Credit EDA Case Study

By

MAYUKH DAS

(DSC 54 – March 2023)

Overview

This is a case study of a bank, which gives loans to their customers. We are here to check the output that which customers are eligible and who are not and many more.

Here, we will use Exploratory Data Analysis to solve the Case Study. So, without wasting any time, Let’s START.

Problem Statement

* ***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.

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.

Problem Statement

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

Approved: The Company has approved loan Application

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.

Refused: The company had rejected the loan (because the client does not meet their requirements etc.).

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***

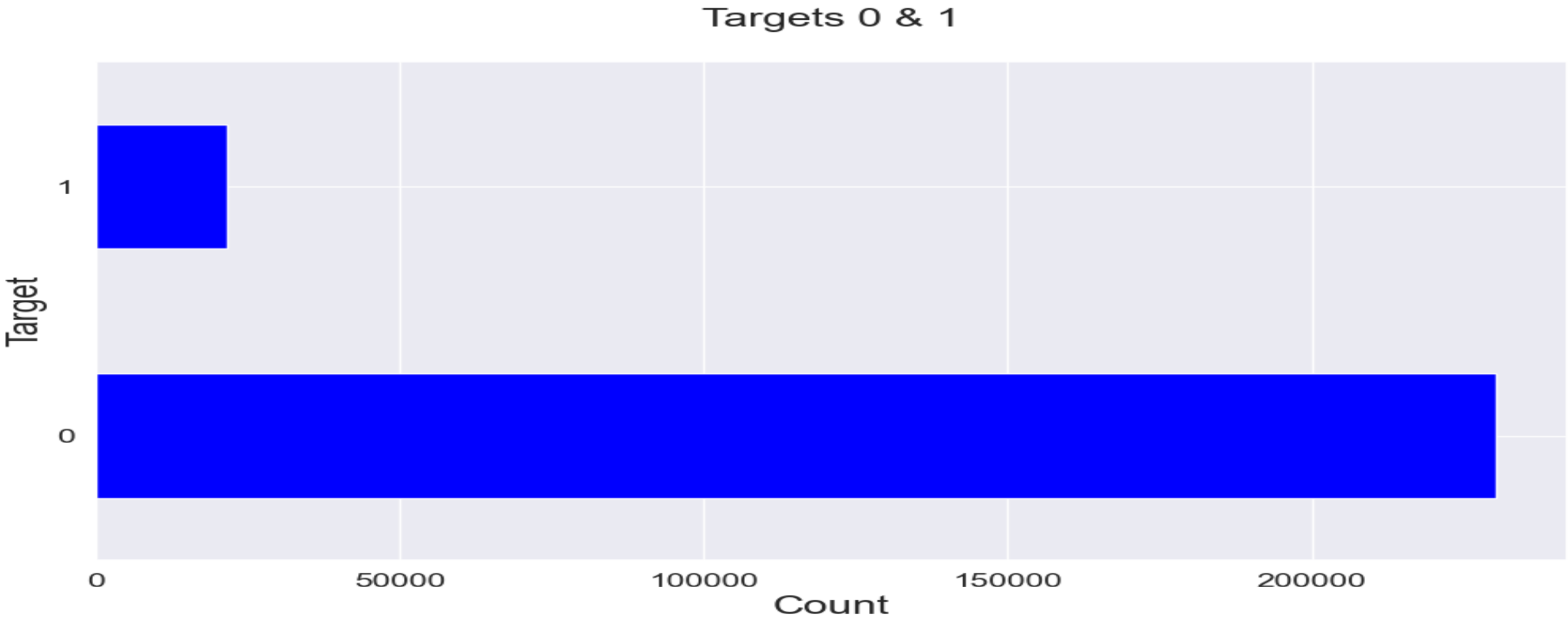
This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise 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.

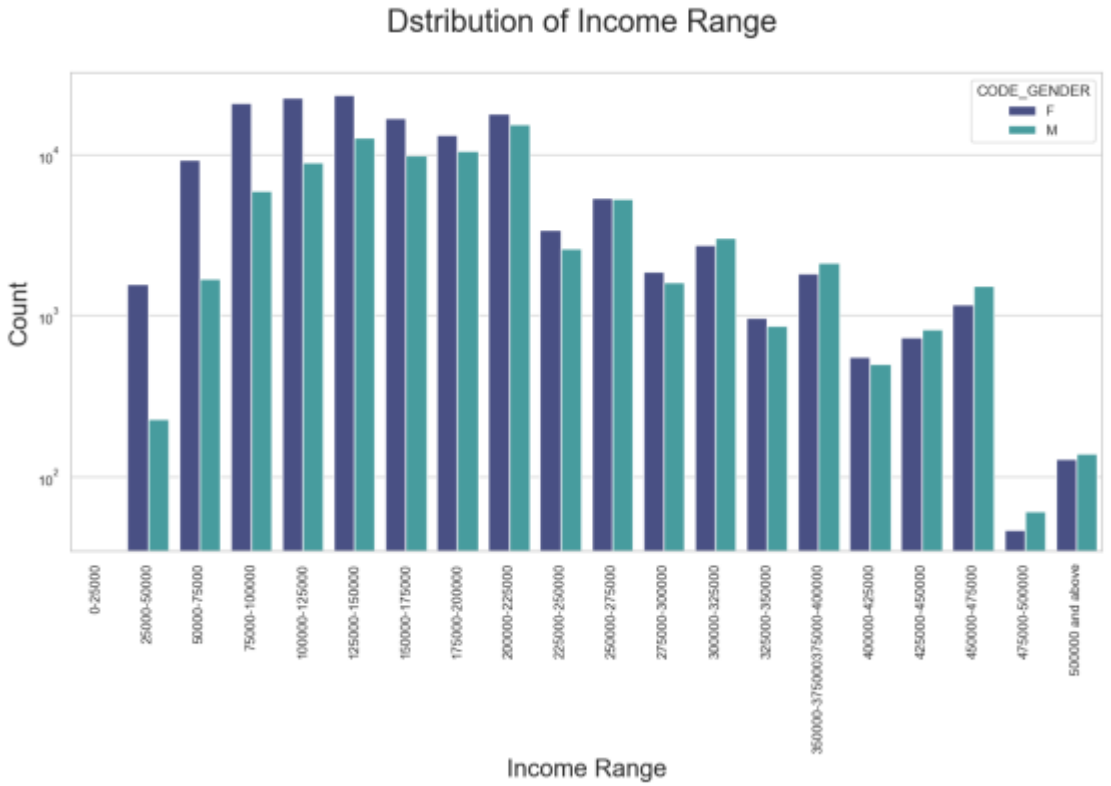
APPROACHES

- Upload the Application DataSet and Previous DataSet in Jupyter Notebook.
- Data Cleaning in Application DataSet – Describe > Null values > Null Values Percentage > Mean Values > Recheck.
- Data Cleaning in Previous Application DataSet – Null values > Null Values Percentage > Removing Null values more than 35% > Recheck.
- Inputting Missing values in Application DataSet > Checking the Data-Types of the columns > Removing Unwanted columns from Application DataSet > Gender & Organization > Checking total **INCOME** and **CREDIT >** Imbalance ratio.
- **Bivariate Analysis** of Numerical columns (**Target_0** and **Target_1**).
- **Univariate Analysis** with **Target_0** and **Target_1** people.
- Finding Outliers
- Defining the Correlation of **Target_0** and **Target_1** > Top 10 Correlation between **Target_0** and **Target_1.**
- **Multivariate Analysis**
- Merging two datasets
- Combo work in new DataSet.
- Conclusion.



