EDA Case Study
Risk Analytics in banking and financial services

Presented by,
Mr. Raviraj Suresh Kangle

- **Subject: The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history.**
- **Aim: We have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.**
- **When a client applies for a loan, there are four types of decisions that could be taken by the client/company):**
- **Approved: The Company has approved loan Application**
- **Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want.**

- **Refused**: The company had rejected the loan (because the client does not meet their requirements etc.).
- **Unused offer**: Loan has been cancelled by the client but at different stages of the process.
- In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency to default.
- At the first we will take the necessary steps in Jupiter notebook
- Importing the required library is important for further analysis refer Jupiter notebook

- Step one we have to analysis the application dataset
- After loading the dataset I found that there are total 307511 rows and 122 columns
- Then I moved to take the information of the dataset
- Where I found the data type is float,int,object
- Dataset contains 122 columns but some data are object so we have to convert it further analysis

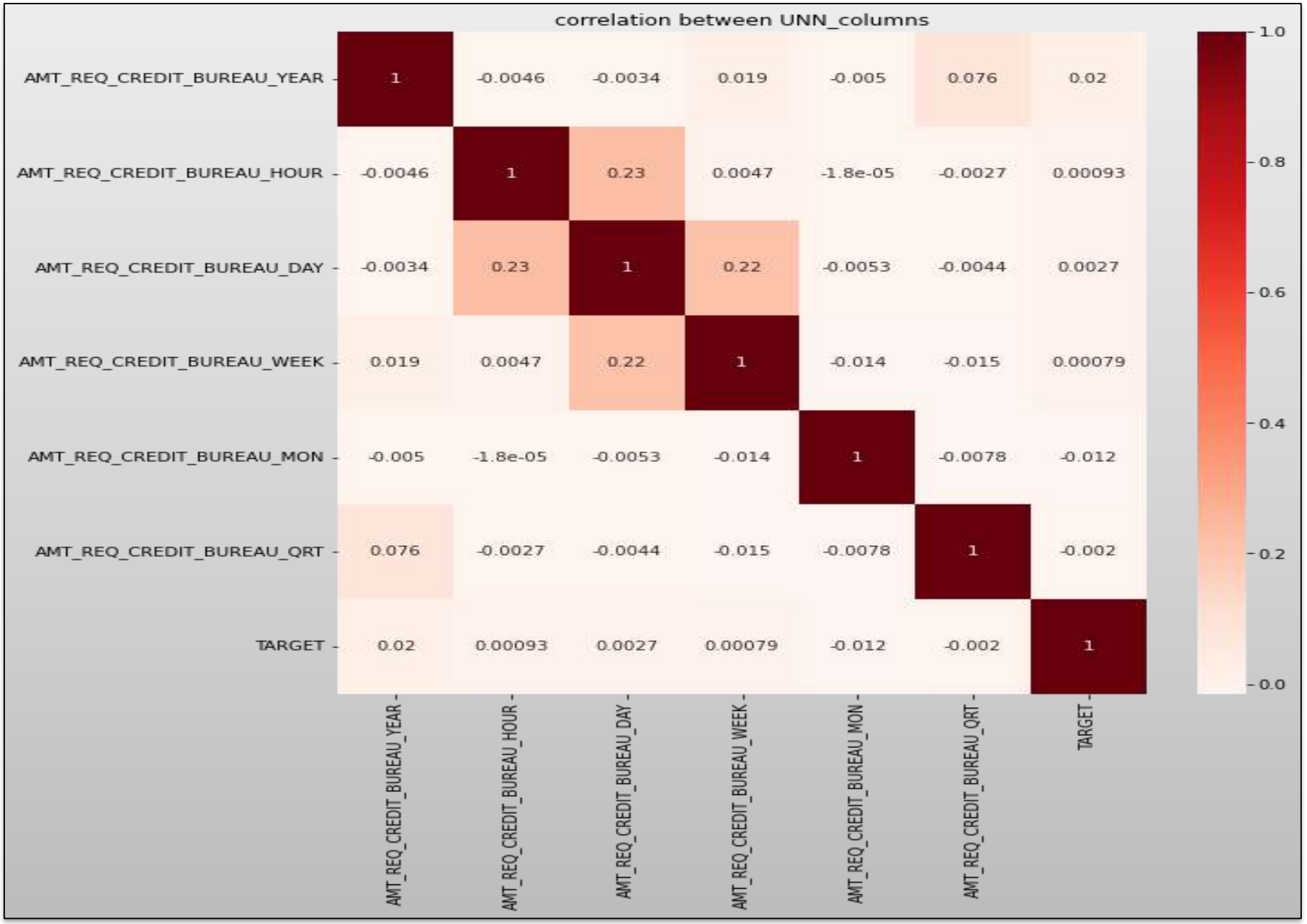
- Then I move to describe option for analysis of numerical values by using describe option.
- Describe option shows that there are ome negative values and positive values so we are going to fix this values in coming codes for suitability.
- To control the high values we have to perform the Standardising method .

- The next step will be data cleaning
- In data cleaning I found that there are so many null values in data set
- So decided to check the null val;ues in percentage
- And found that 69.87 is the highest and contain many rows with the same percentage
- As per the industry criteria mostly we consider that above 50% of the null value is deleted

- Before that I have to identify the no of row having more than 50% of null values by using length function
- Observed that 41 columns having null values more than 50% which are in different category.
- It is important to fix the null values
- After using drop option I fount 307511 rows and 81 columns as 41 columns are deleted
- The next step is to deal with the null values having above 10 %

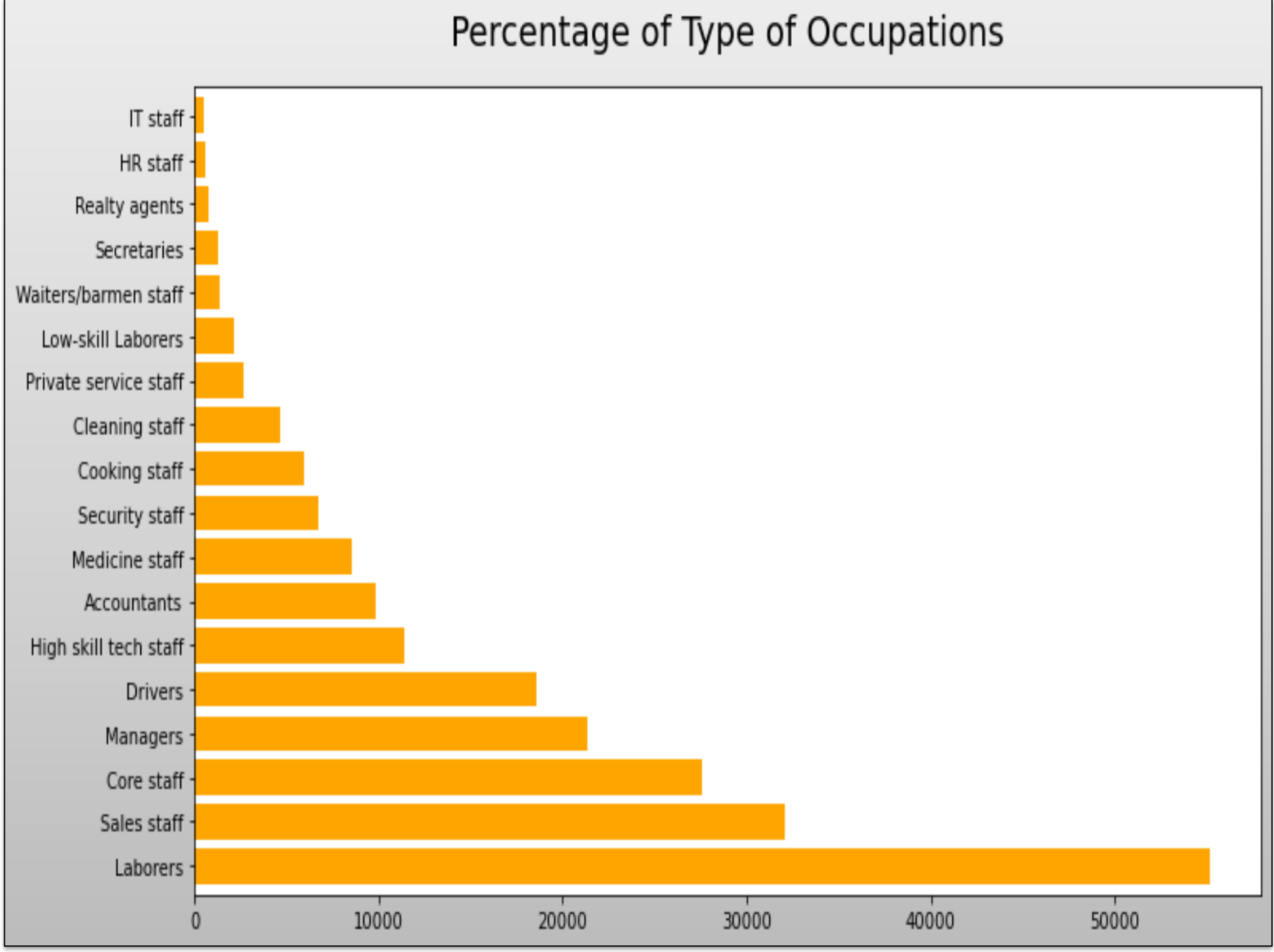
- In this case we use 10% because we dont want to loose data we can observe that occupation type, EXT_SOURCE_3,AMT_REQ_CREDIT_BUREAU_YEAR ,AMT_REQ_CREDIT_BUREAU_HOUR,AMT_REQ_CREDIT_BUREAU_DAY,AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT are the further columns we can use as a target columns for further analysis

- Now in dataset we can see missing values is only 8 columns which is above 10%
- After droping the missing value left with 307511 rows and 73 columns
- Now the data is partially clean and required to remove unwanted columns
- Before that we have to check the correlation between them and easy to take decision
- There seems to be no linear correlation and also from columns description we decided to remove these columns. Please refer heat map


correlation between UNN_columns
AMT_REQ_CREDIT_BUREAU_YEAR
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
TARGET
1 -0.0046 -0.0034 0.019 -0.005 0.076 0.02
-0.0046 1 0.23 0.0047 -1.8e-05 -0.0027 0.00093
-0.0034 0.23 1 0.22 -0.0053 -0.0044 0.0027
0.019 0.0047 0.22 1 -0.014 -0.015 0.00079
-0.005 -1.8e-05 -0.0053 -0.014 1 -0.0078 -0.012
0.076 -0.0027 -0.0044 -0.015 -0.0078 1 -0.002
0.02 0.00093 0.0027 0.00079 -0.012 -0.002 1
1.0
0.8
0.6
0.4
0.2
0.0

- So decided to drop the columns which are not correlated
- After dropping it shows that there are 67 columns and 307511 rows
- Now we will check columns with FLAGS and their relation with TARGET columns to remove irrelevant ones
- We can replace the flag column by 0 as repayer and 1 for target because flag columns are categorical and it is ease to analysis after converting to numerical

- 29 columns are observed as flag columns we can remove it because it will not contribute in decision making for a bank
- After dropping flag columns the dataset left with 42 columns
- Now we can consider occupation_tyre column for further analysis
- There are some null values in occupation but we cannot delete the null value we can use fill -Na option
- We can see that 'occupation_type is categorical and having missing 13.33 % values so now we will fix this'


Percentage of Type of Occupations
IT staff
HR staff
Realty agents
Secretaries
Waiters/barmen staff
Low-skill Laborers
Private service staff
Cleaning staff
Cooking staff
Security staff
Medicine staff
Accountants
High skill tech staff
Drivers
Managers
Core staff
Sales staff
Laborers
0
10000
20000
30000
40000
50000

- Highest percentage of values belongs to Unknown group and Secons belongs to Labourers
- Now taking a look to numerical values on dataset
- The data set shows that some of the values is having positive and negative value so we ghave to fix it with contains 31 coumns.
-

- we have to consider the columns like AMT_INCOME_TOTAL, AMT_CREDIT, AMT_GOODS_PRICE because this column help us to find the insights more deeply
- We have to bin the numerical columns to creat the categorical columnswe have to consider the columns like AMT_INCOME_TOTAL
- Creating bins for income amount in term of Lakhs
- Now the result should be normalize of column AMT_CREDIT and AMT_GOOD-PRICE_RANGE with 2 decimal point

- Now we are dealing with numerical columns by using describe option
- We can see that after removing positive and negative value data is clear
- In data set we can see that many values are at higher end
- It may be outliers so need to fix it and find the unique values.
- While check the data type we can see that many columns contains object data type so we need to change the data type to categorical

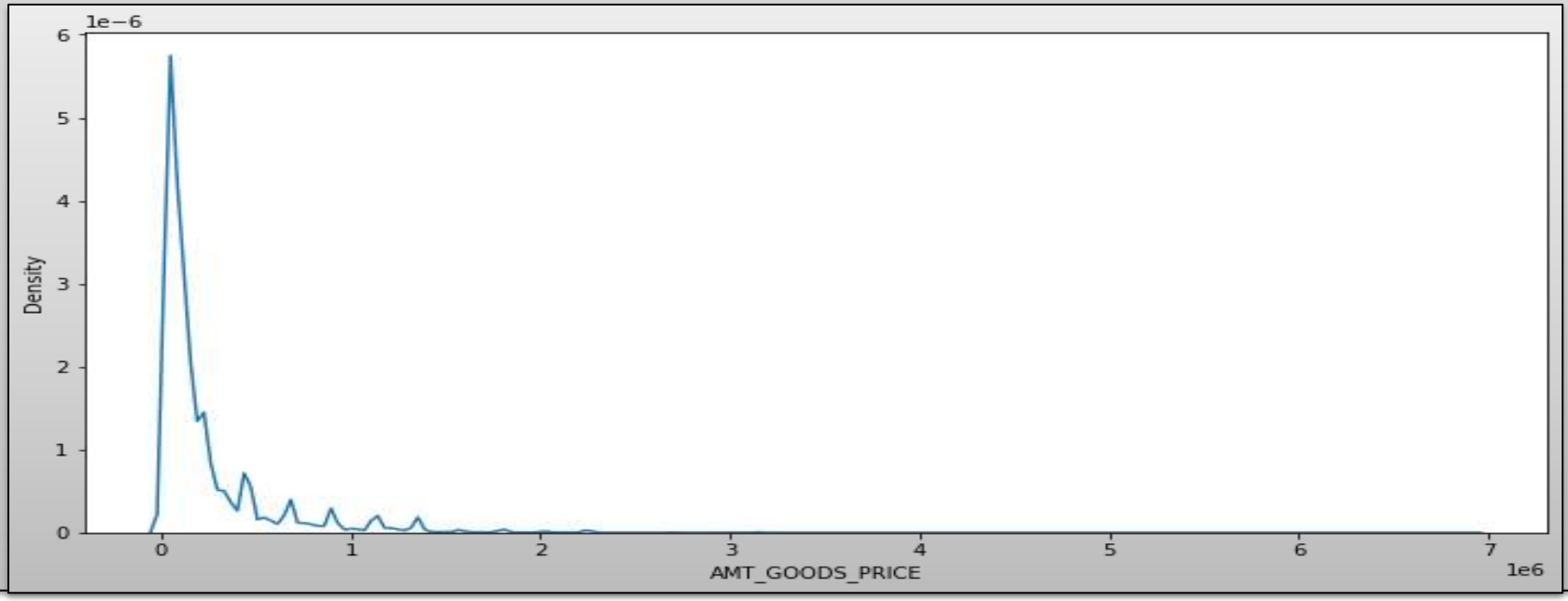
- 21 columns are converted to categorical for the further analysis
- Now the data type show categorical,int,float
- now the object type i converted to category
- now we have 49 columns for analysis this 49 columns will continue with merging previous data set
- Now we are loading the previous data set which indicate that what are the previous application for loan

