EDA CASE STUDY
Credit Risk Analysis

SUDEEP DWIVEDI

MIHIKA SHENDAGE

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.

The client with payment difficulties :-

(Target variable that tells 1 is)

- 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 :-

(Target variable that tells 0 is)

- All other cases when the payment is paid on time.

Analyzed Accordingly -

Steps:-

1. **Data Cleaning** - Handling null values and garbage values,
   Finding out which operations to perform to treat this values.

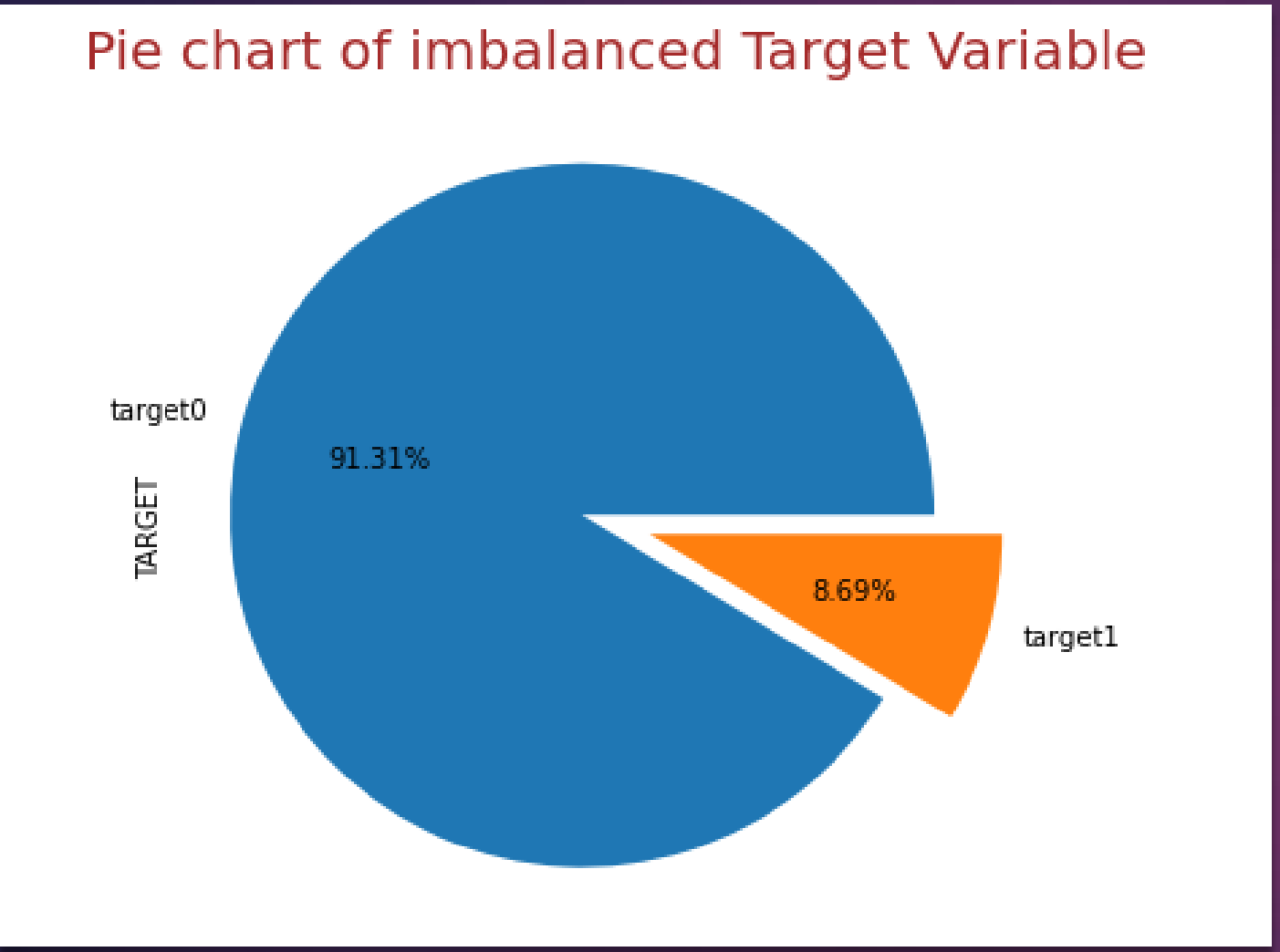
2. **Checking for Outliers** - Checking for Data Imbalance, Ratio of Imbalance.

3. **Top 10 correlations for client with Payment Difficulties, other variables within Application Dataset & Previous Application Dataset .**

4. **Finding which Correlation is more relevant.**

5. **Univariate and Bivariate analysis and their visualizations.**

Univariate Analysis.



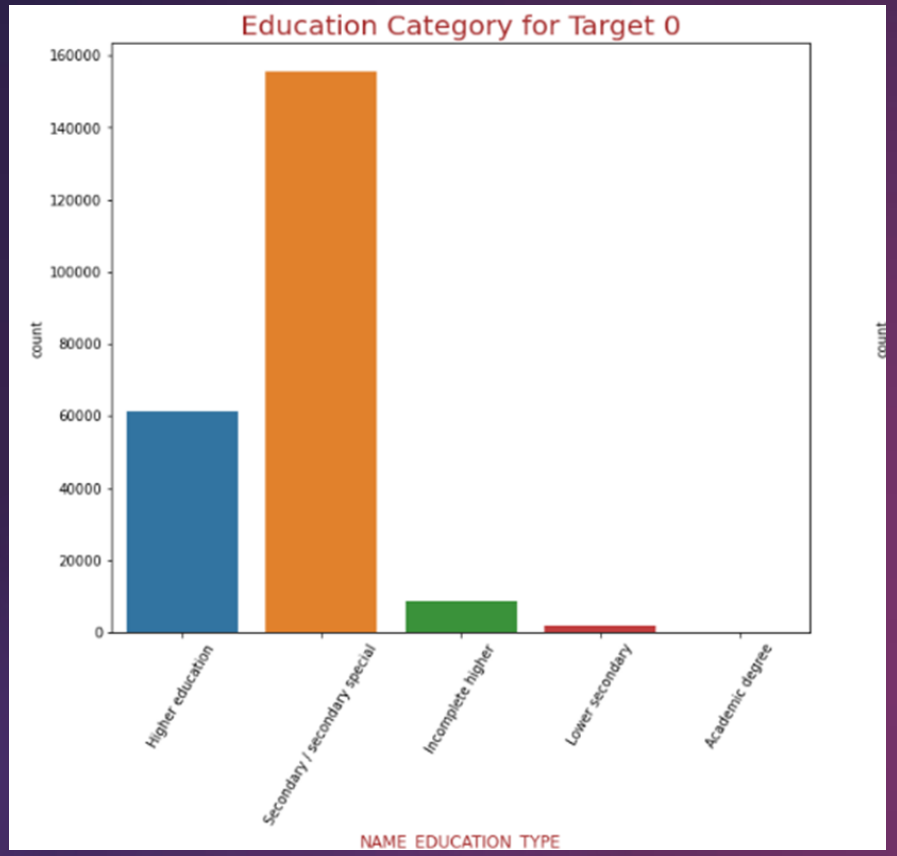
Proportion of defaulters in Dataset.

8.69% of Defaulters are present in Total Loans.

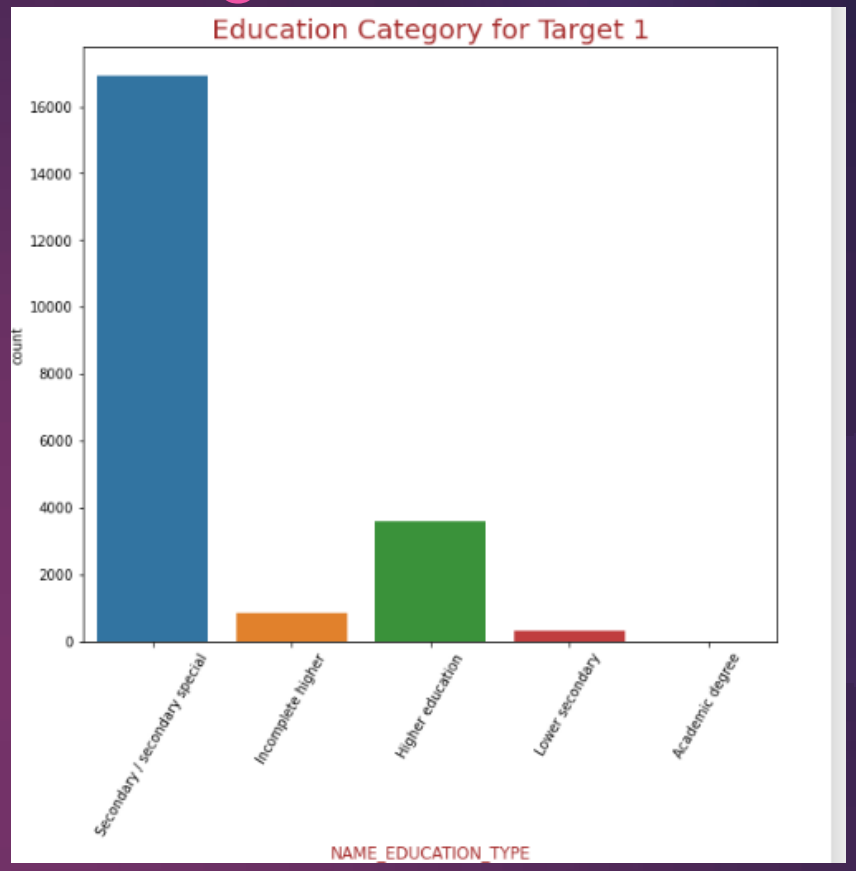
Therefore we can say that there is a data imbalance in the dataset because there is too much difference between the count of the values.

