EDA ASSIGNMENT PRESENTATION

BY - HARISH PRATAP

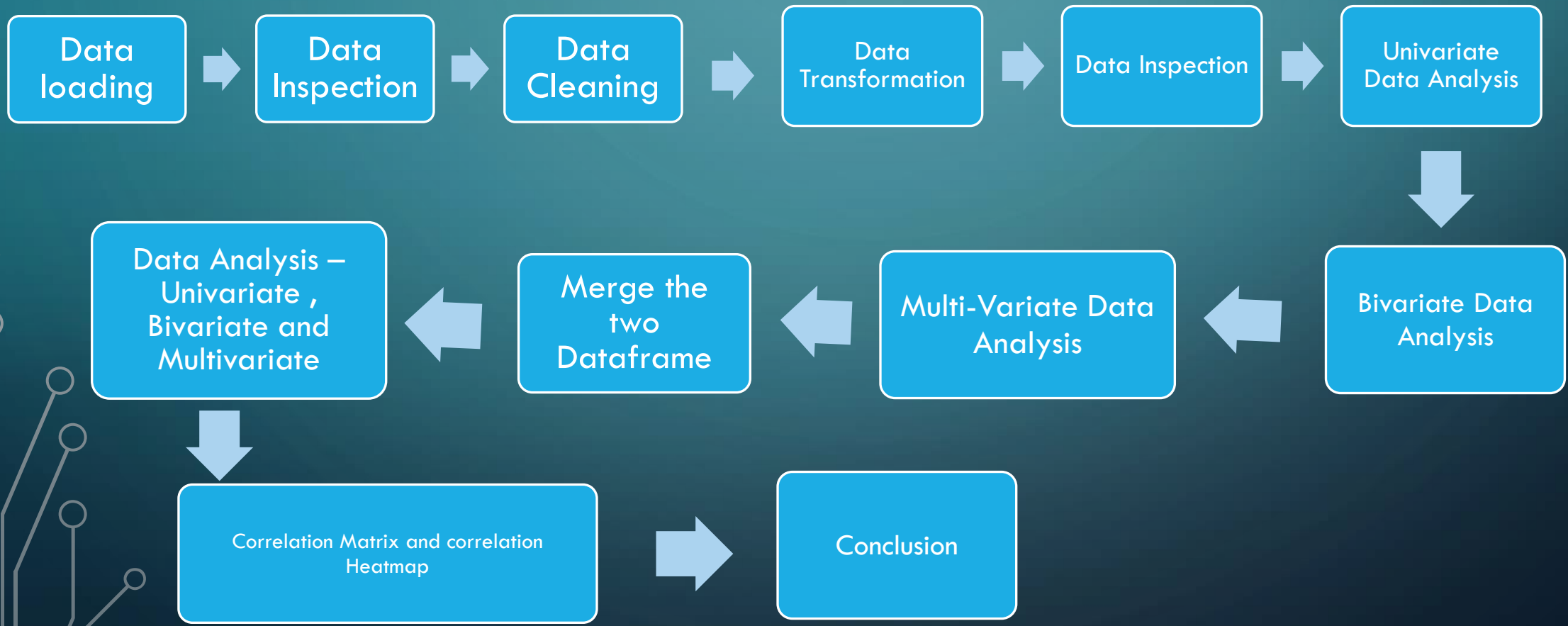
PROBLEM STATEMENT

- To find driving factors behind loan default, i.e. the variables which are strong indicators of default. The bank can utilize this knowledge for its portfolio and risk assessment.
- To identify patterns that indicate if a client has difficulty paying their installments 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.

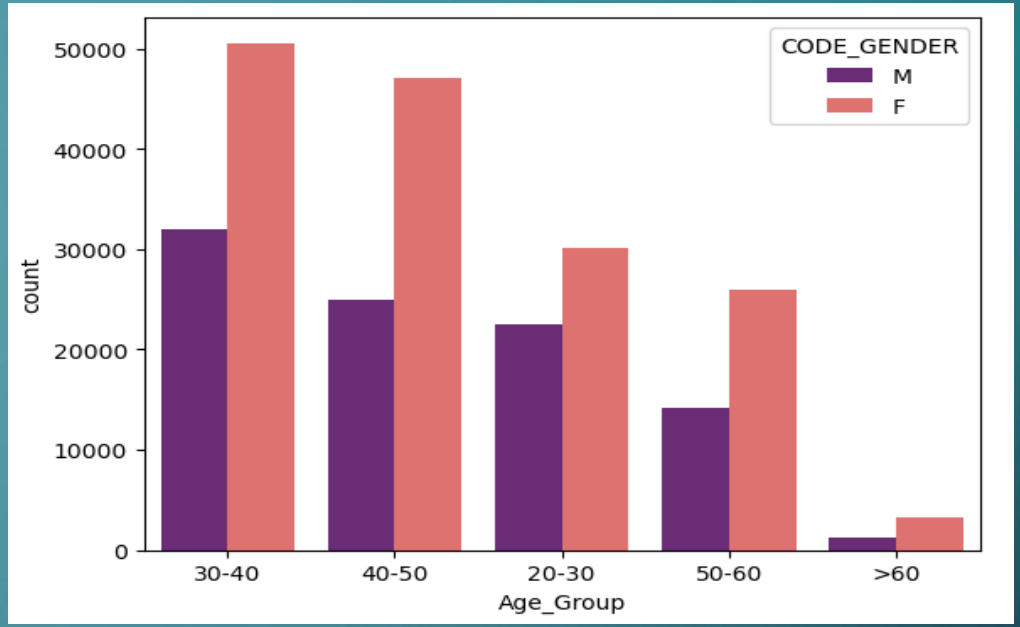
ASSUMPTIONS TAKEN DURING DATA ANALYSIS

- We have dropped columns that have missing percentages greater than 40% in the application dataset and above 30% in the previous _application dataset(We are doing this to avoid incorrect calculation and analysis or inaccurate prediction because of these missing values.)
- Variables with less than 30% missing values are been filled by appropriate values of mean, median, and mode based on the outliers present, variable type(categorial or numerical), and data distribution of the variable.
- We have removed some variables which are irrelevant to our data analysis and which will not make any impact on our findings.

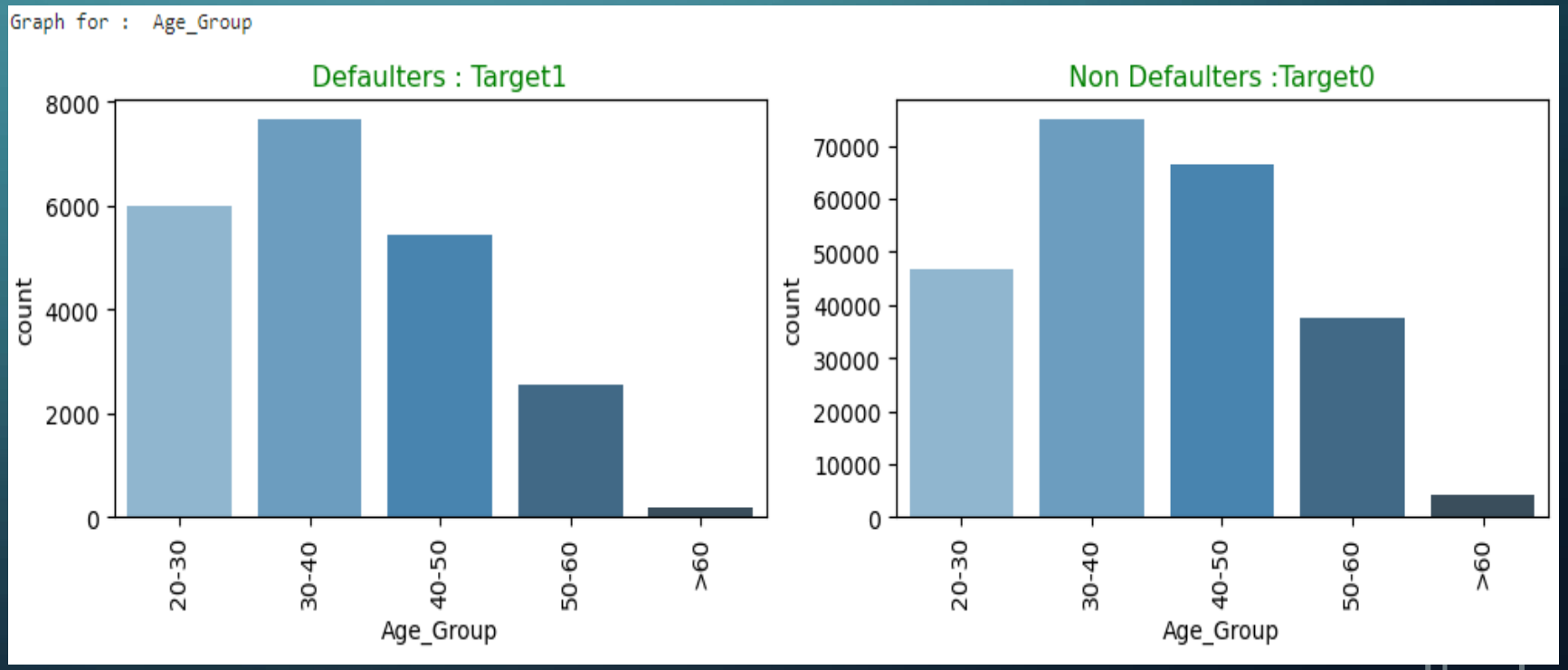
EDA APPROACH AND METHODOLOGY



DISTRIBUTION OF AGE GROUP

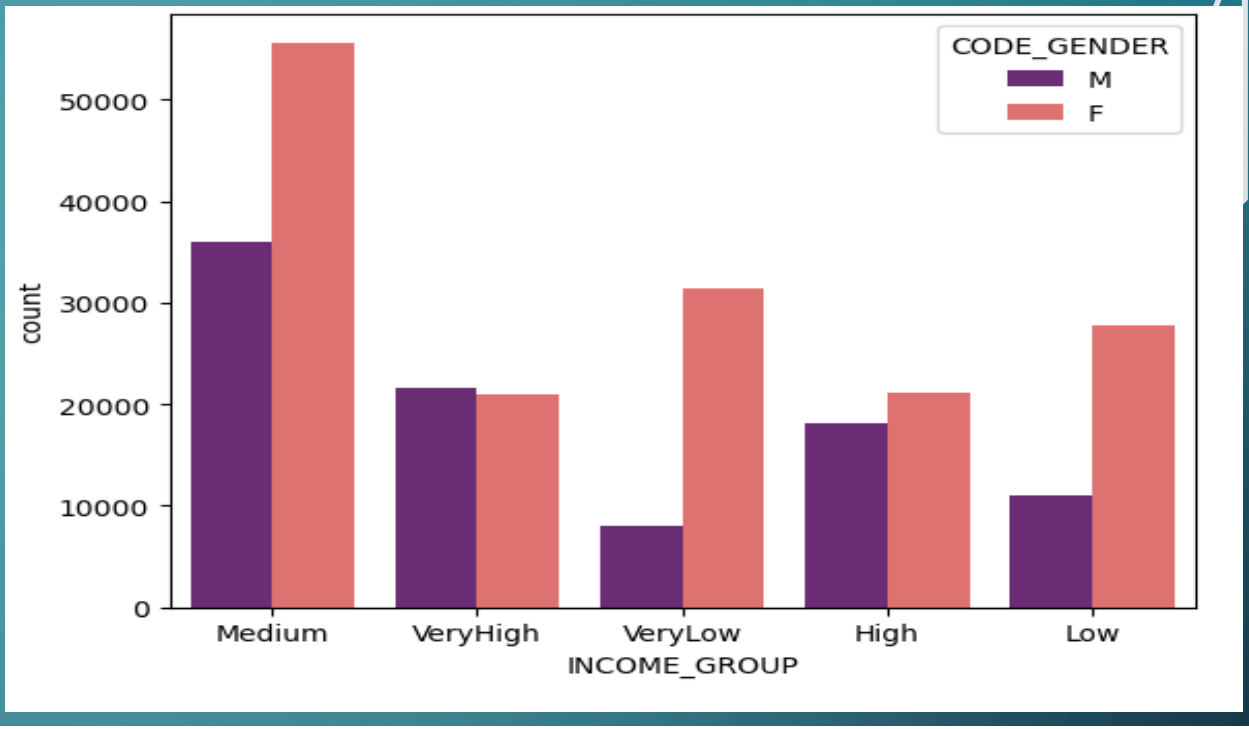


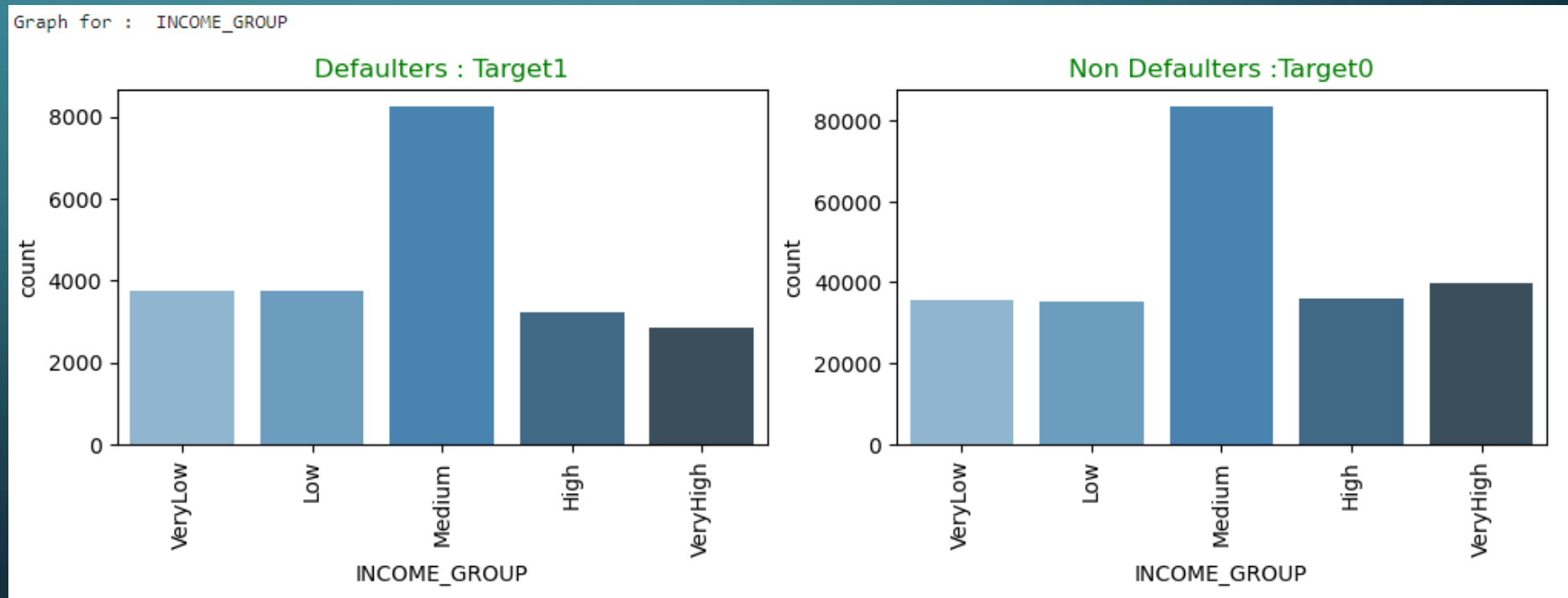
- Females have higher number of loan applications among all age categories
- Age Group - Highest applicants are from age group 30-40.
- It seems Age group 20-30 is having payment difficulties
- Age group 30-40,40-50 is less likely to default on loan repayment - >60 is very less to be defaulters so it can be a good target audience



