CREDIT EDA CASE STUDY ASSIGNMENT

BY GARIMA RAI

INTRODUCTION

In this EDA assignment we are provided with the two types of data sets one is an application data which contains all the information of the client at the time of application, on the other hand the other data is the previous application that has client previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused.

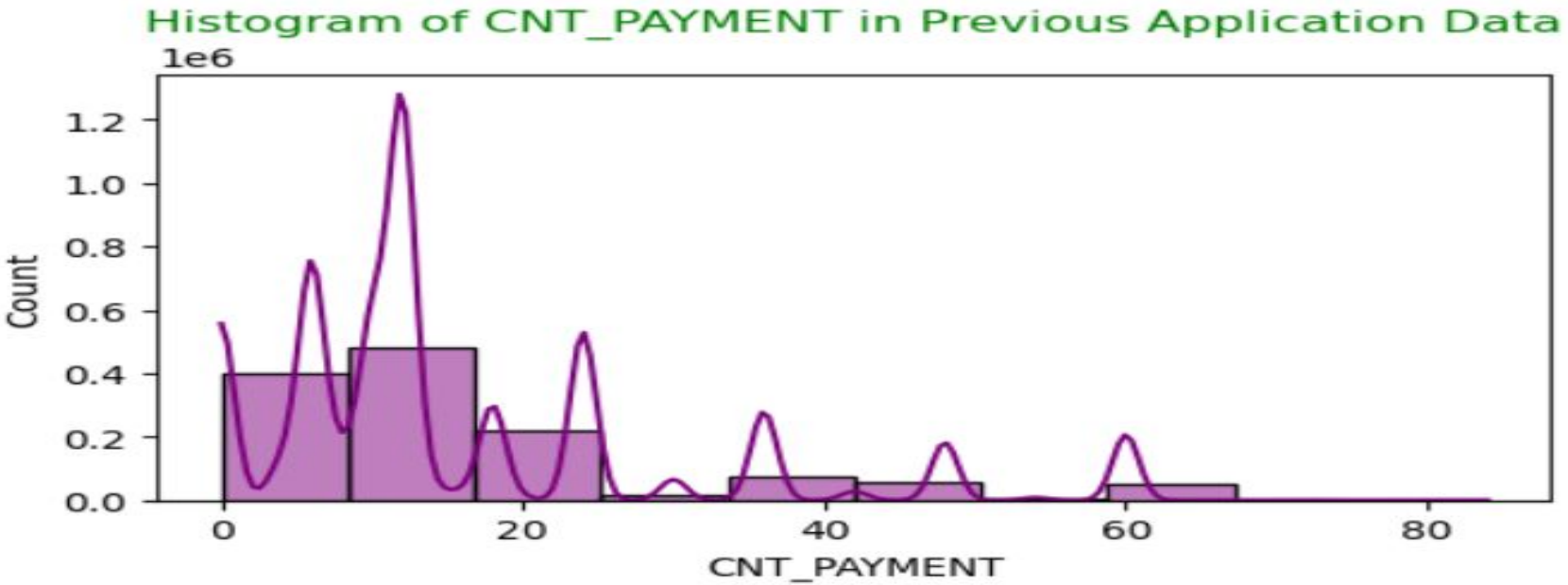
PROBLEM STATEMENT 😢

The provided data is related to loan application. The aim is to perform Exploratory Data Analysis (EDA) on this data to understand the features and their relationships, and to identify important driver variables that can predict whether a client will have difficulty repaying their loan.

- We are expected to find out the if a client has any difficulty in paying their installments which may leads to taking actions like denying the loan, reducing the amount of loan.
- We need to make sure that the clients who are capable of repaying loan are not rejected. Here we need to identify the driving factors/variables behind loan default.

