CREDIT EXPLORATORY DATA ANALYSIS ASSIGNMENT

By- DOLLY PANDEY

TITLE: CREDIT EDA ASSIGNMENT
SUBTITLE: INSIGHTS AND RECOMMENDATIONS

Objective:

- To analyze credit data using EDA techniques.
- To identify key factors influencing payment difficulties.
- To identify variables which are strong indicator of default.
- Key features – Age, Income ,Employment status ,Payment difficulties, Credit amount etc…

DATA CLEANING

- APPLICATION DATA
- Handling missing values
- Columns with over 30% missing values were dropped
- Mean value imputation
- Median imputation for numerical features
- Mode imputation for categorical features
- Removing unwanted data
- Rows with over 50% missing values were dropped

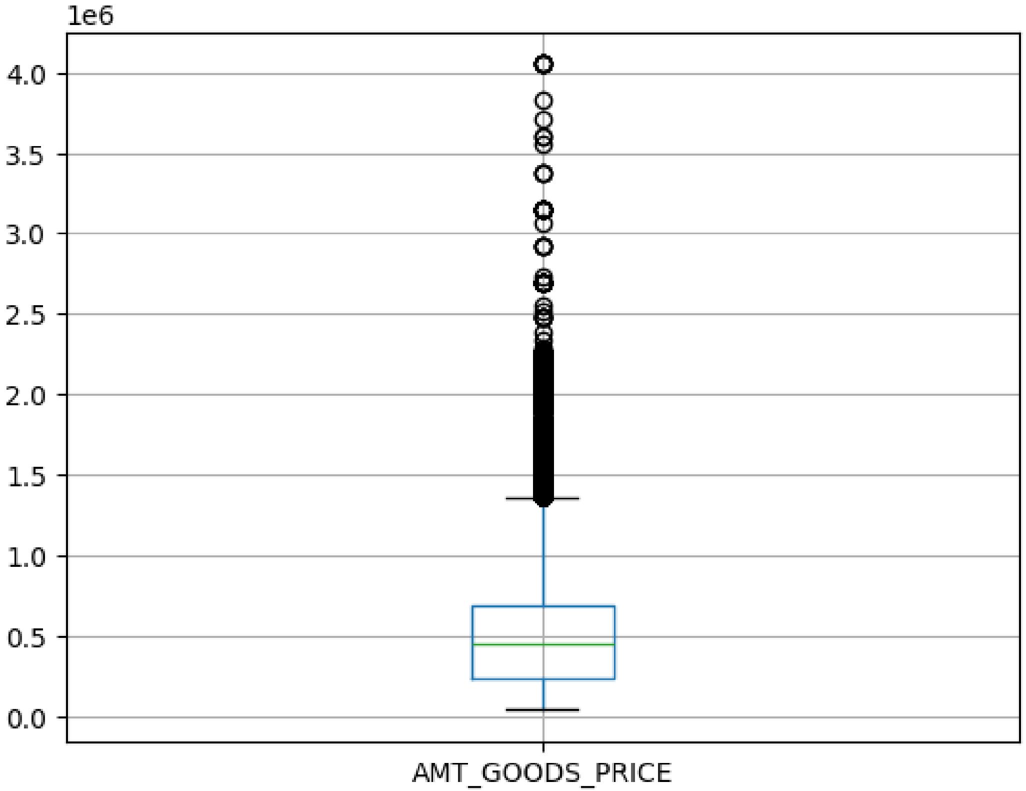
- Correcting data types
- Converted DAYS_BIRTH column to AGE

WE HAVE ZERO NULL VALUES AFTER DROPPING AND IMPUTATION PROCESS

E.g-AMT_GOODS_PRICE

(BOXPLOT)

- Standard deviation of AMT_GOODS_PRICE IS VERY HIGH
- Outliers are present in huge amount it needs to be treated



