Credit
Exploratory Data Analysis
(Case Study)

By Mohammed Suleman

BUSINESS OBJECTIVE

Objective:

- The case study aims to identify patterns indicating clients' difficulty in paying installments.
- These patterns will guide actions such as loan denial, risk reduction, or higher interest rates.
- The goal is to avoid rejecting capable loan applicants.
- **Approach**: Exploratory Data Analysis (EDA) to identify such applicants.

Company Goal:

- Understand key factors driving loan default.
- Identify strong indicators of default.
- Utilize this knowledge for portfolio management and risk assessment.

Problem Statement

Data Description:

- The dataset contains information related to loan applications at the time of applying for the loan.
- It includes two distinct scenarios:

Scenario 1: All Other Cases (Target = 0):

- This category encompasses all remaining instances where payments were made on time.
- In our analysis, we label these cases as “target = 0.”

Scenario 2: Clients with Payment Difficulties (Target = 1):

These clients experienced late payments exceeding a certain threshold (X days) on at least one of the initial Y loan installments.

In our analysis, we label these cases as “target = 1.”

ANALYSIS DONE Steps

- Data Understanding & preparation
- Data cleaning & Manipulation
- Data imbalance & Binning
- Data Analysis- Application data

  Univariate

  Bivariate &

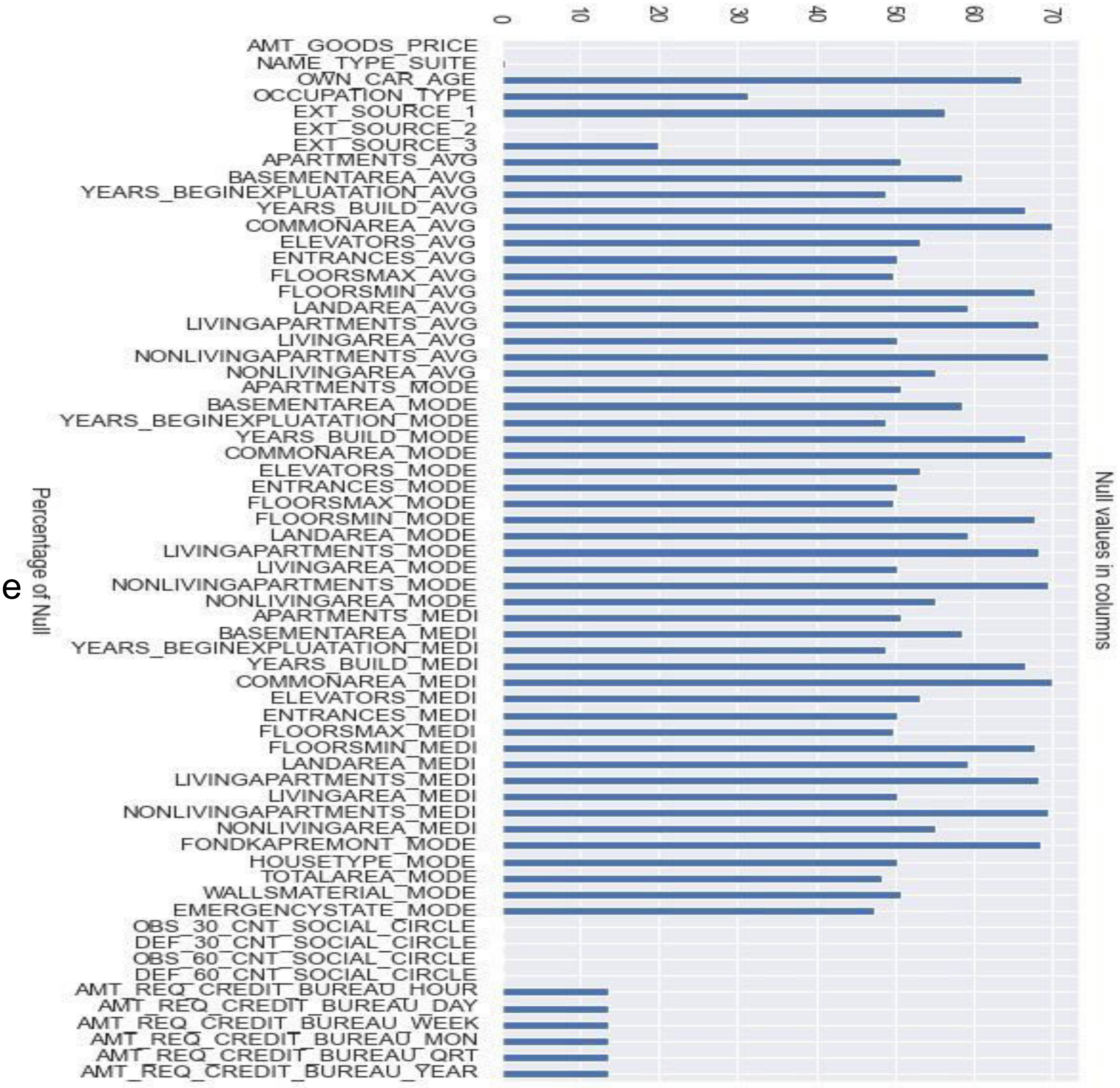
  Correlation
- Merging of application data with previous application data
- Recommendations and Risks

Columns having null values

Column Null Values:
Columns with more than 40% null values should be
dropped to maintain data integrity.

Imputation Strategy:
Columns with less than 13% null values should be
imputed with suitable values, the strategy for which
will be explained subsequently.

Imputation Analysis:
For the purpose of analyzing the imputation strategy,
5 variables have been selected.



Outliers & Binning

Inference -

AMT_CREDIT has little bit more outliers

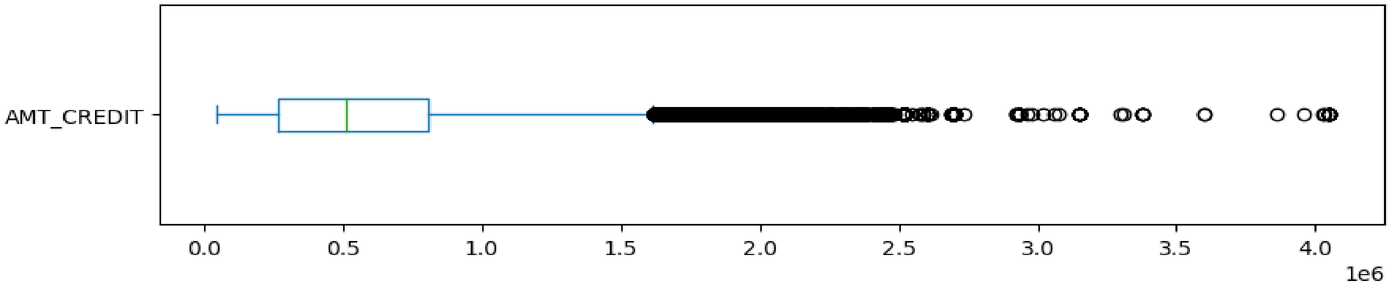
1st quartiles and 3rd quartile for AMT_ANNUITY is moved towards first quartile.

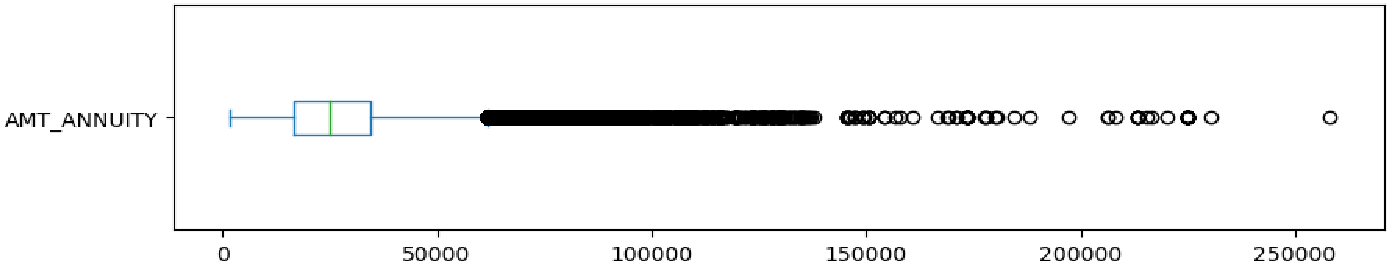
1st quartiles and 3rd quartile for DAYS_EMPLOYED is stays first quartile.