And the data imbalance ratio is **10.5:1**

Target 0 = non-defaulters

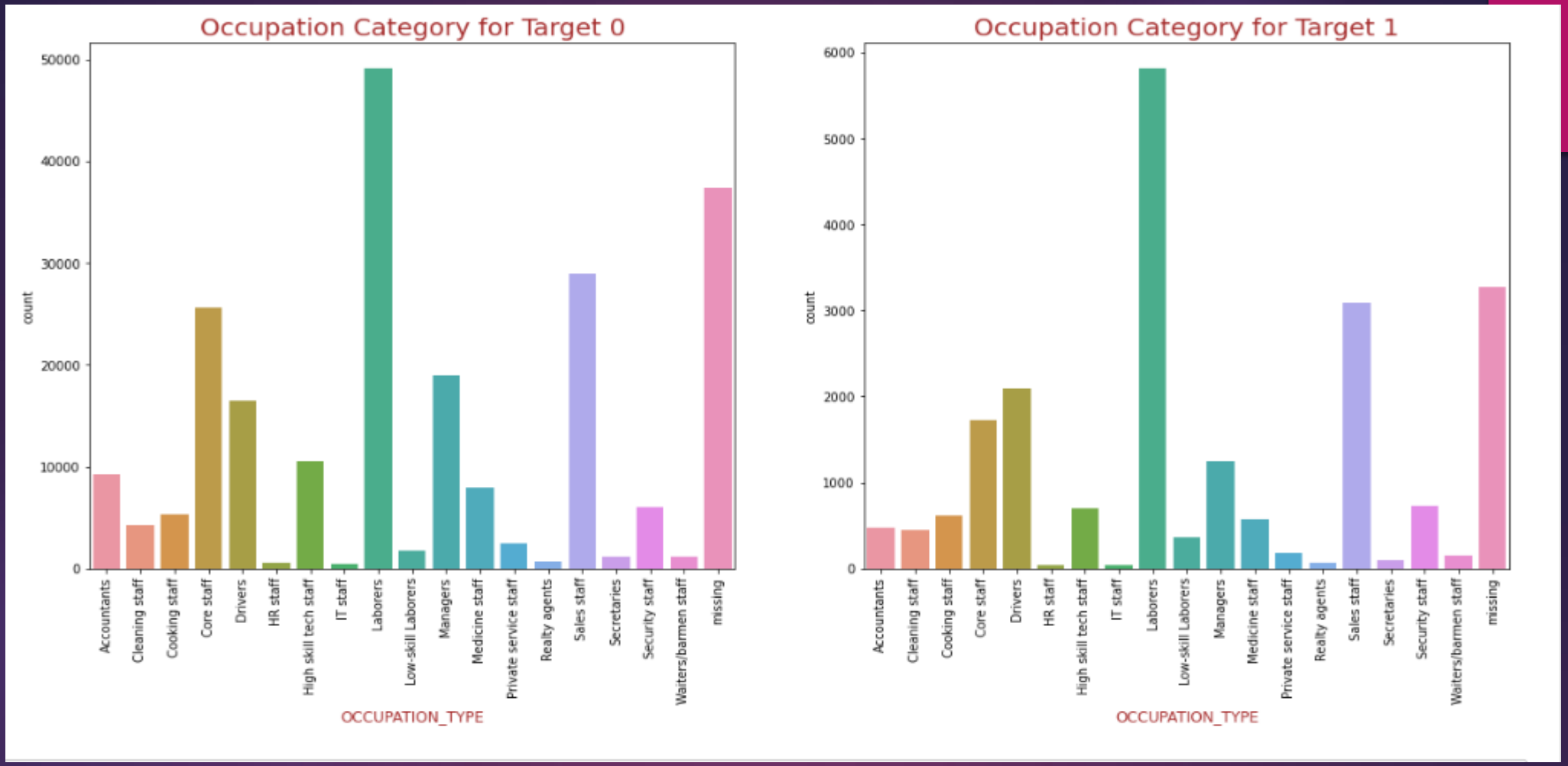


Target 1 = defaulters

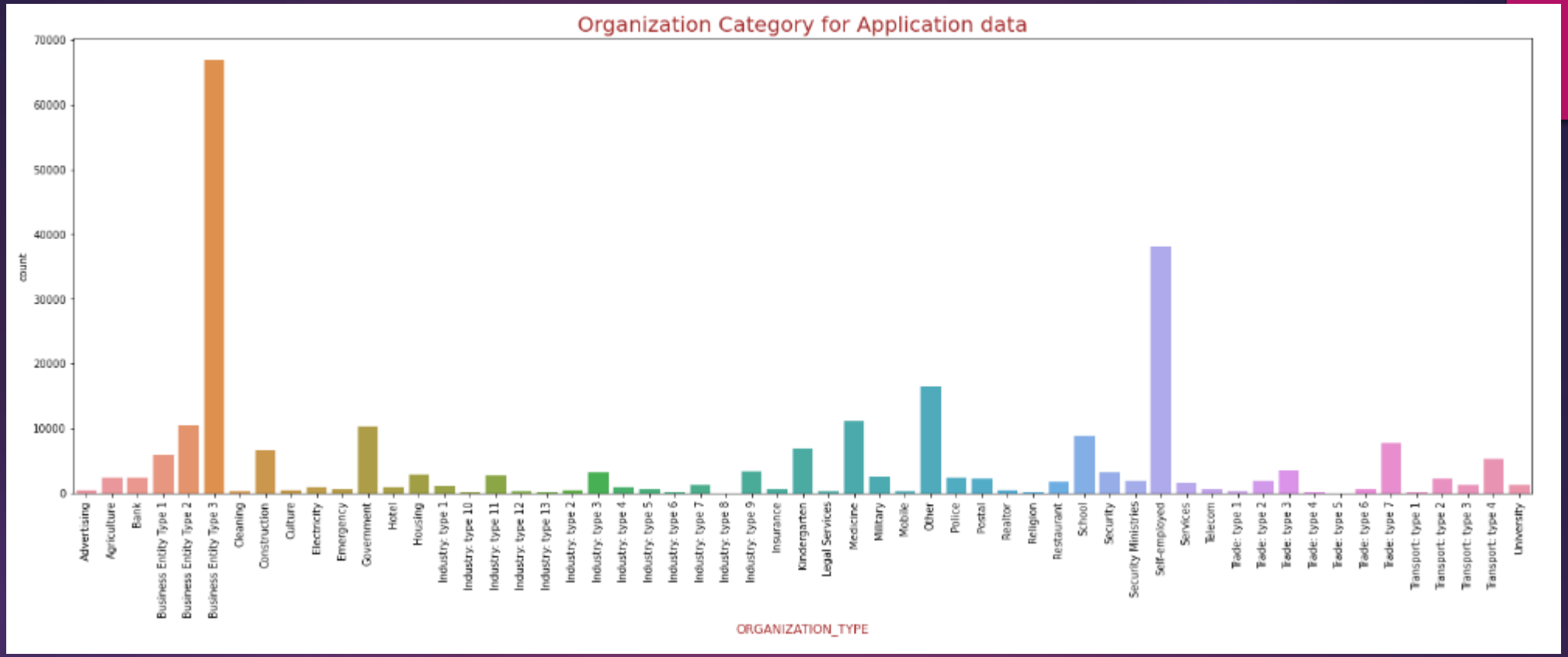


Analysis of Education category for Target 0 and Target 1.

- **Non-defaulters** with **Secondary education type** show most number of counts.
- **Defaulters** with **Secondary education type** also show most number of counts.
- This means that **most number of loans** are taken for **Secondary education irrespective of Target variable.**



- The missing values are large in quantity, we need to ignore them.
- The maximum number of people are **Laborers** and minimum are from **IT** and **HR staff** from the **Current Application**.



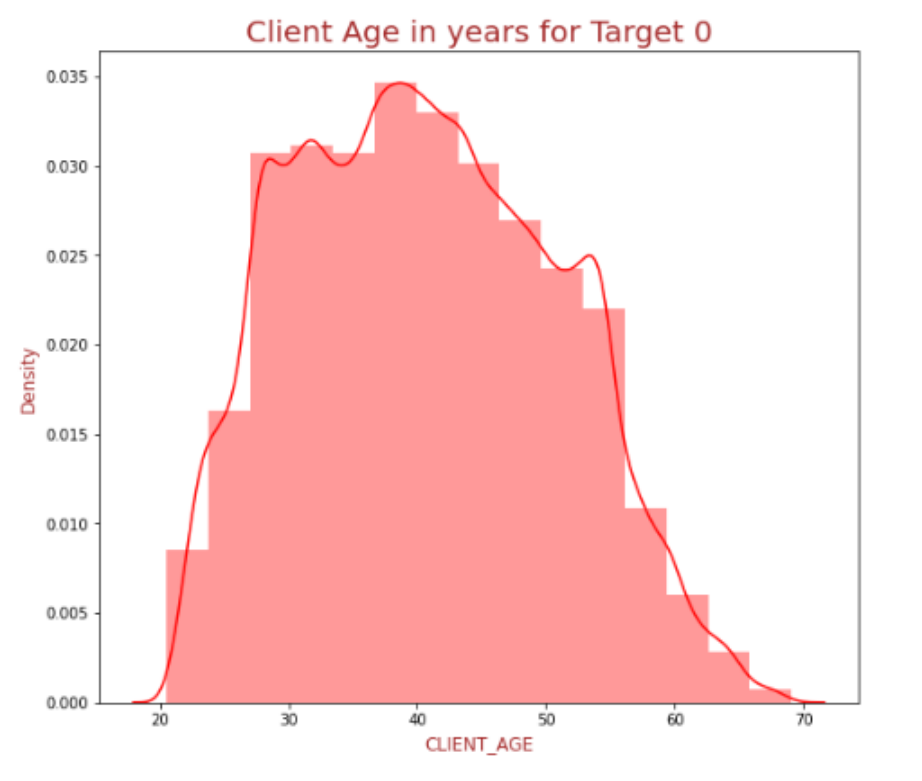
From Current Application Data the organization type of '**Business Entity type 3**' followed by '**Self Employed**' are the **most** number of applicants present in the dataset.

-For **Target 0** variable, the **maximum** of Client age is present between approximately 30 – 45 years old .

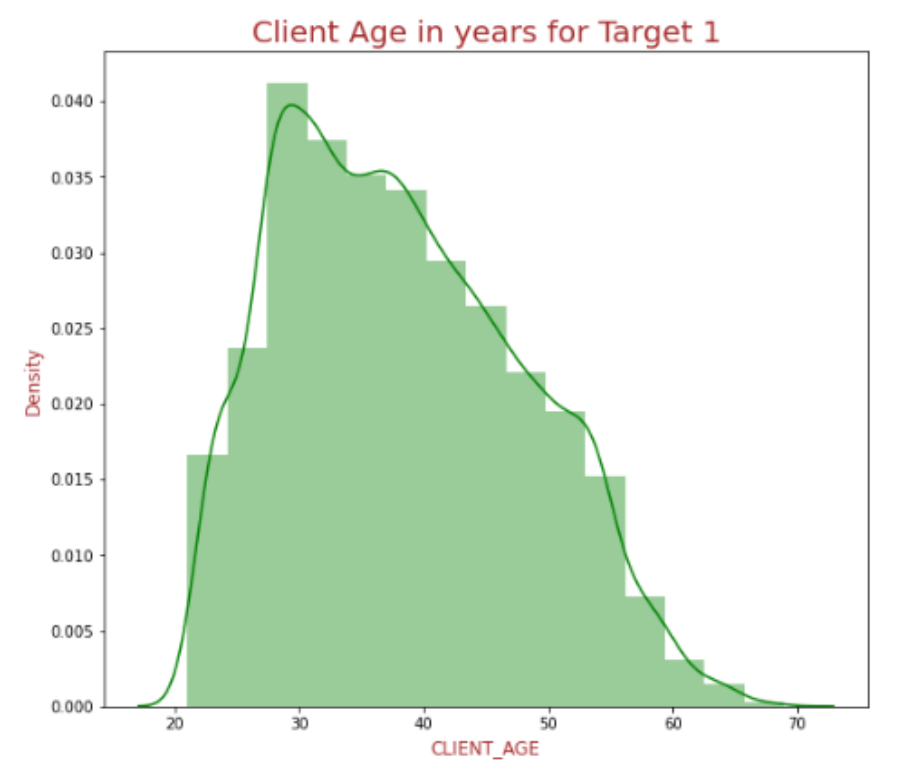
-For **Target 1** variable the **maximum** of Client age is present between approximately 28 – 40 years old shows the Highest Density.

- **This means that client more than 40 years of age is less likely to make a default.**

For Target 0.

