CREDIT EDA ASSIGNMENT ON BANK LOAN DATA

Krithika Murugesan

PROBLEM STATEMENT

The given problem statement is

To give an idea of applying EDA in a real business scenario.

We will also be able to develop our basic understanding of risk analytics in banking and financial services.

Understand how the data is used to minimize the risk of losing money when lending to customers.

BUSINESS UNDERSTANDING

The loan providing companies find it hard to give loans to the people due to their insufficient or non existent credit history.

We have 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 data given 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 installments of the loan.

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

BUSINESS UNDERSTANDING(CONTINUED…)

When a client applies for a loan, there are four types of decisions that could be taken by the client or company.

Approved:
The company has approved loan application.

Cancelled:
the client cancelled the application sometime during approval. Either the client has changed her/his mind about the loan or in some cases due to higher risk of the client , he received worse pricing which he did not want .

Refused:
The company has rejected the loan (because the client doesnot meet their requirements etc..)

Unused Offer:
Loan has been cancelled by the client but at different stages of the process.

BUSINESS OBJECTIVES

Aims to identify patterns which indicate if a client has difficulty paying their installments 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..

This will ensure that the customers are capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, company wants to understand the driving factors behind loan default. Ie) the variables which are strong indicators of default.

Company can utilize this knowledge for its portfolio and risk assessment.

DATA UNDERSTANDING

This dataset has 2 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.**

APPROACH

Data Collection and Understanding:

Obtained the dataset containing relevant information about the bank loans.
Understanding the data dictionary to know the meaning of each variable.
Identifying the target variable (e.g., whether the loan was repaid or defaulted).

Data Cleaning:

Handling missing values, outliers, and duplicates.
Checking for data integrity issues, such as inconsistencies or errors in data entry.
Converting data types if necessary.

Descriptive Statistics:

Computing basic statistics (mean, median, mode, range) for key variables.
Generating summary statistics to get an overview of the dataset.

Univariate Analysis:

Analyzing the distribution of the target variable (e.g., loan status).
Exploring the distribution of key numerical variables (loan amount, interest rate) using histograms and summary statistics.
Examining the distribution of categorical variables (loan type, purpose) using bar plots.

Bivariate Analysis:

Exploring relationships between the target variable and other variables.
Using scatter plots, or correlation matrices to understand the impact of numerical variables on loan status.

Multivariate Analysis:

Exploring Interactions:

Use techniques like heatmaps to visualize interactions between multiple variables.

METHODOLOGY

Visualizations and Insights:
a. Use Seaborn and Matplotlib for Visualizations:
Create informative visualizations to communicate key insights.
Customize plots for better interpretability.
b. Document Key Findings:
Summarize your observations, patterns, and trends.
Document any data quality issues or potential areas for feature engineering.
Missing Values and Outliers:
a. Handle Missing Values:
Decide on strategies for imputing or dropping missing values.
b. Treat Outliers:
Identify and decide on methods for handling outliers.
Data Preprocessing:
a. Encoding Categorical Variables:
Convert categorical variables into numerical format if needed.
b. Feature Scaling:
Scale numerical variables if necessary.

Data Quality Checks:

a. Validate Data Integrity:

Ensure that transformed data retains its integrity and consistency.

Reporting:

a. Create a Detailed Report:

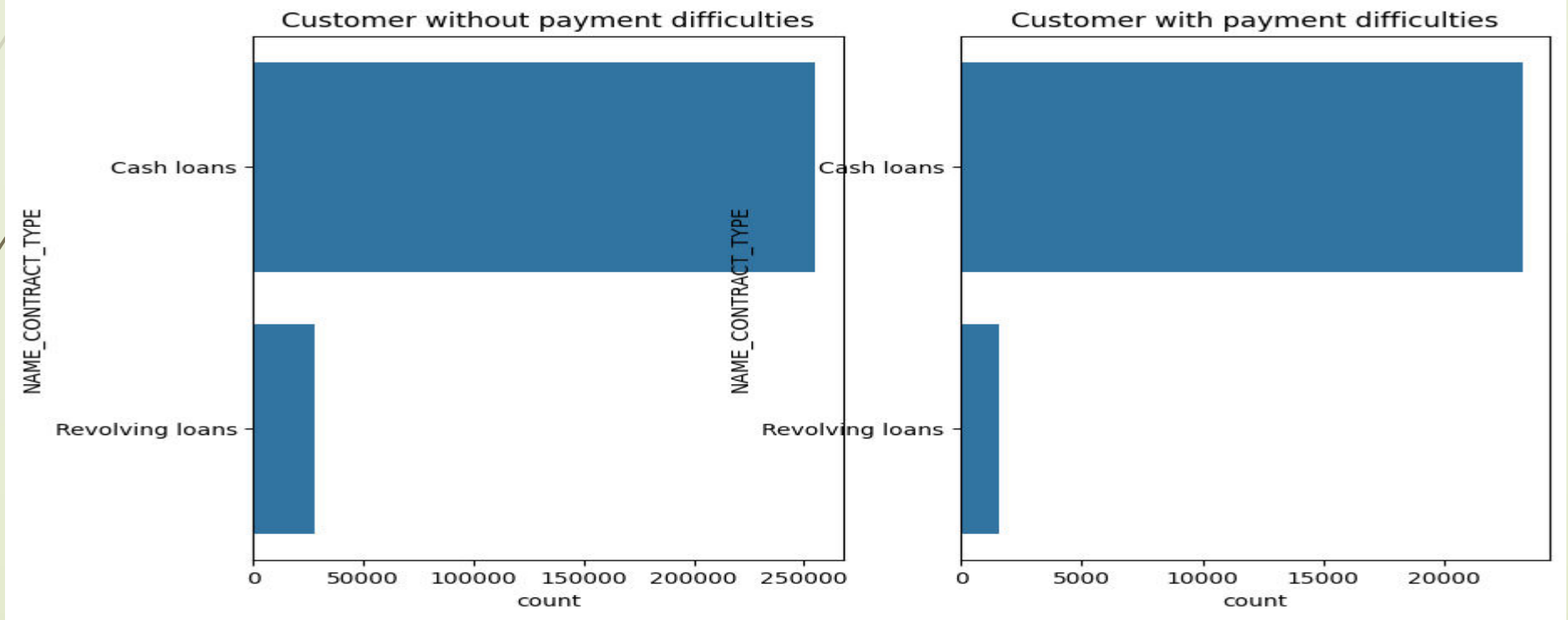
Summarize key findings, insights, and preprocessing steps.

Include visualizations and statistical measures.

