Credit EDA Case Study

By: Sakshi Gupta

Introduction

This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, 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

- 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. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze 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.

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 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. Identification of such applicant's 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.

Data Understanding

This dataset has 3 files as explained below:

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.**

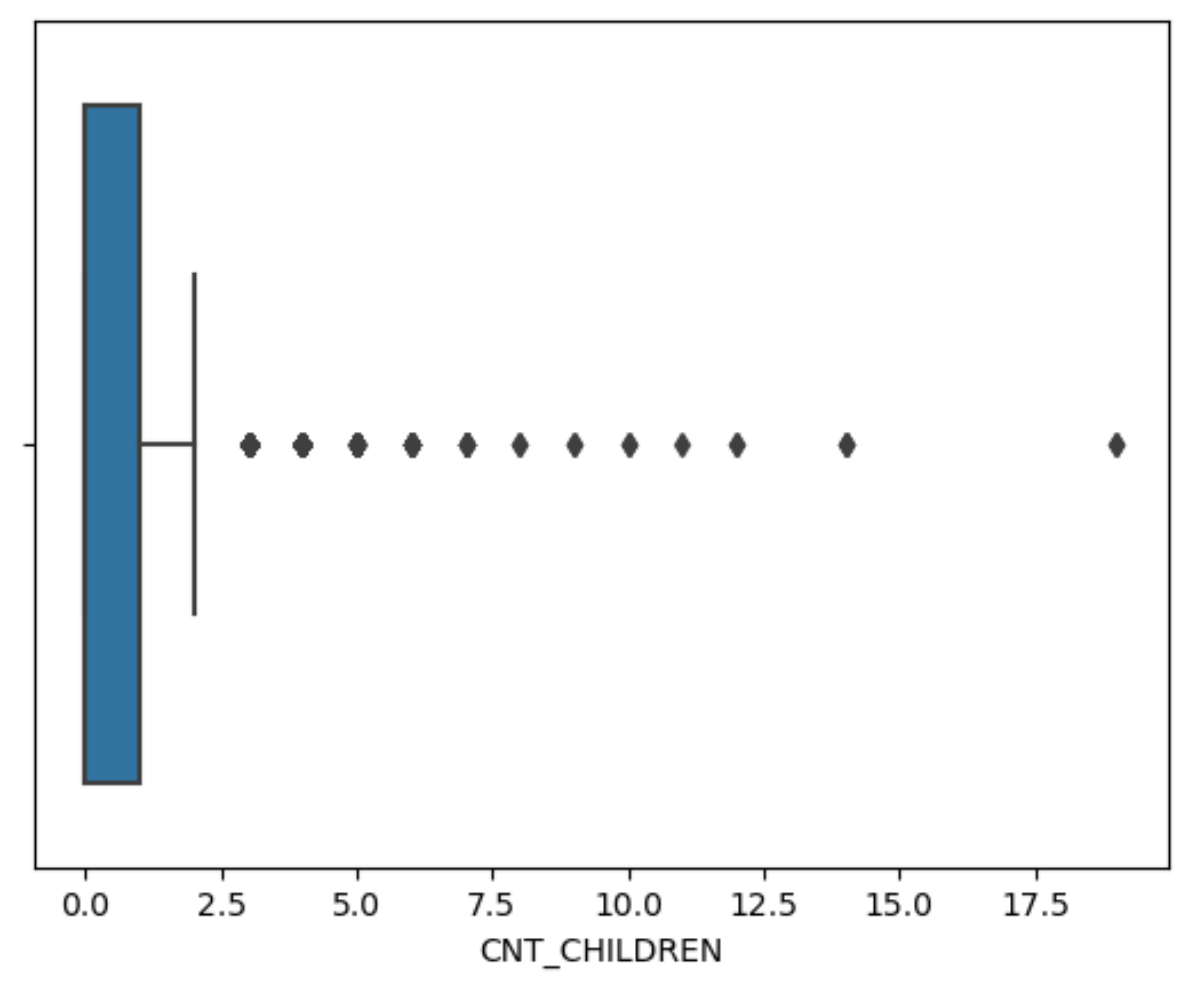
2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data on whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

3. 'columns_description.csv' is data dictionary which describes the meaning of the variables.

Analysis of information of the clients at the time of application

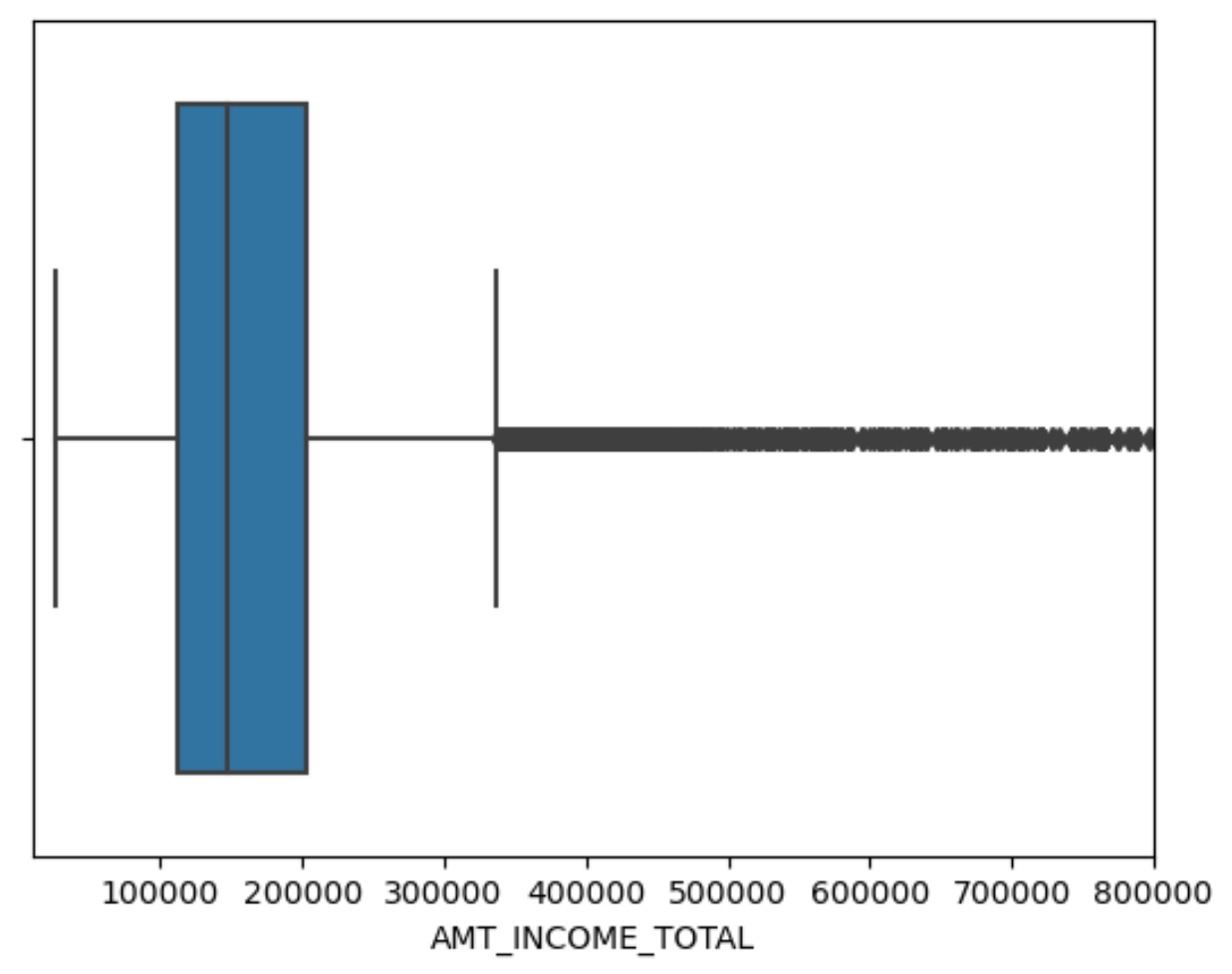
Outlier Analysis

Analysis of ‘CNT_CHILDREN’



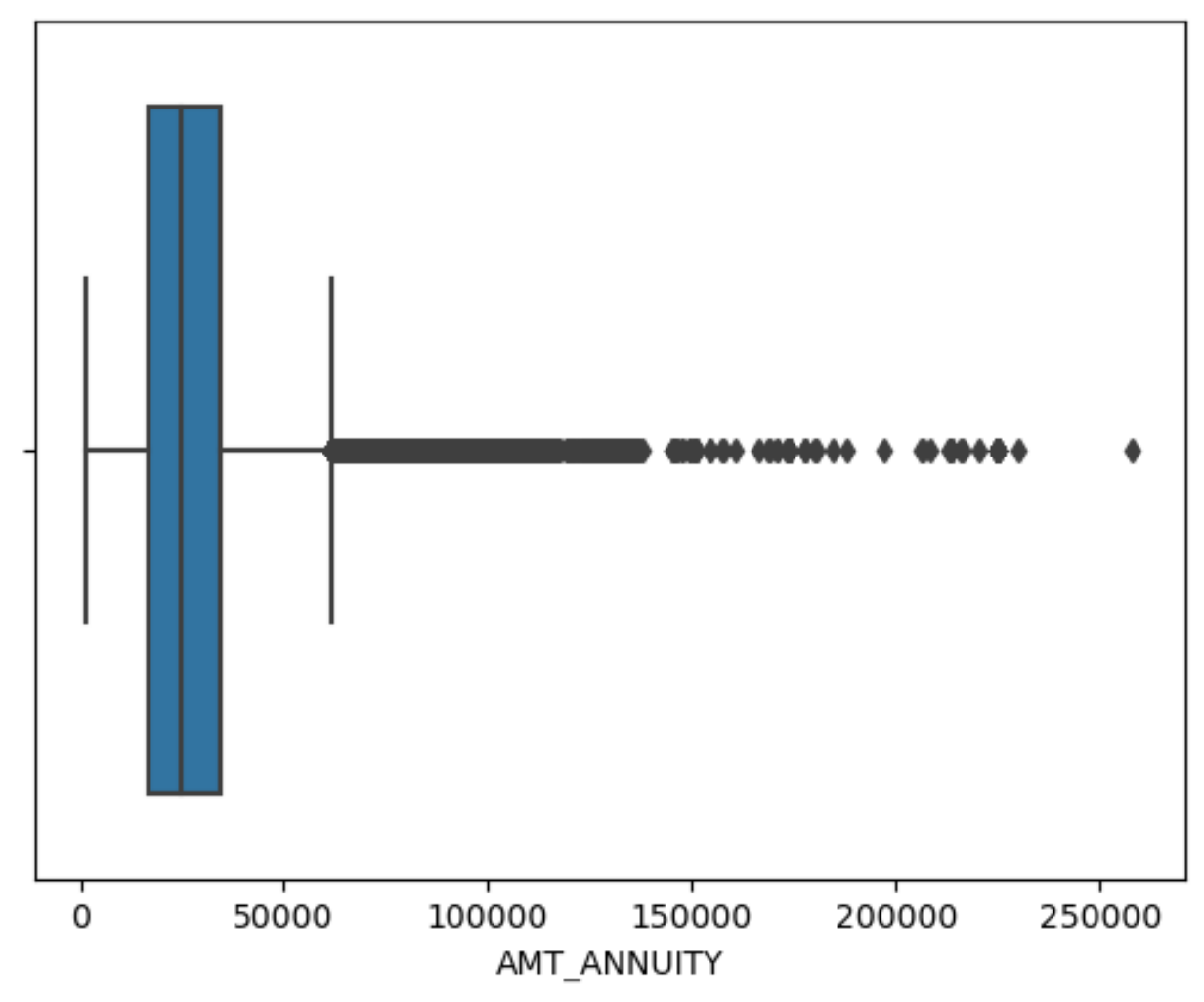
- Looking at the data, we can see that above 7 children, applicants are very minimal (2 0r 3 in each category)
- Boxplot clearly shows the values above 2.5 as being outliers.
- Applicants with 3 or more cases are outliers cases.

