



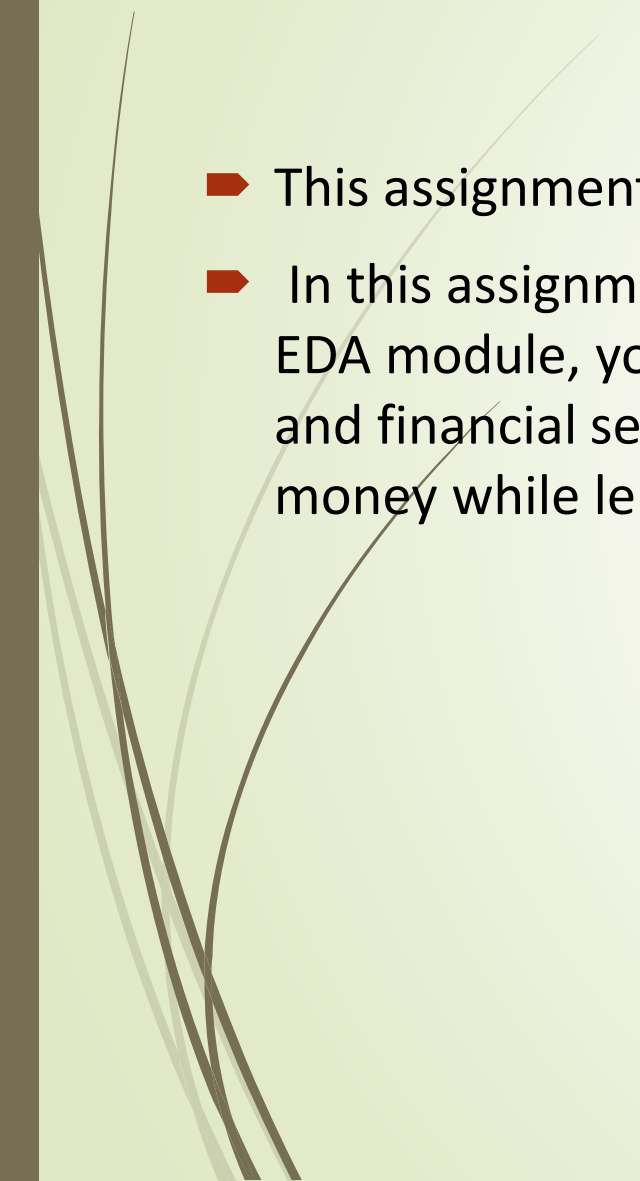
Credit EDA Assignment

ROHAN SHARMA

DSC-49



Problem statement

- This assignment aims to give you an idea of applying EDA in a real business scenario.
 - In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.
- 

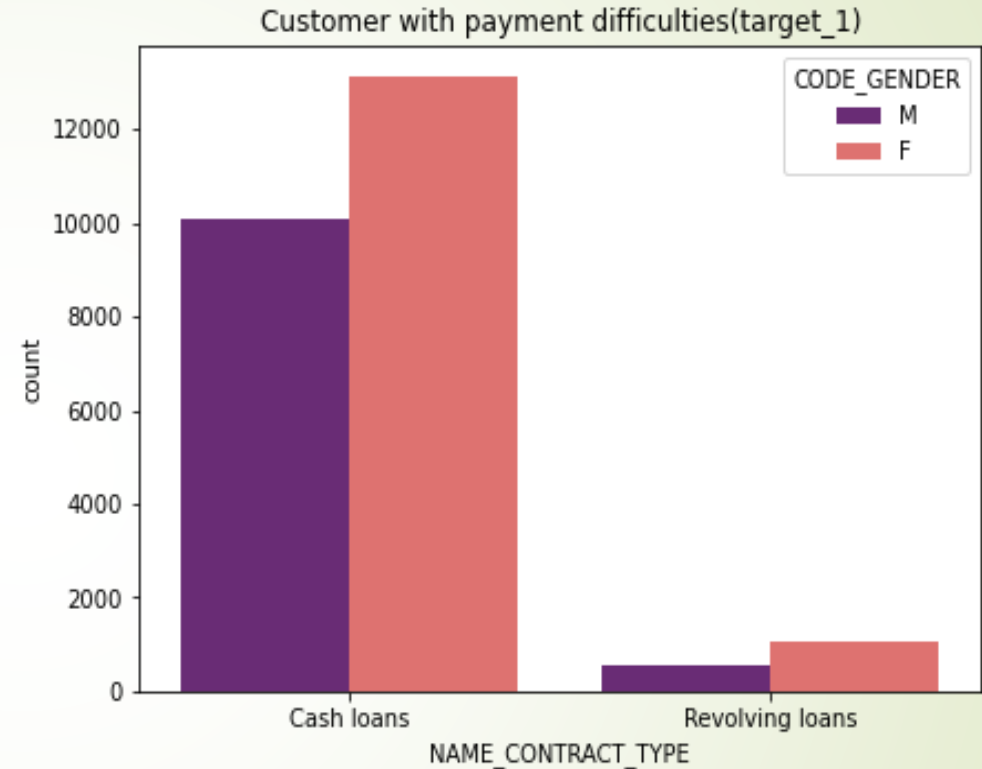
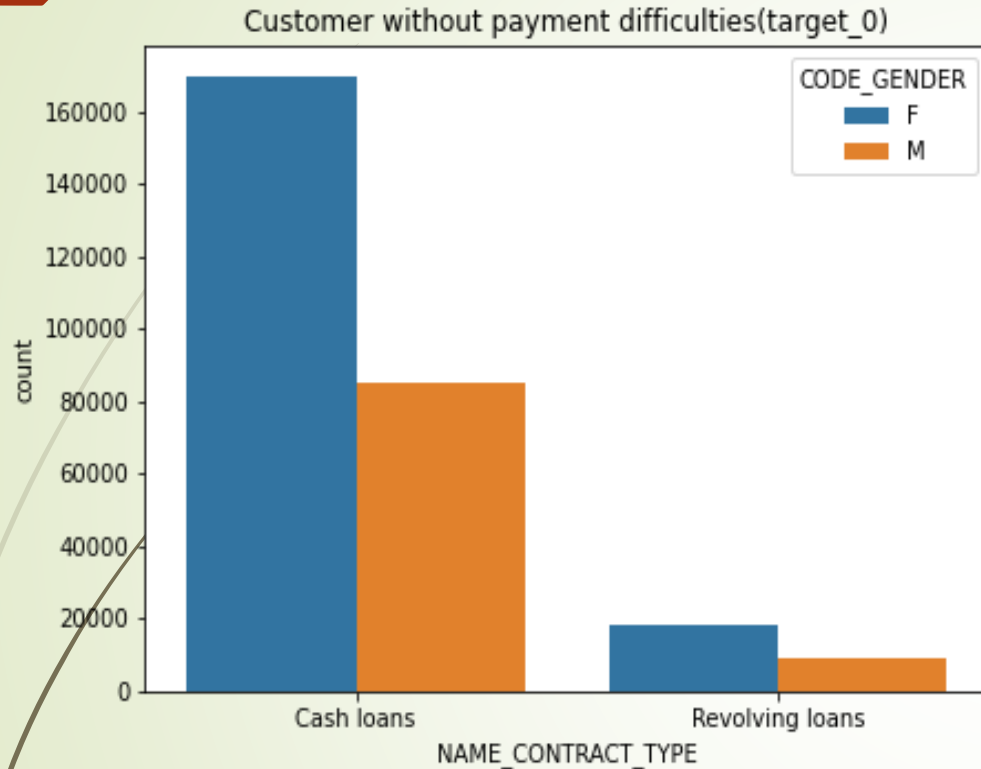


Business Objective

- 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.

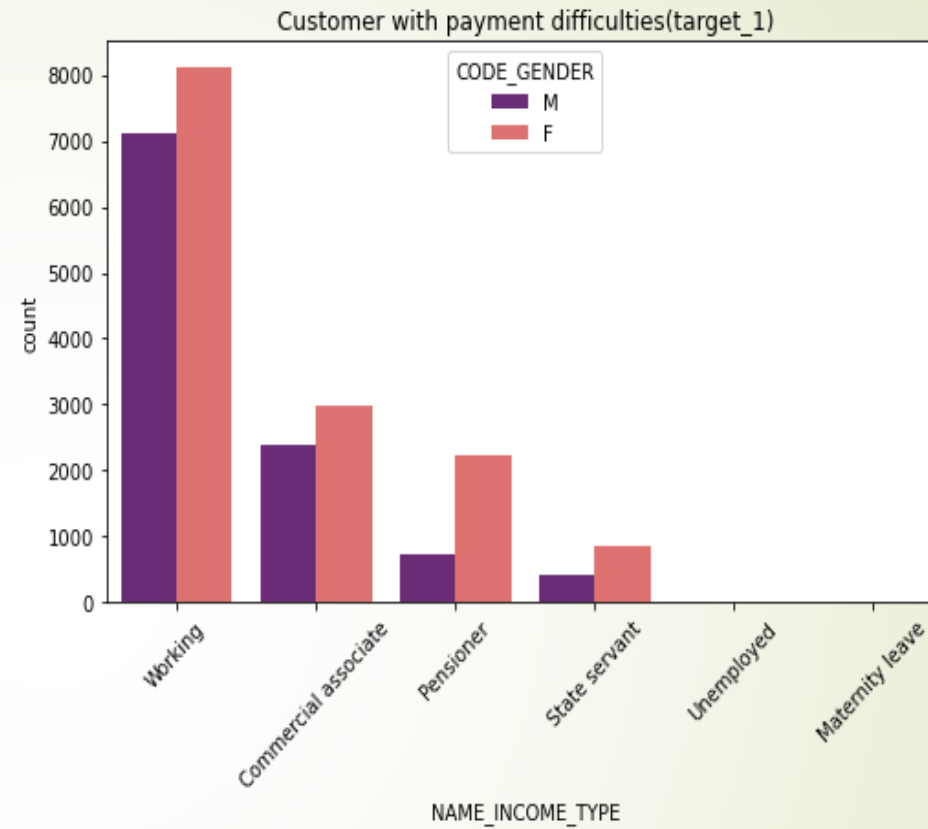
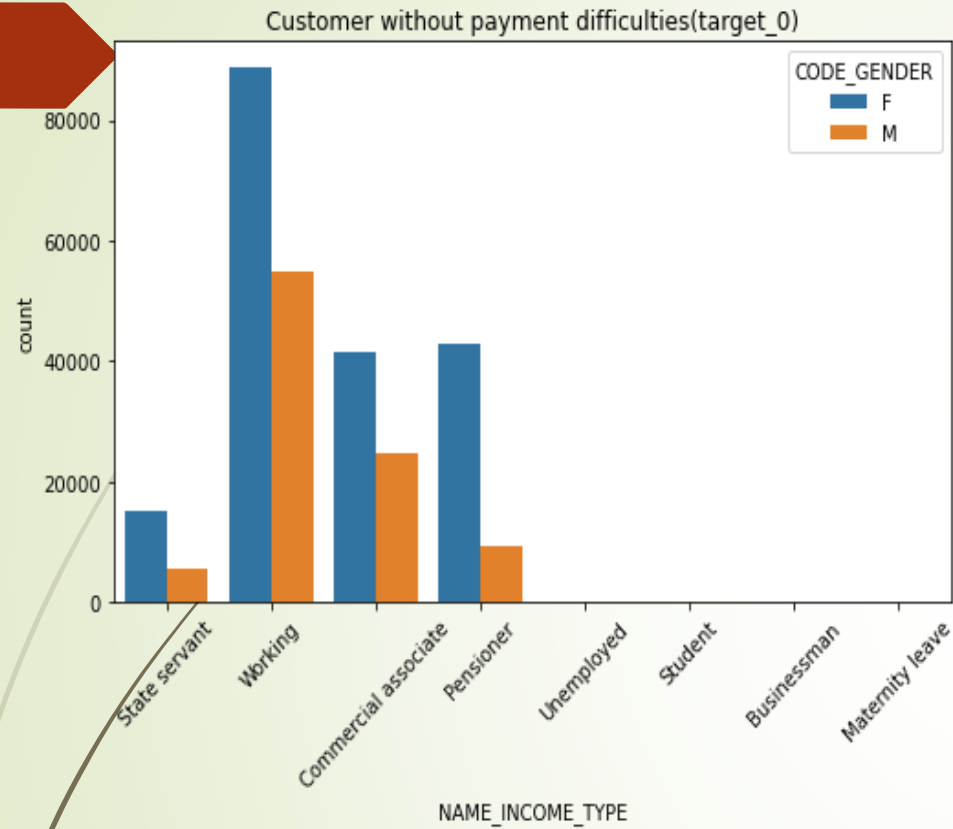
Application Data Analysis

