Bank Loan Risk Analysis

AMAZING BANK

Business Objective

- Identify the defaulters, i.e. the clients that have difficulties in paying their instalments
- Possibly use the given information to take actions such as deny loans and reduce the amount of the loans (lending) to risky applicants.

This problem statement comes under **classification** – as we need to decided whether a client is **defaulter** or not? There are 1 million customers and only 5% defaulters, it is **a highly imbalanced dataset**.

Data pre-processing & Data cleaning

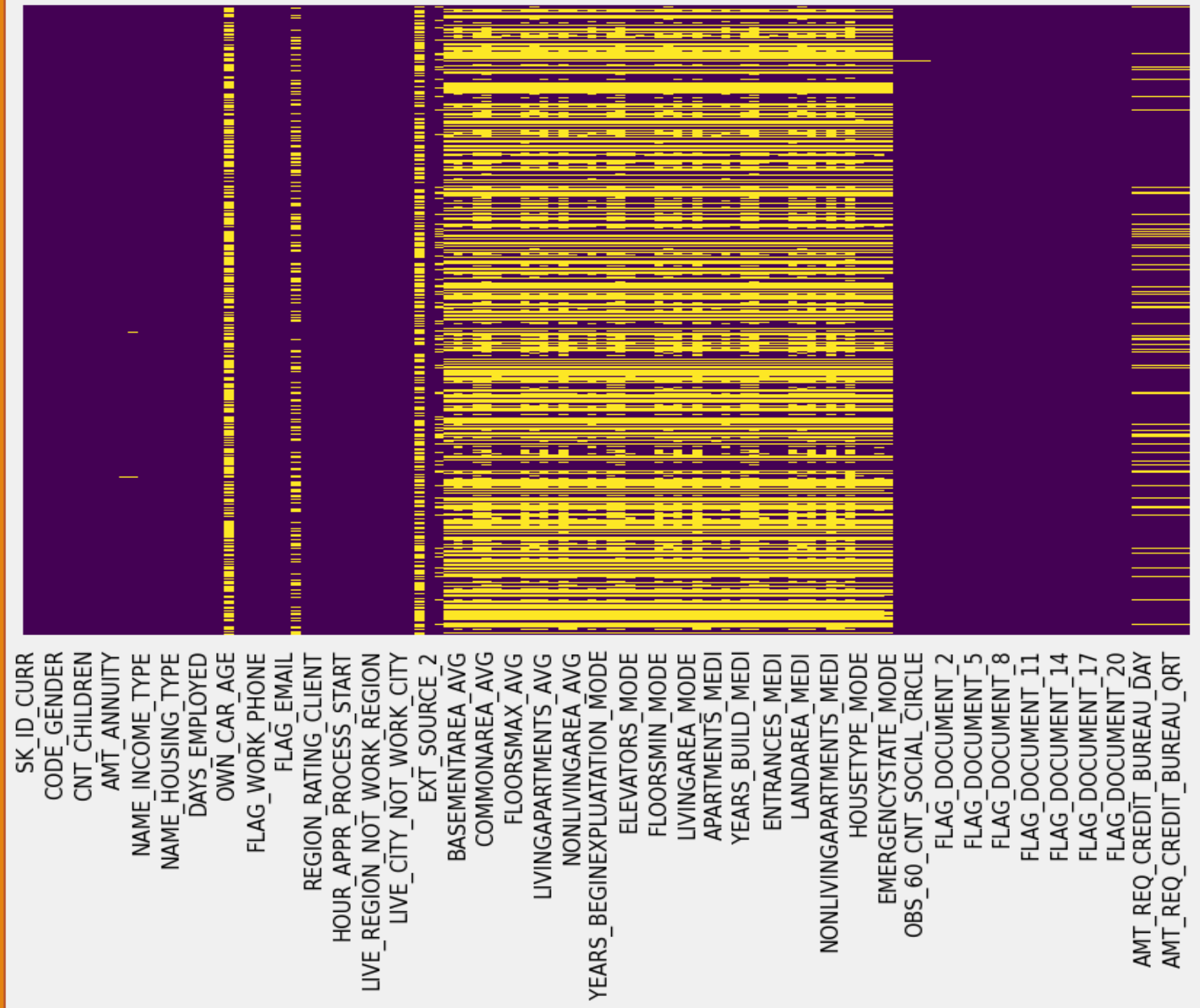
Data pre-processing & Data cleaning steps

1) Import the required libraries.
2) Importing the dataset.
3) Read the data frame and try to understand the attributes and datatypes, and we have to check for any missing values.
4) Data cleaning steps:-
   - We will check for any missing values and how much percentage (%) of the data is missing.
   - For missing numerical data, we apply the mean, median or mode strategy.
   - For missing categorical data, we use imputation techniques – dummy variables, one hot encoder or label encoder to convert it into numerical data.
5) Analyse and delete unnecessary columns, Feature engineering is required to do that means try to find out the corelation between variables using heatmap visualization.
6) Data type conversion (all the object datatype will convert from object data type to int, float or string datatype based on the requirement.

Analyse & Delete Unnecessary Columns

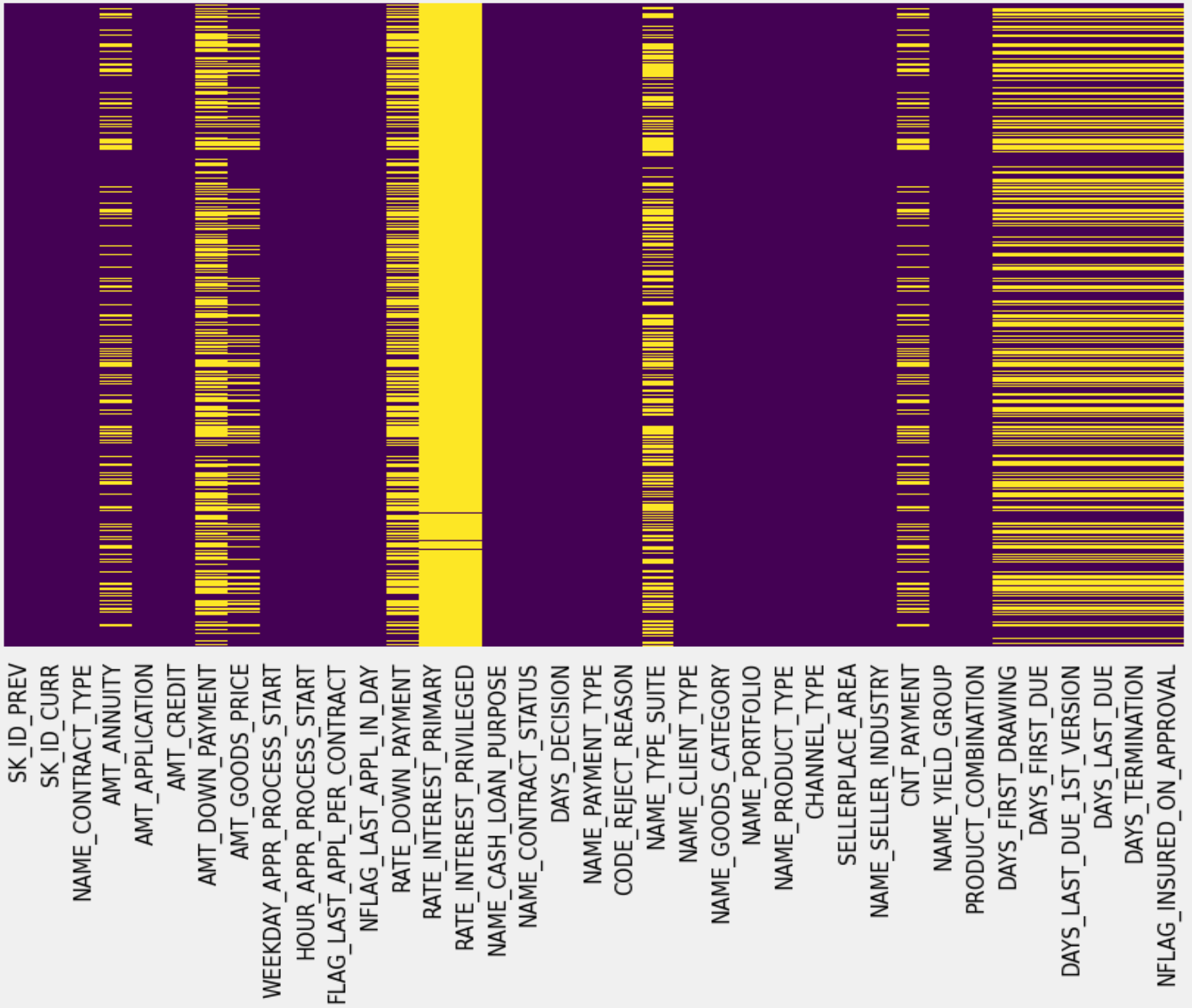
Null Values Heat Map – Application Dataframe

- There are 49 columns in applicationDF dataframe where missing value is more than 40%.



Null Values Heat Map – Previous Dataframe

- There are 11 columns in previousDF dataframe where missing value is more than 40%.



Correlation between the EXT_SOURCE_X columns and target column

Based on the above Heatmap, we can see there is almost no correlation between EXT_SOURCE_X columns and target column, thus we can drop these columns. EXT_SOURCE_1 has 56% null values, where as EXT_SOURCE_3 has close to 20% null values

Documents Provided vs Loan Repayment

- The above graph shows that in most of the loan application cases, clients who applied for loans has not submitted FLAG_DOCUMENT_X except FLAG_DOCUMENT_3.
- Thus, Except for FLAG_DOCUMENT_3, we can delete rest of the columns. Data shows if borrower has submitted FLAG_DOCUMENT_3 then there is a less chance of defaulting the loan.

Correlation Between Mobile, Email, Phone, etc and Target

There is no correlation between flags of mobile phone, email etc with loan repayment; thus these columns can be deleted

Summary: Unnecessary Columns

ApplicationDF

- Total 76 columns can be deleted from applicationDF.
- After deleting unnecessary columns, there are 46 columns remaining in applicationDF.

PreviousDF

- Total 15 columns can be deleted from previousDF.
- After deleting unnecessary columns, there are 22 columns remaining in previousDF.
- WEEKDAY_APPR_PROCESS_START', 'HOUR_APPR_PROCESS_START', 'FLAG_LAST_APPL_PER_CONTRACT' ,'NFLAG_LAST_APPL_IN_DAY' were identified as unnecessary columns.

Standardize Values

Standardize Values

Strategy for applicationDF:

- Convert DAYS_DECISION,DAYS_EMPLOYED, DAYS_REGISTRATION,DAYS_ID_PUBLISH from negative to positive as days cannot be negative.
- Convert DAYS_BIRTH from negative to positive values and calculate age and create categorical bins columns
- Categorize the amount variables into bins
- Convert region rating column and few other columns to categorical

Strategy for previousDF:

- Convert DAYS_DECISION from negative to positive values and create categorical bins columns.
- Convert loan purpose and few other columns to categorical.

Null Values Data Imputation

Null Values Data Imputation

Strategy for applicationDF:

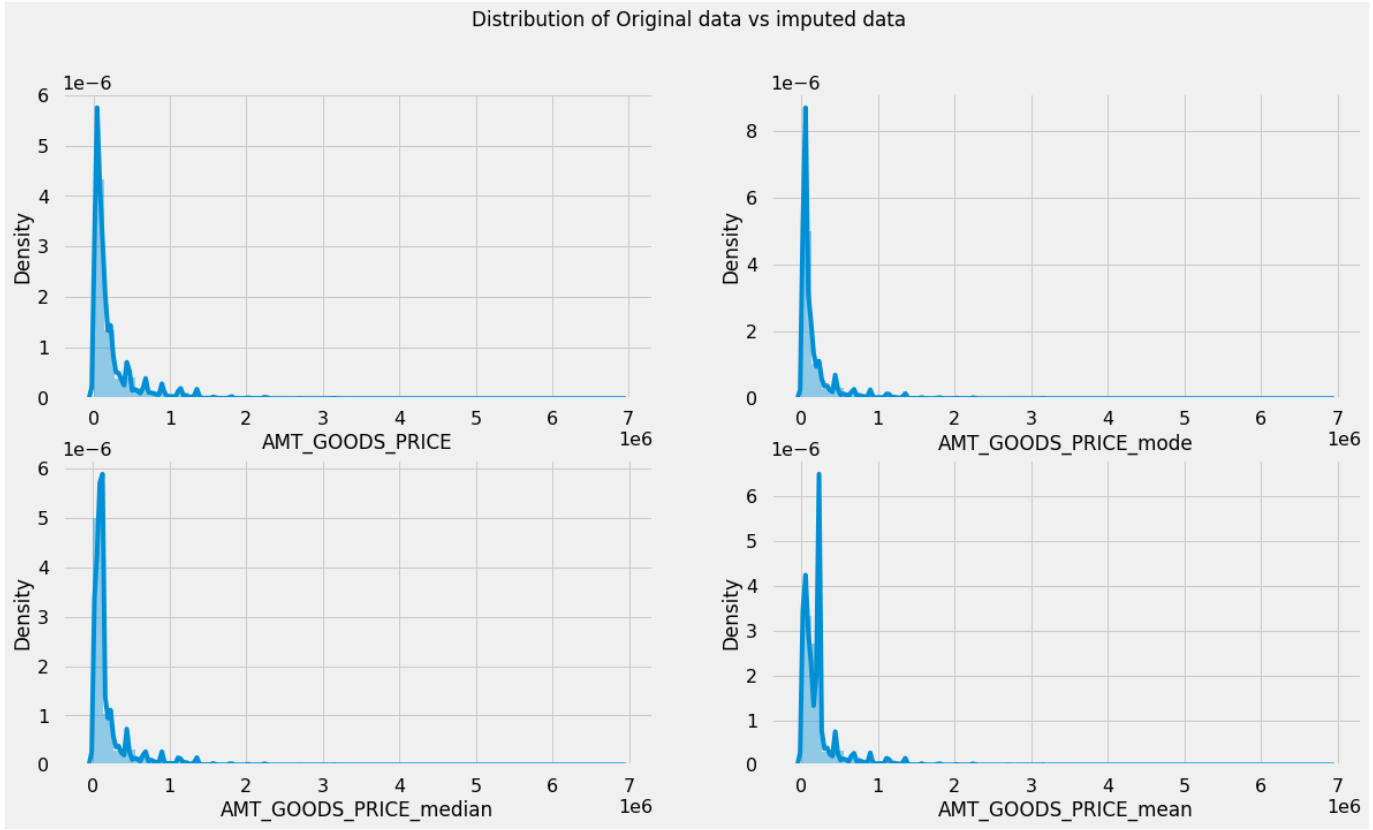
- To impute null values in categorical variables which has lower null percentage, mode() is used to impute the most frequent items.
- To impute null values in categorical variables which has higher null percentage, a new category is created.
- To impute null values in numerical variables which has lower null percentage, median() is used as
  - There are no outliers in the columns
  - Mean returned decimal values and median returned whole numbers and the columns were number of requests

Strategy for applicationDF:

- To impute null values in numerical column, we analysed the loan status and assigned values.
- To impute null values in continuous variables, we plotted the distribution of the columns and used
  - median if the distribution is skewed
  - mode if the distribution pattern is preserved.

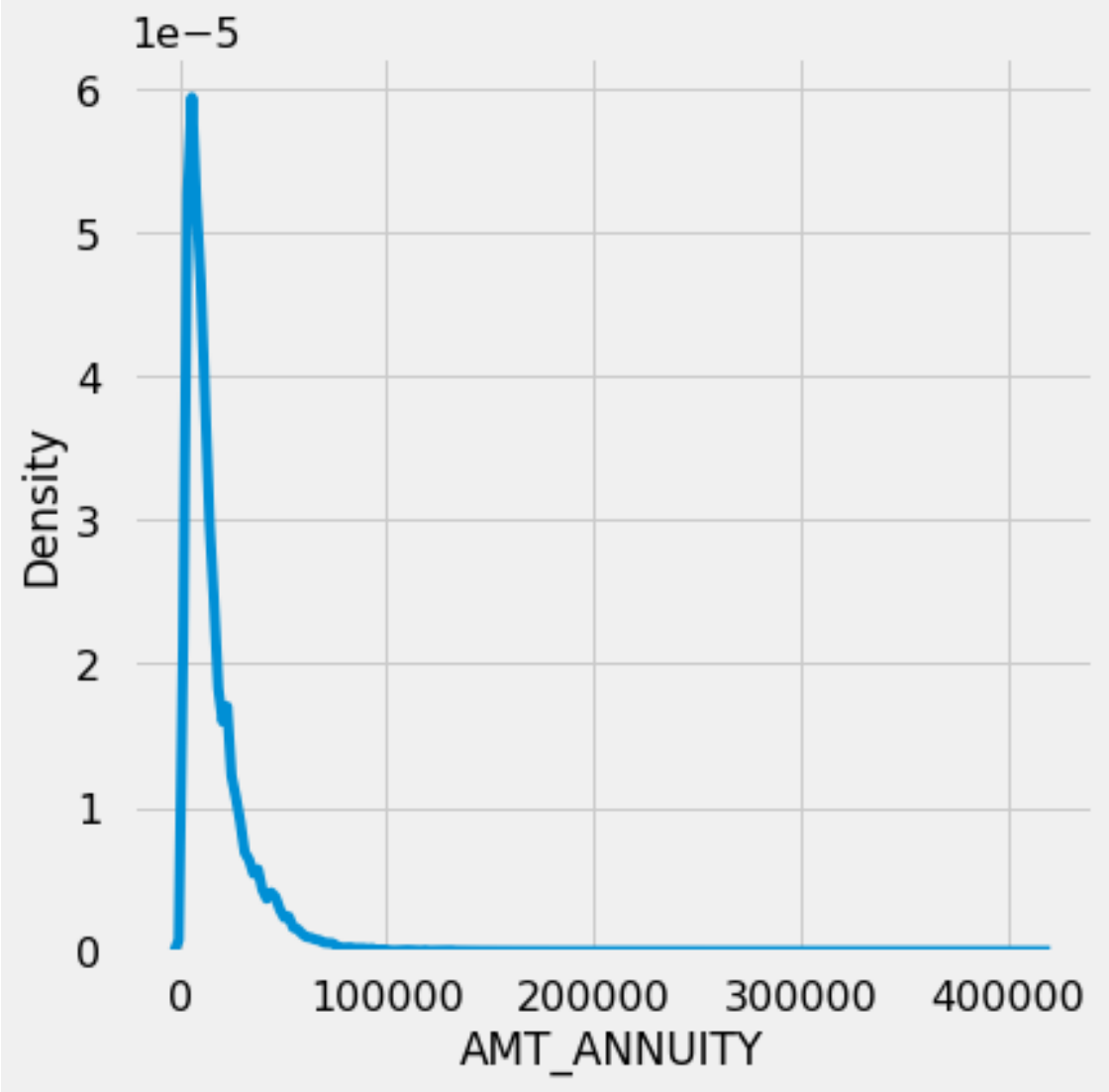
Distribution of AMT_GOODS_PRICE originally and after imputation.

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



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.



Summary: Data Imputation

- Impute categorical variable 'NAME_TYPE_SUITE' which has lower null percentage(0.42%) with the most frequent category using mode()[0].
- Impute categorical variable 'OCCUPATION_TYPE' which has higher null percentage(31.35%) with a new category as assigning to any existing category might influence the analysis.
- Impute numerical variables with the median as there are no outliers that can be seen from results of describe() and mean() returns decimal values and these columns represent number of enquiries made which cannot be decimal.
- Impute 'AMT_REQ_CREDIT_BUREAU_HOUR', 'AMT_REQ_CREDIT_BUREAU_DAY','AMT_REQ_CREDIT_BUREAU_WEEK','AMT_REQ_CREDIT_BUREAU_MON', 'AMT_REQ_CREDIT_BUREAU_QRT','AMT_REQ_CREDIT_BUREAU_YEAR' with median as mean has decimals and this is number of requests.
- Impute AMT_ANNUITY with median as the distribution is greatly skewed.
- Impute AMT_GOODS_PRICE with mode as the distribution is closely similar.
- Impute CNT_PAYMENT with 0 as the NAME_CONTRACT_STATUS for these indicate that most of these loans were not starte.