DISTRIBUTION BY INCOME GROUP

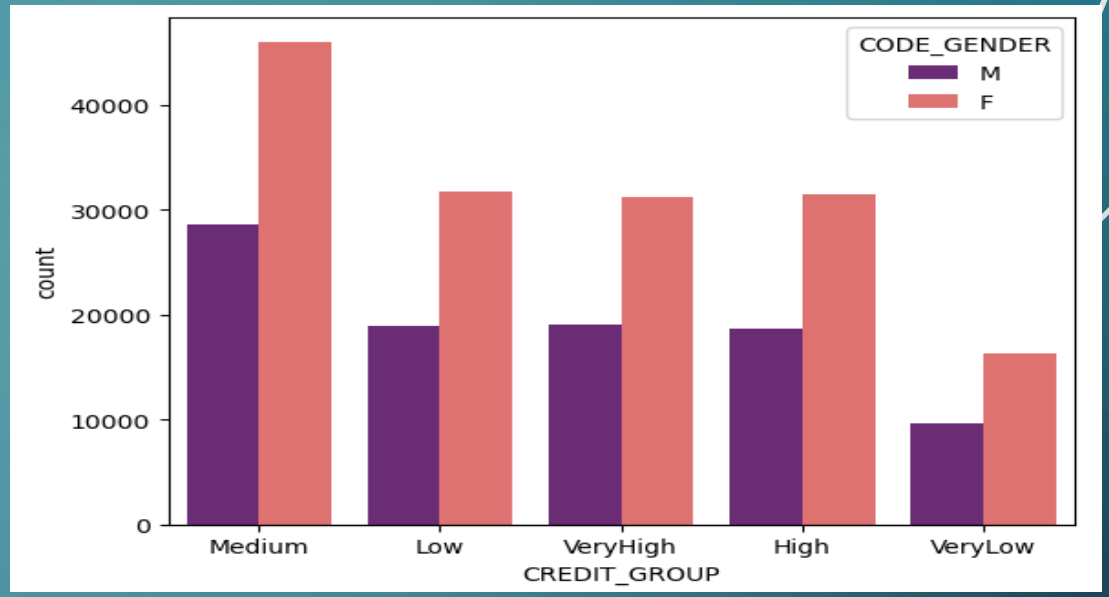
- Medium-income applicants have a higher number of loan applicants and are very high among defaulters.
- Females have higher income compared to Males except Very high-income group except for the very high-income group.
- Income Group - People with very high incomes are very likely to be defaulters.

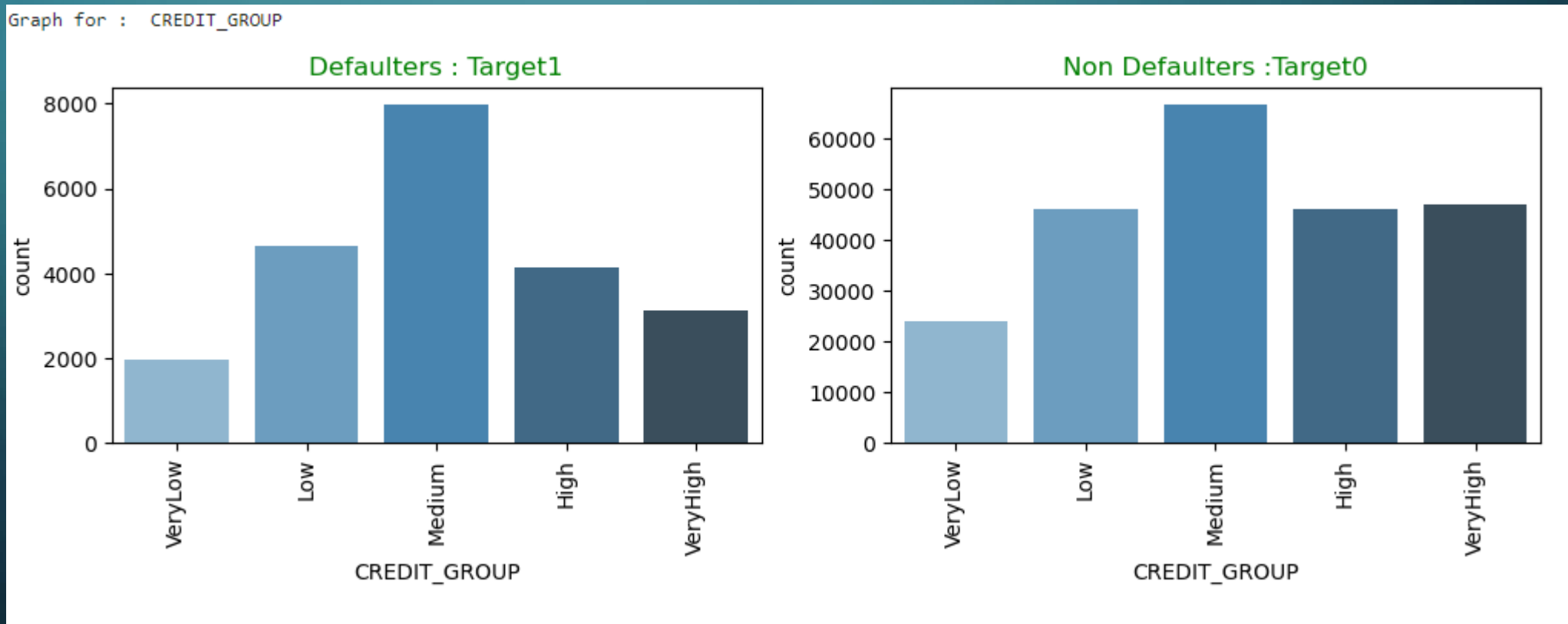


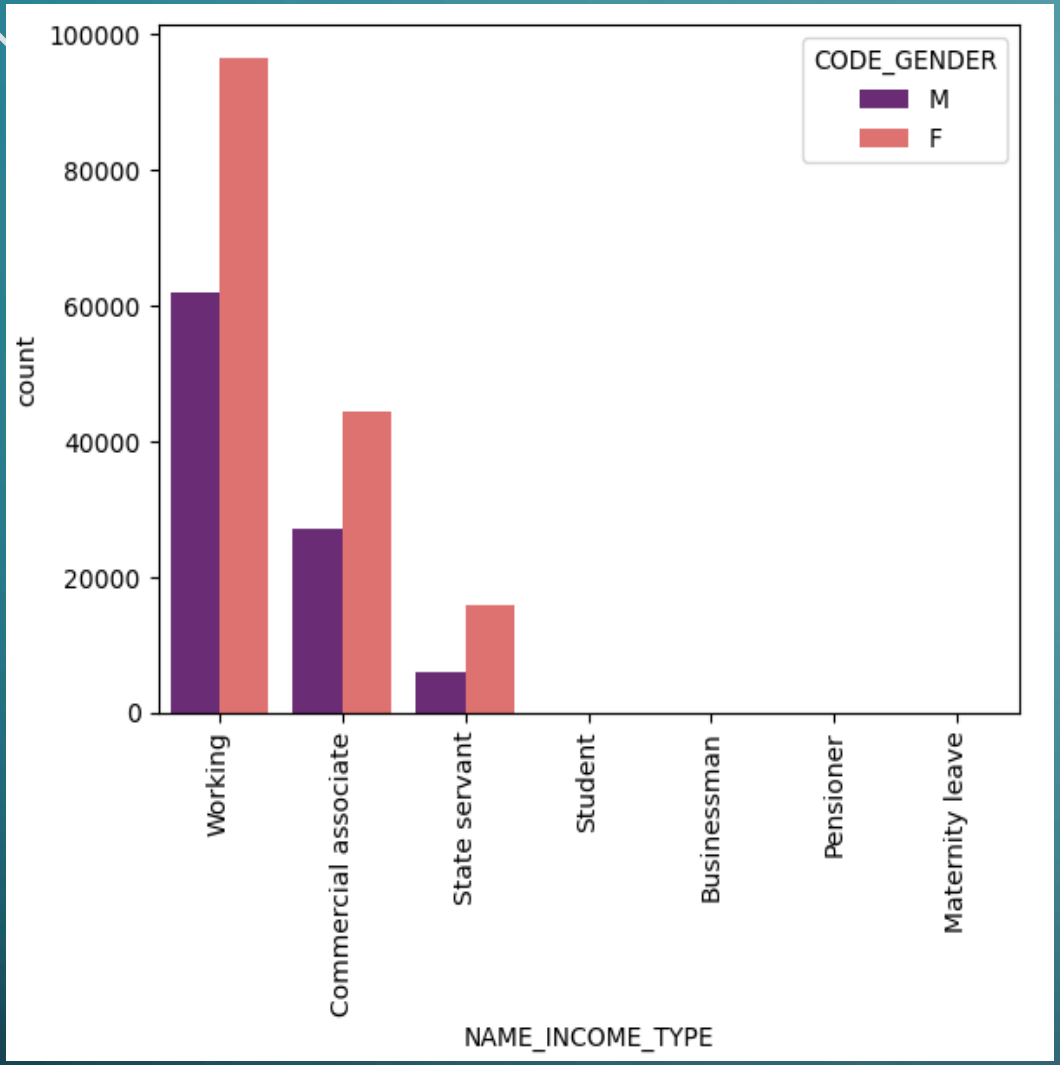


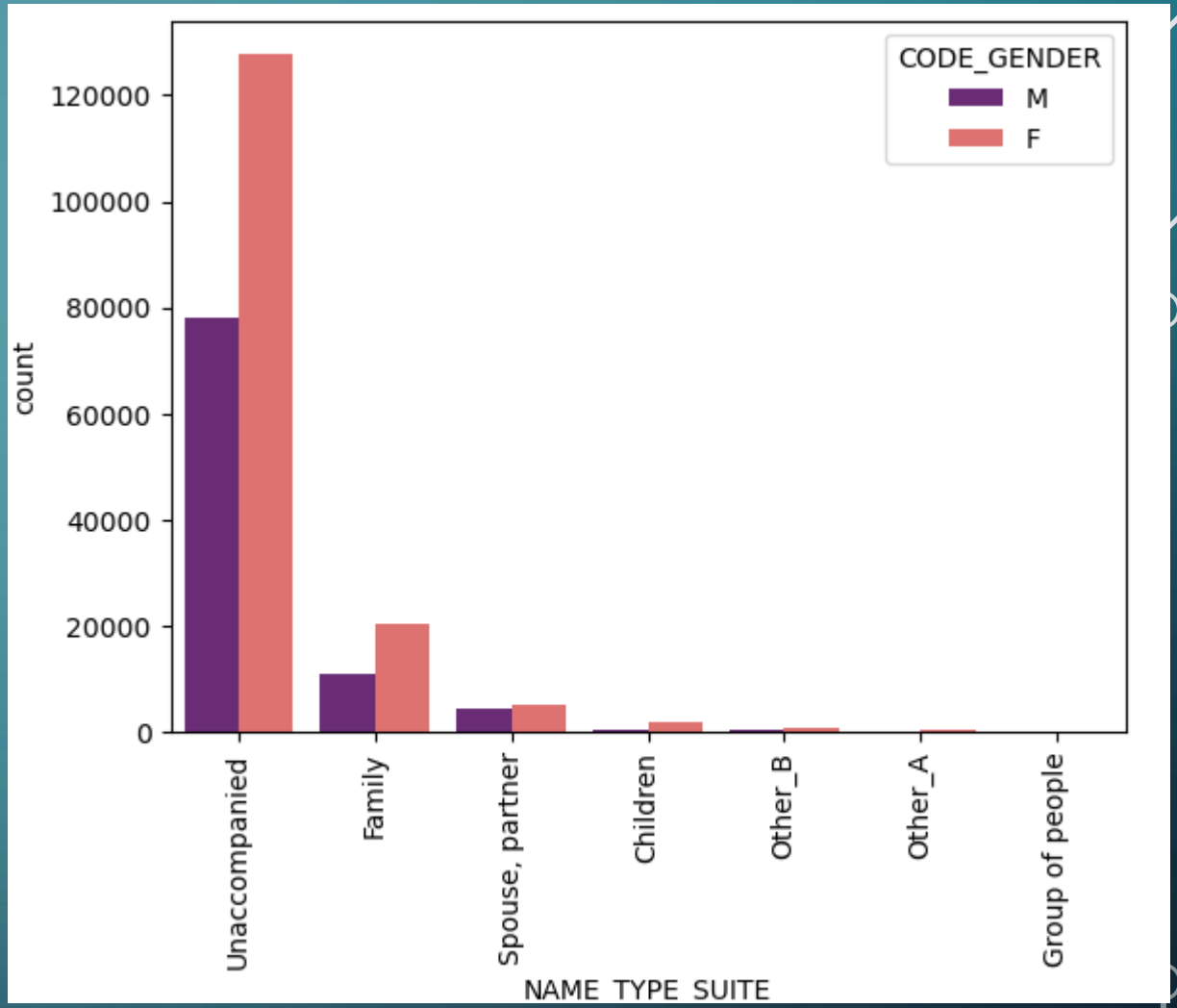
DISTRIBUTION OF CREDIT GROUP

- Credit Group - for Target 1 it is less compared with Targets 0 which is a good sign and company will have lesser defaulters.
- Mostly Medium group ranging from 0.3-0.6 quantile has highest number