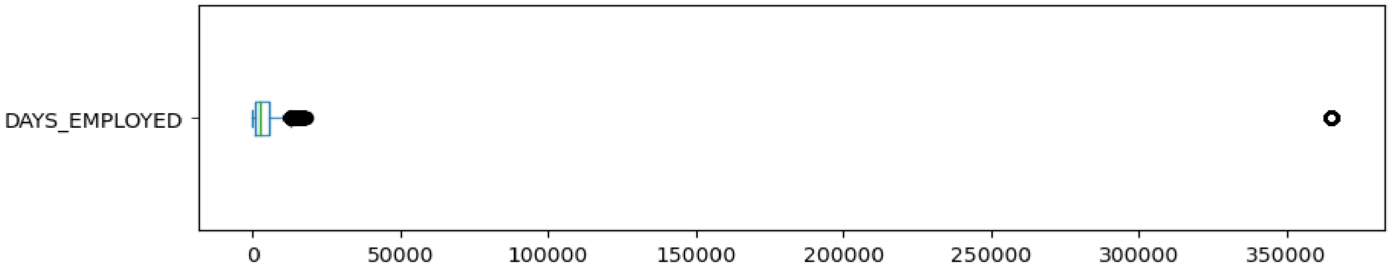
1st quartile is missing for CNT_CHILDREN which means most of the data are present in the 1st quartile.

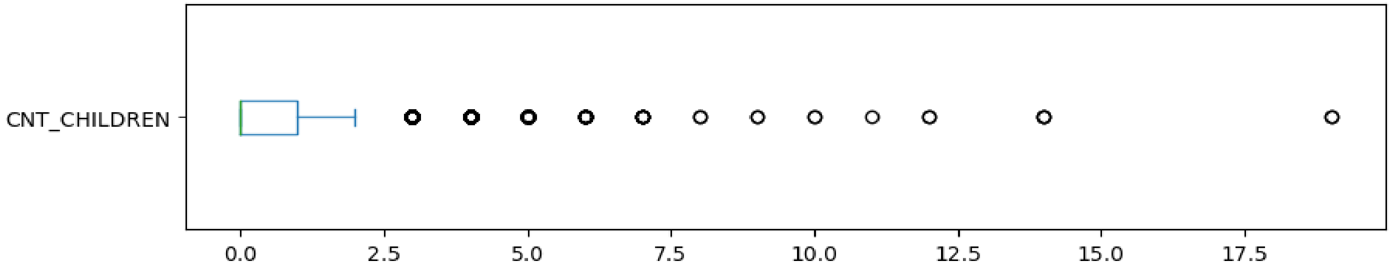
In AMT_INCOME_TOTAL only single high value data point is present as outlier

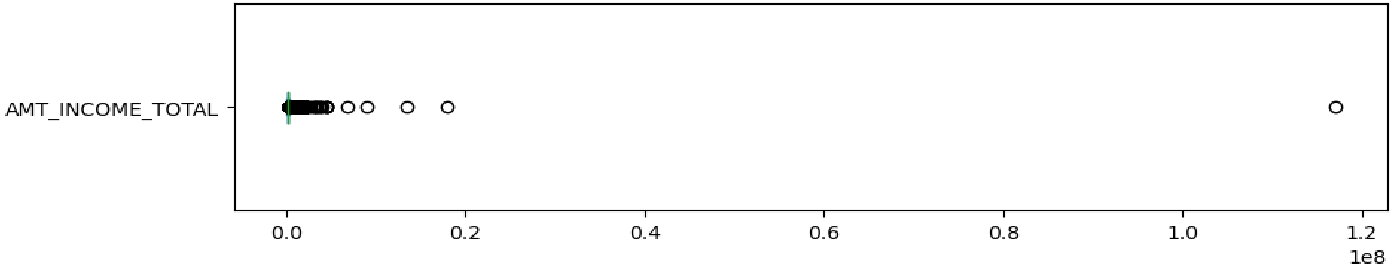
Binning on income & Credit columns











Data analysis

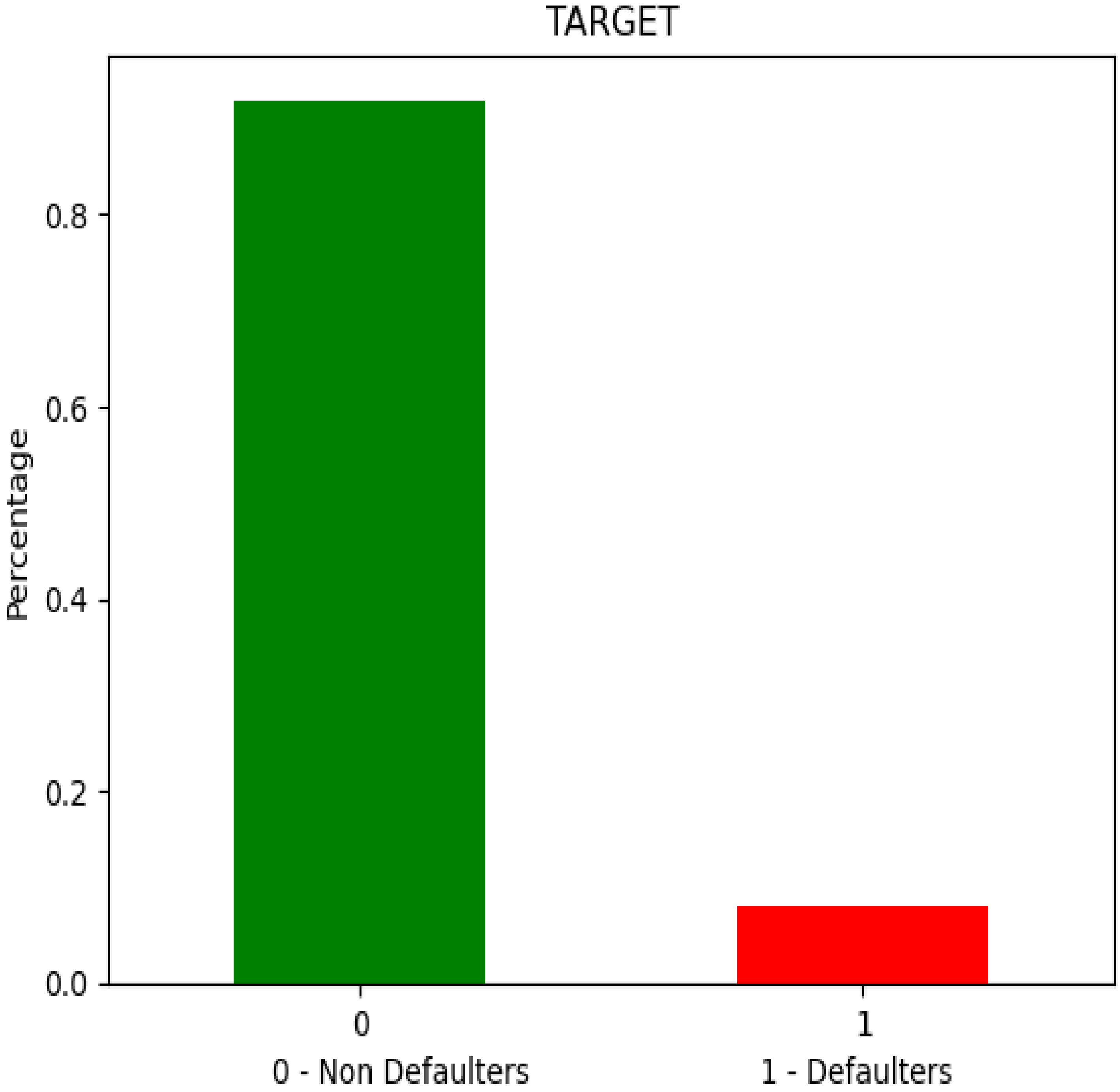
Bar Chart: The graph is a bar chart representing two categories: Non Defaulters (0) and Defaulters (1).

