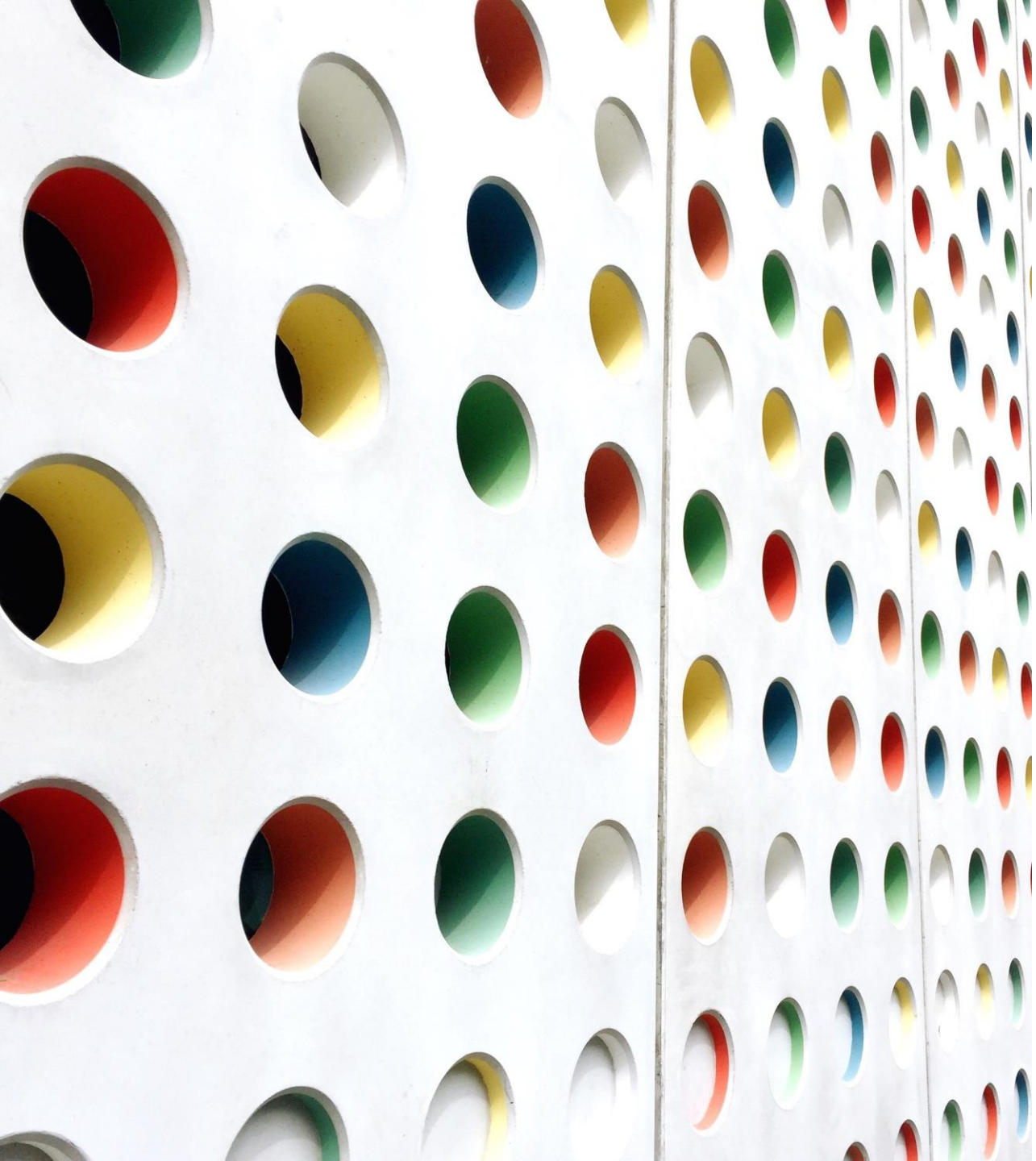


Credit EDA Assignment

SUBMITTED BY

PRIYANKA KUMARI

DS C46

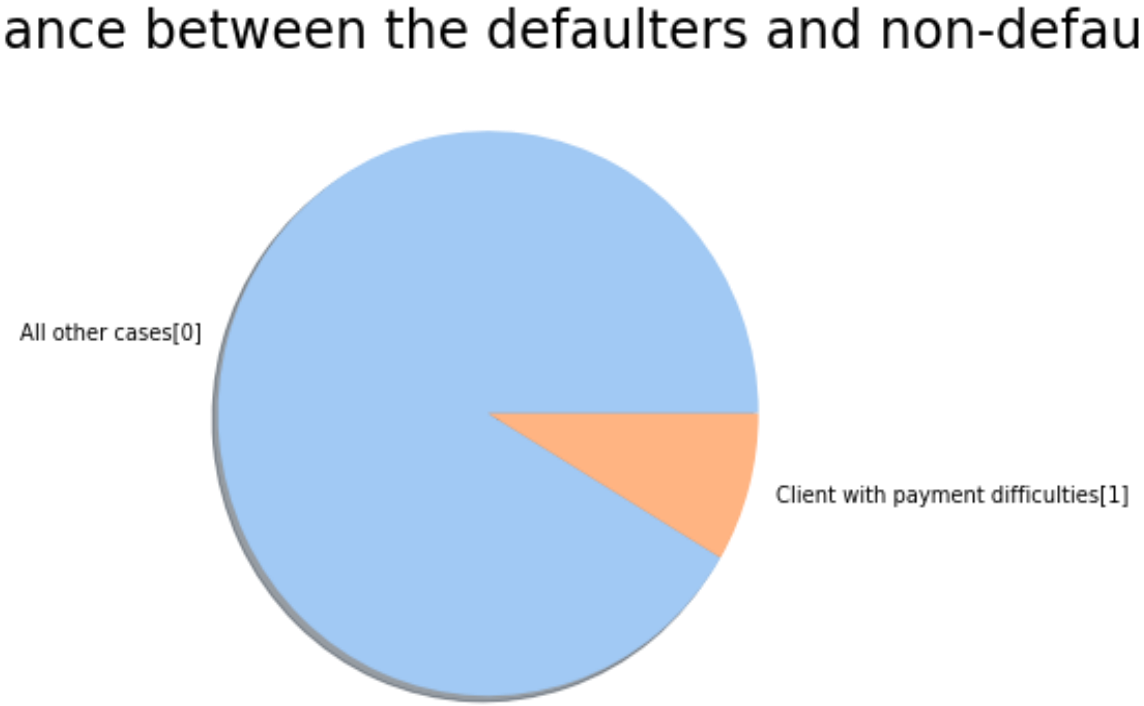
Problem Statement

- This EDA main objective is to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.
- The above mentioned objective has been achieved by performing EDA (Exploratory Data Analysis) on the dataset provided to us. The data has been analyzed and help the bank in noticing the risk which are associated with it. Such risks can be:
- If the applicant is likely to repay the loan, then not approving the loan might result in a loss to the company.
- If the applicant is likely to default and the company approves its application for the loan , then it may lead to huge losses for the company.

Steps Included

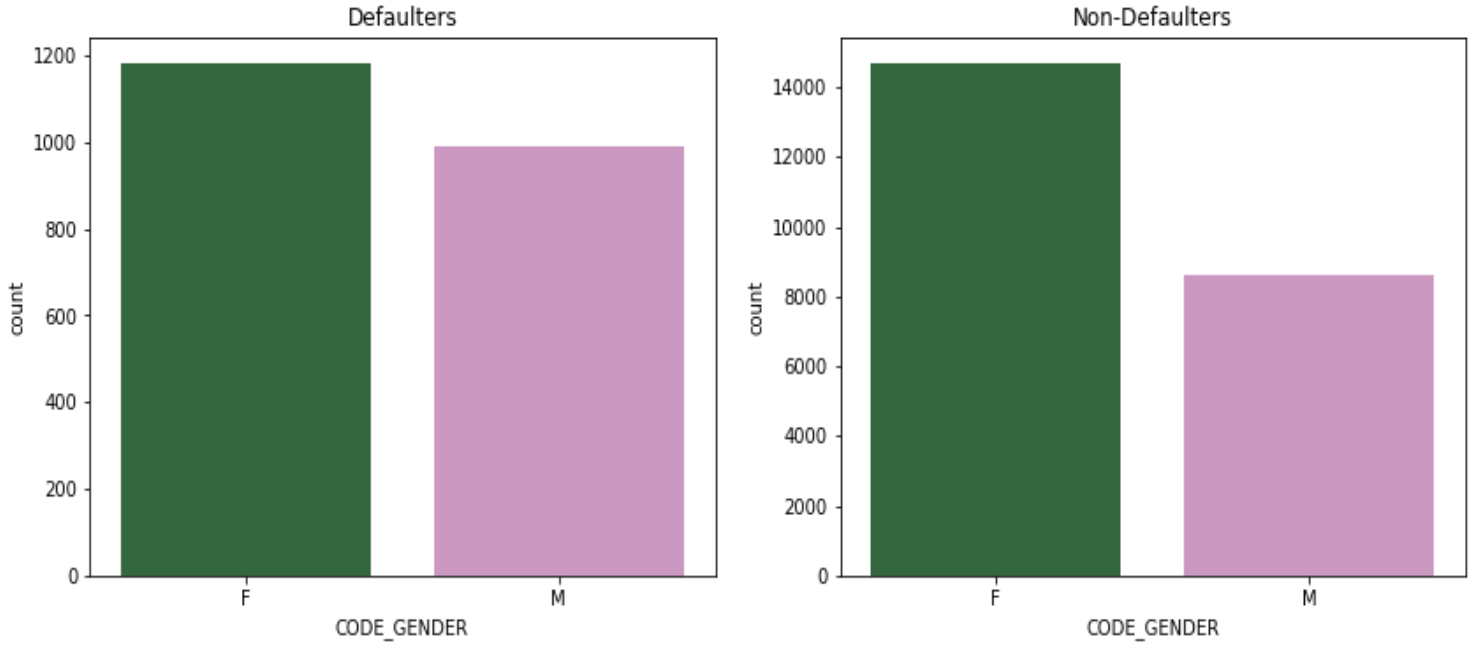
- Data Loading
- Data cleaning
- Handling missing values
- Checking Outliers
- Univariate Analysis
- Bivariate Analysis
- Correlation
- Merging of data set

