EDA ASSIGNMENT

- BY VISHAL KUMAR

Objective :-

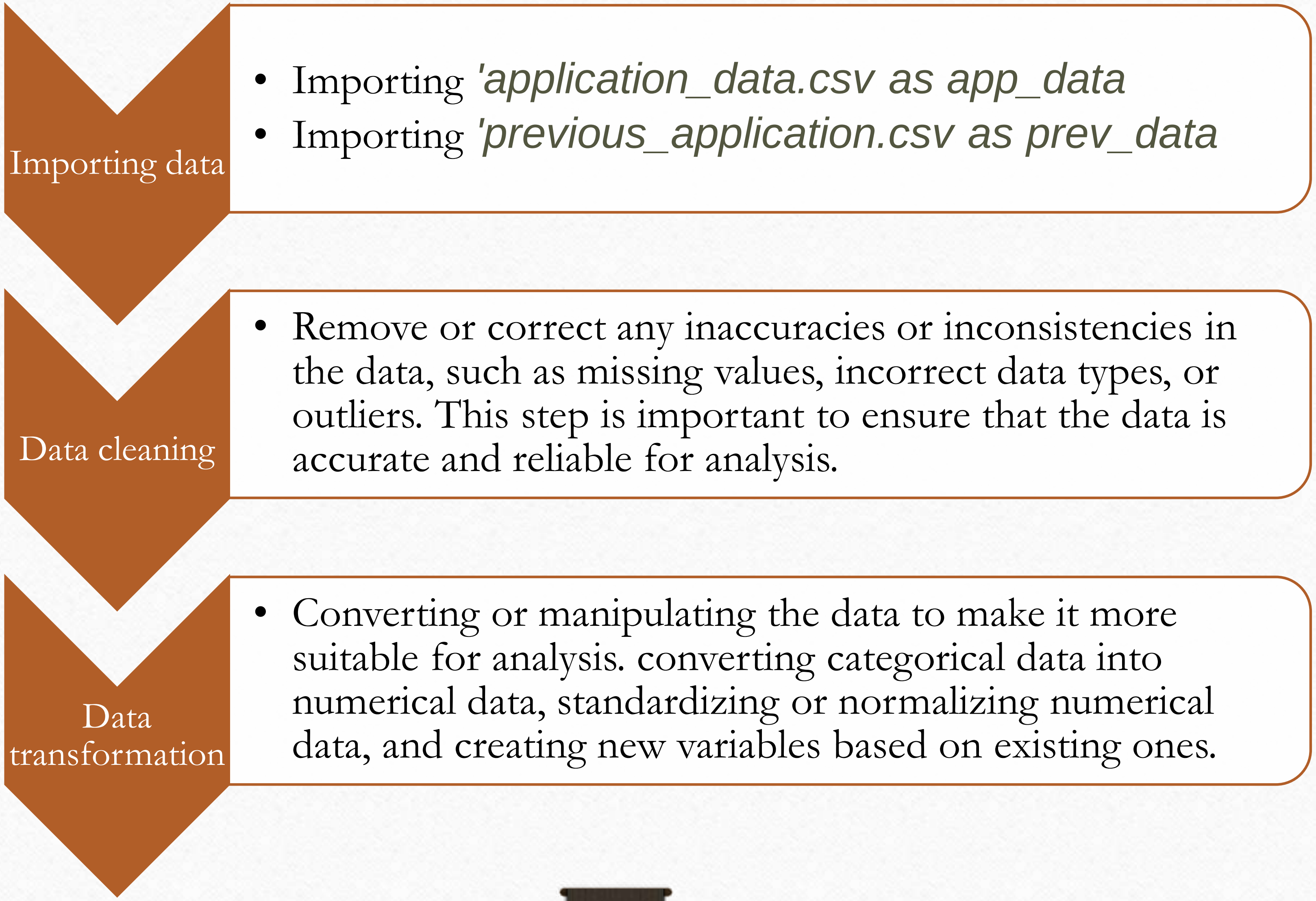
The aim is to identify patterns that show whether a client is having trouble making their payments, which may be used to decide whether to grant the loan, reduce its size, charge riskier applicants a higher interest rate, etc. By doing this, it will be ensured that only borrowers who can repay the loan will be accepted. The objective of this case study is to identify such applications using EDA.

Data Understanding

- *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

- Gathered and imported loan data from various sources
- Performed data cleaning and preprocessing to ensure data quality
- Conducted exploratory data analysis (EDA) using visualizations and statistical analysis
- Identified potential predictors of loan default, such as income, credit history, and loan amount
- Examined the relationships between the potential predictors and loan default status
- Identified any patterns or trends in the data that may be useful for identifying high-risk loan applicants
- Developed hypotheses about the factors that contribute to loan default and validated them through further analysis
- Summarized the key findings and insights from the EDA, including any potential action items that may be taken to mitigate loan default risk.



Importing data

- Importing *'application_data.csv as app_data*
- Importing *'previous_application.csv as prev_data*

Data cleaning

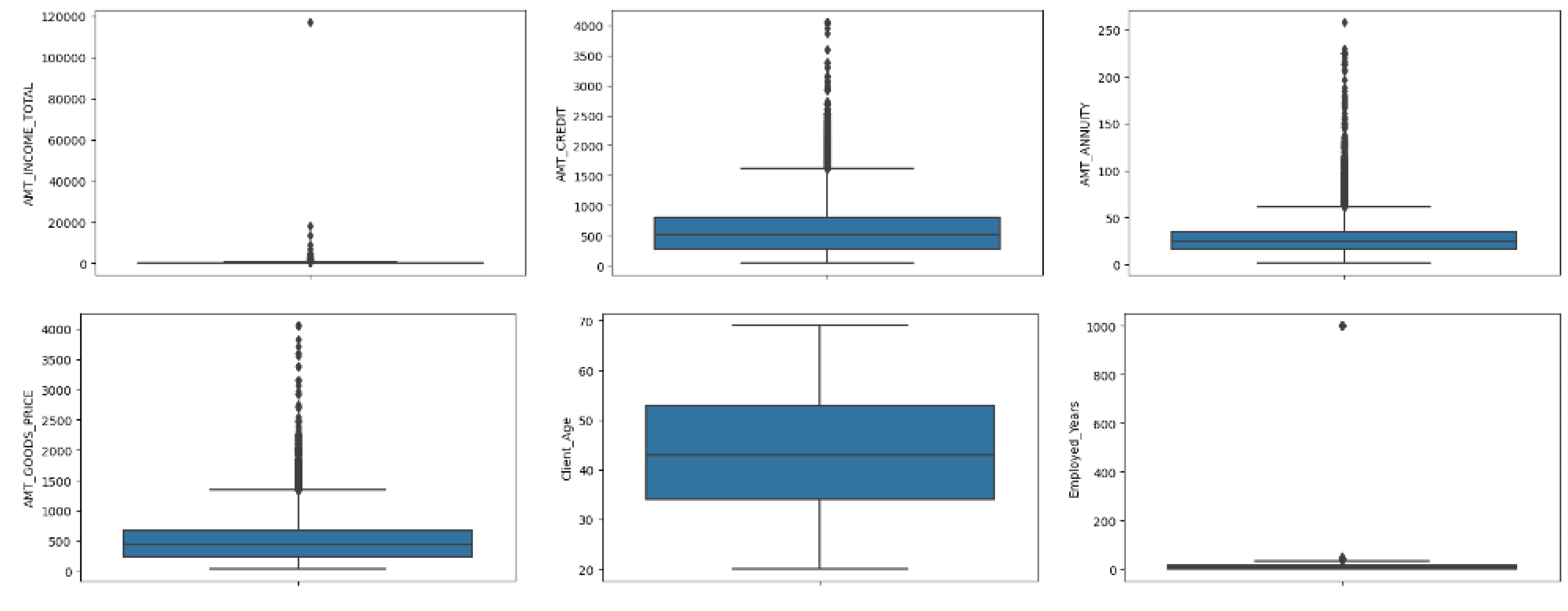
- Remove or correct any inaccuracies or inconsistencies in the data, such as missing values, incorrect data types, or outliers. This step is important to ensure that the data is accurate and reliable for analysis.

