EXPLORATORY DATA ANALYSIS IN BFS

Presented by:

Shakti Singh

Introduction to Exploratory Data Analysis

- “While fraud reduction is a common goal for banks and financial institutions, analytics can be used to manage risk instead of simply detecting fraud” (Janaha, *Data Analytics in banking and Financial Services* 2023).
- “Analytics can be used to identify and rate individual customers who are at risk of fraud and then apply different levels of monitoring and verification to those accounts. Analyzing the risk of the accounts allows banks and financial institutions to know what to prioritize in their fraud detection efforts” (Janaha, *Data Analytics in banking and Financial Services* 2023).
- “With the rise of computing power and new analytical techniques, banks can now extract deeper and more valuable insights from their ever-growing mountains of data.” Moreover, “The recent dramatic increases in computing power have allowed banks to deploy advanced analytical techniques at an industrial scale” (Dash et al., *Risk analytics enters its prime* 2017).

Introduction to Exploratory Data Analysis

- “Machine-learning techniques, such as deep learning, random forest, and XGBoost, are now common at top risk-analytics departments” (Dash et al., *Risk analytics enters its prime* 2017).
- “The new tools radically improve banks’ decision models. And techniques such as natural-language processing and geospatial analysis expand the database from which banks can derive insights” (Dash et al., *Risk analytics enters its prime* 2017).
- “This means that risk teams can increasingly measure and mitigate risk more accurately and faster” (Dash et al., *Risk analytics enters its prime* 2017).
- “Banks that are fully exploiting these shifts are experiencing a “golden age” of risk analytics, capturing benefits in the accuracy and reach of their credit-risk models and in entirely new business models” (Dash et al., *Risk analytics enters its prime* 2017).

Introduction to the Case Study

- “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” (upGrad, *Credit EDA Assignment* 2023).
- “When the company receives a loan application, the company has to decide for loan approval based on the applicant’s profile” (upGrad, *Credit EDA Assignment* 2023).
- “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.” (upGrad, *Credit EDA Assignment* 2023).
- Problem Statement: “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 utilise this knowledge for its portfolio and risk assessment” (upGrad, *Credit EDA Assignment* 2023).

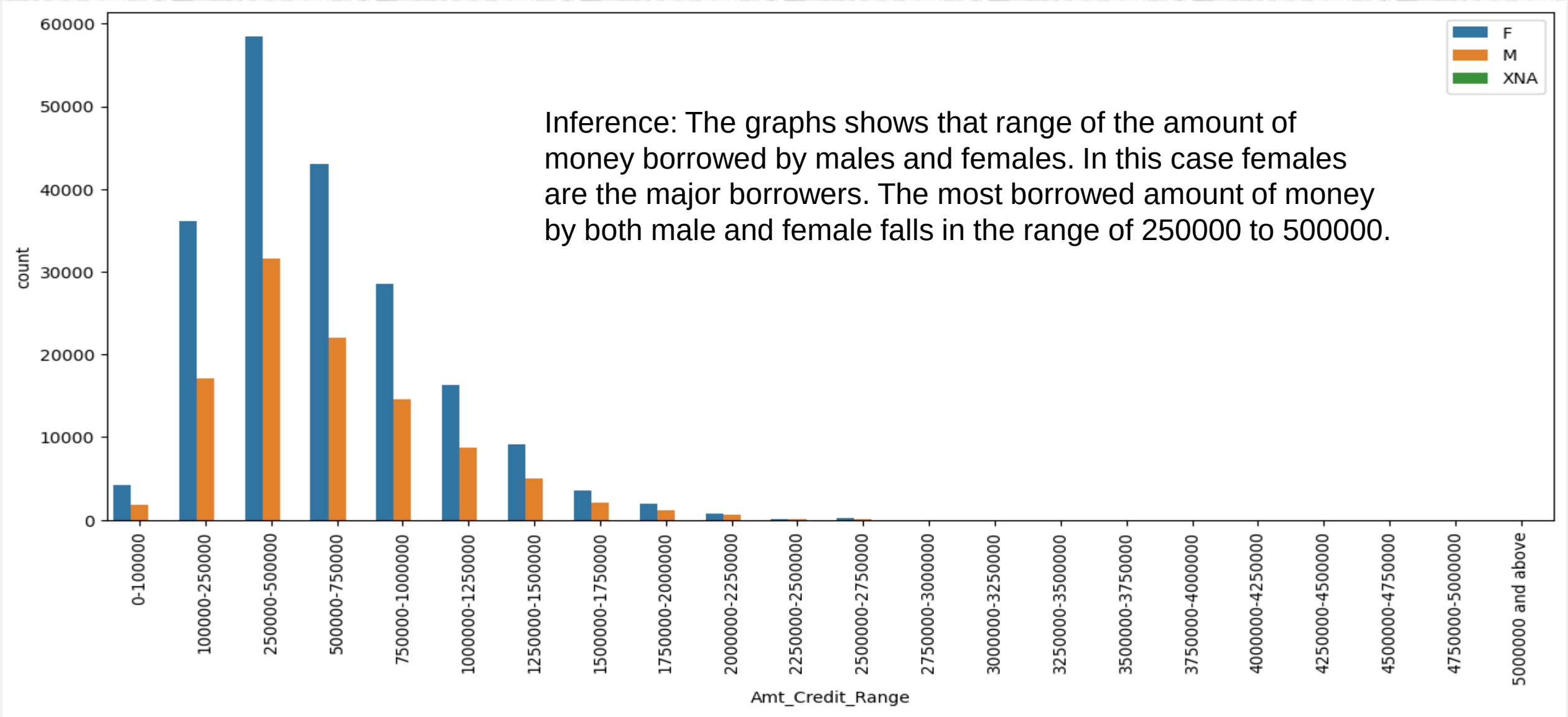
Procedure followed from the Beginning

- All the warning and necessary libraries were imported such as import warnings, numpy, pandas, matplotlib.pyplot and seaborn.
- Then the CSV file called ‘application_data.csv’ was imported along with CSV file ‘previous_application.csv’.
- ‘application_data.csv’ was converted to a pandas data frame and stored in a variable and same was done for ‘previous_application.csv’.
- First analysis was done on current application or application data set and then later it was done on the previous application dataset.
- File was read using the head( ) and tail( ) function
- After that the columns were checked, and info was printed.
- Then the data set was checked for null values.

Null Value Treatment

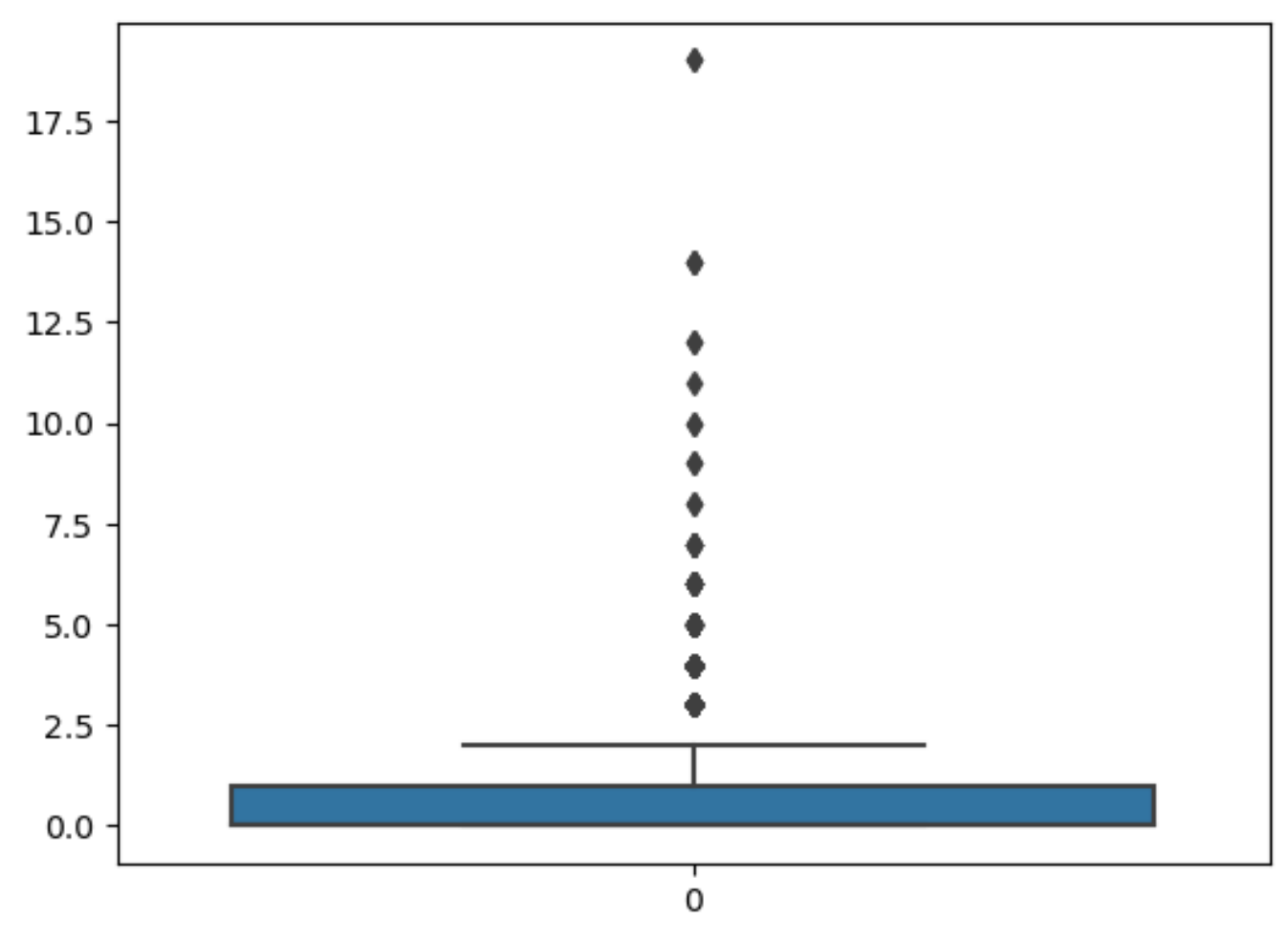
- Then the null values were converted to a percentage format.
- Columns that were having null values more than 35 percent were removed from the dataset.
- Then null values that were less than 13 to 20 percent that were missing in the dataset were identified labeled as minor_missing_values.
- Then the unique values of the columns were identified using the nuinque ( ) function and was determined that they were of categorical nature.
- Some unnecessary columns were dropped that were not needed for the analysis.
- Then data type correction was performed making columns that are of object type or categorical or numerical type and were converted accordingly.
- Columns having negative values were converted to positive where it was required.

Data Binning



Inference: The graphs shows that range of the amount of money borrowed by males and females. In this case females are the major borrowers. The most borrowed amount of money by both male and female falls in the range of 250000 to 500000.

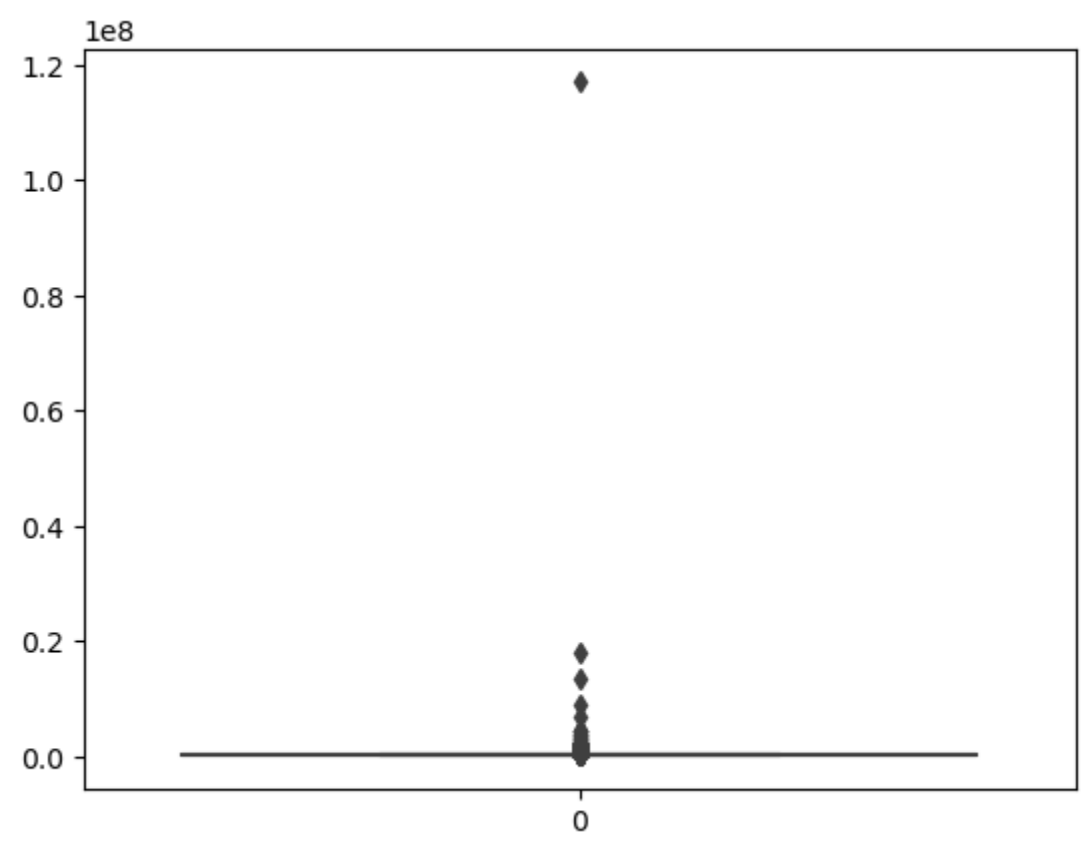
Outlier Identification and Suggested Treatment



Box Plot for **CNT_CHILDREN**

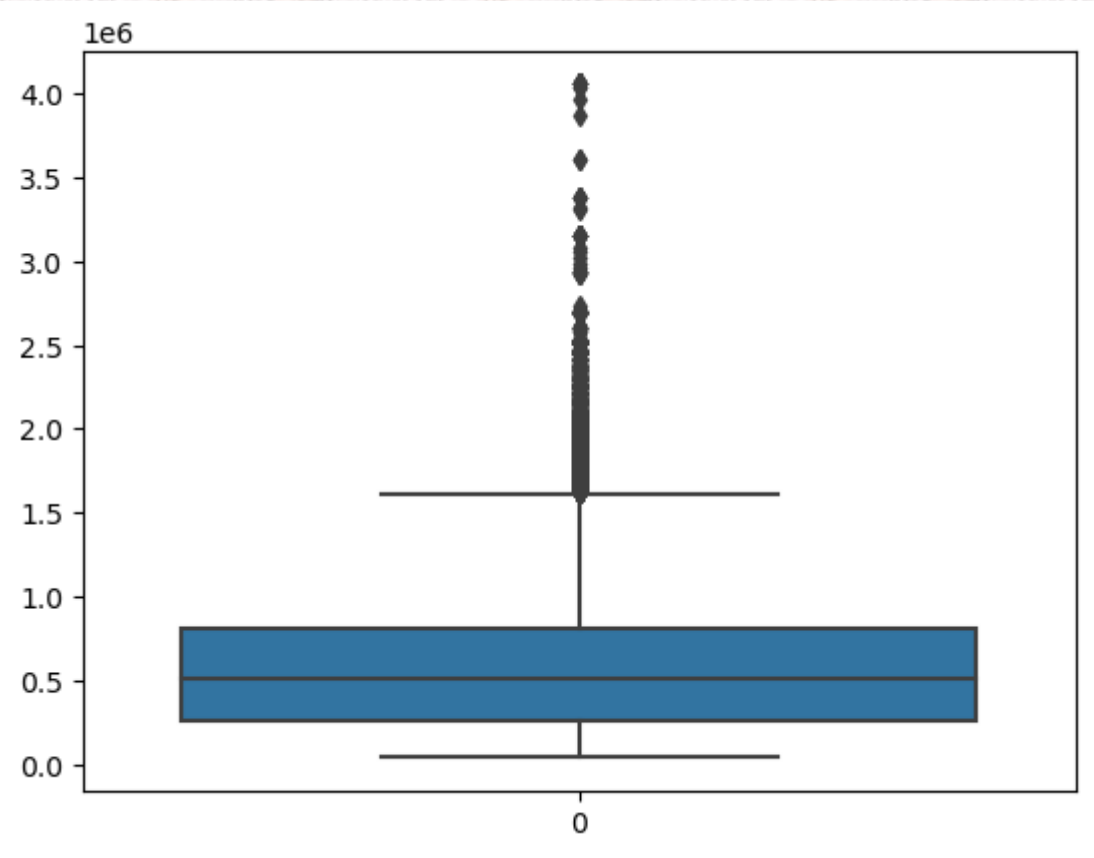
Inference: We can clearly see that there are many outliers present, we can use median or mode to impute the outliers because mode or median will give an accurate representation of the whole dataset.

Box Plot for AMT_INCOME_TOTAL



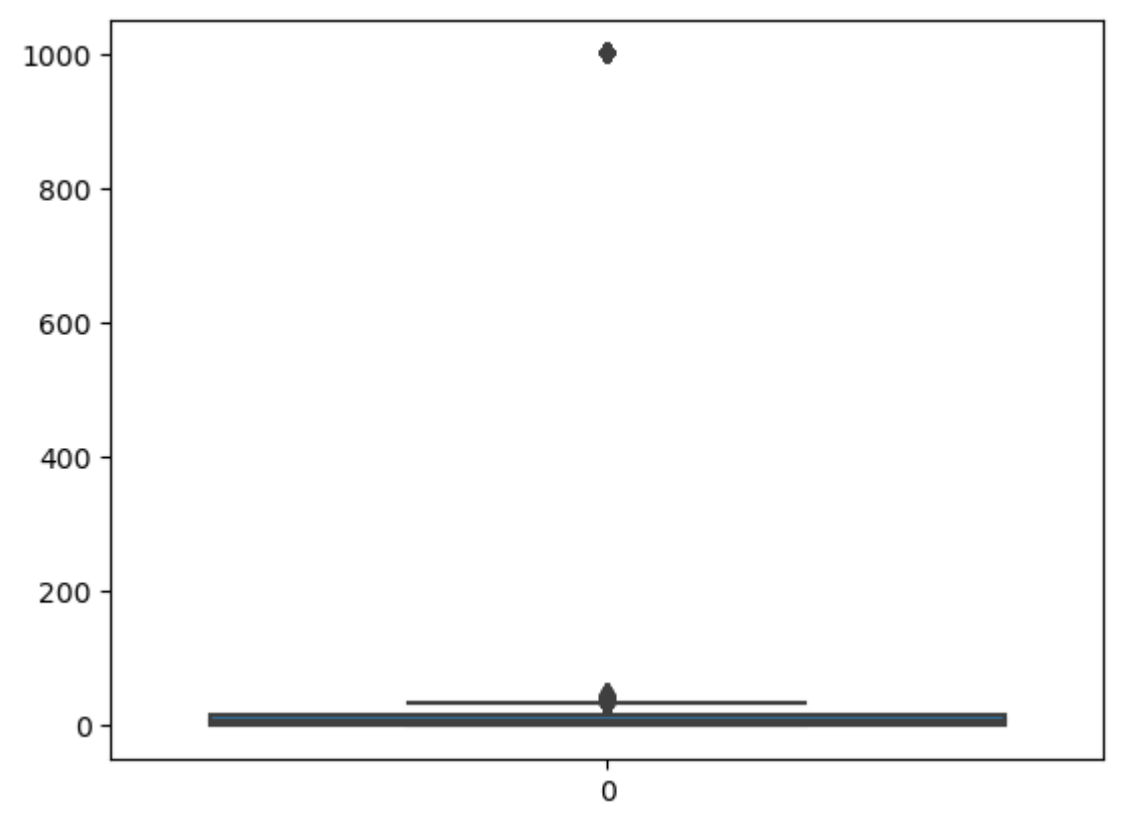
Inference: We can observe that there are outliers present here as well. We can use median or mode to impute the outliers because mode or median will give an accurate representation of the whole dataset

Box Plot for AMT_CREDIT



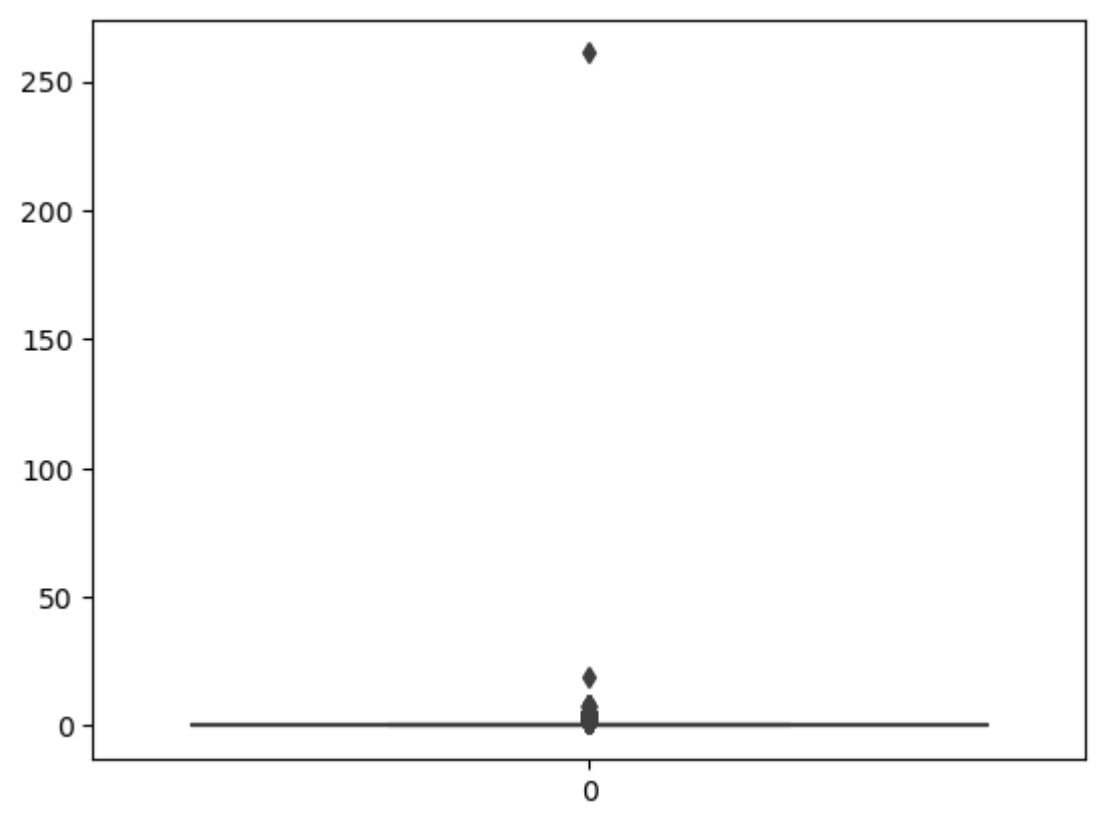
Inference: There are outliers present as well. We can use median or mode to impute the outliers because mode or median will give an accurate representation of the whole dataset

