CREDIT EDA ASSIGNMENT

Submitted by
Abhilasha Garg
Batch ID: 1972
IIT-B UpGrad DS C39 Dec 2021

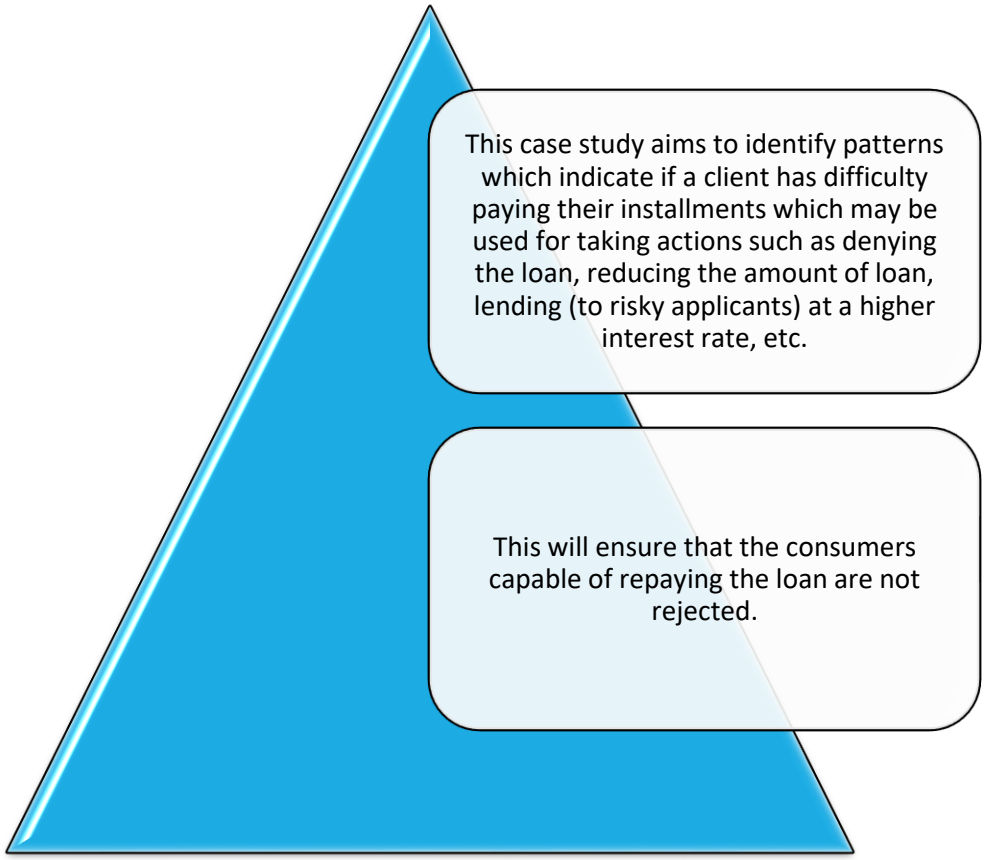
OBJECTIVE

Understanding of risk analytics in banking and financial services

Knowing how data is used to minimize the risk of losing money while lending to customers.

Understand how consumer attributes and loan attributes influence the tendency of default.

TARGET


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

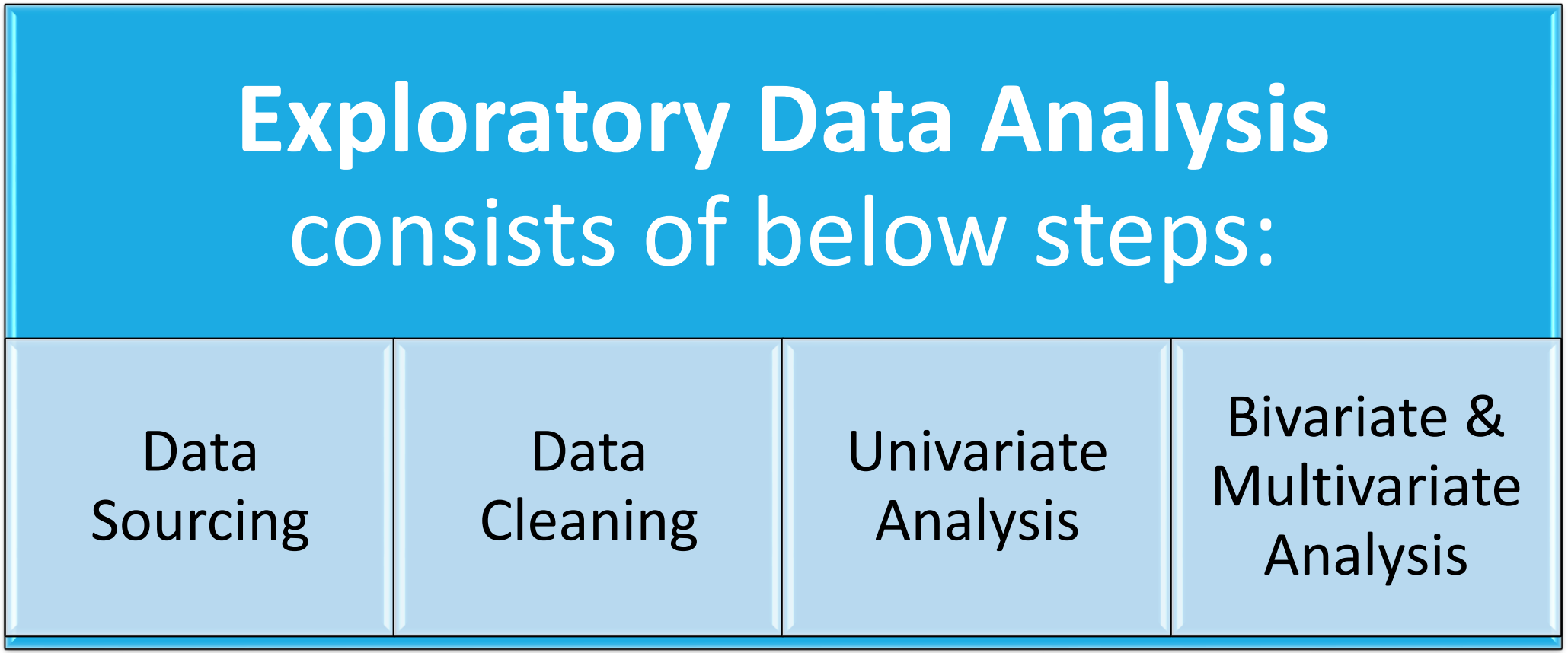
DATASET USED

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.**

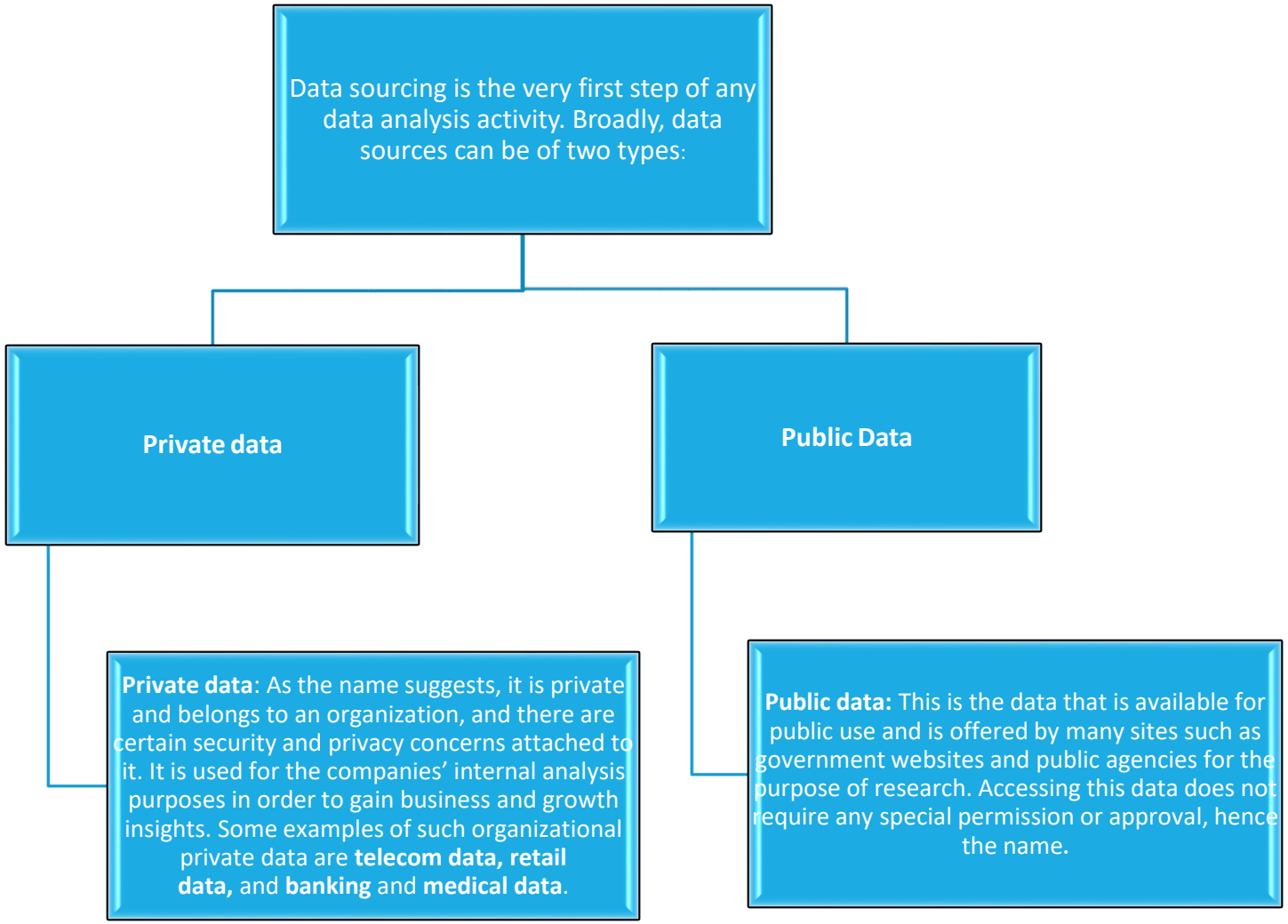
Previous_application.csv contains information about the client's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

Columns_description.csv is data dictionary which describes the meaning of the variables.

EDA PROCESS

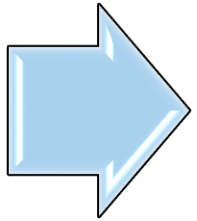

Exploratory Data Analysis
consists of below steps:
Data Sourcing
Data Cleaning
Univariate Analysis
Bivariate & Multivariate Analysis

STEP 1. DATA SOURCING



STEP 2: DATA CLEANING

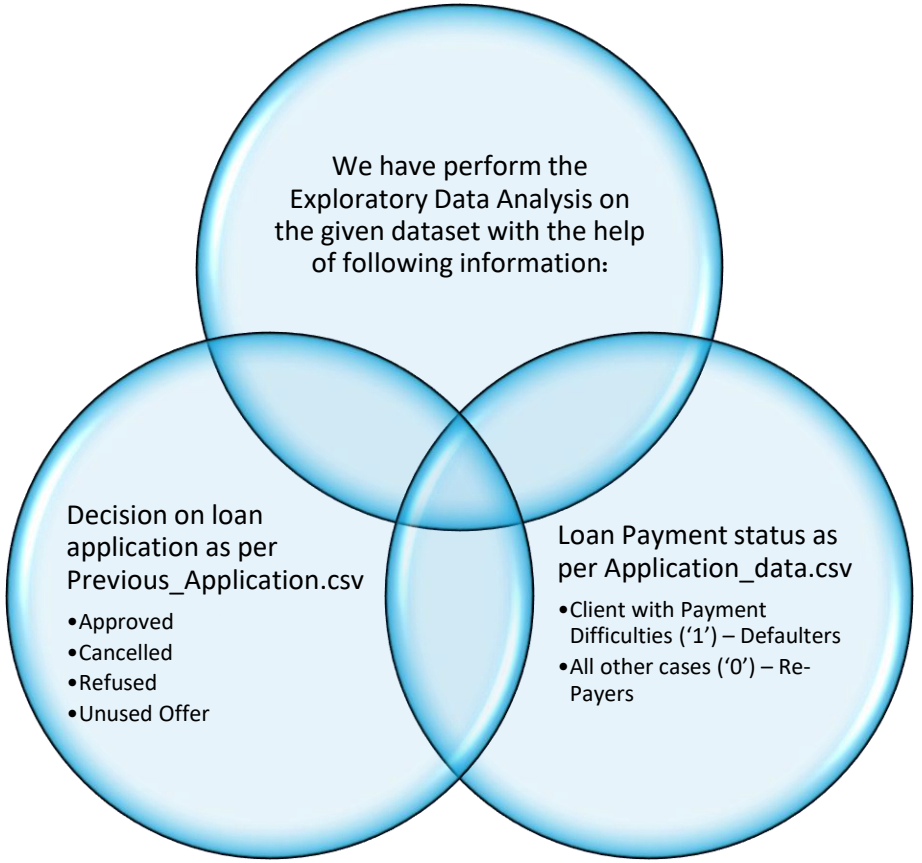
Irregularities may appear in the form of **missing values, anomalies/outliers, incorrect format** and **inconsistent spelling**, etc. These irregularities may propagate further and affect the assumptions and analysis based on that dataset and may hamper the further process of machine learning model building. Hence, data cleaning is a very important step in EDA.



Consists of below steps:

- Identifying data type
- Fixing Rows and Columns
- Imputing / Renaming missing values
- Handling Outlier
- Standardising values
- Fixing invalid values
- Filtering data

APPROACHES TAKEN



APPROACHES TAKEN

Handling Null Values:

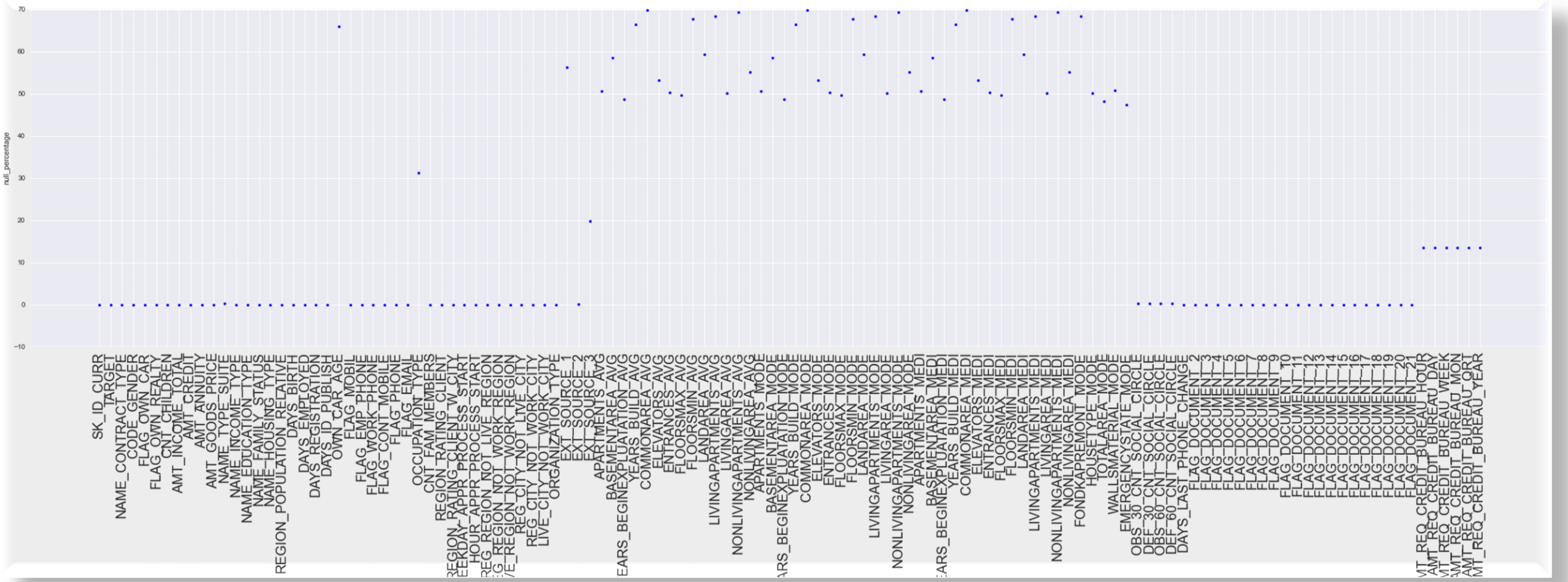
- Check for missing values % in each column of Application_data and create a new Data Frame with no column having more than 40% of missing data. So that we are not dropping anything from the main dataset.
- Imputing data like mean , median , mode or new category depending if the Data type is Numerical or Categorical. Whether it has Outlier or not.

Handling Outliers:

- Capping the outliers more than Maximum values with Max value and outlier less than minimum value with minimum value after plotting it on BoxPlot/ScatterPlot

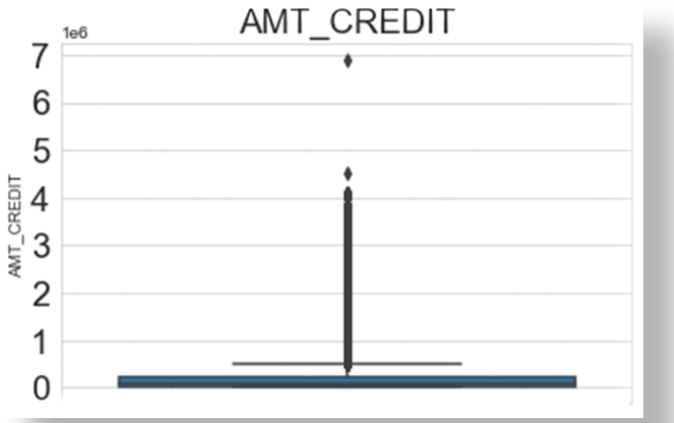
Standardize the values

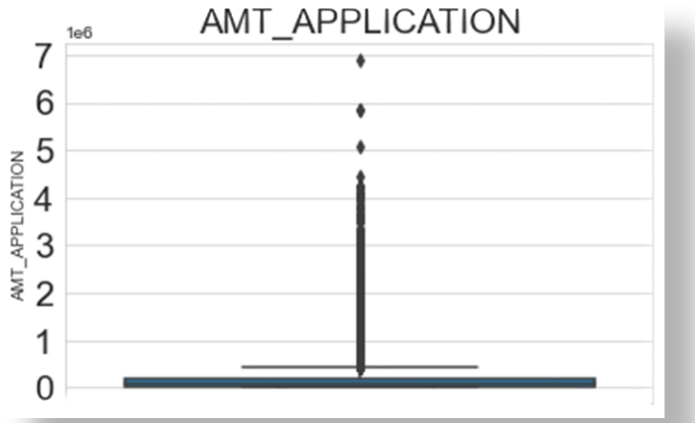
- Checking the Data type in case of Dates, Age and the value stored should make sense. Like in our case the days_birth was negative so making them positive.
- Imputing the values in case of null values in rows.