- Loading the previous application data set
- Data set show that 37 columns and 1670214 rows
- We will follow the same method as previous to remove the missing values
- The missing values will be drop as per industry standard which is more than 50%
- It show only 4 columns which is having null values
- Now we can take the step that more than 12 % of column should drop

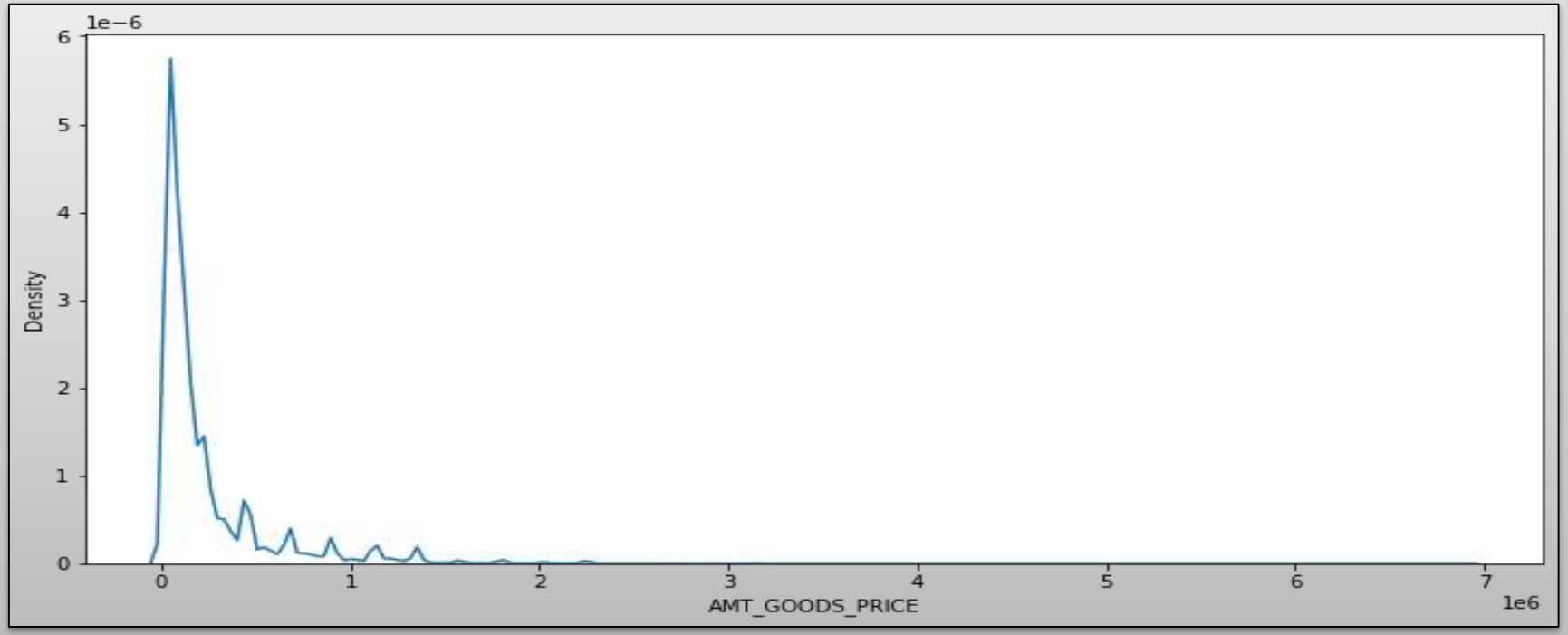
- Now after dropping more than 50% and 12 % percent it shoe 33 columns and 1670214 rows
- Again unwanted column are dropped because there is no use in further analysis
- Now the columns are 29 and the rows are 1670214
- missing values in columns 'DAYS_FIRST_DUE', 'DAYS_TERMINATION', 'DAYS_FIRST_DRAWING','DAYS_LAST_DUE_1ST_VERSION', 'DAYS_LAST_DUE' and these columns count days thus will keeping null values as it is

- In prev_app data set there are some negative values so we have to convert to positive values then and only then we can do analysis with normal values
- On days _decision column we can normalize the values by multiplying by 100 to do the smooth calculation
- here we can clearly see that 35% applicant have applied for new loan in 1 year which shows yearly decision

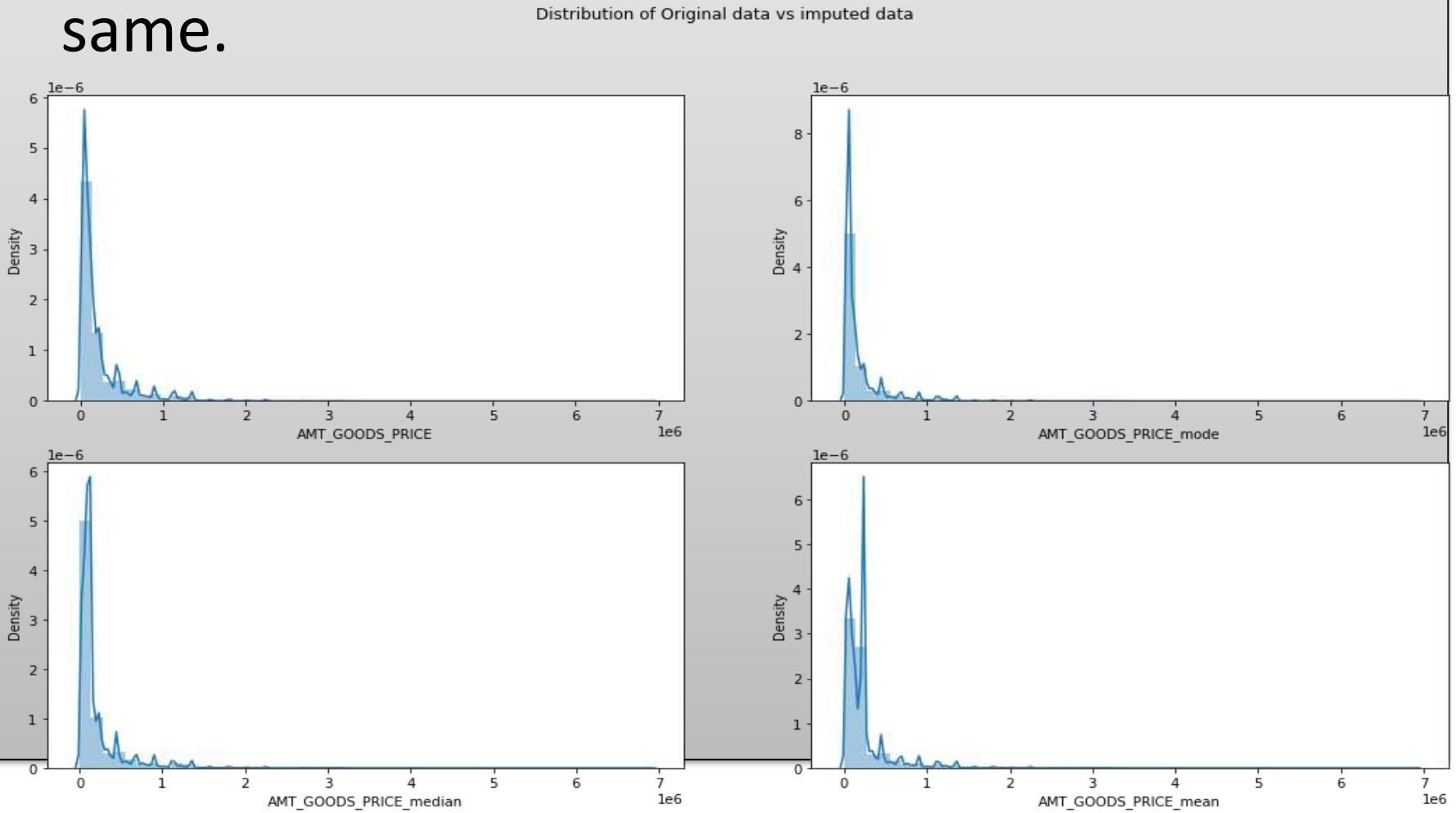
- consider the continous variable"AMT_ANNUITY", "AMT_GOODS_PRICE“
- in continous variable we can see null values by using plot
- if the median if the distribution is skewed
- mode if the distribution pattern is preserved.



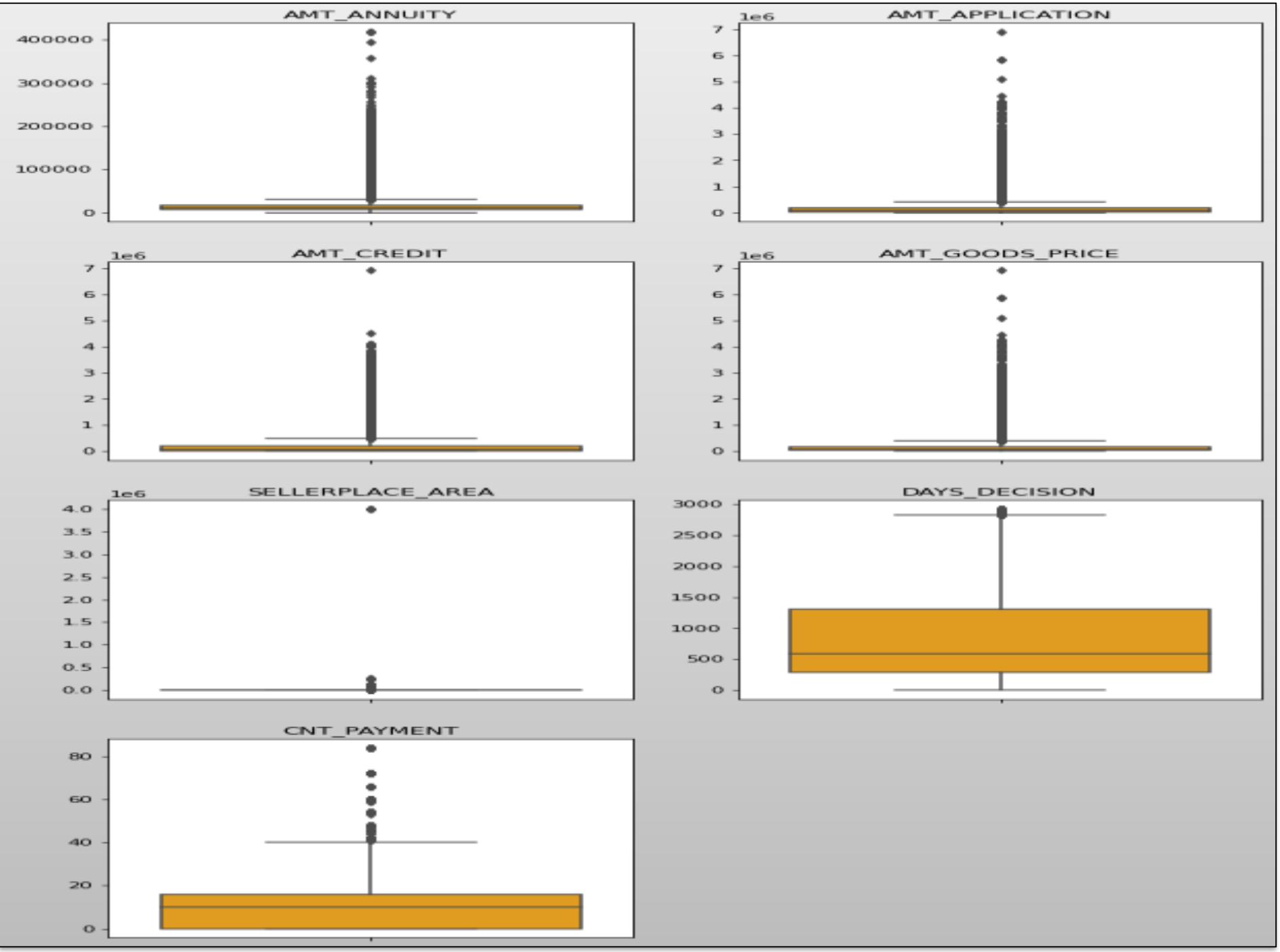
- we can clearly oserved that the skewness is at left side so there is some outliers so we will go with median
- imputing missing values with median
- for "AMT_GOODS_PRICE" to understand the distribution refer graph.