Box Plot for YEARS_EMPLOYED



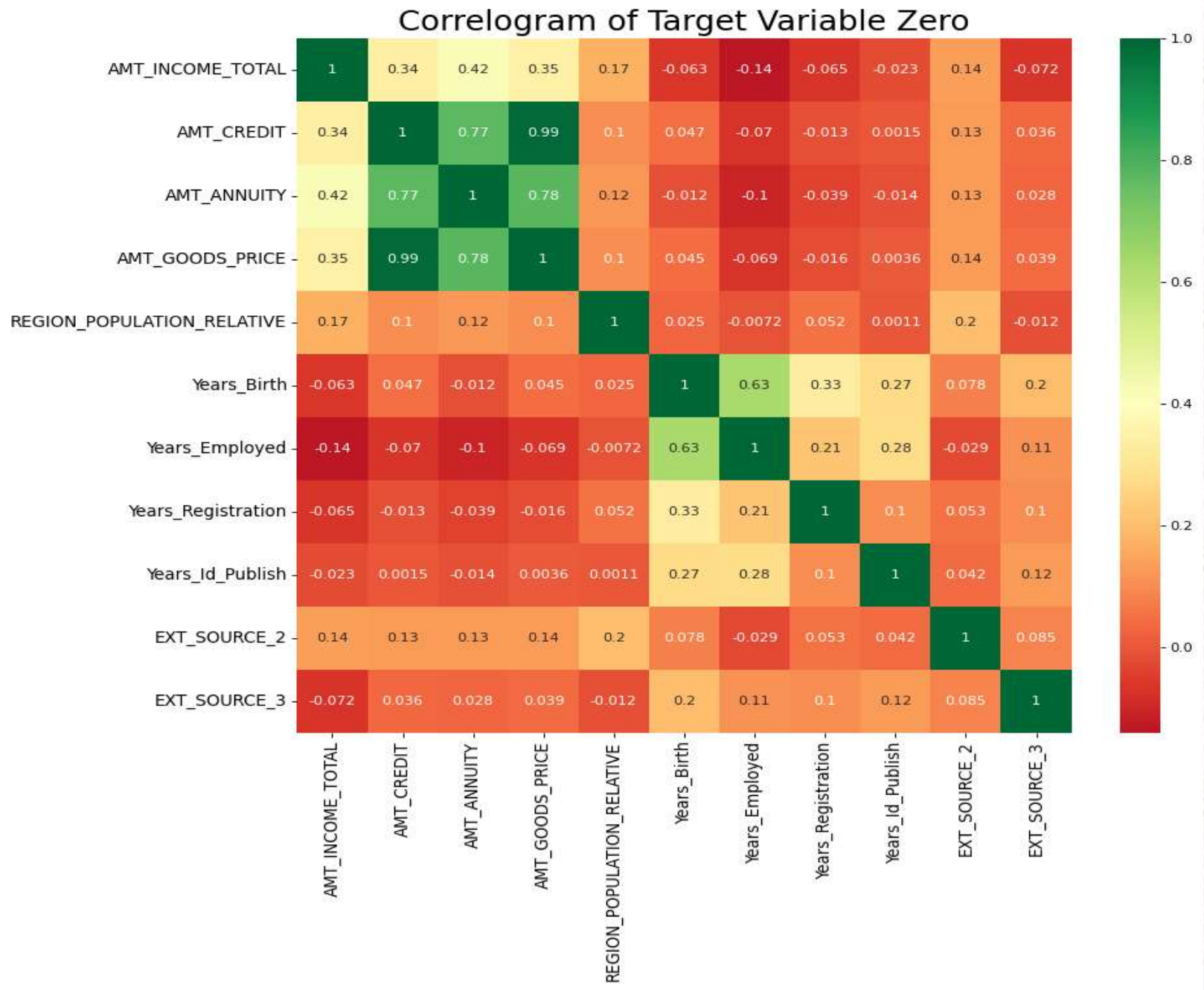
Inference: We can see that there is an outlier here as well which is at 1000 years realistically it is not possible. we can use median or mode to impute the outliers because mode or median will give an accurate representation of the whole dataset

Box Plot for AMT_REQUIRED_CREDIT_BUREAU_QRT



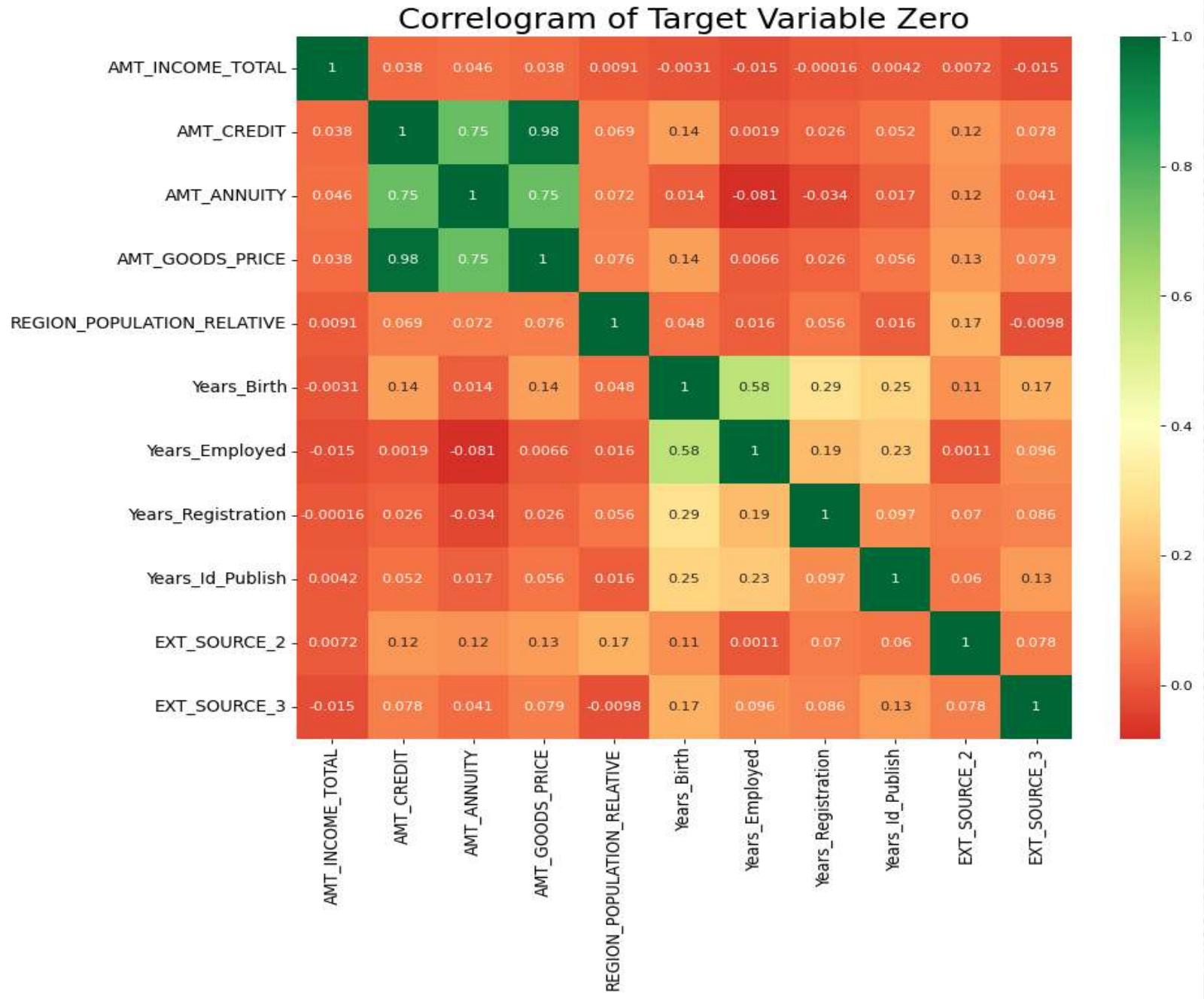
Inference: There is an outlier here as well. We can use median or mode to impute the outliers because mode or median will give an accurate representation of the whole dataset

Heat Map of Target Variable Zero



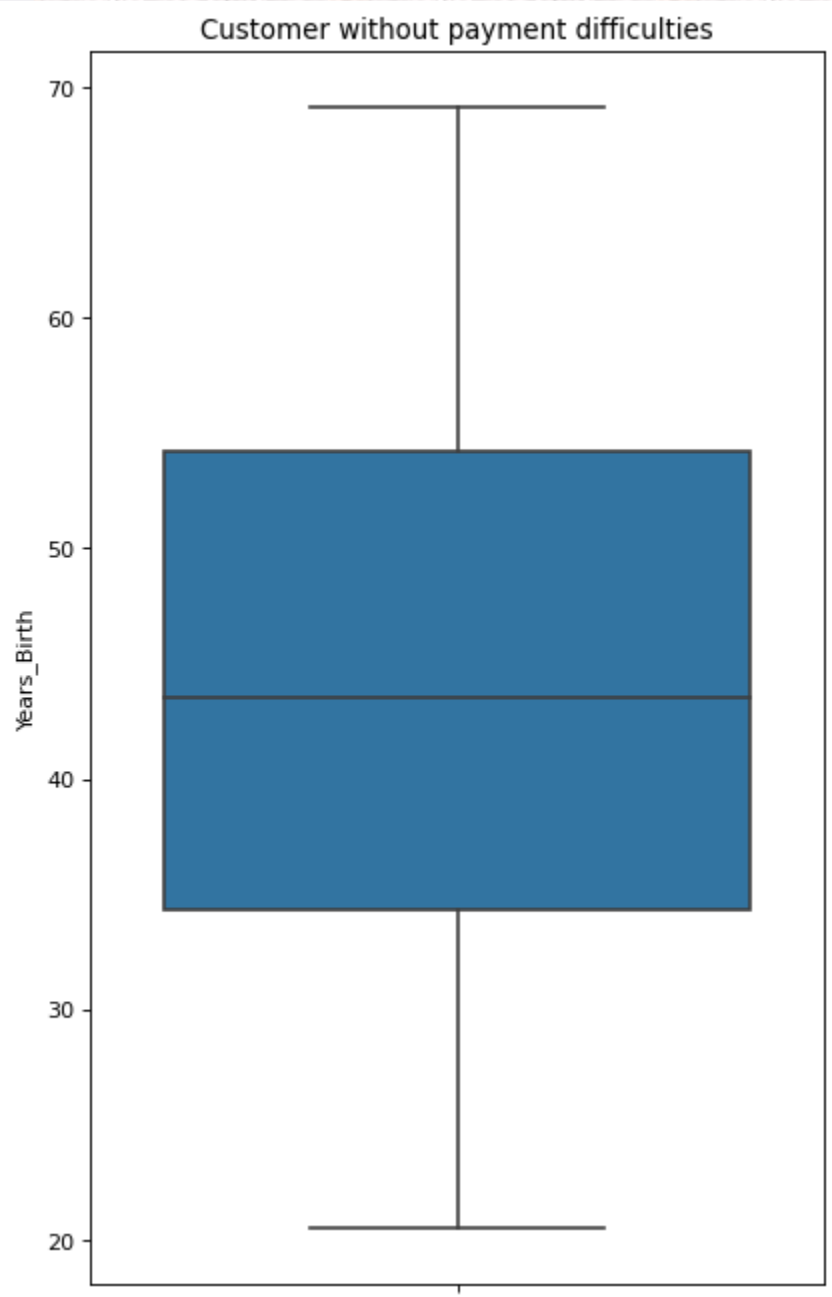
Inference: We can see clearly that our upper matrix is symmetrical and the amount credit, and amount goods price have the highest positive correlation 0.99 which means they are positively linearly correlated. However, the ext. source 3 and amount total income have a negative correlation. Additionally, the amount good price and amount annuity are also presenting with a good correlation.

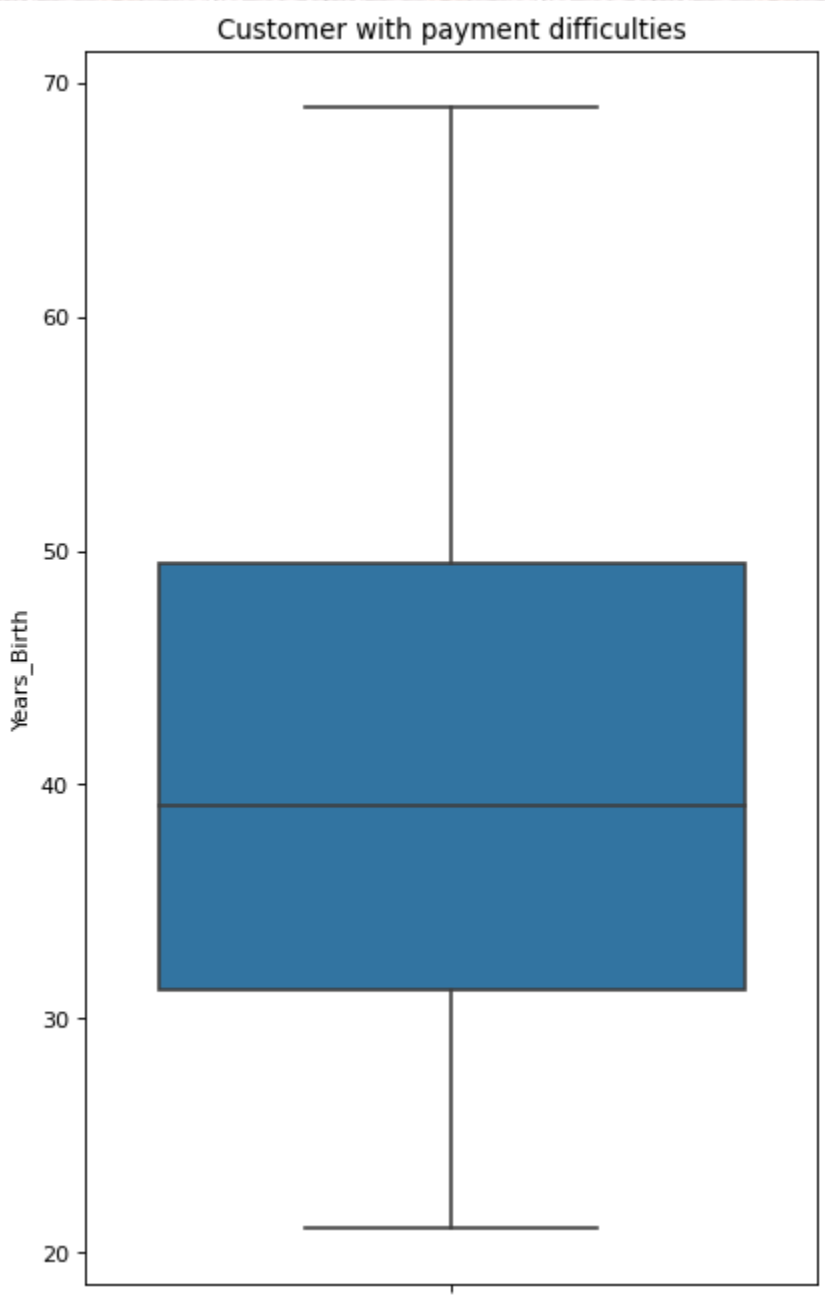
Heat Map of Target Variable Zero



Inference: We can see from the above heat map that the amount credit and amount goods price have highest positive correlation which shows the positive linear relation of the two variables. However, ext. source 3 and amount income total have a negative correlation. This just shows that two variables that have negative correlation are not necessarily dependent on each other.

Numerical Univariate Analysis

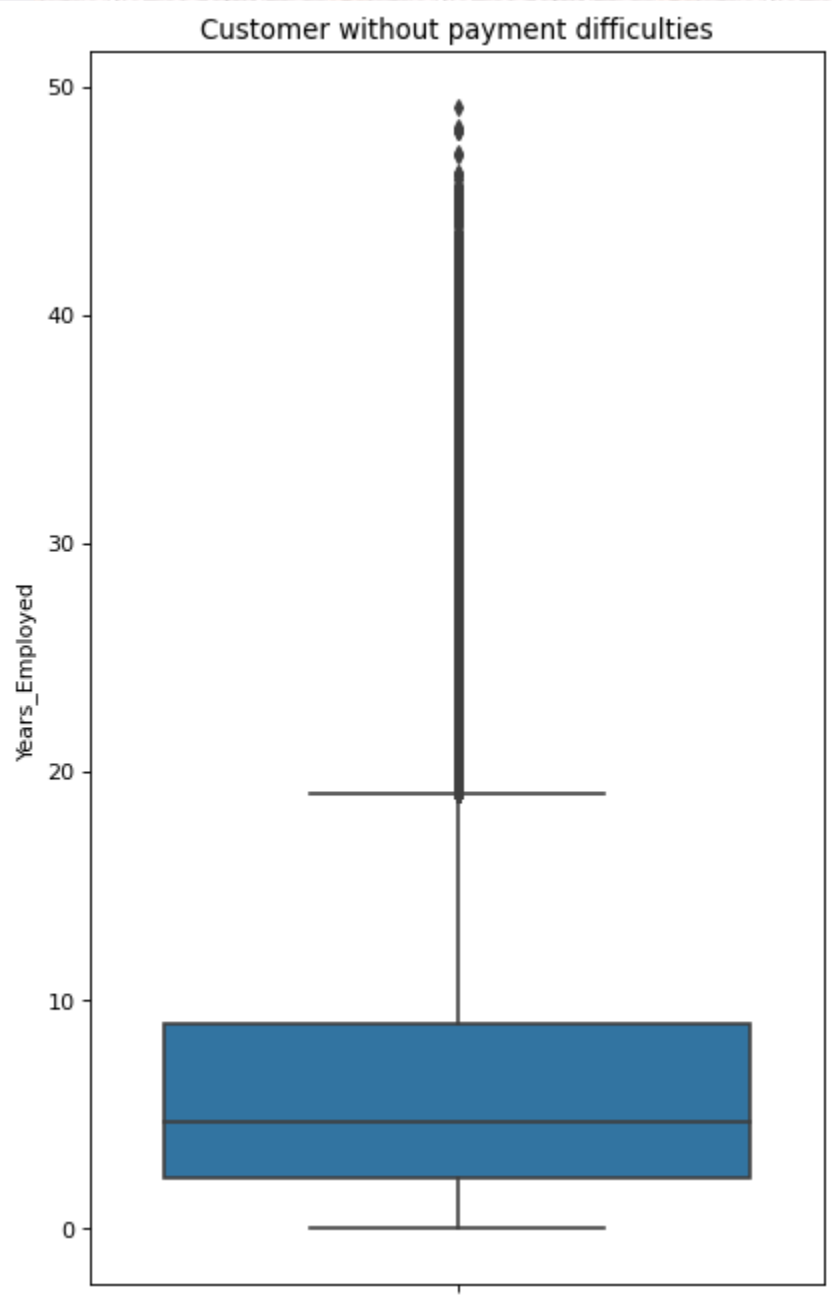


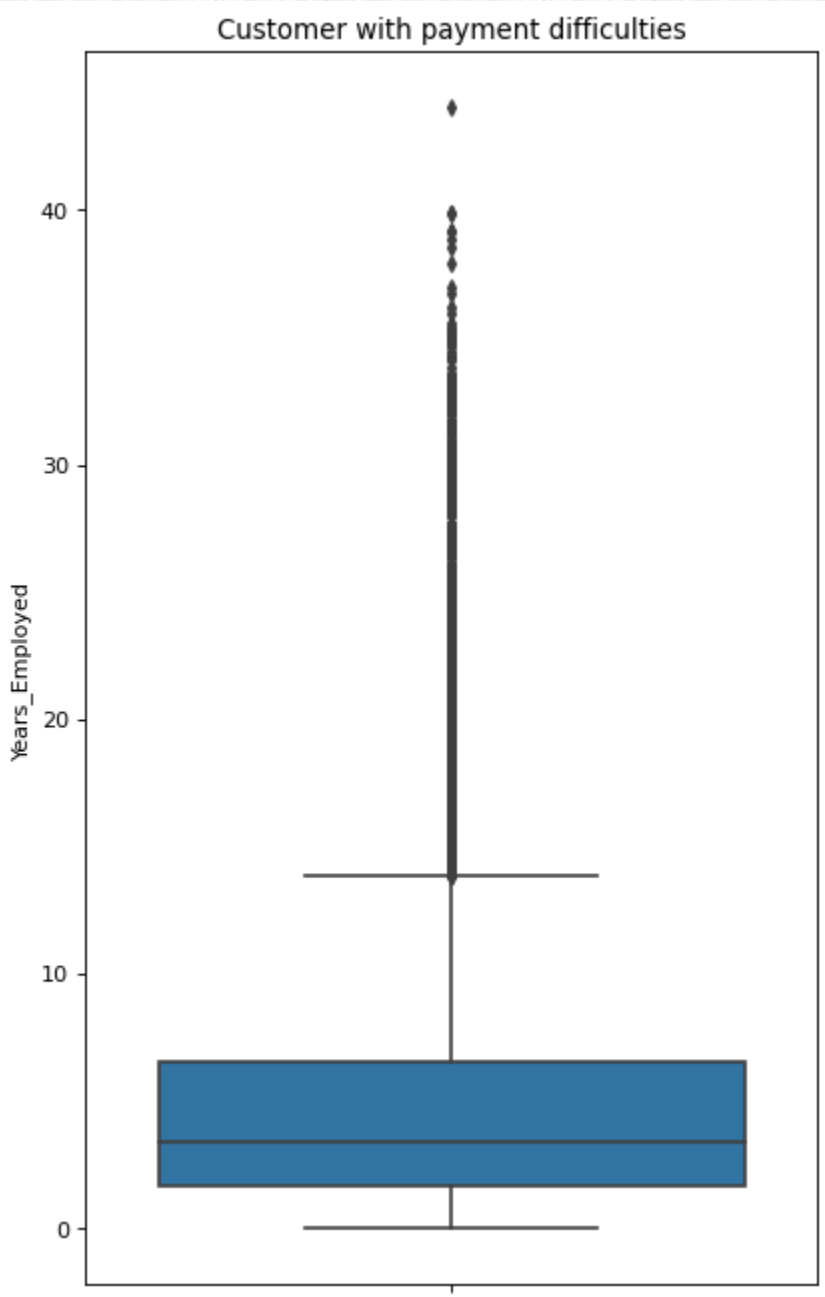


Box Plot comparing YEARS_BIRTH by using clients from dataset of Target 0 and 1

Inference: From the box plot we observe that clients without payment difficulties are of the age 35 to 55 years , And clients with payment difficulties are between ages of 32 to 48 years. Also, the clients that do not have difficulty paying have a higher IQR.

