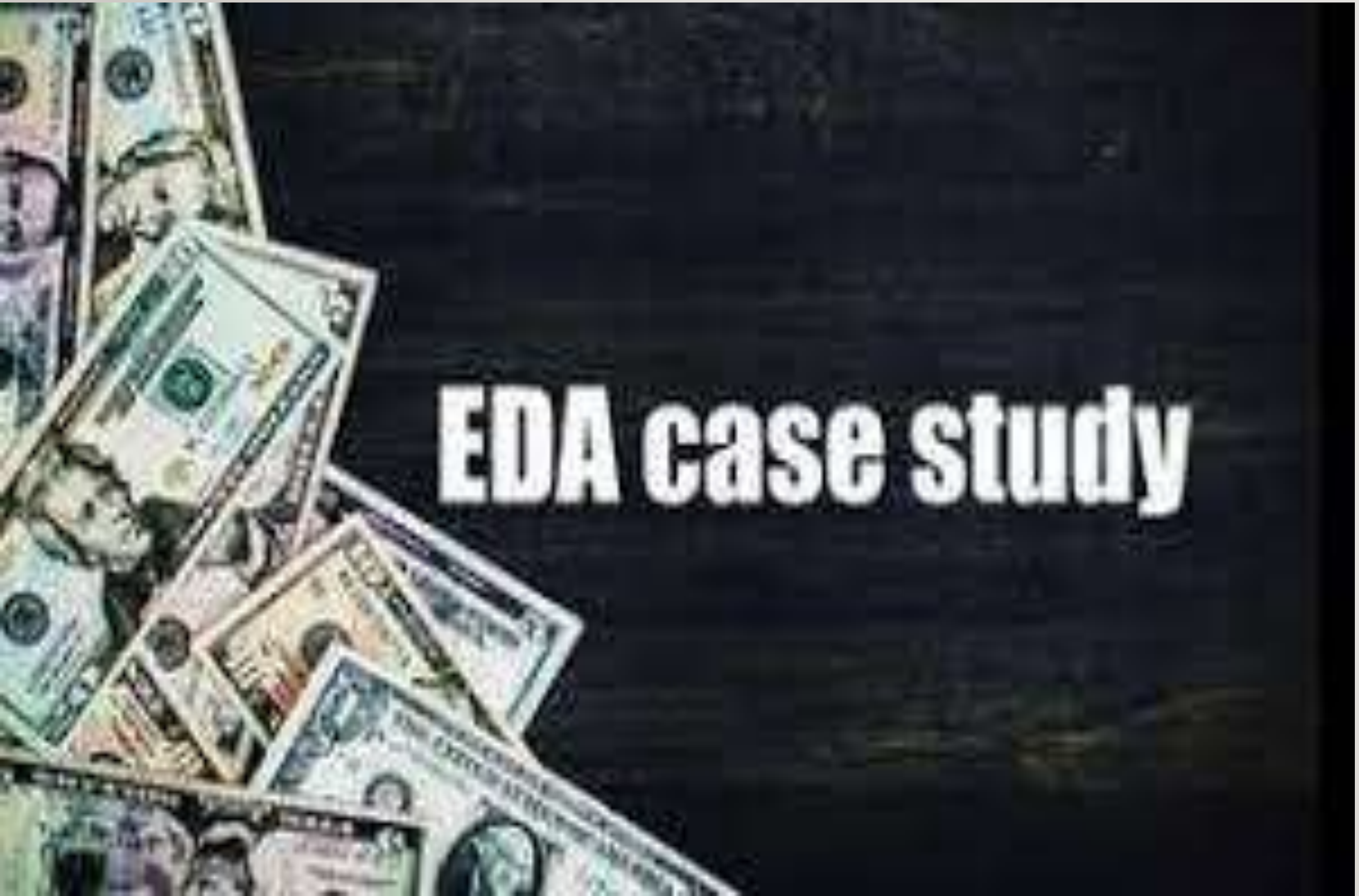
CREDIT EDA CASE



By Pankaj Yadav

INTRODUCTION:

This case study is designed to give an understanding of the application of EDA in a real business scenario. In this case study, in addition to applying the techniques you learned in the EDA module, you will gain a basic understanding of risk analysis in banking and financial services and understand how data can be used to reduce the risk of loss, while lending to customers.

BUSINESS UNDERSTANDING - 1

- Lenders find it difficult to lend to them because their credit history is poor or non existent. As a result, some consumers use this to their advantage by becoming defaulters.
- Suppose you work for a consumer finance company that specializes in providing various types of loans to urban customers. You should use EDA to analyze the pattern found in the data. This will ensure that applicants who are able to repay their loans will not be rejected.
- When a company receives a loan request, the company must decide whether to approve the loan based on the applicant's profile. The bank's decision carries two types of risks:

1. If the applicant is unlikely to repay the loan, not approving the loan will result in a loss to the business.
2. If it is unlikely that the applicant repays the loan The loan will not be repaid. ready, that is it is likely to default, approval of loan may result in financial loss to the business.

BUSINESS UNDERSTANDING - 2

The data below contains loan application information at the time of loan application. It contains tw
o types of
 scenarios:
• Customer in difficulty with payment: he is late in payment by more than X days on at least one of the Y previous
 installments of the loans in our sample.
• All other cases: All other customers whose payment status is on time.

When a customer applies for a loan, the customer/company can make four types of decisions:

1. **Approval**: The company has approved the loan request.

2 .
Canceled: The customer canceled the request at some point during the approval period. Either the
 client changed
 their mind about the
loan or, in some cases, they got a lower price that they didn't want due to the client's higher risk.

3.
Rejected: The company refused the loan (because the customer did not meet its requirements, etc.
).

4. **Unused Offers**: Loans canceled by customers, but at different stages of the process.

BUSINESS OBJECTIVES:

- This case study seeks to identify patterns that indicate whether customers are ha ving difficulty paying installments, which can be used to take action such as refusing loans reducing loan amounts, lending (to high-risk applicants) for plus high interest rates this will ensure that consumers who are able to repay loans are not rejected. Identifying these candidates using EDA is the goal of this case study.

- In other words, the business wants to understand the drivers (or determining vari ables) behind defaults, that is, the variables that are good indicators of defaults. C ompanies can use this knowledge for their portfolio and risk assessments.