CATAGORICAL UNIVARIATE ANALYSIS ON THE BASIS OF INCOME

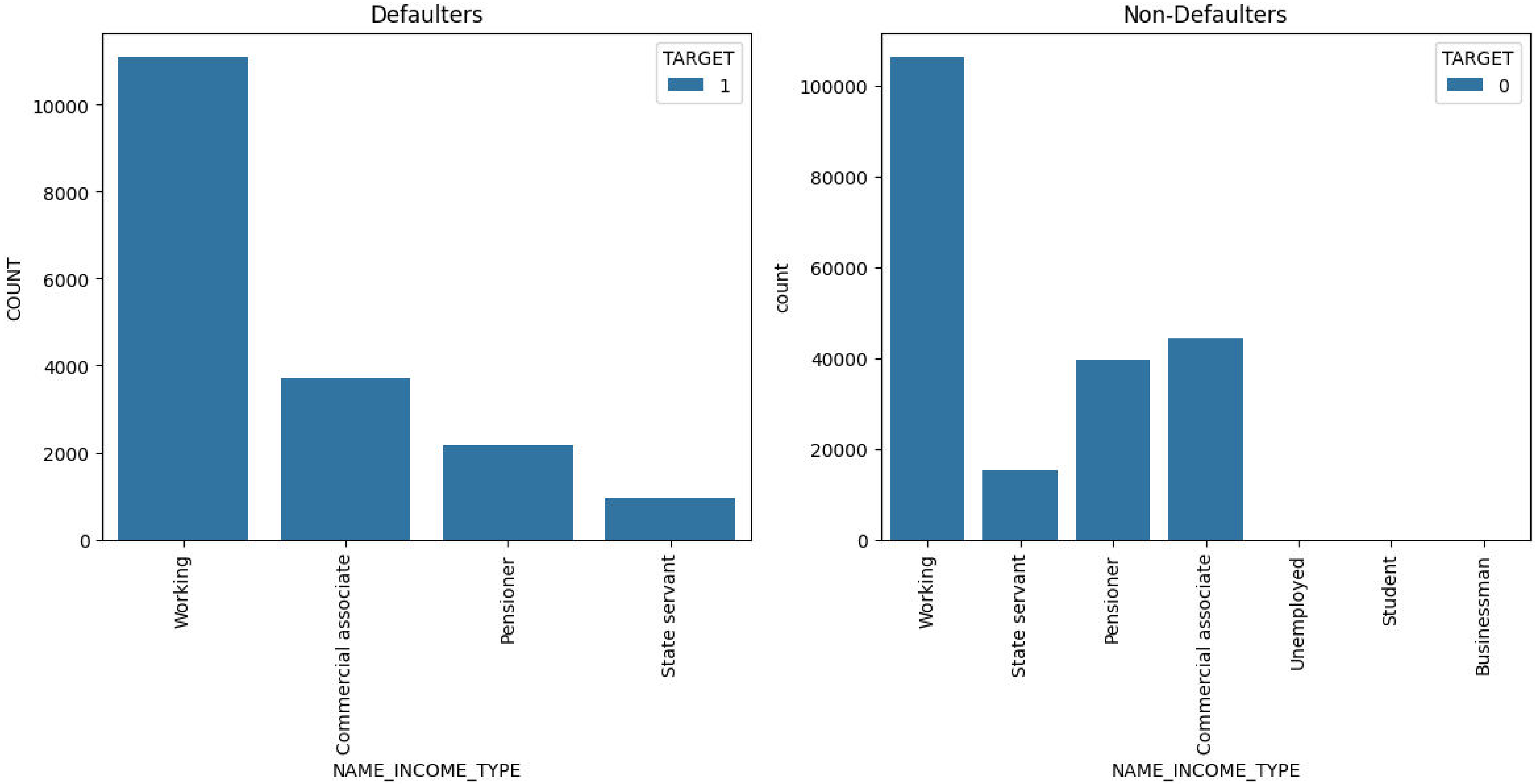
Target 0

- Points that we observe in graph:
- Working clients are mostly Non-defaulted in this case.
- Bussinessman's has no records of defaulter.
- Commercial associates and pensioner has almost equal ratio there is a very minor difference between them.

Target 1

- Points that we observe in graph:
- Working clients are more defaulted then other cases.
- State servents are less defaulted then others.
- The difference between commercial associate and pensioner in this region slightly increased