Numerical Univariate Analysis

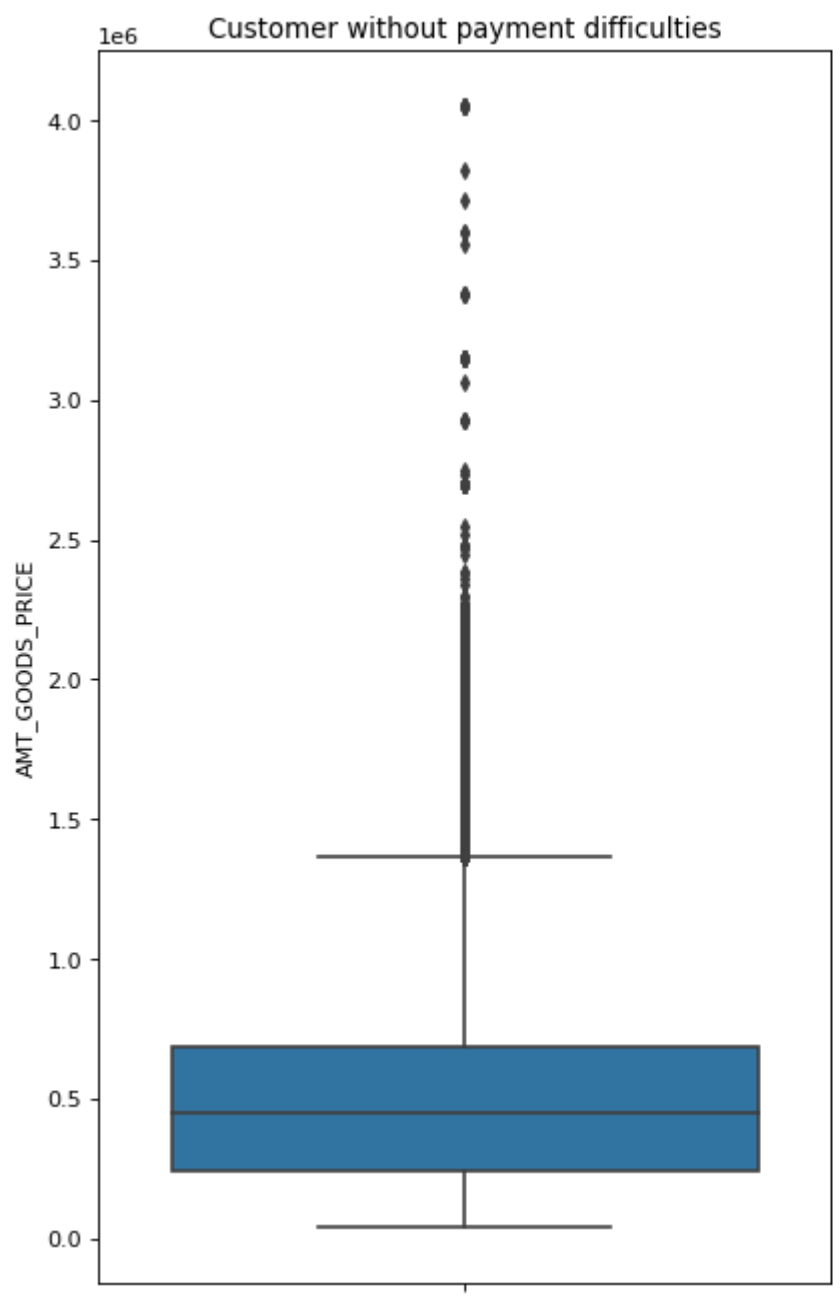


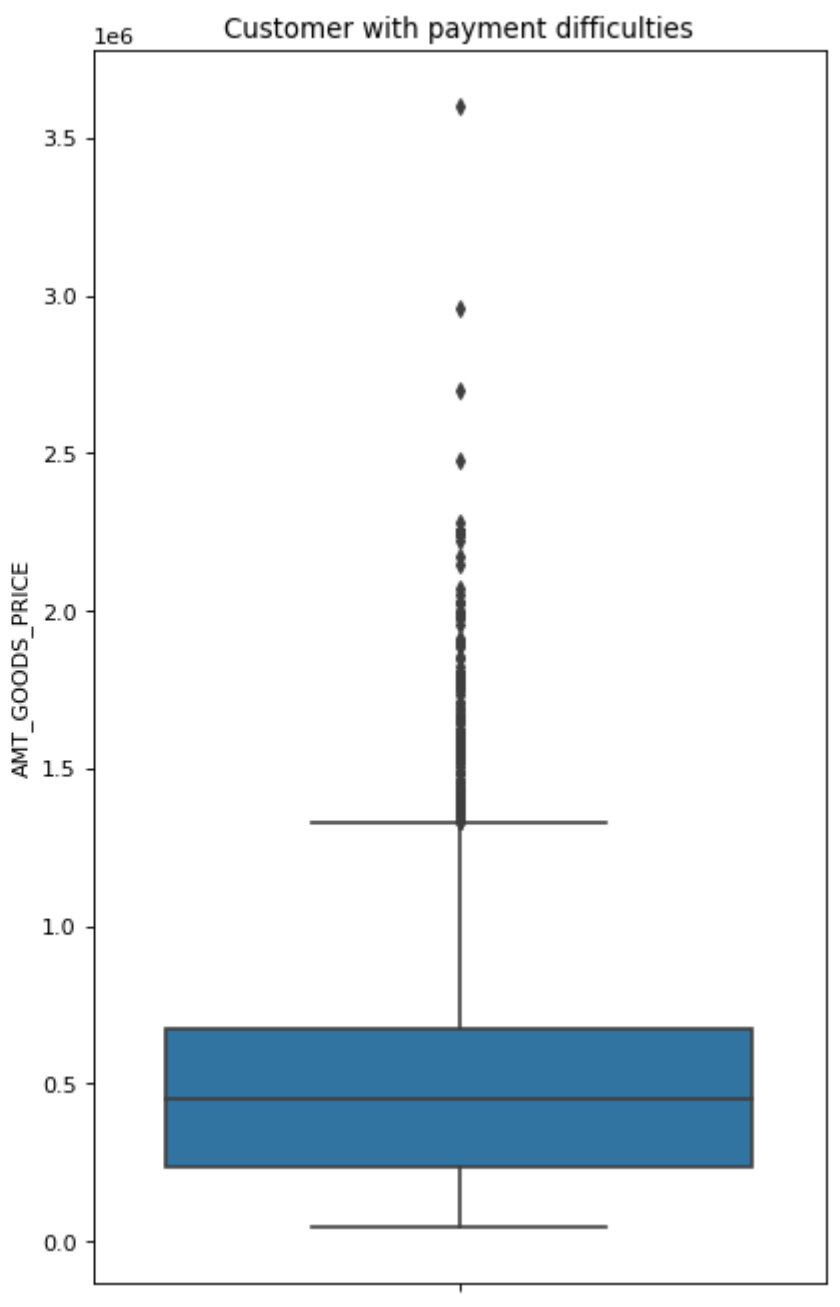


Box Plot comparing YEARS_EMPLOYED by using clients from dataset of Target 0 and 1

Inference: The box plot shows that clients without payment difficulties have been employed between 2 to 9 years whereas, clients with payment difficulties have been employed between 2 to 7 years.. Also, the clients that do not have difficulty paying have a higher IQR.

Numerical Univariate Analysis

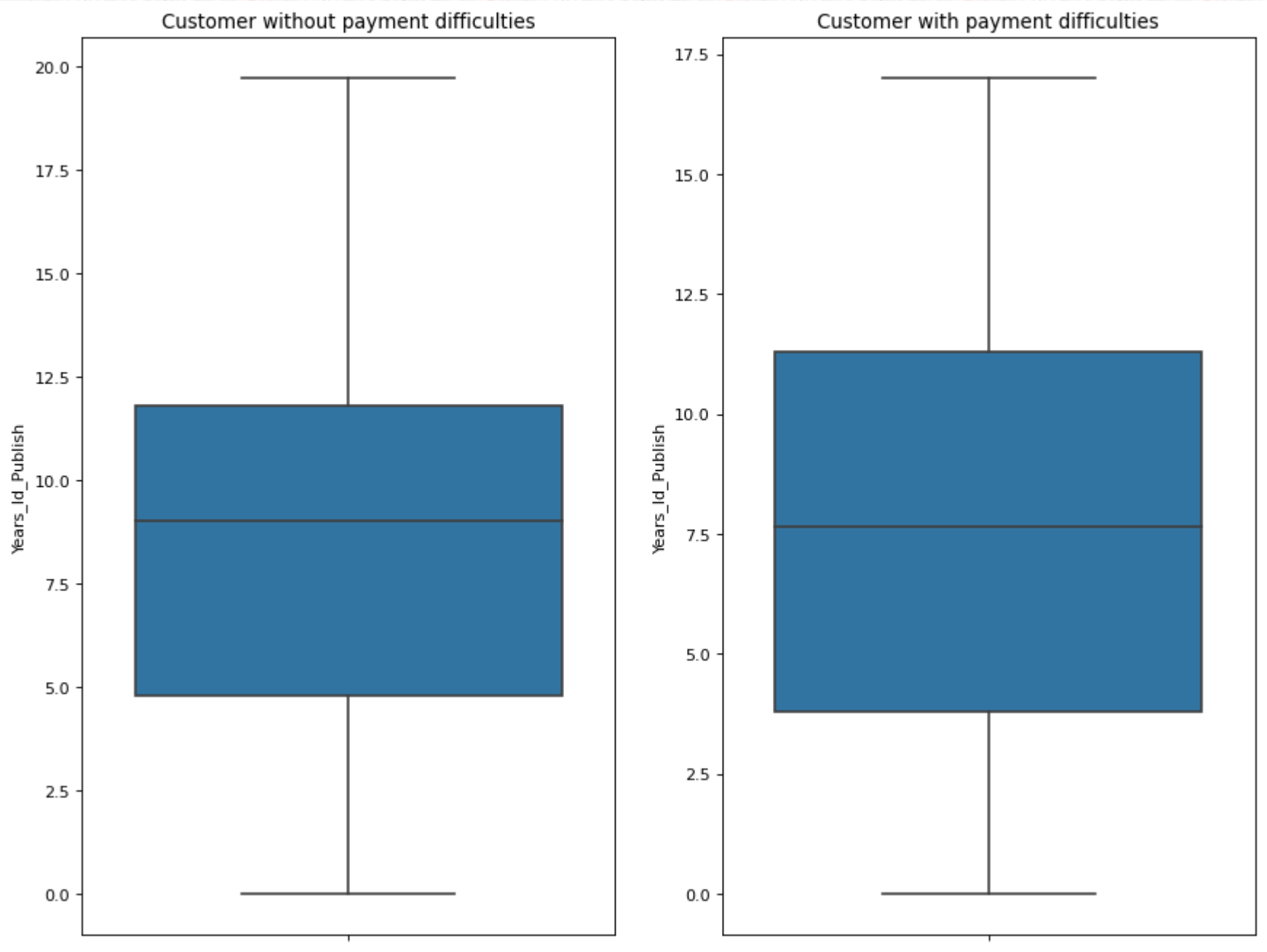




Box Plot comparing AMT_GOODS_PRICE: clients from dataset of Target 0 and 1

We can see that clients who have not had difficulty paying the loans have been granted a lower price when compared to the clients that have had difficulties in making the payments. To put it numbers, clients that didn't have difficulties they have been granted the loan at a low cost of .25 to 1.25 whereas for clients that have had difficulties they have been granted also similar rates of interest. Now this can be a result of data not being clean as expected but generally people with payment difficulties are not granted loans at such low percentage and if they are granted, they are usually higher

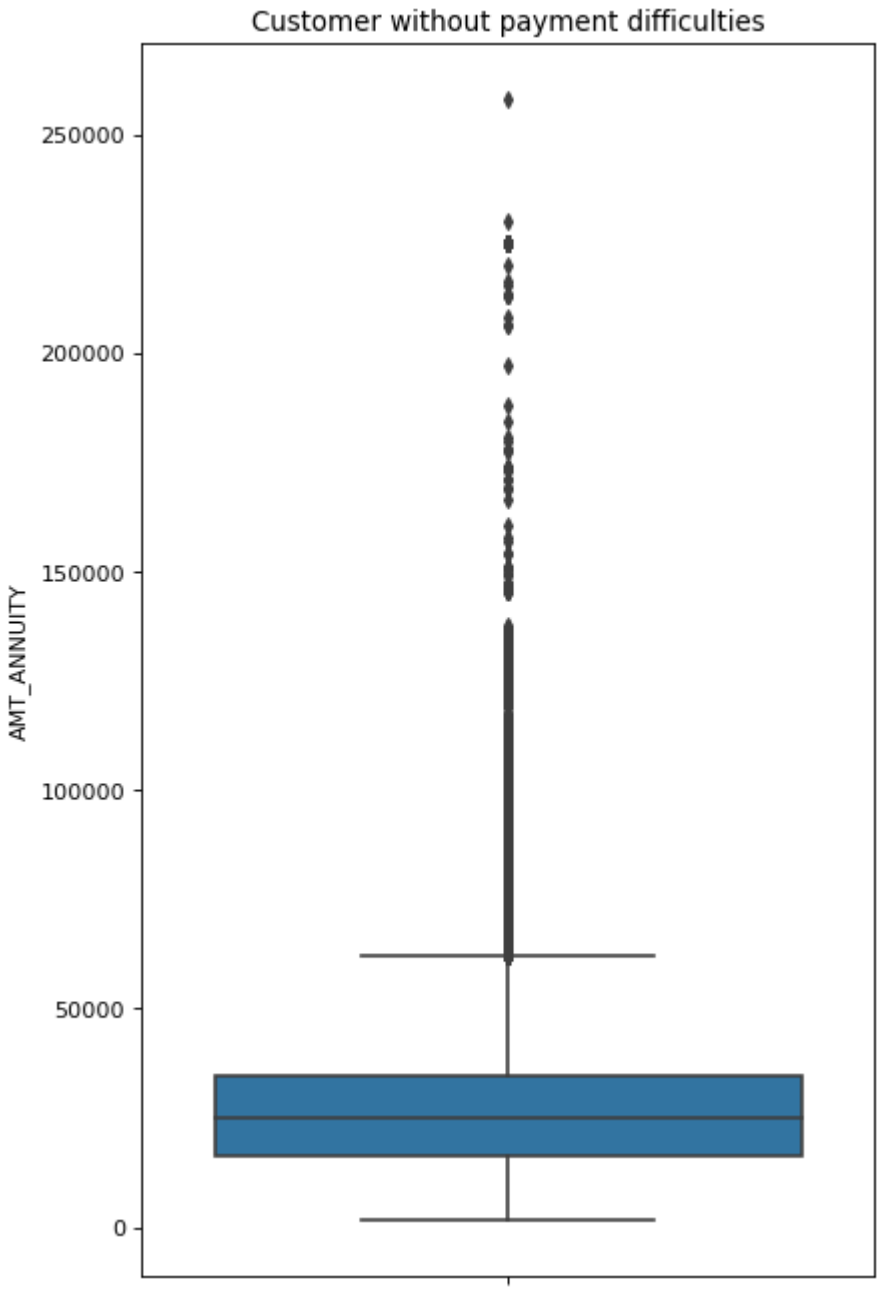
Numerical Univariate Analysis

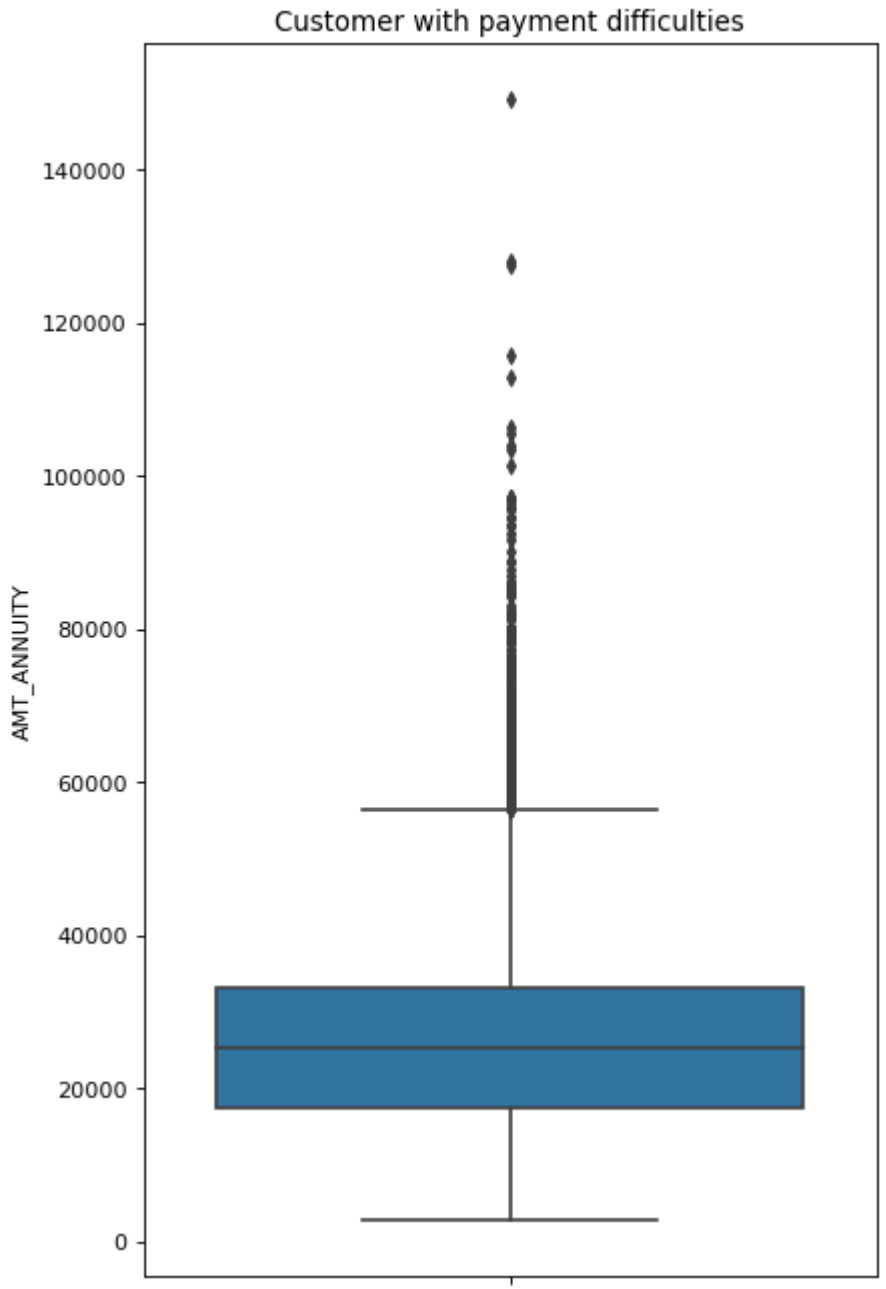


Box Plot comparing YEARS_ID_PLUBLISHED and clients from dataset of Target 0 and 1

It is an interesting observation here one can see that clients that have been paying regularly they have moved more frequently and clients that have had problems with payment have moved less often.

Numerical Univariate Analysis

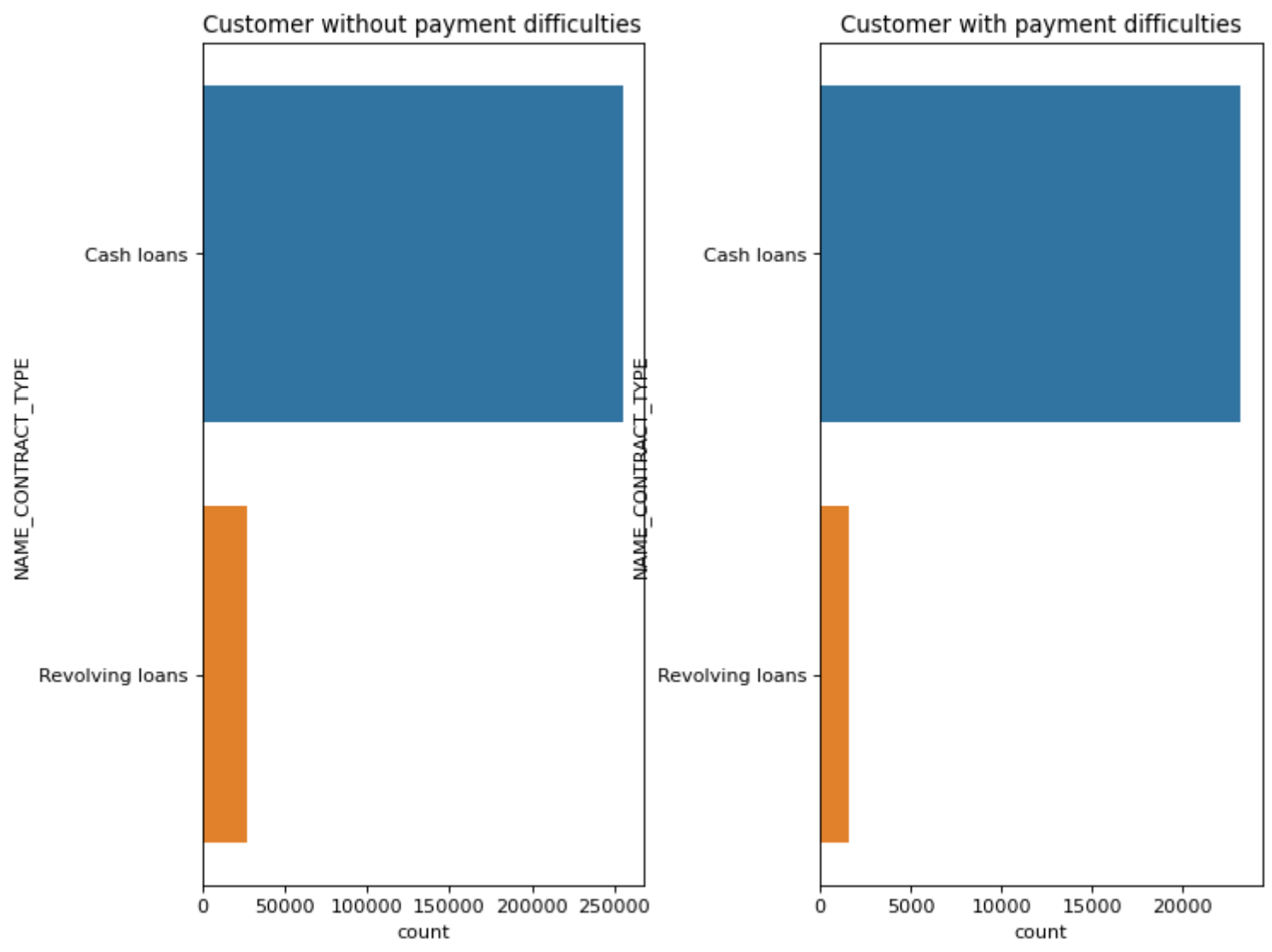




Box Plot comparing AMT_ANNUITY and clients from dataset of Target 0 and 1

We can clearly observe that both clients with and without difficulties have almost the same Annuity but there is a difference in their IQR, clients that have had difficulties have a higher IQR when compared to clients that haven't had problems.

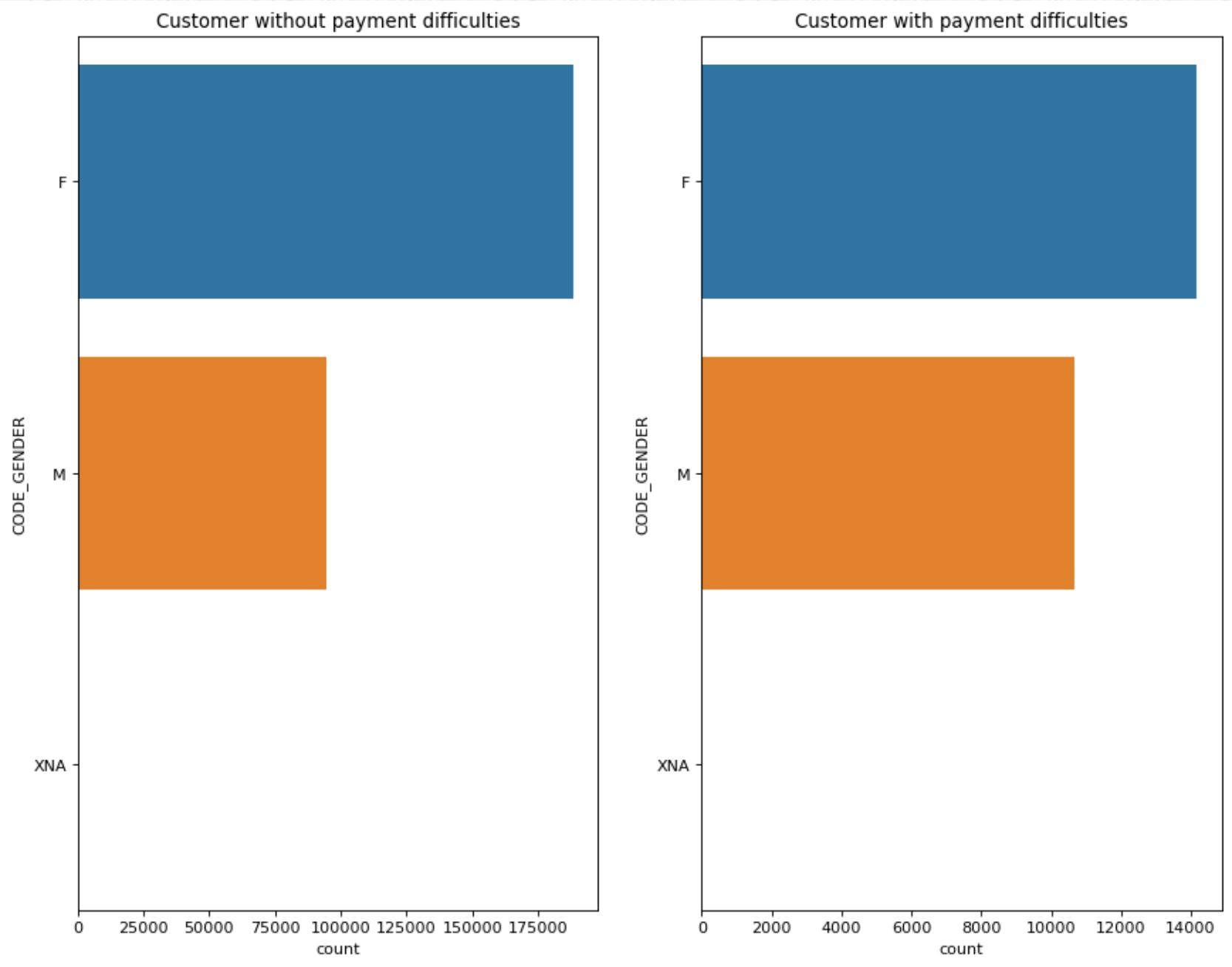
Categorical Univariate Analysis



Count Plot comparing CONTRATC_TYPE and clients from dataset of Target 0 and 1

Both the client base people with difficulty or without have taken more cash loans rather than the revolving loans. These depend on many factors such as the income of the populous that have the greatest number of people in that category or their education.