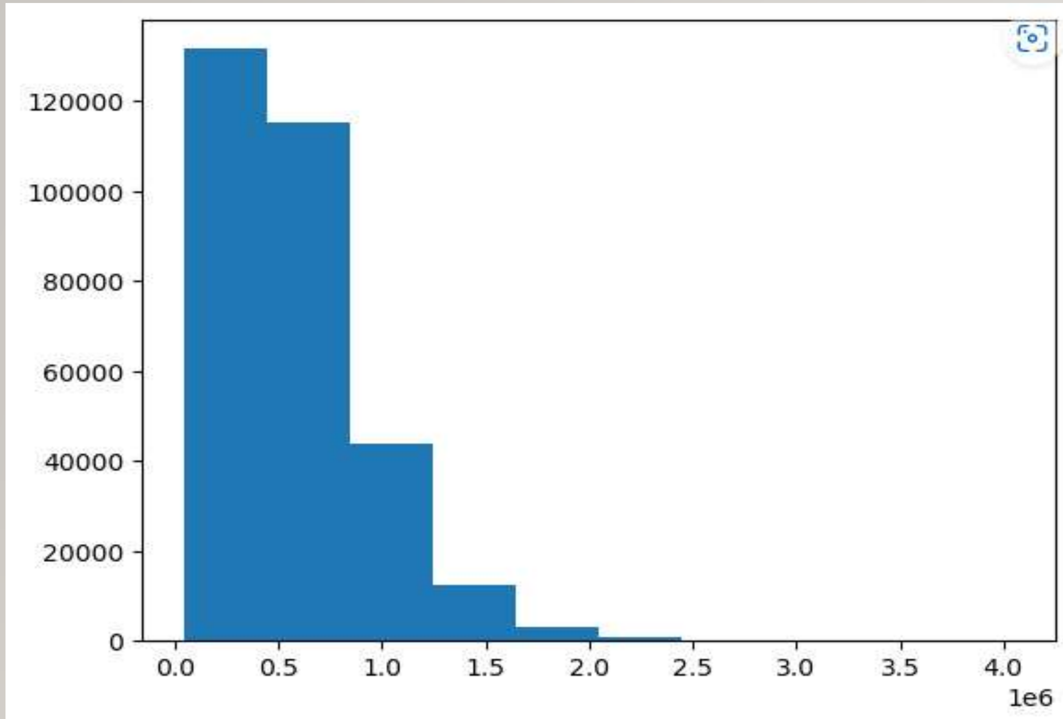
DATA UNDERSTANDING:

This data set has 3 files, explained as follows:
1. 'application_data.csv' contains all customer information at the time of application. If customer has payment problem.
2. 'previous_application.csv' contains information about the customer's previous loan data. It contains data indicating whether the previous applicati on was an **approved, canceled, declined, or unused offer.**
3.'columns_description.csv' is a data dictionary describing the meaning of the variable.

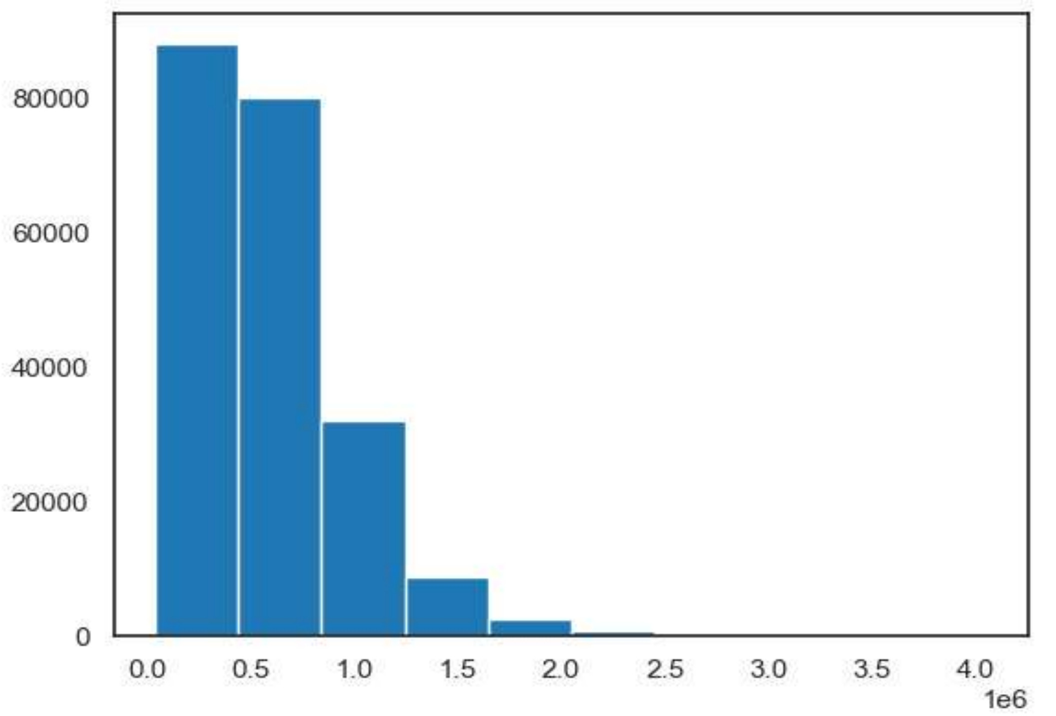
DEALING WITH MISSING VALUES:

Missing values are a common issue in datasets, and they can cause problems in data an and modeling. There are several ways to impute missing values in Python, some of whic

- **Mean/Median/Mode Imputation.**
- **Forward/Backward Filling.**
- **Linear Interpolation.**
- **Multiple Imputation.**
- **K-Nearest Neighbour Imputation.**
- **Model-Based Imputation.**



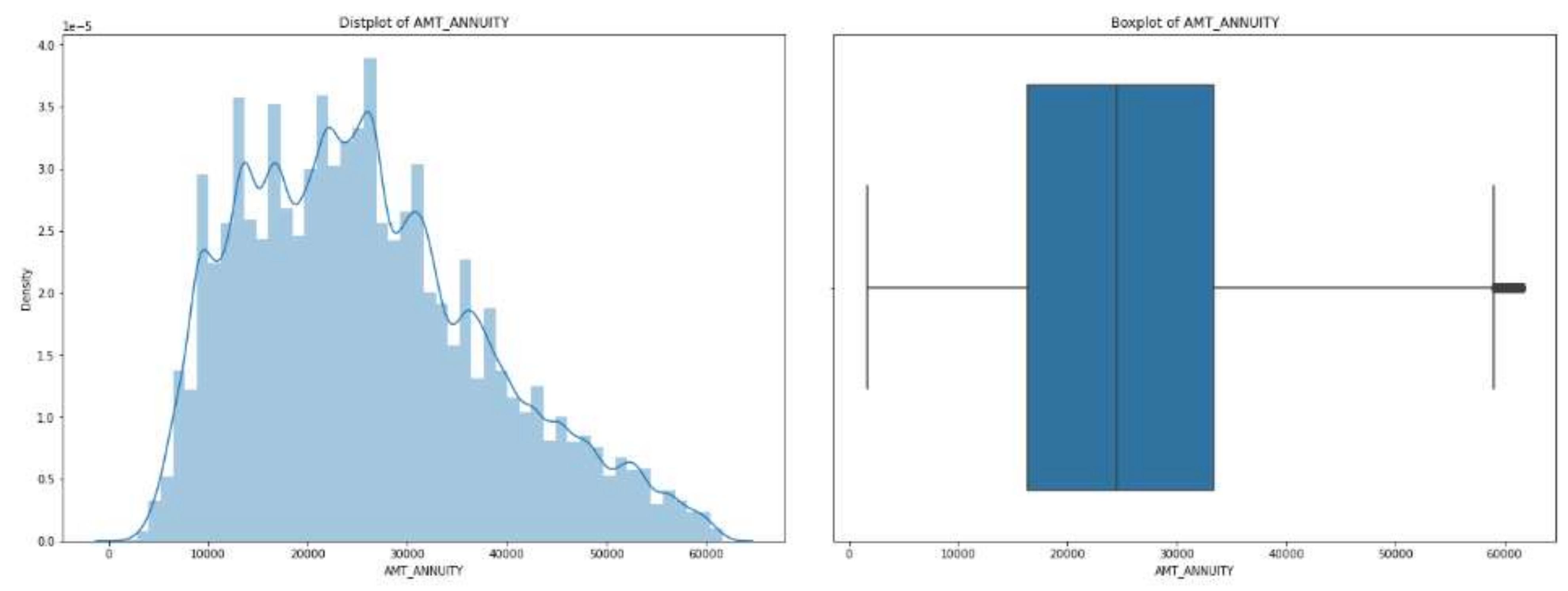
BEFORE



AFTER

Analysis of [AMT_GOODS_PRICE]