Non Defaulters: A significant majority of the data, represented by the green bar, are Non Defaulters.

Defaulters: A small percentage of the data, indicated by the red bar, are Defaulters.

Percentage Representation: The y-axis represents the percentage of each category in the dataset, providing a clear comparison between the two groups.

Data Imbalance Detected (11.39%) - Ratio of Non Defaulters to Defaulters is 11.39:1



Univariate Analysis

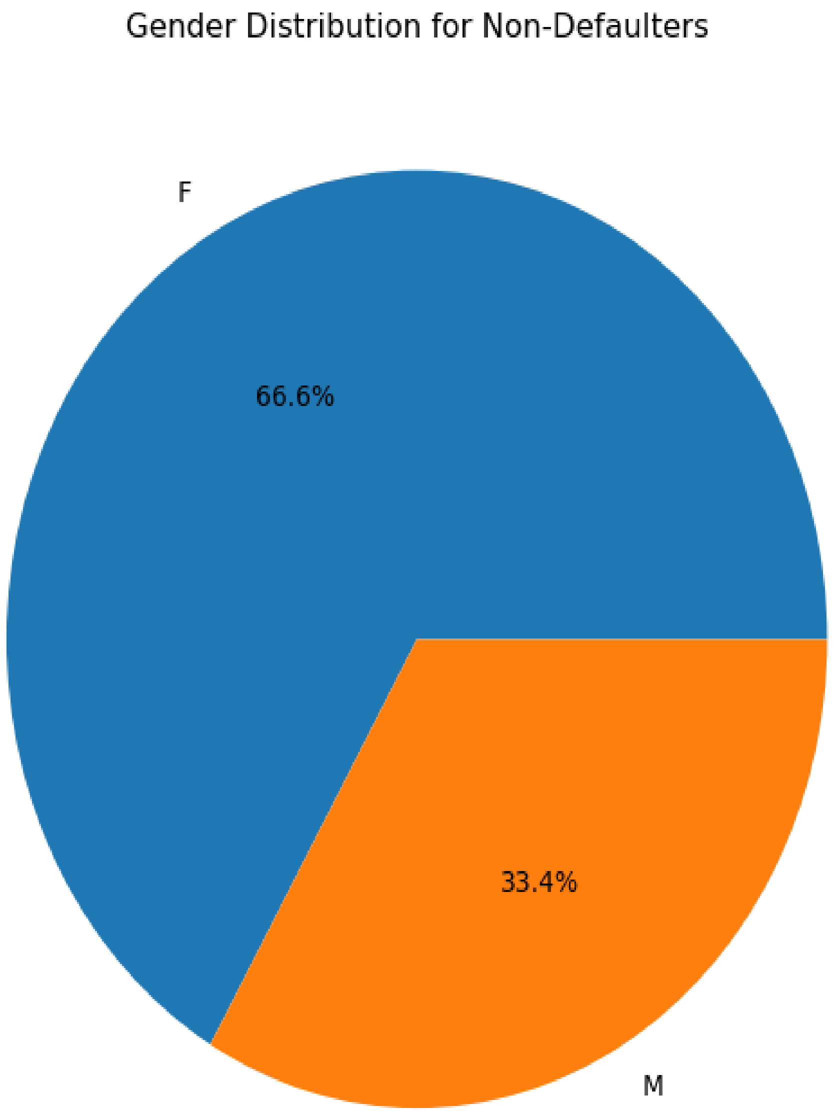
