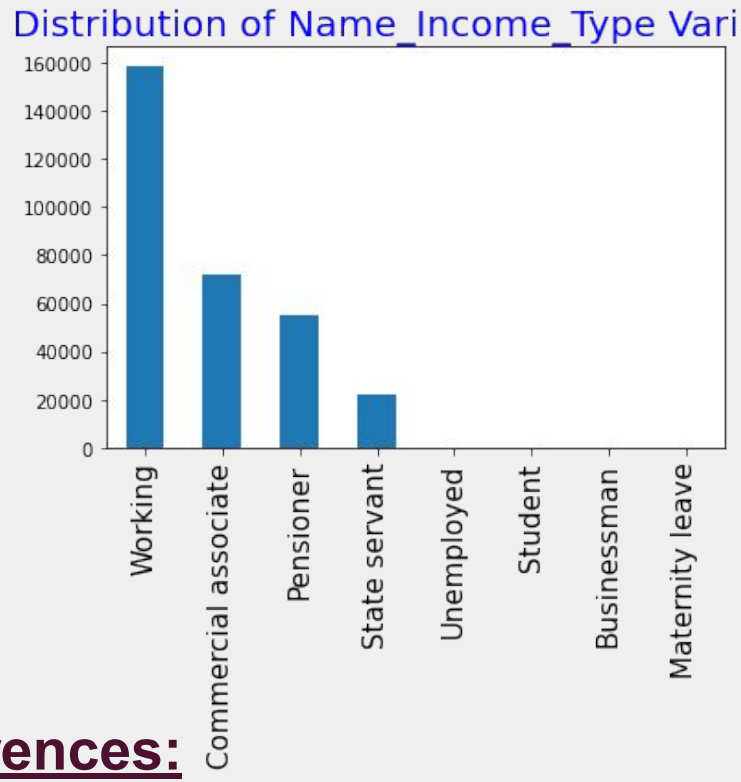


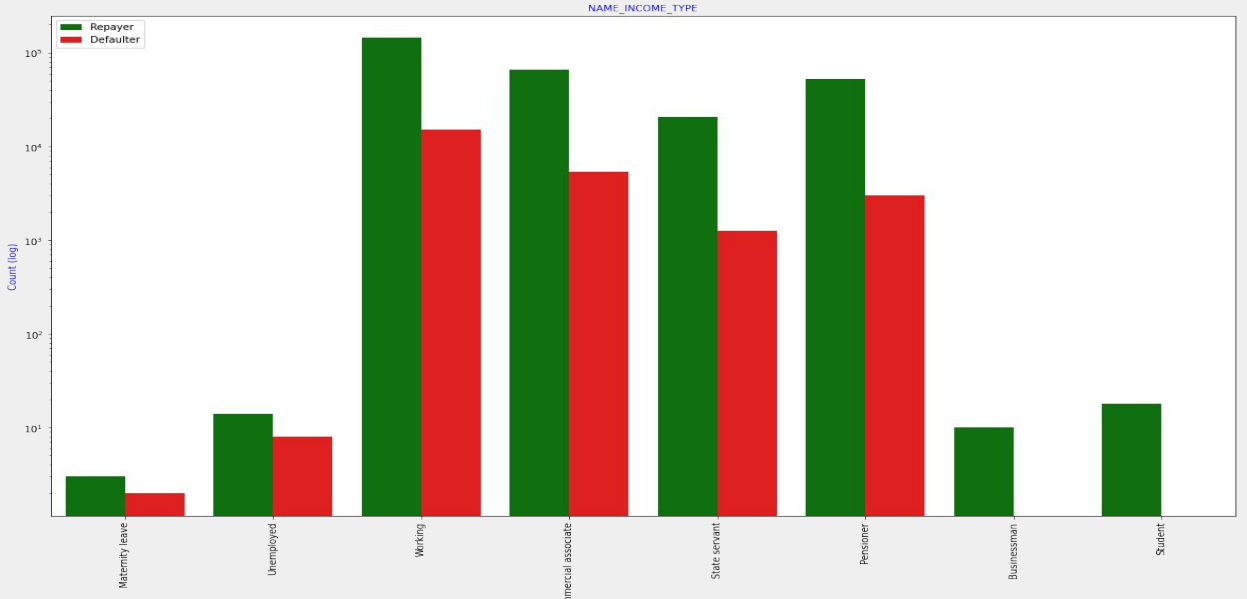


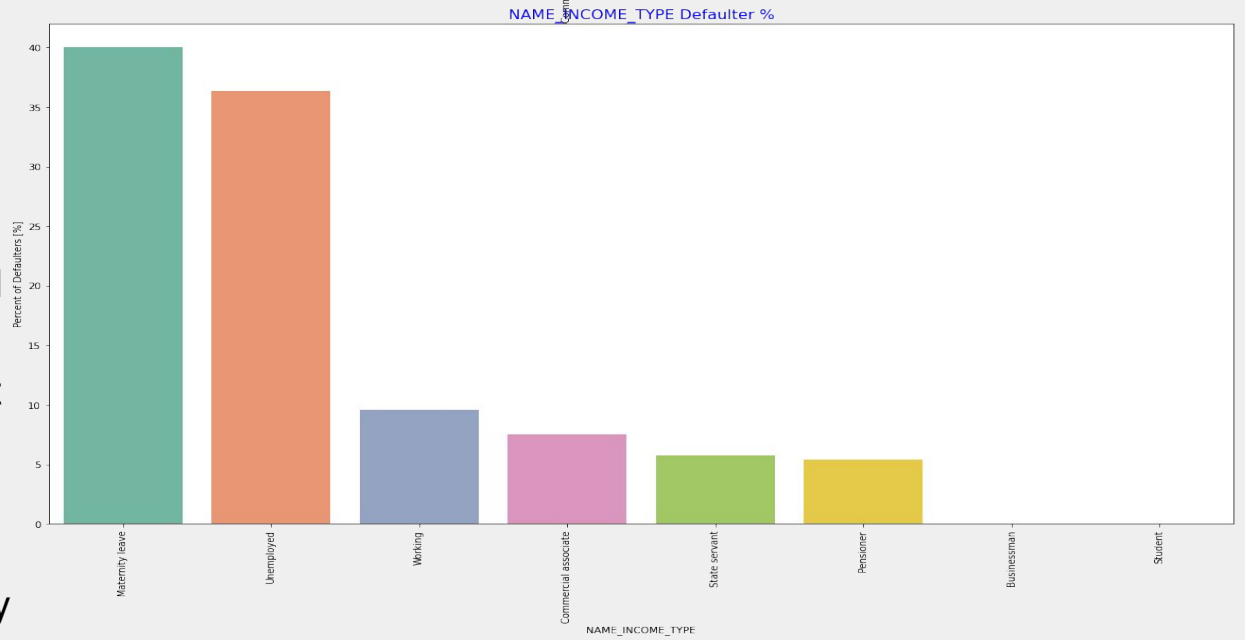
Inferences:

- Majority of the clients have Secondary / secondary special education, followed by clients with Higher education. Only a very small number having an academic degree
- The Lower secondary category, although rare, have the largest rate of defaulters (11%). The people with Academic degree have the lowest defaulting rate(around 2%).

Name_Income_Type



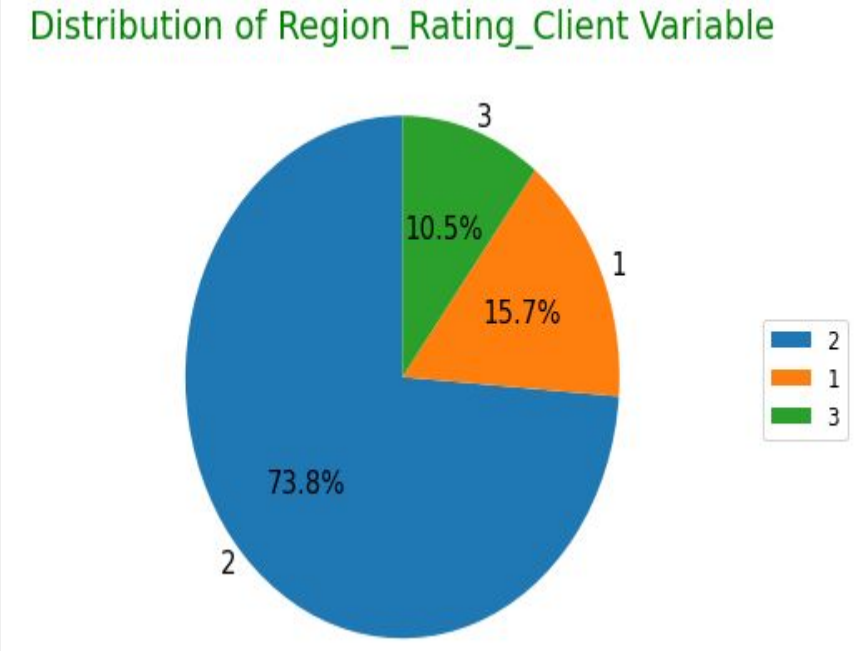




Inferences:

- Most of applicants for loans have income type as Working, followed by Commercial associate, Pensioner and State servant.
- The applicants with the type of income Maternity leave have almost make 40% ratio of the defaulters, followed by Unemployed (37%). The rest of types of incomes are under the average of 10% for not returning loans.
- Student and Businessmen, though less in numbers do not have any default record. Thus these two category are safest for providing loan.

