CREDIT EDA CASE STUDY

Vikas Ware_DS53 Batch

INTRODUCTION

- 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 aconsumer finance company which specialises in lending various types of loans to urban customers.
- 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.*
- 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 loanApplication*
  - ***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.*
- The data has 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 firstY instalments of the loan in our sample,*
  - ***All other cases:*** *All other cases when the payment is paid on time.*

BUSINESS OBJECTIVE

- This 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 utilise this knowledge for its portfolio and risk assessment.

DATA SET

- The dataset has 3 files as explained below:

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

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

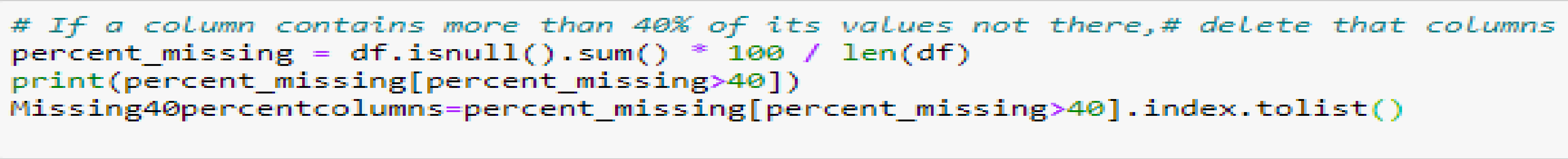
- '**columns_description.csv**' is data dictionary which describes the meaning of the variables.

EDA STUDY

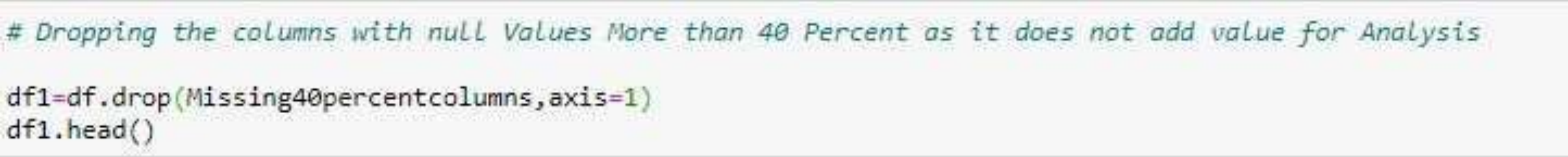
- For this problem statement,we will be adapting the Exploratory data analysis using data visualisation techniques to draw inferences and obtain insights from them.As we are aware that besides plotting graphs or visualising data,EDA is more about understanding and studying the given data in detail.Visualisation of data into plots/graphs can be termed one of the tools in the EDA process.
- We will be using the following approach:
- ***Data Cleansing:*** Irregularities may appear in the form of missing values,anomalies/outliers,incorrect format and inconsistent spelling,etc.Hence, this is an important step in our case study.
- ***Fixing the Rows and Columns:*** Analysing the rows and columns and removing the unwanted columns which are not required for the analysis.
- ***Handling /Replacing Missing Values:*** Identify values that indicate missing data. Deleting rows if the number of missing values is insignificant. Fill partial missing values using business judgement.
- ***Handling Outliers:*** Outliers are values that are much beyond or far from the next nearest data points. So handling them is also important.
- ***Standardising derived variables***:To observe all the variables for standard unit, make relevant changes for Boolean data,etc.
- ***Univariate and MultivariateAnalysis***:Analysing a single column/variable or multiple column/variable to show better insights.

DATA CLEANSING - HANDLING MISSING VALUES

- We deleted the columns which contained more than 40% of its values as it does not add value forAnalysis.



```
# If a column contains more than 40% of its values not there,# delete that columns
percent_missing = df.isnull().sum() * 100 / len(df)
print(percent_missing[percent_missing>40])
Missing40percentcolumns=percent_missing[percent_missing>40].index.tolist()
```



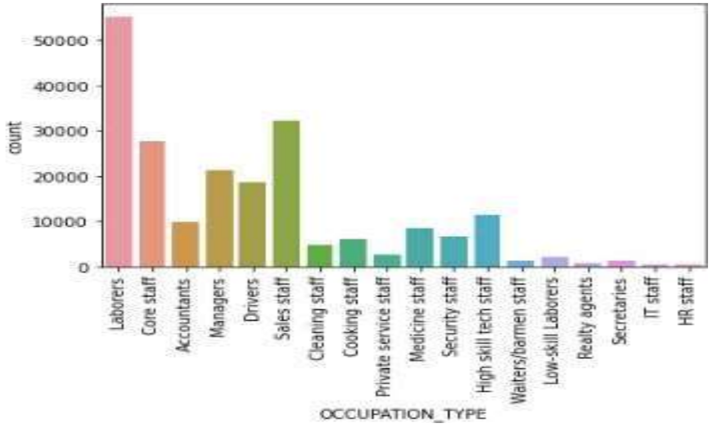
```
# Dropping the columns with null Values More than 40 Percent as it does not add value for Analysis

df1=df.drop(Missing40percentcolumns,axis=1)
df1.head()
```

- After removing the 40% values,there were 73 columns and 307511 Rows remaining in the dataset.

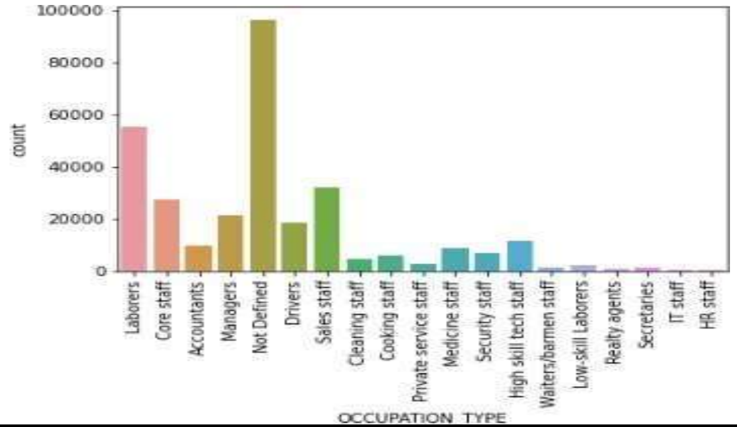
HANDLING THE MISSING VALUES IN CATEGORICAL COLUMNS

- The Null value count for OCCUPATION_TYPE is 96391 and around 31 Percent values are missing .We decided to impute the missing values by adding text as 'Not Defined' as imputing with mode or any other value may lead to incorrect analysis.

