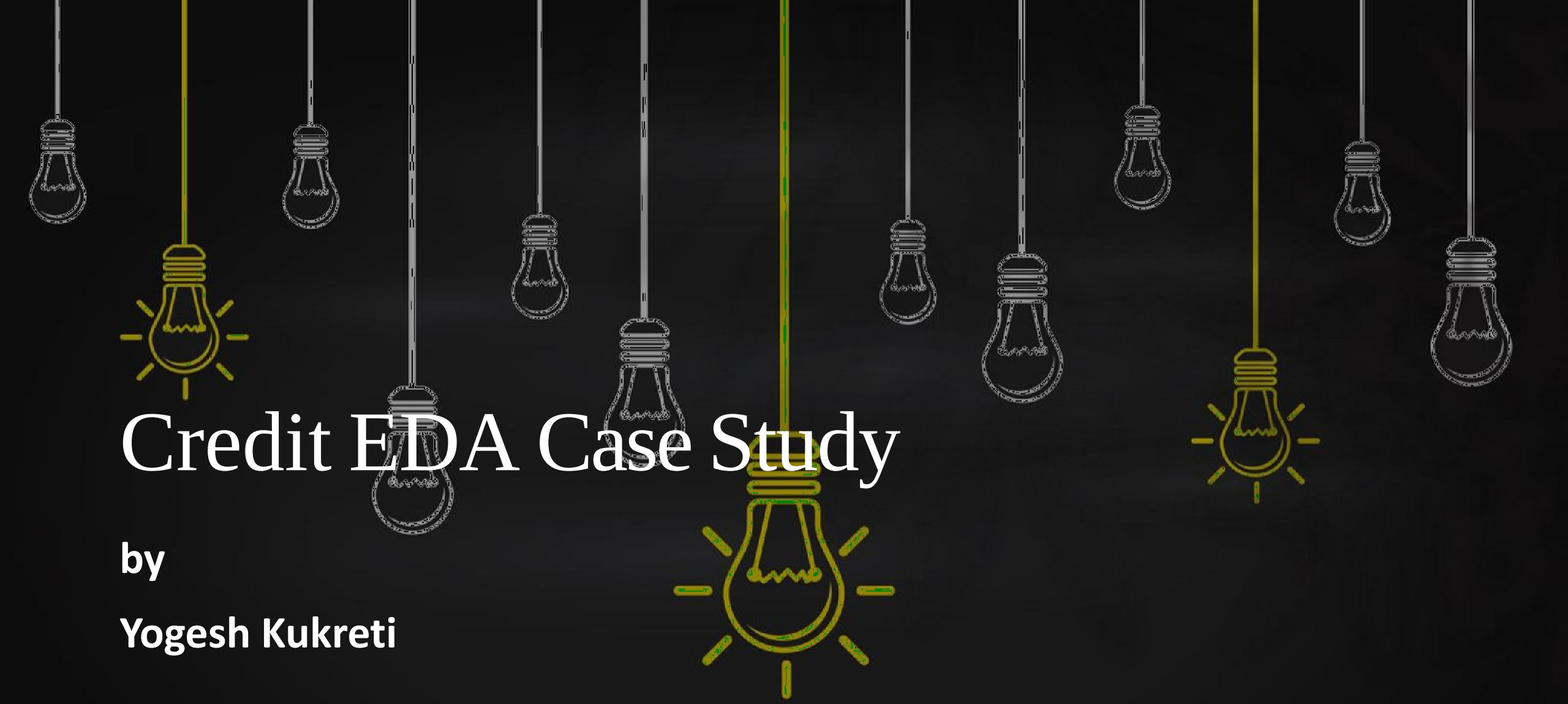


Credit EDA Case Study

by

Yogesh Kukreti

Introduction

- This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Understanding - 1

- 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 as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.
- 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:
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- 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.

Business Understanding - 2

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:

- **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,
- **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):

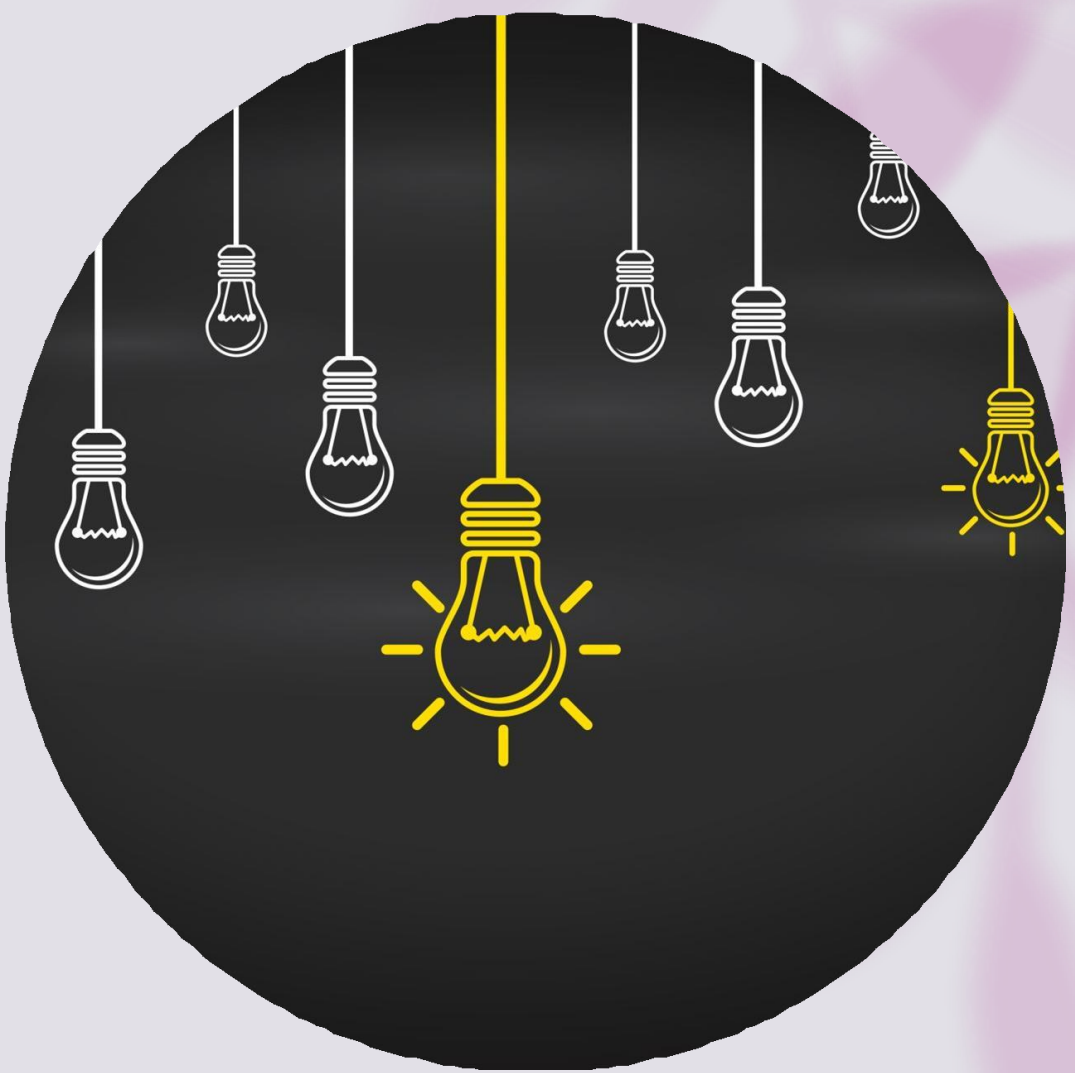
1. **Approved:** The Company has approved loan Application
2. **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
3. **Refused:**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

