EDA PROJECT

Risk Analytics in Banking and Financial Services

BY

KARTHIGGEYAN MAVALAVAN

PROBLEM STATEMENT

- This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicant's using EDA is the aim of this case study.
- The dataset given by the client contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:
    1) The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample.
    2) All other cases: All other cases when the payment is paid on time.
- We will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

ANALYSIS APPROACH

- Following steps done for the analysis purpose:

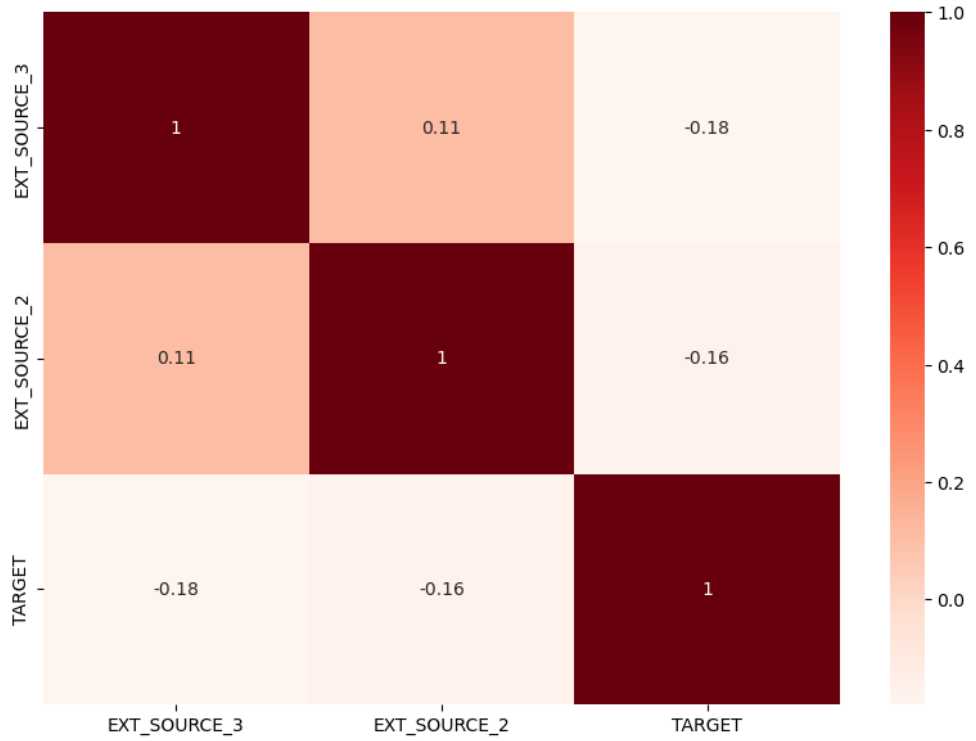
1) Imported the NumPy, pandas, matplotlib and seaborn python libraries.
2) Imported the datasets (Application_Data & Previous_Application)
3) Identification: We have identified how we will approach the data, finding missing dataset and working on it accordingly to gain the required results.
4) Outliers: Identified outliers and showed how they play if any role in our dataset.
5) Imbalance: Understanding the ratio of imbalance in our data.
6) Correlation Analysis: Finding the correlation between the variables with respect to the target variables and find the top three correlation.
7) Visualisation: Visualized the data with the help of charts and graphs.

IDENTIFYING MISSING DATA

- There are 122 columns and 307511 rows.
- There columns having negative, positive values which includes days. Hence, fixing is required.
- There are 41 columns having null values more than 50% which are related to different area sizes on apartment owned/rented by the loan applicant which will be removed.
- After after dropping 41 columns we are left with 81 columns.
- Dealing with null values more than 15%, from the columns dictionary we can conclude that only 'OCCUPATION_TYPE', 'EXT_SOURCE_3 looks relevant to TARGET column. Thus, dropping all other columns except 'OCCUPATION_TYPE','EXT_SOURCE_3.
- After dropping null_col_15, we have left with 73 columns.

ANALYZE & REMOVING UNNECESSARY COLUMNS

ɔrrelation between EXT_SOURCE_3, EXT_SOURCE_2, TARGET



- There seems to be no linear correlation and from columns description we decided to remove these columns.
- Also, we are aware correlation doesn't cause causation.
- Now we are left with 71 columns.

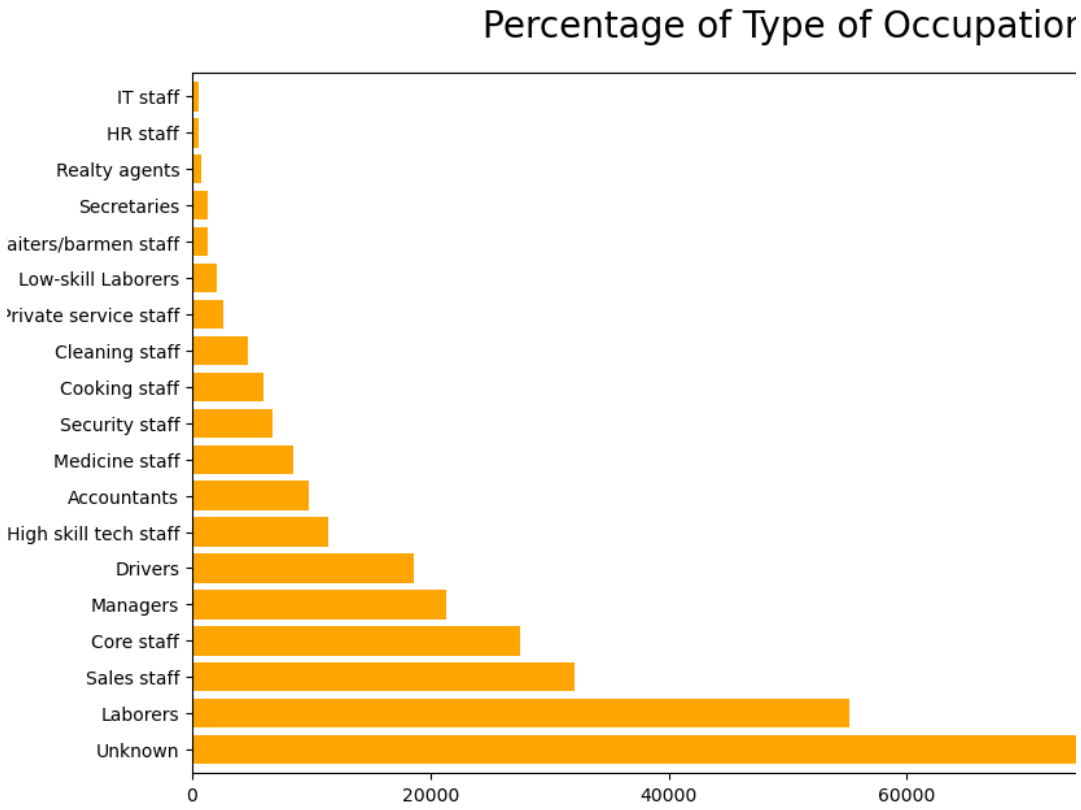
CHECK COLUMNS WITH FLAGS AND THEIR RELATIONSHIP WITH TARGET COLUMNS TO REMOVE IRRELEVANT ONES

- Columns (FLAG_OWN_REALTY, FLAG_MOBIL ,FLAG_EMP_PHONE, FLAG_CONT_MOBILE, FLAG_DOCUMENT_3) have more repayers than defaulter and from these keeping FLAG_DOCUMENT_3,FLAG_OWN_REALTY, FLAG_MOBIL more sense thus we can include these columns and remove all other FLAG columns for further analysis.
- After removing unnecessary, irrelevant and missing columns. We are left with 46 columns.

IMPUTING VALUES

- Now that we have removed all the unnecessary columns, we will proceed with imputing values for relevant missing columns wherever required.
- Now we have only 7 columns which have missing values more than 1%. Thus, we will only impute them for further analysis and such columns are: OCCUPATION_TYPE, AMT_REQ_CREDIT_BUREAU_YEAR, AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_HOUR

IMPUTING FOR "OCCUPATION_TYPE" COLUMN



- It looks like this column is categorical and have missing values of 31.35%. To fix this we will impute another category as "Unknown" for the missing values.
- Highest percentage of values belongs to Unknown group and Second belongs to Laborers

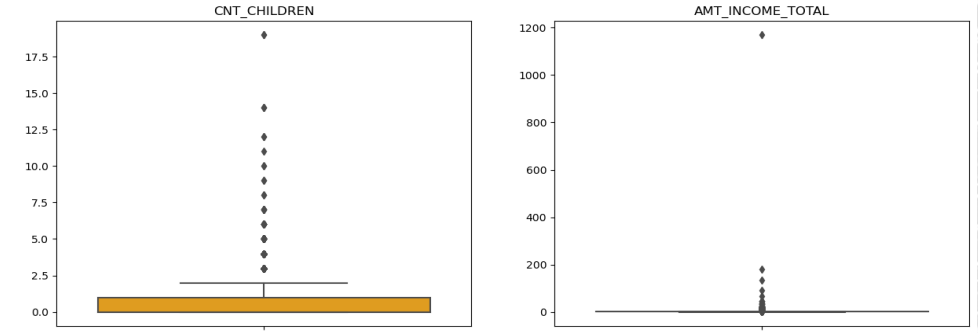
```
NAME_TYPE_SUITE              0.42
DEF_60_CNT_SOCIAL_CIRCLE     0.33
OBS_60_CNT_SOCIAL_CIRCLE     0.33
DEF_30_CNT_SOCIAL_CIRCLE     0.33
OBS_30_CNT_SOCIAL_CIRCLE     0.33
AMT_GOODS_PRICE              0.09
AMT_ANNUITY                  0.00
CNT_FAM_MEMBERS              0.00
DAYS_LAST_PHONE_CHANGE       0.00
ORGANIZATION_TYPE            0.00
dtype: float64
```

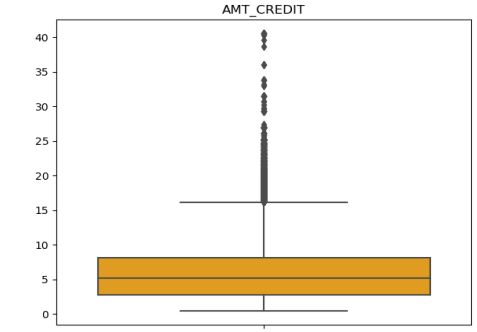
IMPUTING FOR OTHER 6 COLUMNS

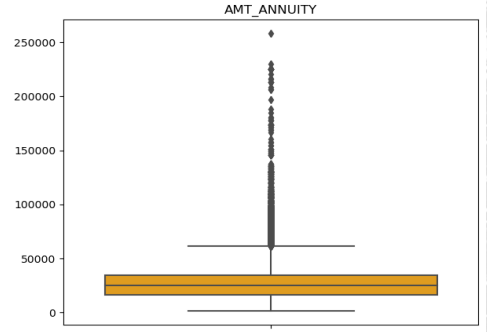
- Imputing #AMT_REQ_CREDIT_BUREAU_YEAR", "AMT_REQ_CREDIT_BUREAU_QRT","AMT_REQ_CREDIT_BUREAU_MON", "AMT_REQ_CREDIT_BUREAU_WEEK","AMT_REQ_CREDIT_BUREAU_DAY", "AMT_REQ_CREDIT_BUREAU_HOUR".
- These columns represent number of enquiries made for the customer (which should be discrete and not continuous).
- From describe results, all values are numerical and can conclude that for imputing missing values, we should not use mean as it is in decimal form, hence for imputing purpose we will use median for all these columns.
- Still there some missing value columns but we will not impute them as the missing value count very less.

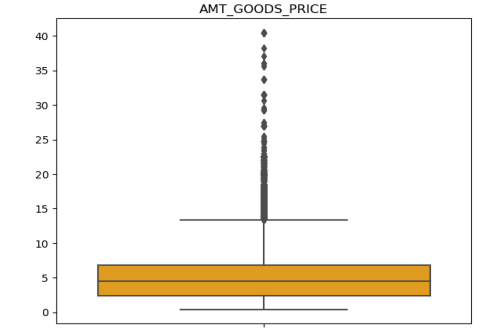
STANDARDISING VALUES

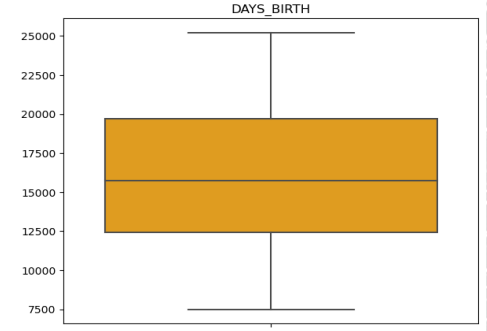
- Columns AMT_INCOME_TOTAL, AMT_CREDIT, AMT_GOODS_PRICE have very high values, thus will make these numerical columns in categorical columns for better understanding.
- Columns DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, DAYS_LAST_PHONE_CHANGE which counts days have negative values. Thus, will correct those values.
- Convert DAYS_BIRTH to AGE in years , DAYS_EMPLOYED to YEARS EMPLOYED.
- Binning Numerical Columns to create a categorical column.

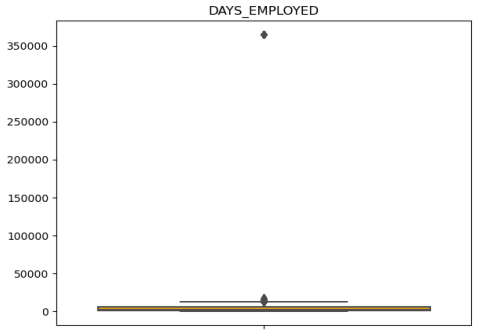


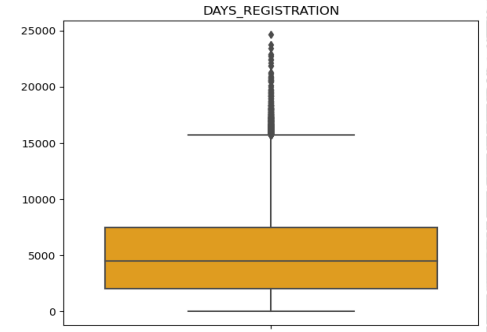












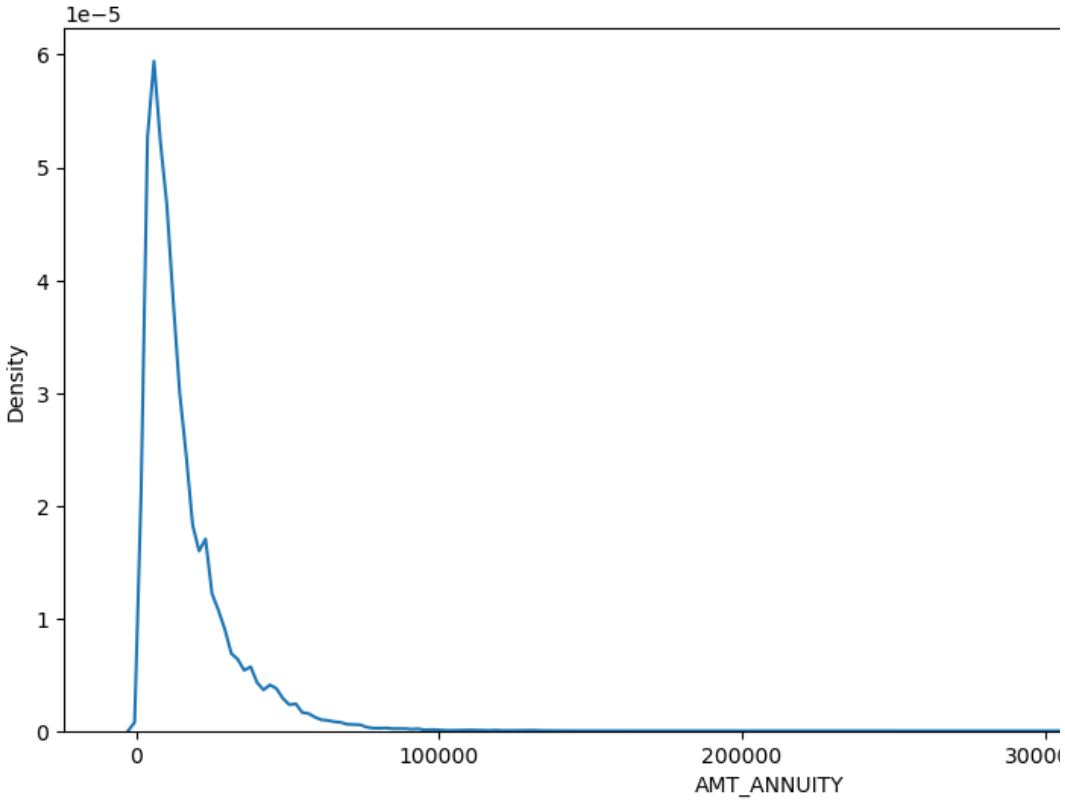
IDENTIFYING OUTLIERS

- In current application data:
- #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 must be incorrect entry.

