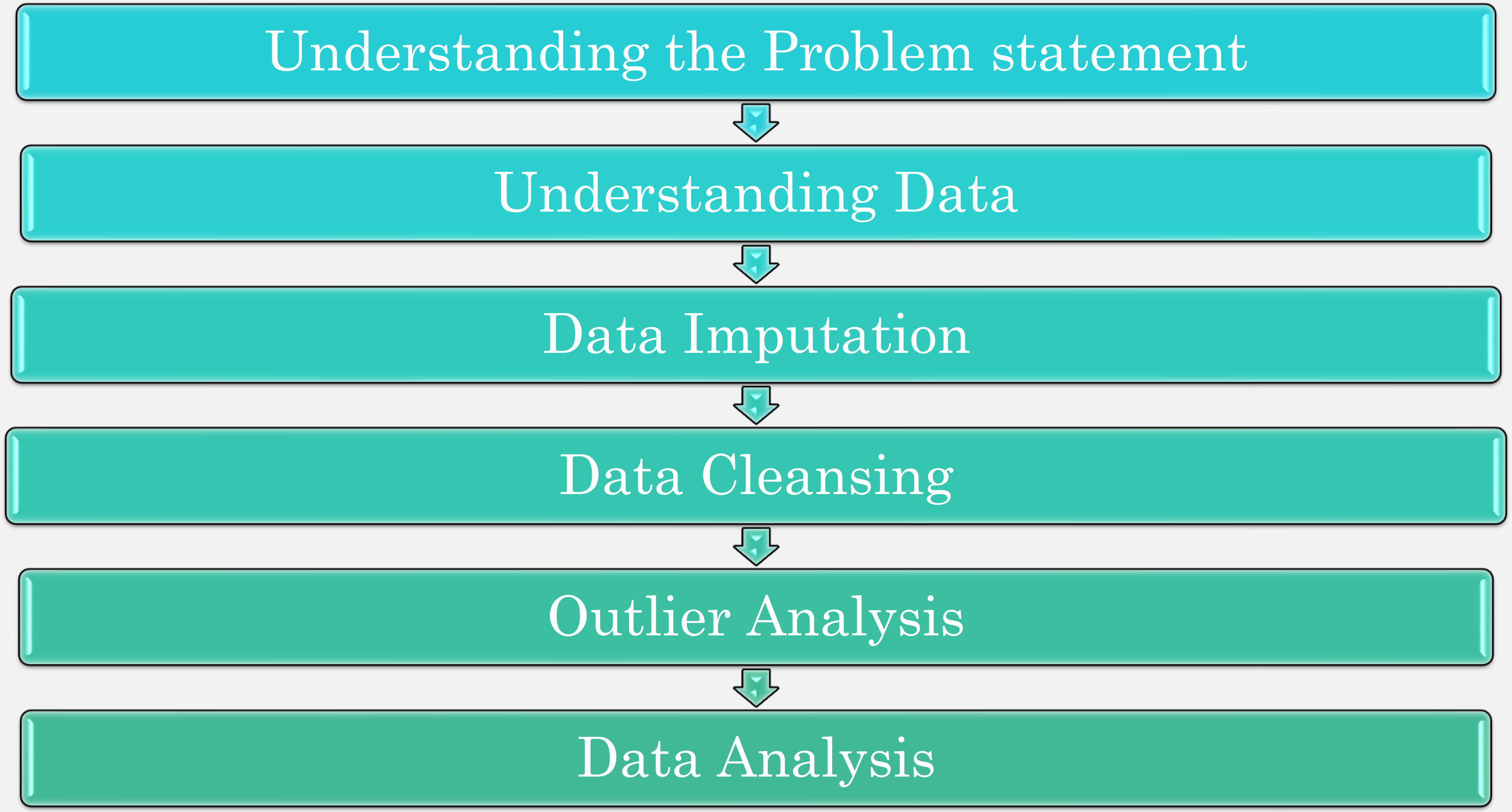
CREDIT EDA ASSIGNMENT

By SUMAN KUMAR

Problem Statement

- A consumer finance company specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile.
- Two types of risks are associated with the bank's decision:
    1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
    2. 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
- The company wants to understand the driving factors 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.

Analysis Approach



Understanding the Problem statement

Understanding Data

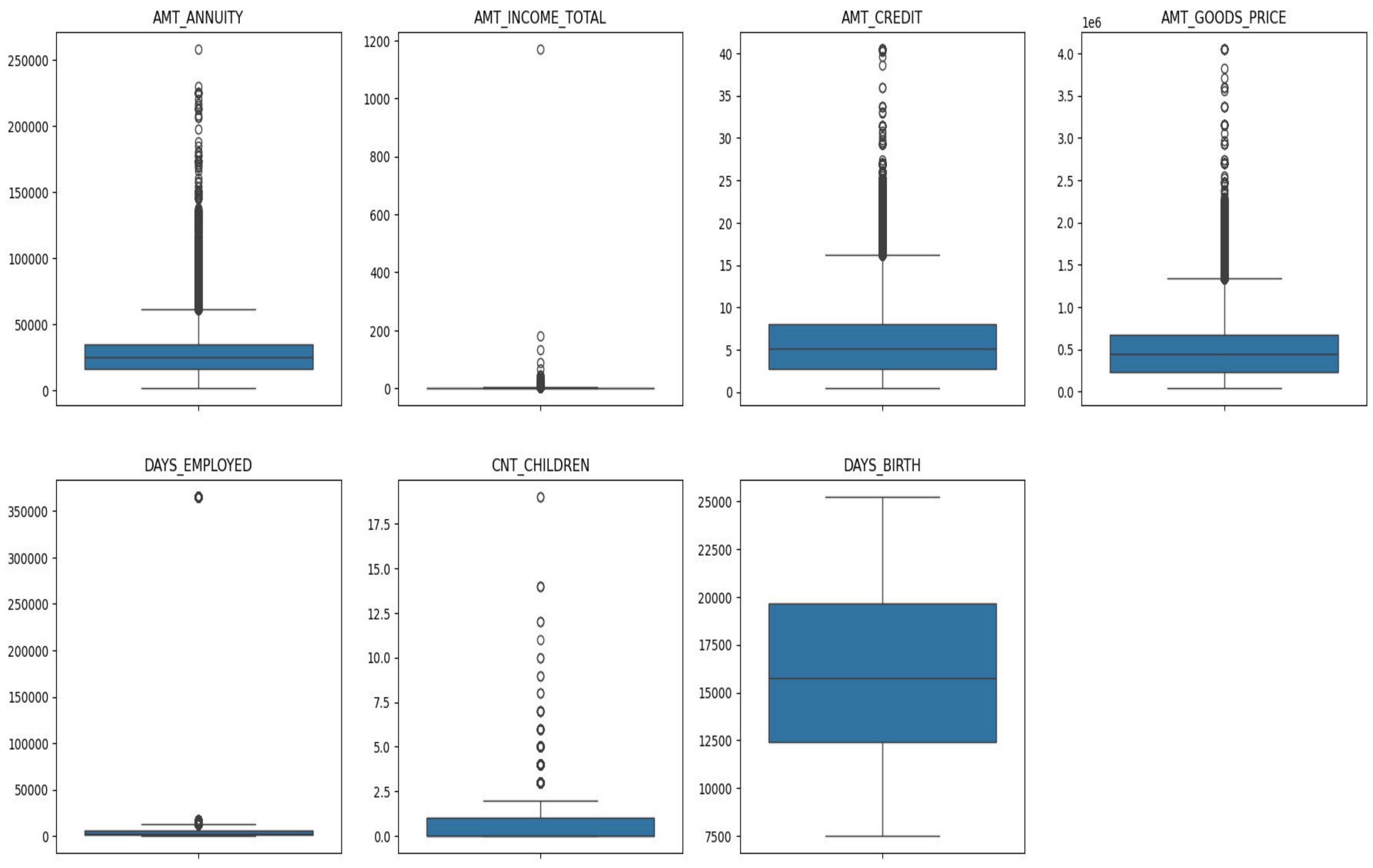
Data Imputation

Data Cleansing

Outlier Analysis

Data Analysis

Outlier Analysis

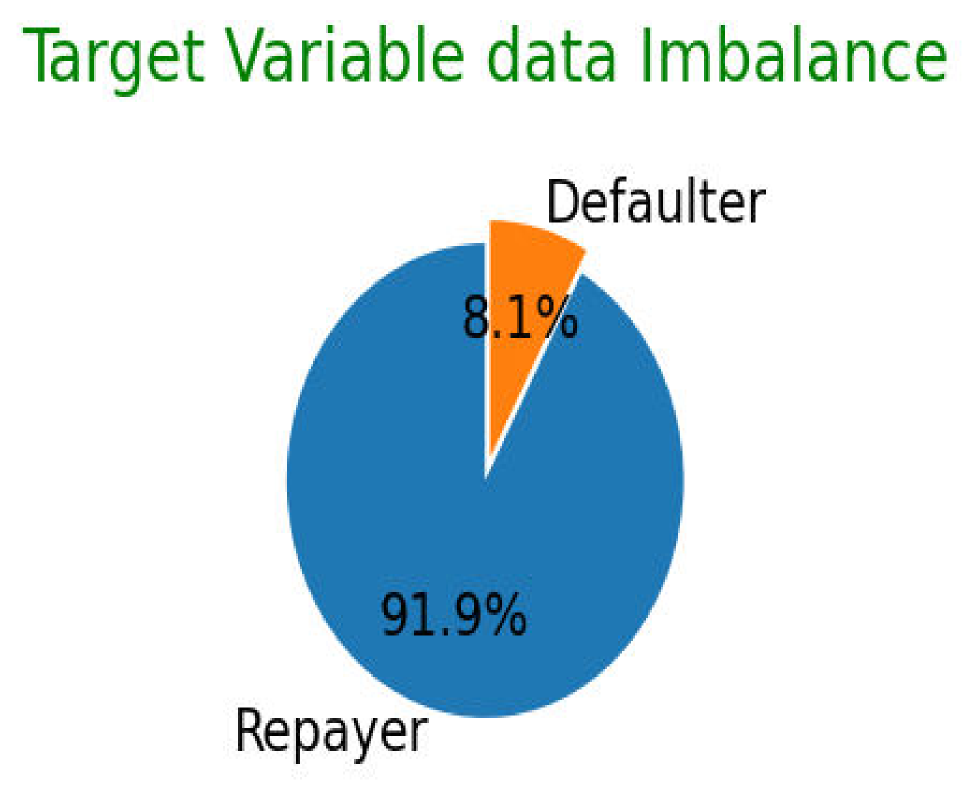


In APP Data new dataset

Inference :-

- _AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.
- CNT_PAYMENT has few outlier values
- SK_ID_CURR is an ID column and hence no outliers.
- DAYS_DECISION has little number of outliers indicating that these previous applications decisions were taken long back.
- We can see the stats for these columns below as well

Data Imbalance

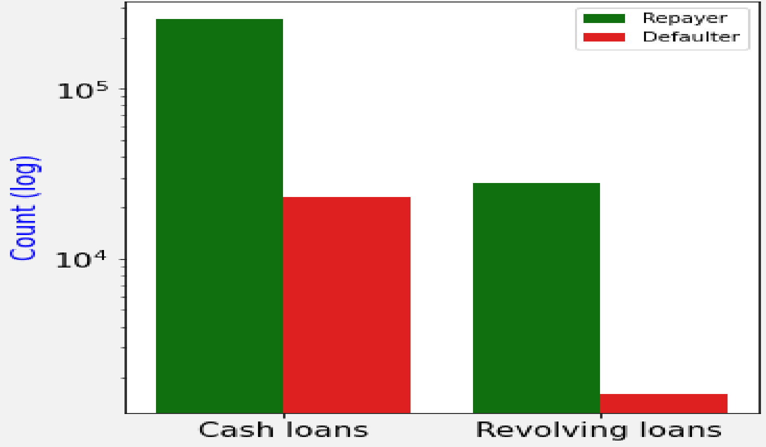


- The ratios of data imbalance for Repayer and Defaulter in percentage is :- 91.9% and 8.1%.
- The ratios of data imbalances for Repayer : Defaulter is 11.34 : 1 (approx.)

