EDA Assignment

Problem statement

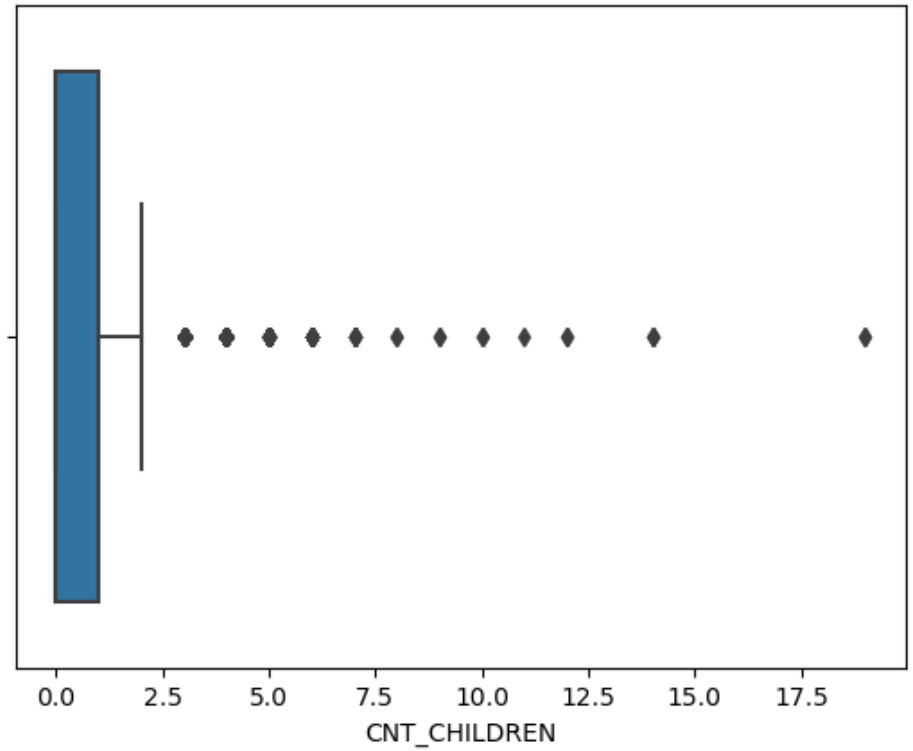
Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly.Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)

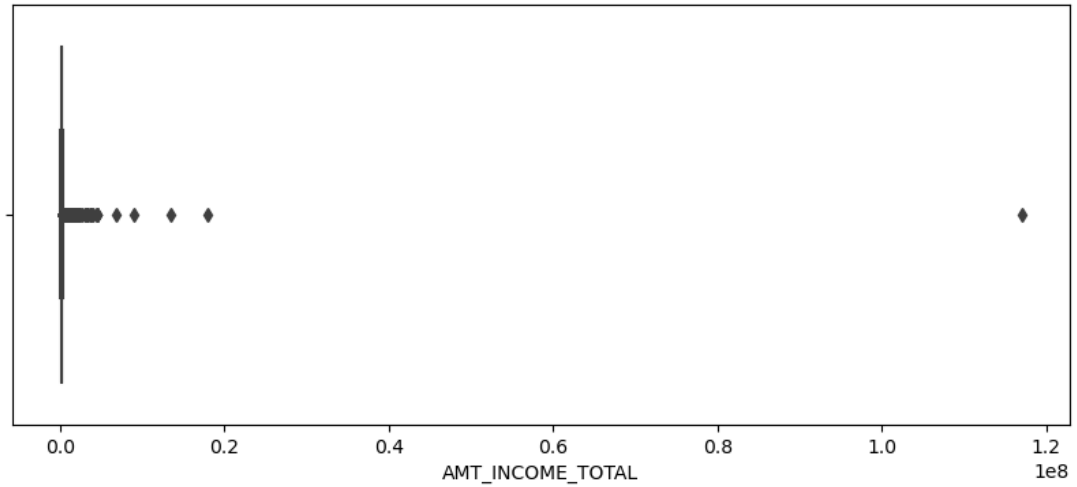
Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.

Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5, Target. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.

Include visualisations and summarise the most important results in the presentation.

Create box plot for 'CNT_CHILDREN’ and 'AMT_INCOME_TOTAL'

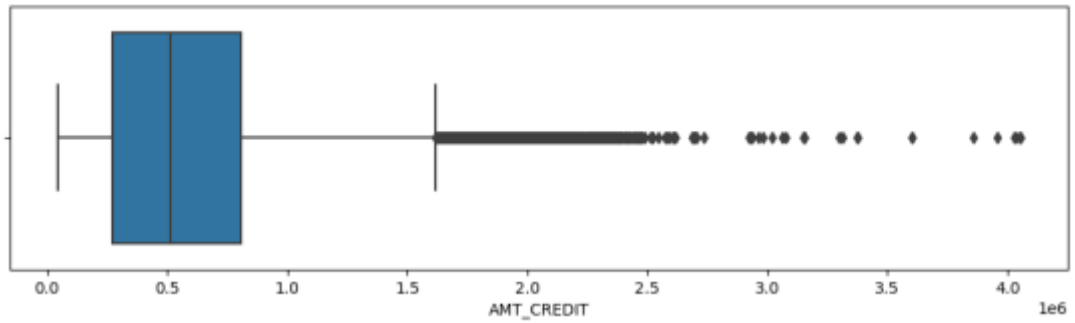


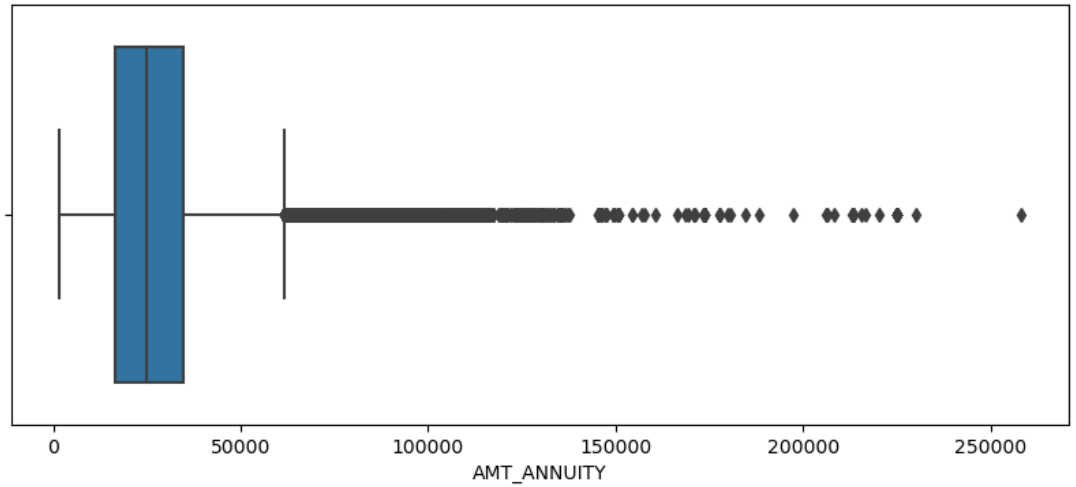


inference: high value can be seen only near single data which shows an outlier in 'AMT_INCOME_TOTAL'

inference: data points are seen in the first quartile as there is not presence of first quartile 'CNT_CHILDREN'

