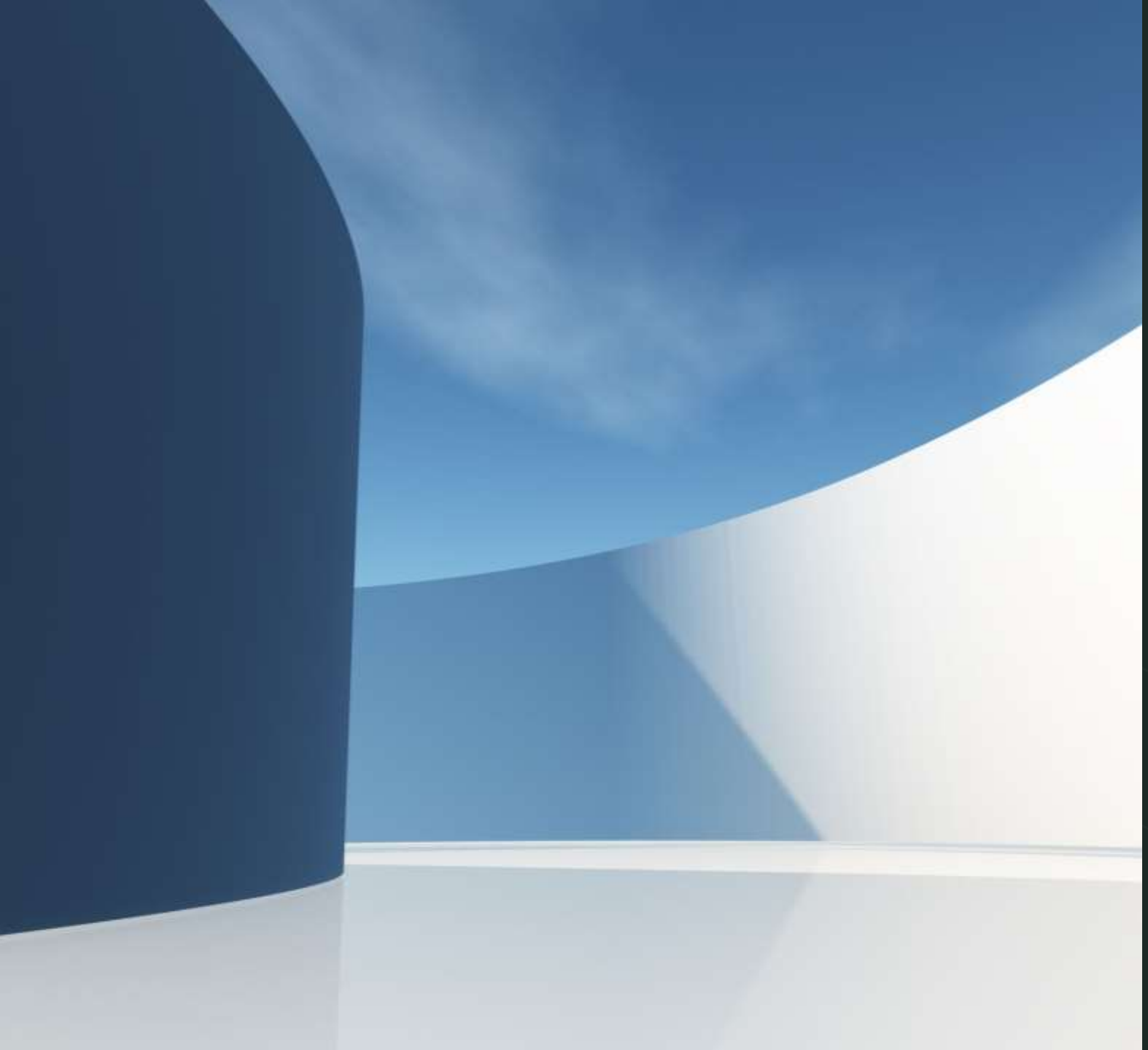


BANK CREDIT EDA – A CASE STUDY

---- by

Arnav Sinha

Introduction

- When a bank reviews a loan application, it must determine whether to approve the loan based on the applicant's profile. There are two main risks involved in this decision:
- If the applicant is expected to repay the loan, rejecting the application could mean losing out on potential business for the bank.
- If the applicant is unlikely to repay the loan and is expected to default, approving the loan might result in financial losses for the bank.

Key factors of the data

The data given here contains the information about the loan application at the time of applying for the loan. It contains two scenarios:

- The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
- All other cases: All other cases when the payment is paid on time.

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

- Approved: The bank 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 bank 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.

Objectives

- This case study seeks to pinpoint patterns that suggest a client may struggle with paying their installments.
- These insights can inform decisions such as denying the loan, adjusting the loan amount, or charging higher interest rates for riskier applicants.
- The bank aims to uncover the key factors behind loan defaults—those variables that most strongly predict default. By understanding these factors, the bank can better manage its portfolio and assess risk.

Approach

For Application Data:

- Missing value treatment
- Data Imbalance analysis – Imbalance of data in terms of Target customers (Imbalance of the number of customers with payment difficulties with those without difficulties)
- Segregation of the dataset into two – One with Target = 0, another with Target = 1
- Finding outliers for the two datasets
- Univariate analysis for the two datasets separately
- Multivariate analysis for the datasets
- Bivariate analysis for separate datasets

Approach

For Previous Application Data:

- Missing value treatment
- Finding outliers
- Univariate analysis
- Bivariate analysis
- Multivariate Analysis

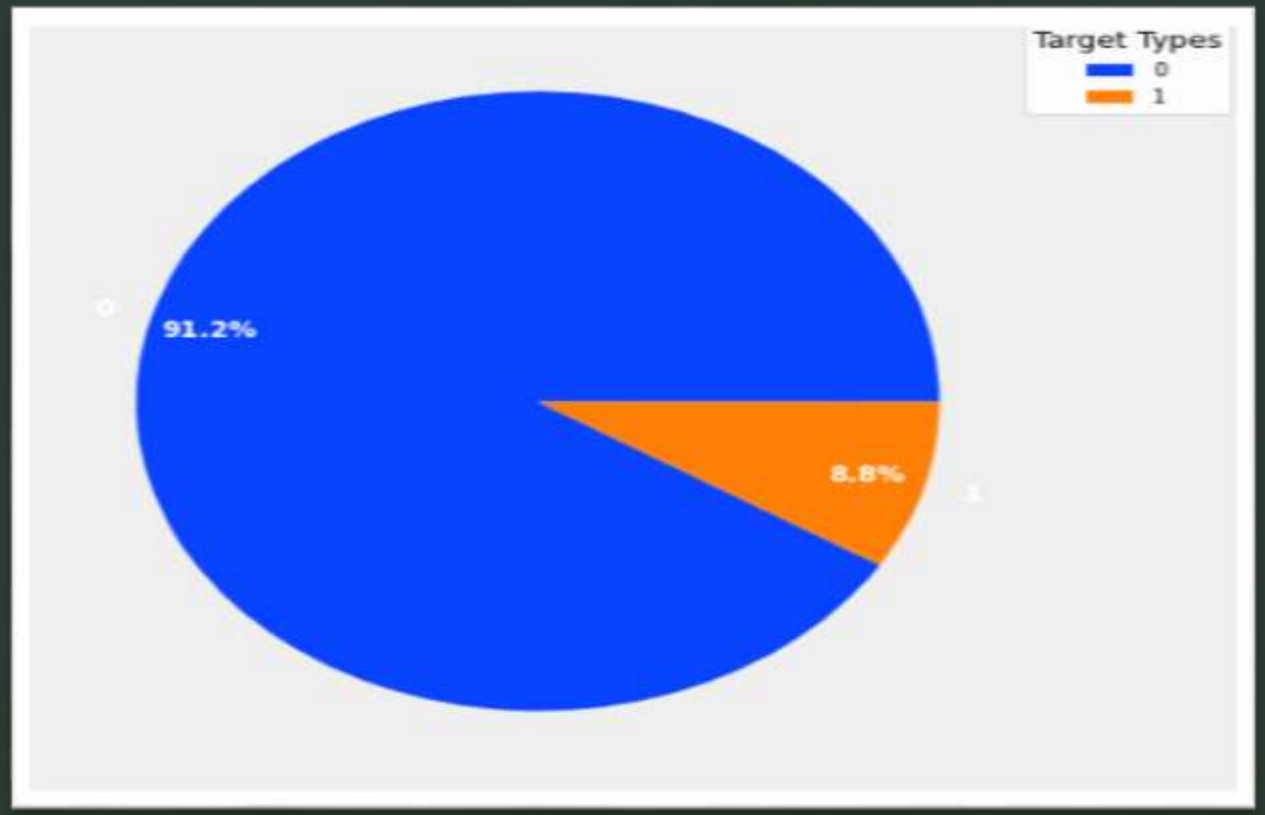
For Merged datasets (application data and previous data):

- Contract Status analysis for applicants
- Univariate analysis
- Bivariate analysis
- Multivariate Analysis
- Conclusion

APPLICATION DATASET

Data Imbalance

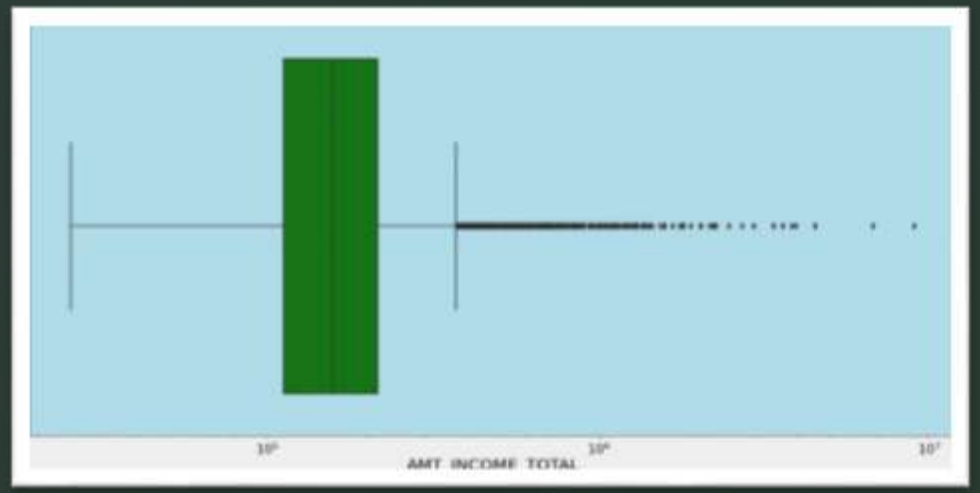
Target Segregation Imbalance



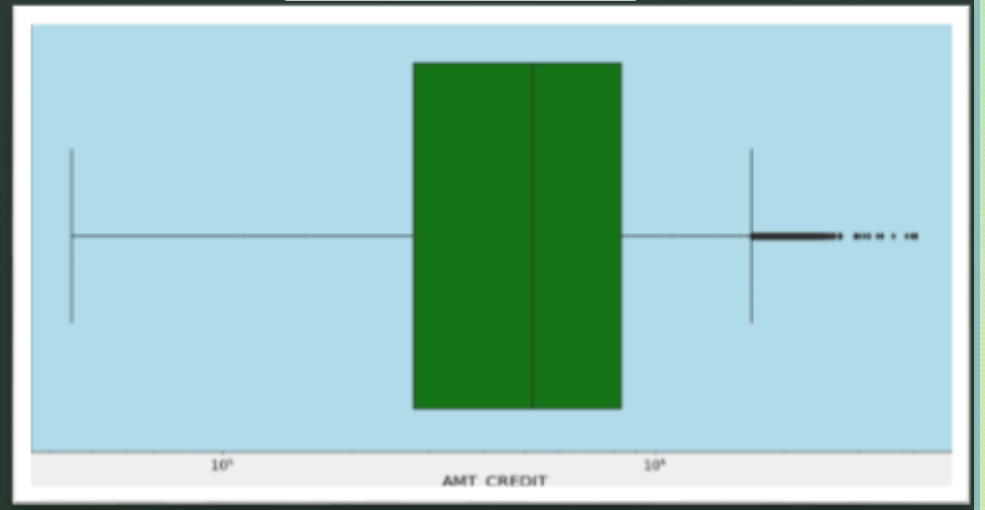
- This shows that the customers with payment difficulties are less than those without payment difficulties.
- The imbalance ratio is 10.35

Outliers for Target = 0 (Boxplot Method)

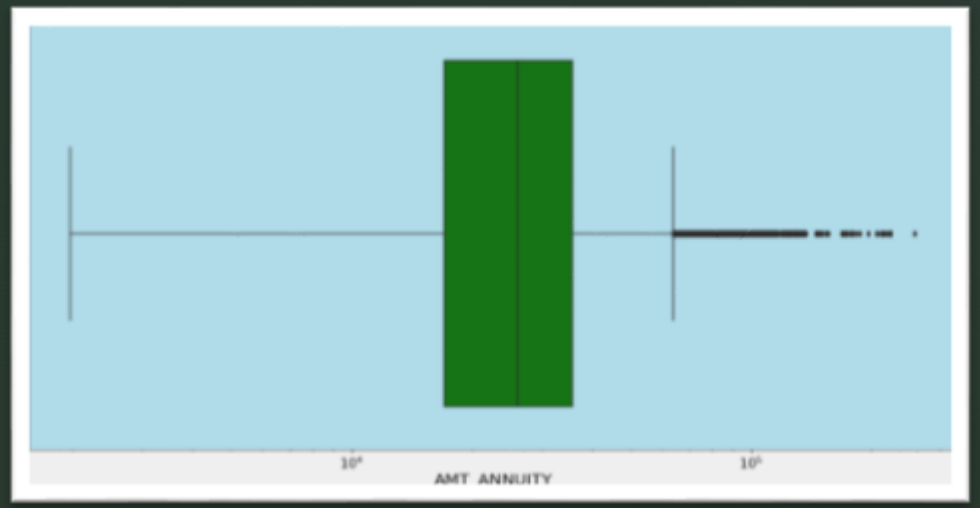
Income Amount



Amount Credit

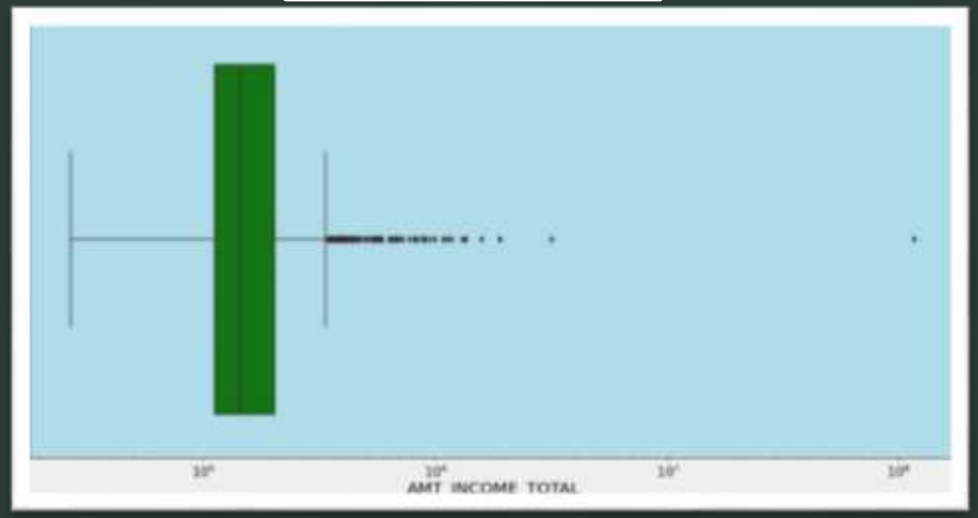


Amount Annuity

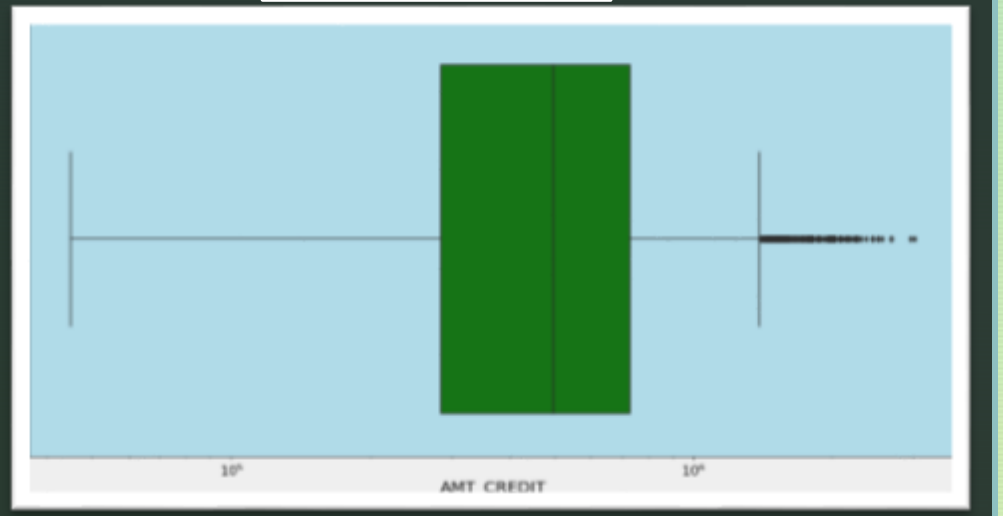


Outliers for Target = 1 (Boxplot Method)