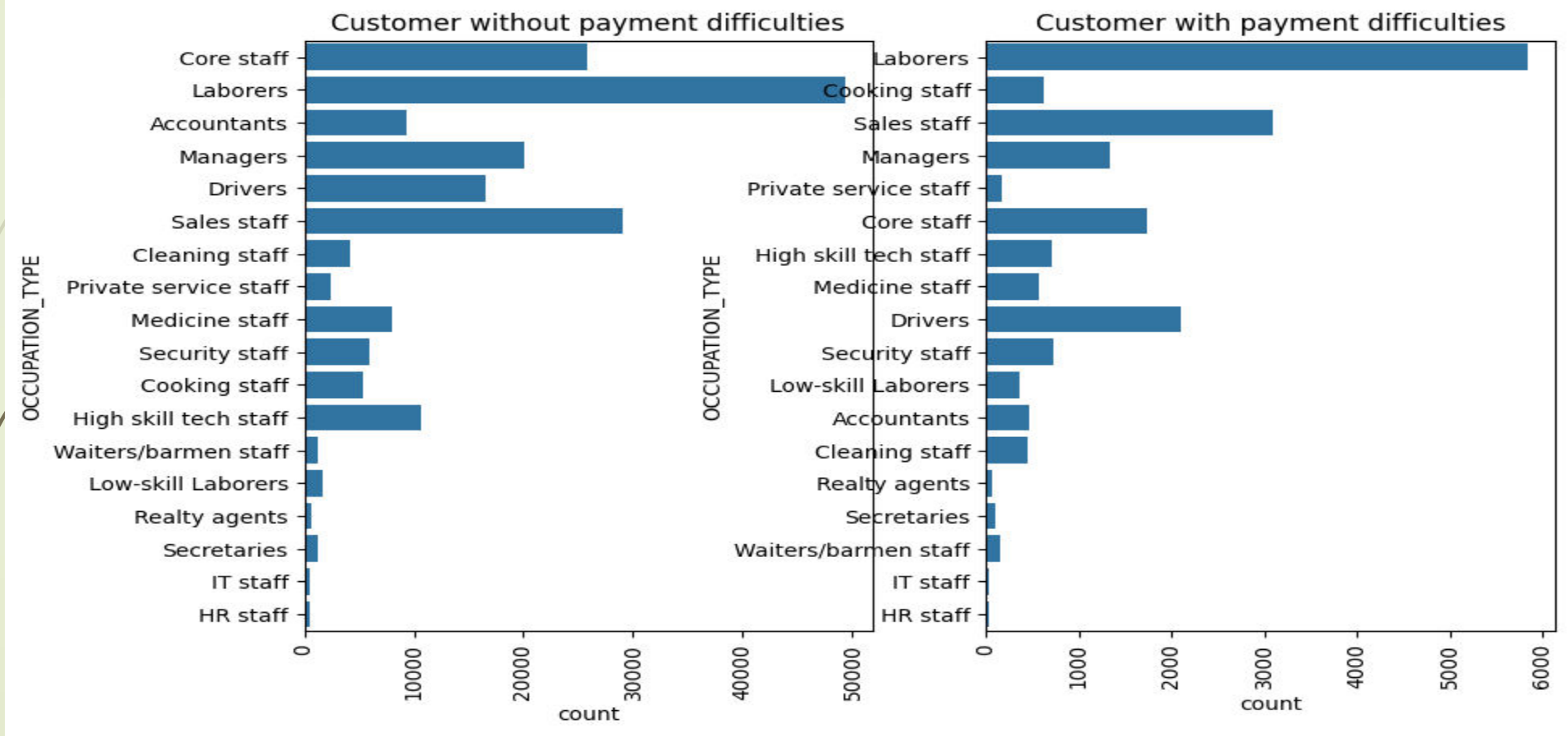
GRAPHS & INSIGHTS

A) UNIVARIATE ANALYSIS

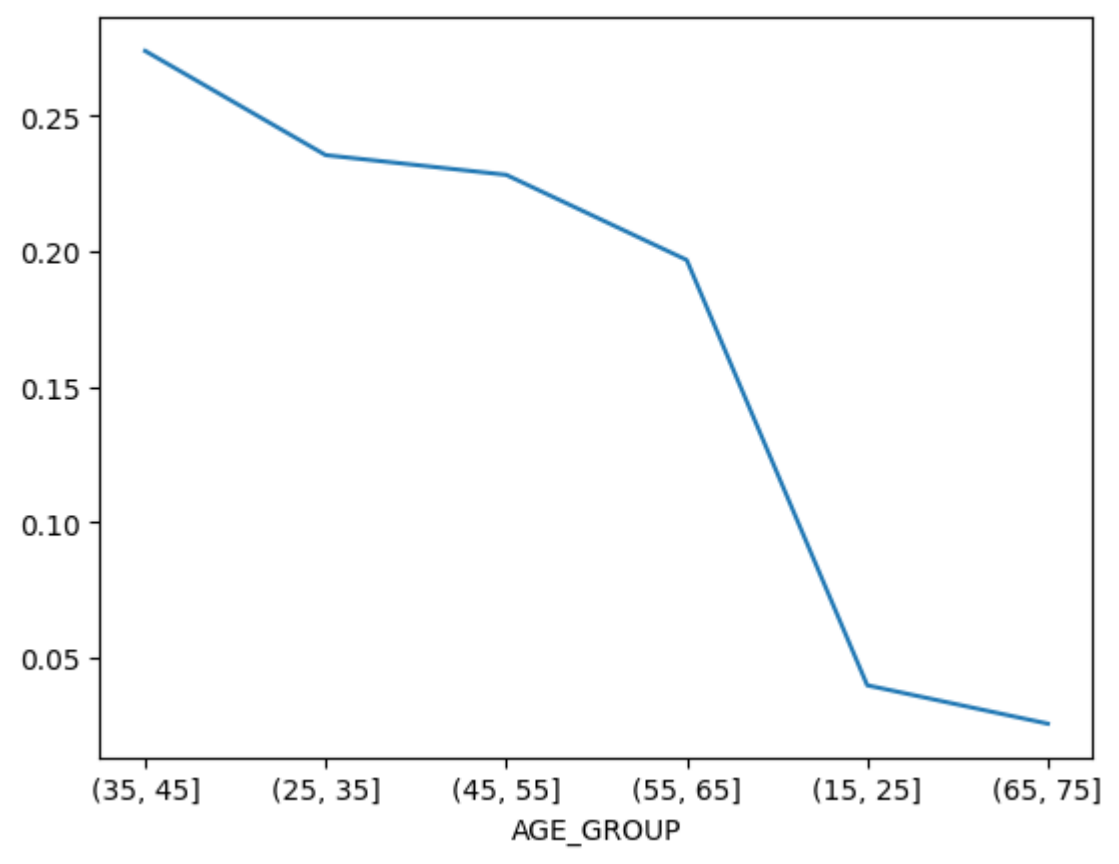
More cash loans go to default. Bank should give more Revolving loans.



