BANK LOAN SERVICES CASE STUDY

EDA BY KOMAL NALAWADE AND PRABHJOT SINGH

INTRODUCTION

This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Understanding

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants are capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

The data given below 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.

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 on different stages of the process.

Bank_loan_services_Casestudy

Business Objective

The case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Data Understanding

1. ***'application_data.csv'*** It contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.

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

3. ***'columns_description.csv'*** It is data dictionary which describes the meaning of the variables.

Reading and Understanding the dataset

Inspecting dataframes

Shape and size

1. **application data**
   Dimensions of application_data : (307511, 122)
   Size of application_data : 37516342

   There are total of 122 attributes out of which

   - 65 columns are of datatype float
   - 41 columns are of datatype int
   - 16 columns are of datatype object
   - There are no duplicate records in application_data.

2. **previous data**
   Dimensions of previous_data : (1670214, 37)
   Size of previous_data : 61797918
   There are total of 37 variables out of which

   - 15 columns are of datatype float
   - 06 columns are of datatype int
   - 16 columns are of datatype object
   - There are no duplicate records in previous data.

After primarily inspecting application_data without examining null values,

1. CNT_PAYMENT describes the term of payment in months. Hence, should be in integer datatype
2. There are no duplicate records in application and previous data.
3. DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH columns(attributes) are with negative sign. Need to convert them to absolute values for further analysis.
4. Maximum value for DAYS_EMPLOYED is 365243 which when converted to years gives 1000. It is an outlier as no person would serve so long.

Data Cleaning and Imputations

Null Value Calculation

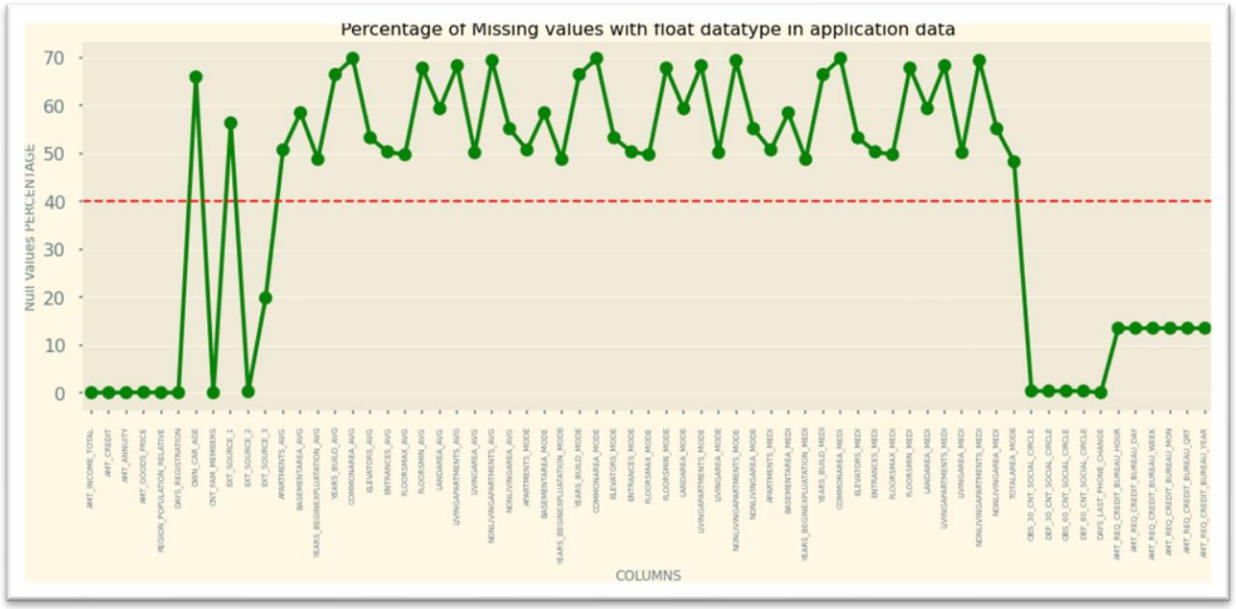
In application data,

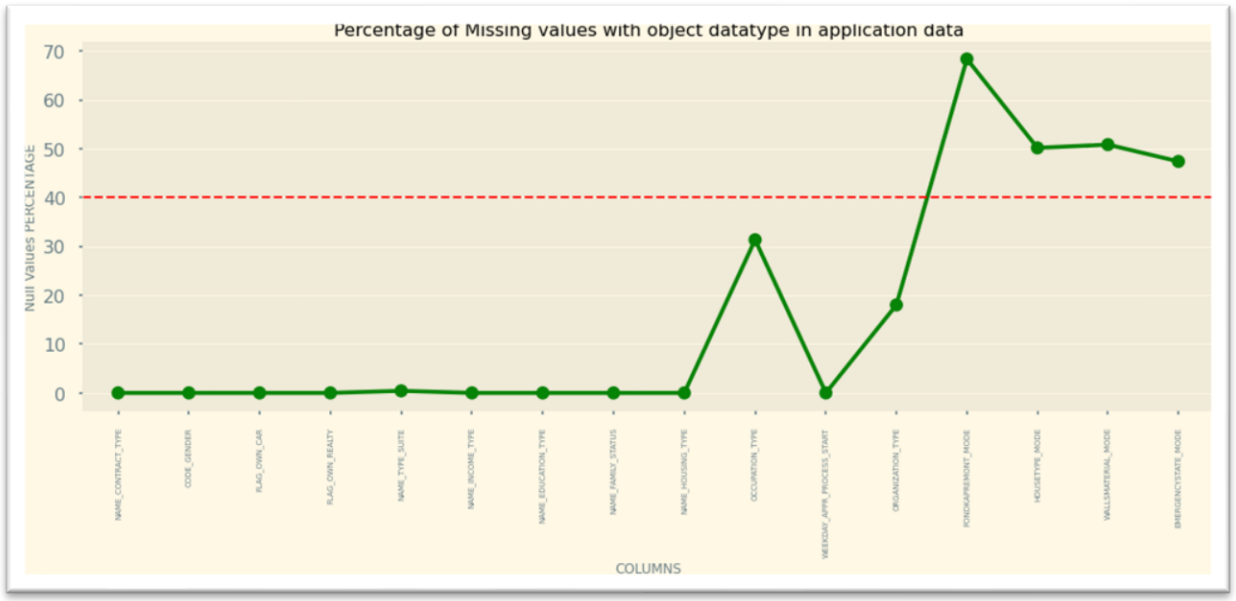
1. There are no null values in the integer datatype columns those are 41 columns
2. Variables with 40% and above missing values with float datatype can be dropped.

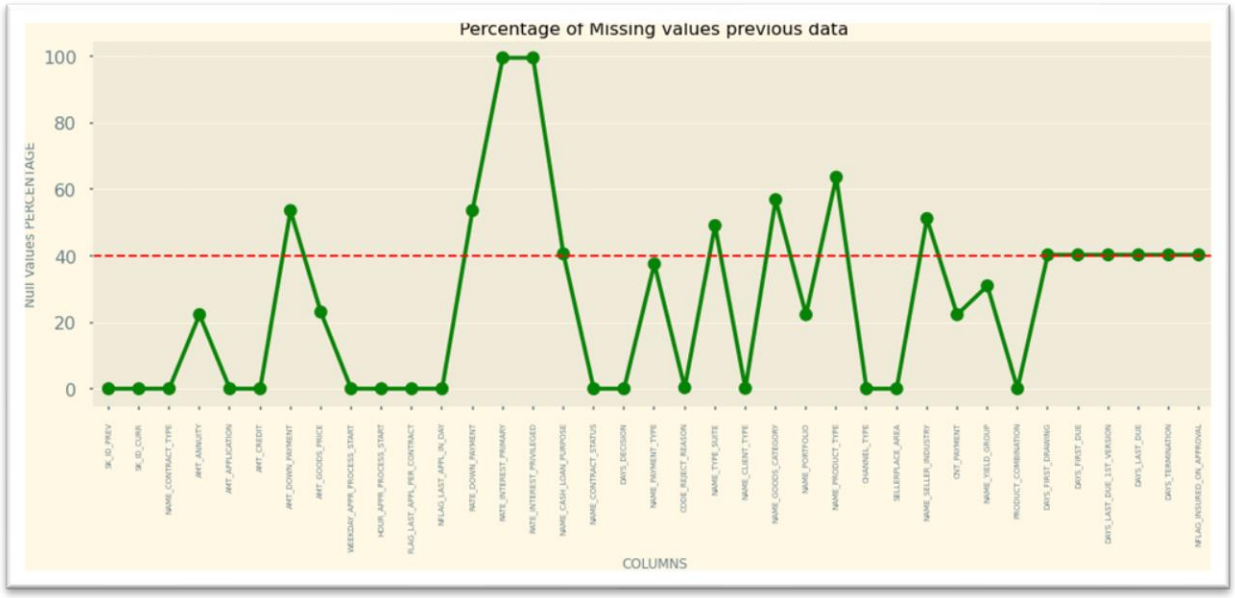
- FONDKAPREMONT_MODE ----- 68.39 %
- HOUSETYPE_MODE ----- 50.18 %
- WALLSMATERIAL_MODE -----50.84 %
- EMERGENCYSTATE_MODE ----- 47.40 %

3. Number of null values below 40% with float datatype. These values can be imputed

- CODE_GENDER ---- 4
- NAME_TYPE_SUITE ---- 1292
- OCCUPATION_TYPE ---- 96391
- ORGANIZATION_TYPE ---- 5537


Percentage of Missing values with float datatype in application data
Null Values PERCENTAGE
70
60
50
40
30
20
10
0
COLUMNS


Percentage of Missing values with object datatype in application data
Null Values PERCENTAGE
70
60
50
40
30
20
10
0
NAME_CONTRACT_TYPE
CODE_GENDER
FLAG_OWN_CAR
FLAG_OWN_REALTY
NAME_TYPE_SUITE
NAME_INCOME_TYPE
NAME_EDUCATION_TYPE
NAME_FAMILY_STATUS
NAME_HOUSING_TYPE
OCCUPATION_TYPE
WEEKDAY_APPR_PROCESS_START
ORGANIZATION_TYPE
FONDKAPREMONT_MODE
HOUSETYPE_MODE
WALLSMATERIAL_MODE
EMERGENCYSTATE_MODE
COLUMNS



In previous data,

From the plot we can see the columns in which percentage of null values more than 40% are marked above the red line and the columns which have less than 40 % null values below the red line.

AMT_DOWN_PAYMENT	53.64
RATE_DOWN_PAYMENT	53.64
RATE_INTEREST_PRIMARY	99.64
RATE_INTEREST_PRIVILEGED	99.64
NAME_CASH_LOAN_PURPOSE	40.59
NAME_TYPE_SUITE	49.12
NAME_GOODS_CATEGORY	56.93
NAME_PRODUCT_TYPE	63.68
NAME_SELLER_INDUSTRY	51.23
DAYS_FIRST_DRAWING	40.30
DAYS_FIRST_DUE	40.30
DAYS_LAST_DUE_1ST_VERSION	40.30
DAYS_LAST_DUE	40.30
DAYS_TERMINATION	40.30
NFLAG_INSURED_ON_APPROVAL	40.30

There are 15 columns in previous_data dataframe where missing value is more than 40%.

Analyzing and Dropping Irrelevant Variables in application data

There are 45 columns with float datatype with 40% and more missing values. These are mostly related to the locality of the applicant.so these columns can be dropped.

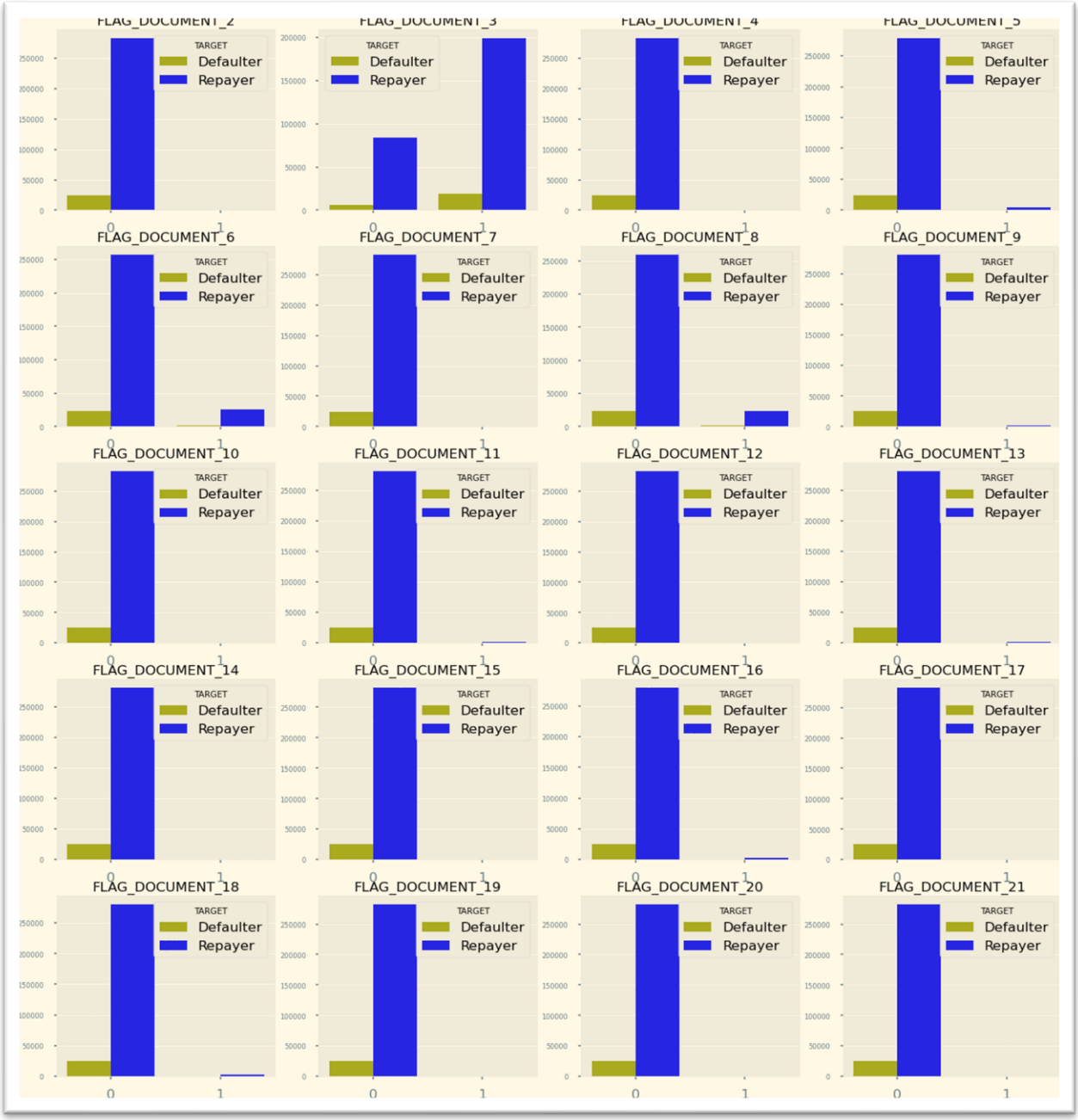
EXT_SOURCE_1 gives the normalized(between 0 and 1) credit score obtained from the credit bureau which can be important driving factor for loan defaulters.(applicant with the low credit score is likely to default).

- source There are three credit reporting bureaus that give individual's credit score(ability to repay the loan).

