CREDIT EDA ASSIGNMENT

Submitted by,
Aparna Menon S
Bhagyesh Shah
PGDDS,IIIT-B(February 2021)

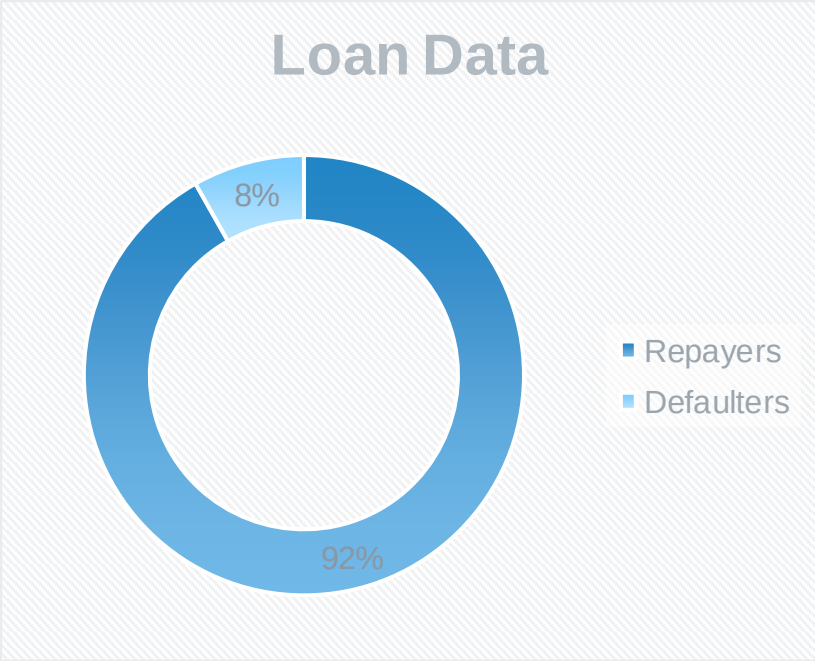
Index

Problem Statement

- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
  - If the applicant is not likely to repay the loan, i.e., customer is likely to default, then approving the loan may lead to a financial loss for the company

The main aim of the EDA is to draw insights from the data sets to understand what are the attributes that determine loan default.

Data Imbalance

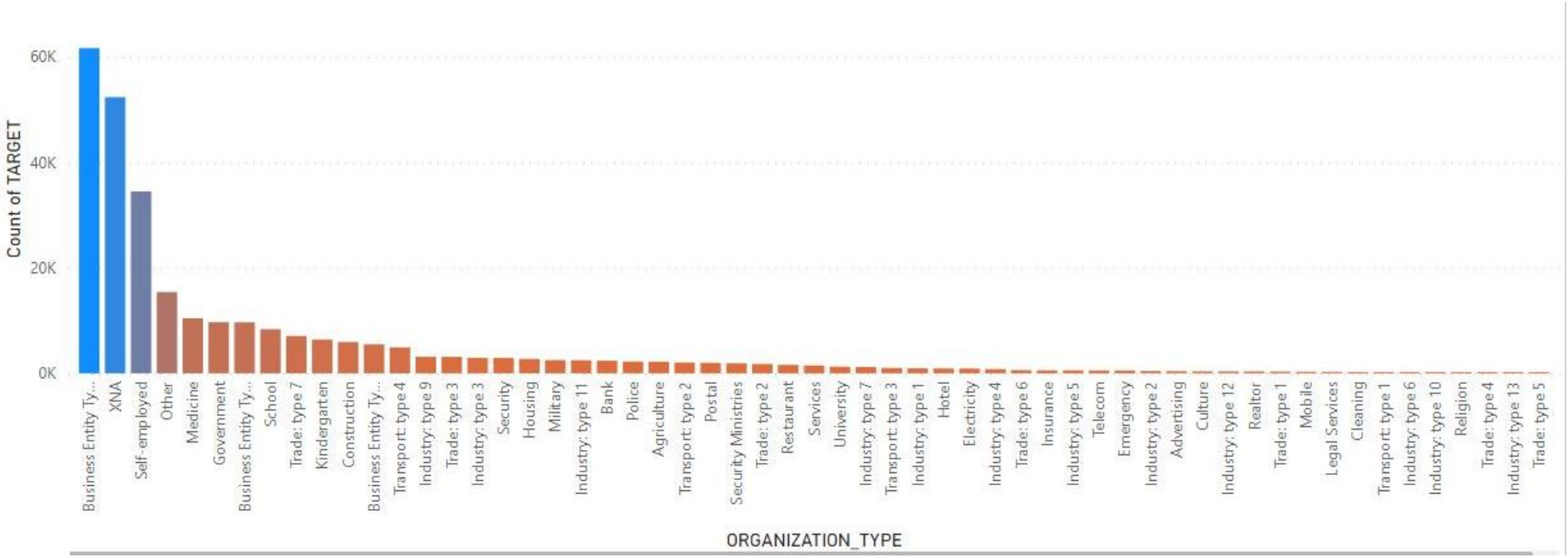


The distribution of records with Target==0 (Re-payers) makes up 92 % of the records in the data set , on the other hand , for Target==1(Defaulters) it only constitutes to 8% of the dataset records.

- Methods to fix data imbalance
  - Oversampling can be used when the quantity of data is insufficient. It tries to balance dataset by increasing the size of fewer samples(Target==1). Rather than getting rid of abundant samples, new rare samples are generated by using e.g., repetition, bootstrapping.
  - Under-sampling balances the dataset by reducing the size of the abundant class. This method is used when quantity of data is sufficient. By keeping all samples in the rare class and randomly selecting an equal number of samples in the abundant class, a balanced new dataset can be retrieved.

Univariate Analysis

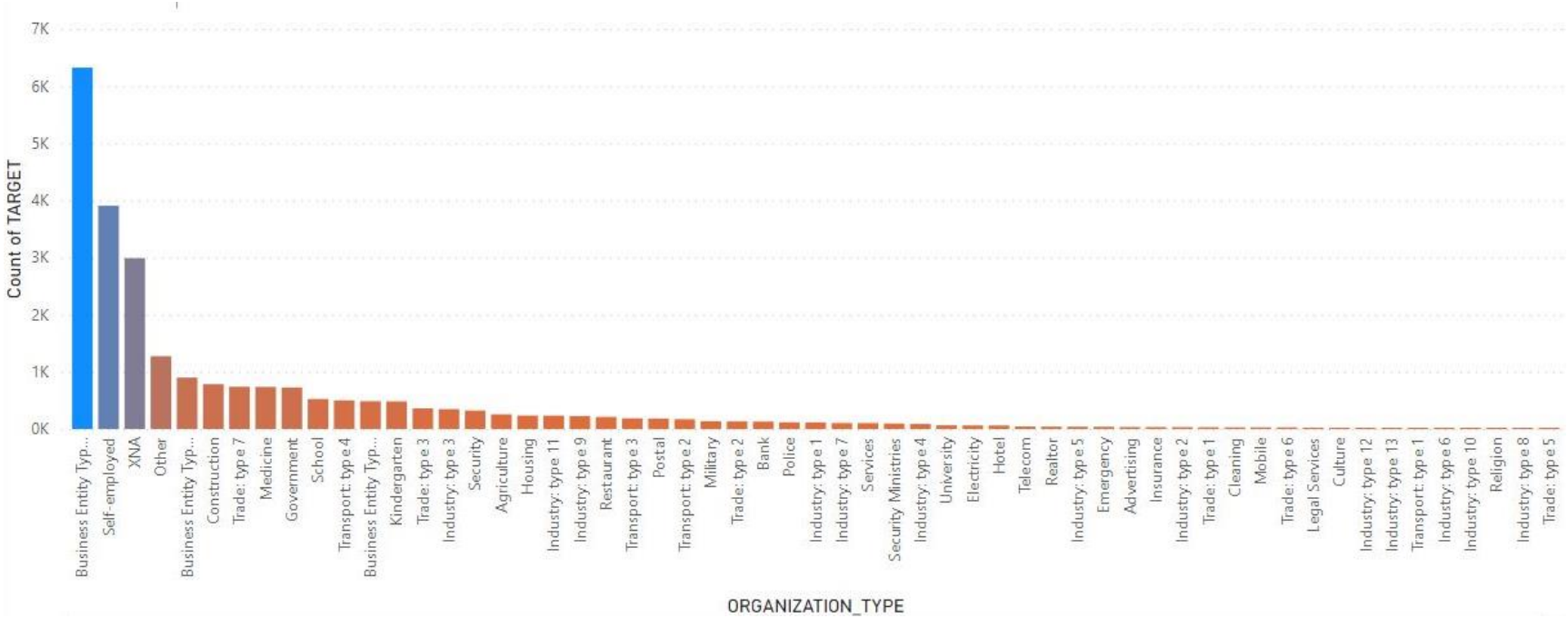
Organization Type (Re-payers)



Inferences from the graph 'Target0 - Organization Type'

•Customers who belong to Industry Type 13 , Trade Type 4,5 are the least likely to default.

