



EDA CREDIT CASE STUDY

SUBMITTED BY:

MAHESH BABU R (DDS2090165)

SRAVANA KALYANI VUNDELA (DDS2090322)

DATA EXPLORATION

Problem Statement

- ▶ 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 .

Understanding consumer attributes and loan attributes

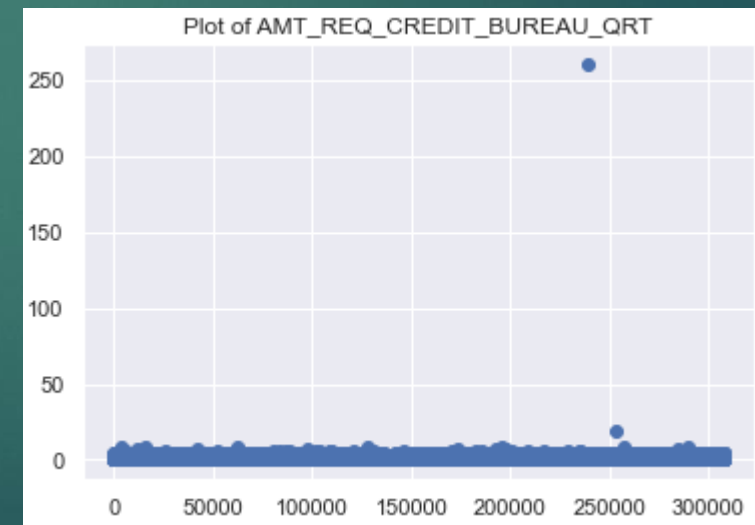
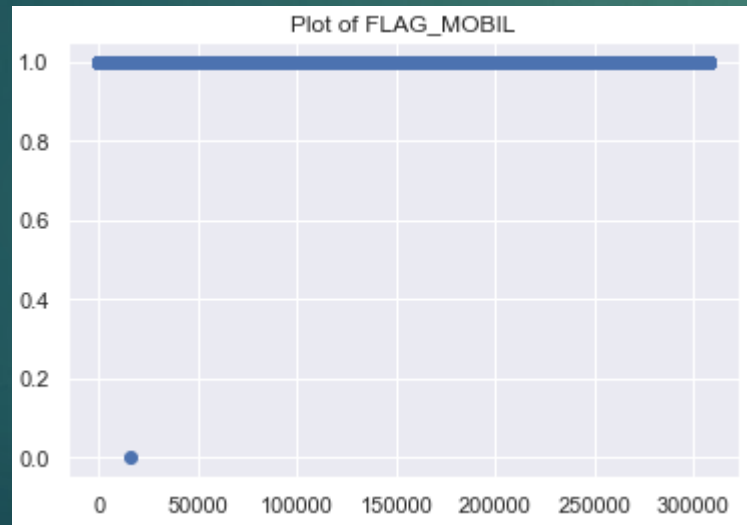
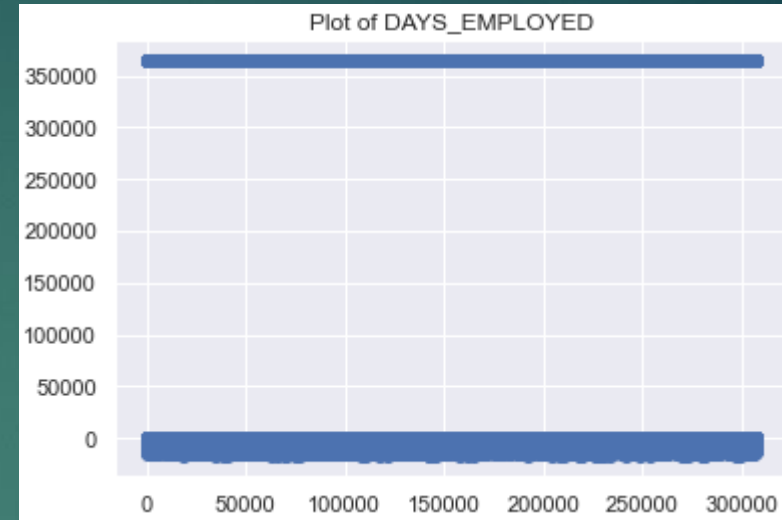
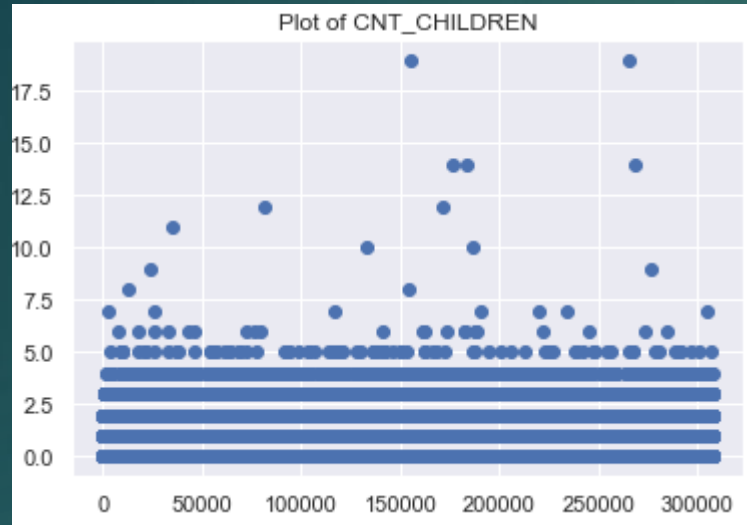
Focus on not rejecting the consumer applications who can repay the loans on time which will be an asset for the banking and financial companies.

- ▶ Need to buckle down on the default cases where consumers are facing difficulties to pay the loan.