Null values from EXT_SOURCE_1 can be replaced with corresponding EXT_SOURCE_2 or EXT_SOURCE_3 as they are nearly same

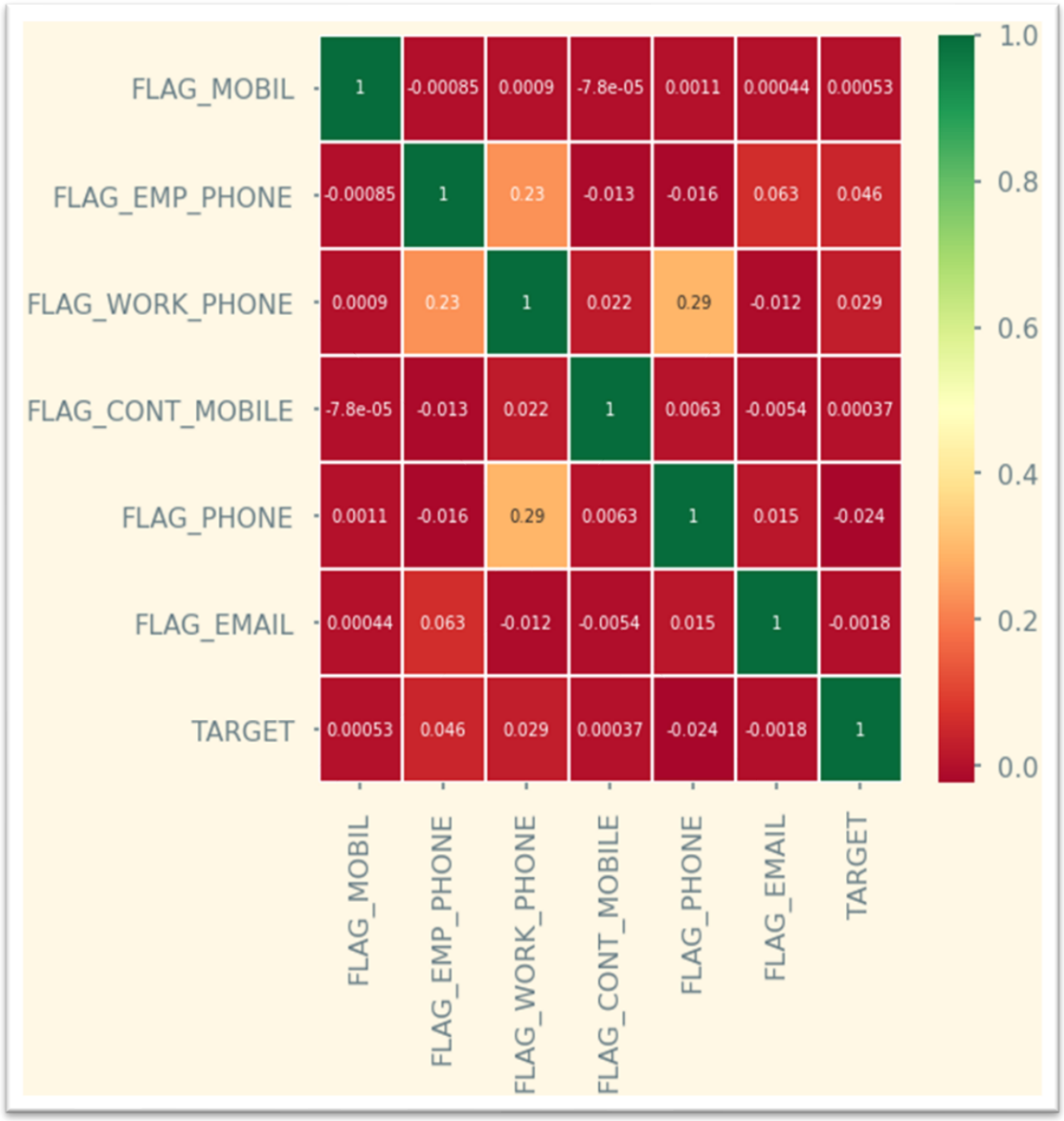
44 columns float datatype with 40% and above missing values columns(EXT_SOURCE_1 excluded) ['OWN_CAR_AGE', 'APARTMENTS_AVG', 'BASEMENTAREA_AVG', EARS_BEGINEXPLUATATION_AVG', 'YEARS_BUILD_AVG','COMMONAREA_AVG', 'ELEVATORS_AVG', 'ENTRANCES_AVG', 'FLOORSMAX_AVG','FLOORSMIN_AVG', 'LANDAREA_AVG','LIVINGAPARTMENTS_AVG', 'LIVINGAREA_AVG','NONLIVINGAPARTMENTS_AVG','NONLIVINGAREA_AVG', APARTMENTS_MODE', 'BASEMENTAREA_MODE','YEARS_BEGINEXPLUATATION_MODE','YEARS_BUILD_MODE',

'COMMONAREA_MODE','ELEVATORS_MODE','ENTRANCES_MODE','FLOORSMAX_MODE', 'FLOORSMIN_MODE','LANDAREA_MODE','LIVINGAPARTMENTS_MODE','LIVINGAREA_MODE', 'NONLIVINGAPARTMENTS_MODE','NONLIVINGAREA_MODE','APARTMENTS_MEDI', 'BASEMENTAREA_MEDI','YEARS_BEGINEXPLUATATION_MEDI','YEARS_BUILD_MEDI','COMMONAREA_MEDI','ELEVATORS_MEDI','ENTRANCES_MEDI','FLOORSMAX_MEDI','FLOORSMIN_MEDI', 'LANDAREA_MEDI','LIVINGAPARTMENTS_MEDI','LIVINGAREA_MEDI','NONLIVINGAPARTMENTS_MEDI','NONLIVINGAREA_MEDI','TOTALAREA_MODE'] +



19 FLAG_DOCUMENT_X columns ['FLAG_DOCUMENT_2,'FLAG_DOCUMENT_4', 'FLAG_DOCUMENT_5', 'FLAG_DOCUMENT_6','FLAG_DOCUMENT_7', 'FLAG_DOCUMENT_8', 'FLAG_DOCUMENT_9','FLAG_DOCUMENT_10', 'FLAG_DOCUMENT_11', 'FLAG_DOCUMENT_12','FLAG_DOCUMENT_13', 'FLAG_DOCUMENT_14', 'FLAG_DOCUMENT_15','FLAG_DOCUMENT_16', 'FLAG_DOCUMENT_17', 'FLAG_DOCUMENT_18',FLAG_DOCUMENT_19', 'FLAG_DOCUMENT_20', 'FLAG_DOCUMENT_21'] +

4 columns with the object type variables possessing 40% and more missing values ['FONDKAPREMONT_MODE', 'HOUSETYPE_MODE', 'WALLSMATERIAL_MODE', 'EMERGENCYSTATE_MODE']



6 columns of contact parameters['FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE', 'FLAG_EMAIL']=

73 unwanted_application columns

122 total columns -73 irrelevant columns = 49 columns in application data

Analyze and Dropping Irrelevant Variables in previous_data

There are 15 columns with 40% and more missing values.

- Though AMT_DOWN_PAYMENT and RATE_DOWN_PAYMENT have more than 40% missing values, those columns can be important driving factors for loan defaulters. And can be imputed
- **source**
- RATE_DOWN_PAYMENT values are the normalized rate calculated from respective AMT_DOWN_PAYMENT column values
- Hence, there are 13 columns with 40% and more missing values. And 4 unnecessary columns.
- Total 17 irrelevant columns

[RATE_INTEREST_PRIMARY', 'RATE_INTEREST_PRIVILEGED', 'NAME_CASH_LOAN_PURPOSE',
'NAME_TYPE_SUITE', 'NAME_GOODS_CATEGORY', 'NAME_PRODUCT_TYPE', 'NAME_SELLER_INDUSTRY',
'DAYS_FIRST_DRAWING','DAYS_FIRST_DUE', 'DAYS_LAST_DUE_1ST_VERSION', 'DAYS_LAST_DUE',
'DAYS_TERMINATION', 'NFLAG_INSURED_ON_APPROVAL''WEEKDAY_APPR_PROCESS_START','HOUR_APPR_PROCESS_START','FLAG_LAST_APPL_PER_CONTRACT','NFLAG_LAST_APPL_IN_DAY']

- Two records with negative AMT_DOWN_PAYMENT deleted
- 37 total columns - 17 irrelevant columns= 20 columns

Standardize Values

Strategy for application_data:

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

After binning,

1. More than 50% loan applicants have income amount in the range of 100K-200K. Almost 92% loan applicants have income less than 300K
2. More Than 16% loan applicants have taken loan which amounts to more than 1M. And 17.82% is in 200K-300K
3. 31% loan applicants have age above 50 years. More than 70% of loan applicants have age over 40 years.
4. 55% of the loan applicants have work experience within 0-5 years and almost 80% of them have less than 10 years of work experience
5. New 7 columns are added after binning , so the total column count is 56. AMT_INCOME_RANGE, AMT_CREDIT_RANGE, AGE, AGE_GROUP ,YEARS_EMPLOYED, EMPLOYMENT_YEAR,CREDIT_SCORE.

Data Type Conversion

In application_data,

23 columns are made categorical

['NAME_CONTRACT_TYPE','CODE_GENDER','NAME_TYPE_SUITE','NAME_INCOME_TYPE','NAME_EDUCATION_TYPE','NAME_FAMILY_STATUS','NAME_HOUSING_TYPE','OCCUPATION_TYPE','WEEKDAY_APPR_PROCESS_START','ORGANIZATION_TYPE','FLAG_OWN_CAR','FLAG_OWN_REALTY','LIVE_CITY_NOT_WORK_CITY','REG_CITY_NOT_LIVE_CITY','REG_CITY_NOT_WORK_CITY','REG_REGION_NOT_WORK_REGION','LIVE_REGION_NOT_WORK_REGION','REGION_RATING_CLIENT','WEEKDAY_APPR_PROCESS_START','REGION_RATING_CLIENT_W_CITY']

In previous_data,

10 columns are made categorical
['NAME_CONTRACT_STATUS','NAME_PAYMENT_TYPE','CODE_REJECT_REASON',
'NAME_CLIENT_TYPE','NAME_PORTFOLIO','CHANNEL_TYPE','NAME_YIELD_GROUP',
'PRODUCT_COMBINATION','NAME_CONTRACT_TYPE','DAYS_DECISION_GROUP']

Null Value Data Imputation

Strategy for application_data:

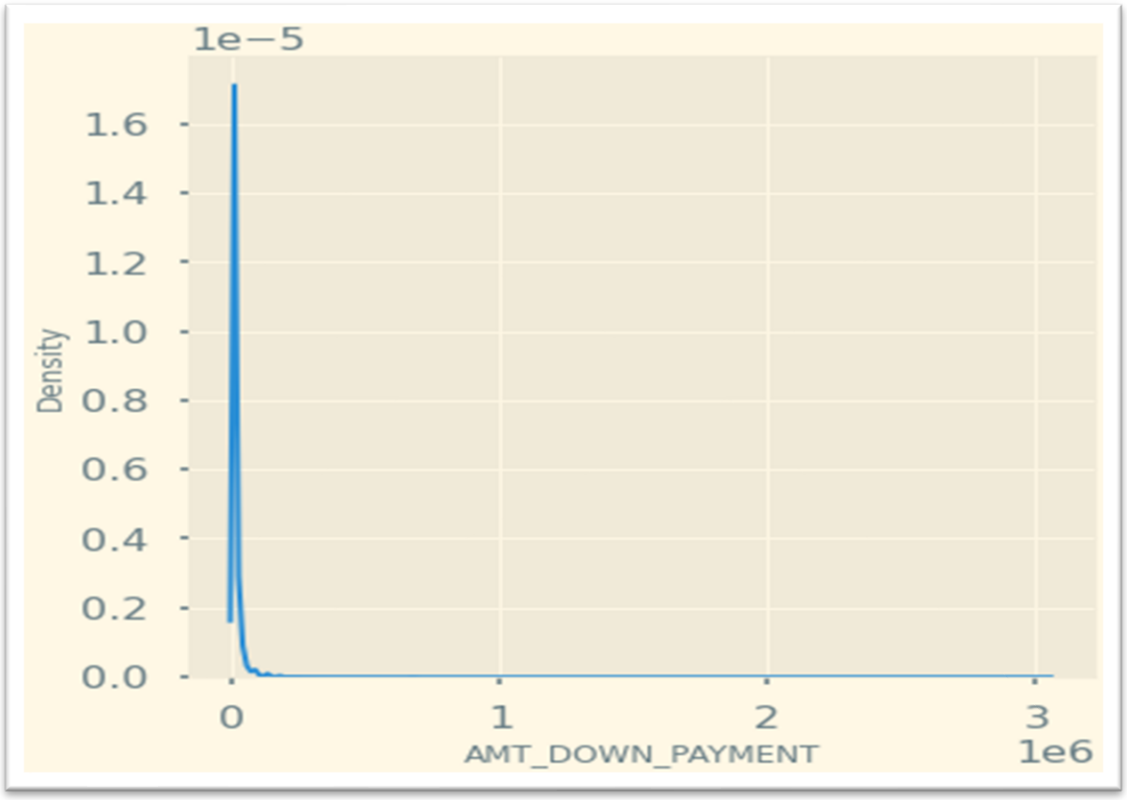
1. To impute null values in categorical variables which has lower null percentage, mode() is used to impute the most frequent items.
2. To impute null values in categorical variables which has higher null percentage, a new category is created.
3. To impute null values in numerical variables which has lower null percentage, median() is used as
4. There are no outliers in the columns

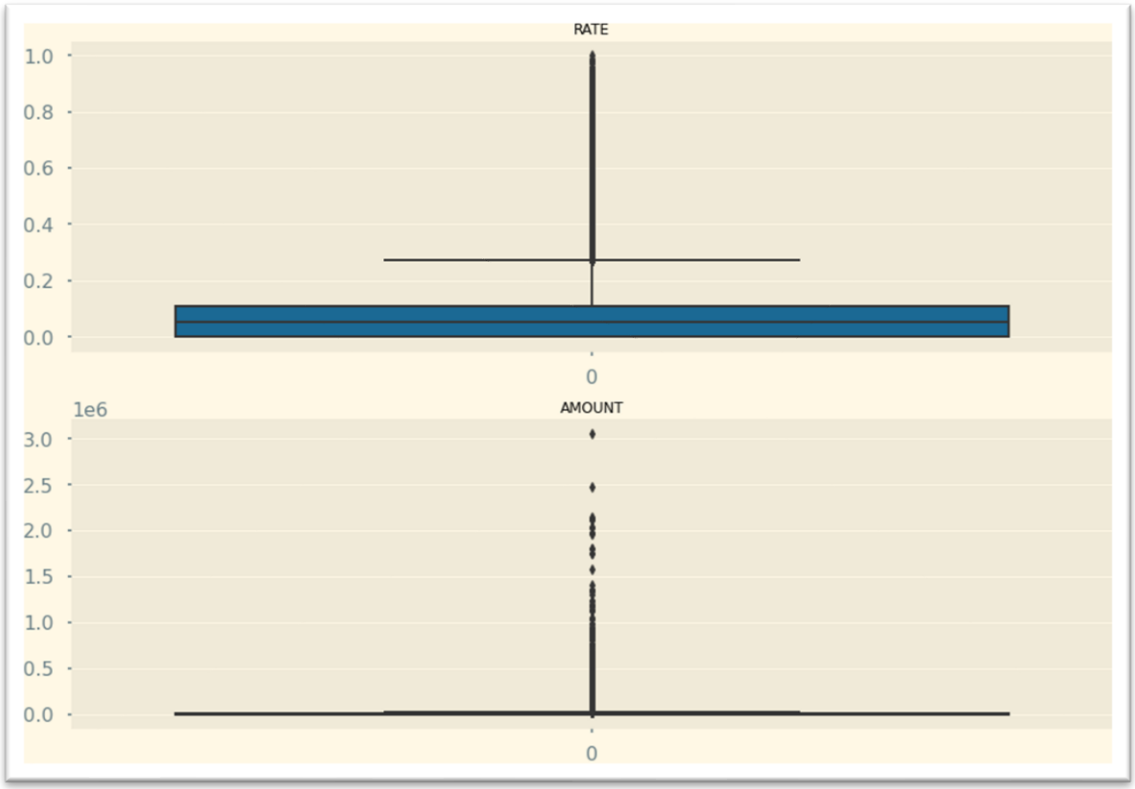
- Mean returned decimal values and median returned whole numbers and the columns were number of requests

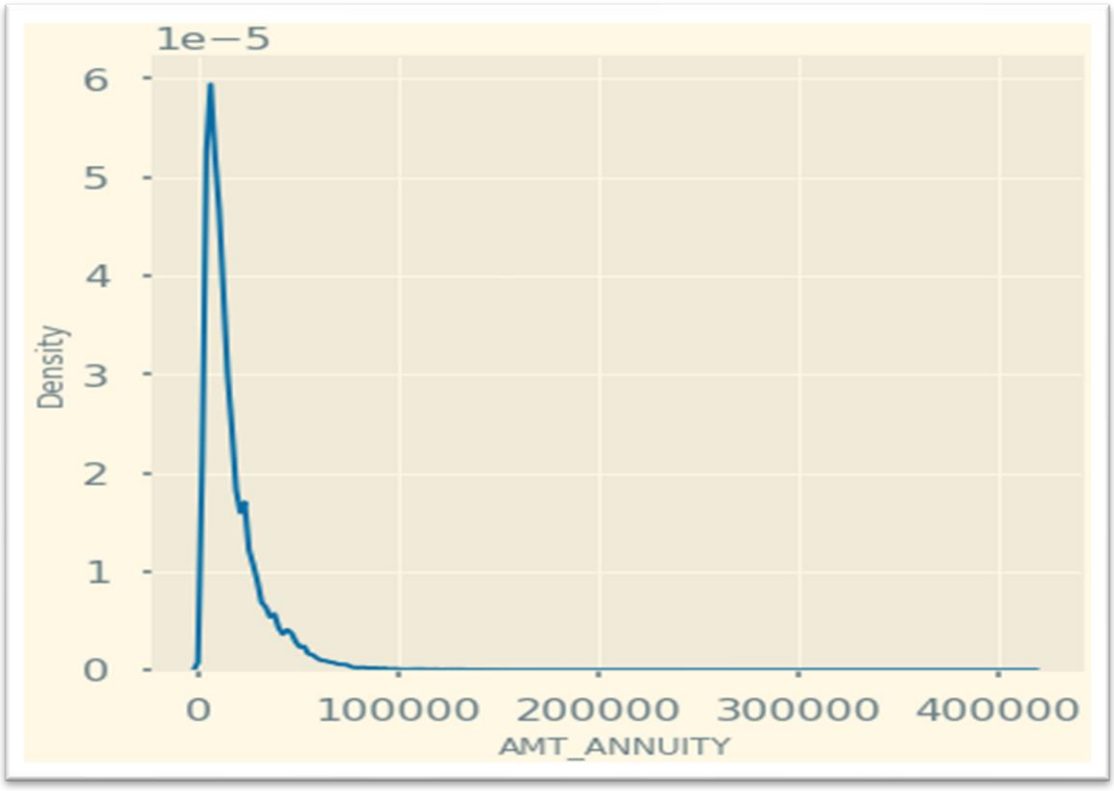
1. Impute categorical variable 'NAME_TYPE_SUITE' which has lower null percentage(0.42%) with the most frequent category using mode()[0]
2. Impute categorical variable 'OCCUPATION_TYPE' which has higher null percentage(31.35%) with a new category as assigning to any existing category might influence the analysis.
3. Impute numerical variables with the median as there are no outliers that can be seen from results of describe() and mean() returns decimal values and these columns represent number of enquiries made which cannot be decimal.
4. Impute values for EXT_SOURCE_1 with respective values of EXT_SOURCE_2 and EXT_SOURCE_3.Impute remaining null values with mean.