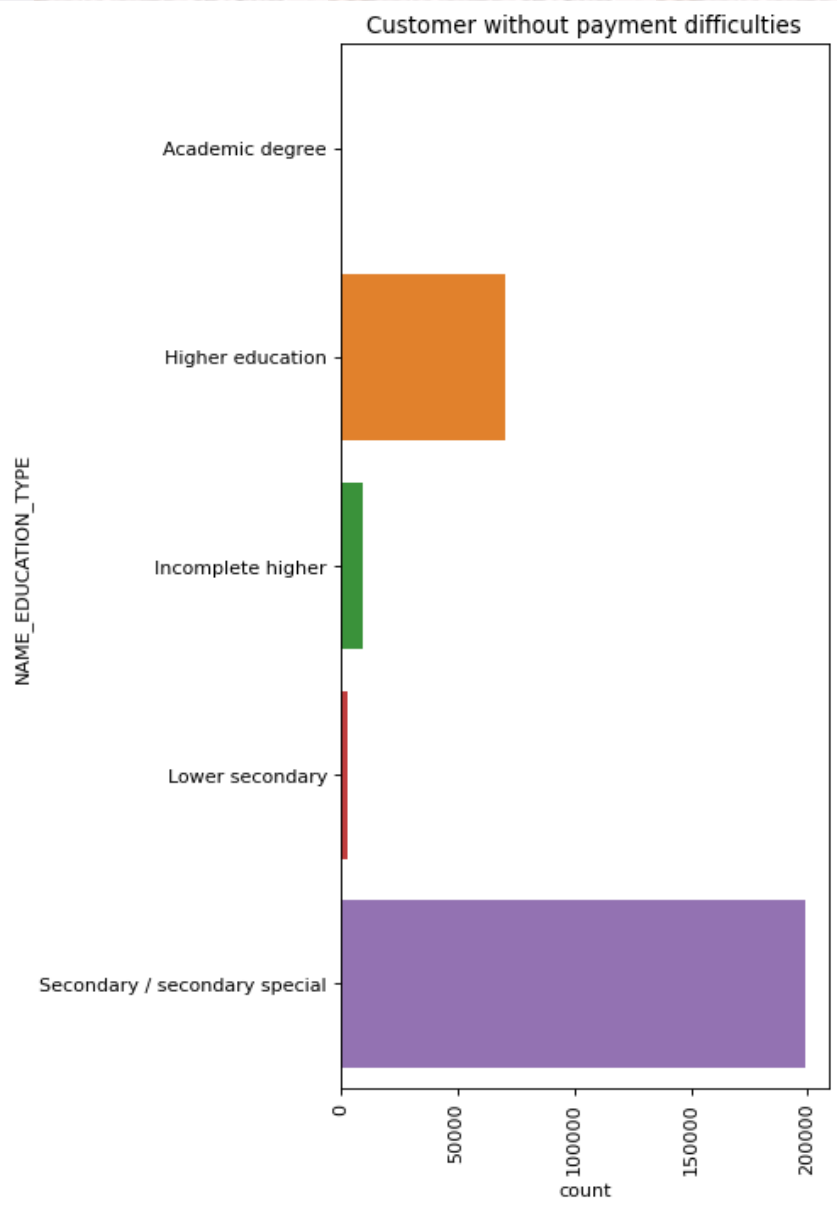
Categorical Univariate Analysis

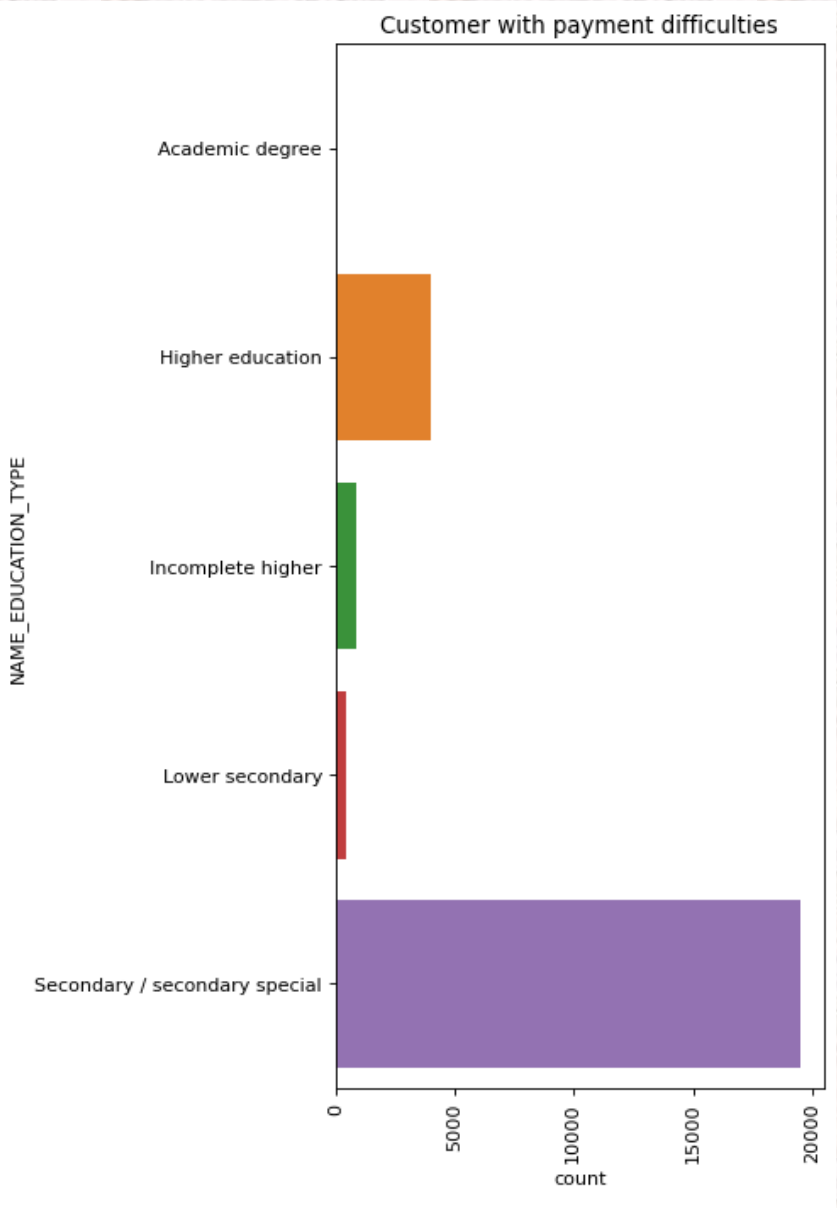


Count Plot comparing CONTRATC_TYPE and clients from dataset of Target 0 and 1

An interesting observation, Males in both the categories have defaulted less than the females. However, the number of females that have taken a loan is also more than males.

Categorical Univariate Analysis

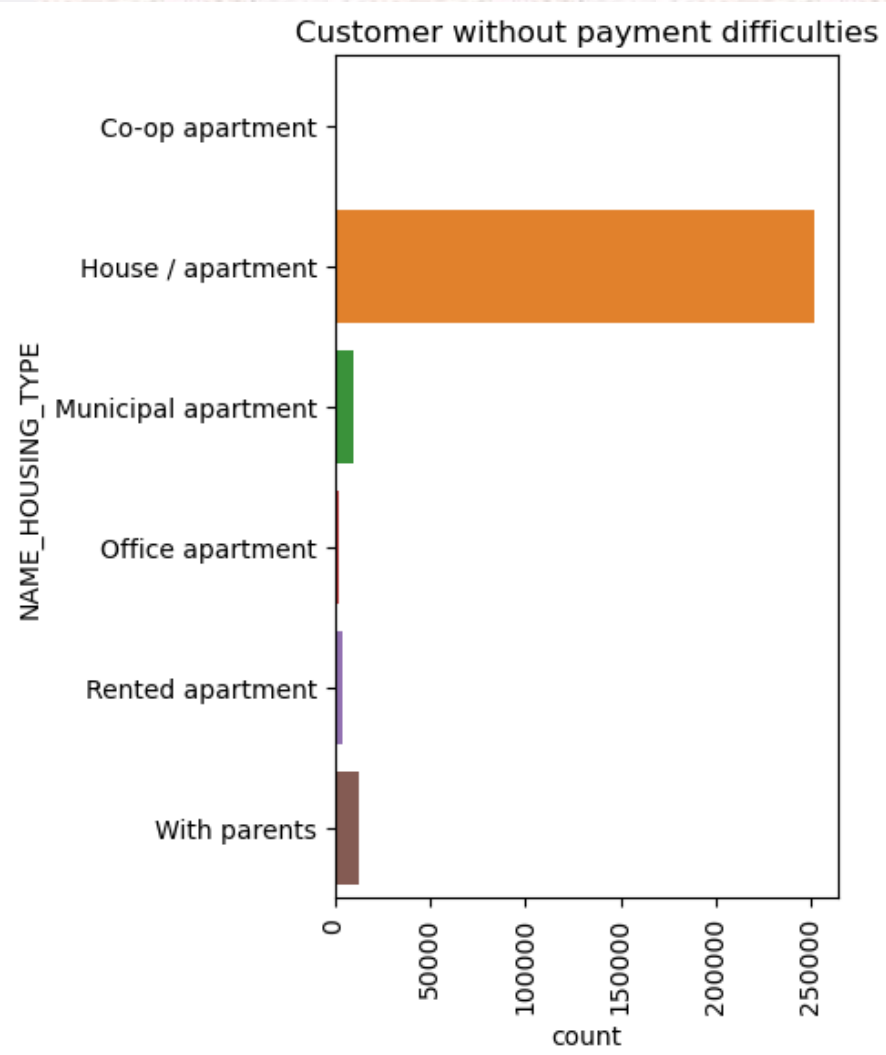


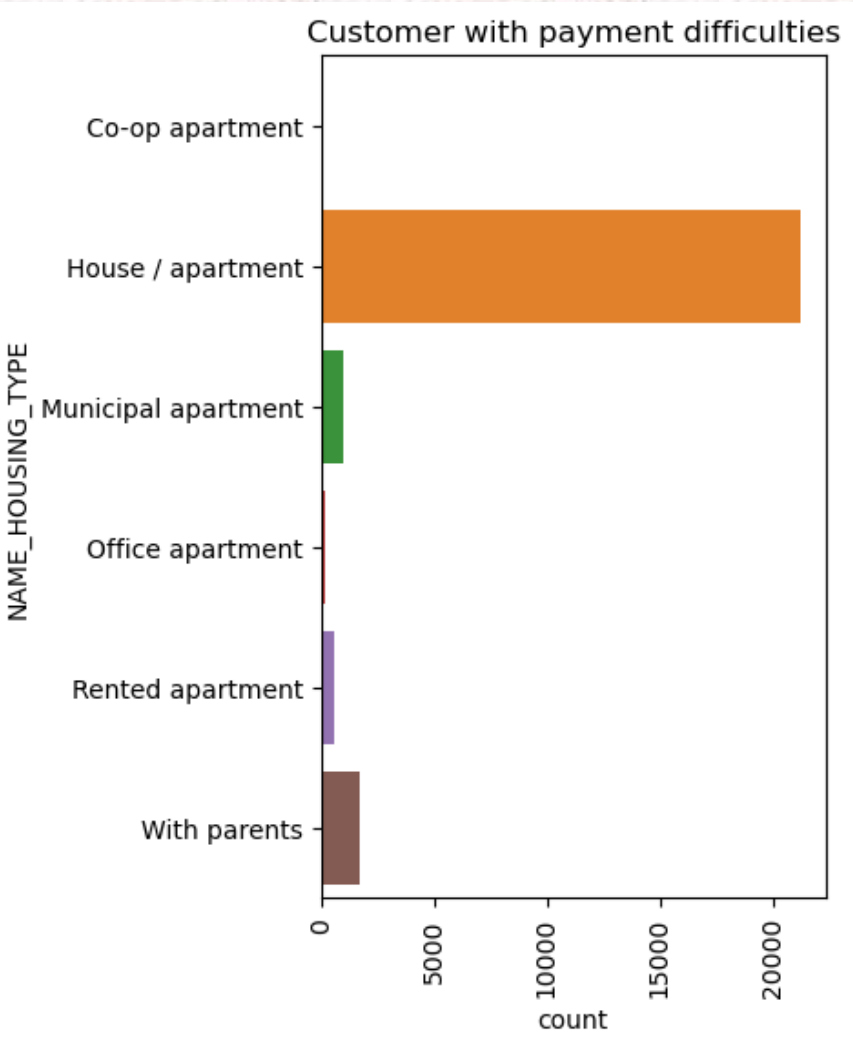


Count Plot comparing EDUCATION_TYPE and clients from dataset of Target 0 and 1

Here we can see that most of the population or the database has a secondary/secondary special education, and this could be a driving factor that we were seeing earlier with people taking more cash loans rather than revolving loans.

Categorical Univariate Analysis

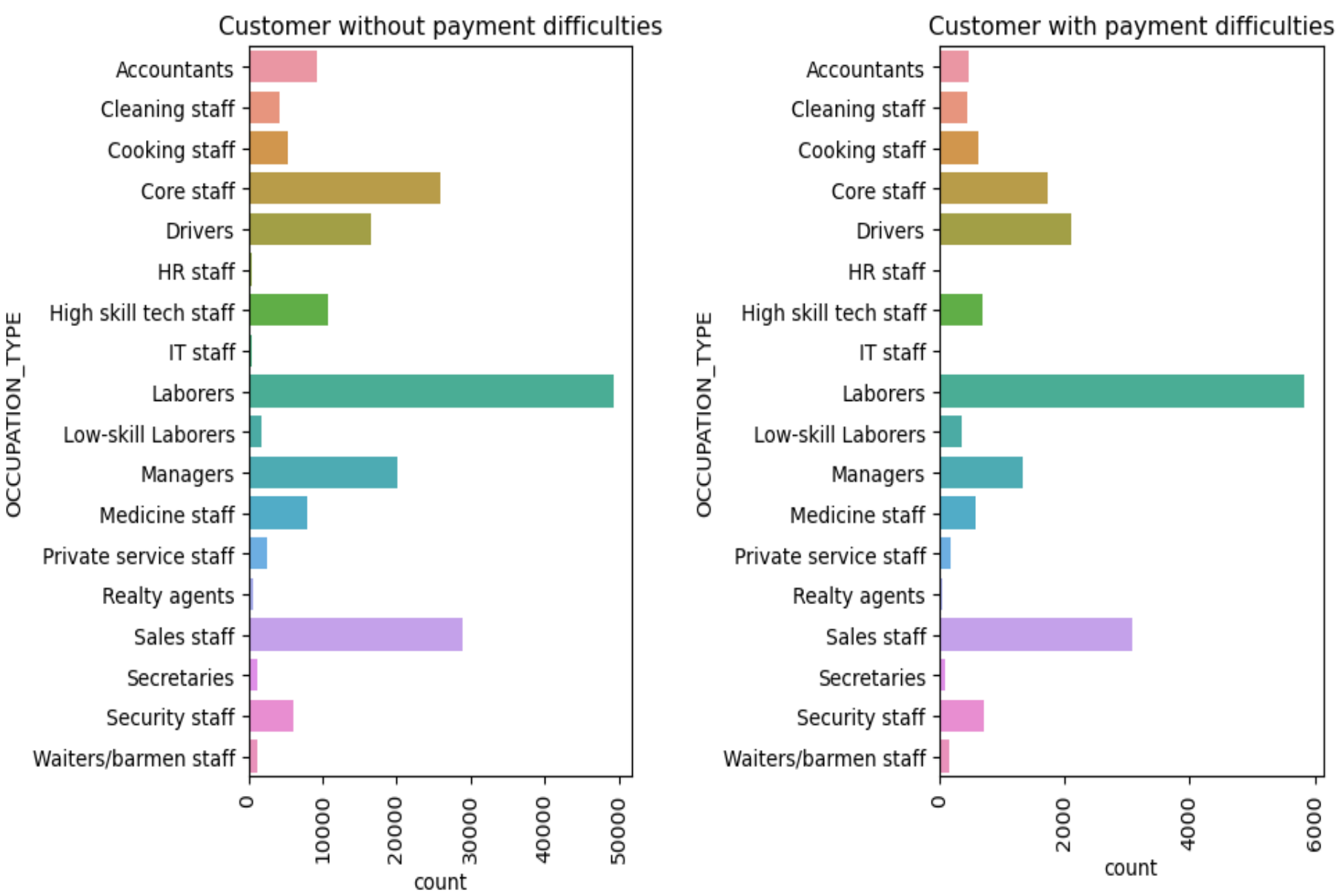




Count Plot comparing HOUSING_TYPE and clients from dataset of Target 0 and 1

Both the categories of customers have their own houses. But we can also see that people with difficulties are way less than people without difficulties, which is good news. As less defaulters are present, and bank can make recovery of their money by asking the defaulters to keep their houses as collateral.

Categorical Univariate Analysis

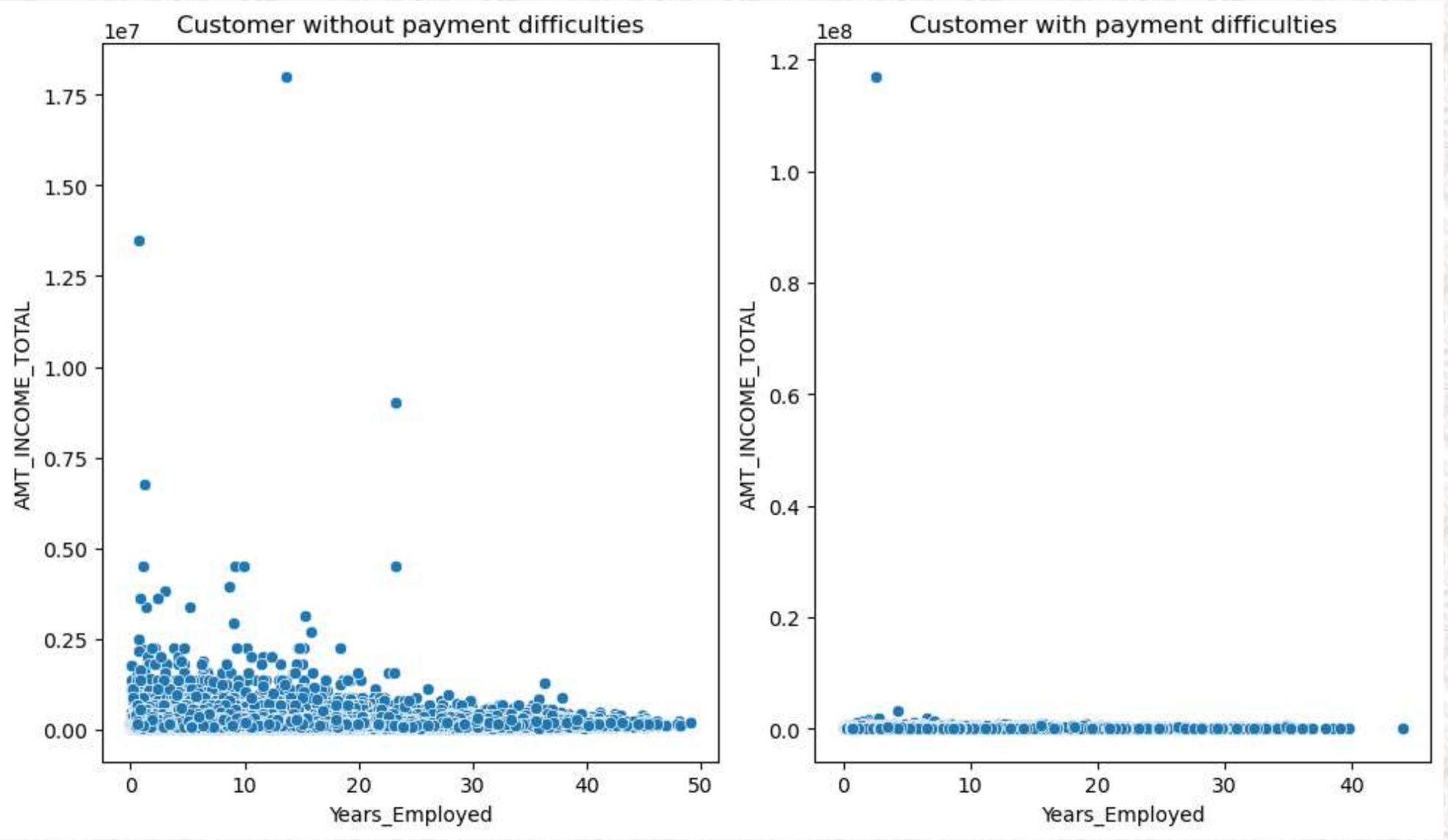


Count Plot comparing OCCUPATION_TYPE and clients from dataset of Target 0 and 1

Categorical Univariate Analysis

We can see in the count plot OCCUPATION_TYPE in previous slide that in both categories that is people who have difficulty paying the loans and people who don't have difficulty paying the labor class is at the highest because they are mostly dependent on daily wages and seasonal work and when they don't get paid, they default However, when they do get paid, they tend to pay back because of the small loan amount has small installment as well. Additionally, we can see that the sales category has the second highest numbers because they are there in majority as in the number of people who are in the sales field are comparatively more in numbers. Lastly, we can see that employees from the IT sector and HR sector are the least people to default as they have better paying jobs and less need for borrowing money.

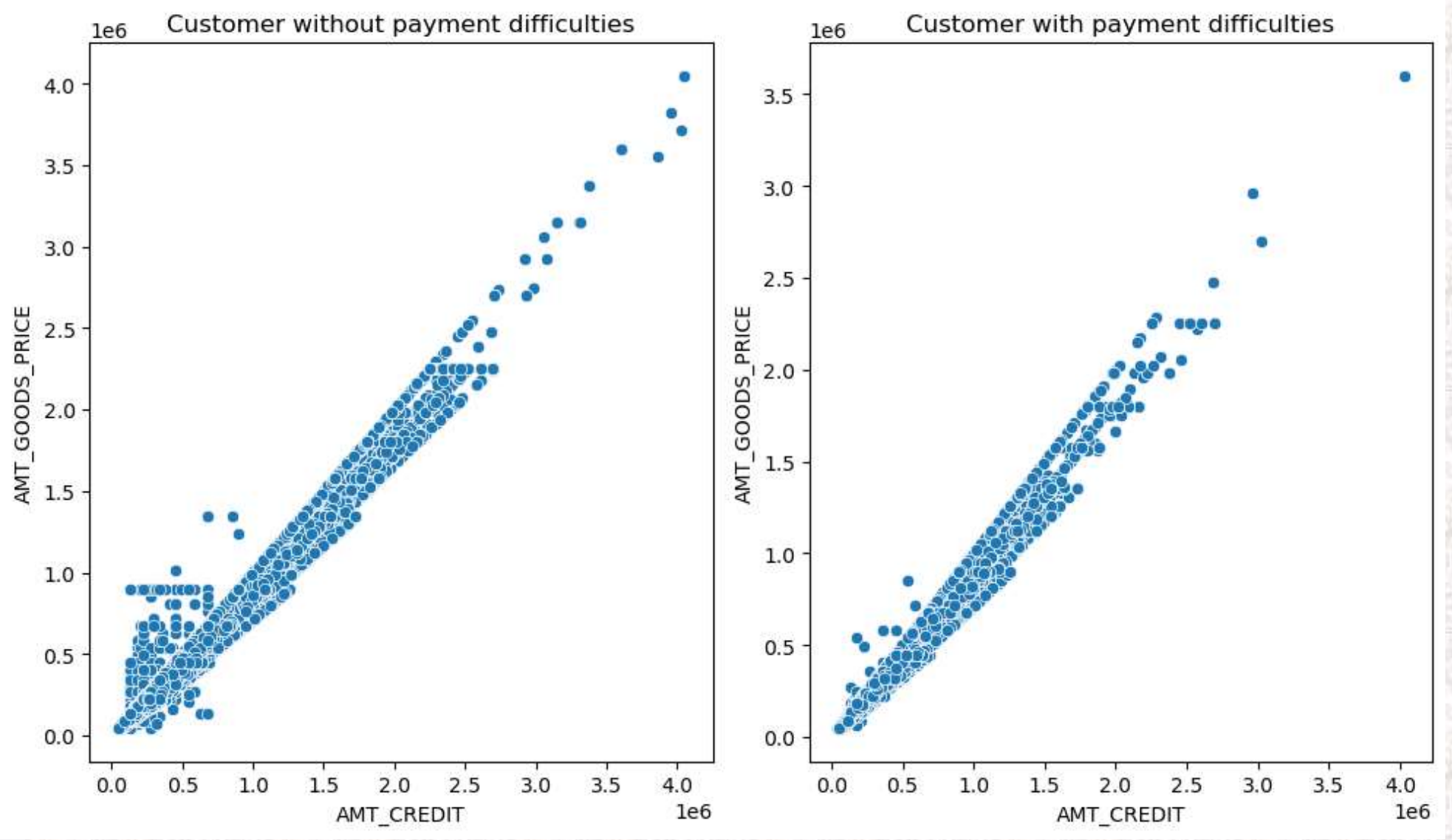
Bi-variate Analysis



Scatter Plot on Years Employed and Total Income

In this plot we can observe that people with payment problems are more consistently in the low-income group that may be the result of the labor and sales category. However, people with less payment difficulty are ranging from low income to high income.