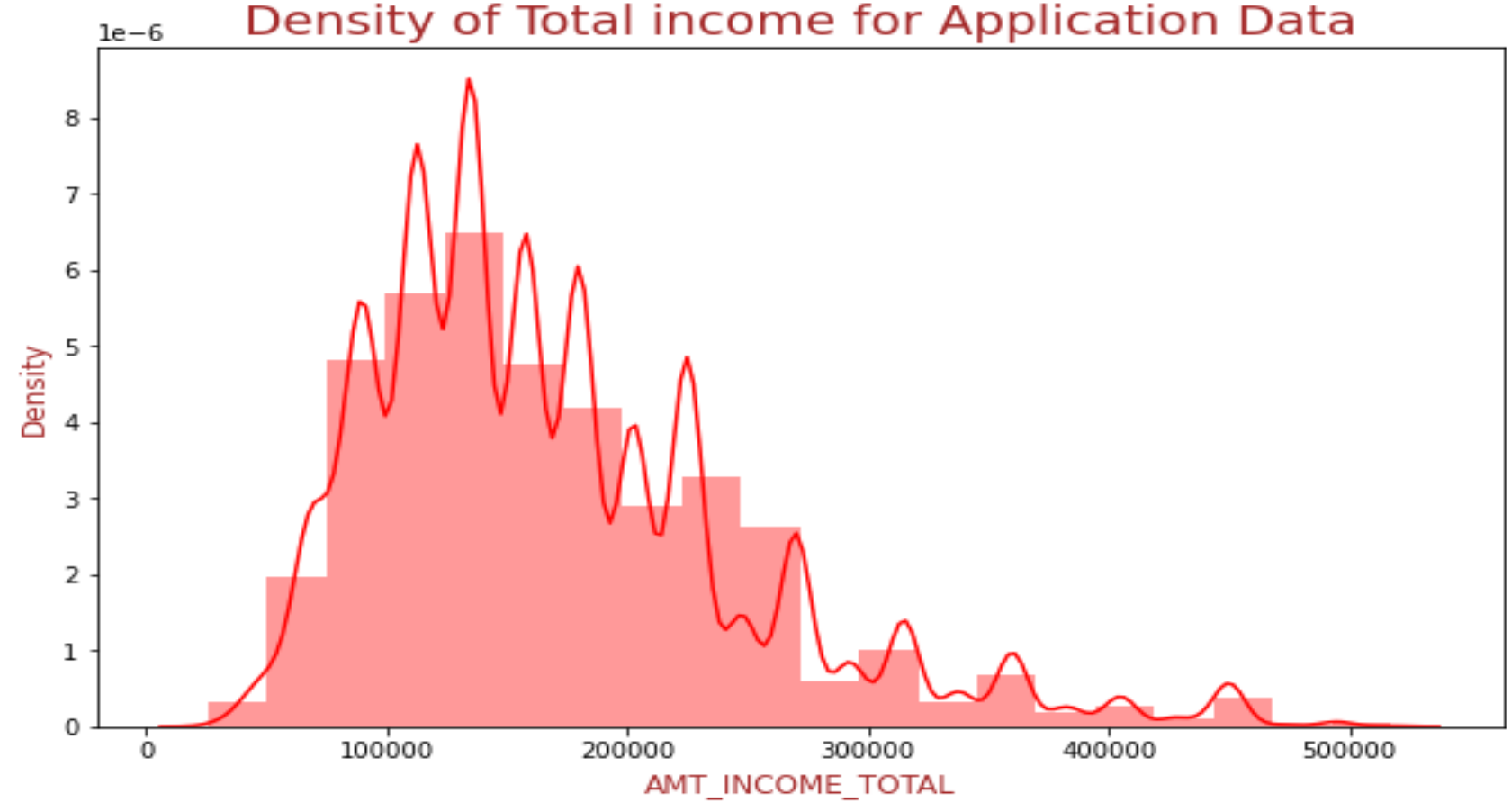


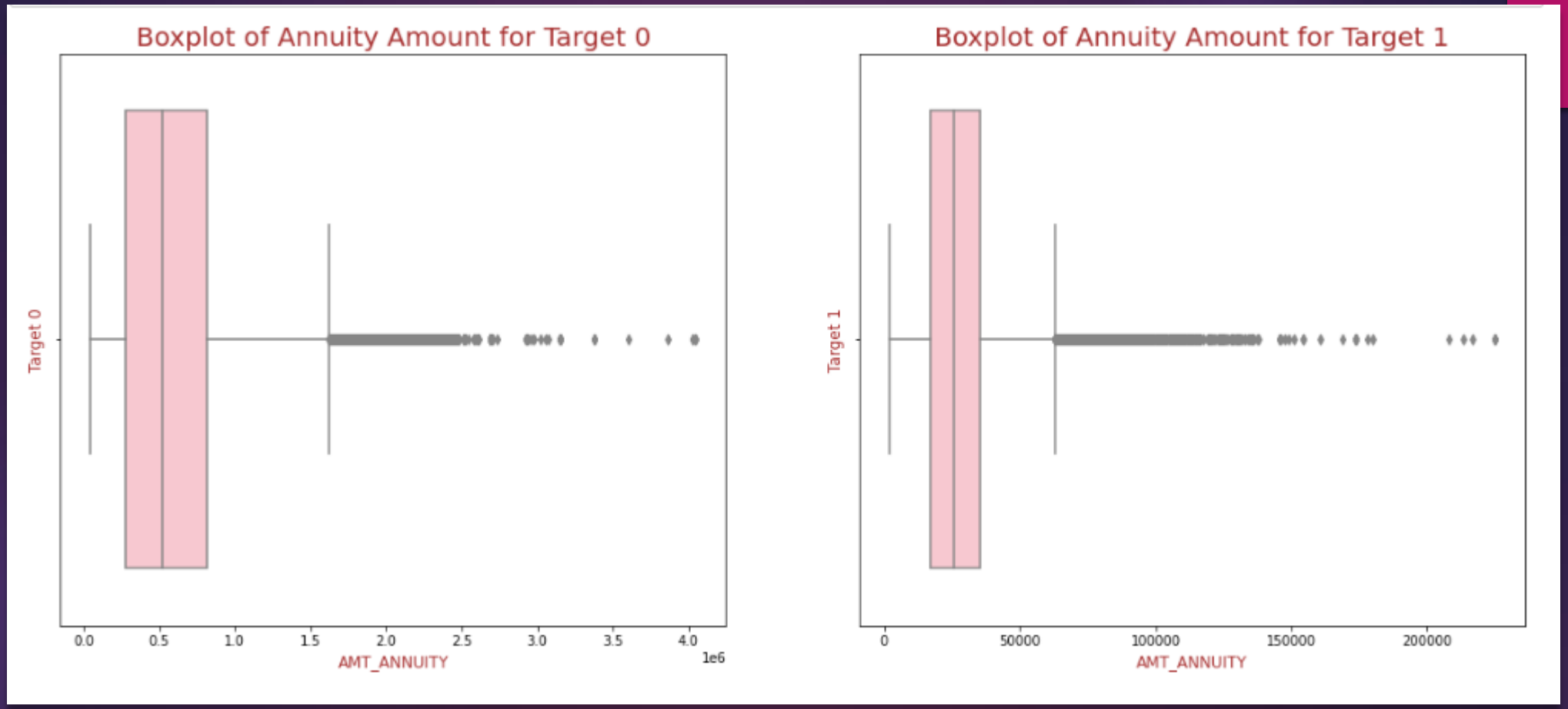
For Target 1.



Income column after removing the outliers.

The Application Data shows that people with income range between 1 lakh - 2 lakhs show the **Highest Density Range**.