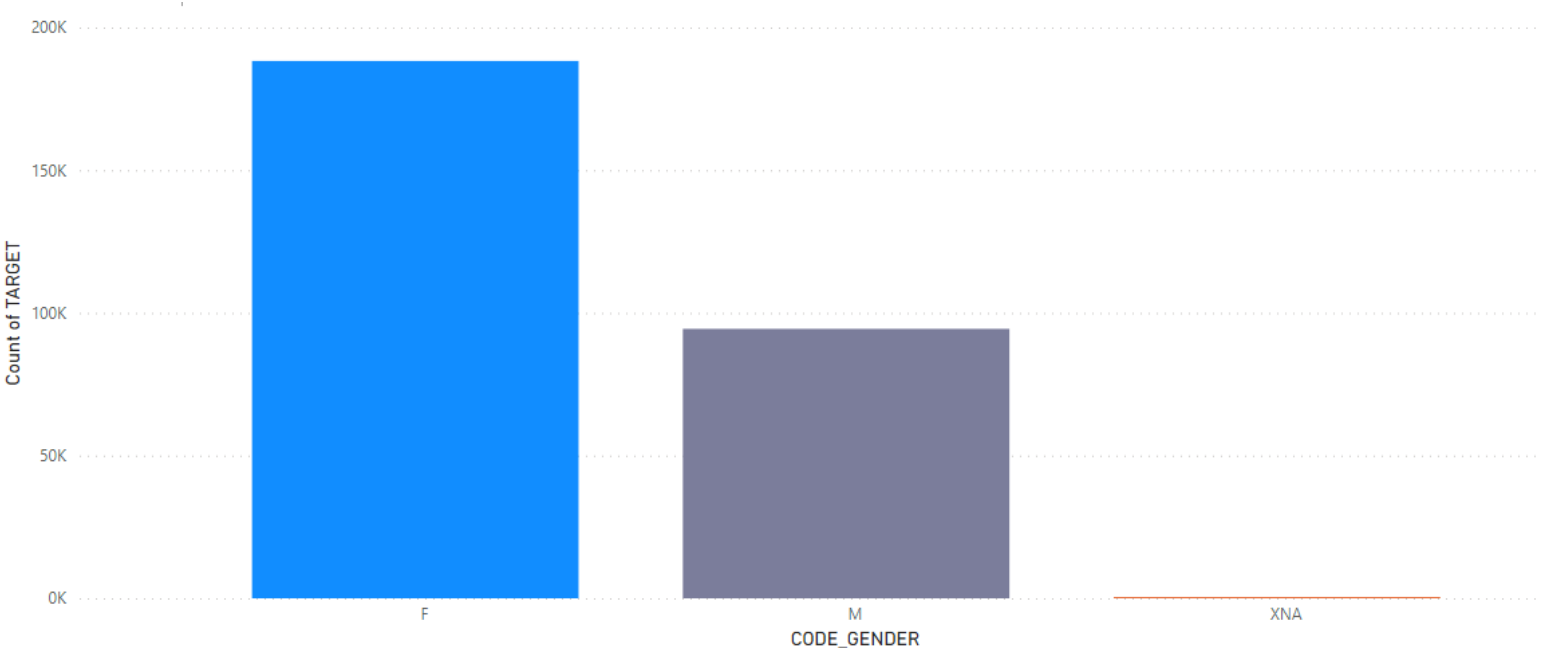
Organization Type (Defaulters)



Inferences from the graph 'Target1 - Organization Type'

•Customers who belong to Business Entity type 3 and Self-Employed Organization Type have highest percentage of defaulters.

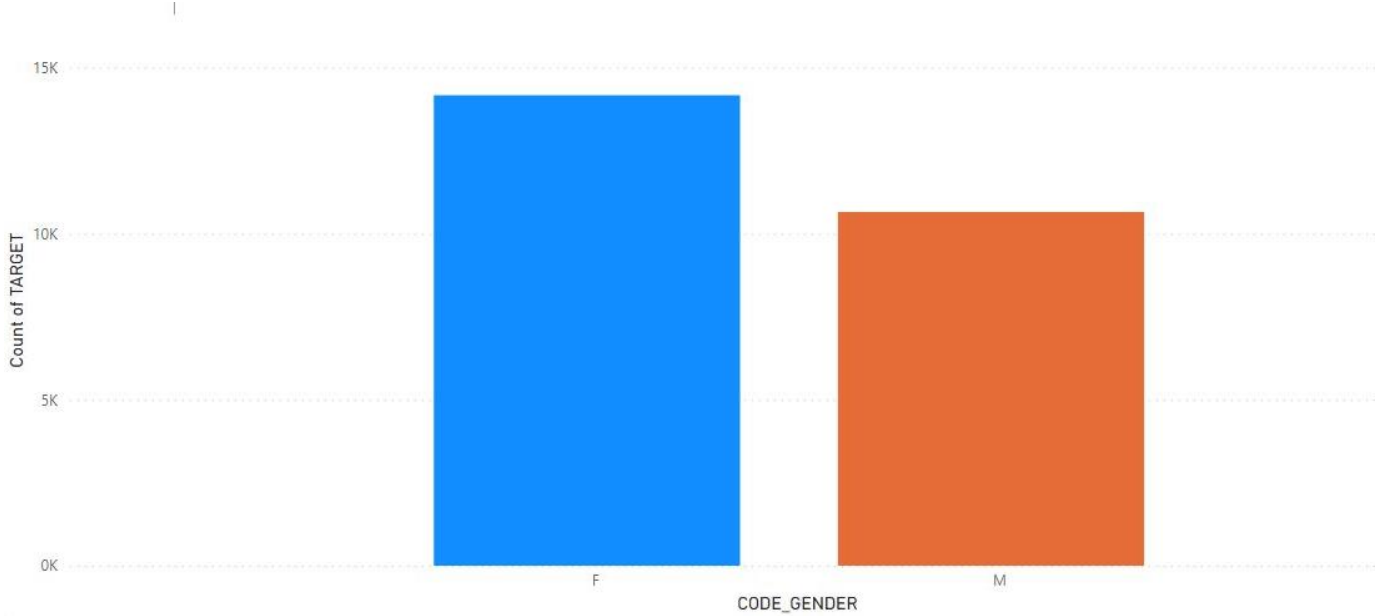
Gender (Re-payers)



Inferences from the graph 'Target0 - Gender'

- Female customers are in a greater percentage among customers who pay back without defaulting.

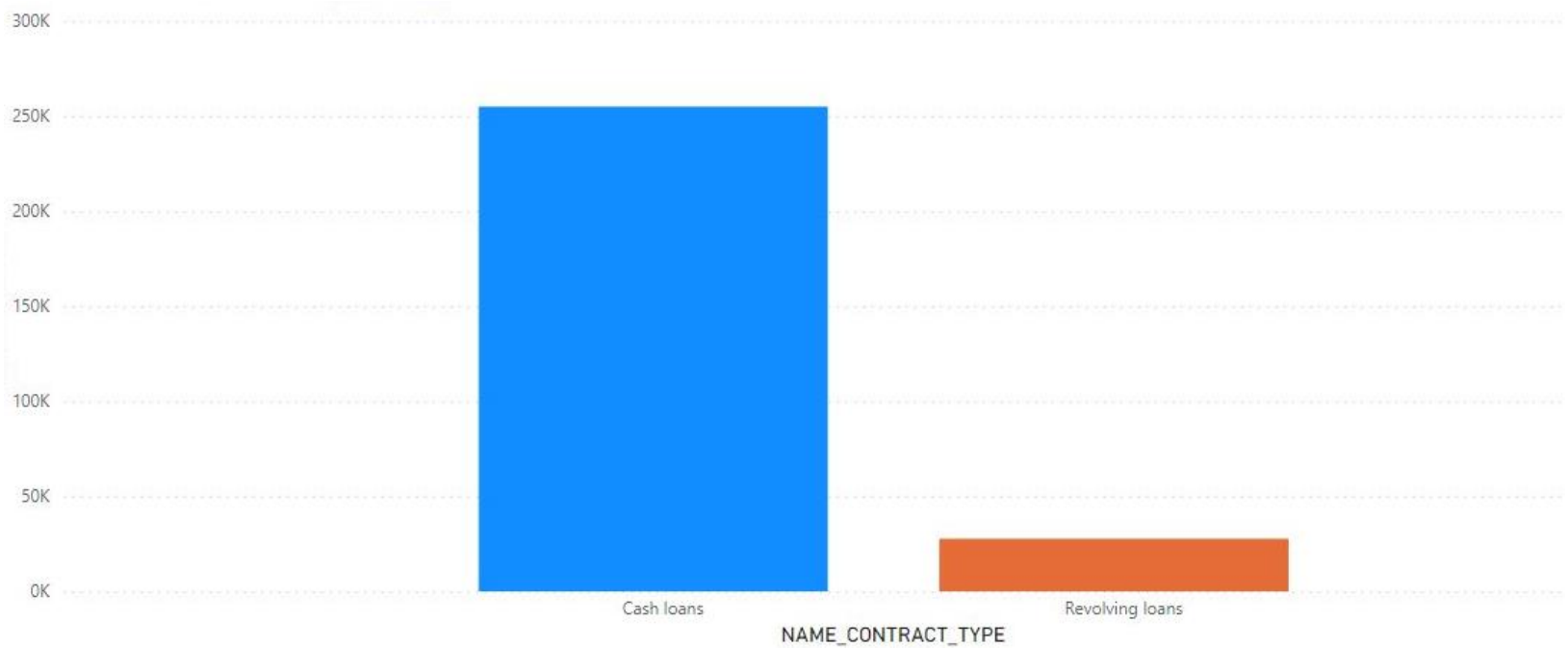
Gender (Defaulters)



Inferences from the graph 'Target1 - Gender'

•Since the ratio of female customers are more compared to male customers, female customers make up the larger percentage of defaulters as well.