Data transformation

- Converting or manipulating the data to make it more suitable for analysis. converting categorical data into numerical data, standardizing or normalizing numerical data, and creating new variables based on existing ones.

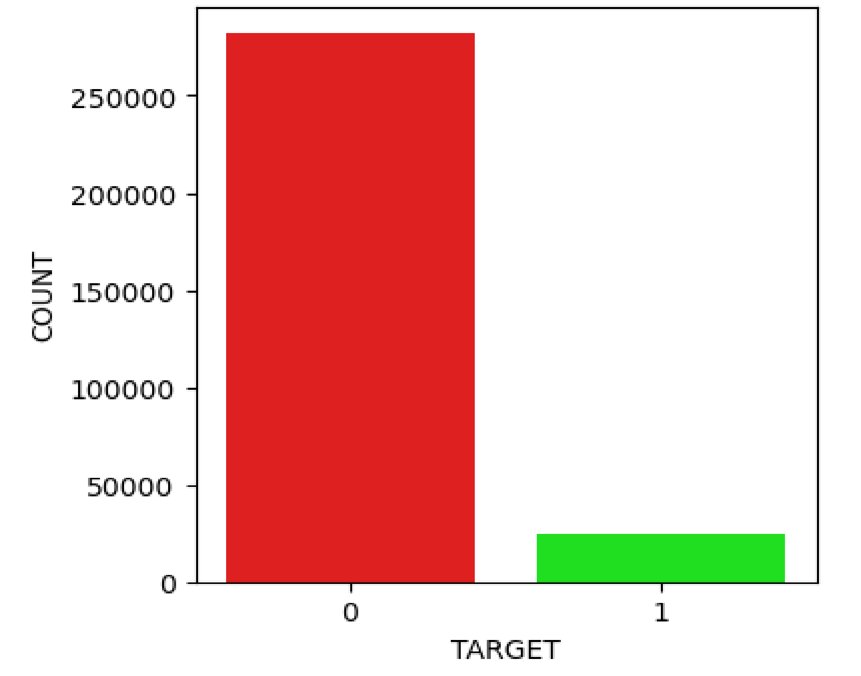
Outliers in variable found during box plot and quantile categorizing

AMT_GOODS_PRICE above 13 lkh
Employed_Years above 19 years
AMT_ANNUITY above 60,000
AMT_INCOME_TOTAL above 2 lkh
AMT_CREDIT above 15 lkh



DATA IMBALANCE

- 🟥 = Target with payment difficulties (0)
- 🟩 = All other cases (1)