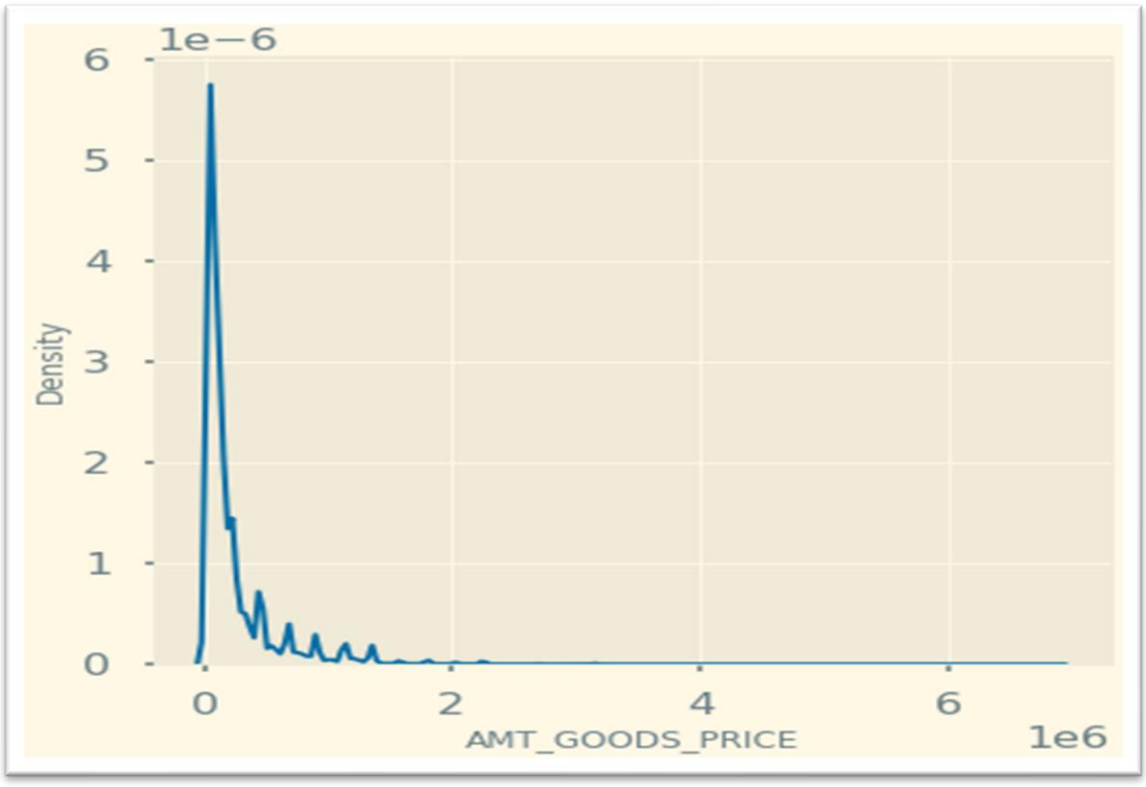
Strategy for Previous _data:

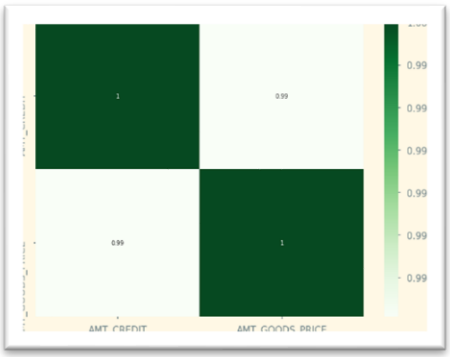
1. To impute null values and negative values in AMT_DOWN_PAYMENT
2. To impute null values in numerical column, we analysed the loan status and assigned values.
3. To impute null values in continuous variables, we plotted the distribution of the columns and used median if the distribution is skewed and mode if the distribution pattern is preserved
4. To impute null values in categorical columns mode is used


1e−5
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Density
0
1
2
3
AMT_DOWN_PAYMENT
1e6

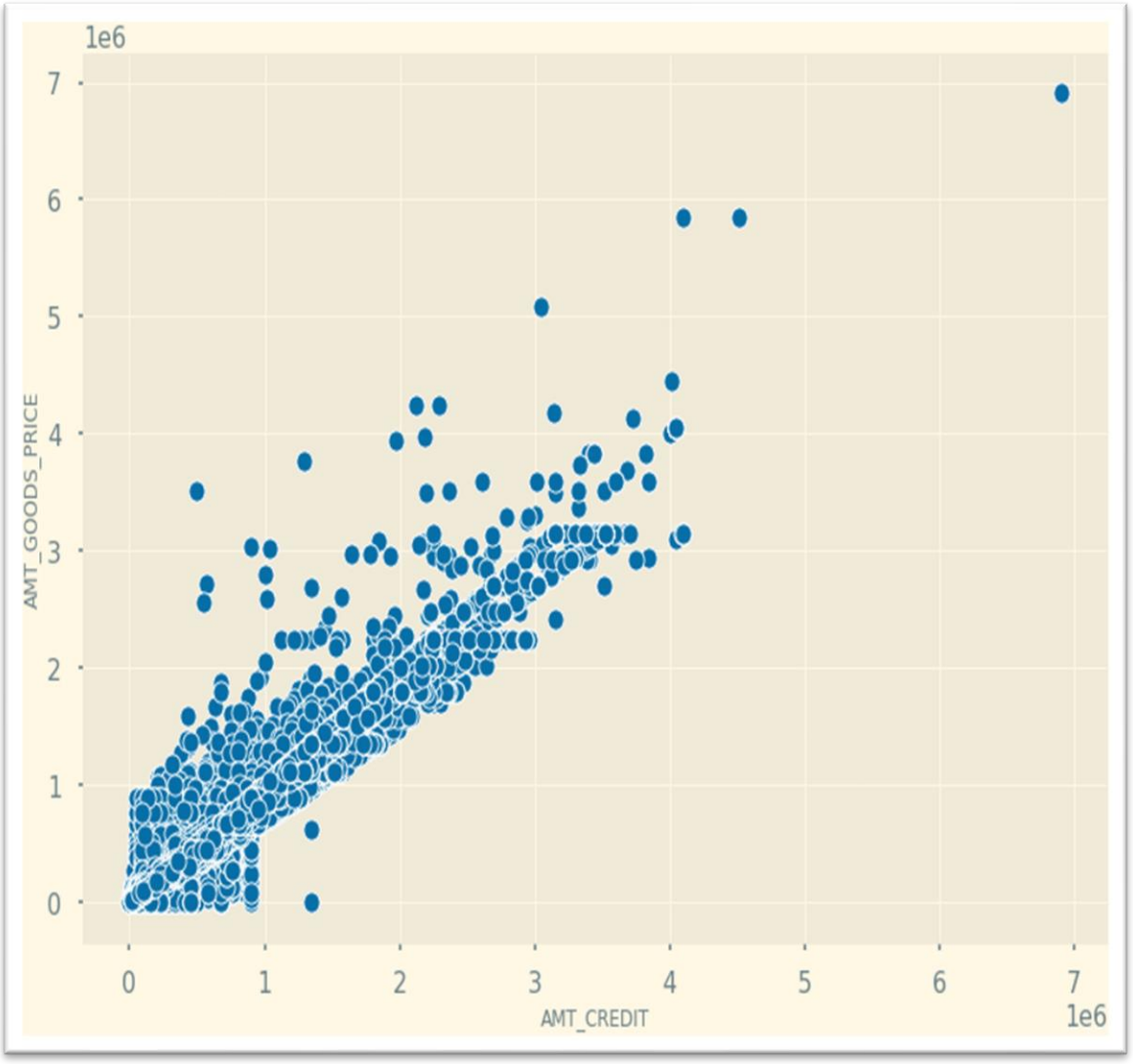

RATE
1.0
0.8
0.6
0.4
0.2
0.0
0
1e6
AMOUNT
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0

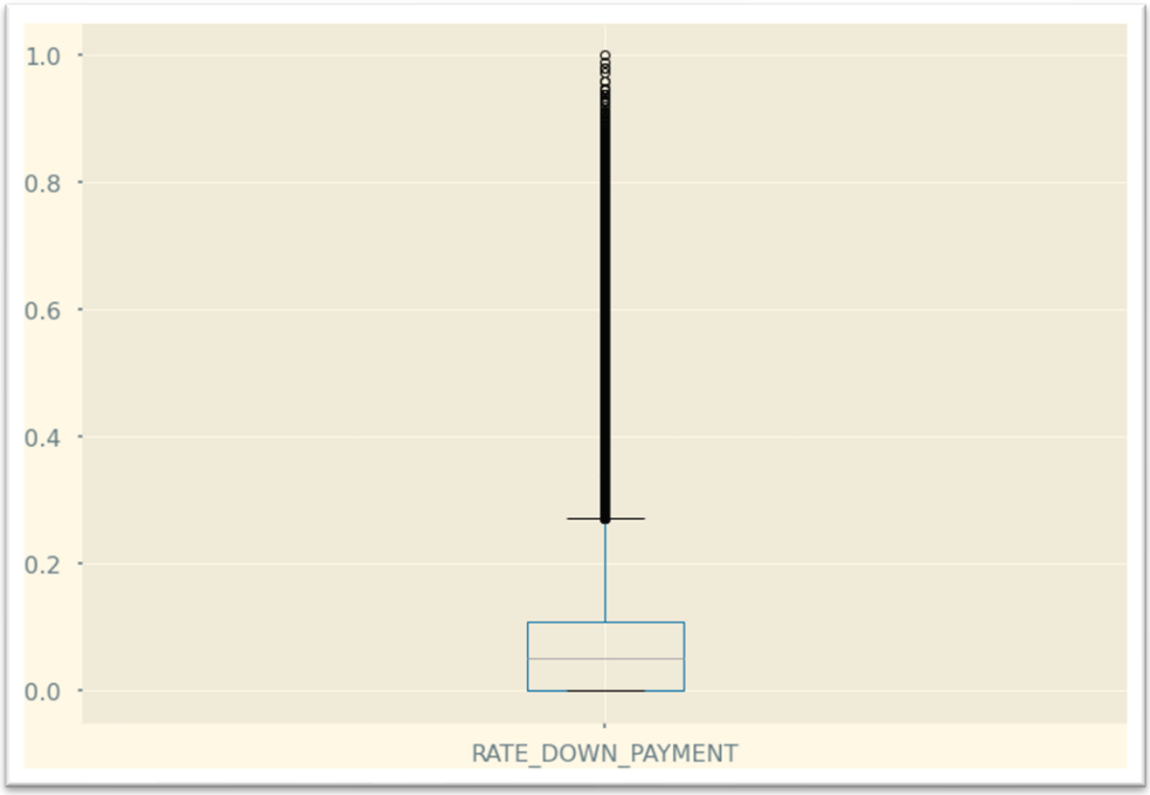

1e−5
6
5
4
3
2
1
0
Density
0
100000
200000
300000
400000
AMT_ANNUITY

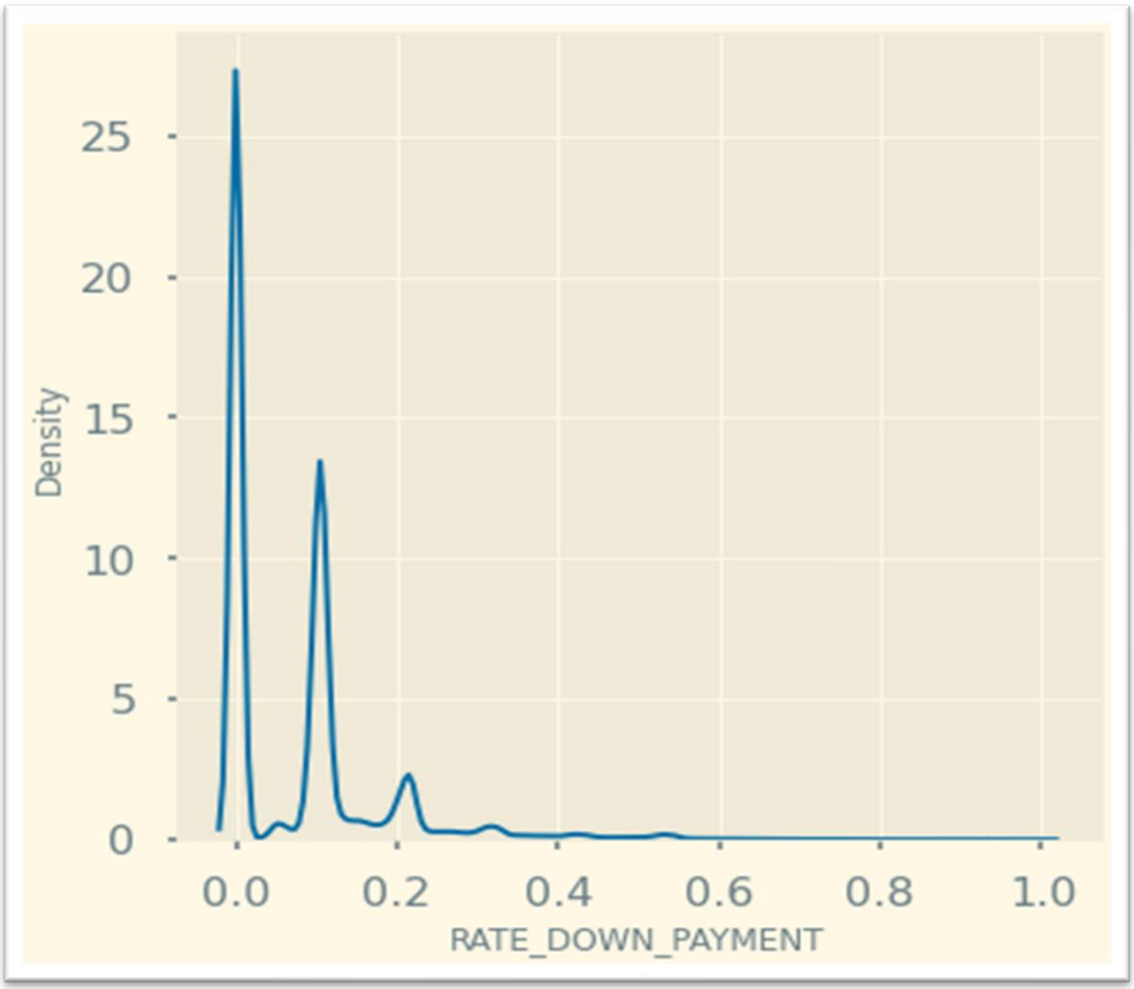

1e−6
6
5
4
3
2
1
0
Density
0
2
4
6
AMT_GOODS_PRICE
1e6


1
0.99
0.99
1
AMT_CREDIT
AMT_GOODS_PRICE


1
0.79
0.79
1
AMT_CREDIT
AMT_DOWN_PAYMENT


1e6
7
6
5
4
3
2
1
0
AMT_GOODS_PRICE
0
1
2
3
4
5
6
7
AMT_CREDIT
1e6


1.0
0.8
0.6
0.4
0.2
0.0
RATE_DOWN_PAYMENT


25
20
15
10
5
0
Density
0.0
0.2
0.4
0.6
0.8
1.0
RATE_DOWN_PAYMENT

Bank_loan_services_Casestudy

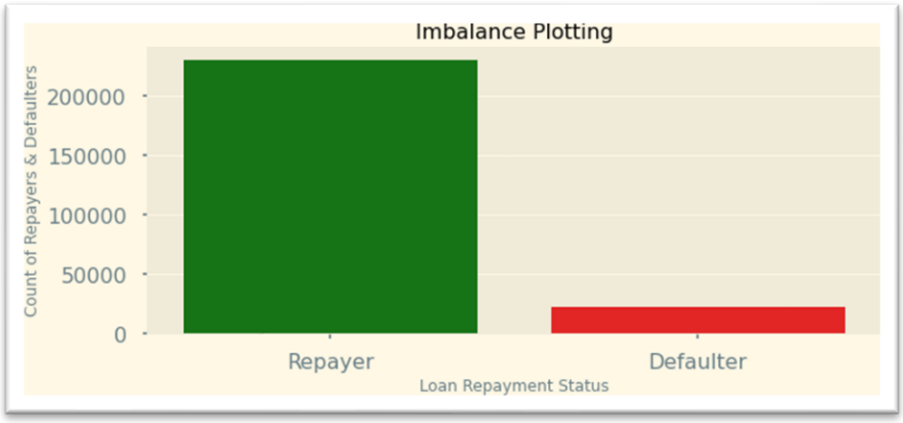
Data Analysis

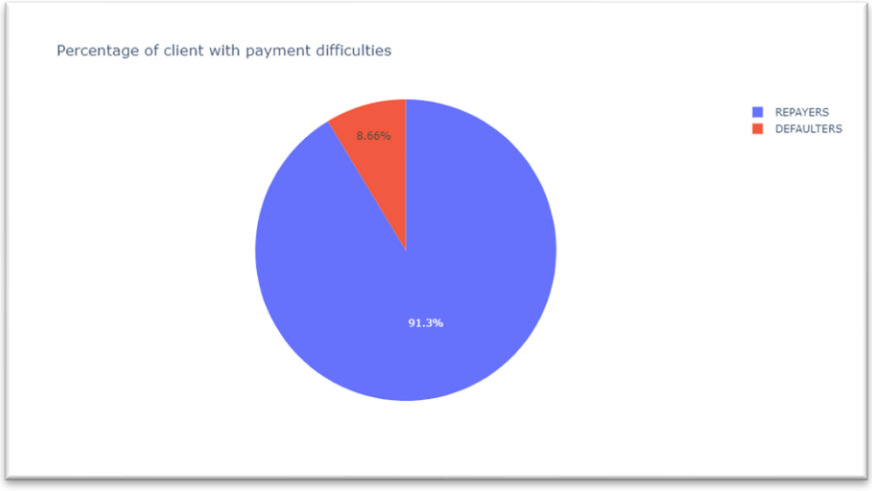
Strategy:

The data analysis flow has been planned in following way :

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

Imbalance Analysis





Bank_loan_services_Casestudy

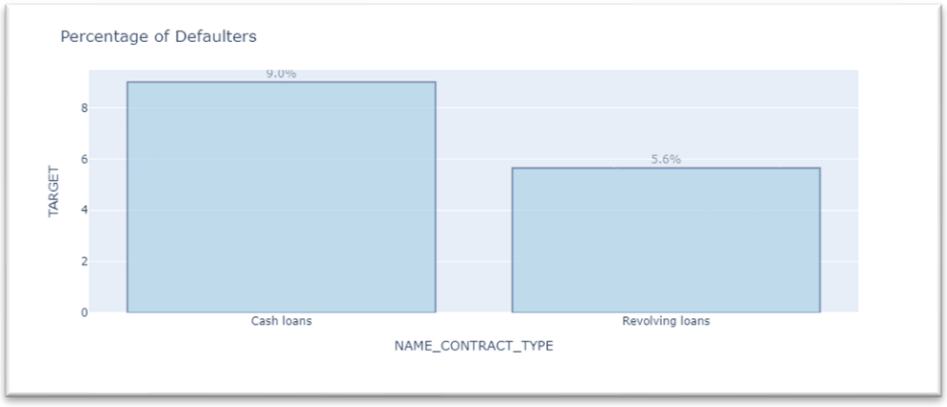
Plotting Functions

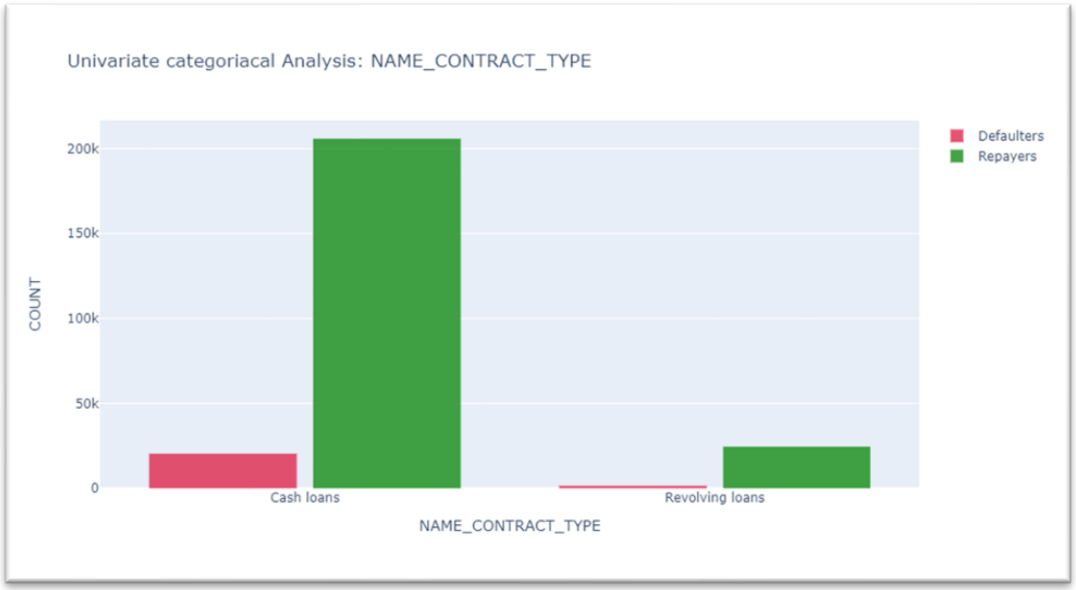
1. function for plotting repetitive histograms(count) in univariate categorical analysis on application_data
2. function for bar chart to plot percentage Defaulters
3. bivariate_bar function bivariate analysis
4. function for plotting repetitive relation plots in bivariate numerical analysis on application_data
5. univariate_merged function for univariate analysis on merged dataframe

Categorical Variables Analysis

1. Categorical segmented Univariate Analysis

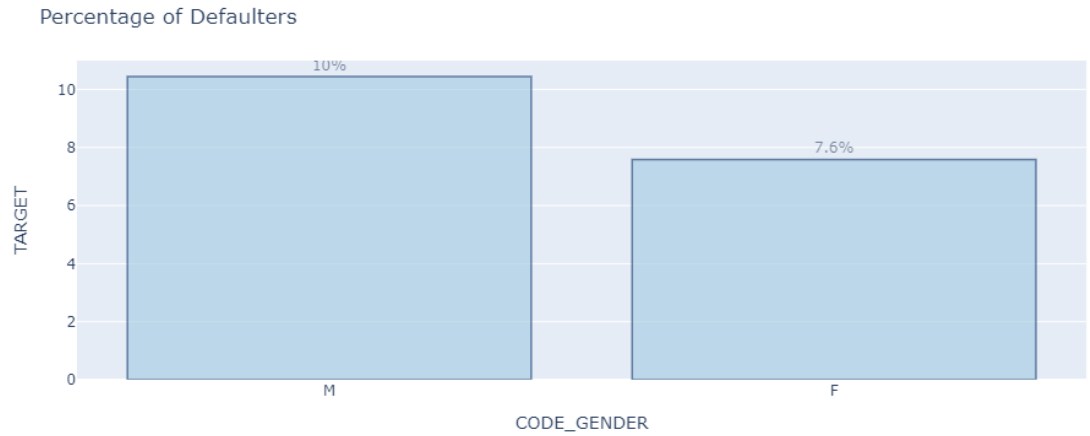
For application data:

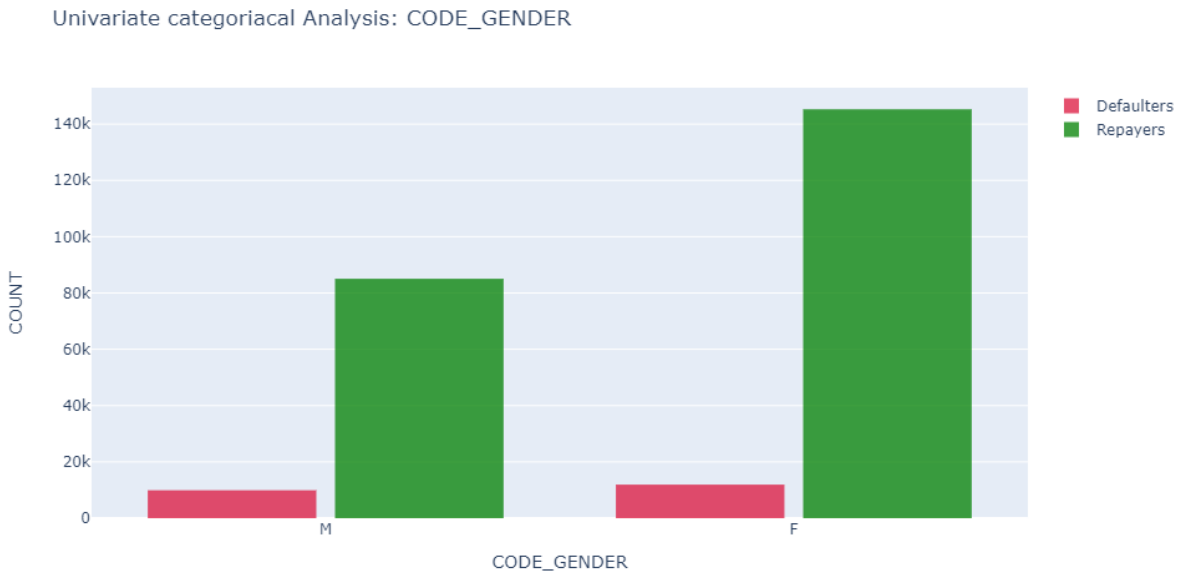




Inferences:

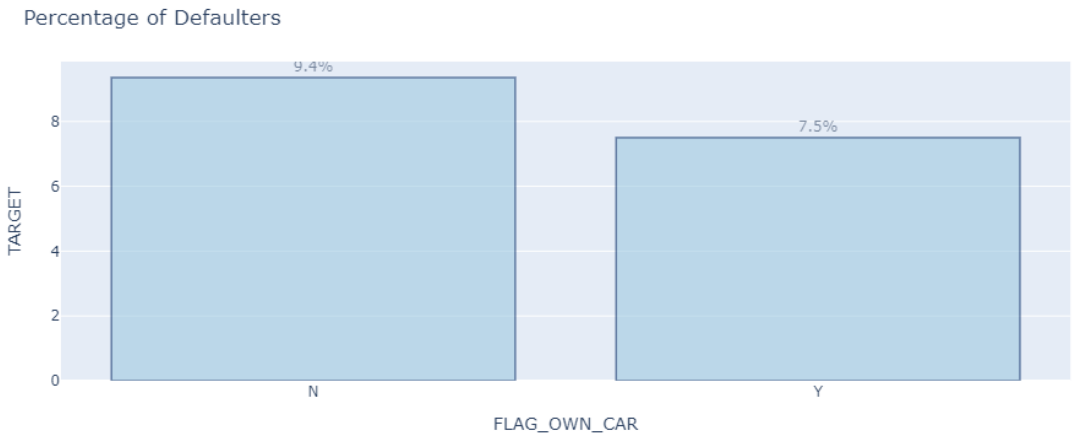
1. Contract type: Revolving loans are just a small fraction (10.3%) from the total number of loans; in the same time,5% defaulters are for Revolving loans, comparing with their frequency.
2. Cash loans have 9% defaulting rate which is more than for revolving loans comparatively

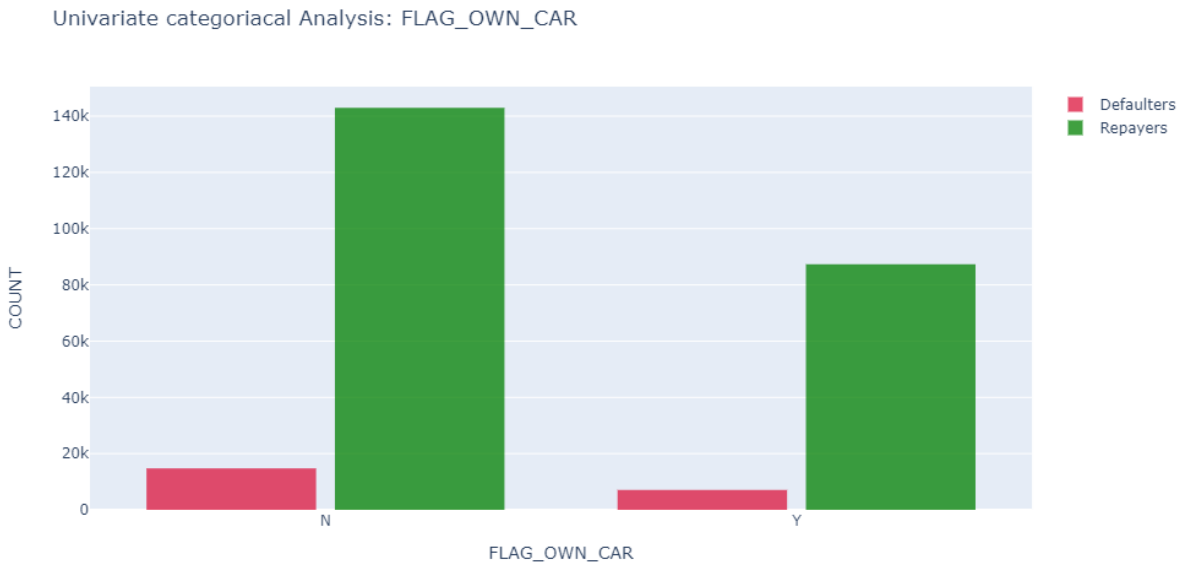




Inferences:

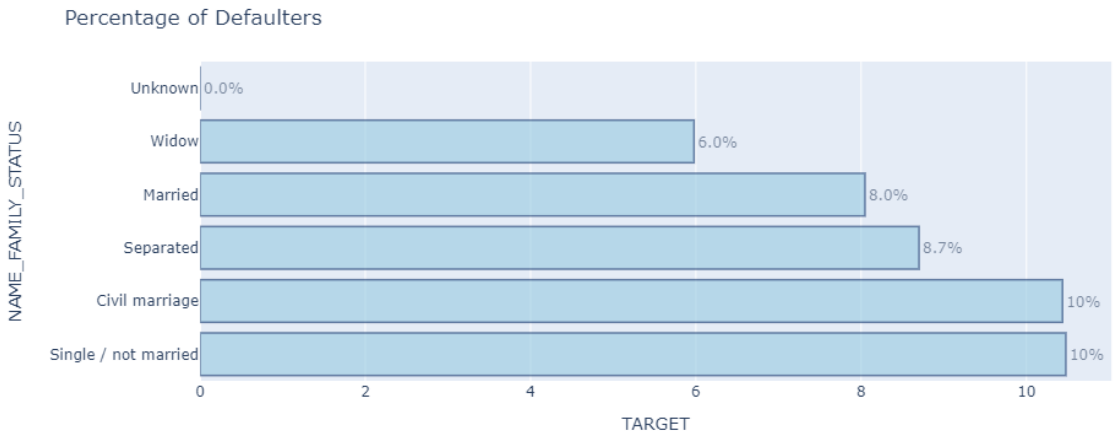
- The number of female clients is almost double the number of male clients. Based on the percentage of defaulted credits, males have a higher chance of not returning their loans (nearly 10%), comparing with women (nearly 7%)

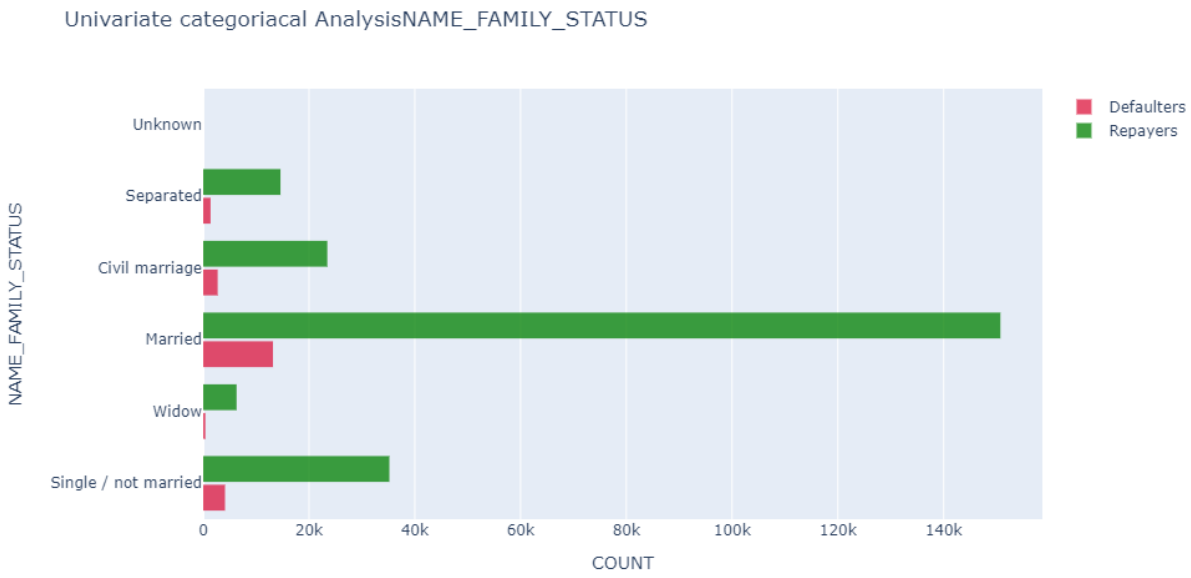




Inferences:

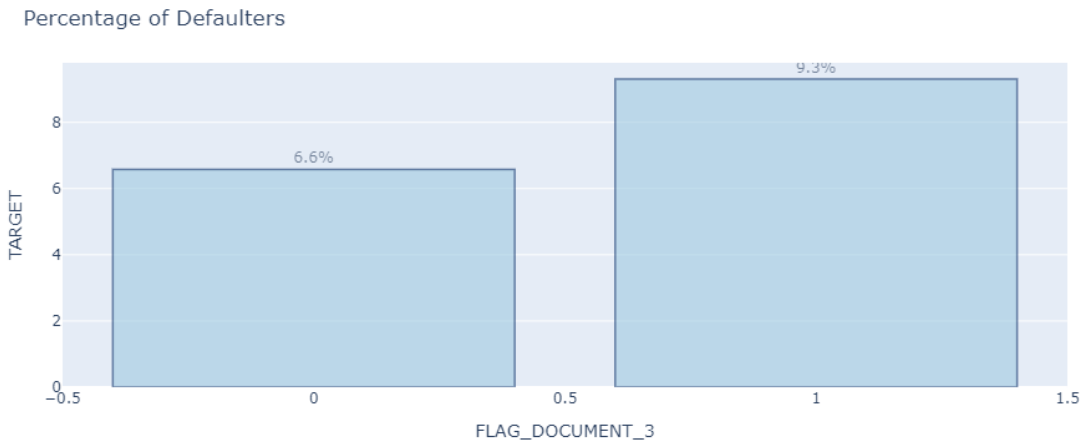
Clients who own a car are half in number of the clients who don't own a car. But based on the percentage of default, there is no correlation between owning a car and loan repayment as in both cases the default percentage has just a difference of 2%.