There is a difference between the min and max values, so we will use the median to impute missing values since the mean will skew the data.



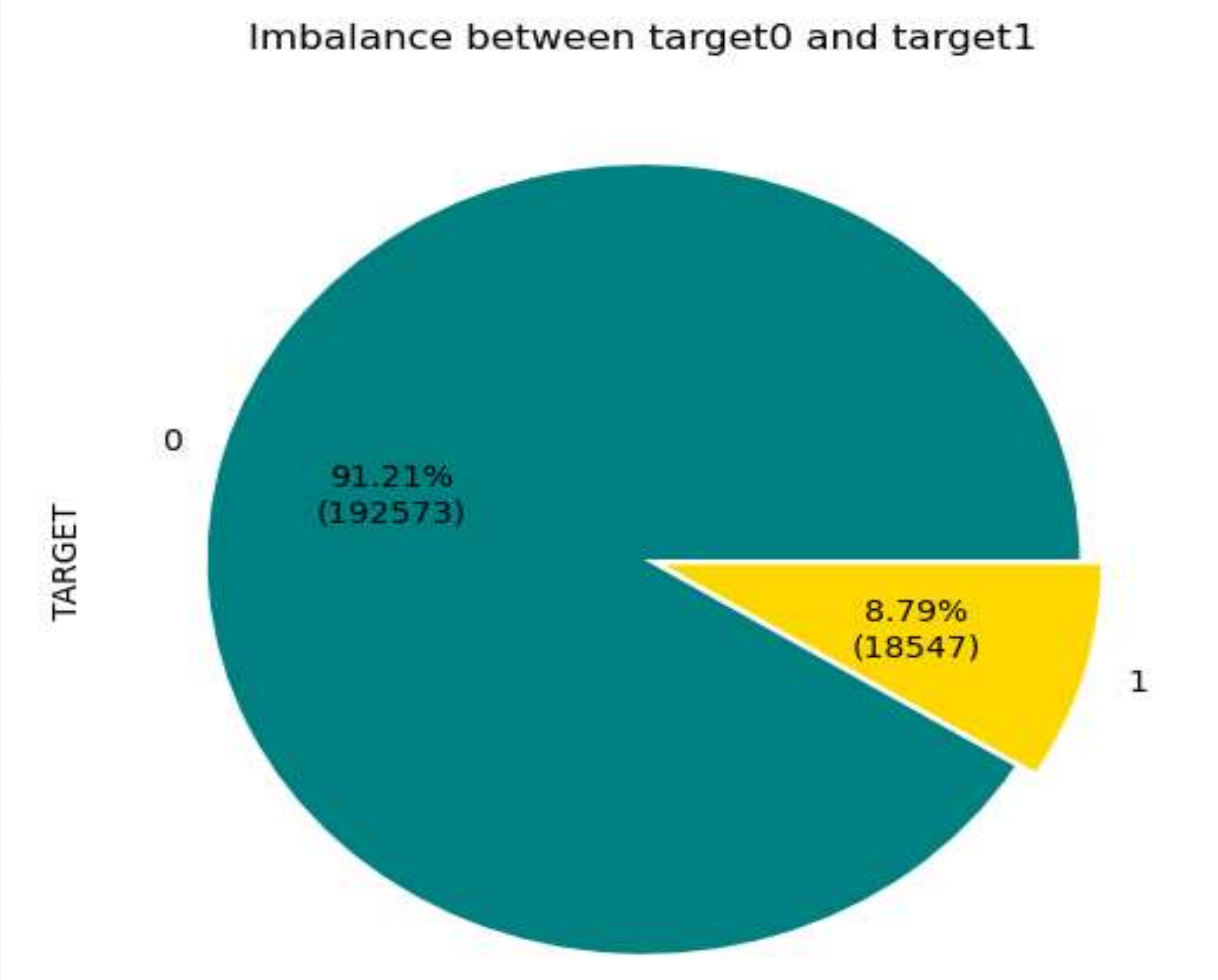
Analysis of `AMT_ANNUITY`

As observed from distplot and boxplot, the outliers tend to exist after 61704.

Applicants with `AMT_ANNUITY` above 61704 (calculated using IQR) are outliers.