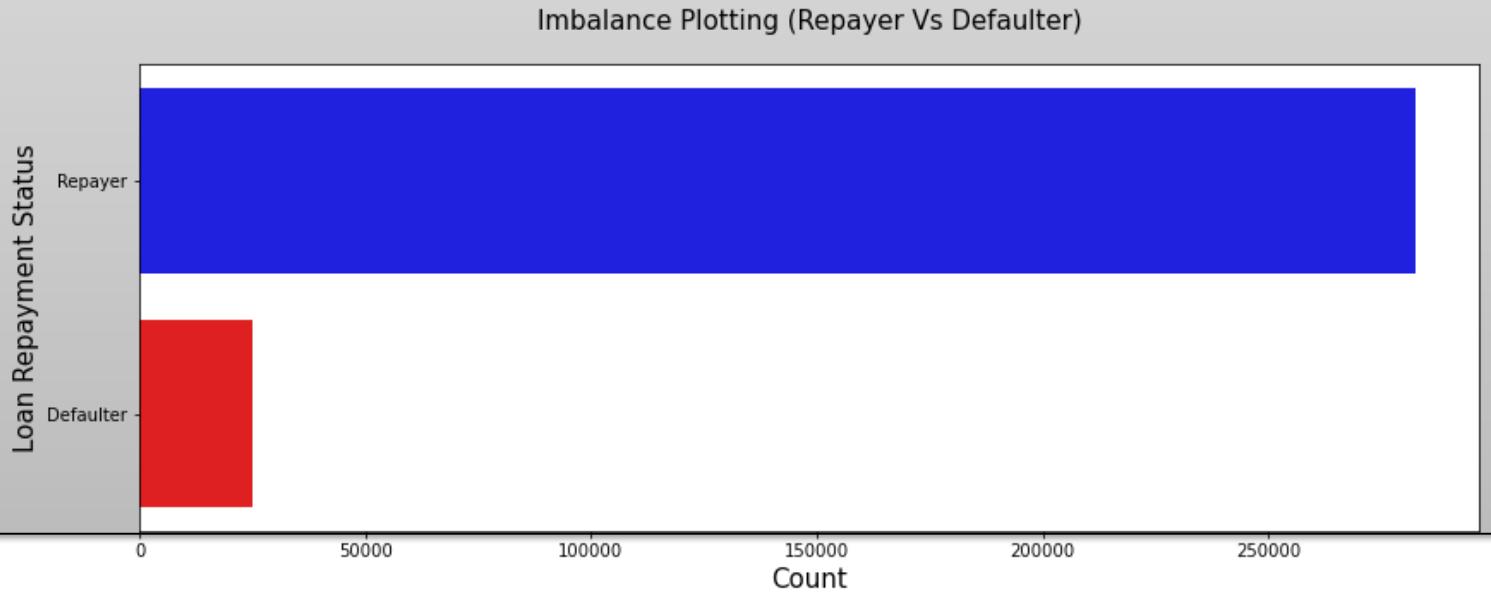
- There are several peaks along the distribution. Let's impute using the mode, mean and median and see if the distribution is still about the same.



- replace null values with mode
- find the valu count for name_contract_status
- required categoical columns converted from Object to categorical for further analysis
- And after converting we are doing outlier analysis because after using describe we can see that there are negative and positive values
- To check the outlier we can use box plt

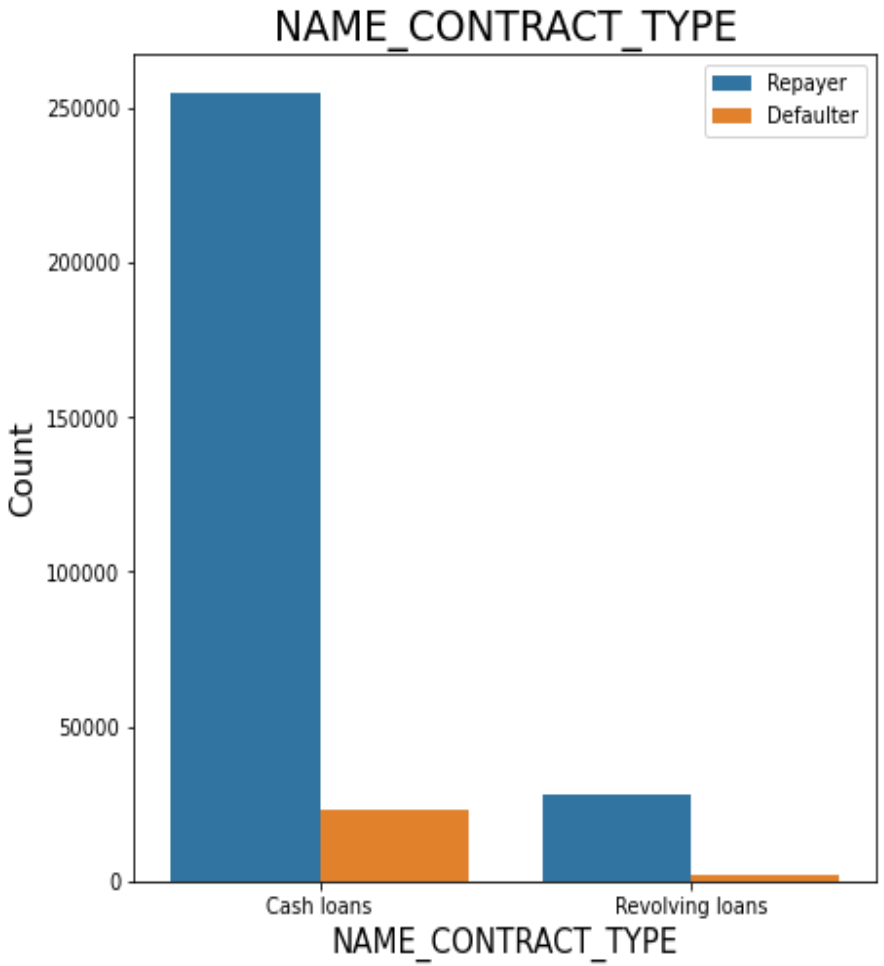

AMT_ANNUITY
400000
300000
200000
100000
0
AMT_APPLICATION
1e6
7
6
5
4
3
2
1
0
AMT_CREDIT
1e6
7
6
5
4
3
2
1
0
AMT_GOODS_PRICE
1e6
7
6
5
4
3
2
1
0
SELLERPLACE_AREA
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
DAYS_DECISION
3000
2500
2000
1500
1000
500
0
CNT_PAYMENT
80
60
40
20
0

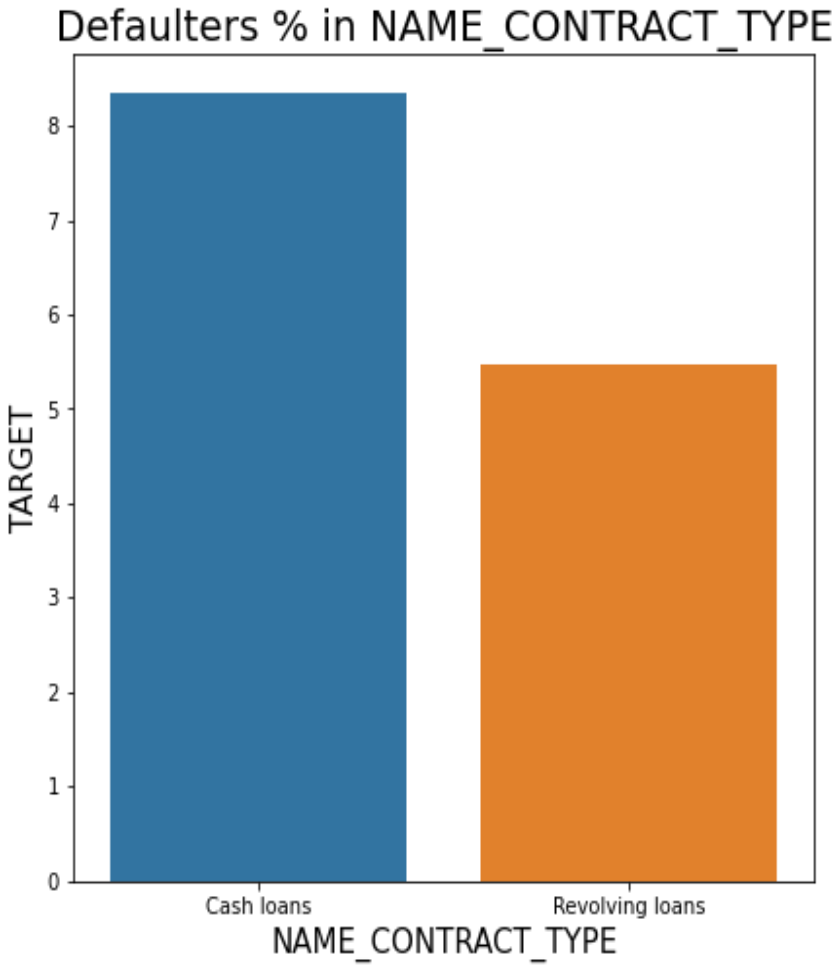
- Observation
- 1) previous application data contains outliers as per above graph
- 2) CNT_PAYMENT has few outlier values.
- 3) DAYS_DECISION has little number of outliers
- Now we are analyzing imbalance data



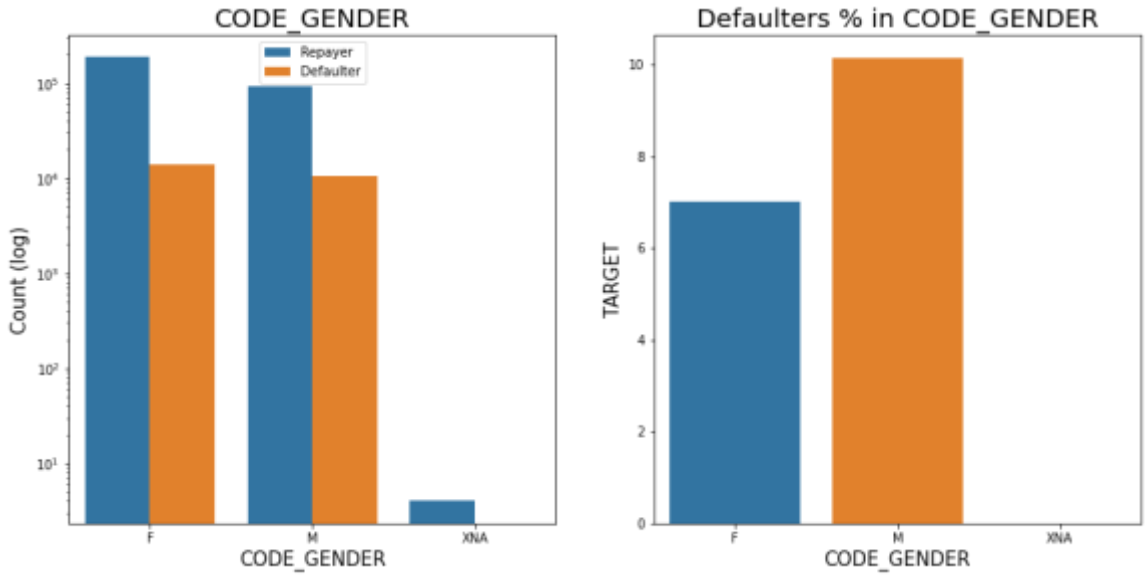
- Repayer Percentage is 91.93% Defaulter Percentage is 8.07% Imbalance Ratio with respect to Repayer and Defaulter : 11.39/1 (approx)
- Now will go with single univariate analysis to find if the column is categorical or numerical
- Parallel we can see bivariante analysis by using various type of graps

plotting a graph contract type based on loan repayment status

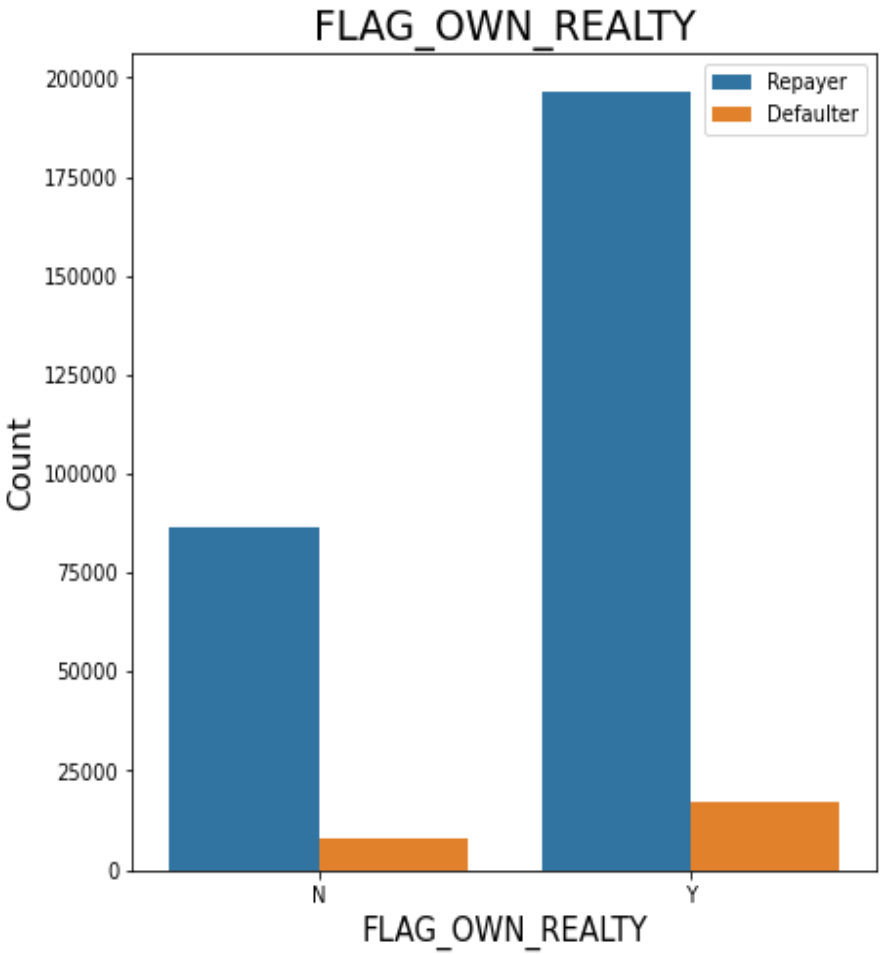


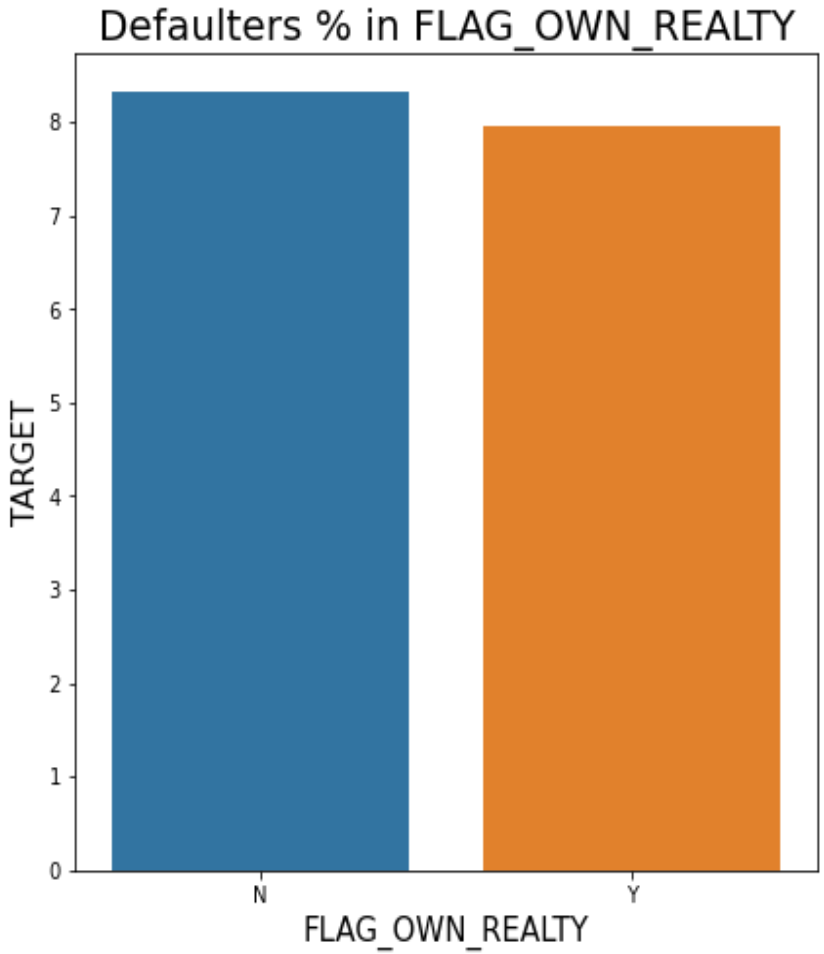


- here on above graph we can see that revolving lone is just 10% from the total loan
- # cash loan applicant show 8.5% and in defaulter it shows range between 5% to 6%
- plotting the graph for the type of Gender on loan repayment status