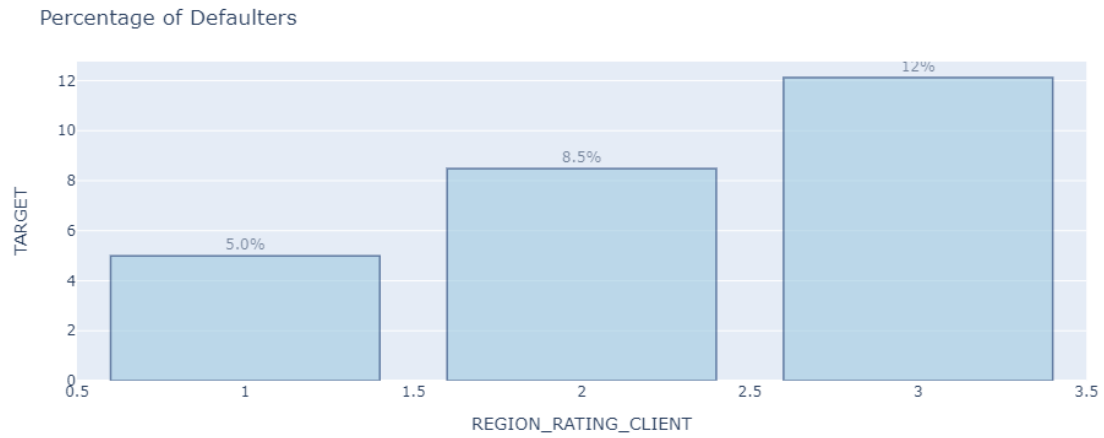
Inferences:

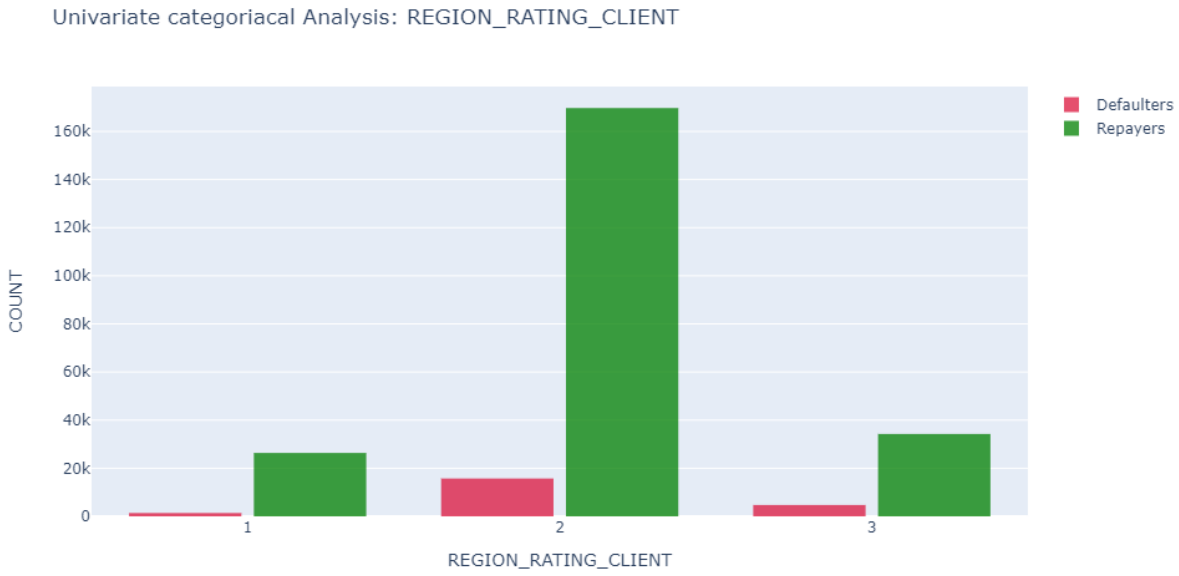
1. Most of the people who have taken loan are married, followed by Single/not married and civil marriage
2. In terms of percentage of defaulters, Civil marriage and single/not married has the highest percent of no repayment (10%), with Widow the lowest (exception being Unknown).



Inferences:

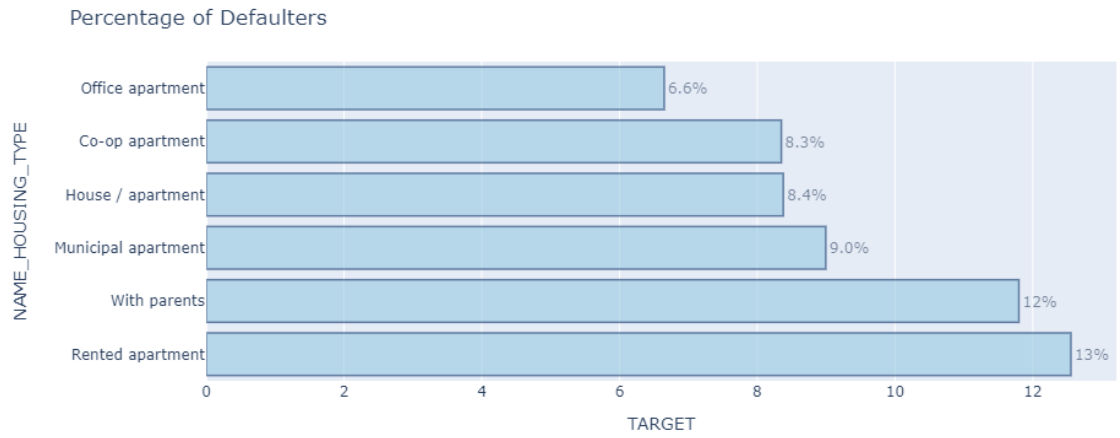
1. There is no significant correlation between repayers and defaulters in terms of submitting document 3 as we see even if applicants have submitted the document, they have defaulted a slightly more (9.3%) than who have not submitted the document (6.6%)

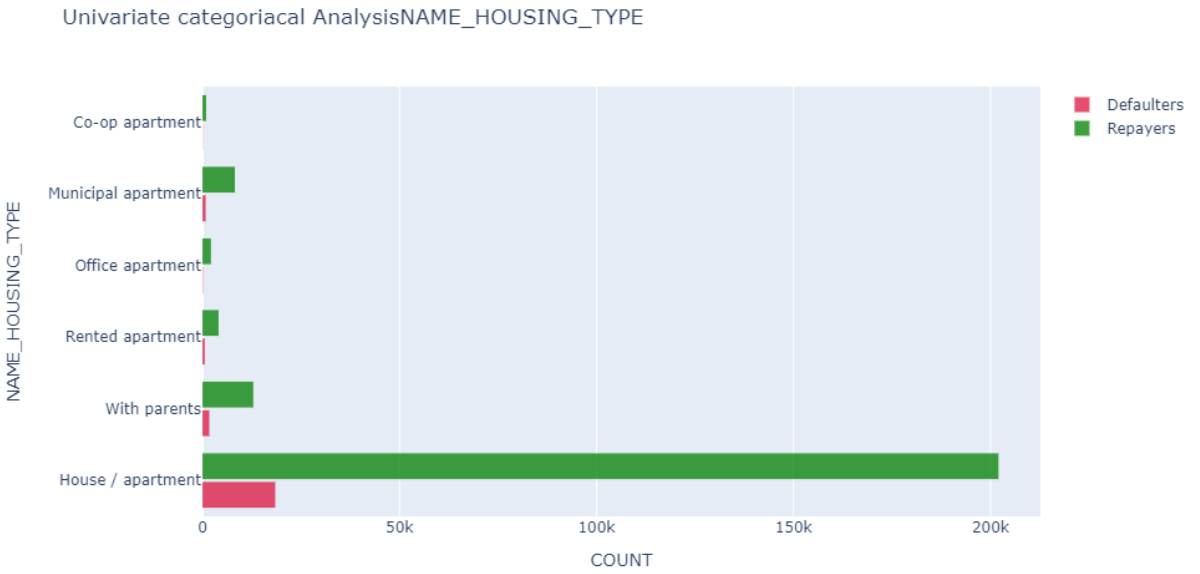




Inferences:

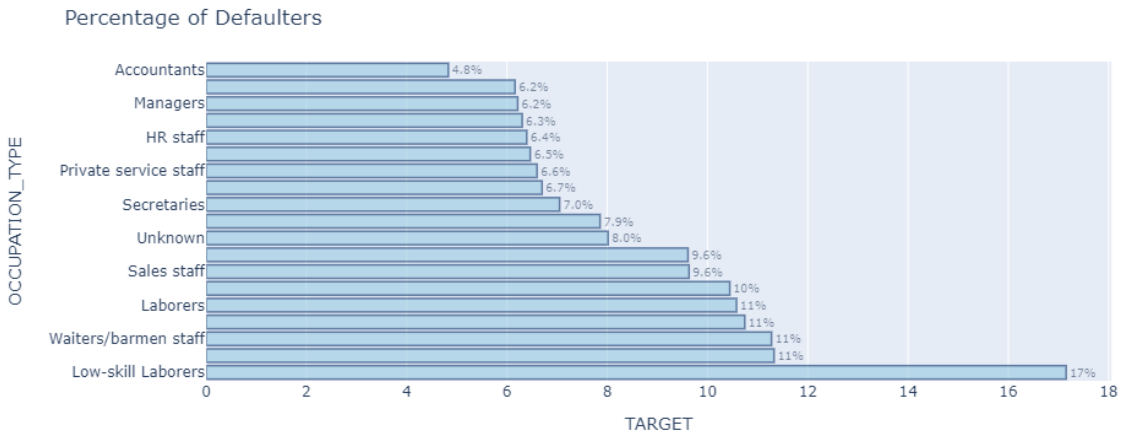
1. Most of the applicants are living in Region_Rating 2 place.
2. Region Rating 3 has the highest default rate (12%)
3. Applicant living in Region_Rating 1 has the lowest probability of defaulting, thus safer for approving loans

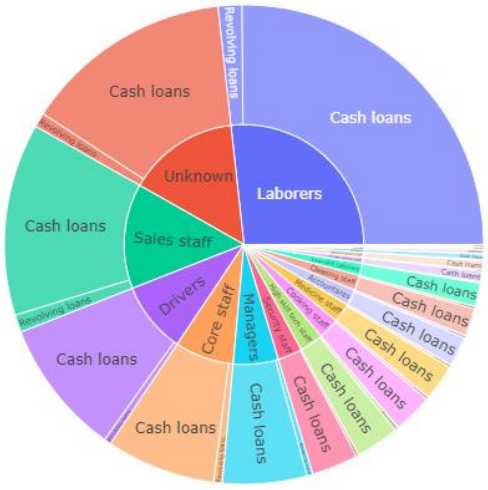




Inferences:

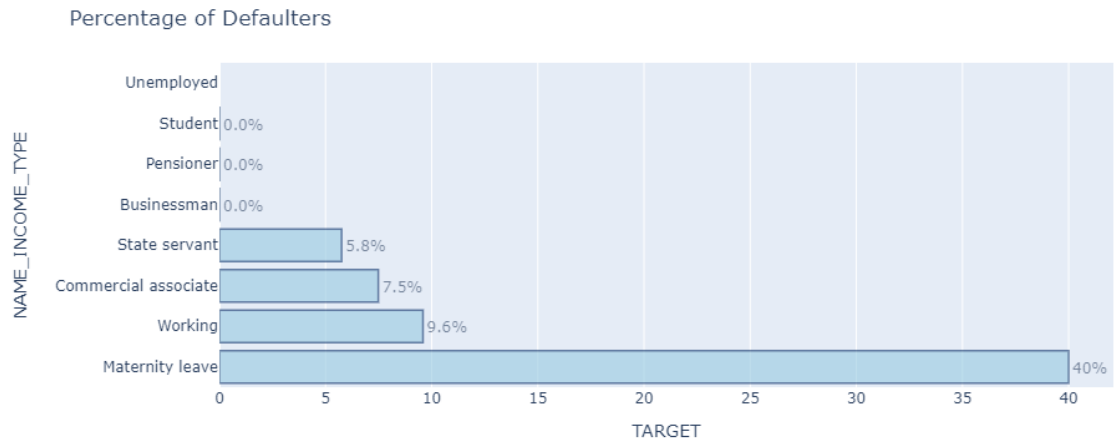
1. Majority of people live in House/apartment
2. People living in office apartments have lowest default rate
3. People living with parents (12%) and living in rented apartments (13%) have higher probability of defaulting

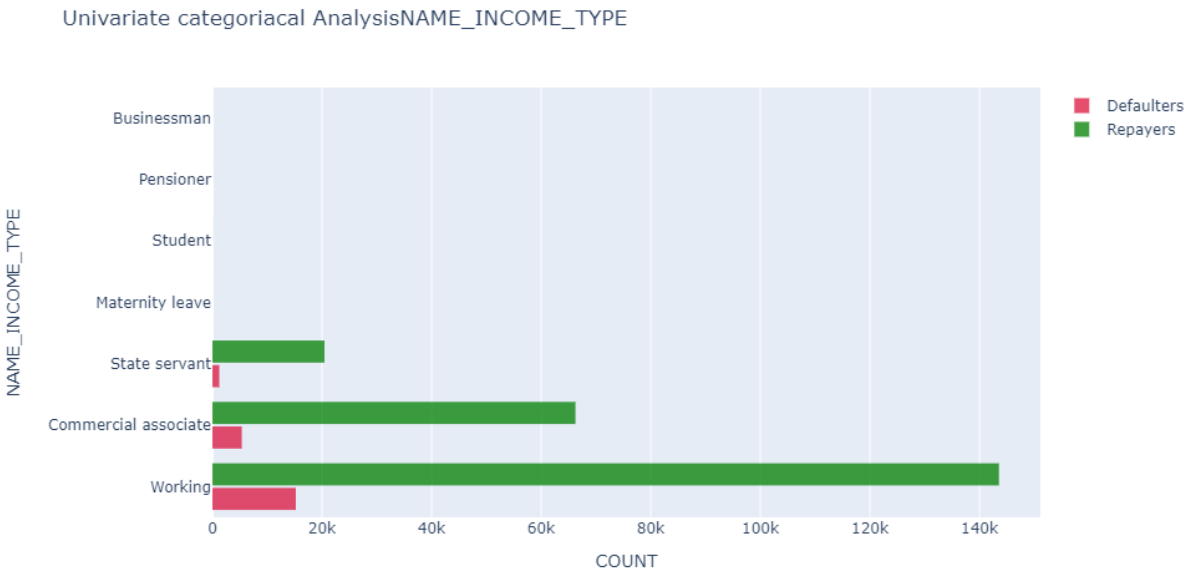




Inferences:

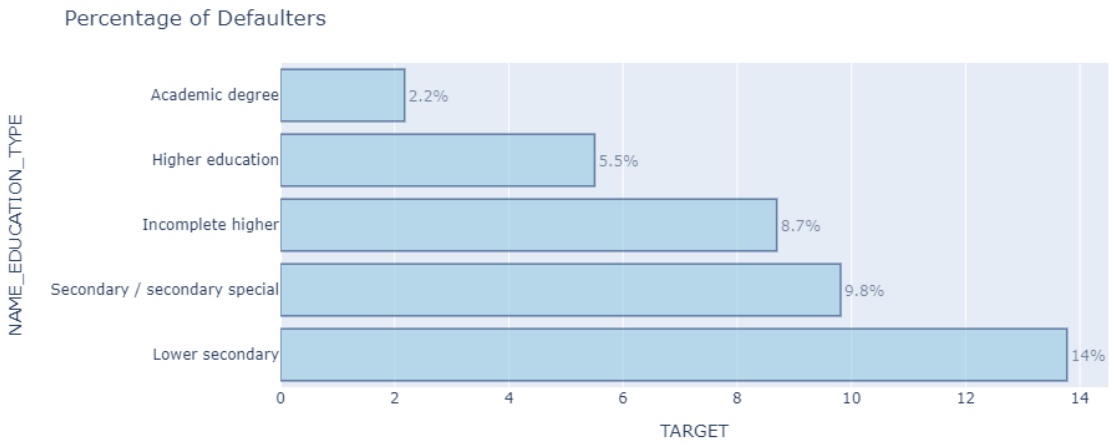
1. Most of the loans are taken by Laborers, followed by Sales staff. IT staff take the lowest amount of loans.
2. The category with highest percent of not repaid loans are Low-skill Laborers (above 17%), followed by Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff.