UNIVARIATE ANALYSIS OF TARGET 0 AND 1



UNIVARIATE ANALYSIS ON THE BASIS OF EDUCATION

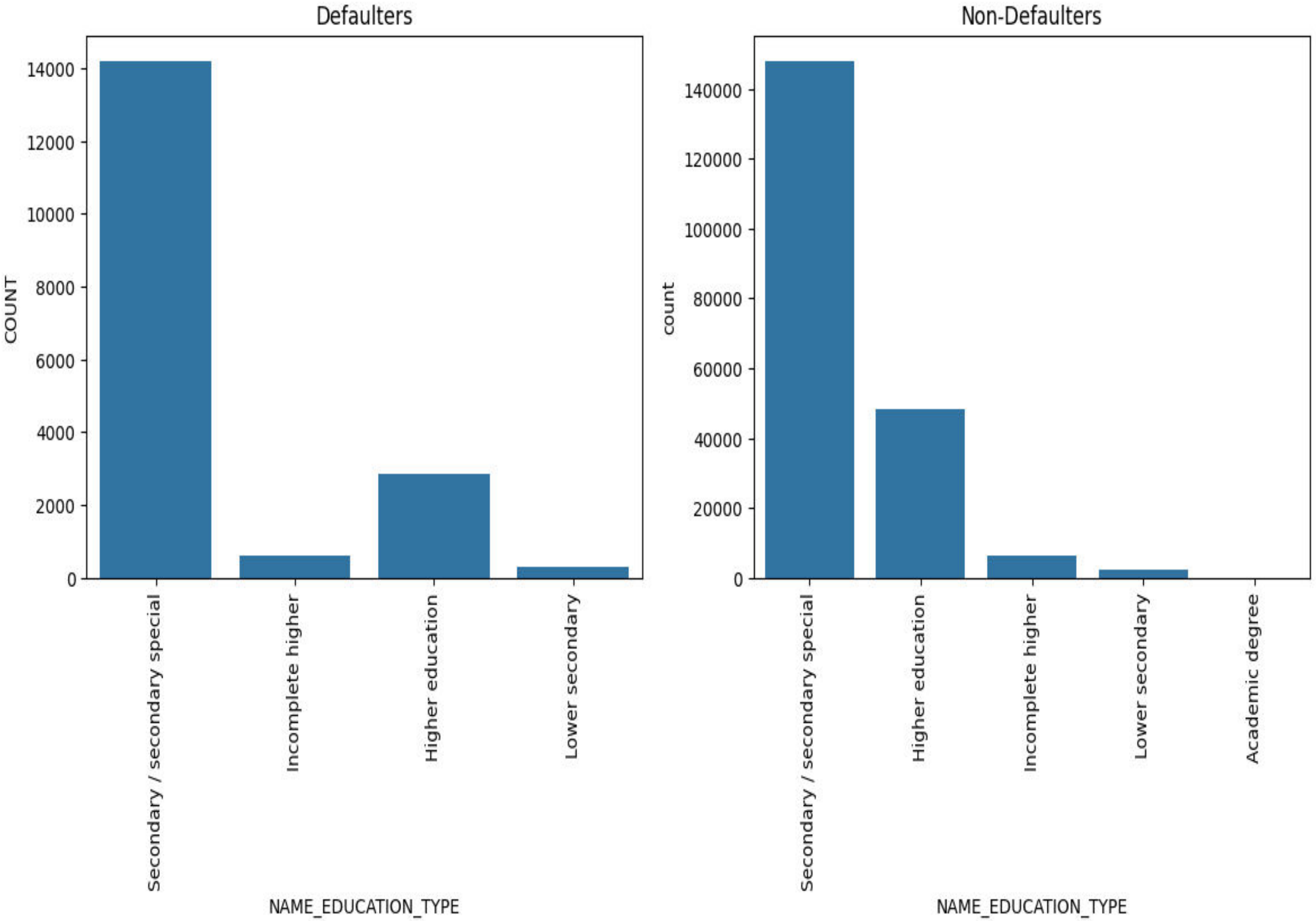
❖**Defaulters-**

- Secondary/Secondary special are mostly defaulted in

Target 1 case

- Non-defaulters-
- Secondary/Secondary special are mostly non-defaulted in

Target 0 case



SEGMENTED UNIVARIATE ANALYSIS FOR ORDERED CATEGORICAL VARIABLES

❖Credit amount groups

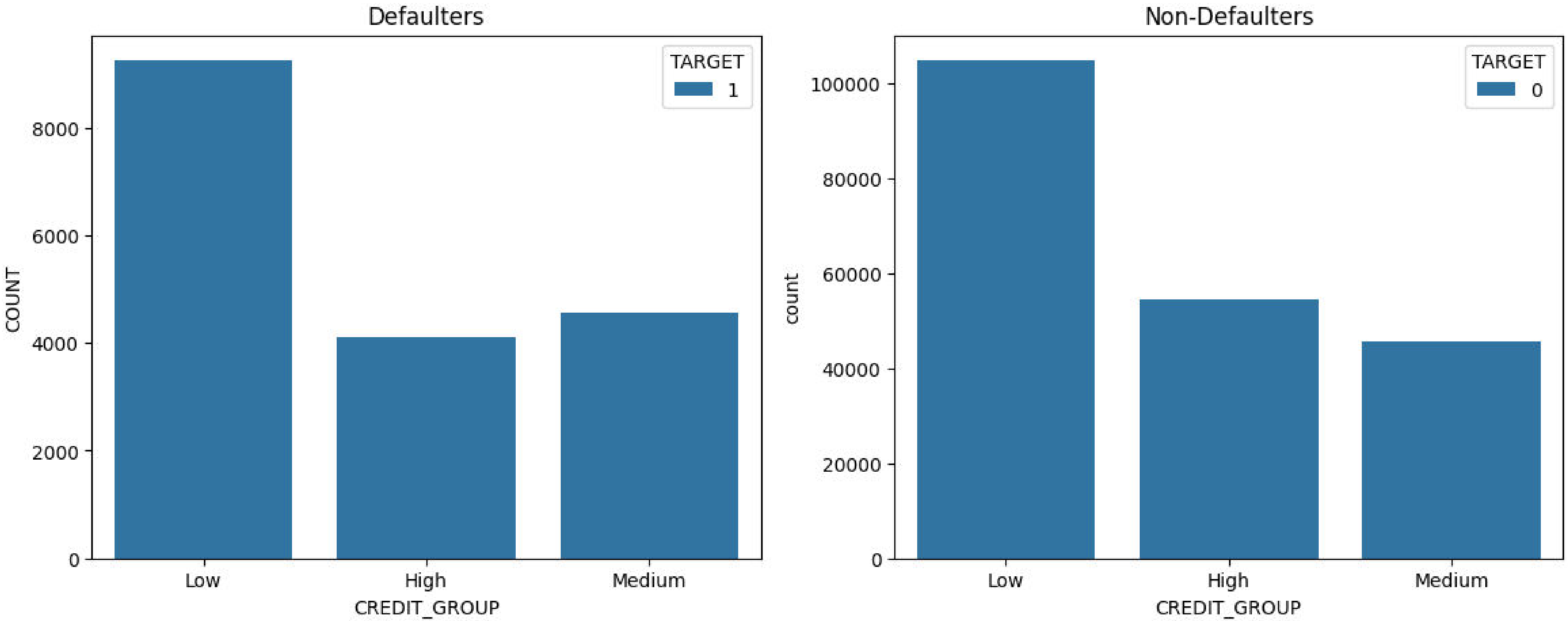
➢Defaulters-

- Low credited amounts group are more defaulted in Target 1 case.

➢Non – defaulters-

- As expected, Low credit amount groups are more non-defaulted in Target 0 case.

CREDIT AMOUNT GROUP



BIVARIATE ANALYSIS ON CONTINUOUS VARIABLE

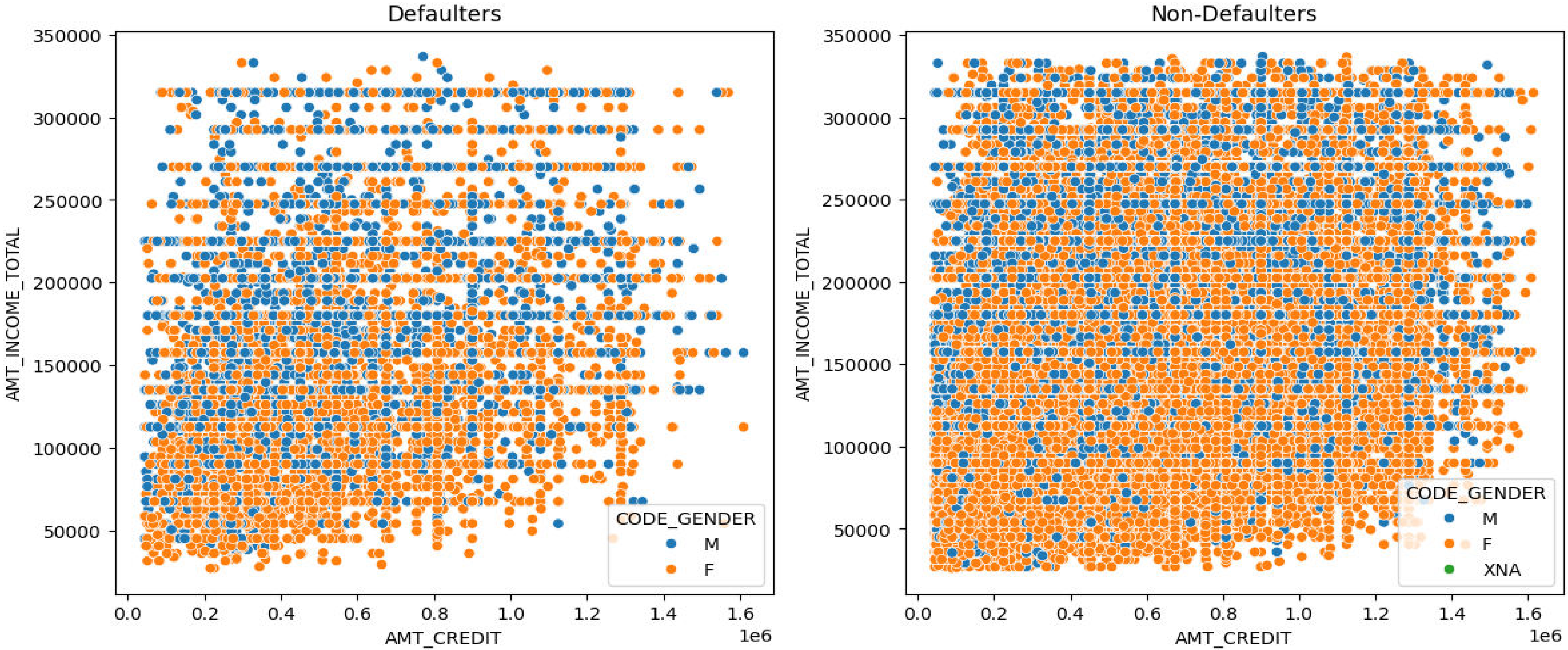
- **Credit amount of the loan on the basis of client income for both male and female**