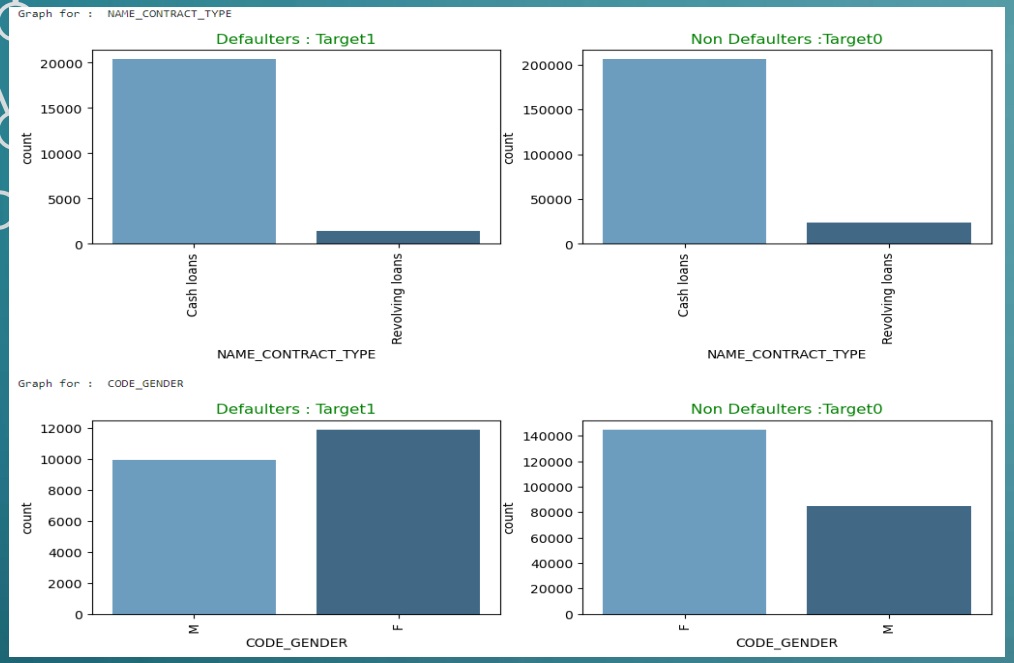


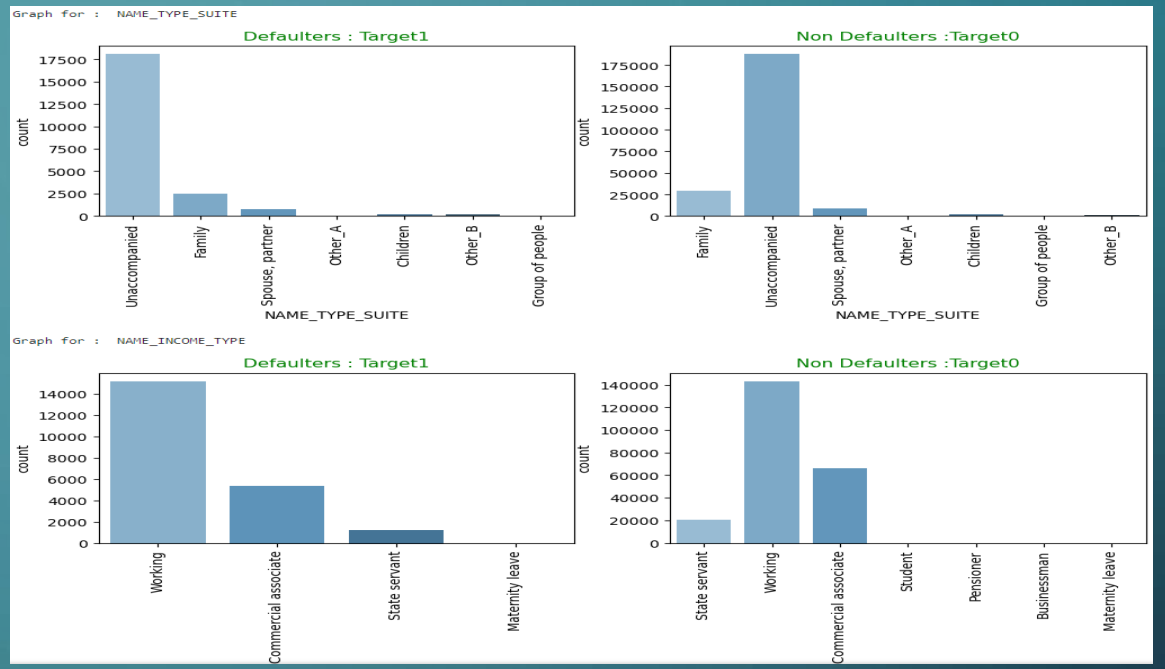


Points to be concluded

- For income types ‘working’, ’commercial associate’, and ‘State Servant’ the number of credits are higher than others.
- Females have more credits compared to males.
- Less number of credits for income type ‘student’ ,’ pensioner’, ‘Businessman’, and ‘Maternity leave’.
- For Name Type Suite Unaccompanied are the majority loan applicants followed by Family and Female are the highest among all groups.

Analysis on Contract Type , Name Type , Income type count plots for Target vs Non Targets

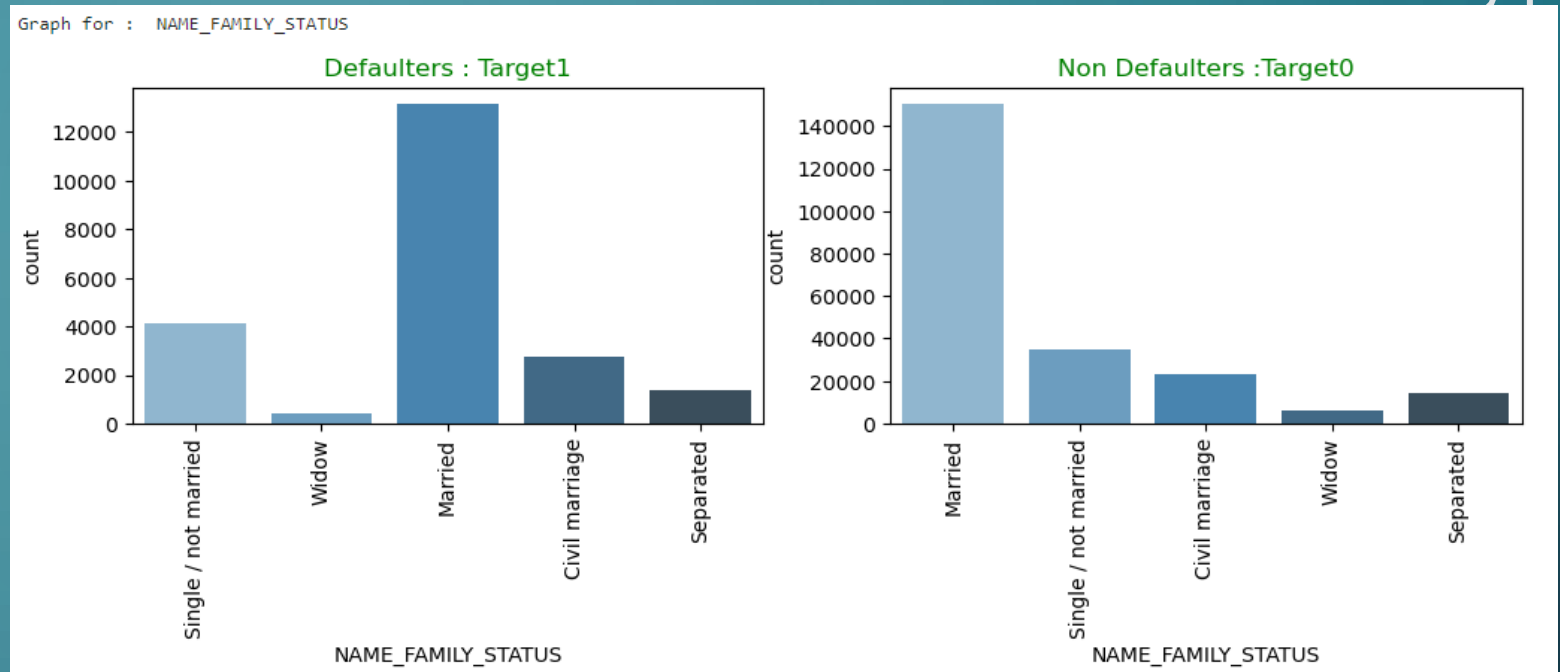


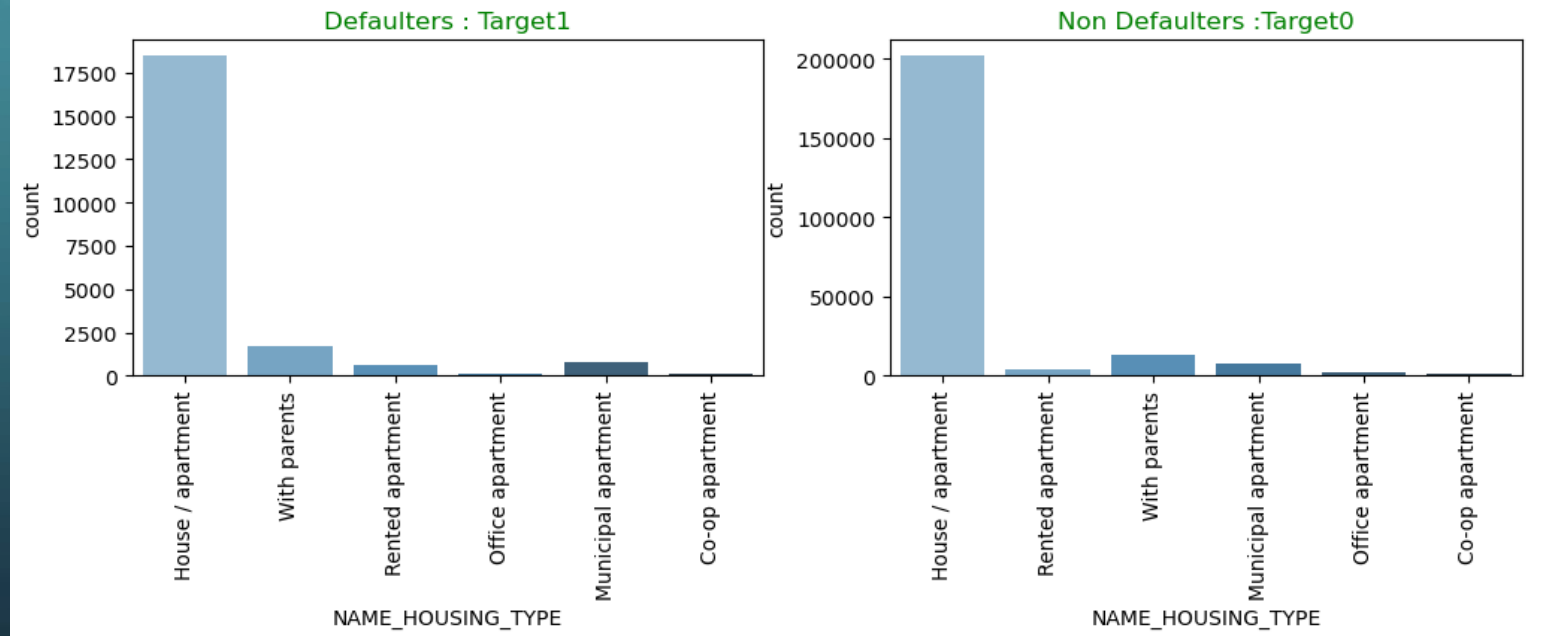


Few notable points

- NAME_CONTRACT TYPE- Cash Loans are large part of the company's portfolio. For Target 1 - 95% and almost 85% for Target-0
- CODE_GENDER - It seems there are large number of female applicants for Target 1 and Target 0 and Males are defaulting more than females
- FLAG_OWN_CAR - we can see applicants with car ownership are less likely to be dafaulters
- NAME_TYPE_SUIT - almost 80-90% in both Target 0 and Target 1 are applying are Unaccompanied. Indicating, this is not a parameter that can influence loan payment default.
- NAME_INCOME_TYPE - Working class is the highest percentage applied for loan. - if applicant on Maternity Leave then she will mostly going to default , also State Servant are less likely to default. - Pensioner, businessman and student have less payment defficulties.

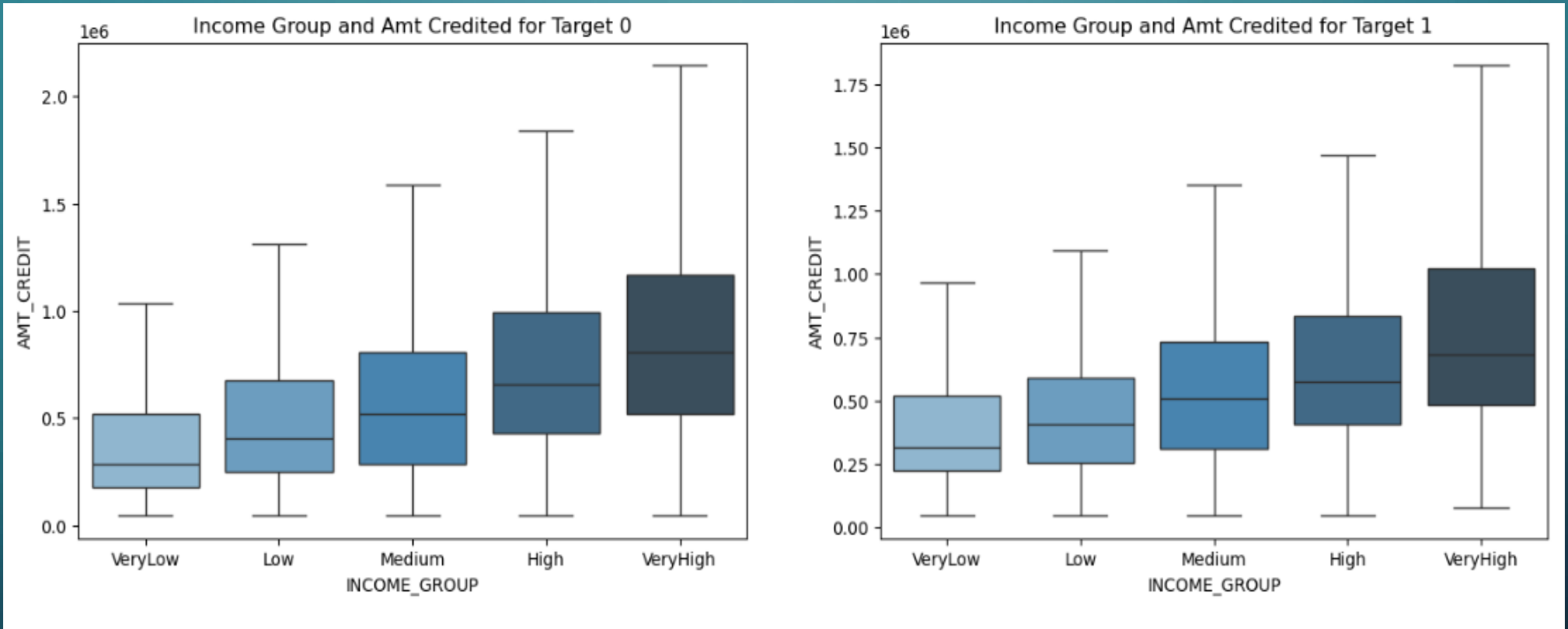
- NAME_FAMILY_STATUS - Married and widowed applicants are less likely to default on the loan payments.
- NAME_HOSUING_TYPE -85-90% in Target 0 and Target 1 applicants are staying in "House/apartment“
- We observe an increase in the percentage of Payment Difficulties who live with their parents when compared to the percentages of Payment Difficulties and non-Payment Difficulties.