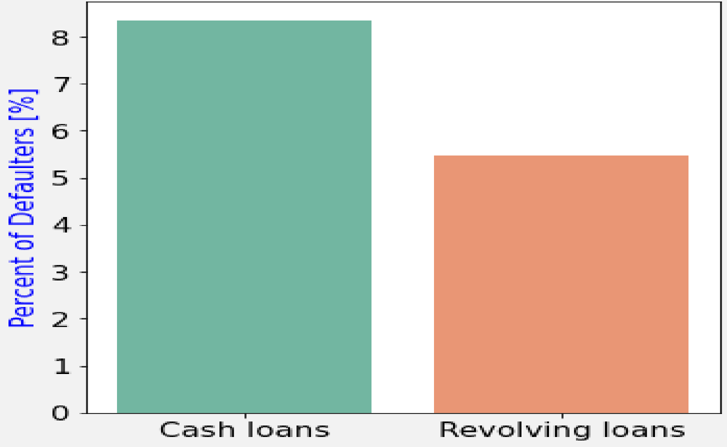
UNIVARIATE ANALYSIS

Contract Type

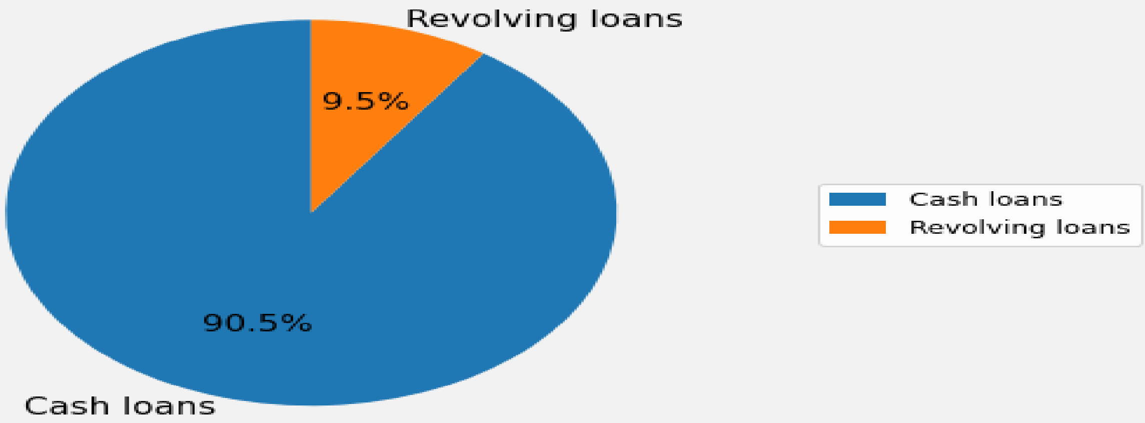
NAME_CONTRACT_TYPE



NAME_CONTRACT_TYPE Defaulter %



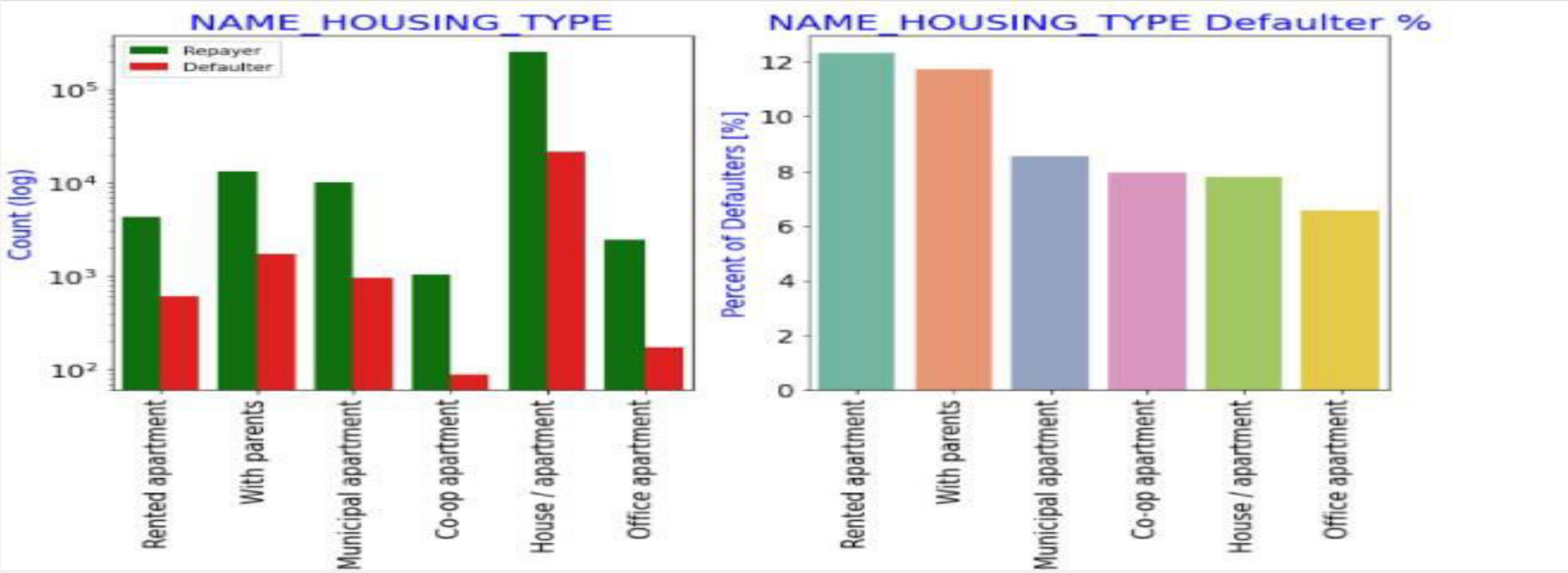
Distribution of Name_Contract_Type Variable

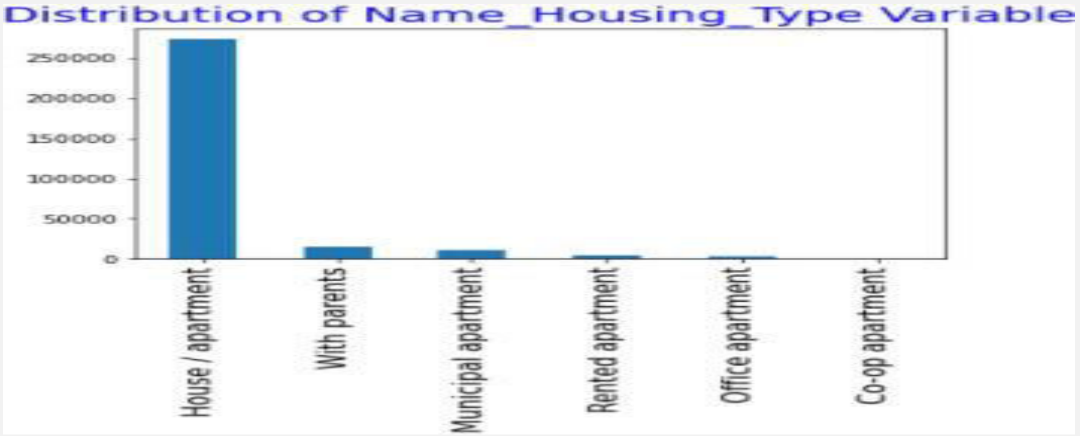


Inferences:

- □ We can see that revolving loans are only 10% that of total number of loans.
- □ The defaulted rate of revolving loan is quite high.

Housing Type

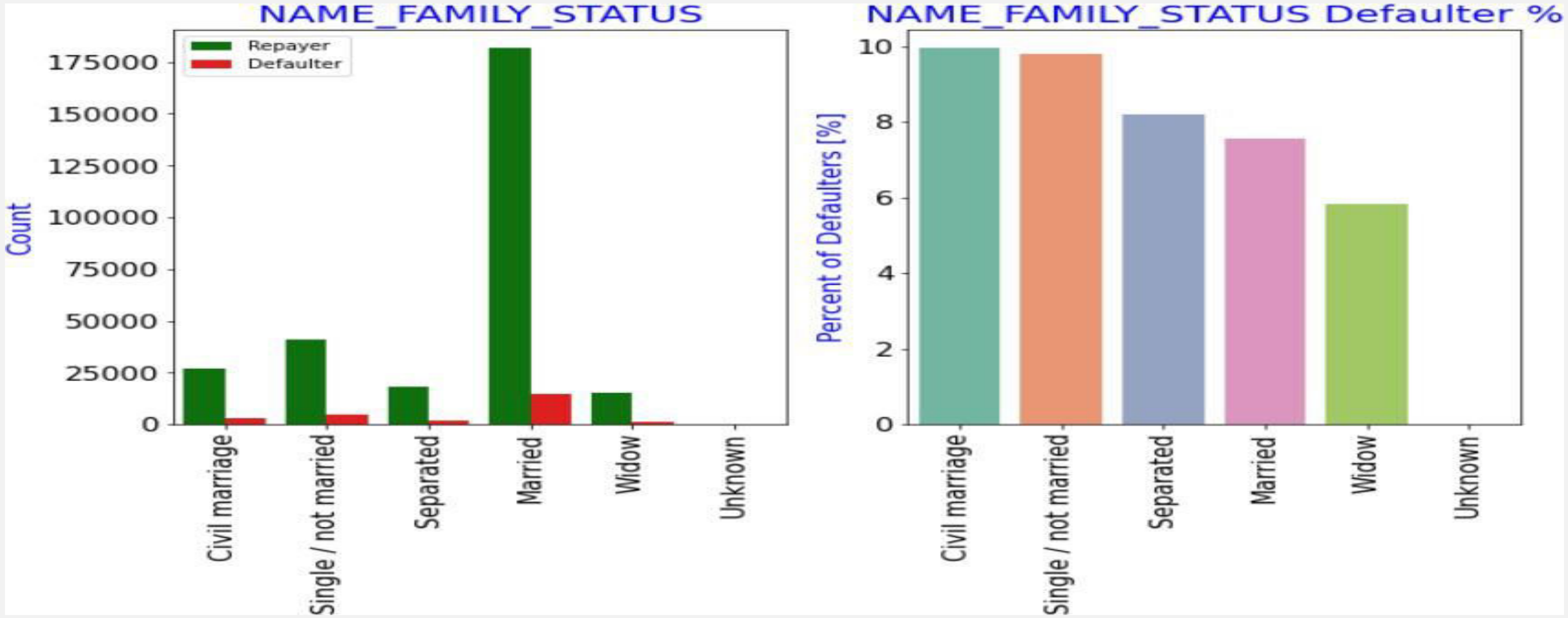


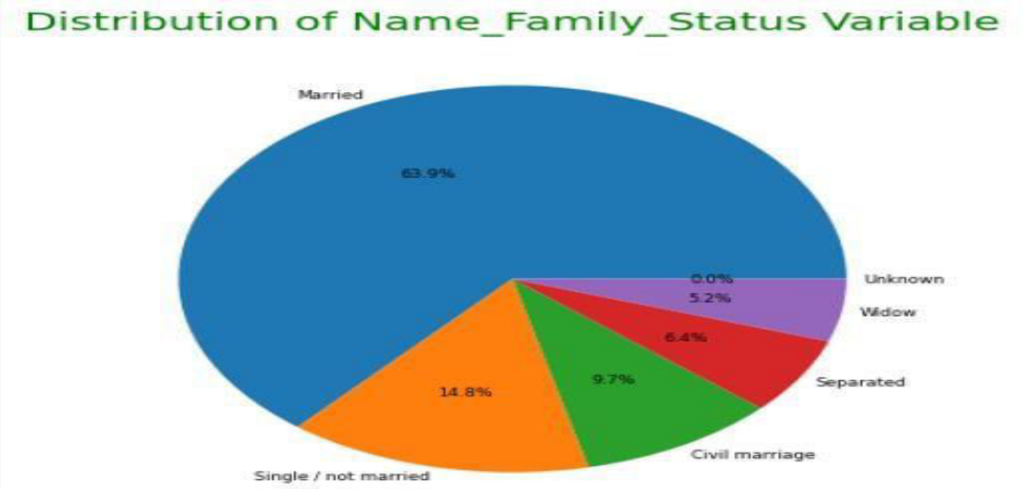


Inferences:

❖ The number of loan applicants are high from the category of people who either live in the rented apartments or with parents, so offering them the loan would result in loss if any of those default.