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

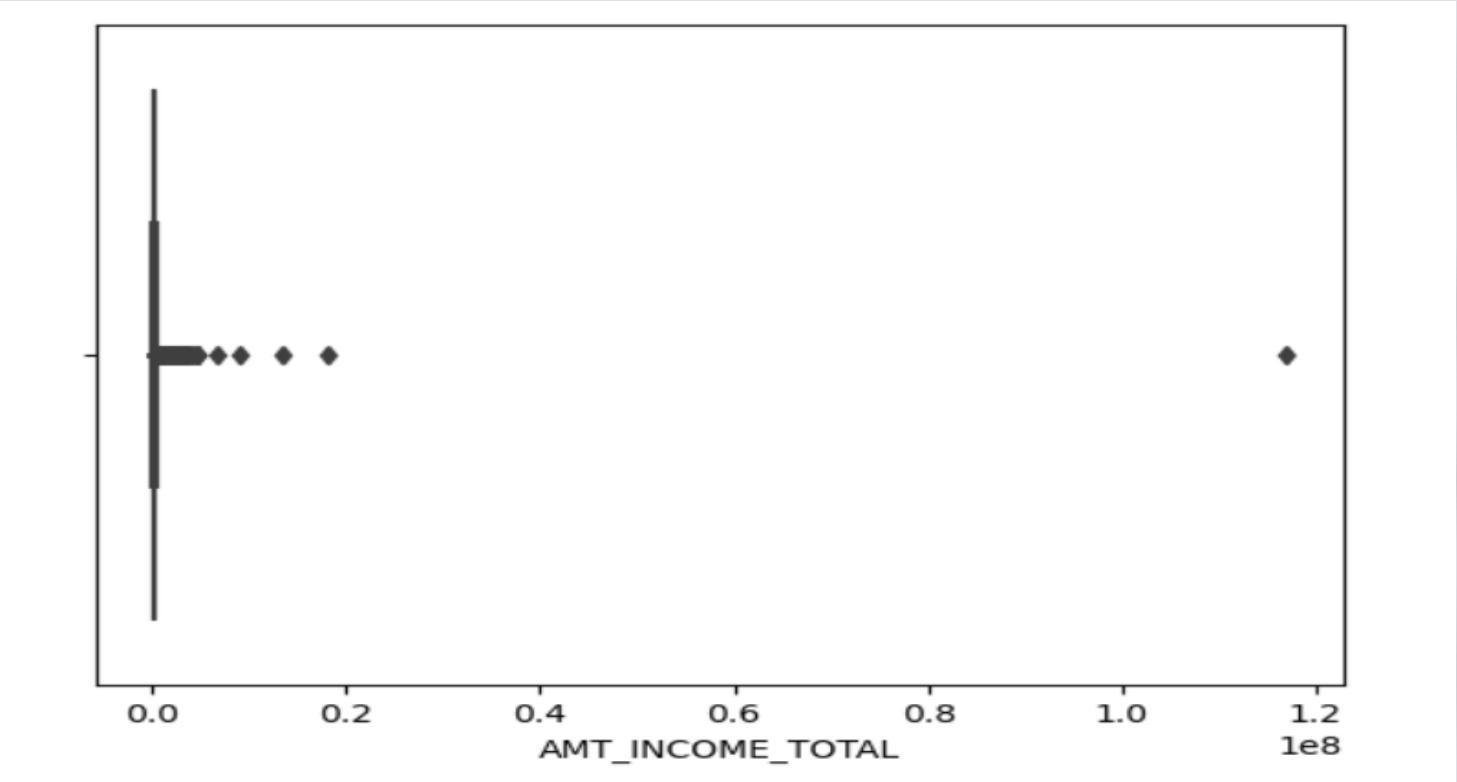
Business Objectives

- This case study 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 consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.
- In other words, 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.
- To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