Graph for : FLAG_MOBIL

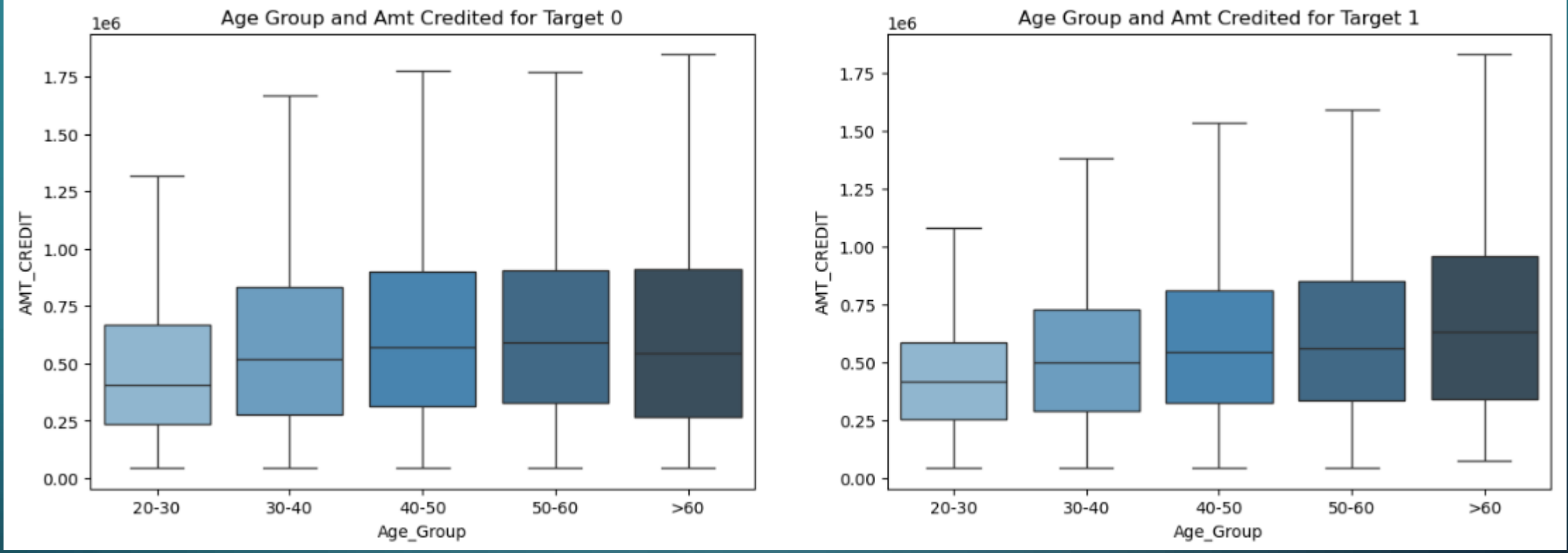
ANALYZING – INCOME GROUP VS AMT_CREDIT FOR BOTH DEFAULTER (TARGET 1) AND NON-DEFAULTER (TARGET 0)



Insights :

- We can infer that the maximum number of loans is given to the medium-income group.
- Default value per loan is highest in the High-income group as the AMT_CREDIT is higher too, so the loan book of the financial institution can get affected due to the higher amount not being paid back.
- Amount Credit is showing a positive trend with the Income group in both data sets

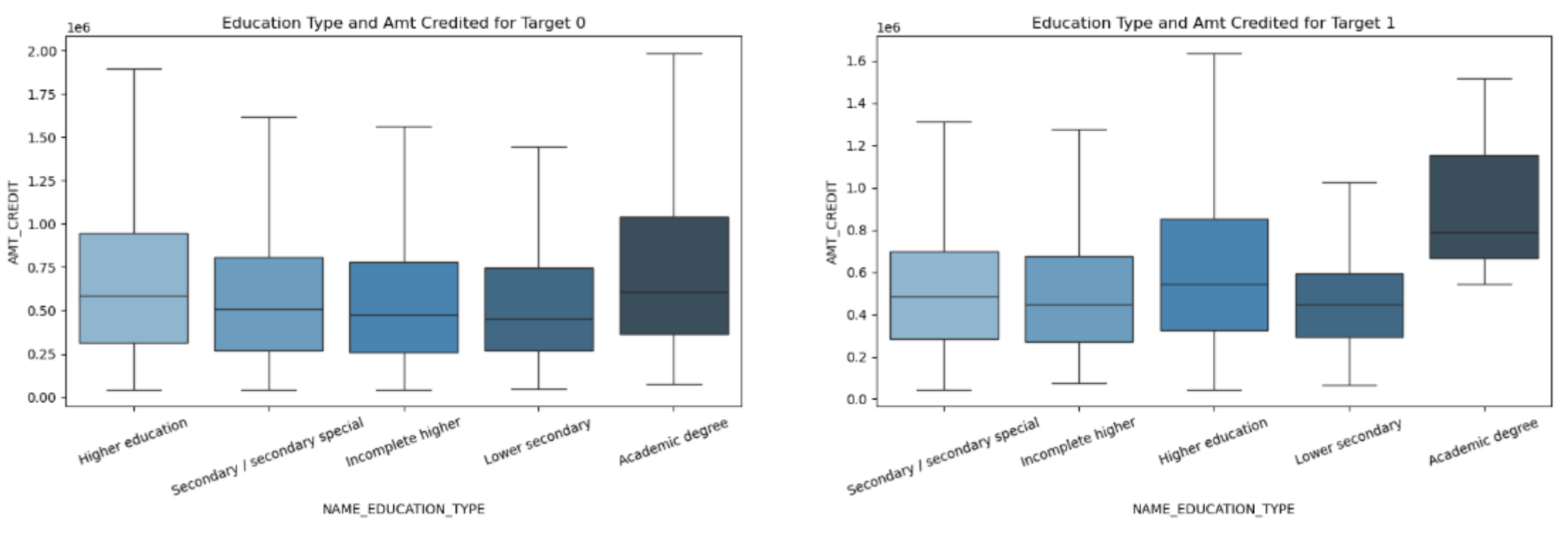
ANALYZING - AGE_GROUP VS AMT_CREDIT FOR BOTH DEFAULTER (TARGET 1) AND NON-DEFAULTER (TARGET 0)



Few notable points

- AGE_GROUP - 30-40 are more among Non-Defaulter.
- Age does seem like influencing defaulters.
- Most of the amount is credited to age groups 30-40 and 40-50, seems these two age groups are the bank's primary target section.
- Median Amount of credit for age group >60 is higher for Defaulters.

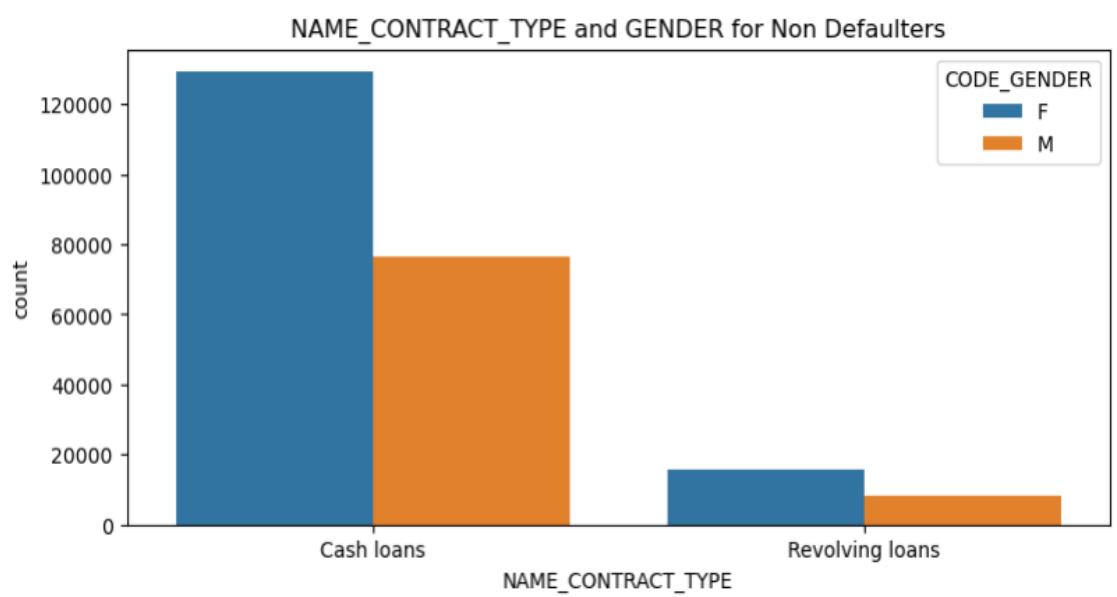
ANALYZING - EDUCATION_TYPE VS AMT_CREDIT FOR BOTH DEFAULTER (TARGET 1) AND NON-DEFAULTER (TARGET 0)

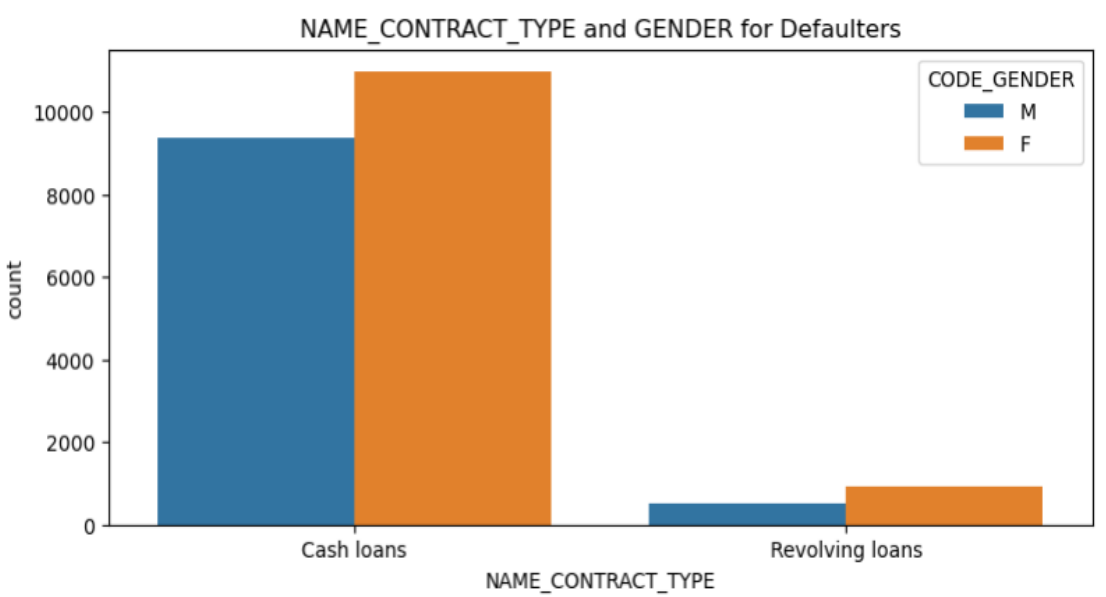


Insights:

- Amount credit median for Academic degrees is higher among defaulters this indicates loans to Academic degrees holders should be given after taking all precautions and mandatory background checks.
- The number of applicants is increasing with the increase in education categories.
- Higher and Secondary education group seems to be applicants with good amount credits.

BIVARIATE ANALYSES – NAME OF CONTRACT AND GENDER(FOR BOTH DEFAULTER (TARGET 1) AND NON- DEFAULTER (TARGET 0)

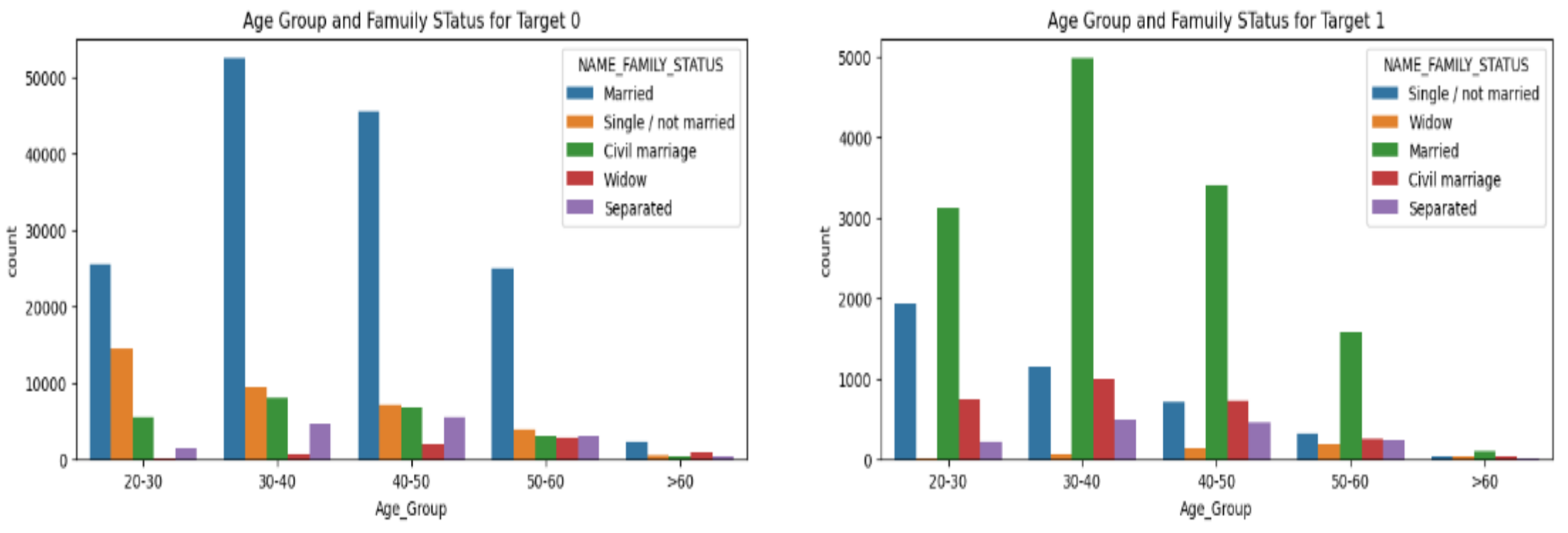




Insights:

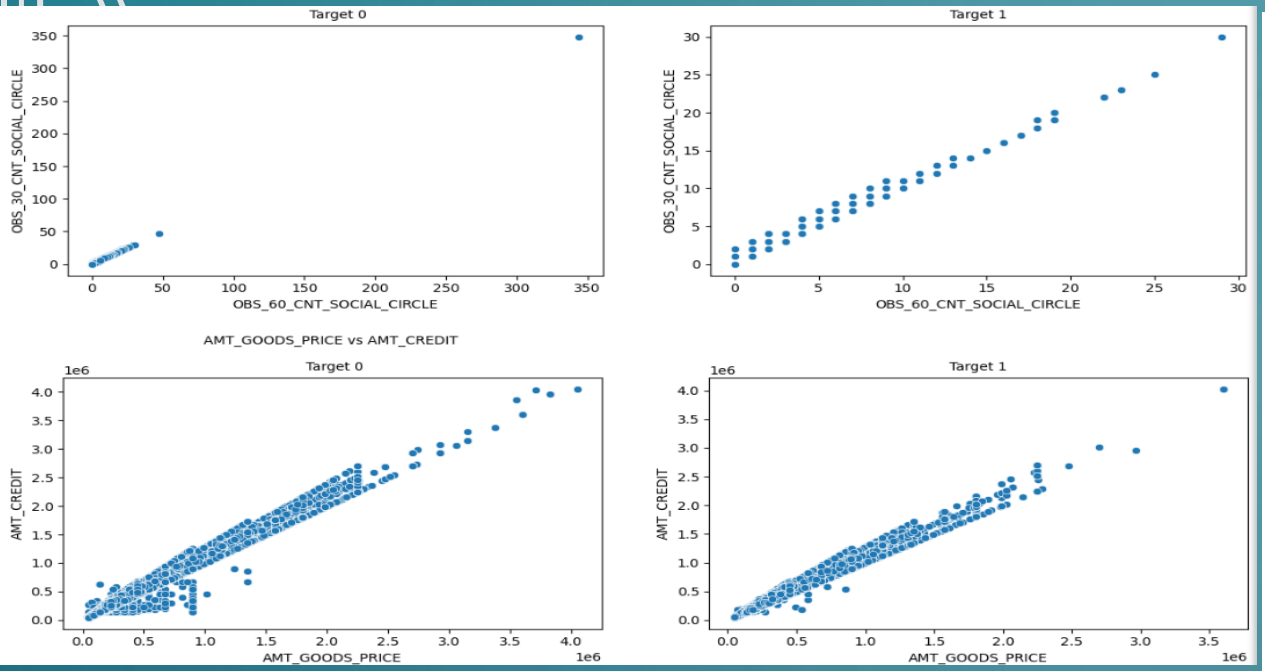
- Cash loans percentage is higher among all loan types for both the Target 0 and Target 1 datasets.
- Females are higher defaulters under cash loans and Revolving loans than male applicants.

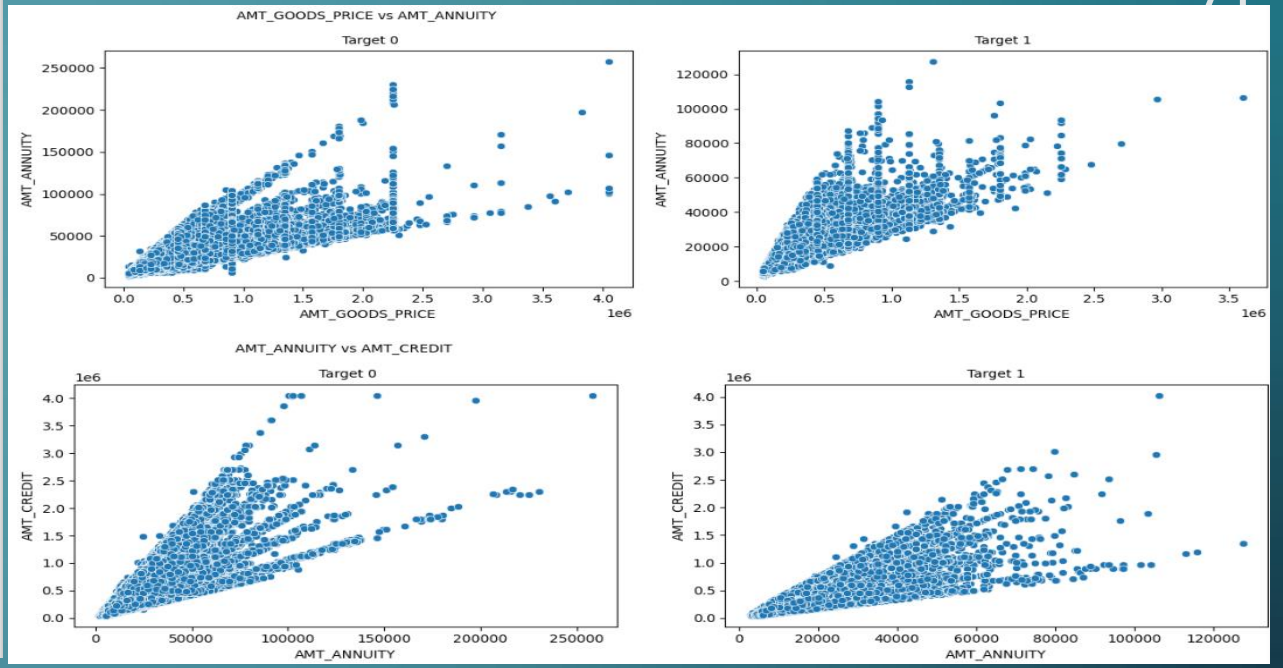
BIVARIATE ANALYSES – NAME OF CONTRACT AND GENDER FOR BOTH DEFAULTER (TARGET 1) AND NON- DEFAULTER (TARGET 0)

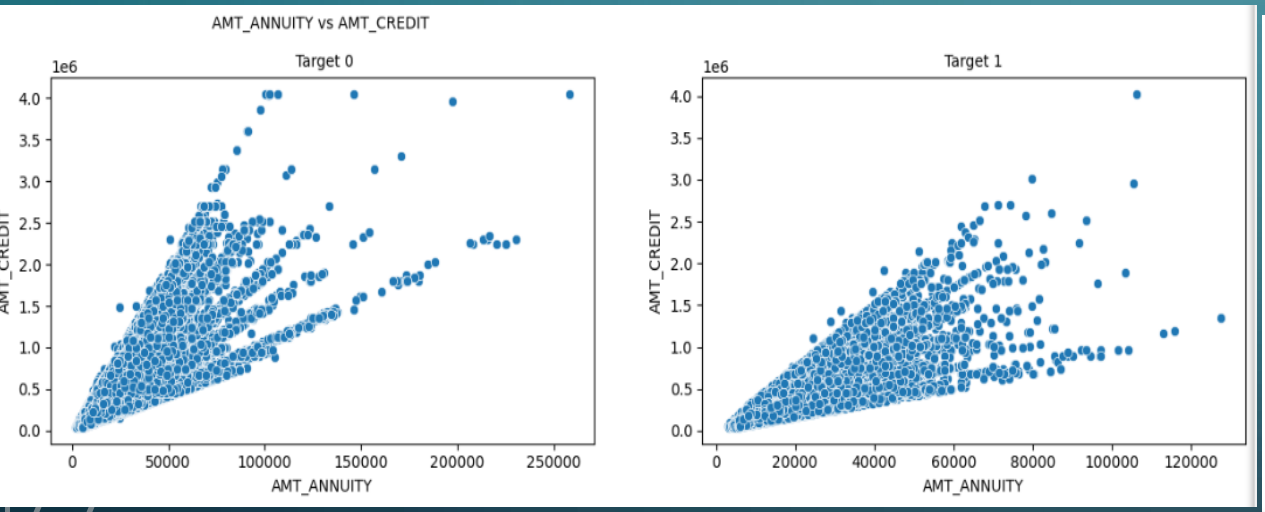


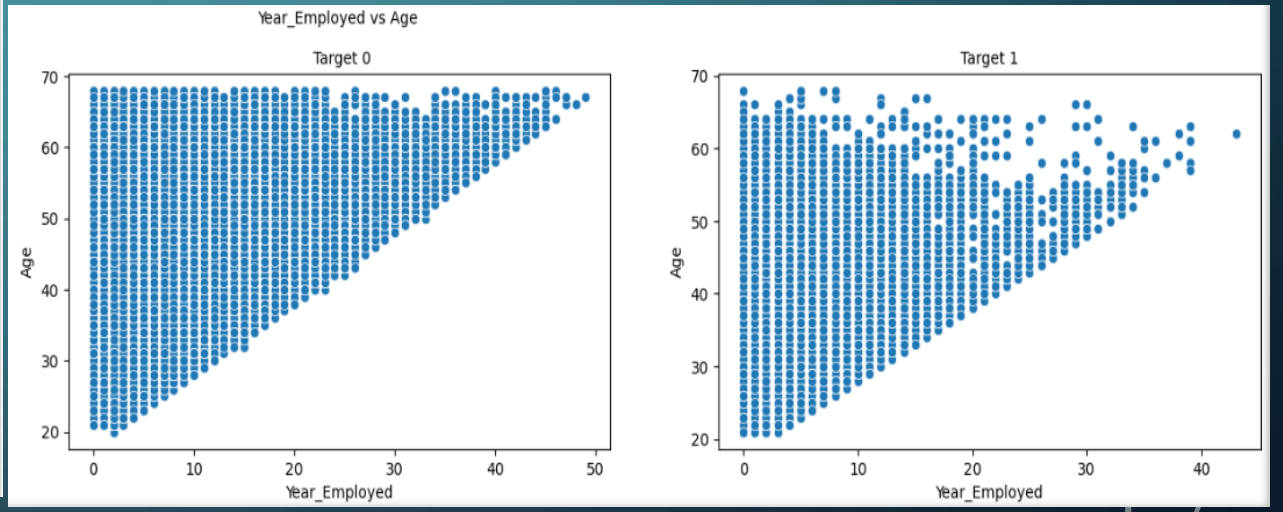
Insights:

- Married applicants are the highest percentage to apply for loan and they are one to be most defaulters across all Families Status.
- Widow are among very less applying for the loan but their defaulting number is higher.
- The age group of >60 has fewer defaulters.









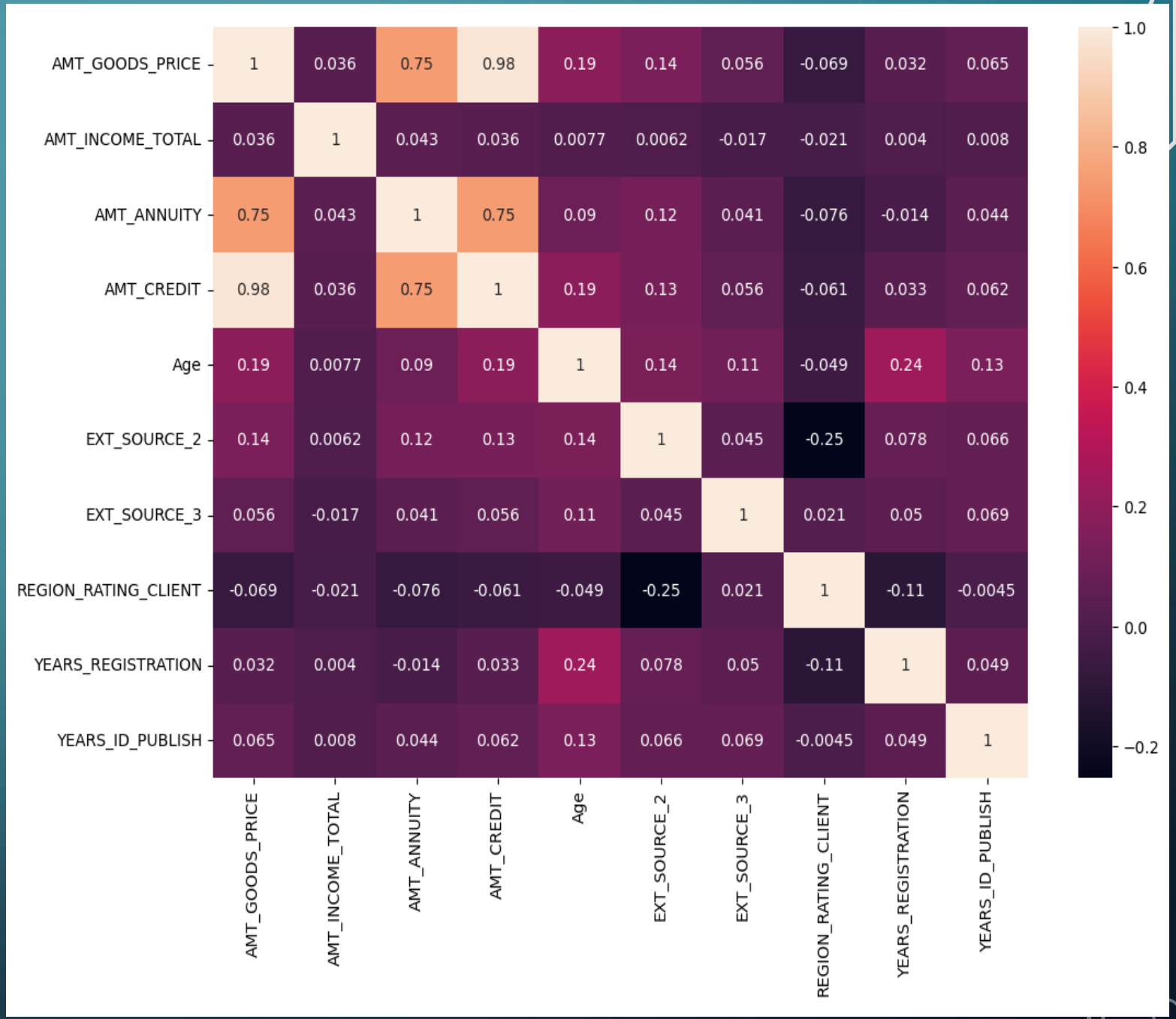
####### Insights

- OBS_30_CNT_SOCIAL_CIRCLE ,OBS_60_CNT_SOCIAL_CIRCLE - we find a strong linear positive correlation with clients obser social circle for 30/60 in Target 1.
- DEF_30_CNT_SOCIAL_CIRCLE with DEF_60_CNT_SOCIAL_CIRCLE- Have linear positive weak correlation with few points.
- Years employed has an outlier value of 999 and this is skewing the graph
- AMT_CREDIt and AMT_GOOD PRICE dont seem to be increasing proportionately with AMT_INCOME for TARGET 1, thus possibly leading to default