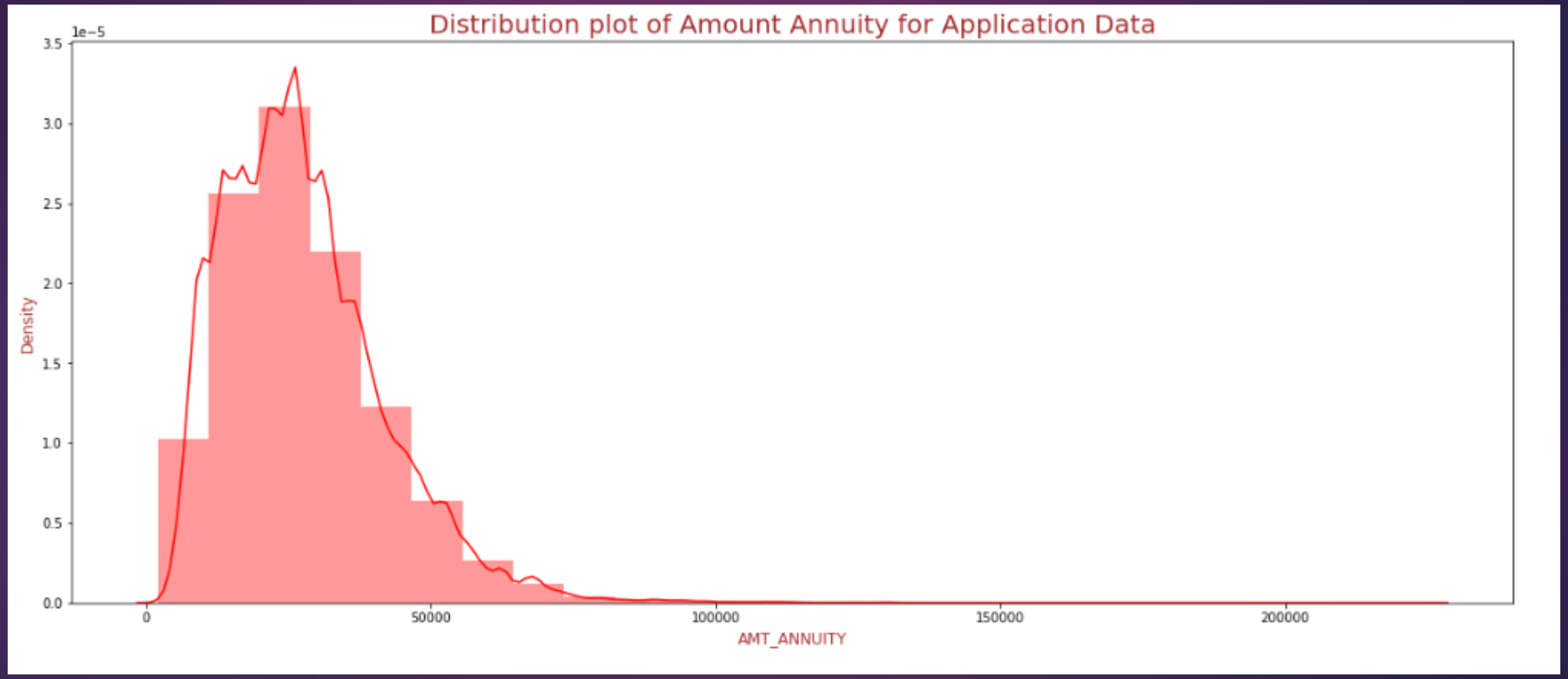


Comparison for Annuity Amount for Target 0 and Target 1

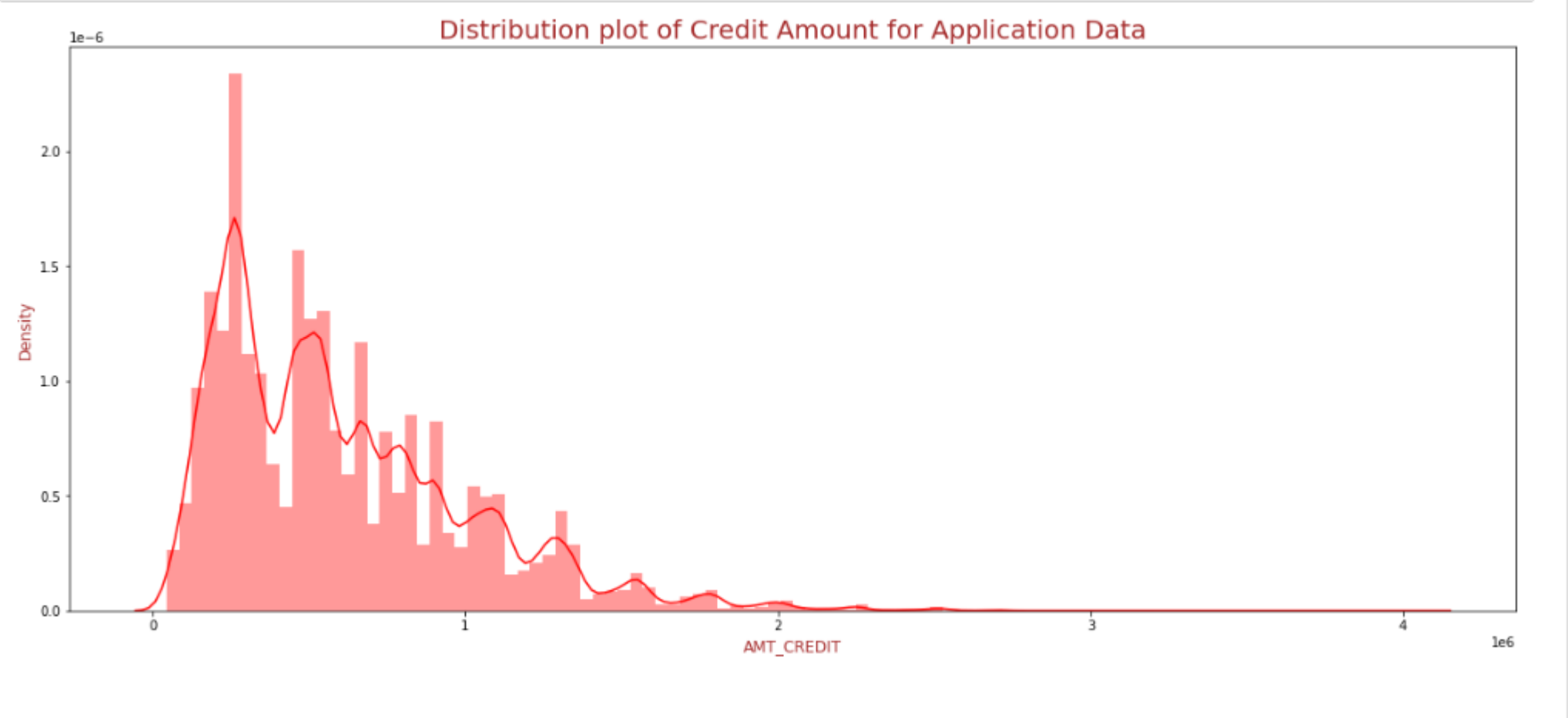
These are Correct Entries that will help us to reassure our analysis on Annuity Amount for both Target 0 and Target 1.

No need to remove outliers.

Amount Annuity for **Application Data** shows that Highest Range of Density lies between range of **0 – 50000** rupees.

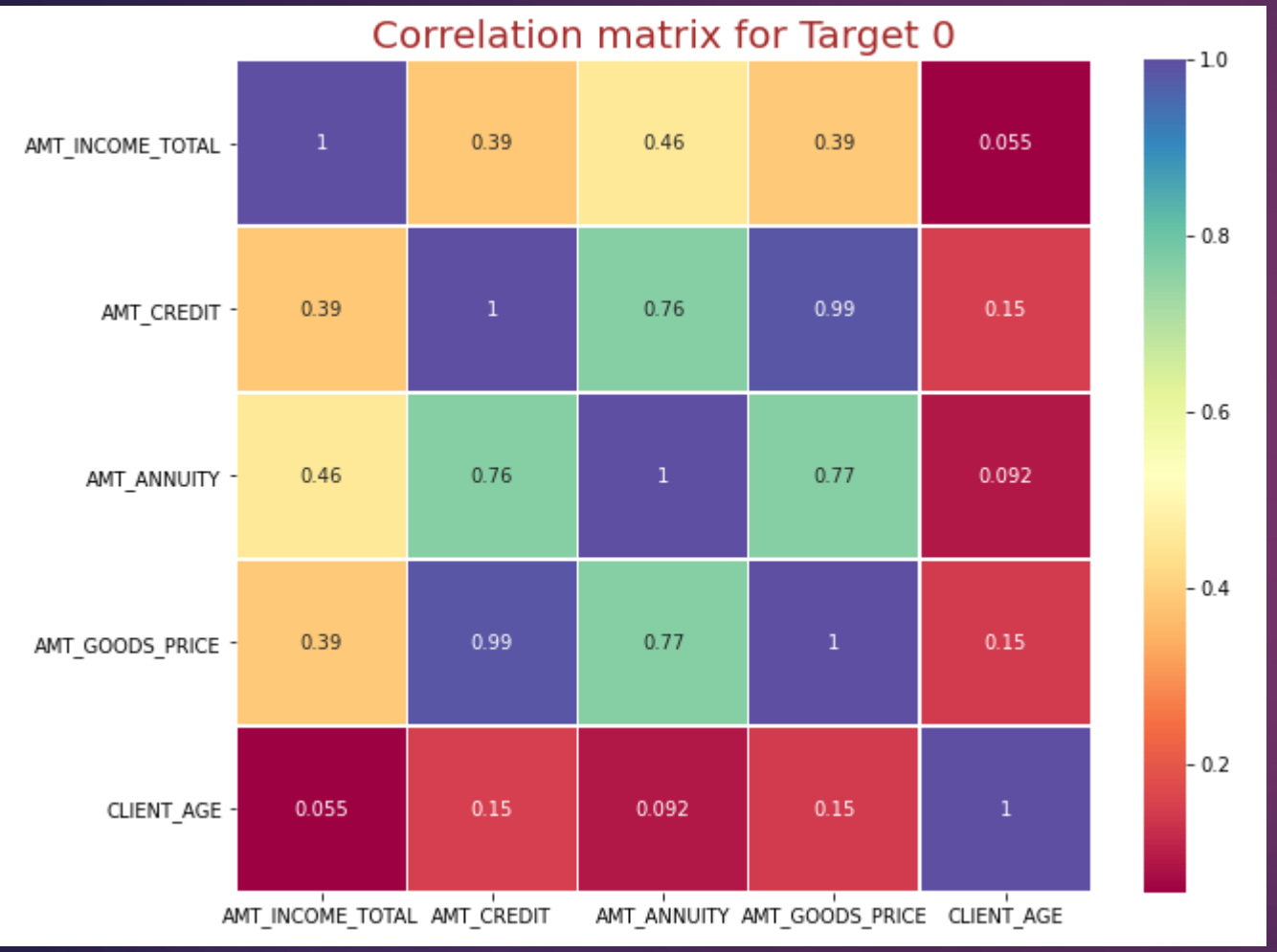


Credit Amount For **Application Data** shows Highest Density between range of 2 Lakhs rupees – 6 Lakhs rupees.



Bivariate Analysis.

Selecting Top 10 variables to draw a correlation matrix for bivariate and multivariate analysis.



AMT_INCOME_TOTAL

AMT_CREDIT

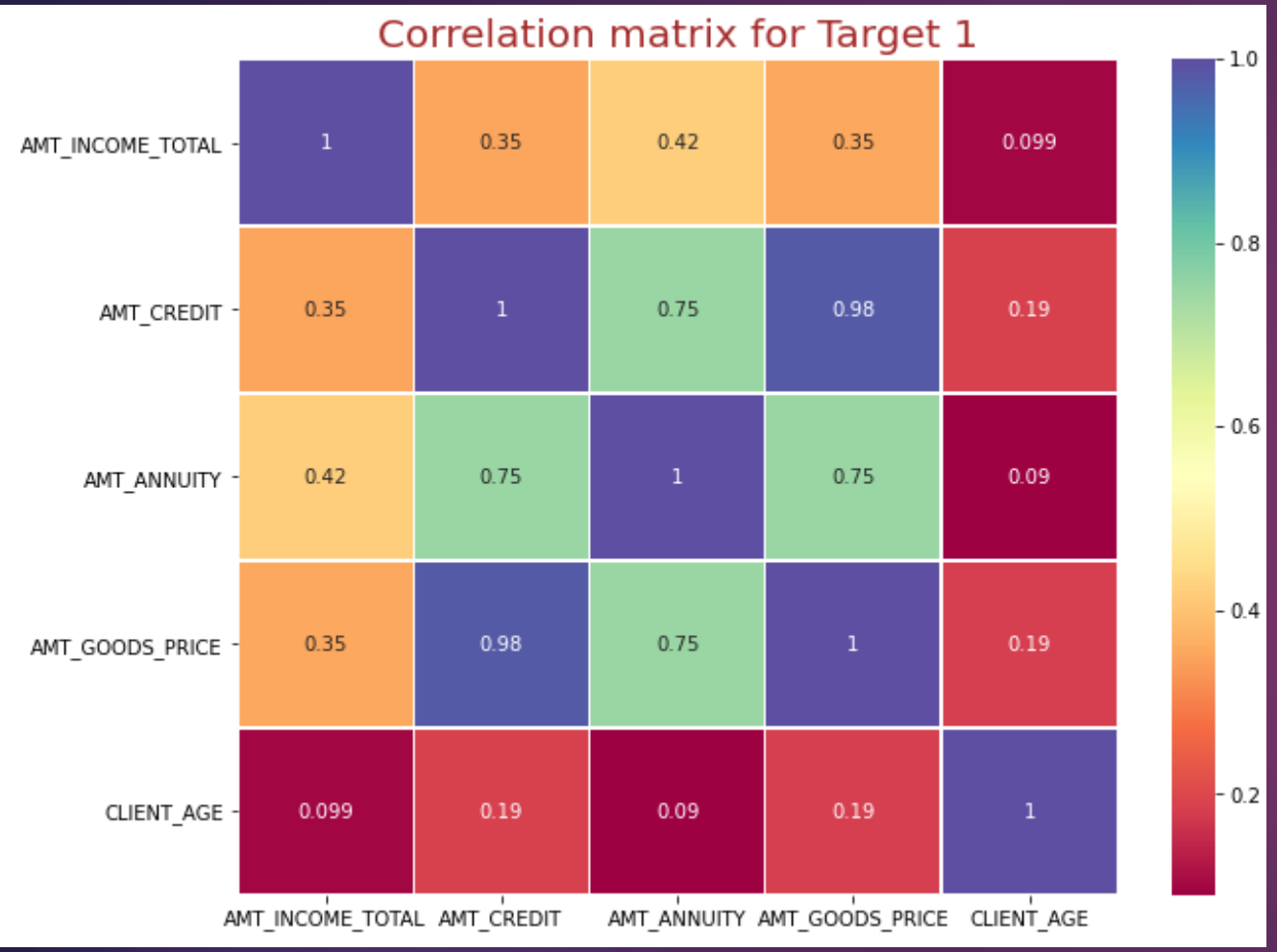
AMT_ANNUITY

AMT_GOODS_PRICE

CLIENT_AGE

These Variables show the highest correlation for **Target 0.**

Correlation matrix for Target 1.



