EDA ON LOAN PROVIDING COMPANY

By:- Zeeshan Maindargi.

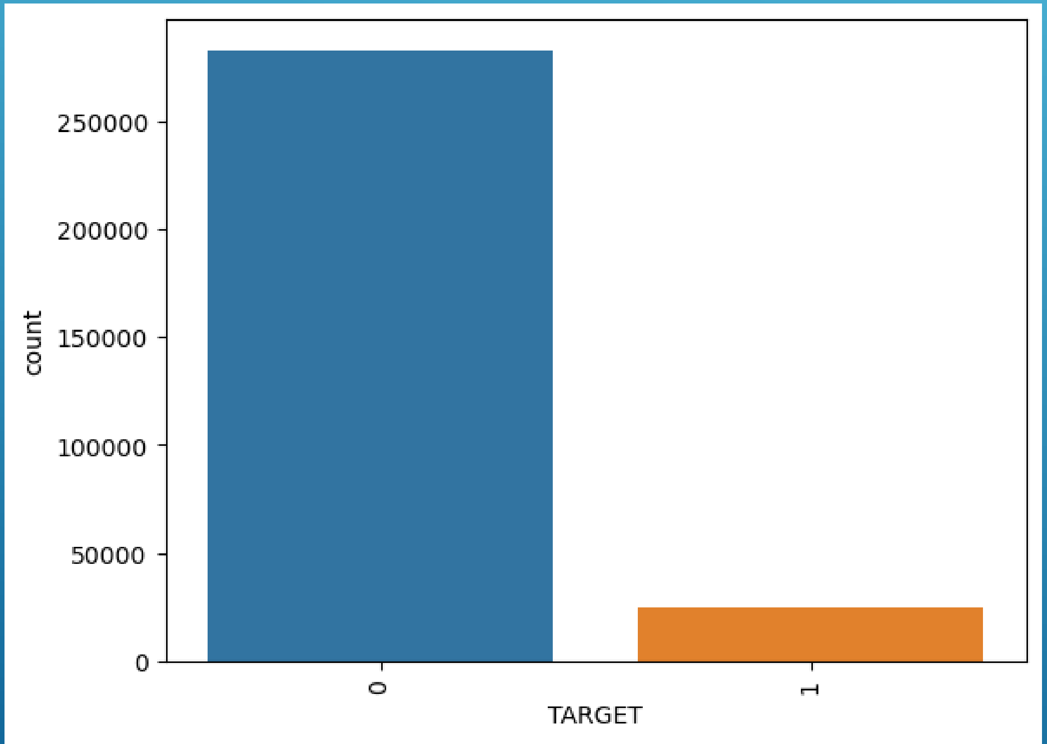
PROBLEM STATEMENT

- The data given is form a loan providing company.
- The company receives applications from clients. All the applications have be analysed.
- All the applications can’t be accepted, as it can be a risk to company if the client doesn’t repay the loan.
- All application can’t be rejected, because losing good clients will reflect on financial loses to the company.
- Hence analyzing the data to ensure that capable client are not rejected.

MISSING VALUES AND OUTLINERS IN DATA

- The missing values in the application data with more than 40% threshold are dropped.
- The column with less than 40%threshold values are imputed using statistical methods.
- Mode for categorical and median for continuous columns.
- Outlines/ anomalies: there are outliners in the data, in columns such as count of children, income, region population, etc.
- The outliners are detected using boxplot and quantiles .

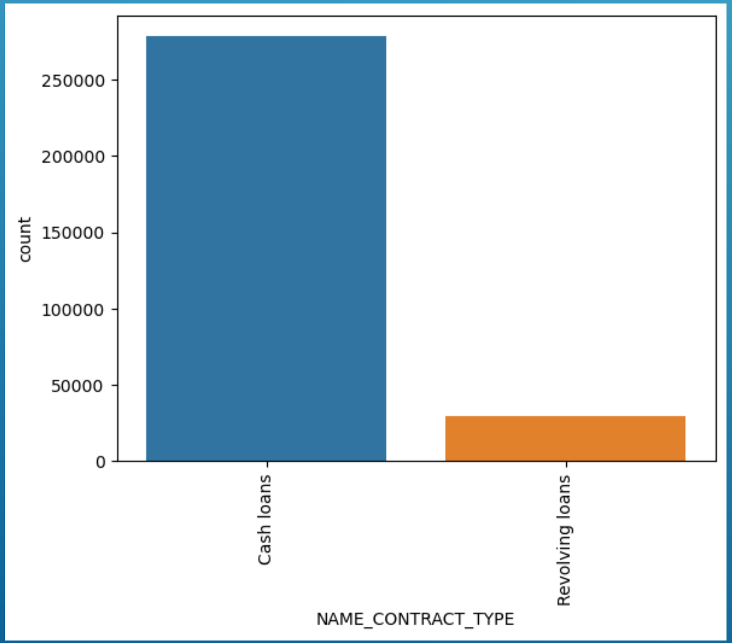
UNIVARIATE ANALYSIS OF CATEGORICAL VARIABLES ~ USING COUNTPLOT



- Here, 0- clients with no payment difficulties. 1-clients with payment difficulties. The count of clients with no payment difficulties is much higher than clients with payment difficulties.

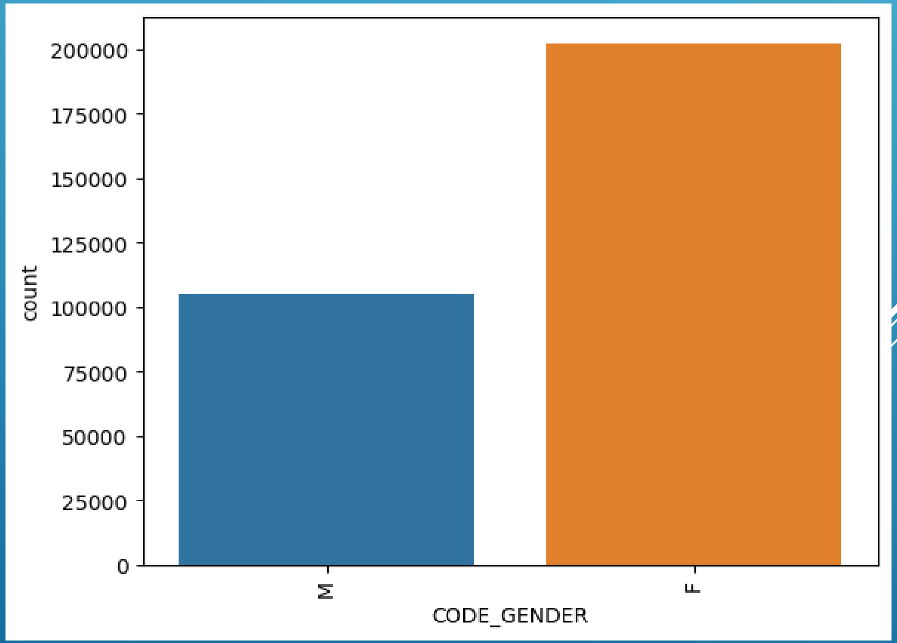
NAME_CONTRACT_TYPE

NAME_CONTRACT_TYPE: A COLUMN STATING TYPE LOAN CLIENTS ARE OPTING. FROM THE DATA AND GRAPH IT CAN BE SEEN THAT CASH LOANS ARE MUCH HIGHLY PREFERRED.