Current Application



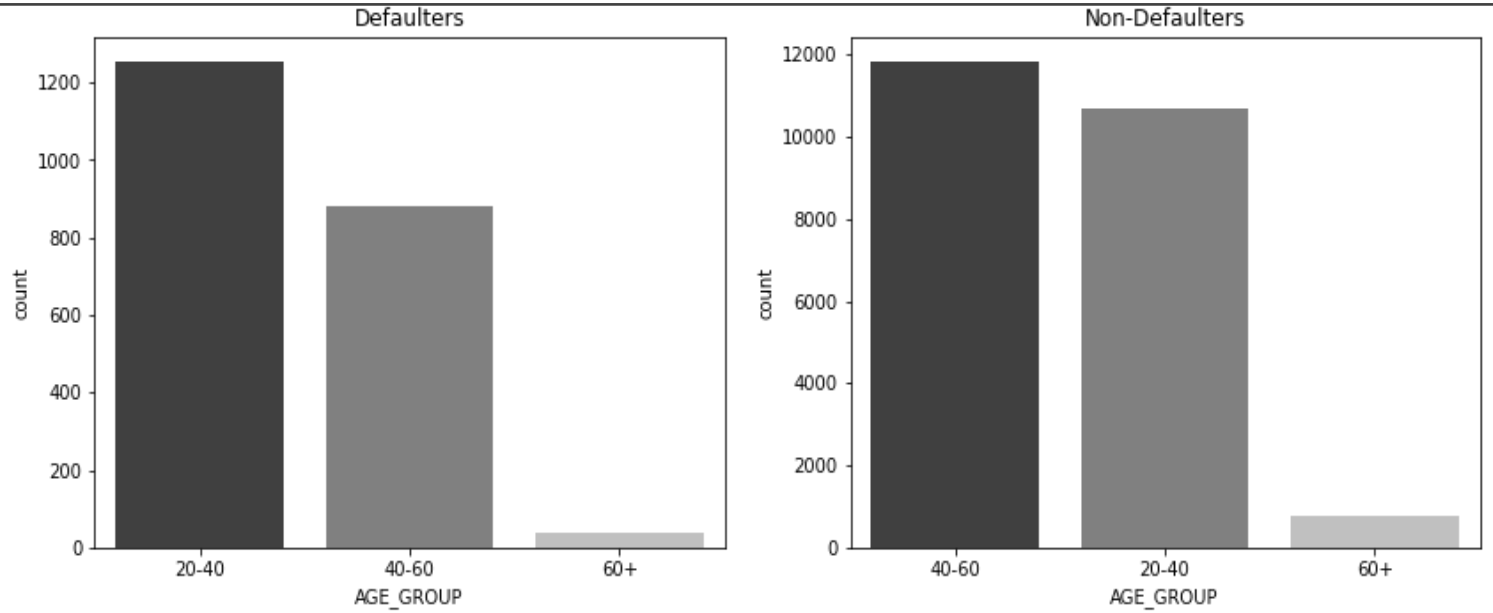
The above analysis shows us that there is a data imbalance between defaulters & non-defaulters.

Univariate Analysis On CODE_GENDER



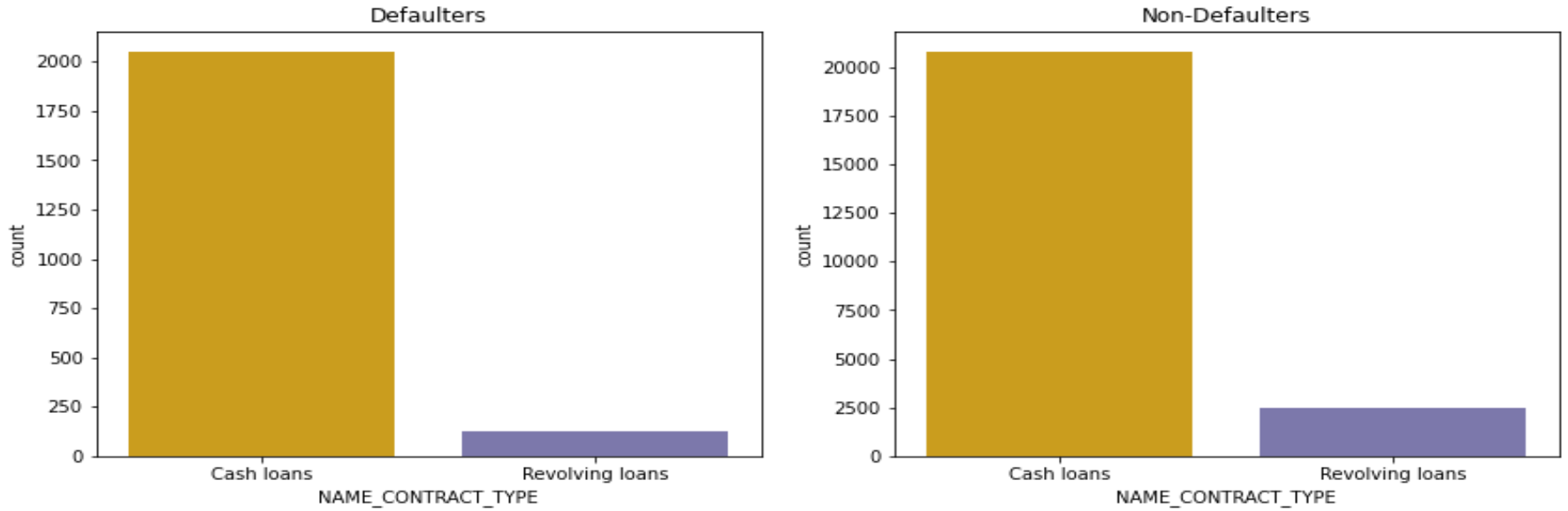
- This count plot depicts the univariate analysis of CODE_GENDER
- This shows that female count is higher as non-defaulters also they are most riskiest to give loans.

Univariate Analysis On AGE_GROUP



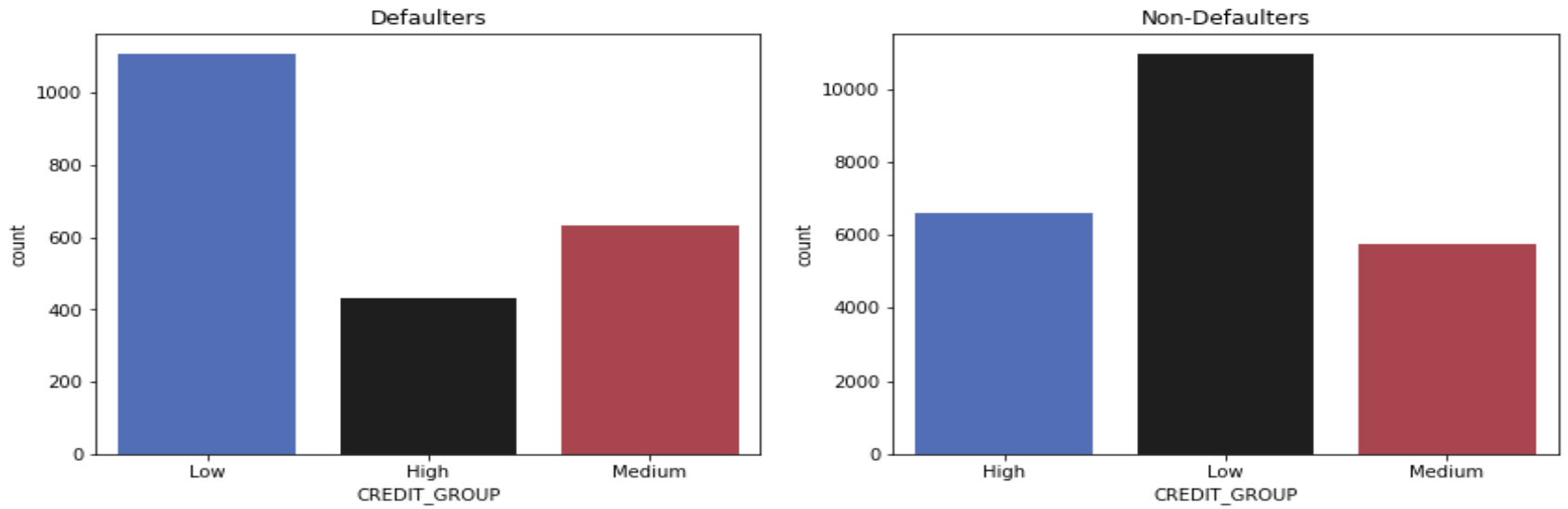
- This count plot above shows the univariate of AGE_GROUP variable.
- Clients in the age group of 40-60 are the clients who should be focused on for giving loans as they are the maximum non-defaulters. Clients above the age of 60+ also should be focused on as they are the least defaulters. Giving out credit to the age group of 20-40 is very risky, they should be given loans but at a very high interest rate

