EDA CASE STUDY

CREDIT RISK ANALYSIS

Aniket Pande
Muskan Agrawal

PROBLEM STATEMENT

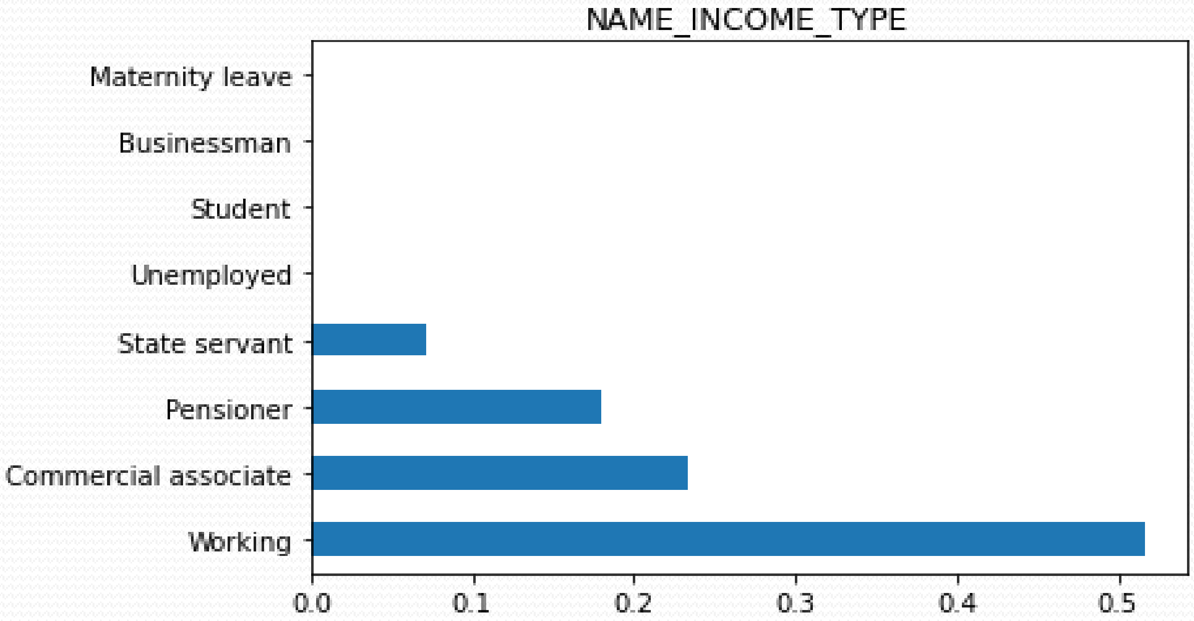
- This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.
- This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

Steps To Follow

- Read the data files and import by using pandas.
- Check for the missing data and try to impute them with Mean or median or mode values.
- Check for the data quality issues and start binning into groups for easy analysis.
- Check for data imbalance for univariate, bivariate analysis and correlation.
- Now merge the application data with previous data.
- Do data analysis univariate, bivariate analysis and correlation.
- Making inferences by using data.
- Categorizing applicants applicable for loan and risks.

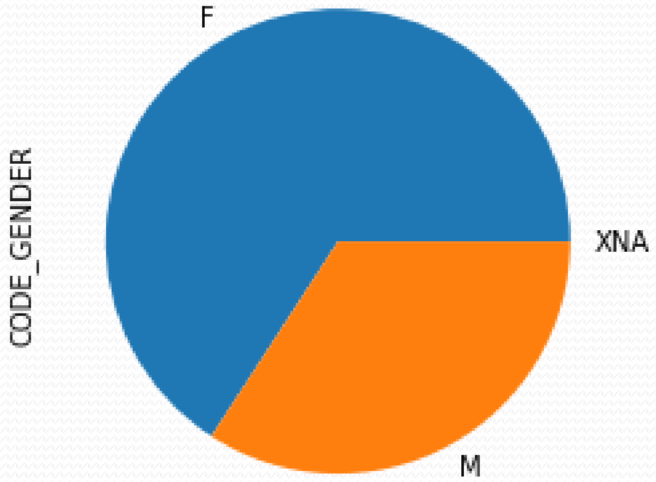
Categorical Unordered Univariate Analysis

- Unordered data do not have the notion of high-low, more-less etc.



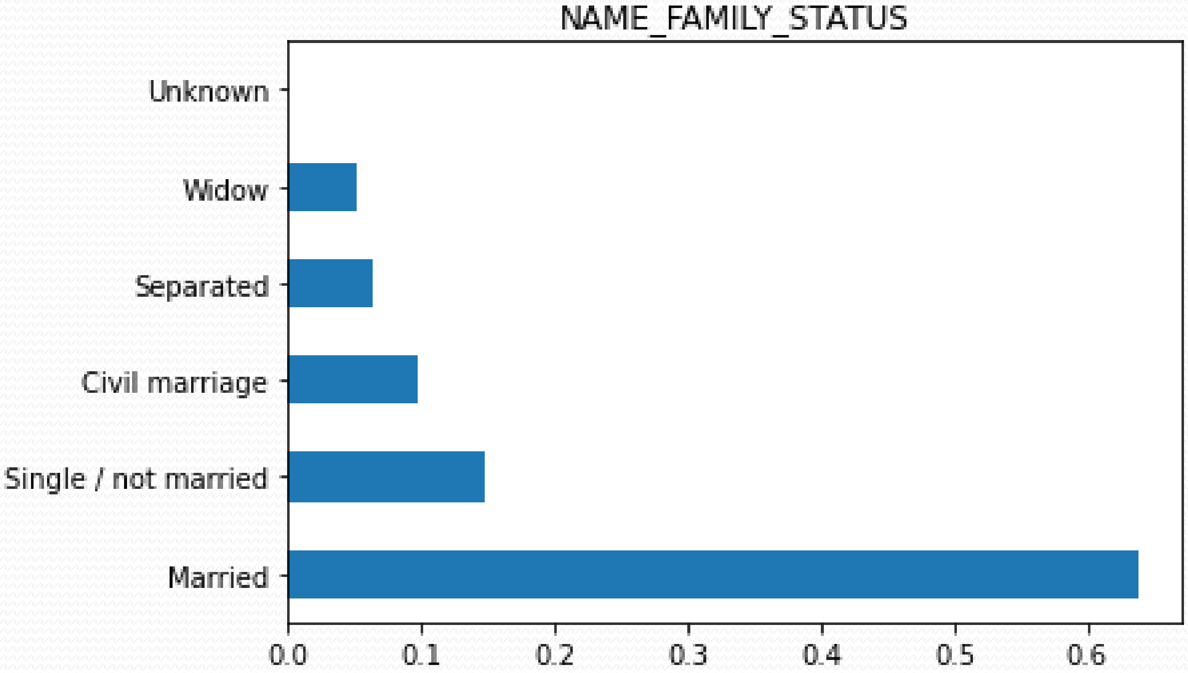
- **Inference: Most clients are working professionals with working income and least are on income of maternity leave.**

Pie Chart of CODE_GENDER Categories



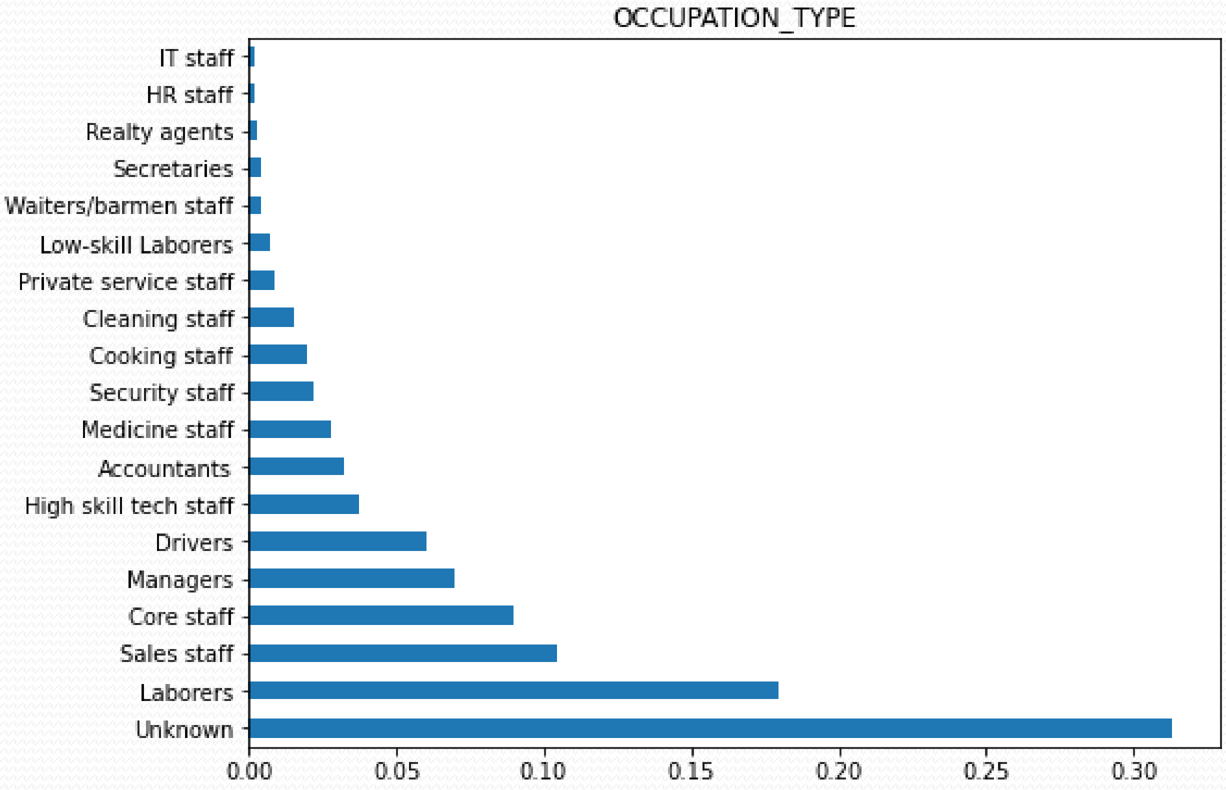
- **Inference: Female clients are more than male clients but still there are least clients with income type of 'maternity leave'. Therefore we can say that when any family is expecting they avoid applying for loans as they have different financial commitments due to the coming new member of the family.**

Bar Graph of Percentage Name_Family_Status Categories



- **We can infer that married people apply for loan more as compared to single people**