CODE_GENDER

CODE GENDER: M=MALE, F=FEMALE THE NUMBER OF FEMALES OPTING FOR LOANS IS MORE THAN MALES. THIS MAYBE DUE TO SUBSIDIES PROVIDED FOR WOMEN.



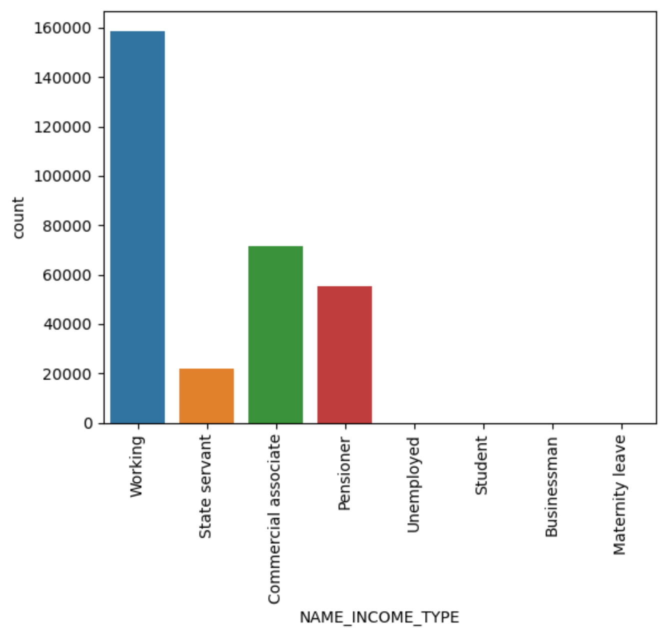
NAME_INCOME_TYPE & OCCUPATION_TYPE

NAME_INCOME_TYPE:

CLIENTS FROM WORKING CLASS OPT FOR

MORE LOANS. FOLLOWED BY COMMERCIAL ,

PENSIONER THEN STATE SERVANTS.

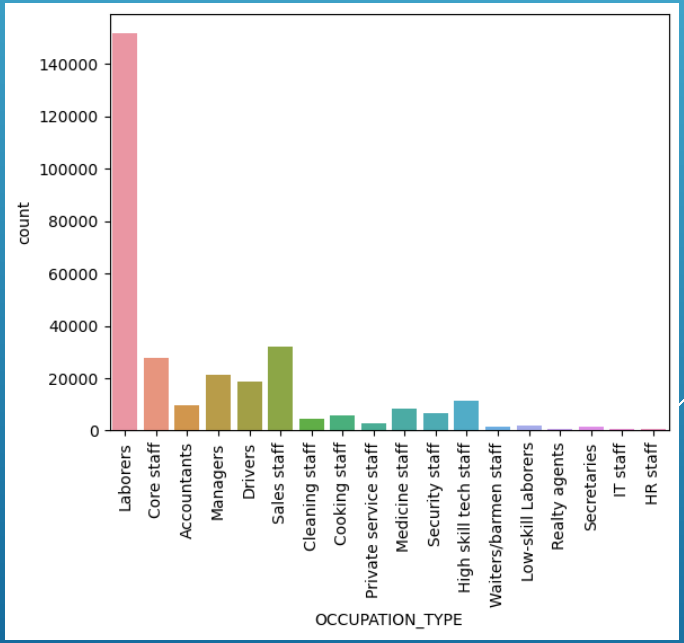


OCCUPATION_TYPE :

MAJORITY OF LABORERS FOLLOWED BY SALES

STAFF, CORE STAFF, MANAGERS, DRIVERS, AND

SO ON GO FOR LOAN APPLICATIONS THE MOST.

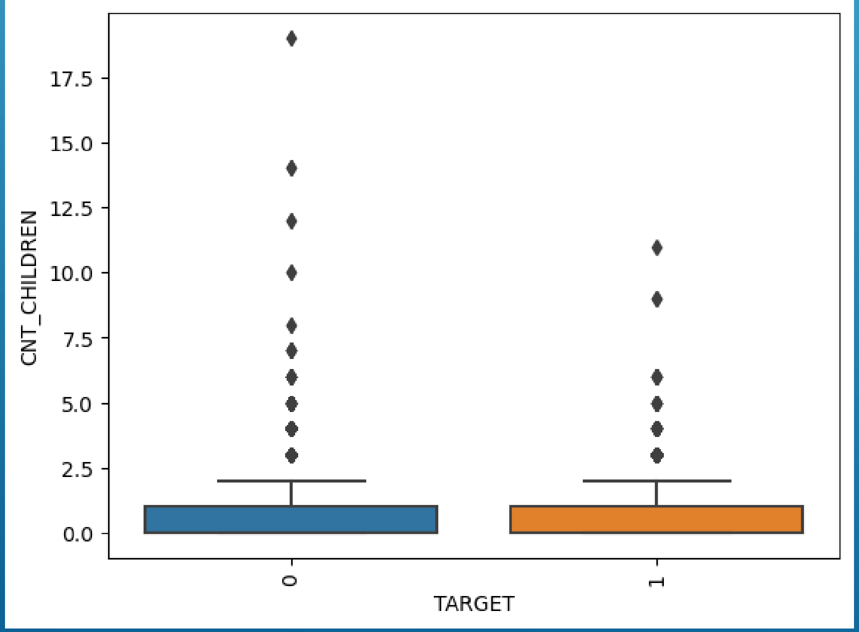


BIVARIATE ANALYSIS OF TARGET WITH CONTINUOUS VARIABLES.

CNT_CHILDREN :

CLIENTS WITH MORE NUMBER OF CHILDREN ARE SEEN TO

BE ONES PAYING BACK THE LOAN.