Analysis of ‘AMT_INCOME_TOTAL’



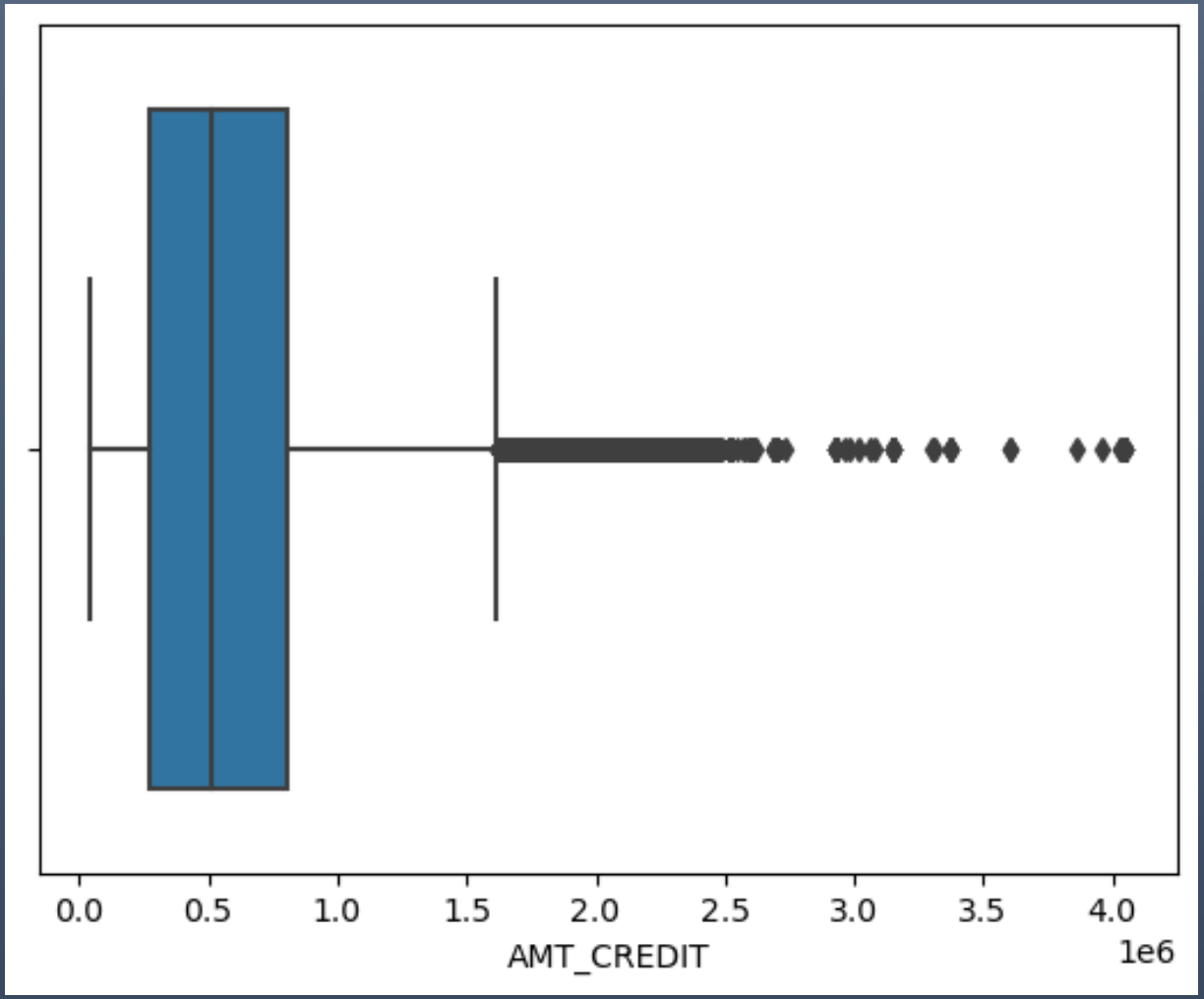
Applicants with income above 3,50,000 are outliers.

Analysis of ‘AMT_ANNUITY’



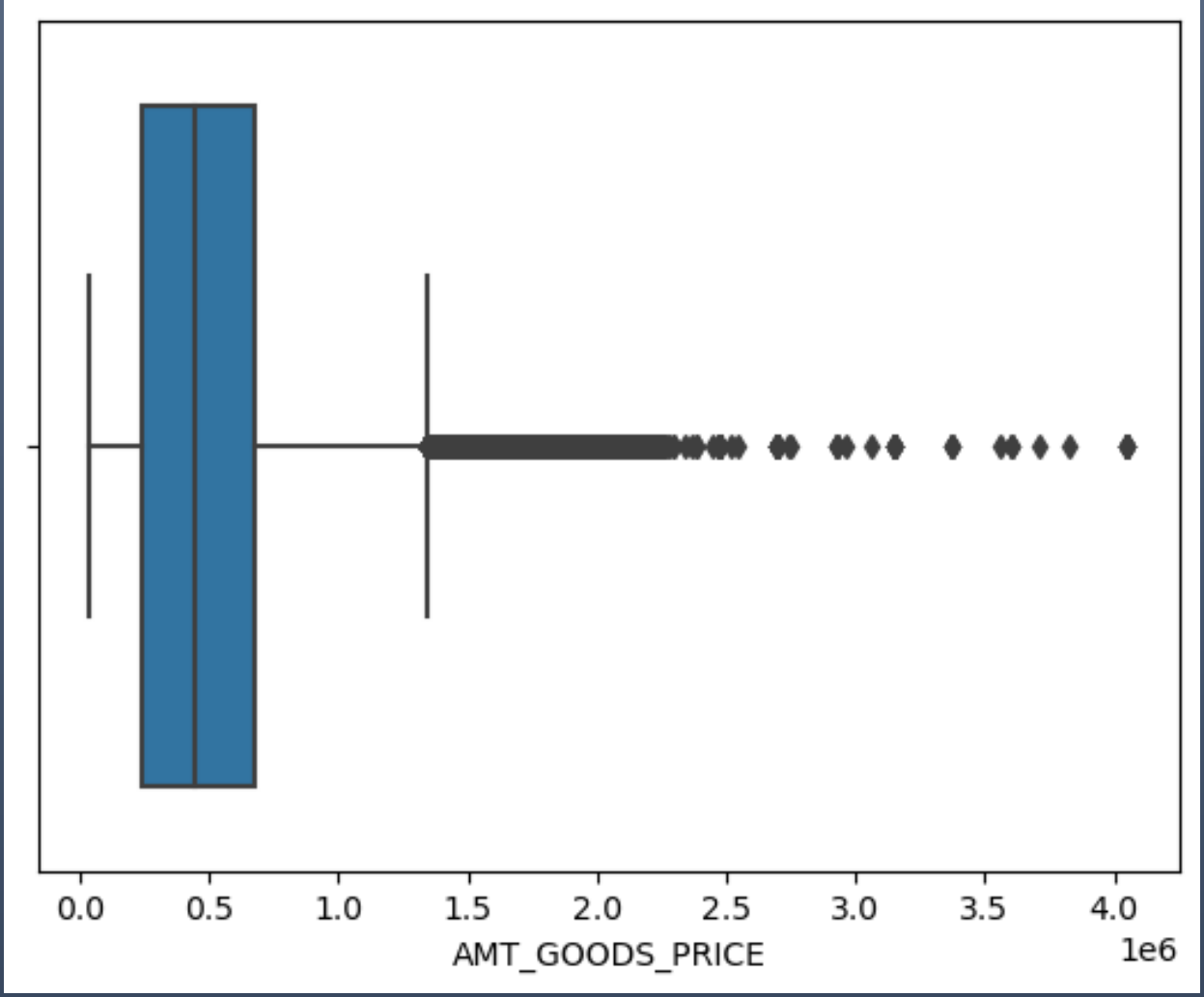
Applicants with AMT_ANNUITY above 60,000 are outliers.

Analysis of ‘AMT_CREDIT’



Applicants with ‘AMT_CREDIT’ above 1616625.0(calculated using IQR) are outliers.

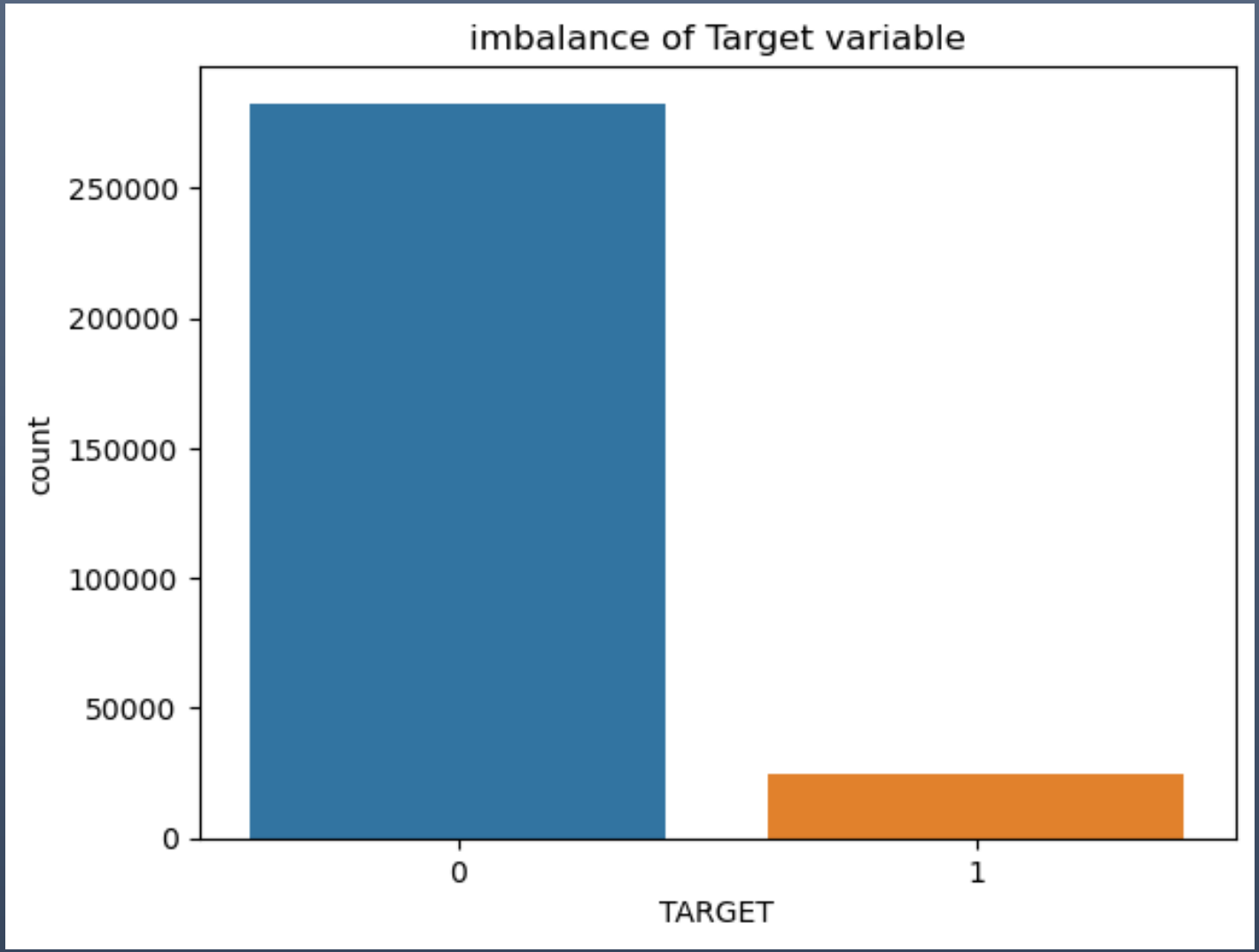
Analysis of ‘AMT_GOODS_PRICE’



Applicants with ‘AMT_GOOD_PRICE’ above 1341000.0(calculated using IQR) are outliers.

Checking imbalance of ‘Target’

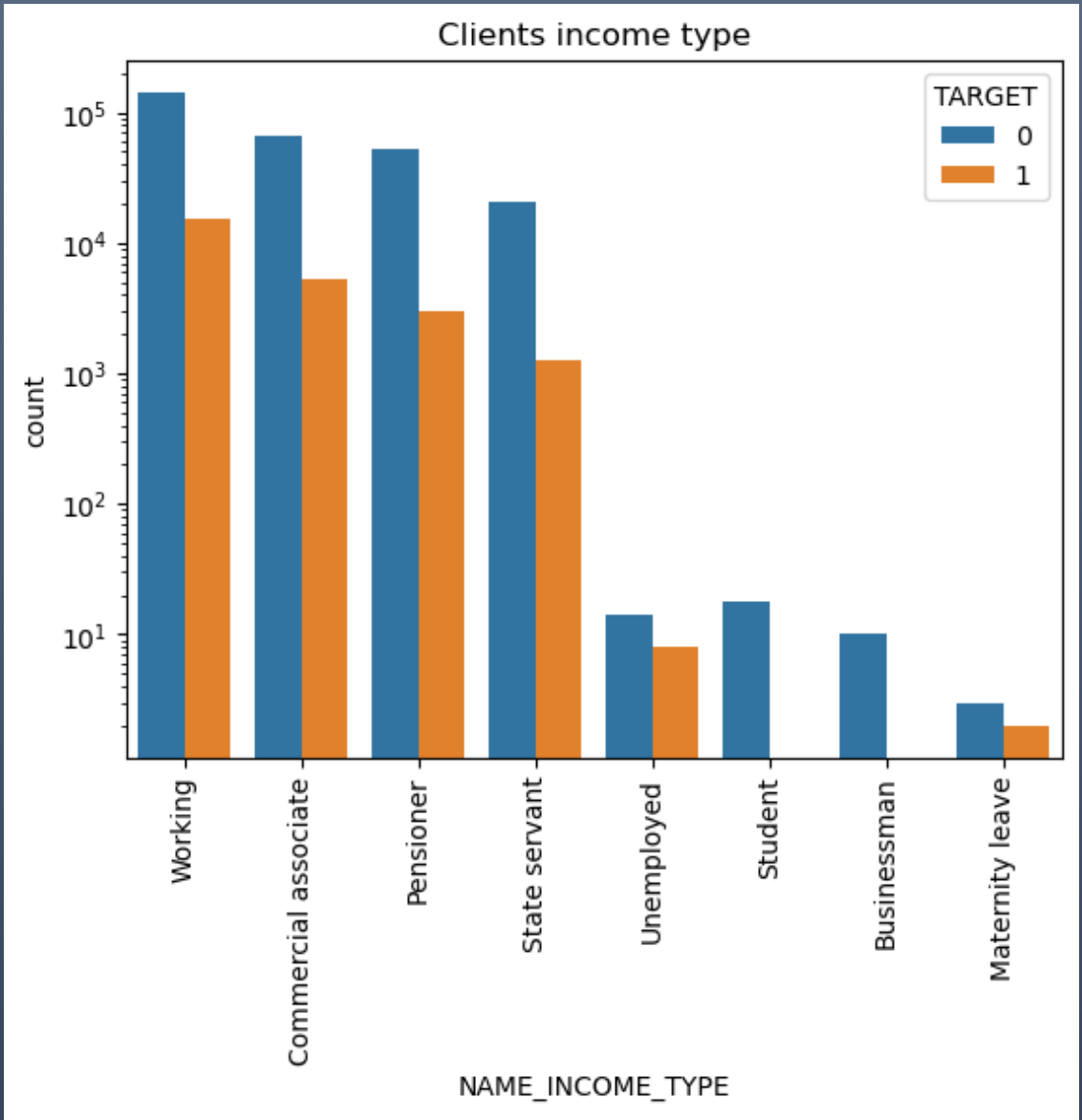
Checking imbalance of Target variable



- We have imbalance in ‘TARGET’ variable based on the % of observations.
- ‘TARGET’ value-1- represents defaulters(clients with payment difficulties). This is only 8.07% of the data.
- ‘TARGET’ value-0- represents non-defaulters(all-other cases than 1). This is 91.93% of the data.