Here, ‘Target=0’ means, the people those who are non-defaulters

And, ‘Target=1’ means, the people those who are defaulters

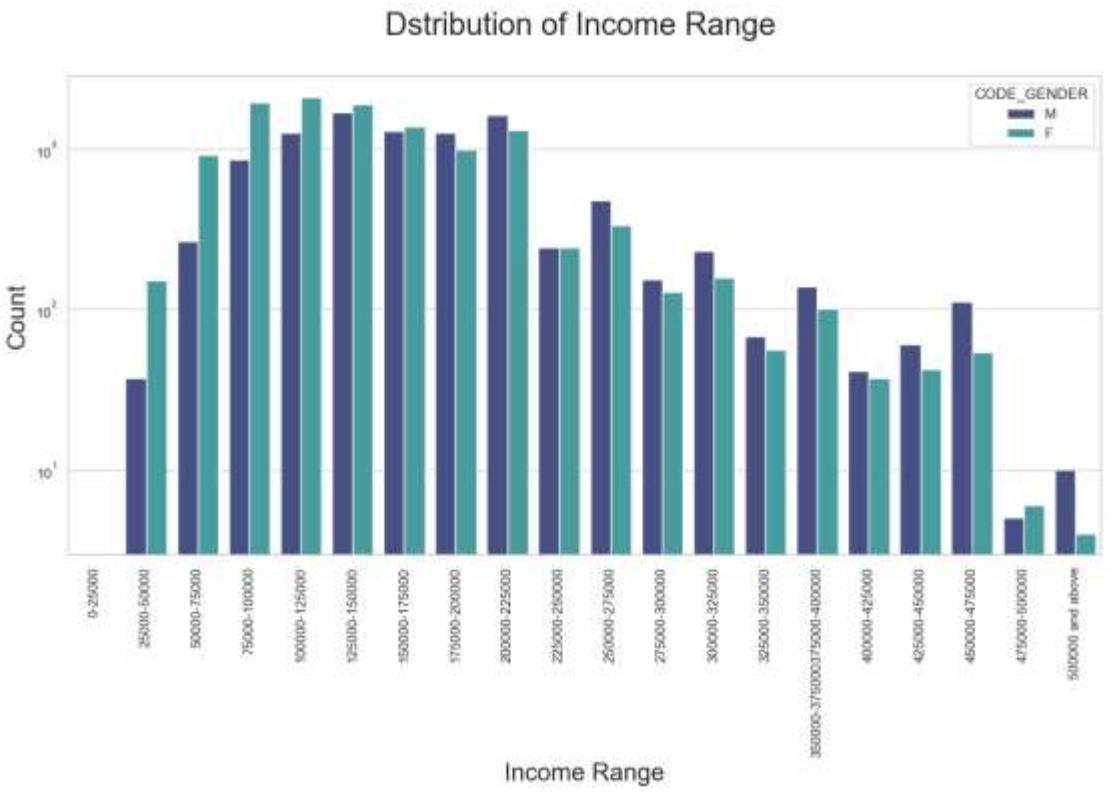


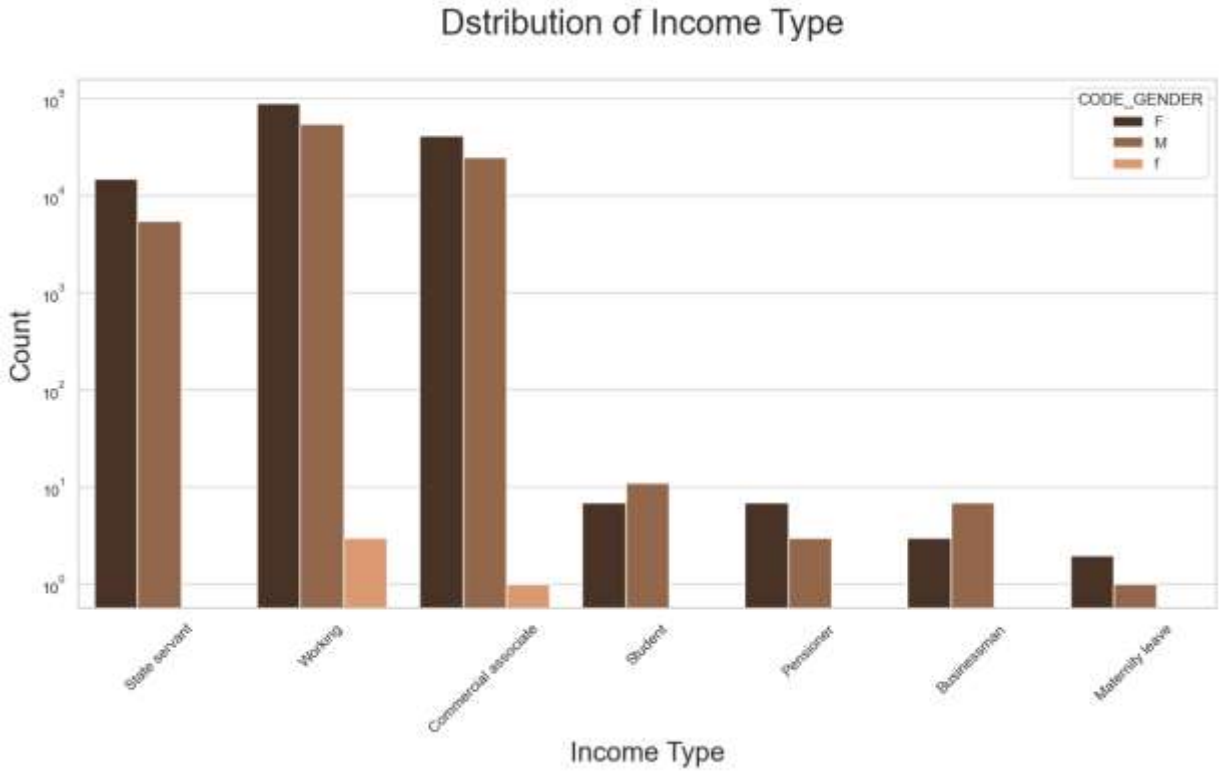
Target 0

1. Income range from 125000 to 150000 is having more number of credits.
2. Very less count from range 450000 to 475000.
3. It seems that the females are more than male is having credit of range 125000 to 150000.

1. Male accounts are higher.
2. Income range from 100000 to 200000 is having more number of credits.
3. Les count for income range 450000 to 475000.

Target 1



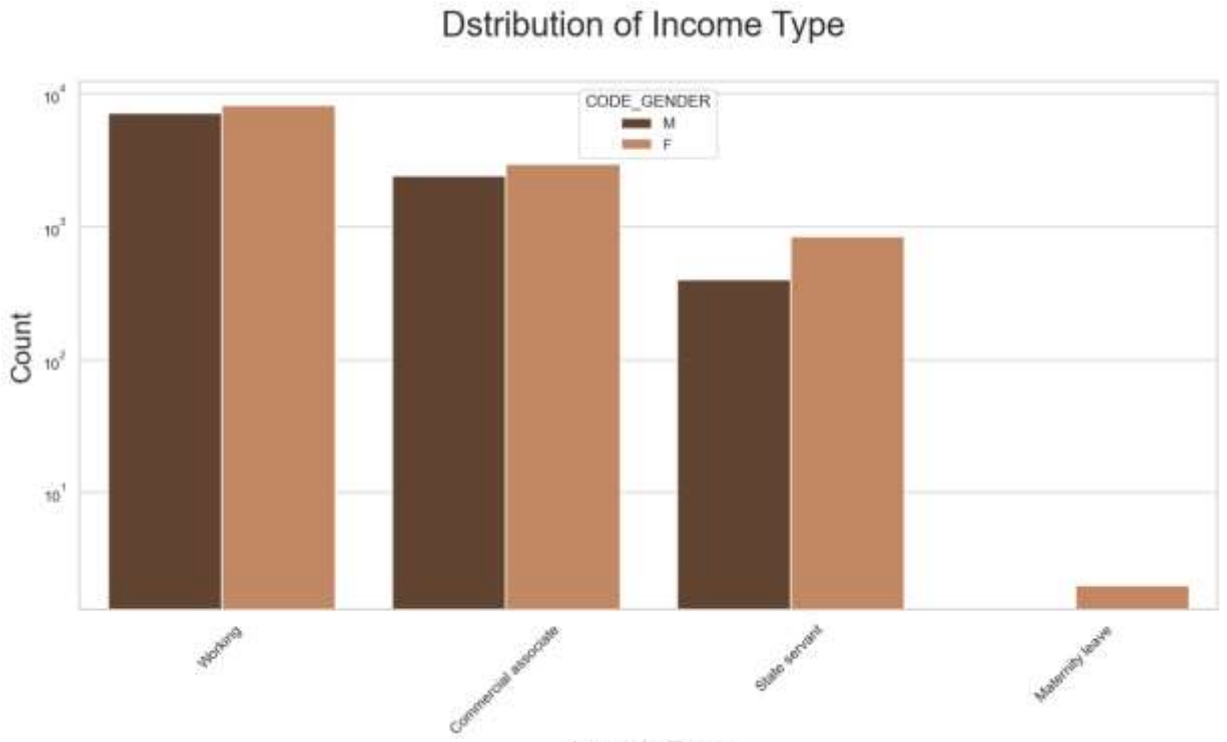


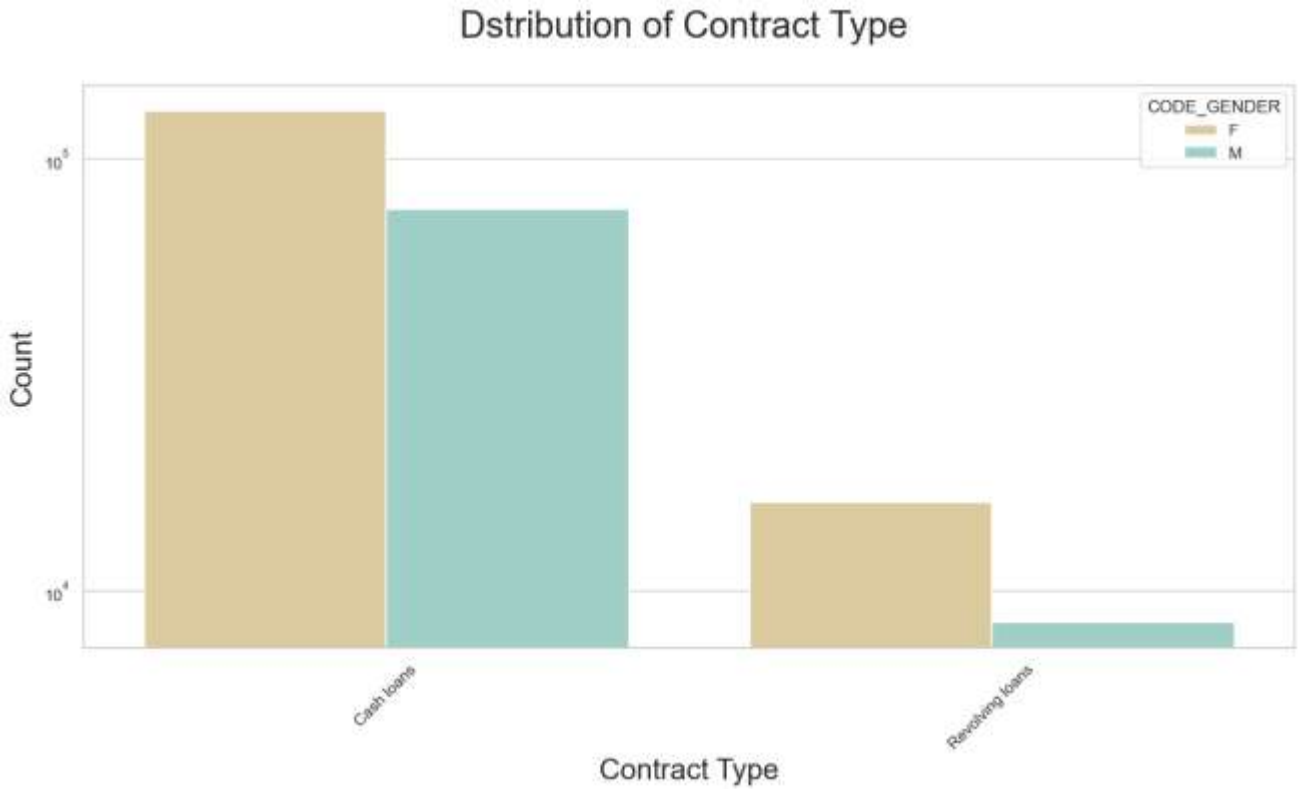
Target 0

1. It seems that working women have most credit than others.
2. It seems that state servant ‘working’ and ‘commercial associate’ have more credit counts compared to others.
3. It seems that women in maternity leave has less credit in comparison to others.

1. For income type ‘working’, ‘commercial associate’ and ‘state servant’ the number of credit are higher than other. just for example ‘maternity leave’.
2. For these females are having mode number of credits than male.
3. Less number of credits for income type ’maternity leave’.