APPROACH

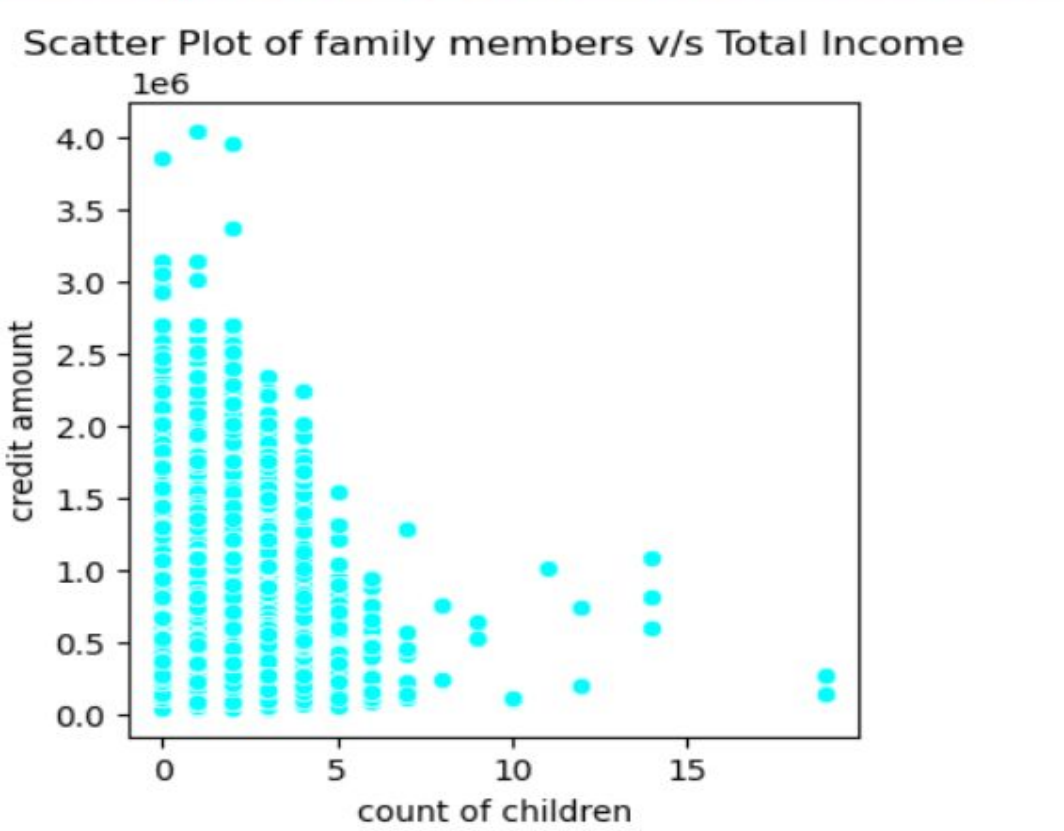
- Understanding the data
- Loading the data and checking data structure
- Data cleaning
- Dealing with missing values
- Handling outliers
- Univariate and bivariate analysis on both data
- Merging data
- Finding correlation matrix
- conclusion



Univariate analysis

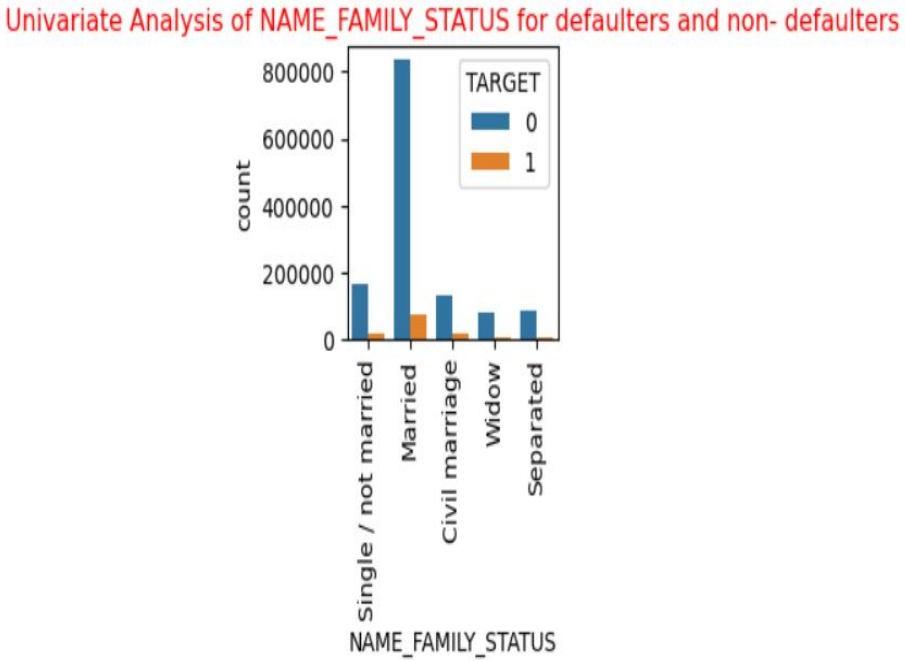
This shows that people want short term loans so banks can provide more such offers in future. As here we can assume that clients opting for longer term installments lack money.