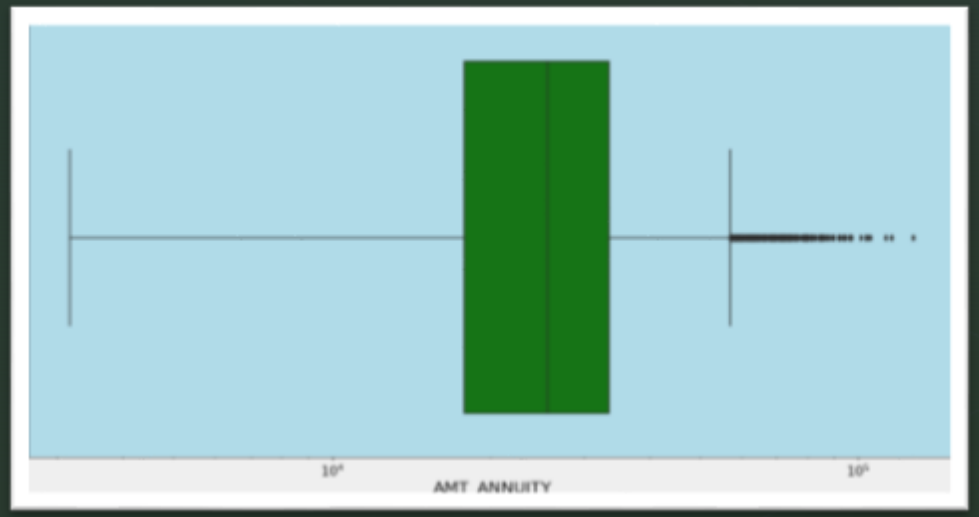
Income Amount



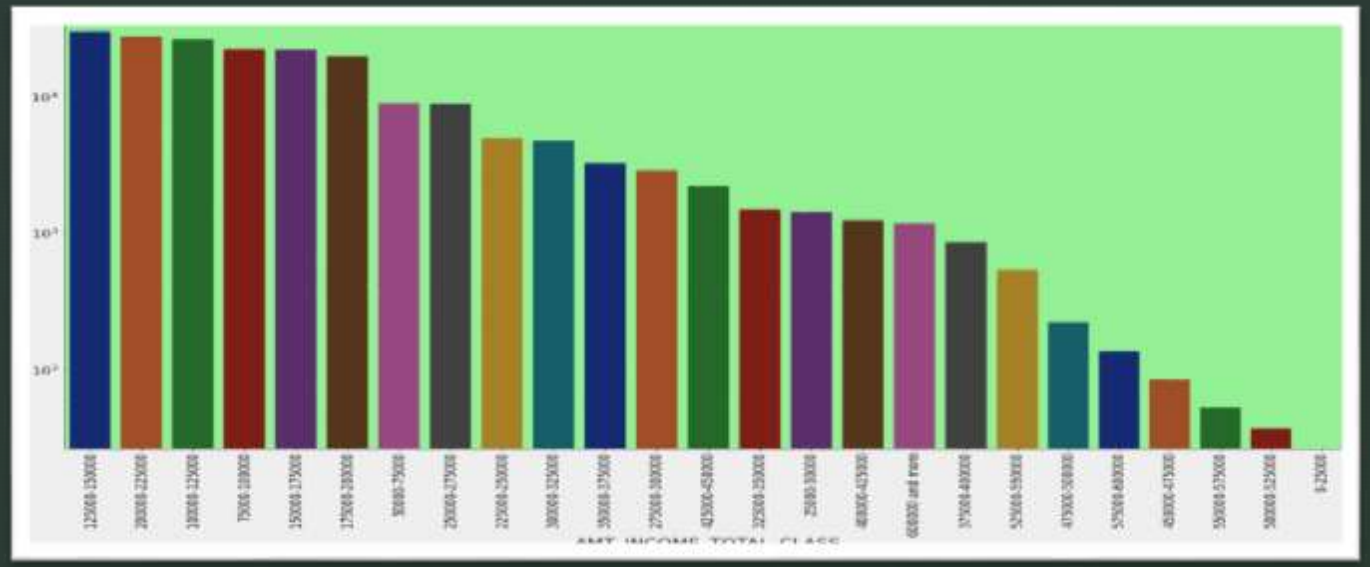
Amount Credit



Amount Annuity

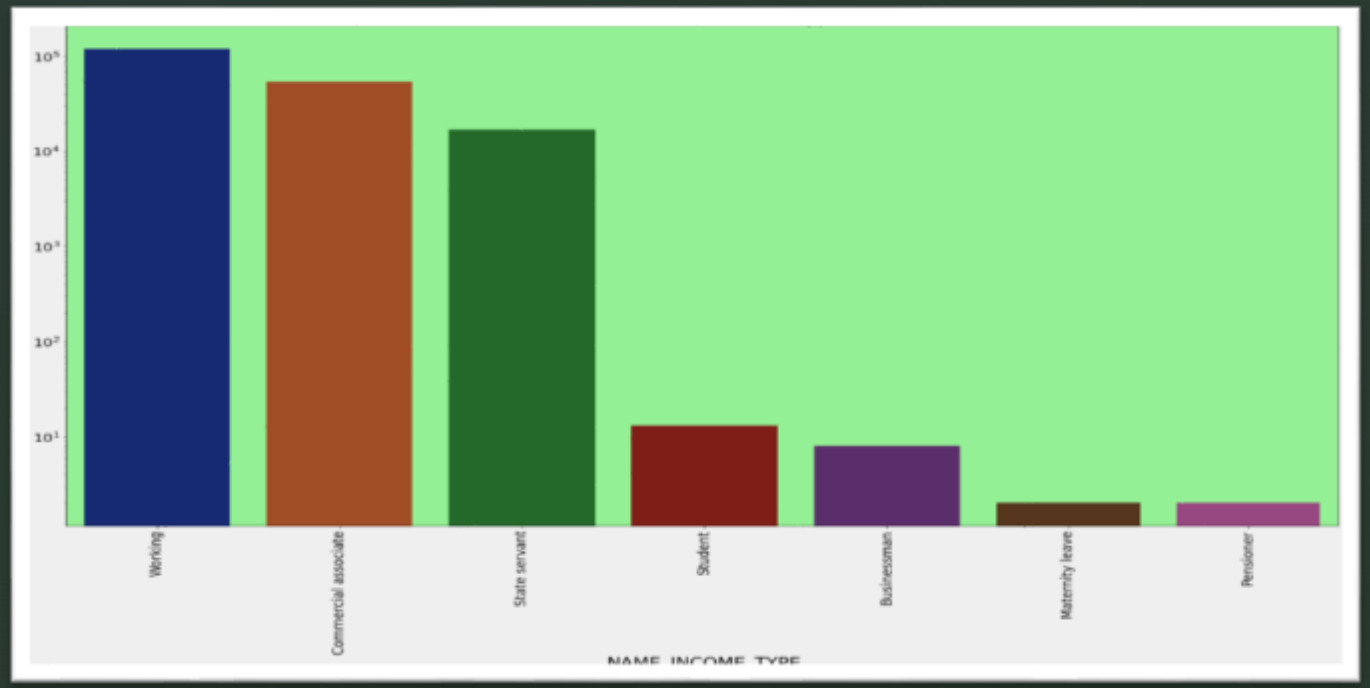


Univariate analysis for Target = 0



Income class

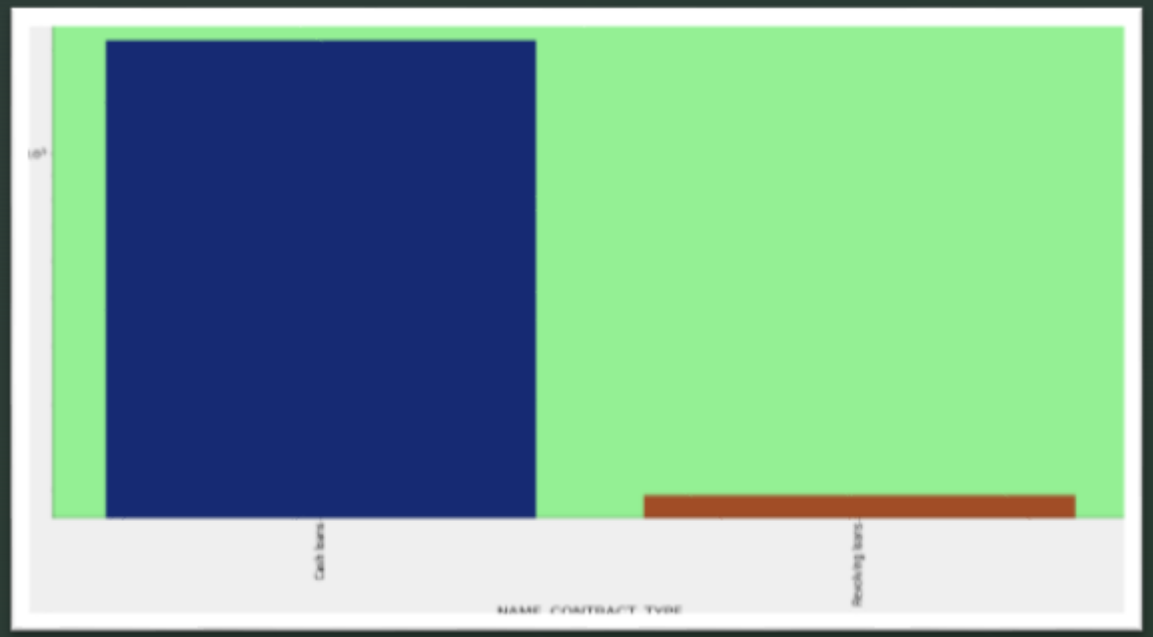
- Income class from 75K to 200K customers are tending to repay the loan more frequently than others.
- Income class of 3750K and above have very less chance of loan repayment on time.



Income type

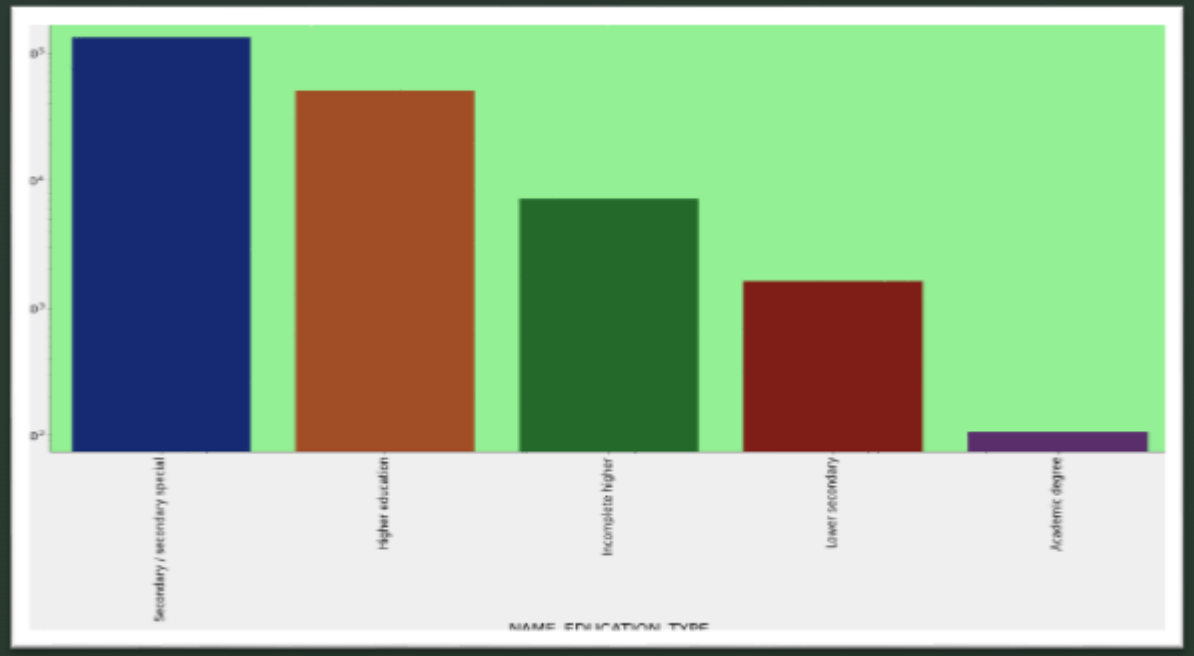
- For 'Working', 'Commercial associate' and 'State servant' income types, customers are mostly reliable to repay loan on time.

Univariate analysis for Target = 0



Contract type

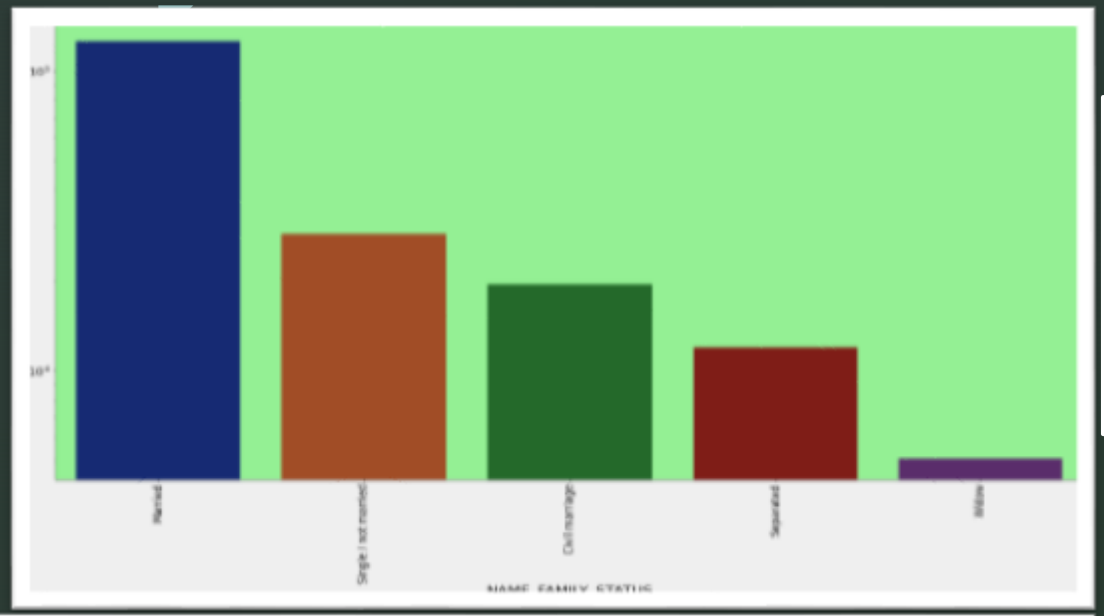
- Cash loans accounts for majority of customers than revolving loans.



Education type

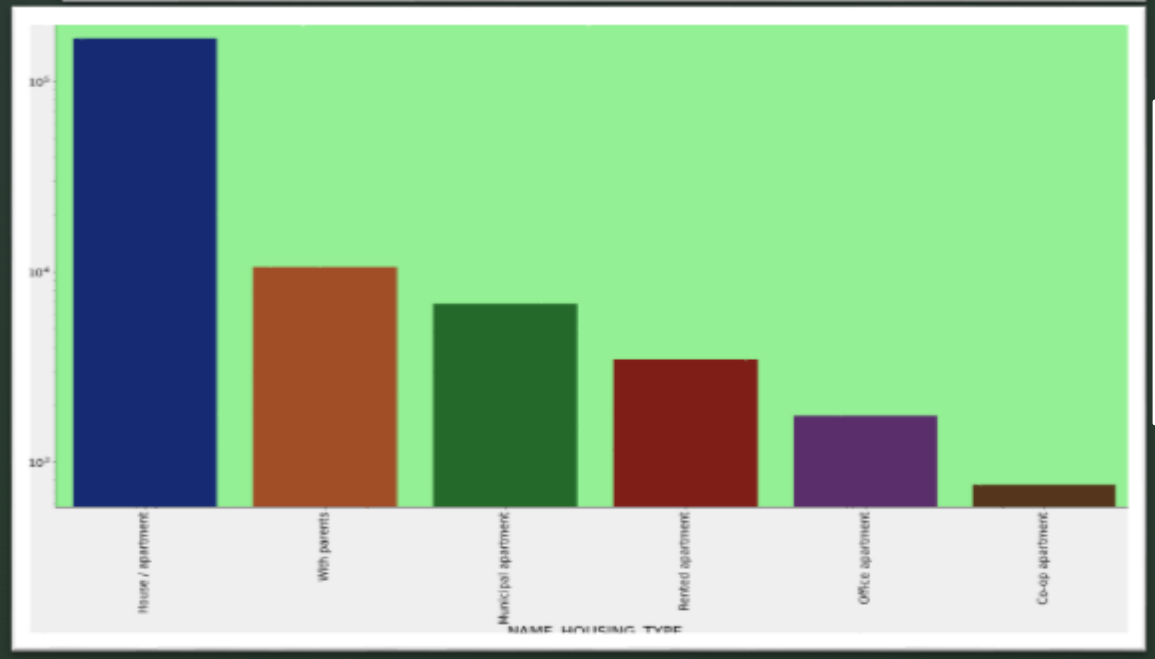
- Customers mostly whose education types are secondary/special and higher education tend to repay loan on time.

Univariate analysis for Target = 0



Family status

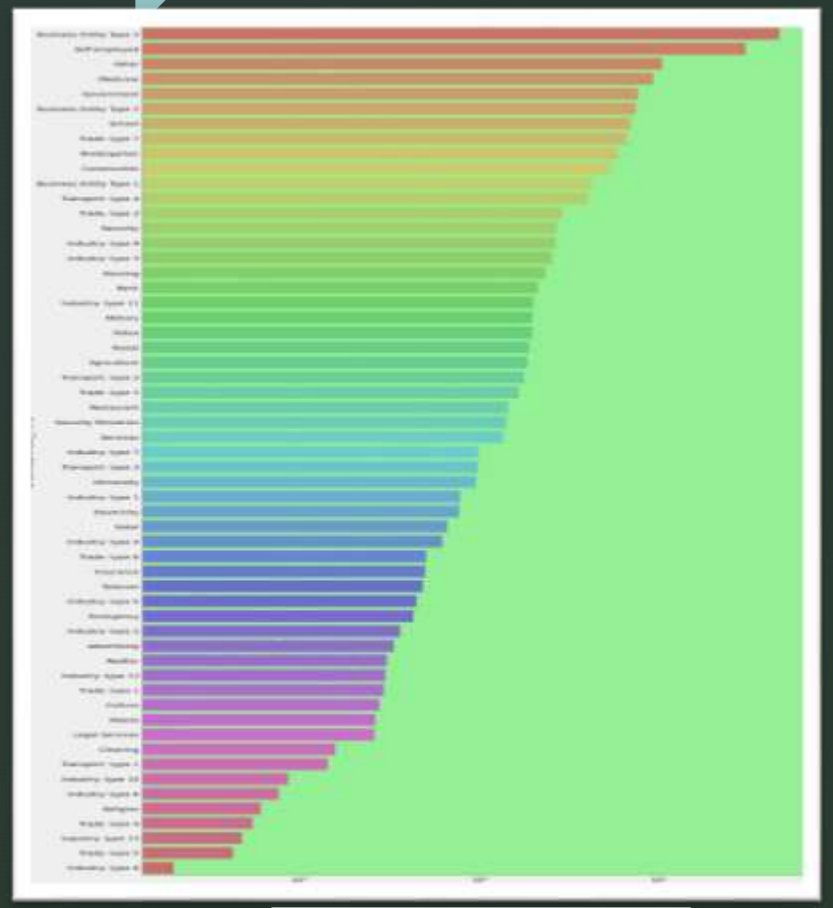
- Mostly married customers tend to repay loan on time.



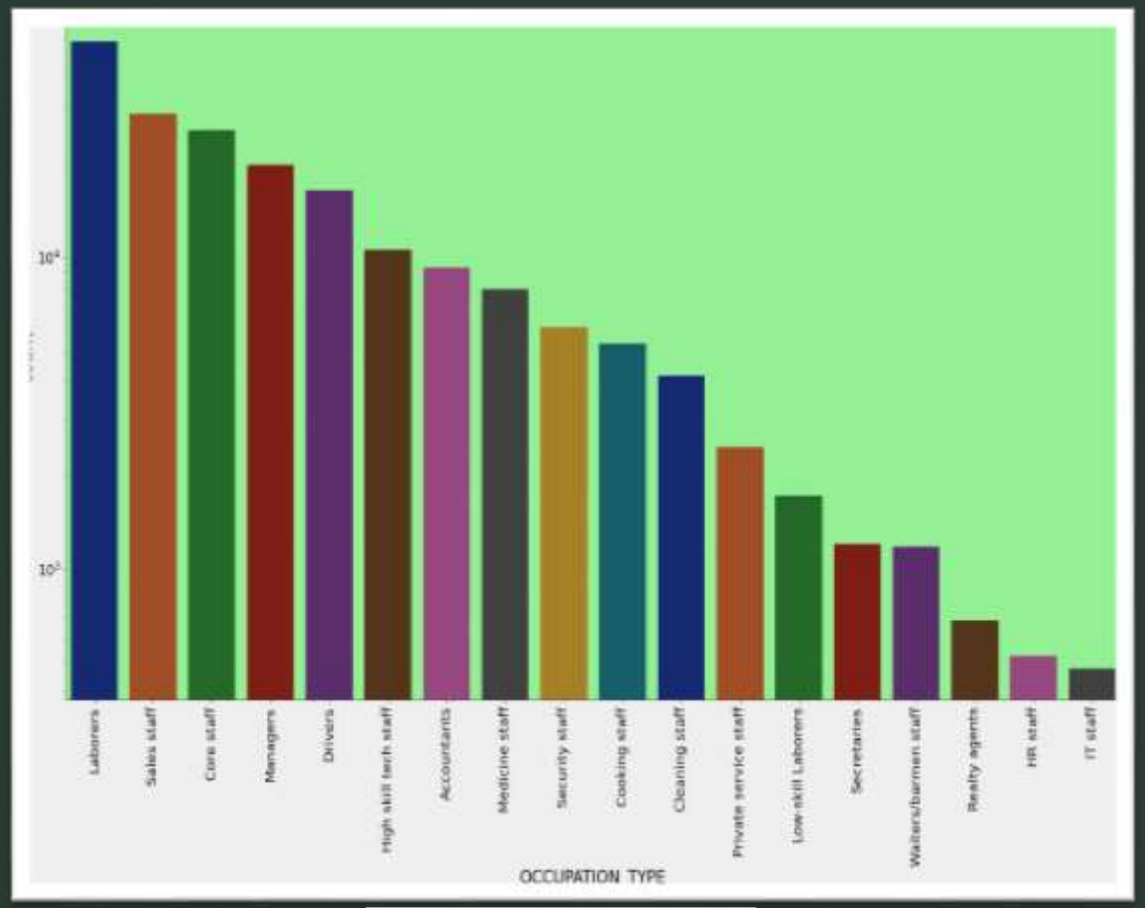
Housing type

- Customers who lives in house/apartment tend to have no payment difficulty.
- Customers who lives in co-op apartment or office sponsored apartment have a little payment difficulty.

Univariate analysis for Target = 0



Organization type



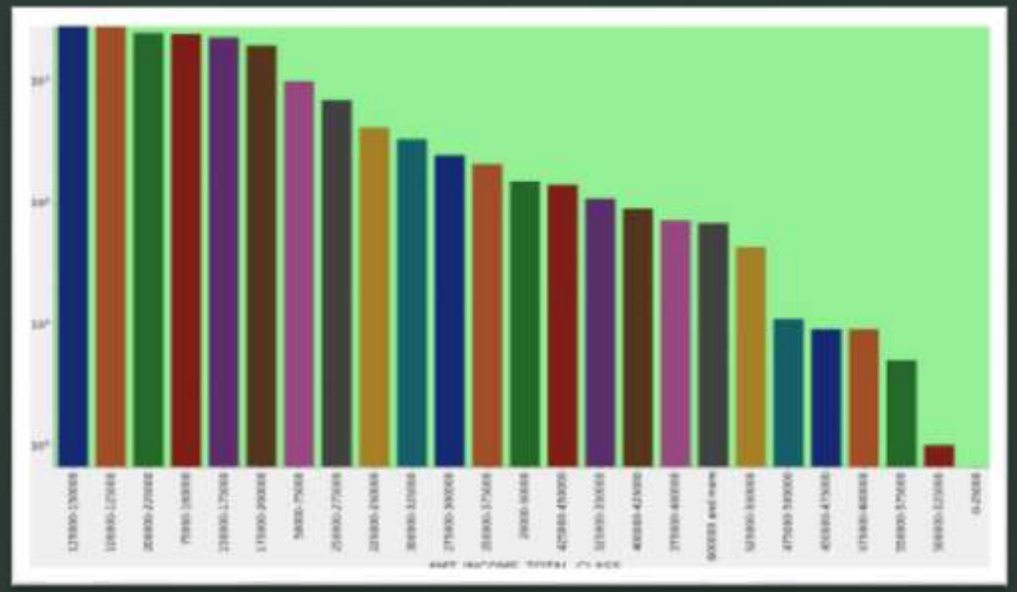
Occupation type

- Customers of the organization type ‘Business entity Type 3’ , ‘Self employed’, ‘Other’ , ‘Medicine’ and ‘Government’ tend to have very less to no difficulty in repayment.