AMT_INCOME_TOTAL

AMT_CREDIT

AMT_ANNUITY

AMT_GOODS_PRICE

CLIENT_AGE

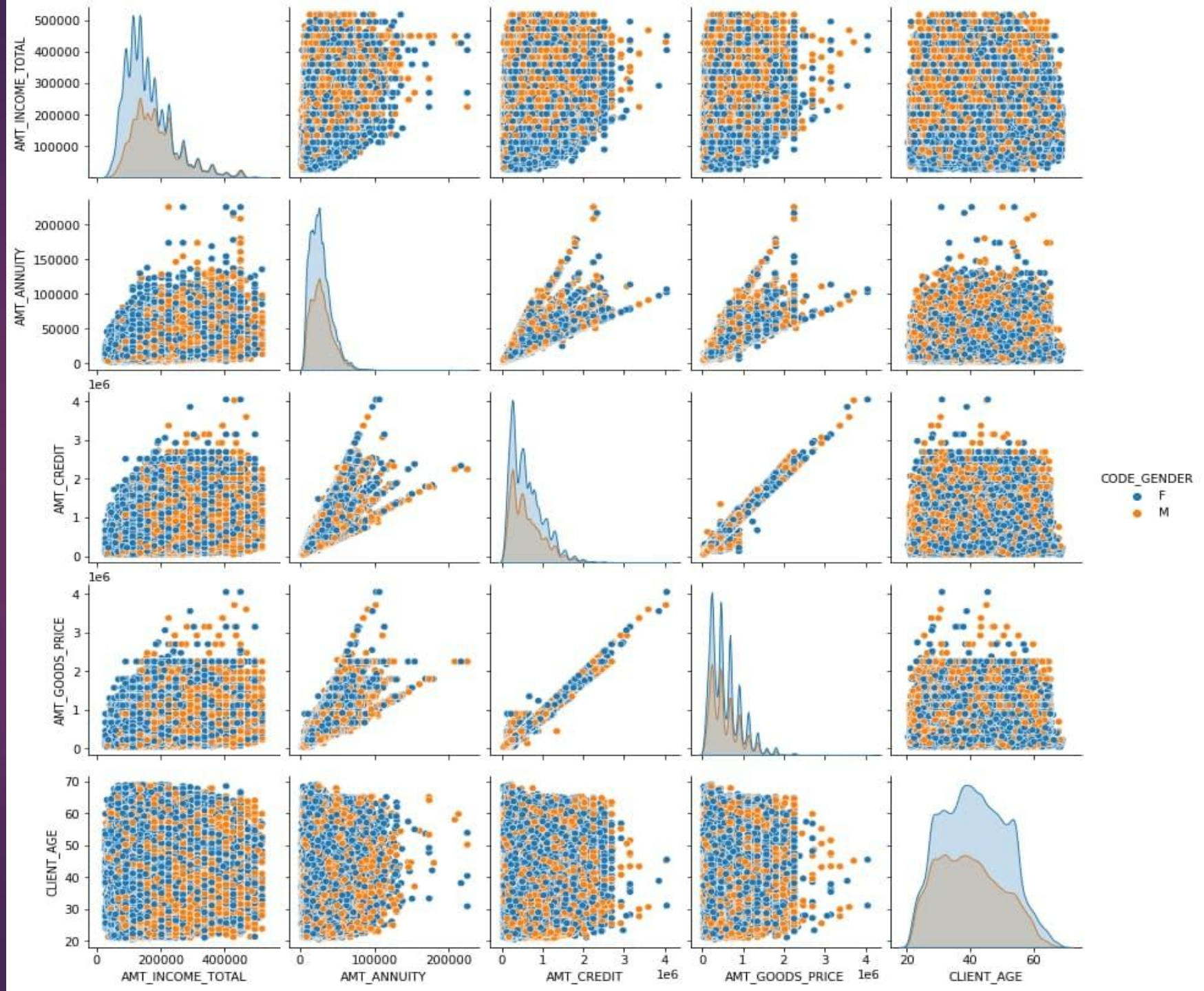
These Variables show the highest correlation for **Target 1.**

Correlation Matrix Pairplot for Target 0 -

- We can see in the pair plot that when the **Credit amount increases** the **Goods price** also **increase**.

- As the **Total income increases** we can see that there are some clients who take higher Credit amount, higher Goods price and higher Annuity.

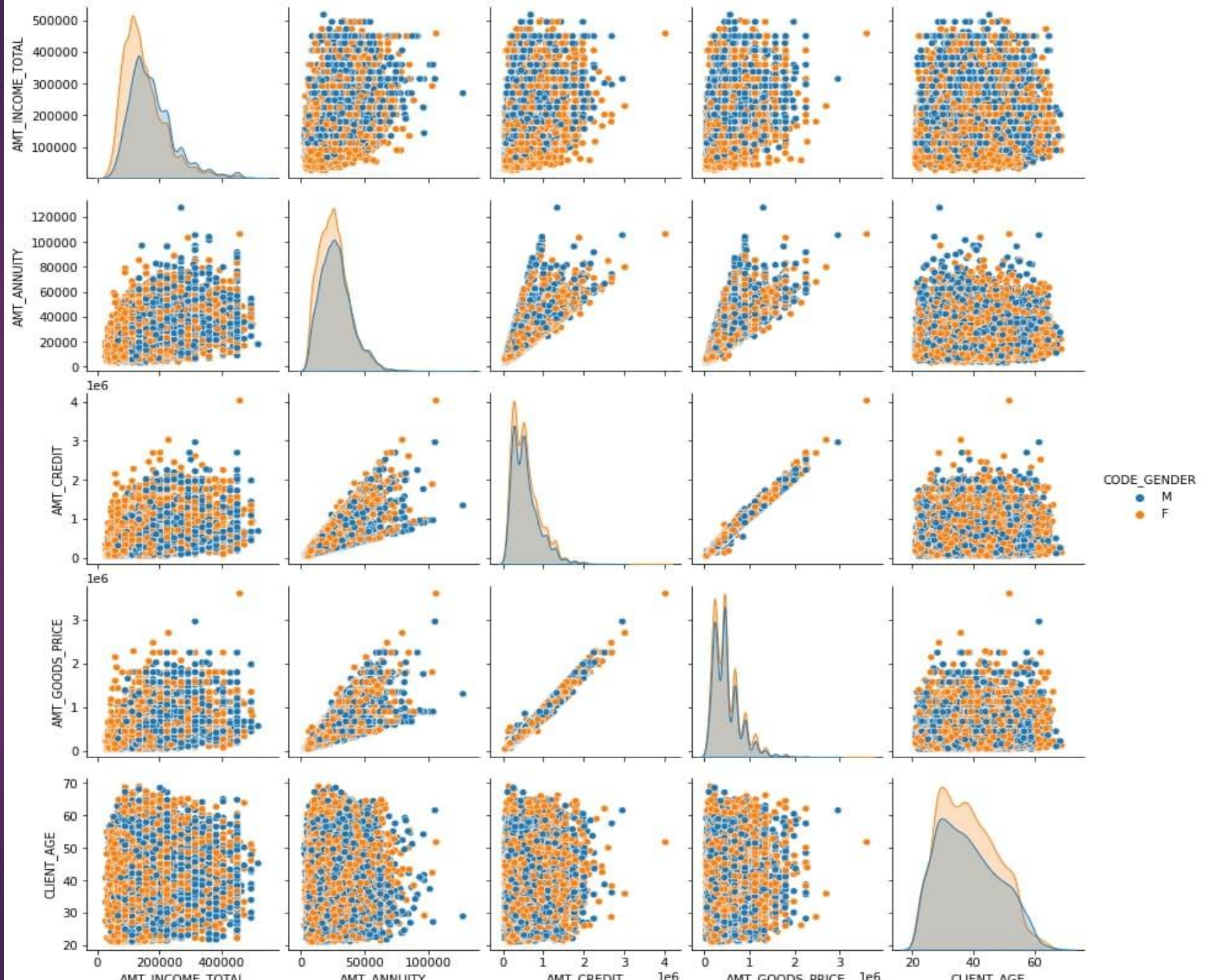
- There is a **large difference** between male and female ages.



Correlation Matrix Pairplot for Target 1 -

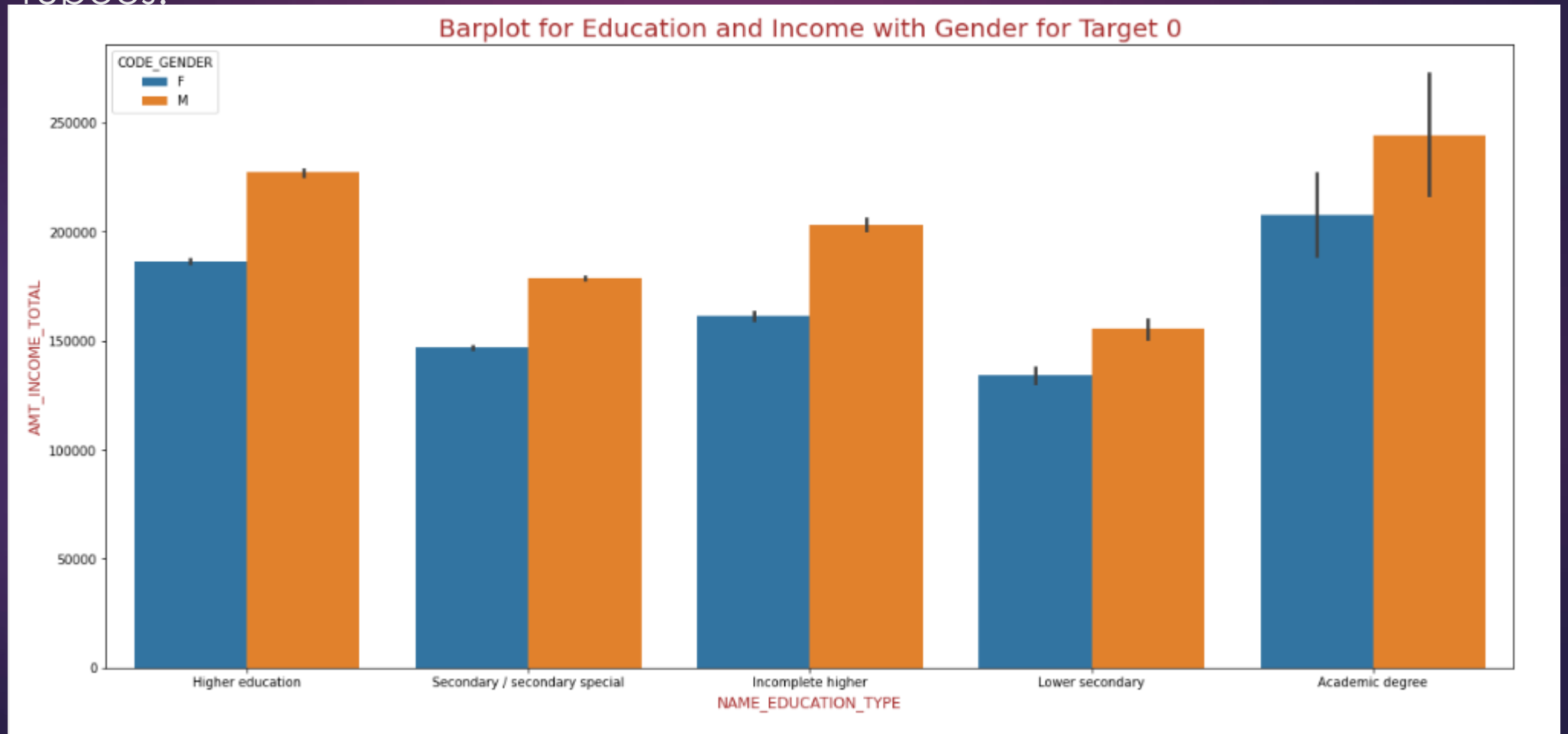
- We can see in the pair plot that when the **Credit amount increases** the **Goods price** also **increase.**

- As the **Total income** increases we can see that there are some clients who take higher Credit amount, higher Goods price and higher Annuity and not able to pay on time.