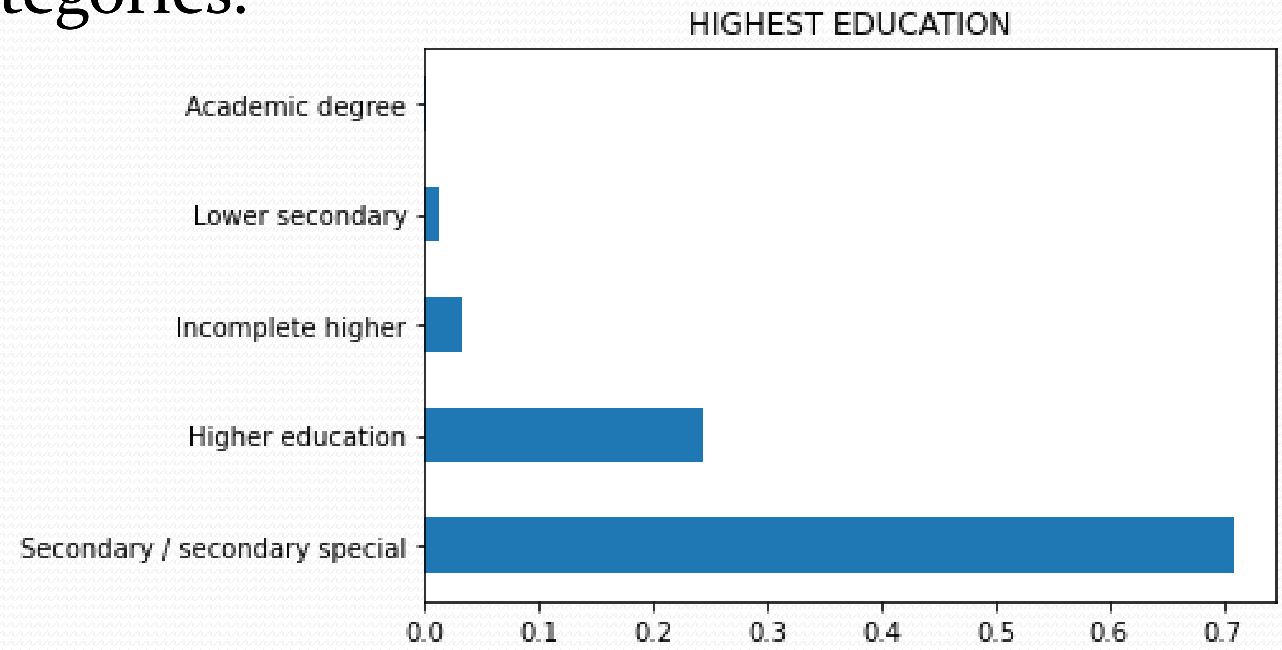
Bar Graph of Percentage Occupation_Type Categories



- **Inference: Maximum clients are of labourer occupation and minimum is IT staff.**

Categorical Ordered Univariate Analysis

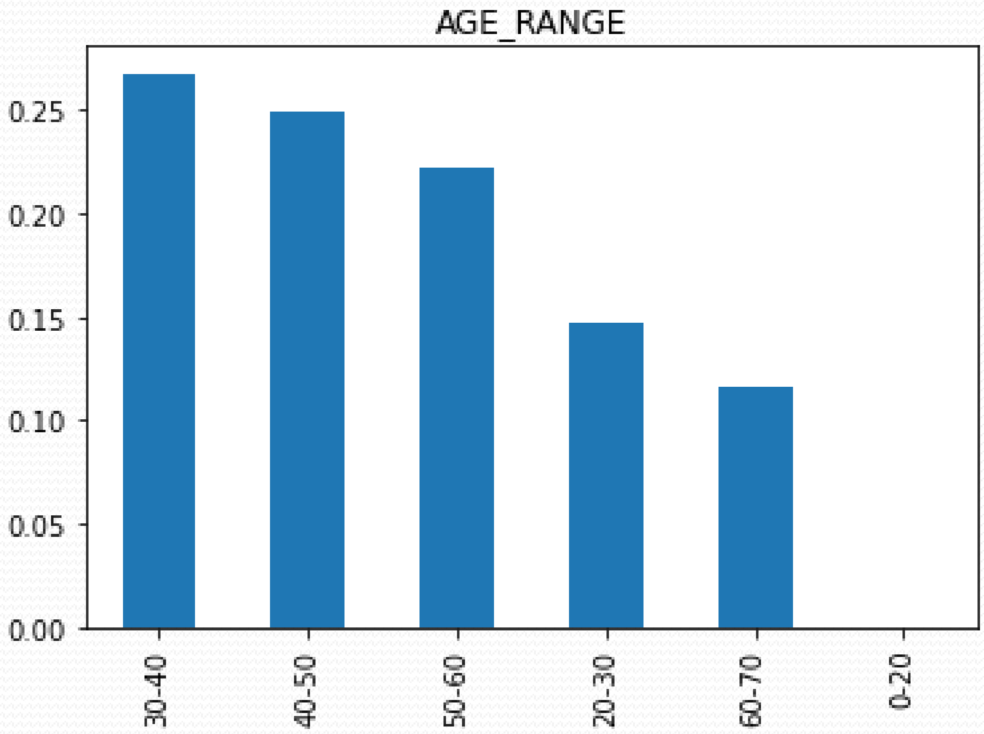
- Bar graph of Percentage Name_Education_Type Categories.



- **Inference: Most of the clients are educated till secondary level and clients with academic degree are the least**

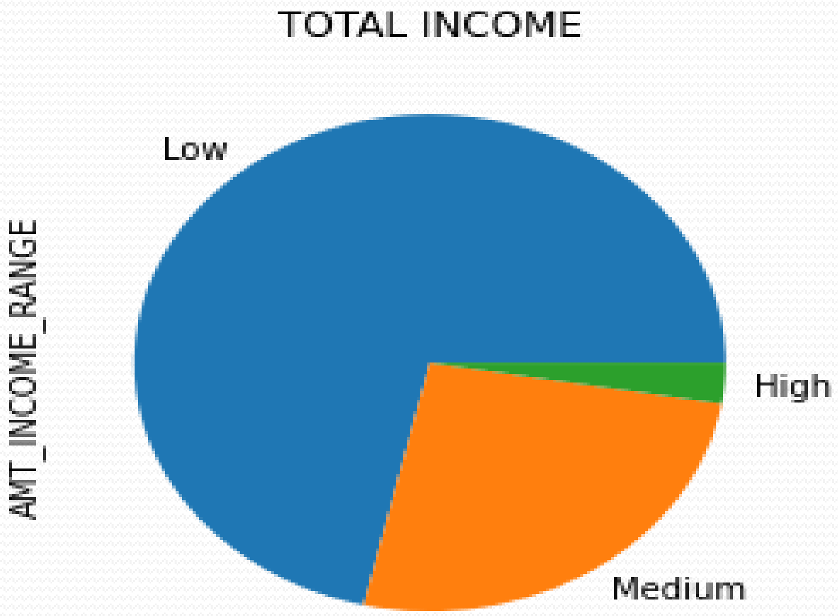
Numerical Univariate Analysis

- **AGE**



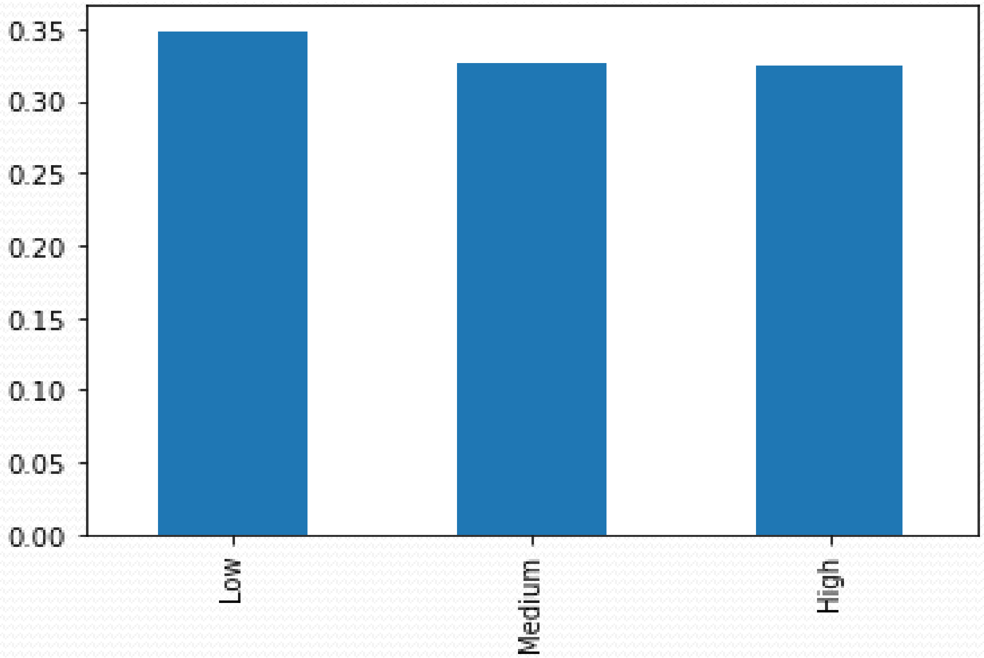
- **Inference: Maximum clients belongs to 30-40 age group**

Pie Chart of Amt_Income_Range Categories



- **Inference: Most customers belong to low range of total income. As it is understood people with higher income don't need loans that much.**

Bar Graph of Percentage Amt_Credit_RANGE Categories

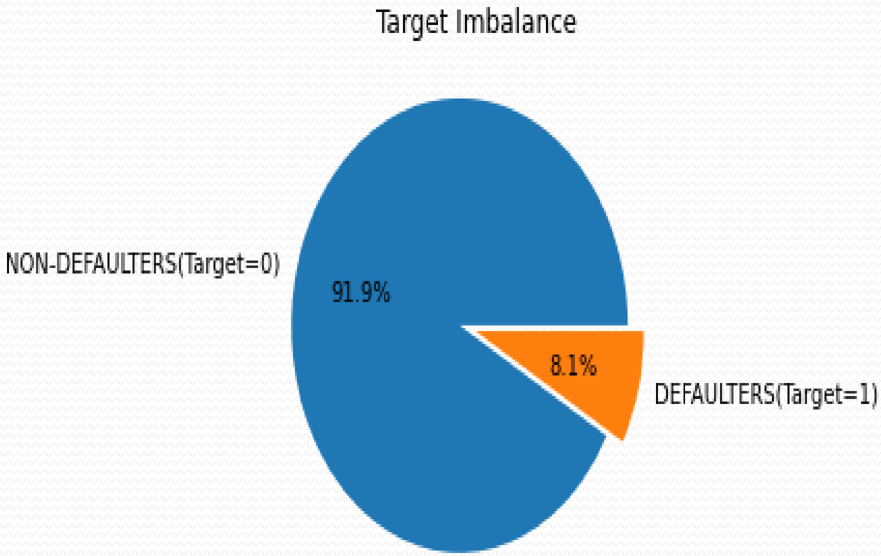


- **Inference: Most of the customers have applied for low amount of credit for loan.**