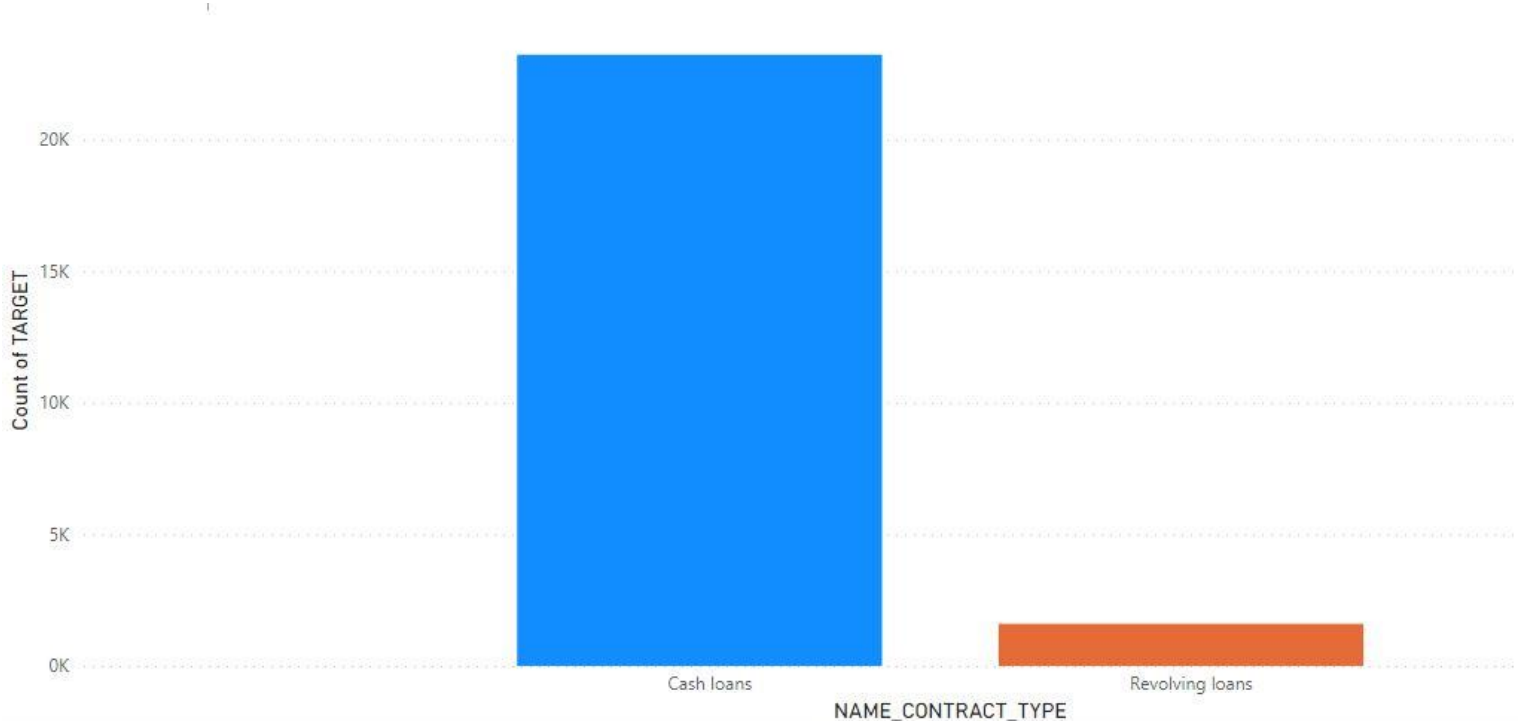
Loan Type(Re-payers)



Inference from the graph 'Target0-LOAN TYPE'

•Cash loans account for 255011 (90%) of the total loans that are repaid on time by customers, and 9% of revolving loans.

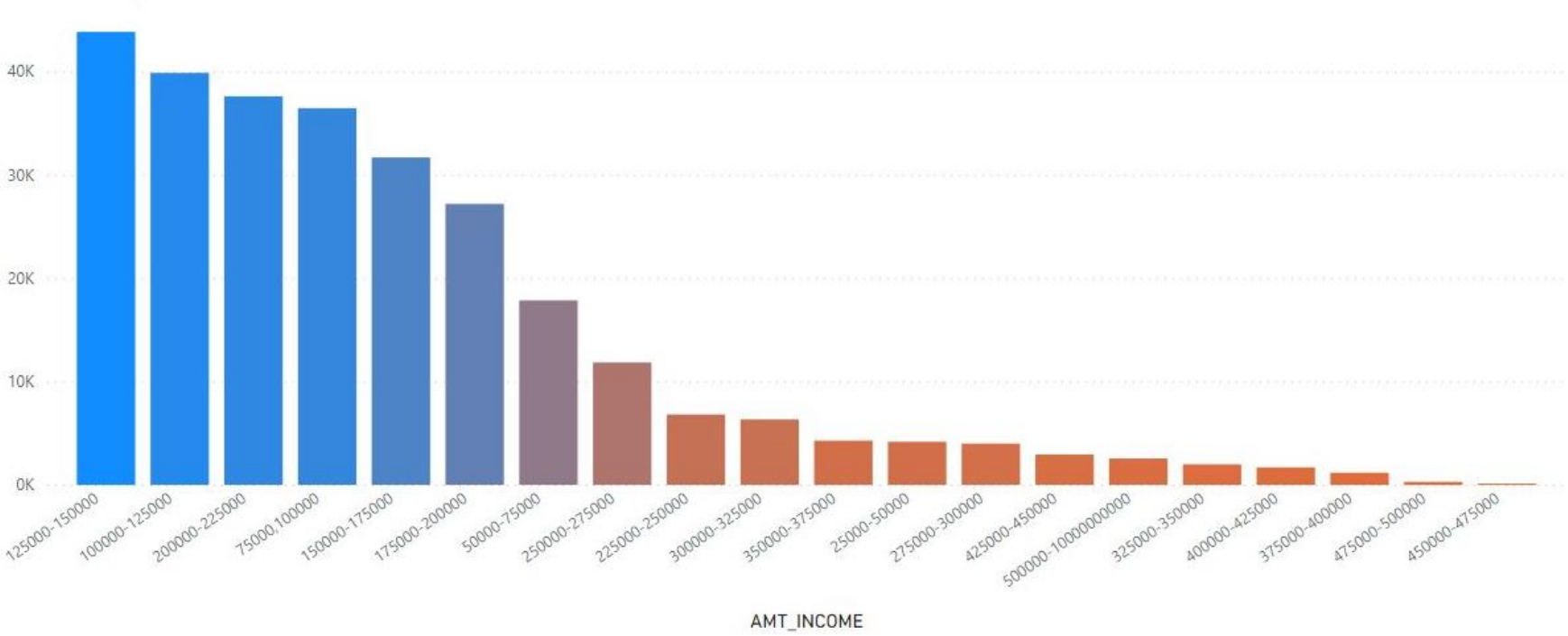
Loan Type(Defaulters)



Inference from the graph 'Target1-LOAN TYPE'

- Cash Loans make 93% of the Total Loan applications that are defaulted by users.