Target 1



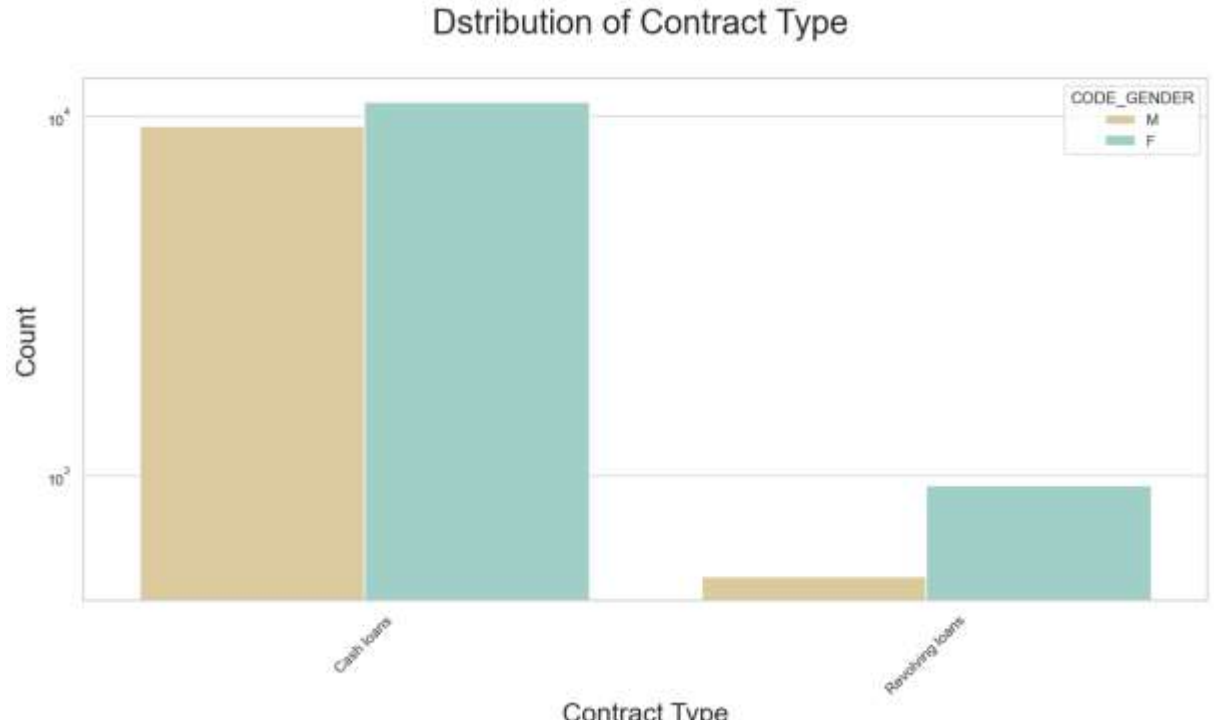


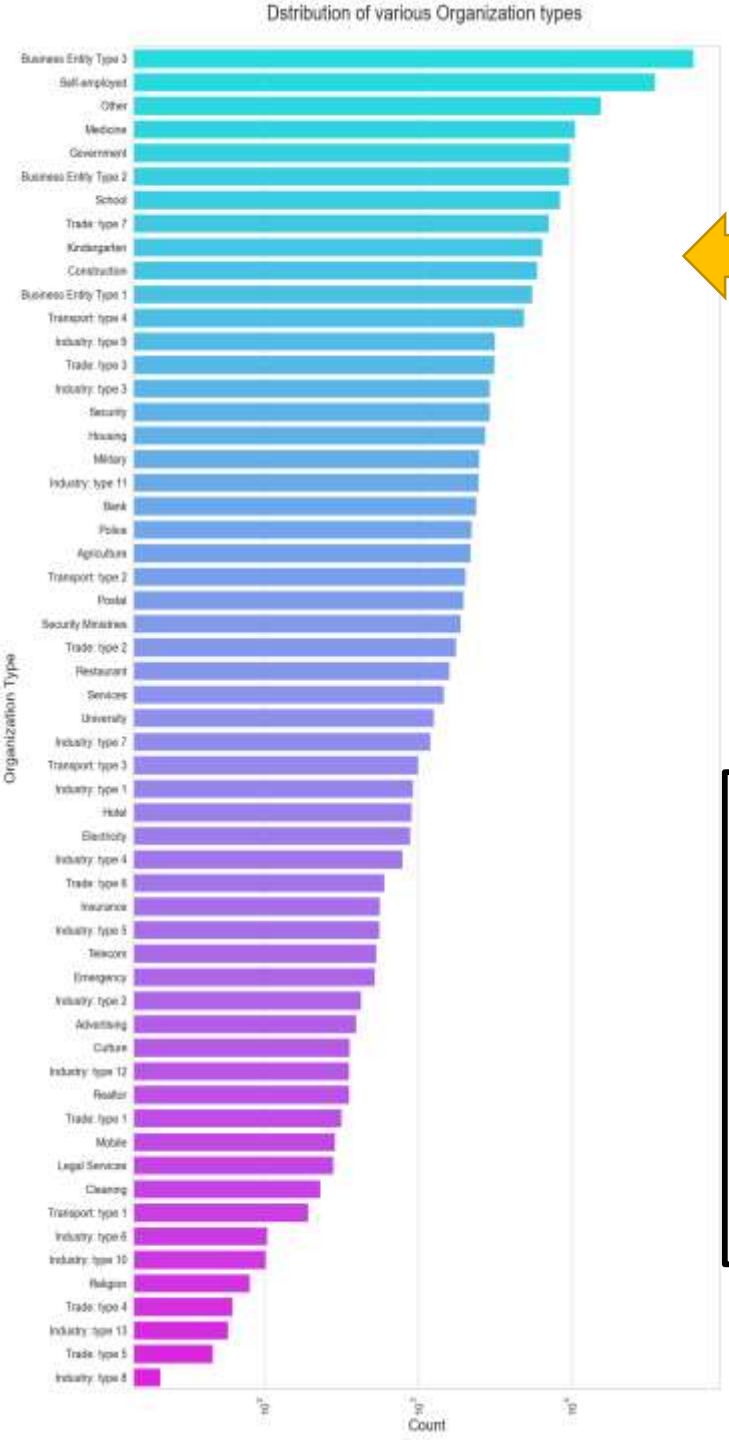
1. For contract type ‘cash loans’ is having hard number of credits than ‘revolving loans’ contract type.
2. For this also female is leading for applying credits.

Target 1

Target 0

1. It seems that ‘cash loans’ is having high number of credits then 'revolving loans' contract type.
2. Also, female applies for more than credit.



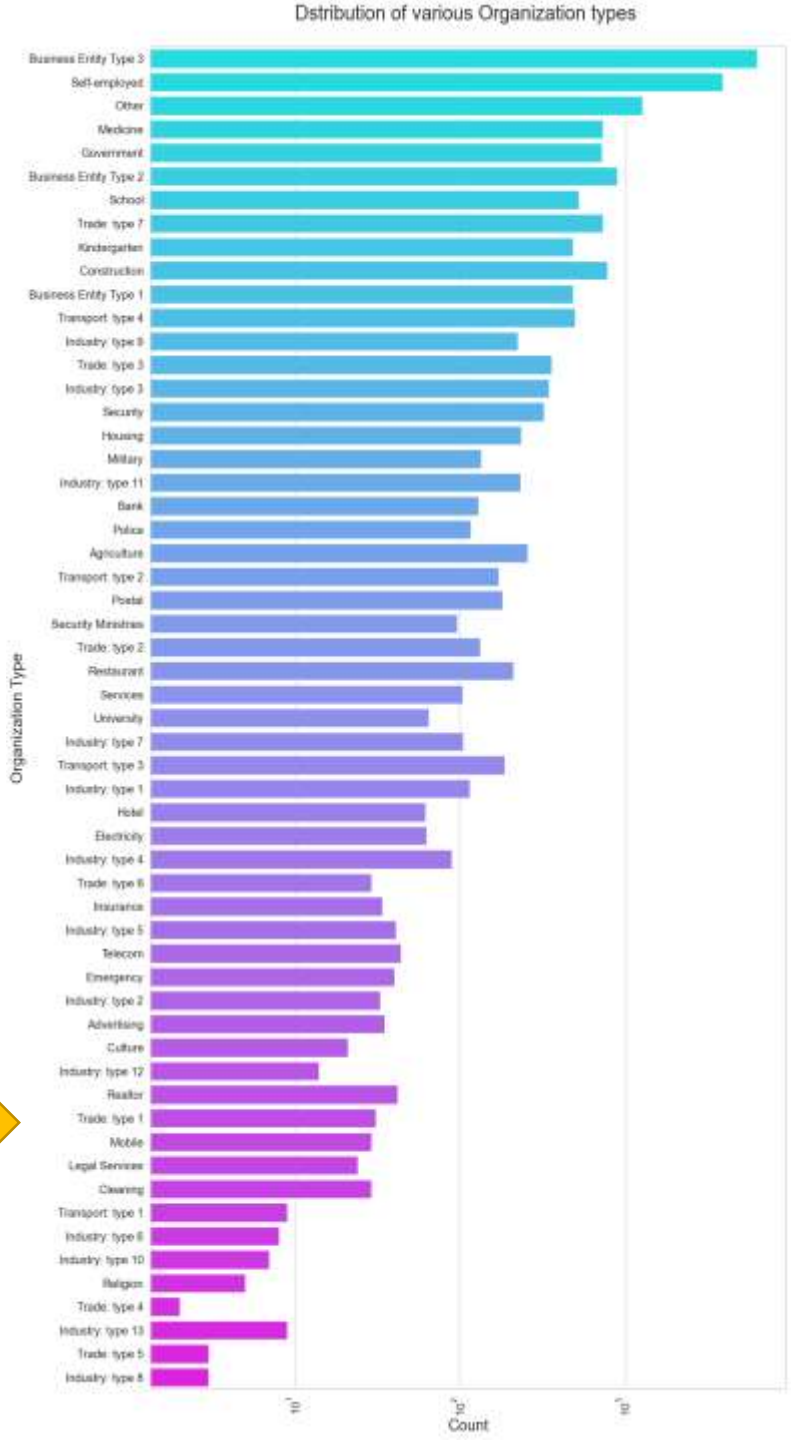


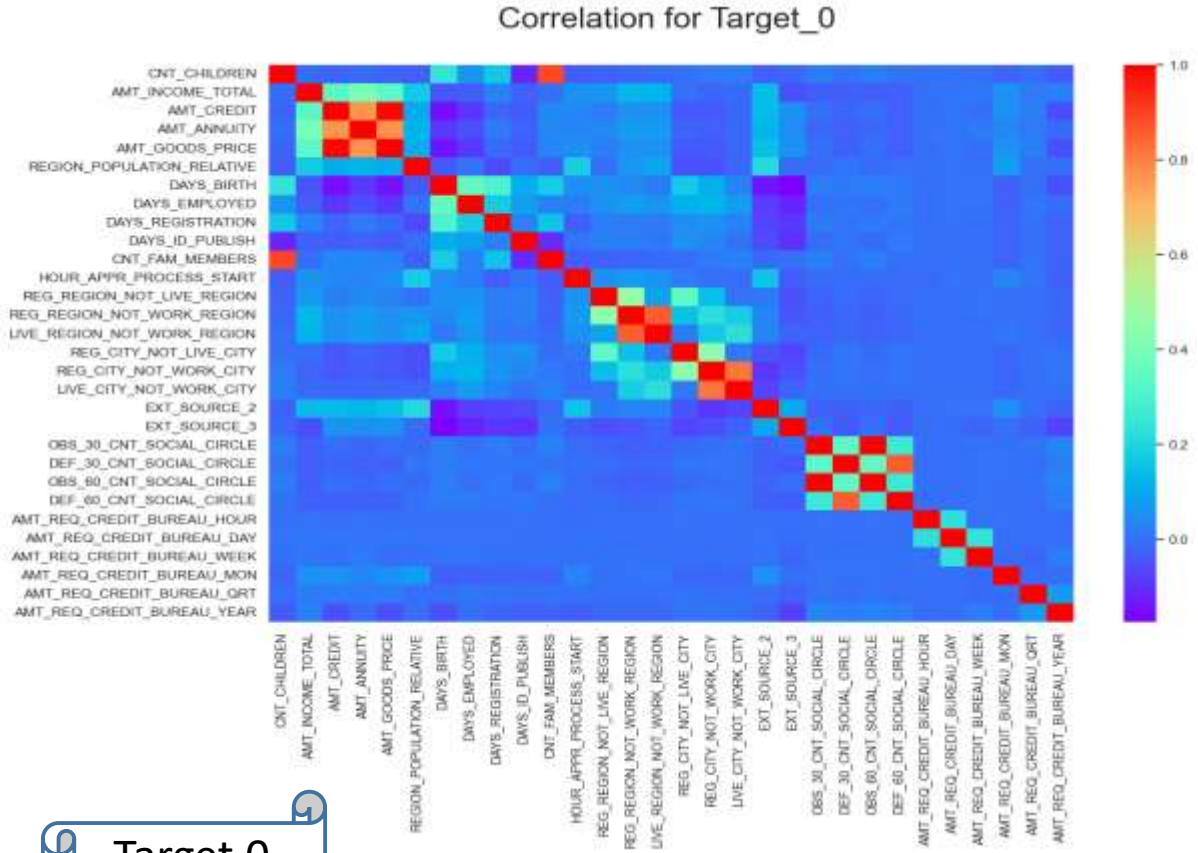
1. Clients which have applied for credits are from most of the organization type ‘business entity type 3’, ‘self employed’, ‘other’, ‘medicine’ and ‘government’.
2. Less clients are from industry type 8, type 6, type 10, religion and trade type 5 ,type 4.

Target 0

1. Clients which have applied for credits and from most of the organiztion type ‘business entity type 3’, ‘self employed’, ‘other’, ‘medicine’ and ‘government’.
2. List clients are from industry type 8, type 6, type 10, religion and trade type 5, type 4.
3. Same as type zero is distribution of organisation type.

