CREDIT EDA CASE STUDY

By Tushar Joshi and Smrity Panda.

PG Diploma Data Science- DS C27 Batch.

PROBLEM STATEMENT

This case has been undertaken to perform EDA in a real business scenario. In this case study, we apply techniques of Exploratory Data Analysis.We will also Highlight the risk analytics in banking and financial services, and understand how data is used to minimise the risk of losing money while lending to customers.

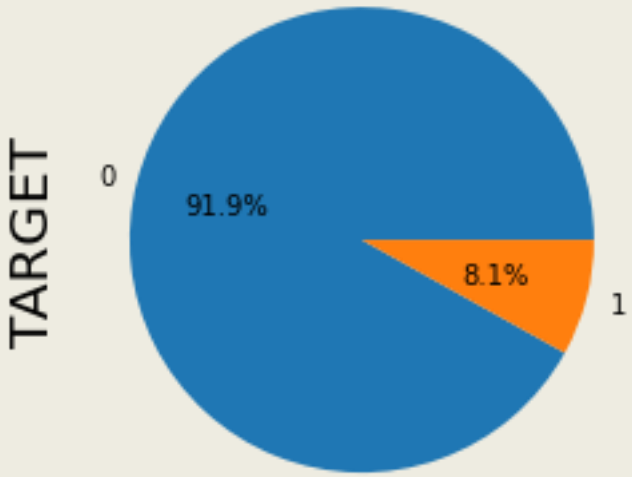
Business Understanding:

Using EDA from to analyse the patternsin the dataset, and figure out, which all clients can repay the loan, based on their credit history and various other factors. The loan providing companies can use this data and come to an inference about whether the lient will be defaulter or not.

Finding out if the client has any payment difficulties.

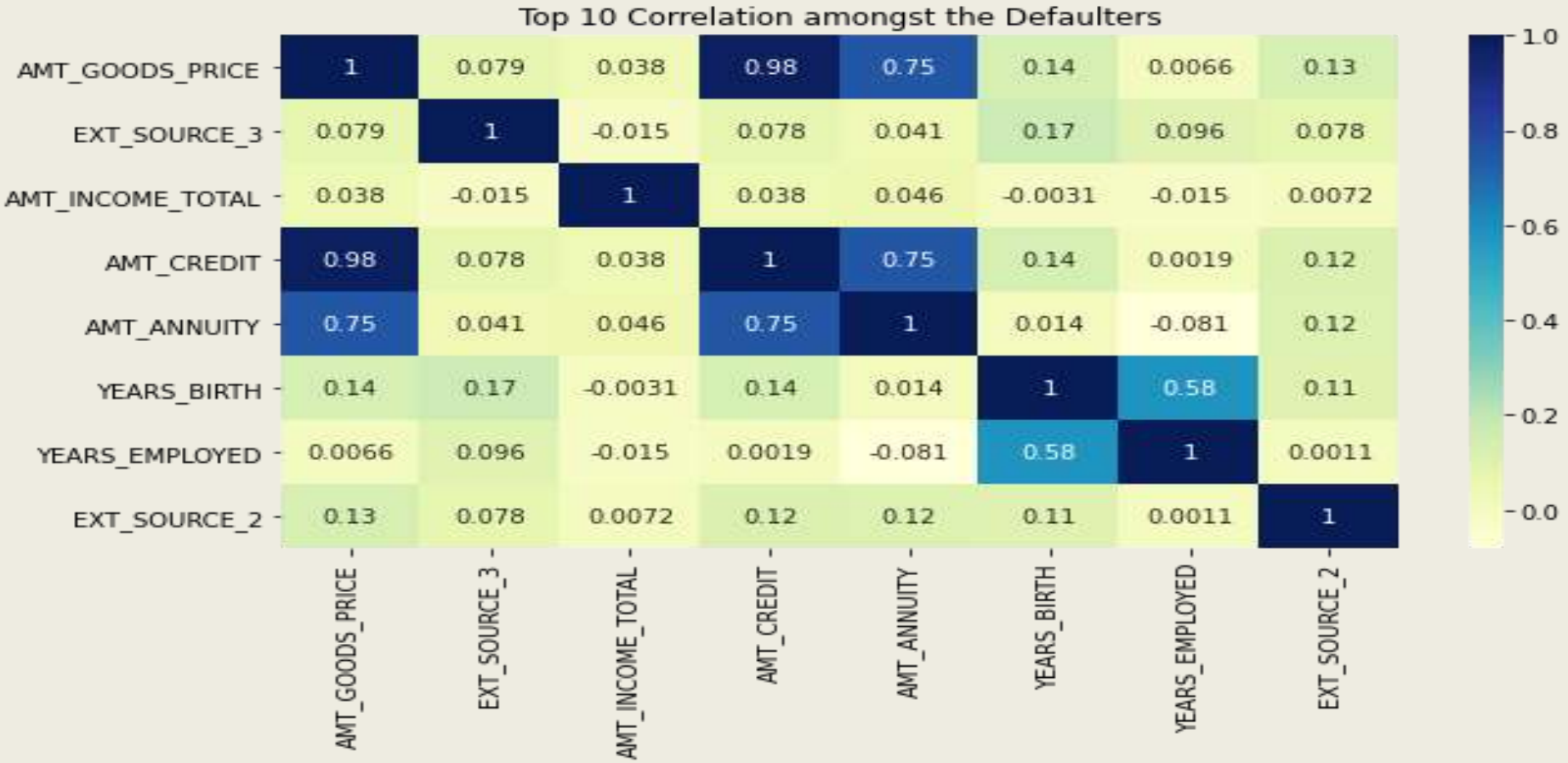
And making the reuired decision vis a vis the loan, which is to **Approve, Cancel, Refuse** or remains **Unused**.

ANALYSING THE DATA IMBALANCE



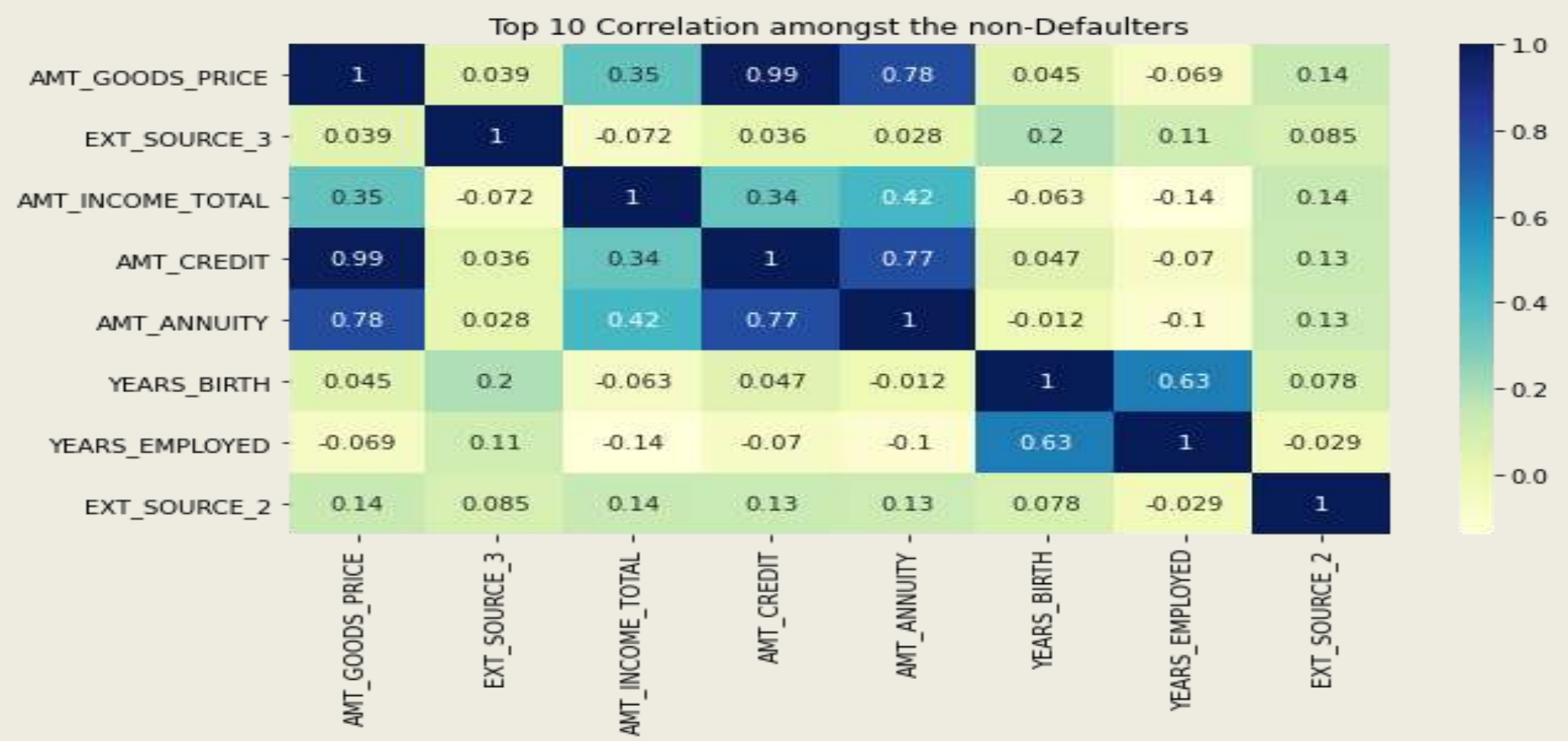
FROM THE PIE CHART WE OBSERVE THAT THERE IS HUGE IMBALANCE BETWEEN THE PERCENTAGE OF DEFAULTERS WITH RESPECT TO NON DEFAULTERS.

SO THE DATA IMBALANCE RATIO IS COMING TO BE 11.39:1



TOP CORRELATIONS ARE

1)AMT_CREDIT & AMT_GOODS_PRICE
2)AMT_CREDIT & AMT_ANNUITY
3)AMT_GOODS_PRICE & AMNT_ANNUITY
4)YEARS_BIRTH & YEARS_EMPLOYED
5)AMNT_ANNUITY & AMNT_INCOME_TOTAL
6)AMNT_INCOME_TOTAL & AMNT_CREDIT

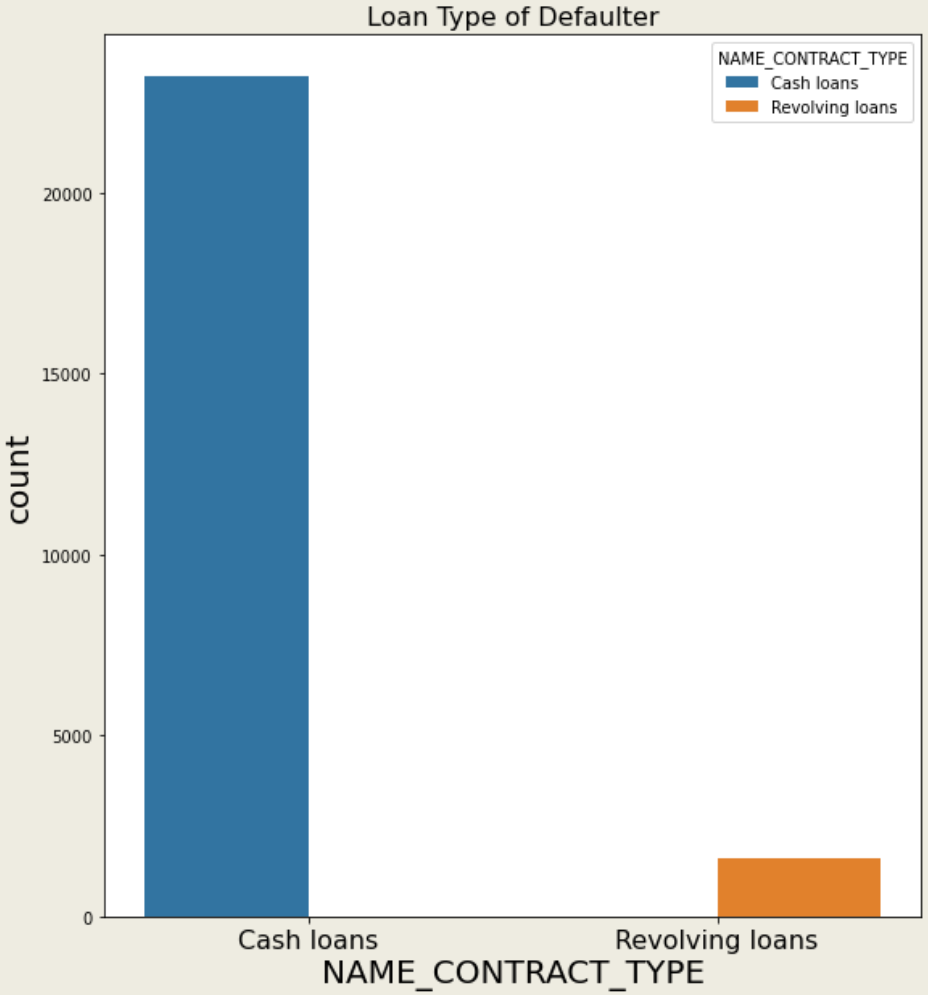


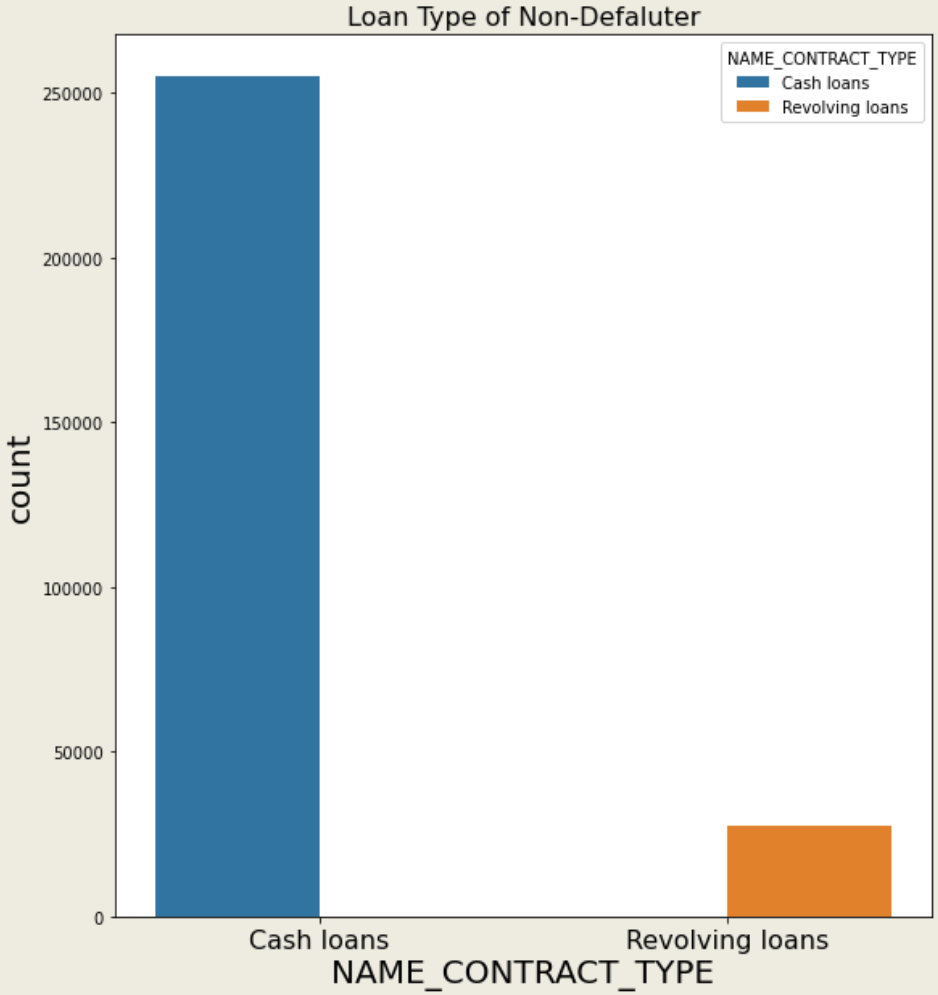
TOP CORRELATIONS ARE

1)AMT_CREDIT & AMT_GOODS_PRICE
2)AMT_GOODS_PRICE & AMNT_ANNUITY
3)AMT_CREDIT & AMT_ANNUITY
4)YEARS_BIRTH & YEARS_EMPLOYED
5) AMNT_INCOME_TOTAL & AMNT_CREDIT

UNIVARIATE ANALYSIS OF CATEGORICAL COLUMNS

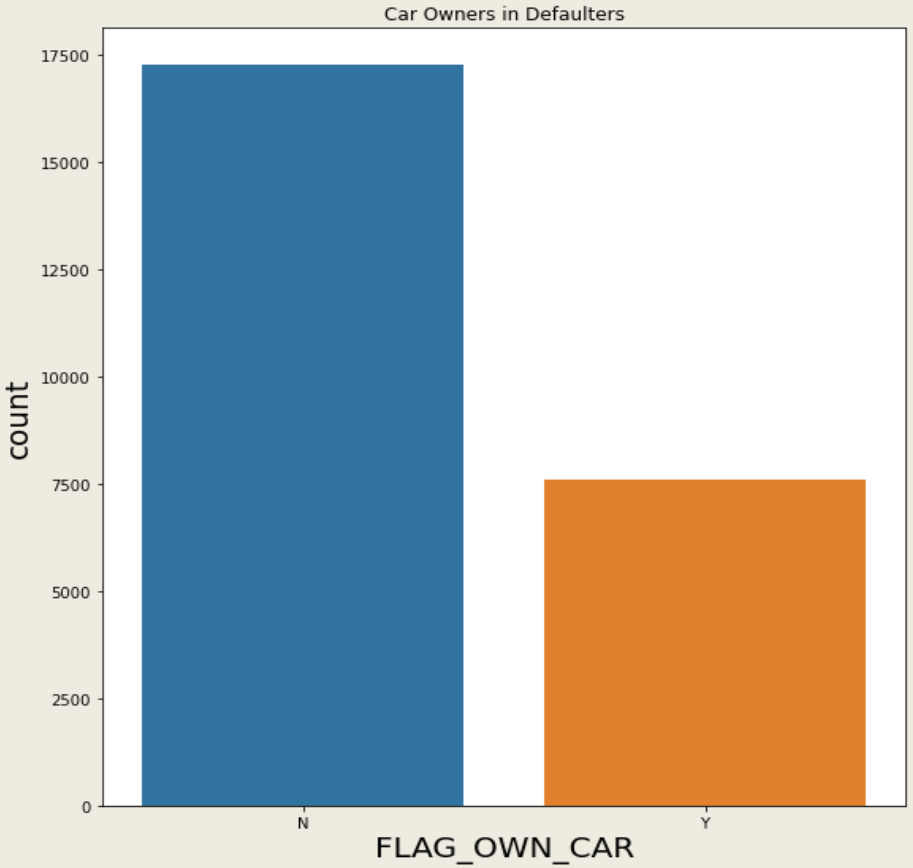
plotting based on 'NAME_CONTRACT_TYPE'

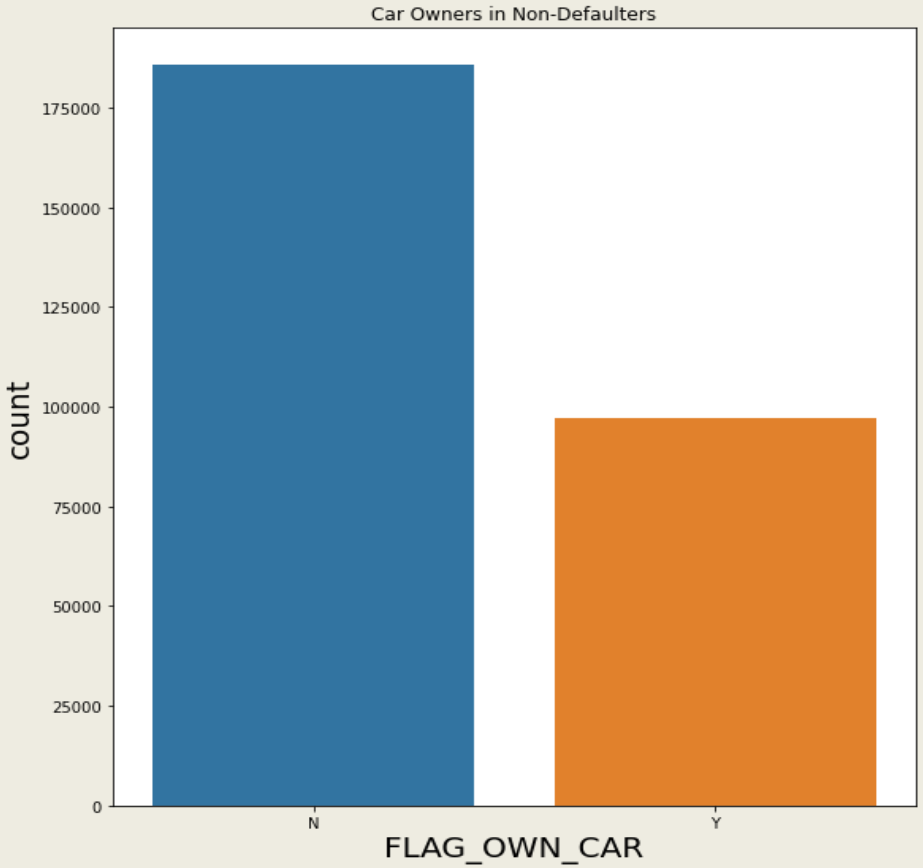




WE CAN SEE FROM THE GRAPH THAT CASH LOANS ARE HIGH IN NUMBER IN THOSE WHO HAVE DEFAULTED AND IN THOSE WHO HAVE NOT DEFULTED. REVOLVING LOANS ARE PREFERRED MUCH LESS BY THE CUSTOMERS