DATA CLEANING AND MANIPULATION

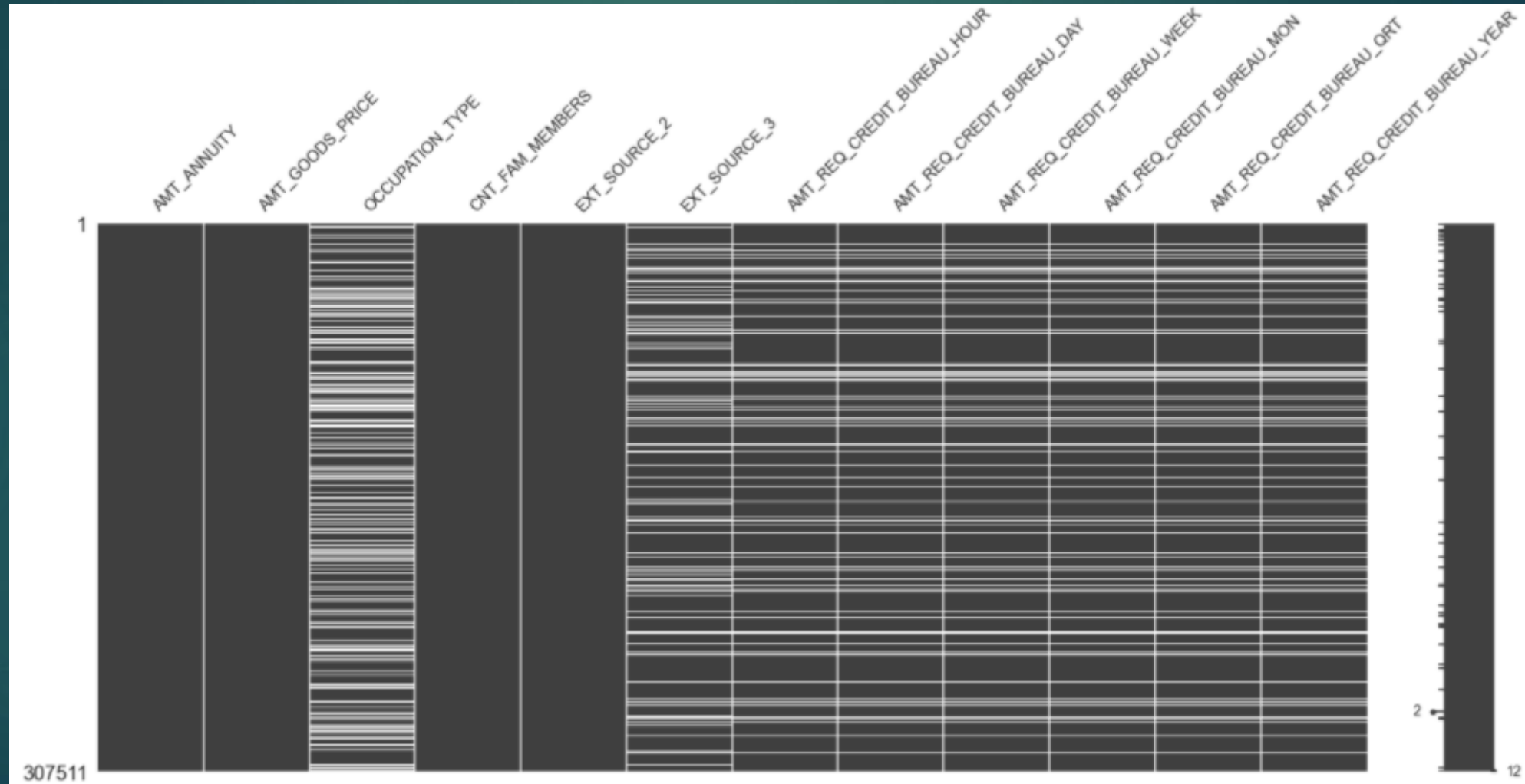
DEALING WITH MISSING VALUES

- ▶ In this dataset 49 variables which contains more than 47% NA are Dropped.
- ▶ 39 irrelevant variables which are showing less impact on meaningful insights are dropped .
- ▶ Outliers are verified before imputing null values.
- ▶ High count of children is not a valid observation therefore it is considered as a extreme outlier.
- ▶ Days employed consists extreme outlier e.g. 1000 years.
- ▶ Flag mobile not having a value is impractical as everyone consists of mobile.
- ▶ AMT_RQD_CREDIT_BEAURO_QTR variable having high outliers.

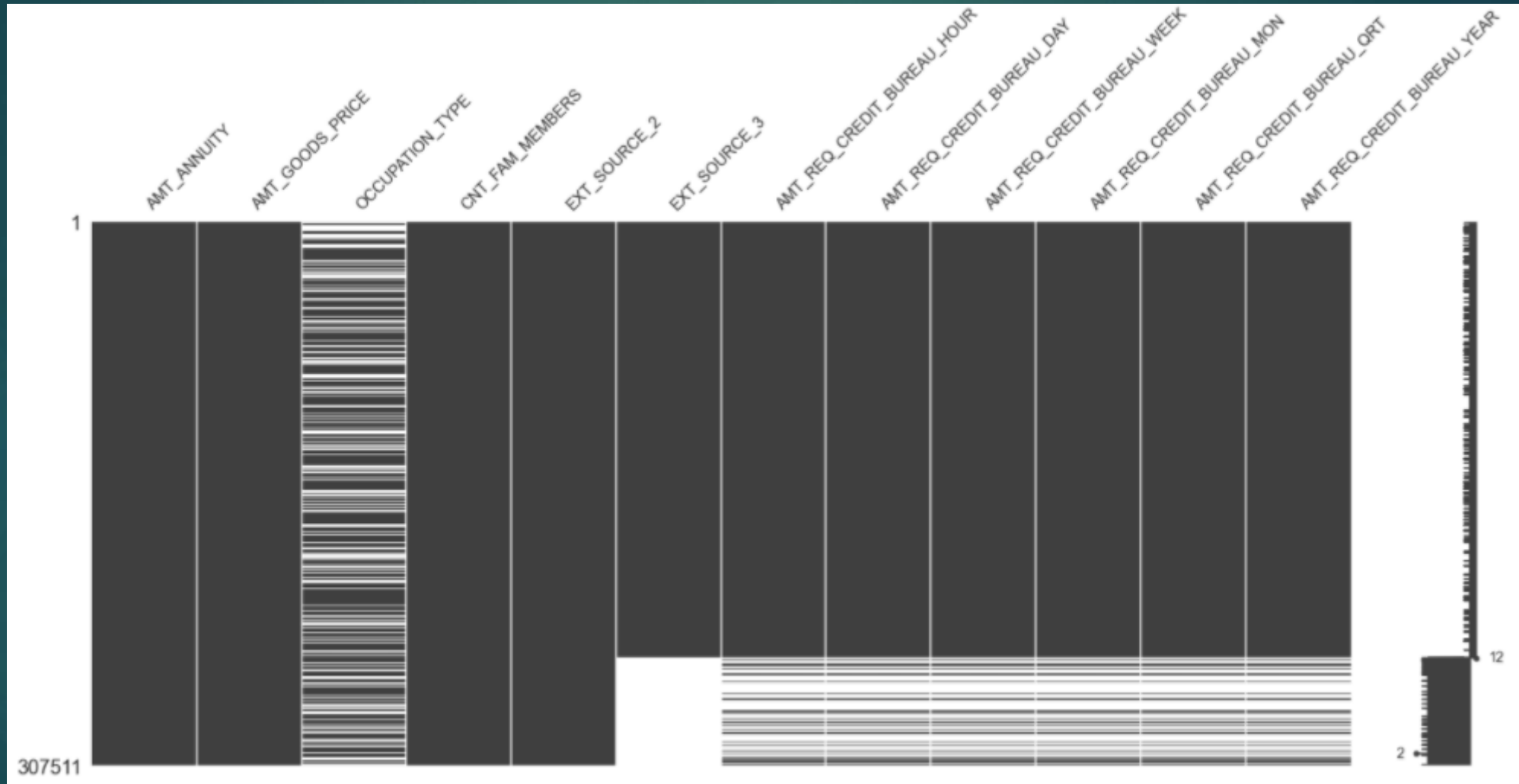



MISSING VALUES

White lines in the below figure represent the missing values of their corresponding columns



Missing values went to the extreme bottom corner when sorted w.r.t to the EXT_SOURCE_3



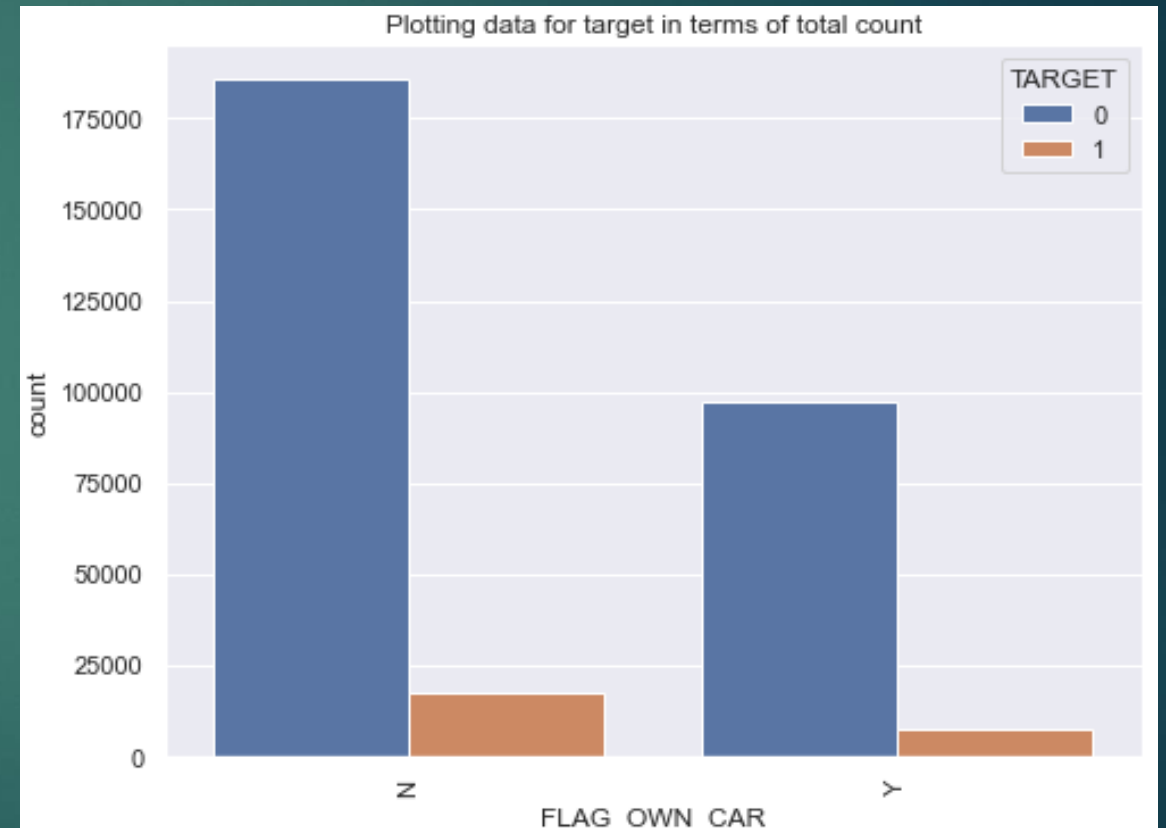
- 
- ▶ Imputed all categorical columns with mode, except occupation type column as imputing this column with mode will mislead our analysis.
 - ▶ Imputed Numerical columns with mean if mean and median values are closely precise or median if they are not since we can't remove outliers .
 - ▶ Converted data types of categorical columns into categories and data types of numerical columns into integer or float to get the required format for analysis.
 - ▶ Converted DAYS_BIRTH column into AGE , AMT_INCOME_TOTAL into Income range groups, AMT_CREDIT into Credit range for segmented analysis.