DATASET 2 - "PREVIOUS_APPLICATION.CSV"

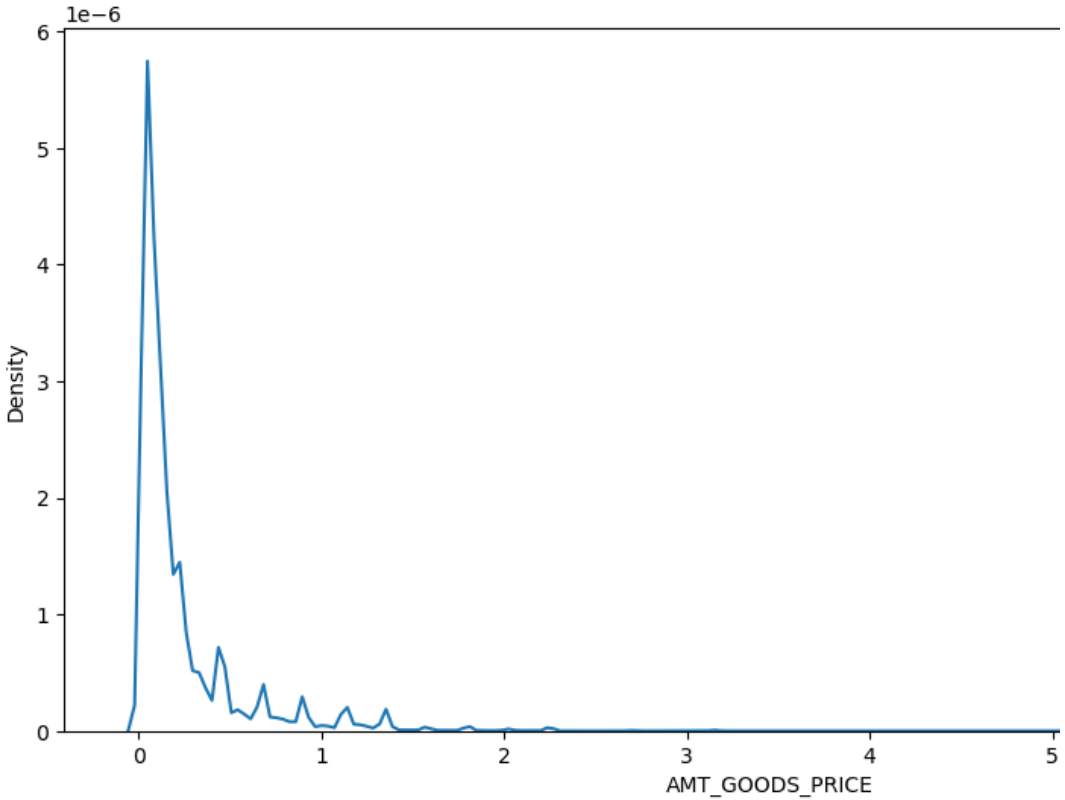
- There are 37 columns having various data types like object, int, float and 1670214 rows.
- There columns having negative, positive values which includes days. Fixing is required.
- There are missing values in columns 'DAYS_FIRST_DUE', 'DAYS_TERMINATION', 'DAYS_FIRST_DRAWING','DAYS_LAST_DUE_1ST_VERSION', 'DAYS_LAST_DUE' and these columns count days thus will keeping null values as they are.
- Almost 35% loan applicants have applied for a new loan within 1 year of previous loan decision.

DISTRIBUTION PLOT

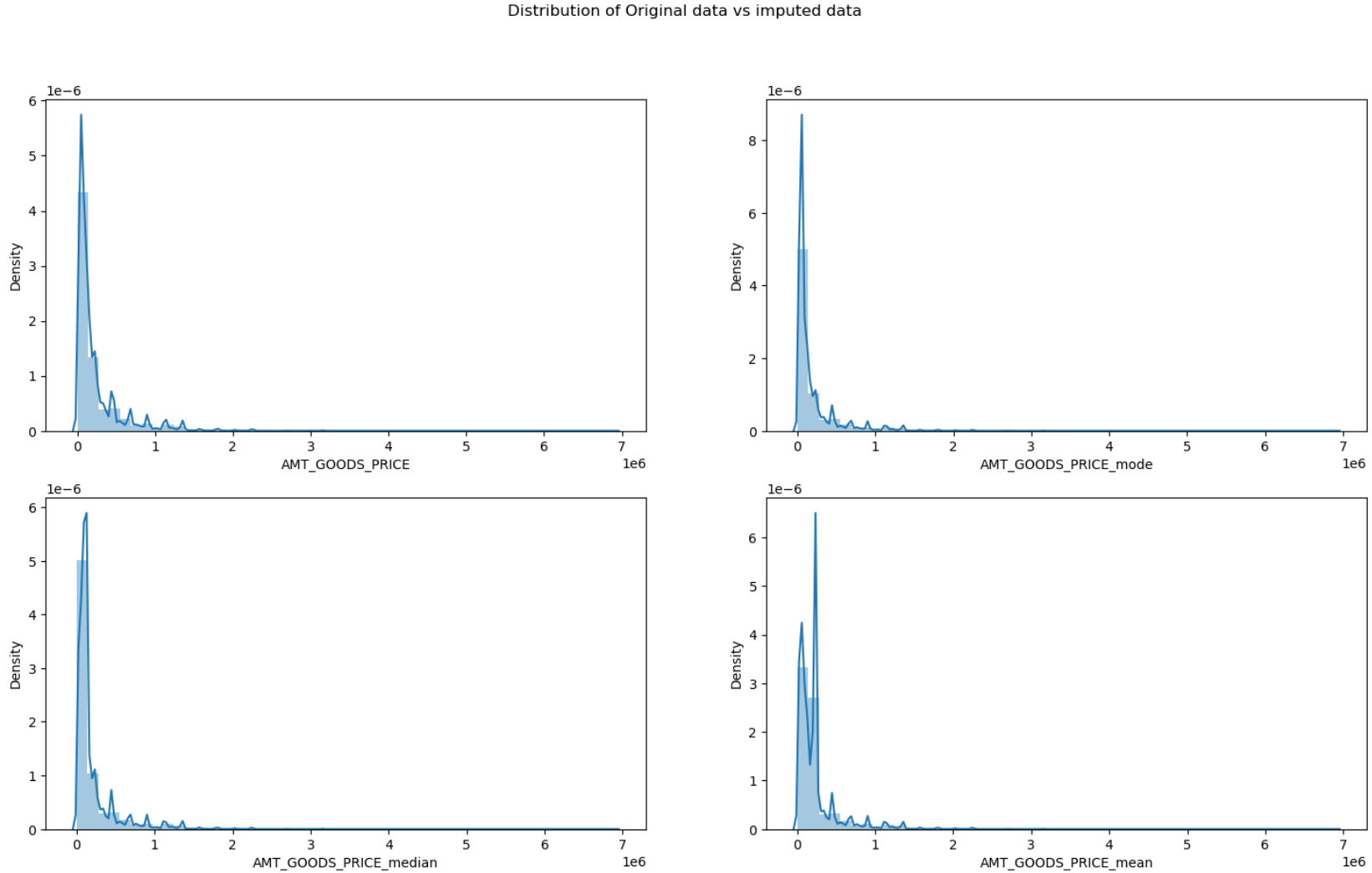


- Now dealing with continuous variables "AMT_ANNUITY", "AMT_GOODS_PRICE"
- To impute null values in continuous variables, we plotted the distribution of the columns and used:
- median if the distribution is skewed
- mode if the distribution pattern is preserved.
- #Insight:
- There is a single peak at 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 MISSING VALUES WITH MEDIAN

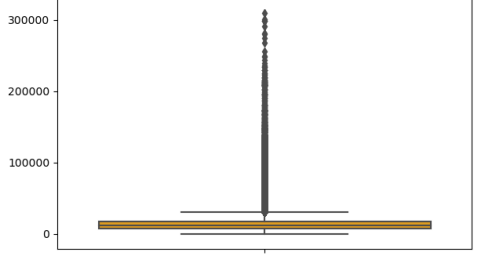


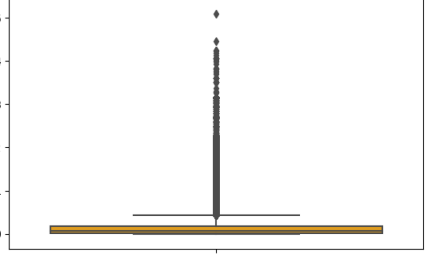
- Plotting kde plot for "AMT_GOODS_PRICE" to understand the distribution.
- There are several peaks along the distribution. Let's impute using the mode, mean and median and see if the distribution is still about the same.

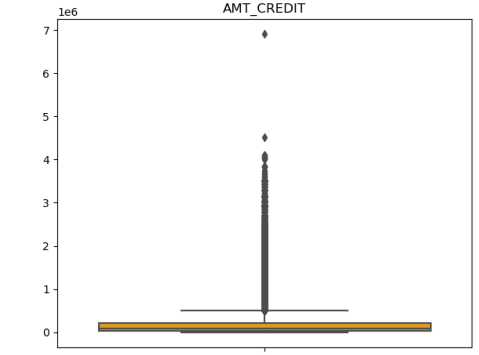


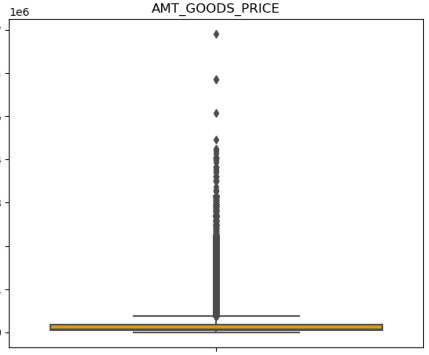
CREATING NEW DATAFRAME FOR "AMT_GOODS_PRICE" WITH COLUMNS IMPUTED WITH MODE, MEDIAN AND MEAN

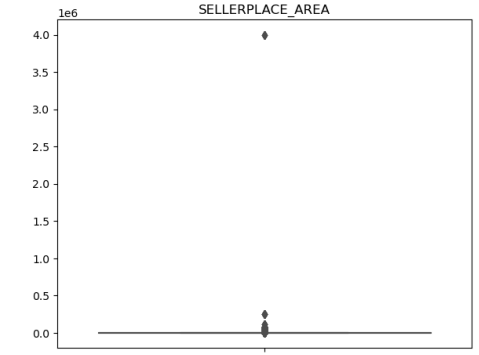
- The original distribution is closer with the distribution of data imputed with mode in this case, thus will impute mode for missing values.
- Imputing CNT_PAYMENT with 0 as the NAME_CONTRACT_STATUS for these indicate that most of these loans were not started.

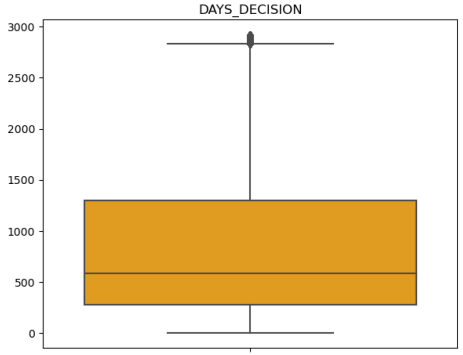


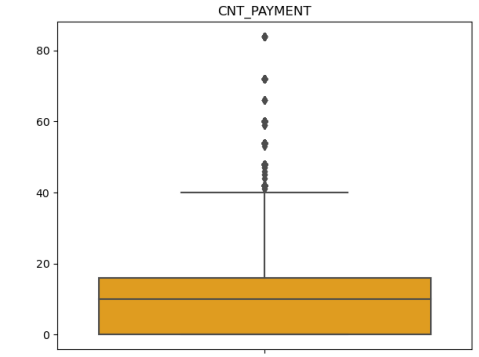










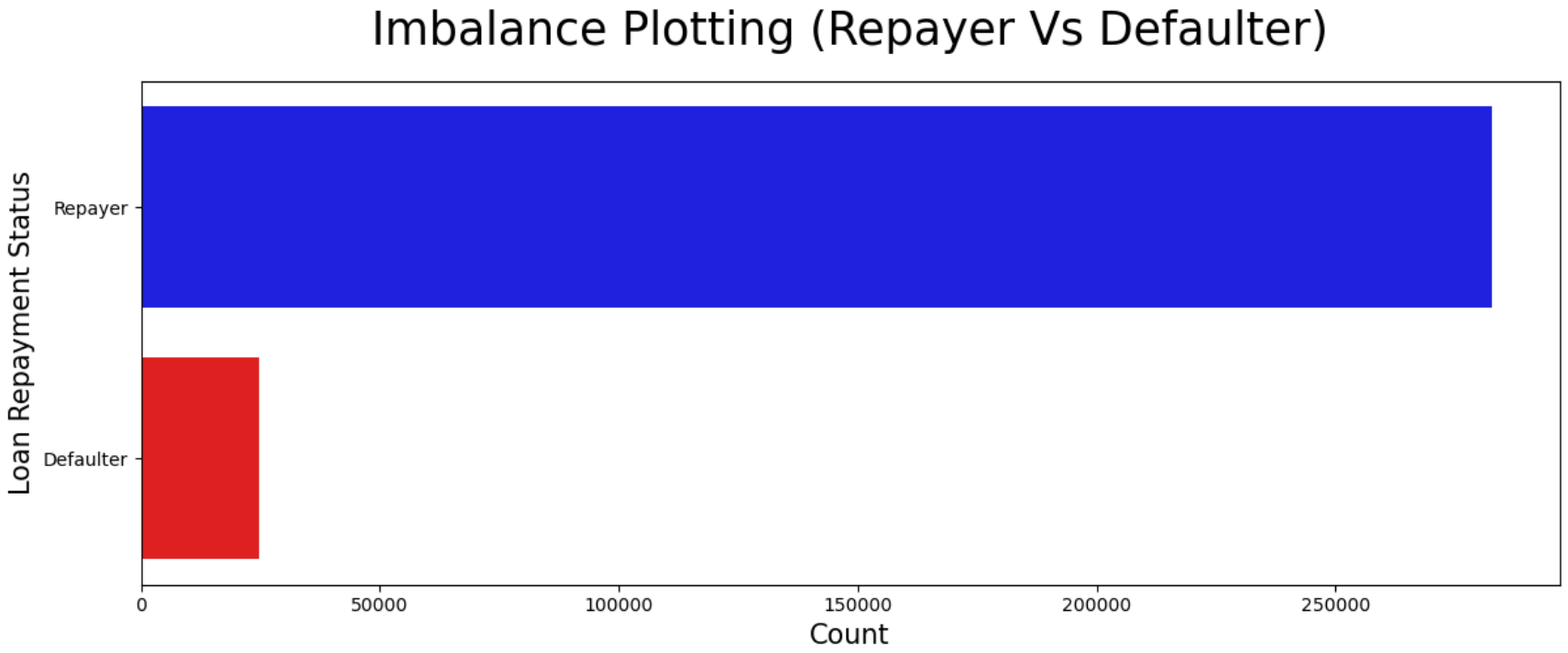


FINDING OUTLIERS

- From describe, we could find all the columns those wo have high difference between max and 75 percentile and the ones which makes no sense having max value to be so high are captured below.
- It can be seen that in previous application data:
- #AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.
- #CNT_PAYMENT has few outlier values.
- #DAYS_DECISION has little number of outliers indicating that these previous applications decisions were taken long back.