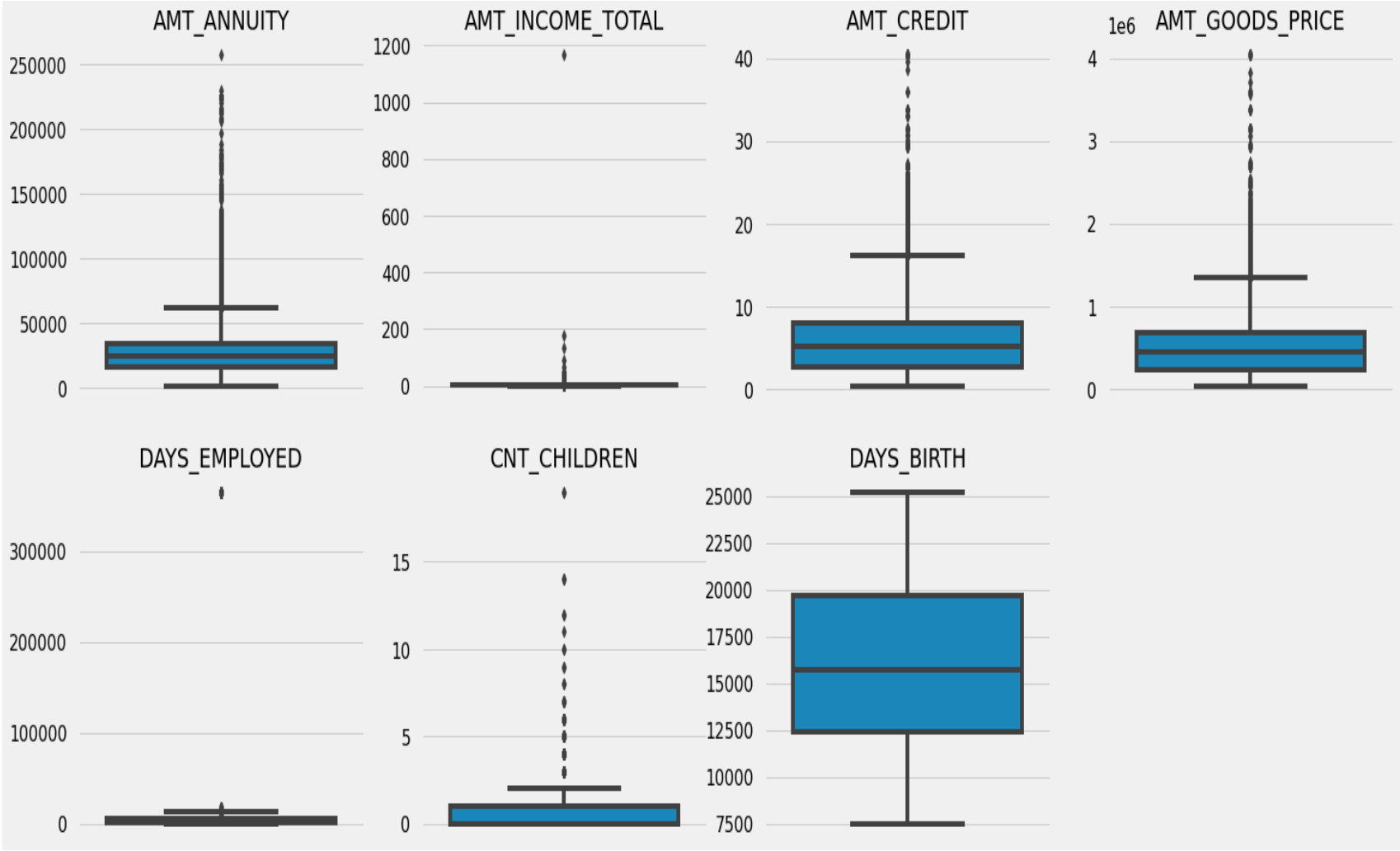
Identifying the outliers (applicationDF)

1. AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.

2. AMT_INCOME_TOTAL has huge number of outliers which indicate that few of the loan applicants have high income when compared to the others.

3. DAYS_BIRTH has no outliers which means the data available is reliable.

4. DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and hence this has to be incorrect entry.



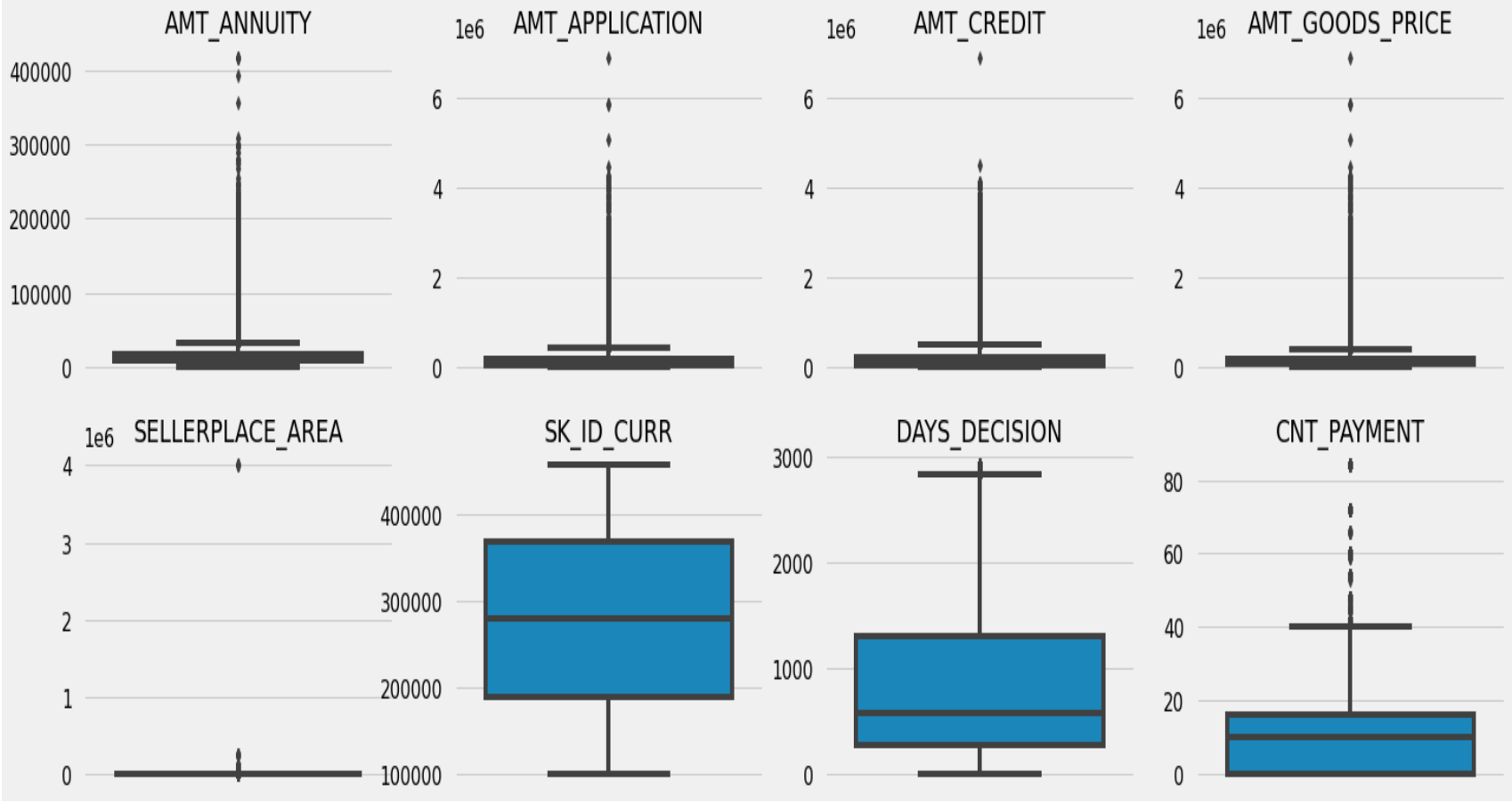
Identifying the outliers (previousDF)

1. AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.

2. CNT_PAYMENT has few outlier values.

3. SK_ID_CURR is an ID column and hence no outliers.

4. DAYS_DECISION has little number of outliers indicating that these previous applications decisions were taken long back.



Data Analysis

Data Analysis

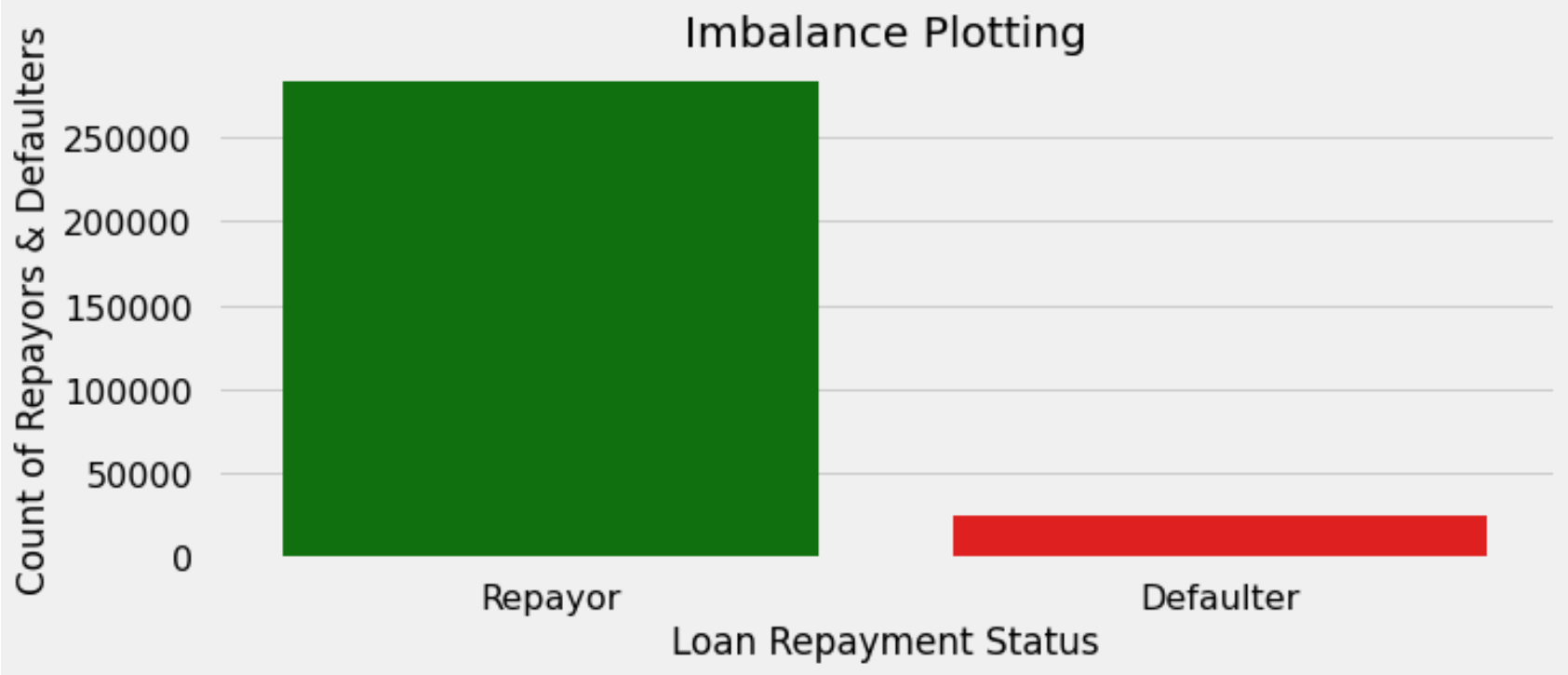
Strategy:

The data analysis flow has been planned in following way :

- Imbalance in Data
- Categorical Data Analysis
  - Categorical segmented Univariate Analysis
  - Categorical Bi/Multivariate analysis
- Numeric Data Analysis
  - Bi-furcation of databased based on TARGET data
  - Correlation Matrix
  - Numerical segmented Univariate Analysis
  - Numerical Bi/Multivariate analysis

Imbalance Analysis

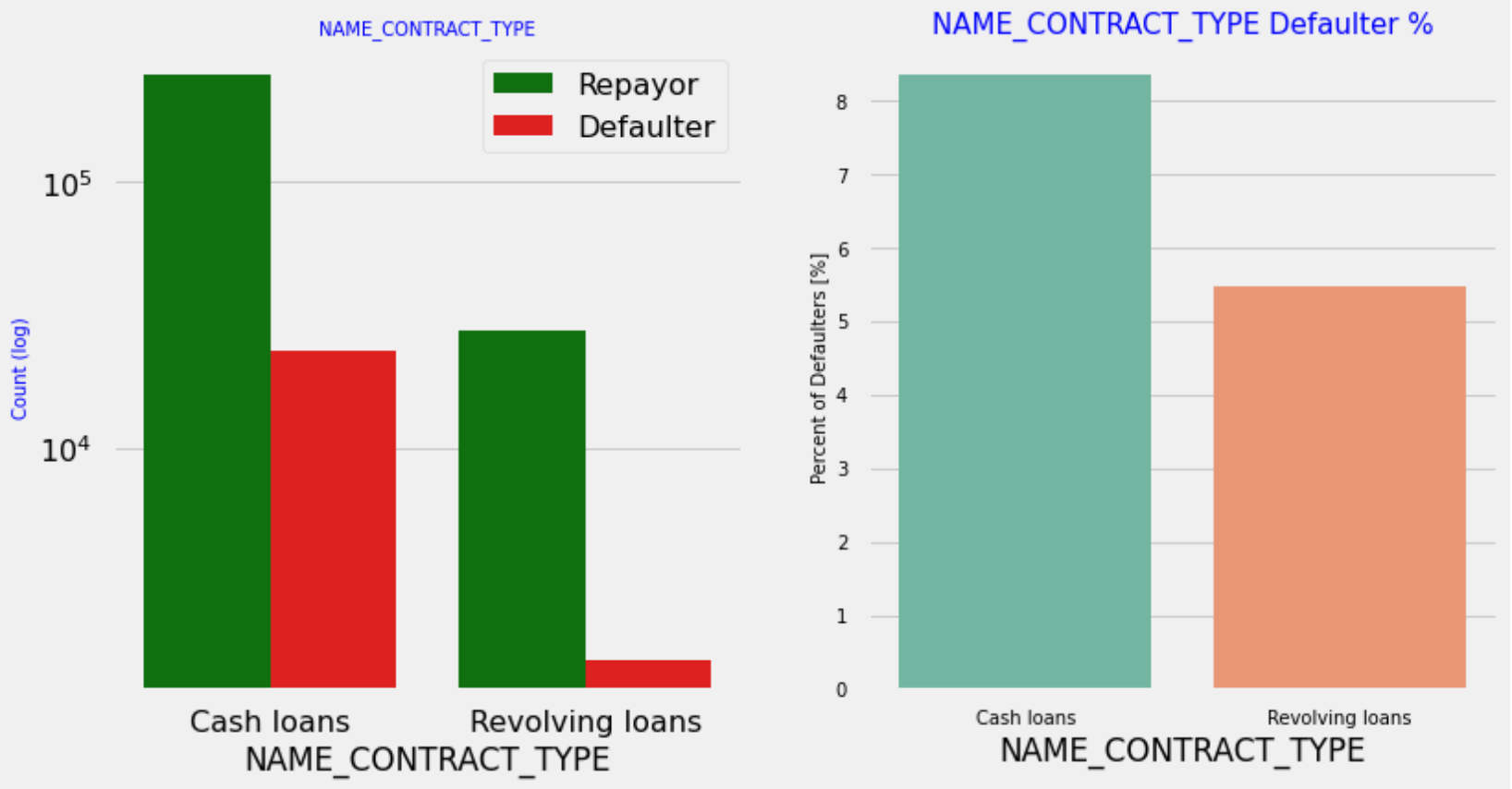
- Ratios of imbalance in percentage with respect to Repayer and Defaulter datas are: 91.93 and 8.07
- Ratios of imbalance in relative with respect to Repayer and Defaulter datas is 11.39 : 1 (approx.)



Segmented Univariate Analysis

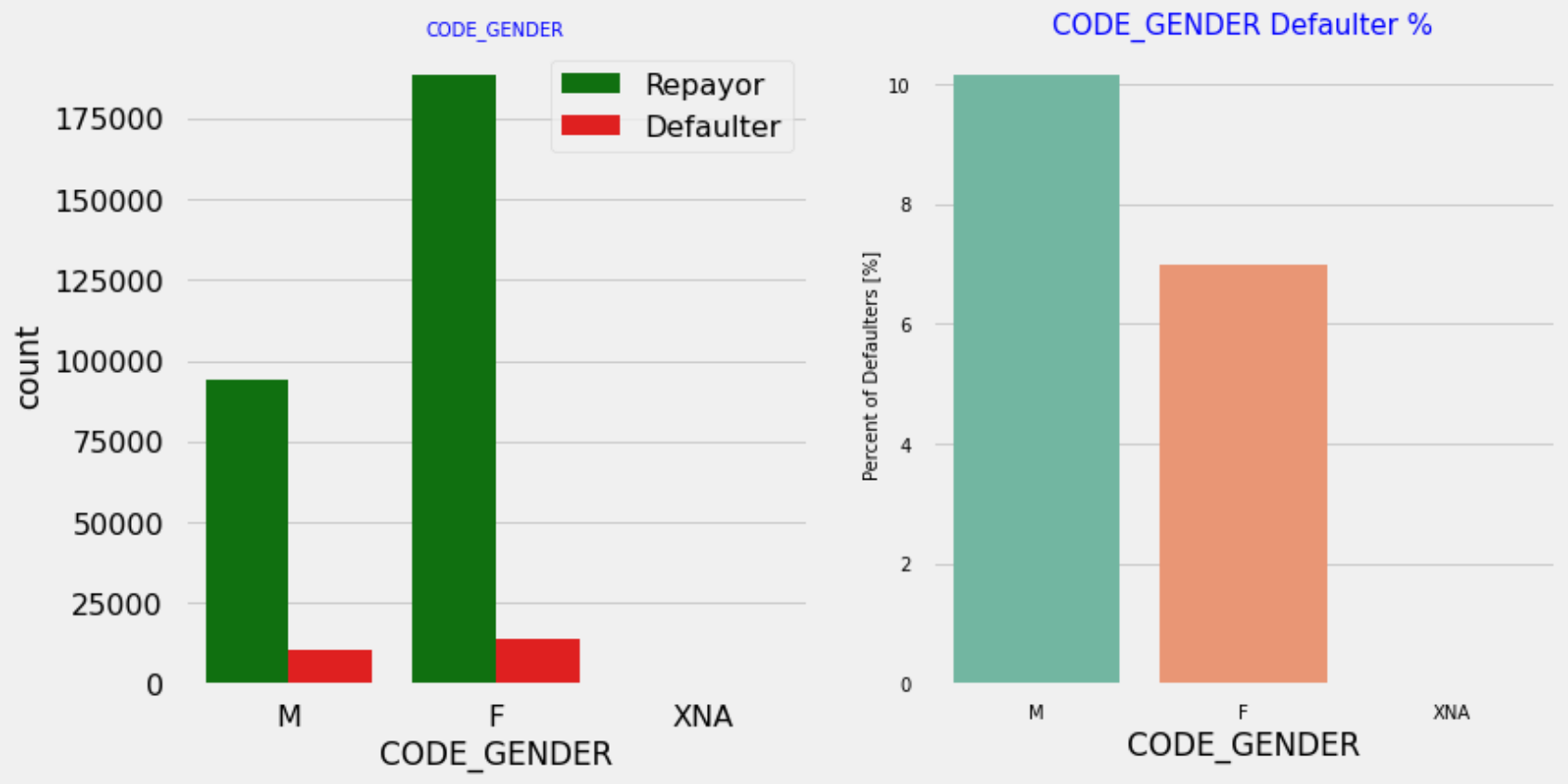
Checking the contract type based on loan repayment status

Contract type: Revolving loans are just a small fraction (10%) from the total number of loans; in the same time, a larger amount of Revolving loans, comparing with their frequency, are not repaid.