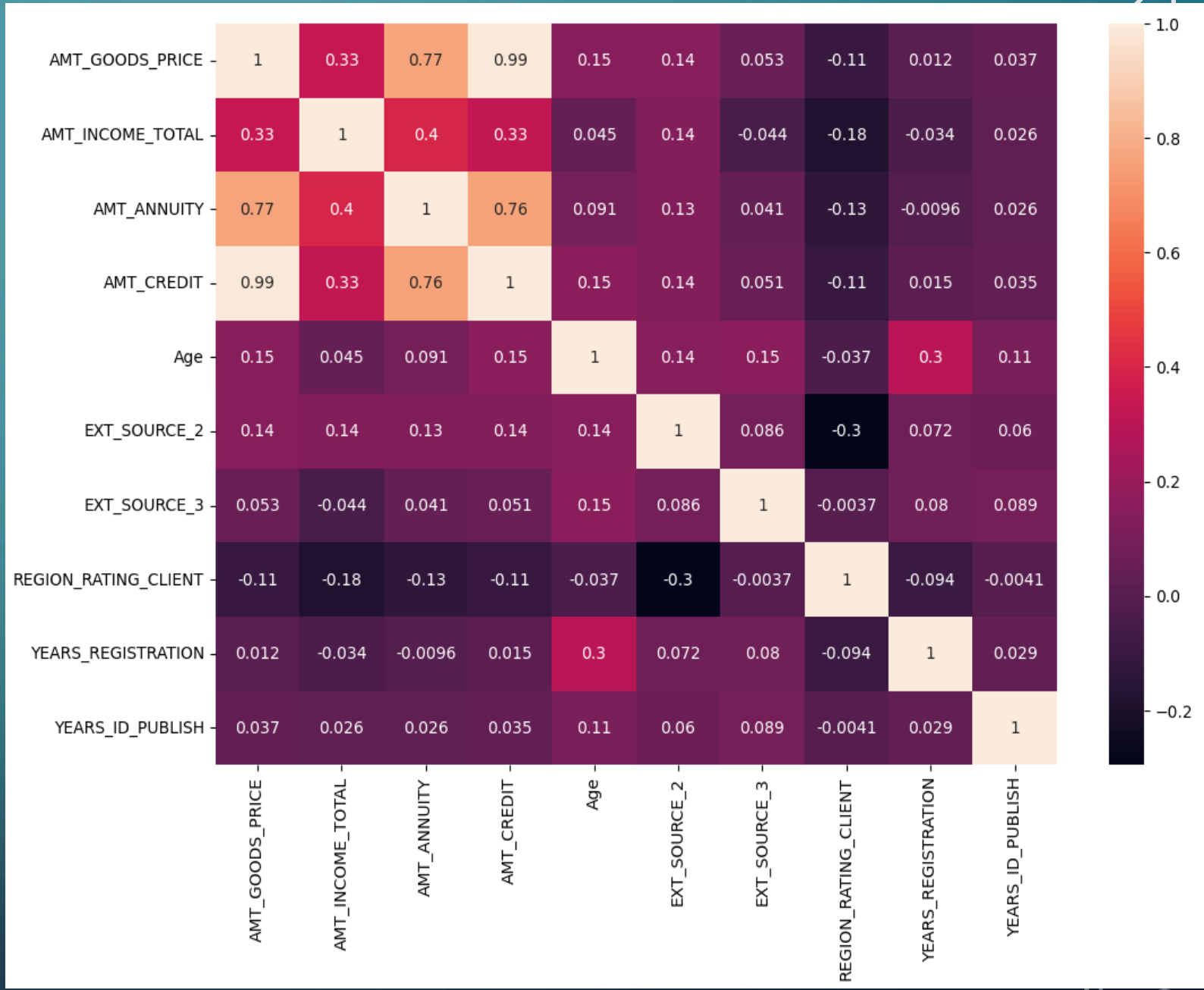
Correlation for target 1 – Defaulters

Insights

- AMT_CREDIT and AMT_GOODS_PRICE are highly correlated --> .98
- AMT_CREDIT and AMT_ANNUITY are highly correlated --> .75
- AMT_ANNUITY and AMT_GOOD_PRICE are highly correlated -->.75



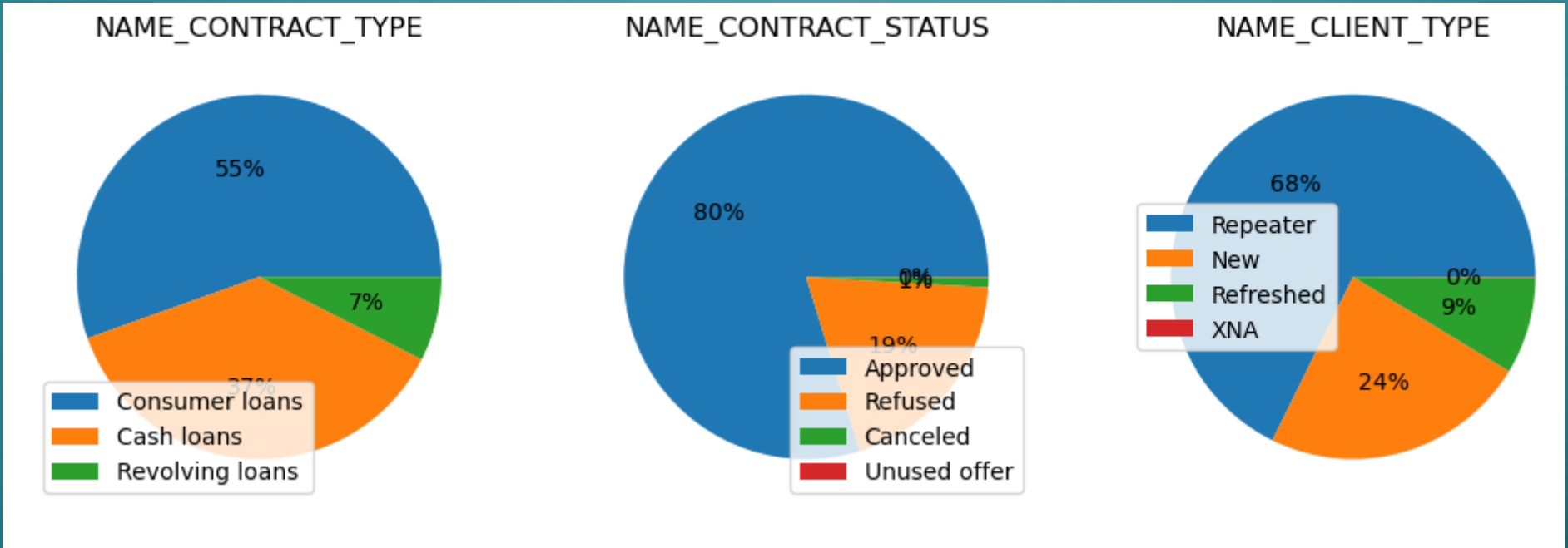
Correlation for target 0
Non- Defaulters



Insights

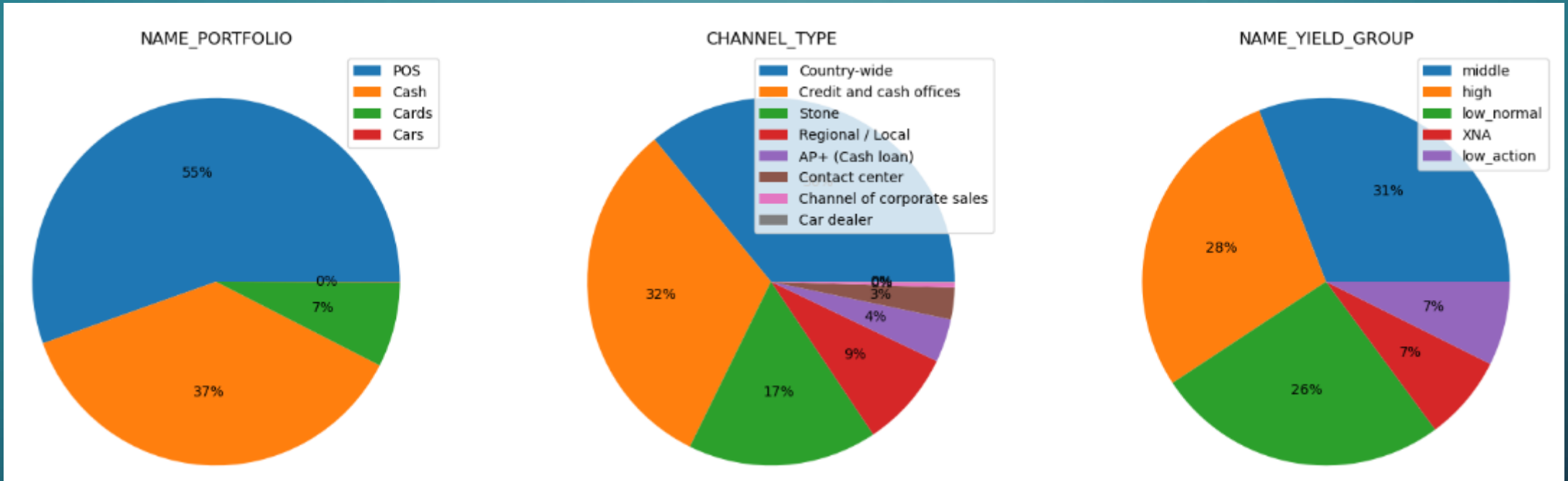
- AMT_GOOD_PRICE and AMT_CREDIT is having high correlation -->.99
- AMT_GOOD_PRICE and AMT_ANNUITY is having high correlation -->.77
- AMT_ANNUITY and AMT_CREDIT is having high correlation -->.76
- AMT_ANNUITY and AMT_CREDIT is having high correlation -->.76
- REGION_RATING_CLIENT is Negatively correlated with all variable listed in the app_corr_target_0 dataset

Analyzing Previous Application dataset



Insights:

- Consumer Loans are the highest Contract Type with 55% and Cash loans are 37%.
- Approved are 80% of total Contract status
- Repeater are 68% of Clients Type



Insights:

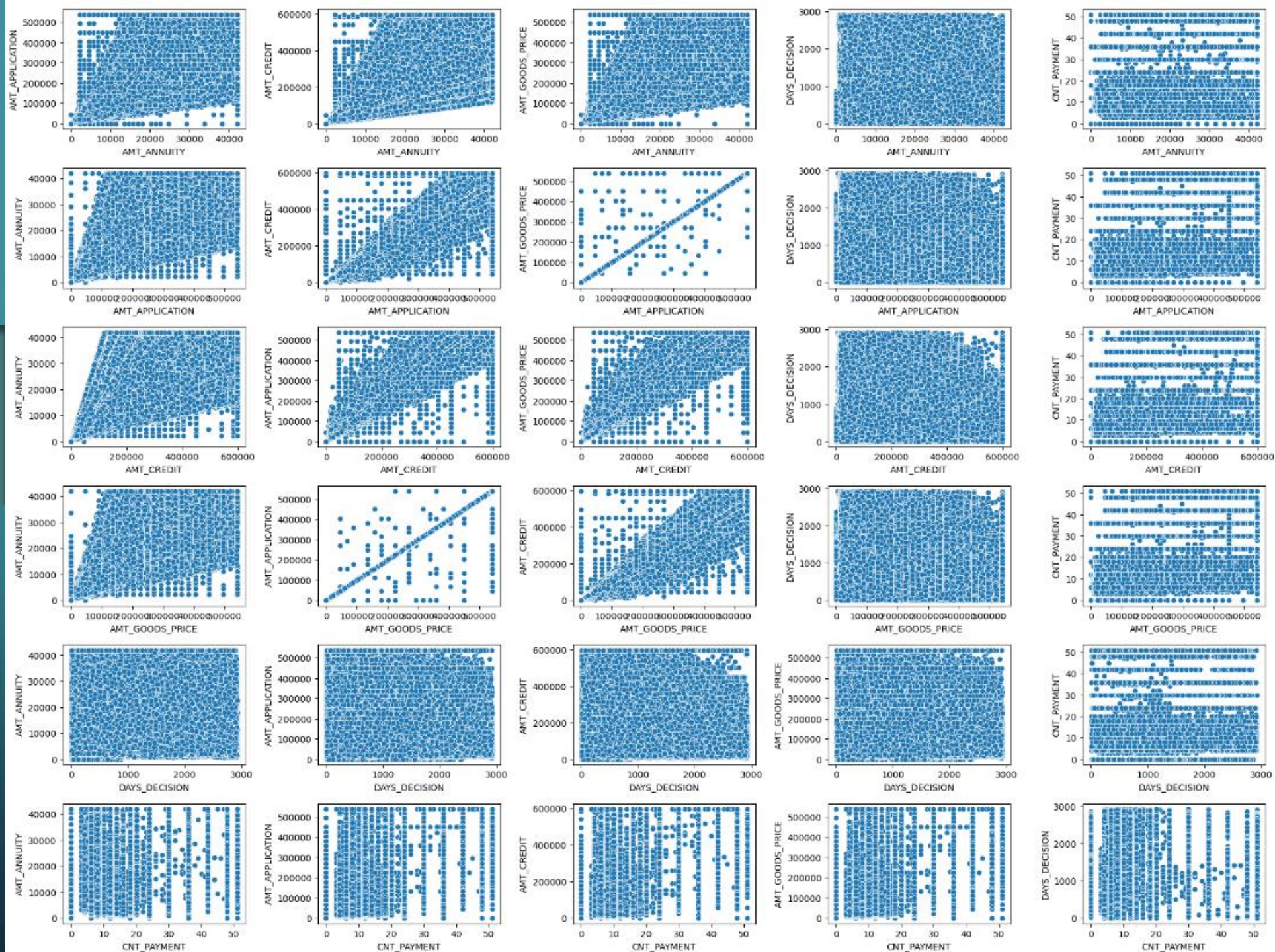
- 55% of applicants have taken POS purchase.
- Country wise and Credit/cash offices have 70% of channel type.
- Among Yield groups Middle have 31% followed by 28% high and 26% low_normal.