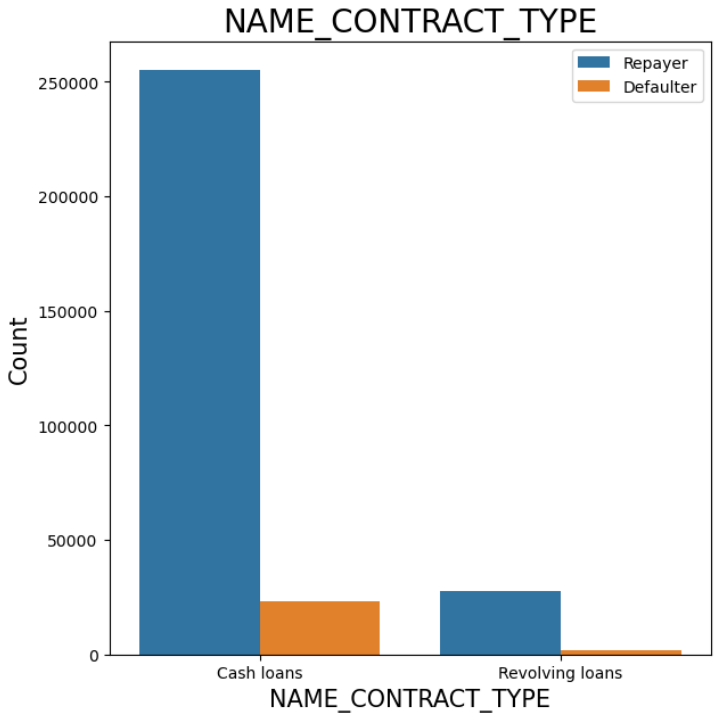
THE DATA ANALYSIS FLOW

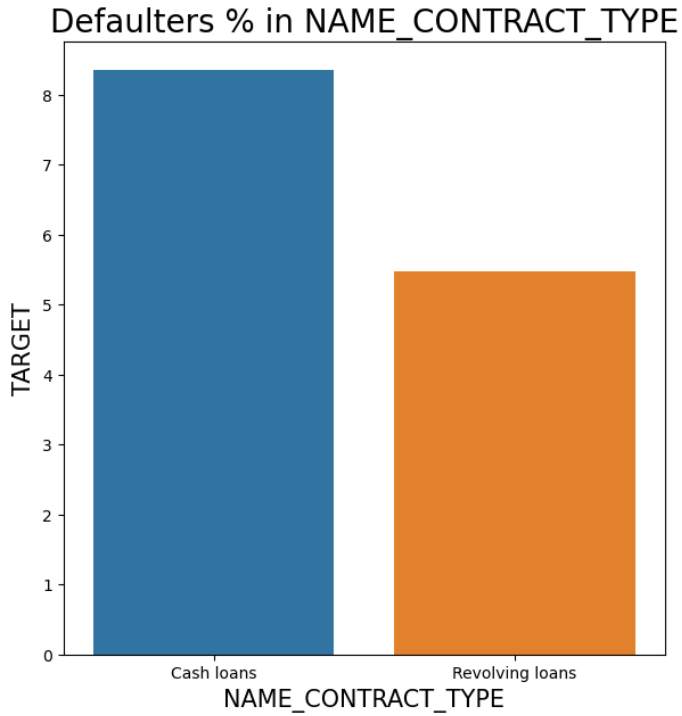
- Imbalance in Data
- Categorical Data Analysis
- Categorical segmented Univariate Analysis
- Categorical Bi/Multivariate analysis
- Numeric Data Analysis
- Bi-furcation of database based on TARGET data
- Correlation Matrix
- Numerical segmented Univariate Analysis
- Numerical Bi/Multivariate analysis



IMBALANCE DATA

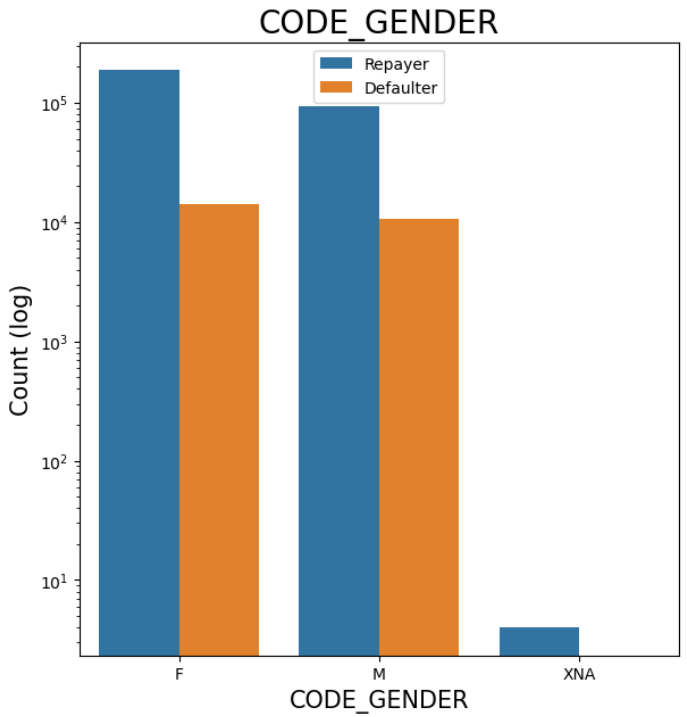
- Repayer Percentage is 91.93% Defaulter Percentage is 8.07% Imbalance Ratio with respect to Repayer and Defaulter is given: 11.39/1 (approx)

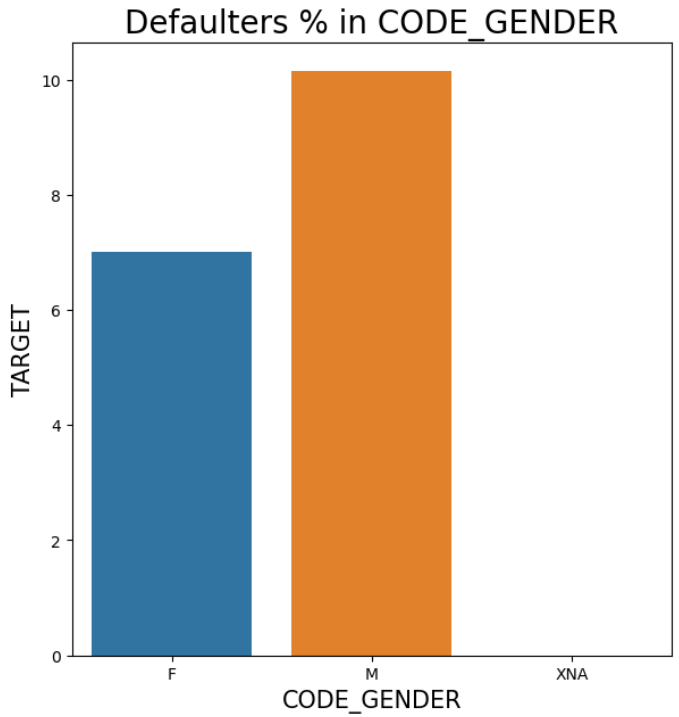




SEGMENTED UNIVARIATE ANALYSIS

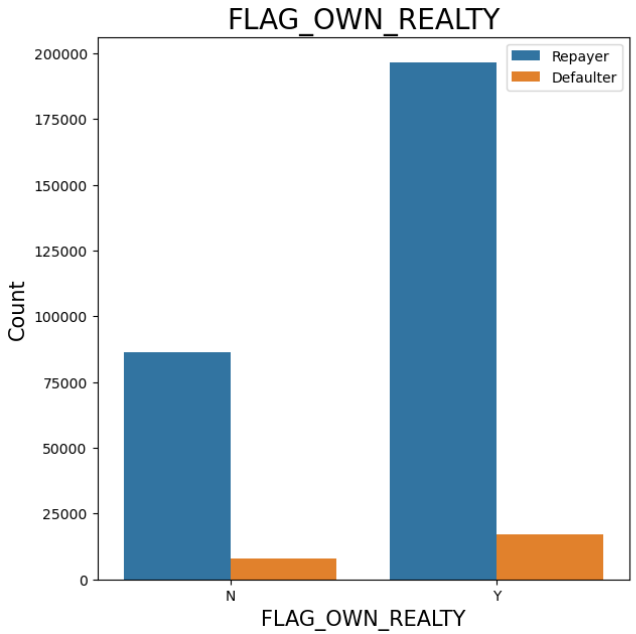
- Inferences: Contract type
- Revolving loans are just a small fraction (10%) from the total number of loans
- Around 8-9% Cash loan applicants and 5-6% Revolving loan applicant are in defaulters

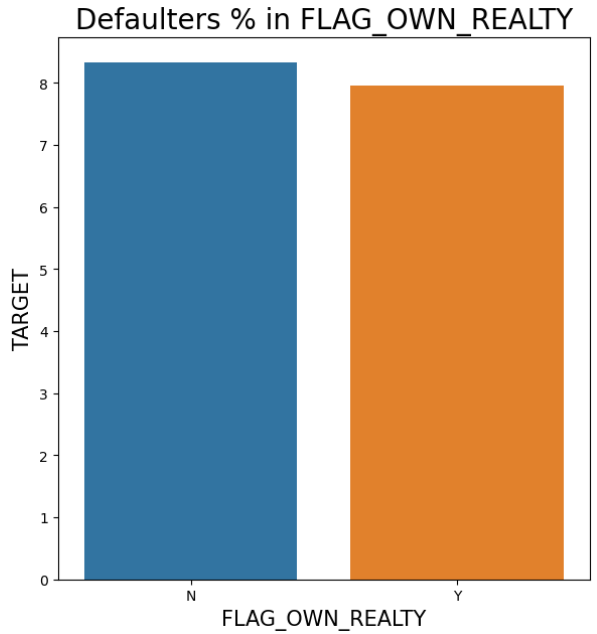




CHECKING THE TYPE OF GENDER ON LOAN REPAYMENT STATUS

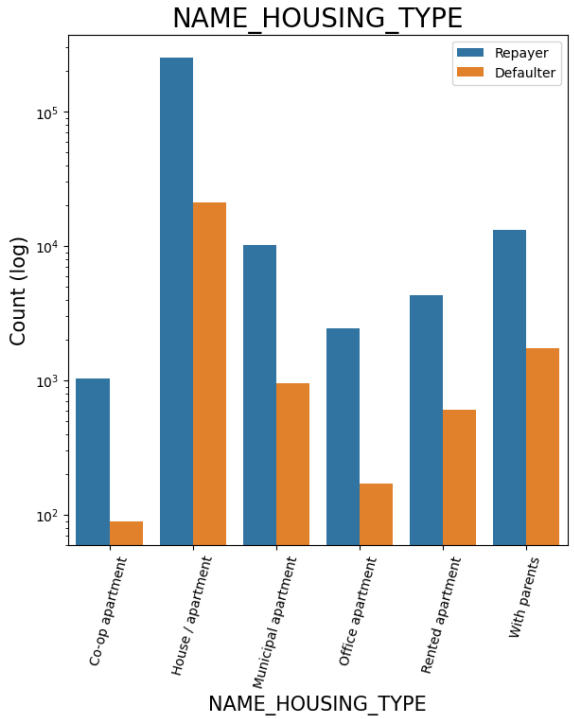
- Inferences: Gender Type
- The number of female clients is almost double the number of male clients.
- Based on the percentage of defaulted credits, males have a higher chance of not returning their loans about 10%, comparing with women about 7%

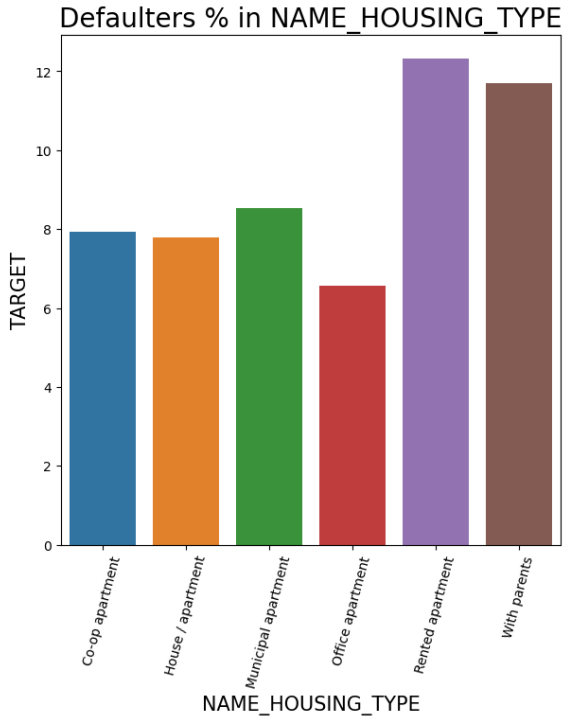




CHECKING IF OWNING A REAL ESTATE IS RELATED TO LOAN REPAYMENT STATUS

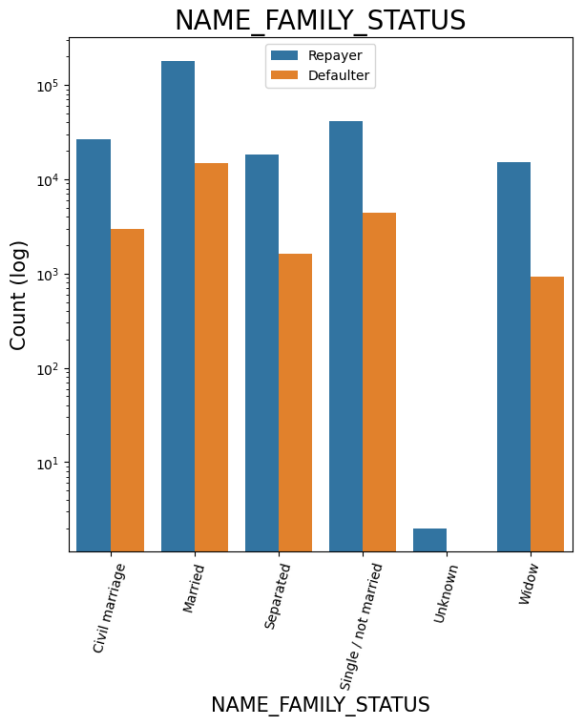
- Inferences:
- The clients who own real estate are more than double of the ones that don't own.
- The defaulting rate of both categories are around the same (~8%). Thus, we can infer that there is no correlation between owning a reality and defaulting the loan.