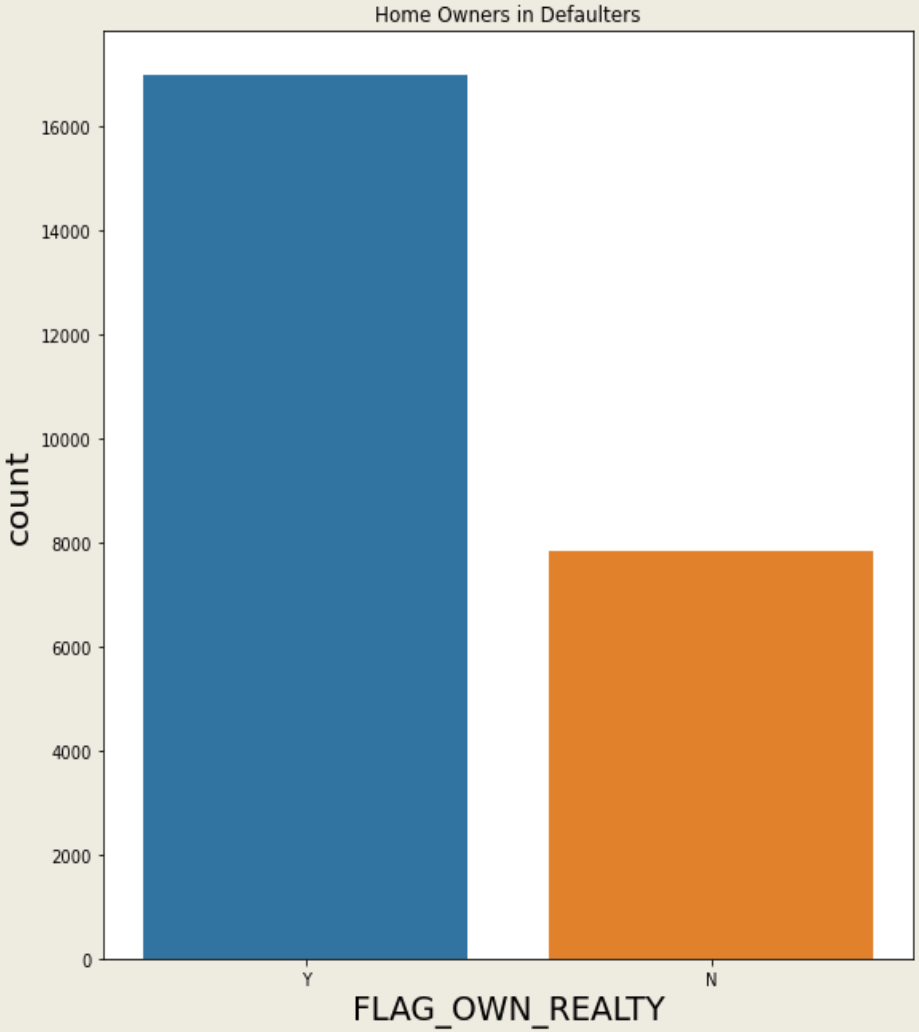
CAR OWNERSIP OF THE APPLICANTS

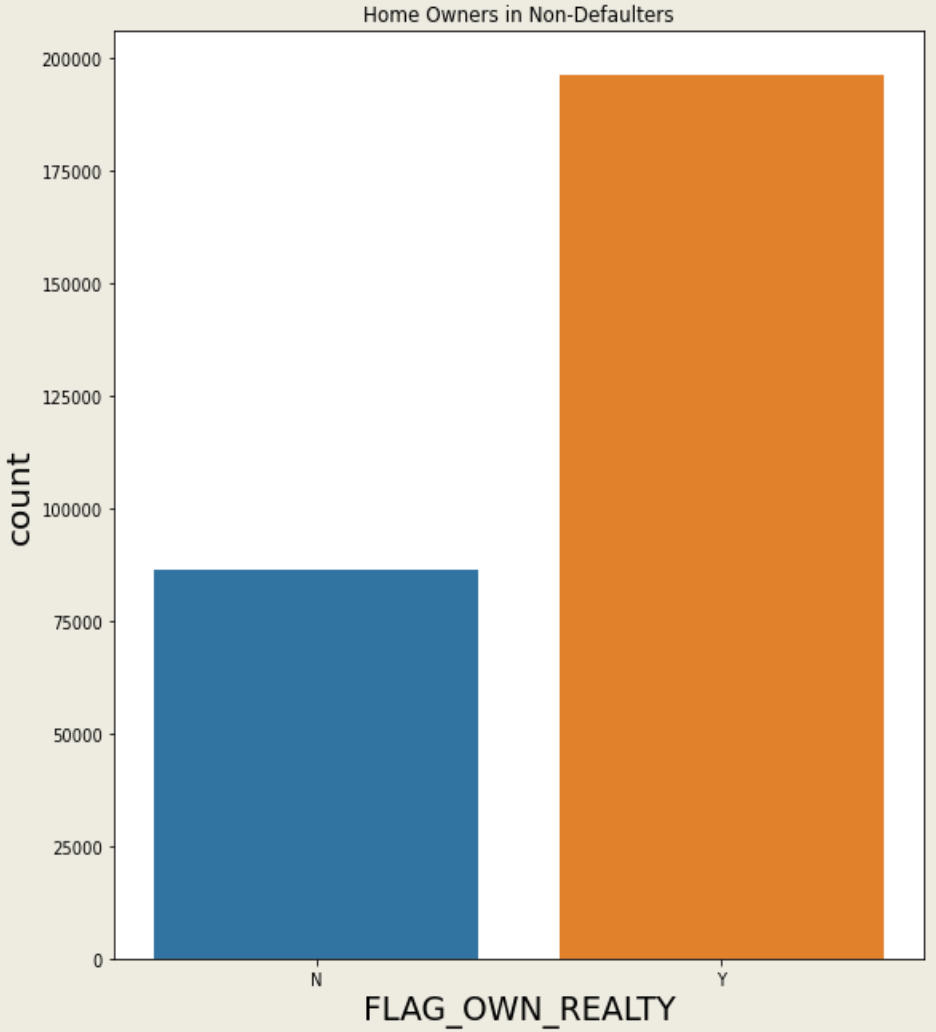




WE CAN SEE FROM THE GRAPH THAT A VERY HIGH NUMBER OF PEOPLE, AMONGST THE DEFAULTERS AND NON-DEFAULTERS DO NOT OWN A CAR

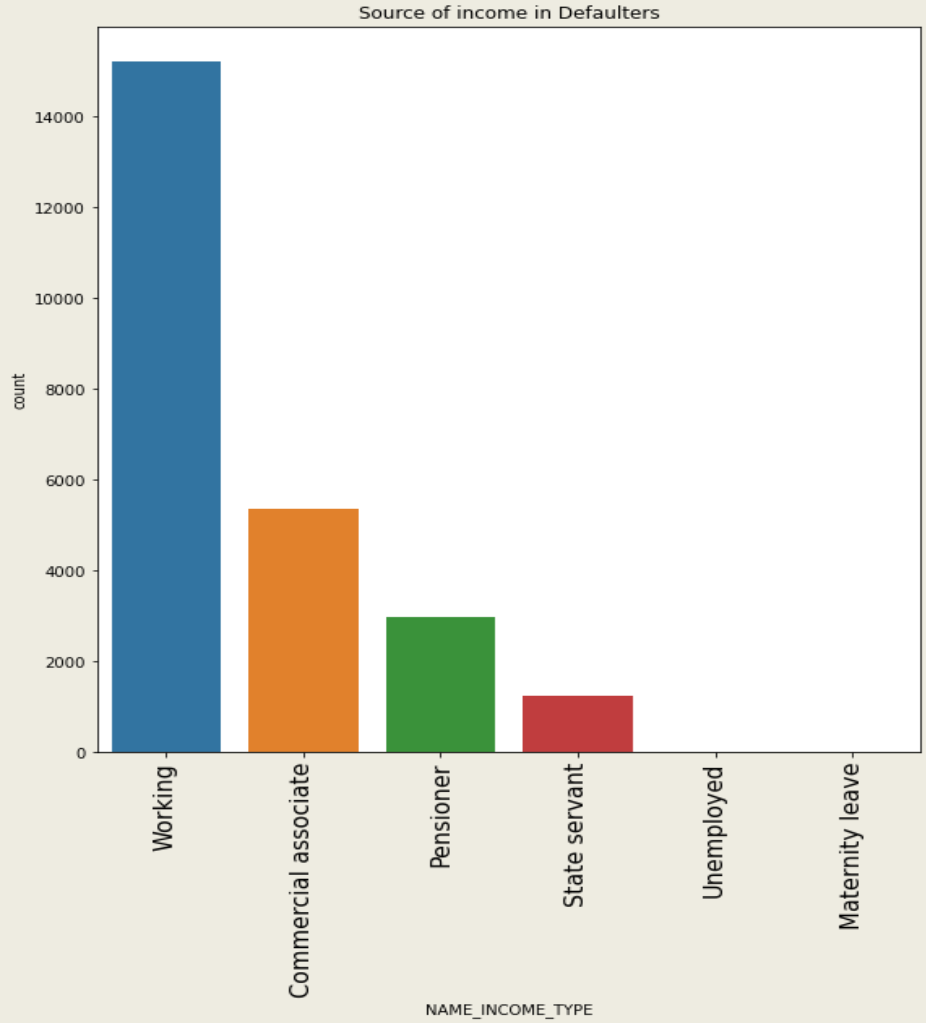
HOME OWNERSIP OF THE APPLICANTS

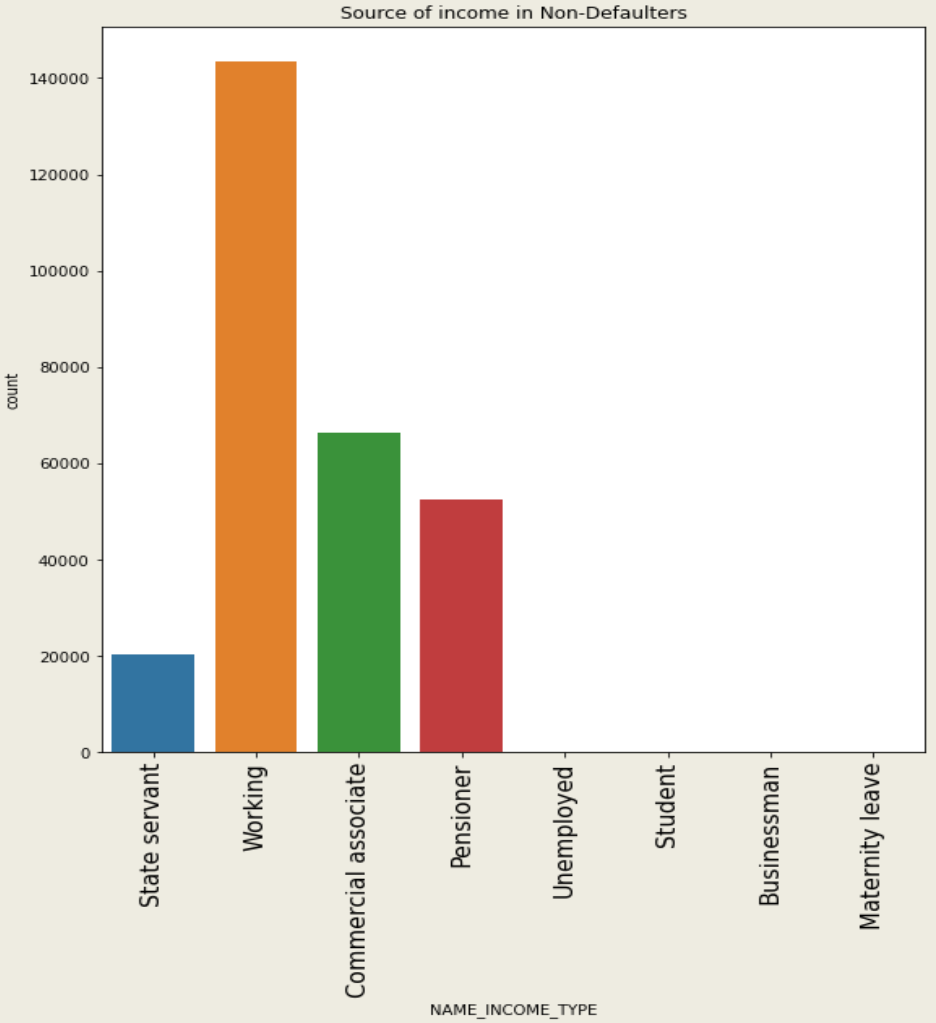




WE CAN SEE THAT AMONGST THE NON DEFAULTERS, THERE ARE MORE PEOPLE WHO OWN A HOME, BOTH AMONG DEFAULTERS AND NO-DEFAULTERS.

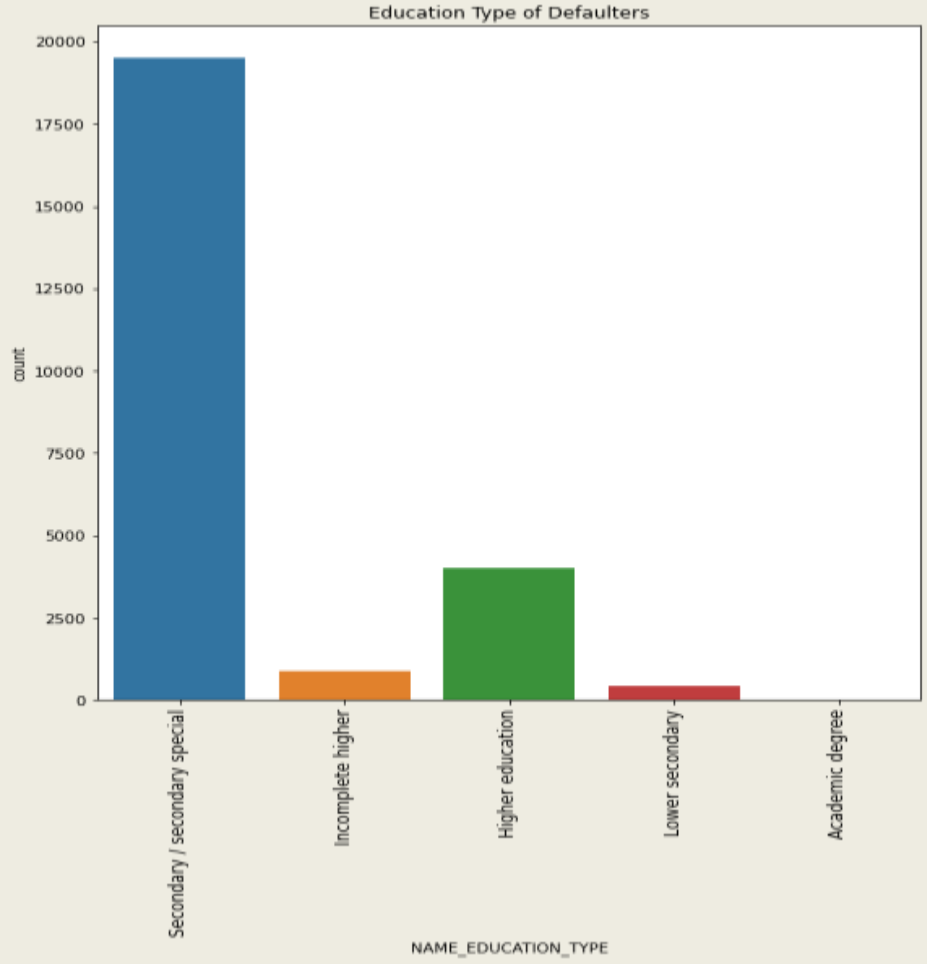
INCOME TYPE OF THE APPLICANTS

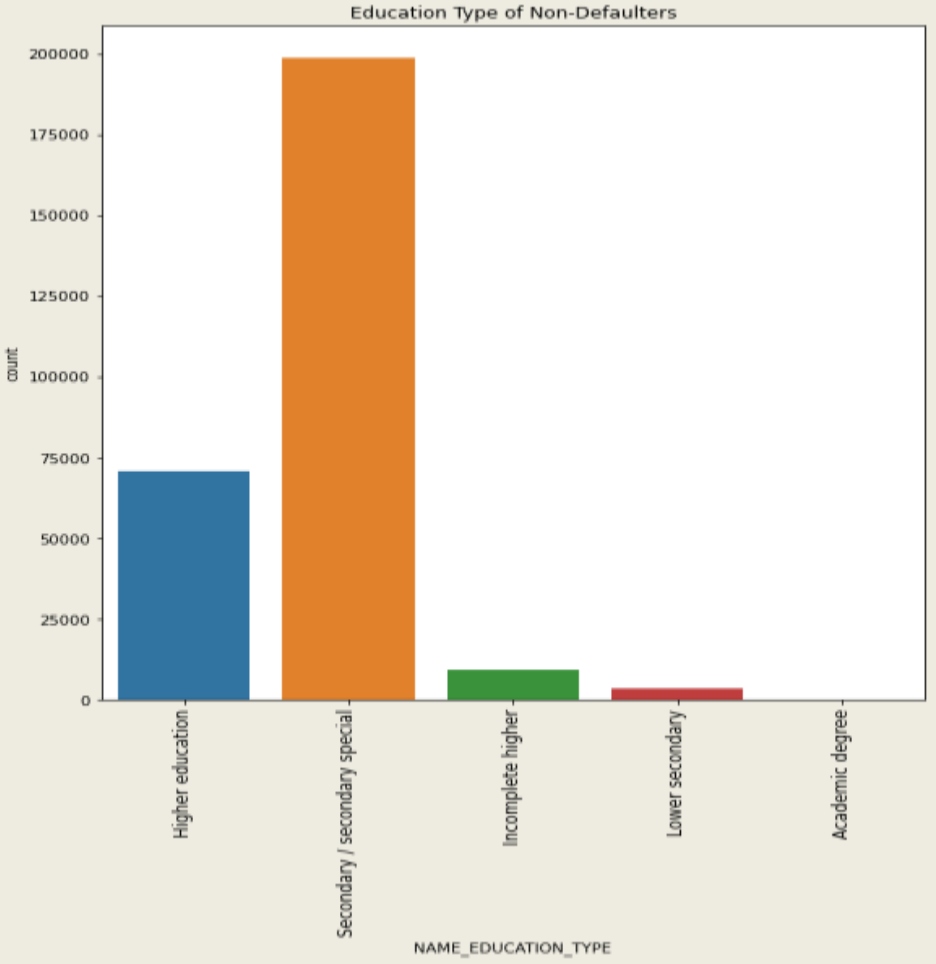




WE SEE THAT BOTH AMONGST THE DEFAULTERS AND NON DEFAULTERS THE PEOPLE WITH DAY JOBS ARE HIGHEST.

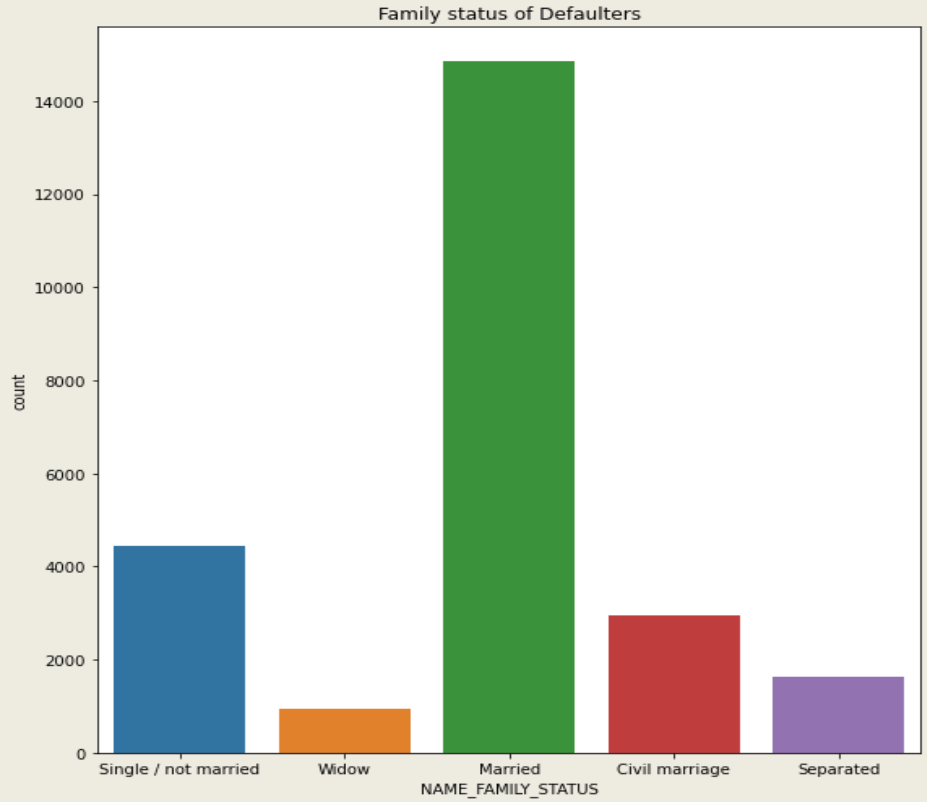
EDUCATION LEVEL OF THE APPLICANTS

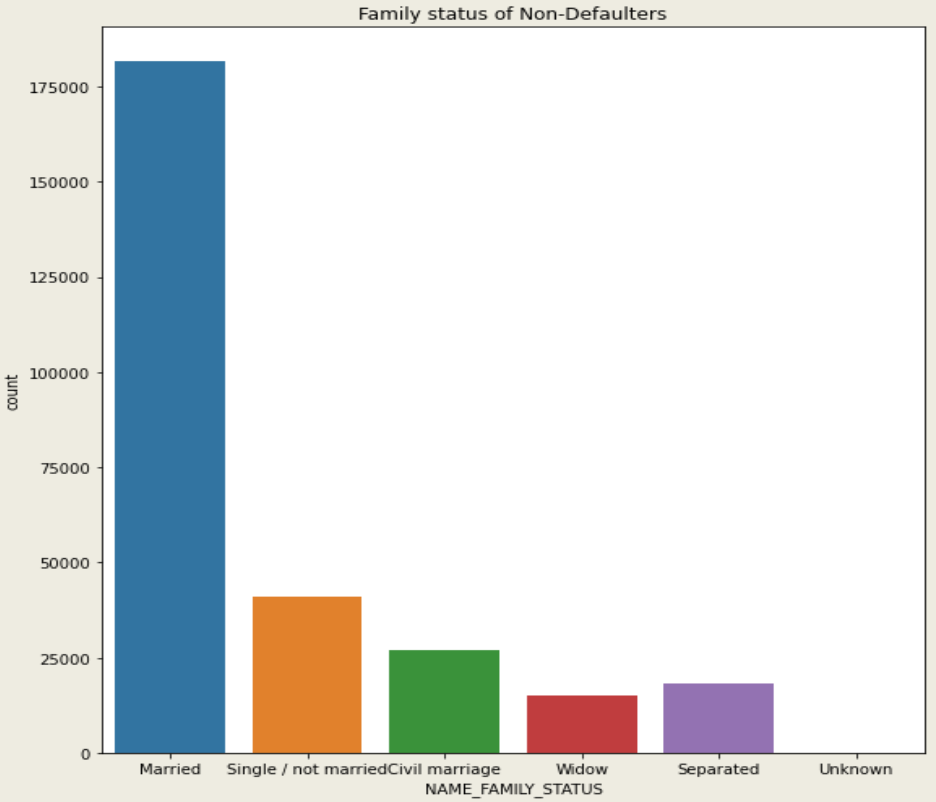




THE EDUCATION TYPE OF BOTH THE DEFAULTERS AND NON- DEFAULTERS SEEMS TO BE SECONDARY/SECONDARY SPECIAL. WHERE AS FOR NON DEFAULTERS IT IS LOWER SECONDARY OR INCOMPLETE HIGHER.