Create box plot for AMT_CREDIT and AMT_ANNUITY

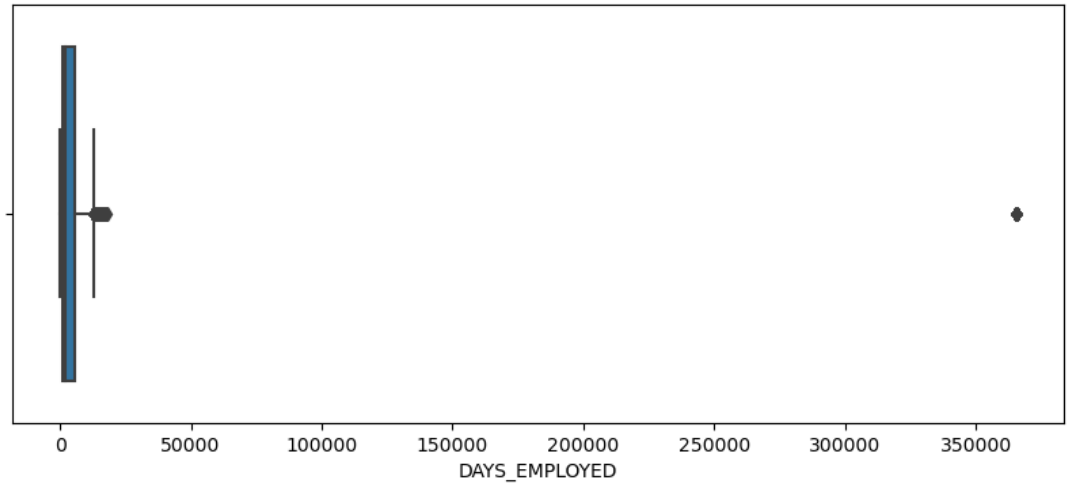




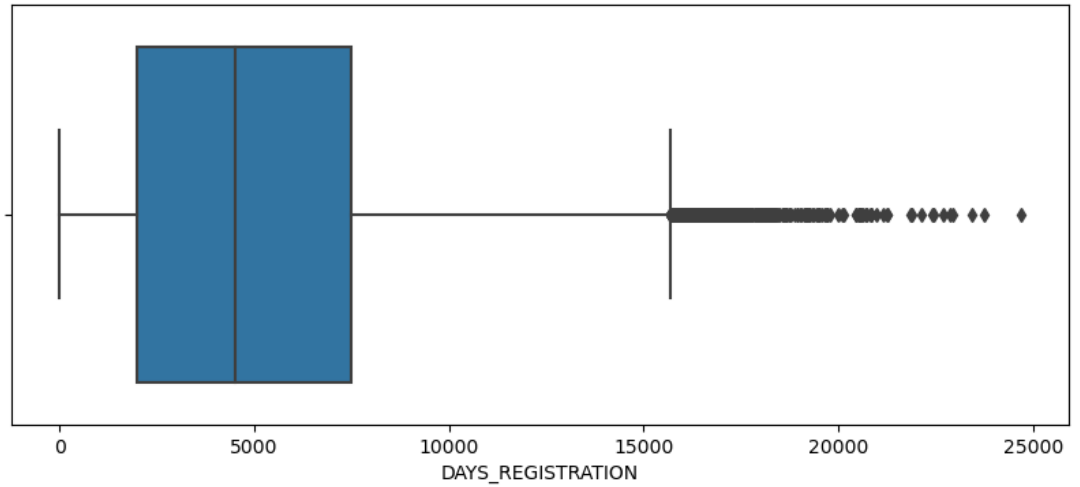
inference: if other vatriables are being considered there is an elevation in the outliers of'AMT_CREDIT'

inference : maximum range seen in lower quartile and so 3rd is being shifted towards earlier quartile

Create box plot DAYS_EMPLOYED and DAYS_REGISTRATION

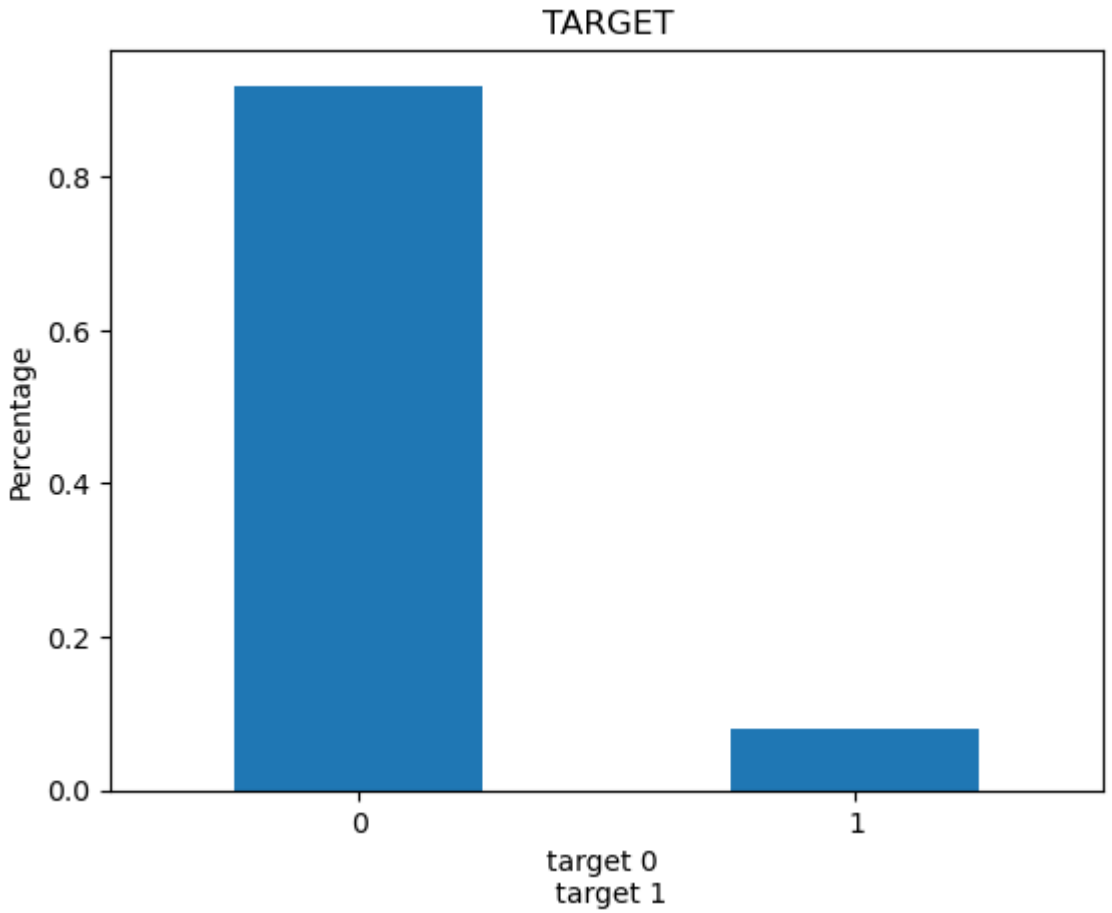


inference: more concentration seen in lower end showing small employment duration



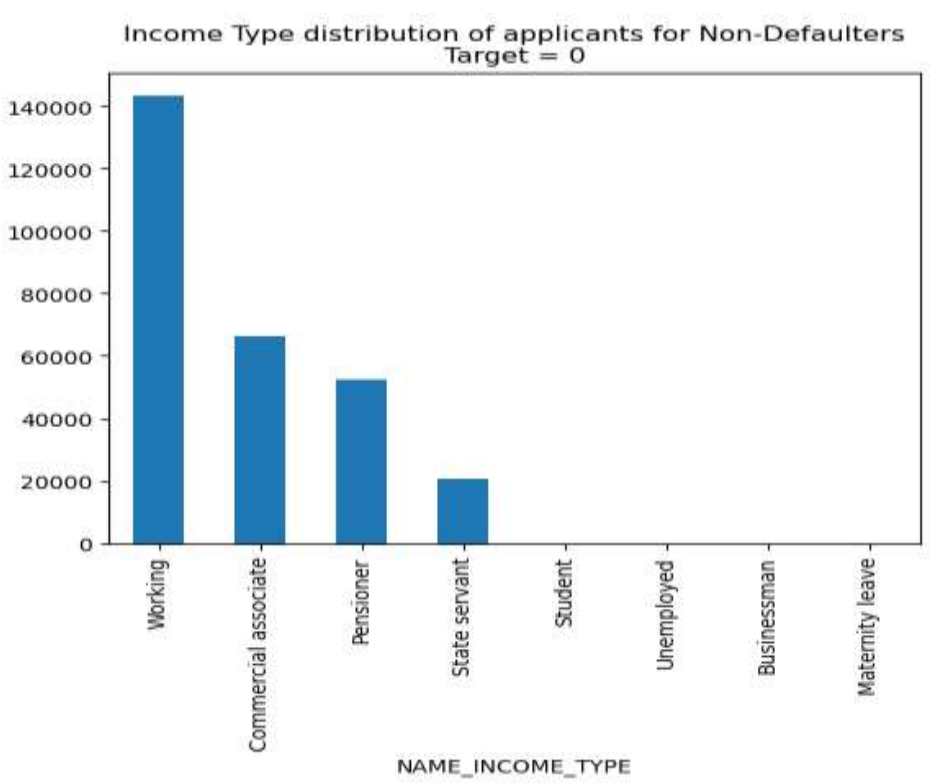
inference: for the DAYS_EMPLOYED variable, both the first quartile and third quartile remain near to the lower quartile distribution and moretowards shorter employment durations numeric columns showing outliers