Target 1



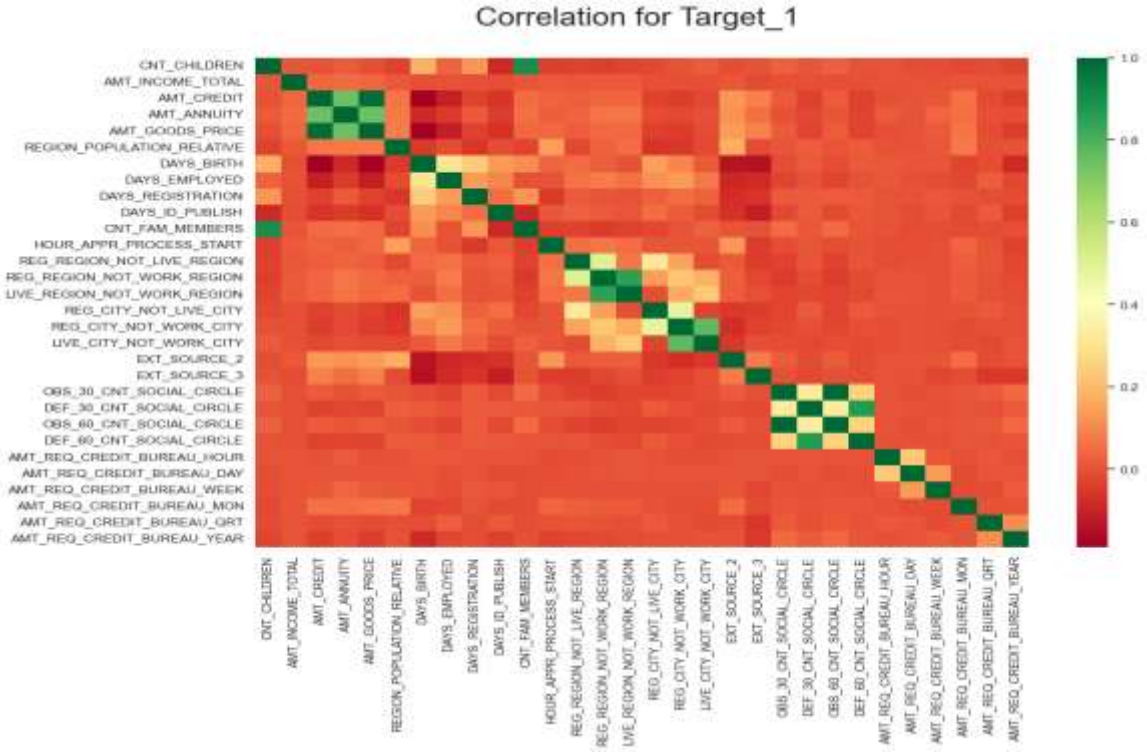


Target 0

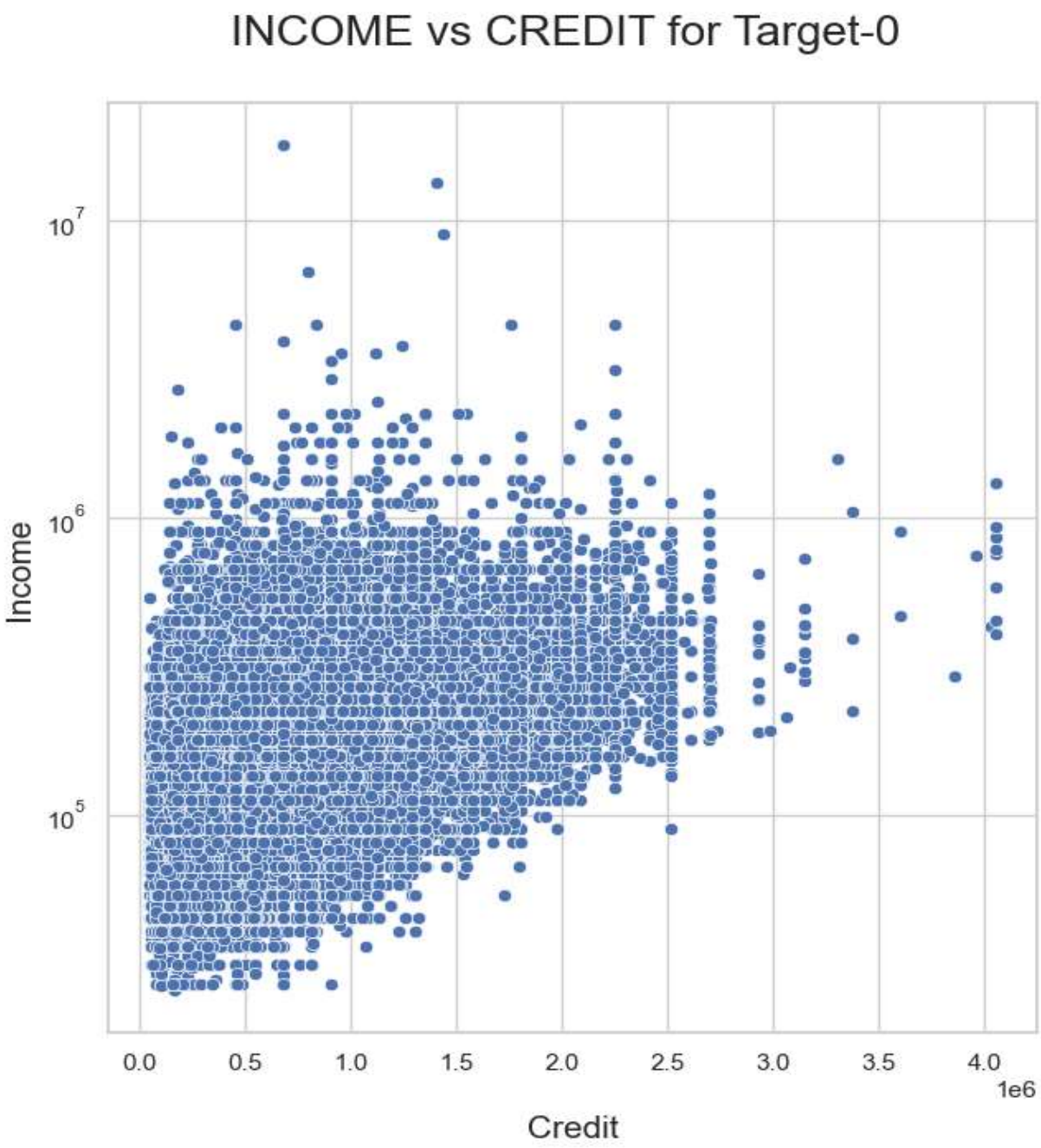
1. Credit amount is inversely proportional to the date of birth which means credit amount is higher for low age and vice-versa.
2. Income amount is inversely proportional to number of children client have means more income is for less children client have and vice versa.
3. Less children client have in densely populated area.
4. Credit amount is higher to densely populated area.
5. The income is also higher intensely populated area.

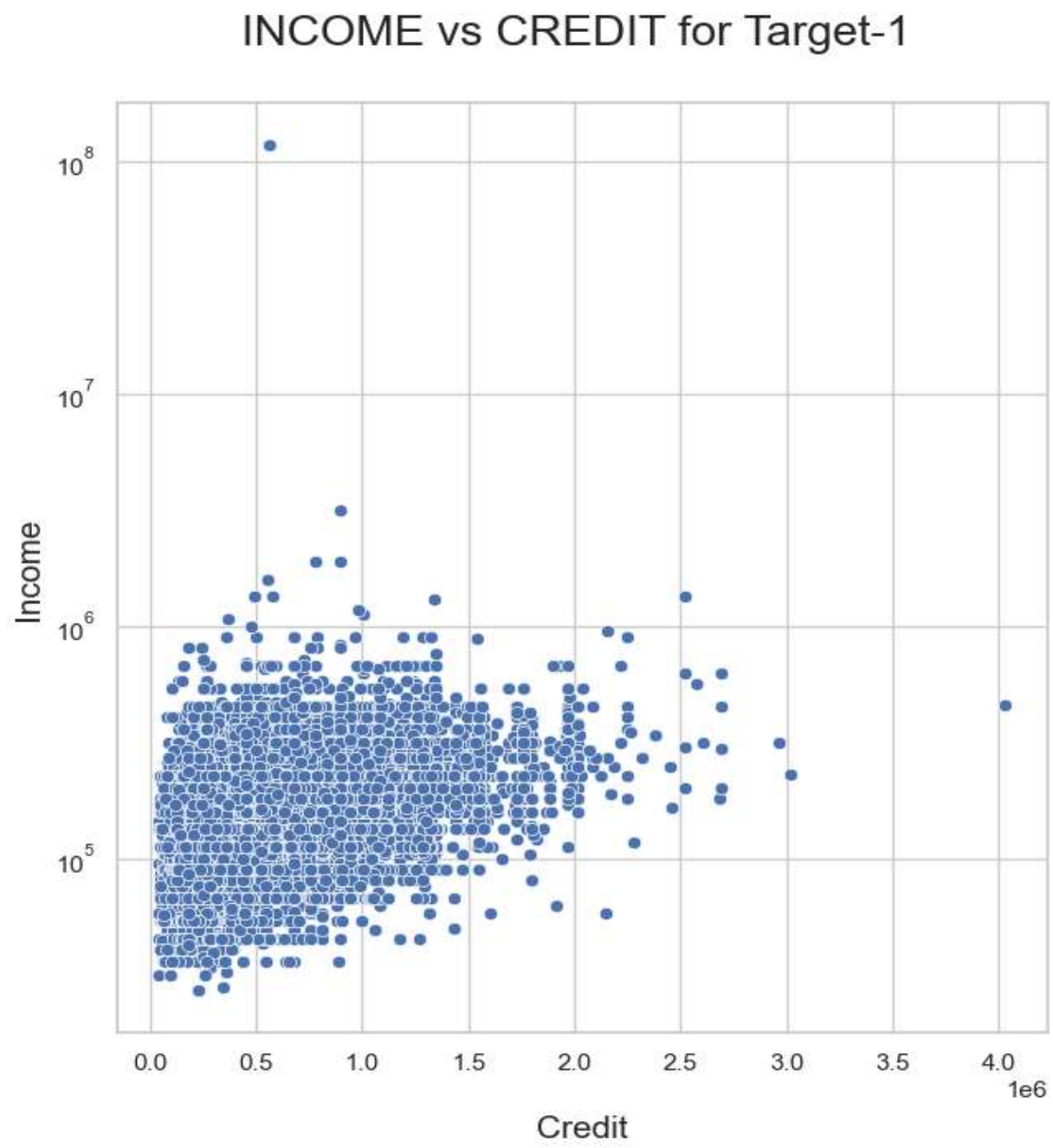
1. The clients permanent address does not match contact address and having less children and vice versa.
2. The clients permanent address does not match work address and having less children and vice versa.

