EXPLORATORY DATA ANALYSIS: CREDIT

KUNAL KASHYAP

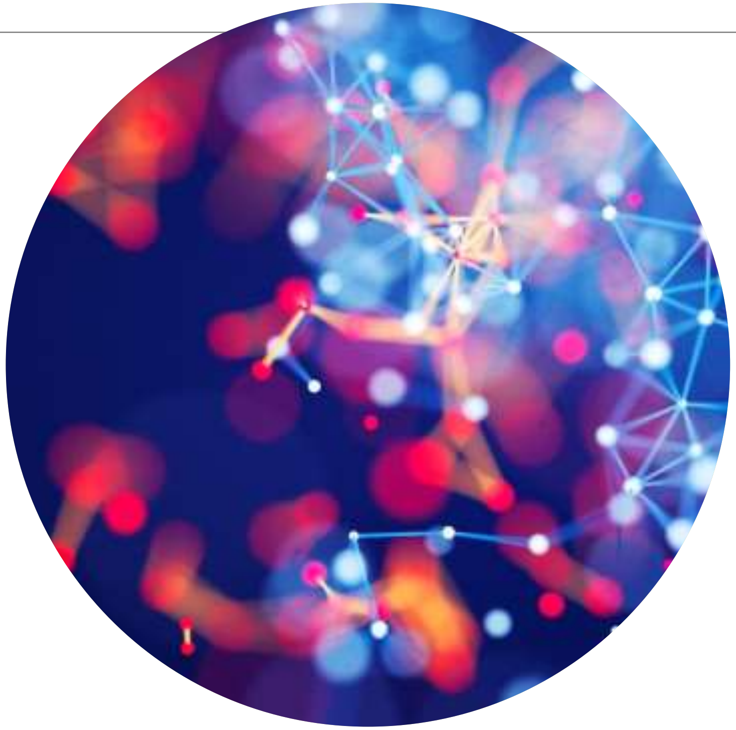
DSC:40

Agenda

PROBLEM STATEMENT

EDA APPROACH:

- MISSING VALUES AND OUTLIER
- DATA IMBALANCE
- ANALYSIS
- OBSERVATION






Introduction

Problem Statement

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.

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.

Resources / Dataset provided

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 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.

Without data you’re just another person with an opinion

W. EDWARDS DEMING

EDA APPROACH

Process taken todays analysis of the datasets:

1. Understanding the dataset.
2. Importing datasets :
    1. NumPy
    2. Pandas
    3. Seaborn
    4. Warnings (Need to import due to seaborn library)
    5. Missingno (for graphically representing the missing values in the dataset.
    6. Matplotlib
3. Checking the Structure of the Dataset:
    1. Shape of the data
    2. Detail information of the data to know about the data types and its null values
    3. Describe the statistical values from the data.

4. Inspecting the null values
5. Dropping those columns with high null values. And using interpolation to modify the remaining null values.
6. Splitting the data set in Target variable 0 and 1.
    1. 0   : Non Defaulter
    2. 1   : Defaulter
7. Visualising the data set using seaborn and matplotlib.

Analysis and Visualization

Missing Values and Outliers

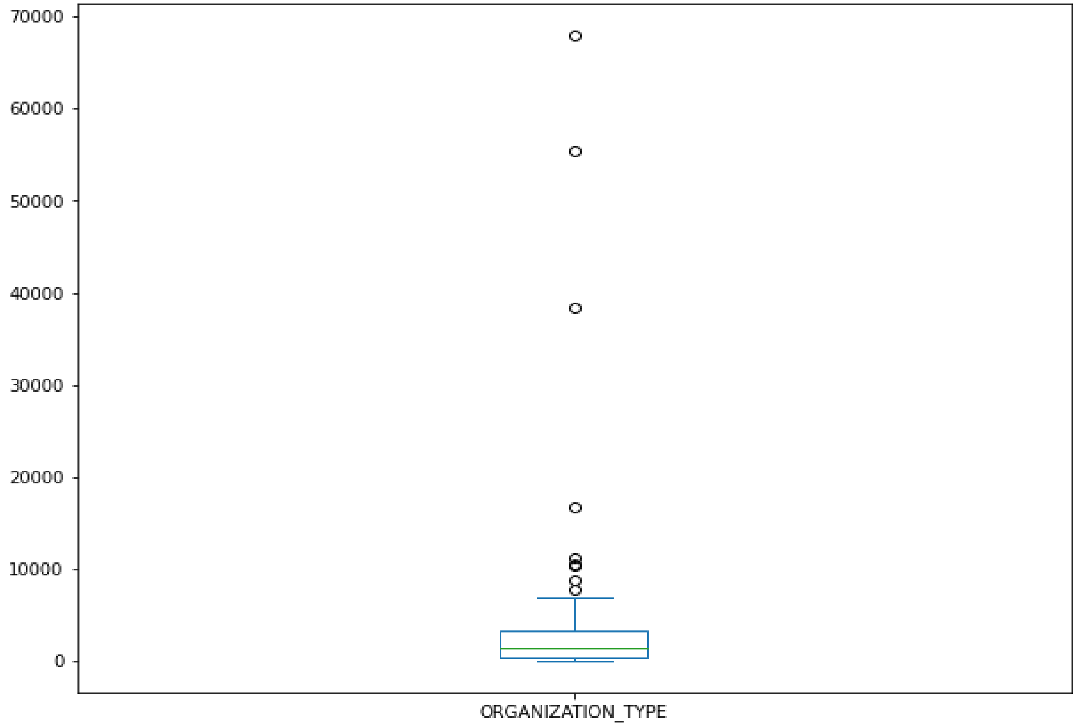
Missing Values in the data set above 50% have been dropped.

And the remaining missing values are filled using interpolation method.

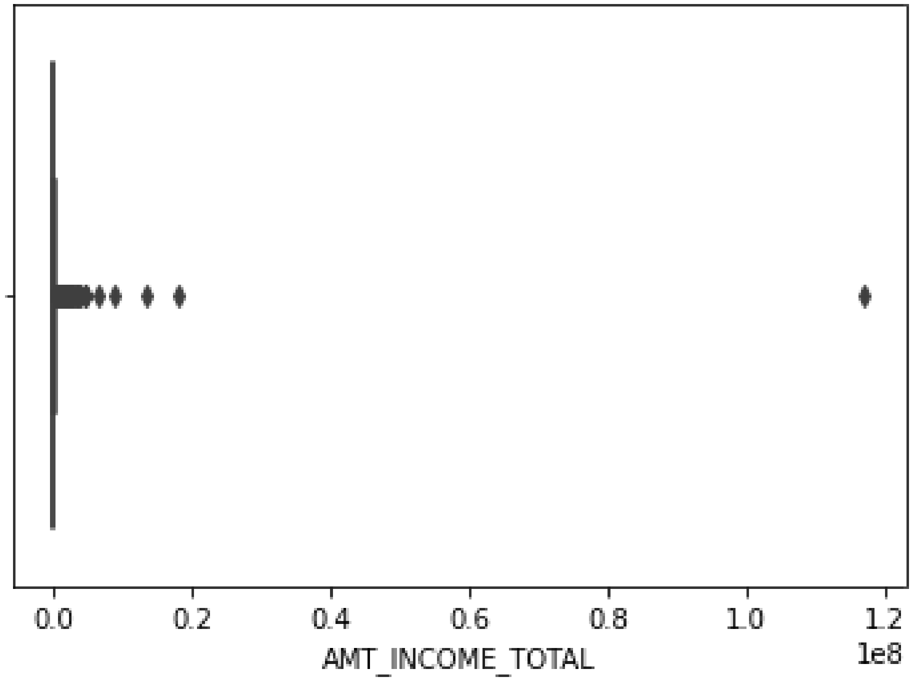
Note: Interpolation : The method of producing new data points based on the range of a discrete set of known data points is known as interpolation.