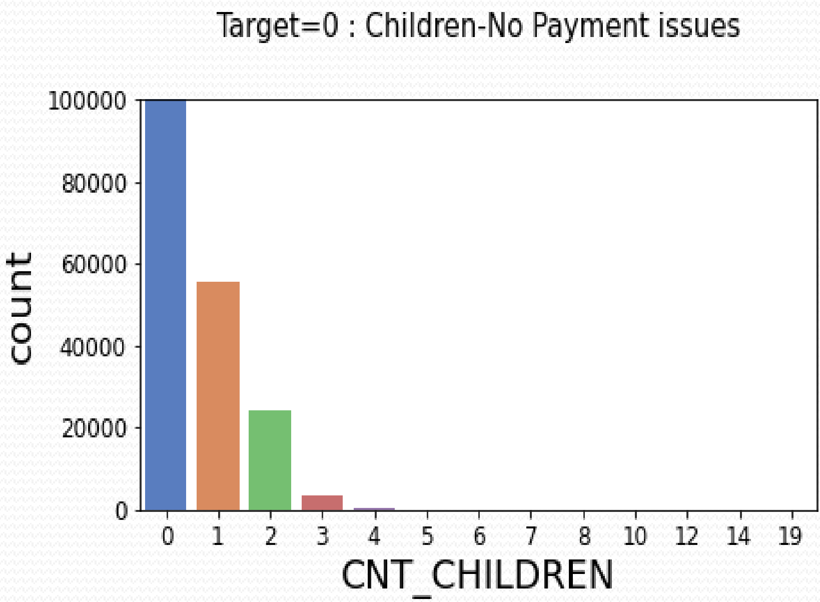
Proportion of Data in Target Column

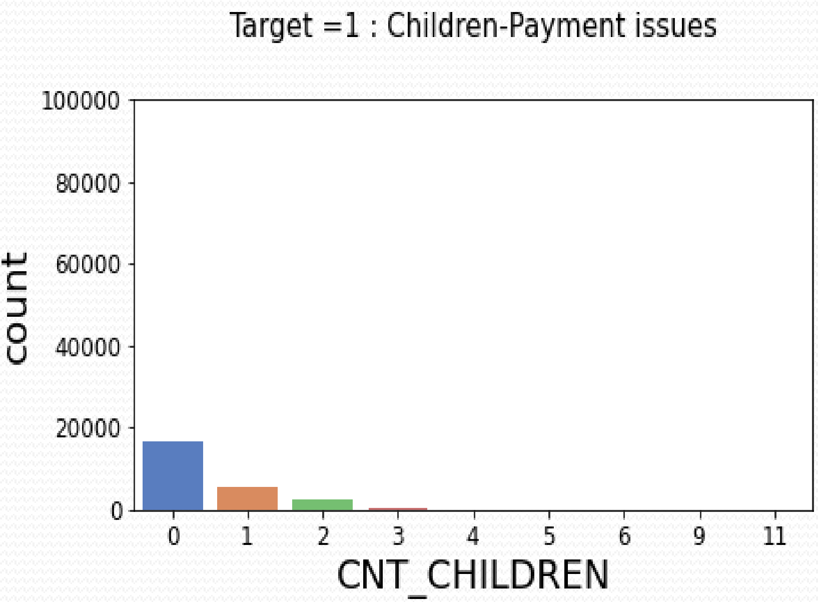
➢ TARGET column has 8.07% of 1's(target_1) which means 8% clients have
payment difficulties.

➢ And remaining 91.93% of 0's(target_0)
which means 91.93% clients doesn't have
any payment difficulties



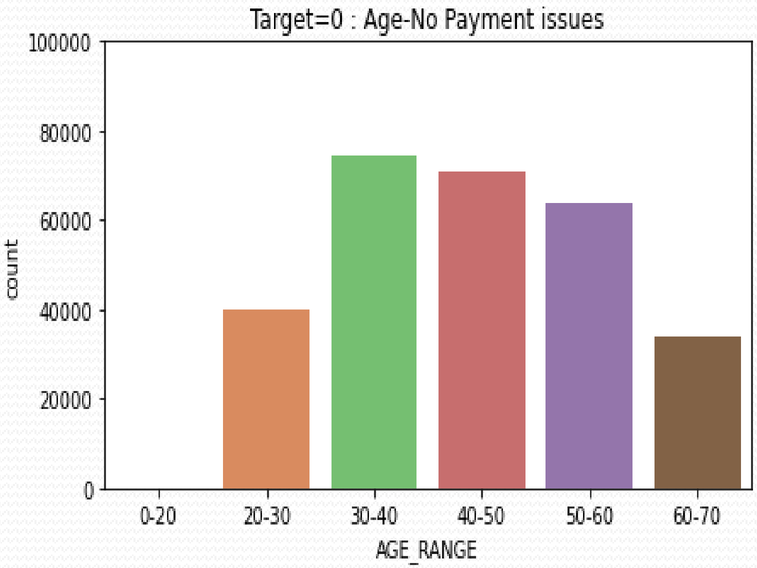
Univariate Analysis For Target Variable



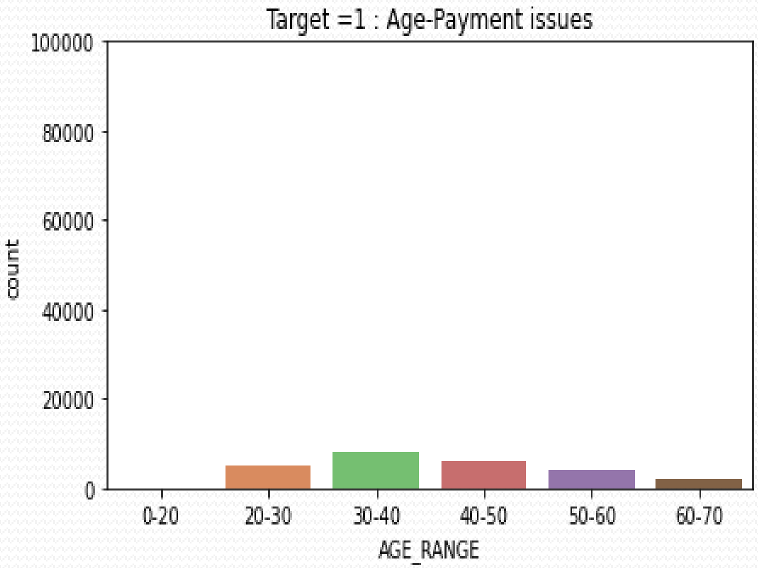


- **Most of the clients have 0 children and very few have more than 3 children. Banks can consider lending loans to people with no children.**

AGE Analysis for Target 0 & Target 1 Data frame



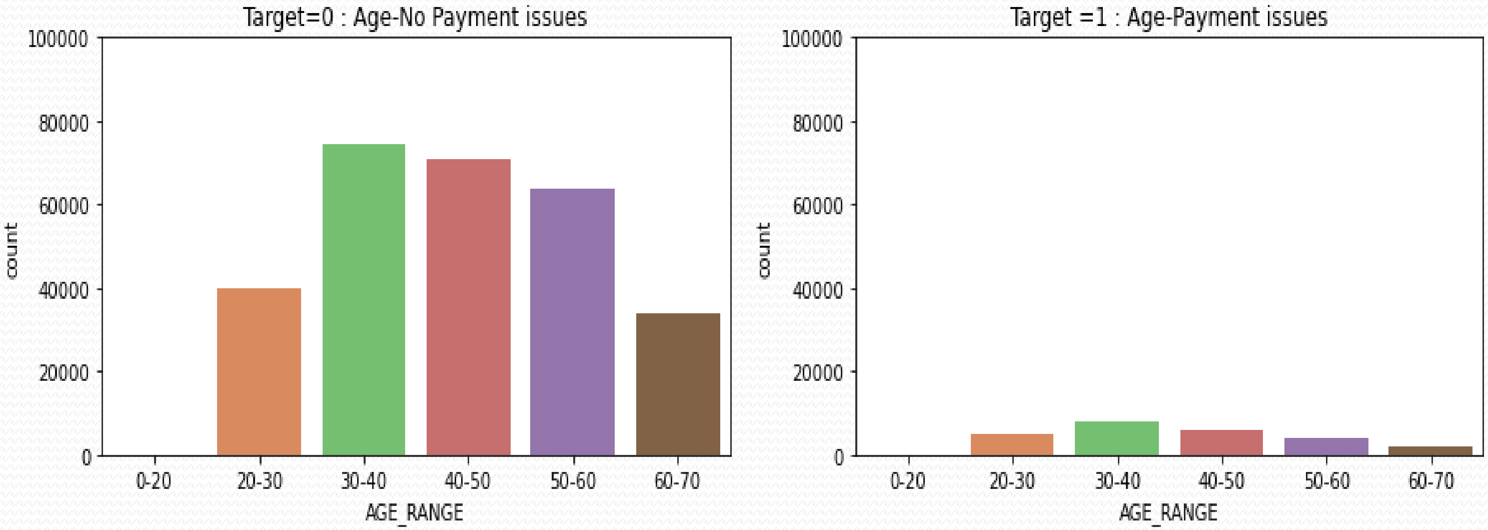
Target=0 : Age-No Payment issues



Target =1 : Age-Payment issues

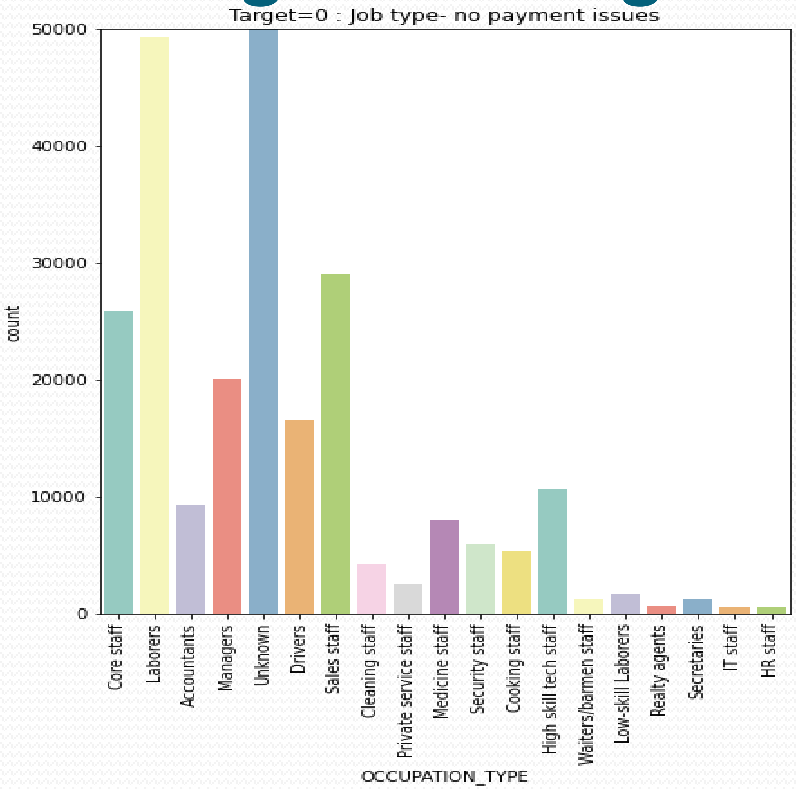
- **As maximum customers belong to 30-40 age group which we observed earlier, we can infer from this graph that customers of age group 50-60 should be given loans as compared to other age groups because they have low probability of defaulting.**