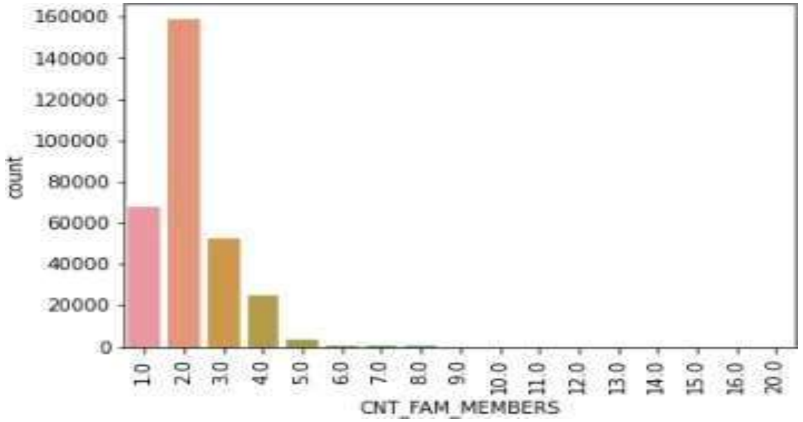
BEFORE

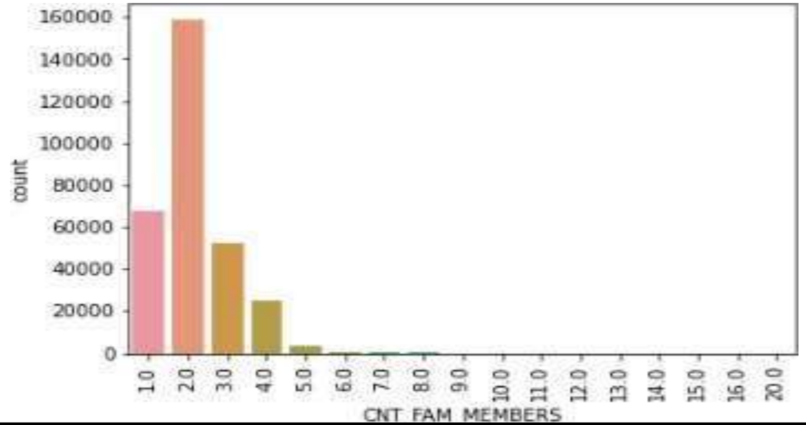


AFTER

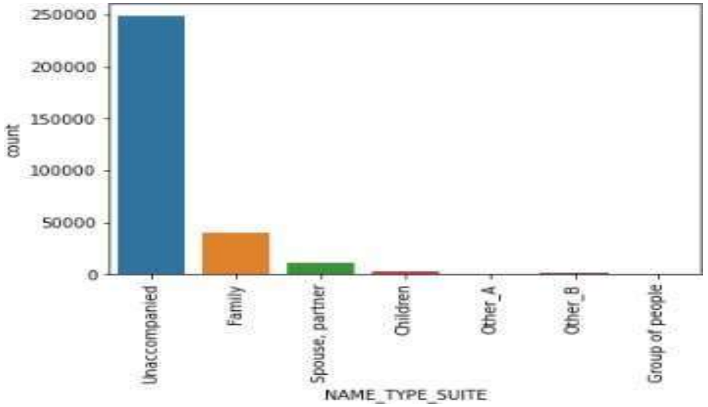


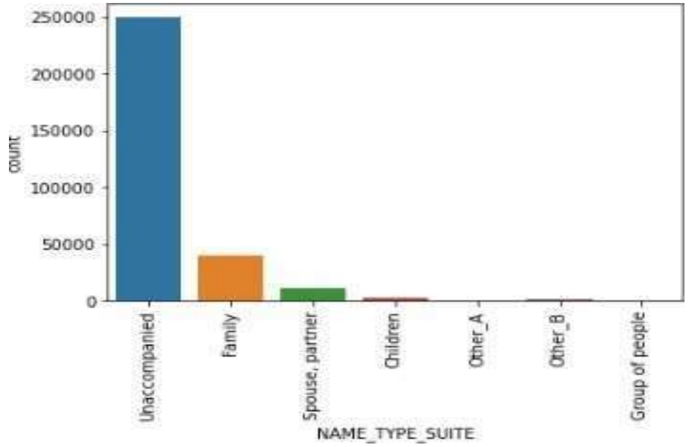
- The Column CNT_FAM_MEMBERS has a Null Value count of 2 .So imputing the missing values with mode of the distribution would be ideal as the number of missing values is less and it would not impact the distribution.





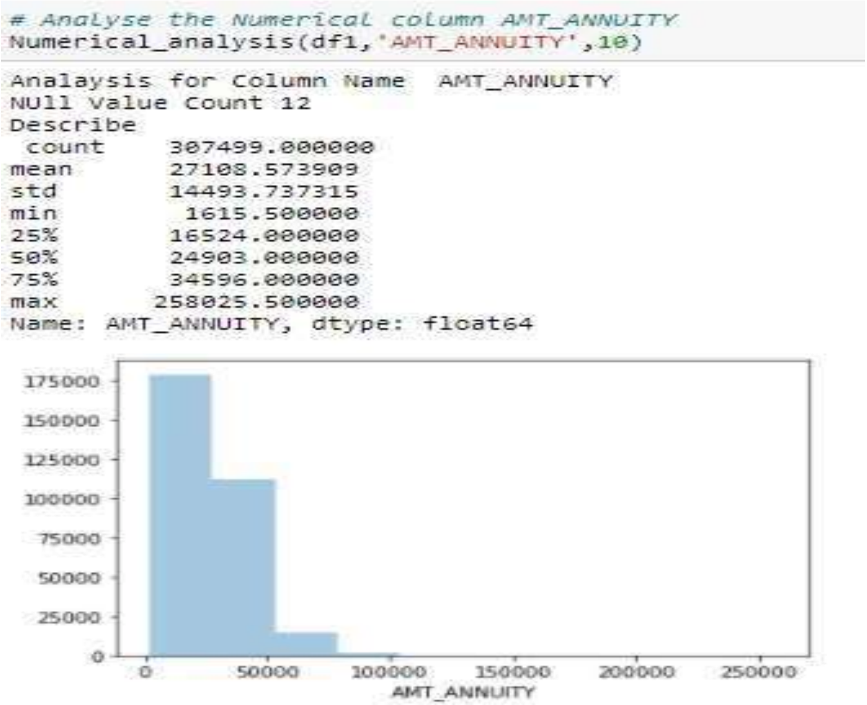
- The Column NAME_TYPE_SUITE has a Null Value count of 1292 and missing percent is around 0.332021 .So imputing the missing values with mode of the distribution would be ideal as the percentage missing values is less.





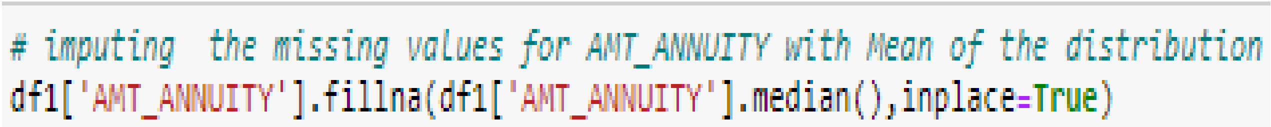
HANDLING THE MISSING VALUES IN NUMERICAL COLUMNS IMPUTING NULL VALUES WITH MEAN

- The Null value count in most of the numerical variables were replaced with mean values to get better insights of the data.
- Columns were *EXT_SOURCE_3, AMT_REQ_CREDIT_BUREAU_YEAR, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_QRT, OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE,OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE, EXT_SOURCE_2, AMT_GOODS_PRICE,AMT_ANNUITY.*
- The example screenshot below explains the technique and code used for replacing null values with the mean values.



```
# Analyse the Numerical column AMT_ANNUITY
Numerical_analysis(df1,'AMT_ANNUITY',10)
```

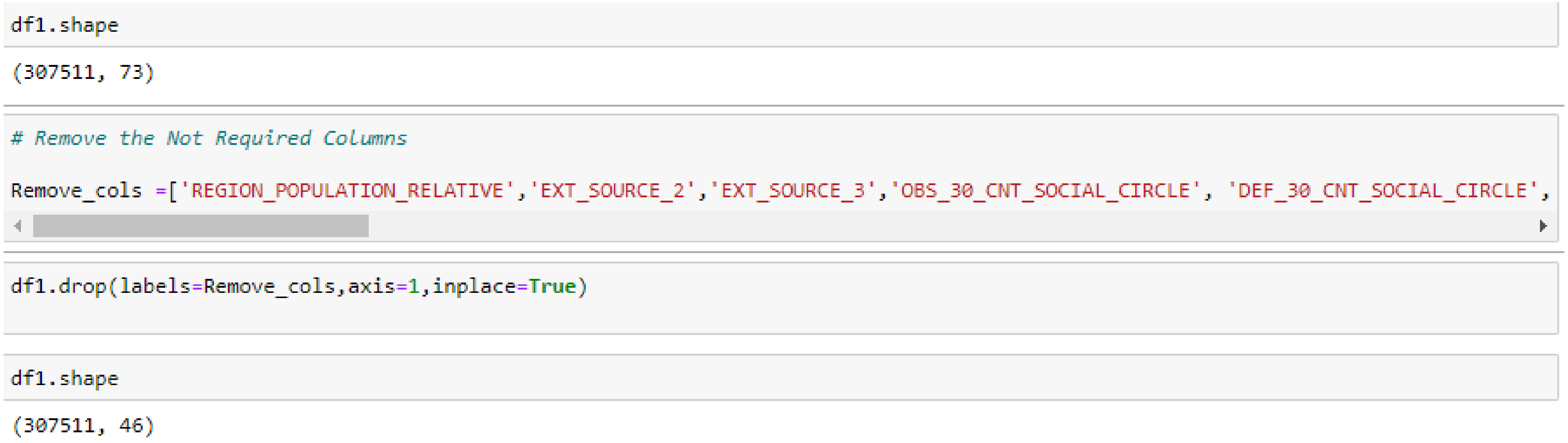
```
Analaysis for Column Name  AMT_ANNUITY
NUll Value Count 12
Describe
 count     307499.000000
mean       27108.573909
std        14493.737315
min         1615.500000
25%        16524.000000
50%        24903.000000
75%        34596.000000
max       258025.500000
Name: AMT_ANNUITY, dtype: float64
```



```
# imputing  the missing values for AMT_ANNUITY with Mean of the distribution
df1['AMT_ANNUITY'].fillna(df1['AMT_ANNUITY'].median(),inplace=True)
```

FIXING THE ROWS AND COLUMNS

- After a deep analysis of the rows and columns, we decided to remove the columns which are not required for Analysis.This will help us to get better insights w.r.t data.Earlier there were 73 columns. After executing the drop query, the numbers of columns were decreased to 46.



```
df1.shape
```

```
(307511, 73)
```

```
# Remove the Not Required Columns

Remove_cols =['REGION_POPULATION_RELATIVE','EXT_SOURCE_2','EXT_SOURCE_3','OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE',
```

```
df1.drop(labels=Remove_cols,axis=1,inplace=True)
```

```
df1.shape
```

```
(307511, 46)
```

STANDARDISING DERIVED VARIABLES

- We reviewed the flag cols and other categorical variables and found that they had numerical values.Hence converted them to string values. Like 0-1 toY-N.

- It was also observed that there are some values by name XNA in CODE_GENDER,ORGANIZATION_TYPE columns . The value counts of XNA in organization_type was replaced with an already available option called‘Other’.