Categorical Variables

Gender

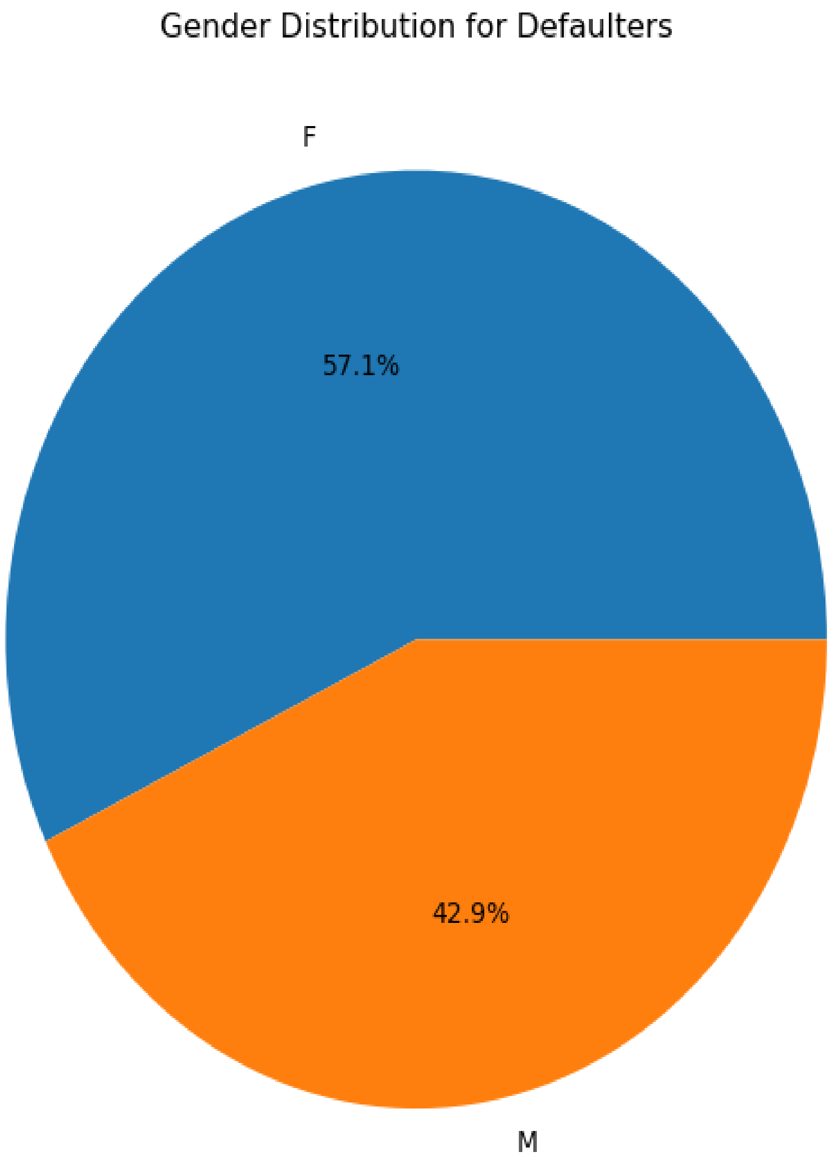
Inference -

Close to 57% of the applicants are Females in Defaulters

Close to 67% of the applicants are Females in Non-Defaulters



Gender Distribution for Non-Defaulters



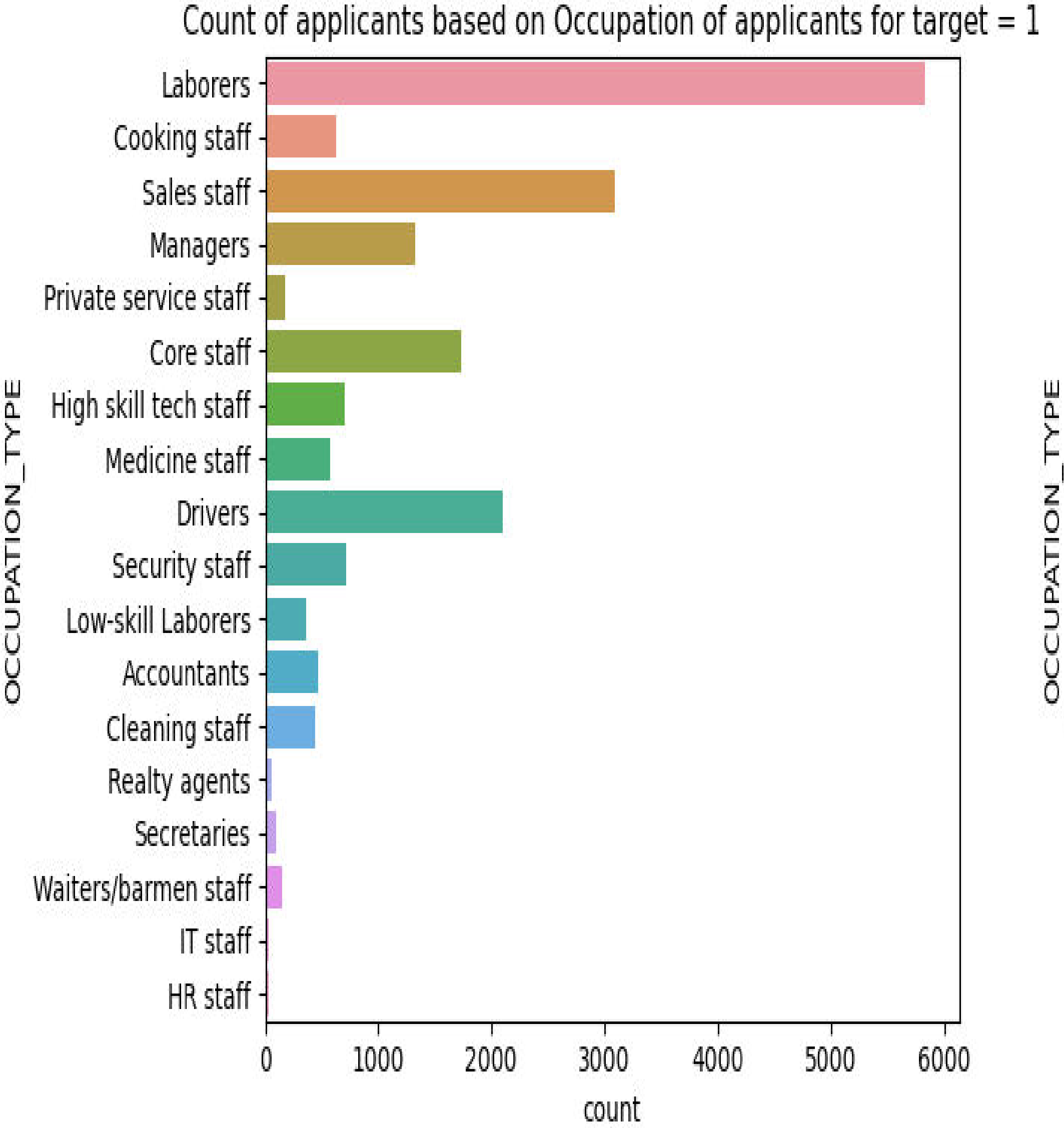
Gender Distribution for Defaulters

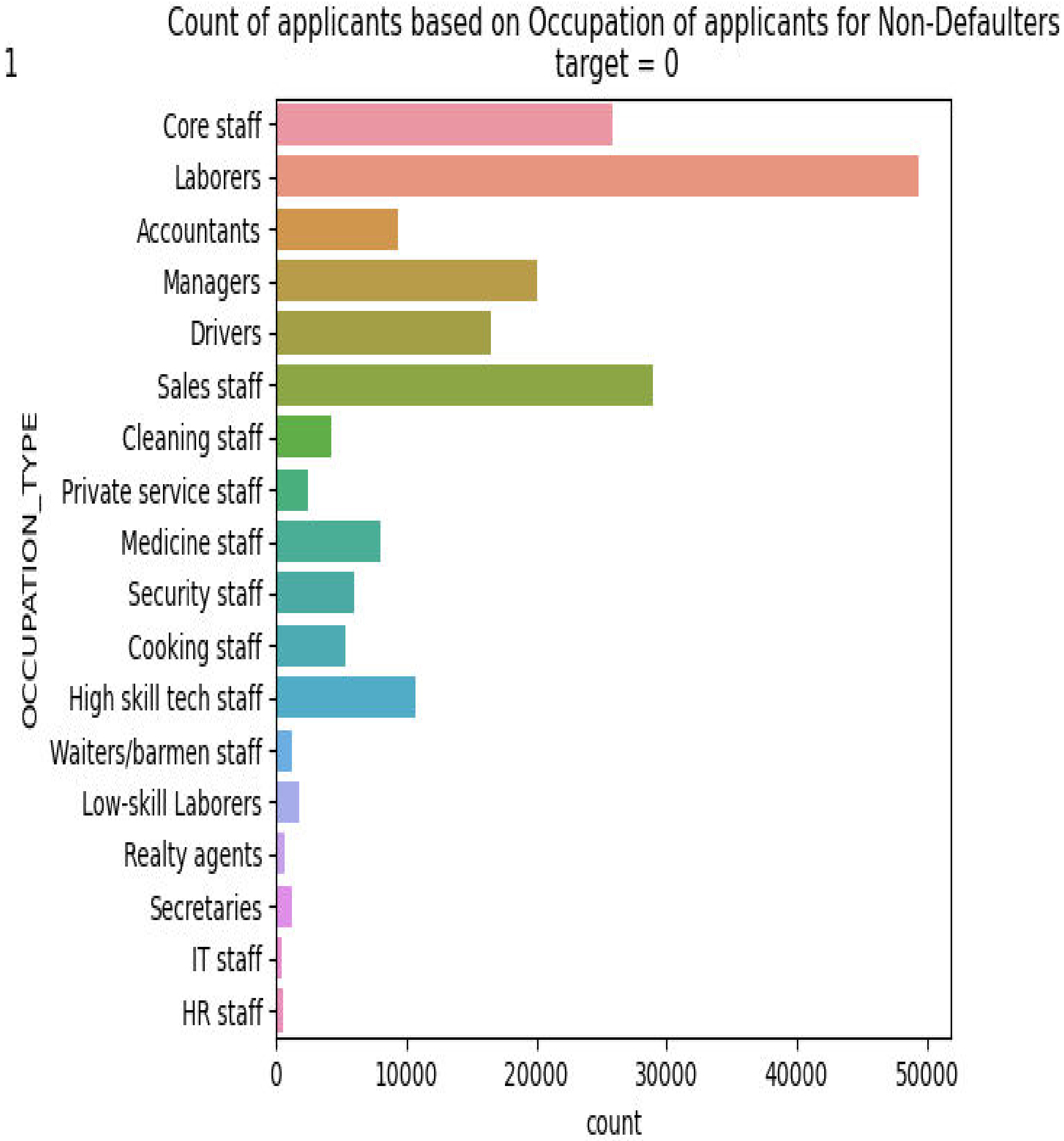
Univariate Analysis

OCCUPATION

Inference -

Most of the applicants belong to Laborer as Occupation


