CREDIT EDA CASE STUDY

BY
PRANJIL MARTIN

INTRODUCTION

This case study aims to give an idea of applying EDA in a real business scenario. In this case study, we will 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 to 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 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.

INTRODUCTION

BUSINESS UNDERSTANDING:

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:
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,
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 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.

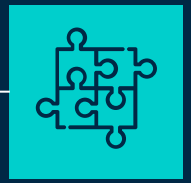
INTRODUCTION

BUSINESS OBJECTIVE :

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

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.
To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

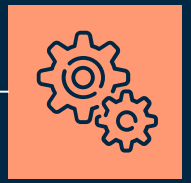
CONTENTS



01

PROBLEM

Analyzing the Problem Applicant Can repay a loan or not



02

OUR PROCESS

- Data Cleaning
- Univariate Analysis
- Bi & Multi variate Analysis

03

TARGET

Bringing out the insights as much as from the data on basis of visualization

Application Data

Columns Dropped Which Has NULL values More Than 40%

['OWN_CAR_AGE', 'EXT_SOURCE_1',
'APARTMENTS_AVG', 'BASEMENTAREA_AVG',
'YEARS_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', 'FONDKAPREMONT_MODE',
'HOUSETYPE_MODE', 'TOTALAREA_MODE',
'WALLSMATERIAL_MODE', 'EMERGENCYSTATE_MODE']

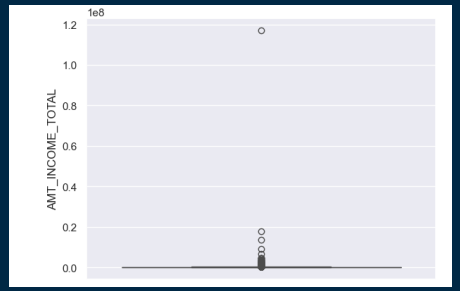
Removing Outliers

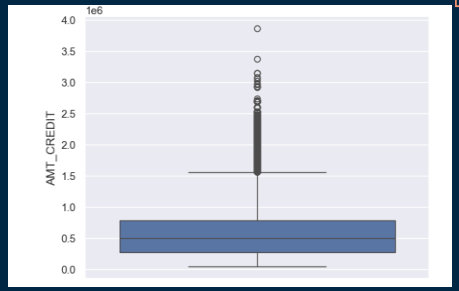
In AMT_INCOME_TOTAL :- As we can see above the person income is 117000000.0 and have 2 years of employed where age is 34 this i think won't be possible where occupation type is laborers definitely this is an outlier. So, we need to remove an outlier for amt_incoe_total .

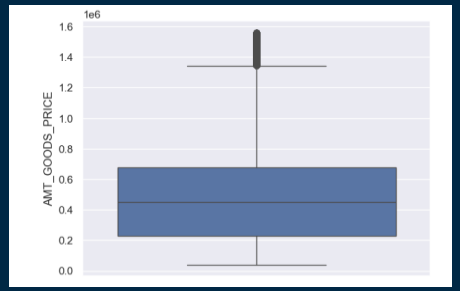
Dropped Outliers From AMT_CREDIT

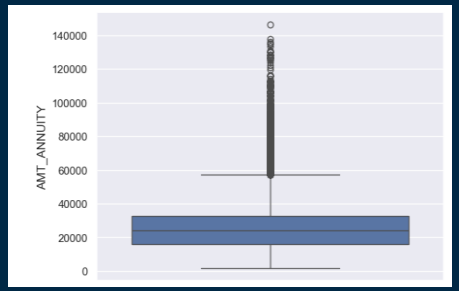
There are outliers in AMT_ANNUITY but we cannot drop these columns because amt annuity is the payment made by the client on the regular basis

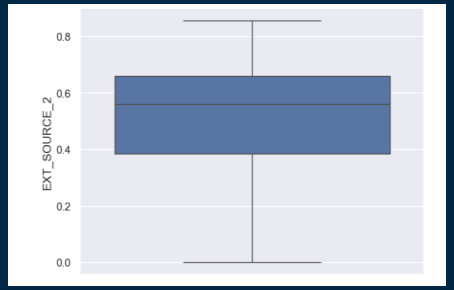
There are outliers in AMT_GOODS_PRICE but we cannot drop these columns because AMT_GOODS_PRICE is the price of goods for which client has taken the loan

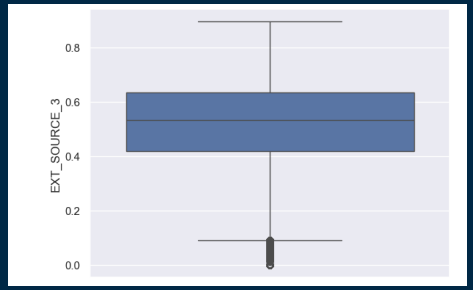






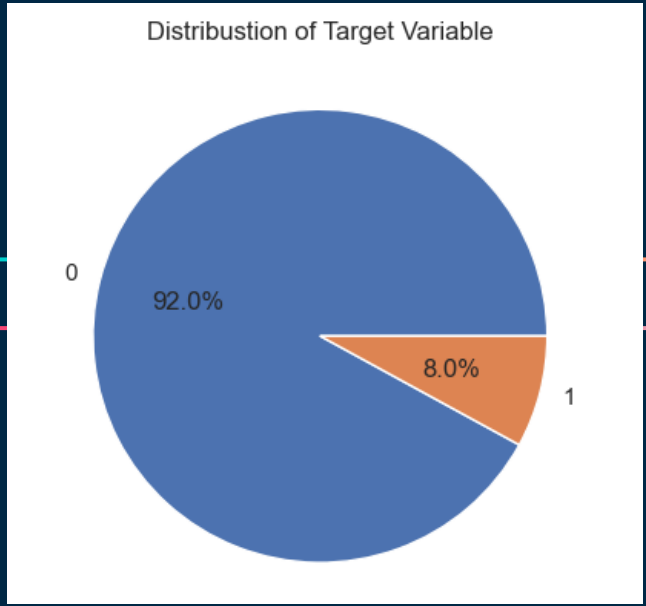






Distributon Of Target Variable

As we can see below only 8.0 % of the people are the defaulters as compare to repayor's that is 92.0% this is an imbalanced data where we can see there only few defaulters



1 represents the defaulters who get difficulty to repay the loan

0 represents the repayor who was able to replay the loan

Binning Some columns

Biining the columns give more analysis on the data as we can see binning gives more insights to the particular columns