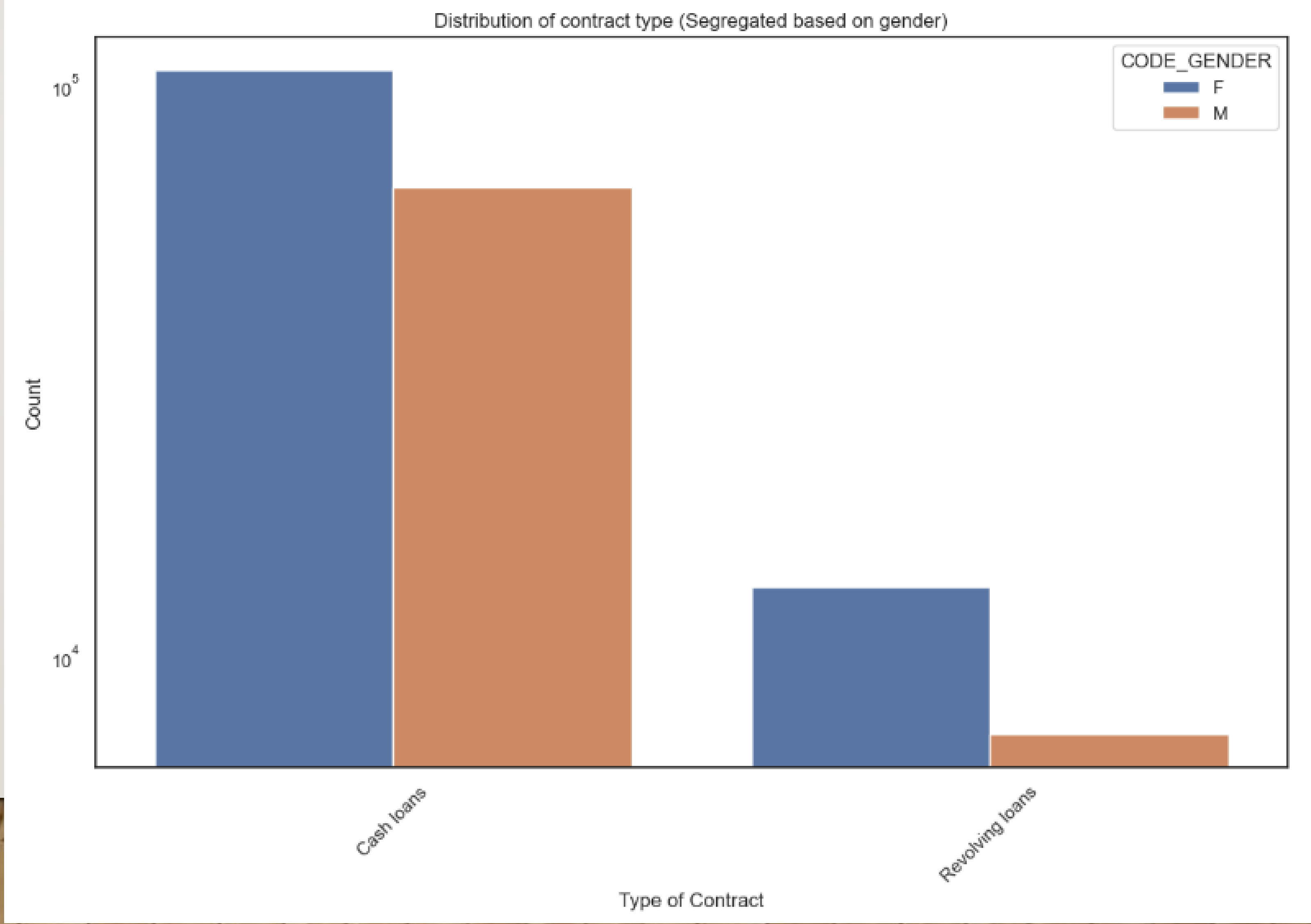
Analysis of INCOME_RANGE

8.78% of customers are customers with payment problems. 91.21% of customers fall into the "all other cases" category



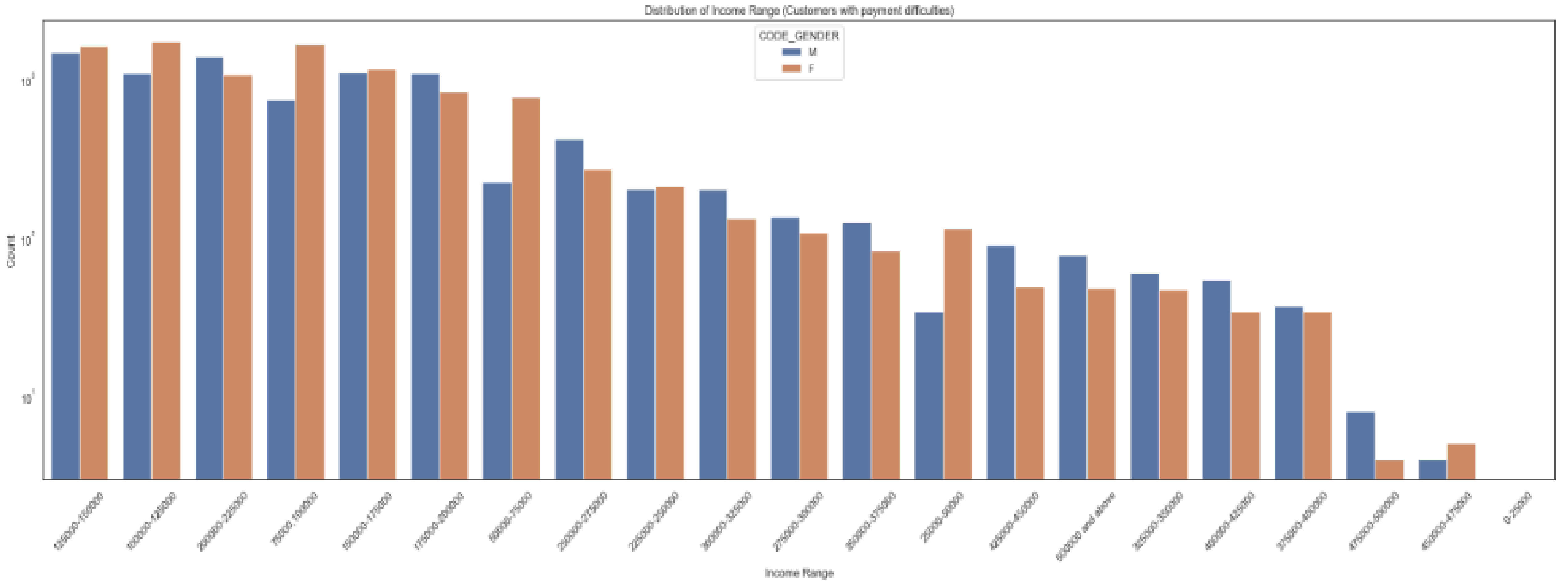
Analysis of NAME_CONTRACT_TYPE

It is obvious that cash loans have more customers per loan than revolving loans. In both cases, there were more female customers than male customers.