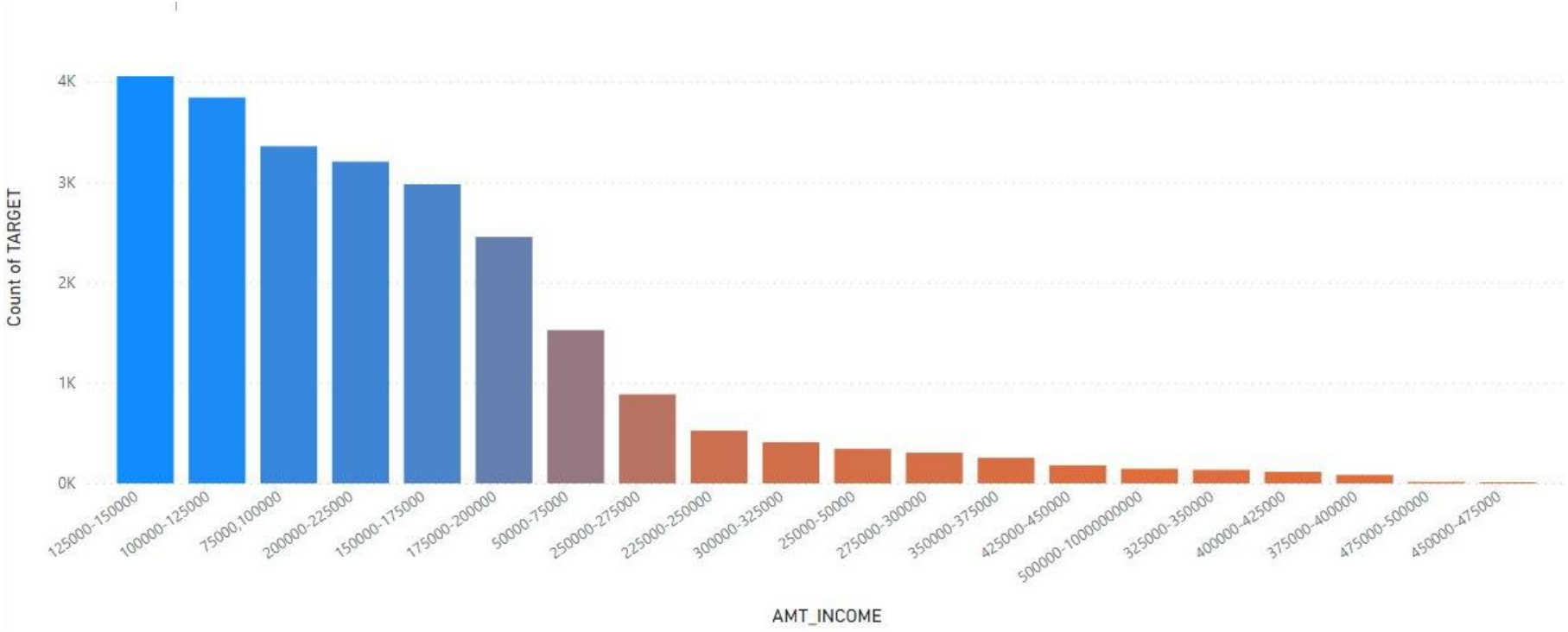
Amount Income (Re-payers)



Inference from the graph 'Target0 - Income Amount'

•Customers in the Income brackets of 750000-225000 make up the most no of safe borrowers.

Amount Income (Defaulters)

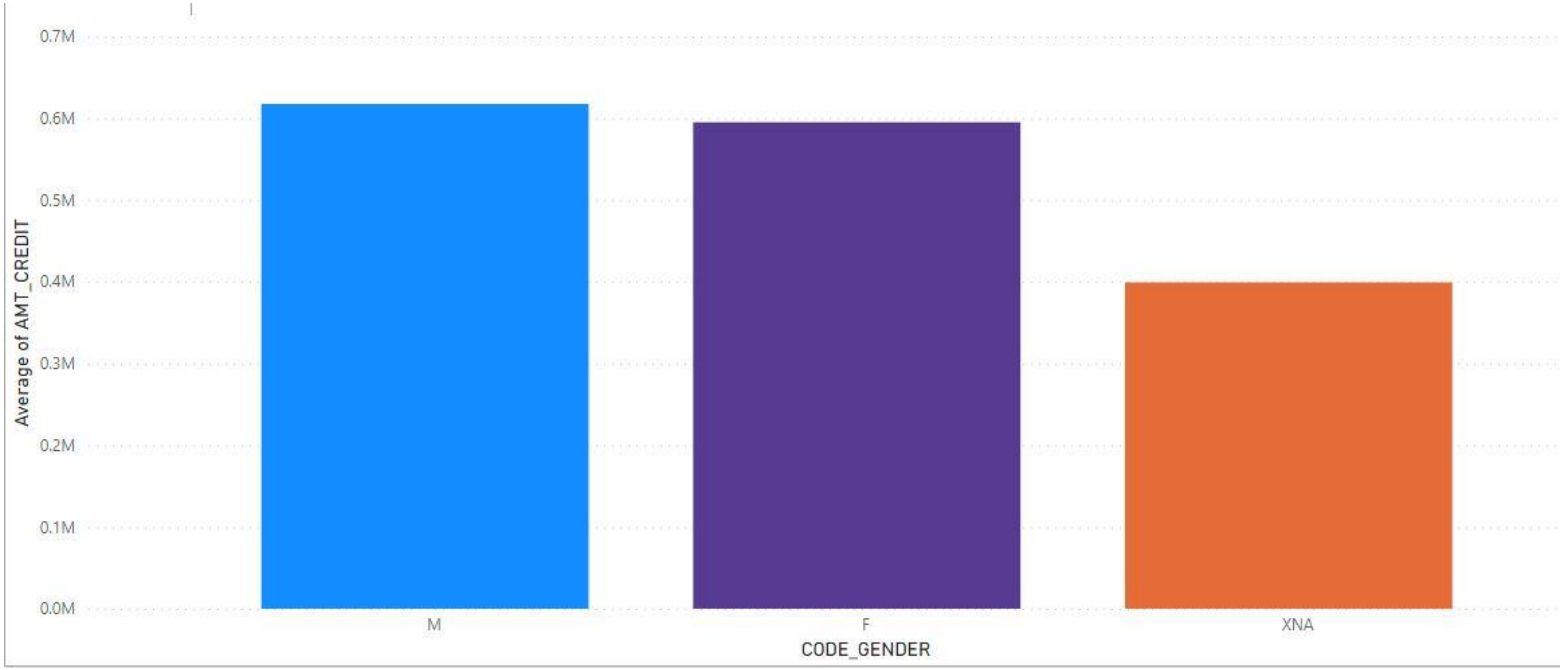


Inference from the graph 'Target1 - Income Amount'

- Most of the defaulters fall into the Income bracket of 1 lac to 1.5 lac

Bivariate Analysis

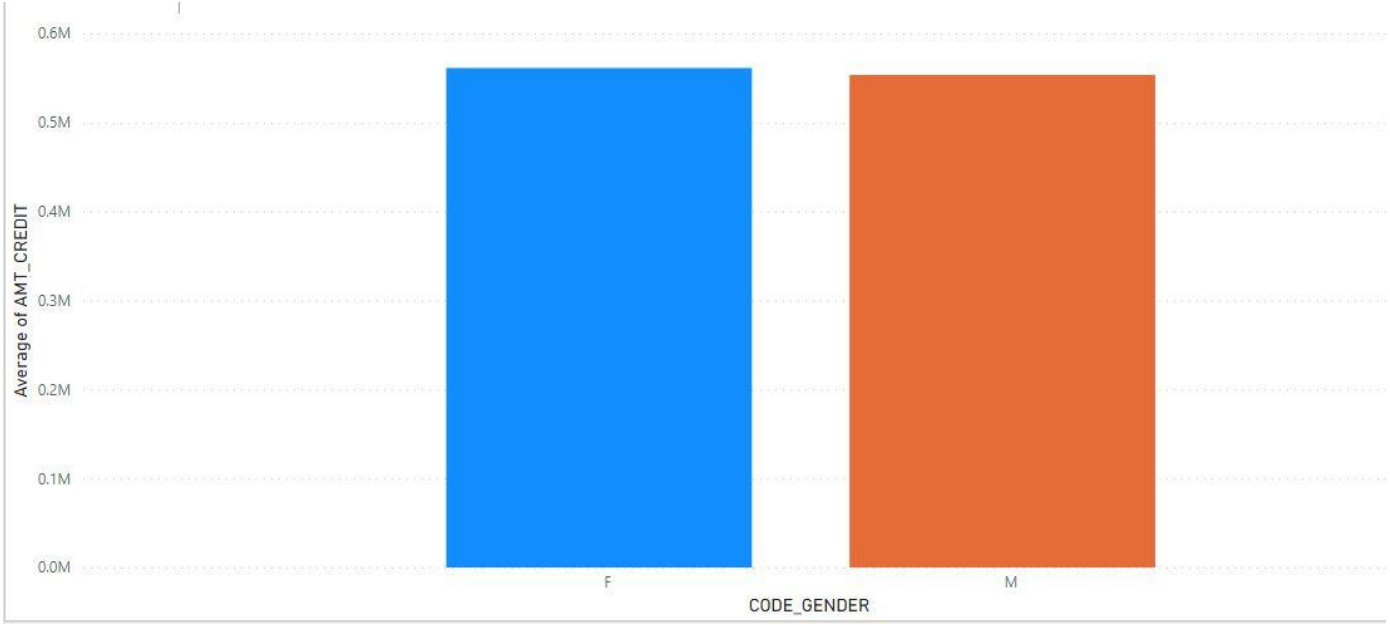
Gender vs Credit Amount(Re-payers)