Family/ Marital Status

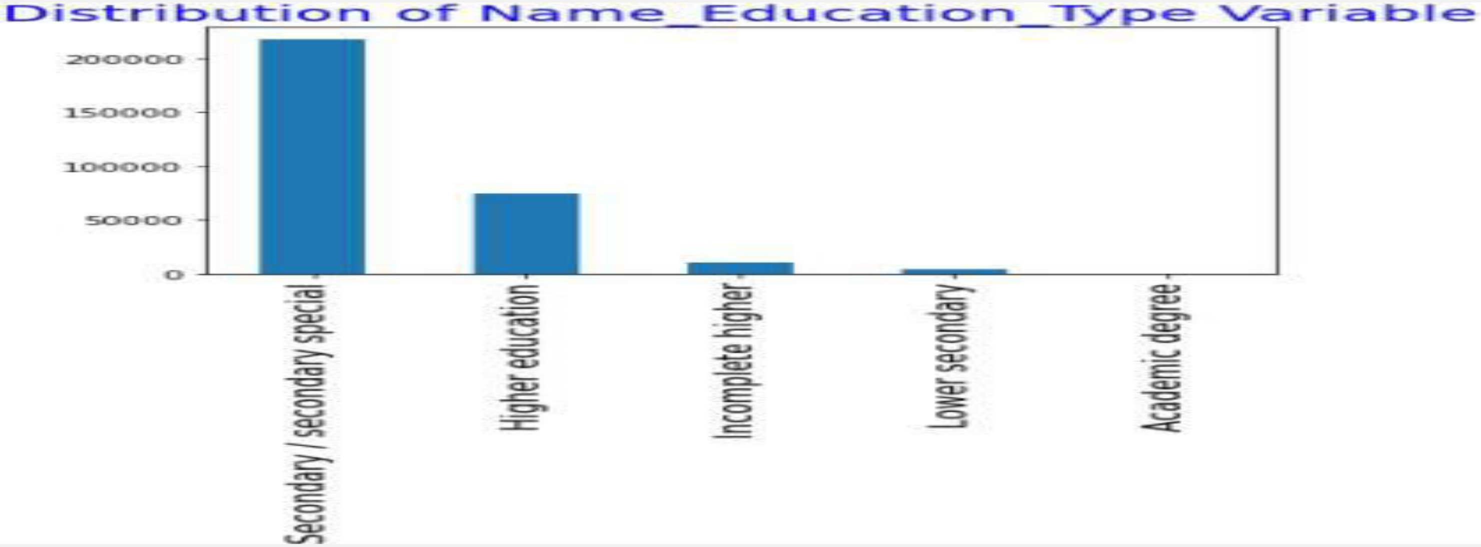


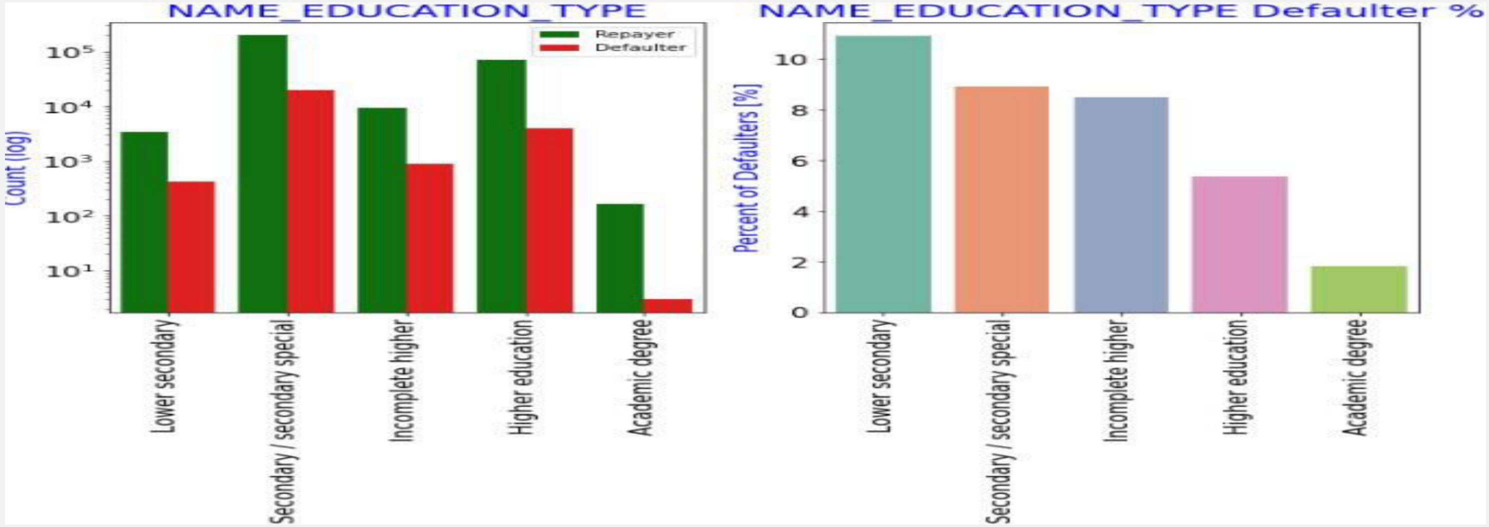


Inferences:

- 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, with Widow the lowest.

Education Type

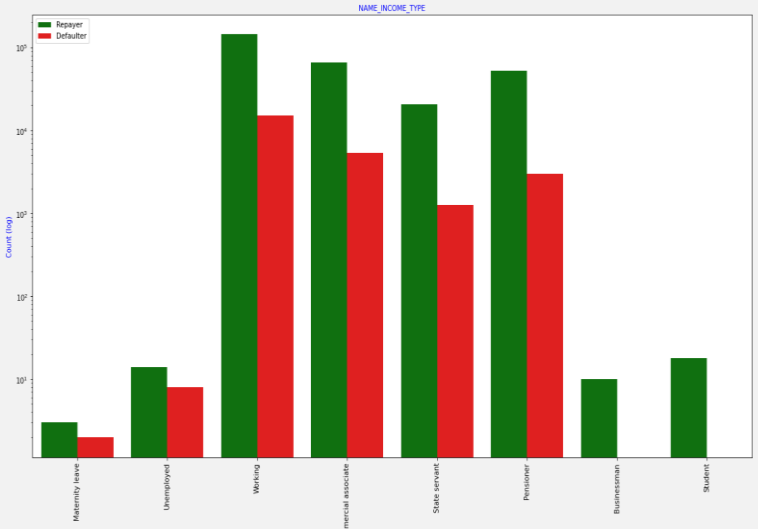


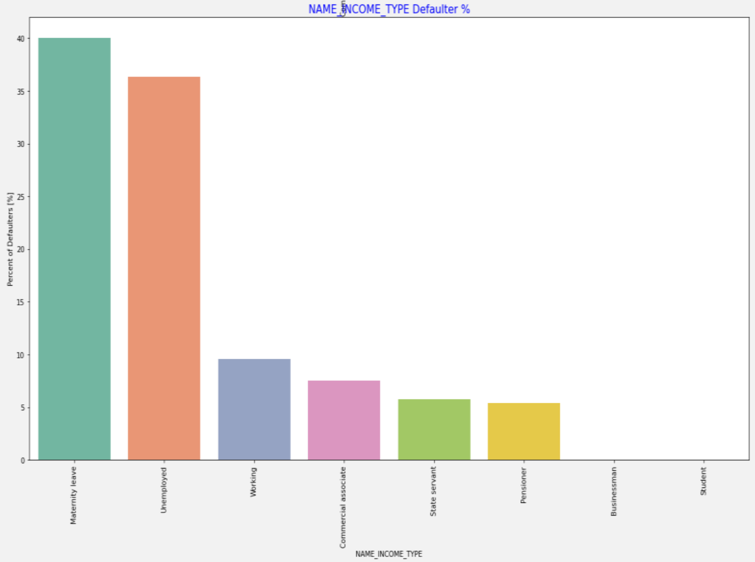


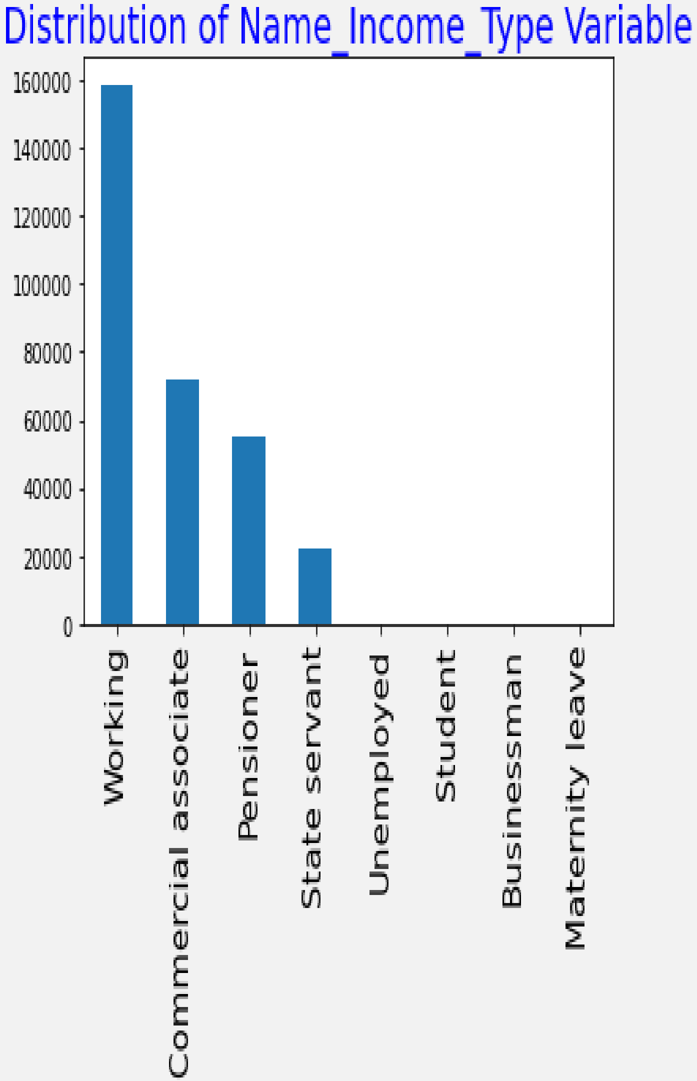
Inference:

- 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. The people with Academic degree have the lowest defaulting rate.