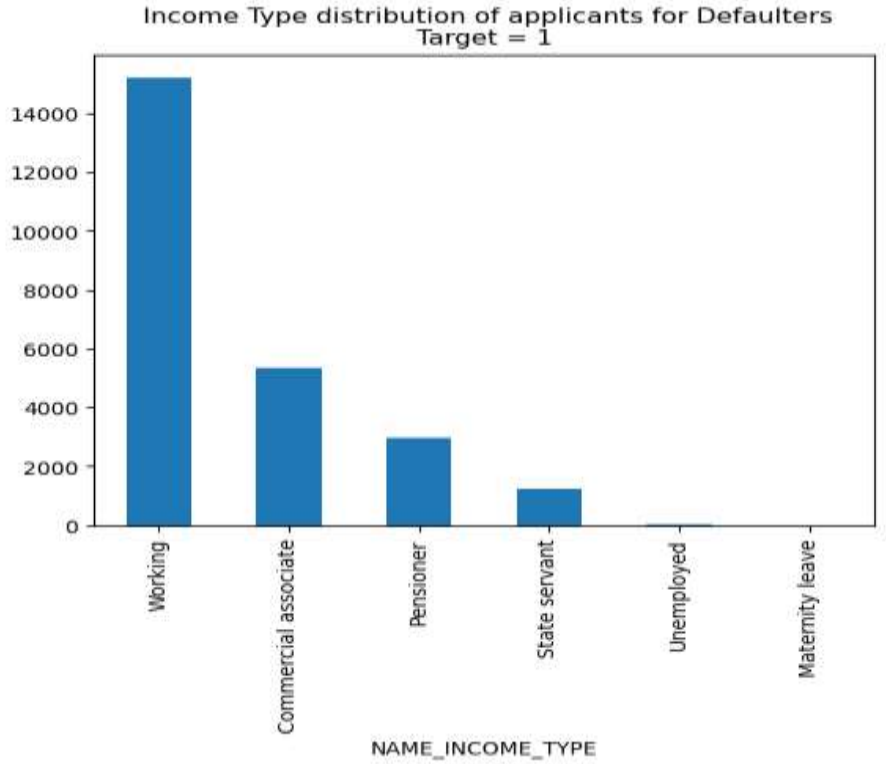
Analysis



univarate analysis

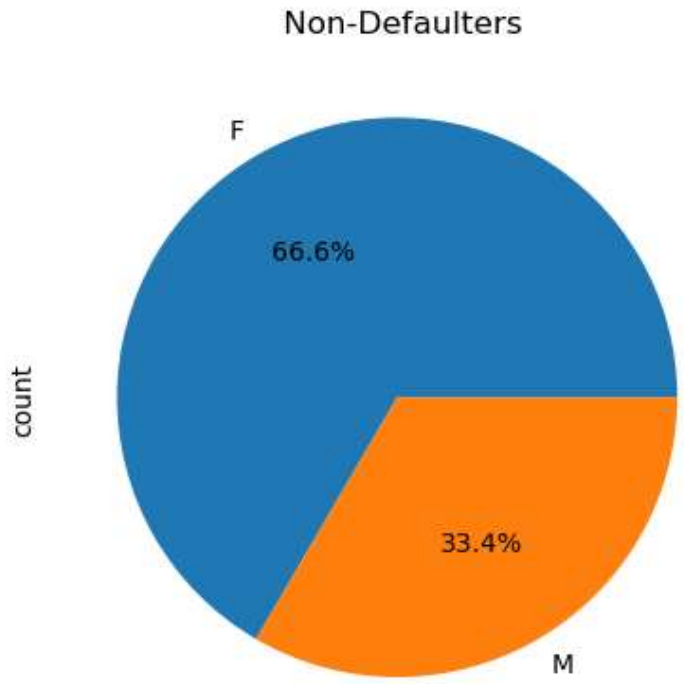
Income type

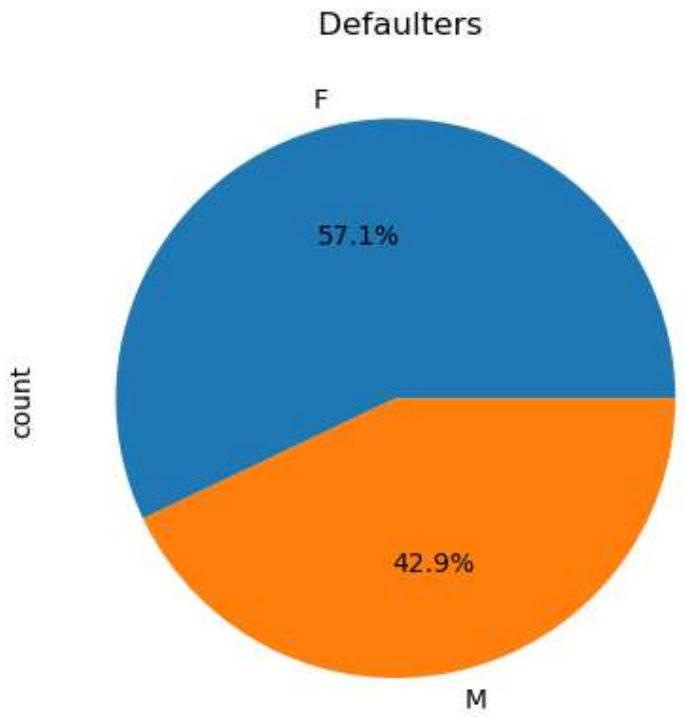




inference: highest number is seen in working
there is negligible number for unemployed and maternity leave

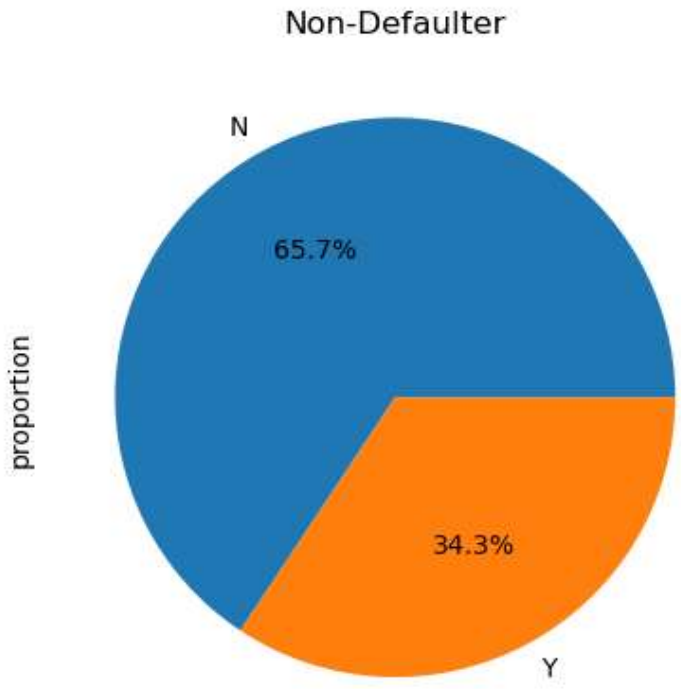
code gender

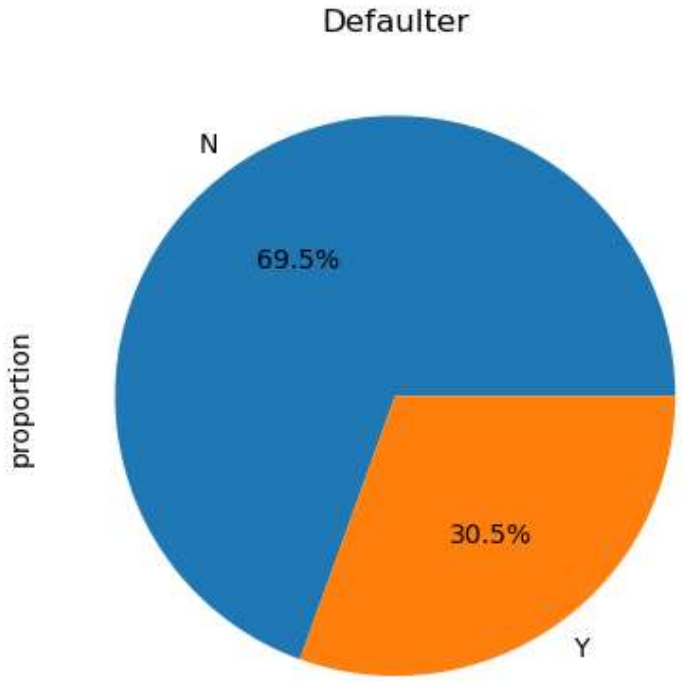




inference 66.6% is seen females in non defaulters and 57.1% seen in defaulters it is seen that there is more percent of females than males in defaulters application.

