CREDIT EDA ASSIGNMENT

BY VRUSHALI RANE

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:

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
2. 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.

The data given below contains the information about the loan application at the time of applying for the loan.

It contains two types of scenarios:

1. **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,
2. **All other cases:** All other cases when the payment is paid on time.

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

1. **Approved:** The Company has approved loan Application
2. **Cancelled:** The client cancelled the application sometime during approval.
3. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want.
4. **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
5. **Unused offer:** Loan has been cancelled by the client but at different stages of the process.

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency to default.

Business Objectives

This case study aims 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. This will ensure that the consumers capable of repaying the loan are not rejected.

The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Approach

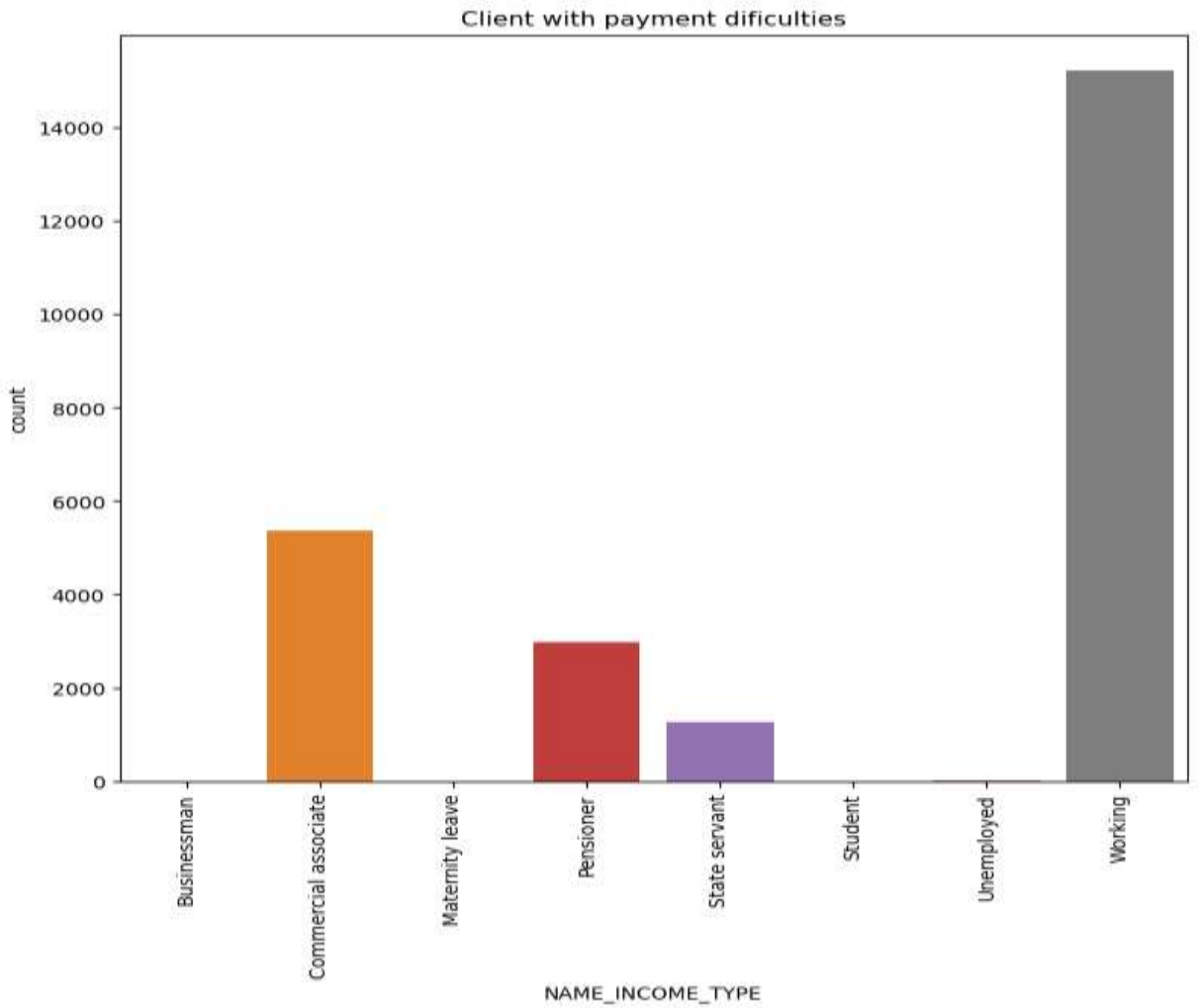
1. Identify and deal with missing data: We will identify the missing data and decide on the appropriate method to deal with it. We will remove columns or replace the missing values with appropriate values.
2. Identify outliers: We will identify the outliers in the dataset and determine why they are outliers. We will not remove any data points.
3. Identify data imbalance: We will identify any data imbalance and find the ratio of data imbalance. We will analyze the data using a mix of univariate and bivariate analysis techniques.
4. Explain analysis results: We will explain the results of univariate, segmented univariate, and bivariate analysis in business terms.
5. Find top 10 correlations: We will find the top 10 correlations for clients with payment difficulties and all other cases by segmenting the data frame with respect to the target variable.
6. Visualizations and results: We will include visualizations and summarize the most important results in the presentation.

