CREDIT EDA ASSIGNMENT

Muniba Naushad

“NO DATA IS CLEAN,
BUT MOST IS USEFUL.”

PROBLEM STATEMENT

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile.

Two types of risks are associated with the bank's decision:

- 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.

OBJECTIVE

To identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc**. through different driver variables behind loan default, i.e. the variables which are strong indicators of default.**

To use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

The company can utilize this knowledge for its portfolio and risk assessment. This will ensure that the consumers capable of repaying the loan are not rejected.

DATA

The data contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,

All other cases: All other cases when the payment is paid on time.

DATA UNDERSTANDING

This dataset has 3 files as explained below:

1. 'application_data.csv' contains all the information of the client at the time of application.
The data is about whether a **client has payment difficulties.**

2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data on whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

3. 'columns_description.csv' is data dictionary which describes the meaning of the variables.

APPROACH

Data Cleaning

- Columns having more than 45% missing values were removed.
- Then columns that seemed irrelevant for analysis were identified and dropped.

Missing Values Treatment

- Category type columns were imputed with most frequent values.
- Numerical columns were imputed with median as the outliers were detected.

Outliers

- Outliers were detected , in some of the columns, outliers could be present due to data mis-entry and would require further analysis.
- Suggesting, to treat with the upper or lower limit values in numerical column types.

APPROACH

Data Standardization

- Ensuring all observations under one variable are expressed in a common and consistent unit.
- Some of the observational columns like flag type with 0 and 1 values were converted to categorical type to be represented as ‘yes’ and ‘no’ for better visualization.
- Some of the columns with negative values were converted to positive and standardized across all columns.

Binning

- Created bins/range for certain columns like Age, Income Total, Family status for better visualization and reaching meaningful insights.

APPROACH

Data Cleaning

Columns having more than 45% missing values were removed.

Then columns that seemed irrerelevant for analysis were identified and dropped.

Missing Values Treatment

Category type columns were imputed with most frequent values.

Numerical columns were imputed with median as the outliers were detected.

Outliers

Outliers are replaced with the upper or lower limit values in numerical column types.

ANALYSIS

Univariate Analysis

a) Used count plot to check frequency distribution of categorical variables.

b) Used box plot to see distribution of numerical variables.

Bivariate Analysis

a) Used pair plot for numerical values relational distribution.

b) Used bar graph for by grouping on categorical variable and aggregating numerical variable.

c) For both categorical variables

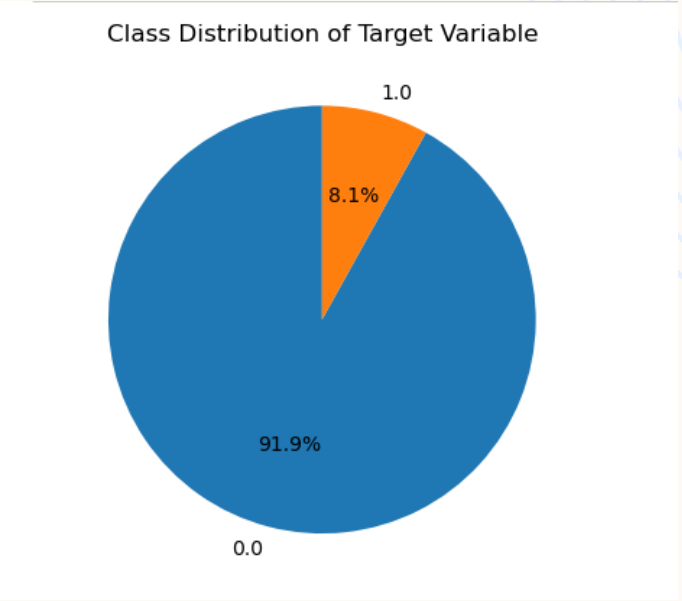
Created barplot with hue on TARGET variable to see which category faces more difficulty in payment.

Multivariate Analysis

Used heat map by creating pivot table for 2 columns and aggregating target on the category.