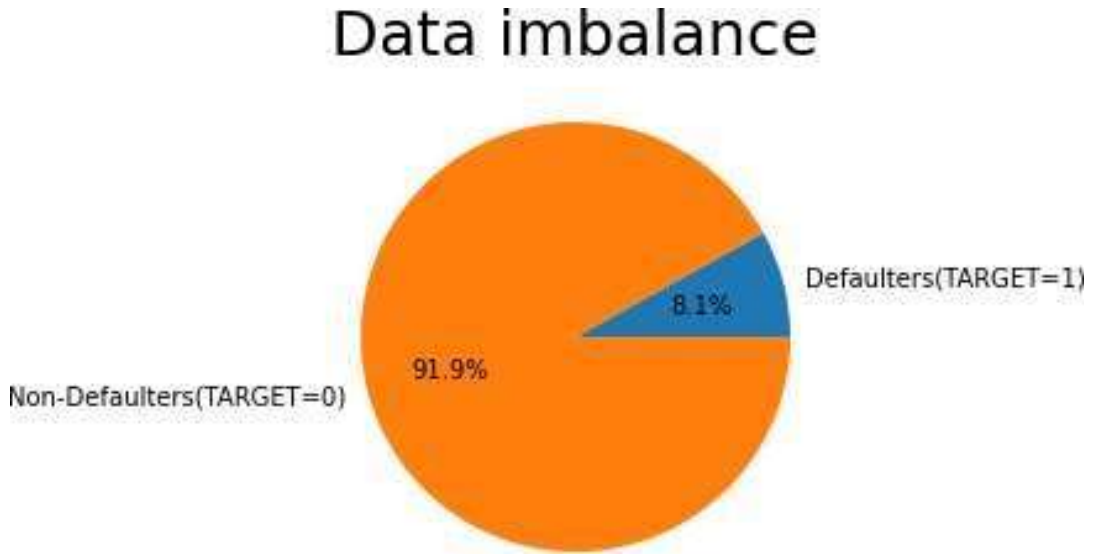
- The gender column had 4 XNA values which was replaced with ‘F’being the majority.

- We created bins for theAMT_INCOME_TOTAL columns as 'Lower', 'Middle', 'Higher', 'Very High‘.

- DAYS_BIRTH column determines the age of the individual at the time of loan application. Hence,we converted the DOB variable into age and put them into buckets of <30, 30-40, 40-50,50-60 and 60+.

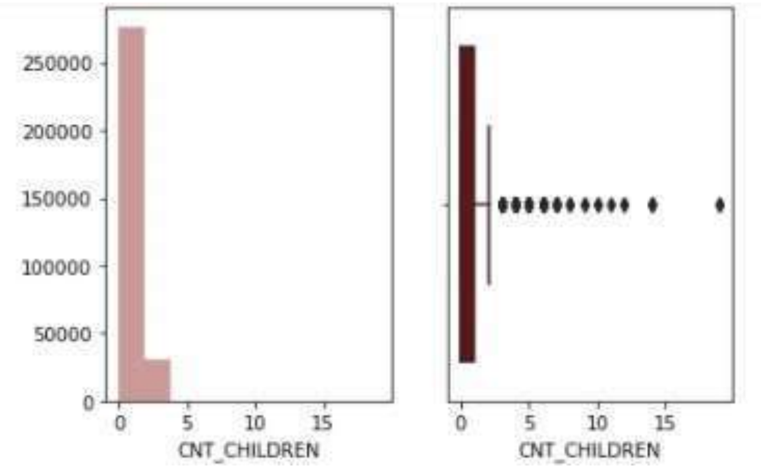
CHECKING DATA IMBALANCE

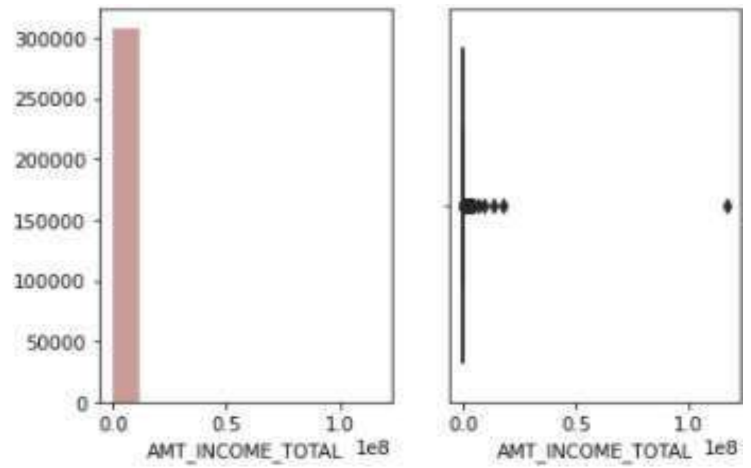
- The ratio of data imbalance in the target variable was 91.9% for Non-defaulters and 8.1% for Defaulters. Hence, to minimize the imbalance between the two variables, they were separated into two train_0 and train_1 variables.

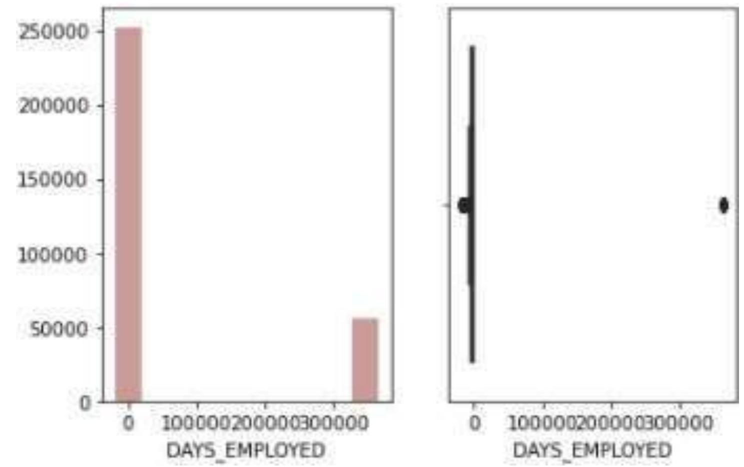


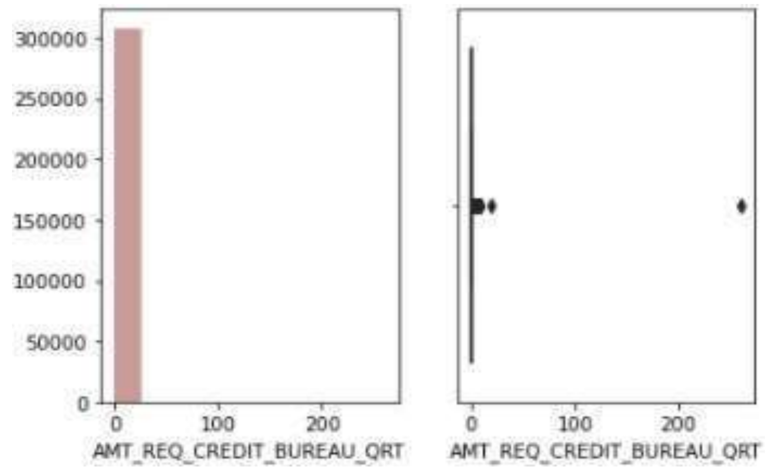
OUTLIER ANALYSIS

- We found out the outliers by using a distribution plot and box plot for different columns. The columns which have outliers were 'CNT_CHILDREN', 'AMT_INCOME_TOTAL', 'DAYS_EMPLOYED' 'AMT_REQ_CREDIT_BUREAU_QRT’.
- We removed the outliers from all the identified rows so that the data is normalised.





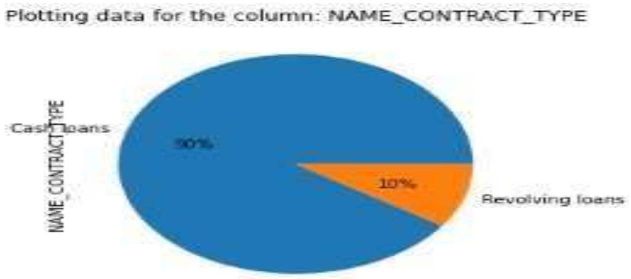


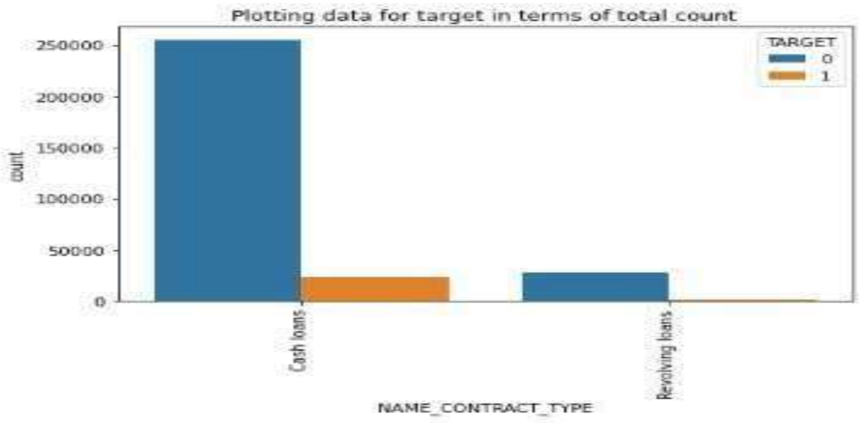


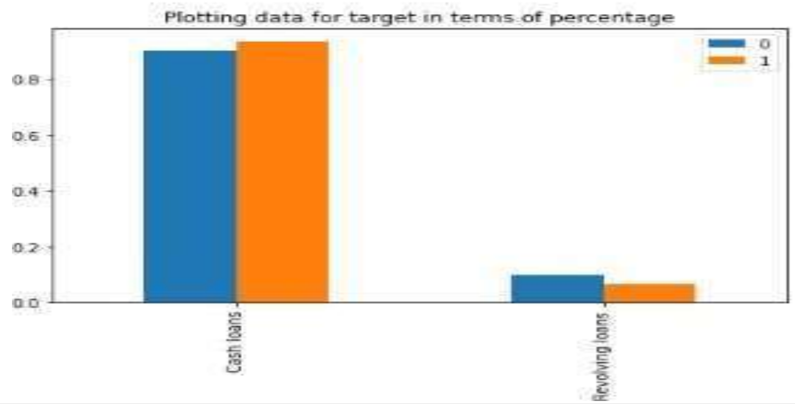
UNIVARIATE ANALYSIS OF THE CATEGORICAL DATA

Unordered Categorical Variables

- Name_Contract_Type is a column which gives the type of loan whether it is a Cash Loan or Revolving Loan .As per the Pie Chart ,we see that 90% of the loans are Cash loans and 10 % are Revolving Loans.The percentage of Defaulters(Target=1) for cash loans is more as compared to Revolving Loans .This can be attributed to revolving loans as more risk Free.