Methods

1. Dealing with missing data: We will use appropriate methods to deal with missing data, such as imputation or removal.
2. Identifying outliers: We will use box plots to identify outliers. We will also consider the context of the data to determine if an outlier is significant.
3. Analyzing data imbalance: We will use visualizations such count plot to analyze data imbalance. We will also use statistical techniques to determine if the data is significantly imbalanced.
4. Explaining analysis results: We will use clear language and visualizations to explain the results of our analysis. We will also provide insights and recommendations based on our findings.
5. Finding top 10 correlations: We will use correlation matrices and heatmaps to find the top 10 correlations for clients with payment difficulties and all other cases.
6. Visualizations and results: We will use a variety of visualizations such as scatter plots, histograms, and bar charts to present our findings. We will also provide a summary of our most important results and insights.

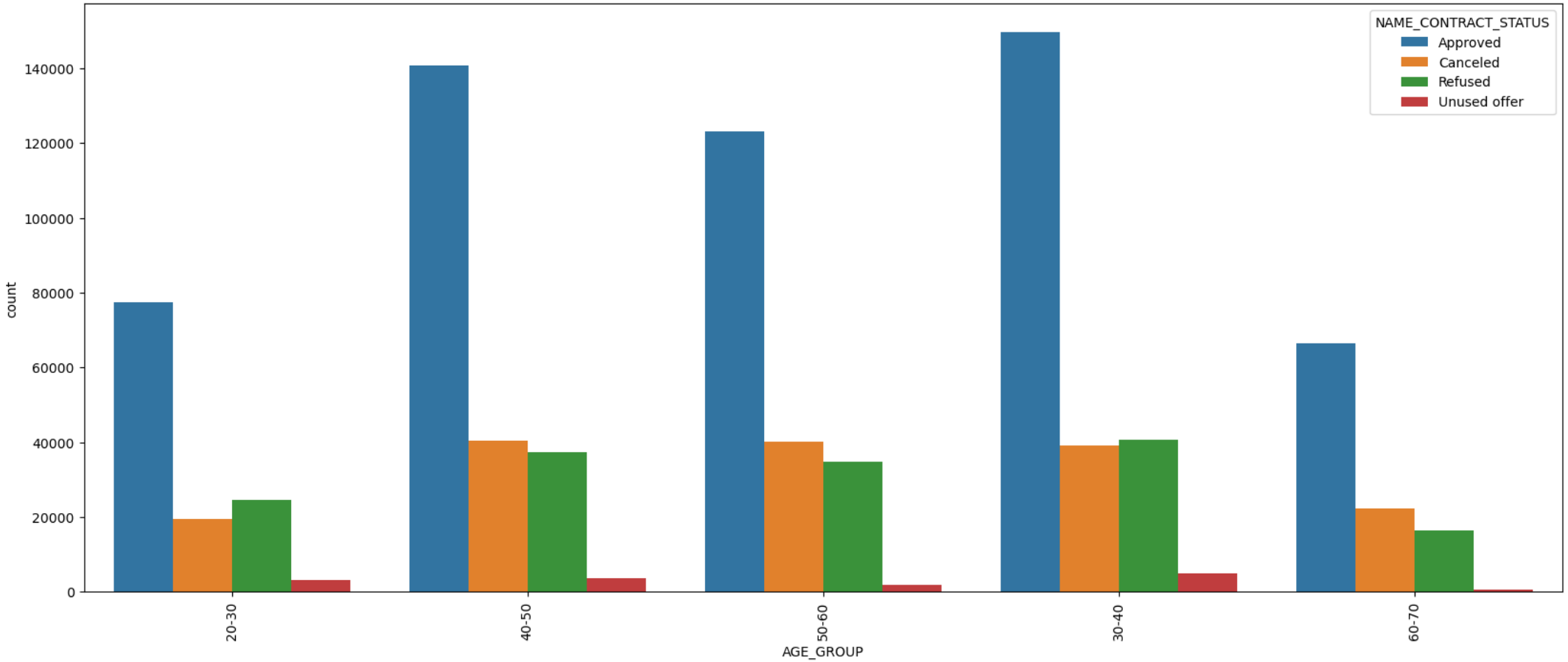
Graphs AND Insights

Univariate Analysis on Categorical variable ‘NAME_INCOME_TYPE



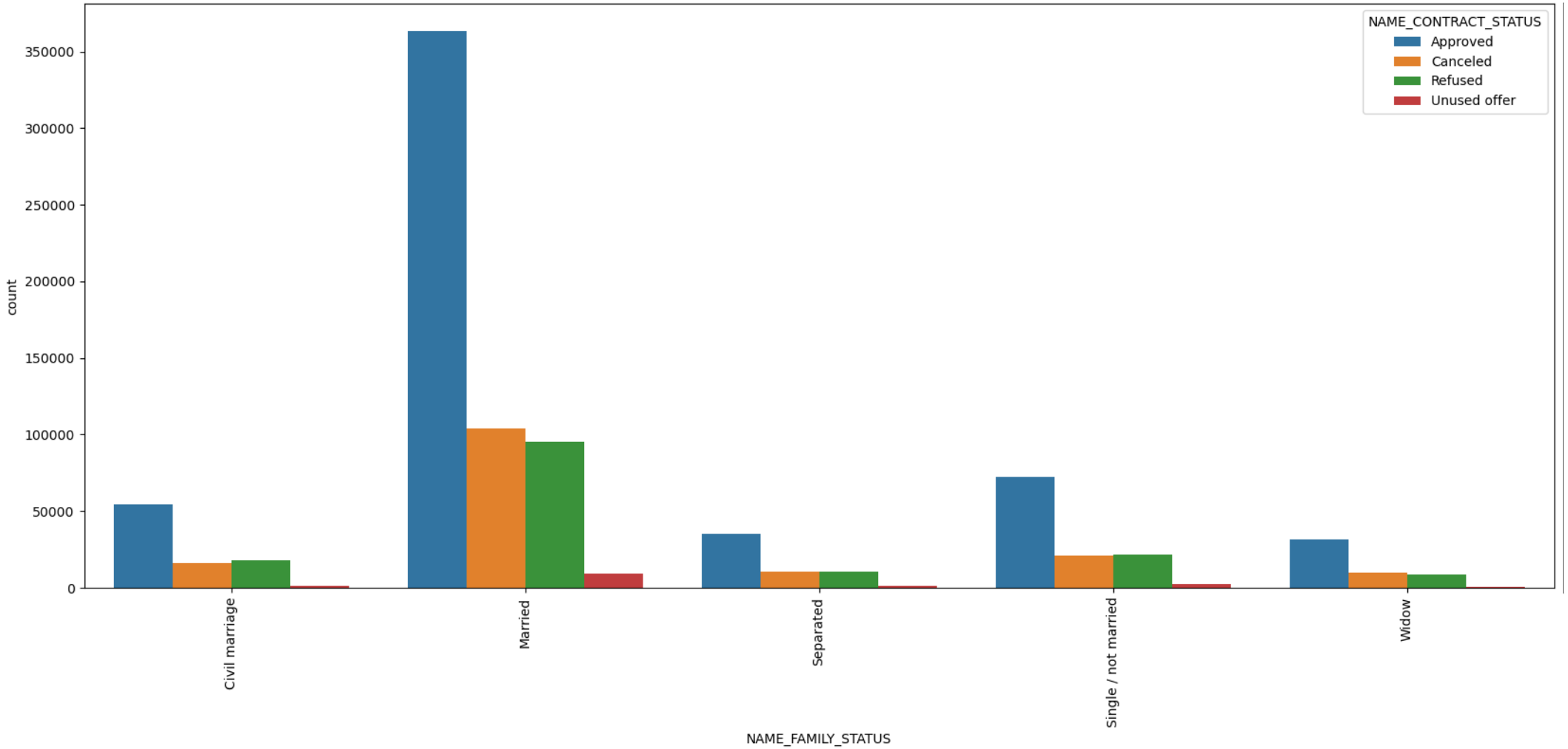
Students and Businessman don't have Payment difficulties

