Exploratory Data Analysis : Problem Statement

- The purpose of this analysis is to identify the driving factors, variables or group of variables that influence the repayment of loans.
- My job is to find out the variables in the datasets that has correlation or causation with the TARGET variable.
- Overall purpose is to find out the factors that leads to loan default in order to take better data-driven decisions while approving loans, to minimize business loss either in the form of loan default or not approving loans to eligible applicants.

Assumptions:

- There are columns in the dataset which are not clear in meaning and it is difficult to decipher any purpose of those columns in the analysis. Hence I have decided to drop those columns from the datasets.
- It is possible that an industry expert might extract some meaning or purpose out of those columns and use them in the analysis. However, taking the benefit of doubt I have not used these columns in the analysis.

Approach & Methodology:

- I have taken a bottom-up approach in this analysis.
- I have taken one dataset at a time and did my analysis on the dataset by loading the data, understanding it, cleaning it, handling the null values and segmenting the columns based on the nature of columns.
- Initially I have checked for data imbalance specially in the TARGET variable. I have found data is imbalanced and skewed for the variables like TARGET and Gender.
- I have standardized the data wherever necessary and dropped columns having more than 40% null values and those columns that I considered to be non-influencial in the analysis.
- I have segmented the columns into numerical and categorical columns based on the number of unique values taking as 30 as threshold. Although I had to make exceptions for a few columns.

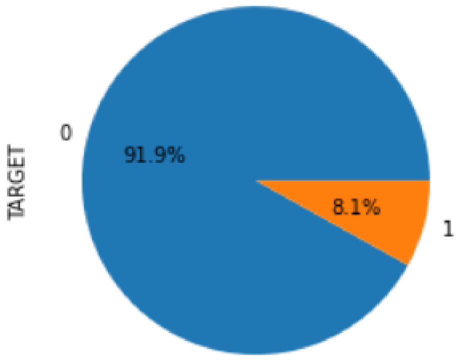
Approach & Methodology:

- I have analyzed the two groups of columns individually as Univariate analysis and also against each other as Bivariate analysis.
- I have grouped the dataset by the categorical columns and taken the mean or median of the numerical column against each categorical column. For e.g. Gender vs Income analysis.
- I have done Univariate Analysis of each categorical and numerical variable and analyzed it’s impact on the TARGET variable by taking the mean of the TARGET variable.
- In the numerical Univariate analysis I have identified the outliers by using box plots and describe function and capped the outliers wherever necessary to ignore the outliers.
- I have used binning technique to divide the numerical column into groups and plotted the mean of the TARGET variable as segmented univariate analysis.
- I have used scatterplots, pairplots and boxplots in the Bivariate analyses.

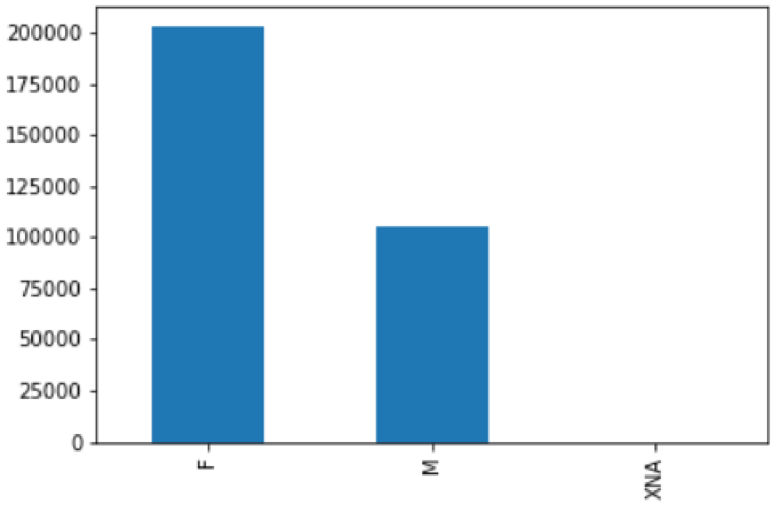
Approach & Analysis:

- I have used stacked bar chart to analyze the influence of categorical variables on loan approval, cancellation, refusal or unused in the previous application dataset.
- In the end I have merged the 2 datasets and did a Bivariate analysis of the new columns added to the application data from the previous data, on the TARGET variable.
- I have used HeatMap to do a Multivariate analysis to identify the correlation in the combined dataset.