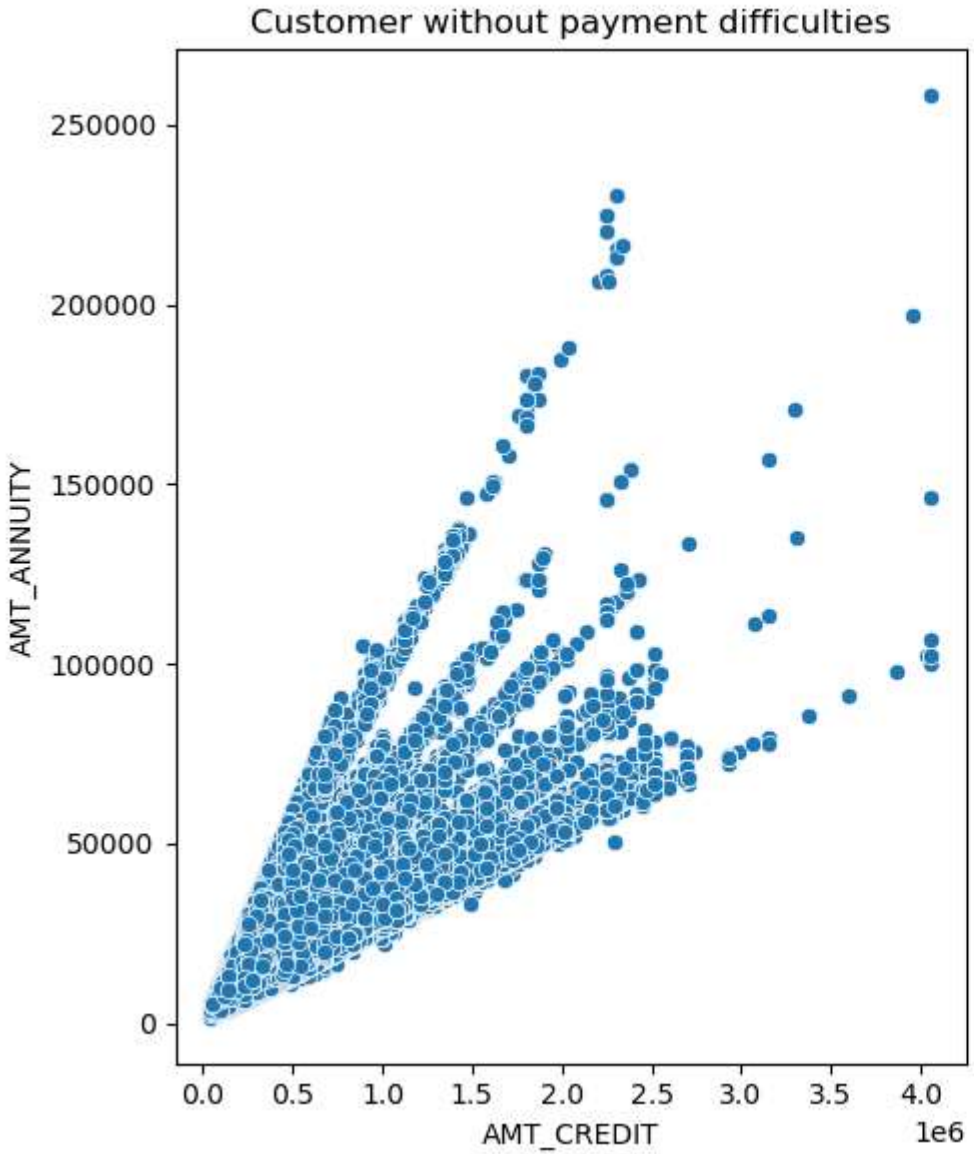
Bi-variate Analysis

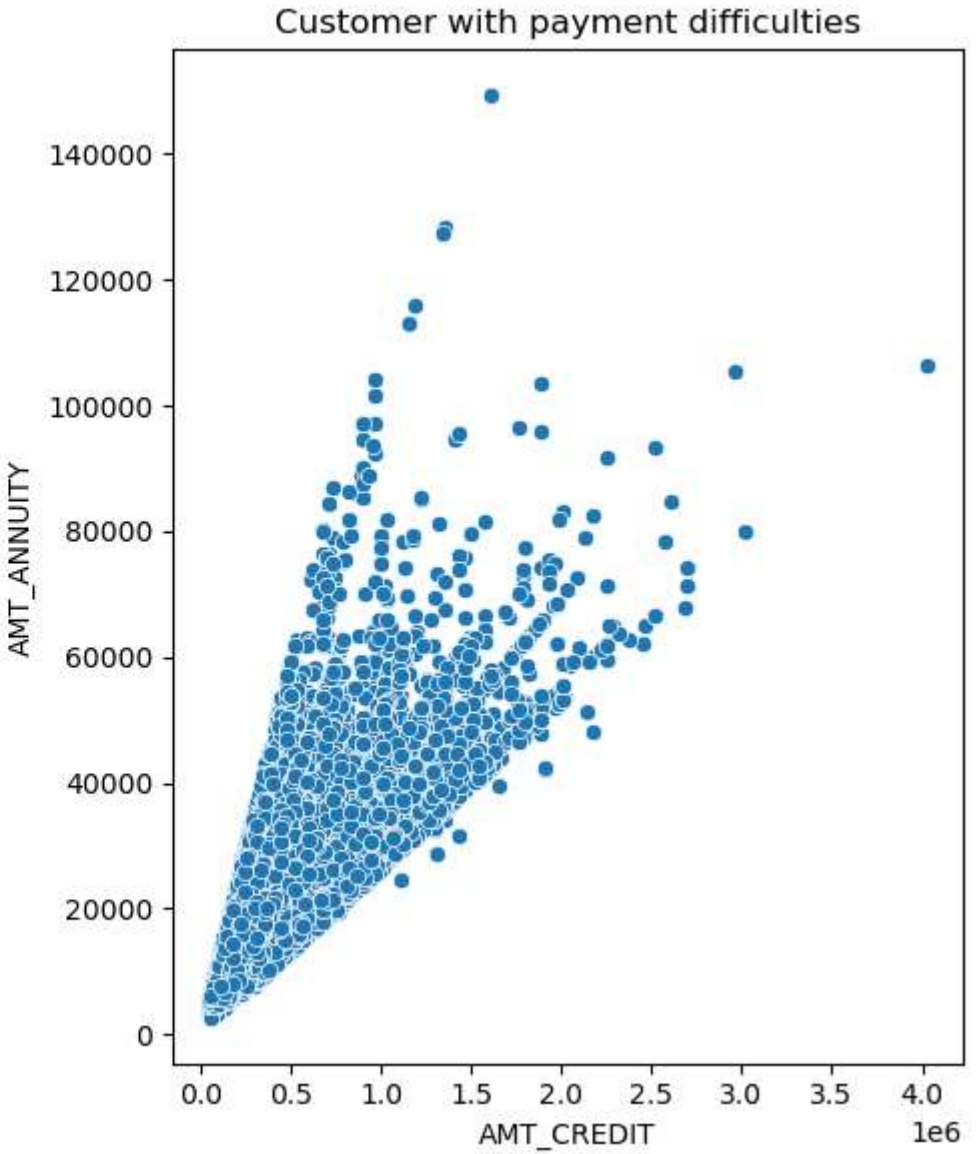


Scatter Plot on Credit amount and Goods Price

Here we see the positive correlation of amt credit to goods price in both the cases.

Bi-variate Analysis

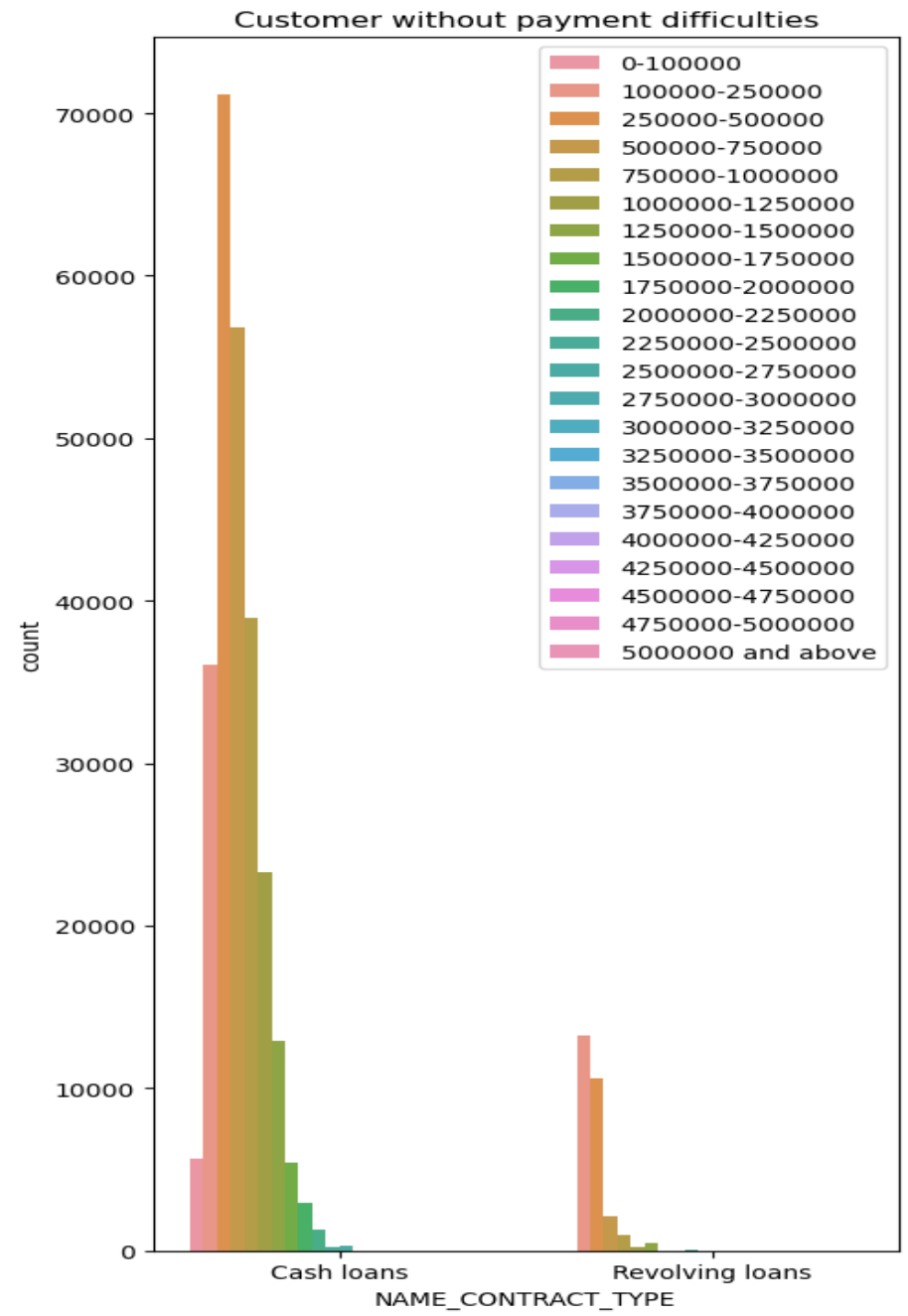


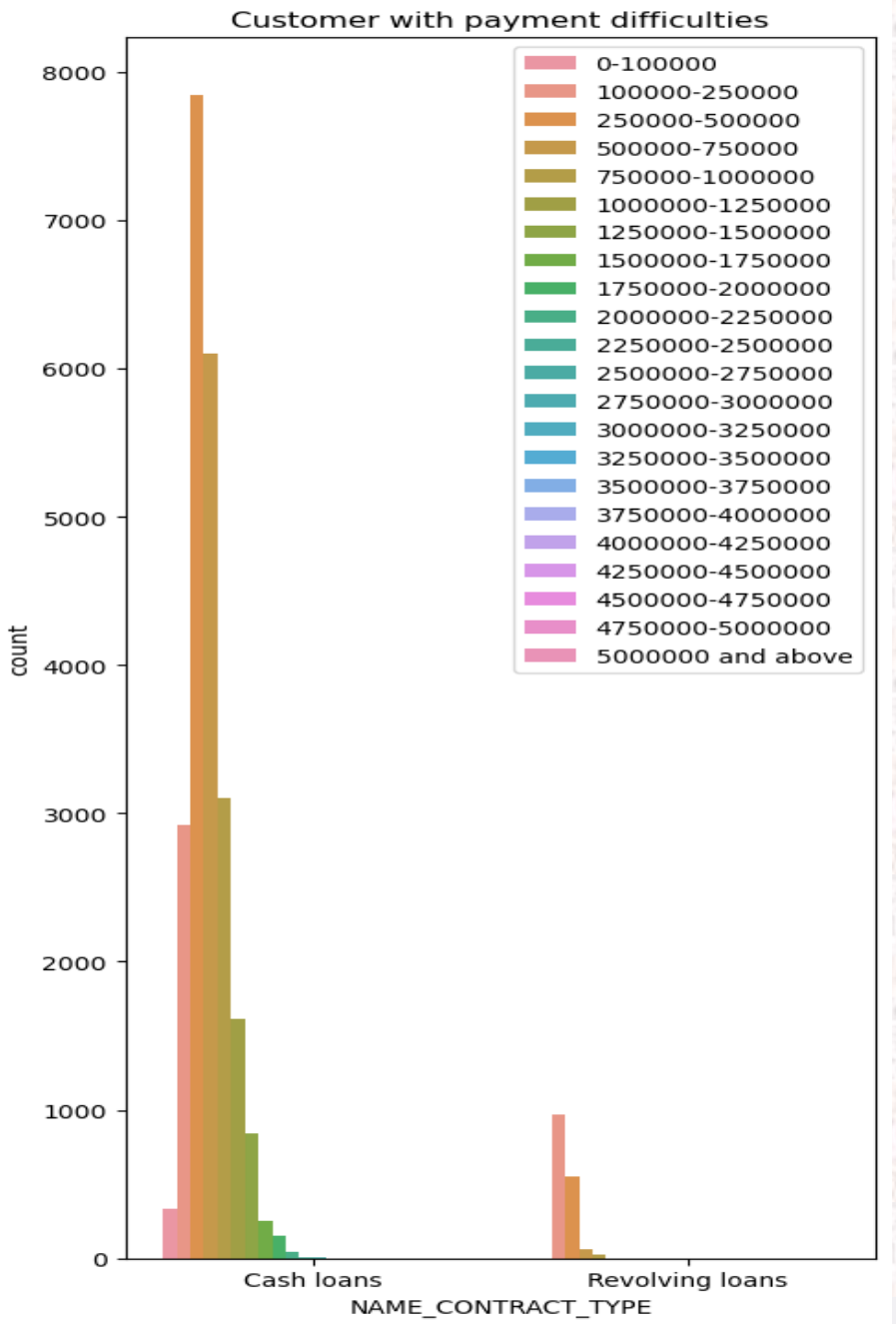


Scatter Plot on Credit amount and Annuity

Inference:
People without payment difficulties take more credit for the annuity

Bi-variate Analysis

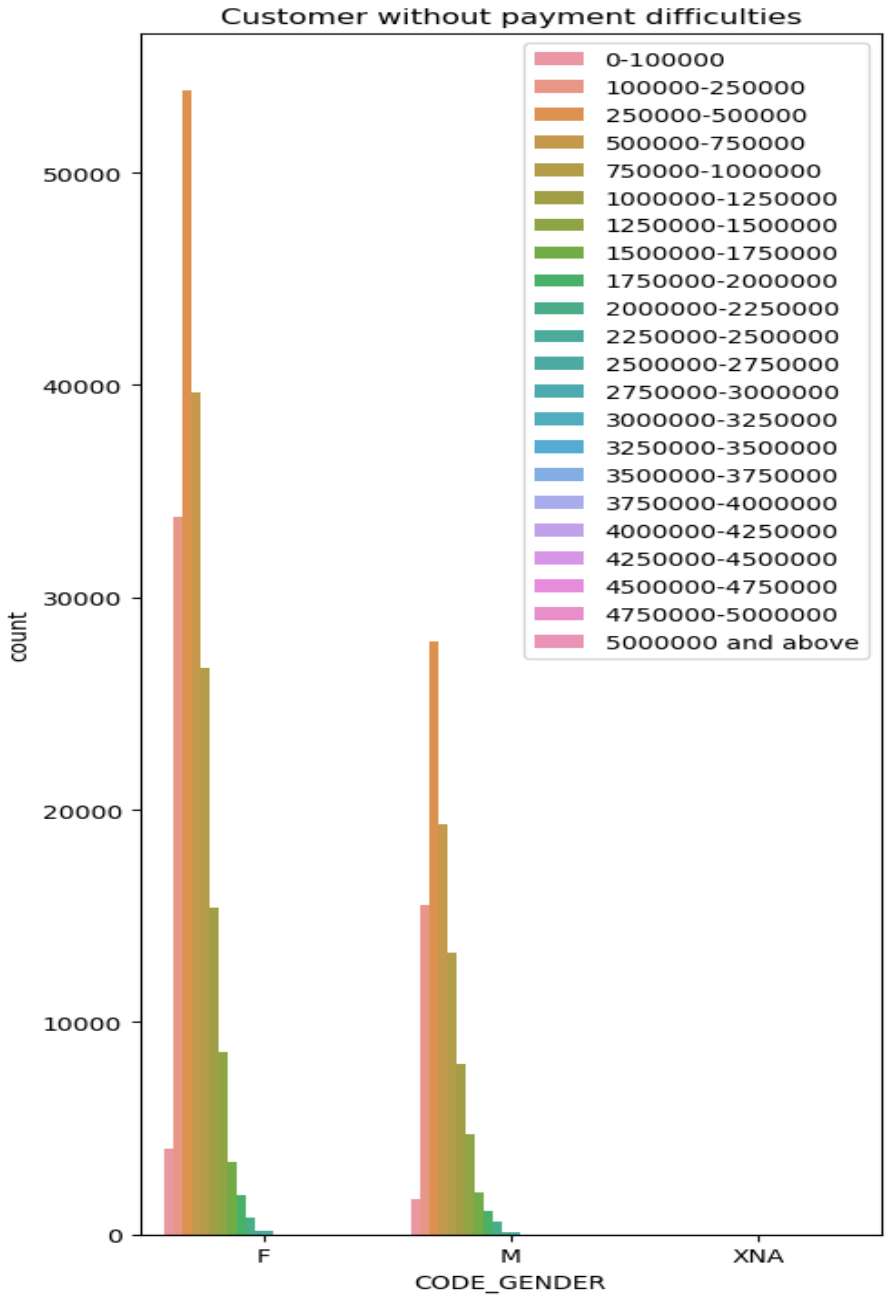


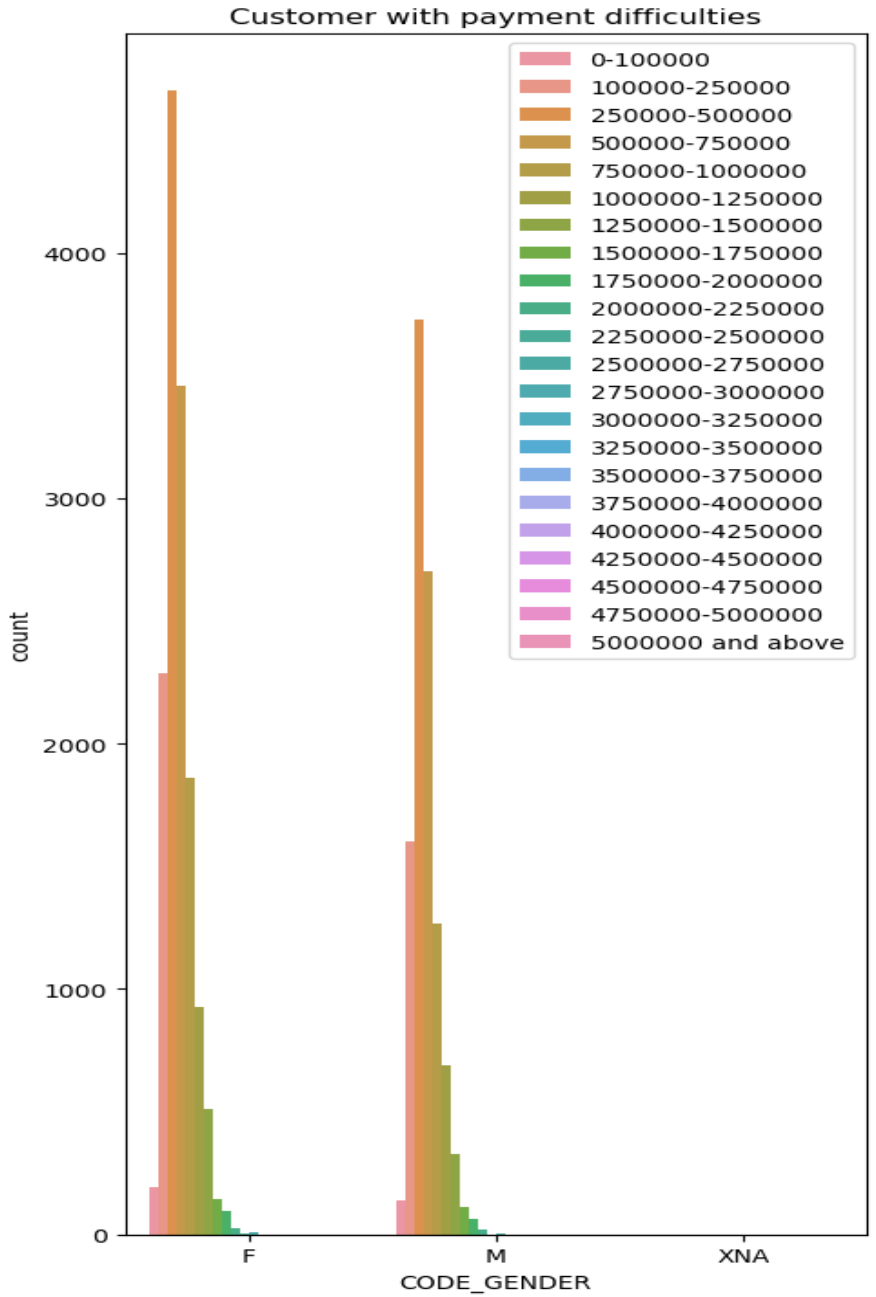


Count Plot on Contract Type and Credit Range

Inference:
Here we see that people from both the category take loans of cash type more than the revolving loans but that can be credited to most people in Labor and sales class.