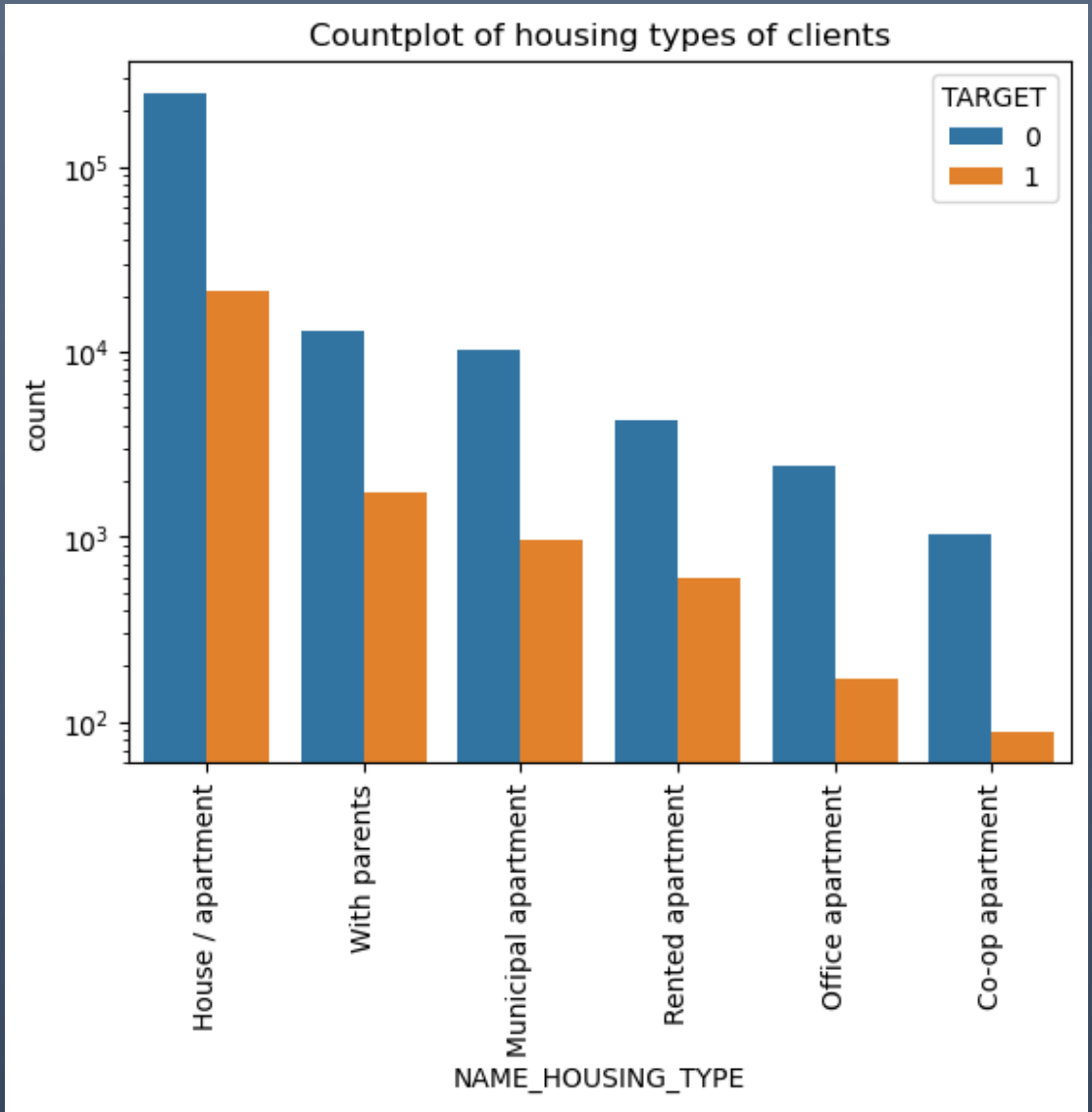
Univariate analysis of categorical variables

Analysis of ‘NAME_INCOME_TYPE’



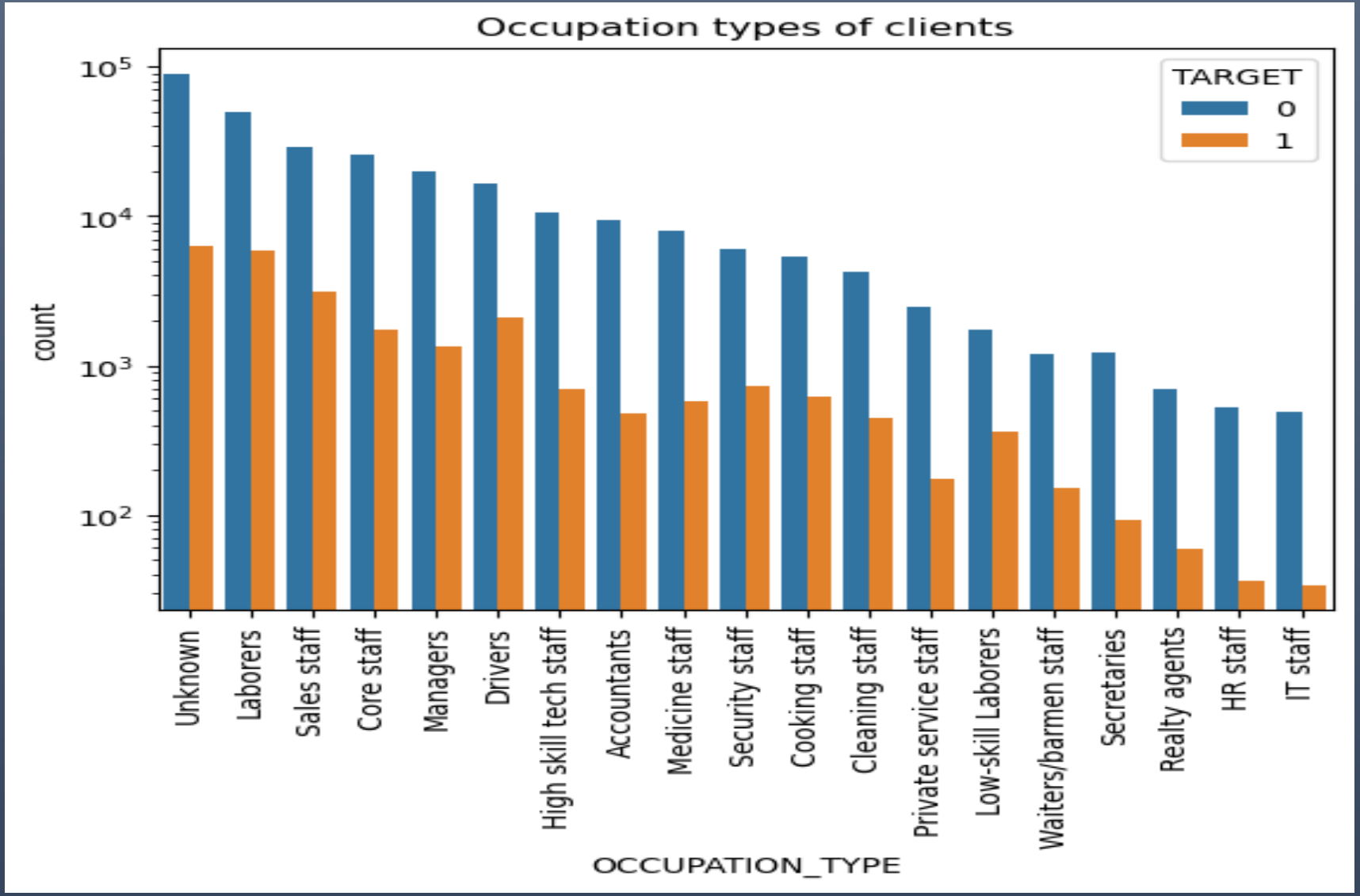
- Working, Commercial associate, Pensioner and State servant are safest to give loans.
- Unemployed and Maternity leave faces difficulties.

Analysis of ‘NAME_HOUSING_TYPE’



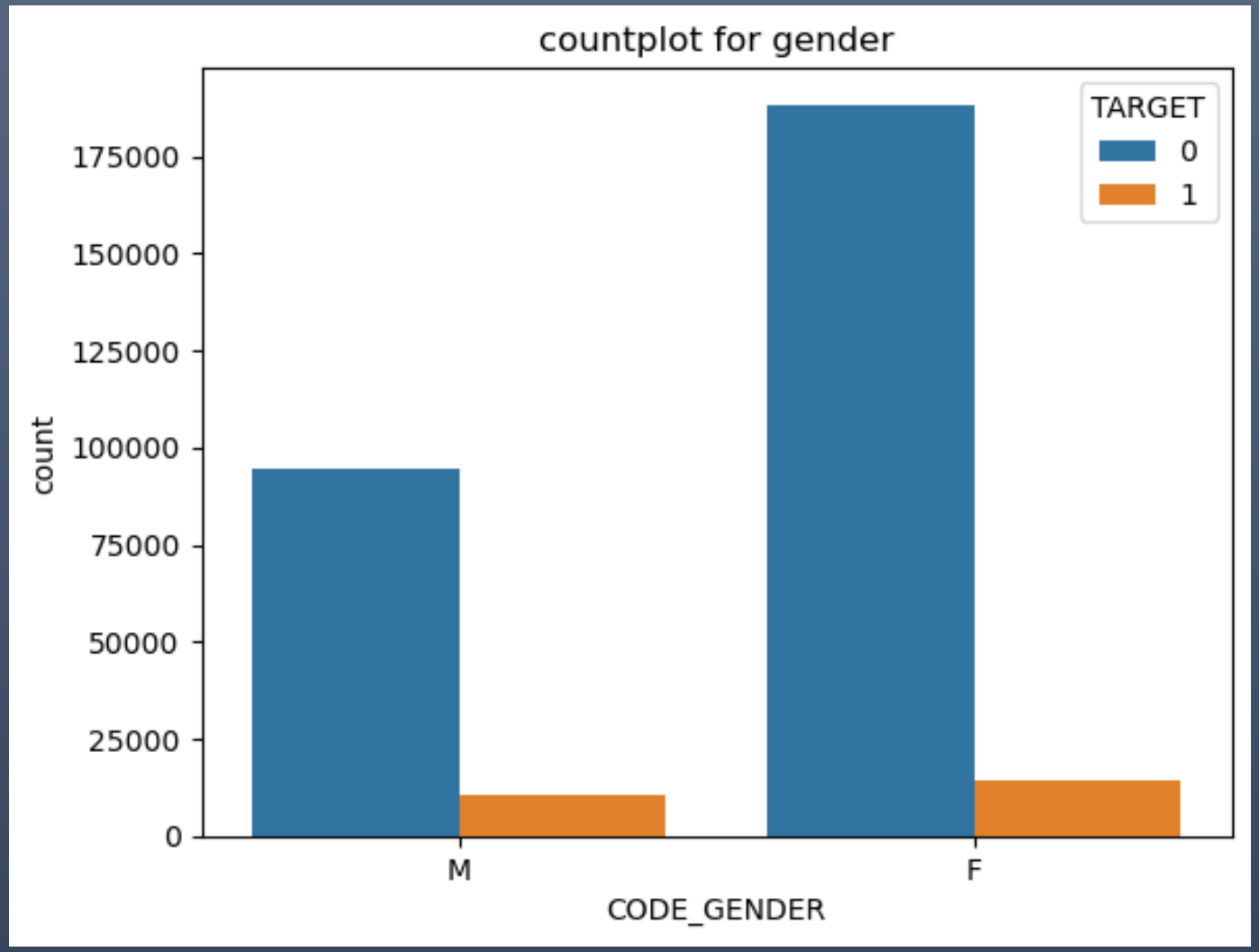
- Clients who has their own house/apartment are safest to target.