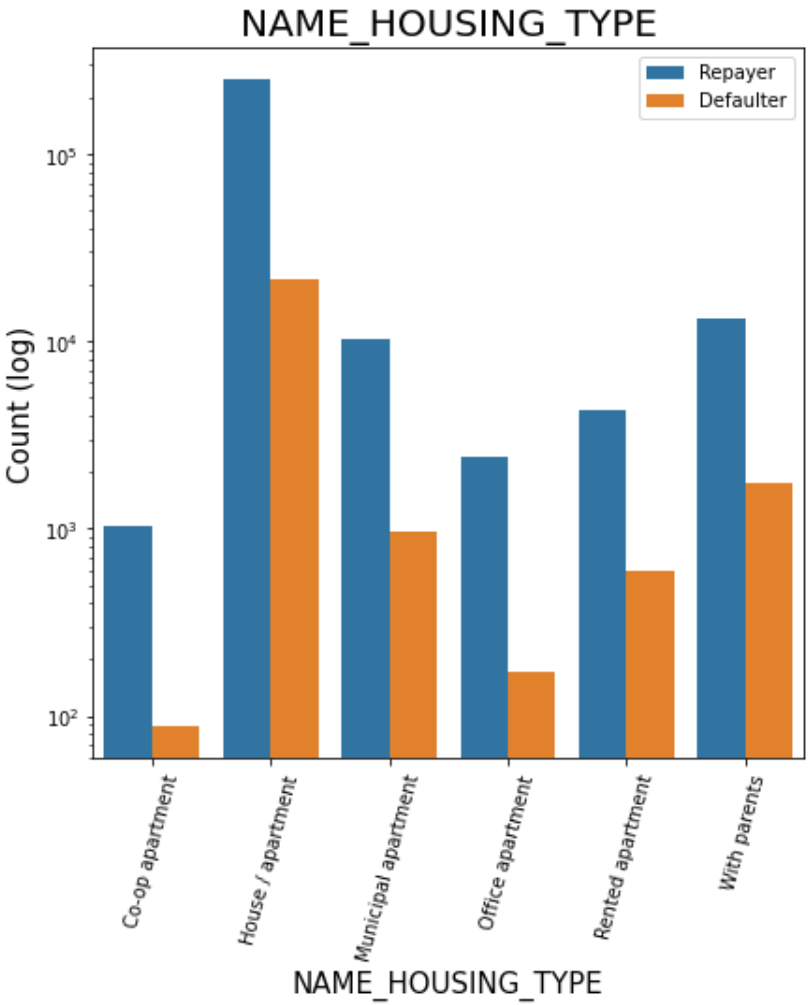
comparing with owning a real estate is related to loan repayment status

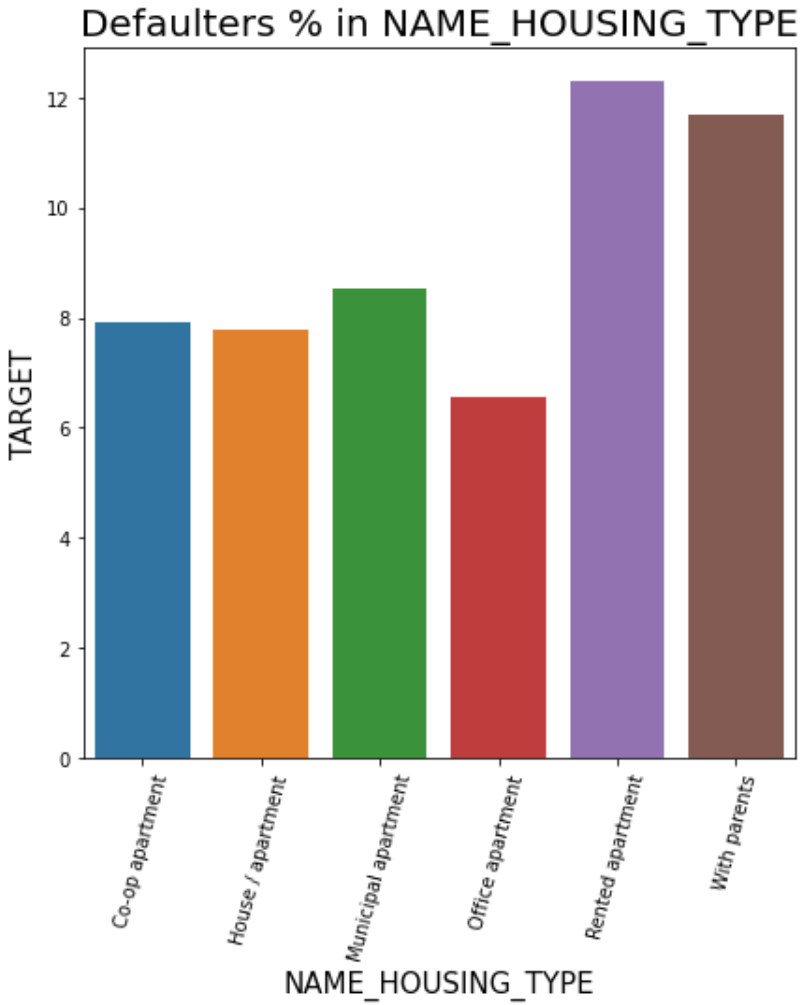




- observation:1) we can see in above graph that who own the real estate are more compare to dont own
- 2) defaulter show near 8% so it conclude that there is not strong relation between own reality and defaulter

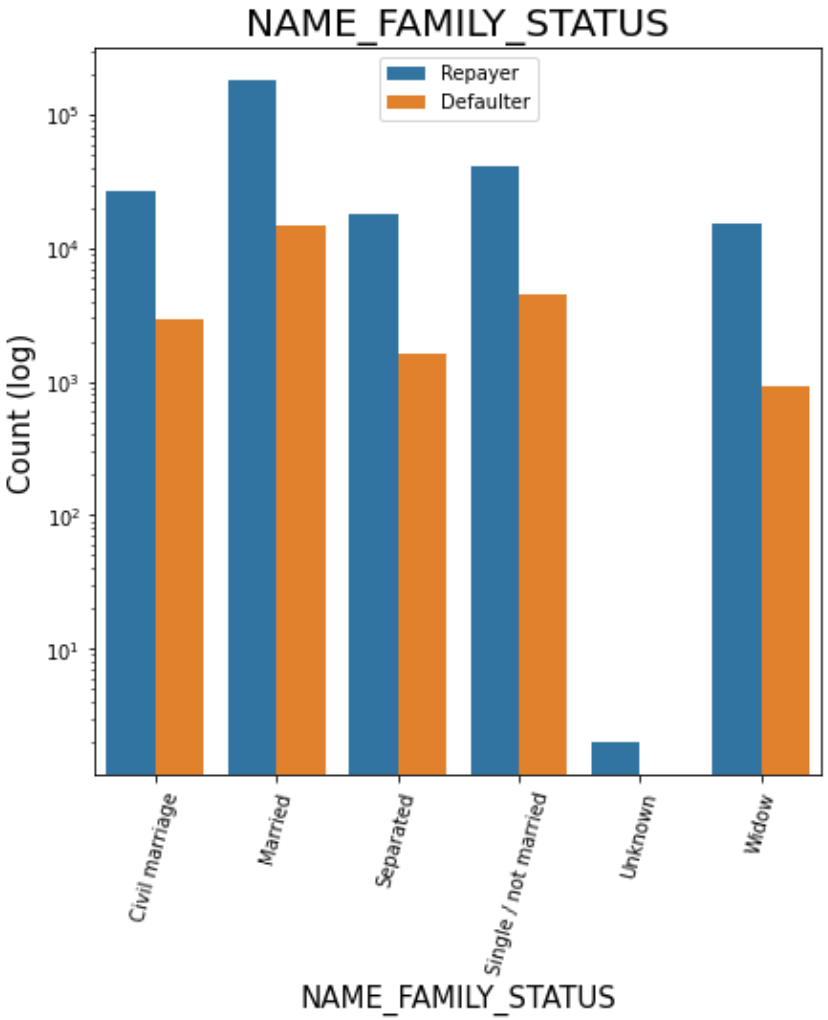
comparing between Housing Type based on loan repayment status

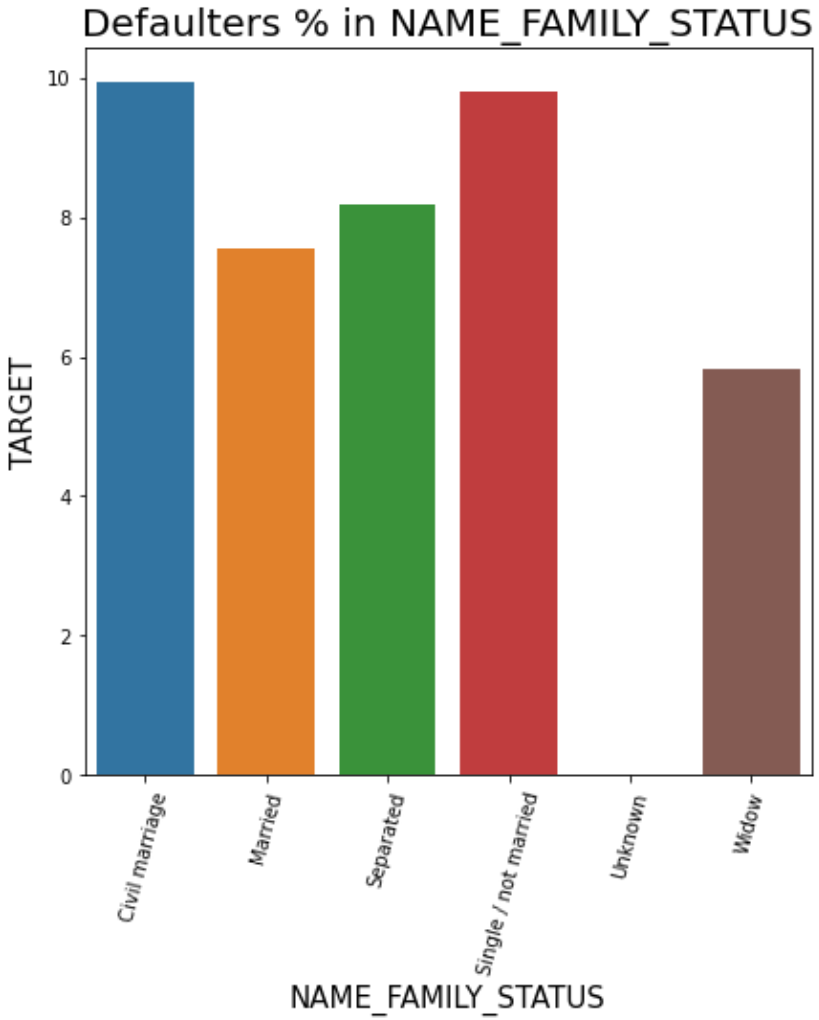




- the above graph show living in house apartment is high
- people living in co- apartment is low
- people living with parents is near to 12 % and living in rented is more than

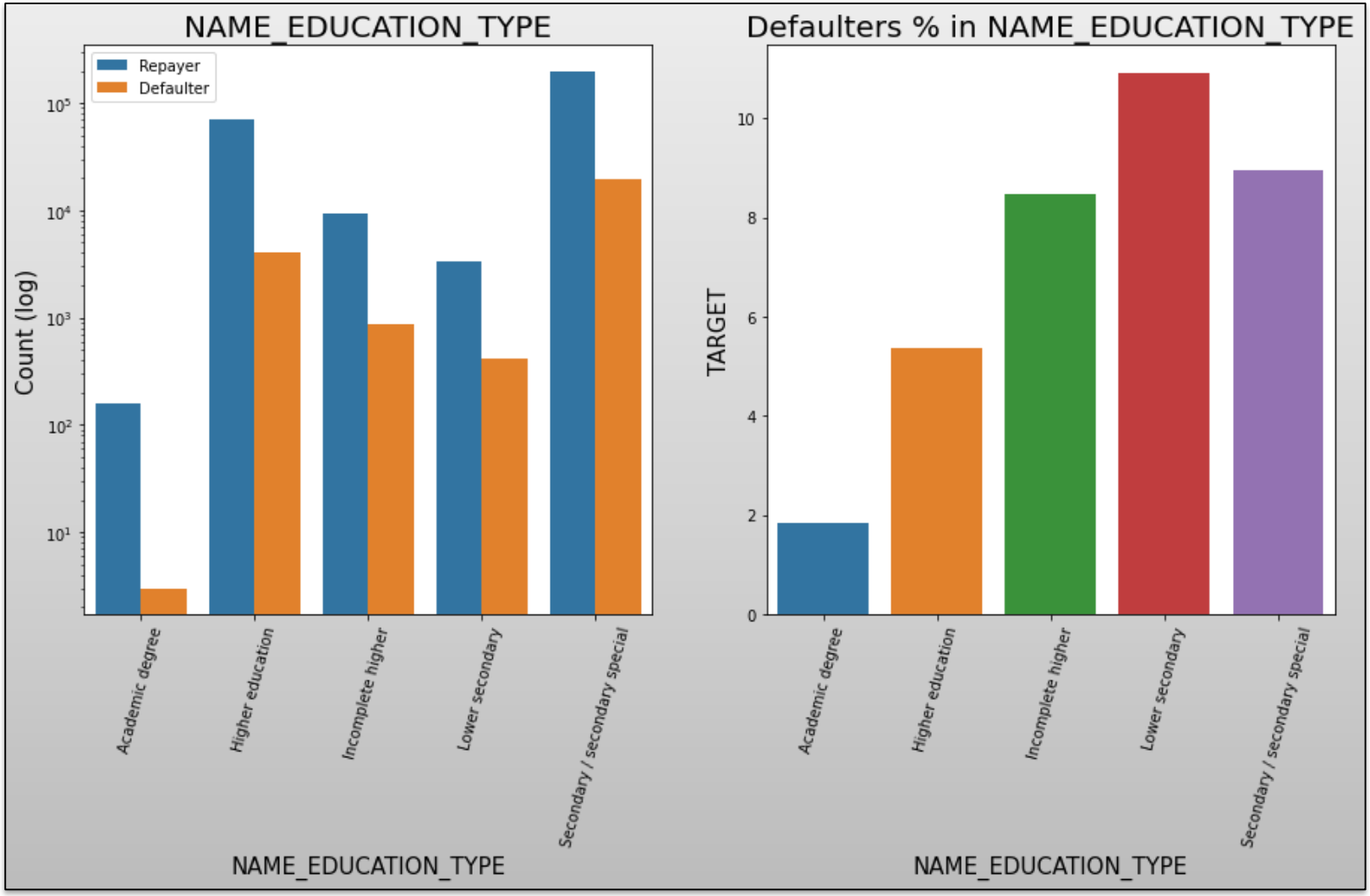
comparing with Family status based on loan repayment status