Analysis of information of the client at the time of application

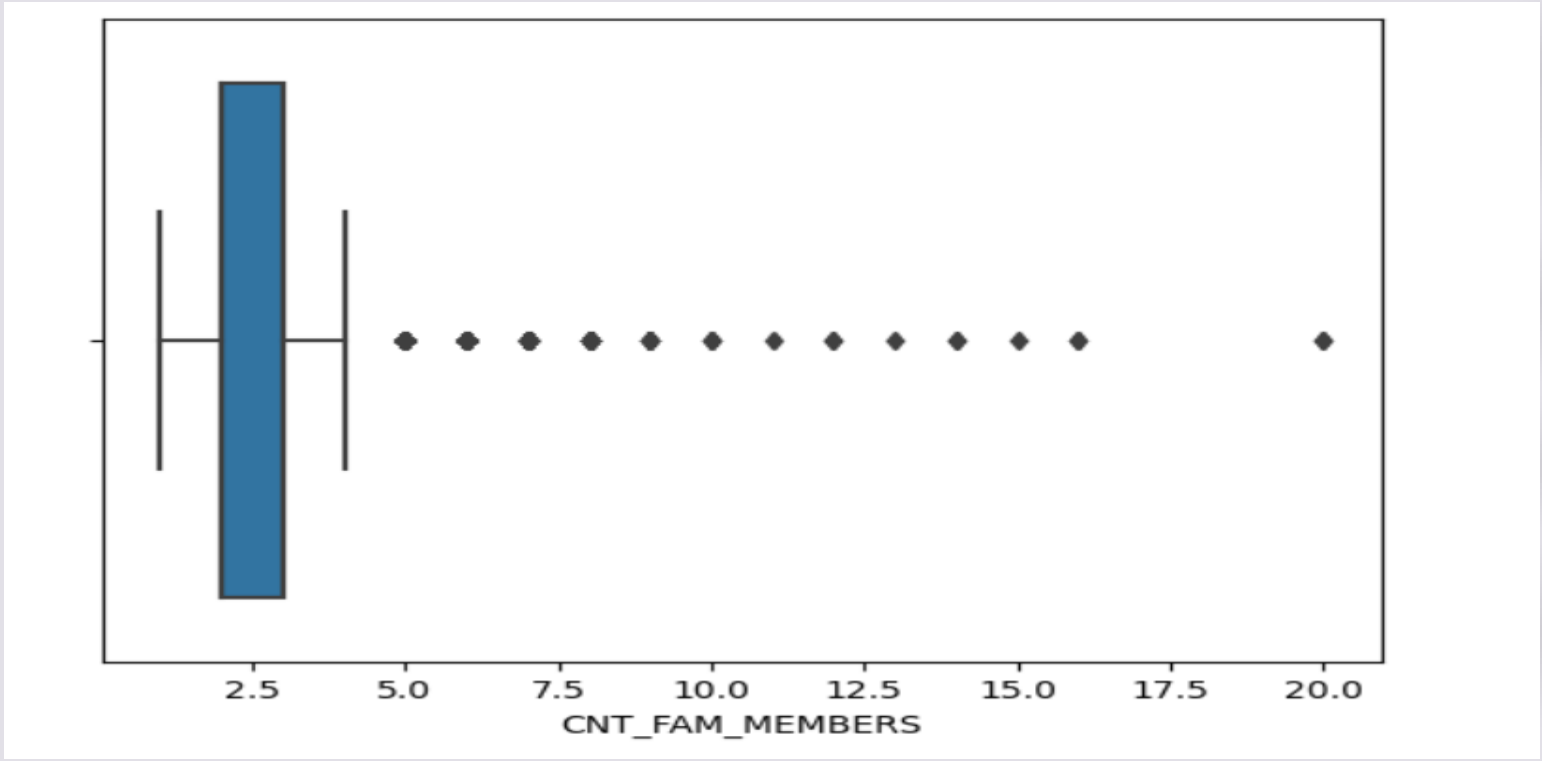


Outlier analysis



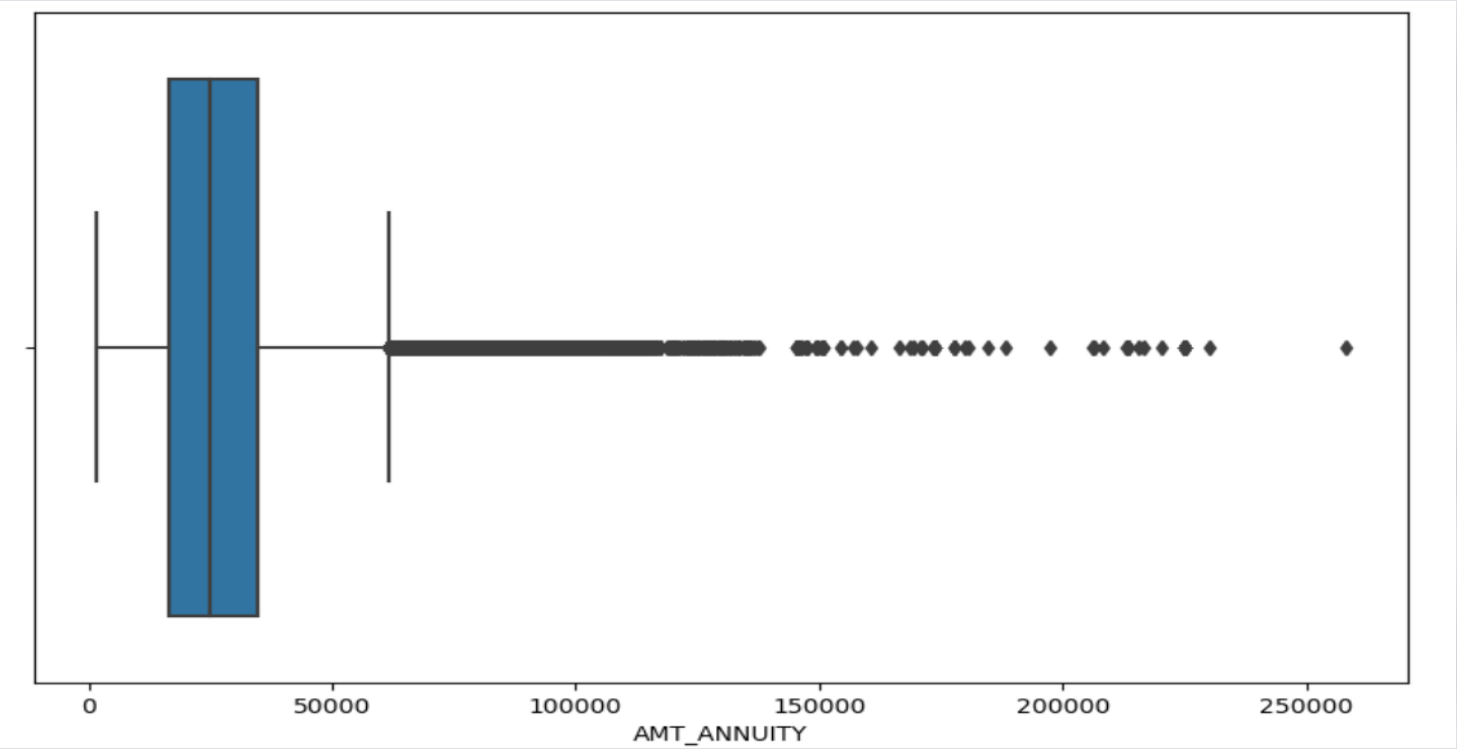
Analysis of `AMT_INCOME_TOTAL`

- Applicants with Income above 900K (99.9% value) are outliers



Analysis of `CNT_FAM_MEMBERS`

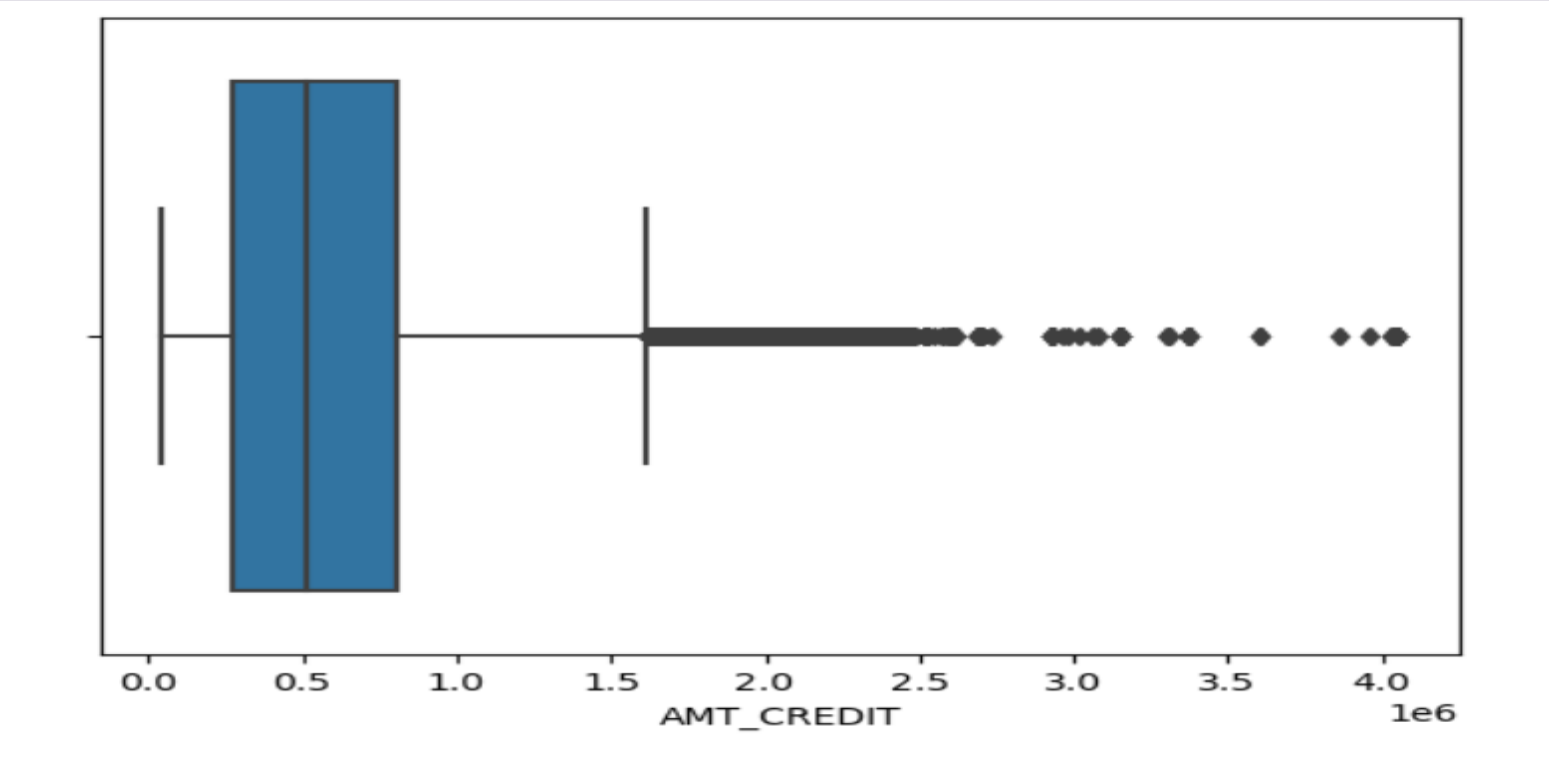
- Applicants with 5 or more family members are clearly outliers