Age : - We can there are more number of people between 30- 40 age group have applied for the loan

AMT_INCOME_GROUP :- we see from the data that whose income is between 100k – 200k has applied for the loan

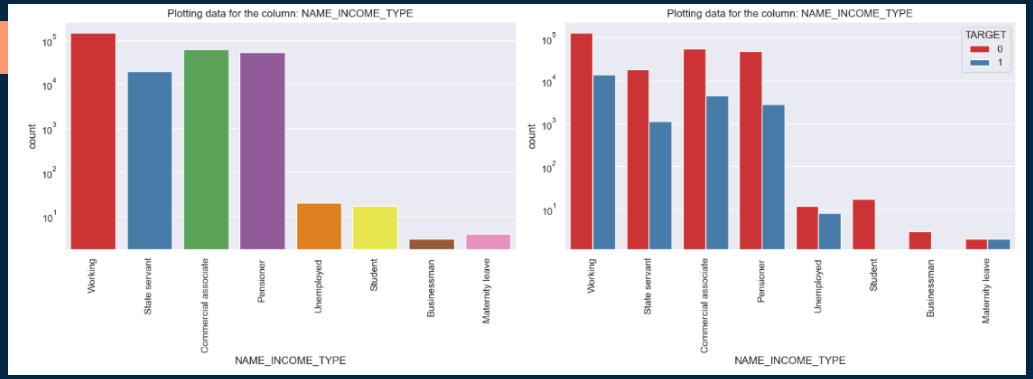
AMT_CREDIT_GROUP :- we can see here that there is maximum 45k – 270k amount mostly Credited

```
Age
30-40     75076
40-50     66883
50-60     62511
20-30     49094
60-70     28127
10-20         1
0-10          0
70-80         0
80-90         0
90-100        0
Name: count, dtype: int64
```

```
AMT_INCOME_GROUP
100k - 200k    151617
200k - 300k     61073
80k - 90k       30147
60k - 70k       12733
300k - 400k      6133
70k - 80k        5841
90k - 100k       5674
50k - 60k        3995
40k - 50k        3522
30k - 40k         818
25k - 30k         139
Name: count, dtype: int64
```

```
AMT_CREDIT_GROUP
45k-270k          75968
513.5k-808.65k    73002
270k-513.5k       72055
808.65k-1.2M      43273
1.2M-1.6M         17170
1.6M-2M               0
2M-2.5M               0
2.5M-4.05M            0
Name: count, dtype: int64
```

Univariate Analysis with Target column



Count of the people taken loan

As we can see mostly all week's clients has taken the loan except Sunday because mostly Sunday are holidays

Loan Defaulter

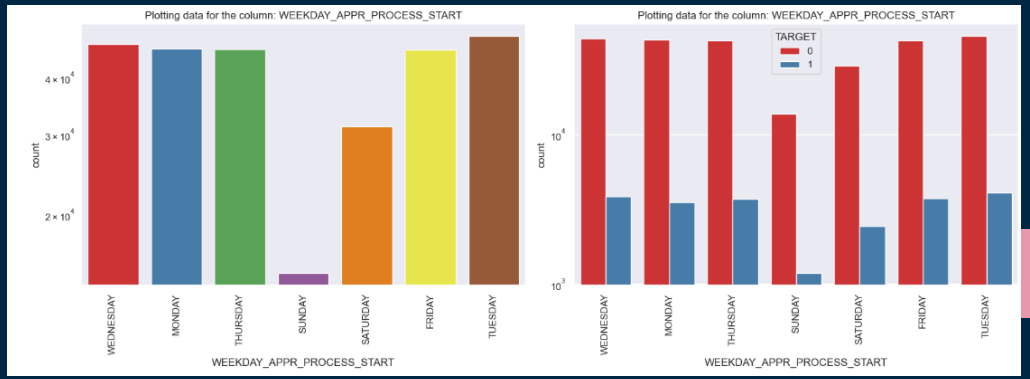
1. As I can see the plot with the default loan repay there nothing much we can see in the weekday_app_process_start column I can also drop this column from my analysis

Count of the people taken loan

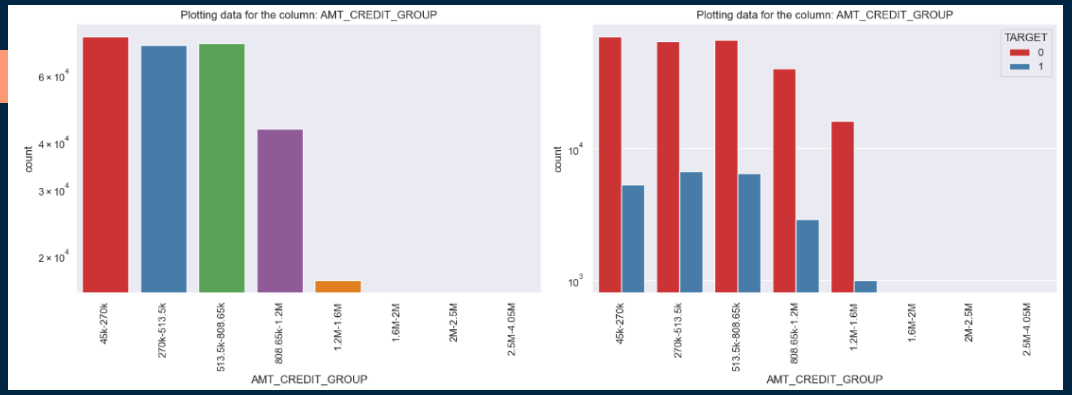
Most of client for loans have income type as Working, followed by Commercial associate, Pensioner and State servant

Loan Defaulter

1. if we see the client who are on maternity leave has the highest possibility not to repay the loan, followed by unemployed.
2. Students and businessman are not done in default ratio these are group which safer to provide the loan.