Univariate Analysis



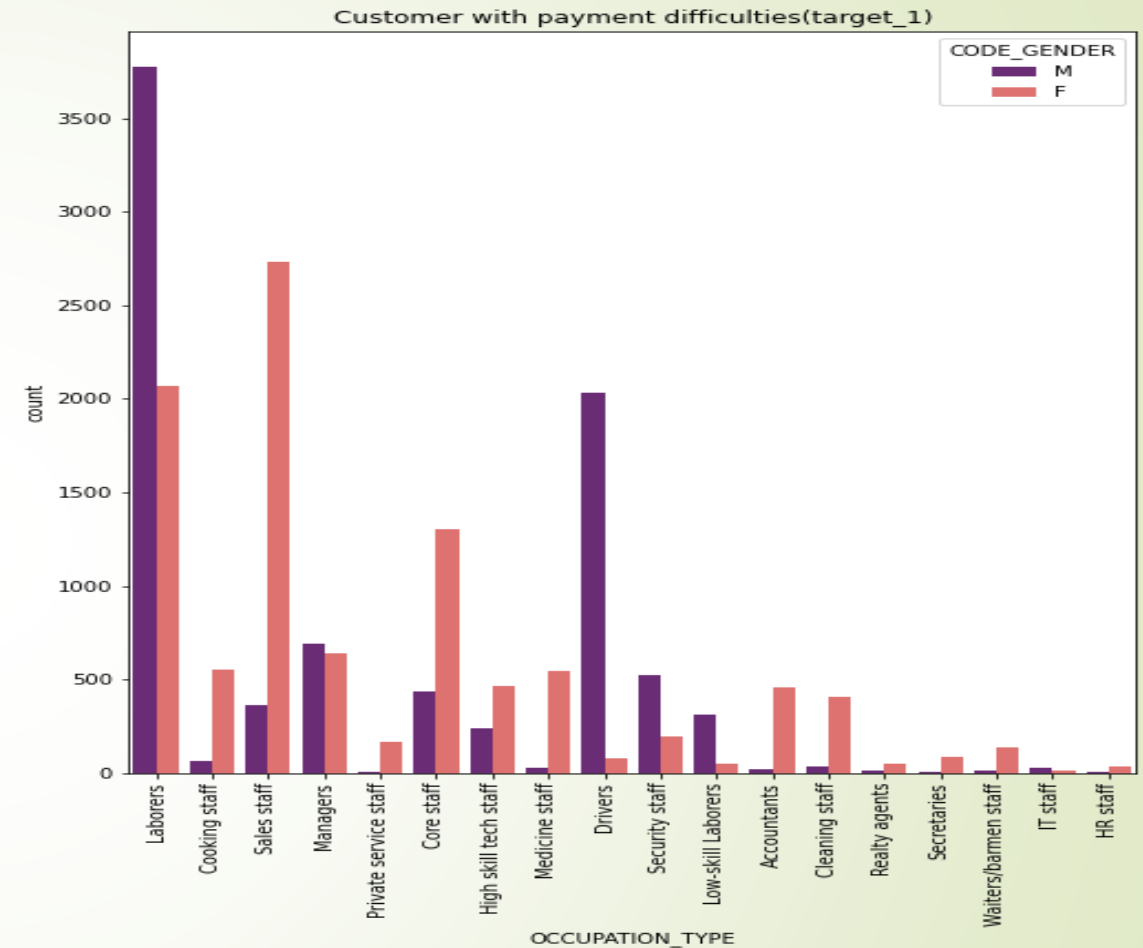
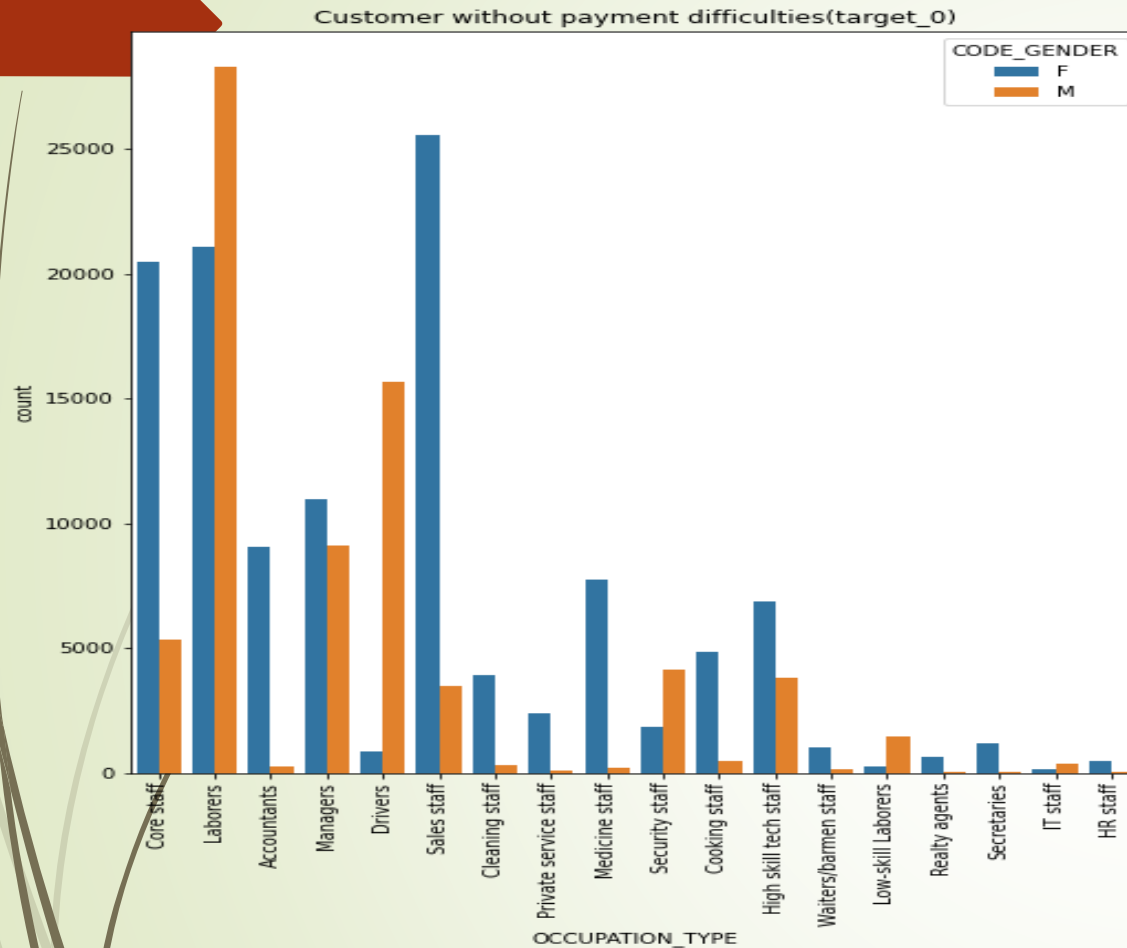
1. For both type of customers i.e. defaulters or non-defaulters, people with cash loans are higher.
2. Female counts are higher than male, for both customers.

Income type count



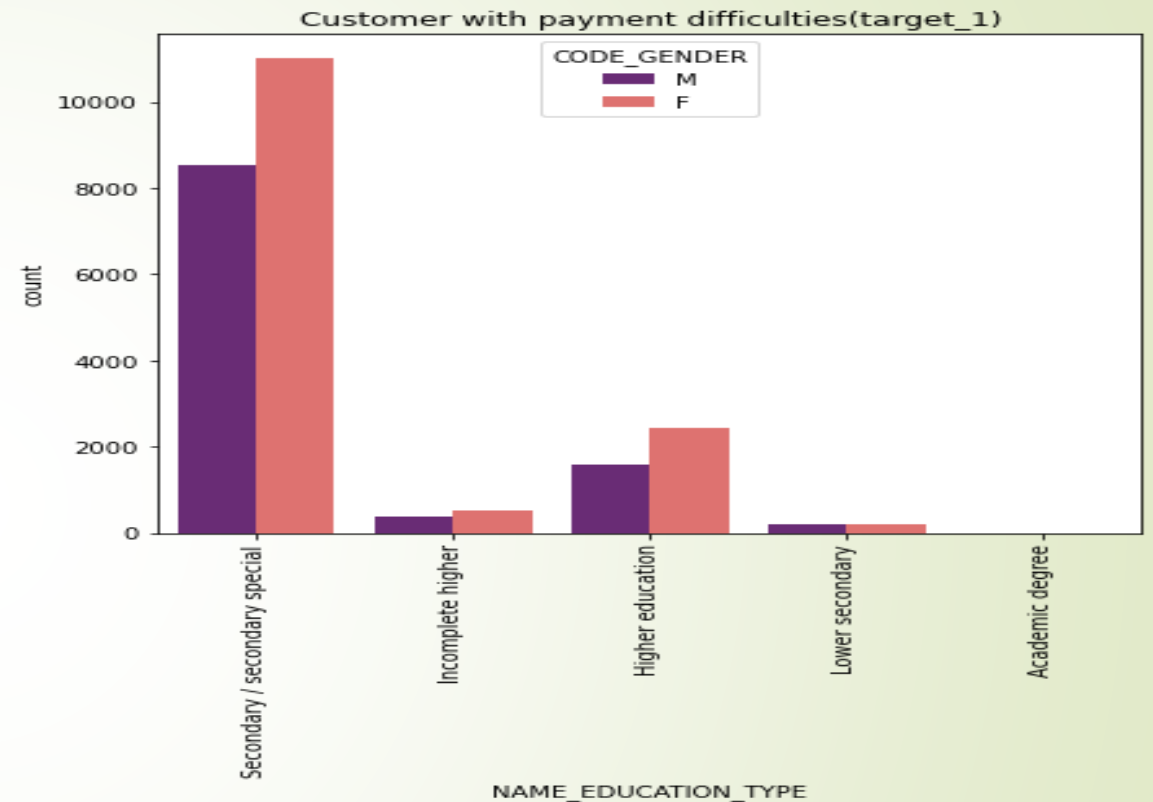
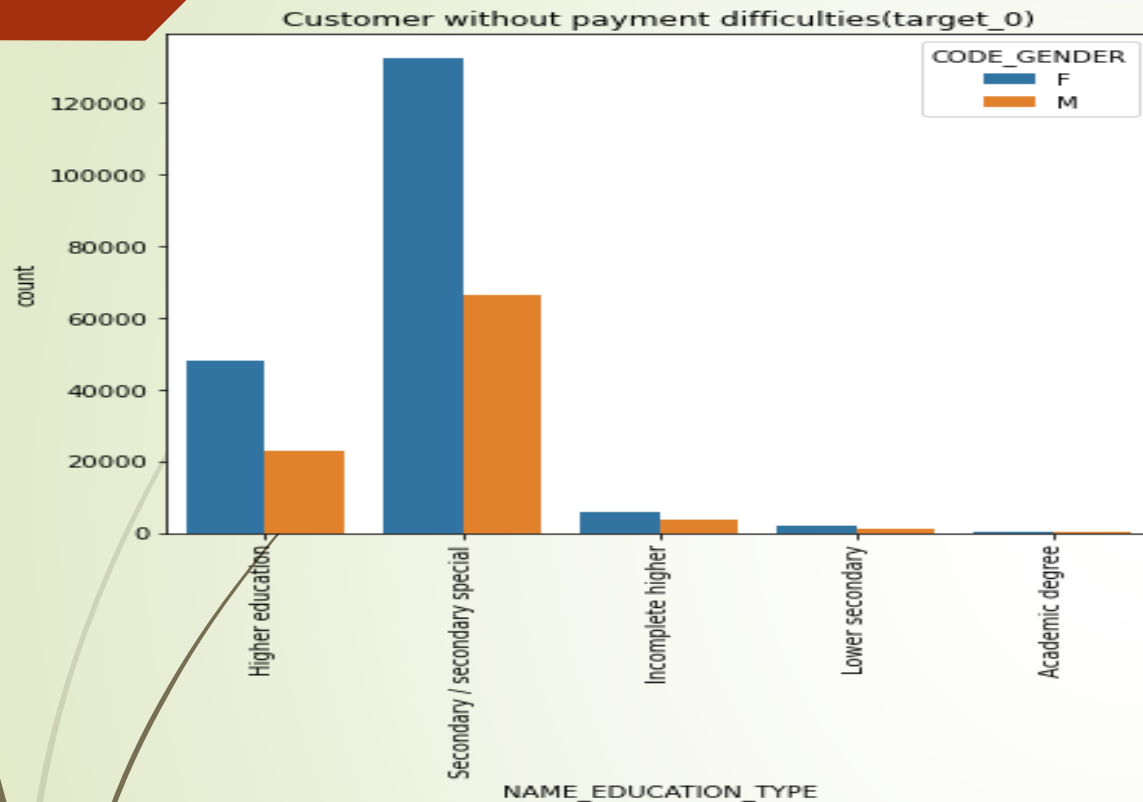
1. Working & commercial associate are higher than others.
2. Female counts are higher than male.
3. Less number of credits for student and Maternity leave.

Occupation type count



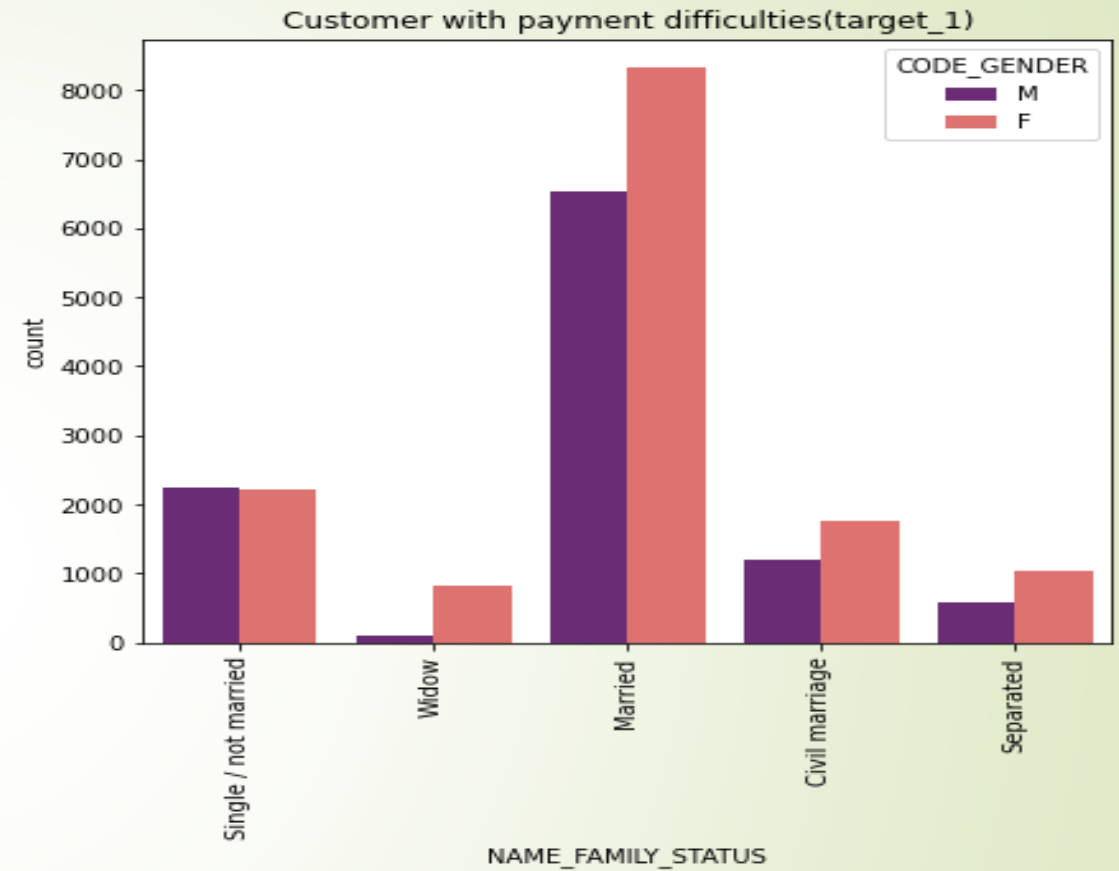
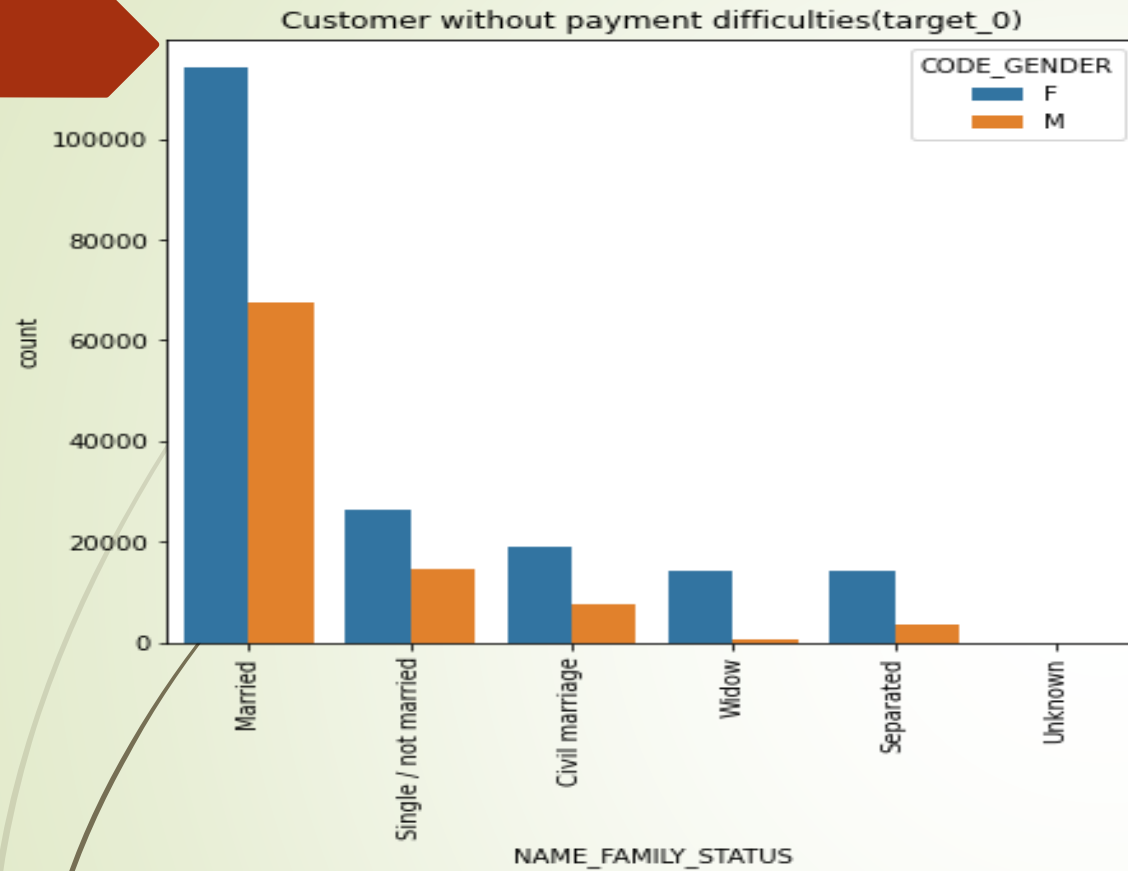
1. Laborers, sales staff and core staff people are defaulters.
2. In labour occupation, male counts is more then female.
3. Secretaries, HR and IT staff counts are very minimal.