Count of applicants based on Occupation of applicants for target = 1

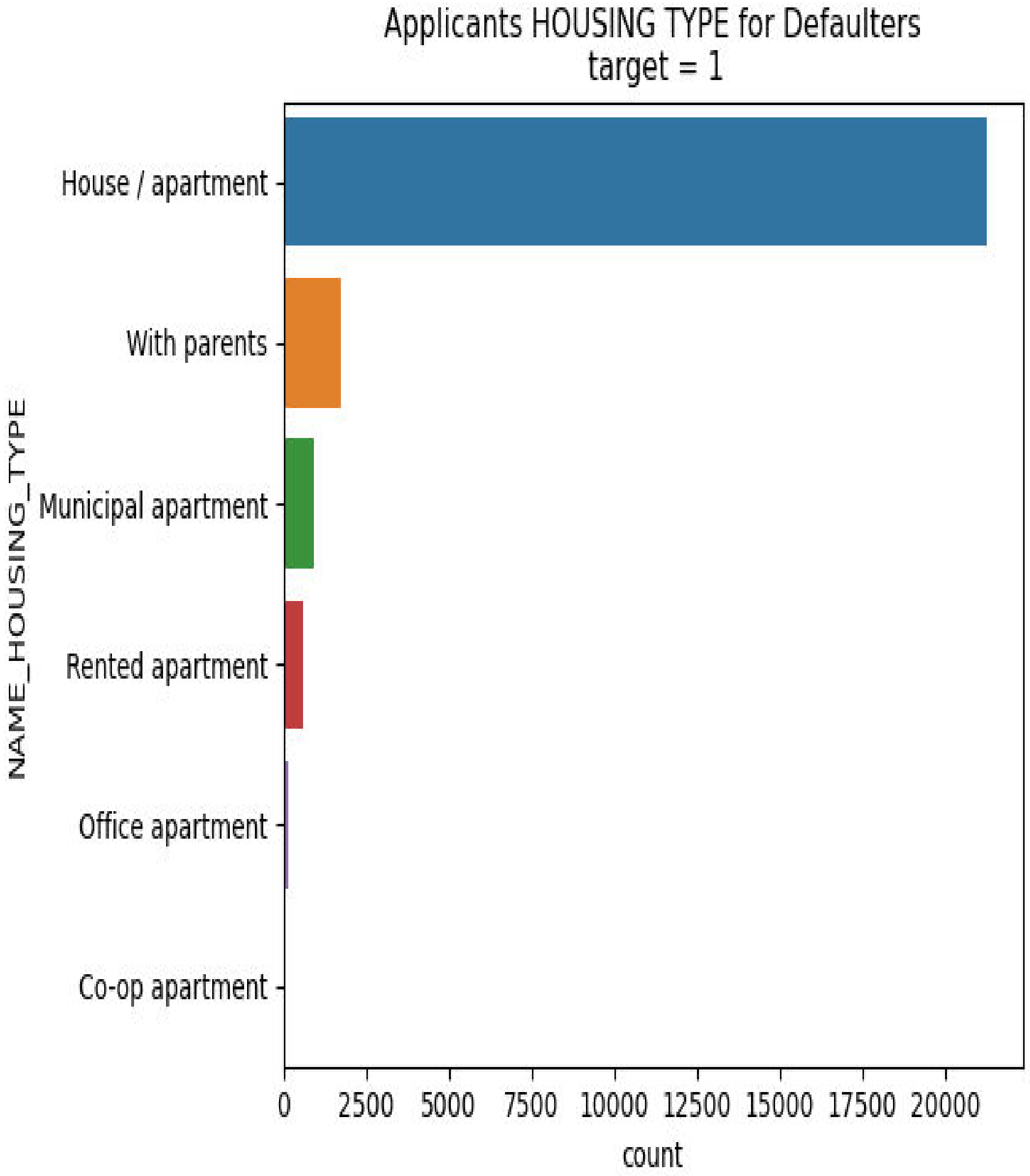

Count of applicants based on Occupation of applicants for Non-Defaulters target = 0

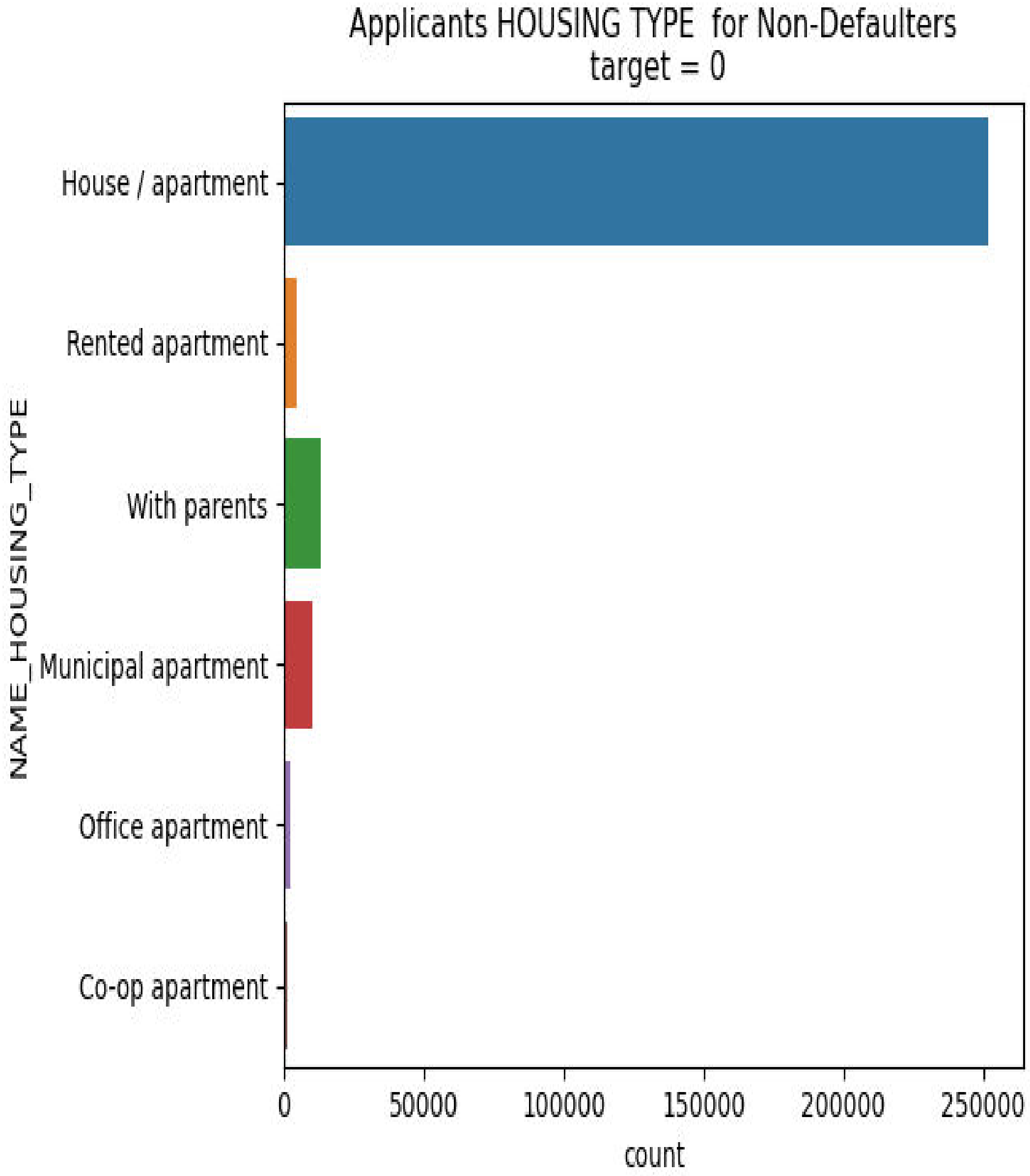
Univariate Analysis

HOUSING

Inference -

Most of the the applicants who own a house are non-defaulters and who don't own a house are defaulters. Its a very intresting trend here. We can say that applicants who own a house are tend to be non-defaulters