Univariate Analysis with Target column



Count of the people taken loan

As we can see mostly of the people taken loan is between 45 k - 270k

Loan Defaulter

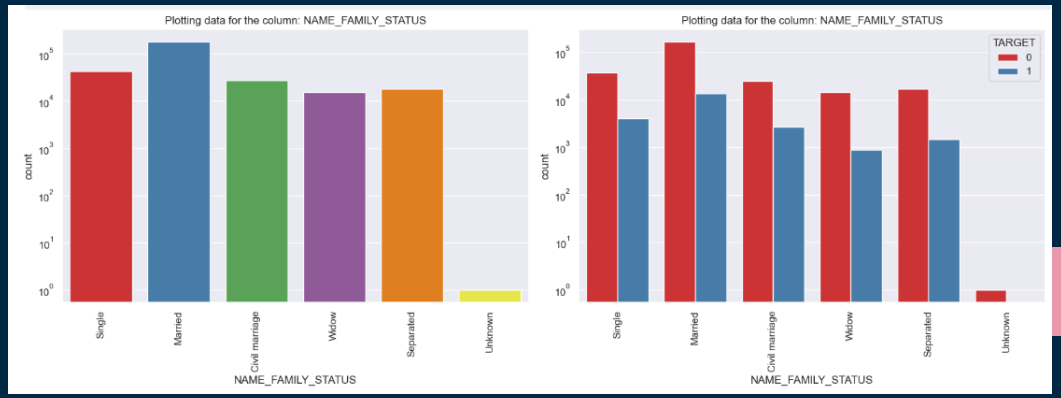
1. People taken loan between 1.6 M - 4.05 M are less defaulters

Count of the people taken loan

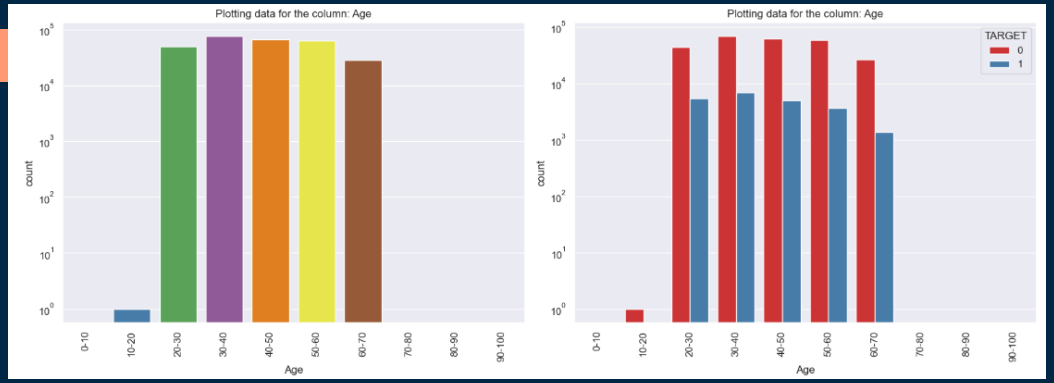
Most of the people who has taken loan are married followed by single and civil marriage

Loan Defaulter

1. if we see civil marriage has the highest percent of not repayment with Widow the lowest.



Univariate Analysis with Target column



Count of the people taken loan

1. We can see all the age group has taken loan
2. .10 - 20 age group are very less who has taken the loan because majority of the people are students or and who has taken the loan has taken education loan who are above 18 years of the age group

Loan Defaulter

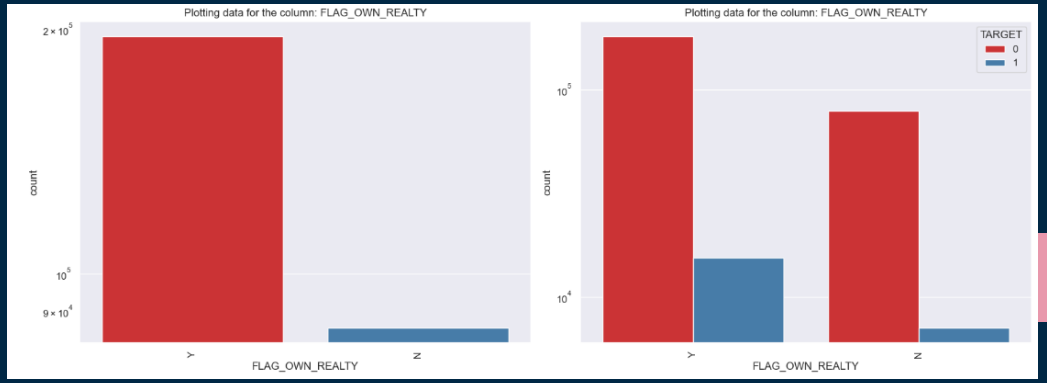
1. Majority of the people who are under 30- 40 has done loan default
2. As the age increases there are less chance to do the default in loan > 70 year of age

Count of the people taken loan

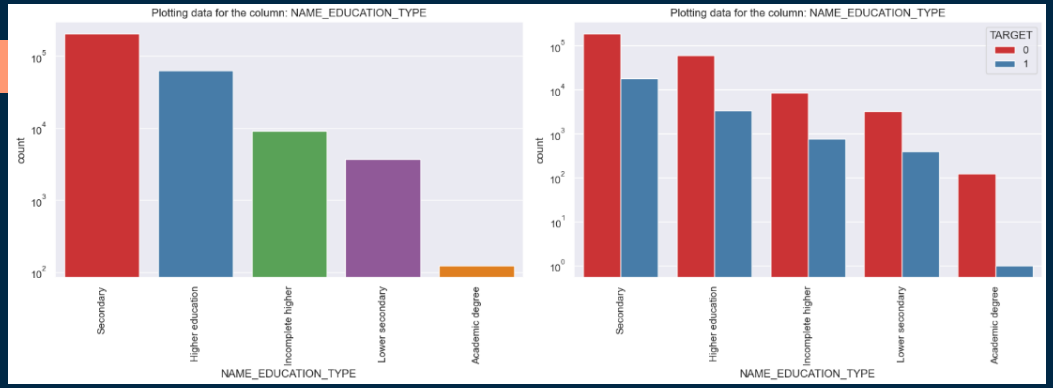
1. The majority of clients in the dataset own real estate.
2. Clients who own real estate are more than double in number compared to those who don't own

Loan Defaulter

1. Based on the data, there doesn't appear to be a significant correlation between owning real estate and loan defaulting.



