EDA CREDIT CASE STUDY

BY –

SMARANIKA SAHOO & NITISH RATHORE

Load Required Library & Filter Warnings

- This Step involve includes importing of some libraries of python along with library to ignore Warnings.

Load the Data

- This Step includes reading the .csv file through Panda Library .

Check the Data

Glimpse The Data

- Using Shape, describe, info functions to check No. of column, No. of Rows, Data types and Few Mathematical Stats like Standard Deviation, mean, max, various Quartiles.

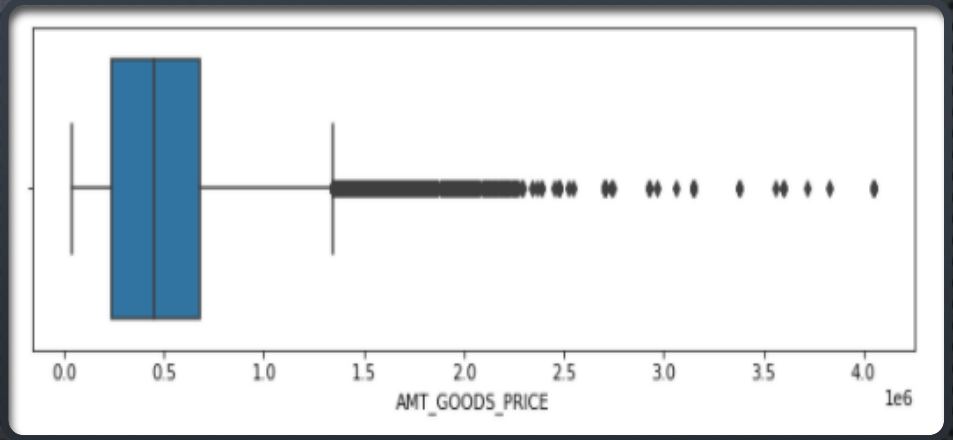
Check missing Values

- This Steps Involve Checking the Null Values Count & Percentage of Null Values in Every Columns.
- Moving Ahead, We'll Remove the Columns where Percentage of Null Values is Greater Than 50 %.
- Check the data and Remove Columns which are non of Use for to make Inference Further.

IMPUTATION OF NULL VALUES

Basically, there are two ways to Impute Null Values from Respective Columns.

- For Continuous variables :- To impute Null Values from Column with Continuous variables, Generally Median is used if There is Outlier in Data otherwise Mean can be used.
- For Categorical variables :- To impute Null Values from Column with Categorical variables, Generally Mode is used.



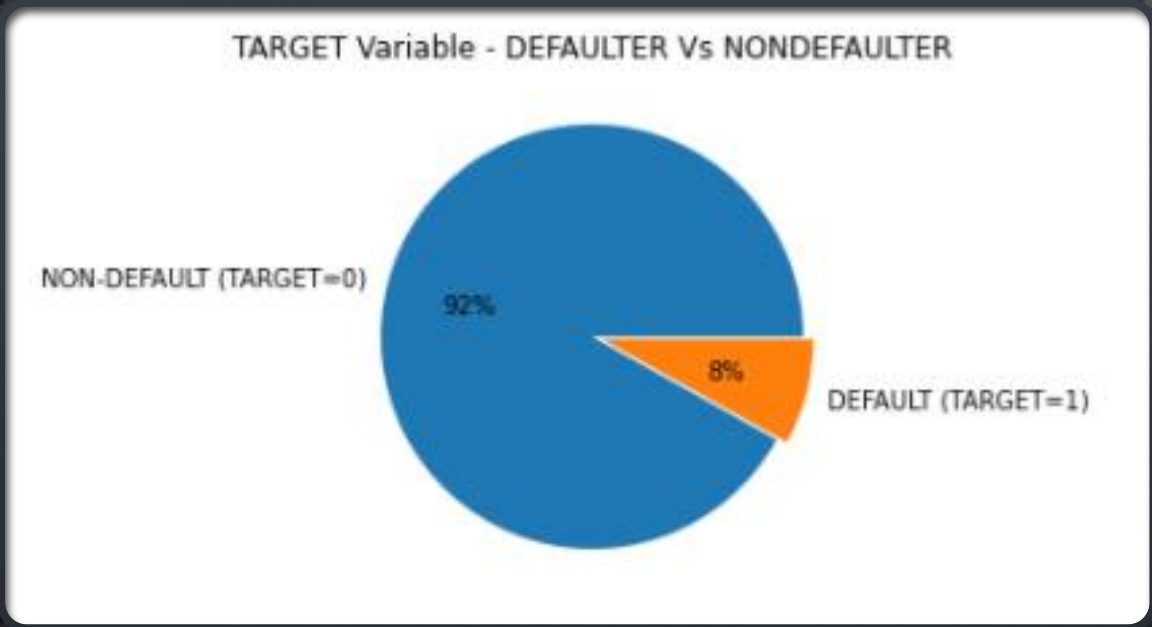
We can see in the Above Diagram, there are Outliers, so Here Median will be used for Imputation.

Note :- For the Given Set of Data, we Have not done Null Value Imputation, As Mentioned in Problem Statement.

Correction in Columns

- Binning of AMT_INCOME_TOTAL and Created New Column as INCOME_GROUP which include Sub-point to Denote Income Range. Similarly Binned AGE column and Created AGE_GROUP.
- Created new Column as AGE from DAYS_BIRTH then Dropped DAYS_BIRTH Column.
- There are few Column with Days as Unit, that having some negative values, so converted those Columns to Positive.

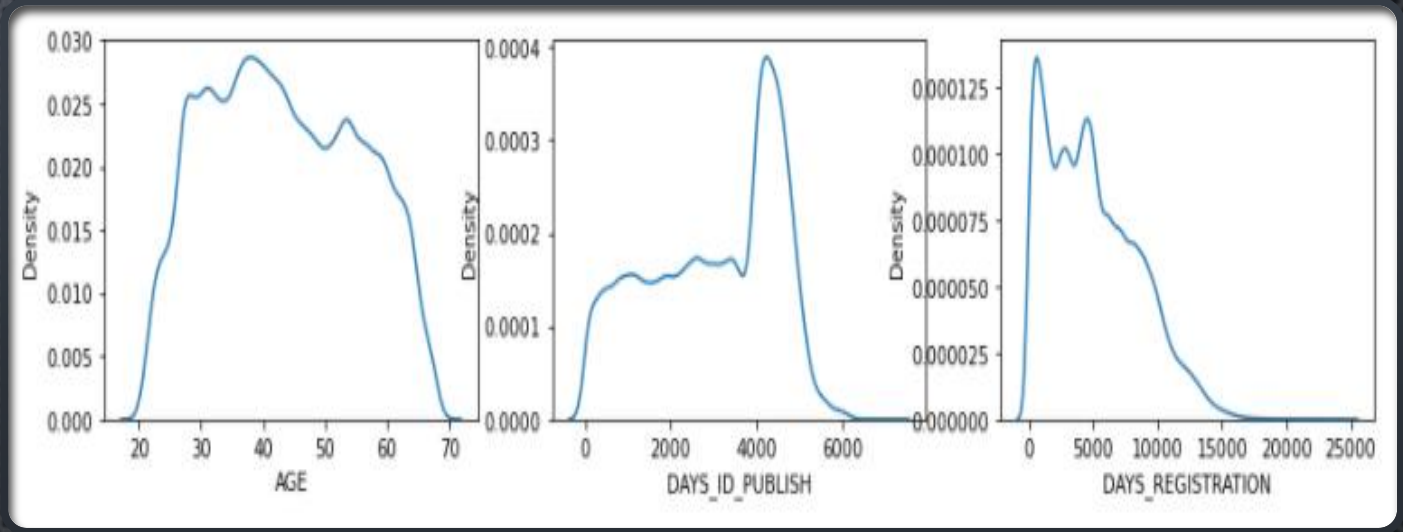
CHECK DATA UNBALANCE



- We are checking the data unbalance for the **TARGET** variable.
- This is to find the ratio between the Non-Default clients & Default clients .
- **Target 1** means Default clients which is 8%
- **Target 0** means Non- Default clients which is 92%
- Ratio between both the cases is 11.39

