Exploratory Data Analysis

Bank Loan Credit Analysis

A Case study by :

Amrita Chatterjee
Padma A.

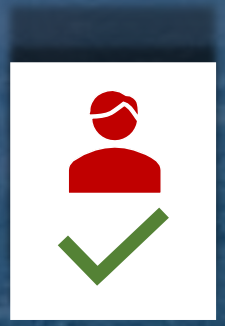
Business Objective

Approve the loan of applicants who can repay the loan

Analysis Approach

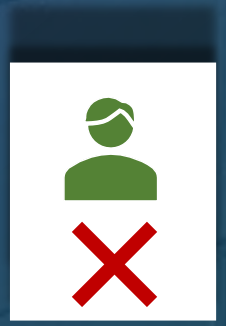
Identify the driving factors behind loan default

Risk Associated with Decision



Financial Loss

If the application is not likely to repay the loan, i.e. he/she is likely to default; then approving the loan may lead to financial loss for the bank



Opportunity Loss

If the application is likely to repay the loan, then rejecting the application will result in loss of business for the bank

Dataset

1. 'application_data.csv'
   - contains all the information of the client at the time of application
   - shows whether a client has payment difficulties

2. 'previous_application.csv'
   - contains information about the client's previous loan data
   - contains the decision of the previous application - Approved, Cancelled, Refused or Unused offer.

application_data

Dataframe Name : applicationDF
Column Count : 122
Row Count : 307511
Total no. of Elements : 37516342
Data types : float64(65), int64(41), object(16)
Columns with >40% null values: 49

Previous_application

Dataframe Name : previousDF
Column Count : 37
Row Count : 1670214
Total no. of Elements : 61797918
Data types : float64(65), int64(41), object(16)
Columns with >40% null values: 11

Data Cleaning Steps

1. Remove columns with null value >40%
2. Remove unrequired columns
3. Convert 'Object' type data columns to numerical /categorical columns
4. Use nunique() function to identify numerical columns with very less data type variable and convert them to categorical
5. Imputing Null values based on Median (to avoid outliers) for numerical and Mode for Categorical columns
6. Converting Days columns from negative to positive and then to year
7. Converting some numerical columns to categorical by binning process

Total Columns dropped from Data frames :

- applicationDF : 76
- previousDF : 15

Remaining Column count by data type for analysis :

- applicationDF :
  - Categorical : 19
  - Numerical : 28 (float64 – 18 , int64 – 10)
- previousDF :
  - Categorical : 14
  - Numberical : 9 (float64 – 5 , int64 – 4)

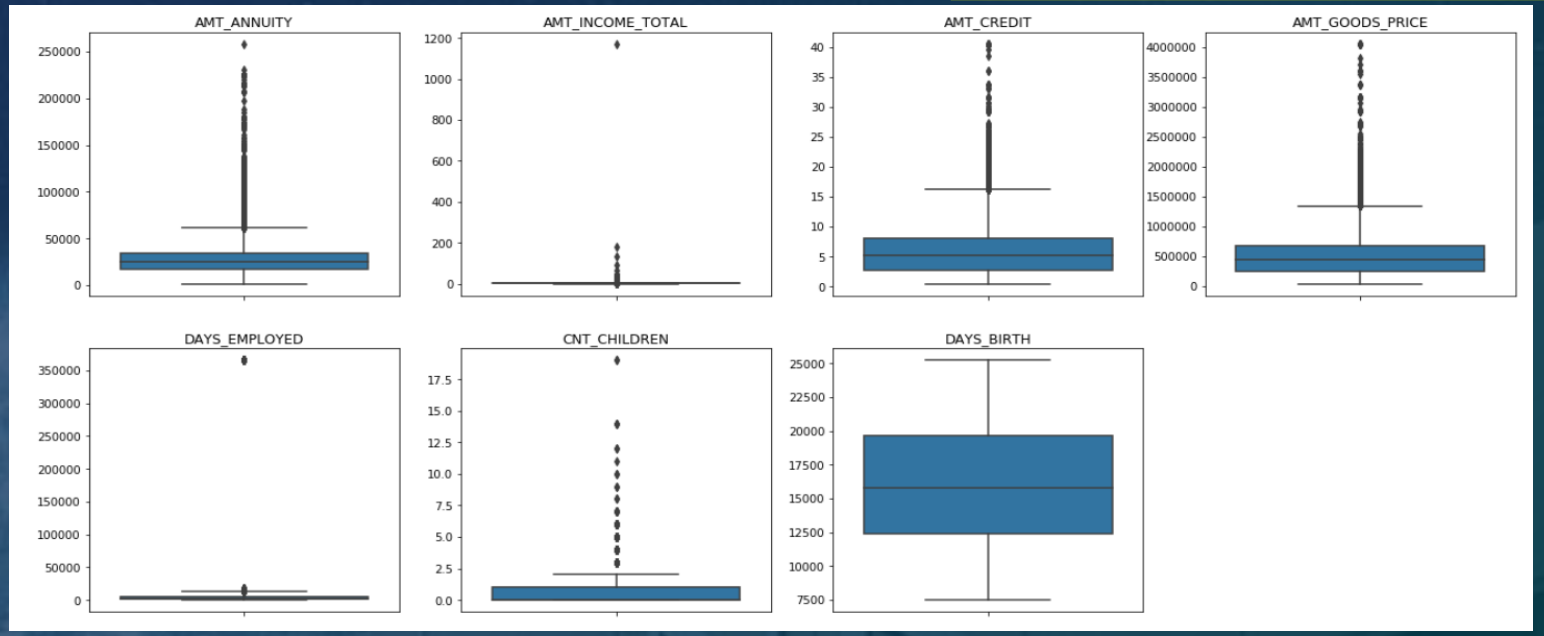
Outliers

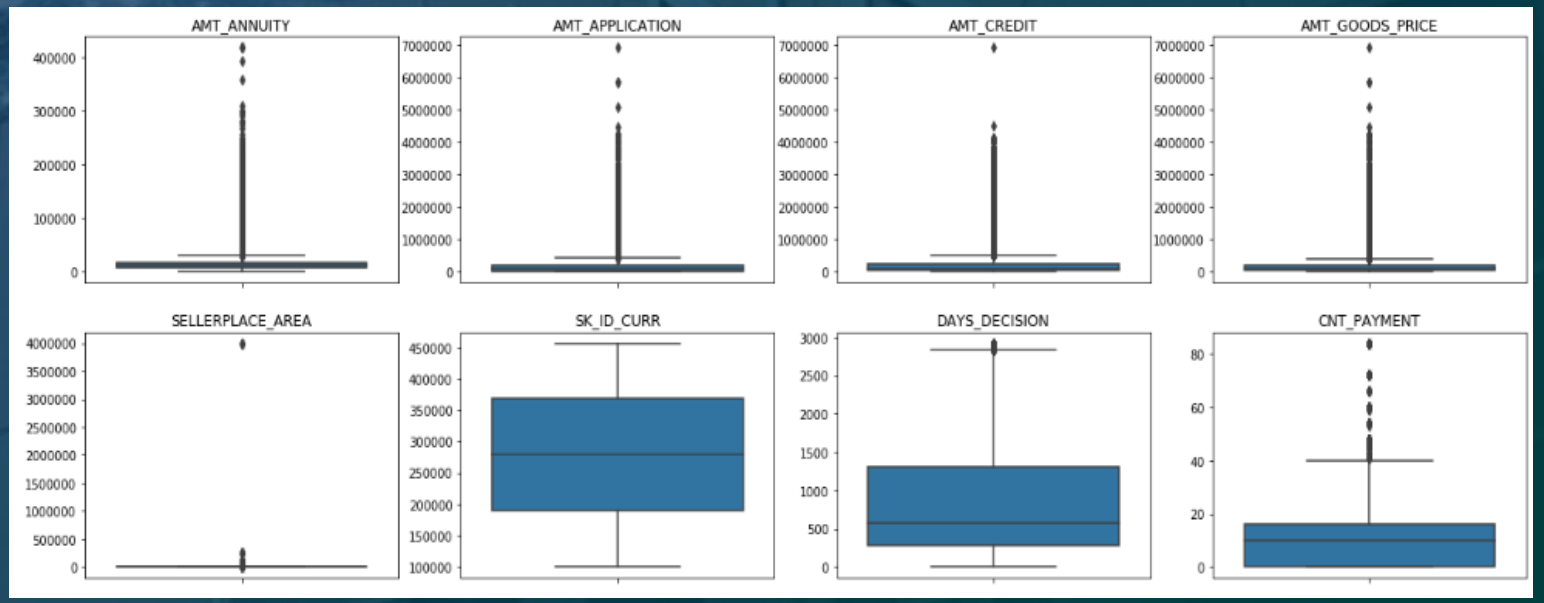
applicationDF:

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.