Education Type and Total Income (Target 0)

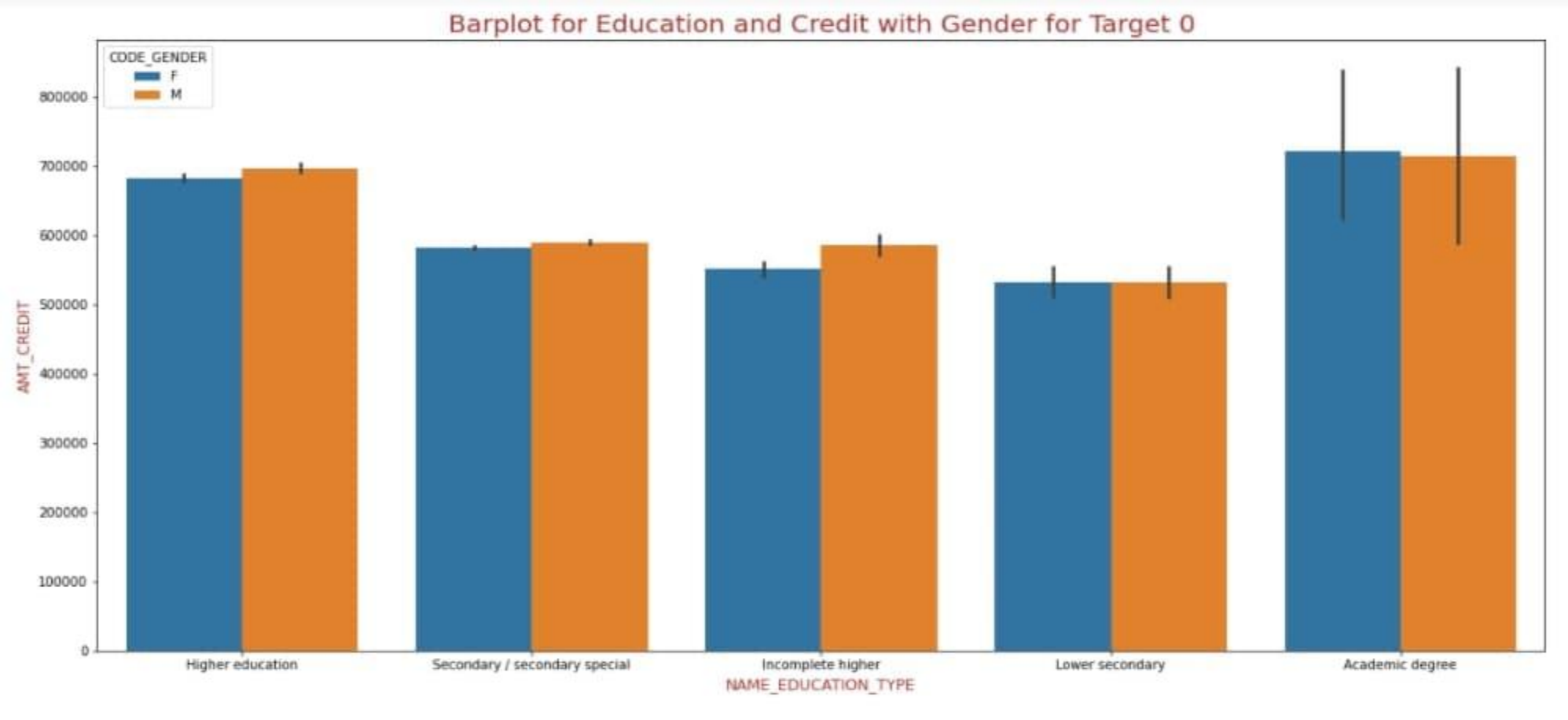
- **Males** with Academic Degree seem to have Highest Income Total above 2 lakh rupees which are non defaulters.
- **Females** with Academic Degree seem to have Highest Income Total above 2 lakh rupees.



Education Type and Total Income (Target 1)

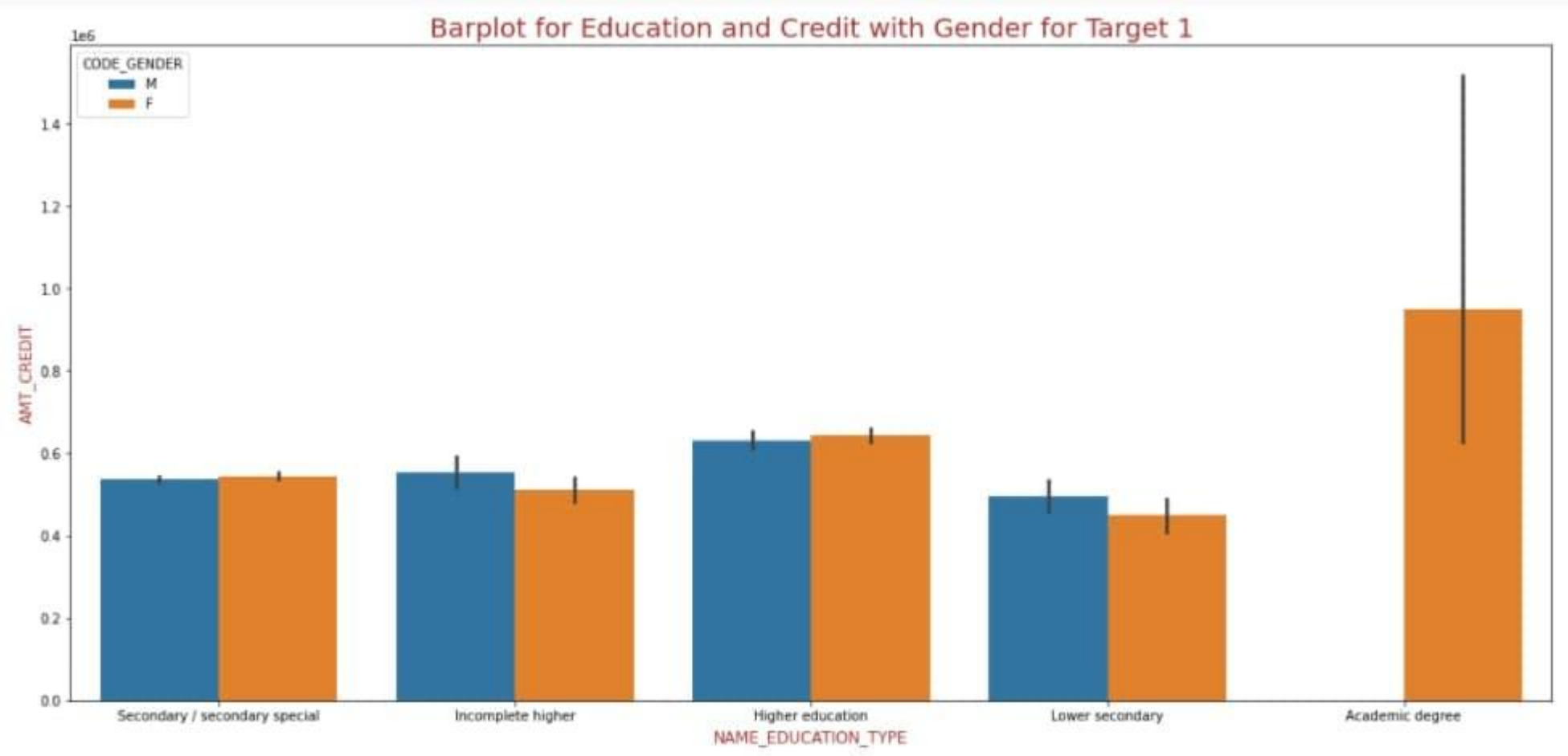
Males with Higher education seem to have Highest Income Total above 2 lakh rupees.

Females with Academic Degree seem to have Highest Income Total above 3 lakh rupees, they form the **most** number of **defaulters** w.r.t education.



Education Type and Credit Amount (Target 0)

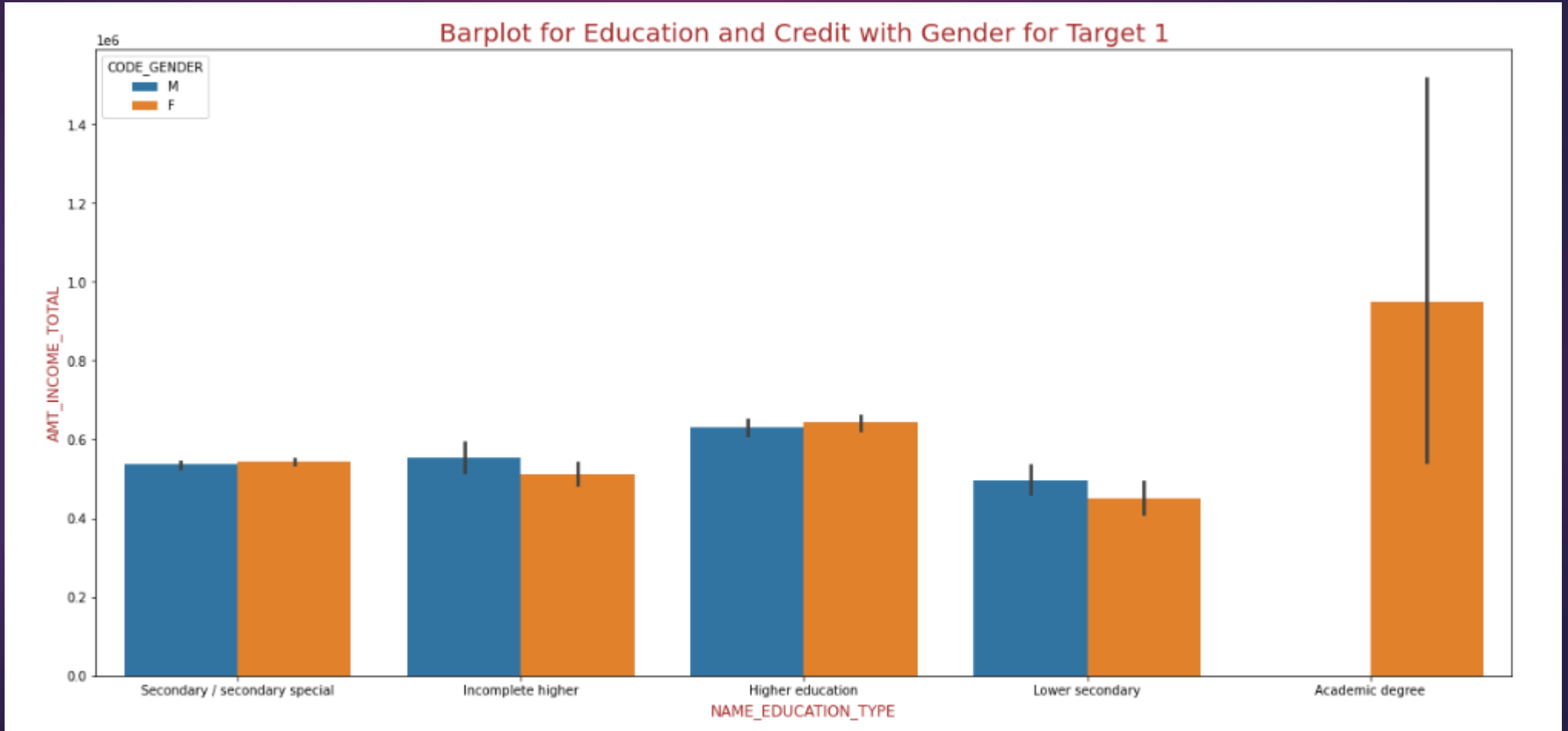
Males and **Females** with **Academic Degree** seem to have Highest Credit above **700000.**