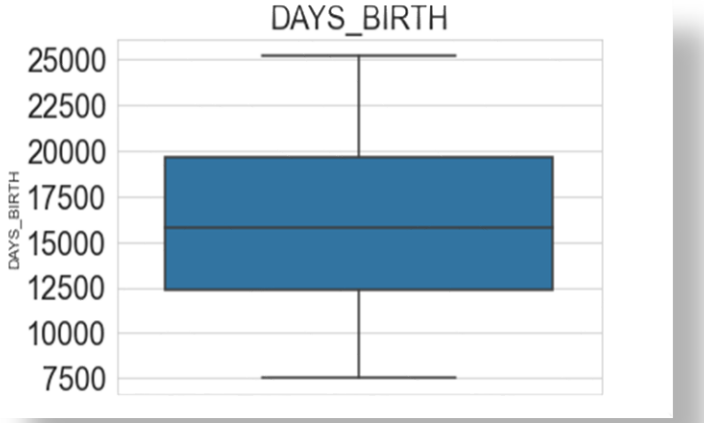


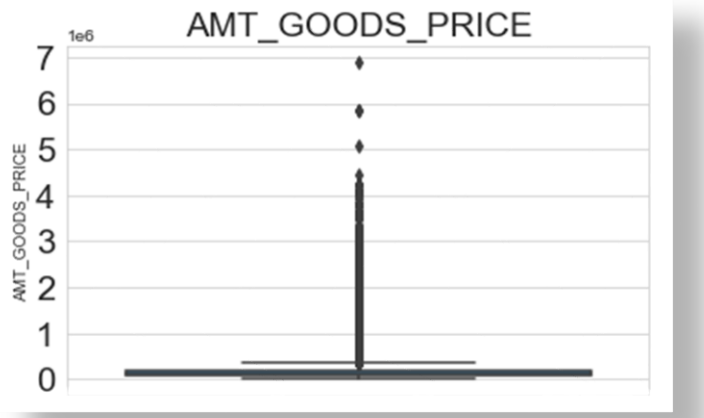
SCATTER PLOT FOR MISSING VALUES

We have lots of columns with more than 40% missing values









OUTLIERS

In the above boxplot we could see:

- Days_Birth do not have outlier
- Amt_income_total and Sellerplace_area has huge outliers
- Amt_Credit,Amt_Application, Amt_Goods_Price have some outliers

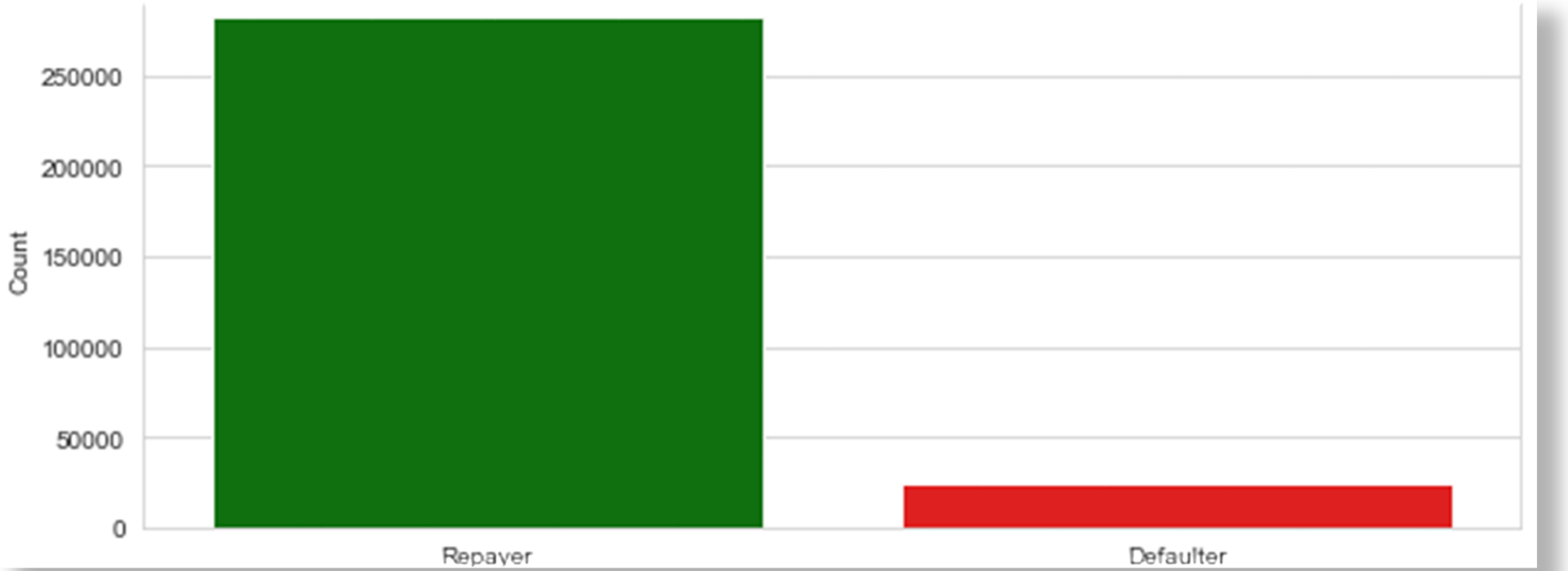
STEP 3: DATA ANALYSIS

Univariate Analysis

- Categorical Unordered Analysis
- Categorical Ordered Analysis
- Statistics on Numerical Variable
- Numerical Variable Analysis

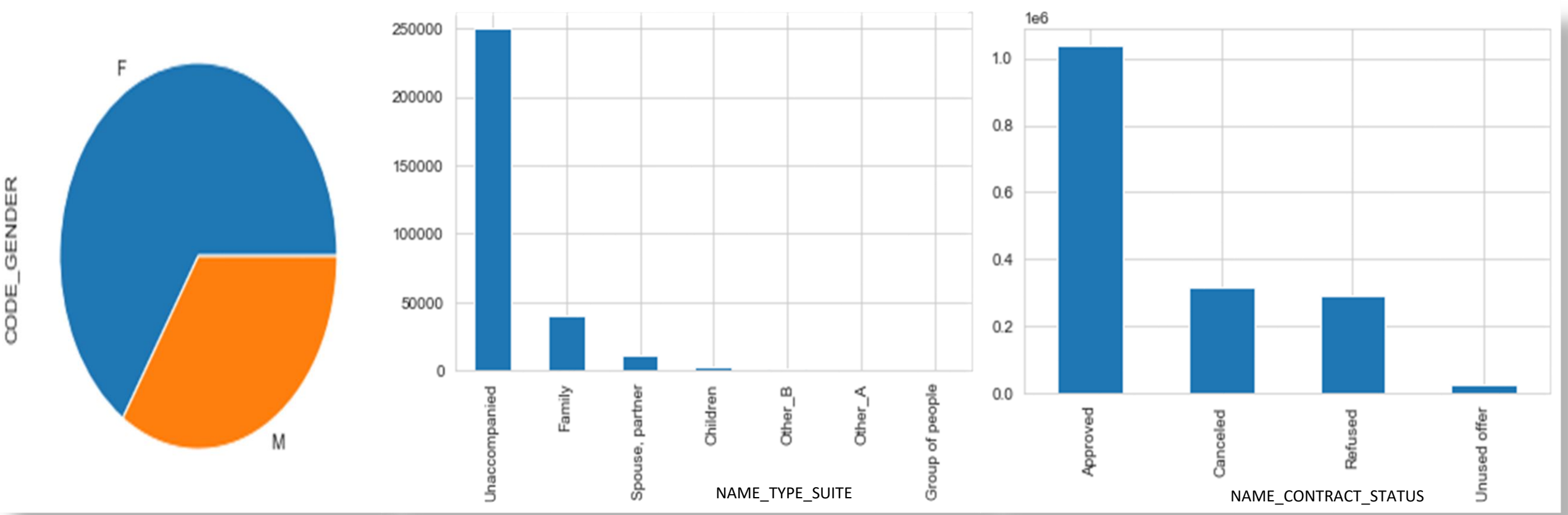
Bivariate and MultiVariate Analysis

- Numeric-Numeric Analysis
- Numeric-Categorical Analysis
- Correlation versus Causation
- Categorical-Categorical Analysis



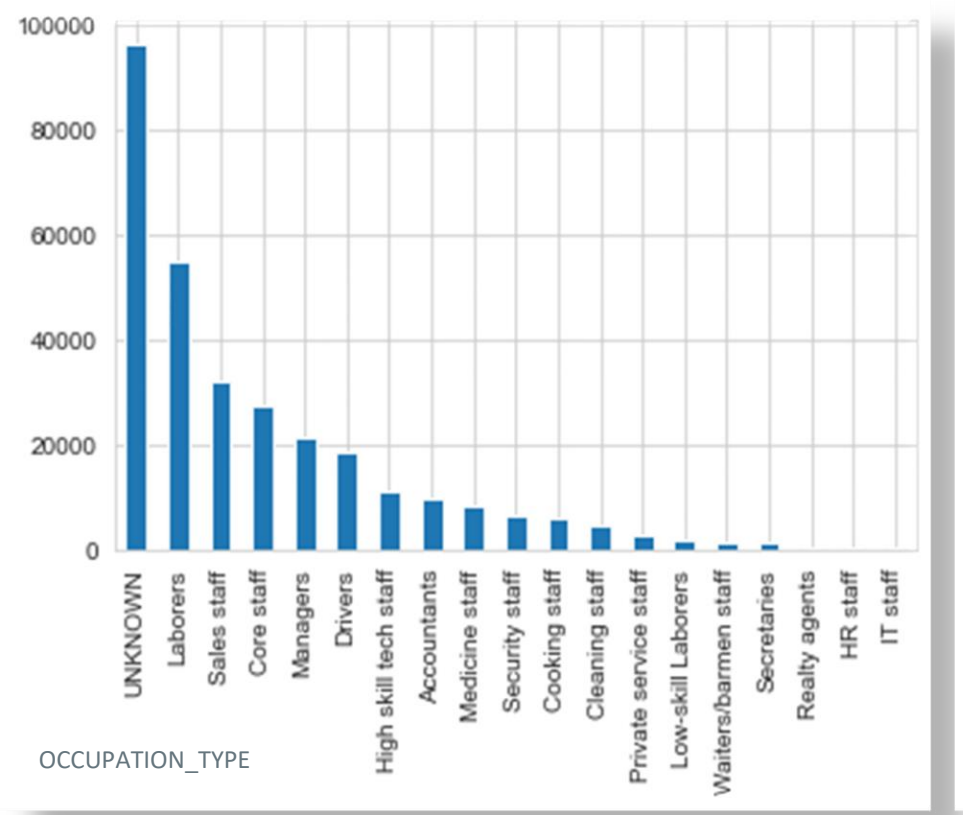
DATA IMBALANCE

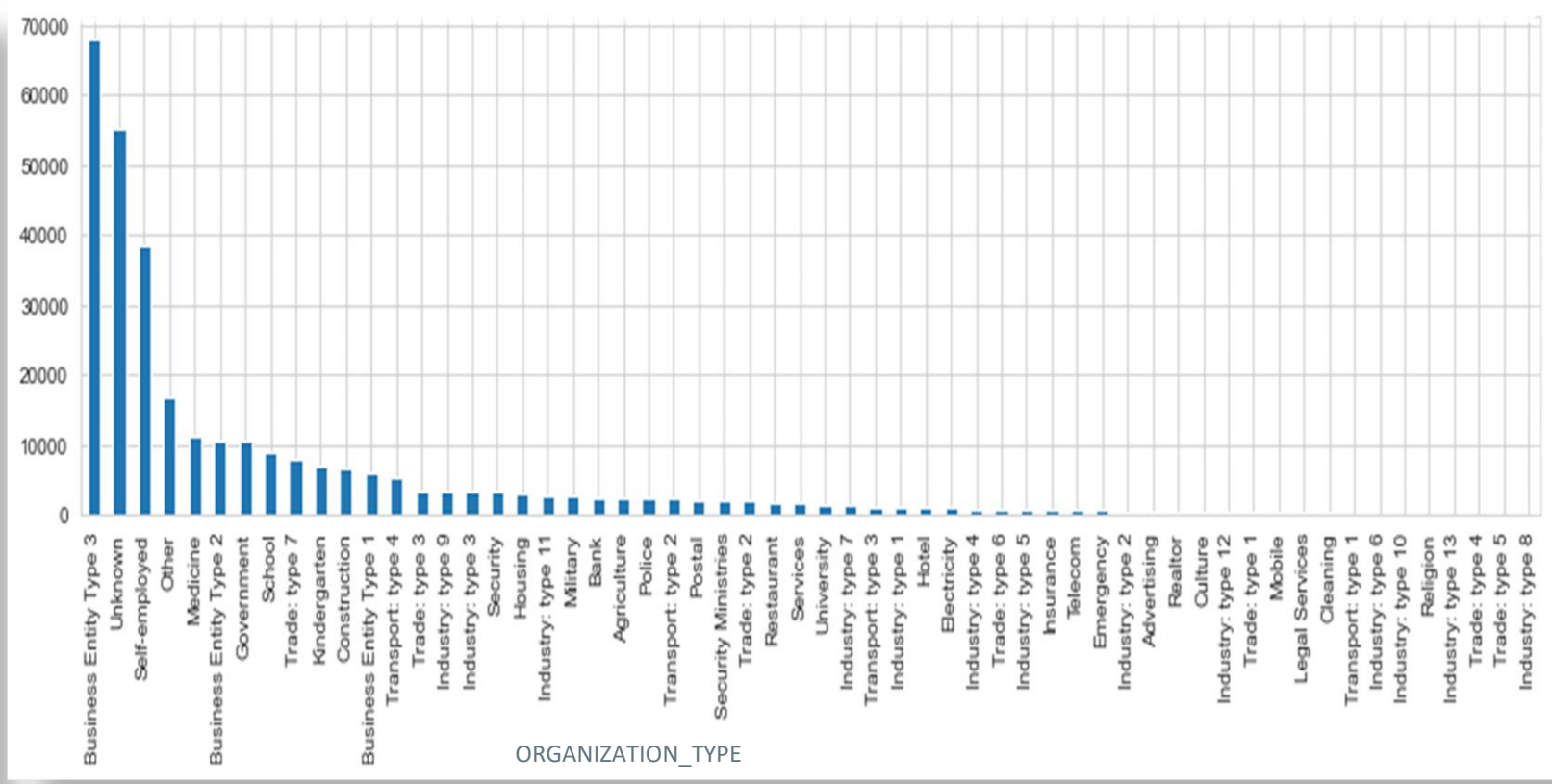
Re-Payer - 91.93%
Defaulter - 8.07%
Ratios of imbalance is 11.39 : 1



UNIVARIATE ANALYSIS

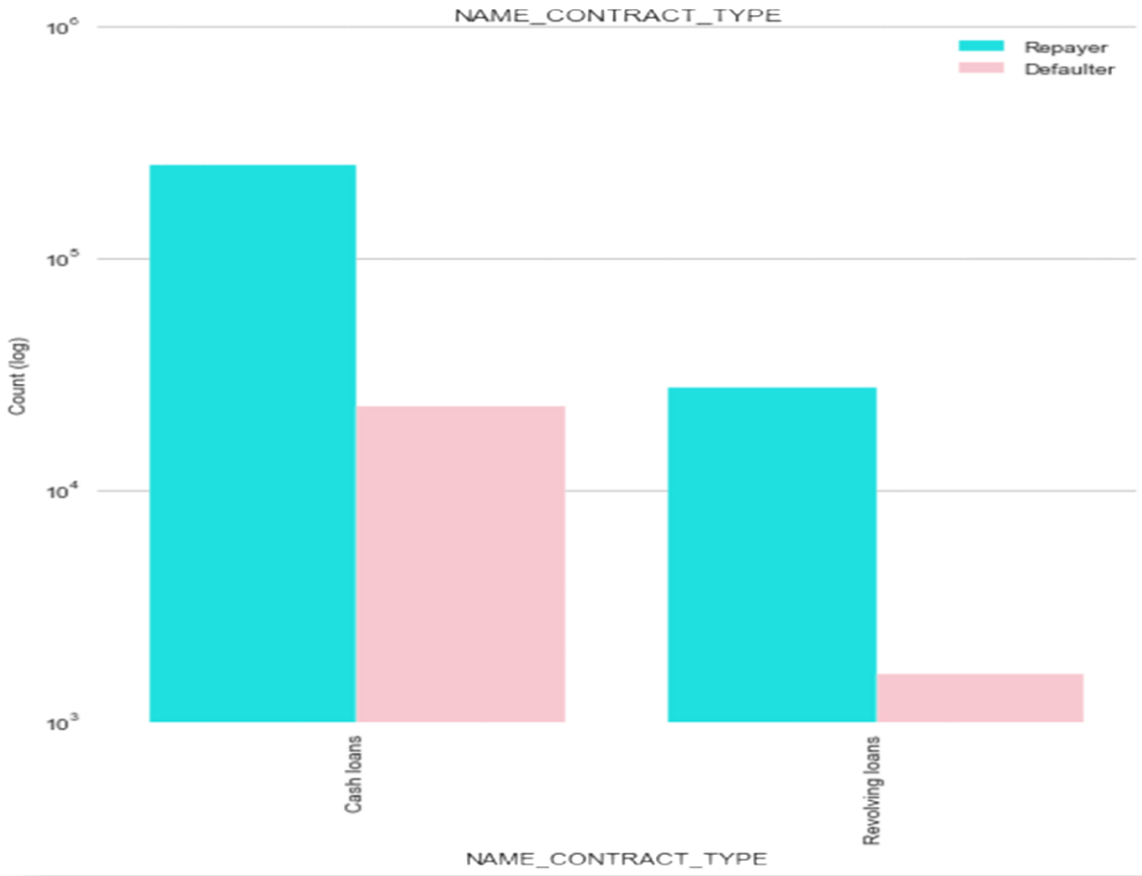
- Applicants are more Female as compare to Male
- Mostly applicants were alone during filling the application and some were with there families
- Most of the application were Approved

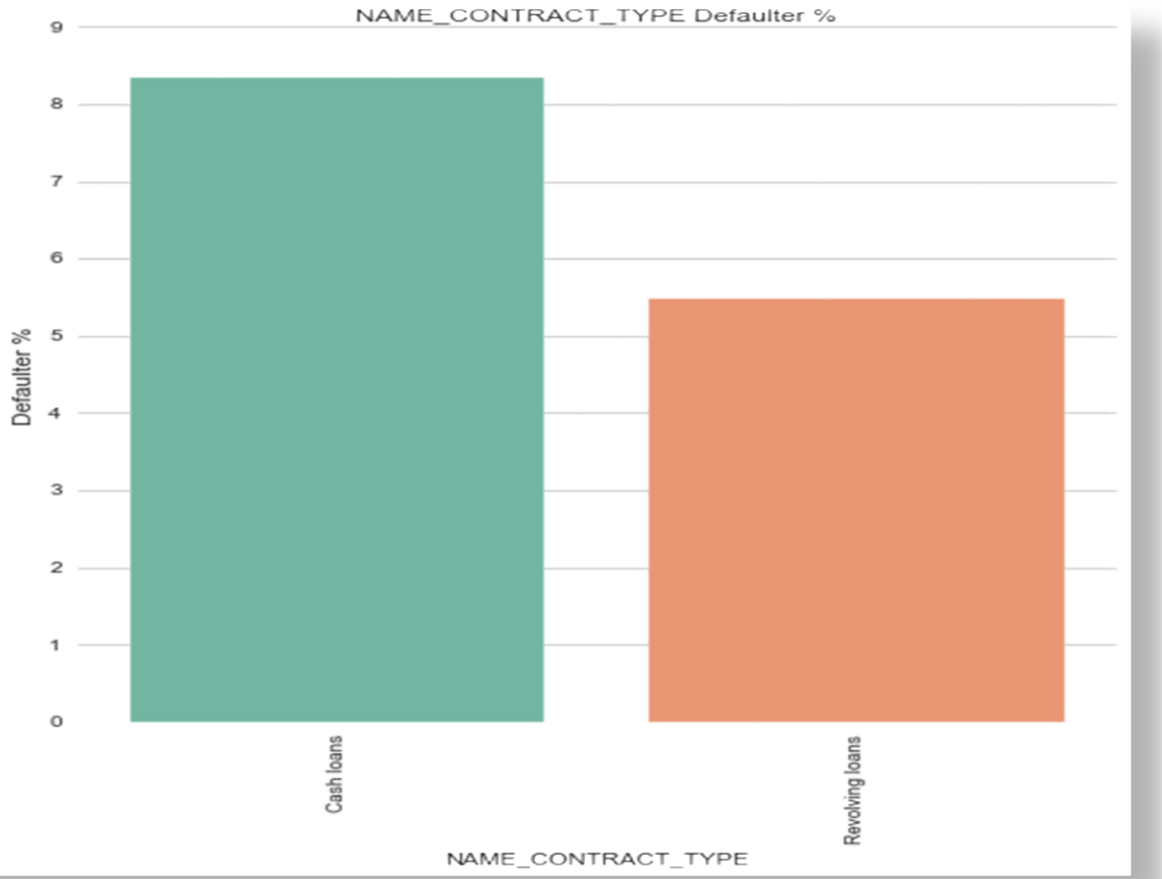




UNIVARIATE ANALYSIS

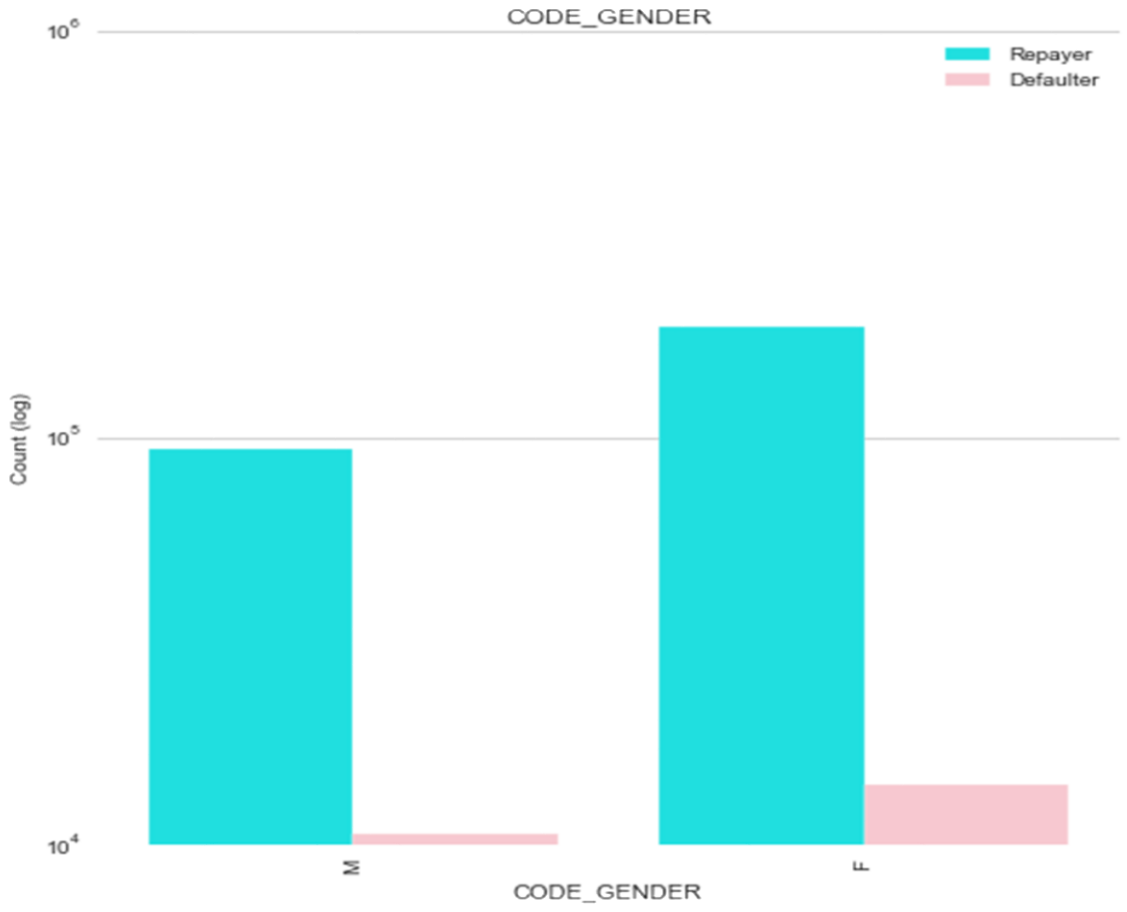
- Most of the applicants we are not aware of the occupation type but the second largest is Laborers and least is IT staff
- Mostly applicants belongs to Business Entity Type 3 Organization and Self Employed

