CREDIT EDA CASE STUDY

NIKUNJ PATEL

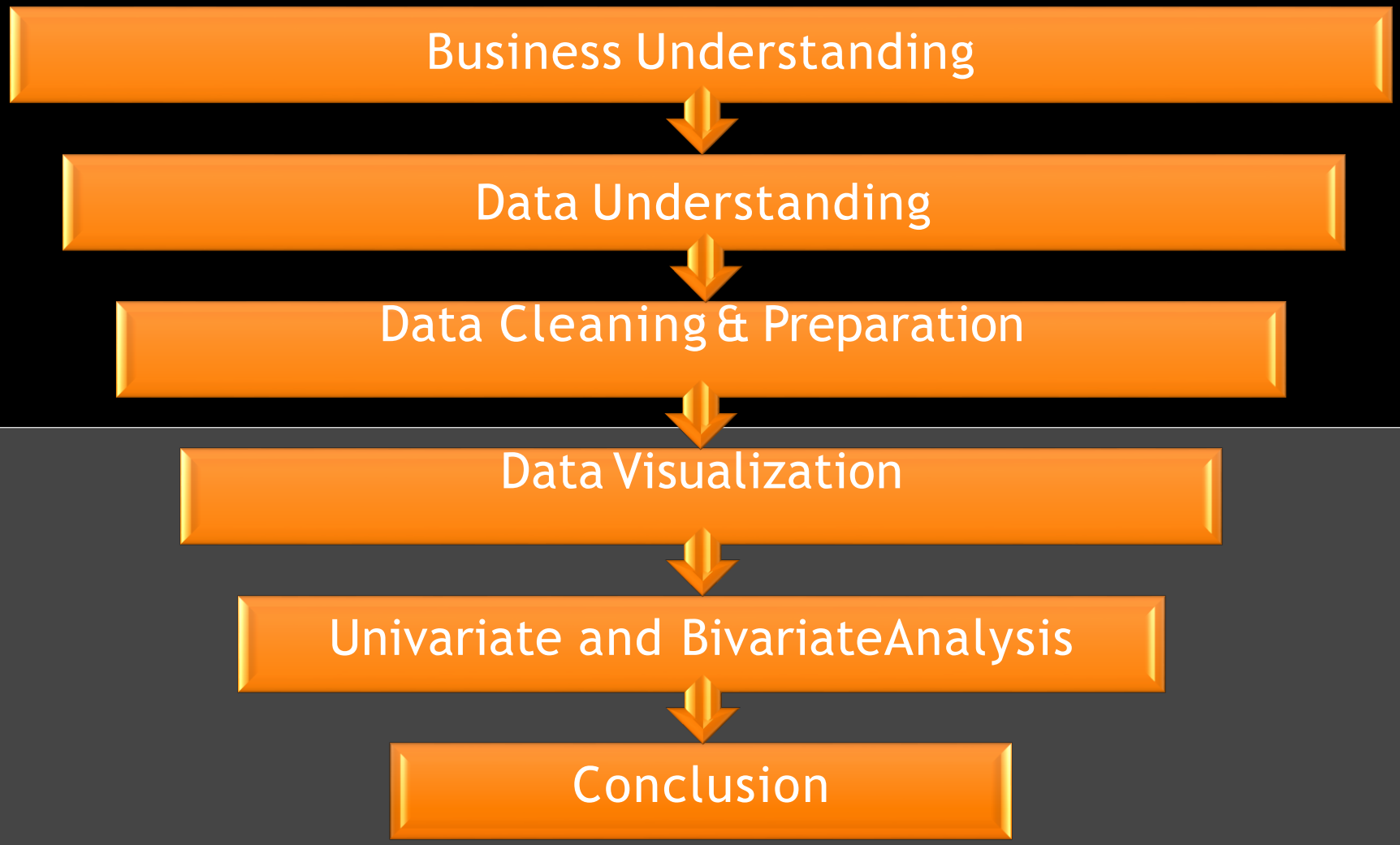
LAXMAN RAO

ABSTRACT

The aim of Loan Case Study is to 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 using EDA.

The data given contains the information about past loan applicants and whether they 'defaulted' or not. The aim is to identify patterns which indicate if a person is likely to default, 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.

PROBLEM SOLVING METHODOLOGY



BUSINESS UNDERSTANDING

A Consumer finance company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

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 contains the information about previous loan applicants and application whether they 'defaulted' or not. The aim is to identify patterns which indicate if a person is likely to default, 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.

When a person applies for a loan, there are two types of decisions that could be taken by the company:

Loan Status	Description
Loan Accepted (Fully Paid)	Applicant has fully paid the loan (the principal and the interest rate)
Loan Accepted (Current)	Applicant is in the process of paying the instalments.
Loan Accepted (Charged-Off)	Applicant has not paid the instalments in due time for a long period of time
Loan Rejected	The company rejected loan (as candidate didn't meet their requirements)

DATA UNDERSTANDING

Pre Loan Attributes

SK_ID_PREV
CONTRACT_TYPE
AMT_ANNUITY
AMT_APPLICATION
AMT_CREDIT
AMT_DOWN_PAYMENT
AMT_GOODS_PRICE
RATE_INTEREST
CASH_LOAN_PURPOSE
CNT_PAYMENT

Post Loan Attributes

SK_ID_CURR
TARGET
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
Inq _Before_loan
HOUR_APPR_PROCESS _START
DAYS_REGISTRATION
NAME_INCOME_TYPE
RATE_INTEREST_PRIVILEGED

Applicant Attributes

FAMILY_STATUS
CNT_CHILDREN
FLAG_DOCUMENT
INCOME_TYPE
EDUCATION_TYPE
OCCUPATION_TYPE

Background Check Attributes

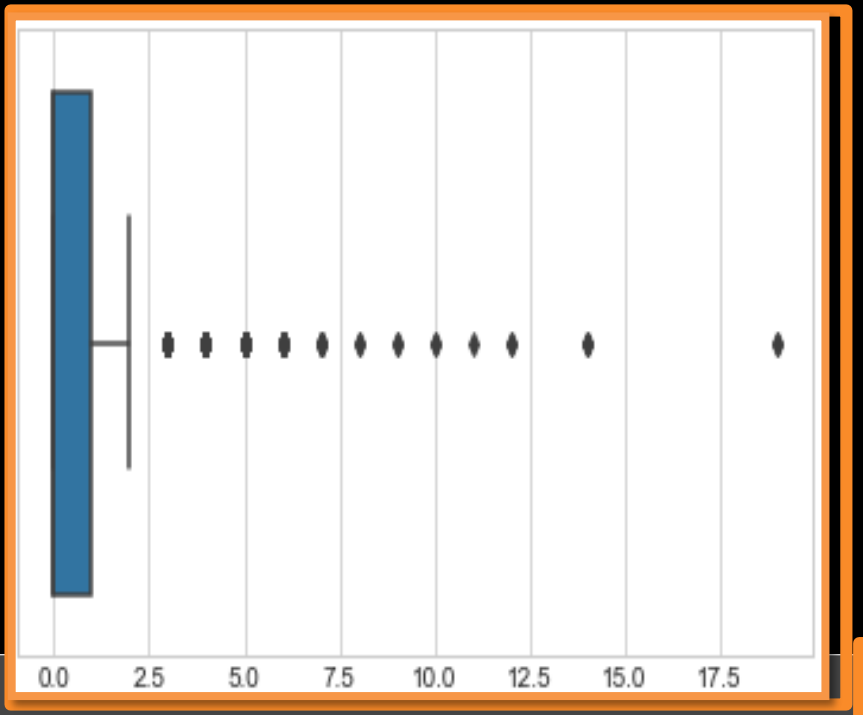
NAME_CLIENT_TYPE
NAME_PORTFOLIO
AMT_CREDIT
NAME_COTRACT _STATUS
DAYS_DECISION
NAME_PAYMENT _TyPE
CODE_REJECT_REASON
DAYS_TERMINATION
DAYS_FIRST_DUE
DAYS_LAST_DUE

We have categorised all the attributes as Pre loan , Curr Loan, Applicant and Background check to better understand the data.

DATA CLEANING & PREPARATION

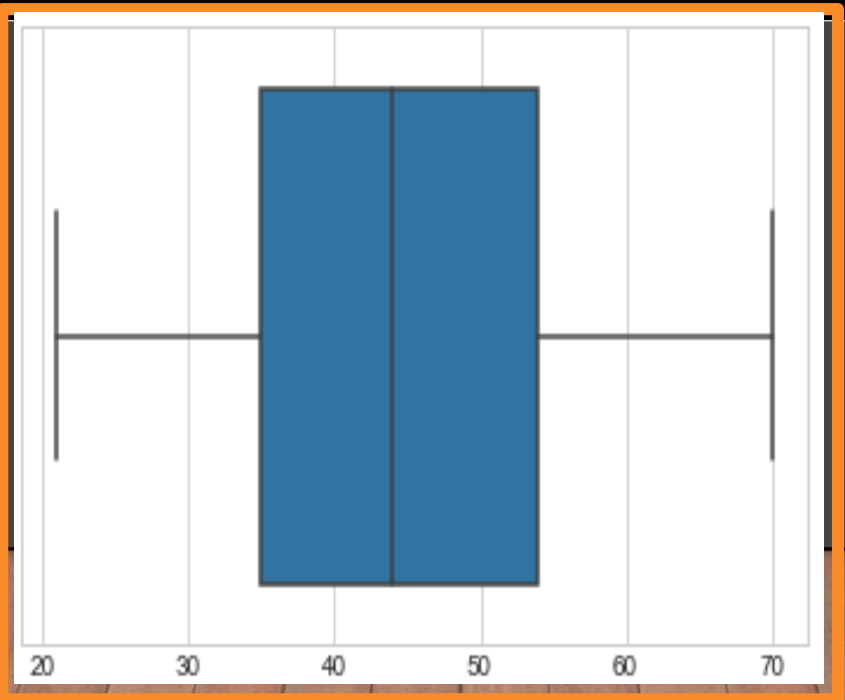
- Identification of missing values and duplicate in data set.
- Treatment of missing values by dropping columns having missing values more than 50% & 30% from other file,
- Replace some categorical column with Top Frequency, We can also Treat missing value with Mean or Median or zero Also. There is a XNA value in some column so we leave that value as it is and considered as categorical data ‘Non- Availability’.
- Dropping off few irrelevant columns (which contain extra information related to loan) for analysis.
- Cleaning of special characters in rows of specific.
- As aim is to identify patterns which indicate if a person is likely to default so we would drop the values for loans which are current as the applicant is in the process of paying the instalments.
- Creating the new columns and bucketing the values (creating slots) for easy understanding of data for analysis.