DATA IMBALANCE

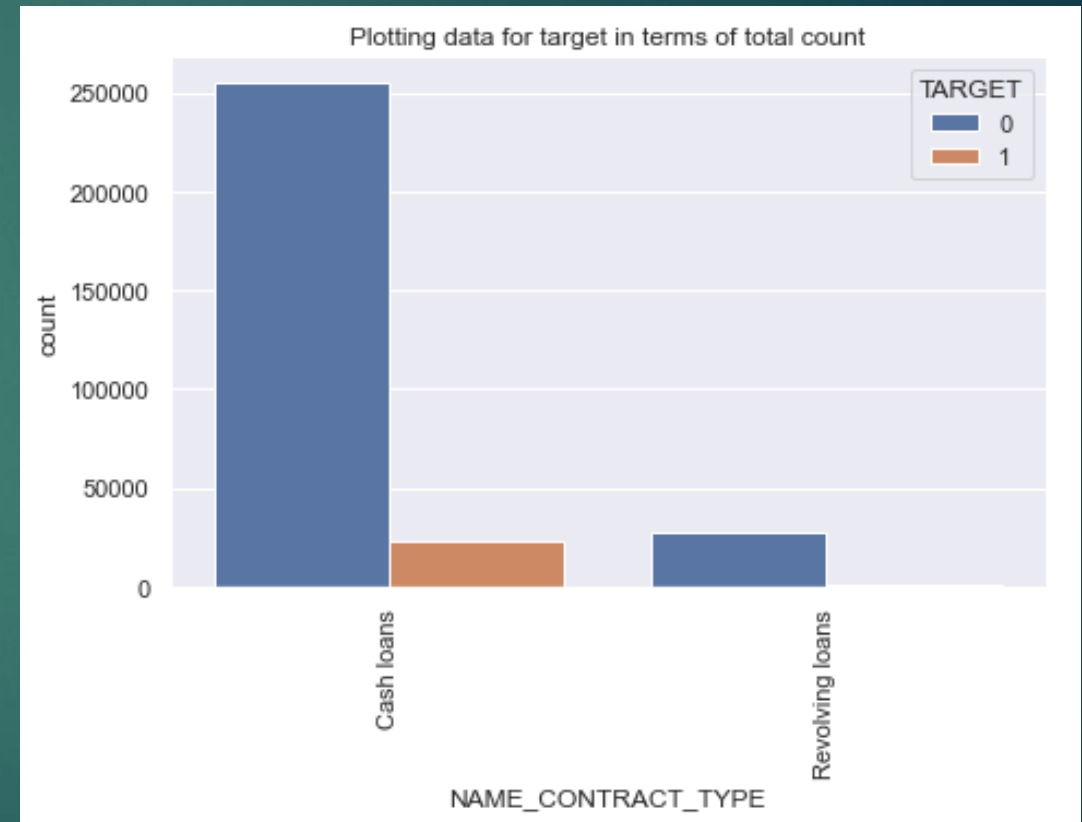
- ▶ As the proportion of data is distributed unevenly w.r.t to loan target column resulting data imbalance is validated in a specific way by bifurcating dataframe into tar_0,tar_1.
- ▶ Further Categorical column analysis is done based on tar_0,tar_1.
- ▶ Derived insights are discussed in detail in the further slides with valid reasoning and performing demographic analysis.

Univariate Analysis for categorical columns

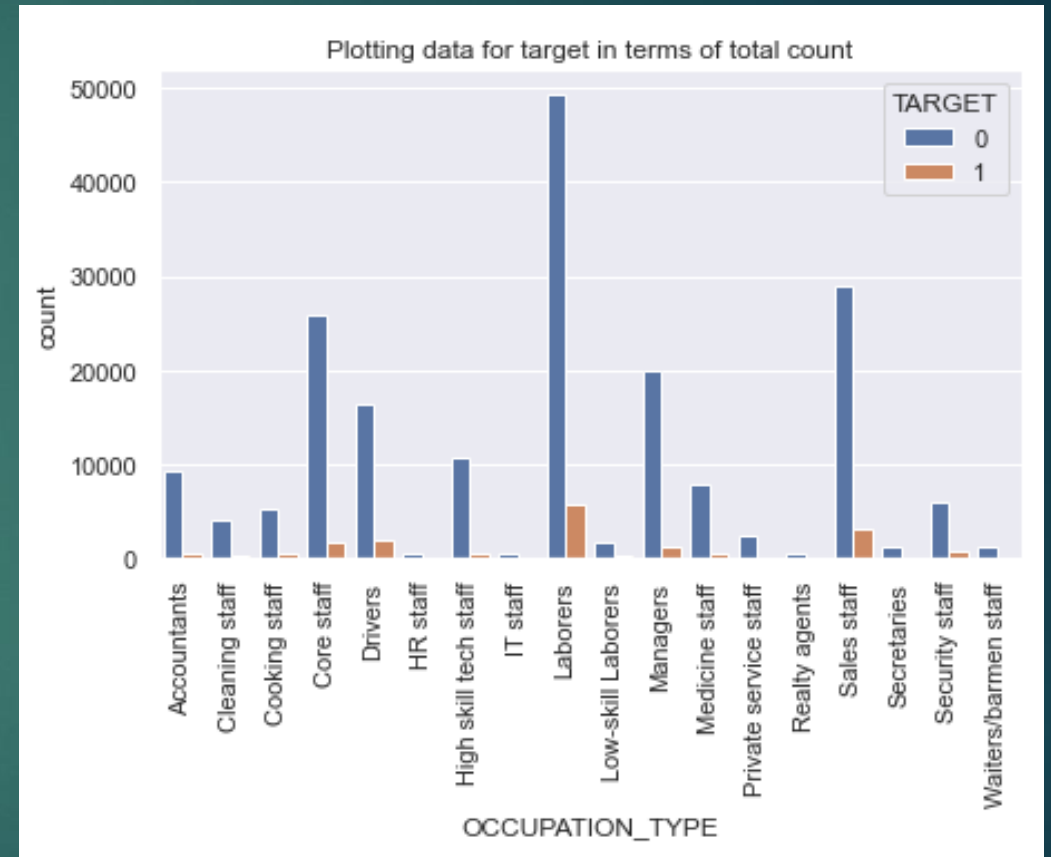
- ▶ Categorized the dataset w.r.t to the target values .
- ▶ People not having car are more inclined towards cash loans with a decent repay rate.