- When there is negative value there is an inverse relation between variables as for e.g if the price of product increased the demand for it decrease.
- Here we can slightly figure out that the values are more concentrated on the lower income and lower credit of the loan. That means as the income is increased, the amount of loan is also increased.
- This is same for both genders

- Here we can observe-

- Males are more defaulted
- Females are much non-defaulted

CREDIT AMOUNT OF THE LOAN ON THE BASIS OF CLIENT INCOME FOR BOTH MALE AND FEMALE¶

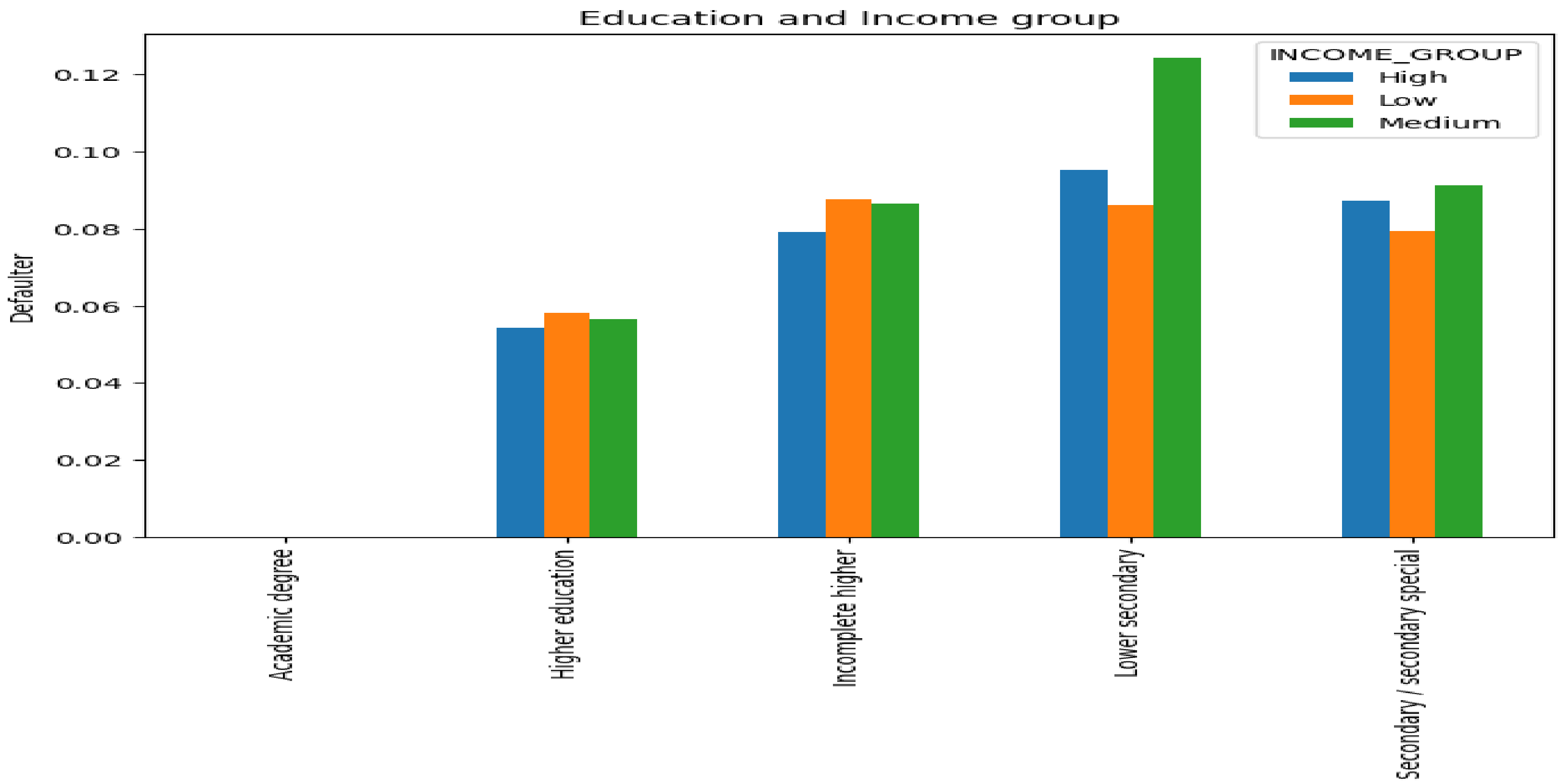


MULTIVARIATE ANALYSIS

❖ **Education and Income group-**

- Clients with lower secondary education and medium income group are mostly defaulted.
- Client with academic degree has no records of default.
- Clients with higher education has low records of default.