Analysis of `AMT_ANNUITY`

- As observed from boxplot, the outliers tend to exist after 61704

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

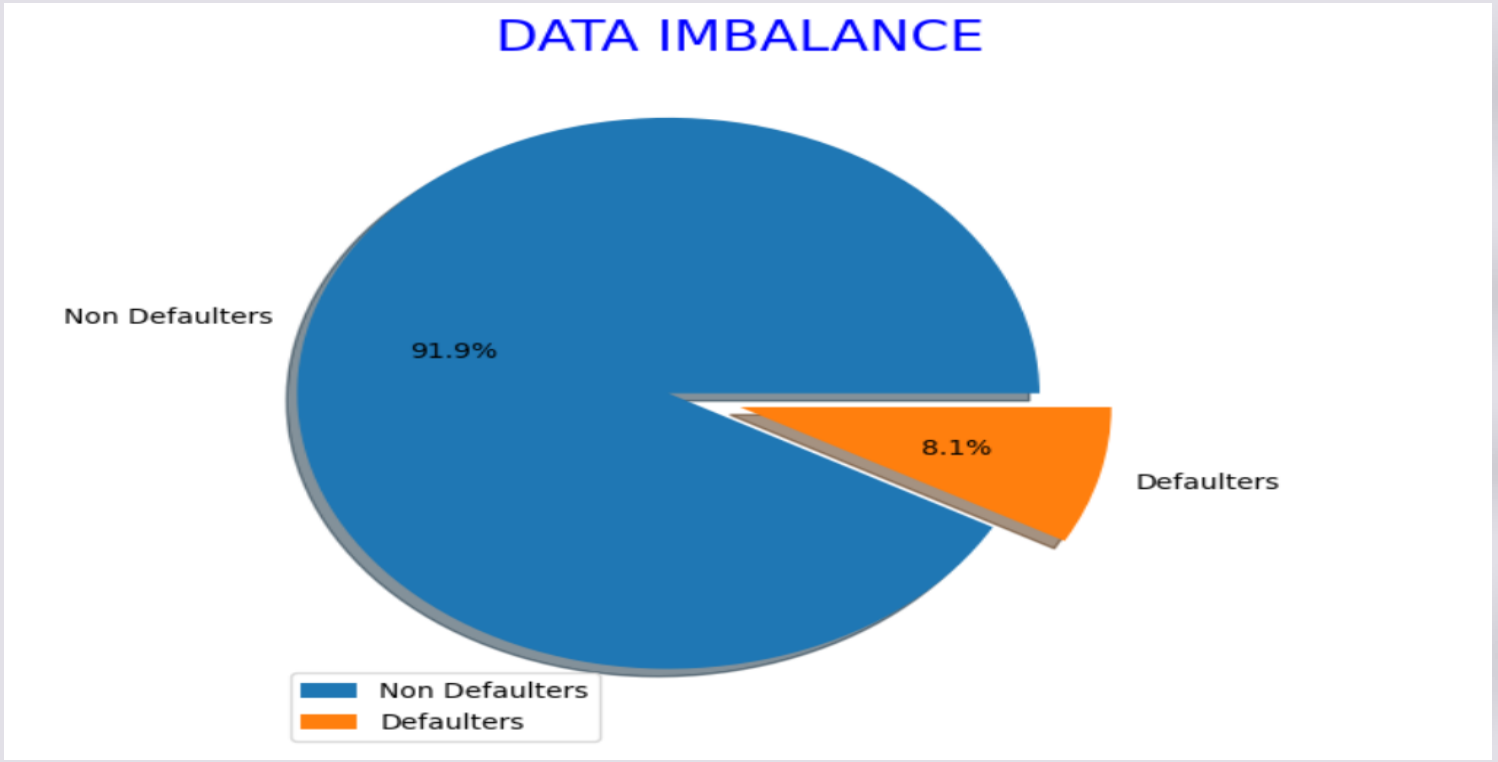


Analysis of `AMT_CREDIT`

- As observed from the boxplot, the outliers tend to exist after 1616625.0

- Applicants with `AMT_CREDIT` above 1616625.0 (calculated using IQR) are outliers

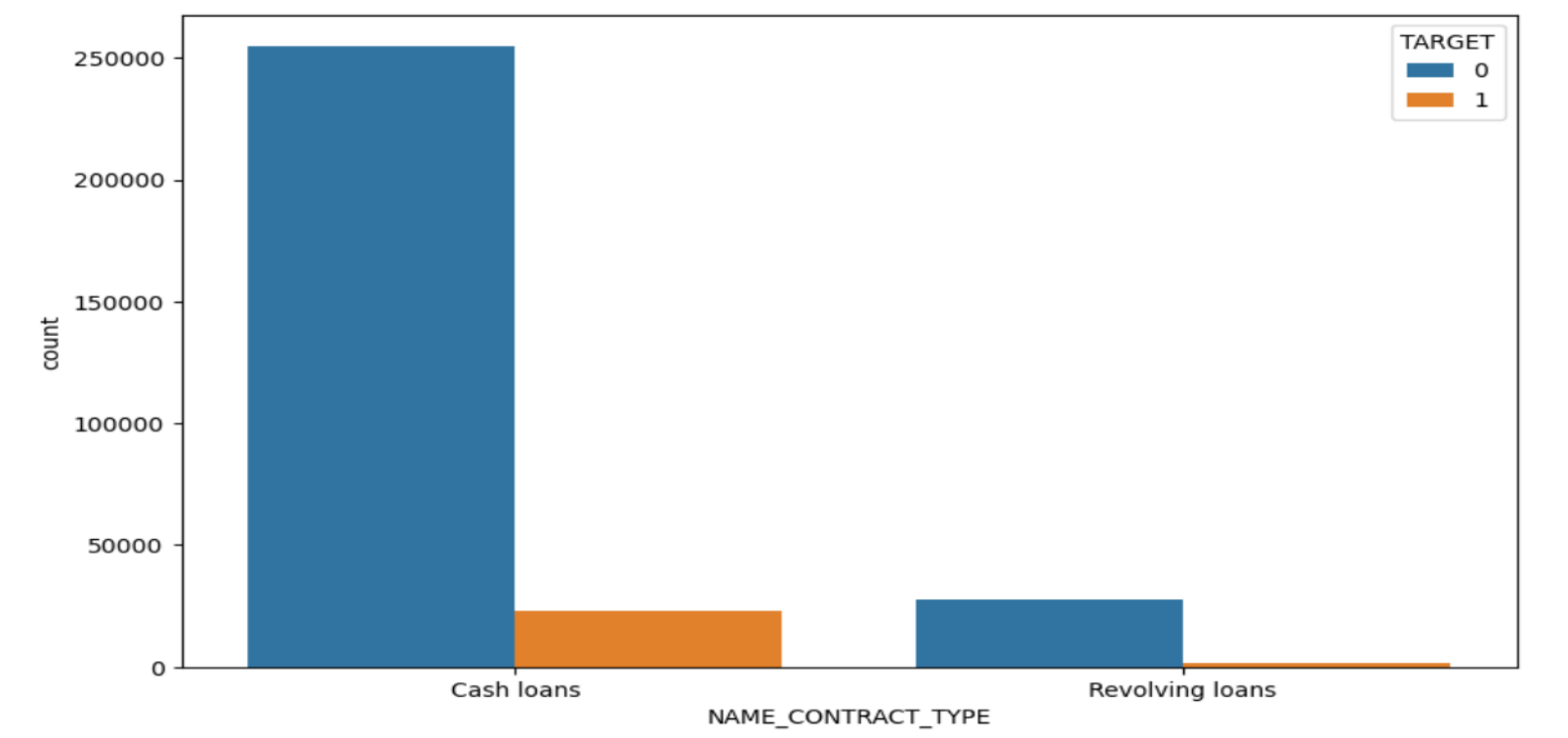
Checking imbalance for Target



Analysis of imbalance for `TARGET`

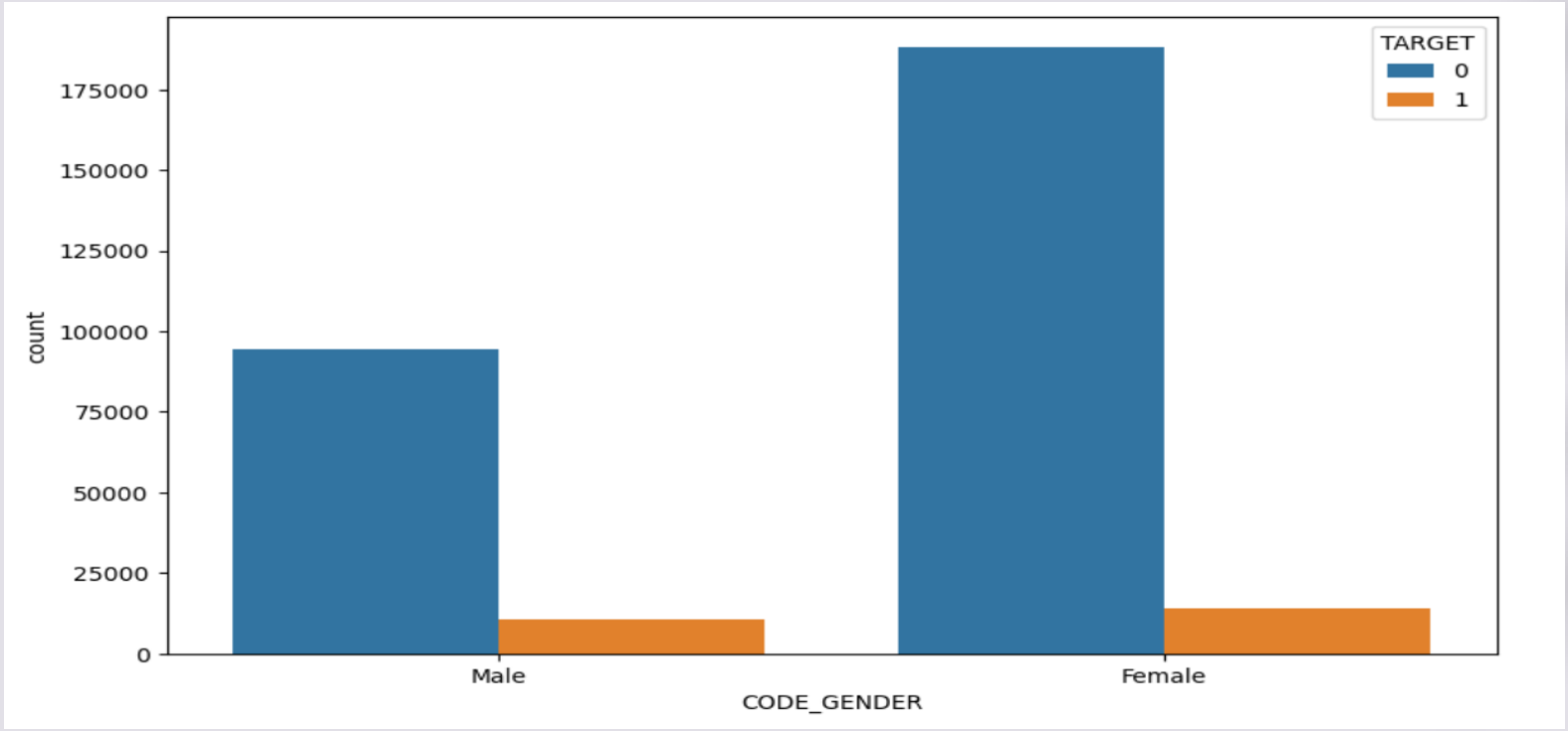
- We have imbalance in `TARGET` variable based on the % of observations
- `TARGET` value 1 represents client with payment difficulties (he/she had late payment more than X days on at least one of the first Y installments of the loan). This is only 8.1% of the data
- `TARGET` value 0 represents all other cases than 1. This is 91.9% of the data