Univariate Analysis On NAME_CONTRACT_TYPE



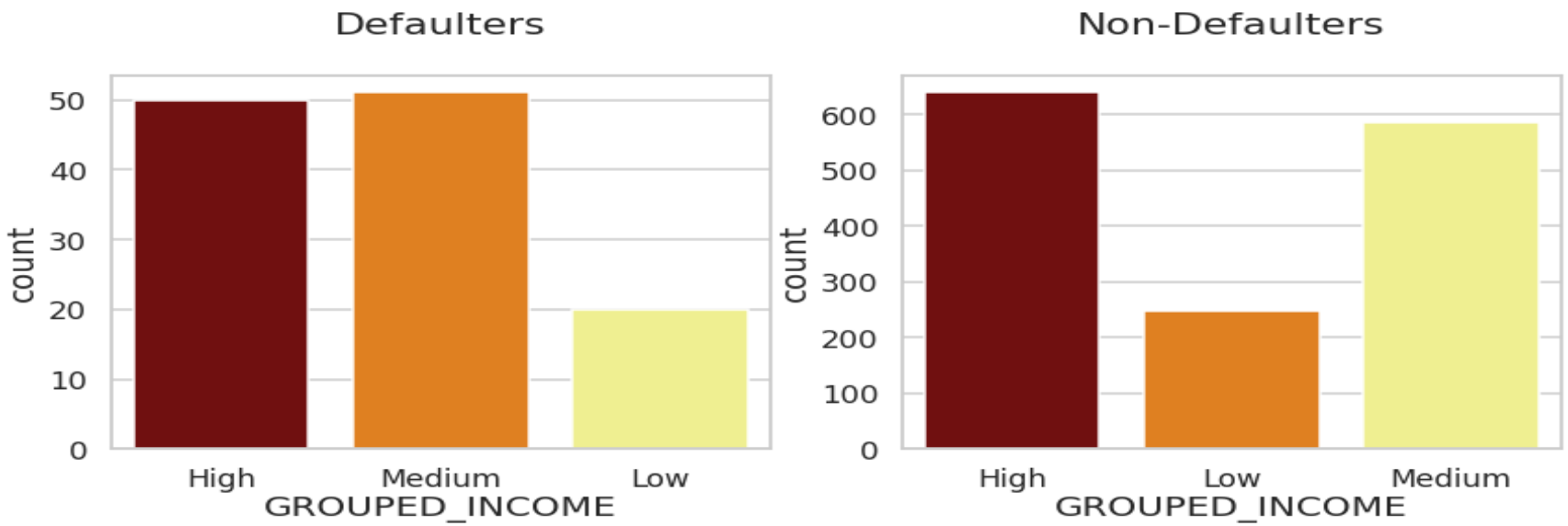
- The above count plot shows the univariate analysis of NAME_CONTRACT_TYPE
- As we can see in both graphs defaulters & non-defaulters make high cash loans in both cases.

Univariate Analysis On CREDIT GROUP



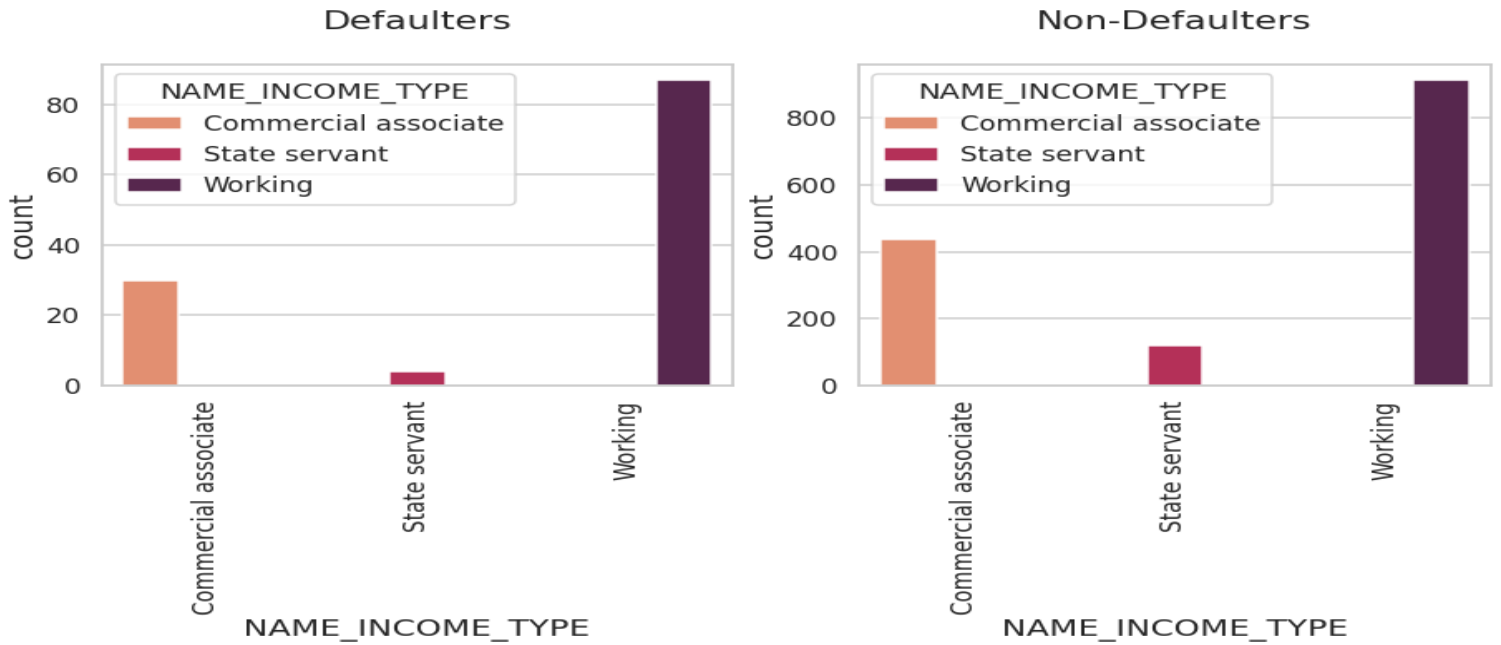
- Low credit group are maximum defaulters
- The trend is the same, low credit amount clients are the maximum non-defaulters too followed by high and medium