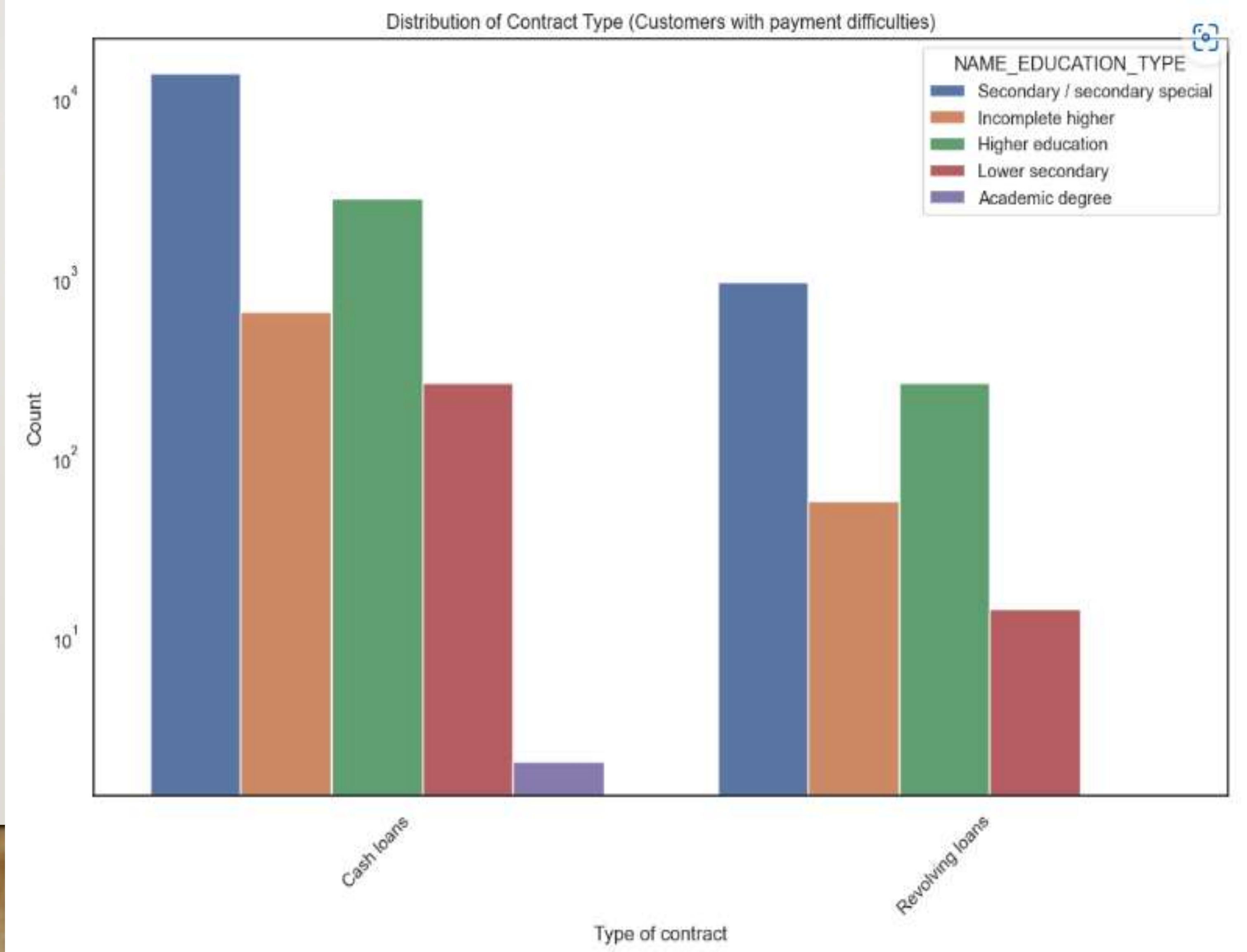
Analysis of AMT_INCOME_RANGE

100,000 to 200,000 income, the highest credit limit. There are far fewer than income brackets of 400,000 and above.
On average, there are more male customers with less credit.



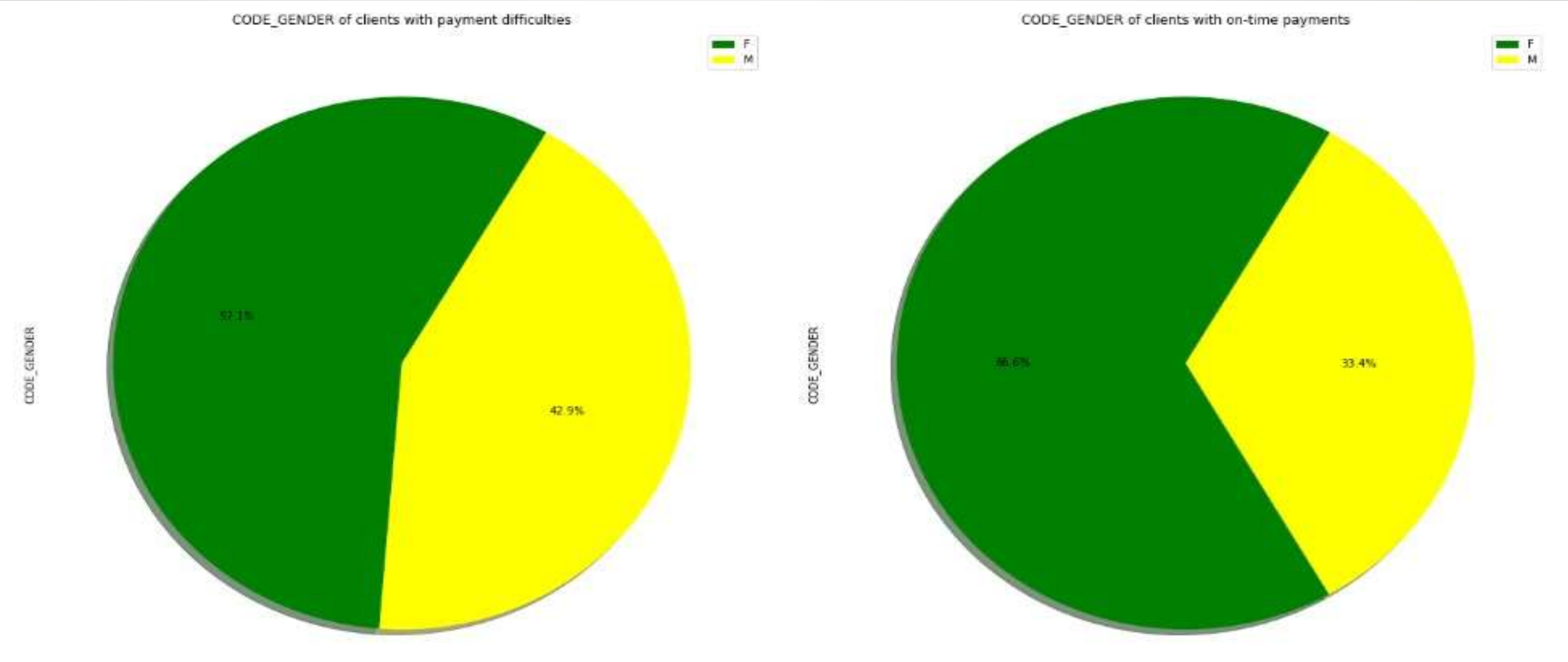
Analysis of NAME_EDUCATION_TYPE

1. As we have seen, cash loans are overwhelmingly preferred by customers of all educational levels.
2. People who only have diplomas don't like revolving loans at all.