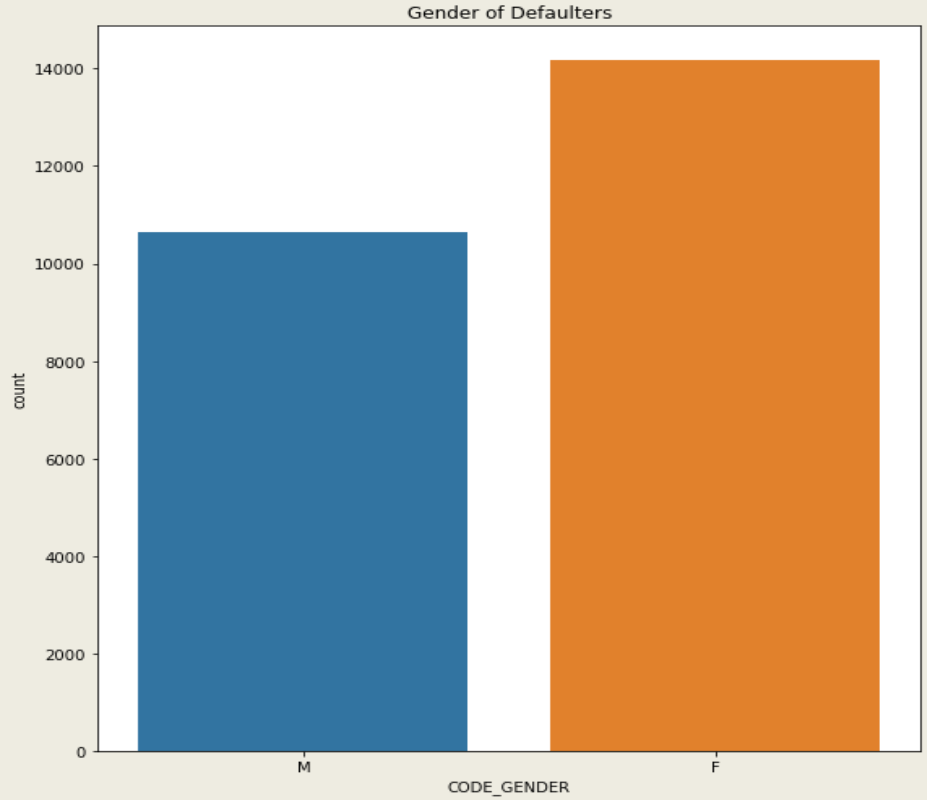
FAMILY STATUS OF THE APPLICANTS

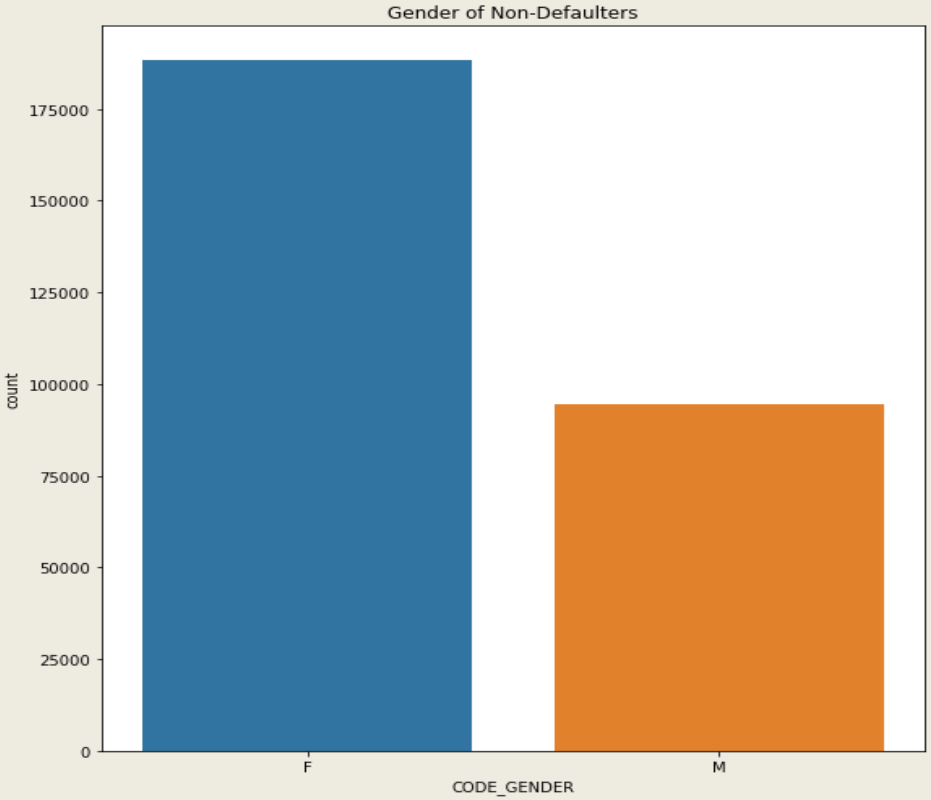




THE APPLICANTS WHO ARE MARRIED ARE THE HIGHEST NUMBER OF DEFAULTERS AND NON DEFAULTERS. WHEREAS WIDOWS ARE THE LOWEST NUMBERS OF DEFAULETRS AS WELL AS NON- DEFAULTERS.

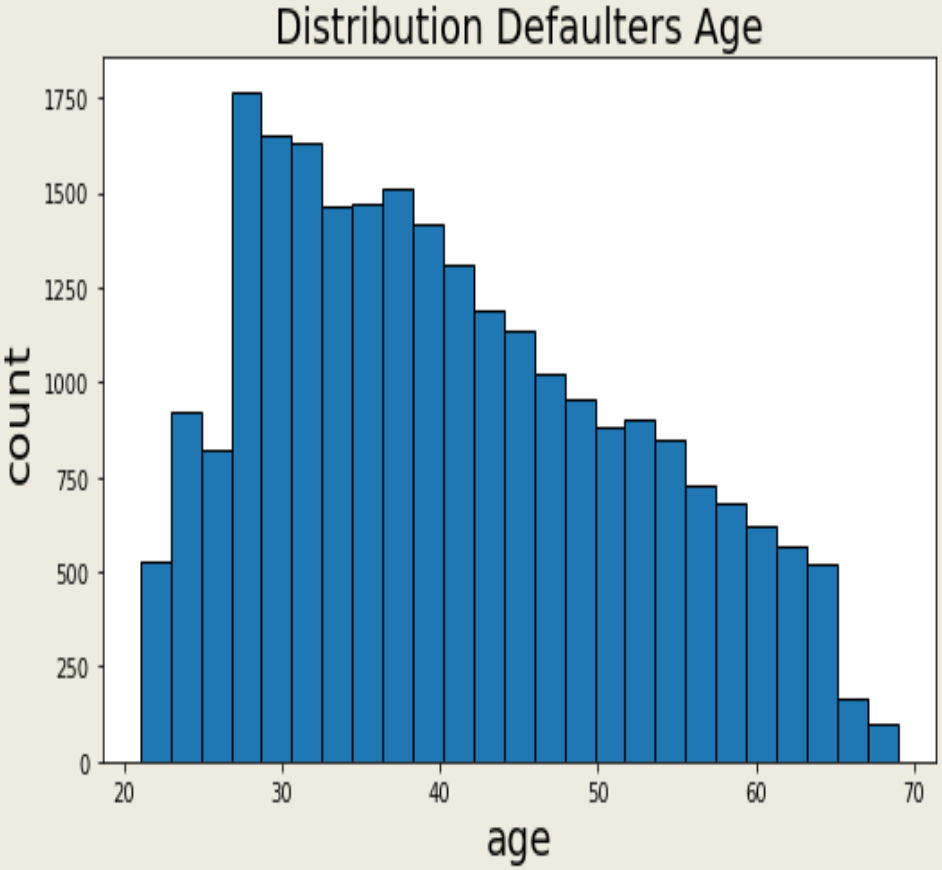
GENDER OF APPLICANTS

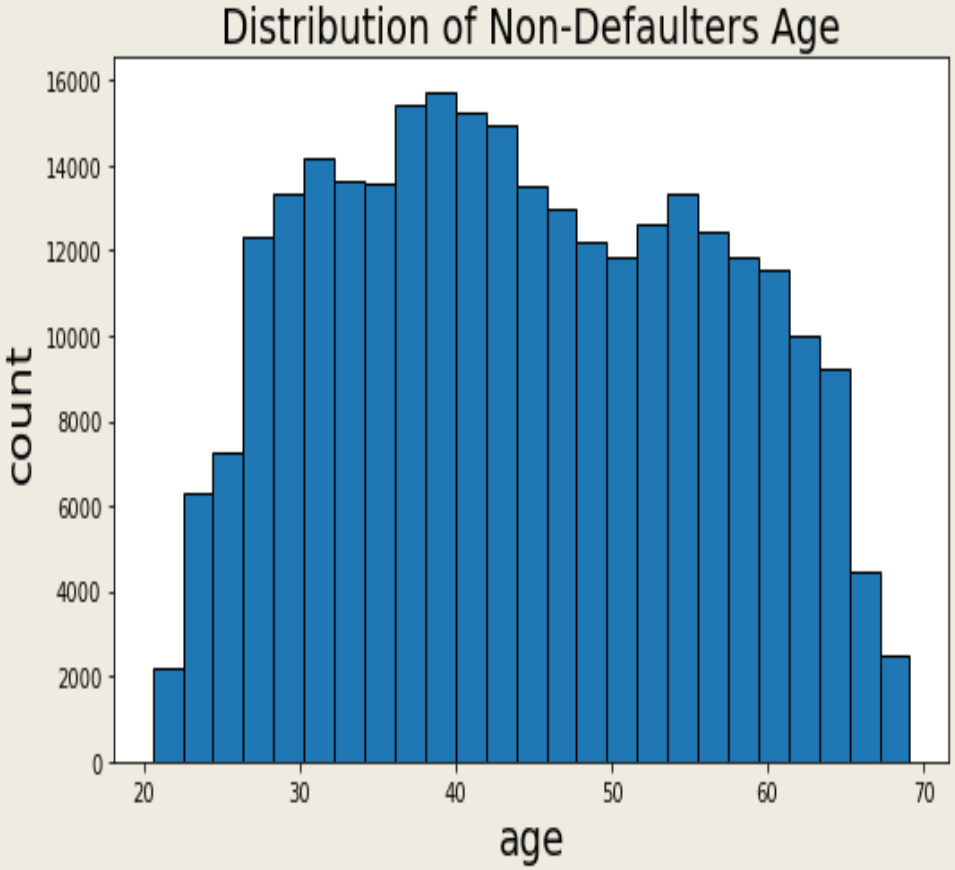




IN BOTH THE CASES THE NUMBER OF FEMALES IS HIGH, MEANING, IT IS THE FEMALES THAT HAVE APPLIED FOR MOST LOANS

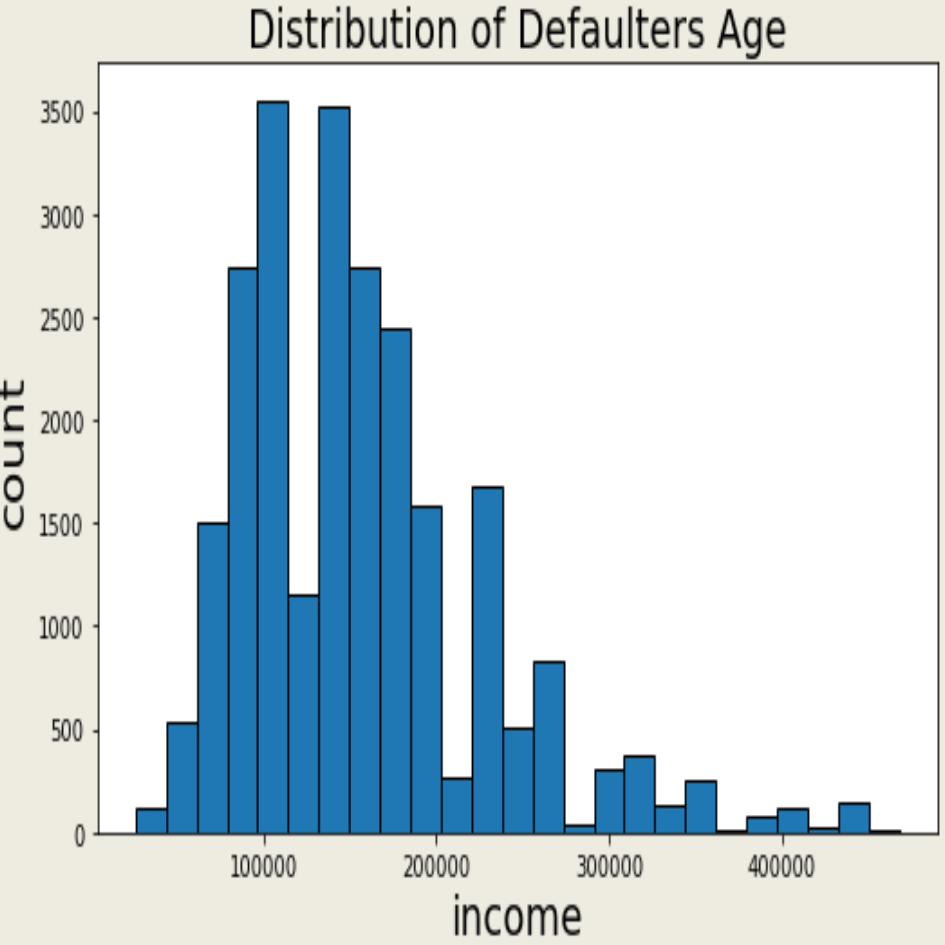
AGE OF THE APPLICANTS.

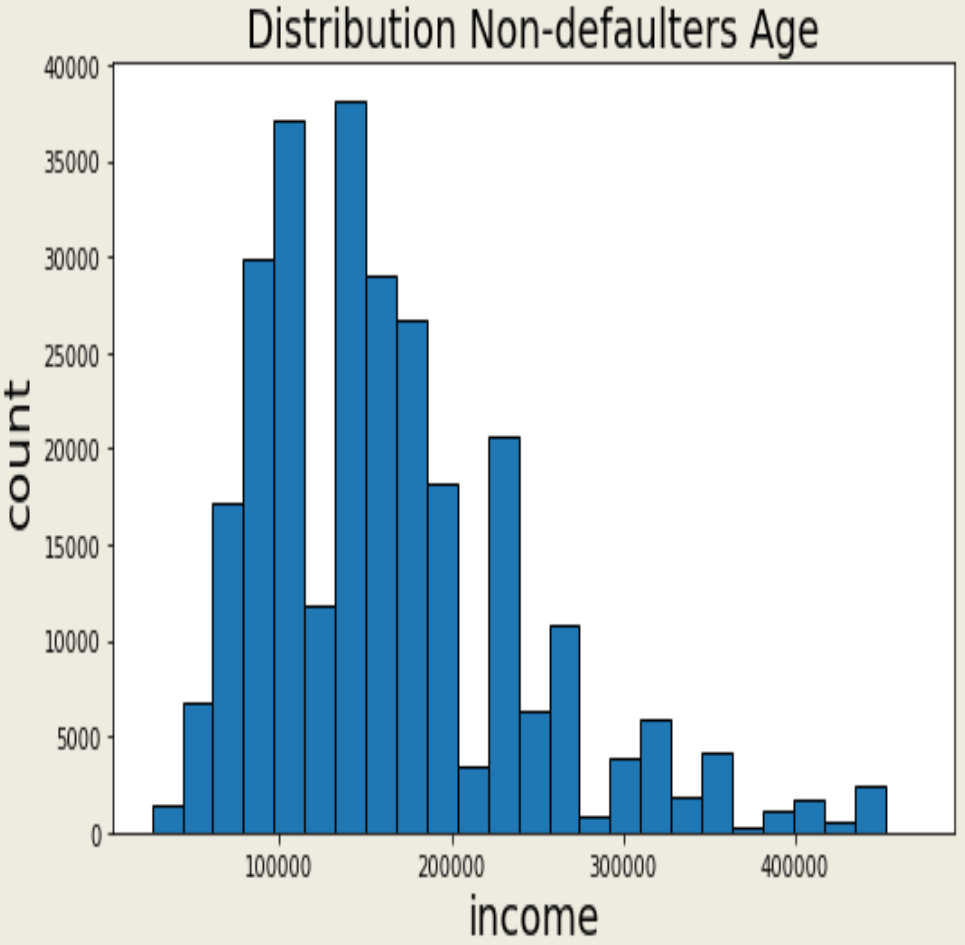




FROM THE HISTOGRAM WE SEE THAT THE HIGHEST NUMBER OF NON-DEFAULTERS ARE BETWEEN 25-30 YEARS, AND AS THE AGE INCREASES THE NUMBER OF NON- DEFALUTERS ALSO INCREASES.PEOPLE WITH AGE GROUP FROM 25-30ARE PREFERRABLE FOR CREDIT. FOR THE DEFAULTERS PEOPLE BETWEEN THE AGE GROUP OF 35-45 TEND TO DEFAULT THE MOST, SO CREDIT LOANS TO THIS AGE GROUP IS NOT PREFERRABLE

ANALYSING THE INCOME RANGE.

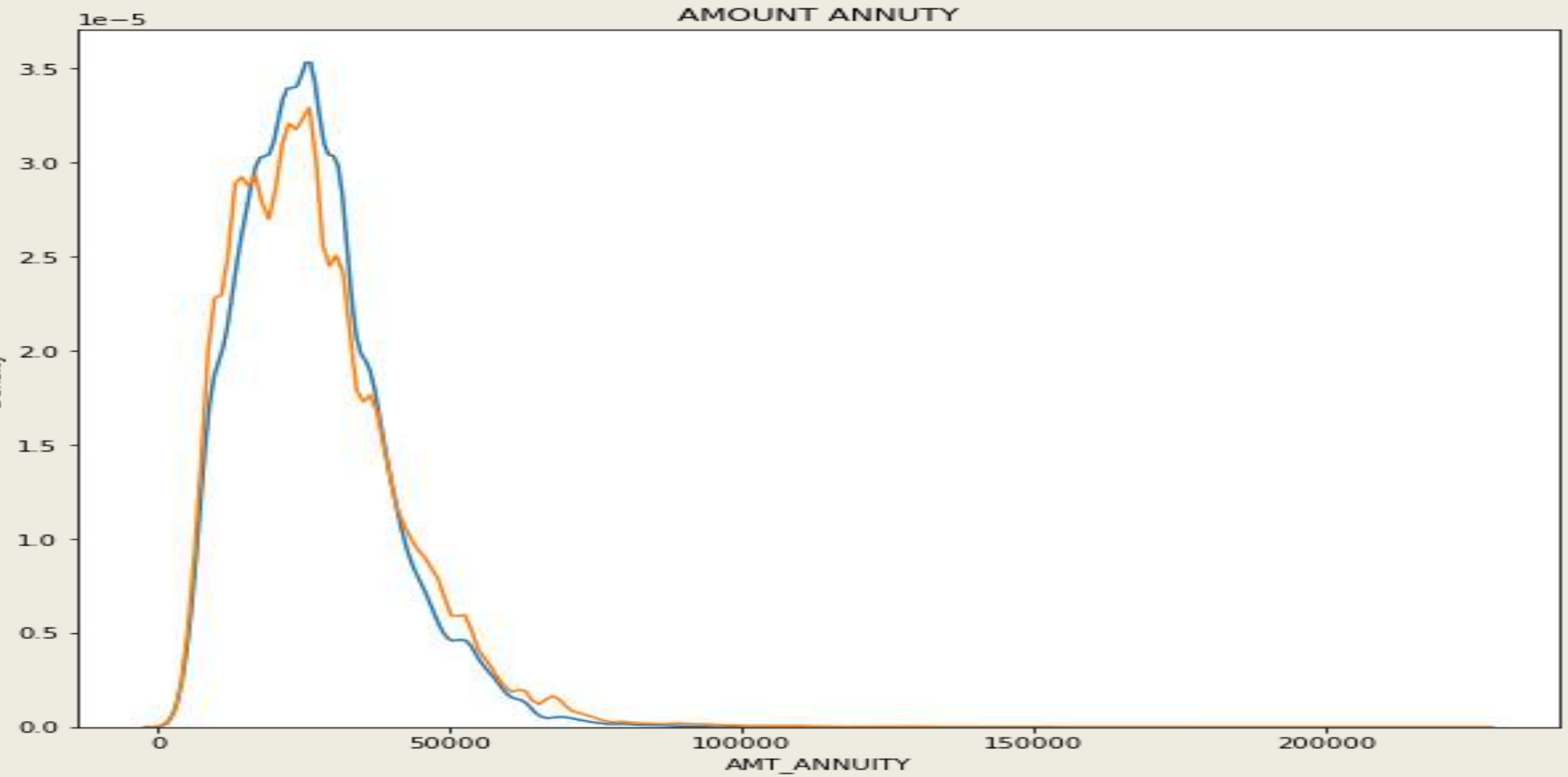




MOST OF THE INCOME RANGES ARE BETWEEN 50,000- 200000 WE SEE THAT THE INCOME RANGES OF THE BOTH DEFAULTERS AND NON- DEFAULTERS ARE ALMOST SAME. SO GOING FOR FURTHER ANALYSIS.

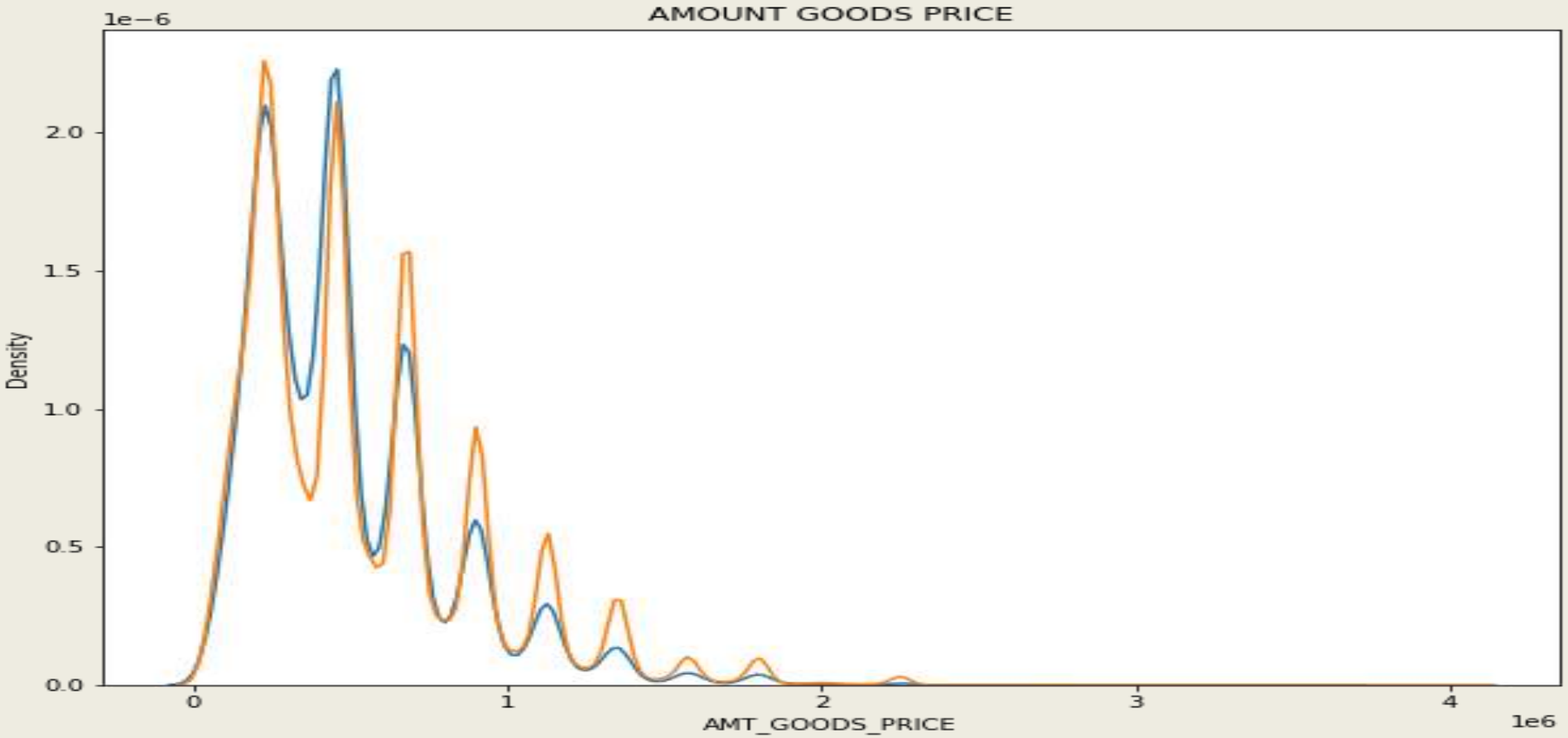
UNIVARIATE ANALYSIS OF THE NUMERICAL COLUMNS.

PLOTTING KDE FOR LOAN ANNUITY.



LOAN ANNUITY IS MOSTLY CONCENTRATED BETWEEN 0-50000. THE GRAPH PATTERN IS SAME FOR BOTH DEFAULTERS AND NON- DEFAULTERS.