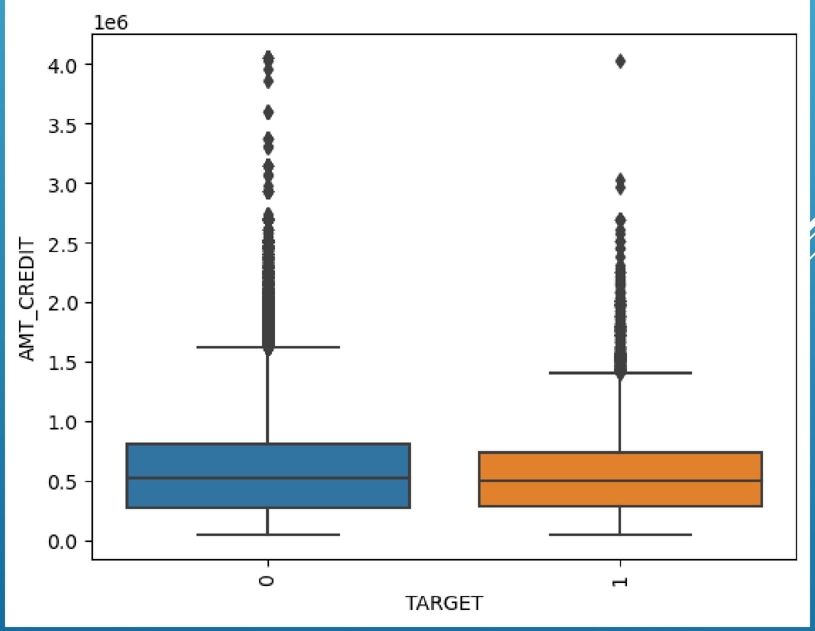


AMT_CREDIT:

CLIENTS WITH NO PAYMENT DIFFICULTIES ARE CLIENTS WITH HIGH

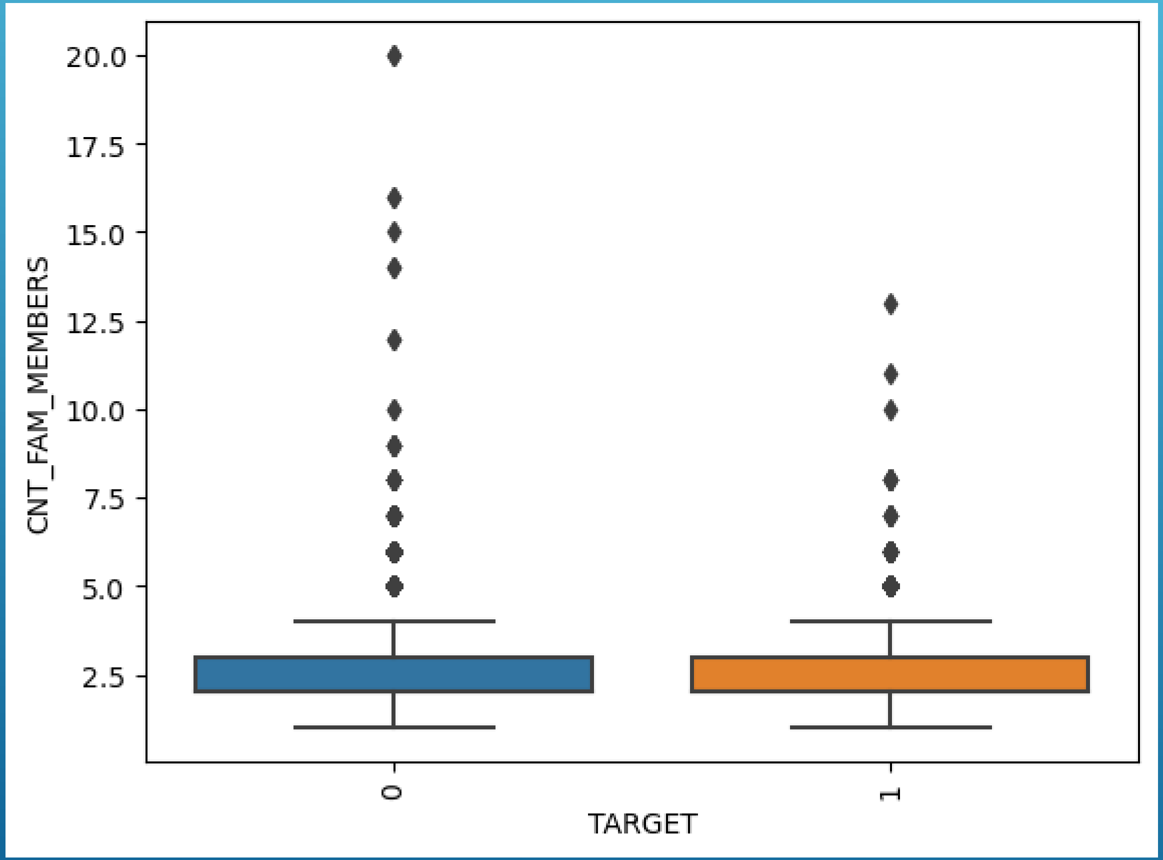
CREDIT AMOUNTS. THOUGH THE MEDIANS ARE VERY CLOSE FOR

BOTH THE CONTINUOUS OUTLIER NUMBER SEEMS TO BE HIGH IN 0.



CNT_FAM_MEMBERS:

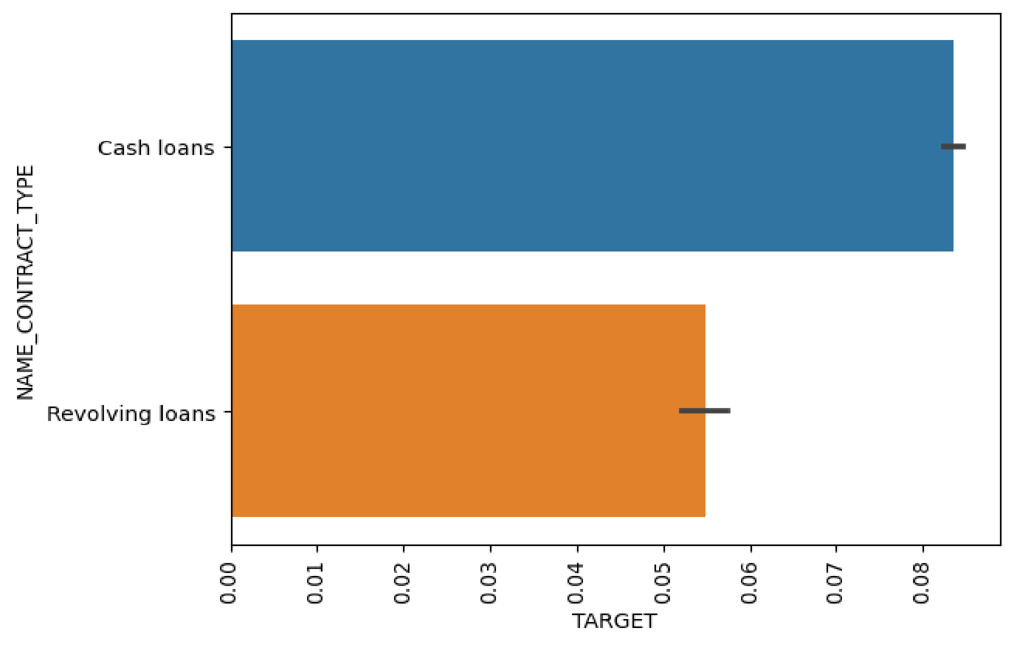
THE CUSTOMERS WITH MORE FAMILY MEMBERS ARE SEEN TO BE PAYING ON TIME. IT CAN BE RELATIVE TO AN ASSUMPTION THAT THE FAMILY MEMBERS WILL HAVE MORE EARNING HANDS.