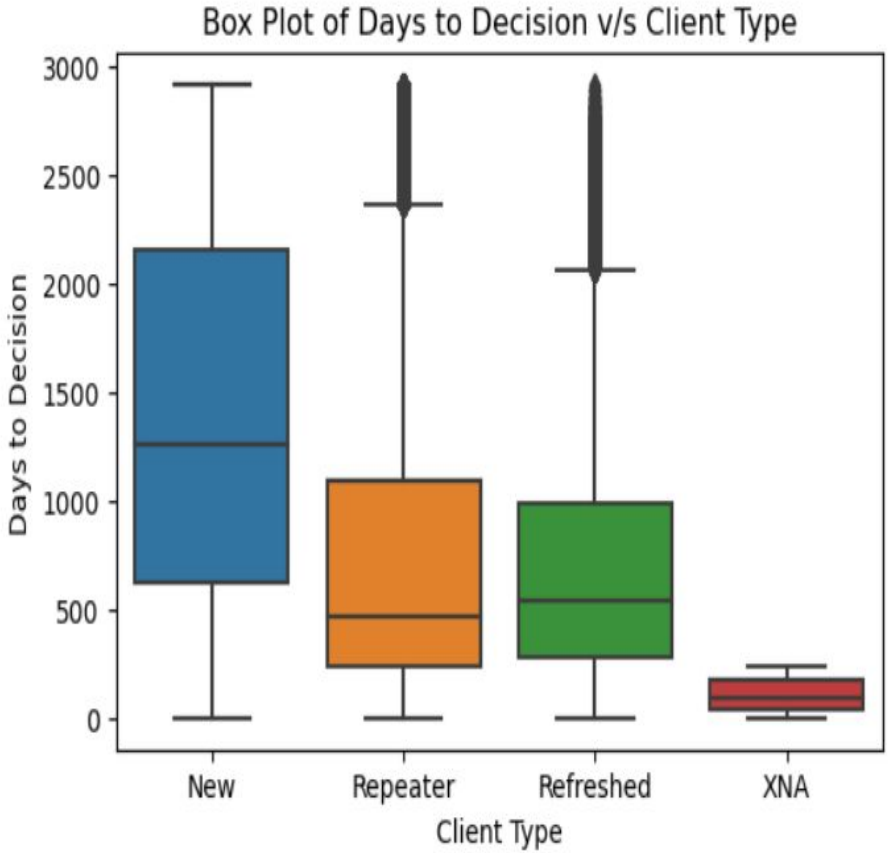
people with generally small family means lesser number of children are taking more loans, as we can assume that family with more children have more responsibilities.



UNIVARIATE ANALYSIS WITH THE TARGET VARIABLE:

We can observe that married people are more likely to be defaulters as in case if they have more expenses than others would not be able to repay loan amount on time.