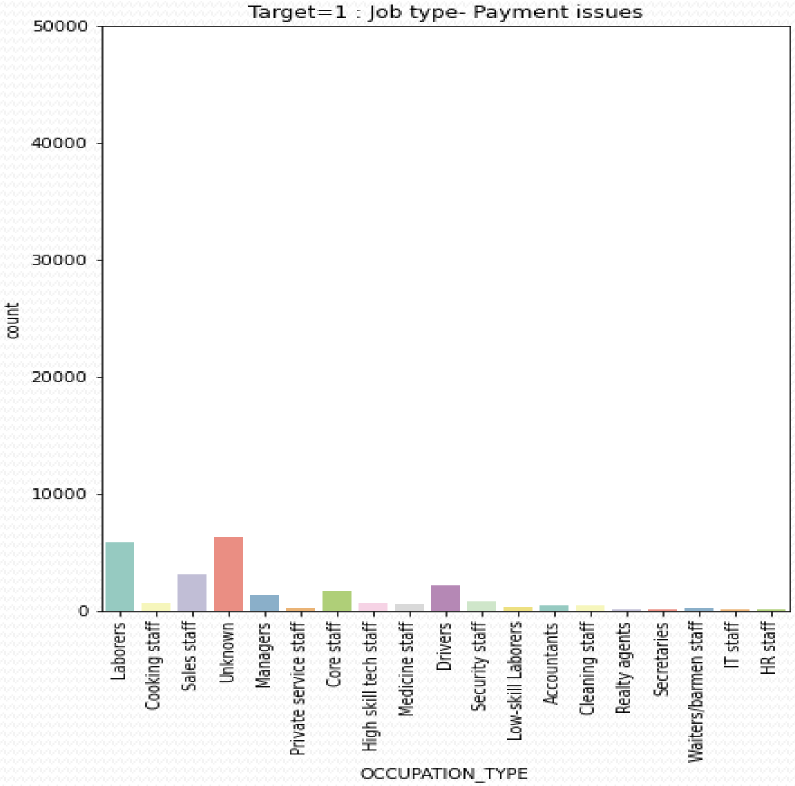
Amount Credit Analysis for Target 0 & Target 1 Data frame



- **We can observe from this plot that low credit amounts are paid back with less issues. So banks can consider them while lending loans.**

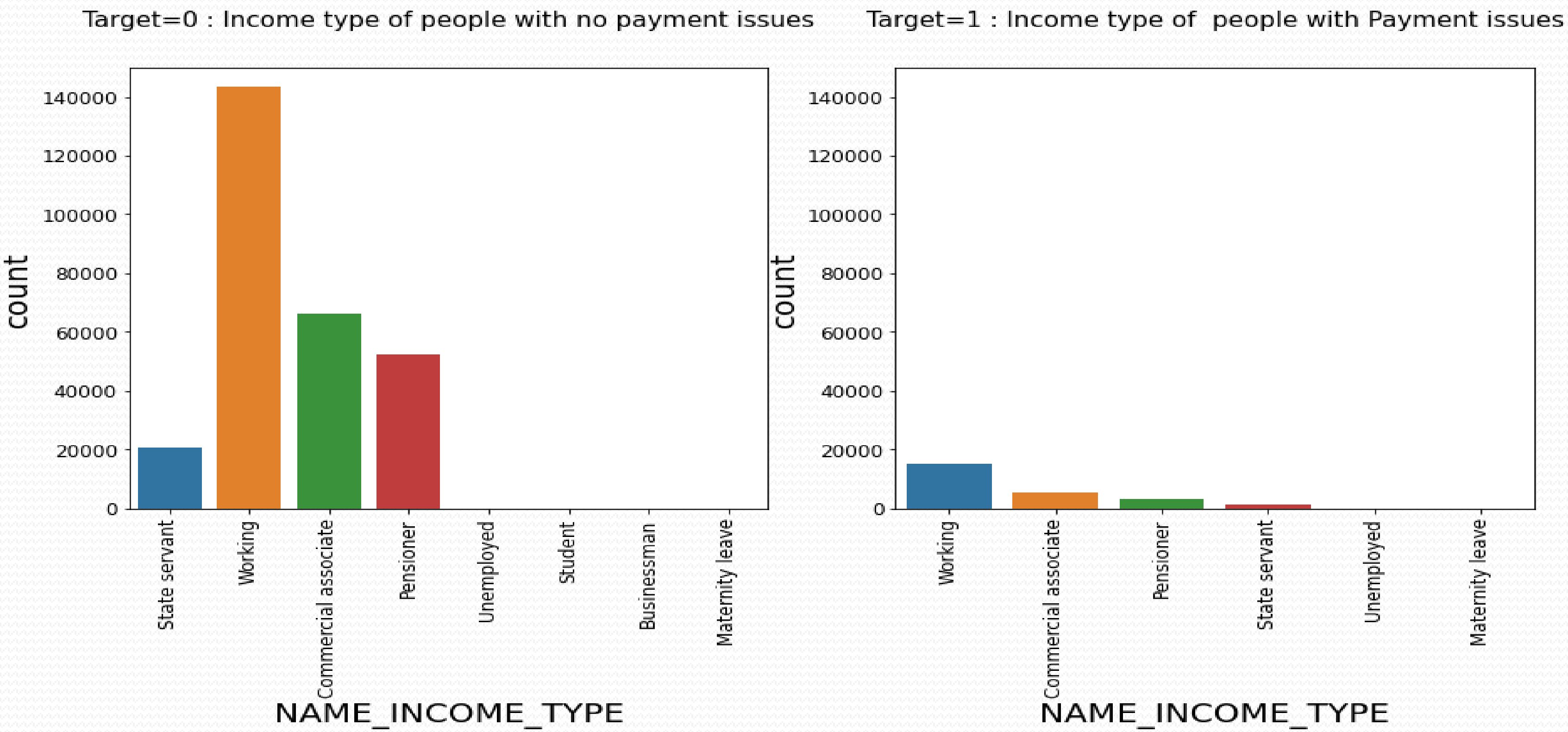
Categorical variable(OCUPATION_TYPE) Analysis for Target 0 & Target 1 data frame





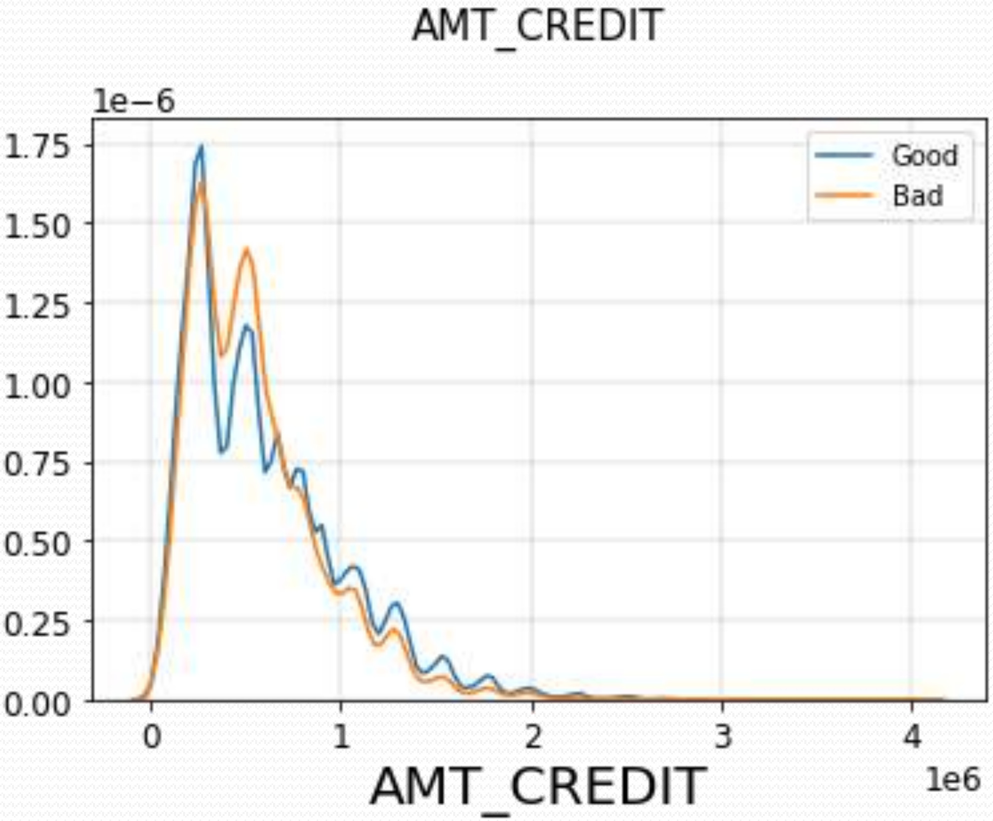
From the above graphs we can observe that count of labourers is high in both payment issue and no payment issue. Even the credit amount of labourers is low. So, we cannot surely predict that labourers cannot be defaulters but they can be given small amount of loan. On the other hand, sales staff and core staff have less percentage of defaulting so, they can be considered.