UNIVARIATE ANALYSIS

(DISTPLOT FOR **AGE**, **DAYS_ID_PUBLISH** & **DAYS_REGISTRATION**)



INFERENCES

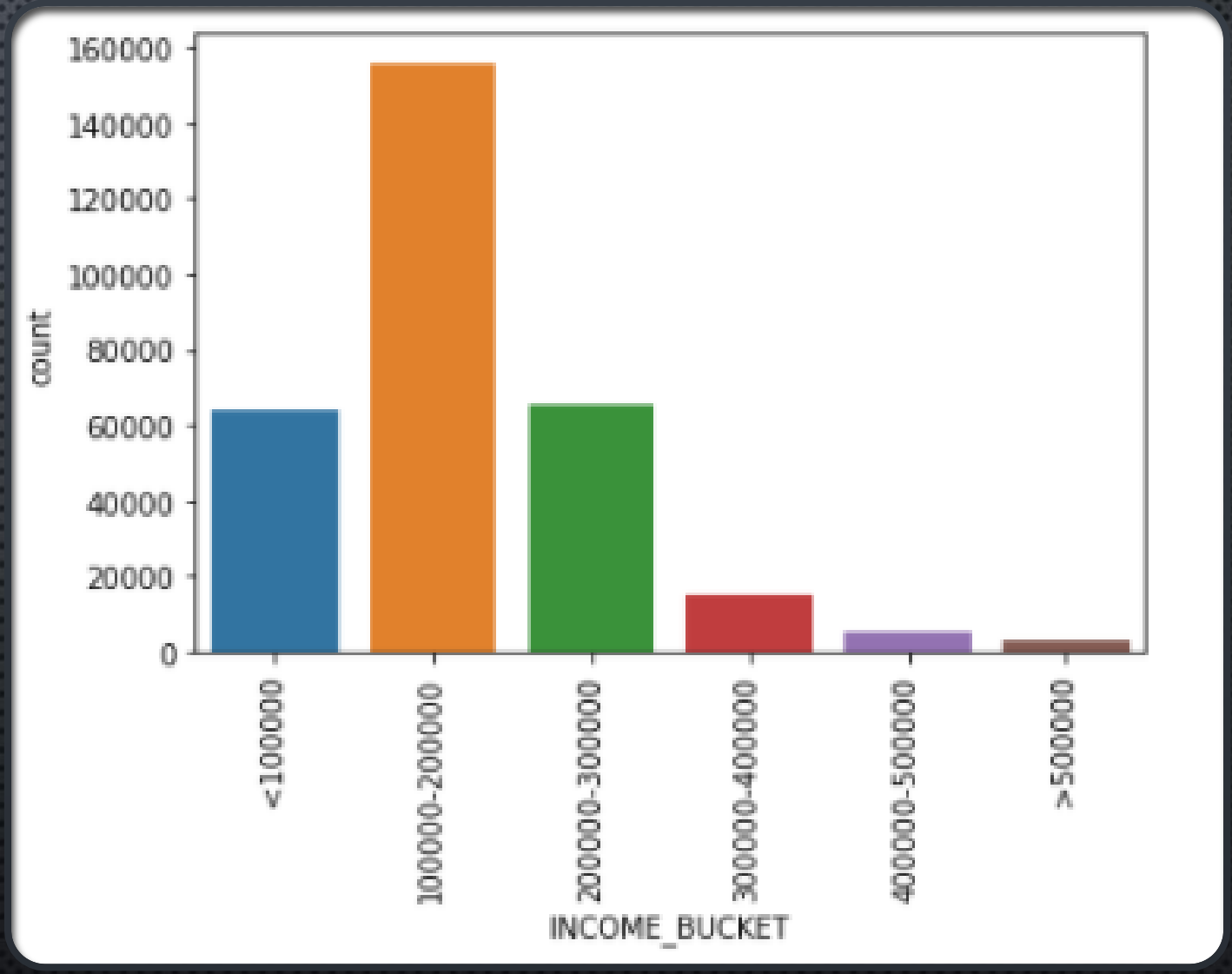
1. MOST PEOPLE WHO HAVE TAKEN LOANS ARE IN B/W 30-40 YEARS OF AGE.
2. CLIENTS WHO HAVE CHANGED THEIR ID DOCUMENTS DID IT MOSTLY 4000-5000 DAYS BEFORE APPLYING THE LOAN.
3. CLIENTS CHANGING THEIR REGISTRATION WITHIN 5000 DAYS ARE TAKING MOST LOANS.

UNIVARIATE ANALYSIS

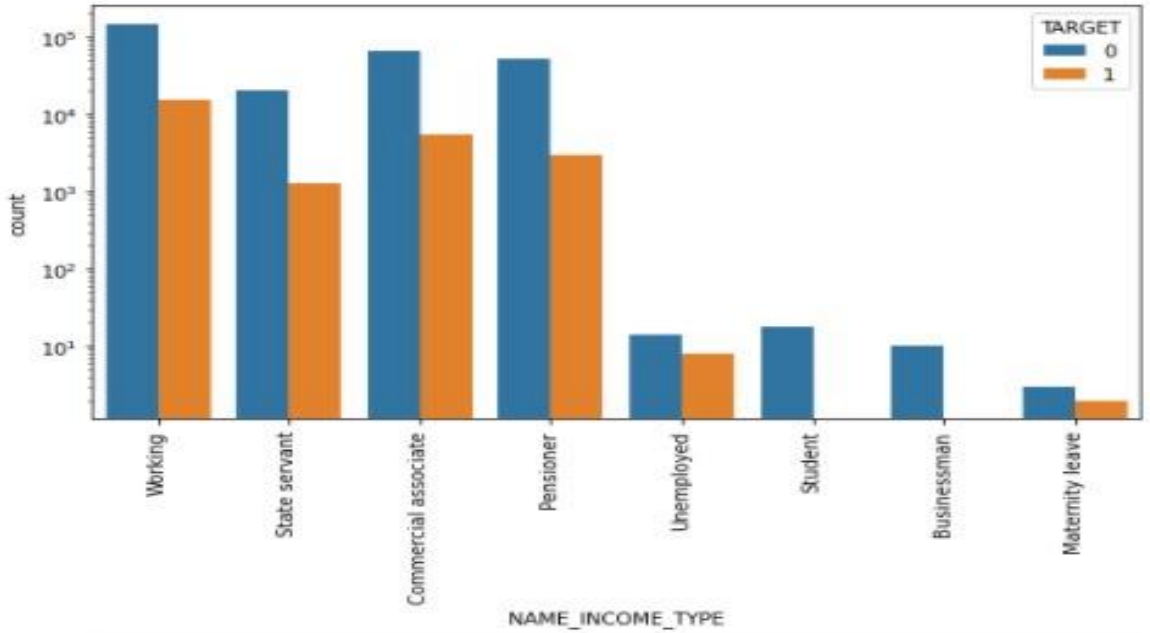
(PLOTTING SUBPLOT FOR **INCOME_BUCKET**)