Plotting data for the column: NAME_CONTRACT_TYPE

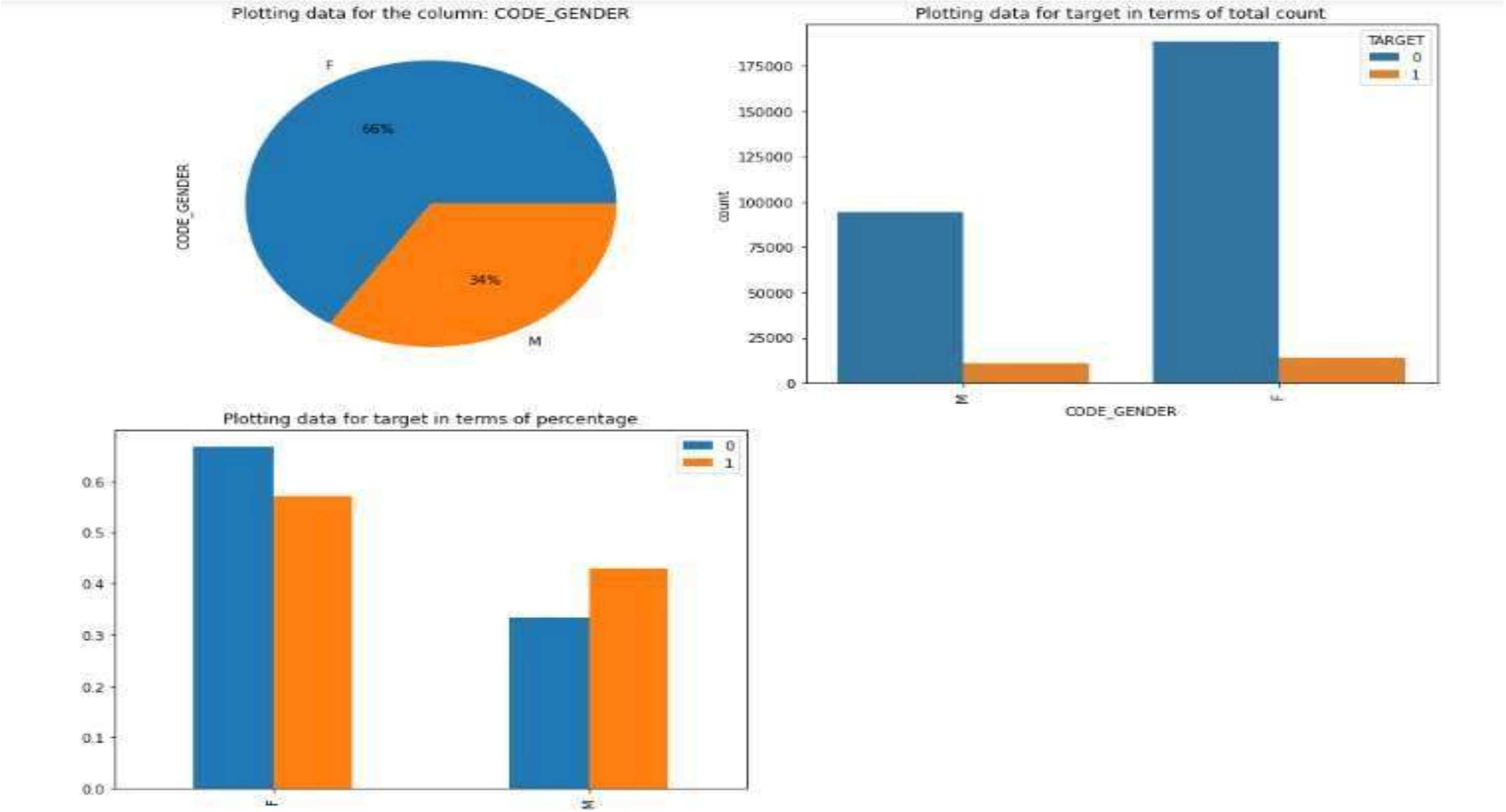

Plotting data for target in terms of total count


Plotting data for target in terms of percentage

UNIVARIATE ANALYSIS OF THE CATEGORICAL DATA

Unordered Categorical Variables

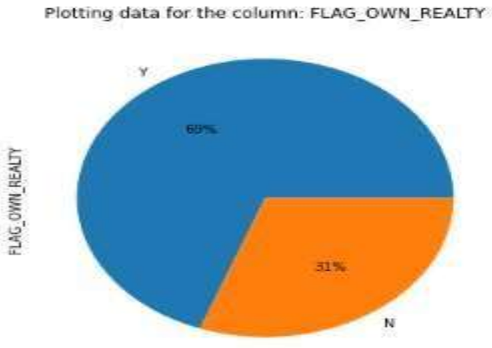
- CODE_GENDER which tells whether theApplicant is Male or Female.AS per the Pie Graph ,it is observed that 66% are females Applicants and 34% are MaleApplicants .Even though the female applicants are more ,MaleApplicants default more than Females.This puts giving loans to females is safer.

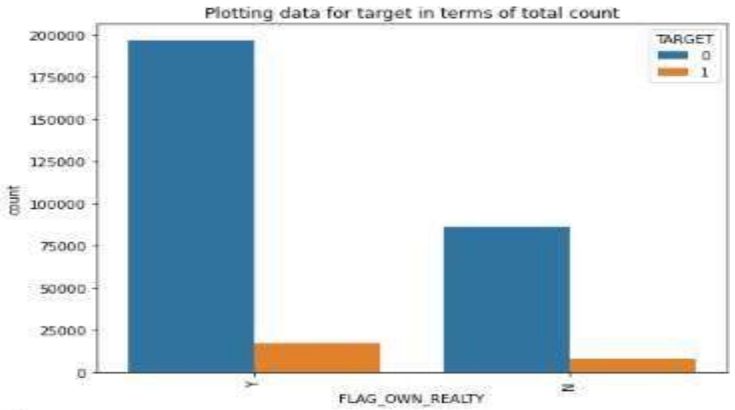


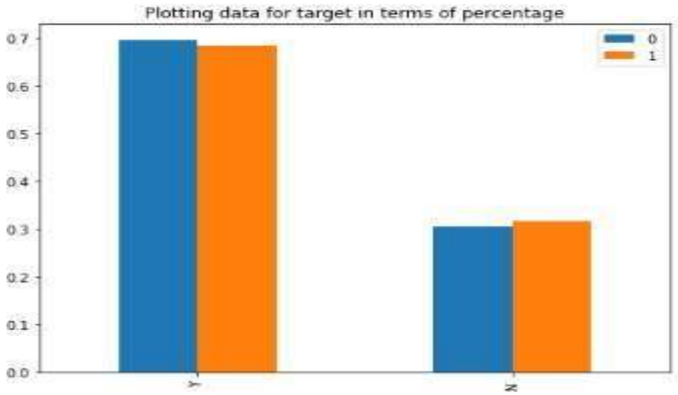
UNIVARIATE ANALYSIS OF THE CATEGORICAL DATA

Unordered Categorical Variables

- FLAG_OWN_REALTY is a column which tells wehhter the applicant owns house/flat.As per the pie Graph ,we can infer that 69% own homes .However as the margin is less between Defaulters and Non Defaulters ,we can say that is difficult to take a loan decision based on this Metric.