Income Type


NAME_INCOME_TYPE

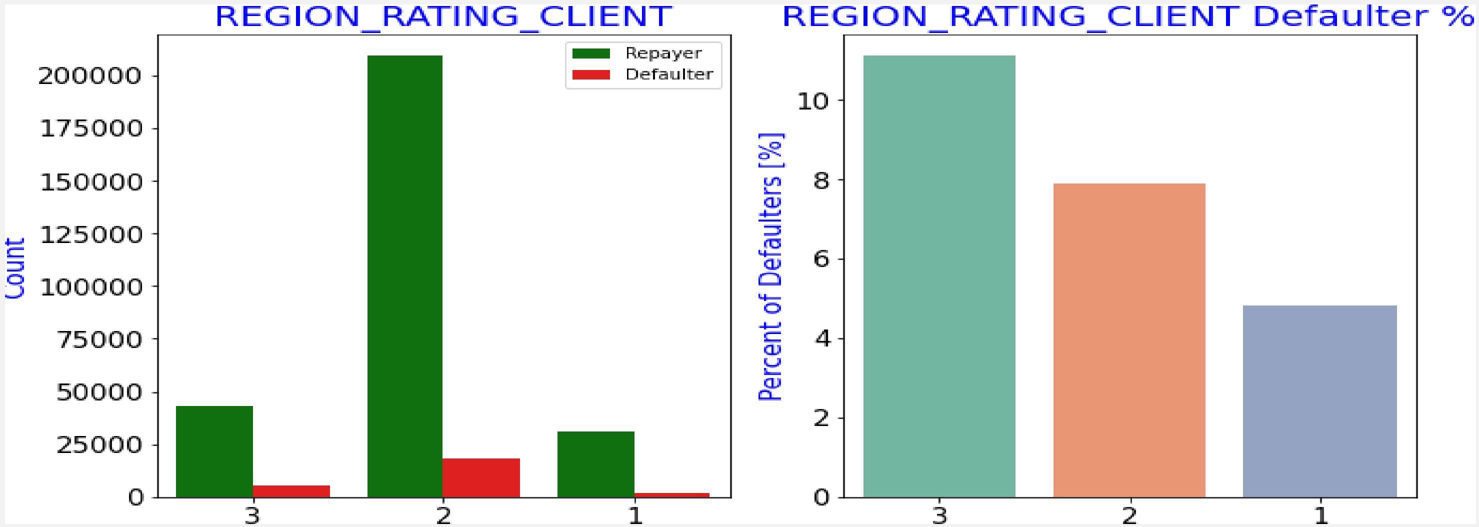

NAME_INCOME_TYPE Defaulter %

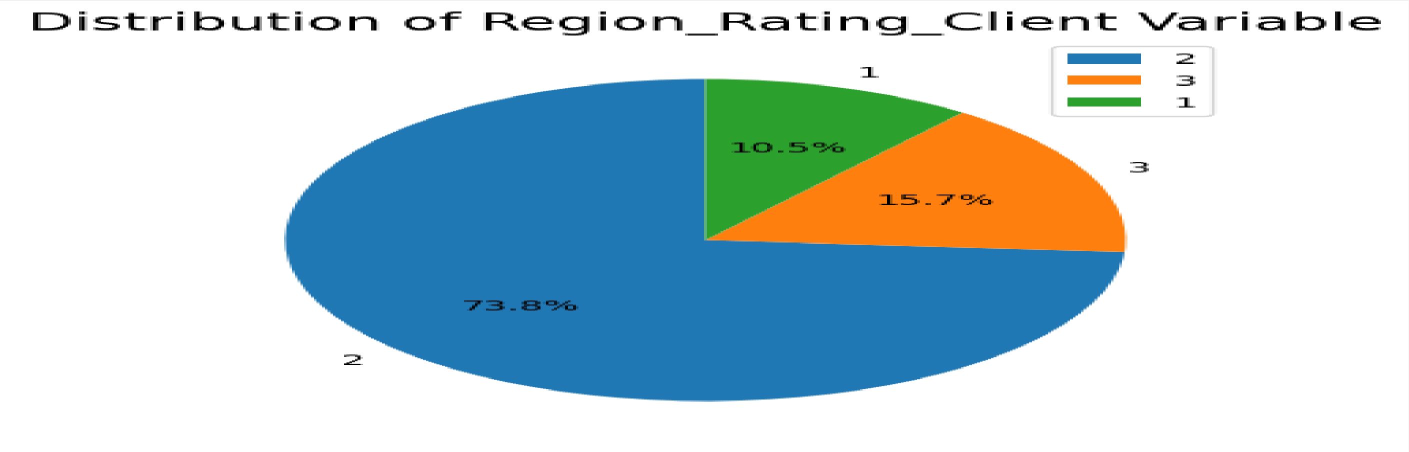

Distribution of Name_Income_Type Variable

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 Rating





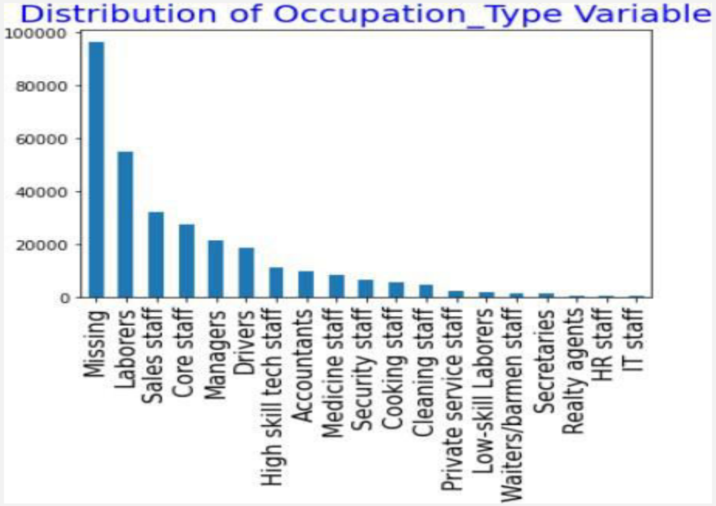
Inference:

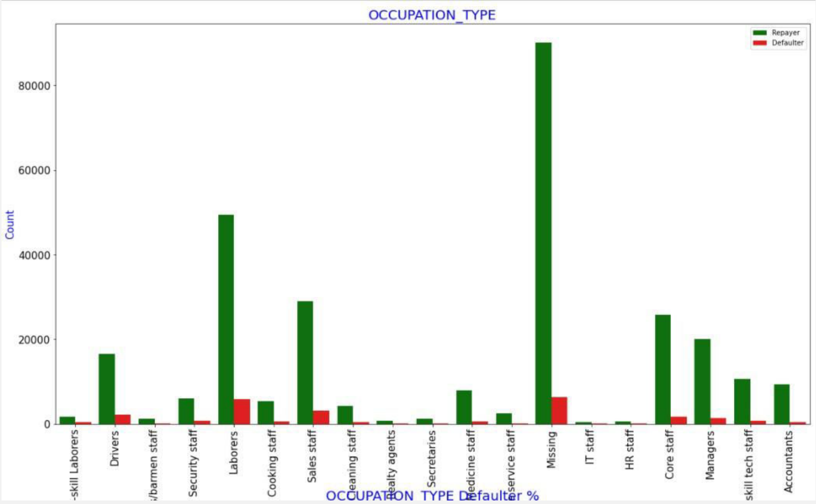
❑Most of the applicants are living in Region Rating 2 place.

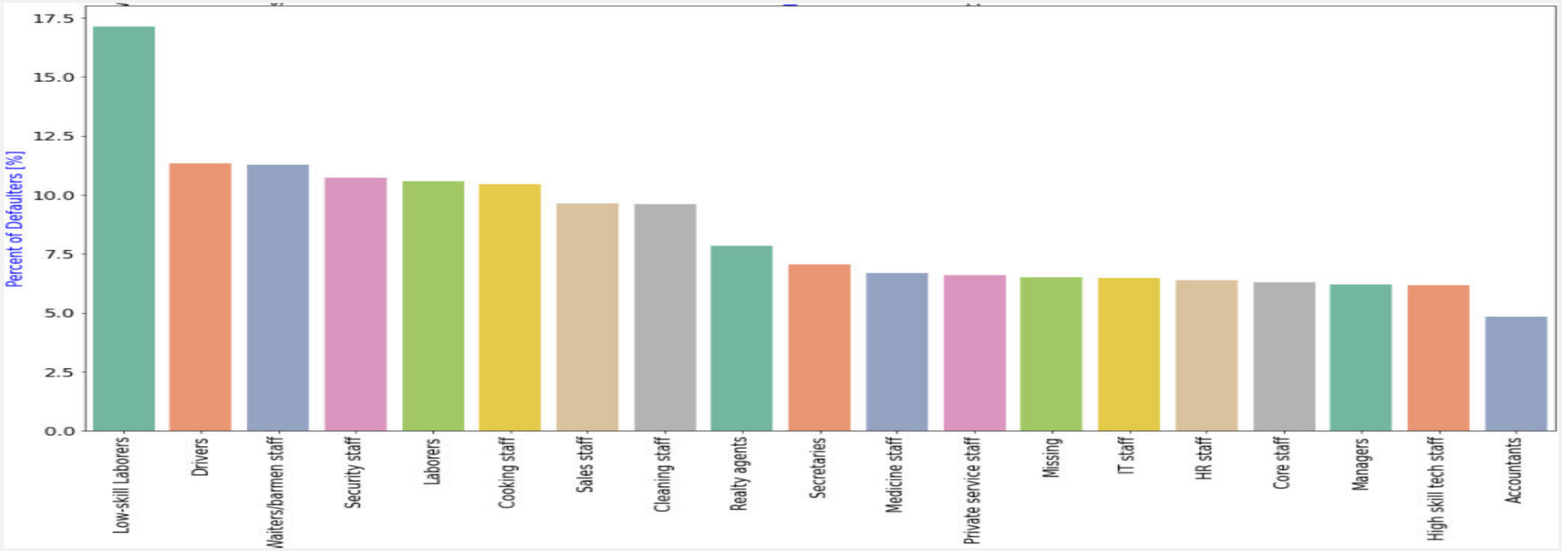
❑Region Rating 3 has the highest default rate (11%) , followed by 2(around 8%) and 1(around 5%)

❑Applicant living in Region Rating 1 has the lowest probability of defaulting, thus safer for approving loans

Occupation Type



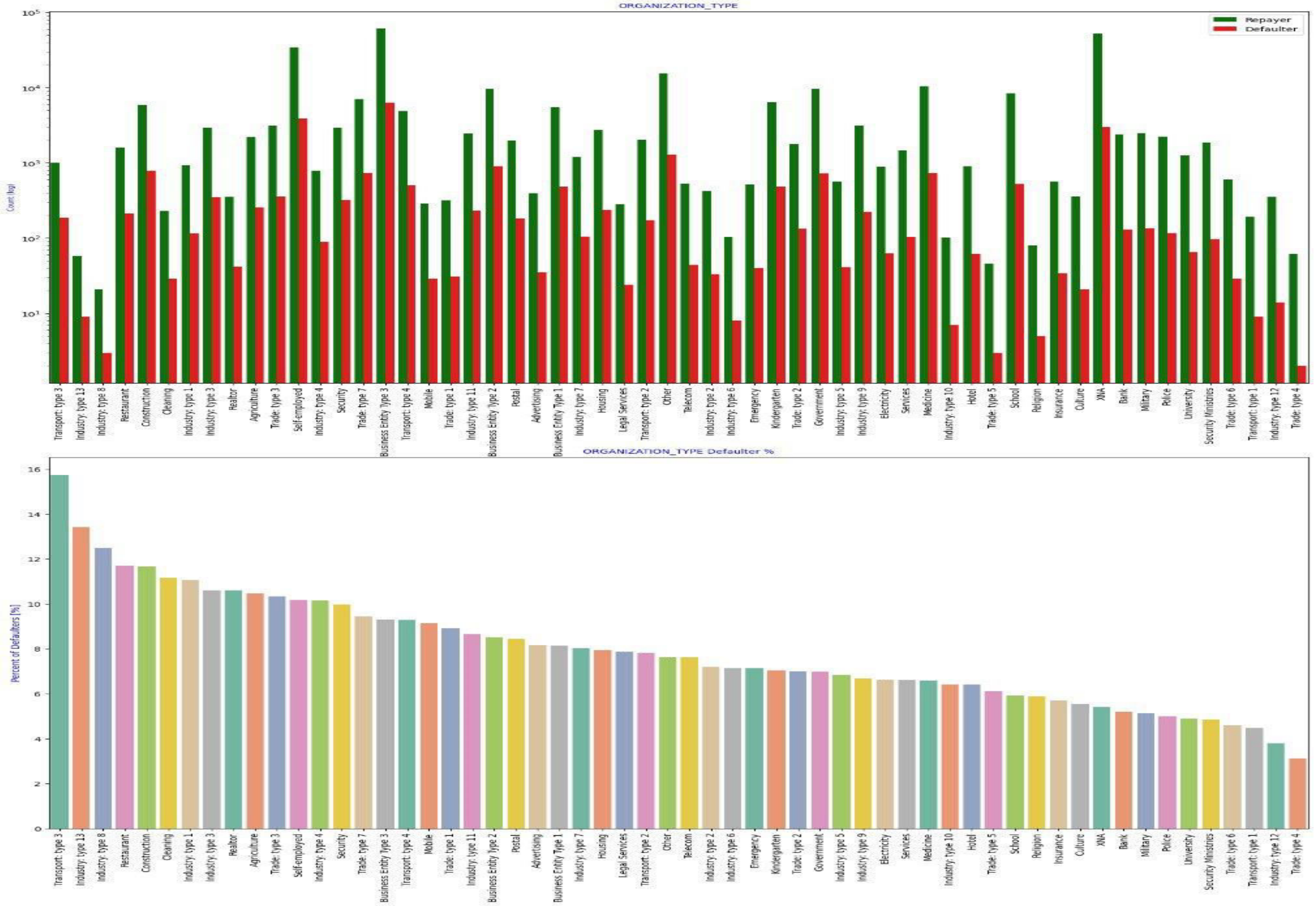




Inference:

- ➢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.