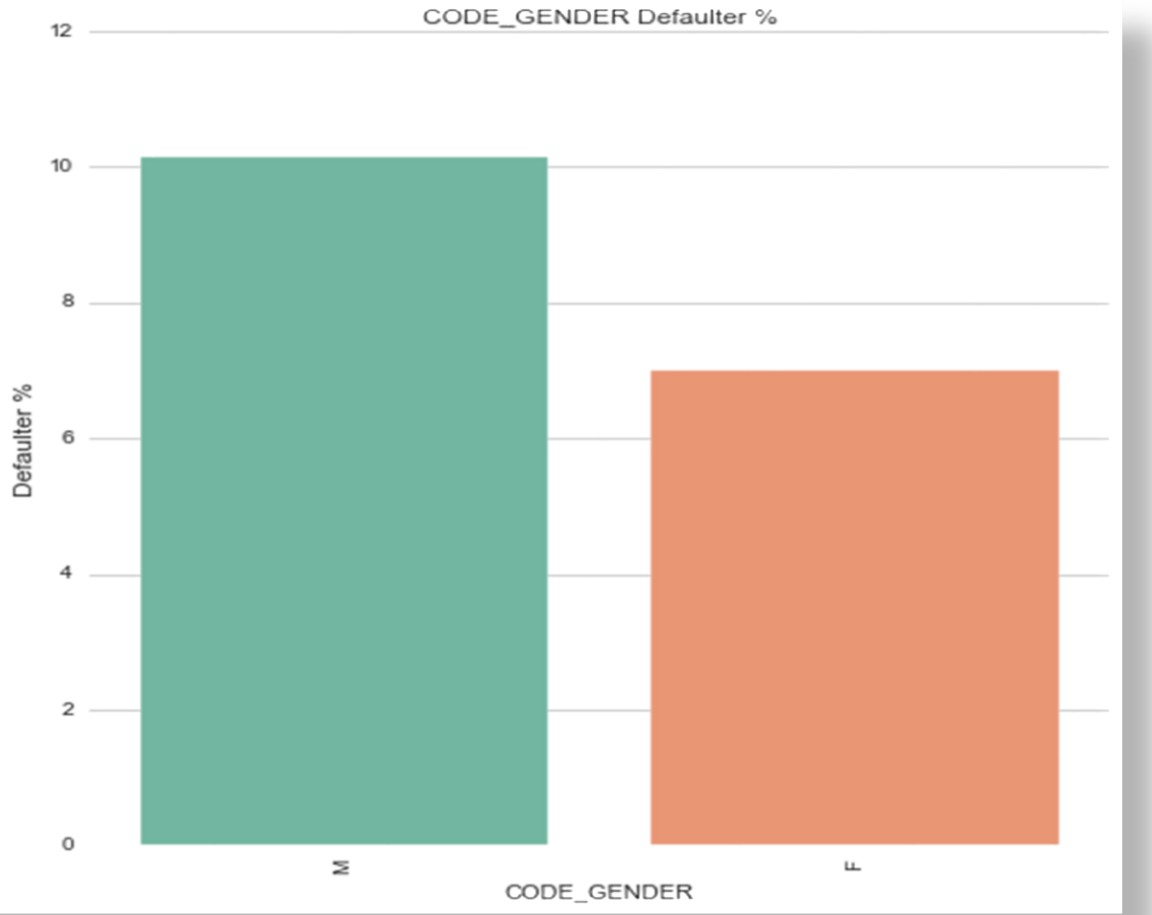




BIVARIATE ANALYSIS

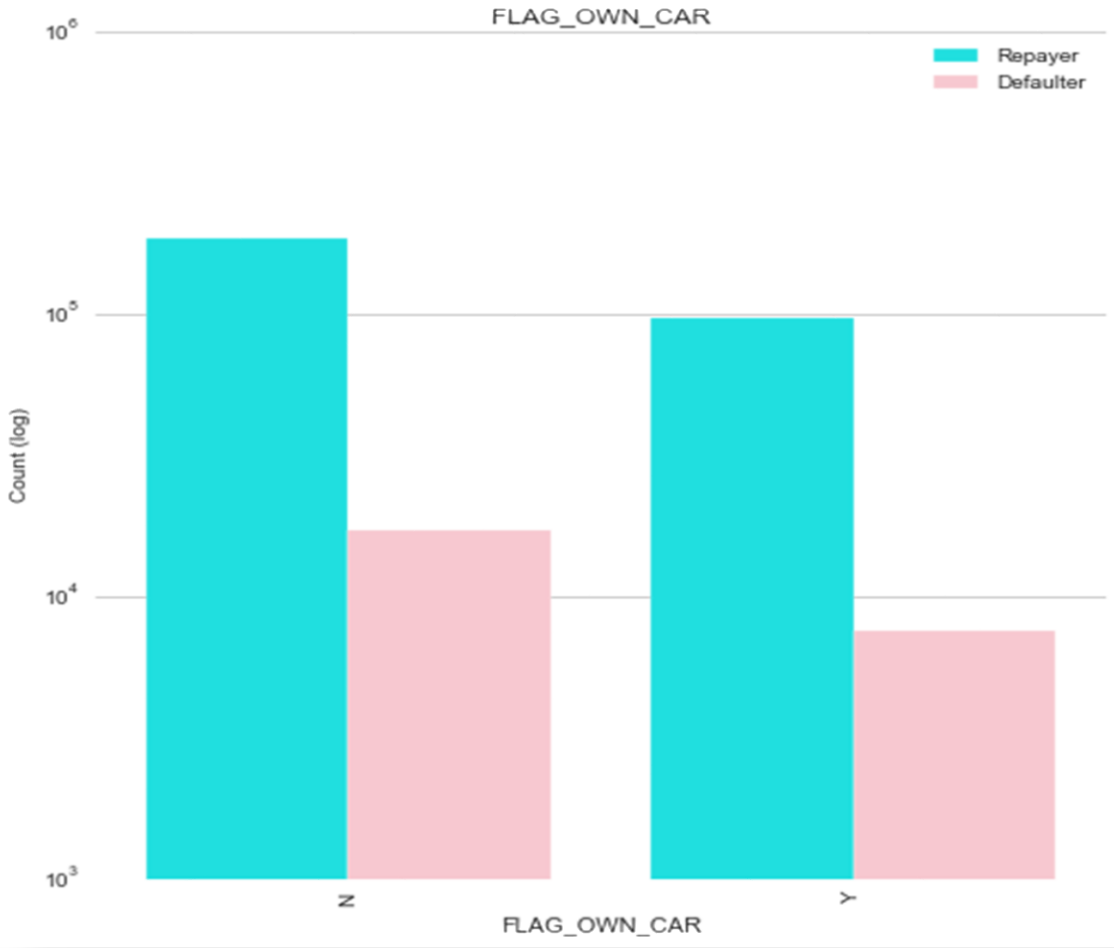
- We could see the Revolving loans are in very less percentage still the defaulter percentage is high

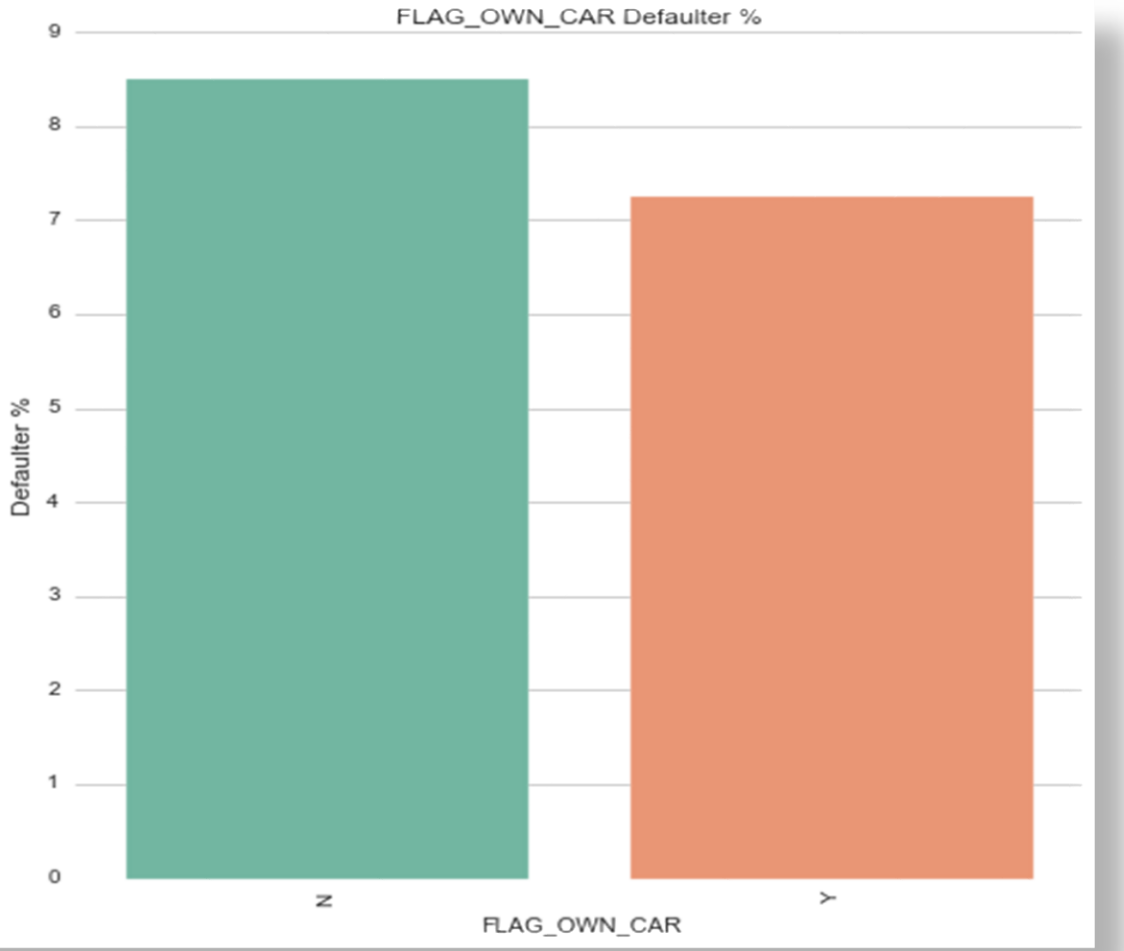




BIVARIATE ANALYSIS

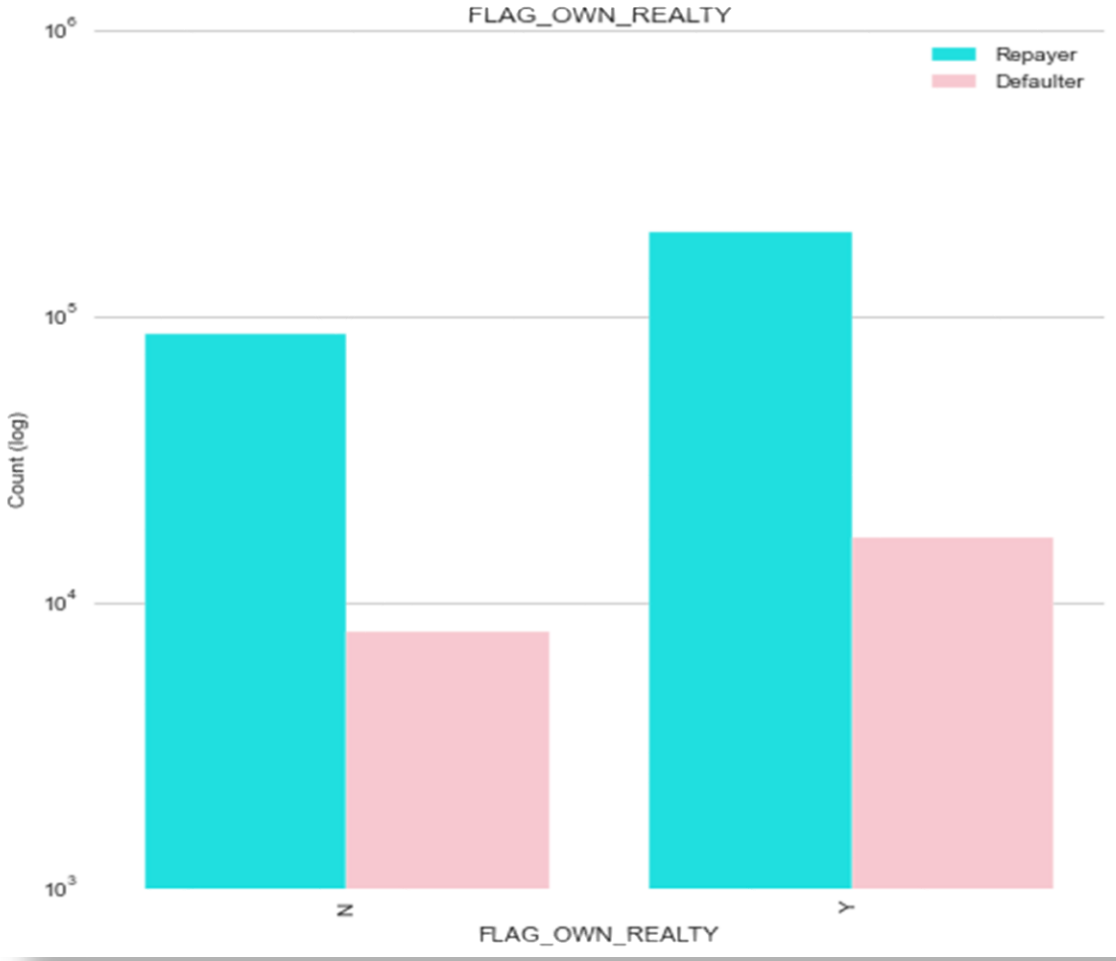
- We could see female applicants double as compare to male applicants but still Defaulter % is high among male as compared to female.

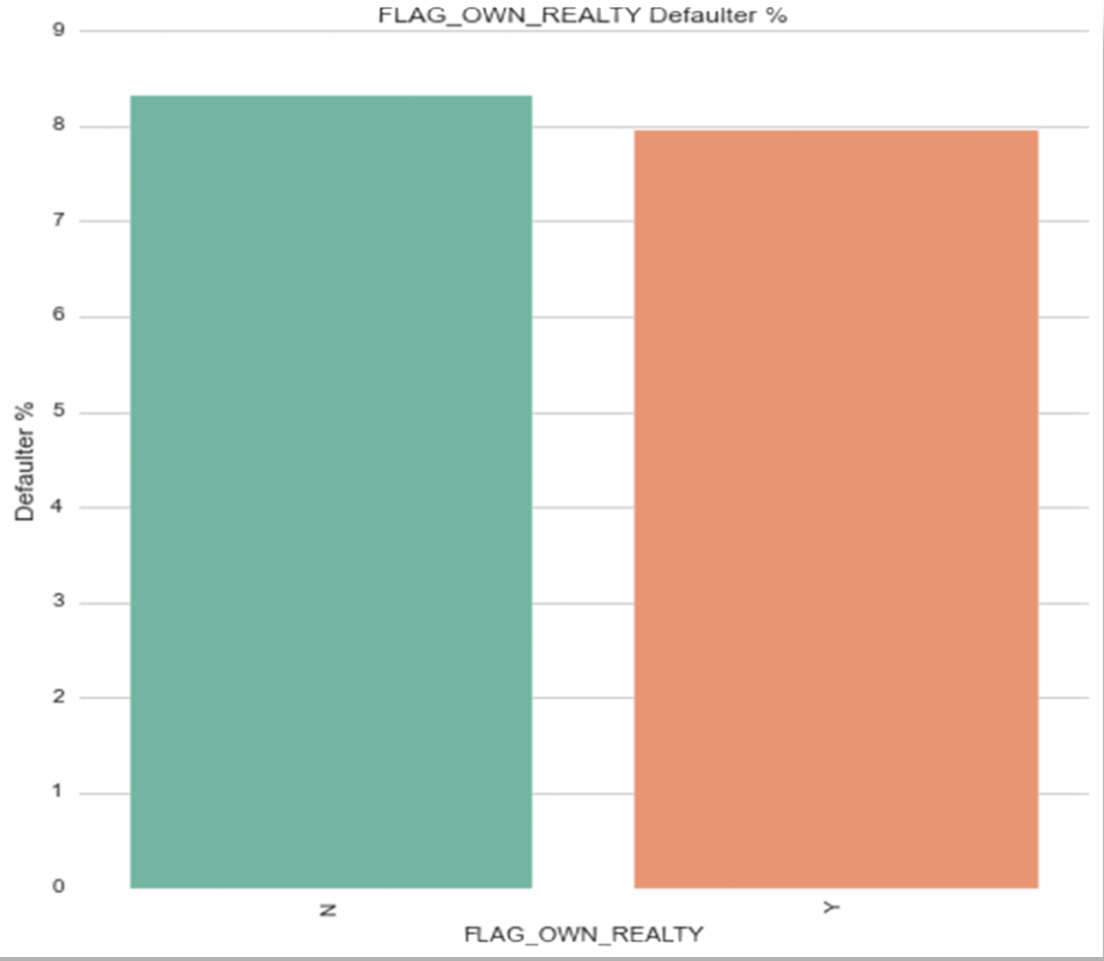




BIVARIATE ANALYSIS

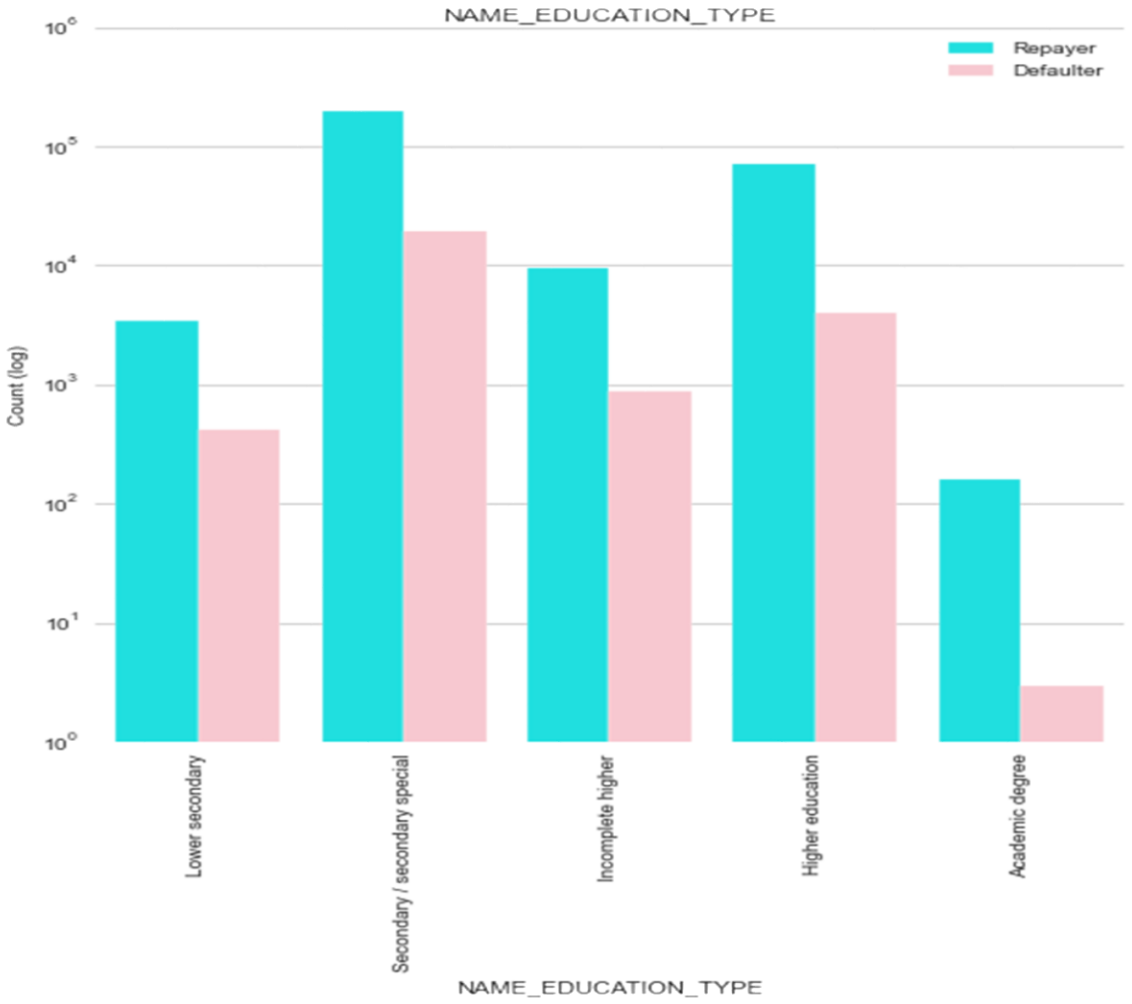
- People who do not own a car are more as compare to applicants who own the car but most likely both have equal chances of being a defaulter

