EDA CASE STUDY ON LOAN ELIGIBILITY REFINEMENT

BY- MUKESH BHATT

PROBLEM STATEMENT

- Our Study focuses on identifying indicators suggesting potential difficulties in making loan payments.
- These indicators aid in decision-making processes such as rejecting loan applications, adjusting loan amounts, or offering loans to risky applicants at higher interest rates.
- The primary objective is to prevent deserving borrowers from being denied loans while effectively identifying individuals who may encounter repayment challenges.
- We will use Exploratory Data Analysis (EDA) to achieve these goals and inform our decision-making processes in lending.

STRATEGY AND METHODOLOGY

- Understanding the Problem Statement
- Understanding the 'Previous Application' data with the help of a data dictionary. Also, check the shape, info, d-type, and statistical values to get an idea about the data
- Data Cleaning:
  - ✓ Removed a few columns that were not required for the analysis. i.e. 'FLAG_DOCUMENT_2','FLAG_DOCUMENT_3'.
  - ✓ Checked for the outliers using a boxplot. However, did not treat them as not necessary for this case study.
  - ✓ Checked the missing values, and deleted all the columns which have more than 40% of null values. Treated the remaining values with the help of mean, and median, and also left some columns where it is not required.

- ✓ Did the sanity check and changed the values to positive as they were given in negative. Also, changed the values from days to years. For ex- 'days birth'
- ✓ Checked if there are any duplicate values in the data set.
- ✓ Binned a few columns for better understanding. i.e. 'year birth', 'year employed'

- ➢ Checked Data Imbalance: While checking for the same we found that more than 91.9 %have no payment difficulty however, 8.1% have payment difficulty

- ➢ Univariate analysis: Analysed one variable at a time using a count plot on the application data

- ➢ Segmented Univariate: Analysed one variable at a time with respect to the 'target' column (clients with payment difficulty)

- ➢ Bivariate analysis: Analysed two variables with respect to the 'target' column(clients with payment difficulty)

- **Correlation:** Checked the correlation of variables with respect to clients who defaulted and who did not default.

- Loading and understanding of the 'Previous Data'

- Data cleaning:

(a) Few columns have 'XNA' and 'XAP' so changed them to null values.

(b) Checked the null values and deleted the columns that have more than 30% of null values.

(c) Treatment of null values where required with the help of median and mode.

(d) Removed a few columns that are of no use for the analysis

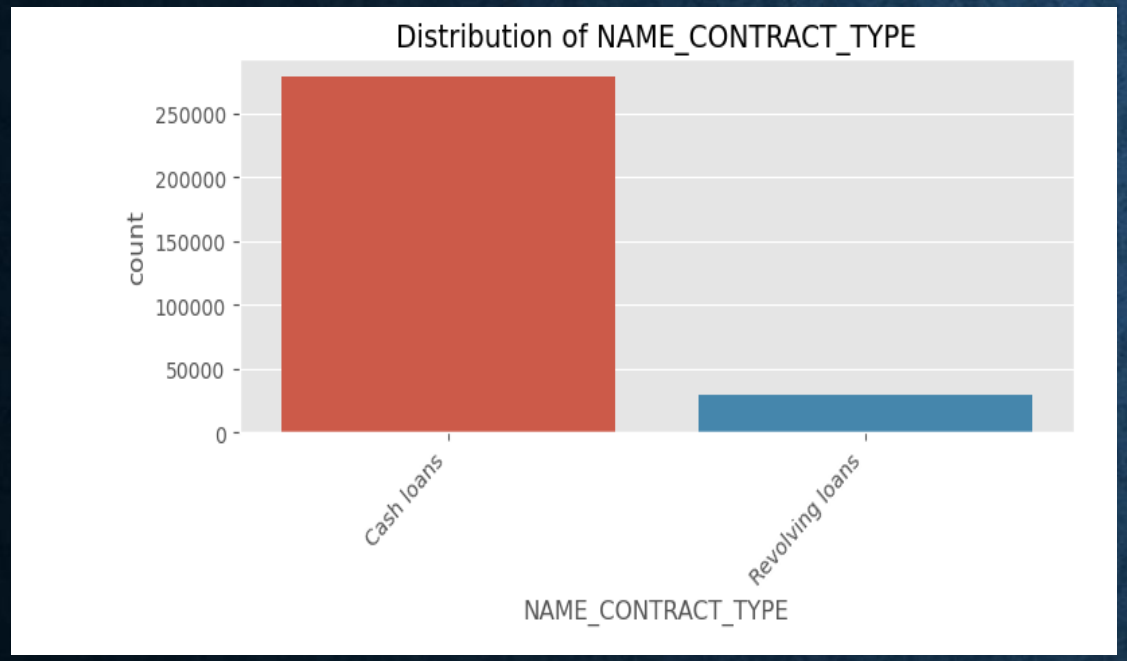
(e) Changed the values from negative to positive