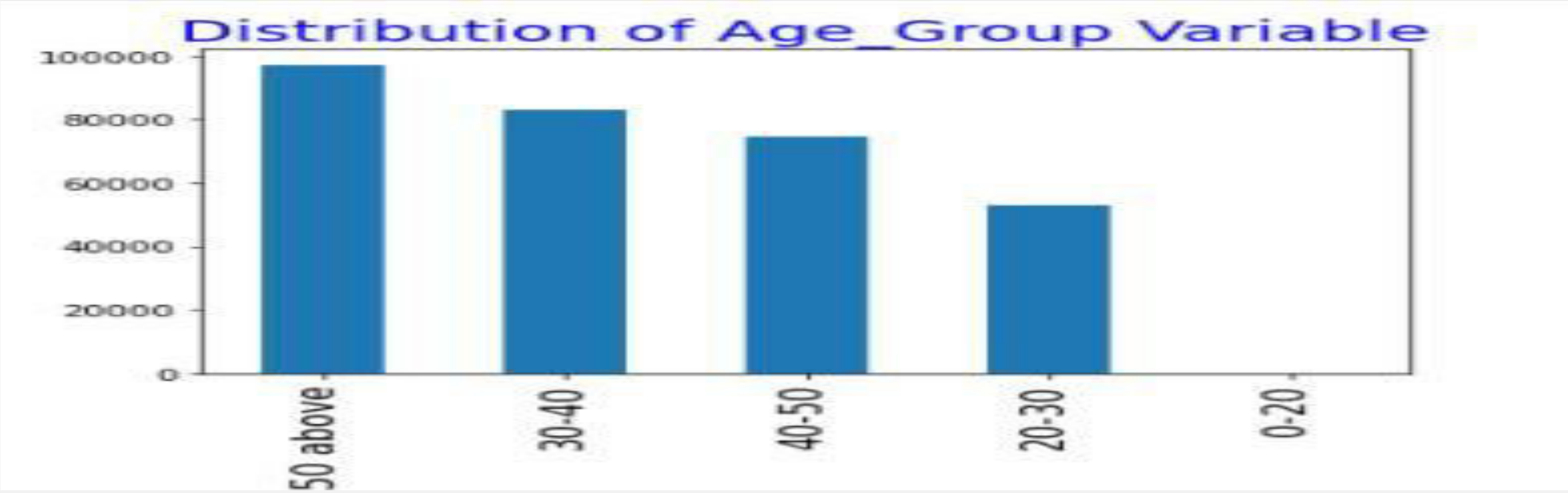
Organization Type

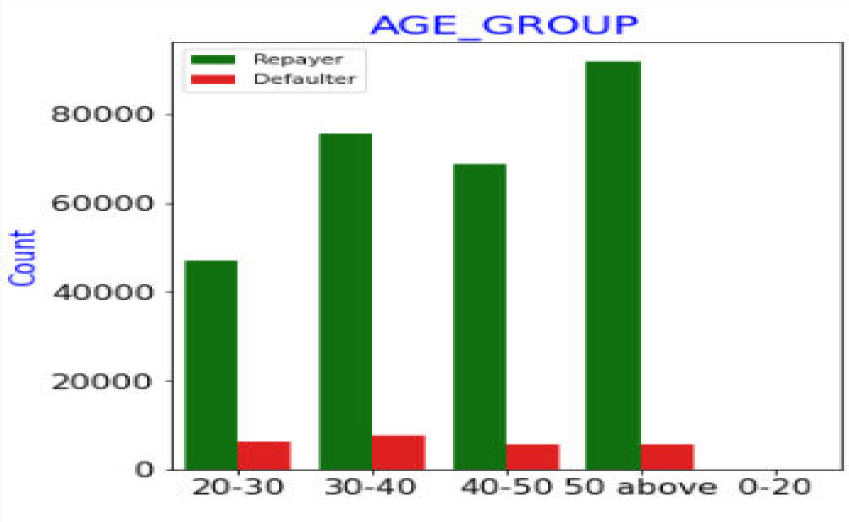


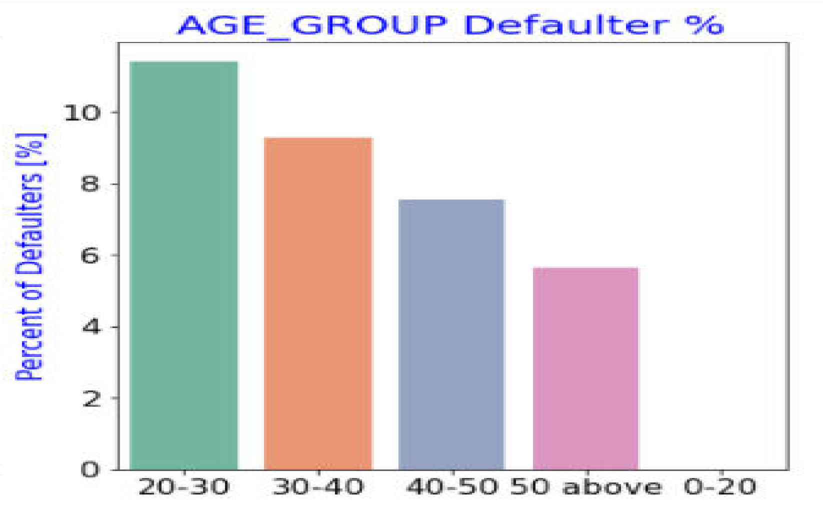
Inferences:

- 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%).
- Self employed people have relative high defaulting rate (10%), and thus should be toughly securitized before being approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting.
- 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



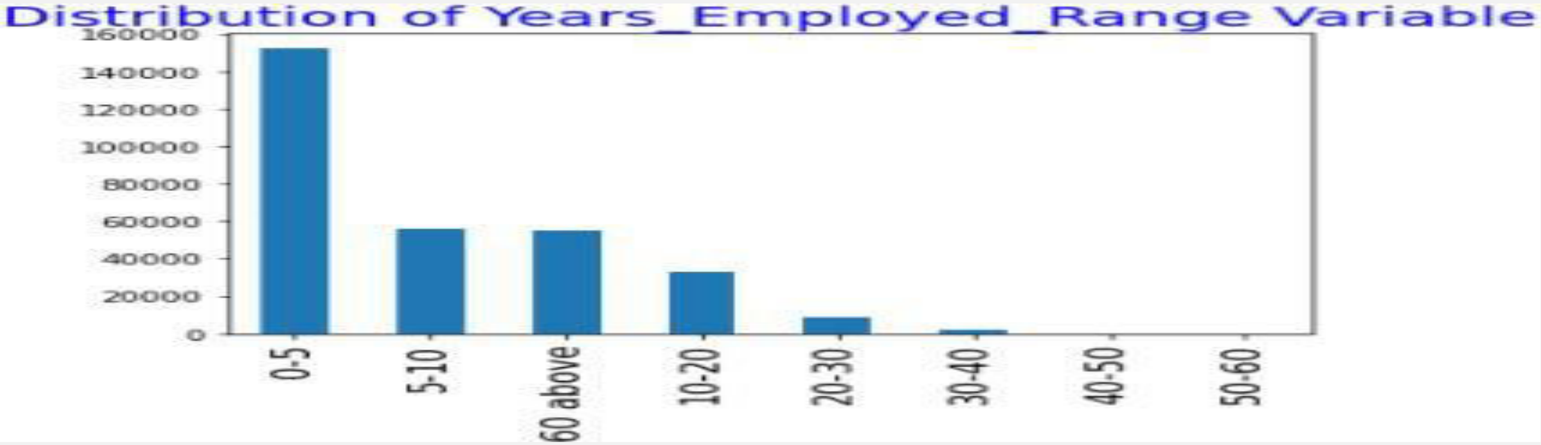


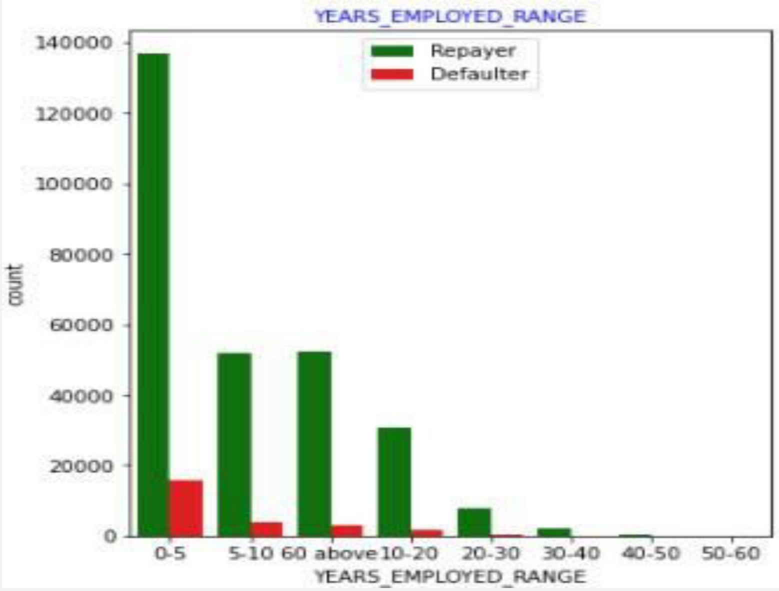


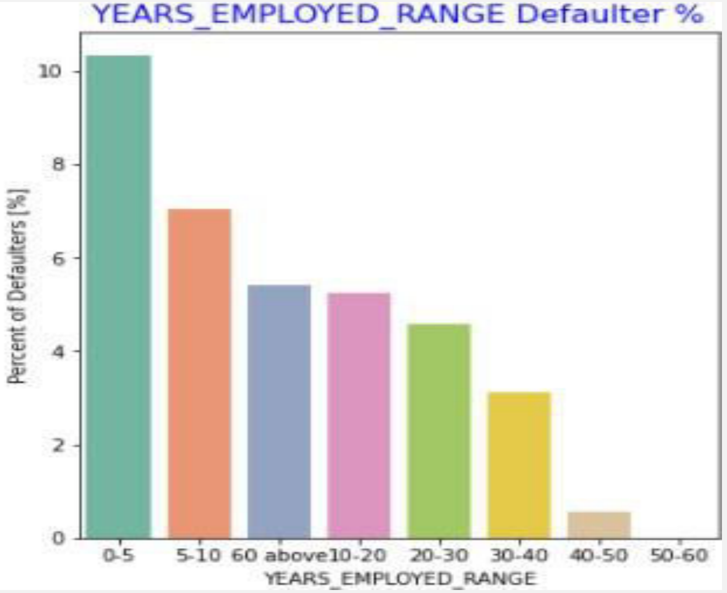
Inferences:

- ❖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





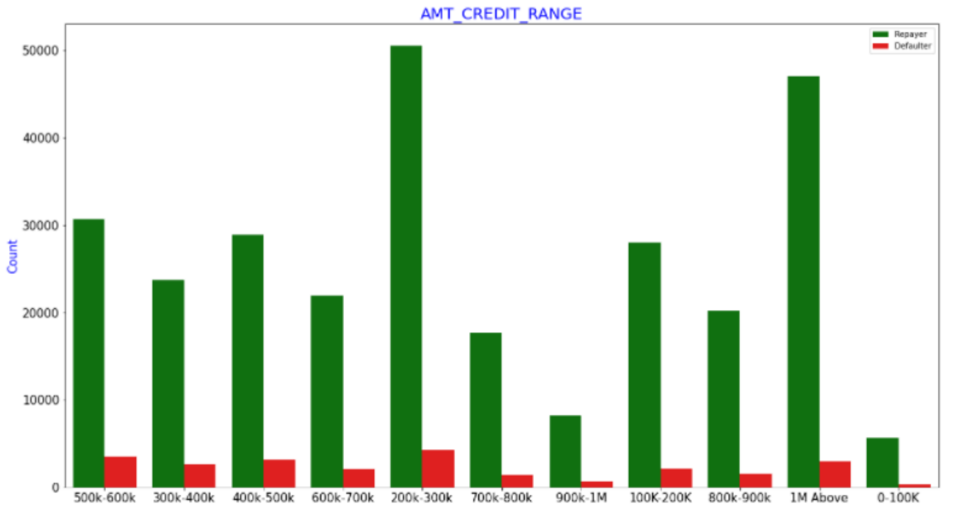


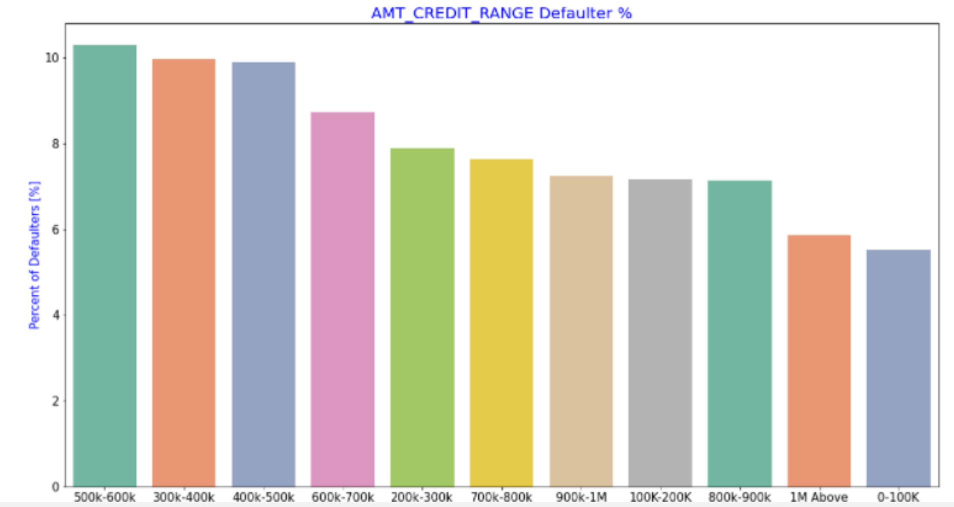
Inference:

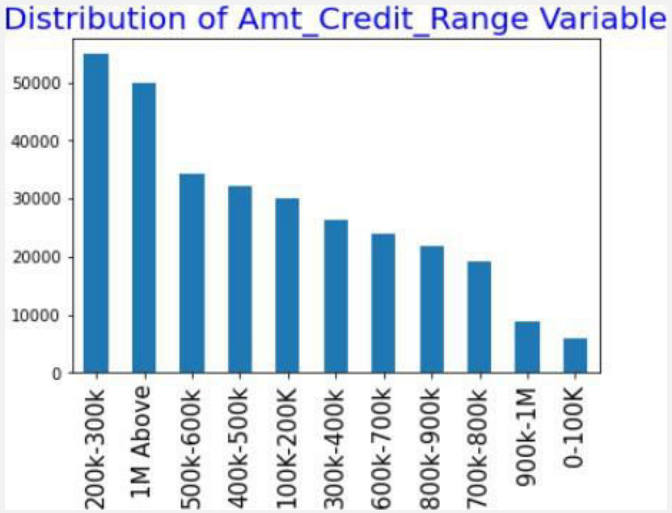
❑Majority of the applicants have been employed in between 0-5 years. The defaulting rating of this group is also the highest which is More than 10%.

❑With increase of employment year, defaulting rate is gradually decreasing with people having 40+ year experience having less than 1% default rate.

Credit Amount


AMT_CREDIT_RANGE

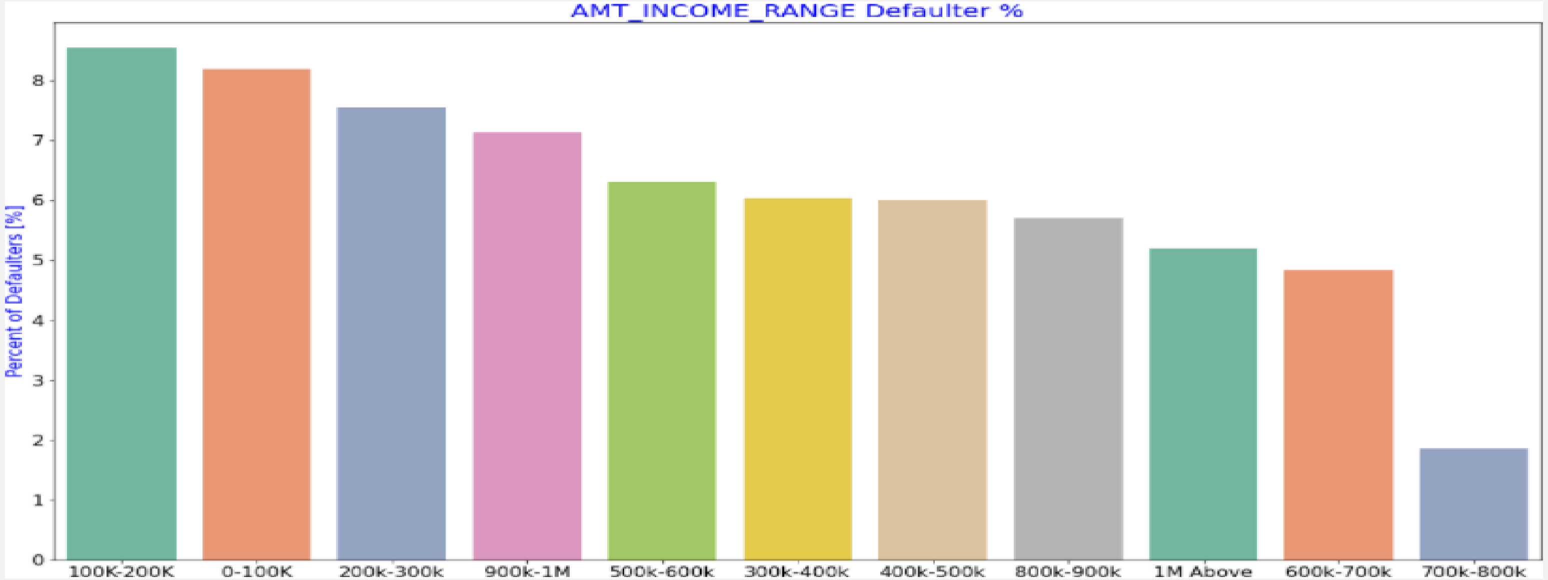

AMT_CREDIT_RANGE Defaulter %

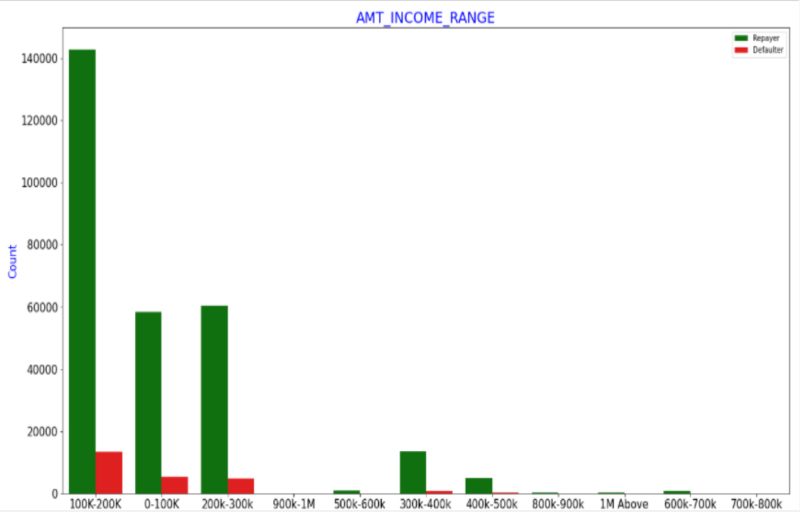

Distribution of Amt_Credit_Range Variable

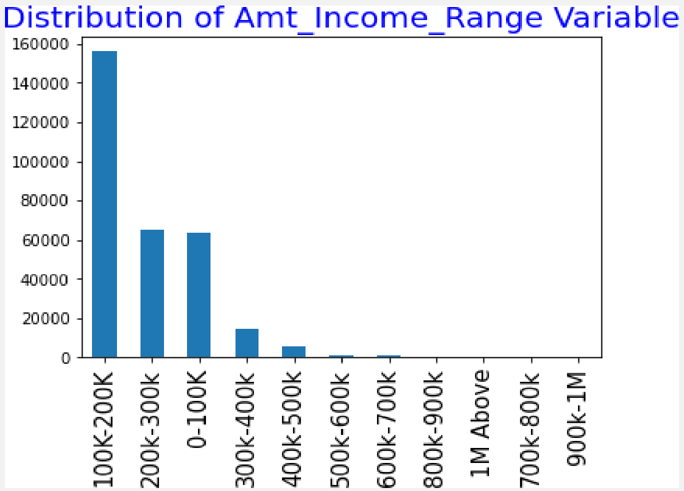
Inference:

- ❖ 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.

Amount Income Range



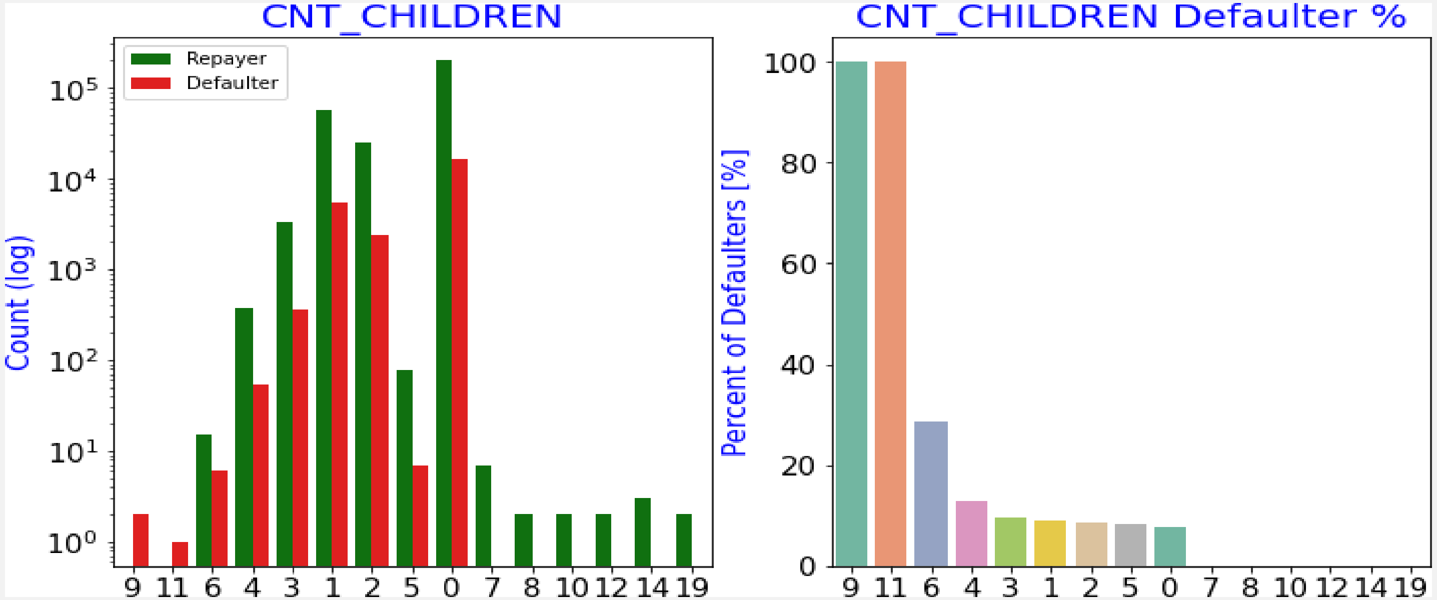


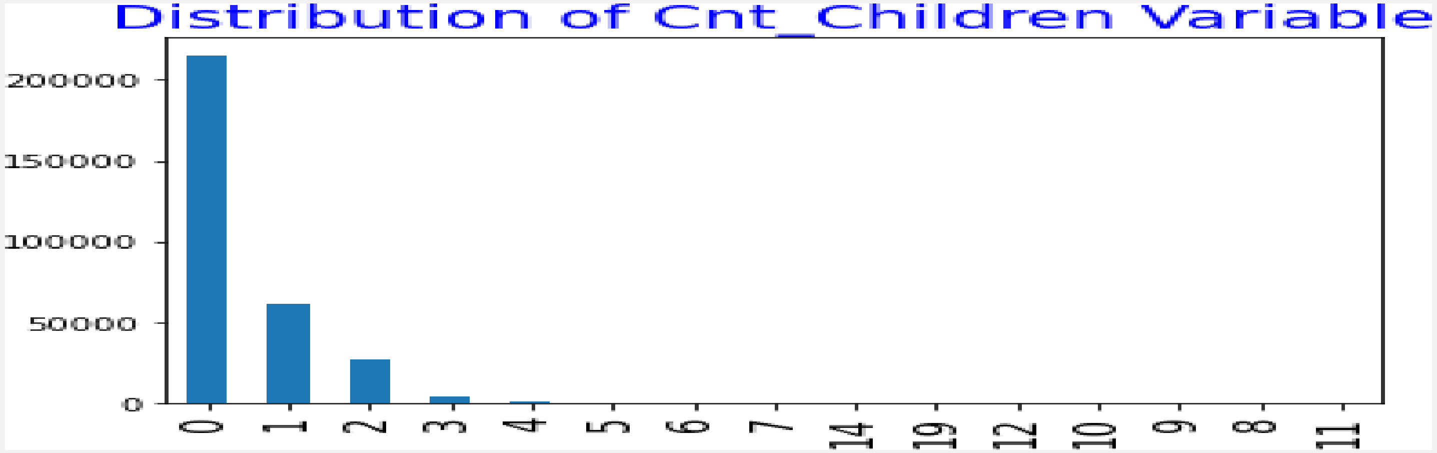


Interference:

- 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.