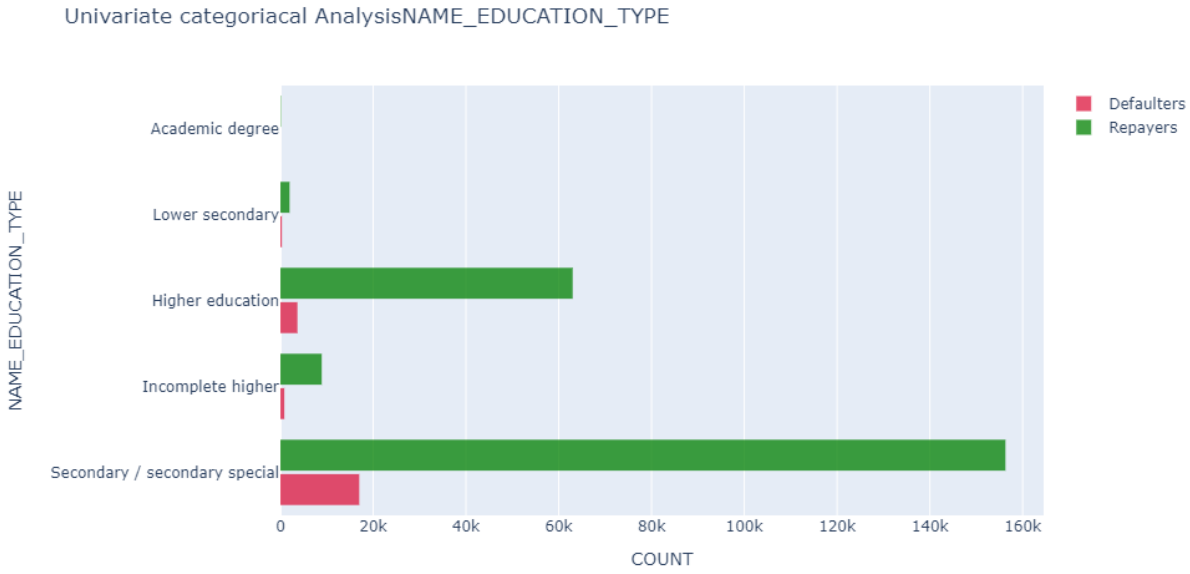




Inferences:

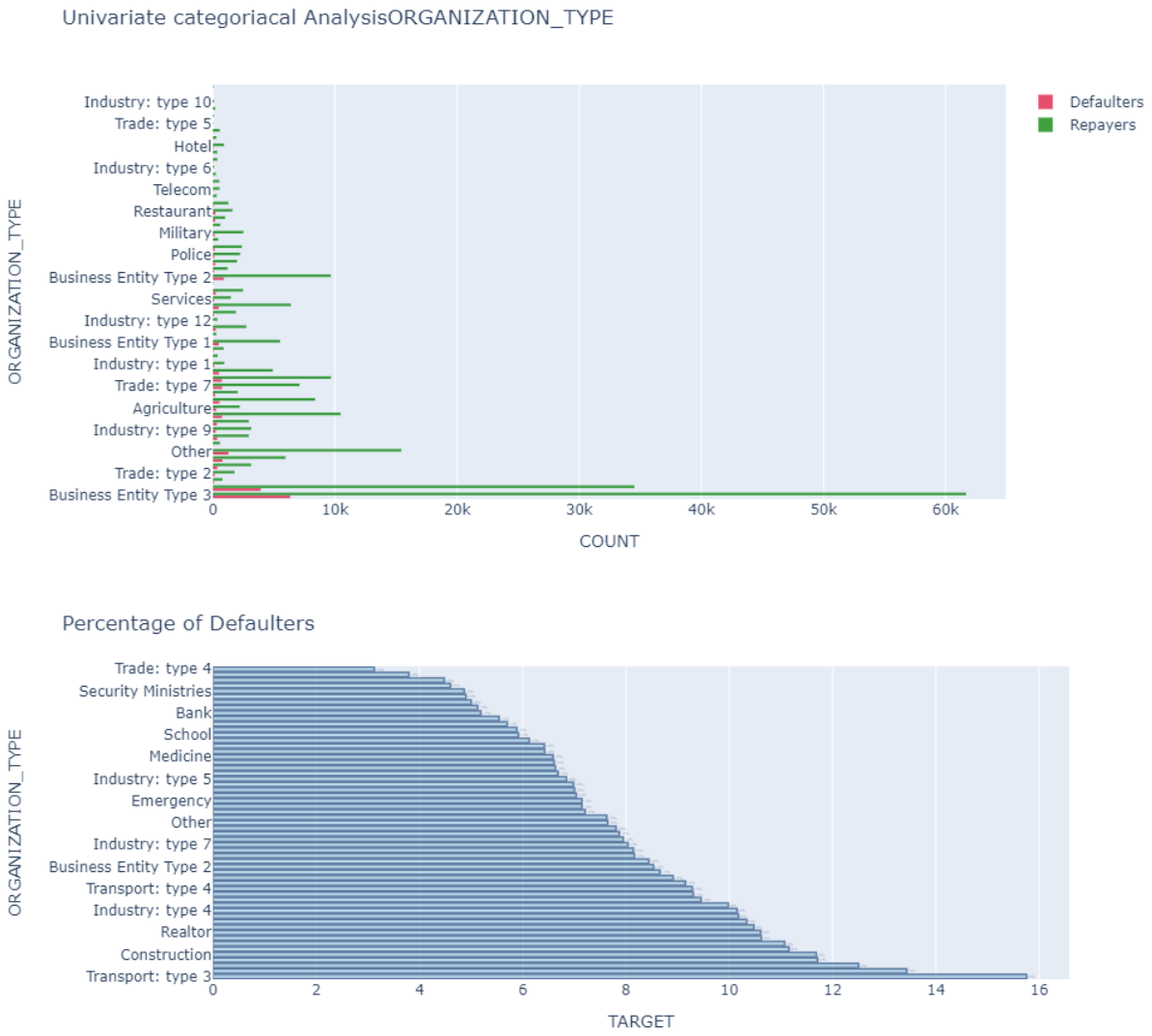
1. Most of applicants for loans have income type as Working, followed by Commercial associate, Pensioner and State servant.
2. The applicants with the type of income Maternity leave have almost 40% ratio of not returning loans, followed by working (9.6%).
3. Student and Businessmen, Pensioners though less in numbers do not have any default record. Thus these three categories are safest for providing loan.





Inferences:

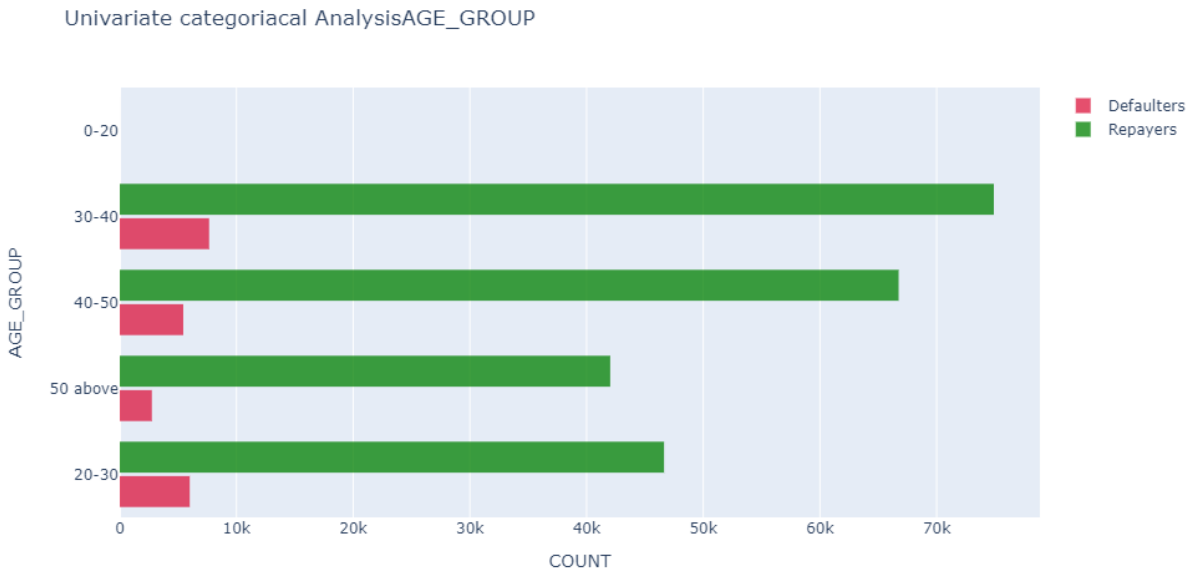
1. Majority of the clients have Secondary / secondary special education, followed by clients with Higher education. Only a very small number having an academic degree
2. The Lower secondary category, although rare, have the largest rate of not repaying the loan (14%). The people with Academic degree have 2.2% defaulting rate.

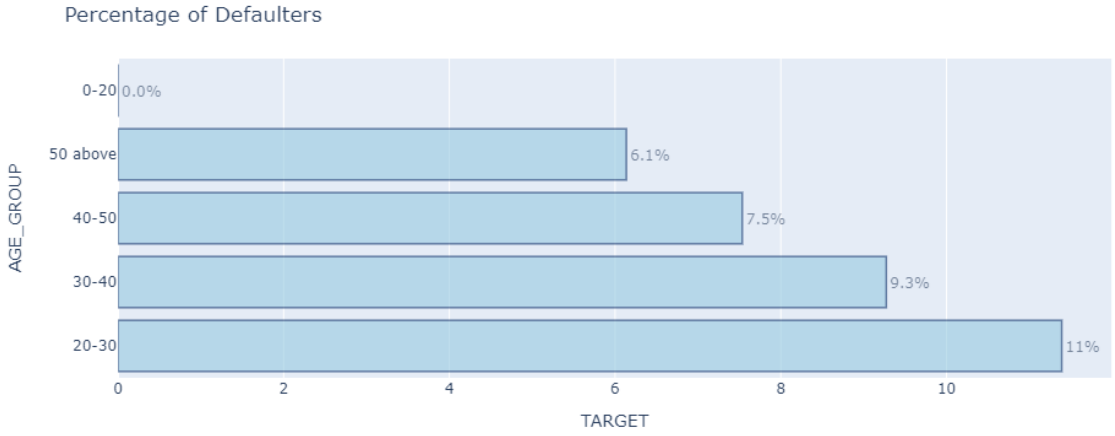


Inferences:

1. Organizations with highest percent of loans not repaid are Transport: type 3 (15.75%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self employed people have relative high defaulting rate (10.17%), and thus should be avoided to be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting.
2. Most of the people application for loan are from Business Entity Type 3
3. Though business entity type 2 has more loans disbursed, they have comparatively lees default rate(8.5%)
4. It can be seen that following category of organization type has lesser defaulters thus safer for providing loans:

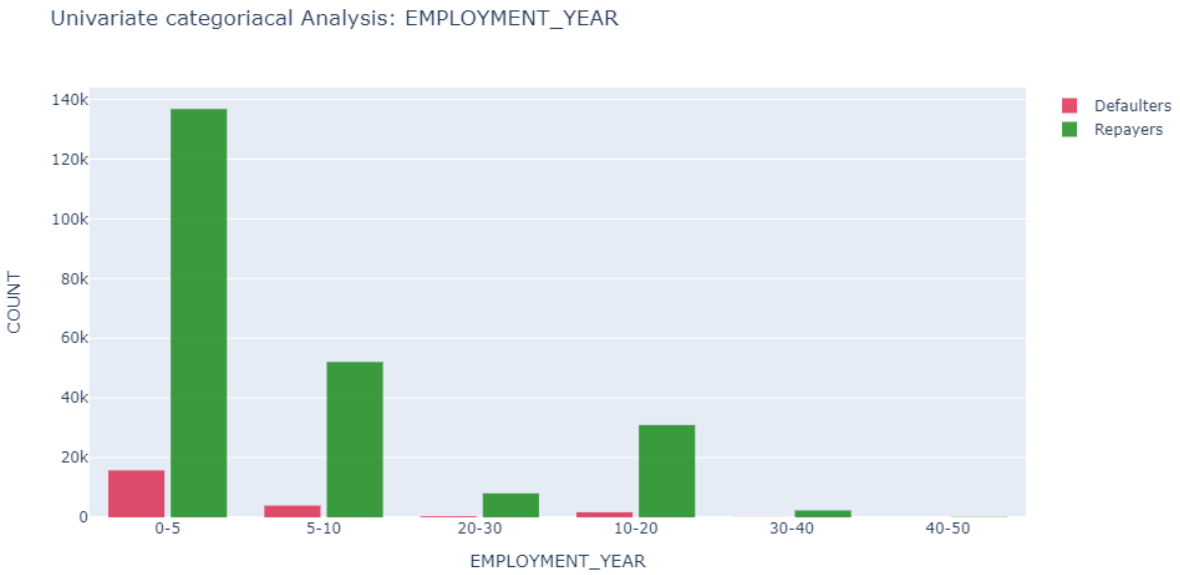
- Trade Type 4
- Securities Ministries

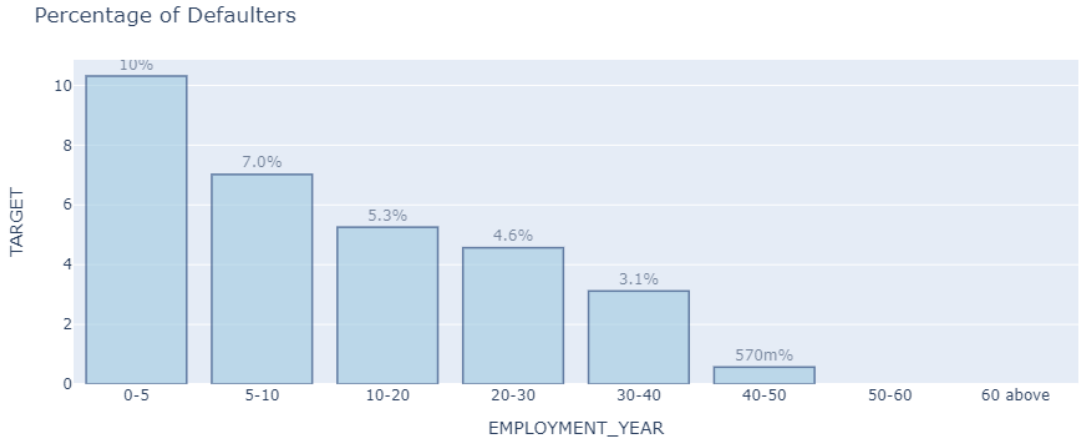




Inferences:

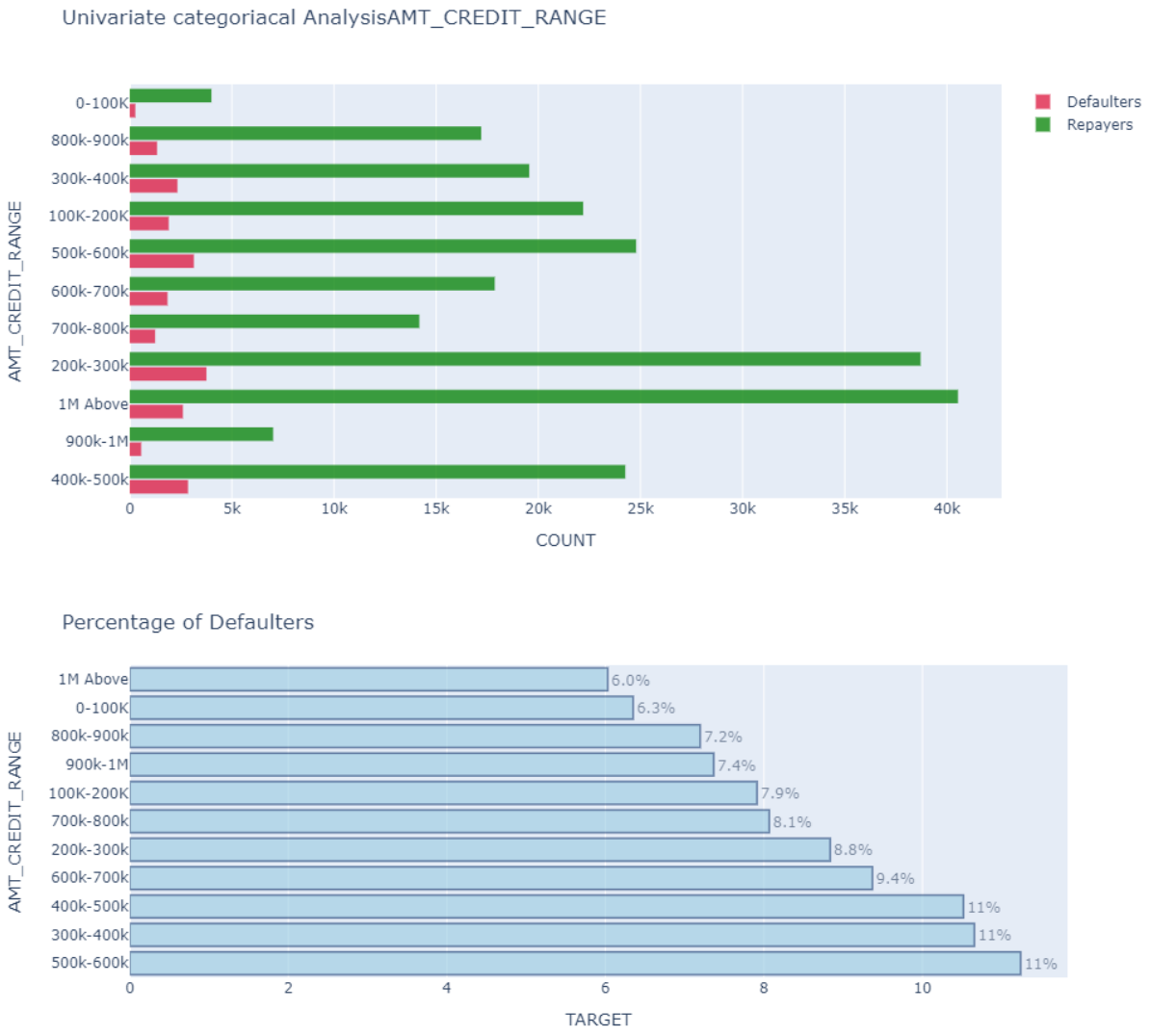
1. People in the age group range 20-40 have higher probability of defaulting
2. People above age of 50 have low probability of defaulting
3. loan given to above age 40 is safer as defaulting rate is comparatively less.





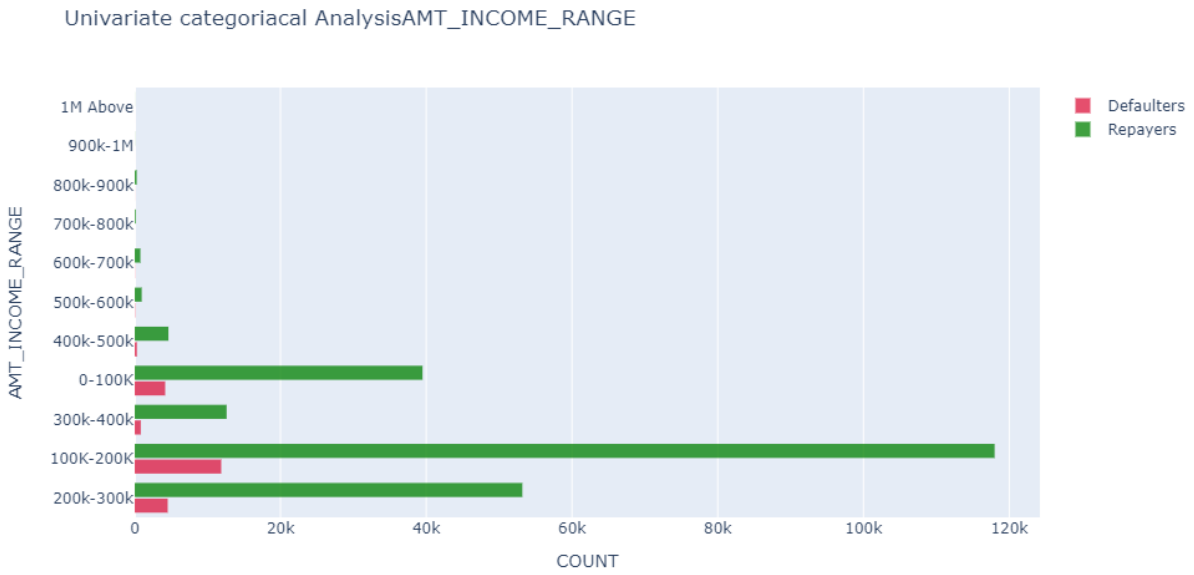
Inferences:

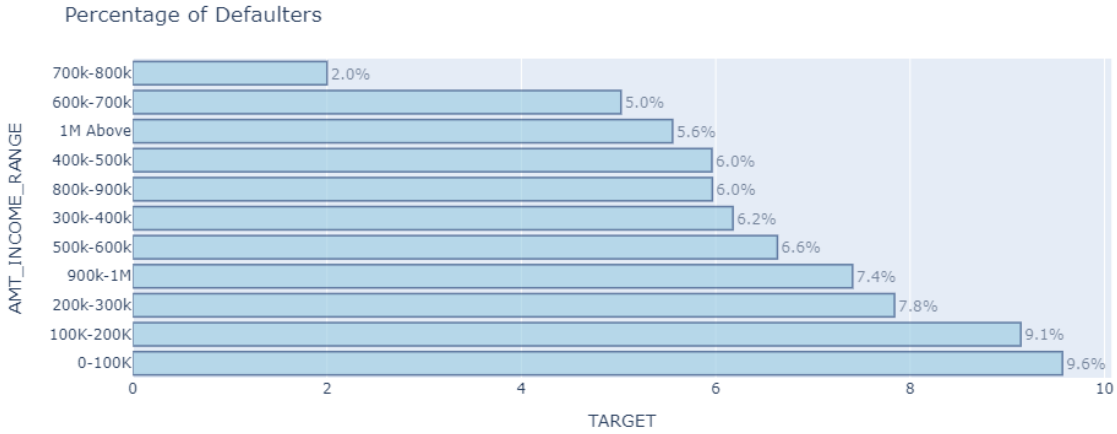
1. Majority of the applicants have been employed in between 0-5 years. The defaulting rating of this group is also the highest which is 10%
2. With increase of employment year, defaulting rate is gradually decreasing with people having 40+ year experience having less than 1% default rate



Inferences:

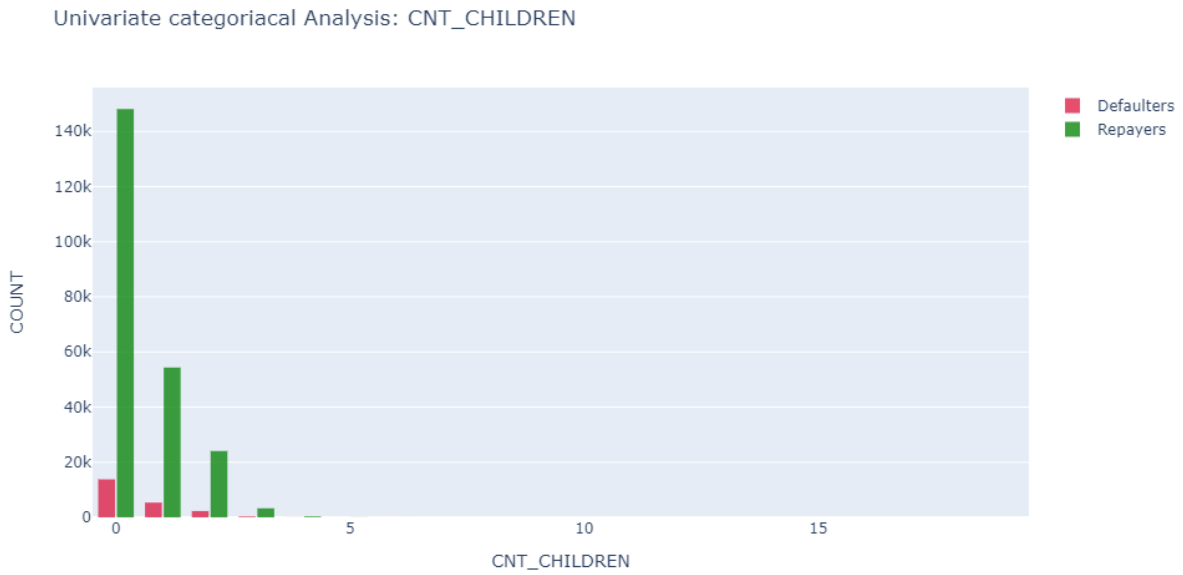
1. More than 80% of the loan provided are for amount less than 900,000
2. People who get loan for 300-600k tend to default more than others.

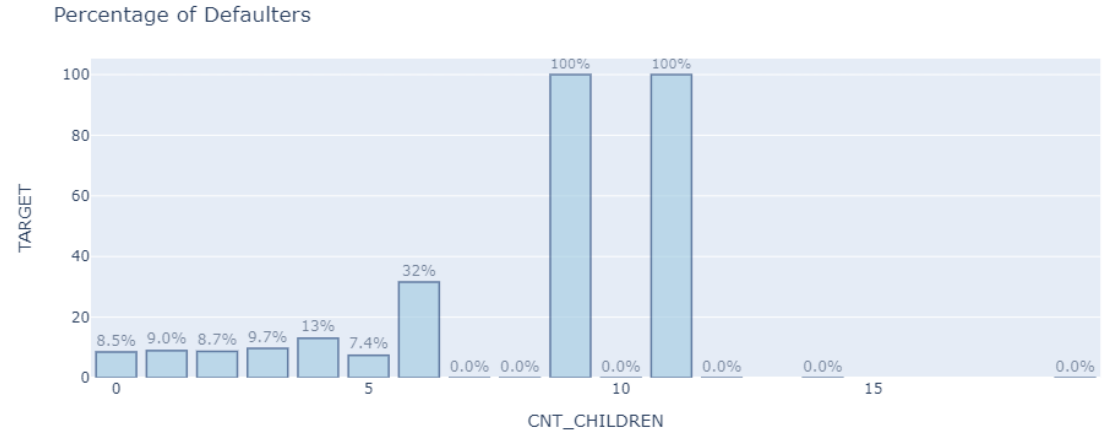




Inferences:

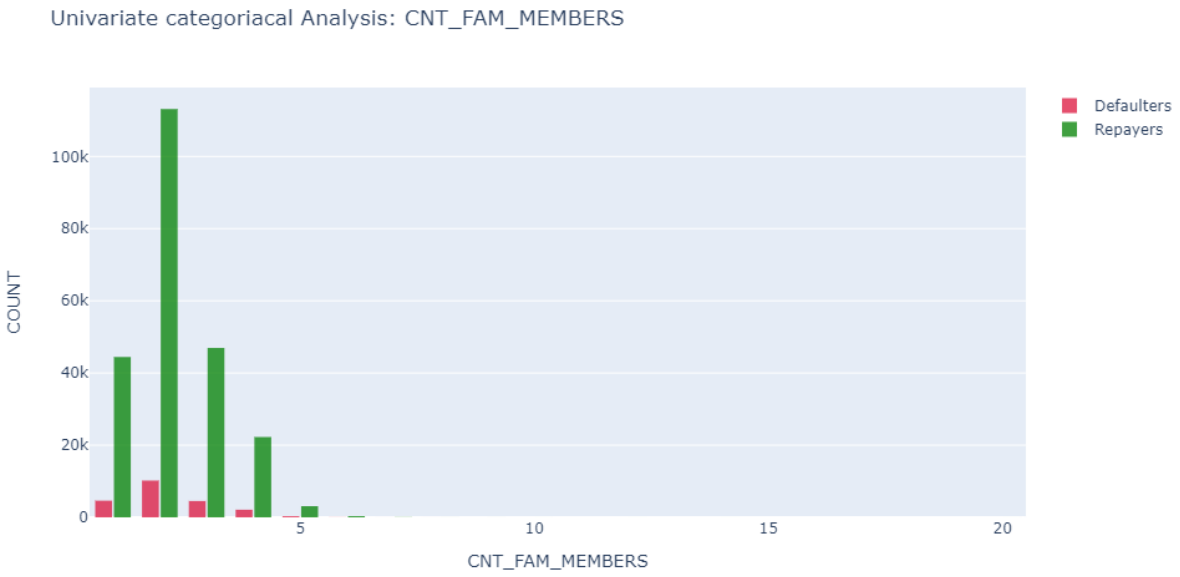
1. 90% of the applications have Income total less than 300,000
2. Application with Income less than 300,000 has high probability of defaulting
3. Applicant with Income more than 700,000 are less likely to default

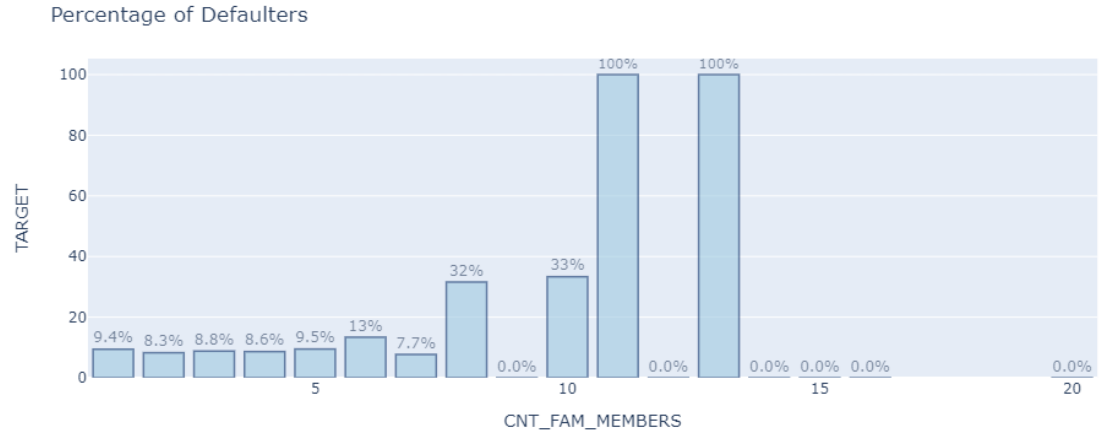




Inferences:

1. Most of the applicants do not have children
2. Very few clients have more than 3 children.
3. Client who have more than 4 children has a very high default rate with child count 9 and 11 showing 100% default rate



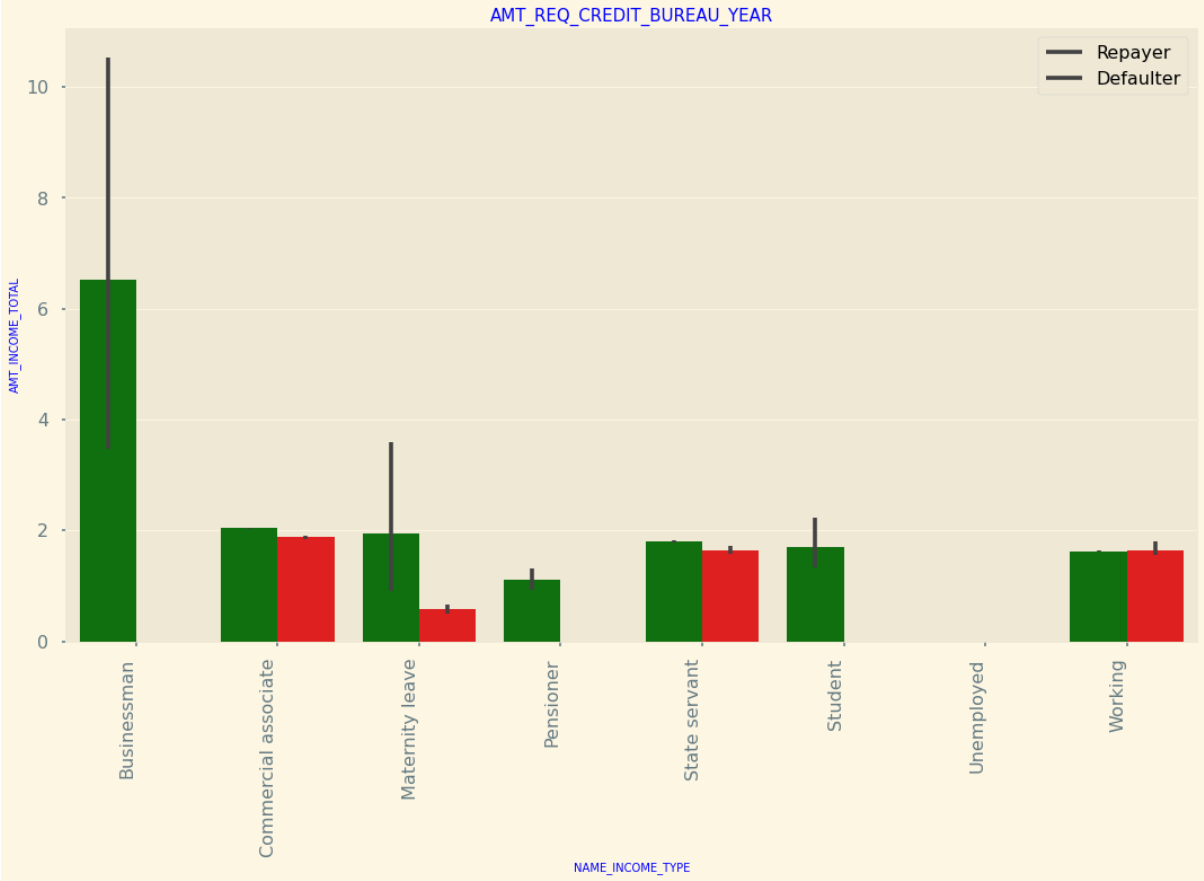


Inferences:

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

2. Categorical Bi/Multivariate analysis

For application data:



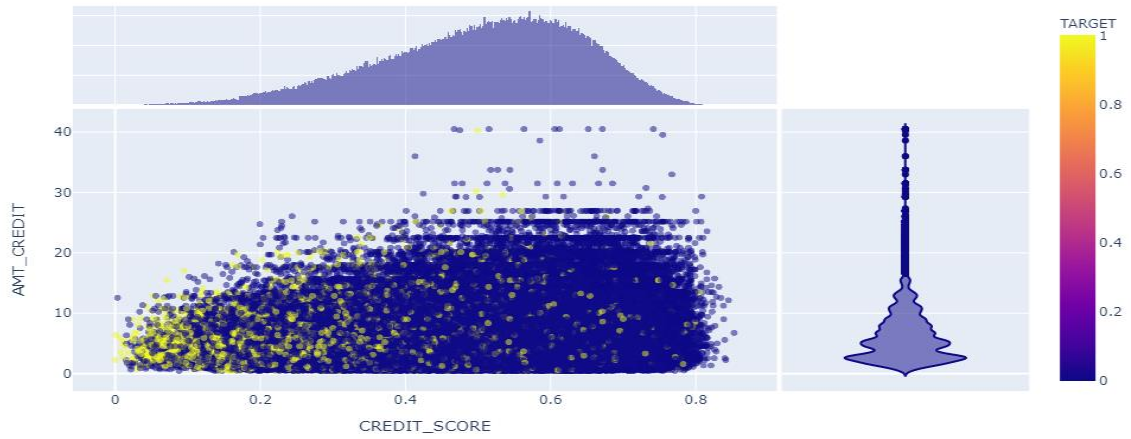
Inferences:

1. It can be seen that business man's income is the highest and the estimated range with default 95% confidence level seem to indicate that the income of a business man could be in the range of slightly close to 4 lakhs and slightly above 10 lakhs

Bank_loan_services_Casestudy

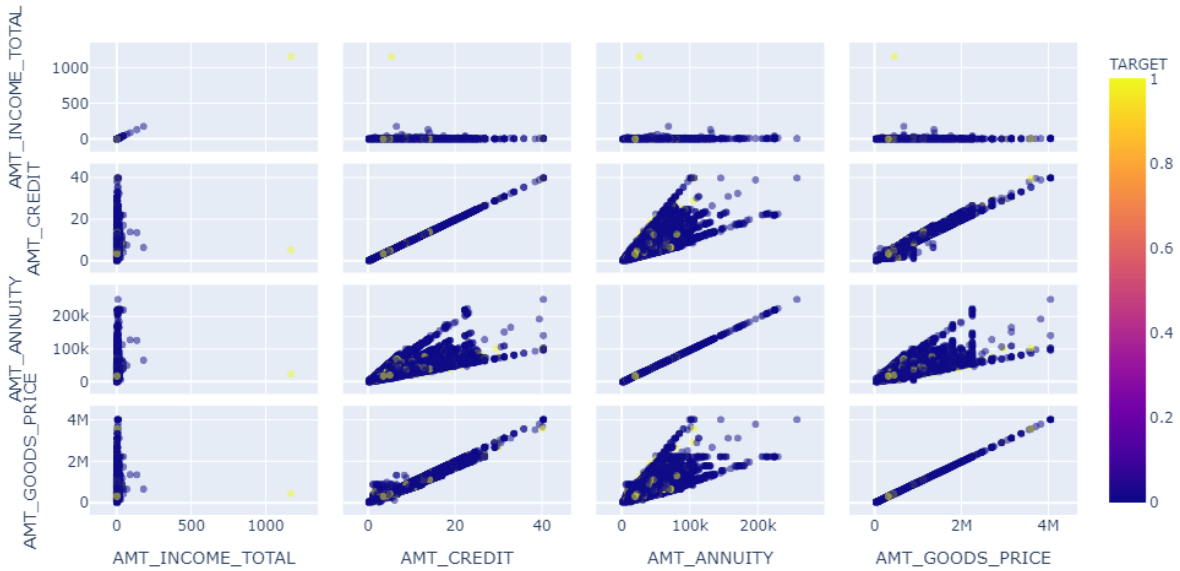
Numerical Variables Analysis

Numerical univariate Analysis



Inferences

- Credit score doesn't seems to have impact on defaulters. But density of repayers with more than 0.5 normalised credit score is high irrespective of loan amount

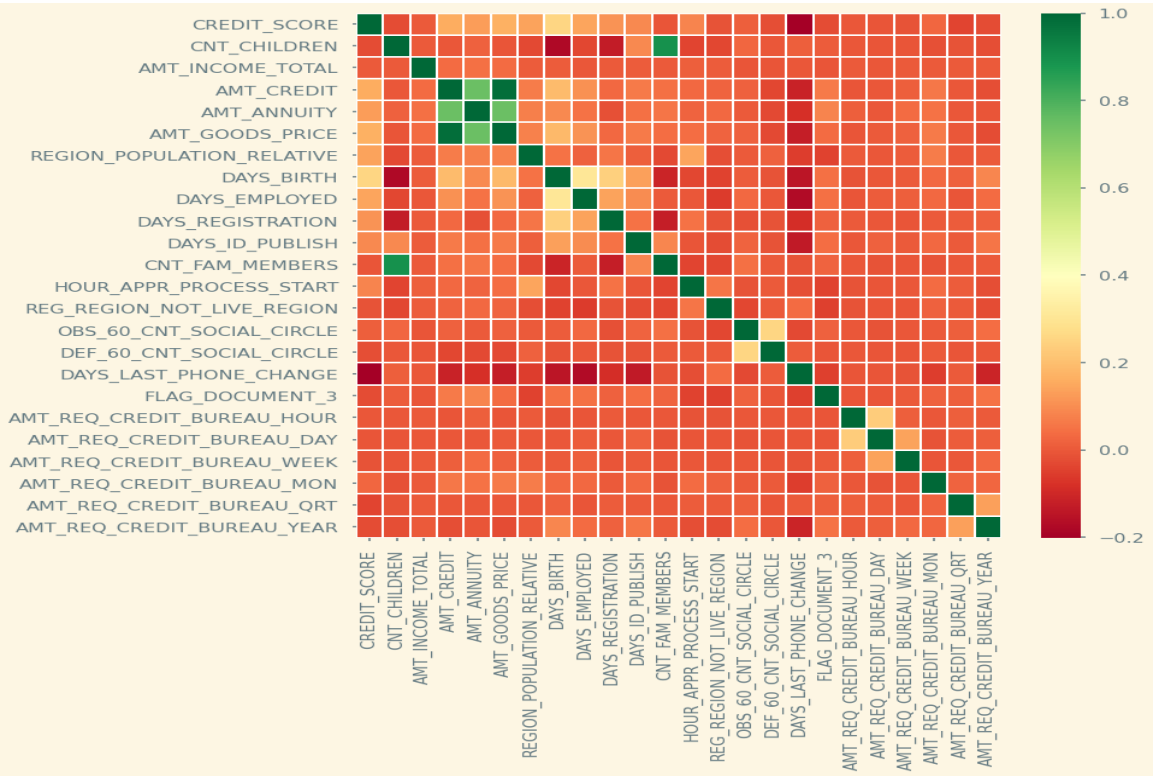


Inferences:

1. When amt_annuity >15000 amt_goods_price> 3M, there is a lesser chance of defaulters
2. AMT_CREDIT and AMT_GOODS_PRICE are highly correlated as based on the scatterplot where most of the data are consolidated in form of a line
3. There are very less defaulters for AMT_CREDIT >3M

Numerical Multivariate Analysis

Top ten correlation between variables of repayer



	VAR1	VAR2	Correlation
1	AMT_GOODS_PRICE	AMT_CREDIT	0.99
2	CNT_FAM_MEMBERS	CNT_CHILDREN	0.89
3	AMT_GOODS_PRICE	AMT_ANNUITY	0.77
4	AMT_ANNUITY	AMT_CREDIT	0.76
5	AMT_ANNUITY	AMT_INCOME_TOTAL	0.40
6	DAYS_EMPLOYED	DAYS_BIRTH	0.35
7	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.33
8	AMT_CREDIT	AMT_INCOME_TOTAL	0.33
9	DAYS_BIRTH	CREDIT_SCORE	0.30
10	DAYS_REGISTRATION	DAYS_BIRTH	0.30

Top ten correlation between variables of defaulters

	VAR1	VAR2	Correlation
1	AMT_GOODS_PRICE	AMT_CREDIT	0.98
2	CNT_FAM_MEMBERS	CNT_CHILDREN	0.89
3	AMT_GOODS_PRICE	AMT_ANNUITY	0.75
4	AMT_ANNUITY	AMT_CREDIT	0.75
5	DAYS_EMPLOYED	DAYS_BIRTH	0.31
6	DAYS_BIRTH	CREDIT_SCORE	0.26
7	DEF_60_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.26
8	DAYS_REGISTRATION	DAYS_BIRTH	0.24
9	AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_HOUR	0.23
10	DAYS_LAST_PHONE_CHANGE	CREDIT_SCORE	0.20

Inferences:

1. Credit amount is highly correlated with amount of goods price.
2. But the loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to repayers(0.77)
3. We can also see that repayers have high correlation in number of days employed(0.62) when compared to defaulters(0.58).
4. There is a severe drop in the correlation between total income of the client and the credit amount amongst defaulters whereas it is 0.33 among repayers.
5. Days_birth and credit score correlation has reduced to 0.26 in defaulters when compared to 0.30 in repayers.
6. There is a decrease in correlation between age and days registration : defaulters(0.24),repayers(0.30)

Merged Dataframe Analysis

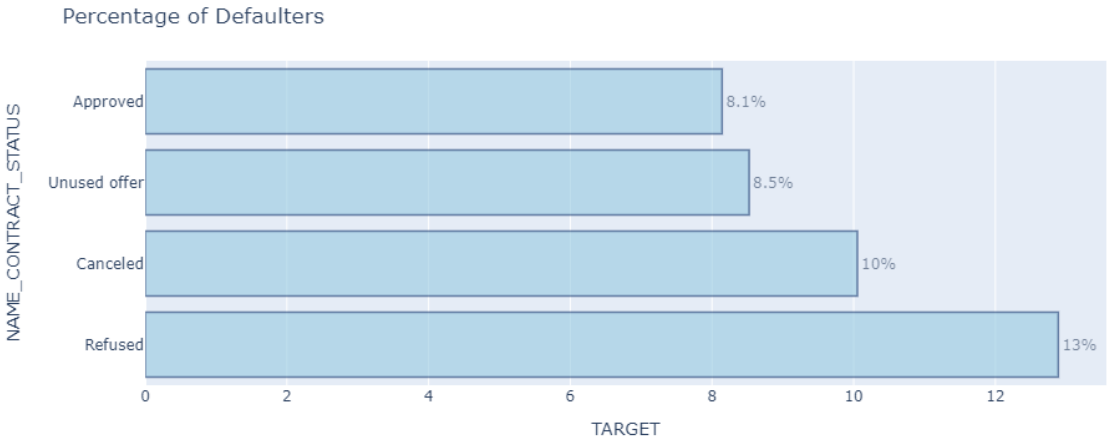
Strategy:

New data frame is created after having inner join on application_data and previous_data as we require all the columns from both dataframes with matching records on unique_id.

Shape of merged dataframe : (1140116, 76)

Size of merged dataframe : 86648816

Information about new dataframe : dtypes: category(33), float64(29), int64(14)

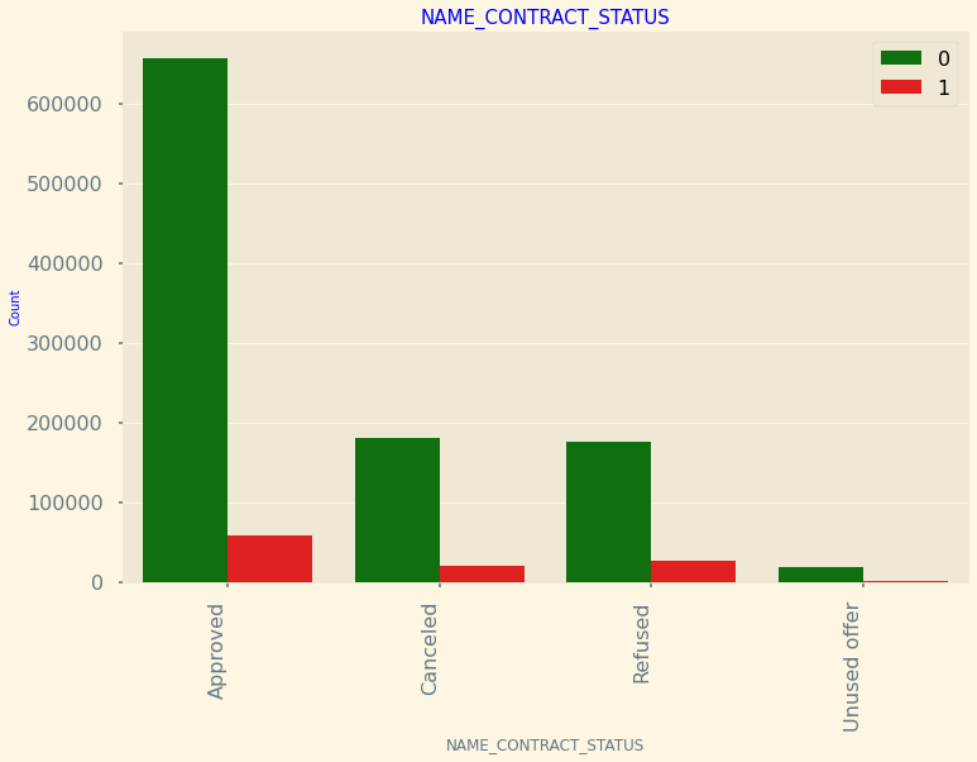


Inferences

1. Repeaters with large quantity of cash loans given and approved are defaulters. At the same time there is nearly equal amount of repeaters are refused and cancelled
2. Large number of new applicants approved with revolving loans are defaulters
3. 13% clients were previously refused are defaulters

Bank_loan_services_Casestudy

NAME_CONTRACT_STATUS	TARGET	Counts	Percentage
Approved	0	657632	91.86%
	1	58297	8.14%
Cancelled	0	181483	89.95%
	1	20282	10.05%
Refused	0	175597	87.11%
	1	25979	12.89%
Unused offer	0	19069	91.48%
	1	1777	8.52%



Inferences:

1. 90% of the previously cancelled client have actually repayed the loan. Revisiting the interest rates would increase business opoortunity for these clients
2. 88% of the clients who have been previously refused a loan has payed back the loan in current case.
3. Refusal reason should be recorded for further analysis as these clients would turn into potential repaying customer.

Conclusions

- 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 consised as below with the contributing factors and categorization:

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. CNT_CHILDREN: People with zero to two children tend to repay the loans.
9. CREDIT_SCORE : People with average to high credit score are less likely to default

Decisive Factor whether an applicant will be Defaulter:

1. CODE_GENDER: Men are at relatively higher default rate
2. NAME_FAMILY_STATUS : People who have civil marriage or who are single default a lot.
3. NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education
4. NAME_INCOME_TYPE: Clients who are either at Maternity leave OR Unemployed default a lot.
5. REGION_RATING_CLIENT: People who live in Rating 3 has highest defaults.
6. OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as the default rate is huge.
7. ORGANIZATION_TYPE: Organizations with highest percent of loans not repaid are Transport: type 3 (15.75%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self employed people have relative high defaulting rate (10.17%), and thus should be avoided to be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting.
8. DAYS_BIRTH: Avoid young people who are in age group of 20-40 as they have higher probability of defaulting
9. DAYS_EMPLOYED: People who have less than 5 years of employment have high default rate.
10. CNT_CHILDREN & CNT_FAM_MEMBERS: Client who have children equal to or more than 9 default 100% and hence their applications are to be rejected.
11. AMT_GOODS_PRICE: When the credit amount goes beyond 3M, there is an increase in defaulters.
12. EXT_SOURCE_1,_2,_3: People with low credit score must be avoided .