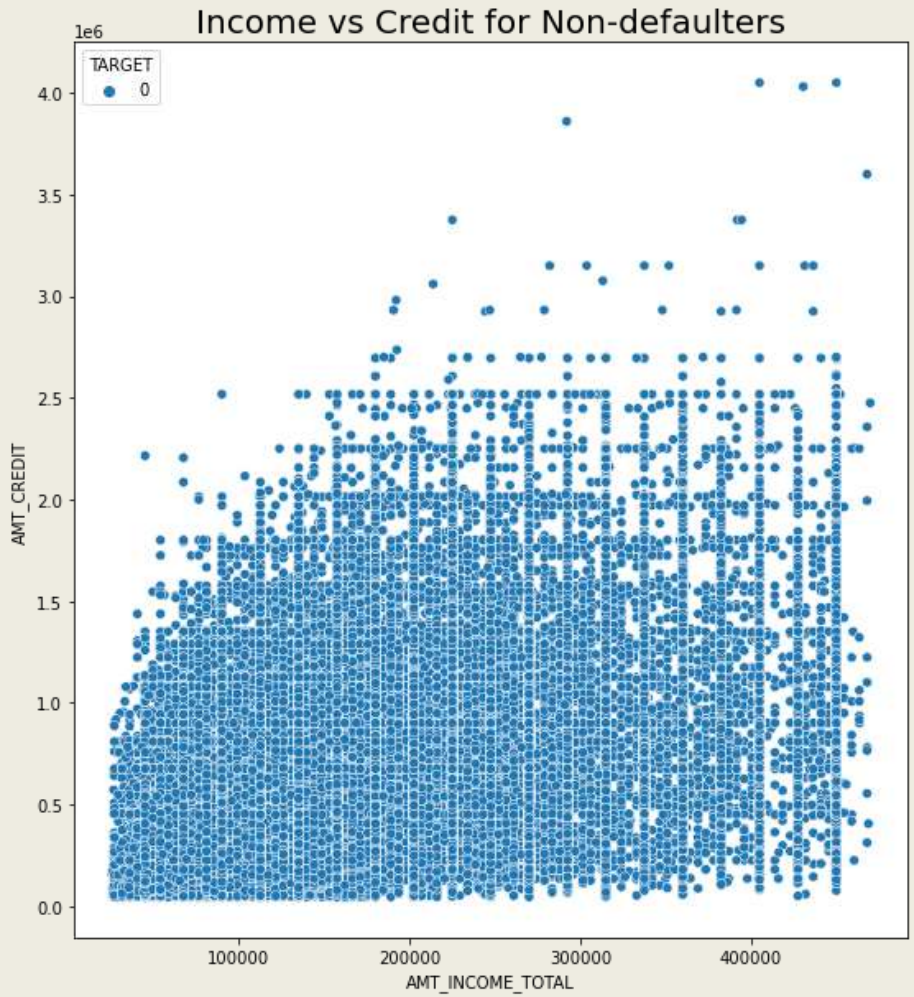
KDE PLOT FOR THE VALUE OF THE GOODS, FOR WHICH THE LOAN WAS AVAILED.

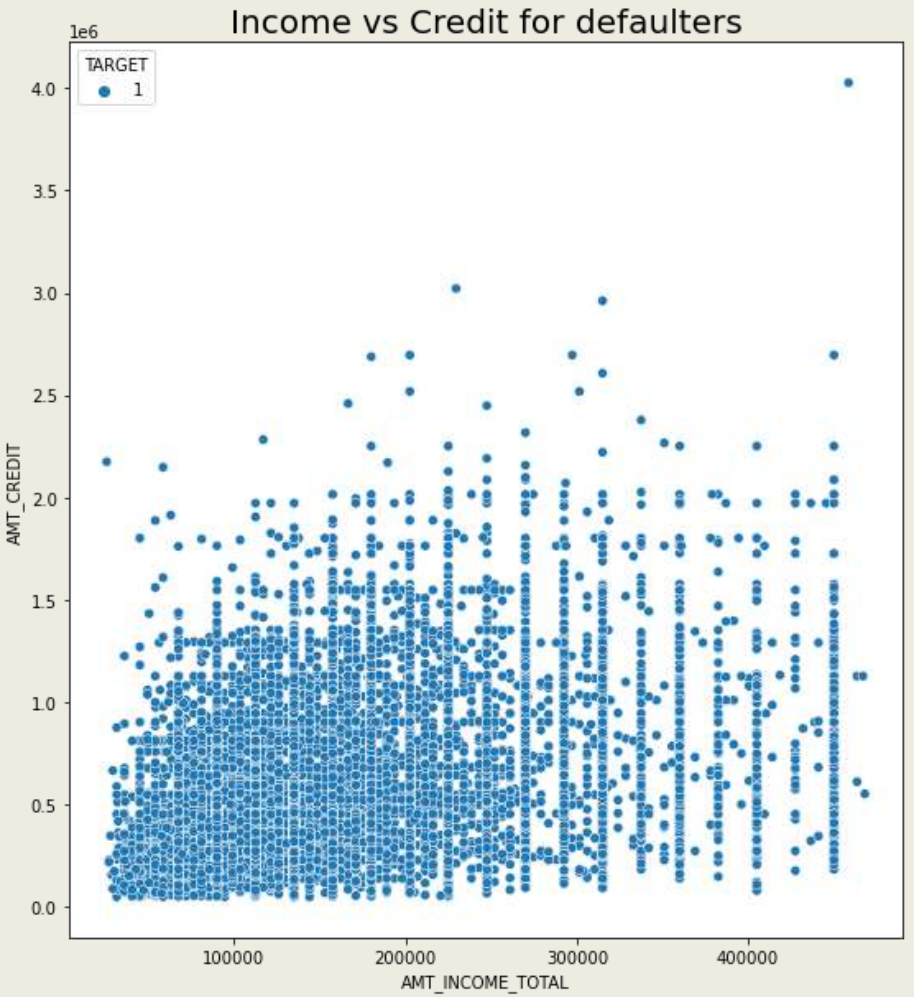


EVEN IN THIS KDE PLOT WE SEE THAT THE GRAPH ALMOST FOLLOWS A SIMILAR PATTERN DISTRIBUTION. THERE ARE SOME SIMILARITIES BETWEEN 150000-220000, HENCE MORE EXPENSIVE THE GOODS, LESS CHANCES OF DEFAULT.

BIVARIATE ANALYSIS

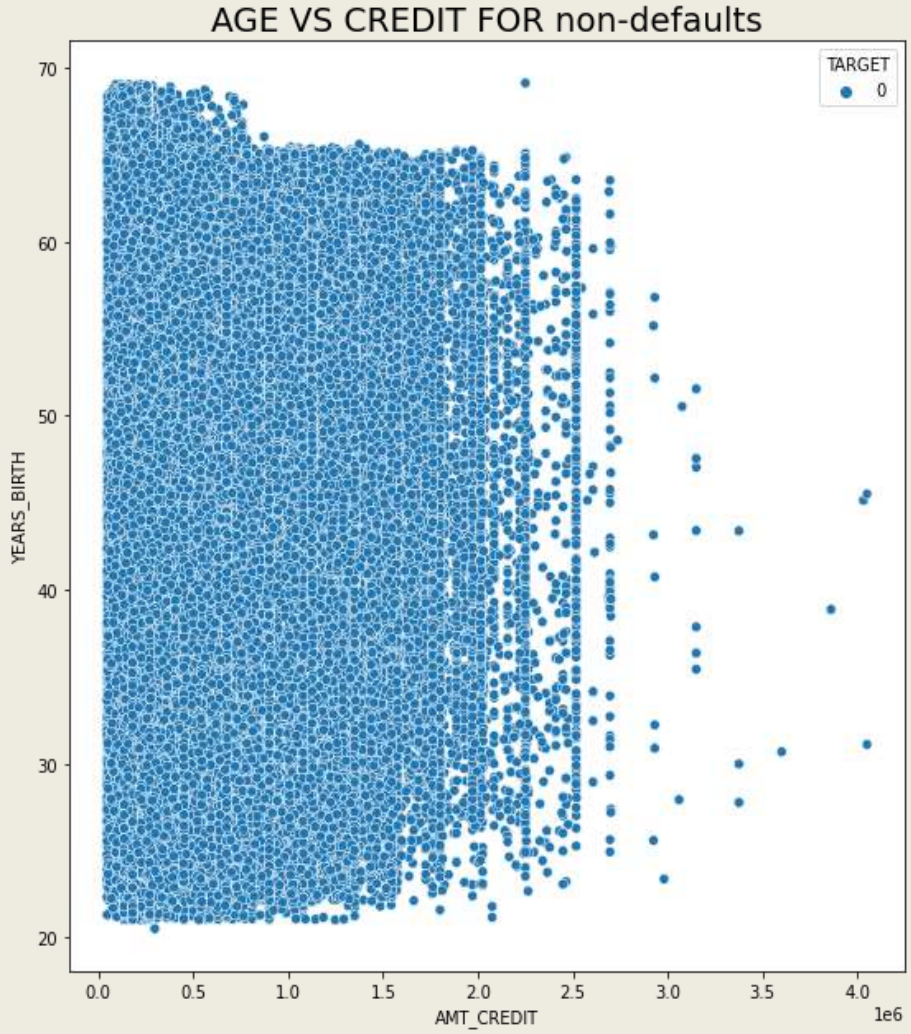
SCATTER PLOT FOR INCOME AND CREDIT.

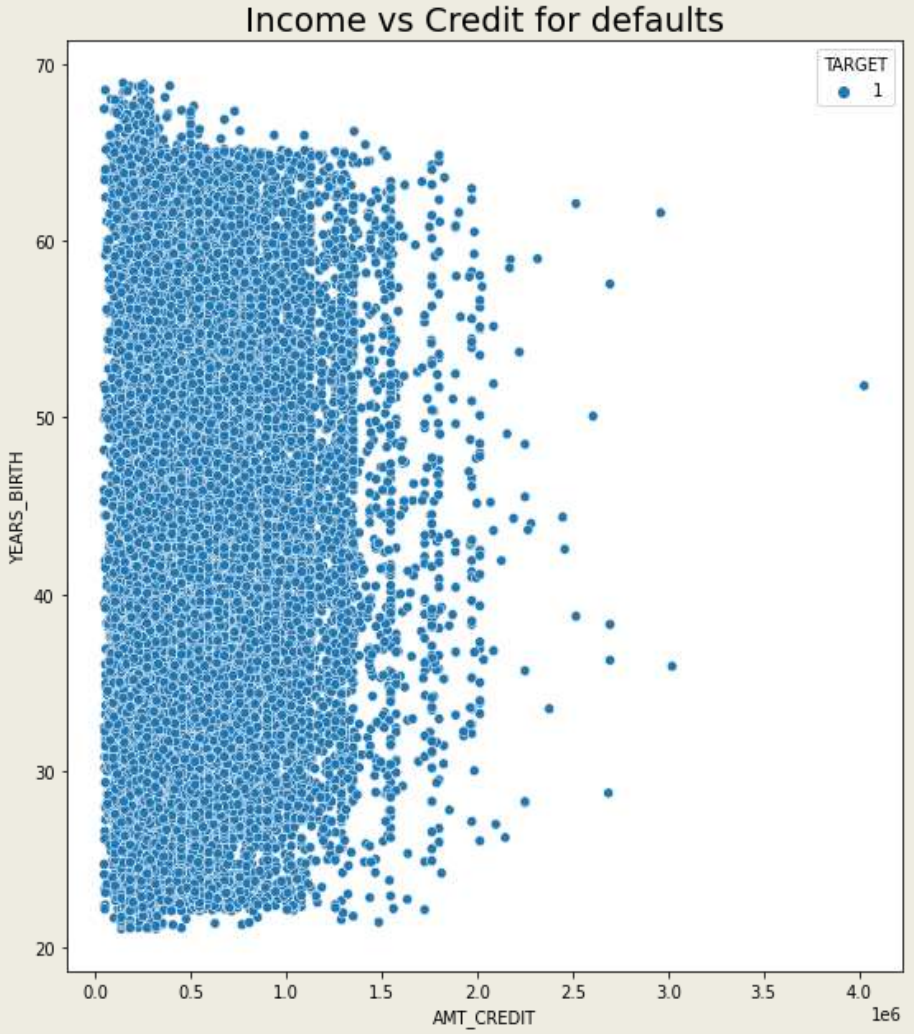




THE DENSITY OF DEFAULTS IS LOW WHEN THE INCOME IS HIGH, THE DENSITY OF NON DEFAULTERS DECREASES AS THE CREDIT AND INCOME BOTH INCREASE

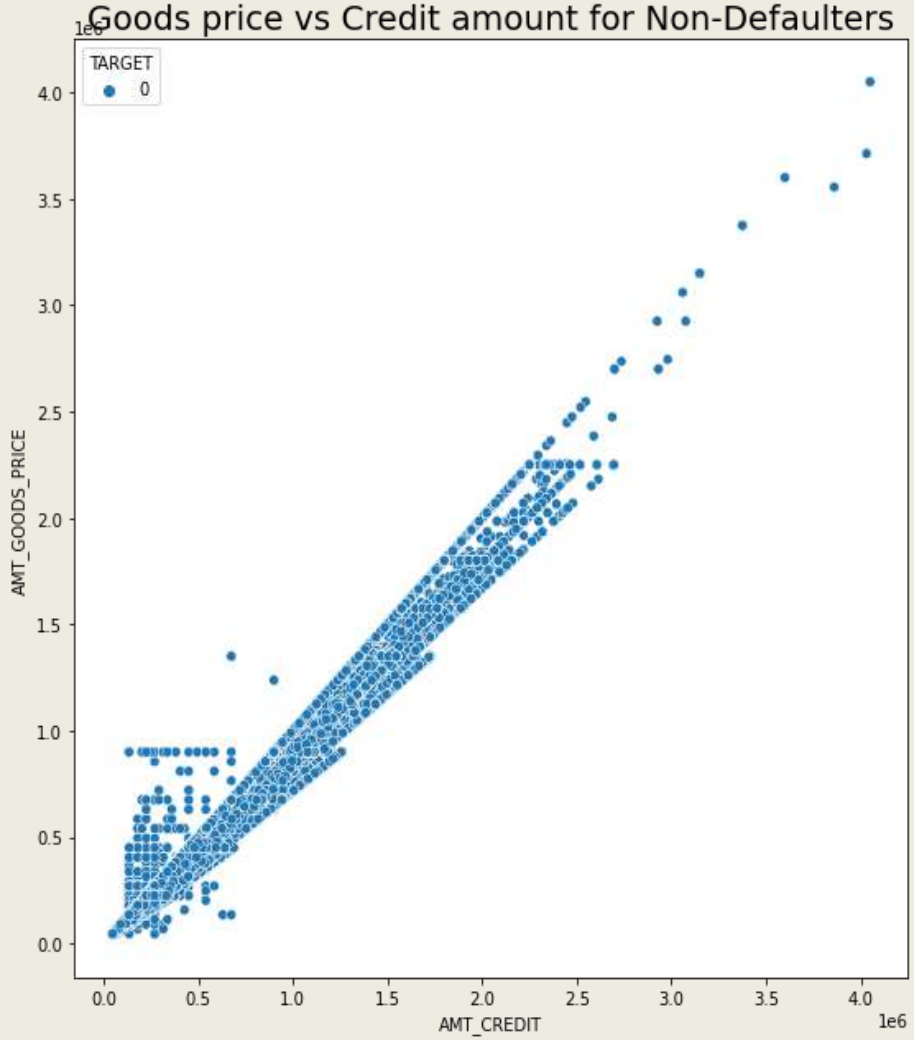
SCATTER PLOT FOR AGE AND CREDIT AMOUNT.

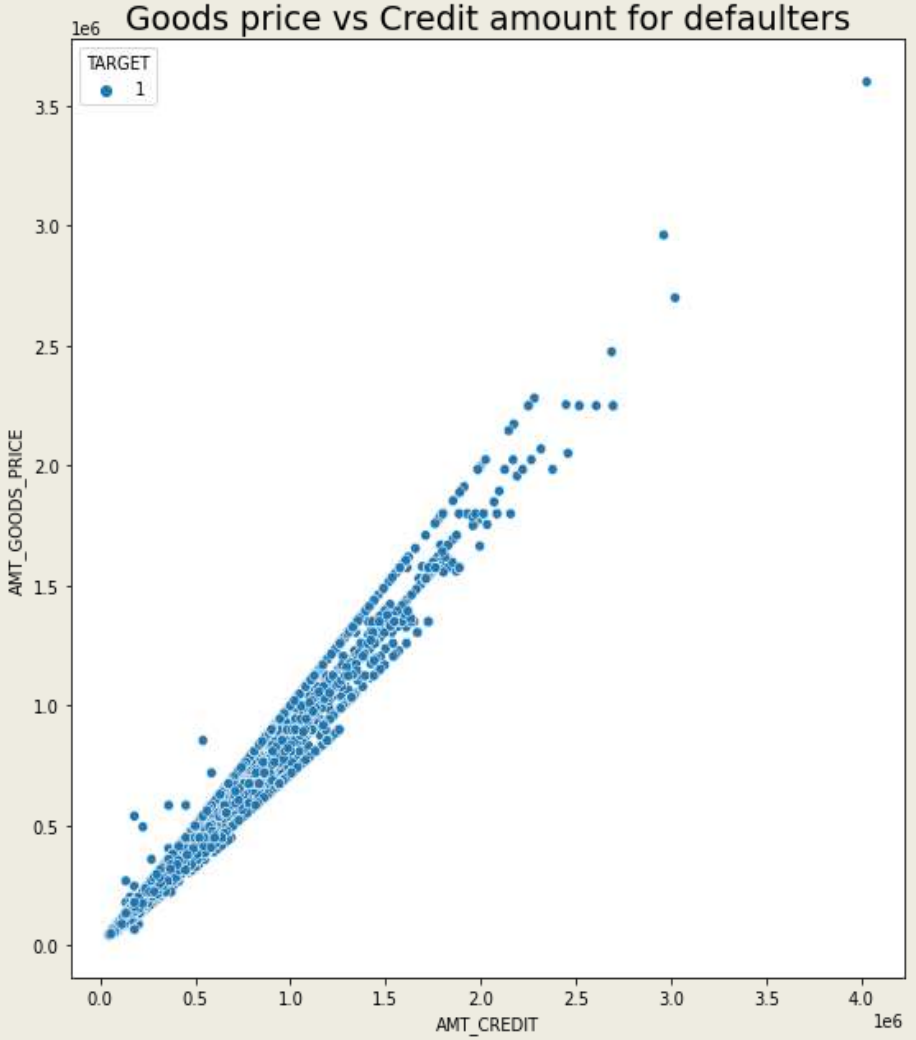




THE DEFAULTERS AND NON DEFAULTERS DENSITY IS SPRAD EVENLY ACROSS, FOR MOST OF THE DEFALUTERS THE CREDIT AMNT IS LESS THAN 1.5, WHEREAS THAT FOR NON DEFAULTERS IS LESS THAN 2

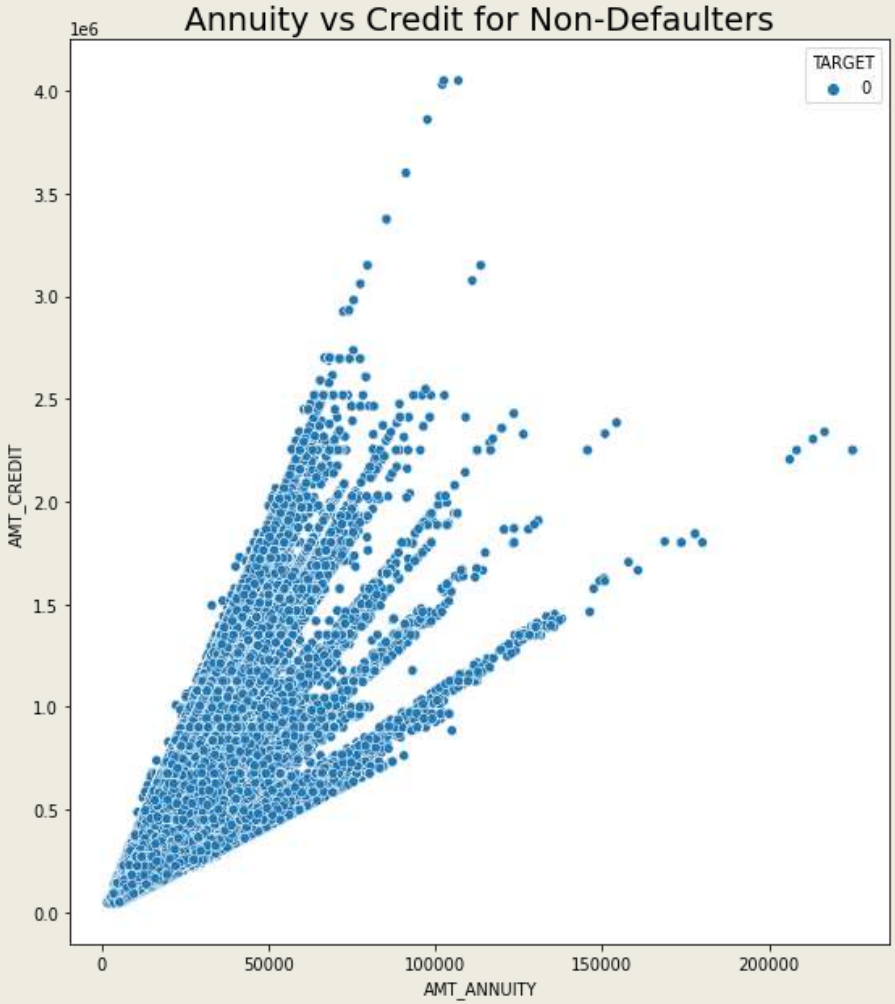
SCATTER PLOT FOR CREDIT AMOUNT AND GOOD PRICE.

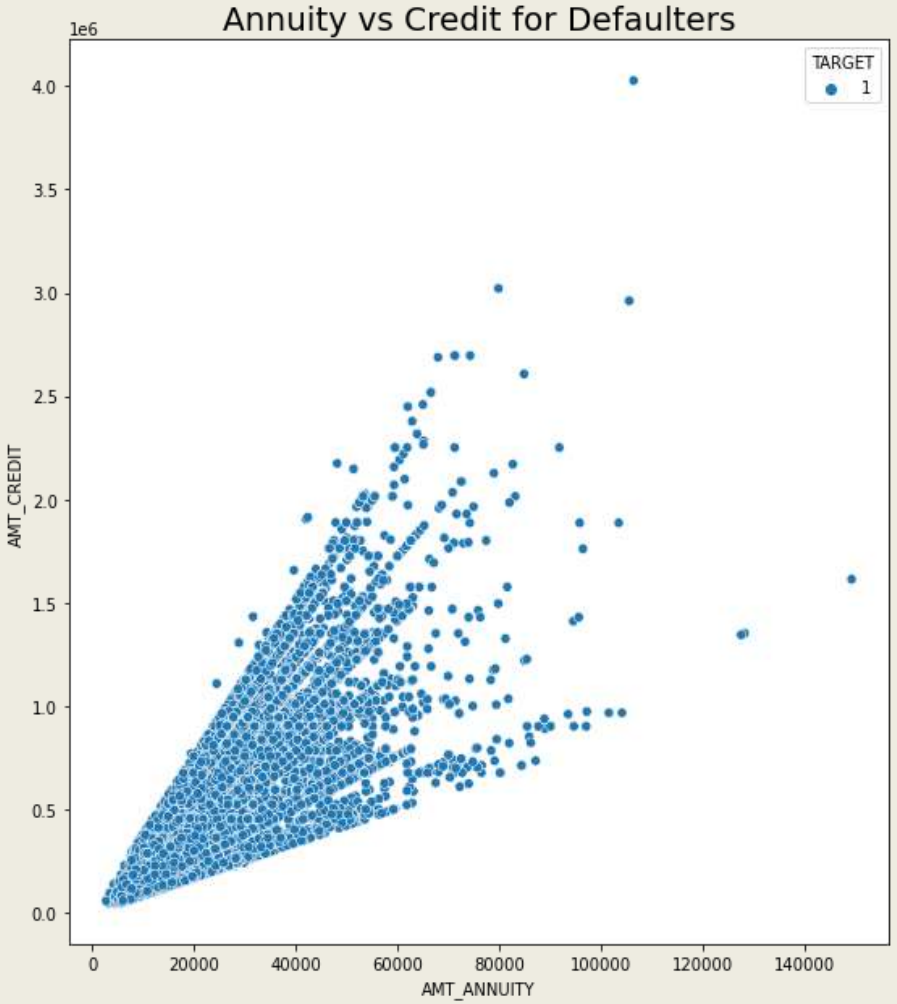




FOR BOTH THE DEFALUTERS AND NON DEFAULTERS, AS THE VALUE OF THE GOODS INCRESES, THE LOAN VALUE INCREASES

SCATTER PLOT FOR ANNUITY AND CREDIT AMOUNT.