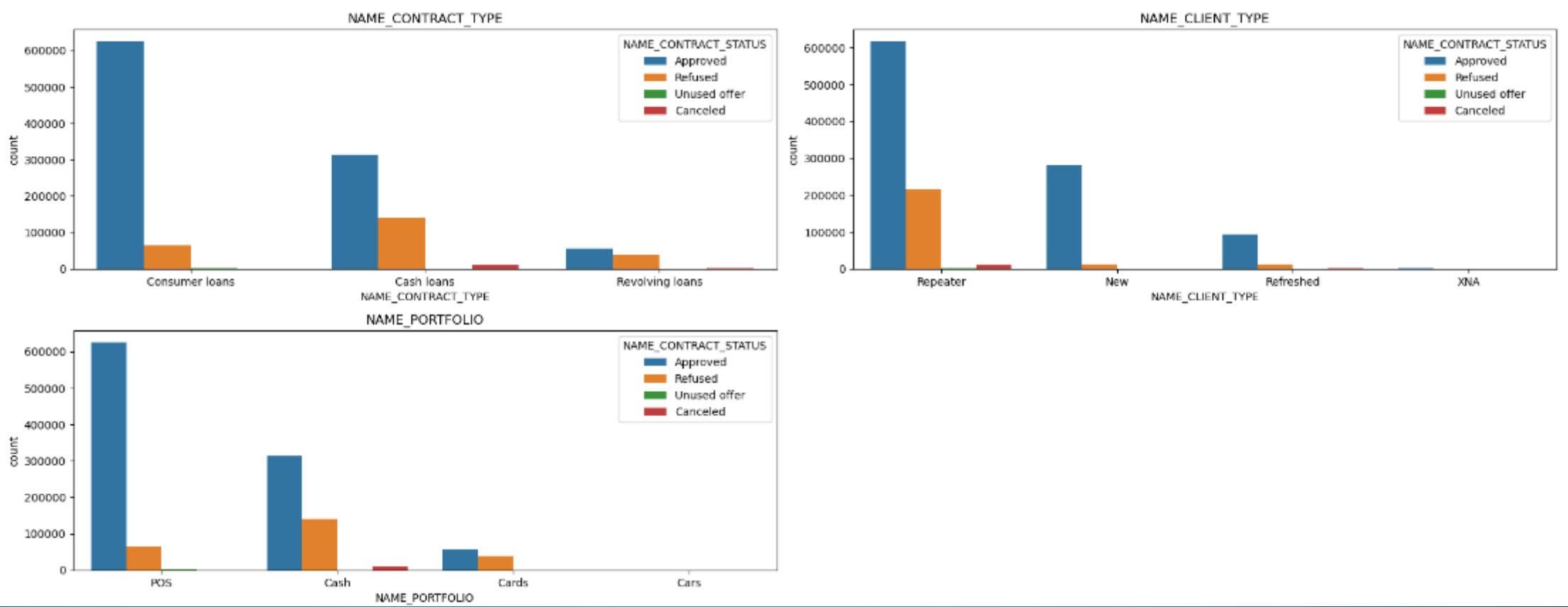
Bivariant numerical column Analysis

Strong correlation

1. **AMT_CREDIT and AMT_ANUUITY**
2. **AMT_GOOD_PRICE and AMT_ANNUITY**
3. **AMT_CREDIT and AMT_APPLICATION**
4. **AMT_GOOD_PRICE and AMT_APPLICATION**

- Variable -->AMT_GOODS_PRICE, AMT_ANNUITY, AMT_APPLICATION - as expected have high correlation.
- Variable AMT_Credit to AMT_GOOD_PRICE also shows high correlation.
- Variable CNT_Payment should ideally have had a high correlation with AMT_credit, but no such correlation can be seen.



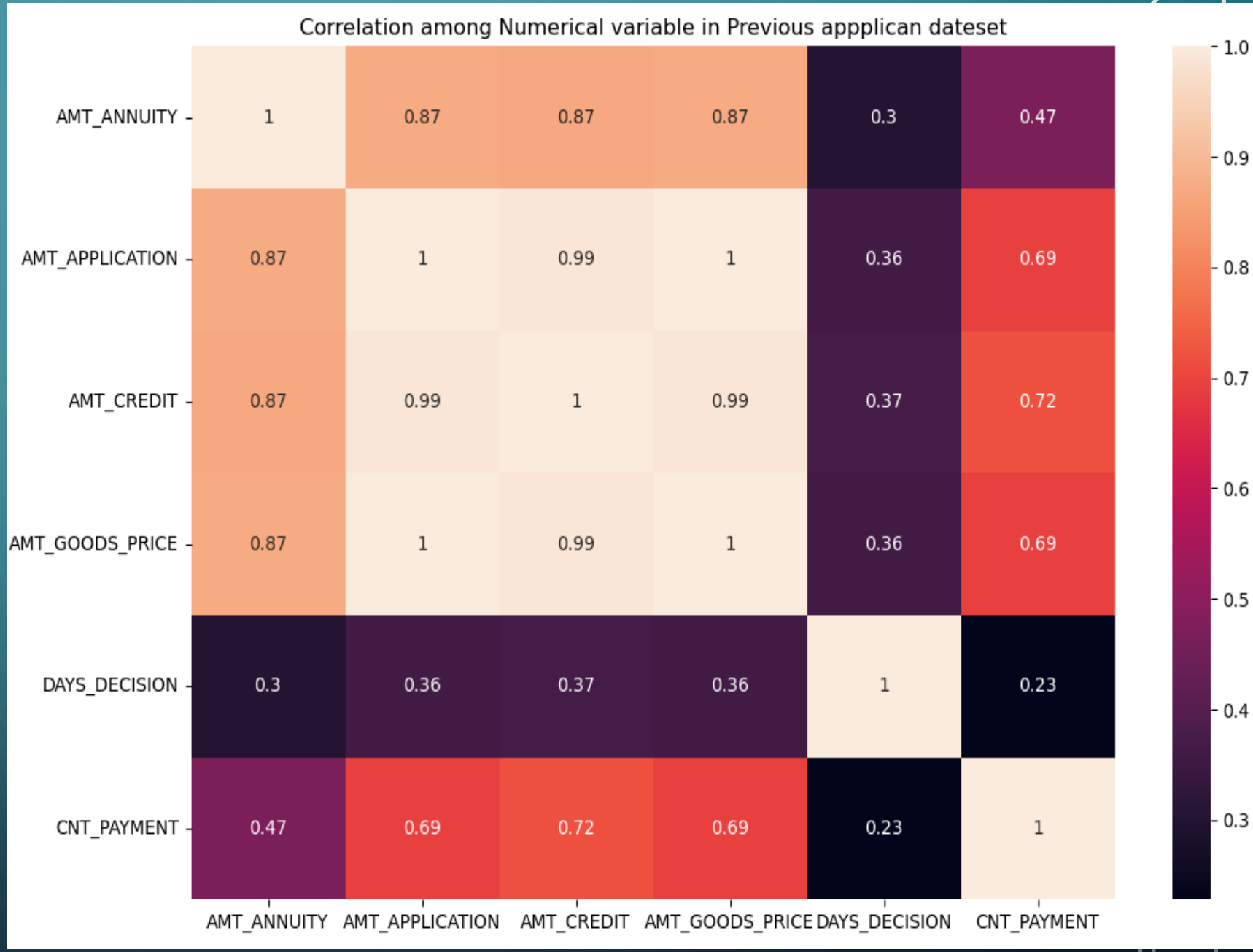


Insights

1. In approved category, consumer loan has largest no of applicants.
2. There seem to be no cancelled loans in cash loan category than consumer loan.
3. More cash loans have been refused than consumer loans.
4. The bank has more repeaters in all approved, refused, unused, cancelled categories
5. POS transactions seem to be consumer loans and similar to point 2 - more cash loans have been refused than POS.

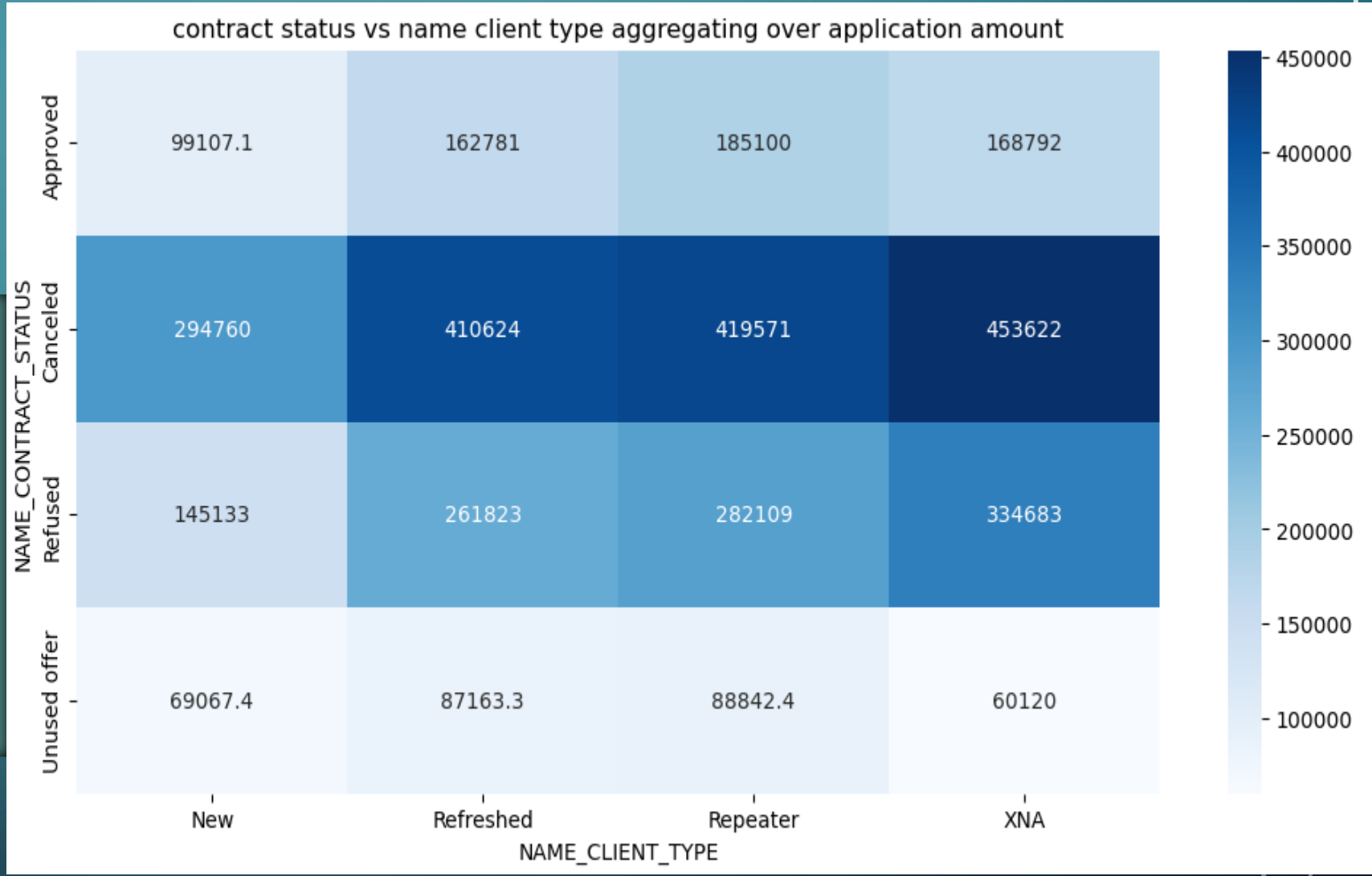
Insights

1. AMT_ANNUITY have high positive correlation with
2. AMT_APPLICATION,AMT_CREDIT,AMT_GOODS
3. - AMT_CREDIT have .99 correlation with AMT_APPLICATION, AMT_GOODS
4. - CNT_PAYMENT have good correlation with AMT_APPLICATION,AMT_CREDIT,AMT_GOODS



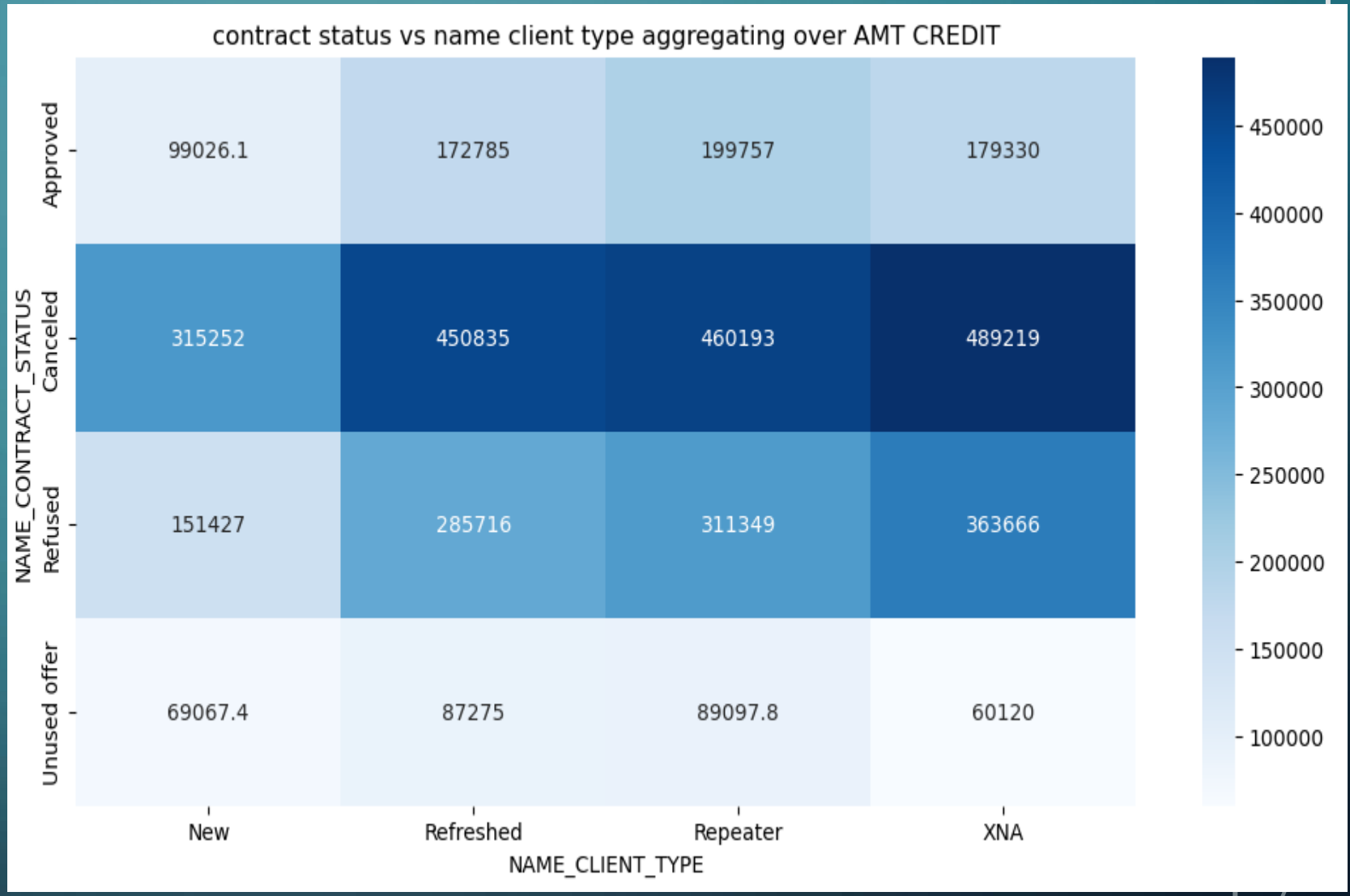
Insights

1. Unused offer have low percentage for all types of client
2. Cancelled application are highest as Bank may be taking precaution as these applicant may be a defaulters or with low credit score
3. Repeater number is higher than New applicant , seems bank is having good relationship with affordable interest rate