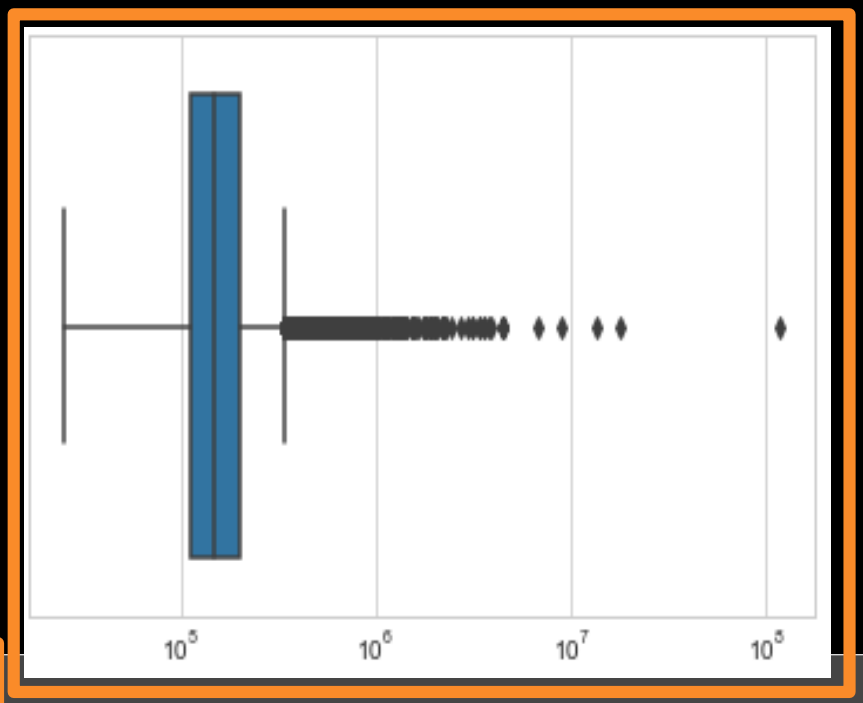
OUTLIER ANALYSIS



It can be seen that Number of children the client has outlirs in CNT_CHILDREN Columns

It can be seen Client's age column has no Outlier in Dataset

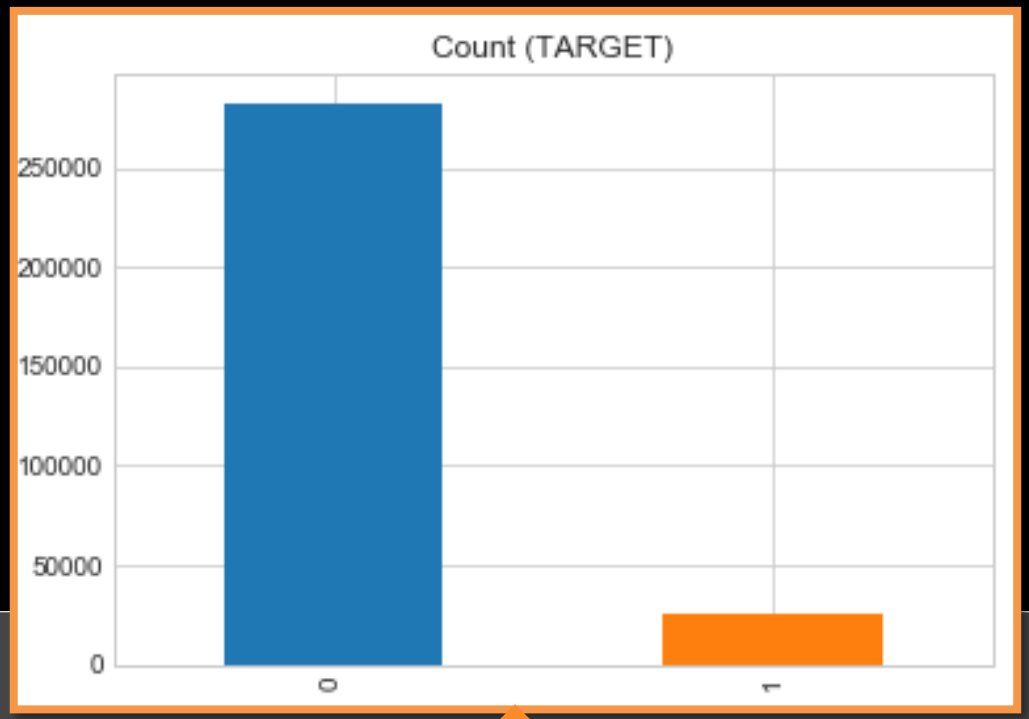




It can be seen that Income of the Client has Outlier in AMT_INCOME_TOTAL column

DATA IMBALANCE

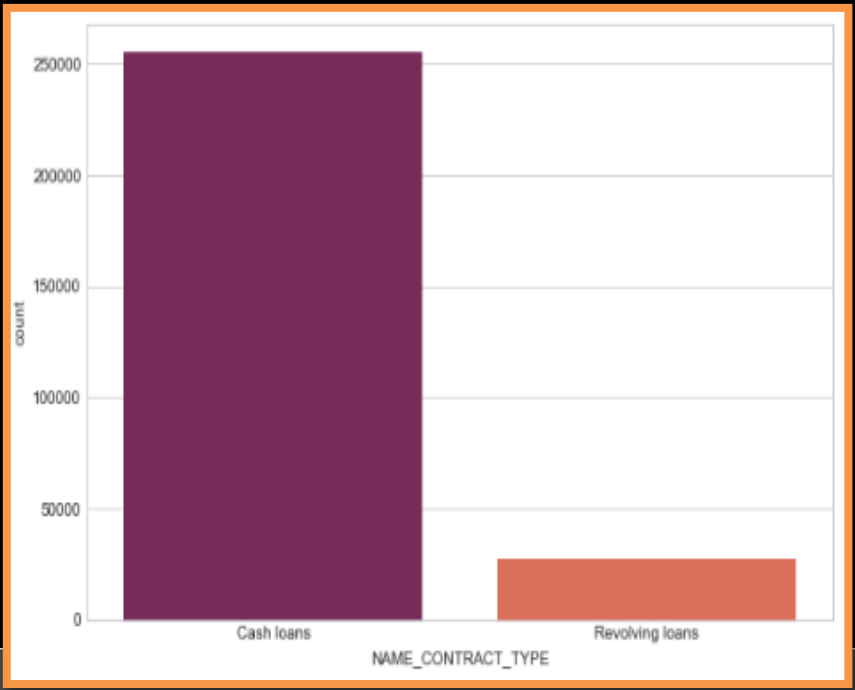
1) It can be seen that TARGET Columns has the Data Imbalance.



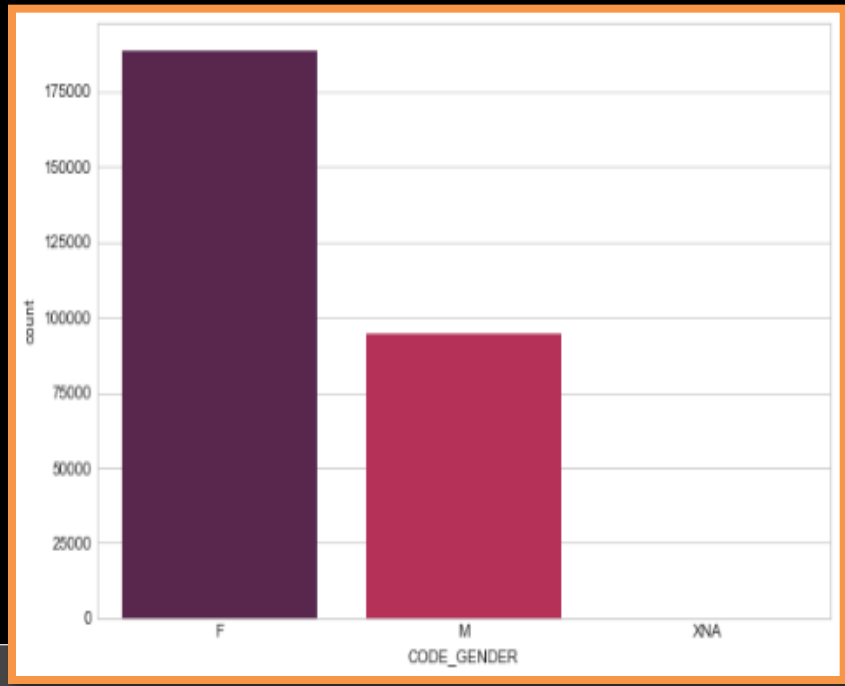
2) Total Ratio of data imbalance In TARGET column is : 11.39 : 1

3) Percentage of no difficulty faced is : 91.92
percentage of payment difficulties is : 8.07

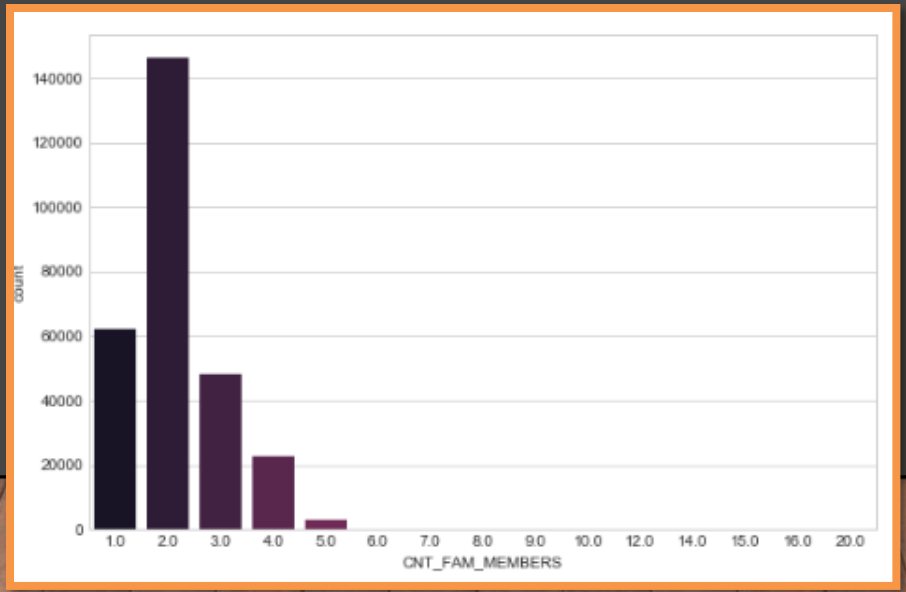
UNIVARIATE ANALYSIS



It can be seen that client Who has 2 family members they apply highest no for loan

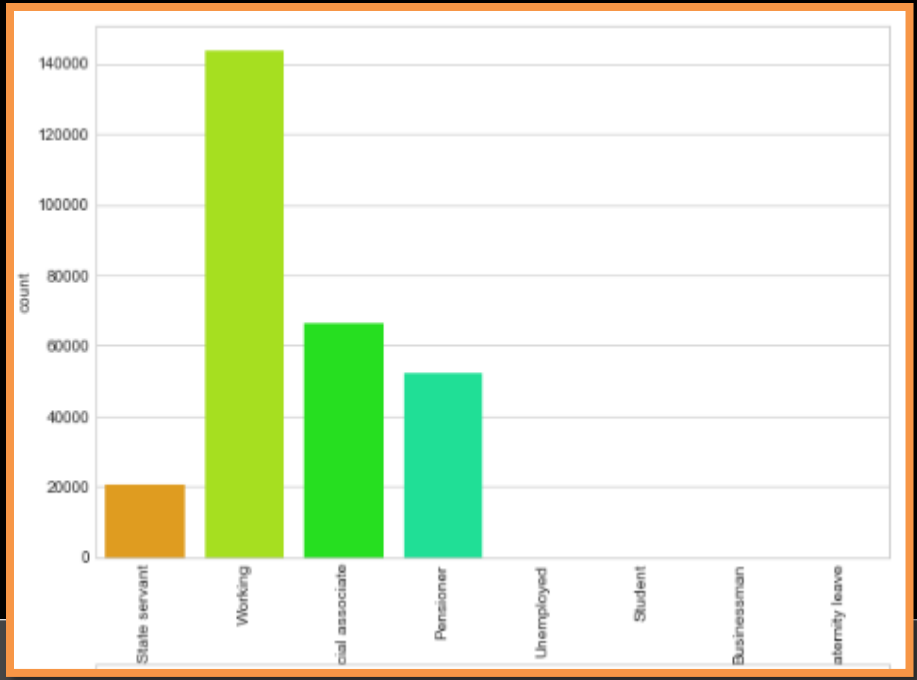


It can be seen that applicant majorly apply for Cash loan i.e : 2,52,000 approx. applicant