Education type count



1. Secondary education people are defaulters.
2. Lower secondary and academic degree people are non-defaulters.
3. Female again leading the count.

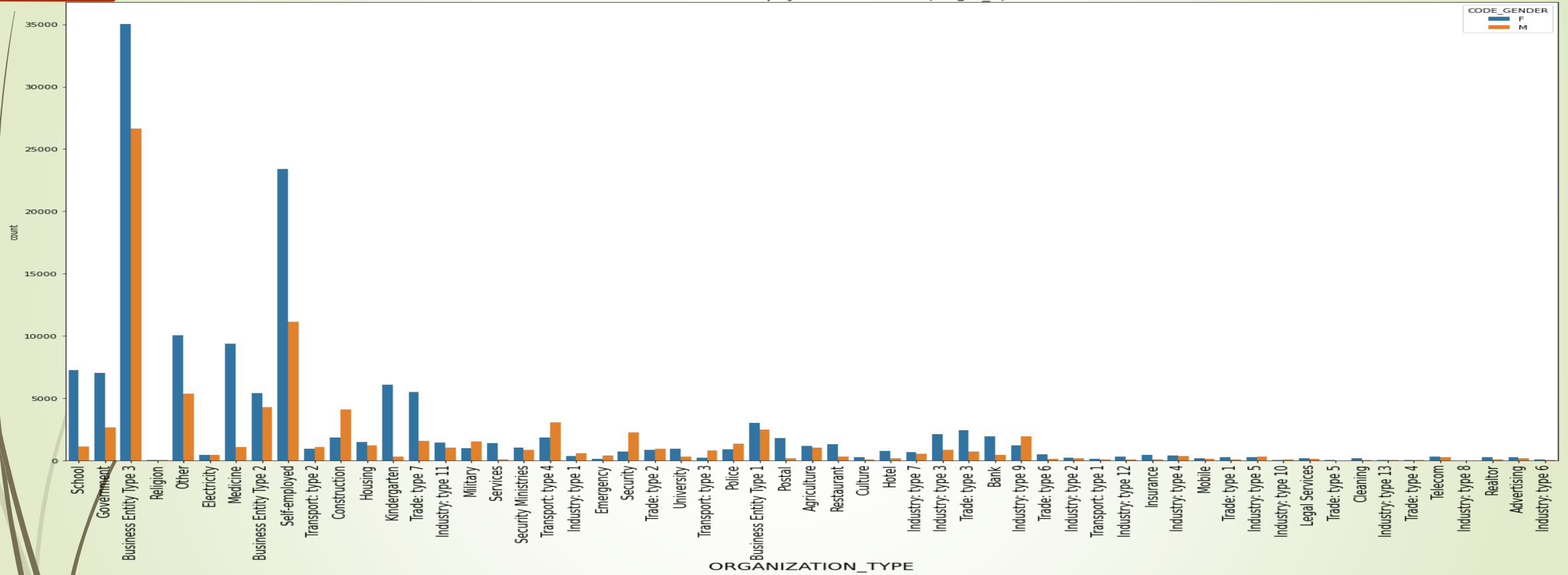
Family status count



1. Married people are defaulters.
2. Separated and widow people are non- defaulters.
3. Females again leading the count.

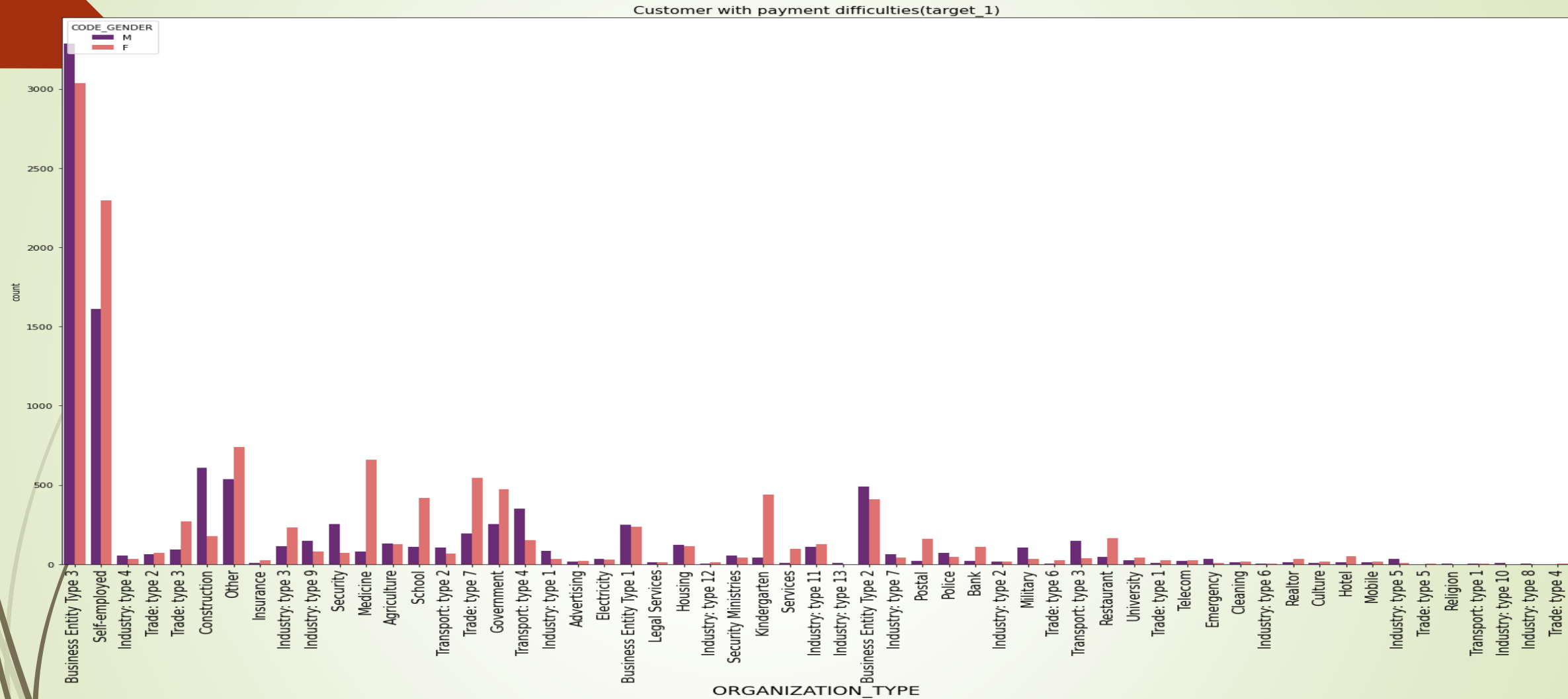
Organisation type non-defaulters

Customer without payment difficulties(target_0)



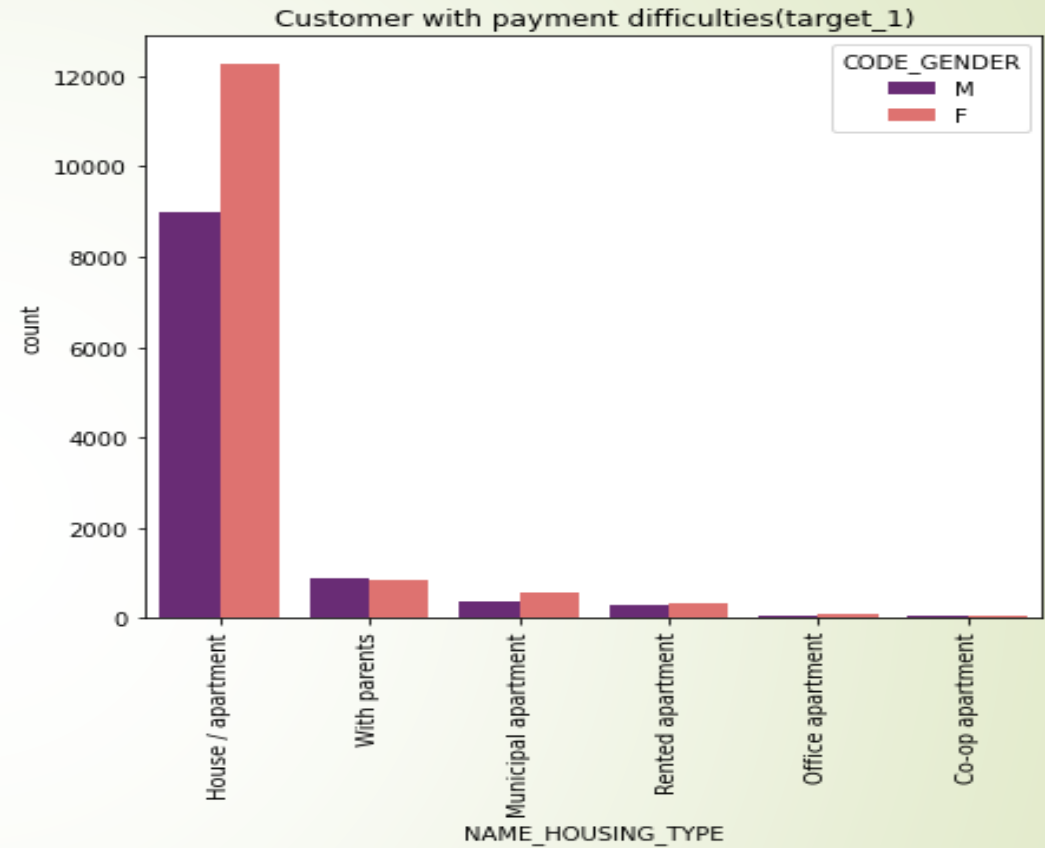
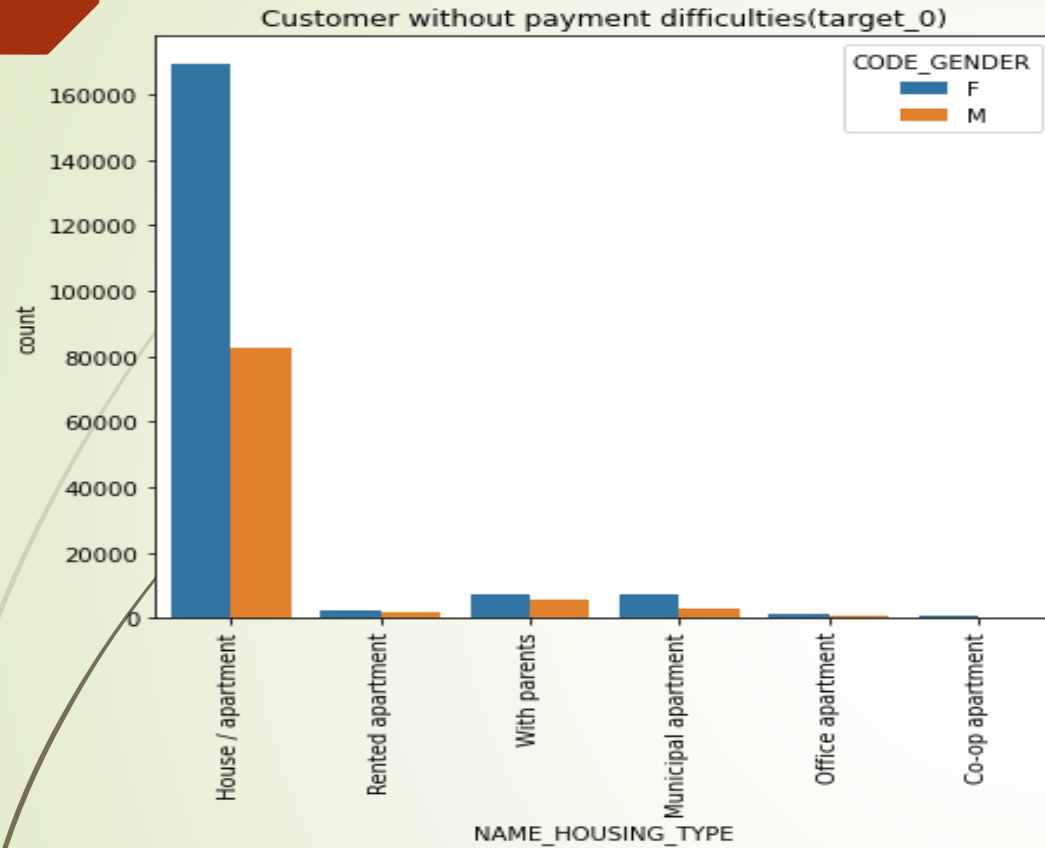
1. Business entity type 3 are non-defaulters.