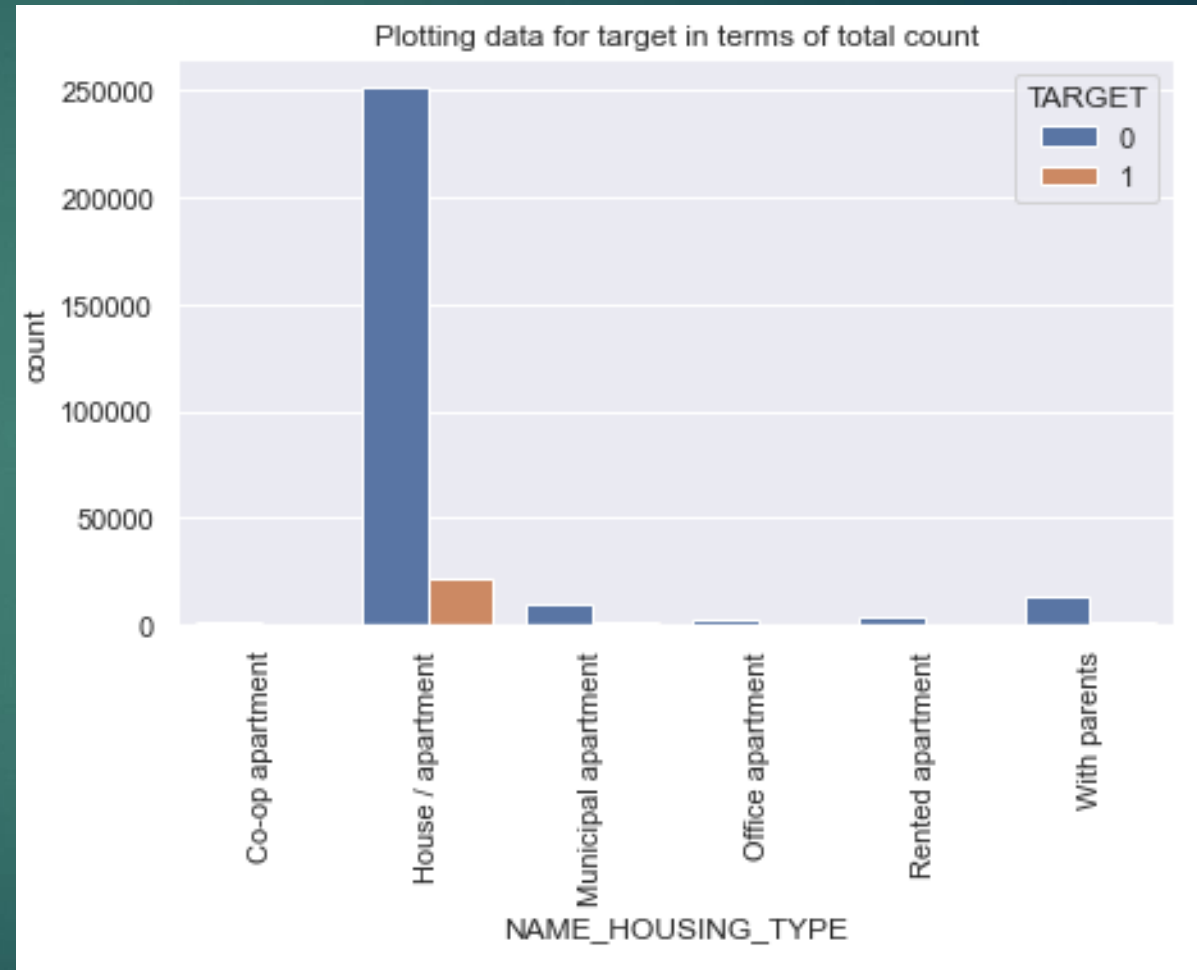
- ▶ Categorized the dataset w.r.t to the target values .
- ▶ Count of cash loans approved is more than the revolving loans .
- ▶ Defaulters count is more for the cash loans compared to revolving loans.



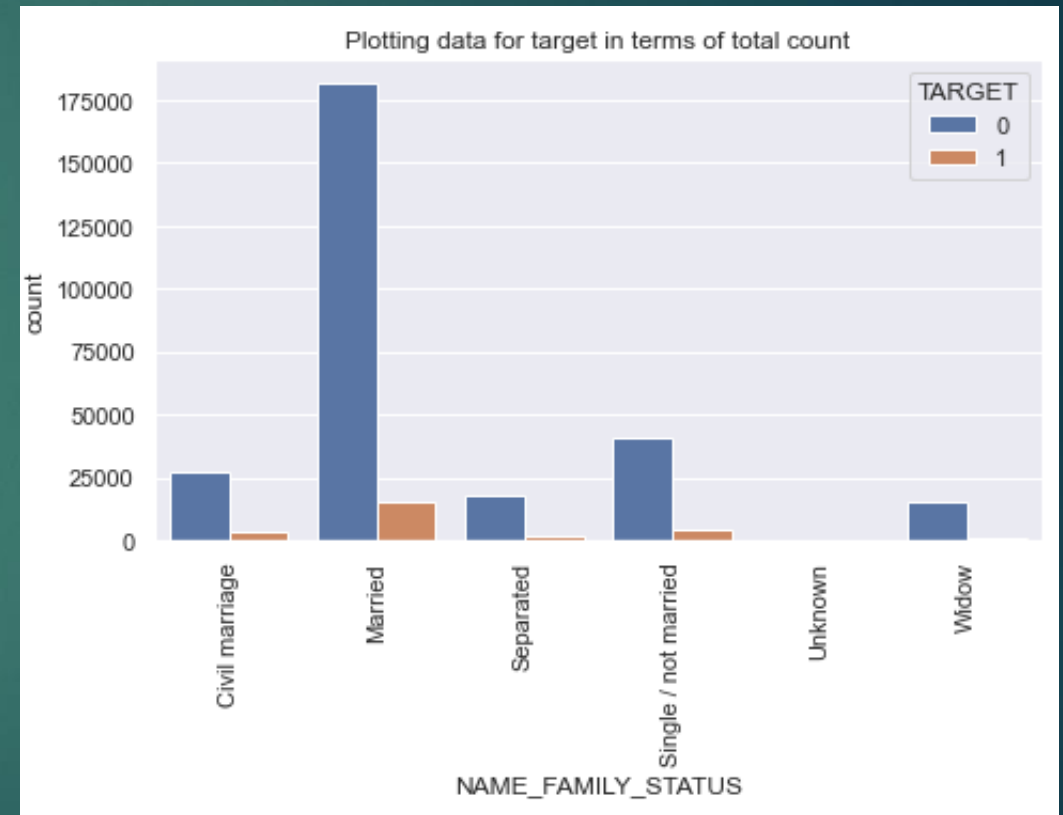
- ▶ As per target 1, Count of Laborers occupation type is more compared to the other occupation types which means low skilled labor are facing more difficulties to pay the loan .
- ▶ Around 31% people have not provide the data of their occupation type, majority of the occupation type in this data frame is Laborers then comes Sales staff followed by Core staff .
- ▶ As per target 0 laborers are the people who are mostly likely to pay the loans .