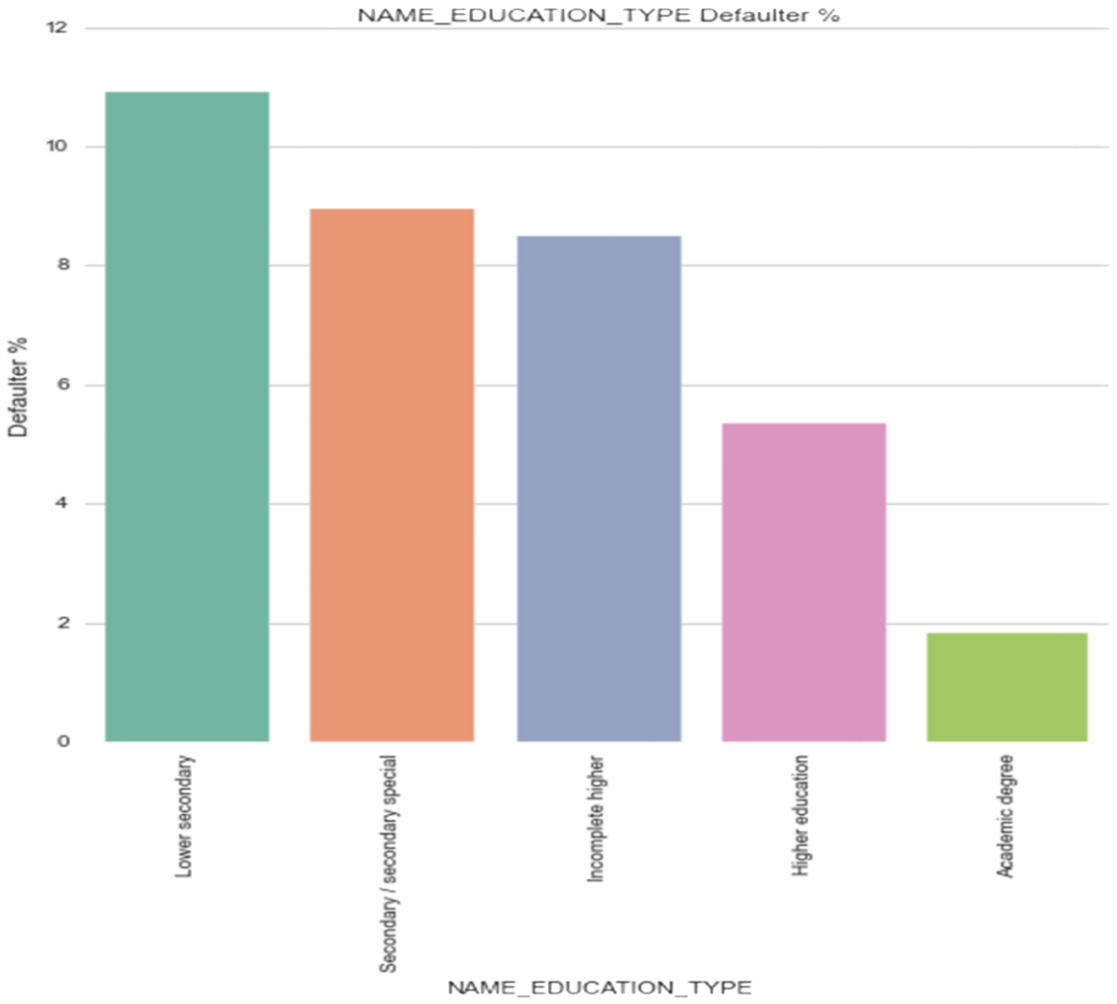




BIVARIATE ANALYSIS

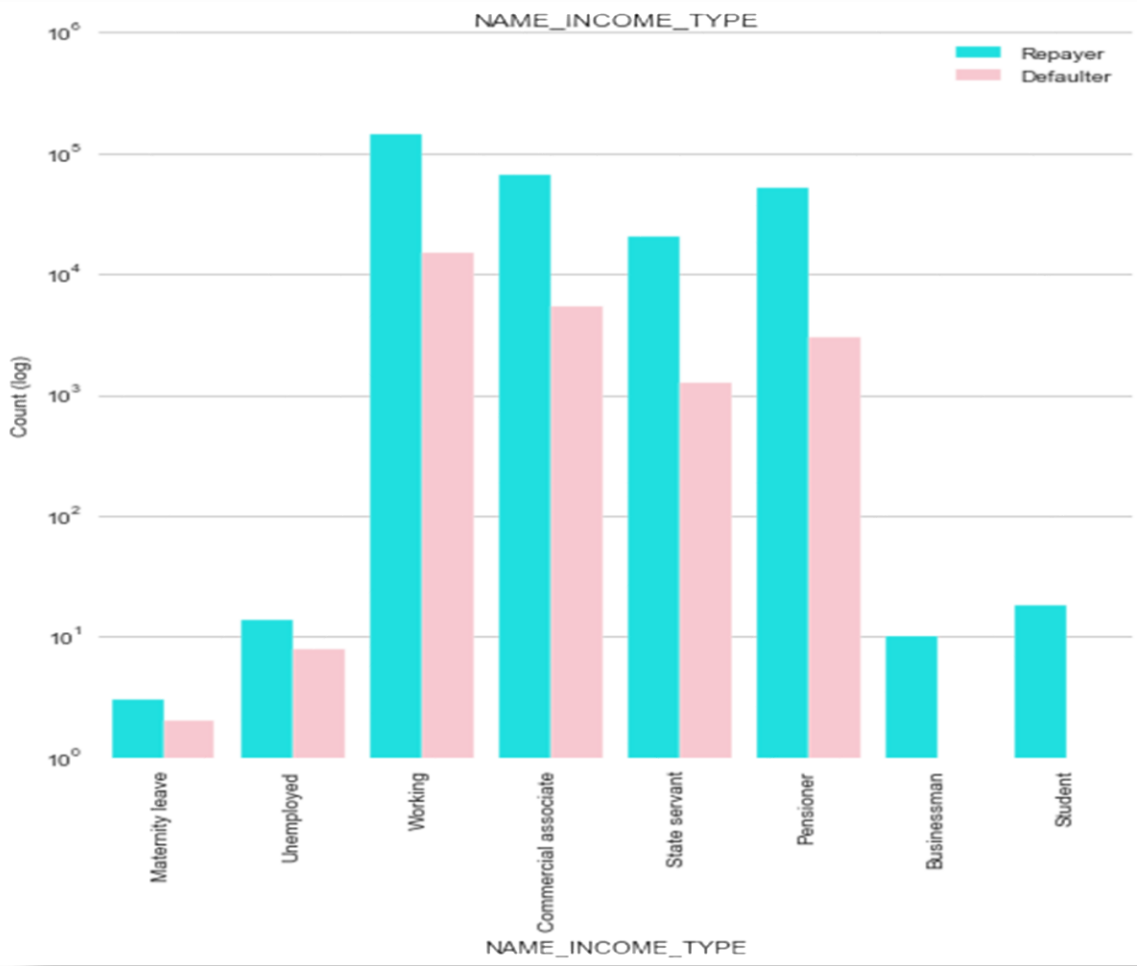
- Most of the applicants own a real estate but still the changes of defaulter is same among both categories

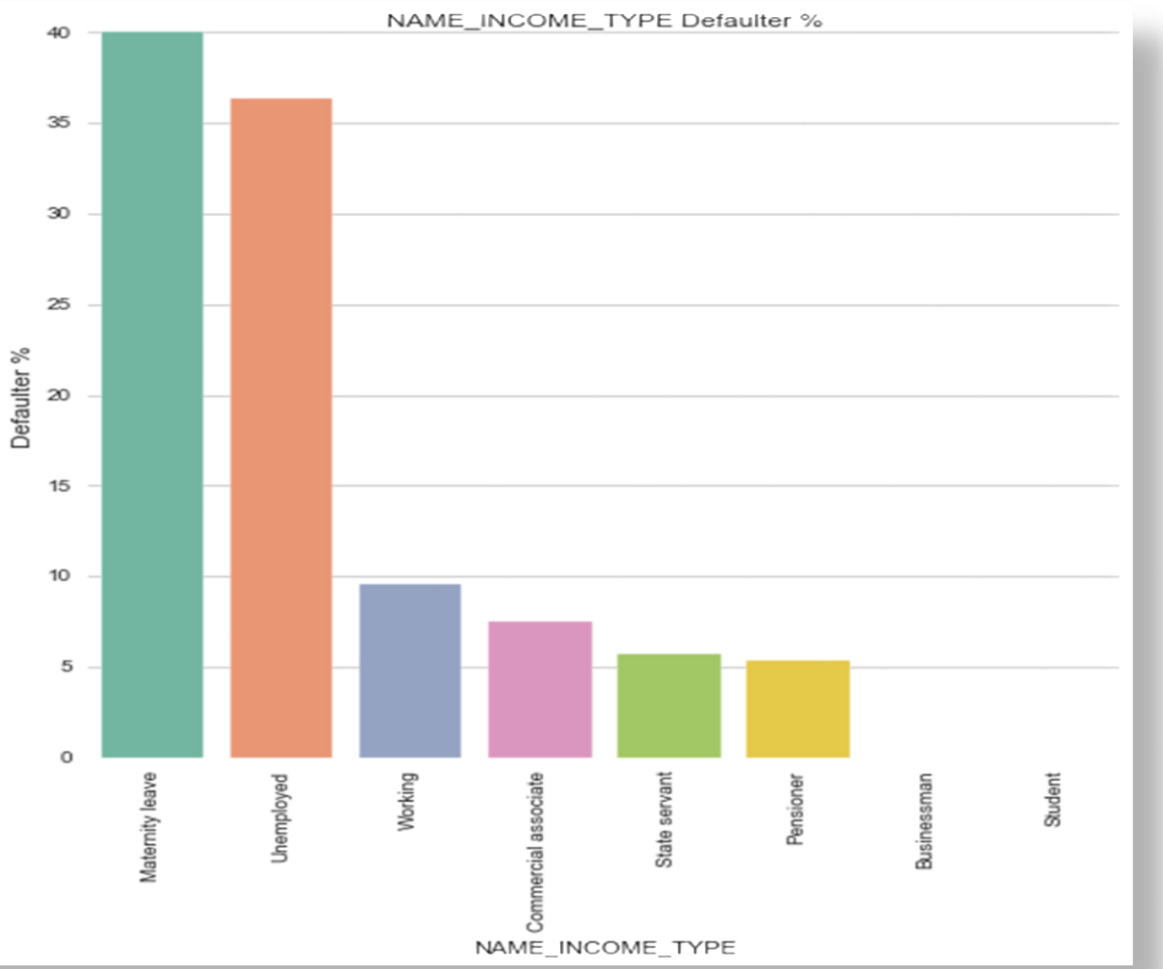




BIVARIATE ANALYSIS

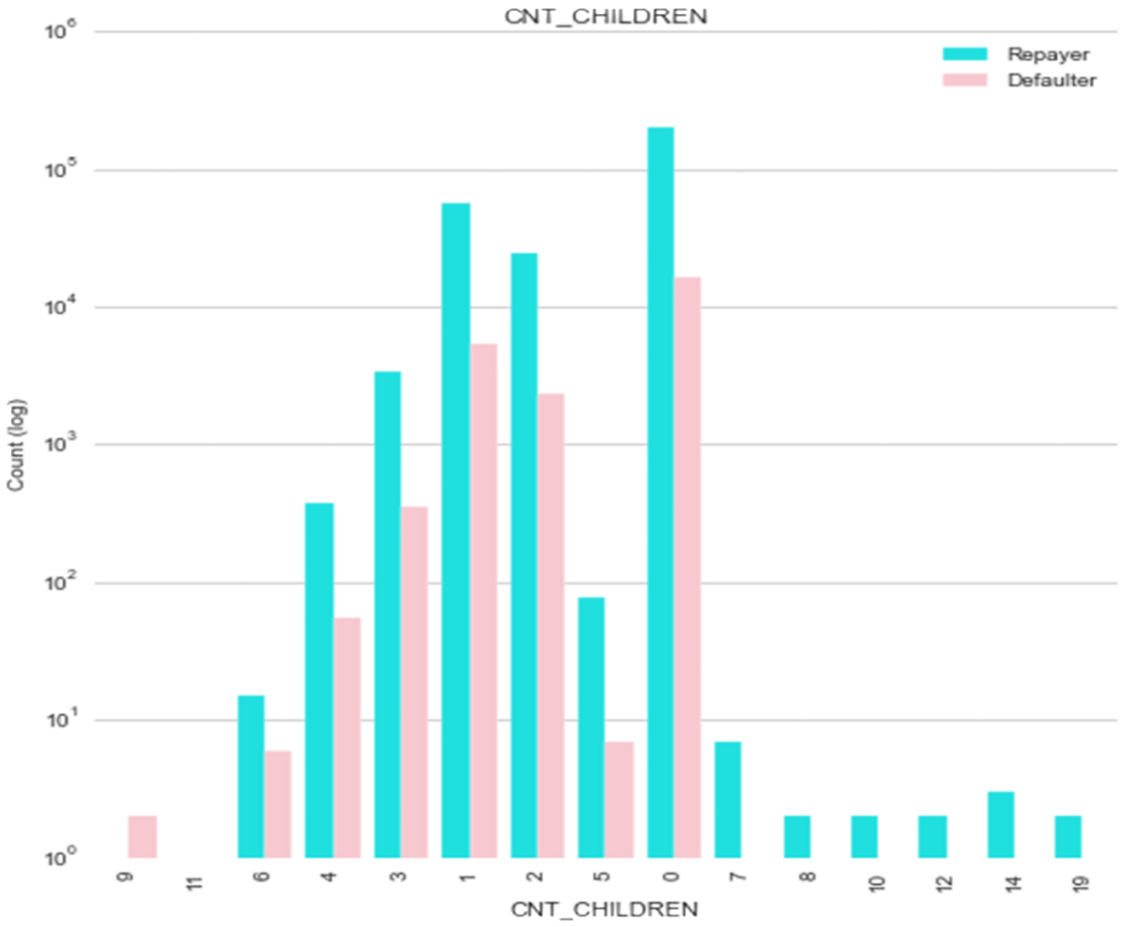
- Lower Secondary has the highest Defaulter percentage
- Most of the application are from Secondary/Secondary Special education.
- Rarest are from Academic degree with less % of defaulter as well