Organisation type defaulters



1. Business entity type 3 are defaulters.
2. In Business entity type 3 male are defaulters, that's interesting to see.

Housing type count

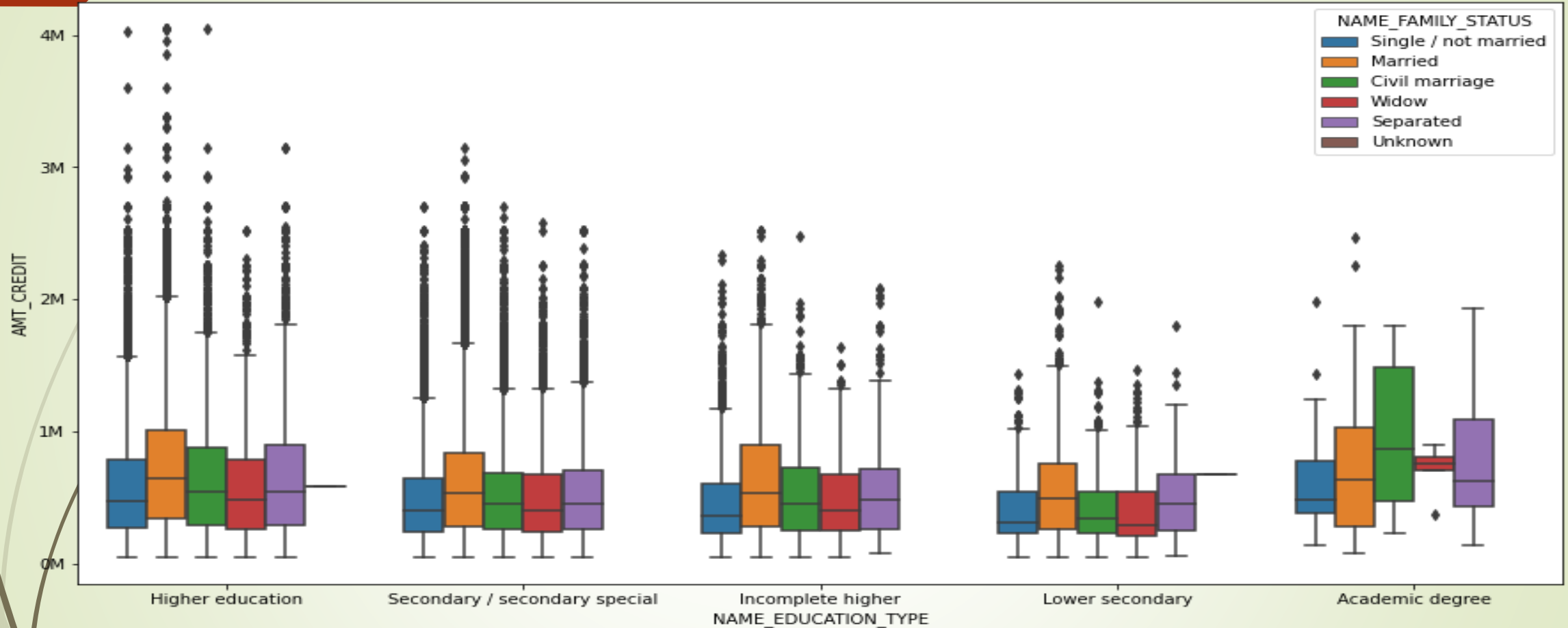


1. Those who have their own house/apartment are defaulters.
2. Females are defaulters.

Bivariate Analysis

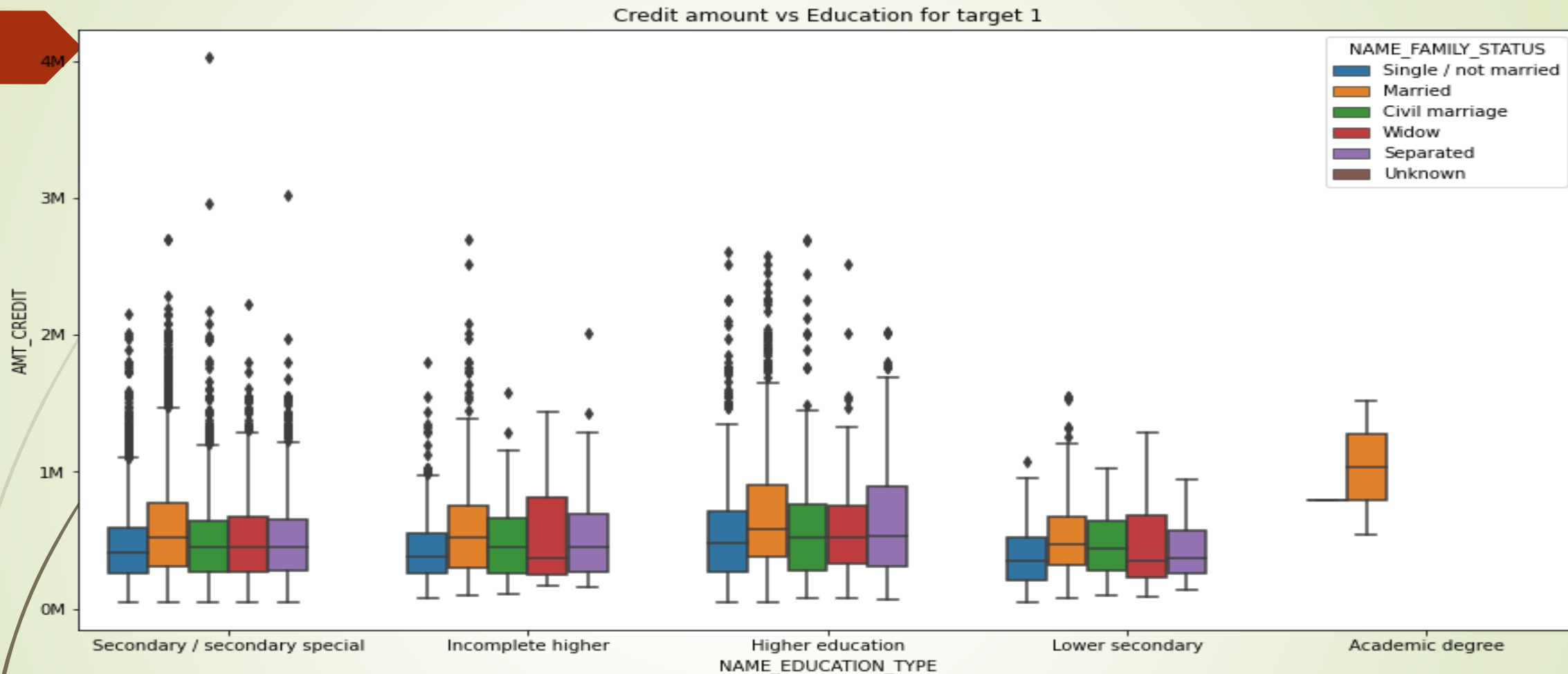
Education type vs credit amount for non-defaulters

Credit amount vs Education for target 0



1. Family status of 'civil marriage', 'married' and 'separated' of Academic degree education are having higher number of credits than others.
2. Also, higher education of family status of 'marriage', 'single' and 'civil marriage' are having more outliers.
3. Civil marriage for Academic degree is having most of the credits in the third quartile.

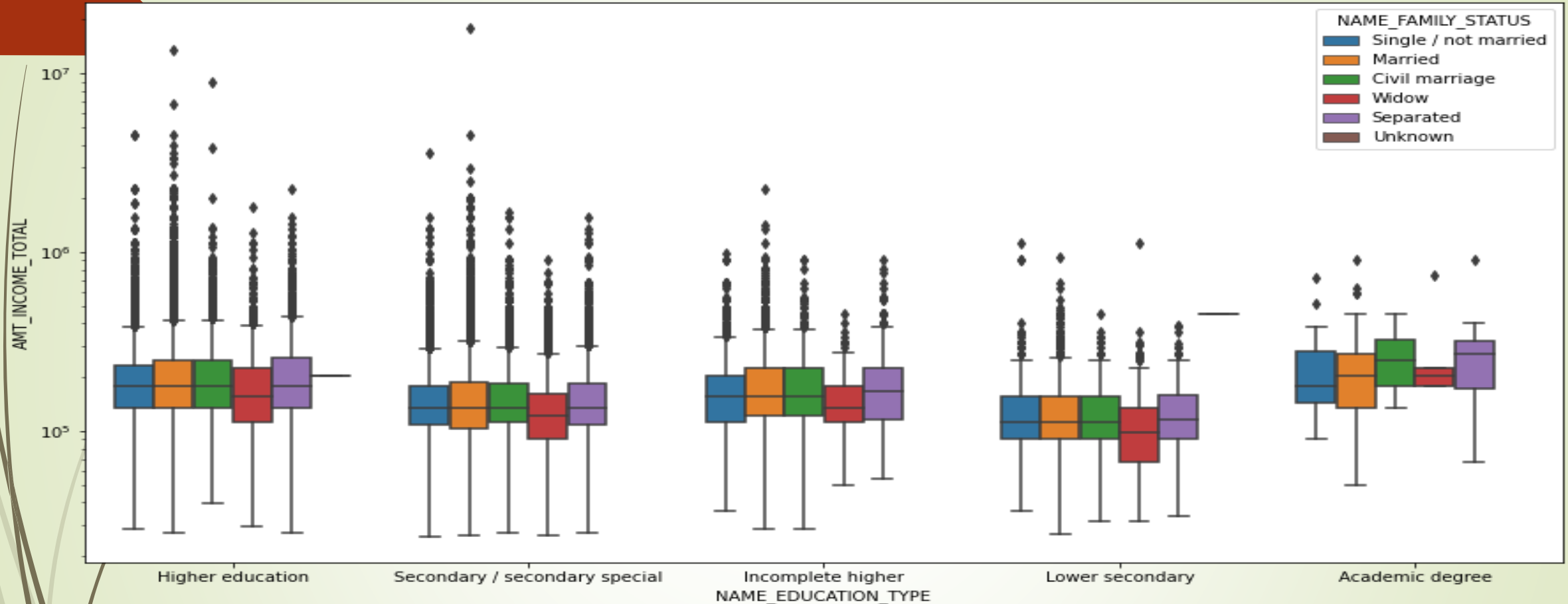
Education type vs credit amount for defaulters



1. Quite similar with Target 0 From the above box plot we can say that Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others.
2. Most of the outliers are from Education type 'Secondary'.
3. In Education type 'Academic degree' only married people are having difficulty in paying loan.

Education type vs Total income for non-defaulters

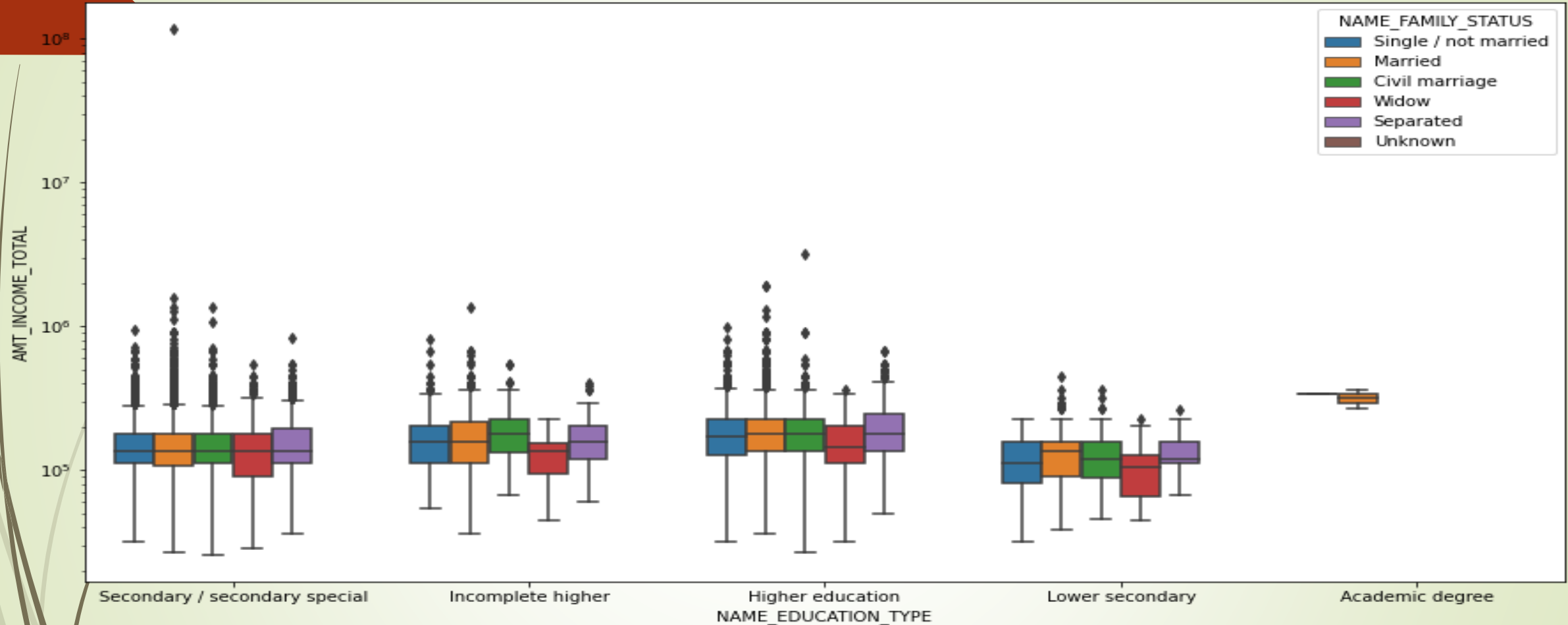
Total Income vs Education for target 0



1. For Education type 'Higher education' the income amount is mostly equal among all family status. It does contain many outliers.
2. Less outlier are having for Academic degree but there income amount is little higher than Higher education.
3. Lower secondary of civil marriage family status have less income amount than others.
4. Except Widows all family status have same income in 'Higher education', 'Secondary', 'Incomplete higher' and 'Lower secondary'.
5. Secondary education also have outliers.