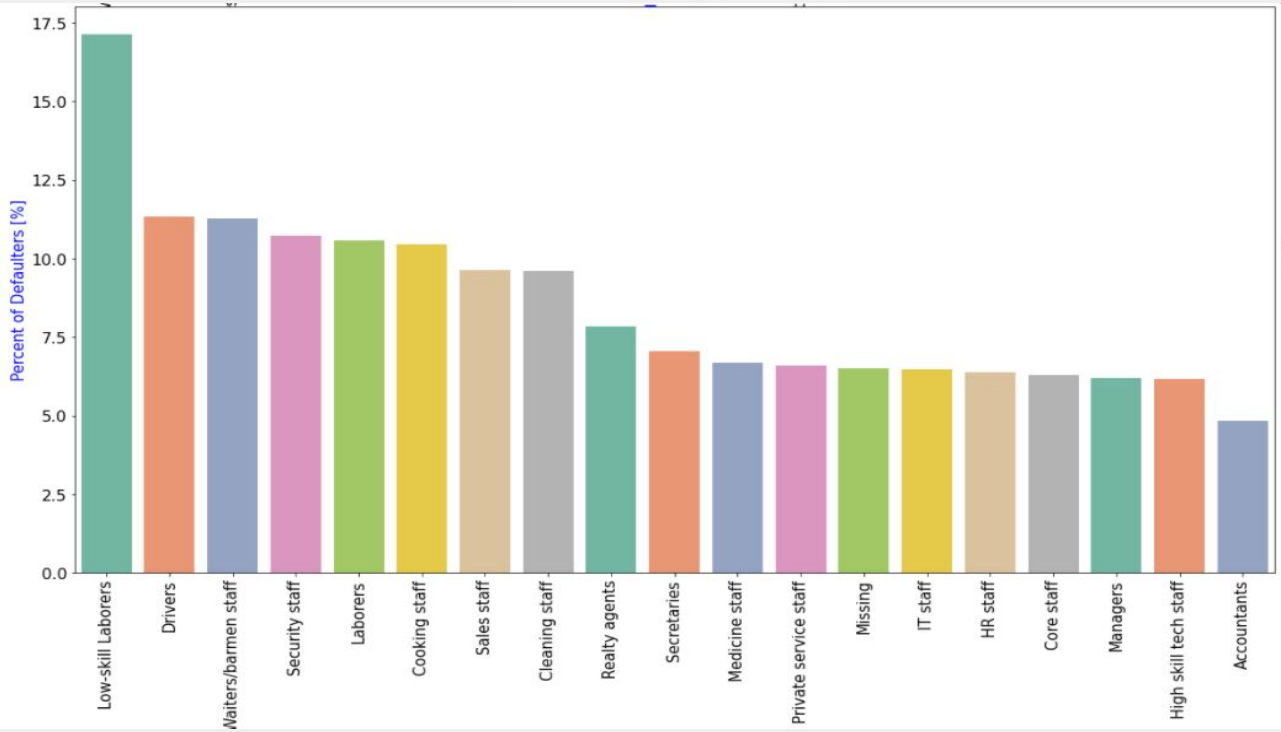
Region_Type

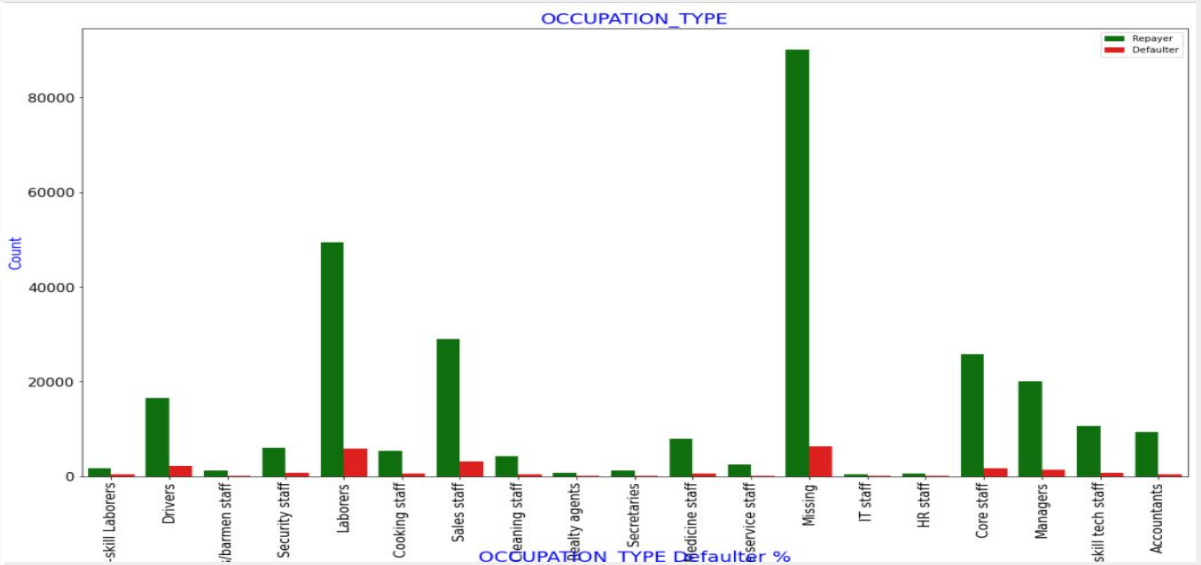


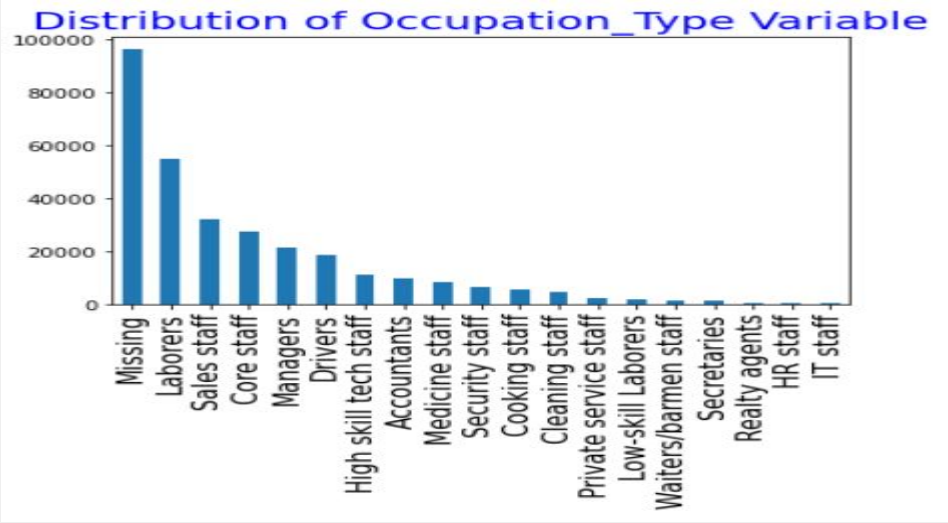
Inferences:

- Most of the applicants are living in Region_Rating 2 place.
- Region Rating 3 has the highest default rate (11%) , followed by 2(around 8%) and 1(around 5%)
- Applicant living in Region_Rating 1 has the lowest probability of defaulting, thus safer for approving loans

Occupation_Type



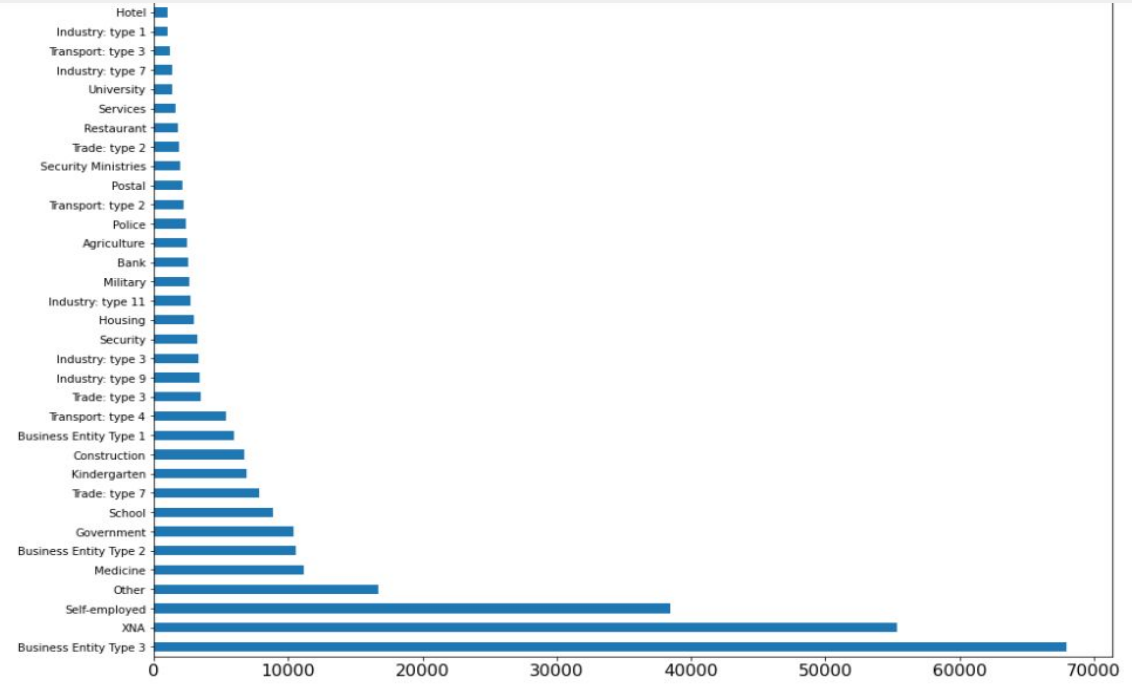


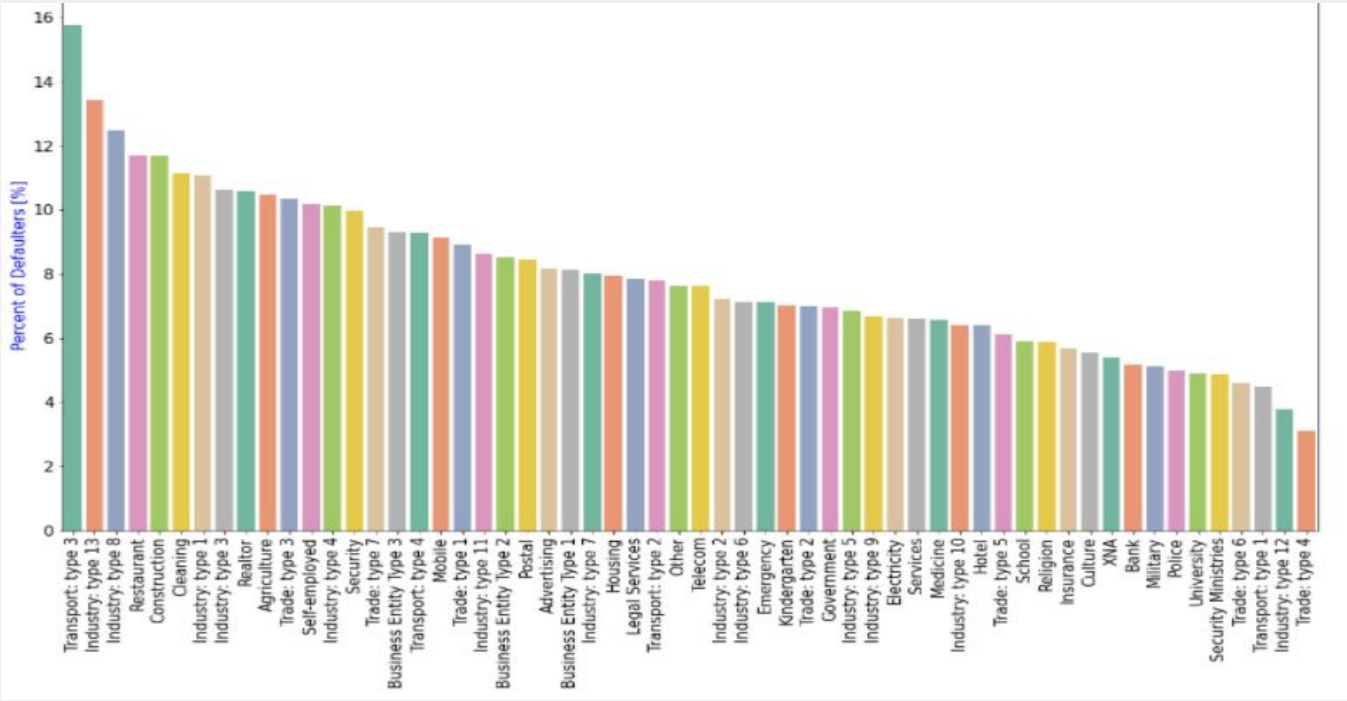


Inferences:

- Most of the loans are taken by people whose Occupation is "Missing" in the dataset followed by Laborers, Sales staff. IT staff take the lowest amount of loans.
- The category with highest percent of not repaid loans are Low-skill Laborers (above 17%), followed by Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff.