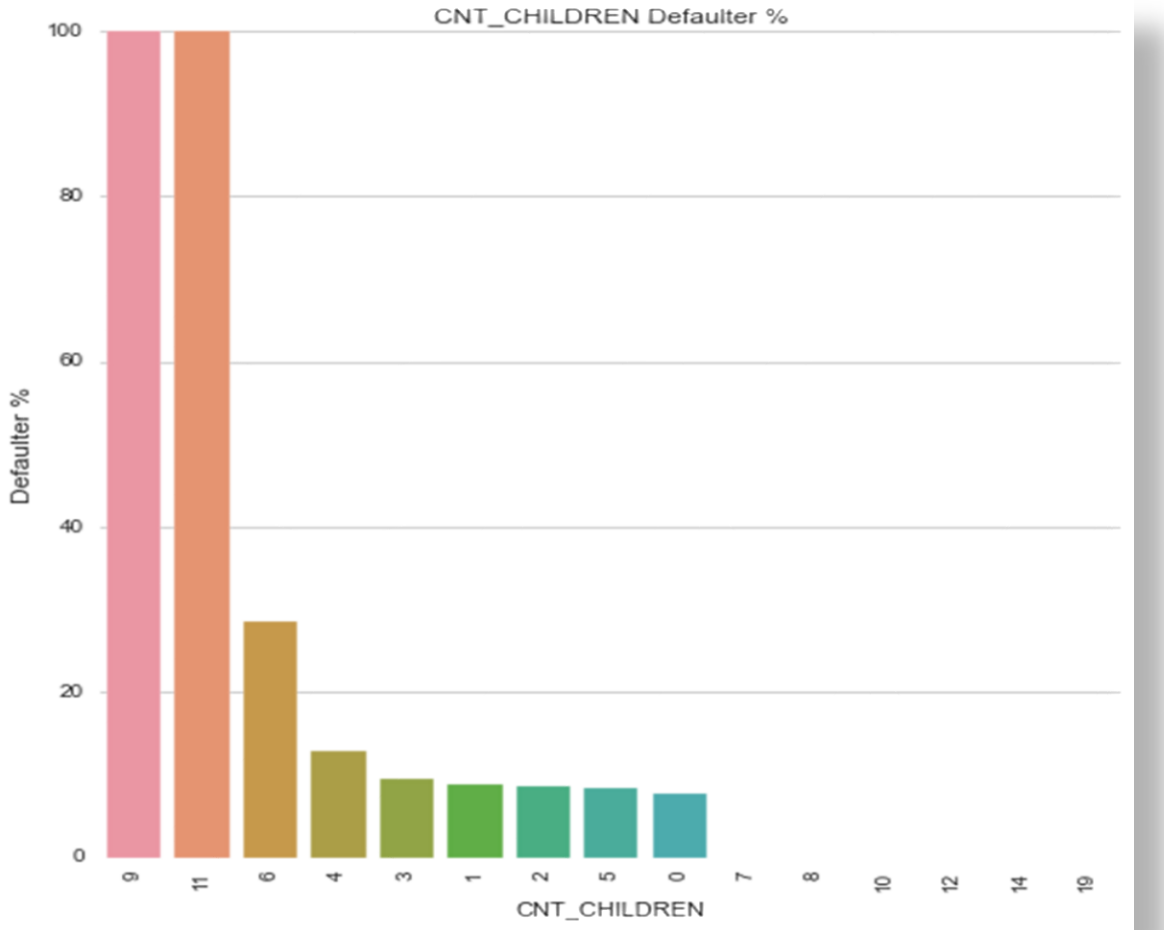




BIVARIATE ANALYSIS

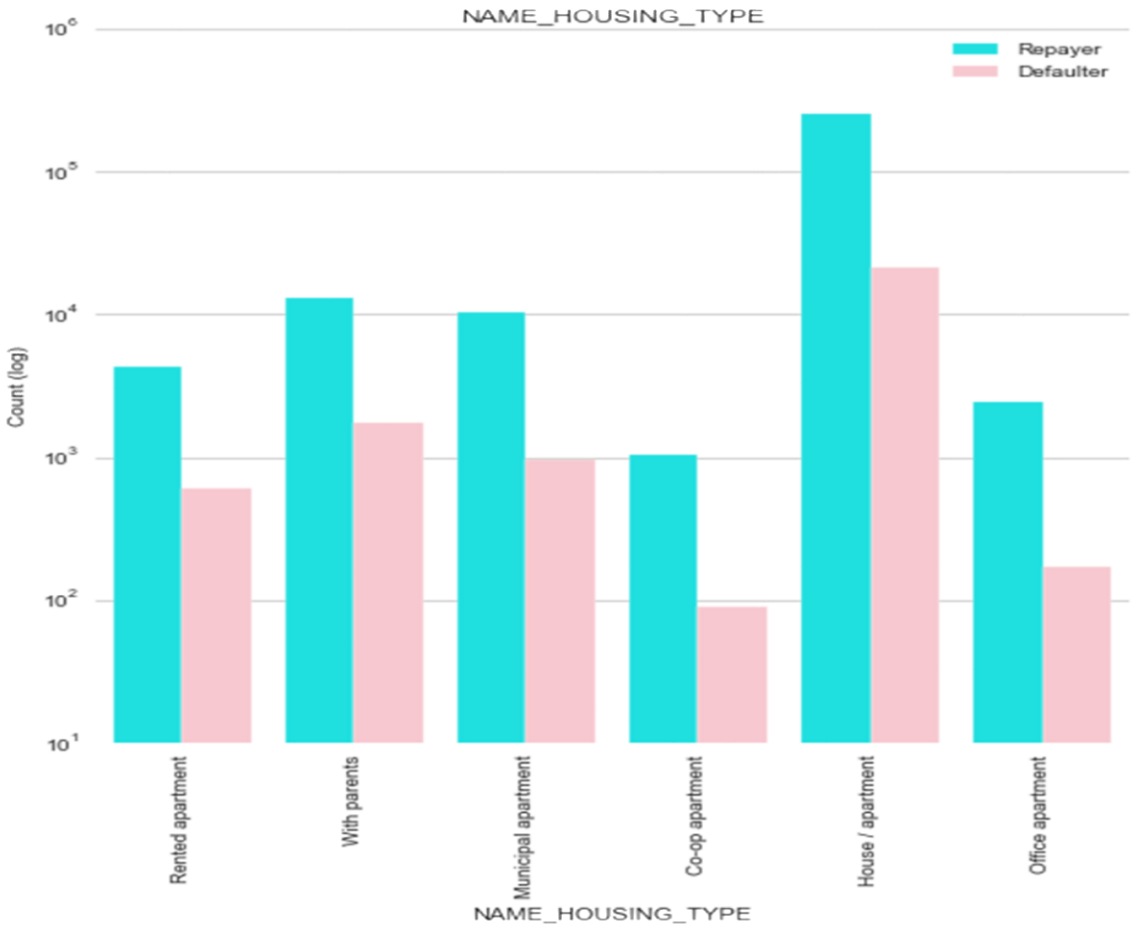
- Businessman and Student are the safest Income_Type to loan as always repay the loan back
- Maternity and Unemployed leave are the one with maximum percentage of defaulter with 30-40% of defaulter even though the count of the application with those type is not large.
- Most of the application are from Working followed by Commercial Associate , State Servant and pensioner.

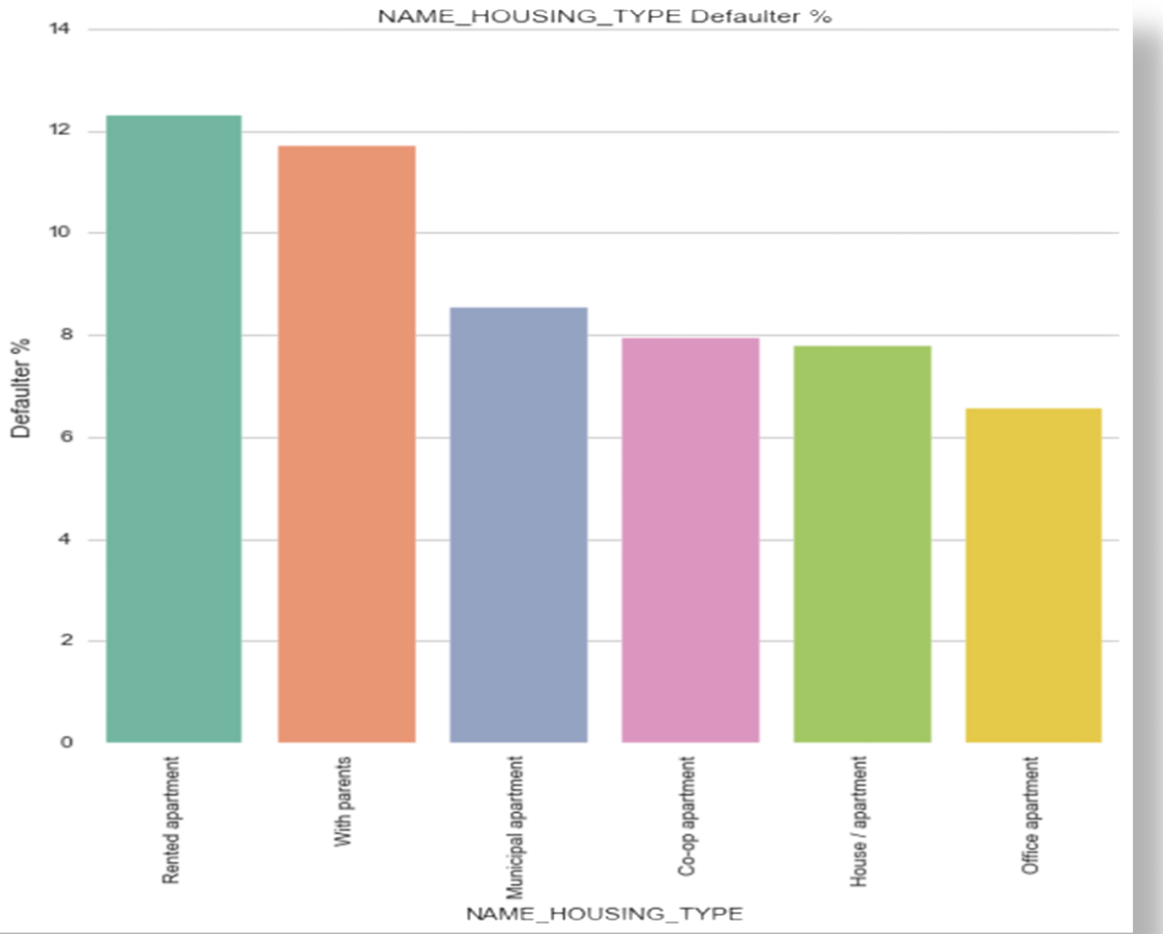




BIVARIATE ANALYSIS

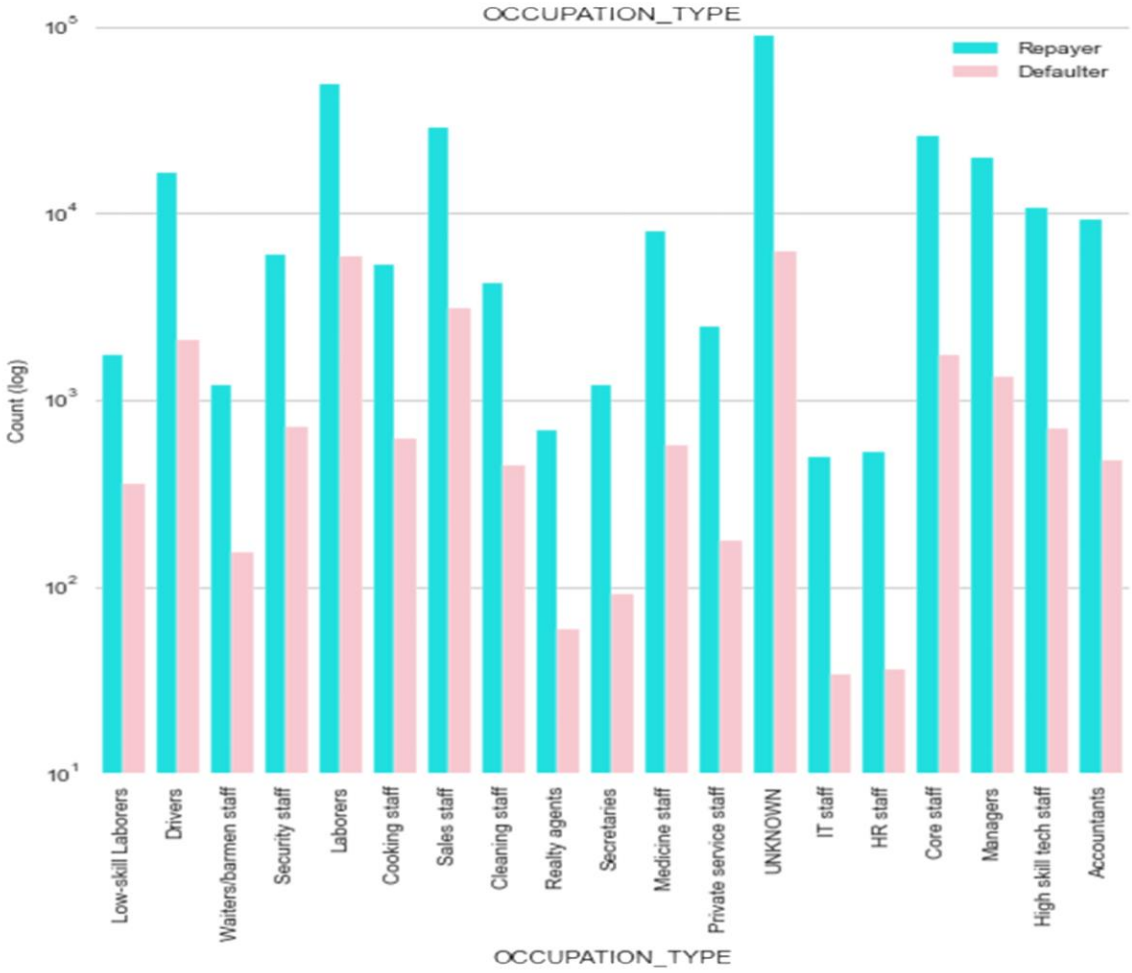
- Most applicants don't have children or just have 1 child , and also have higher chances of repaying the loan
- We also observe that applicants having children between 9-12 has high % of being a defaulter with less applications
- Also applicants with more than 10 children have no records as a defaulter.

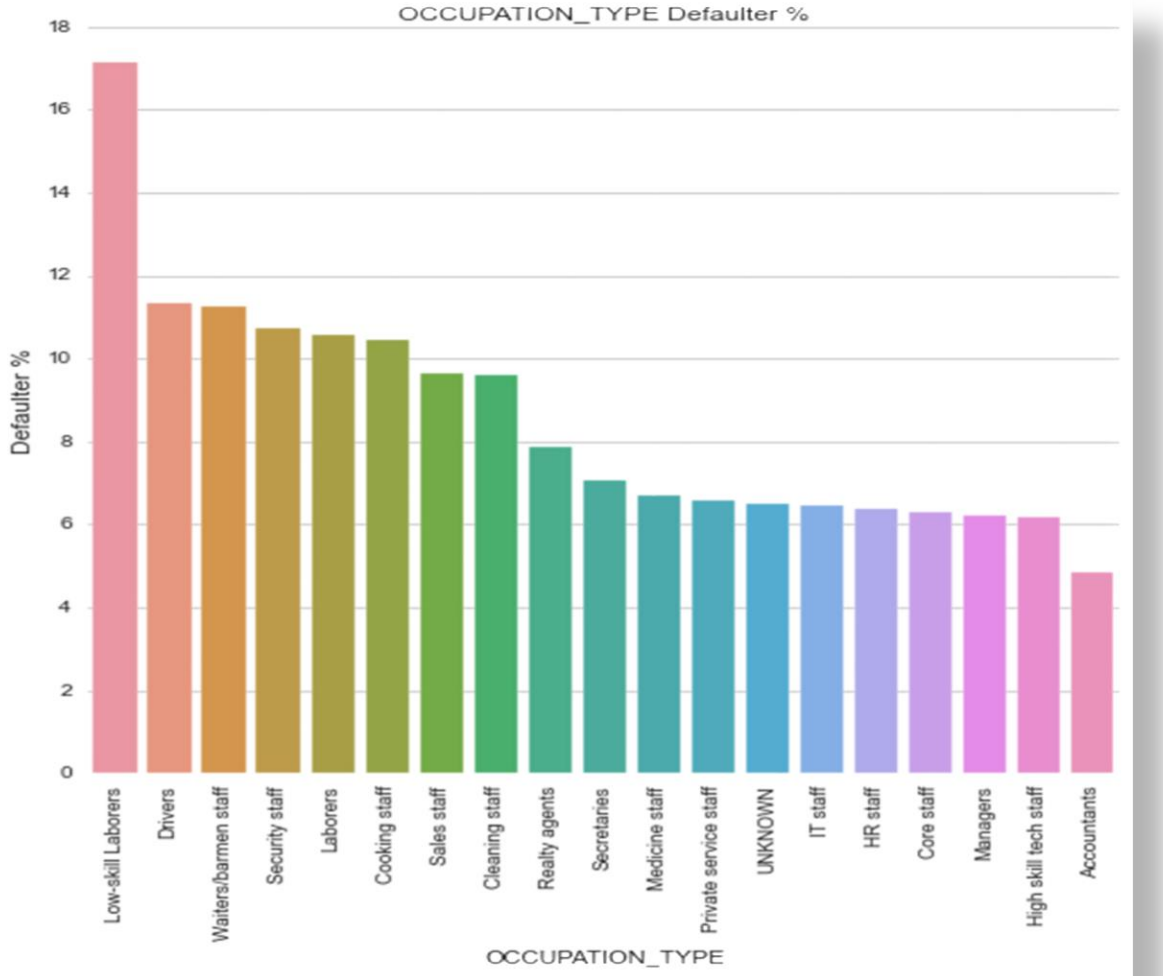




BIVARIATE ANALYSIS

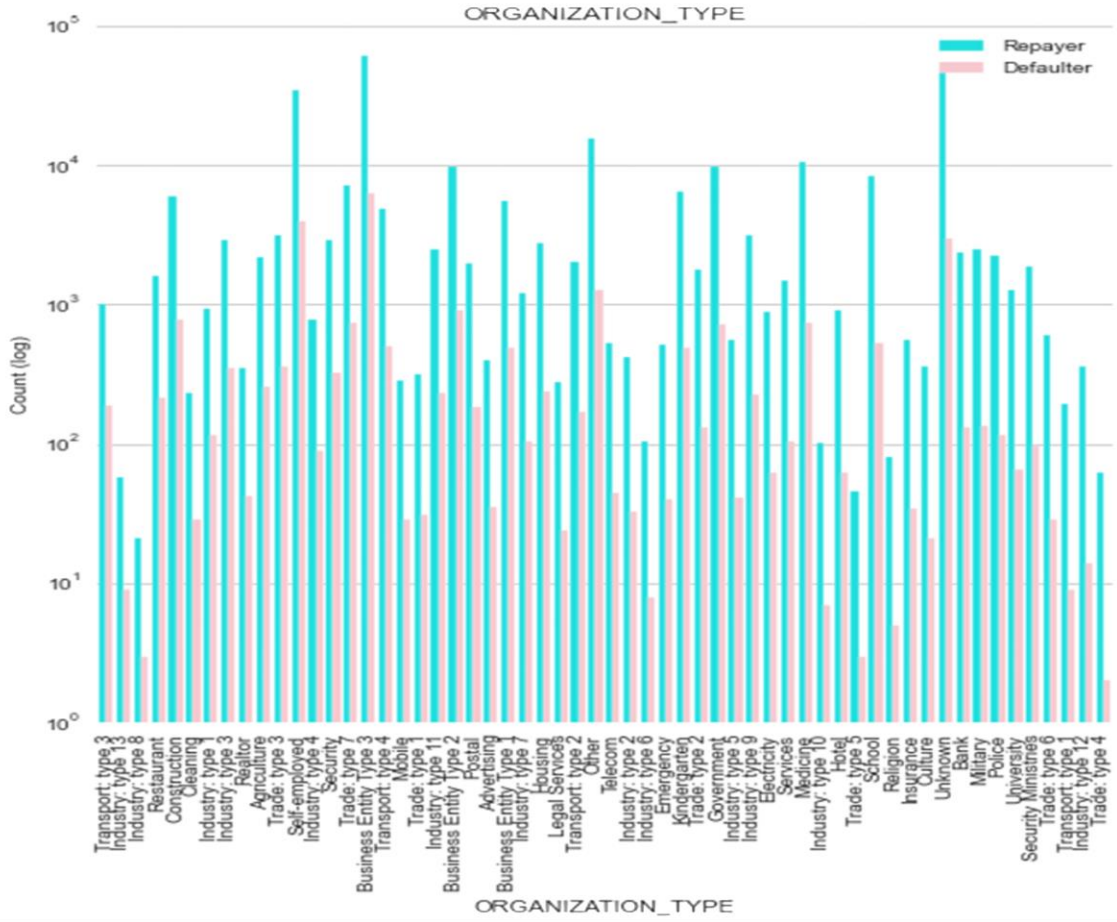
- Most of the applicants have House/Apartment and comparatively less defaulter %
- Most defaulter lies under the category of Rented Apartment followed by Living with Parents