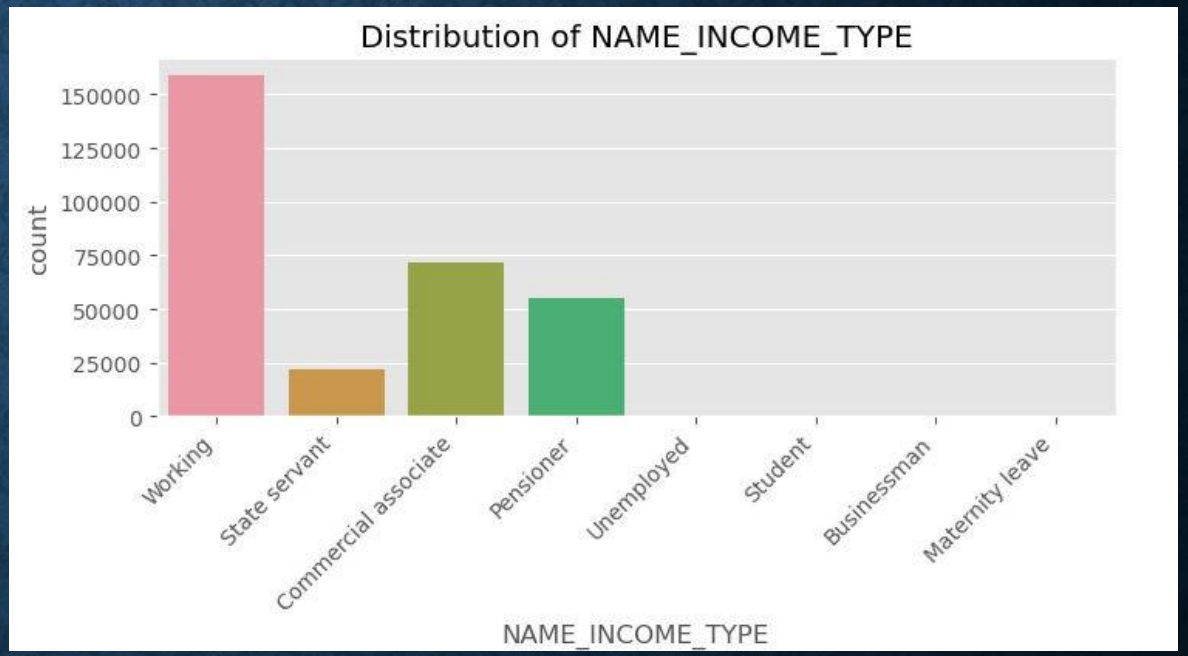
(f) Checked the outliers with the help of a boxplot but did not treat them as not required.

- Merged the dataset using left join so that we do not lose values.