Bivariate Analysis on Categorical variables AGE_GROUP V/S NAME_CONTRACT_STATUS



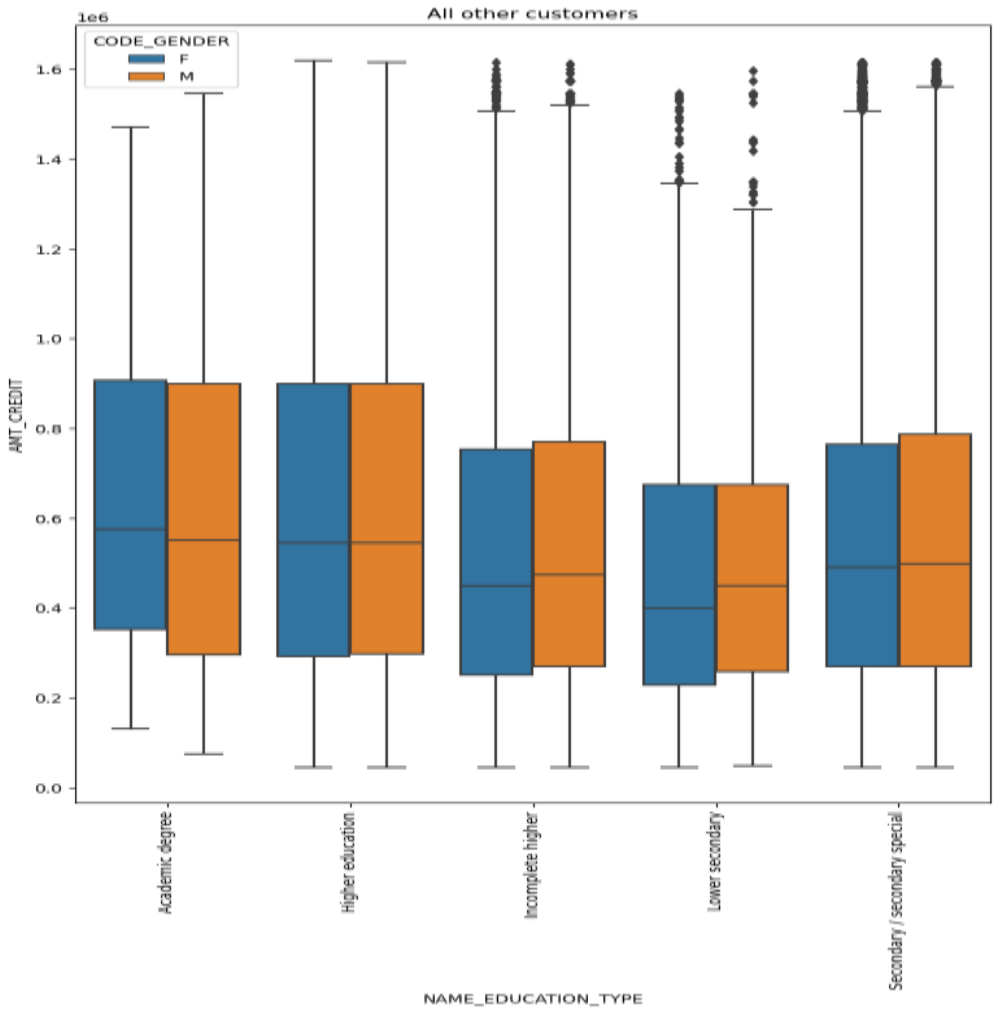
Clients who are in the age range 30-40 get most approval followed by clients in 40-50 age range