Univariate analysis of categorical variables



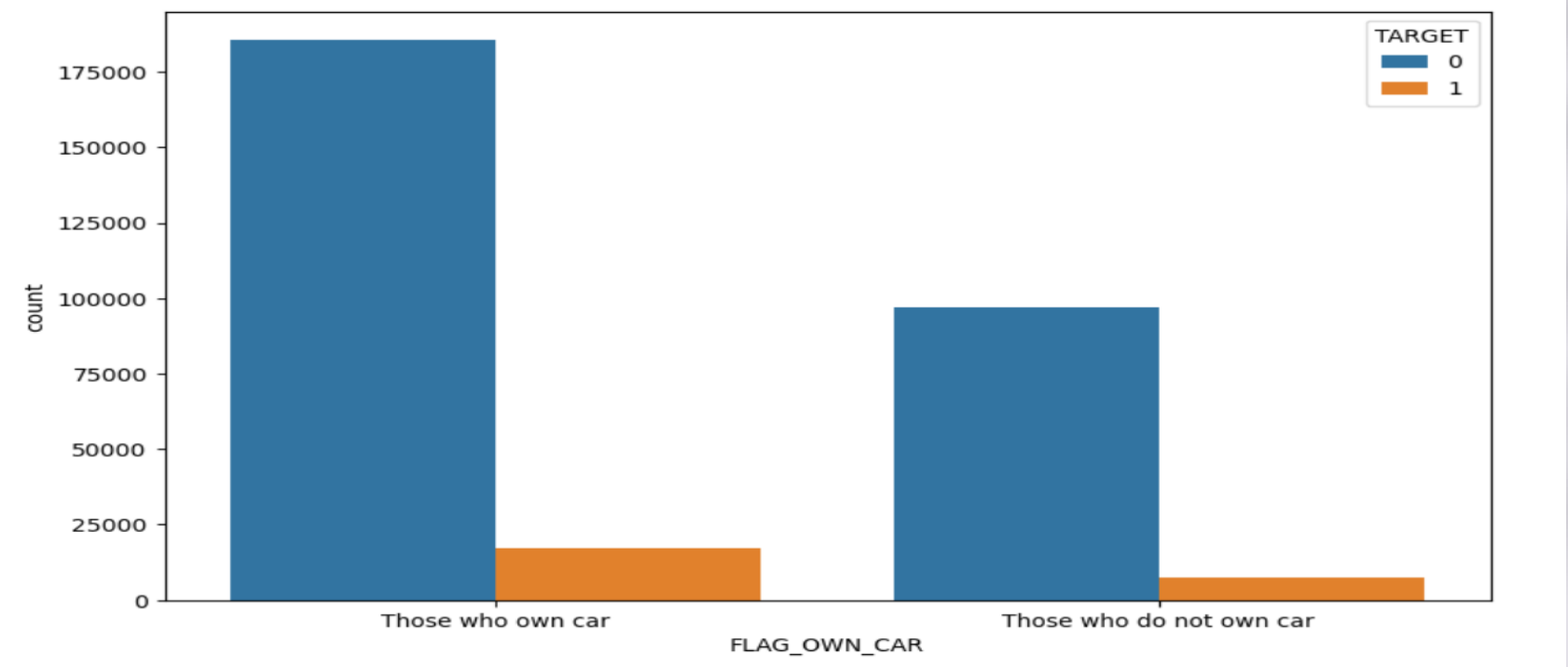
Analysis of `NAME_CONTRACT_TYPE`

- `NAME_CONTRACT_TYPE` column does not provide any conclusive evidence in favor of clients with payment difficulties OR on-time payments