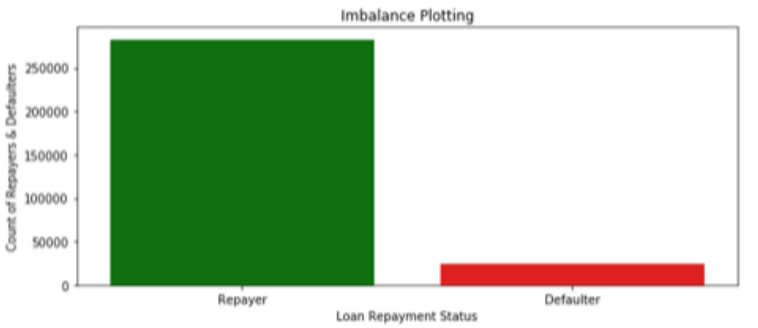
previousDF:

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 outliers on the negative side as the number of days is in negative





Imbalance



Current Application :

Repayers:92%
Defaulters: 8%

Repayer : Defaulter ratio is 11.31:1

Parameters for Analysis

Categorical

- NAME_CONTRACT_TYPE
- CIDE_GENDER
- FLAG_OWN_CAR
- FLAG_OWN_REALTY
- NAME_HOUSING_TYPE
- NAME_FAMILY_STATUS
- NAME_EDUCATION TYPE
- NAME_INCOME_TYPE
- REGION_RATING_CLIENT
- OCCUPATION TYPE
- ORGANIZATION TYPE
- CNT_CHILDREN
- CNT_FAM_MEMBER

Numerical converted into Categorical

- AGE GROUP
- EMPLOYMENT YEAR)
- AMT_CREDIT_RANGE
- AMT_INCOME_RANGE

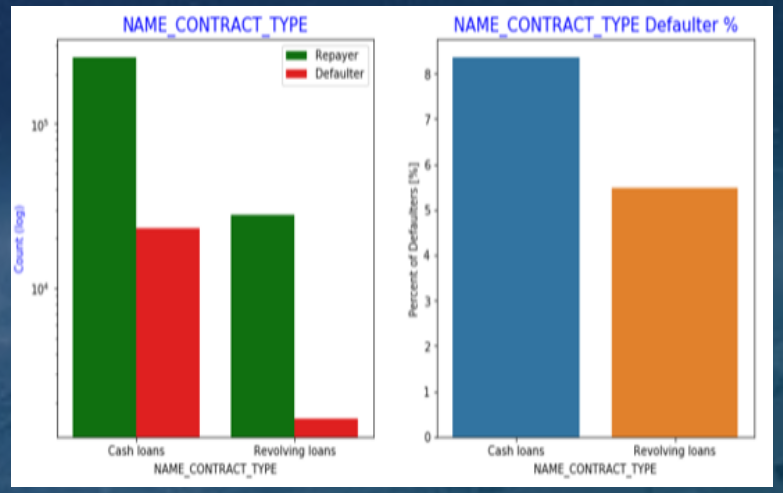
Numerical

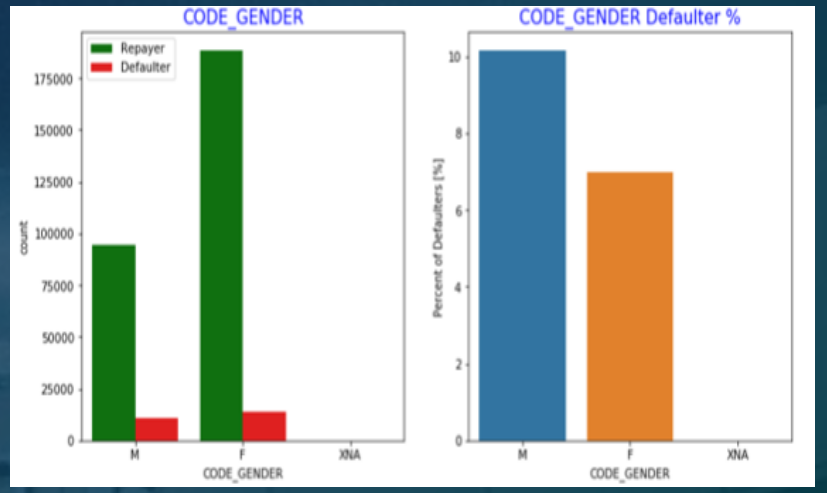
- AMT_INCOME_TOTAL
- AMT_CREDIT
- AMT_ANNUITY
- AMT_GOODS_PRICE
- REGION_POULATION_RELATIVE
- DAYS_BIRTH
- DAYS_EMPLOYED
- DAYS_REGISTRATION
- OBS_60_CNT_SOCIAL_CIRCLE
- DEF_60_CNT_SOCIAL_CIRCLE
- AMT_REQ_CREDIT_BUREAU_*

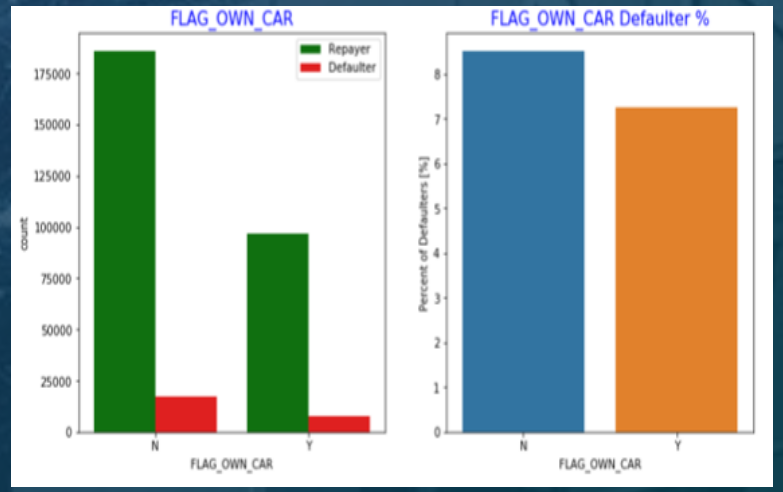
Categorical Analysis

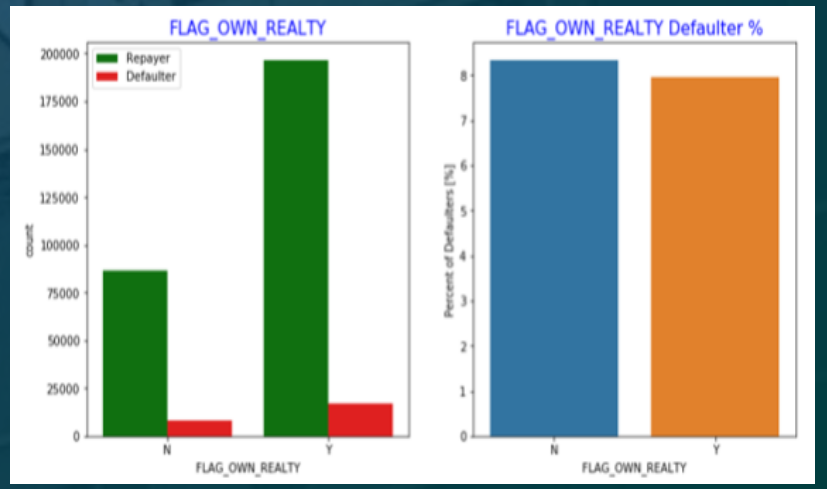
Segmented Univariate

1. Contract Type Revolving loans are 10% of all loans but high default rate
2. Males default (~10%) more than the females(~7%)
3. Clients who own a car is half of who don't own a car
4. Majority of the clients own Realty
5. Owning Car or Realty are not very much correlated to loan repayment status as default rate is almost same for both categories