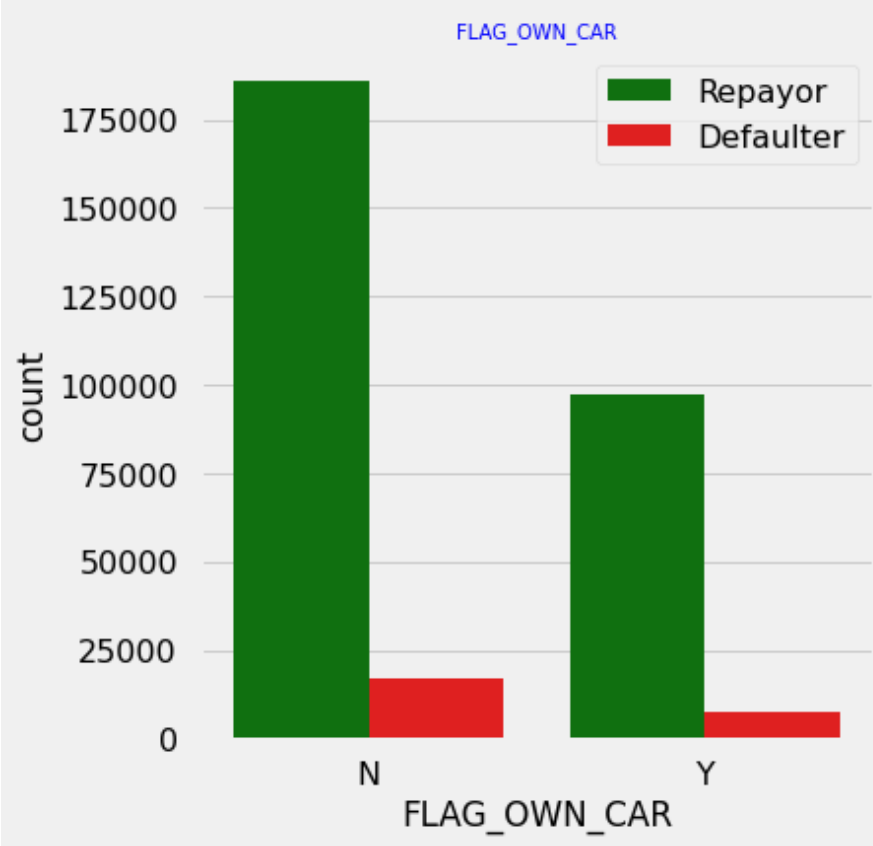
Checking the type of Gender on loan repayment status

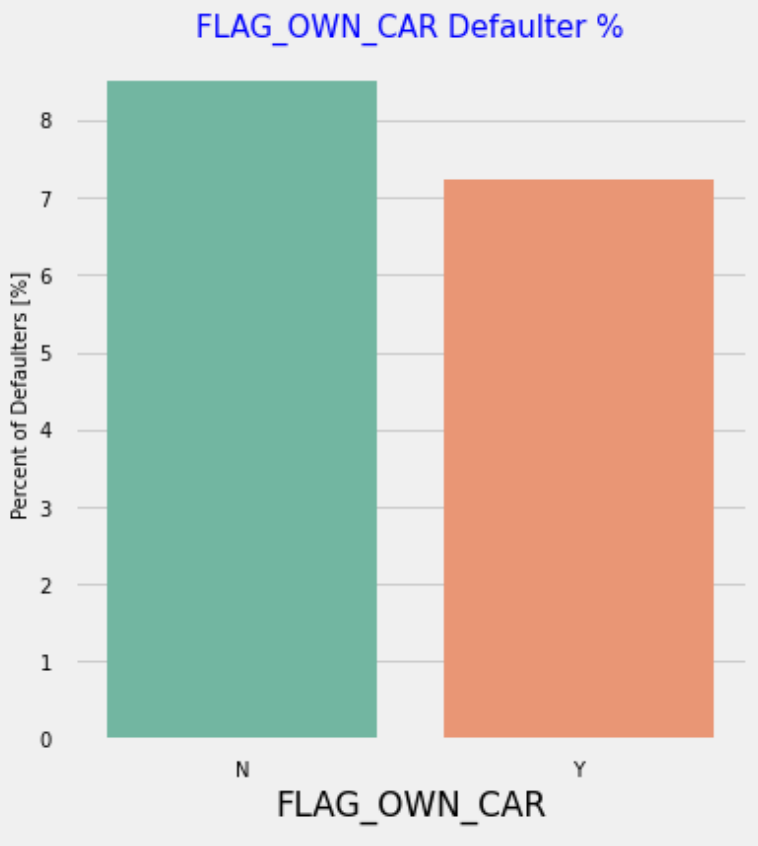
The number of female clients is almost double the number of male clients. Based on the percentage of defaulted credits, males have a higher chance of not returning their loans (~10%), comparing with women (~7%)



Checking if owning a car is related to loan repayment status

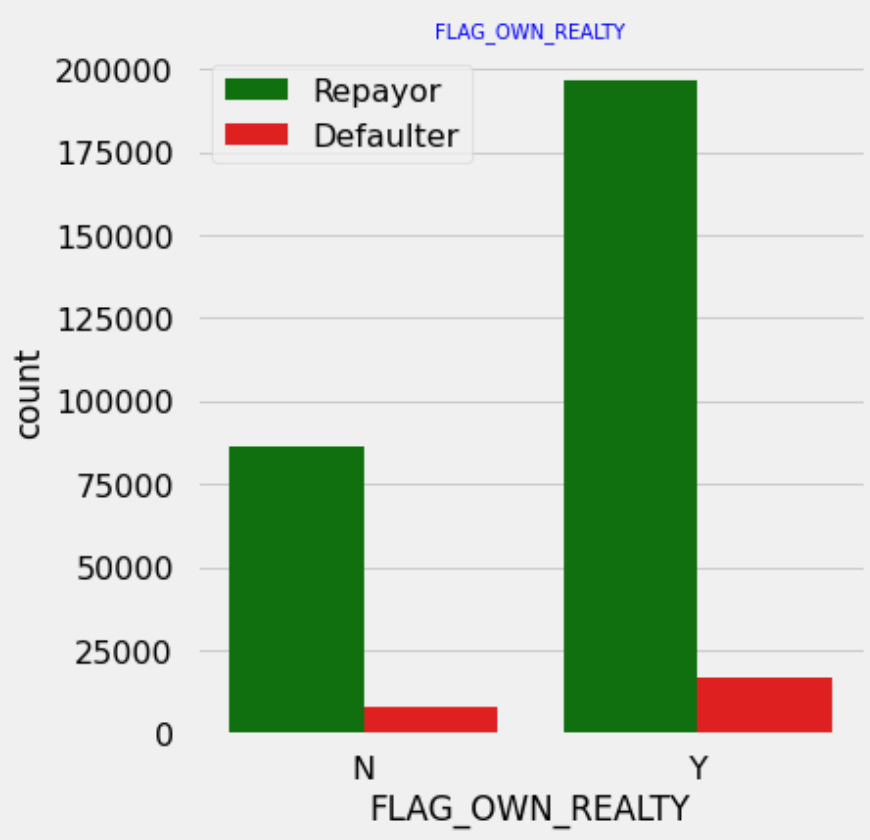
Clients who own a car are half in number of the clients who dont own a car. But based on the percentage of deault, there is no correlation between owning a car and loan repayment as in both cases the default percentage is almost same.

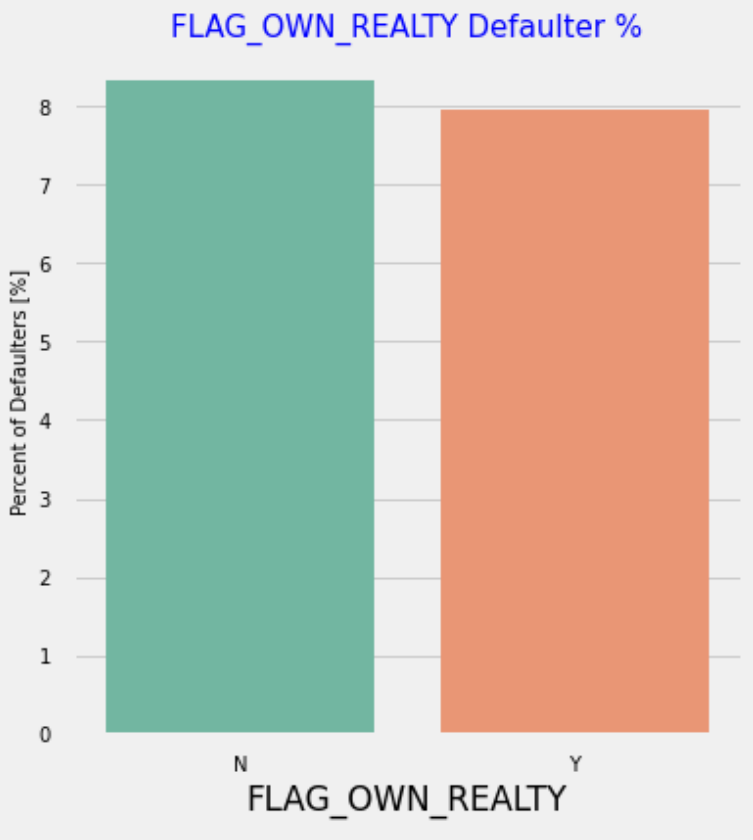




Checking if owning a realty is related to loan repayment status

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.



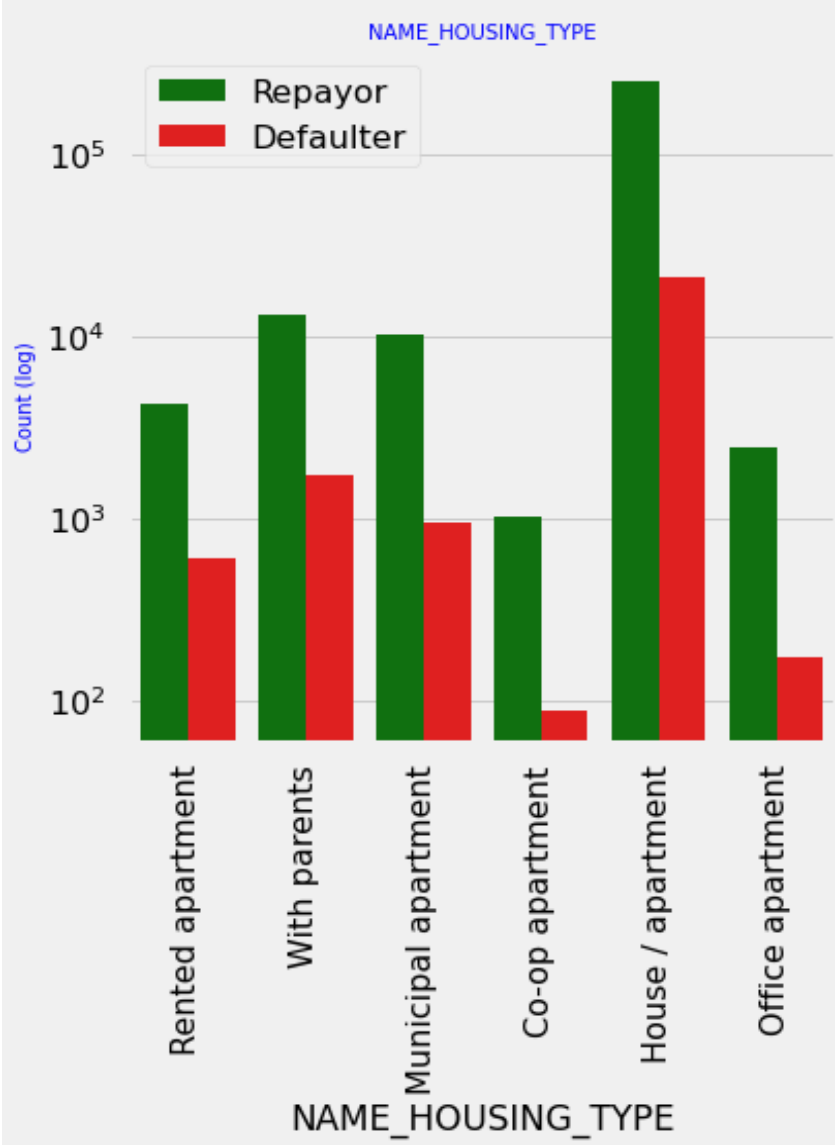


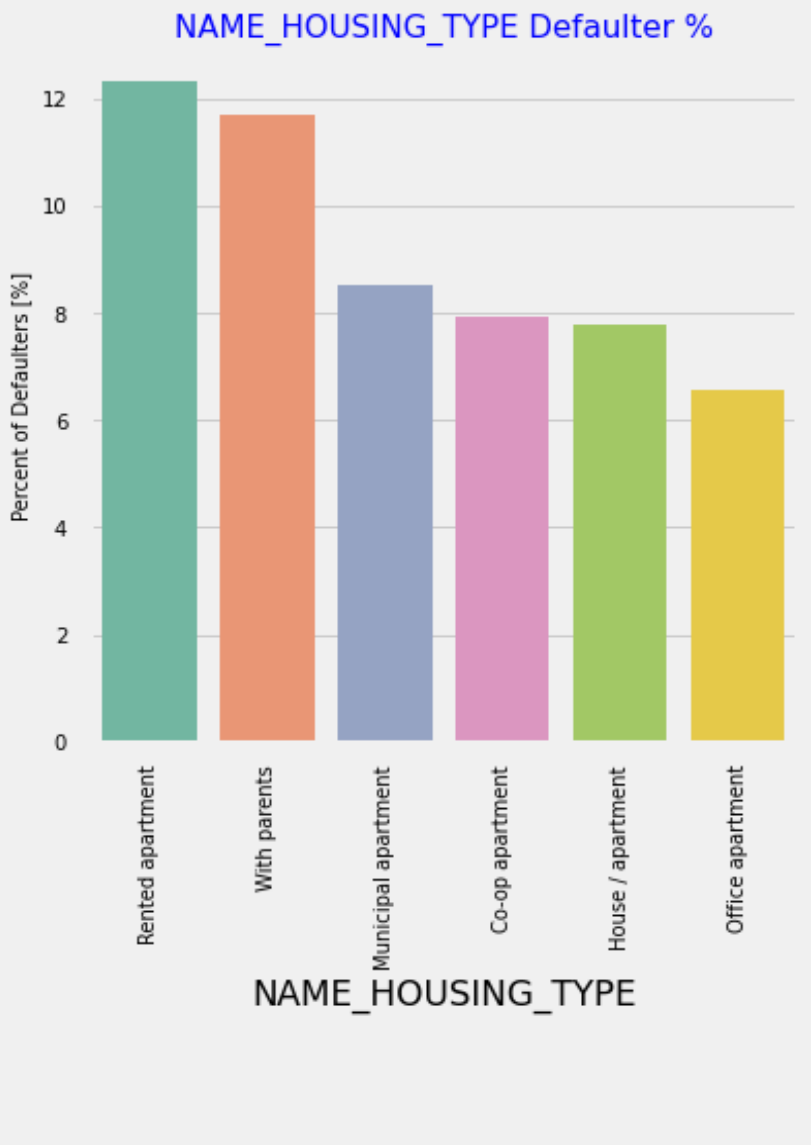
Analysing Housing Type based on loan repayment status

1. Majority of people live in House/apartment

2. People living in office apartments have lowest default rate

3. People living with parents (~11.5%) and living in rented apartments(>12%) have higher probability of defaulting

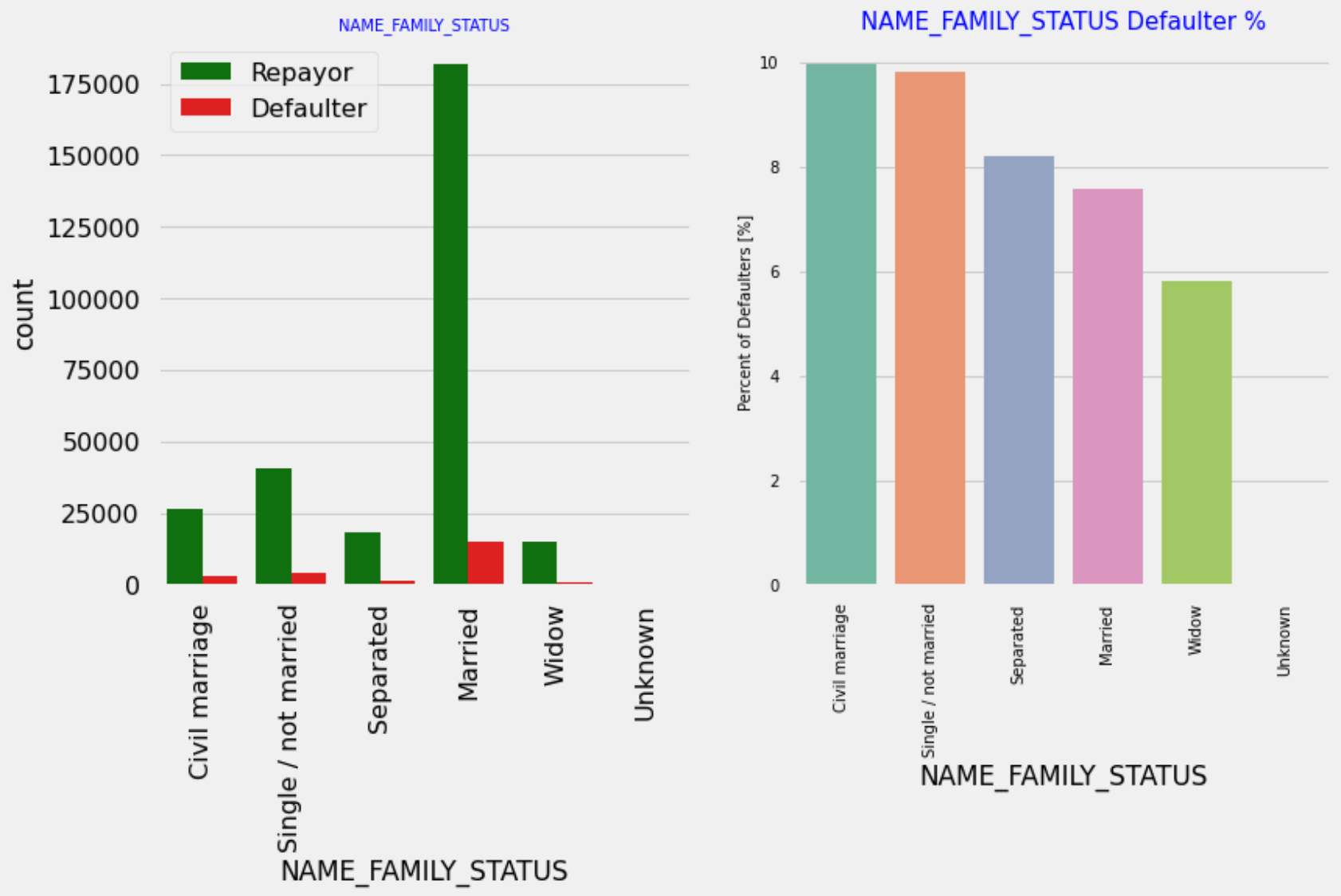




Analysing Family status based on loan repayment status

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

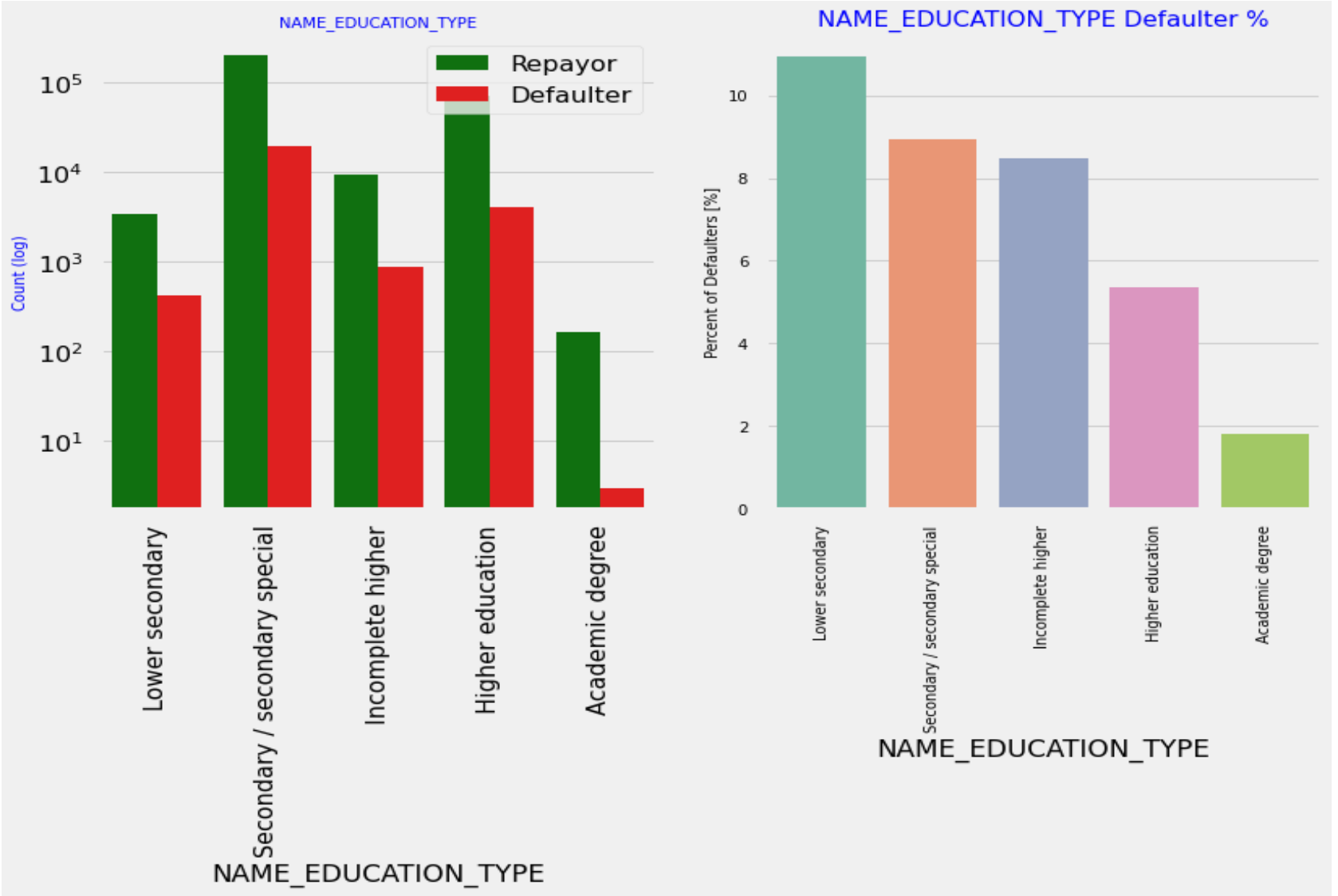
2. In terms of percentage of not repayment of loan, Civil marriage has the highest percent of not repayment (10%), with Widow the lowest (exception being Unknown).