Univariate Analysis with Target column



Count of the people taken loan

1. The most of the people are secondary educated followed by higher education very a smaller number of academic degree holder

Loan Defaulter

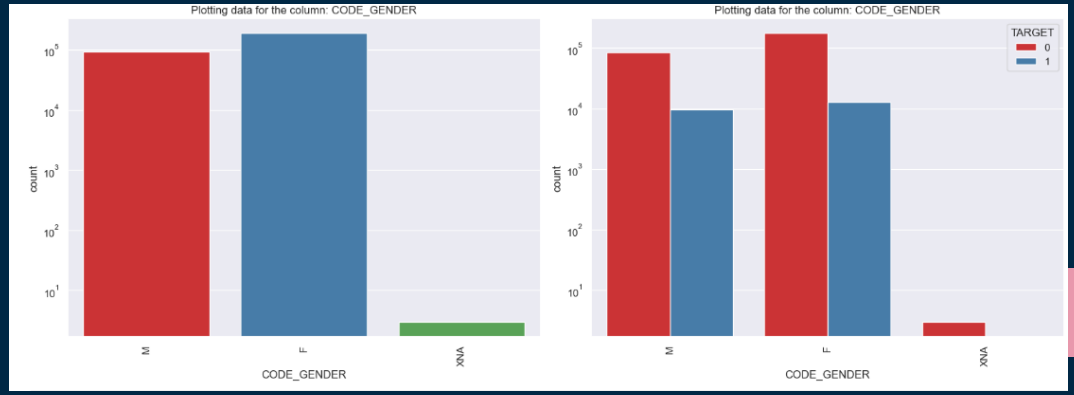
1. lower secondary, secondar and incomplete higher are highest number of not repaying the loan

Count of the people taken loan

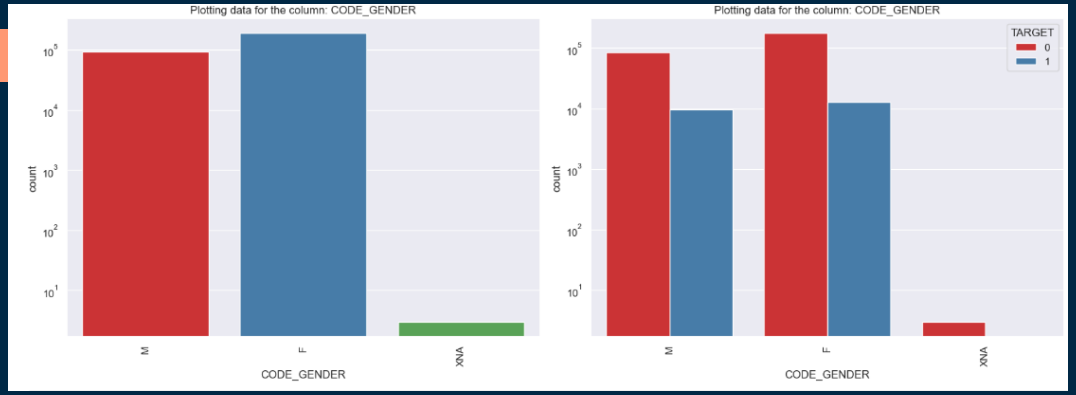
1. As compare to male female has more number loans taken

Loan Defaulter

1. As we can see there male as compare to female who has not replayed the loan



Univariate Analysis with Target column



Count of the people taken loan

1. As compare to male female has more number loans taken

Loan Defaulter

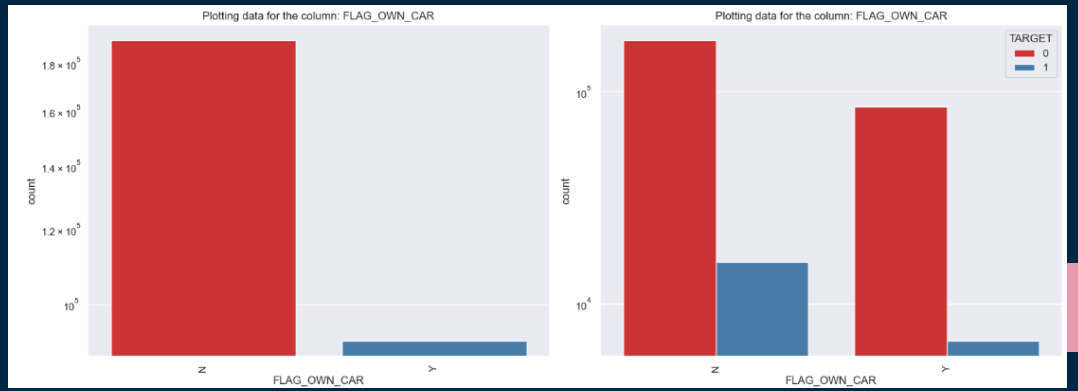
1. As we can see there male as compare to female who has not replayed the loan

Count of the people taken loan

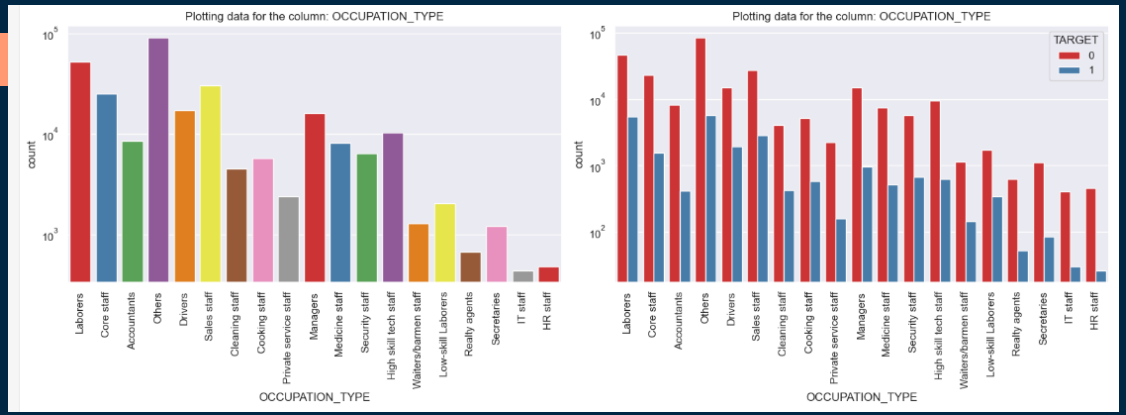
1. Clients who own a car are half in number of the clients who don't own a car.

Loan Defaulter

1. But based on the percentage of deault, there is no correlation between owning a car and loan repayment as in both cases the default percentage is almost same.



