CREDIT - EDA ASSIGNMENT

BY - VIKHYAT NEGI

Introduction

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

BUSINESS UNDERSTANDING-I

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 as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

- *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.*

BUSINESS UNDERSTANDING-2

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

1. *Approved: The Company has approved loan Application*

2. *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.*

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

4. *Unused offer: Loan has been cancelled by the client but on different stages of the process.*

In this assignment, we will use EDA to understand how consumer attributes and loan attributes influence the tendency of 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 utilize 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).*

DATA UNDERSTANDING

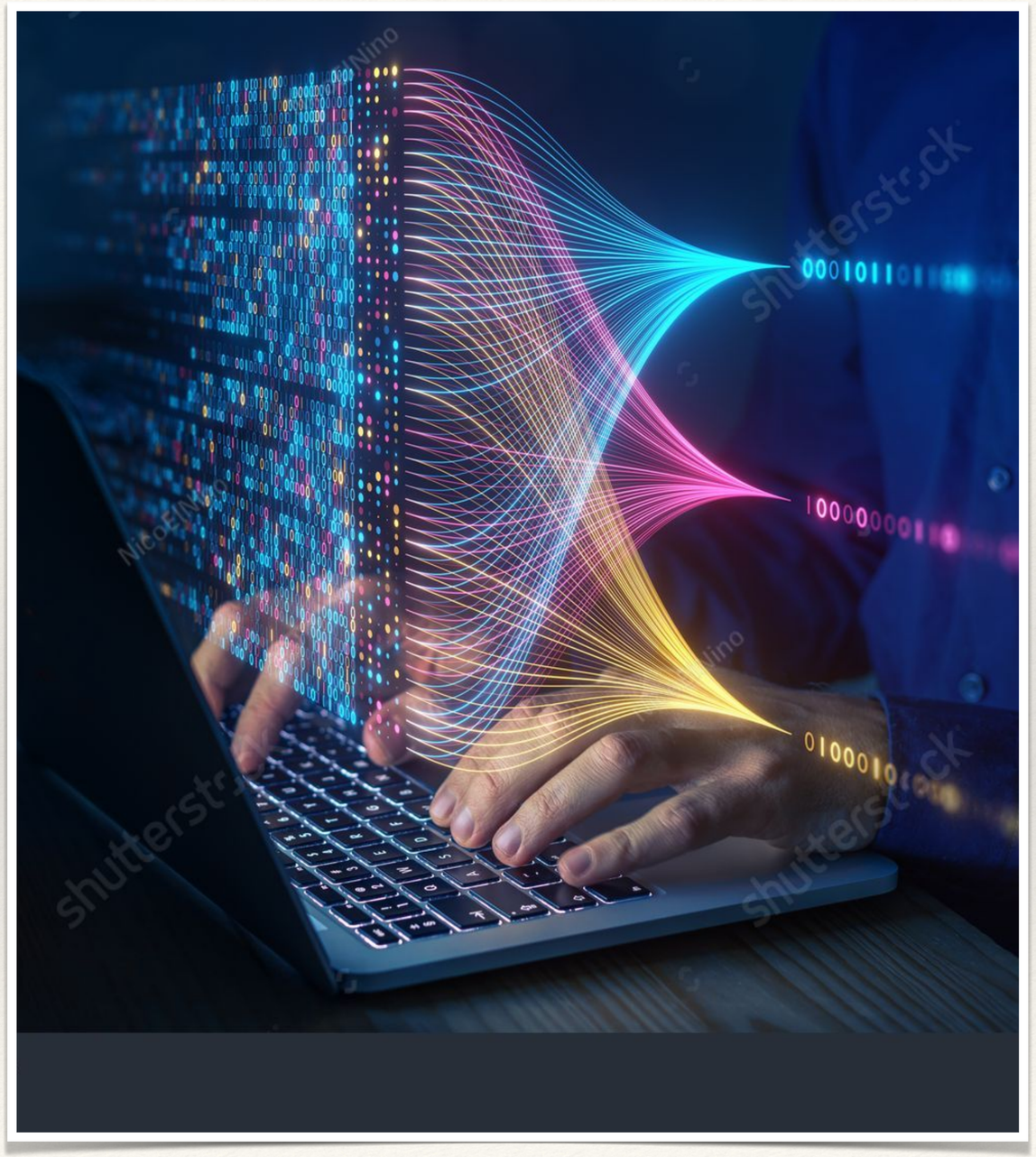
This dataset has 3 files as explained below:

- *1. 'application_data.csv' contains all the information of the client at the time of application.*

The data is about whether a client has payment difficulties.

- *2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.*
- *3. 'columns_description.csv' is data dictionary which describes the meaning of the variables.*

ANALYSIS

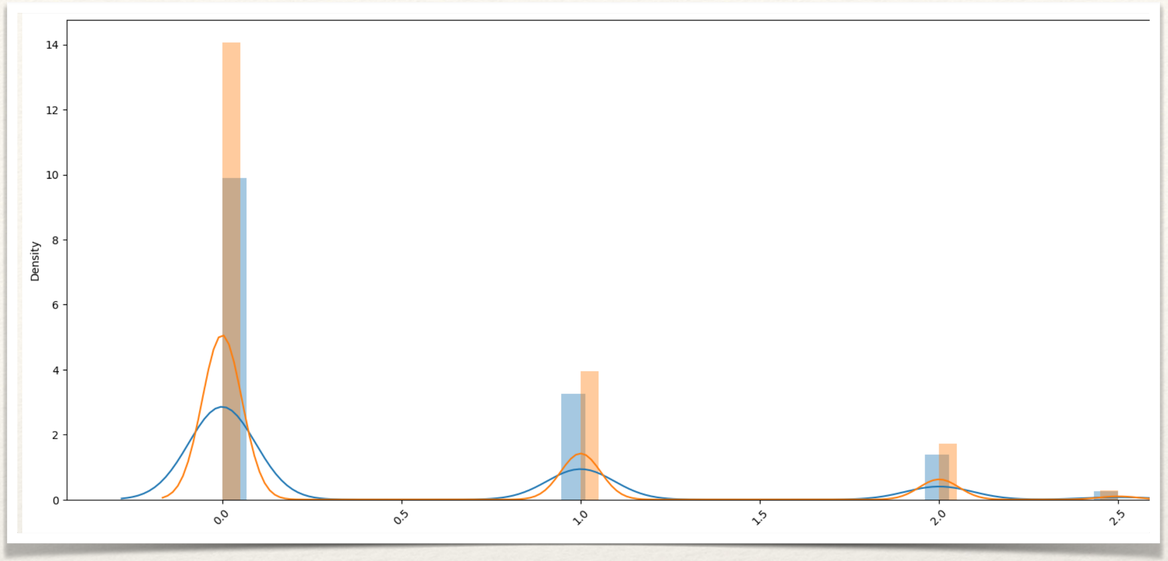


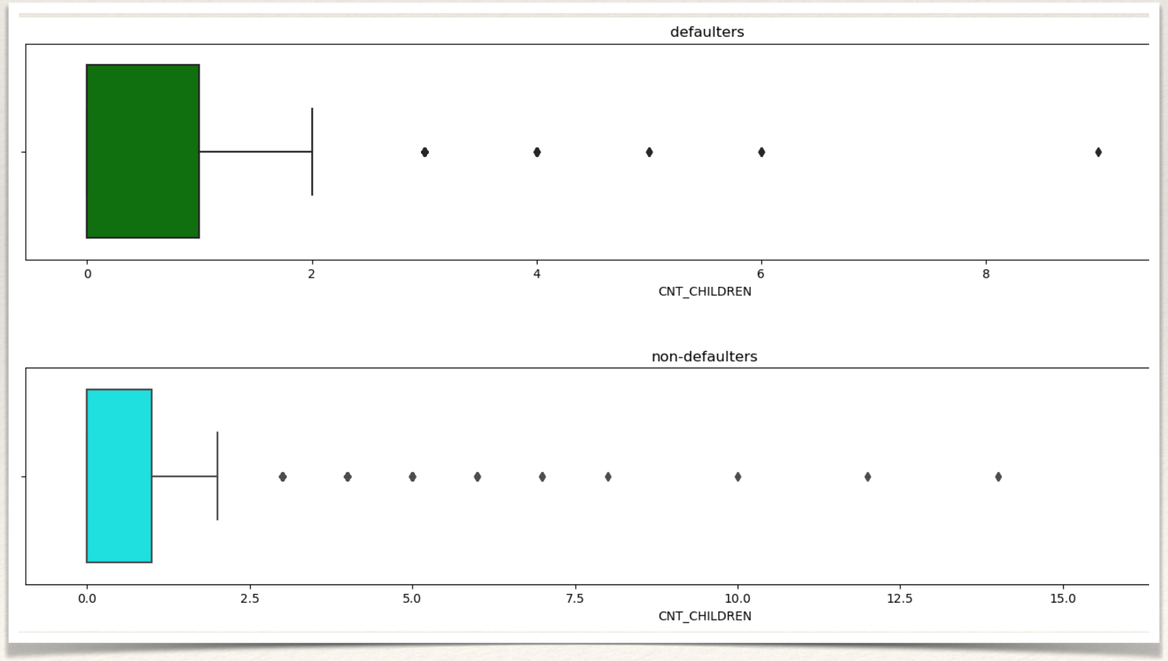
OUTLIERS ANALYSIS FOR TARGET -1, TARGET -0

ANALYSIS OF 'CNT_CHILDREN'

Many clients with 0 children are non defaulters whereas some clients with 1 and 2 childen are defaulters as complare to non defaulters.

- *Both distplots and boxplots clearly show the values above 2.5 as being outliers*
- *Applicants with 3 or more children are outlier case*

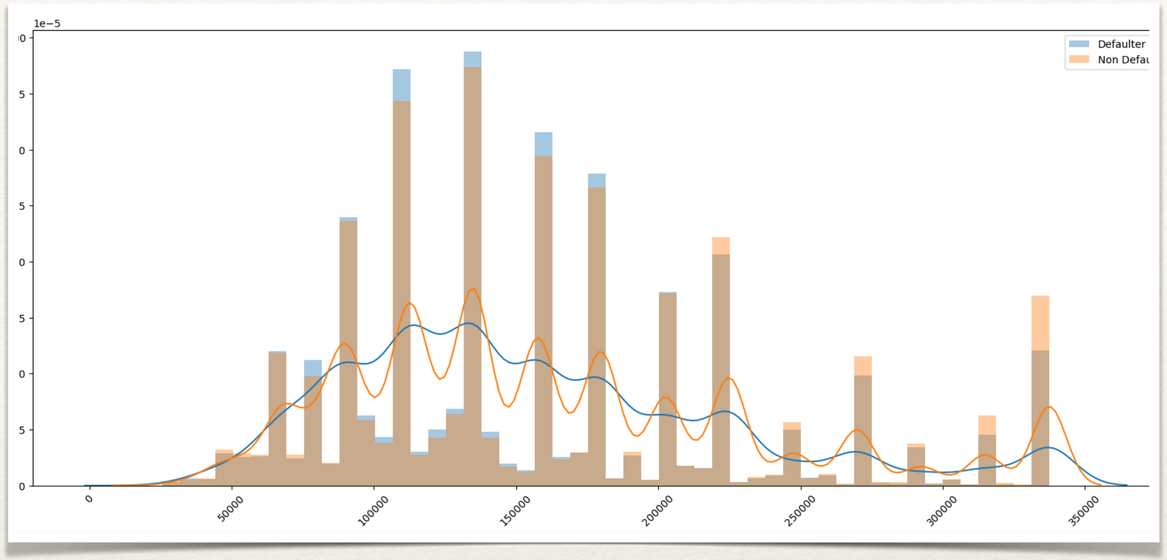


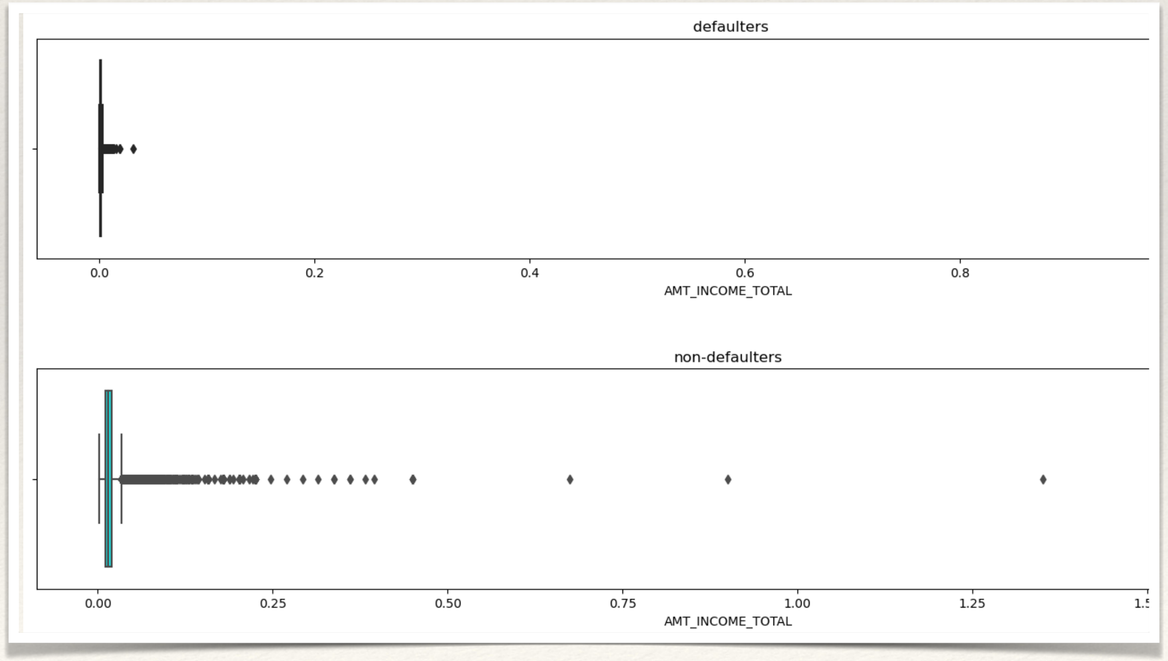


ANALYSIS OF ‘AMT_INCOME_TOTAL’

Not clear observation with AMT_INCOME_TOTAL

Applicants with `AMT_INCOME_TOTAL` above 337500.0 calculated using IQR) are outliers .
As observed from displot and boxplot, the outliers tend to exist after 337500.0



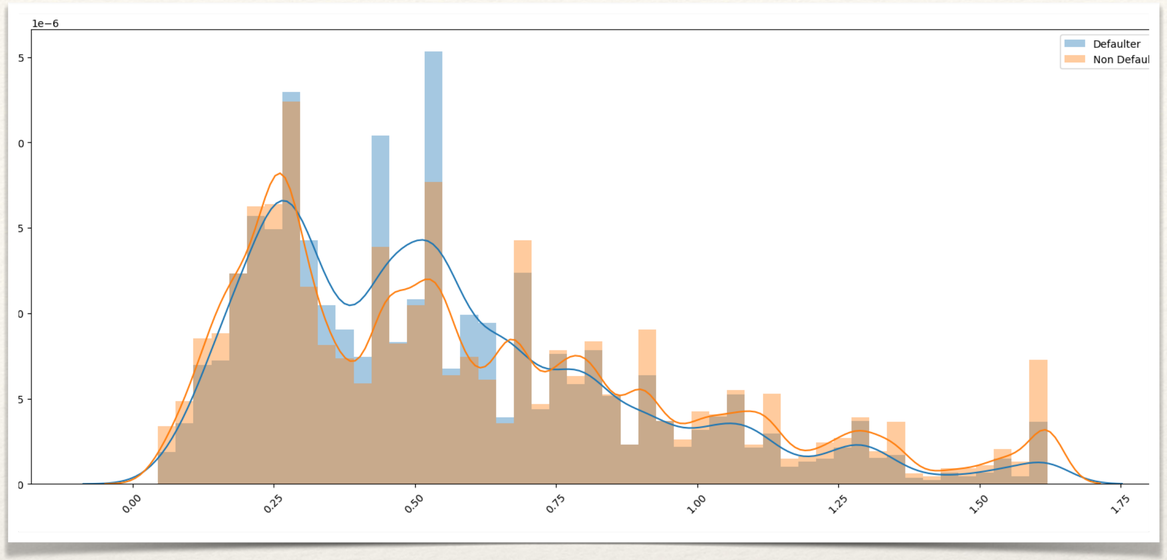


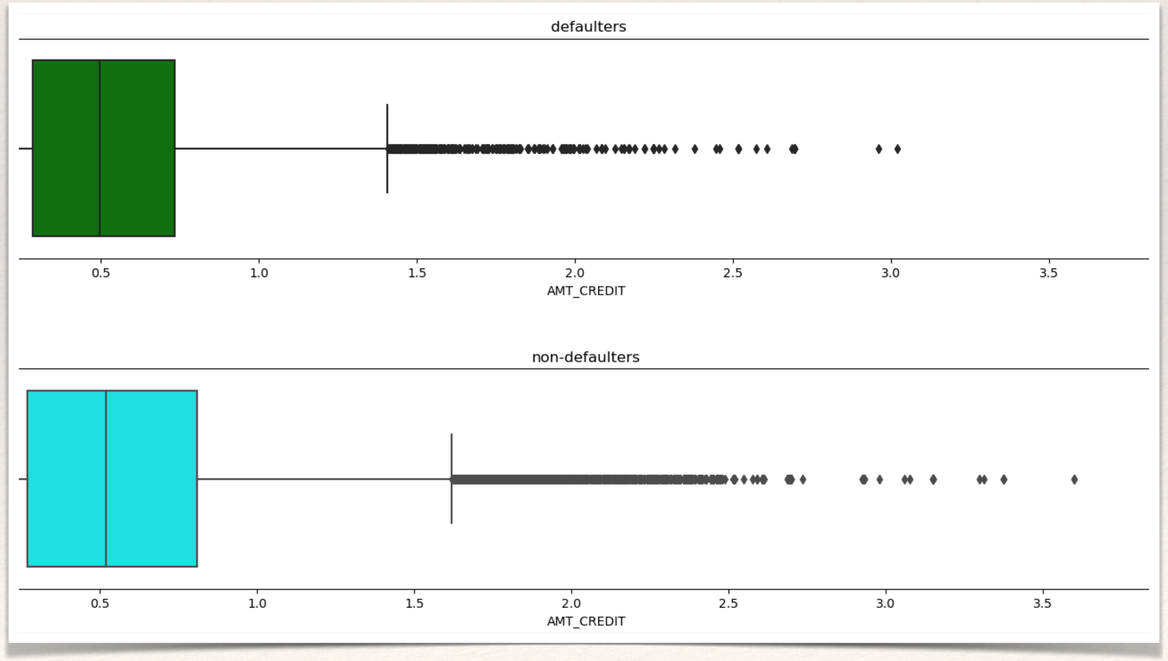
ANALYSIS OF ‘AMT_CREDIT’

Applicants with`AMT_CREDIT`above1620000.0 (calculated using IQR) are outliers .