NAME_INCOME_TYPE analysis for Target 0 & Target 1 data frame

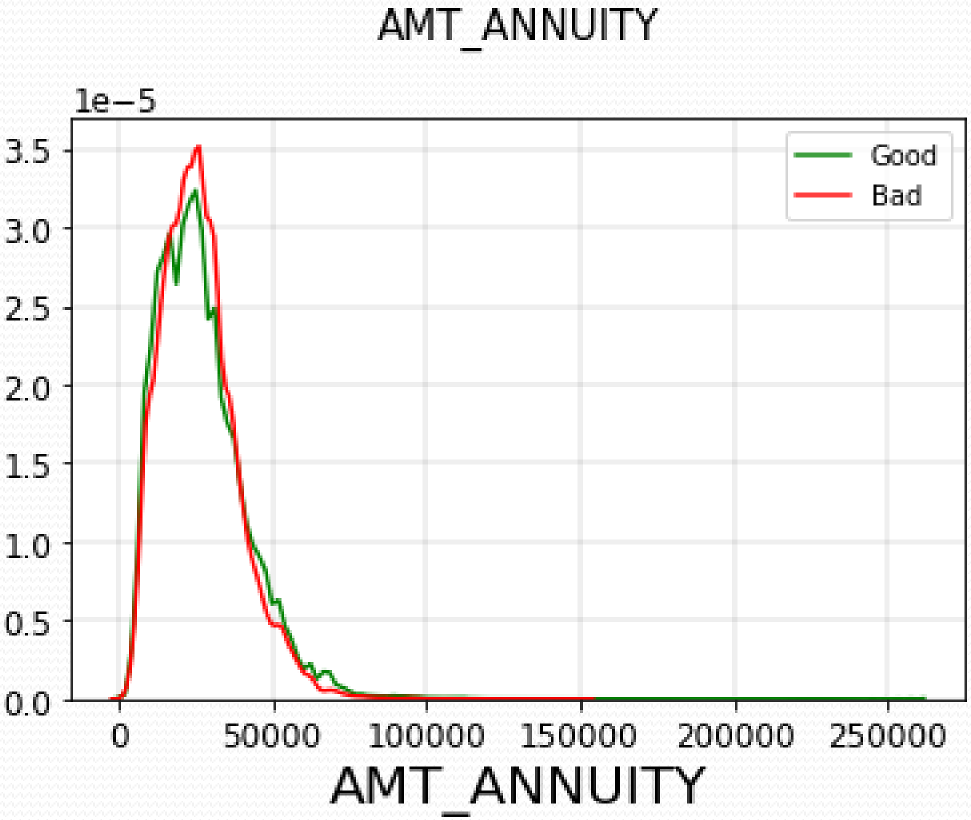


- **We can observe that working professionals and state servants are more likely to pay back the loan on time.**

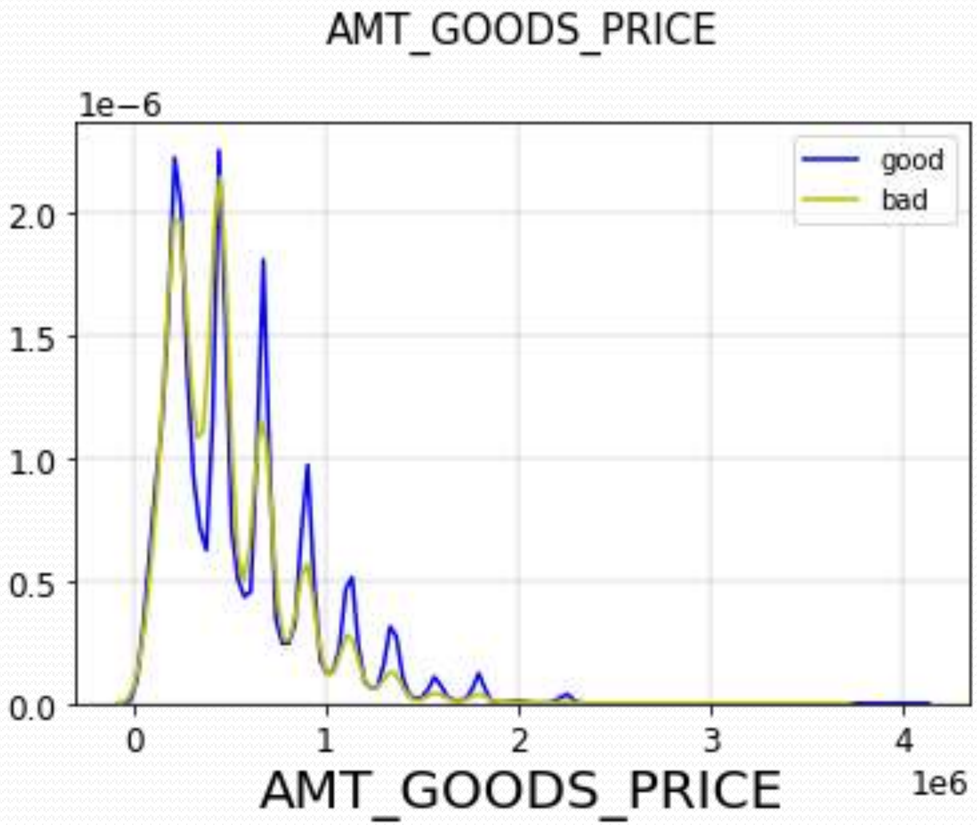
Analysis Of Continuous column with respect to the Target column for Amt_Credit

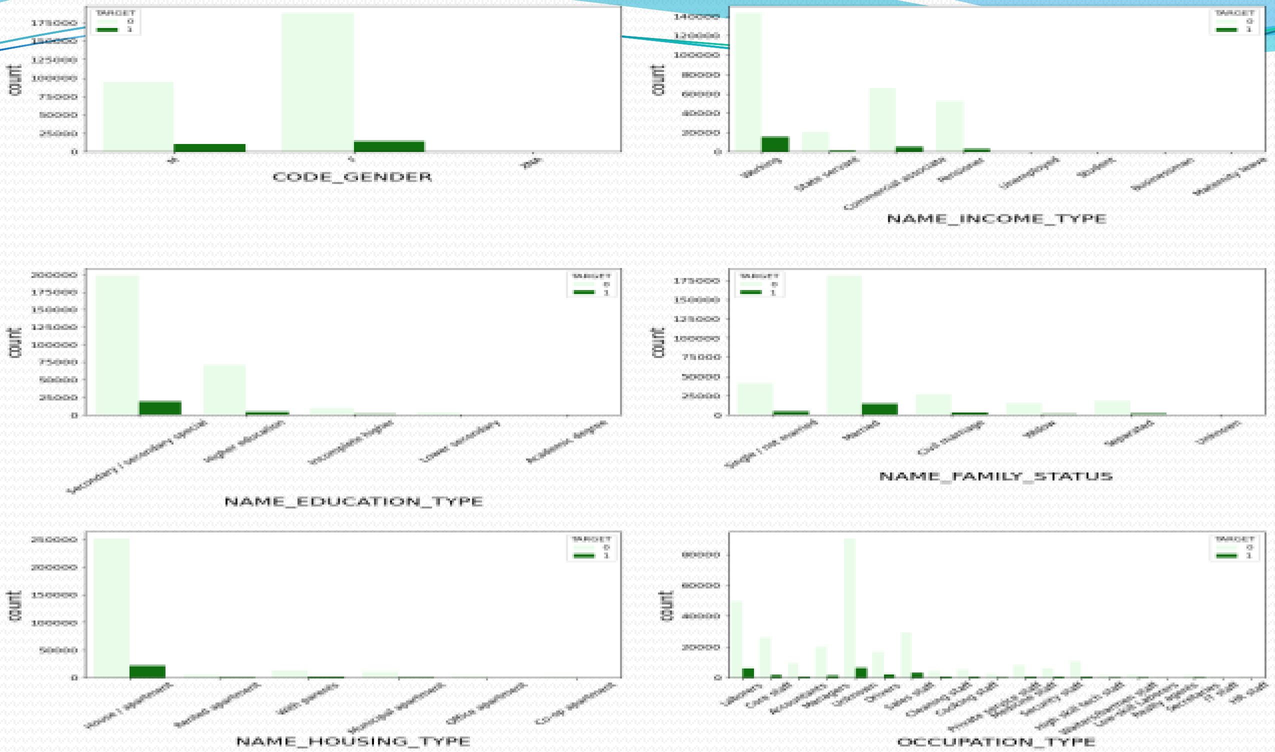


Analysis Of Continuous column with respect to the Target column for Amt_Annuity



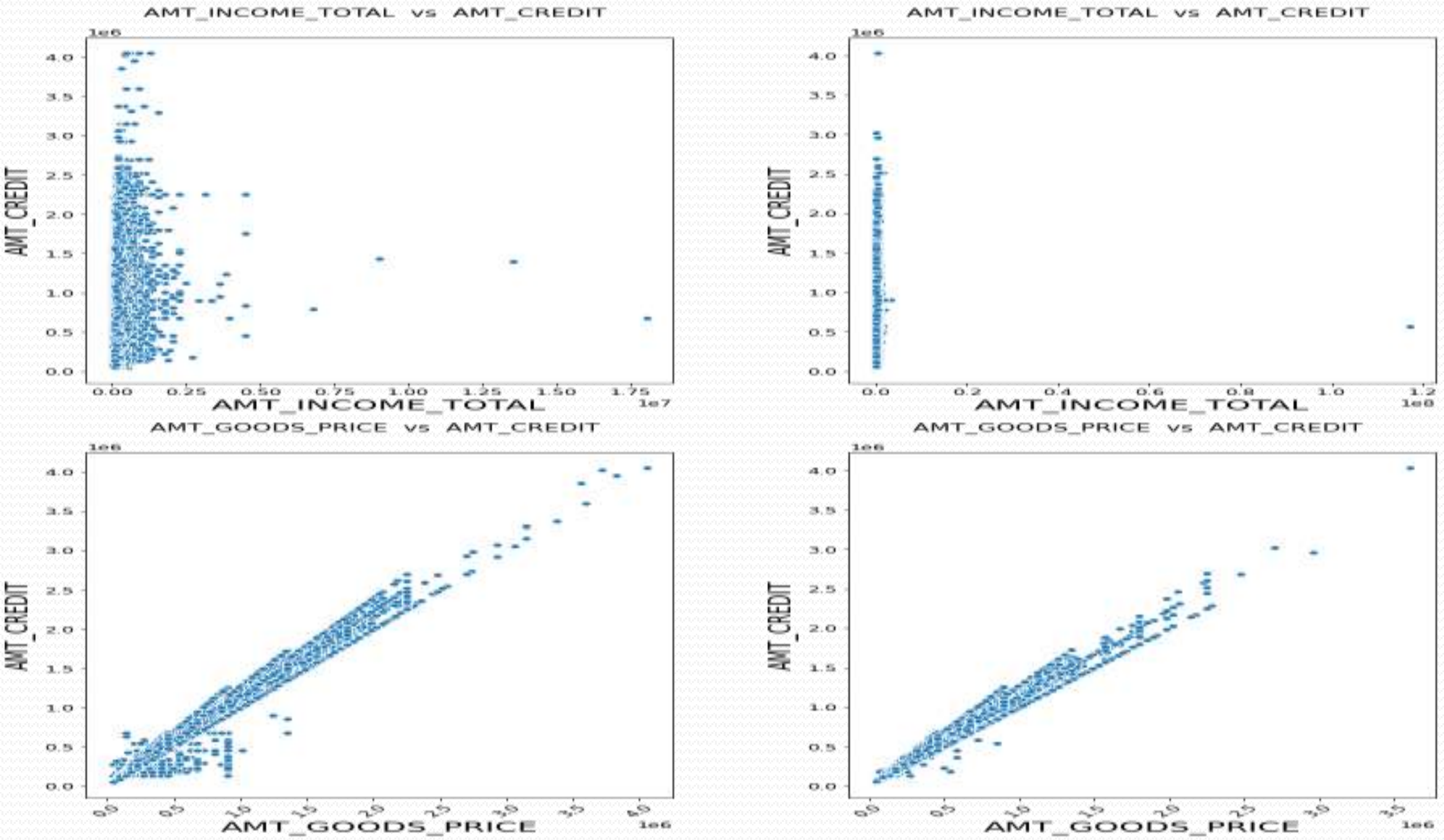
Analysis Of Continuous column with respect to the Target column for Amt_Goods_Price