- Did univariate analysis

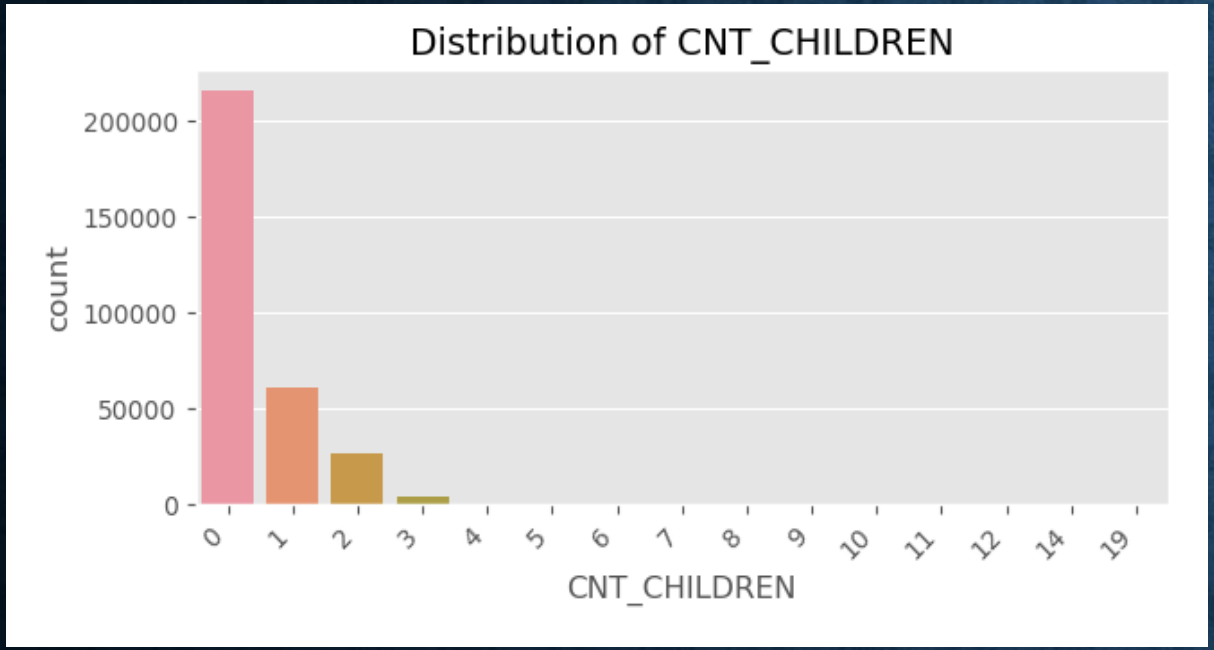
- Did bivariate analysis

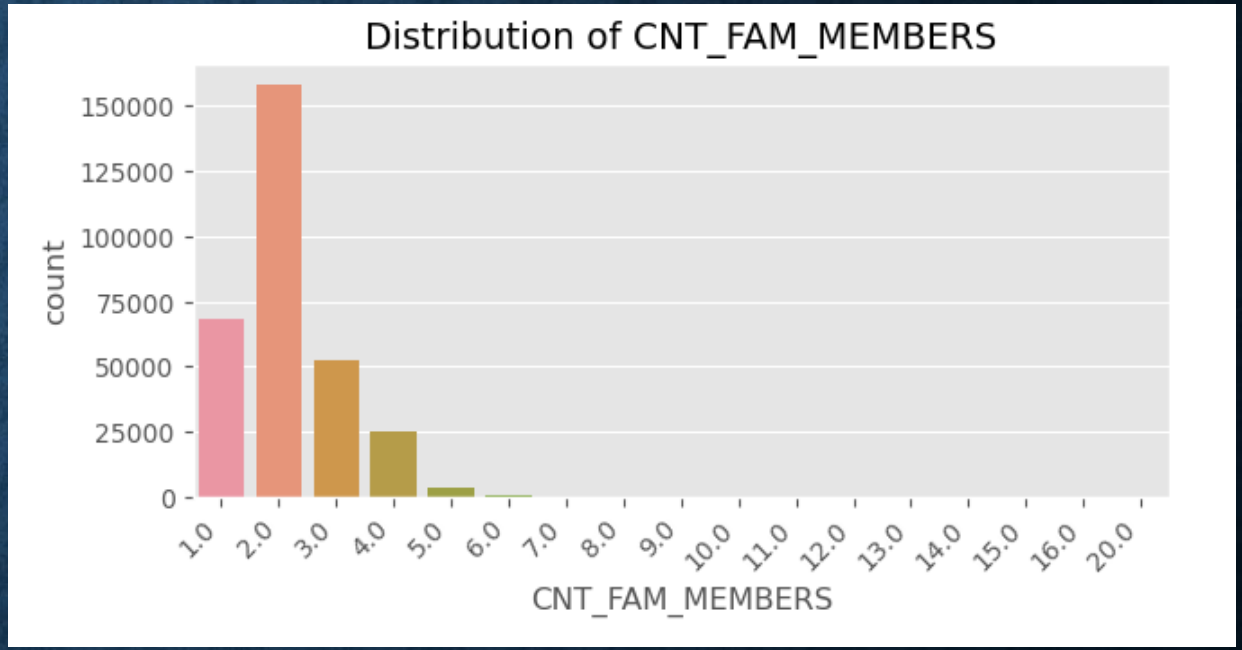




Most of the clients have applied for cash loans instead of Revolving loans

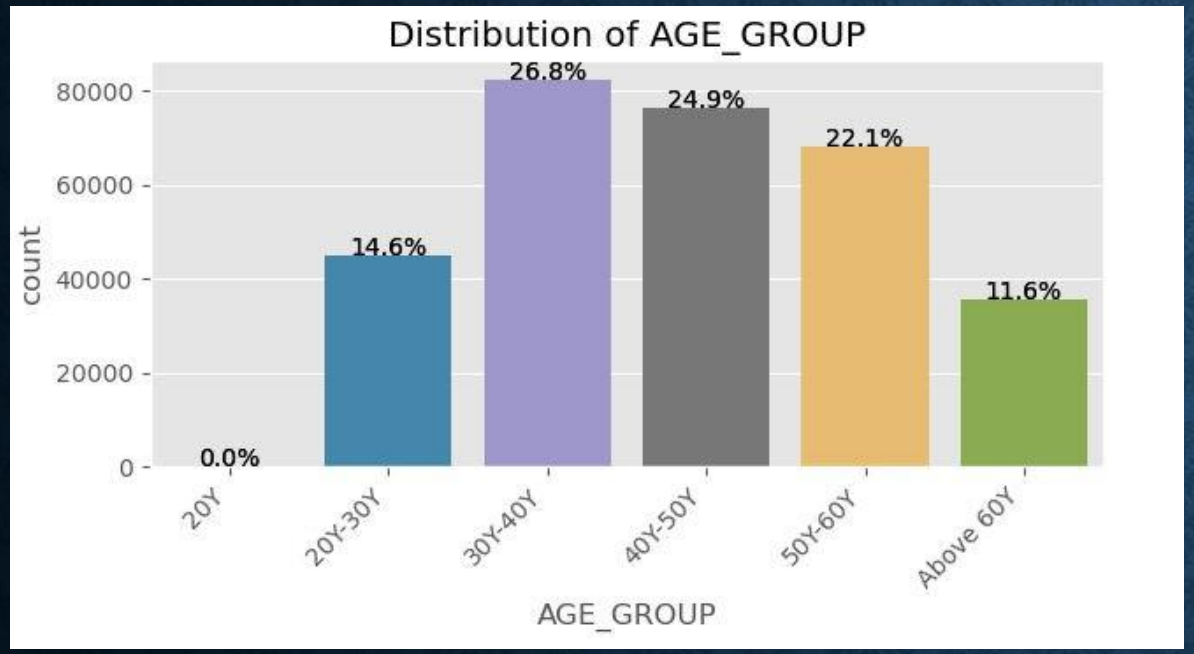
The working class is applying for a loan in comparison to other type