Outliers

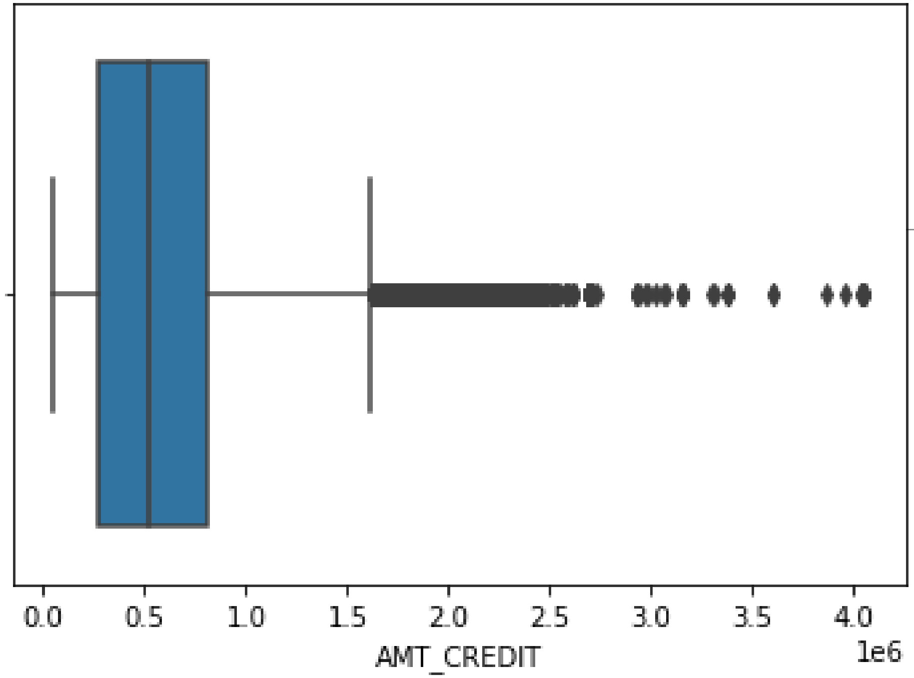
ORGANIZATION OUTLIER



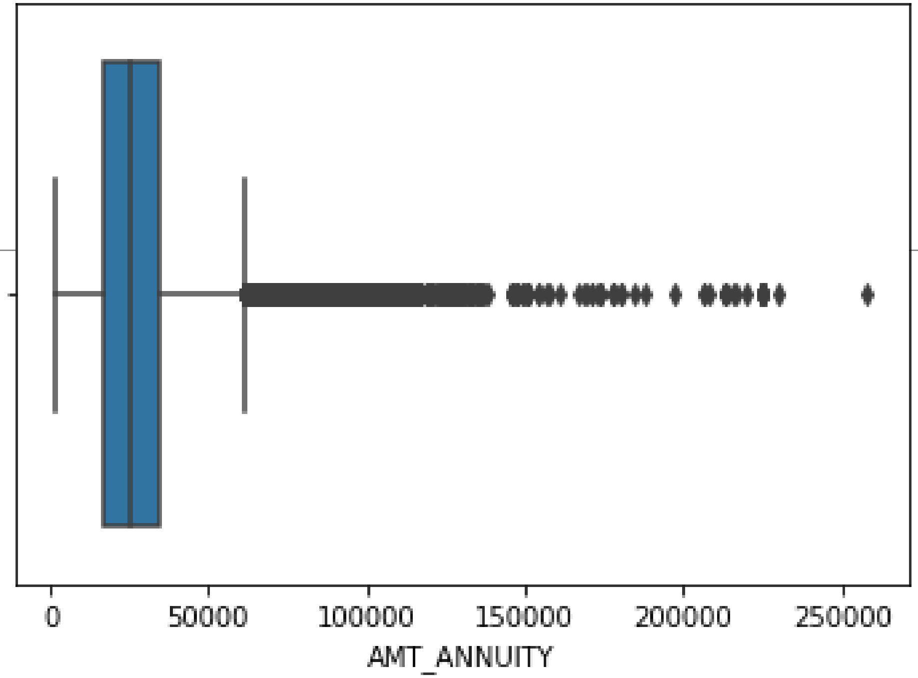
INCOME OUTLIER



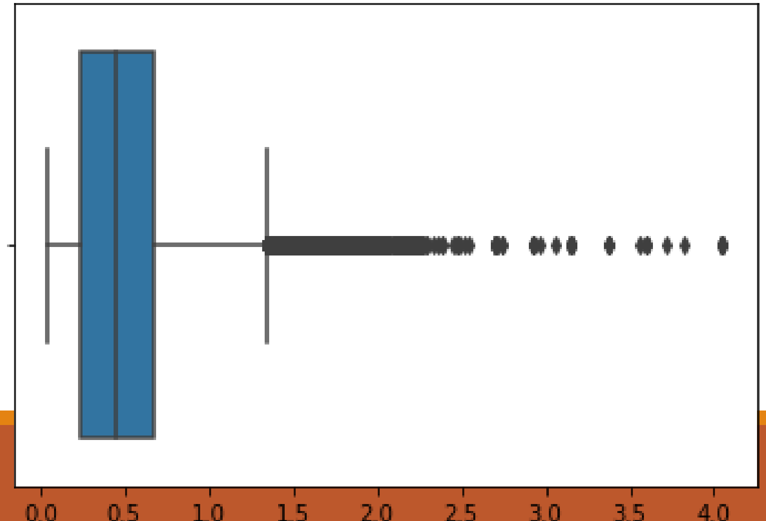
CREDIT OUTLIER



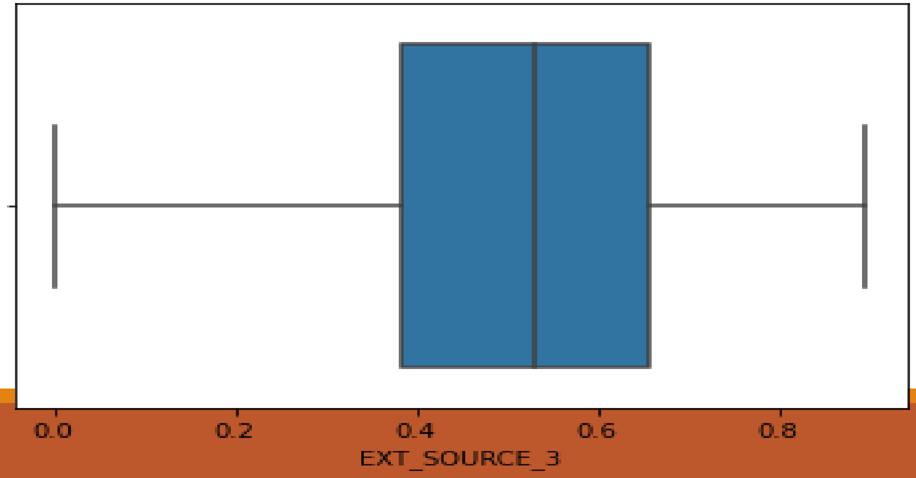
ANNUITY OUTLIER



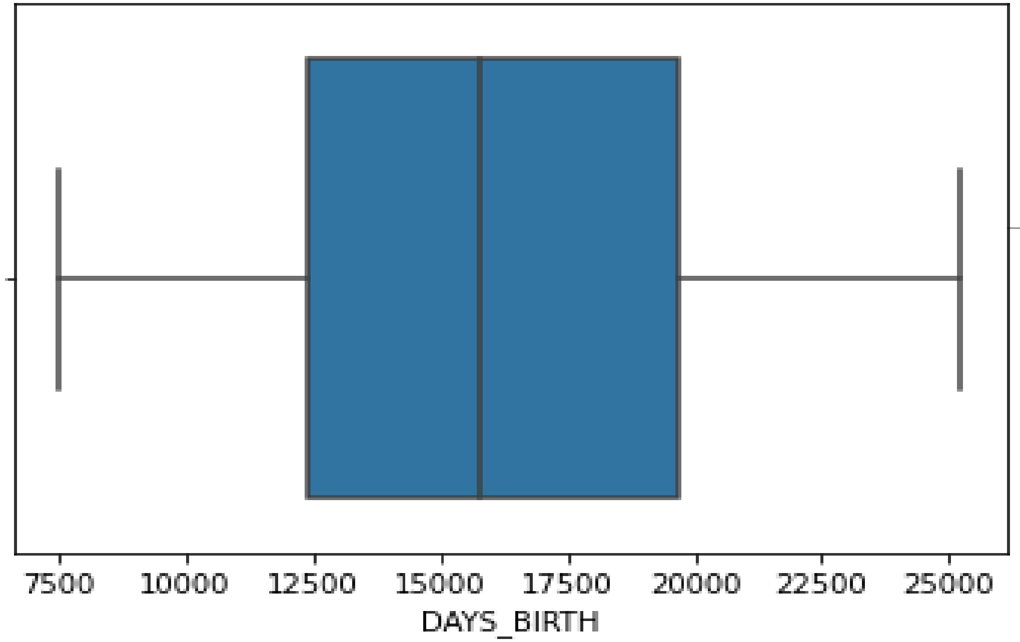
GOODS PRICE OUTLIER



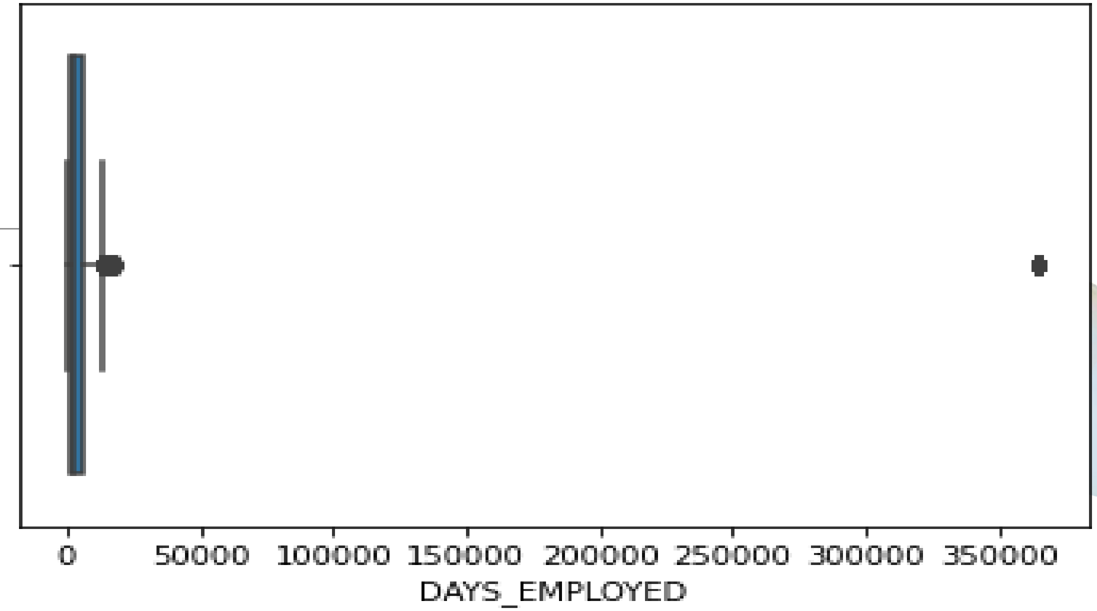
EXTERNAL SOURCE OUTLIER



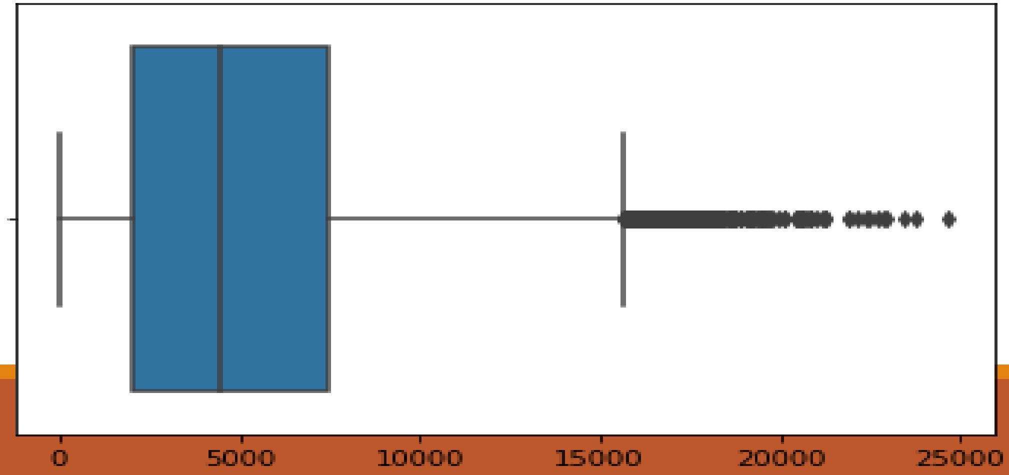
BIRTH OUTLIER



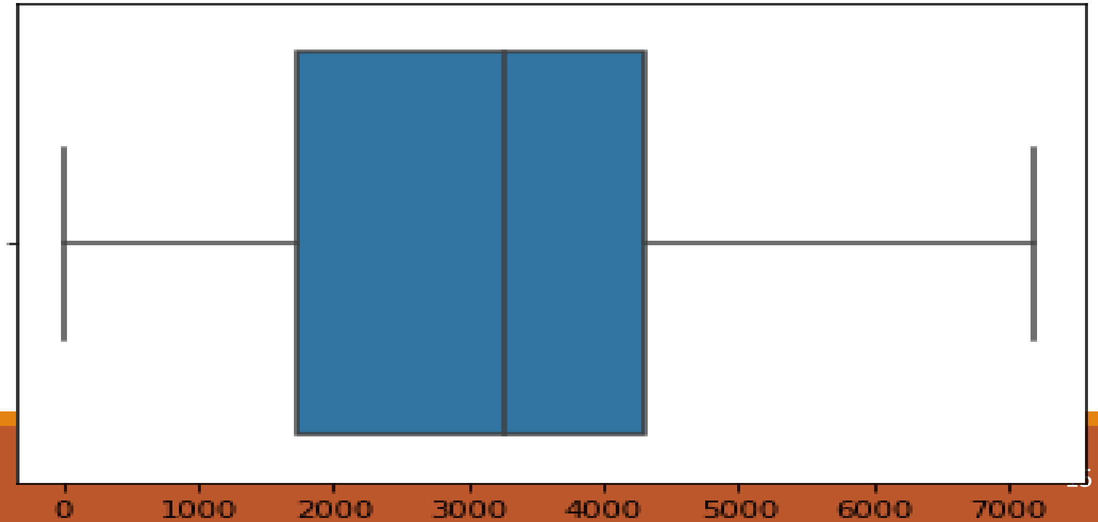
DAYS EMPLOYED OUTLIER



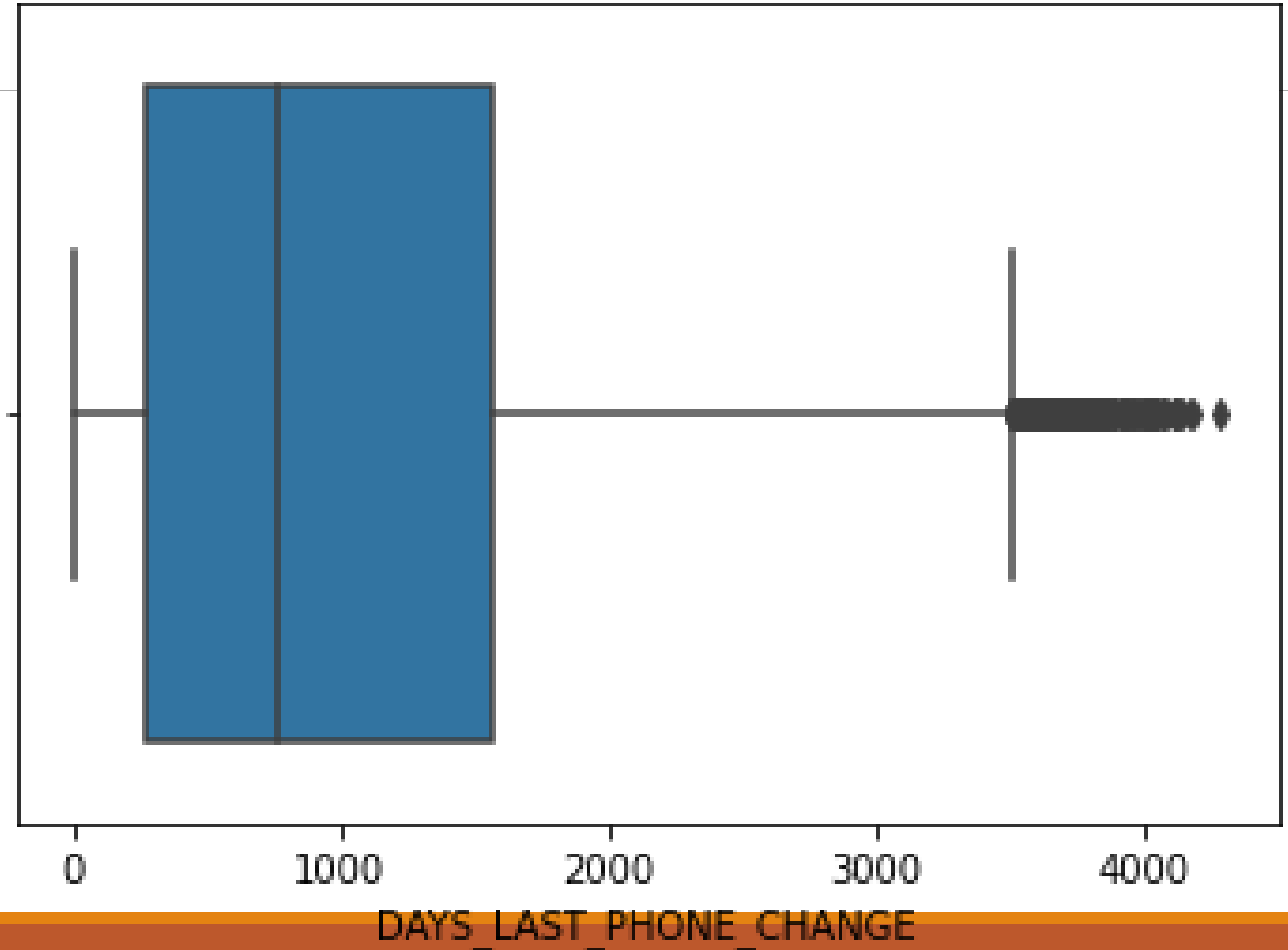
DAYS REGISTRATION OUTLIER



DAYS ID PUBLISHED OUTLIER

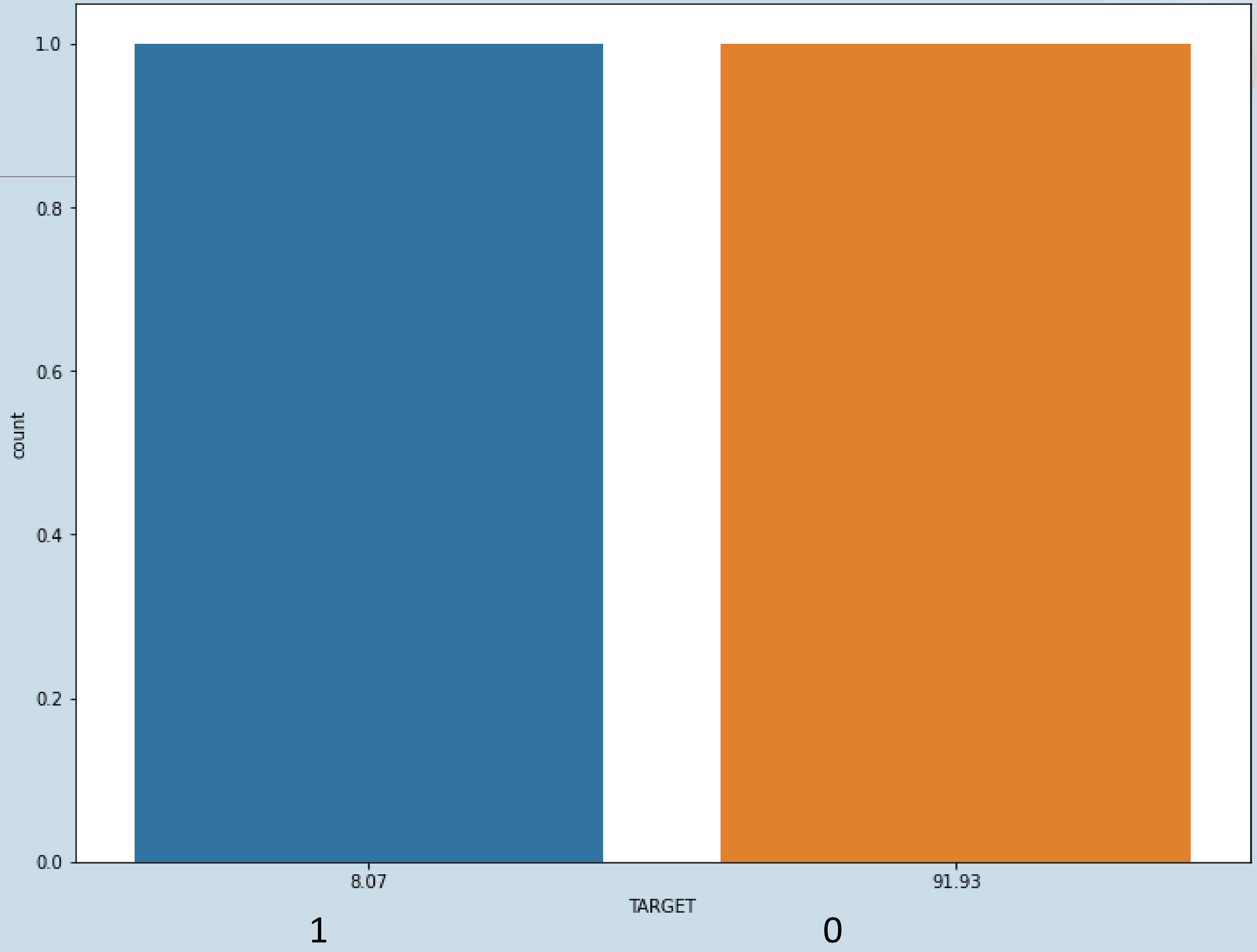


PHONE NUMBER CHANGE OUTLIER



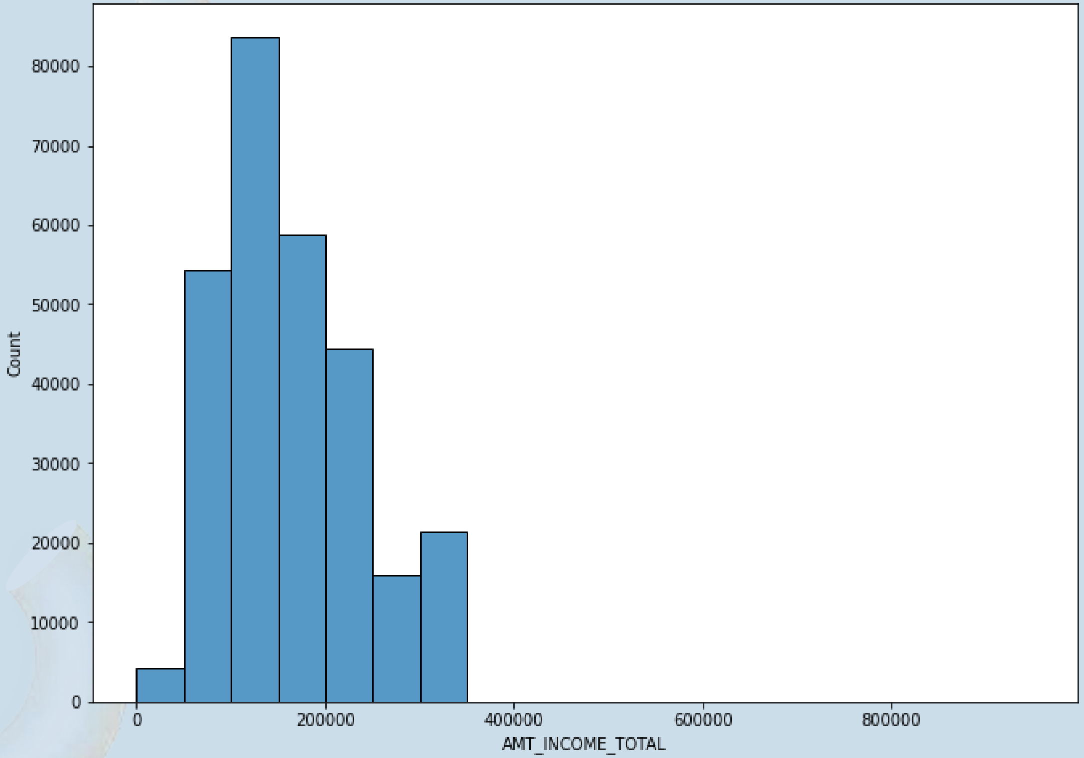
Data Imbalance

TARGET COLUMN IS BEING USED TO DIVIDE THE DATASET INTO TWO CATEGORIES.

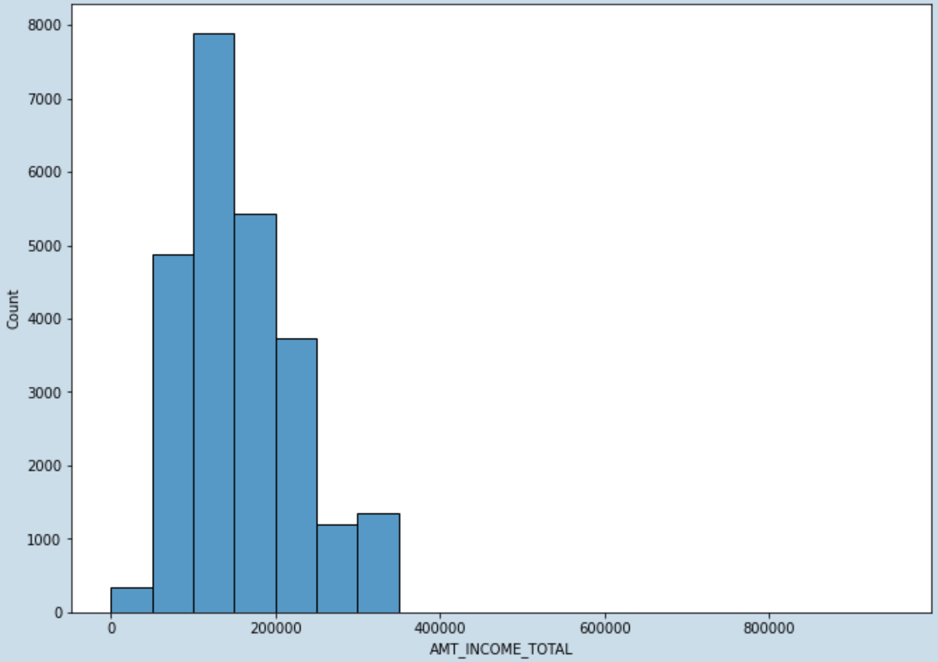


Distribution of Income in Target 1 and Target 0 client.

Target 0

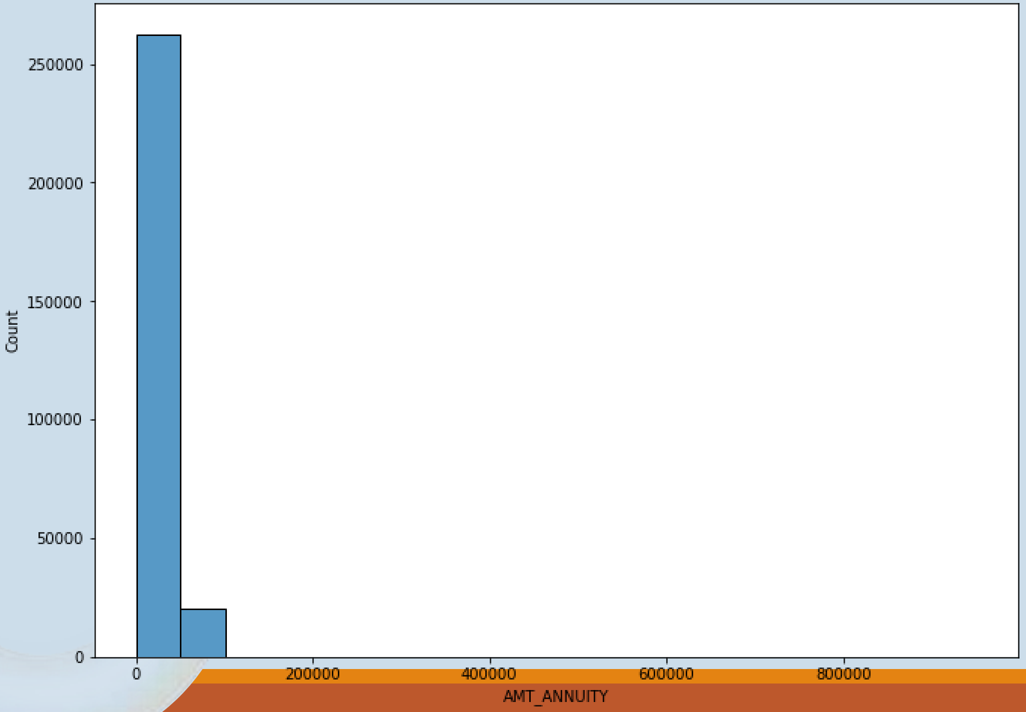


Target 1

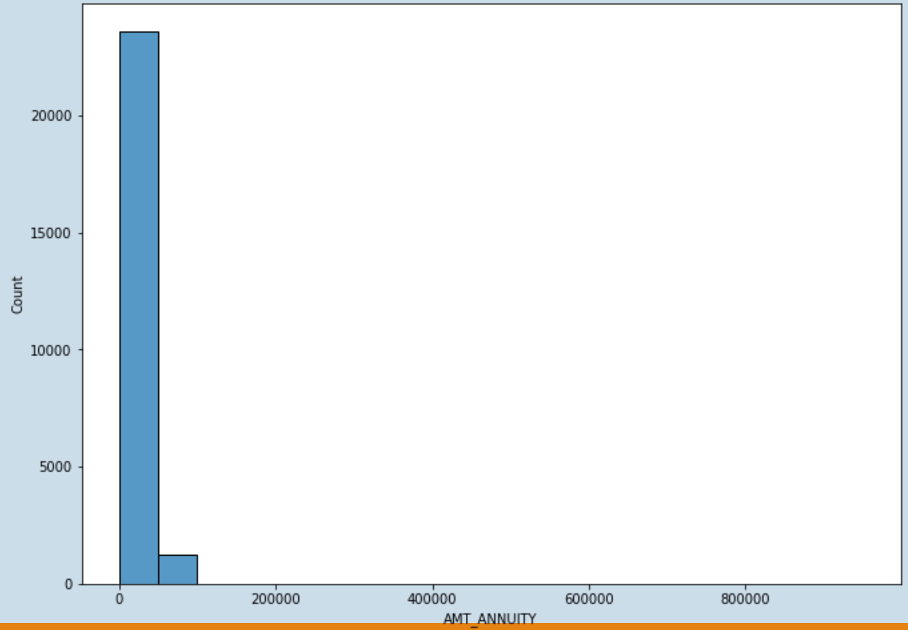


Distribution of Annuity in Target 1 and Target 0 client.

Target 0

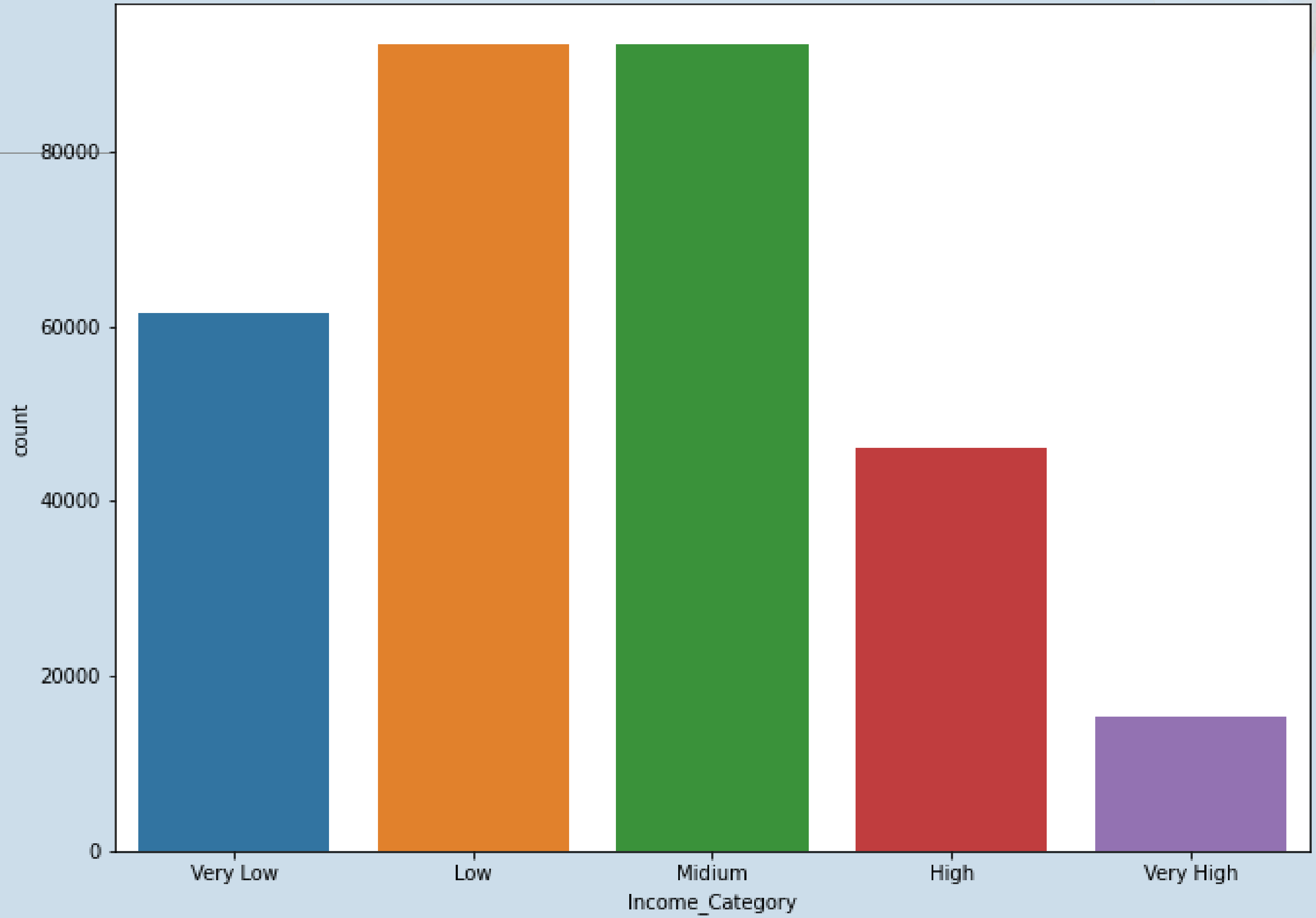


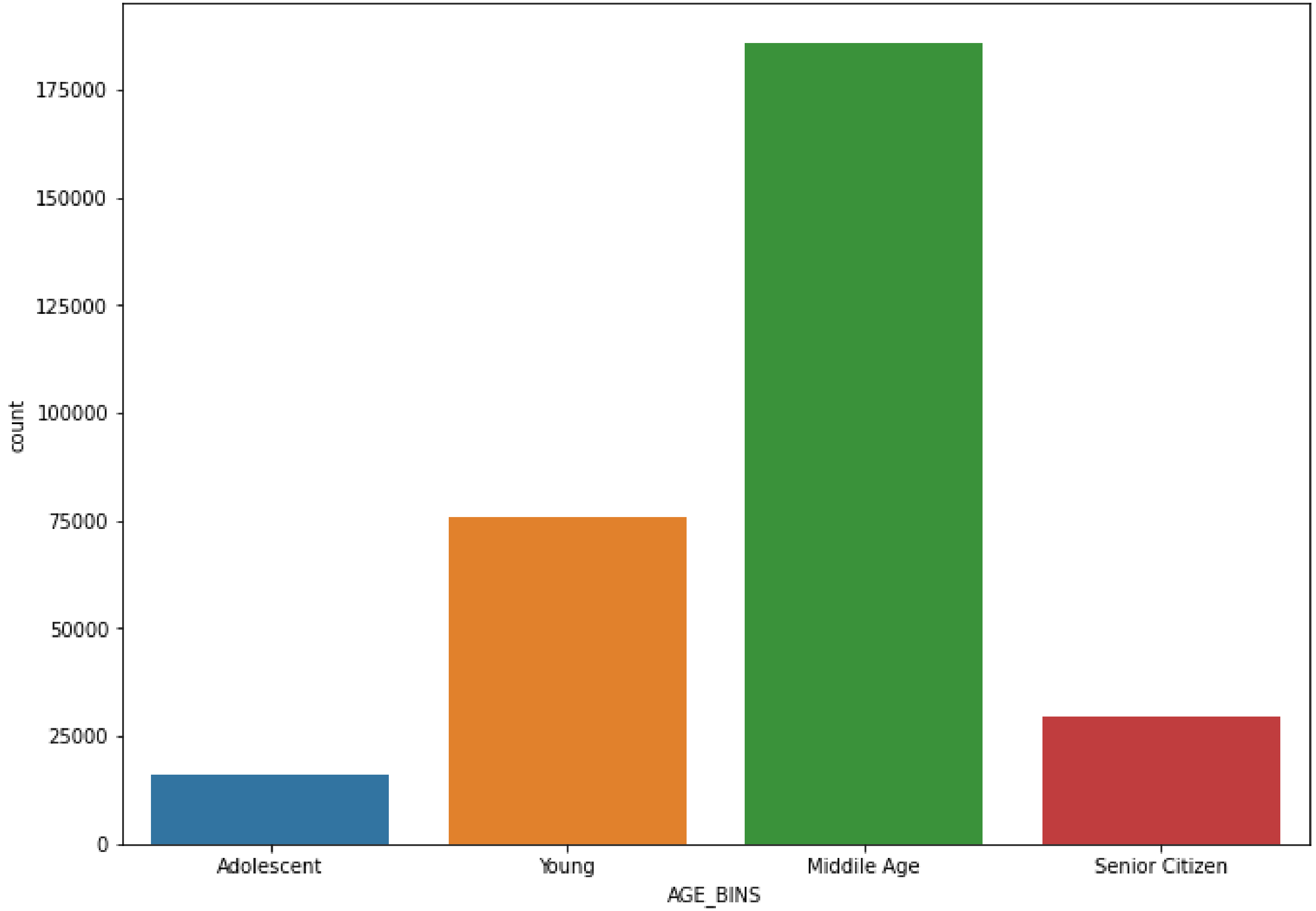
Target 1



Continuous variable:

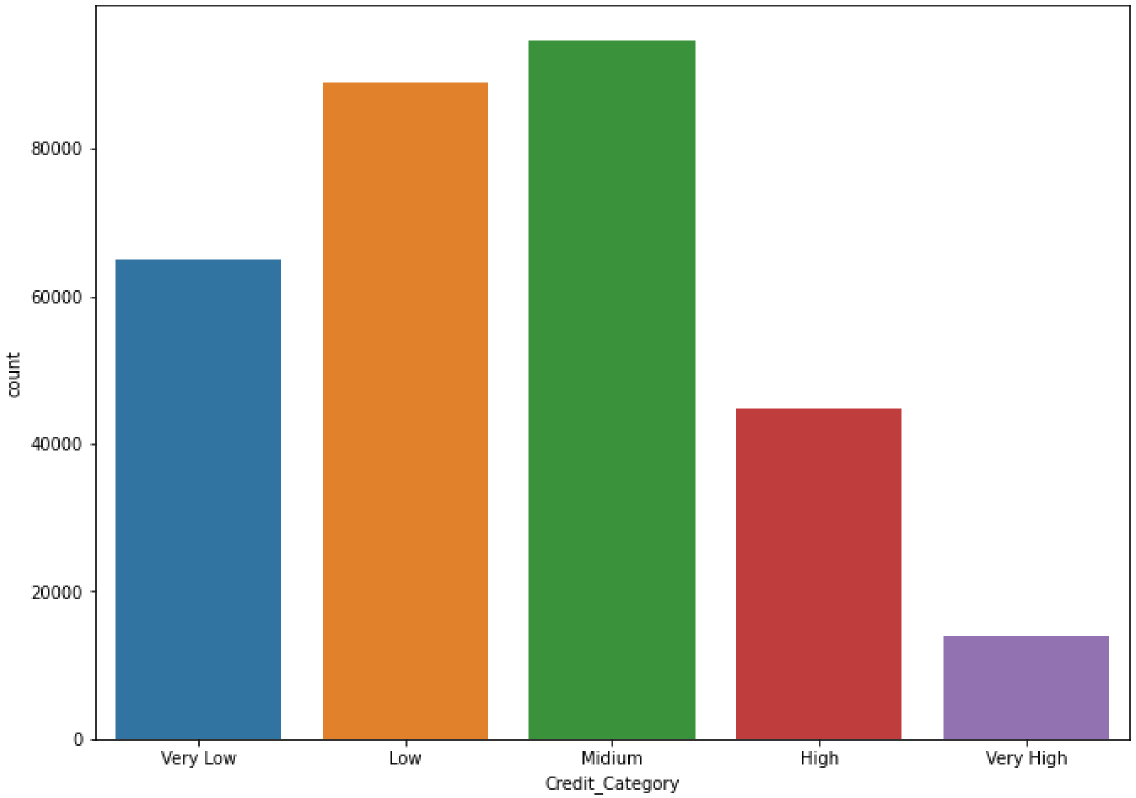
Binning of Income.