As observed from displot and boxplot, the outliers tend to exist after 1620000.0

Clearly between 25k to 65k there are defaulters and clients >75k hare non defaulters



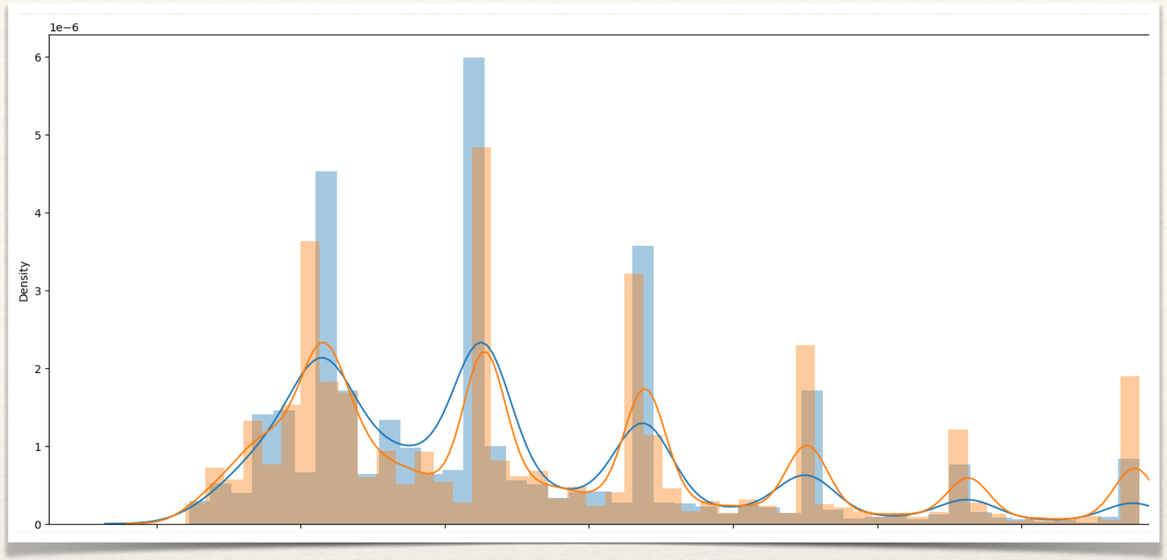


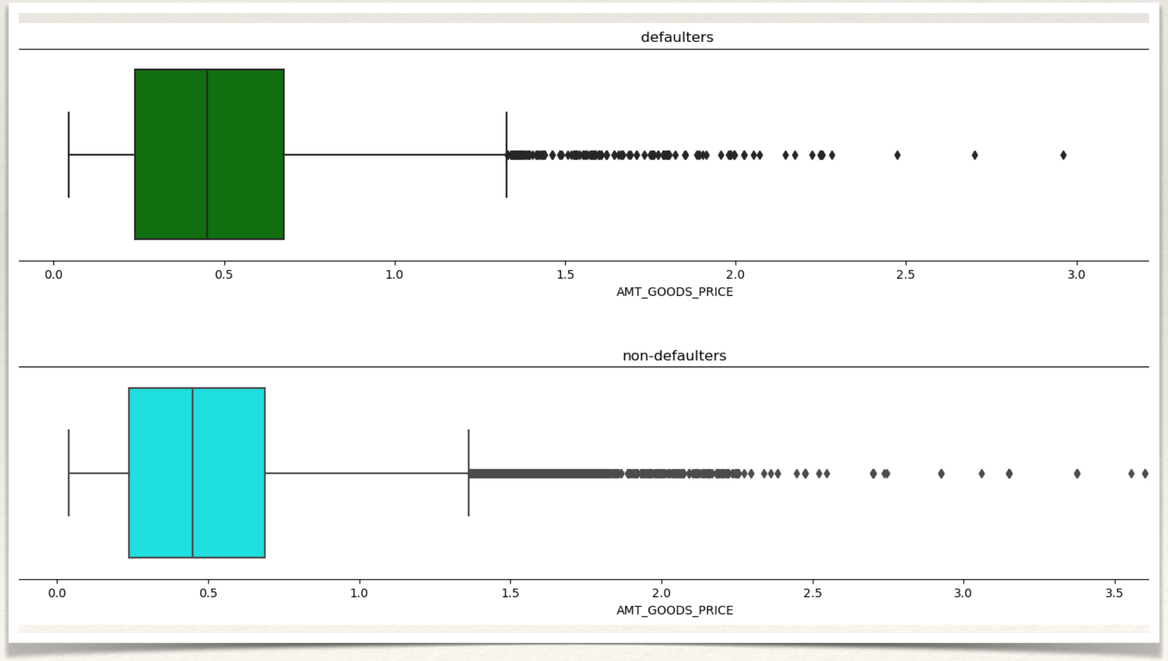
ANALYSIS OF ‘AMT_GOODS_PRICE’

As observed from displot and boxplot, the outliers tend to exist after 1363500

Applicants with `AMT_GOODS_PRICE` above 1363500 (calculated using IQR) are outliers

AMT_GOODS_PRICE between 25k and 55k there are more clients who are defaulters





CHECKING IMBALANCE DATA

DATA IMBALANCE

We can see that there is data imbalance in columns:-

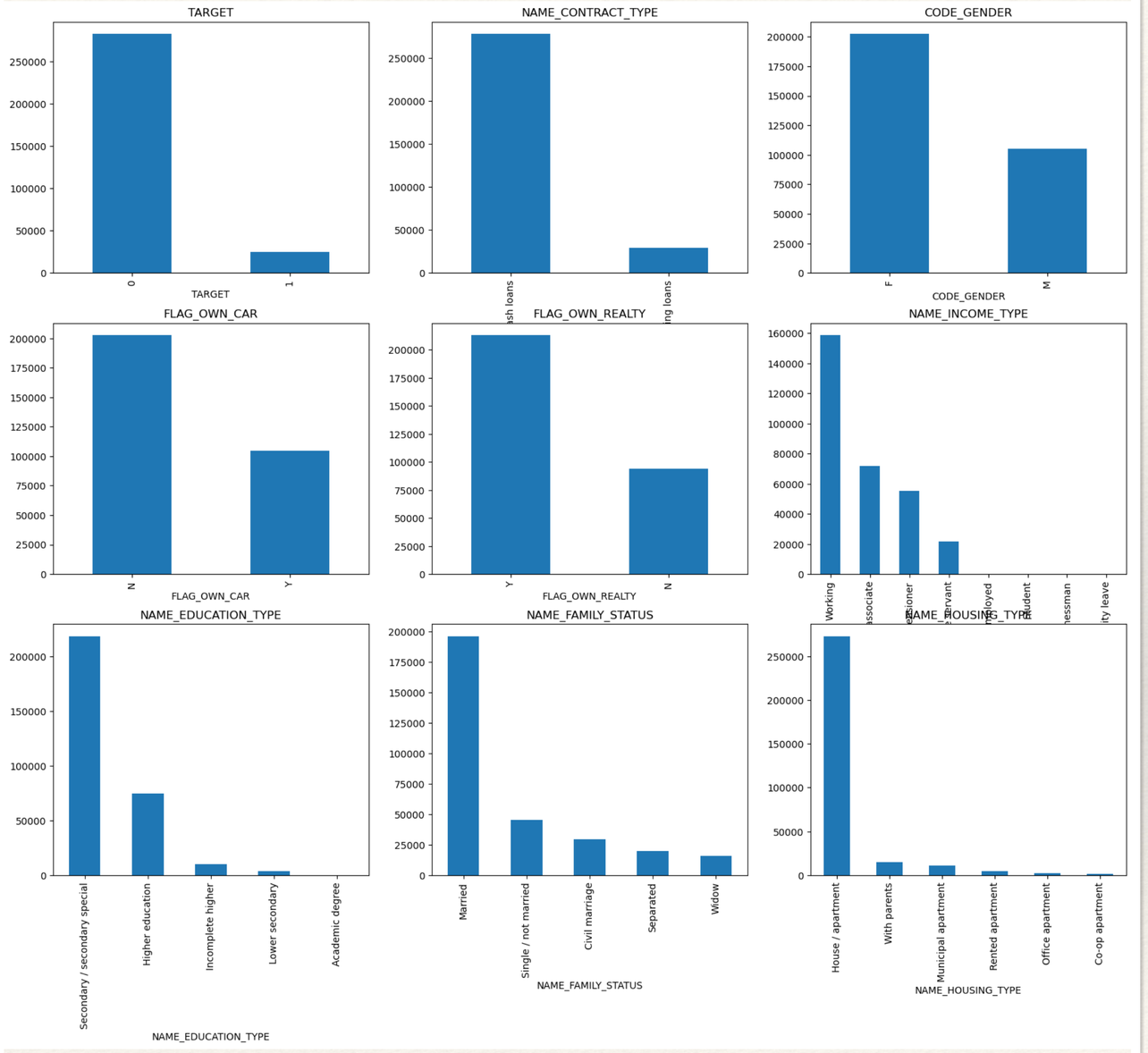
TARGET - There are very few defaulters(1) compare to non defaulters(0)

NAME_CONTRACT_TYPE - There are very few Revolving loans than Cash loans

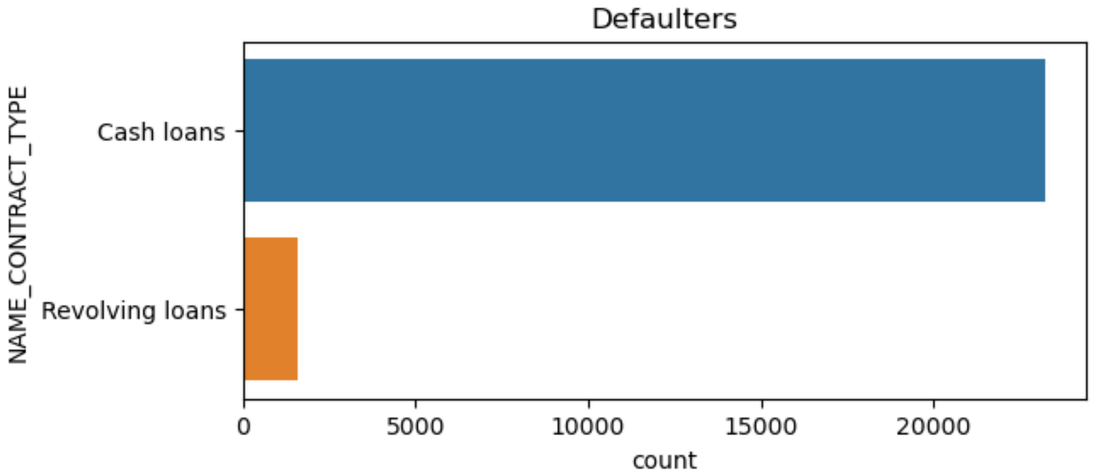
NAME_EDUCATION_TYPE - Most of the loans applied by Secondary/Secondary special educated people

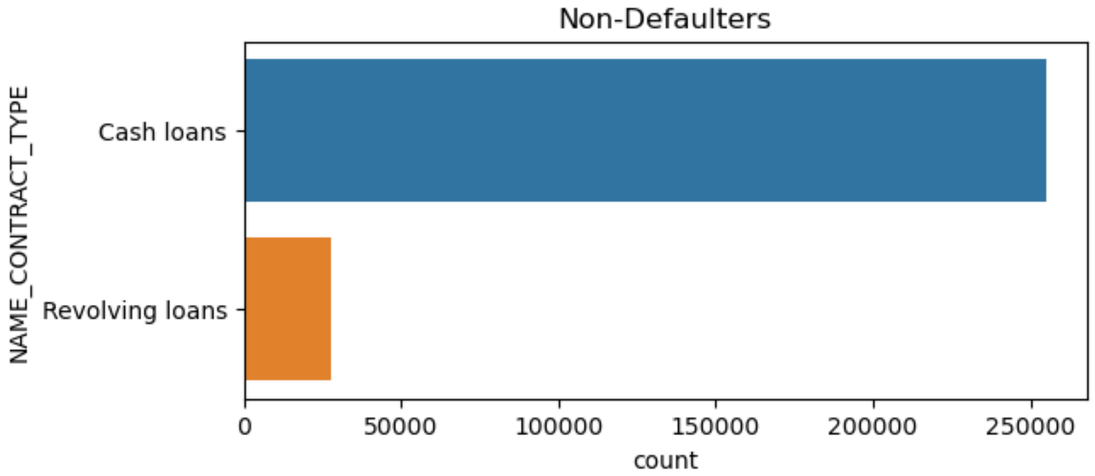
NAME_FAMILY_STATUS - Most of the loans applied by Married people.

NAME_HOUSING_TYPE - Most of the application came from Home / appartment owner



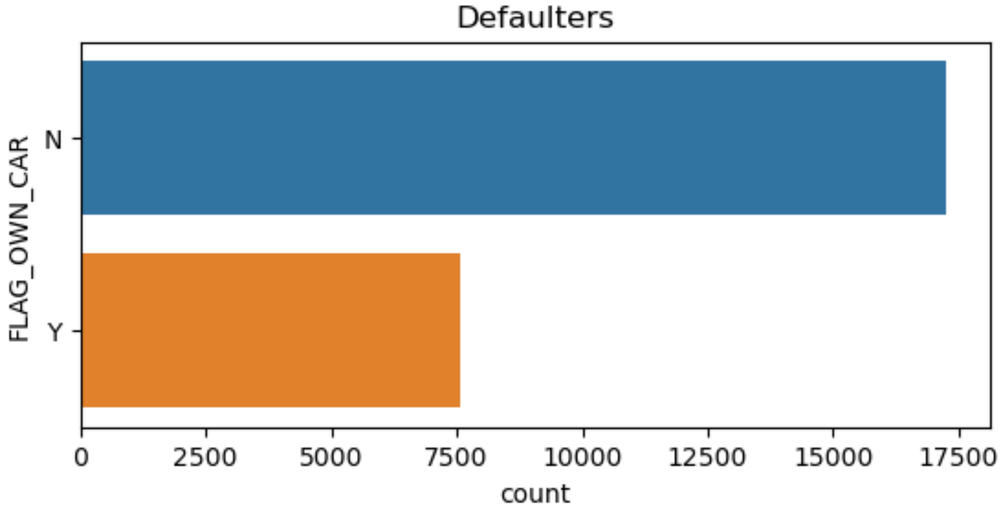
UNIVARIATE ANALYSIS

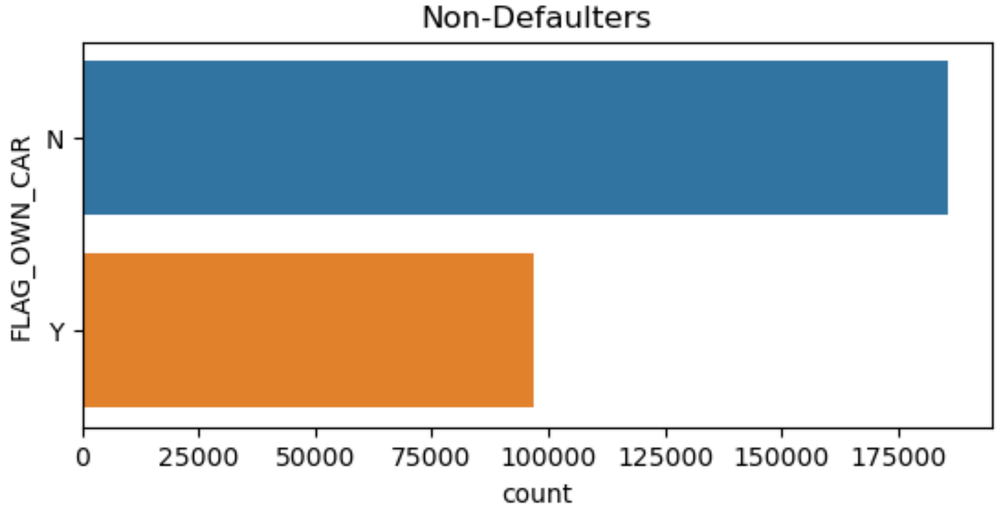




ANALYSIS OF 'NAME_CONTRACT_TYPE'

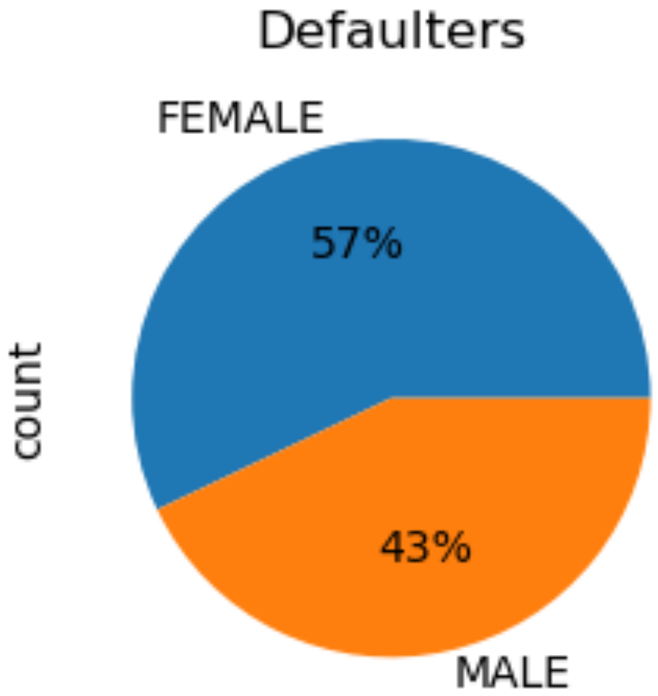
NAME_CONTRACT_TYPE` column does not provide any conclusive evidence in favor of clients with payment difficulties OR on-time payments .