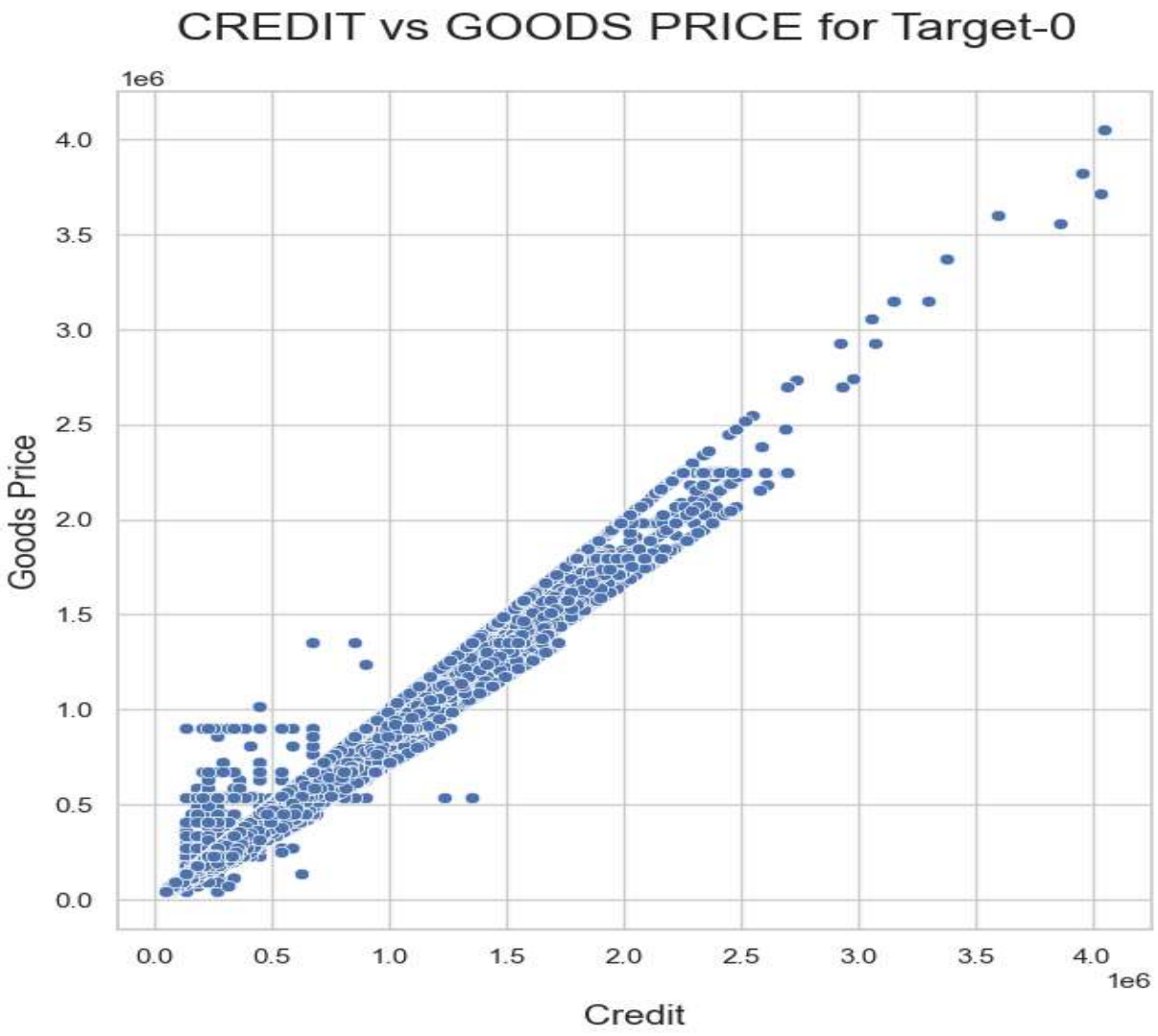
Target 1

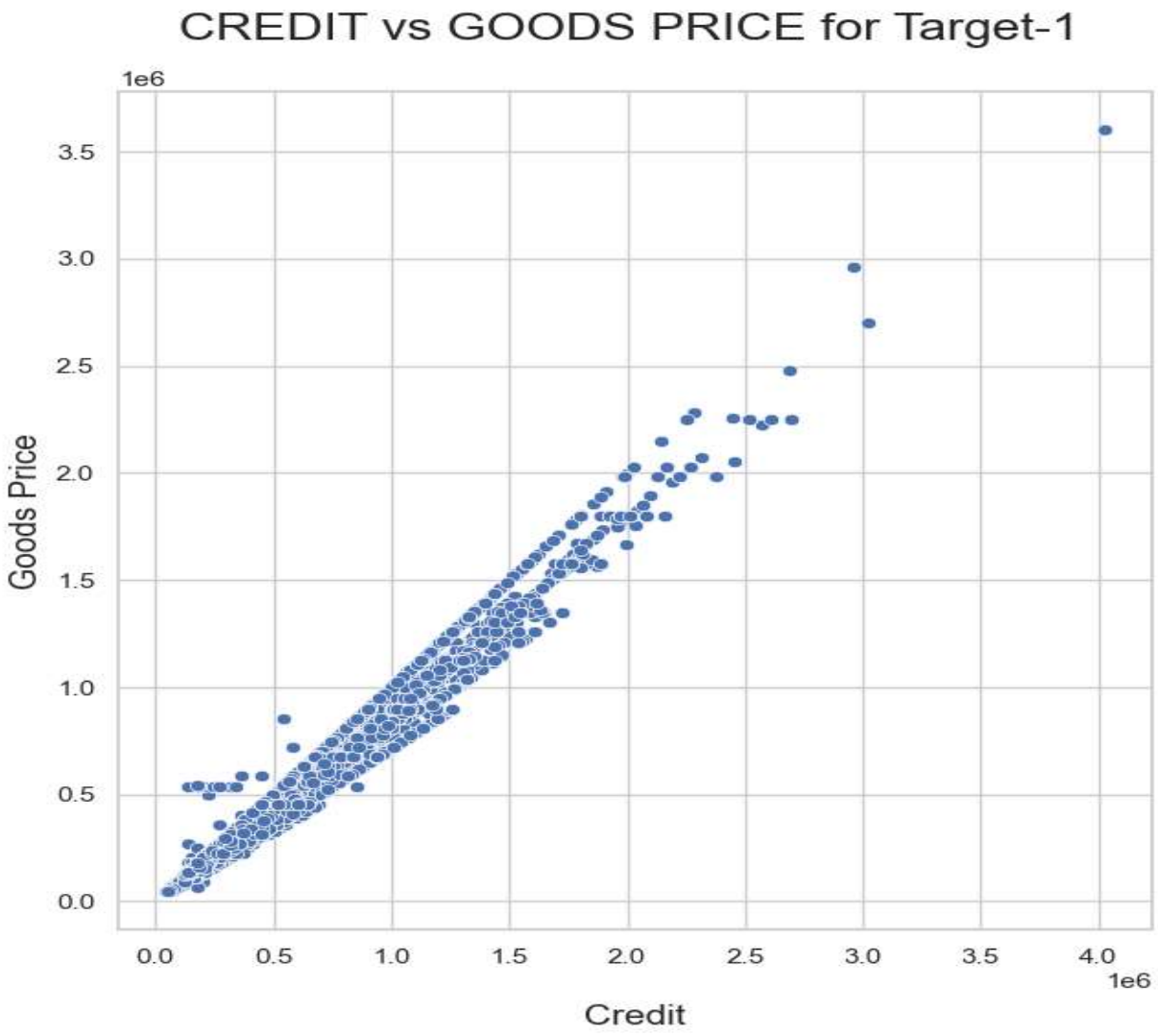


Bivariate Analysis


INCOME vs CREDIT for Target-0
Income
Credit
1e6


INCOME vs CREDIT for Target-1
Income
Credit
1e6





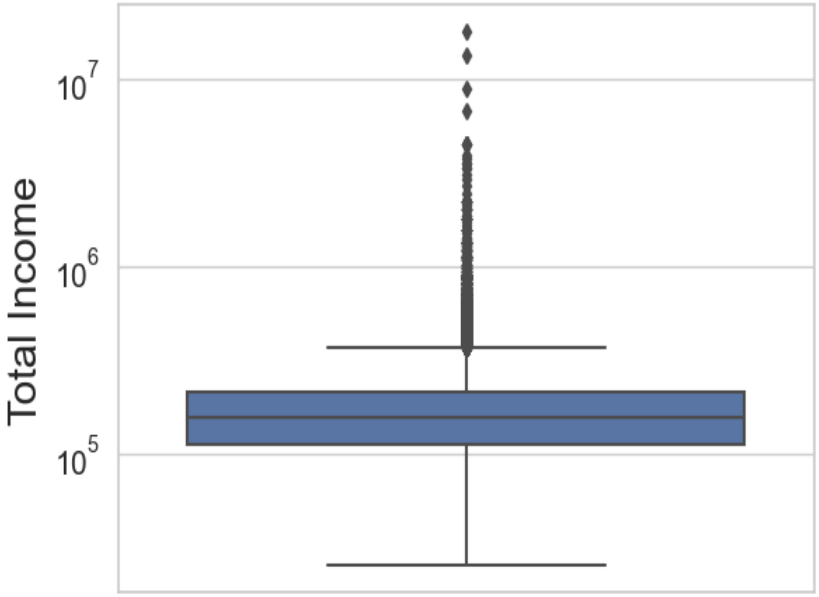
With the scatter plot we can determine that AMT credit and AMT goods price and highly correlated with means if increase in goods price the credit increased directly and vice versa.

Finding Outliers

Univariate Analysis

OutLiers

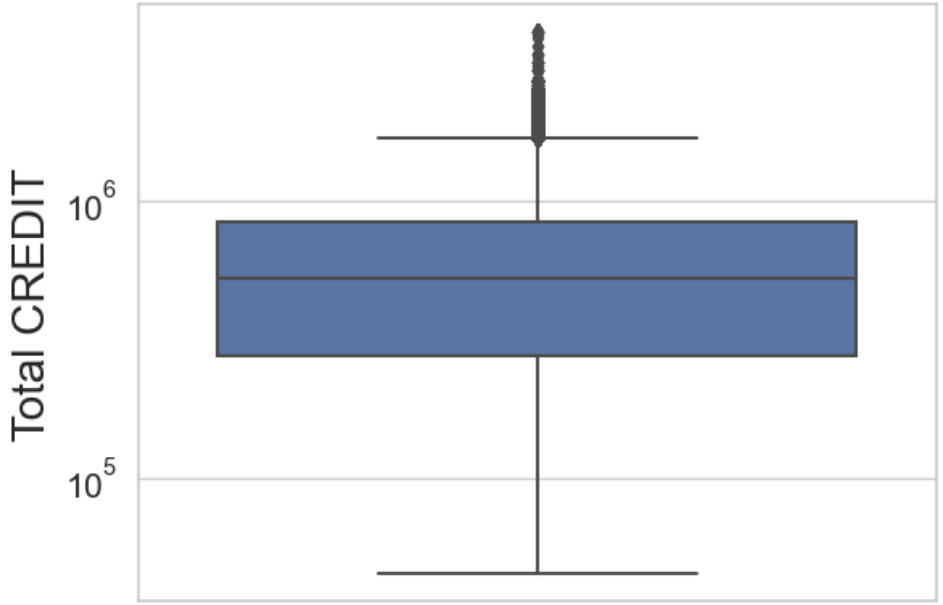
Distribution of Income Amount



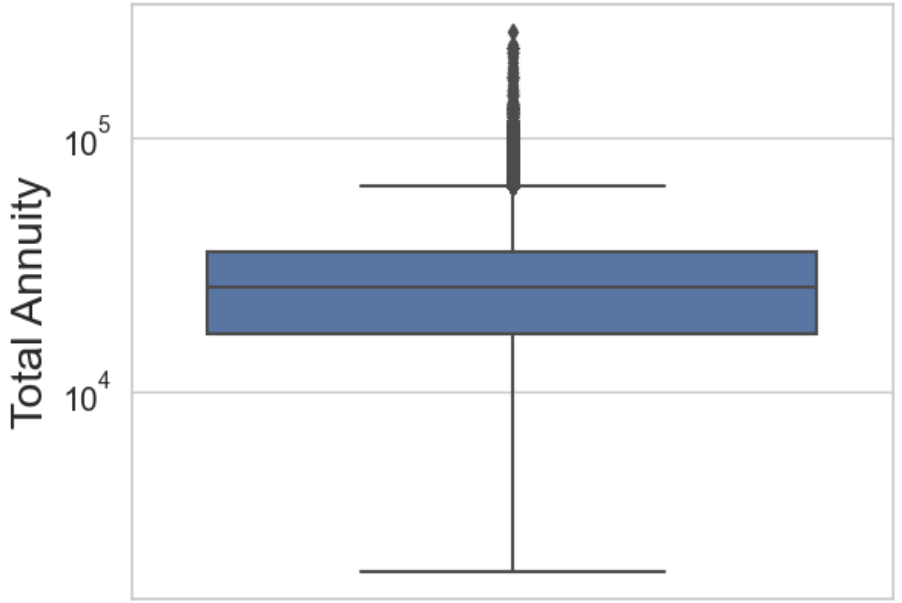
1. There seems to be an equal distribution of the income amount of the clients.
2. Also some of the outliers present in the data set.