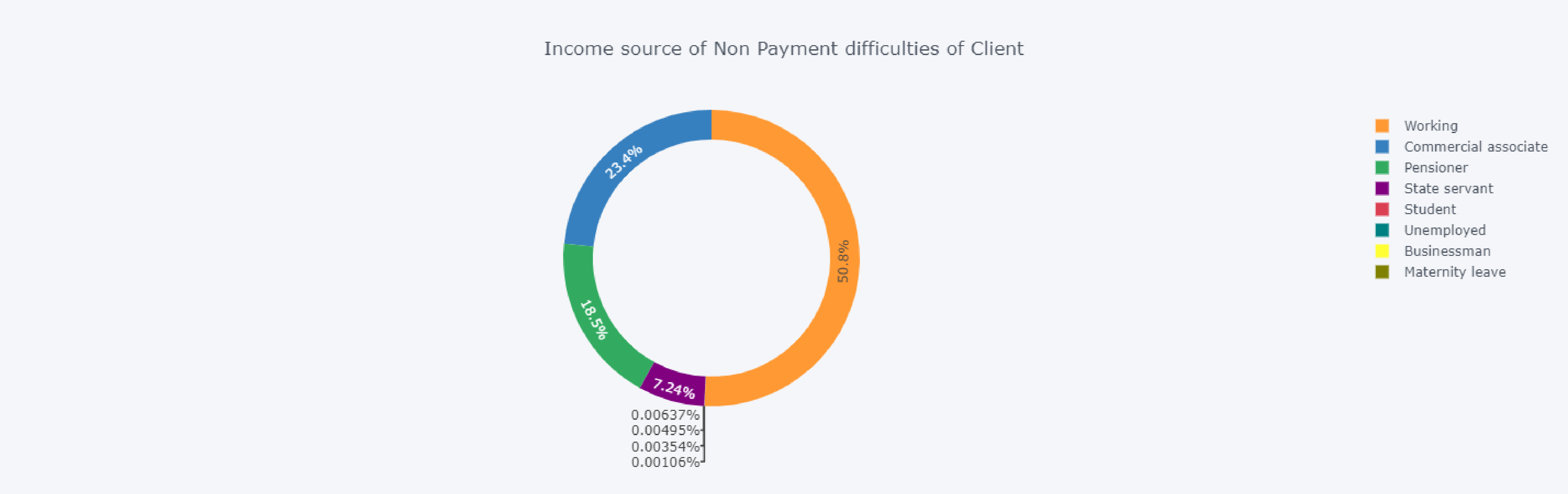
Continuous variable:

Binning of Age (in Years).

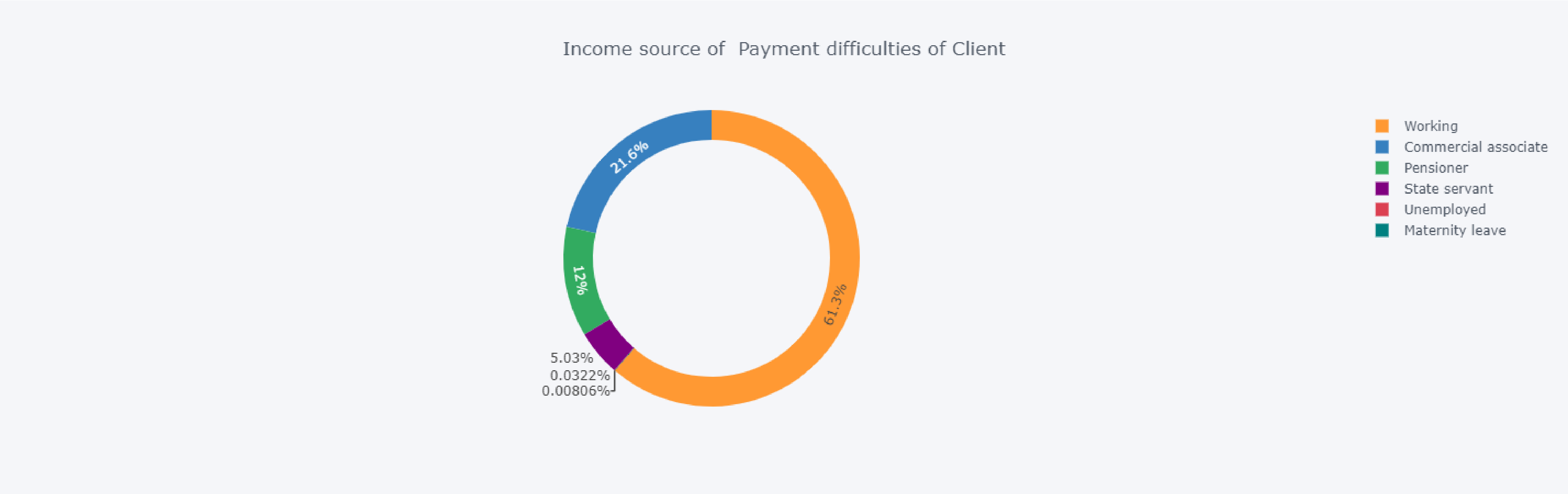


Continuous variable:

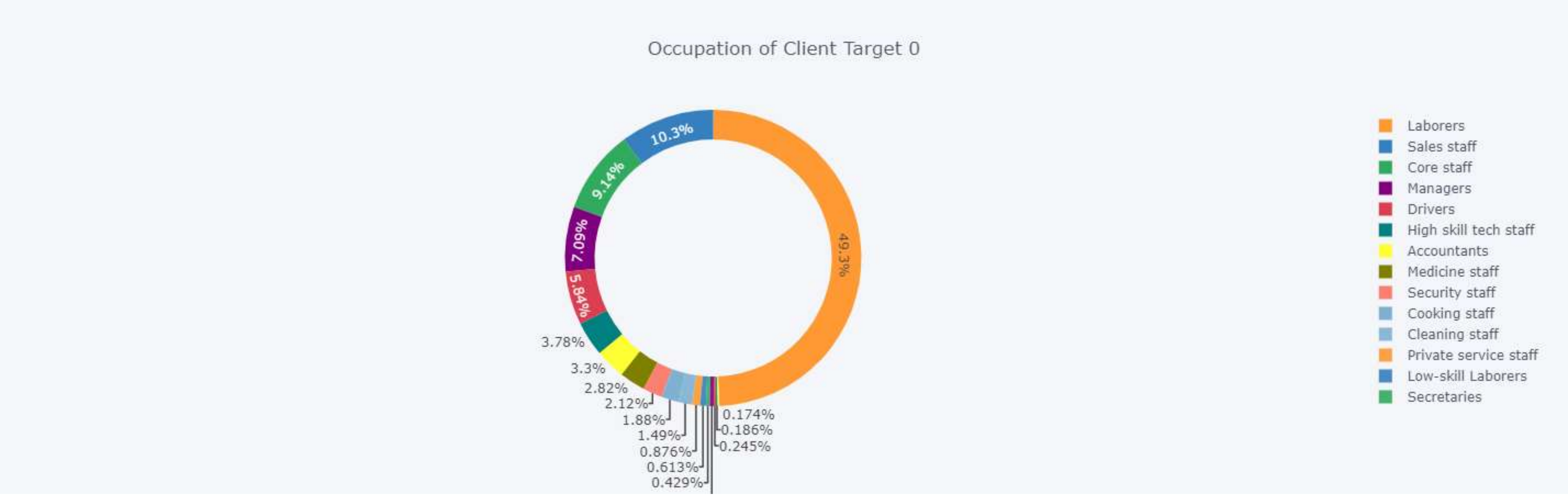
Binning of CREDIT.



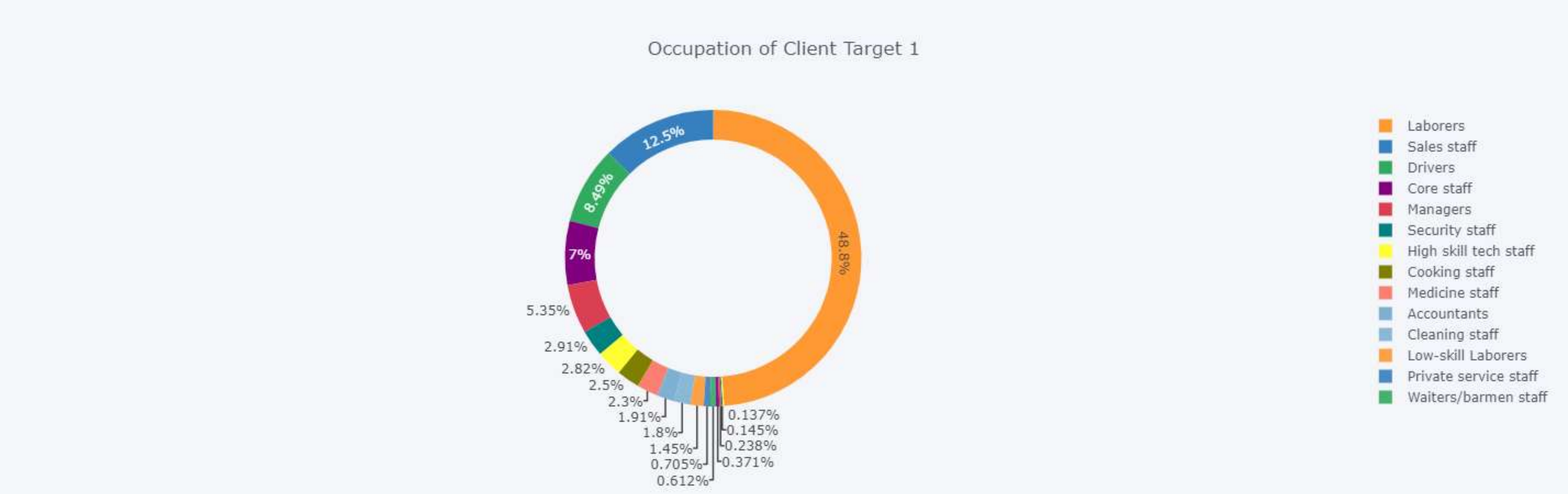
TARGET 0:
INCOME SOURCE



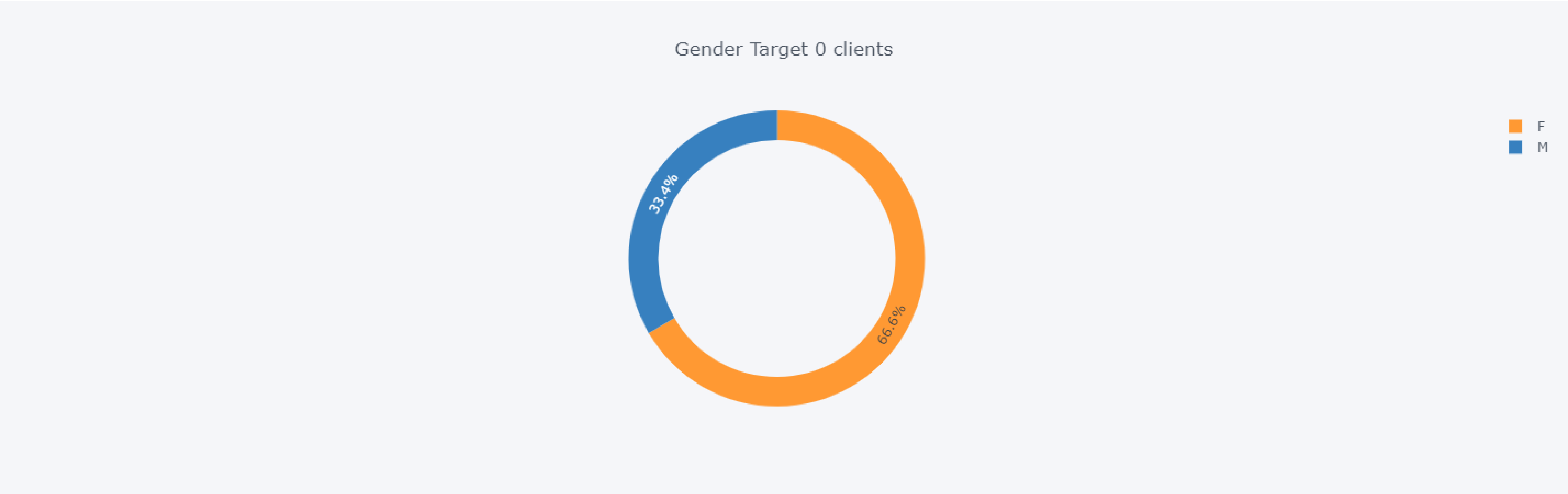
TARGET 1: INCOME SOURCE



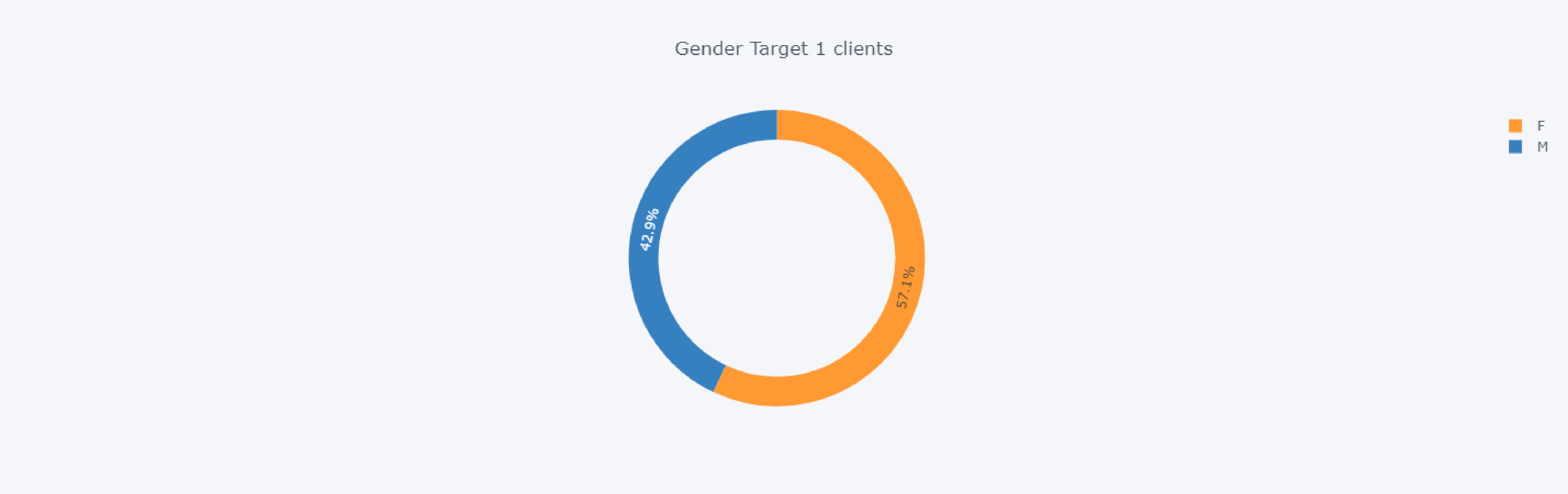
TARGET 0: OCCUPATION



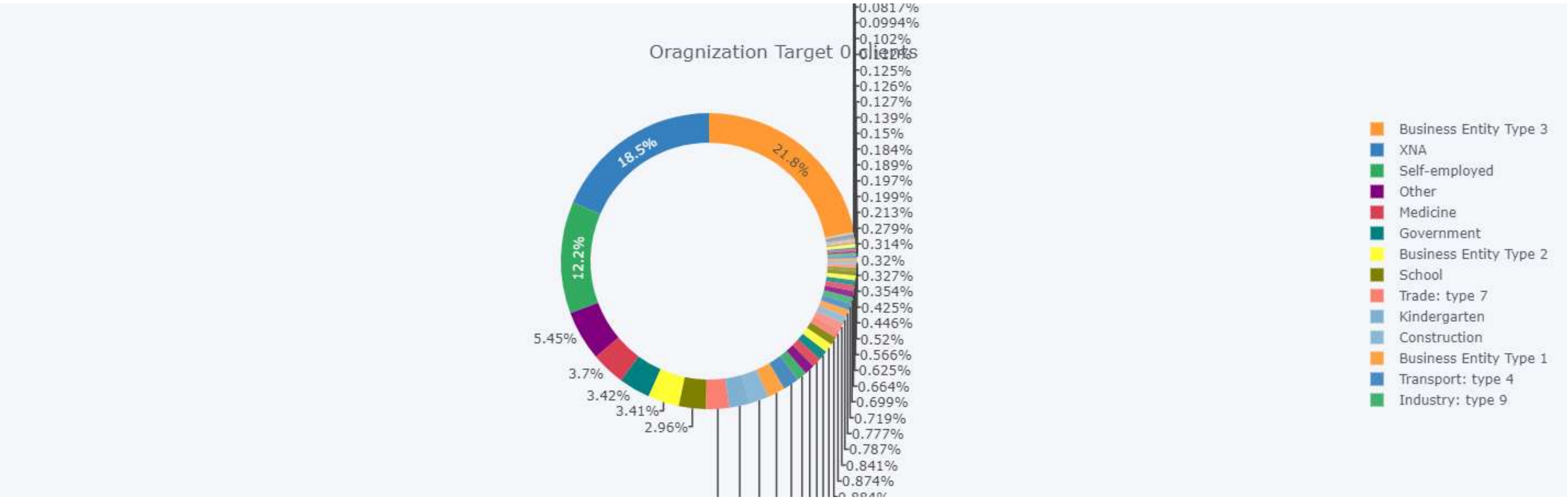
TARGET 1: OCCUPATION



TARGET 0:
GENDER

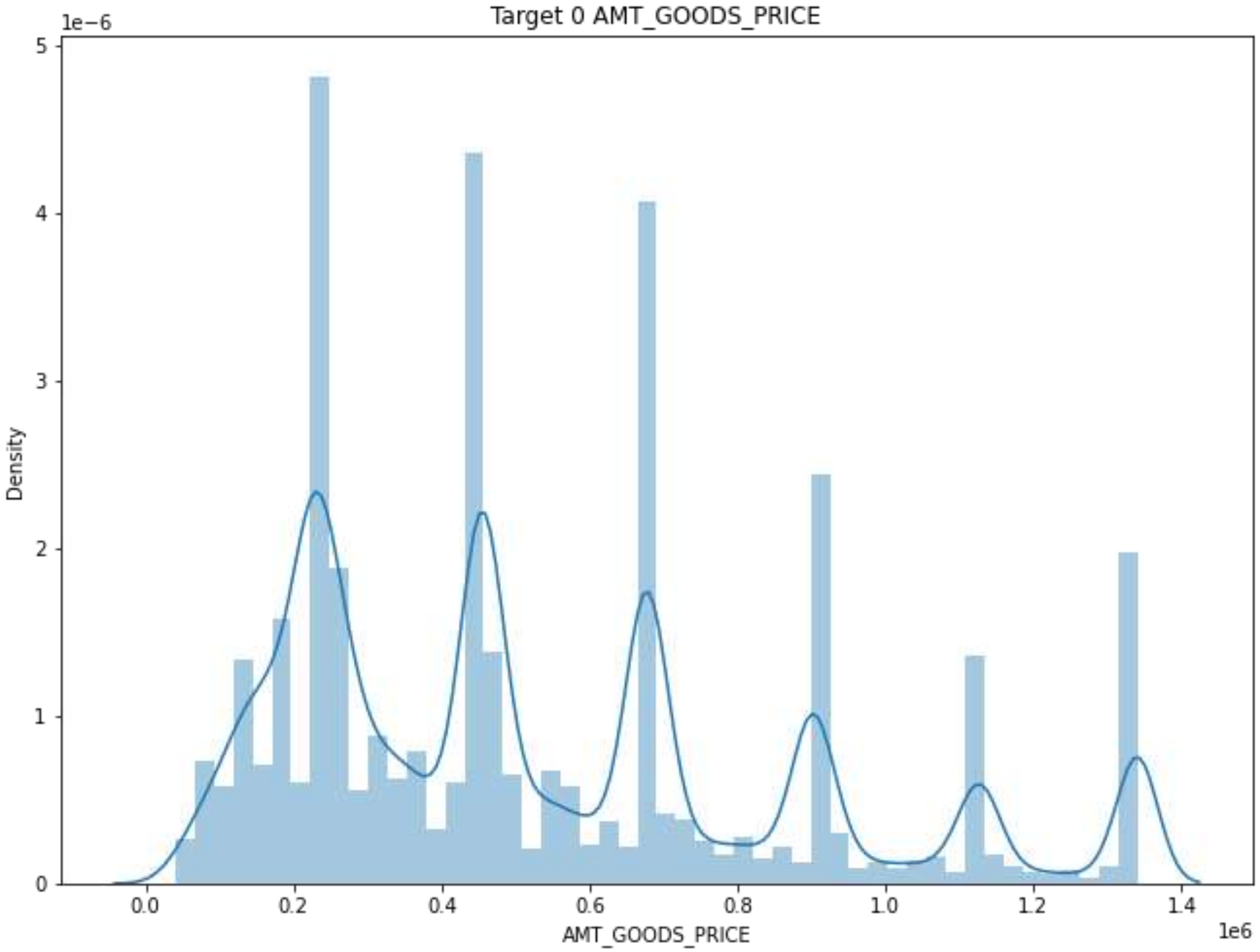


TARGET 1:
GENDER

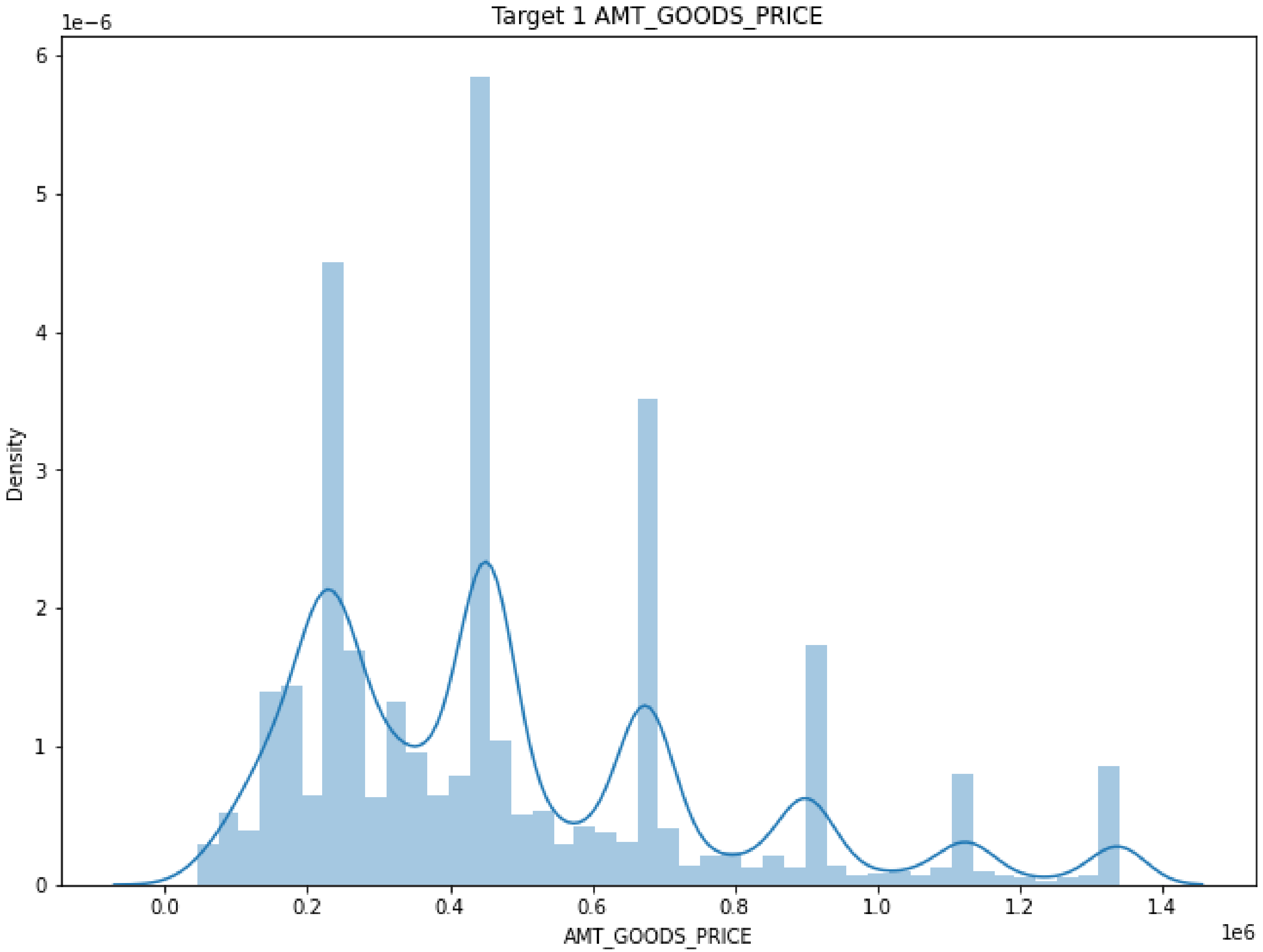


TARGET 0:
ORAGANIZATION THAT THE CLIENT WORKS IN.