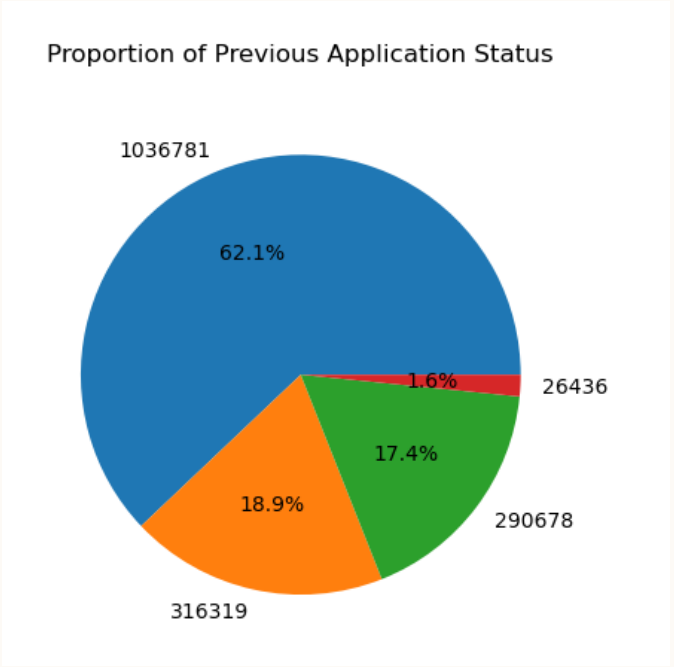
INSIGHTS

- Target data is not balanced, only 8% clients are shown who defaults.



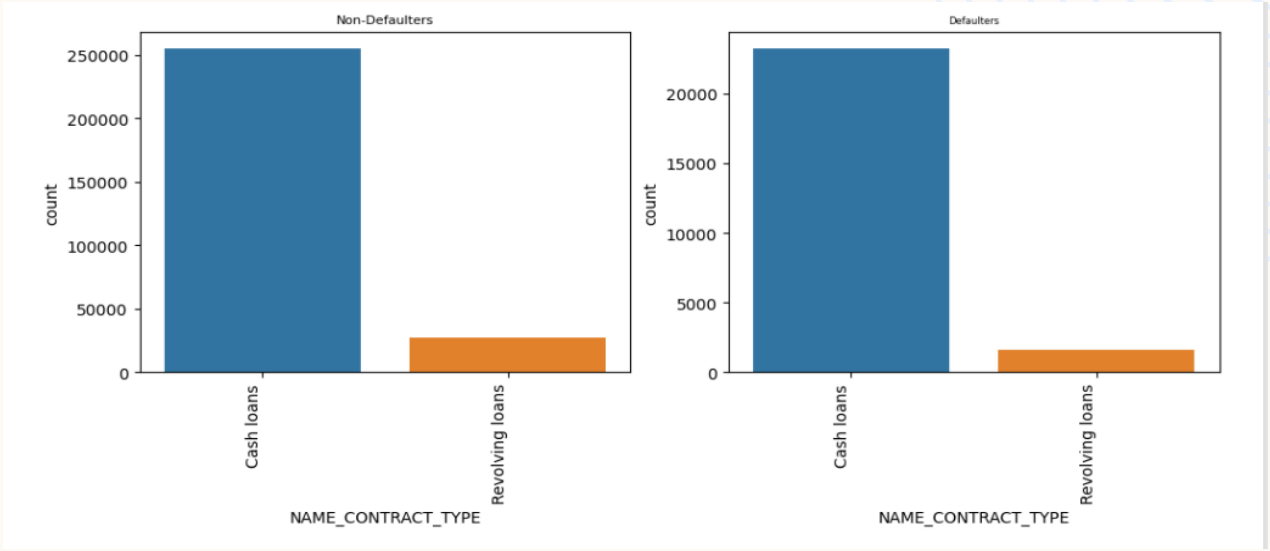
Data Imbalance Ratio – approximately 11%

- Almost 62% applications are approved in previous application data and 18% and 17% are Cancelled and refused respectively.