EDUCATION AND INCOME GROUP

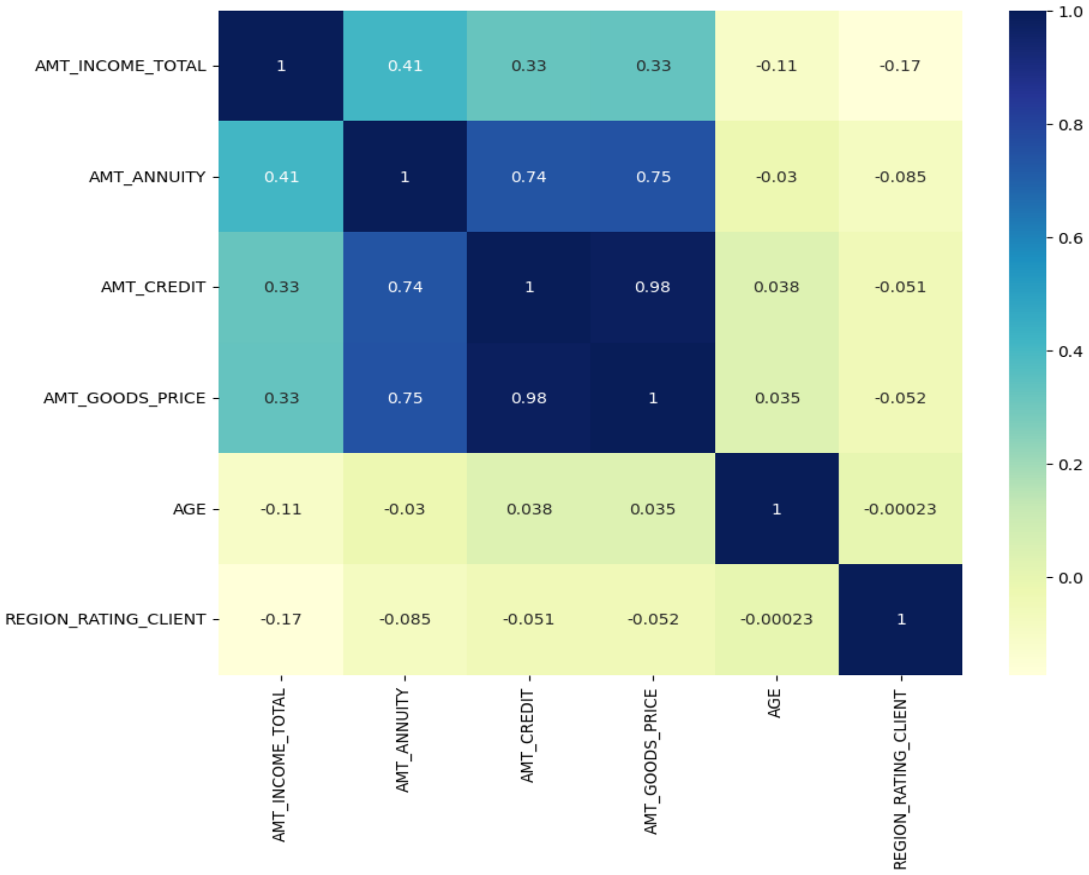


CORRELATION FOR TARGET 0

❖Correlates for Non-defaulter

➢Highly correlated column for non-defaulter-

1. AMT_CREDIT and AMT_ANNUITY (0.74)
2. AMT_CREDIT and AMT_GOODS_PRICE (0.98)
3. AMT_ANNUITY and AMT_GOODS_PRICE (0.75)



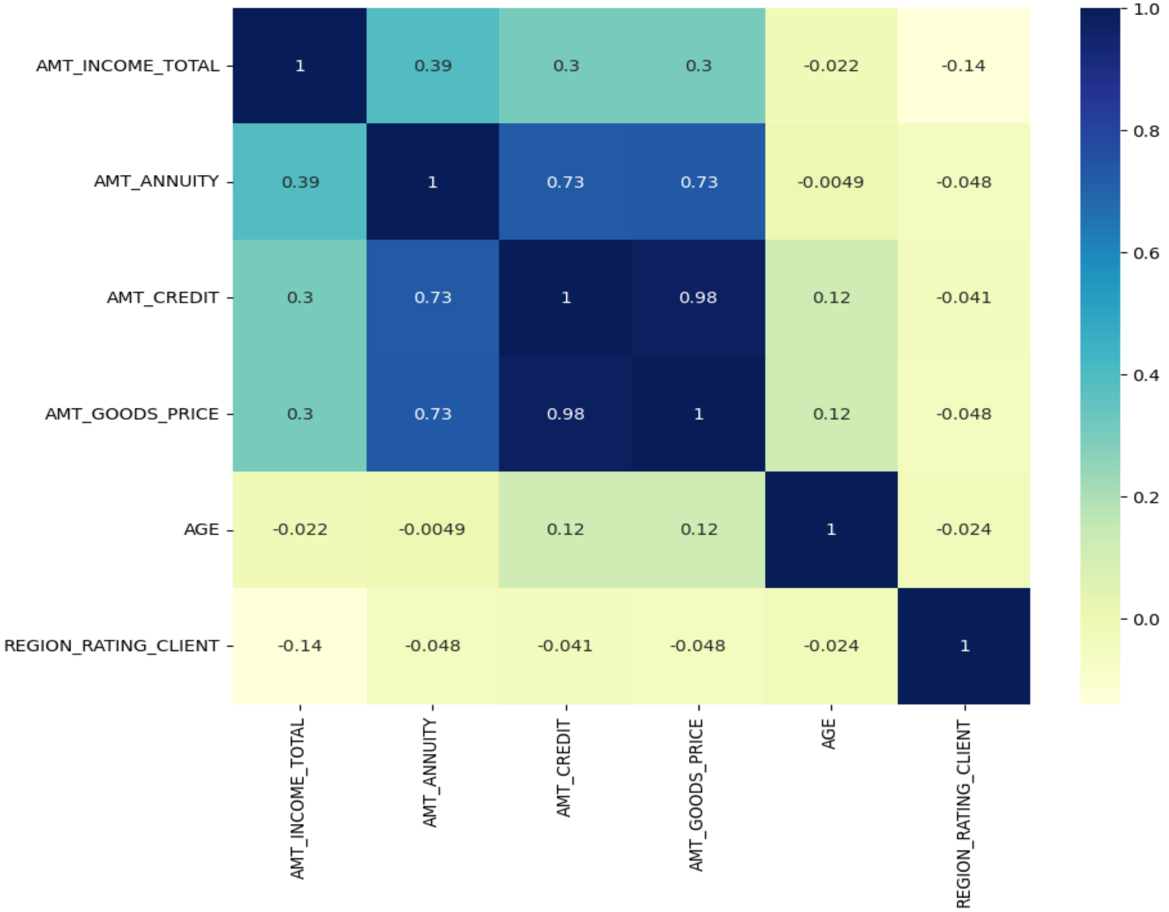
❑**Conclusion** - We can see that for both defaulters and non defaulters the same pairs of columns are highly corelated.