1. The first quartile is bigger than the third quartile. That means most of the client's credit lies in the first quartile.
2. There seems some outliers in the credit box plot.

Distribution of CREDIT Amount



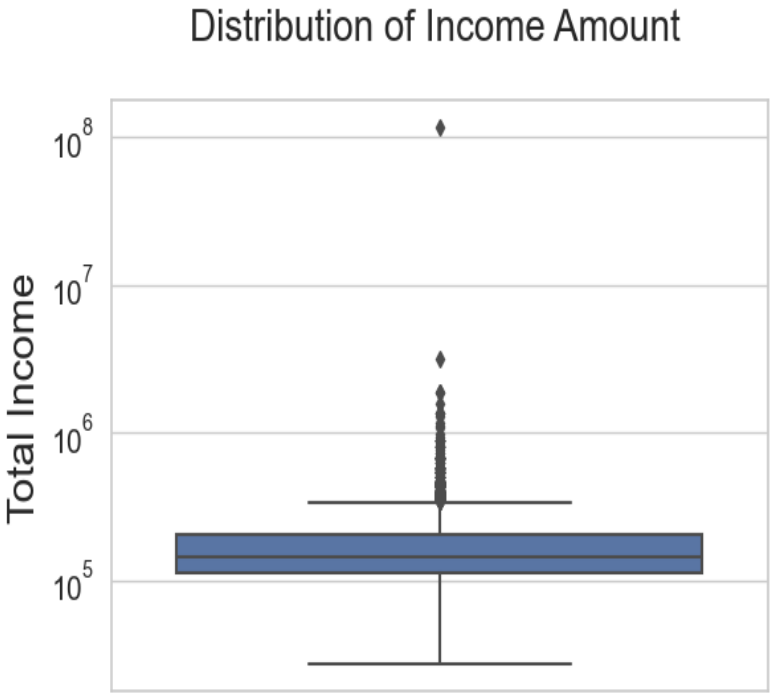
Distribution of Annuity Amount



1. The first quartile is bigger than the third quartile.
2. Their seems some outliers in the annuity box plot.

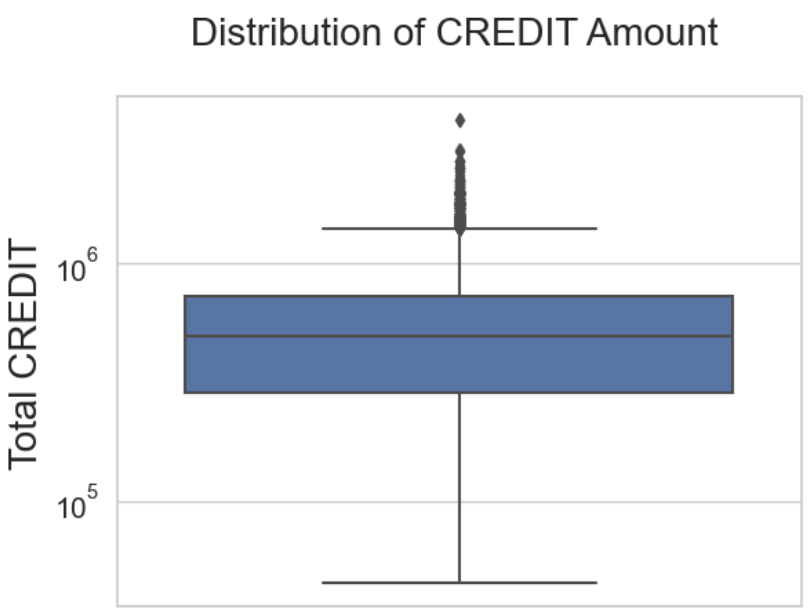
Target 0

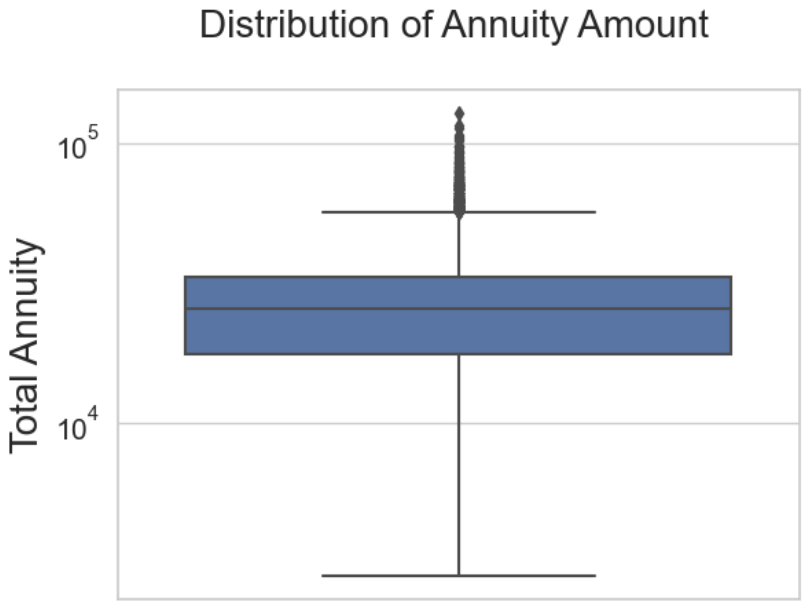
OutLiers



1. Their seems a significant outlier in the income data set.
2. Most of the income of the client lies on the third quartile.