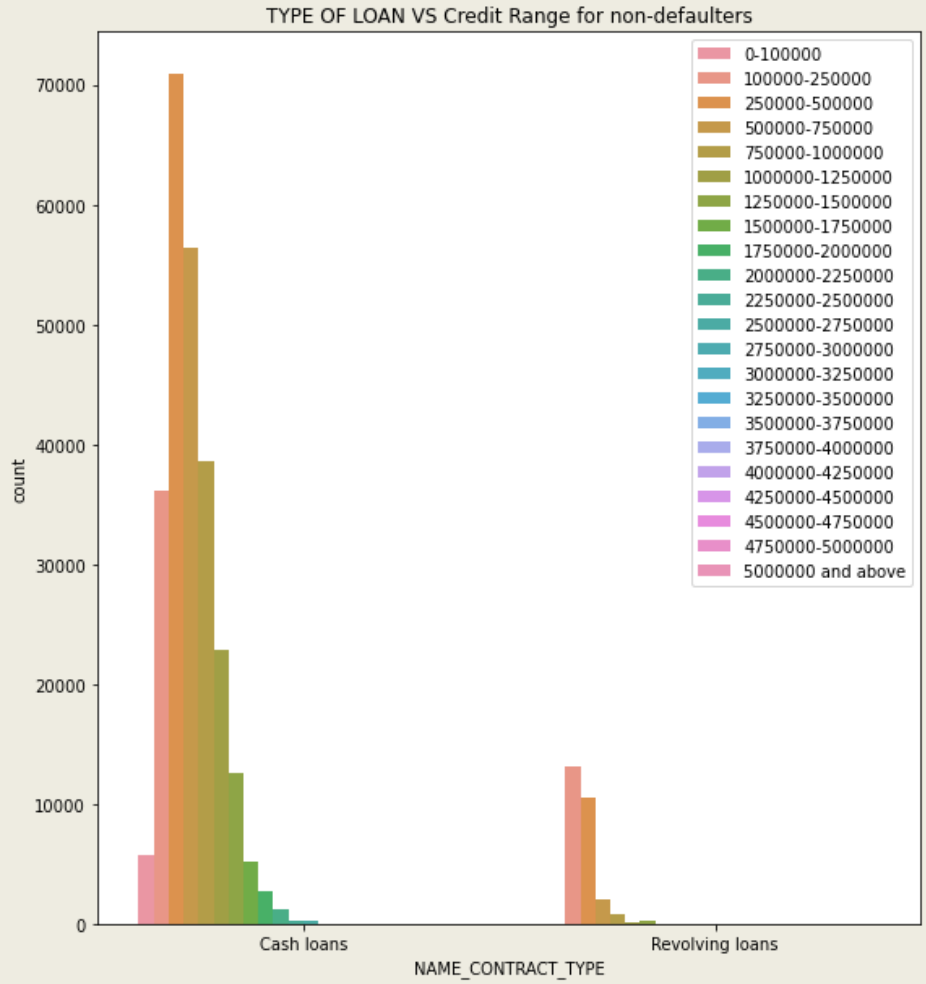


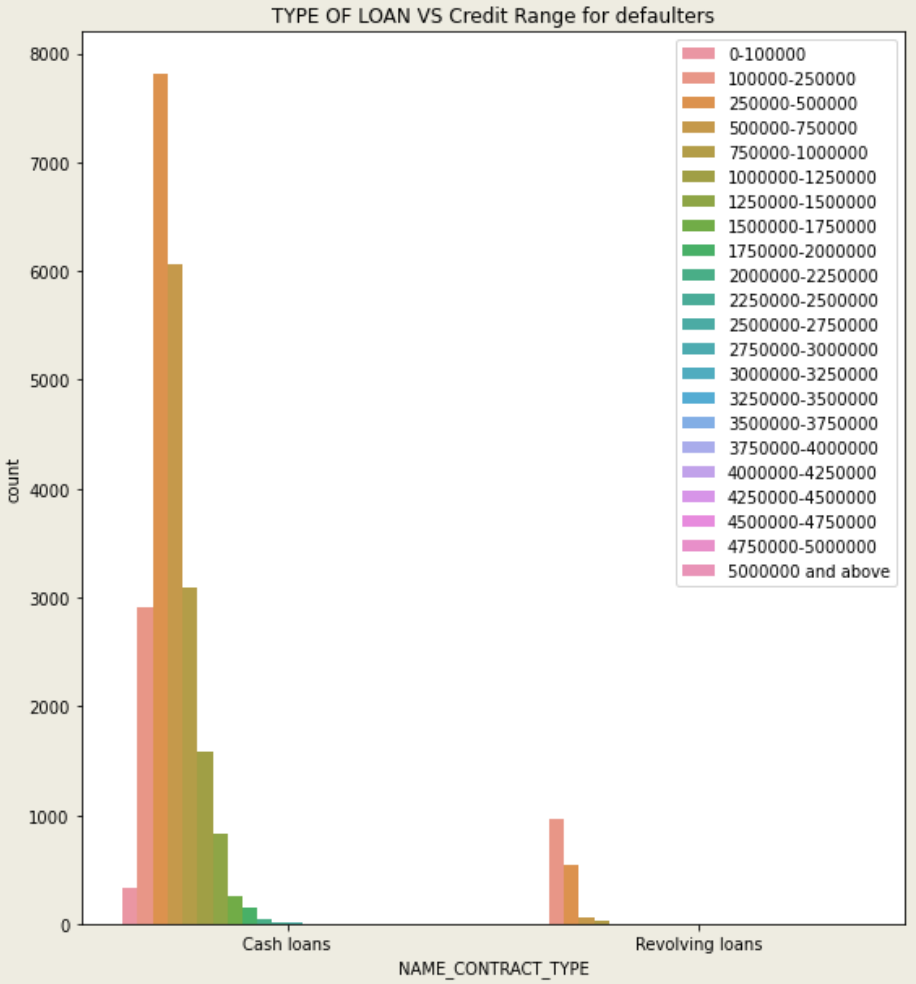


FOR BOTH DEFAOLTERS AND NON- DEFAULTERS, AS THE CREDIT INCREASES, THE ANNUITY ALSO INCREASES

BIVARIATE ANALYSIS BETWEEN NUMERICAL AND CATEGORICAL COLUMN

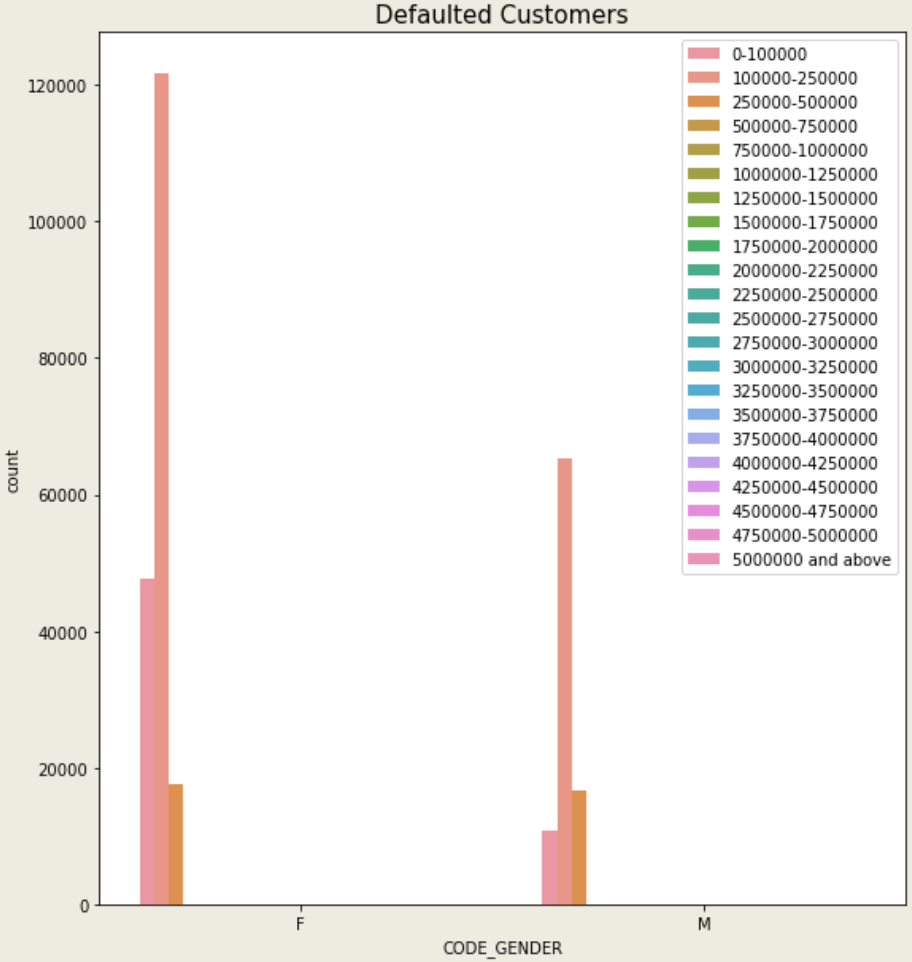
COUNT PLOT FOR TYPE OF LOAN AND CREDIT RANGE.

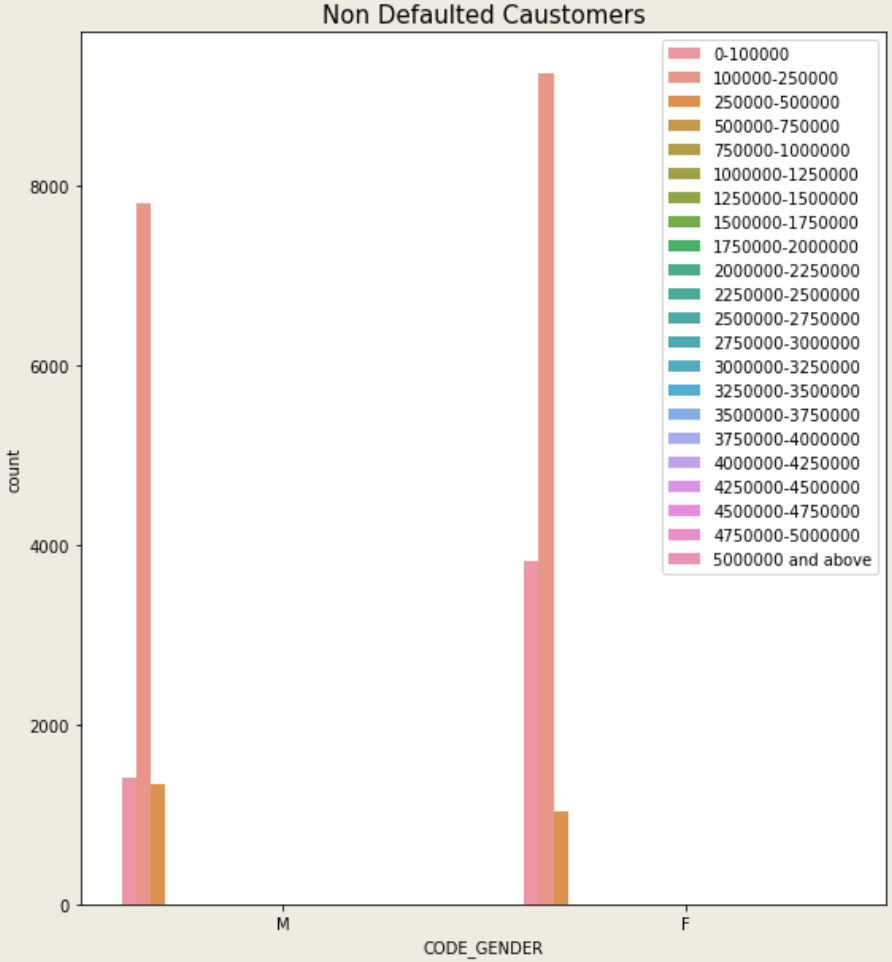




WE CAN SEE THAT THE CASH LOANS ARE PREFERRED IRRESPECTIVE OF THE CREDIT RANGE. AND THE MOST PREFERRED CREDIT RANGE IS 250000-500000

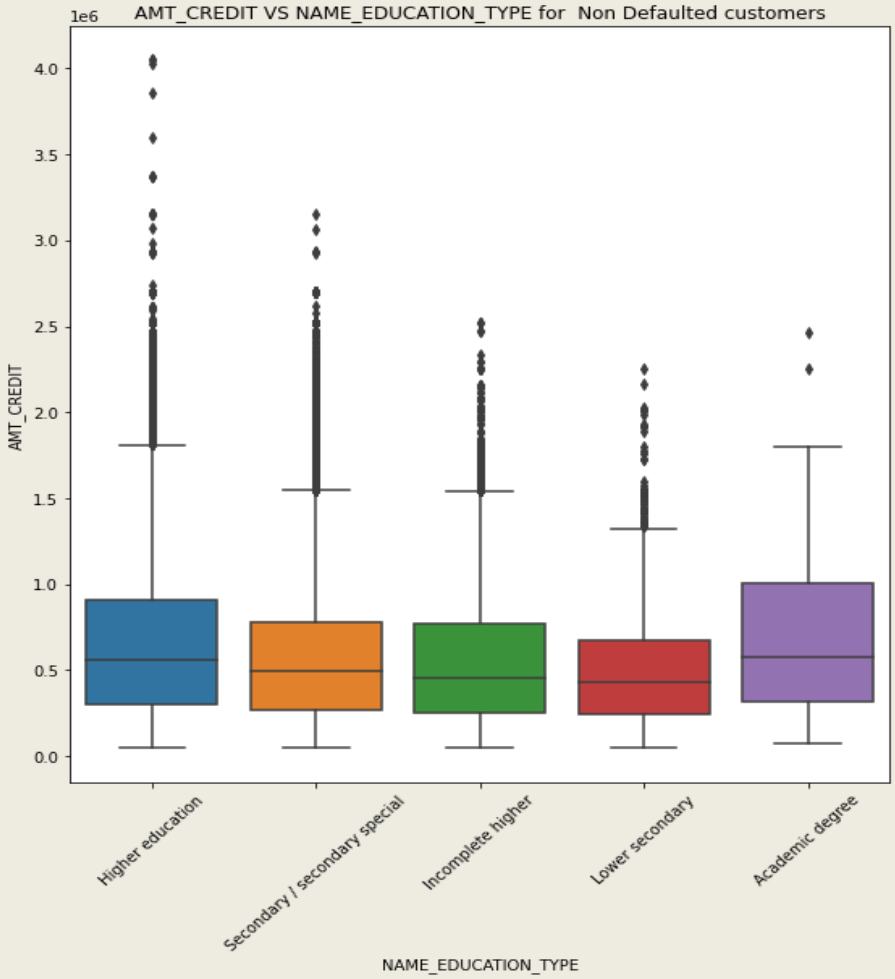
ANALYSIS BETWEEN GENDER AND INCOME.

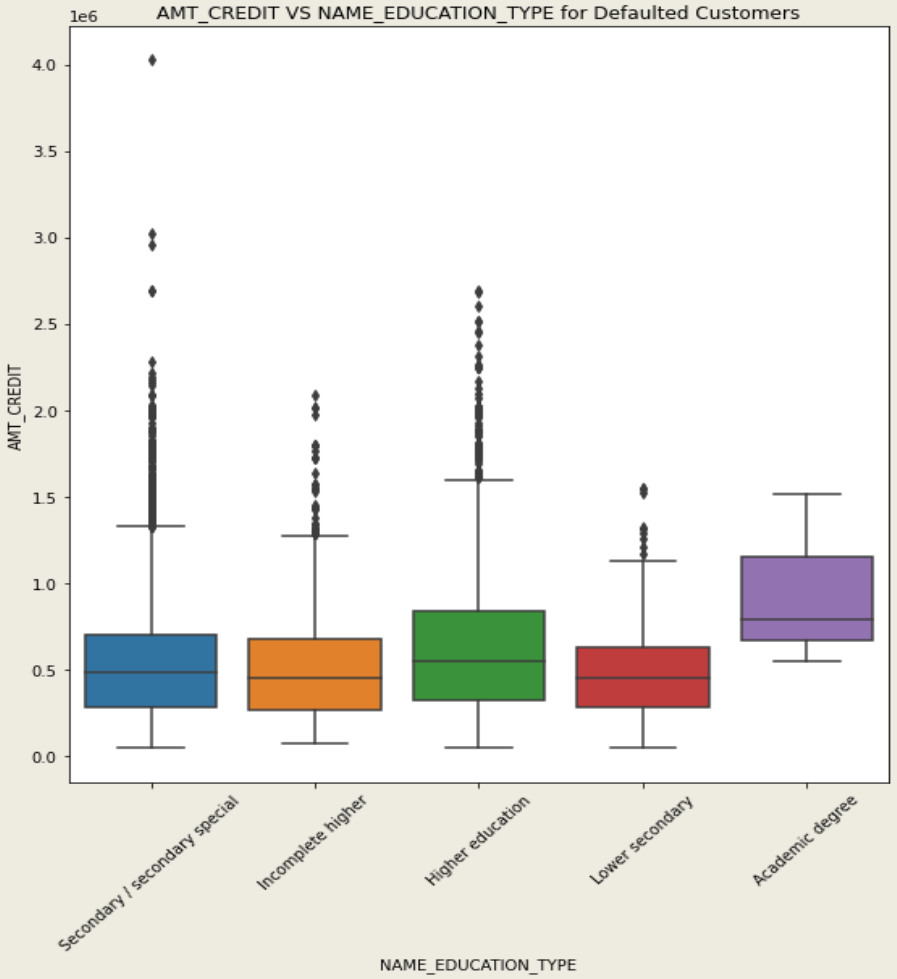




WE SEE THAT BOTH AMONG DEFALTERS AND NON DEFAULTERS FEMALES HAVE THE HIGHEST AMOUNT OF INCOME RANGE

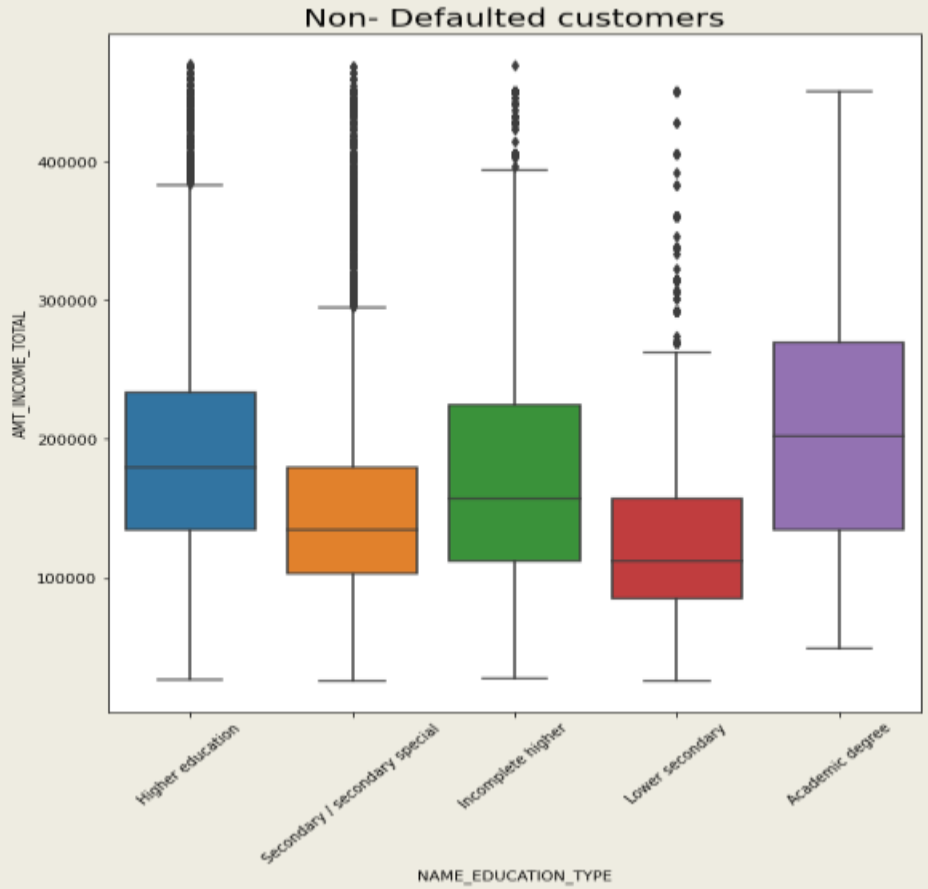
BOX PLOT FOR CREDIT AMOUNT AND EDUCATION TYPE.

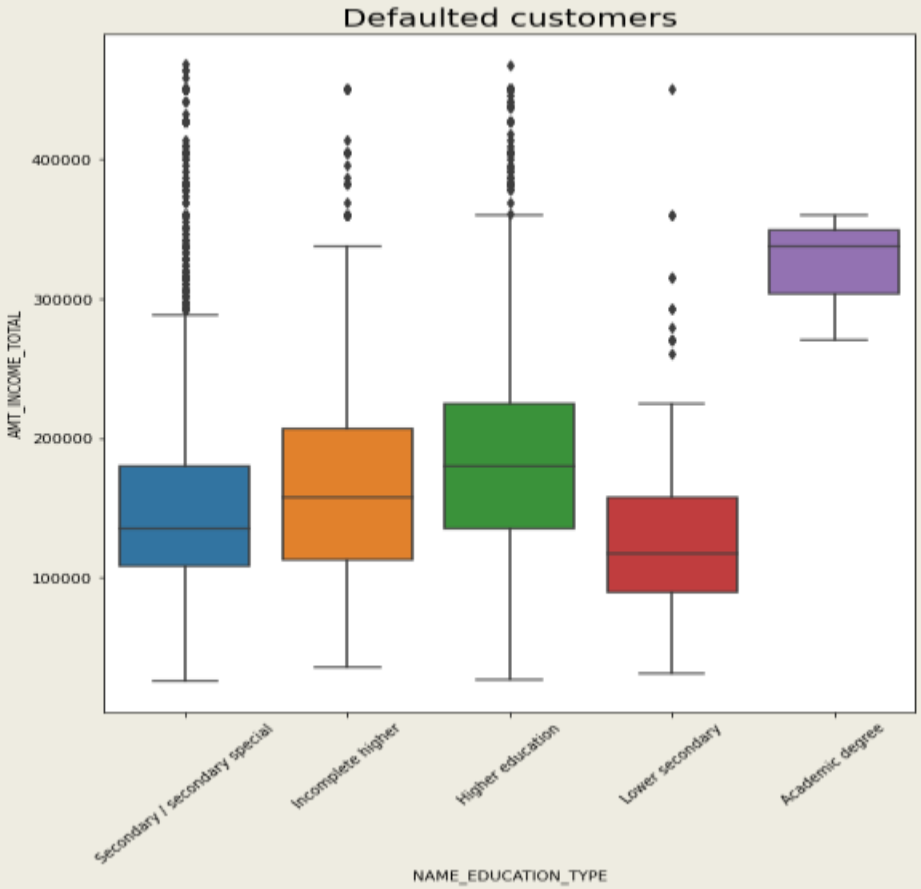




WE SEE THAT AMONGS THE NON DEFAULTED CUSTOMERS, PEOPLE WITH HIGHER EDUCATION HAVE AVAILED MORE CREDIT COMPARED TO OTHER EDUCATION CATEGORIES. AMONGST DEFAULTED CUSTOMERS THE HIGHEST CREDIT AVAILED IS BY SECONDARY EDUCATION TYPE.