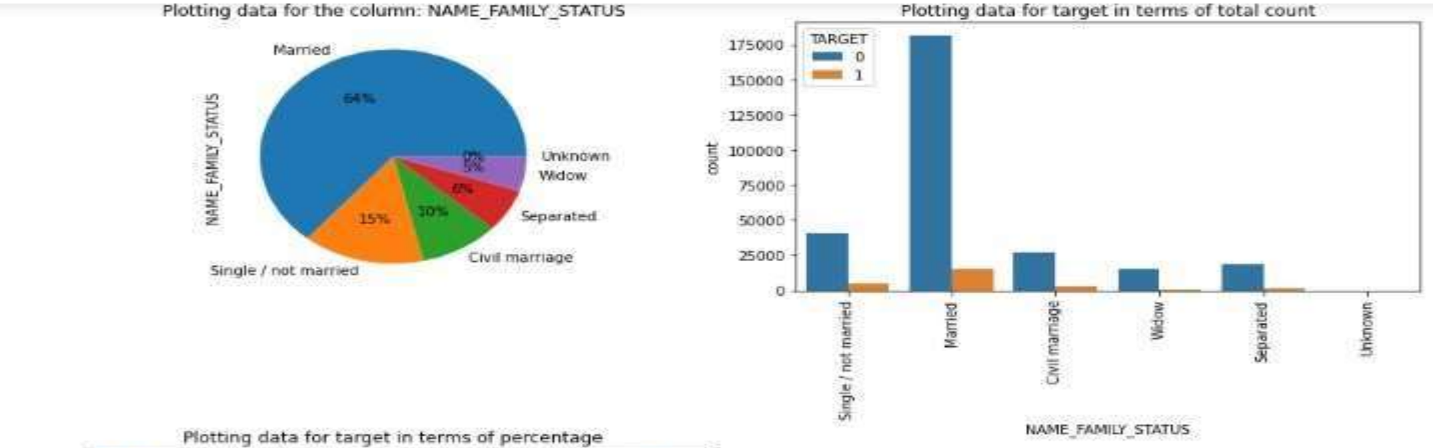


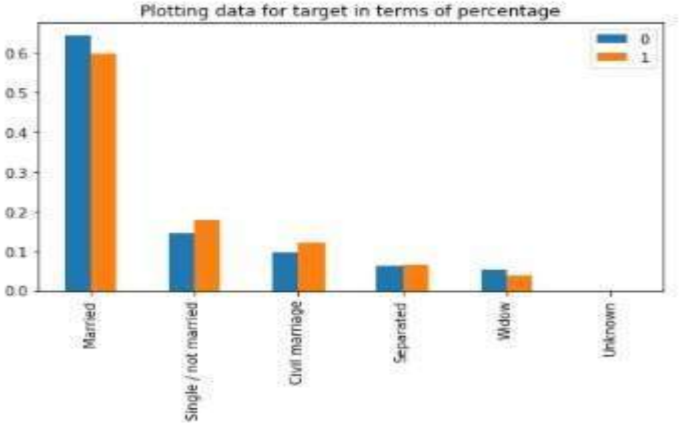


UNIVARIATE ANALYSIS OF THE CATEGORICAL DATA

Unordered Categorical Variables

- NAME_FAMILY_STATUS is a column which gives the Family status of the applicant.As per the pie Graph ,we can infer that 64% of the applicants are Married .Applicants who are single or not Married are Defaulting more than other Categories .Hence We can say it is more Risky to give a Loan to Single/Not Married Applicants.

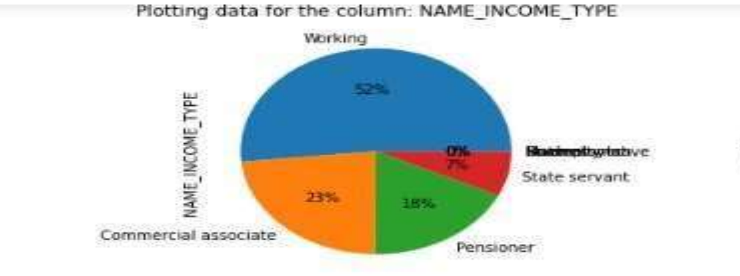


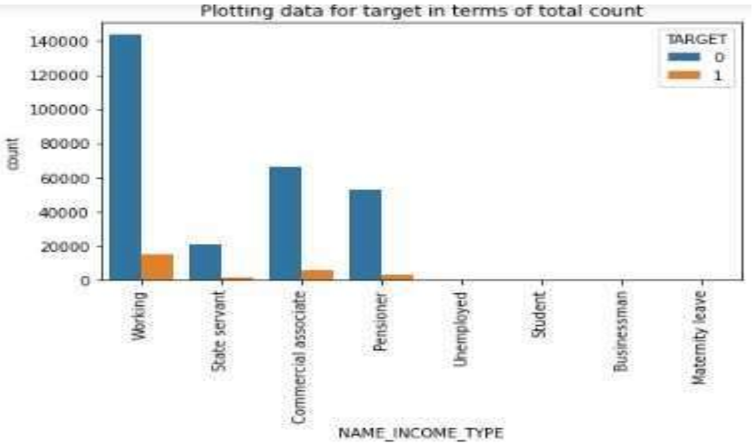


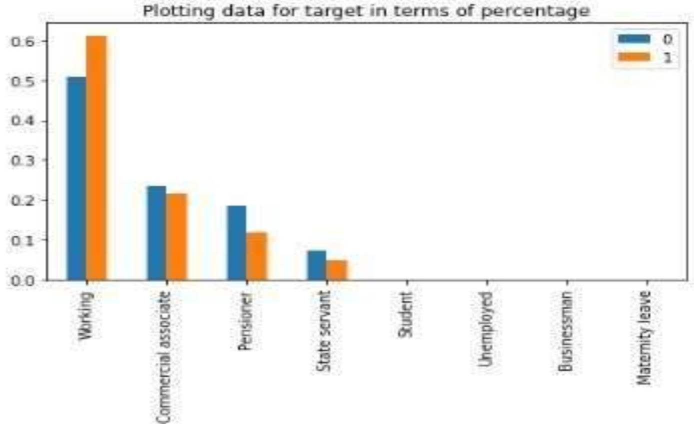
UNIVARIATE ANALYSIS OF THE CATEGORICAL DATA

Unordered Categorical Variables

- NAME_INCOME_TYPE is a column which gives the Income Type of the applicant. As per the pie Graph, we can infer that 52% of the applicants are Working and they are highest defaulters.



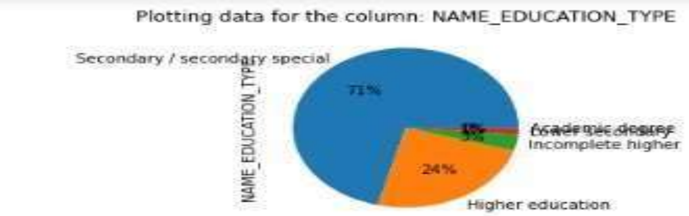


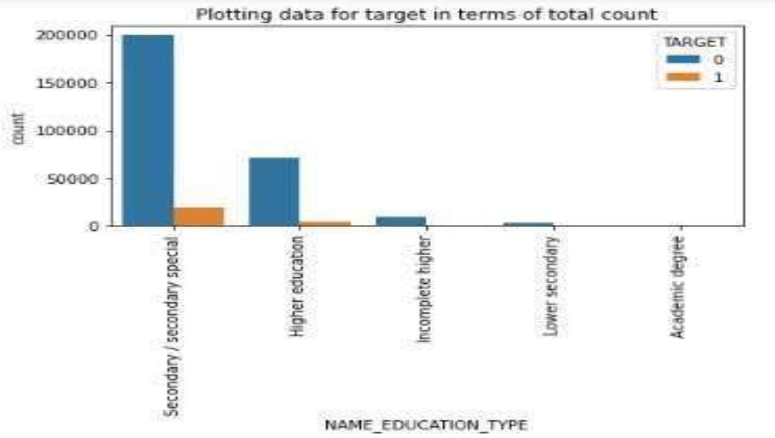


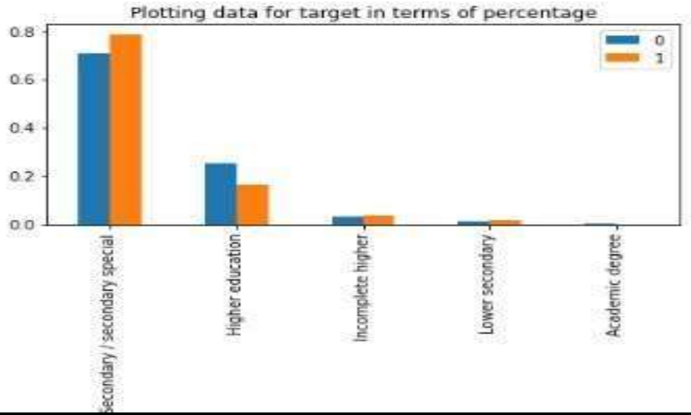
UNIVARIATE ANALYSIS OF THE CATEGORICAL DATA

Ordered Categorical Variables

- NAME_EDUCATION_TYPE is a column which gives the Education Level of the applicant.As per the pie Graph, we can infer that 71% of the applicants are having an education of Secondary/Secondary Special.Only for the Higher Education the default Rate is less. Hence we can say Higher the Education lesser the Default Risk.