Analysing Education Type based on loan repayment status

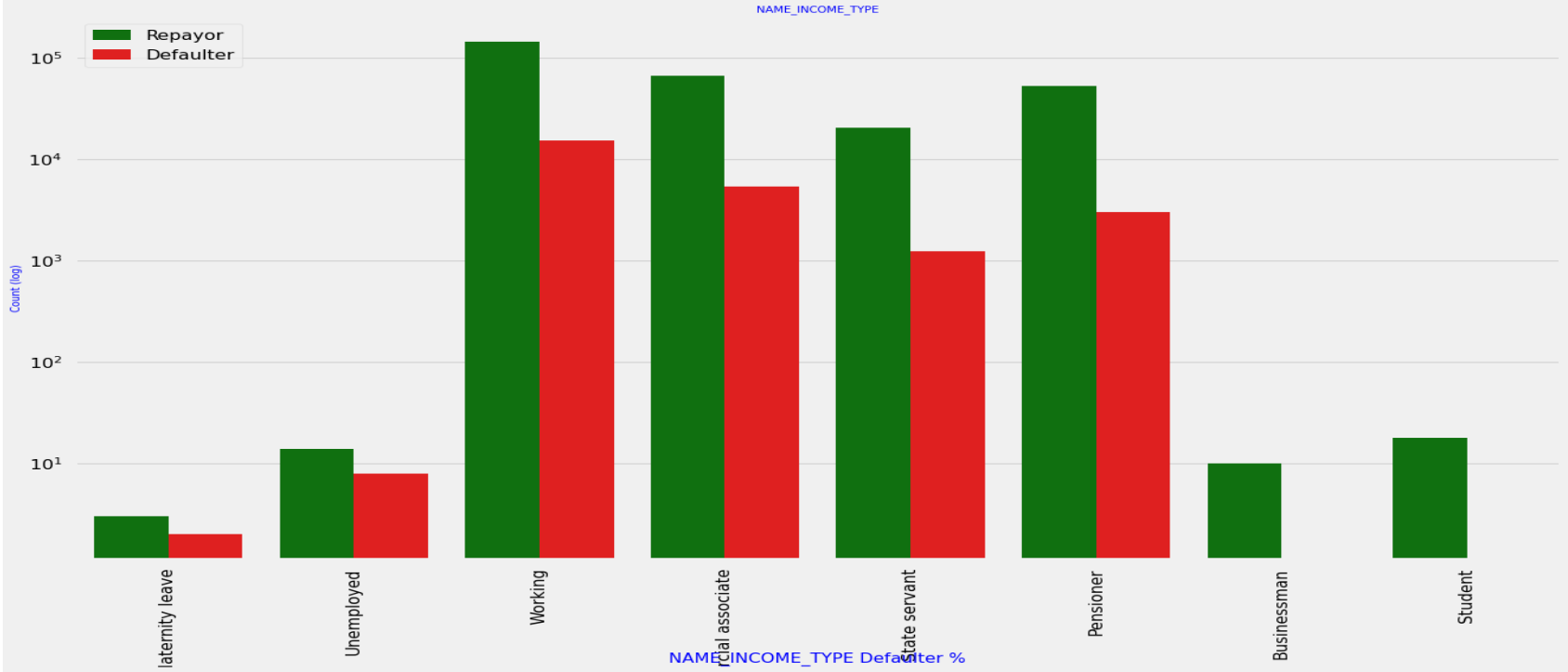
1. Majority of the clients have Secondary / secondary special education, followed by clients with Higher education. Only a very small number having an academic degree

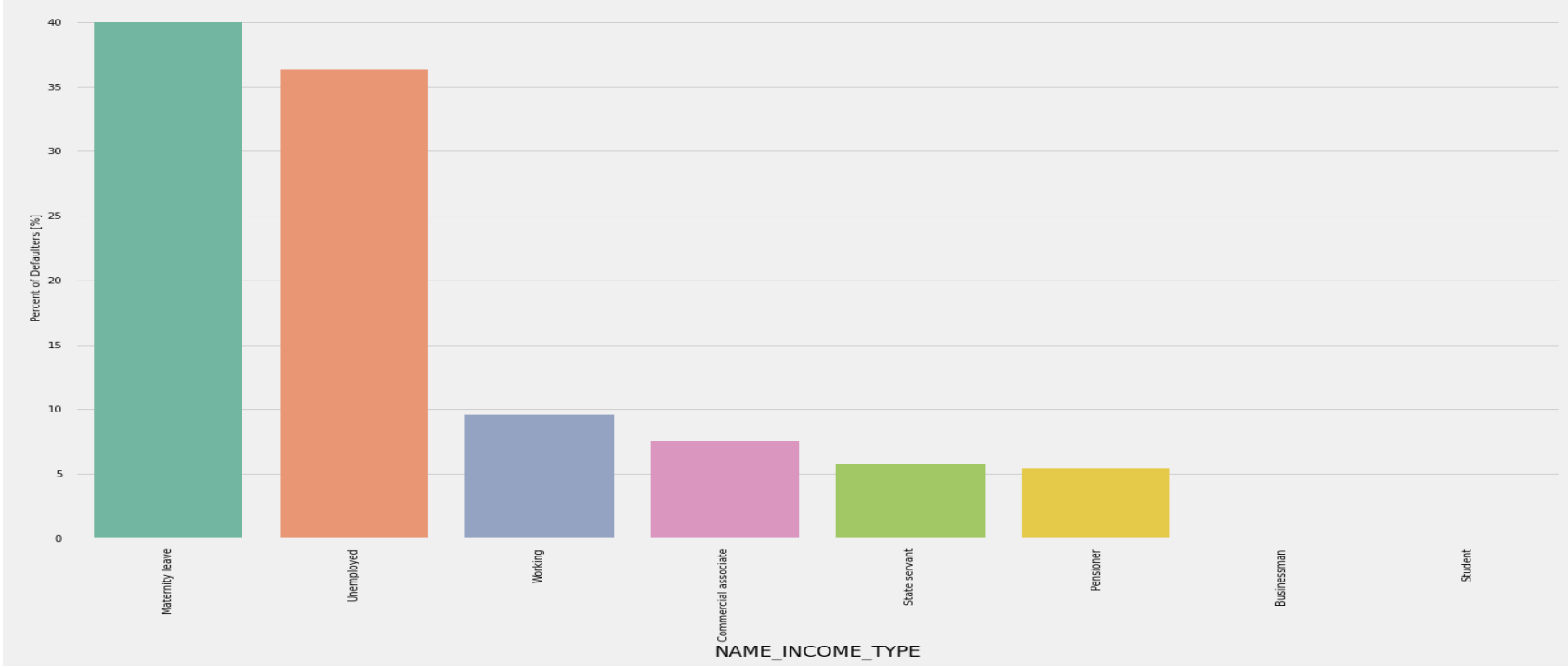
2. The Lower secondary category, although rare, have the largest rate of not returning the loan (11%). The people with Academic degree have less than 2% defaulting rate.



Analysing Income Type based on loan repayment status

- 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 40% ratio of not returning loans, 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.



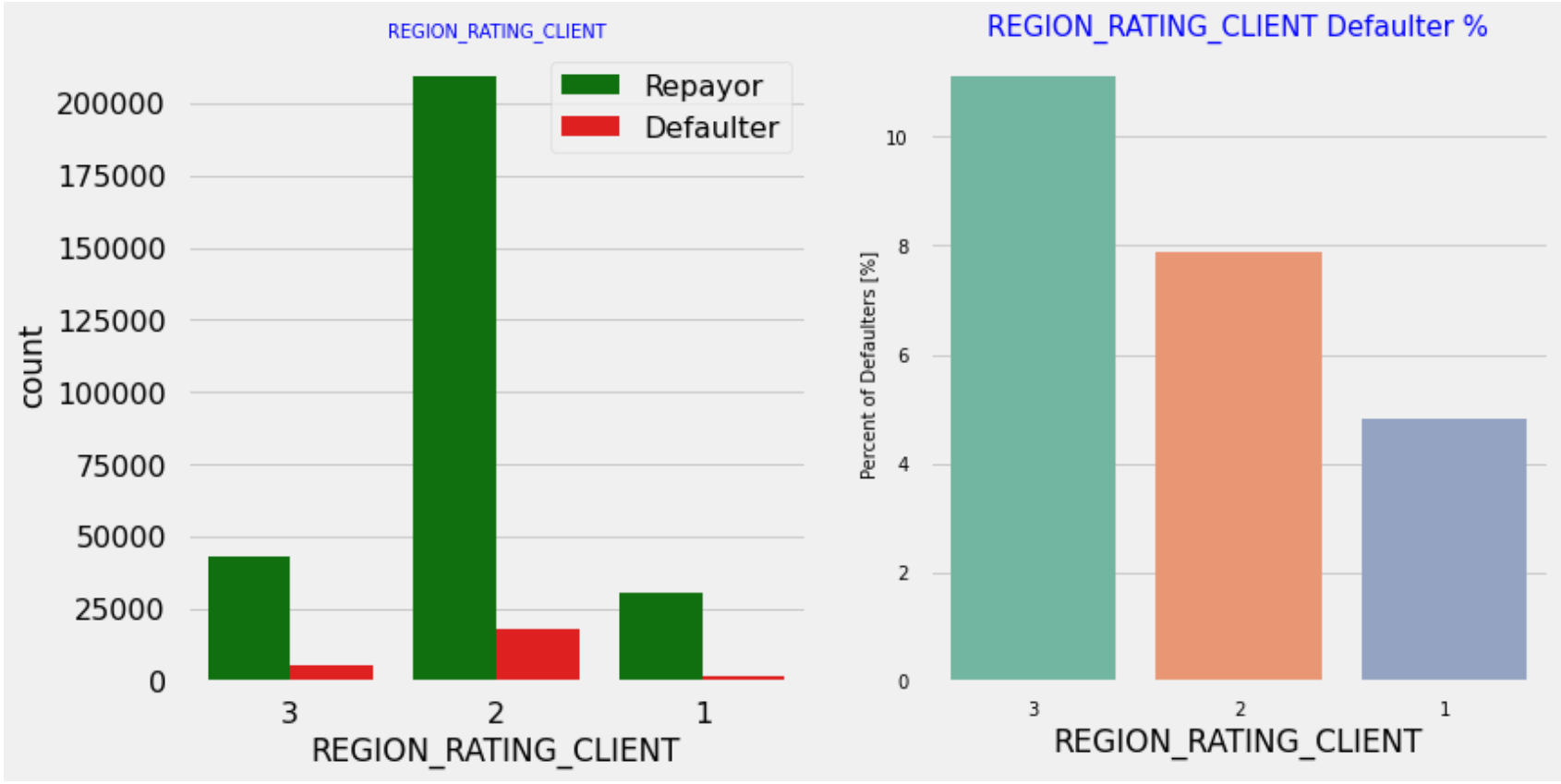


Analysing Region rating where applicant lives based on loan repayment status

1. Most of the applicants are living in Region_Rating 2 place.

2. Region Rating 3 has the highest default rate (11%)

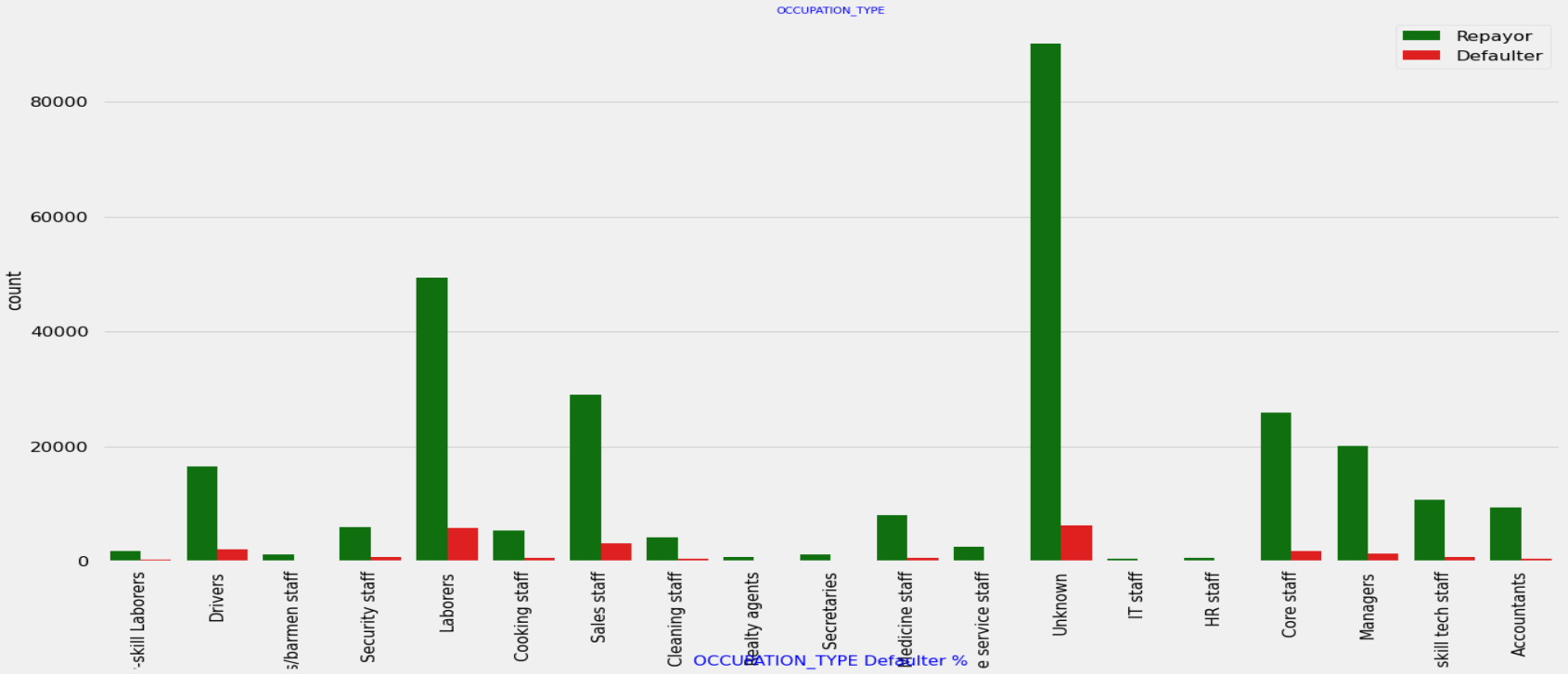
3. Applicant living in Region_Rating 1 has the lowest probability of defaulting, thus **safer** for approving loans

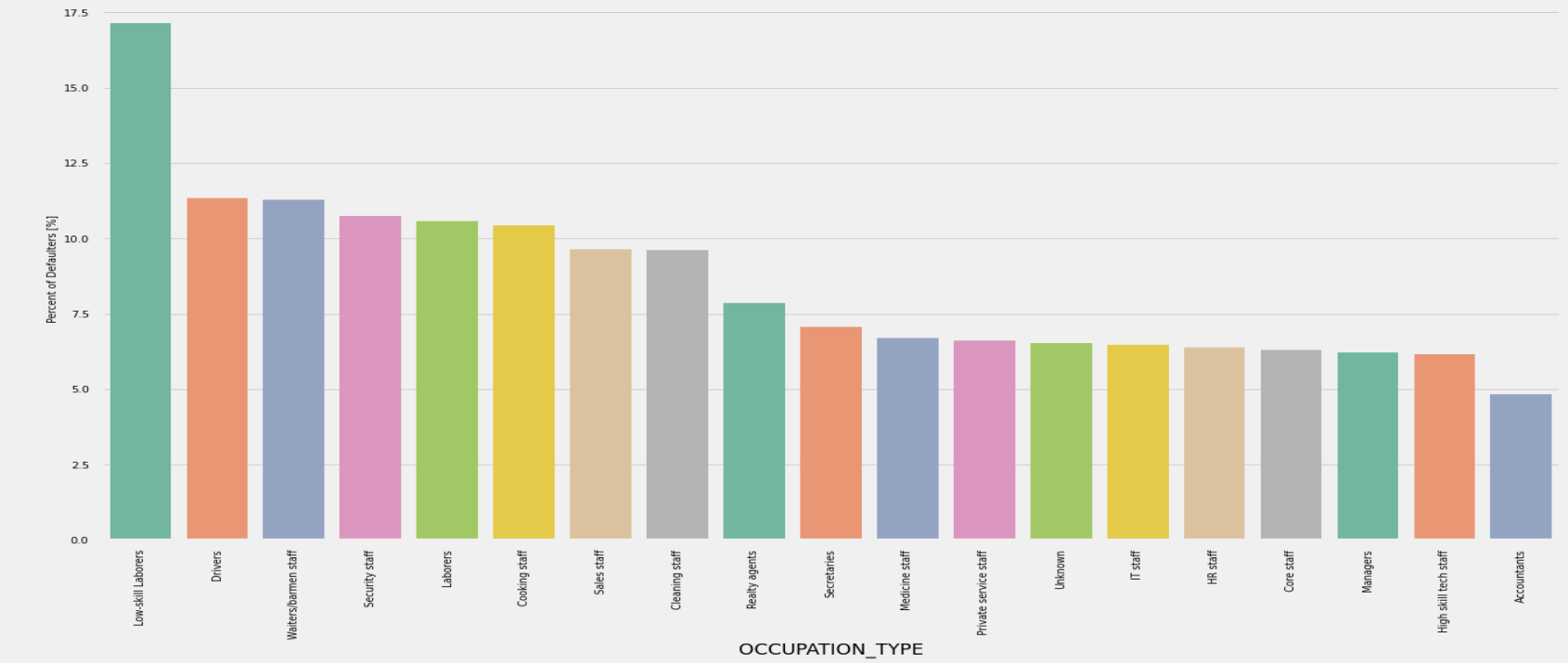


Analysing Occupation Type where applicant lives based on loan repayment status

1. Most of the loans are taken by Laborers, followed by Sales staff. IT staff take the lowest amount of loans.

2. 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.