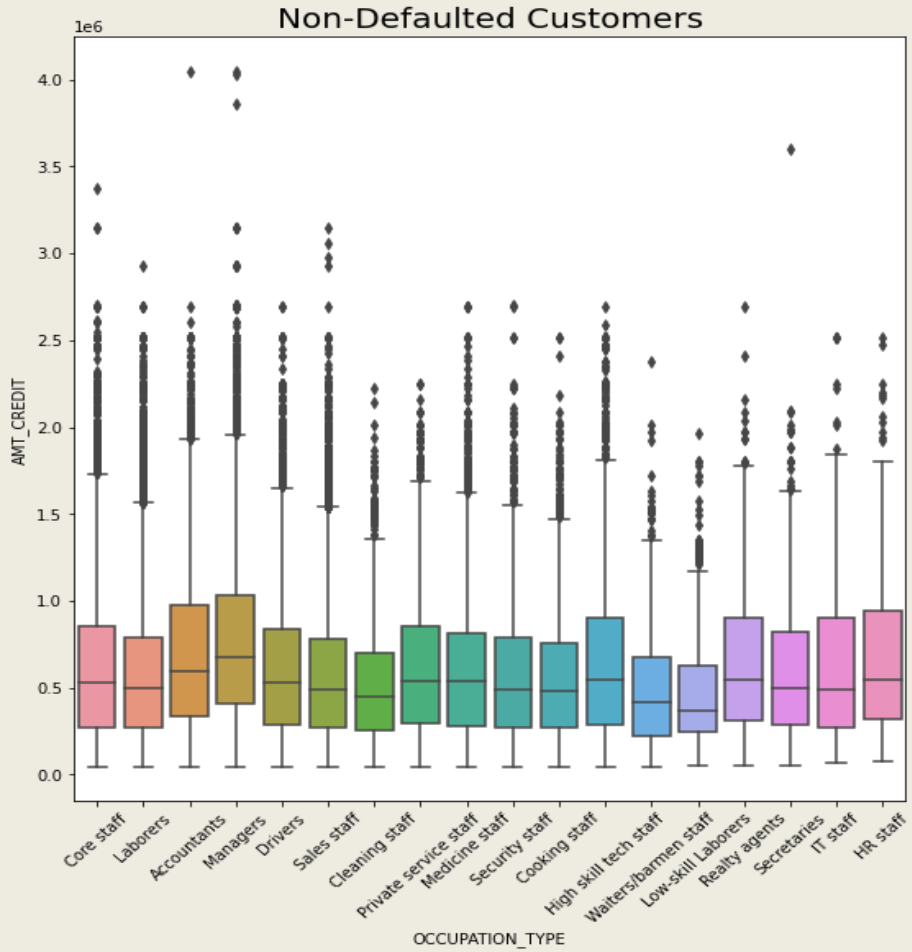
BOXPLOT FOR INCOME AND EDUCATION TYPE.

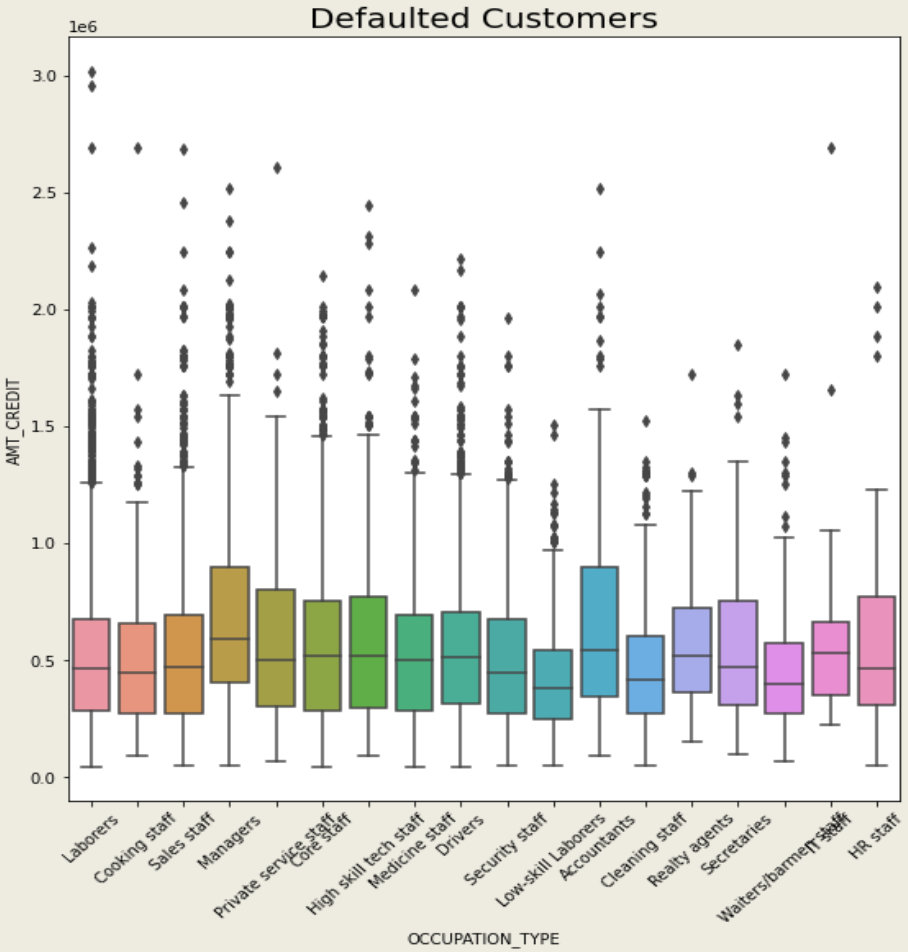




FROM THE PLOT WE SEE THAT THE ACADEMIC DEGREE TYPE OF CLIENTS HAVE AVAILED A WIDE RANGE OF CREDITS AMONGST THE NON DEFAULTERS, WHERE AS AMONGST THE DEFAULTERS ITS CONCENTRATED IN A CERTAIN RANGE.

BOXPLOT FOR OCCUPATION AND THE CREDIT AMOUNT OF THE CLIENTS.



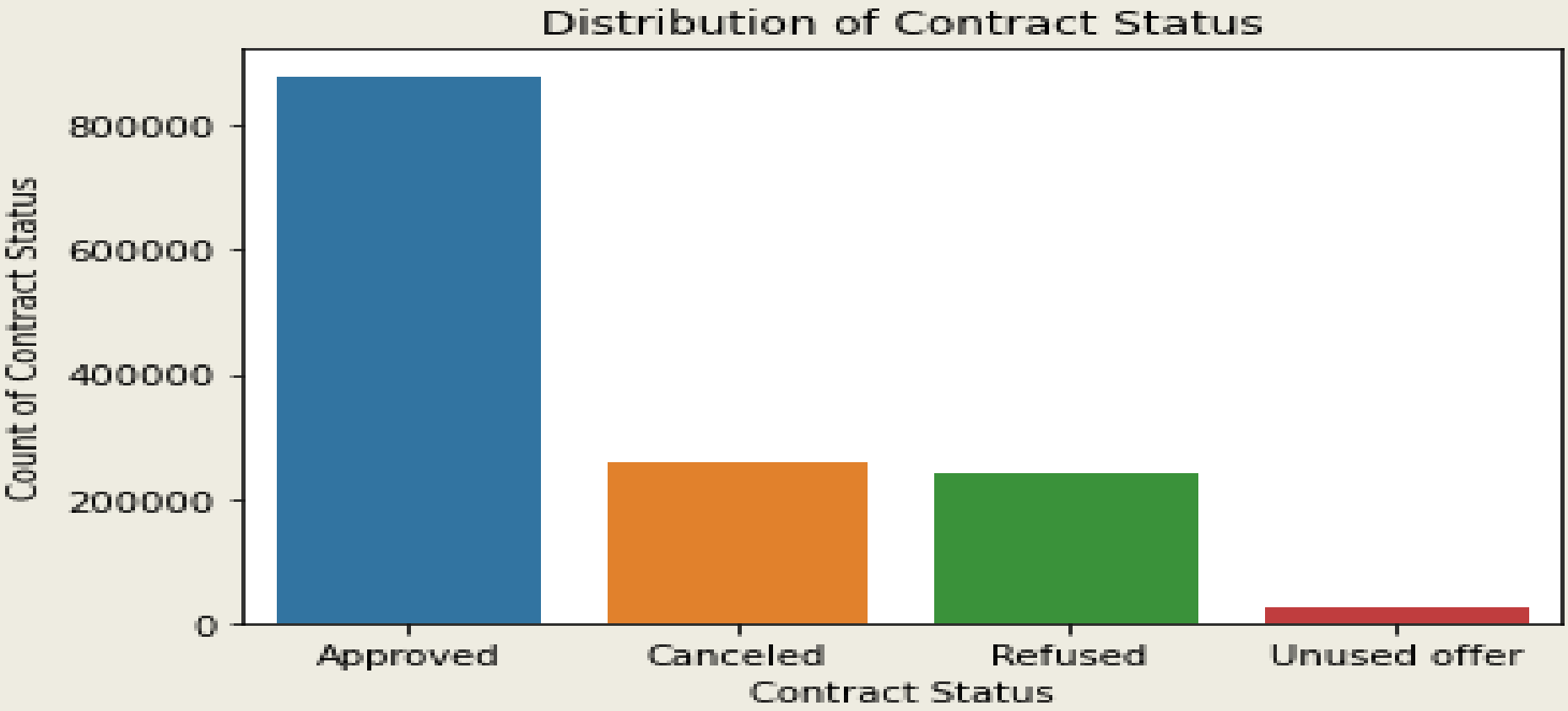


THE OCCUPATION TYPE OF DEFAULTED CUSTOMERS ARE MORE COMPARED TO NON DEFAULTED CUSTOMERSM, LOOKS LIKE MANAGERS HAVE AVAILED MORE LOAN IN BOTH THE CASES, ANOTHER INTERESTING POINT IS LOW SKILL LABOURERES HAVE ALSO AVAILED SIGNIFANT LOANS AMONGST DEFAULTERS.

ANALYSIS AFTER COMBINING THE DATA SET

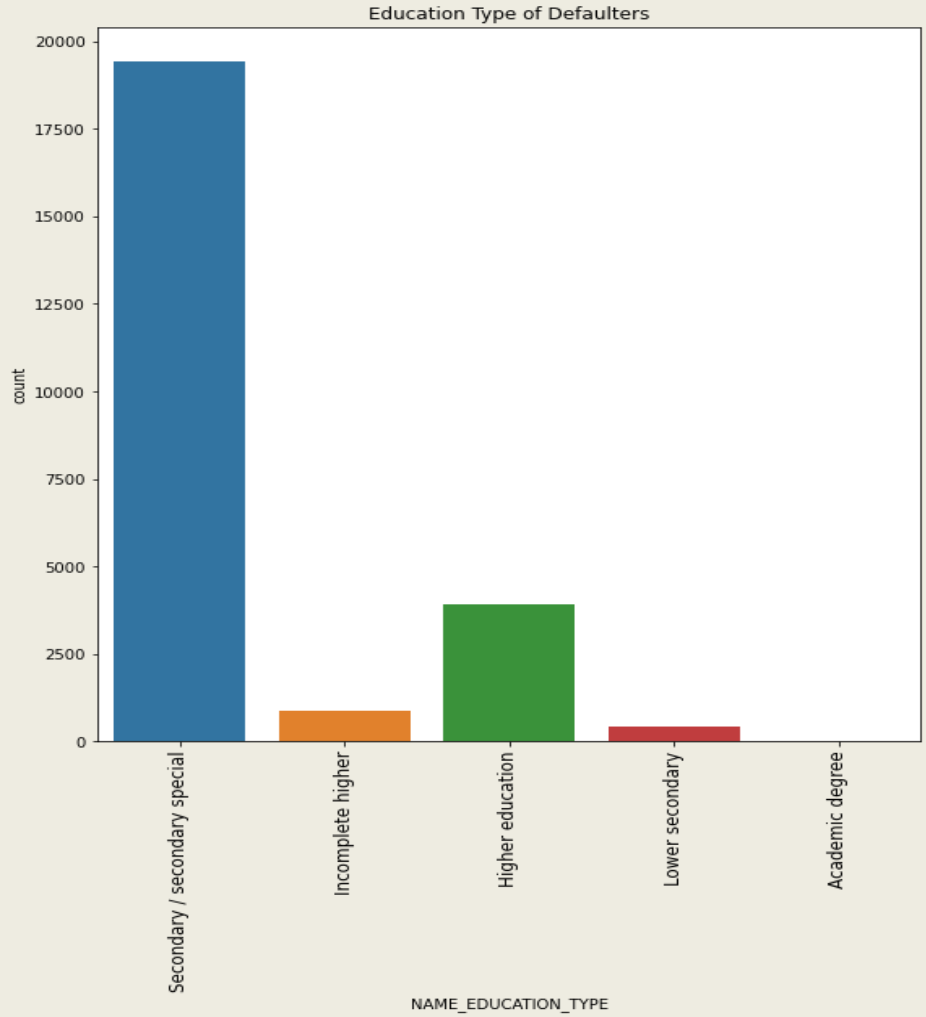
UNIVARIATE ANALYSIS

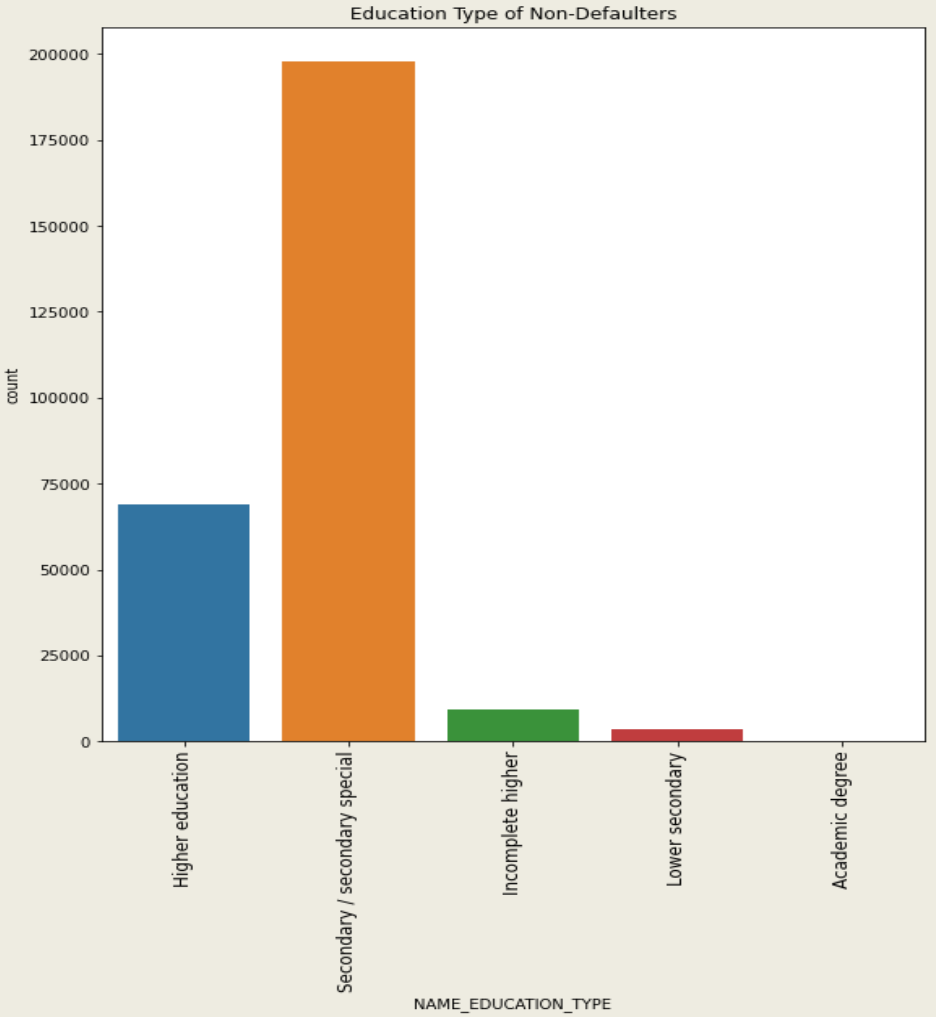
ANALYSIS OF THE CONTRACT STATUS.



WE SEE FROM THE COMBINED DATA THAT THE NUMBER OF APPROVED LOANS ARE MORE THAN CANCELLED OR REFUSED. DOING FURTHER ANALYSIS ON THIS TO GET SOME INFERENCE.

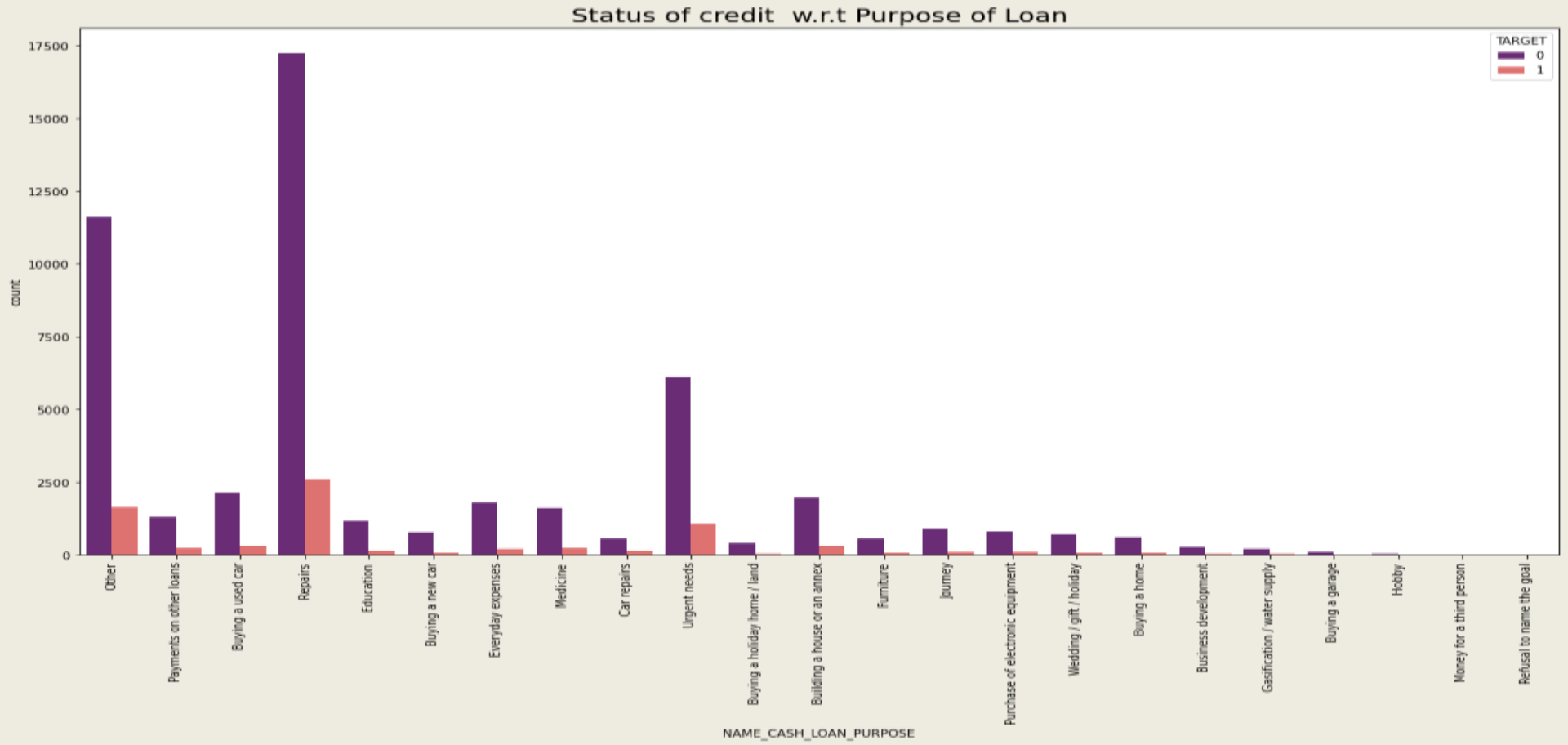
ANALYSIS OF DEFAULTERS AND NON-DEFAULTERS W.R.T THEIR EDUCATION TYPE.





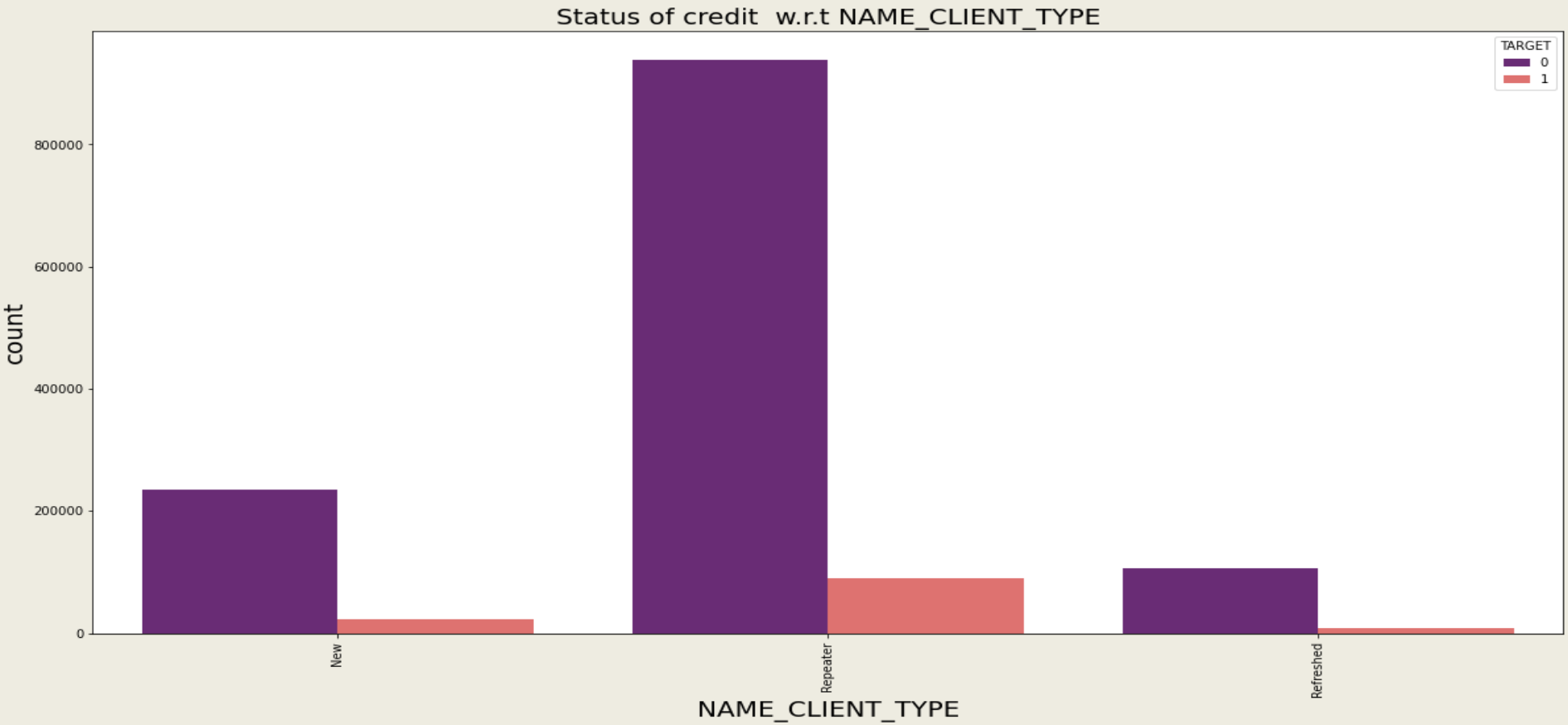
THE SECONDARY EDUCATION IS HIGHEST AMONG BOTH THE SITUATIONS, FOLLOWED BY HIGHER EDUCATION AND INCOMPLETE HIGHER. SO BOTH THE GRAPHS ARE PRETTY SIMILAR.

ANALYSIS OF PURPOSE OF THE LOAN W.R.T DEFAULTERS AND NON-DEFAULTERS.