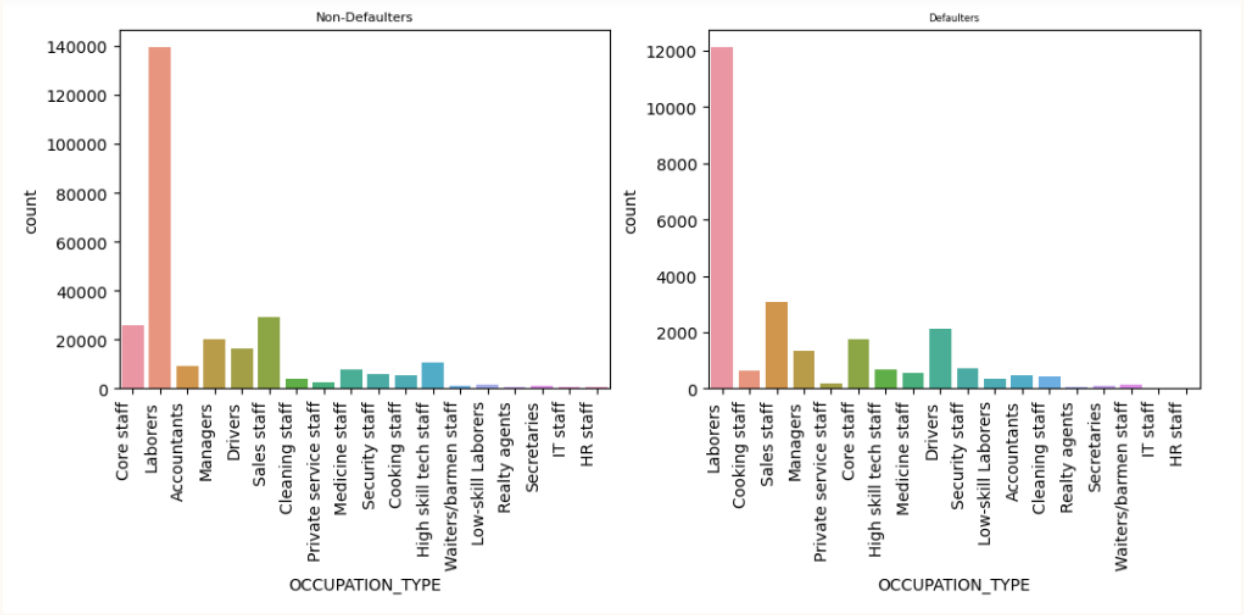


INSIGHTS

- Cash loan Contract type clients are majorly doing defaults compared to Revolving loan type