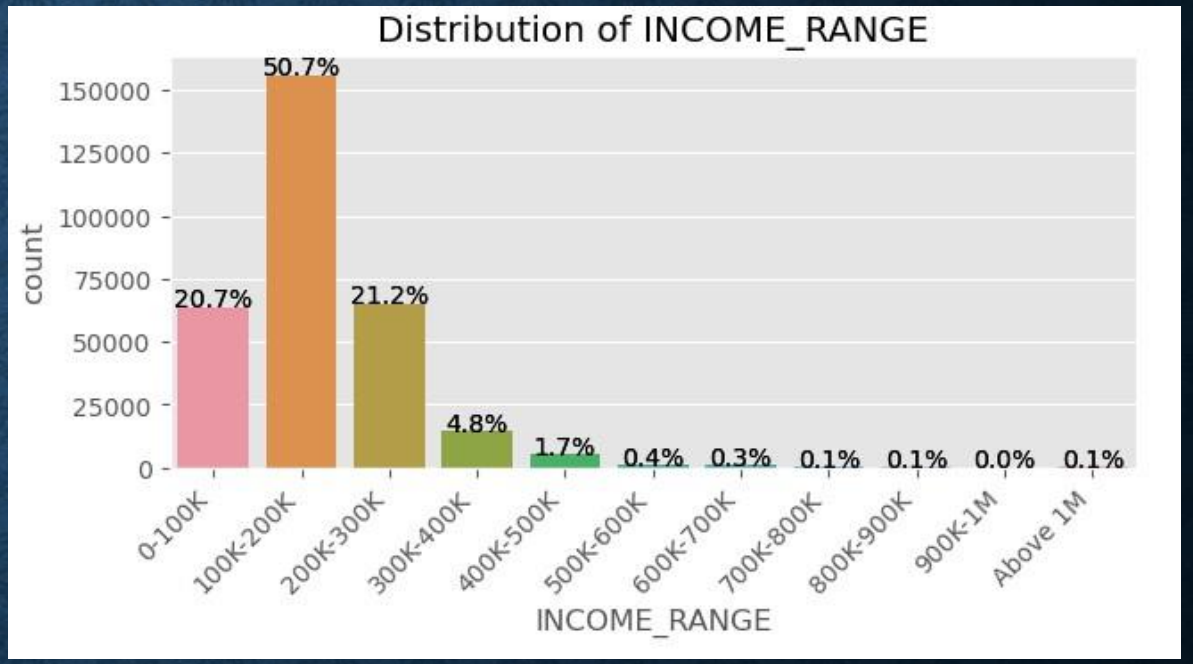




Labors have applied for most of the loans, followed by sales staff, Managers, and drivers.