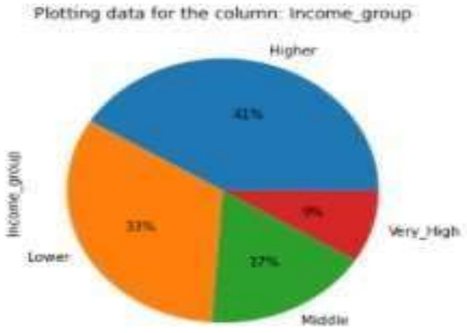


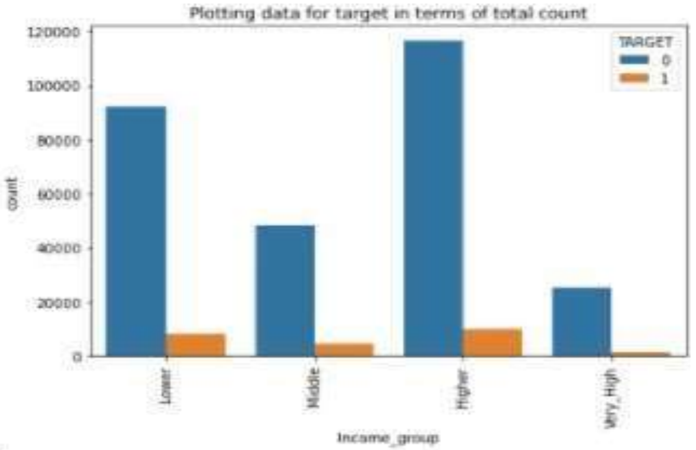


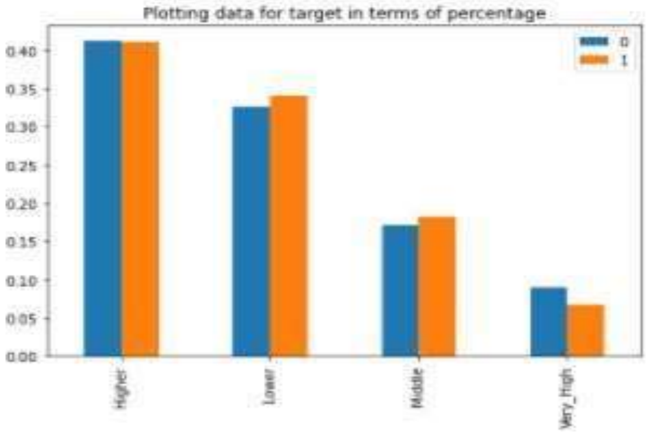
UNIVARIATE ANALYSIS OF THE CATEGORICAL DATA

Ordered Categorical Variables

- 'Income_Group' is the Derived column from AMT_INCOME_TOTAL. As per the pie Chart ,Higher income Group are applying for loans at 41 % .However In Lower income Group ,the percentage of Defaulters are High making them more risky to issue a Loan.

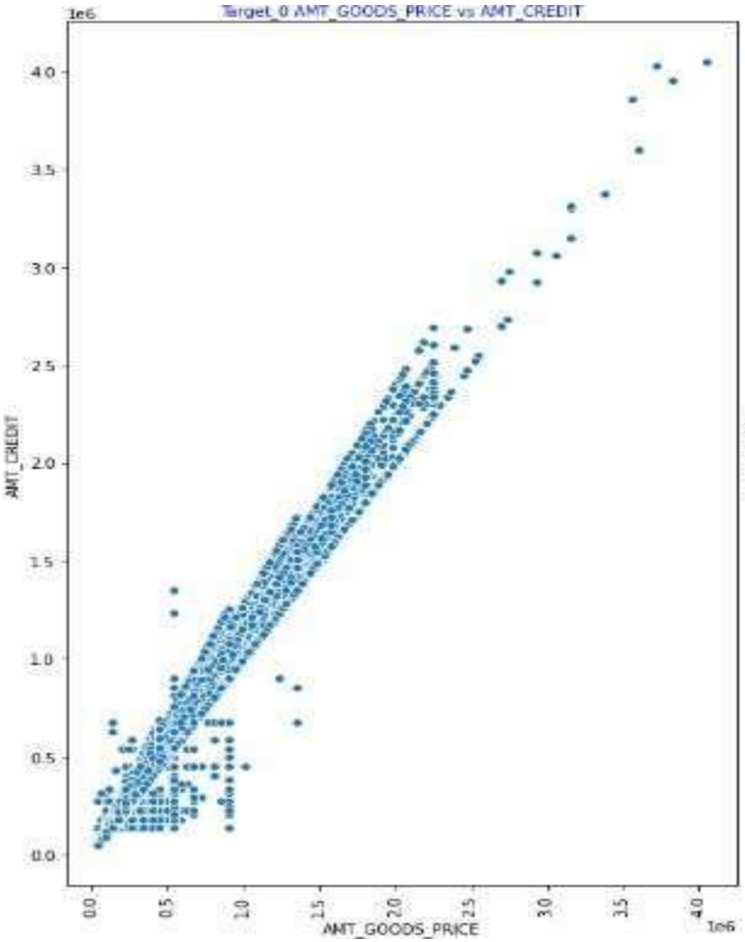


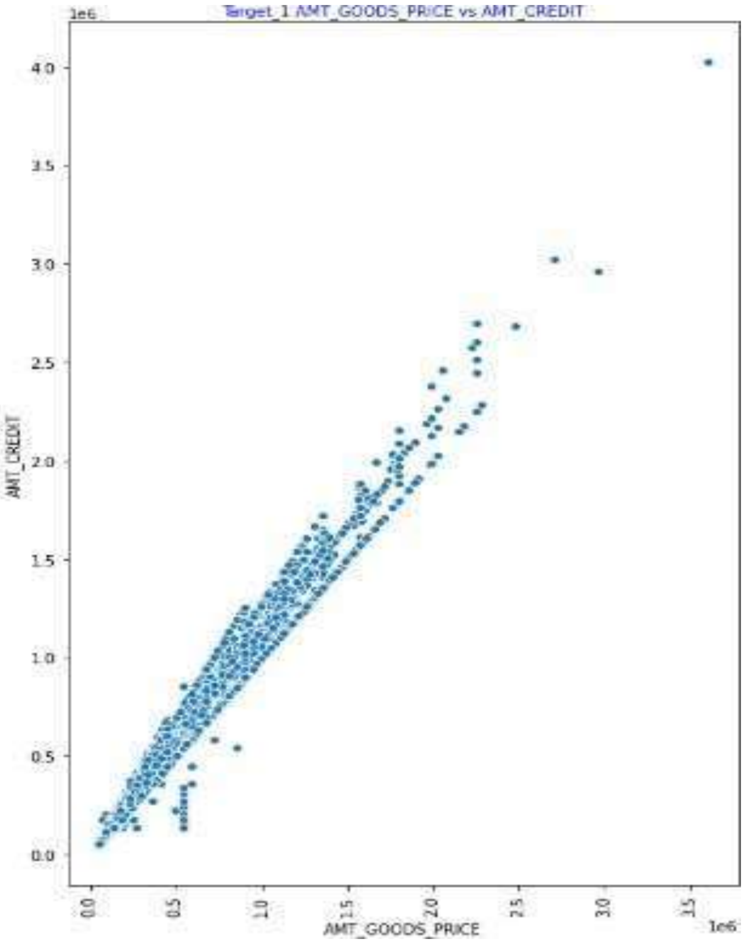




BIVARIATE ANALYSIS OF NUMERICAL VARIABLES FOR TARGET =0 AND TARGET =1

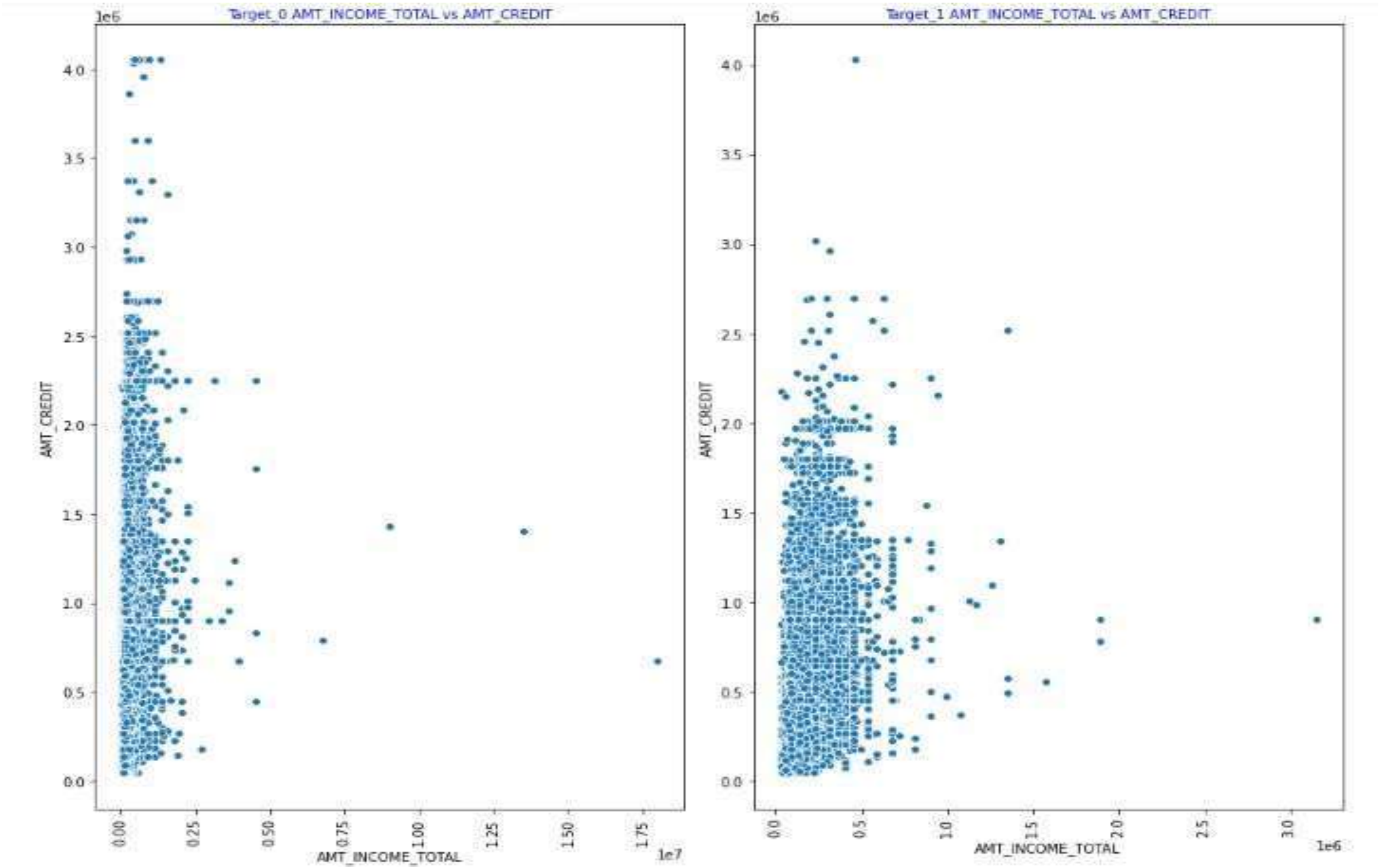
- For an Applicant with Goods Price < 500000 and credit Amount < 500000 is less likely to have Payment Defaults as the Density is Very less