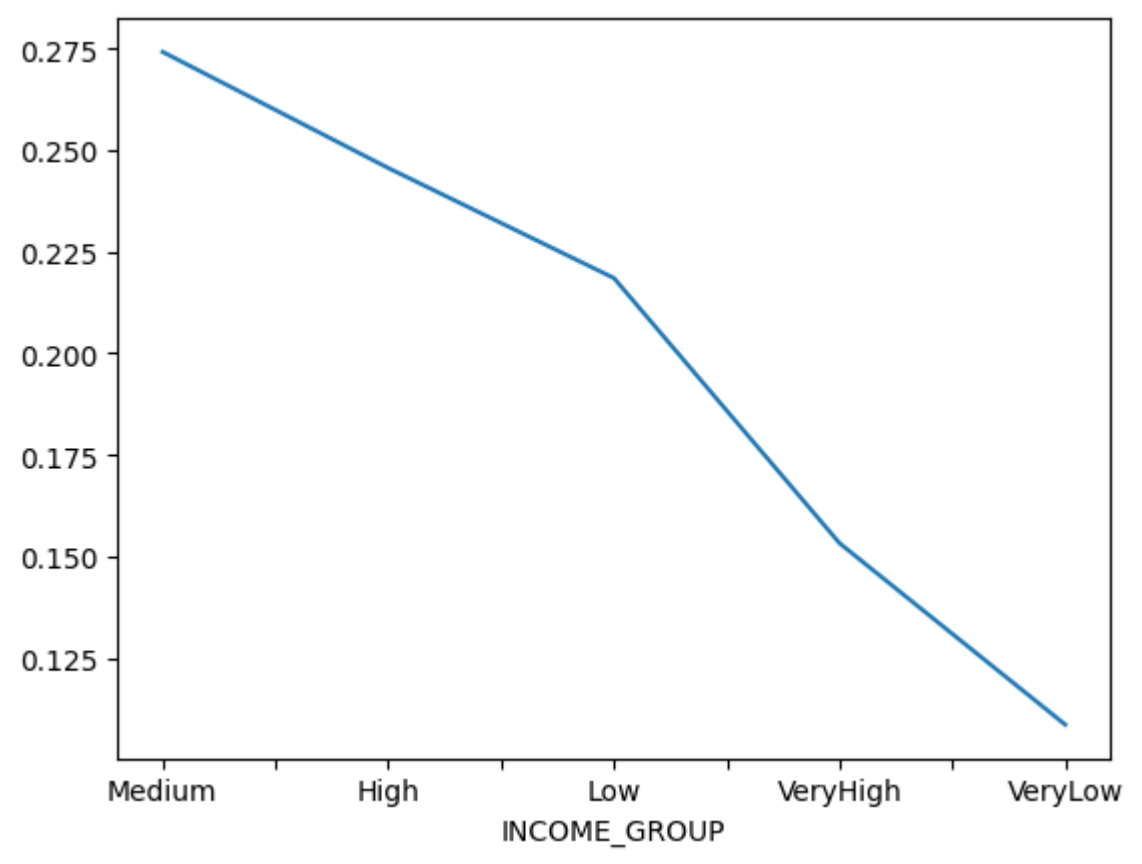
labourers are facing more difficulty in paying loans,sales staff and core staff too facing difficulty in paying loans



35-45 Age group is the largest Group of Age applying for loans.