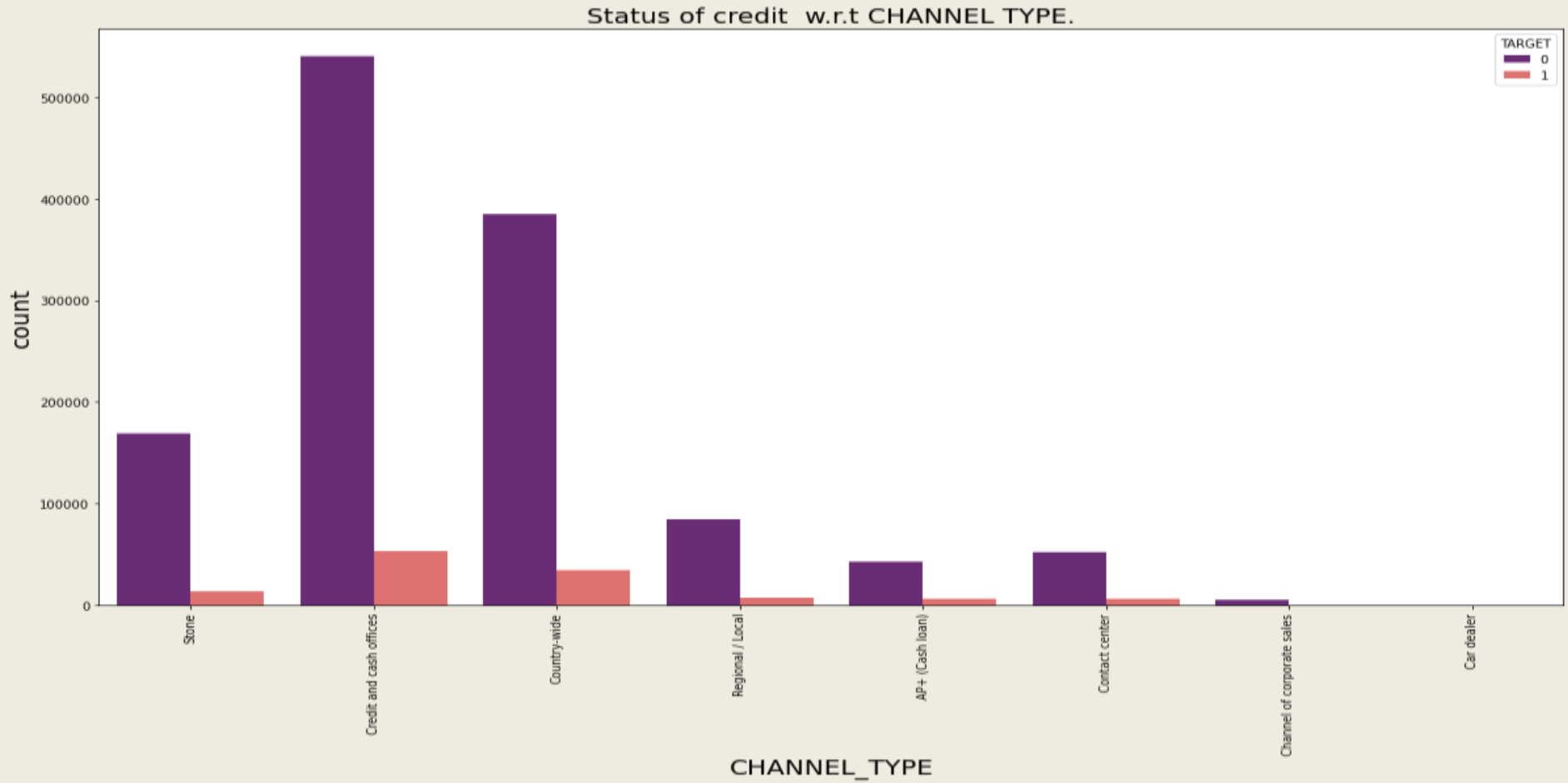
LOAN APPLIED FOR REPAIRS IS THE HIGHEST IN BOTH THE DEFAULTERS AND NON-DEFAULTERS. WHERE AS IT IS FOLLOWED BY OTHER NEEDS AND URGENT NEEDS. EDUCATION, HOUSE, MEDIACL AND CAR LOANS ARE FOLLOWED BY NEXT.

STATUS OF CREDIT W.R.T CLIENT TYPE.



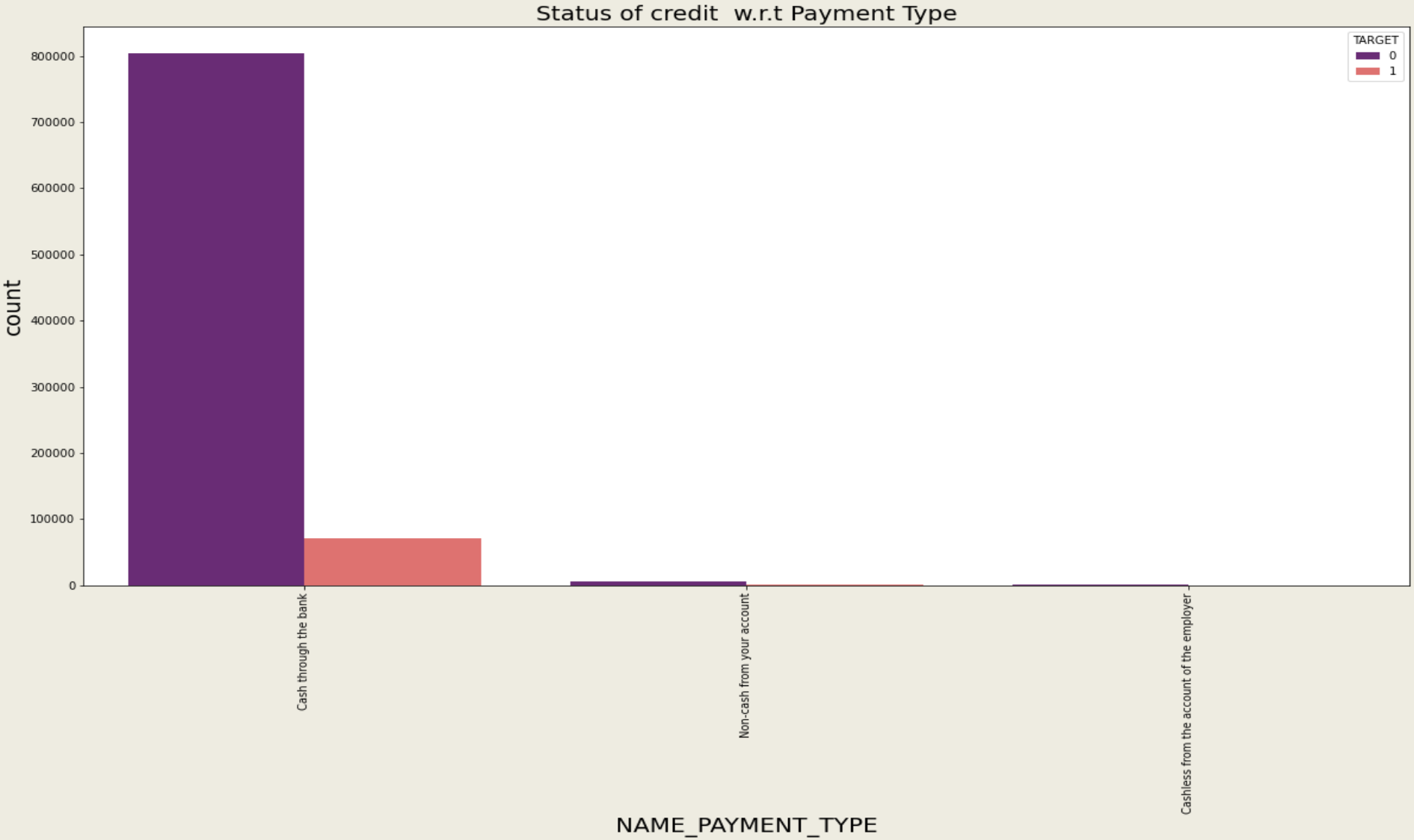
REPEATERS AMONG THE DEFAULTERS AND NON- DEFAULTERS IS HIGHEST, FOLLOWED BY NEW AND REFRESHED. BUT THE DIFFERENCE BETWEEN BOTH THE DEFAULTERS AND NON-DEFAULTERS IS HUGE.SO, IF YOU ARE AVAILING A REPEAT CREDIT, THE CHANCES OF CREDIT BEING APPROVED IS HIGH.

Status of credit w.r.t CHANNEL TYPE.



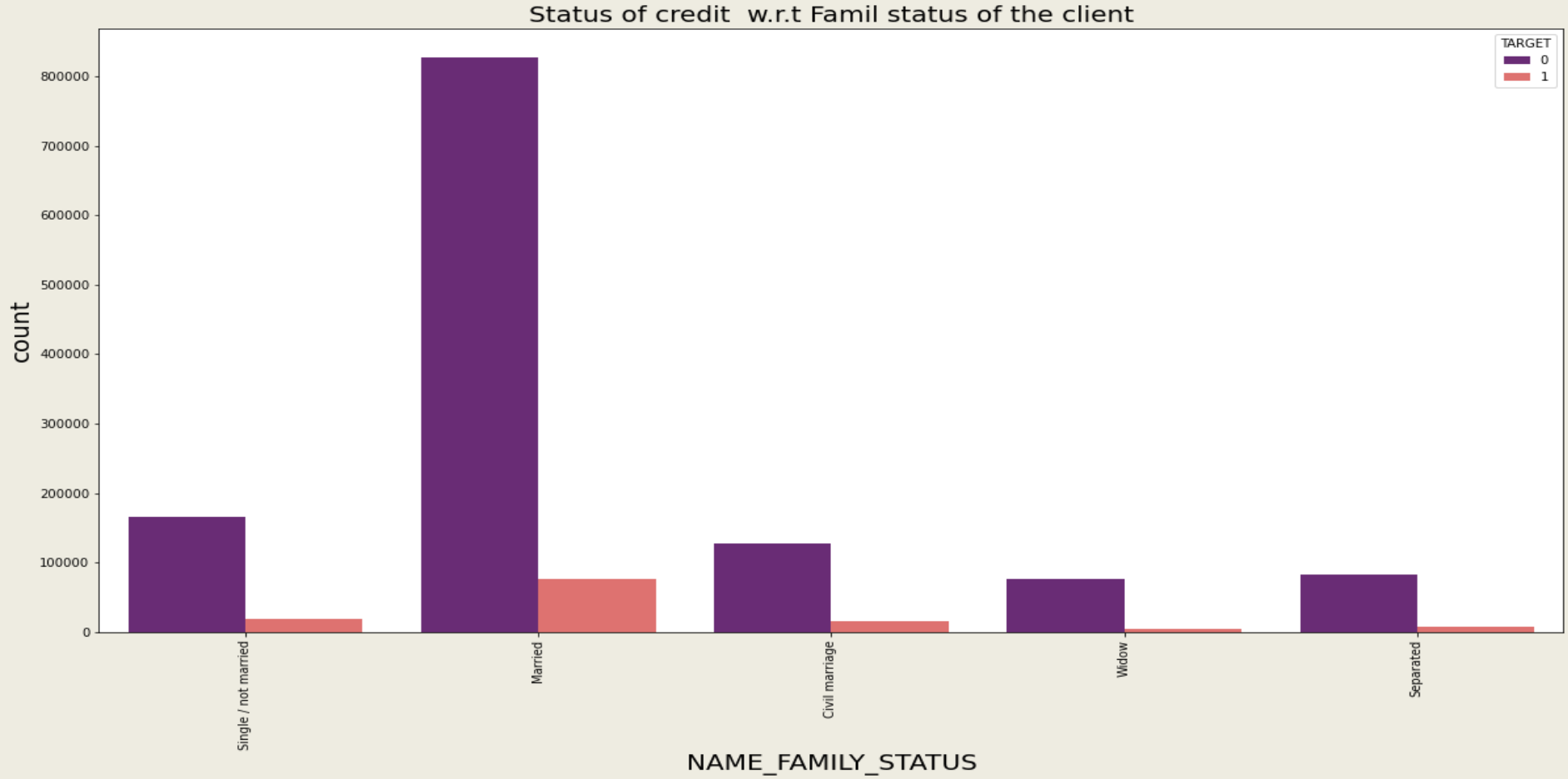
FOR NON DEFAULTED CLIENTS, THE CHANNEL OF ACQUISITON WAS THROUGH CASH AND CREDIT OFFICES, FOLLOWED BY COUNTRY WIDE ACQUISITON.

STATUS OF CREDIT W.R.T PAYMENT TYPE.



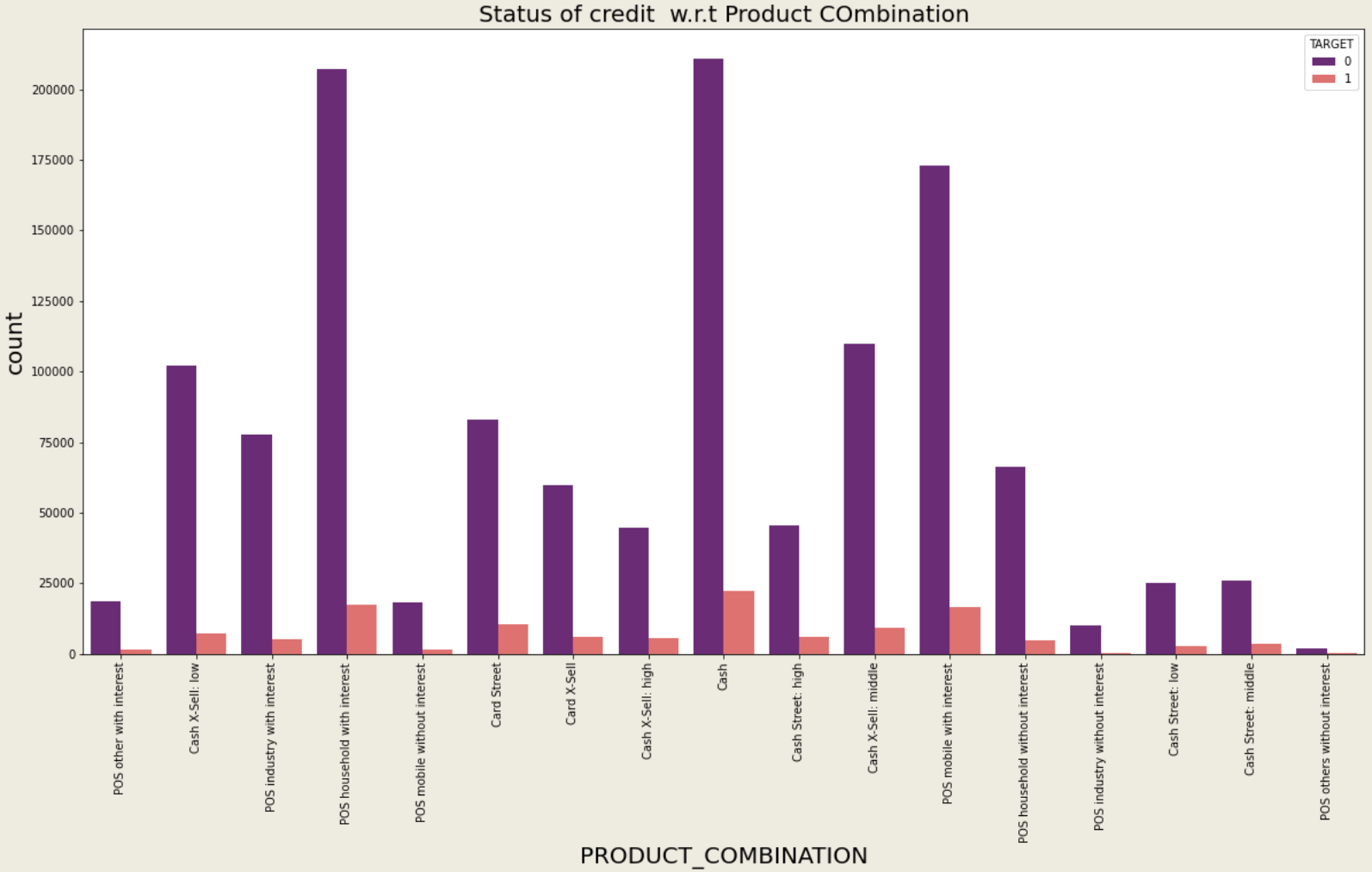
CASH PAYMENT SEEMS TO BE THE MOST PREFERRED WAY FOR THE BANKS, AS MOST OF THE NON- DEFAULTERS HAVE PAID VIA CASH.

ANALYSIS OF THE FAMILY STATUS OF THE CLIENTS.



MARRIED COUPLES ARE THE MOST IN NON-DEFAULTERS. WHICH ARE FOLLWED BY SINGLE, CIVIL MARRIAGE, WIDOW AND SEPERATED

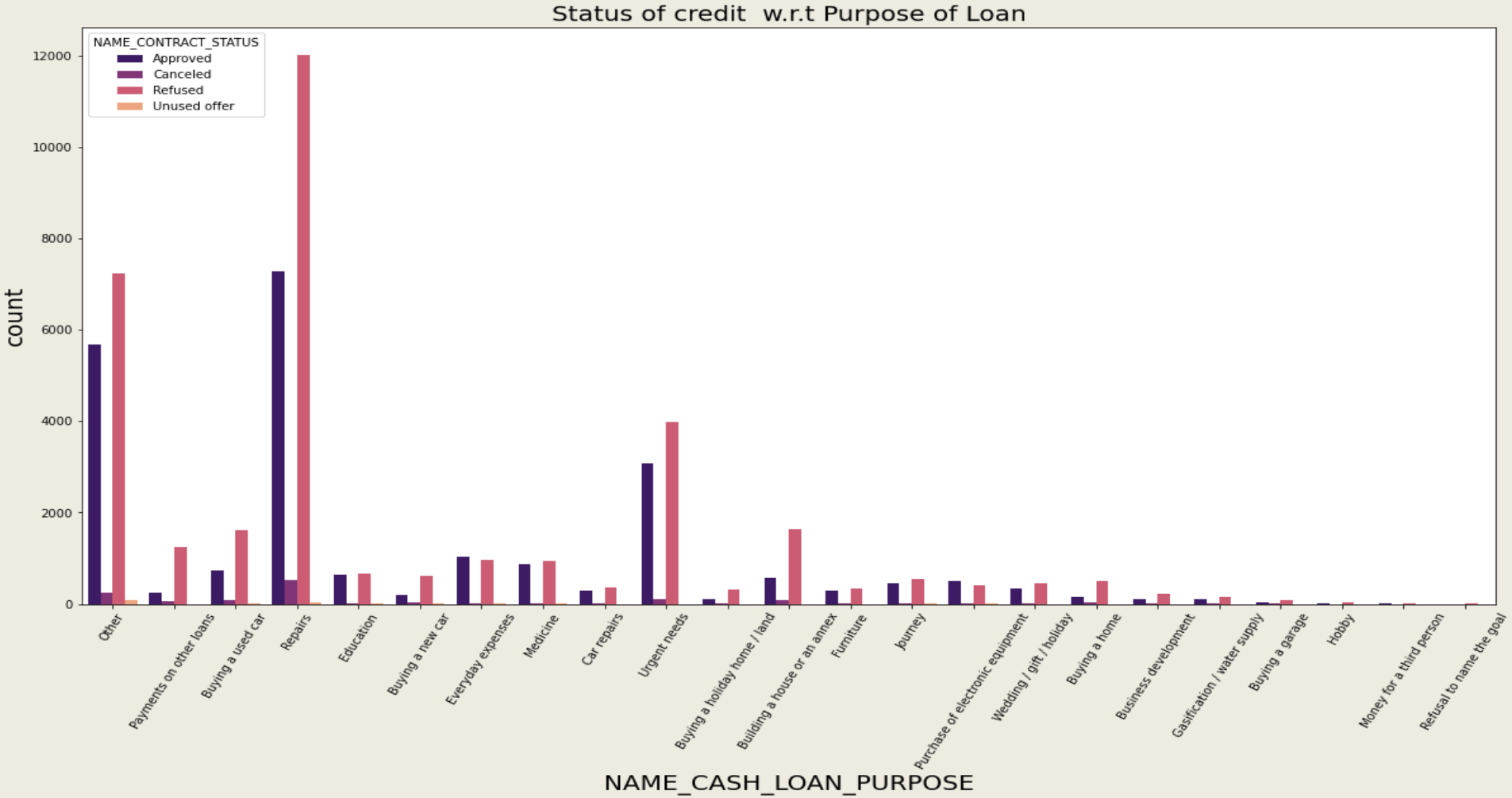
ANALYSIS OF PRODUCT COMBINATION W.R.T DEFAULTERS AND NON-DEFAULTERS.



THE PRODUCT COMBINATION CASH AND POS HOUSEHOLD WITH INTEREST HAS THE HIGHEST VALUE OF NON-DEFAULTERS.

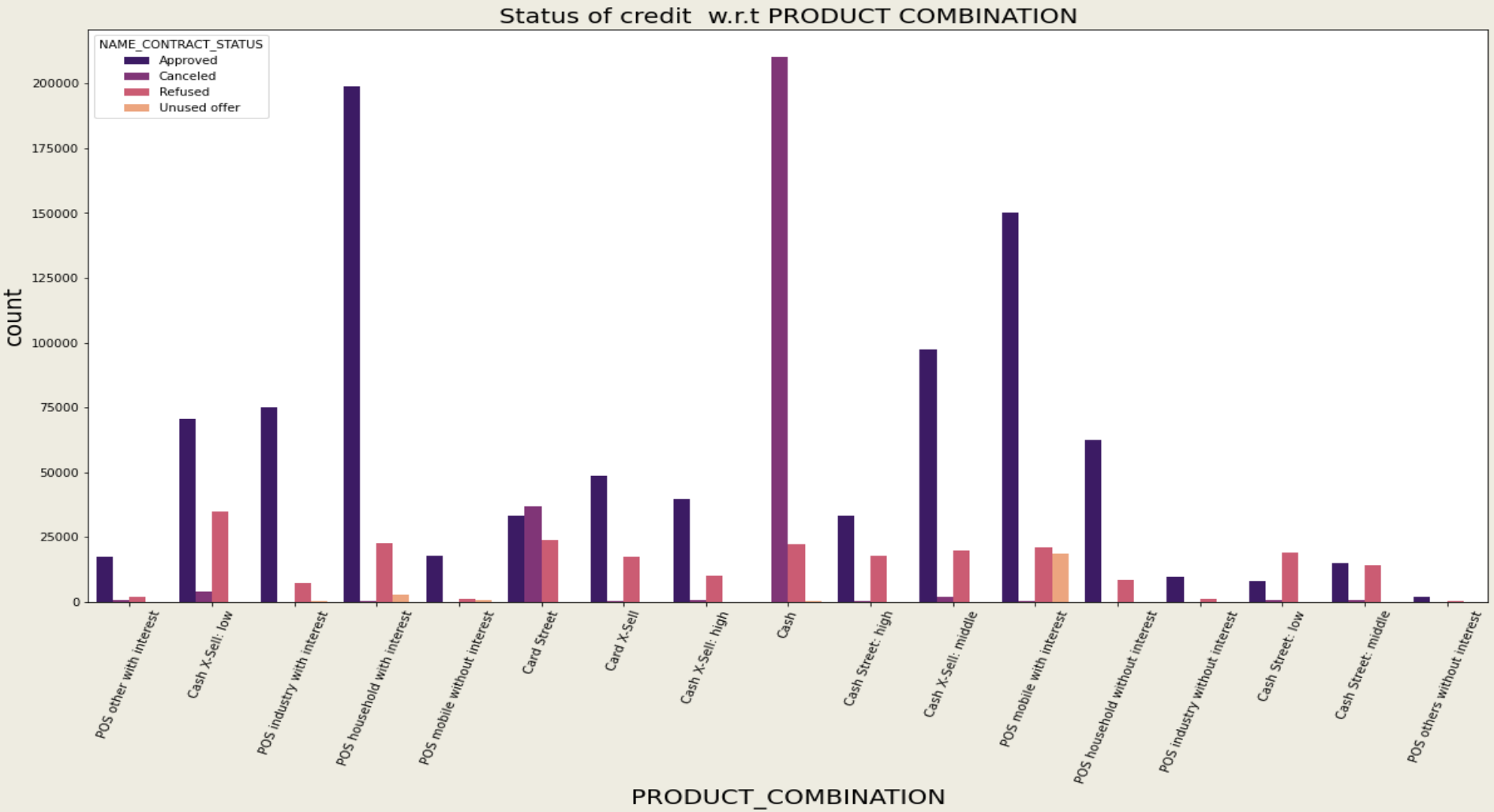
BIVARIATE ANALYSIS

STATUS OF CREDIT W.R.T LOAN PURPOSE.