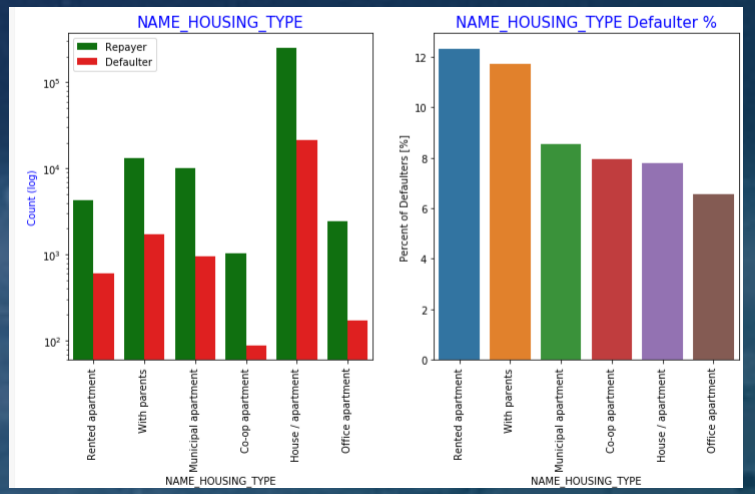


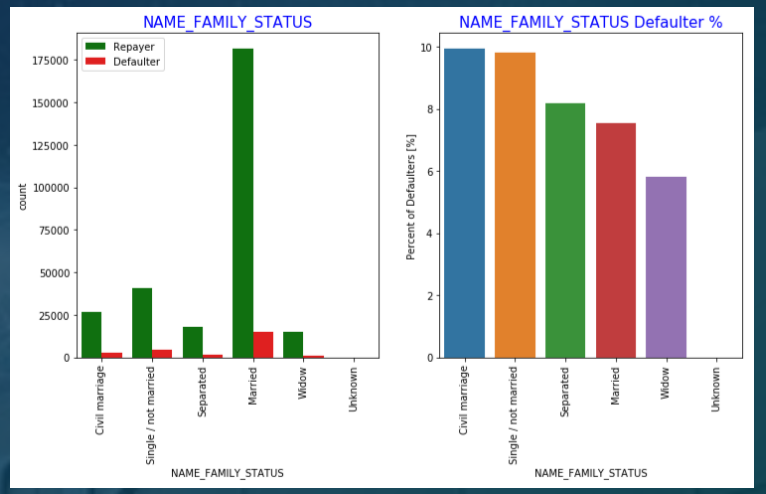


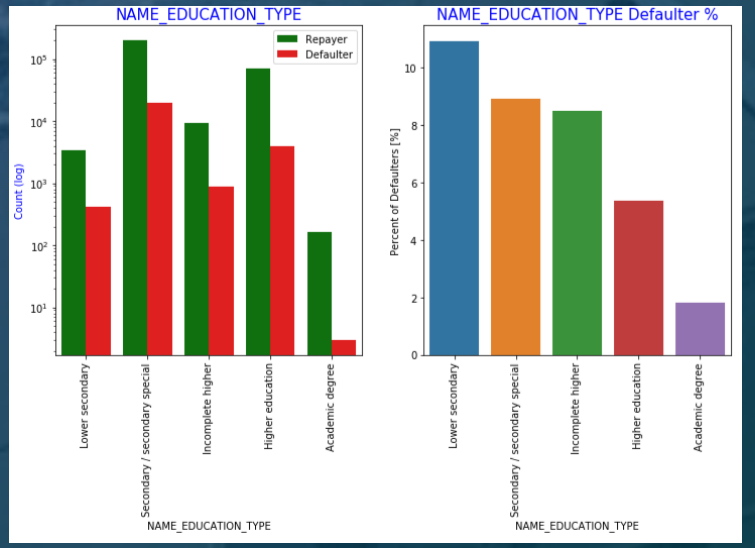
Categorical Analysis

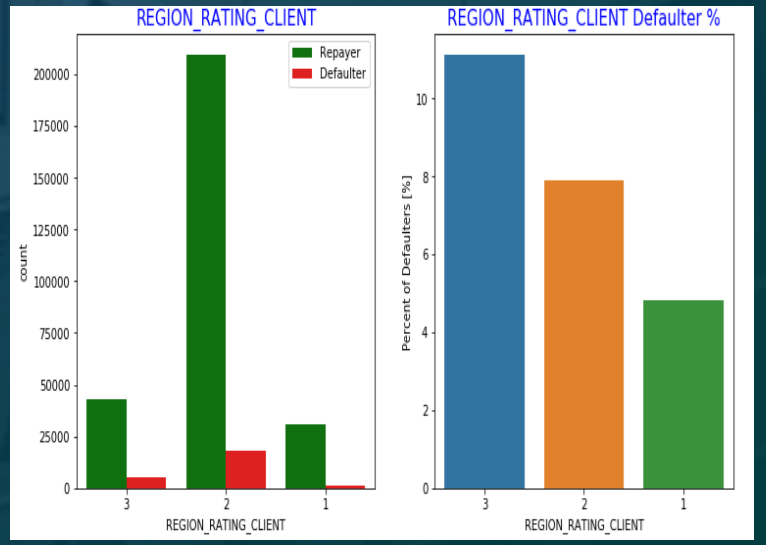
Segmented Univariate

1. The clients who live in rented apartments or with parents have higher default rate
2. The clients who are single/ not married or who had civil marriage have higher default rate
3. Lower Secondary & Secondary education category should be avoided for loan as default rate is >10%
4. The clients with academic category has very less default rate, thus should be given loan
5. The clients who are living in Region Rating 3 area has very high default, thus should be avoided or should be given loan with higher interest rate.