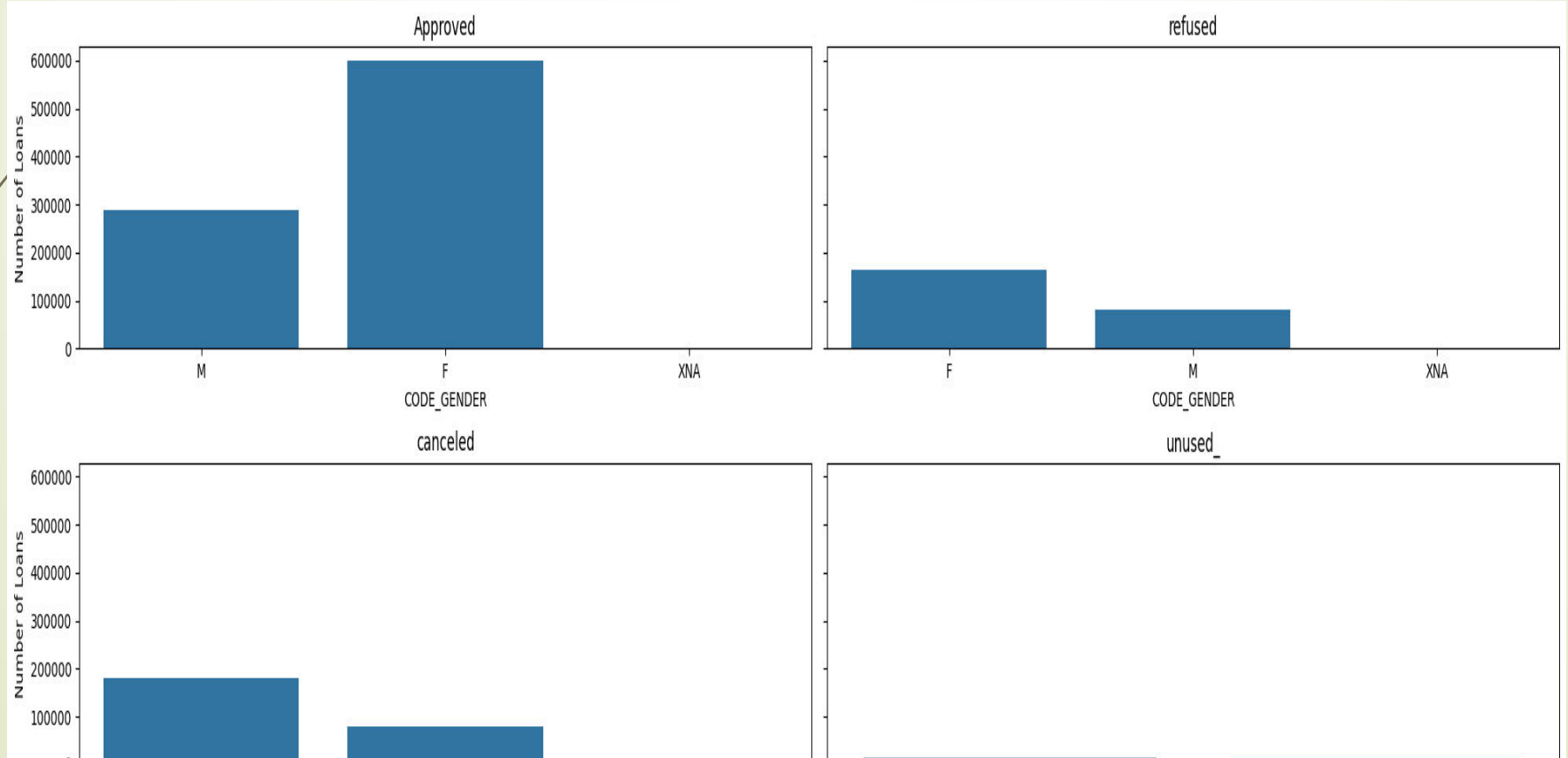
Medium Income group is the largest Group applying for loans.