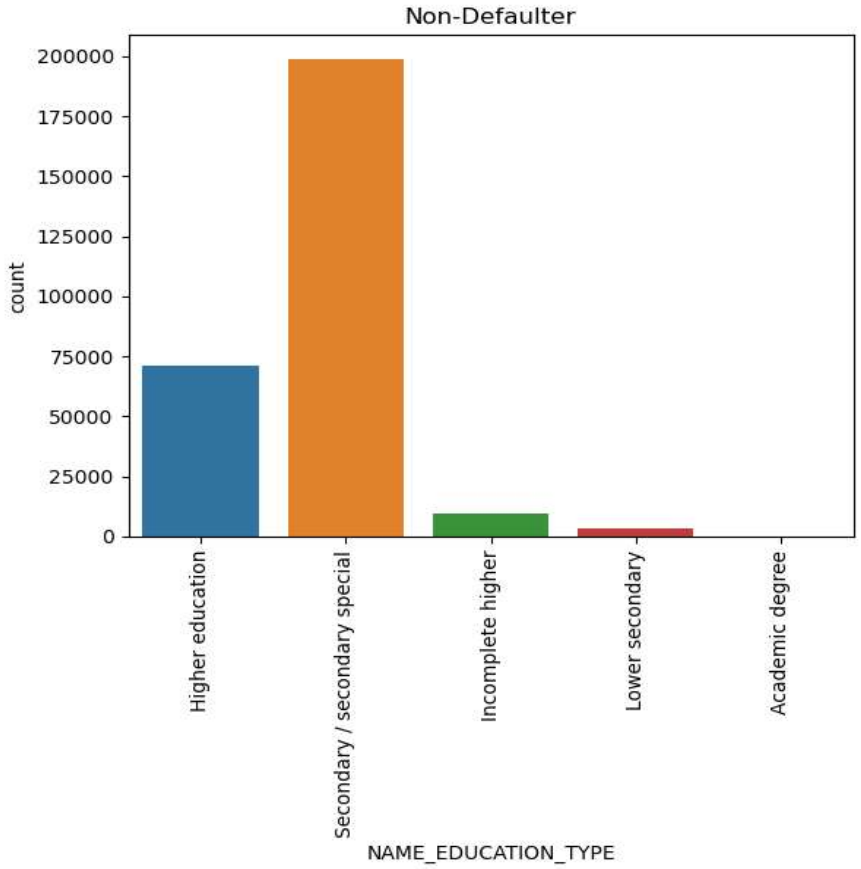
flag own car

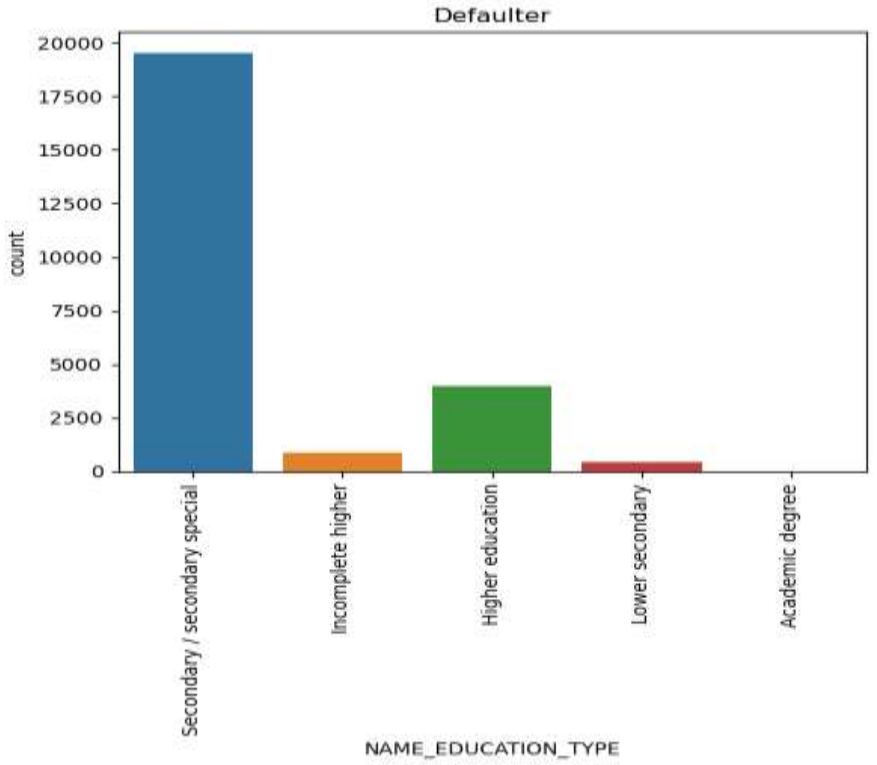




inference: more percentage of the applicants do not own a car. people who own cars is low compared to those who dont own cars.