Univariate Analysis with Target column



Count of the people taken loan

1. The majority of people who has taken a loan comes under "laborers" , "Others" (These are the clients whose occupation is not mentioned in the data) ,"Core staff", "Sales Staff“
2. It and Hr. Staff comes under small group who has taken the loan

Loan Defaulter

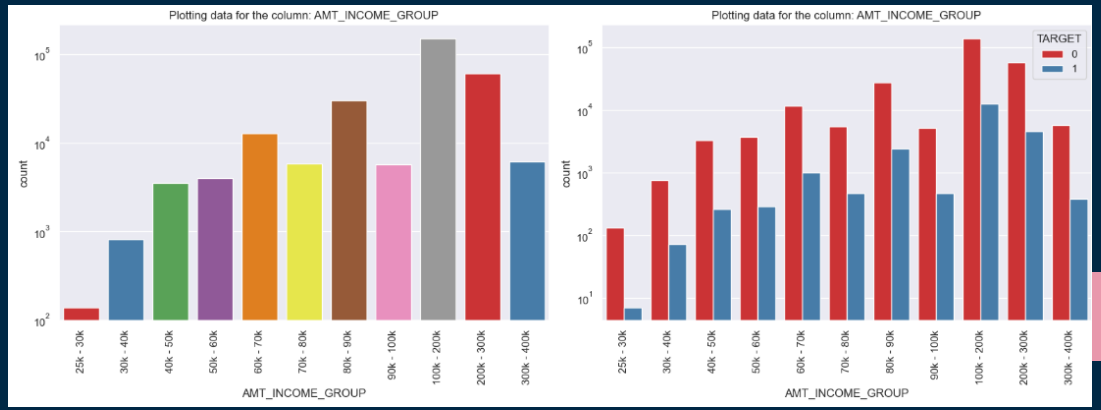
1. Majority of the people who has taken the loan are comes under the loan defaulter
2. We can also see to highest category is "Laborers" and "Others" and "Drivers" ,"Sales Staff", "Cleaning Staff" who are not high skilled to their education

Count of the people taken loan

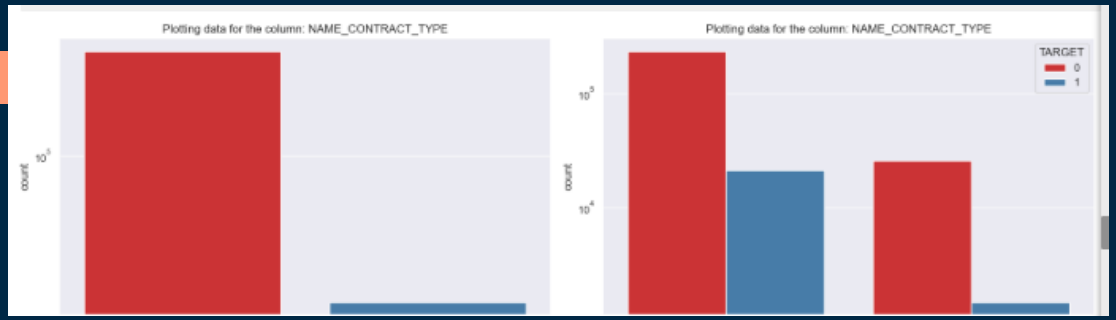
1. As we can see mostly of the people taken loan is between 45 k - 270k

Loan Defaulter

1. People taken loan between 1.6 M - 4.05 M are less deaulters



Univariate Analysis with Target column



Count of the people taken loan

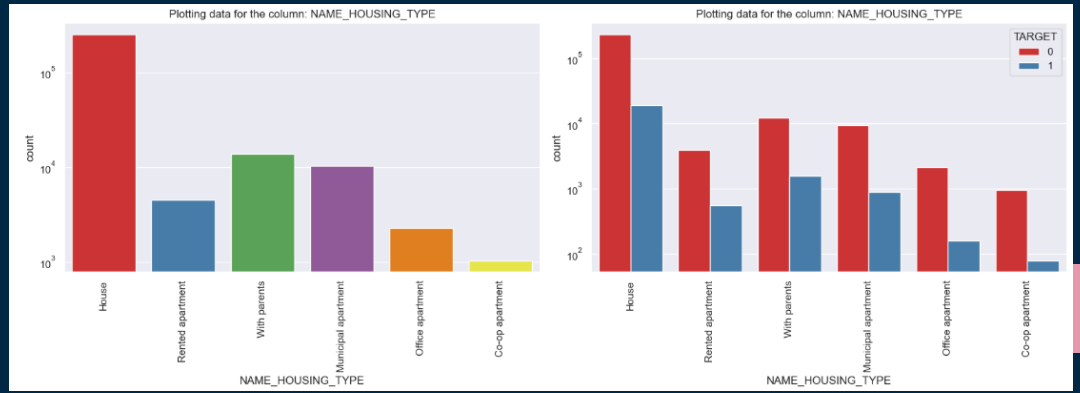
1. The majority of the client own a house. Flat that is just a double who don't owe a flat.
2. As compare to target variable we see that there is no much difference between the owning the flat and loan defaulters it can seen that the people who don't owe a flat also considered as a defaulter

Count of the people taken loan

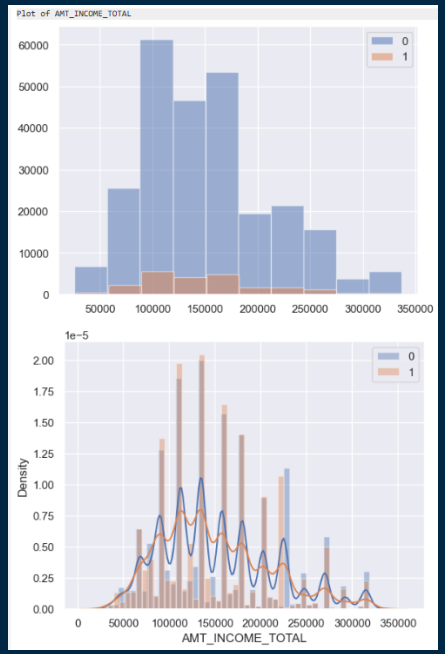
1. the Majority of the people has taken loan who owns house, and leaving with parent