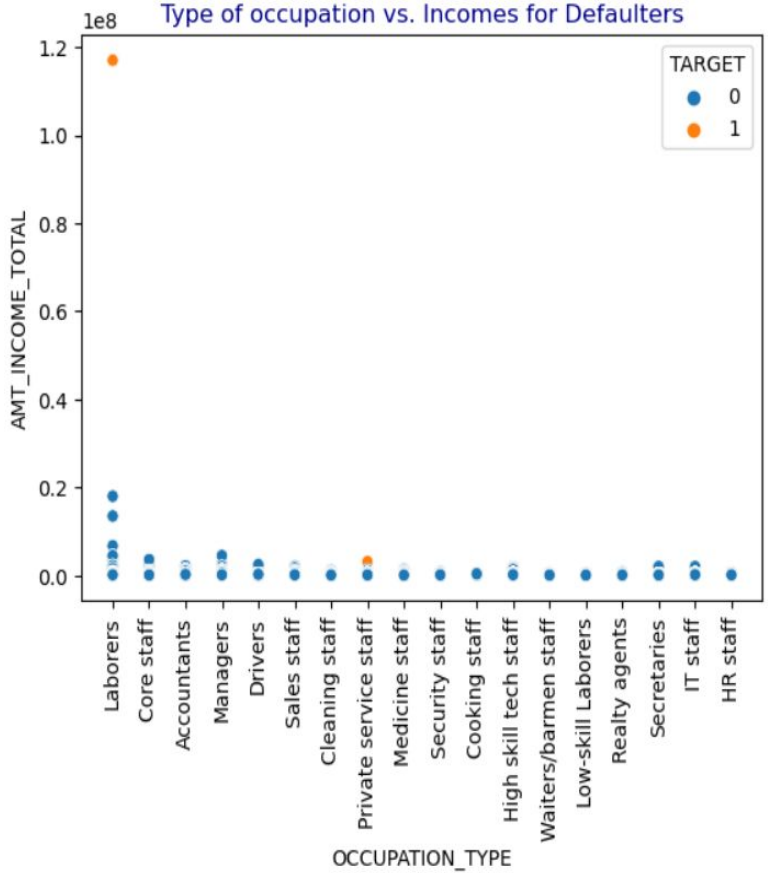




Bivariate analysis between days to decision and client type:

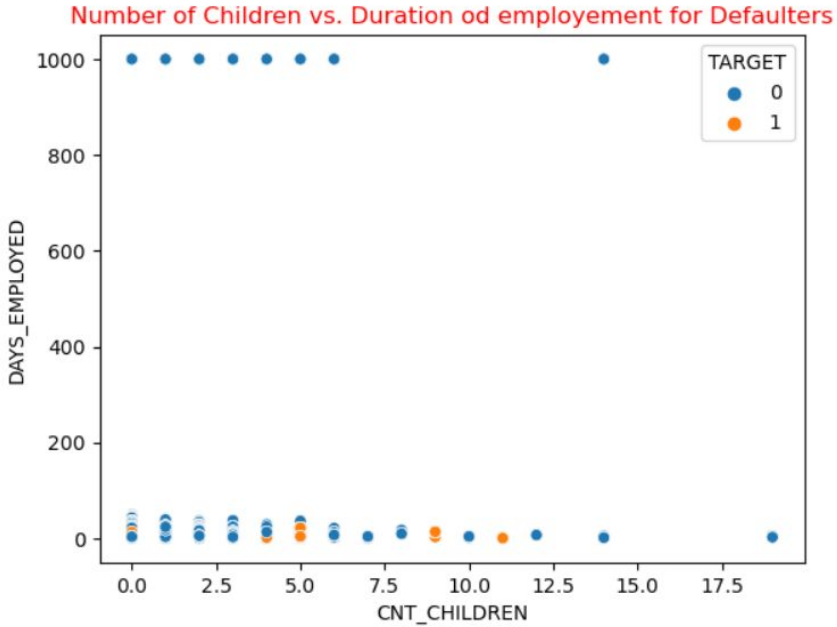
We can observe that bank istaking more time to decide to give loan to new clients as there is more information to be verified than repeaters.