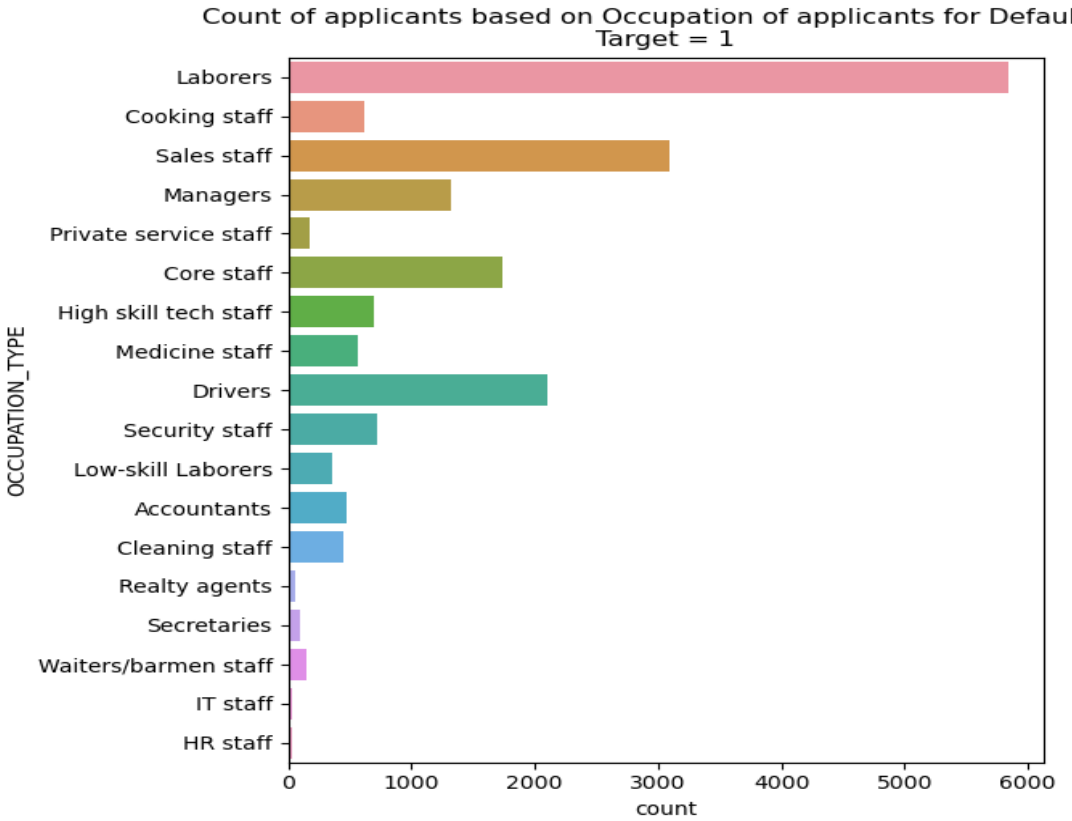
education type

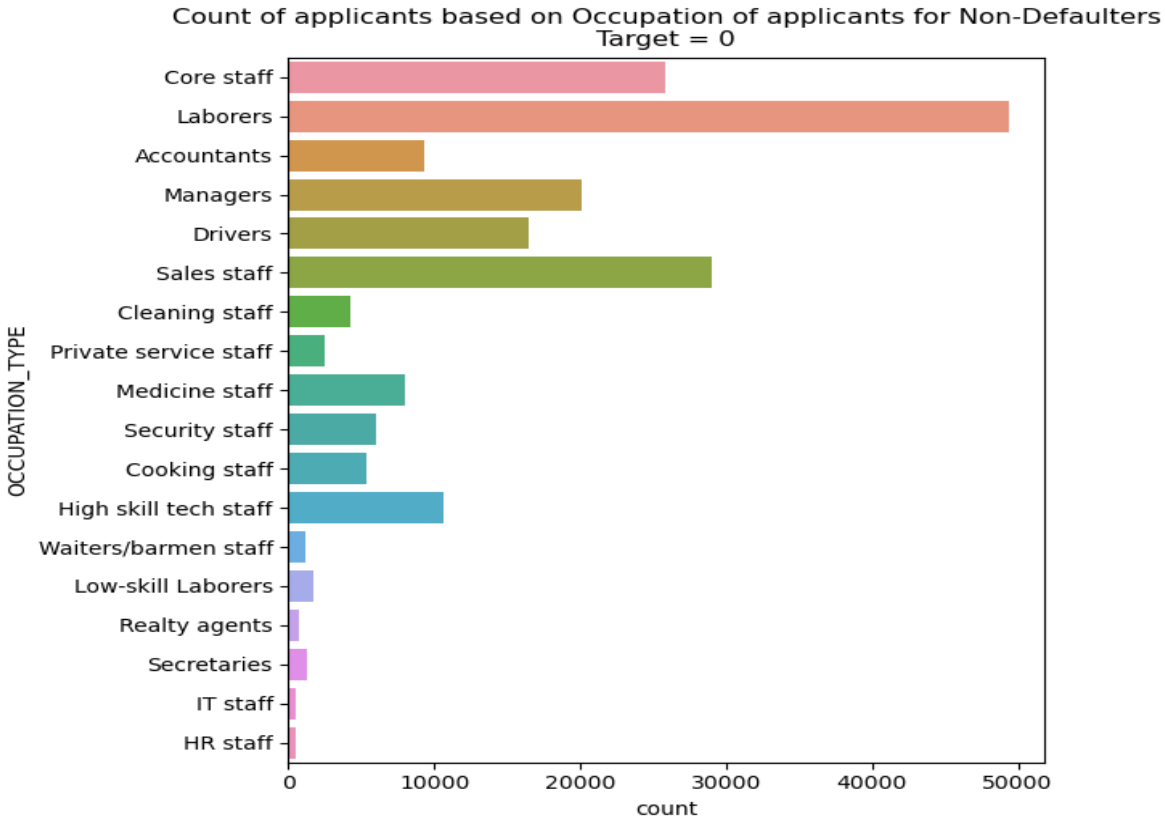




inference: academic degree shows no number in both categories. majority of applicants in both categories, have completed Secondary Education. then Higher Education is complete in both defaulter and non-defaulter categories

occupation type

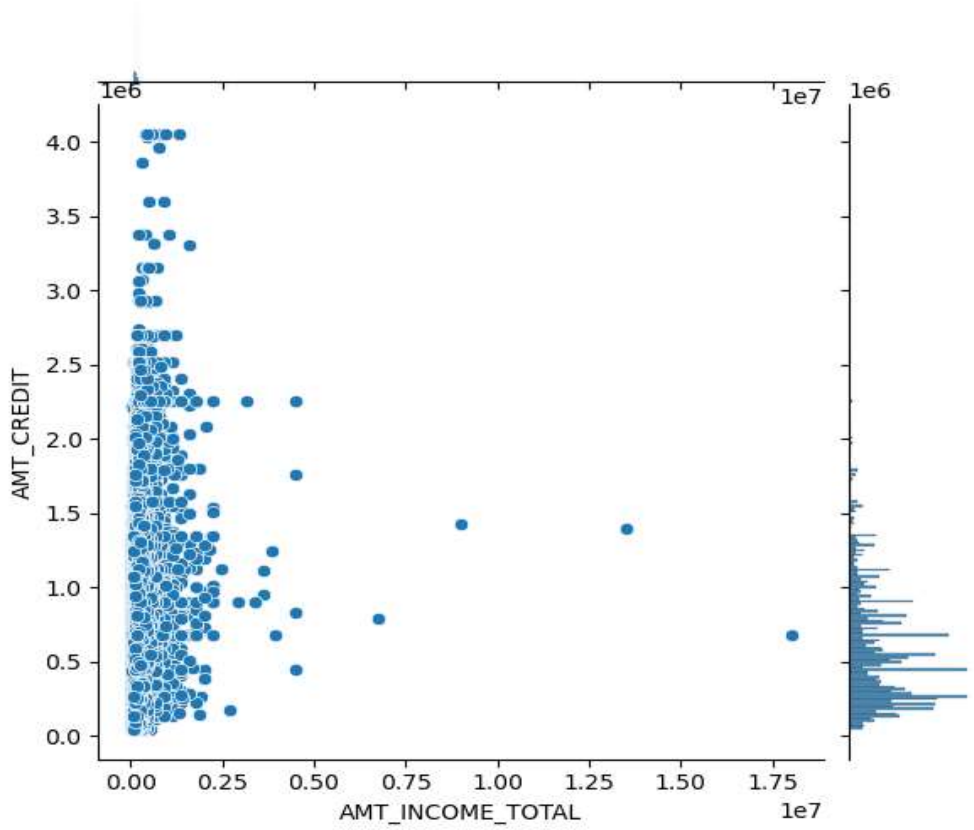


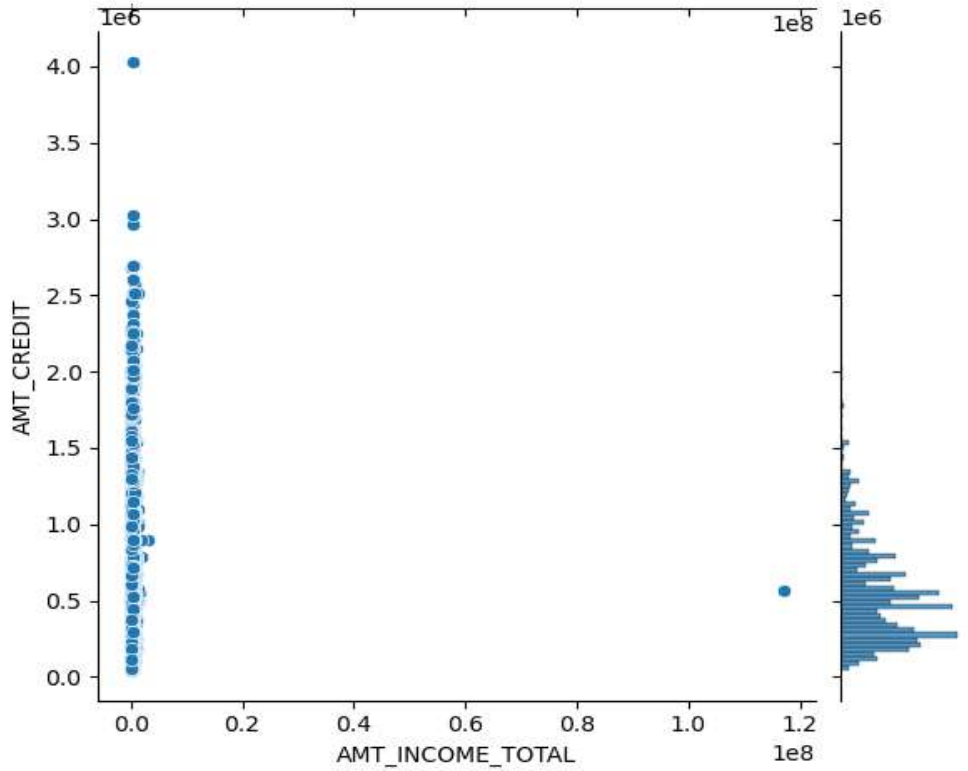


inference: HR staff is seen less in number whereas number of applicants are as Laborers from their occupation.