Inferences from the graph 'Target0 - Gender vs Credit Amount'

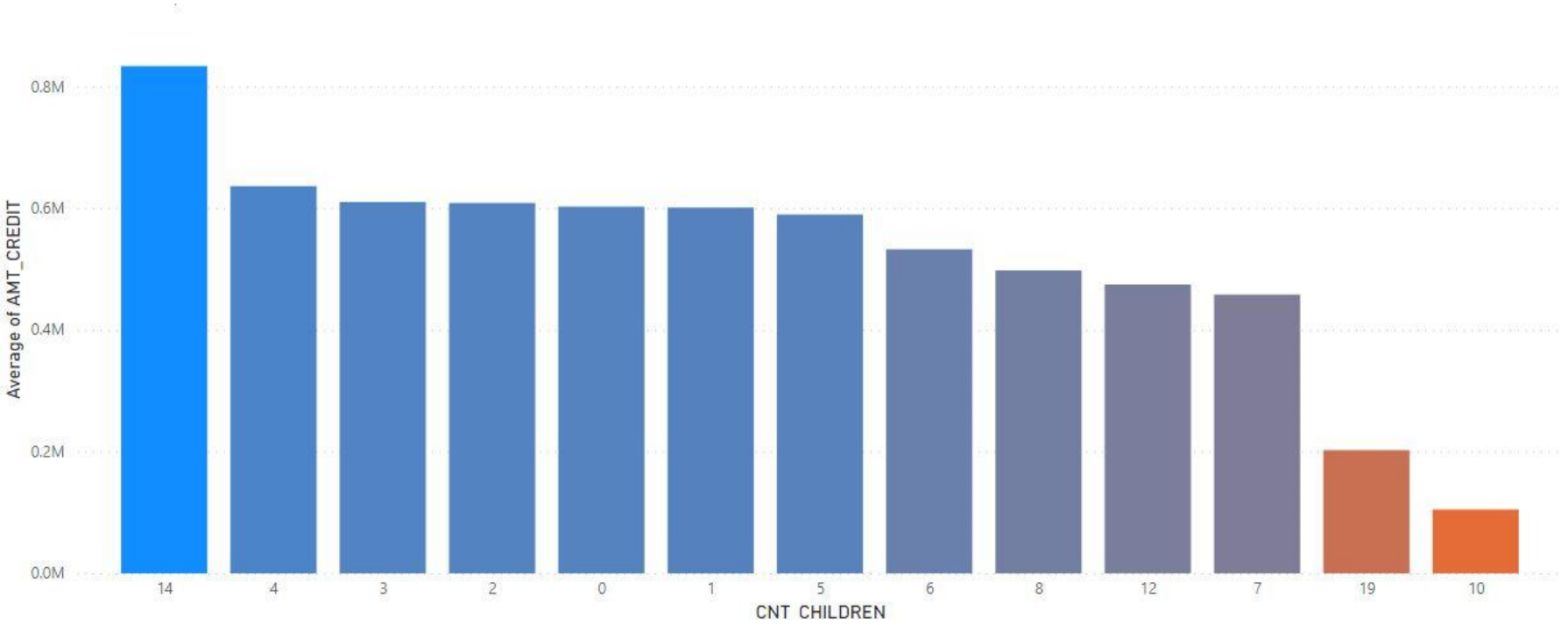
- Among customers who pay back on time i.e. (Target0) , male customers have a higher credit amount compared to female and other gender customers.
- Male customers have a highest Amount Credit over 6 lacs whereas female customers have only less than 6 lacs and other gender only less than 4 lacs.

Gender vs Credit Amount(Defaulters)



Inferences from the graph 'Target1 - Gender vs Credit Amount'
Female customers with payment difficulty has higher Amount Credit against them compared to male customers.
It can be noted that the Amount Credit for male and female customers with payback difficulty is almost identical, in contrary to customers with no default history.

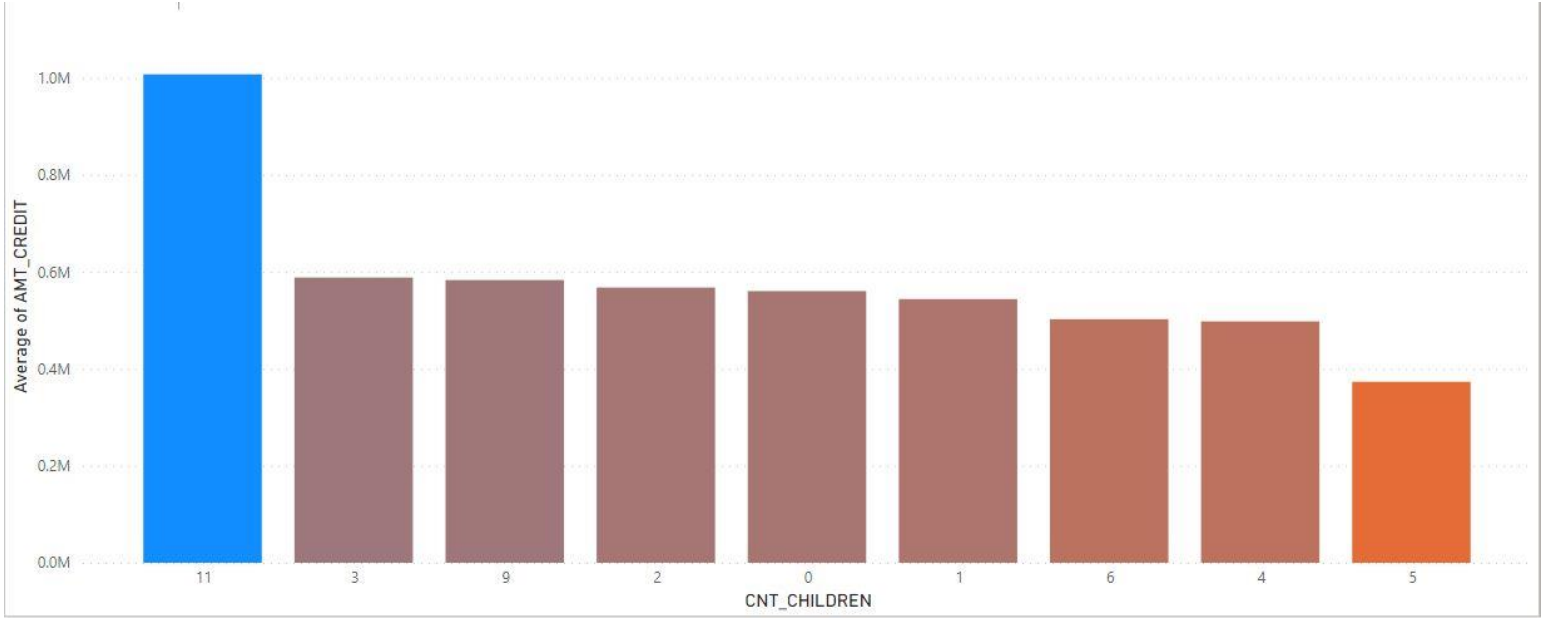
No of children vs Credit Amount(Re-payers)



Inferences from the graph 'Target0 - No of children vs Credit Amount'