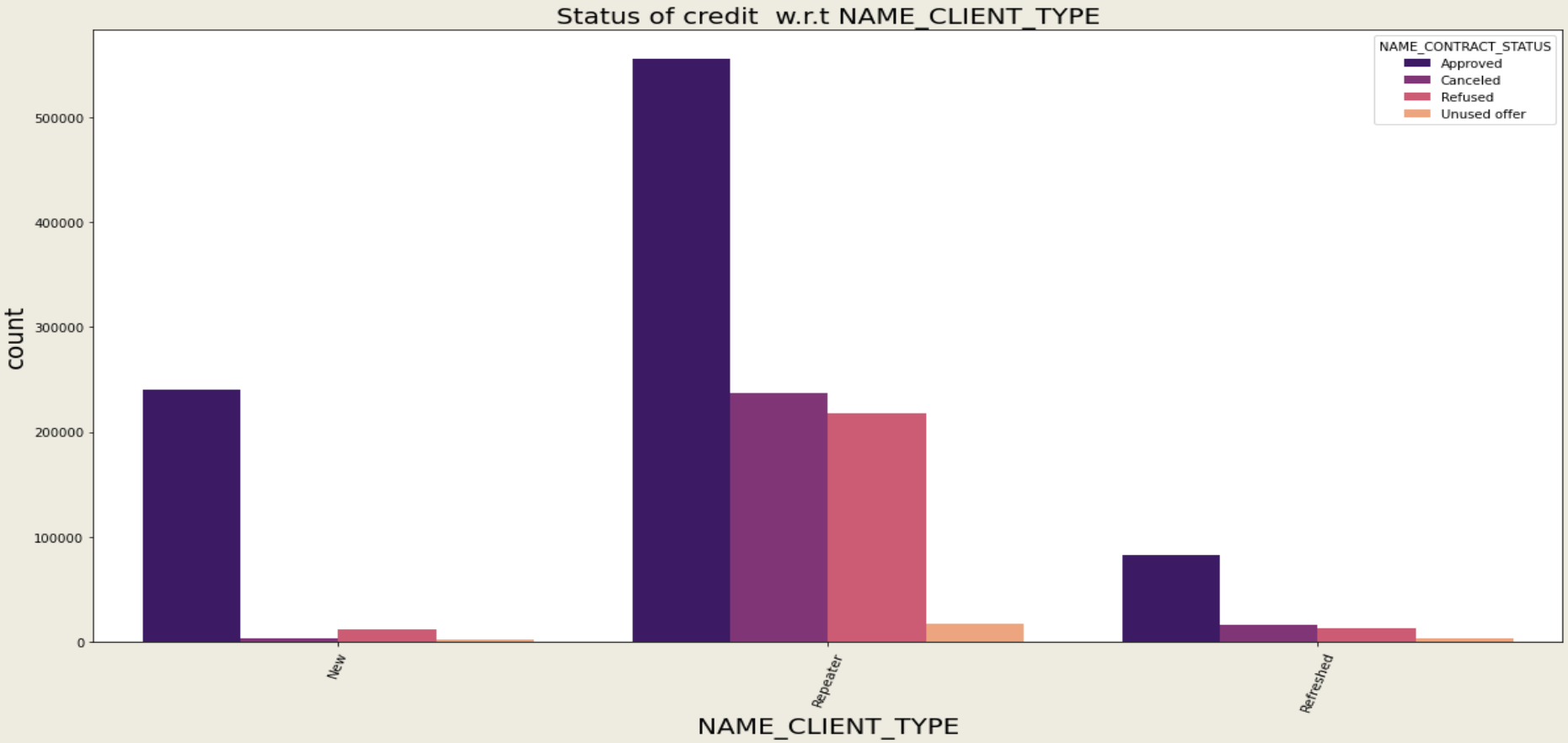
REPAIRS HAVE THE HIGHEST APPROVED AND REFUSED LOANS. THE SECOND HIGHEST IS OF OTHERS CATEGORY. AND IT IS FOLLOWED BY URGENT NEEDS

Status of credit w.r.t PRODUCT COMBINATION



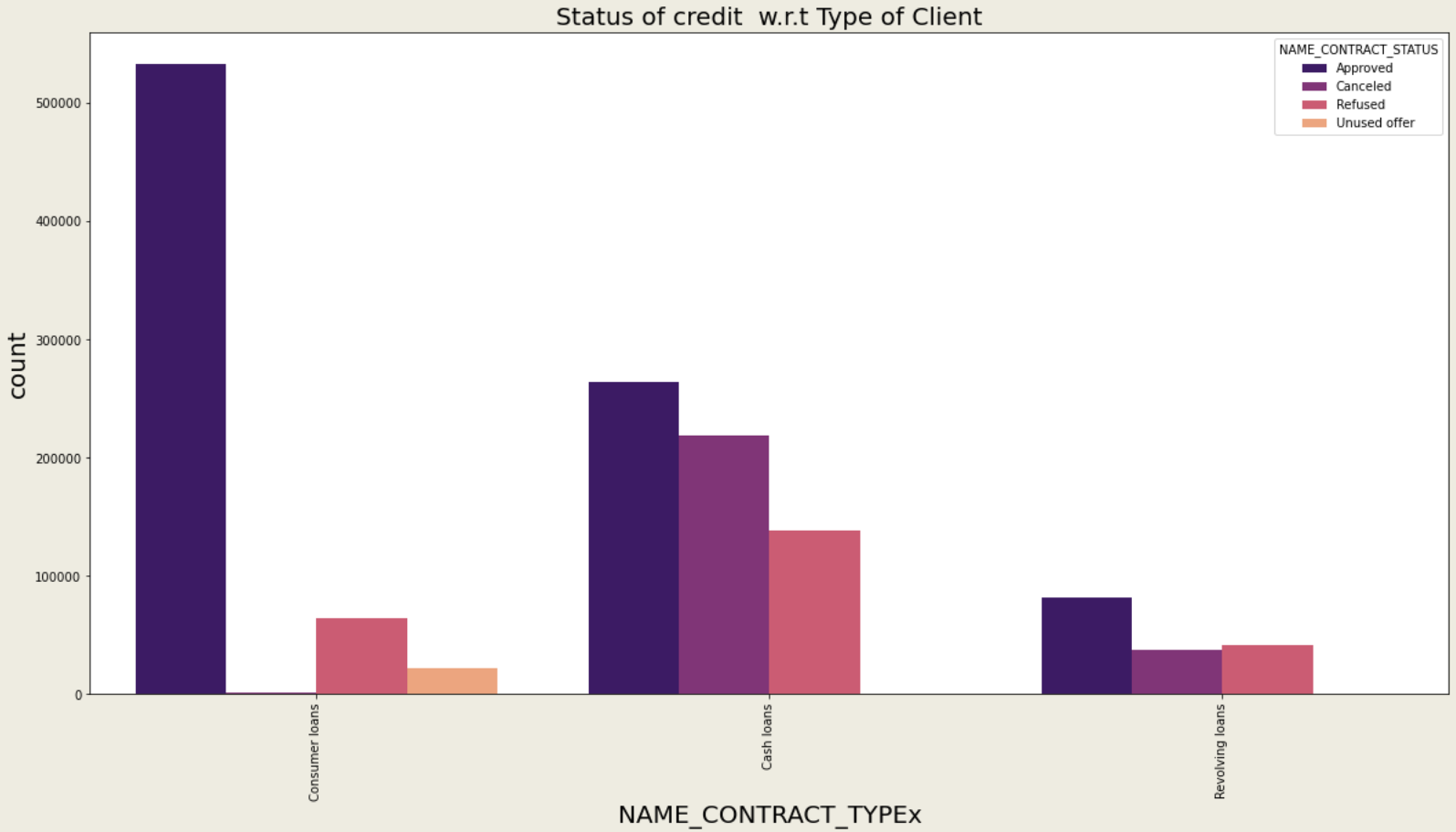
HIGHEST APPROVED IS OF POS HOUSEHOLD WITH INTEREST, HIGHEST CANCELLED IS CASH,HIGHEST REFUSED IS OF CASH, X-SELL.

Status of credit w.r.t NAME_CLIENT_TYPE



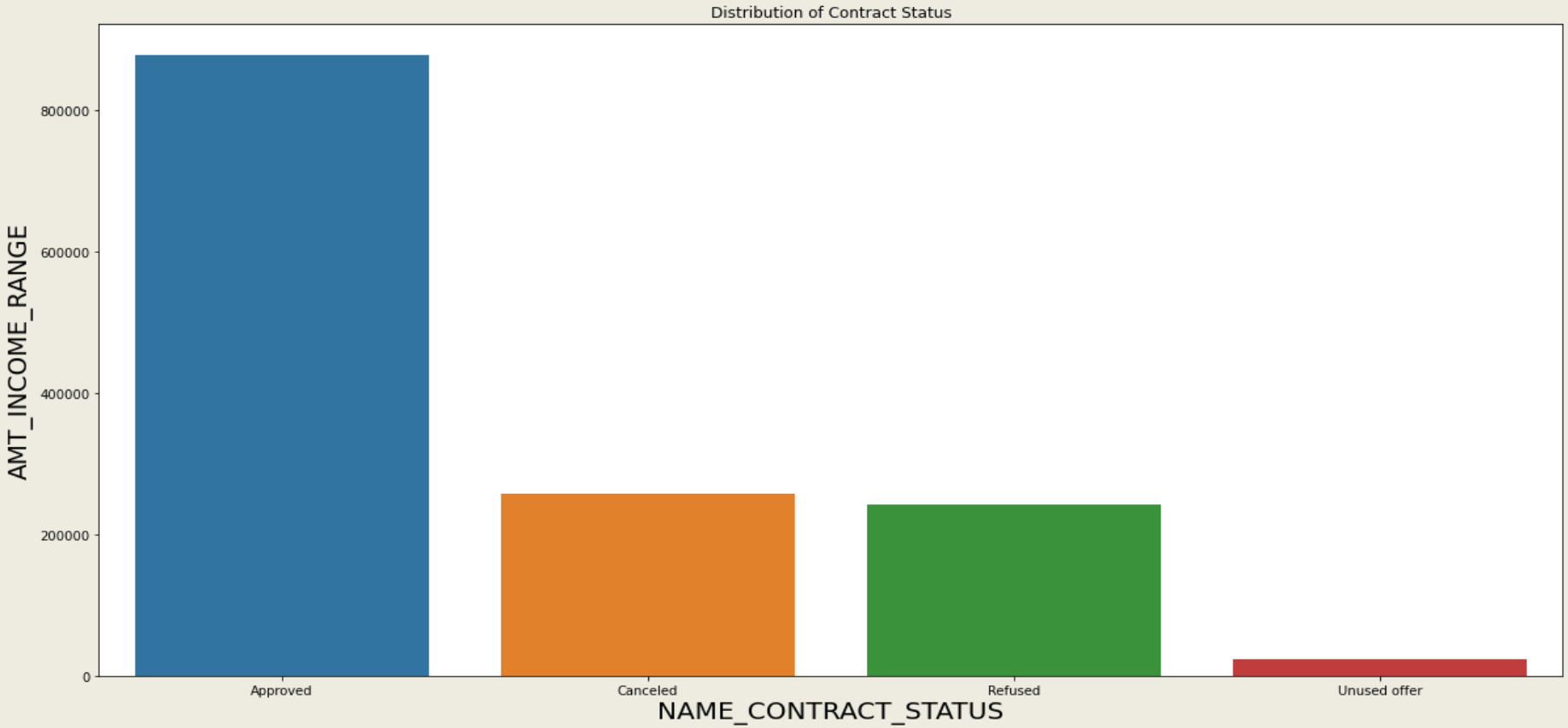
HIGHEST APPROVED IS OF REPEATER WHILE SECOND HIGHEST IS THAT OF NEW CLIENTS.
REPEATERS ONLY HAVE HIGHEST AMOUNT OF CANCELLED, REJECTED AND UNUSED OFFERS.

CONTRACT STATUS W.R.T THE CONTRACT TYPE OF THE PREVIOUS DATA.



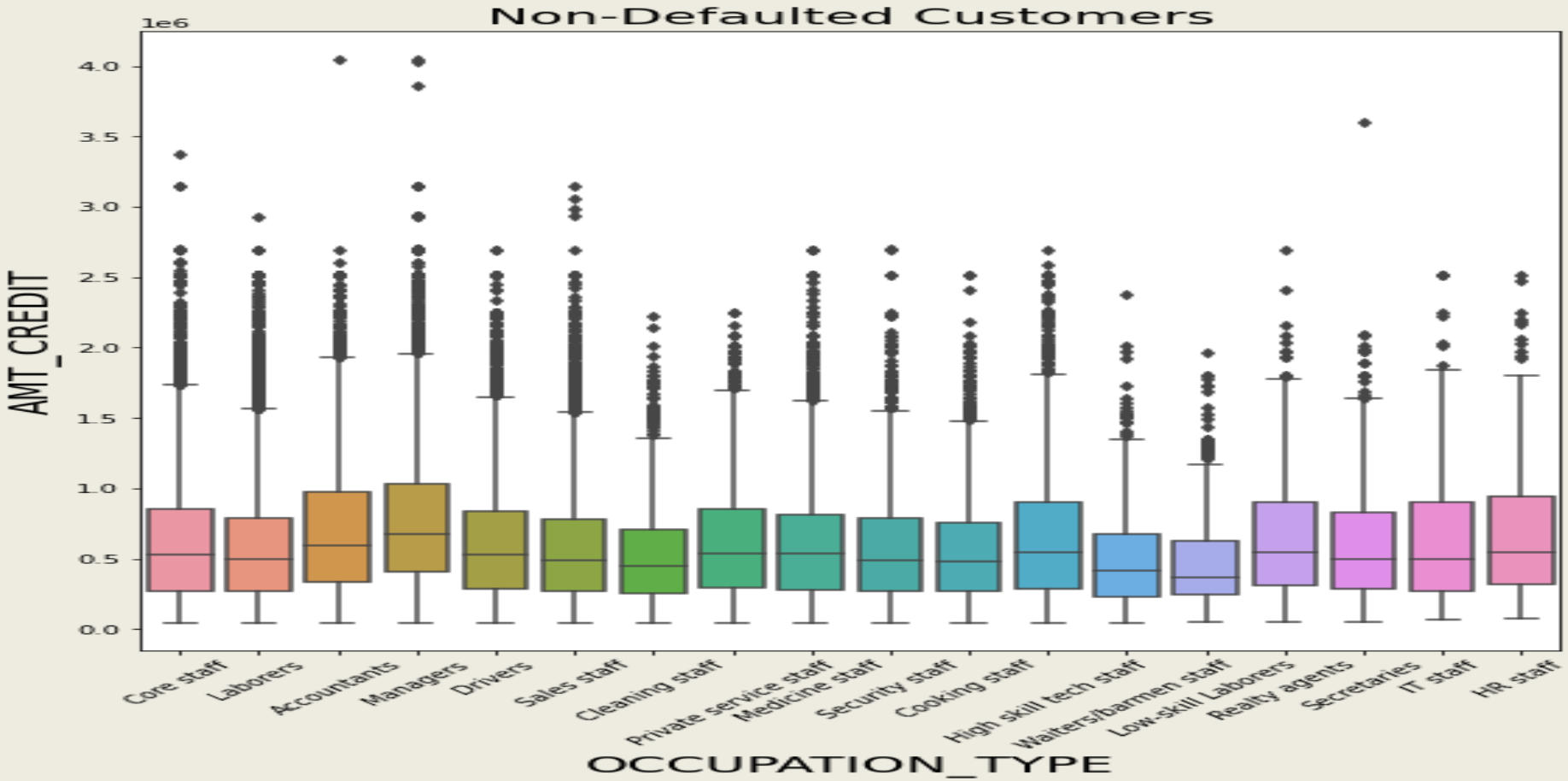
CONSUMER LOANS ARE THE HIGHEST APPROVED IN PREVIOUS DATA, FOLLOWED BY CASH AND REVOLVING LOANS.THE HIGHEST CANCELLED AND REFUSED ARE OF CASH LOANS.

ANALYSIS OF CONTRACT STATUS W.R.T TO INCOME RANGE.



THE HIGHEST INCOME FOR THE APPROVED LOANS ARE MORE THAN 800000, WHILE THAT FOR CANCELLED AND REFUSED LOANS ARE LESS THAN 300000, WHILE UNUSED OFFERS ARE LESS THAN 50000

analysis of occupation type w.r.t the credit amount.

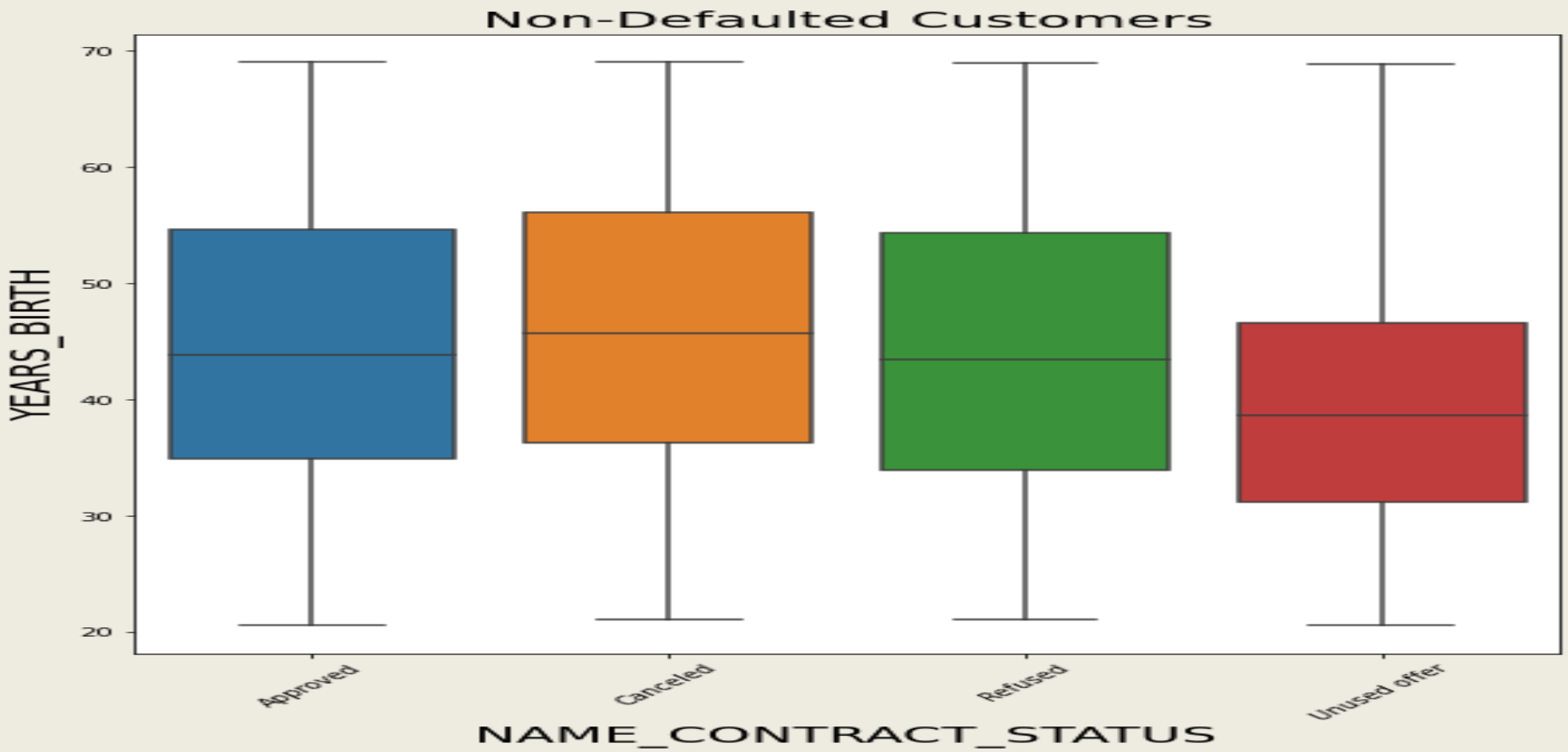


WE SEE THAT THE MANAGERS AND ACCOUNTANTS TEND TO HAVE LARGE OUTLIERS, THEY ALSO SEEM TO HAVE THE HIGHEST CREDIT AMOUNT COMPARED TO OTHERS.

CONTRACT STATUS W.R.T CNT PAYMENT.

APPROVED CNT_PAYMENTS ARE BETWEEN 0-40, WITH SOME OUTLIERS, WHERE AS CNACELLED LOANS EXIST ACROSS ALL CNT_PAYMENTS. MOST OF THE REFUSED LOANS EXIST IN THE RANGE OF 10-40.

ANALYSIS W.R.T AGE OF THE CLIENT AND THE CONTRACT STATUS.



THE AGE GROUP OF MOST OF THE CLIENTS LIES ACROSS 35-55. THE YOUNGEST AGE SEEMS TO BE IN THE UNUSED OFFER, MEANING, IF THE CLIENTS ARE YOUNG, THEY MAY NOT REALLY, GO FOR THE LOAN. THERE IS NOT MUCH DIFFERENCE IN THE CLIENT AGE DISTRIBUTION GROUP OF APPROVED, CACELED AND REFUSED

INFERENCE

1) Cash Loans are much preferred by the clients, and a significant number, of no-defaulters have paid cash loans.
2) A lot of non- defaulters are not car owners, we can assume that, the non-defaulters pay, when they have less overheads.
3) A lot of home owners have repaid their loan.
4) Most working professionals and business owners are non-defaulters.
5) Secondary education dominates in most defaulters and non-defaulters.
6) Married couple tend to repay their loan, with widows being the least chance of paying their loan.
7) Majority of loan applicants and non-defaulters are females.
8) Age group of 35-45 tend to default less.
9) High earners tend to default less.
10) As the value of the item for ehich the loan increases, the loan value also increases.
11) Credit range of 250000 to 500000 have availed cash loans.
12) Females earn more than males and their income ranges are between 100000=250000
13) Higher education and secondary education have availed more credit compared to other categories.
14) People with academic degree have higher incomes, so chances of them defaulting is low.
15) Accountants, managers, high skill tech staff have availed higher credit compared to other professions, and have a less chance of defaulting.

17) Approved loans are the highest compared to refused and cancelled loans, for the previous applications.
18) Secondary and secondary special education types tend to default less, compared to others.
19) Non-Defaulters have availed loans for repairs, oter needs and urgent needs, compared to others.
20) Repeaters tend to pay their loans successfully.
21) Repaying the loan through cash is much preferred by the non-defaulters.
22) If loans are availed citing repair reasons, the chances of the loans being approved are high.
23) Higher the income range, higher the chance of loans being approved.
24) Consumer loans are highest approved in the previous data, followed by cash and revolving.

DRIVER VARIABLES.

Driver Variables to Identify Defaulters/ Non Defaulters.

1) NAME_EDUCATION_TYPE
2) NAME_INCOME_TYPE
3) DAYS_BIRTH
4) AMT_INCOME_TOTAL
5) NAME_CASH_LOAN_PURPOSE
6) CODE_GENDER
7) NAME_FAMILY_STATUS
8) OCCUPATION_TYPE
9) AMT_CREDIT
10) AMT_INCOME