Univariate Analysis On Grouped Income



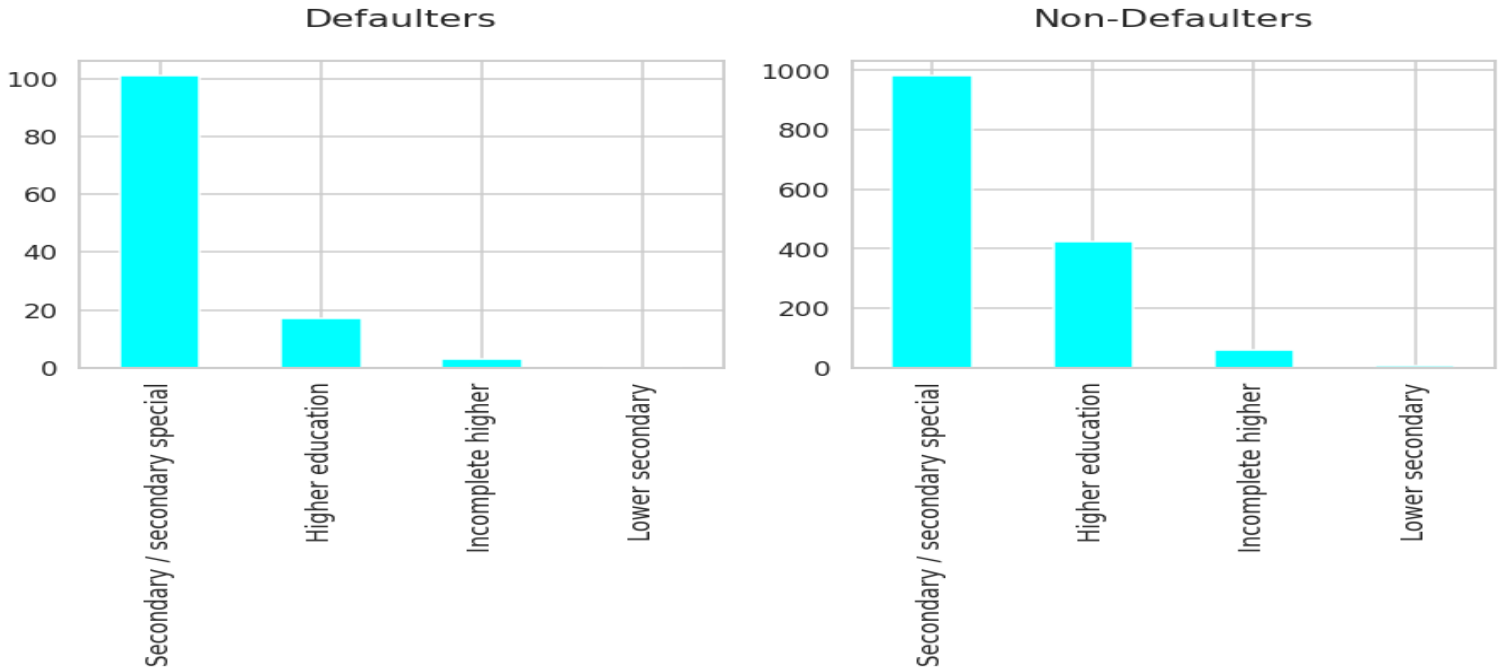
- Giving loans to medium income is riskiest as compared to high & low income groups.
- High income group are non-defaulters & can easily back the loan.

Univariate Analysis On Income Type



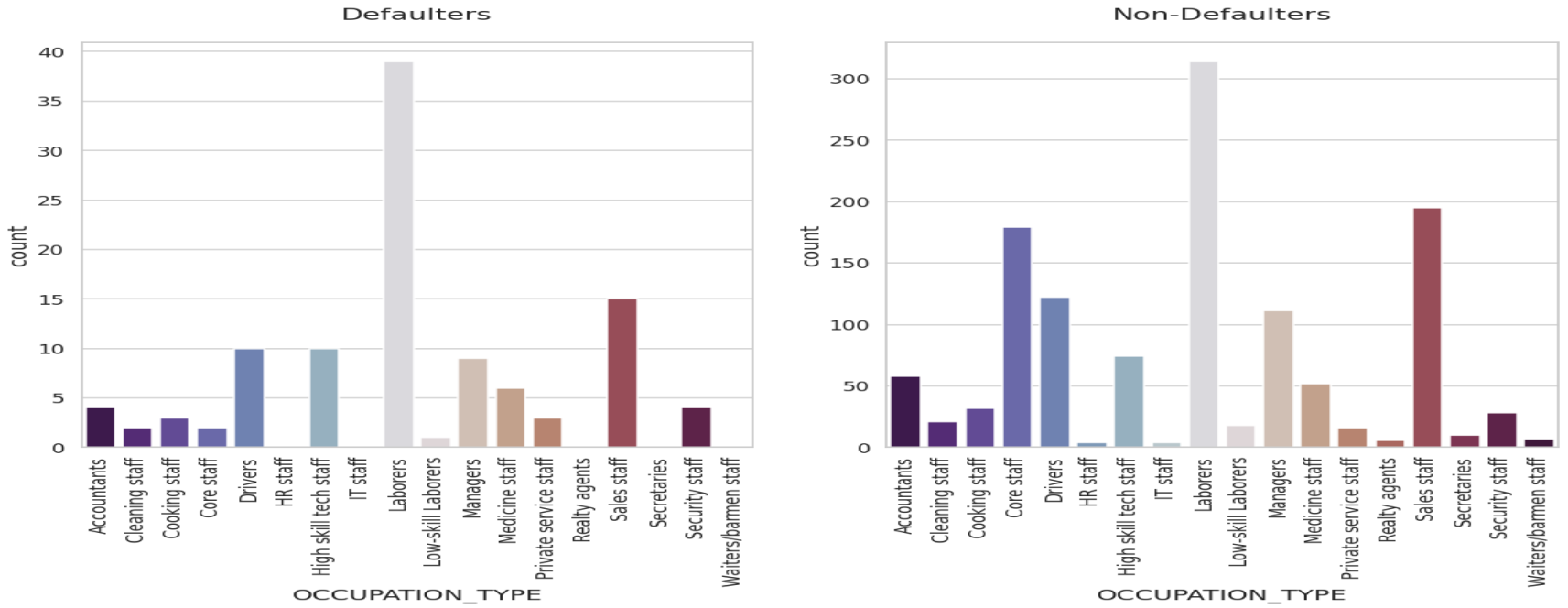
- Working class , commercial associate & state servants are most defaulters
- Working class , commercial associate & state servants are most non-defaulters, as per of their income range.

Univariate Analysis On Education Type



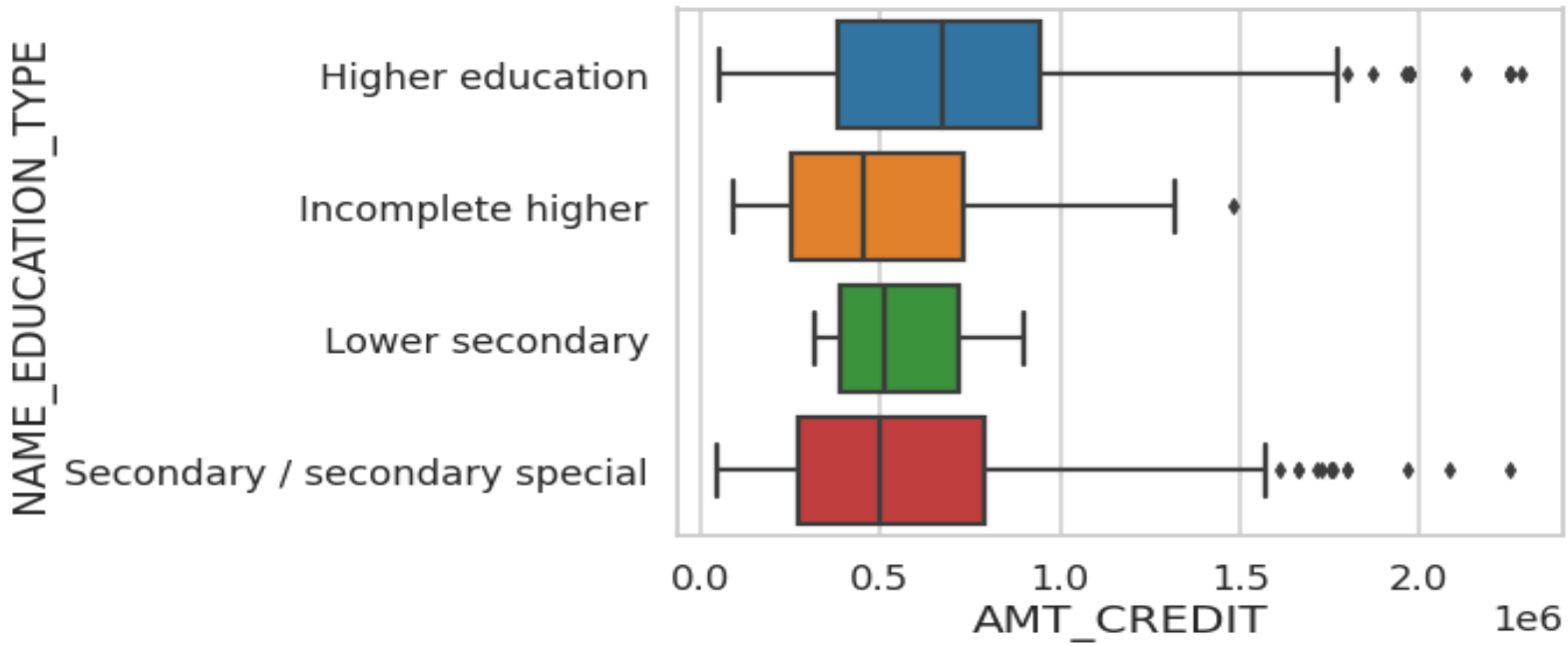
- Secondary/secondary special group of clients are riskier customers to give loans to as they are the most defaulters. Higher education completed clients are less defaulters, thus we can think of giving loans to them.