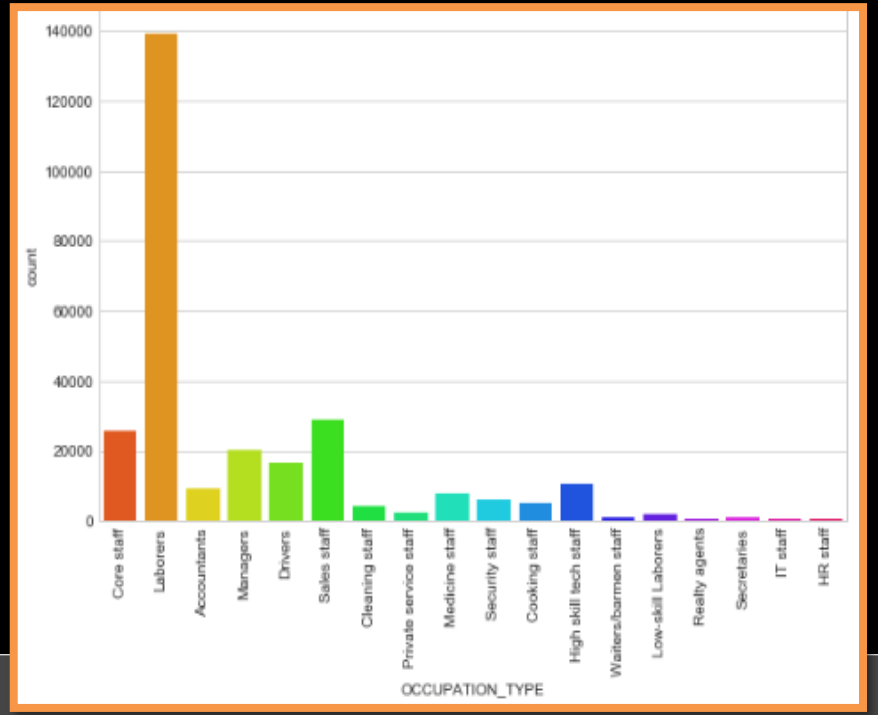


It can be seen that major applicant is Female who taking/apply loan

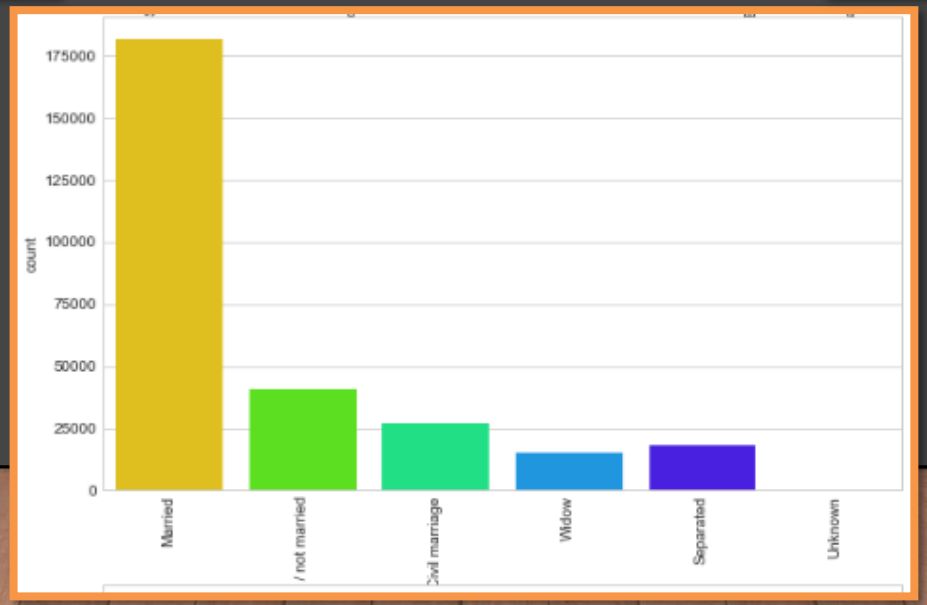
UNIVARIATE ANALYSIS



It can be seen that client Who martial status is married they apply more for loan compare to other status

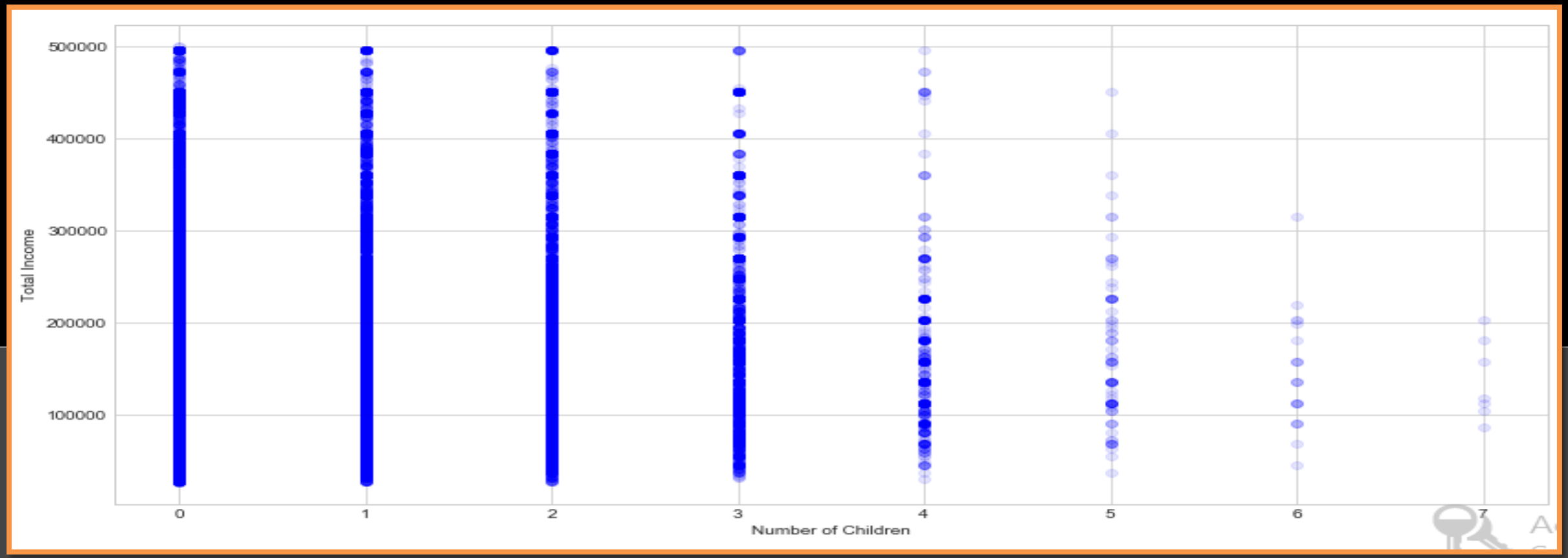


It can be seen that applicant Whose income type has been working they apply highest for loan



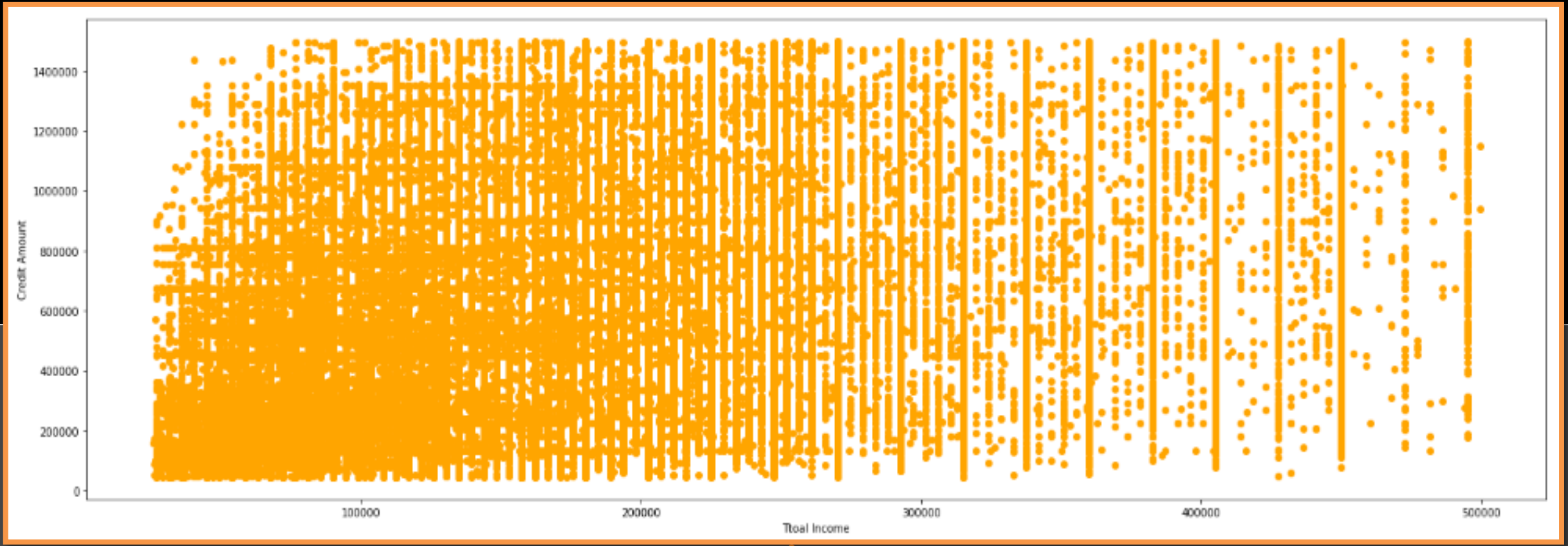
It can be seen that whose occupation type is Laours they apply highest number for loan