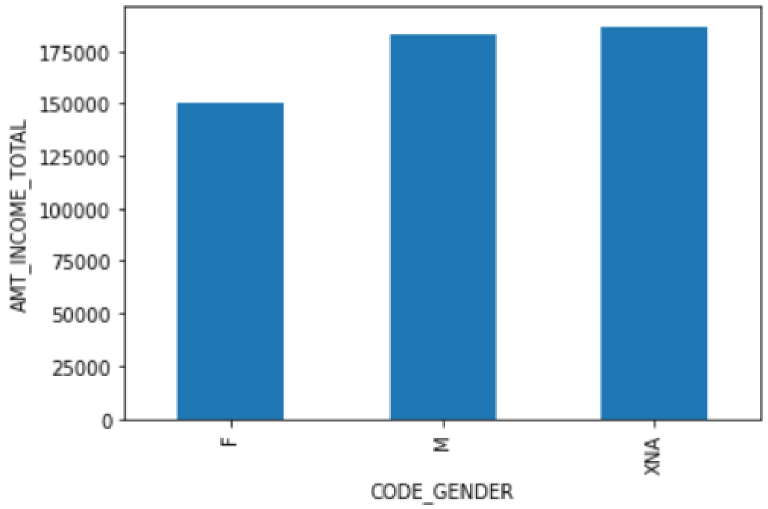
Plots & Insights:



We identify data imbalance in TARGET variable

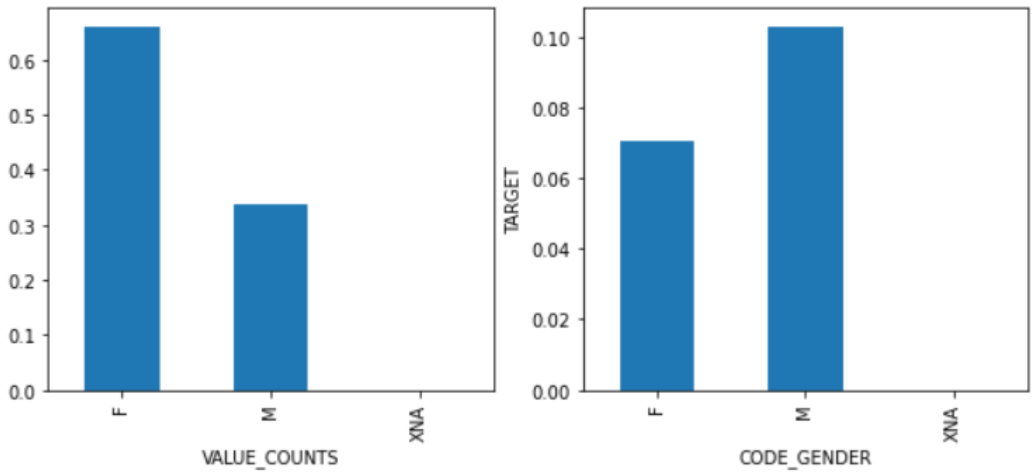


We see another data imbalance in gender. Records of males are 50% of that of females.

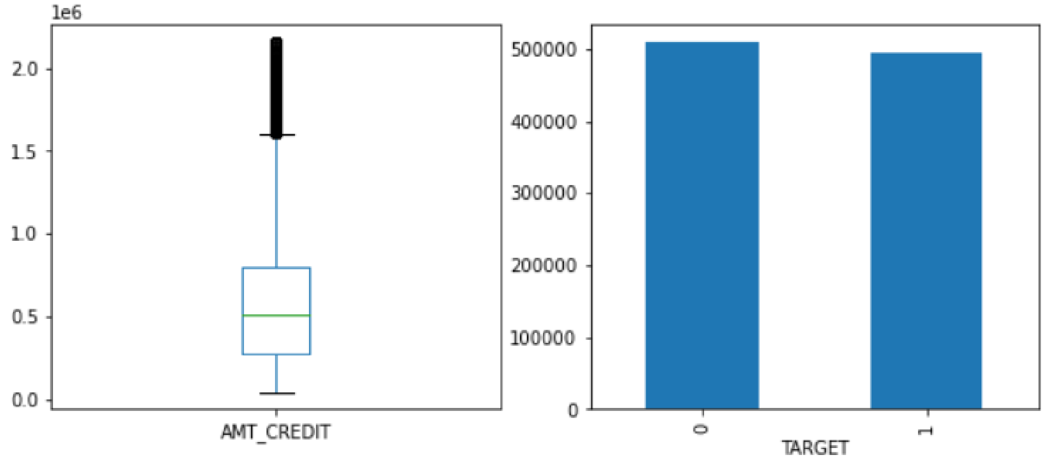


Males have a higher mean income than Females.