1. The first quartile is bigger than the third quartile.
2. most of the client lies in the first quartile.

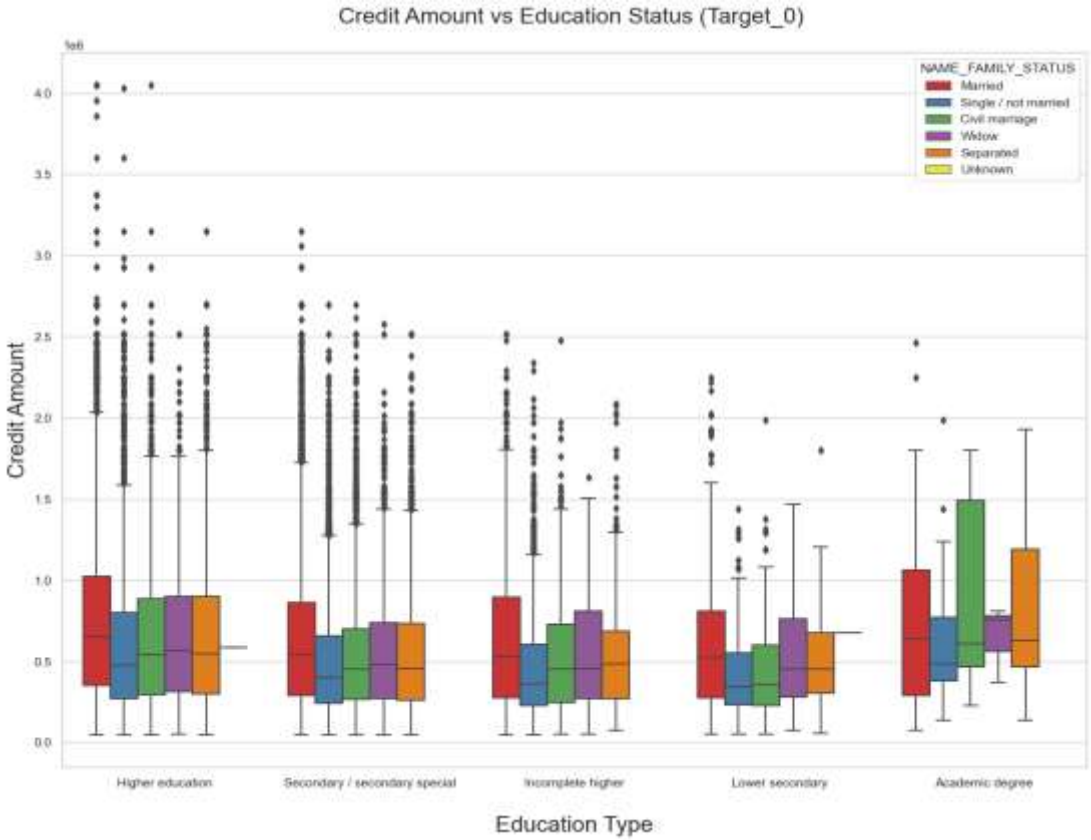




1. The first quartile is bigger than the third quartile.
2. there seems some outlets in the annuity box plot.

Target 1

Multivariate Analysis

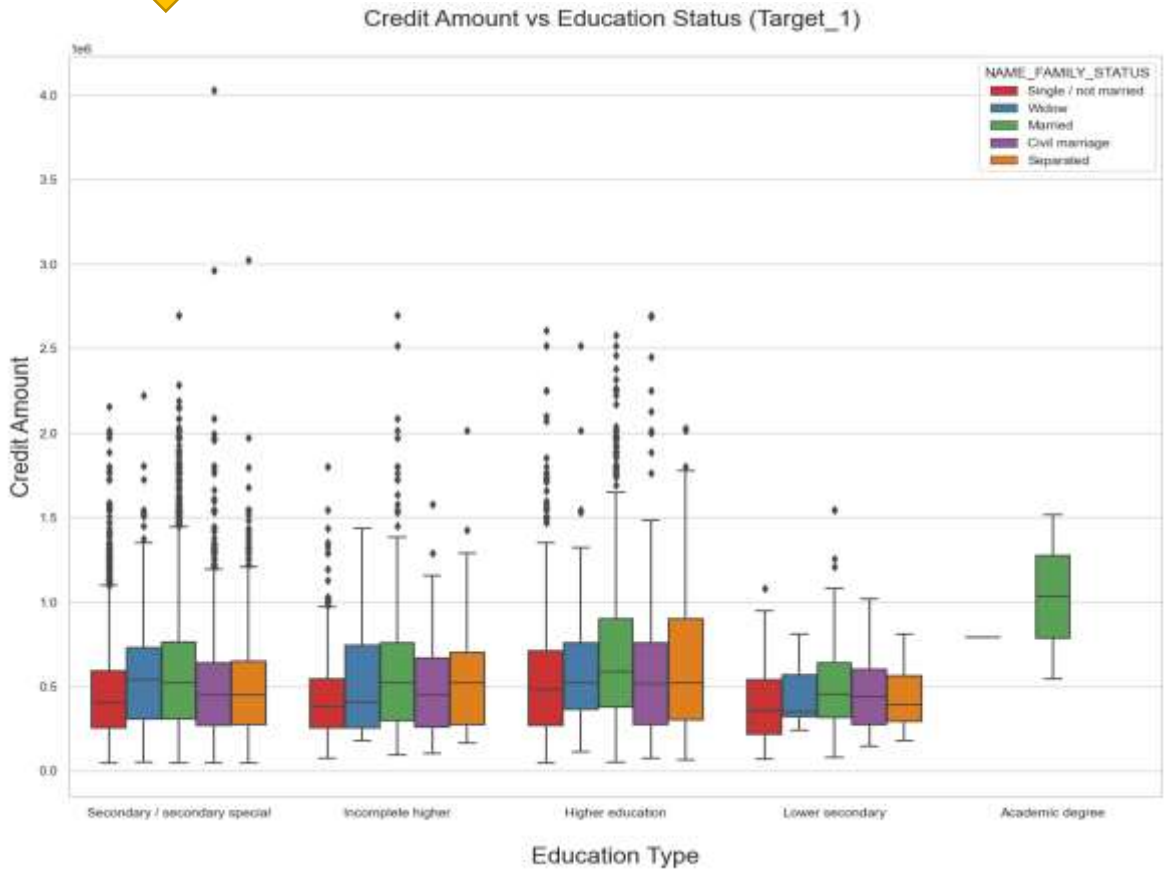


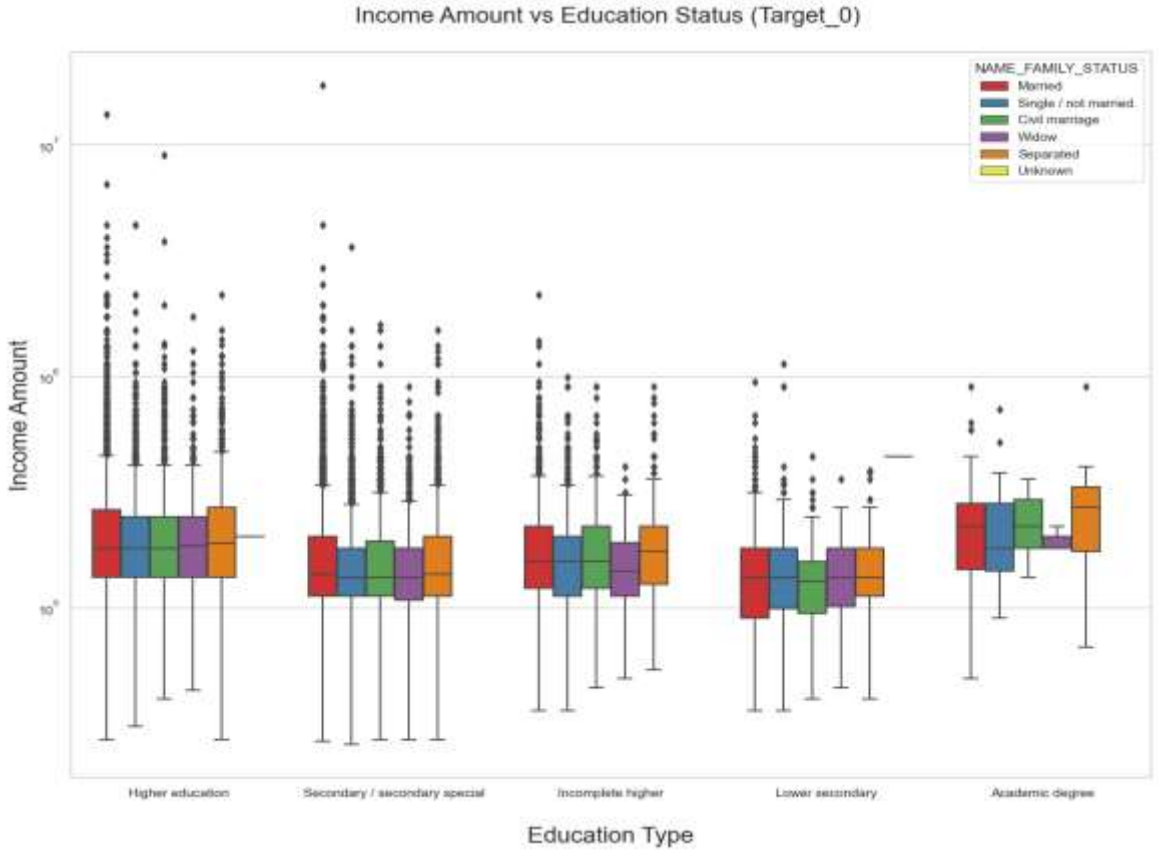
Target 0

For the above box plot we can conclude that Family status of ‘civil marriage’, ‘marriage’, and ‘separated’ of Academic degree education are having higher number of credits than others. Also, higher education of family status of ‘marriage’, ‘single’ and ‘civil marriage’ are having more outliers.

From the above box plot we can say that family status of ‘civil marriage’, ‘marriage’ and ‘separated’ of academic degree education and having higher number of credits than others. Most of the outliers are from education type ‘higher education’ and ‘secondary’. Civil marriage for academic degree is having most of the credits in the third quartile.

Target 1



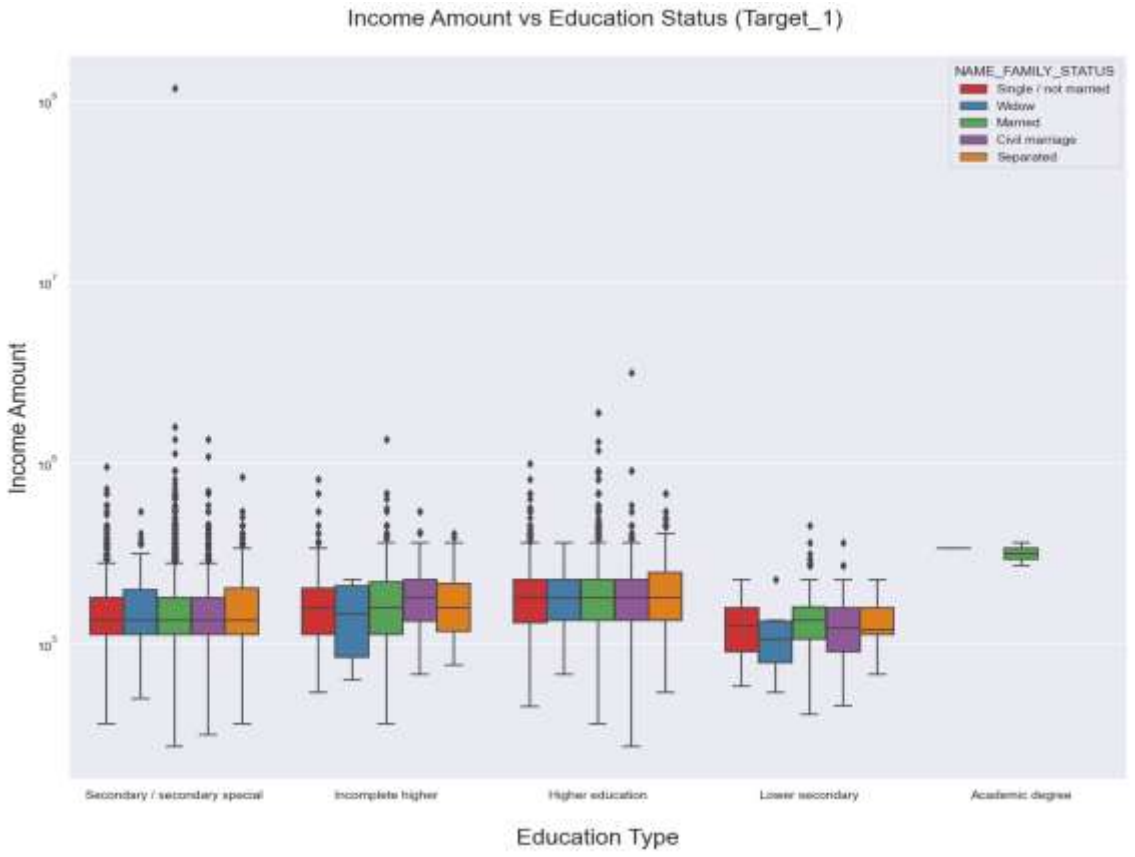


Target 0

From the above box plot for education type higher education the income amount is mostly equal with family status. It does contain many outliers. Less outliers having for academic degree but there income amount is little higher than higher education. Lower secondary of civil marriage family status I have less income amount than others.

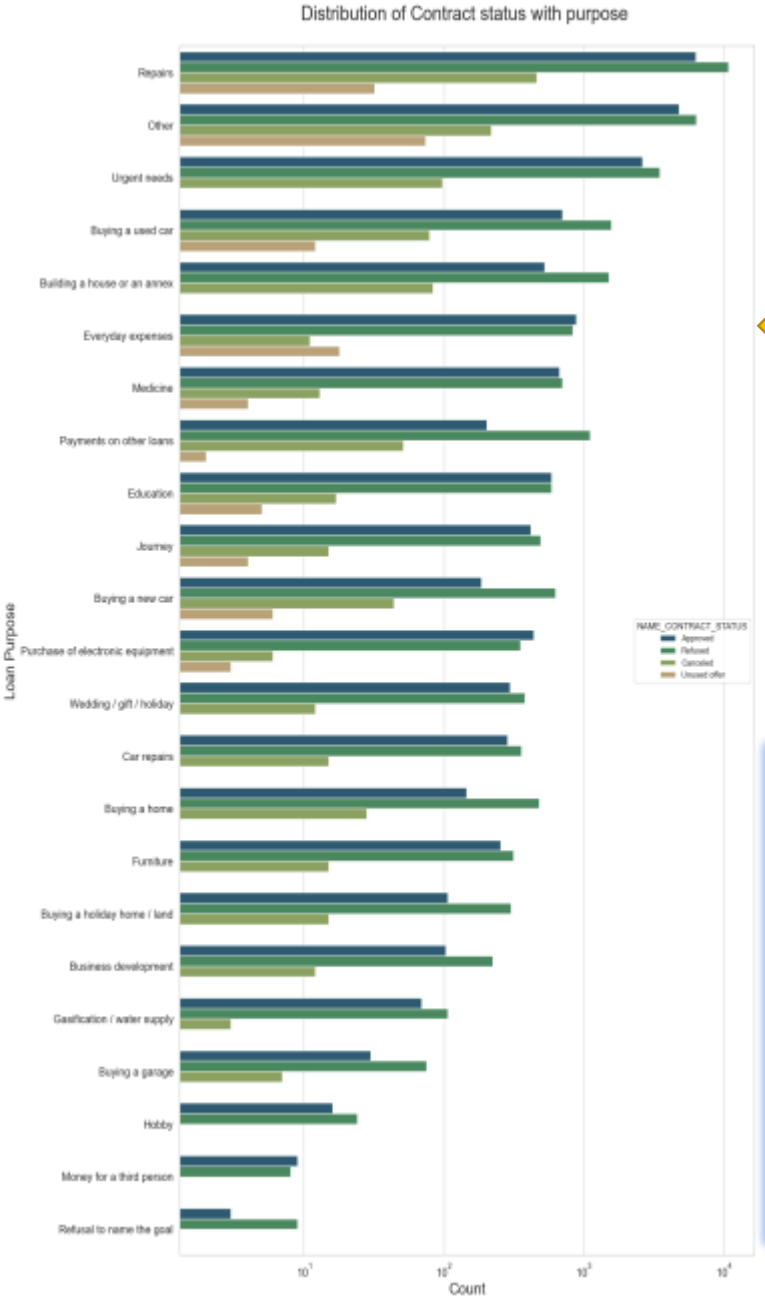
From bellow boxplot for education type ‘Higher Education’ the income amount is mostly equal with family status. It does contain many outliers. Less outlier are having for academic degree but their income amount is little higher . Lower secondary of civil marriage family status are have less income amount than others.