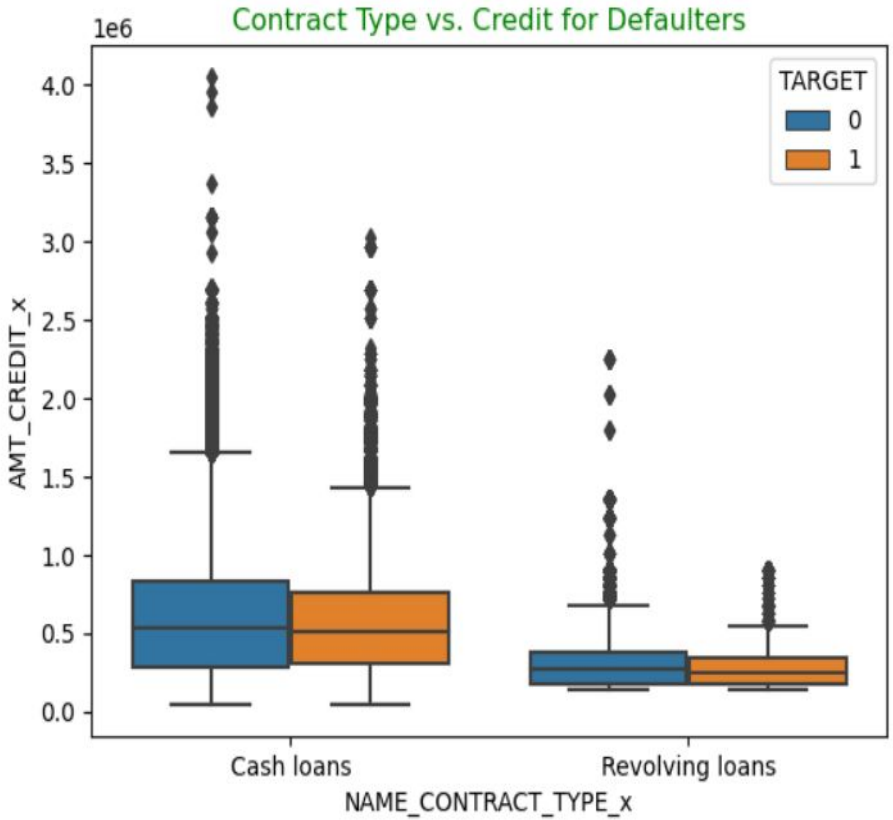
Here we can observe that the occupation (i.e labourers, private service staff) comes under defaulters list as less paying jobs is one of the key concern.



Bivariate analysis with the target variable:

Shows that the people with more children are defaulters as it might be difficult for them for handling expenses of more number of people.