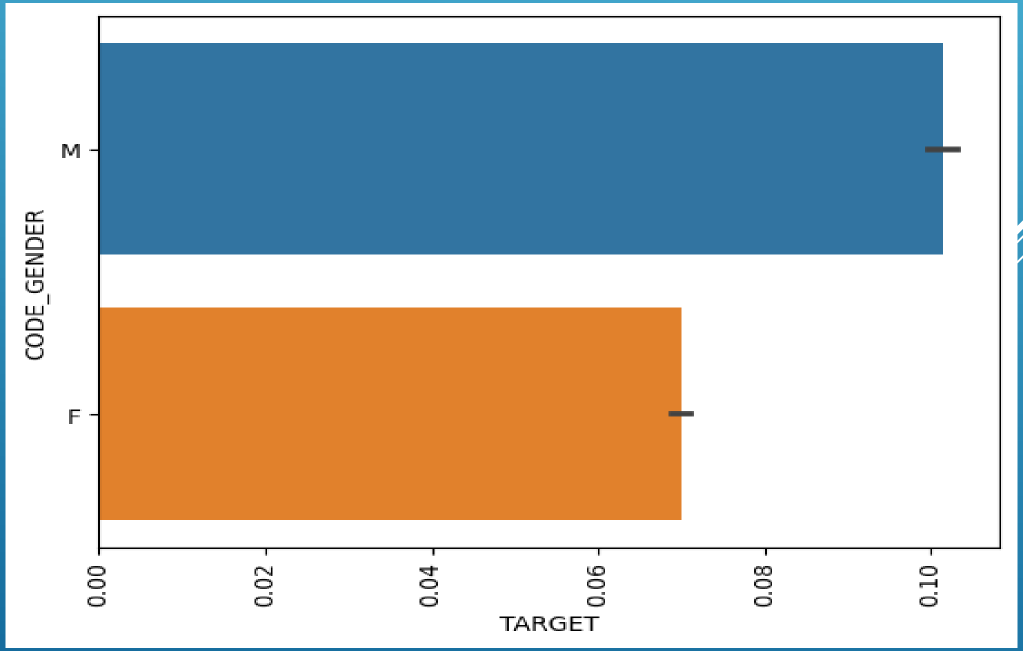


Bivariate analysis of target with categorical variables

NAME_CONTRACT TYPE:
CASH LOANS ARE MORE PREFERRED THAN REVOLVING
LOANS BY CLIENTS WITH GOOD PAYMENT CAPABILITIES

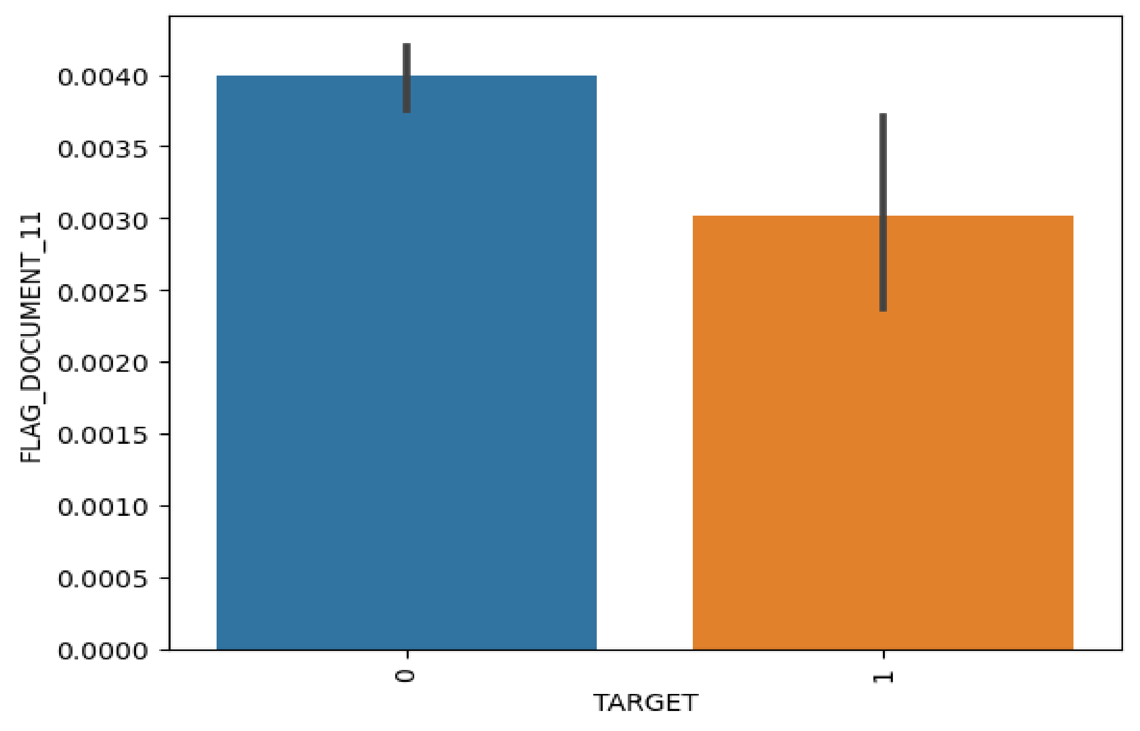
CODE_GENDER:
MALE CLIENTS ARE MORE DILIGENT IN PAYING ON
TIME THAN FEMALE CLIENTS.