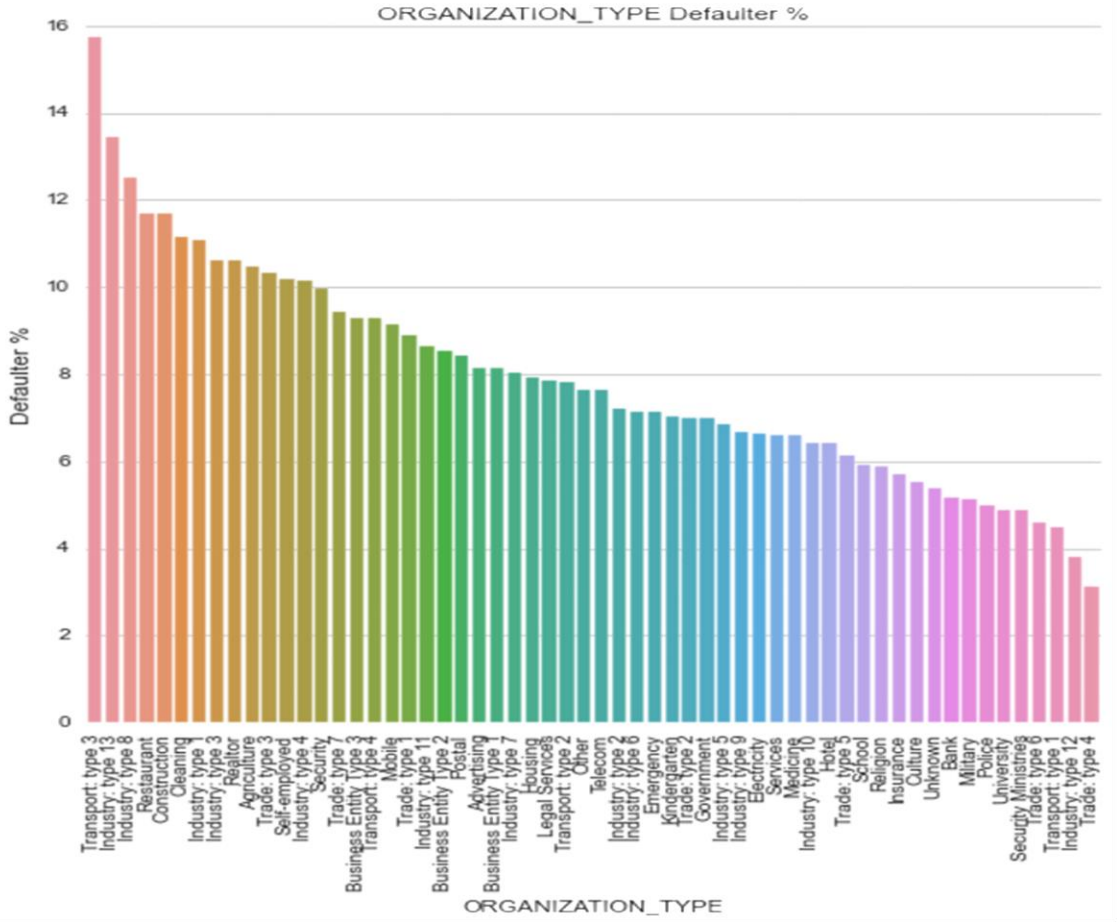




BIVARIATE ANALYSIS

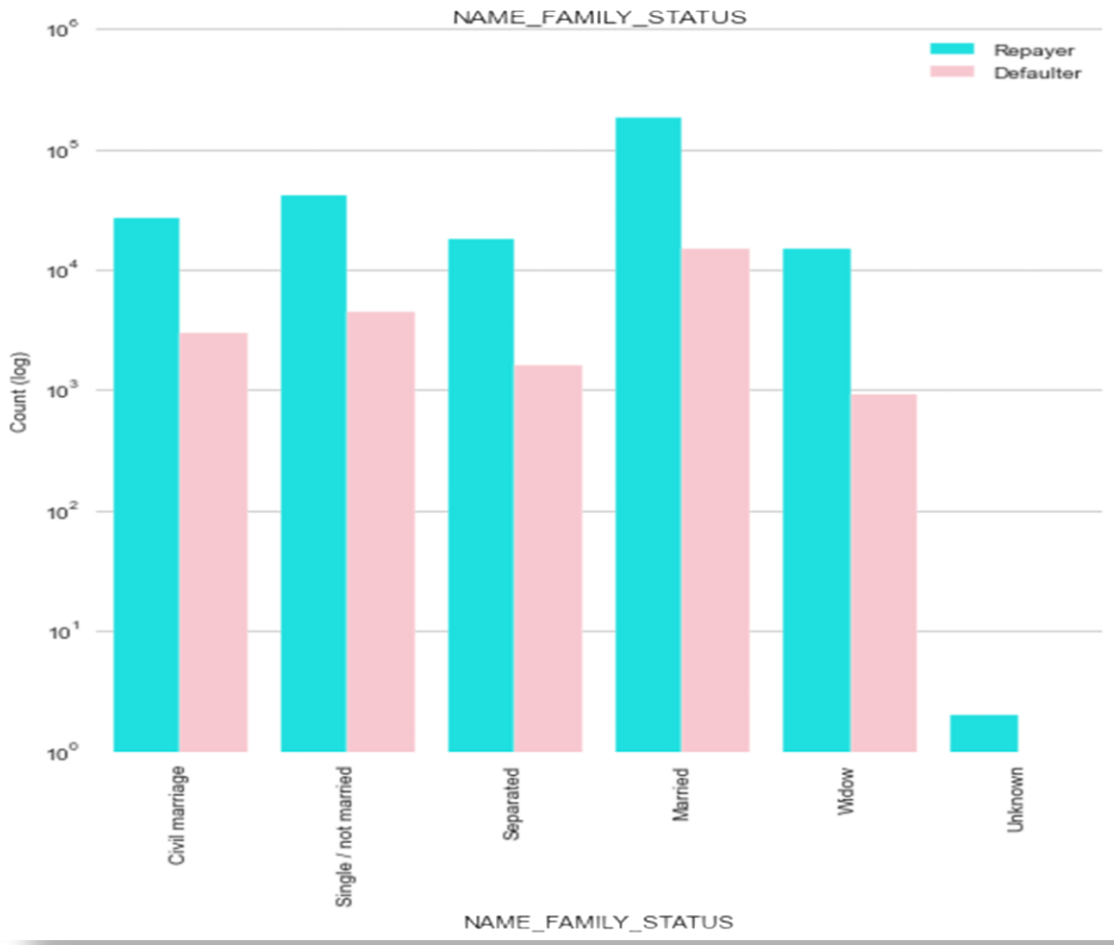
- Most of the applicants have not revealed their occupation but have less % of defaulter
- The occupation with most Default % is Low-skill Laborers followed by Drivers and Waiters/barmen staff.

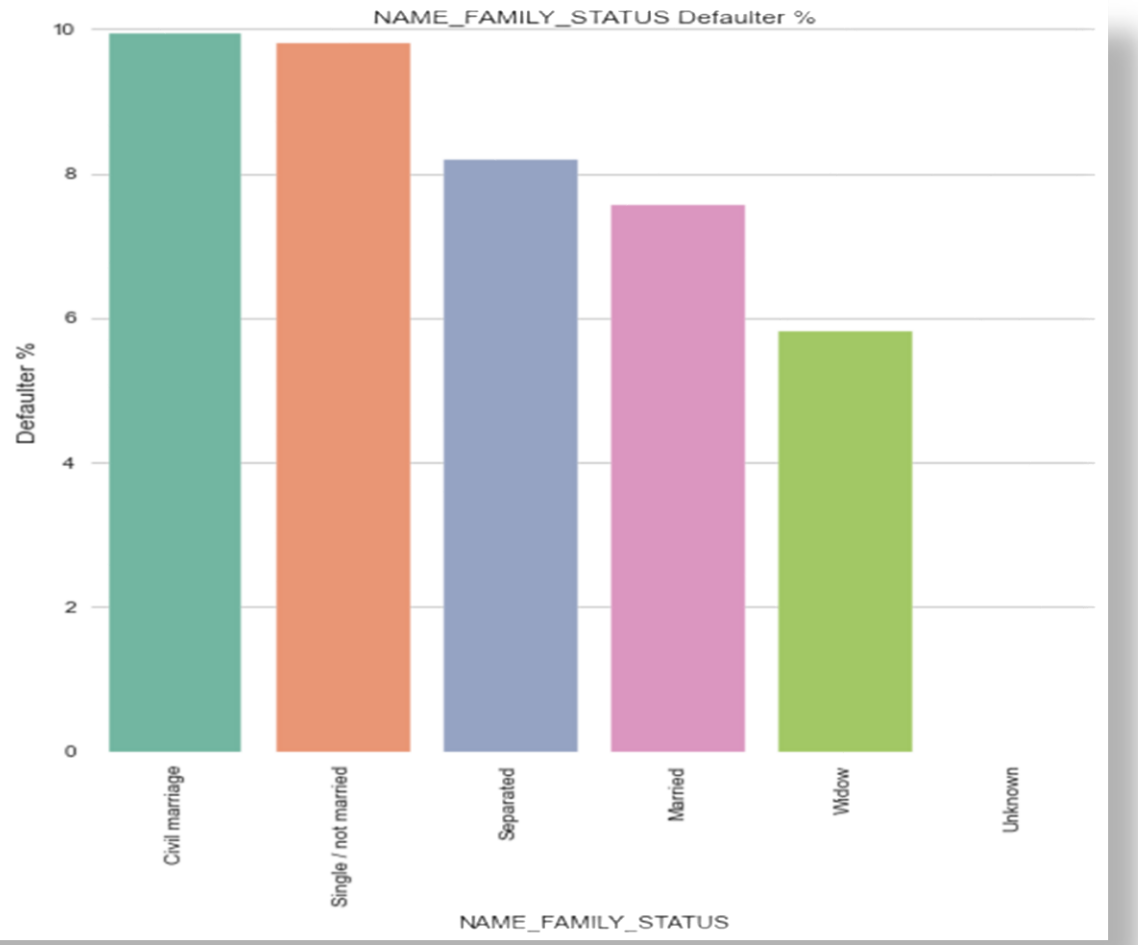




BIVARIATE ANALYSIS

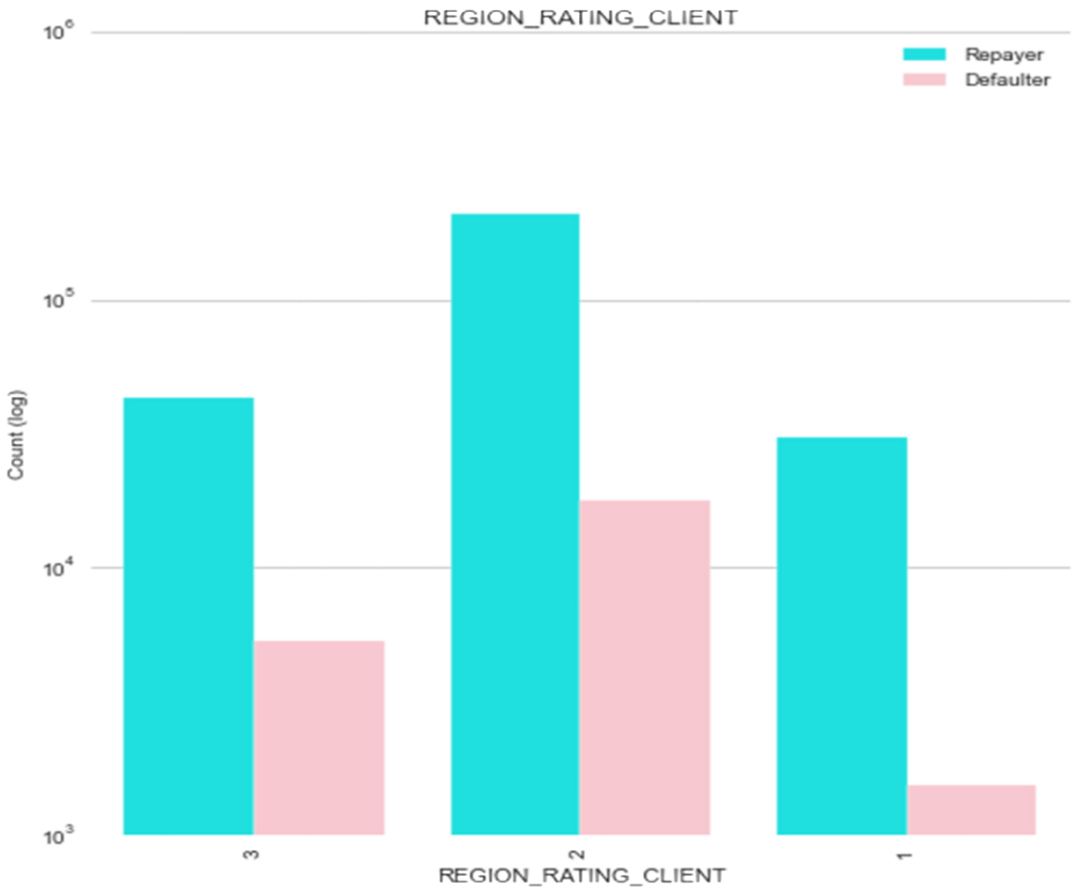
- Most of the applicants have the category unknown under organization type
- The category with most default are Transport Type 3, Industry Type 13, 8, Restaurants, Construction
- The Category with less defaulter are Transport Type 1, Industry Type 12, Trade Type 4

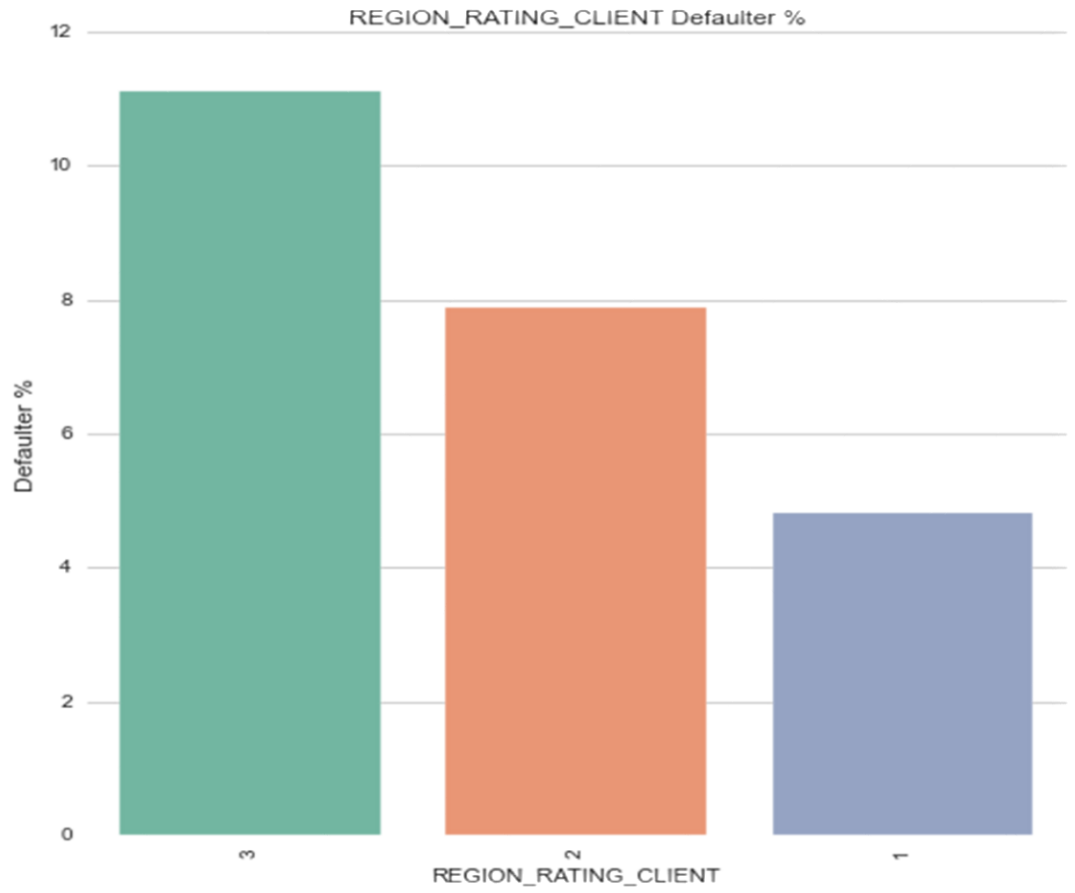




BIVARIATE ANALYSIS

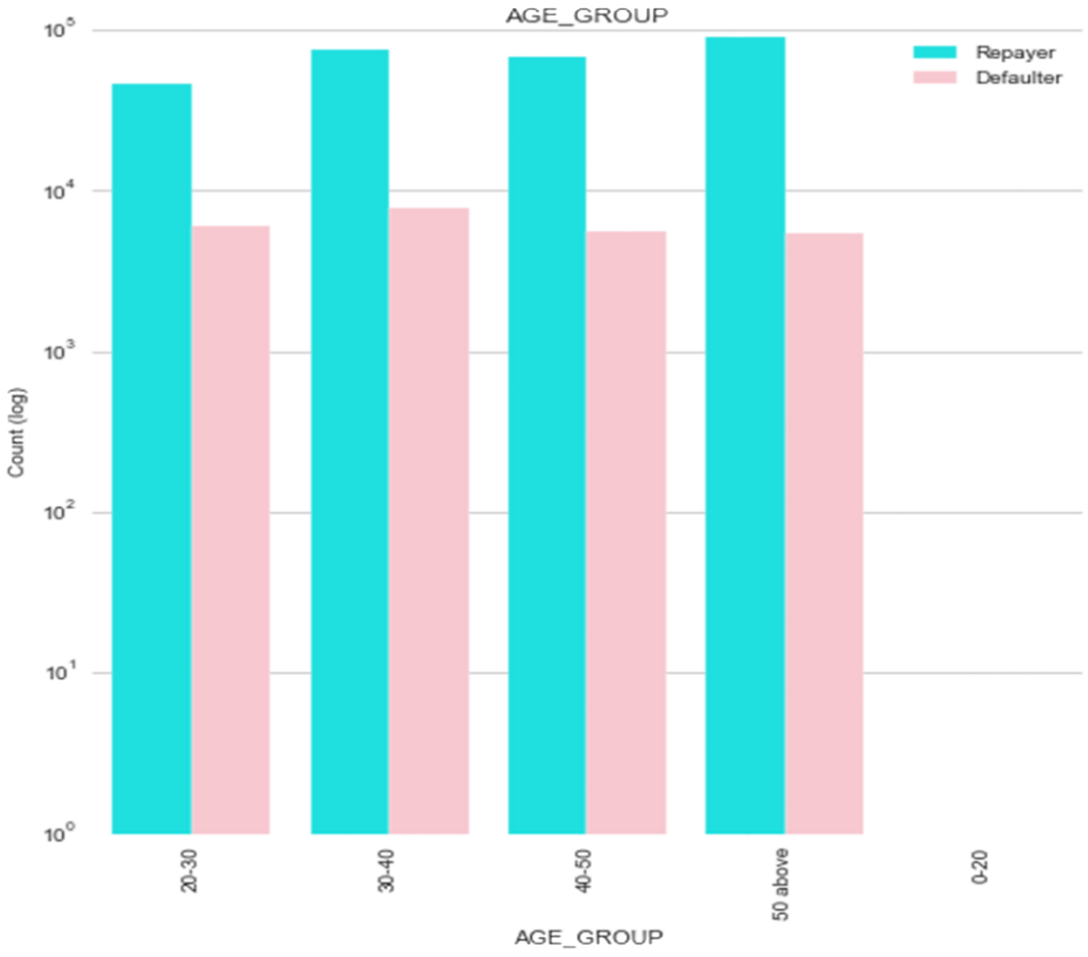
- Most of the applicants who applied for loan are married followed by single/not married
- We observed that people with civil marriage or single are tends to be default
- Comparatively people who are married or window tends to repay the loan