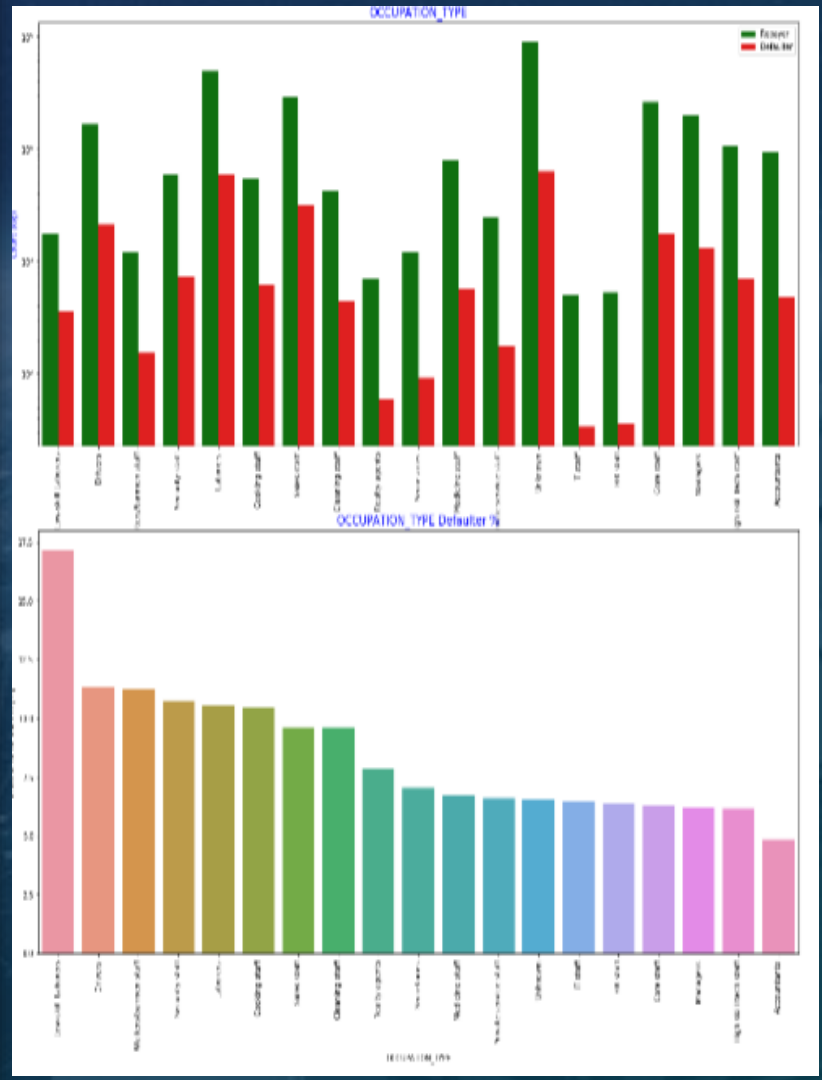


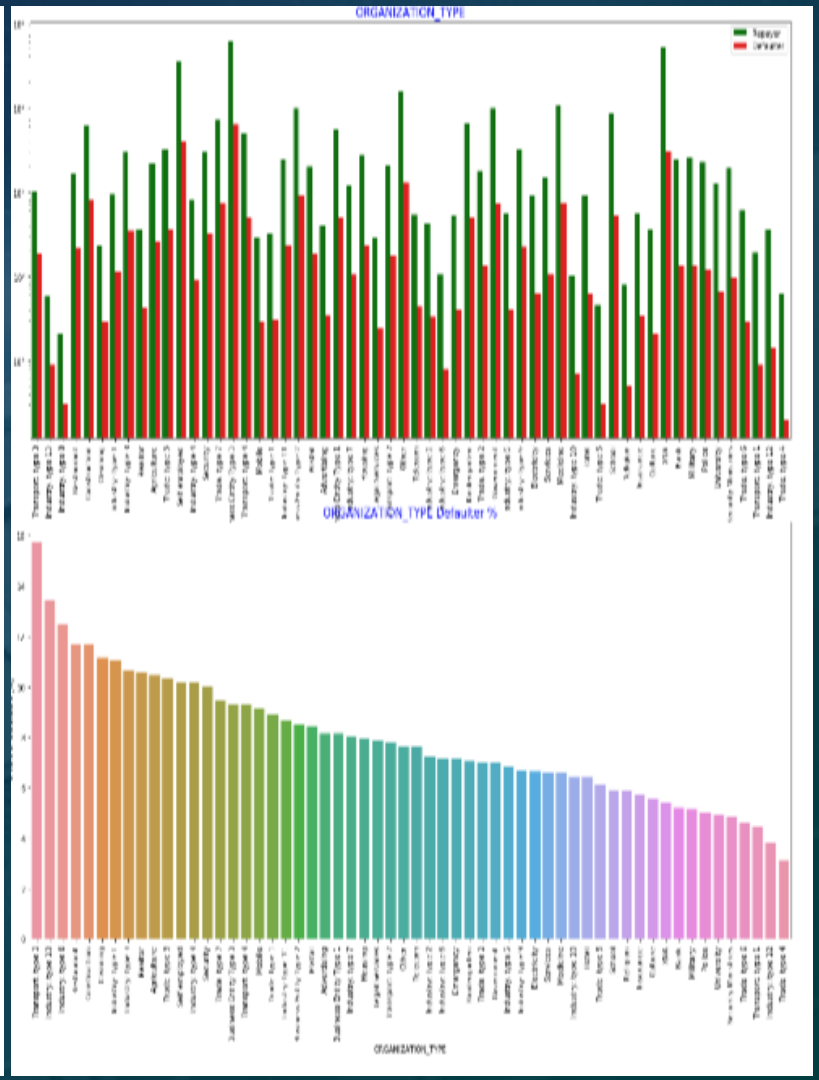


Categorical Analysis

Segmented Univariate

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 default loans are Low-skill Laborers (above 17%), followed by Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff.
3. 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%).
4. Self employed people have relatively higher defaulting rate

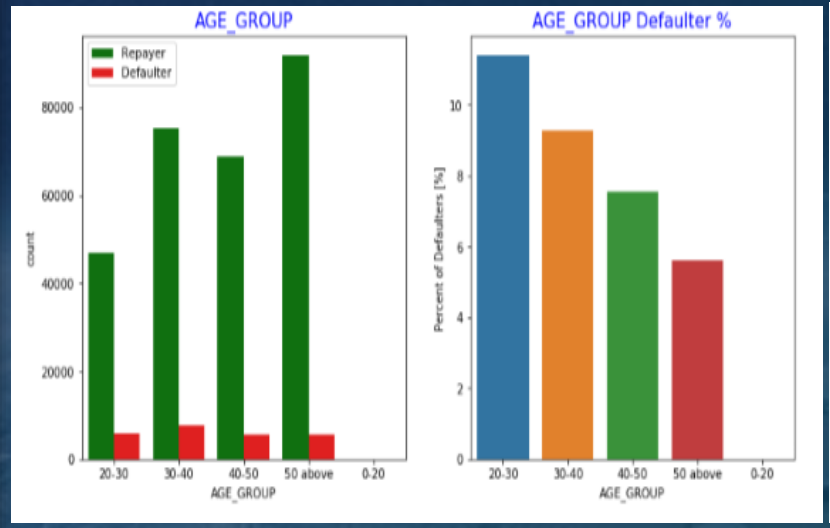


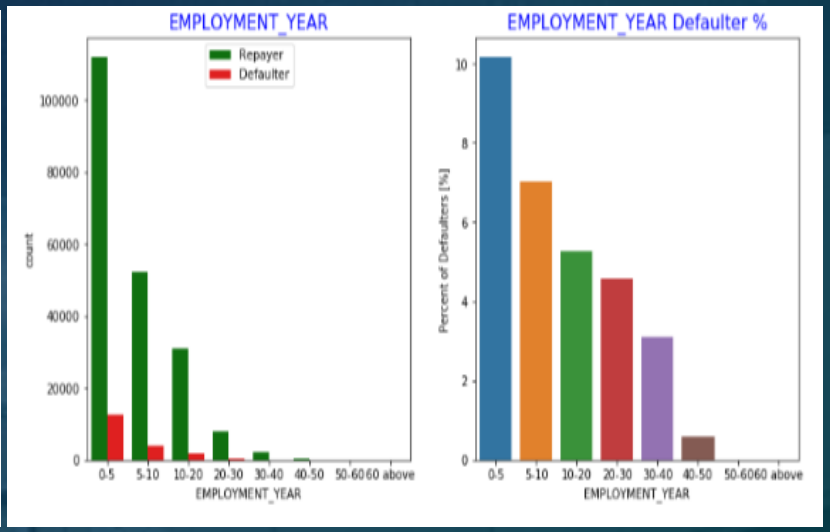


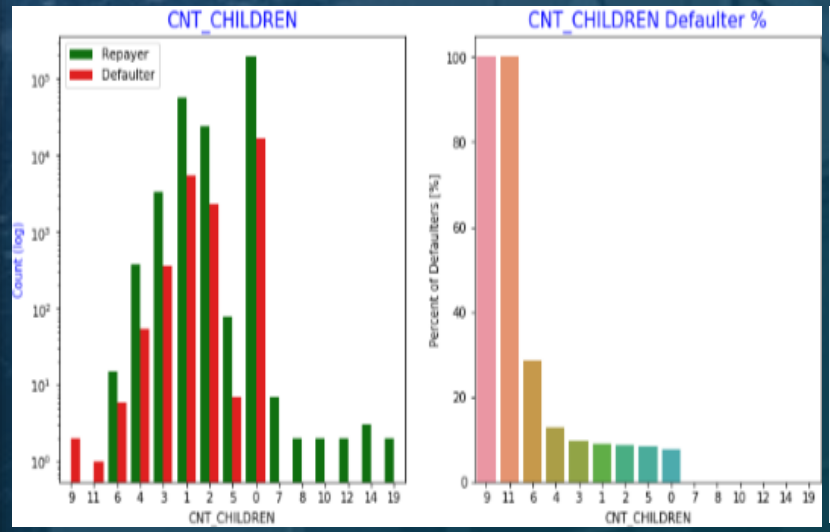
Categorical Analysis

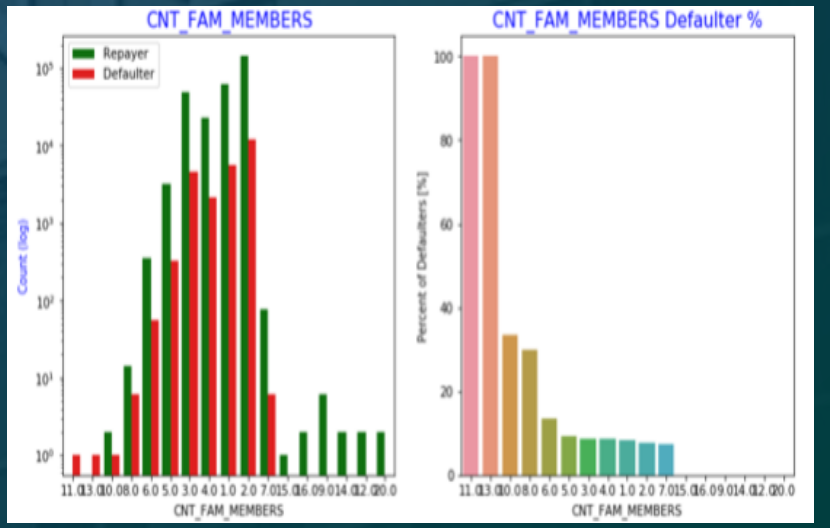
Segmented Univariate

1. Clients in the age group range 20-40 have higher probability of defaulting
2. Clients above age of 50 have low probability of defaulting
3. Clients with 0-5 years of work experience has ~10% defaulting rate
4. Clients with 40+ years experience have very low default rate
5. Most of the applicants do not have children. Very few clients have more than 3 children.
6. Client who have more than 4 children has a very high default rate
7. Family count follows the same trend of children count and they are highly correlated