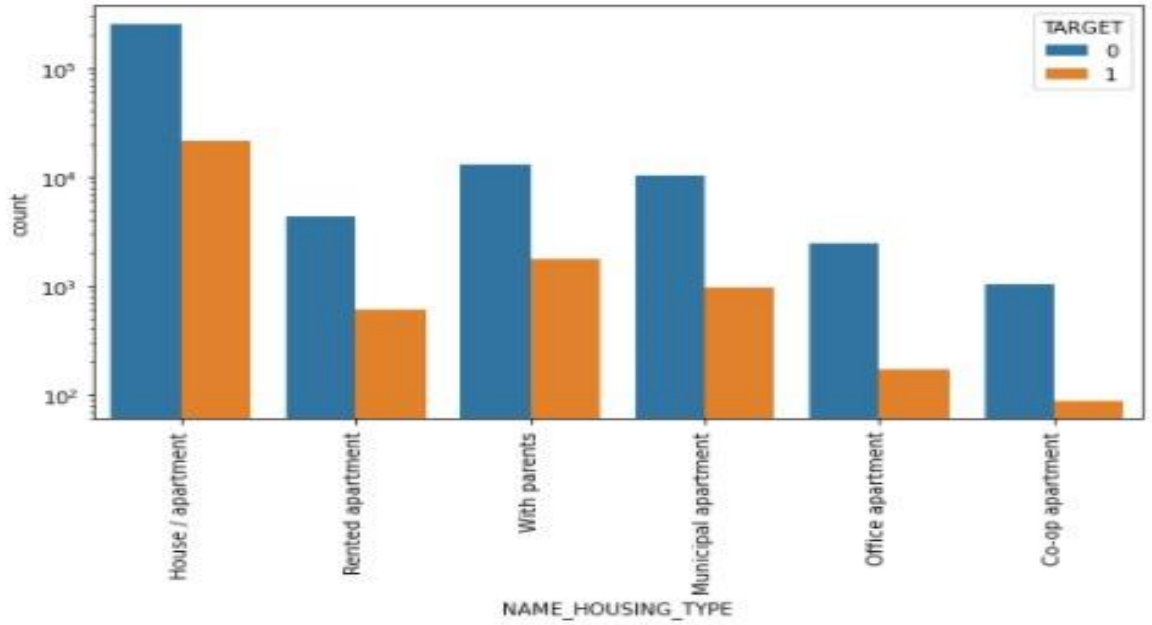
INFERENCE

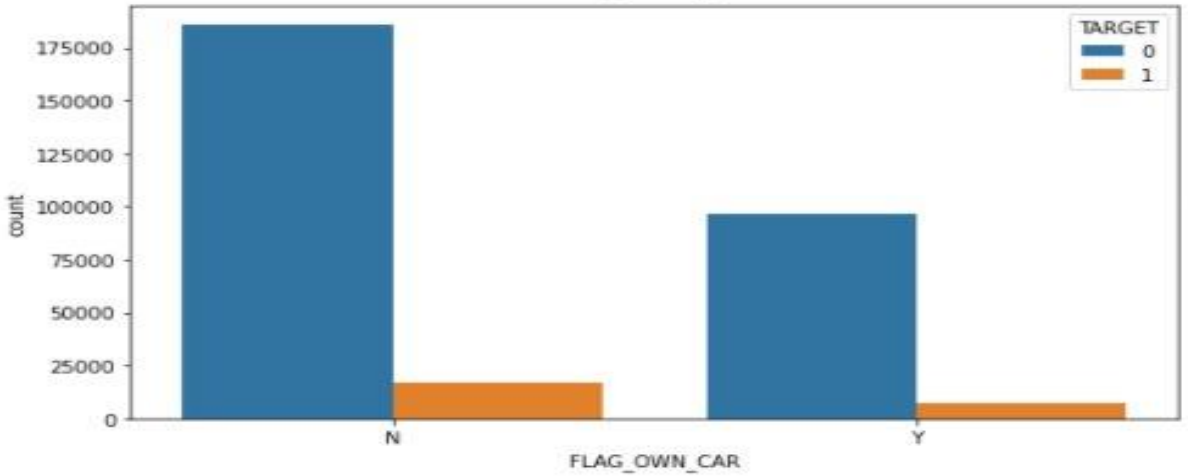
- PEOPLE IN 100000-200000 HAVE TAKEN MOST LOANS
- PEOPLE WITH SALARY 200000 AND MORE ARE LESS LIKELY TO TAKE LO ANS COMPARISON TO PEOPLE WITH SALARY LESS THAN 200000

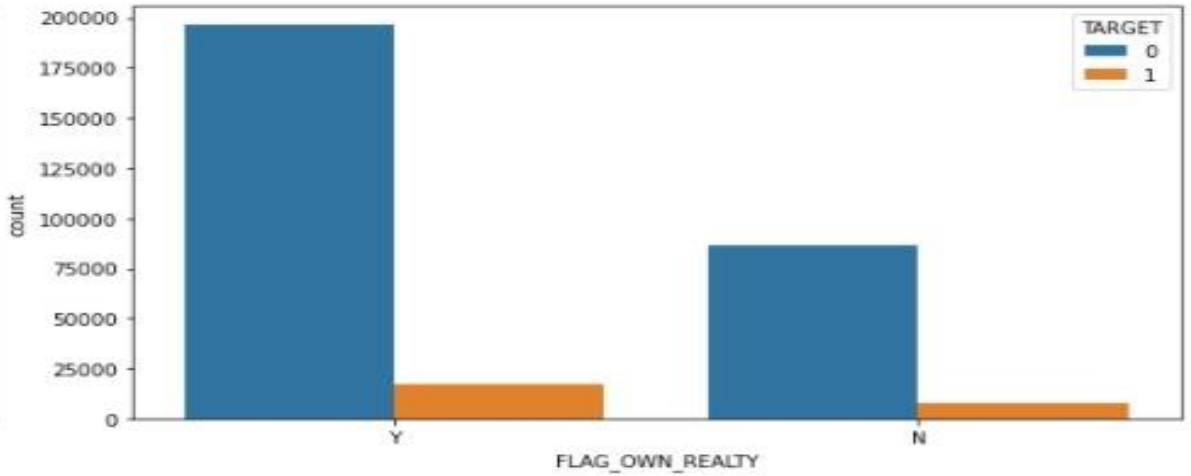


COMPARISON WITH DEFAULT AND NON-DEFAULT









INFERENCES

1. The income bucket 100000-200000 have highest defaulters also highest non defaulters.
2. The income bucket >500000 have comparatively less defaulters which means bank will be safer with these people.
3. Working people are most inclined to taking loans also, they are highest in Defaulter group as well
4. Student and Businessman are the safer group to give loans.

HOUSING TYPE:

1. People who live in house/apartment have taken maximum loans and their defaulter percentage is high.
2. People who live in co-op apartment have significantly less defaulters

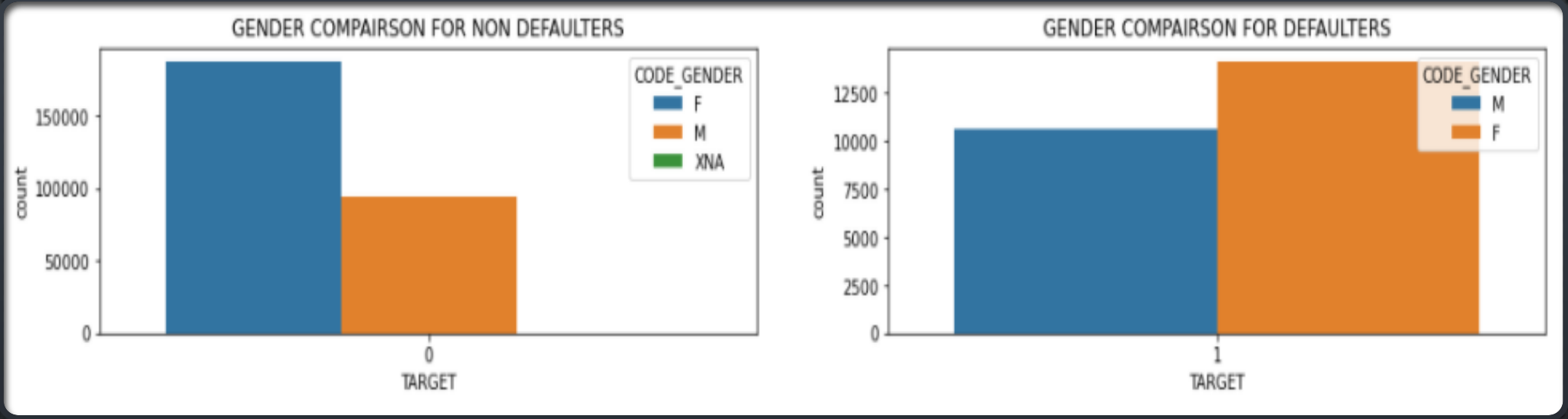
OWNING CAR

1. People who doesn't owns a car have less possibility to become Defaulter.

OWNING A FLAT

1. People who owns a flat have less possibilities to become defaulters.

BIVARIATE ANALYSIS (BETWEEN TARGET AND CODE_GENDER)