Univariate Analysis On Education Type



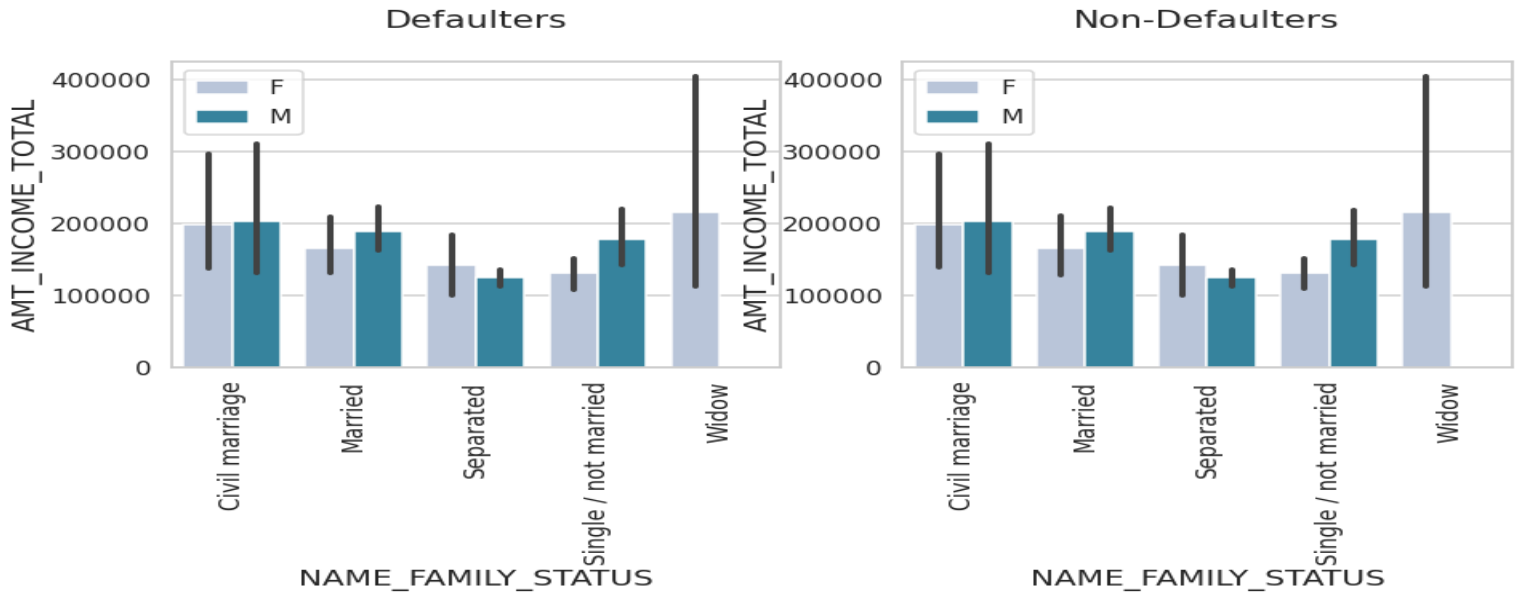
- Labourers are the type which take most loans and as per the count plot, they are the most riskier group to lend loans to followed by sales staff and drivers..

Bivariate Analysis On Amt Credit Vs Education Type



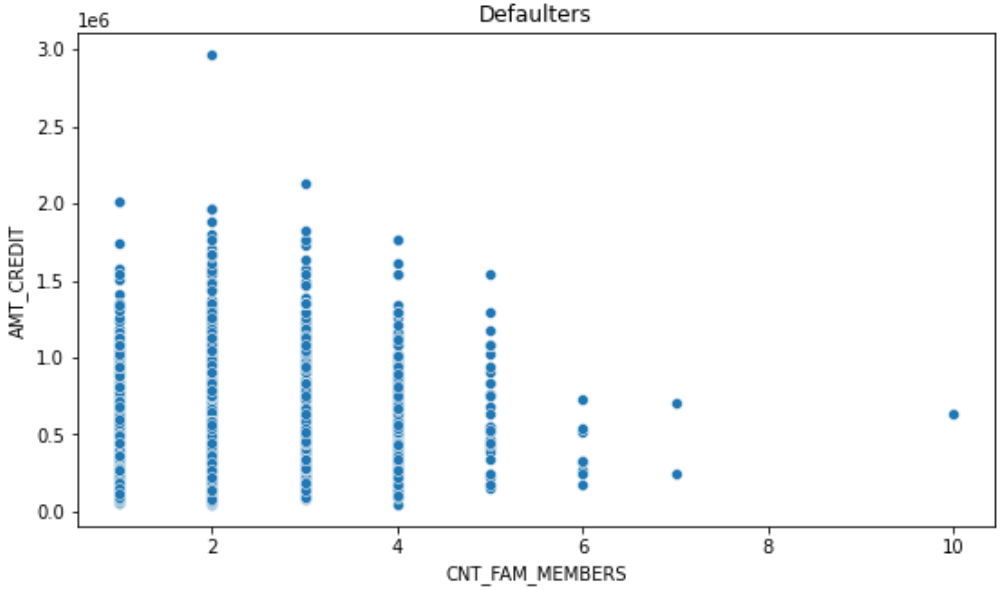
- Client having higher education have more credit than others. Secondary/Special clients have high outliers.

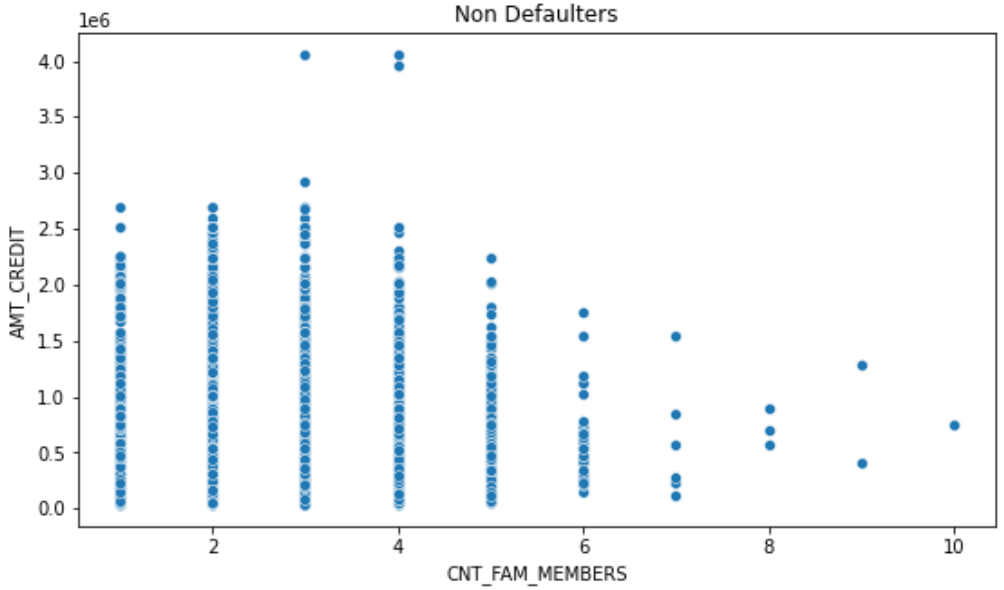
Bivariate Analysis On Income Type Vs Family Status



- Civil marriage males and Married males have the highest income total. Only widowed men and women have same income status. Thus, this group can be focused on when giving credit.