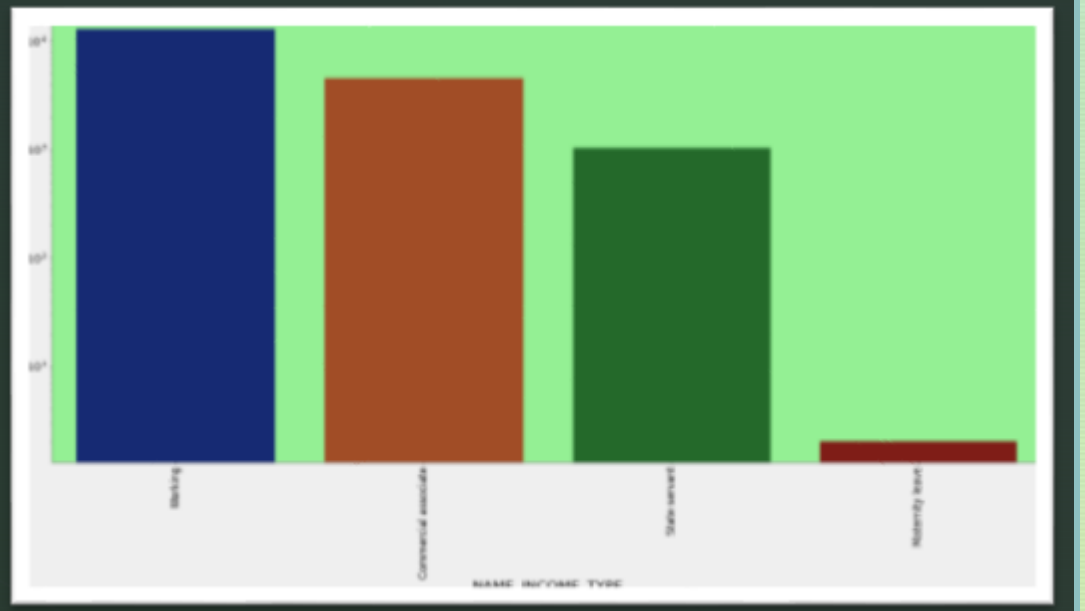
- Customers who are laborers have no difficulty in loan repayment.

Univariate analysis for Target = 1

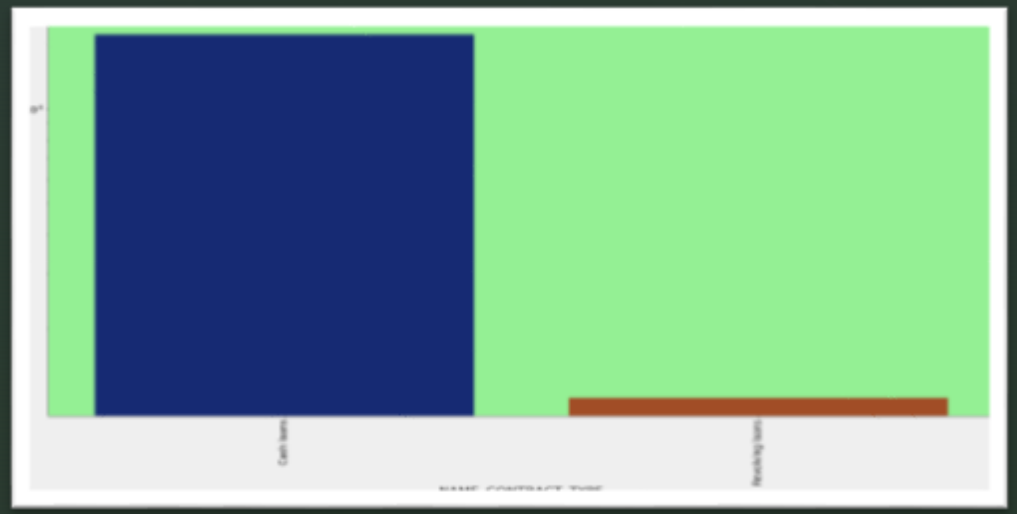
- Income class



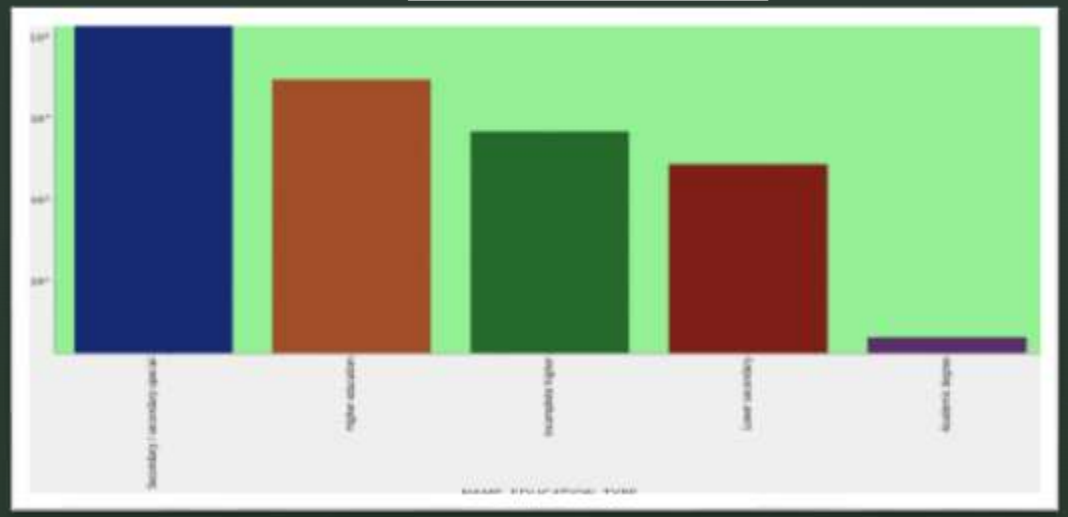
Income type



Contract type

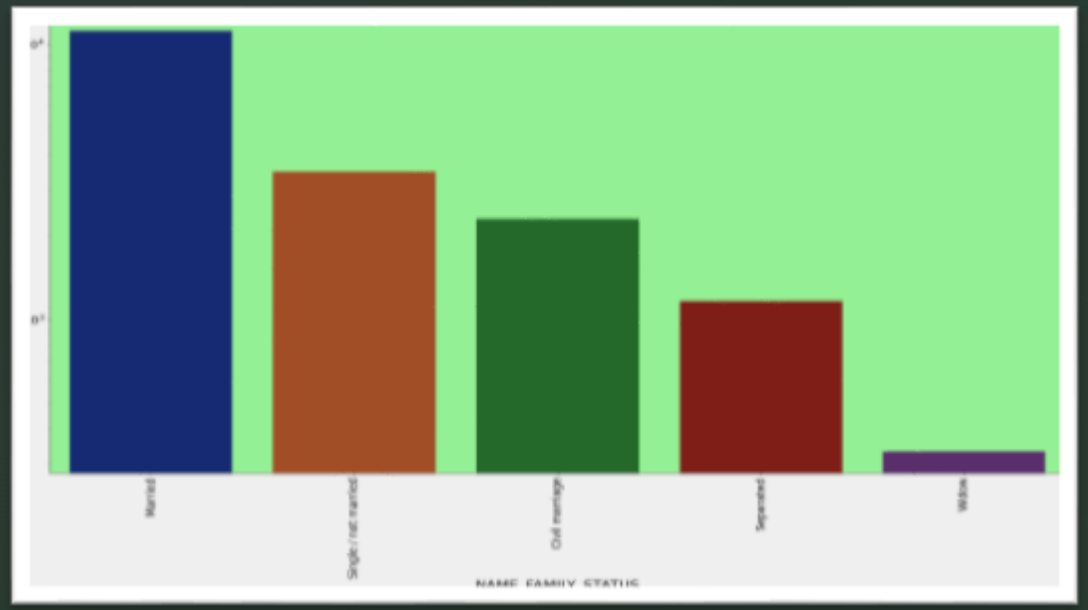


Education type

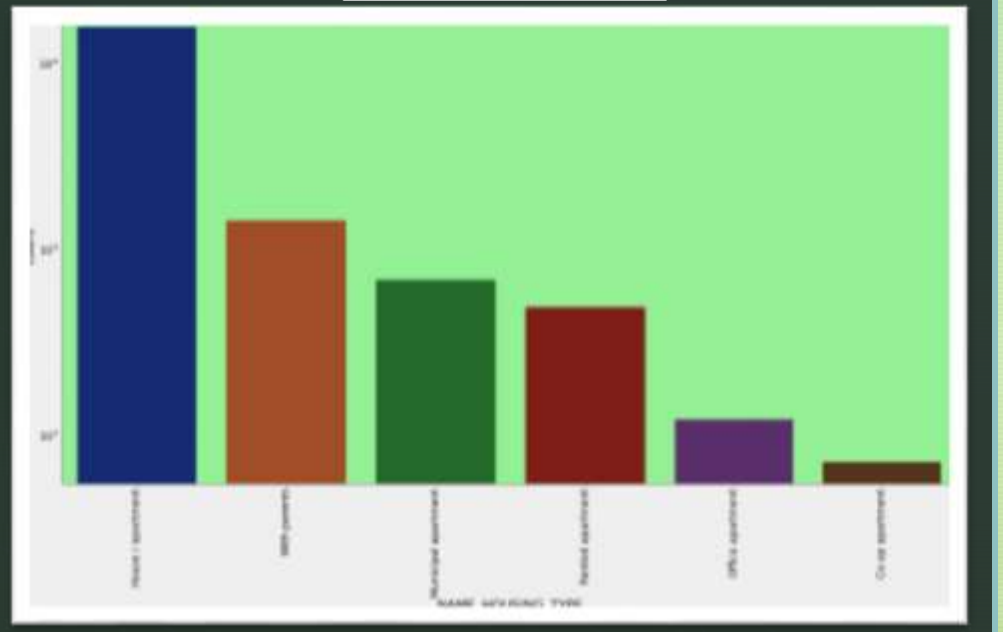


Univariate analysis for Target = 1

Family status

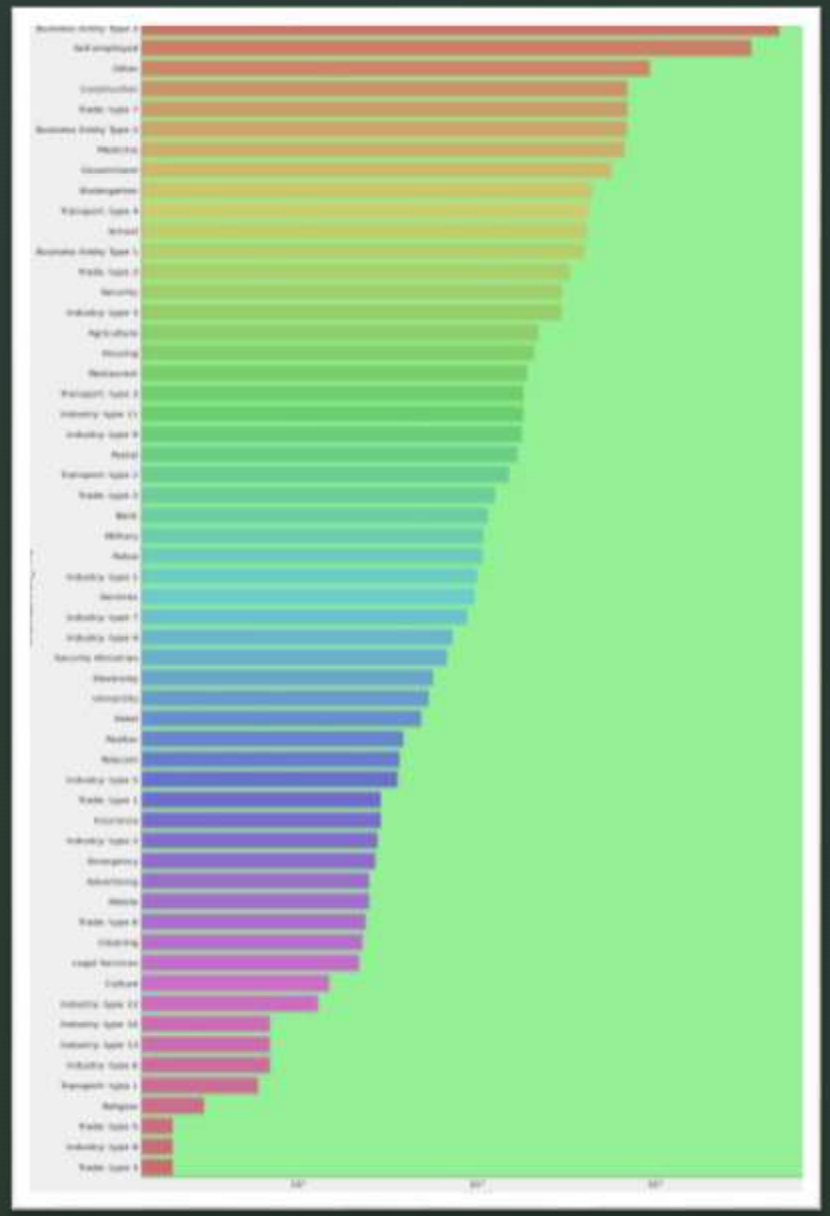


Housing type

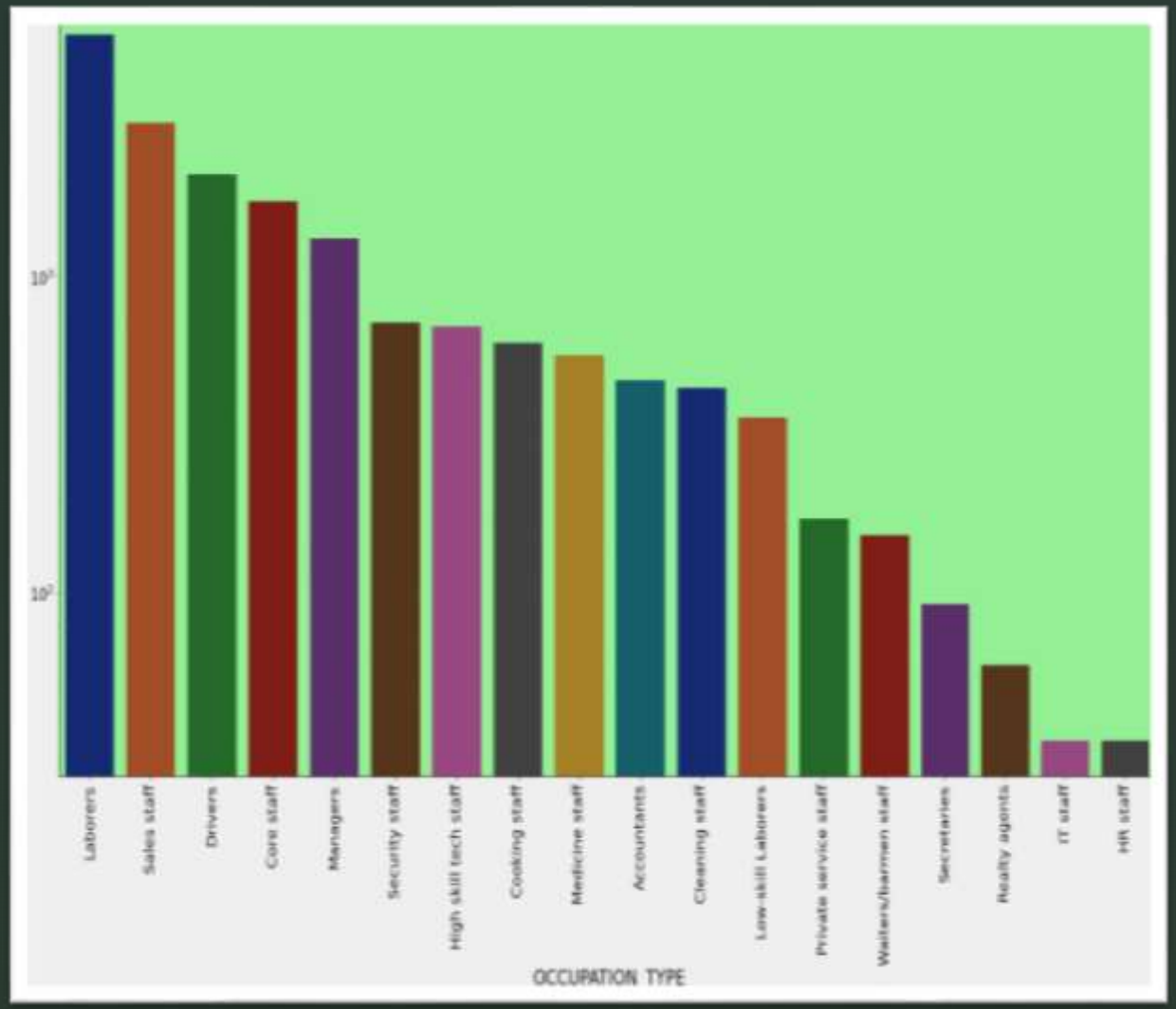


Univariate analysis for Target = 1

Organization type

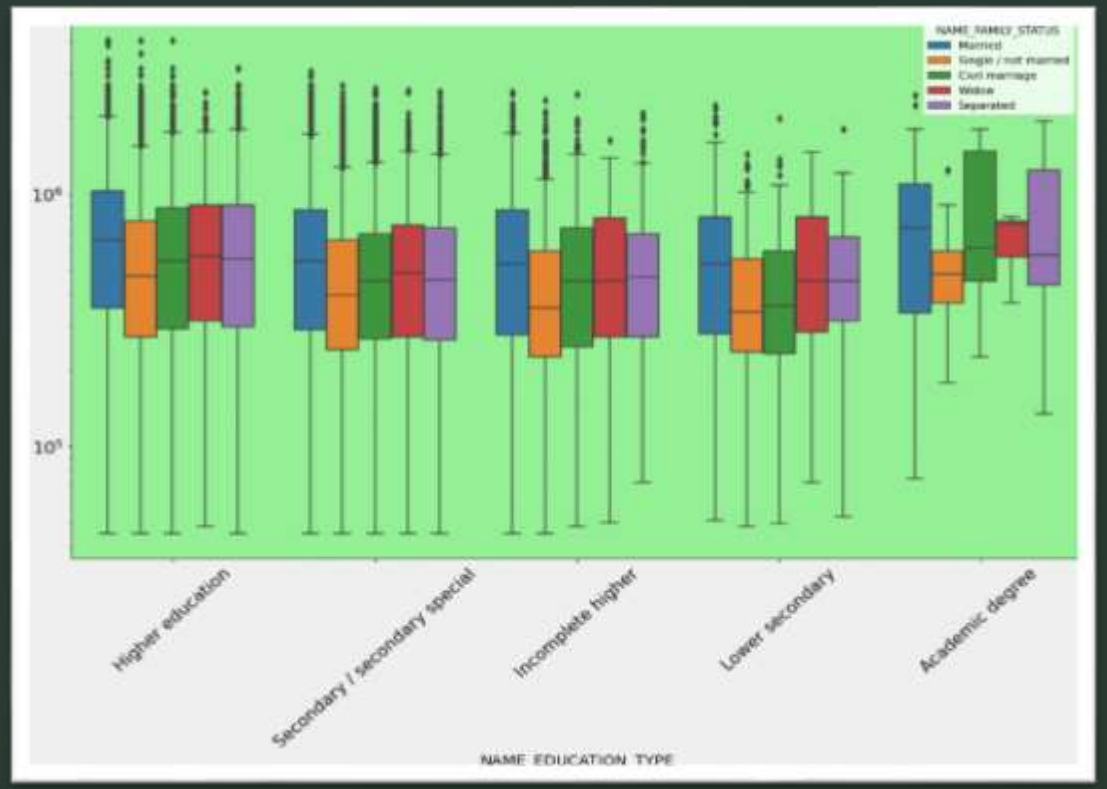


Occupation type



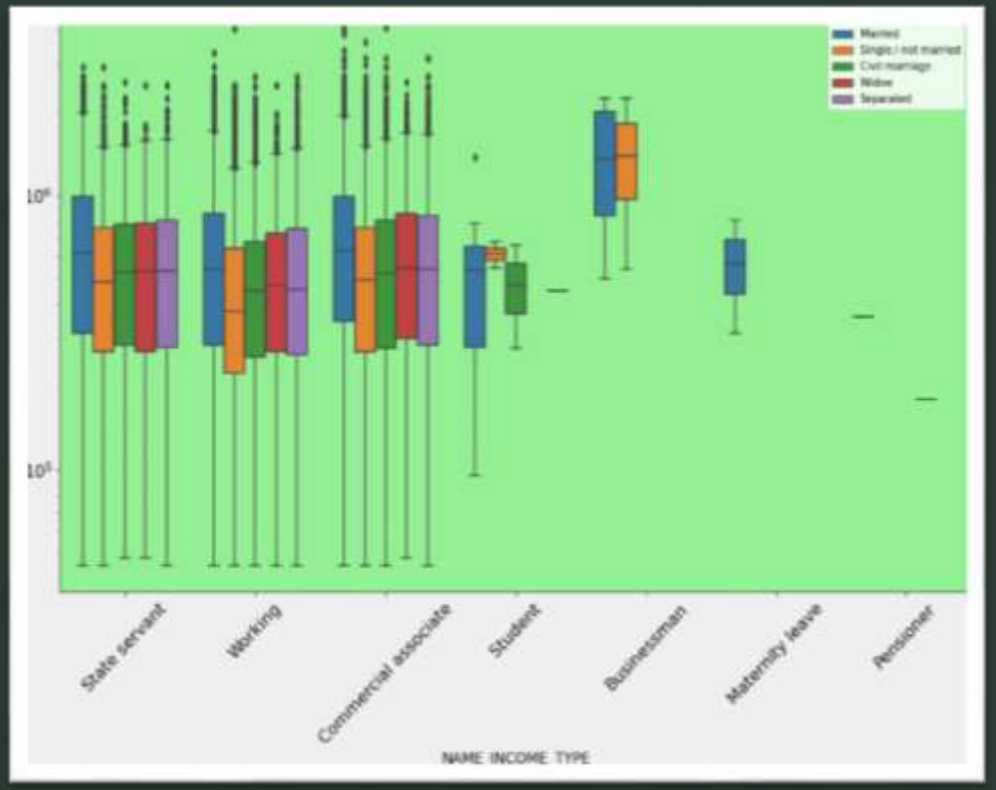
Bivariate analysis for Target = 0

Credit amount vs Education Status



- The customers who are married, single and civil marriage for Higher Education have more outliers in the dataset.
- For Academic degree holding customers, who are married, civil marriage and separated, they have higher amount of credit.
- For all the education types, married customers have more credit amount lying between 1st and 2nd quartiles.
- For Academic degree holders, who are civil marriage and separated, they have more credit amount between 2nd and 3rd quartiles.