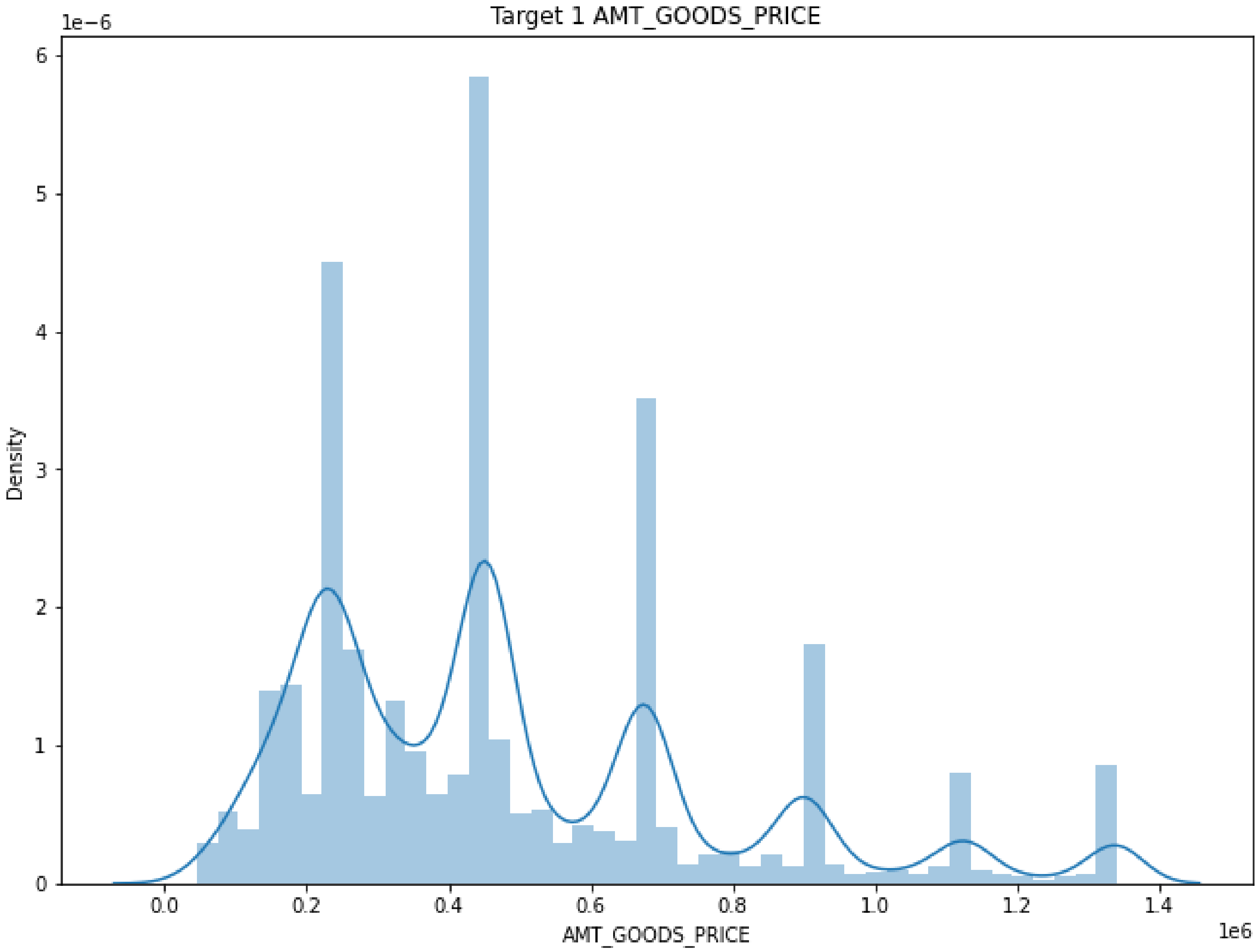
TARGET 1:
ORAGANIZATION THAT THE CLIENT WORKS IN.



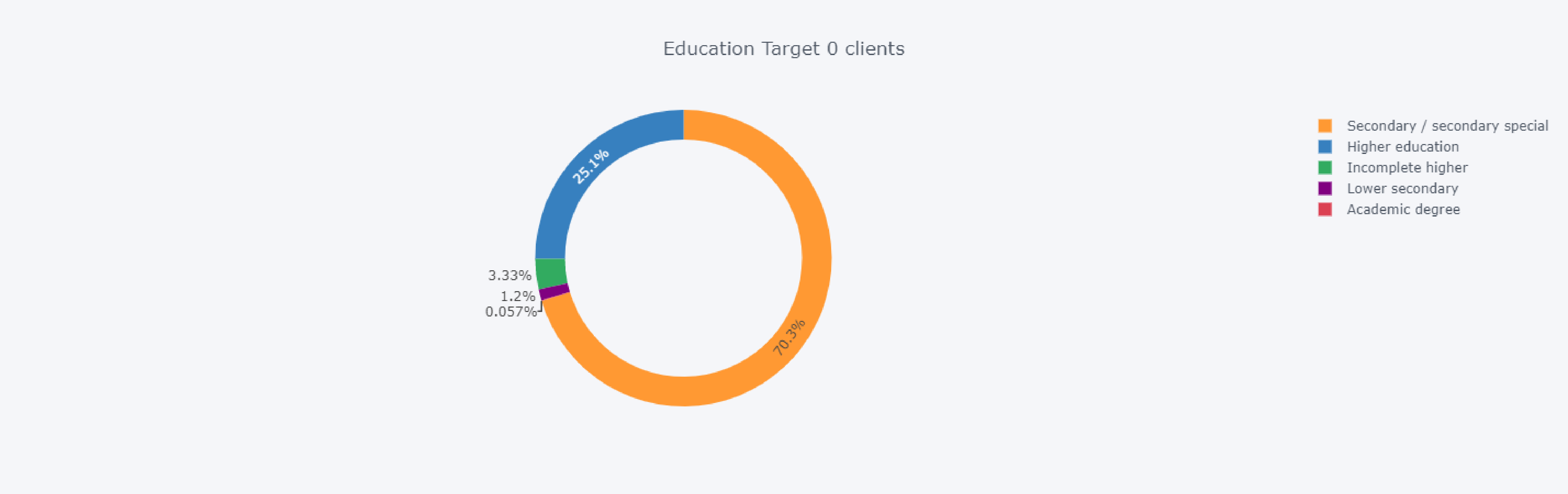
TARGET 0: GOODS PRICE



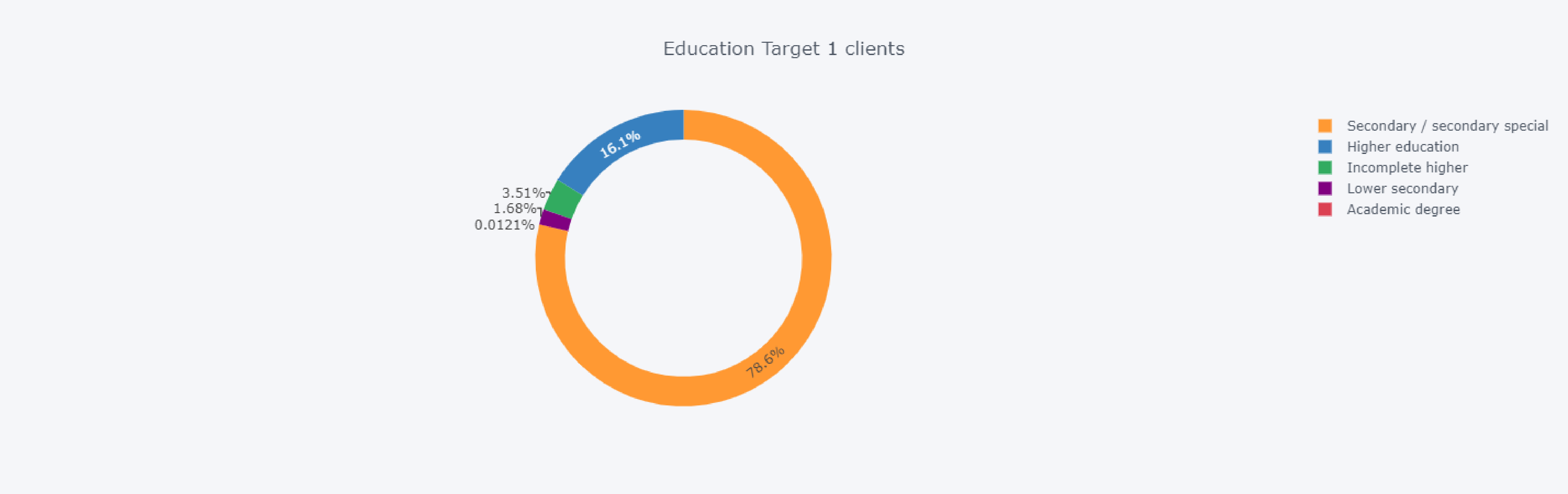
TARGET 1: GOODS PRICE



TARGET 1: GOODS PRICE

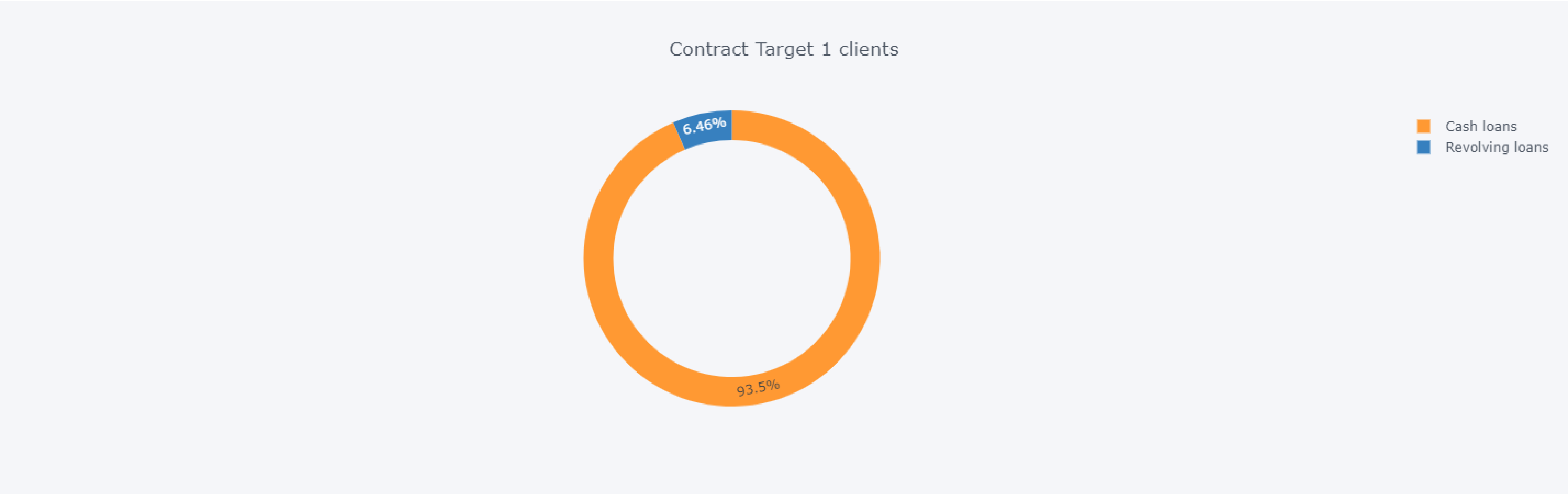


TARGET 0: EDUCATION

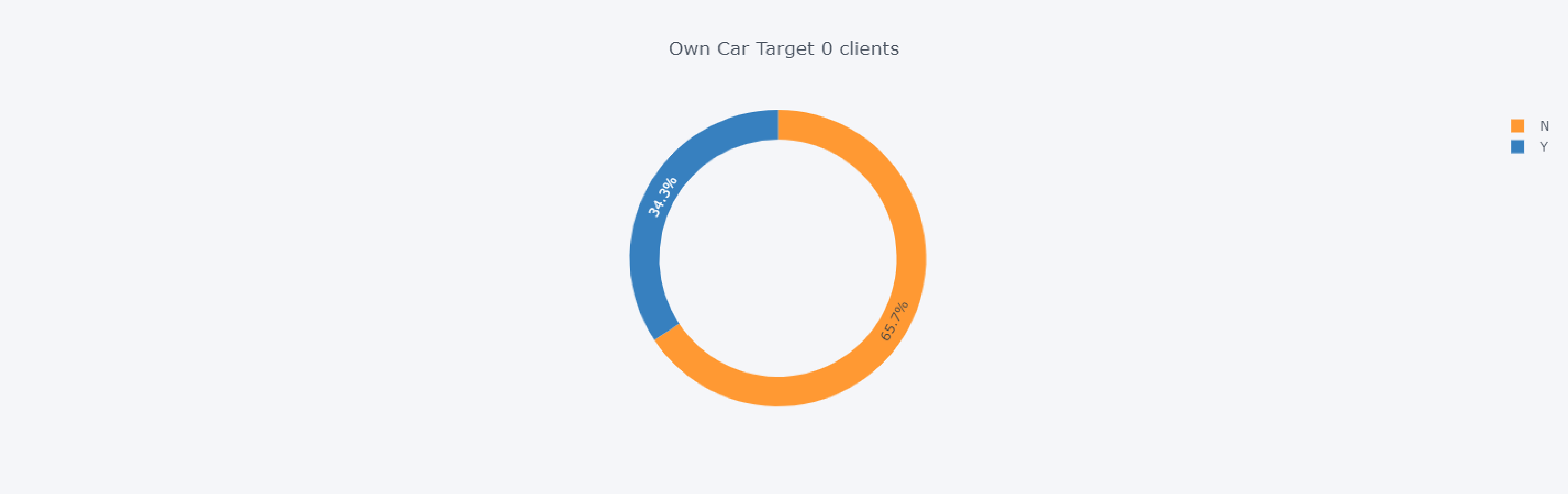


TARGET 1: EDUCATION

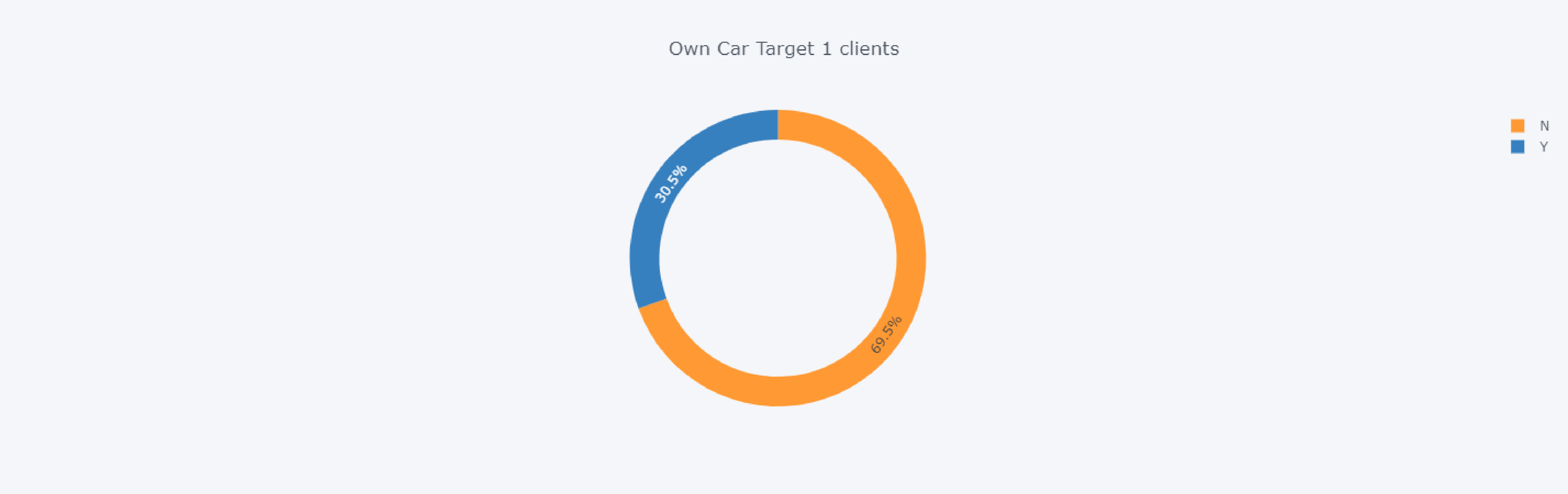
TARGET 0: CONTRACT



TARGET 1: CONTRACT

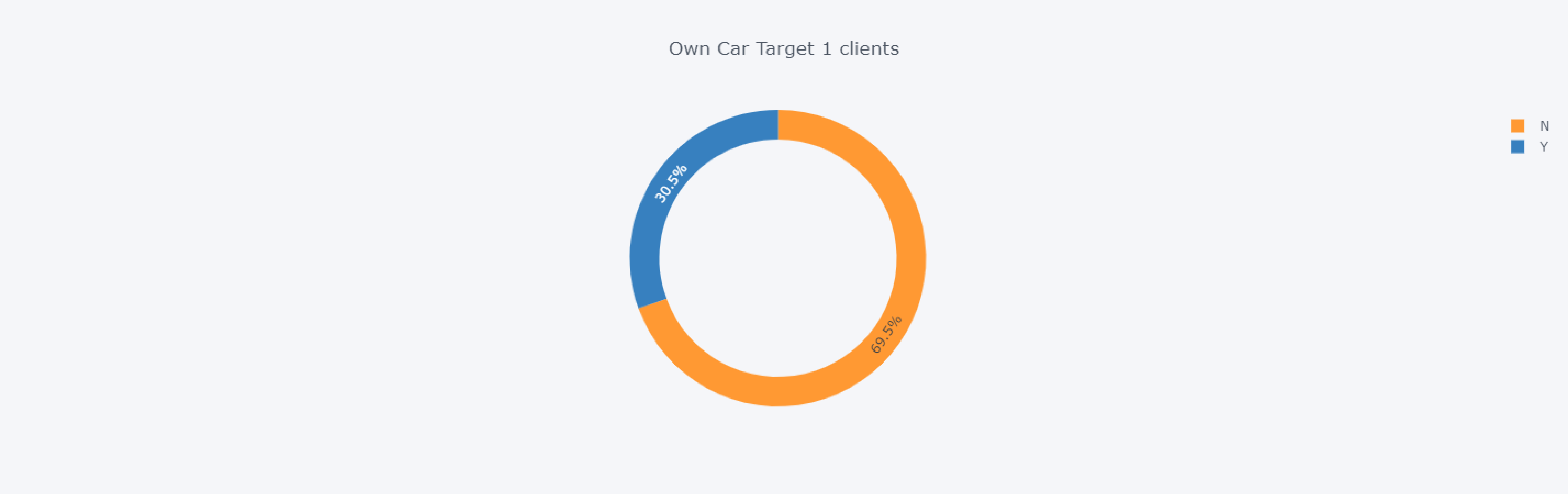


TARGET 0:
CLIENT THAT OWNS CAR

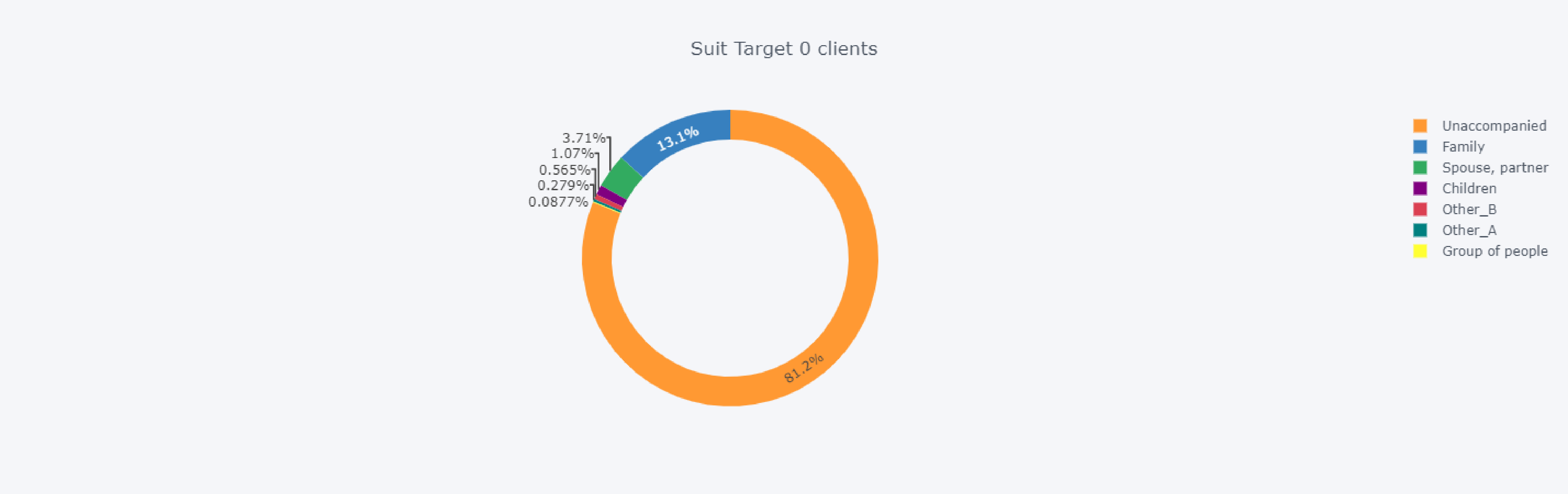


TARGET 1:
CLIENT THAT OWNS CAR

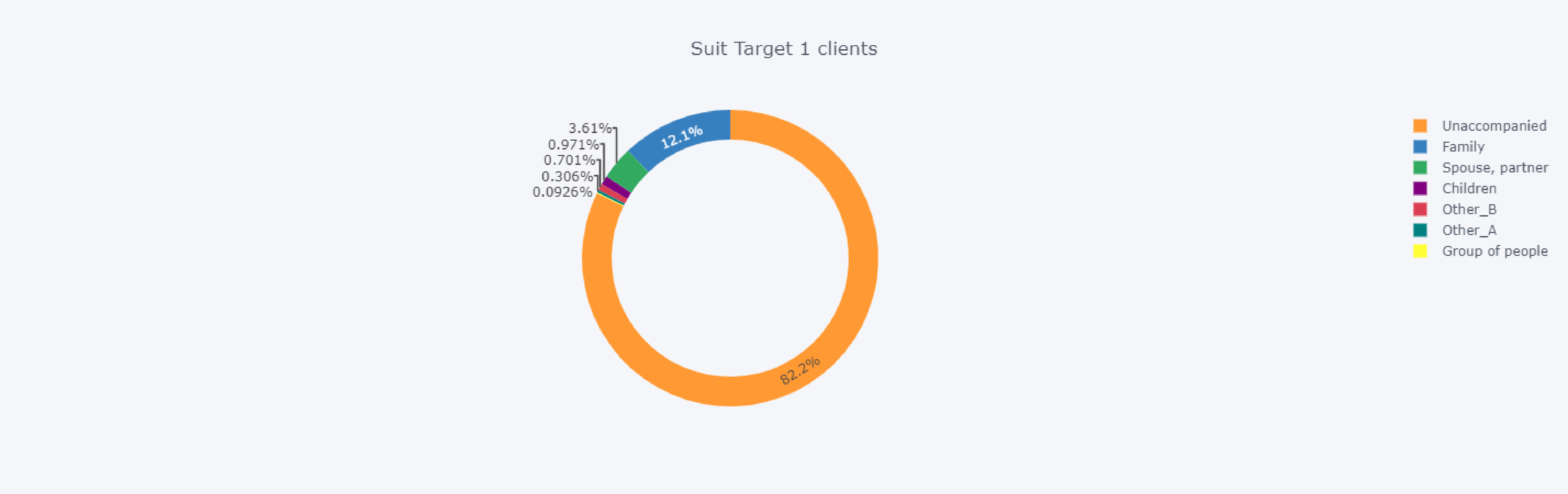
TARGET 0:
CLIENT THAT OWN REALTY OR PROPERTY