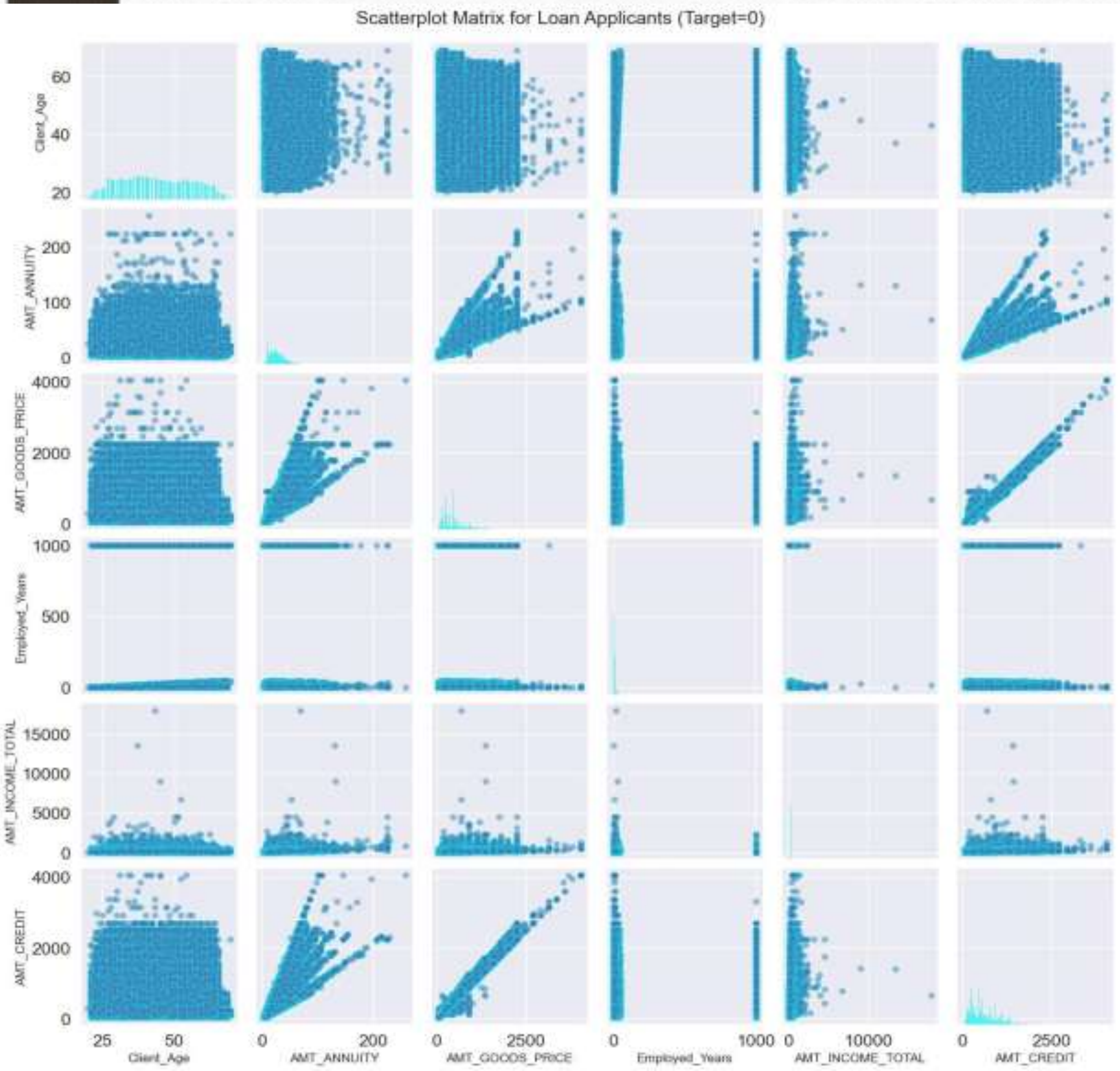
Bivariate analysis when TARGET = 0

The same variables have the highest correlations in both scenarios, with TARGET = 0 slightly higher than TARGET =1:

1. The strongest positively linked variables are AMT_GODS_PRICE and AMT_CREDIT, with coefficients of 0.98 when TARGET = 1 and 0.99 when TARGET = 0. Since people receive loans in a quantity corresponding to the purpose for which they would use them, this is to be expected.

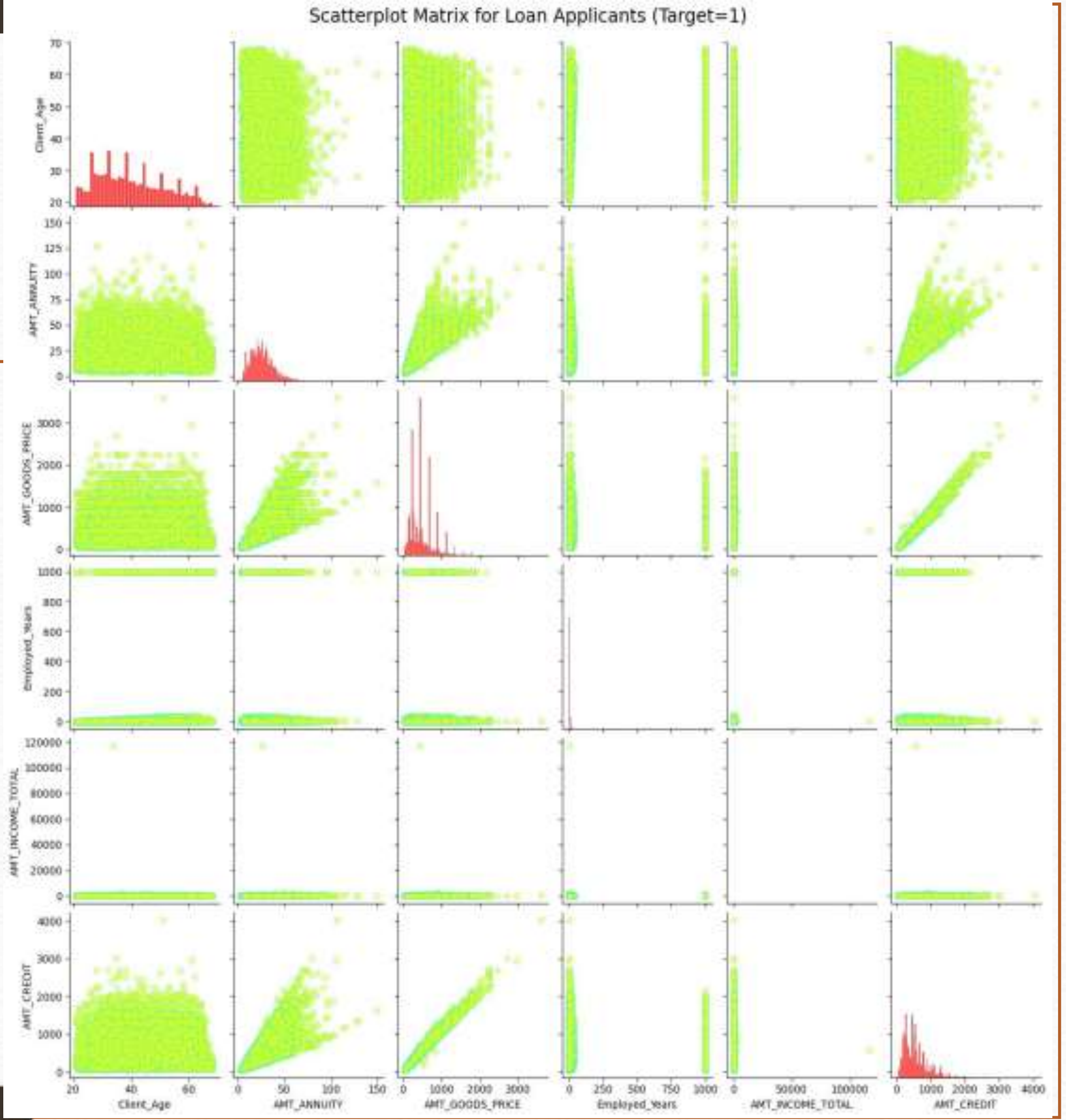
2.Since the annuity amount will be proportionate to the loan amount, AMT_ANNUITY and AMT_CREDIT also have a significant correlation.