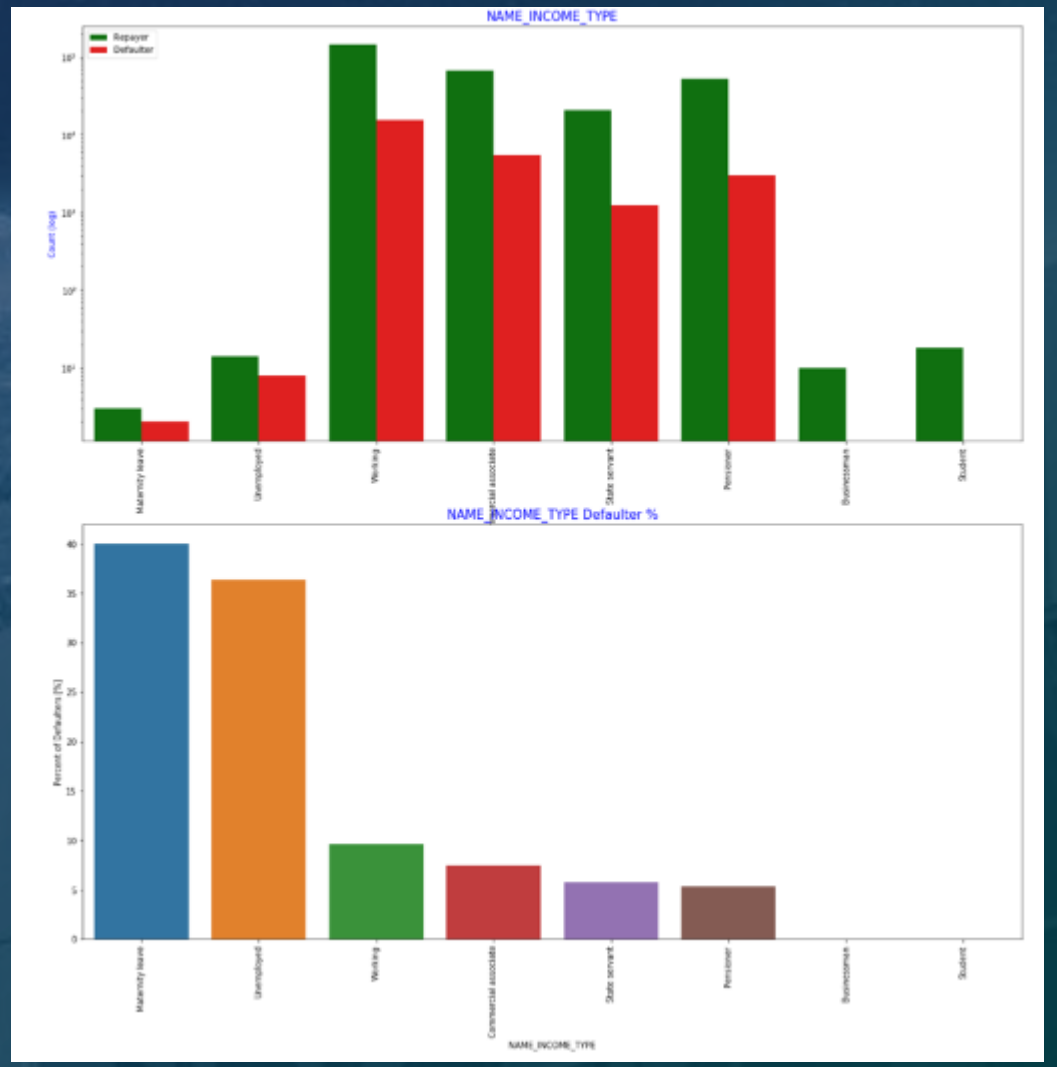




Categorical Analysis

Segmented Univariate

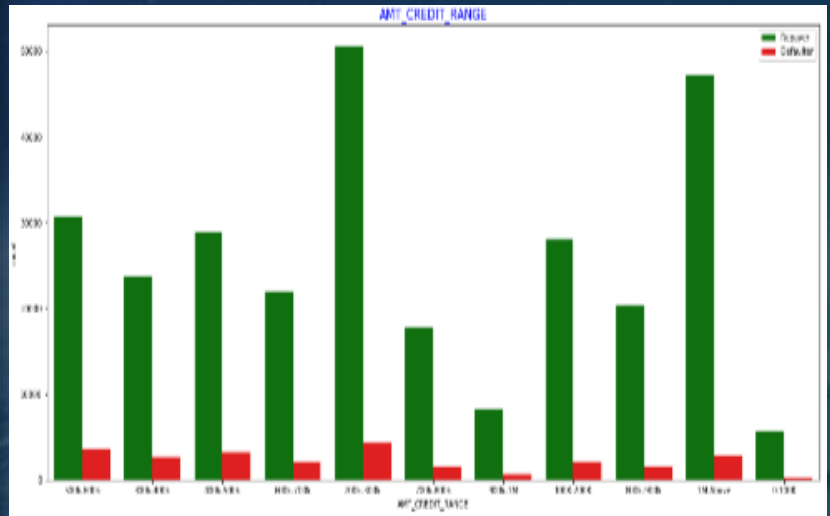
1. Working, followed by Commercial associate, Pensioner and State servant are highest number of clients.
2. The clients with the type of income Maternity leave have almost 40% ratio of not returning loans, followed by Unemployed (37%). These two category should be avoided
3. The rest of types of incomes are under the average of 10% for not returning loans.
4. Student and Businessmen, though less in numbers do not have any default record. Thus these two category are safest for providing loan

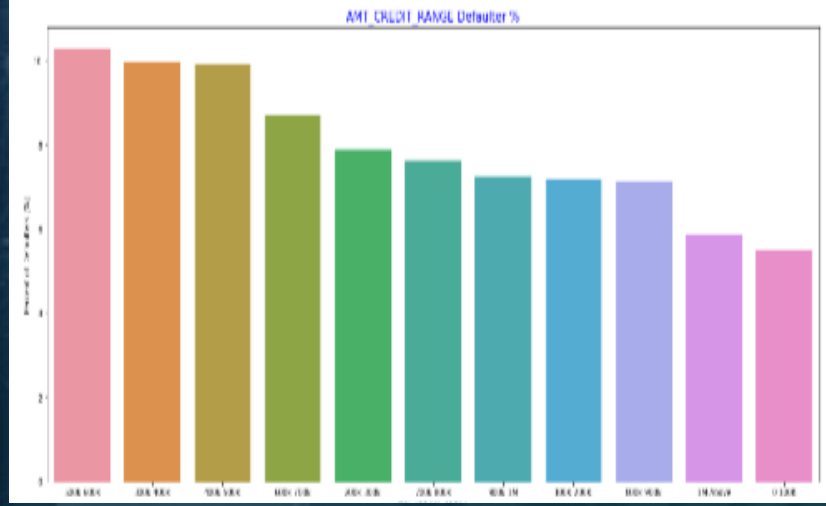


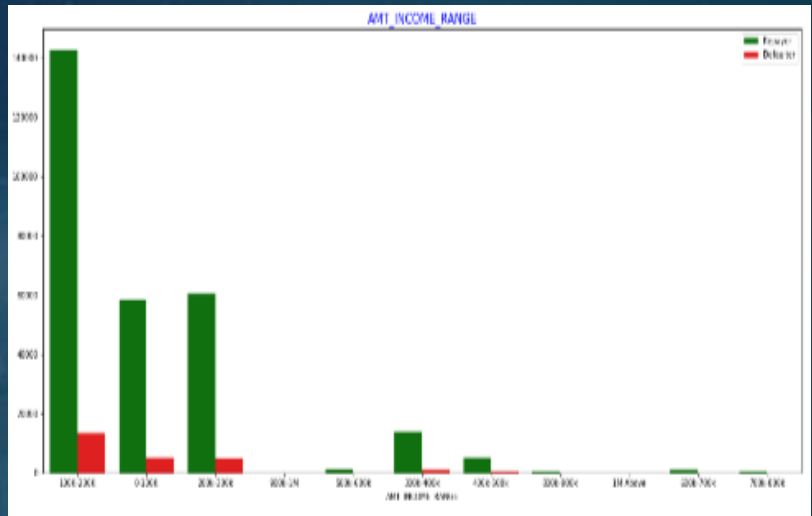
Categorical Analysis

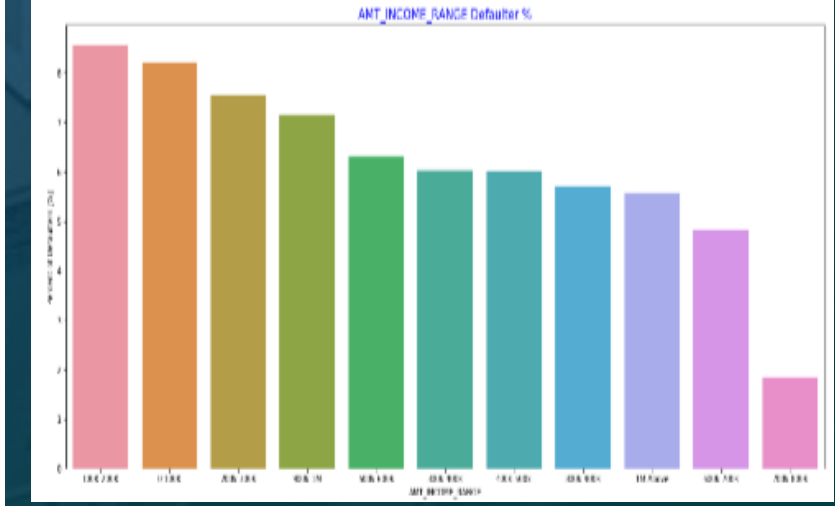
Segmented Univariate

1. More than 80% of the loan provided are for amount less than 900,000
2. People who get loan for 300-600k tend to default more than others.
3. 90% of the applications have Income total less than 300,000
4. Application with Income less than 300,000 has high probability of defaulting
5. Applicant with Income more than 700,000 are less likely to default