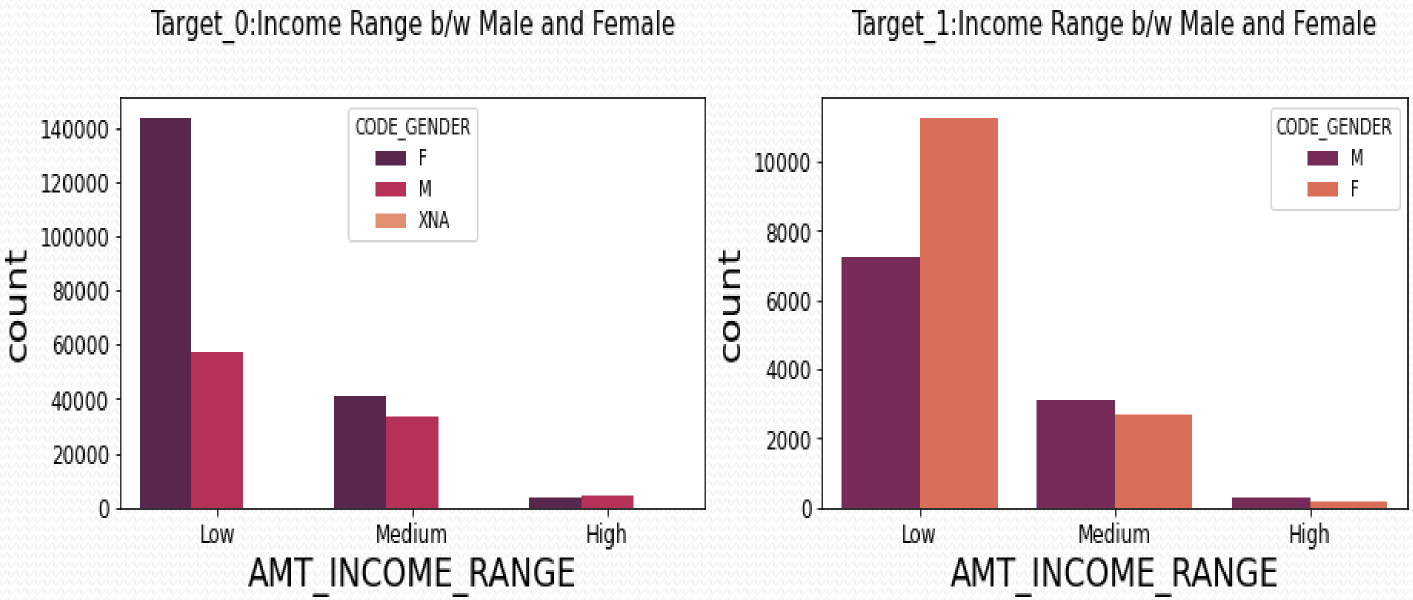
- From plot 1: we can observe that banks can target more female customers for lending loans as they pay loan on time.
- From plot 2: we can observe working/state servant clients can be targeted to lend loans as they've higher percent of giving payments on time.
- From plot 3: we can observe Customers with higher education/ academic are maximum possibly to make payments while as compared to customers with academic degree.
- From plot 4: we can observe Married customers had paid loan on time as compared to widow and separated.
- From plot 5: we can observe customers with own house/apartment are most likely to make payment on time as compared to other customers.
- From plot 6: we can observe sales staff and core staff have high percentage of being a repayer.

Bivariate Analysis for target 0 and target 1



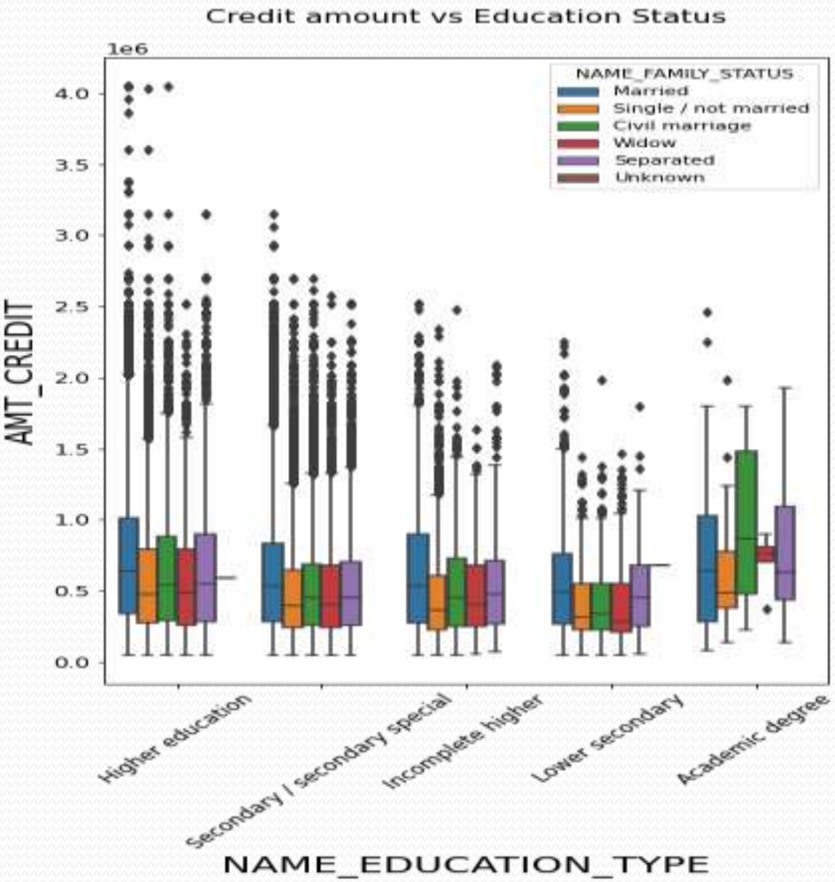
- Those who payback the loan on time can get higher credits. Mostly low income clients have difficulties in paying instalments and therefore they have low credits.
- Goods price and credit have linear relationship. Clients who have paid back on time with higher goods price have higher credit

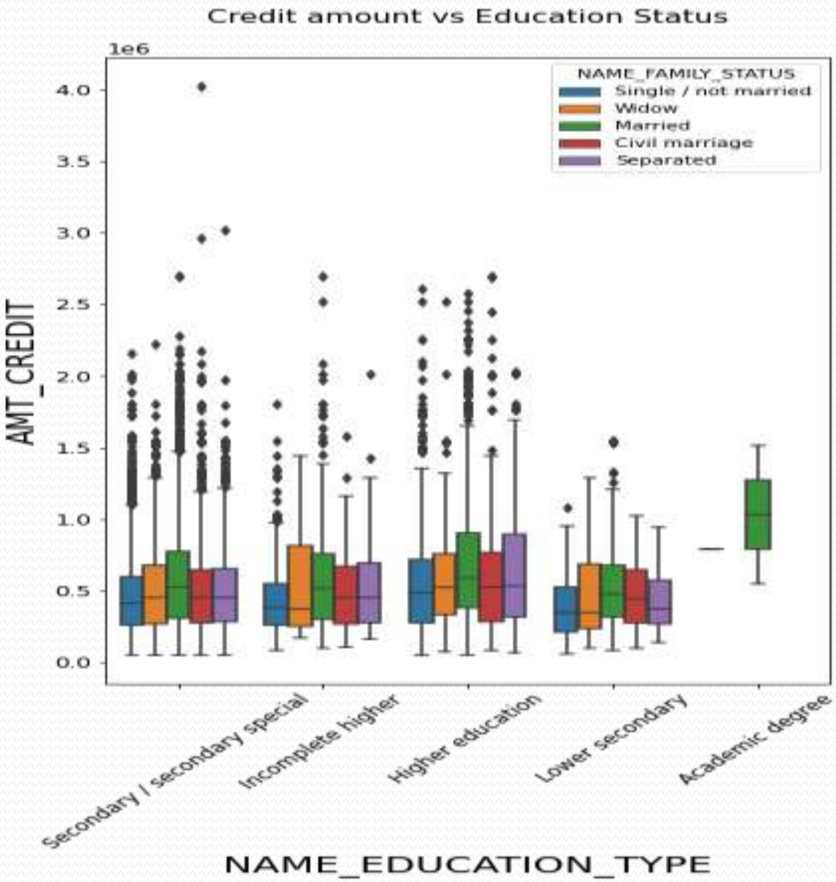
Numerical categorical analysis for Target 0 and Target 1



- **From the above analysis we can say that females pay back the loan on time as compared to males**