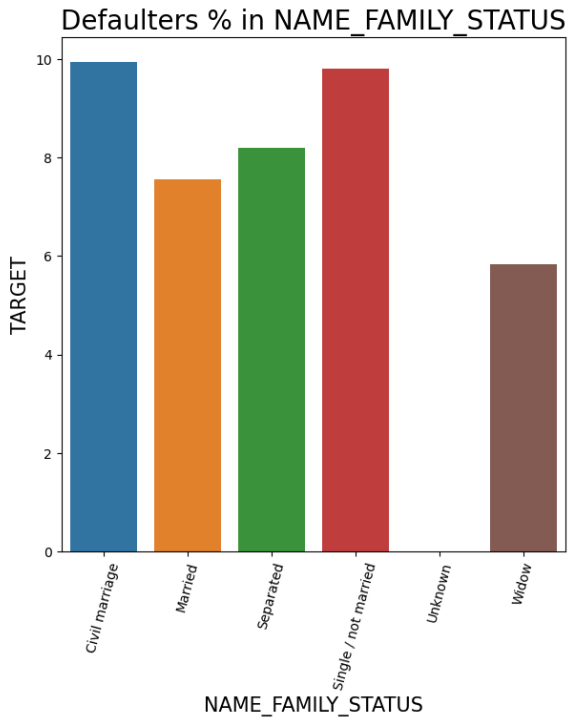




ANALYZING HOUSING TYPE BASED ON LOAN REPAYMENT STATUS

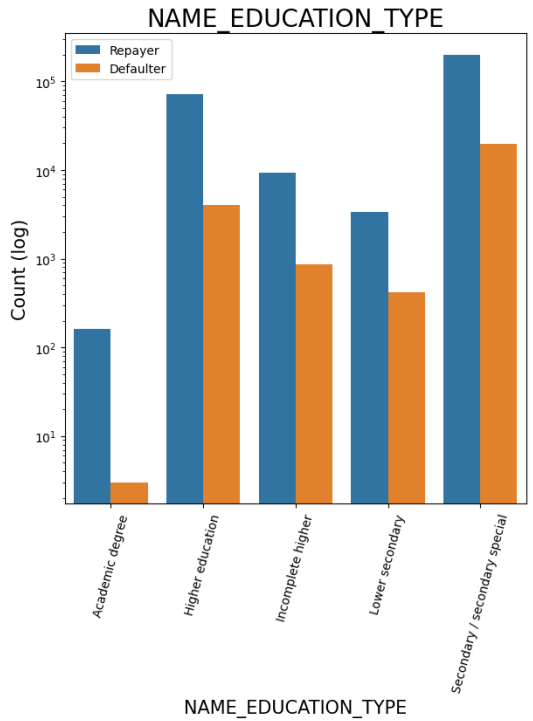
- Inferences: Applicant House type
- Majority of people live in House/apartment
- People living in office apartments have lowest default rate
- People living with parents (~11.5%) and living in rented apartments(>12%) have higher probability of defaulting

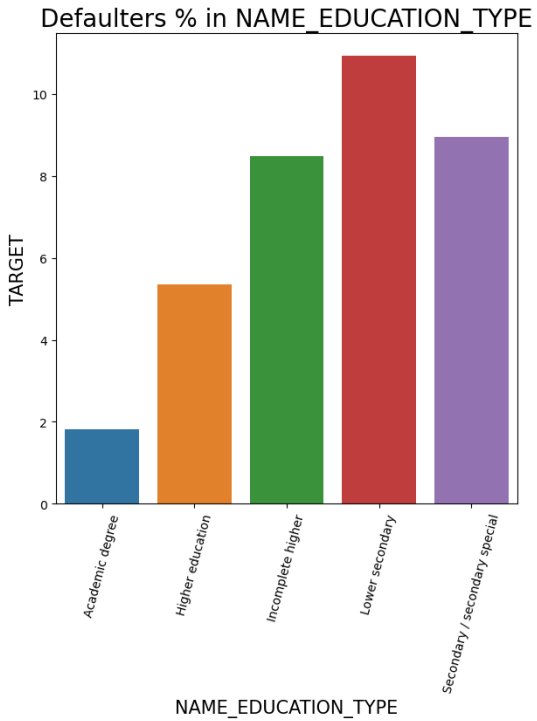




ANALYZING FAMILY STATUS BASED ON LOAN REPAYMENT STATUS

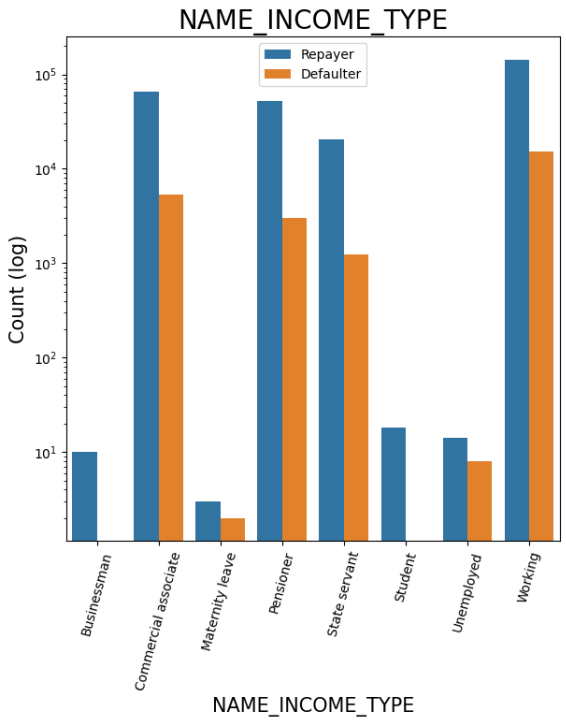
- Inferences:
- Most of the people who have taken loan are married, followed by Single/not married and civil marriage
- In Percentage of defaulters, Civil marriage has the highest percent around (10%) and widow has the lowest around 6% (exception being Unknown).

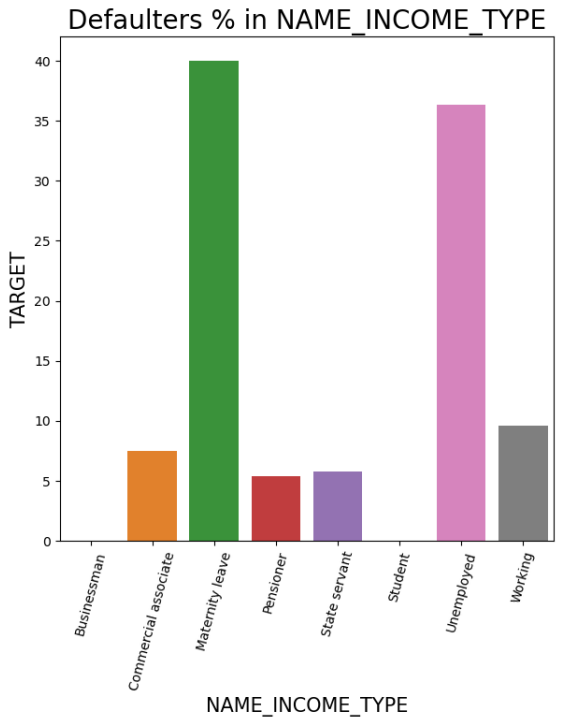




ANALYZING EDUCATION TYPE BASED ON LOAN REPAYMENT STATUS

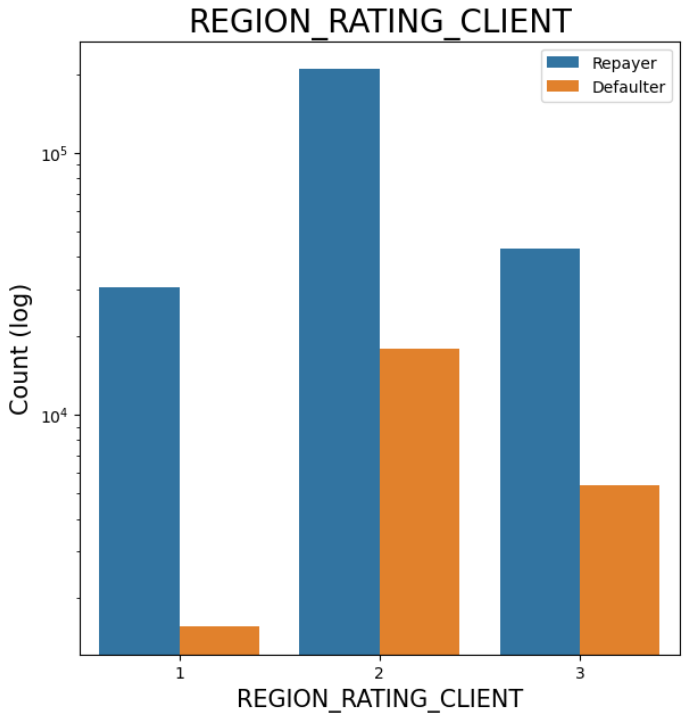
- Inferences: Education Type
- Majority of clients have Secondary/secondary special education, followed by clients with Higher education.
- Very few clients have an academic degree
- Lower secondary category have highest rate of defaulting around 11%.
- People with Academic degree are least likely to default.

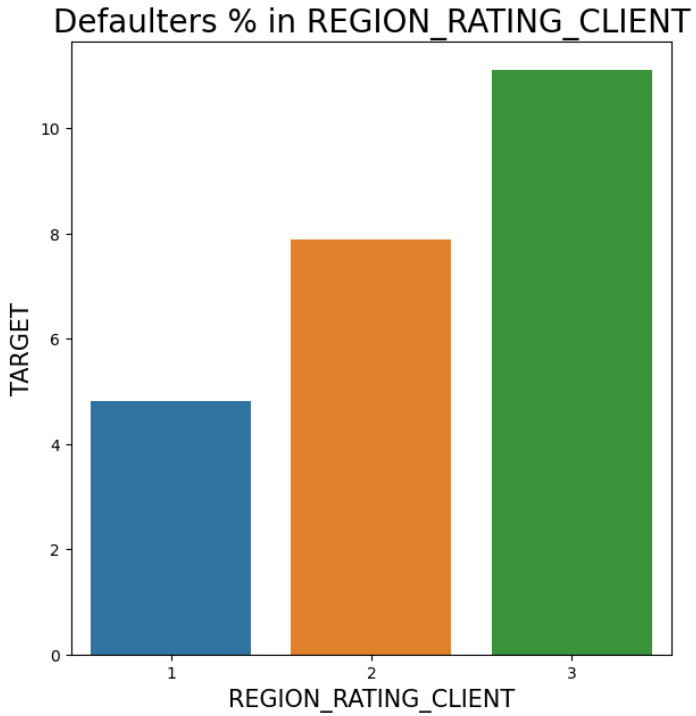




ANALYZING INCOME TYPE BASED ON LOAN REPAYMENT STATUS

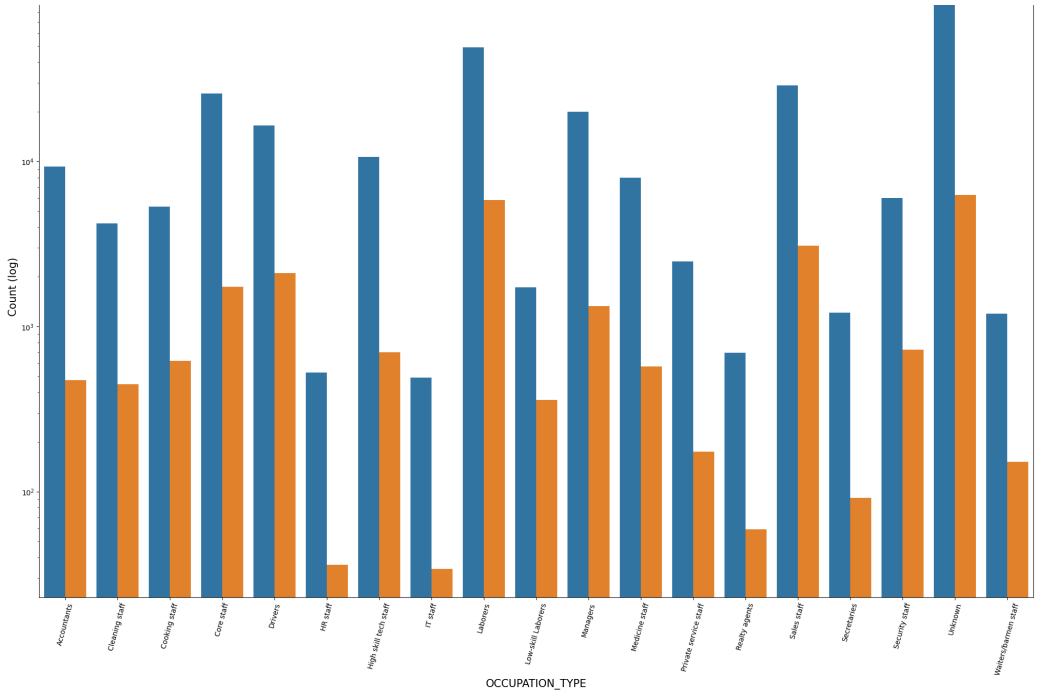
- Inferences:
- Most of applicants for loans income type is Working, followed by Commercial associate, Pensioner and State servant.
- The applicants who are on Maternity leave have defaulting percentage of 40% which is the highest, followed by Unemployed (37%). The rest under average around 10% defaulters.
- Student and Businessmen though less in numbers, do not have default record. Safest two categories for providing loan.

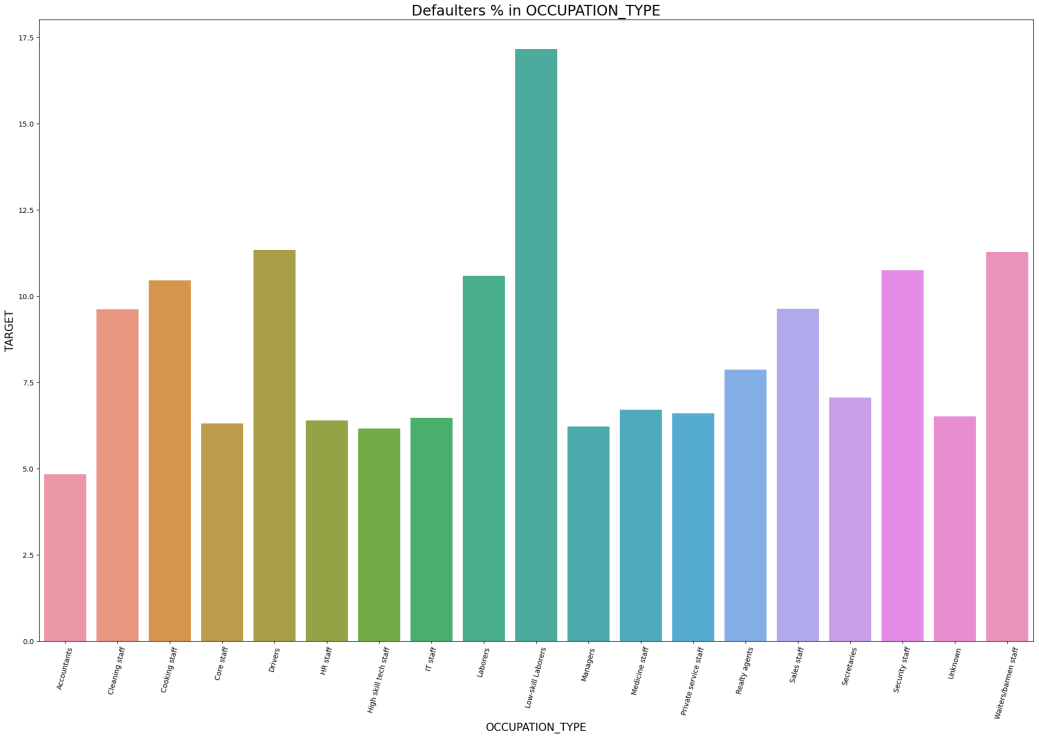




ANALYZING REGION RATING WHERE APPLICANT LIVES BASED ON LOAN REPAYMENT STATUS

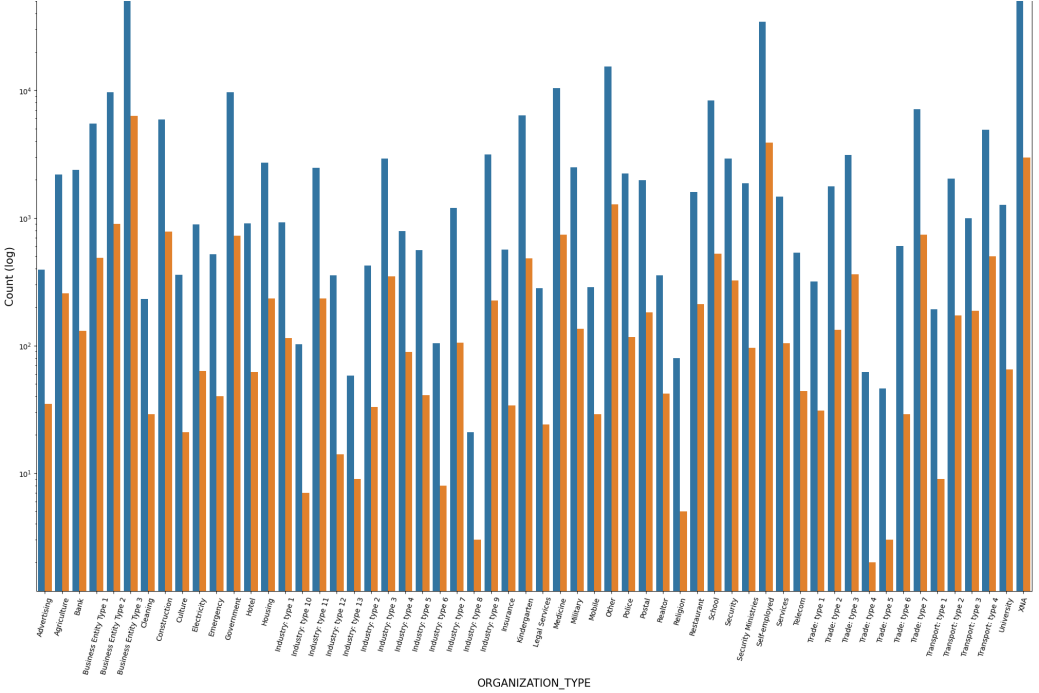
- Inferences:Client Region Rating
- Most of the applicants are living in Region with Rating 2 place.
- Region Rating 3 has the highest default rate (11%)
- Applicant living in Region Rating 1 has the lowest probability of defaulting, thus safer for approving loans