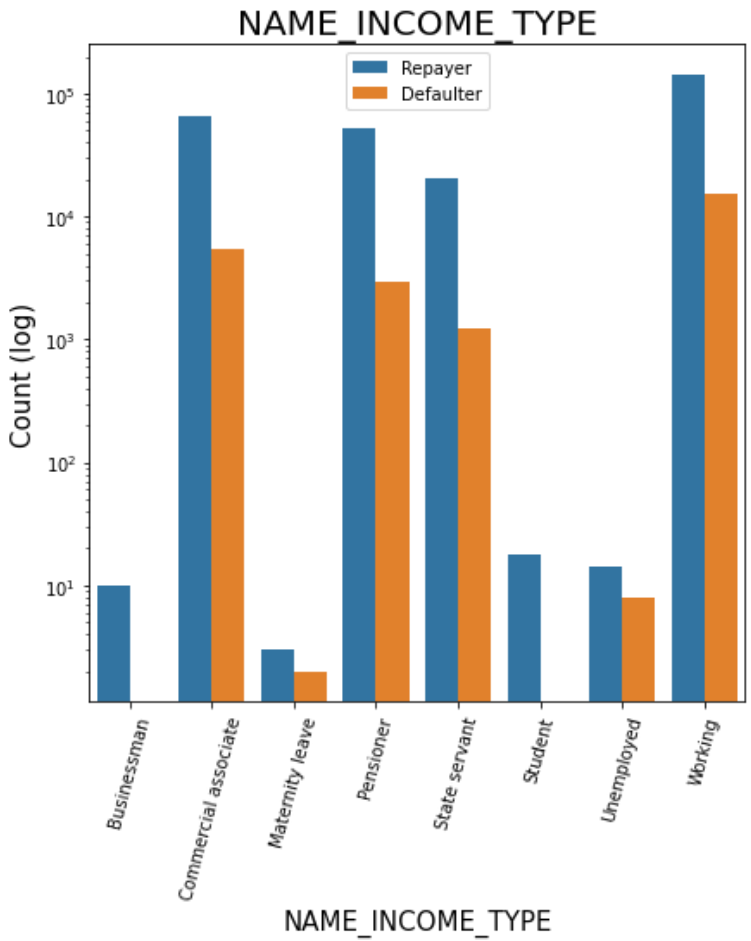
- the above graph show that most of the people is married who is taking lone
- # civil marriage in defaulter is near to 10% which is high as compare to other
- # widow is at lower percentage near to 5 %

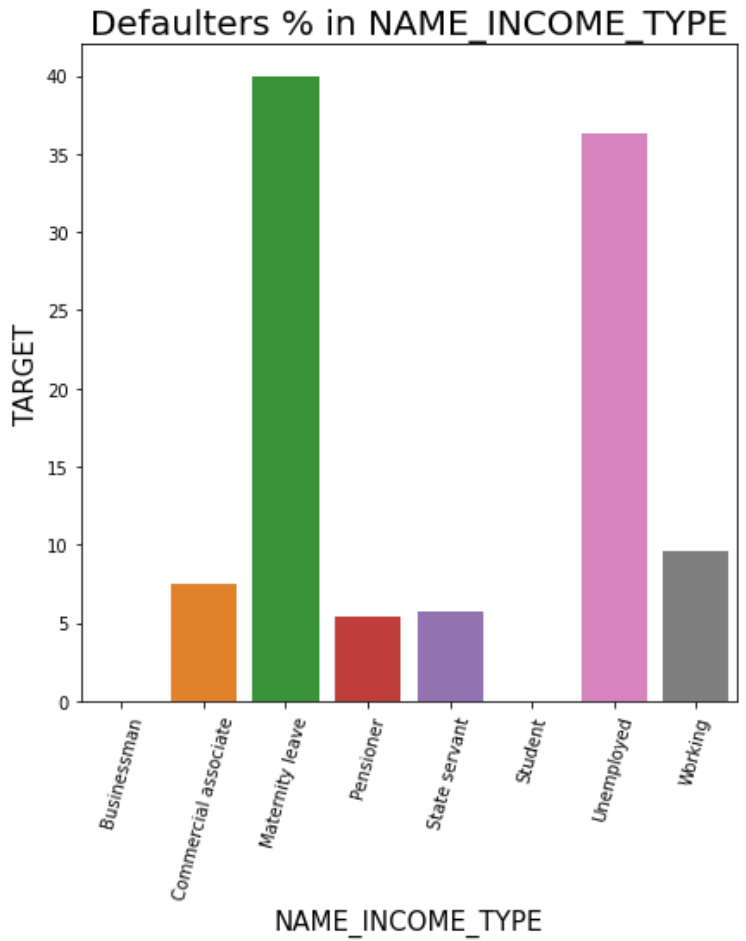
comparing with Education Type based on loan repayment status



- the above graph show that secondary special education, followed by clients with Higher education.
- academic degree is at lower side who carry the degree
- in defaulter lower secondary is at higher range
- in defaulter academic degree is at lower side of range

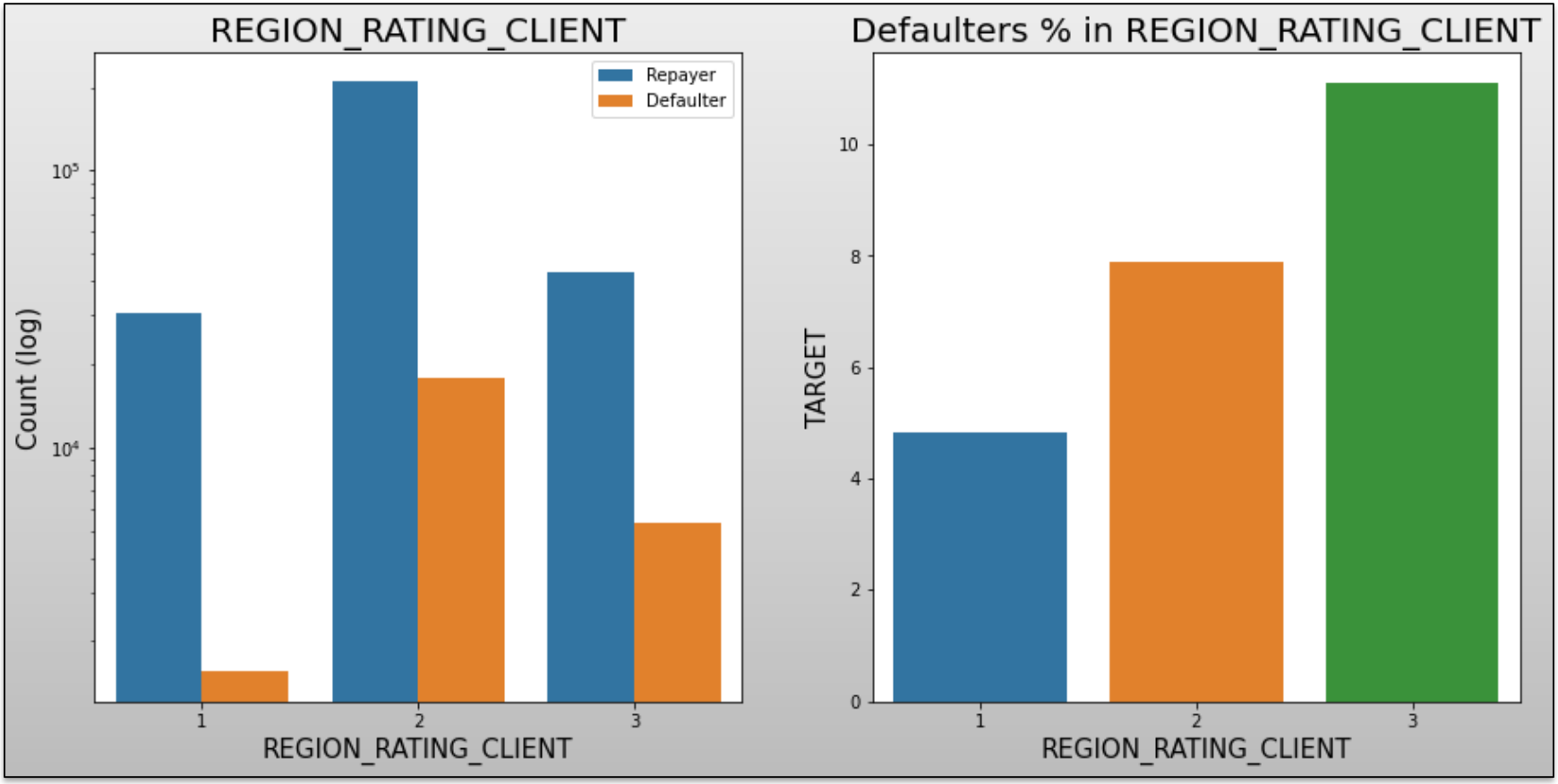
comparing with Income Type based on loan repayment status





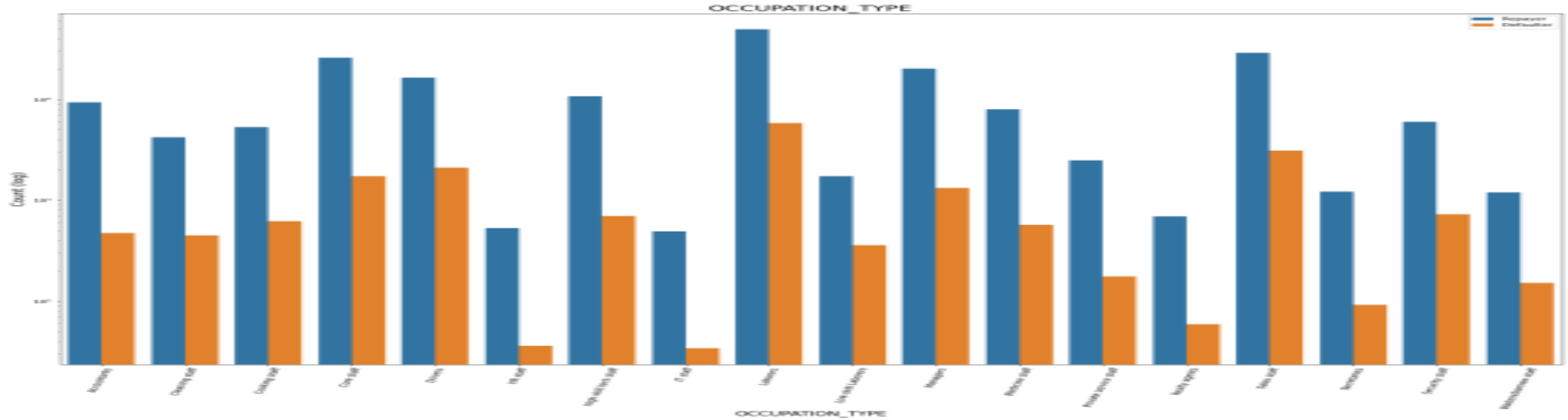
- the above graph show that most of the applicant are working
- maternity leave is at higher side in defaulter near to 40%
- unemployed are at 36 % in defaulter
- at the second position we can see the commercial associate

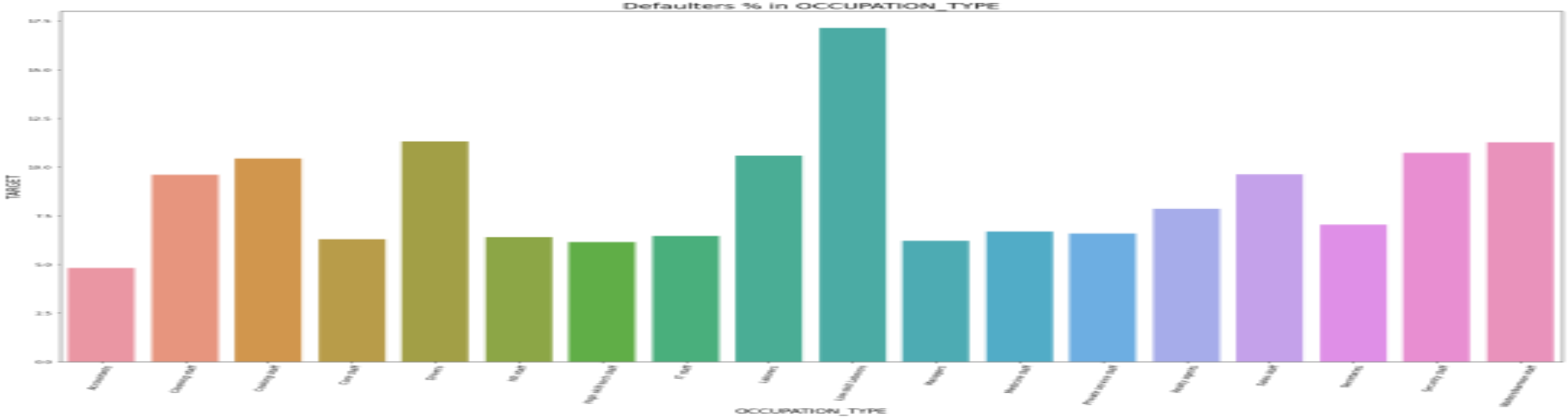
comparing with Region rating where applicant lives based on loan repayment status



- most of the applicant are living in region with 2 rating
- but in defaulter with rating 3 is at higher side
- at the lower side is region rating 1