Applicants HOUSING TYPE for Defaulters
target = 1


Applicants HOUSING TYPE for Non-Defaulters
target = 0

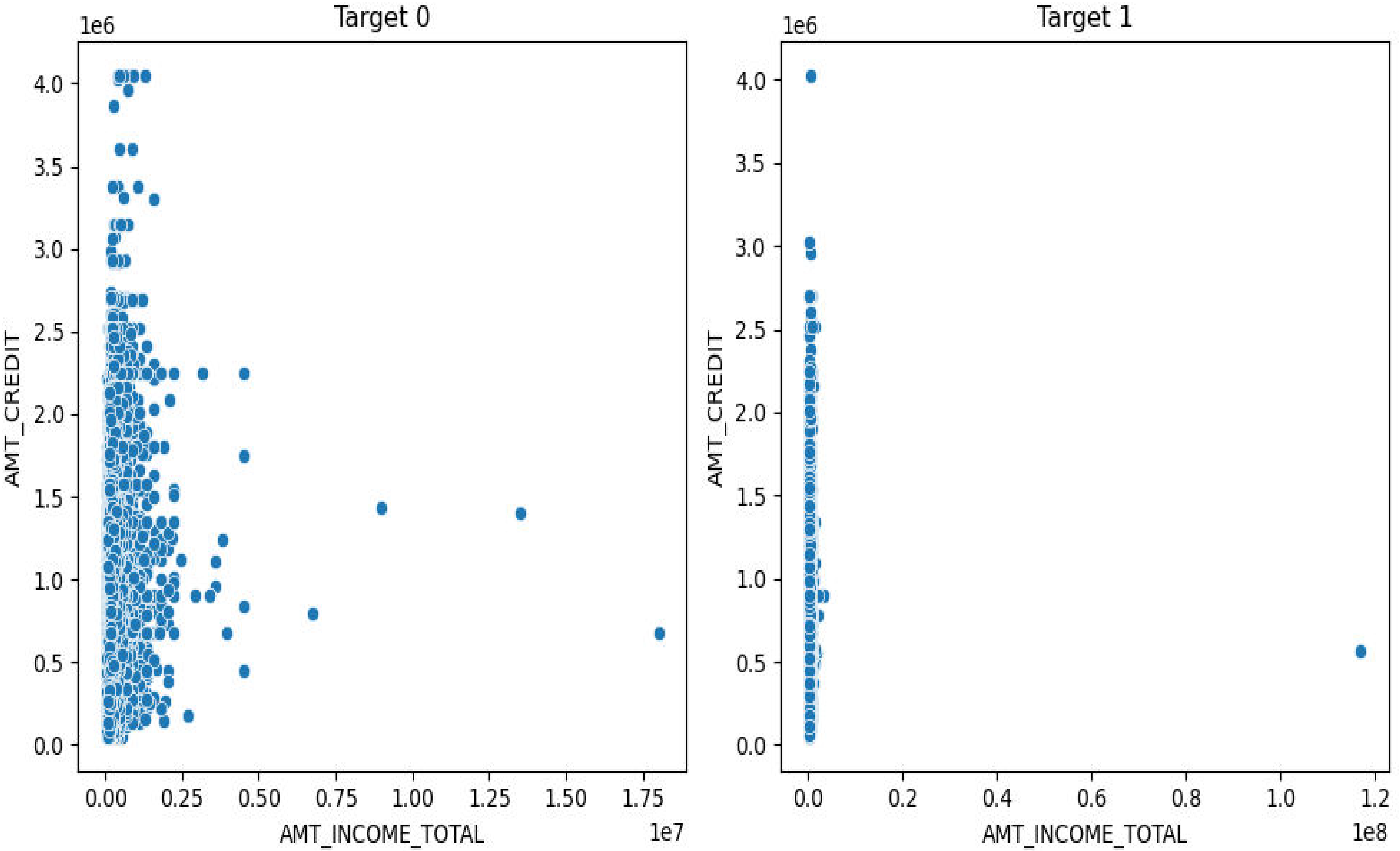
Bivariate Analysis

INCOME & CREDIT

Inference -

There is a positive correlation between amt_credit and amt_income_total. This means that as the credit amount increases, the income total also tends to increase.

The data points are spread out, indicating a weak to moderate correlation. This means that the increase in income is not always consistent with the increase in credit amount.



Correlation

Inference:

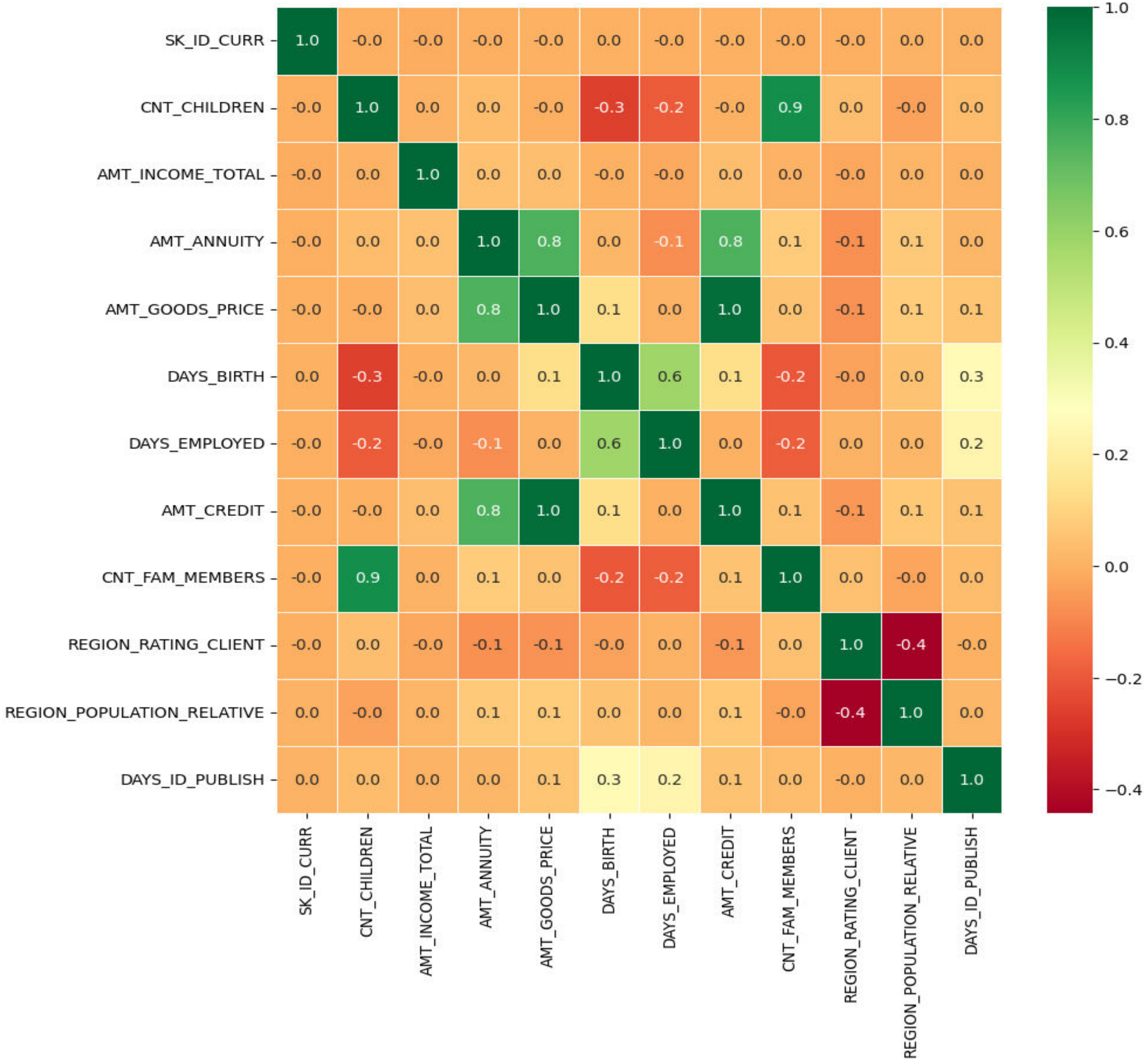
For both Target 0 and Target 1, these columns exhibit significant correlation values.