CORRELATION FOR TARGET 1

- **Correlates for defaulters**

- **Highly correlated columns for defaulters**

1. AMT_CREDIT and AMT_ANNUITY (0.73)
2. AMT_CREDIT and AMT_GOODS_PRICE (0.98)
3. AMT_ANNUITY and AMT_GOODS_PRICE (0.73)



❑ **Conclusion** - We can see that for both defaulters a highly corelated

PREVIOUS APPLICATION

❑DATA CLEANING

❖Handling missing values-

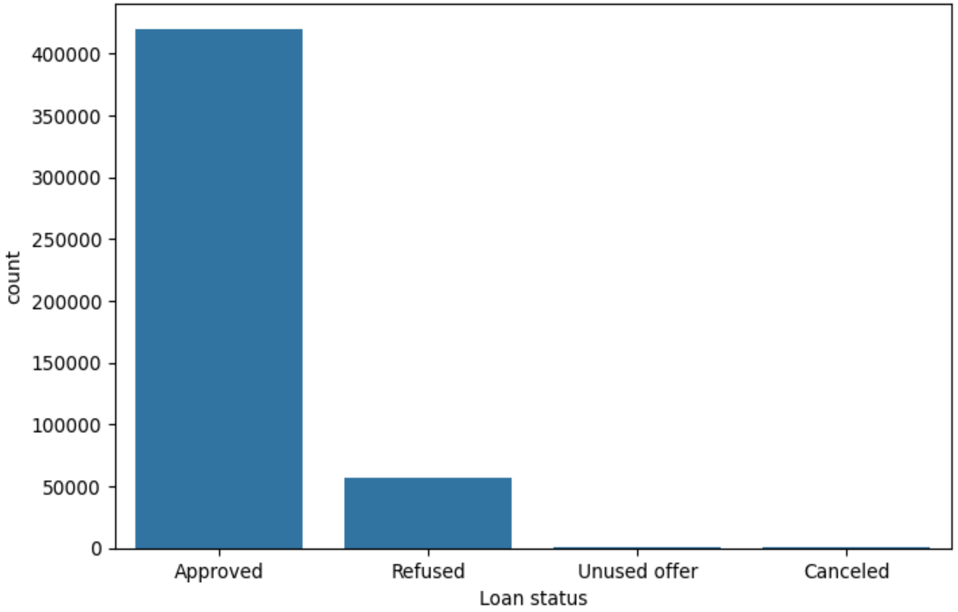
- More than 30% null values were dropped.
- Unwanted columns were dropped.
- Imputation with mean median mode
- Removing Outliers
- Converting DAYS_COLUMN into MONTH_DECISION for readability and analysis.
- Dropping DAYS_DECISION column.

UNIVARIATE ANALYSIS AFTER MERGING DATA

Previous loan status

- **Analysis-**

- Here is a huge difference between Approved loan and Refused loans.
- The graph of Approved loan is slightly higher than refused loans.