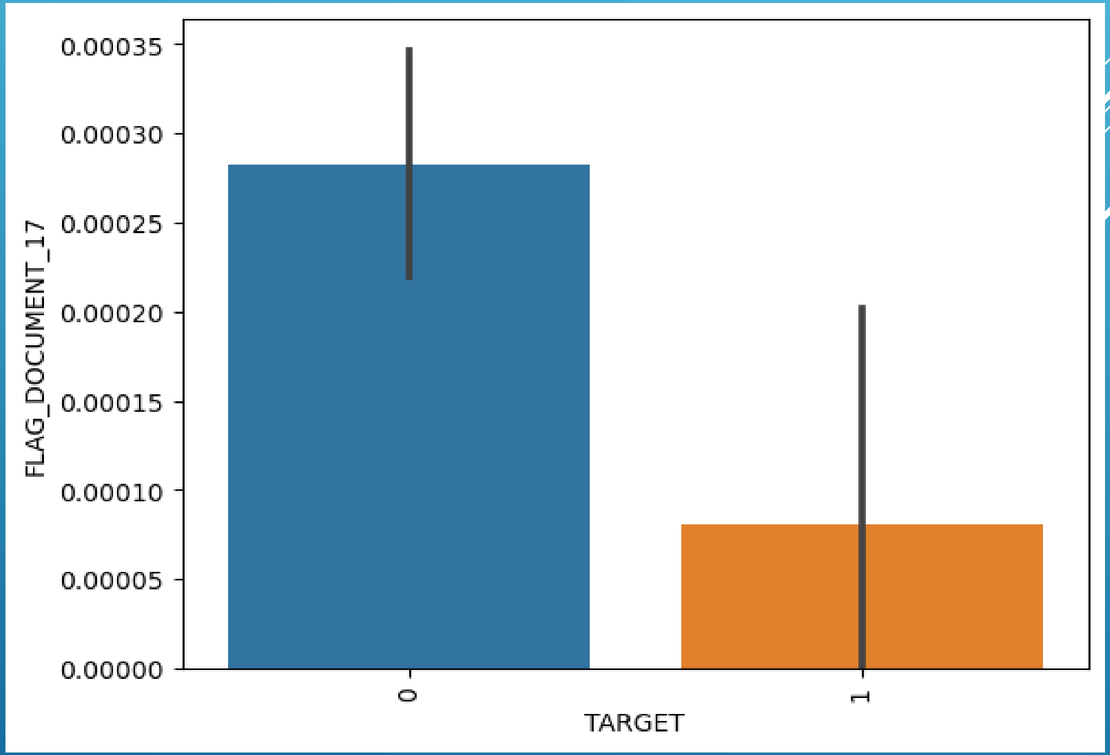




FLAG_DOCUMENT:

CLIENTS WHO HAVE NOT SUBMITTED MOST OF THE DOCUMENTS ARE THE ONES WITH MORE PAYMENT DIFFICULTIES. THIS PATTERN CAN BE SEEN FOR MOST OF THE DOCUMENTS.





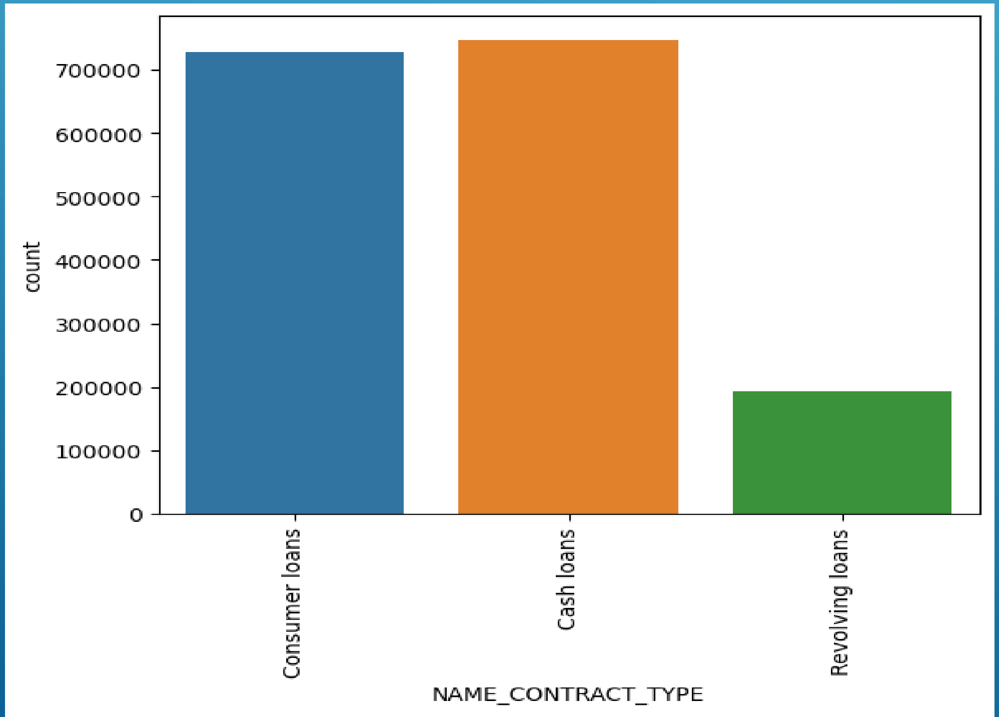
EDA ON PREVIOUS_APPLICATION DATASET

- Certain data cleaning steps had to be taken in this dataset.
- Some columns had values such as XNA, XAP which are assumably missing values.
- So replaced them as missing value so that the statistical measures output is not hampered.
- Statistical measures simply ignore the missing values and compute on the rest data.
- Outliers check was done using boxplot.
- Columns with missing values more than 40% threshold were dropped.

Univariate analysis

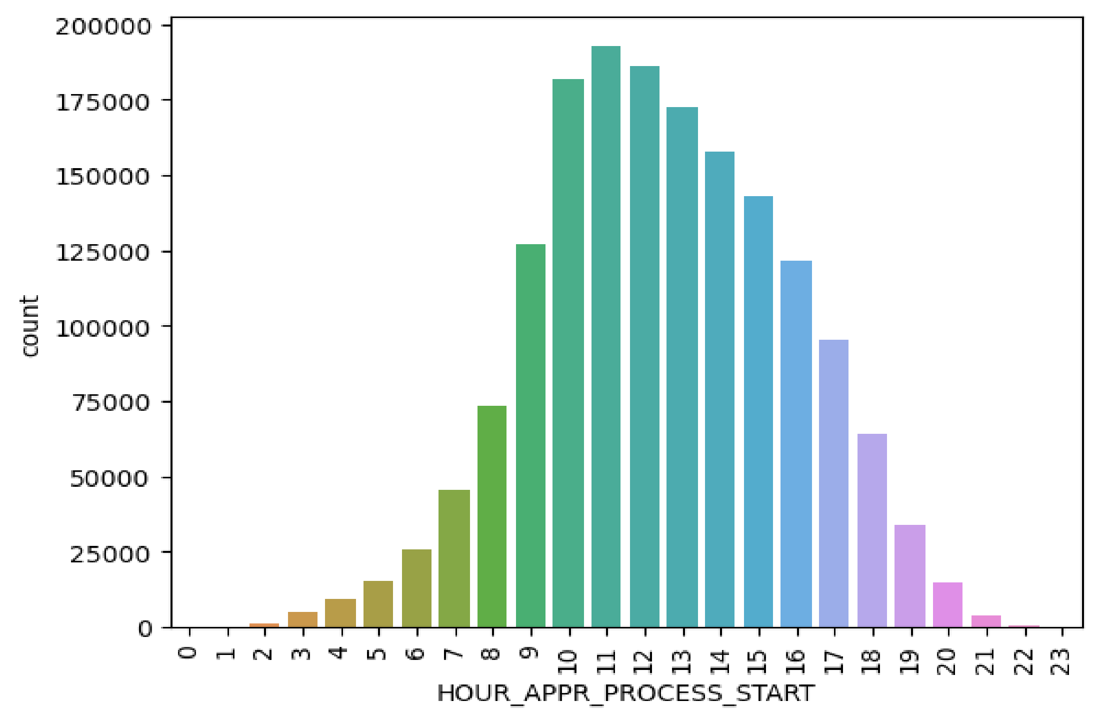
NAME_CONTRACT TYPE:

THE COUNT OF CASH LOANS IS THE MOST FOLLOWED BY CONSUMER LOANS. REVOLVING LOANS ARE MUCH LESS PREFERRED.