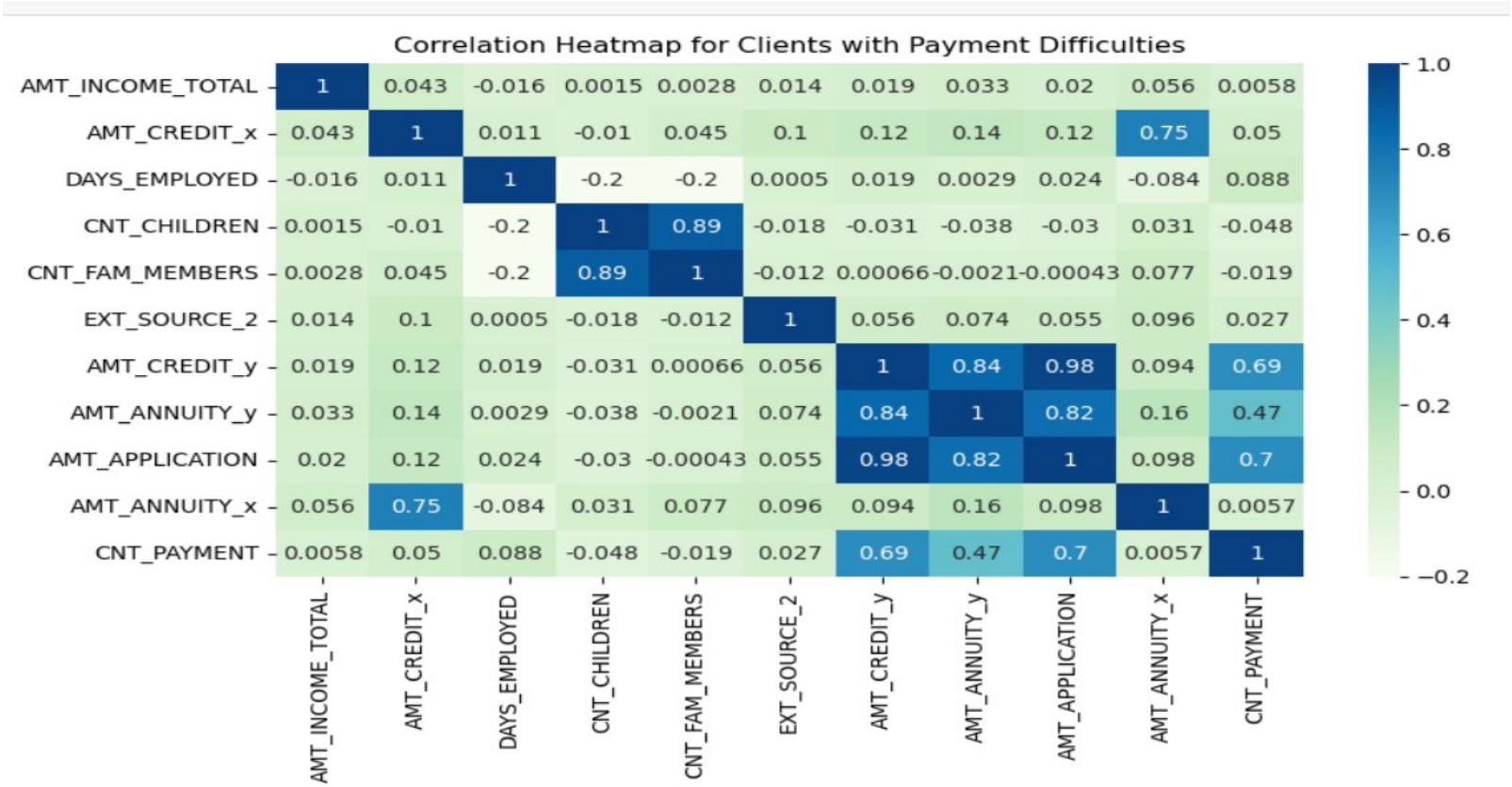




Bivariate analysis with the target variable.

Indicates people who opted for cash loans are more likely to be defaulters, it may be because the terms and conditions of cash type is not well suited for the clients financial condition.

Correlation heatmap for the clients with payment difficulties



Conclusions (defaulters)😠

Parameters to be considered before lending loan

TOP 5 CORRELATIONS OF DEFAULTERS

- **CNT_CHILDREN and CNT_FAM_MEMBERS:** indicates that people having more children or large family has more difficulty in paying loans, as large families have extra expenses to deal with.