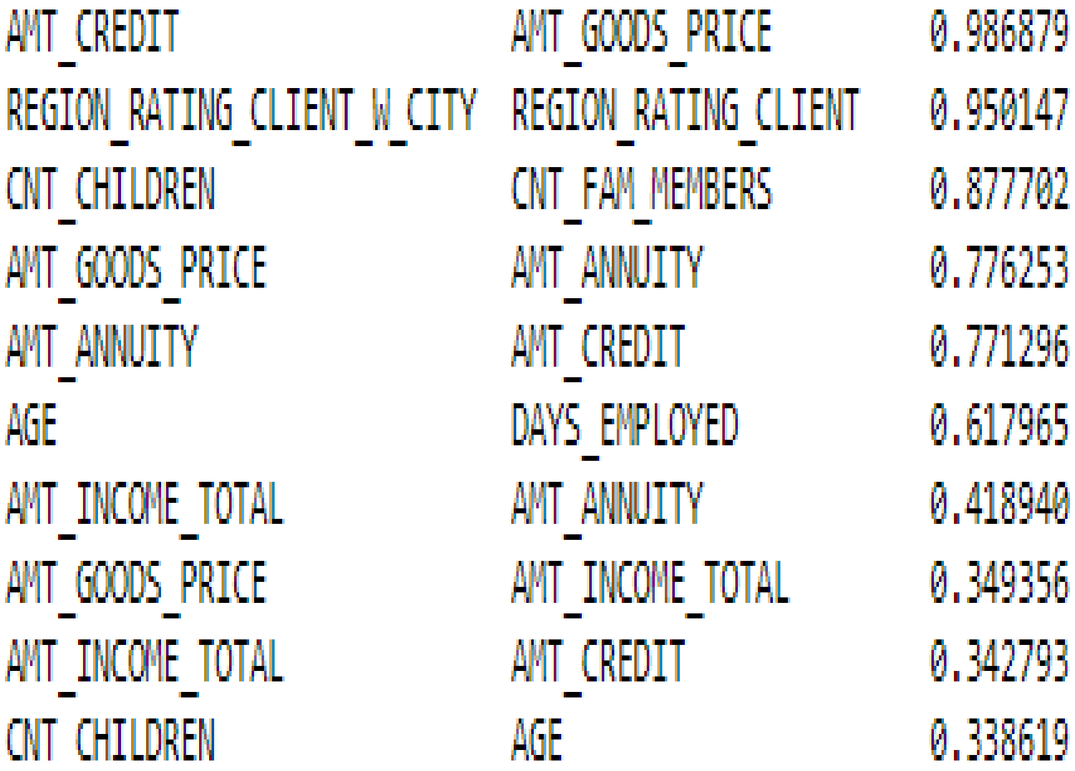
BIVARIATE ANALYSIS OF NUMERICAL VARIABLES FOR TARGET =0 AND TARGET =1

- For an Applicant with Income < 50000 and credit Amount < 150000 is more like to have Payment Defaults as the Density is Very high.



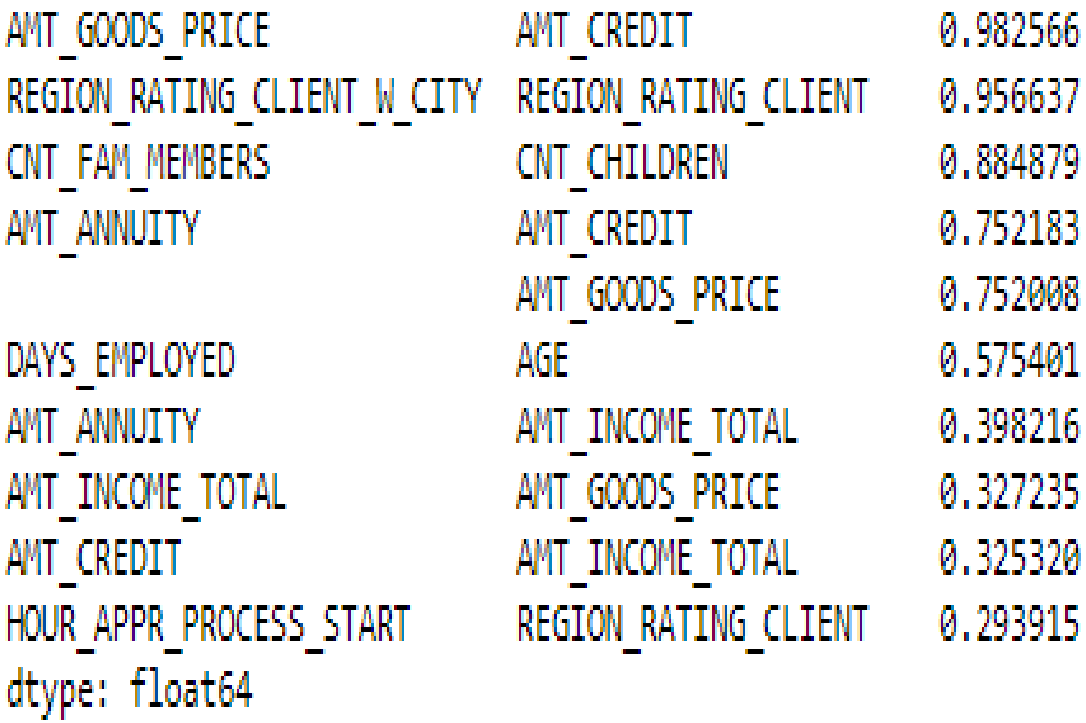
TOP 10 CORRELATION OF NUMERICAL COLUMNS

Target==0 or Non-Defaulters



```
AMT_CREDIT                   AMT_GOODS_PRICE       0.986879
REGION_RATING_CLIENT_W_CITY  REGION_RATING_CLIENT  0.950147
CNT_CHILDREN                 CNT_FAM_MEMBERS       0.877702
AMT_GOODS_PRICE              AMT_ANNUITY           0.776253
AMT_ANNUITY                  AMT_CREDIT            0.771296
AGE                          DAYS_EMPLOYED         0.617965
AMT_INCOME_TOTAL             AMT_ANNUITY           0.418940
AMT_GOODS_PRICE              AMT_INCOME_TOTAL      0.349356
AMT_INCOME_TOTAL             AMT_CREDIT            0.342793
CNT_CHILDREN                 AGE                   0.338619
```