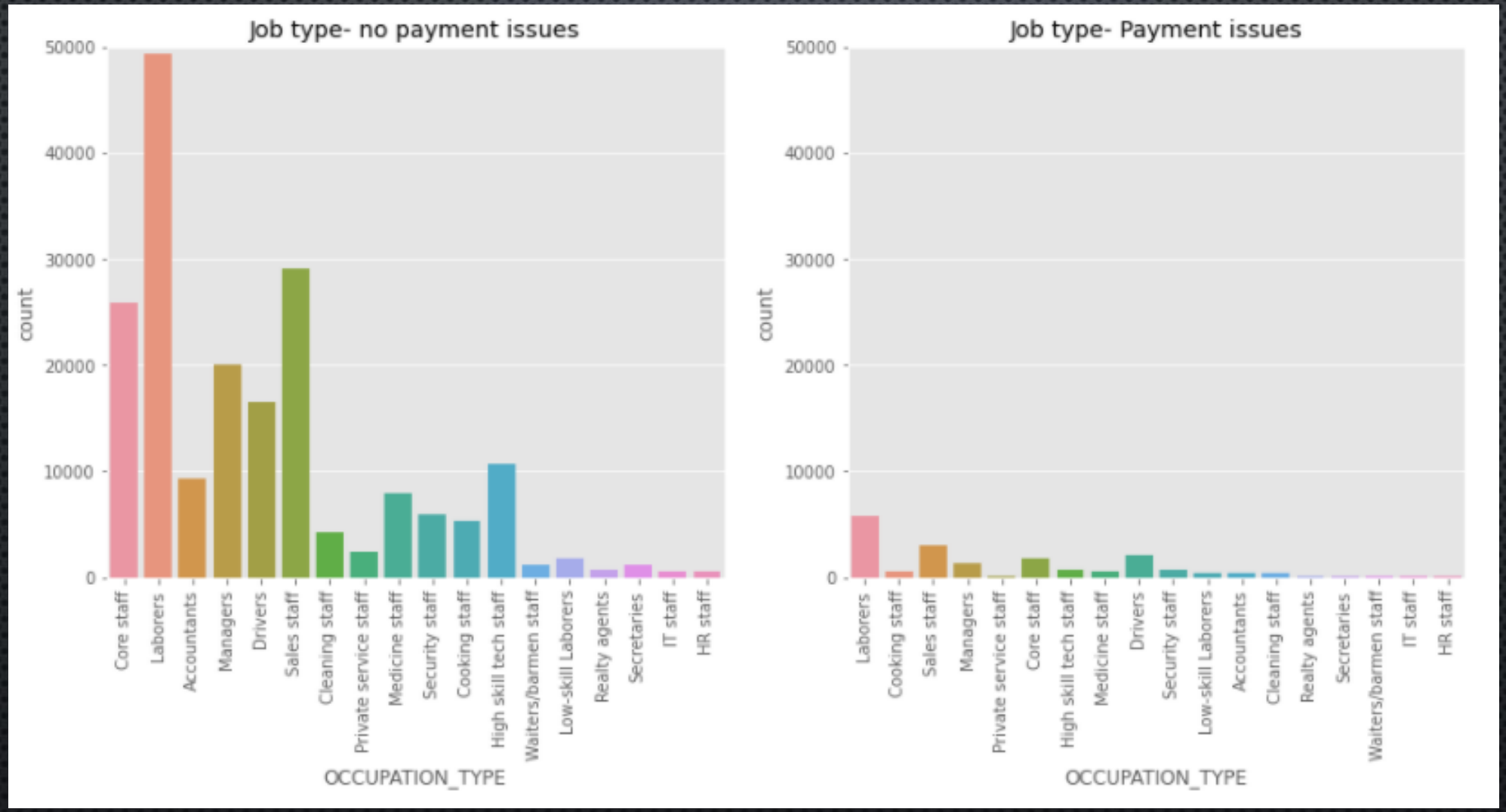


INFERENCES-

FEMALES ARE THE MAJORITY IN BOTH DEFAULTERS AND NON-DEFAULTERS. UNDOUBTEDLY, MOST FEMALES HAVE APPLIED FOR LOAN THAN MALES.

MALES ARE MORE IN DEFAULT THAN IN RATIO WITH MALES IN NON-DEFAULT.

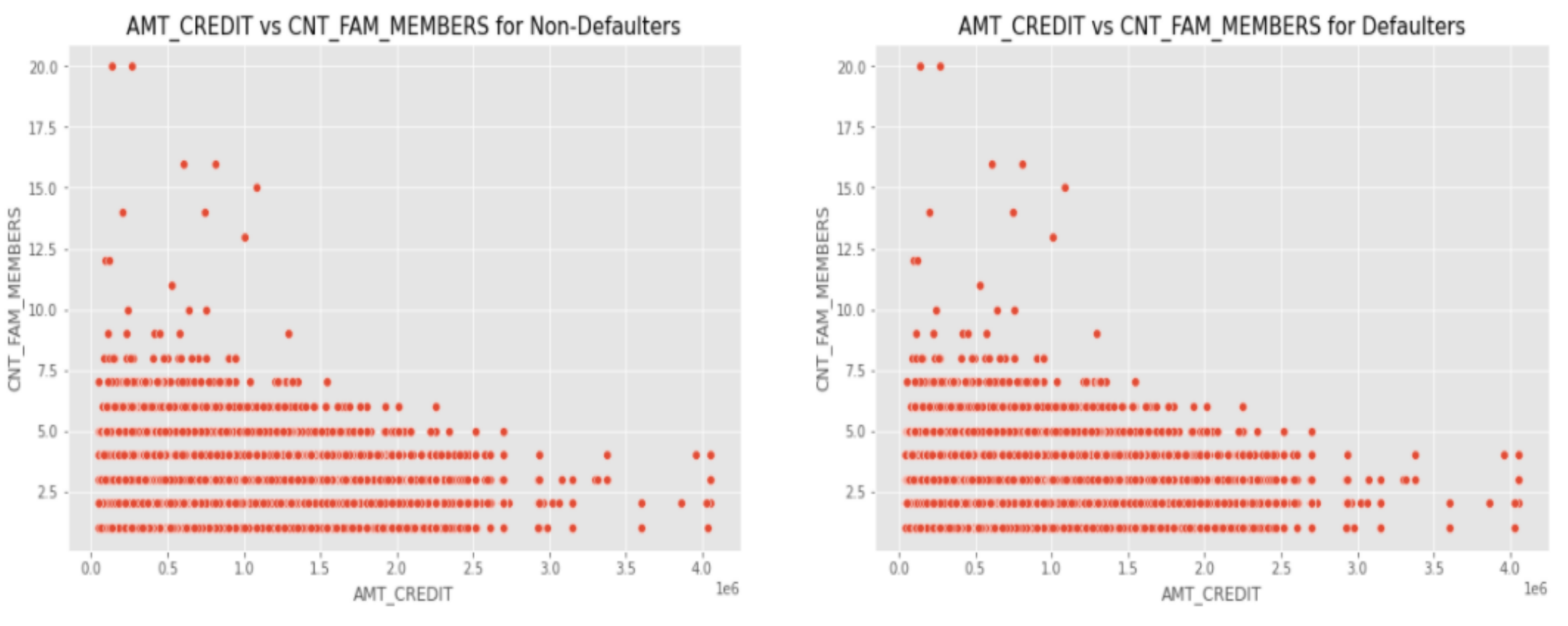
BIVARIATE ANALYSIS (BETWEEN TARGET AND OCCUPATION_TYPE)



INFERENCE-

FROM THE ABOVE GRAPH, WE COULD SEE THAT LABORERS ARE MOST LIKELY TO MAKE PAYMENT ON TIME WHEREAS HR STAFF ARE LESS LIKELY TO MAKE PAYMENT ON TIME.