Education Type and Credit Amount (Target 1)-

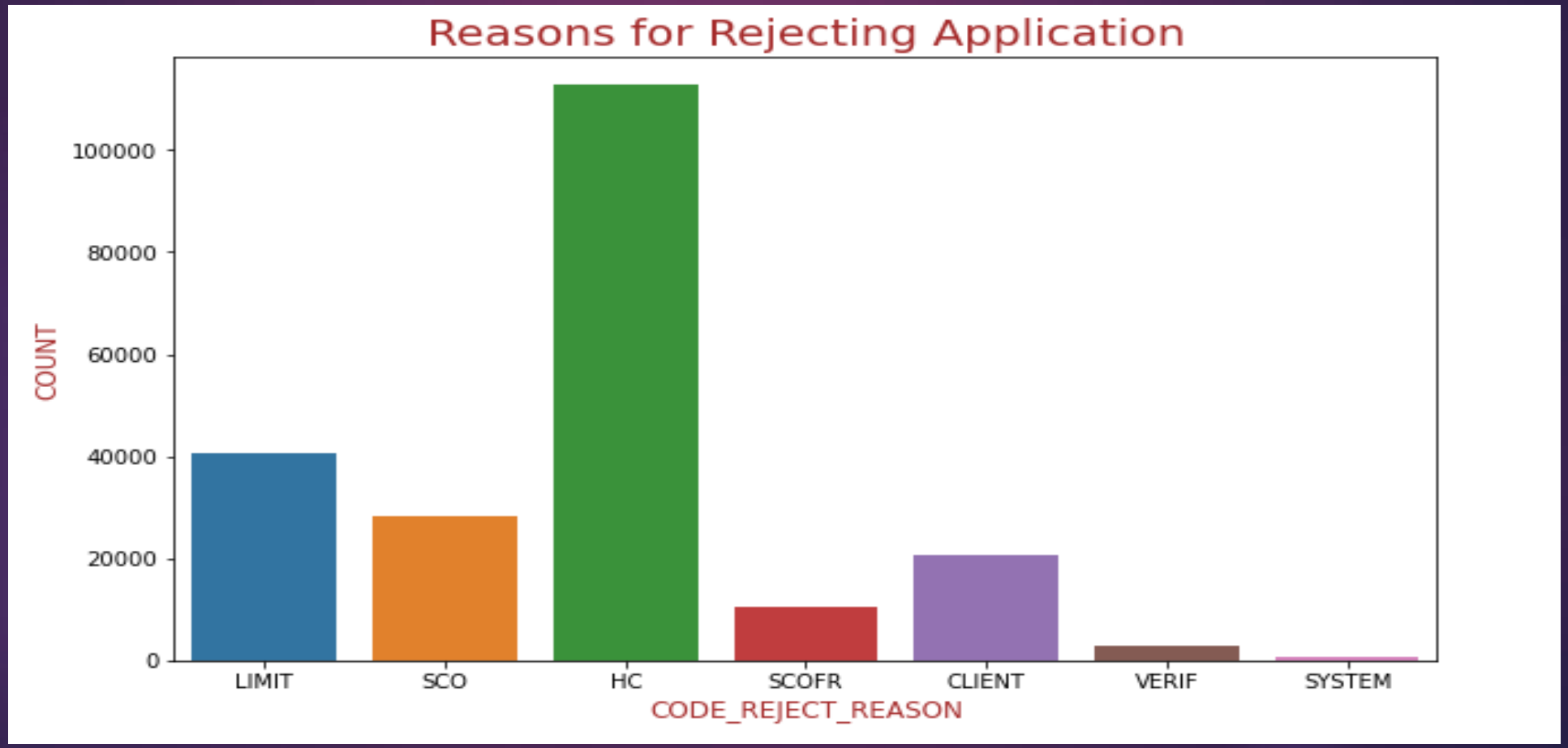
-Males with Higher Education seem to have Highest Credit above 0.6.
- Females with Academic Degree seems to have Highest Credit above 0.8.

We can see that there is a lot of variation in credit amount of Females in Academic Degree



Reasons for rejecting application.

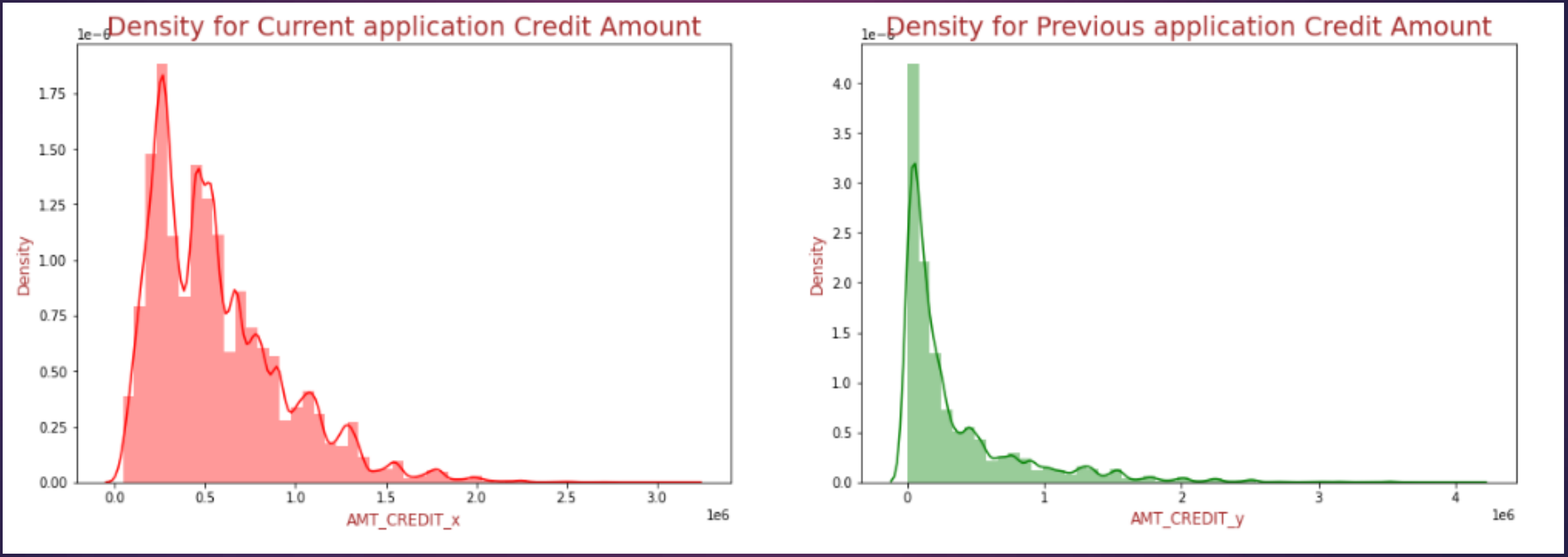
HC seems to be the reason most of the applications are rejected, above 100000 applications are rejected.



Observations.

Credit Amount between **2 lakhs – 6 lakhs** seem to have Highest Density for **Current Applications**.

Credit Amount between **2 lakhs – 4 lakhs** seem to have Highest Density for **Previous Applications**.

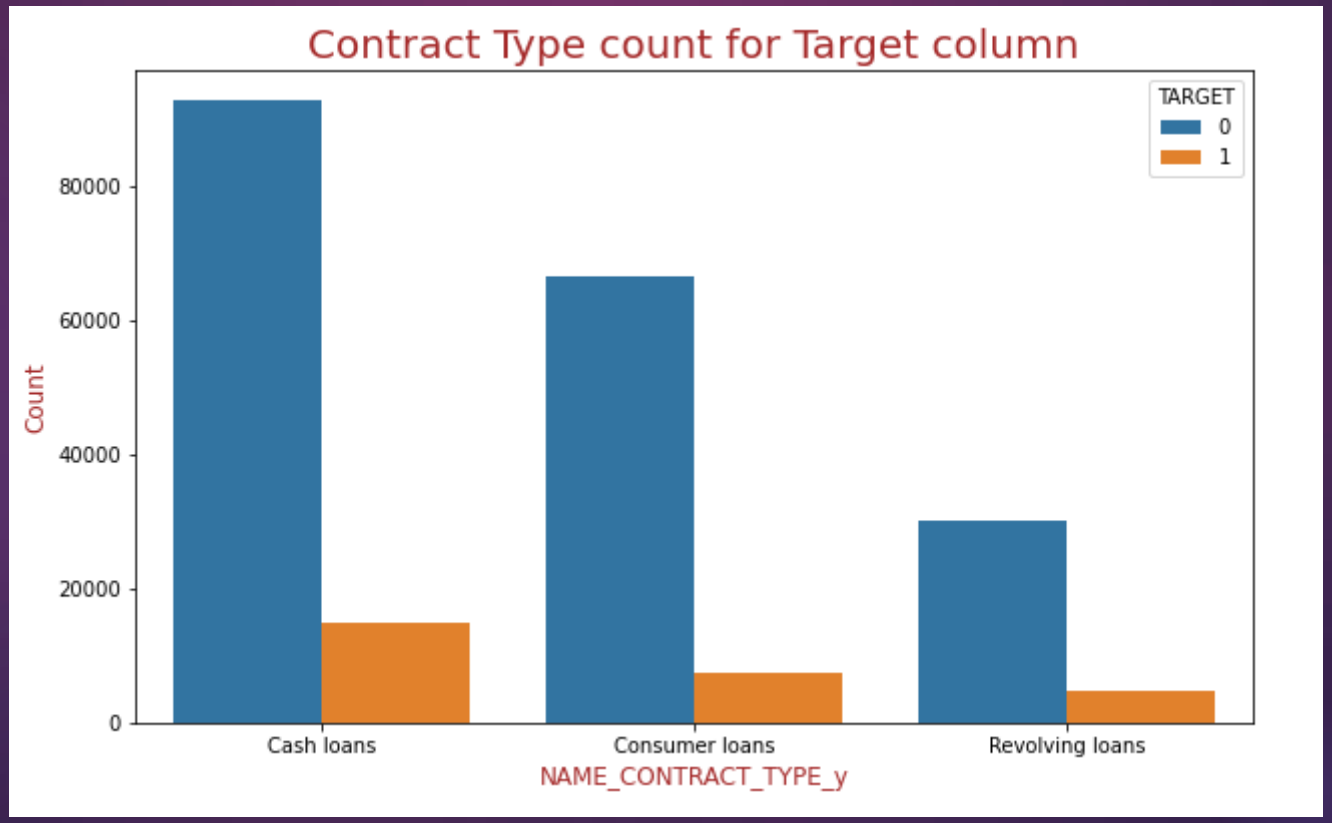


Contract Type count for Target Column.