Analysis of ‘OCCUPATION_TYPE’



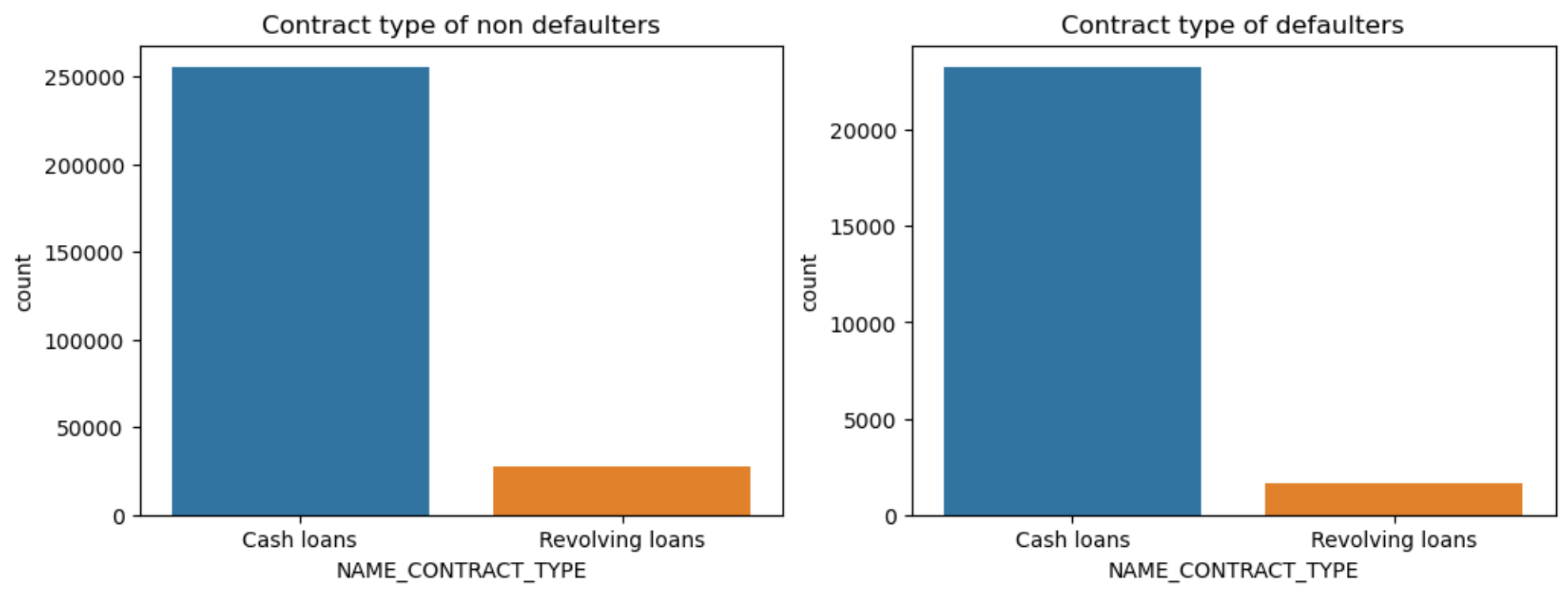
- Laborers, Accountants are safest to target.
- Drivers and low skill labourers are defaulters.

Analysis of ‘CODE_GENDER’



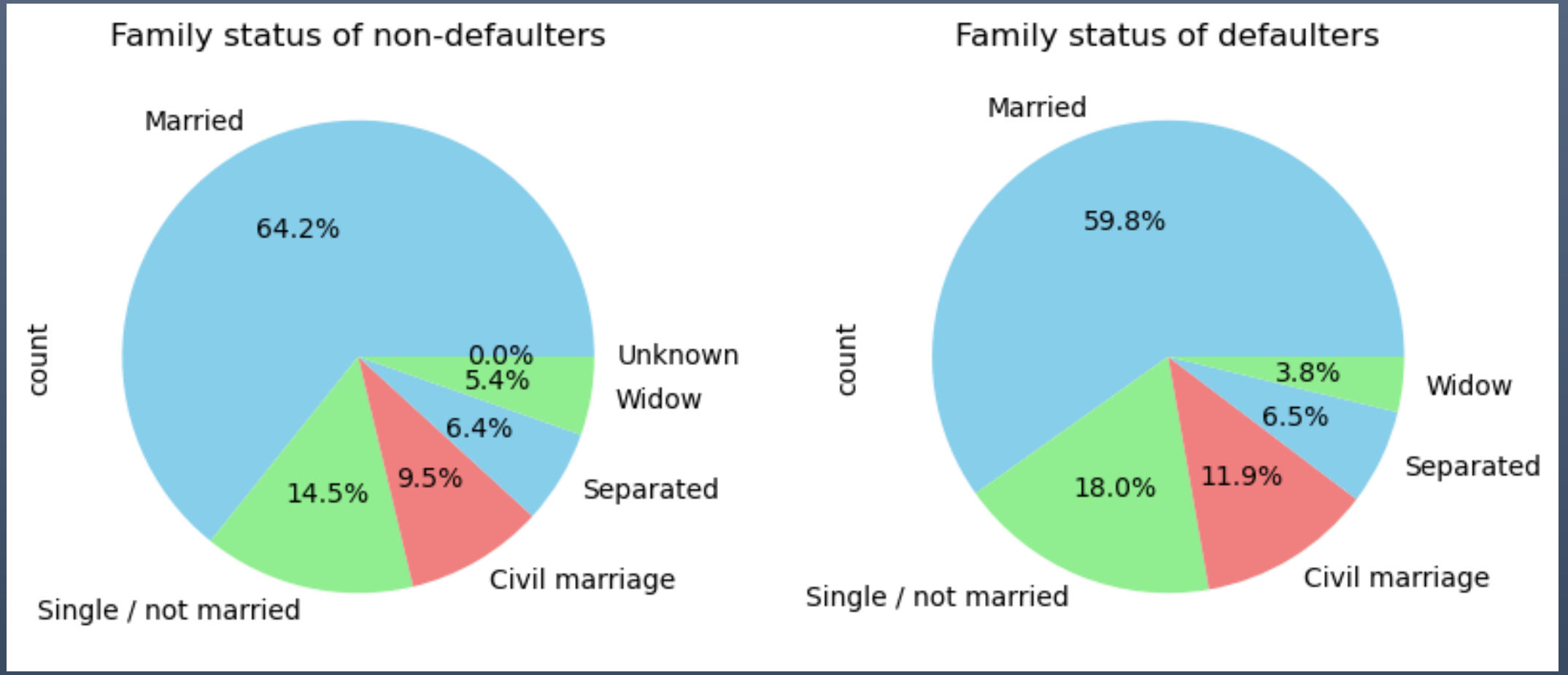
- Default rate of female clients are lower than male.

Analysis of ‘NAME_CONTRACT_TYPE’



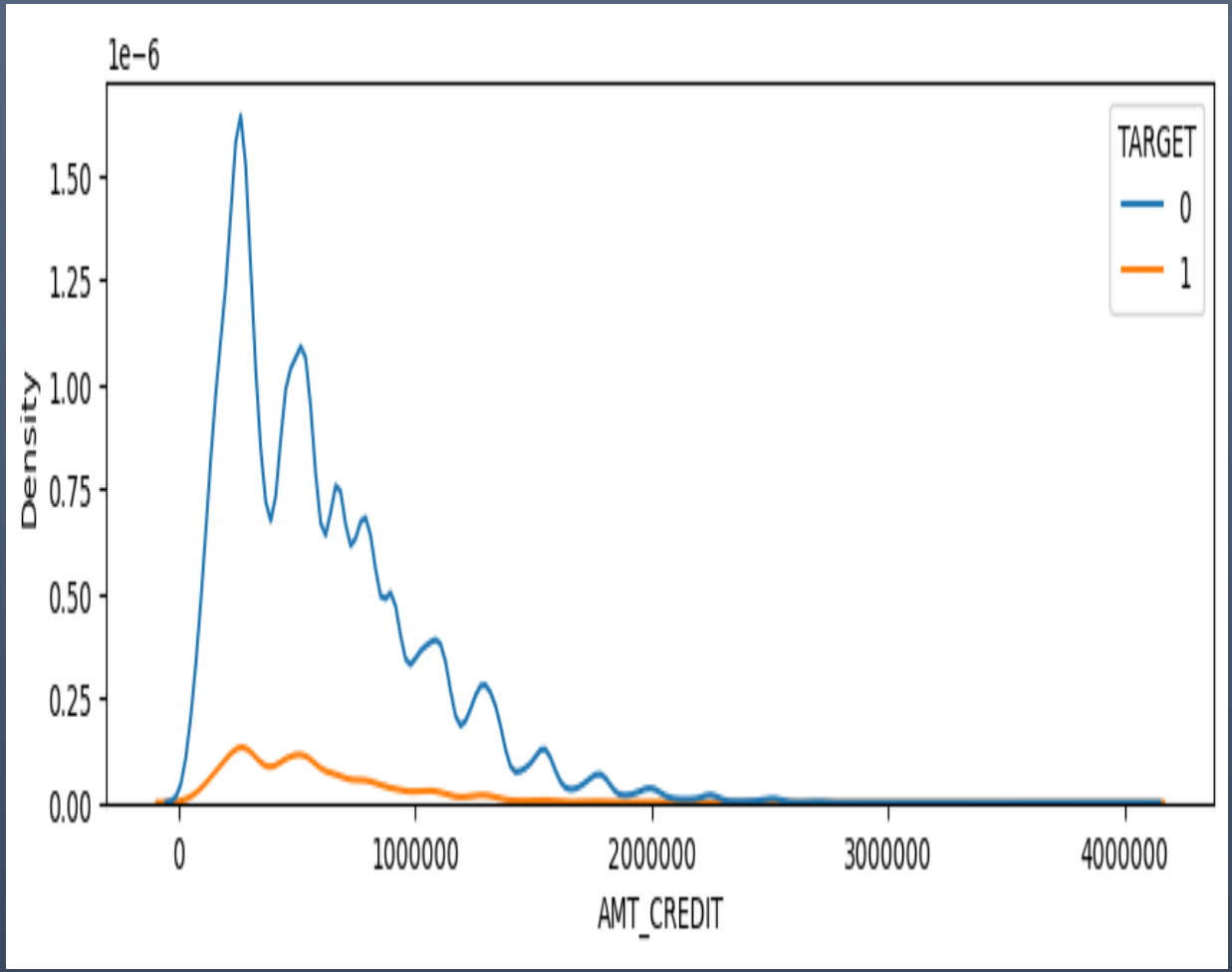
- Most of the customers has Cash loan.
- Customers who have taken cash loan are most likely to default.

Analysis of ‘NAME_FAMILY_STATUS’



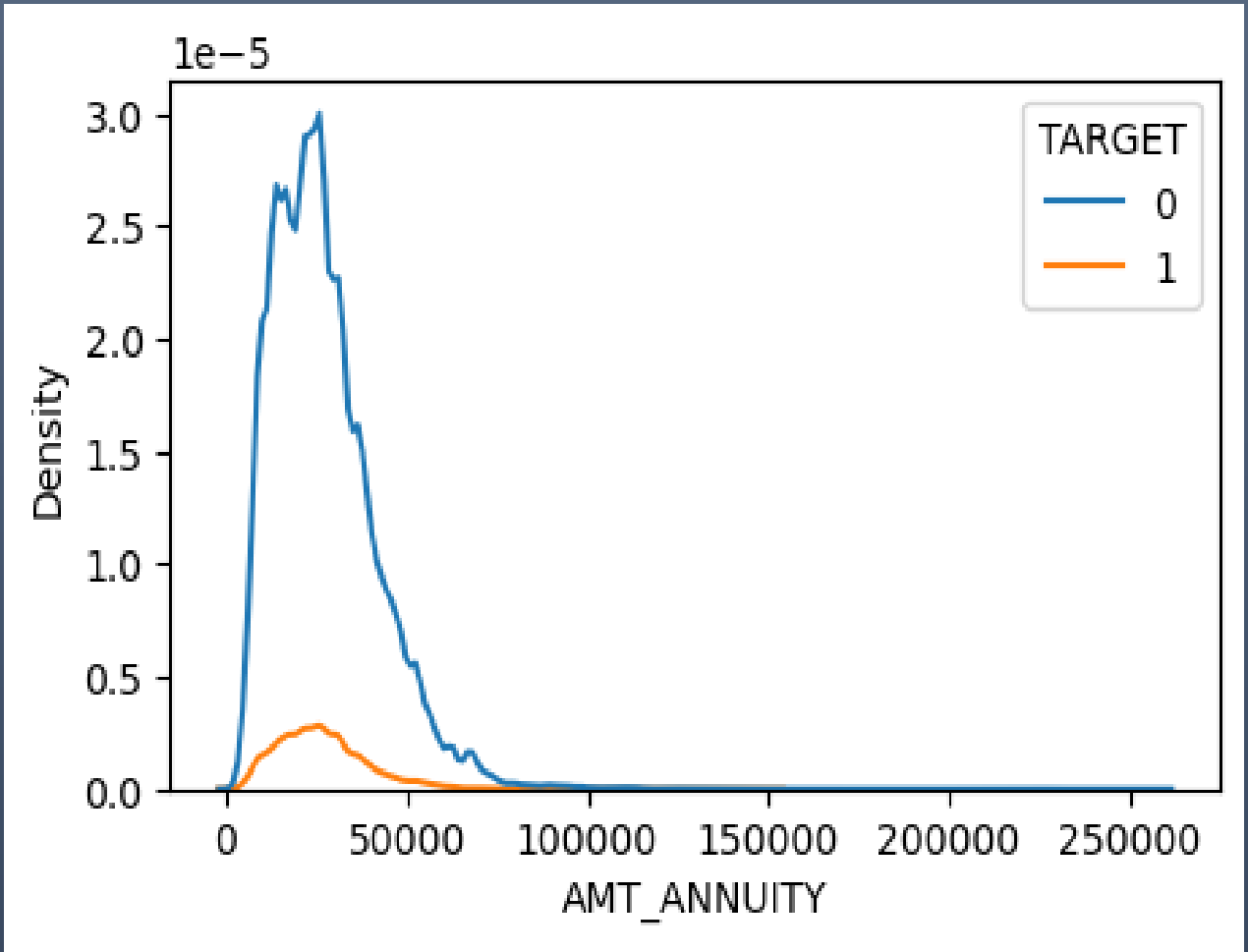
Univariate analysis of numerical variables