3. As a result, there is a strong correlation between AMT_ANNUITY and AMT_GOODS_PRICE indirectly.



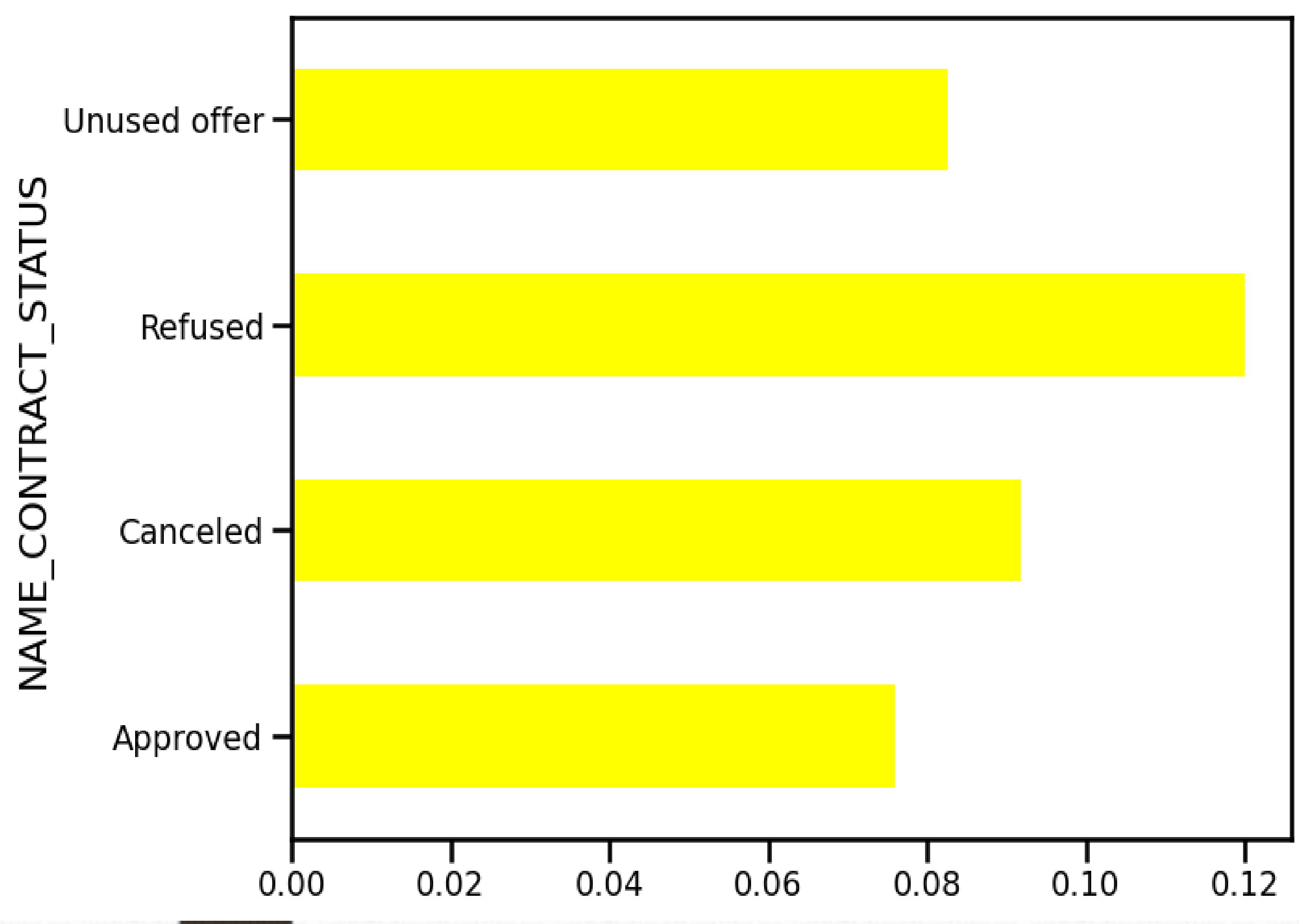
Bivariate analysis when TARGET = 1

When total revenue exceeds 10 Cr., there is one obvious exception that could skew our results. Consequently, we should examine the portfolio without that client and instead focus on that client's debt.



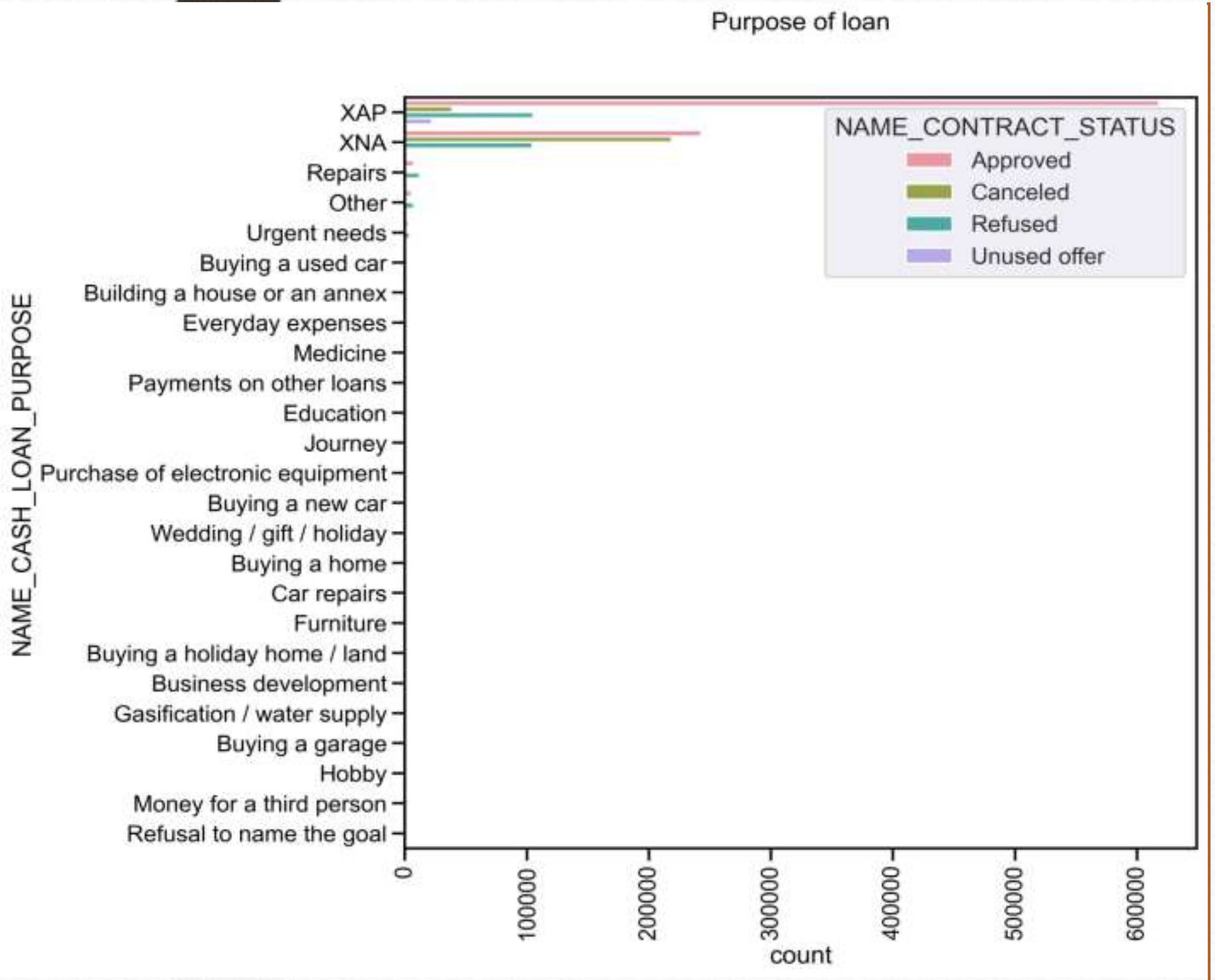
NAME_CONTRACT_STATUS versus the proportion of clients experiencing payment issues

Previously denied customers now fall into the 12% category of those having the most difficulties making payments.



