

EDA CASE STUDY

By Aashka Vijapura and Luqman

Introduction on Credit EDA Case Study

In this case study we are applying the concept of exploratory data analysis on the two given data sets having the data of application of loans done by clients and the data of previous loan applications and the results (default and non-default). We are trying to find the variables which affects the target variable through data visualisation.

Problem Statement on Credit EDA

CASE STUDY

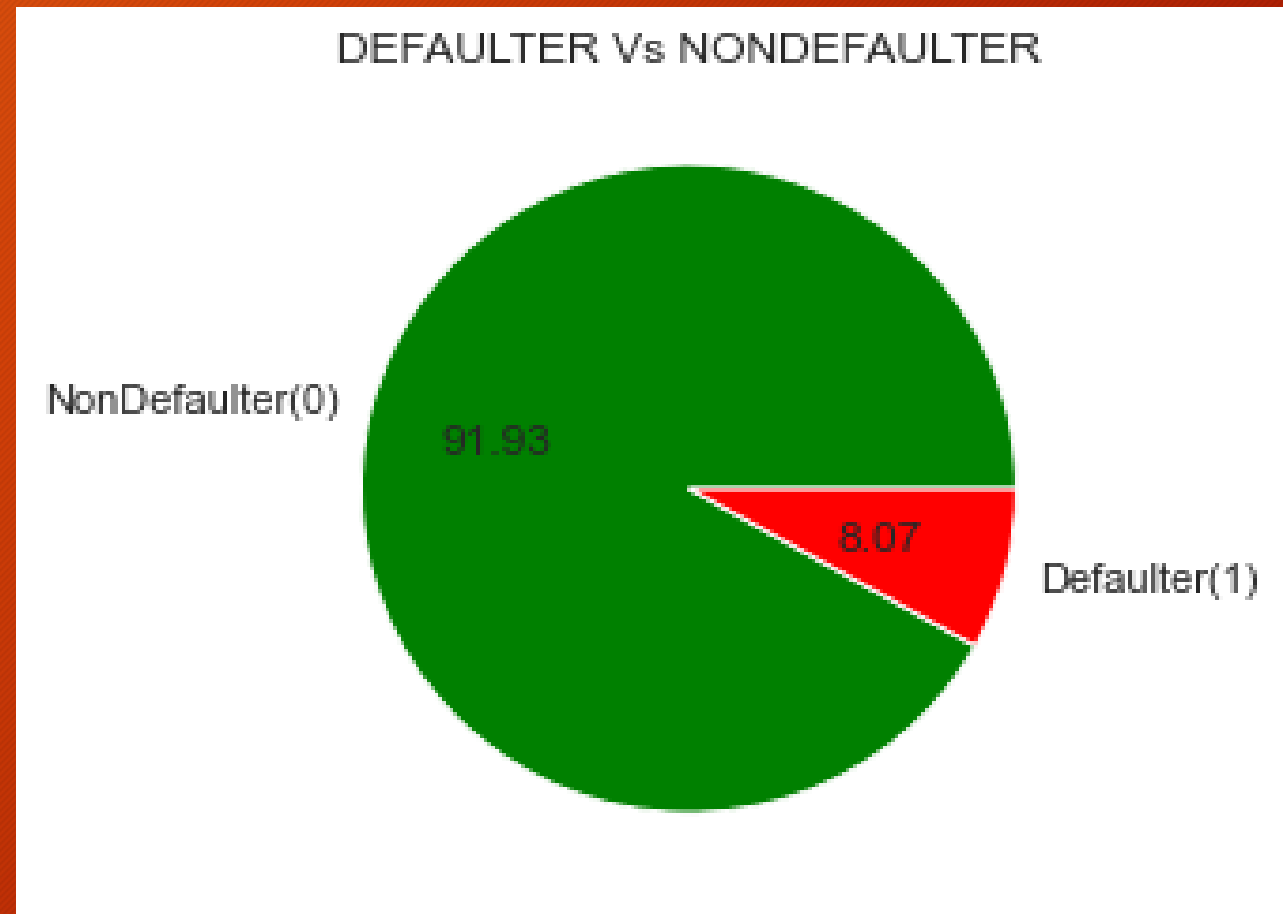
When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:


- 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. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Distribution of values in "Target" column

Inference :