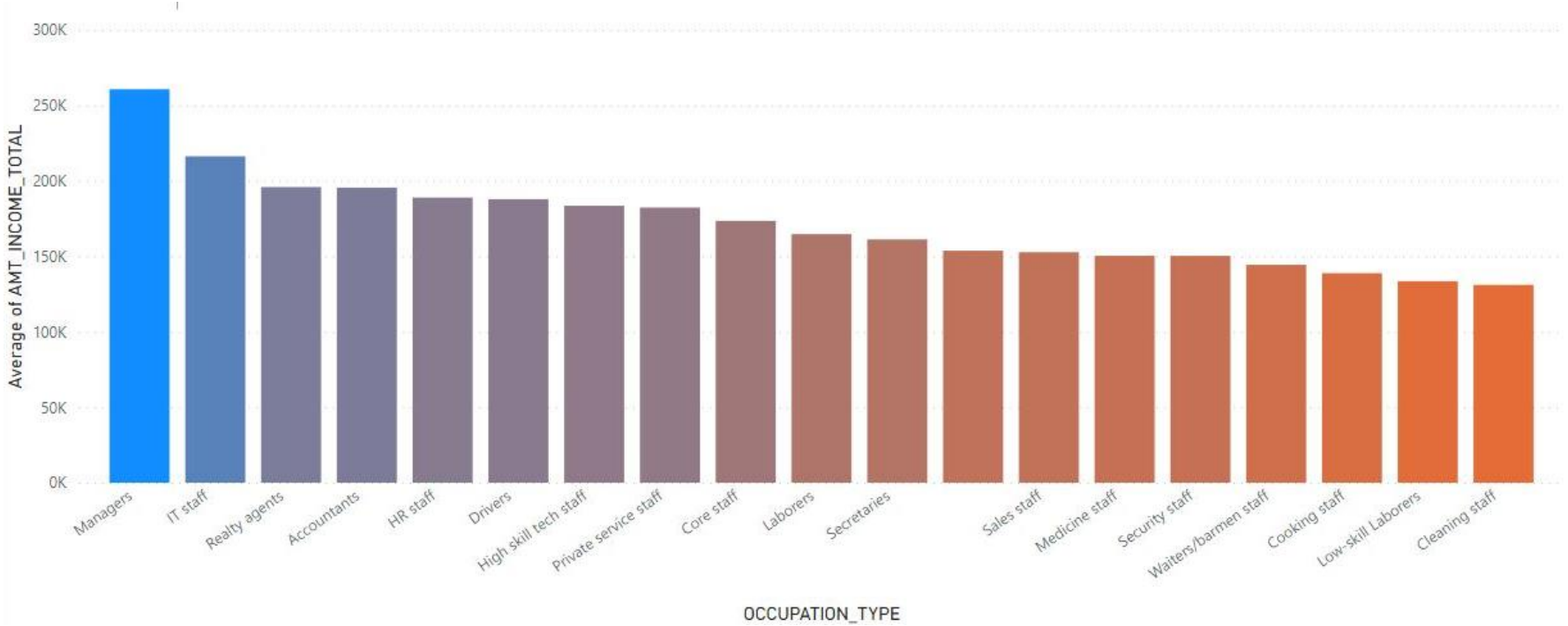
- Among defaulters, customer with 14 children seems to have the highest credit amount.
- The steady pattern in the Credit Amount and No of children suggests that the attribute holds comparatively insignificant relation to the Credit amount attribute.

No of children vs Credit Amount(Defaulters)



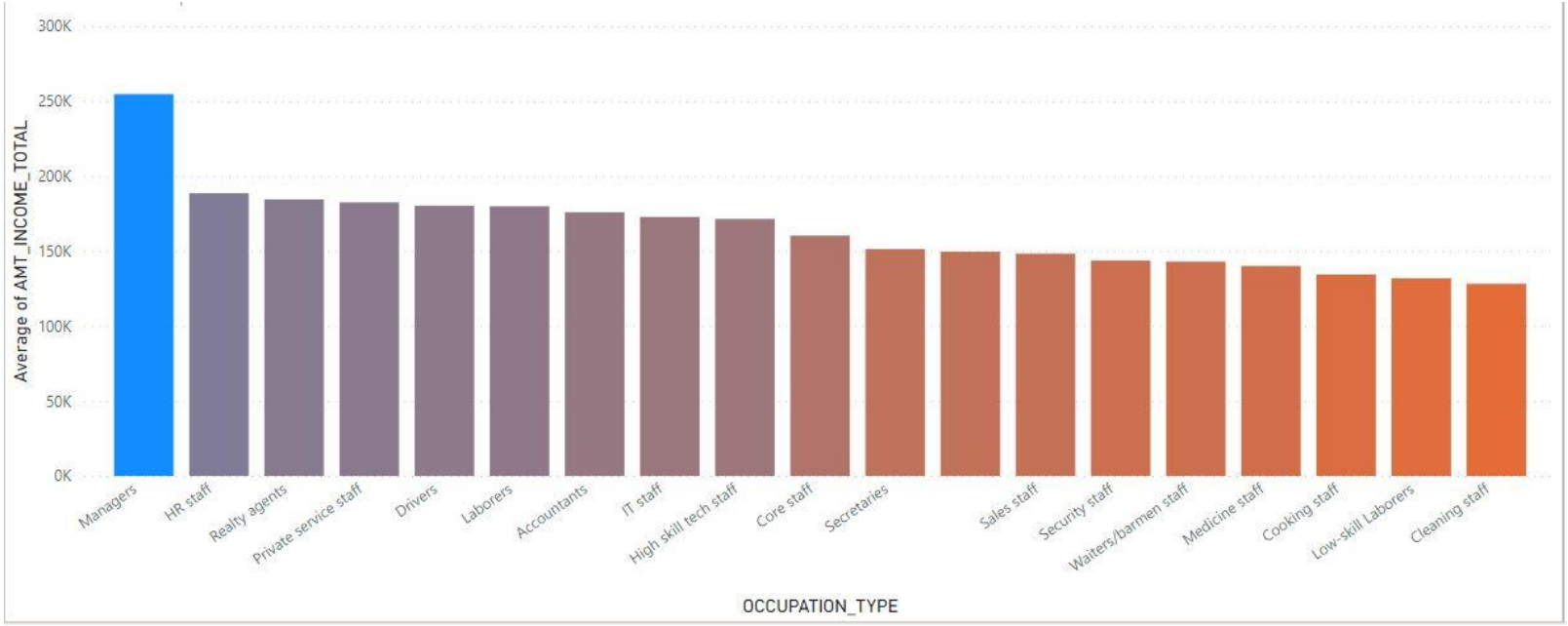
Inferences from the graph 'Target1 - No of children vs Credit Amount'
Among defaulters, customer with 14 children seems to have the highest credit amount.
The steady pattern in the Credit Amount and No of children suggests that the attribute holds comparatively insignificant relation to the Credit amount attribute.

Occupation Type vs Income (Re-payers)



Inferences from the graph 'Target0 - Occupation Type vs Total Income'
Managers are the highest paid among customers who have no payment difficulty.
Managers earn the most 2.5 lacs and Cleaning staff and Low-skill Laborers are among the customers who earn the least.

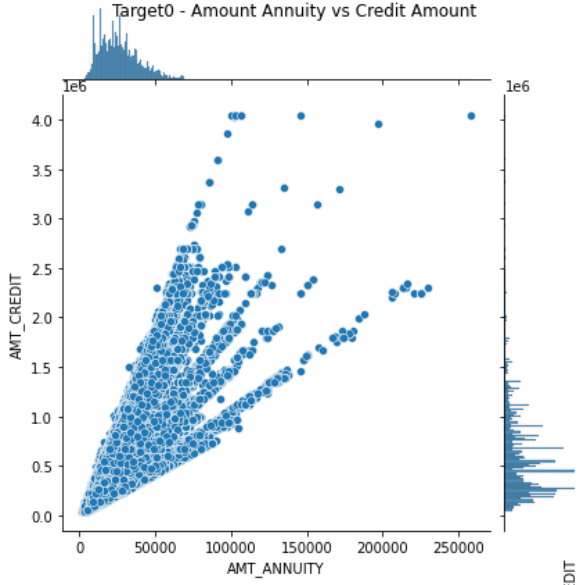
Occupation Type vs Income (Defaulters)

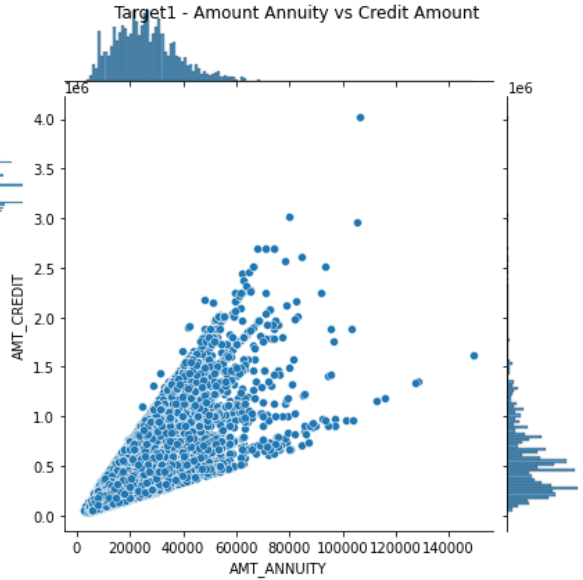


Inferences from the graph 'Target1 - Occupation Type vs Total Income'

- Among defaulters, Cleaning staff has the least Total Income

Amount annuity vs Credit Amount

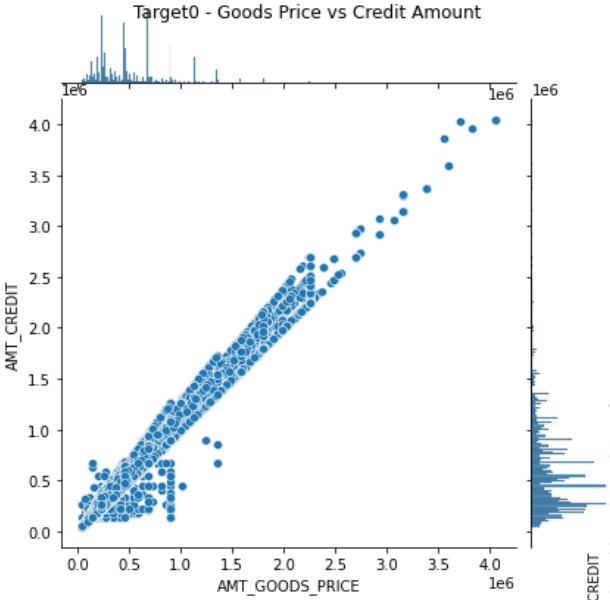




Inferences from the graph 'Target0 - Goods Price vs Credit Amount' and 'Target1 - Goods Price vs Credit Amount'

- Since the graph shows high positive correlation, it indicates that as the Goods Price increases the Credit Amount increases, i.e., if a customer is pledging on a goods that has a higher price, higher credit amount can be provided to them.