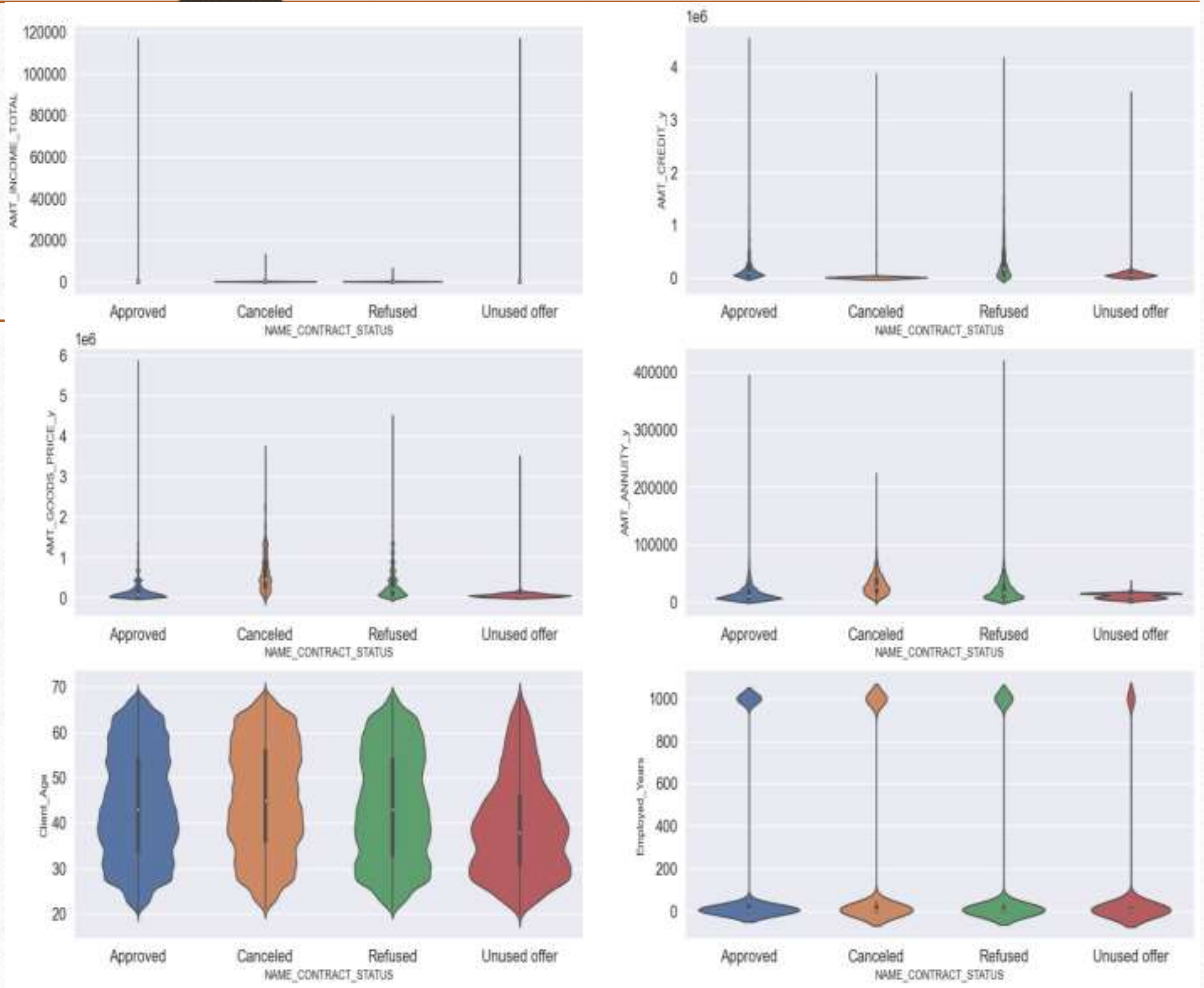
Purpose of the loan

'Repairs' accounted for the majority of loan rejections.



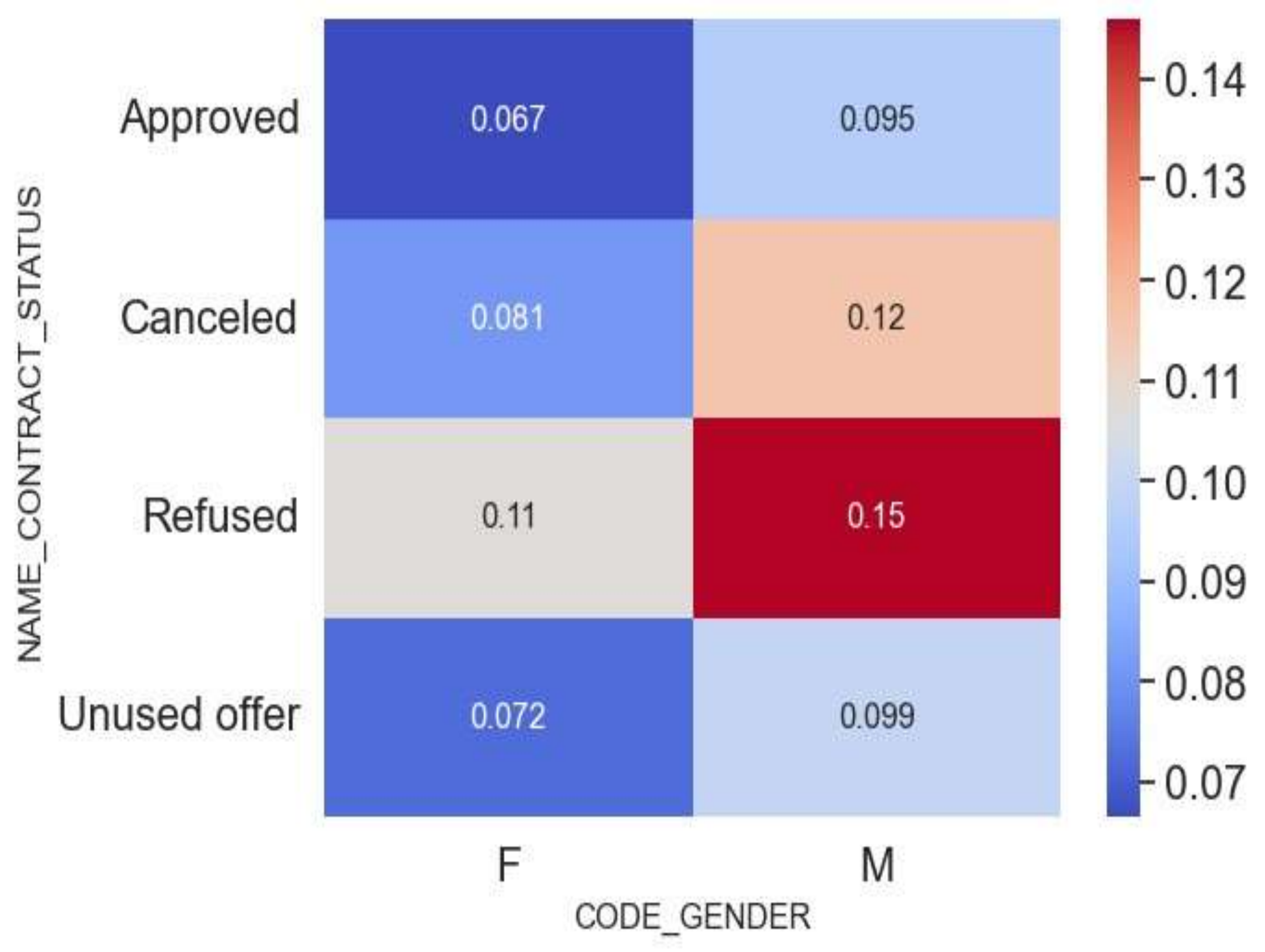
Univariate analysis on numerical variables