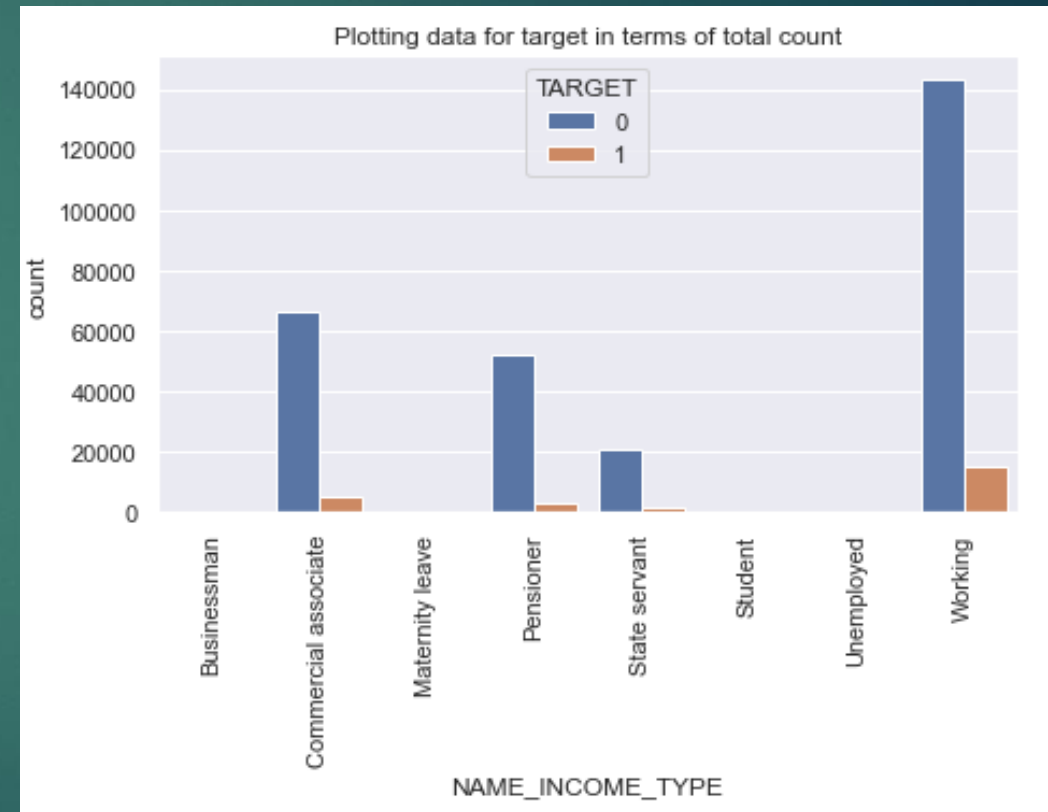
- ▶ Categorized the dataset w.r.t to the target values
- ▶ NAME_HOUSING_TYPE is good enough for the analysis and the people who have taken loan for the House/Apartment are more compared to the other types.



- ▶ Married people are likely to repay the loans
- ▶ Married people are the primary consideration for the companies to approve loans since their repaying rate is high
- ▶ Remaining categories are also showing good impact on the analysis.

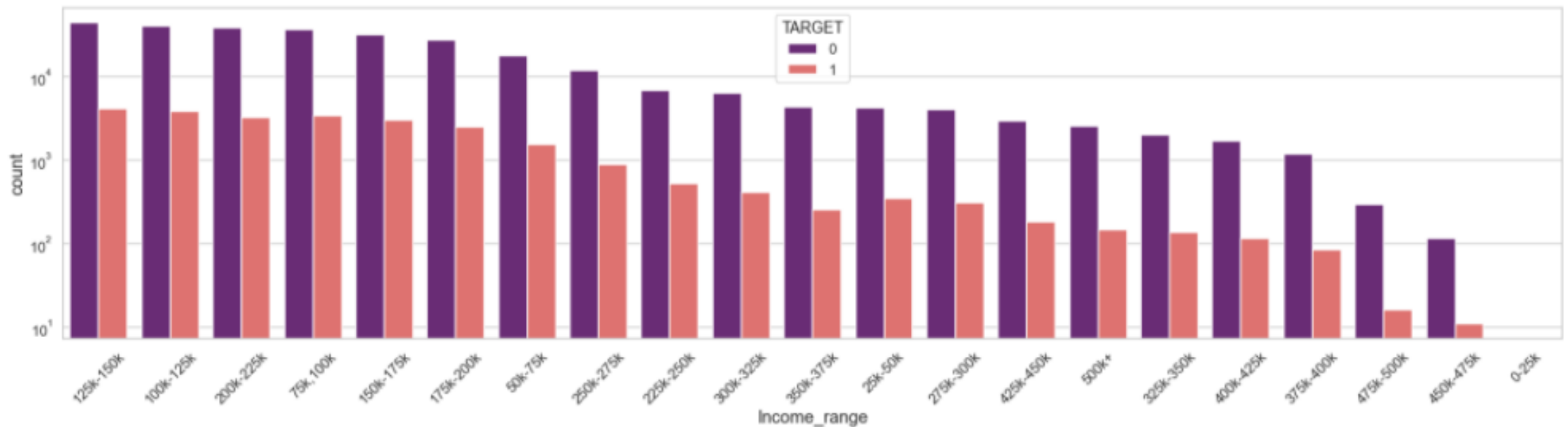


- ▶ Working people has good rate of repaying the loans followed by commercial associates and pensioners.
- ▶ Also defaulters count is more for working and the least for social servants.



- Income range played a vital role in analyzing the data ,people with less income range has repaid the loans correctly when compared with their income .