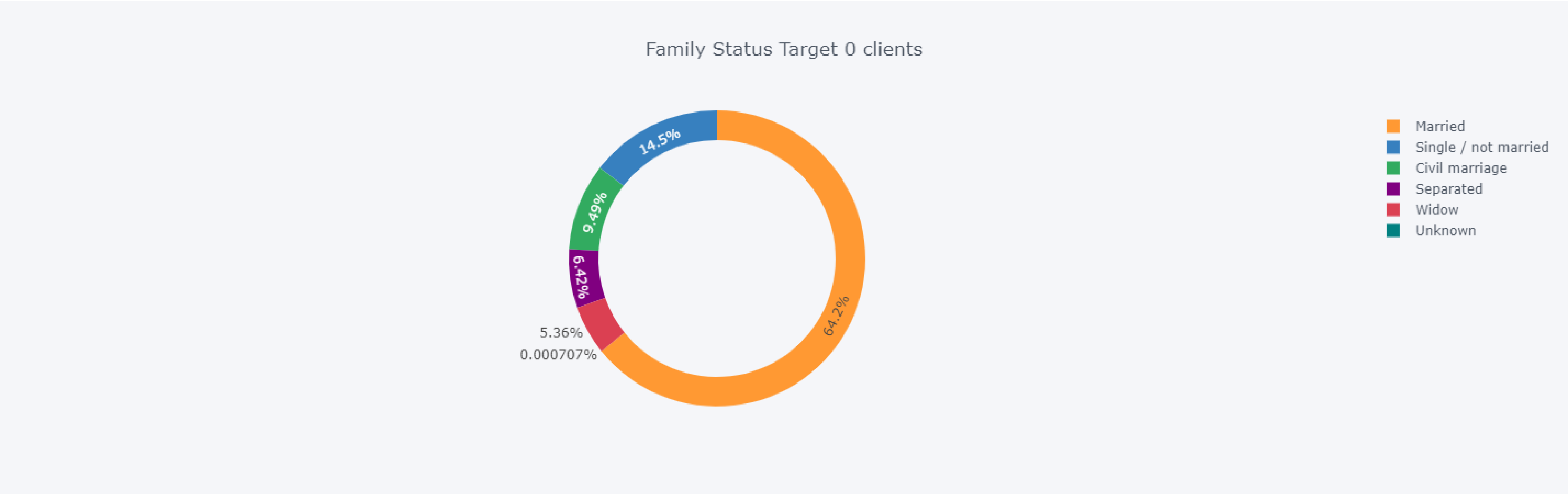
TARGET 1:
CLIENT THAT OWN REALTY OR PROPERTY