BIVARIATE ANALYSIS ON COUNT_CHILDREN & TOTAL_INCOME



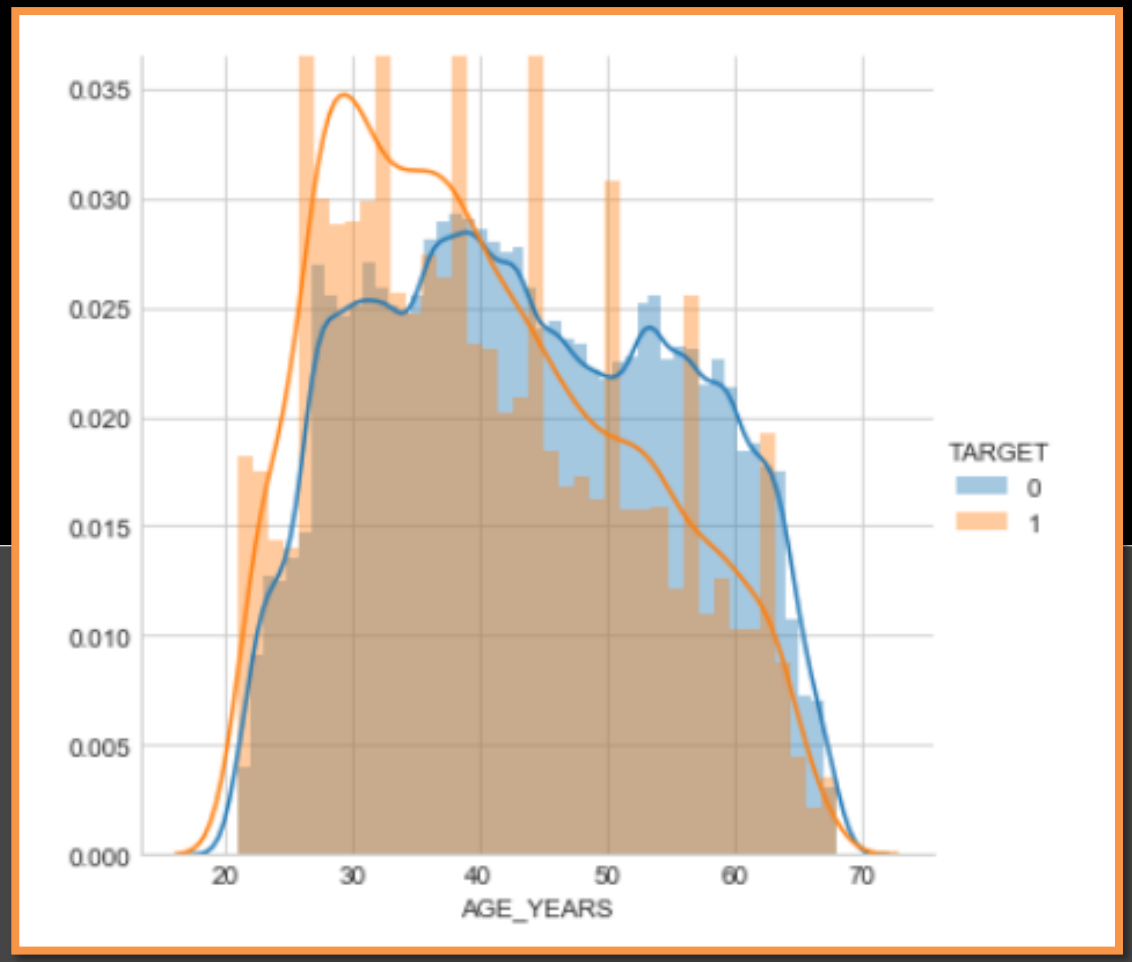
Bivariate analysis is used to find out if there is a relationship between two sets of values.
It can clearly seen that total income and number of children relationship for loan purpose. Who has more than 3 children they show more interest for loan.

BIVARIATE ANALYSIS ON TOTAL_INCOME & AMOUNT_CREDIT



We can seen that whose total income is approx. above 10 lakh they apply more for total credit amount loan and they apply around 20lakh amount of loan.

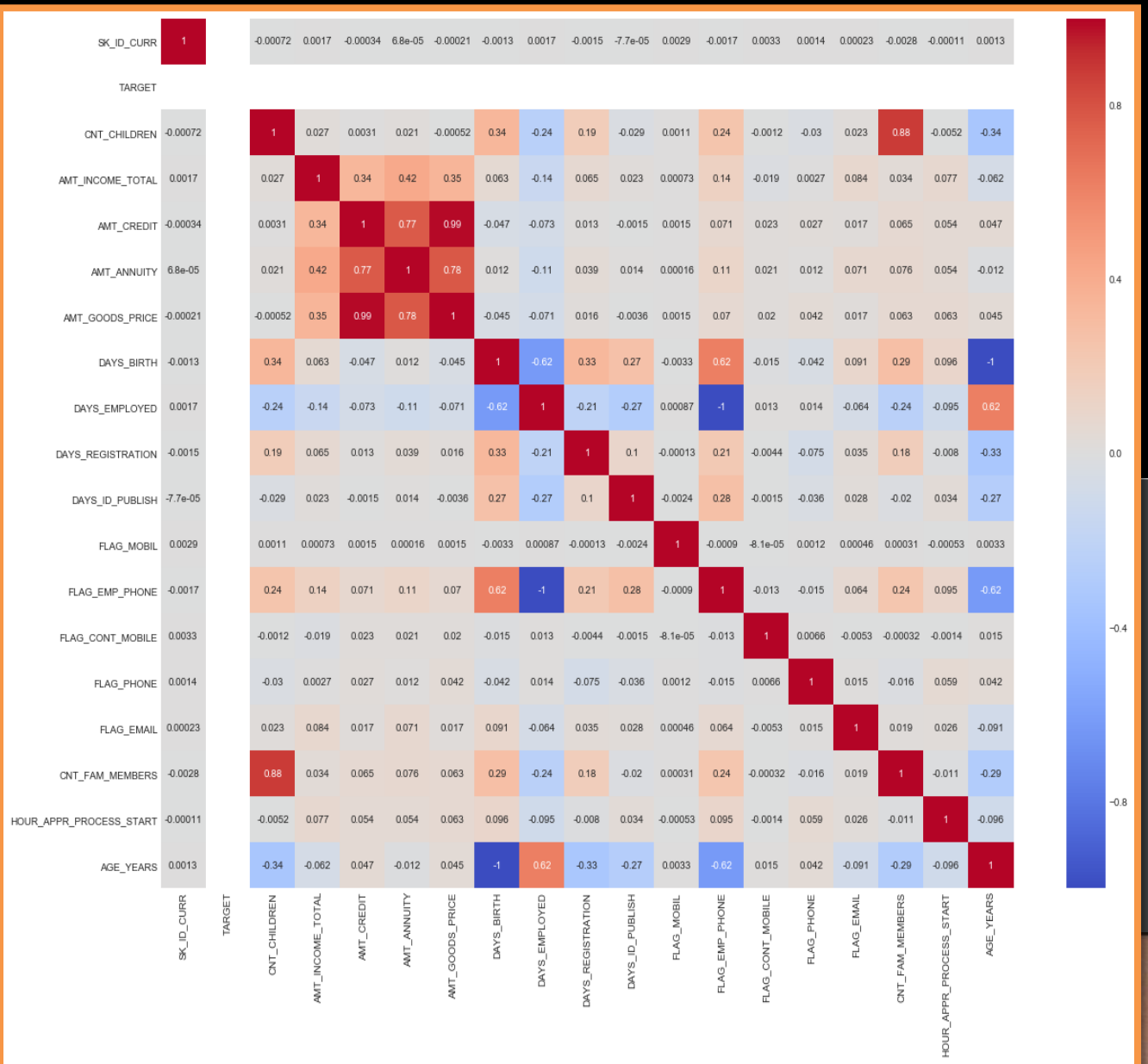
Distribution of loan applicants is as shown below.



The Figure Shows the distribution of Applicants loan Applicants Base on age.

As we can see in the figure the No. of applicants is more between the age group of 20 and 40. in both defaulters and non defaulters

Correlation between various attributes of loan applicants of non payments Defaulters

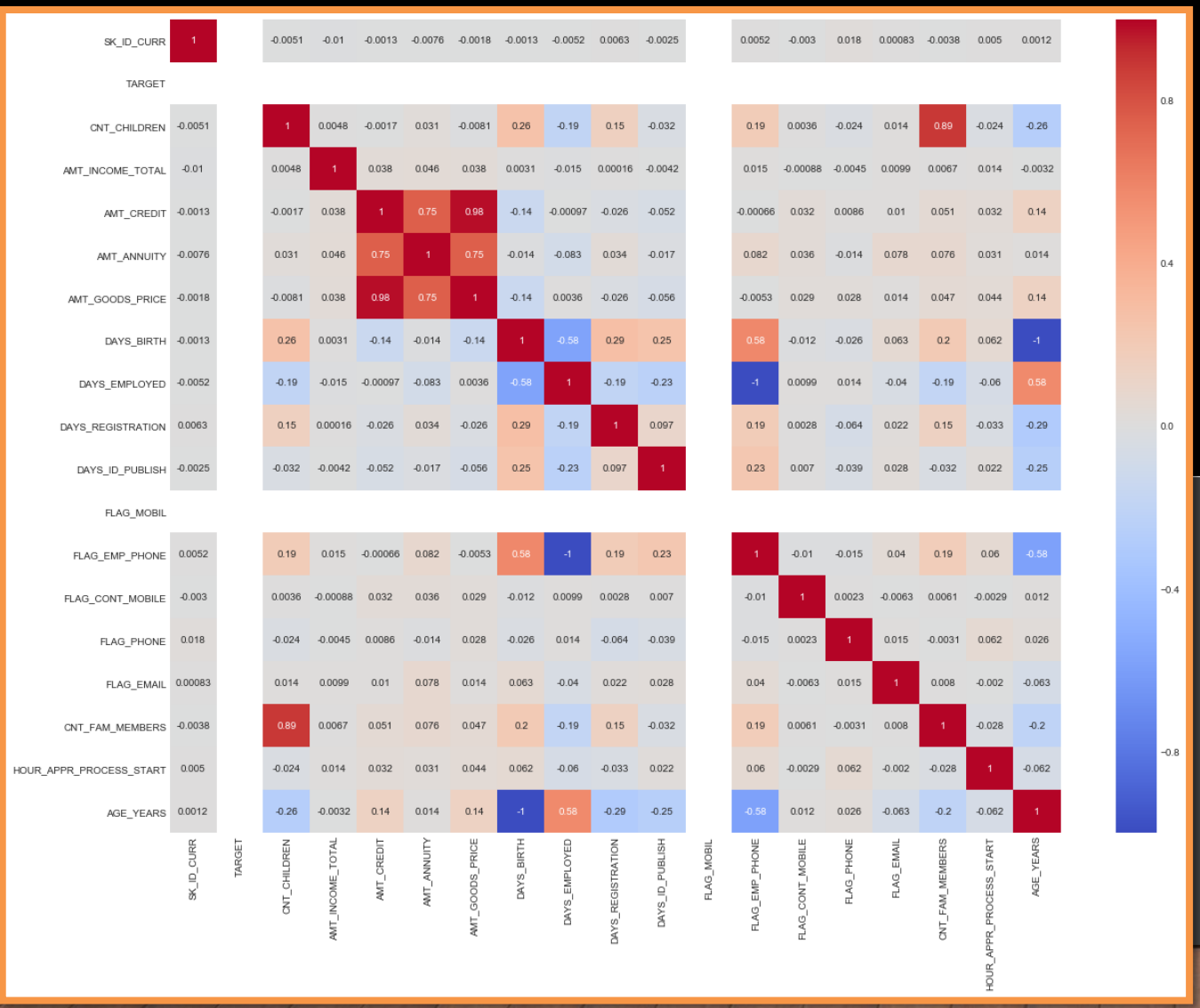


The figure Shown here is the heat map of Correlation matrix of Various Attributes in payments defaulters data.

1) AMT_credit
2) Amount_goods_price
3) Amount_Annnuity
4) CNT_FAM_Members
5) Days of Birth

The Above mentioned some of the attribute having more correlation together in non payment defaulters

Correlation between various attributes of loan applicants of payments Defaulters



The figure Shown here is the heat map of correlation matrix of Various Attributes in non payments defaulters data.