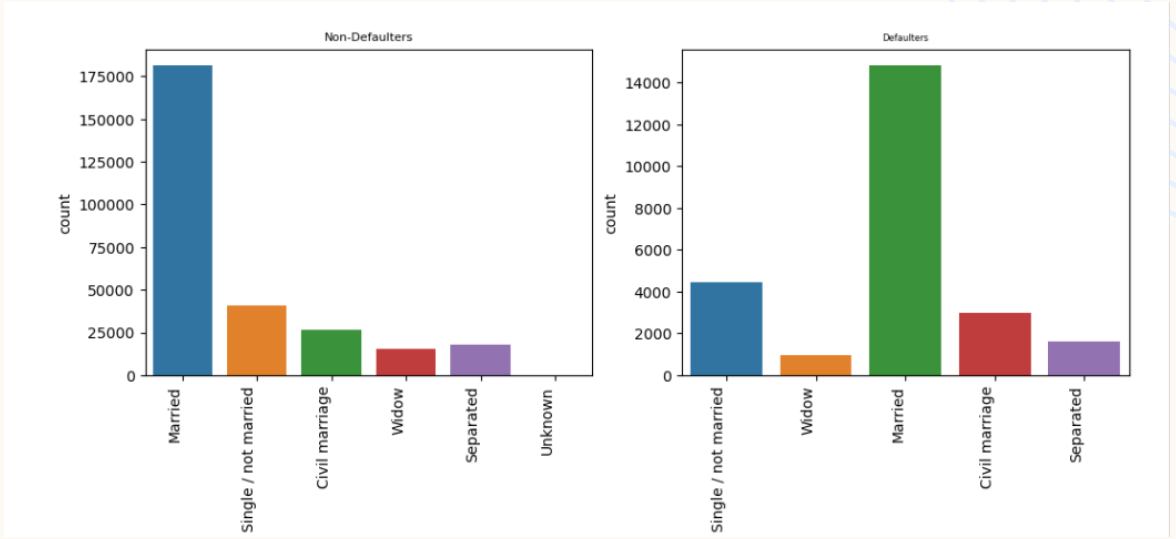


- Unskilled citizens or laborer are taking more loans than skilled clients like Managers, Accountants etc.

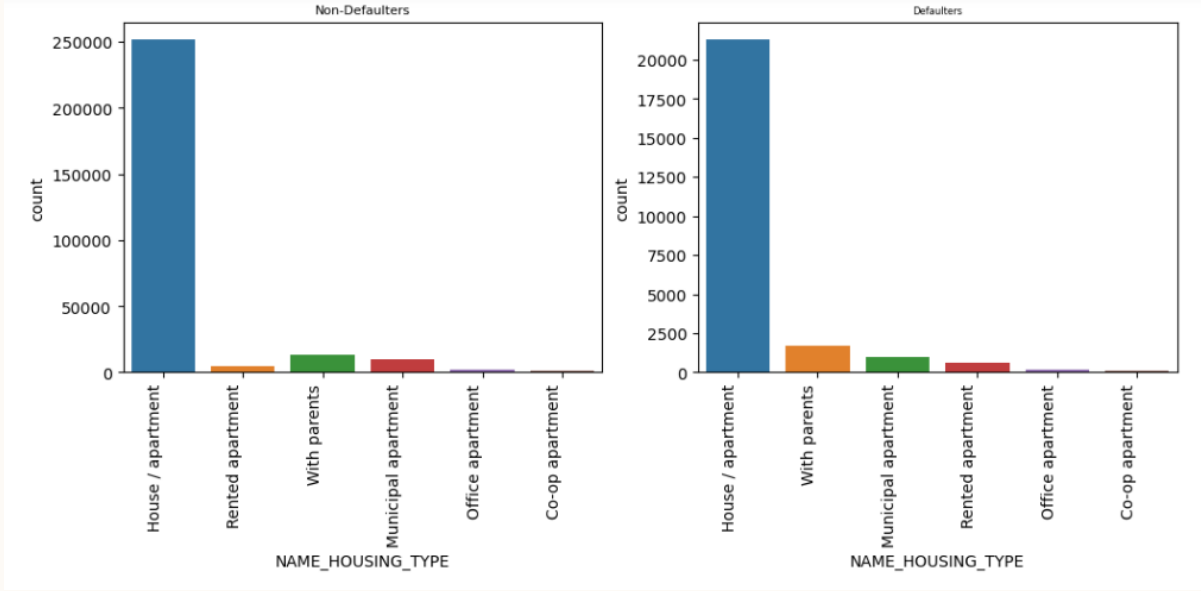


INSIGHTS

- Married clients are doing more default compared to Single or Unmarried.