Number of Children

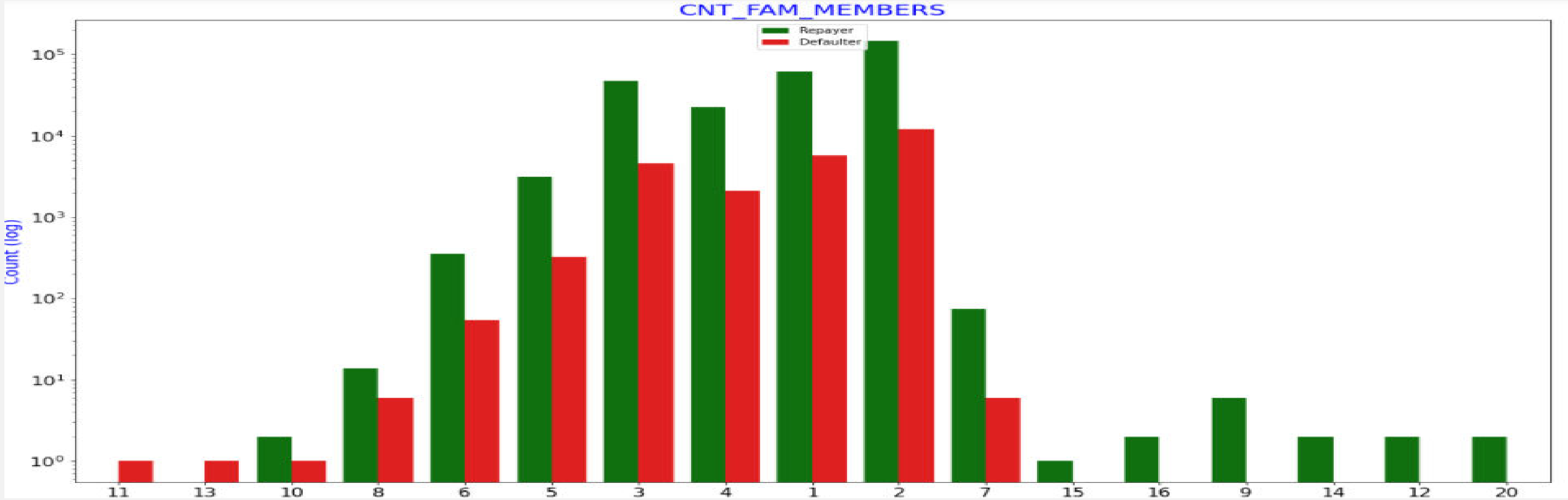


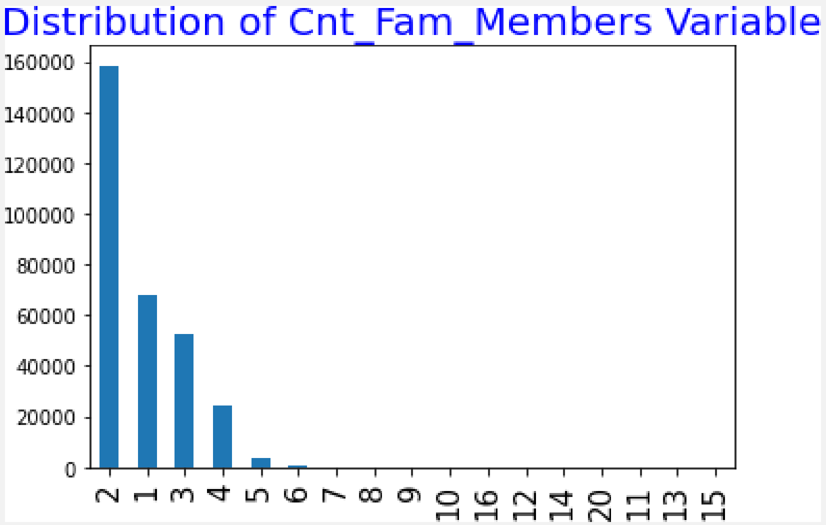


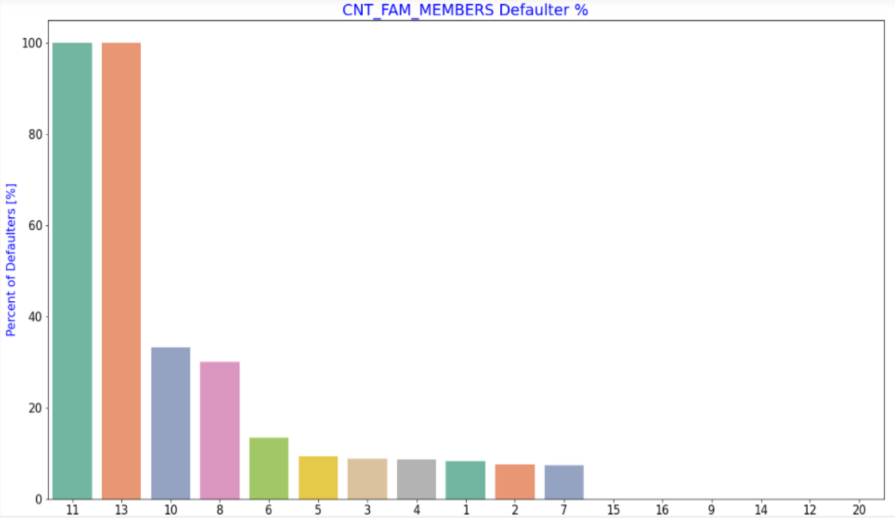
Inference:

- ➢ 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

Number of Family Members





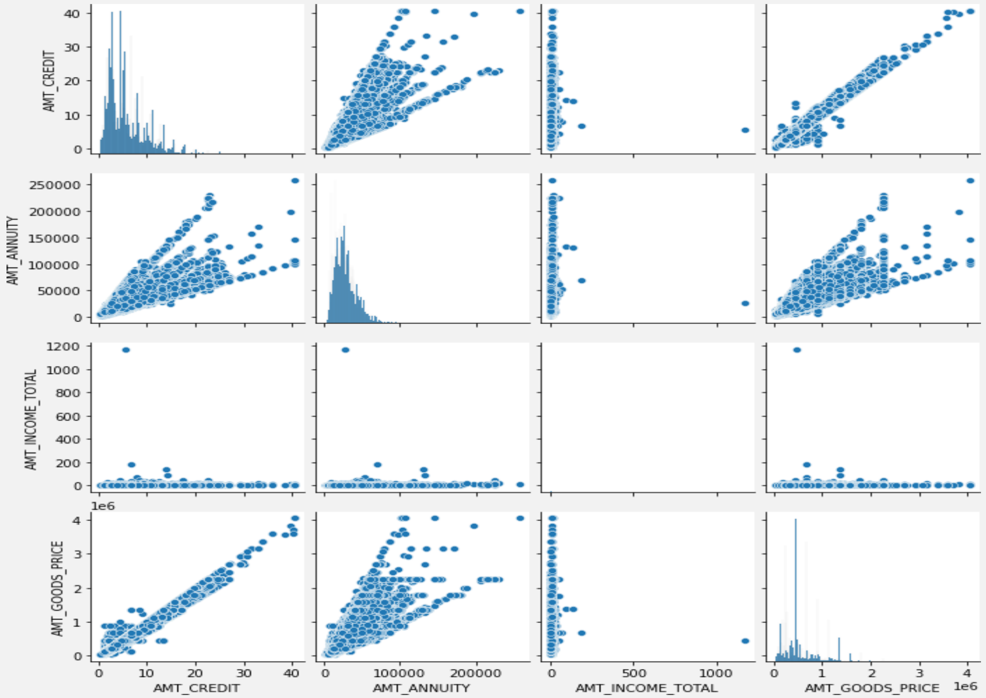


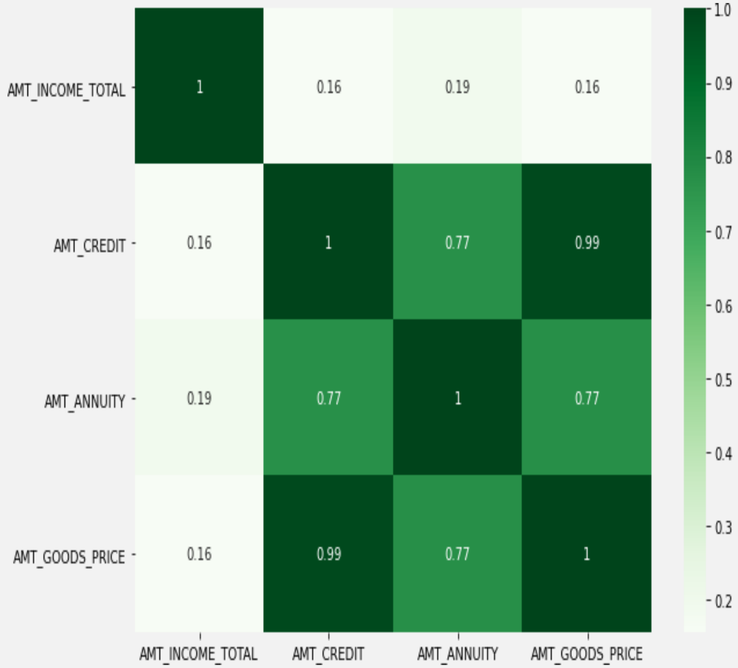
Inference:

❑ Applicants who have higher family members i.e. more than or equal to 11 have higher default rate so their application can be rejected.

BIVARIATE $ MULTIVARIATE ANALYSIS

Amount credit, Amount annuity, Amount income total, Amount goods price

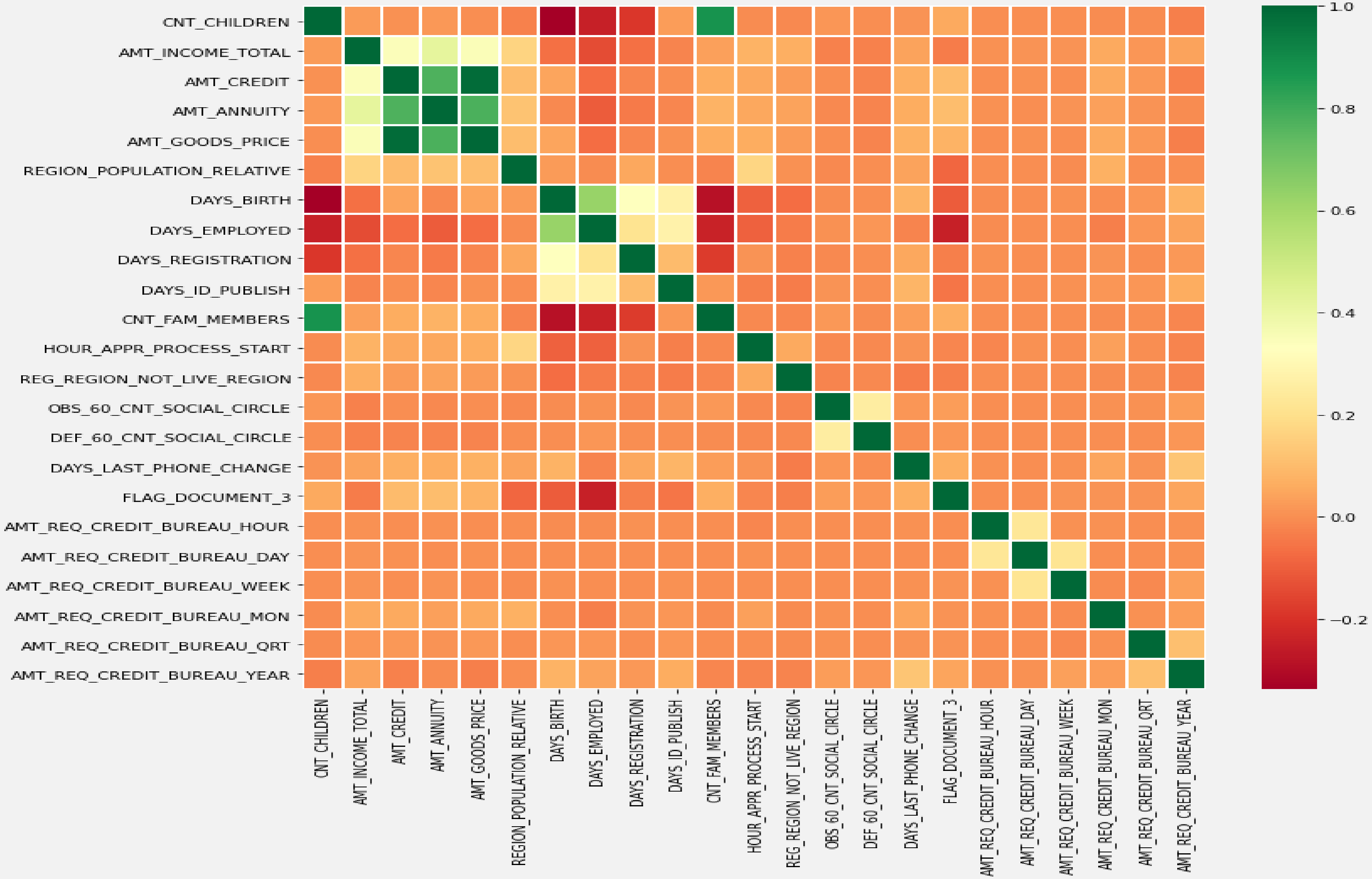




Inferences:

- We can see that there is a very high correlation between amount credit and amount goods price i.e. those, applicants who own goods of high value can take loans of high amount.