bivariate

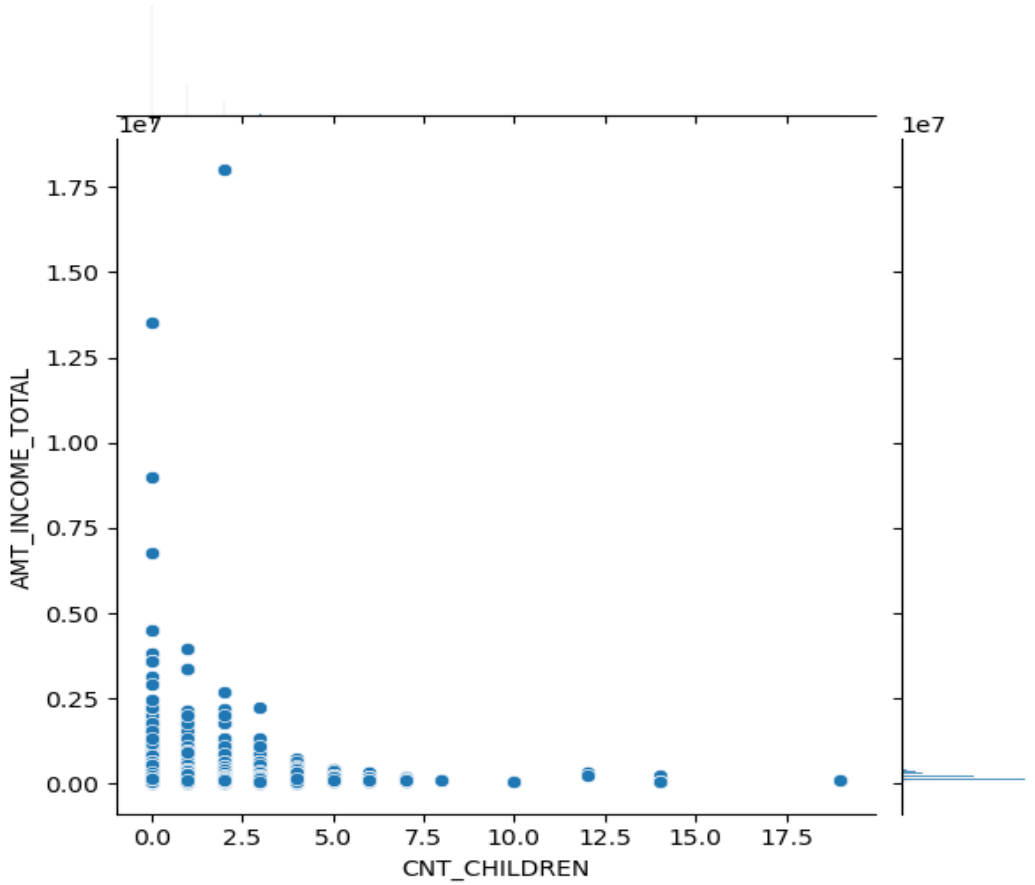
plotting income vs credit for Target 0 and 1

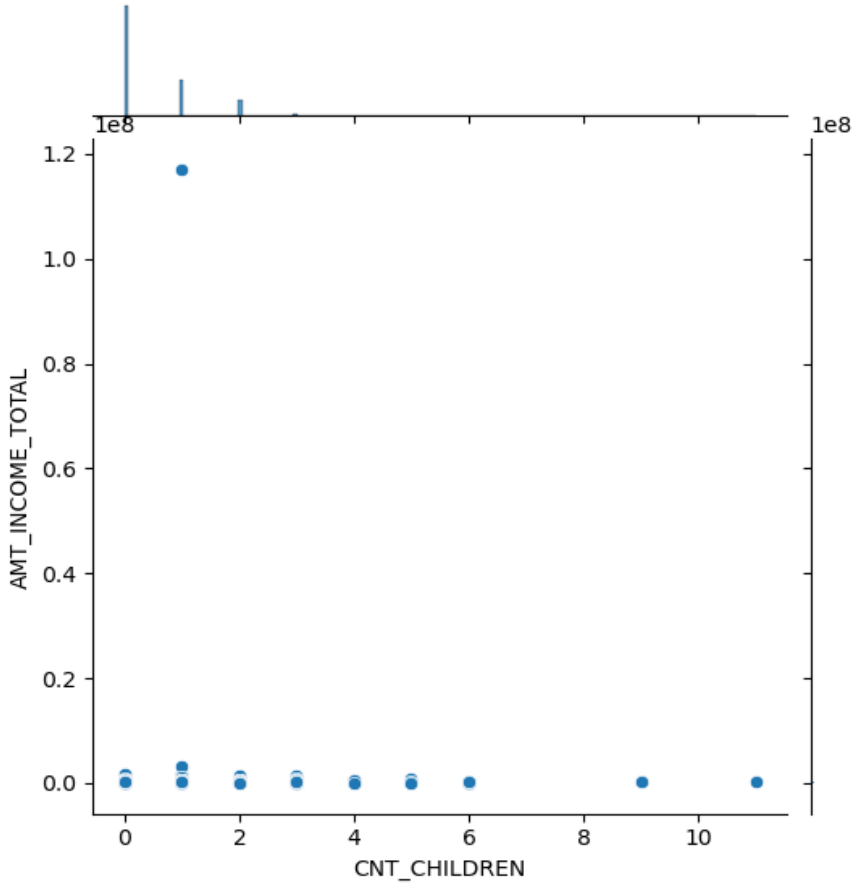




inference: income total increases for credit of defaulters near low value whereas for nondefaulter is goes on decreasing

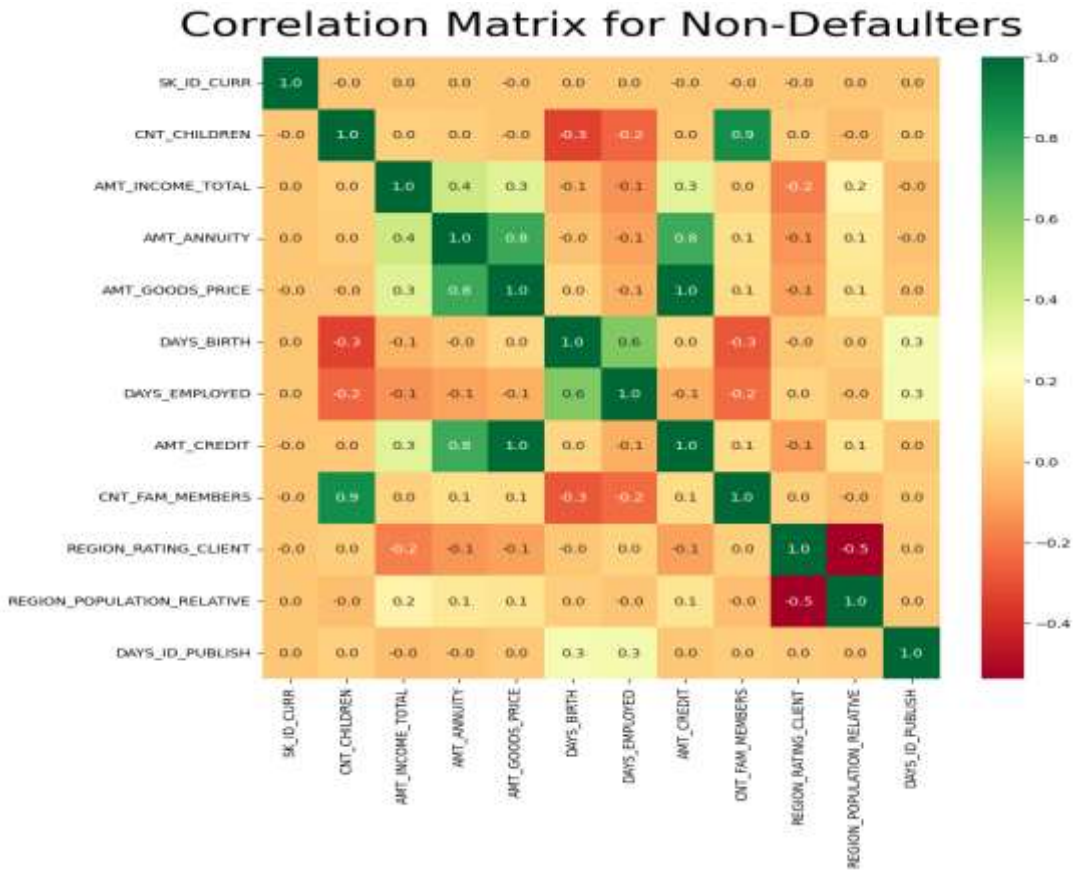
plotting AMT_INCOME_TOTAL vs CNT_CHILDREN for Target 0