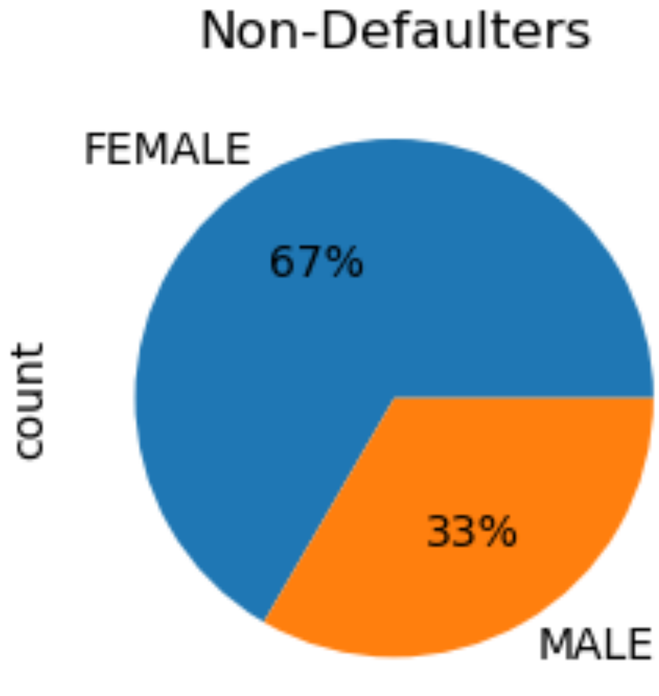




ANALYSIS OF 'FLAG_OWN_CAR'

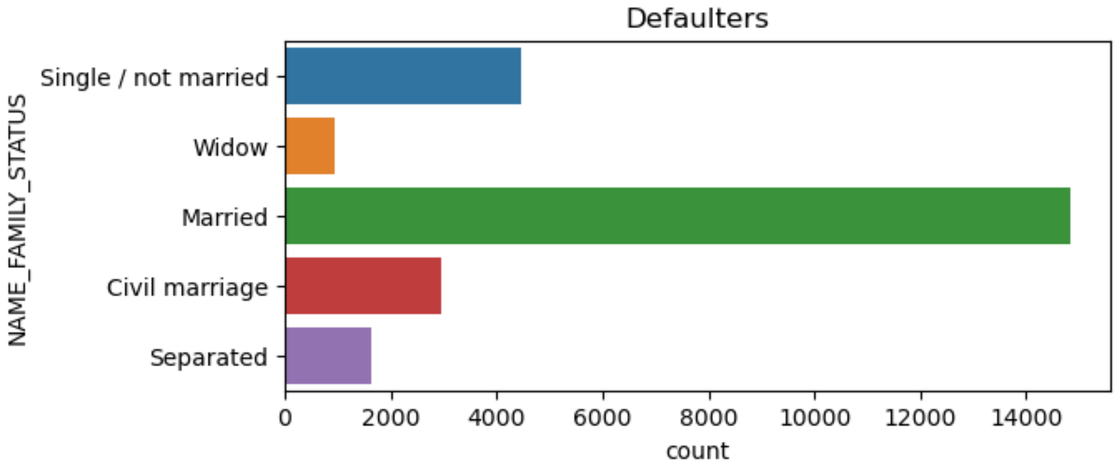
FLAG_OWN_CAR column does not provide any INSIGHT of clients with DEFAULTERS OR NON DEFAULTERS

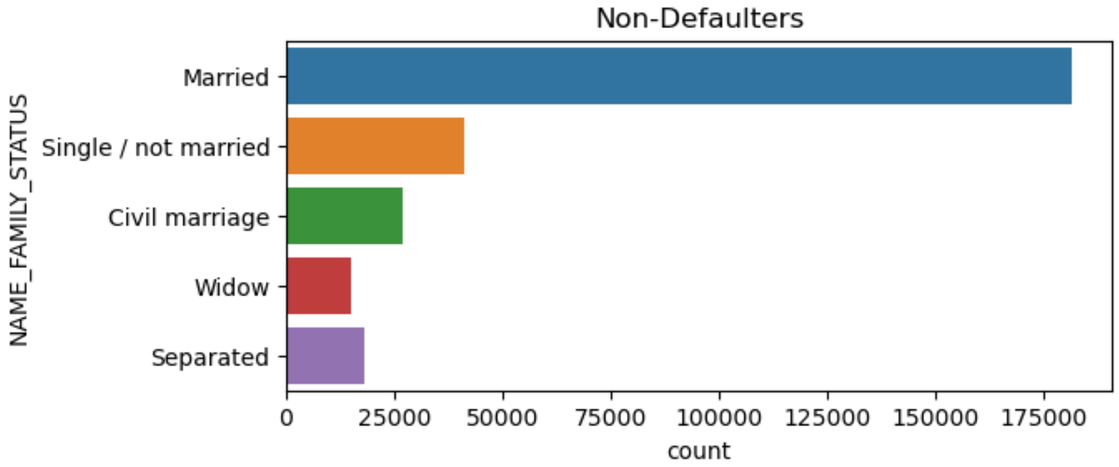




ANALYSIS OF 'CODE_GENDER'

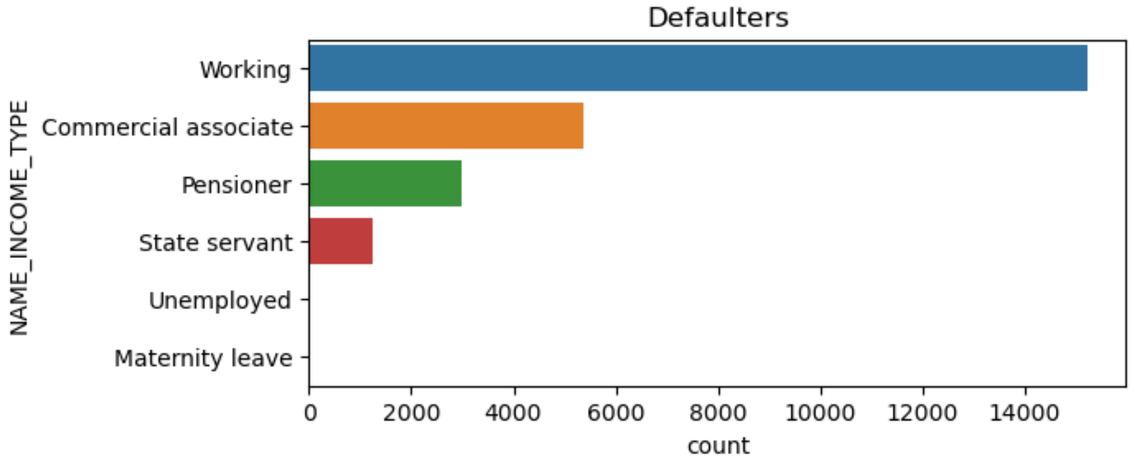
`CODE_GENDER` column provides a weak inference that "Male" clients have more payment difficulties

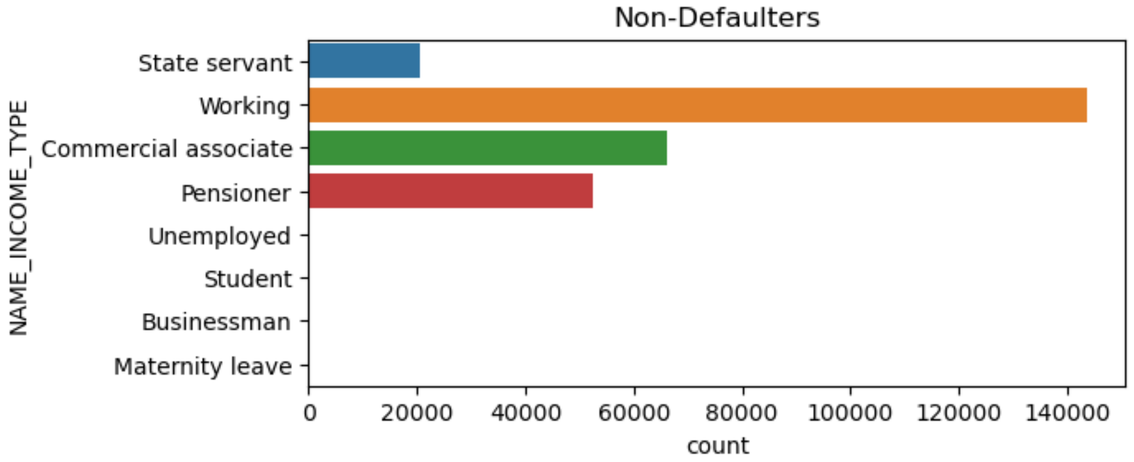




ANALYSIS OF 'NAME_FAMILY_STATUS'

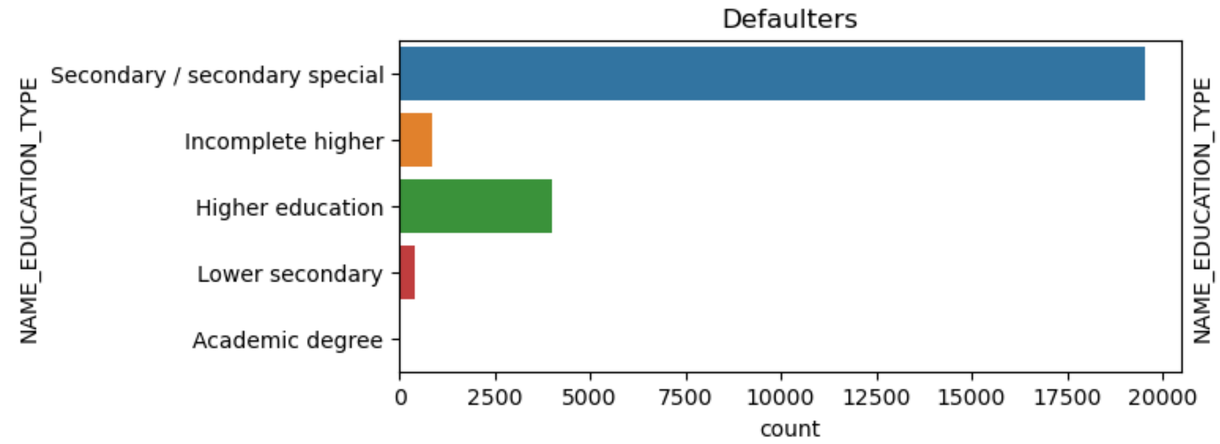
Most of the the applicants are married 2nd highest applicants are Single/Non-married.Most of the the applicants are married in both defaulter and non-defaulter categories

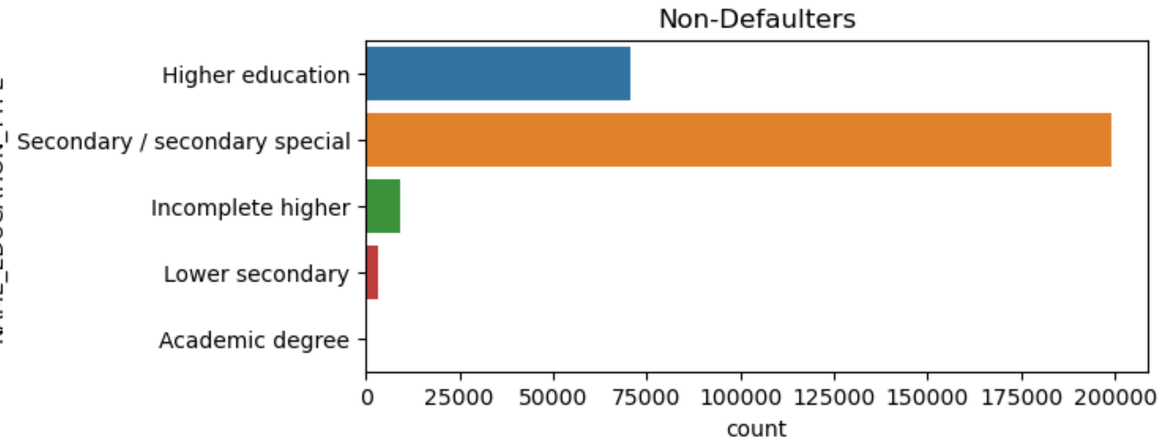




ANALYSIS OF 'NAME _INCOME_TYPE'

Pensioners ARE non defaulters simillarly students and Businessmen are also non defaulters.Whereas working category are falls under defaulters

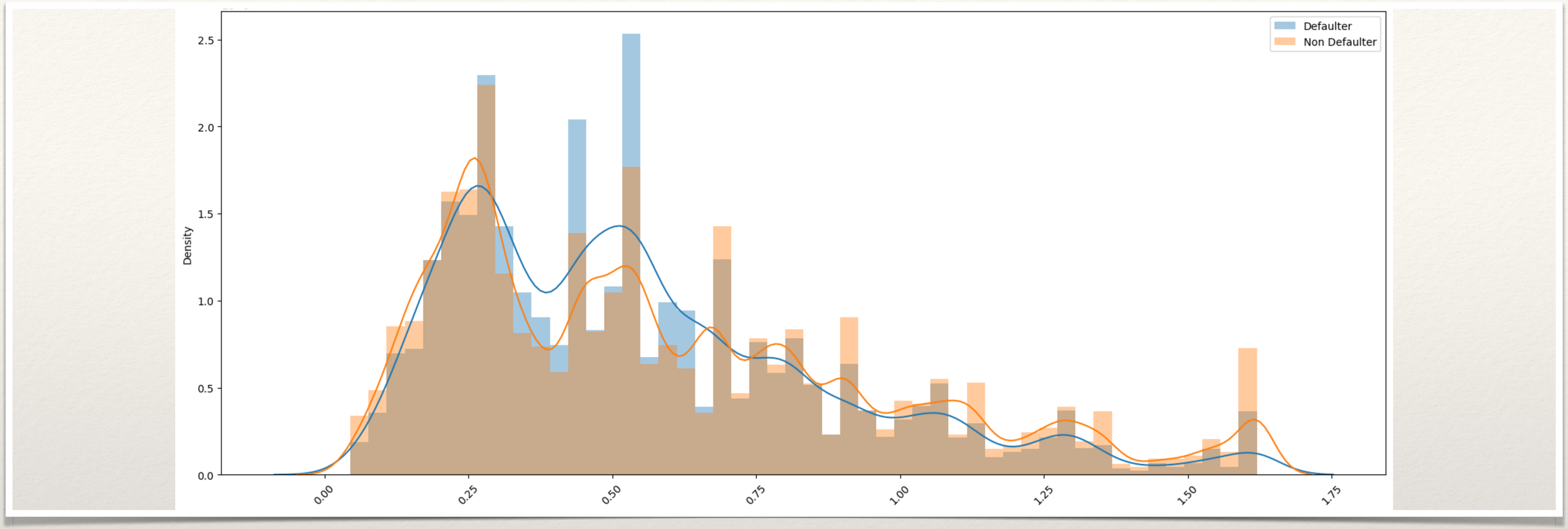




Lorem Ipsum Dolor

ANALYSIS OF 'NAME_EDUCATION_TYPE'

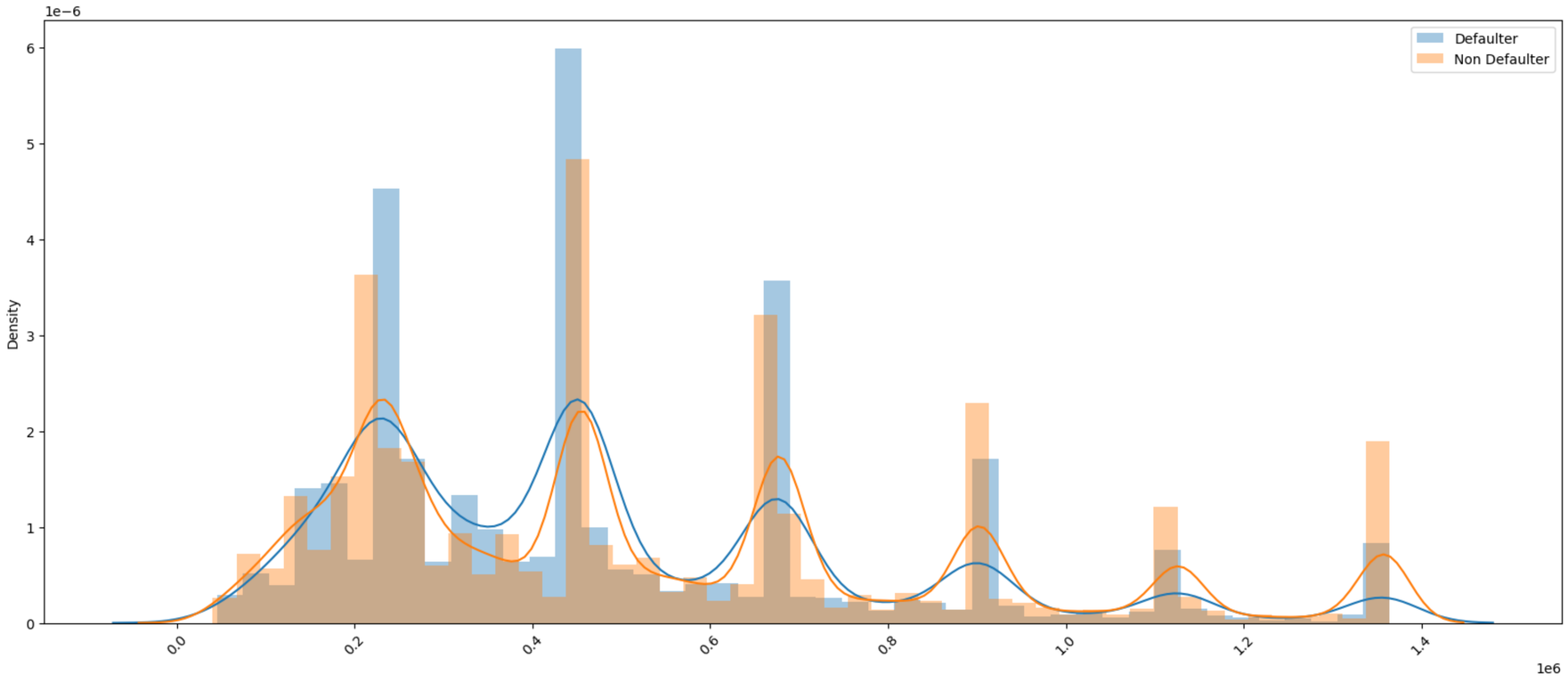
Clients with higher education are non defaulters



ANALYSIS OF 'AMT_CREDIT'

For `AMT_CREDIT` between 250000 and approximately 650000, there are more clients seems to be in default zone

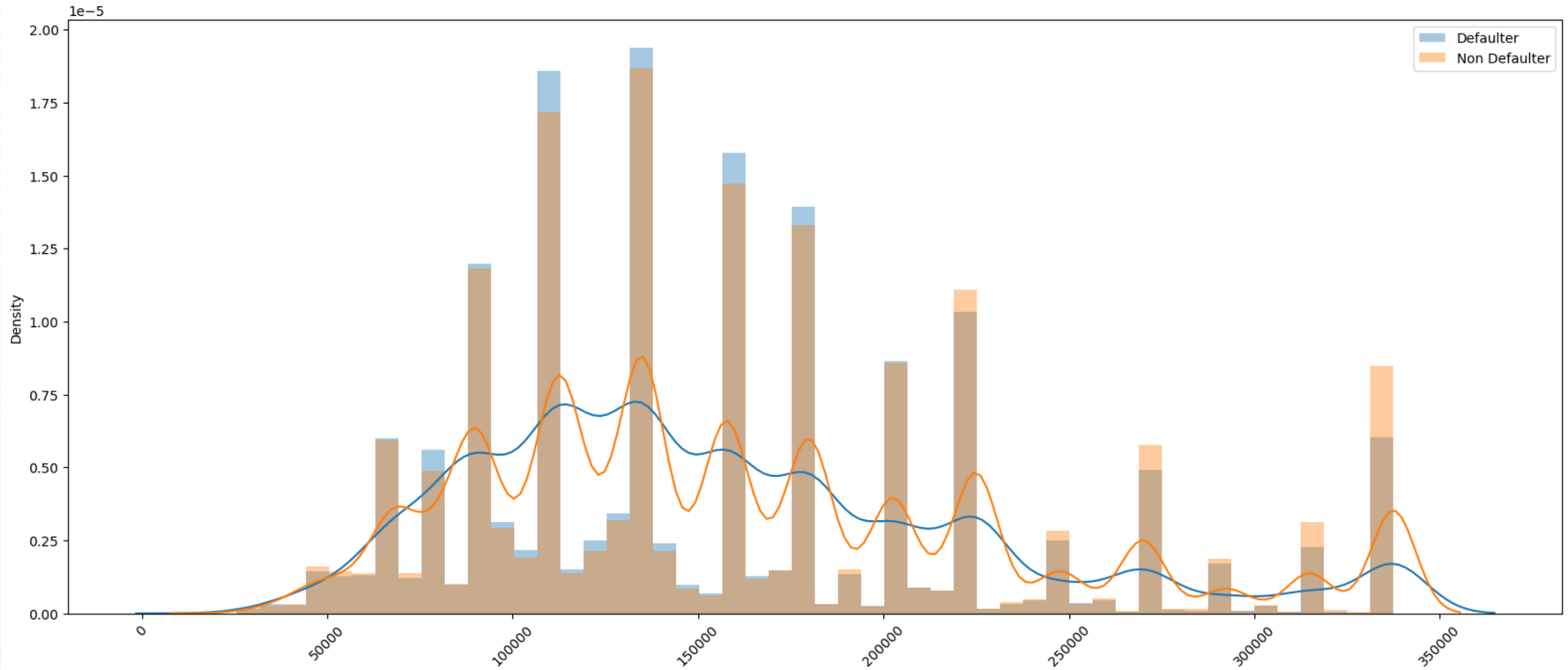
- For `AMT_CREDIT` > 750000 , there are more clients who are NON DEFAULTERS