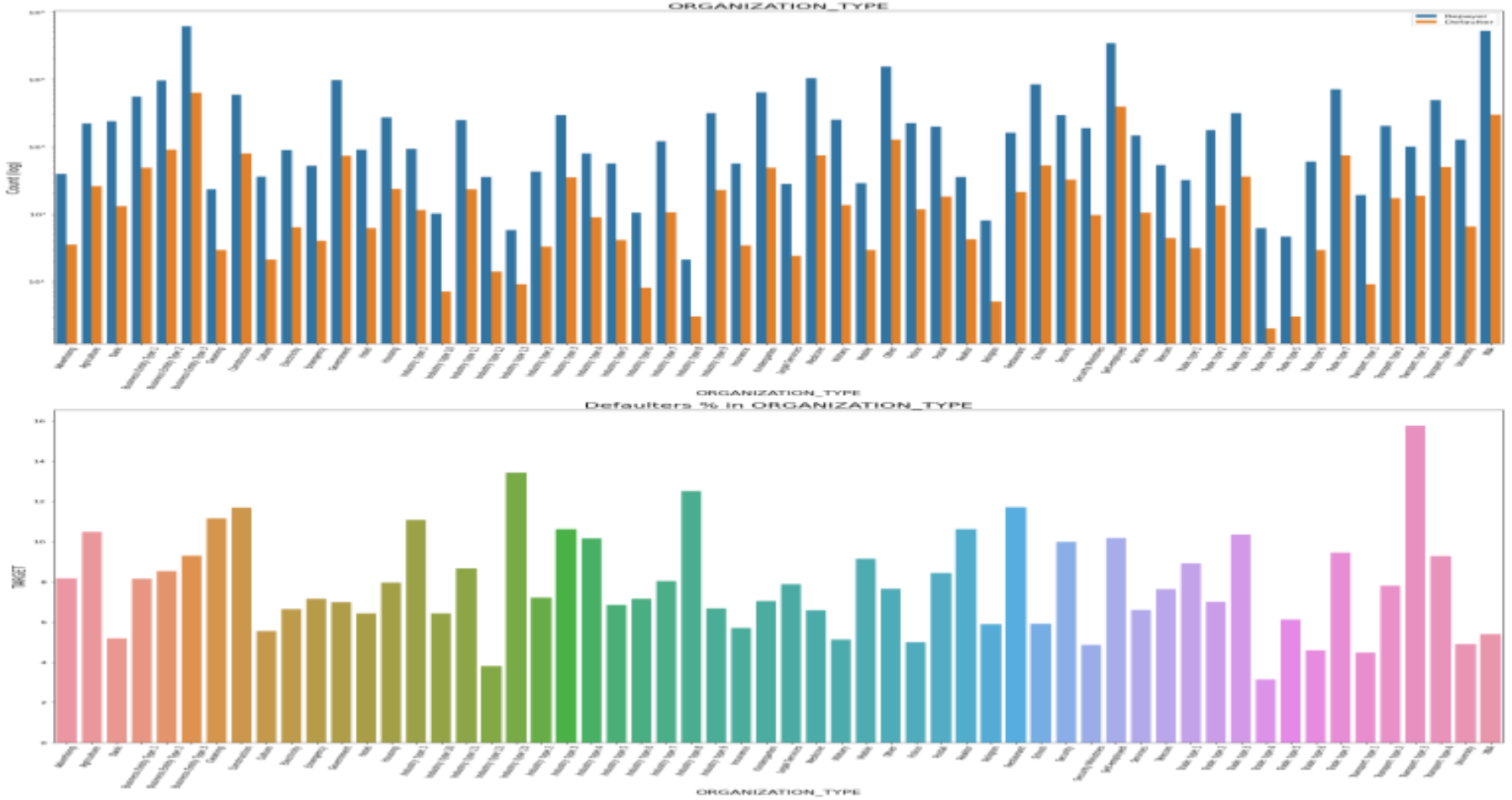
comparing with Occupation Type where applicant lives based on loan repayment status





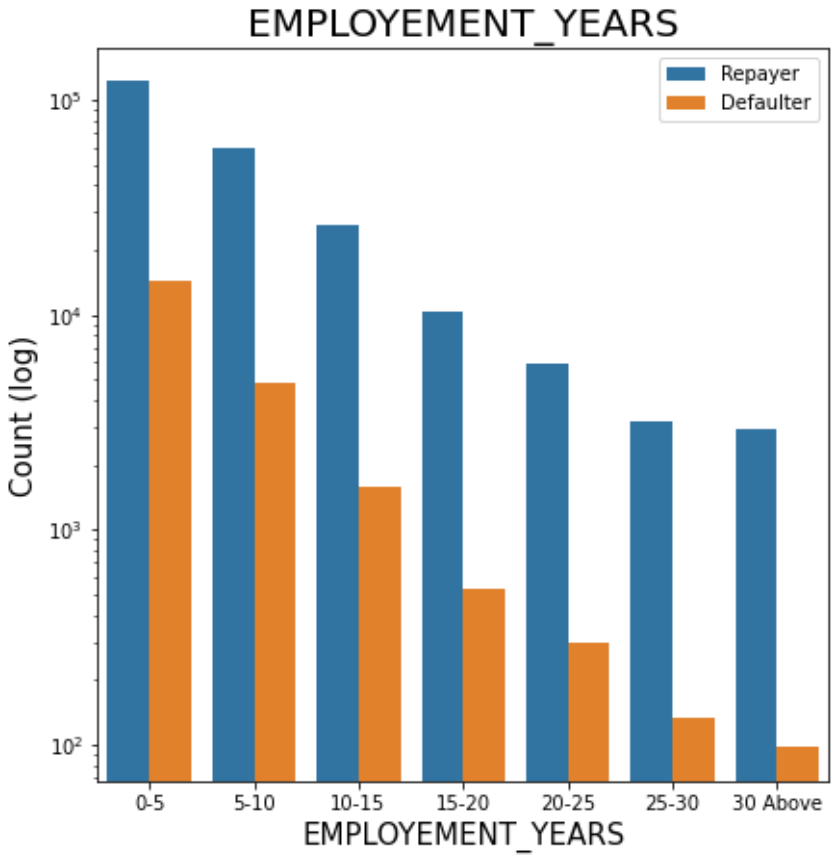
- from the above graph we can see that taking lone by labour is at higher
- IT staff is at lower side
- in defaulter low skill labour is at higher side

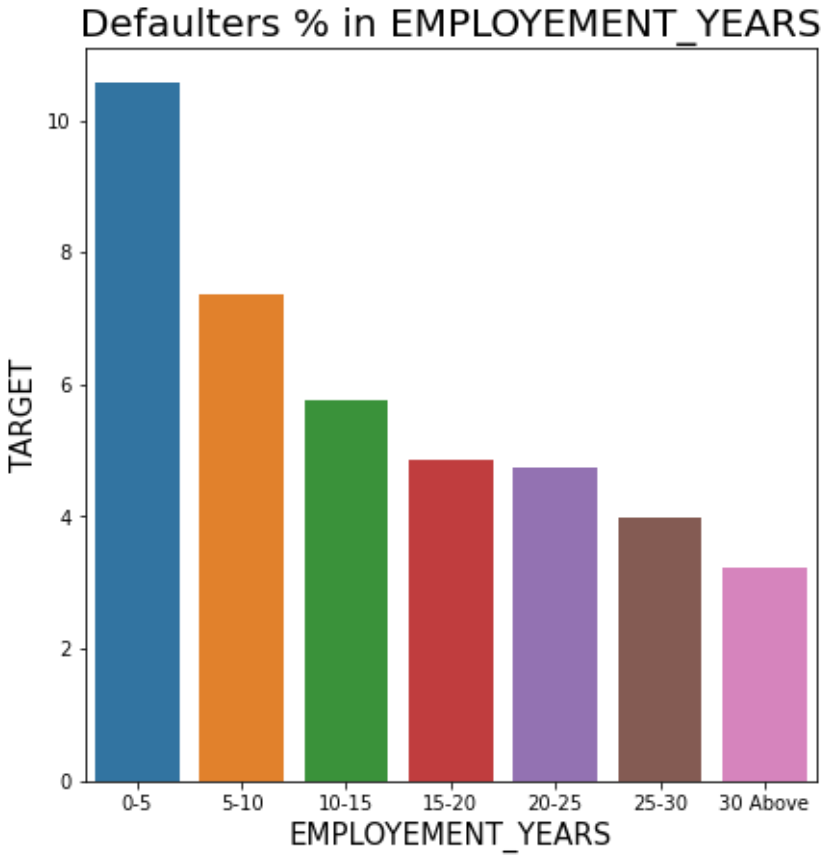
comparing with Loan repayment status based on Organization type



- from the above grapg we can see that in defaulter organization is at higer side that is transporter type 3 with 15 % ndustry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%).
- applicant take lone from business entity type 3 is high
- xna is also show high percentage where information is unavailable
- category of organization type has lesser defaulters thus safer for providing loans: Trade Type 4 and 5, Industry type 8

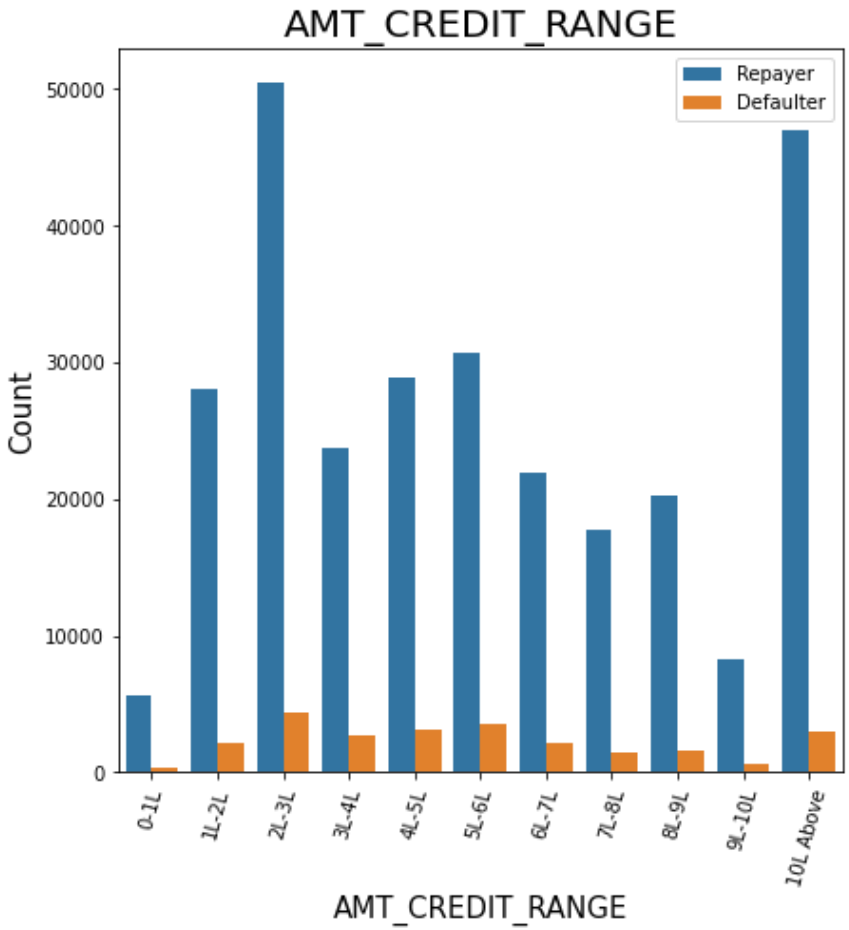
comparing with Employment_Year based on loan repayment status

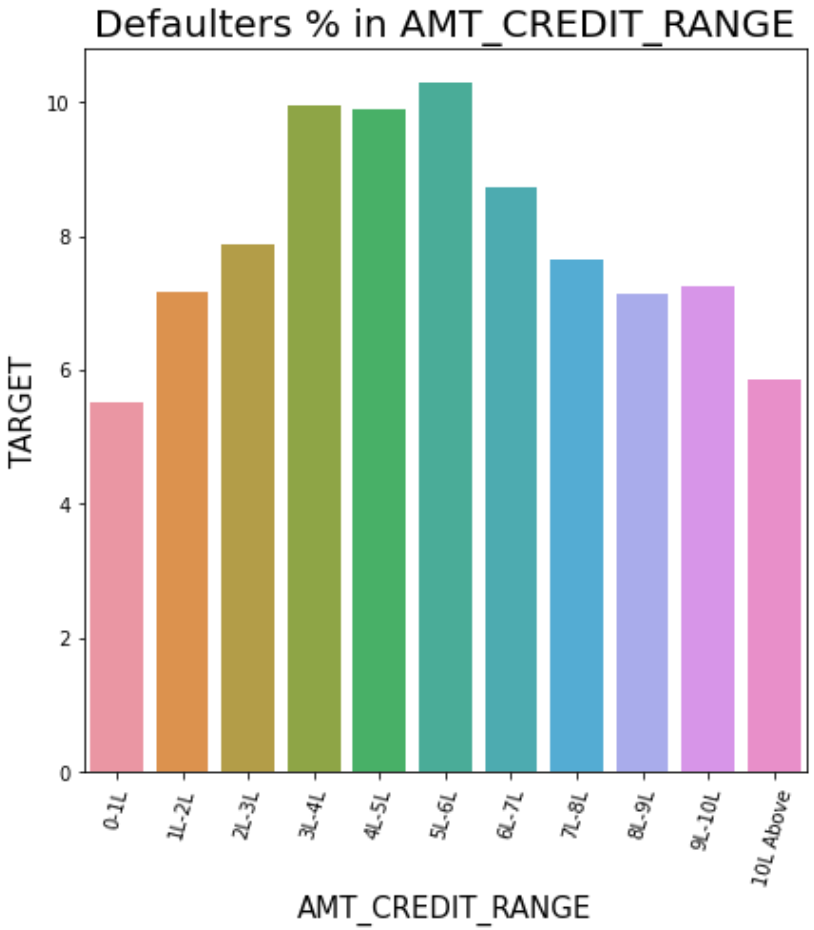




- working experience sow higher from range 0-5
- and in defaulter working is also higher
- as the year increase the defaulter rate decreases
- 30 plus year eperience is less than 1 %