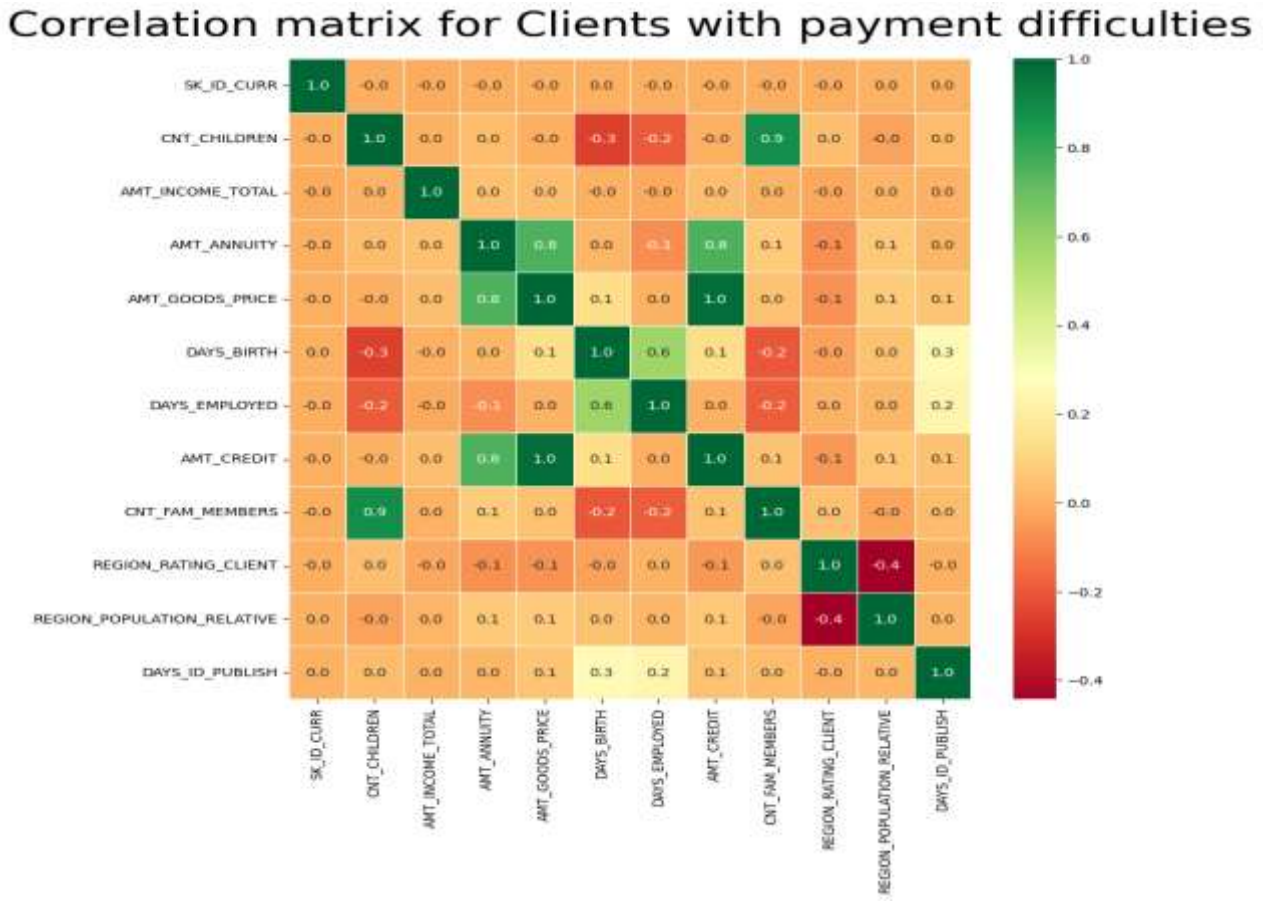


inference: it can be seen that 'CNT_CHILDREN' have a decreasing income total.

correlation



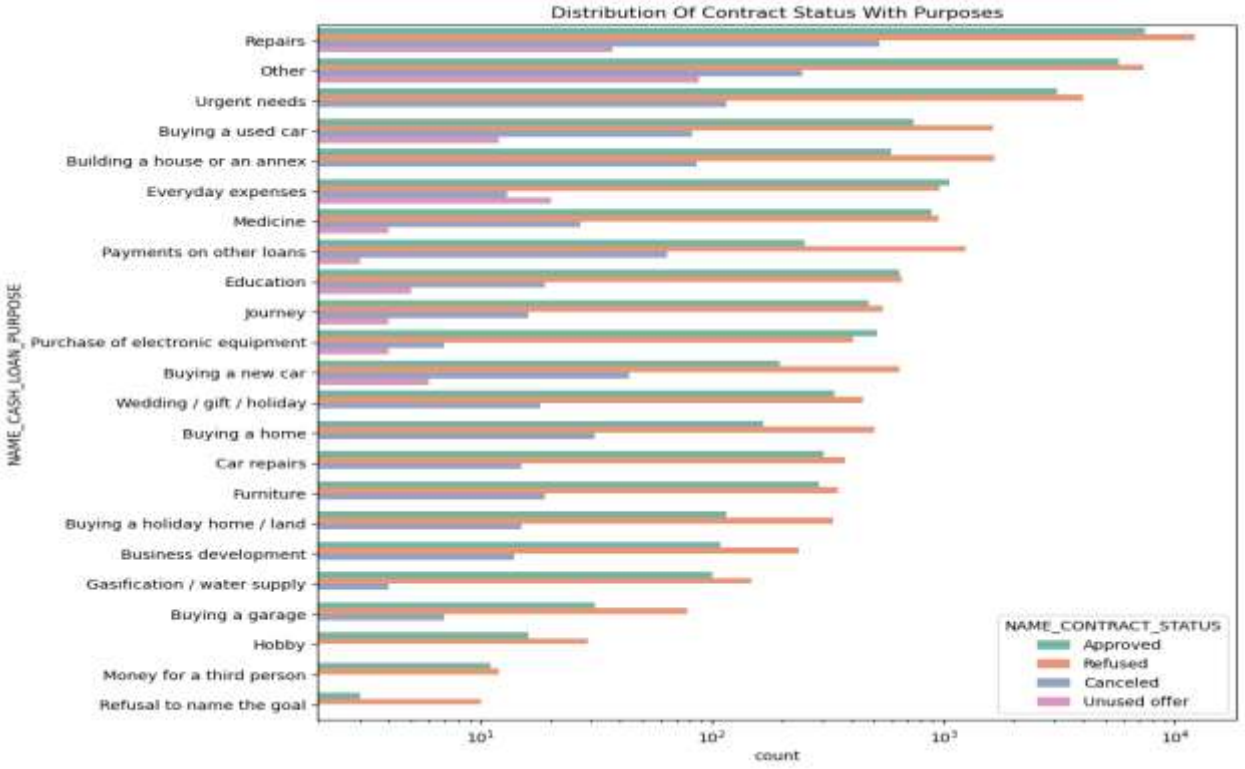
Inference: correlation values are seen hig for Target 0
AMT_ANNUITY and AMT_CREDIT
CNT_FAM_MEMBERS and CNT_CHILDREN
AMT_ANNUITY and AMT_INCOME_TOTAL



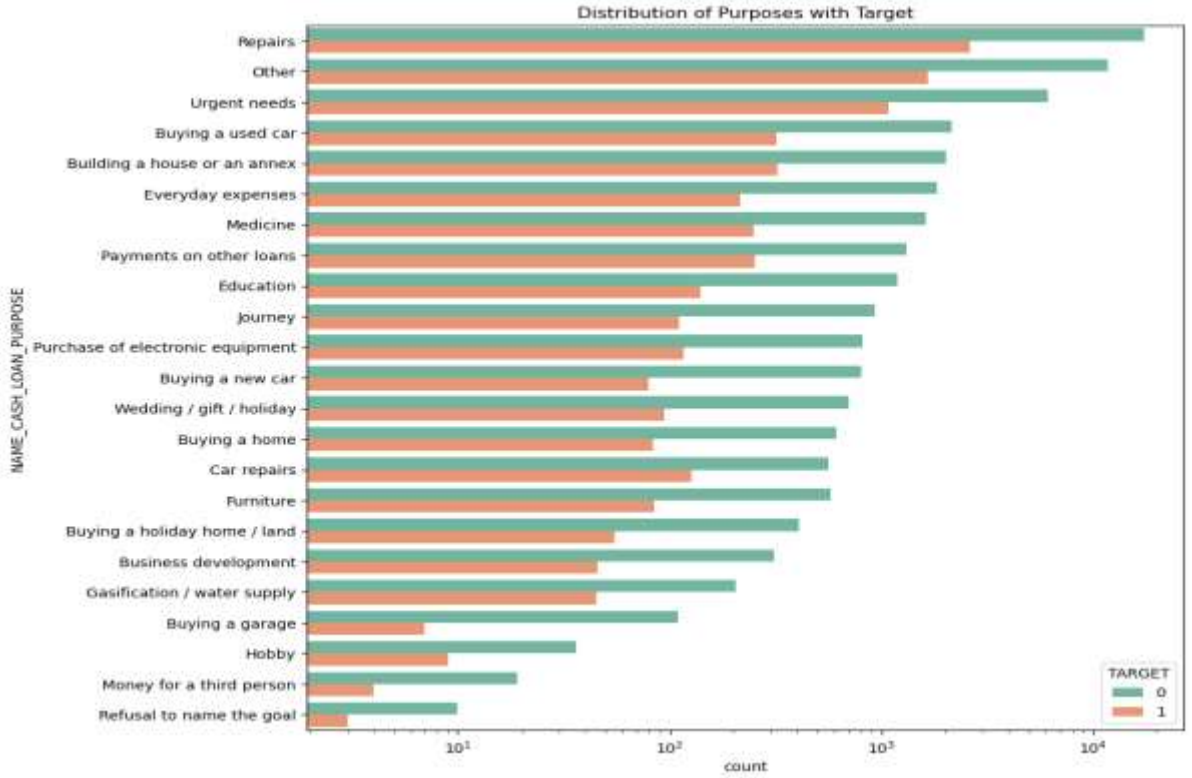
Inference: high correlation values seen for target 0 and 1 in
AMT_GOODS_PRICE and AMT_CREDIT and
CNT_FAM_MEMBERS and CNT_CHILDREN AMT_INCOME_TOTAL and AMT_GOODS_PRICE