Bi-variate Analysis

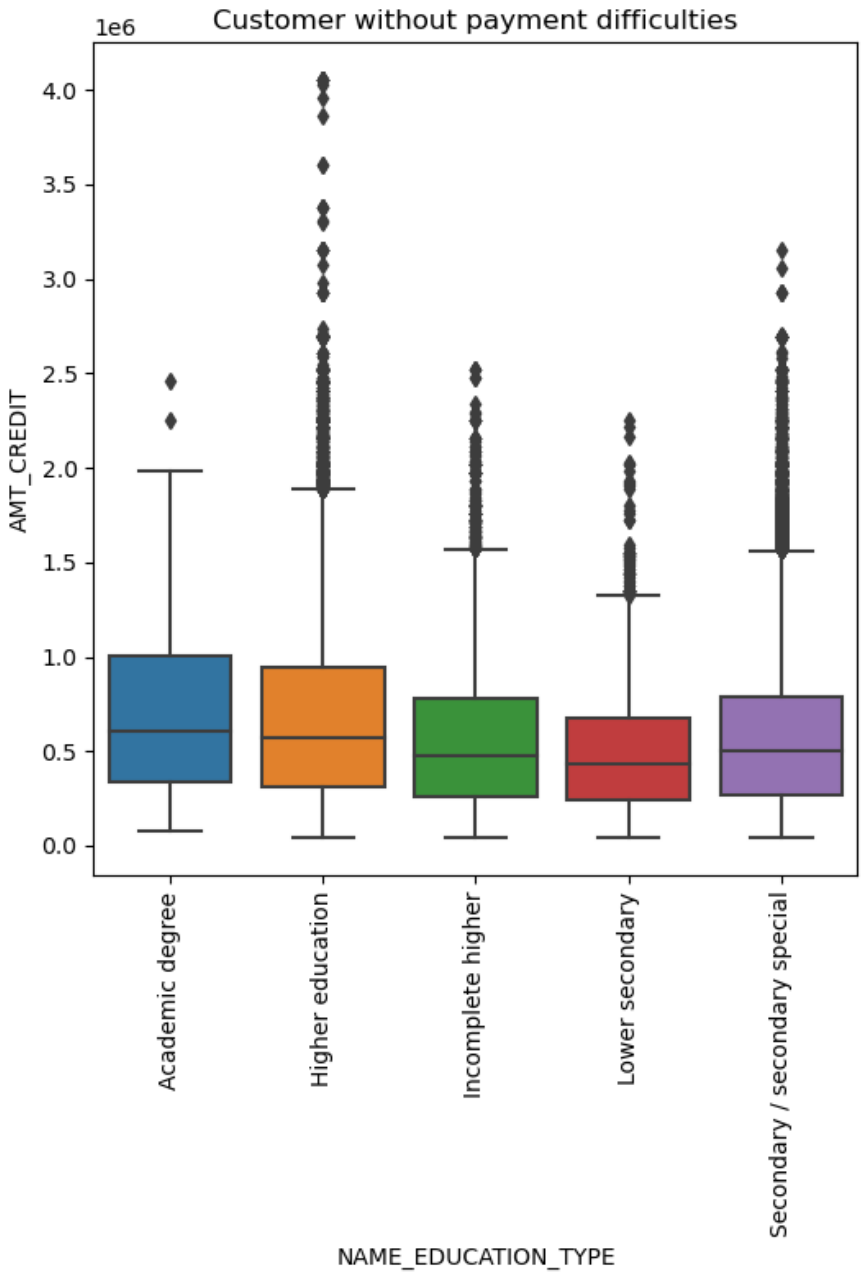


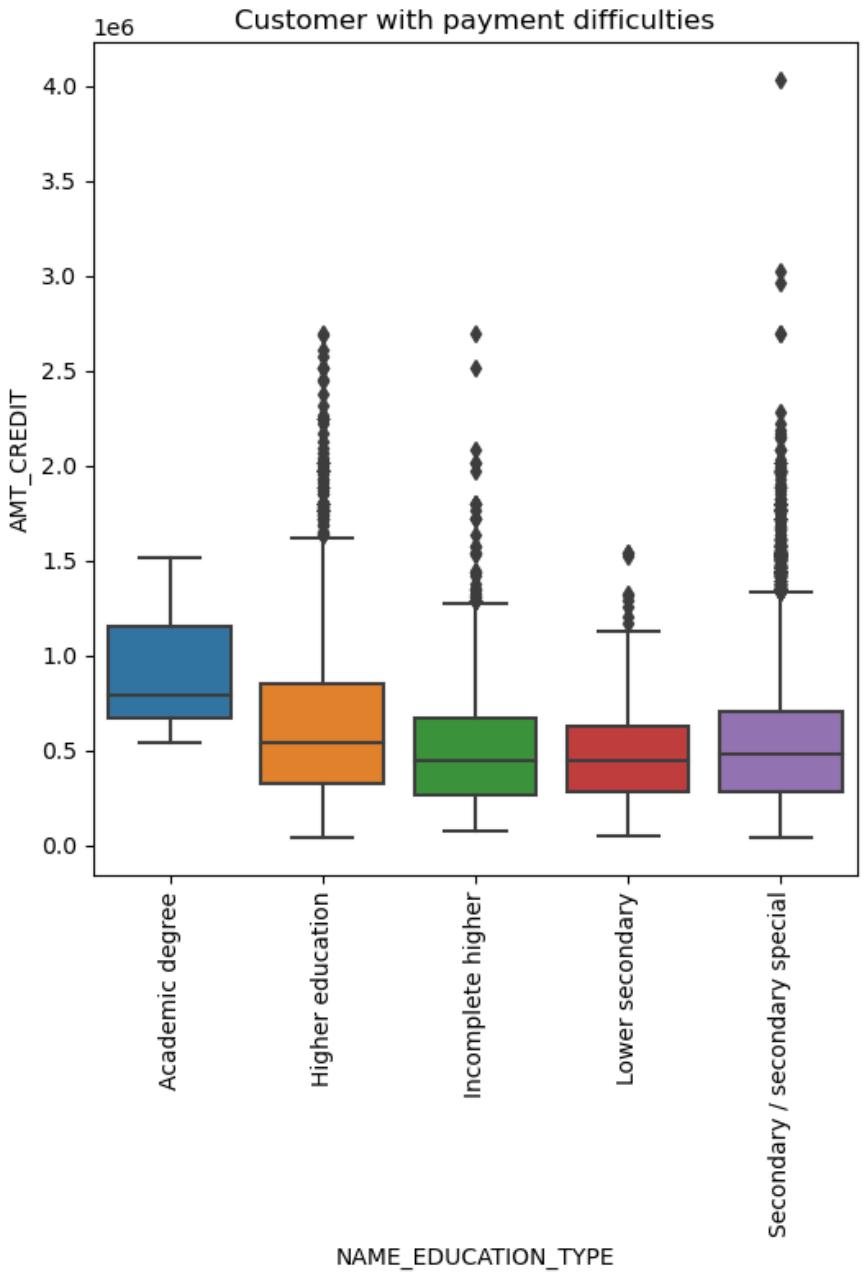


Count Plot on Gender and Credit Range

Inference:
In both categories females have taken out more loans and even the amount is greater in both cases in females.

Numerical and Categorical Bi-variate Analysis

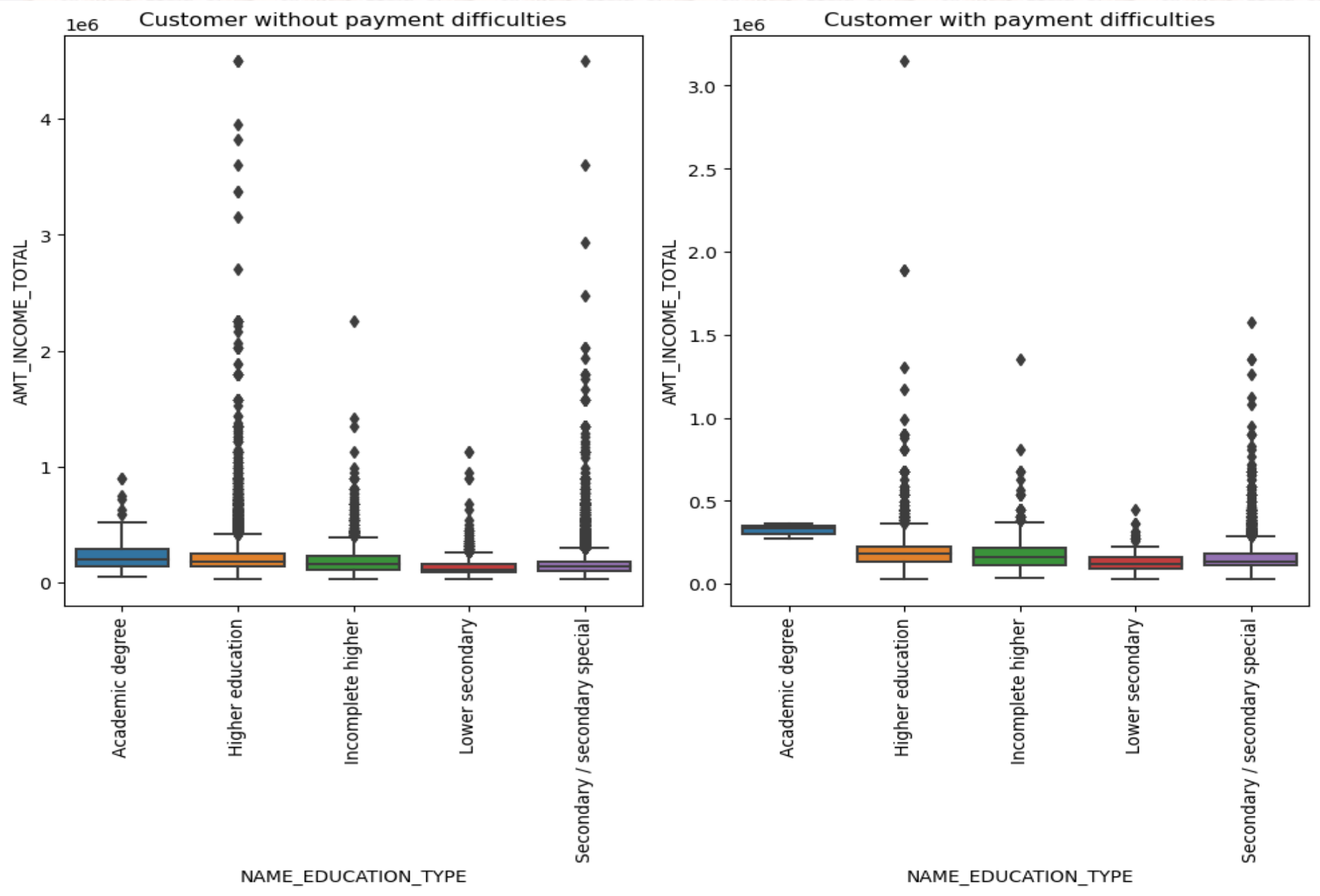




Box Plot on Credit Amount and Education Type

Inference: In case of clients with difficulty paying the loan, we can see that people with higher education struggled more in repayment it can be because of the situation of job market and the amount of loan or their employment status. On the other hand, clients who don't have payment difficulty is also leading with higher education.

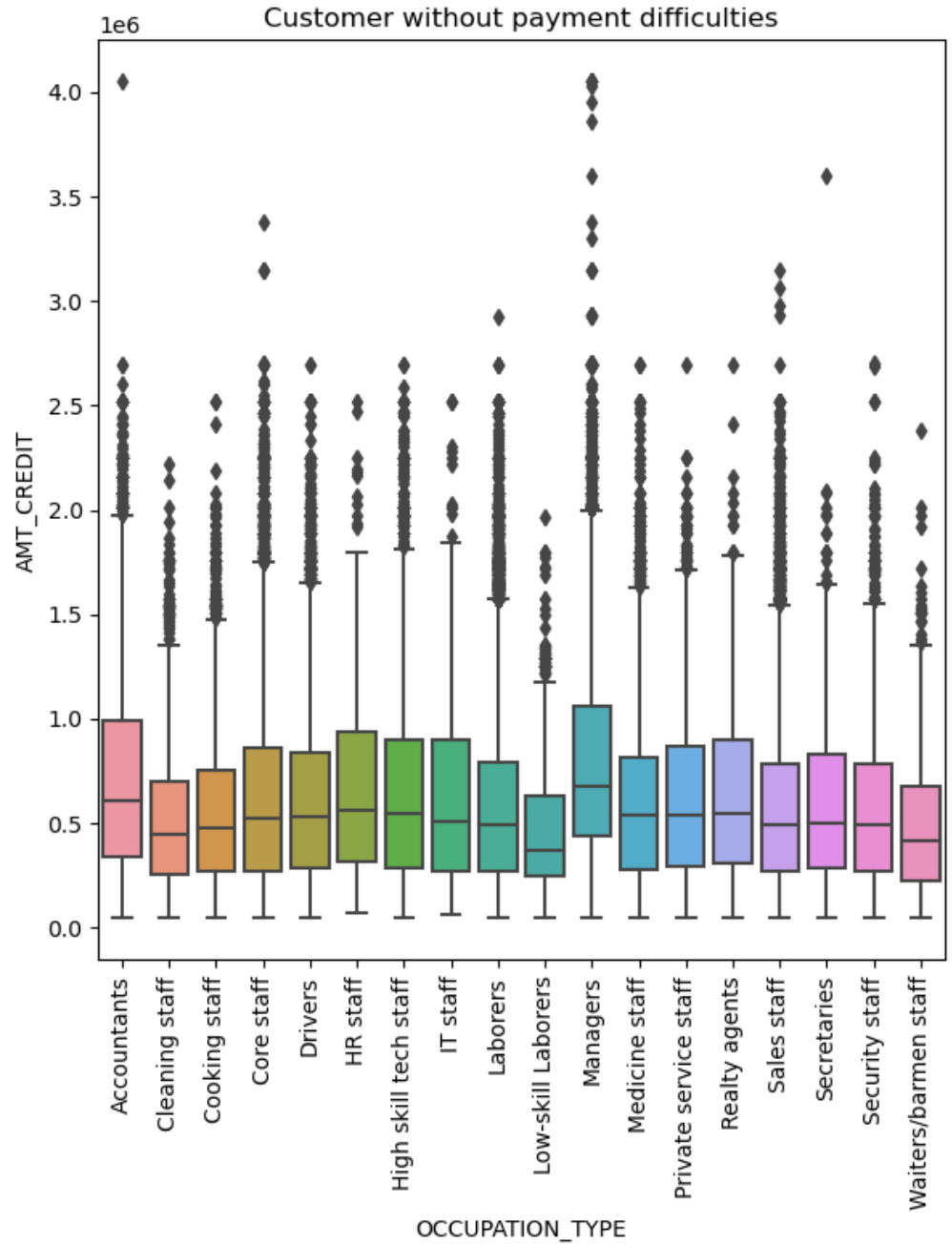
Numerical and Categorical Bi-variate Analysis

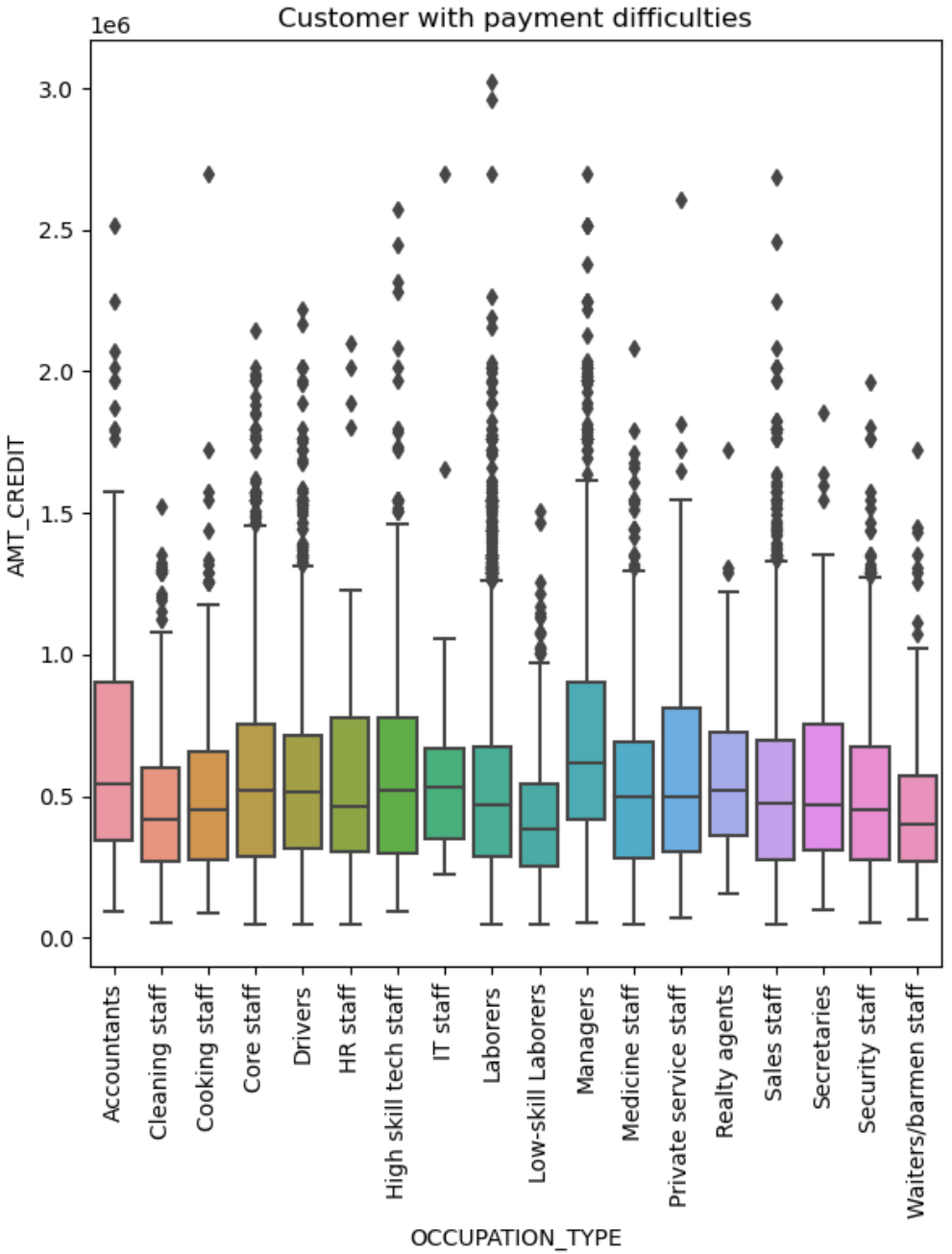


Box Plot on Total Income and Education Type

Inference:
In both cases there are numerous outliers, but academic degree is the least in both the cases because of the client base of the dataset which is less.

Numerical and Categorical Bi-variate Analysis





Box Plot on Credit Amount and Occupation Type

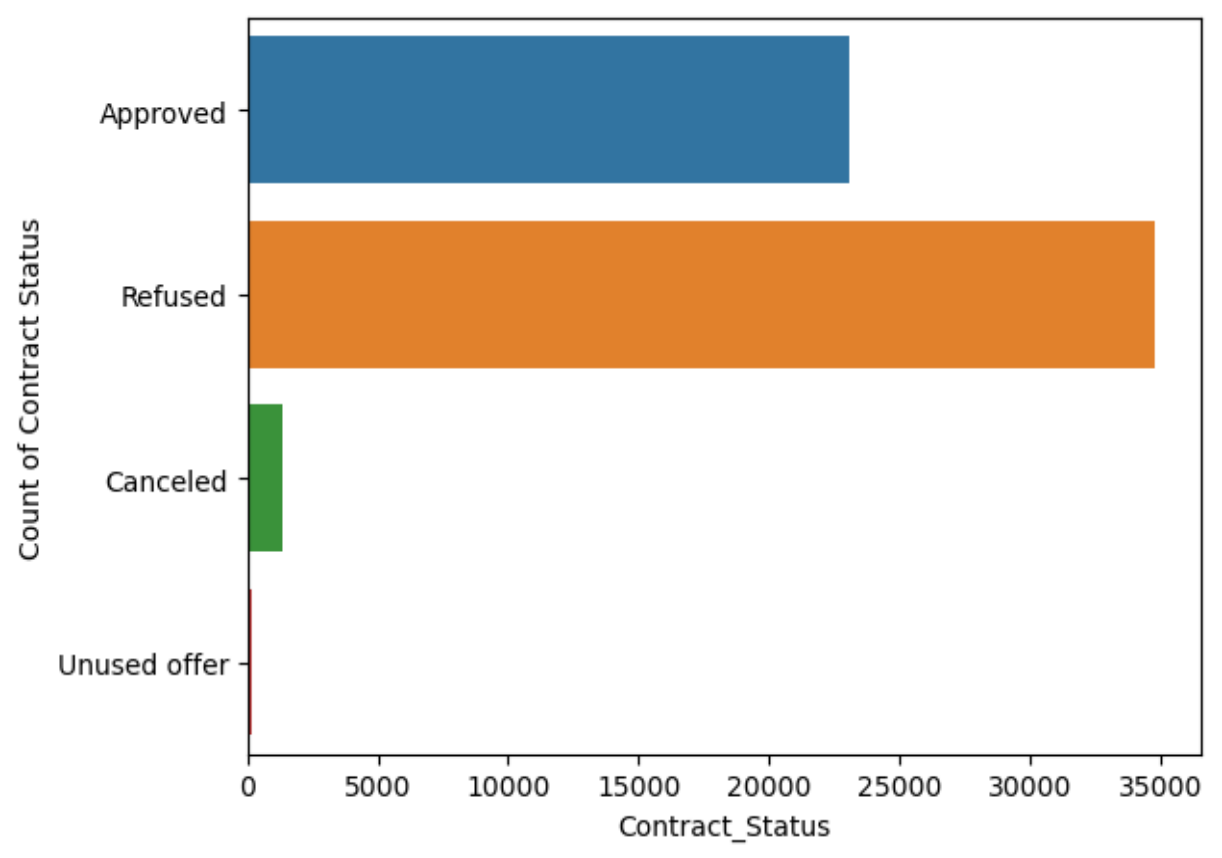
Inference:
We can observe that the amount of credit taken is more in clients with no difficulties and defaulters tend to take less amount of credit in any occupation type.
Also, we can see that Accountants and Mangers tend to take more loans and have more difficulty paying back as well.

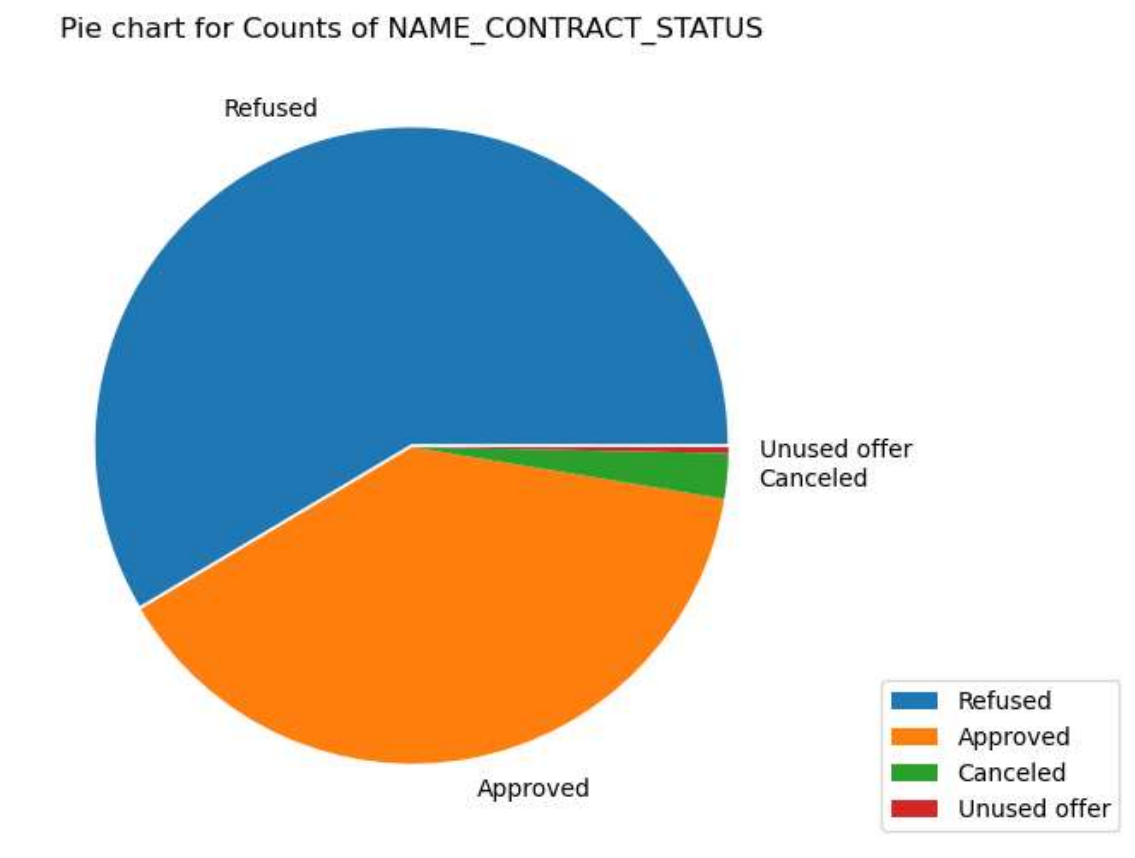
Previous Application Data

- We have followed the same steps as the current application data for Data Cleaning.
- After Cleaning the data, we have merged both the current and previous application data to perform the final analysis.