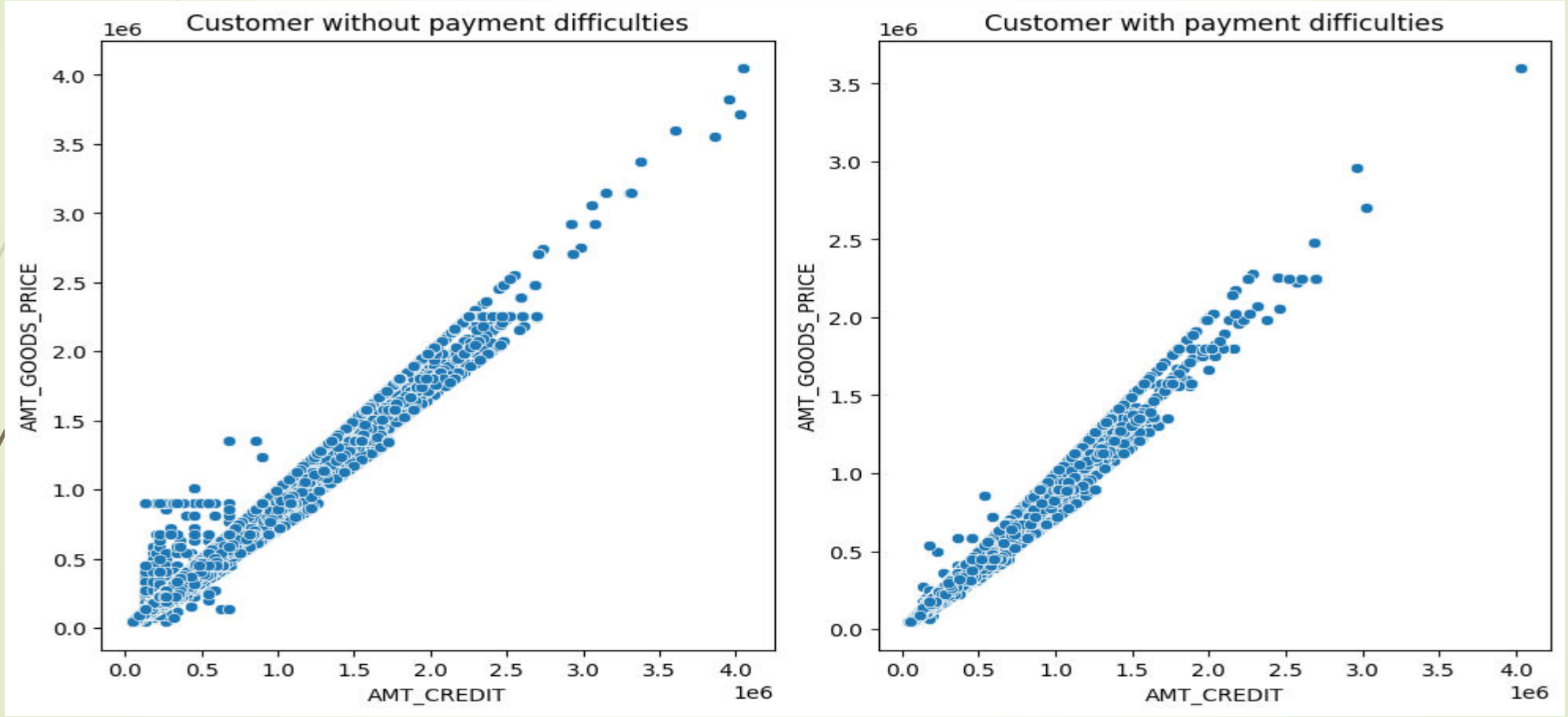
GRAPHS & INSIGHTS

B) BIVARIATE ANALYSIS

Female is getting more Refused more approved more canceled more unused.,
#but in case of male it is having average in every category.