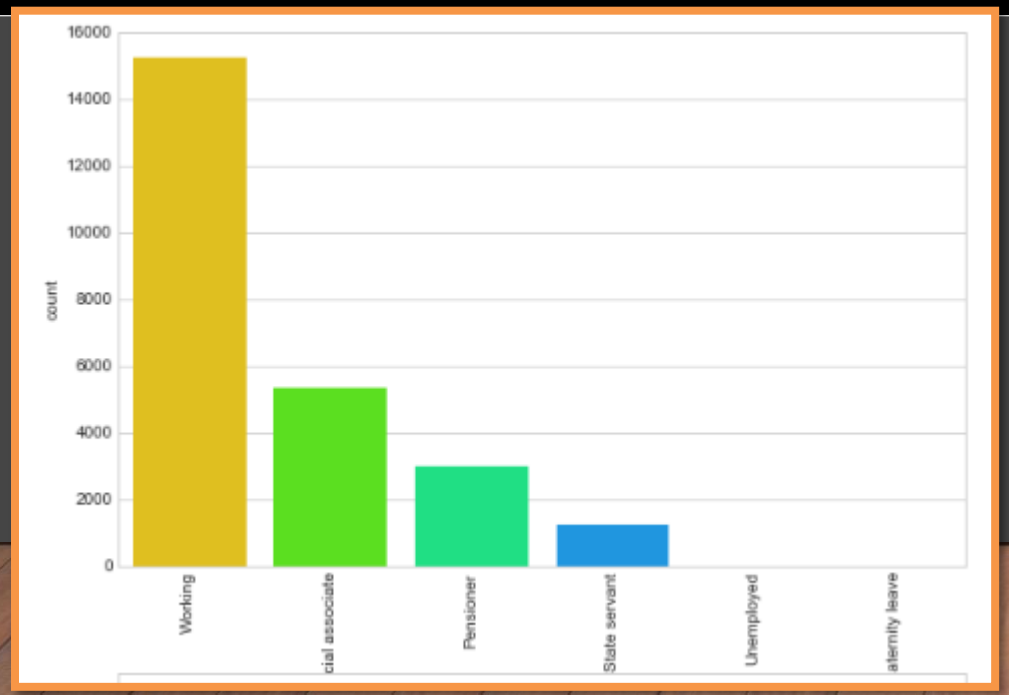
1) AMT_credit
2) Amount_goods_price
3) Amount_Annnuity
4) CNT_FAM_Members
5) Days of Birth

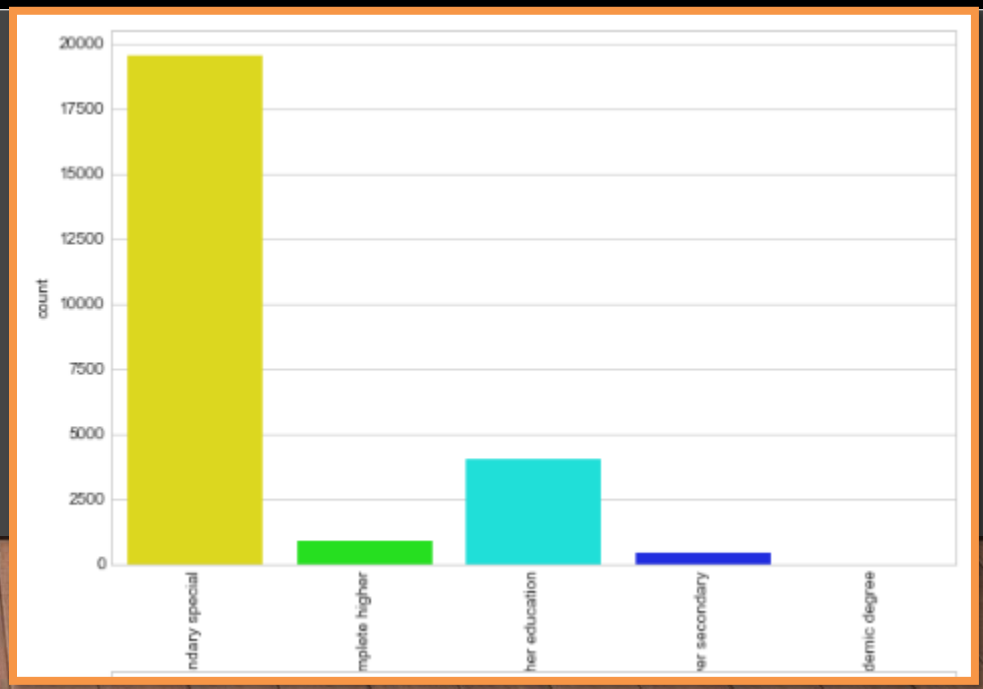
Correlation between the attributes is same in the case of payments defaulters .

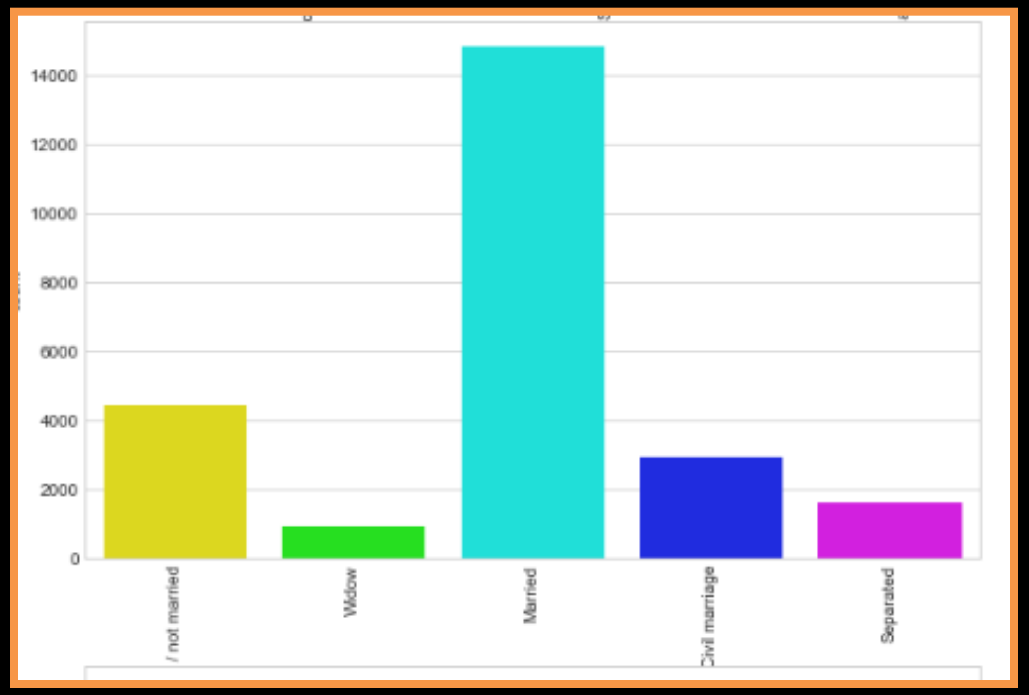
Observations of the Given Data (Applicantion_data):

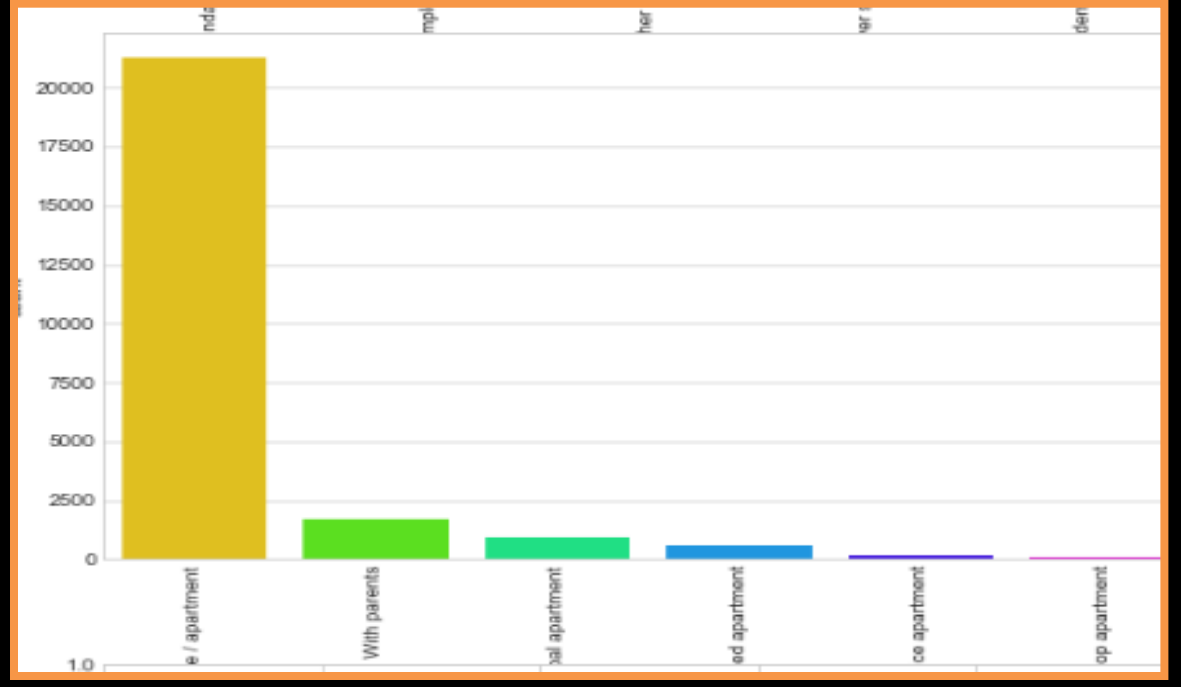
1) The No.of applications for the cash loan is 8 Times more than the revolving loans in case of non payment defaulters where as its is 15 Times in payment defaulter
2) The No.of applications from females is 2 Times more than the male applicants in case of non payment defaulters where as its is 1.3 Times in payment defaulter.
3) The No.of applications having car is 2 Times more than the applicants not having car in case of non payment defaul where as its is 2.3 Times in payment defaulter.
4) The No.of applications having realty is 2.3 Times more than the applicants not having realty in case of non payment defaulters and payment defaulters