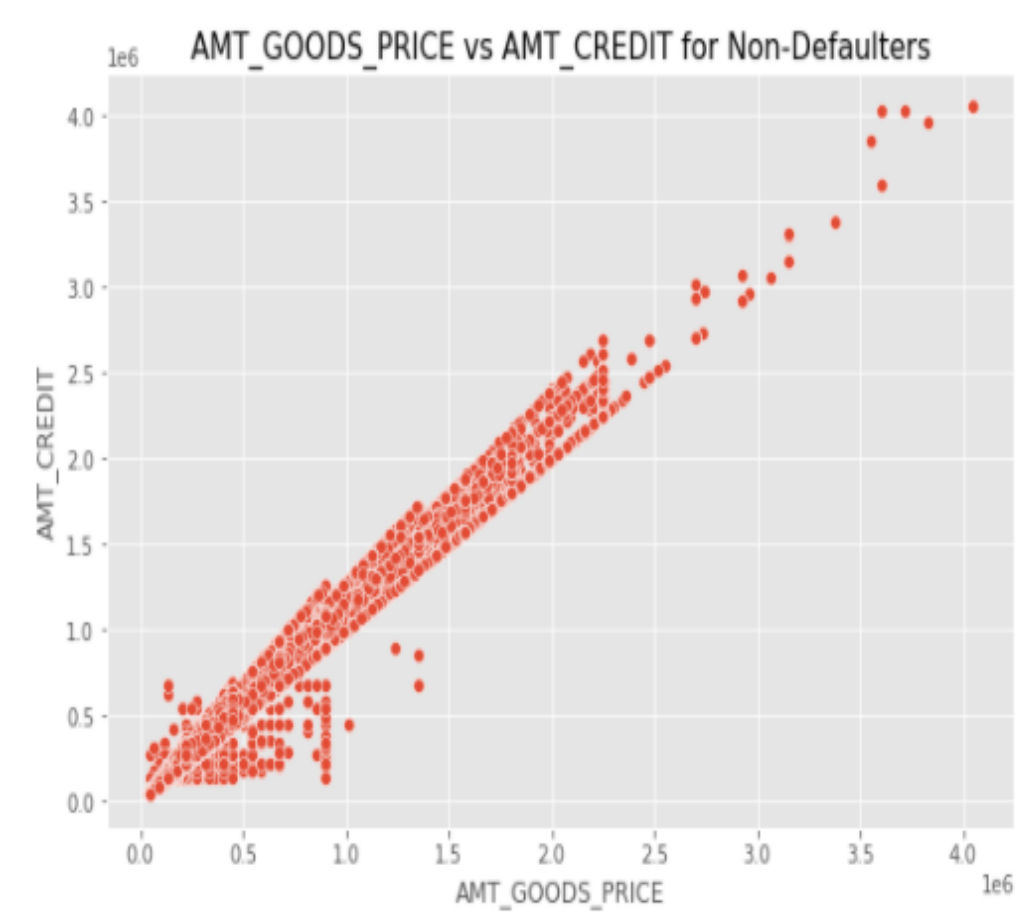
BIVARIATE ANALYSIS (WITH AMT_CREDIT & CNT_FAM_MEMBERS)

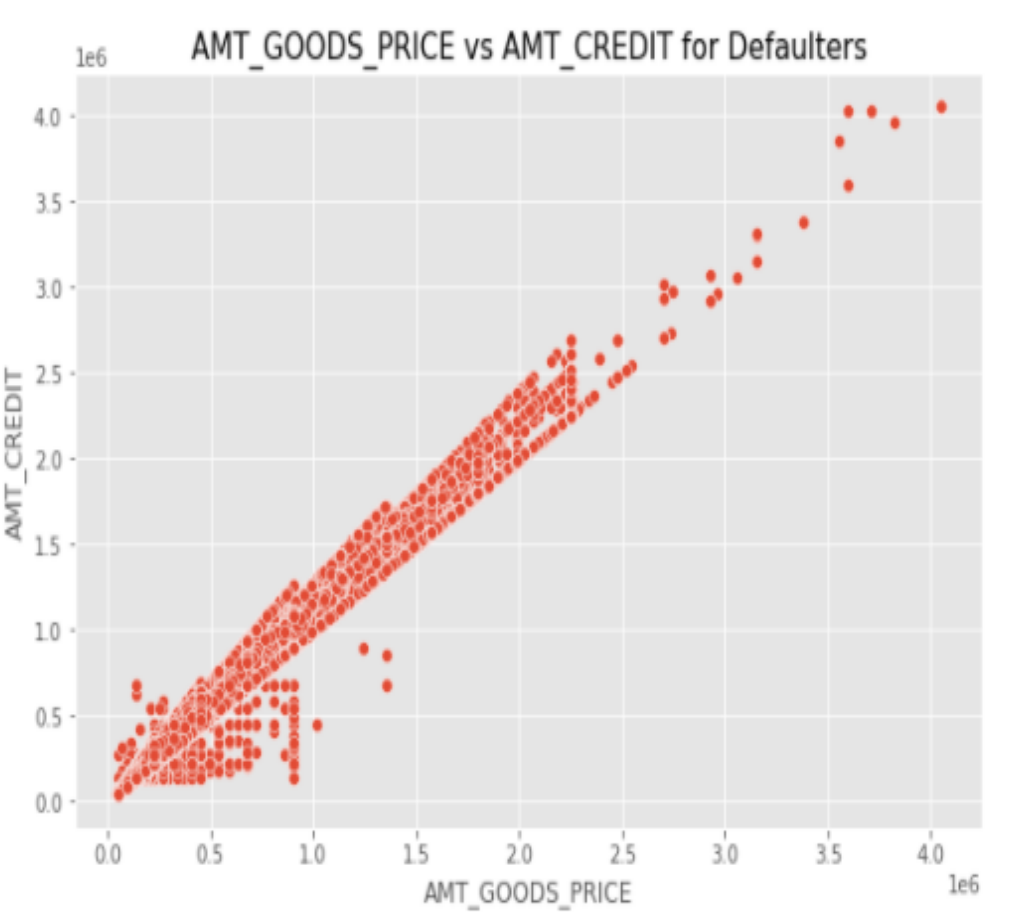


INFERENCES

- WE CAN SEE THAT THE DENSITY IS IN THE LOWER LEFT CORNER, SIMILAR IN BOTH THE CASE. SO, THE PEOPLE ARE EQUALLY LIKELY TO DEFAULT IF THE FAMILY IS SMALL AND THE AMT_CREDIT IS LO W. WE CAN OBSERVE THAT LARGER FAMILIES AND PEOPLE WITH LARGER AMT_CREDIT DEFAULT LESS OFTEN.

BIVARIATE ANALYSIS (WITH AMT_GOODS_PRICE & AMT_CREDIT)