- The Percentage of defaulters are 8% and percentage of non-defaulters are 92%



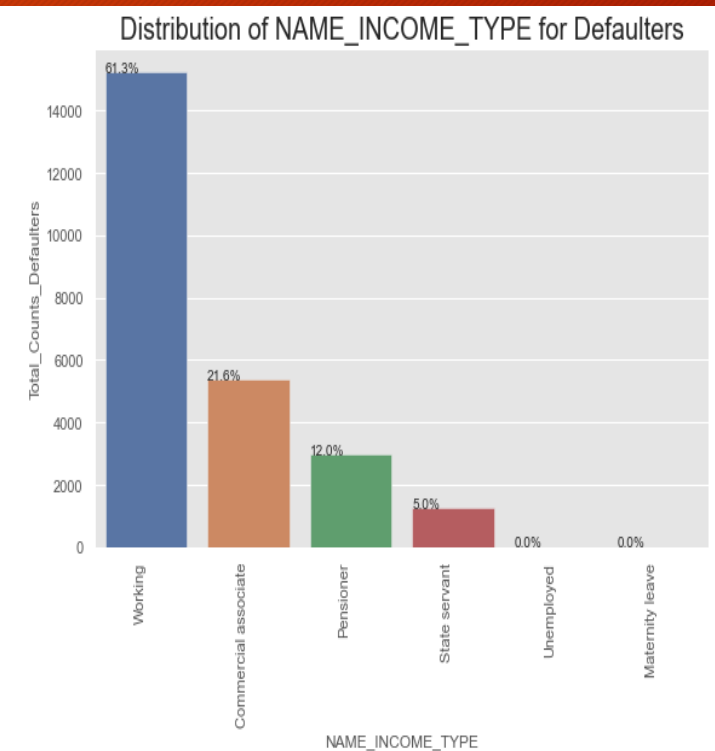
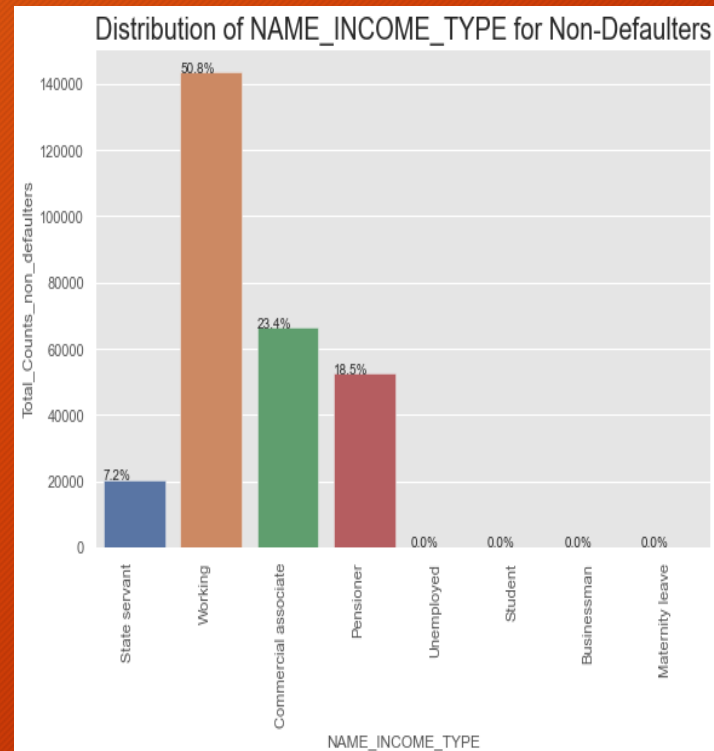


Univariate Analysis on Non-
Default(0) and Default(1) data
frames split based on the target
variable

To Check and Compare defaulters and Non-Defaulters based on the type of income

Inference :