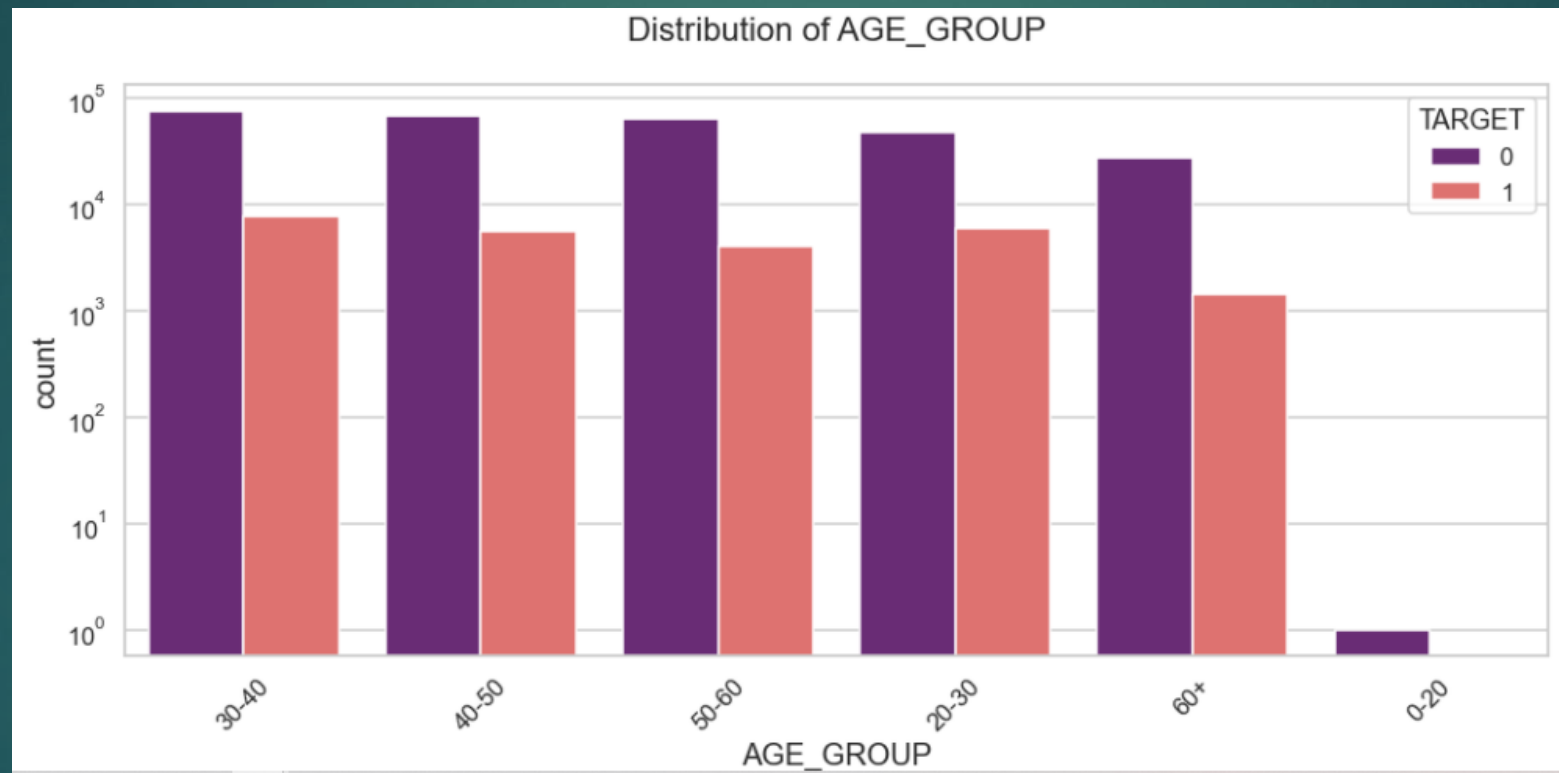
Distribution of income range



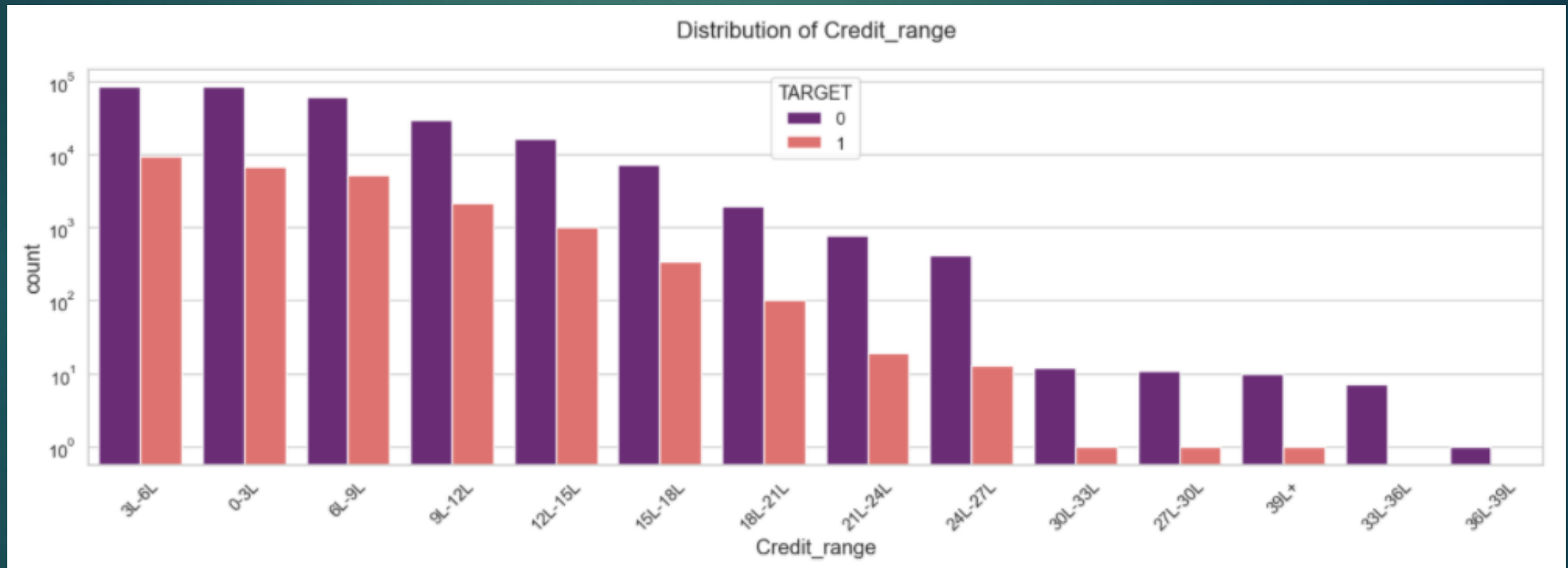
People in age group of (0-20) have taken less number of loans with a good repay rate.

Age group between (30-50) are more likely taking the loans also with

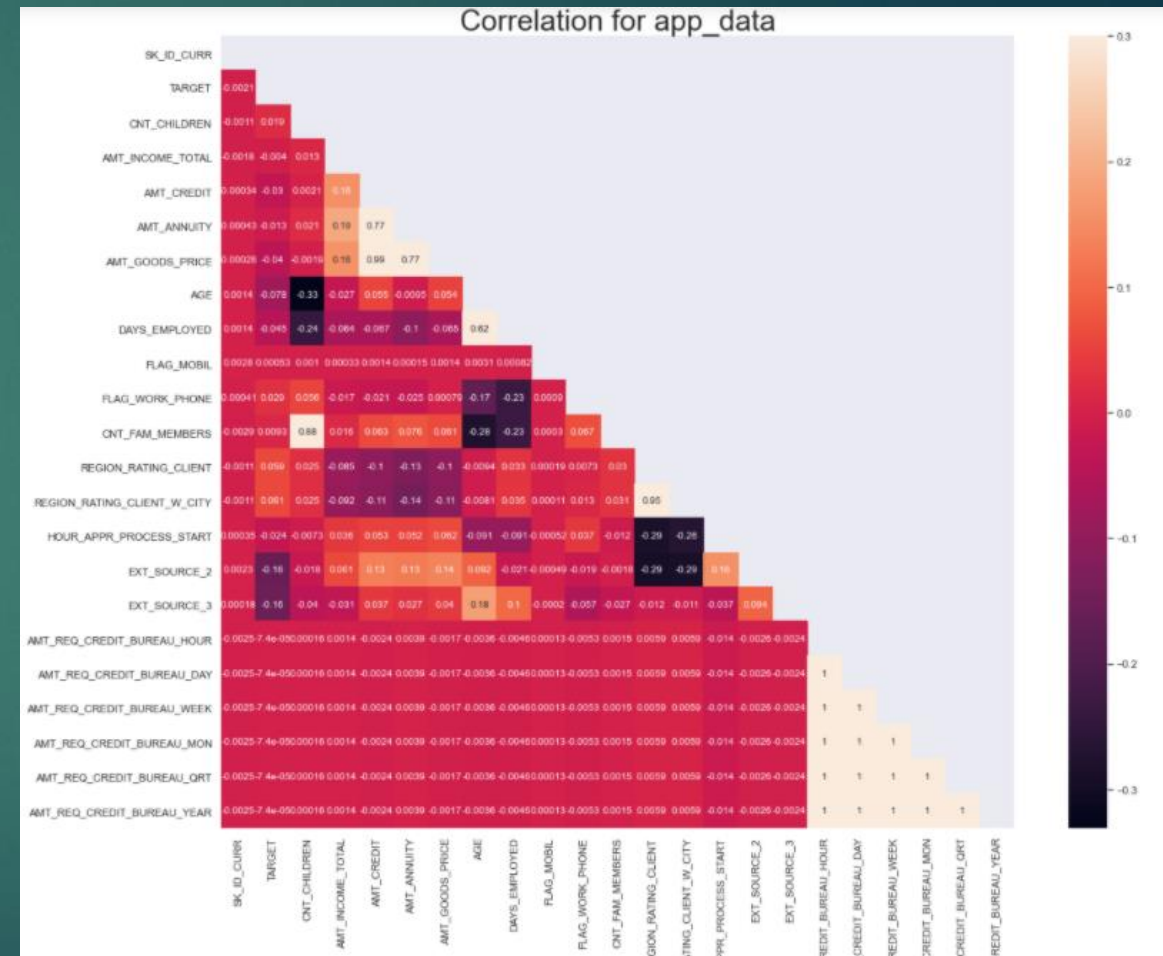
More defaulters.



- ▶ People opting the loans in range of (0-12L) are more and also more amount of defaulters are recorded in the same area.



- ▶ Heatmap plotted for correlation between the columns
- ▶ Correlation shows the strength between the columns how the values are changing w.r.t to each other either negatively or positively.
- ▶ AMT_CREDIT_HOUR, DAY, WEEK, MONTH, QTR, YEAR are positively linear correlated.
- ▶ Days_Employed and EXT_SOURCE_3 are negatively linear correlated.
- ▶ Most of the other fields have no linear correlation between them.
- ▶ Correlation matrices are similar for Tar_0 and Tar_1 dataframes.