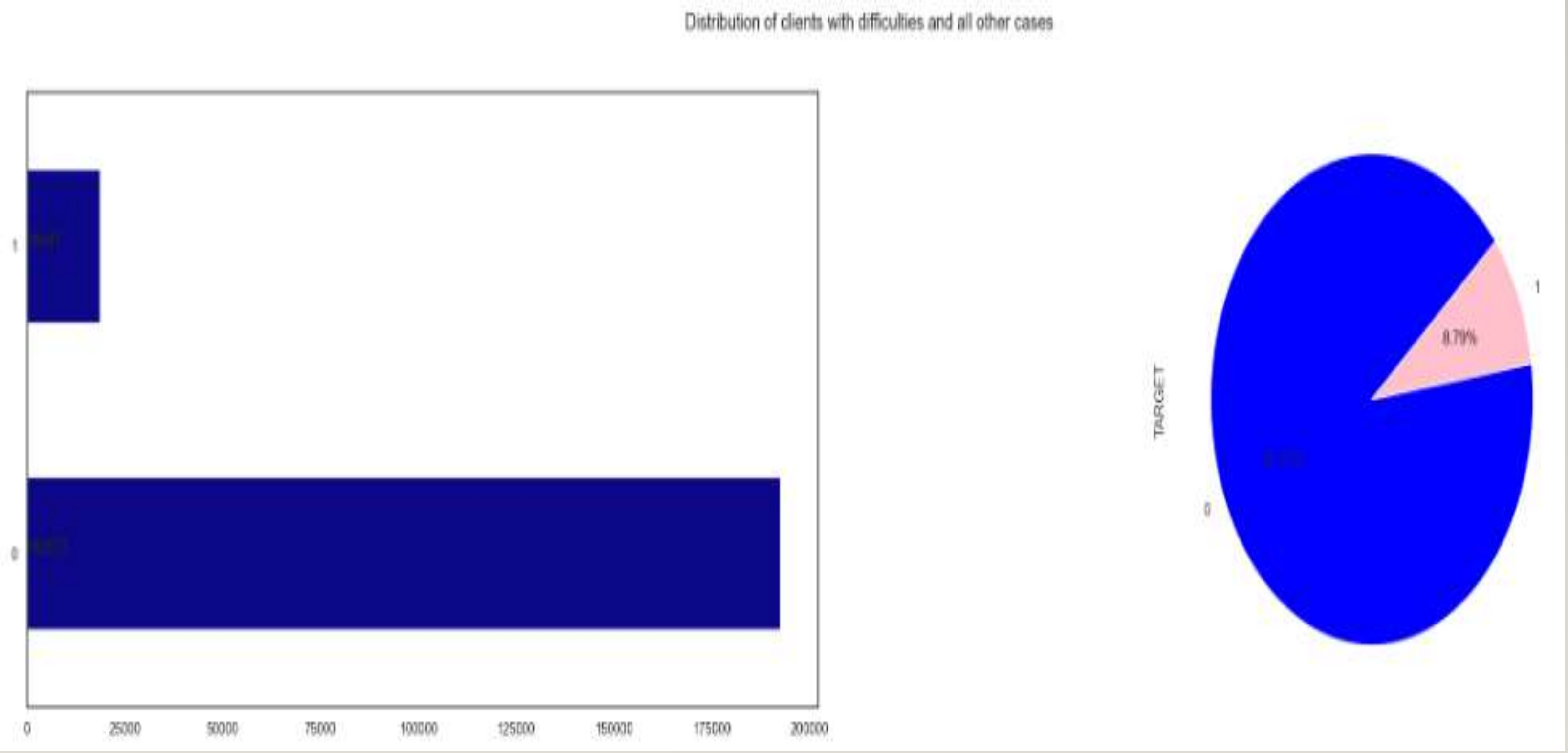
Analysis of `CODE_GENDER`

`CODE_GENDER` column provides a weak inference that "Male" clients have more payment difficulties



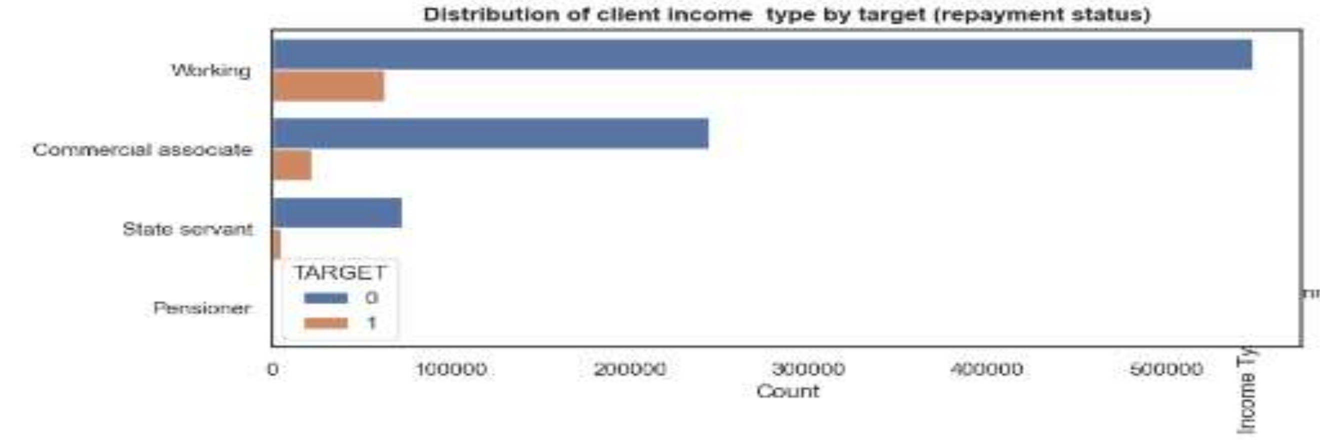
Analysis of
Distribution of clients with difficulties and all other cases

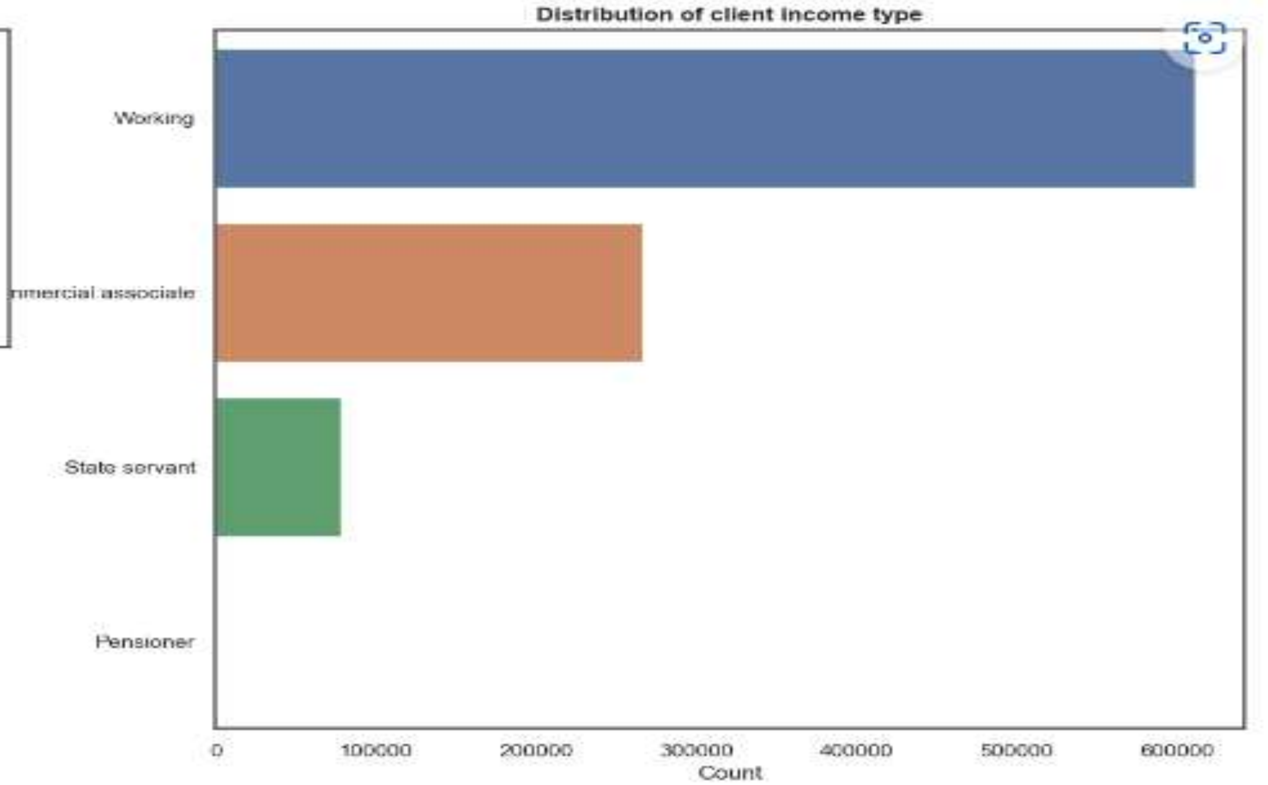
8.79% (18,547) of the total number of customers (192,573) are struggling to repay their loans.