Insights

1. Unused offer have low percentage for all types of client
2. Cancelled application are highest as Bank may be taking precaution as these applicant may be a defaulters or with low credit score
3. Repeater number is higher than New applicant , seems bank is having good relationship with affordable interest rate

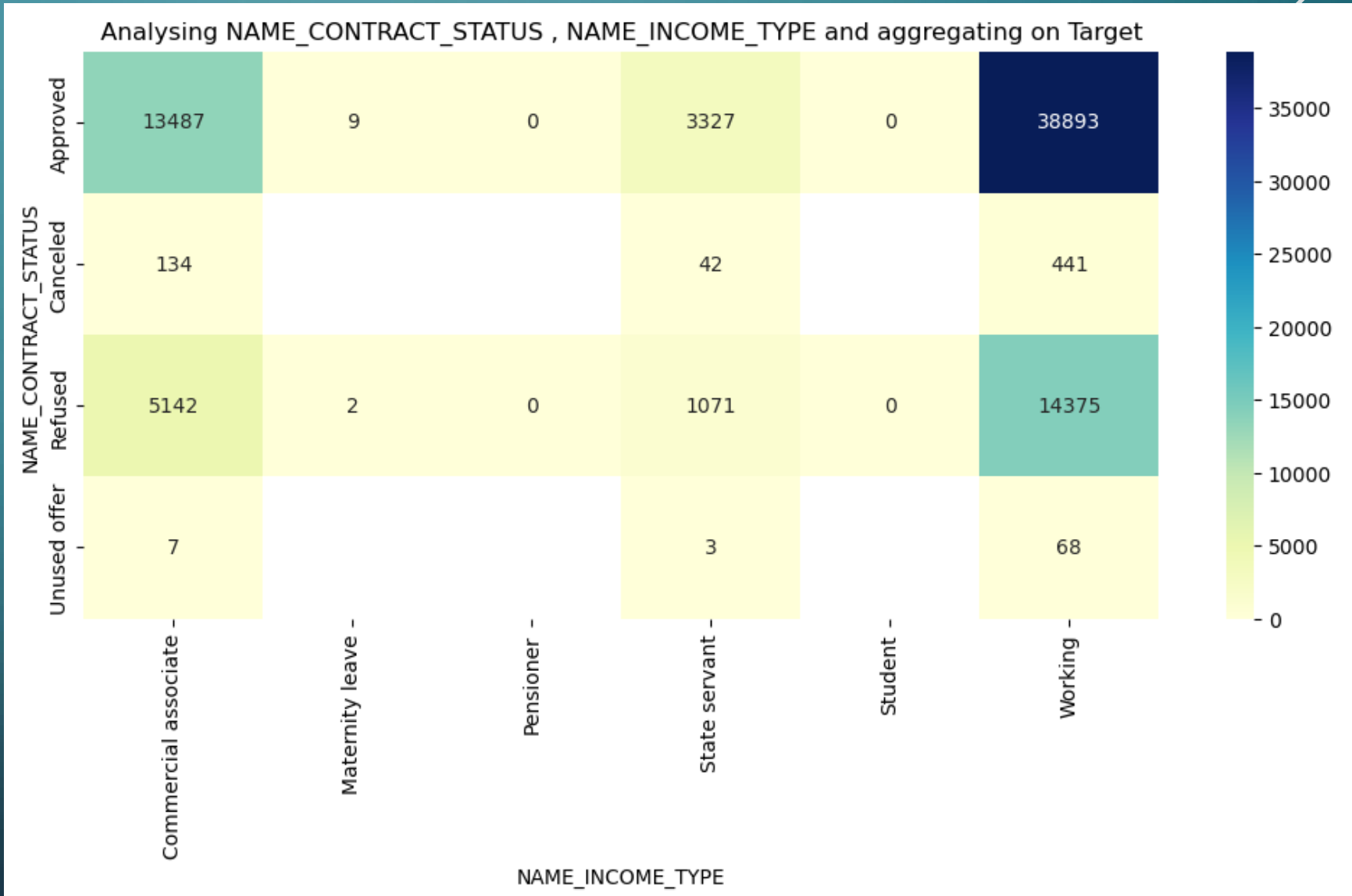


Analyzing the data after merging the Application and previous application dataset

Correlation matrix – Categorical vs categorical and aggregation on Target

Insights

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. 14,375 applicants of working class were REFUSED earlier and now have defaulted.

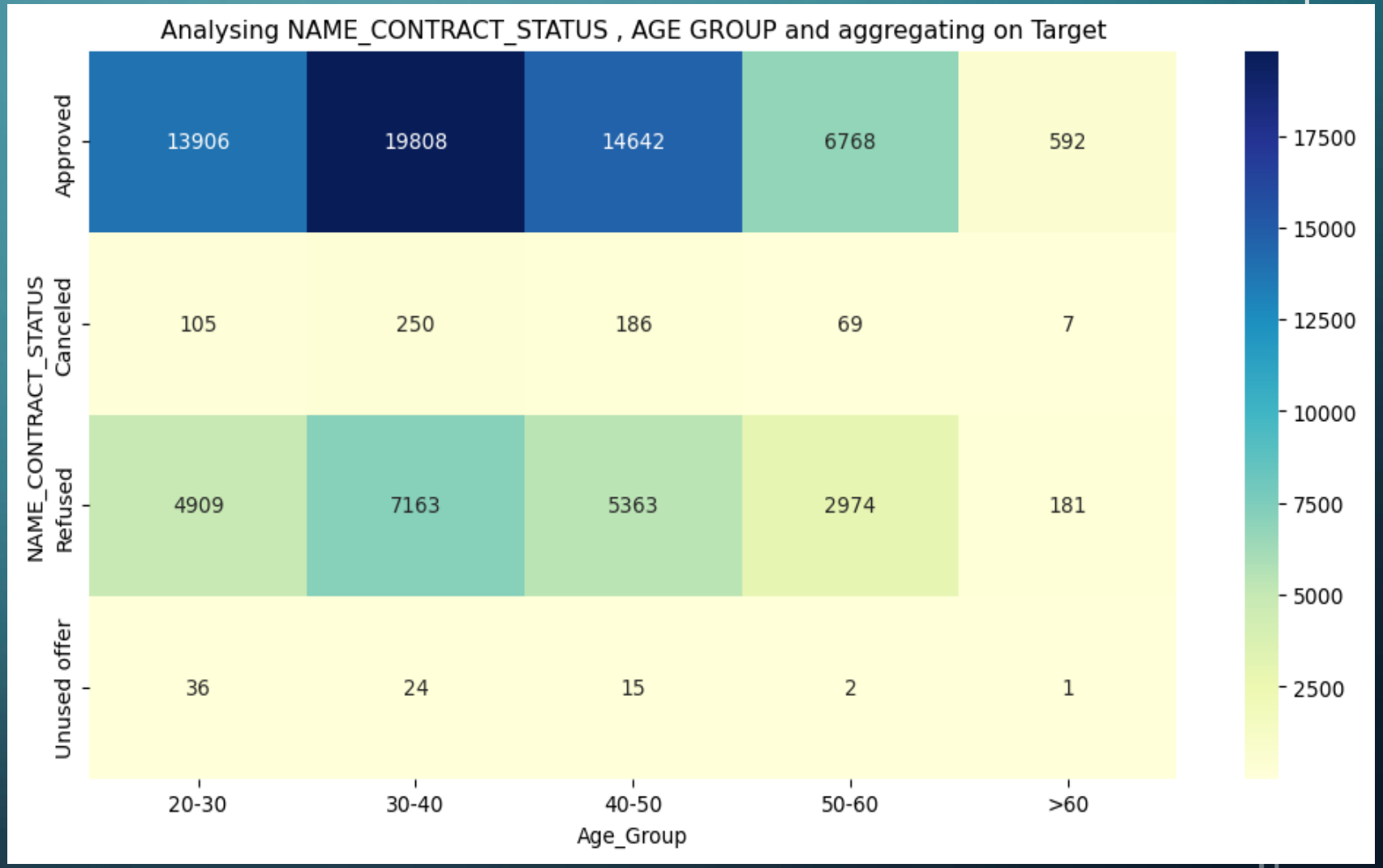


Analyzing the data after merging the Application and previous application dataset

Correlation matrix – Contract Status vs Age Group and aggregation(sum) on Target

Insights

1. Approved loans of age group 30-40 and 40-50 have higher default rates.
2. Refused, canceled, loans in the previous application have defaulted in the current one.
3. Applicants with age >60 have higher approval percentage across contract Status

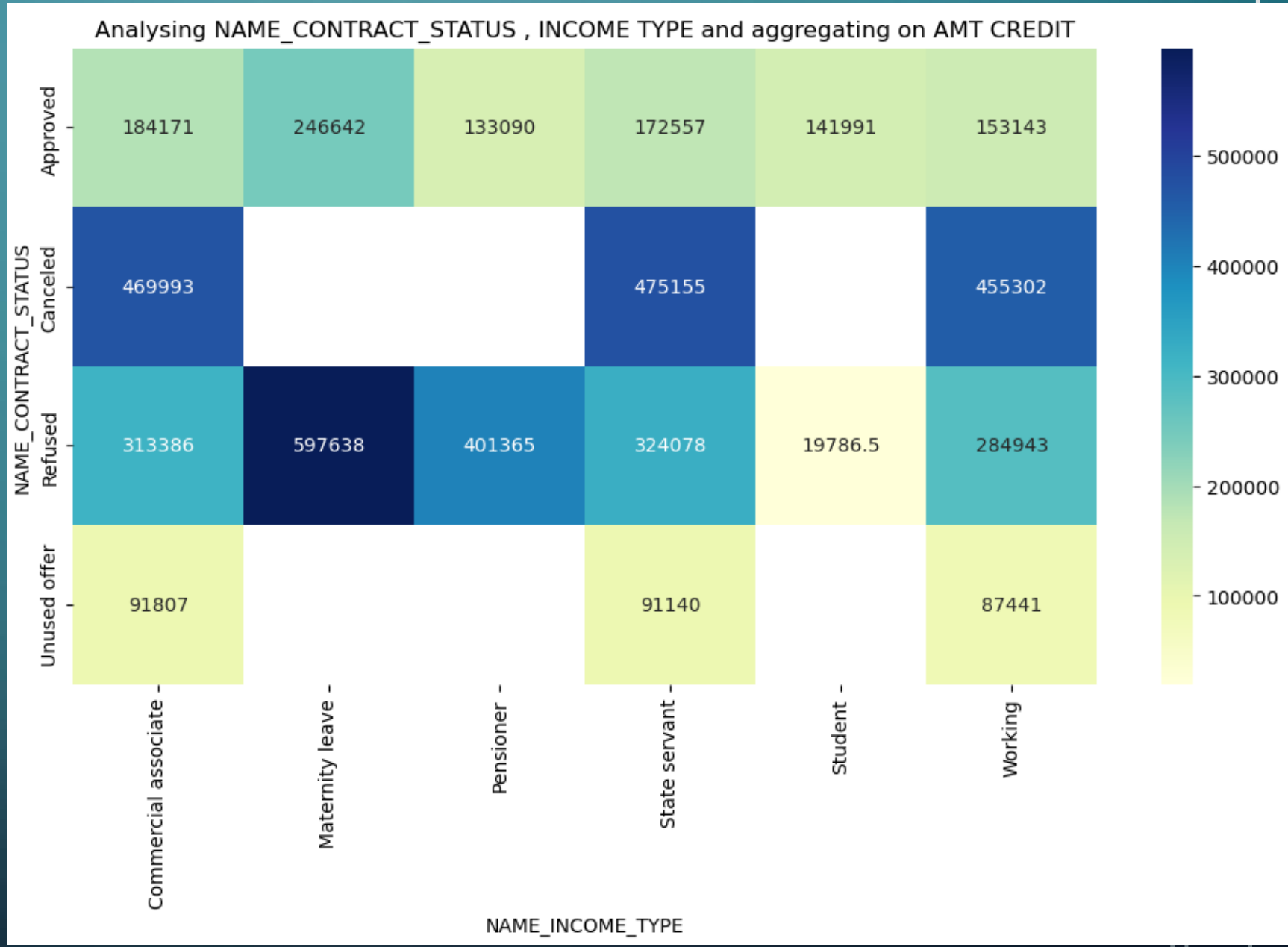


Analyzing the data after merging the Application and previous application dataset

Correlation matrix – Contract Status vs Income Type and aggregation on AMT Credit

Insights

1. Applicants with Maternity leave have the highest refusals and still banks approved them and now most of them are defaulters - Banks should avoid approving loans to Maternity leave applicants
2. Pensioners are less defaulter.



CONCLUSION

- Banks should focus more on contract types – Student, Pensioner and Businessman with housing type other than Co-op apartment for successful payments.
- Banks should focus less on income type - Working as they have a higher defaulter percentage.
- Banks should avoid Loans with the purpose Repair as they have higher number of unsuccessful payments on time.
- For Banks clients from housing type With parents are good as they are having least number of unsuccessful payments.
- We also see there were many applications with refused status in the previous application dataset but in the current application, they are paying regular payments which indicates that rejecting these applications was a wrong decision initially or there may be data entry issue while collecting the data.
- Below are the indicators for loan defaulters, bank should take extra precautions while approving loan to applicants with below variables
  - Applicants with medium income but without Own House.
  - Unemployed Male
  - Females on maternity leave
  - 25-35 years applicants , followed by 35-45 years age group
  - Labourers, Salesman, Drivers
  - Business type 3