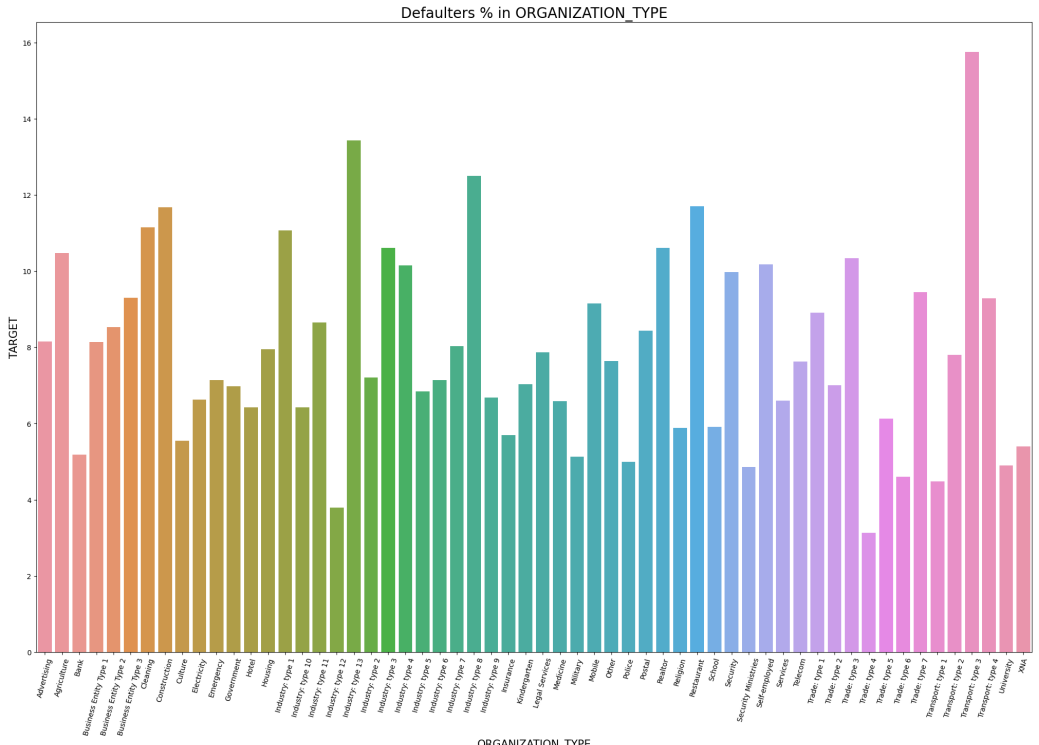




ANALYZING OCCUPATION TYPE WHERE APPLICANT LIVES BASED ON LOAN REPAYMENT STATUS

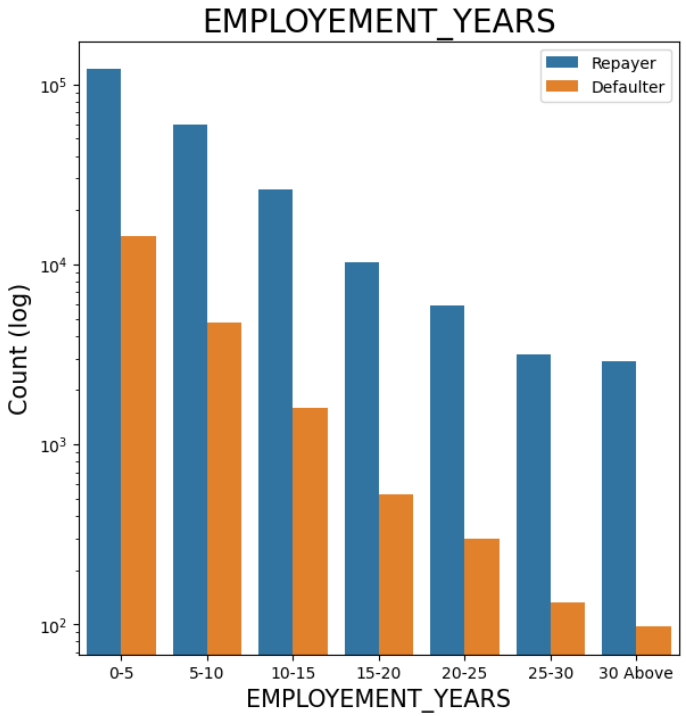
- Inferences:
- Most of the loans are taken by Laborers, followed by Sales staff.
- IT staff are less likely to apply for Loan.
- Category with highest percent of defautess are Low-skill Laborers (above 17%), followed by Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff

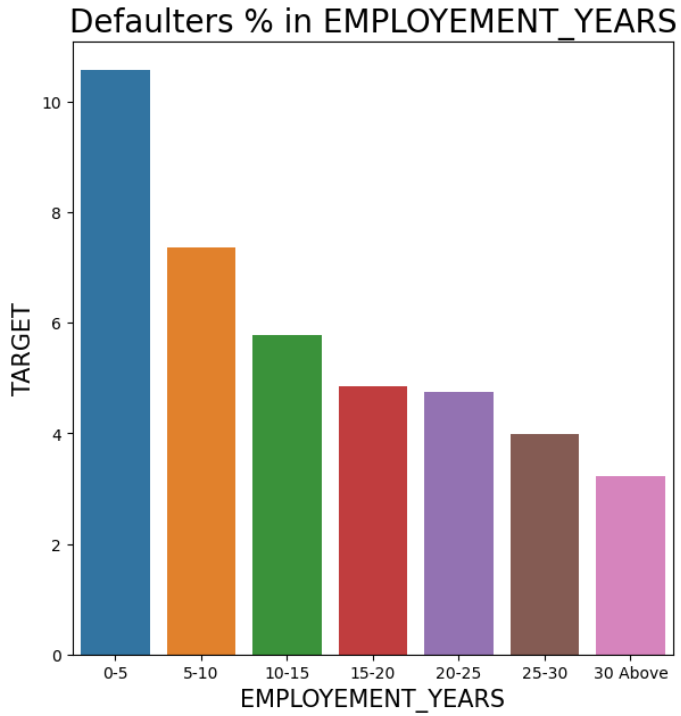




CHECKING LOAN REPAYMENT STATUS BASED ON ORGANIZATION TYPE

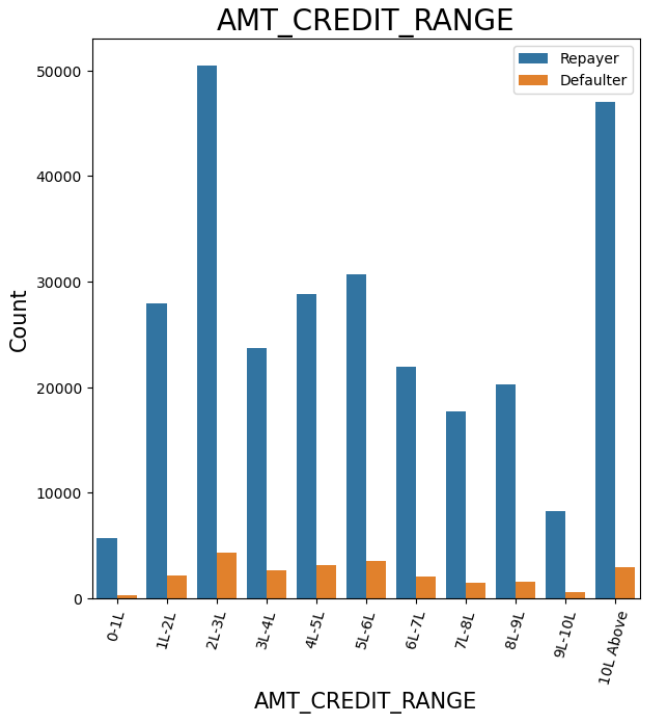
- Inferences: Organization Type
- Organizations with highest percent of defaultess are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%).
- Self employed people have relative high defaulting rate, to be safer side loan disbursement should be avoided or provide loan with higher interest rate to mitigate the risk of defaulting.
- Most of the people application for loan are from Business Entity Type 3
- For a very high number of applications, Organization type information is unavailable(XNA)
- Following category of organization type has lesser defaulters thus safer for providing loans: Trade Type 4 and 5, Industry type 8

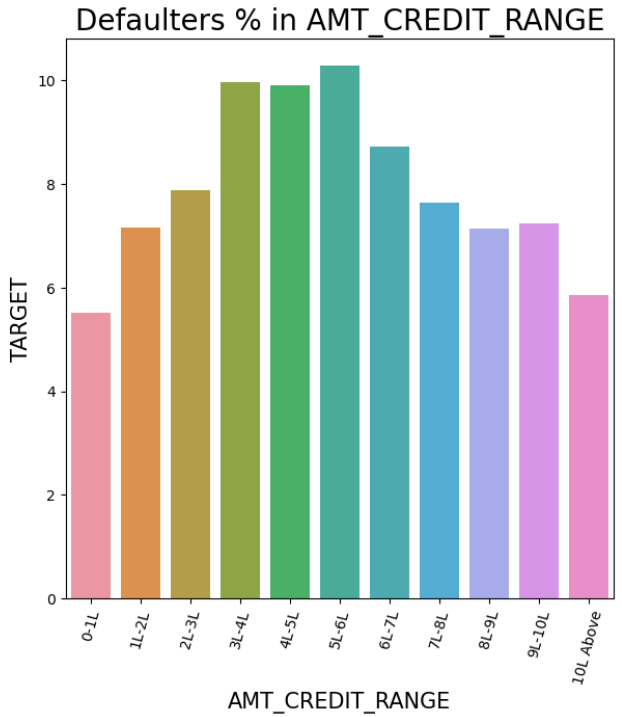




ANALYZING EMPLOYMENT_YEAR BASED ON LOAN REPAYMENT STATUS

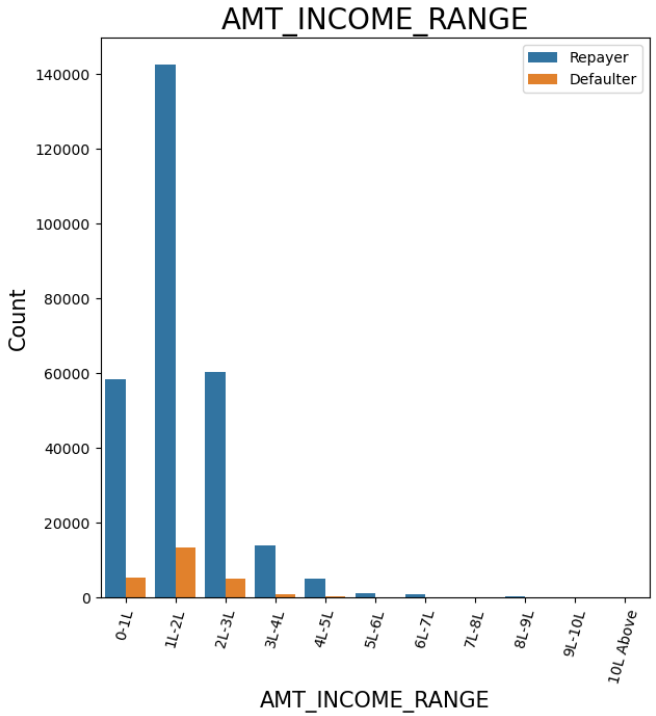
- Inferences: Employment in Years
- Majority of the applicants having working experience between 0-5 years are defaulters. The defaulting rating of this group is also the highest which is around 10%
- With increase of employment year, defaulting rate is gradually decreasing.
- With people having 40+ year experience have less than 1% default rate

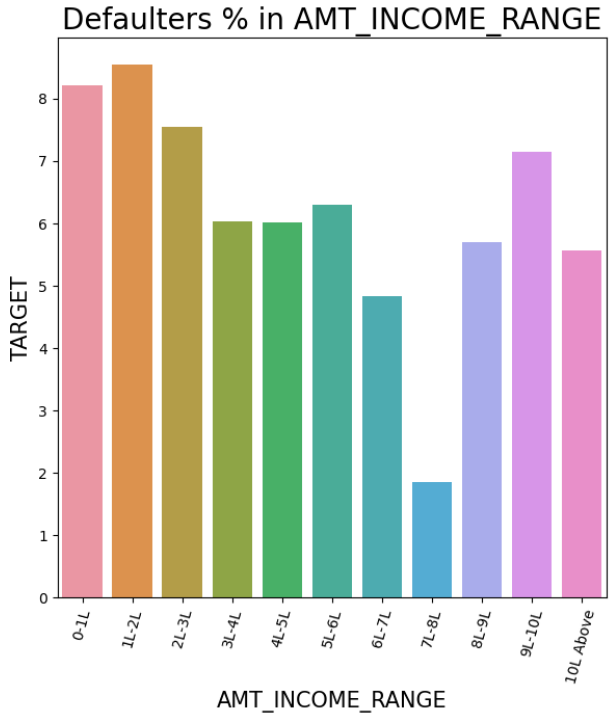




ANALYZING AMOUNT_CREDIT BASED ON LOAN REPAYMENT STATUS

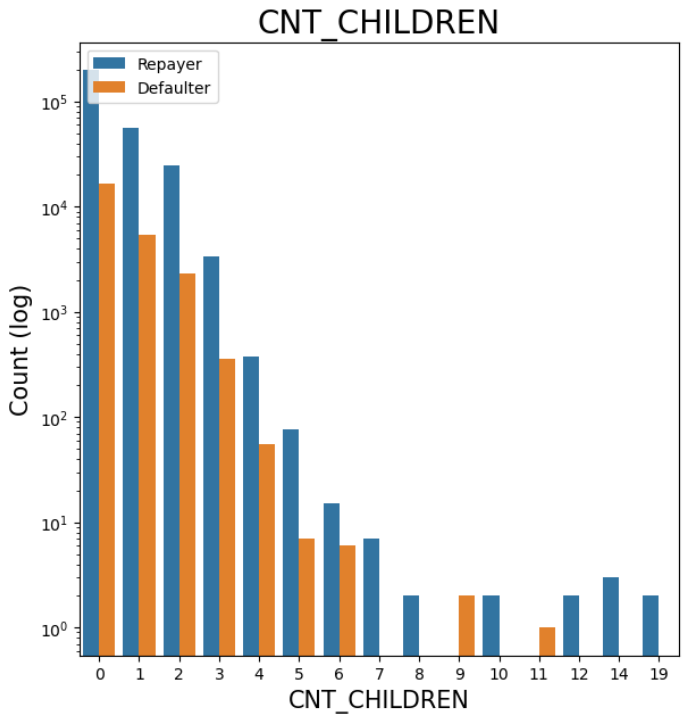
- Inferences: Loan Amount
- There are high number of applicants have loan in range of 2-3 Lakhs followed by 10 Lakh above range
- People who get loan for 3-6 Lakhs have the greatest number of defaulters than other loan range.