Education type vs Total income for defaulters

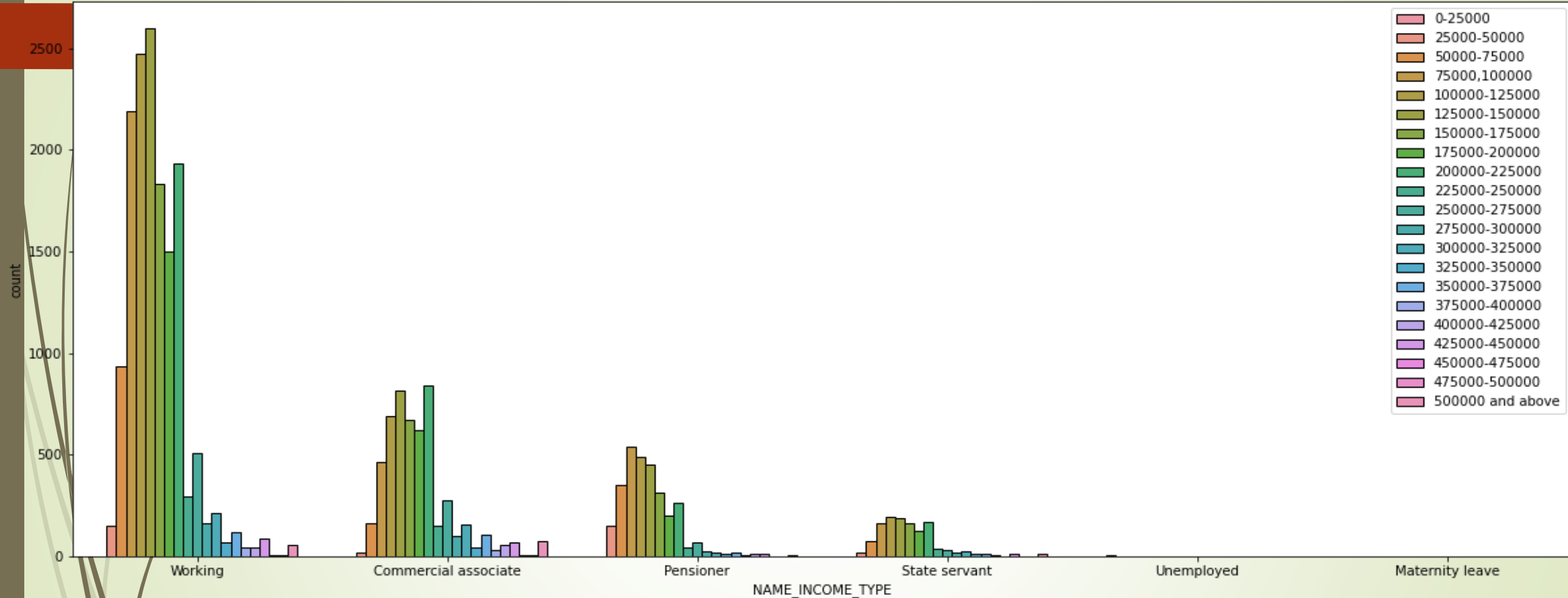
Total Income vs Education for target 1



1. Have some similarity with Target0, for Education type 'Higher education' the income amount is mostly equal with family status.
2. No outlier for Academic degree but its income amount is little higher than Higher education.
3. Lower secondary are having less income amount than others.
4. Except Widows all family status have same income in 'Higher education', 'Incomplete higher' and 'Lower secondary'.
5. Outliers in 'Secondary' and 'Higher education'.

Categorical-Categorical Analysis

Customer with payment difficulties



1.Among the INCOME_TYPE , Working are defaulters.

2.In Working customers, those who are having total income between 50k-200k having more difficulty.
i.e. those who are having **Medium income** are defaulters.

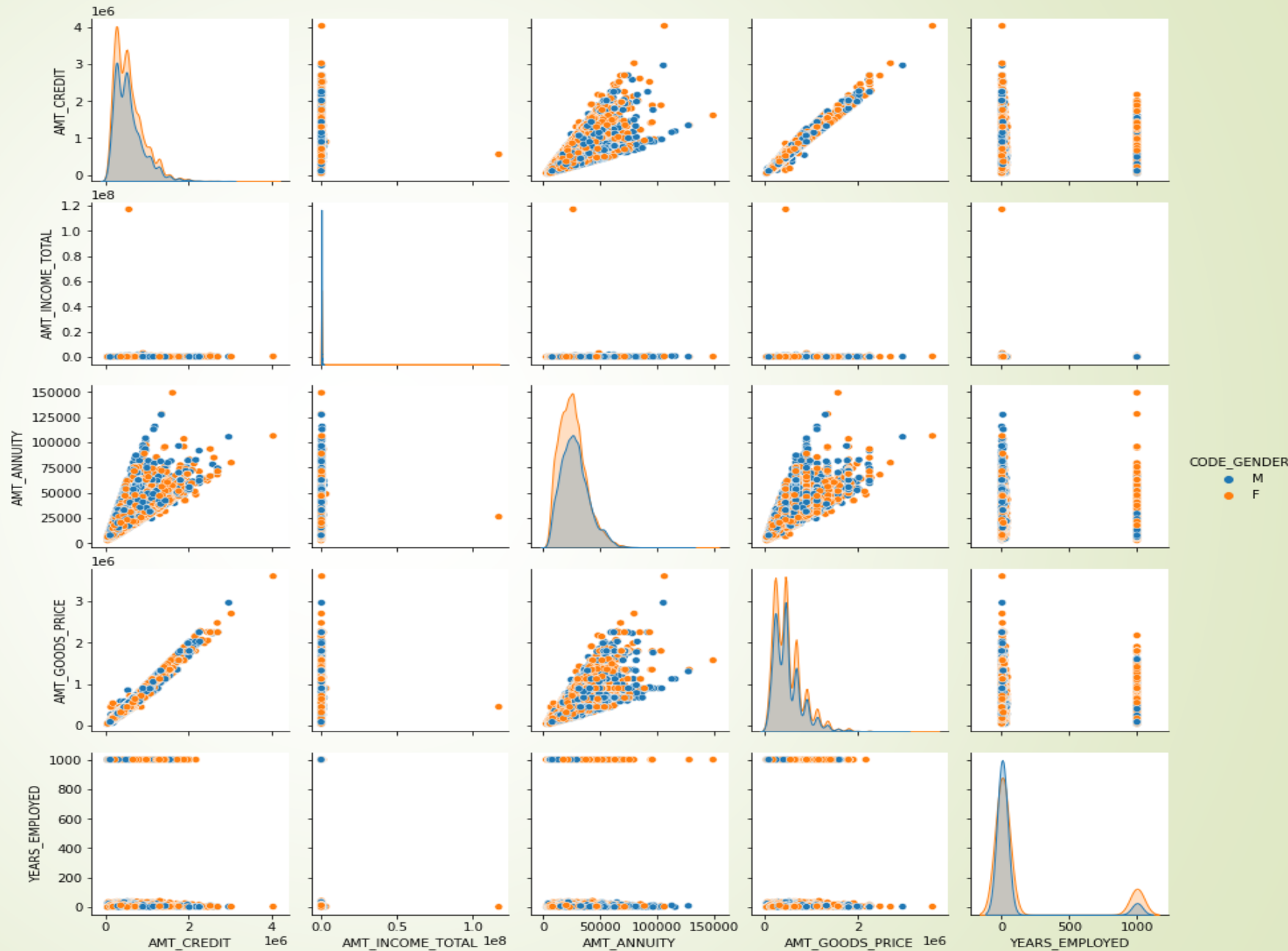
Numerical-Numerical Analysis for non-defaulters

1. Those who have less total income credited all amount from starting to maximum. some people who have very high income credited less, so it may count as an outlier and mostly are males.
2. Loan annuity and credit amount has a linear relationship. Those who have less credit limit has less loan annuity. It has few outliers.
3. Price of goods for which loan is given has a linear relationship with credit amount.
4. In years_employed and amt_credit graph few have 1000 year of employment so that makes no sense in real life. That is outliers for sure. Mostly females have credited the loan.

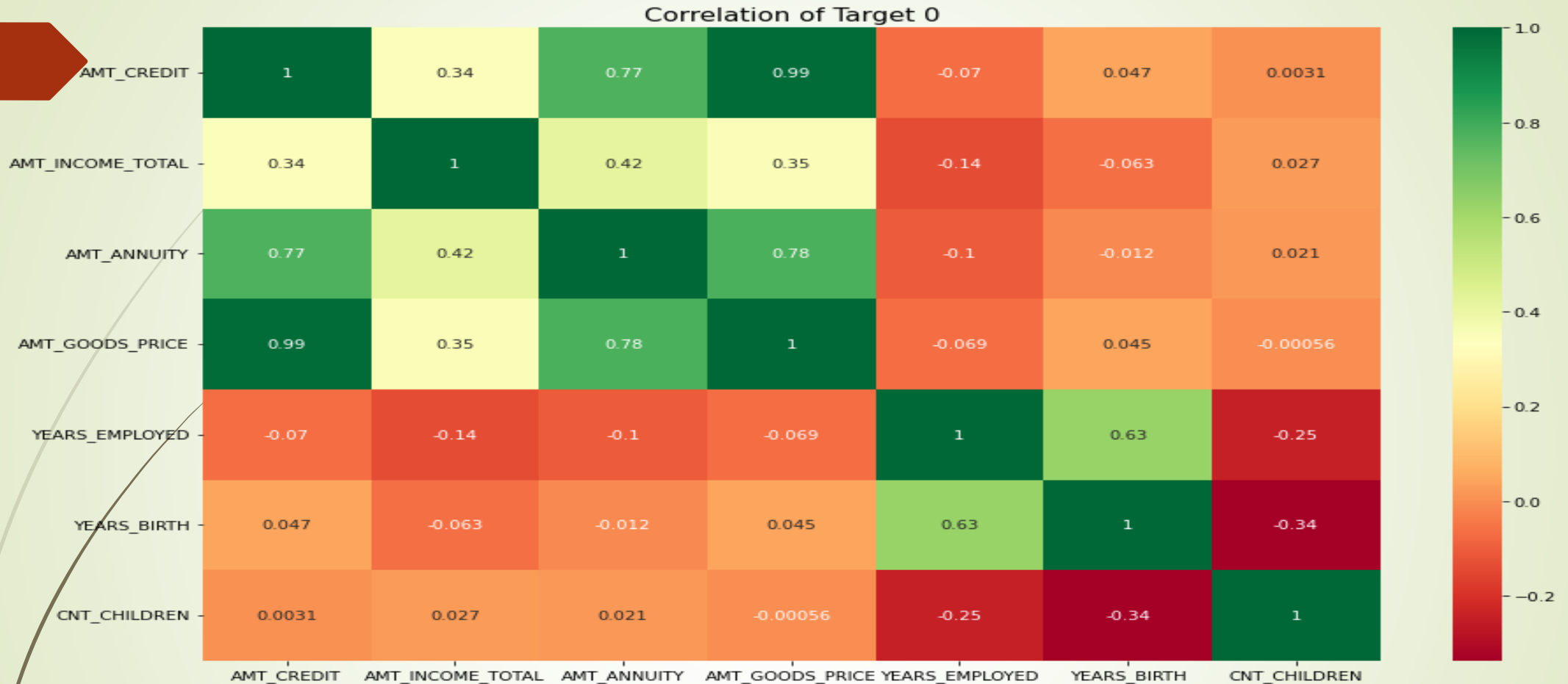


Numerical-Numerical Analysis for defaulters

1. Same as target 0, those who have less total income credited all amount from starting to maximum. one customer who has very high income credited less, so it may count as an outlier.
2. Loan annuity and credit amount has a linear relationship. Those who have less credit limit has less loan annuity. It has few outliers.
3. Price of goods for which loan is given has a linear relationship with credit amount.
4. In years_employed and amt_credit graph few has 1000 year of employment so that makes no sense in real life. That is outliers for sure. Mostly females have credited the loan.



Multivariate Analysis



1. As we have seen earlier in bivariate analysis, amount annuity and credit amount has a good linear relationship.
2. Price of goods for which loan is given has a good linear relationship with credit amount.
3. years_employed and amt_credit dose not have a good linear relationship.
4. Loan annuity and goods price has a good linear relationship.
5. Credit amount having inverse linear relationship with number of children client have, means Credit amount is higher for less children count client have and vice-versa.
6. Income amount having inverse linear relationship with number of children client have, means more income for less children client have and vice-versa.

Multivariate Analysis

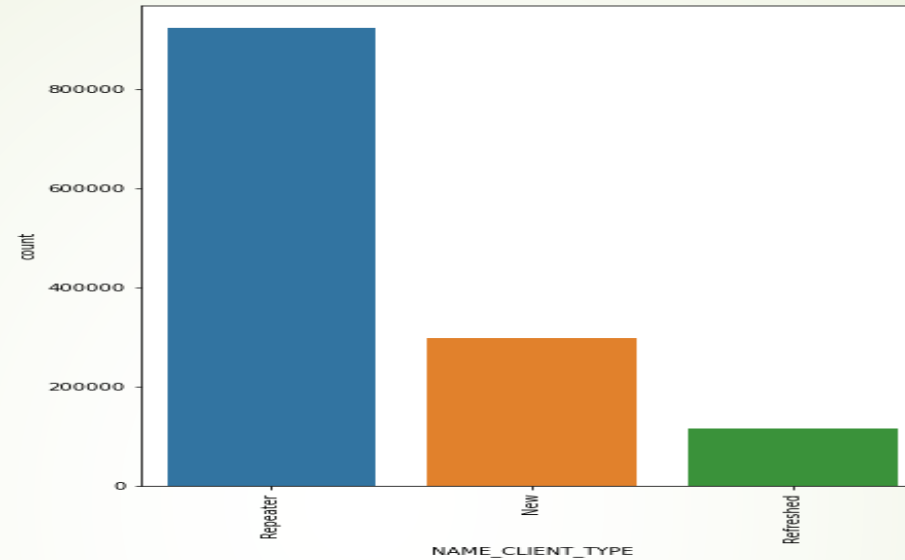
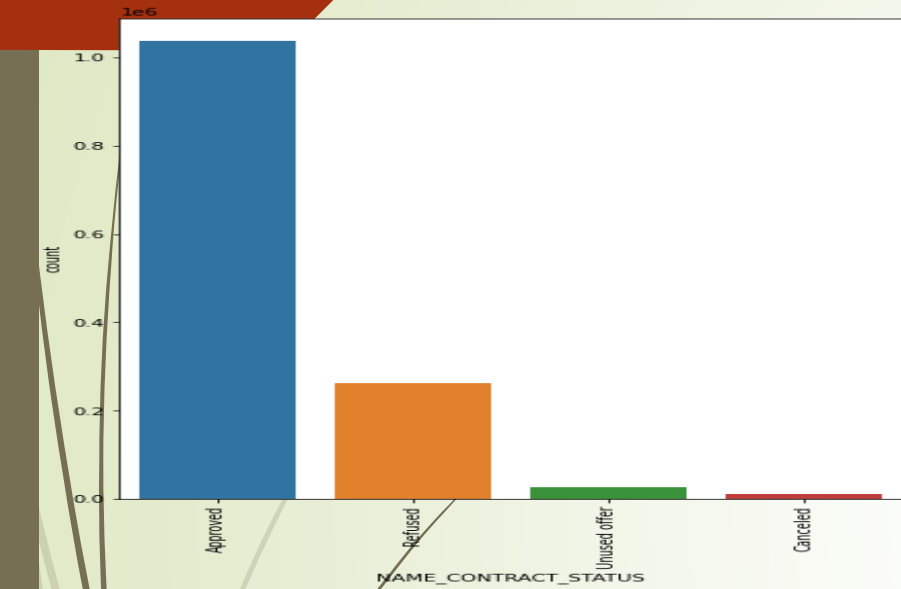
Correlation of Target 1



1. As we have seen earlier in bivariate analysis, amount annuity and credit amount has a good linear relationship.
2. Price of goods for which loan is given has a good linear relationship with credit amount.
3. years_employed and amt_credit dose not have a good linear relationship.
4. loan annuity and goods price has a good linear relationship.
5. Here total income with loan annuity & goods price does not have a good linear relationship.
6. Credit amount having inverse linear relationship with number of children client have, means Credit amount is higher for less children count client have and vice-versa.
7. Income amount having inverse linear relationship with number of children client have, means more income for less children client have and vice-versa.