Final Dataset Analysis

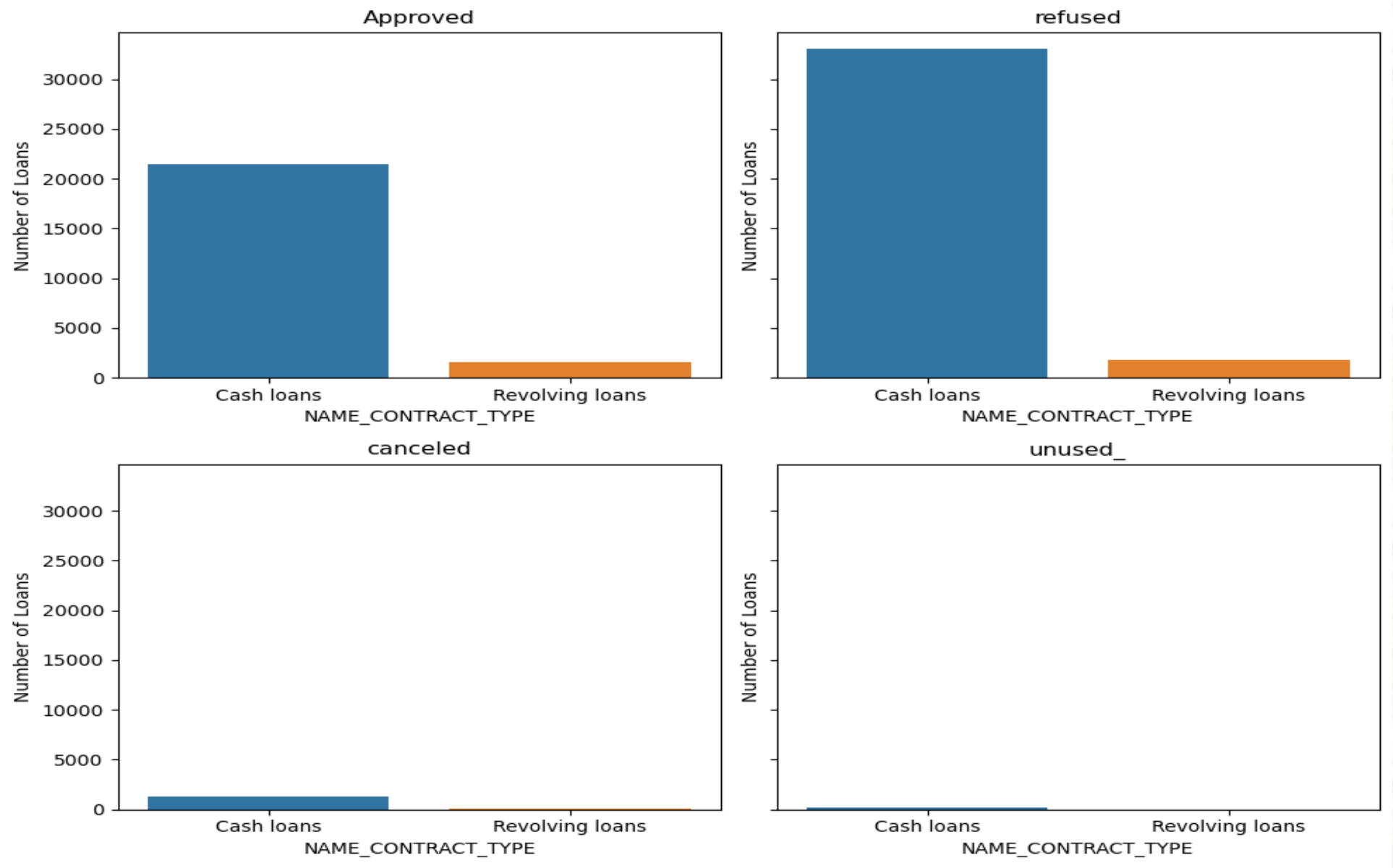
Count Plot and Pie Plot showing Different status of Loan Offered





Final Dataset Analysis

Count Plot for Contract Type with four subcategory

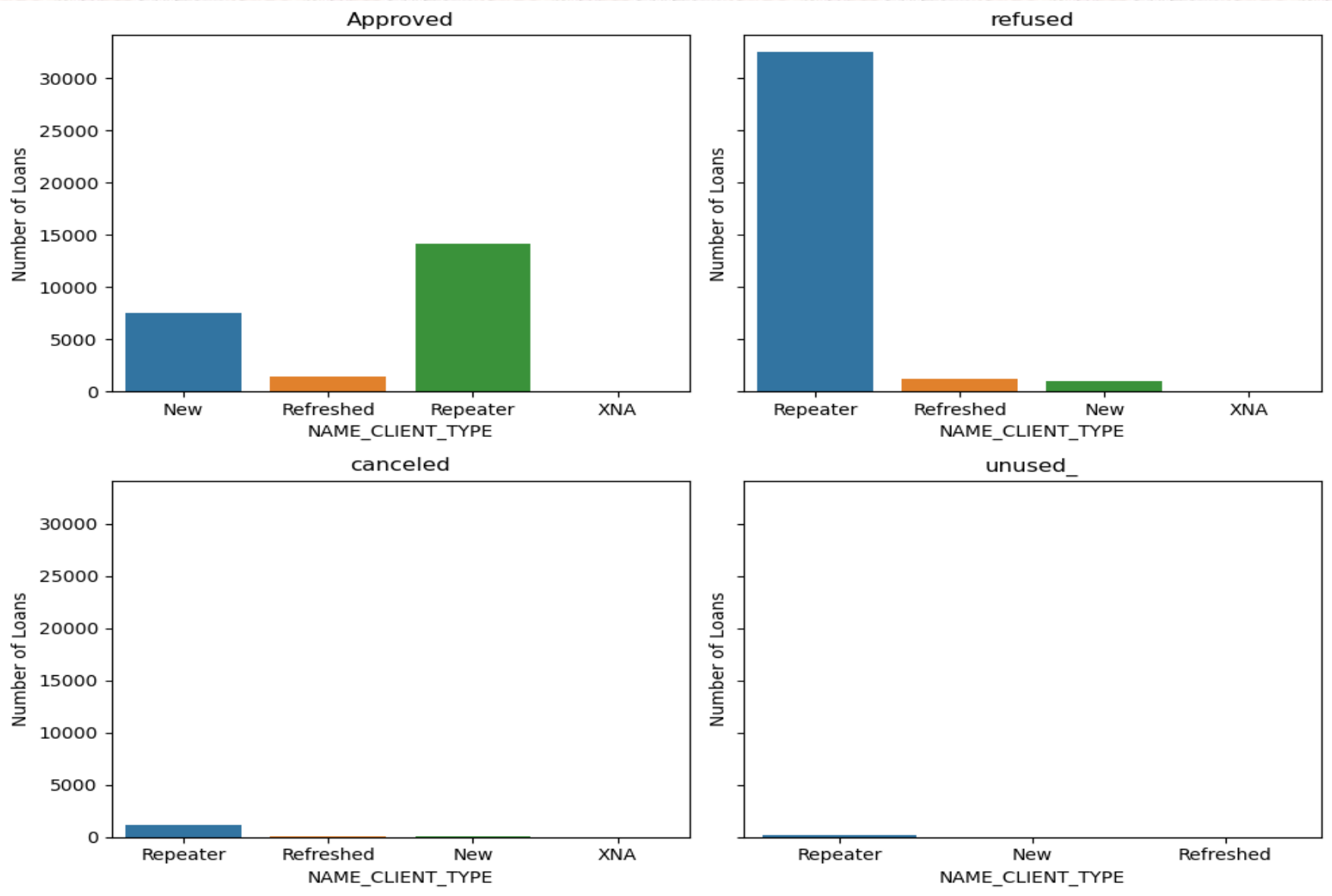


Inference:

We can see that the Cash loans are more commonly taken than the revolving loan. This can be since Laborer and low-income type people are in majority in the dataset.

Final Dataset Analysis

Count Plot for Client Type with four subcategory

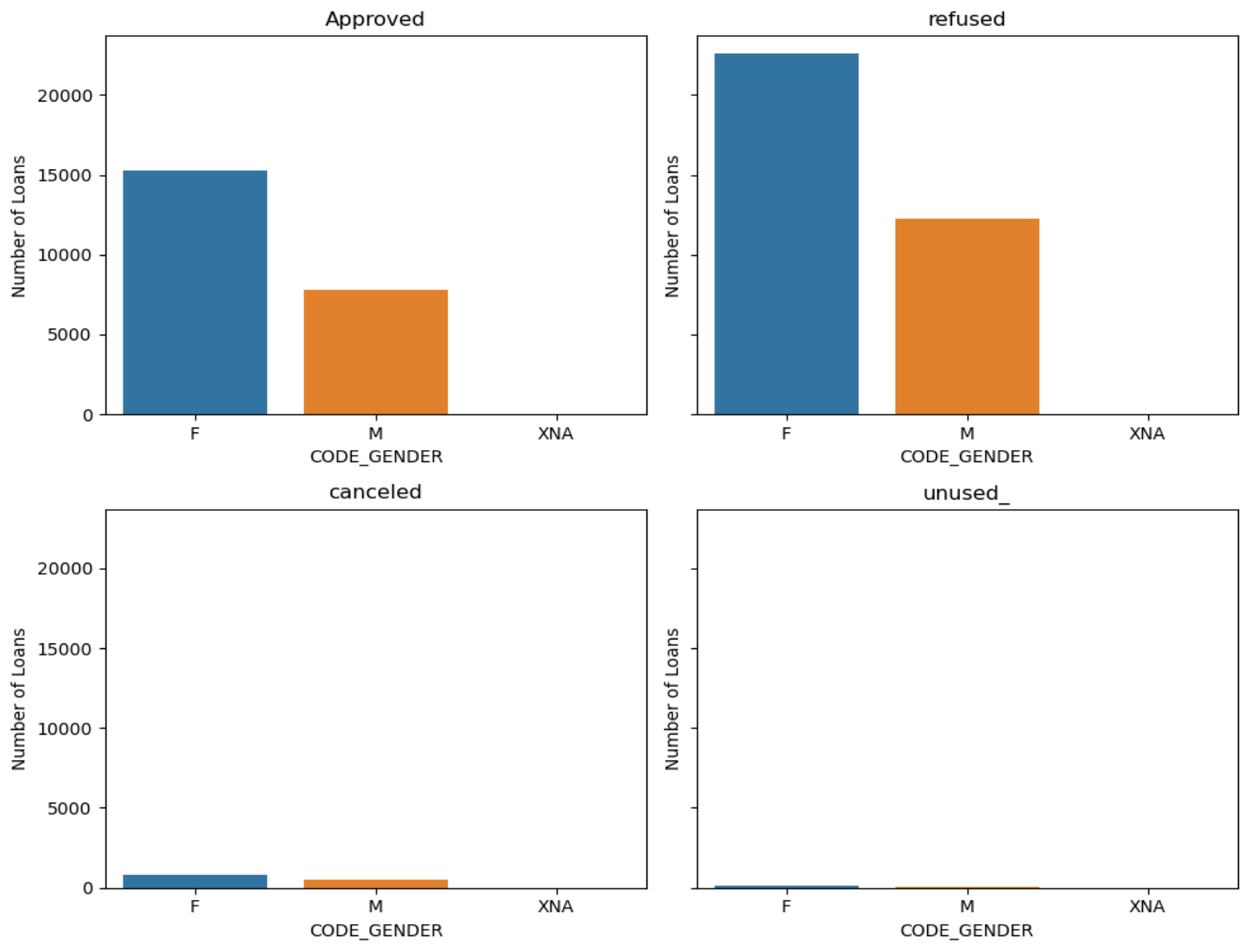


Inference:

In all the cases above the repeating borrower is getting rejected more often followed by new applicants

Final Dataset Analysis

Count Plot for Gender with four subcategory

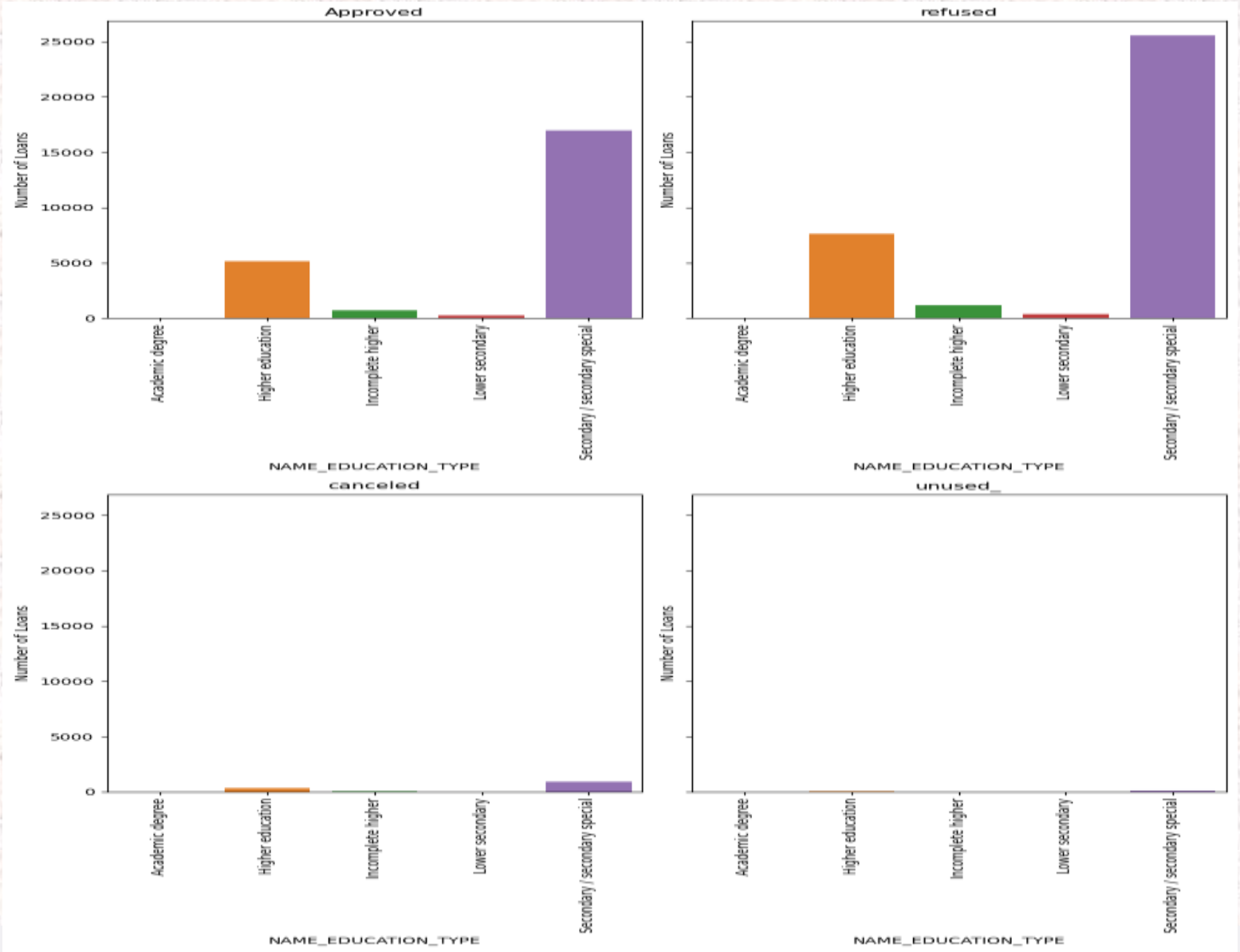


Inference:

In all the cases females have been refused more as we have seen above it is because they are more in numbers when compared to males hence the disparity

Final Dataset Analysis

Count Plot for Education Type with four subcategory

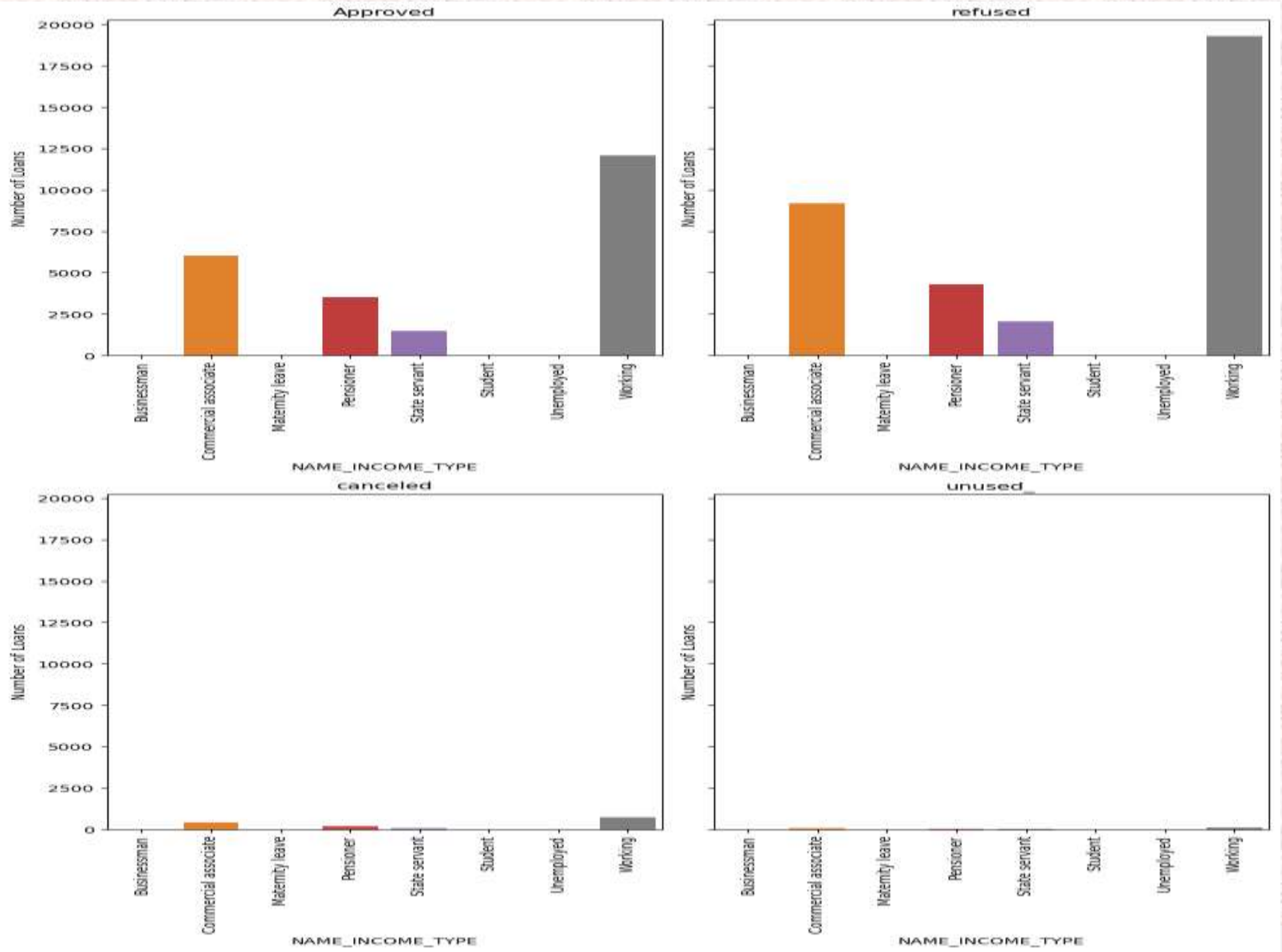


Inference:

Here people from less educated background that is secondary and secondary special have taken more loans compared to others

Final Dataset Analysis

Count Plot for Income Type with four subcategory

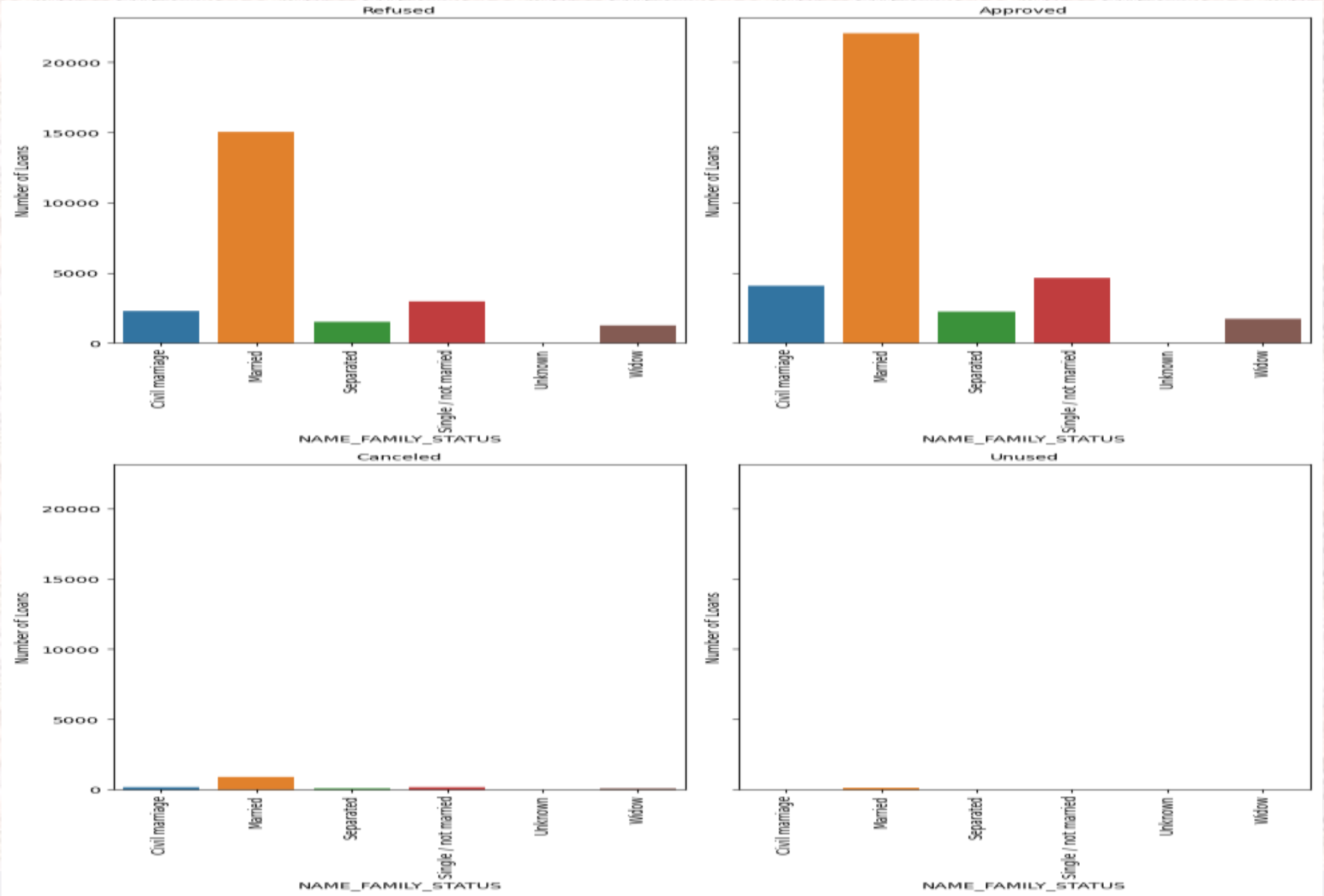


Inference:

In all cases working people have been approved more and have more loans when compared to other classes of people followed by commercial associates

Final Dataset Analysis

Count Plot for Family Status with four subcategory

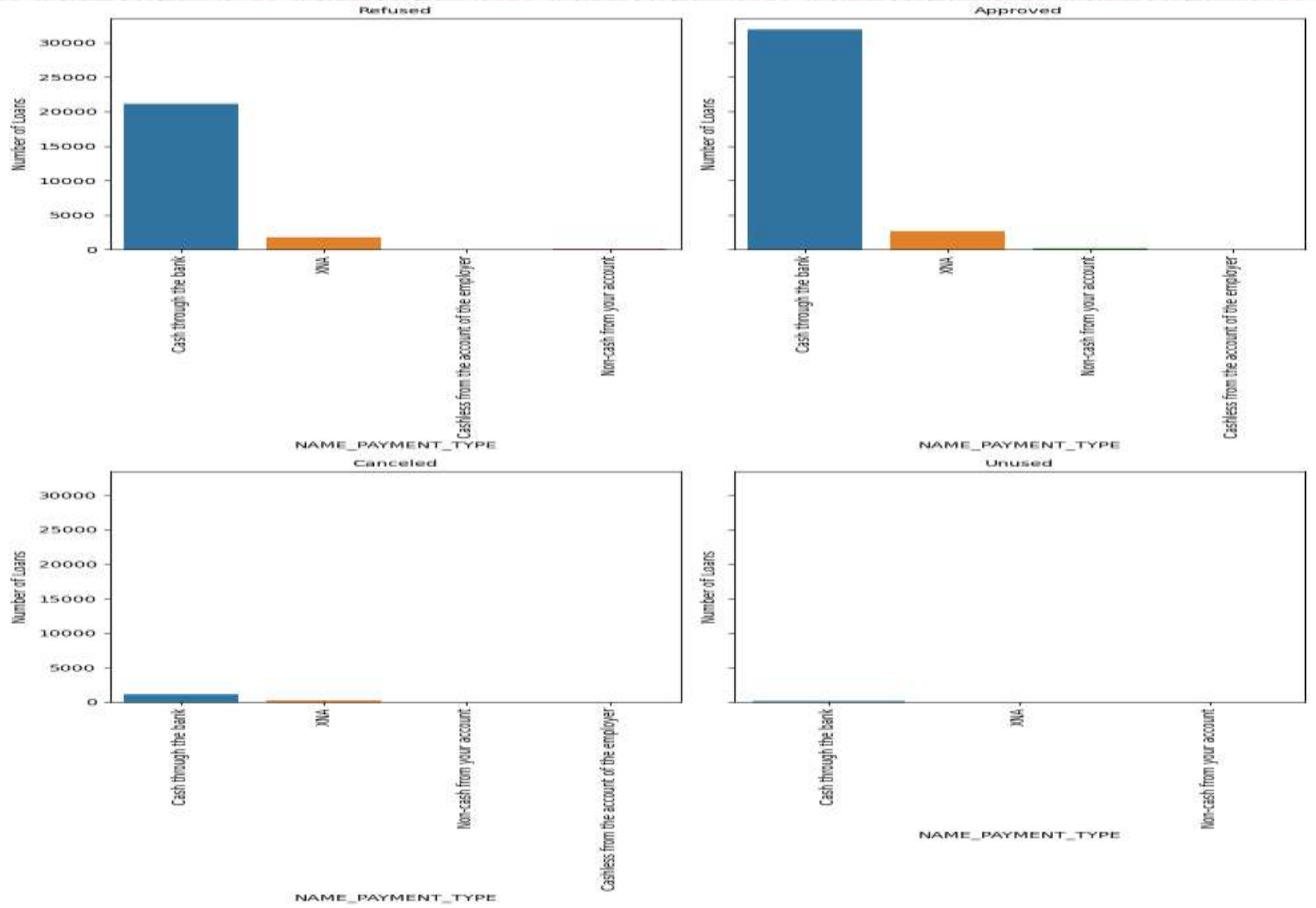


Inference:

We can observe that married clients are the biggest borrowers from the bank.