For **Target 0**, Contract Type '**Cash Loans**' has the Highest Count above 80000.

For **Target 1**, Contract Type '**Cash Loans**' has the Highest Count below 20000.

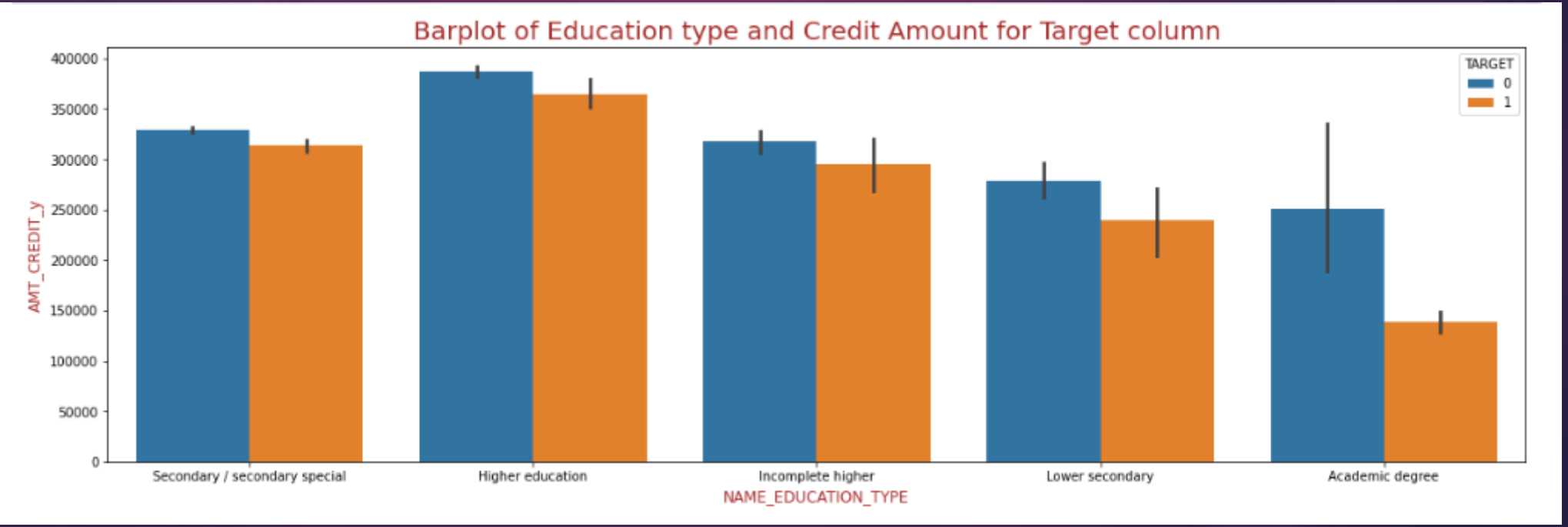
We can see that we can rely on cash loans because the defaulter ratio is very less.



Education Type and Previous Credit Amount for Target Column.

- For **Target 0**, Education Type '**Higher Education**' seems to have highest Credit amount above 3,50,000.
- For **Target 1**, Education Type '**Higher Education**' seems to have highest Credit amount above 3,50,000.

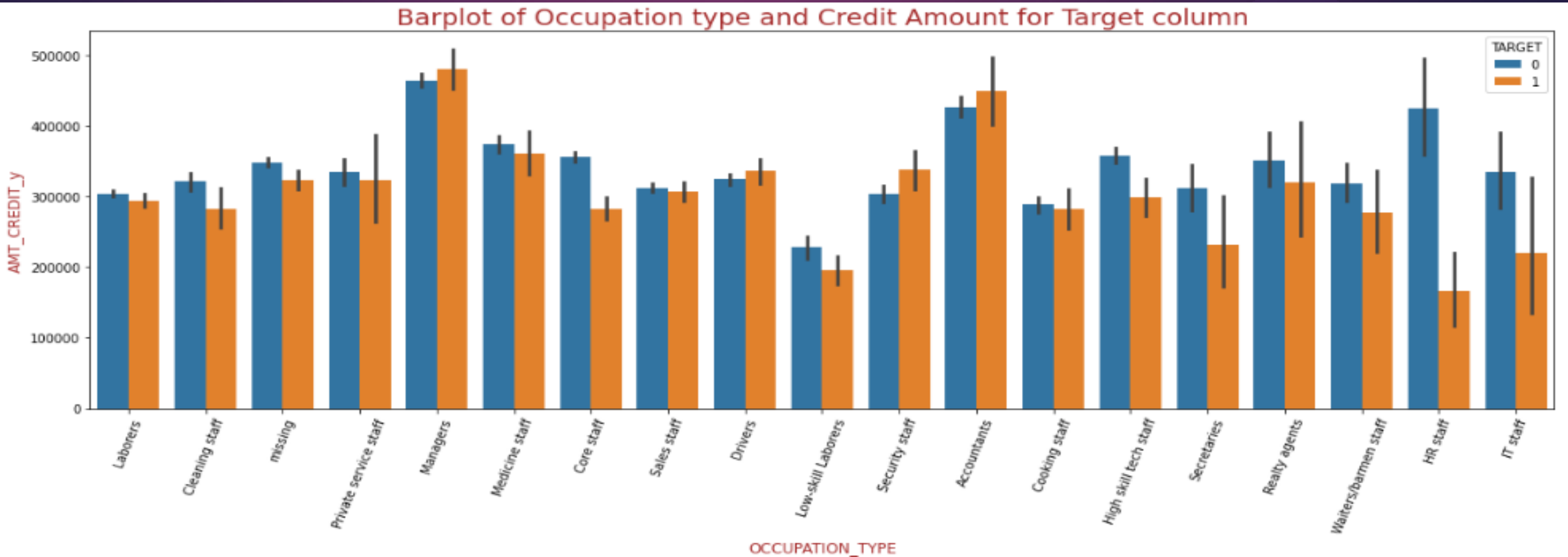
- We can see that we can rely on Academic degree pursuers to make less default.



Occupation Type and Credit Amount for Target Column.

-For **Target 0**, Occupation Type of "**HR staff**" and "**IT staff**" are **most reliable** ones to make less default in payment.

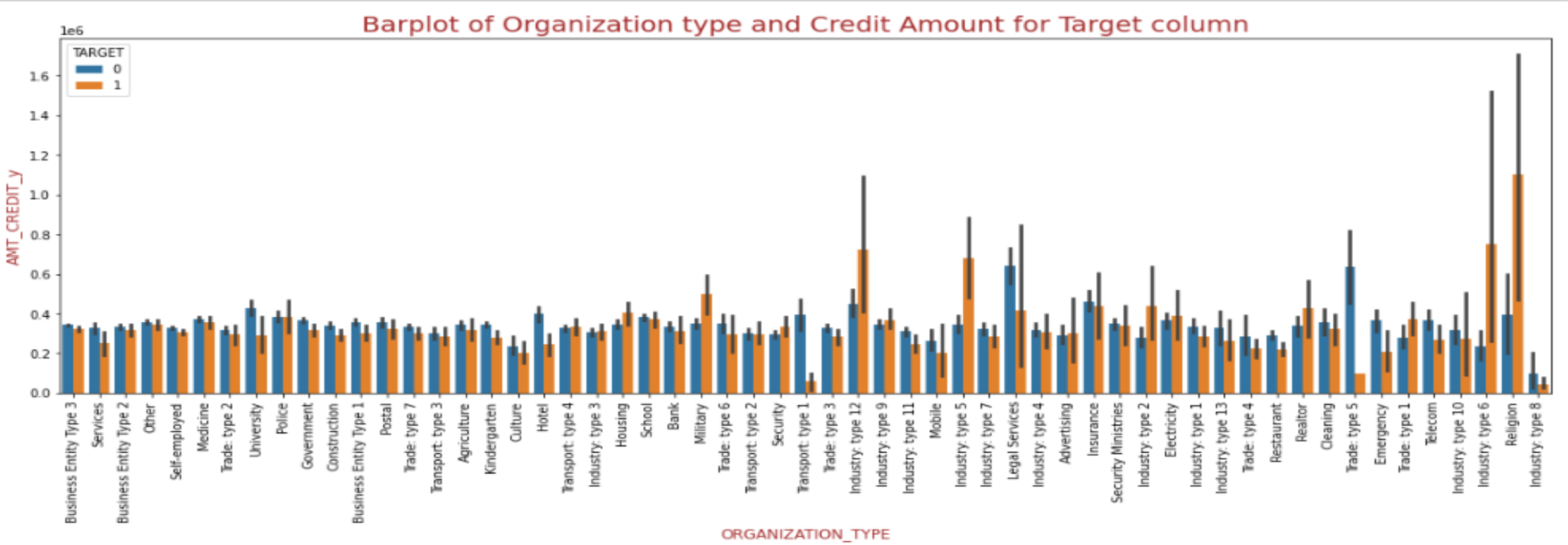
-For **Target 1**, Occupation Type of "**Managers**" and "**Accountants**" are more **likely to make a default** in payment.



Organization Type and Credit Amount for Target Column.

For **Target 0**, Organization Type of "**Trade Type 5**", "**Transport: Type 1**", "**Legal Services**" and "**Emergency**" seems to be **more reliable** in making **less default** in payment.

For **Target 1**, Organization Type of "**Religion**", "**Industry type 5**", "**Industry type 6**", "**Industry type12**" and "**Military**" are **more likely** to **make a default** in payment.



Conclusions

EDA for Banking Dataset revealed that –

- The percentage of Defaulters is **8.69%**.
- The data imbalance ratio is **10.5:1**
- **Most** Number of loans are taken for **Secondary Education** irrespective of Target Variable.
- **Maximum** number of people are **Laborers** and **minimum** are from **IT** and **HR** staff from the **Current Application**.
- From **Current Application Data** the **organization type** of **'Business Entity Type-3'** followed by **'Self Employed'** are the **most** number of applicants present in the dataset.
- **Client** more than **40** years of age is **less** likely to make a **default.**
- **Females** with Academic degree form the most number of defaulters.
- A lot variation can be observed in Credit amount of Females with academic degree.
- **HC** seems to be reason why most of the applications are rejected.
- We can rely on Cash Loans because the **defaulter ratio is very less**.
- We can rely on Academic Degree Pursuers **to make less default.**
- **For Target 0**, HR staff, IT staff are most **reliable** to make **less default** in payment.
- **For Target 1**, Managers, Accountants are **more likely** to make a **default** in payment.
- **For Target 0**, Organization Type, Transport Type Trade Type 5, Transport Type 1, Legal Services, Emergency are more reliable for **less default**.
- **For Target 1**, Organization Type, Religion, Industry type 5, industry type 6, industry type 12, military are more likely to make a **default.**
- For **Target 1**,We can see in the pair plot that when the **Credit amount increases the Goods price also increase**.
- For **Target 0**, We can see in the pair plot that when the **Credit amount increases the Goods price also increase.**