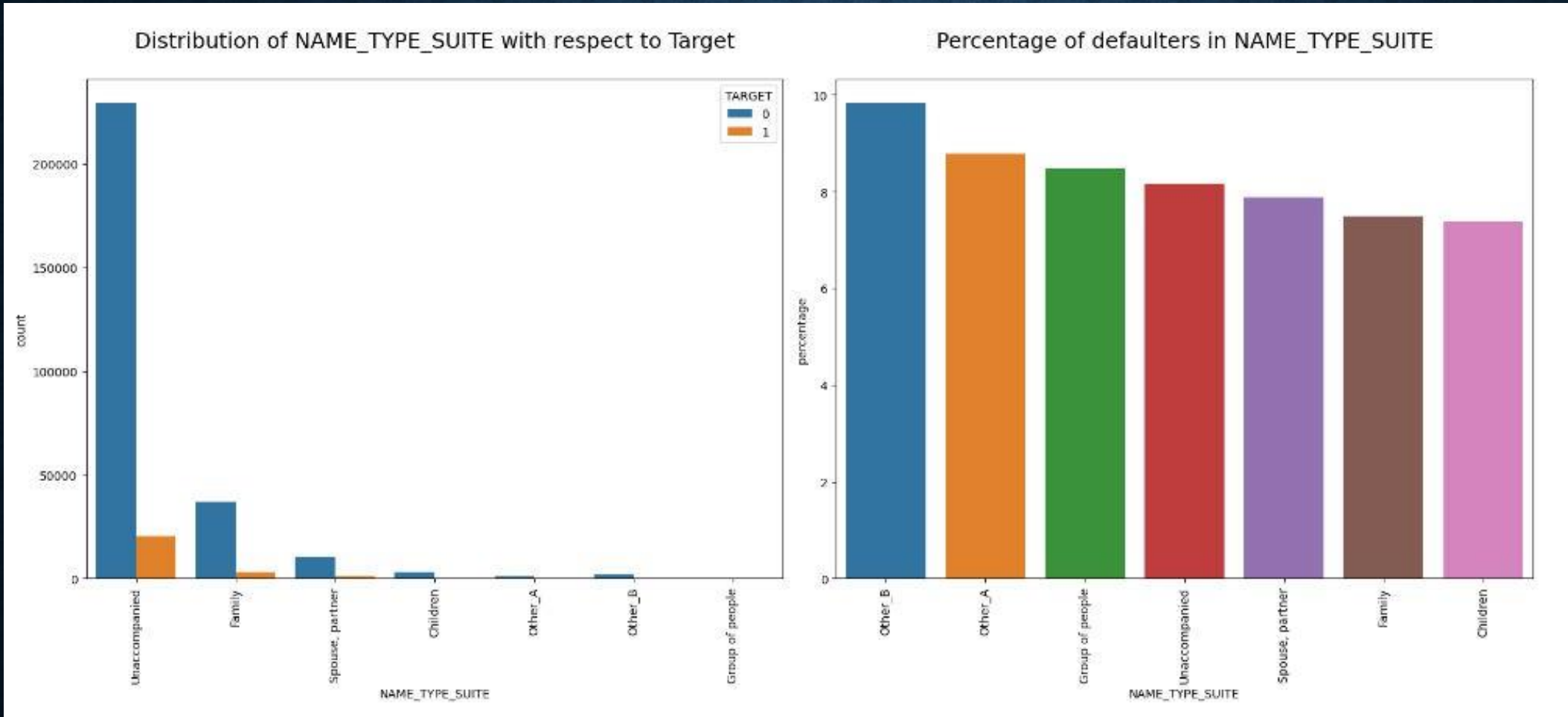
Most of the clients who applied for loans have two family members followed by one and three.