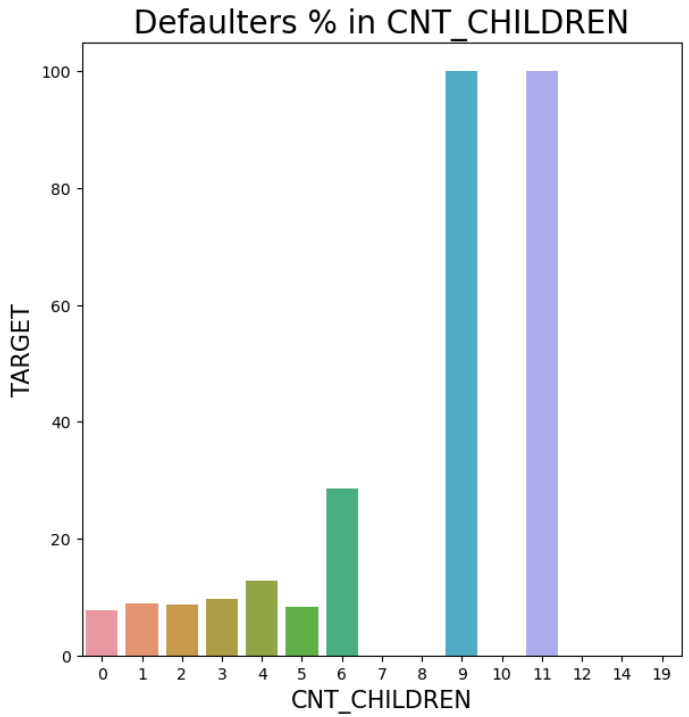




ANALYZING AMOUNT_INCOME RANGE BASED ON LOAN REPAYMENT STATUS

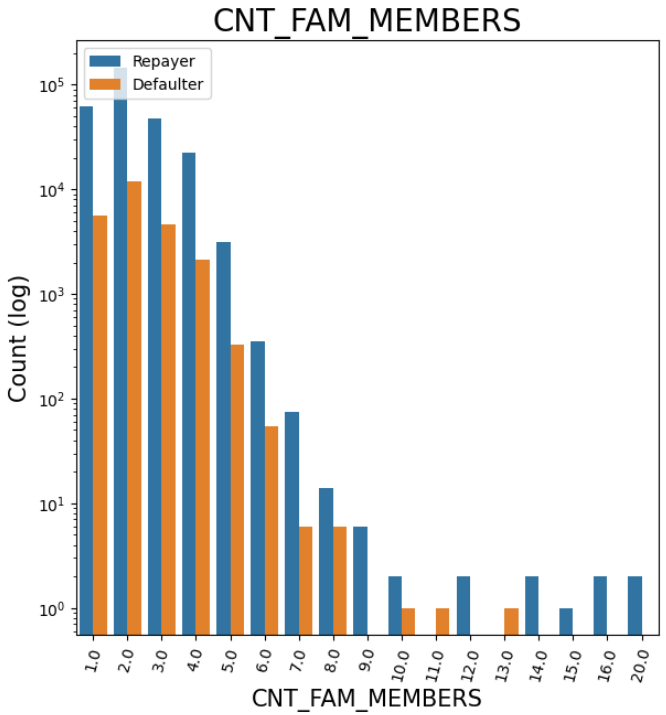
- Inferences: Applicant Income
- Majority of the applications have Income total less than 3 Lakhs.
- Application with Income less than 3 Lakhs has high probability of defaulting
- Applicant with Income 7-8 Lakhs are less likely to default.

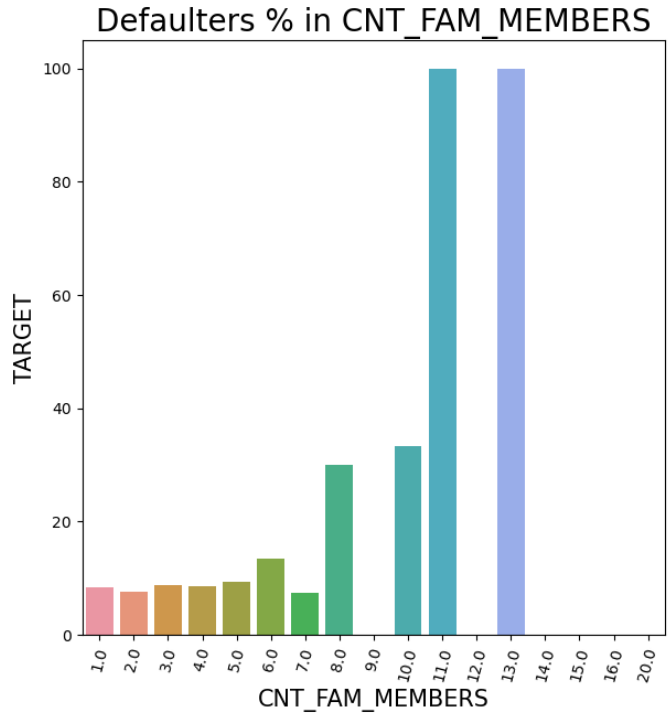




ANALYZING NUMBER OF CHILDREN BASED ON LOAN REPAYMENT STATUS

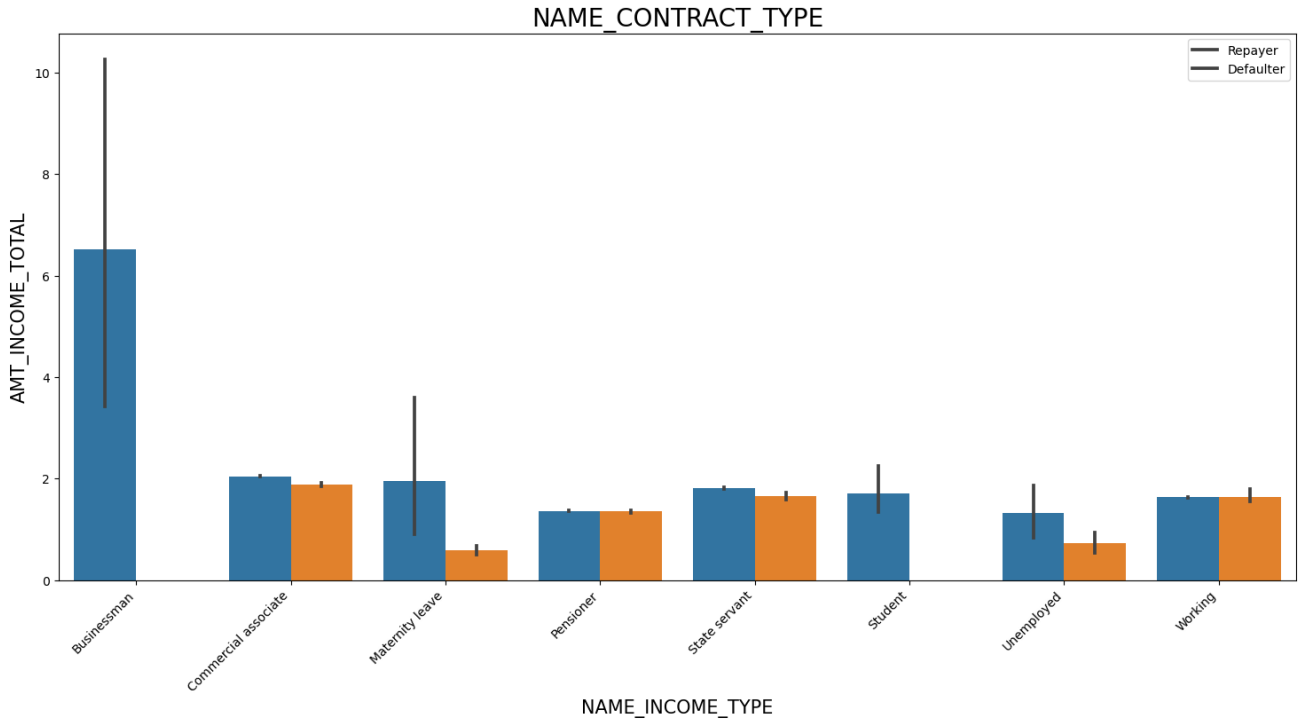
- Inferences: Client Children's Count
- Most of the applicants do not have children
- Very few clients have more than 3 children.
- Client who have more than 4 children has a very high default rate with child count 9 and 11 showing 100% default rate





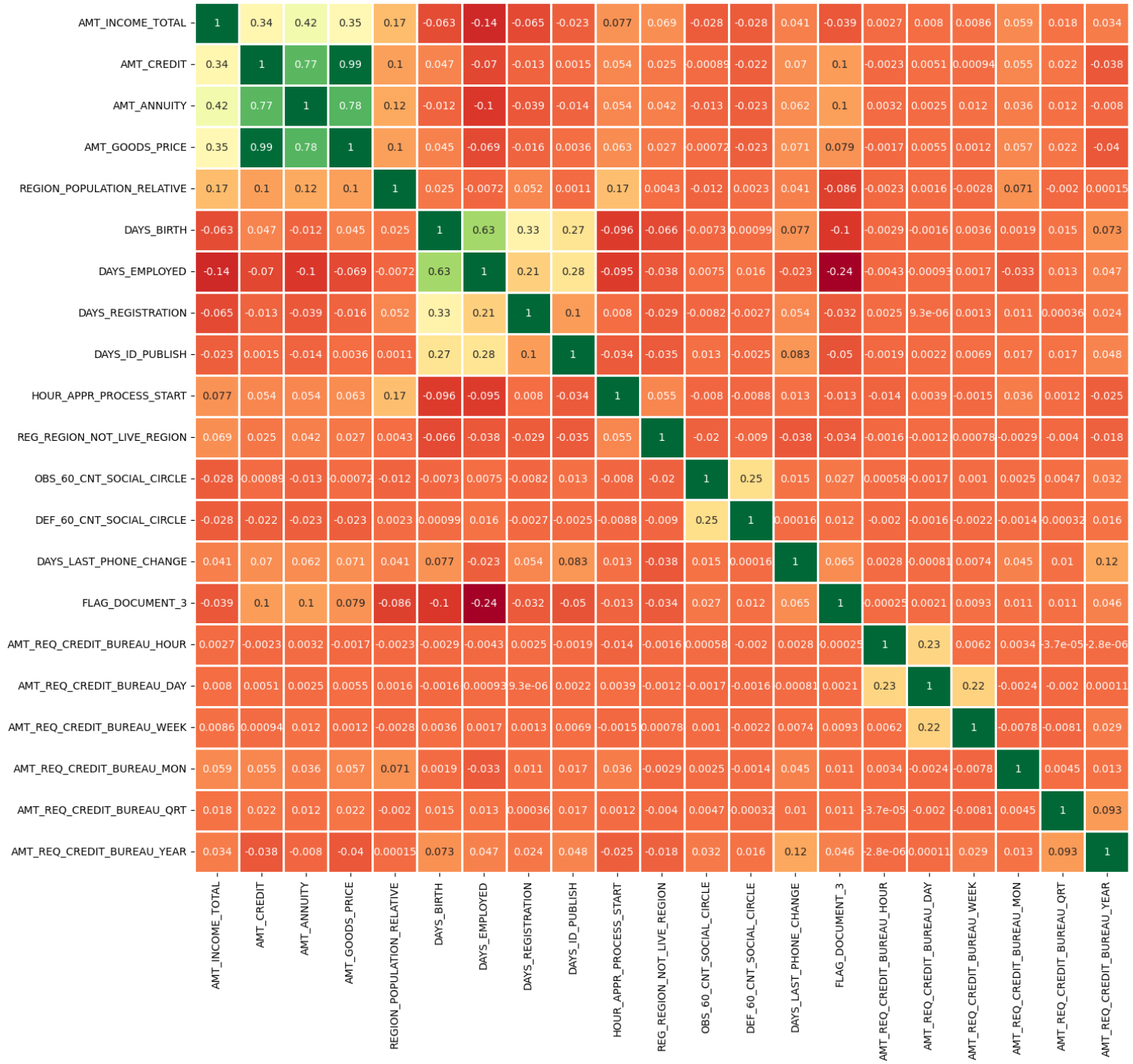
ANALYZING NUMBER OF FAMILY MEMBERS BASED ON LOAN REPAYMENT STATUS

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



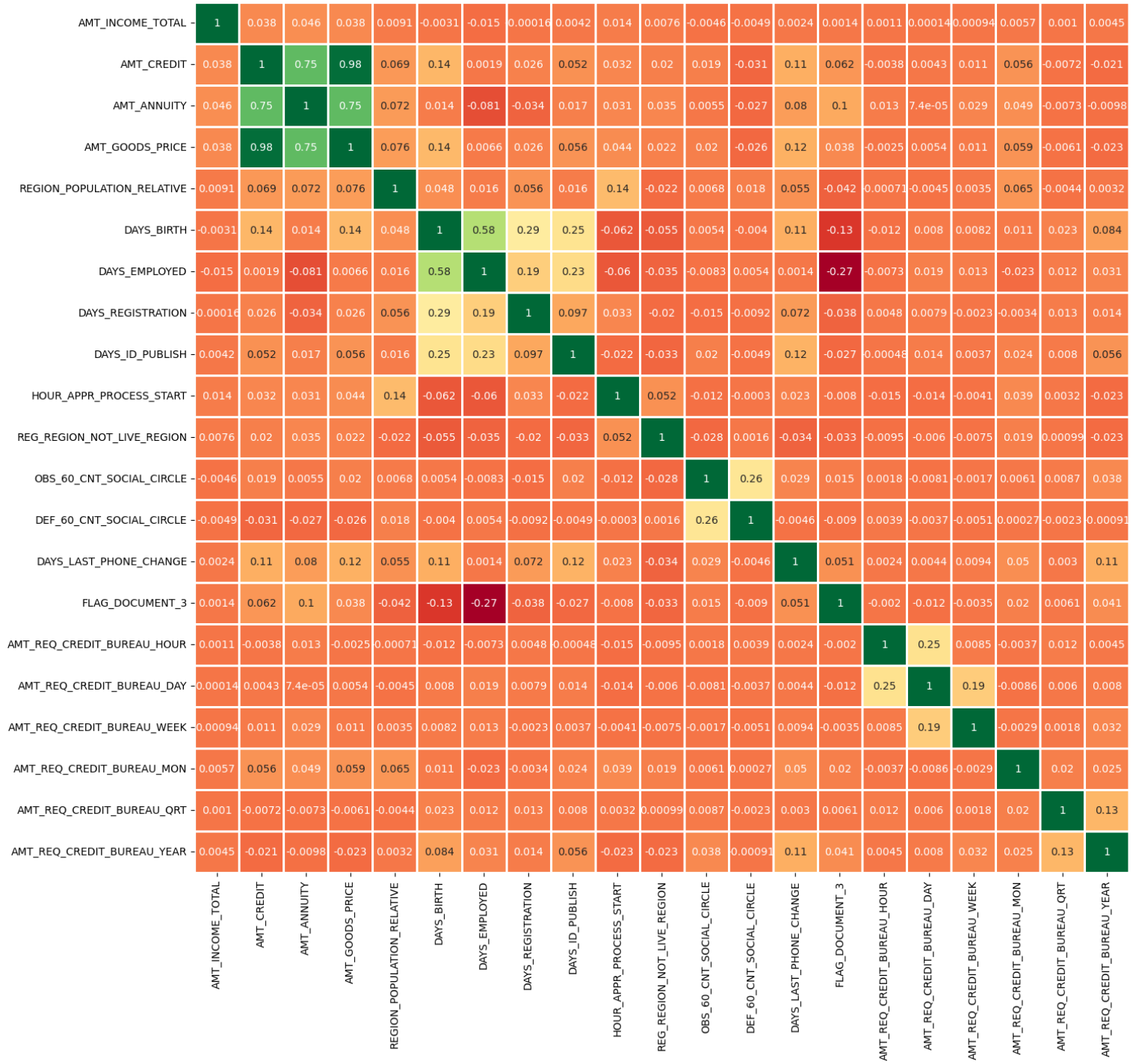
CATEGORICAL BIVARIATE OR MULTIVARIATE ANALYSIS

- Income type vs Income Amount Range on a Seaborn Barplot.
- Inferences:
- Businessman income is the highest and the estimated range with default 95% confidence level seem to indicate that the income of a Businessman could be in the range of slightly close to 4 lakhs and slightly above 10 lakhs



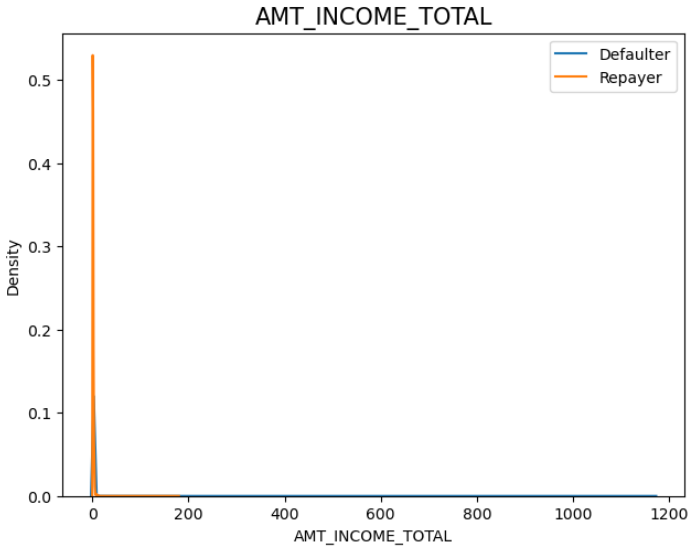
NUMERIC VARIABLES ANALYSIS

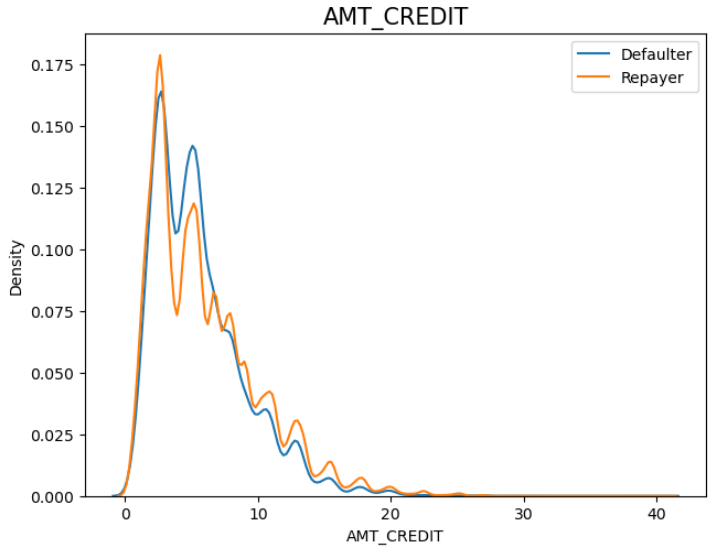
- plotting heatmap to see linear correlation among Repayers.
- Inferences:
- Correlating factors amongst repayers
- 1. Credit amount is highly correlated with:
  - Goods Price Amount
  - Loan Annuity
  - Total Income
- 2. We can also see that repayers have high correlation in number of days employed.