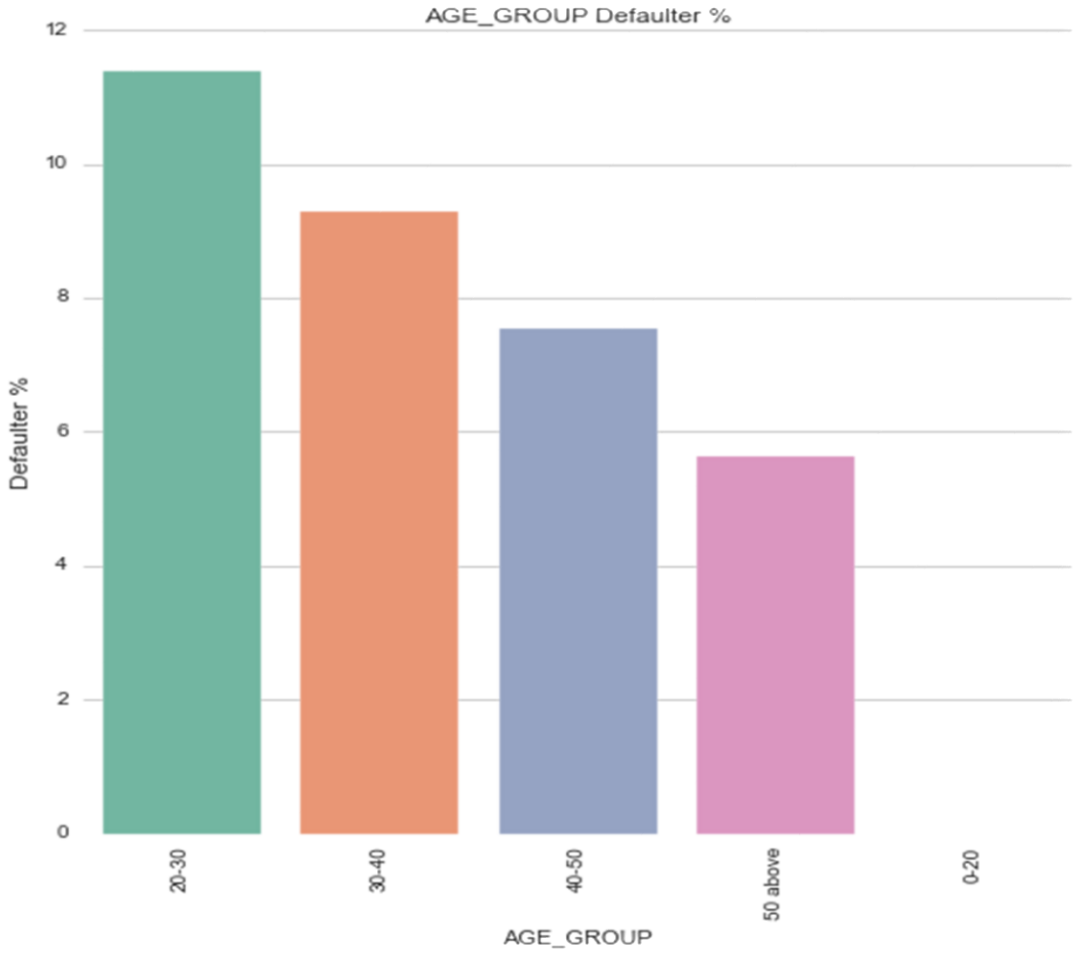




BIVARIATE ANALYSIS

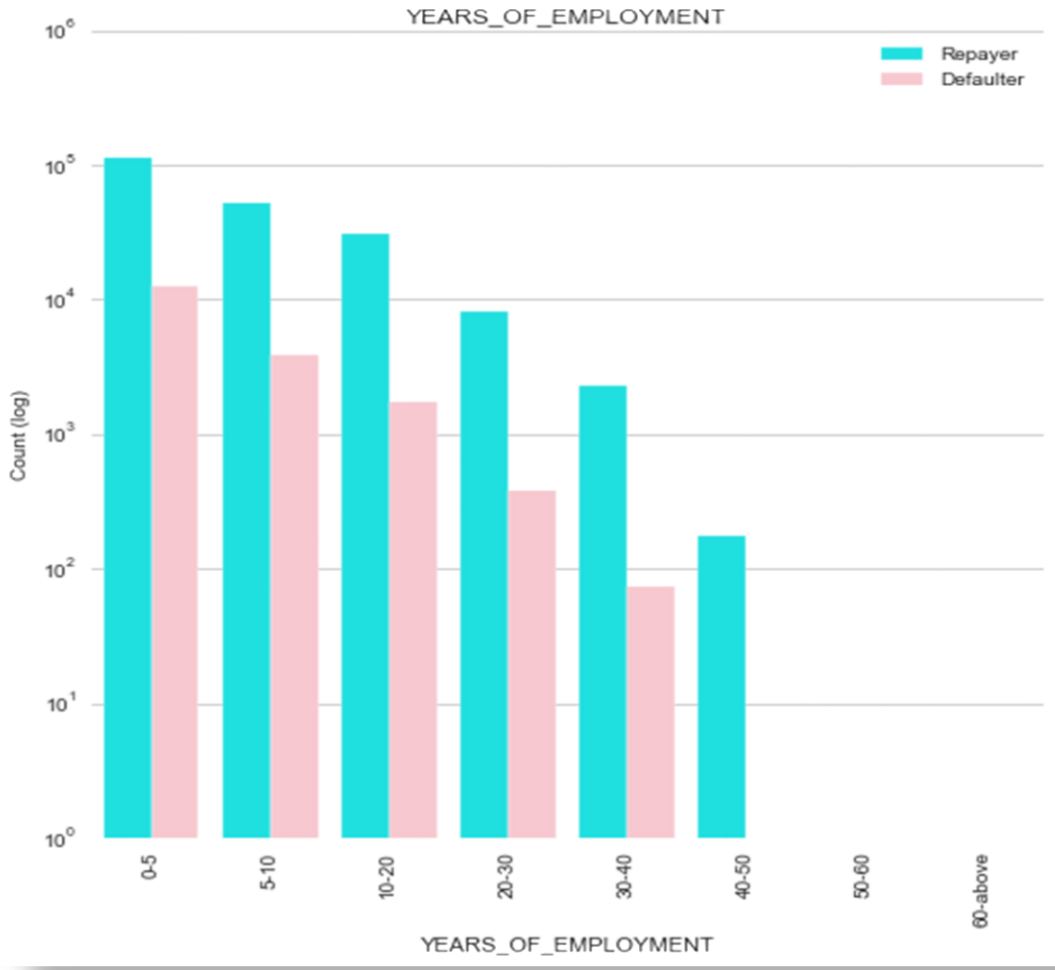
- Applicants with 3 as Region Rating are more likely to be defaulter
- And among rating 1 is less prone to being a defaulter

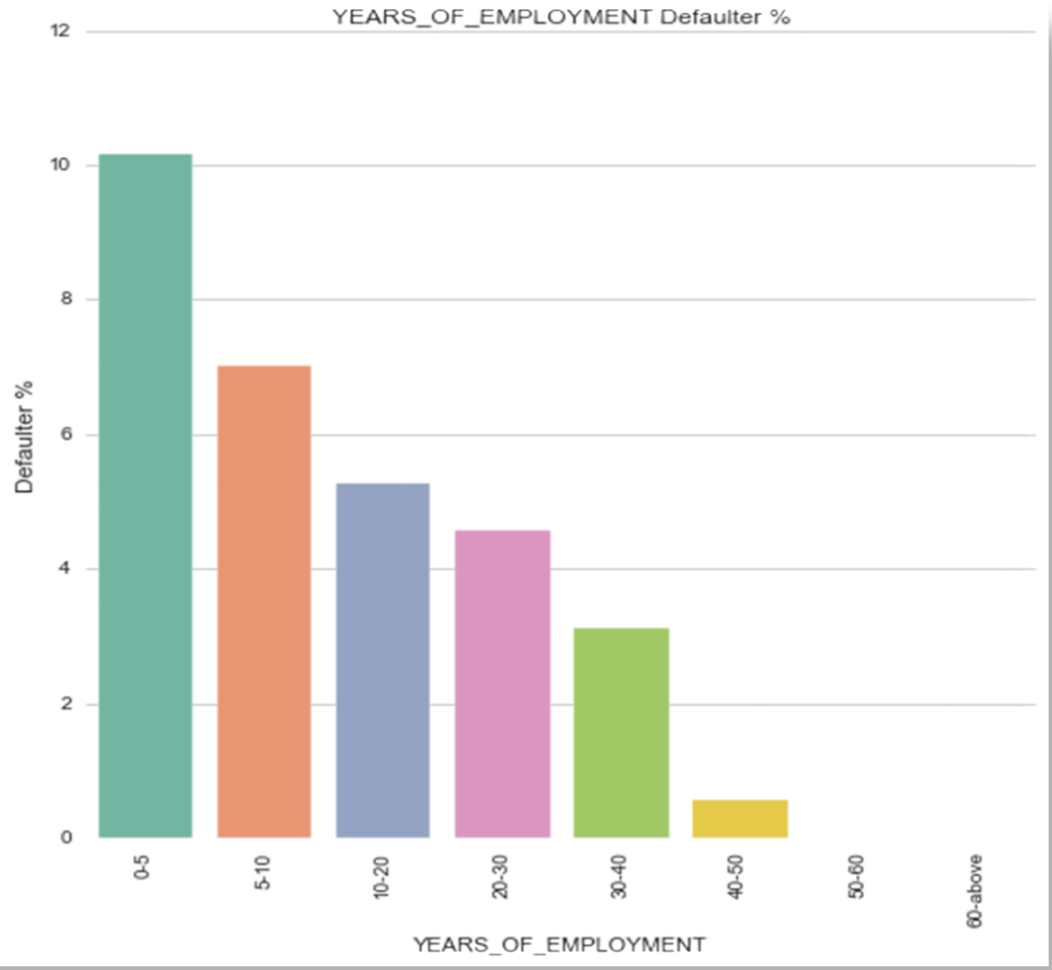




BIVARIATE ANALYSIS

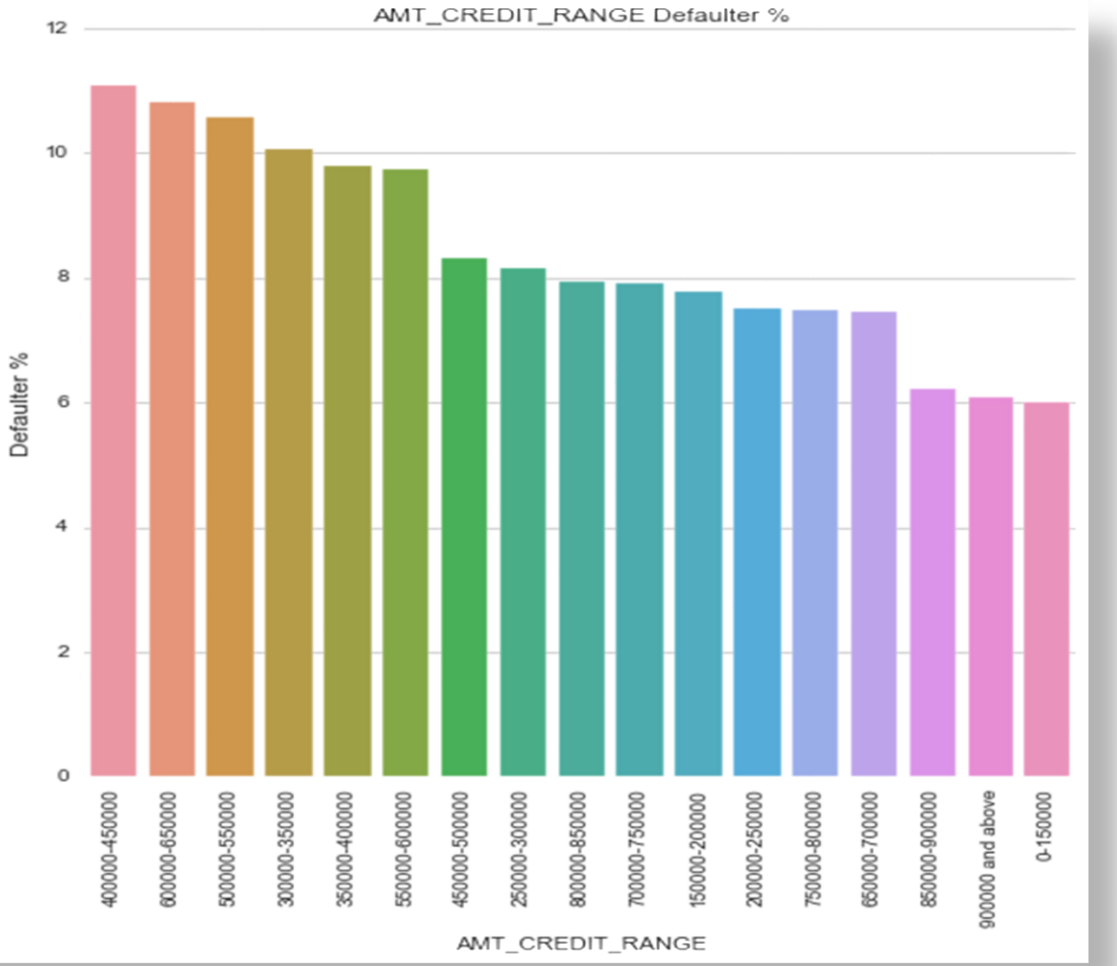
- People among the age group from 20-40 have high % of being a defaulter an the safest are above 50 age group





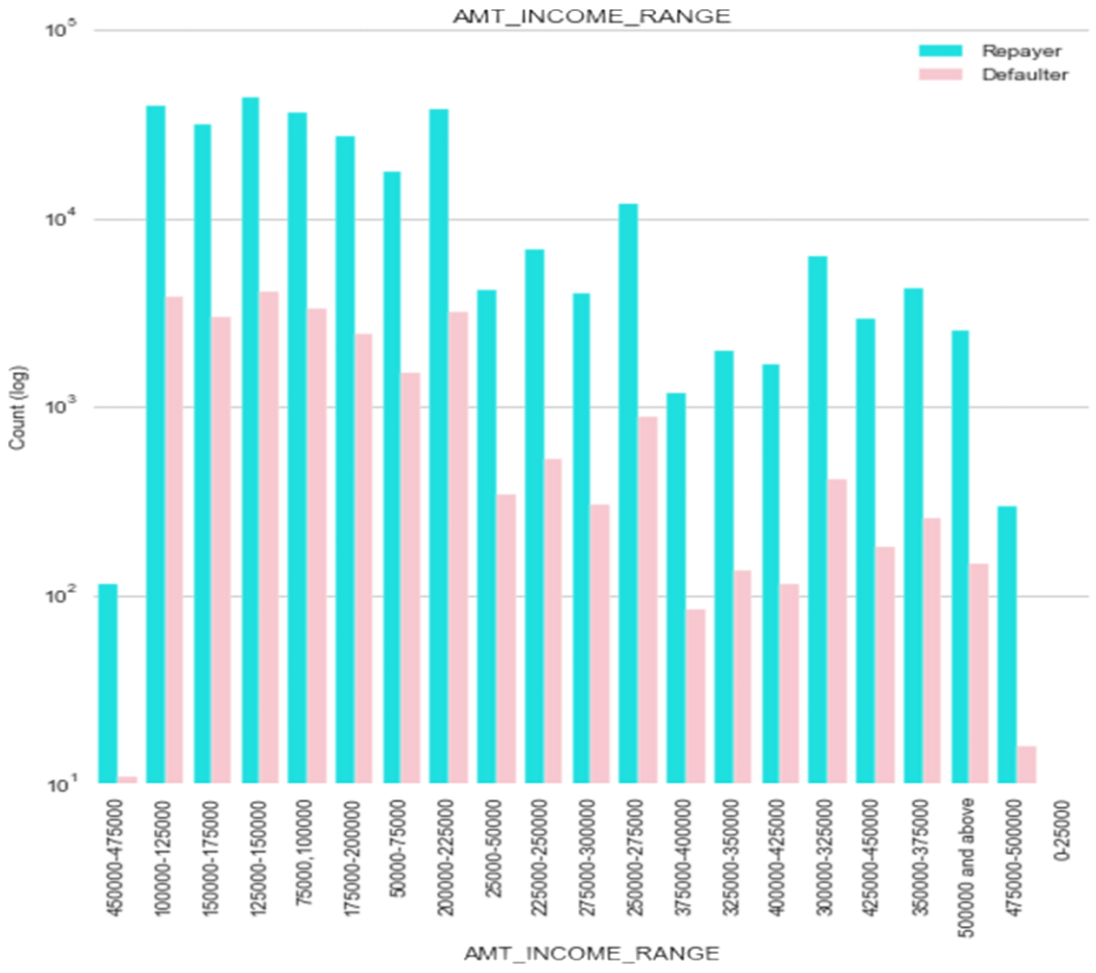
BIVARIATE ANALYSIS

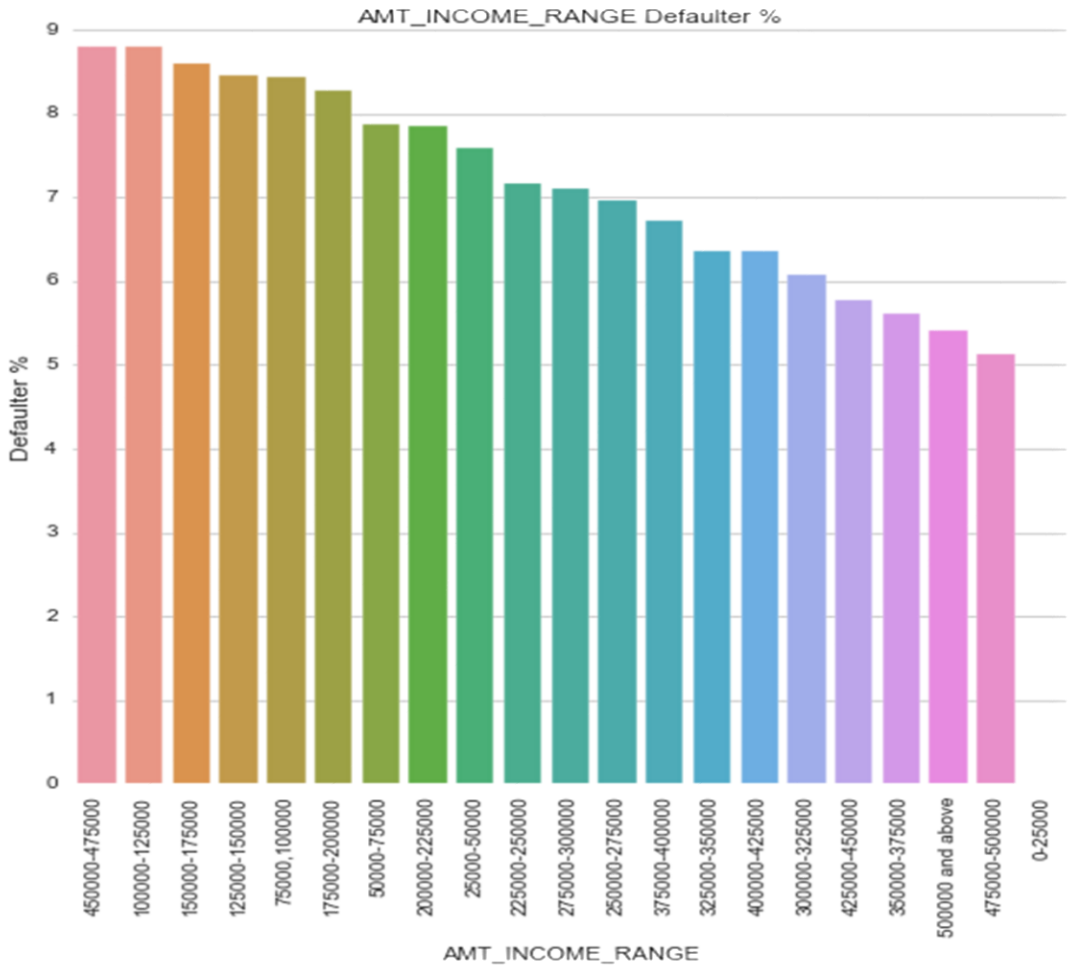
- Most of the applications are from the year of experience 0-5 group but this has highest percentage of being a defaulter.
- People with more years of experience have less percentage of being a defaulter
- Safest is the group of 40-50



BIVARIATE ANALYSIS

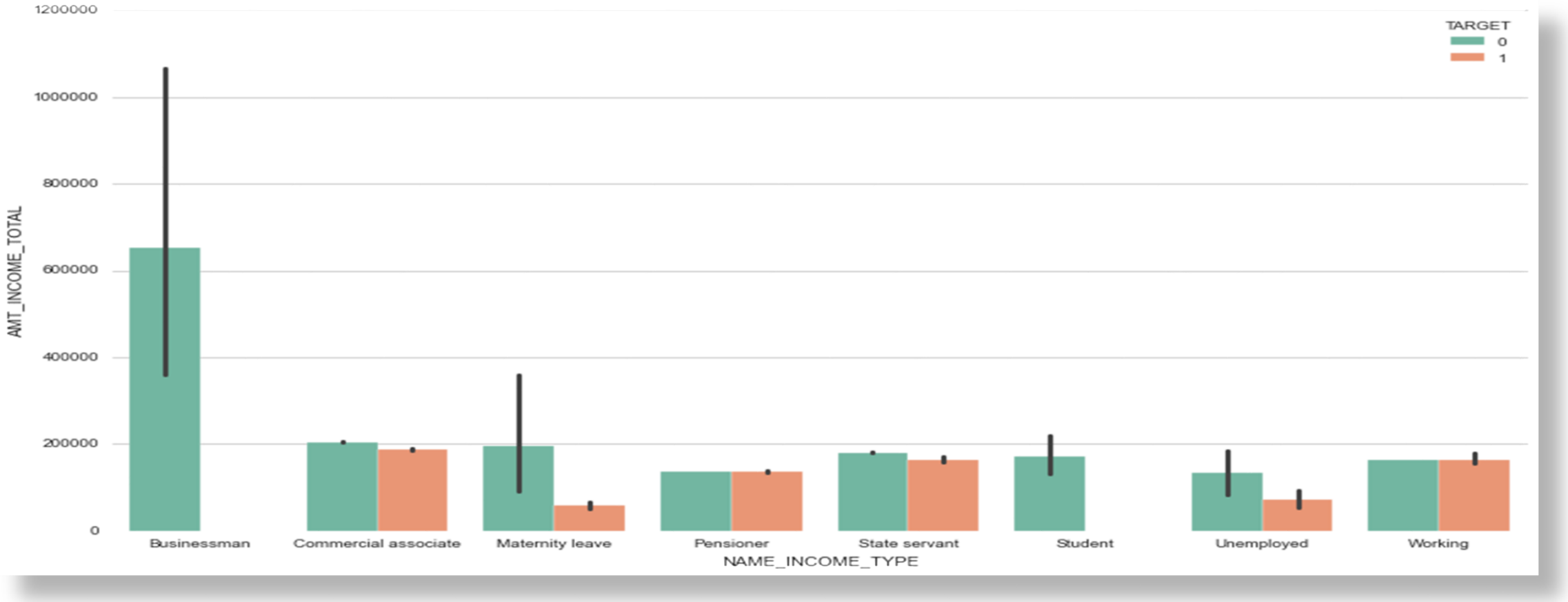
- We could see applicants who applied for loan amount more than 900 K are the most as well as they are more likely to repay the loan amount as well.
- The highest percentage of defaulter lies between the loan amount 400k - 600K
- 90% of the loan applied is less than 900K