With one exception of CLIENT_AGE, all four cases demonstrate a similar distribution but in various age ranges, such as a lower age range for Unused Offers with a small spread and a greater age range for Cancelled Loans with a broader spread. All variables seem to have a similar distribution for all cases.



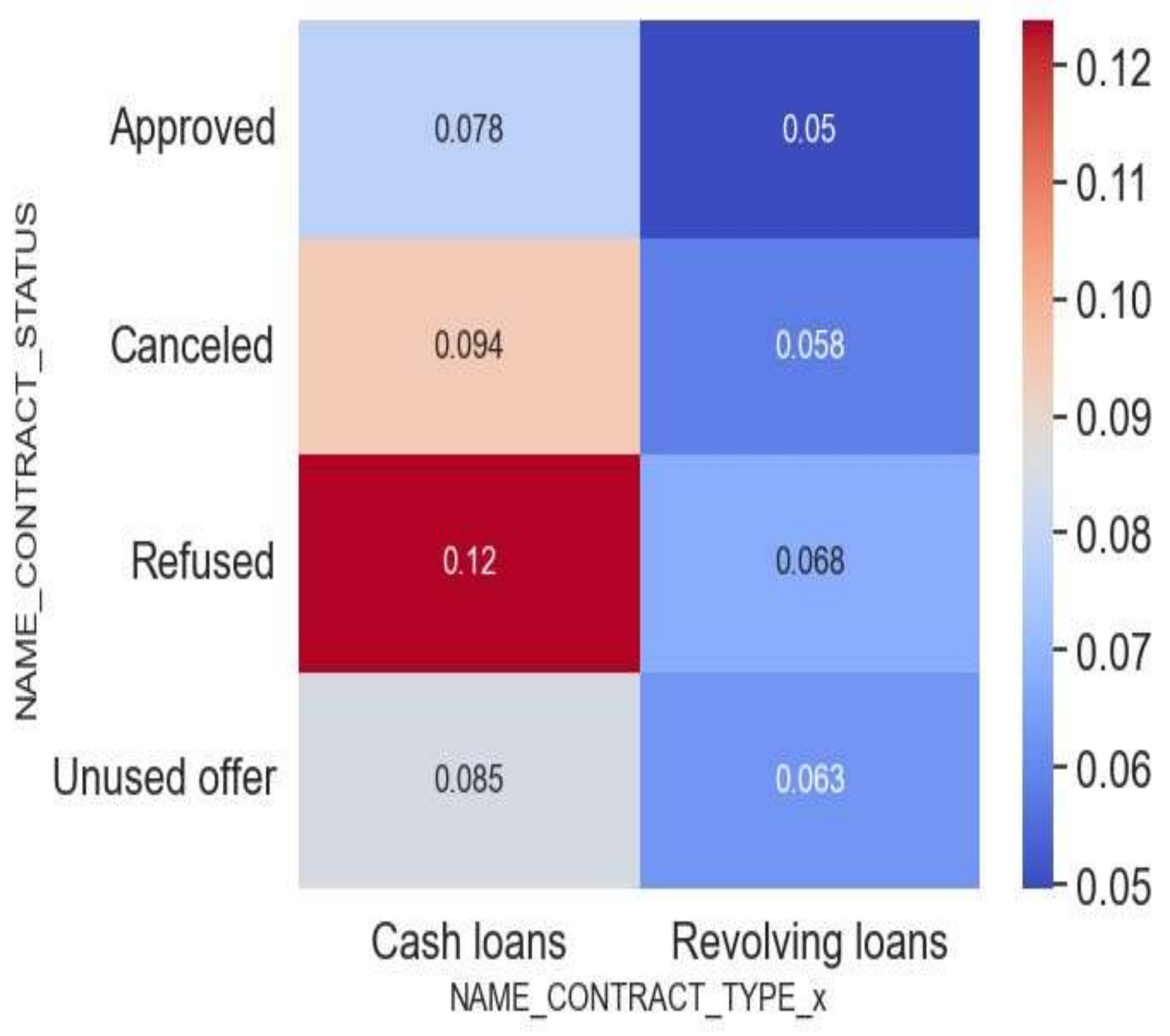
Gender *vs* **Previous** ***contact status*** *vs* ***Target variable***

- 1. Clients of cash loans are more likely to experience payment difficulties, with previously rejected loans having the highest percentage at roughly 12%.
- 2. Since revolving loans have a lower rate of customers having payment difficulties– the lowest rate is for loans that have already been approved–they are a better alternative for the bank to market.