Organization_Type

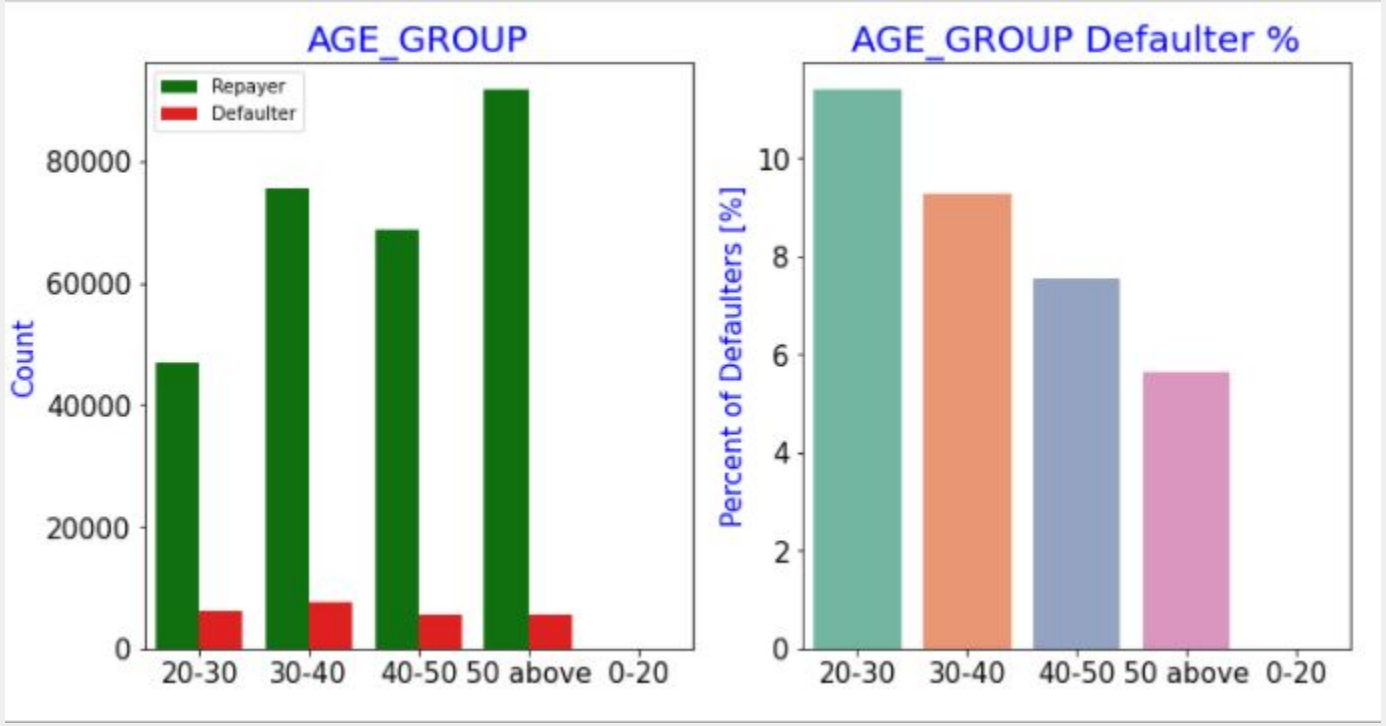


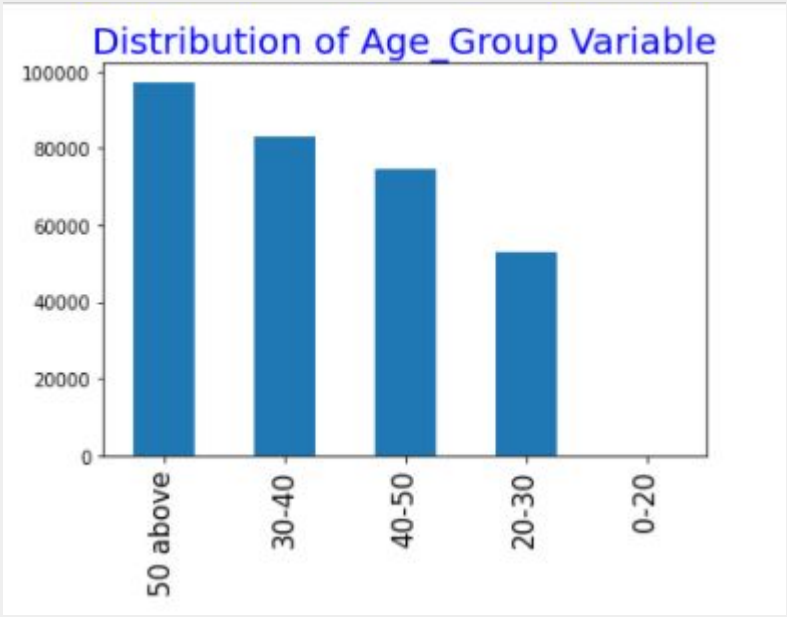


Inferences:

- Most of the applications for loan are from people working in Business Entity Type 3 organization
- Organizations with highest percent of loans not repaid are Transport: type 3 (around 16%), Industry:
- Type 13 (13.5%), Industry: type 8 (around 12.5%) and Restaurant (less than 12%).
- For a very high number of applications, Organization type information is unavailable(XNA)
- It can be seen that following category of organization type has lesser defaulters thus safer for providing loans:
  - Trade Type 4
  - Industry type 12

Age Group

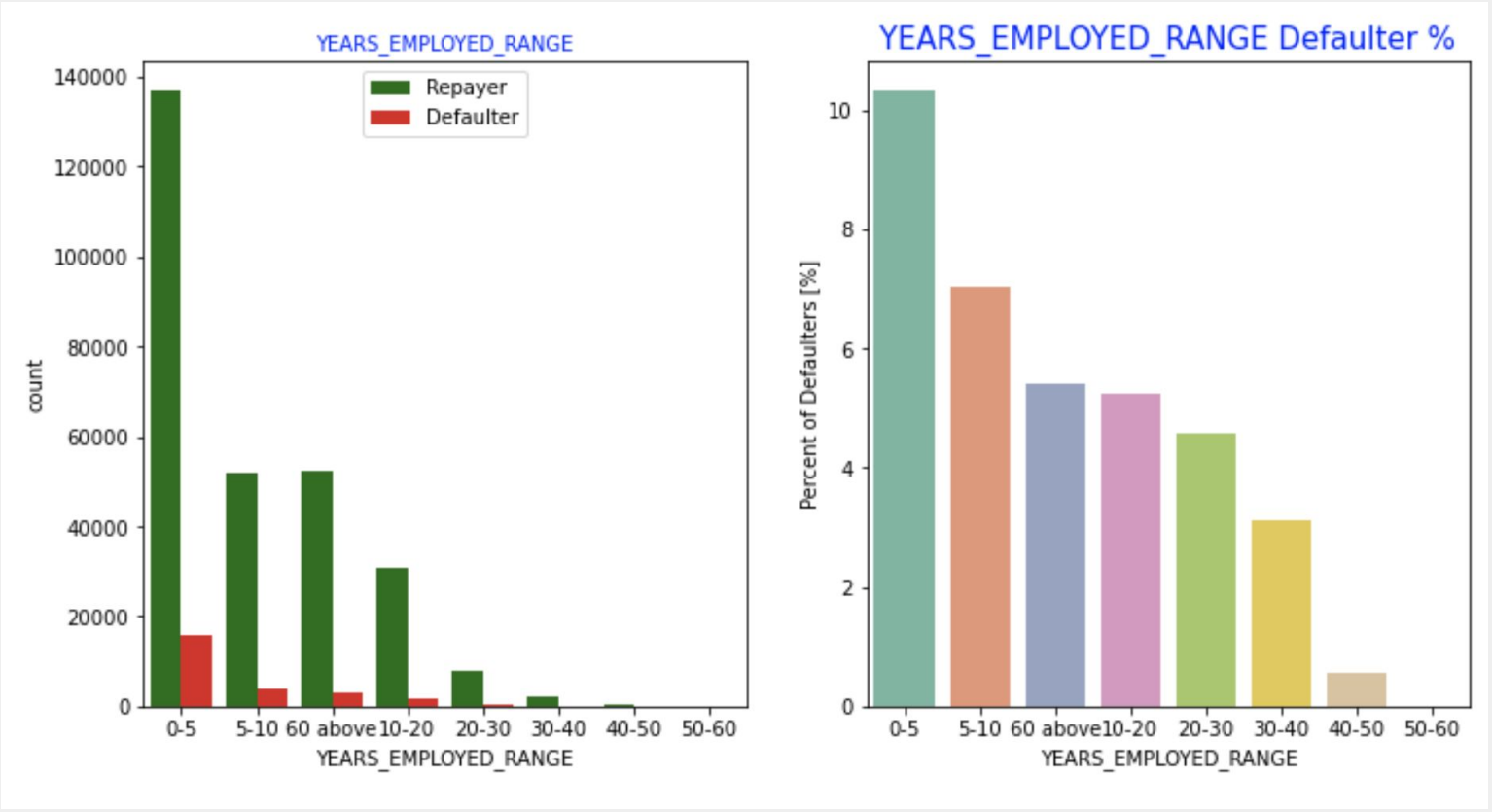


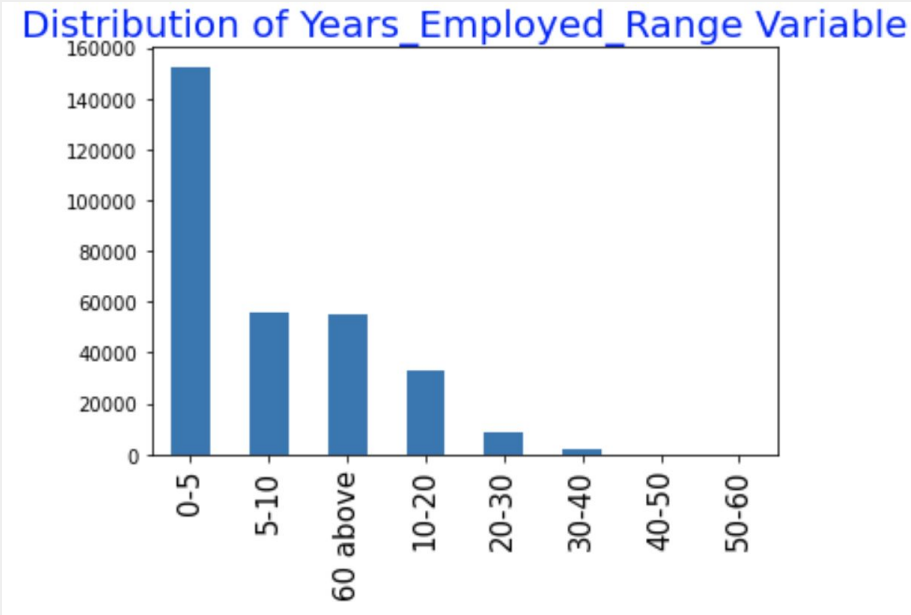


Inferences:

- People in the age group range 20-40 have higher probability of defaulting
- People above age of 50 have low probability of defaulting