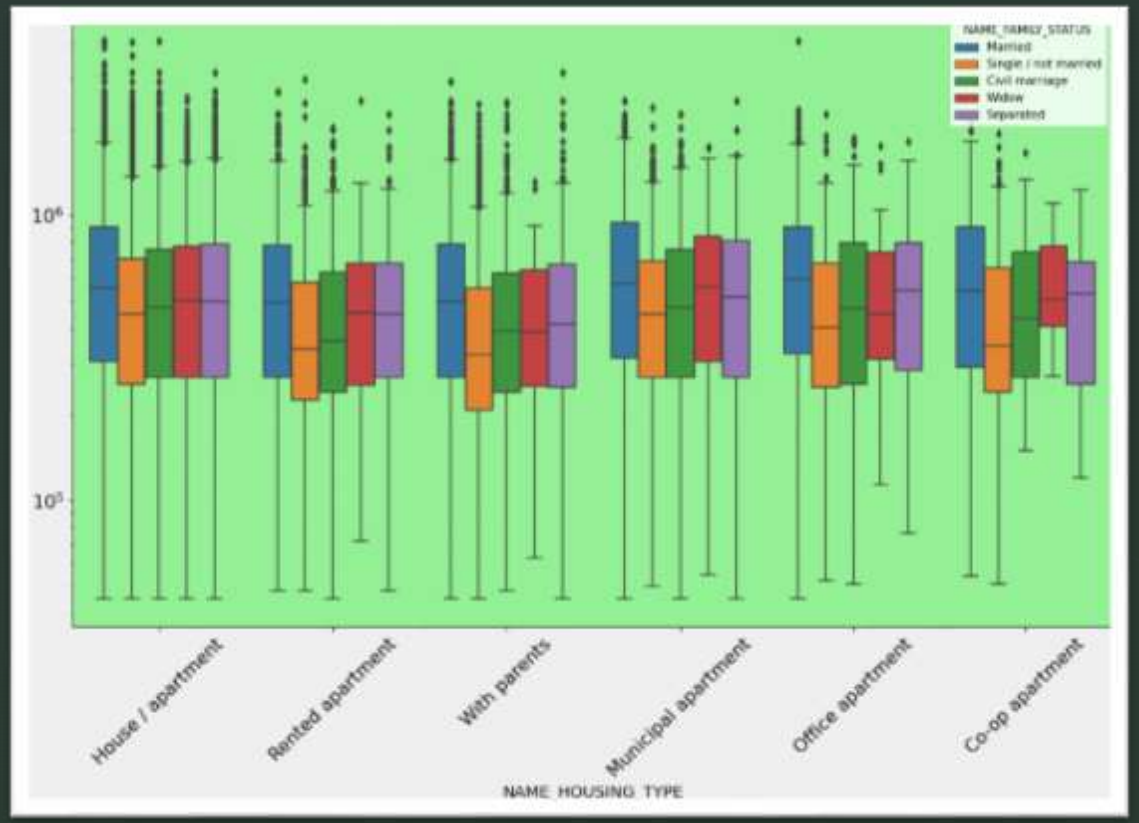
Credit amount vs Income Type



- For State servant, working and commercial associate customers, married ones are having more credit amount.
- For student, pensioner, businessman and maternity leave customers, the minimum credit amount is much higher and is having no outliers.
- For student, businessman and maternity leave customers, maximum credit amount lies between 1st and 2nd quarter.

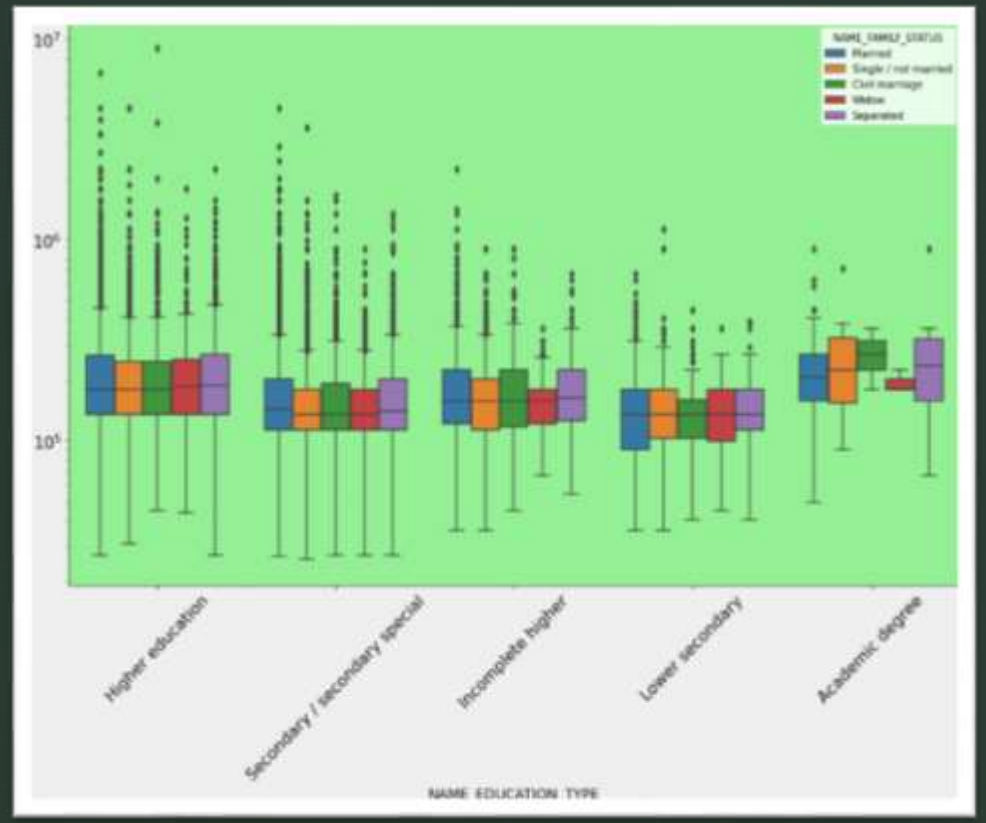
Bivariate analysis for Target = 0

Credit amount vs Housing type



- For all the housing types, married customers are having more credits than others.
- Customers who live in house/apartment have more outliers than others.

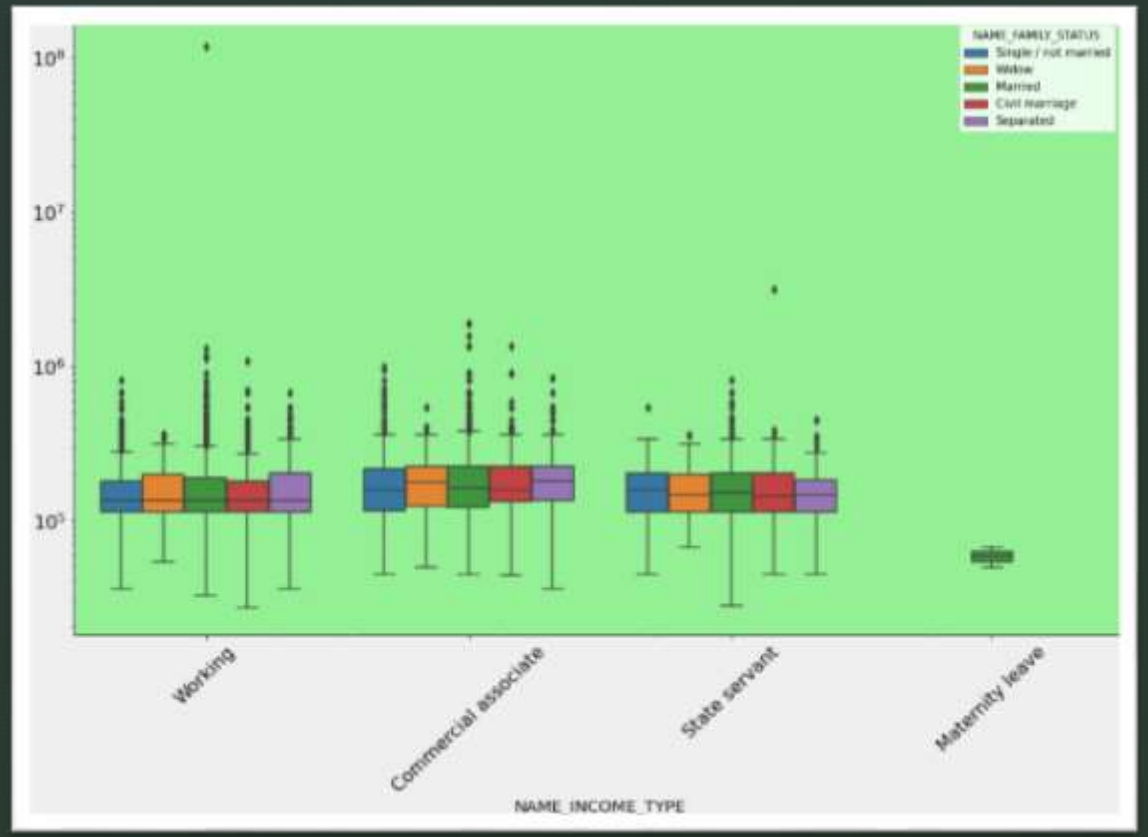
Income amount vs Education Status



- Customers having higher education and secondary/special education have more outliers than others.
- Customers who are separated and having academic degree have more income amount lying between 1st and 2nd quartile.

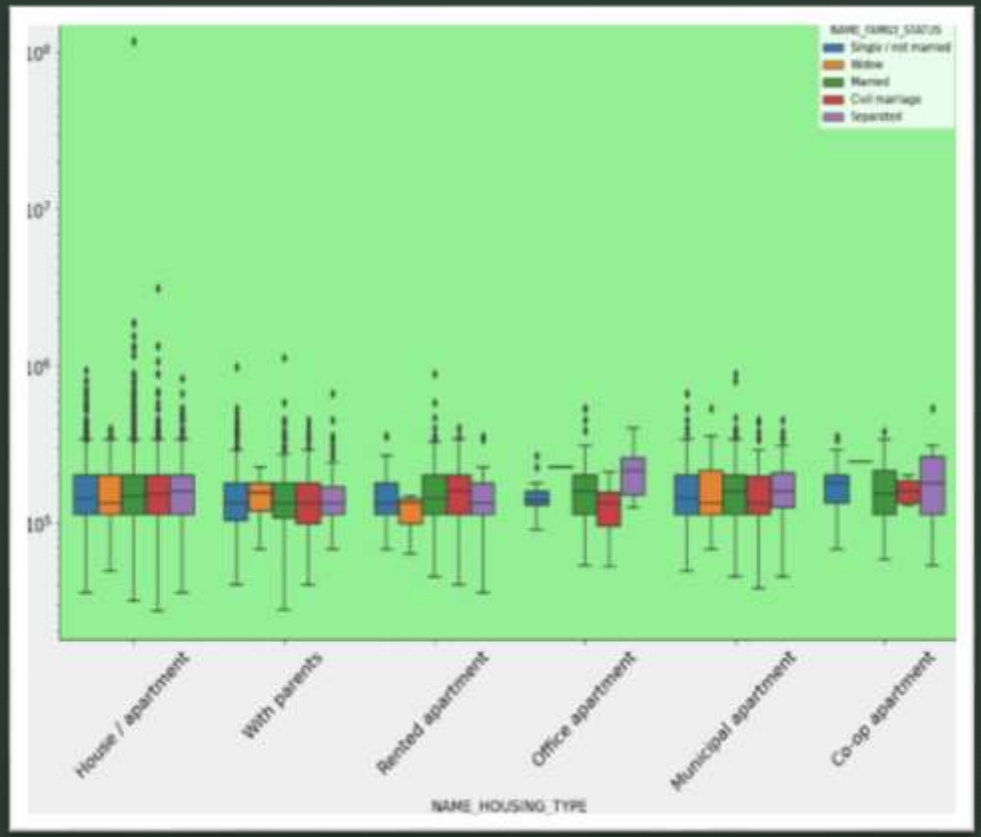
Bivariate analysis for Target = 1

Income amount vs Income type



- Customers who are working, commercial associate and state servant have more outliers in income.

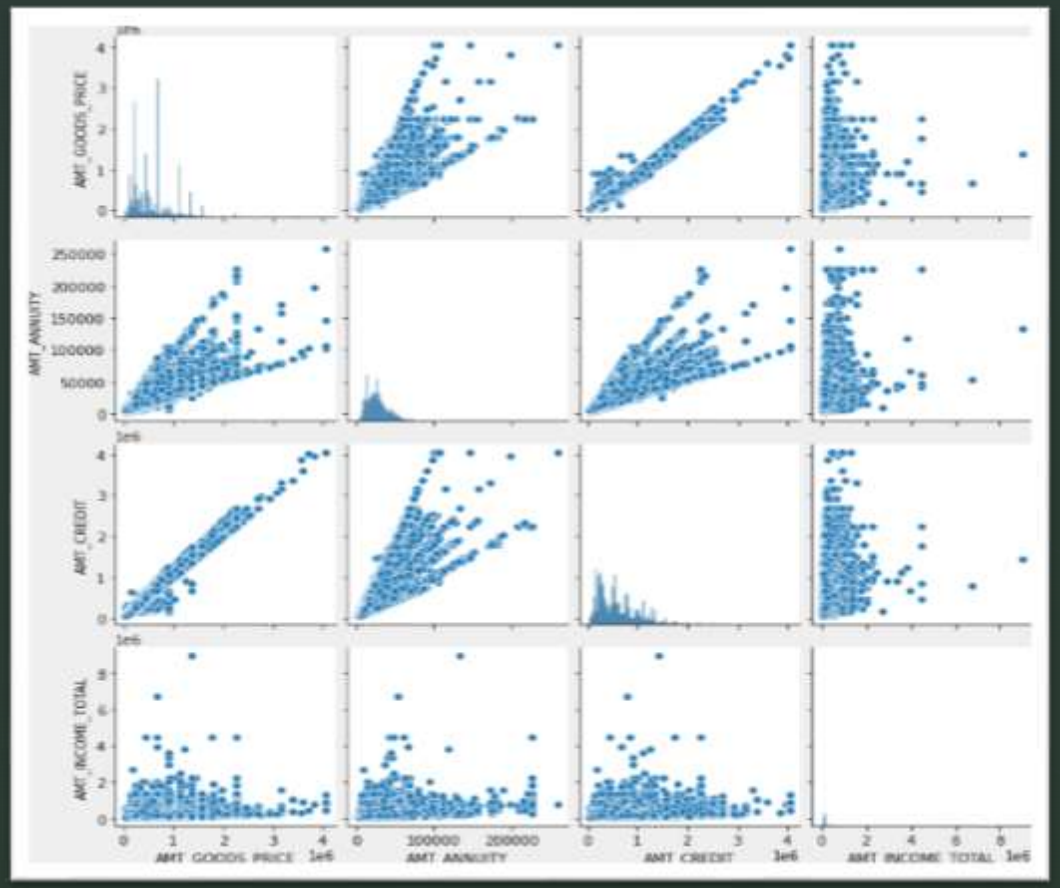
Income amount vs Housing type



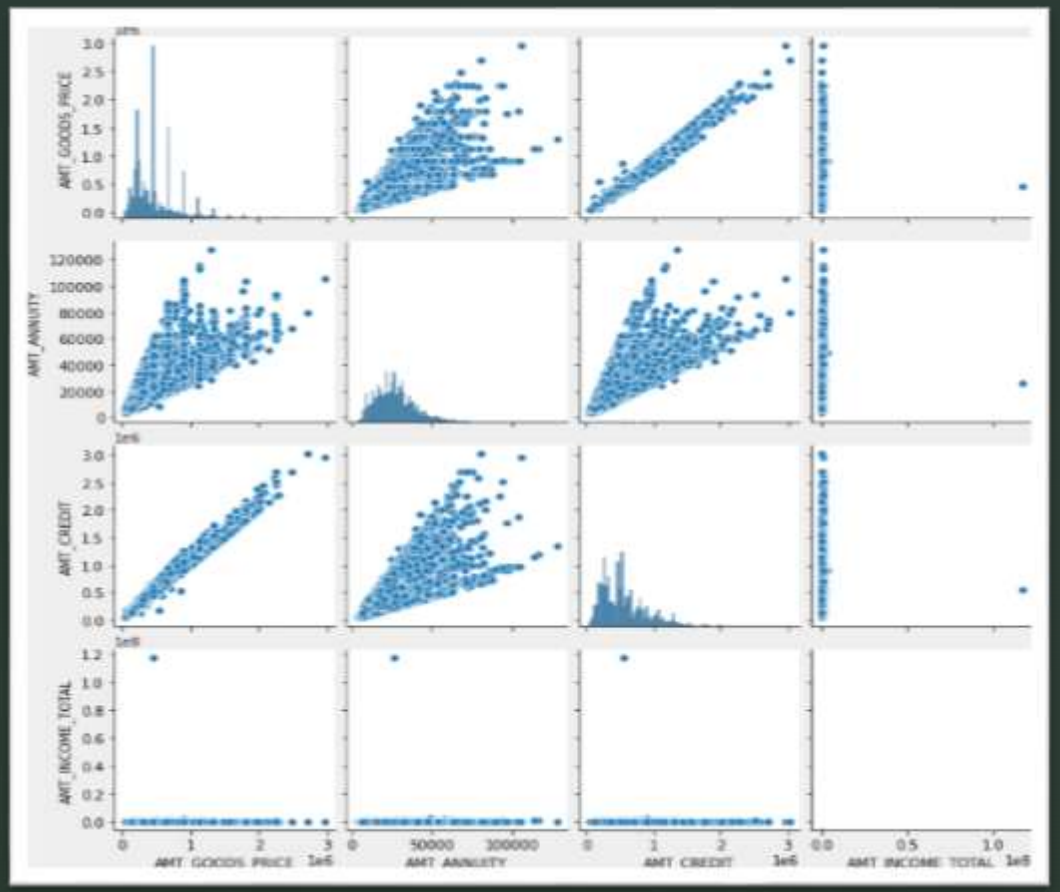
- Customers who live in house/apartment have more outliers in income.

Multivariate analysis

Pair plot for target = 0



Pair plot for target = 1

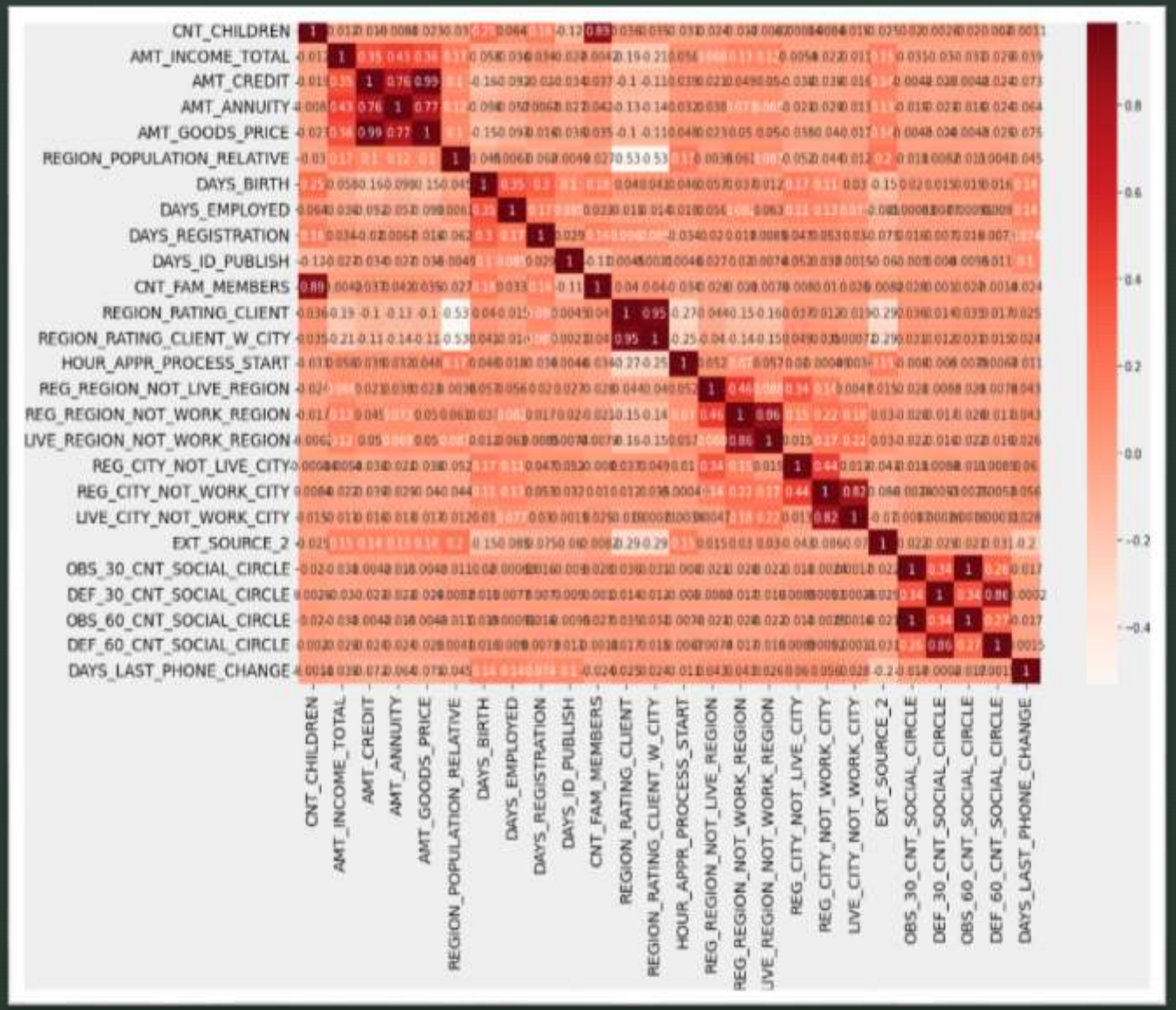


High correlation observed between:

- AMT_CREDIT and AMT_GOODS_PRICE - As the price of goods increases, the loan amount needed also increases.
- AMT_CREDIT and AMT_ANNUITY - As the amount of loan increases, the installment amount needed also increases.
- AMT_GOODS_PRICE and AMT_ANNUITY - As the price of goods increases, the installment amount needed also increases.