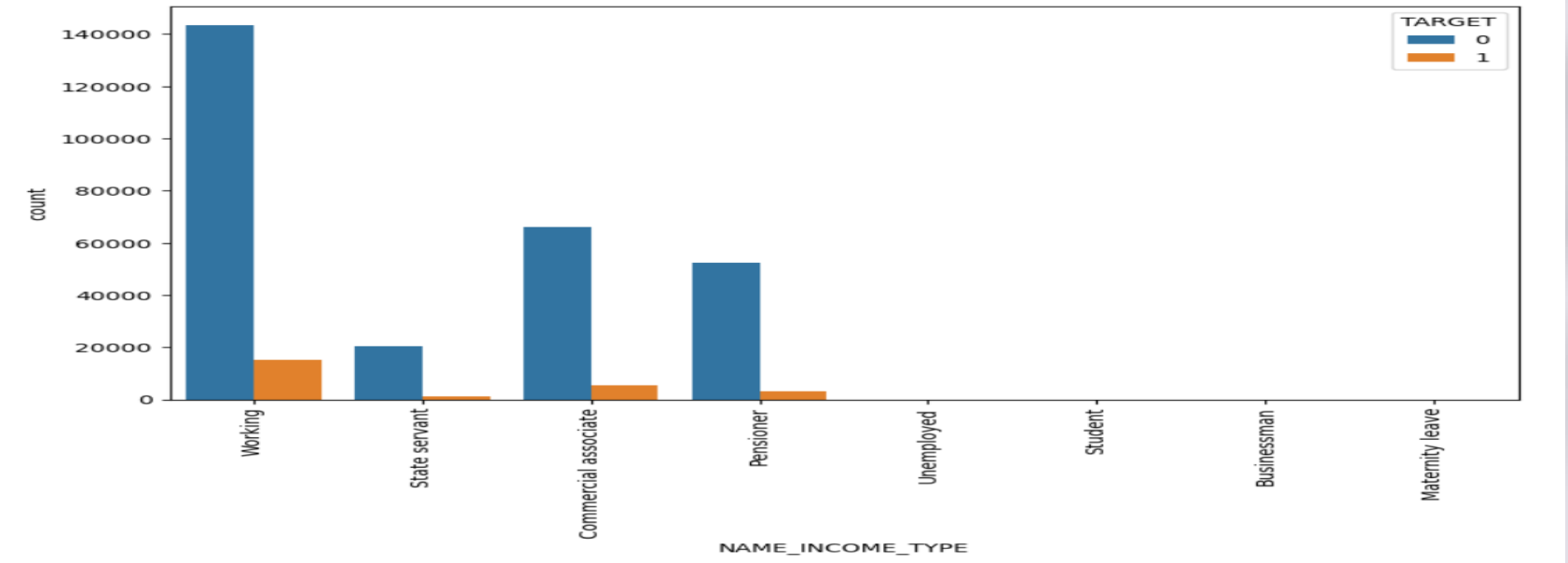
Analysis of `CODE_GENDER`

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



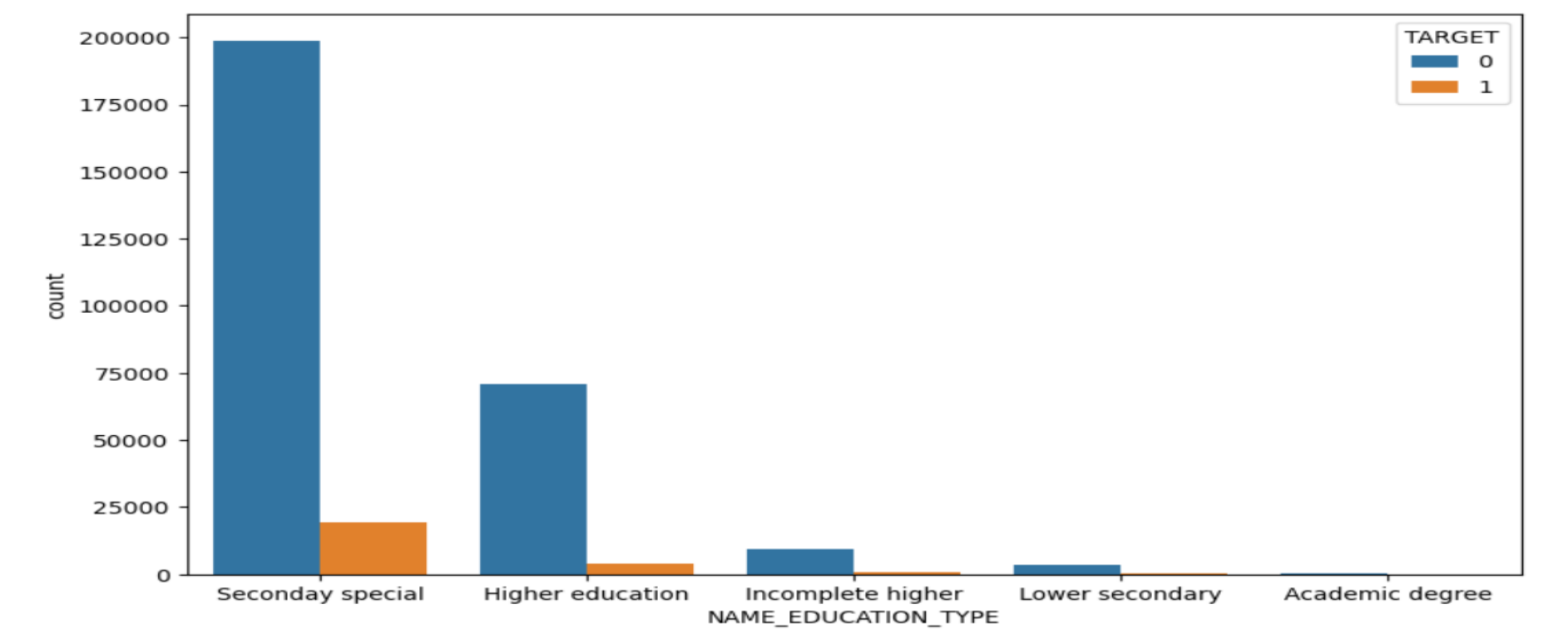
Analysis of `FLAG_OWN_CAR`

- `FLAG_OWN_CAR` column does not provide any conclusive evidence in favor of clients with payment difficulties OR on-time payments



Analysis of `NAME_INCOME_TYPE`

- Pensioners have better on-time payments. This is a weak correlation.
- Students don't have Payment difficulties. In this case, total students have only 18 observations and should be treated as a weak correlation
- Businessmen don't have Payment difficulties. In this case, Businessmen have only 10 observations and should be treated as a weak correlation



Analysis of `NAME_EDUCATION_TYPE`

- Clients with 'Higher education' have less payment difficulties. However, this is a weak correlation

- 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

With Payment difficulties

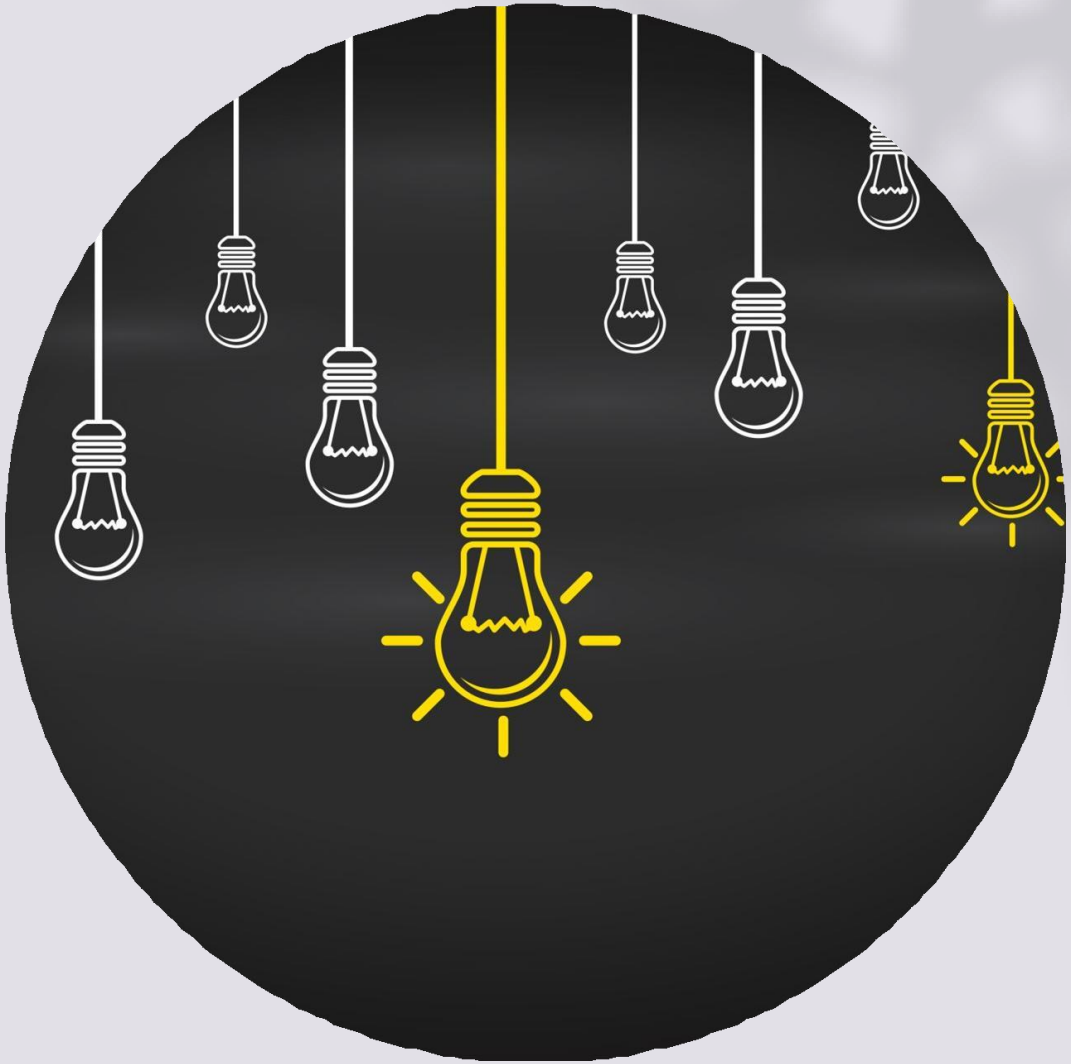
- Getting top 10 correlations

- AMT_GOODS_PRICE AMT_CREDIT 0.99
- REGION_RATING_CLIENT REGION_RATING_CLIENT_W_CITY 0.95
- CNT_FAM_MEMBERS CNT_CHILDREN 0.88
- REG_REGION_NOT_WORK_REGION LIVE_REGION_NOT_WORK_REGION 0.86
- DEF_30_CNT_SOCIAL_CIRCLE DEF_60_CNT_SOCIAL_CIRCLE 0.86
- LIVE_CITY_NOT_WORK_CITY REG_CITY_NOT_WORK_CITY 0.83
- AMT_ANNUITY AMT_GOODS_PRICE 0.78
- AMT_ANNUITY AMT_CREDIT 0.77
- DAYS_BIRTH DAYS_EMPLOYED 0.63
- DAYS_EMPLOYED FLAG_DOCUMENT_6 0.60