Target 1



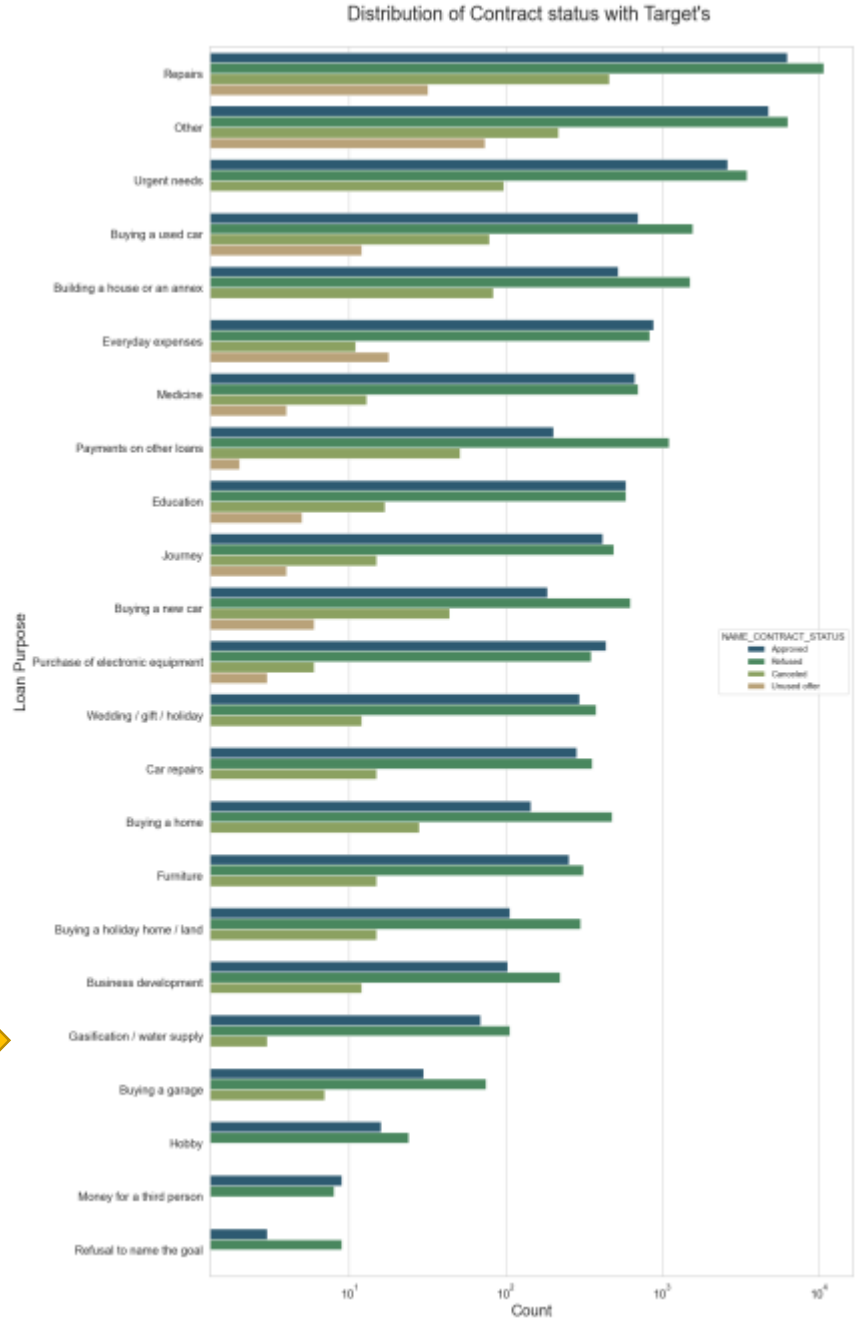
Merging Application DataSet and Previous DataSet

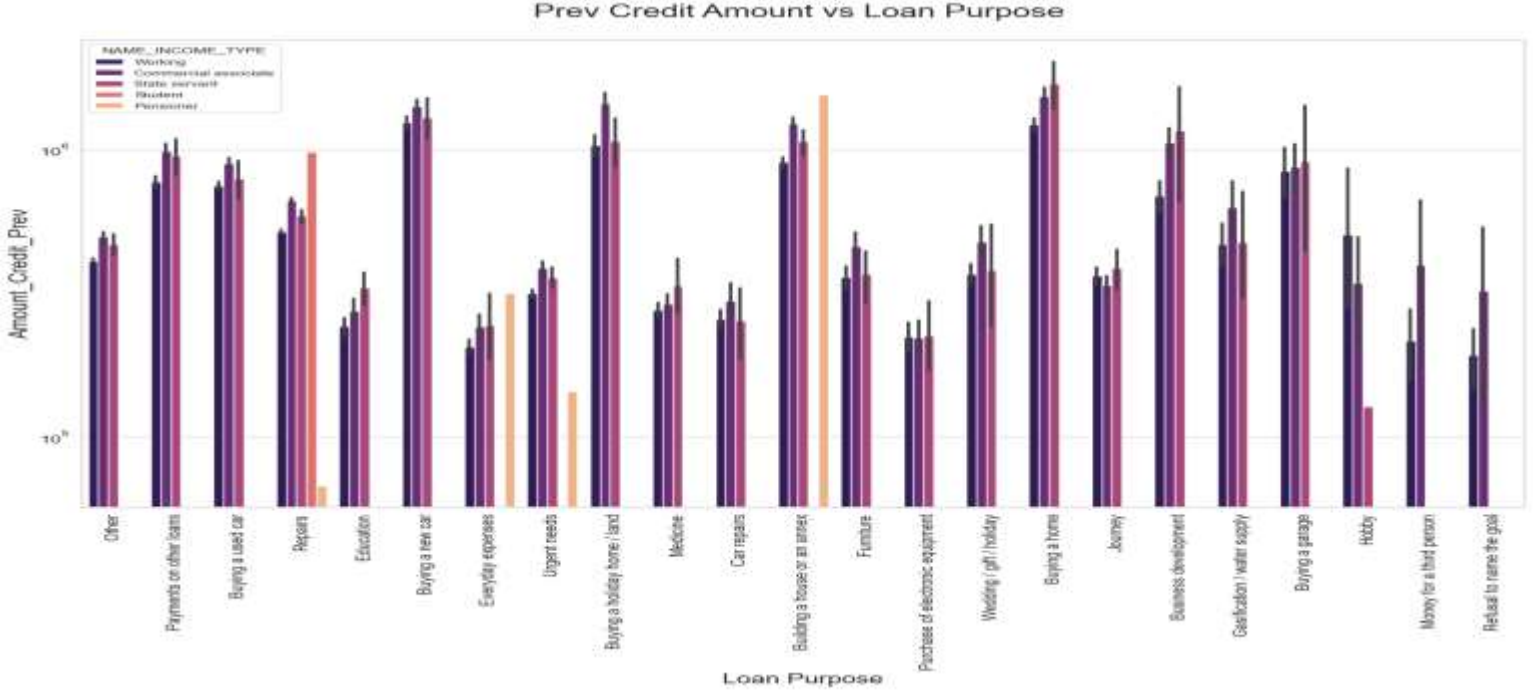
Final Steps



1. Most rejection of loans came from purpose ‘repairs’.
2. For education purposes we have equal number of approves and rejection.
3. Paying other loans and buying a new car is having significant higher rejection than approves.

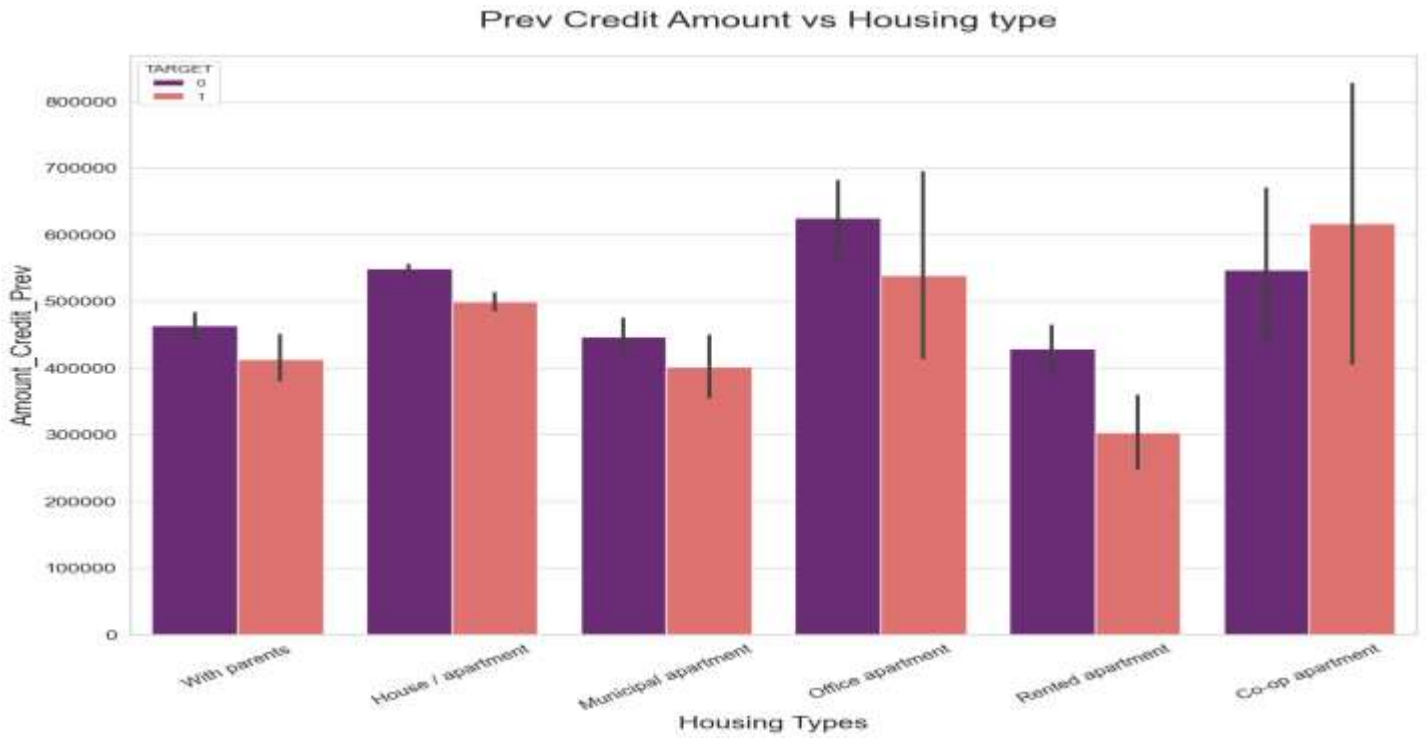
1. Loan purpose with ‘repairs’ are facing more difficulties in payments on time.
2. There are few places where loan payment is significant higher than facing difficulties. They are ‘Buying a garage’, ‘business development’, buying land’, ‘buying new car’ and ‘education’. Hence we can focus on these purposes foe which the client is having for minimal payment difficulties.





Here the Housing Type, office apartment is having higher credit of target_0 and co-op apartment is having higher credit of target_1. So we can conclude the bank should avoid giving loans to housing type of co-op apartment as they are having difficulties in payment. Bank can focus mostly on housing type with parents or house/apartment for successful payments.

1. The credit amount of loan purpose like ‘Buying a land’, ‘buying a car’ and ‘building a house’ is higher.
2. Income types of state servants have a significant amount of credit applied.
3. Money for third person or a a hobby is having less credits applied for.



Conclusion of this Loan Analysis

- Bank should approve loans more for Office Apartment, Co-Op Apartment housing type as there are less payment difficulties.
- Bank should provide loans to ‘Repairs’ and ‘Others’ purposes.
- Bank should provide launched to the ‘Business Entity type 3’ and ‘Self Employed’ persons.
- Working people especially female employees are the best target for the loans.