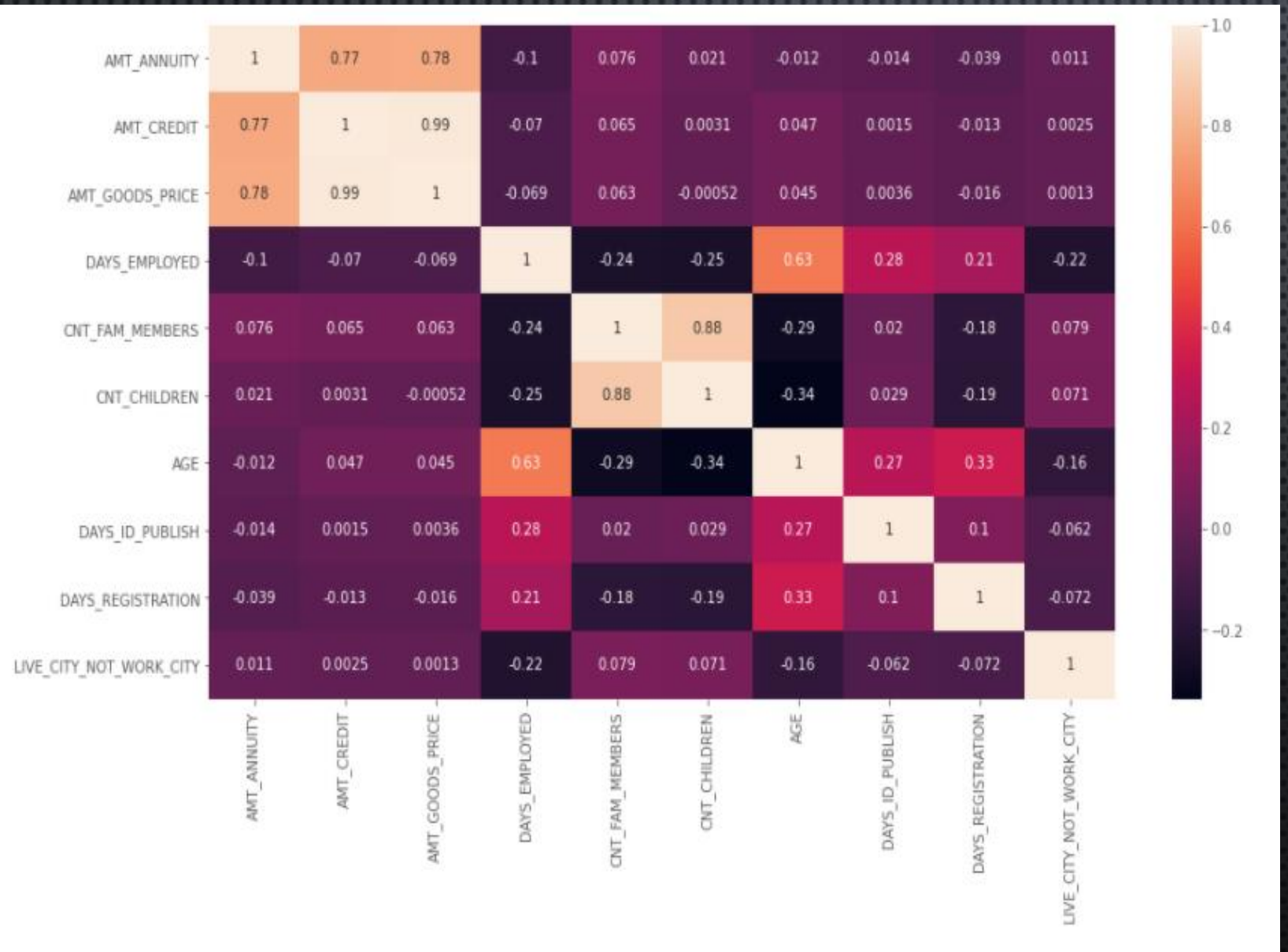




INFERENCE

THE DEFAULTER SEEMS TO BE IN LESS NUMBER WHENEVER THE PRICE OF GOODS IS UPTO 50K AND CREDIT AMOUNT IS EVEN LESSER THAN 50K.

MULTIVARIATE ANALYSIS (FOR VARIOUS COLUMNS IN NON-DEFAULTER DATASET)



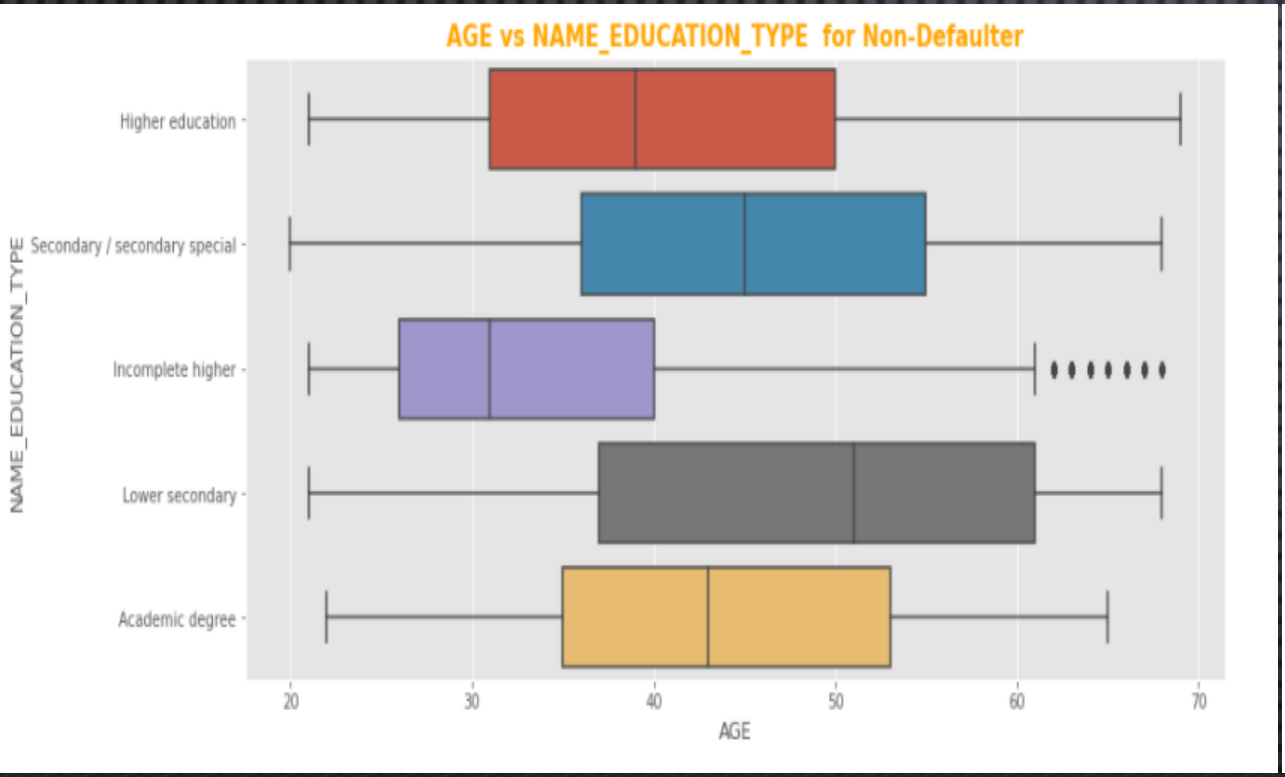
TOP3 CORRELATION IN GIVEN DATASET

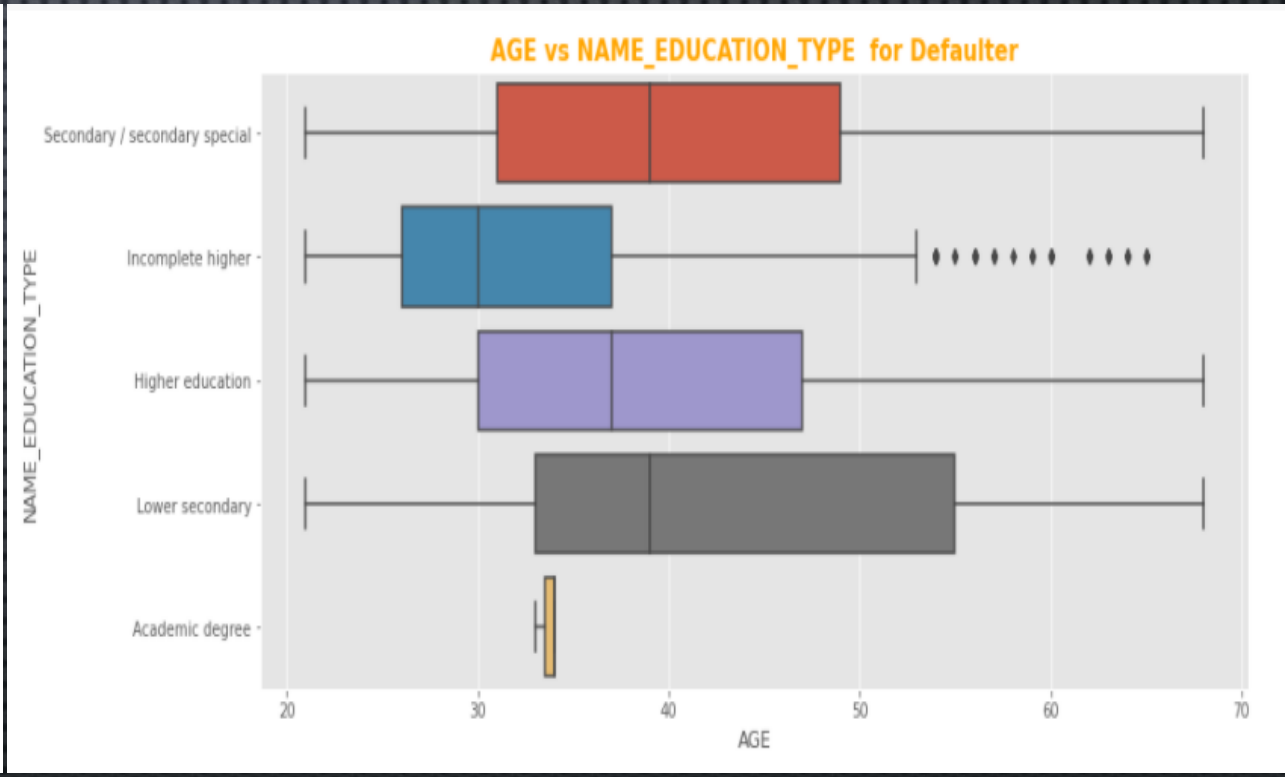
('AMT_GOODS_PRICE', 'AMT_CREDIT')