Goods Price vs Credit Amount



Inferences from the graph 'Target0 - Goods Price vs Credit Amount' and 'Target1 - Goods Price vs Credit Amount'

- Since the graph shows high positive correlation, it indicates that as the Goods Price increases the Credit Amount increases, i.e., if a customer is pledging on a goods that has a higher price, higher credit amount can be provided to them.

Correlation Matrices

Correlation Matrix of numerical columns (target0)

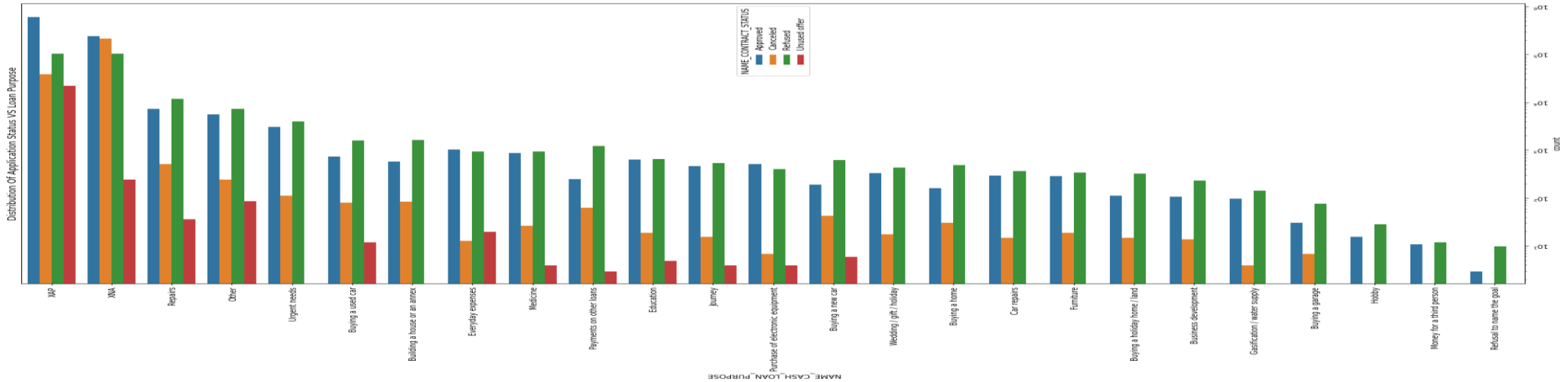
Correlation Matrix of numerical columns (target1)

Inferences from Correlation Matrices

- Columns AMT_GOODS_PRICE AND AMT_CREDIT has a high positive correlation, indicative of the fact that as the price of the goods pledged increases , the AMOUNT_CREDIT increases as well. From a customers' perspective, as they pledge a goods with higher goods price yield more Credit Amount.
- Columns AMT_ANNUITY and AMT_INCOME_TOTAL are correlated, as AMT_ANNUITY represents the amount paid back by the customer in each installment, a customer with higher AMT_INCOME_TOTAL will have a higher AMT_ANNUITY and vice versa.
- The CNT_CHILDREN and CNT_FAM_MEMBERS columns are correlated, because they are dependent on each other.

Analysis on Combined Dataset

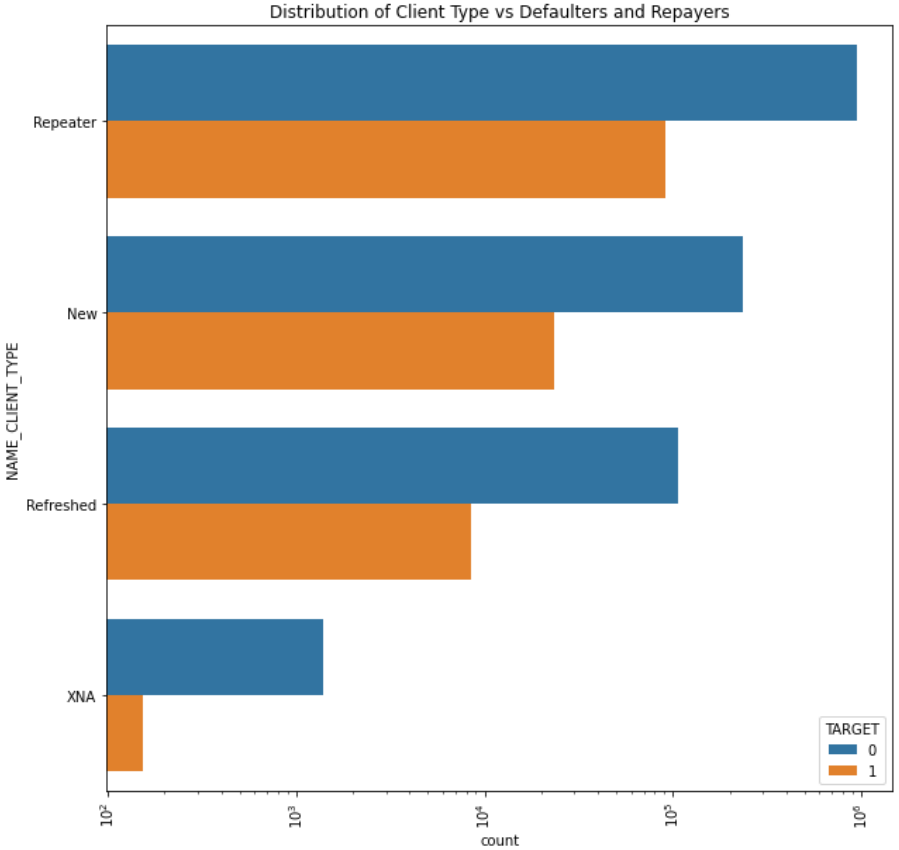
Application Status vs Loan Purpose



Inferences from the graph 'Distribution of Application Status VS Loan Purpose'

- As mentioned earlier , XAP is used as the highest Loan Purpose, which can be hinderance in studying the patterns.
- XAP(Not Applicable) and XNA(Not Available) makes up for most of the Applications.
- Repairs covers the maximum Approved Loan purpose among known applications where the loan purpose is correctly stated.

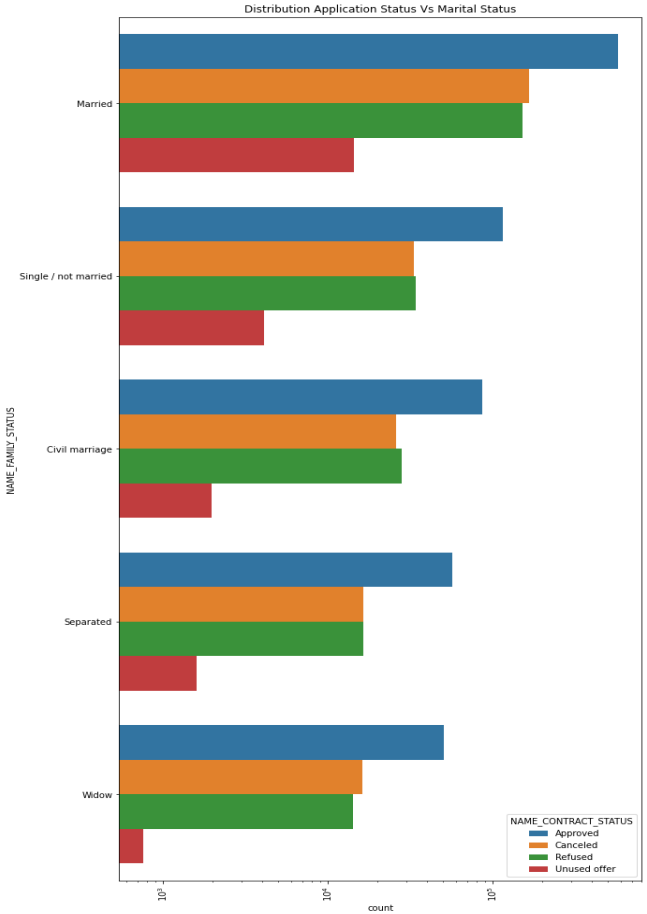
Client Type vs Defaulters and Re-payers



Inference from the graph 'Distribution of Client Type vs Defaulters and Re-payers'

- Repeating customers tend to repay the loan without default more than New and Refreshed customers.