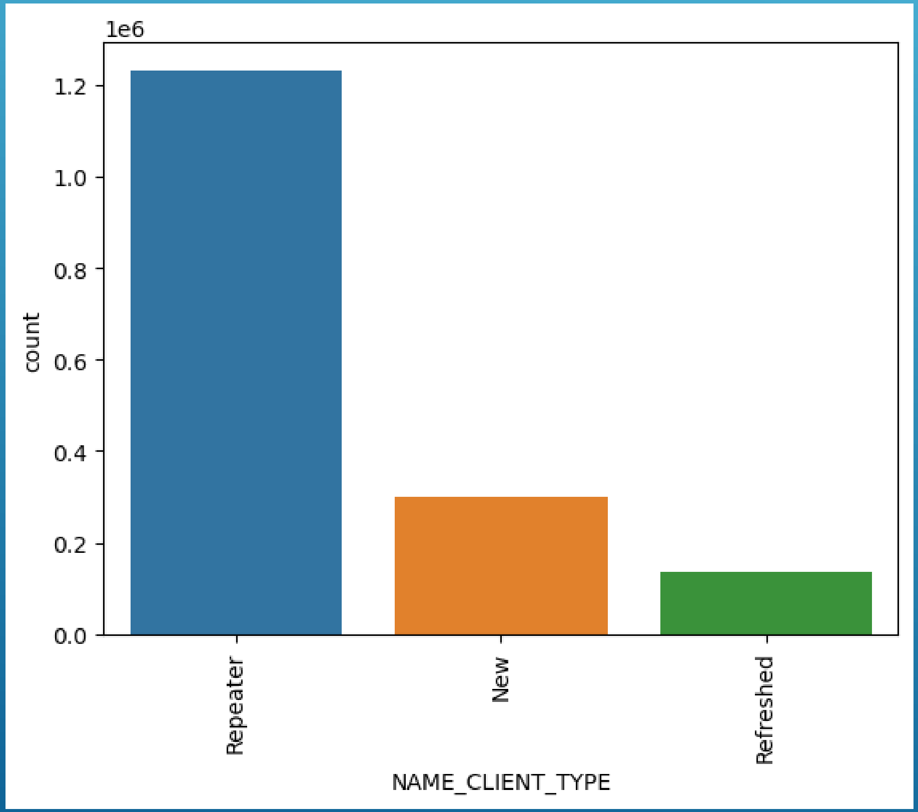
HOUR_APPR_PROCESS_START:

APPROXIMATELY THE COUNT OF APPLICATIONS DONE AT 11AM IS THE HIGHEST FOLLOWED BY 10, 12 AND SO ON. MOSTLY AFTER NOON HAS MORE FREQUENCY



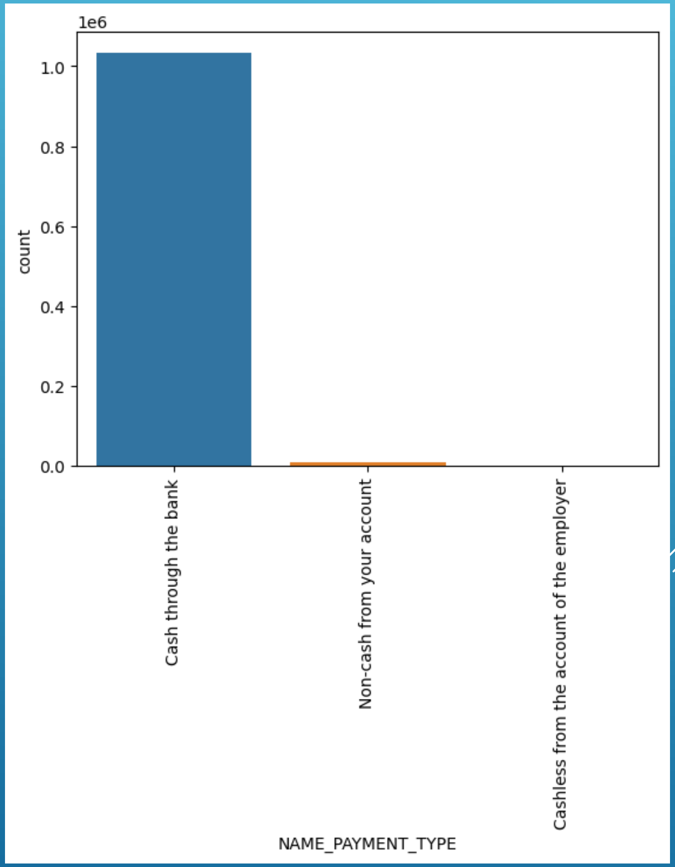
NAME_CLIENT_TYPE:

REPEATER CLIENTS MORE LIKELY TO APPLY FOR LOANS.



NAME_PAYMENT_TYPE:

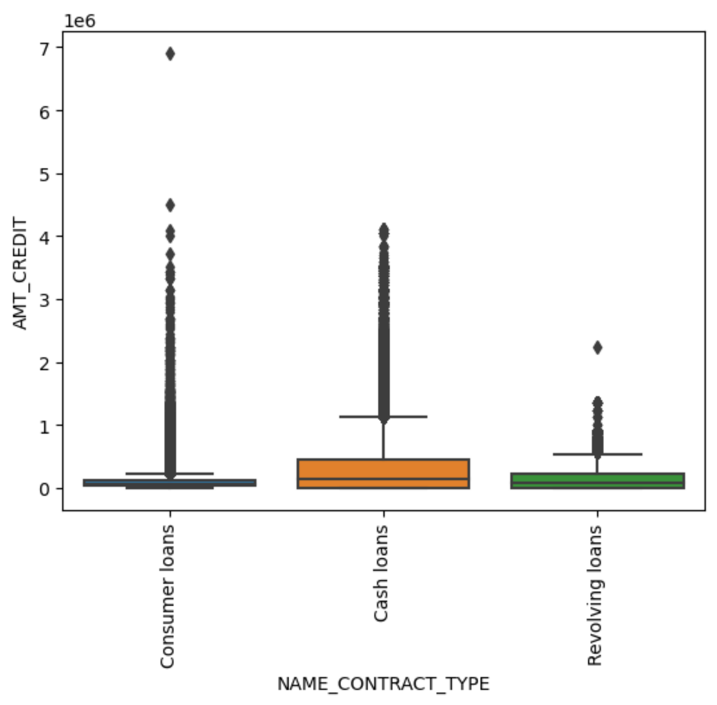
CLIENTS WITH PAYMENT TYPE OF CASH THROUGH BANK ARE THE MOST . NON CASH AND CASHLESS ARE ALMOST NEGLIBLE.