Bivariate Analysis

CORRELATION

Correlating factors amongst repayers:

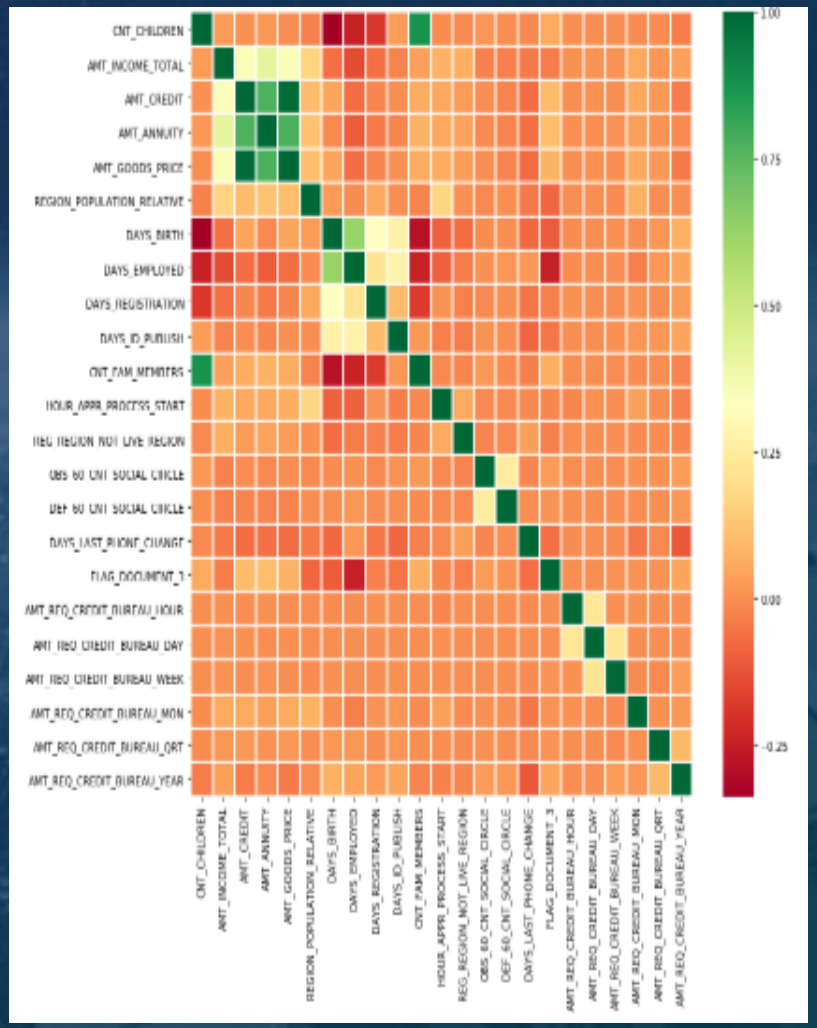
Credit amount is highly correlated with

- Amount of goods price
- Loan annuity
- Total Income

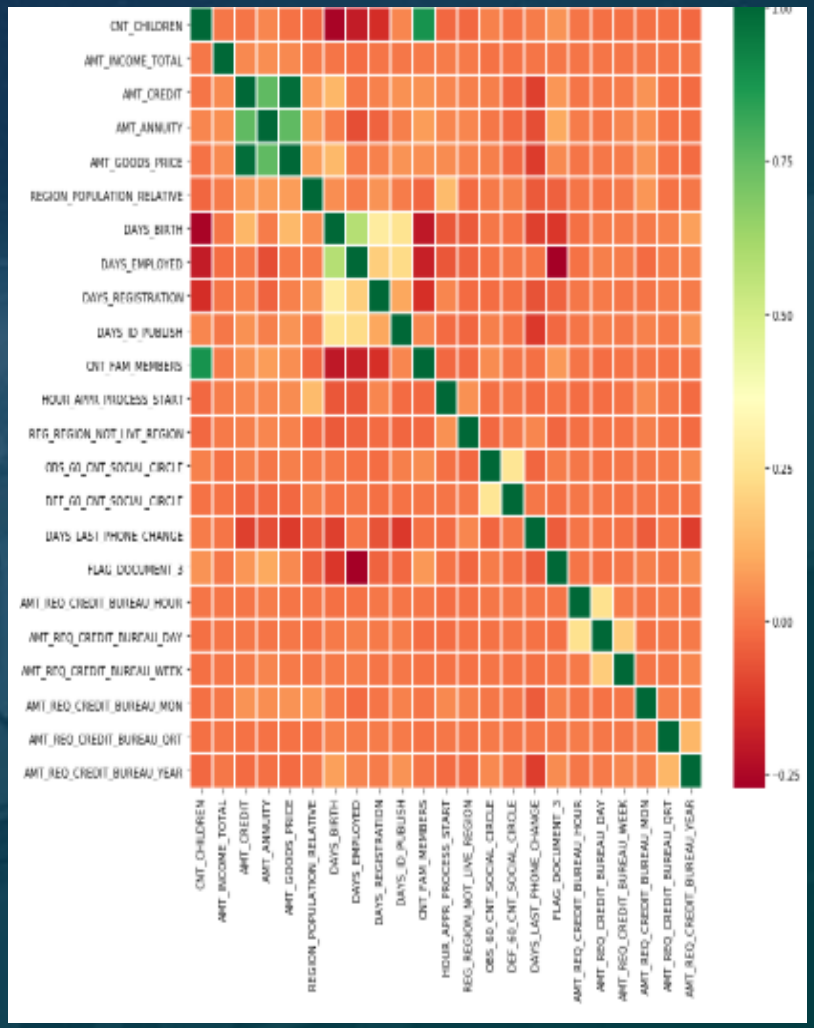
Number of repayers are highly correlated to Days employed

Correlating factors amongst defaulters:

- Credit amount is highly correlated with amount of goods price which is same as repayers.
- But the loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to repayers(0.77)
- There is a severe drop in the correlation between total income of the client and the credit amount (0.038) amongst defaulters vs 0.342 among repayers.
- Days_birth/Age 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 defaulter to observed count in social circle among defaulters(0.264) when compared to repayers(0.254)



Repayer Correlation Matrix



Defaulter Correlation Matrix

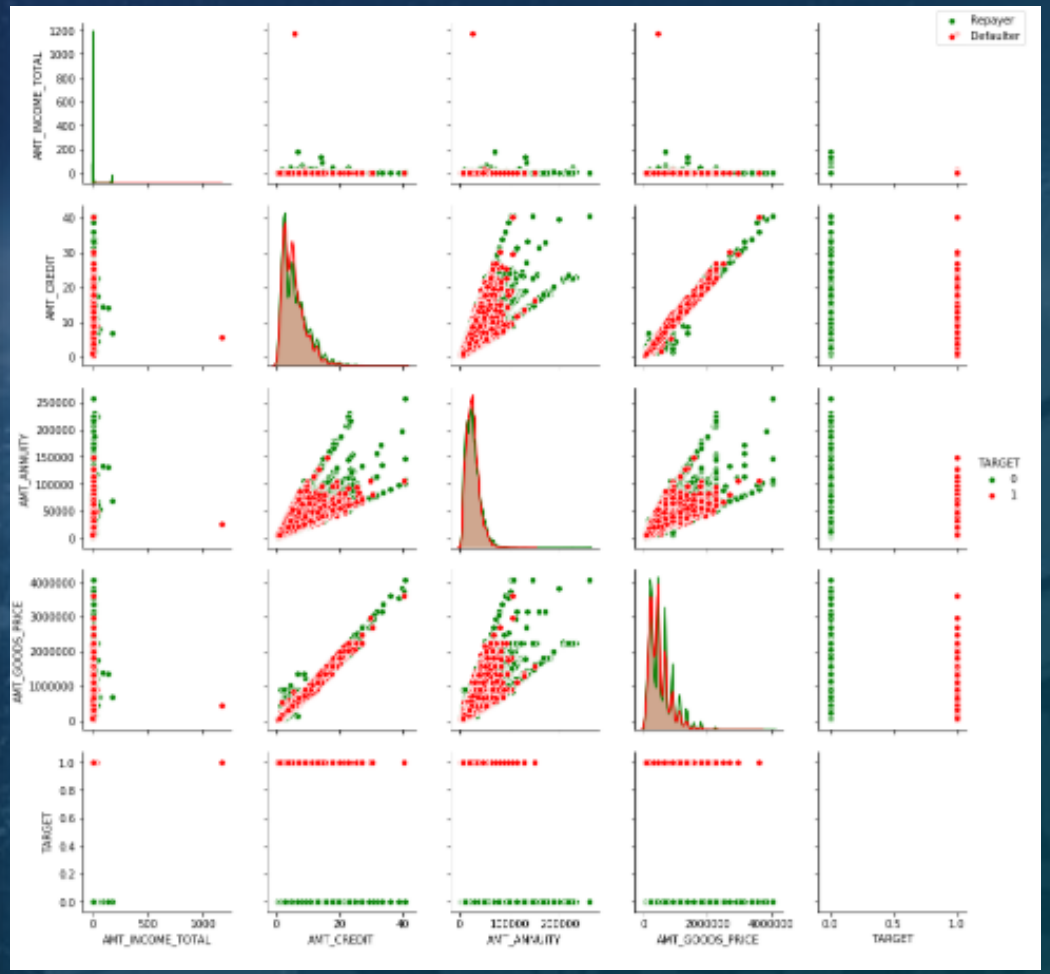
** For Correlation matrix and top 10 correlating factor for Defaulter & Repayer please refer attach .ipynb section 4.2.2*