Analysis of
Distribution of client income type

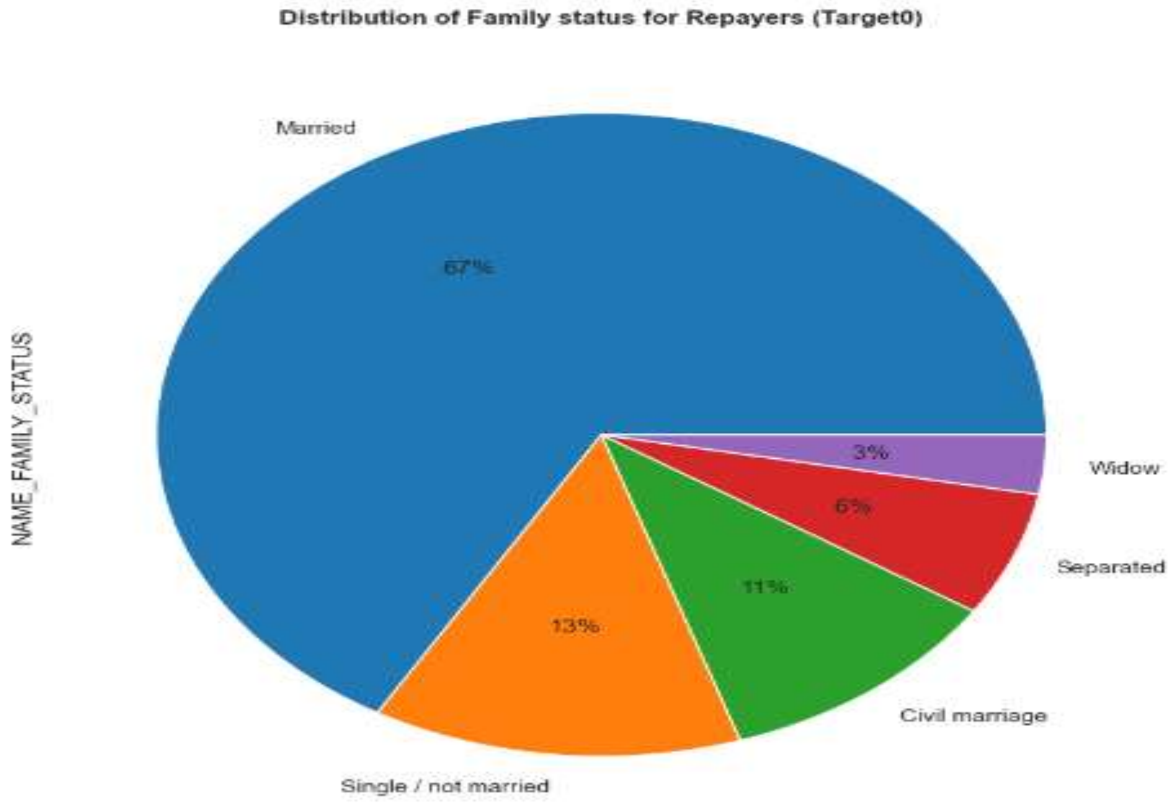
Most customers work based on their refund status in both cases.
Conversely, the fewest customers are retirees (retired customers)

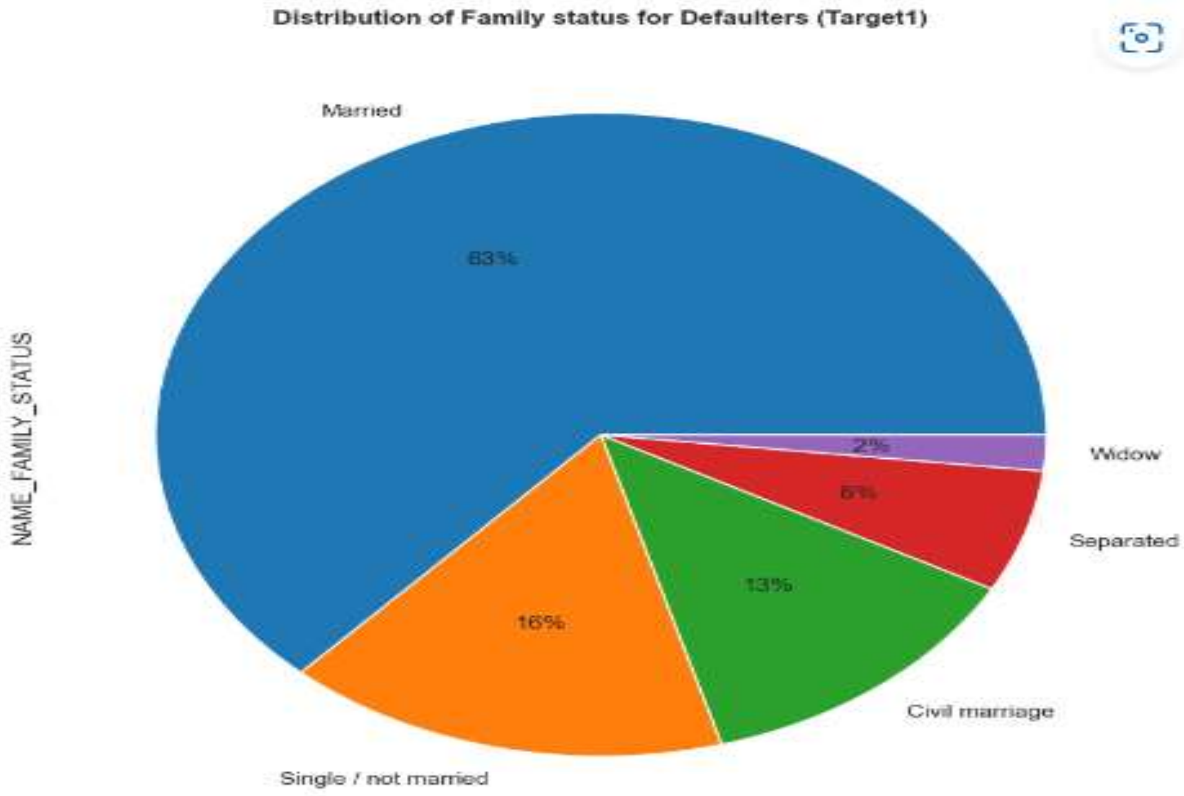




Analysis of NAME_FAMILY_STATUS

There was a -4% difference among married customers who had difficulty paying.
Marital Status in Both Repayment Situations Divided Almost Equally Marital Status (Family Members Living with Client)