Analysis of AMT_CREDIT’



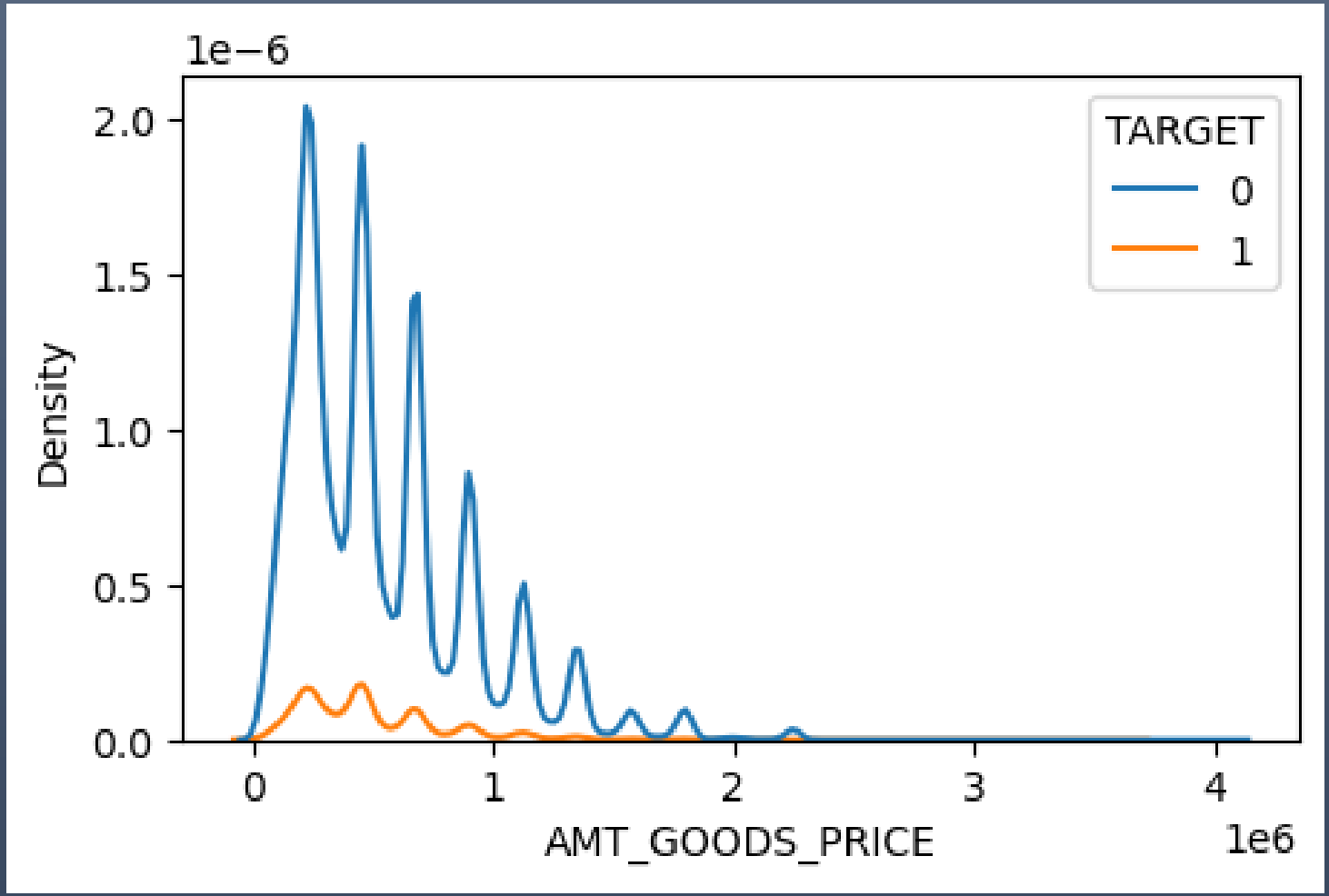
- Most of the loan were given for the amount of 0 to 1million.

Analysis of AMT_ANNUITY'



- Most of the clients are paying annuity between 0 to 1Lakh

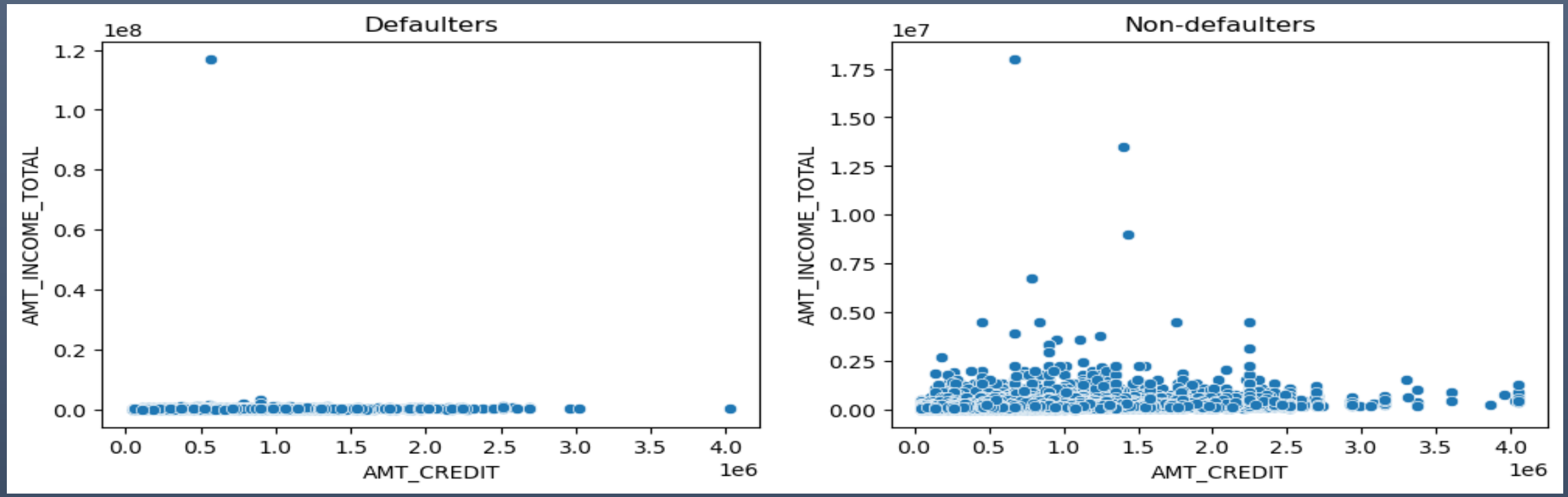
Analysis of AMT_GOODS_PRICE’



- Most of the clients have taken loan for goods price ranging 0 to 2 million.

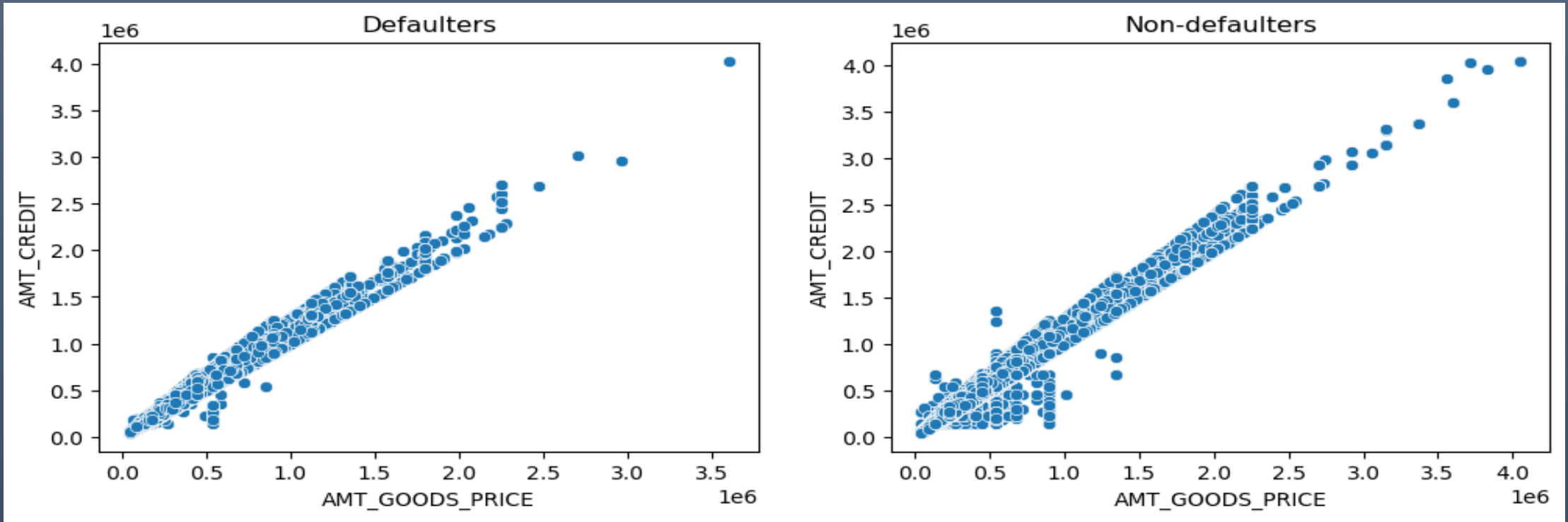
Bivariate analysis Continuous v/s Continuous variable

Analysis of ‘AMT_INCOME_TOTAL’ V/S ‘AMT_CREDIT’



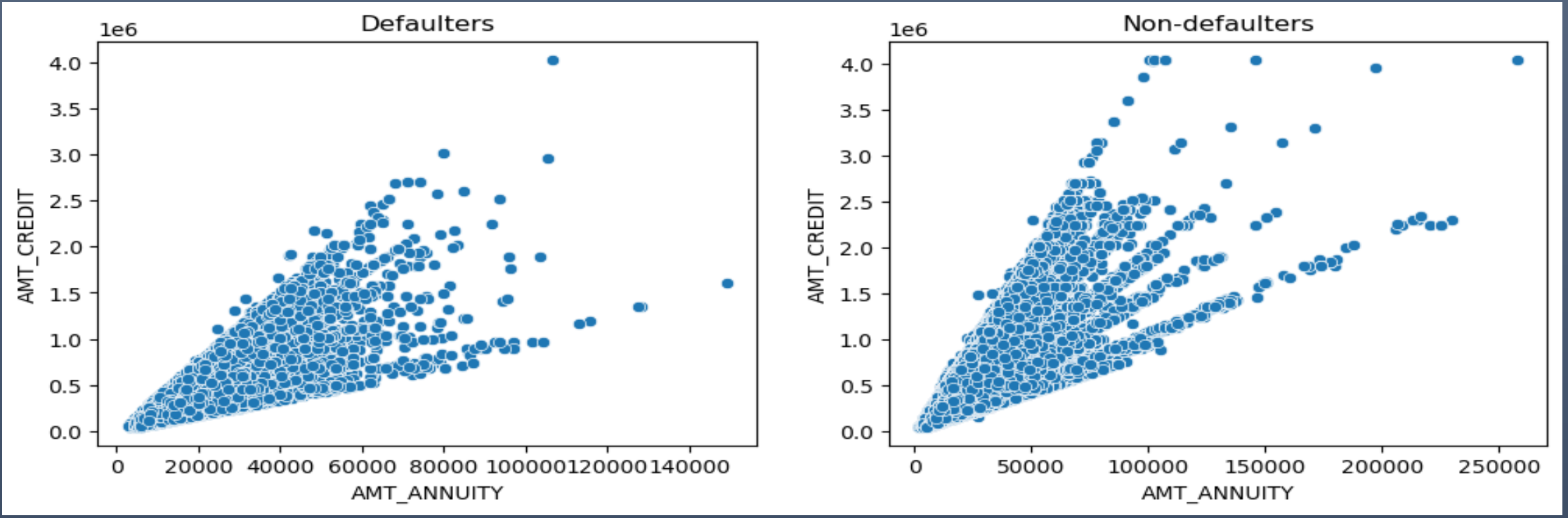
- Almost all client’s income is below 1 million and the chances of default are higher when the credit amount is less than 1.5 million.

Analysis of ‘AMT_GOODS_PRICE’ V/S ‘AMT_CREDIT’



- ‘AMT_GOODS_PRICE’ and ‘AMT_CREDIT’ have strong positive correlation. This means that as Goods price increases, so does Credit amount.