('CNT_FAM_MEMBERS','CNT_CHILDREN')

('AMT_ANNUITY', 'AMT_CREDIT')

BOXPLOT BETWEEN AGE AND NAME_EDUCATION_TYPE
(FOR DEFAULTER AND NON-DEFAULTER)





INFERENCES

- FOR NON-DEFAULTERS:

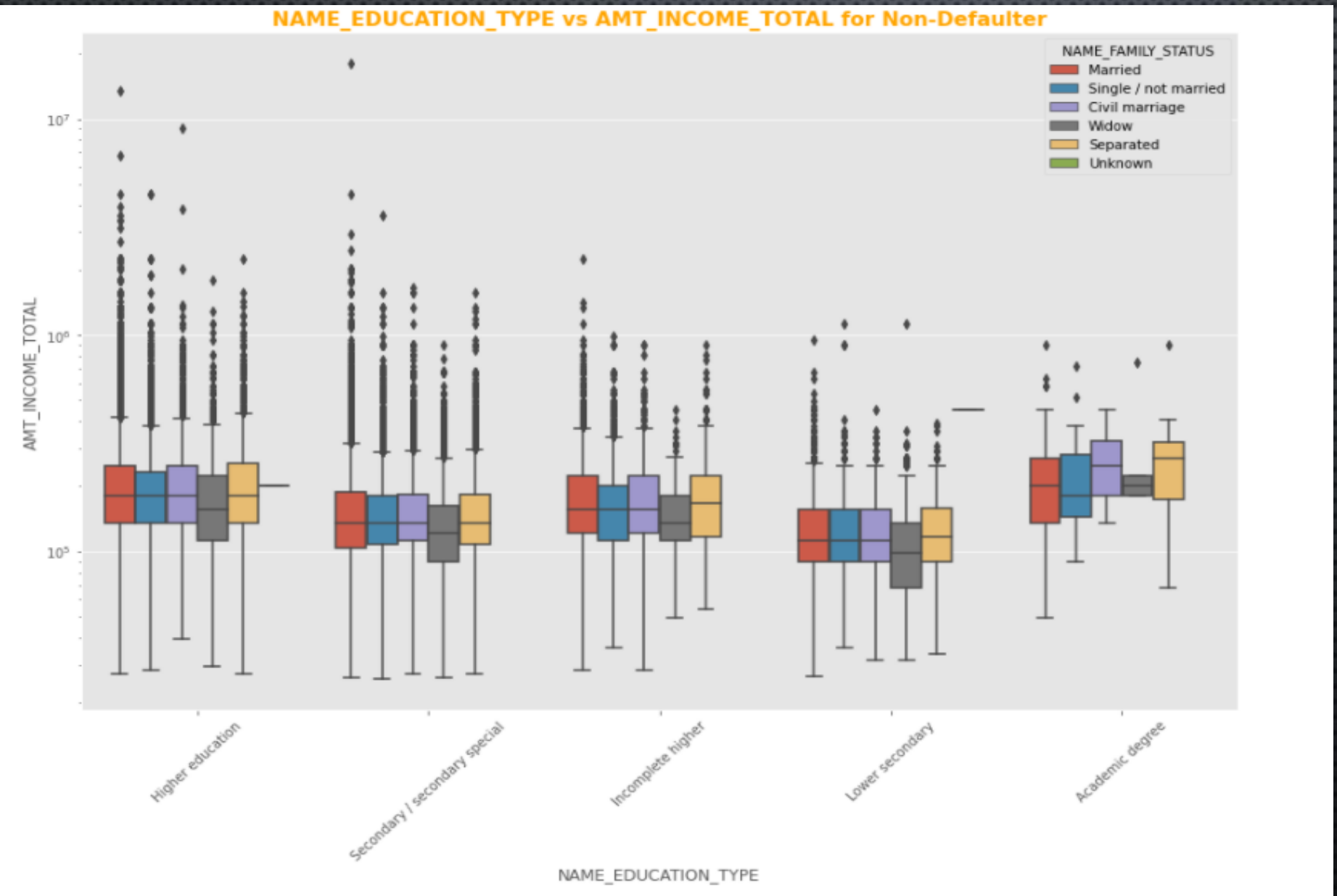
THERE IS AN OUTLIERS IN INCOMPLETE HIGHER IN BETWEEN AGE 60TO70.APART FROM INCOMPLETE HIGHER WHERE FIRST QUARTILE LIES ON AG E 40,OTHER EDUCATION TYPES AREN'T FACING MUCH DIFFICULTIES IN LOAN REPAYMENT.

- FOR DEFAULTERS:

THERE IS A VISIBILITY OF AN OUTLIERS IN INCOMPLETE HIGHER FROM THE AGE GROUP BETWEEN 50-70. PEOPLE WITH AN AGE GROUP 30-40 AND EDUCATION TYPE AS ACADEMIC DEGREE AND INCOMPLETE HIGHER SEEMS TO BE FACING DIFFICULTIES IN LOAN REPAYMENT.