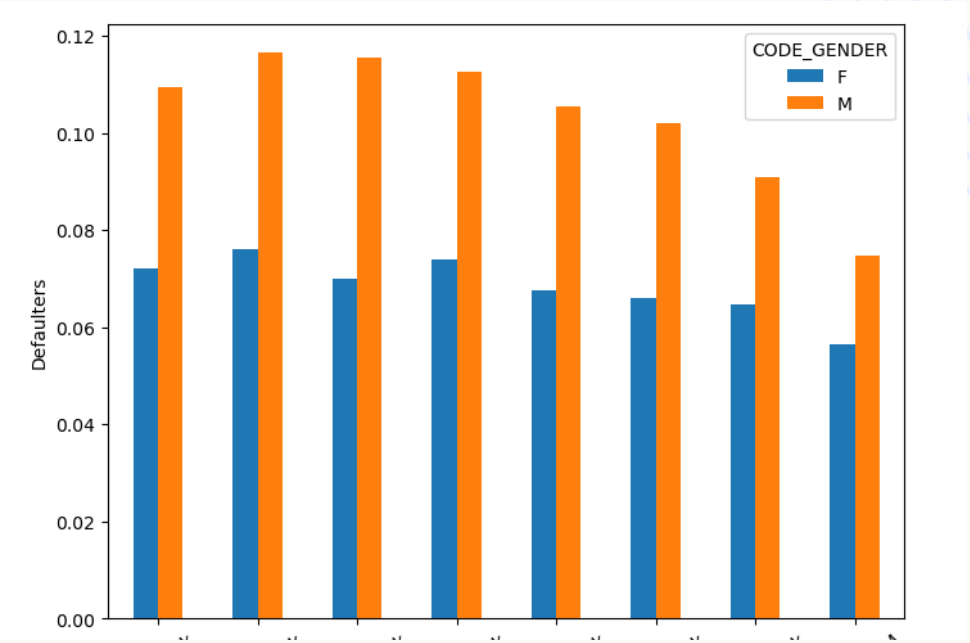


- Clients who owns house/apartment are doing more default compared to other categories

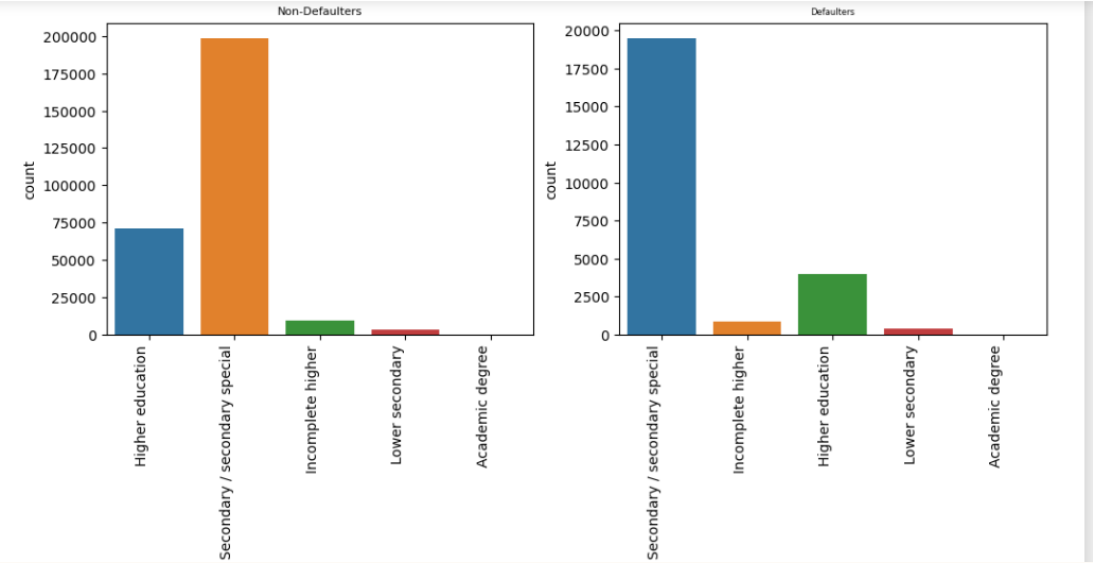


INSIGHTS

- Male candidates are facing more payment difficulty as compared to females.