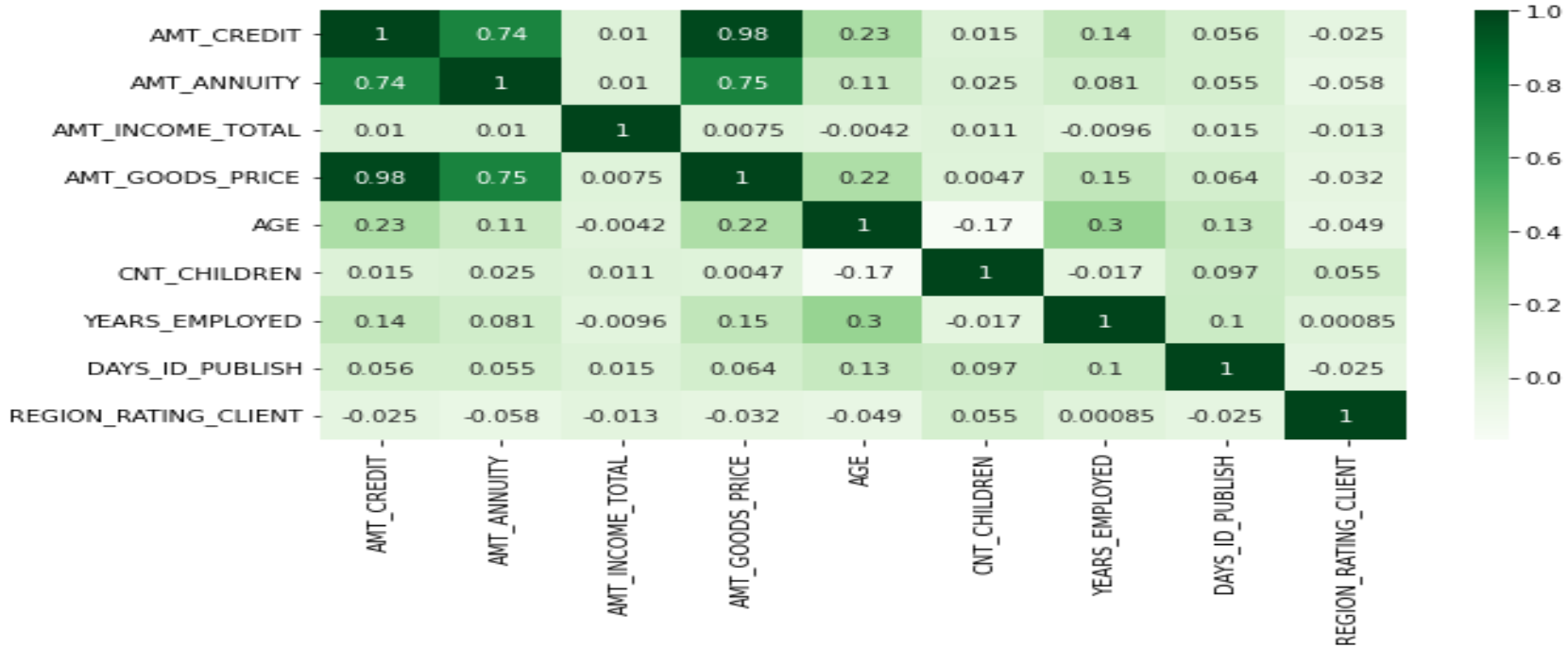
Bivariate Analysis On Fam Members Vs Amt Credit





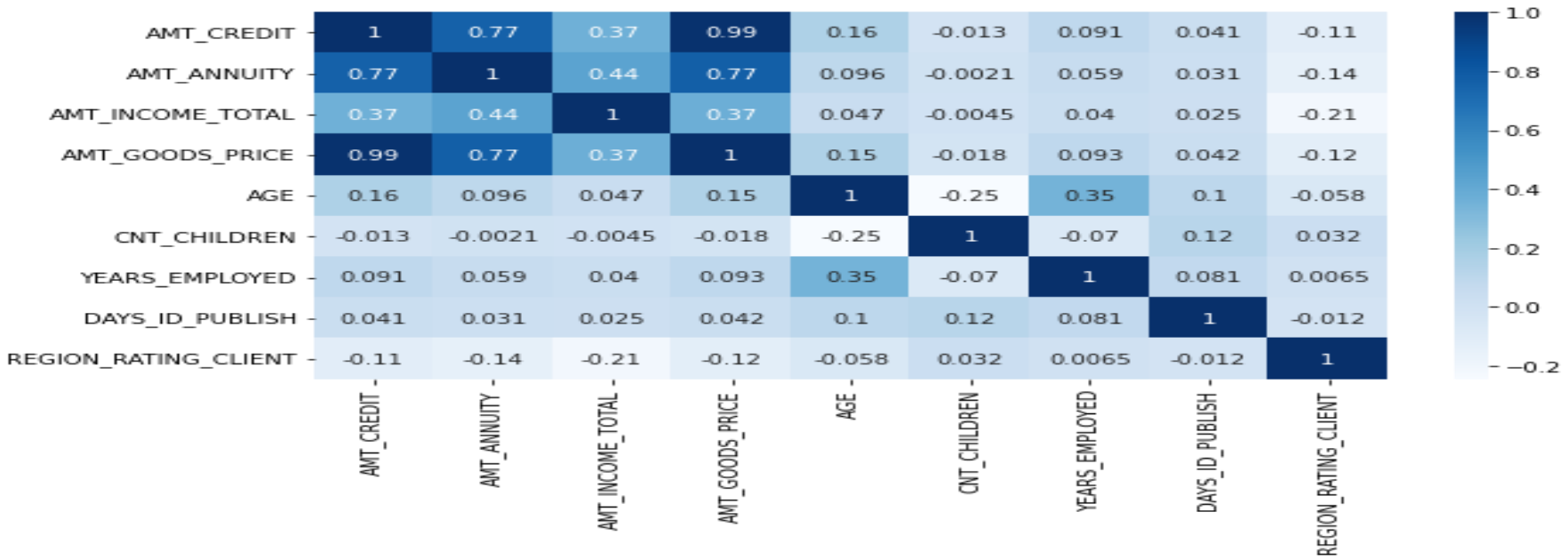
- Scatter plot above shows us that client who have less number of children have high risk of credit loans.

Correlation Of Numeric Columns For Target= 1



- AMT_CREDIT is inversely proportional to the CNT_CHILDREN, means the Credit amount is higher for fewer children count clients have and vice-versa.

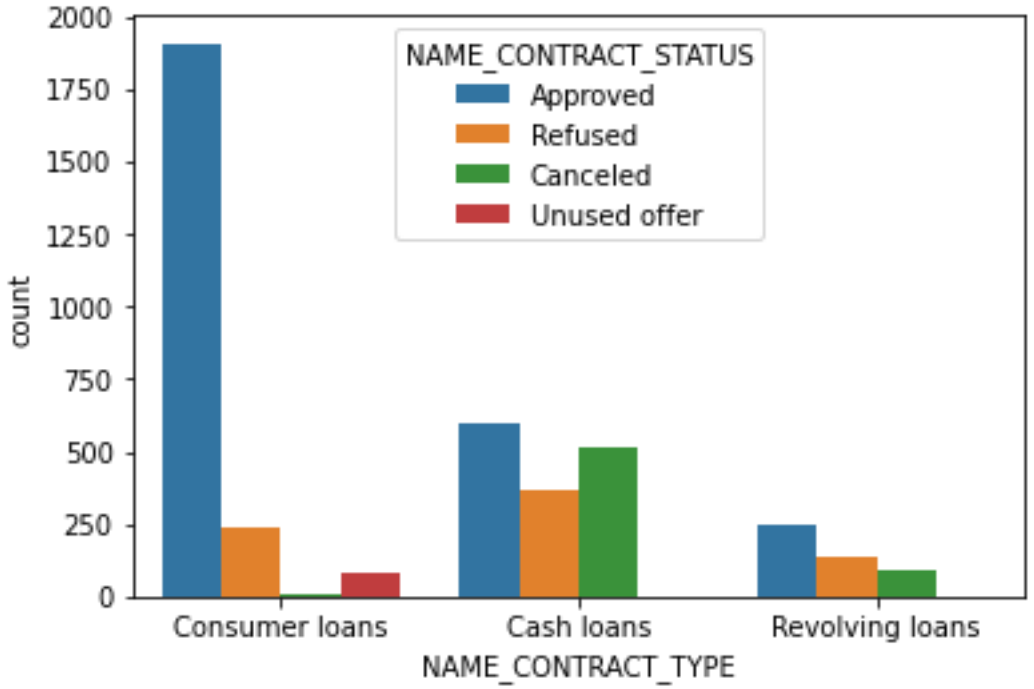
Correlation Of Numeric Columns For Target= 0



- CNT_CHILDREN is positively related to AMT_CREDIT, means the more the CNT_CHILDREN, the more is the AMT_CREDIT.