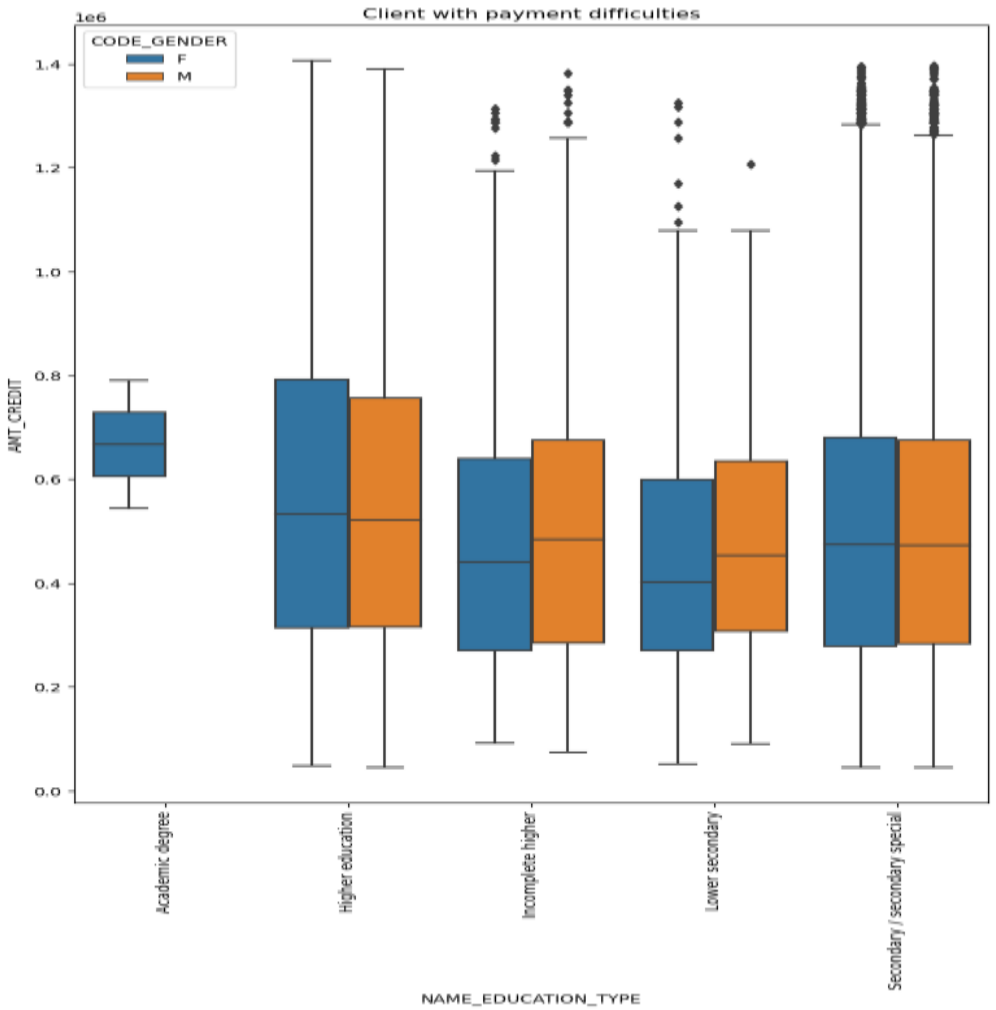
Bivariate Analysis on Categorical variables NAME_FAMILY_STATUS V/S NAME_CONTRACT_STATUS