BIVARIATE ANALYSIS AFTER MERGING DATA

❖Correlation of merged dataset

Highly corelate columns

1.AMT_APPLICATION and AMT_CREDIT

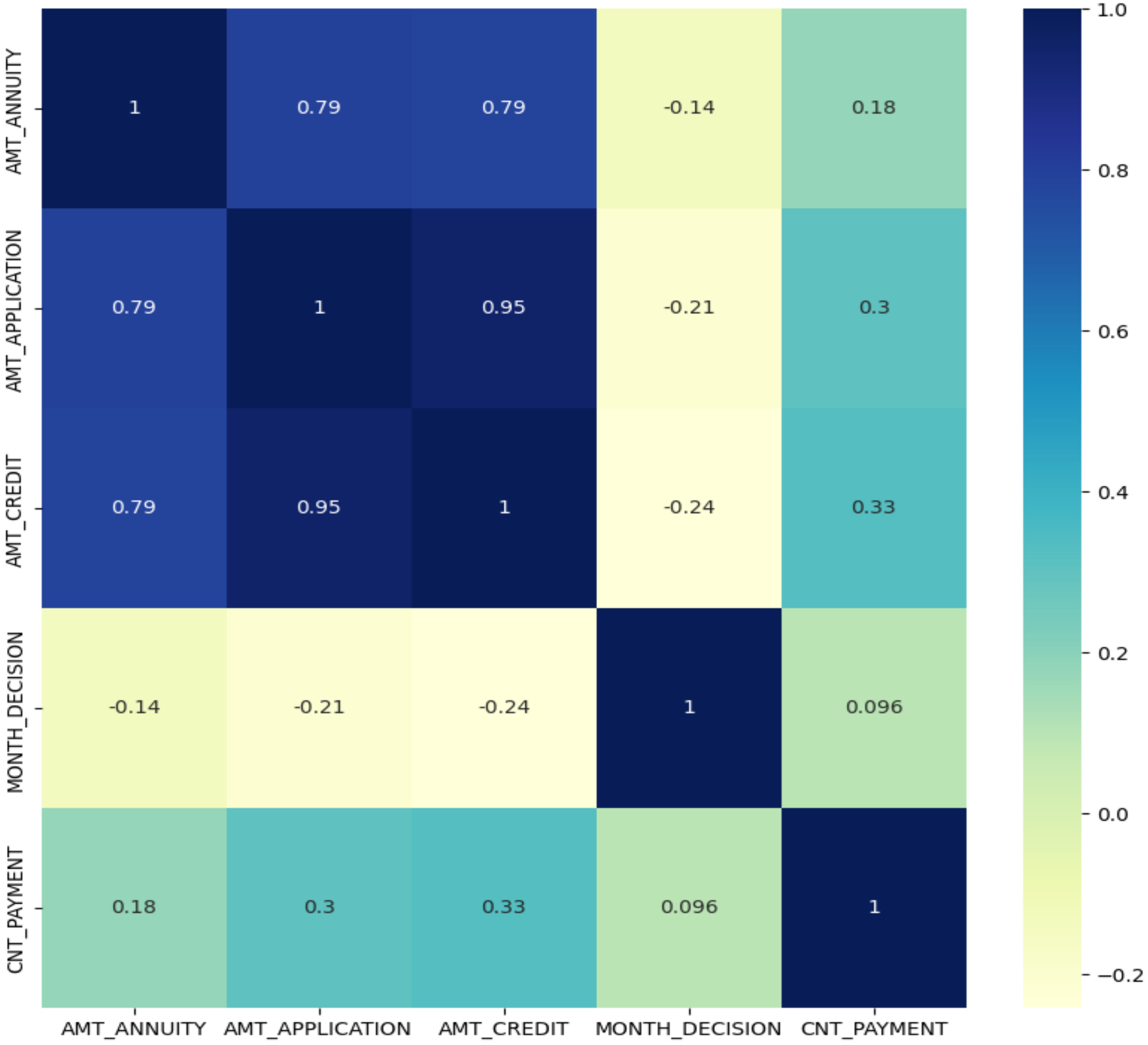
2.AMT_APPLICATION and AMT_ANNUITY

3.AMT_CREDIT and AMT_ANNUITY

Moderately corelated columns

1.AMT_APPLICATION and CNT_PAYMENT

2.AMT_CREDIT and CNT_PAYMENT

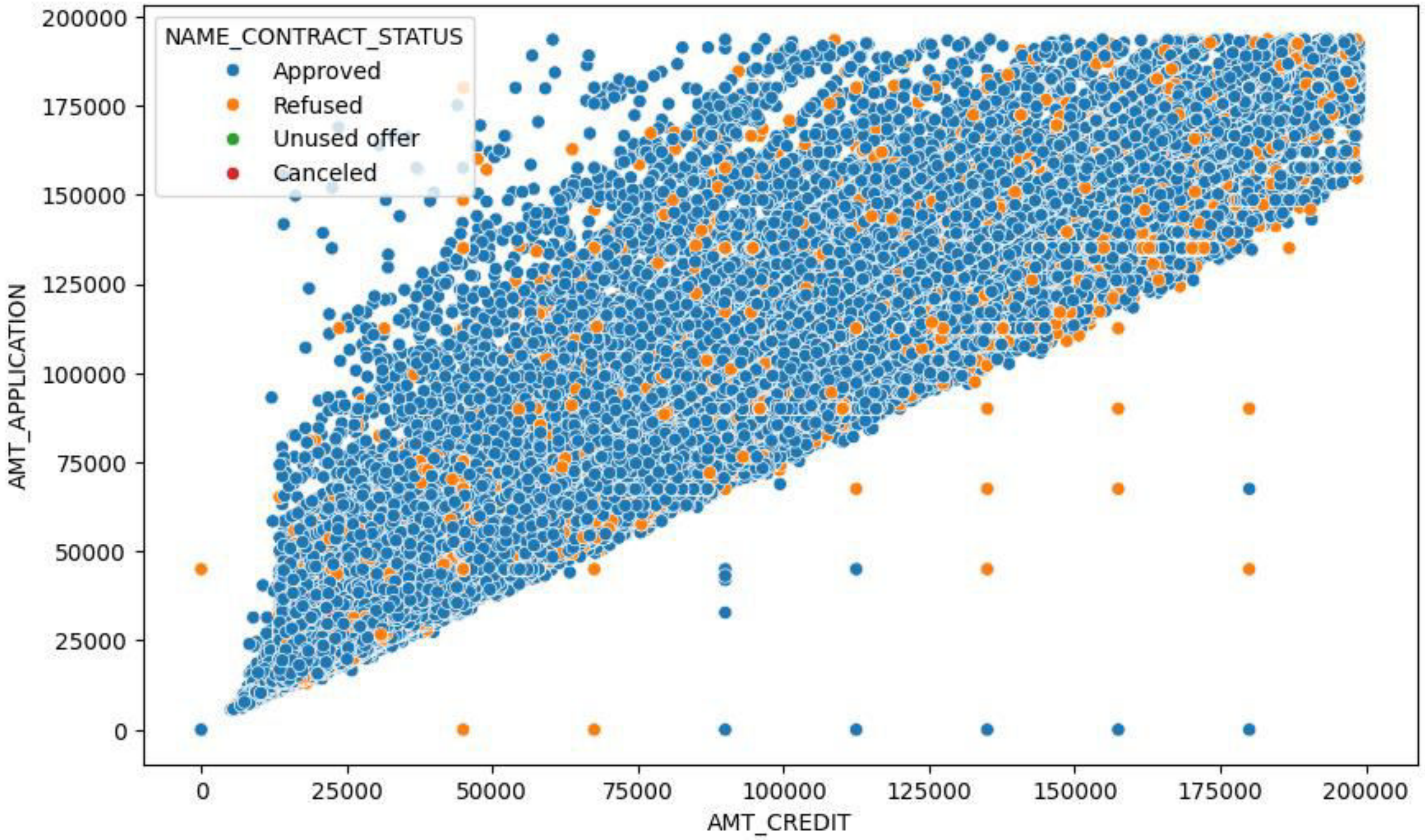


	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	MONTH_DECISION	CNT_PAYMENT
AMT_ANNUITY	1	0.79	0.79	-0.14	0.18
AMT_APPLICATION	0.79	1	0.95	-0.21	0.3
AMT_CREDIT	0.79	0.95	1	-0.24	0.33
MONTH_DECISION	-0.14	-0.21	-0.24	1	0.096
CNT_PAYMENT	0.18	0.3	0.33	0.096	1

BIVARIATE ANALYSIS ON CONTINUOUS VARIABLE

The credited amount become more concentrated with respect to application amount

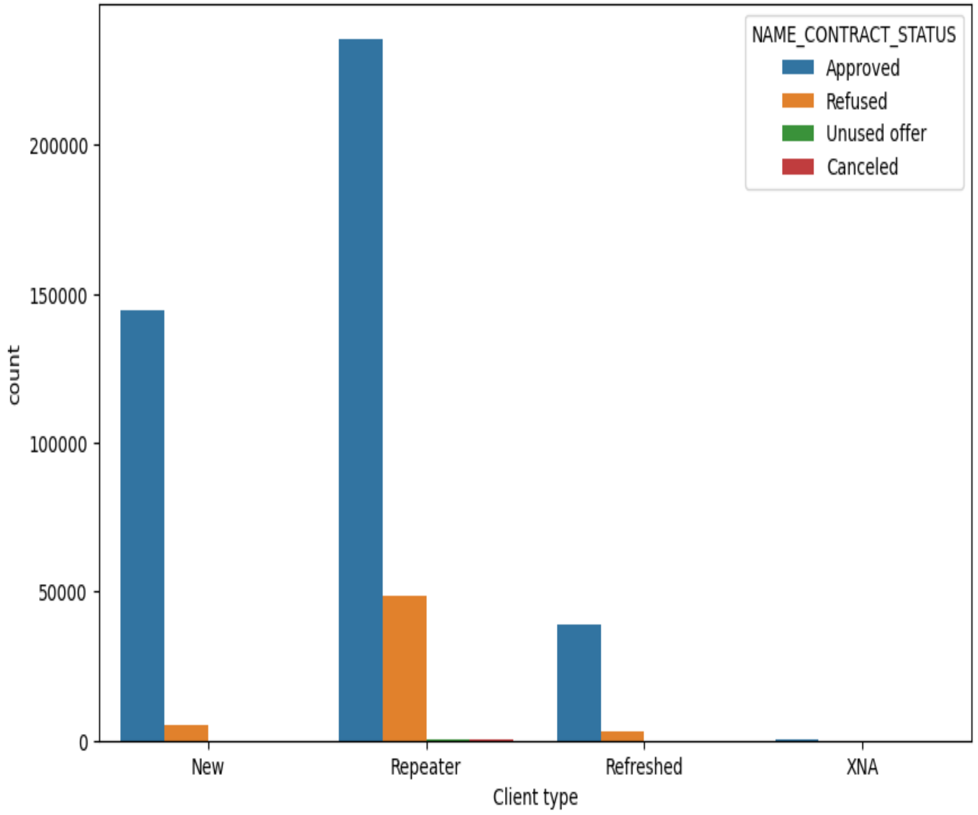
The ratio of approval is very high in scatterplot.