univariate

Distribution of contract status in logarithmic scale



Distribution of purposes with target

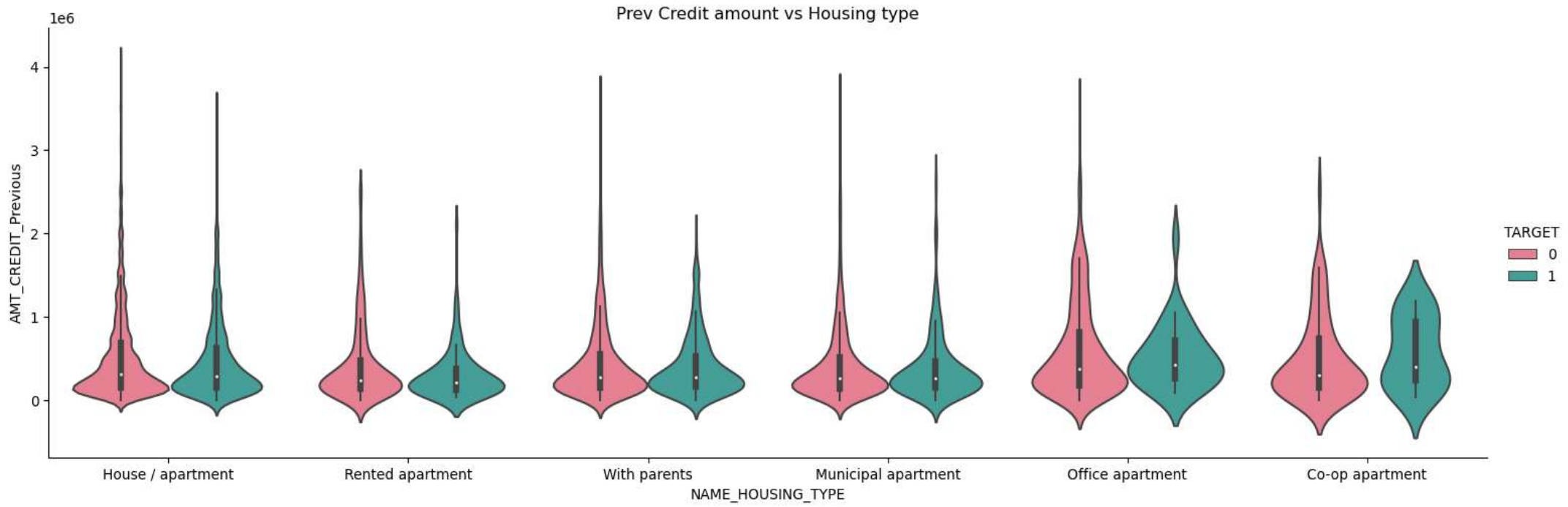


inference: The purpose of 'Repairs' shows a high number of loan rejections. reduction of reduced is seen near journey ,education and purchase of electronic equipments

inference timely payment can be seen from refusal to name the goal and more time is taken by repairs.

bivariate

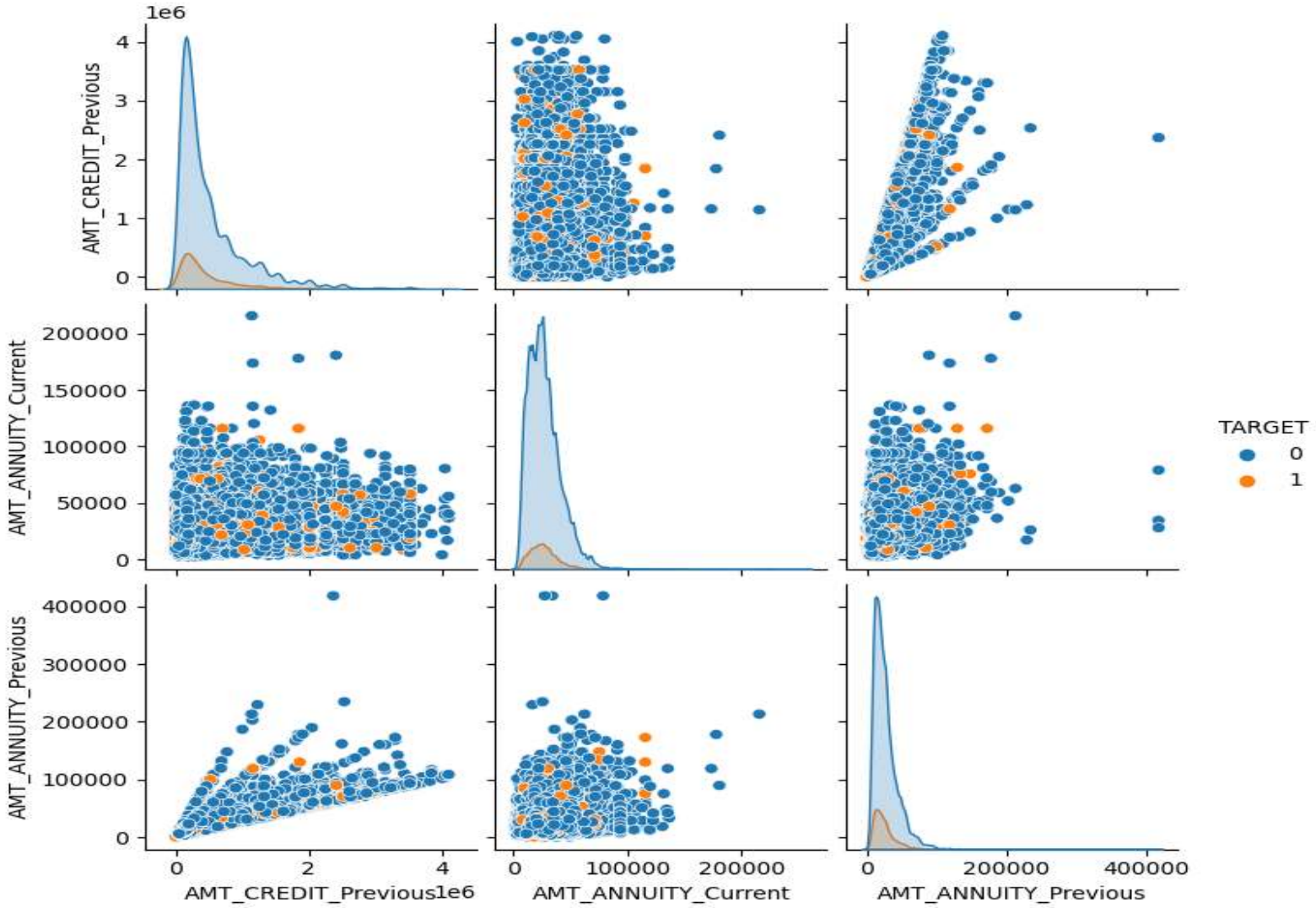
Box plot for Credit amount prev vs Housing type in logarithmic scale



inference: there can be seen a certain difficulty in payment with the housing of co-op apartment so banks must not give loans.

correlation

inference: in 'AMT_CREDIT_PREVIOUS'there is a rise in target 0 and less amount of target 1 for 'AMT_ANNUITY_CURRENT' TARGET 0 contains significant large amount than target 1 in 'AMT_ANNUITY_PREVIOUS' target 1 is less as compared to target 0



Conclusion

Recommended group

- Individuals employed as state employees.
- Seniors from diverse economic backgrounds.
- Clients belonging to the high-income bracket.
- Elderly women clients.
- Female clients with advanced education.
- A widow with a history of untapped loan opportunities.