Years_Employed_Range

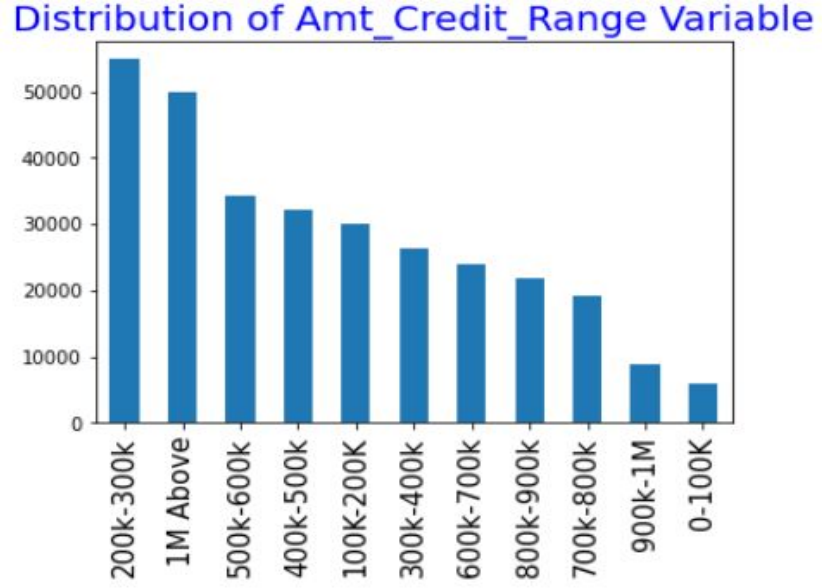


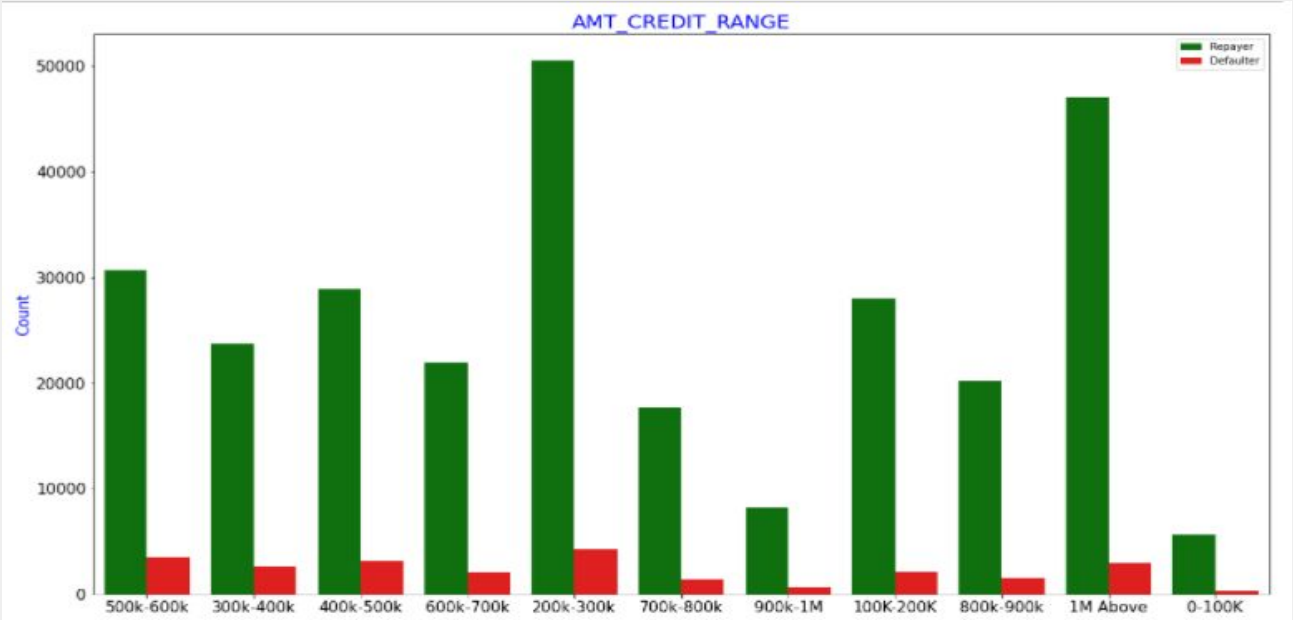


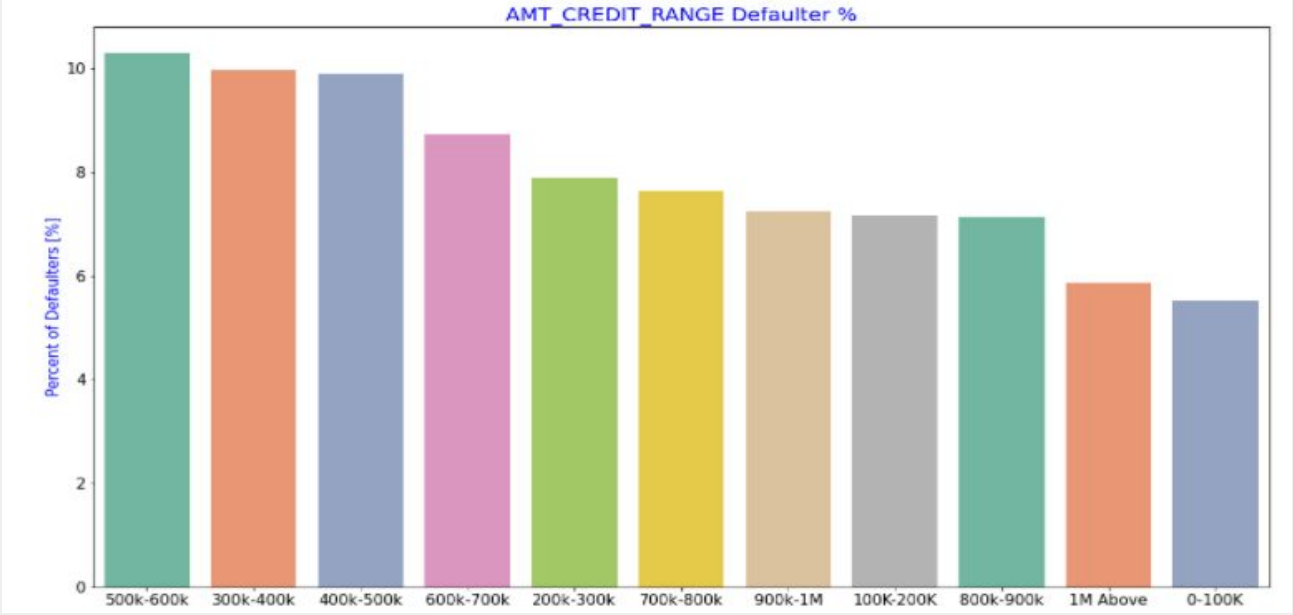
Inferences:

- Majority of the applicants have been employed in between 0-5 years. The defaulting rating of this group is also the highest which is 10%
- With increase of employment year, defaulting rate is gradually decreasing with people having 40+ year experience having less than 1% default rate

Credit Amount



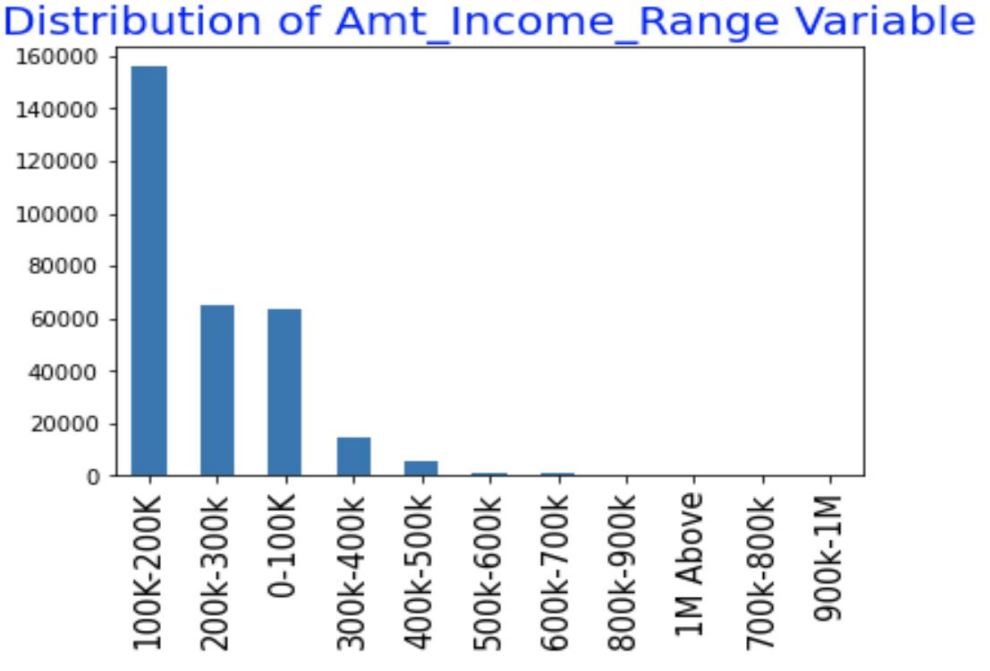




Inferences:

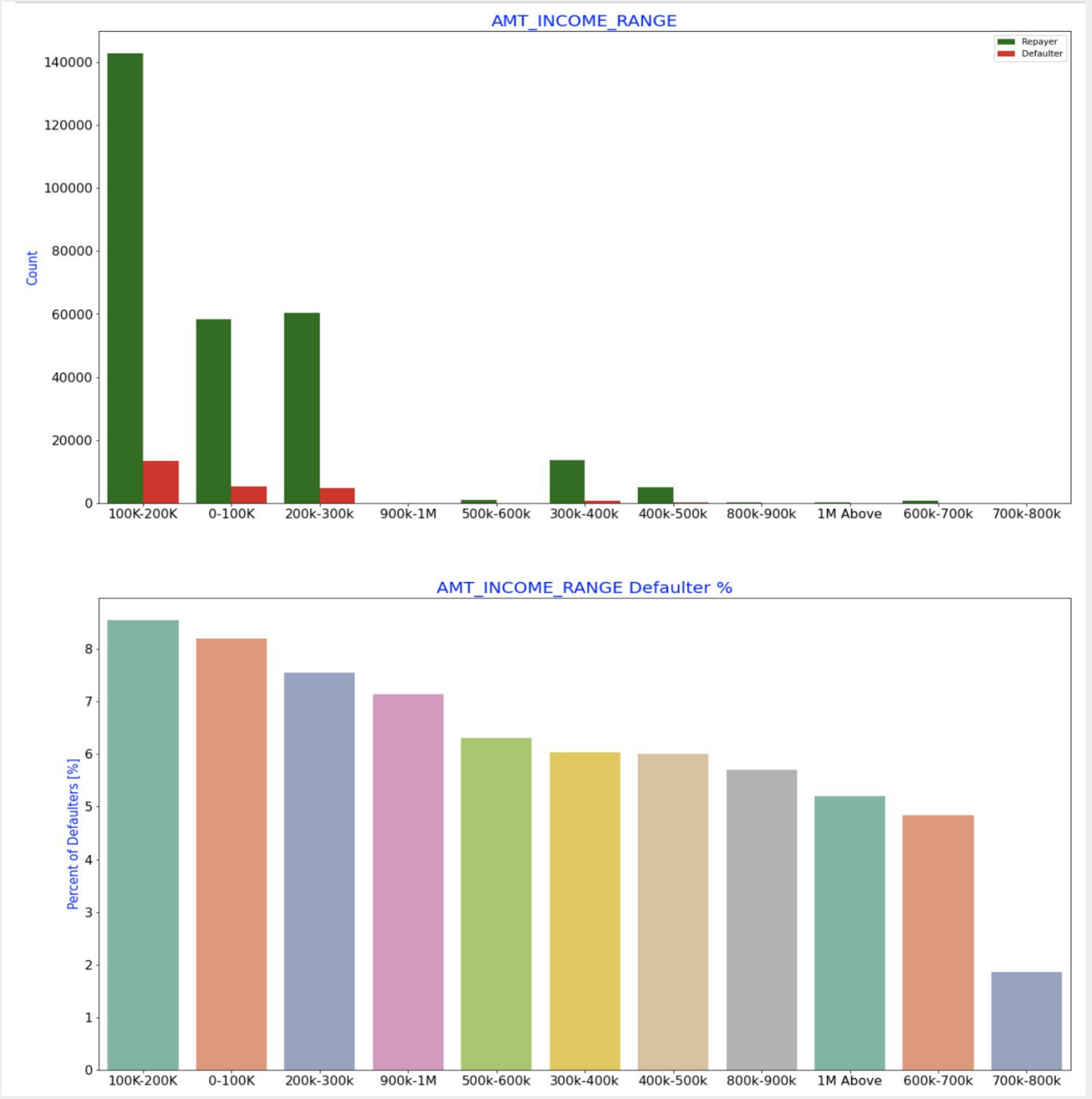
- Majority of the Loan amount is between 200-300K
- More than 80% of the loan provided are for amount less than 900,000
- People who get loan for 300-600k tend to default more than others.