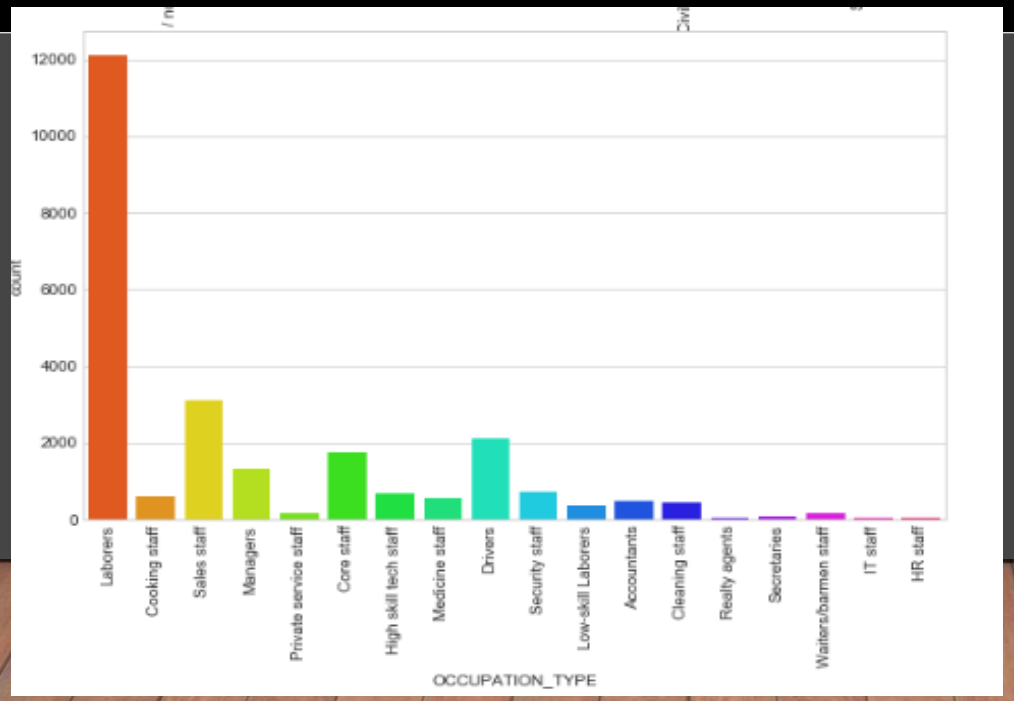
→ The Various observation made through the analysis can be seen in the below attached figures






14000
12000
10000
8000
6000
4000
2000
0
/ not married
Widow
Married
Civil marriage
Separated


20000
17500
15000
12500
10000
7500
5000
2500
0
1.0
With parents

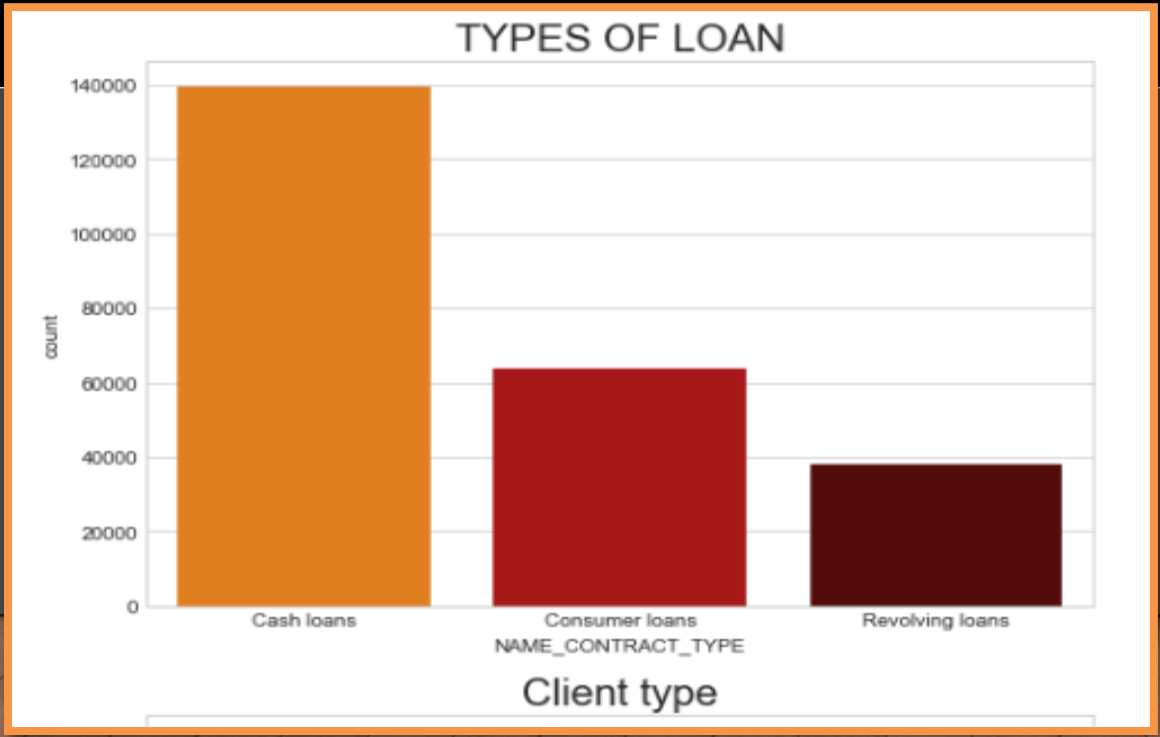

12000
10000
8000
6000
4000
2000
0
Laborers
Cooking staff
Sales staff
Managers
Private service staff
Core staff
High skill tech staff
Medicine staff
Drivers
Security staff
Low-skill Laborers
Accountants
Cleaning staff
Realty agents
Secretaries
Waitors/barmen staff
IT staff
HR staff
OCCUPATION_TYPE

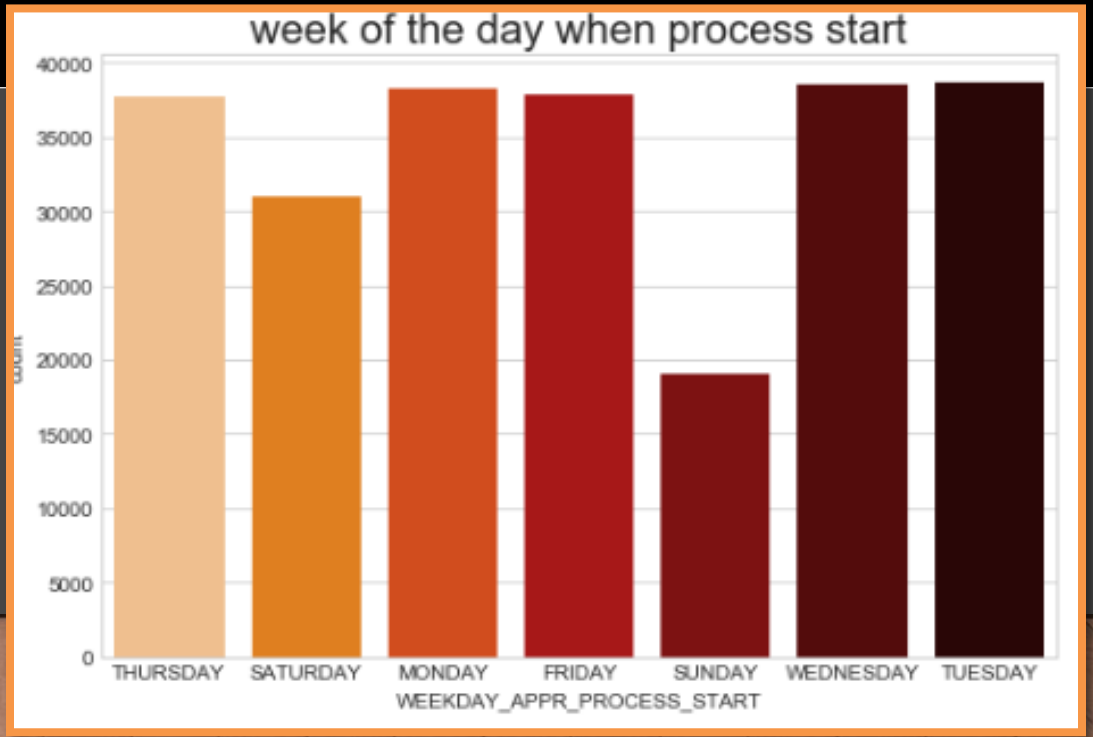
Analysis on previous year Application Data

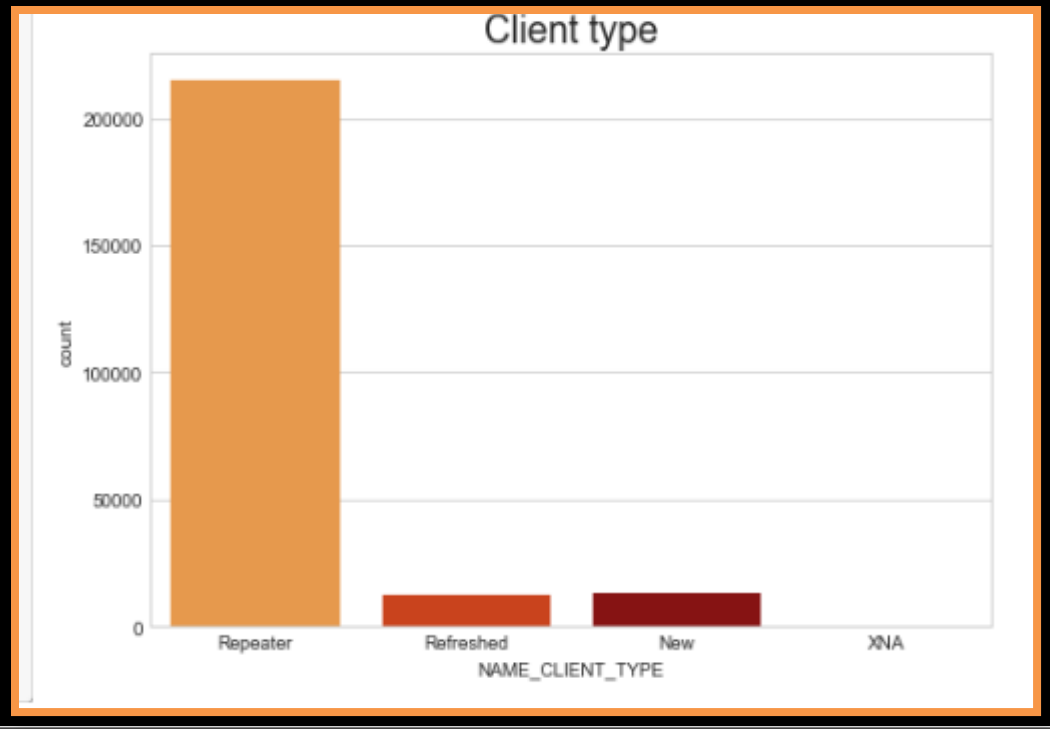
The Analysis was carried out based on the below mentioned attributes of given data :