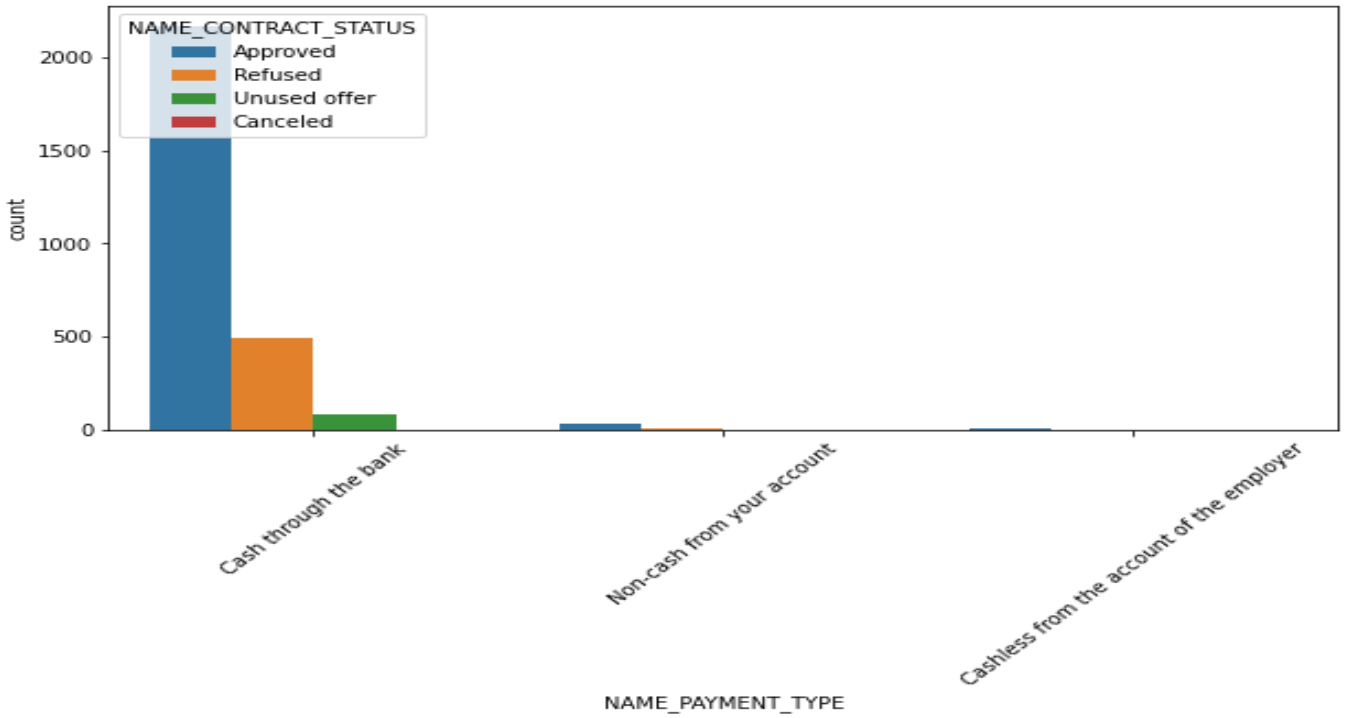
Previous Application Data Analysis

Univariate Analysis On Contract Type



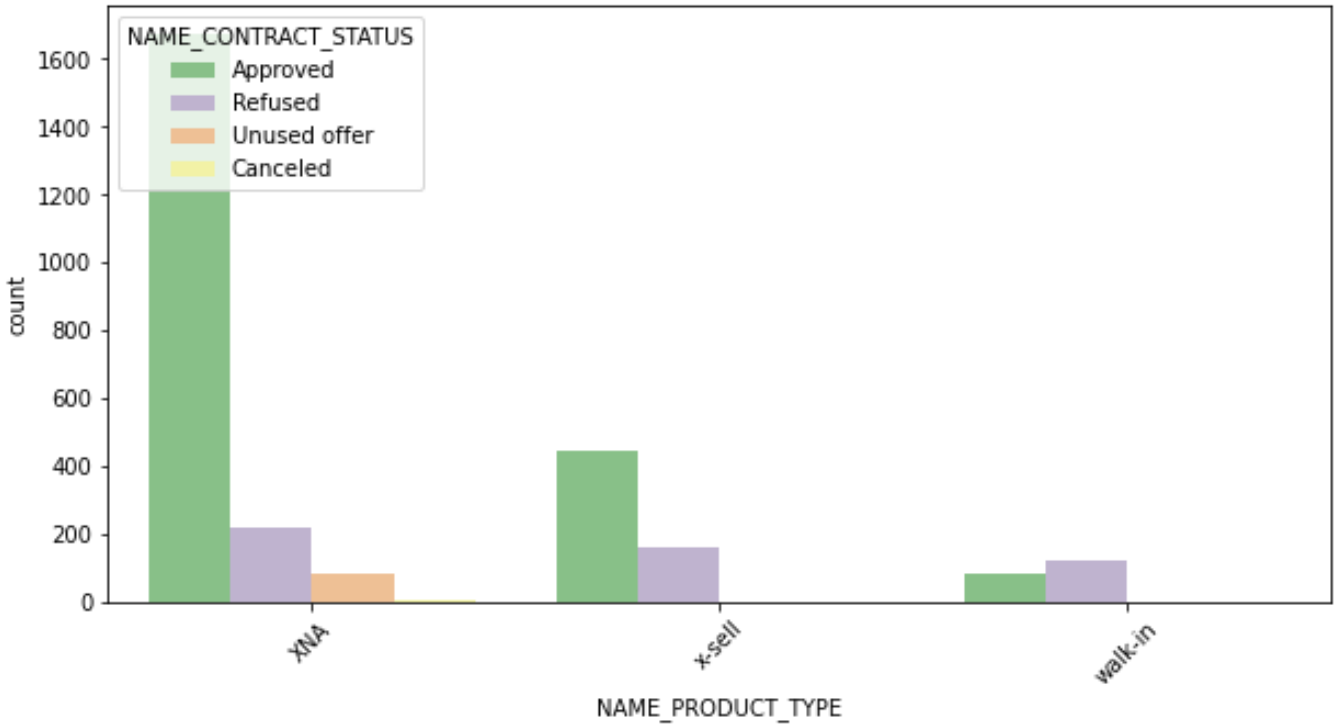
- Approved consumer loans are the highest, followed by cancelled cash loans and refused cash loans. Cash loans have the highest applications followed by consumer loans.

Univariate Analysis On Contract Type



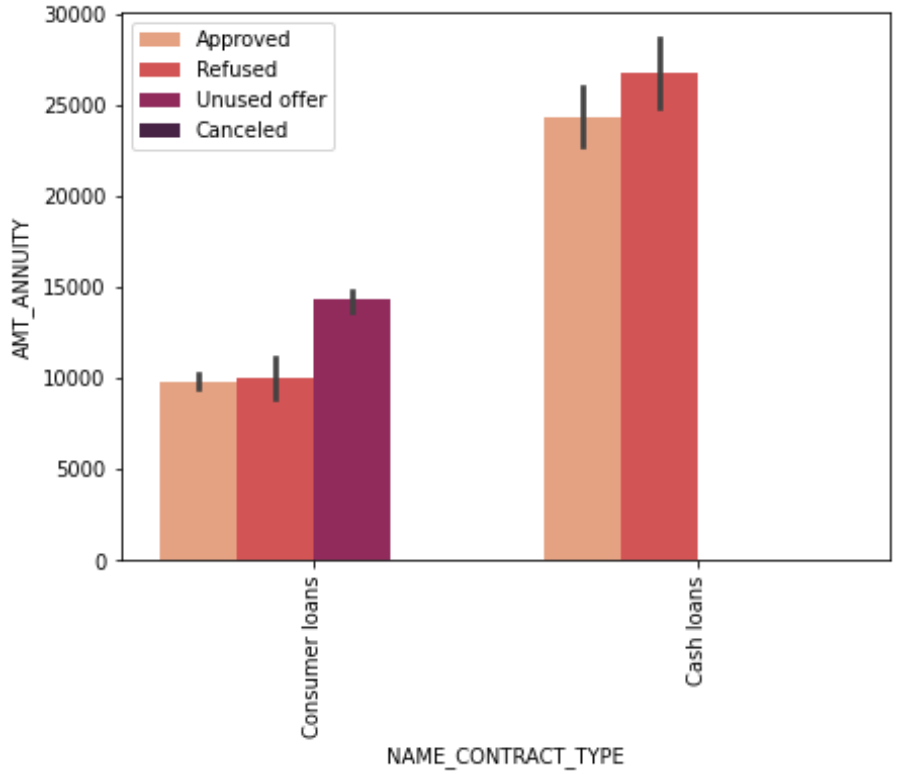
- The most used mode of payment is cash through bank and the least used are non-cash from your account and cashless from the account of the employer.

Univariate Analysis On Product Type



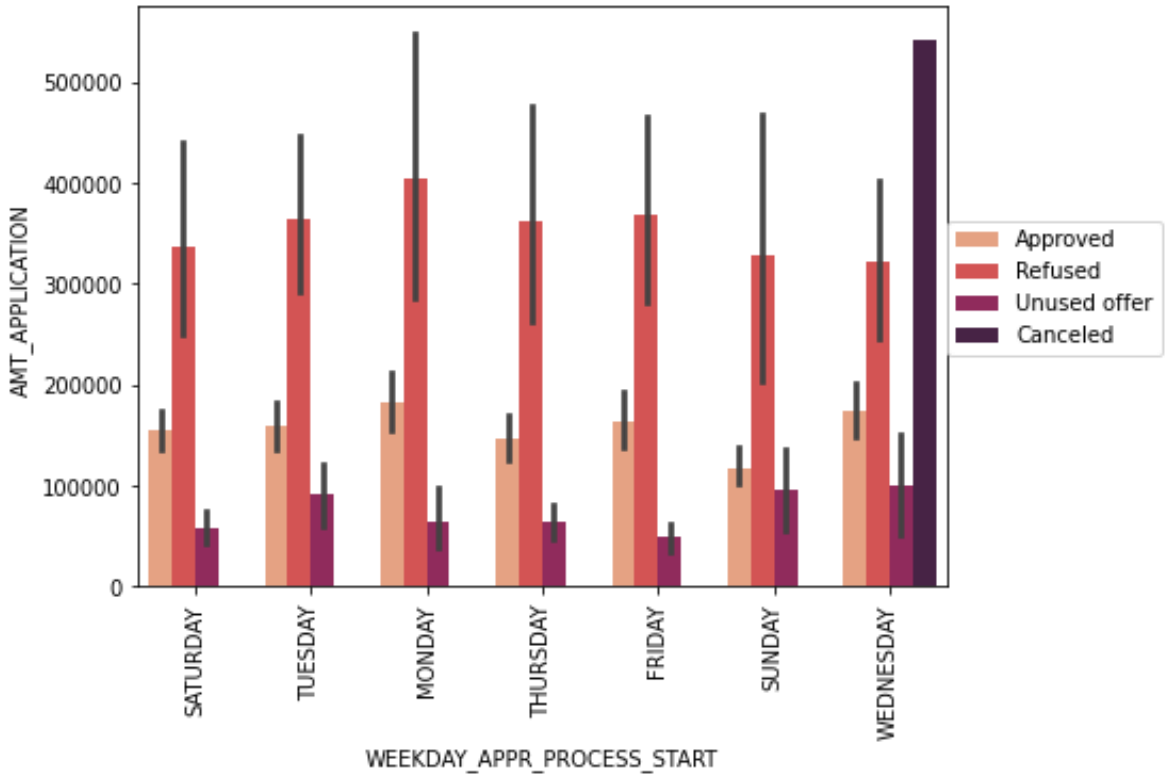
- x-sell applicants requests were approve more then walk-in applicants requests