Multivariate analysis

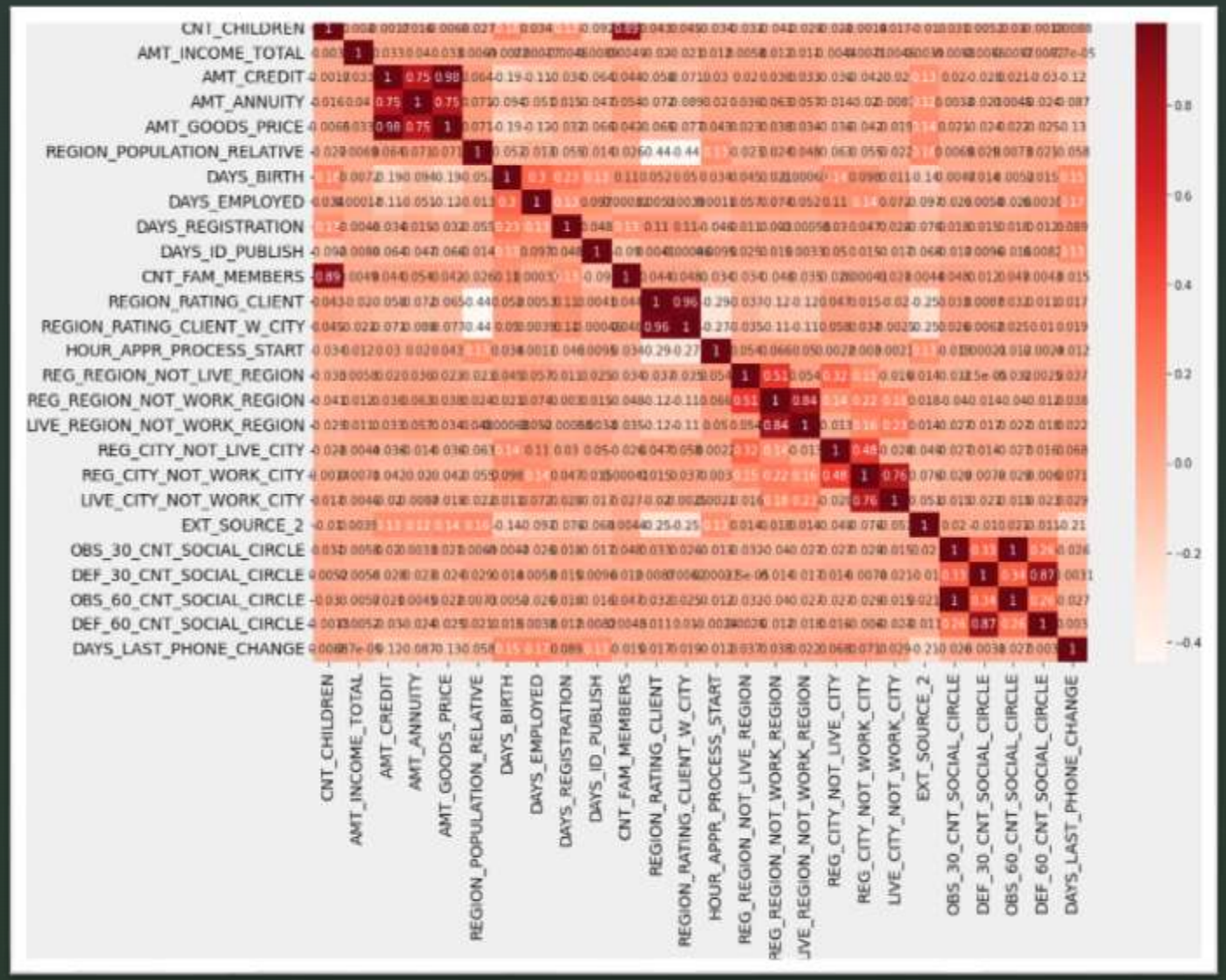
Correlation for target = 0



- Credit amount is higher for low age and vice-versa.
- Credit amount is higher for less children count customers have and vice-versa.
- Income amount is inversely proportional to the number of children client have, means more income for less children client have and vice-versa.

Multivariate analysis

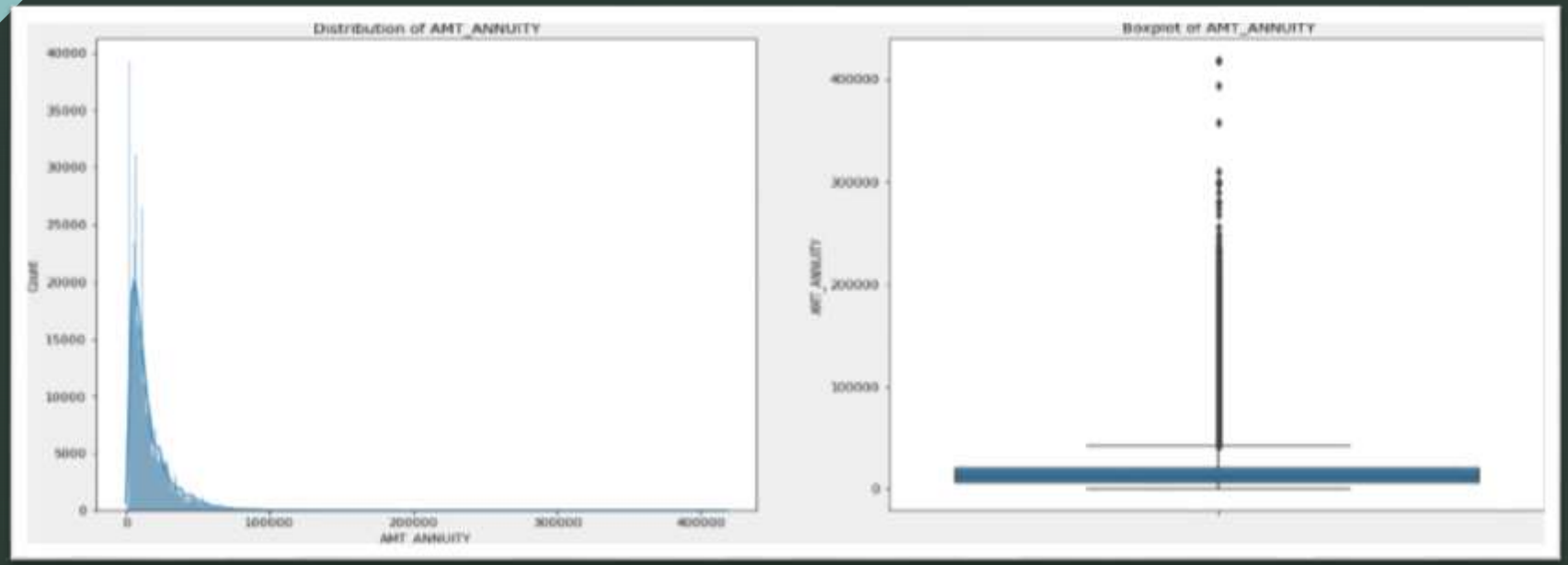
Correlation for target = 1

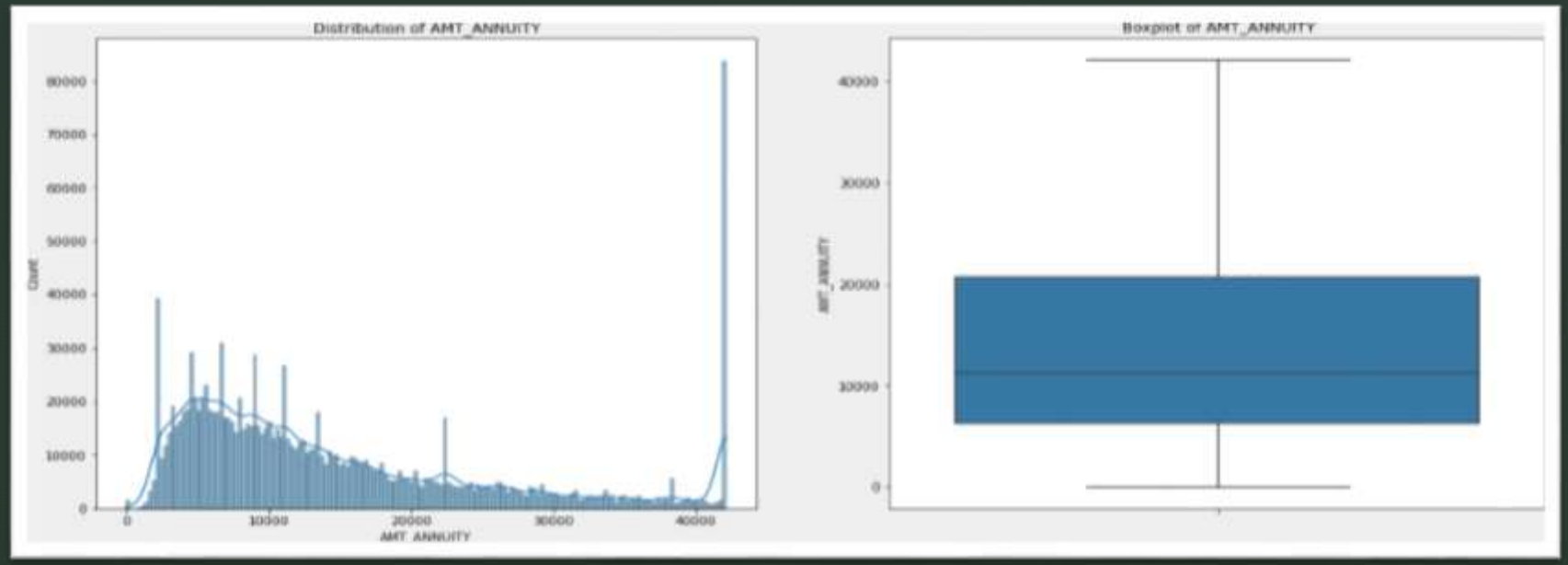


- This also has almost same observations as Target = 0

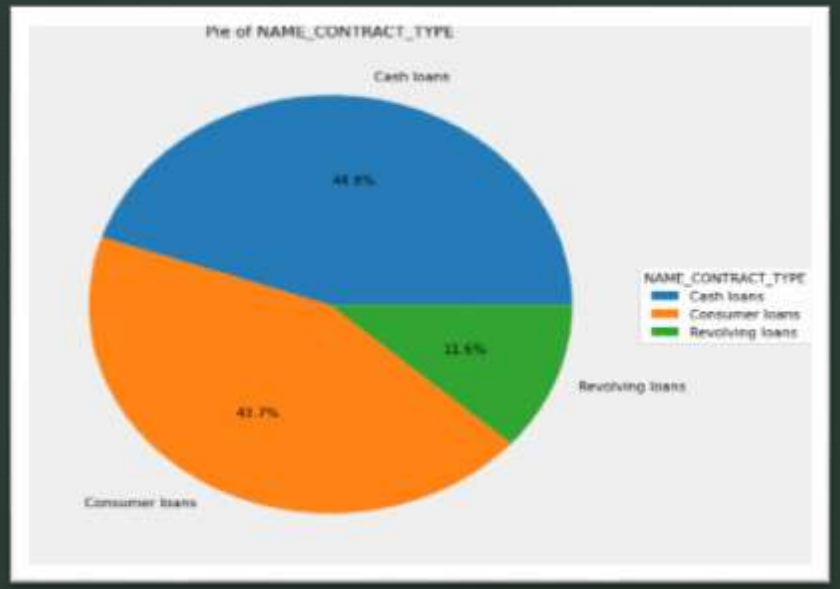
PREVIOUS APPLICATION

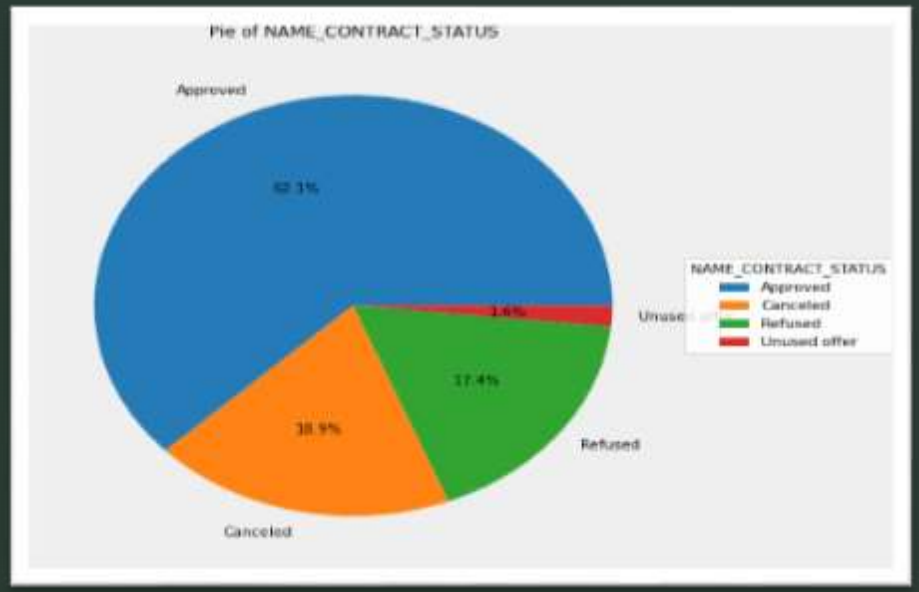
Treatment of Outliers (IQR Method)