Loan Defaulter

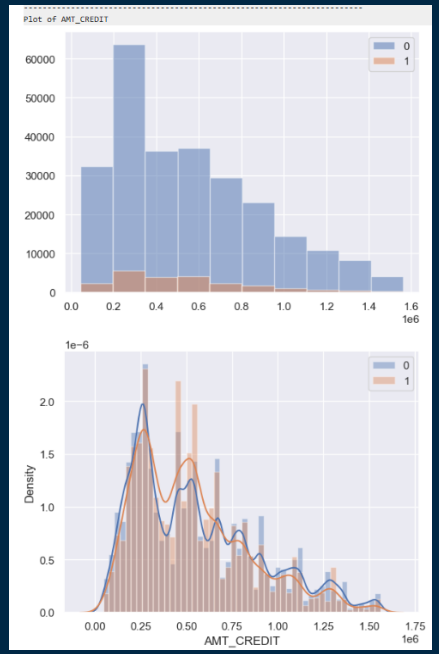
1. People leaving in house and with parents are highest defaulters
2. The people leaving in co apartment rate are less defaulters



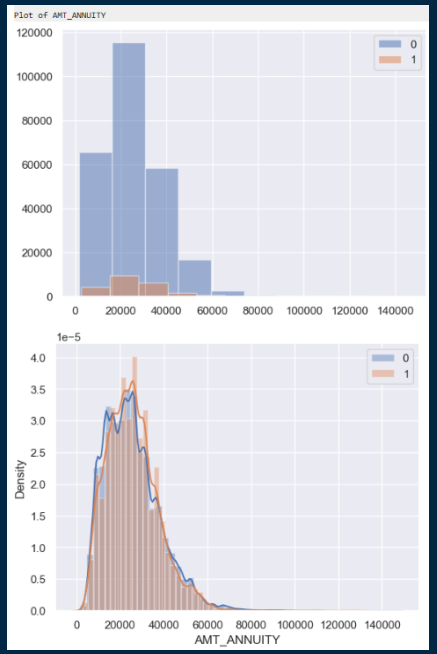
Univariate Analysis with Target column



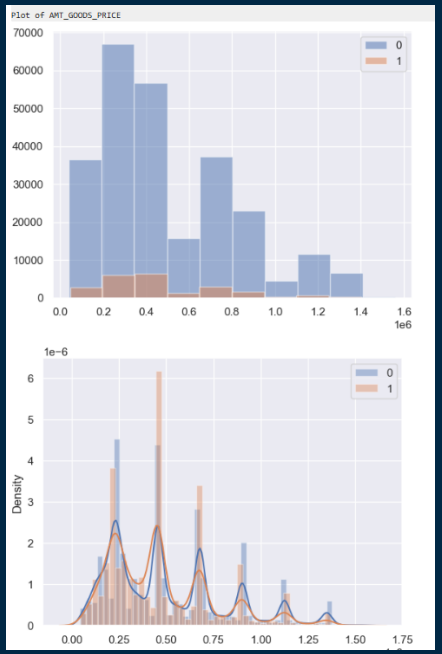
The income type for both targeted type lied around 1000000 to 150000



Most of the credit amount 25k to 50 k



Most client pay an annuity below 60K for their credit loans.

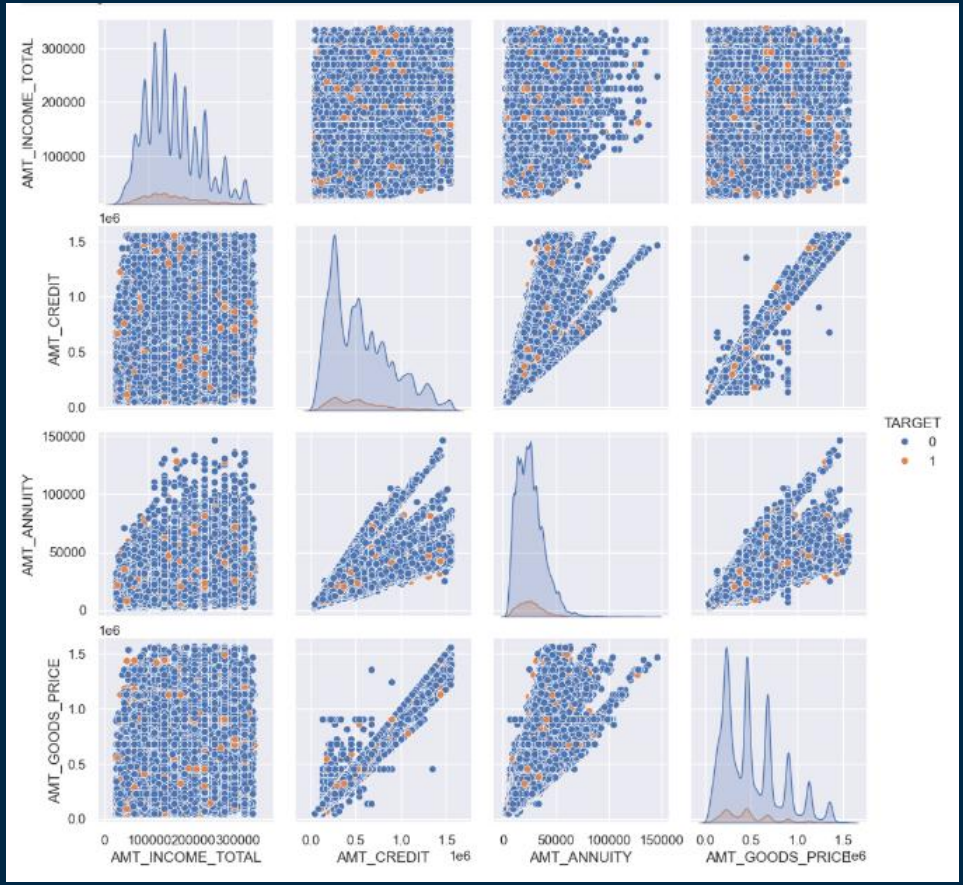


The majority of loans are granted for goods prices below 1M.

Bi & Multivariate Analysis with Target column

AMT_CREDIT and AMT_GOODS_PRICE are highly corelated to each other as based on the scatterplot where most of the data are consolidated in form of a line
There are very less defaulters for AMT_CREDIT > 1M

There are very few defaulters for loan amounts exceeding 1M. This suggests that clients who borrow larger amount are less likely to default on their loans.