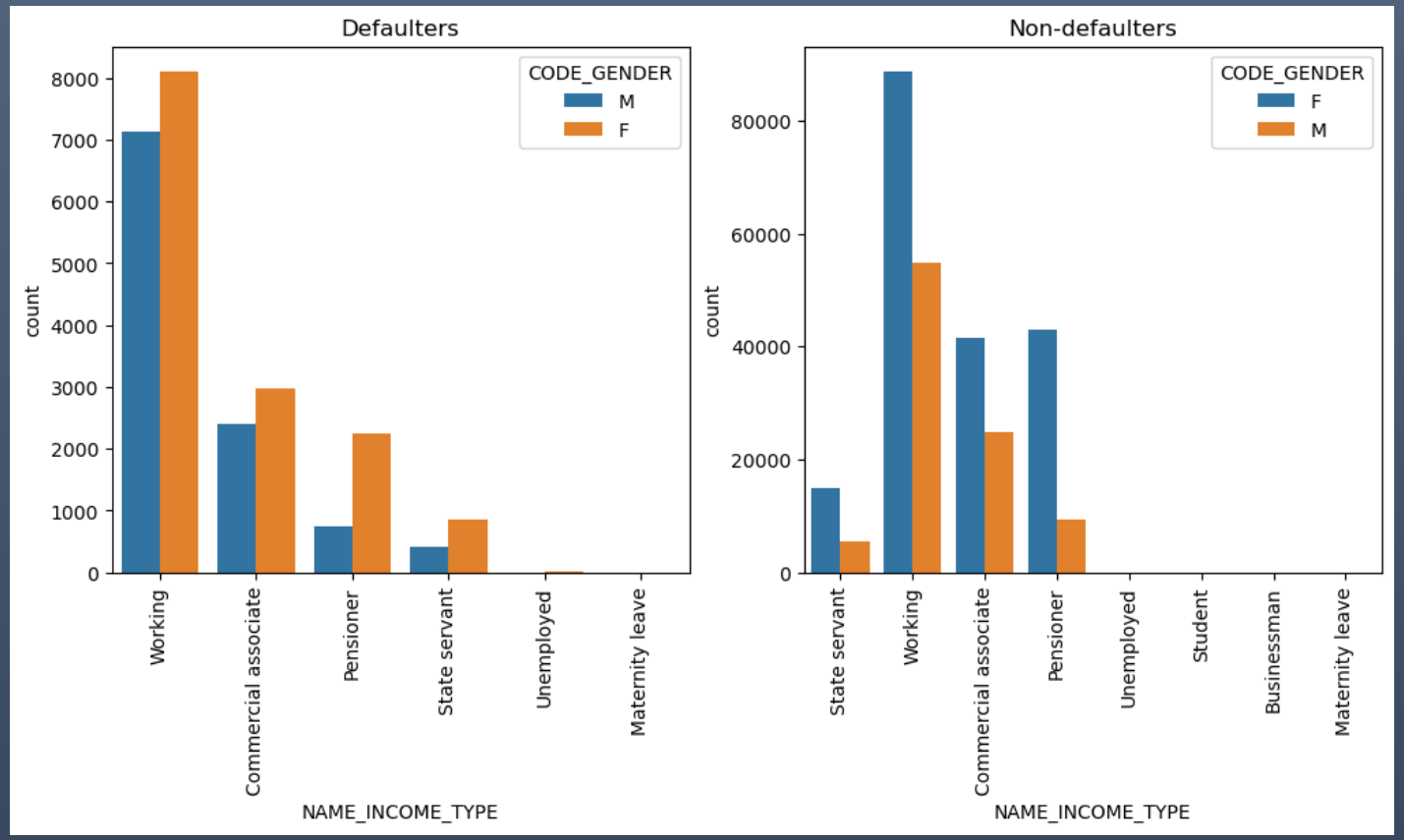
Analysis of ‘AMT_ANNUITY’ V/S ‘AMT_CREDIT’



AMT_ANNUITY’ and ‘AMT_CREDIT’ have strong positive correlation. This means that as Annuity increases, so does Credit amount.

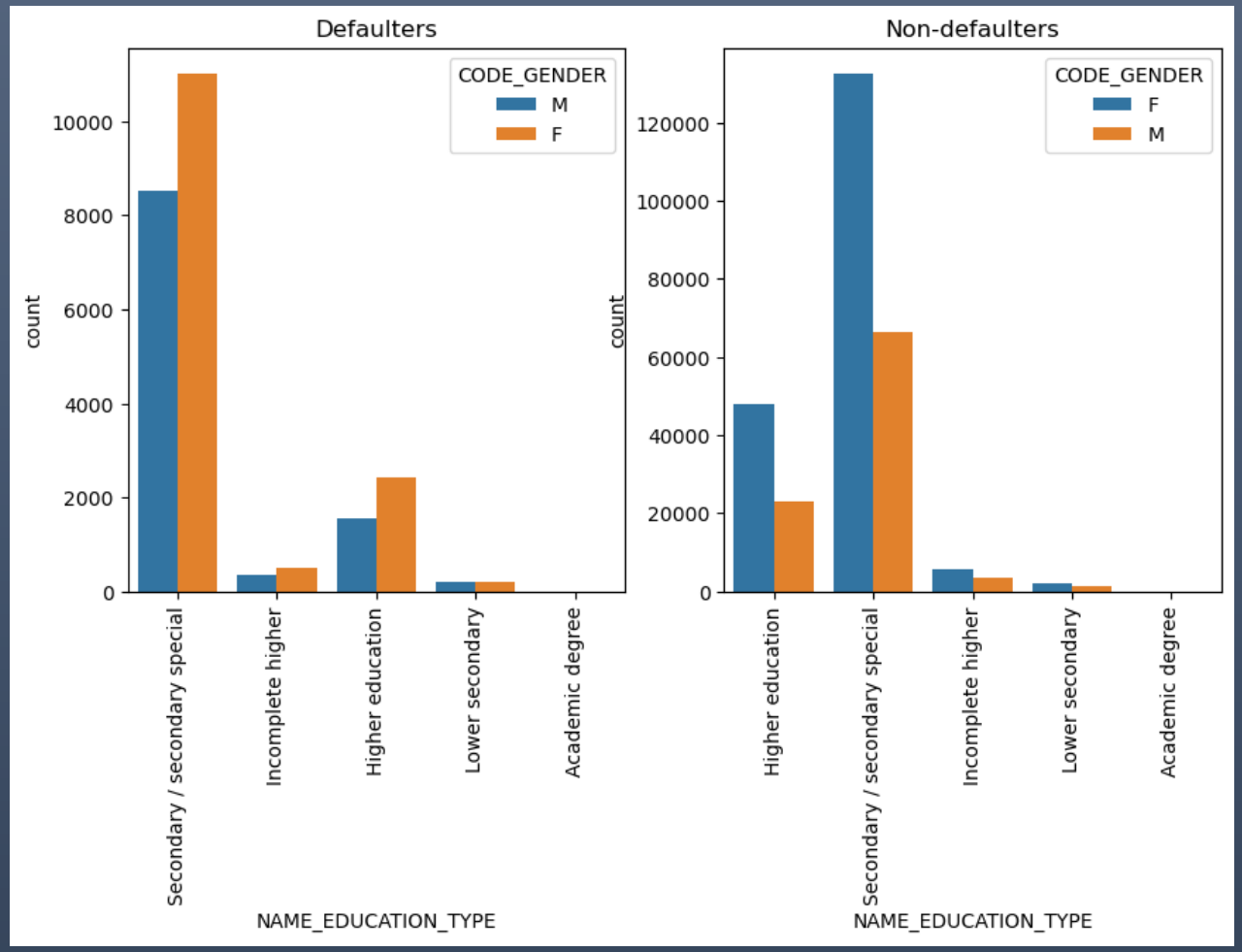
Bivariate analysis Categorical v/s Categorical variable

Analysis of ‘NAME_INCOME_TYPE’ V/S ‘CODE_GENDER’



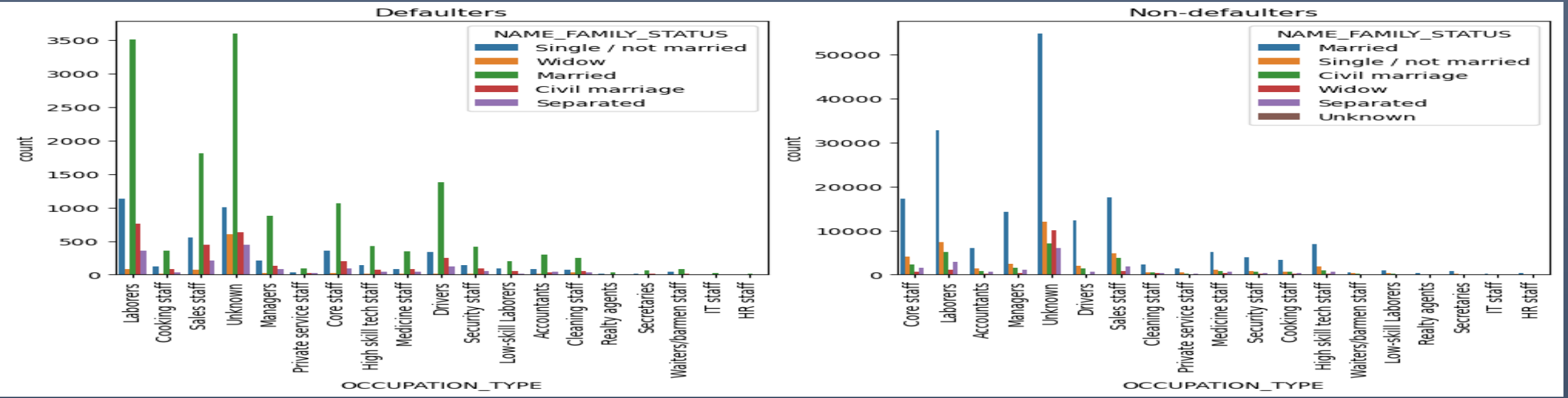
- Clients who are ‘Working’ and “Male’ have more payment difficulties compared to On-Time Payments.
- Clients who are ‘Pensioner’ and ‘Female’ have more Payment difficulties compared to On-Time Payments.
- Clients who are ‘Businessman’ and ‘Students’ do their payments On-Time though their record count is low.

Analysis of ‘NAME_EDUCATION_TYPE’ V/S ‘CODE_GENDER’



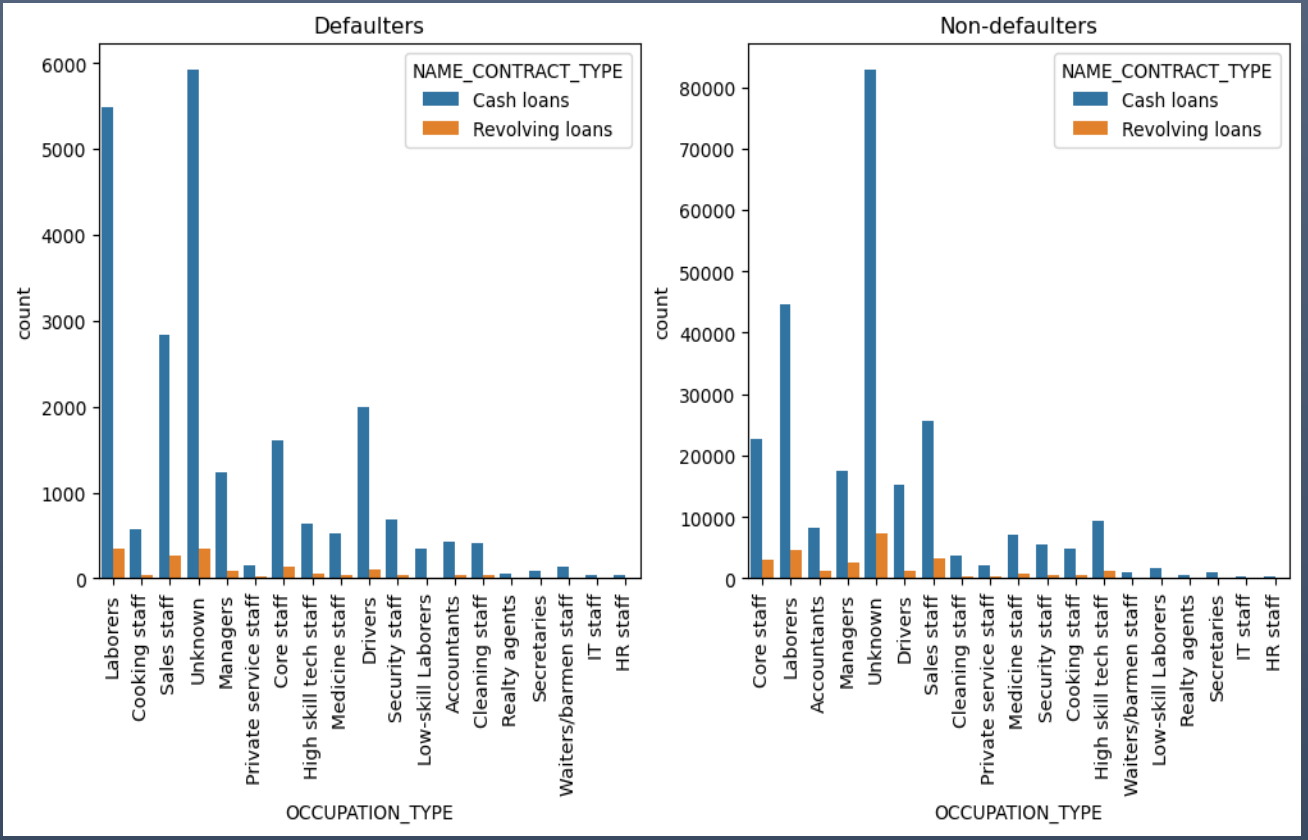
- Clients who have ‘Secondary/Secondary special’ education and ‘Male’ have more Payment difficulties compared to On-Time Payments
- Clients who have ‘Higher education’ and ‘Female’ have more On-Time Payments compared to payment difficulties