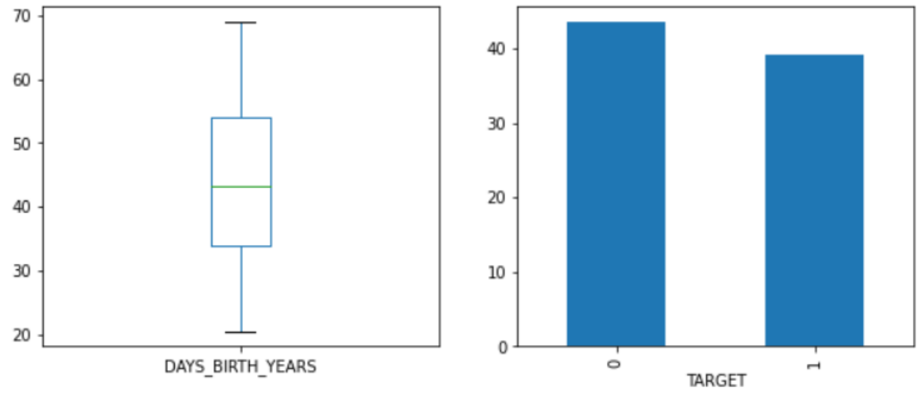
Plots & Insights:



Males have a higher rate of default

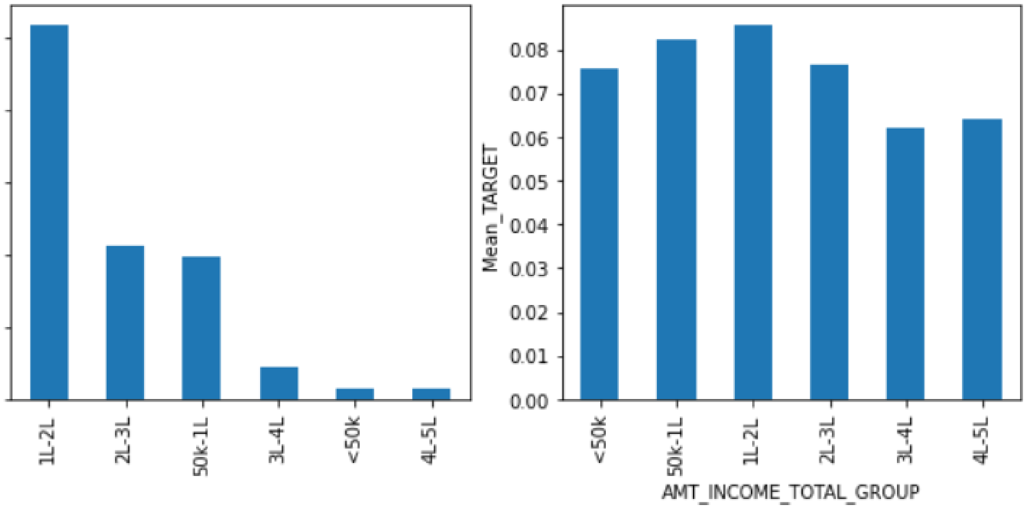


Median Credit amount of loans that have defaulted is slightly lower.