ANALYSIS OF 'AMT_GOODS_PRICE'

For `AMT_GOODS_PRICE` between ~250000 and ~550000, there are more clients with Payment difficulties

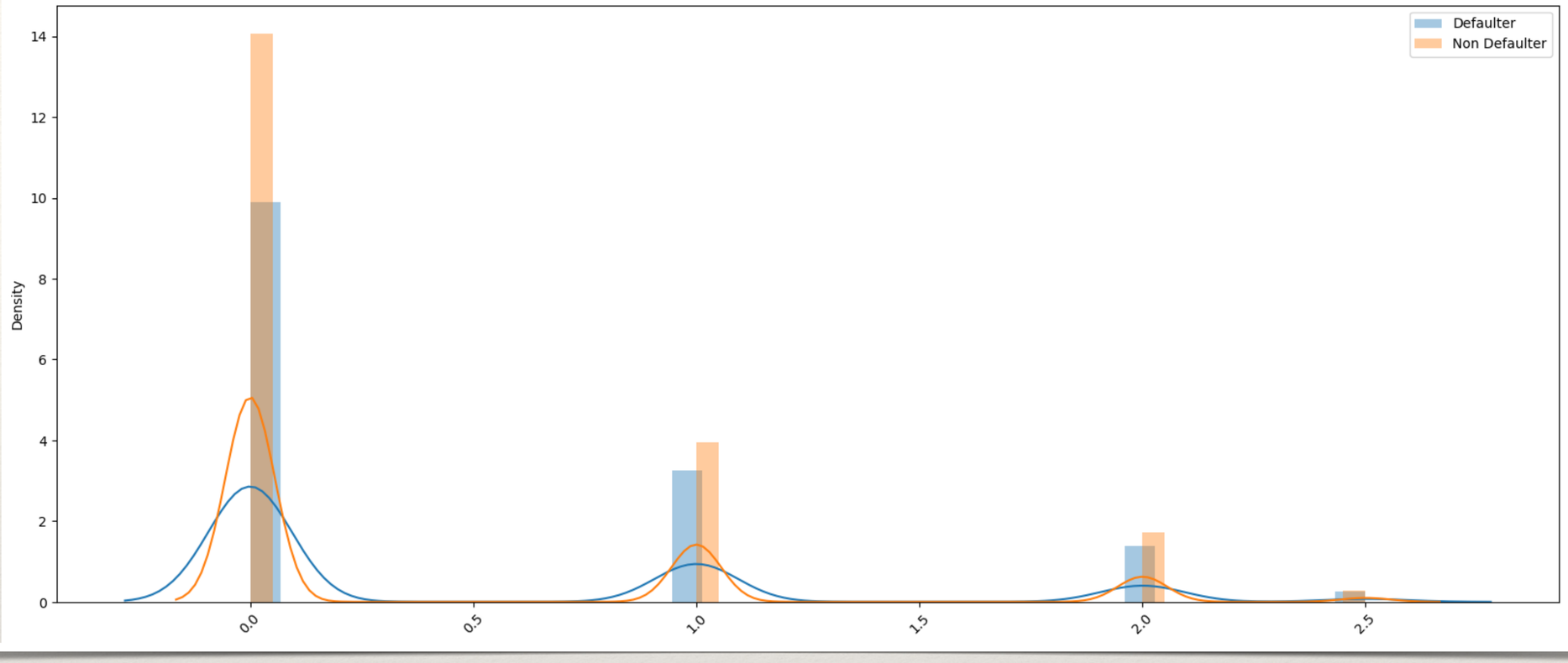
- Otherwise there are spikes on and off but they don't show any conclusive observations



ANALYSIS OF 'AMT_INCOME_TOTAL'

- Based on `AMT_INCOME_TOTAL`, for clients with DEFAULT STATUS, the distribution resembles a normal distribution approximately

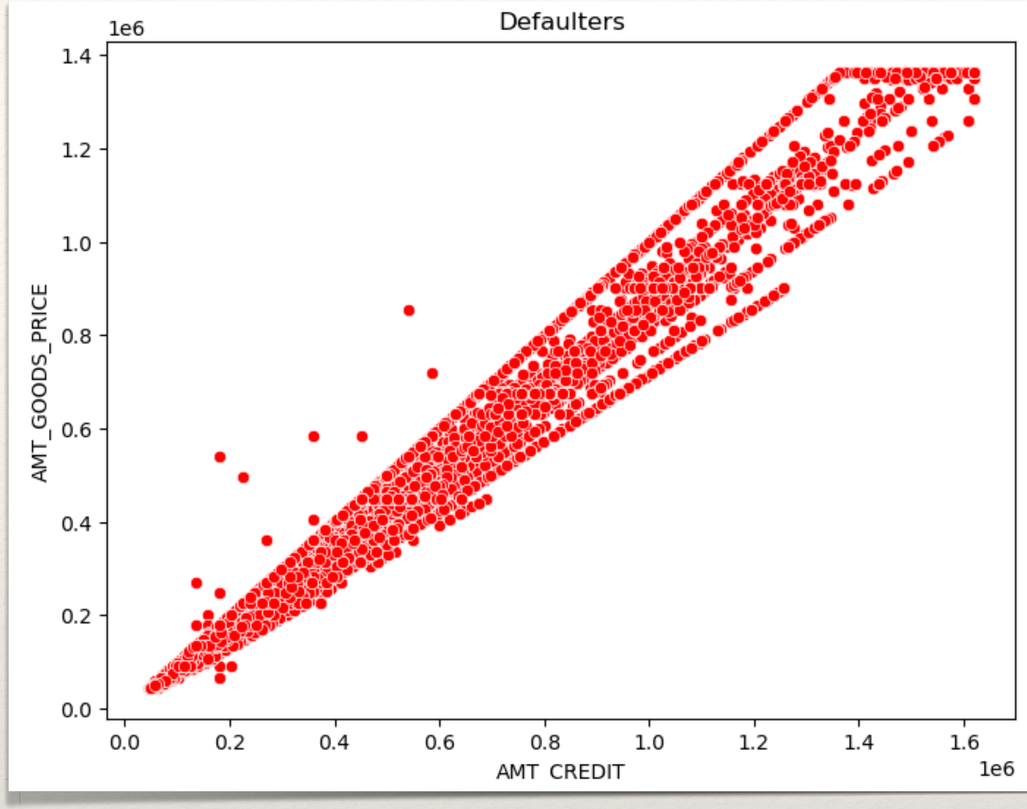
- But for clients with NON DEFAULT, there are erratic spikes in the distribution which doesn't give any valid observations

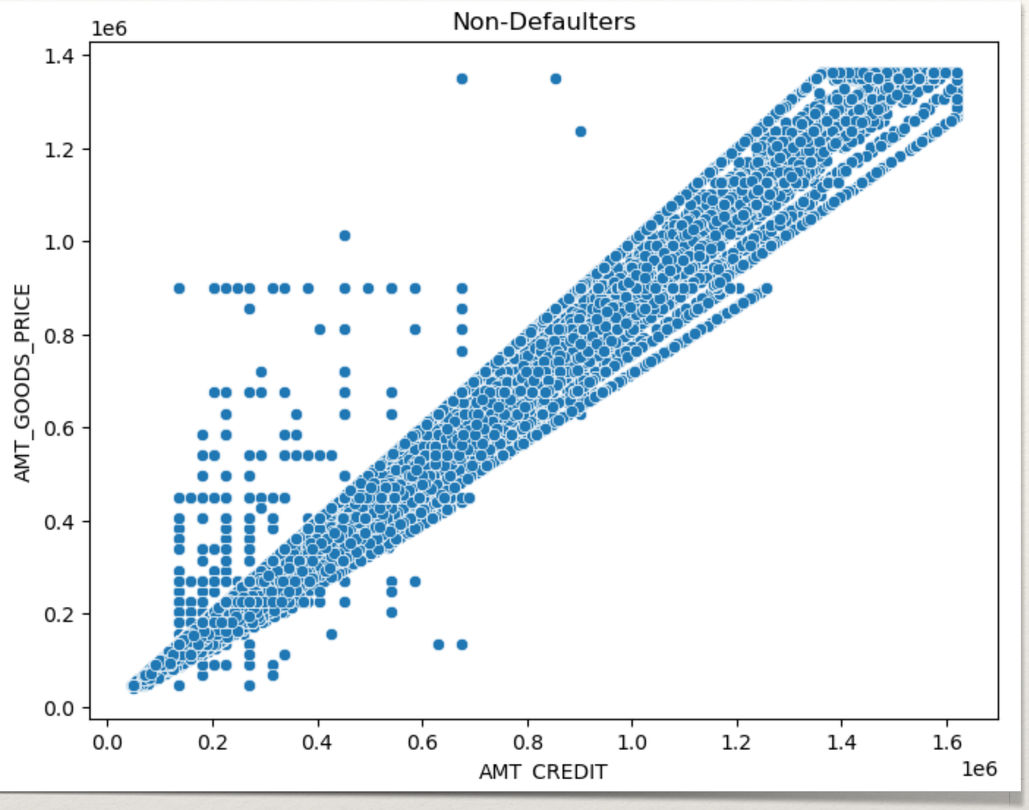


ANALYSIS OF 'CNT_CHILDREN'

For `CNT_CHILDREN` 0 (those with no children), there are lots of clients comes under non defaulters - For `CNT_CHILDREN` with 1 OR 2 (those with 1 or 2 children), there are few more clients with NON defaulters

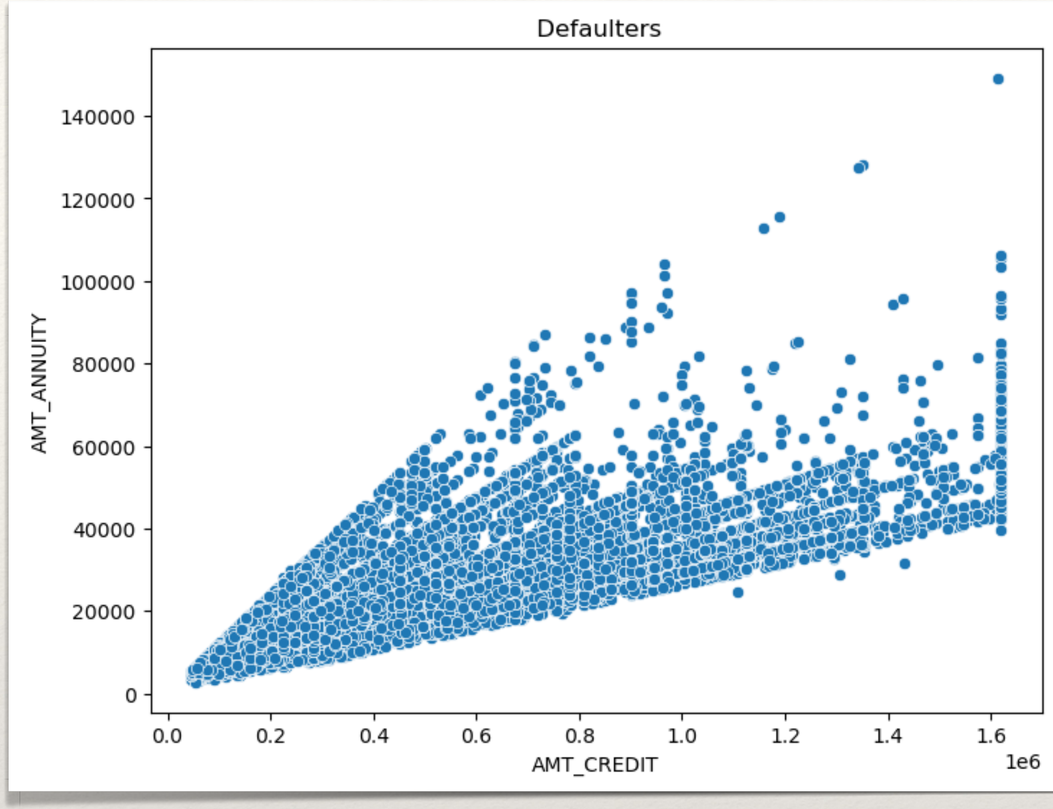
BIVARIATE/MULTIVARIATE ANALYSIS

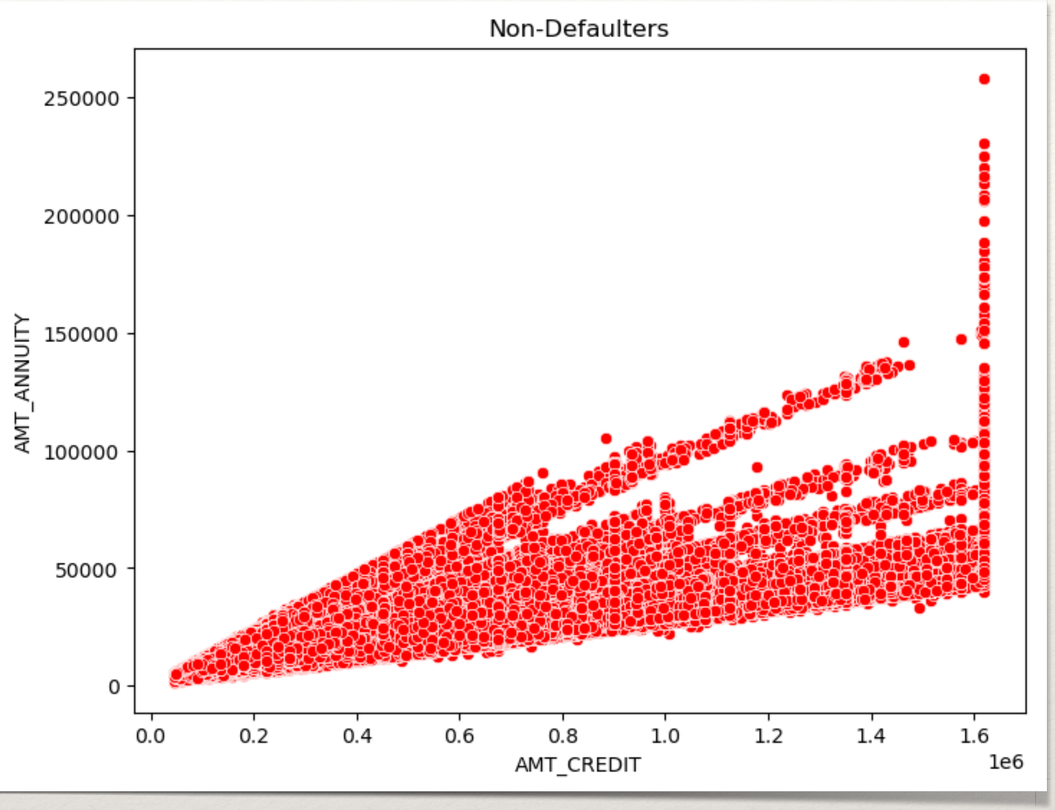




ANALYSIS BETWEEN 'AMT_CREDIT' & "AMT_GOODS_PRICE

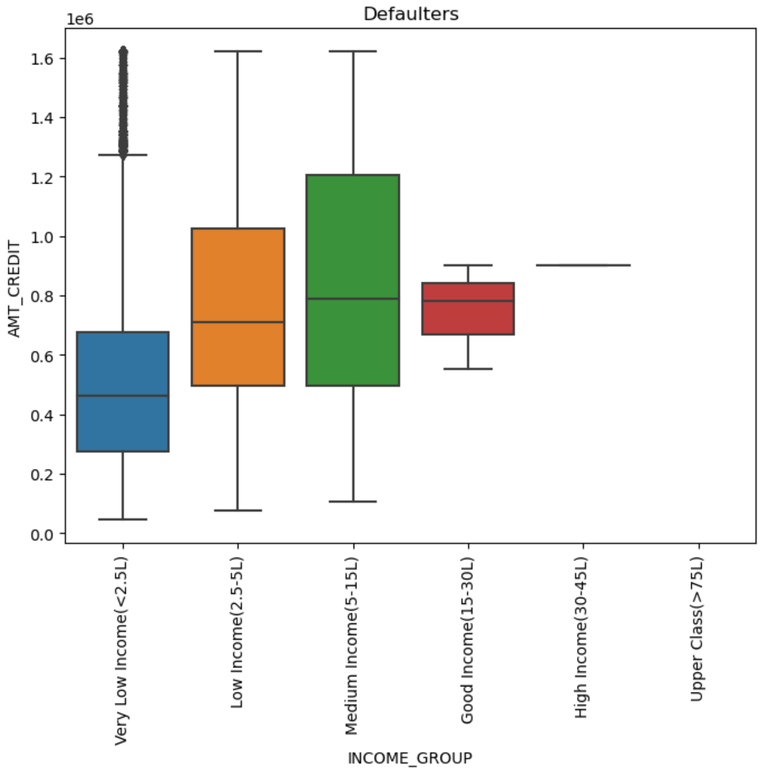
- `AMT_GOODS_PRICE` and `AMT_CREDIT` have strong positive correlation. This means that as Goods price increases, so does Credit Amount