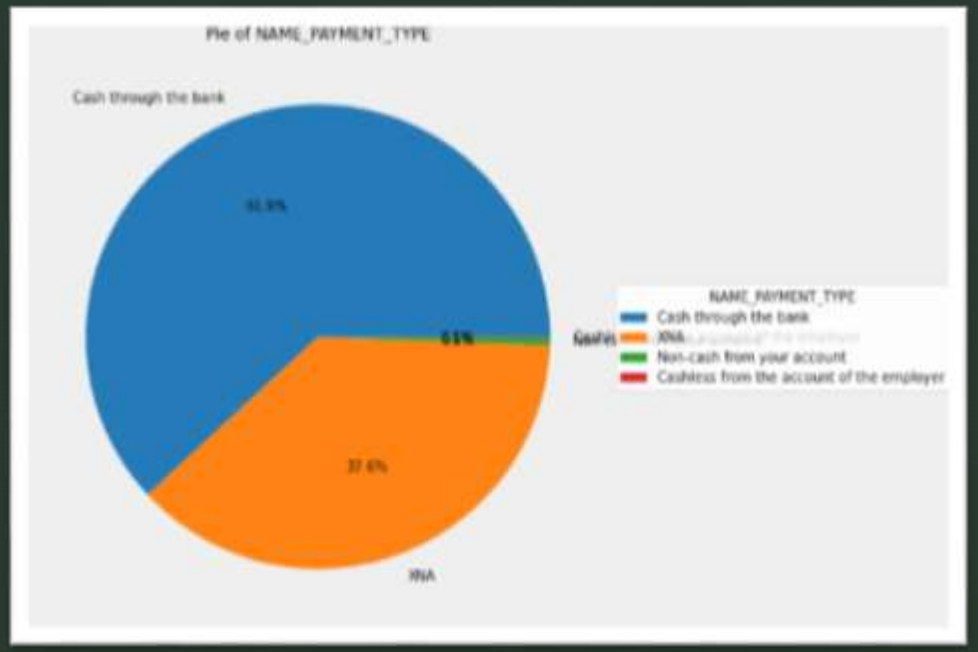




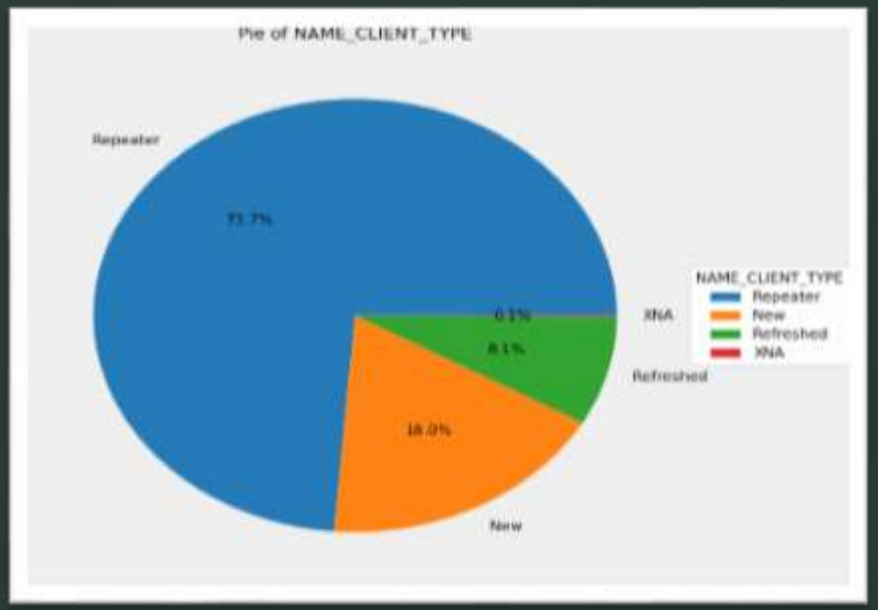
Univariate Analysis

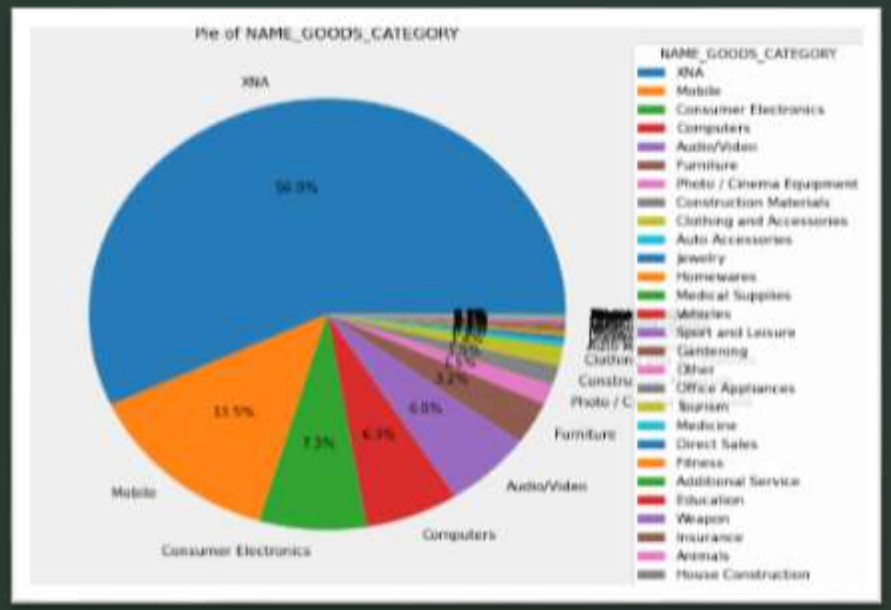


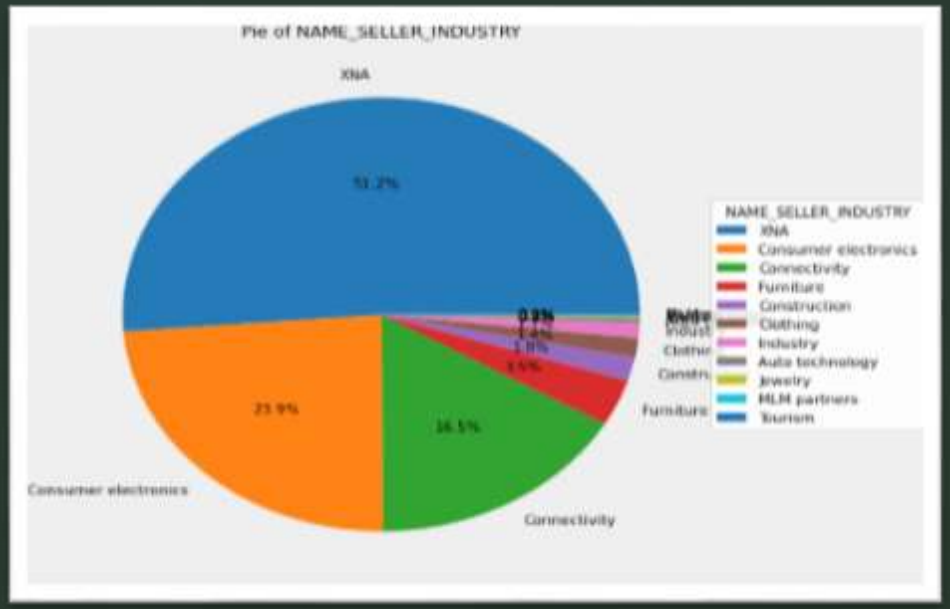




Univariate Analysis



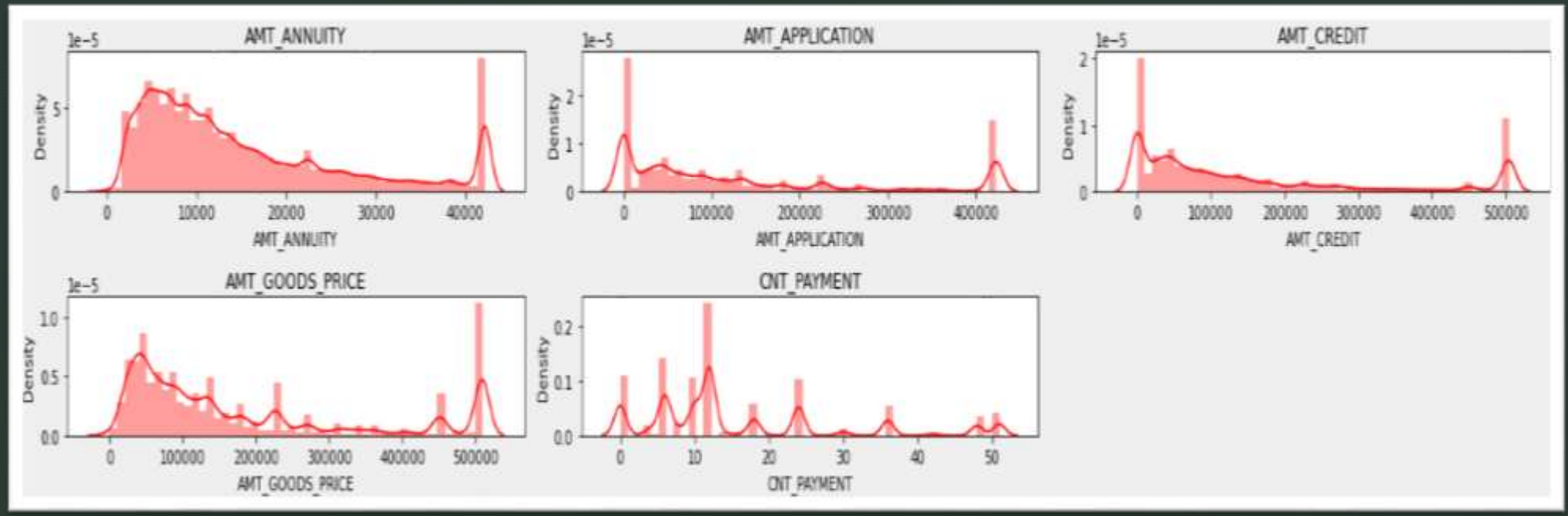




Univariate Analysis

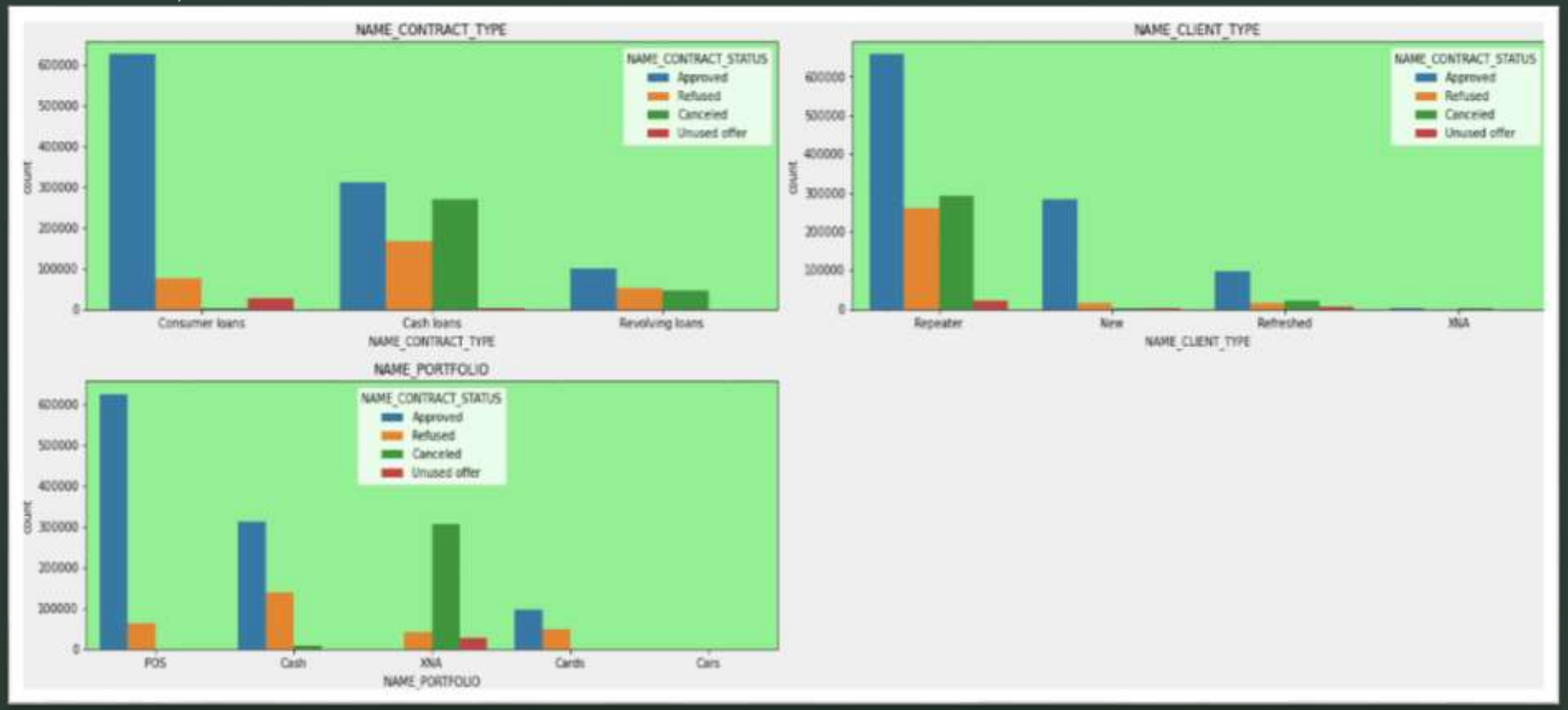
- Consumer loans account for 44% of all loans, while 45% are cash loans, and the remaining are revolving loans.
- Around 74% of the loan applicants are repeaters, while 18% of the applicants have newly applied.
- Approx 62% of the loan applications have been approved while around 19% have been cancelled and around 17% applicants has been refused straightaway for the loan.
- Around 41% of the applicants have taken loan for POS purposes.
- Around 62% of the applicants chose to pay through cash through bank for the loan application.
- Approx 27% of the loan applications was done by cross selling by the loan officials. Here around 64% of the data are missing.
- Around 43% of the client was acquired through credit and cash offices for loans.
- The NAME_SELLER_INDUSTRY column has 51% 'XNA' values, indicating missing or unspecified data. Among the specified categories, the Consumer Electronics industry is the next highest at 24%.
- POS Household with interest has the highest share ~ 16% for the most product combination.

Univariate Analysis of numerical columns



There are a significant share of outliers in the numerical variables.

Bivariate Analysis



- The highest number of applicants in the approved category are for consumer loans.
- There are no cancelled loans in the consumer loan category, but cash loans are significantly cancelled.
- There are a significant number of repeat loan customers across all contract status and it seems that due to the approval rate being higher, they gets pulled towards more loan applications.
- POS transactions are higher in the approval rate. Here also, more cash loans have been refused than other transactions.

Multivariate Analysis

Client type vs contract status with application amount

- For unused offer, the loan applications are low, may be due to the fact that they are not good offers for the customers.
- The application count is higher for the repeaters, but also refused cases for them is also high.
- Mostly refreshed applications are seen to be cancelled, may be due to the fact that bank does not think that they would not be able to repay them.

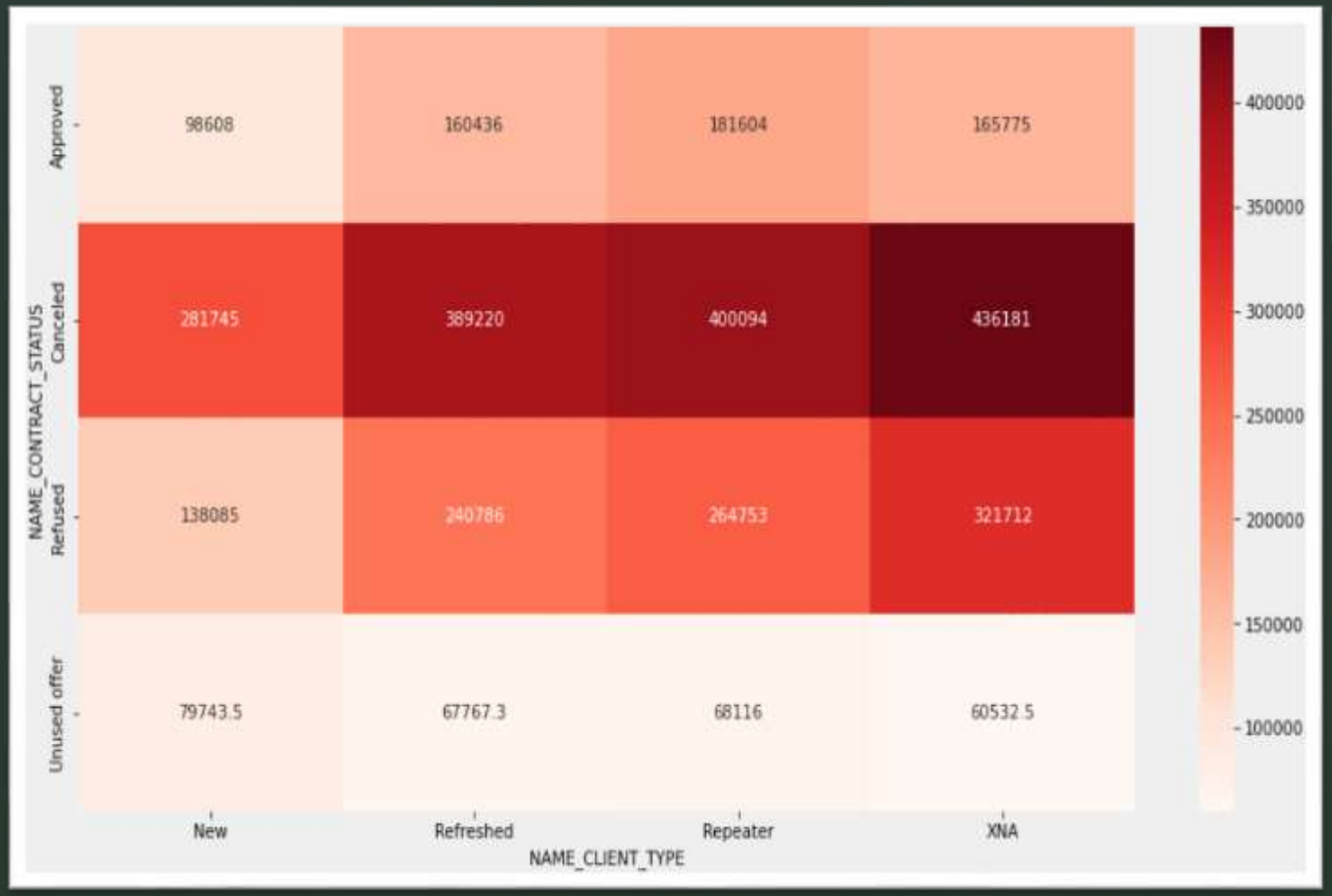
Multivariate Analysis

Client type vs contract status with credit amount

- Less credit amount for unused offers. This may be because the offers are not attractive enough.

Multivariate Analysis

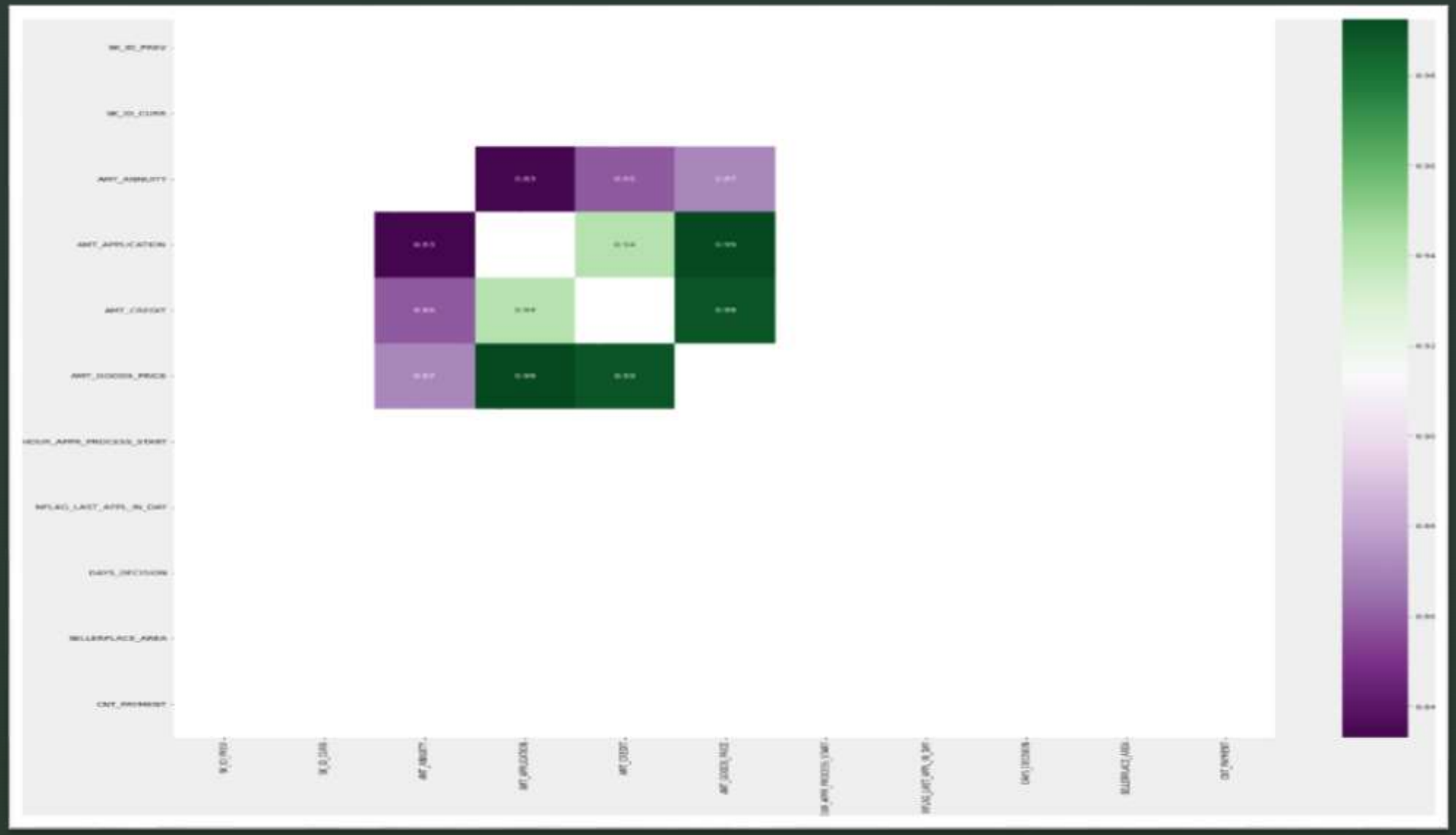
Client type vs contract status with goods price



- For cancelled and refused categories, we see that goods price is higher.

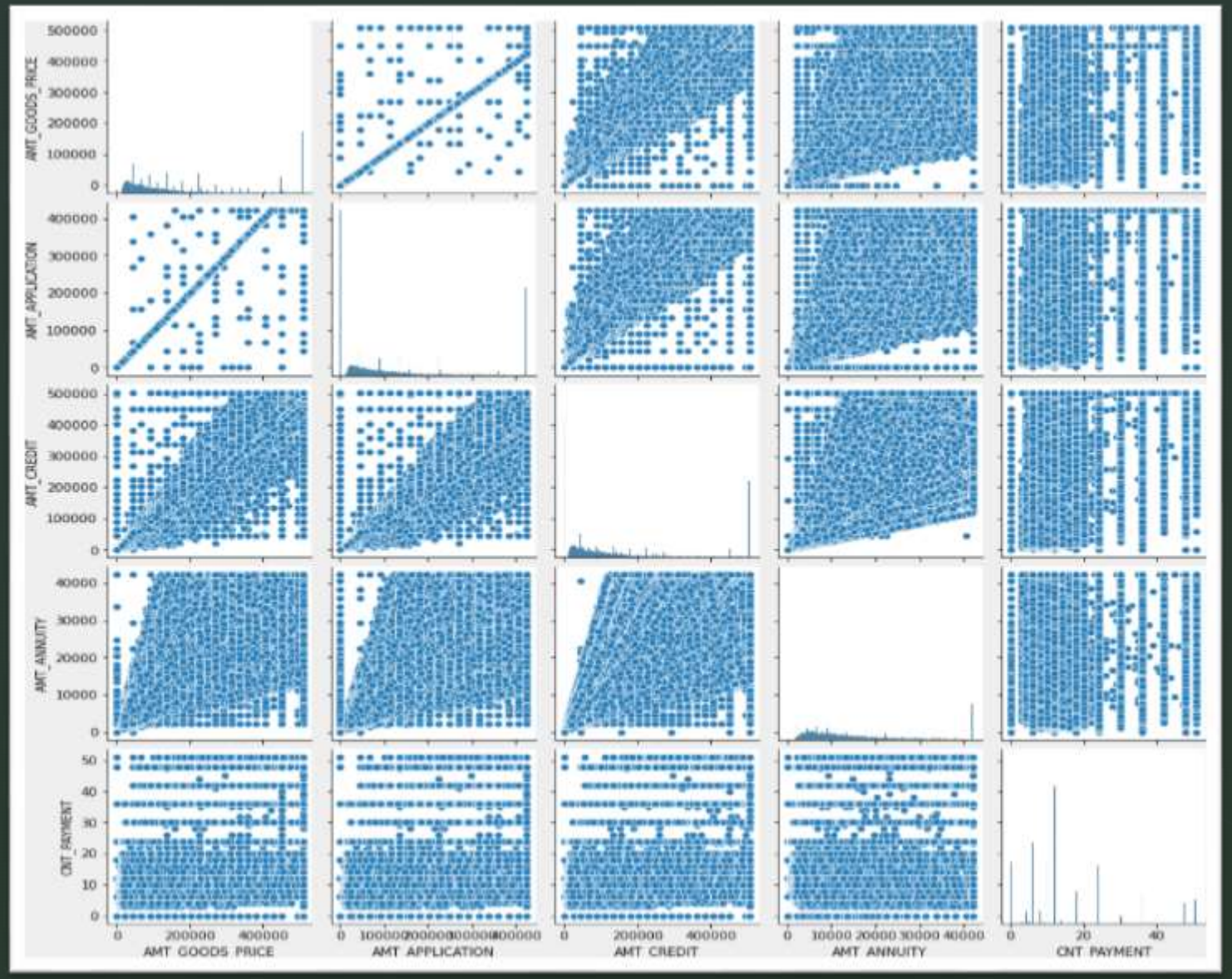
Multivariate Analysis

Correlation



Multivariate Analysis

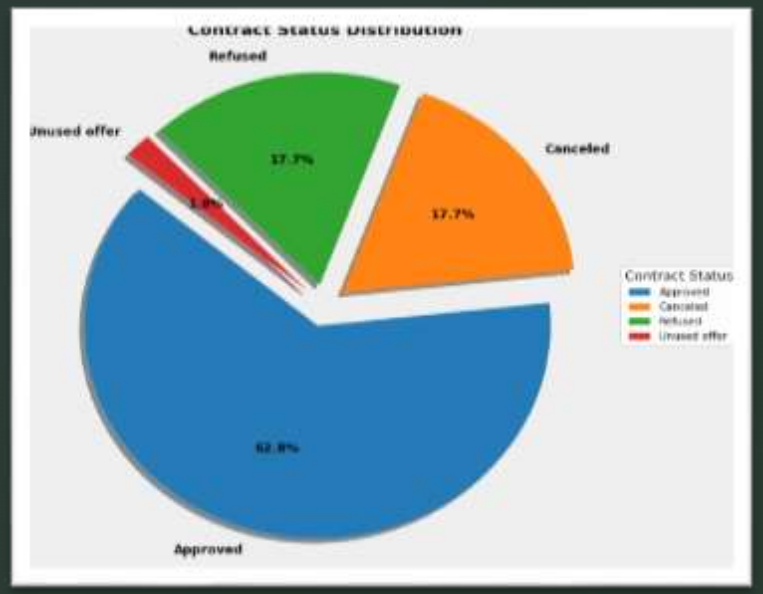
Pair plot between numerical columns



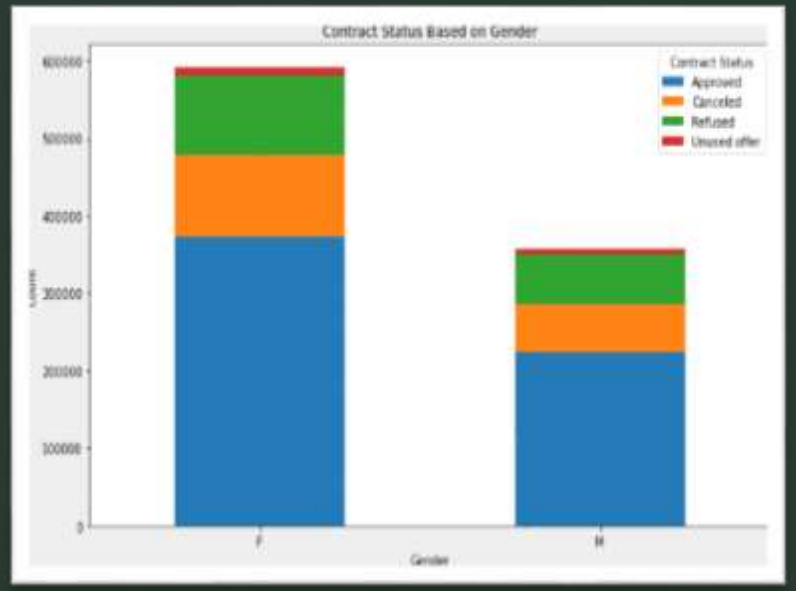
- AMT_GOODS_PRICE, AMT_ANNUITY and AMT_APPLICATION are highly correlated between themselves. This is because, more loan requirement will be there as there will be more purchase of goods.
- High correlation between AMT_GOODS_PRICE and AMT_CREDIT because more finance is required for more goods consumption.