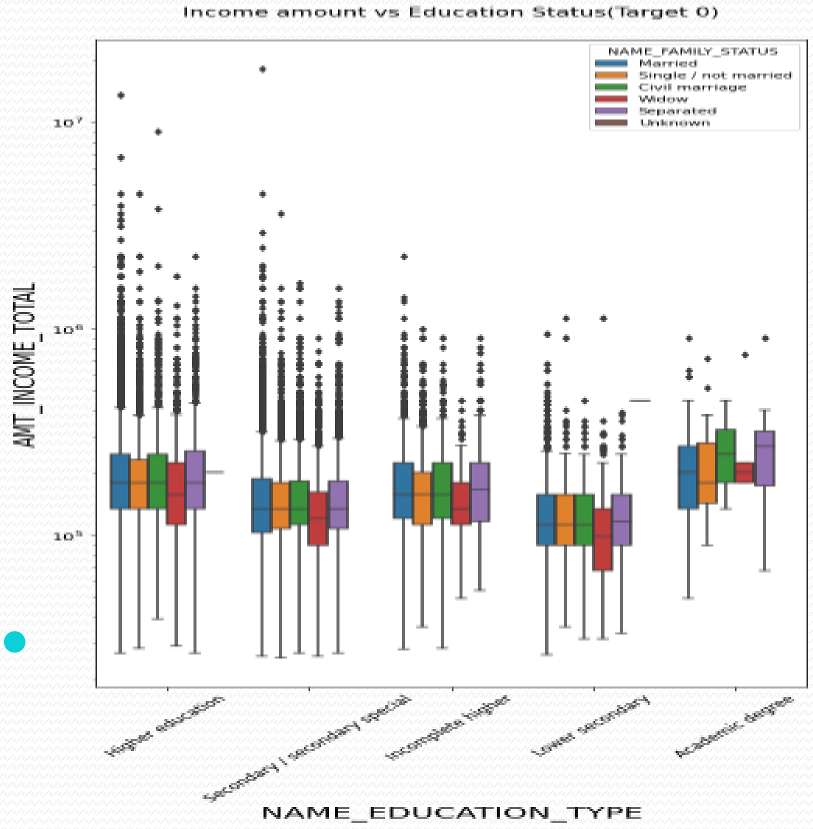
For Credit amount

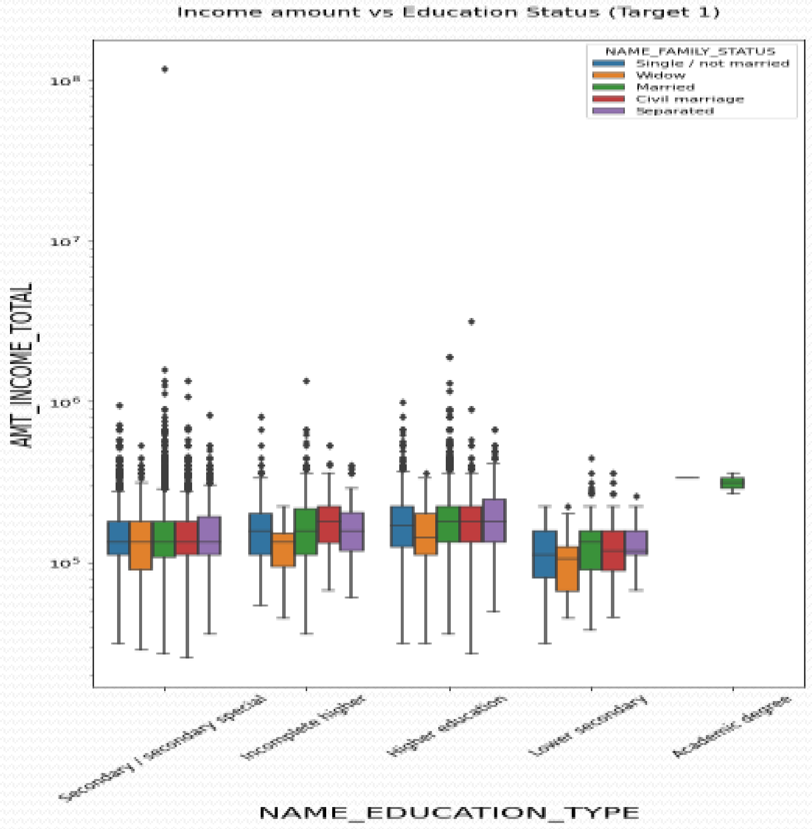




- **Only Married people with academic degree have payment difficulties.**
- **Higher educated clients have higher credit and less payment difficulties.**
- **Secondary special clients have more payment difficulties with more credit.**
- **Lower secondary educated clients have very low credits.**

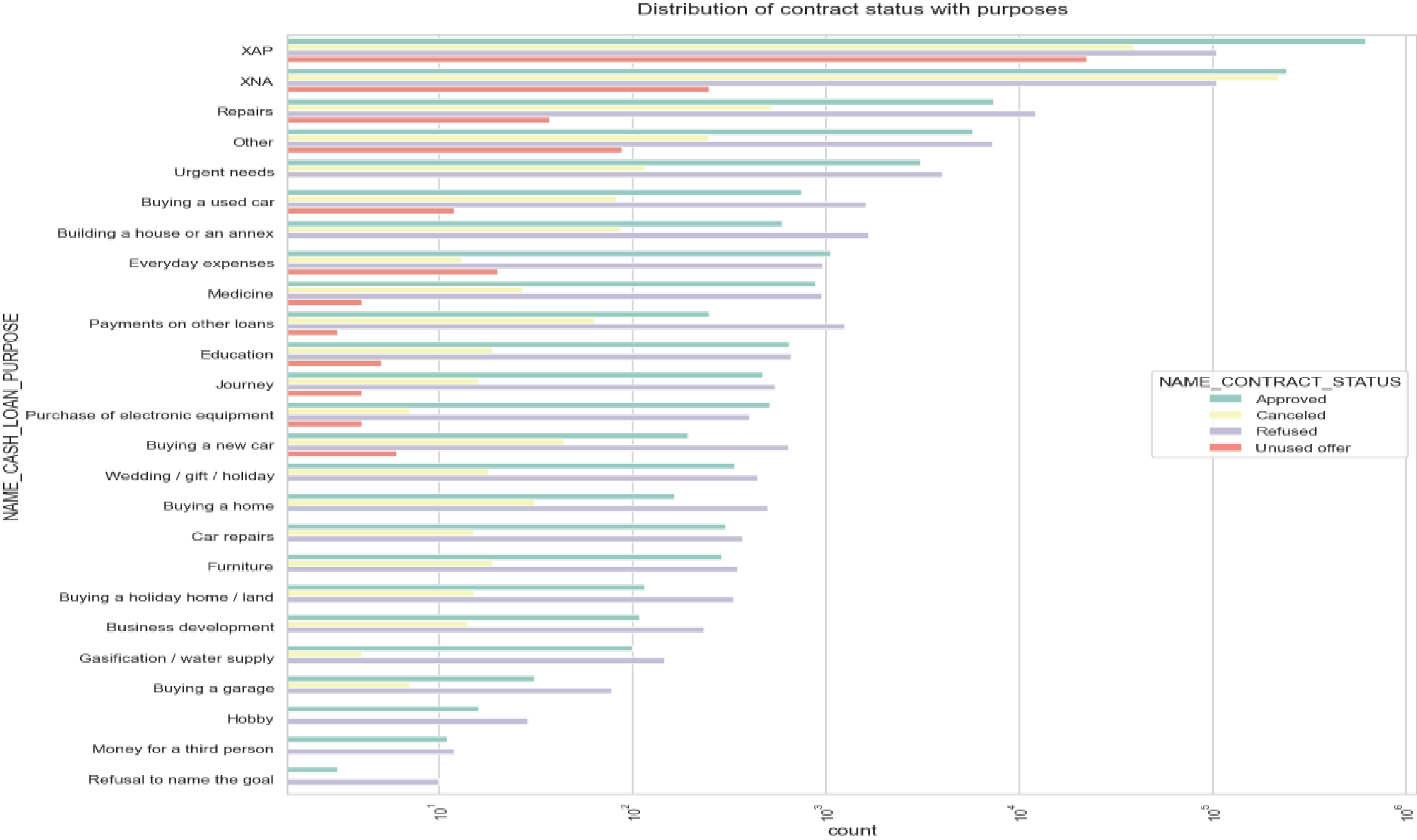
For Income amount in logarithmic scale





- **Higher education clients have highest incomes and low payment difficulties.**
- **People with low income have payment difficulties specially in secondary category.**
- **Higher education has many outliers.**
- **Lower secondary category has low income.**
- **We can conclude that higher educated clients with good income can payback the loan on time.**

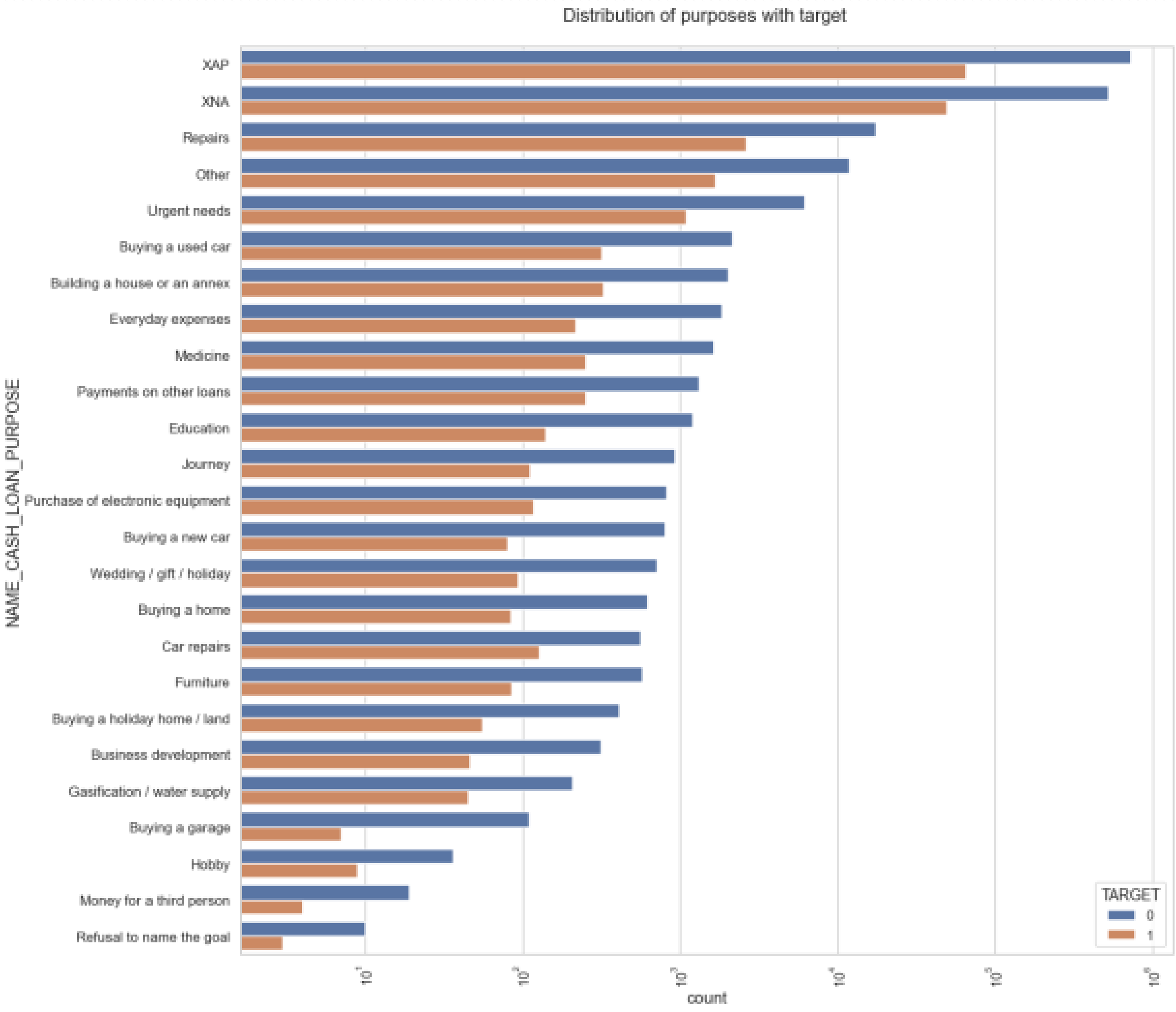
Distribution of contract status



Inferences for Contract Status With Purposes

- Medicine and Education have same number of approved and refused offers.
- Repairs has more number of refused than approved which means it has high default rate.
- Buying a car also has more refused offers than approved which means banks consider it risky.
- It has huge number of missing value