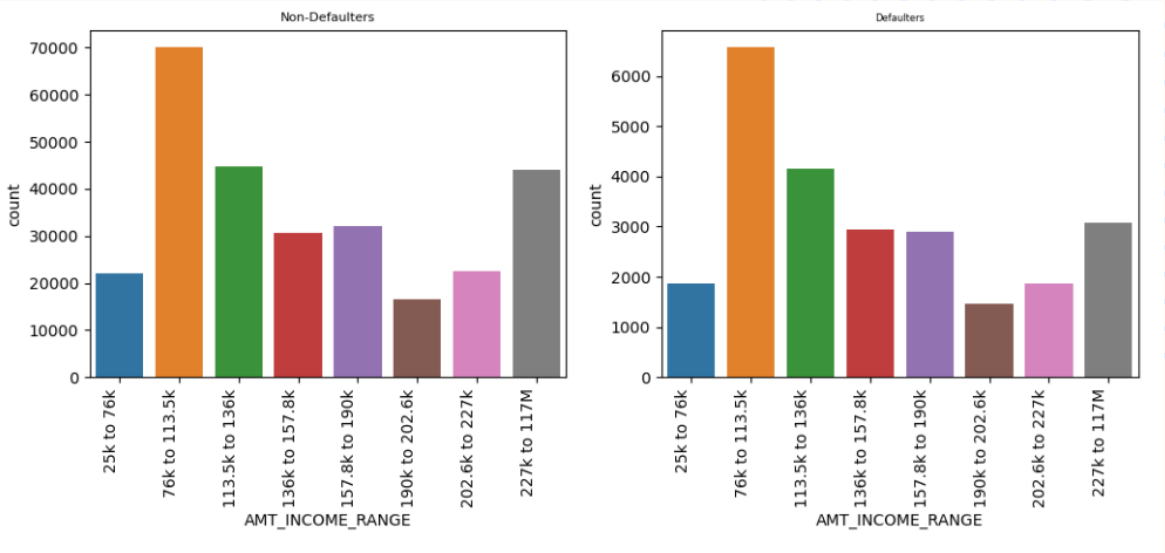


- Secondary Educated type clients are the maximum who have payment difficulties Clients who owns house/apartment are doing more default compared to other categories

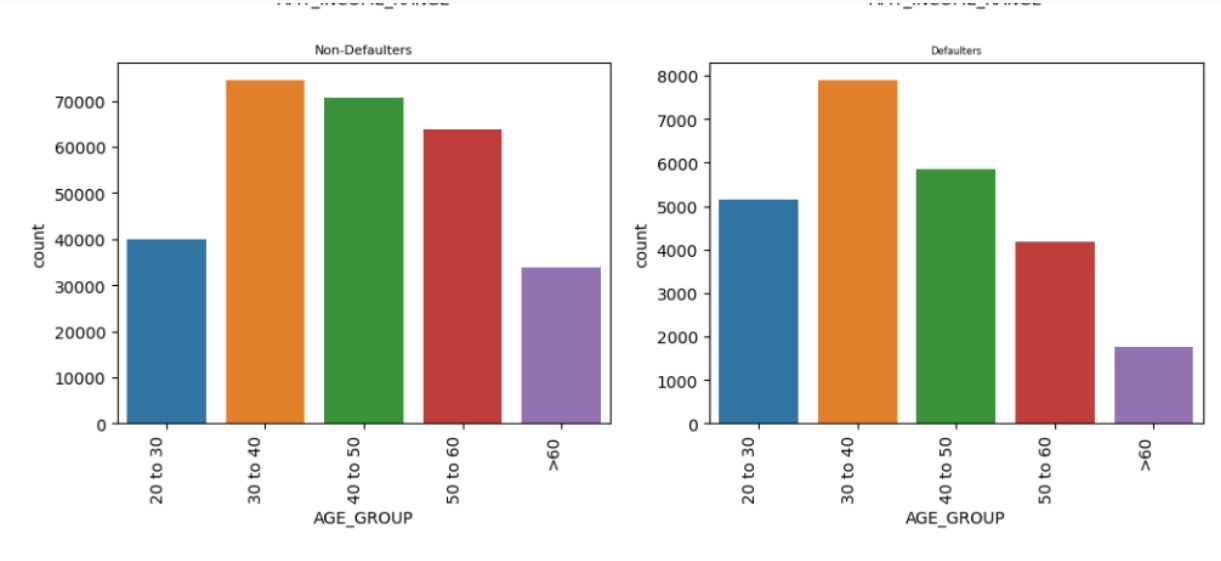


INSIGHTS

- Clients with income range medium to high are more keen towards taking loans.