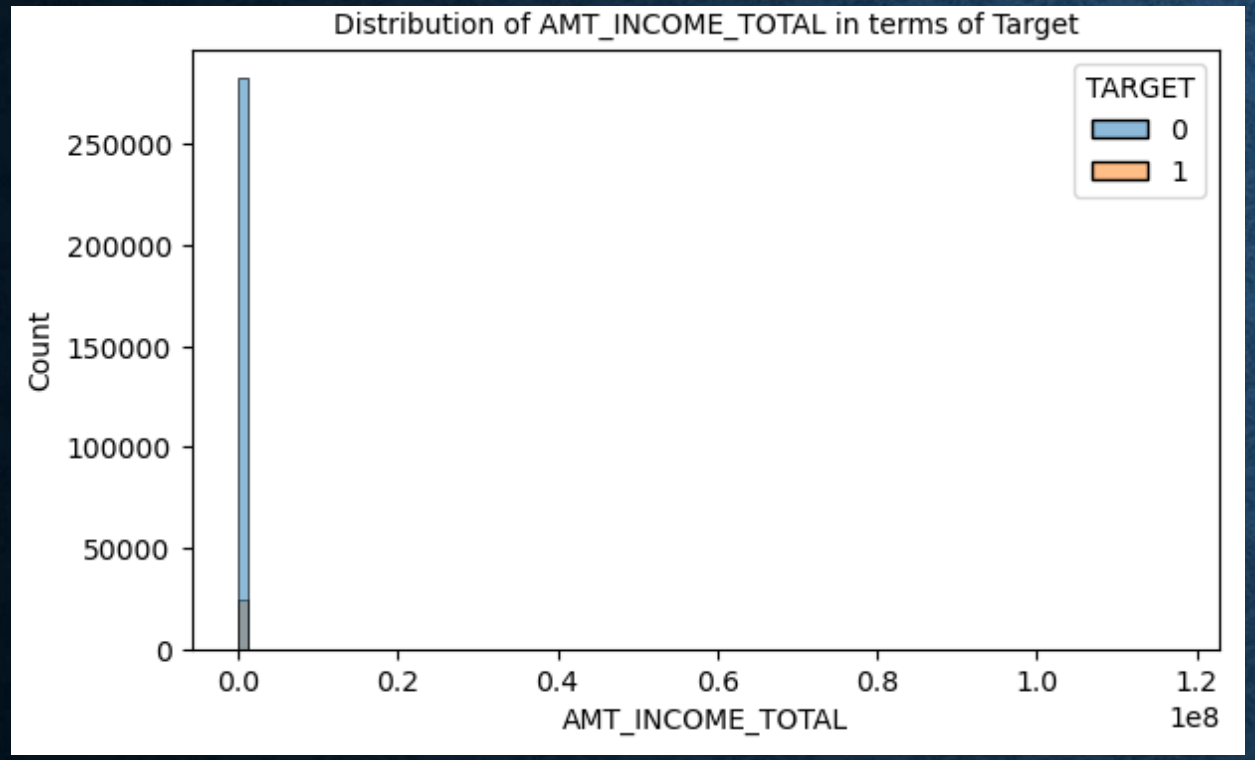


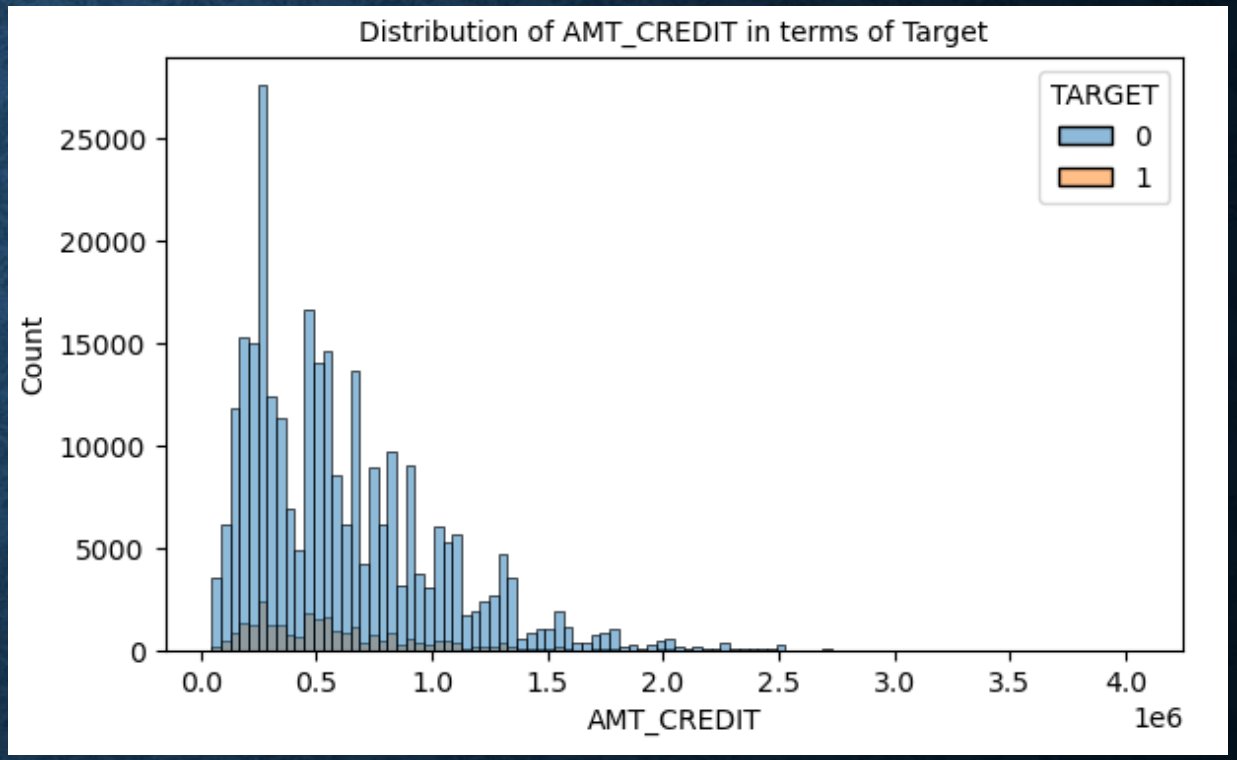
Customers aged between 30-40 years have applied for most of the loans followed by 40-50 years and 50-60 years.

50.7% of clients whose income range is between 200K to 300K has applied for the loan