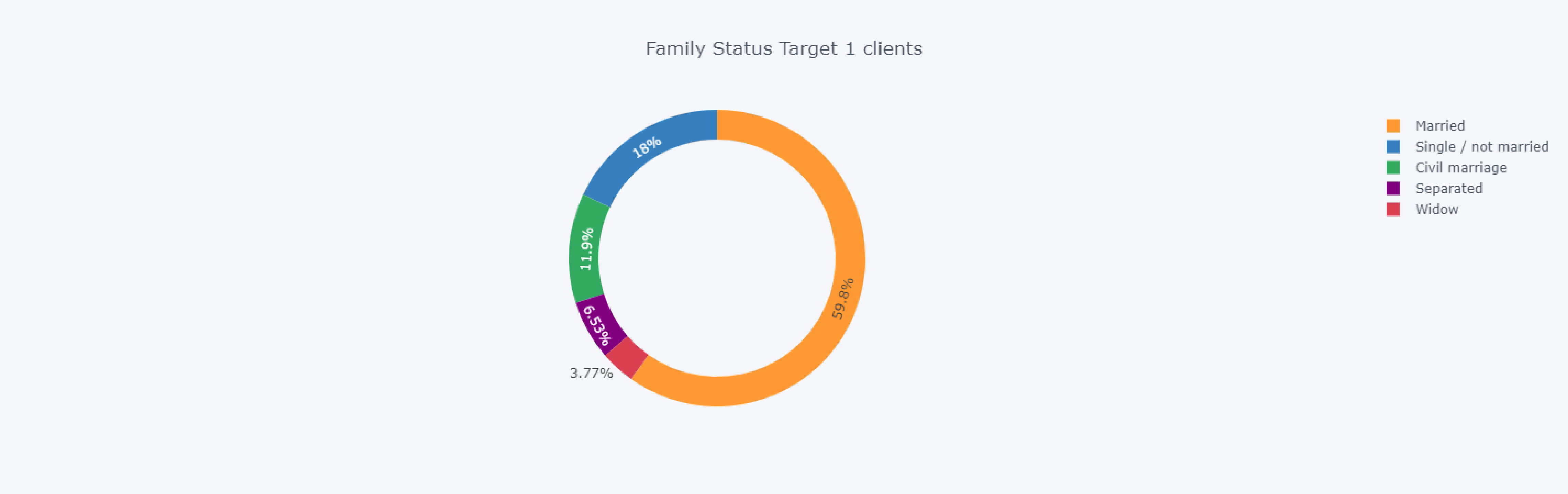
TARGET 0:
CLIENT SUIT



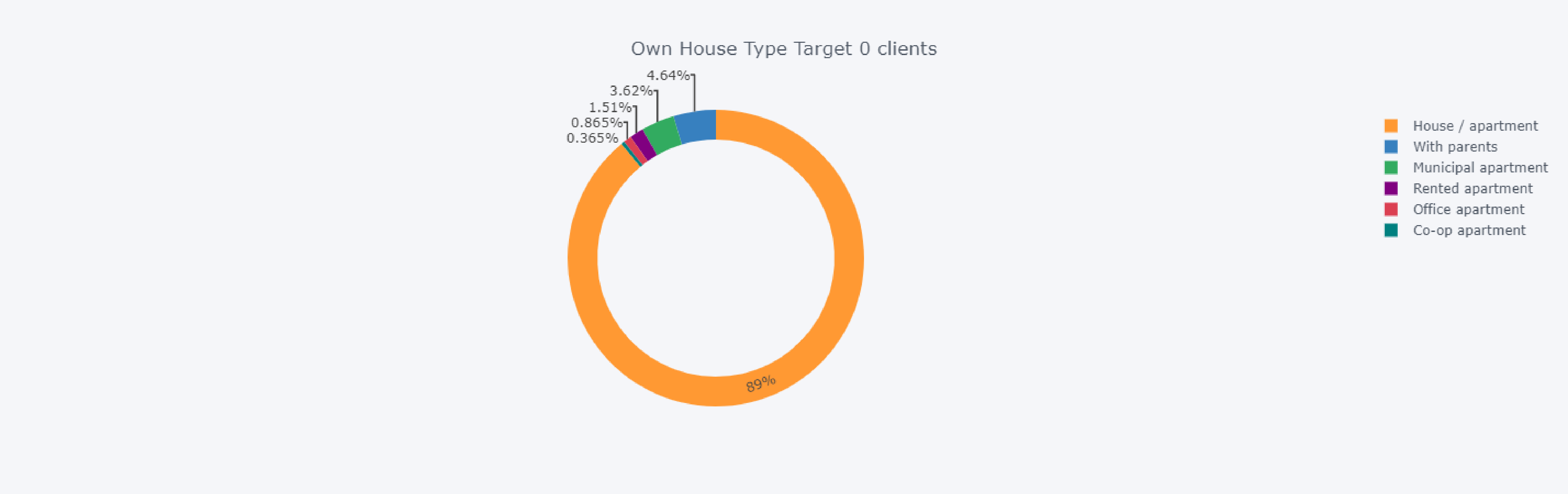
TARGET 1:
CLIENT SUIT



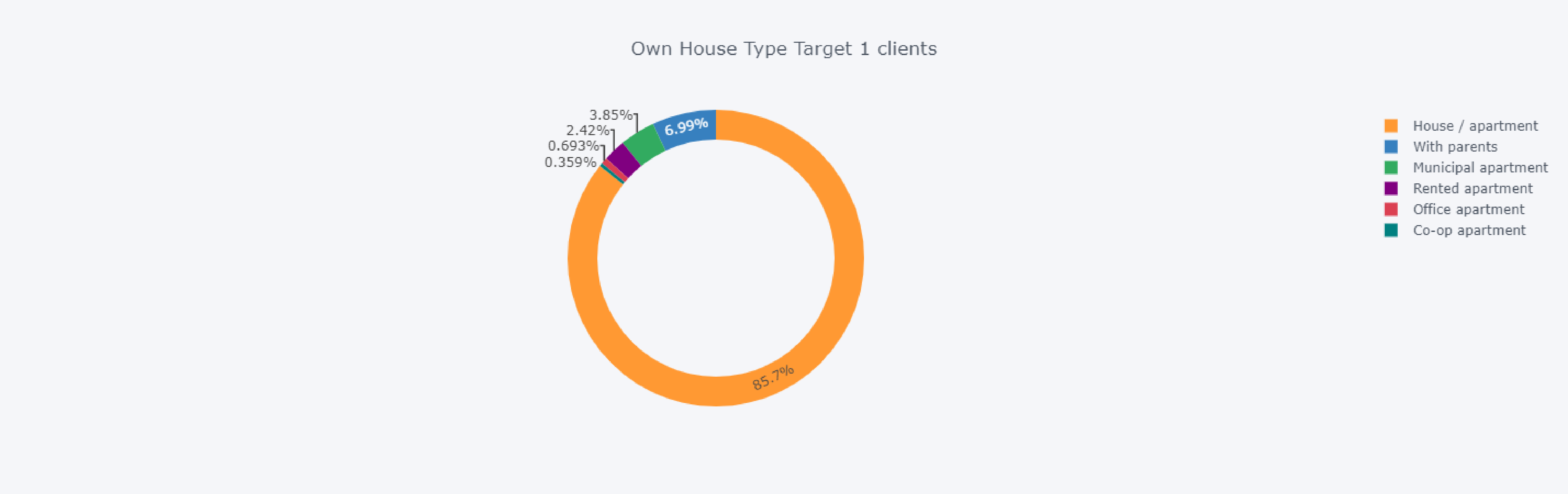
TARGET 0:
CLIENT FAMILY STATUS



TARGET 1:
CLIENT FAMILY STATUS

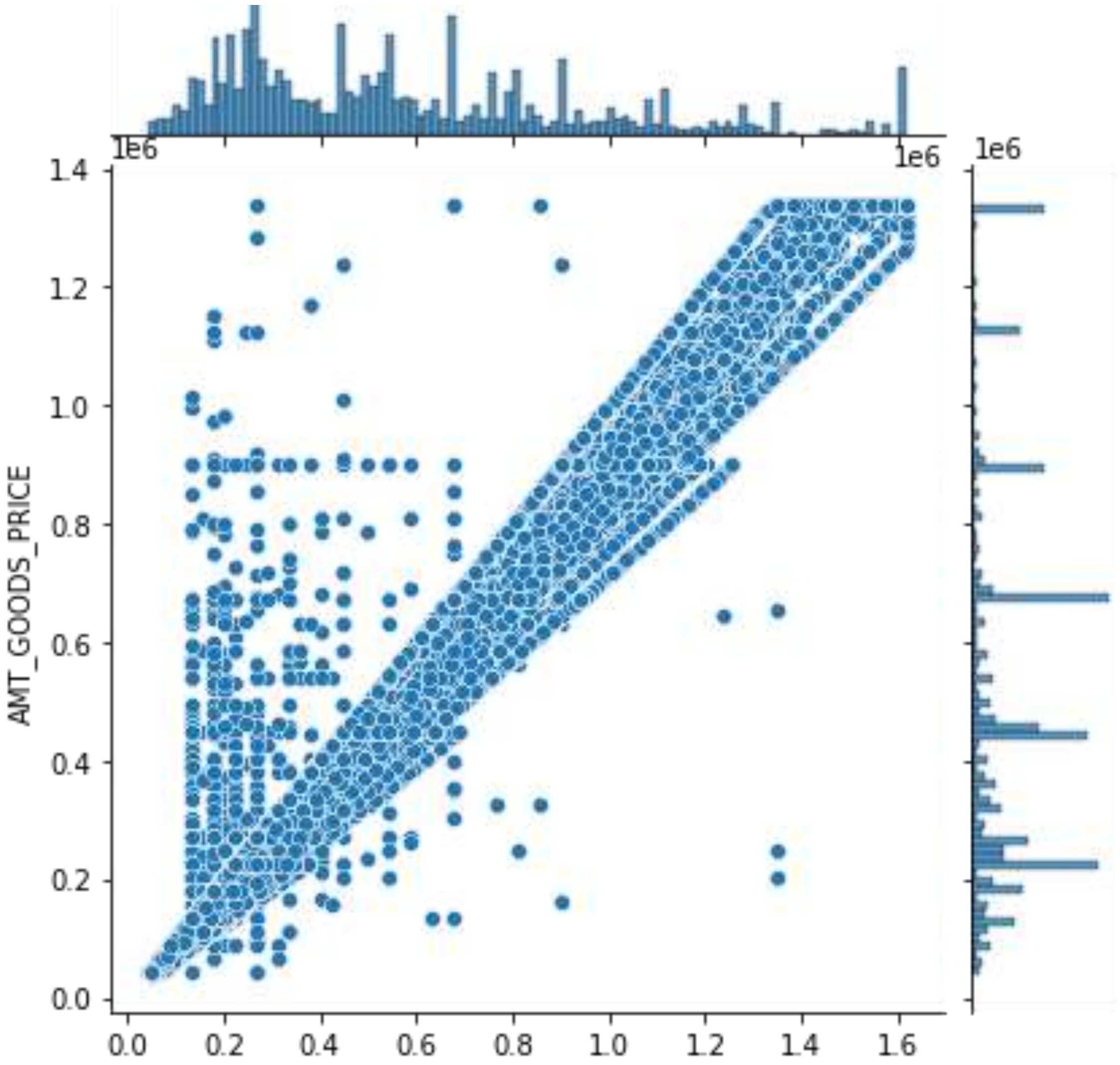


TARGET 0:
CLIENT HOUSE TYPE



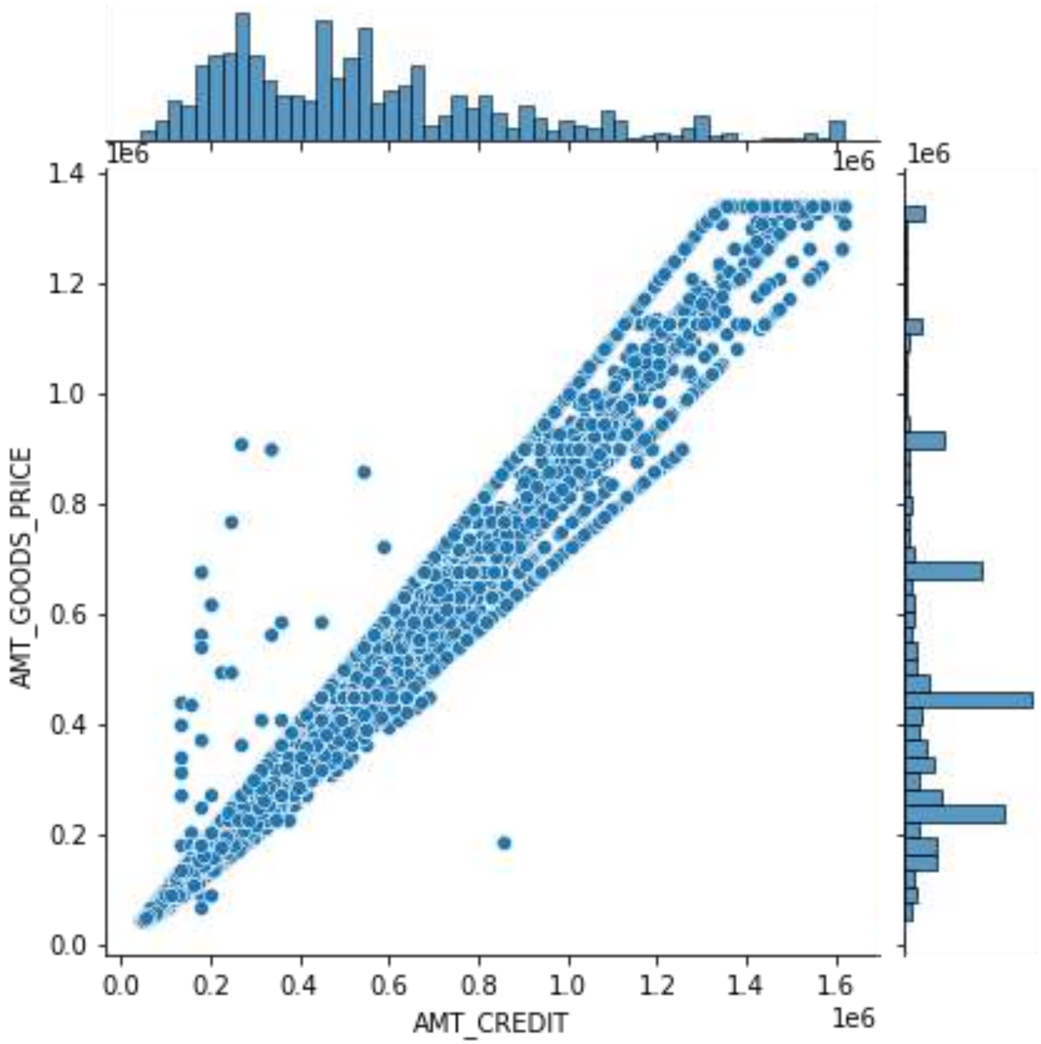
TARGET 1:
CLIENT HOUSE TYPE

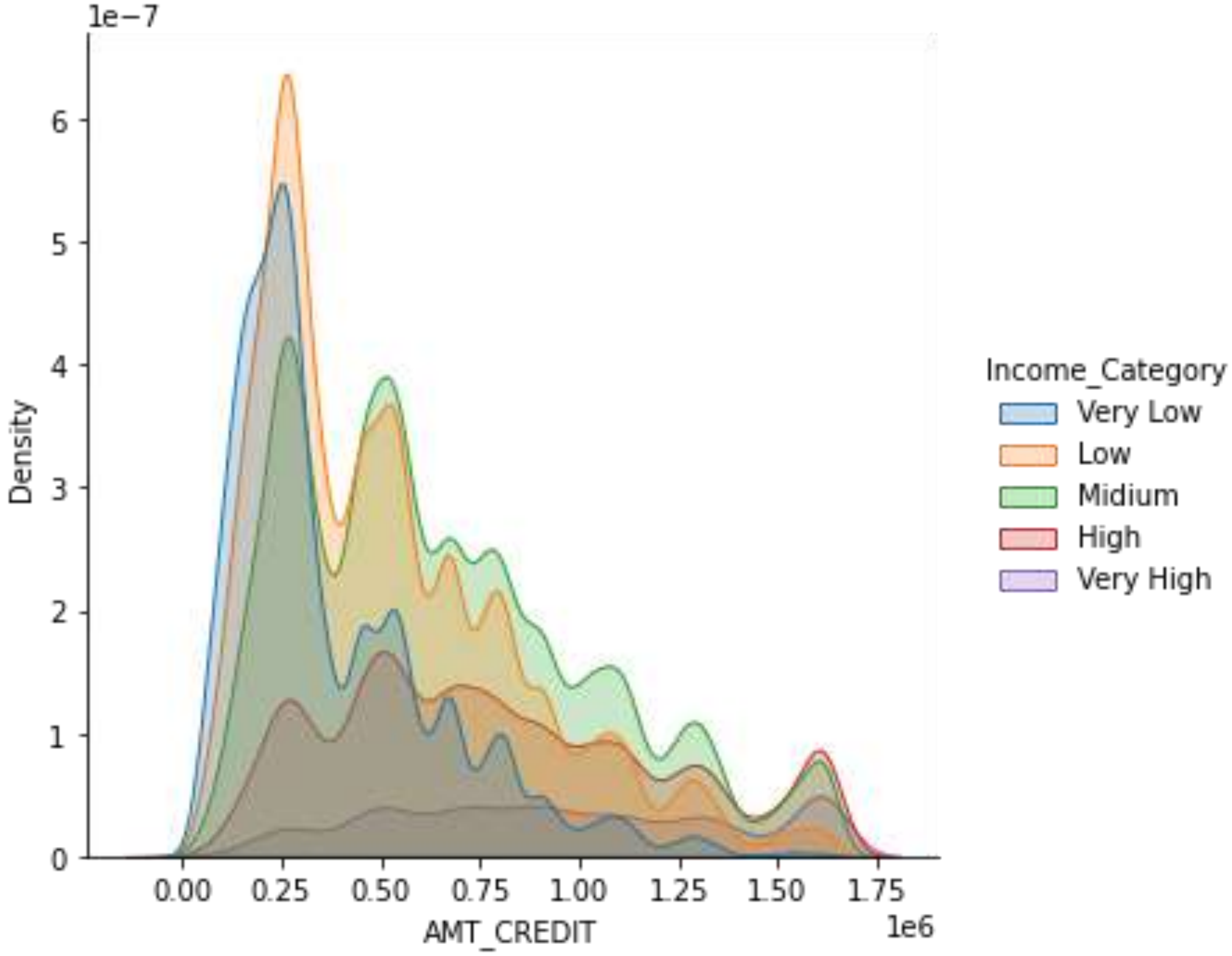
BIVARIATE AND UNIVARIATE ANALYSIS OR TARGET 0 AND TARGET 1



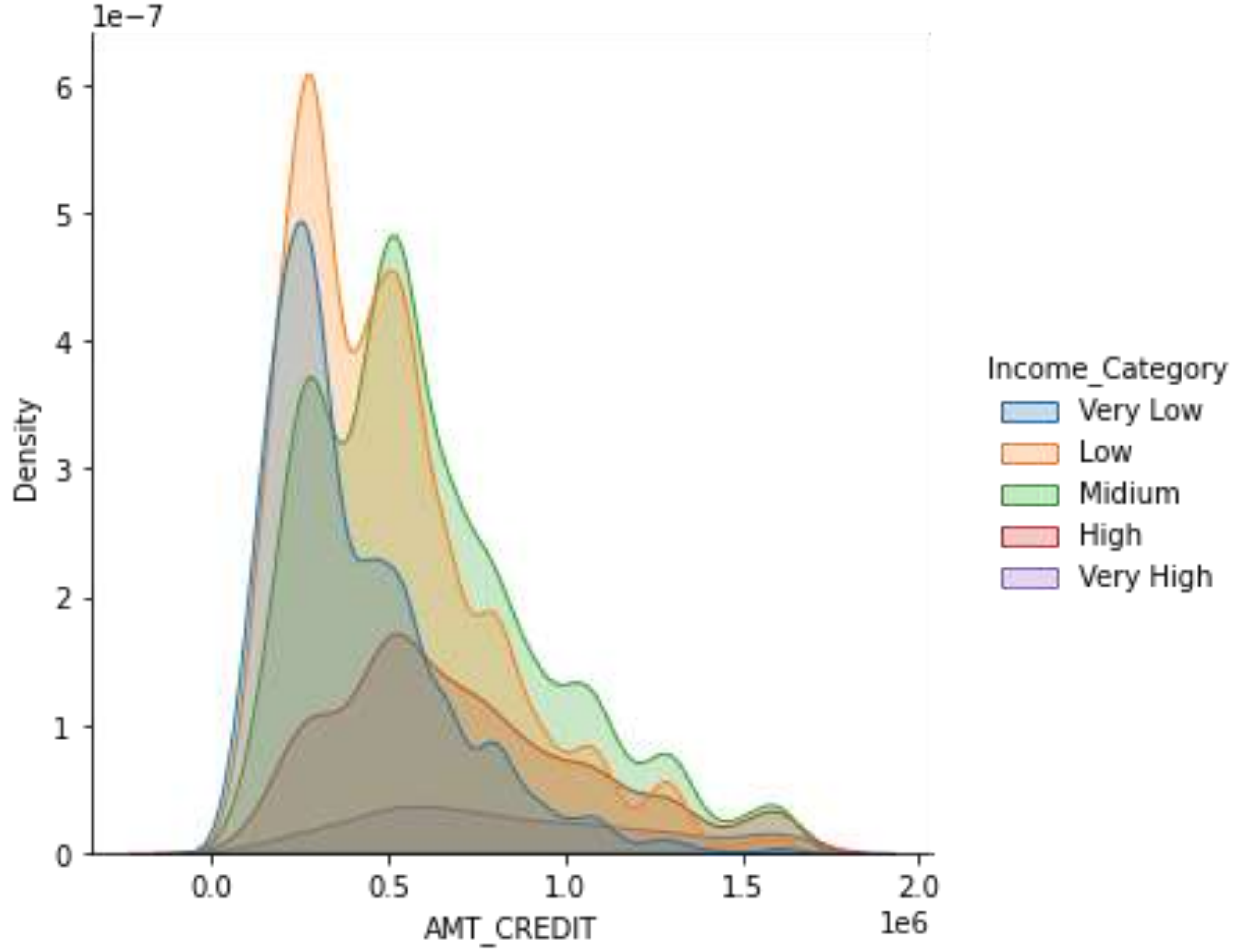
TARGET 0: CREDIT VS GOODS PRICE

TARGET 1: CREDIT VS GOODS PRICE

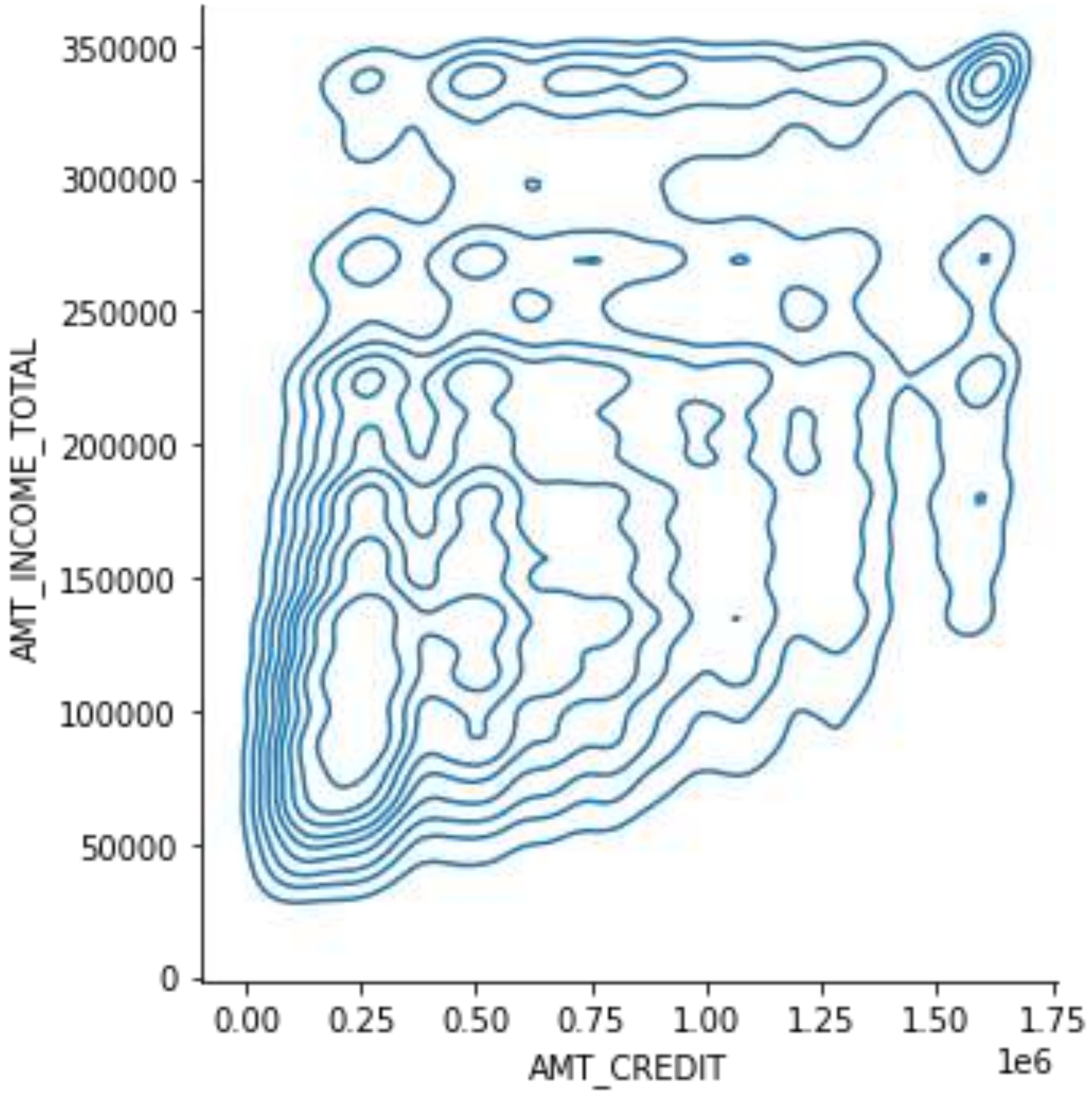




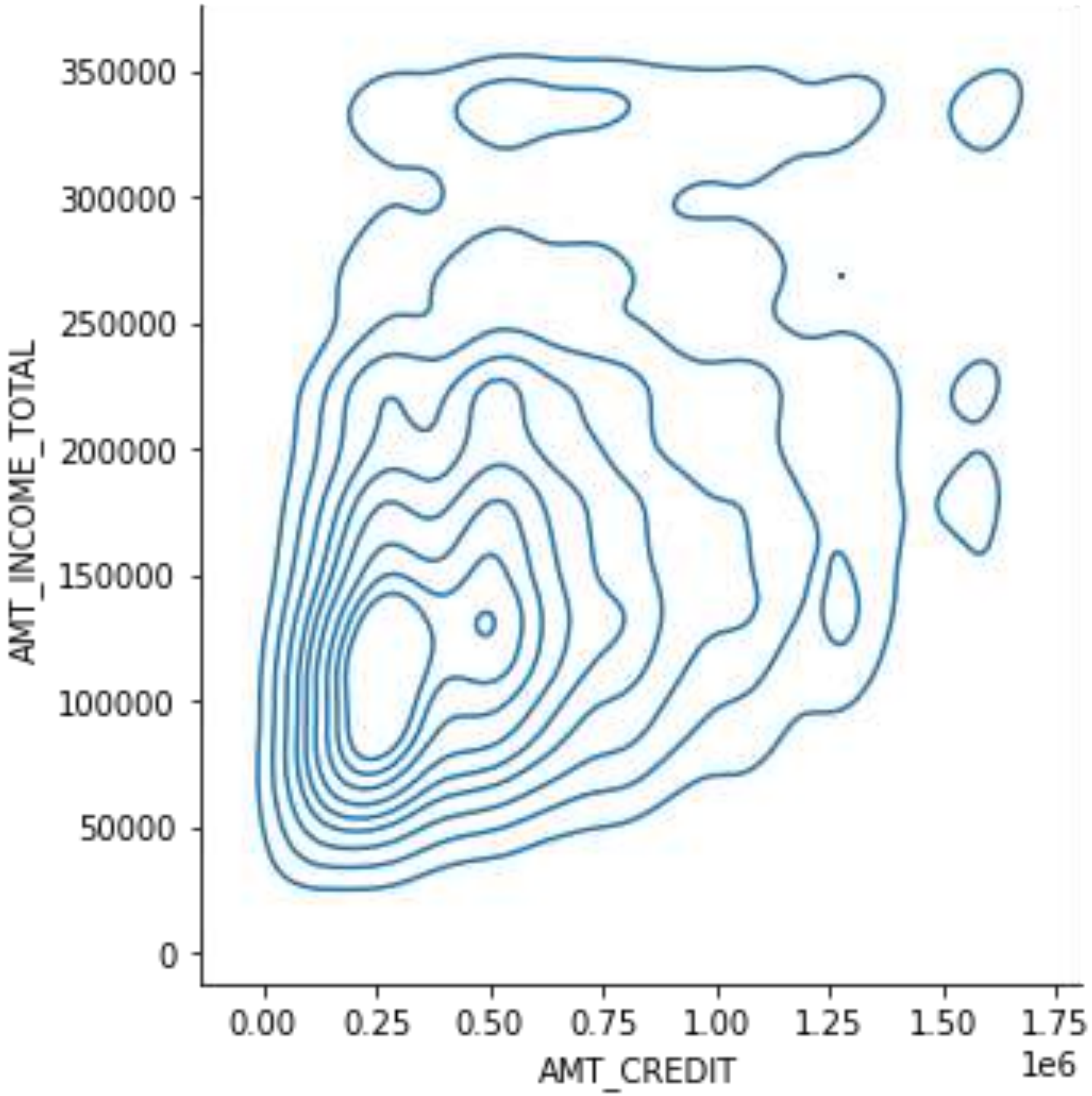
TARGET 0: CREDIT VS INCOME CATEGORY



TARGET 1: CREDIT VS INCOME CATEGORY

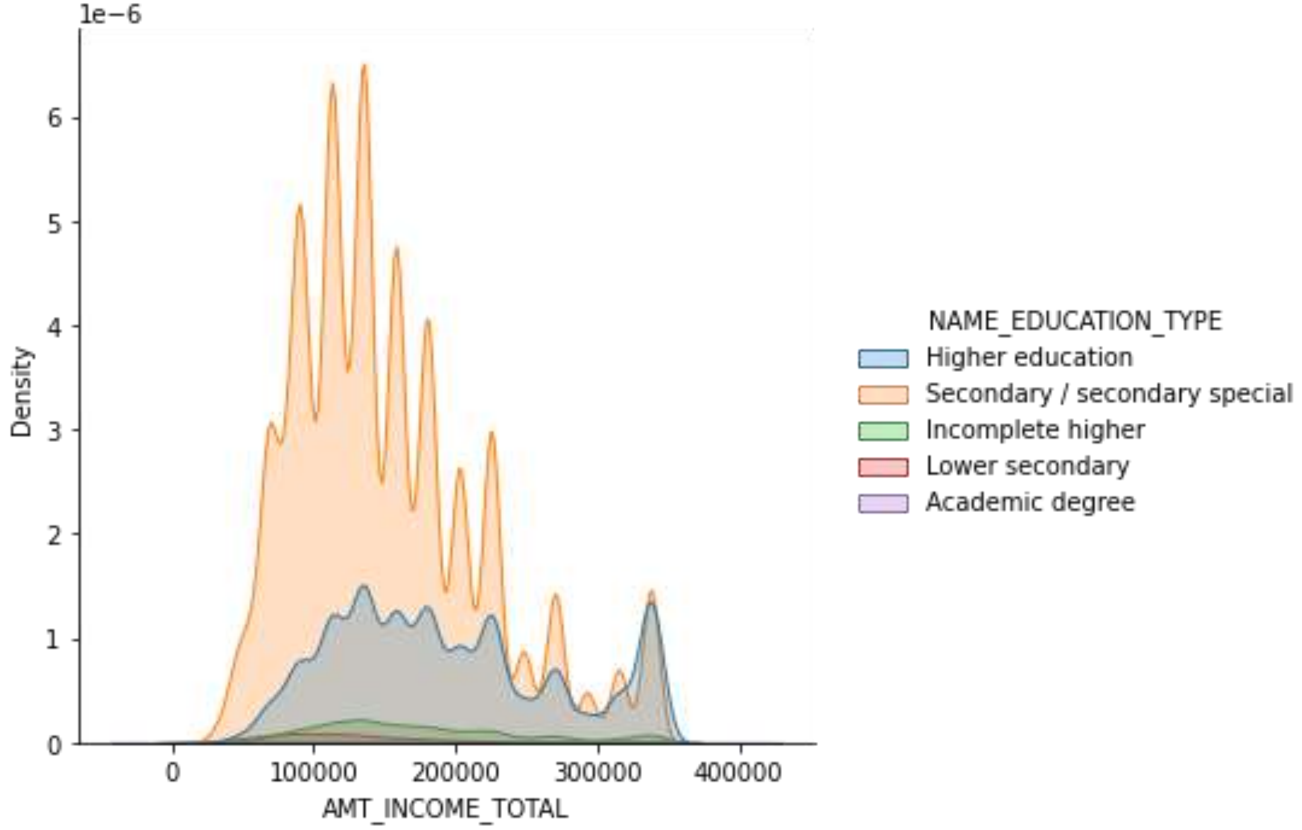


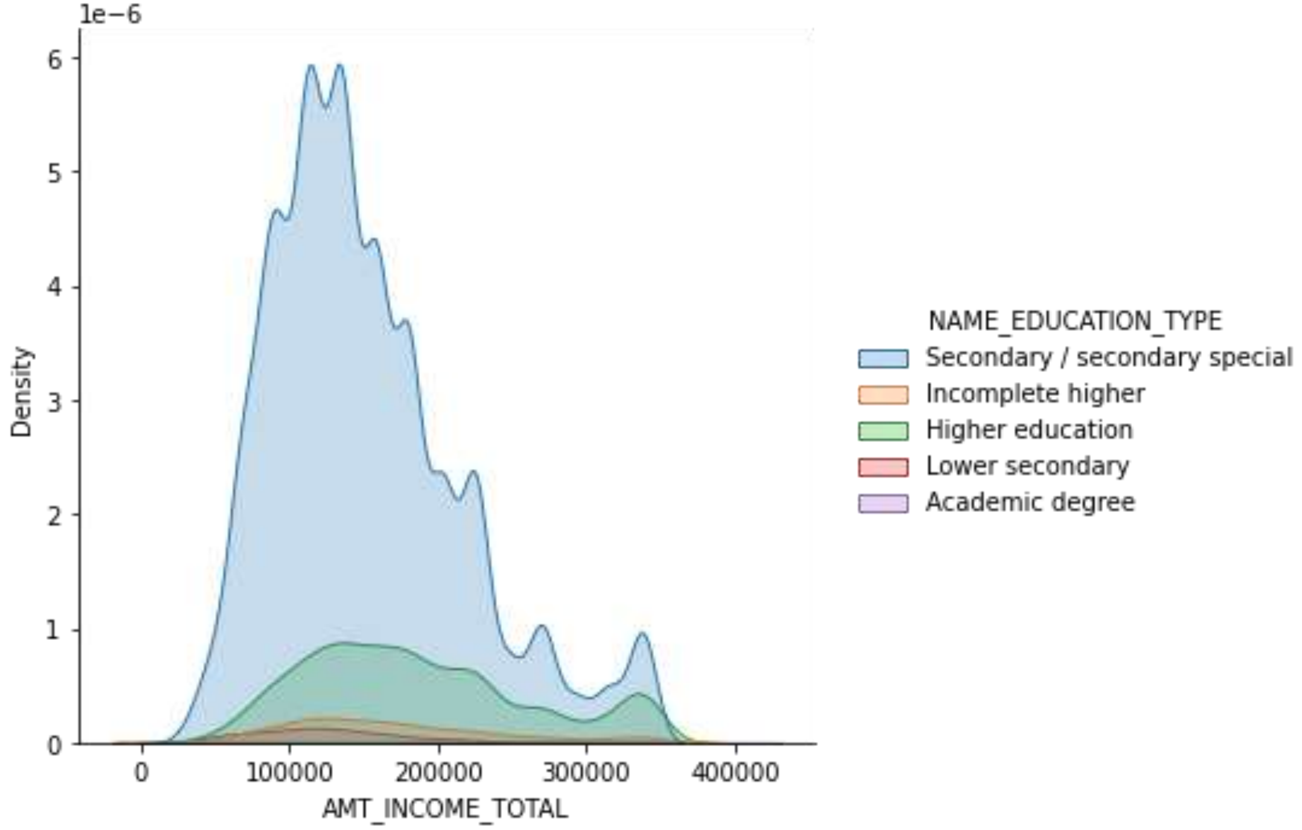
TARGET 0: CREDIT VS TOTAL INCOME OF CLIENT



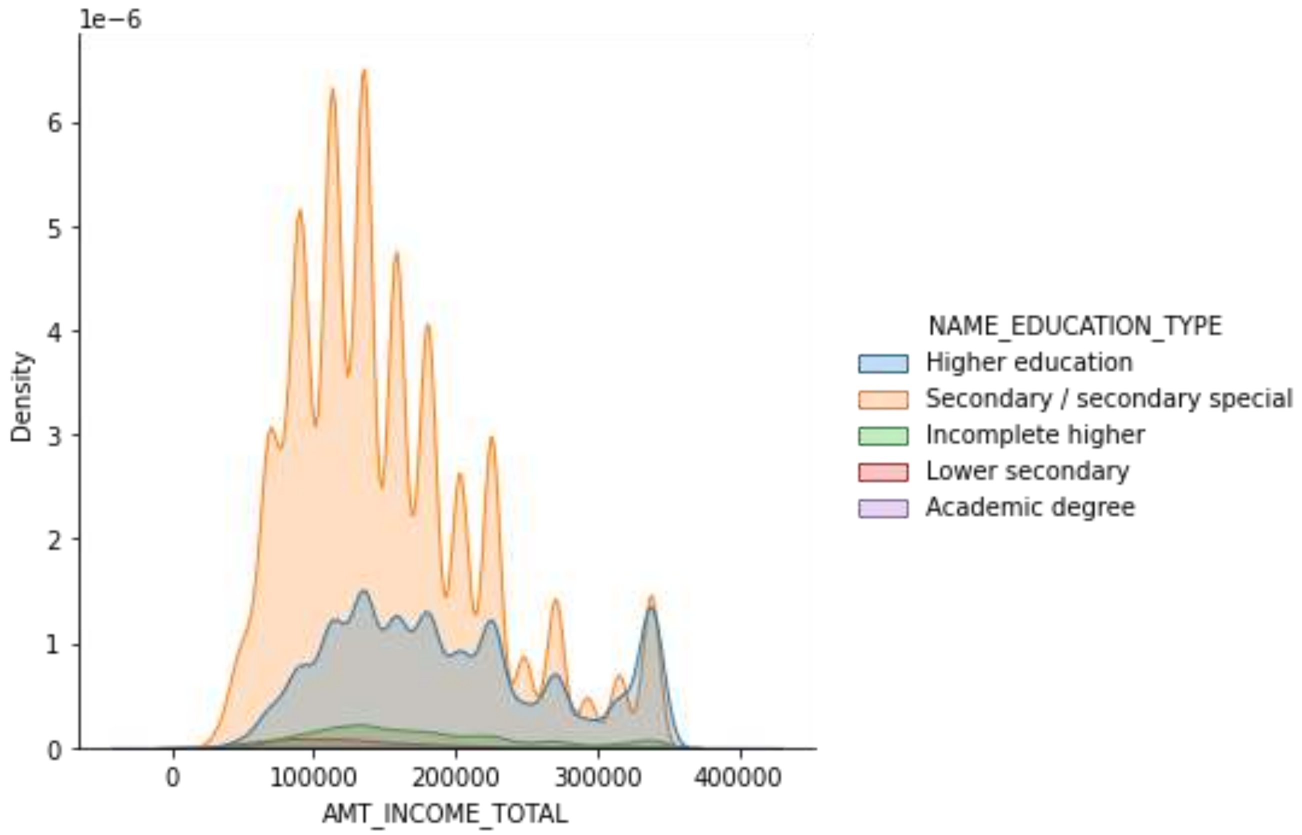
TARGET 1: CREDIT VS TOTAL INCOME OF CLIENT

TARGET 0: INCOME VS EDUCATION

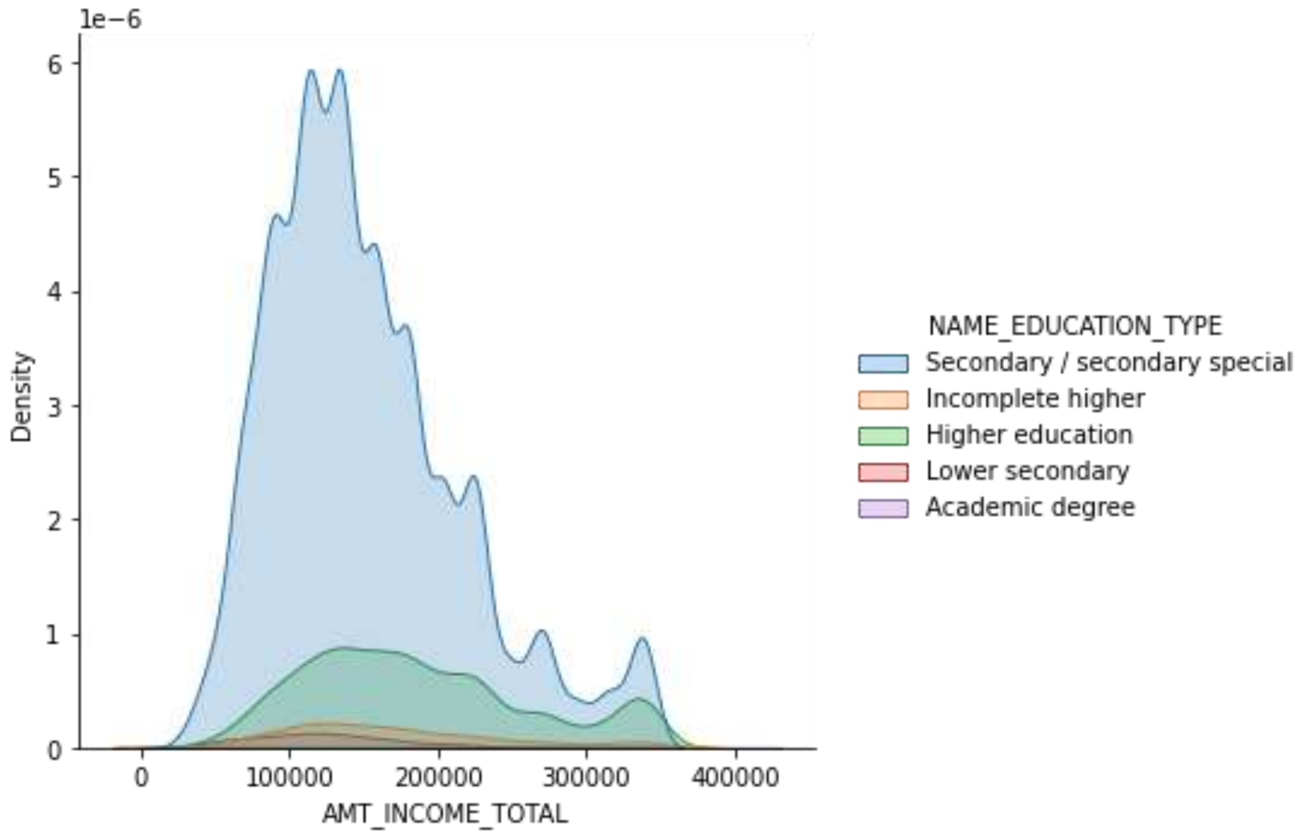




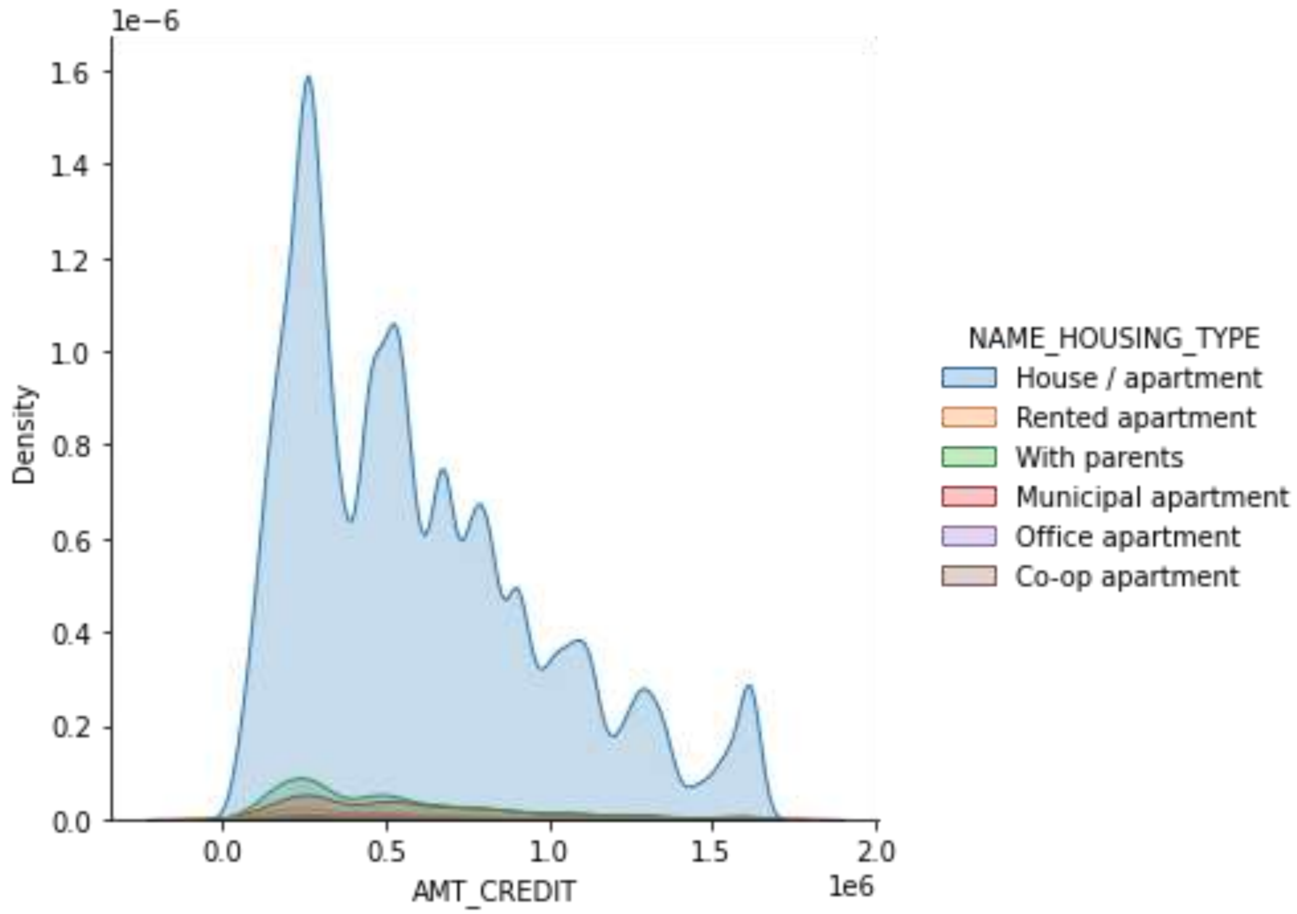