Final Dataset Analysis

Count Plot for Payment Type with four subcategory

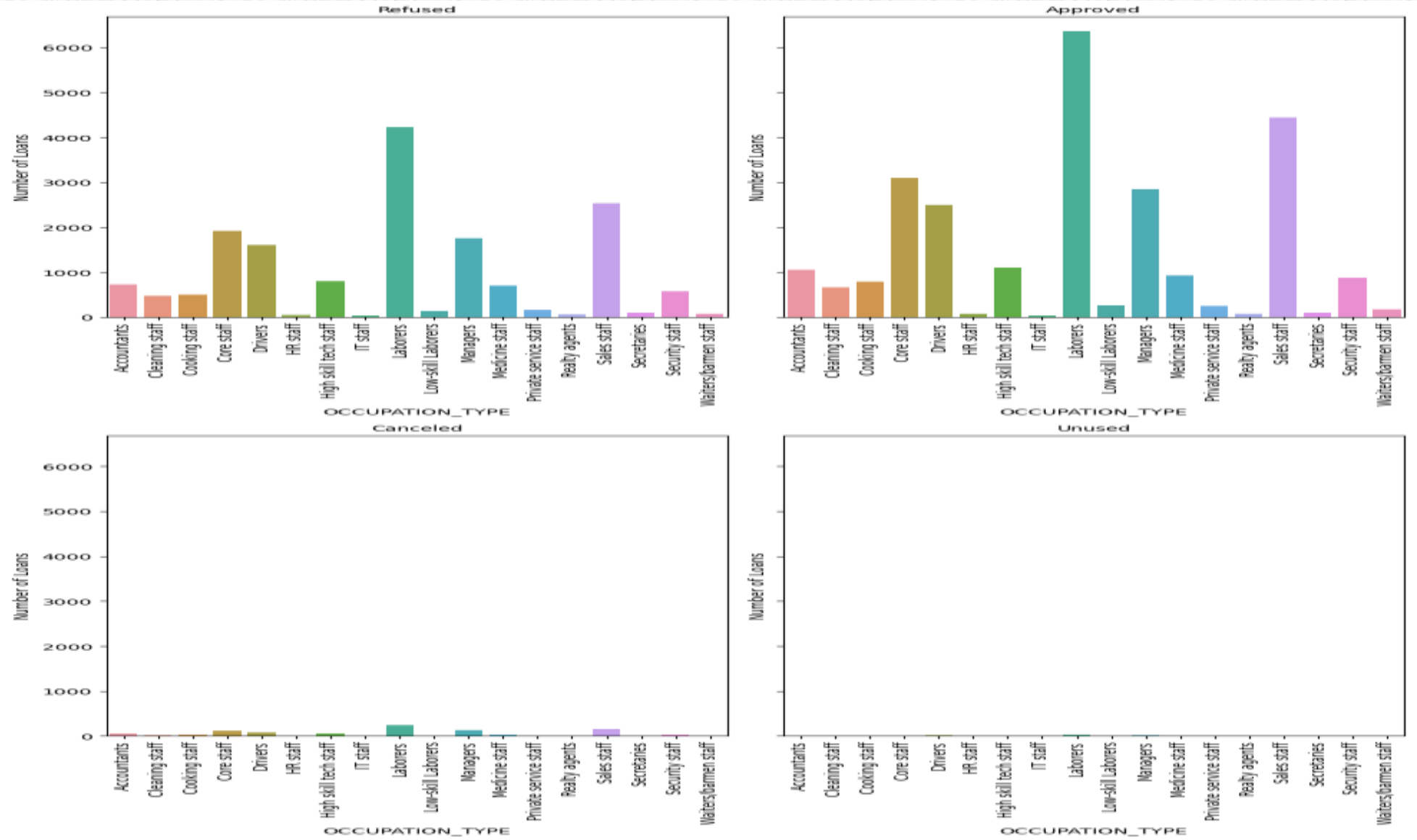


Inference:

Cash through Bank category is the preferred form of borrowing amongst all the clients.

Final Dataset Analysis

Count Plot for Occupation Type with four subcategory



Inference:

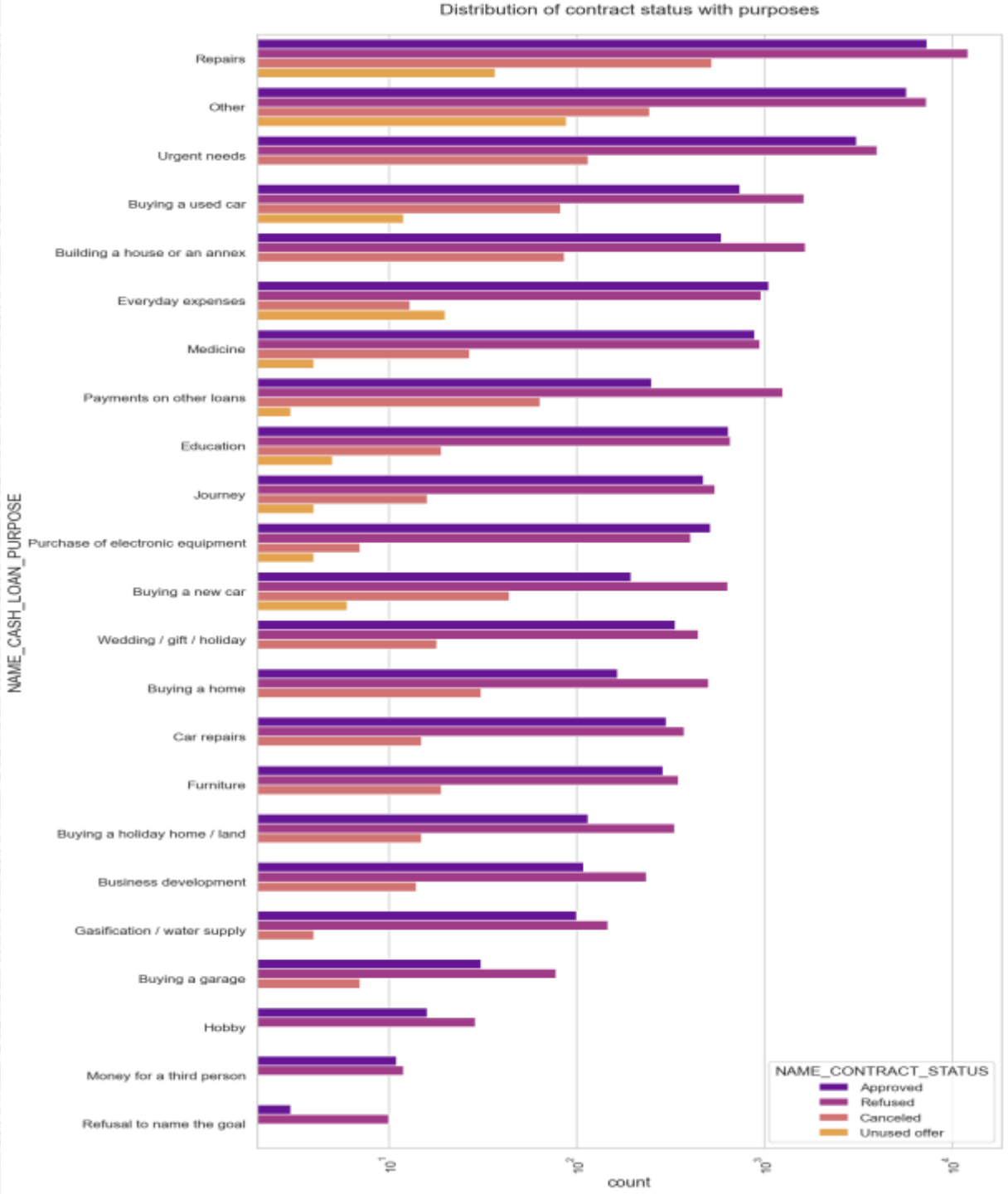
Laborer, sales staff, core staff and drivers are the leading borrowers who highest in approval and rejection.

Univariate Analysis on Final Dataset

Logarithmic Comparison of Purpose of Borrowing

Inference:

Majority of rejected loans are from the category 'repairs'. Also, education has equal number of approves and rejection
Paying other loans and buying a new car is having significant higher rejection than approvals.

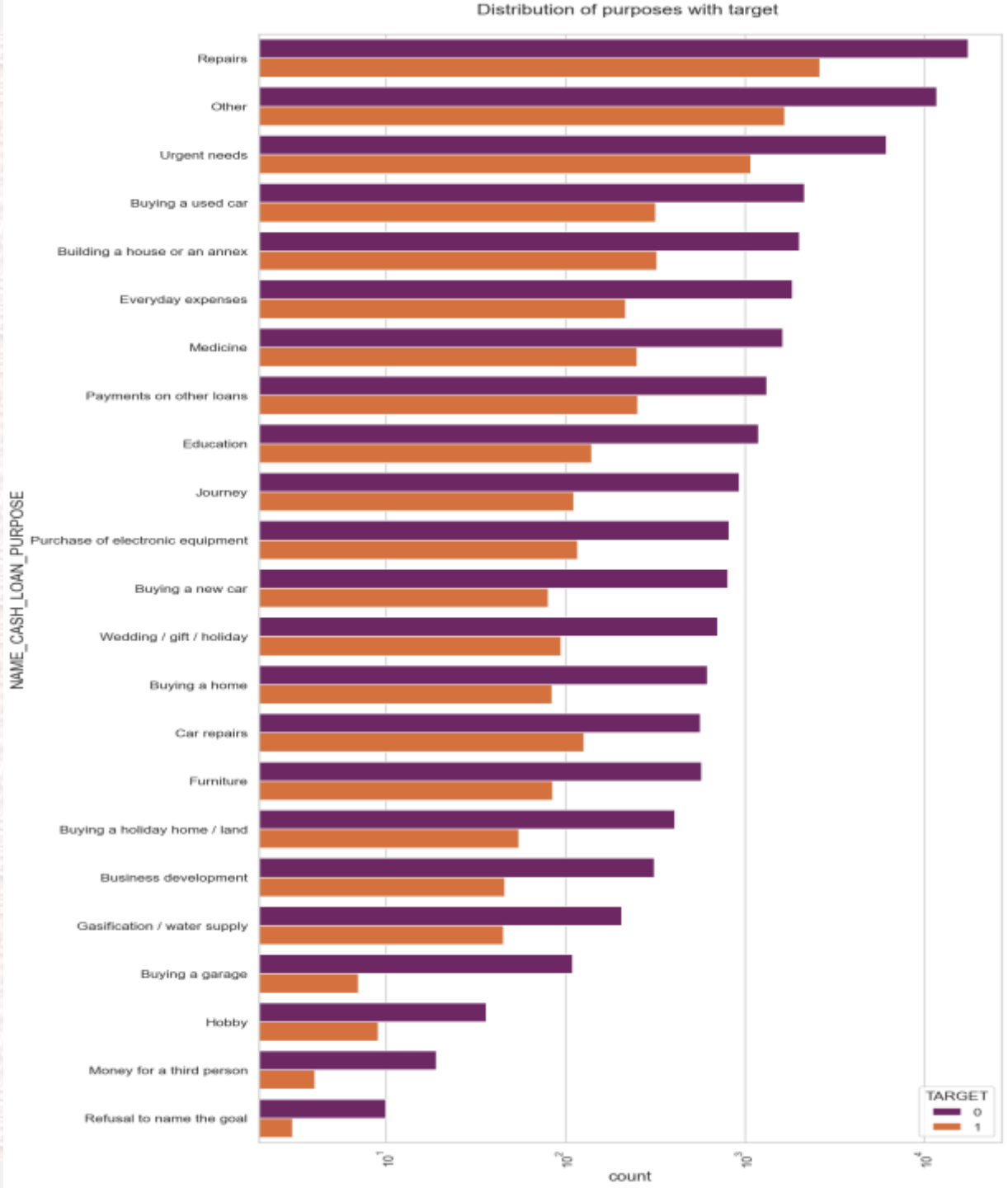


Univariate Analysis on Final Dataset

Logarithmic Comparison of Contract Status

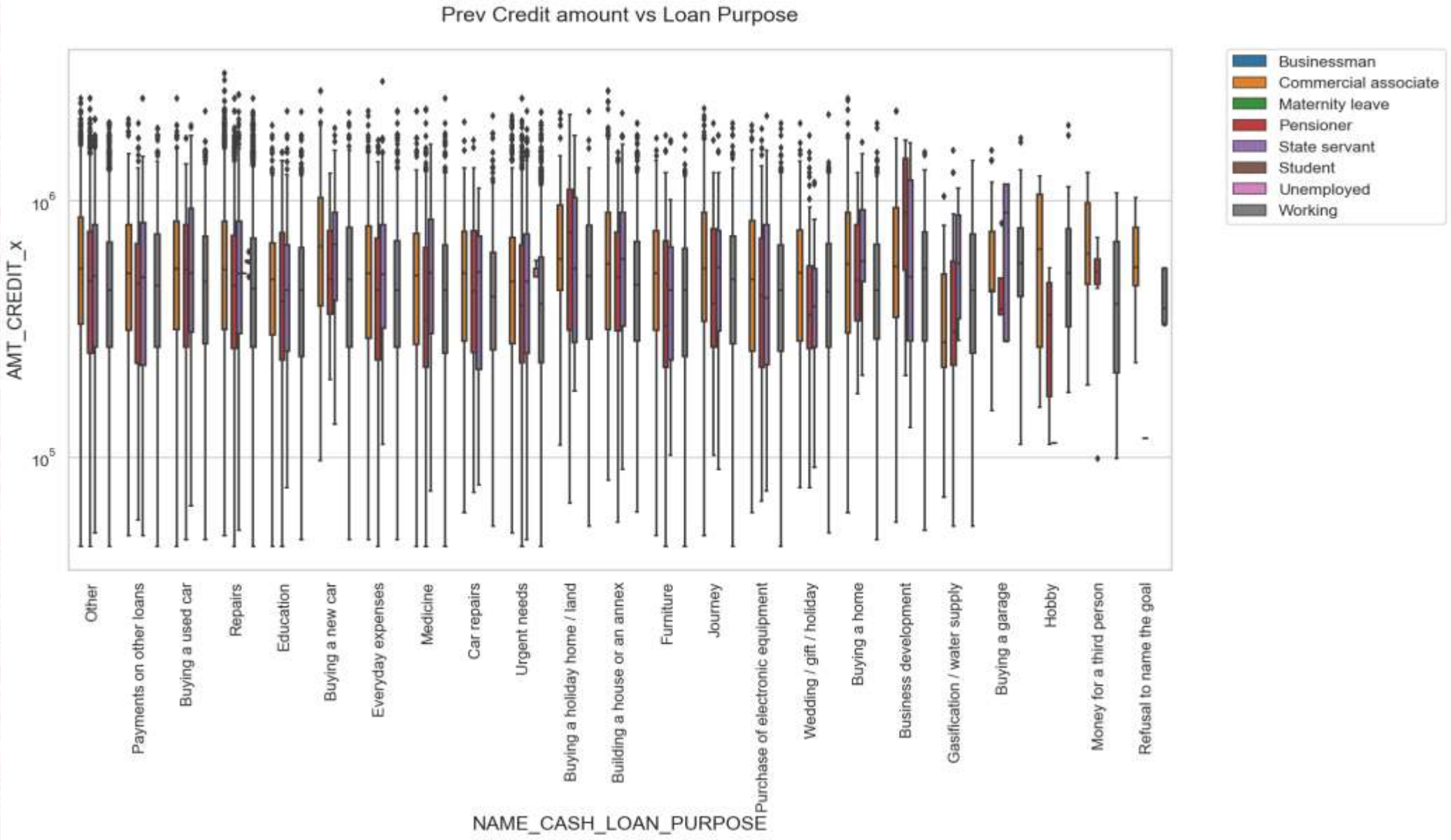
Inference:

As we can see that there is disparity between the two categories target 0 and 1 but loan purposes of 'Repairs' are the highest amongst all purposes. There are few places where loan payment is significantly higher than facing difficulties.
They are 'Buying a garage', 'Business development', 'Buying land', 'Buying a new car' and 'Education'
Hence, we can focus on these purposes for which the client is having for minimal payment difficulties.



Bivariate Analysis on Final Dataset

Logarithmic Comparison of Credit Amount

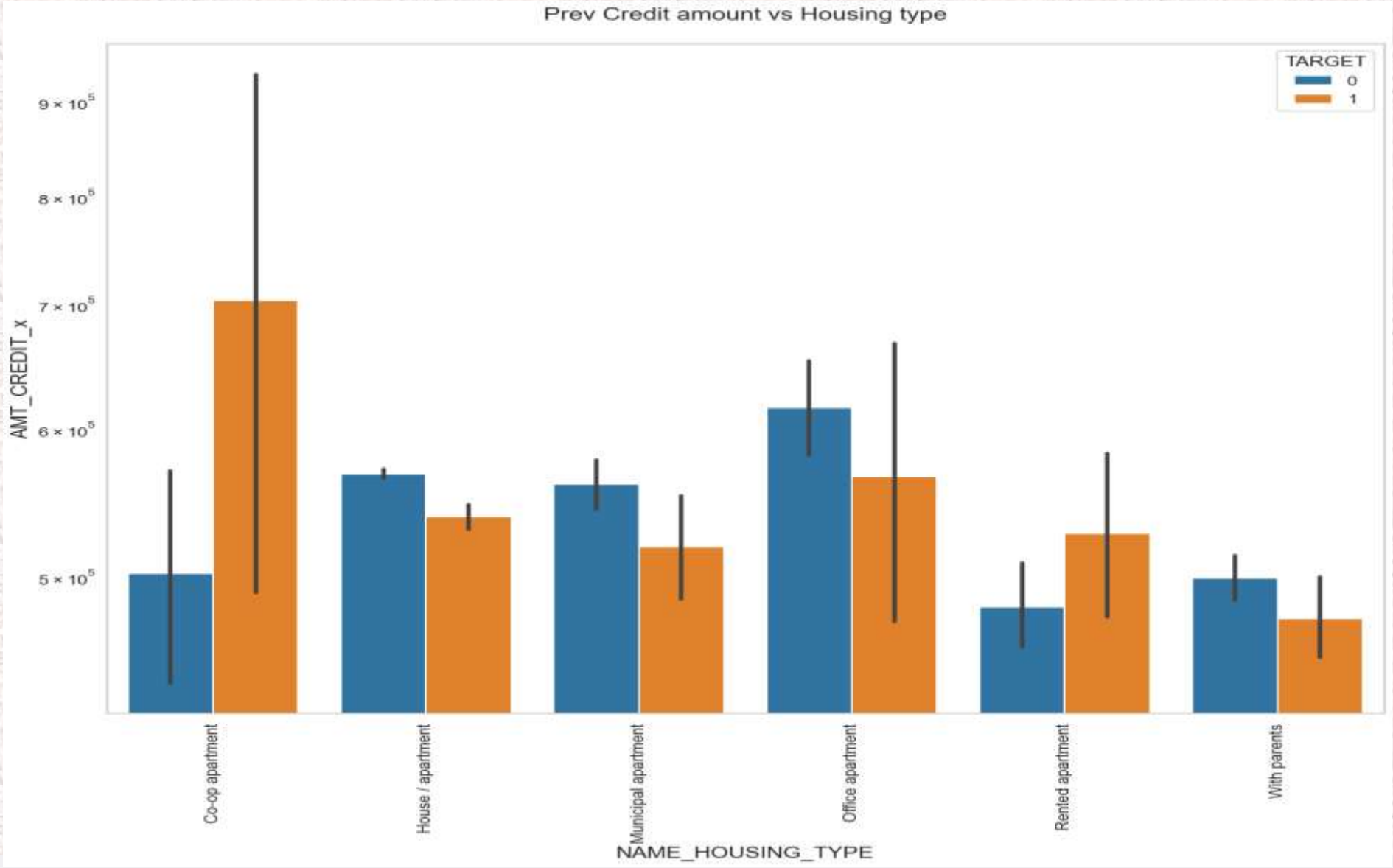


Inference:

The credit amount of Loan purposes like Buying a home, land and car and Building a house is higher.
State servants have a significant amount of credit applied.
Money for third person or a Hobby is having less credits applied for.

Bivariate Analysis on Final Dataset

Logarithmic Comparison of Credit Amount vs Housing type



Inference:

The Housing type, office apartment is having higher credit for target 0 and co-op apartment is having higher credit for target 1. Thus, it is safe to conclude that bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment. Bank can focus mostly on housing type with parents or House or apartment or municipal apartment for successful payments.

Conclusion on Final Dataset

1. Bank should focus on making cash loans more accessible as majority of the people are opting for that also the revolving loans have equal number of approval and refusal which makes the revolving loan less desirable.

2. Banks should also focus on new clients as they are more likely to get approved and less likely to default.

3. Banks should focus more on female candidates as they are greater in number when it comes to applicants.

4.People from secondary background should be the target area where banks should concentrate as they are the leading majority followed by higher education people.

5.Working class people and commercial associates are the highest number of people who should be targeted but they should also put more focus on Pensioners and state servants as they are the new emerging majority who are approved more often and are less likely to default.

6.Married people are leading in the majority when it comes to approval and rejection, but banks should focus on single people separated and widowed.

7. As stated earlier, cash through bank is a leading form or loan giving but bank should also focus on non-cash loan from bank accounts.

8. When we talk about the occupation type the banks should keep in mind the IT and HR staff as they are the least likely to default and target them the most.

9.The banks should shift their focus from urgent needs as they have no unused offers that means they are more likely to have more defaulters.

Conclusion on Final Dataset

10. Banks should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment. Also, Bank can focus mostly on housing type with parents or House/apartment or municipal apartment for successful payments.

11. Banks can also focus on loan purpose of buying house or apartments, cars and additionally, they can come out with creative schemes that helps loan applied for third person as that purpose has significantly low amount of credit applied.

References

- upGrad. (n.d.). *Credit EDA Assignment*. https://learn.upgrad.com/
- Dash, R., Kremer, A., Nario, L., & Waldron, D. (2017, June 6). *Risk analytics enters its prime*. McKinsey & Company. https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/risk-analytics-enters-its-prime
- Janaha, V. (2023, May 3). *Data Analytics in banking and Financial Services*. Zuci Systems. https://www.zucisystems.com/blog/how-is-data-analytics-used-in-finance-and-banking-sector/#:~:text=While%20fraud%20reduction%20is%20a%20common%20goal%20for,levels%20of%20monitoring%20and%20verification%20to%20those%20accounts.

THANK YOU