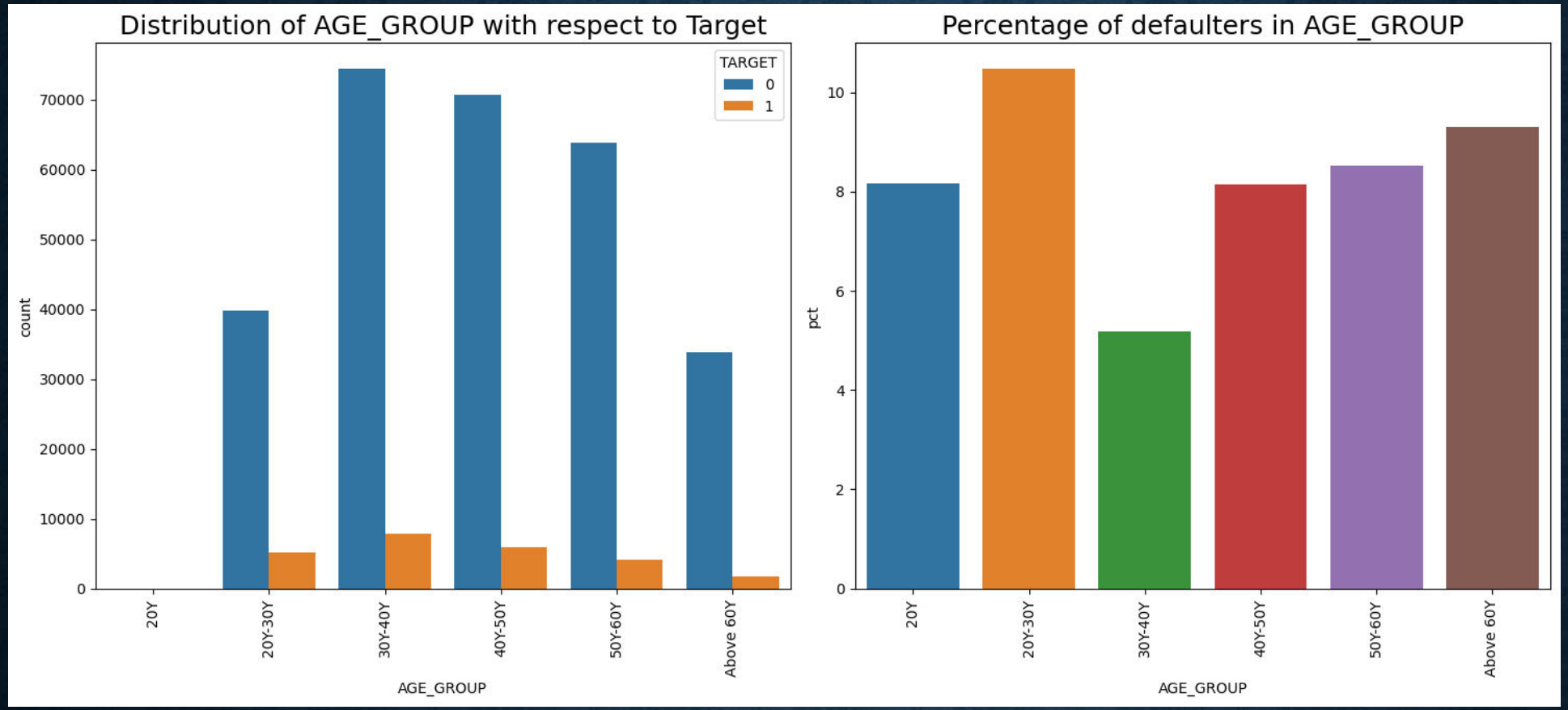
Unaccompanied had taken most of the loans and their default rate is approx. 8% which is still safer. The clients living with children or families have the lowest default percentage. However, their applications are also less.