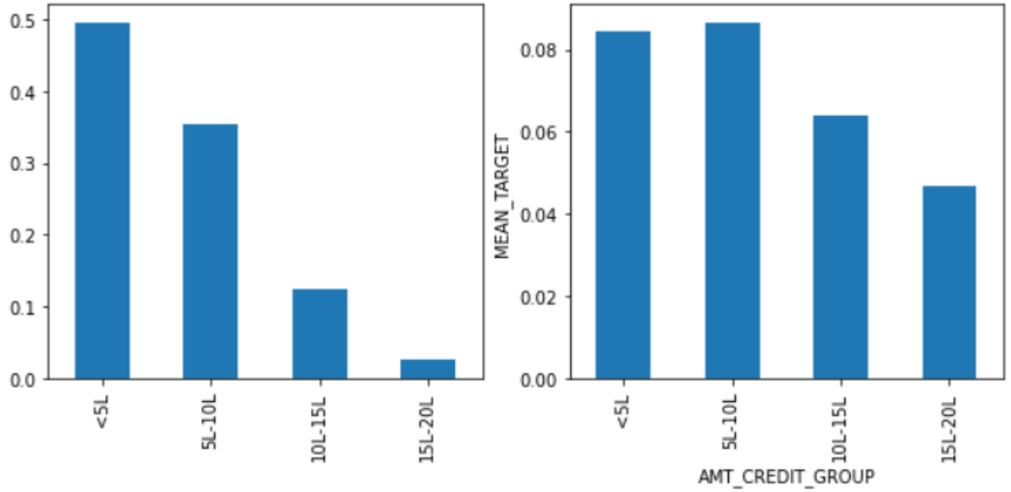


Median age of default customers is lower

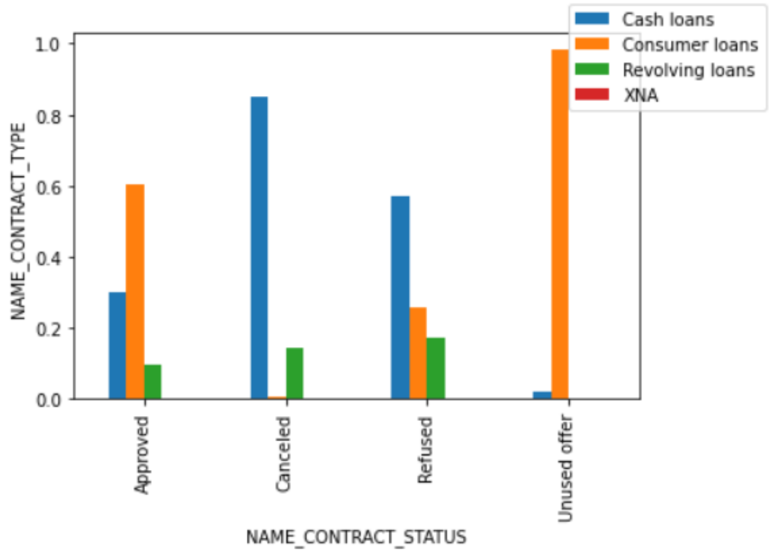
Plots & Insights:



1L-2L Income group have highest rate of default

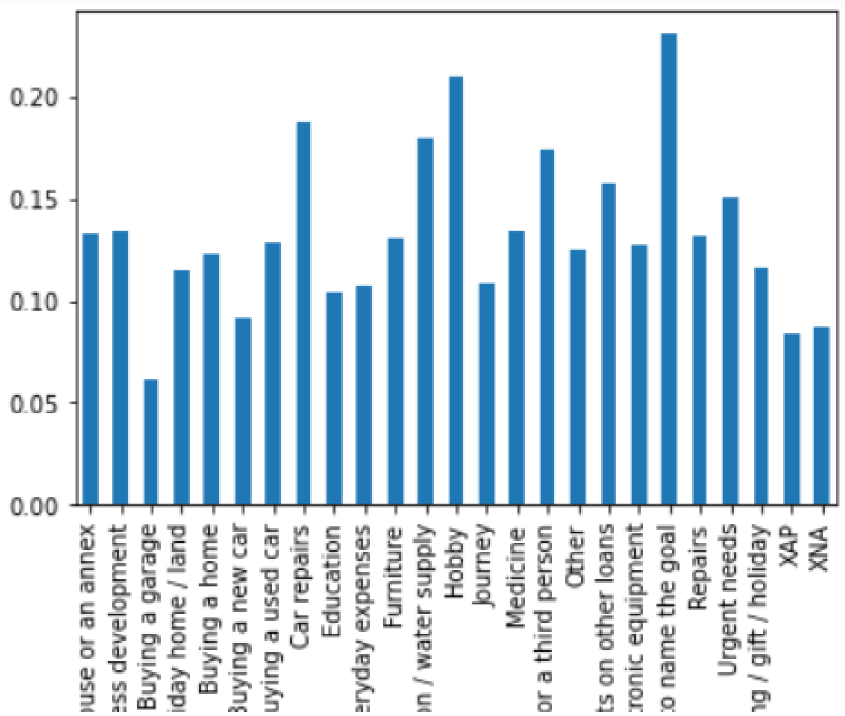


5L-10L credit bucket has highest default rate.



Consumer Loans have highest rate of approval

Plots & Insights:



Applicants who have refused to provide a reason for taking loan or customers who have taken loans to pursue their hobbies have highest rate of default.

Recommendations:

- According to the analysis done I would recommend to forward more loans to females than males as this analysis have shown higher success rate with females in loan repayment than males.
- Consumer loans have a higher rate of approval and successful repayment. Hence I would suggest to focus on consumer loans to increase business.
- For those customers who are refusing to give a reason for taking the loan and those applicants who are taking loans to pursue their hobbies, there should be more scrutiny and mostly in these cases loans should be refused as these cases have a very high rate of default.
- Businessmen and Car dealers have a very good success rate in loan repayment. Hence I would recommend these are hot candidates for loan approval.