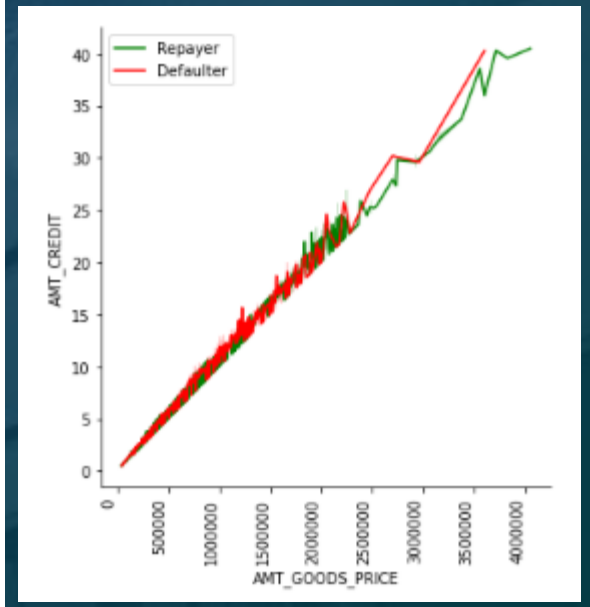
Bivariate Analysis

CORRELATION

1. Majority of the income range is within 200K.
2. Most no of loans are given for goods price below 10 lakhs
3. Most people pay annuity below 50000 for the credit loan
4. Credit amount of the loan is mostly less then 10 lakhs
5. The repayer and defaulter distribution overlap in all the plots and hence we cannot use any of these variables in isolation to make a decision

From rel plot :
When the credit amount goes beyond 3M, there is an increase in defaulters



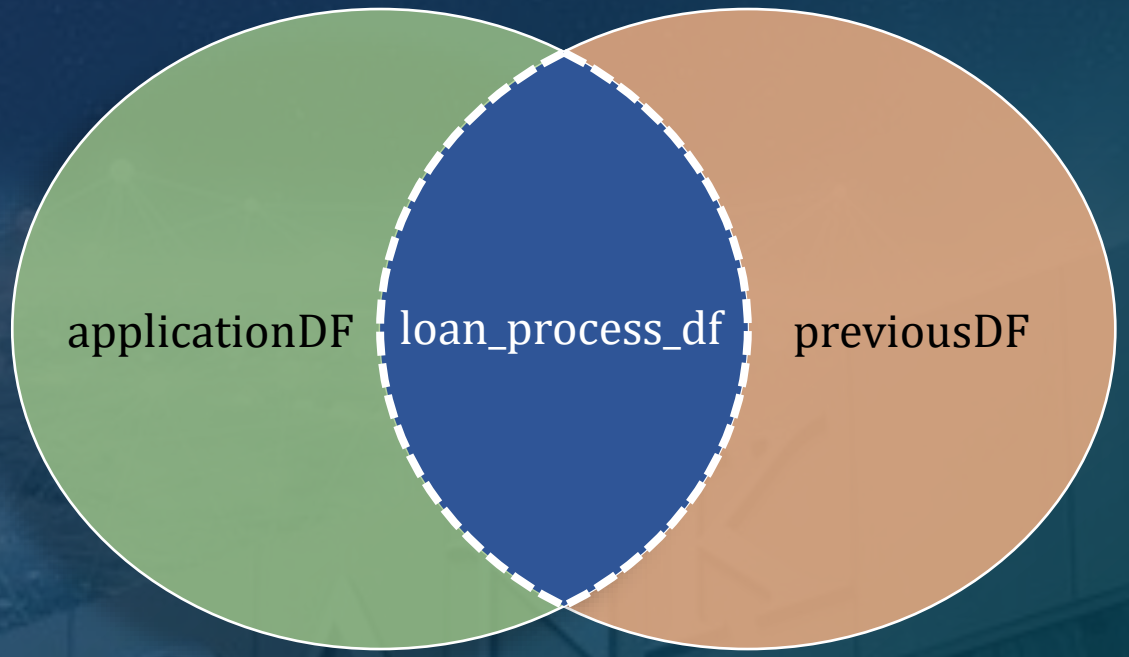


Merging Dataframe

Merged Dataframe

Dataframe Name : loan_process_df
Column Count : 74
Row Count : 1413701
Total no. of Elements : 104613874
Data types :

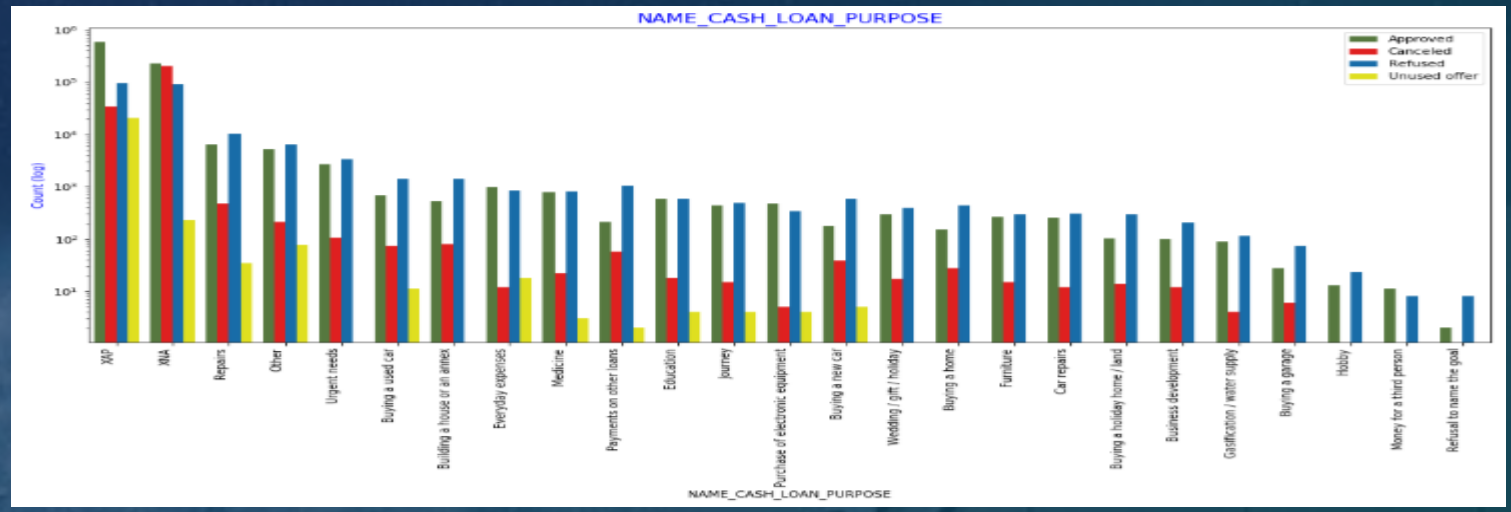
- category(37)
- float64(23)
- int64(14)