Contract type vs Previous contact status vs Target variable

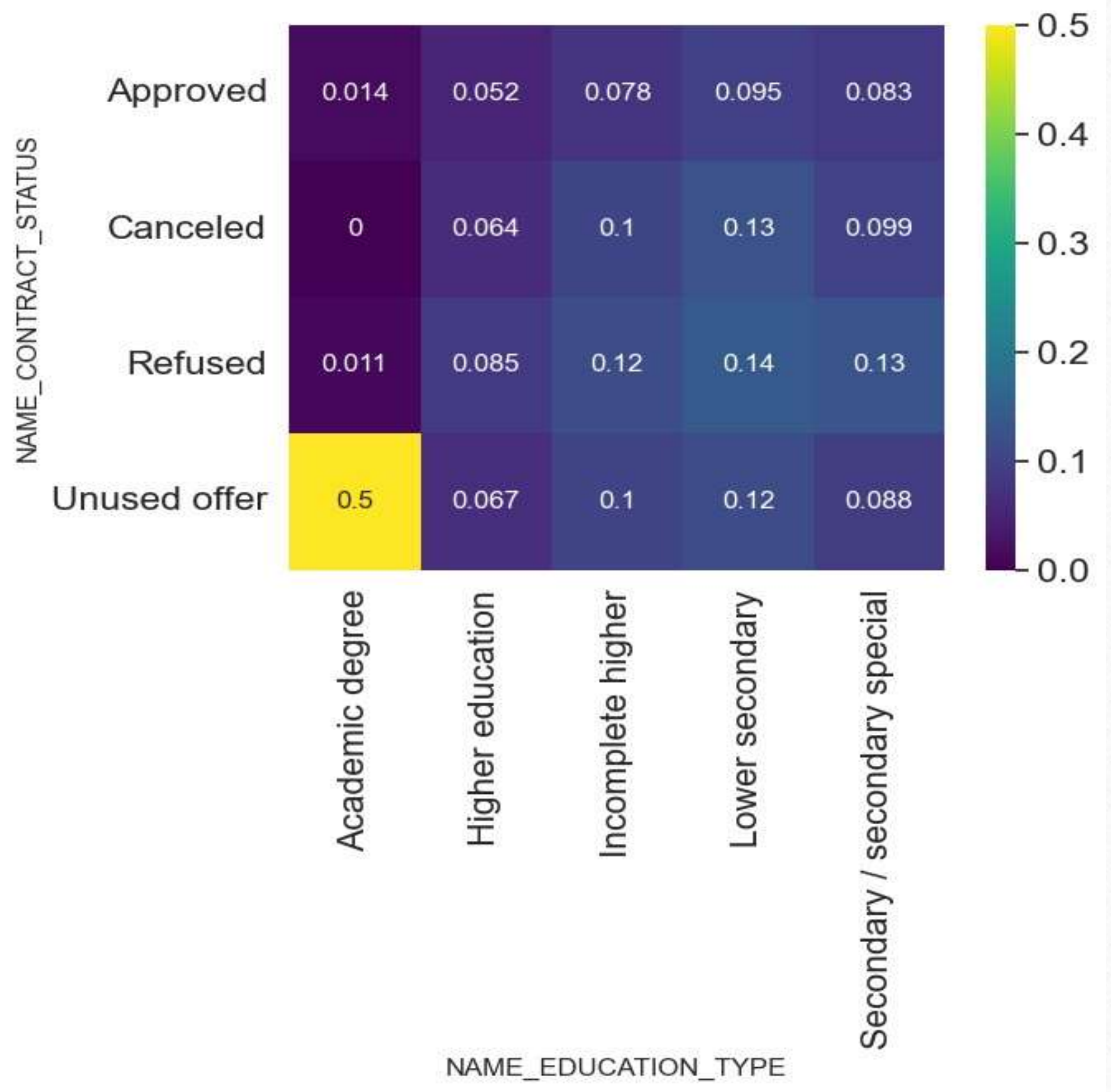
- Clients who take out cash loans are more likely to experience payment difficulties; the greatest percentage is for previously declined loans, which is roughly 12%.
- Revolving loans have a lower rate of customers having payment troubles than other types of loans, with the lowest rate being for loans that have already been approved, making them a better choice for the bank to market



Education type vs Previous contact status vs Target variable

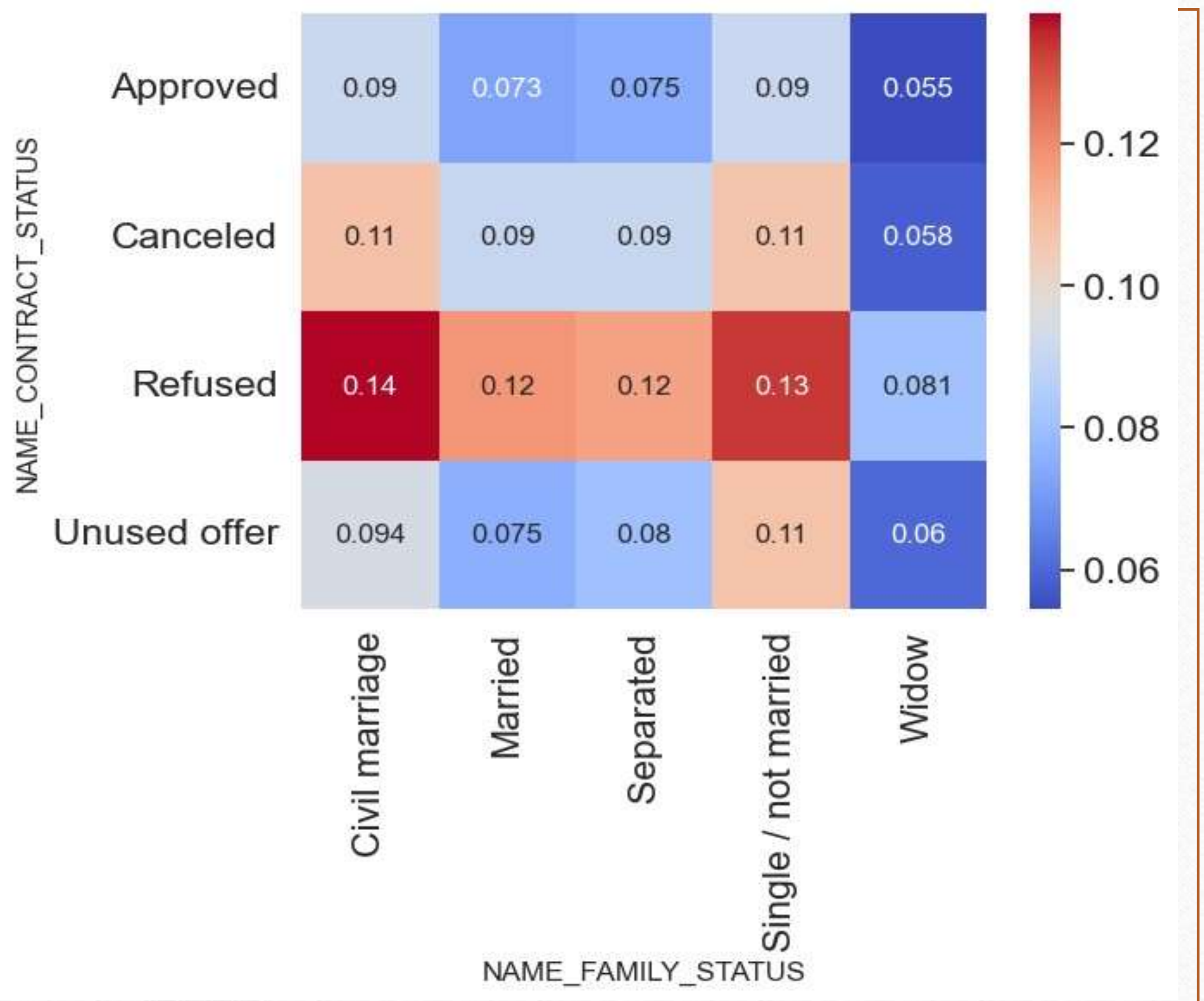
In comparison to other clients, those who earned lesser secondaries pose larger risks to the bank.