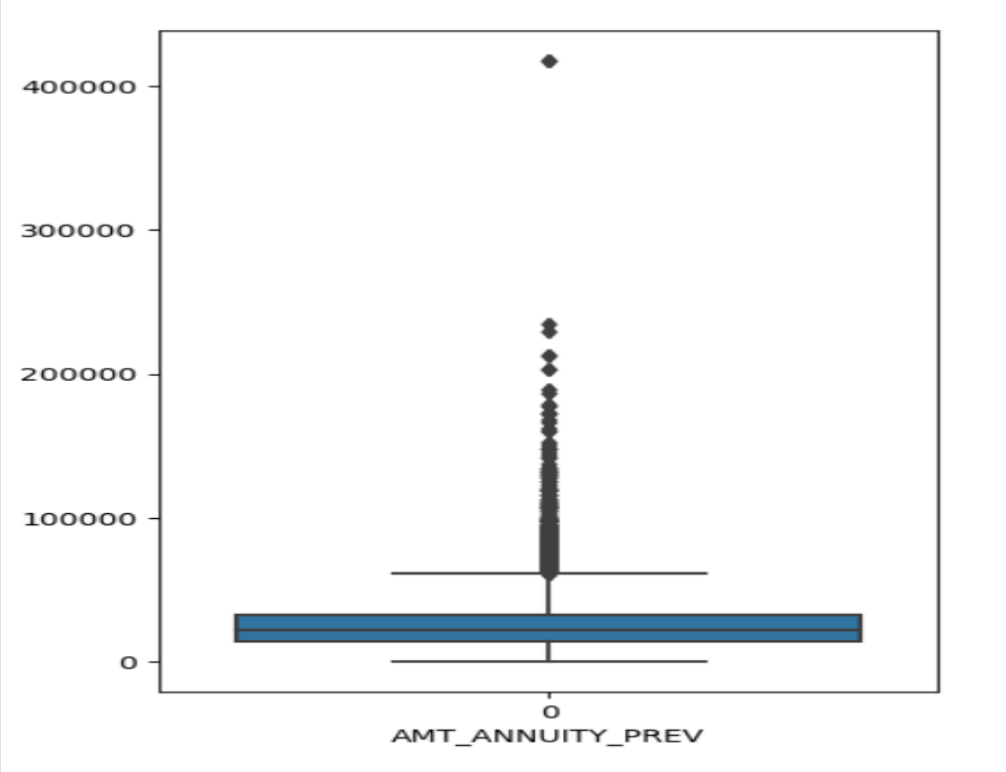
On-Time payments

- Getting top 10 correlations

Analysis of information about the client’s previous loan data

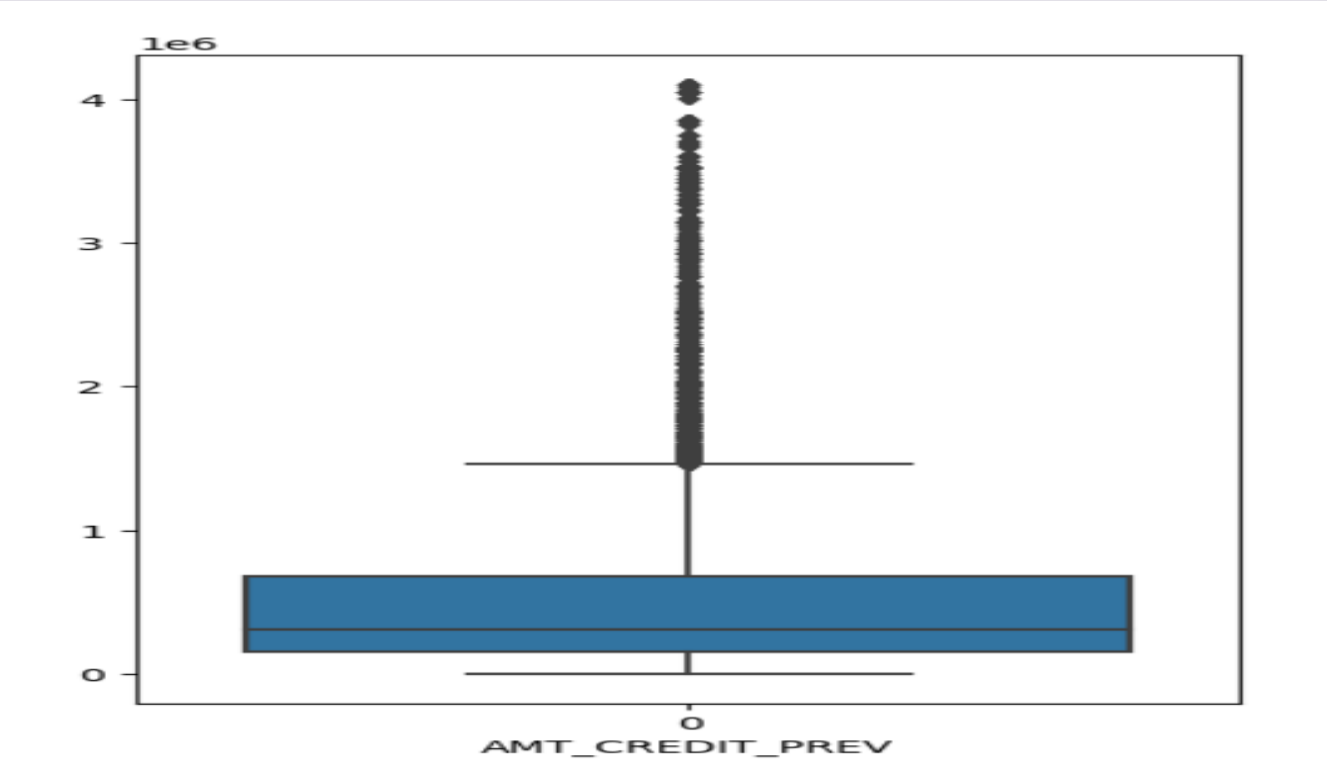


Outlier analysis



Analysis of `AMT_ANNUITY` column

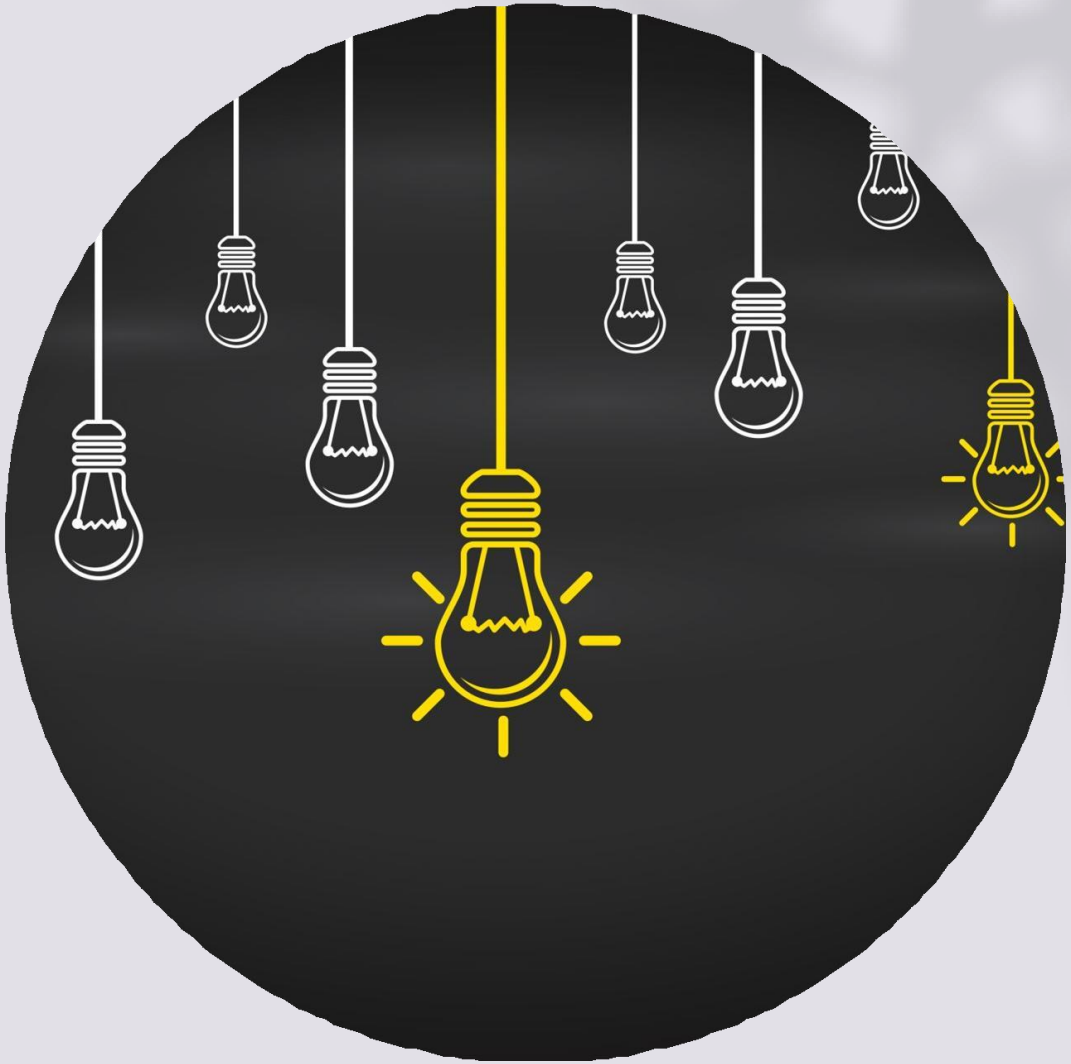
- `AMT_ANNUITY` values above 42163.38 are outliers