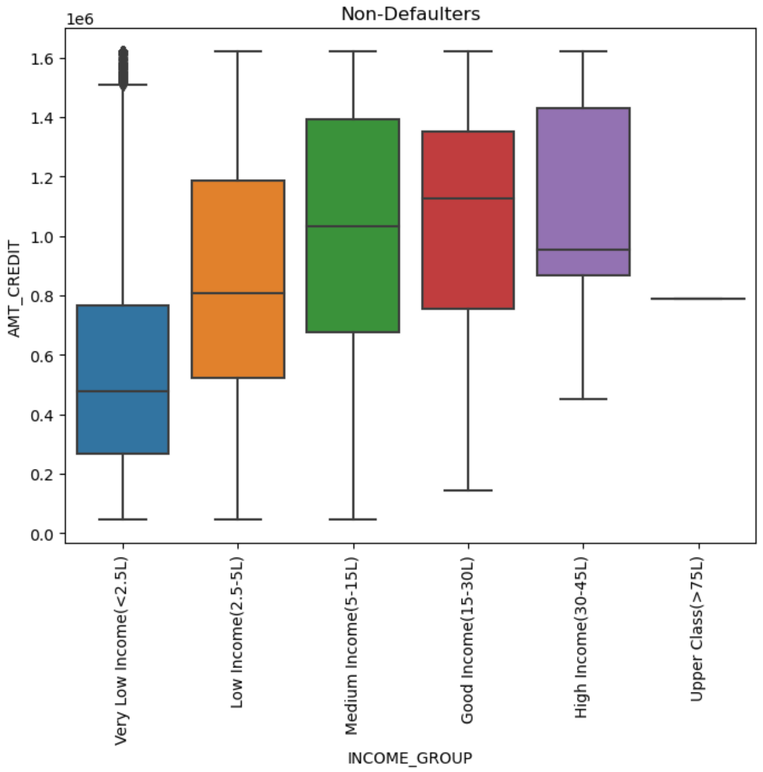




ANALYSIS BETWEEN 'AMT_CREDIT' & "AMT_ANNUITY

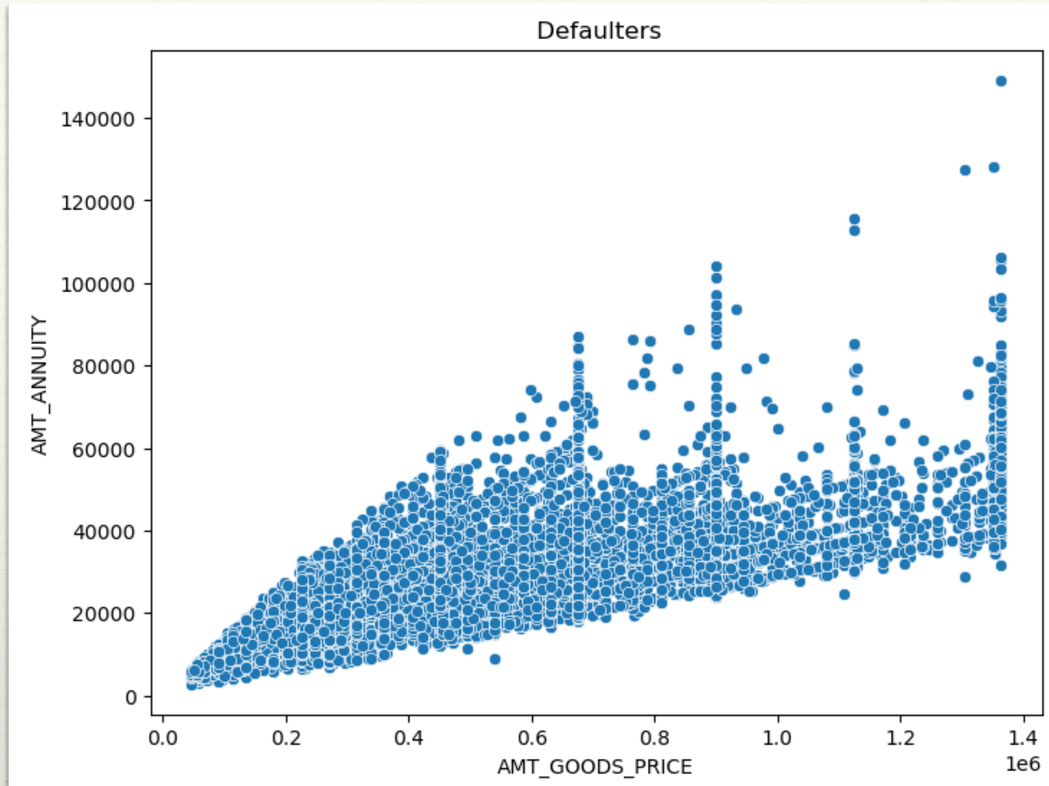
AMT_ANNUITY` and `AMT_CREDIT` have strong positive correlation. This means that as Annuity Amount increases, so does Credit Amount

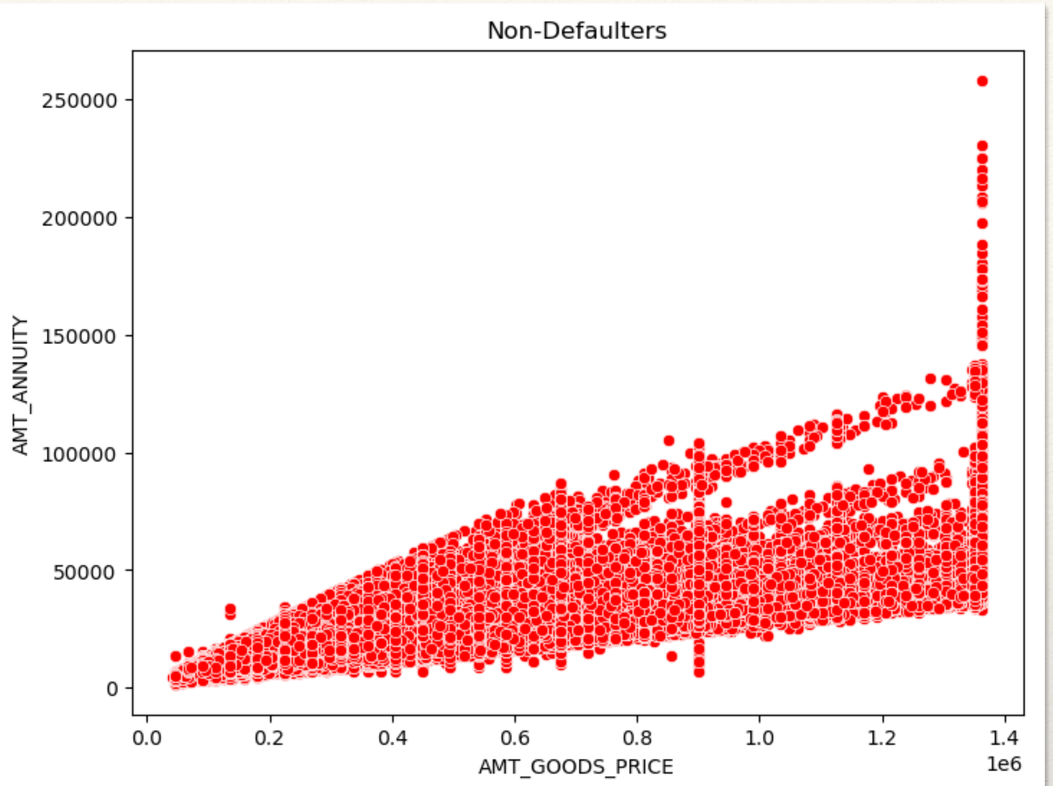




ANALYSIS BETWEEN 'AMT_CREDIT' & "INCOME_GROUP

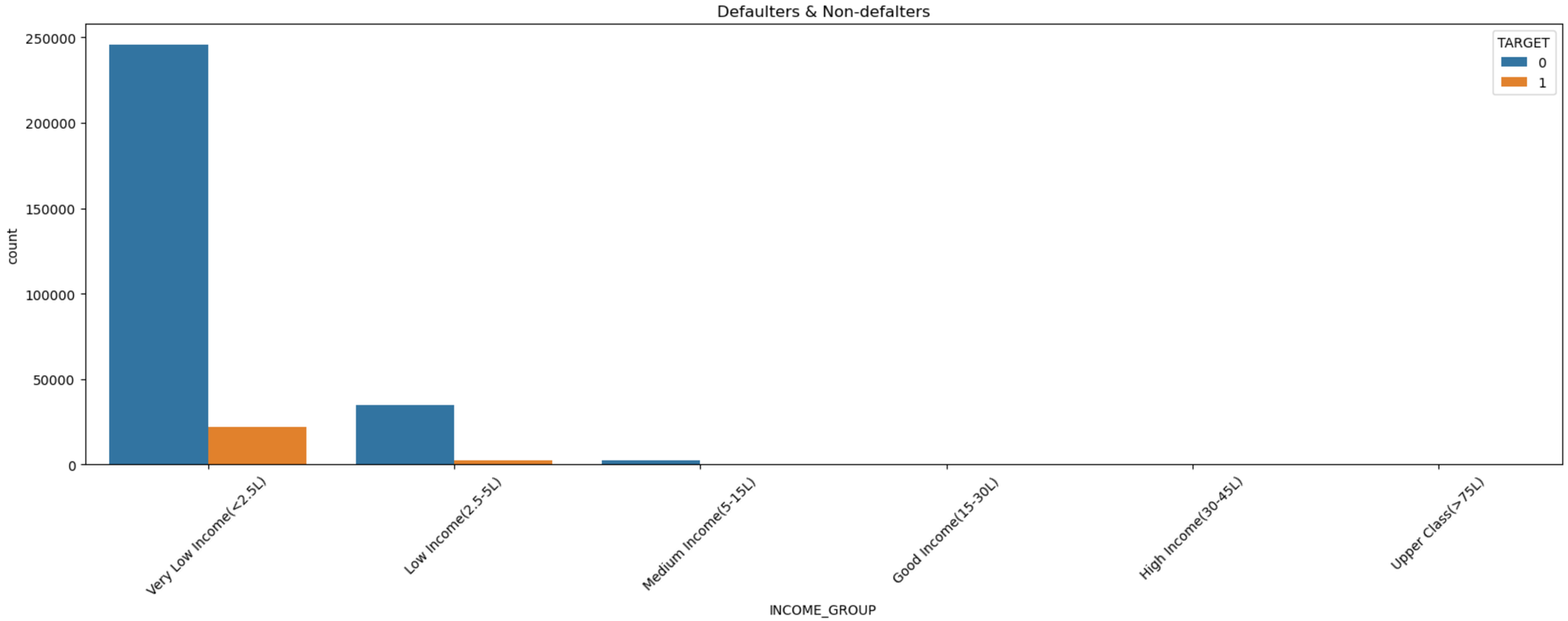
INCOME_GROUP` and `AMT_CREDIT` have strong positive correlation. This means that as INCOME_GROUP increases, so does Credit Amount





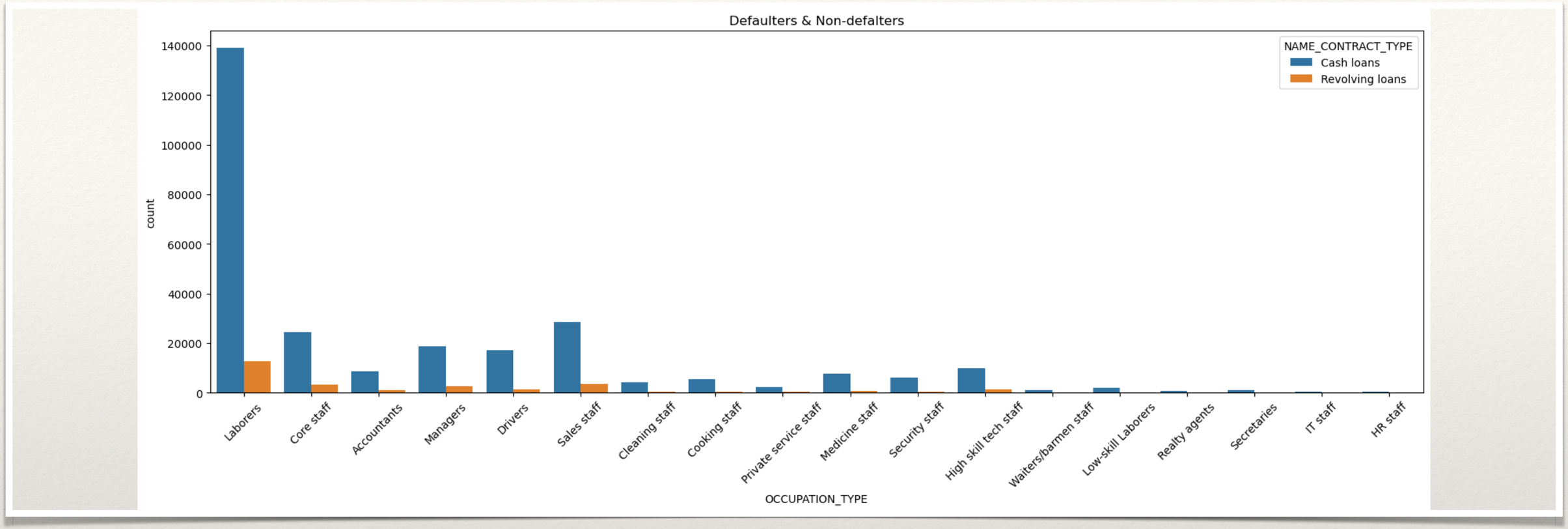
ANALYSIS BETWEEN 'AMT_ANNUITY' & "AMT_GOODS_PRICE

`AMT_ANNUITY` and `AMT_GOODS_PRICE` have strong positive correlation. This means that as Annuity increases, so does Goods Price



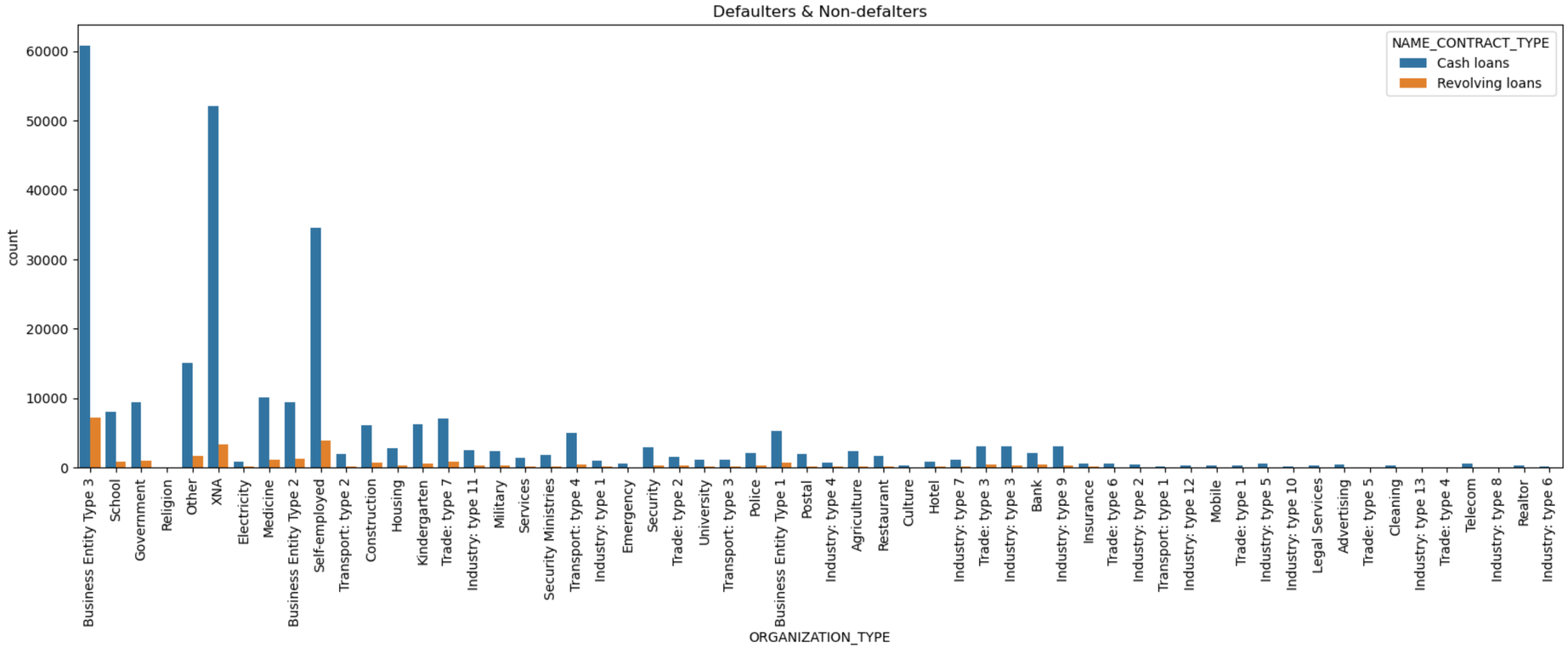
ANALYSIS BETWEEN 'TARGET' & "INCOME_GROUP

we can see maximum numbers of defaulters are of very low income



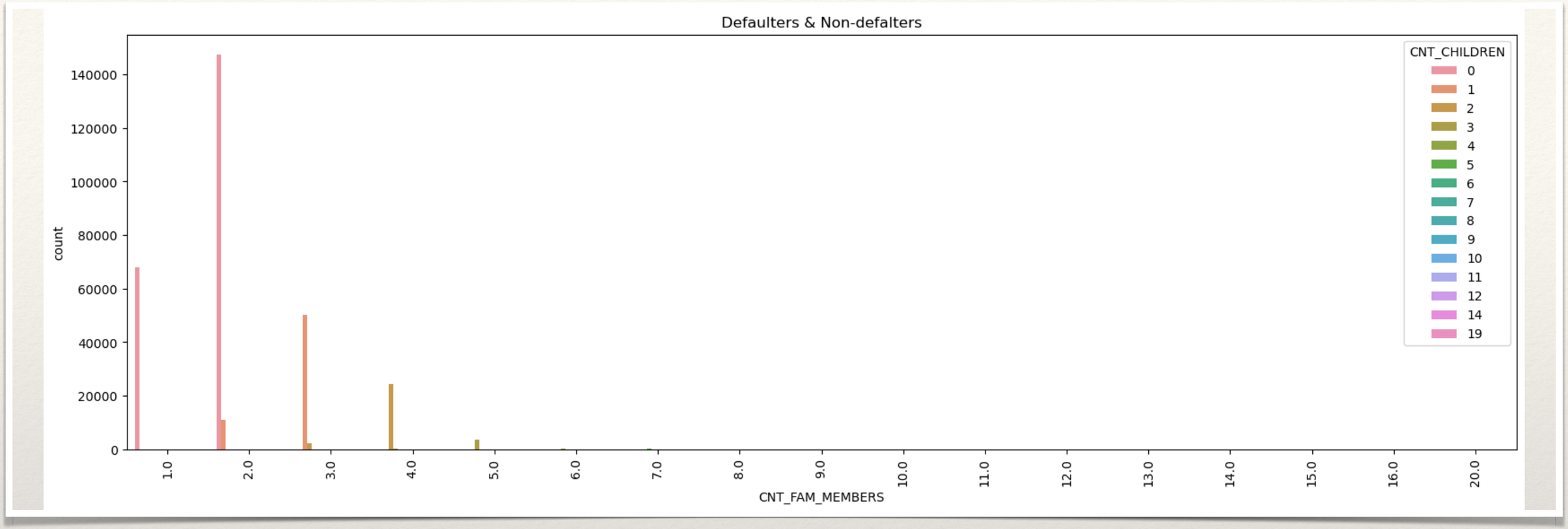
ANALYSIS BETWEEN 'OCCUPATION_TYPE' & "NAME_CONTRACT_TYPE

Clients who are `Sales staff`, `Laborers`, `Drivers` ,Core staff and have `Cash loans` are more in default circle as compared to people comes under non default circle



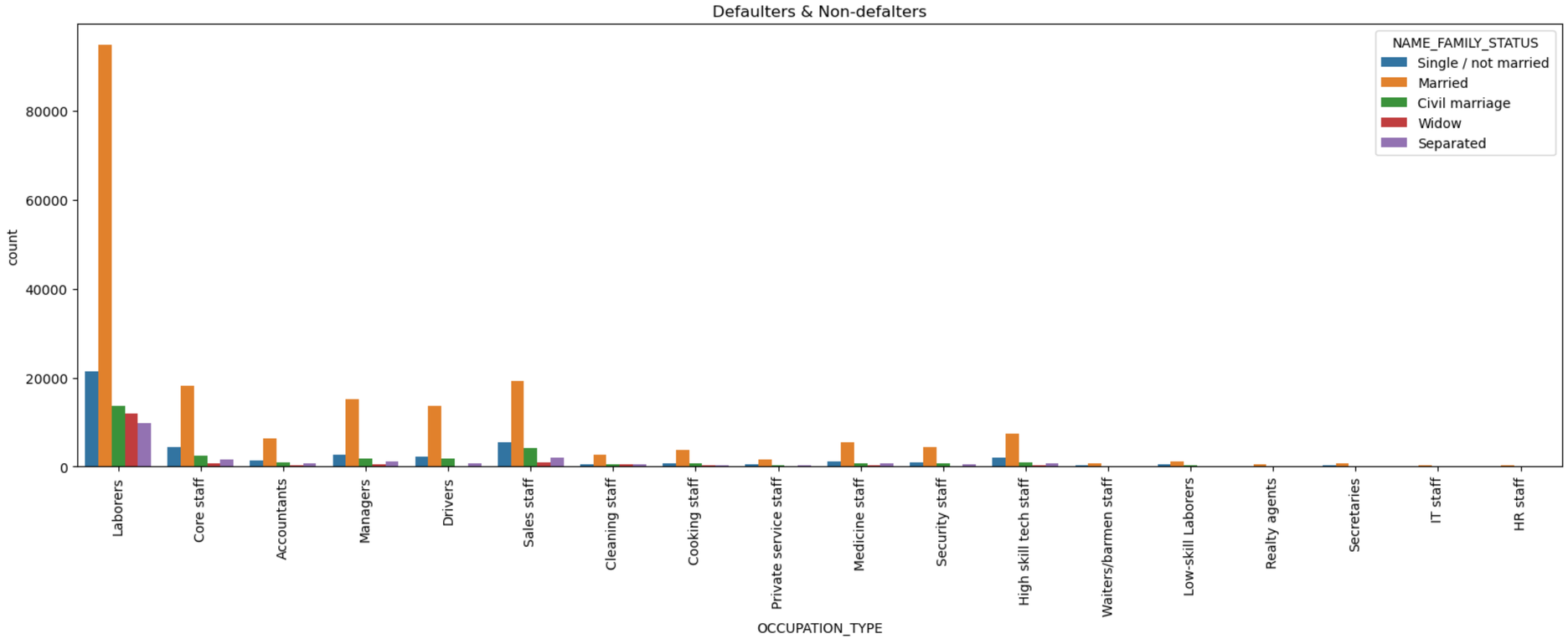
ANALYSIS BETWEEN 'ORGANISATION_TYPE' & "NAME_CONTRACT_TYPE

Clients who are `Self Employed , `Business entity type 3`, `XNA` ,Core and have `Cash loans` are more in default circle as compared to people comes under non default circle