Previous Application Dataset

Univariate Analysis(Categorical variable)

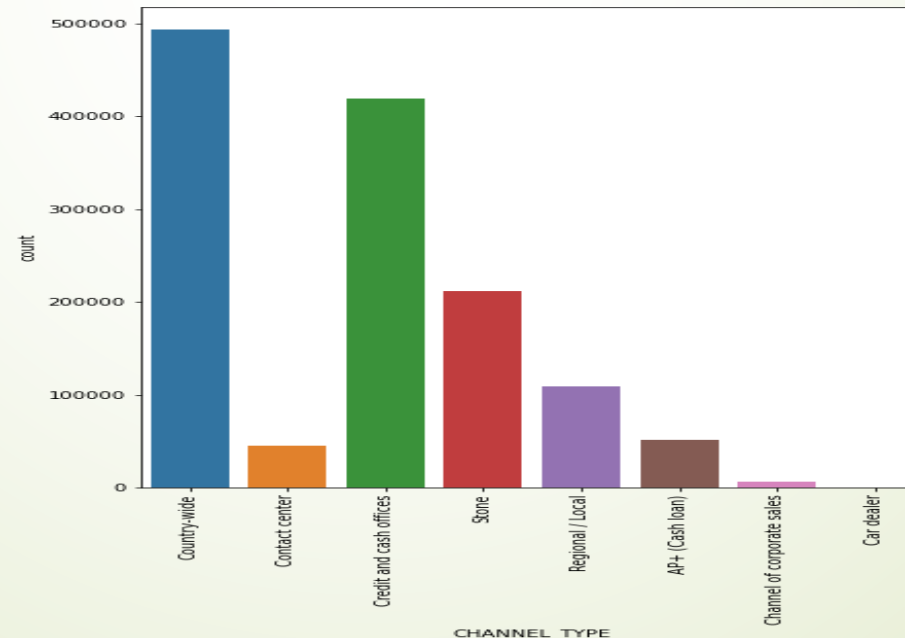
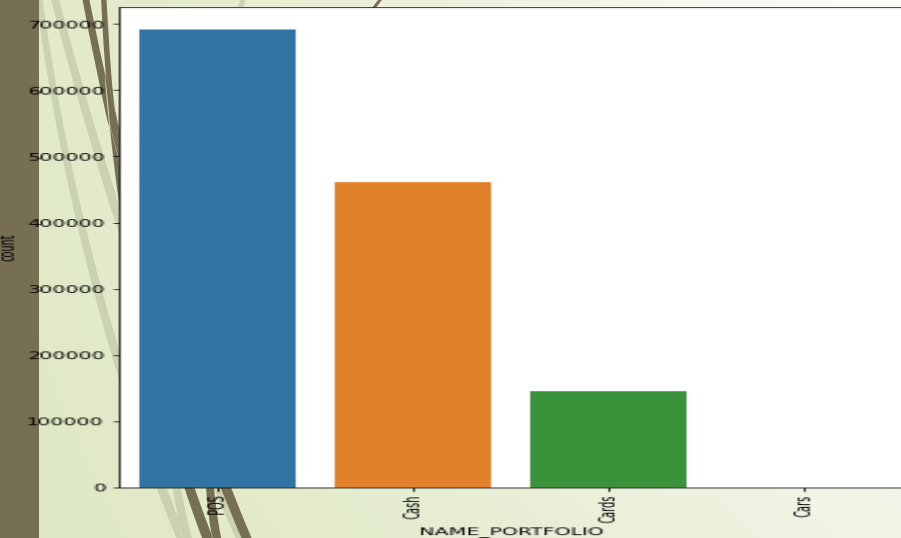


1. Approved loan status is huge than rejected or canceled.

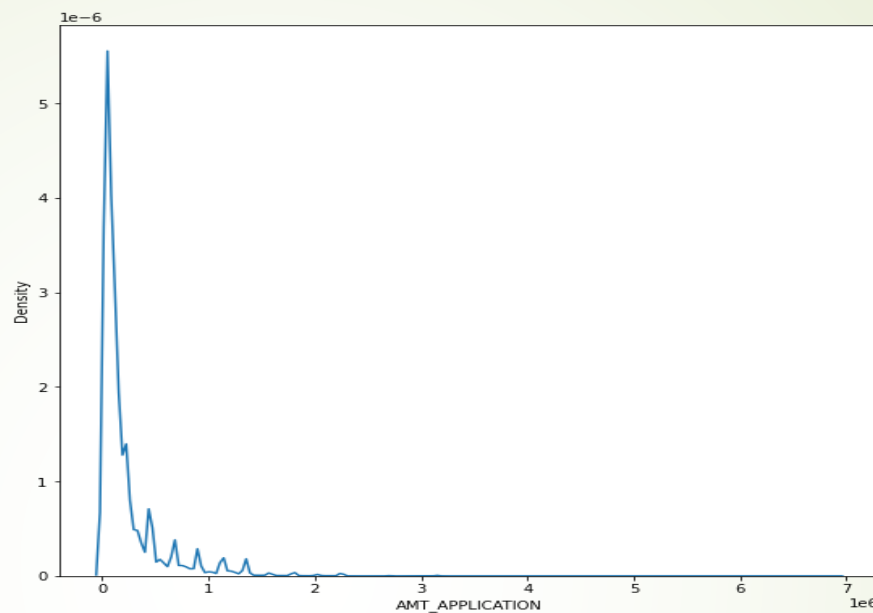
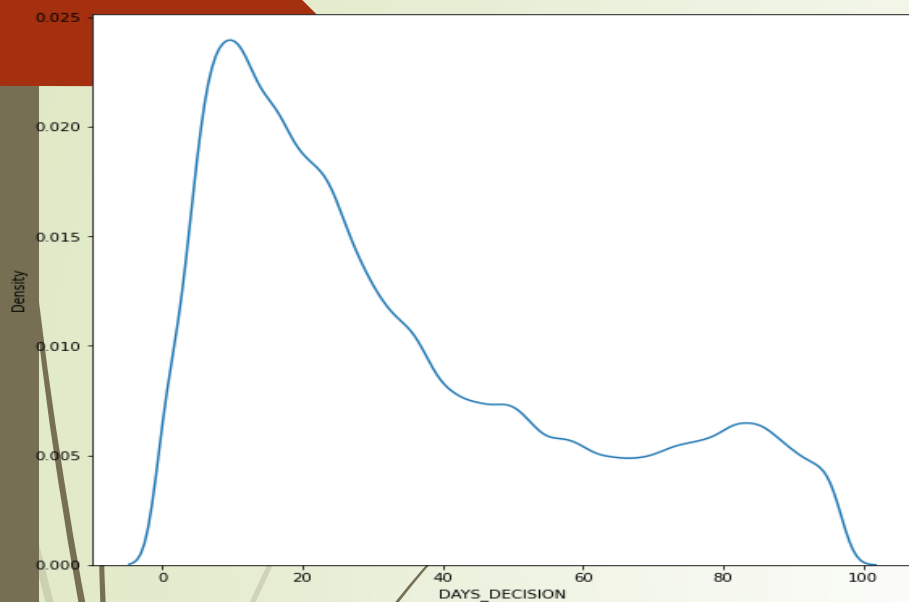
2. Repeater clients are highest in number than new client.

3. POS loans are highest rather than cash loans.

4. Country-wide channel type is the most used channel followed by Credit and cash offices.



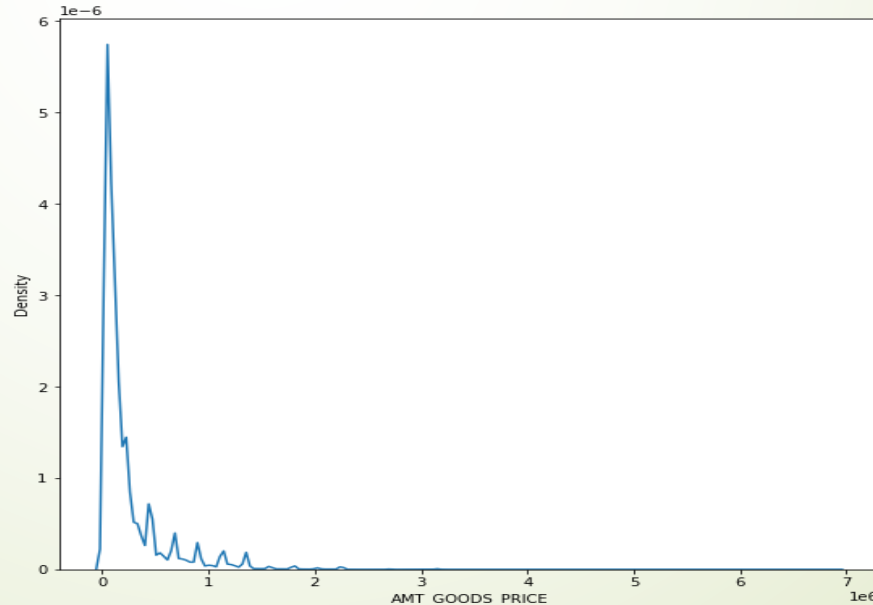
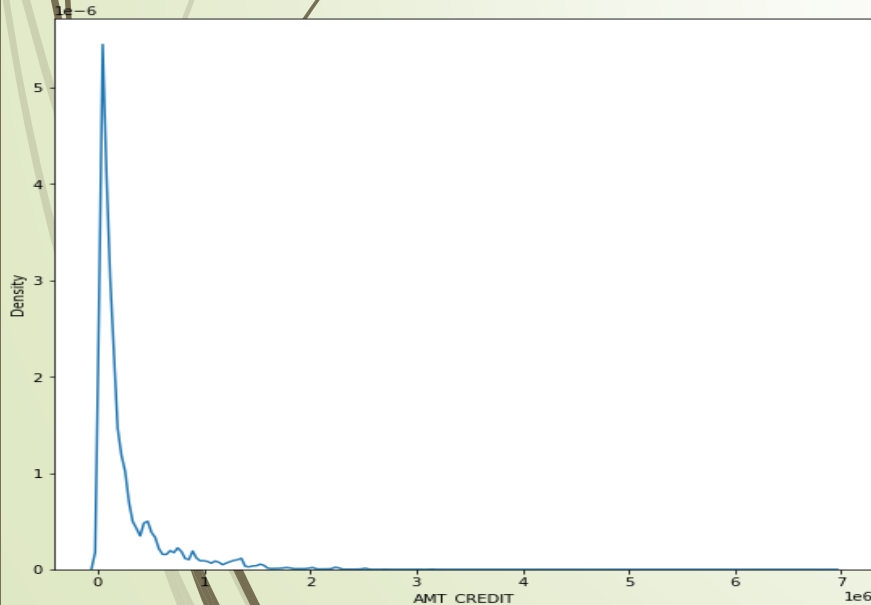
Univariate Analysis(Numerical variable)



1. Most of the applications decision took around 10 to 30 months.

2. Most of the loan application amount were below 500000, we can see a huge spike around 100000 amount.

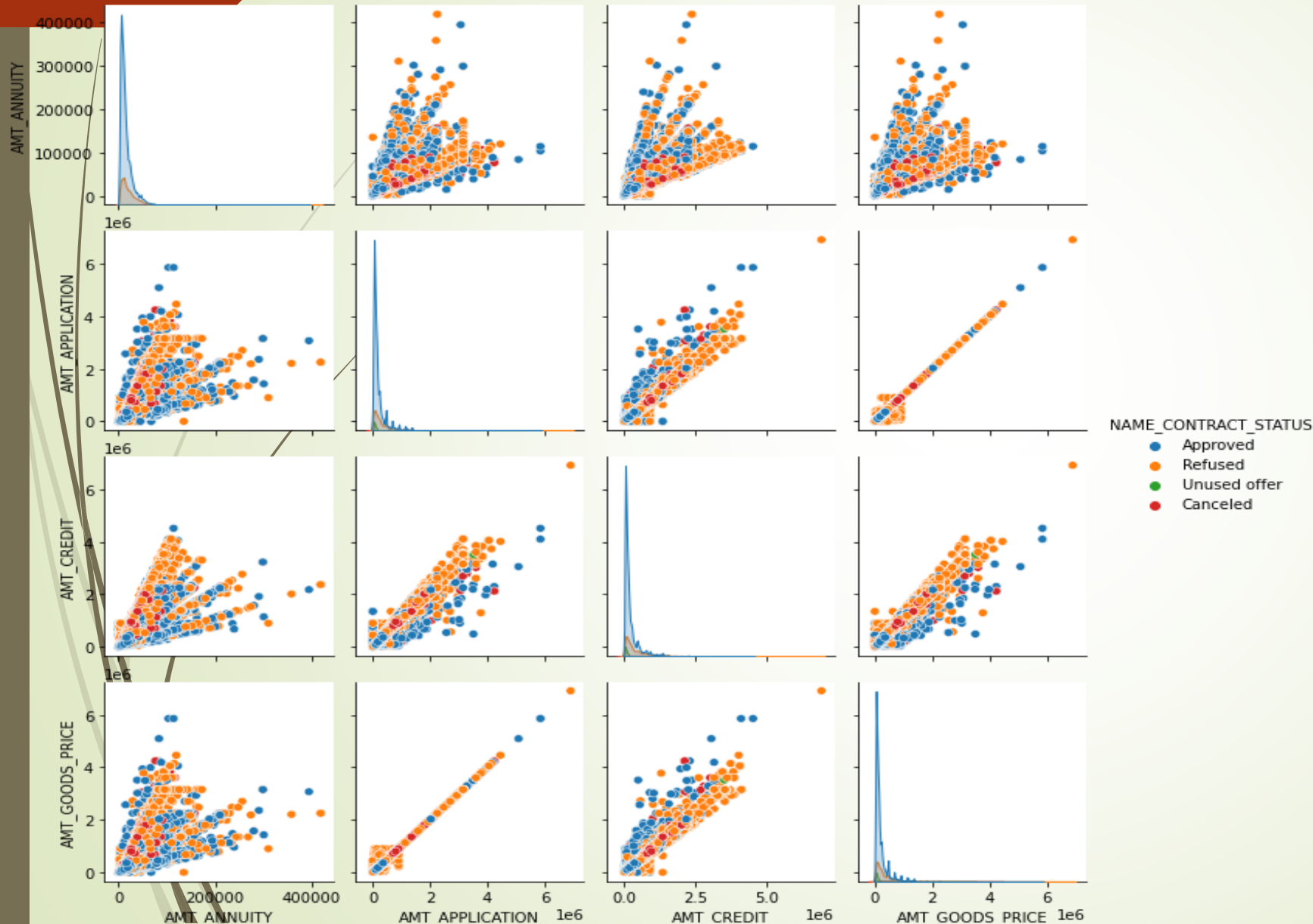
3. Amount credited, is also following the pattern of loan application. We already saw that most of the application was approved in previous plots.