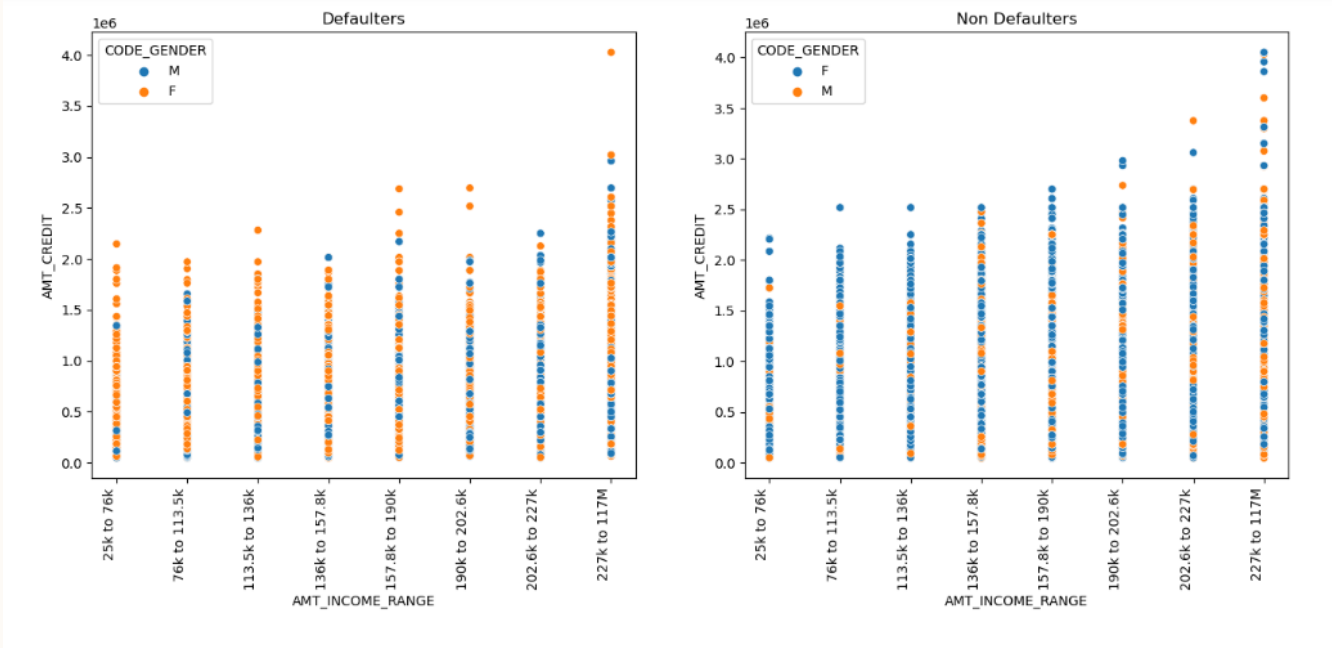


- Young clients of age group 30-40 are having more payment difficulties



INSIGHTS

- We observed that amount credit or loan is increased as the income increased, for both the genders



SUMMARY

Factors indicating possibility of loan default:-

Education level : Less educated people are more likely to default their loan.

Family Status : Married citizens and citizens with more than 2 children, defaulted more in payments due to higher responsibility points.

Occupation : Unemployed, unskilled staff like labors, cleaning staff and people with less stable jobs are more likely to default.

Loan history: Applicant whose previous loans are refused/cancelled are more likely to default their loan.

Loan frequency: Frequent borrowers are also more likely to default

THANK YOU