MERGED DATASET

Contract Status Analysis

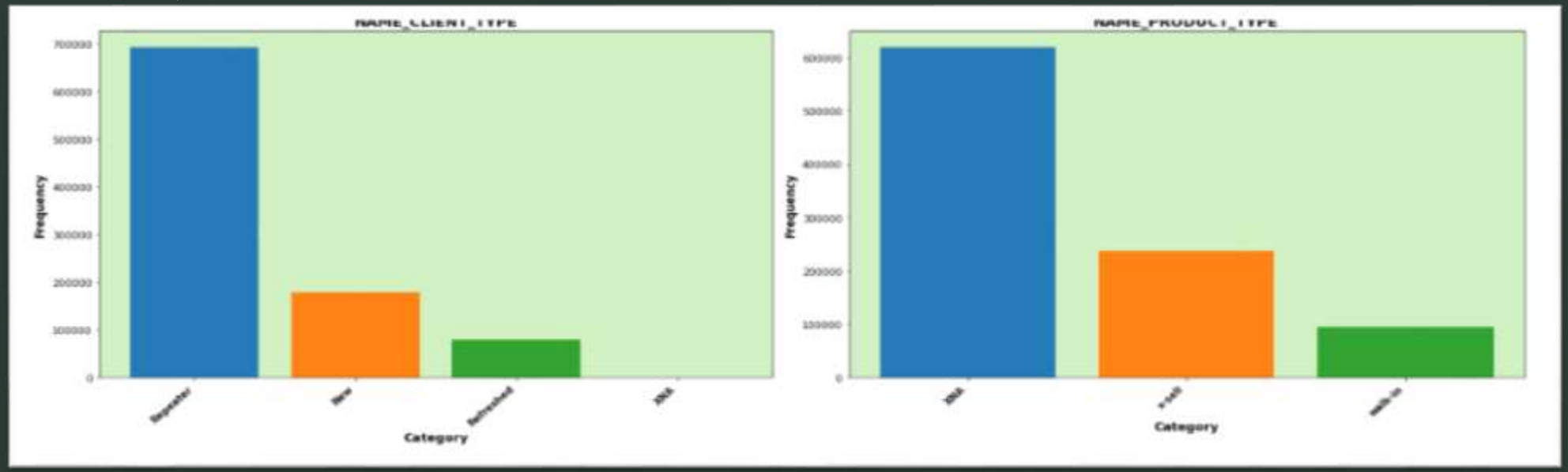


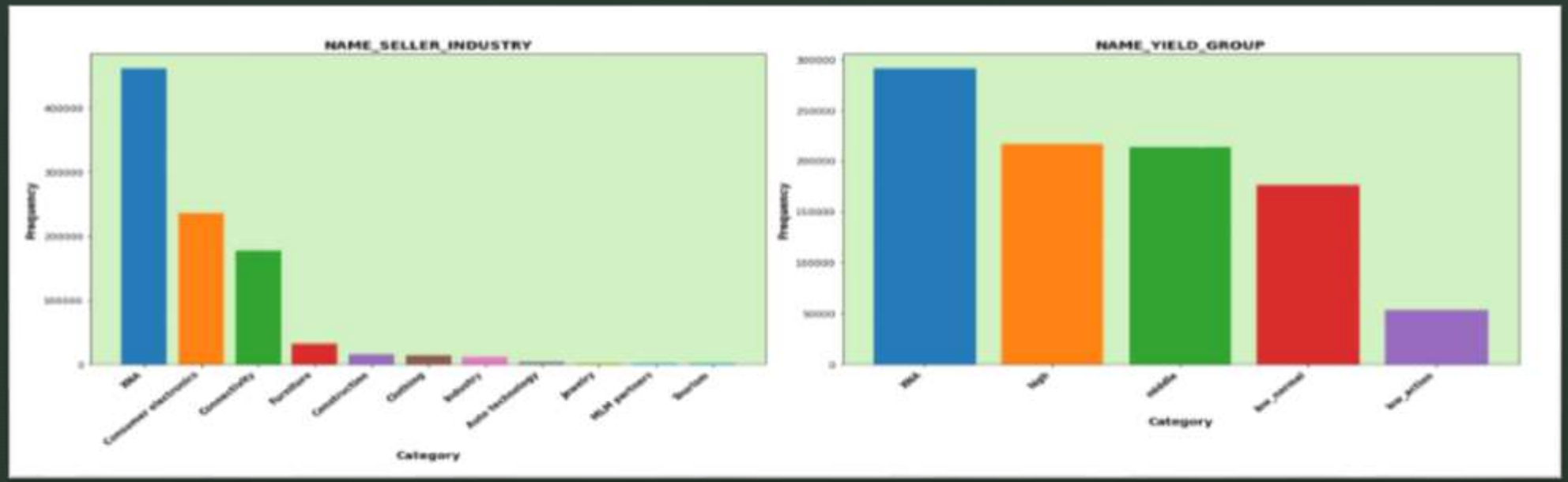
- A significant number of loan applicants (~ 63%) have had their application got approved by the bank.



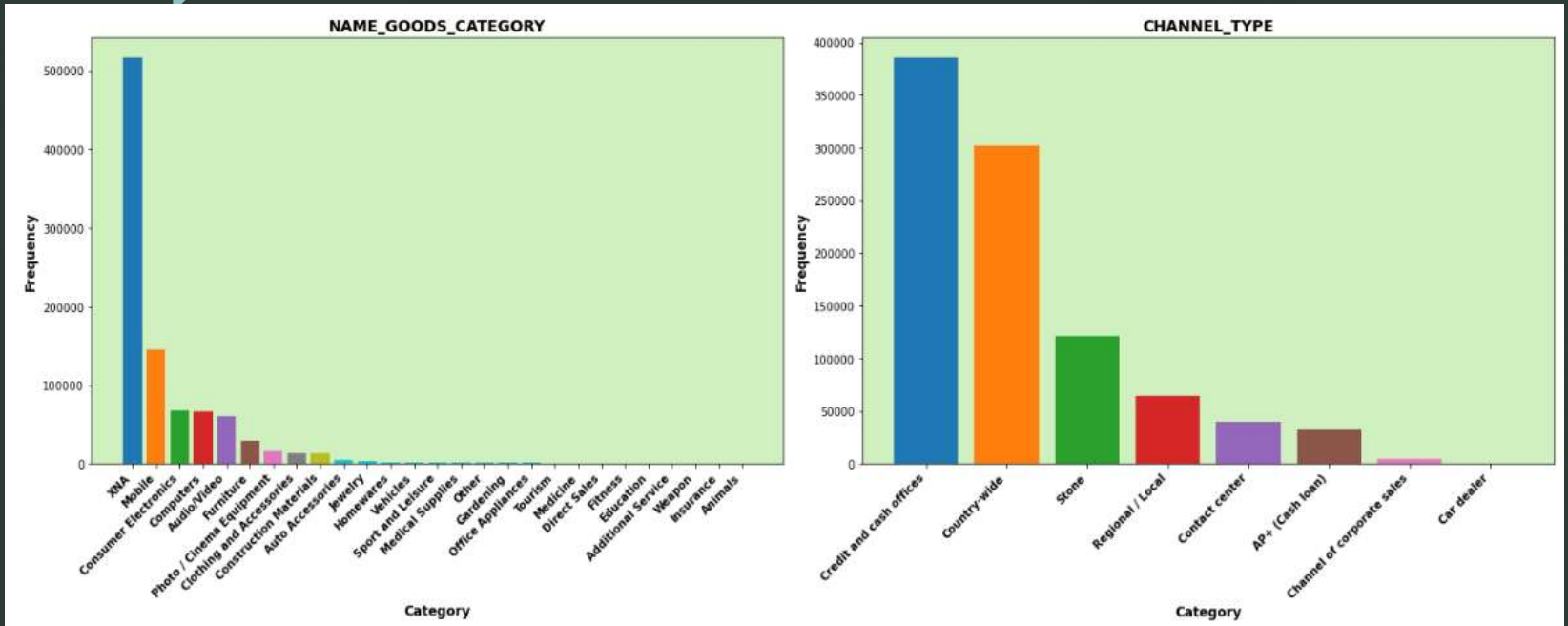
- Mostly the applications by female customers get approved by the bank than males. It may be due to the fact that females maintain a better repayment rate and track record than men.

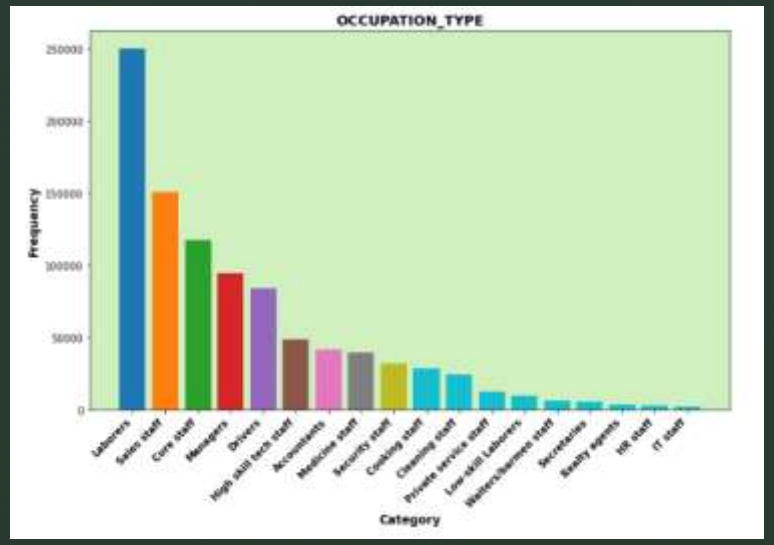
Univariate Analysis





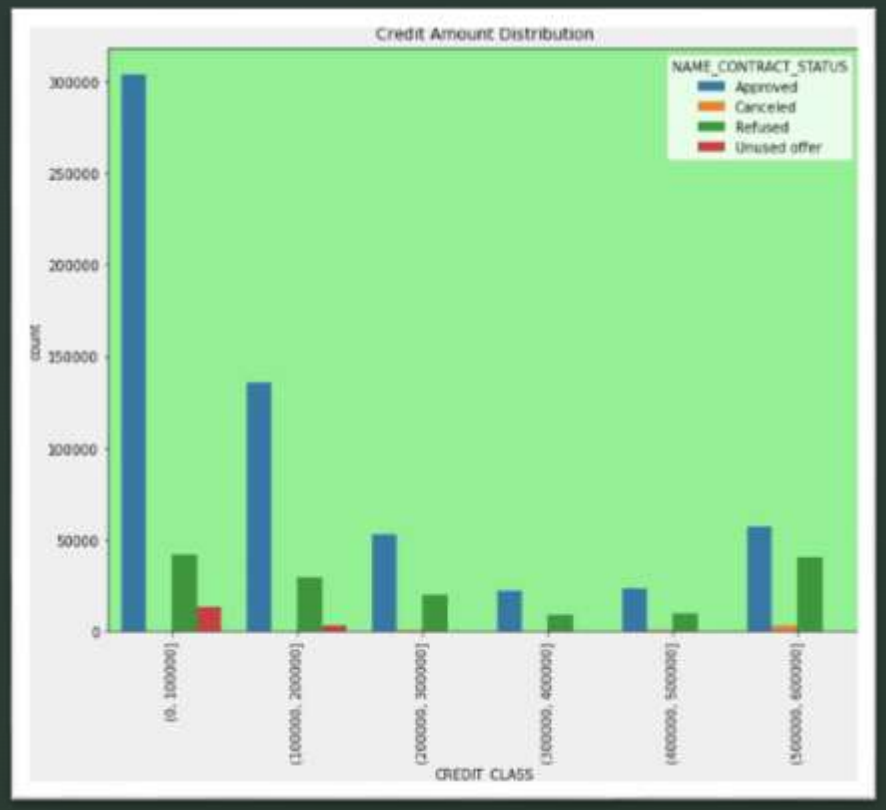
Univariate Analysis



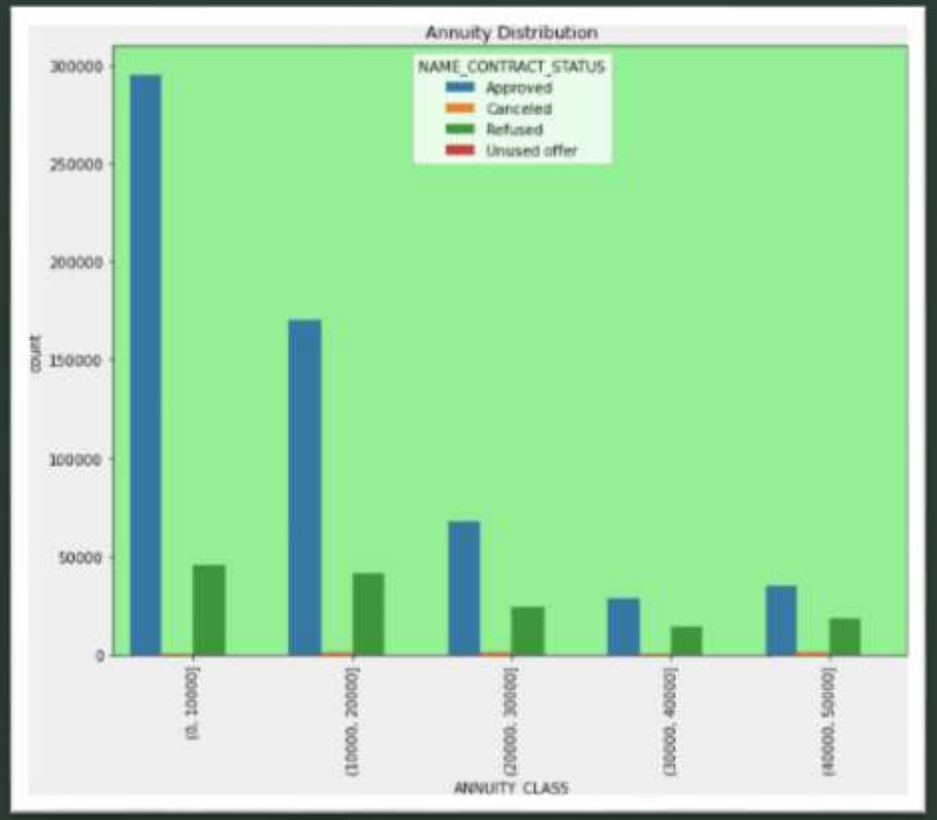


- There are more repeater customers than others.
- A significant number of loan applicants were acquired through cross selling by banks.
- Consumer electronics has the highest share than others.
- Mobile and computer leads the goods category.
- Count of Credit and cash offices is maximum followed by Country-wide.
- Laborers are the highest in occupation type.

Bivariate Analysis of numeric columns

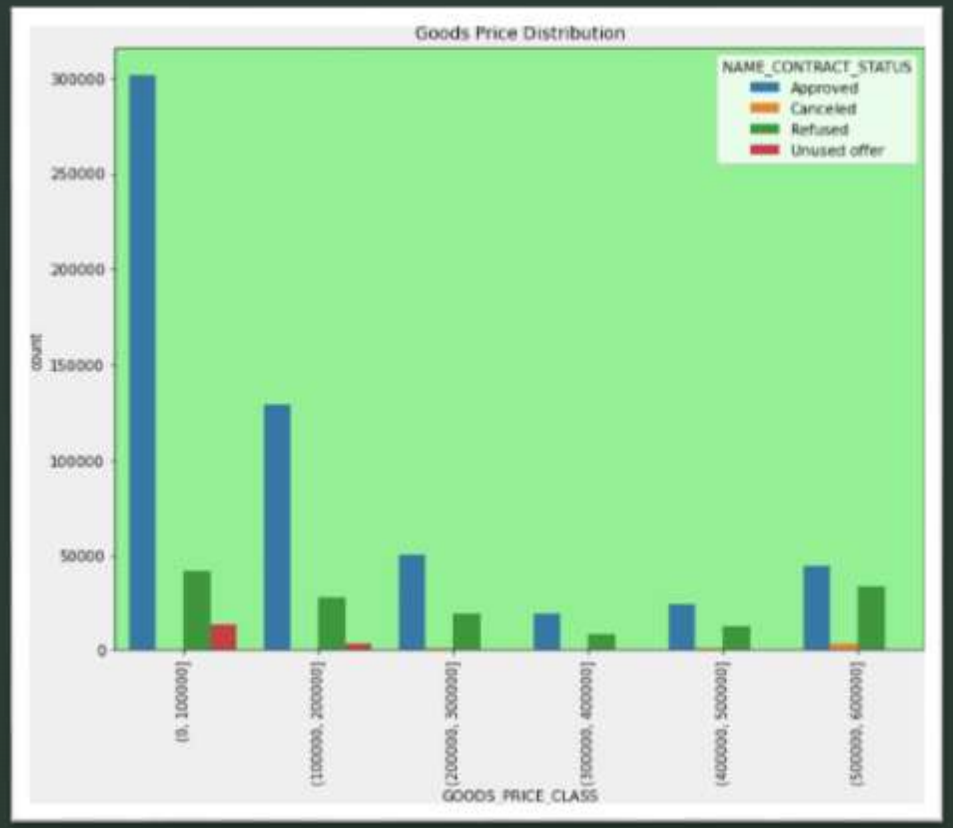


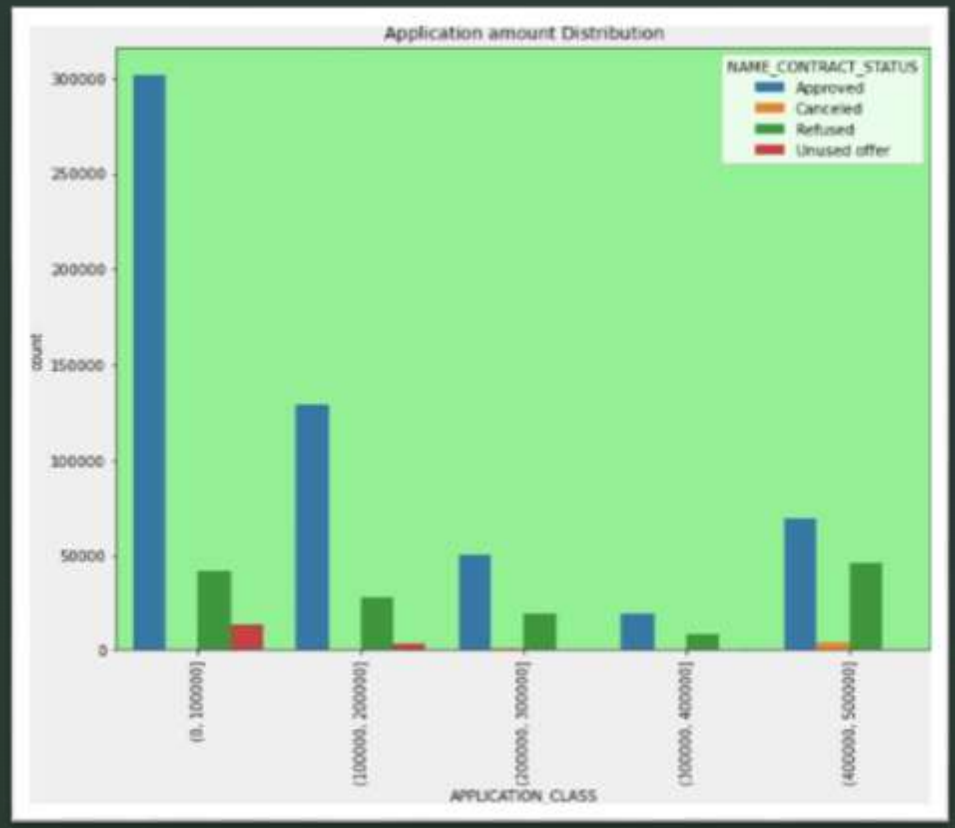
- Credit class from 0 to 100K are generally having their loan applications approved.



- Customers who are having annuity of 10000 got their loans approved.