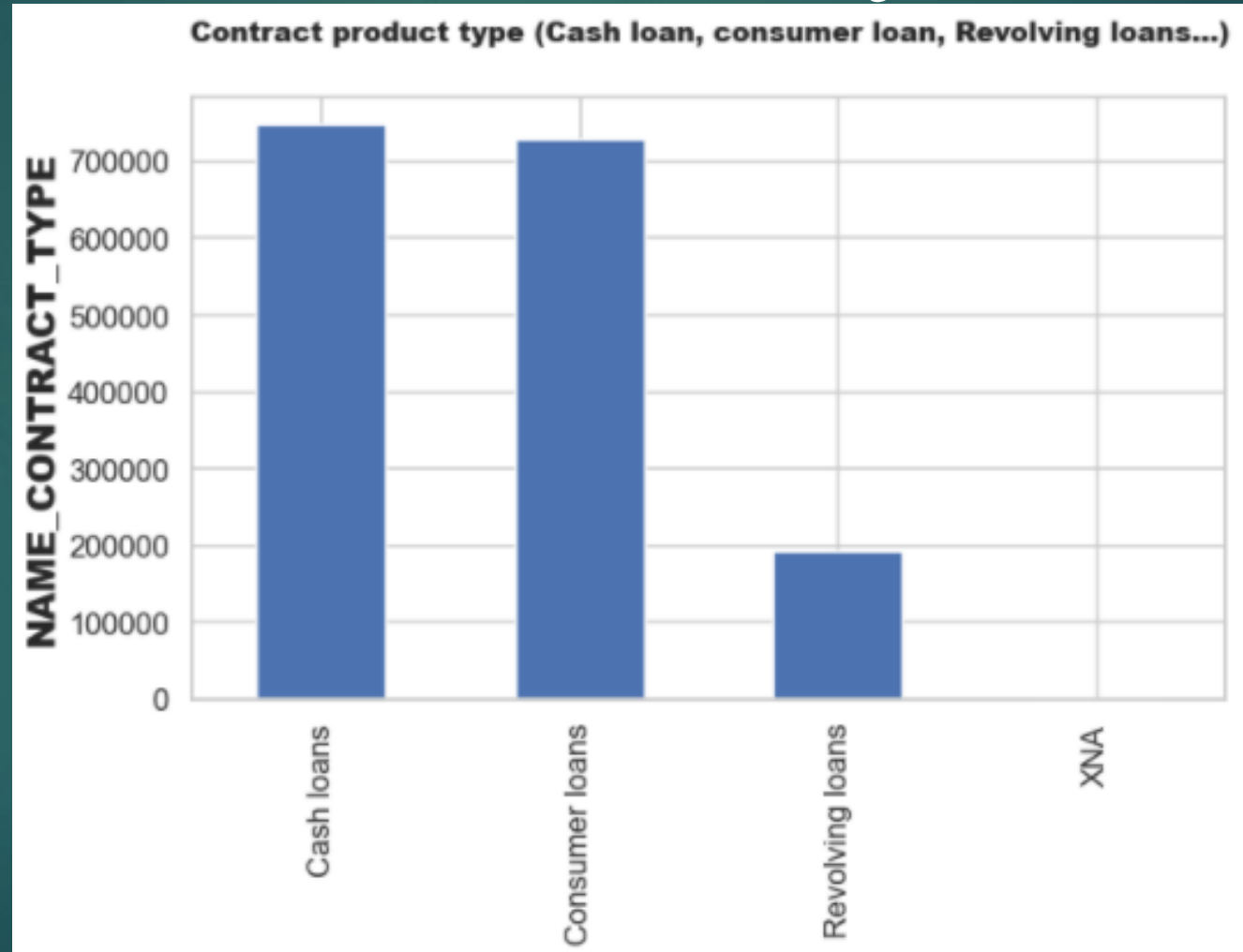


Previous Application Data

- ▶ Loaded the previous application data containing the information about the client's previous loan data whether the previous application had been **Approved, Cancelled, Refused or Unused offer**.
- ▶ Verified the data frame for more than 40% of null values and dropped the columns.
- ▶ Verified the outliers for the remaining columns plotting box plots
- ▶ Imputing the categorical columns with mode and numerical columns with mean or median depending on the area of outliers

Categorical Analysis for NAME_CONTRACT_TYPE

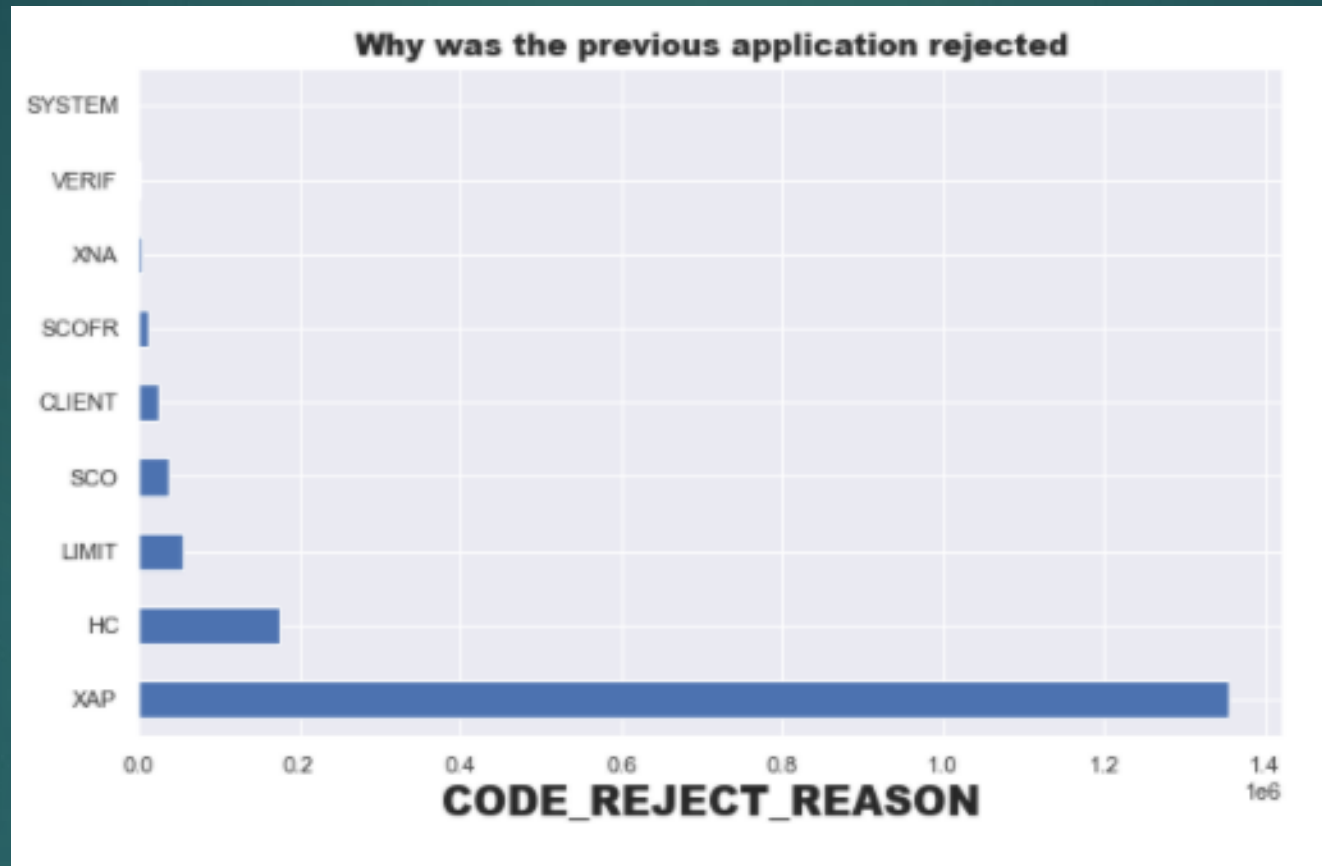
Applications for Cash loans are more where revolving loans comes the least





Inferences for the contract status:

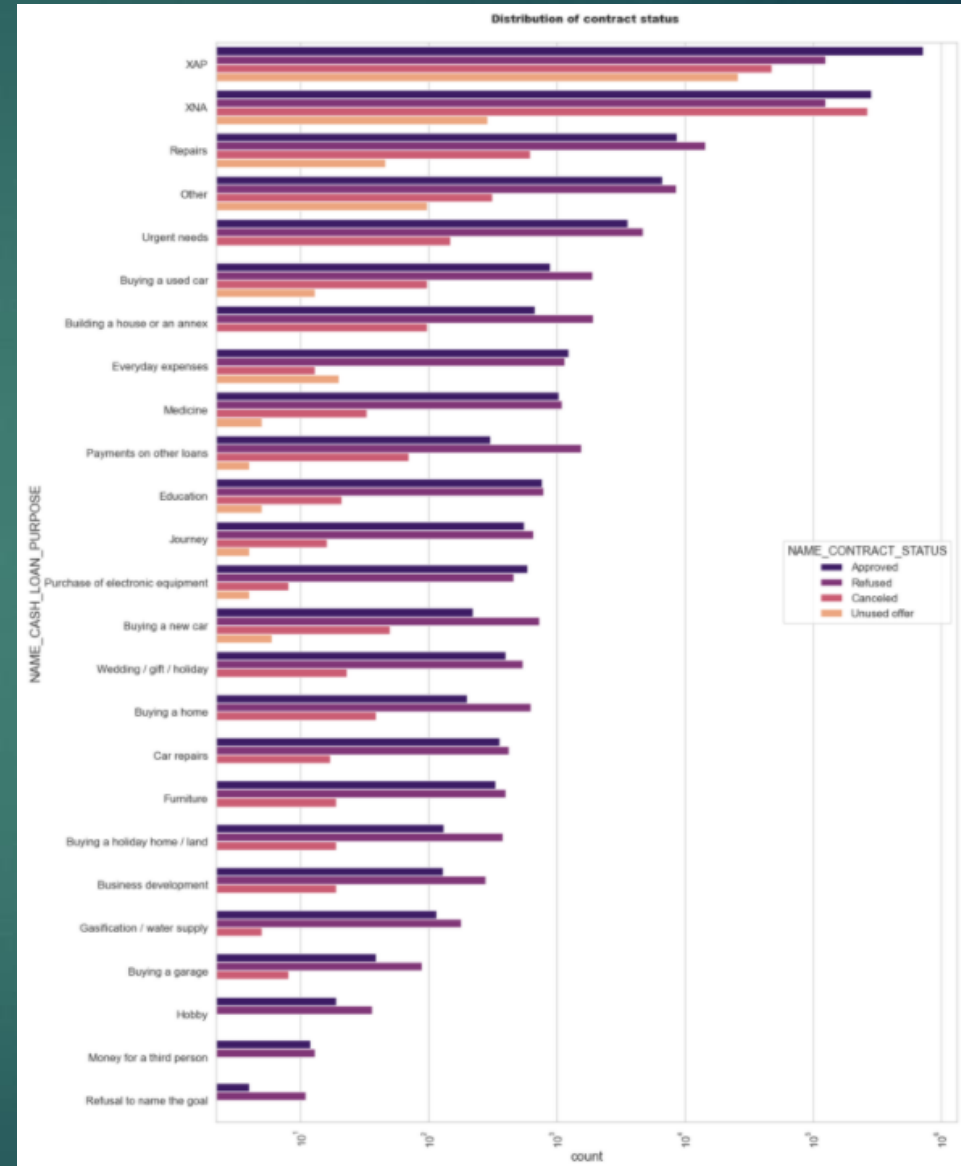
Contract status is approved for most of the applications and also there are very less unused offers



Reason for previous application getting rejected for the consumers most of the times is XAP.

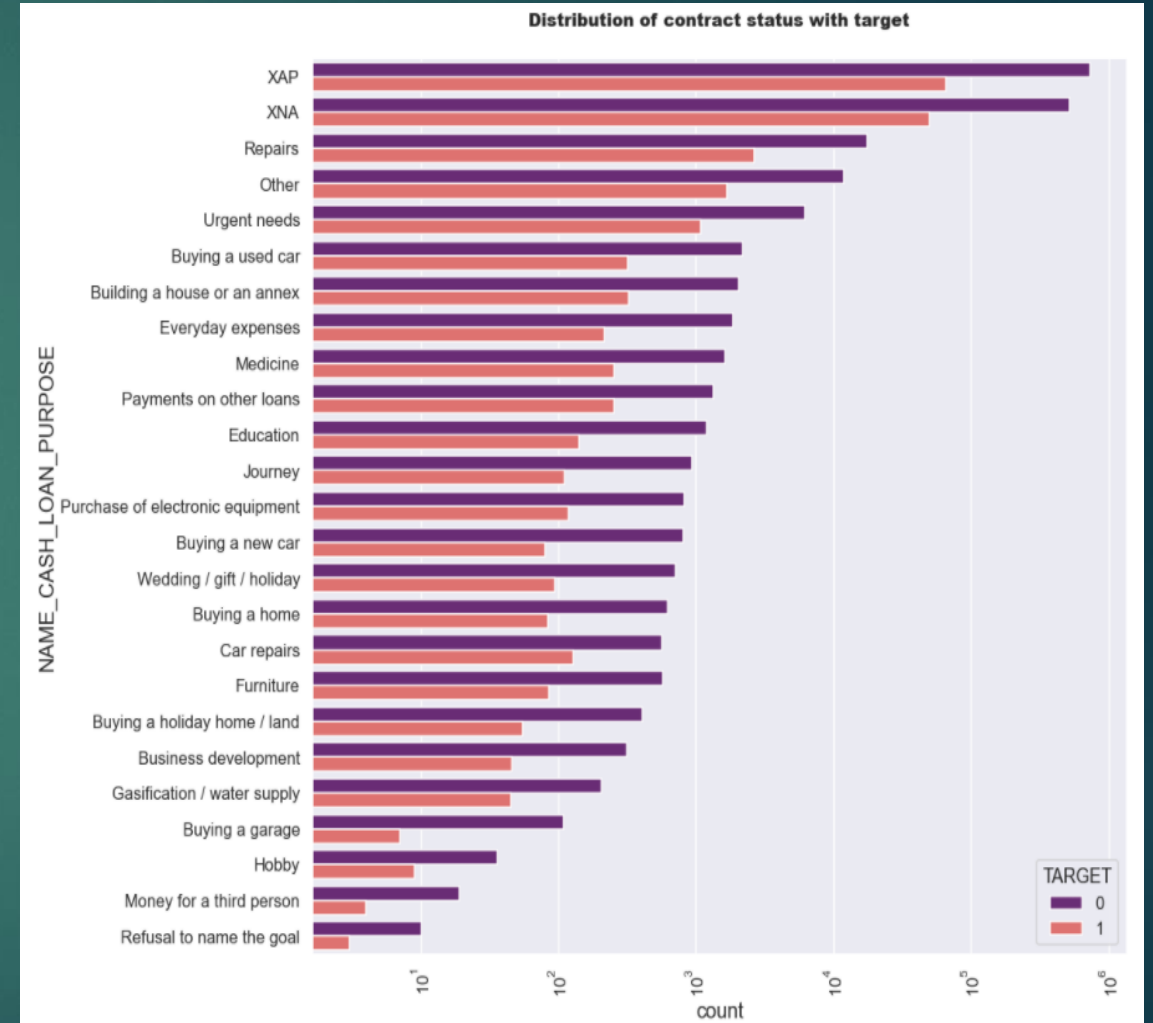
Bivariate Analysis of Cash loan purpose and Name contract Status

Most of the people have not shared the purpose of the loan which is encoded as XNA or XAP. Most common reason for rejection of the loans is repairs. Education purposes have equal number of approvals and rejection. Buying a home and buying a new car are having significant higher rejection than approvals.



Bivariate Analysis between cash loan purpose and Target

Loan purposes with repair are facing more difficulties in payment on time. There are few places where graph of loan payment on time is significantly higher than clients who are facing difficulties to pay loan such as 'Buying a garage', 'Business development', 'Buying land', 'Buying a new car', 'Journey' and 'Education'. Hence bank should focus on these purposes for which the client is having minimal payment difficulties.



Recommendations

- From the inferences following are the recommendations suggested for the banking and financial companies.
- Defaulters are more in the area of cash loans so bank should start giving revolving loans
- Loans which are previously Refused or cancelled have higher default rates.
- People who are working opted more for the loans where as state servants less ,bank should concentrate in these areas,
- Bank approves more loans for Females.