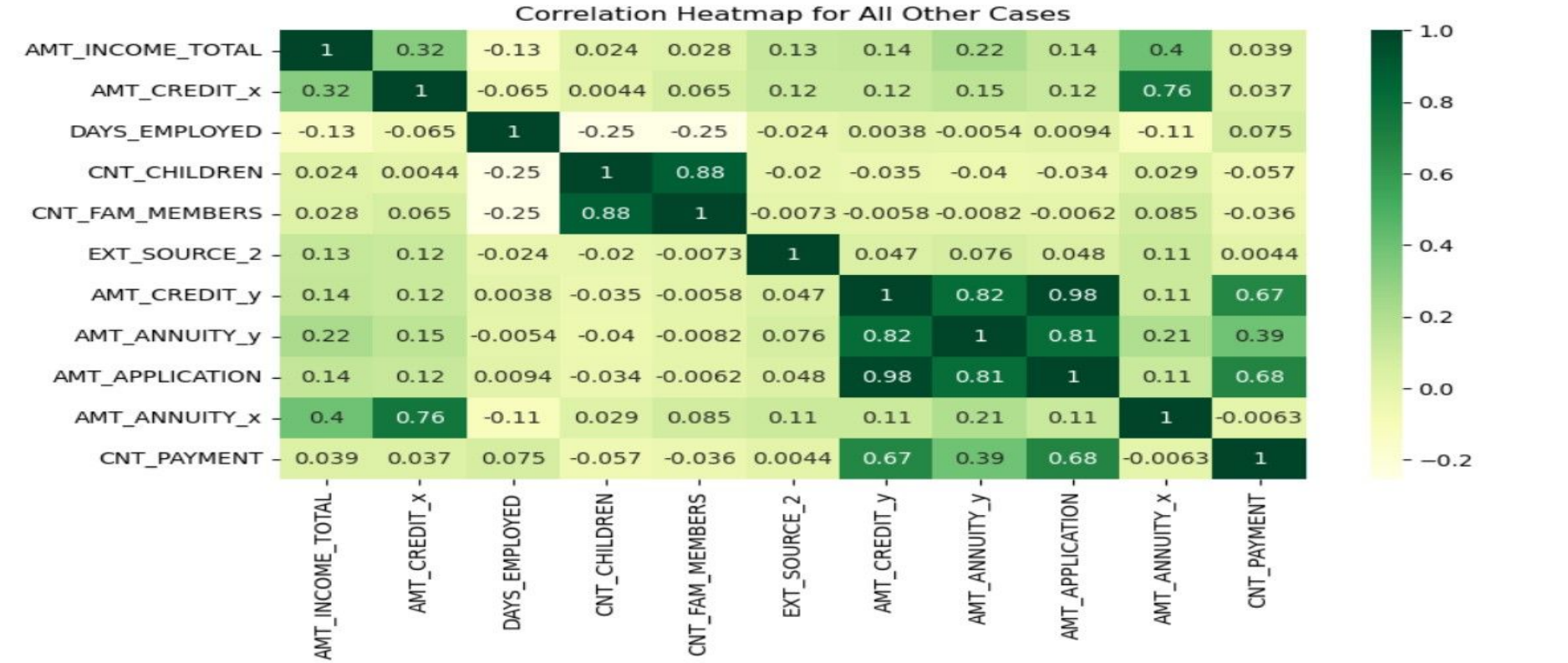
- **AMT_CREDIT_X and AMT_ANNUITY_X :** shows high correlation as larger the amount the bank has approved larger the amount(installment amount)to be paid on time which may become difficult for some clients.

- **AMT_CREDIT_y and CNT_PAYMENT ;** indicates the more the amount borrowed the more number of payments have to be made, but in case of elderly people it can become a risk for lending out large amount for more days.

- **AMT_CREDIT_Y and AMT_ANNUITYY;** this shows if a person has already taken a large amount of loan previously means they are paying huge amount of installments too which makes it riskier for the bank to give them loan again if their income and job is not stable.

- **AMT_APPLICATION and AMT_CREDIT_y :** shows the amount person ask from the bank and the amount bank agreed to give are close. if a person asks for large amount they have to pay more amount in installments too.

Correlation heatmap for the clients with no payment difficulties



Parameters on which bank can lend loan (non- defaulters)😃

Top 5 variables for non defaulters

- **AMT_INCOME_TOTAL and AMT_CREDIT_x:** indicates the more the person earns the more easier for them to pay installment amounts making it easy for bank sto approve their loans.

- **DAYS_EMPLOYED and CNT_FAM_MENBERS:** more the number of days of employment means stable job so even if they have big family it won't be difficult for people to repay loan.

- **AMT_ANNUITY_x and AMT_CREDIT_x:** This reflects a standard loan terms between the two showing that borrowing is well managed.

- **EXT_SOURCE_2 and AMT_ANNUITY_y:** indicates how external sources scores affect the lending process. high the score more easily the person will get the loan.

- **CNT_CHILDREN and AMT_INCOME_TOTAL:** indicates more the children more members in family to earn and thus making it easier to be non defaulter.