4. Amount of the goods price is also following the same distribution like application amount and amount credited. Because, based on the price of the goods, the loan was approved and amount was credited.

Bivariate Analysis

Numerical-Numerical Analysis



1. Annuity of previous application has a very high and positive correlation over:

(a) How much credit did client asked on the previous application.

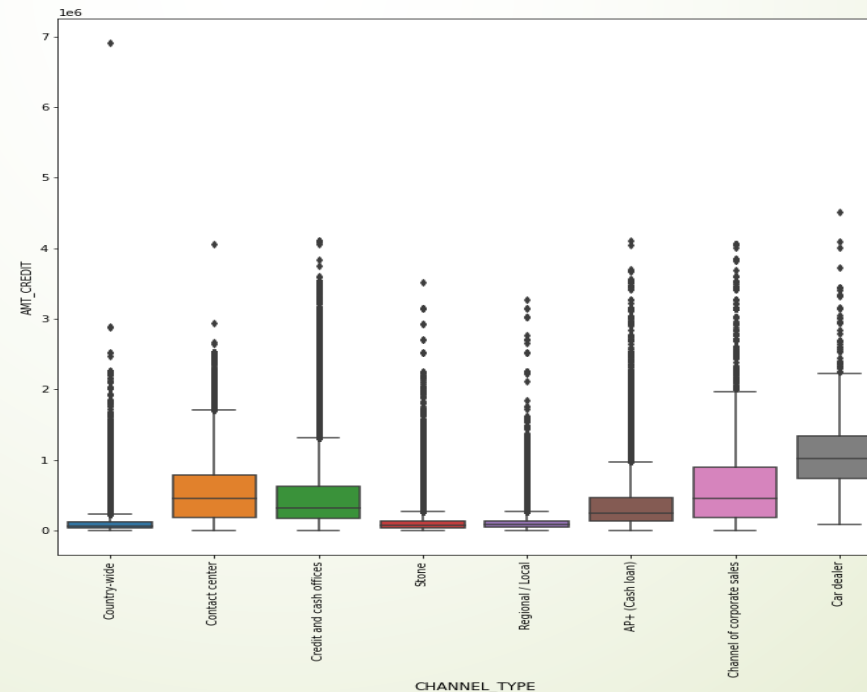
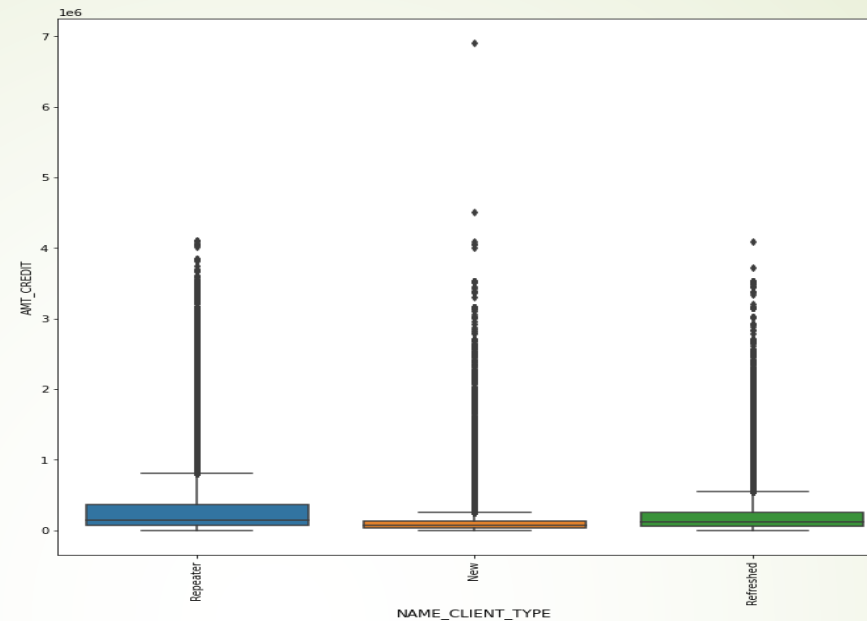
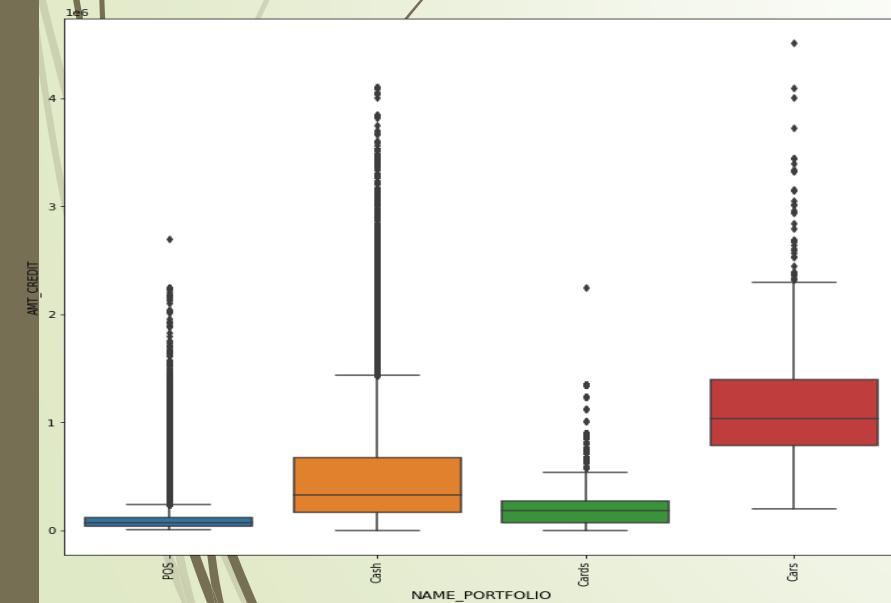
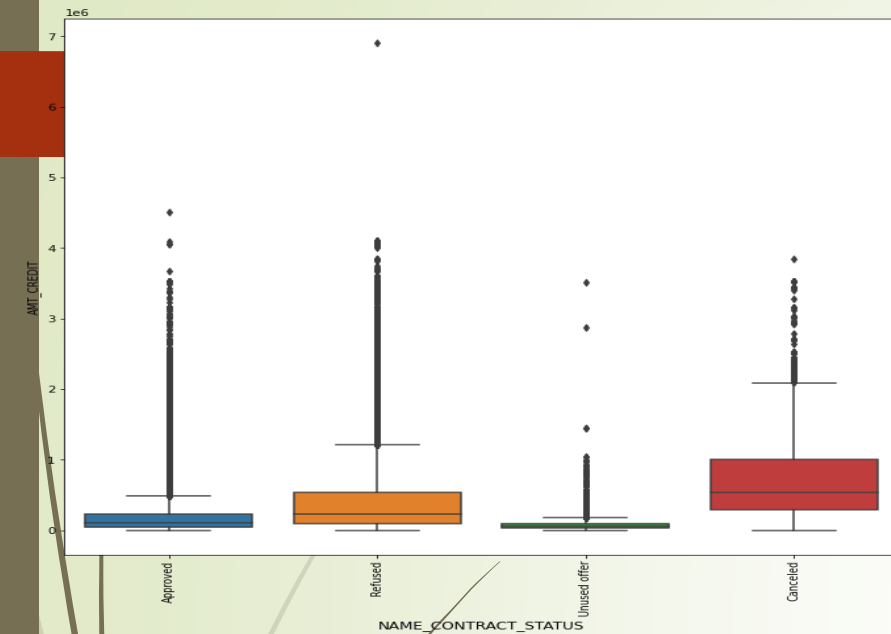
(b) Final credit amount on the previous application that was approved by the bank.

(c) Goods price of good that client asked for on the previous application.

2. For how much credit did client ask on the previous application is highly influenced by the Goods price of good that client has asked for on the previous application. High Credit loans are most likely to be refused.

3. Final credit amount disbursed to the customer previously, after approval is highly influence by the application amount and also the goods price of good that client asked for on the previous application.

Categorical-Numerical Analysis



1. Most of the amount credit was cancelled in contract status. Refused and Unused has an outlier.

2. Repeater client got more loan credit. New client has an outlier.

3. Cash loan got more credited. Cars and cards portfolio has some outliers.

4. Through the contact center and corporate sales channel, more loan got credited. Country wise channel has an outlier.

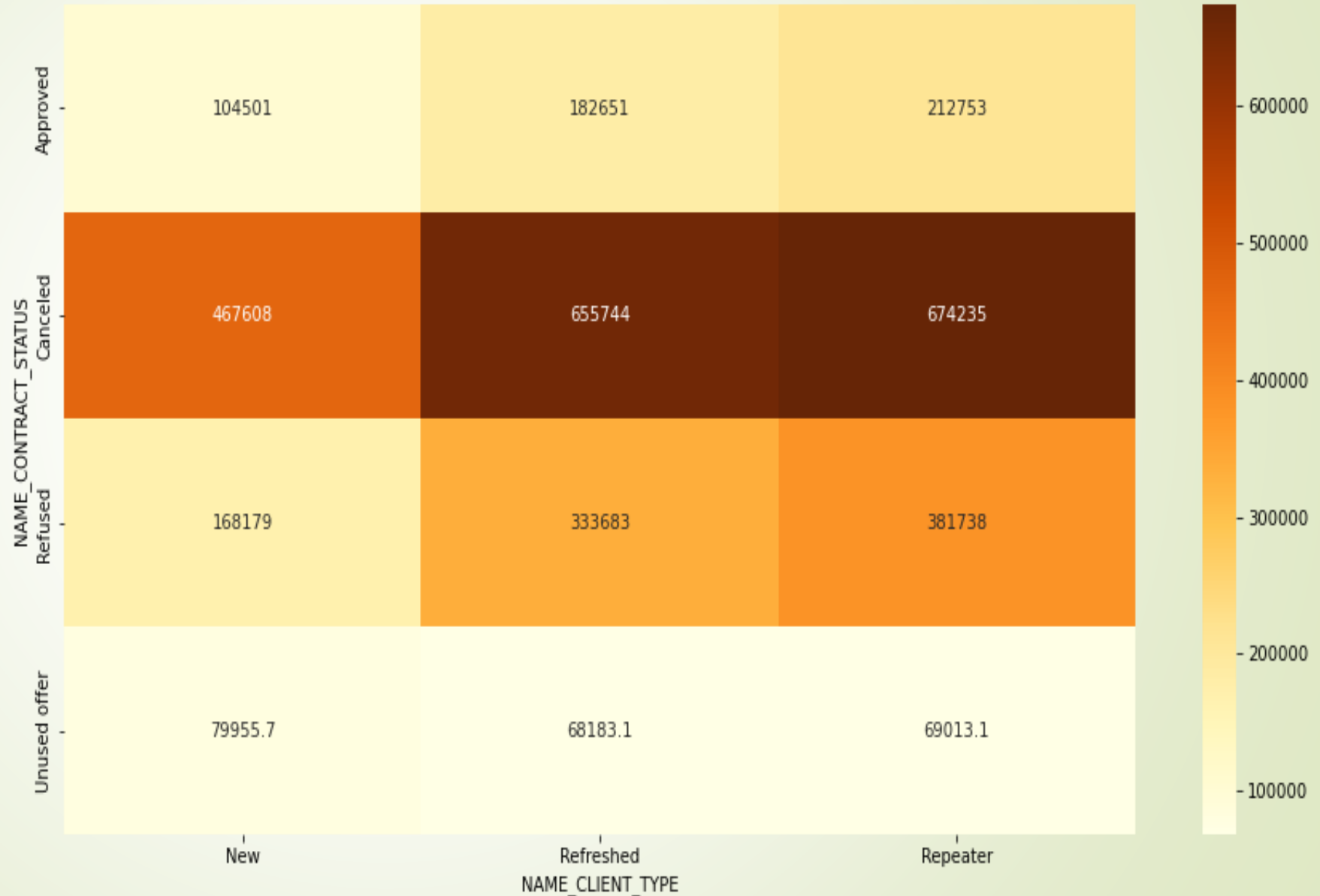
Multivariate Analysis

contract status vs name client type aggregating over application amount

1.Unused offer application amount is low

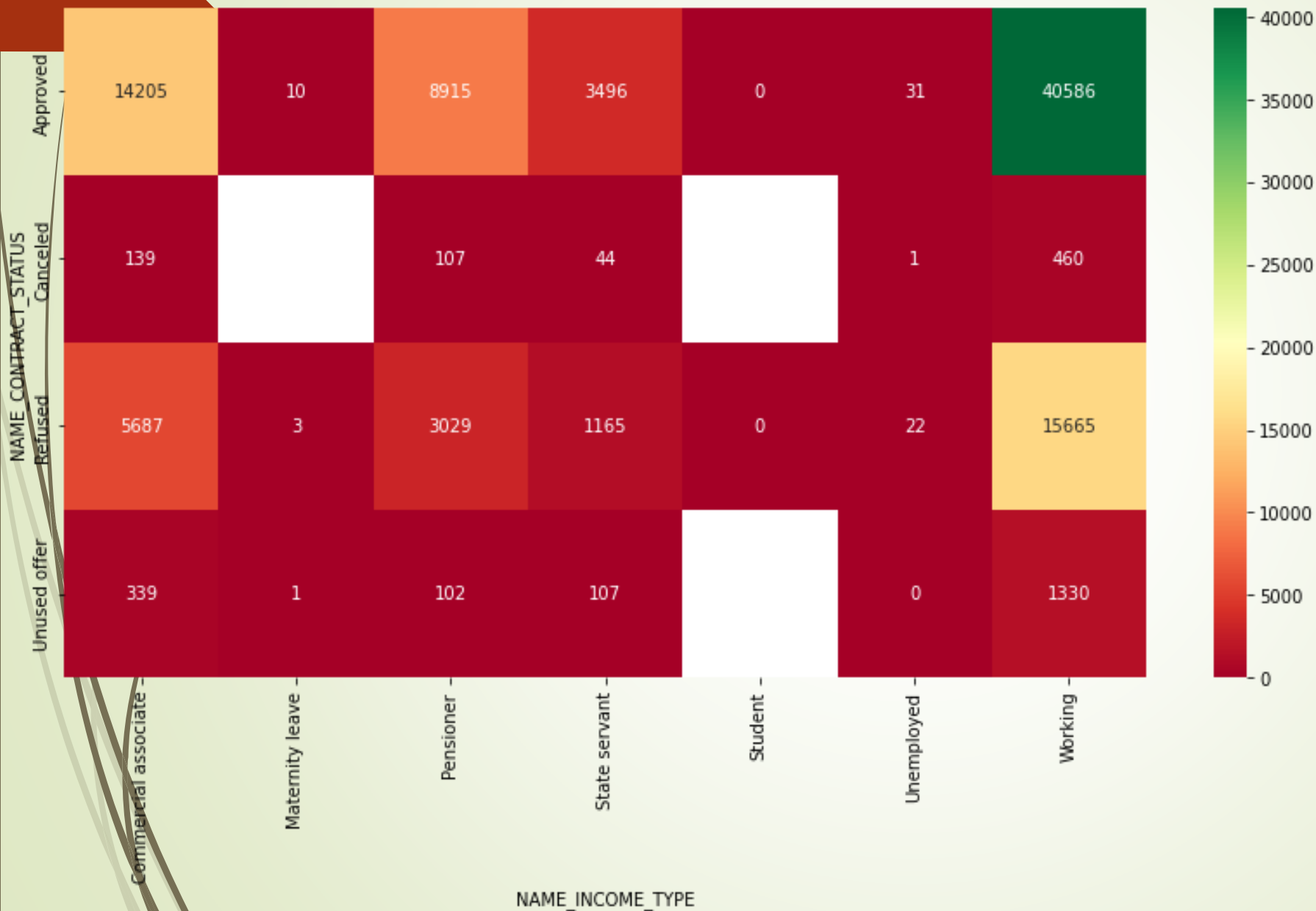
2.Cancelled application amount is high. It means more client cancelled the loan application.