Univariate Analysis On Amt_Annuity Vs Contract_Type



- As we can see, mostly cash loans were approved, more applications were for cash loans rather than consumer loans.

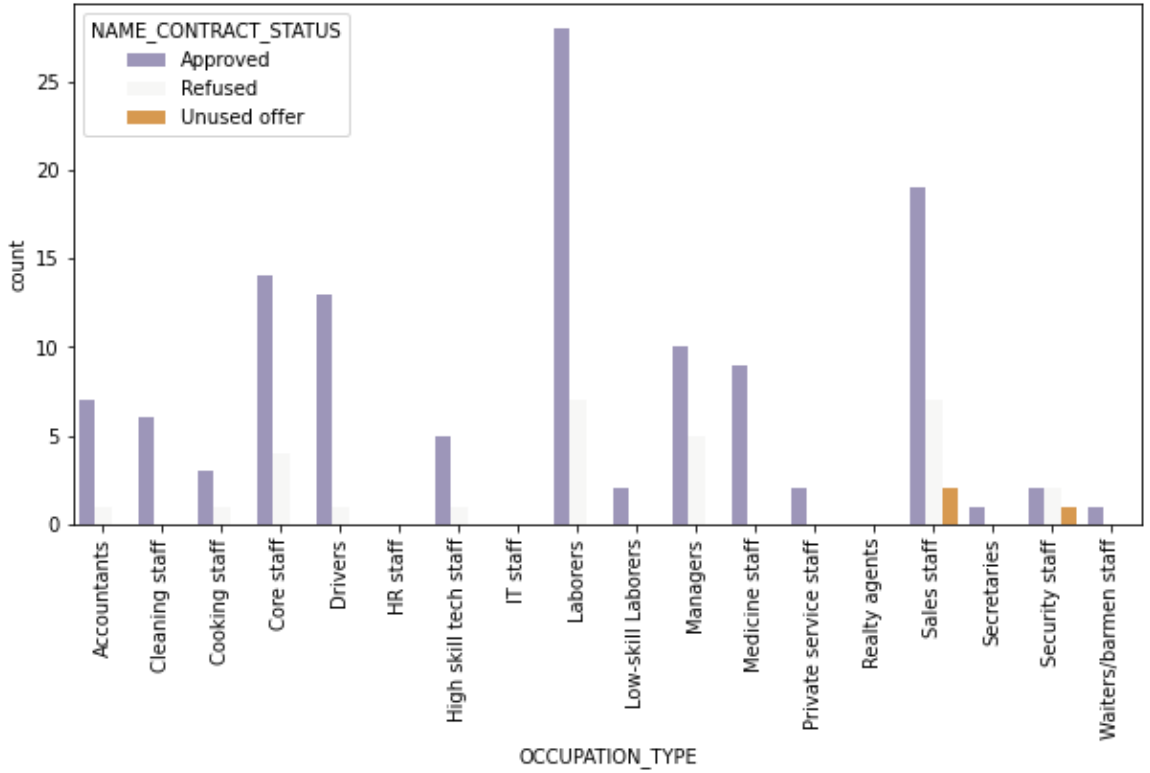
Bivariate Analysis On Weekday_Appr_Process_Start Vs Amt_Application



- Most applications were cancelled on thursdays, most applications were approved on tuesdays and mondays. High amount applications were mostly cancelled

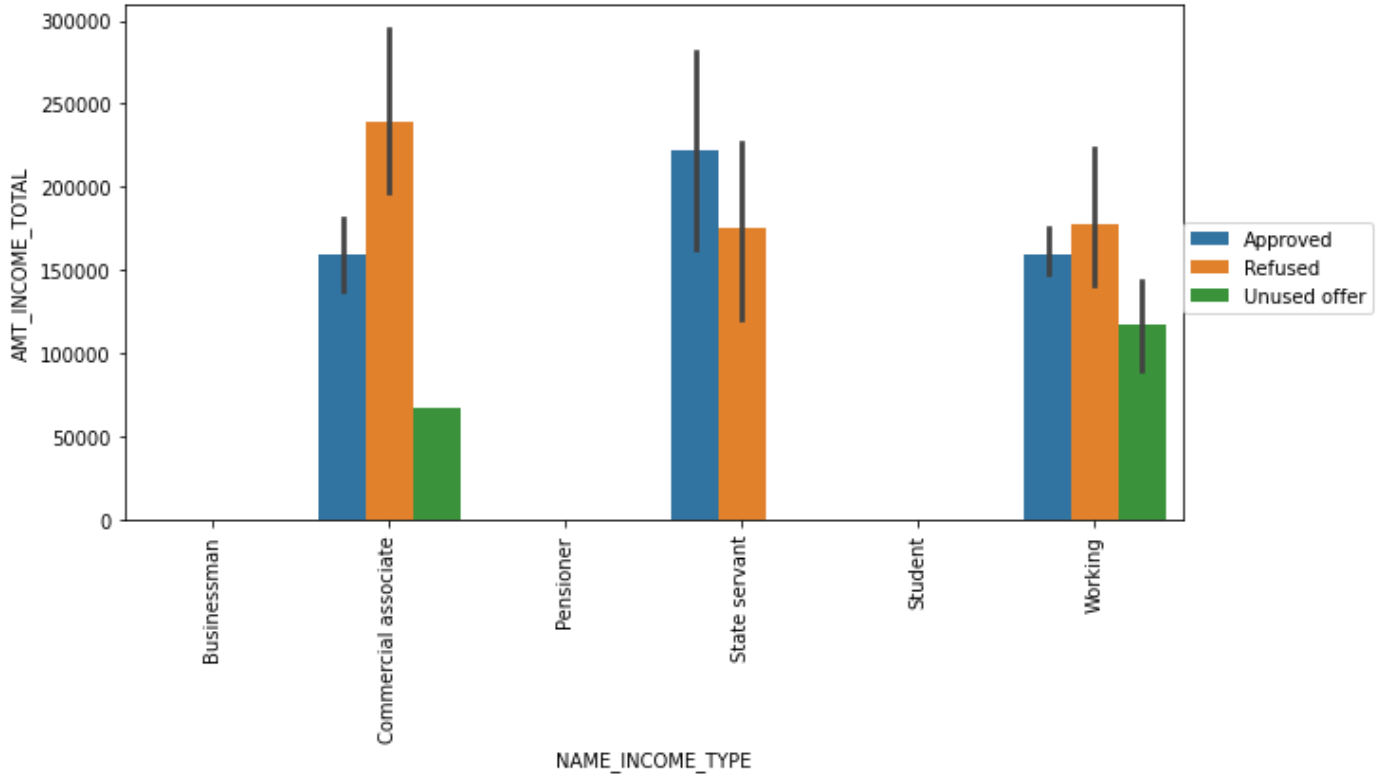
Merged Dataset Analysis

Univariate Analysis On Occupation



- As we can see, loans of laborers were the highest approved followed by sales staff, core staffs and managers.

Bivariate Analysis Of Income Name Vs Income Total



- Students loans were mostly approve and only few were rejected. Pensioner's loans in the income group of 100000-150000 were mostly refused than approved..

Conclusion

After the EDA, the following conclusion can be made:

- Banks should give more revolving loans, as people with revolving loans are less likely to default
  · Consumer loans are more likely to get approved by banks.
- Females are less likely to default, they also provide a good market base for the banks.
- Age is also an important factor, old people and young are less likely to default, so they should be well aware and educated of the various schemes and processes of the bank.
- Education level is a good indicator of a person's ability to pay back the loan. Higher the education ,higher the reliability on the consumer