There is a strong positive correlation (values close to 1) between several pairs of variables, including:

AMT_CREDIT_Current and AMT_CREDIT_Previous
AMT_ANNUITY_Current and AMT_ANNUITY_Previous
AMT_INCOME_TOTAL and AMT_RECEIPT_TOTAL

There is a weak positive correlation (values between 0.2 and 0.4) between AMT_GOODS_PRICE_Current and AMT_INCOME_TOTAL.

Correlation matrix for Clients with payment difficulties



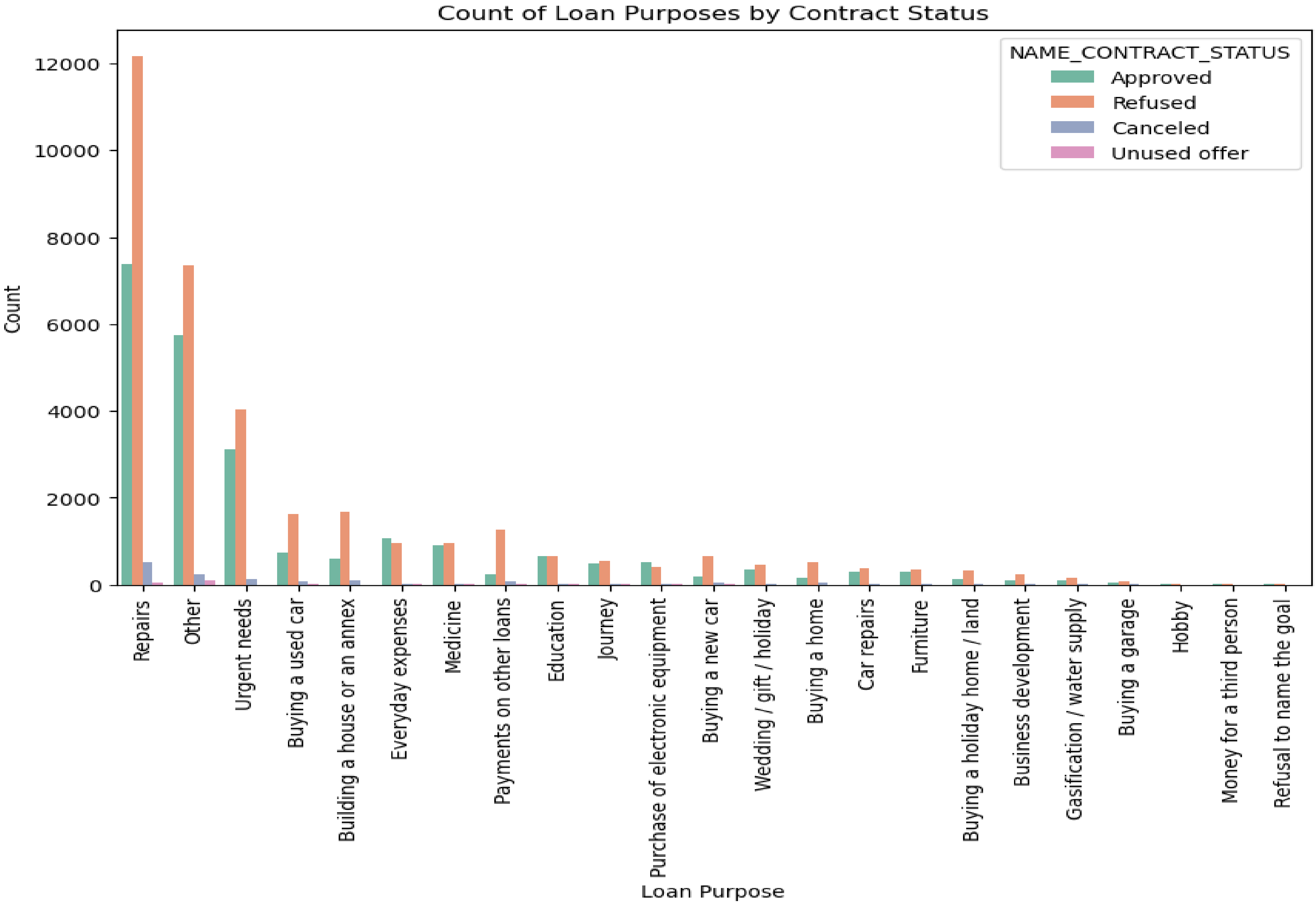
	SK_ID_CURR	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED	AMT_CREDIT	CNT_FAM_MEMBERS	REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	DAYS_ID_PUBLISH
SK_ID_CURR	1.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0
CNT_CHILDREN	-0.0	1.0	0.0	0.0	-0.0	-0.3	-0.2	-0.0	0.9	0.0	-0.0	0.0
AMT_INCOME_TOTAL	-0.0	0.0	1.0	0.0	0.0	-0.0	-0.0	0.0	0.0	-0.0	0.0	0.0
AMT_ANNUITY	-0.0	0.0	0.0	1.0	0.8	0.0	-0.1	0.8	0.1	-0.1	0.1	0.0
AMT_GOODS_PRICE	-0.0	-0.0	0.0	0.8	1.0	0.1	0.0	1.0	0.0	-0.1	0.1	0.1
DAYS_BIRTH	0.0	-0.3	-0.0	0.0	0.1	1.0	0.6	0.1	-0.2	-0.0	0.0	0.3
DAYS_EMPLOYED	-0.0	-0.2	-0.0	-0.1	0.0	0.6	1.0	0.0	-0.2	0.0	0.0	0.2
AMT_CREDIT	-0.0	-0.0	0.0	0.8	1.0	0.1	0.0	1.0	0.1	-0.1	0.1	0.1
CNT_FAM_MEMBERS	-0.0	0.9	0.0	0.1	0.0	-0.2	-0.2	0.1	1.0	0.0	-0.0	0.0
REGION_RATING_CLIENT	-0.0	0.0	-0.0	-0.1	-0.1	-0.0	0.0	-0.1	0.0	1.0	-0.4	-0.0
REGION_POPULATION_RELATIVE	0.0	-0.0	0.0	0.1	0.1	0.0	0.0	0.1	-0.0	-0.4	1.0	0.0
DAYS_ID_PUBLISH	0.0	0.0	0.0	0.0	0.1	0.3	0.2	0.1	0.0	-0.0	0.0	1.0

MERGE

Inference:

The graph indicates that the majority of loan rejections are associated with the purpose of ‘Repairs’.