MULTIVARIATE ANALYSIS

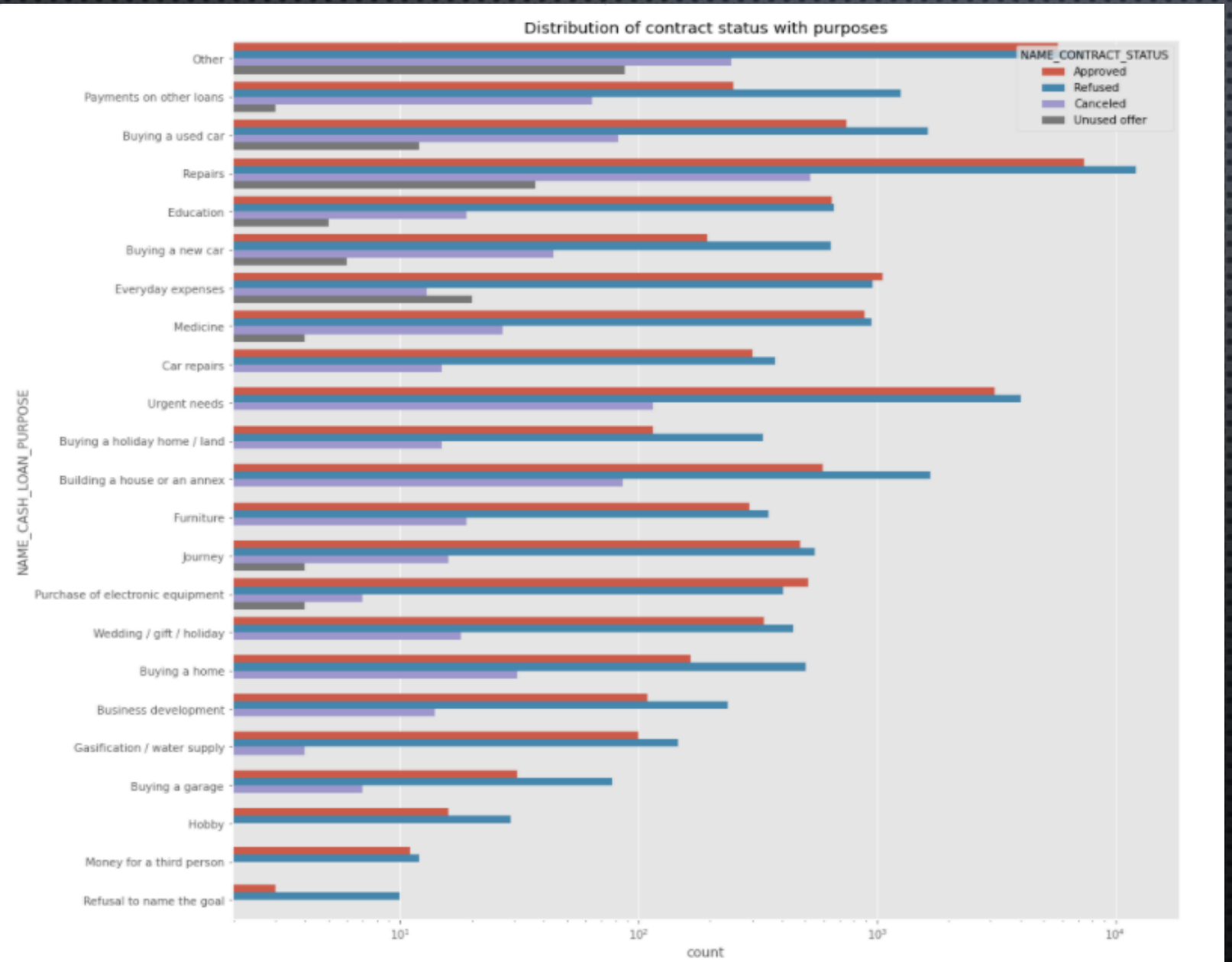
(B/W NAME_EDUCATION_TYPE & AMT_CREDI & NAME_FAMILY_STATUS FOR NON-DEFAULTER)



INFERENCES

- FOR EDUCATION 'HIGHER EDUCATION' TYPE THE INCOME AMOUNT IS MOSTLY EQUAL WITH FAMILY STATUS.
- HIGHER EDUCATION HAS MORE NUMBER OF OUTLIERS.
- ACADEMIC DEGREE TYPE HAVING LESS NUMBER OF OUTLIERS BUT THEIR INCOME AMOUNT IS LITTLE HIGHER THAT HIGHER EDUCATION.
- LOWER SECONDARY OF WIDOW FAMILY STATUS ARE HAVE LESS INCOME AMOUNT THAN OTHERS.

DISTRIBUTION OF CONTRACT STATUS WITH PURPOSE



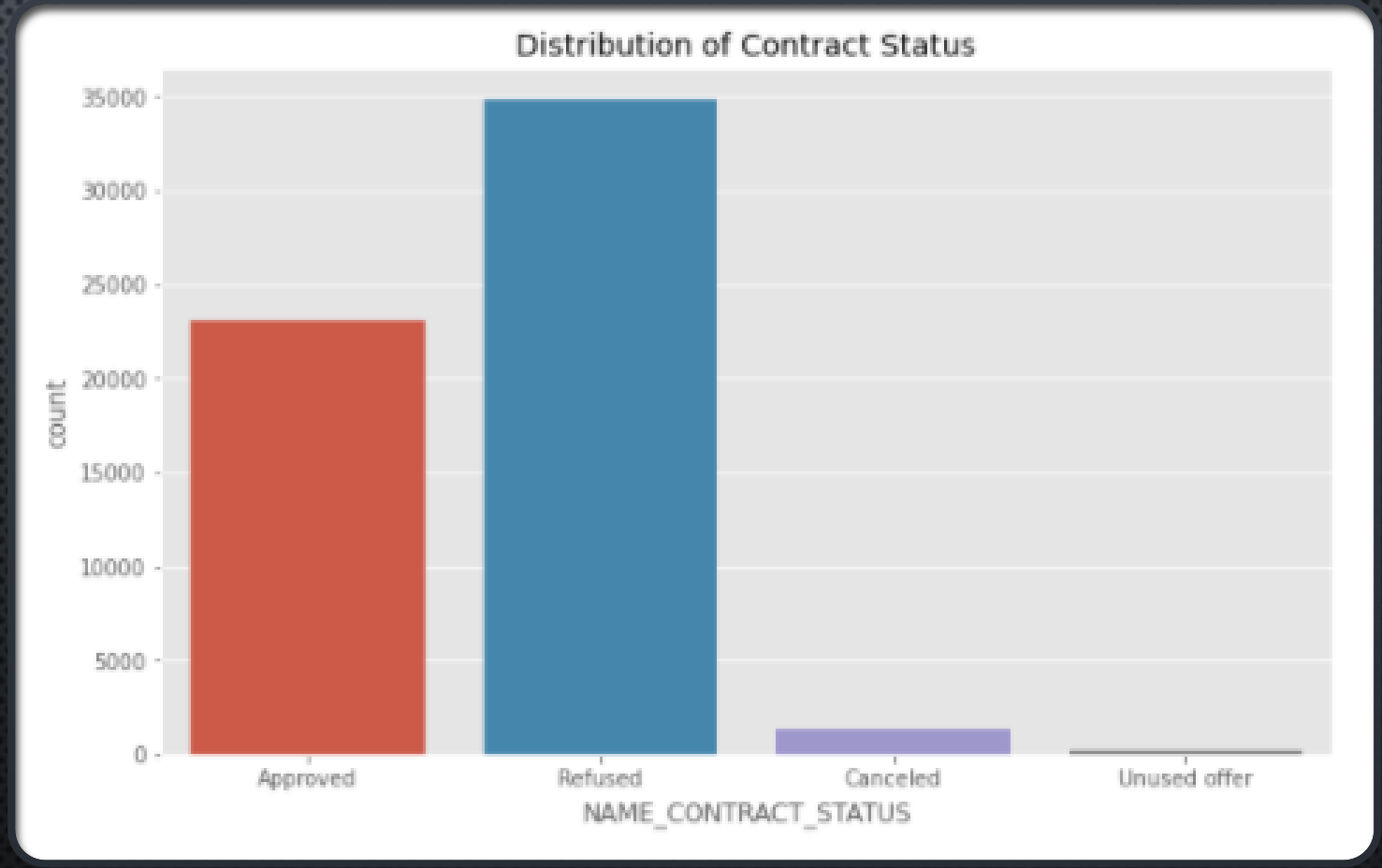
INFERENCES

- MOST OF THE REFUSED APPLICATIONS IS FROM REPAIRS.
- IN PURCHASE OF ELECTRONIC EQUIPMENT AND EVERYDAY EXPENSES, APPROVED APPLICATION STATUS IS MORE THAN REFUSED ONE.
- APPROVED AND REFUSED STATUS SEEMS TO BE ALMOST EQUAL IN EDUCATION.

DISTRIBUTION OF CONTRACT STATUS

INFERENCE

- FROM THE PLOT, REFUSED APPLICATION SEEMS TO BE MORE THAN OTHER STATUS OF APPLICATIONS.



AMT_APPLICATION VS AMT_INCOME_ORIGINAL

Inferences

- Loan request higher than 200k had a higher rejection rate.
- Loan rejection rate was much l ower if the income was higher.

CONCLUSION

1. Bank should focus more on student and businessman for successful payment.
2. Bank should focus less on working type candidates as they are failing to pay dues.
3. Bank should look to give more revolving loans as they are safer.
4. Bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment. Bank can focus mostly on housing type House/apartment, rented apartment and municipal apartment for successful payments.
5. Repairs section in loan purpose should be avoided by the company. Loans for buying land, new car, garage and new land should be approved more.
6. The company should have a closer look at the people who are changing their ID documents recently as they are more prone to default.

THANK YOU