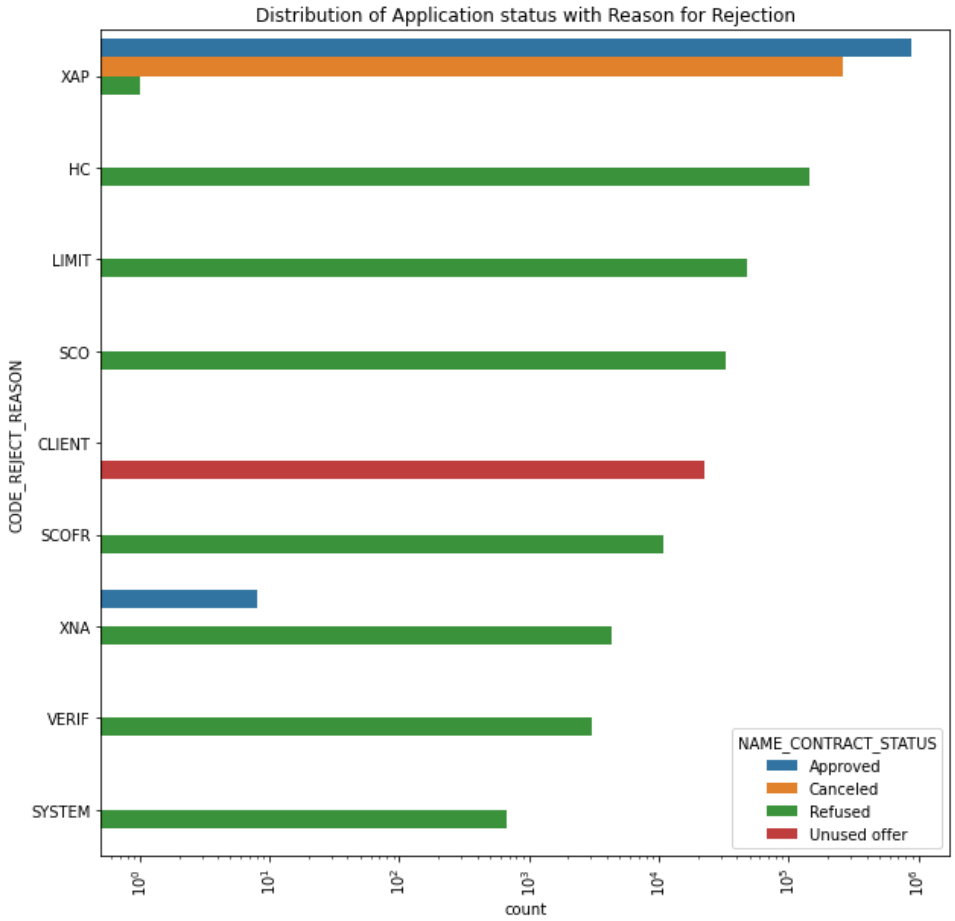
Application Status vs Marital Status



Inference from the graph 'Distribution Application Status Vs Marital Status'

- Married customers have got loan Approved than Single, Separated and Widow customers
- Married people have got most no of loans approved.
- Among separated customers, the percentages of loans cancelled and refused are almost equal.

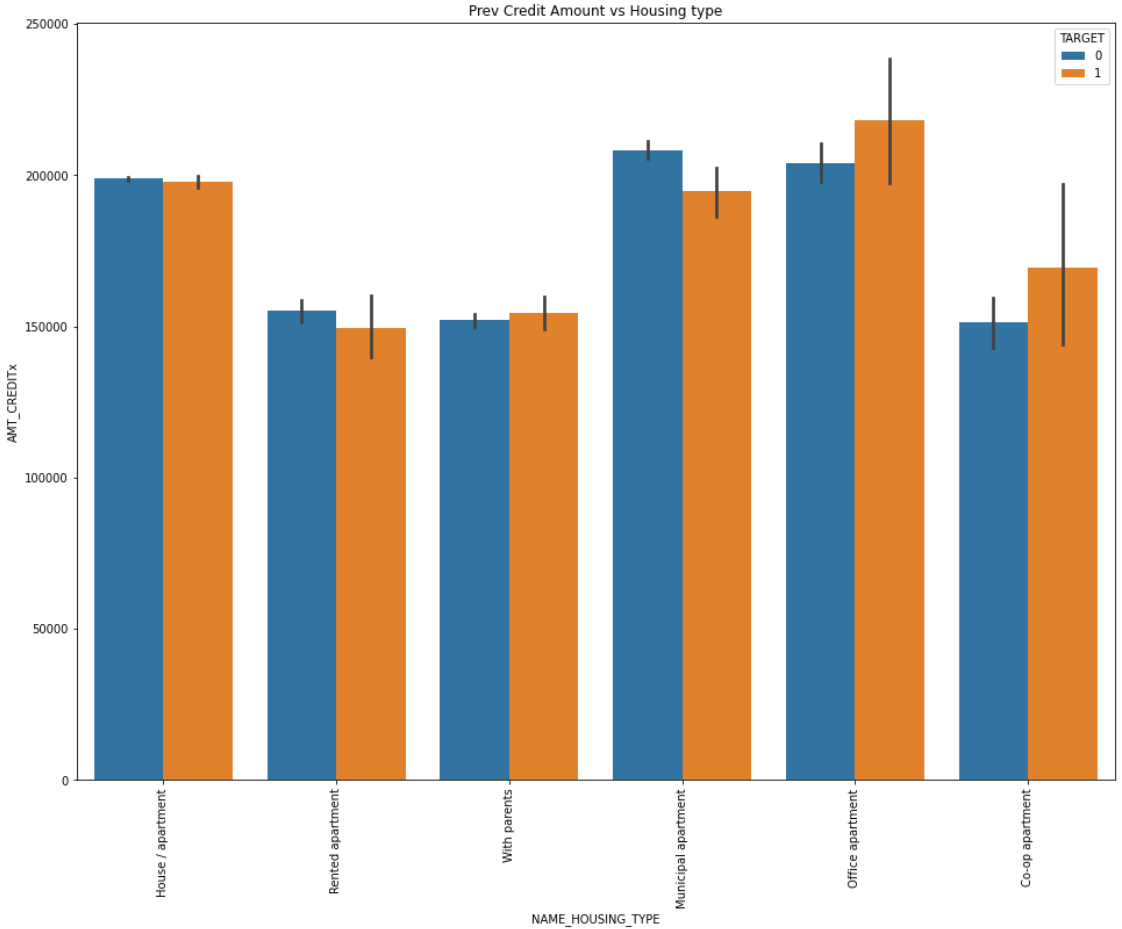
Application Status vs Marital Status



Inferences from the graph 'Distribution of Application status with Reason for Rejection'

- Bank has no rejection reason for most of the cases. (this can be a draw back when considering future studies that can be done on this data set to derive insights)

Previous Credit Amount vs Housing Type



- For Housing type, office apartment is having higher credit of target 1 and co-op apartment is having higher credit of target 1.The bank should avoid giving loans to the housing type of office apartment as they are having difficulties in payment.
- Bank Shall focus mostly on housing type with parents or House\apartment, municipal apartment for successful payments.

EDA Conclusion

Factors determining Re-payers

- Factors that influence a customer to be a Re-payer and so in future the bank can analyze these factors to approve such customers' loan.
  - Customers aged above 50 years are less likely to default , such customers can be given the loan.
  - Customers with no children have less chances of defaulting.
  - Returning customers are more likely to pay the loan without defaulting and thus can be offered loan.
  - Customers from Trade Type 4, 5 and Industry Type 13 are most likely safe borrowers. (less than 1% default rate)
  - Customers with an Amount Income above 100000 are less likely to default.

Factors determining Defaulters

- Factors that influence a customer to be a Defaulter and so in future the bank can analyze these factors and avoid approving loan.
  - When the credit amount exceeds 3000000 , the chances are high that the customer may default.
  - Self Employed customers have a higher chance of default rating, such customers can be avoided.
  - Customers with their marital status as Single are more likely to default.
  - Customers aged between 20-40 are more likely to default.

Factors determining Potential Customer with High Interest Loan

- Factors that influence a customer to be a potential defaulter and so in future the bank can analyze these factors and approve loans with higher interest rates.
  - Most loans are rejected on the grounds of unknown reason, followed by the loans that are applied for Repairs, such loans can be provided with higher interest to make sure the customer wont default.
  - Customers with 1-3 children can be given loan with higher rates since they don't contribute to the highest rate of default

Thank You