Bivariate analysis with name contract type and other columns

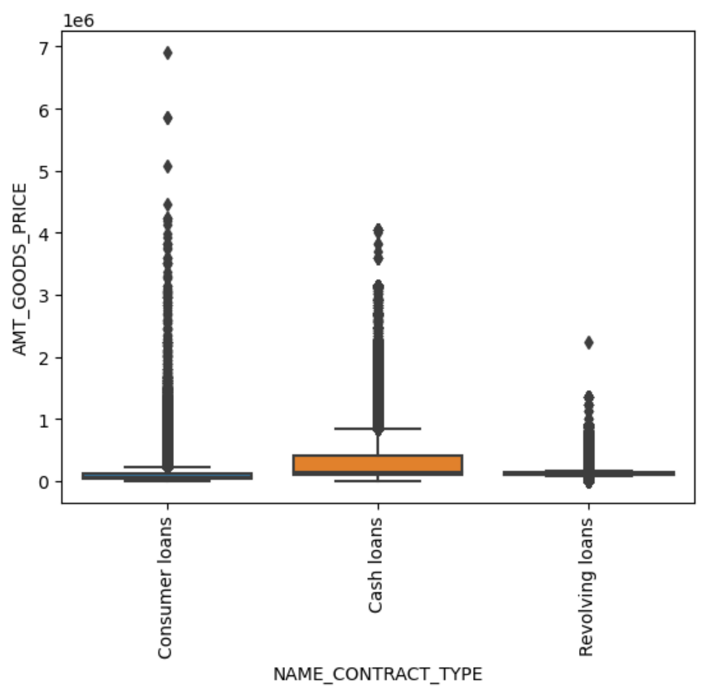
AMT_ANNUITY:

CASH LOANS SEEM TO HAVE MOST ANNUITY FROM PREVIOUS APPLICATION.



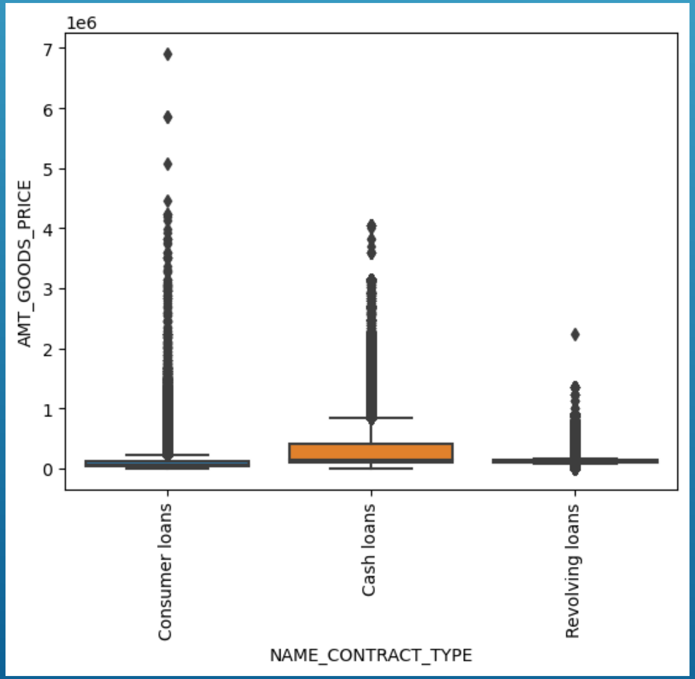
AMT_APPLICATION:

THE AMOUNT OF CREDIT CLIENT ASKED MOST FOR LIES IN CASH LOAN TYPE OF CATEGORY.



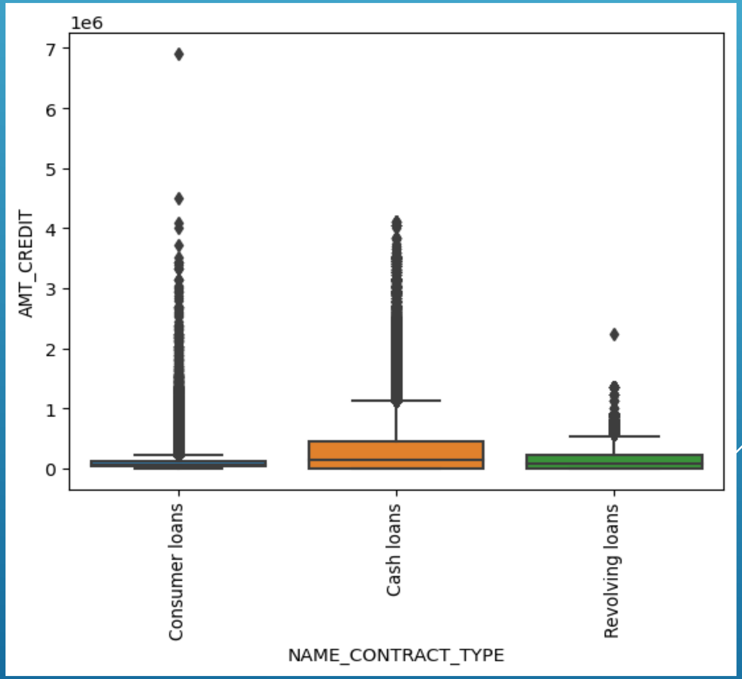
AMT_CREDIT:

THE AMOUNT FINALLY CREDITED TO THE CUSTOMER
FROM WHAT HE/SHE ASKED FROM AMT_APPLICATION
IS AMOUNT CREDIT. THE MAJORITY OF IT IS IN CASH
TYPE OF LOANS.



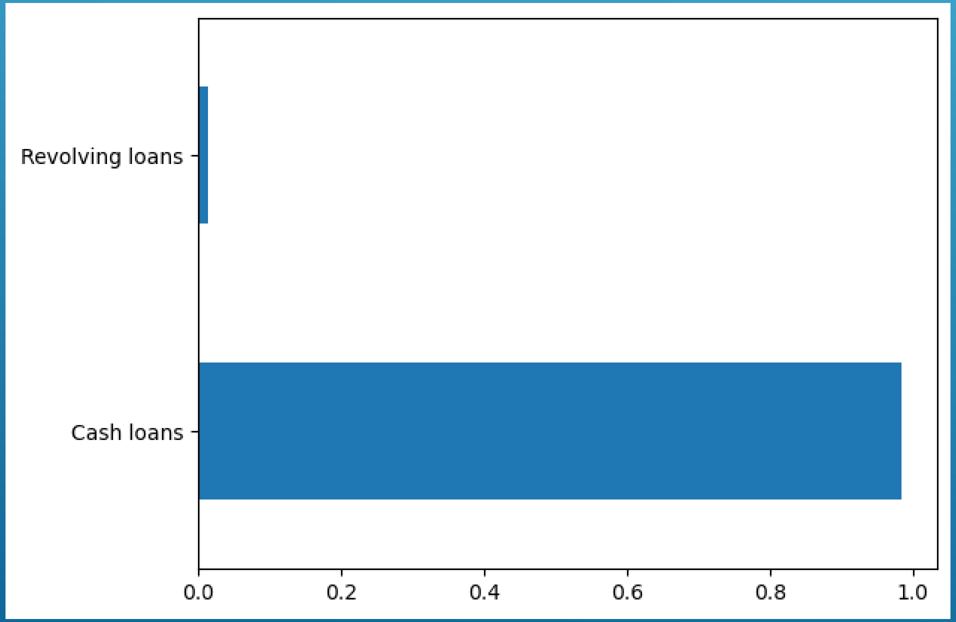
DAYS_DECISION:

THE MORE EARLY APPLICATIONS COME UNDER CASH LOANS, THEN REVOLVING, CONSUMER. THOUGH MORE MASS IS IN CONSUMER, EARLY ONES COME UNDER CASH.

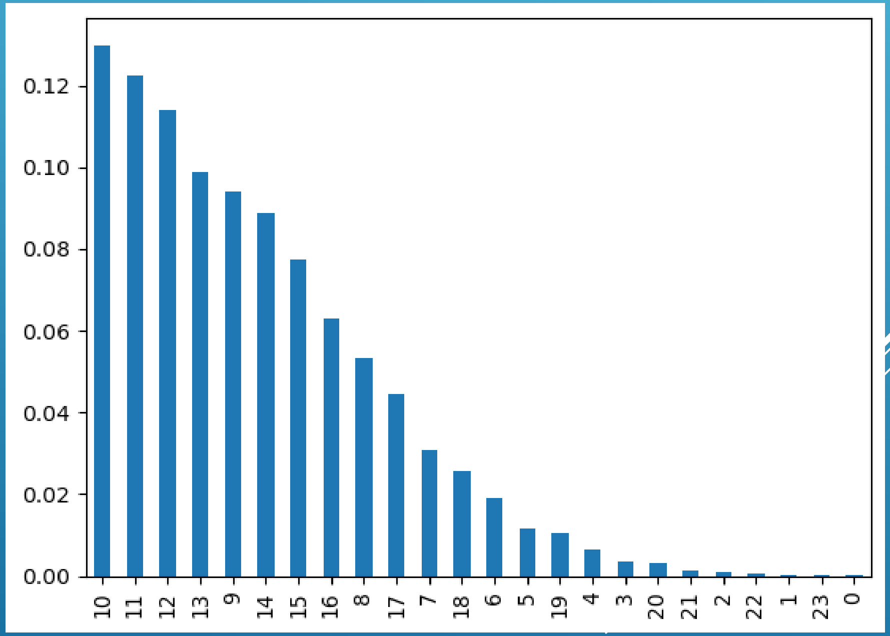


Merged data

NAME_CONTRACT_TYPE
CASH LOANS ARE MORE TARGETED THAN REVOLVING LOANS



HOUR_APPR_PROCESS_START
MOST APPLICATIONS DONE AT 10 AM ARE MORE IN NUMBER.

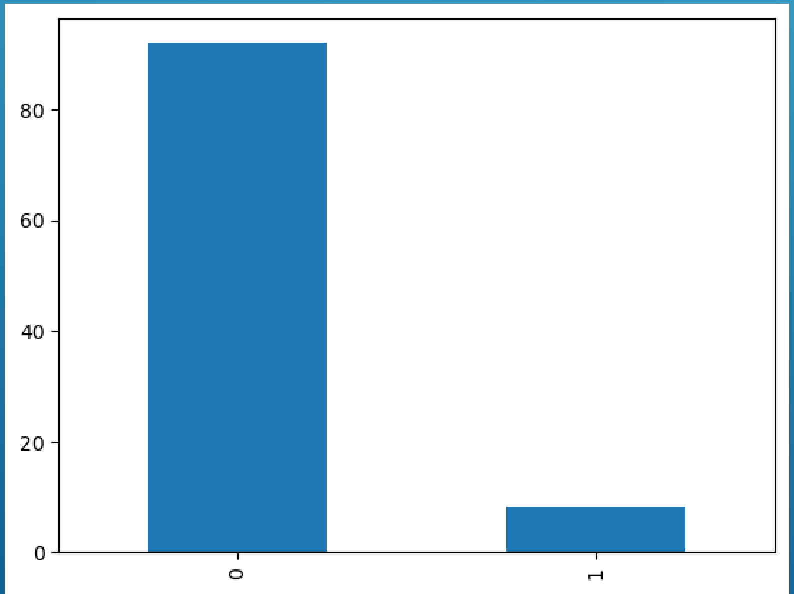


BIVARIATE ANALYSIS

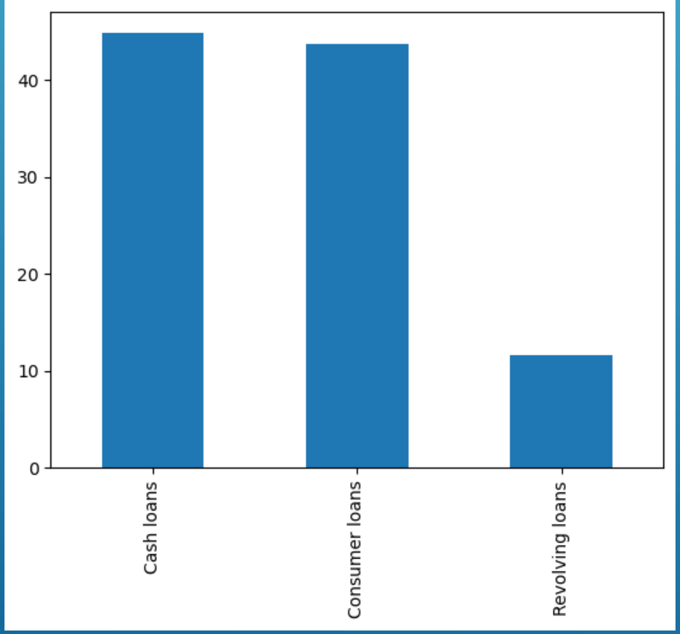
BIVARIATE ANALYSIS OF MERGED DATA IS DONE USING STATISTICAL MEASURES AND QUANTILES.

Data imbalance

DATA IMBALANCE IN APPLICATION DATASET. DATA IMBALANCE OF 92%-8% IS SEEN.



DATA IMBALANCE IN PREVIOUS_APPLICATION DATASET. DATA IMBALANCE OF 44%-43%-11% IS SEEN. THE DATA IS MOSTLY BALANCED IN THIS CASE.


Cash loans
Consumer loans
Revolving loans