Bivariate Analysis of numeric columns

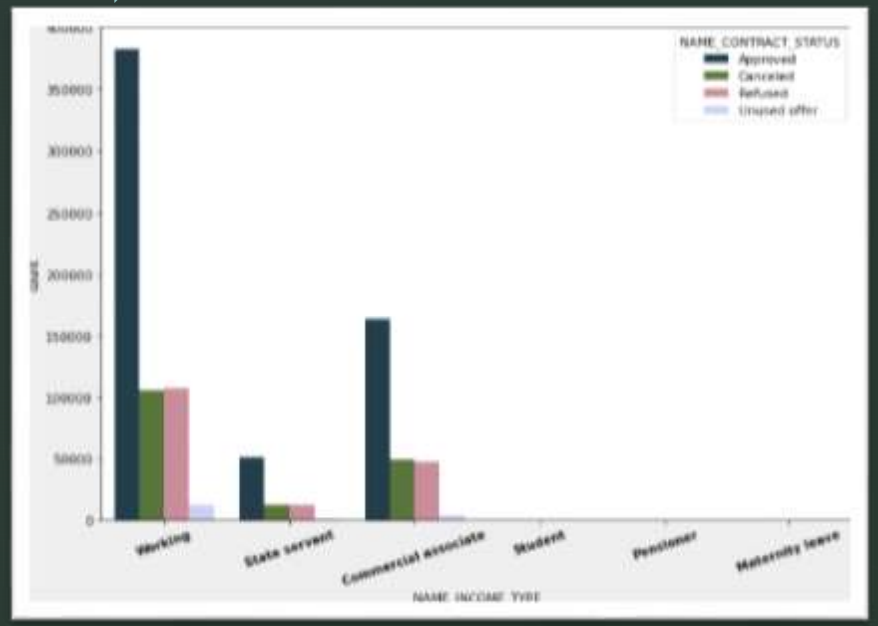


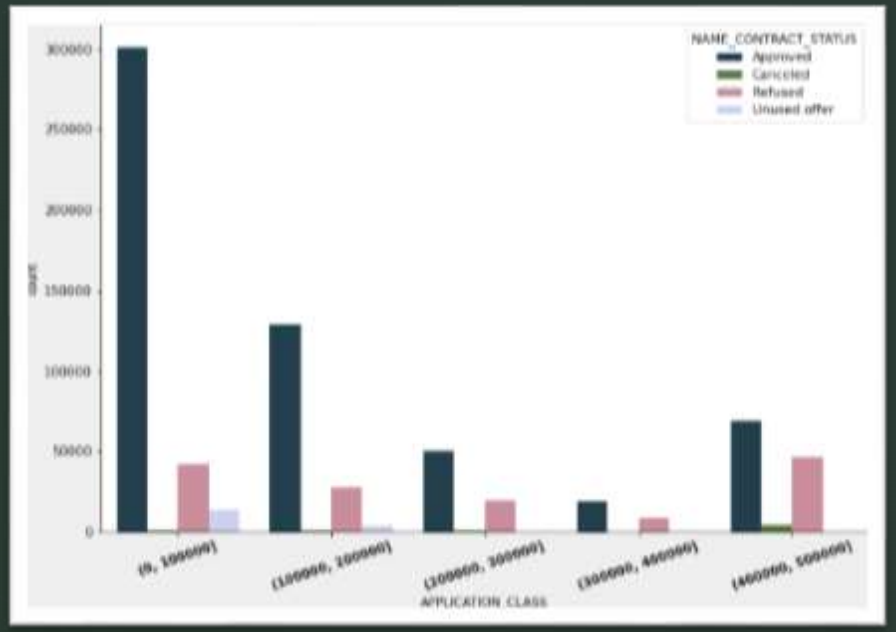


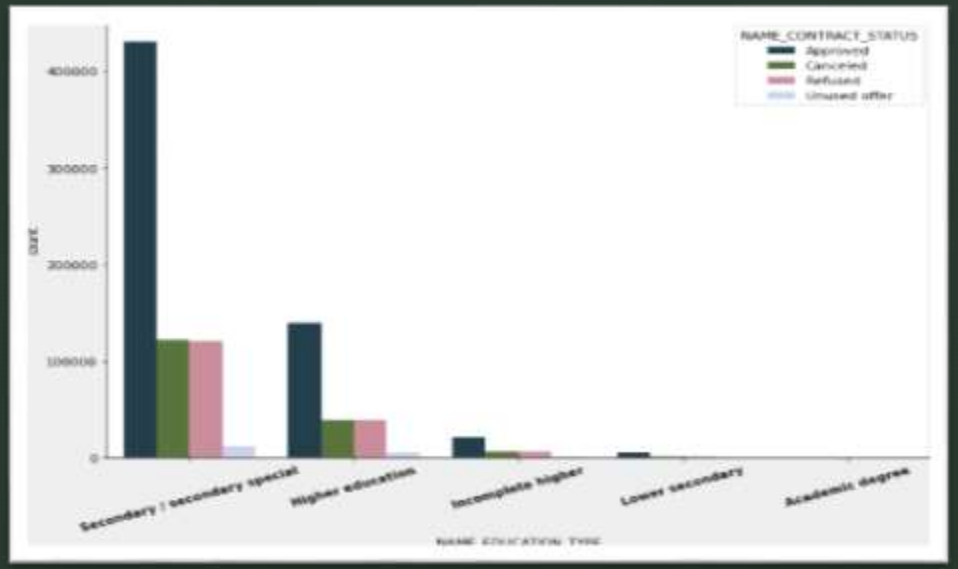
- Goods Price till 100K are generally having their loan applications approved.

- Customers of application amount distribution till 100k got loans approved.

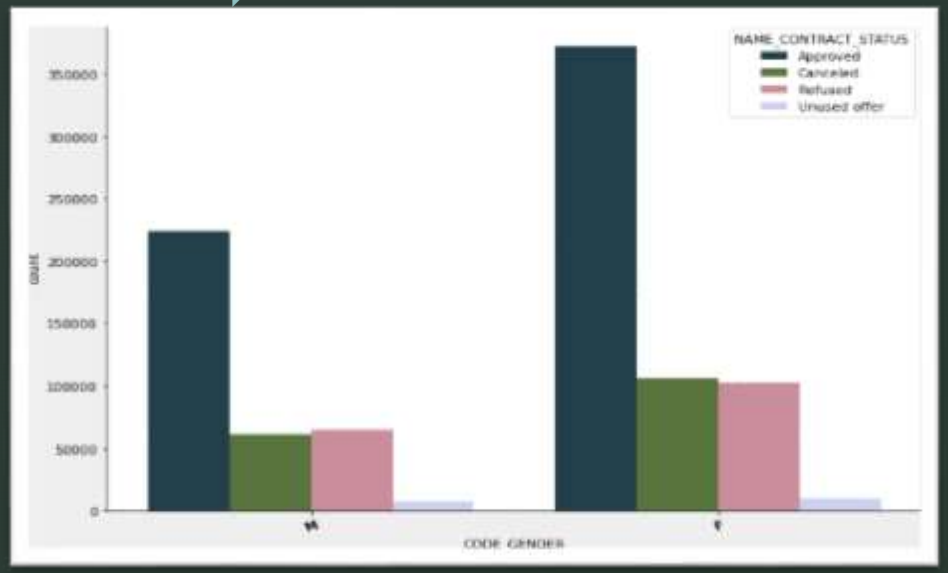
Bivariate Analysis of categorical columns

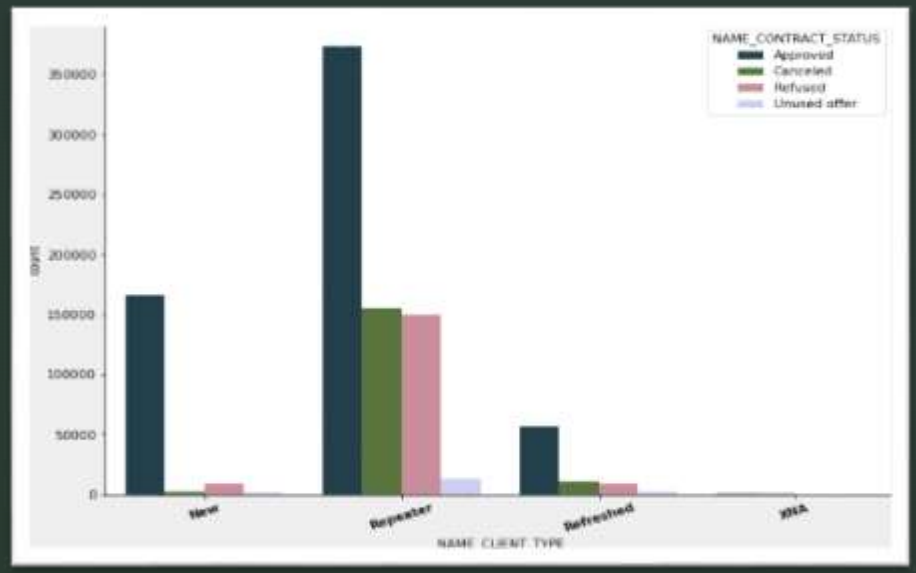


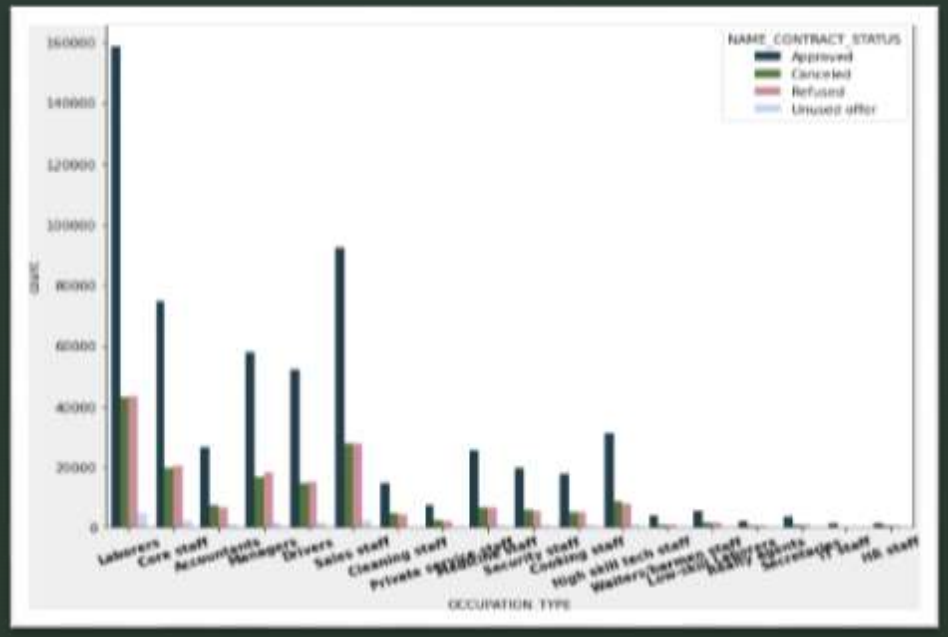


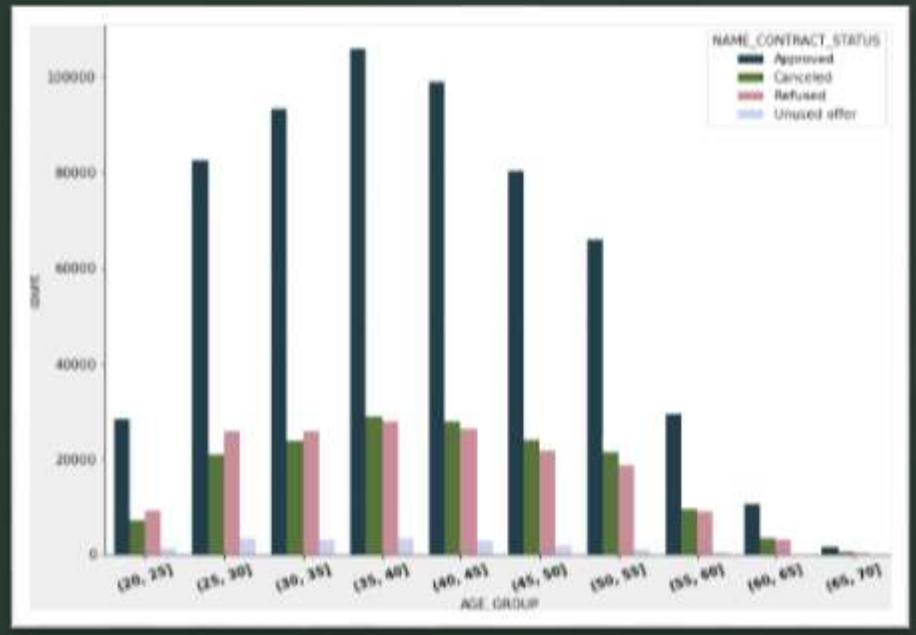


Bivariate Analysis of categorical columns







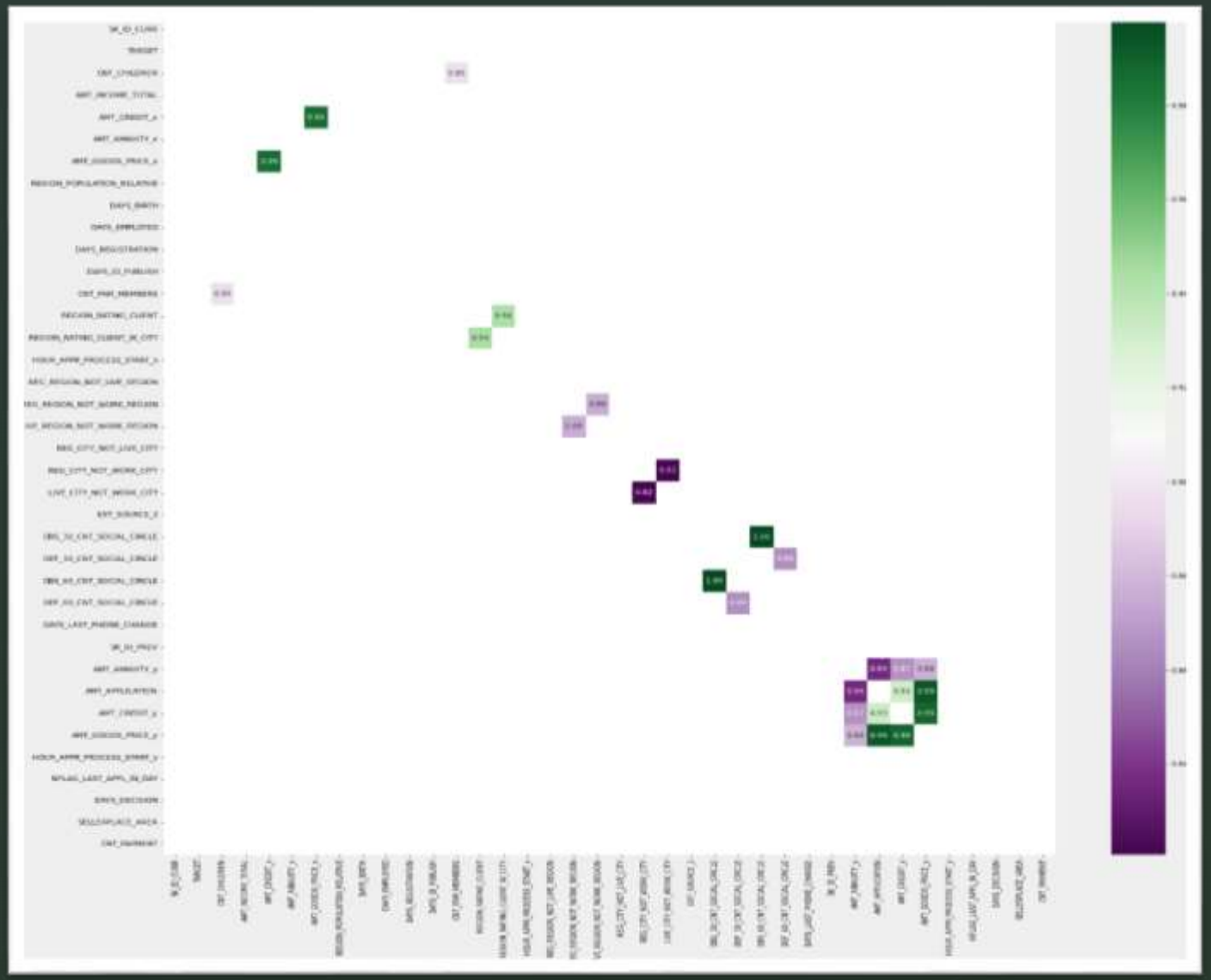


Bivariate Analysis of categorical columns

- Customers under the age bracket of 35-40 years have higher tendency of loan approval
- Secondary/Special education people are having higher loan approval , also females get loan approval more than male
- Repeater customers are having high frequency of approval compared to others.
- Workers are having high chance of getting loan approvals as per previous record.

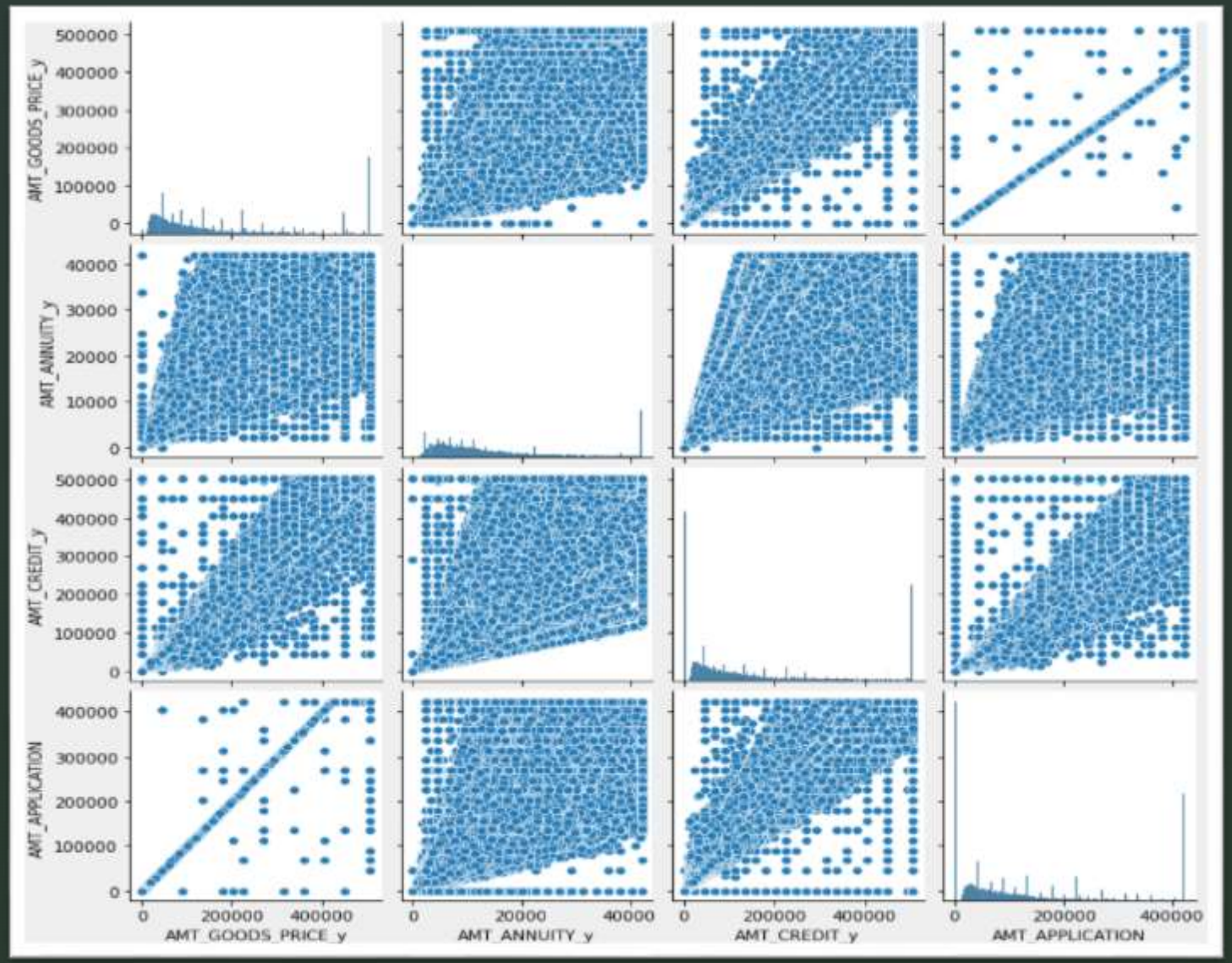
Multivariate Analysis

Correlation



Multivariate Analysis

Pair plot for numerical columns



- As the loan amount increases, the installment amount also increases.
- As the price of goods increases, the installment amount needed also increases.
- As the price of goods increases, the requirement of loan also increases.
- As the loan application increases, the loan amount approved also increases.
- As the loan application increases, the installment amount needed also increases.

Conclusion

- Banks should target more on ‘Student’ ,’pensioner’ and ‘Businessman’ with housing type other than ‘Co-op apartment’ for successful payments.
- Banks should target less on income type ‘Working’ because they tend to miss payments more frequently.
- Banks should target as much as customers from housing type ‘With parents’ as they are having least number of unsuccessful payments.
- Customers under the age of 30 frequently encounter difficulties in making payments, but customers aged over 40 are considered reliable borrowers as their refusal rate tends to fall down. So, banks should target more customers more than age 40.
- Customers with lower secondary education tend not to make payments on time. So, banks should go for an educational background verification check before approving loans.
- Banks should always target married customers as they tend to make loan repayments on time.

Thank You