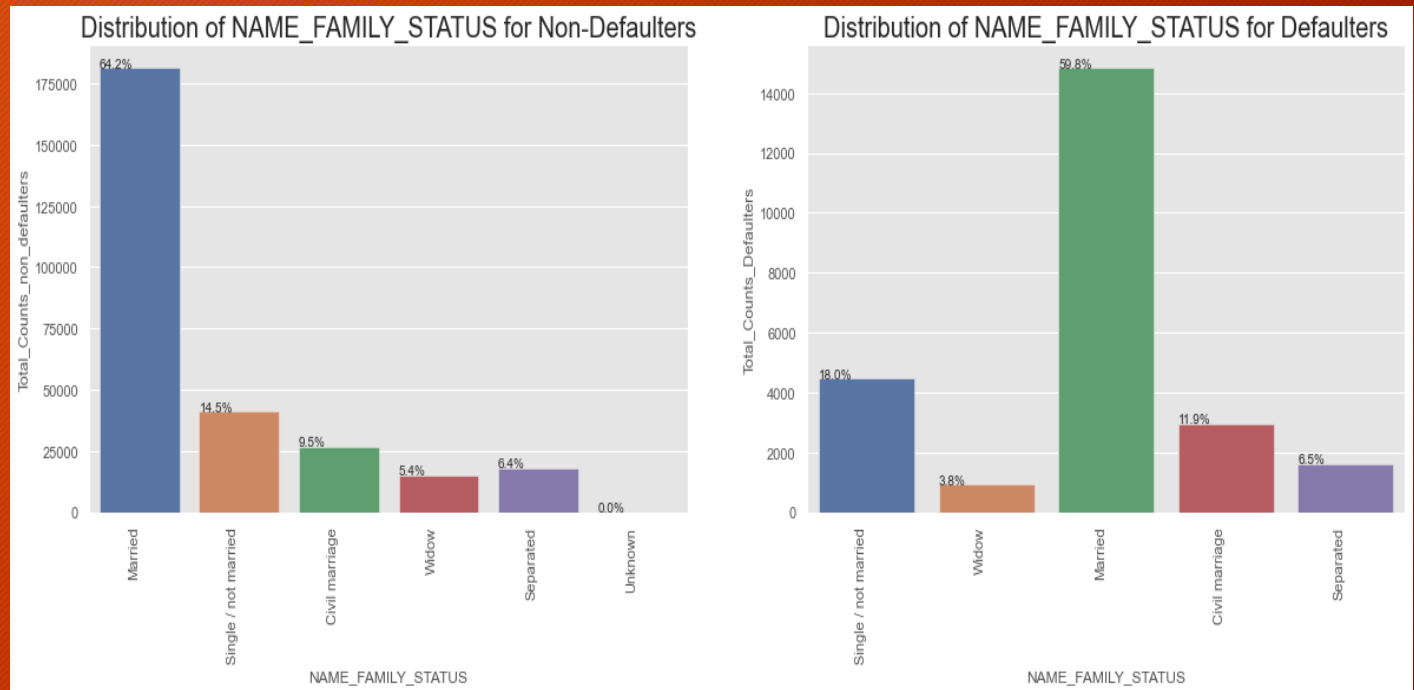
- We can observe from the defaulters graph that the student and businessmen category are not present, hence we can conclude that these two categories do not default.
- Working category contributes equally to defaulters and non-defaulters



To Check and Compare defaulters and Non-Defaulters based on the Family Status

Inference :

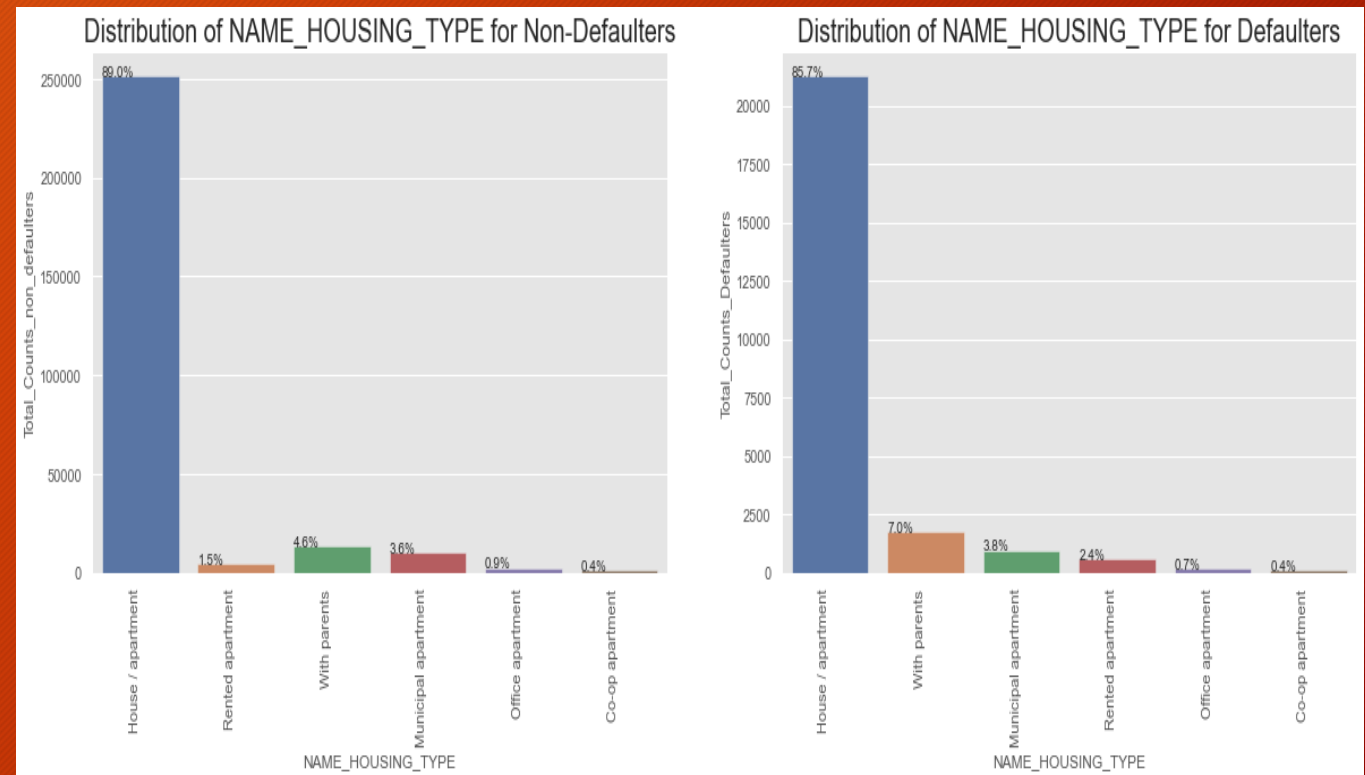
- Single/Unmarried people are the 2nd highest category who are most likely to default compared to other categories.
- Married people applied for most number of loans hence it is high in both defaulter and non-defaulter.