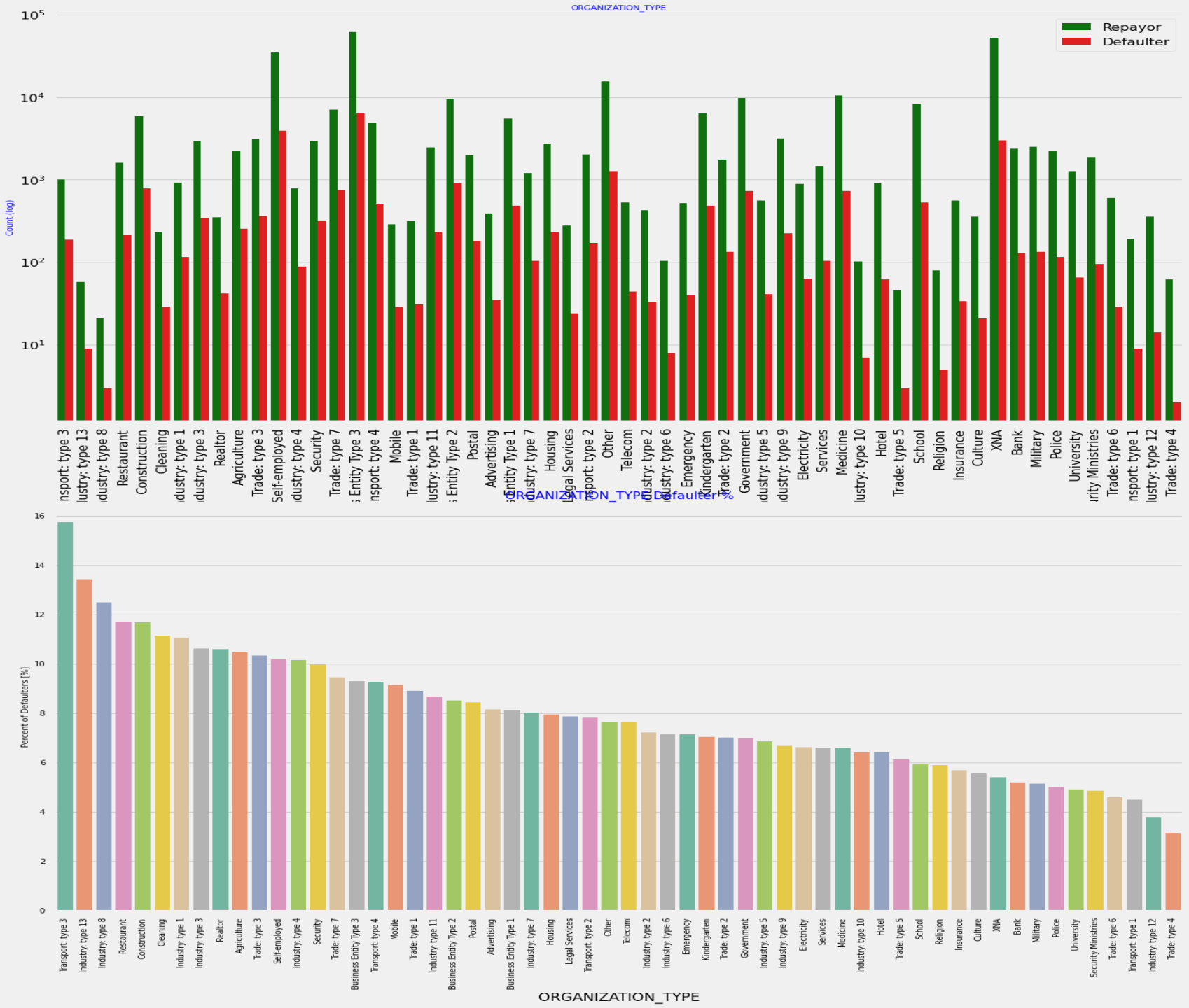
Checking Loan repayment status based on Organization type

1. Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self employed people have relative high defaulting rate, and thus should be avoided to be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting.

2. Most of the people application for loan are from Business Entity Type 3

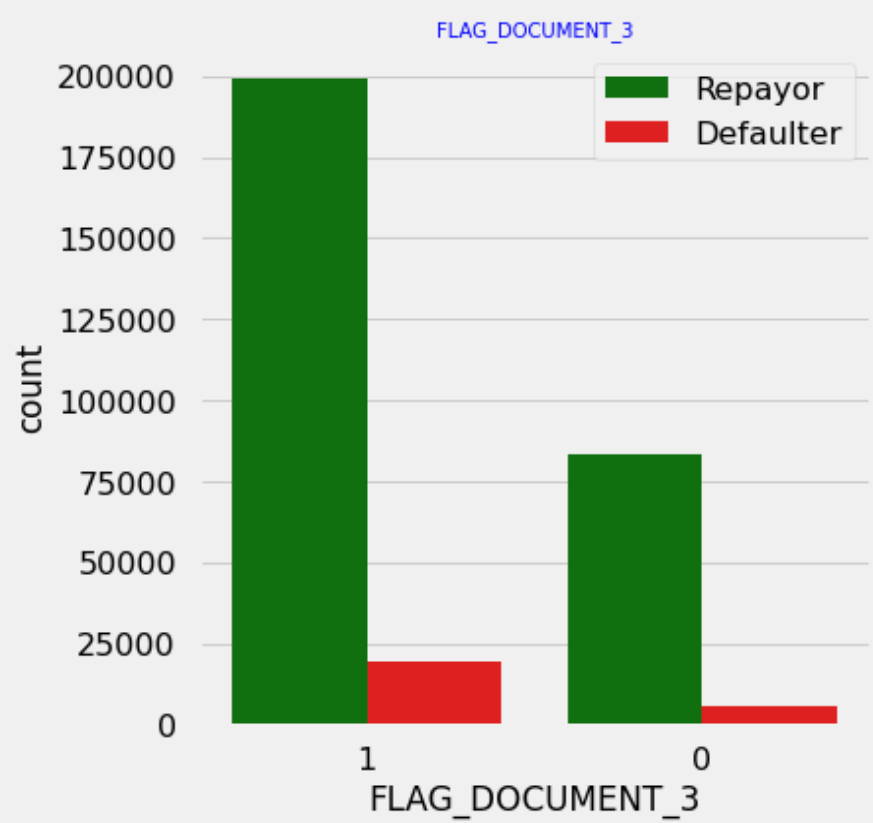
3. For a very high number of applications, Organization type information is unavailable(XNA)

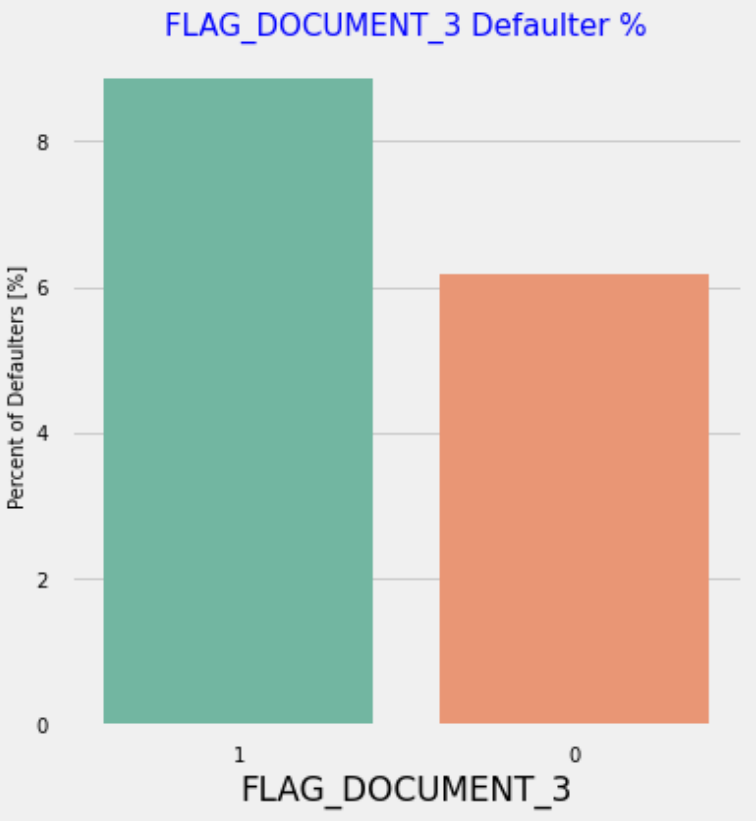
- It can be seen that following category of organization type has lesser defaulters thus safer for providing loans: Trade Type 4 and 5 Industry type 8



Analysing Flag_Doc_3 submission status based on loan repayment status

There is no significant correlation between repayers and defaulters in terms of submitting document 3 as we see even if applicants have submitted the document, they have defaulted a slightly more (~9%) than who have not submitted the document (6%)

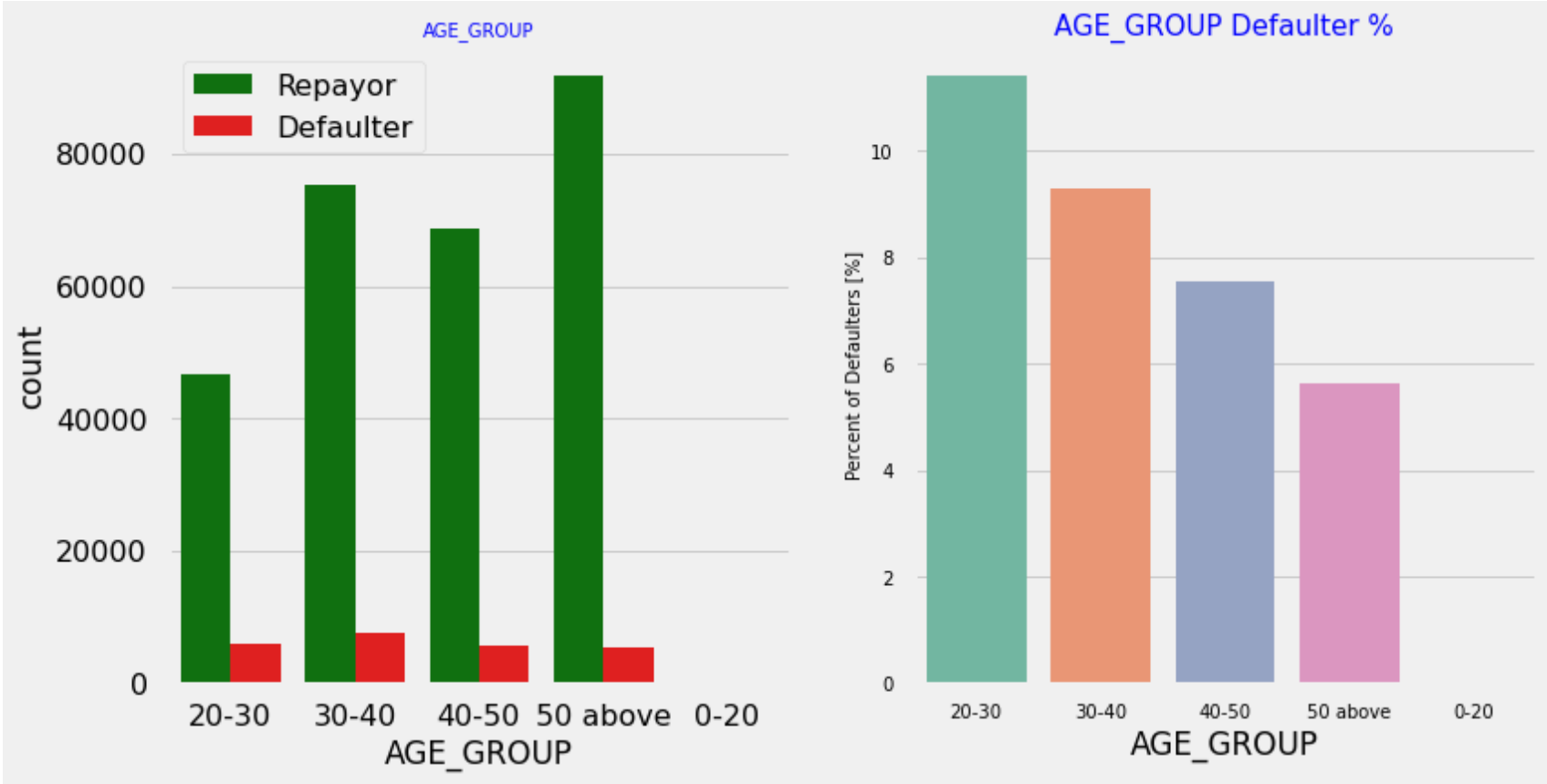




Analyzing Age Group based on loan repayment status

1. People in the age group range 20-40 have higher probability of defaulting

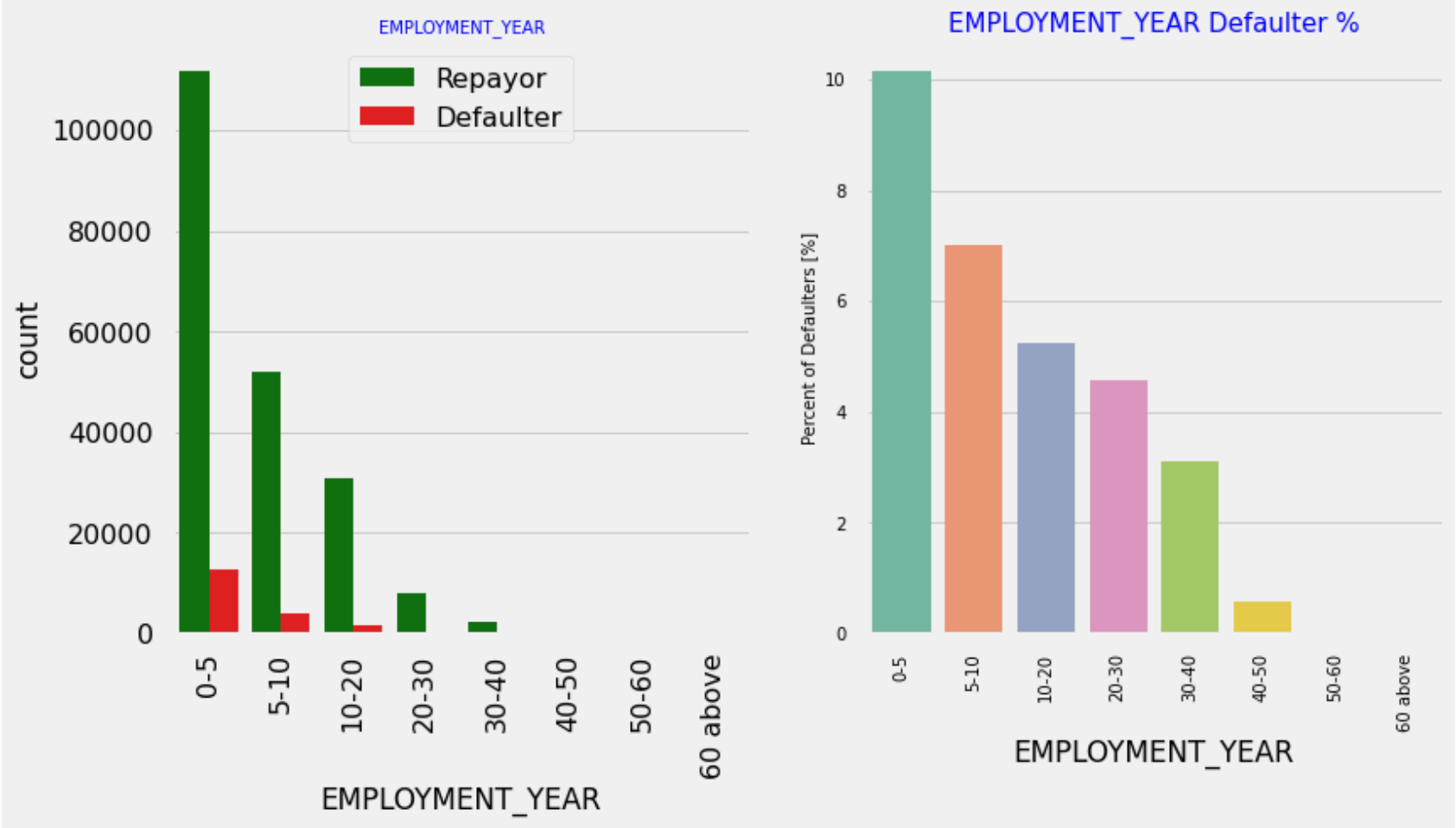
2. People above age of 50 have low probability of defaulting



Analysing Employment_Year based on loan repayment status

1. 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%

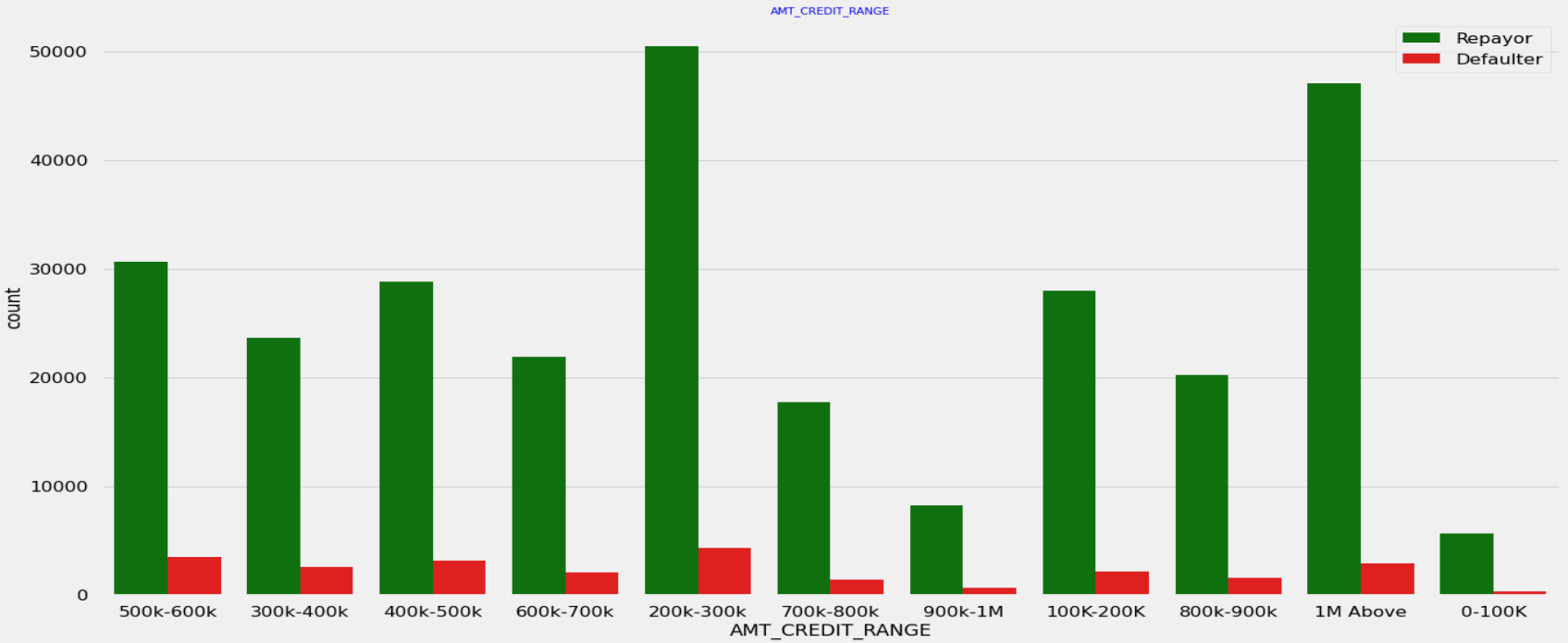
2. With increase of employment year, defaulting rate is gradually decreasing with people having 40+ year experience having less than 1% default rate