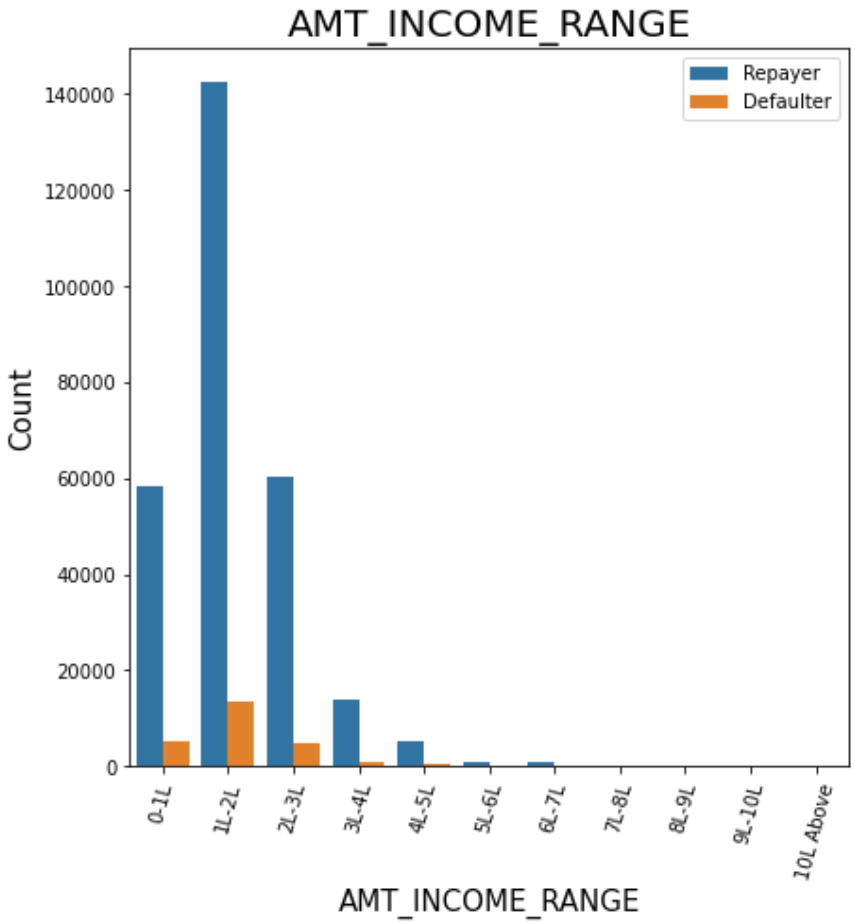
comparing with Amount_Credit based on loan repayment status

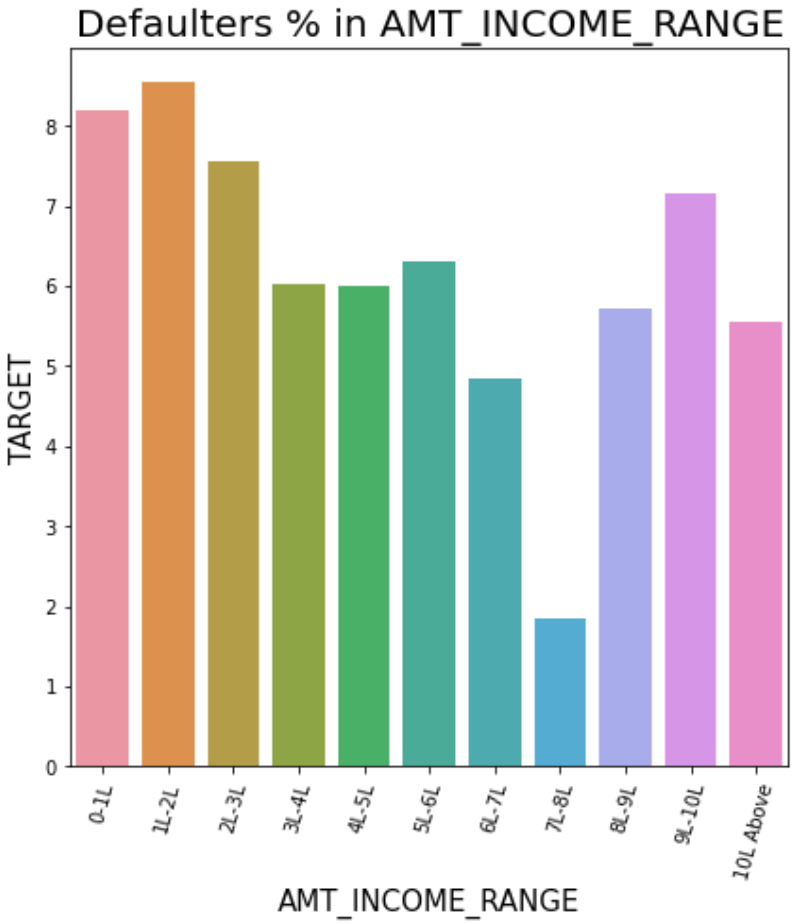




- as per the above graph applicants have loan in range of 2-3 Lakhs
- # who get loan for 3-6 Lakhs have most number of defaulters than other loan range

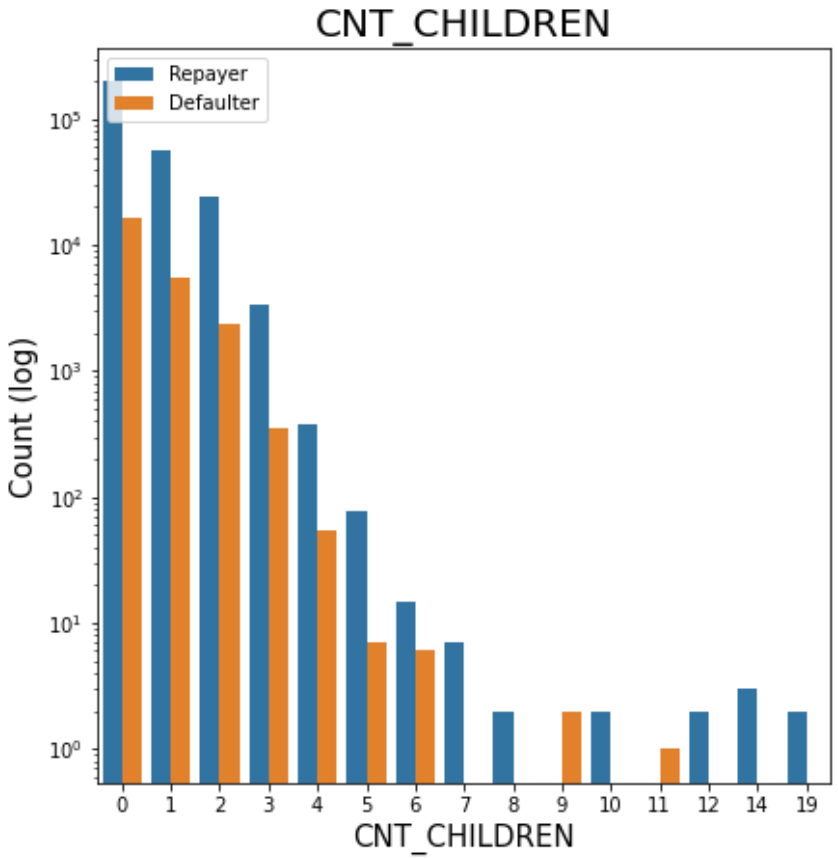
comparing with Amount_Income Range based on loan repayment status

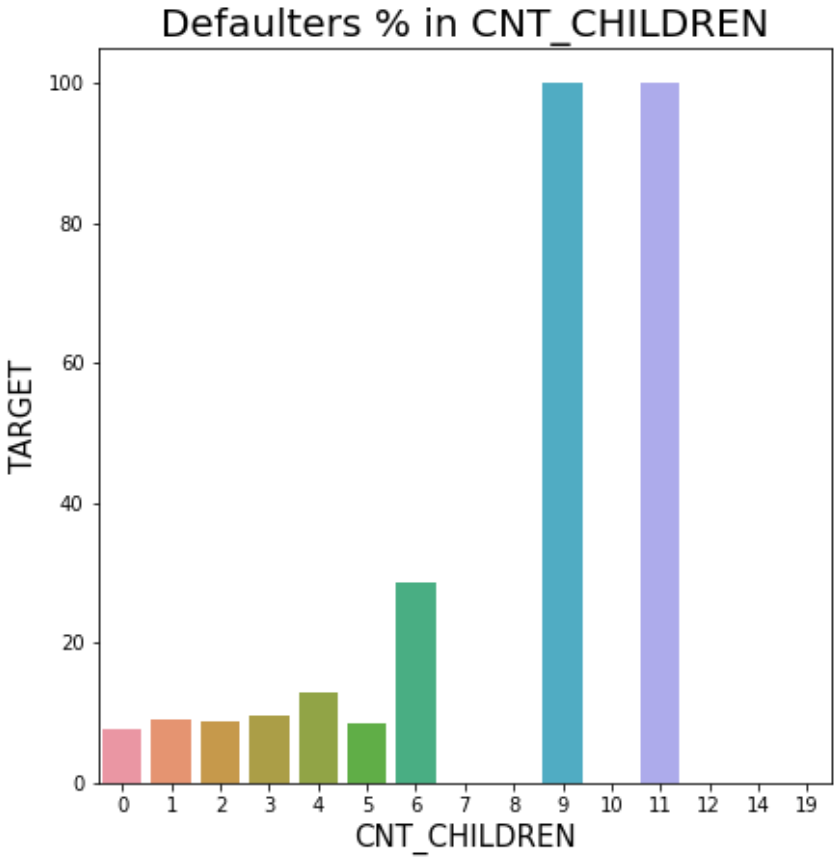




- the above graph shows that applications have Income total in between 1-2 lacks is higher
- in defaulter applications have Income total in between 1-2 lacks is higher
- applicant with 7-8 lacks is lower

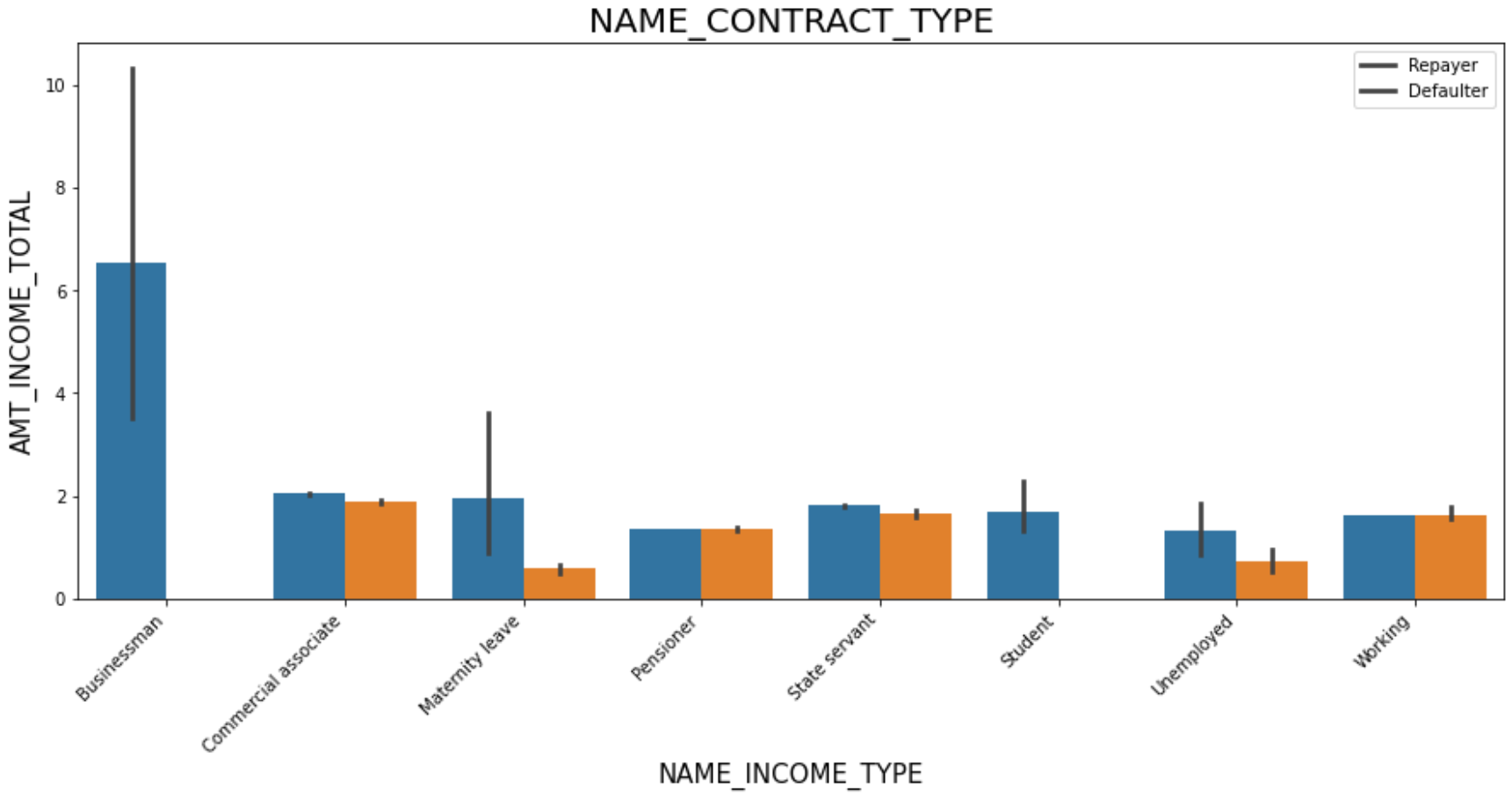
comparing with Number of children based on loan repayment status