3.Repeater's application amount is highest than the New customers. Maybe bank has more policies or rate of interest etc for repeat applicants.



Analysis of Merging Dataset

Income type vs contract status defaulters



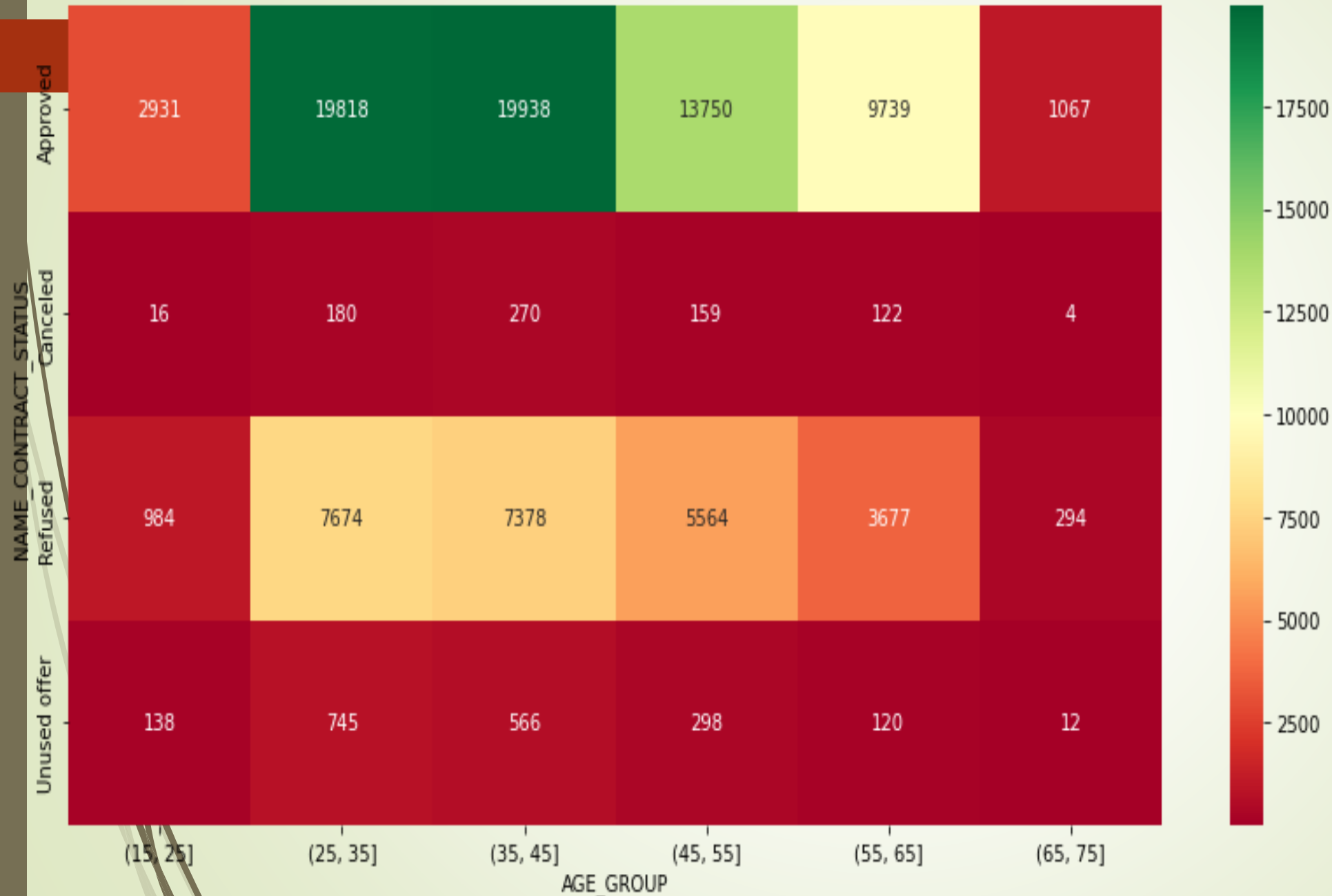
1. Since Target 1 is default, higher on the above matrix shows correlation to default.

2. Working applicant with Approved status have defaulted in highest numbers

3. Previous applications with Refused, Cancelled, Unused loans also have default which is a matter of concern. This indicates that the financial company had Refused/cancelled previous application, but has approved the current, and is facing default on these loans.

4. 15,665 applicants of working class were Refused earlier and now have defaulted.

Age group vs contract status defaulters



1. Since Target 1 is default, higher on the above matrix shows correlation to default.

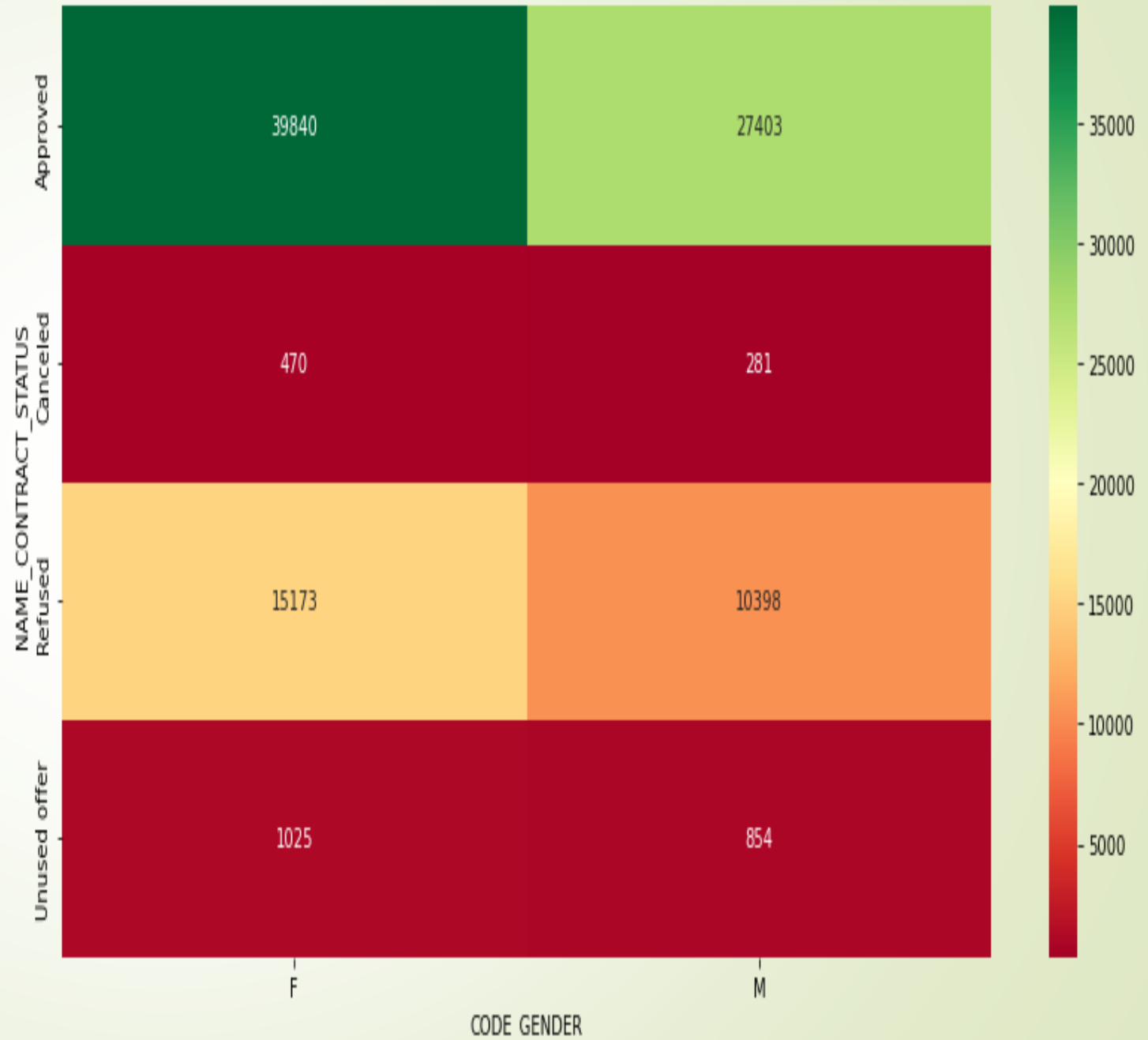
2. Approved loans of age group 25-35 and 35-45 have higher defaults

3. Refused, cancelled, Unused loans in previous application have defaulted in current.

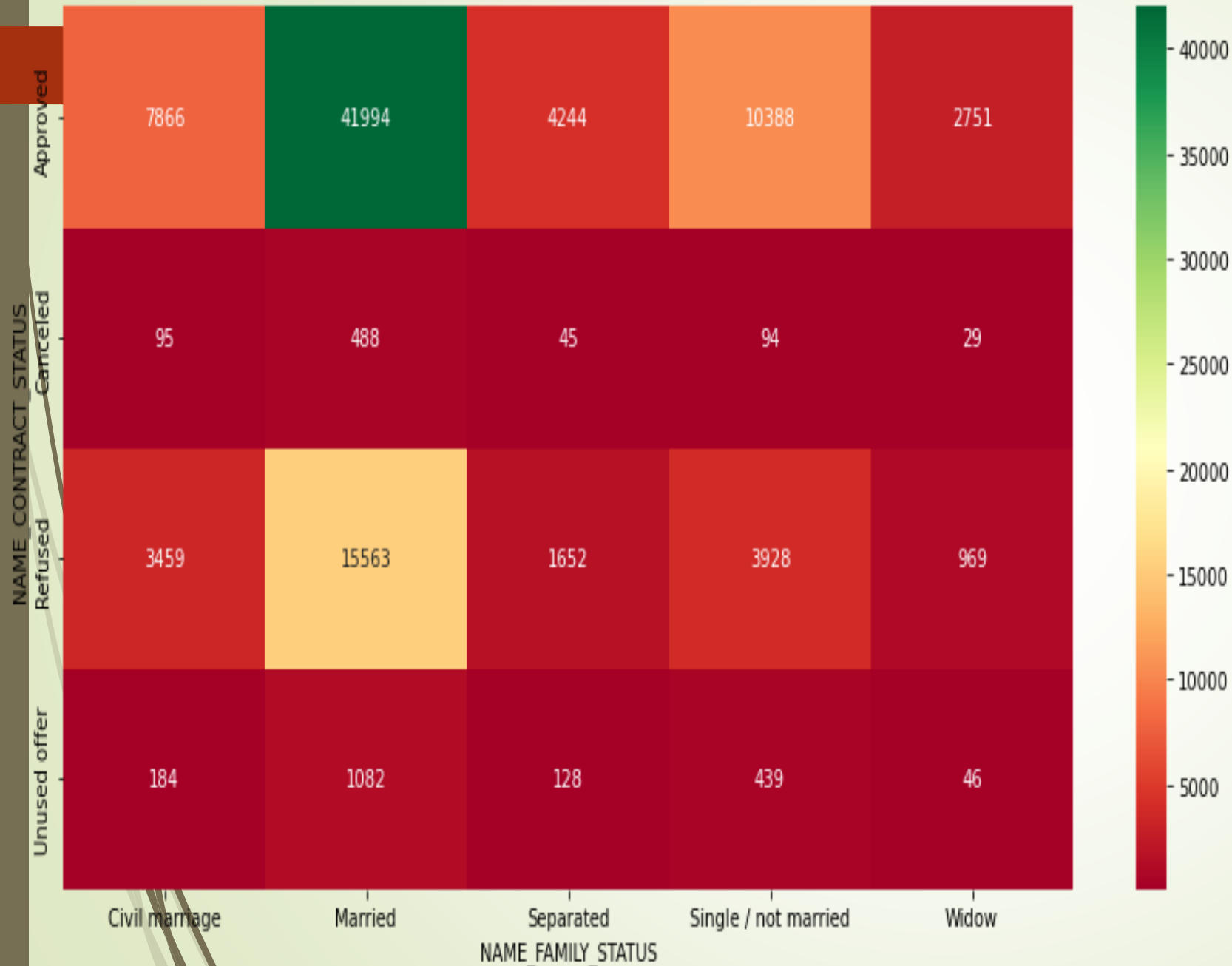
4. Age group 15-25 and 65-75 have the less number of defaulters.

Gender vs contract status defaulters

1. Approved loans of both gender have higher defaults.
2. Cancelled loans in previous application have defaulted in current.
3. Females are more defaulter than males.



Family status vs contract status defaulters



1. Those whose loan is **Approved** and **Married** have higher defaults.

2. **Cancelled** and **Unused** loans in previous application have defaulted in current. This we need to check.

3. **Widows** are defaulted lesser in number compare to all.

Conclusions

Females are having higher rate in both defaulters and non-defaulters. So bank should focus more giving loan to Female gender.

- Banks should focus less on Income type Working category, as they are the highest defaulters.
- Those having Medium income range(50k-200k) are defaulters. Bank should keep eye on them while giving loans.
- In Occupation type - Labourers, Salesman, Drivers are defaulters.
- Bank should focus more on Business type 3 in Organization type, as they have higher rate in both defaulters and non-defaulters.
- In Age group , 25-35 yr old followed by 35-45 are defaulters but bank approved their previous loan.
- Secondary education people are defaulters.
- In Client type, Repeaters client are defaulters but bank approved their previous loan showed bank are giving special offers/rate of interest to them.
- In NAME_FAMILY_STATUS, Married followed by Single/not married has high rate in defaulters and non-defaulters.
- Income amount is inversely proportional to the number of children client have, means more income if client has less children and vice-versa.
- Attract as much clients from housing type with parents as they are having less number of un-successful attempts.
- Price of goods for which loan is given has a good linear relationship with credit amount.