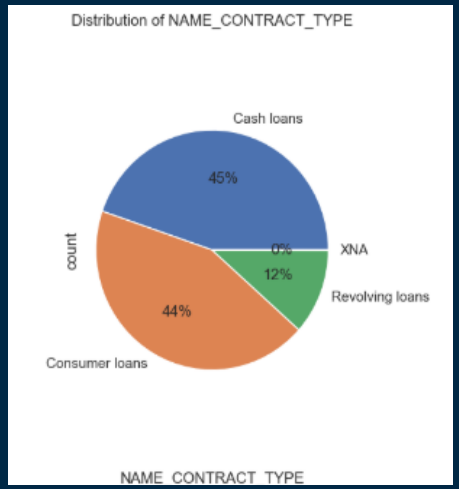


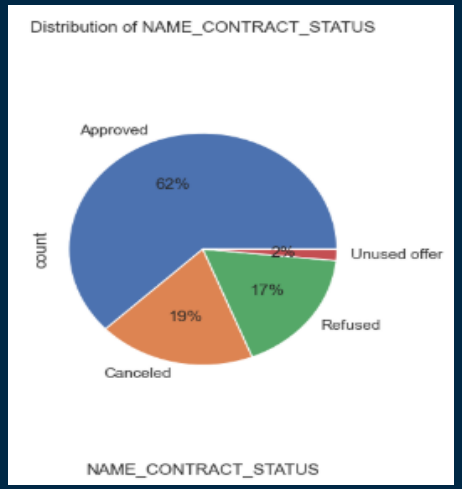
Previous Application Data

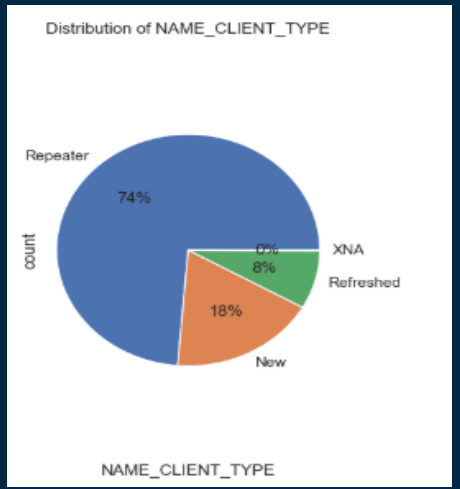
Columns Dropped Which Has NULL values More Than 40%

['AMT_DOWN_PAYMENT',
'RATE_DOWN_PAYMENT',
'RATE_INTEREST_PRIMARY',
'RATE_INTEREST_PRIVILEGED',
'NAME_TYPE_SUITE',
'DAYS_FIRST_DRAWING',
'DAYS_FIRST_DUE',
'DAYS_LAST_DUE_1ST_VERSION',
'DAYS_LAST_DUE', 'DAYS_TERMINATION',
'NFLAG_INSURED_ON_APPROVAL']

Analyzing Columns With Percentage







Name Contract Type

1. Cash loans: The most common contract type, accounting for 45% of all contracts.
2. Consumer loans: A significant portion of contracts, making up 44% of the total.3. Revolving loans: Less common, representing 12% of contracts. Typically associated with credit cards or lines of credit.4. There is 0 % of XNA category

Name Contract Status

1. The majority of loan applications were successfully approved ie 62% approved loan..
2. Approximately 19% of applicants canceled their loan requests
3. About 17% of applications were refused, highlighting the importance of assessing rejection criteria and applicant profile
4. A small percentage received offers but didn't proceed, warranting investigation into factors influencing this decisions.

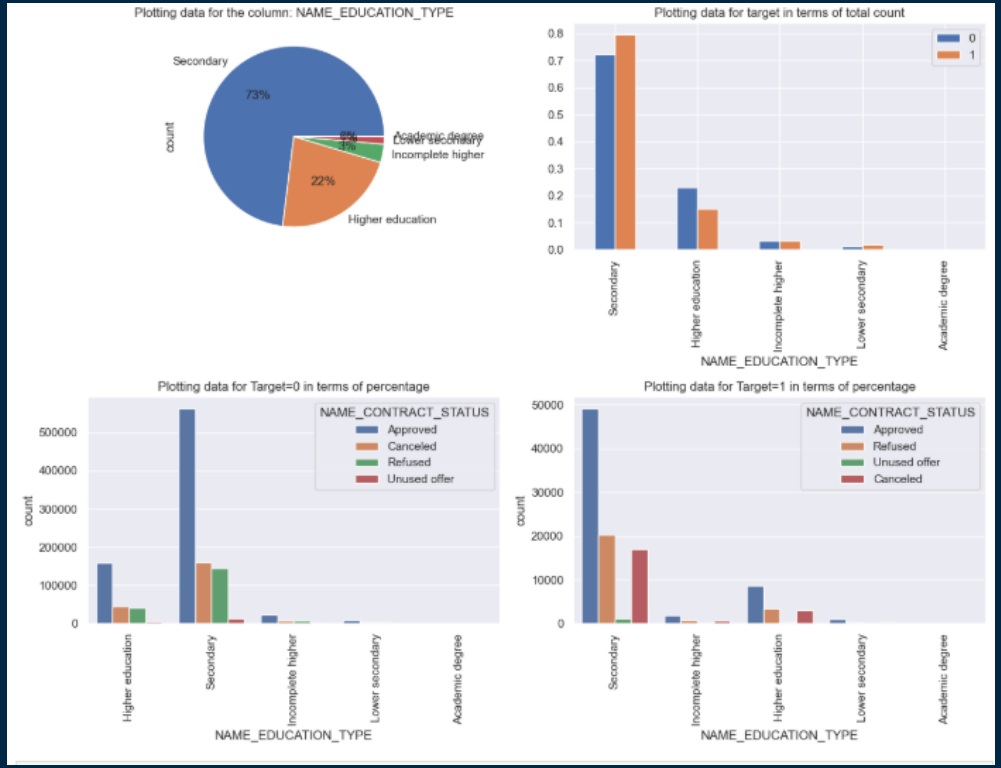
Name Client Type

1. Repeater: 74% of clients fall into this category.
2. New: Approximately 18% of clients are categorized as "New."
3. Refreshed: About 8% of clients are classified as "Refreshed."
4. There is 0 % of XNA percentage