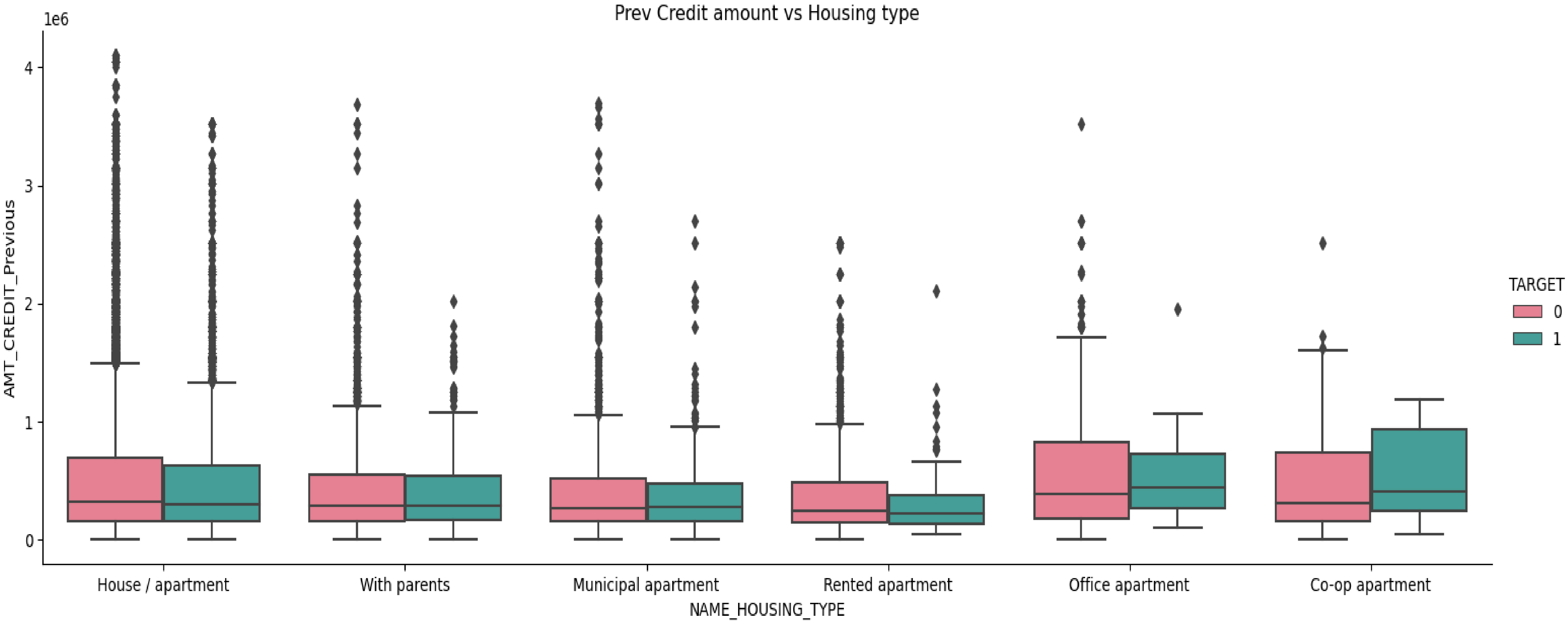
For purposes such as Medicine, Everyday expenses, and Education, the graph shows a nearly equal distribution of loan approvals and rejections.



MERGE

Inference:

It can be inferred that banks should consider refraining from granting loans to co-op apartment housing types, as they appear to encounter difficulties in making payments.



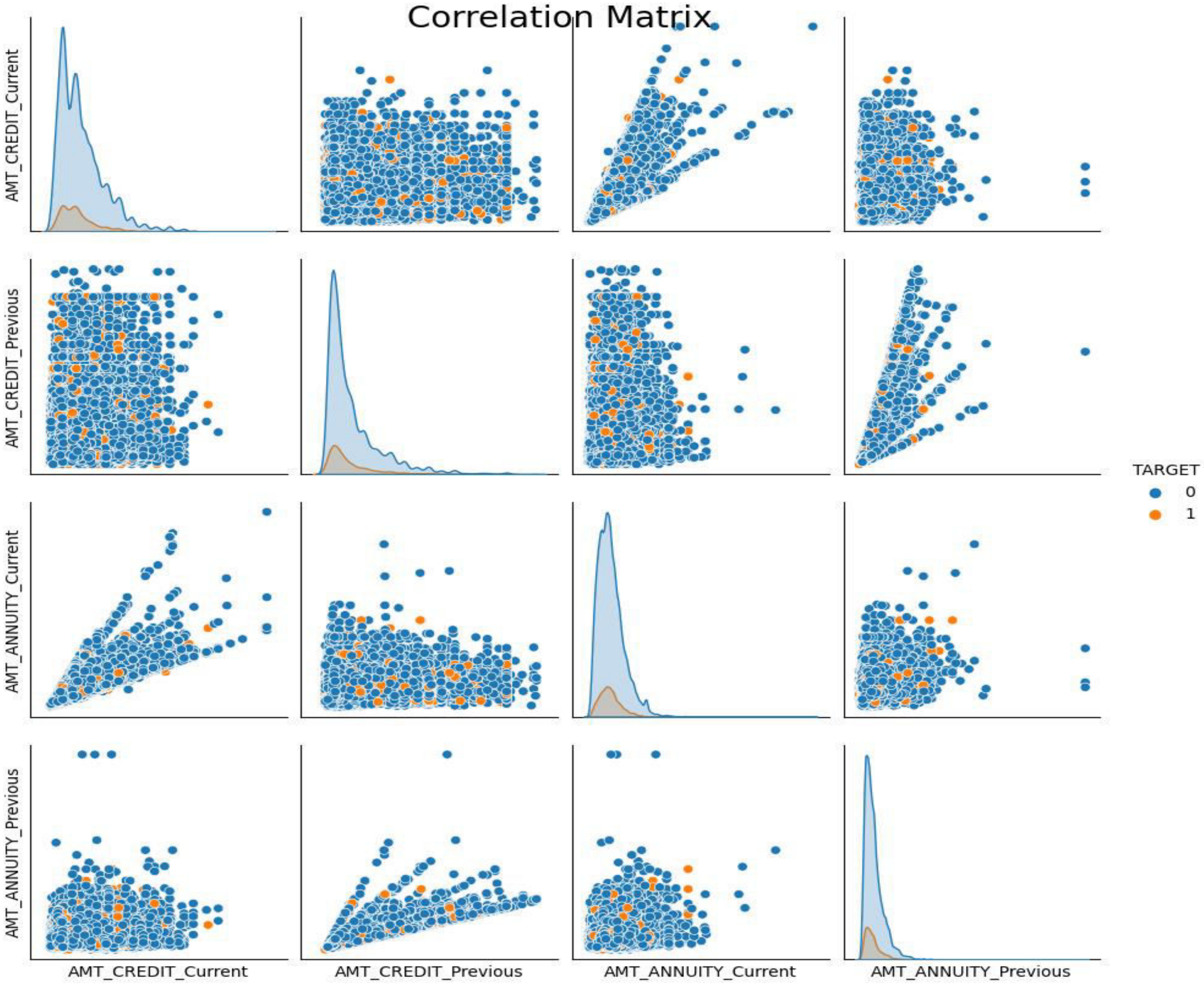
Correlation

Inference:

Correlation Matrix: The graph is a correlation matrix that shows the relationships between different variables: AMT_CREDIT_Current, AMT_CREDIT_Previous, AMT_ANNUITY_Current, and AMT_ANNUITY_Previous.

Scatter Plots and Histograms: Each cell in the matrix shows a scatter plot of two variables with histograms along the diagonal. The color of the points (blue and orange) represents different TARGET values (0 and 1).

Data Distribution: Most data points are concentrated towards the lower values for all variables, indicating a potential skew in the data distribution.



Conclusion & recommendations

Gender and Defaulting: As a higher percentage of defaulters are females, financial institutions might consider developing gender-specific financial literacy programs to help reduce default rates.

Occupation: Since most applicants belong to the laborer occupation, banks could consider offering tailored financial products or advisory services for this group to manage their loans better.

Housing: The trend that non-defaulters are more likely to own a house suggests that home ownership could be a positive factor in loan repayment. Banks might consider this factor in their loan approval process.

Correlation Analysis: The strong positive correlation between several pairs of variables such as AMT_CREDIT_Current and AMT_CREDIT_Previous indicates that these factors could be significant in predicting loan default. Financial institutions could use these insights to enhance their risk assessment models.

The purpose of this analysis is to identify defaulters and understand their characteristics. The steps involved in this process include data cleaning, data analysis and drawing inferences from the data. Relevant graphs have been attached

THANK YOU