ANALYSIS BETWEEN 'CNT_FAM_MEMBERS' & 'CNT_CHILDREN

Clients who have one or two children ` are more likely to come under default circle as compared to other people comes under non default circle

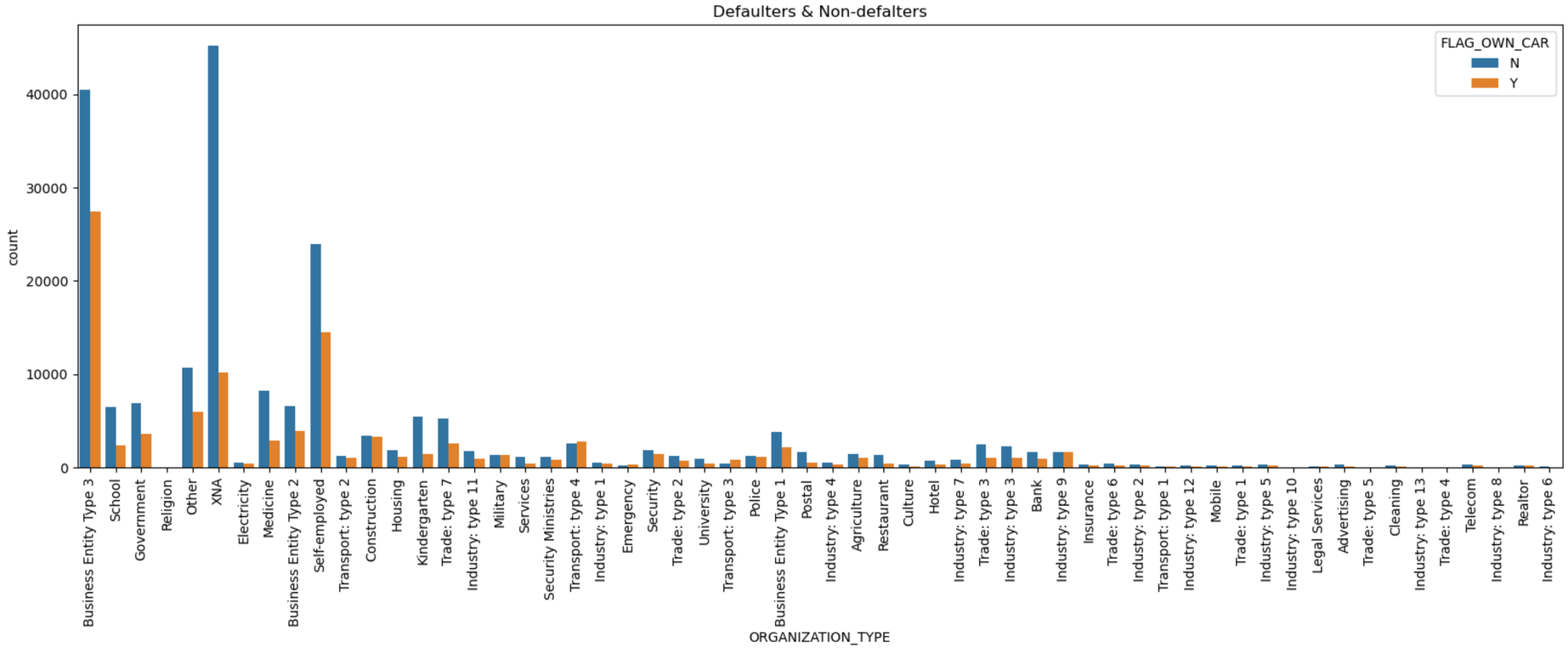


ANALYSIS BETWEEN 'OCCUPATION_TYPE' & 'NAME_CONTRACT_TYPE

Clients who are `Single/not married` & `Married` and are `Laborers` have are more in default circle compared to non defaulters

Clients who are `Married` and are `Drivers` are seems to be defaulters se compared to non default clients

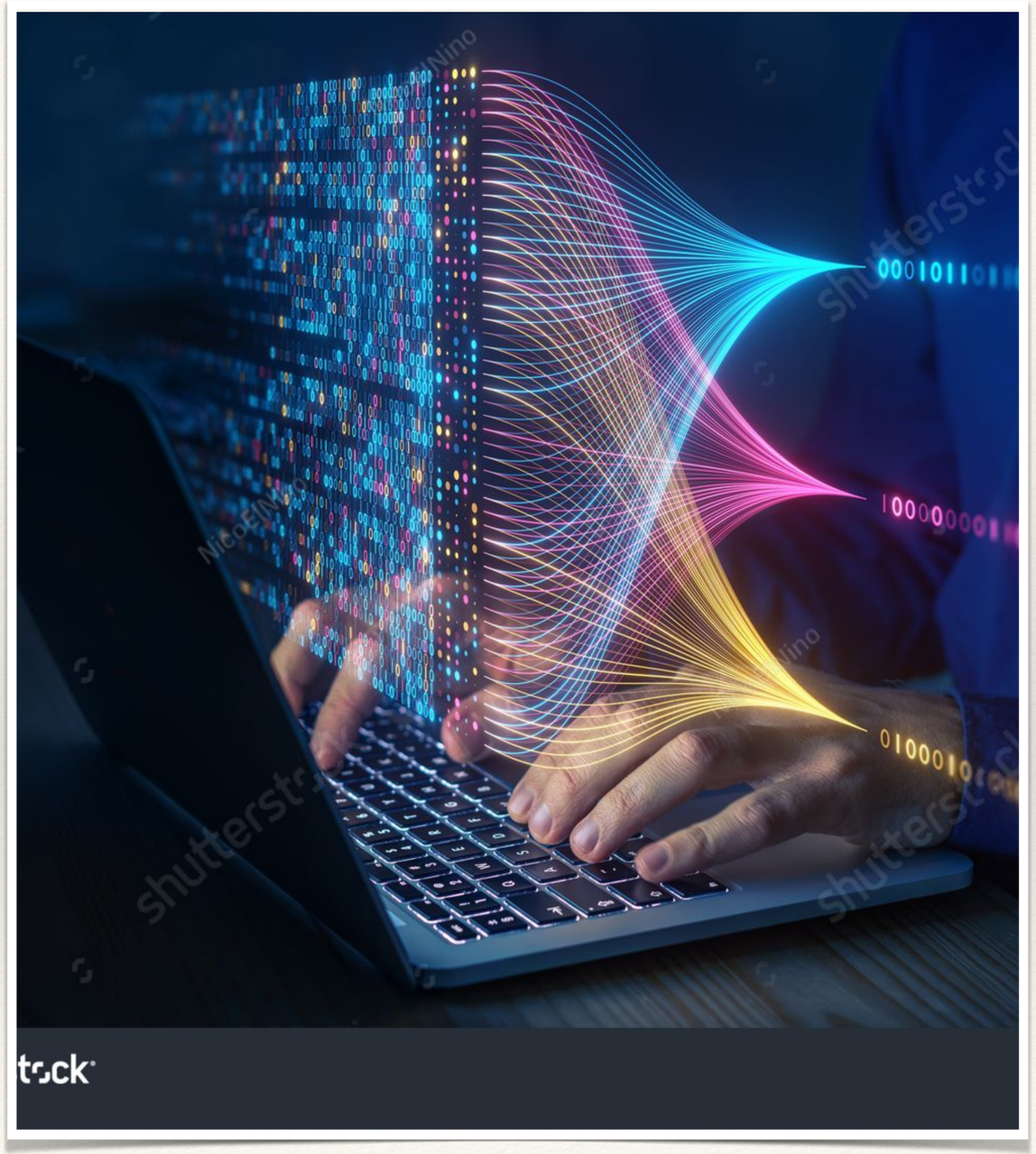
`Married` and `Accountants` are in non default circle



ANALYSIS BETWEEN 'ORGANIZATION_TYPE' & 'FLAG_OWN_CAR

Clients who are `Self-employed` and don't own `Car` are more in default circle as compared to clients who owns the car

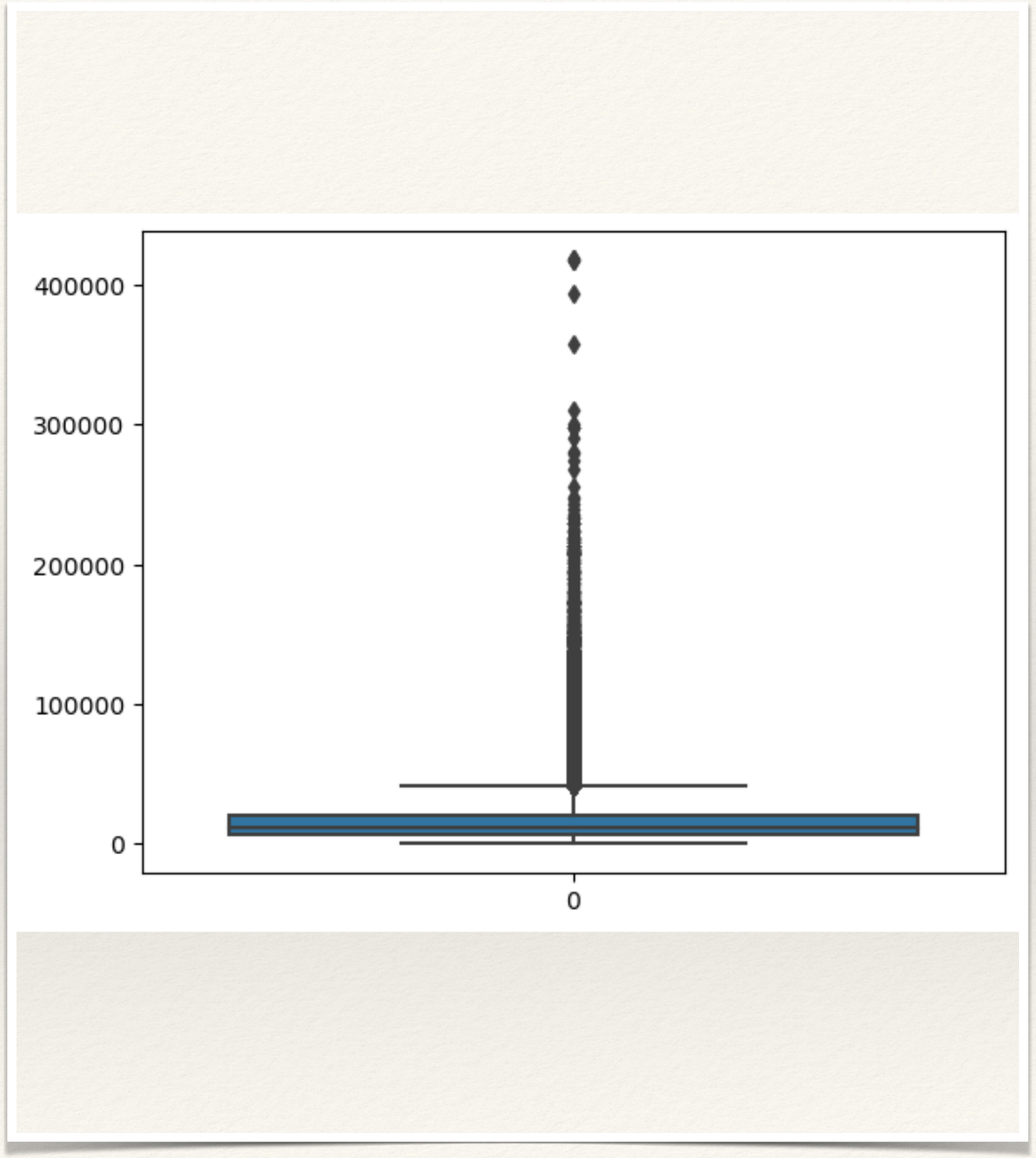
ANALYSIS Of Previous Loan Data



OUTLIERS ANALYSIS

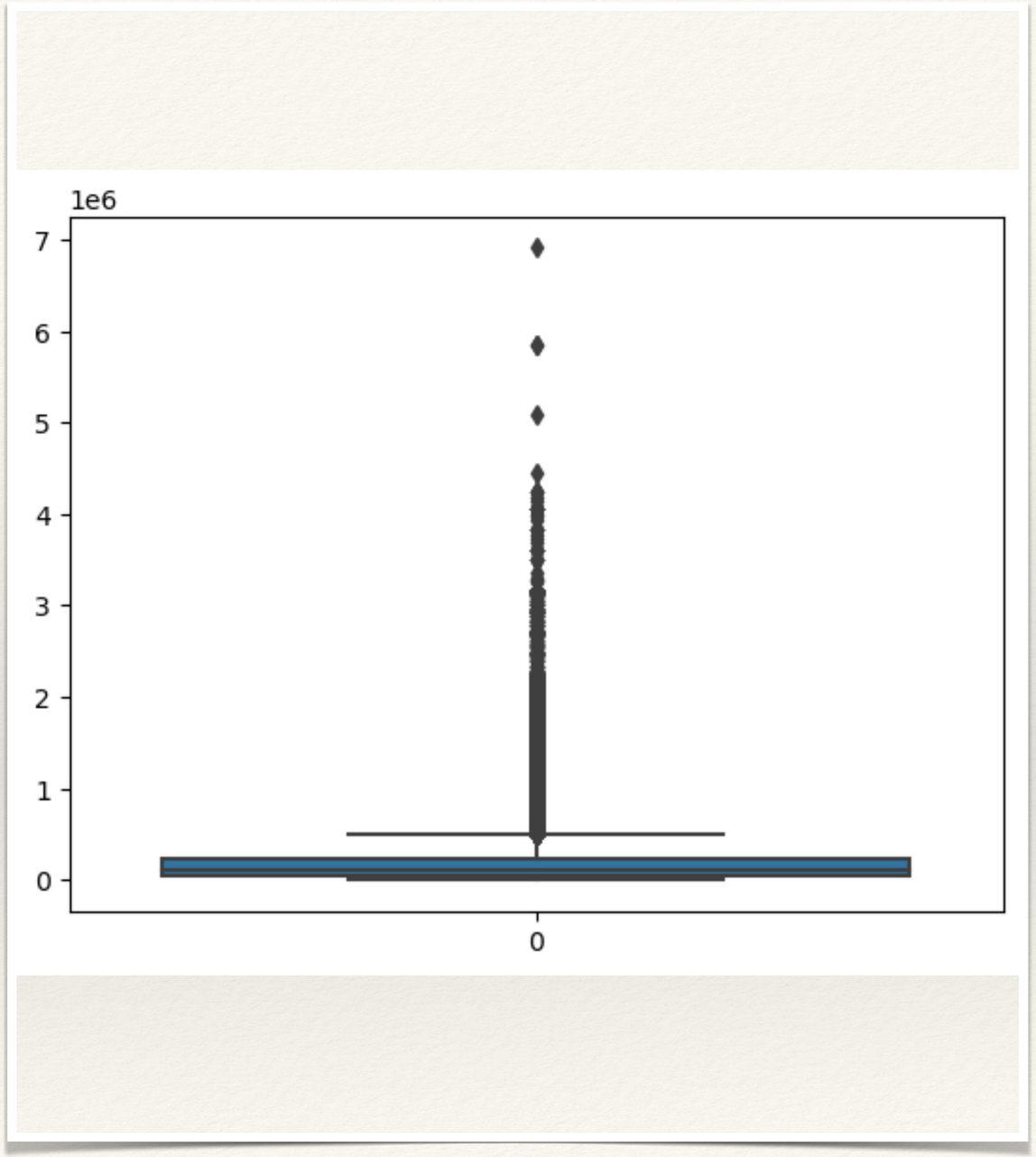
ANALYSIS 'AMT_ANNUITY'

AMT_ANNUITY` values above 42163.38 are outliers



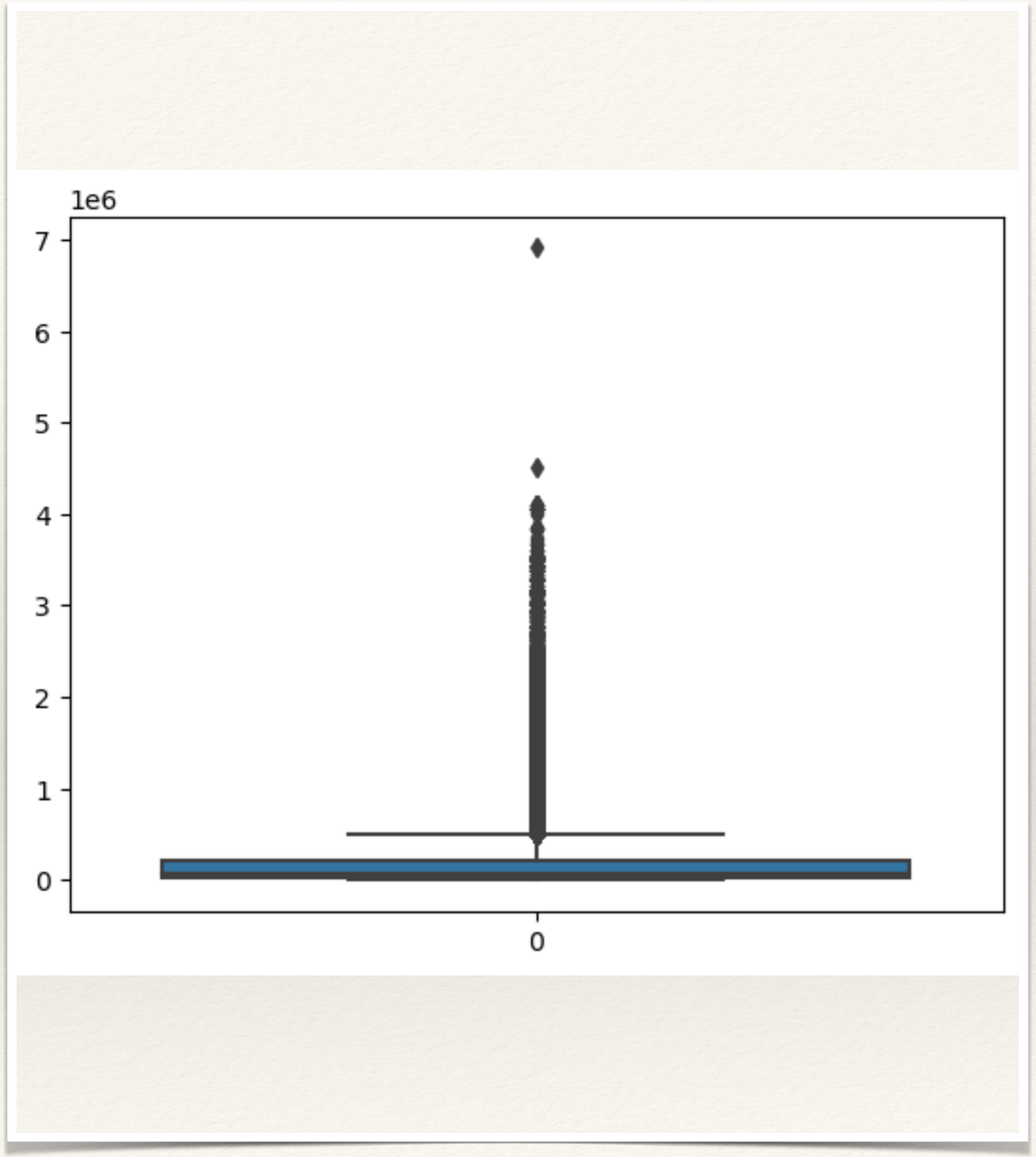
ANALYSIS 'AMT_GOODS_PRICE'

AMT_GOODS_PRICE` values above 508738.5 are outliers



Analysis 'AMT_CREDIT'

AMT_CREDIT` values above 504805.5 are outliers.