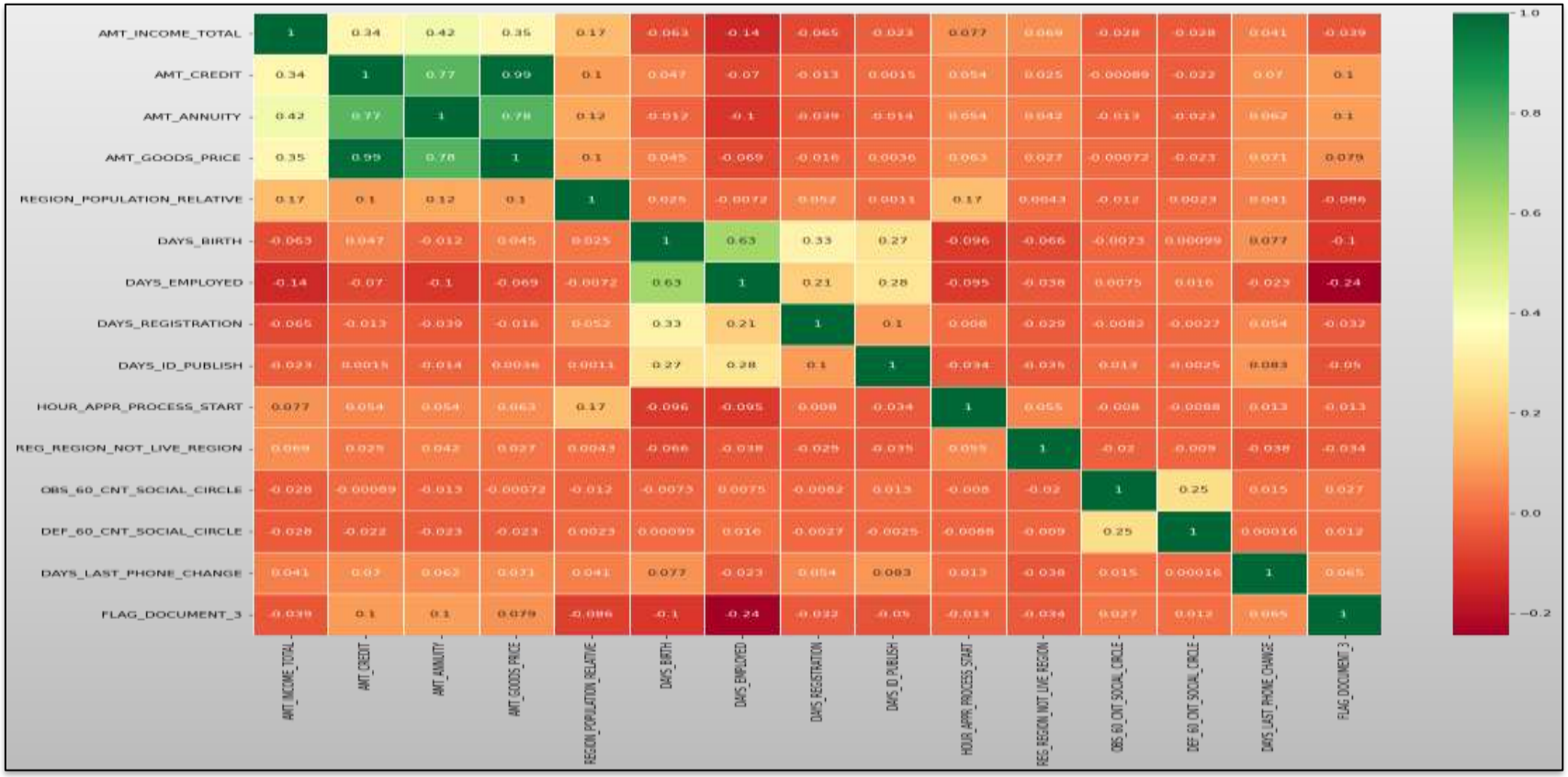
- above graph shows that applicant do not have childerens is higher
- in default graph children more than 4 is high with 9 and 11 and is show 100%

Income type vs Income Amount Range on a Seaborn Barplot



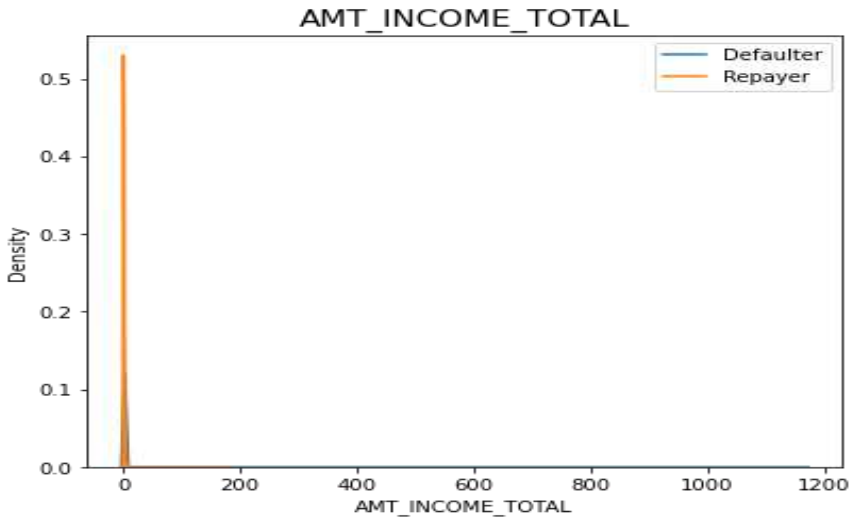
- from the above graph we can see that busines man has higher income range from 3.5 laks to above 10 lacks
- checking for all the columns in data
- We can see top 10 correlation for the Repayers data frame

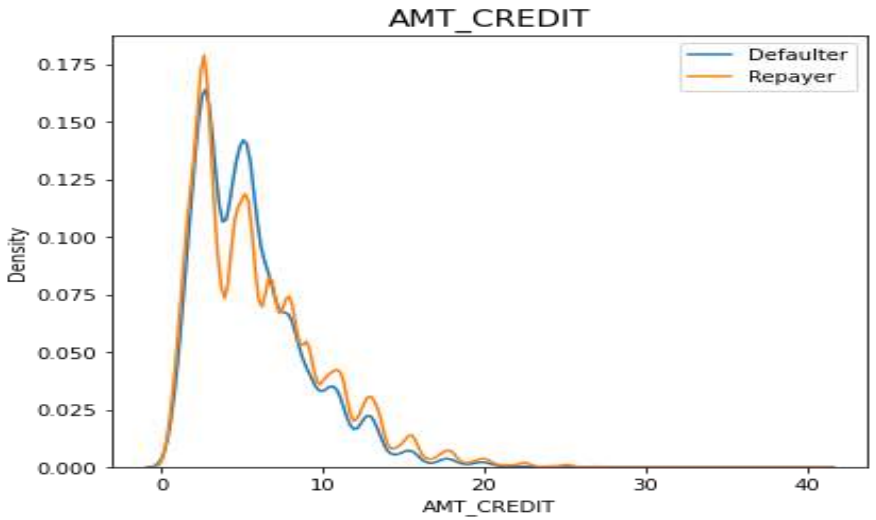
heatmap to see linear correlation amoung Repayers Credit amount is highly correlated with:Goods Price Amount,Loan Annuity,Total Income

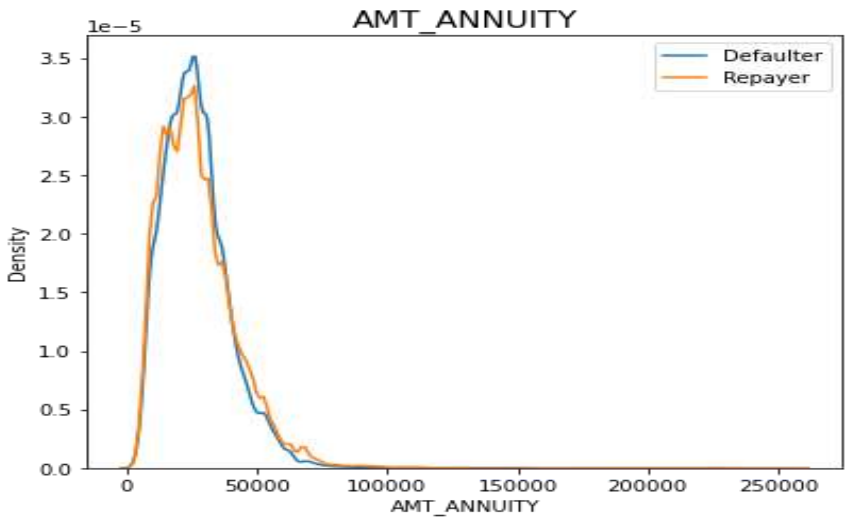


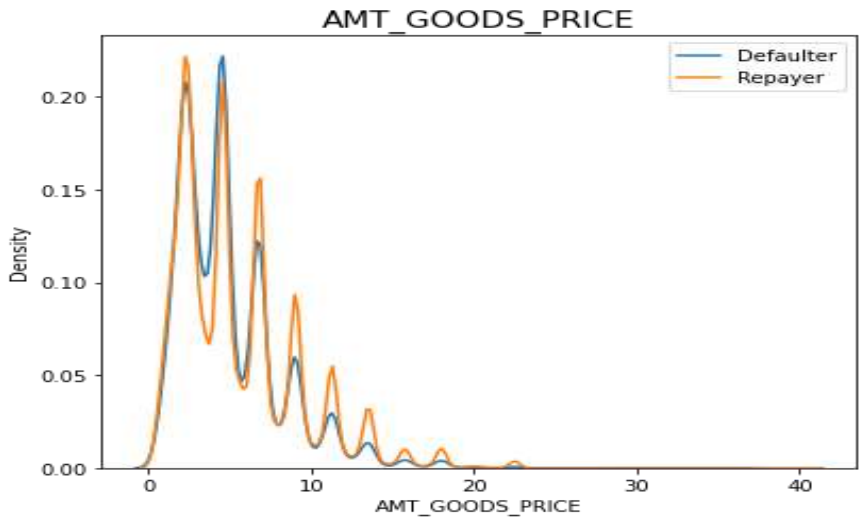
- the above graph show that Credit amount is highly correlated with good price amount which is same as repayers.
- Loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to repayers(0.77)
- We can also see that repayers have high correlation in number of days employed(0.62) when compared to defaulters(0.58).
- Days_birth and number of children correlation has reduced to 0.259 in defaulters when compared to 0.337 in repayers.

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

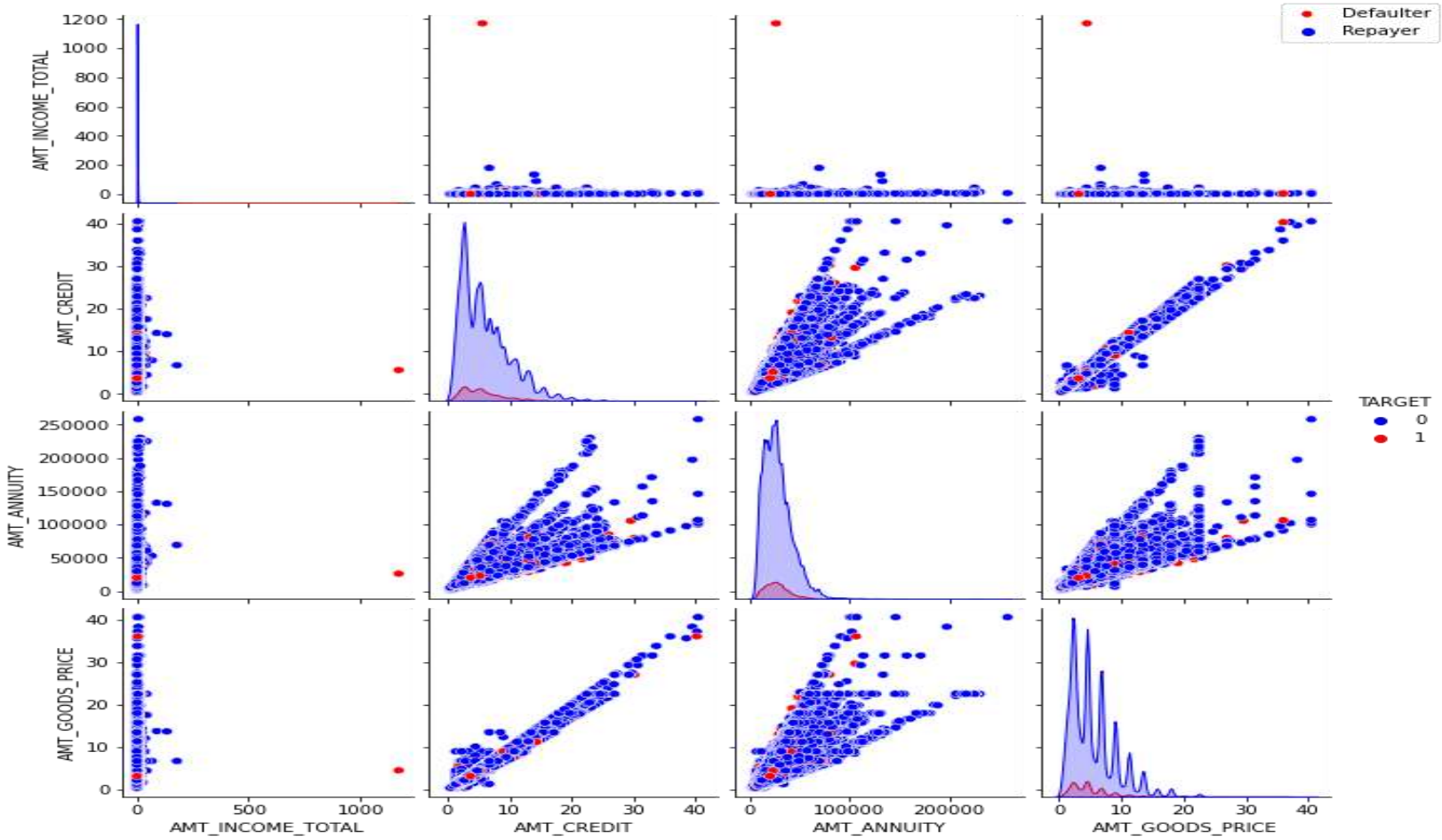








comparing pairplot between amount variable to draw reference against loan repayment status



- **Conclusions**

- The analyses of the datasets indicate that a bank can determine with a few attributes whether a client can repay a loan or not. The analysis is consisted as below with the contributing factors and categorization

- **An applicant's ability to repay**

- NAME_EDUCATION_TYPE: There are fewer defaults for academic degrees.
- # NAME_INCOME_TYPE: There are no defaults for students or businessmen.
- # REGION_RATING_CLIENT: 1 is the safest rating.
- # ORGANIZATION_TYPE: There have been less than 3% defaults for clients with Trade Types 4, 5 and 8 in addition to Industry Type 8.
- # DAYS_BIRTH: It is less likely for people over 50 to default on their loans
- # DAYS_EMPLOYED: Lower default rate among clients with 40+ years of experience
- # AMT_INCOME_TOTAL: Higher incomes are less likely to default than lower incomes
- # NAME_CASH_LOAN_PURPOSE: Hobby loans and garage loans are usually repaid.
- # CNT_CHILDREN: Those with one to two children are more likely to repay the loan.

The following factors determine whether an applicant will be a defaulter:

- CODE_GENDER: Men are more likely to default than women
- NAME_FAMILY_STATUS : Singles and couples with civil marriages default frequently.
- NAME_EDUCATION_TYPE: Individuals with a lower secondary or secondary education
- NAME_INCOME_TYPE: A lot of defaults are attributed to clients on maternity leave or unemployed.
- REGION_RATING_CLIENT: The default rate is highest for people who live in Rating 3.
- OCCUPATION_TYPE: Avoid Low-skill Labourers, Drivers, and Waiters/barmen staff, Security staff, Labourers, and Cooking staff as their default rate are huge.
- ORGANIZATION_TYPE: Organizations with the 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 relatively high defaulting rate, and thus should be avoided to be approved for loan or provide loan with higher interest rates to mitigate the risk of default.

- DAYS_BIRTH: Avoid young people who are in age group of 20-40 as they have higher probability of defaulting
- DAYS_EMPLOYED: People who have less than 5 years of employment have high default rate.
- 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.
- AMT_GOODS_PRICE: When the credit amount goes beyond 3lakhs, there is an increase in defaulters.

Loan can be given on Condition of High Interest rate any default risk leading to business loss:

- 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.
- # 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.
- # 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.
- # 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.
- # 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.

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