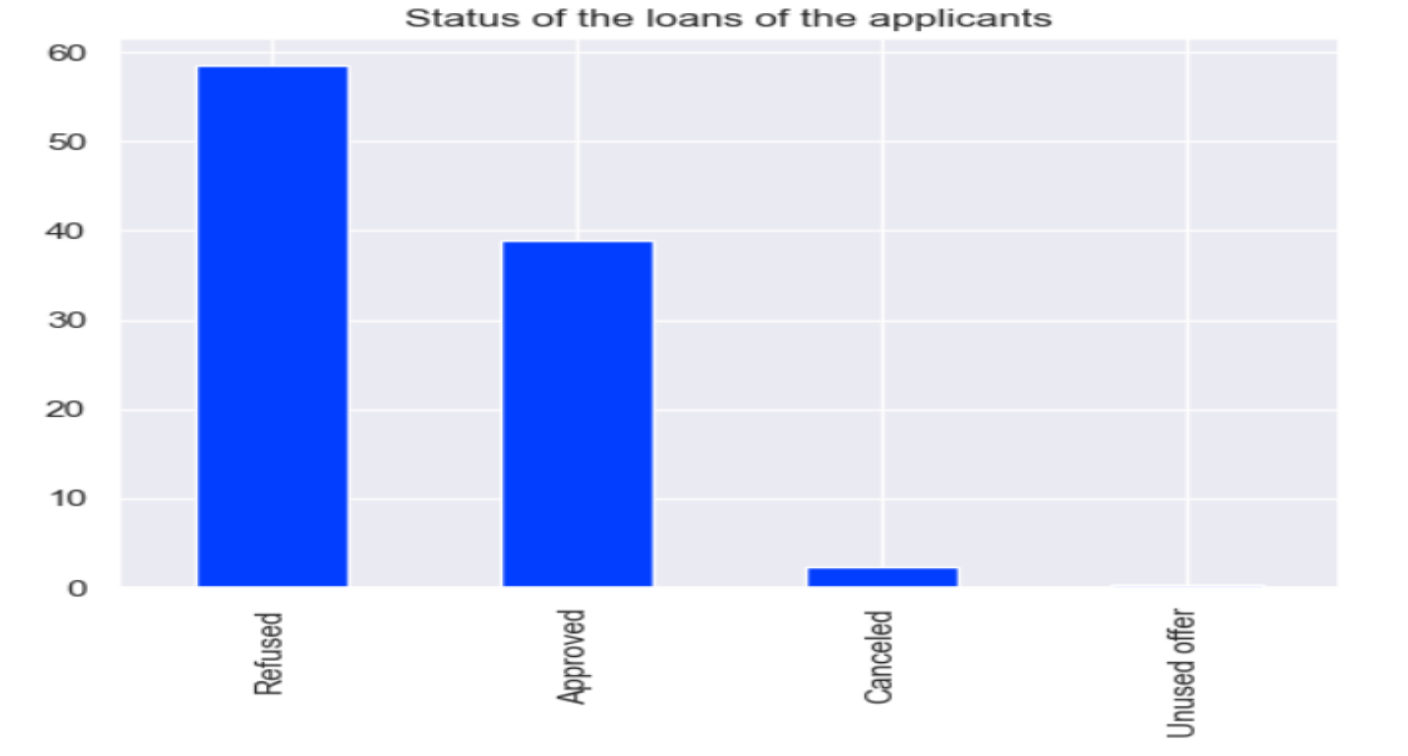
Analysis of `AMT_CREDIT` column

- `AMT_CREDIT` values above 504805.5 are outliers

Analysis of merged information about the client’s previous loan data and current loan application

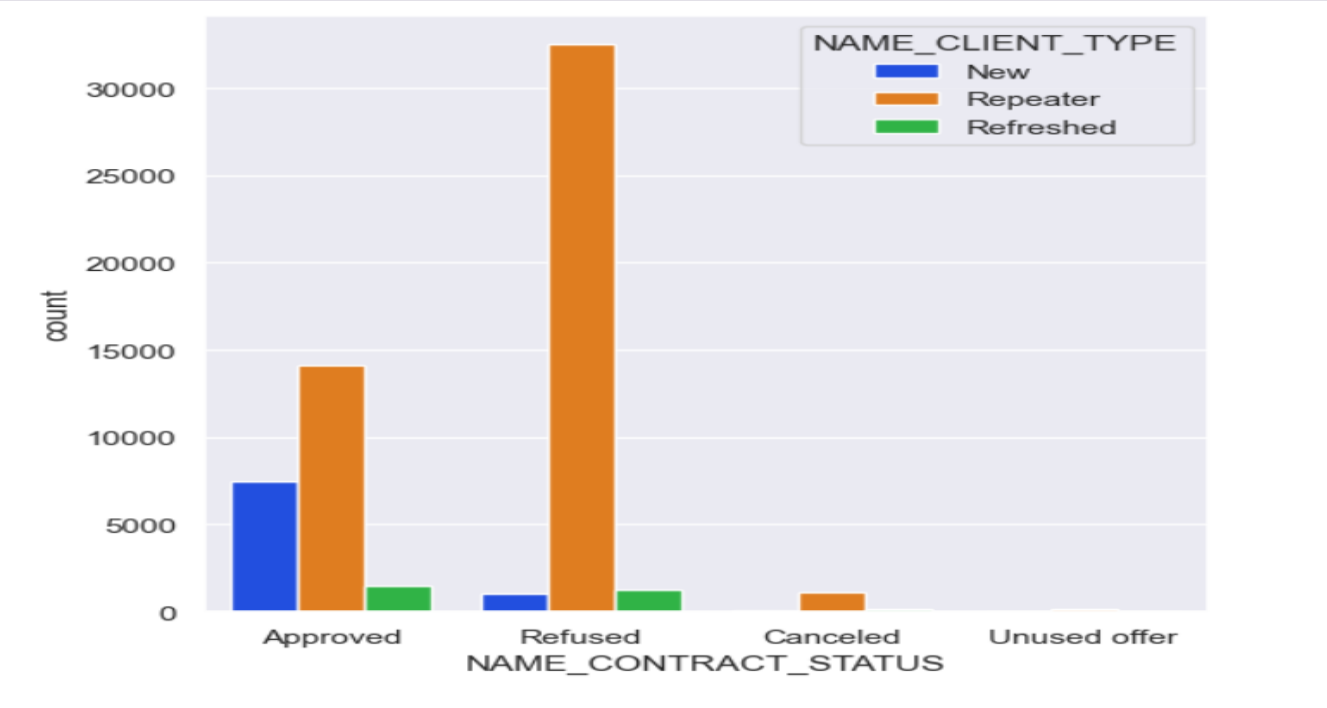


Univariate analysis of categorical variables



Analysis of `NAME_CONTRACT_STATUS`

- `Refused` loan status is the highest among all loan applications
- `Approved` loan status is the second highest among all loan applications



Analysis of `NAME_CONTRACT_STATUS`

- `Repeater` client type is the highest among all loan applications
- `New` client type is the second highest among all loan applications

Conclusion

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