AMT_Income_Range

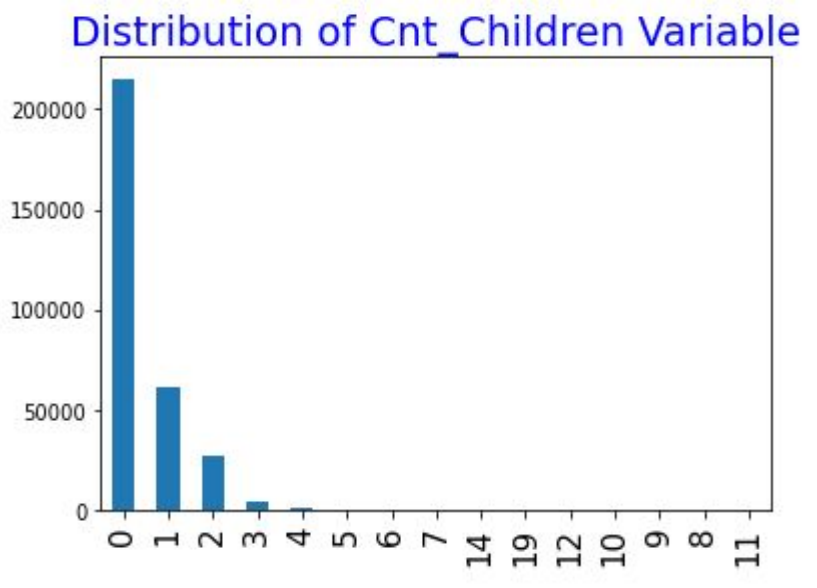


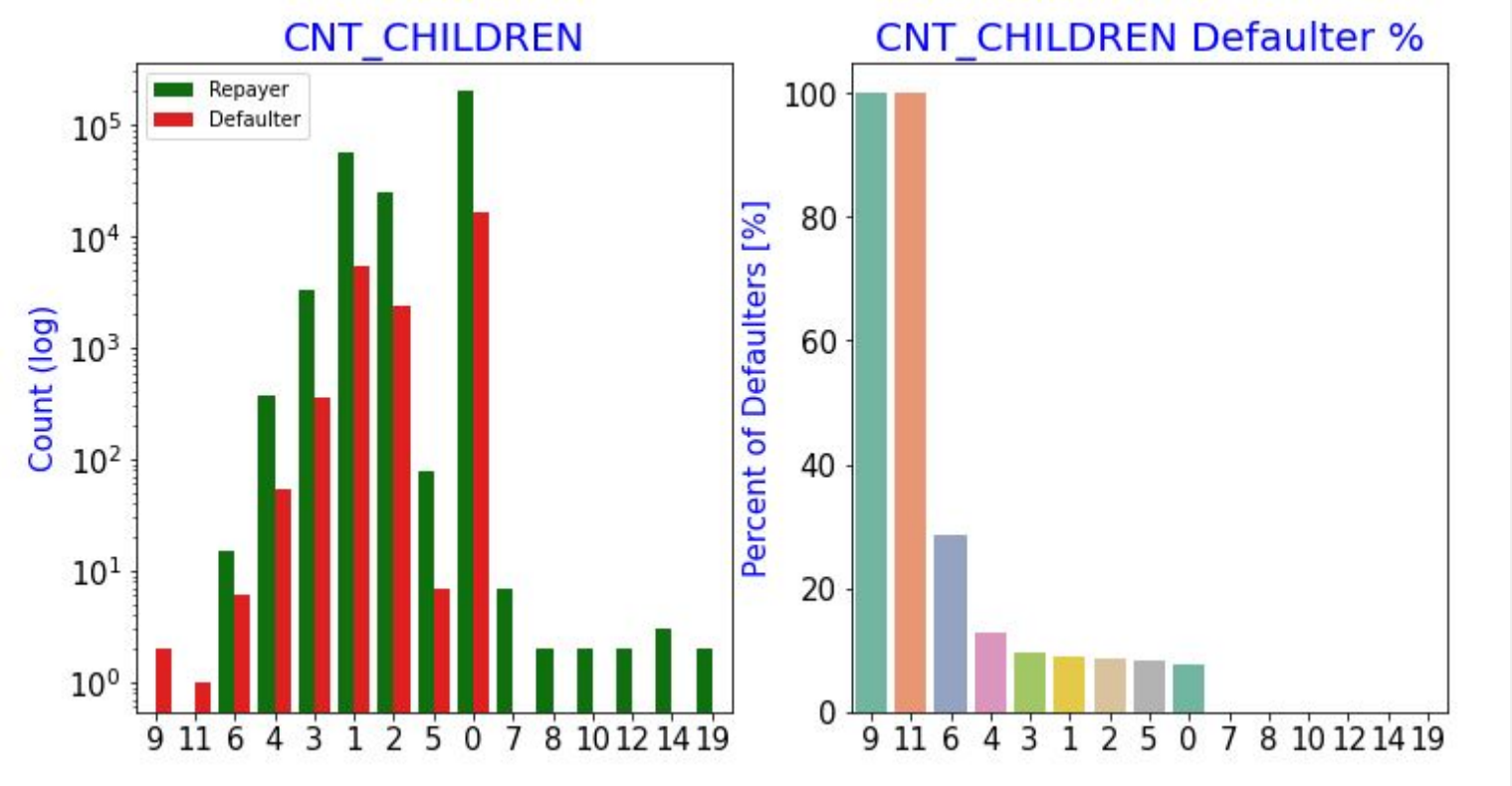
Inferences:

- Majority of the applicants have salary between 100-200K
- Application with Income less than 300,000 has high probability of defaulting
- Applicant with Income between 700-800k are less likely to default



CNT_Children

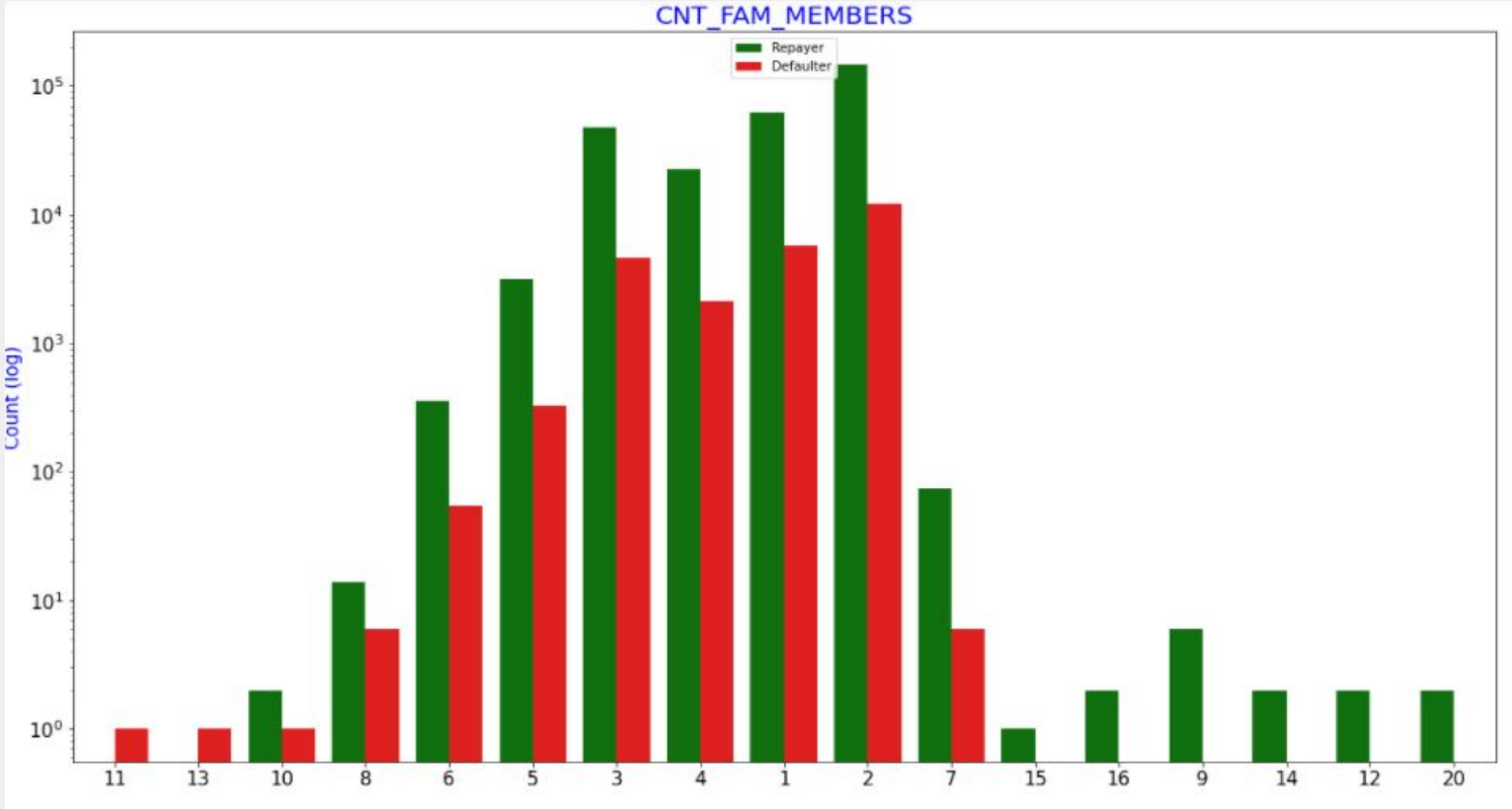


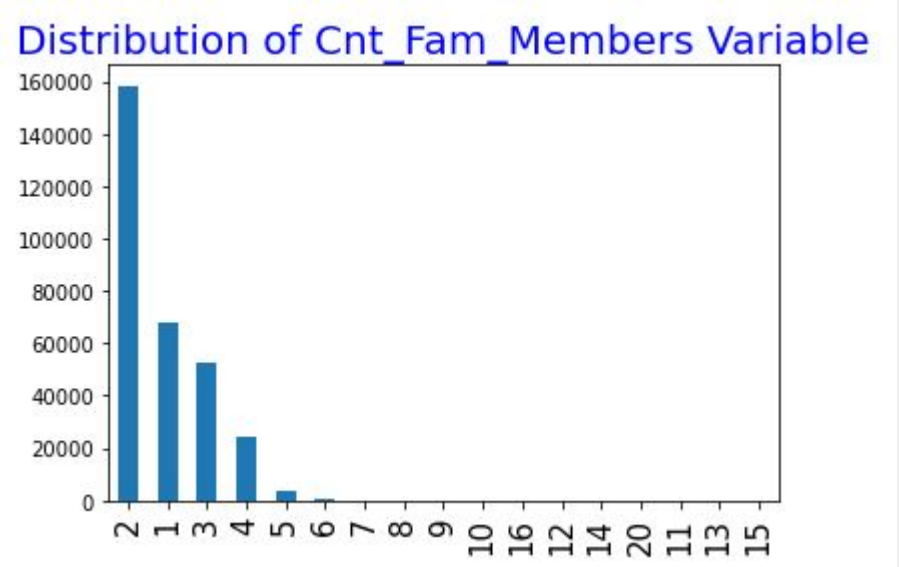


Inferences:

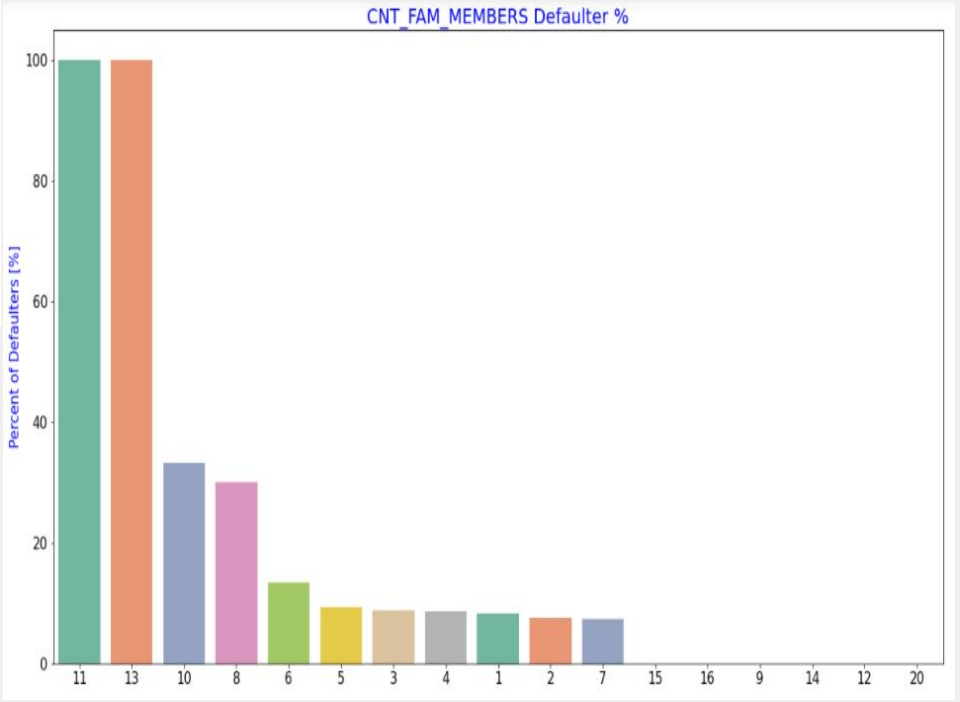
- Most of the applicants do not have children. As applicants in this group are more the no. of defaulters are also more in this group
- Very few clients have more than 3 children.
- Client who have more than 4 children have a very high default rate with child count 9 and 11 showing 100% default rate