TARGET 1: INCOME VS EDUCATION



TARGET 0: CREDIT VS EDUCATION

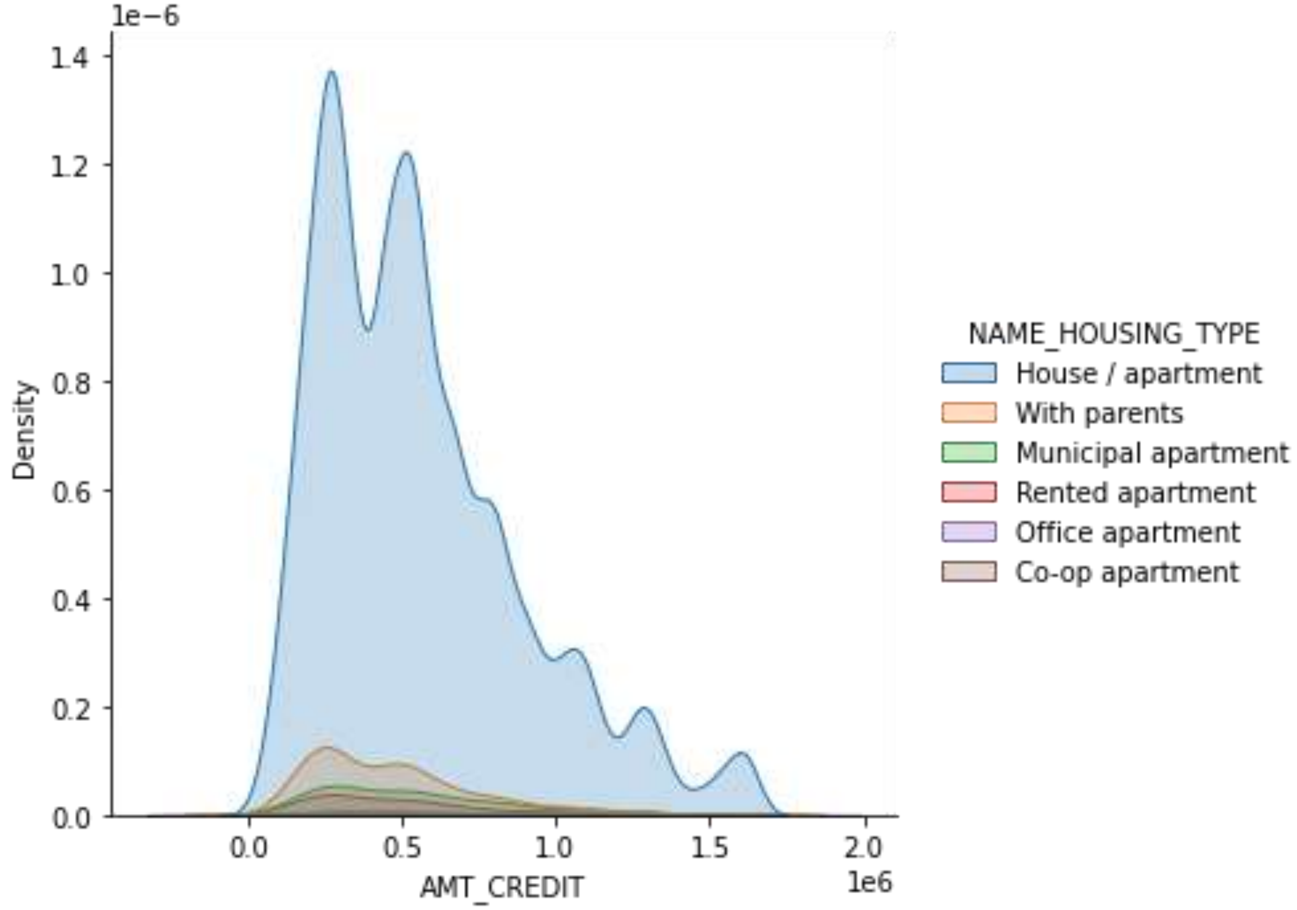


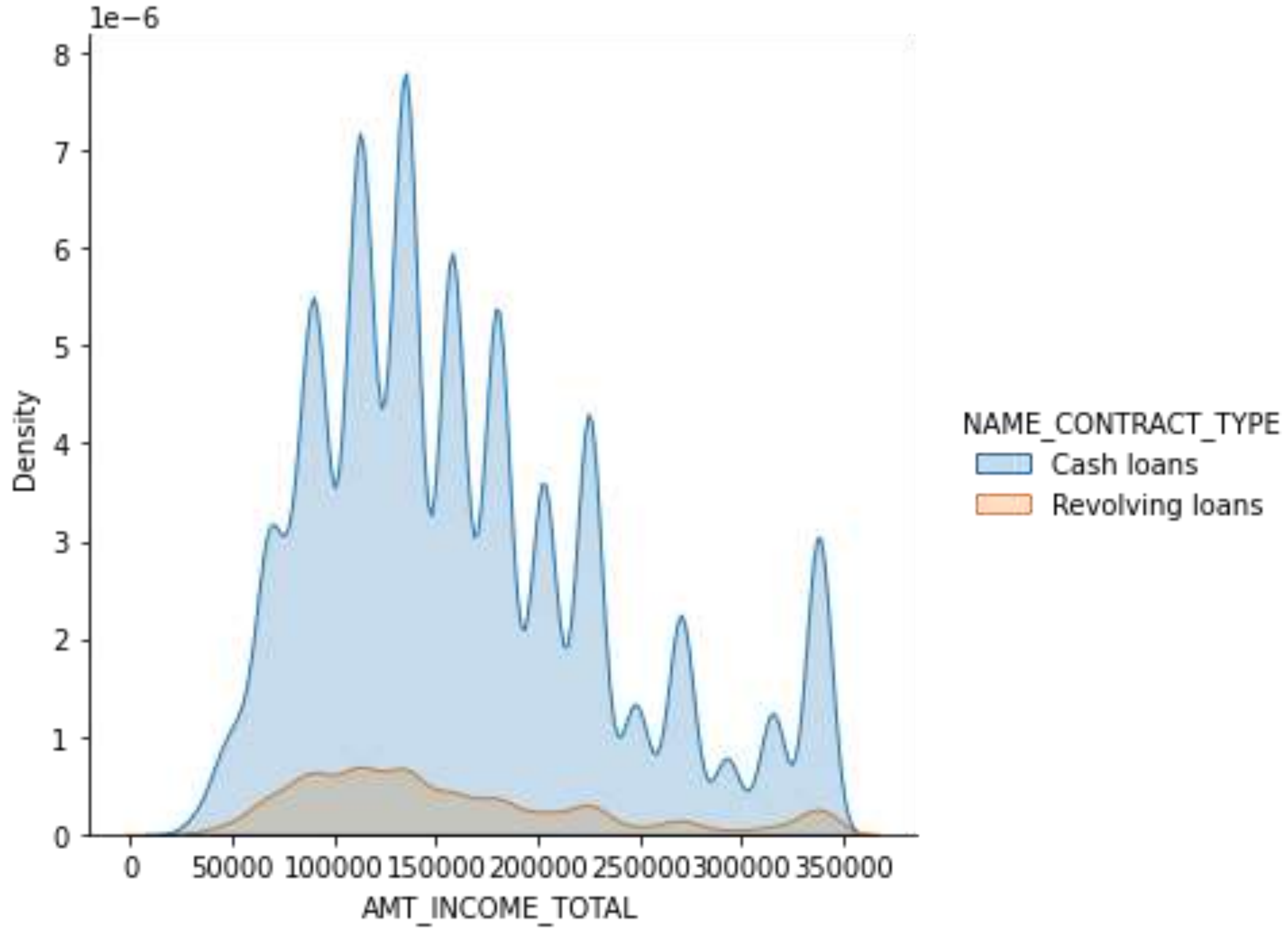
TARGET 1: CREDIT VS EDUCATION



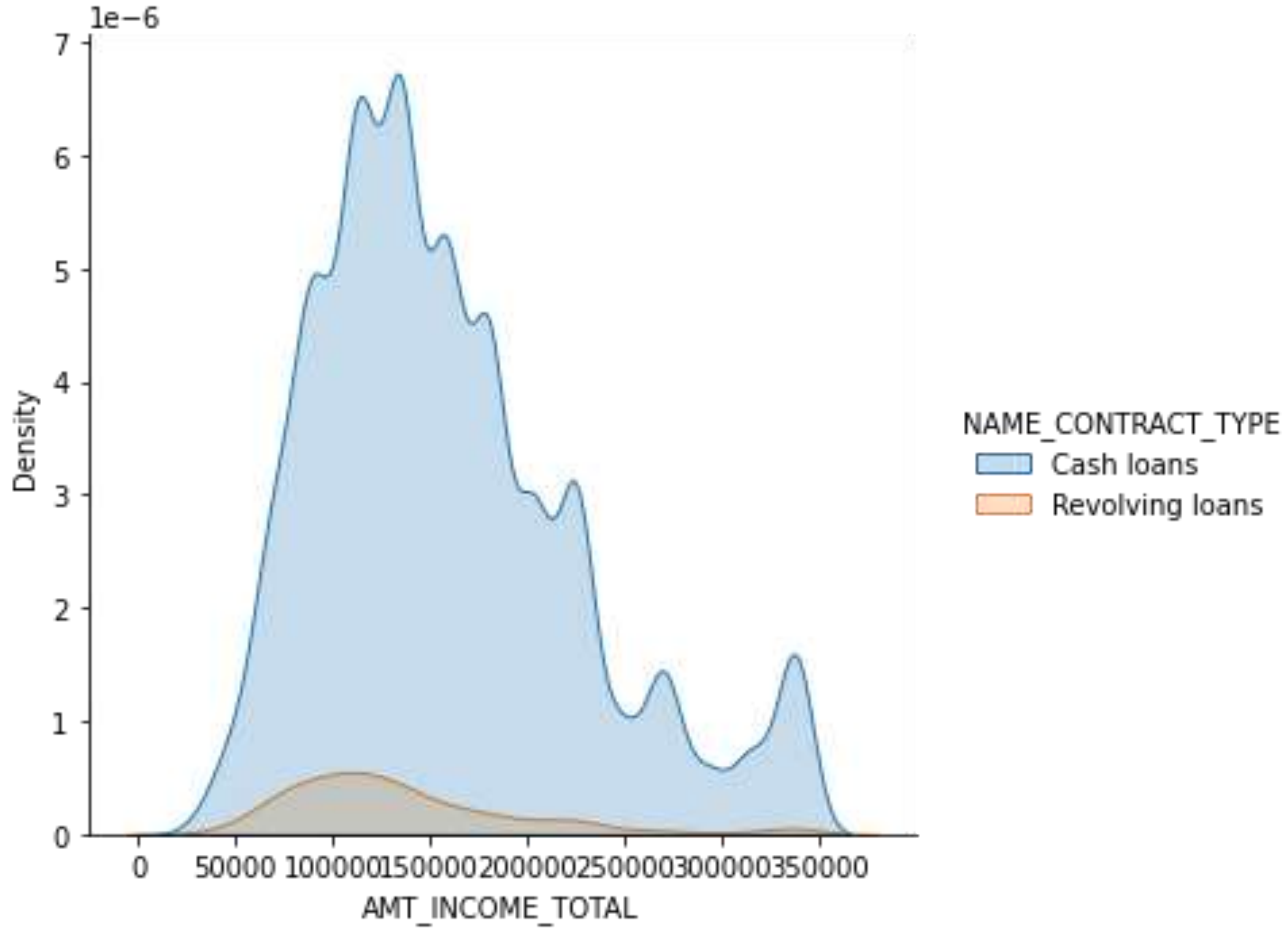
TARGET 0: CREDIT VS HOUSING

TARGET 1: CREDIT VS HOUSING

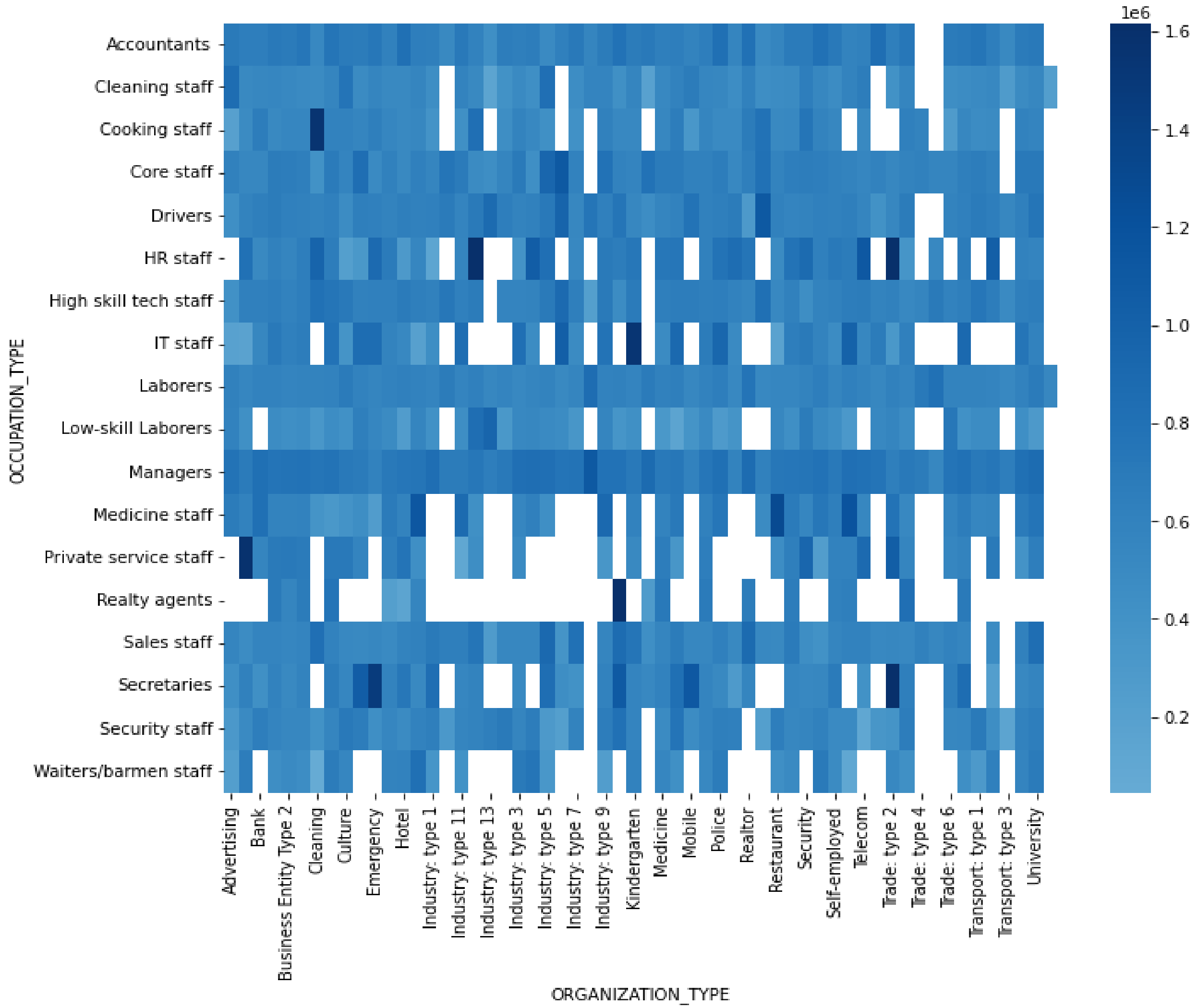




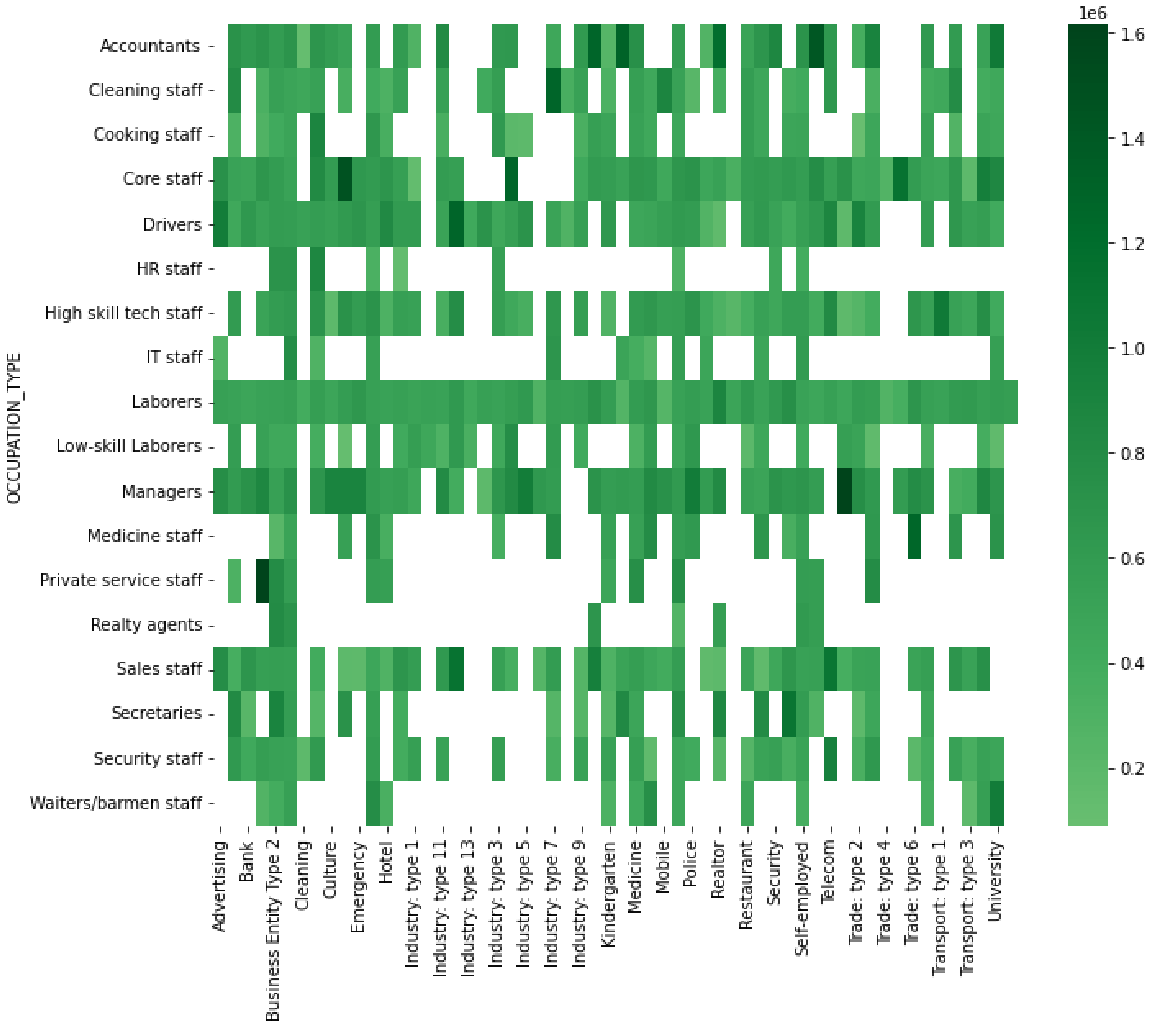
TARGET 0: INCOME VS CONTRACT



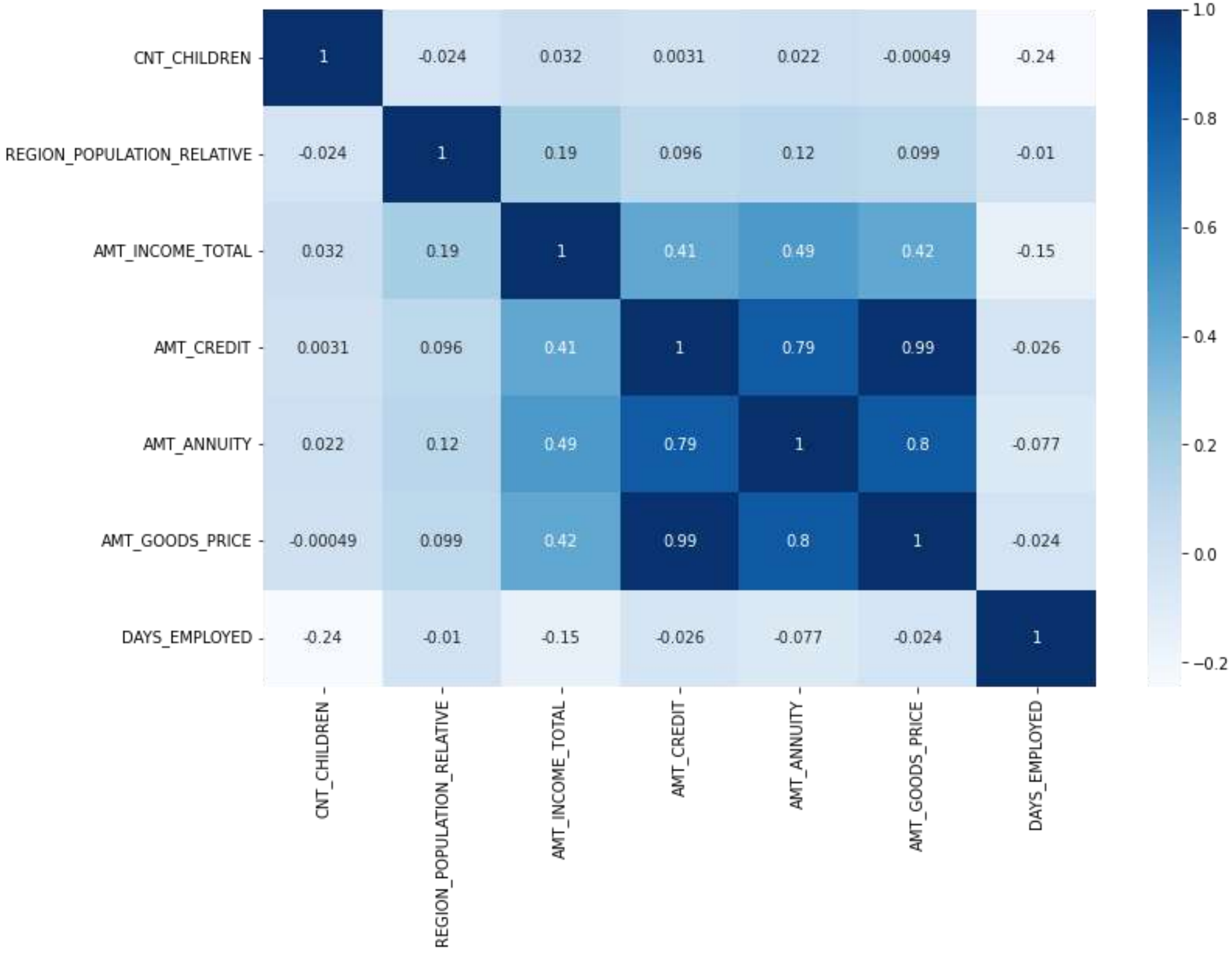
TARGET 1: INCOME VS CONTRACT



TARGET 0: OCCUPATION VS ORGANISATION VS CREDIT

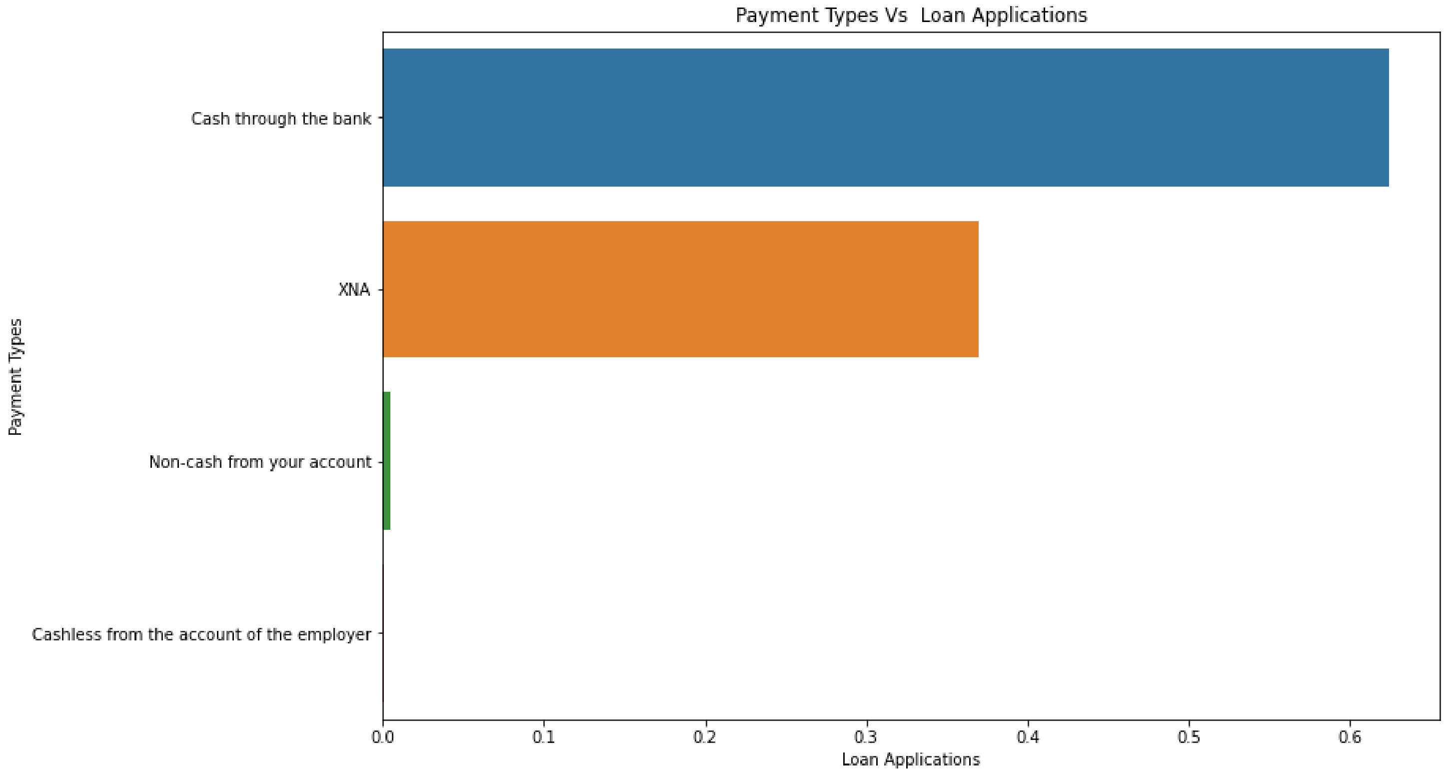


TARGET 1: OCCUPATION VS ORGANISATION VS CREDIT

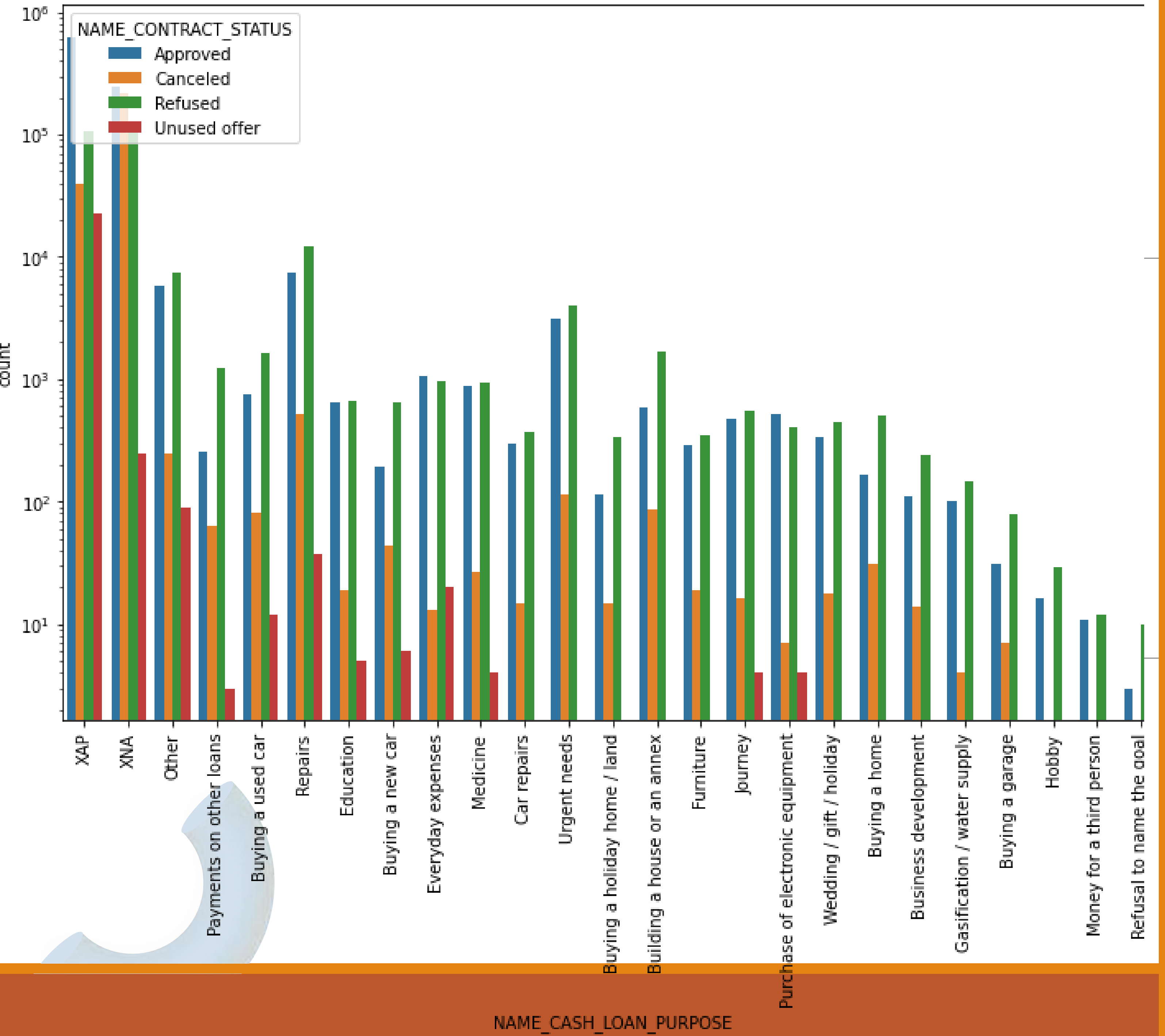


TARGET 0: MULTIVARIATE

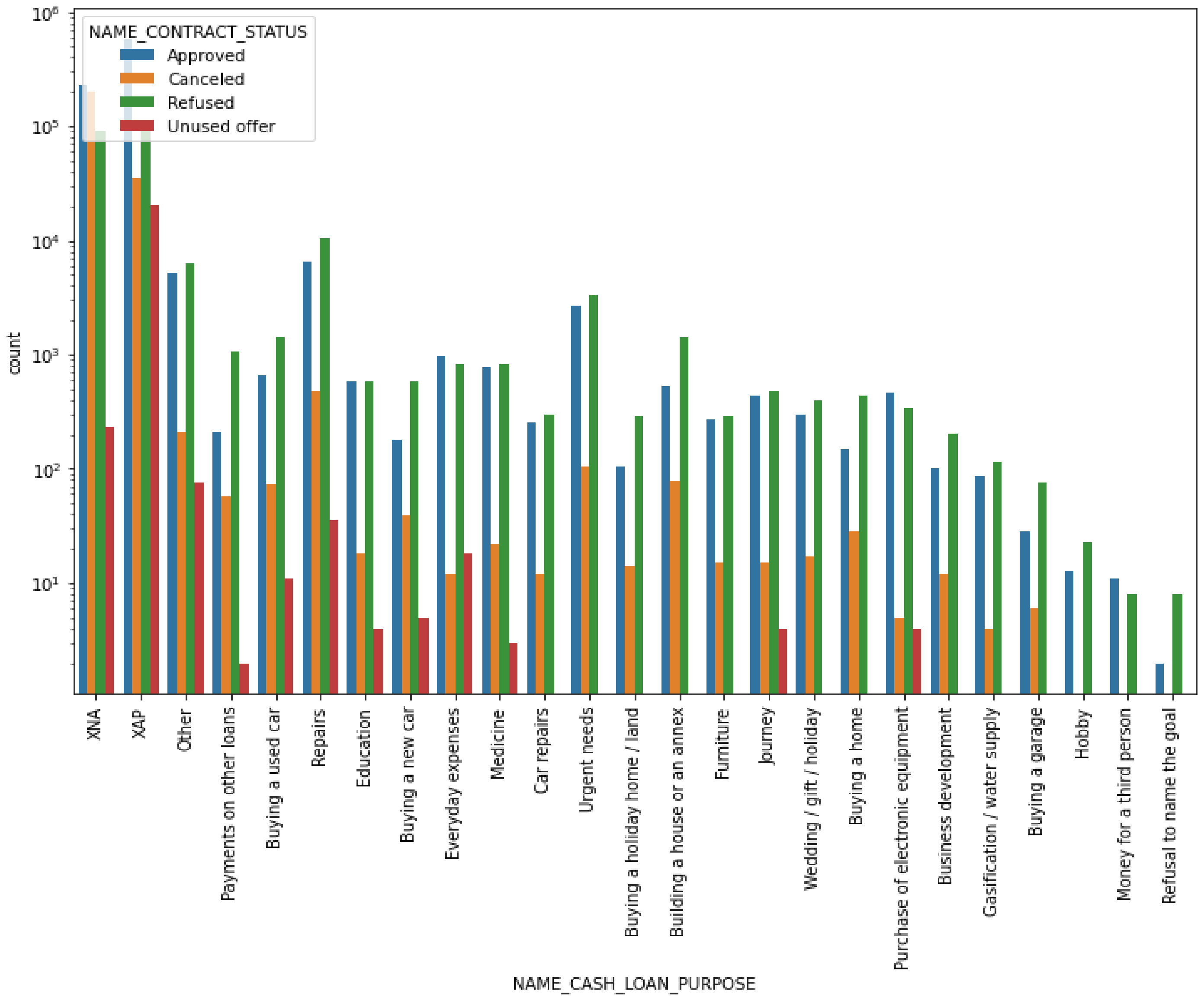
TARGET 1: MULTIVARIATE



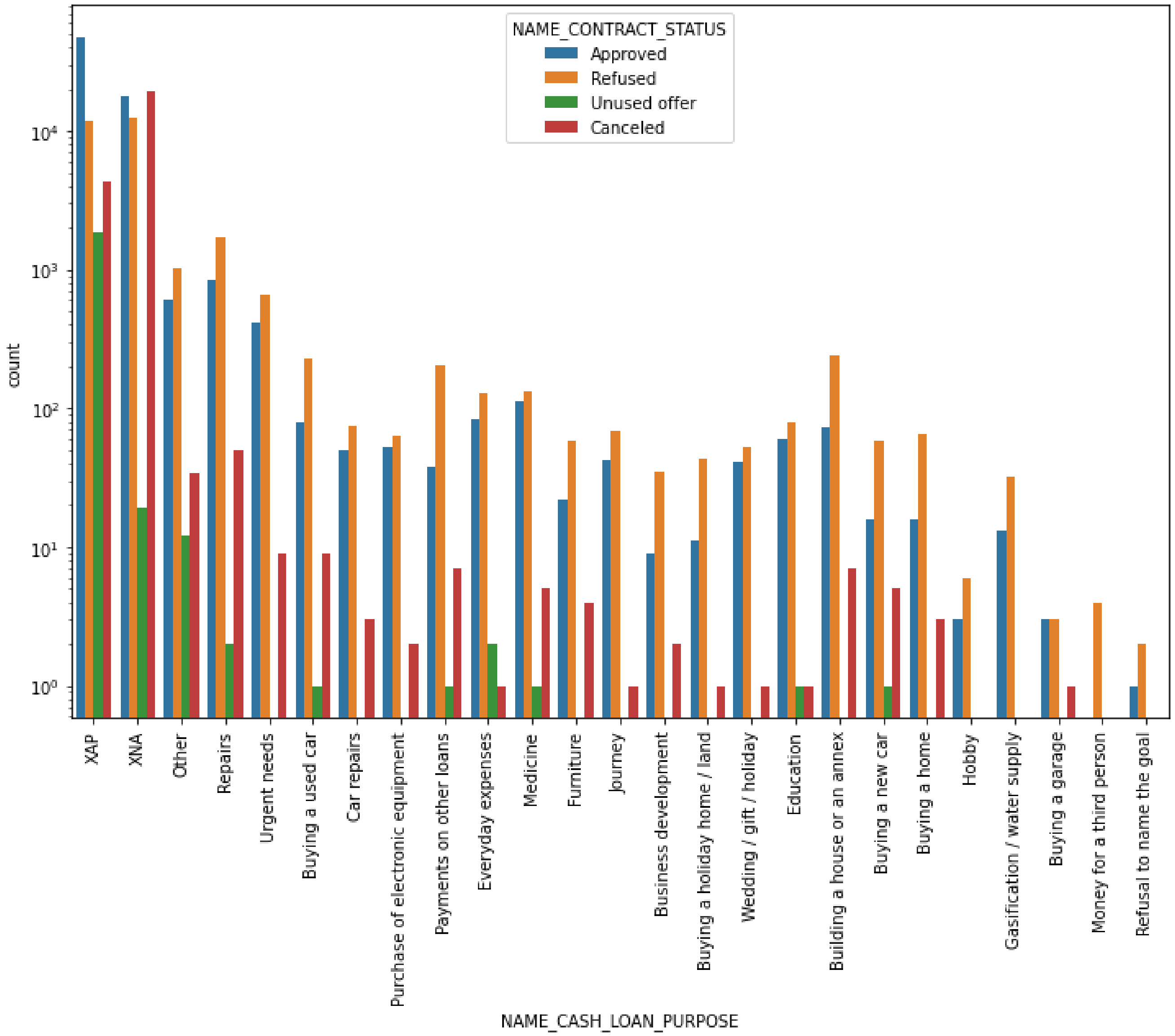
PAYMENT AND LOAN APPLICATION



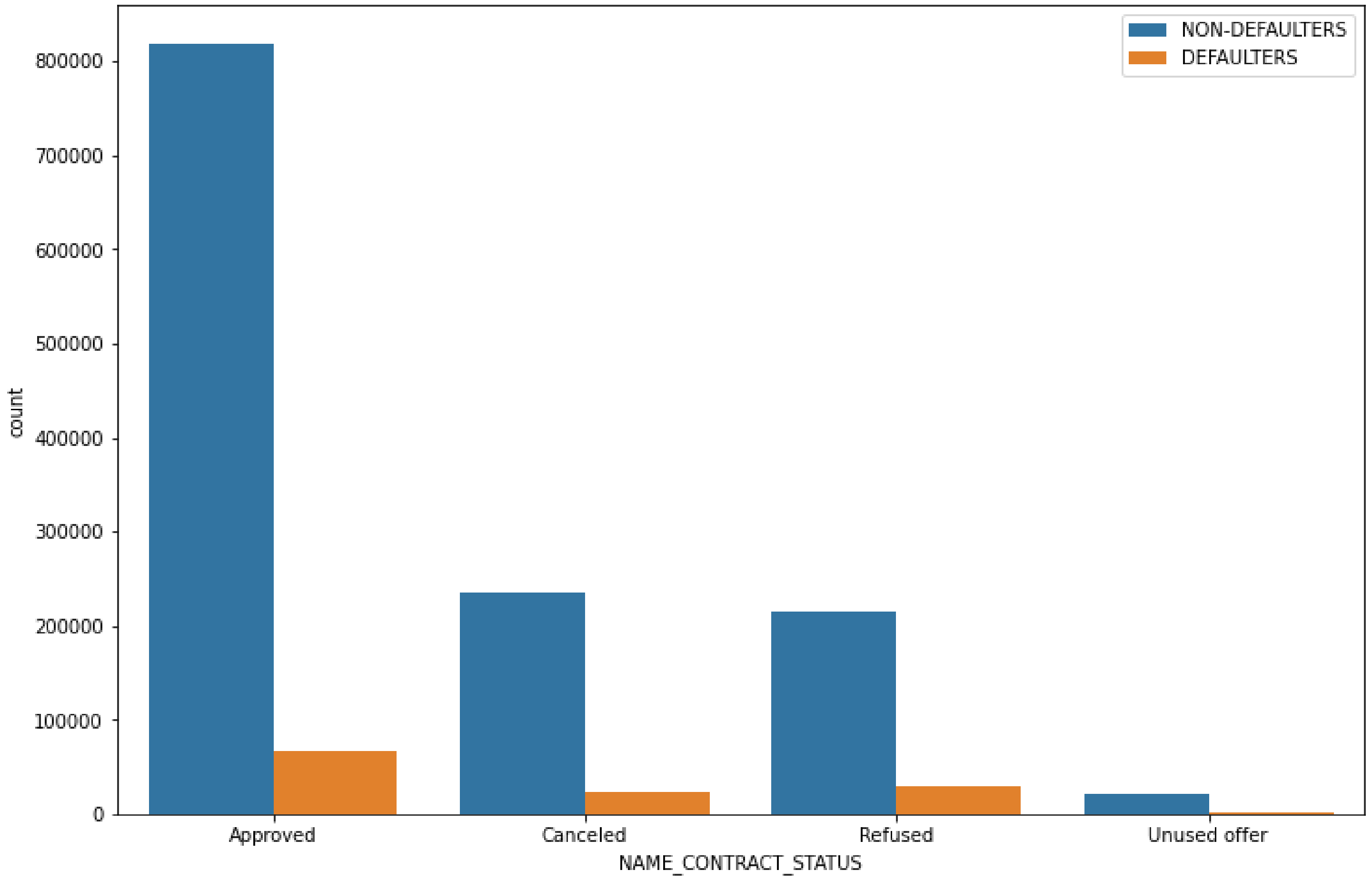
LOAN PURPOSE AND CONTRACT STATUS



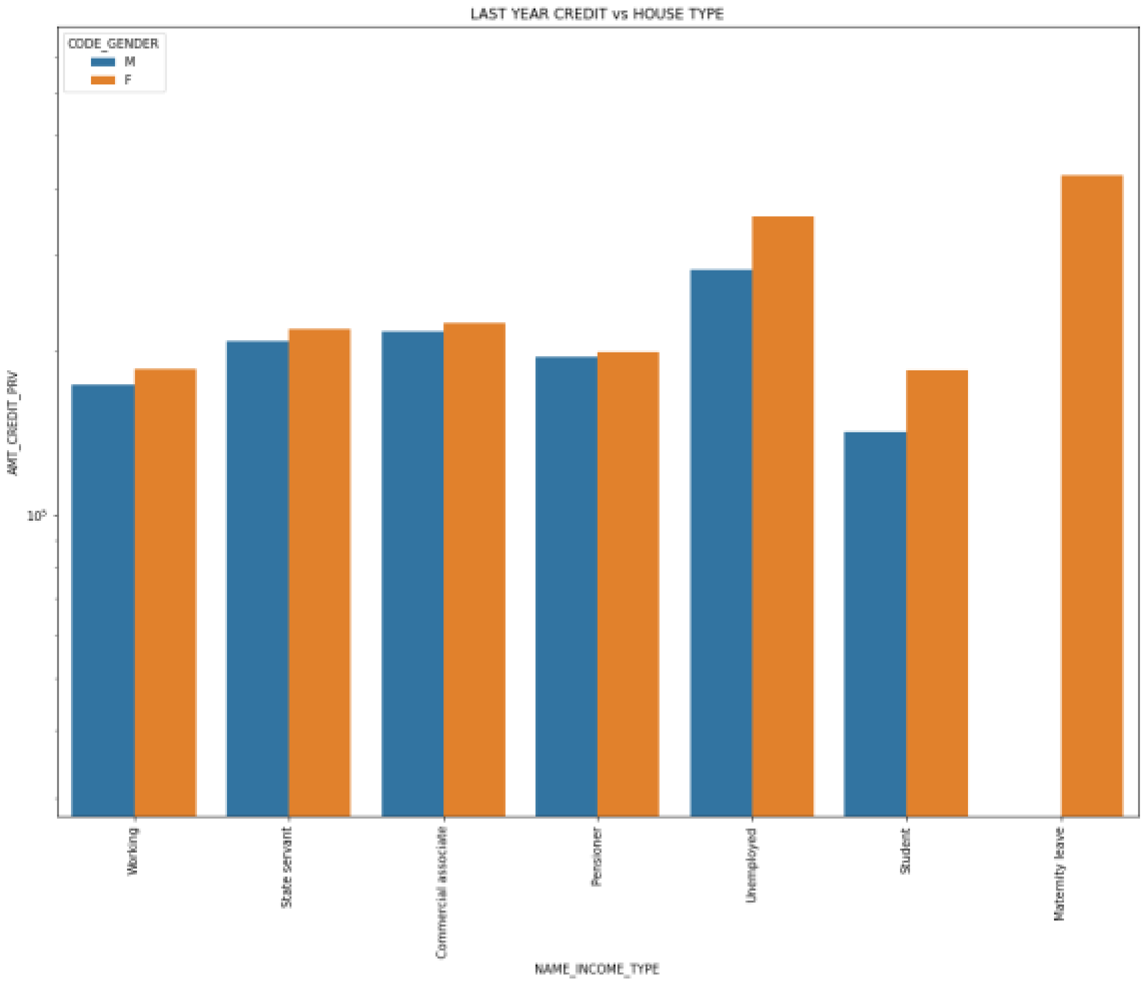
TARGET 0 LOAN PURPOSE AND CONTRACT STATUS



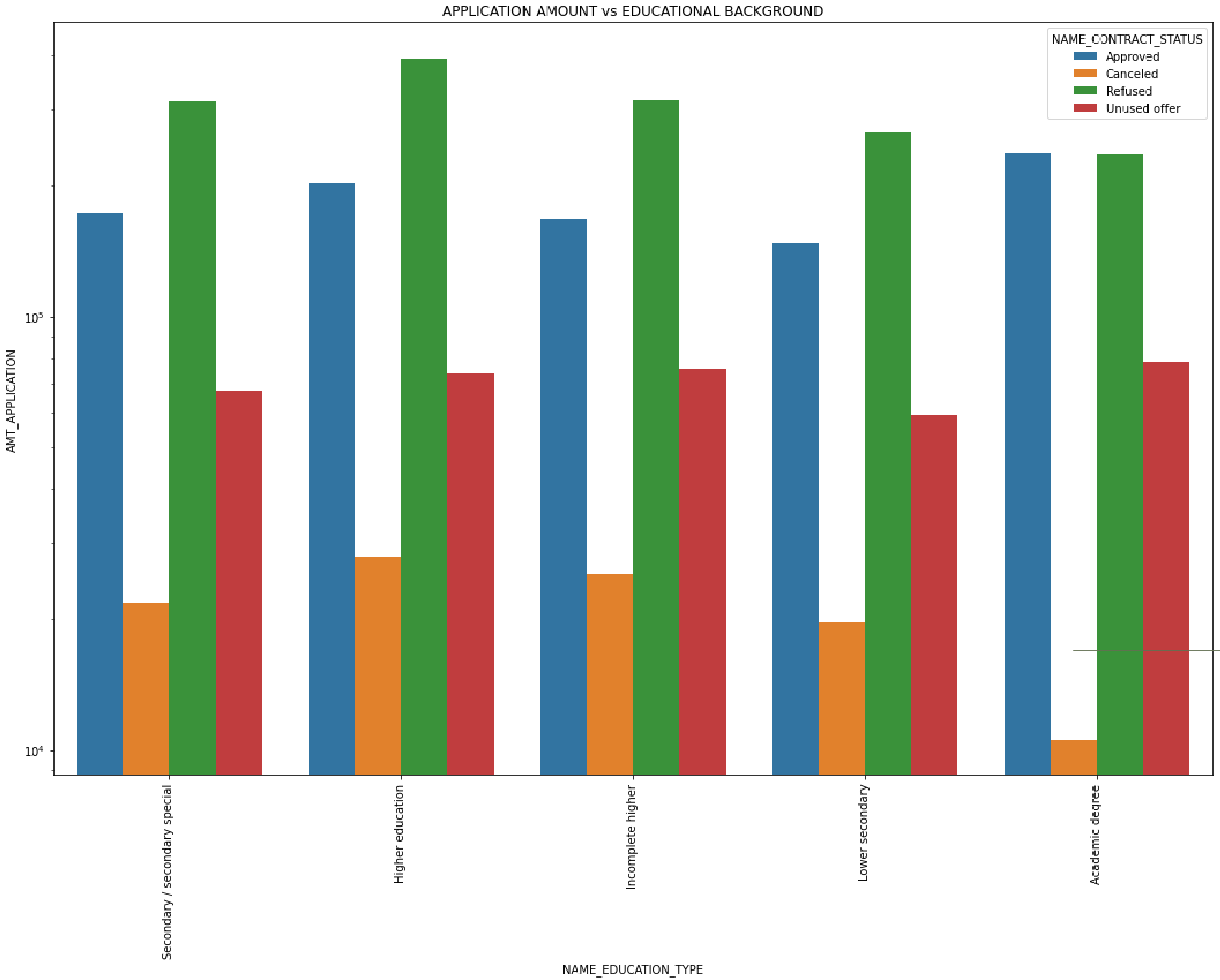
TARGET 1 LOAN PURPOSE AND CONTRACT STATUS



CONTRACT STATUS



PREV CREDIT AND HOUSE TYPE



APPLICATION AND EDUCATIONAL BACKGROUND

OBSERVATION

Following the analysis of the datasets, we can see that there are a number of variables that the bank can use to determine who is most likely to repay the loan. Factors that suggest a Non-Defaulter include:

1) LOANS EDUCATION TYPE: Academic degree holders have a lower likelihood of being approved for a loan.as well as defaulters in comparison to other apps.
2) AMT INCOME TOTAL: Customers with incomes between 700 and 800K are the least likely to use AMT INCOME TOTAL become defaulters
3) CODE GENDER: Male customers are more likely than female customers to default on their payments.

4) NAME FAMILY STATUS: Defaulters are more likely to be single or to have married civilly.

5) NAME INCOME TYPE: Clients on maternity leave or who are unemployed are most likely to be in this category.to those who have fallen behind on their payments.
6) NAME HOUSING TYPE: People who live with their parents or in rented apartments are referred to as Defaulters on loans are more likely.

Thank You

KUNAL KASHYAP