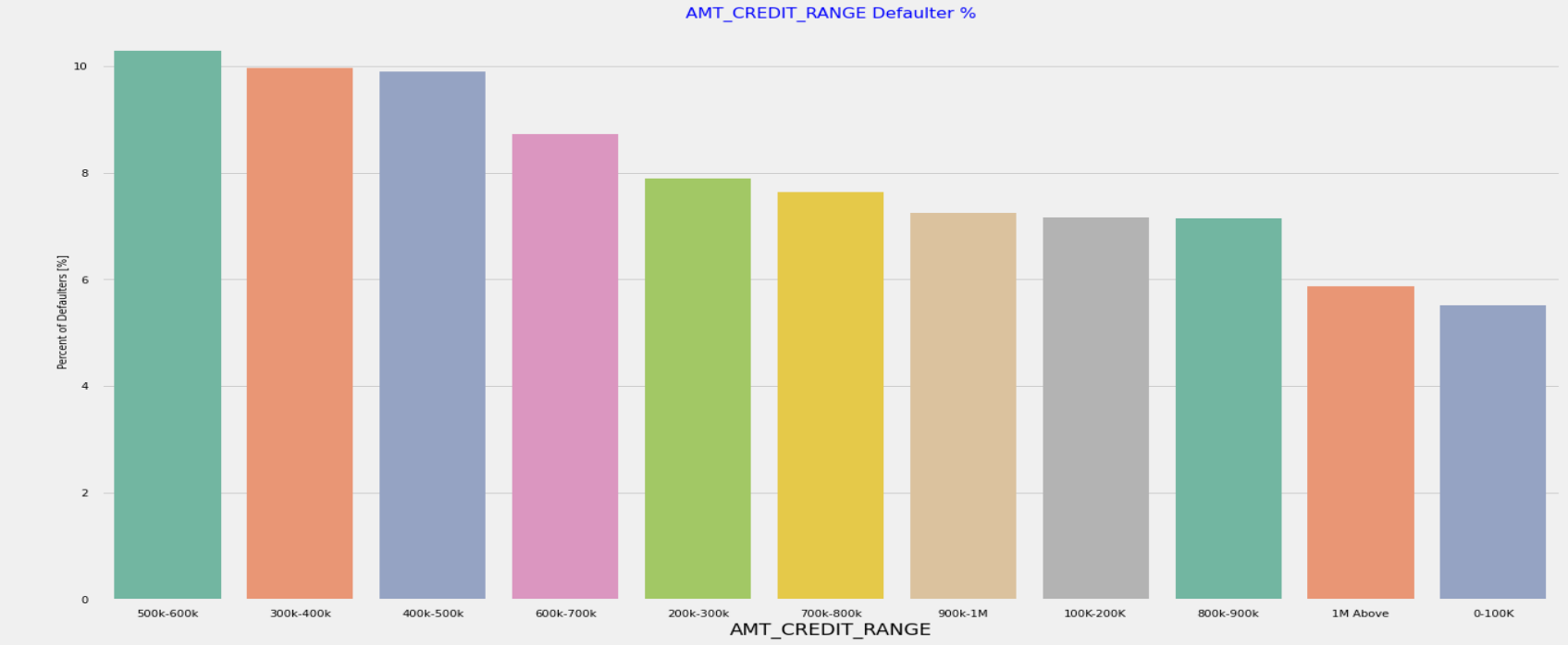


Analysing Amount_Credit based on loan repayment status

1. More than 80% of the loan provided are for amount less than 900,000

2. People who get loan for 300K-600K tend to default more than others.



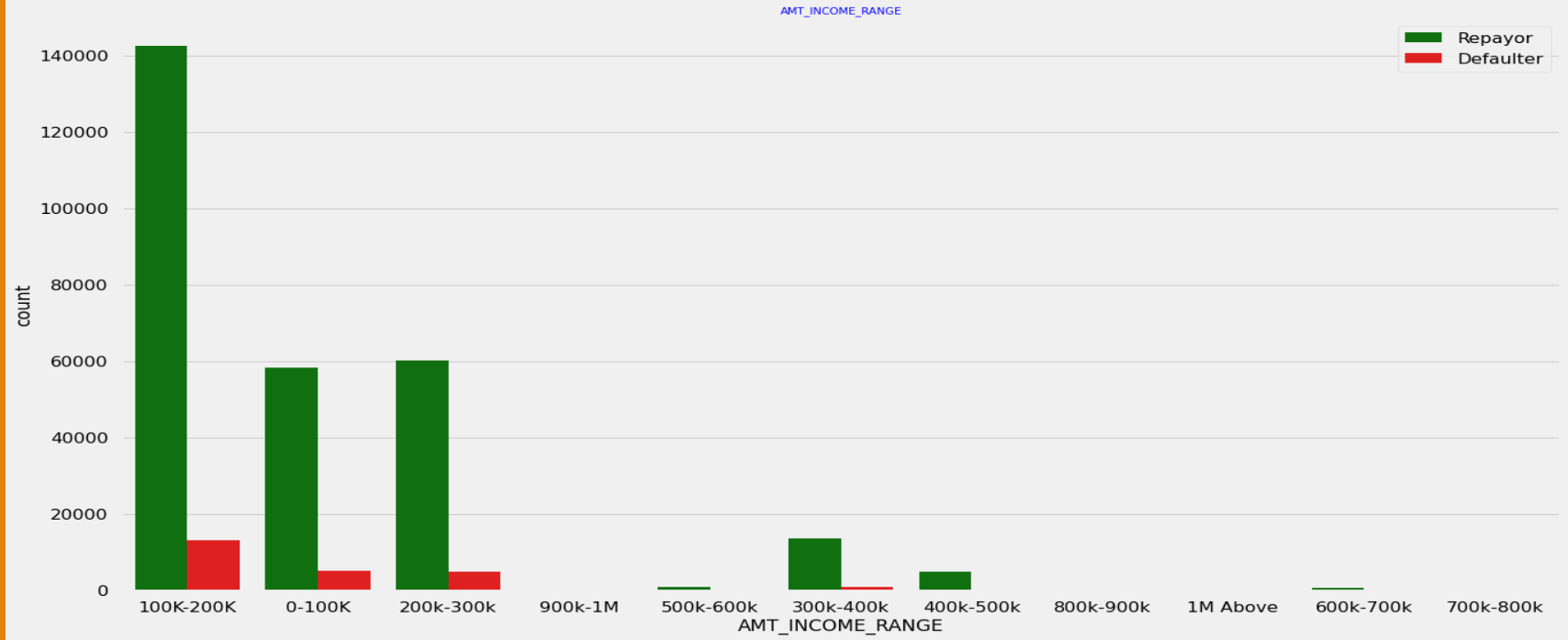


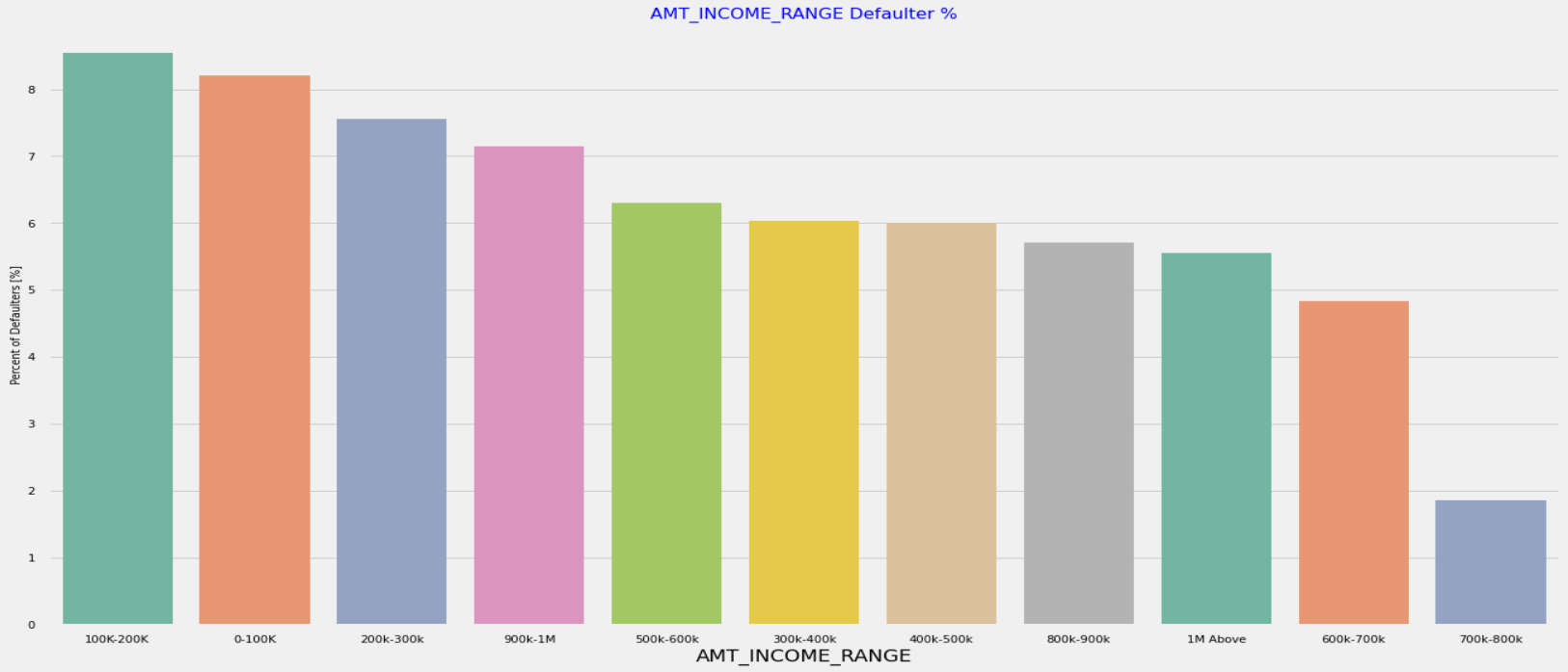
Analysing Amount_Income Range based on loan repayment status

1. 90% of the applications have Income total less than 300,000

2. Application with Income less than 300,000 has high probability of defaulting

3. Applicant with Income more than 700,000 are less likely to default



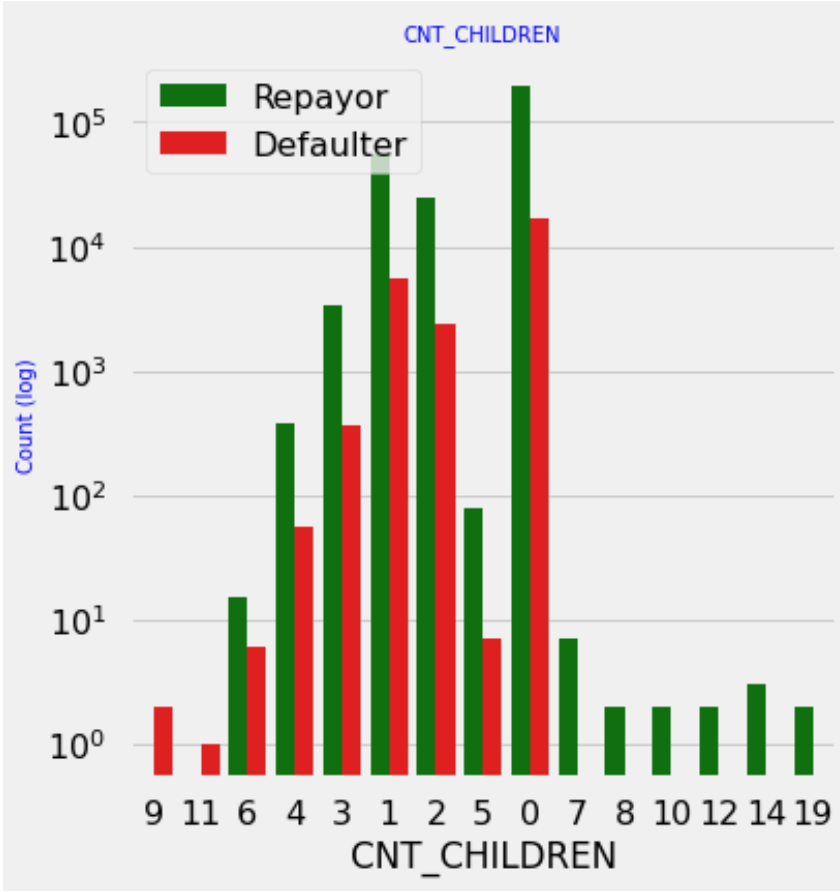


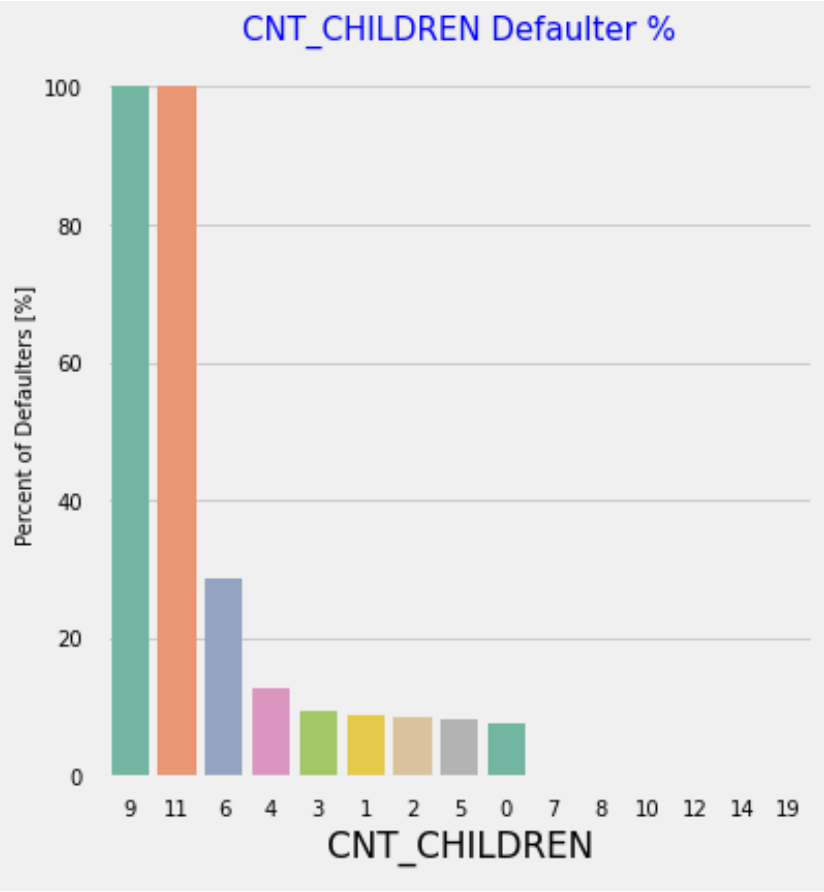
Analysing Number of children based on loan repayment status

1. Most of the applicants do not have children

2. Very few clients have more than 3 children.

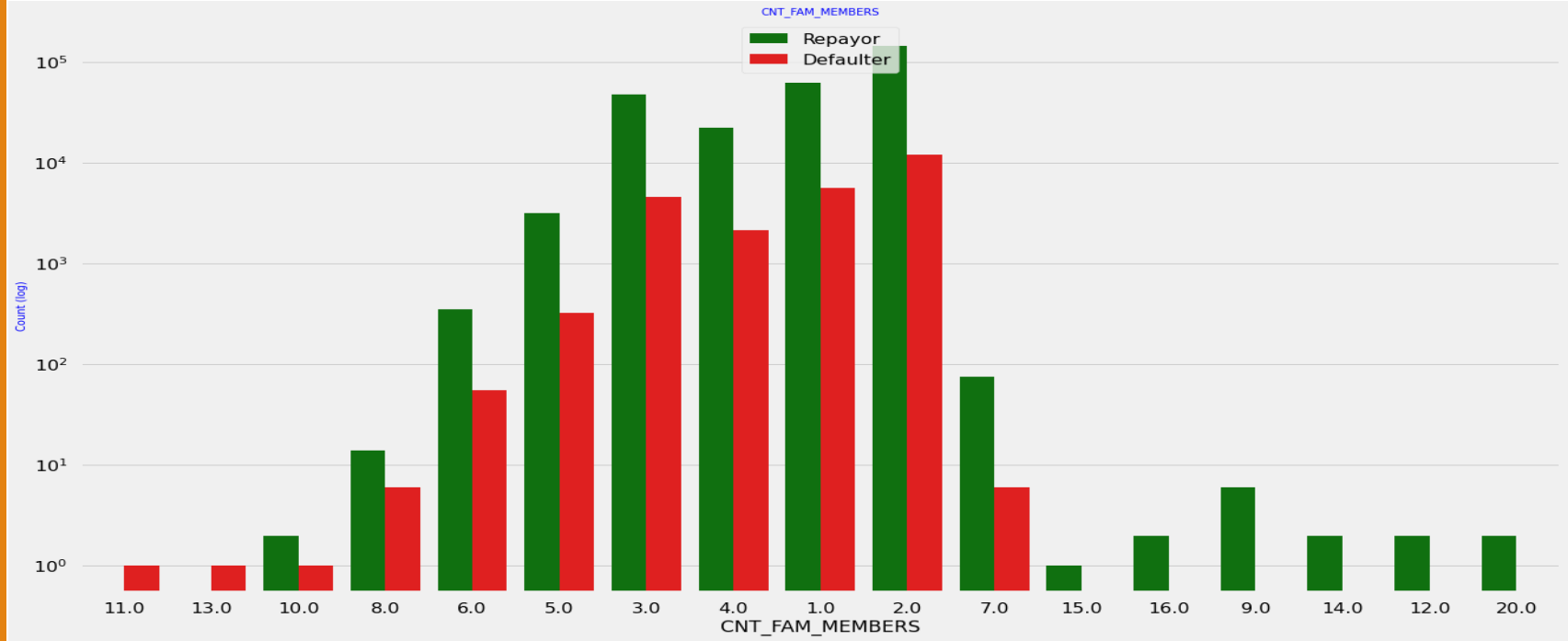
3. Client who have more than 4 children has a very high default rate with child count 9 and 11 showing 100% default rate

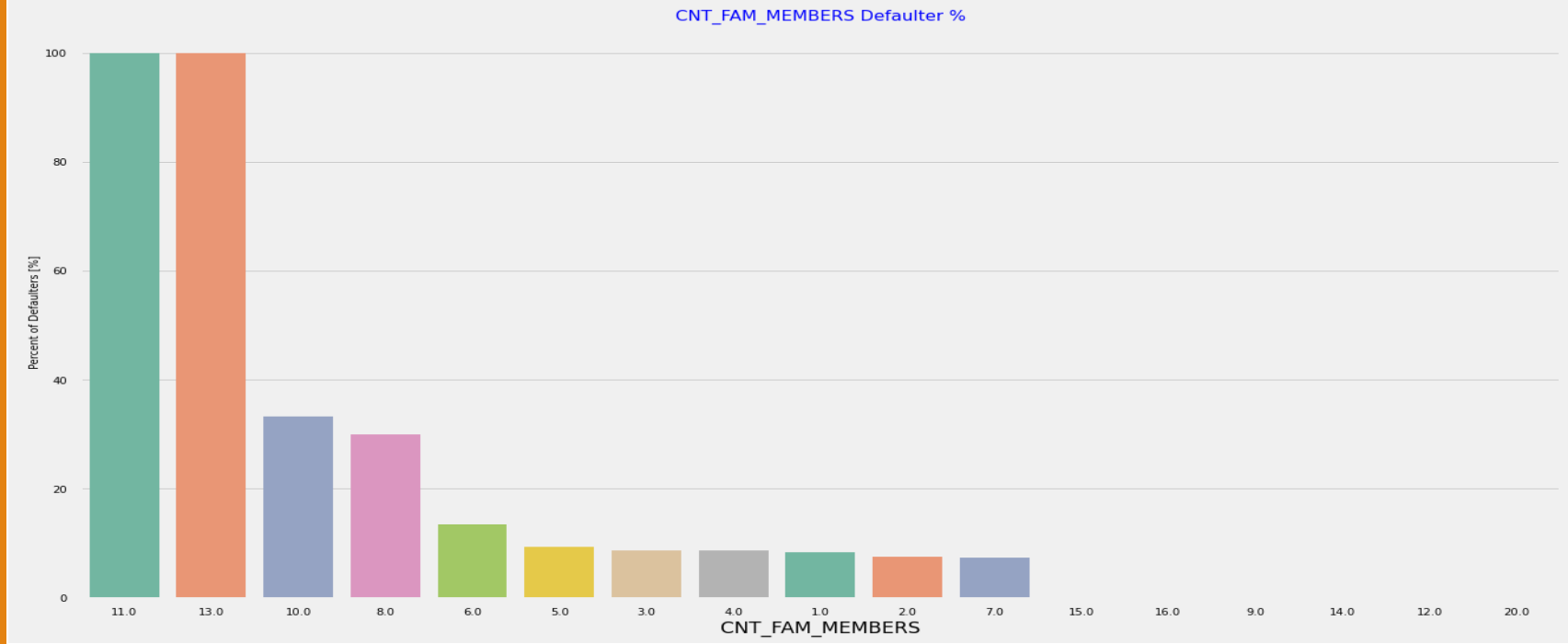




Analysing Number of family members based on loan repayment status

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

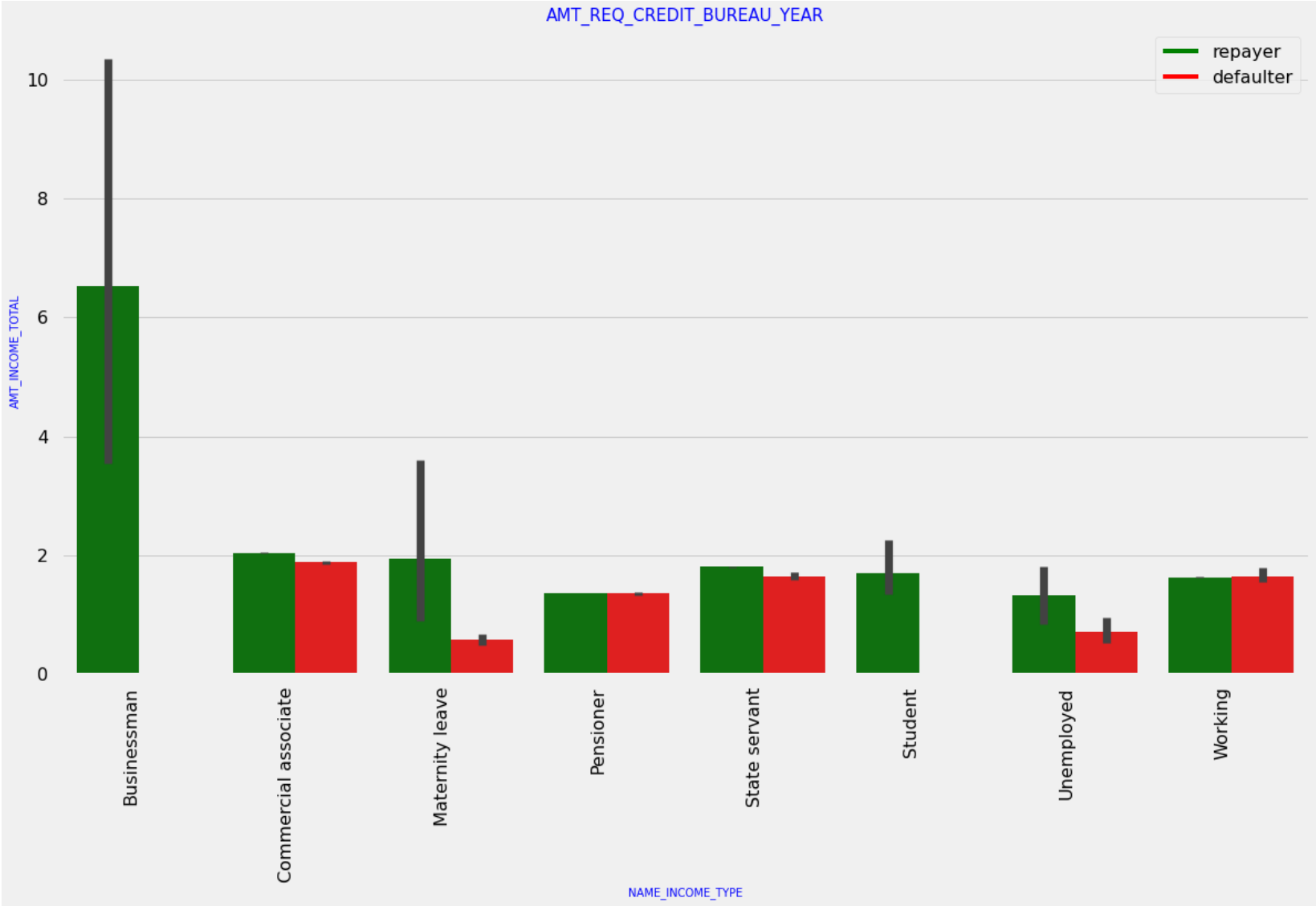




Categorical Bi/Multivariate Analysis

Income type vs Income Amount Range

It can be seen that businessmen'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