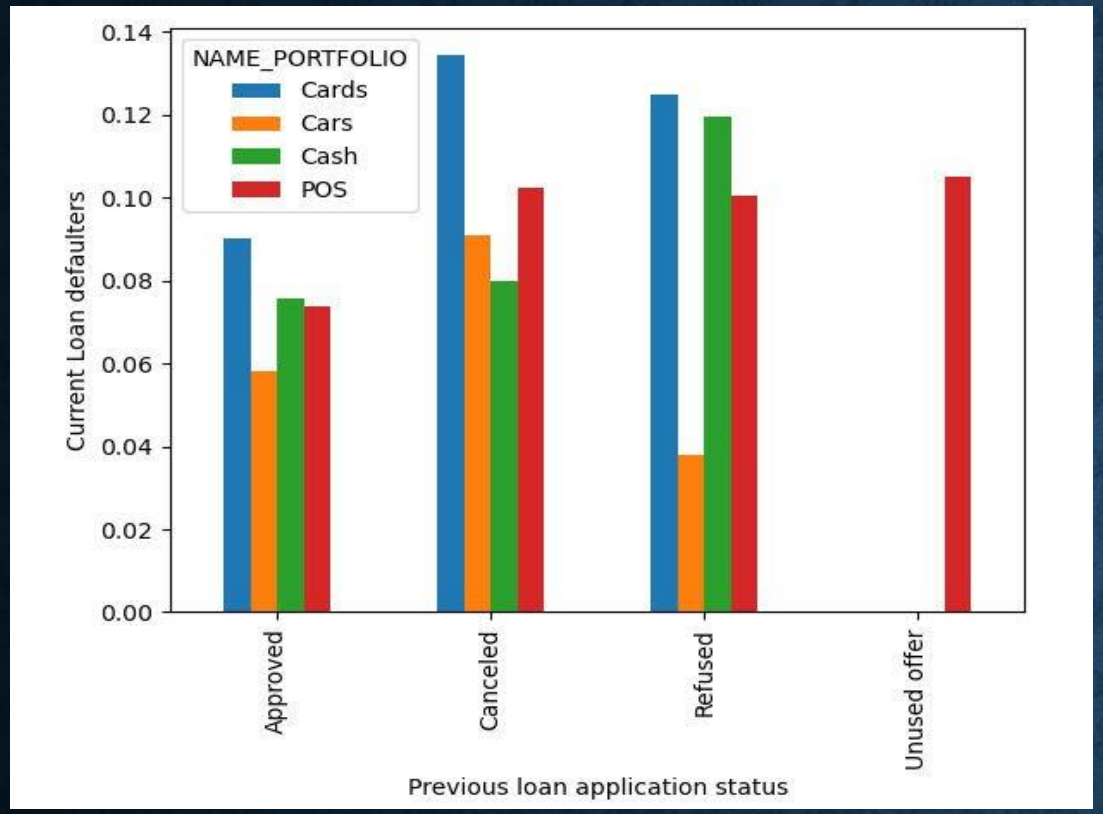


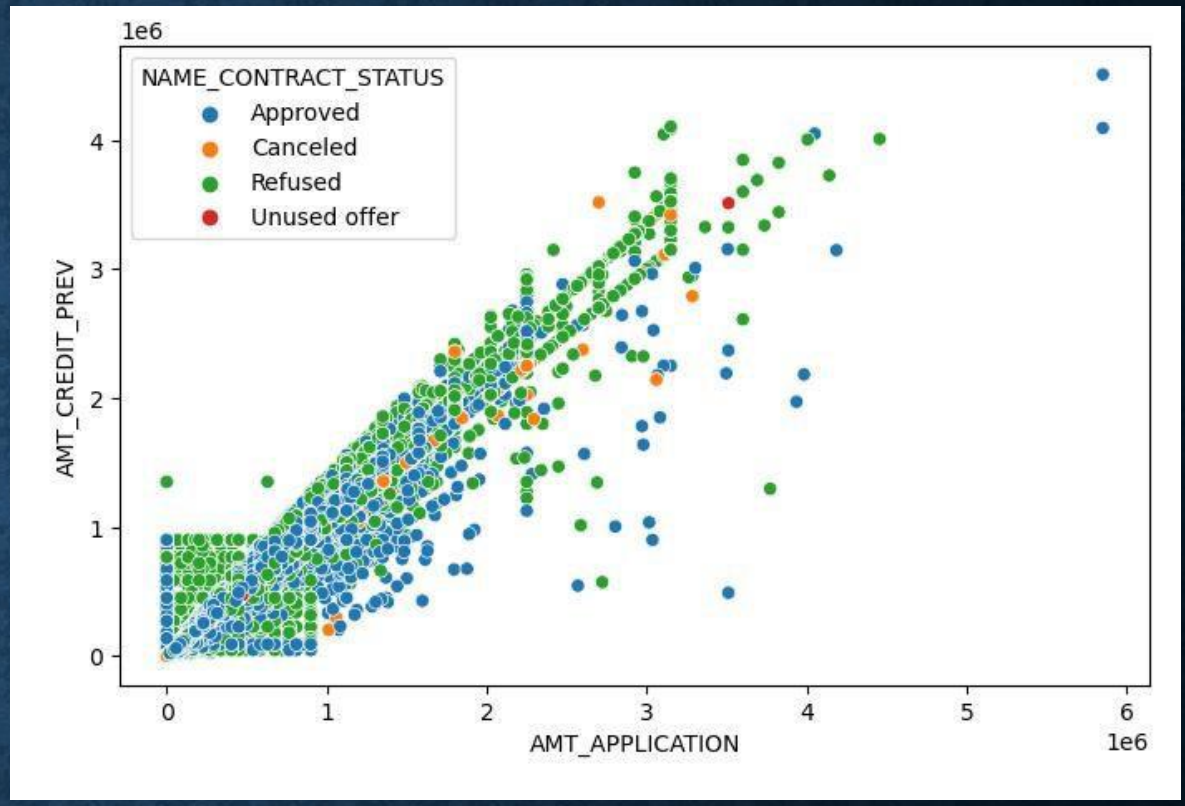
Most of the clients have an income between 0 to 1 Million and their default ratio is very low.

Most of the credit requested is between 0 to 1 Million and also most of the defaulters are between this range only.



Most of the clients who applied for loans are between 30Y-40Y and they also have default values between approximately 5%. 20Y-30Y clients have the highest defaulter percentage i.e. 10.3% which is very high





-Most of the clients were defaulted who previously applied for loans for cards.
-For Refused loans, the clients who applied for CARS are less defaulted.

Looking at the graph most of the clients lie between 0 to 1 Million for both the amount requested and the amount credited. As the application amount is increasing, the credit amount is also increasing.

RECOMMENDATIONS

- Bank should target Cash loans as they are safer to give.
- People who have children less than 1 and greater than 5 are safer to give loans.
- Bank should target clients having children less than 5.
- Clients who are working in others, business entity type 3 are safer to give loans.
- Bank should focus on clients who are working as accountants, and core staff.
- Having house apartments and being married.
- Accompanied people can be safer to give loans.
- Bank can focus more on females as they are less defaulters.
- Amount segment:
  - ✓ The credit amount should not be one Million
  - ✓ Income range should be below one Million