Multivariate (Repayers Columns)



Inferences:

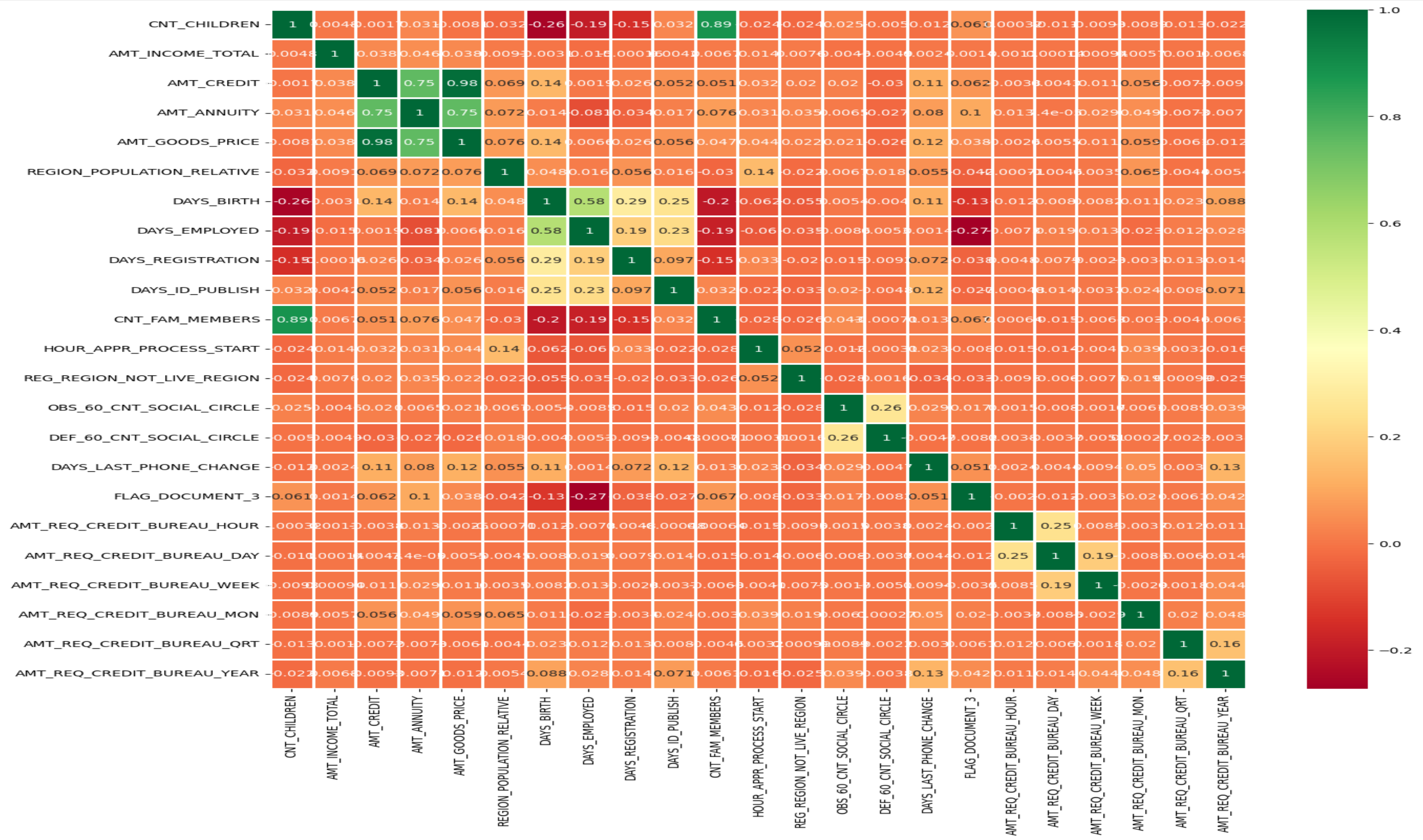
- We can see some correlated factors among re-payers are:

i. Credit amount is highly correlated with the amount of goods price.

ii. Loan annuity, total income is also correlated with the credit amount.

iii. We can even see that repayor's have high correlation with the number of days they are employed .

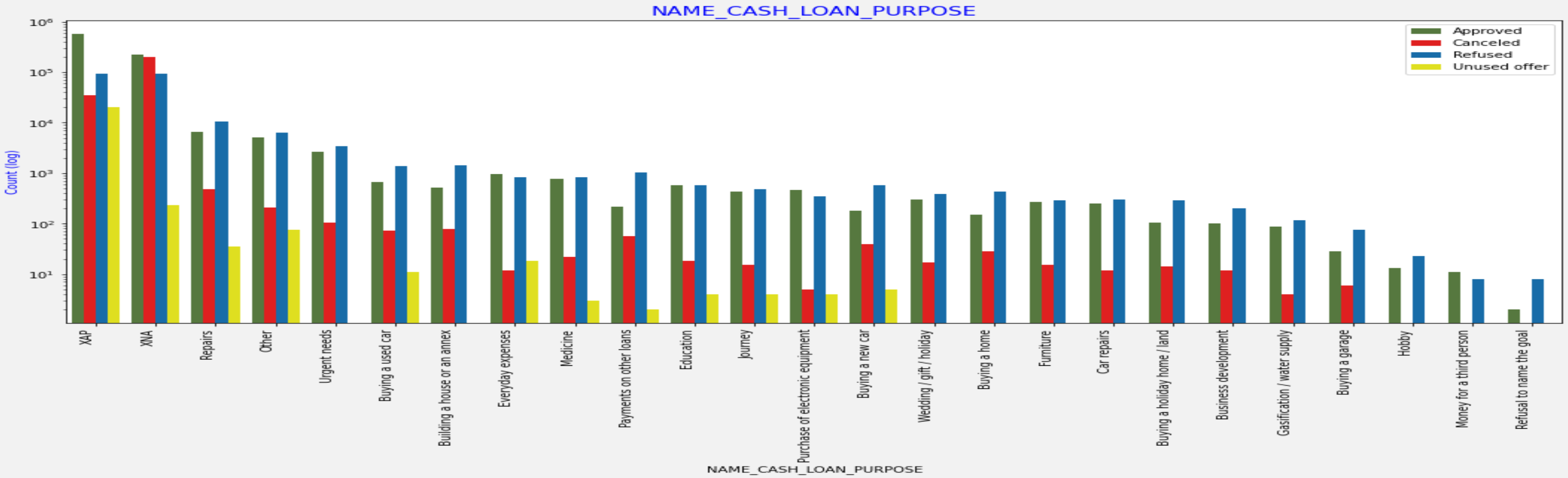
Multivariate (Defaulter Columns)

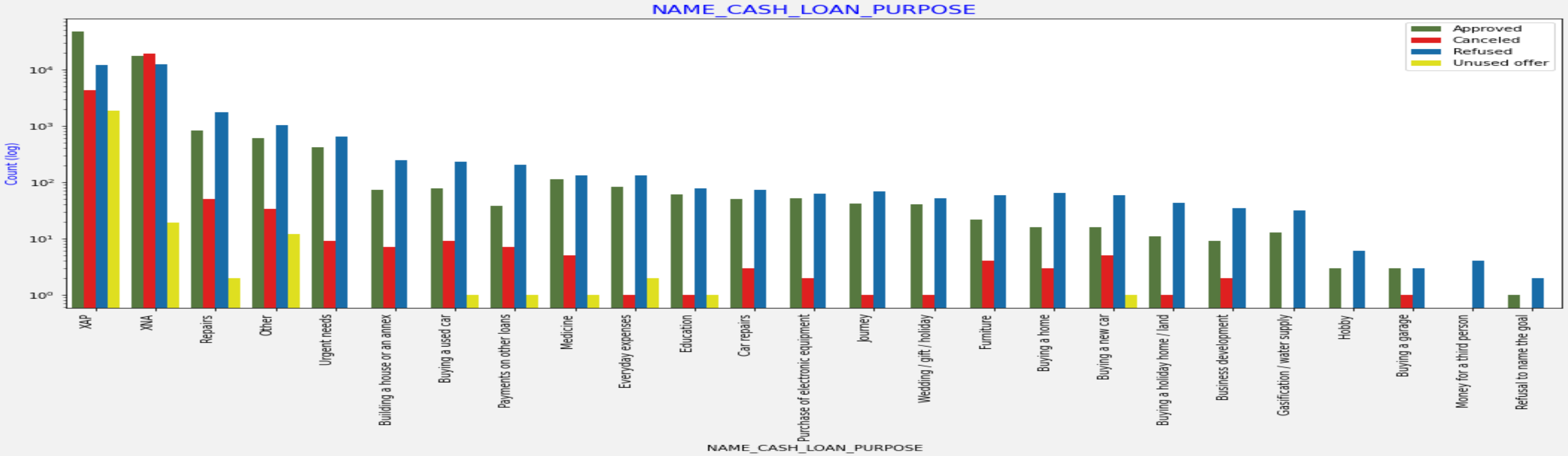


Inferences:

- 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).

Cash Loan Purpose





Inferences:

- ❖The purpose of the loan has a high number of unknown values (XAP, XNA).
- ❖The highest default rate is by those applicants who have taken loan for the purpose of repairs.
- ❖We can also see that bank have rejected a high number of applicants whose purpose is repair or others, or bank charge the high rate of interest so that the applicants refuse the loan.

SUMMARY

Decisive factors for an applicant to be safe borrowers

Loans taken due to hobby, buying garage are repaid.

Those applicants who live in the areas with region rating 1 are known as safe borrower.

The applicants who have income more than 700,000 are less likely to default.

The clients who have academic degree have less default.

Applicants with trade type 4 and 5, industry type 8 have default rate less than 3%.

Those persons who have 0-2 children repay the loans.

The clients who have 40+ year of experience tend to have default rate less than 1%.

Students tend to have no default.

Those age is above 50 have low probability of defaulting.

Decisive factors for an application to be a potential defaulter

The highest number of application rejected by bank are for the house repairs, as they have high default rate.

Industry type 3, type 13 and type 8 has a high default rate.

The client who have family members between 8-10 tend to default so high interest should be charged from them.

Those clients who live in the rented apartments or with their parents tend to mitigate the loss.

Low-skill laborers, drivers and waiters/barmen staff, security staff have high defaulting rate.

Clients who have less than 5 years of employment have high default rate.

As the credit amount increase to 3M, there is an increase in defaulters.

The age group 20-40 have high defaulting rate.

Male applicants have high defaulter rate than female applicant.

Decisive factors for an application to be a potential defaulter

Those applicants who have children equal or more than 9 have 100% default so their application can be rejected.

The person who have studied Lower secondary education, incomplete education have the high defaulting rate.

The client who are unemployed or in maternity leave have high defaulting rate.

Those person who are in civil marriage or single have high default.

Applicants who take loan of 300k-600k should charge high interest as they tends to default.

Those people who live in areas with region rating 3 have the highest defaults.

We can see that the applicants have income less or equal to 300,000 have high probability of defaulting, so they should be offered with high interest rate than others.

Those applicant who have more than or equal to 11 person in a family tend to default so their application can be rejected.

Those applicants should be expelled who have 4-8 children as they tend to default and if loan is given then high interest rate should be charged.

Thank you