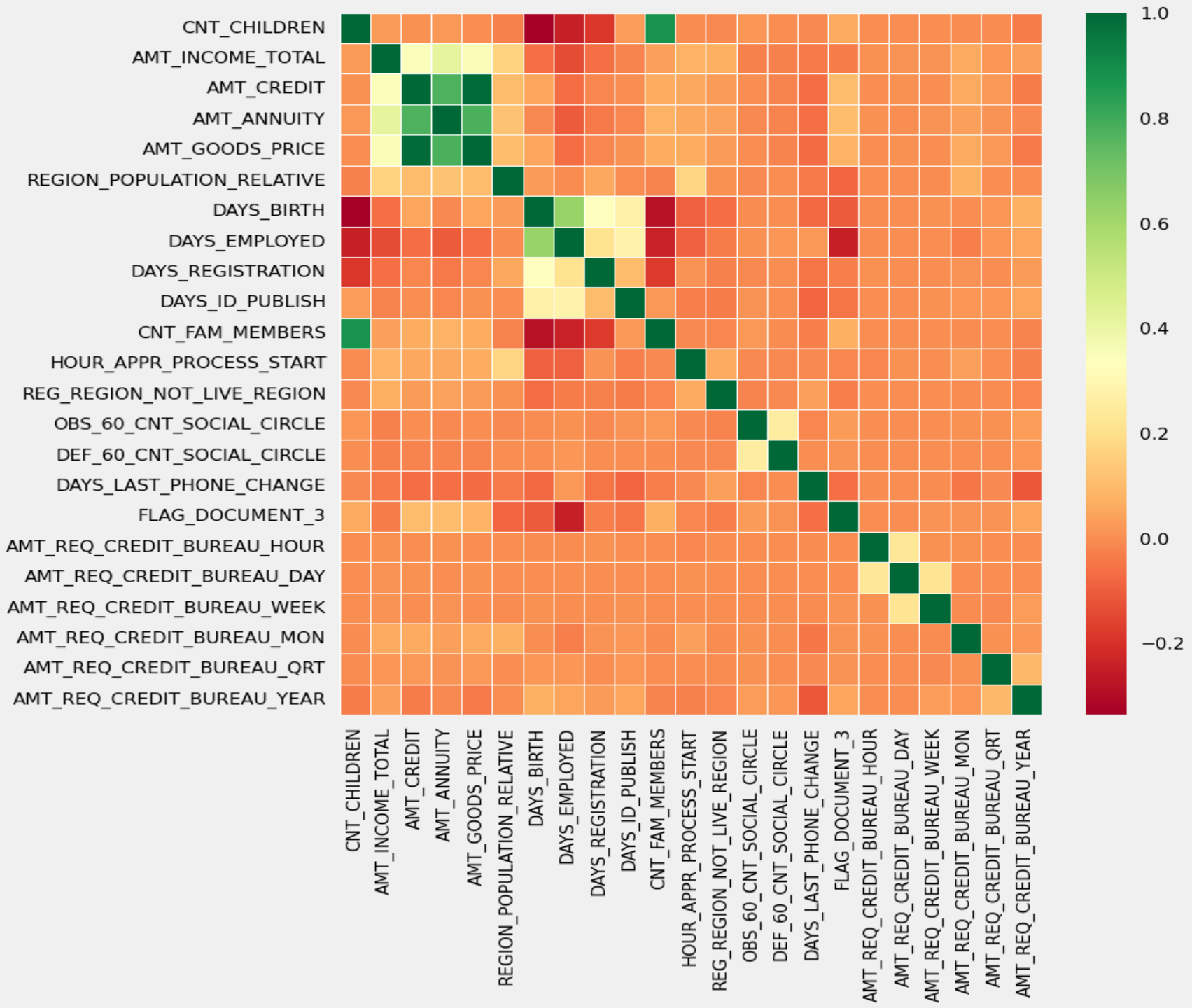


Numeric Variables Analysis

Correlation between numeric variable

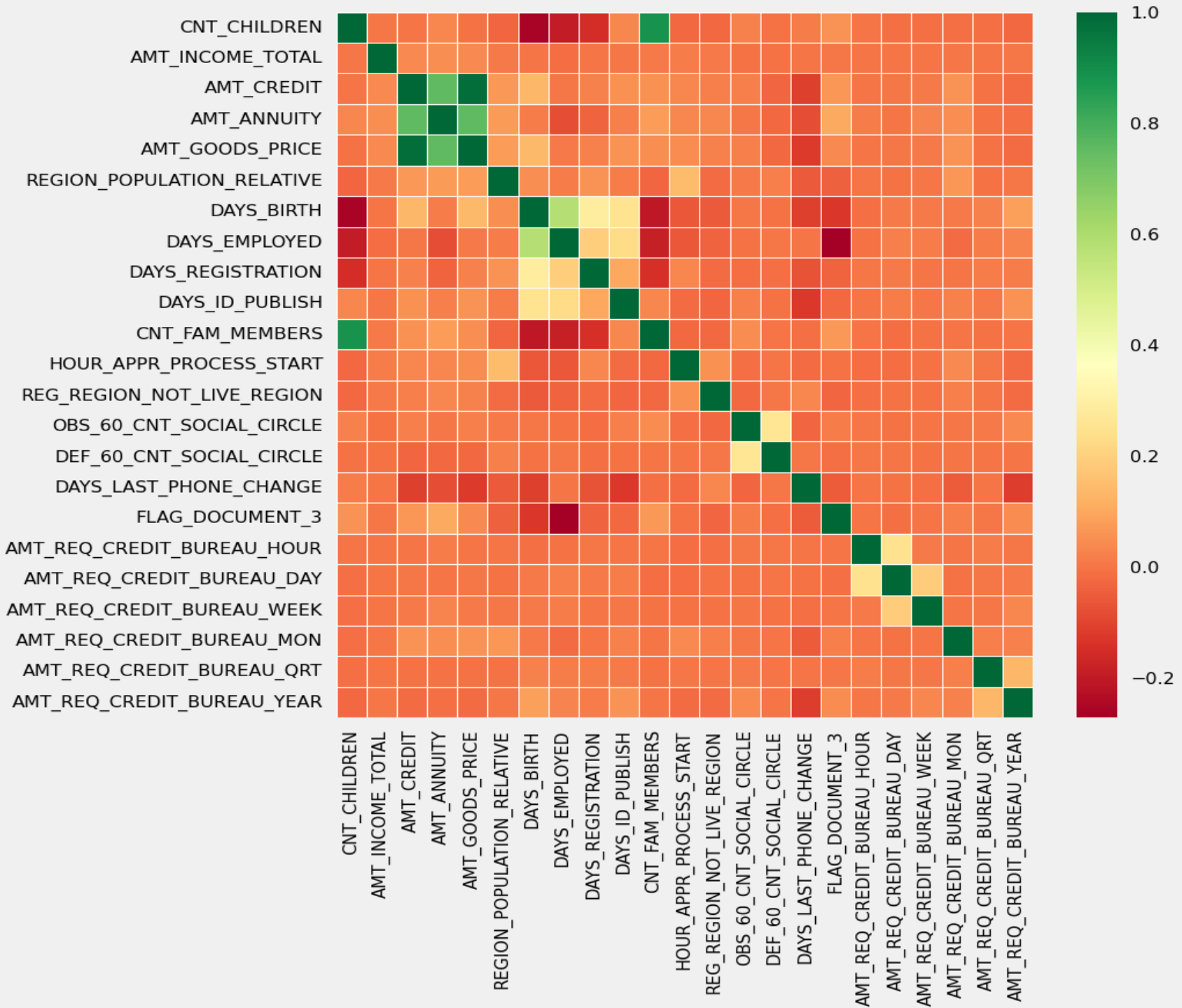
- Correlating factors amongst repayors:
Credit amount is highly correlated with amount of goods price
- loan annuity
- total income

We can also see that repayers have high correlation in number of days employed.



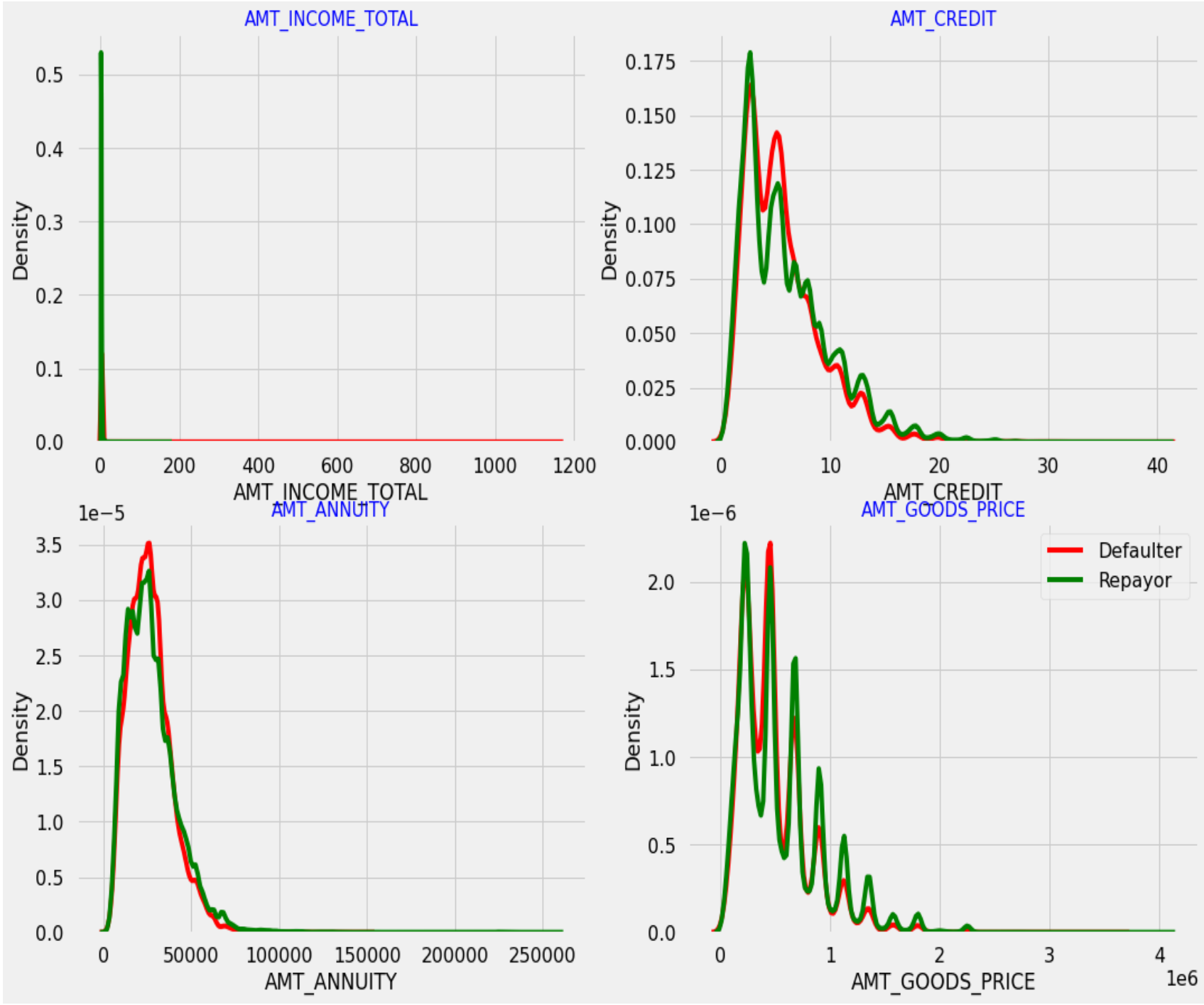
Correlation for the Defaulter data

- Credit amount is highly correlated with amount of goods price which is same as repayors.
- But the loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to repayors(0.77)
- We can also see that repayors 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 repayors.
- Days_birth and number of children correlation has reduced to 0.259 in defaulters when compared to 0.337 in repayors.
- There is a slight increase in defaulted to observed count in social circle among defaulters(0.264) when compared to repayors(0.254)



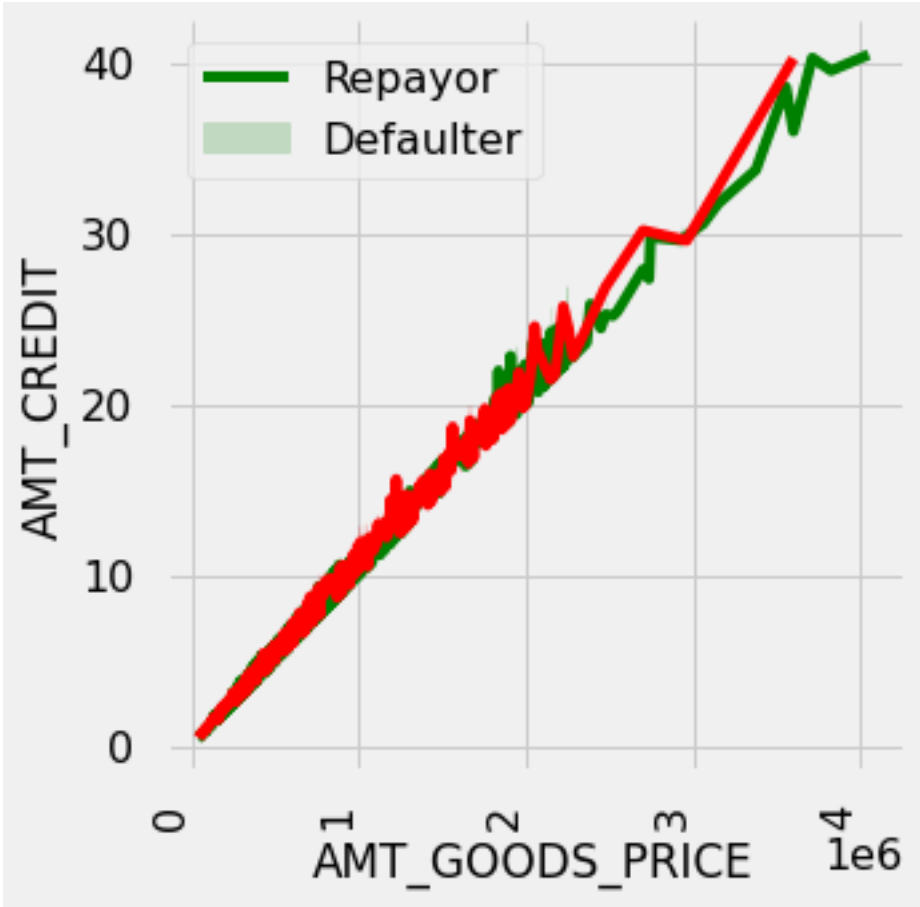
Plotting the numerical columns related to amount as distribution plot to see density

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



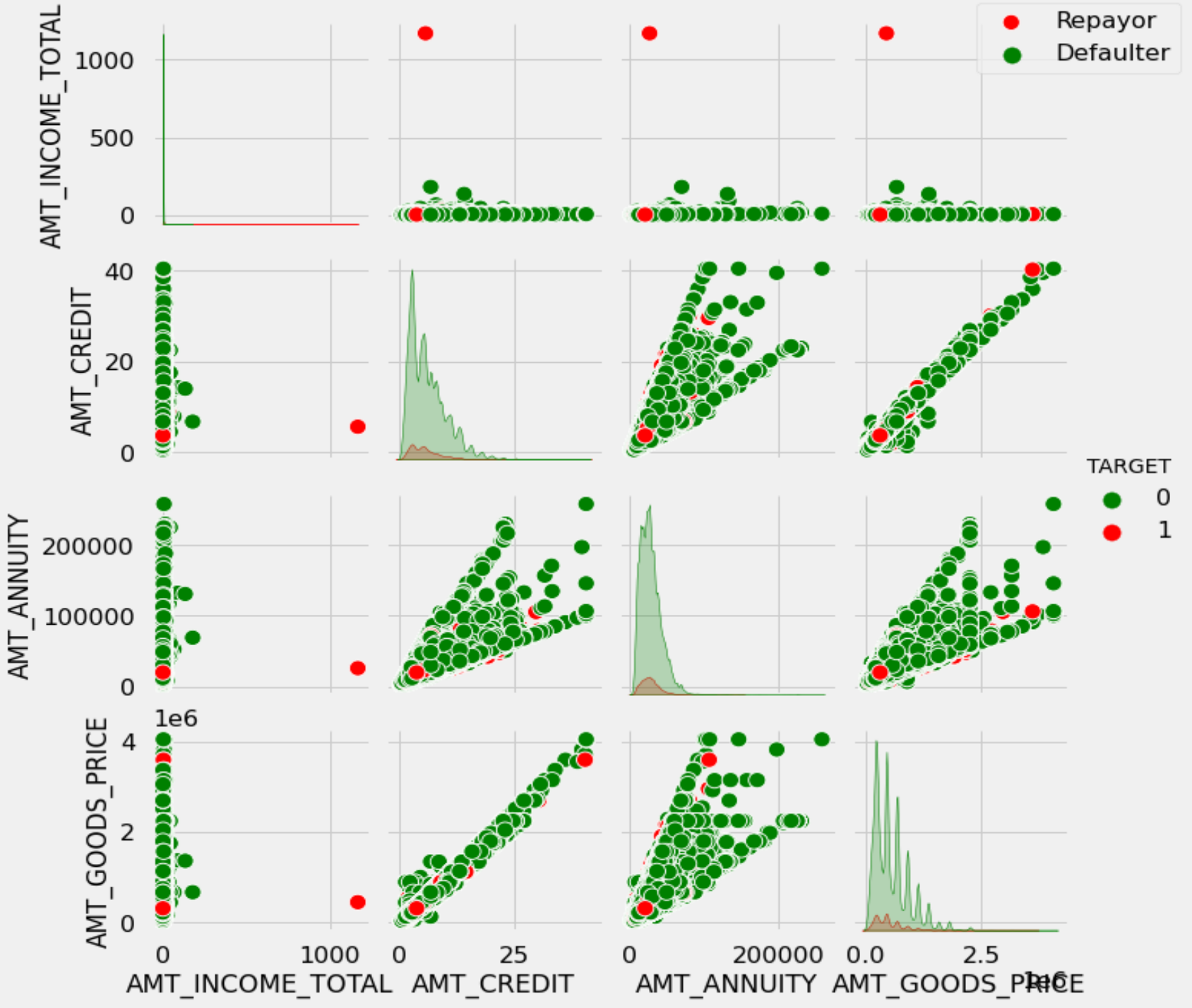
Relationship between Goods price and credit and comparing with loan repayment staus