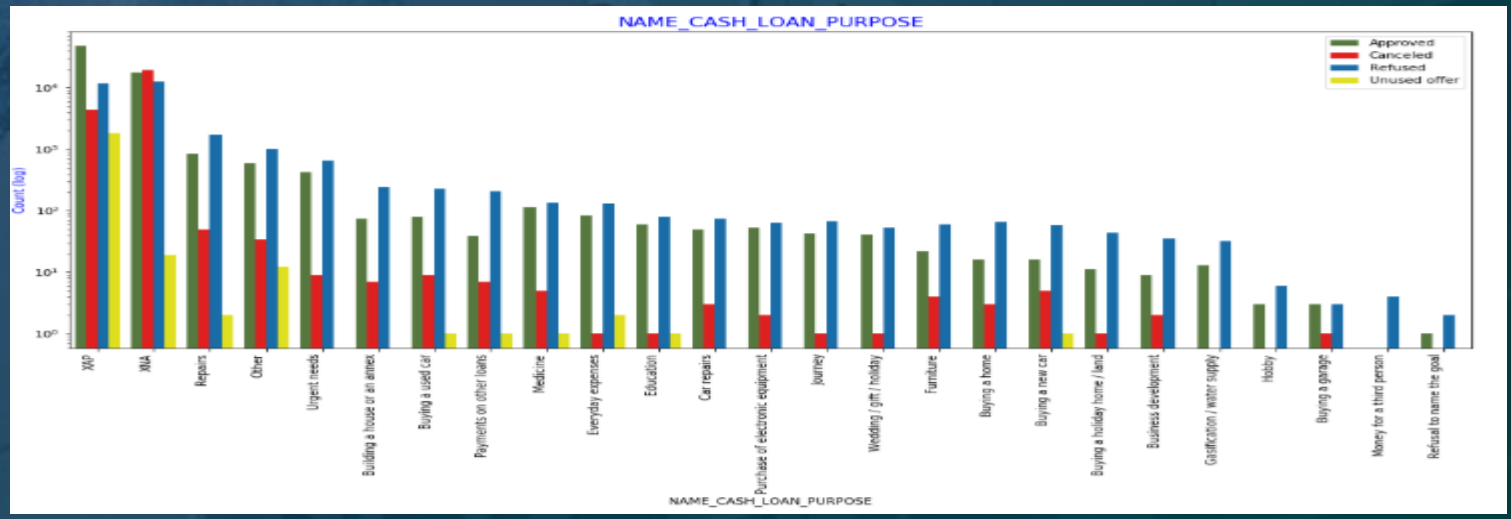
Dataframes merged through inner join

Analysis on Merged Dataframe

1. Loan purpose has high number of unknown values (XAP, XNA)
2. Loan taken for the purpose of Repairs seems to have highest default rate.
3. 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 cancelled the loan



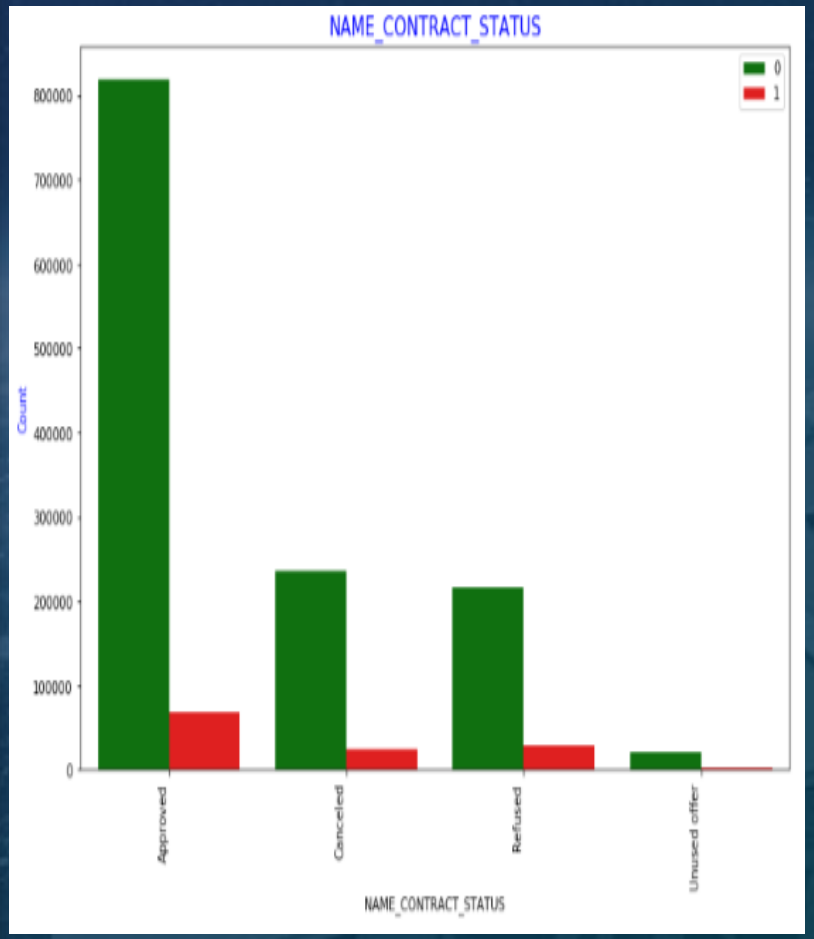
Repayer Cash Loan Purpose vs Loan Decision

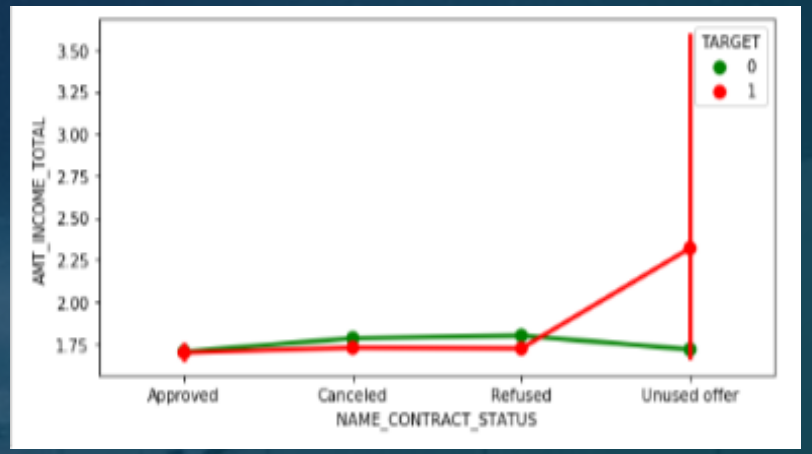


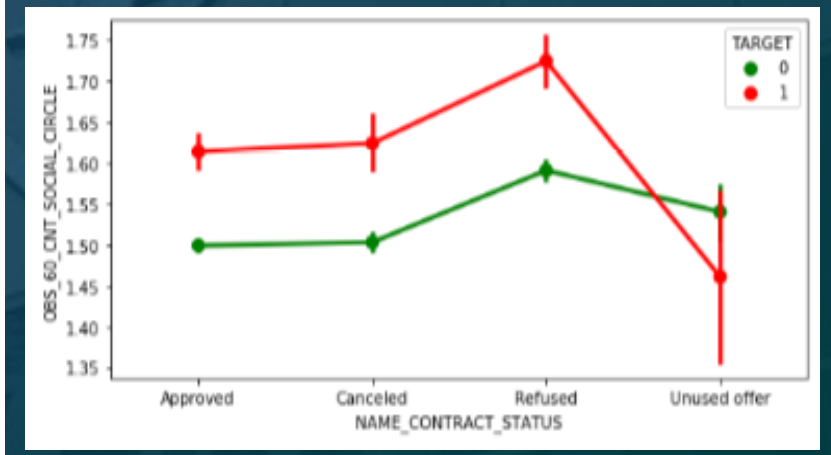
Defaulter Cash Loan Purpose vs Loan Decision

Analysis on Merged Dataframe

1. 90% of the clients who previously cancelled their loan have actually repayed the loan in current application set.
2. 88% of the clients who were previously refused a loan has payed back the loan in current case
3. The point plots show that the people who have not used offer earlier have defaulted even when their average income is higher than others.
4. Clients who have average of 0.13 or higher DEF_60_CNT_SOCIAL_CIRCLE score tend to default more.







Conclusion

Parameter	Drivers for Repayer	Drivers for Defaulter
NAME_EDUCATION_TYPE	Academic degree	Lower Secondary & Secondary
NAME_INCOME_TYPE	Businessman, Student	Maternity Leave, Unemployed
NAME_FAMILY_STATUS		Single/Not married, Civil Marriage
REGION_RATING_CLIENT	Rating 1	Rating 3
NAME_HOUSING_TYPE		Rented apartment, Living with parents
ORGANIZATION_TYPE	Trade type 4, 5, Industry type 8	Transport: type 3, Industry: type 13 ,Industry: type 8 , Restaurant ,Self employed
DAYS_BIRTH	50+ years age	20-40 Years of Age
DAYS_EMPLOYED	40+ years of experience	0-5 years of experience
NAME_CASH_LOAN_PURPOSE	Hobby, Buying Garage	Repair, Others
CNT_CHILDREN	0-2 child(ren)	>4 children
AMT_CREDIT		300k-600K loan amount range
AMT_INCOME_TOTAL	>700K	<300K

- *The drivers for defaulter would result in either rejection or the loan by the bank or it may turn into bank offering higher rate of interest in the loan to mitigate the risk. The client may then refuse the loan due to high interest rate.*
- *For detailed conclusion please refer to the attach .ipynb file Section 6 for Conclusion explanation*

THANK YOU !