Clients who are married got most approvals

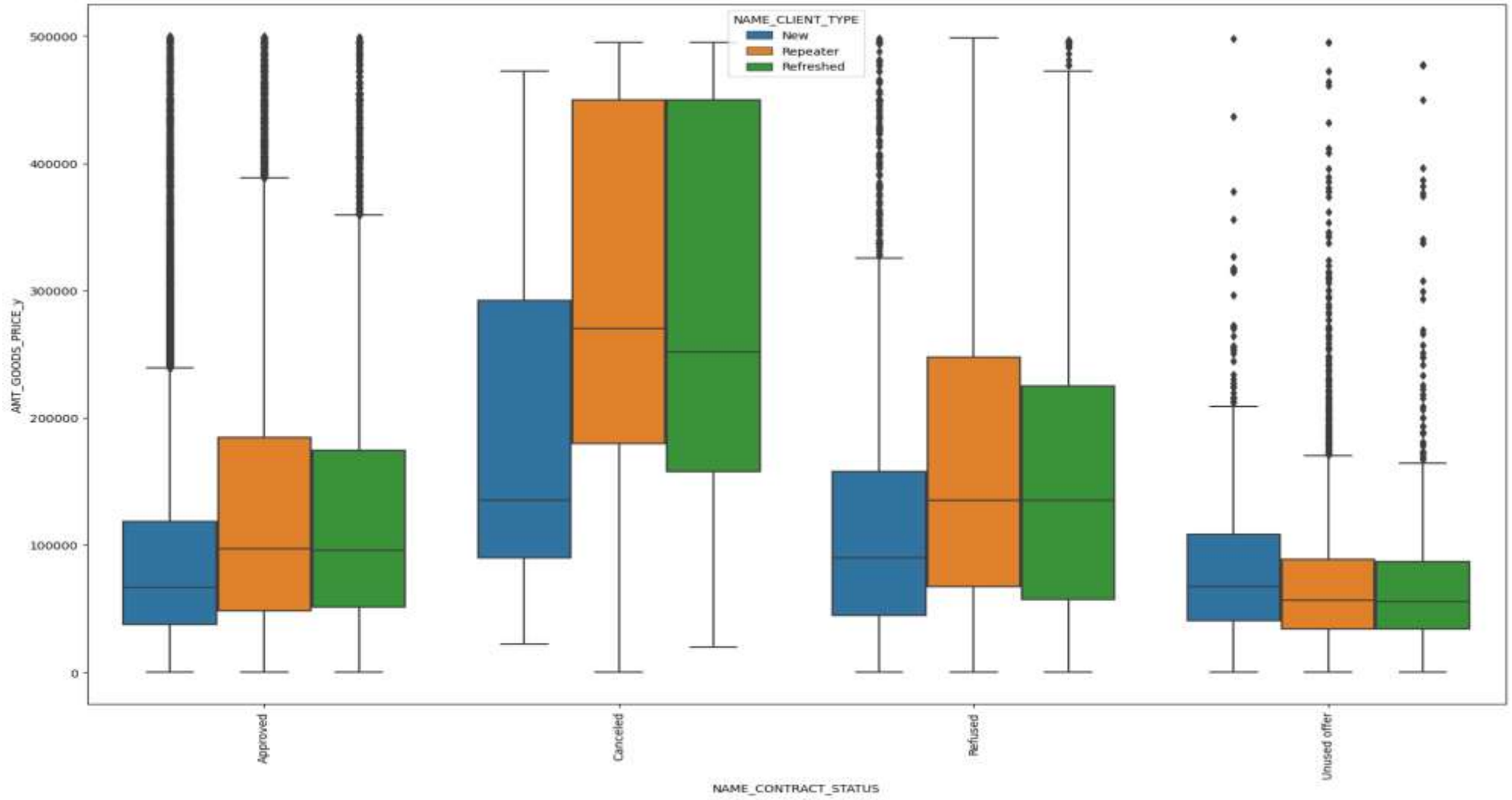
Multivariate Analysis on Categorical and Numerical Variables NAME_EDUCATION_TYPE V/S AMT_CREDIT V/S CODE_GENDER





Male Clients with Academic Degree are good in On-time payments

Multivariate Analysis on Categorical and Numerical Variables NAME_CONTRACT_STATUS V/S AMT_GOODS_PRICE_y V/S NAME_CLIENT_TYPE



Clients who are New and Canceled as well as Approved and New have less median goods price compared to Repeater and Refreshed

Conclusion

Through our analysis, we aim to identify the factors that differentiate clients with payment difficulties from all other cases.

We used exploratory data analysis techniques to identify patterns and relationships in the data, and provided insights and recommendations based on our findings.

Client categories to be targeted for providing loan

- Clients who are employed for more than 19 years
- Clients in the age range 30-40 and 40-50
- Clients who are Married
- Male clients with Academic degree
- Students and Businessman
- Repeater clients