Clients who previously declined the offer are at the greatest danger, since around 50% of them may experience payment difficulties.



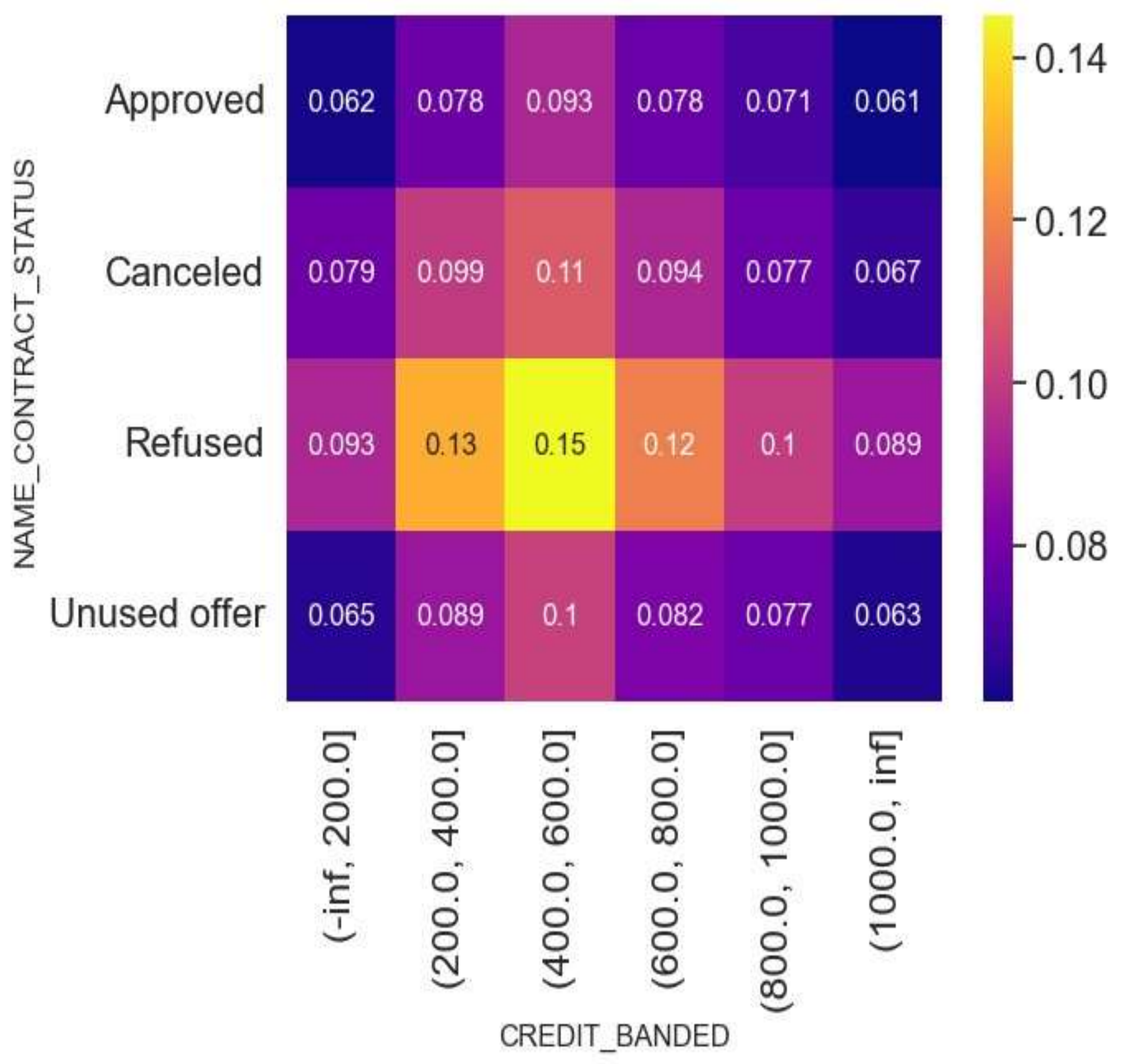
Family status vs Previous contact status vs Target variable

- Civilly wed and single people run the greatest danger.
- In general, widows pose the lowest risk because they consistently result in timely payments.



Credit amount vs Previous contact status vs Target variable

- For the bank, the riskiest loan amount is between 4 lac to 6 lac which bank should avoid.



Conclusion

- Banks should concentrate less on income kinds like "Working," which have a greater rate of failed payments, and more on contract types like "Student," "Pensioner," and "Businessman" for lucrative company. The organization will suffer less financial loss as a result.
- Loans to clients between the ages of 17 and 35 should be avoided, and loans to widows should be favored. Loan applicants whose applications have previously been denied should also be avoided. The 4 to 6 lac loan amount range is the riskiest and should be avoided.
- Cash loans are not the best choice because they carry greater risk, but revolving loans do not. Males should not be chosen above females. Clients with less than a high school diploma should be avoided.
- For effective payments, banks should concentrate on contract kinds like "student," "pensioner," and "businessman," with dwelling types other than "co-op apartments." 'Repair' loans have a higher percentage of late payments that are unsuccessful. The least dangerous housing type is "With parents," thus banks should make an effort to draw more customers from this group.
- These findings generally imply that banks should concentrate on customers who pose less of a risk in terms of their background and financial history. Banks can minimize financial losses and increase profits by avoiding riskier customers and loan types.