MULTIVARIATE ANALYSIS ON MERGED DATA

- Status and client type-

- Here we can clearly see that the repeated clients has high chance of approval with respect to others.
- Refreshed clients has low chance of approval and they May get refused.
- New clients has higher chance of approval as compared To refreshed clients.

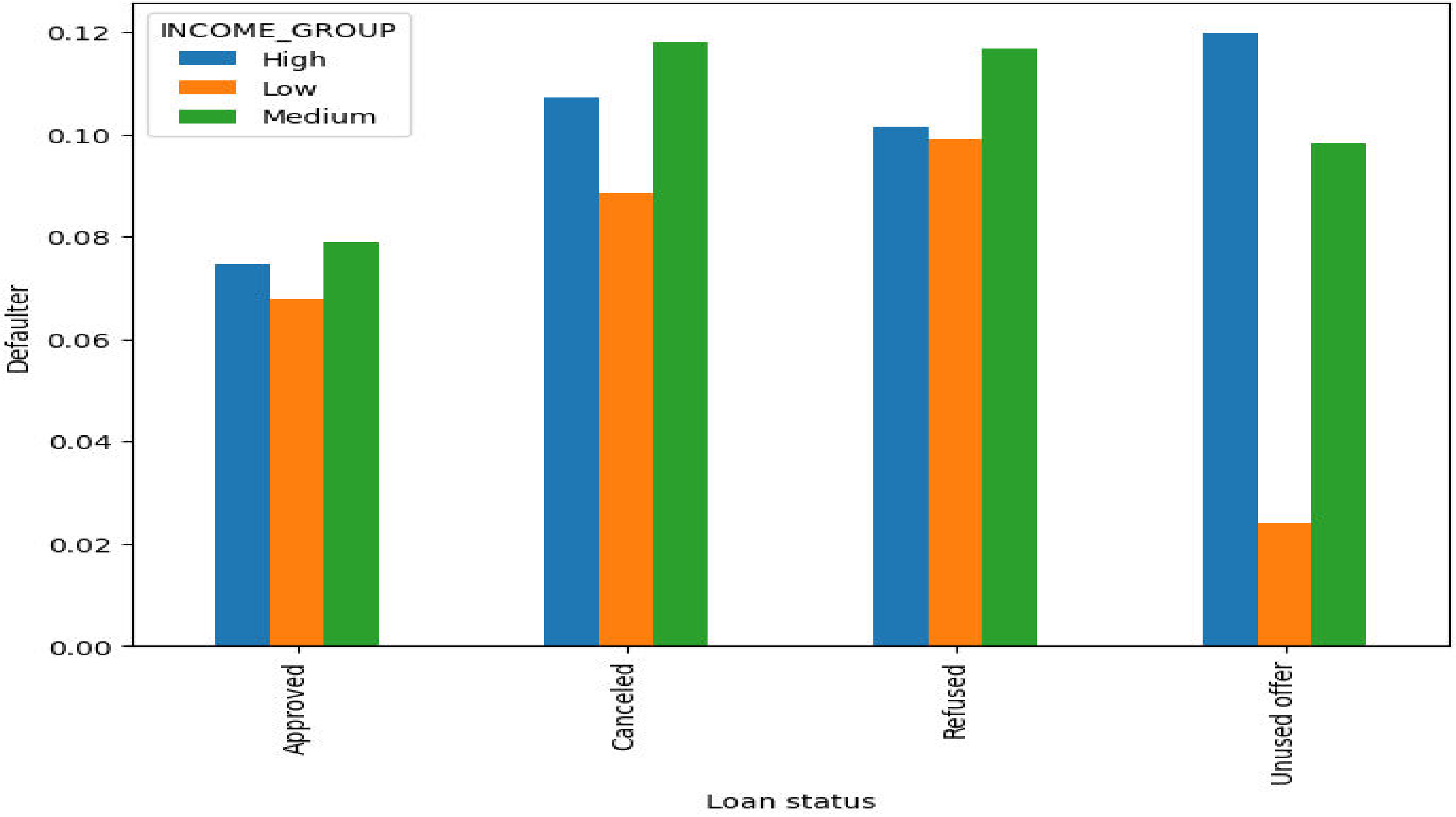


CURRENT LOAN DEFAULTER STATUS WITH RESPECT TO PREVIOUS LOAN APPLICATION STATUS AND INCOME GROUP

Analysis-

- The clients with High and medium income group and cancelled loan application status are much defaulted.
- Refused ones also has almost similar ratio.
- For unused offer high income group are more defaulted and low income group are least defaulted.
- For other application status more or less every income group is equally defaulted.

MULTIVARIATE ANALYSIS OF MERGED DATA



CONCLUSION

- Bank should focus less on working clients as they have most number of unsuccsessful payments
- They are mostly defaulted
- Bank should focus mostly on repeated customers as they have high chance of approval.
- Bank should also focus on clients with higher education status as they have low records of being defaulted.
- Clients with higher income level has low chance of defaults as they are able to pay loan installment on time so bank should also focus on them.

THANKYOU