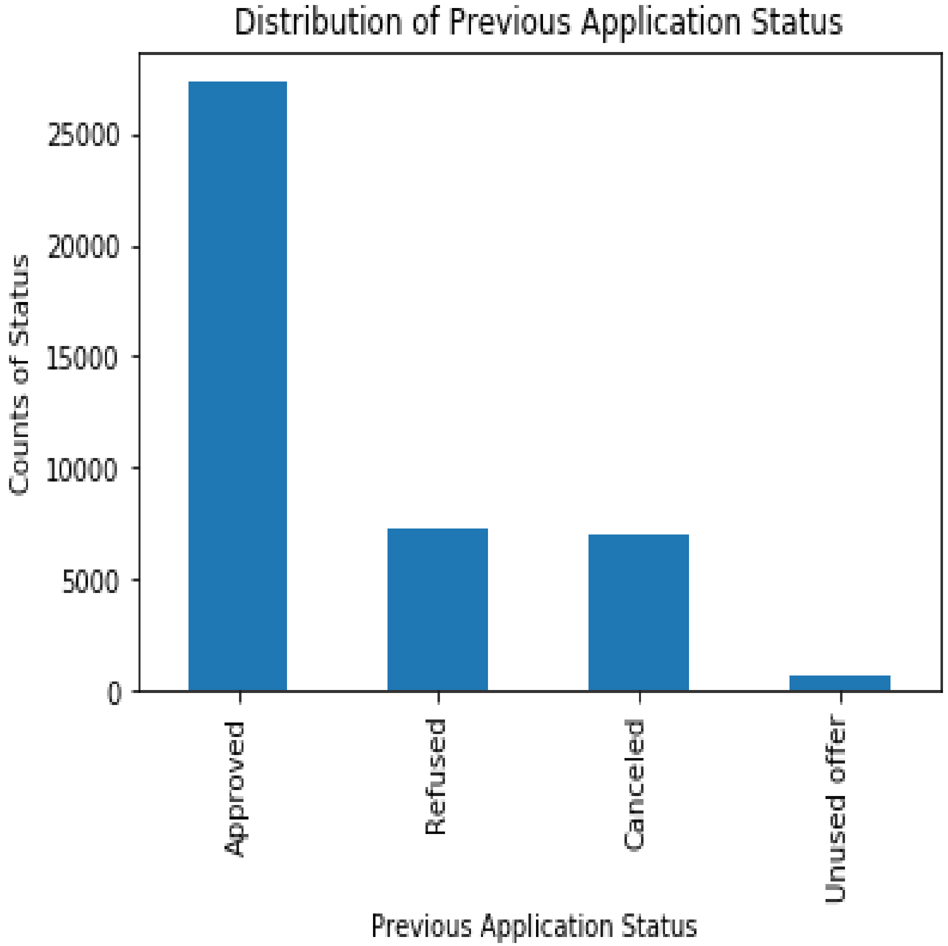
Target==1 or Defaulters



```
AMT_GOODS_PRICE              AMT_CREDIT            0.982566
REGION_RATING_CLIENT_W_CITY  REGION_RATING_CLIENT  0.956637
CNT_FAM_MEMBERS              CNT_CHILDREN          0.884879
AMT_ANNUITY                  AMT_CREDIT            0.752183
                             AMT_GOODS_PRICE       0.752008
DAYS_EMPLOYED                AGE                   0.575401
AMT_ANNUITY                  AMT_INCOME_TOTAL      0.398216
AMT_INCOME_TOTAL             AMT_GOODS_PRICE       0.327235
AMT_CREDIT                   AMT_INCOME_TOTAL      0.325320
HOUR_APPR_PROCESS_START      REGION_RATING_CLIENT  0.293915
dtype: float64
```

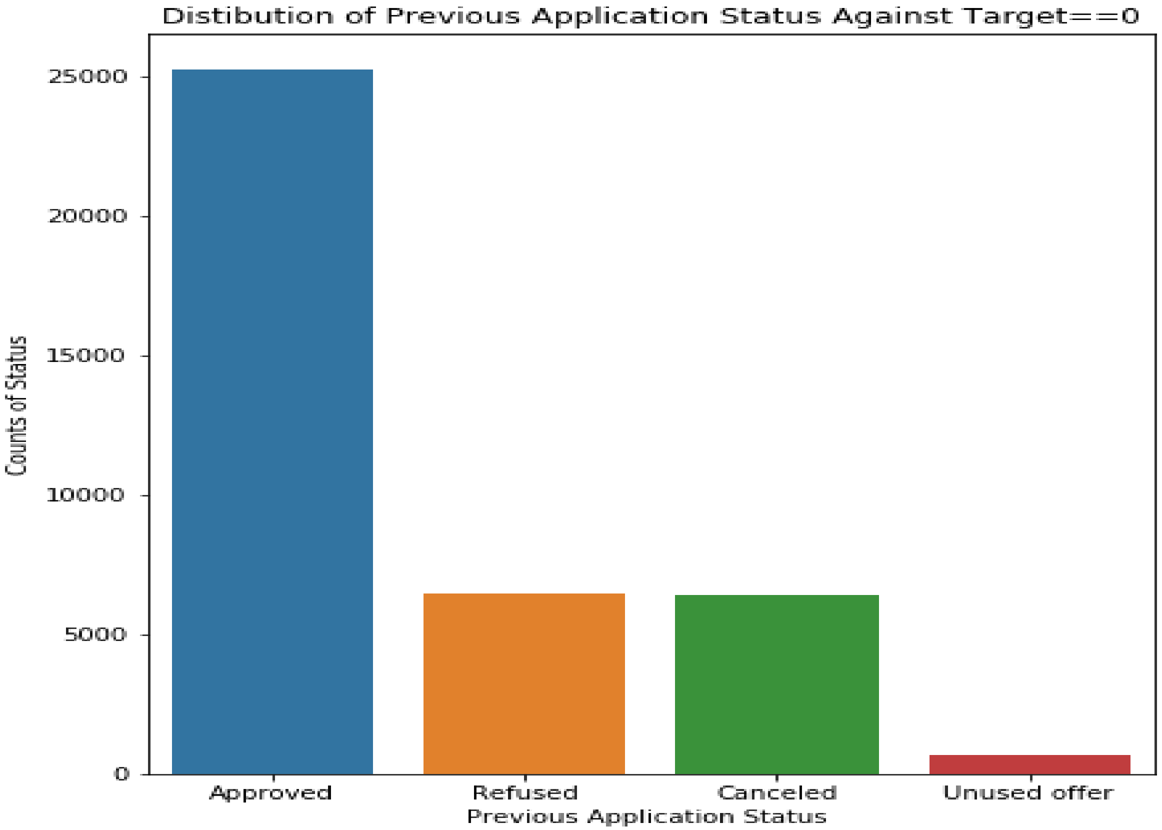
MERGED PREVIOUSAPPLICATION DATAWITHAPPLICATION DATA

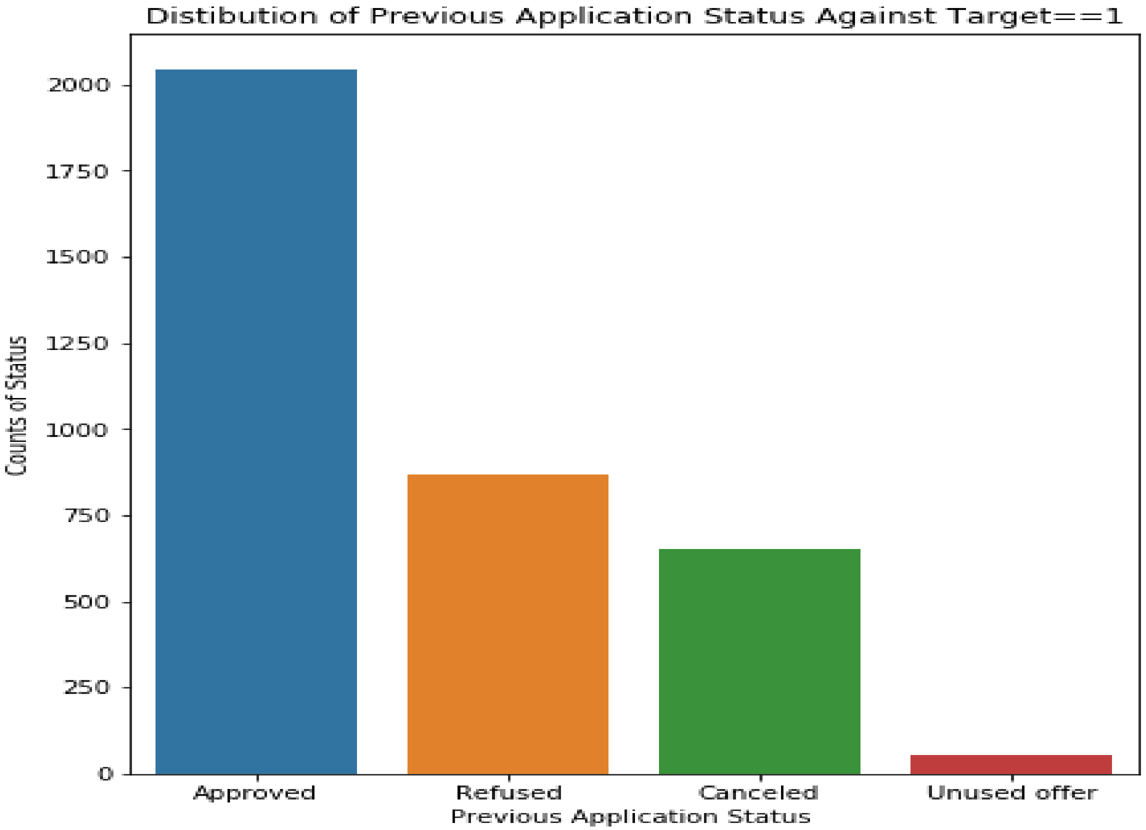
- •Previous Application Data has lot of records , So we have taken only 50,000 records for merging with the Application Data
- •There are D uplicate Values for SK_ID_CURR as a person could have taken loan Multiple Times
- •Missing Values were handled for AMT_GOODS_PRICE_y, CNT_PAYMENT, AMT_ANNUITY_y ,PRODUCT_COMBINATION columns
- •Merged D ata can be D ivided in to 4 columns depending on Previous Application Status
- •Approved, Rejected,Cancelled,Unused Offer



IMPACT OF PREVIOUS APPLICATION STATUS

TheApplicants whose PreviousApplication Status is Refused or Cancelled have high chance of DefaultingAgain





OVERALL INFERENCES OF CREDIT EDA CASE STUDY

- In Lower income Group ,the percentage of Defaulters are High making them more risky to issue a Loan

- Higher the Education lesser the Risk of Defaulting the Loan

- MaleApplicants default more than Females.This puts giving loans to females is safer

- Applicants who are single or not Married are more likely to Default a Loan.

- Applicants of Working class are highest defaulters .

- Age Group of 20-40 have more people as Defaulters.

- TheApplicants whose PreviousApplication Status is Refused or Cancelled have high chance of Defaulting again.