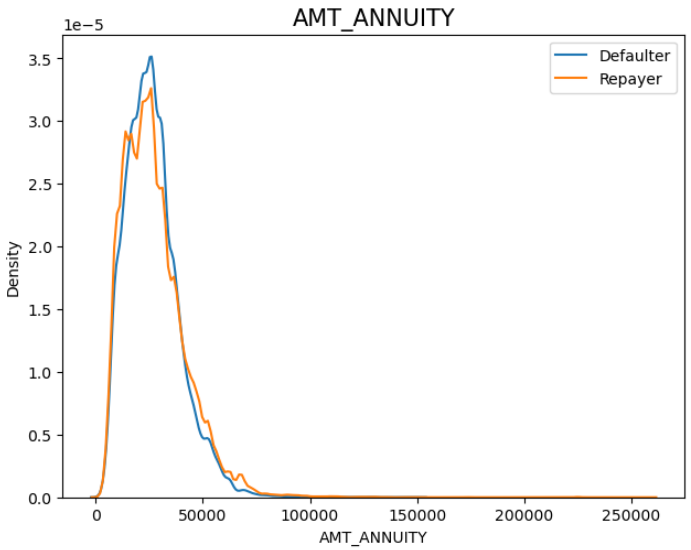


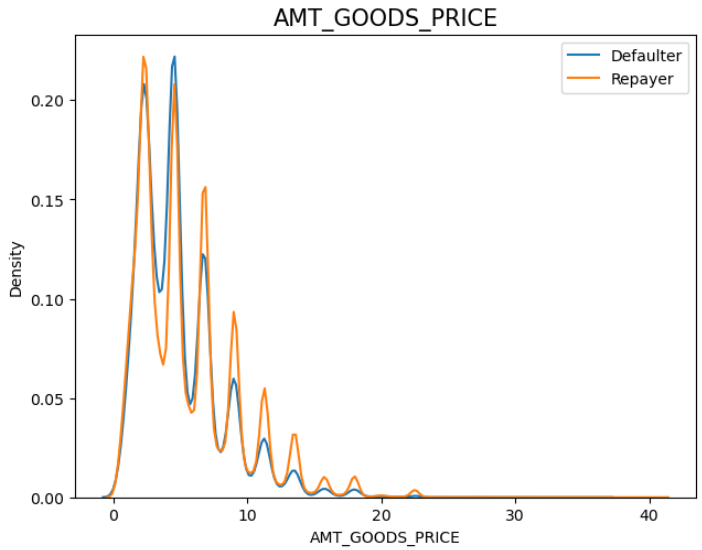
GETTING THE TOP 10 CORRELATION FOR THE DEFAULTER DATA

- Inferences: Correlating factors amongst repayers
- Credit amount is highly correlated with good price amount which is same as repayers.
- 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)



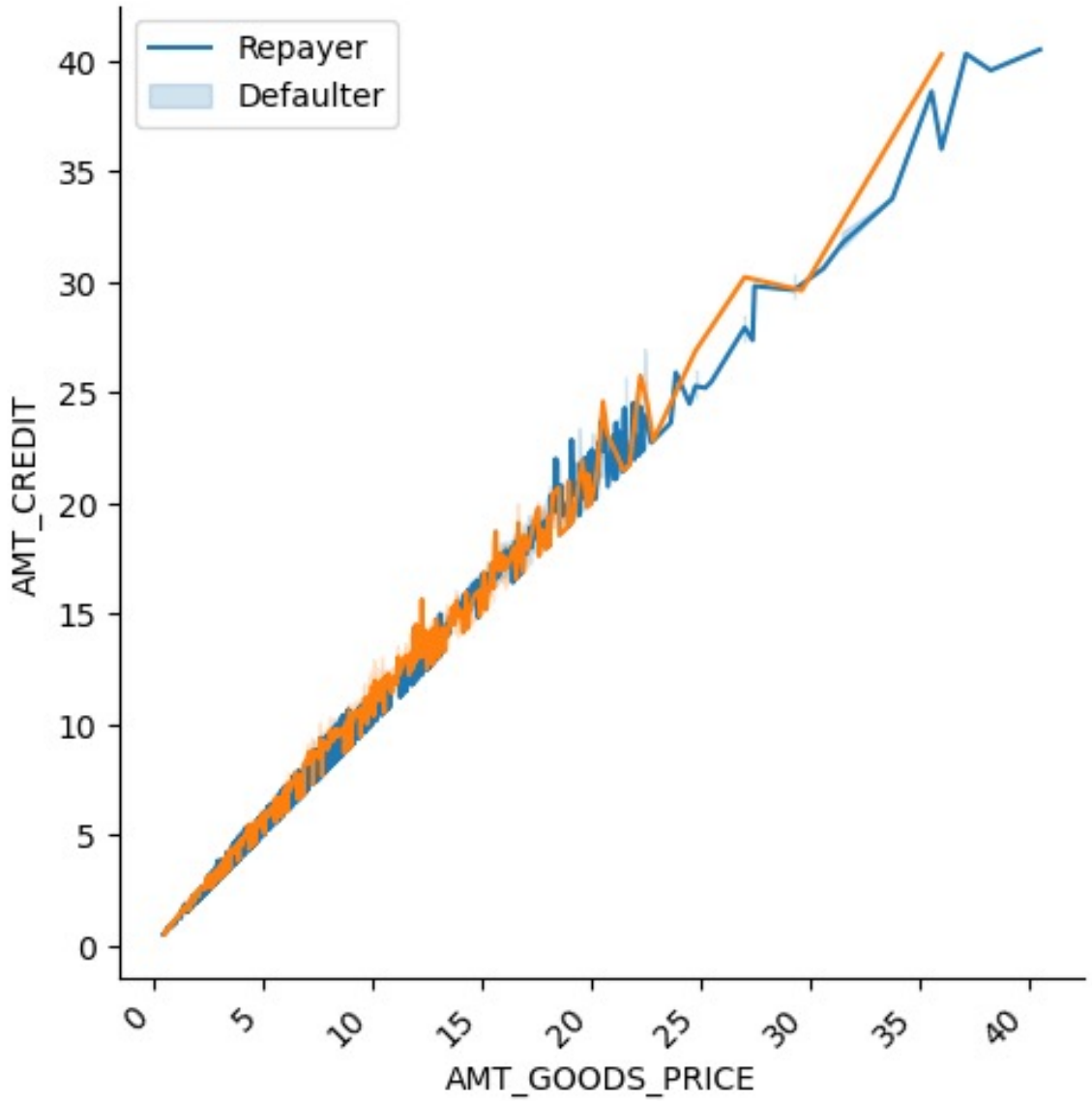






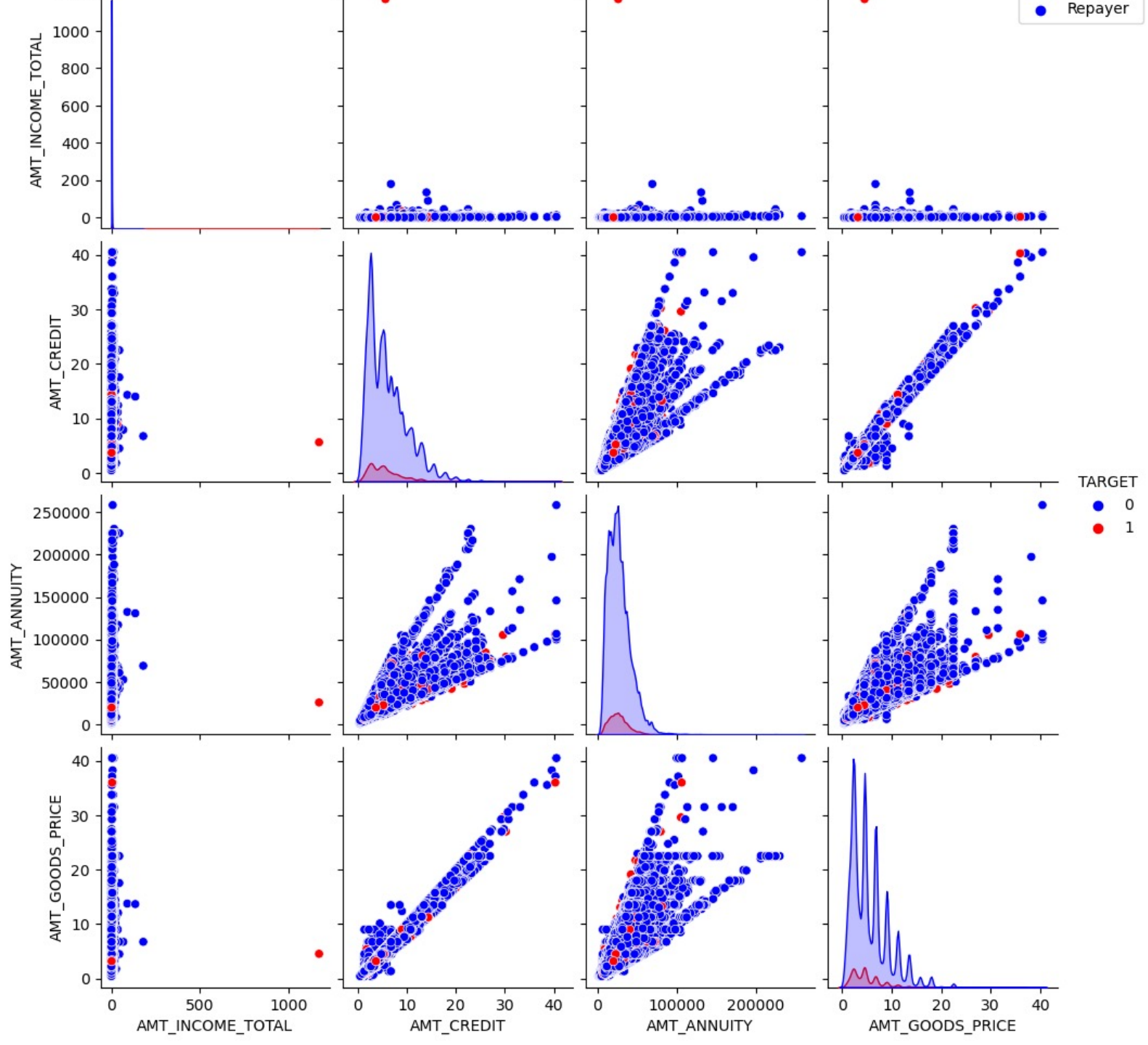
NUMERICAL UNIVARIATE ANALYSIS

- Plotting the numerical columns related to amount as distribution plot to see density.
- Inferences:
- Most no of loans are given for goods price below 10 lakhs
- Most people pay annuity below 50K for the credit loan
- Credit amount of the loan is mostly less then 10 lakhs
- The repayers and defaulters' distribution overlap in all the plots and hence we cannot use any of these variables in isolation to decide