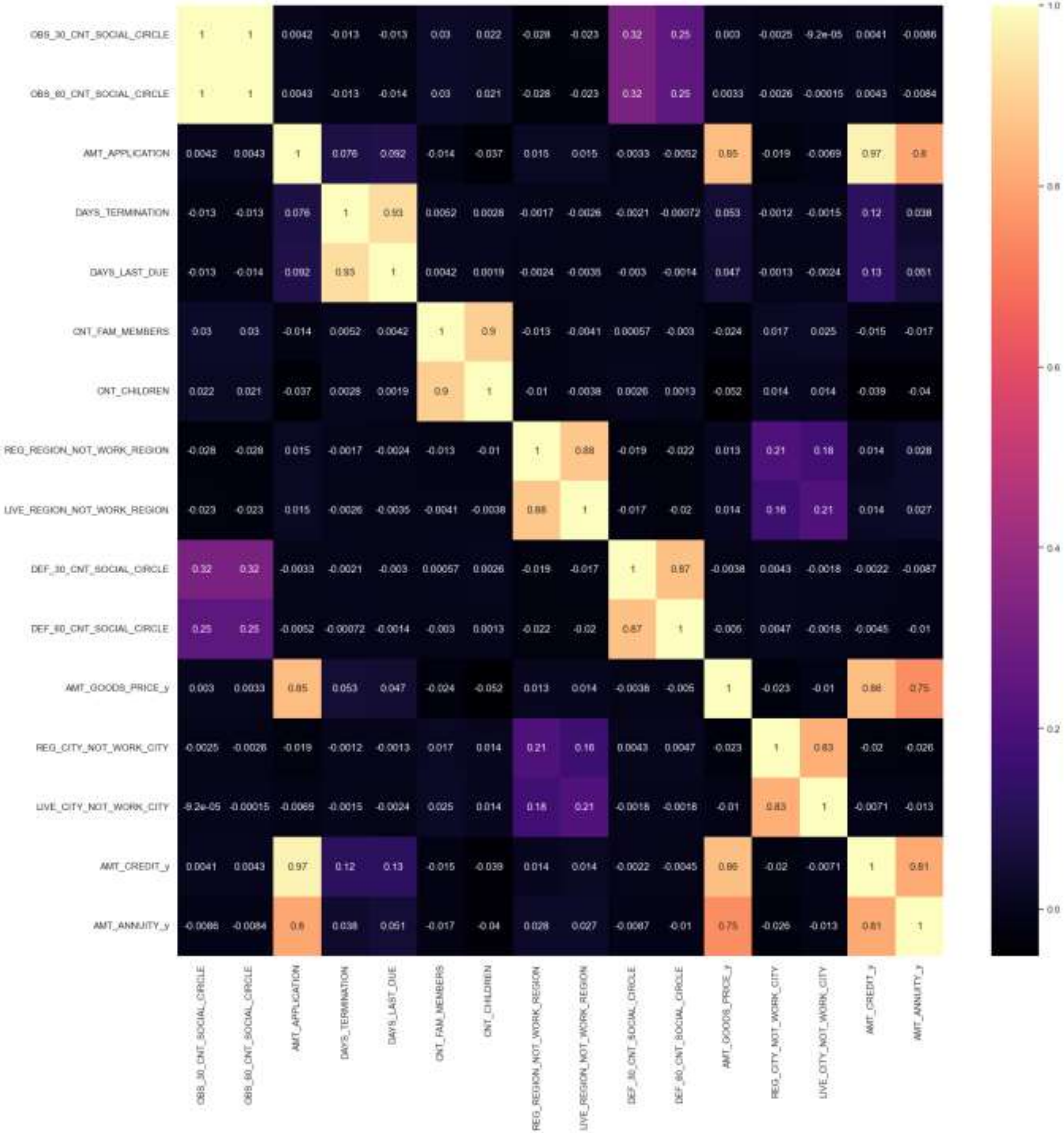


CORRELATION ANALYSIS OF NUMERICAL VARIABLES

Getting Top 10 CorelationWith Payment difficulties

- AMT_GOODS_PRICE AMT_CREDIT 0.98
- REGION_RATING_CLIENT REGION_RATING_CLIENT_W_CITY 0.96
- CNT_FAM_MEMBERS CNT_CHILDREN 0.89
- DEF_60_CNT_SOCIAL_CIRCLE DEF_30_CNT_SOCIAL_CIRCLE 0.87
- REG_REGION_NOT_WORK_REGION LIVE_REGION_NOT_WORK_REGION 0.85
- LIVE_CITY_NOT_WORK_CITY REG_CITY_NOT_WORK_CITY 0.78
- AMT_ANNUITY AMT_GOODS_PRICE 0.75
- AMT_ANNUITY AMT_CREDIT 0.75
- DAYS_EMPLOYED FLAG_DOCUMENT_6 0.62
- DAYS_BIRTH DAYS_EMPLOYED 0.58

Heatmap of corelated variable



OUTCOMES OF CORRELATION ANALYSIS

1. There is a strong correlation between AMT_GOODS_PRICE and AMT_APPLICATION, i.e. the higher the credit previously requested by the customer, the more it is proportional to the price of the product previously requested by the customer.
2. AMT_ANNUITY and AMT_APPLICATION also have a high correlation, meaning that the higher the loan annuity issued, the higher the price of the product the customer previously requested.
3. If the customer's contact address does not match the business address, chances are the customer's permanent address does not match the business address either.
4. The first result of the previous request is strongly correlated with the expected end of the previous request
5. CNT_CHILDREN and CNT_FAM_MEMBERS are strongly correlated, which means that customers with children will also have family members.

CONCLUSION

1. Clients who are Students, Pensioners and Commercial Associates with a housing type such as office/co-op/municipal apartments NEED TO BE TARGETED by the bank for successful repayments. These clients have the highest amount of repayment history.
2. Female clients on maternity leave should NOT be targeted as they have no record of repayments (therefore they are highly likely to default and targeting them would lead to a loss)
3. While clients living with parents have the least amount of repayor's, they also have the least amount of defaulters. So, in cases where the risk is less, such clients can be TARGETED.
4. Clients who are working need to be targeted LESS by the bank as they have the highest amount of defaulters.
5. Clients should NOT be targeted based on their education type alone as the data is very inconclusive.
6. Banks SHOULD target clients who own a car.
7. There are NO repayor's/negligible repayor's when the contract type is of revolving loan.
8. Banks SHOULD target more people with no children.
9. 'Repairs' purpose of loan is the one with the most defaulters and repayor's. Therefore, clients with very low risk SHOULD be given loans for such purpose to yield high profits.
10. Banks SHOULD also target female clients as they are the highest repayor's (almost as double as males) amongst both the genders.

THANK YOU!