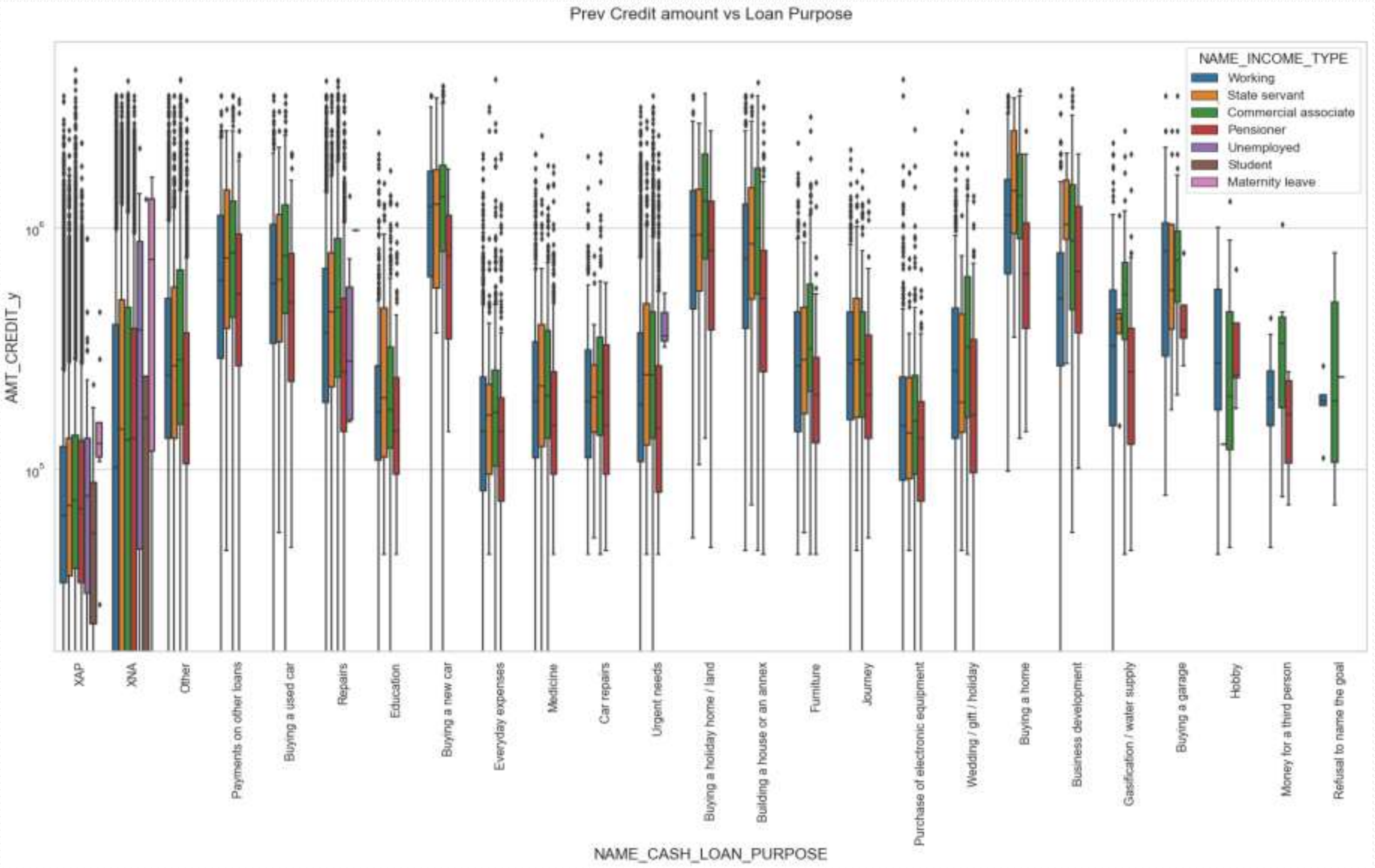
Inferences for Contract Status With Target



Inferences for Contract Status With Target

- Here also clients of loan purpose repairs has the most difficulty to payback the loan. There some cases like education, Buying a Home, Business development where less difficulties are faced for payment so, these categories can be considered for lending loan.

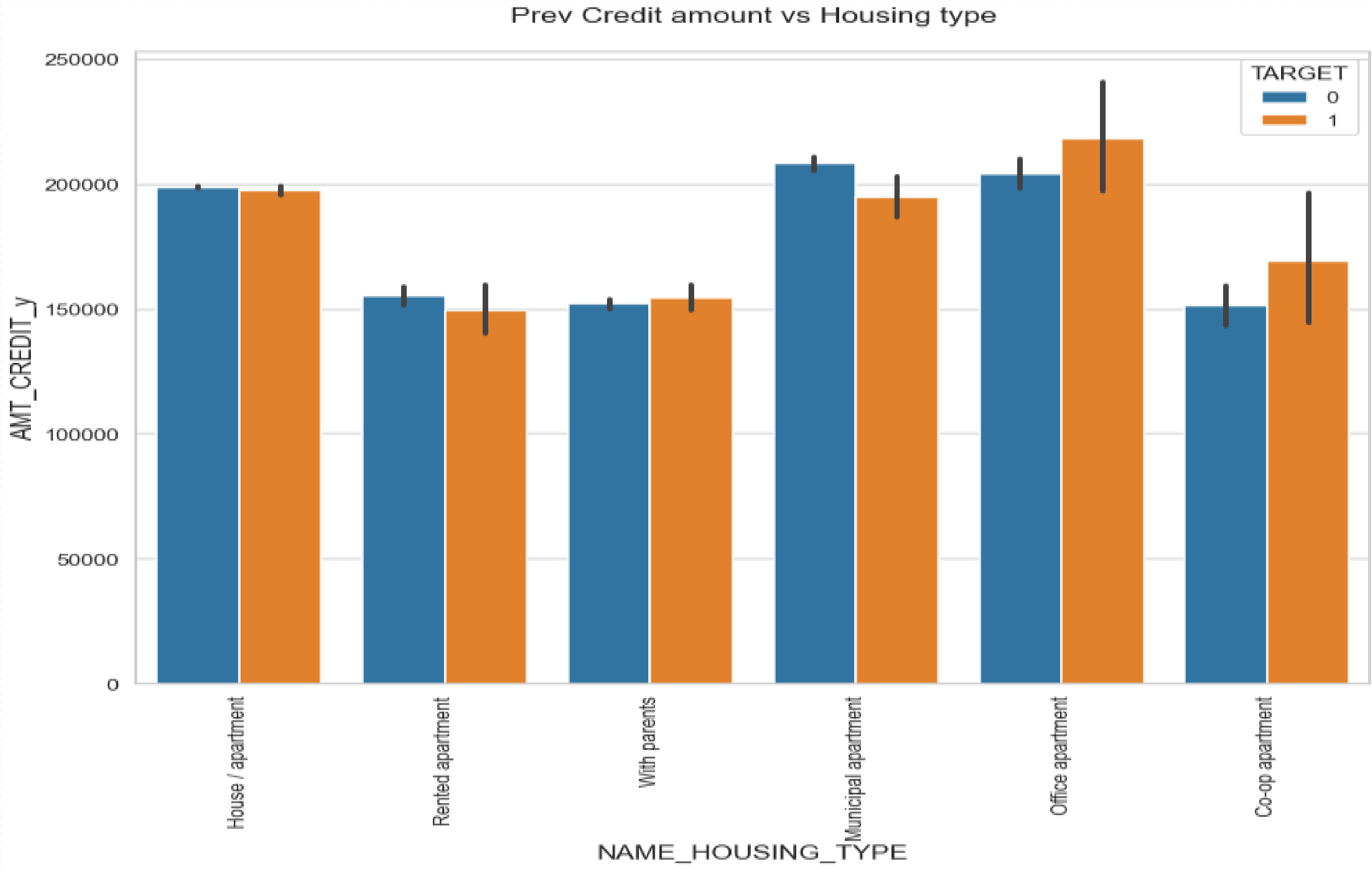
Box plotting for Credit amount in logarithmic scale



Inferences

- The credit amount of Loan purposes like 'Buying a home', 'Buying a land‘ , 'Buying’ a new ‘car' and ’Building’ a house' is higher. Income type of state servants have a high amount of credit for buying a new home and Hobby is having less credits. Working income type have highest credits for buying a new home.

Box plotting for Credit amount Prev vs Housing type in logarithmic scale



Inferences

- **Clients living in office apartment have high difficulties in paying back and those living in municipal apartment don't have much difficulty in paying back. Banks should focus on clients who own apartment or housing type of municipal apartment.**

CONCLUSION

Factors to decide whether an applicant will be repayer:

- *NAME_EDUCATION_TYPE: Academic degree and higher educated clients have less defaults.*
- *NAME_INCOME_TYPE: Student and Businessmen have no defaults.*
- *DAYS_BIRTH: People in the age group 50-60 have low rate of defaulting.*
- *AMT_INCOME_TOTAL: Applicant with Income more than 700,000 are less likely to default*
- *CNT_CHILDREN: People with zero to two children tend to repay the loans.*
- *CREDIT_SCORE : People with average to high credit score are less likely to default*
- *AMT_CREDIT_RANGE: Loan of less credit can be repayer easily*
- *OCCUPATION_TYPE: Sales staff are core staff are less likely to default.*
- *NAME_INCOME_TYPE: State servants or working professionals are less likely to default.*

Factors to decide whether an applicant will be Defaulter:

- *CODE_GENDER: Men are more likely to default.*
- *NAME_FAMILY_STATUS : People who are single or who have civil marriage default a lot*
- *NAME_EDUCATION_TYPE: People with Lower Secondary education have higher default rate*
- *NAME_INCOME_TYPE: Clients who are Unemployed default a lot.*
- *OCCUPATION_TYPE: Avoid Drivers and Waiters/barmen staff, Security staff and Cooking staff as the default rate is huge.*
- *DAYS_BIRTH: Avoid young people or too old people as they have higher probability of defaulting*

THANK YOU