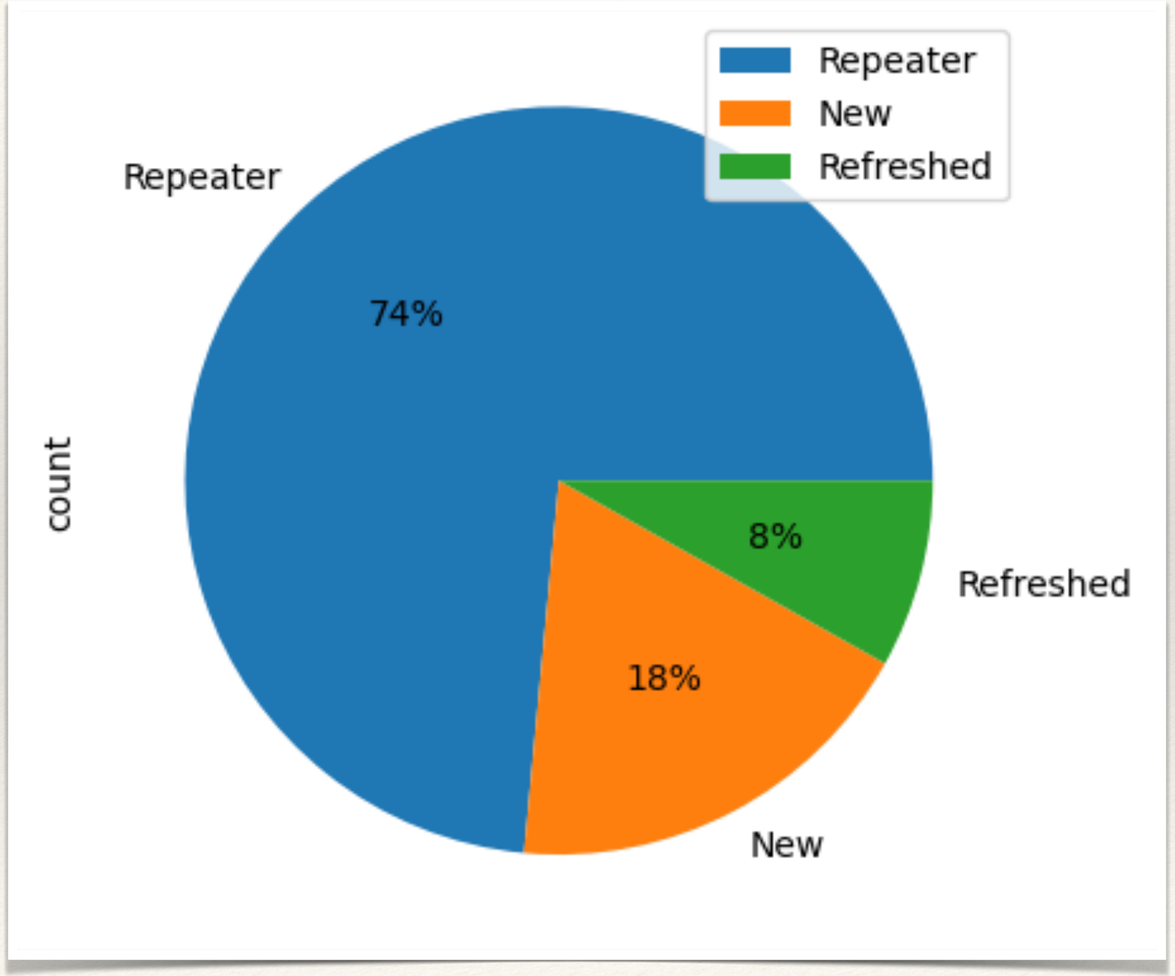
ANALYSIS On Merged Data About Clients Application_data And Previous_data

UNIVARIATE ANALYSIS

ANALYSIS On 'NAME_CLIENT_TYPE'

Repeater` client type is the highest among all loan applications

`New` client type is the second highest among all loan applications

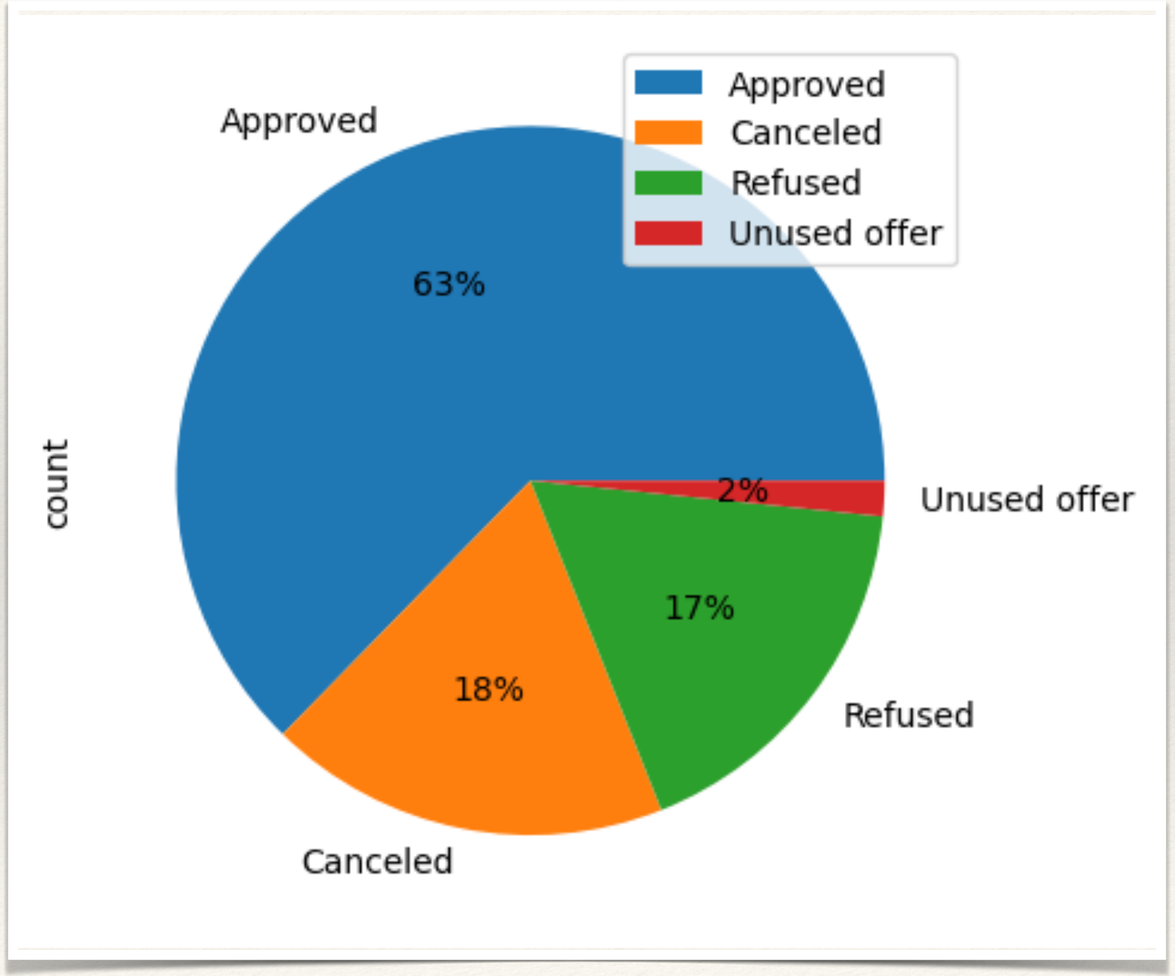


ANALYSIS On 'NAME_CONTRACT_STATUS'

Approved` loan status is the highest among all loan applications

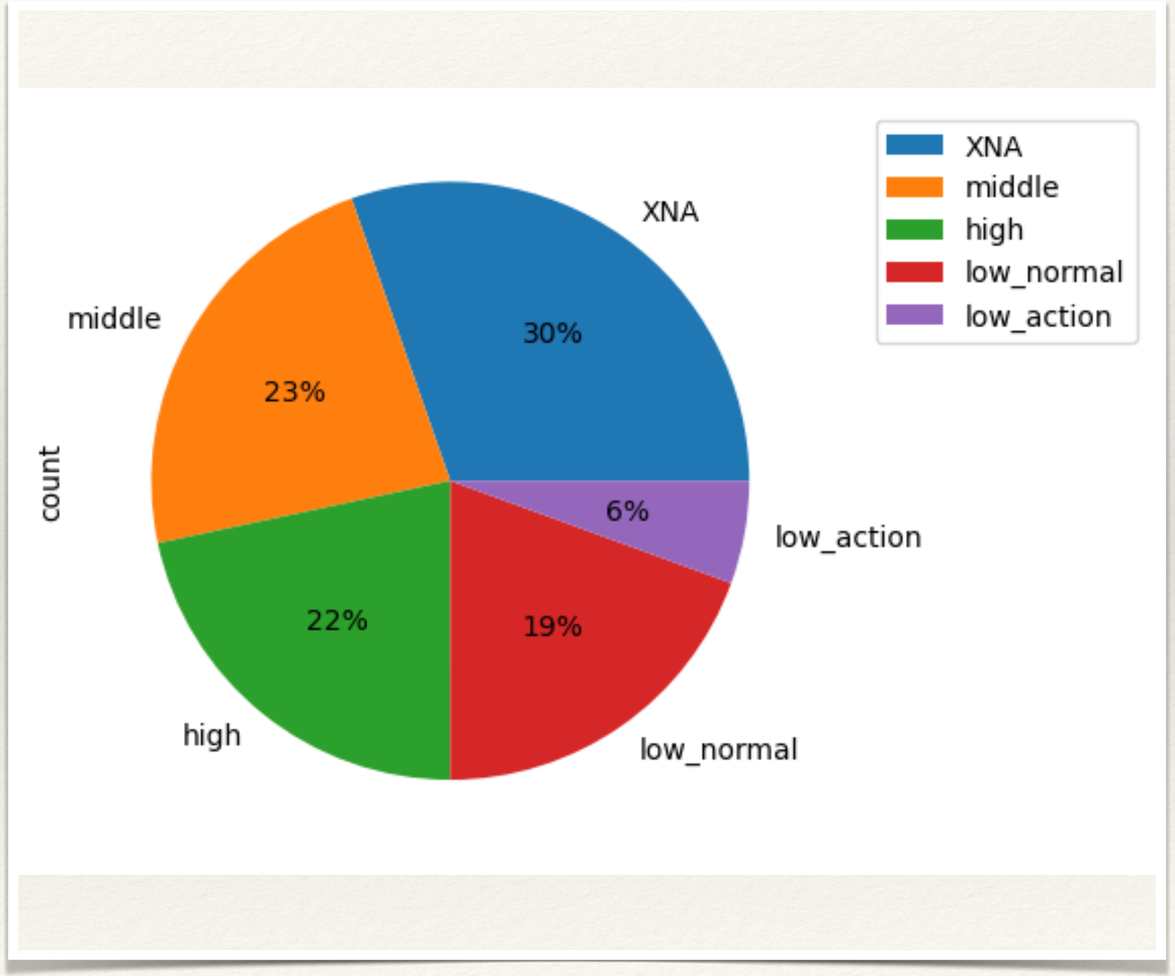
`Canceled` loan status is the second highest among all loan applications

As shown in the PIE plot 63% of the loans are approved 18% of the loans get canceled 17% of the loans gets refused .



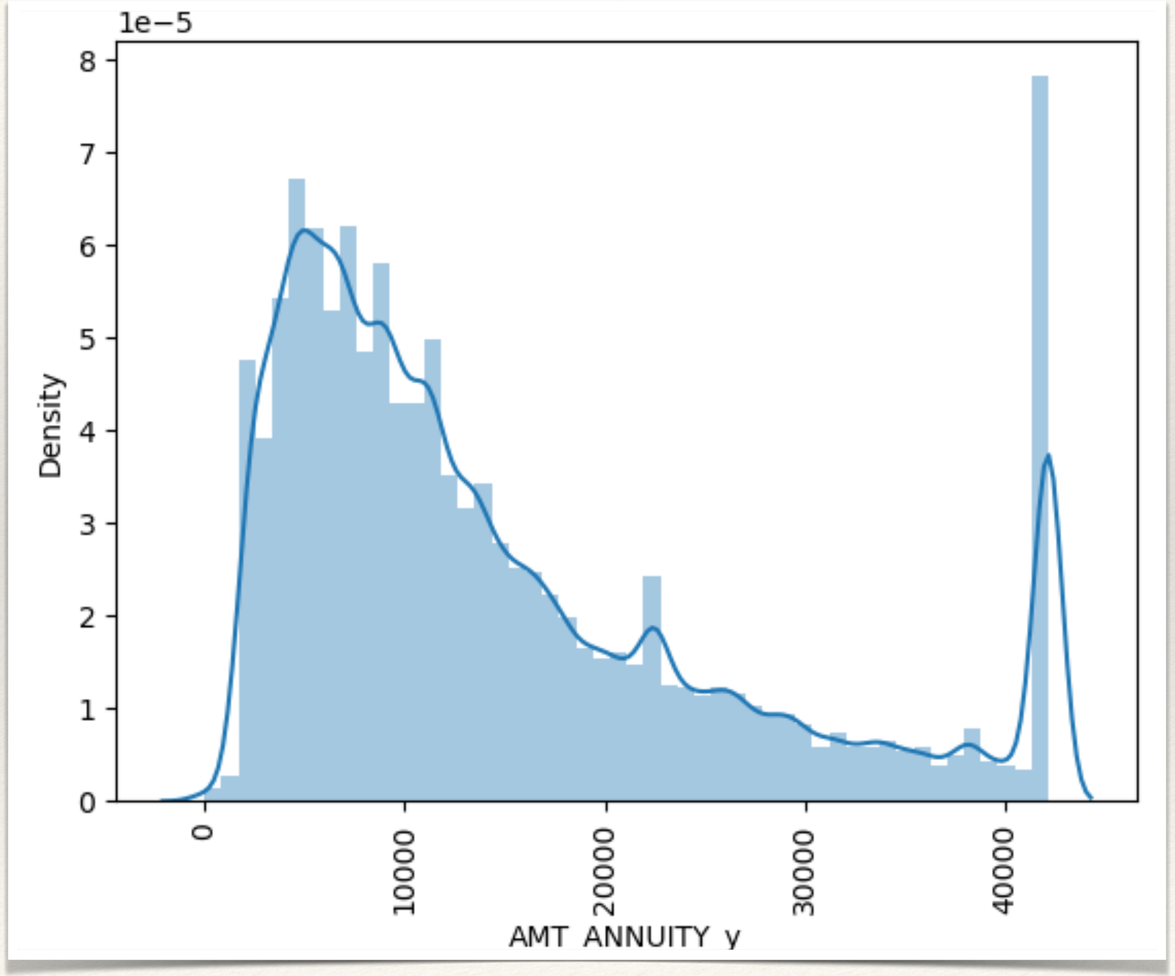
ANALYSIS On 'NAME_YIELD_GROUP'

XNA` interest rate is the highest among all loan applications.
`middle` and `high` interest rates are the second and third highest among all loan applications



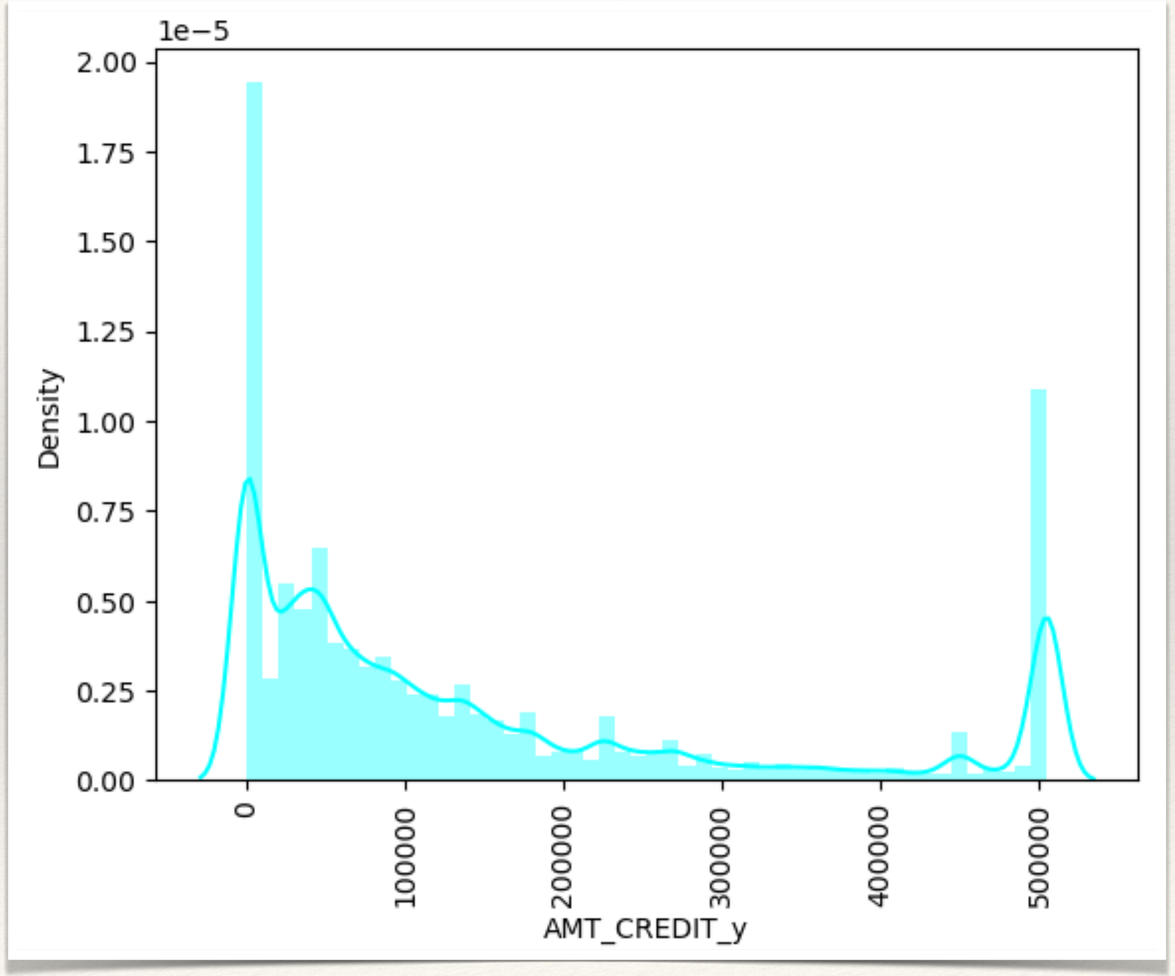
ANALYSIS On 'AMT_ANNUITY_y'