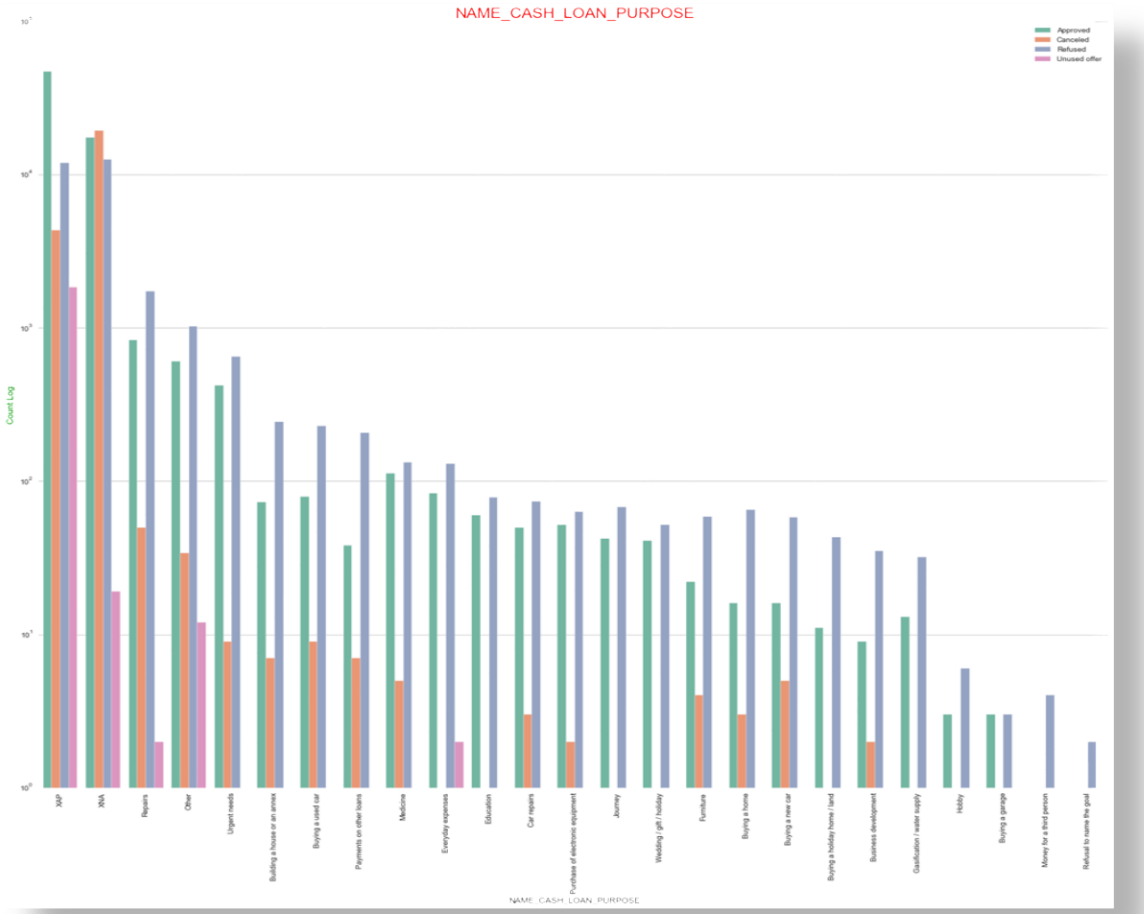
BIVARIATE ANALYSIS

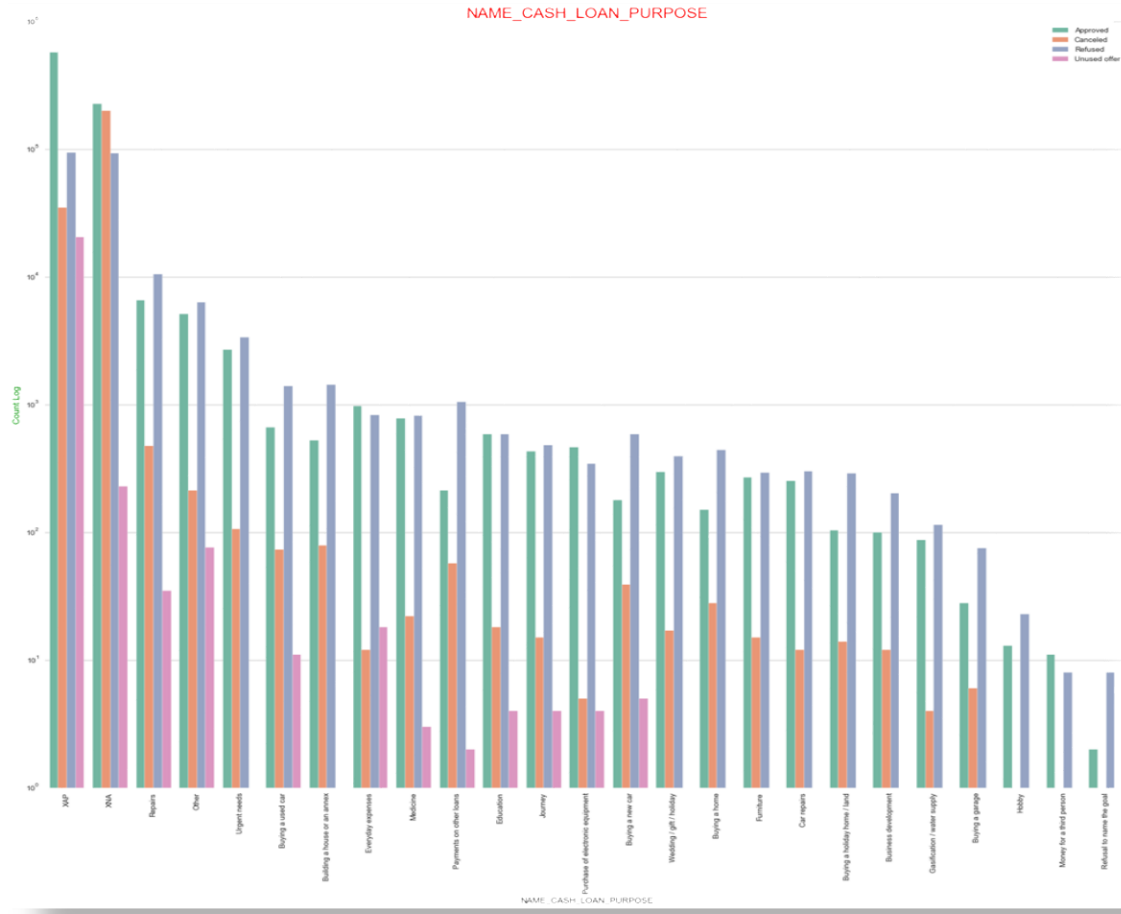
- Maximum applicants who are likely to a defaulter have income ranges from 100k -225K
- Most of the applicants have income less than 300K
- People with more income are tends to be Repayer



MULTI-VARIATE ANALYSIS

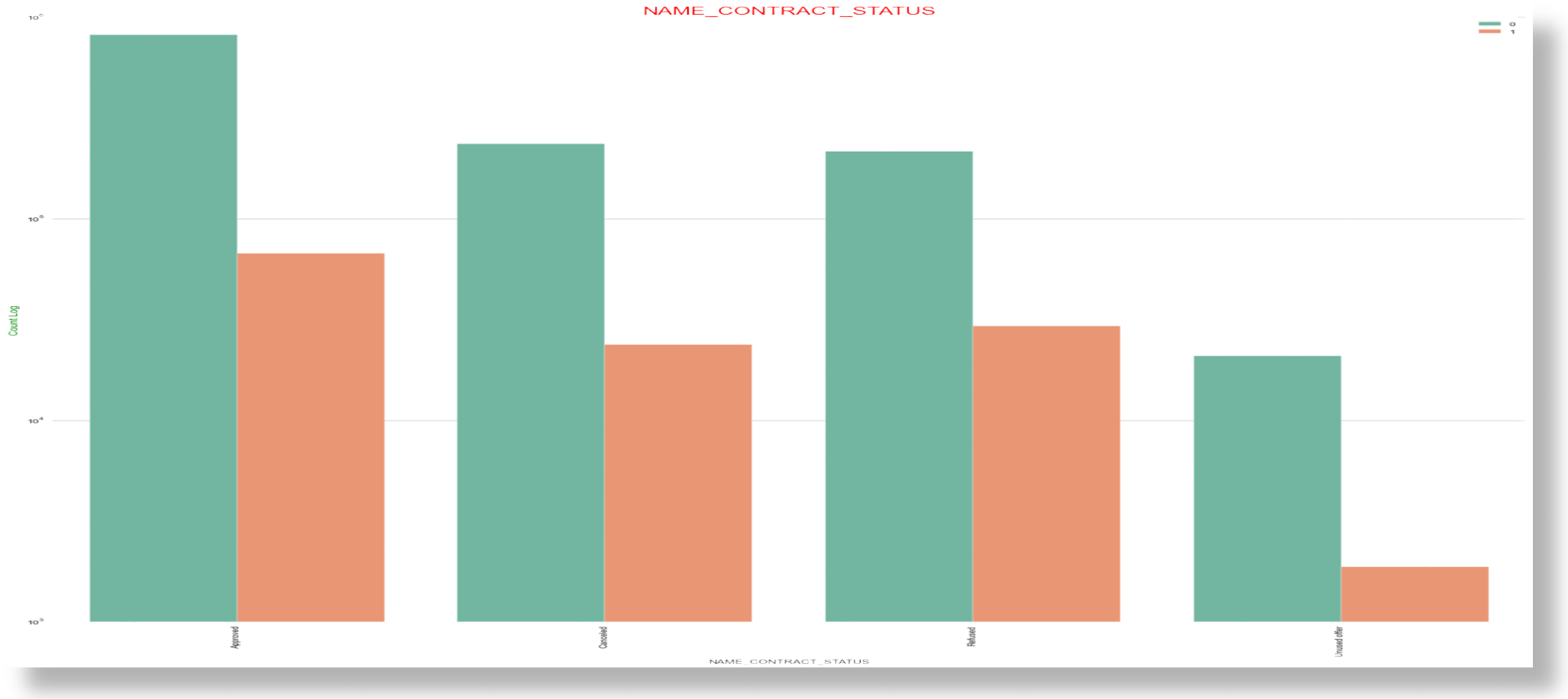
- Income of Businessman is highest and negligible defaulter records, we could see the same graph for Student





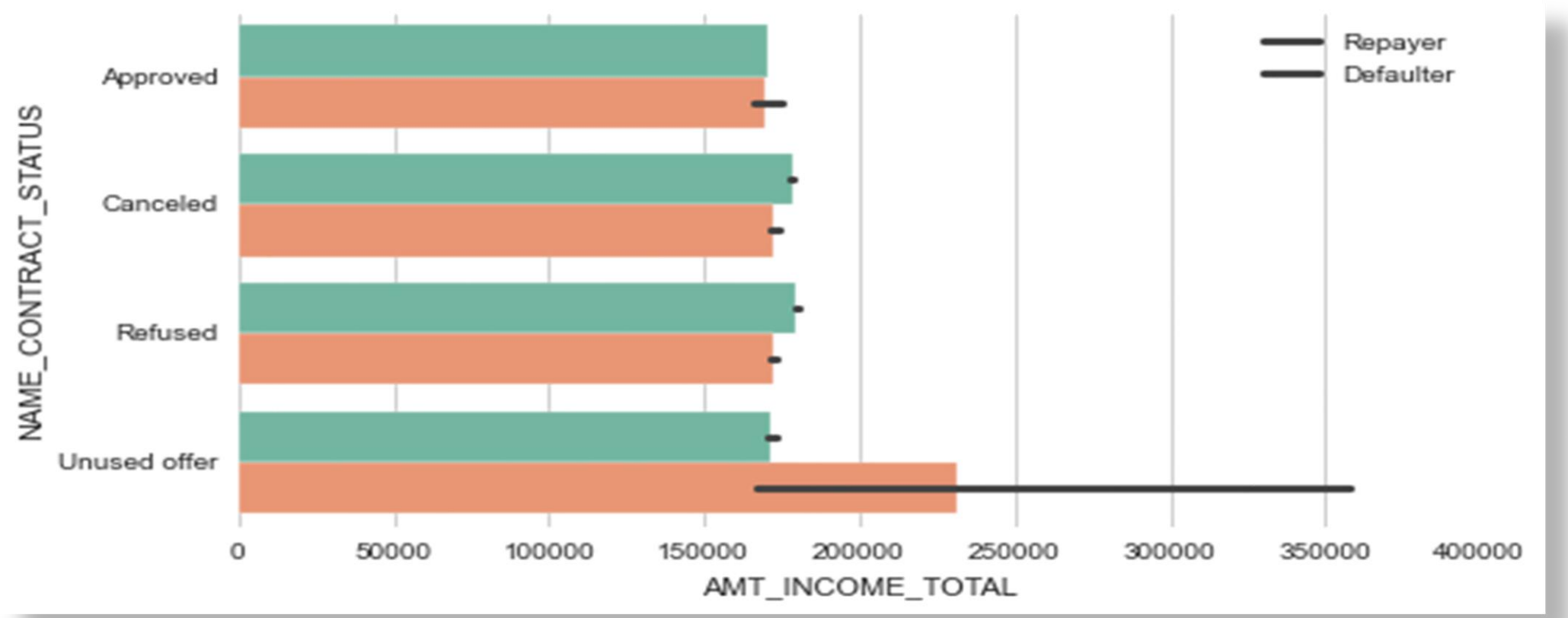
MULTI-VARIATE ANALYSIS

- For most of the applicants the loan purpose is unknown (XAP, XNA)
- Among the unknown purpose we could see a lot of applications has been approved and tends to be defaulter. Hence this gap is needed to be filled
- A lot of applications has been cancelled from bank with purpose Repairs, Others, Urgent Meds but they have high rate of defaults as well , that means these categories are found risky to banks and they offer high interest rate and does the applicants refused the loan.



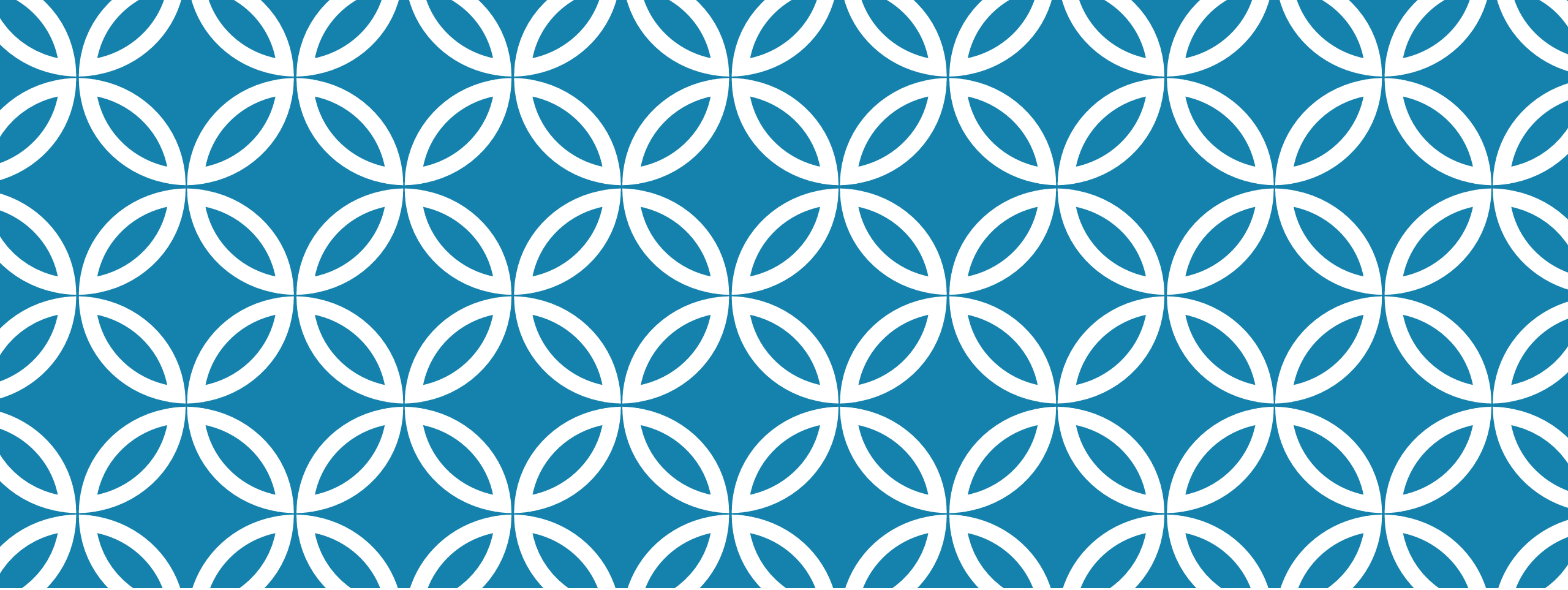
TARGET VARIABLE ANALYSIS

- We could see a lot of applications has been refused or cancelled for the applicants who have repaid their loan



TARGET VARIABLE ANALYSIS

- We could see that the people with higher income than others with unused application previously have default this time



CONCLUSIONS

Upcoming Slide depicts the conclusion derive from the overall analysis

REGARDING DEFAULTERS

- CODE_GENDER: Men are at relatively higher default rate as compare to female
- NAME_EDUCATION _TYPE: Lower Secondary has the highest Defaulter percentage
- NAME_INCOME_TYPE: Maternity leave and Unemployed are the one with maximum percentage of defaulter with 30-40% of defaulter even though the count of the application with those type is not large.
- CNT_CHILDREN & & CNT_FAM_MEMBERS : Applicants more than 9 children has high % of being a defaulter with less applications hence there application needs to be rejected.
- NAME_HOUSING_TYPE: Most defaulter lies under the category of Rented Apartment followed by Living with Parents
- ORGANIZATION_TYPE: The Organization with most default are Transport Type 3, Industry Type 13, 8 Restaurants, Construction
- OCCUPATION_TYPE: The occupation with most Default % is Low-skill Laborers followed by Drivers and Waiters/barmen staff

REGARDING DEFAULTERS

- **DAYS_BIRTH:** People among the age group from 20-40 have high % of being a defaulter
- **DAYS_EMPLOYED:** Most of the applications are from the year of experience 0-5 group but this has highest percentage of being a defaulter.
- **REGION_RATING:** Applicants with 3 as Region Rating are more likely to be defaulter
- **NAME_FAMILY_STATUS:** People with civil marriage or single are tends to be default
- **AMT_INCOME_TOTAL:** Maximum applicants who are likely to a defaulter have income ranges from 100k -225K
- **AMT_GOODS_PRICE:** When the credit amount goes beyond 3M, there is an increase in defaulters.

REGARDING RE- PAYERS

- NAME_EDUCATION _TYPE: Academic degree has less percentage of defaults
- NAME_INCOME_TYPE: Businessman and Student are the safest to repay the loan
- CNT_CHILDREN: Most applicants don't have children or just have 1 child , and also have higher chances of repaying the loan
- NAME_HOUSING_TYPE: Most of the applicants have House/Apartment and comparatively less defaulter %
- ORGANIZATION_TYPE: The Organization with less defaults are Transport Type 1, Industry Type 12, Trade Type 4
- DAYS_BIRTH: People among the age group above 50 are the safest

REGARDING RE- PAYERS

- DAYS_EMPLOYED: People with more years of experience have less percentage of being a defaulter, safest is 40-50 years of experience
- REGION_RATING: Rating 1 is less prone to being a defaulter
- NAME_FAMILY_STATUS: People who are married or window tends to repay the loan
- AMT_INCOME_TOTAL: People with more income are tends to be Repayer. Income of Businessman is highest and negligible defaulter records, we could see the same graph for Student
- NAME_CASH_LOAN_PURPOSE: Loans bought for Hobby, Buying garage are being repayed mostly.

REGARDING REFUSED APPLICATIONS

A lot of applications has been cancelled and refused from bank with purpose Repairs, Others, Urgent Meds but they have high rate of defaults as well , that means these categories are found risky to banks and been offered high interest rate and does the applicants refused the loan.

We could see a lot of applications has been refused or cancelled for the applicants who have repaid their loan in the current application.

REGARDING BUSINESS AND FINANCIAL LOSS

- From the approved applications 92% of the applicants have repaid their loan. But we could also see their are 90% of the applications that were Cancelled in previous loan which have repaid the loan in the current application. That is business loss. Hence it would be good if we could document the reason for cancelling the application so that the terms can be negotiated and can build business opportunities.
- We see that we have 8% financial loss where we approved the application but the applicant was defaulter.
- Also we see 88% of the applications that were Refused in previous loan have repaid the loan in current application, that means the interest might have been provided at higher rate due to which the applicant might have refused the loan. This is also business loss. Hence documenting the refused reason could help negotiate as well to mitigate business loss.
- Among the unknown purpose we could see a lot of applications has been approved and tends to be defaulter. Hence this gap is needed to be filled, where purpose could be documented and then the trend could be analyzed

END

Thank You!