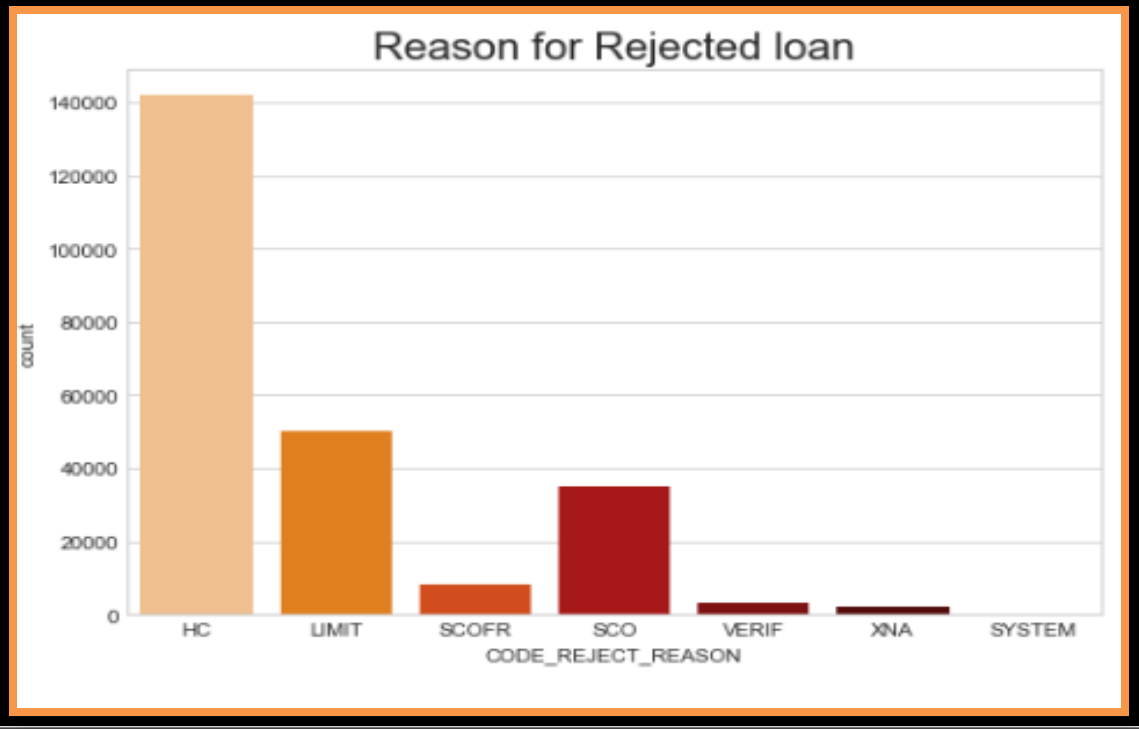
1) Client type
2) Reasons for rejection of Loan
3) The Week day of application process
4) Type of Loans applied
5) Status of Loan application

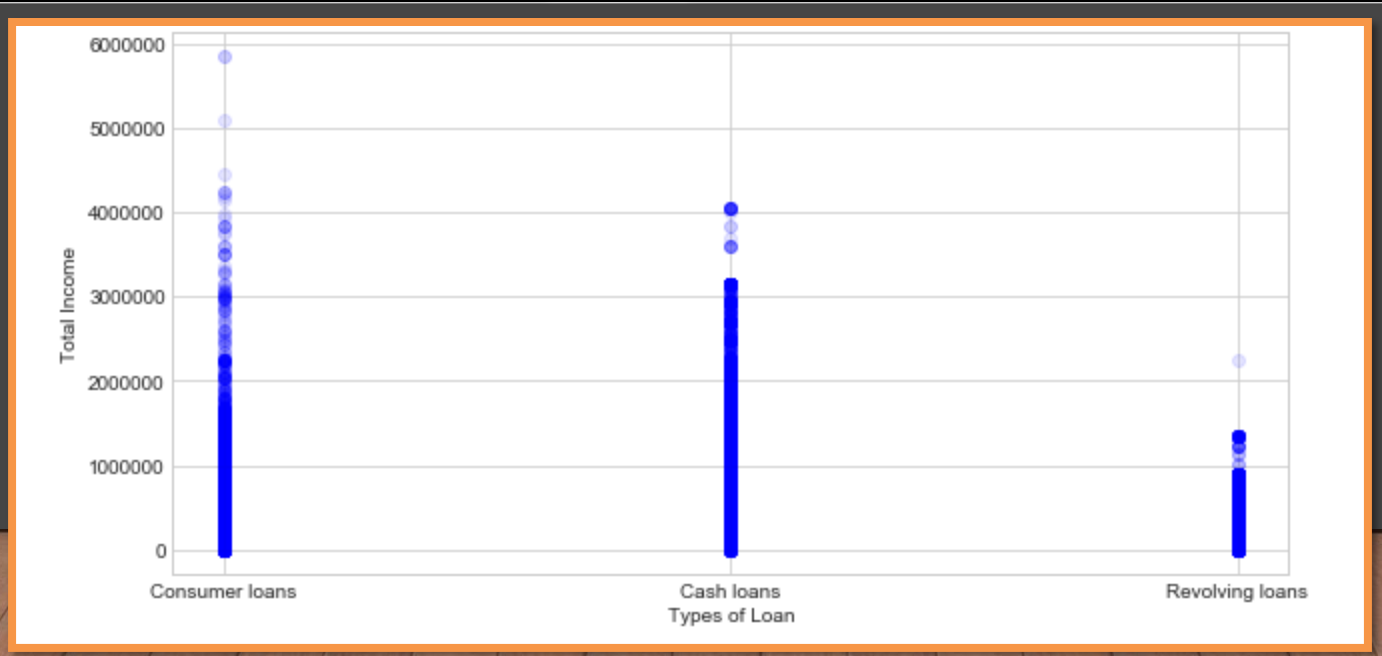
→ The Following plots were drawn to see the distribution pattern that the attributes follow when the application was approved



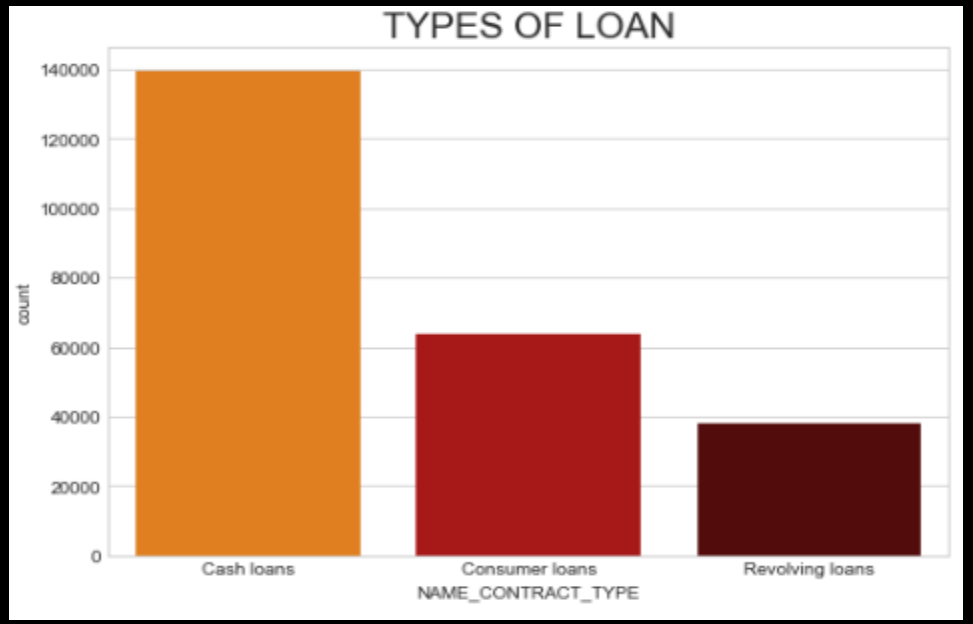


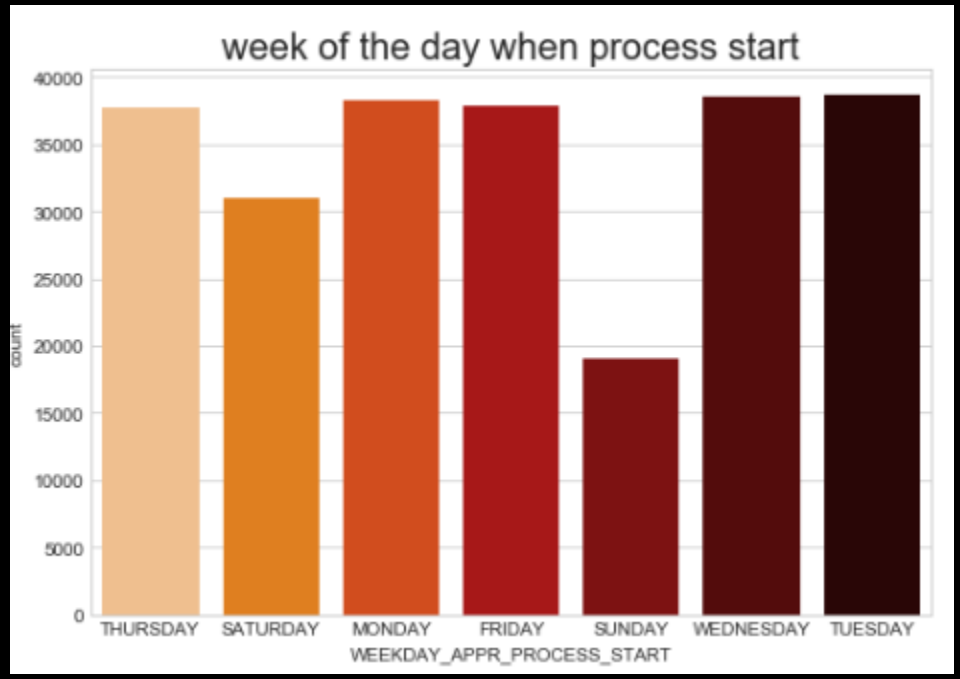

Client type
count
200000
150000
100000
50000
0
Repeater
Refreshed
New
XNA
NAME_CLIENT_TYPE

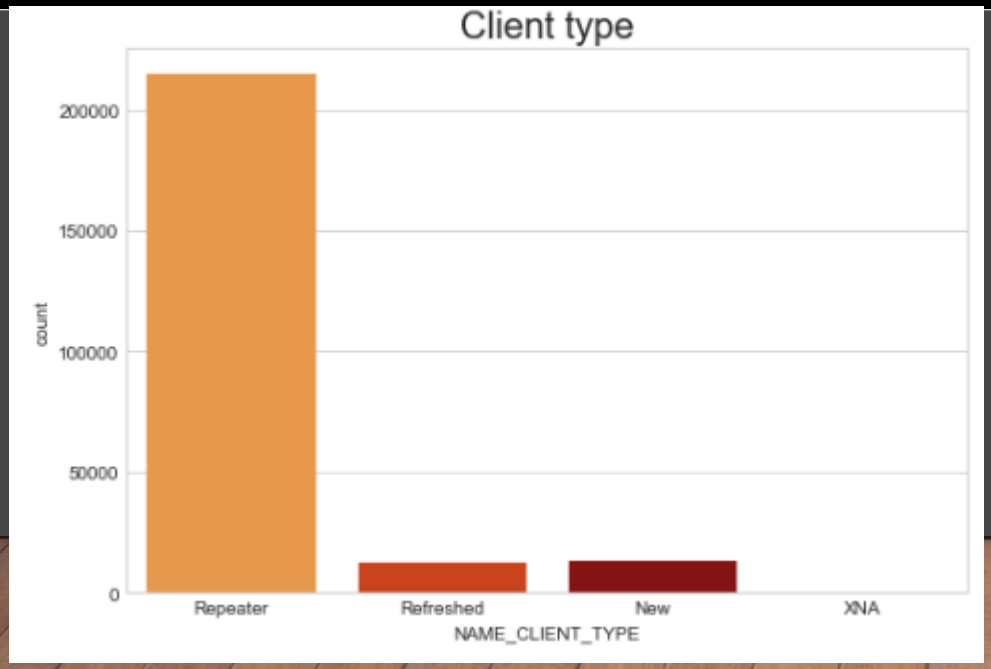

Reason for Rejected loan
count
140000
120000
100000
80000
60000
40000
20000
0
HC
LIMIT
SCOFR
SCO
VERIF
XNA
SYSTEM
CODE_REJECT_REASON