Huge spike on the loan applied can be seen at the beginning Most of the previous loan's annuity from the clients is less than 10,000 as the distribution is high here As previous loan's annuity increases, the no. of clients decreases



ANALYSIS On 'AMT_CREDIT_y'

- *Distribution shows most people received the loan amount that they applied for*



CORRELATION ANALYSIS

★ *AMT_ANNUITY_y has high correlation with AMT_APPLICATION and AMT_CREDIT_y*

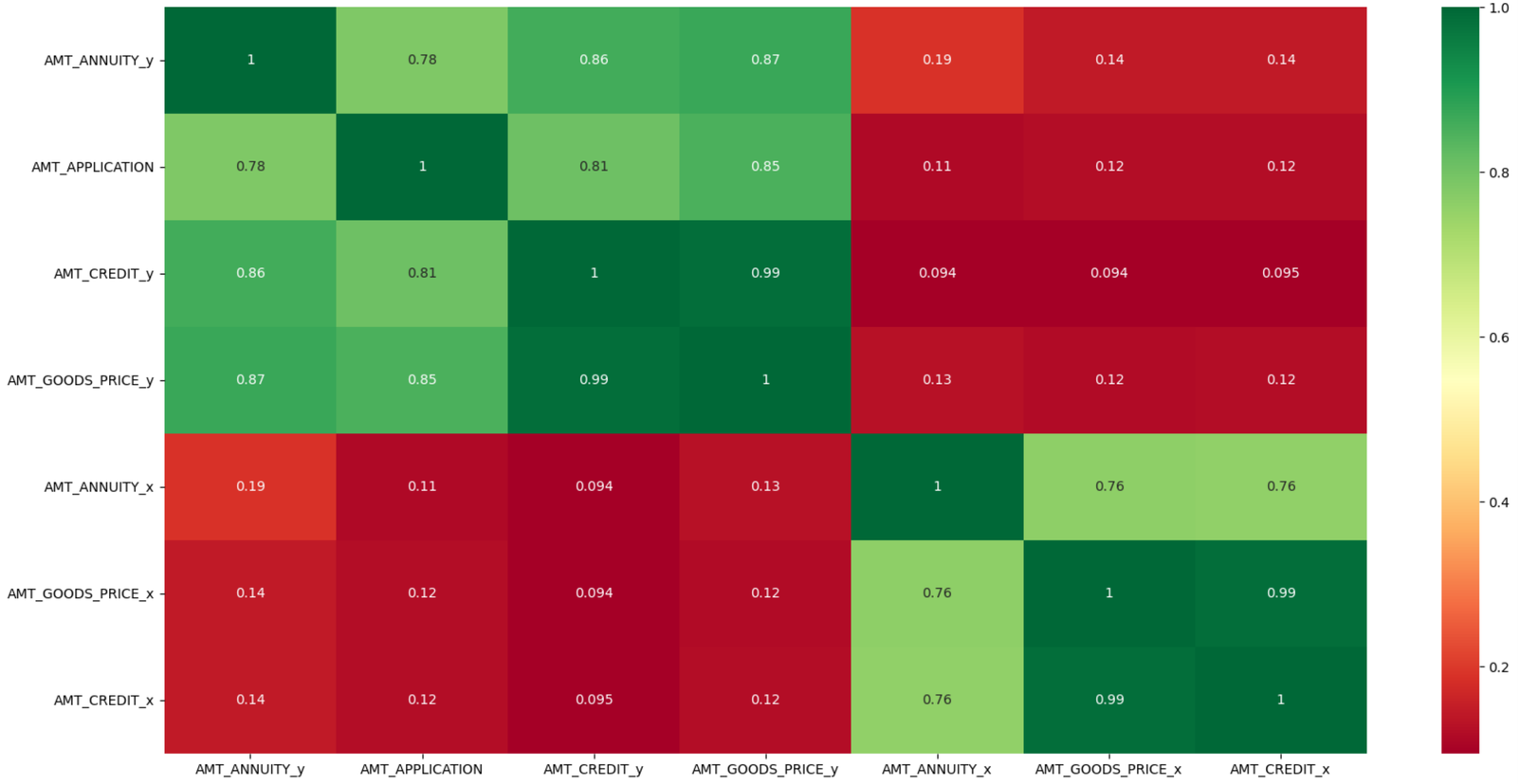
★ *AMT_APPLICATION has high correlation with AMT_GOODS_PRICE_Y and AMT_CREDIT_y*

★ *AMT_CREDIT_y has High correlation with AMT_GOODS_PRICE_Y and AMT_APPLICATION*

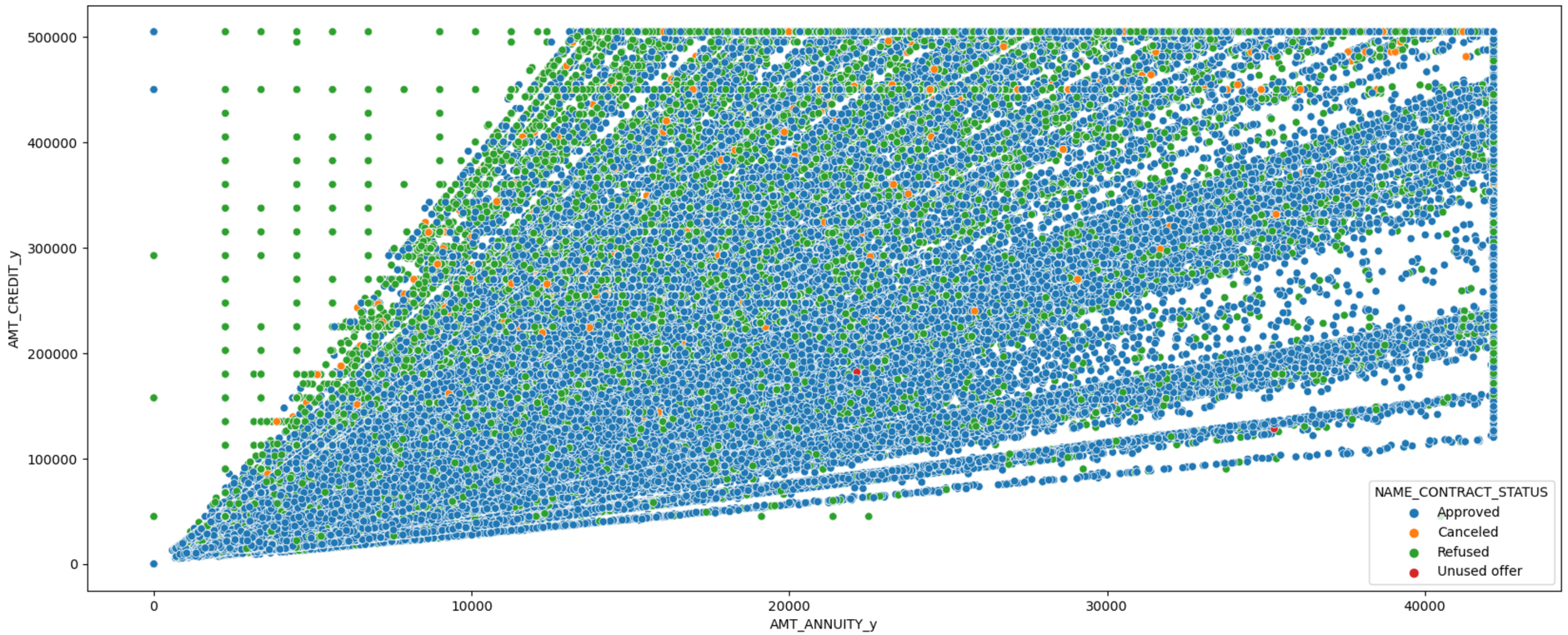
★ *AMT_ANNUITY_x has a high correlation with AMT_GOODS_PRICE_y, AMT_CREDIT_y*

★ *AMT_ANNUITY_x has a high correlation with AMT_GOODS_PRICE_x, AMT_CREDIT_x*

★ *AMT_CREDIT_x has a high correlation with AMT_GOODS_PRICE_x*



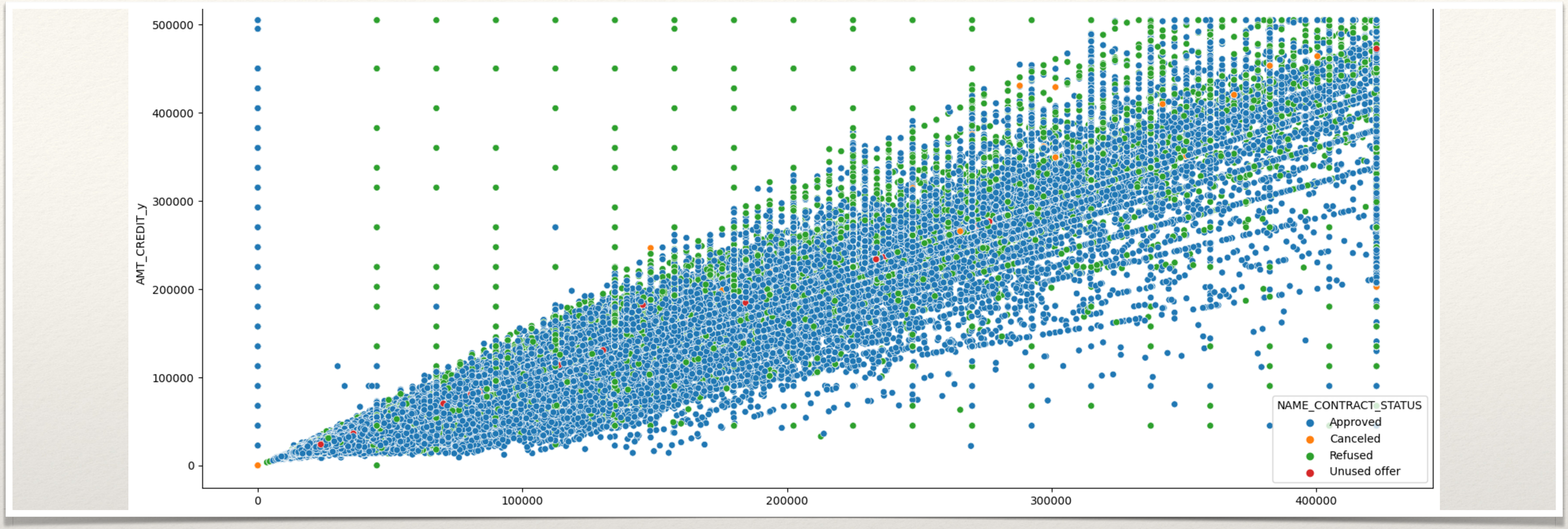
BIVARIATE/MULTIVARIATE ANALYSIS



ANALYSIS

'AMT_ANNUITY_y V/S AMT_CREDIT_y V/S NAME_CONTRACT_STATUS

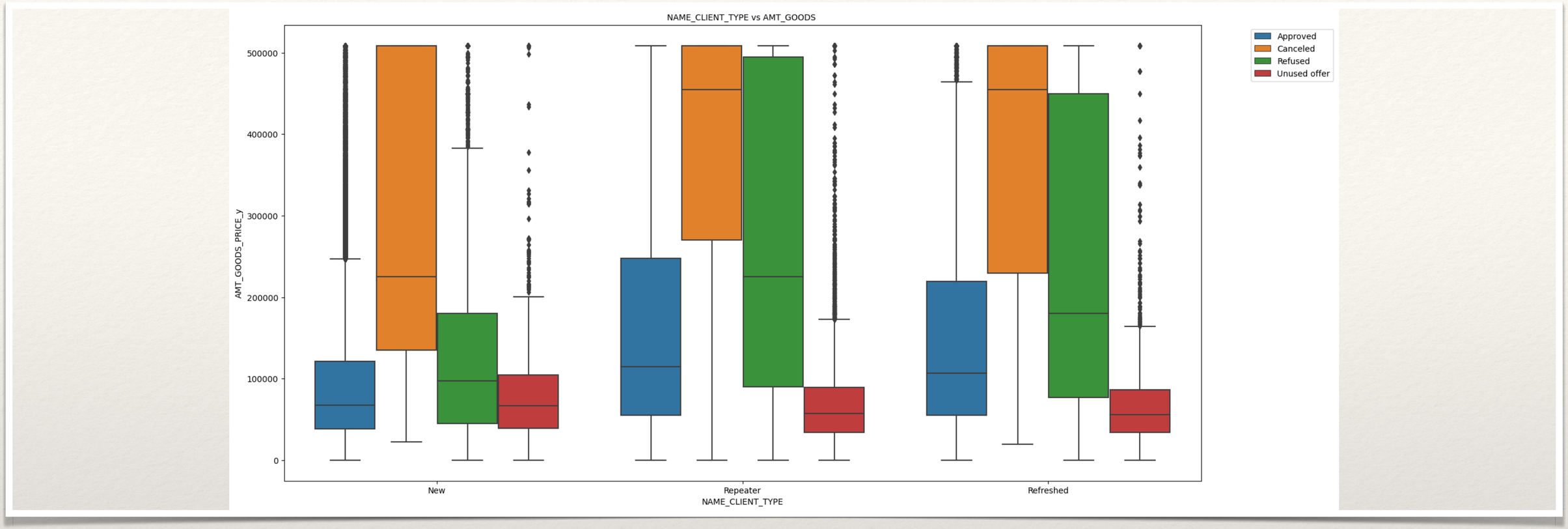
No correlations observed from scatterplot with respect to `NAME_CONTRACT_STATUS`
- `AMT_ANNUITY_y` has a strong correlation with `AMT_CREDIT_y`



ANALYSIS

'AMT_APPLICATION V/S AMT_CREDIT_y V/S NAME_CONTRACT_STATUS

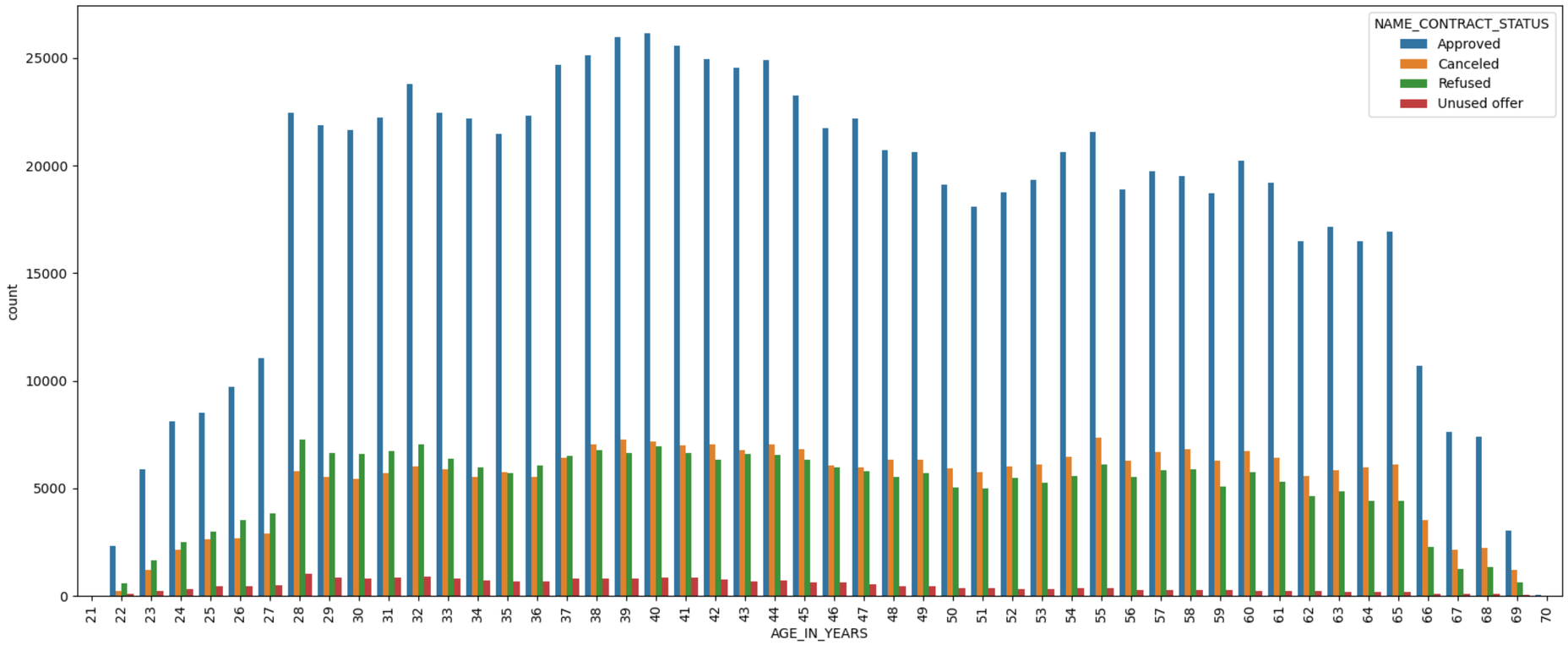
Application amount has strong positive correlation with Credit amount



ANALYSIS

'NAME_CLIENT_TYPE' vs 'AMT_GOODS_PRICE_y' vs 'NAME_CONTRACT_STATUS'

Clients who are Approved and New have less median goods price compared to Repeater and Refreshed



ANALYSIS

AGE_IN_YEARS V/S NAME_CONTRACT_STATUS

age range 30-40 get most approval followed by clients in 40-50 age range whereas age range 60-70 receive least approval followed by 22-27 age range

CONCLUSIONS

Client categories to be targeted for providing loan

1.Clients in the age range 30-40 and 40-50
2.Clients who are Married
3.Male clients with Academic degree
4.Female Clients are more in number for applying loans
5.Repeater clients
6.Banks must target more on contract type ‘Student’ ,’Pensioner’ and ‘Businessman’ for profitable business
7.Banks must focus less on income type ‘Working’ as it is has most number of unsuccessful payments in order to get rid of financial loss for the organisation

Clients categories for denying loans

1.People with more children have greater difficulties in payment
2.People with higher age have greater difficulties in payment (AGE > 60)
3.People with less employment days have greater difficulties in payment
4.People with more family members have greater difficulties in payment
5.People with lower Education have greater difficulties in payment
6.Based on Organisation type Business people and self employed have greater difficulties in payment
7.Loan Purpose on 'Repair' having highest number of unsuccessful repayments.
8.Housing Type 'With Parents' having least number of unsuccessful repayments.

Reducing the amount of loan

1.Based on the Income type the loan amount can be reduced /increased.