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



Plotting pairplot between amount variable to draw reference against loan repayment status

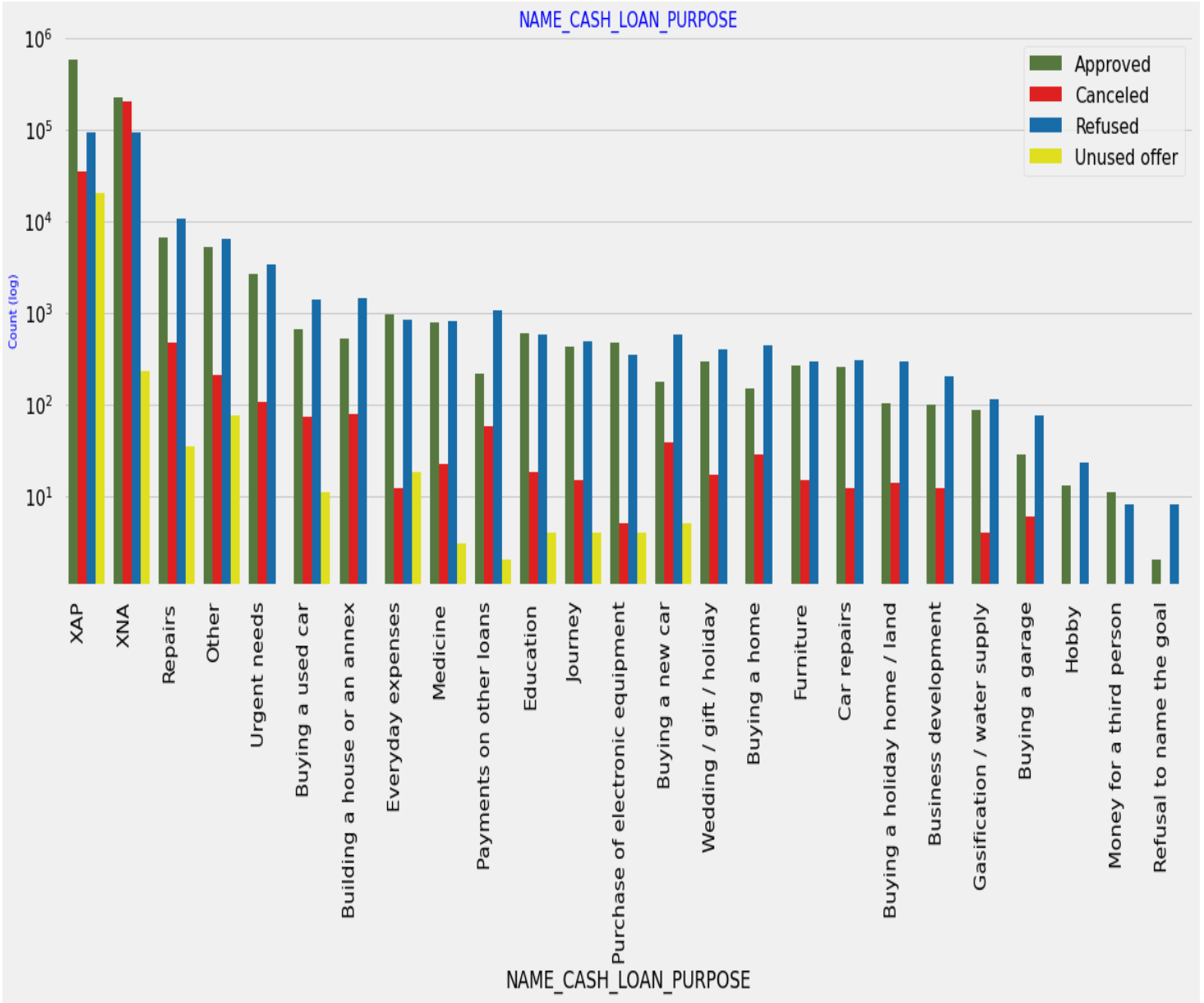
- When amt_annuity >15000 amt_goods_price> 3M, there is a lesser chance of defaulters
- AMT_CREDIT and AMT_GOODS_PRICE are highly correlated as based on the scatterplot where most of the data are consolidated in form of a line
- There are very less defaulters for AMT_CREDIT >3M
- Inferences related to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



Merged Dataframes Analysis

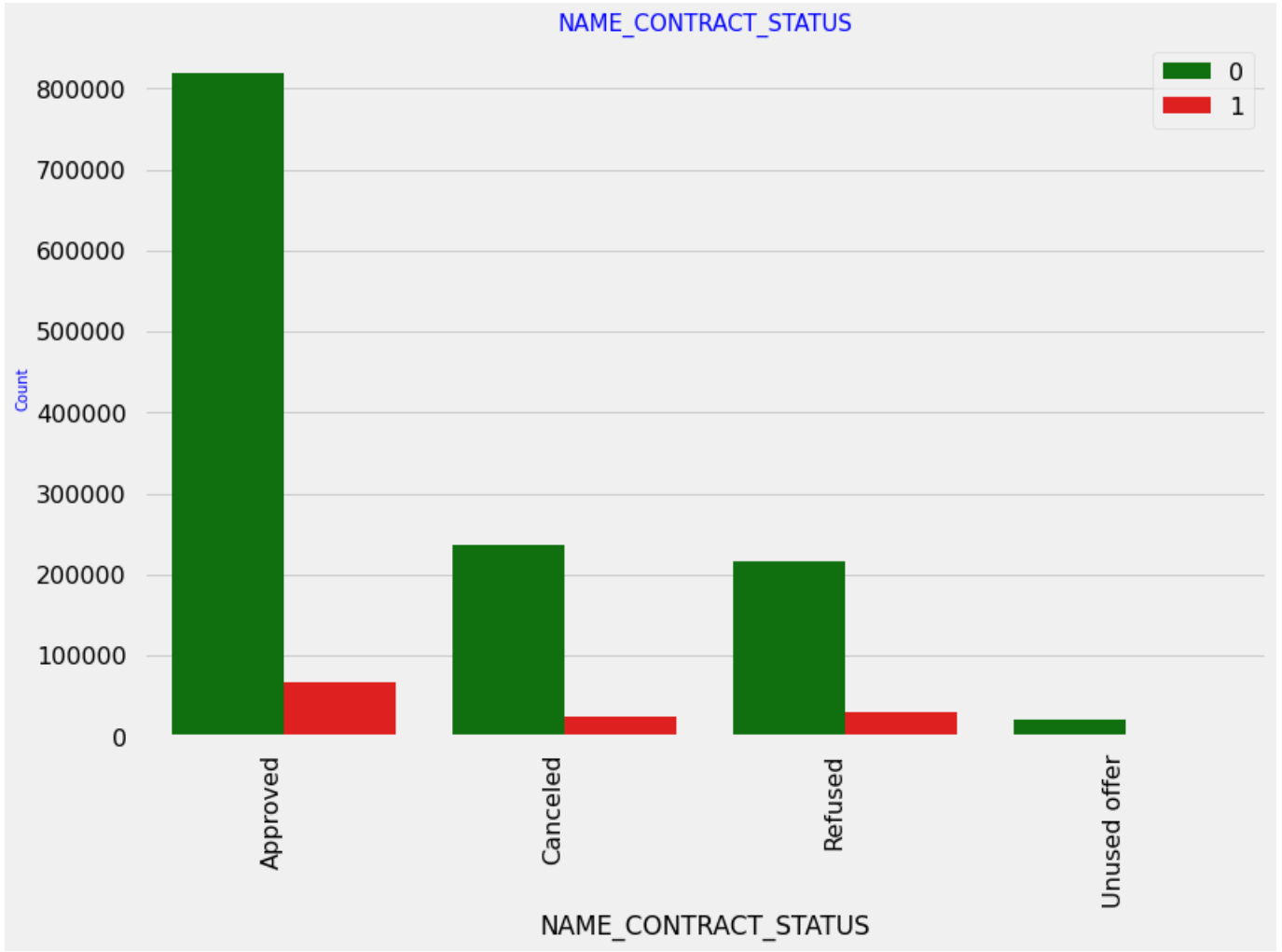
Plotting Contract Status vs purpose of the loan:

- 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 taken as high risk by bank and either they are rejected or bank offers very high loan interest rate which is not feasible by the clients, thus they refuse the loan.



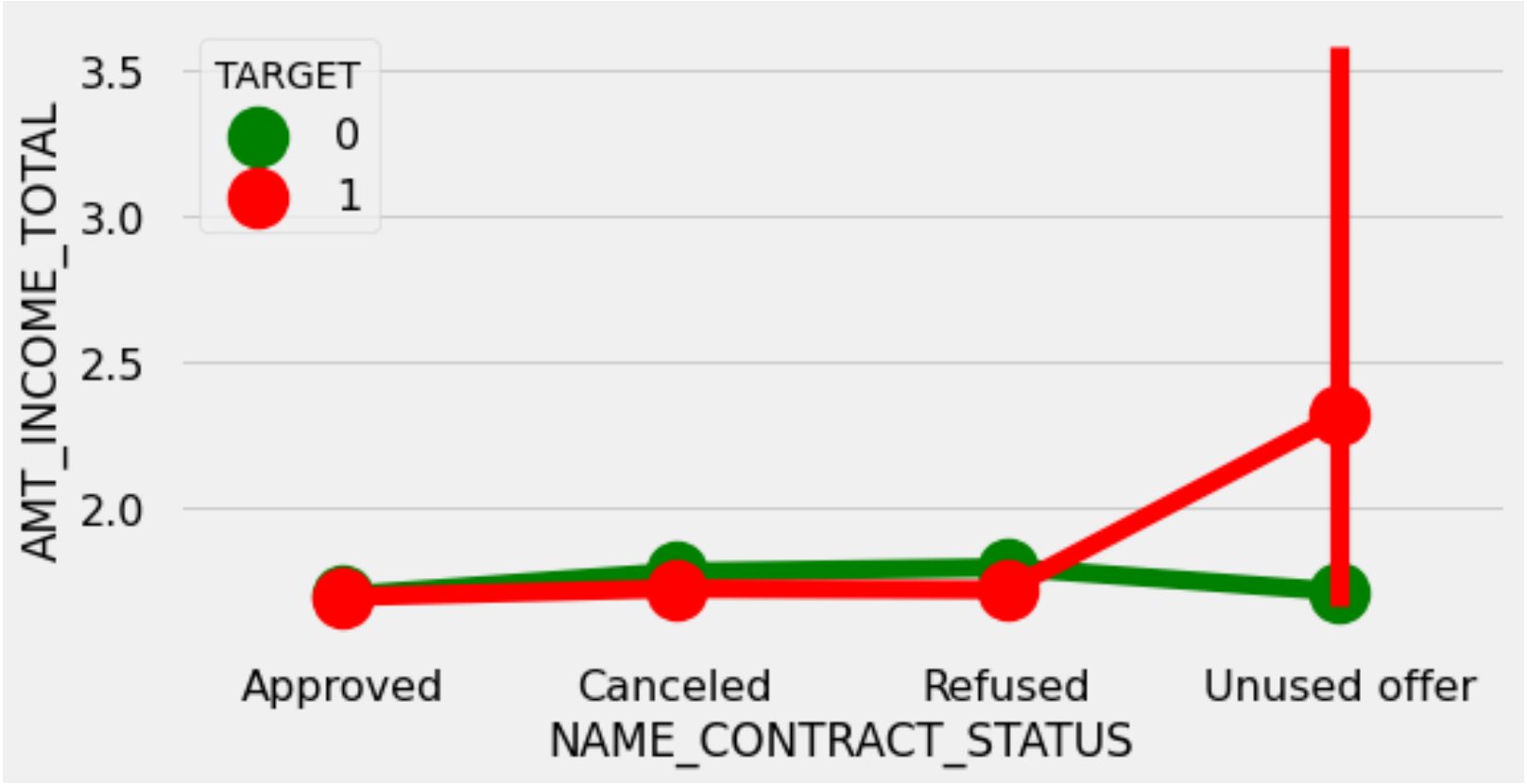
Checking the Contract Status based on loan repayment status and whether there is any business loss or financial loss

- 90% of the previously cancelled client have actually repayed the loan. Revisiting the interest rates would increase business opoortunity for these clients
- 88% of the clients who have been previously refused a loan has payed back the loan in current case.
- Refual reason should be recorded for further analysis as these clients would turn into potential repaying customer.



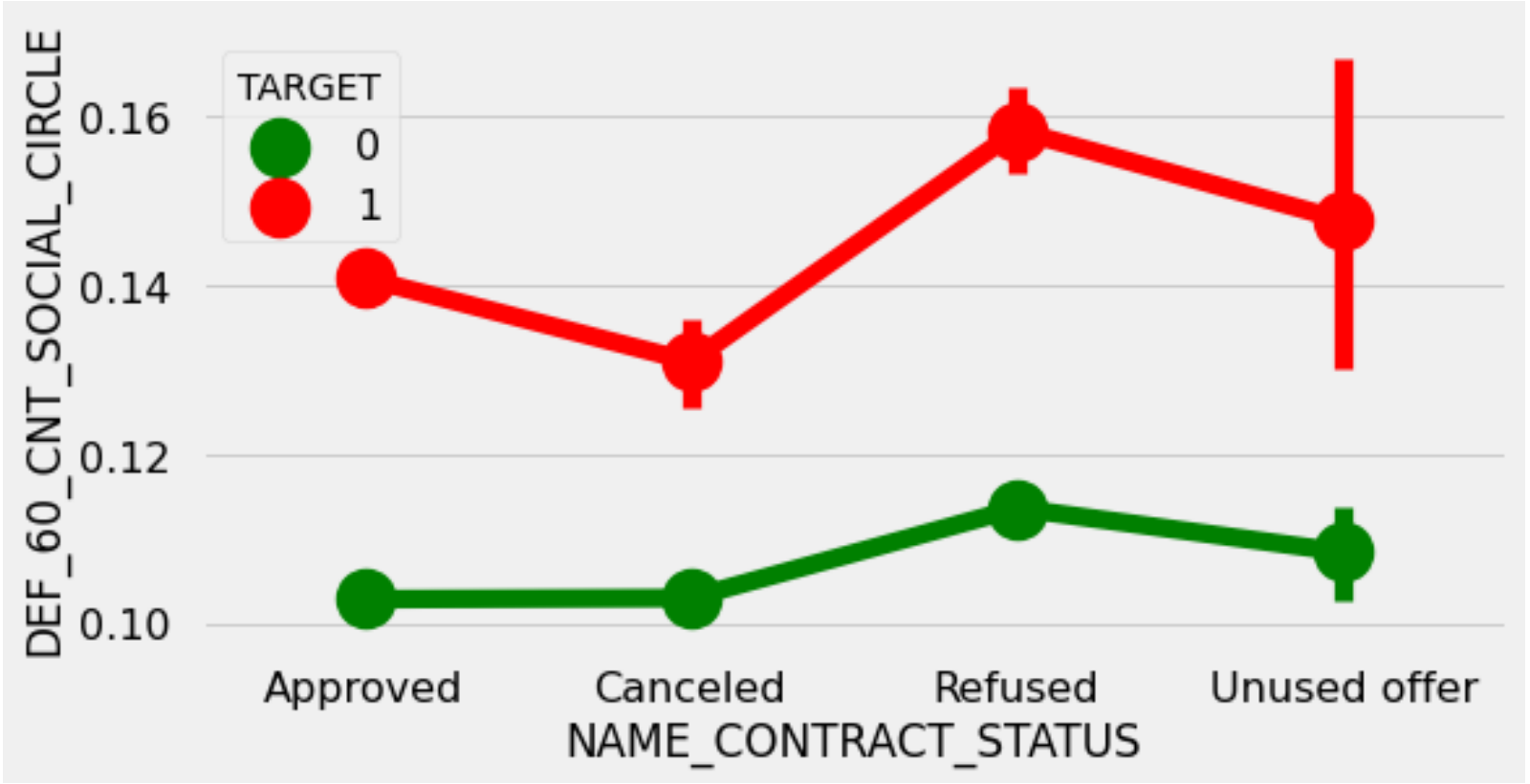
Plotting the relationship between income total and contact status

- The point plot show that the people who have not used offer earlier have defaulted even when there average income is higher than others



Plotting the relationship between people who defaulted in last 60 days being in client's social circle and contact status

- Clients who have average of 0.13 or higher DEF_60_CNT_SOCIAL_CIRCLE score tend to default more and hence client's social circle has to be analysed before providing the loan.



Conclusions

Decisive Factor whether an applicant will be Repayor:-

1. NAME_EDUCATION_TYPE: Academic degree has less defaults.
2. NAME_INCOME_TYPE: Student and Businessmen have no defaults.
3. REGION_RATING_CLIENT: RATING 1 is safer.
4. ORGANIZATION_TYPE: Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%
5. DAYS_BIRTH: People above age of 50 have low probability of defaulting
6. DAYS_EMPLOYED: Clients with 40+ year experience having less than 1% default rate
7. AMT_INCOME_TOTAL:Applicant with Income more than 700,000 are less likely to default
8. NAME_CASH_LOAN_PURPOSE: Loans bought for Hobby, Buying garage are being repayed mostly.
9. CNT_CHILDREN: People with zero to two children tend to repay the loans.

Decisive Factor whether an applicant will be Defaulter:-

1. CODE_GENDER: Men are at relatively higher default rate
2. NAME_FAMILY_STATUS : People who have civil marriage or who are single default a lot.
3. NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education
4. NAME_INCOME_TYPE: Clients who are either at Maternity leave OR Unemployed default a lot.
5. REGION_RATING_CLIENT: People who live in Rating 3 has highest defaults.
6. OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as the default rate is huge.
7. ORGANIZATION_TYPE: Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self-employed people have relative high defaulting rate, and thus should be avoided to be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting.
8. DAYS_BIRTH: Avoid young people who are in age group of 20-40 as they have higher probability of defaulting
9. DAYS_EMPLOYED: People who have less than 5 years of employment have high default rate.
10. CNT_CHILDREN & CNT_FAM_MEMBERS: Client who have children equal to or more than 9 default 100% and hence their applications are to be rejected.
11. AMT_GOODS_PRICE: When the credit amount goes beyond 3M, there is an increase in defaulters.

Decisive Factors that Indicate need to provide loans at higher rates of interest:-

The following attributes indicate that people from these category tend to default but then due to the number of people and the amount of loan, the bank could provide loan with higher interest to mitigate any default risk thus preventing business loss:

1. NAME_HOUSING_TYPE: High number of loan applications are from the category of people who live in Rented apartments & living with parents and hence offering the loan would mitigate the loss if any of those default.

2. AMT_CREDIT: People who get loan for 300-600k tend to default more than others and hence having higher interest specifically for this credit range would be ideal.

3. AMT_INCOME: Since 90% of the applications have Income total less than 300,000 and they have high probability of defaulting, they could be offered loan with higher interest compared to other income category.

4. CNT_CHILDREN & CNT_FAM_MEMBERS: Clients who have 4 to 8 children has a very high default rate and hence higher interest should be imposed on their loans.

5. NAME_CASH_LOAN_PURPOSE: Loan taken for the purpose of Repairs seems to have highest default rate. A very high number applications have been rejected by bank or refused by client in previous applications as well which has purpose as repair or other. This shows that purpose repair is taken as high risk by bank and either they are rejected, or bank offers very high loan interest rate which is not feasible by the clients, thus they refuse the loan. The same approach could be followed in future as well.

Other suggestions:

- 90% of the previously cancelled client have actually repaid the loan. Record the reason for cancellation which might help the bank to determine and negotiate terms with these repaying customers in future for increase business opportunity.
- 88% of the clients who were refused by bank for loan earlier have now turned into a repaying client. Hence documenting the reason for rejection could mitigate the business loss and these clients could be contacted for further loans.

Thank You

- Tools & Technologies Used:
  1. Python
  2. Seaborn
  3. Matplotlib
  4. Numpy
  5. Pandas
  6. Excel
- Author:-

Rishab
rishab260@hotmail.com
https://www.linkedin.com/in/rishab260/