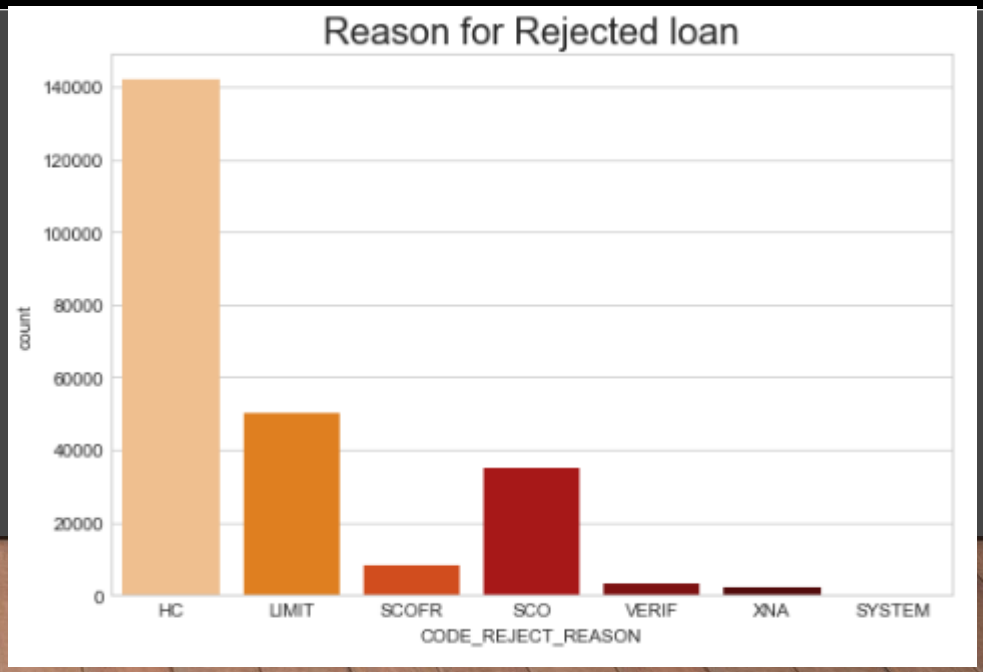

Total Income
6000000
5000000
4000000
3000000
2000000
1000000
0
Consumer loans
Cash loans
Revolving loans
Types of Loan

→ The Following plots were drawn to see the distribution pattern that the attributes follow when the application was Rejected

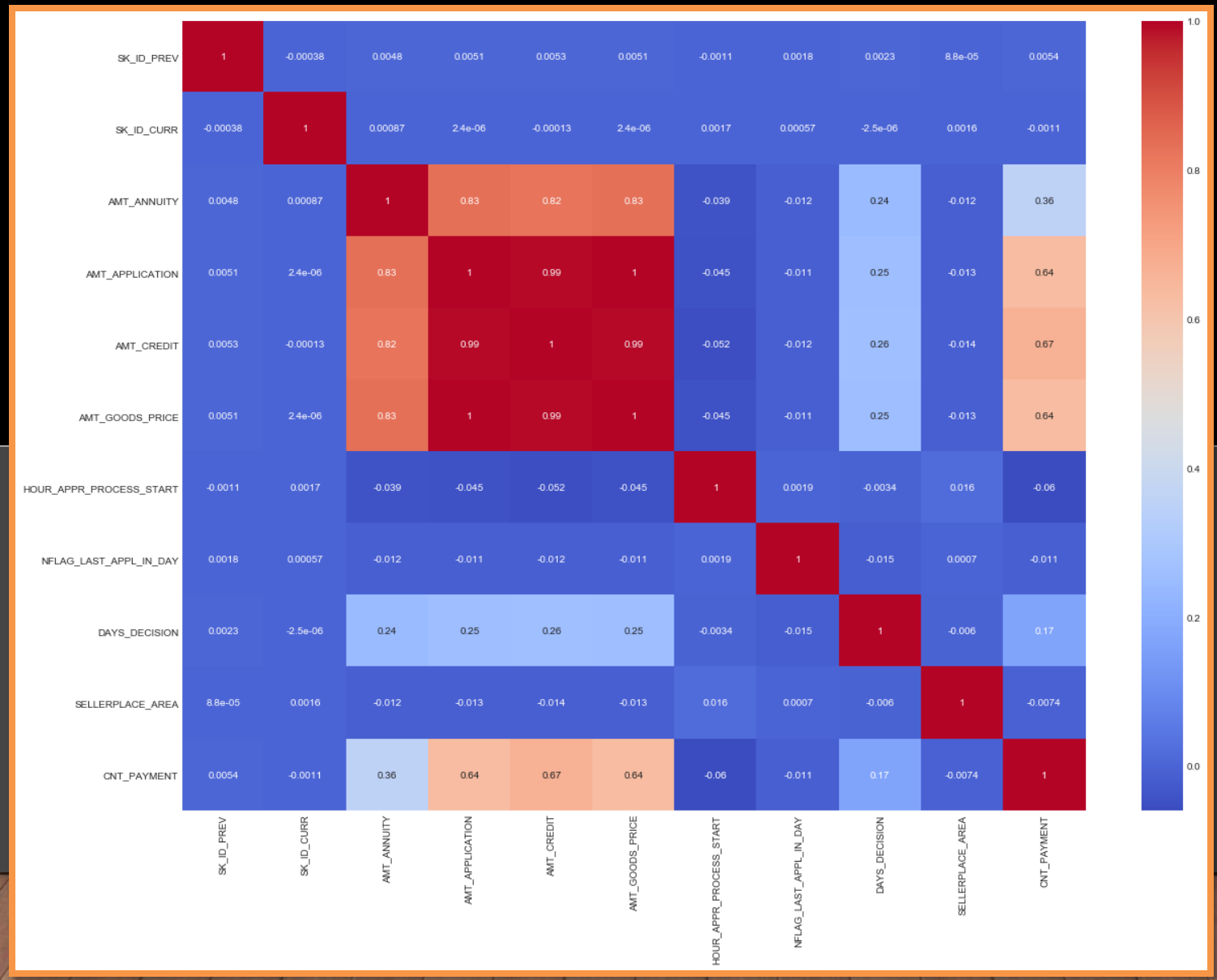




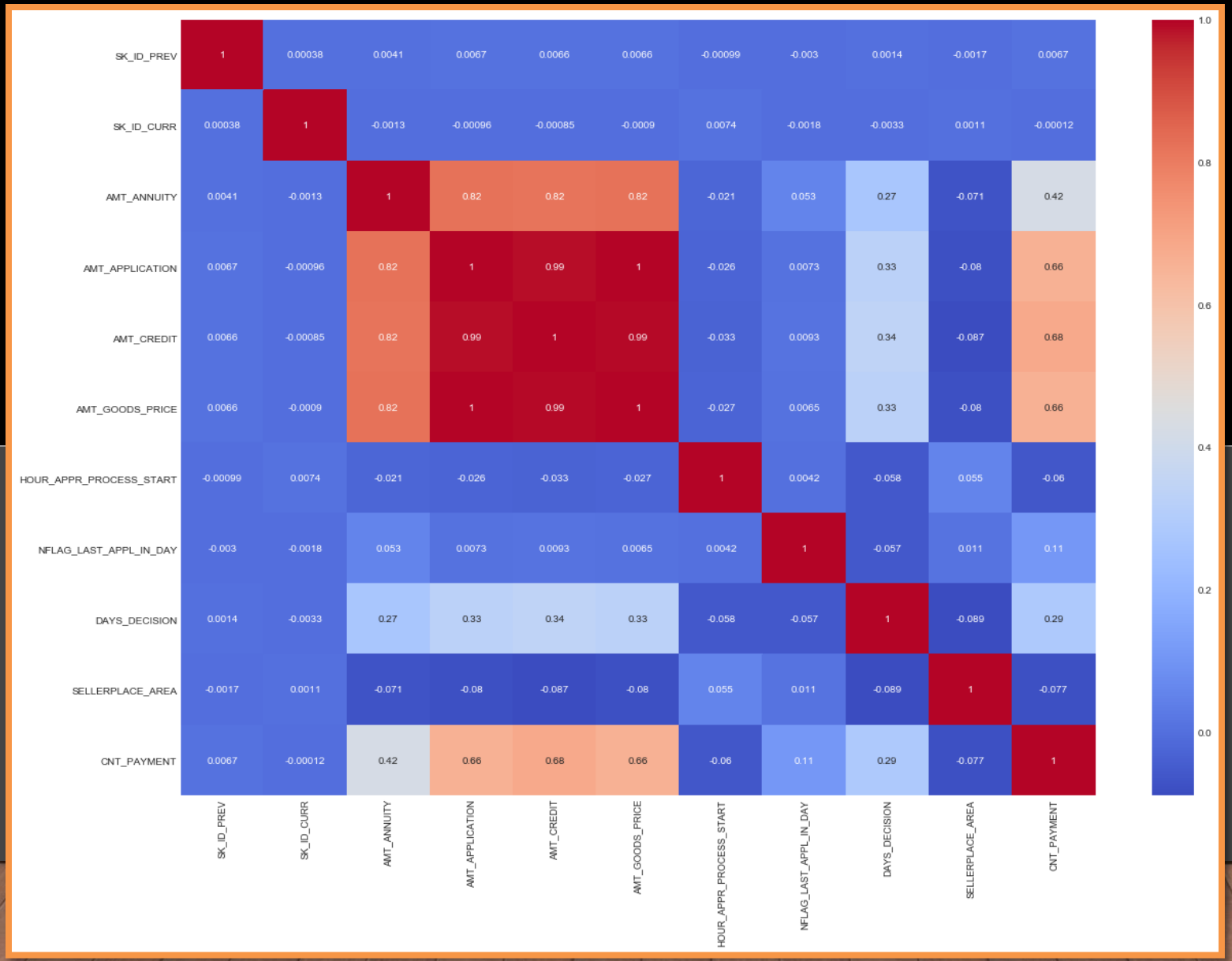




Correlation Matrix of attributes when the loan is approved in the Previous year Application Data is As show in below figure



Correlation Matrix of attributes when the loan is Rejected in the Previous year Application Data is As show in below figure



As per our Understanding and Knowledge, Based on insights drawn out of the data , The following recommendation can be made to the Loan Provider

1) As It can be clearly seen through the Data Imbalance between Target Variable 0 and 1, and the Ratio for No. of Loan non payment defaulters is 11 Times to the payment defaulter. that implies that Loan provider should be more lean towards accepting the loan application

2) Loan Providers shall be more careful when they providing loans to the Married working people , as they tend to be the most in payment defaulters.

3) In Biviariate graph there is a positive result we can get

4) In Housing type column it ll show Defaulter.

CONCLUSION

Important Variables

- CNT_Family_members
- FAMILY_STATUS
- CNT_CHILDREN
- FLAG_DOCUMENT
- INCOME_TYPE
- EDUCATION_TYPE
- Toatl_Income
- OCCUPATION_TYPE
- Applicants Gender
- Applicants Age
- Applicants Occupation

- ✓ Most of the loan applications are for Car loans which is around 272000
- ✓ Females applicants for the Loan Are more than Male applicants
- ✓ Working people are in seeking loan among the all the professions
- ✓ In the univariate analysis we see applicant Whose income type has been working they apply highest for loan
- ✓ It can be seen that client Who martial status is married they apply more for loan compare to other status
- ✓ To,Approved the Loan it is a most important variable to decide loan is approved or not

THANK YOU