Analysis of ‘OCCUPATION_TYPE’ V/S ‘NAME_FAMILY_STATUS’



- Clients who are `Single/not married`, `Married` & `Civil marriage` and are `Waiters/barmen staff` have more Payment difficulties compared to On-Time Payments
- Clients who are `Single/not married` & `Married` and are `Laborers` have more Payment difficulties compared to On-Time Payments
- Clients who are `Married` and are `Drivers` have more Payment difficulties compared to On-Time Payments
- `Married` and `Accountants` have better On-Time Payments

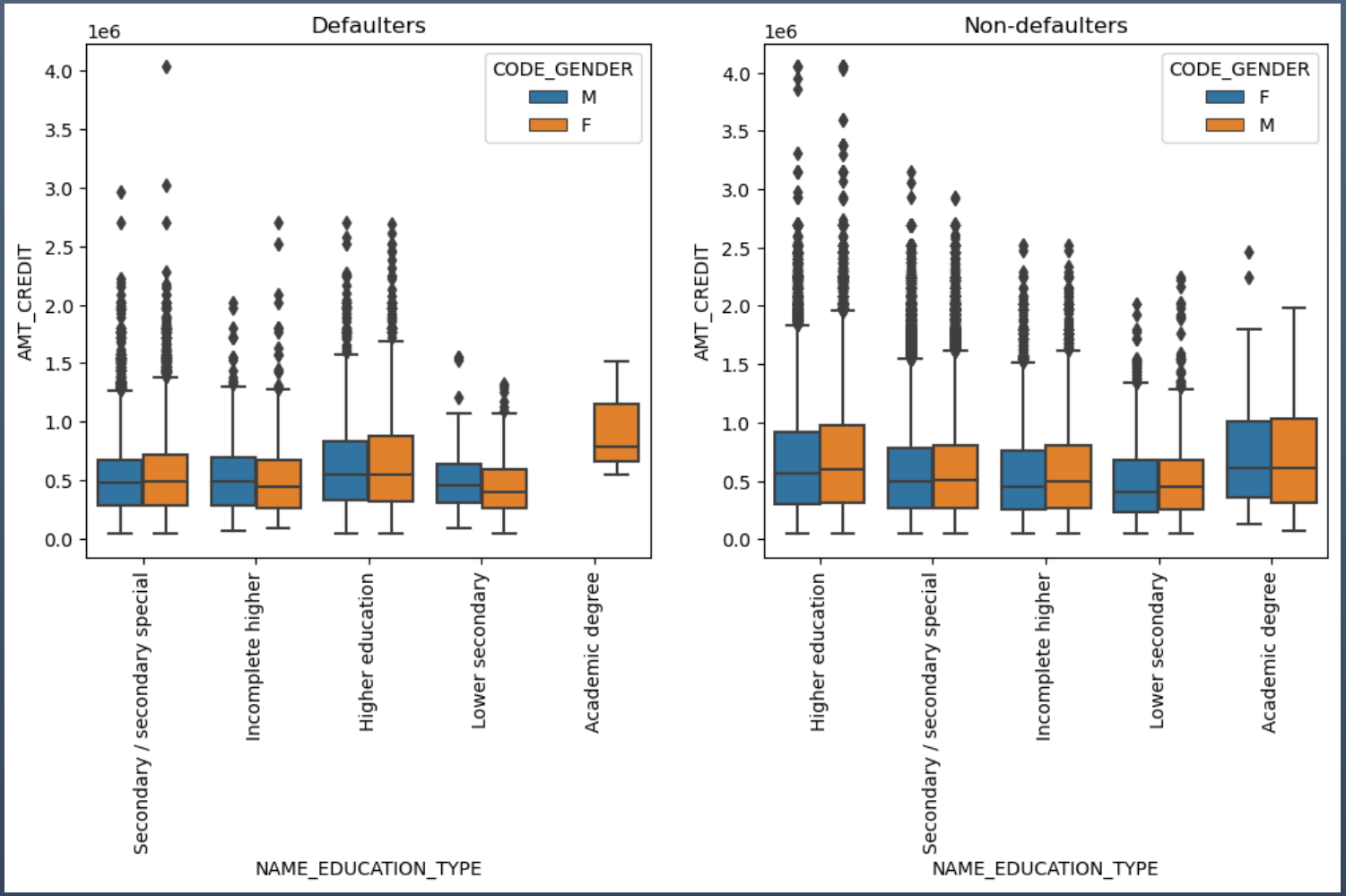
Analysis of ‘OCCUPATION_TYPE’ V/S ‘NAME_CONTRACT_TYPE’



Clients who are `Sales staff`, `Laborers`, `Drivers` and have `Cash loans` have more Payment difficulties compared to On-Time Payments

Multivariate analysis Continuous v/s Categorical variables

Analysis of : AMT_CREDIT V/S ‘EDUCATION_TYPE’ V/S ‘CODE_GENDER’

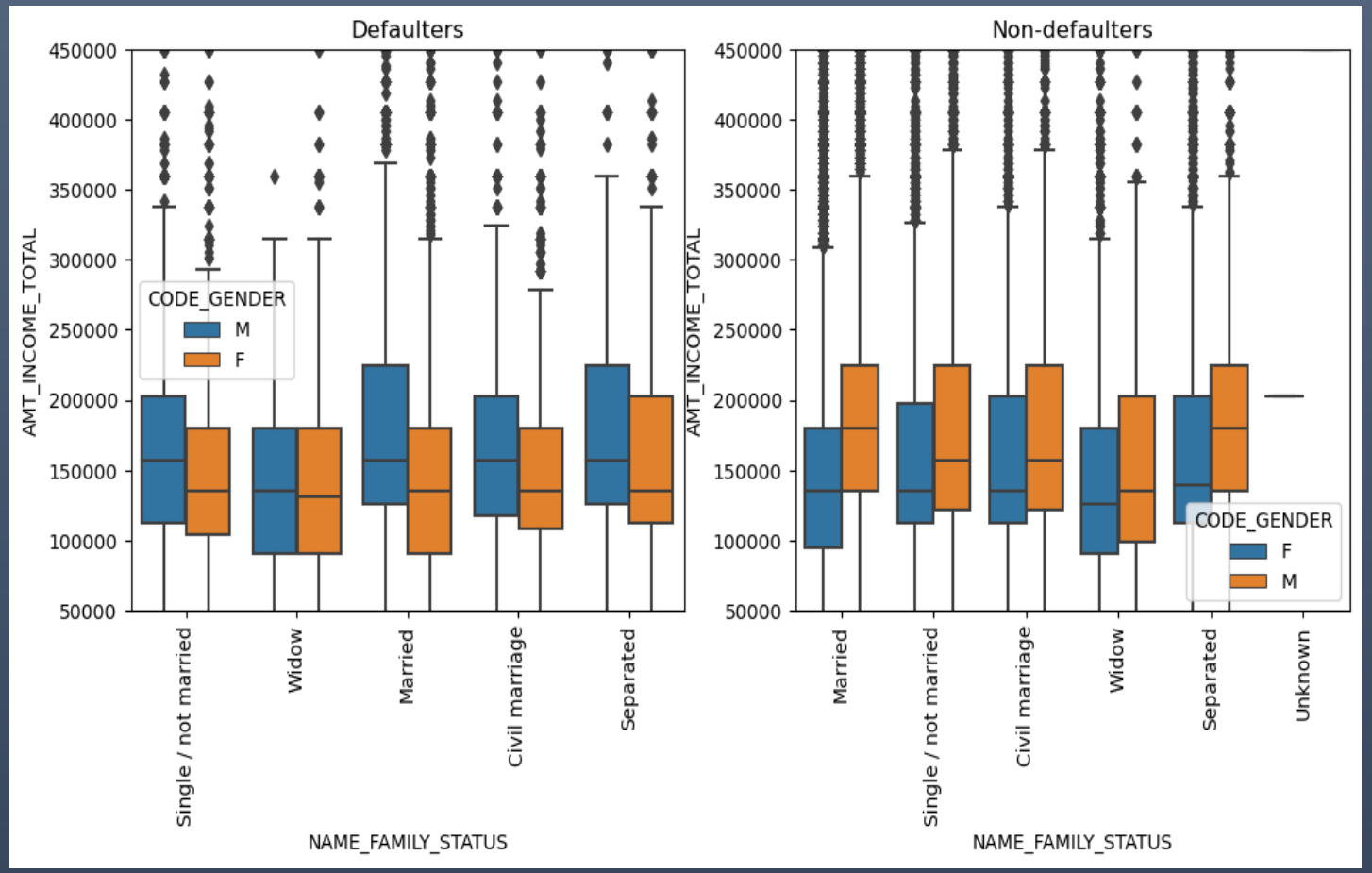


•- Clients with `Academic Degree` have a wide range of credits for On Time Payments whereas the range is much lower for ones with Payment difficulties

•- Looking at summary statistics, Clients with `Academic Degree` and Payment difficulties take mean and median credit at a much higher range than On-Time Payment clients

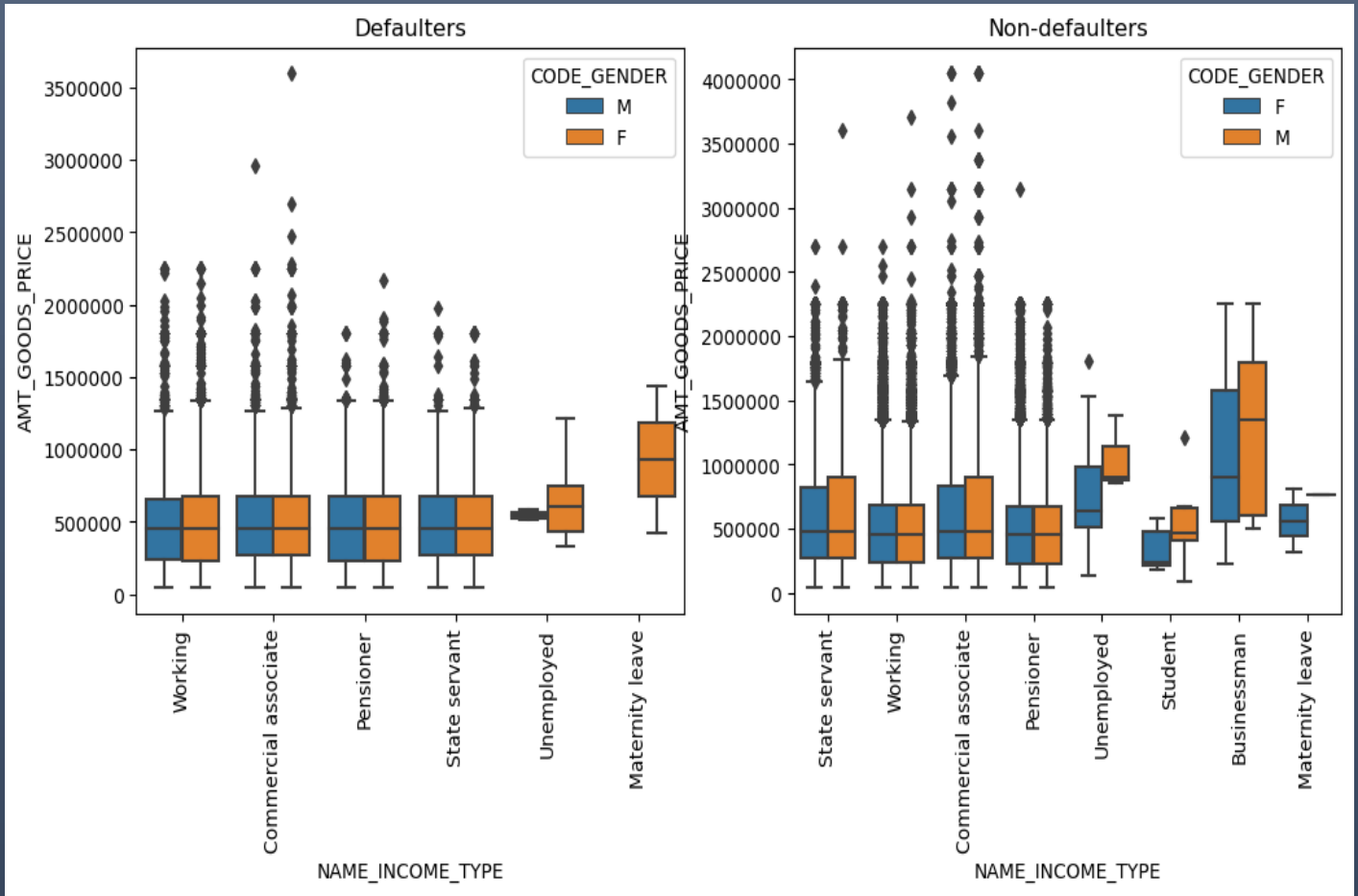
•- `Male` clients with `Academic Degree` always pay the loan on-time