CNT_FAM_Members





Majority of the clients have 2 family members

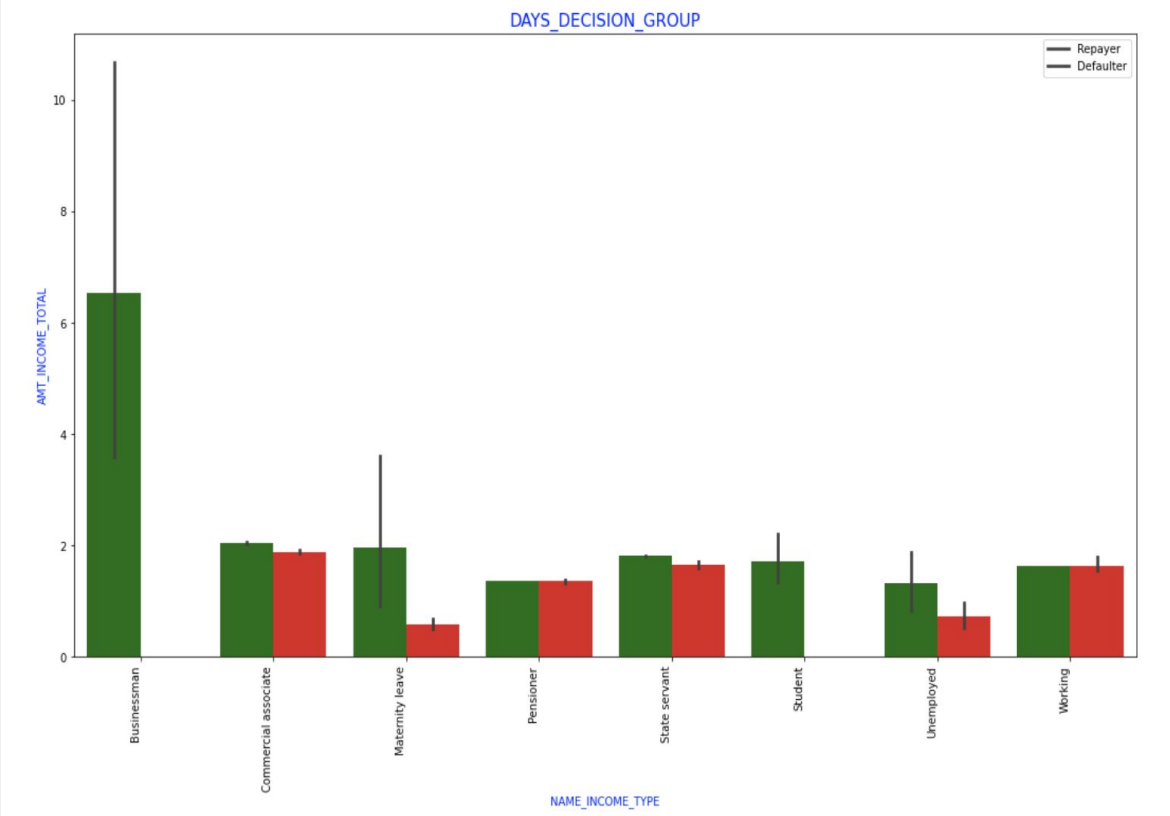


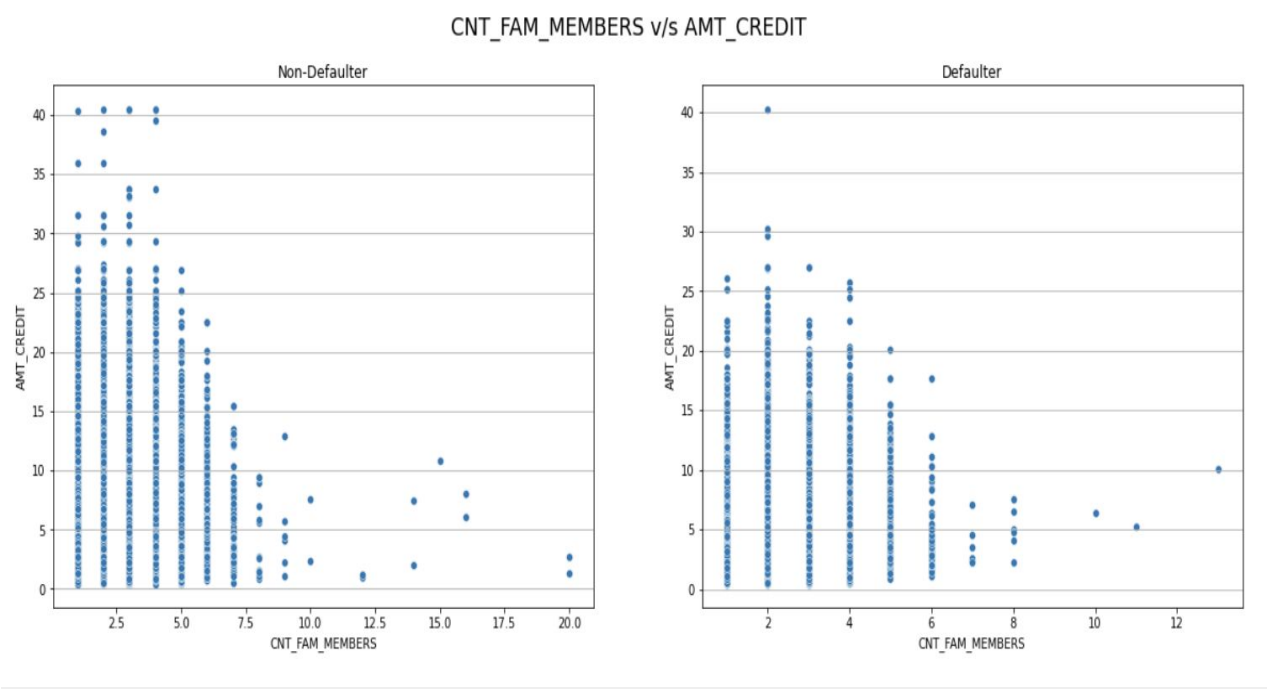
Inferences:

- Family member follows the same trend as children where having more family members increases the risk of defaulting

Bivariate & Multivariate

Categorical Bi/Multivariate Analysis





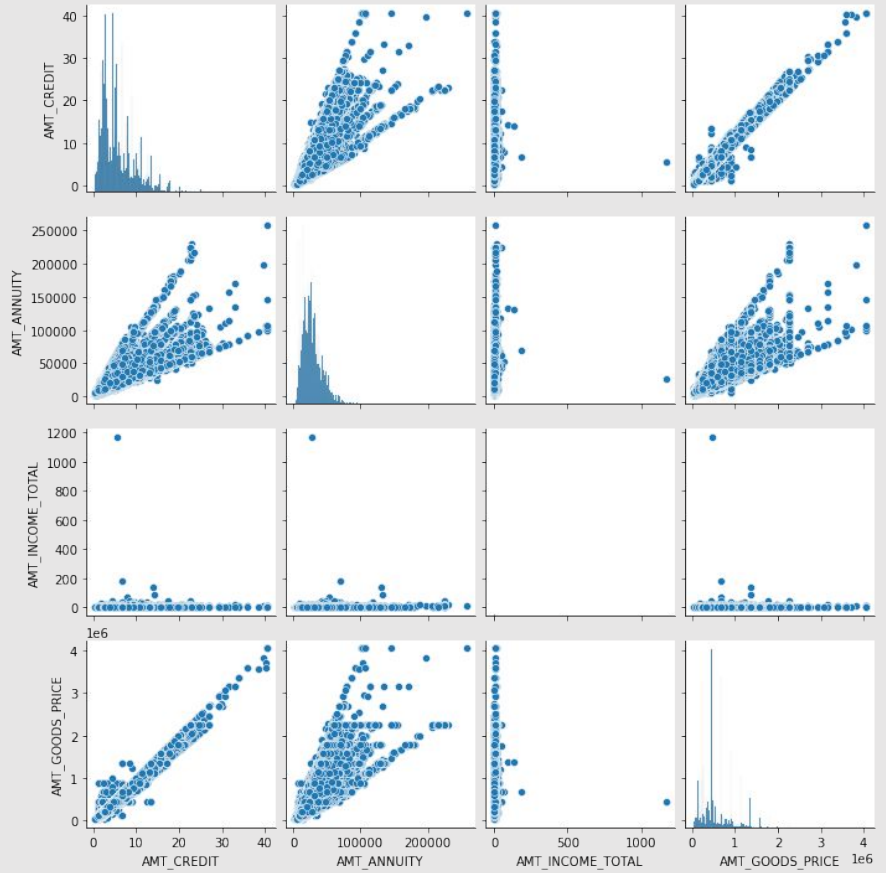
Inferences:

- It can be seen that business man's income is the highest and the estimated range with default 95% confidence level seem to indicate that the income of a business man could be in the range of slightly close to 4 lakhs and slightly above 10 lakhs

Insights:

1. Applicants having larger family size and less Amount credit are likely to default less.
2. Applicants having smaller family size and higher Amount credit are likely to default less.

AMT_CREDIT, AMT_ANNUITY_AMT_INCOME_TOTAL, AMT_GOODS_PRICE



Insights:

Very high correlation between AMT_CREDIT and AMT_GOODS_PRICE - Applicants owning goods of high value can take loans of higher amounts.

Multivariate(Numeric Columns)

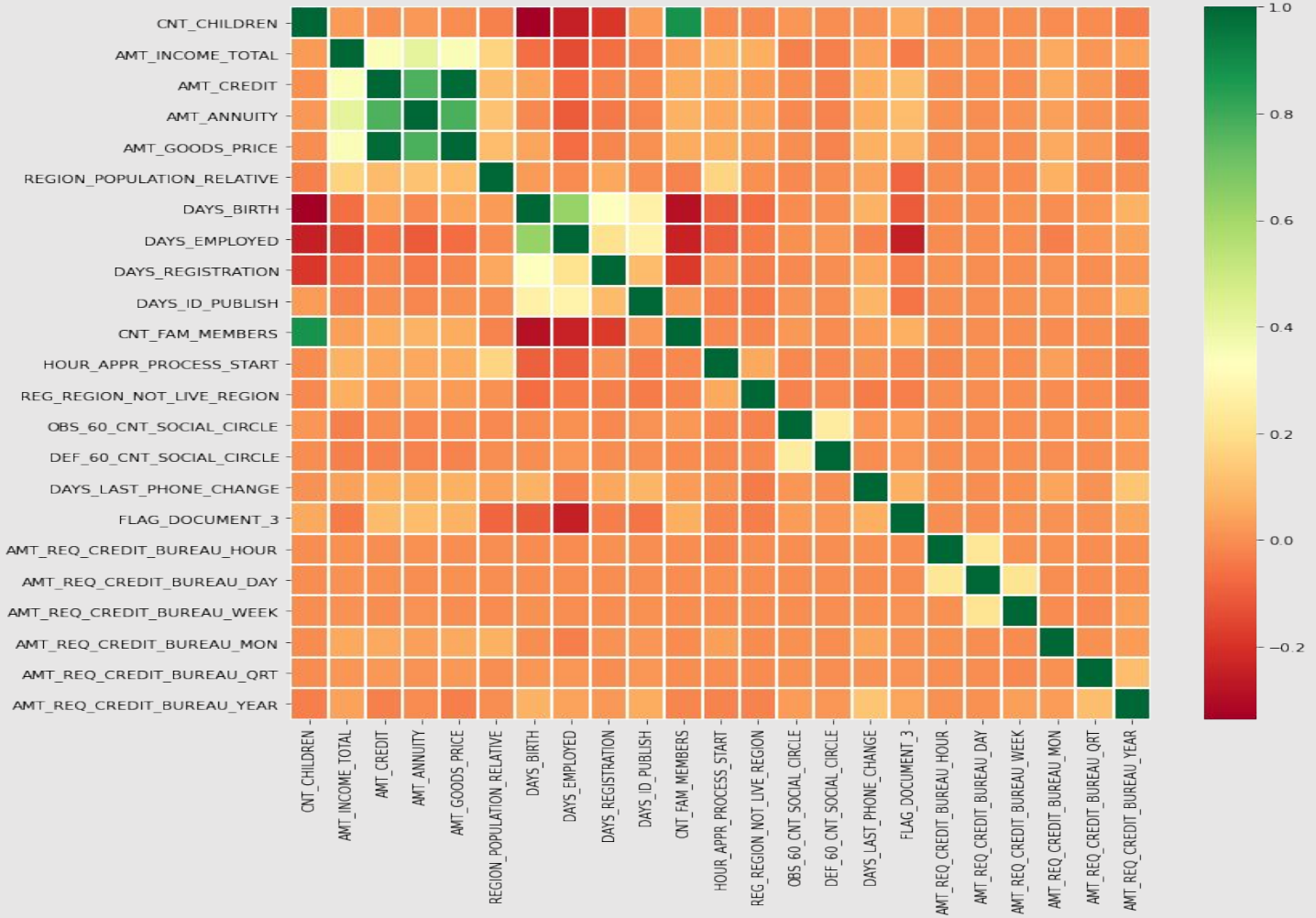
Repayer

Inferences:

Correlating factors amongst repayers:

Credit amount is highly correlated with

- amount of goods price
- loan annuity
- total income

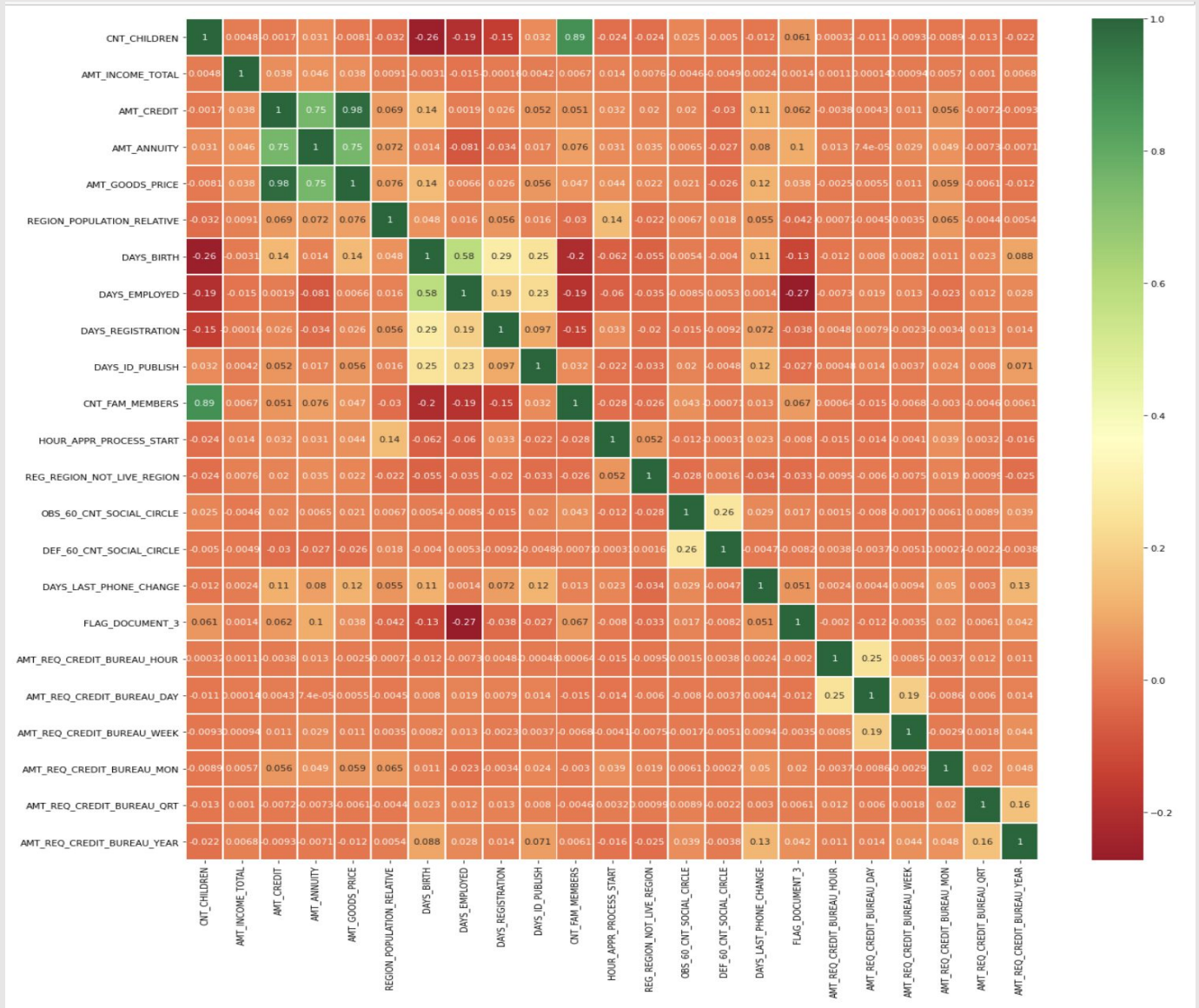


Multivariate(Numeric Columns)

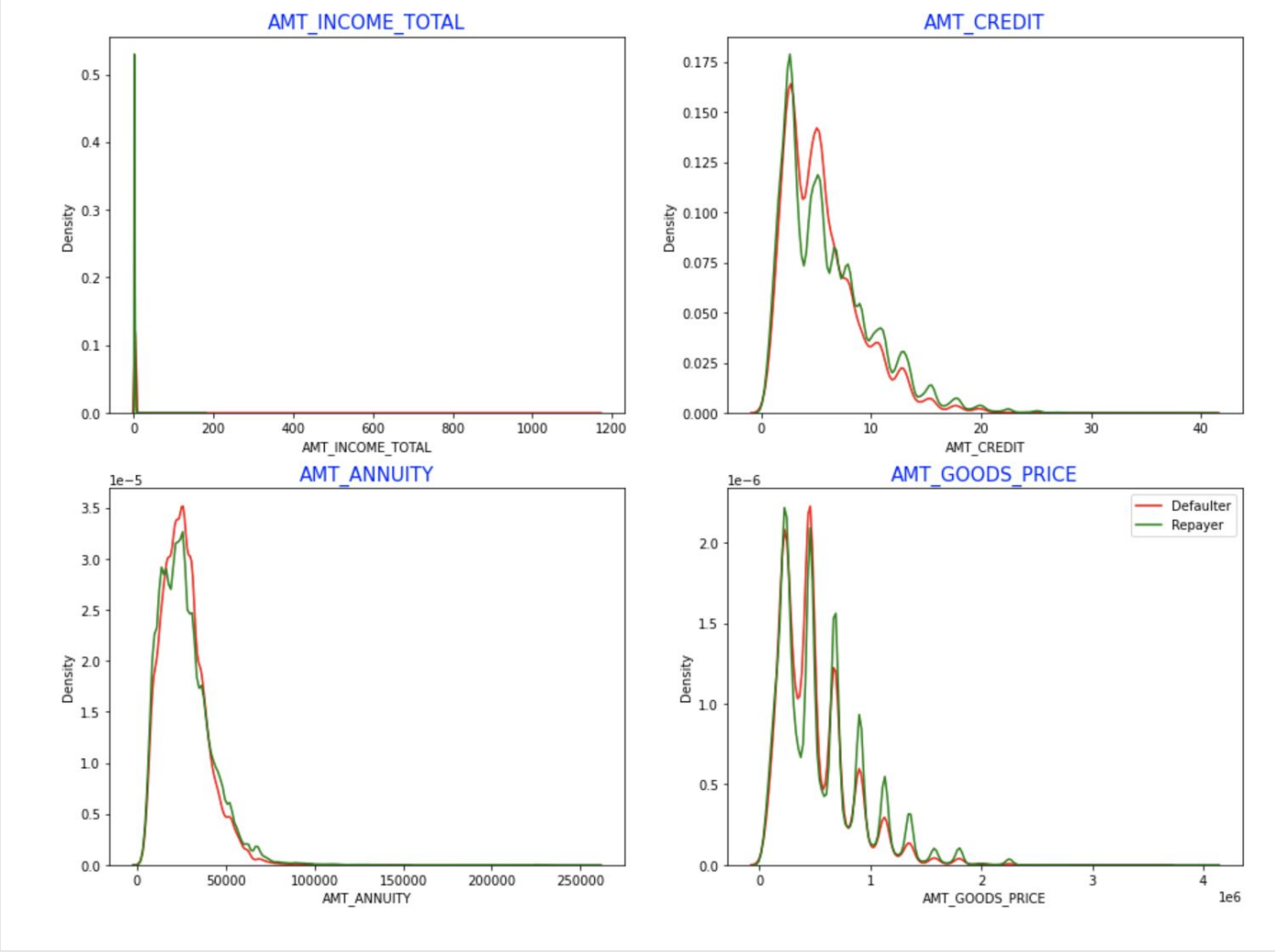
Defaulter

Inferences:

- Credit amount is highly correlated with amount of goods price which is same as repayers.
- But the loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to repayers(0.77)
- We can also see that repayers have high correlation in number of days employed(0.62) when compared to defaulters(0.58).
- There is a severe drop in the correlation between total income of the client and the credit amount(0.038) amongst defaulters whereas it is 0.342 among repayers.
- Days_birth and number of children correlation has reduced to 0.259 in defaulters when compared to 0.337 in repayers.
- There is a slight increase in defaulted to observed count in social circle among defaulters(0.264) when compared to repayers(0.254)