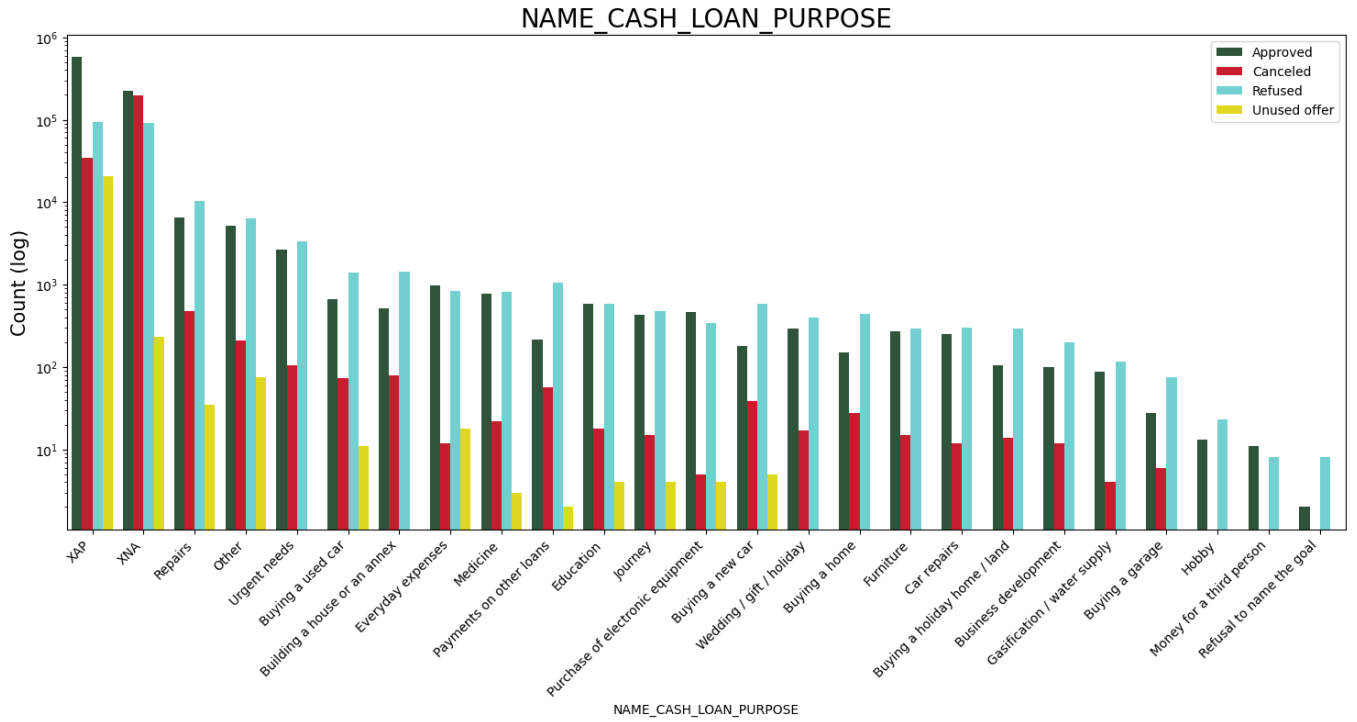
NUMERICAL BIVARIATE ANALYSIS

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



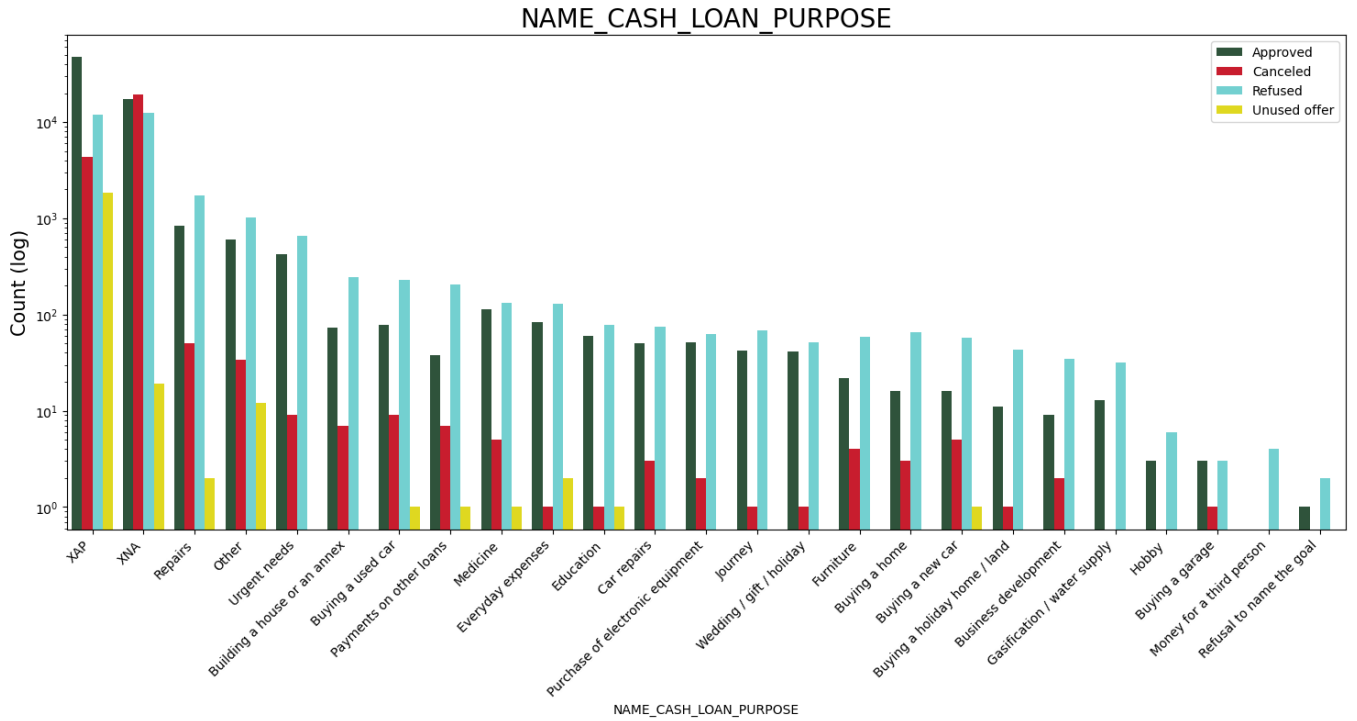
PLOTTING PAIRPLOT BETWEEN AMOUNT VARIABLE TO DRAW REFERENCE AGAINST LOAN REPAYMENT STATUS

- Inferences:
- When Annuity Amount > 15K and Good Price Amount > 20 Lakhs, there is a lesser chance of defaulters
- Loan Amount(AMT_CREDIT) and Goods price(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 >20 Lakhs



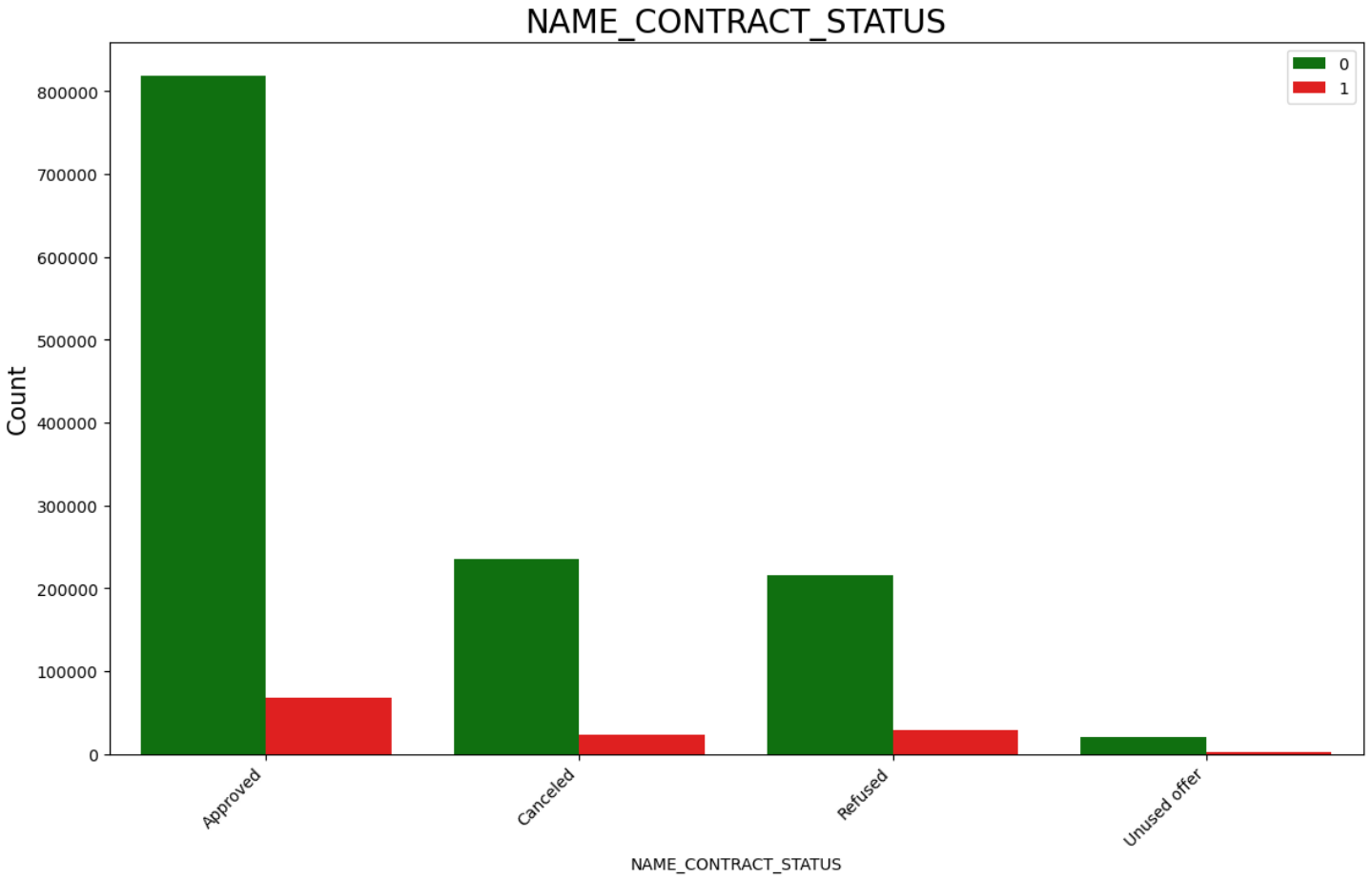
MERGED DATAFRAMES ANALYSIS

- Bisecting the "loan_df" dataframe based on Target value 0 and 1 for correlation and other analysis.
- Plotting Contract Status vs purpose of the loan



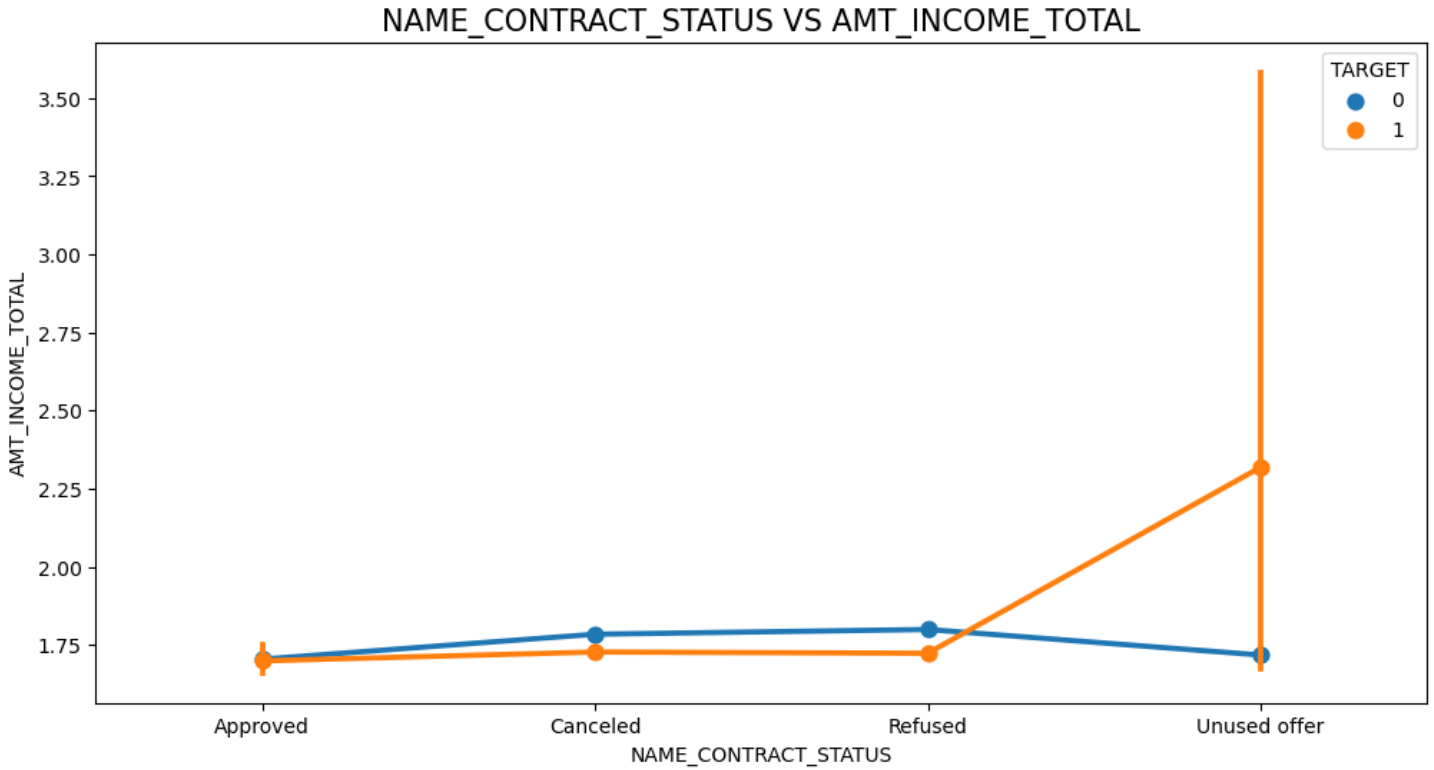
MERGED DATAFRAMES ANALYSIS

- Inferences:
- Loan purpose has high number of unknown values (XAP, XNA)
- Loan taken for the purpose of Repairs looks to have highest default rate
- Huge number application have been rejected by bank or refused by client which are applied for Repair or Other. from this we can infer that repair is considered high risk by bank. Also, either they are rejected, or bank offers loan on high interest rate which is not feasible by the clients, and they refuse the loan.



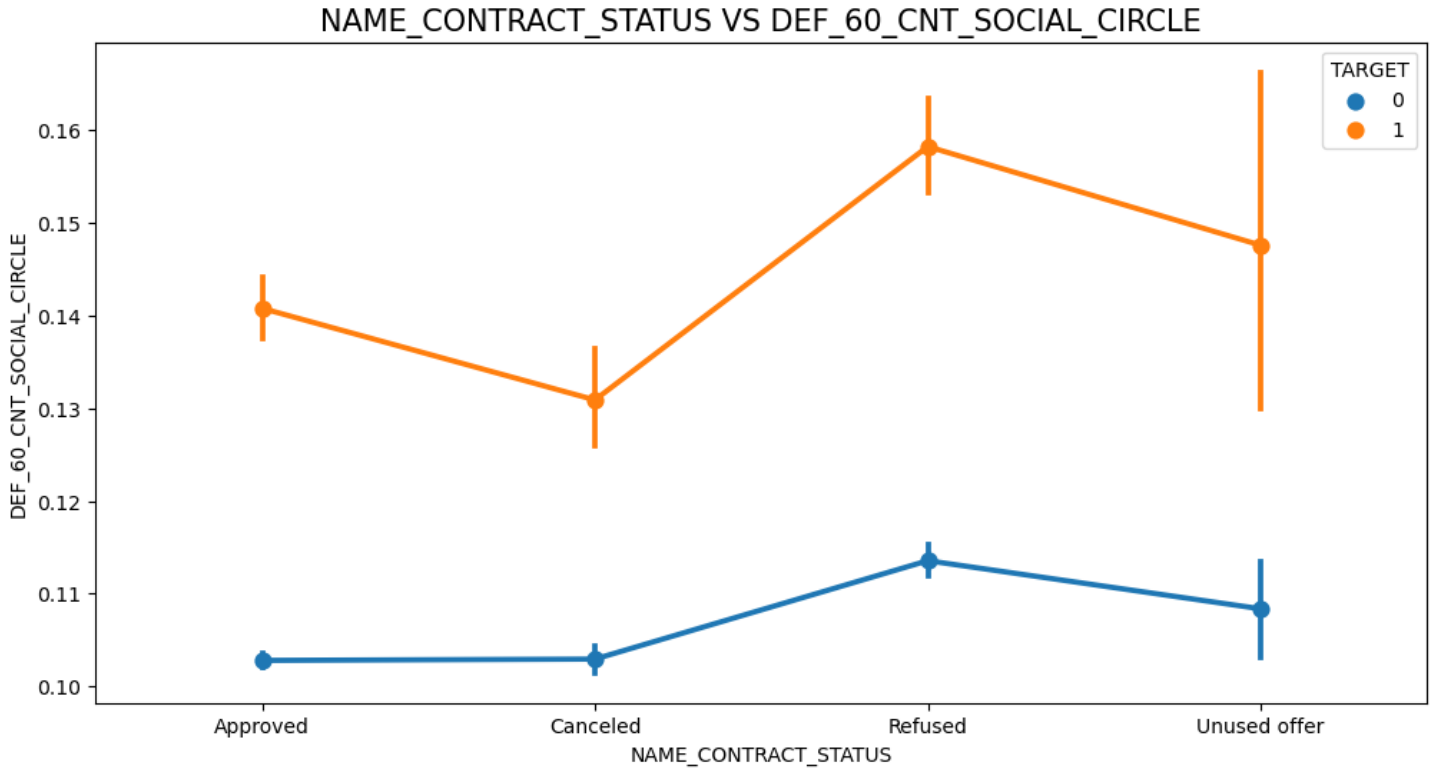
CHECKING CONTRACT STATUS BASED ON LOAN REPAYMENT STATUS WHETHER THERE IS ANY BUSINESS LOSS OR FINANCIAL LOSS

- Inferences:
- 90% of the previously cancelled client have actually repayed the loan. Revising the interest rates would increase business opportunity for these clients
- 88% of the clients who have been previously refused a loan has pay back the loan in current case.
- Refusal reason should be recorded for further analysis as these clients could turn into potential repaying customer.



PLOTTING THE RELATIONSHIP BETWEEN INCOME TOTAL AND CONTACT STATUS

- Inferences:
- The point plot show that the people who have not used offer earlier have defaulted even when their 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

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

CONCLUSION

- After analysing the datasets, there are few attributes of a client with which the bank would be able to identify if they will repay the loan or not. The analysis is consisted as below with the contributing factors and categorization:
  - Decisive Factor whether an applicant will be Repayer
  - Decisive Factor whether an applicant will be Defaulter
  - Factors that Loan can be given on Condition of High Interest rate to mitigate any default risk leading to business loss
  - Suggestions

A. DECISIVE FACTOR WHETHER AN APPLICANT WILL BE REPAYER:

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.

B. 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 their 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 3lakhs, there is an increase in defaulters.

C. FACTORS THAT LOAN CAN BE GIVEN ON CONDITION OF HIGH INTEREST RATE TO MITIGATE ANY DEFAULT RISK LEADING TO 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 3-6 Lakhs 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 3Lakhs 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.

D. SUGGESTIONS:

- 90% of the previously cancelled client have actually repayed 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