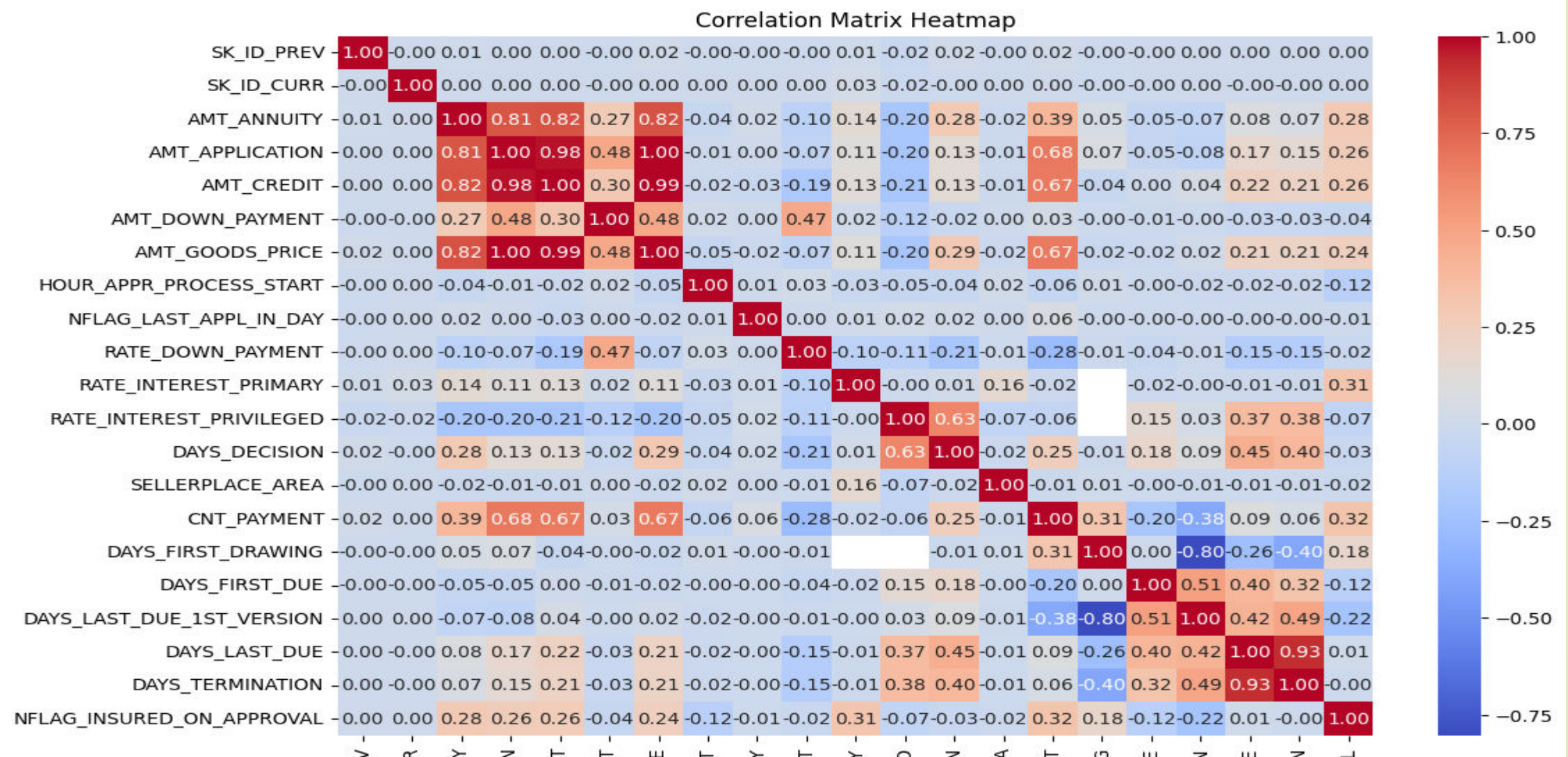


positively correlated(goods price is positivey correlated to credit amount)



GRAPHS & INSIGHTS

C) MULTIVARIATE ANALYSIS



CONCLUSIONS

Exploratory Data Analysis on Bank dataset reveals the following information:

Female is getting more Refused more approved more canceled more unused., but in case of male it is having average in every category.

More cash loans go to default. Bank should give more Revolving loans.

Single people default more and giving loans to married is safer.

The working type people are applying more loans as compare to others.

labourers are facing more difficulty in paying loans,sales staff and core staff too facing difficulty in paying loans.

Medium Income group is the largest Group applying for loans.

35-45 Age group is the largest Group of Age applying for loans.