Bi & Multivariate Analysis with Target column

NAME_EDUCATION_TYPE VS NAME_CONTRACT_STATUS

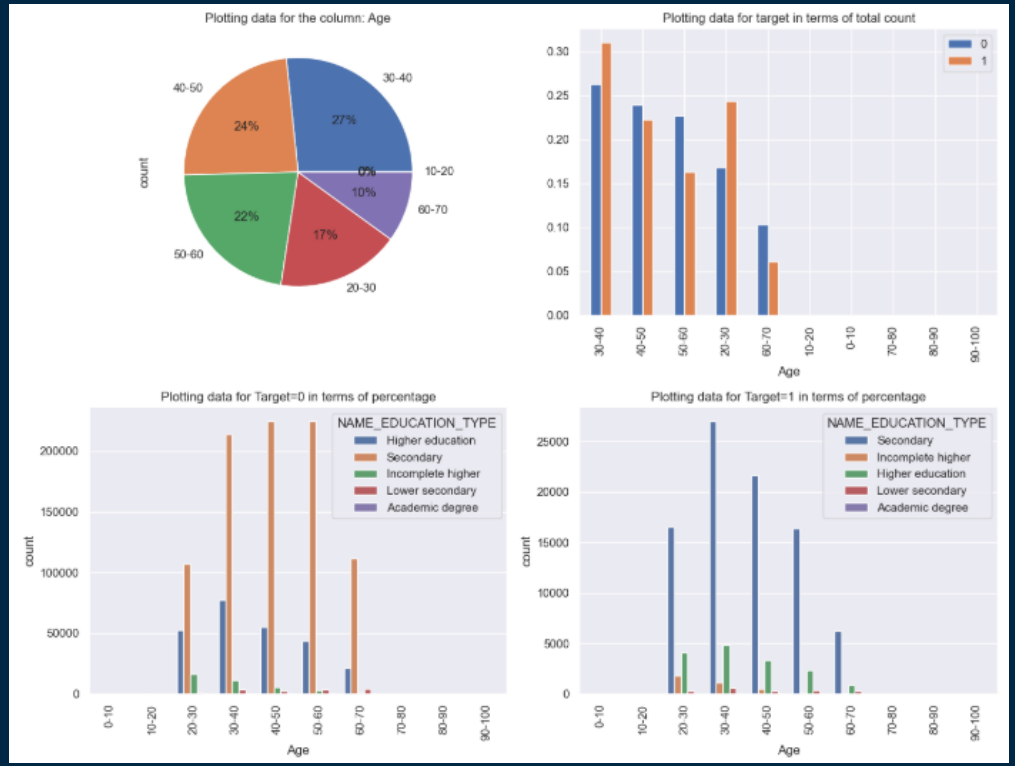
As we can see higher education and secondary education gives more approved loans and as able repay the loan as well There is less defaulter percentage if we give a loan to higher education people



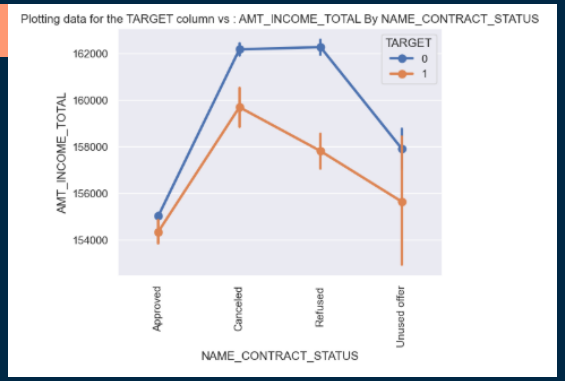
Bi & Multivariate Analysis with Target column

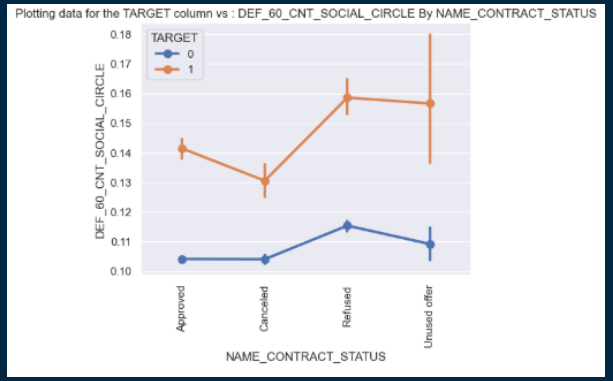
Age VS
NAME_EDUCATION_TYPE

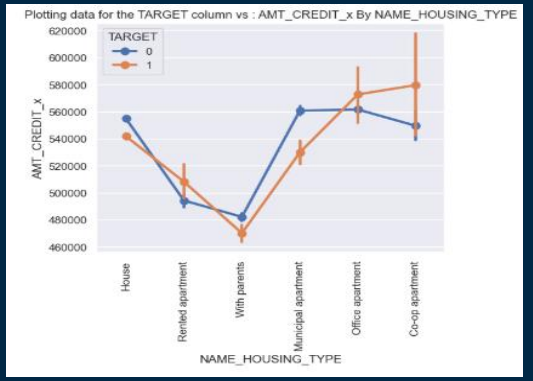
There is a majority of people who are under age of 30- 40 year of age who has repay the loan we can target the people under the age of 30 to 60 year of age groups who are safe to give the loan and higher chances they repay the loan as well with this particular data We have to take care of clients who has not done higher education or whose education are incomplete they can be defaulter



Analuzing Columns With Percentage







we can see here the people whose income is higher than others have not used the offer we can target these kind of people as well

Clients who have average of 0.13 or higher DEF_60_CNT_SOCIAL_CIRCLE score tend to default more and hence client's social circle has to be analyzed before providing the loan.

The People who are leaving in co-op apartment and office apartment has higher chances to be defaulter

Conclusion & Recommendation For Repayers

- AGE : - Loan applicants above the age of 60 has a lower tendency to default.
- NAME_INCOME_TYPE :- Student and Businessmen have no defaults.
- AMT_INCOME_TOTAL :- Applicants with incomes exceeding 10 Millions experience a lower likelihood of default.
- NAME_HOUSING_TYPE :- Applicant leaving in there own house, or parents , or rental apartment are lower likelihood of default
- CNT_CHILDREN:- People with zero to two children tend to repay the loans.

Conclusion & Recommendation For Defaulters

- NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education has higher defaulter
- CNT_CHILDREN: Client who have children equal to or more than 9 default 100% and hence their applications are to be rejected.
- CODE_GENDER: Men are at relatively higher default rate
- NAME_FAMILY_STATUS : People who have civil marriage or who are single default a lot.
- OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters staff, Security staff, Laborers and Cooking staff has the higher default rate .
- DAYS_BIRTH: Avoid young people who are in age group of 20-30 as they have higher probability of defaulting
- AMT_GOODS_PRICE: When the credit amount goes beyond 3M, there is an increase in defaulters.