Analysis of : AMT_INCOME_TOTAL V/S ‘NAME_FAMILY_STATUS’ V/S ‘CODE_GENDER’



`Married` clients have a slightly higher mean/median income with On Time Payments than Payment difficulties category

Analysis of : AMT_GOODS_PRICE V/S ‘NAME_INCOME_TYPE’ V/S ‘CODE_GENDER’

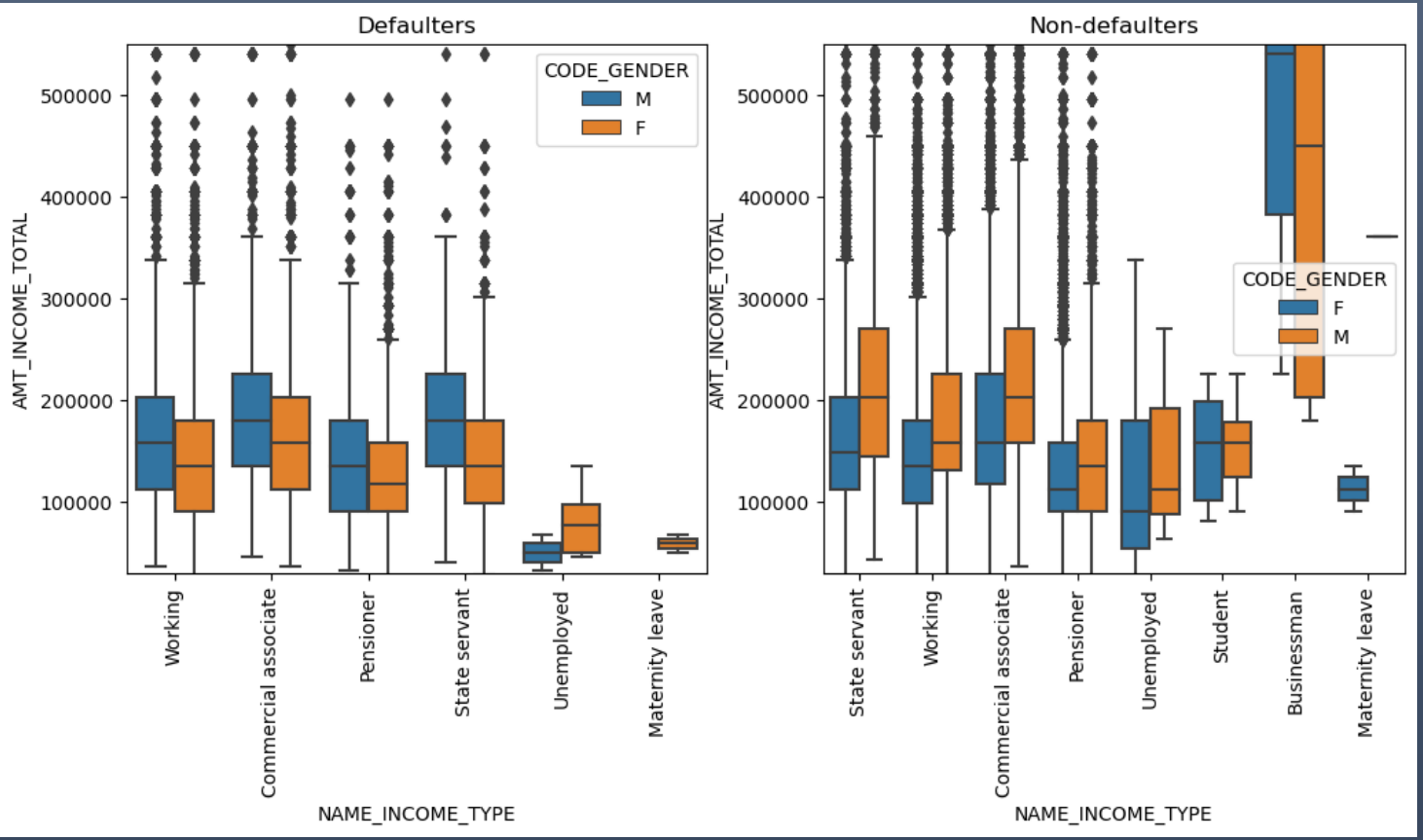


•- Clients who are `Unemployed` and `Male` have a very high price of goods in On-Time Payments than Payment difficulties

•- Clients who are `Student` and either `Male` OR `Female` do their payments On-Time. They are completely missing from Payment difficulties category. `Student` seems to be an attractive category to give loans to.

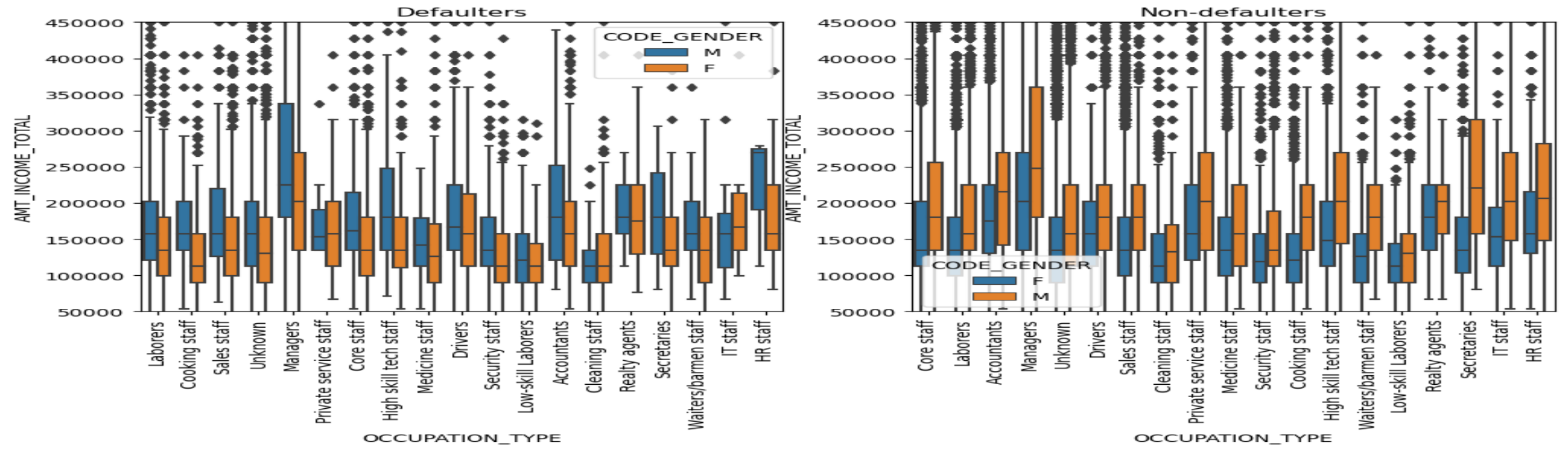
•- Clients who are `Businessman` and either `Male` OR `Female` do their payments On-Time. They are completely missing from Payment difficulties category. `Businessman` seems to be an attractive category to give loans to.

Analysis of : AMT_INCOME_TOTAL V/S ‘NAME_INCOME_TYPE’ V/S ‘CODE_GENDER’



- Clients who are `Student` and either `Male` OR `Female` do their payments On-Time. They are completely missing from Payment difficulties category. `Student` seems to be an attractive category to give loans to.
- Clients who are `Businessman` and either `Male` OR `Female` do their payments On-Time. They are completely missing from Payment difficulties category. `Businessman` seems to be an attractive category to give loans to.
- Clients who are in `Maternity Leave` and `Female` have a very high income in On-Time Payments than Payment difficulties

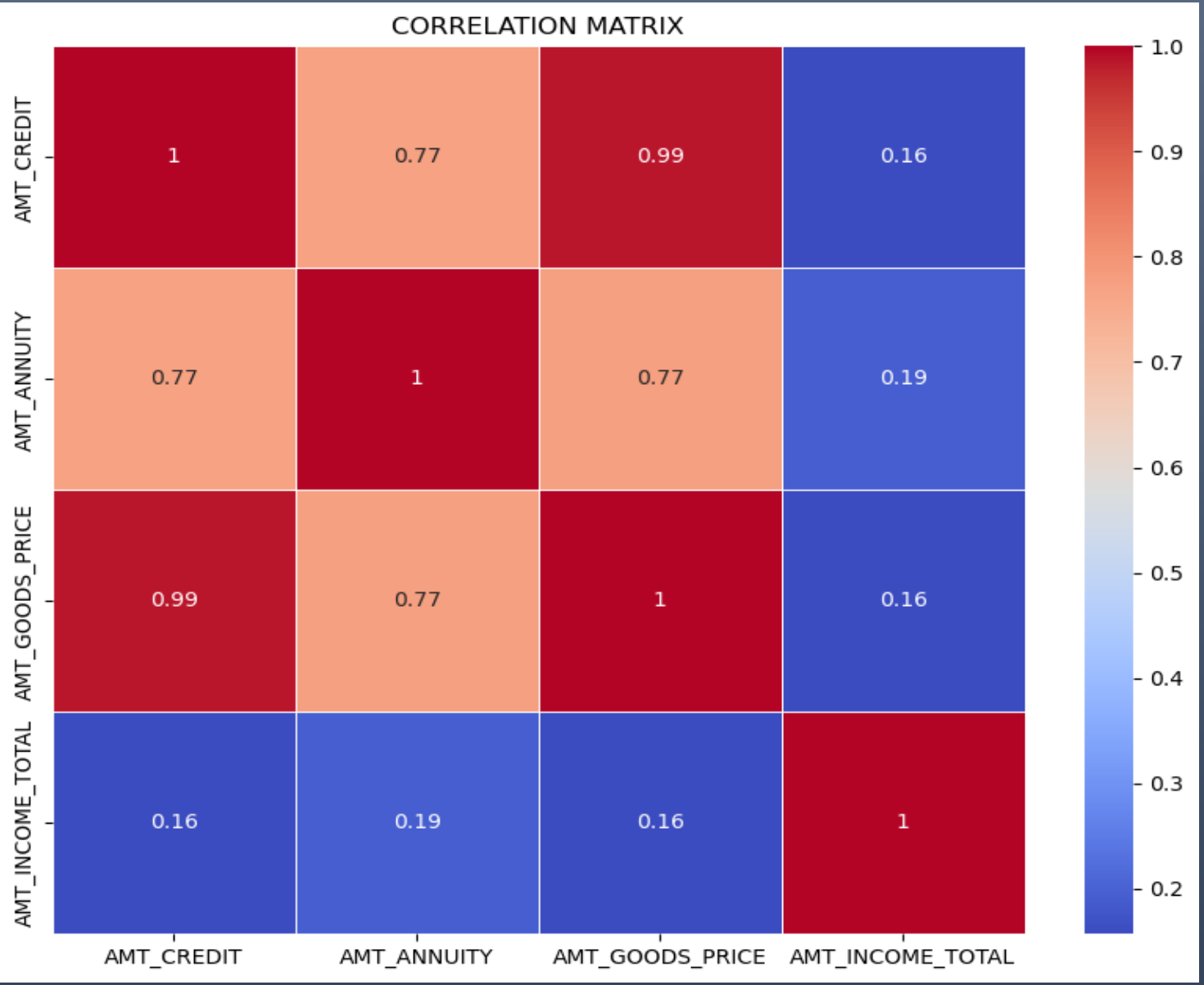
Analysis of : AMT_INCOME_TOTAL V/S 'OCCUPATION TYPE' V/S 'CODE_GENDER'



- Clients who are `Waiters/garment staff` and `female` have less median income in On-Time Payments than Payment difficulties
- Clients who are `Cleaning staff` and `female` have more median income in On-Time Payments than Payment difficulties
- Clients who are `HR Staff` and `Male` have more median income in Payment difficulties than On-Time Payments
- Clients who are `Managers` and `Male` have more median income in On-Time Payments than Payment difficulties

Correlation analysis of numerical variables

Correlation Matrix



- 'AMT_CREDIT' has a high correlation with 'AMT_GOODS_PRICE'
- 'AMT_ANNUITY' has decent correlation with 'AMT_CREDIT' and 'AMT_GOODS_PRICE'.

Conclusion

1.Default rate of female customers is lower than Males.
2.Females are more than male in having credits for that range.
3.Working, Commercial associates, pensioners and state servants are safest to give loans.
4.Unemployed and Maternity leave segments faced difficulty in repaying the loan.
5.Unaccompanied and family has taken more load and repaid the most.
6.Higher education and secondary education are the safest segments.
7.Married people are safer to target.
8.People who has their own house or apartment are safest to target.
9.Laborers, Accountants are safest to give loans.
10.Drivers and low skill laborers are most defaulters.
11.Business Entity Type 3 and self employed are the highest in having loans from the bank.
12.Transportation Type 3 are the highest in default rate.
13.Most of the customers has Cash loan.
14.Customers who have taken cash loan are most likely to default.
15.Most of the clients have income between 1 to 2 million.
16.Most of the loan were given for the amount of 0 to 1 million.
17.Most of the clients are paying annuity between 0 to 1Lakh.
18.Most of the clients have taken loan for goods price ranging 0 to 2 million.
19.Most of the defaulters are lying in the range of 0 to 1 million credit amount.
20.Almost all client's income is below 1 million and the chances of default are hire when the credit amount is less the 1.5 million.
21.Clients who have 0-5 children are safer to give loan.