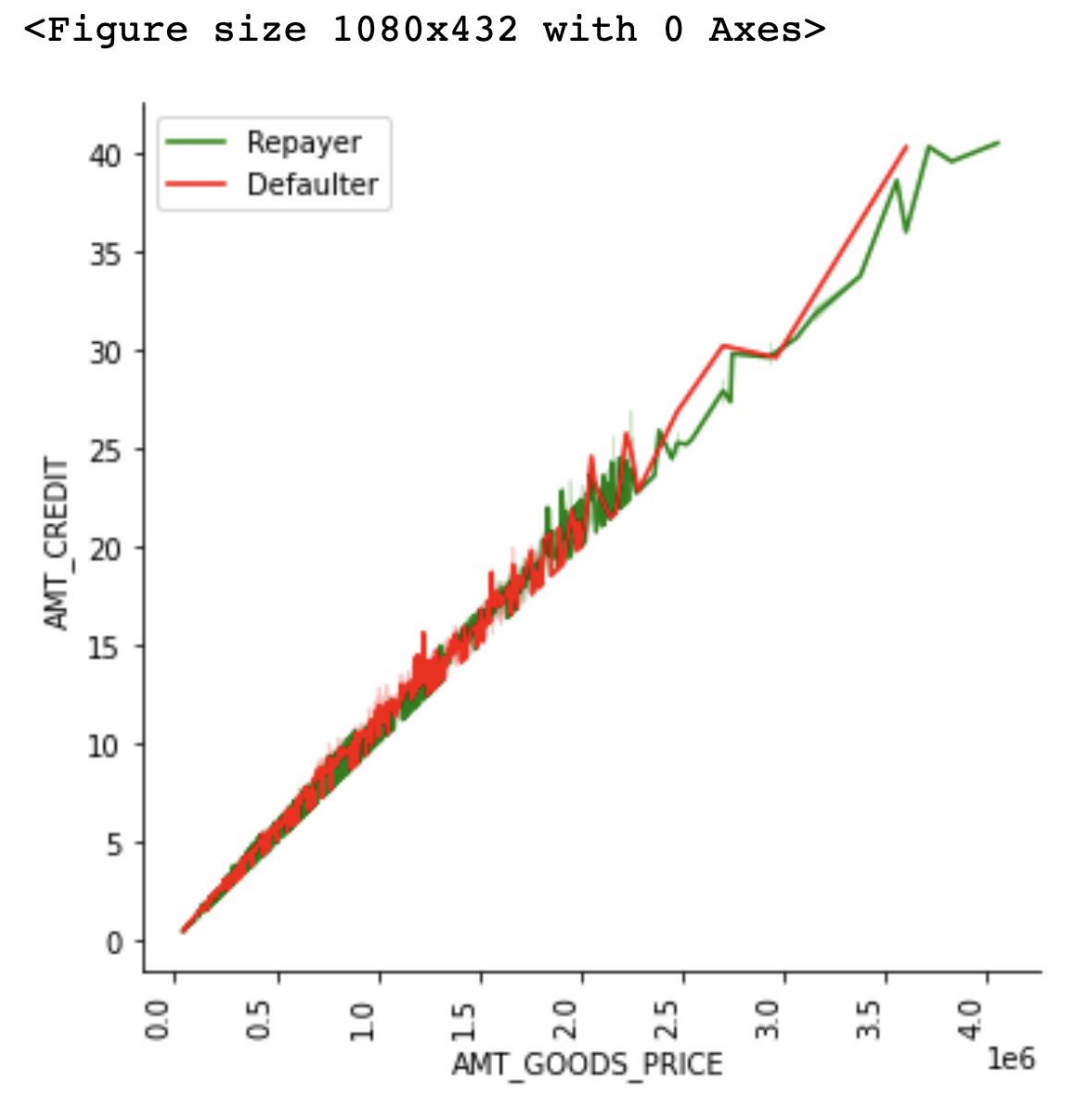
Numerical Univariate Analysis



Inferences:

- Most no of loans are given for goods price below 10 lacs
- Most people pay annuity below 50000 for the credit loan
- Credit amount of the loan is mostly less than 10 lacs
- The re-payers and defaulters distribution overlap in all the plots and hence we cannot use any of these variables in isolation to make a decision

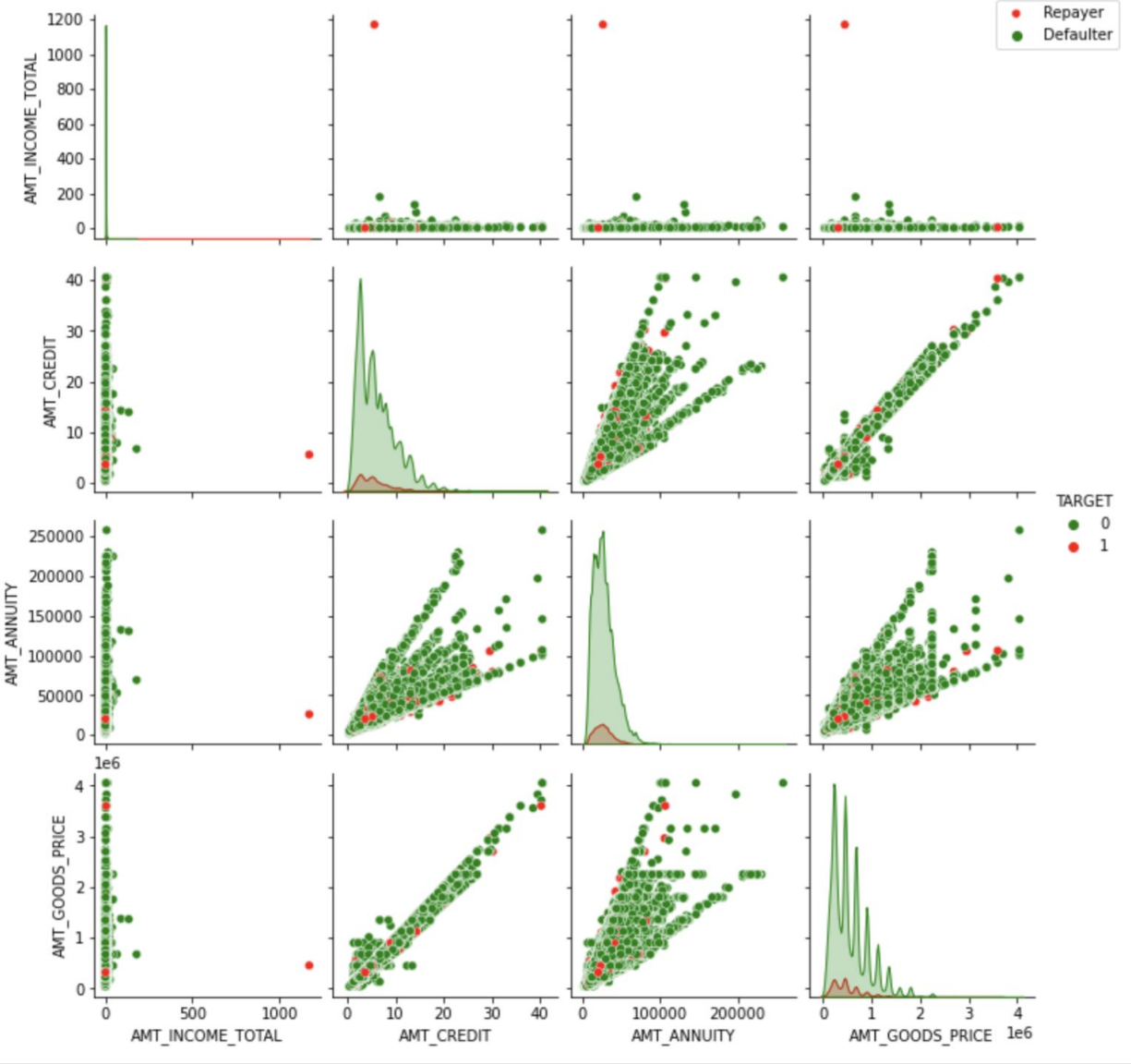
Numerical Bivariate Analysis



Inferences

- **When the credit amount goes beyond 3M, there is an increase in defaulters**

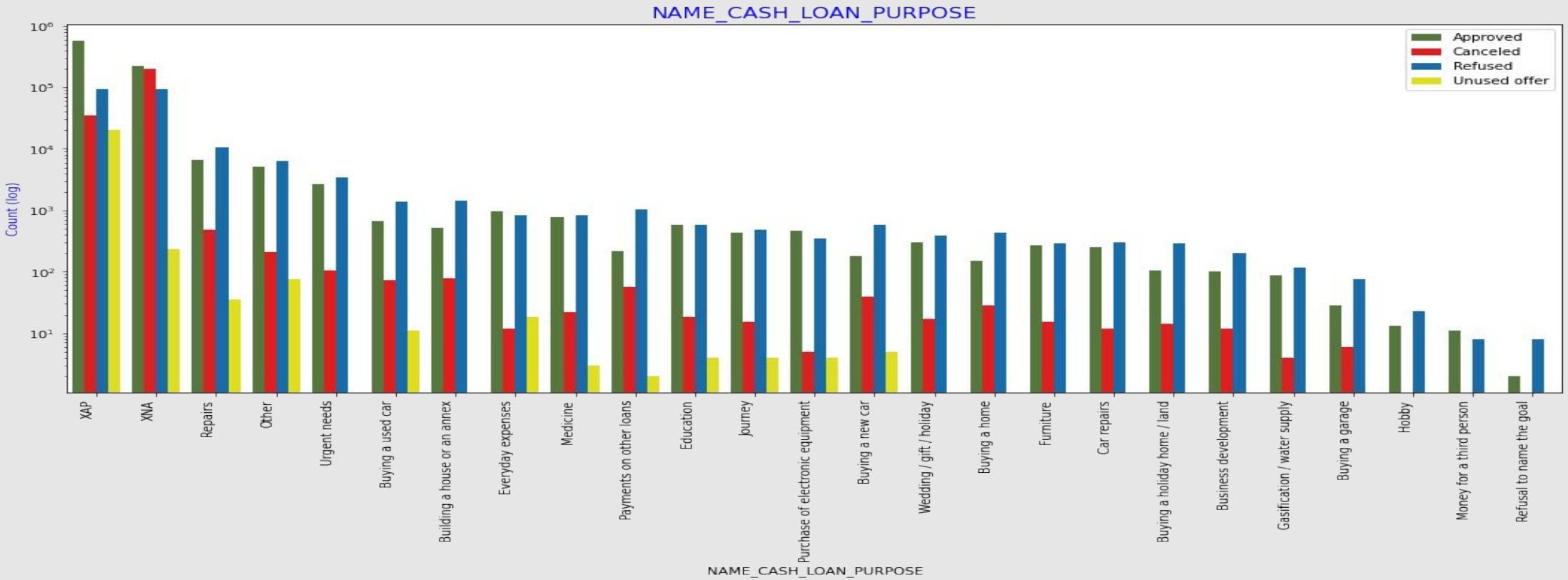
Plotting pairplot between amount variable



Inferences:

- When amt_annuity >15000 amt_goods_price> 3M, there is a lesser chance of defaulters
- AMT_CREDIT and AMT_GOODS_PRICE are highly correlated as based on the scatterplot where most of the data are consolidated in form of a line
- There are very less defaulters for AMT_CREDIT >3M
- Inferences related to distribution plot has been already mentioned in previous distplot graphs inferences section

Name_Cash_Loan_Purpose



Inferences:

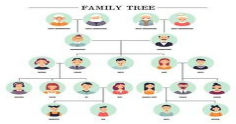
- Loan purpose has high number of unknown values (XAP, XNA)
- Loan taken for the purpose of Repairs seems to have highest default rate
- A very high number application have been rejected by bank or refused by client which has purpose as "repair or other". This shows that purpose "repair" is considered as high risk by bank and either they are rejected or bank offered very high loan interest rate which is not feasible by the clients, thus they refuse the loan.


LOAN
Case Study
Summary

Definitive Factors for an applicant to be safe borrowers

	Applicants with Income more than 700,000 are less likely to default
	Applicants with 40+ year of experience having less than a 1% default rate.
	Applicants with zero to two children tend to repay the loans.
	Applicants above age of 50 have low probability of defaulting.
	Academic degree has less defaults.
	Applicants with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%.
 	Loans bought to pursue Hobby or Buying garage are being repaid mostly.
	Student have no defaults.
	Applicants who live in areas with Region Rating 1 are safe borrowers.

Decisive Factors for an applicant to be a potential Defaulter

	When the loan amount goes beyond 3M, there is an increase in defaulters.
	When we check data gender-wise, there is a high chance of a male applicant being a defaulter.
	Applicants who have children equal to or more than nine default 100%, and hence their applications need to review appropriately before approval.
	Applicants with higher family members (>=11) have higher default rates, and their applications may not repay the amount.
	Avoid young applicants who are in the age group of 20-40 as they have a higher probability of defaulting
	Applicants who have less than five years of employment have a high default rate.
	Applicants with Lower Secondary education, incomplete education have a higher default rate.
	Applicants who are either on Maternity leave or Unemployed have a higher default rate.
	Applicants who live in areas with a Region Rating of 3 has the highest defaults.

Decisive Factors for an applicant to be a potential Defaulter

	Applicants in civil marriage or who are single have higher default rate
	Low-skill Laborers, drivers and Waiters/barmen staff, Security staff have a tremendous default rate.
	Industry type 3, type 13 and type 8 have a high defaulting rate
	Applicants who get a loan for 300-600k tend to default more than others, and hence having higher interest specifically for this credit range would be ideal.
	Since 90% of the applications have an Income total of less than 300,000 and have a high probability of default, they could be offered a loan with higher interest than other income categories.
	Applicants who have 4 to 8 children have a very high default rate, and hence higher interest can be imposed on their loans.
	Loan for house Repairs seems to have the highest default rate. As a result, the bank has rejected a very high number of applications.
	People living in rented apartments & living with parents and hence offering the loan would mitigate the loss if any of those default