To Check and Compare defaulters and Non-Defaulters based on the Housing Type

Inference :

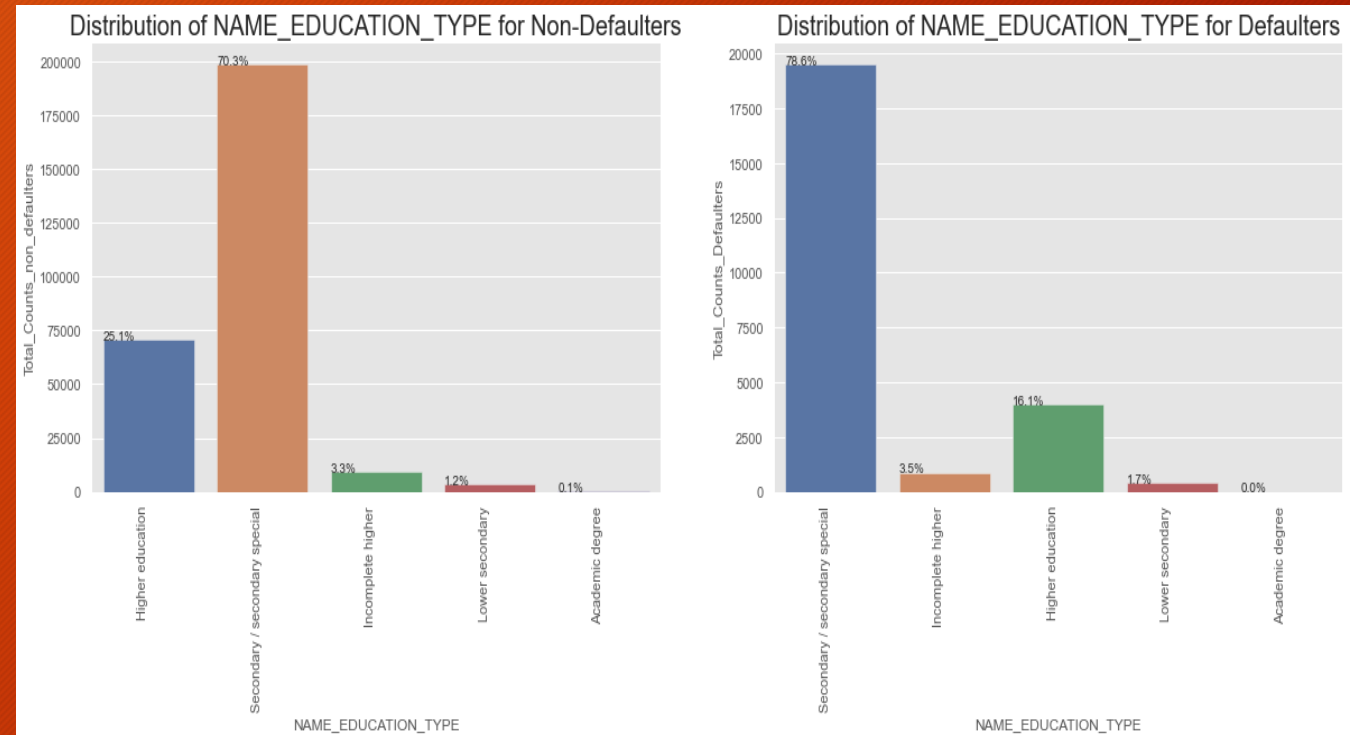
- Most of the People living in apartment and house apply for loan.
- People living with their parents tend to default more.
- People living in municipal apartment and rented apartment are equally likely to default.



To Check and Compare defaulters and Non-Defaulters based on the Education Type

Inference :

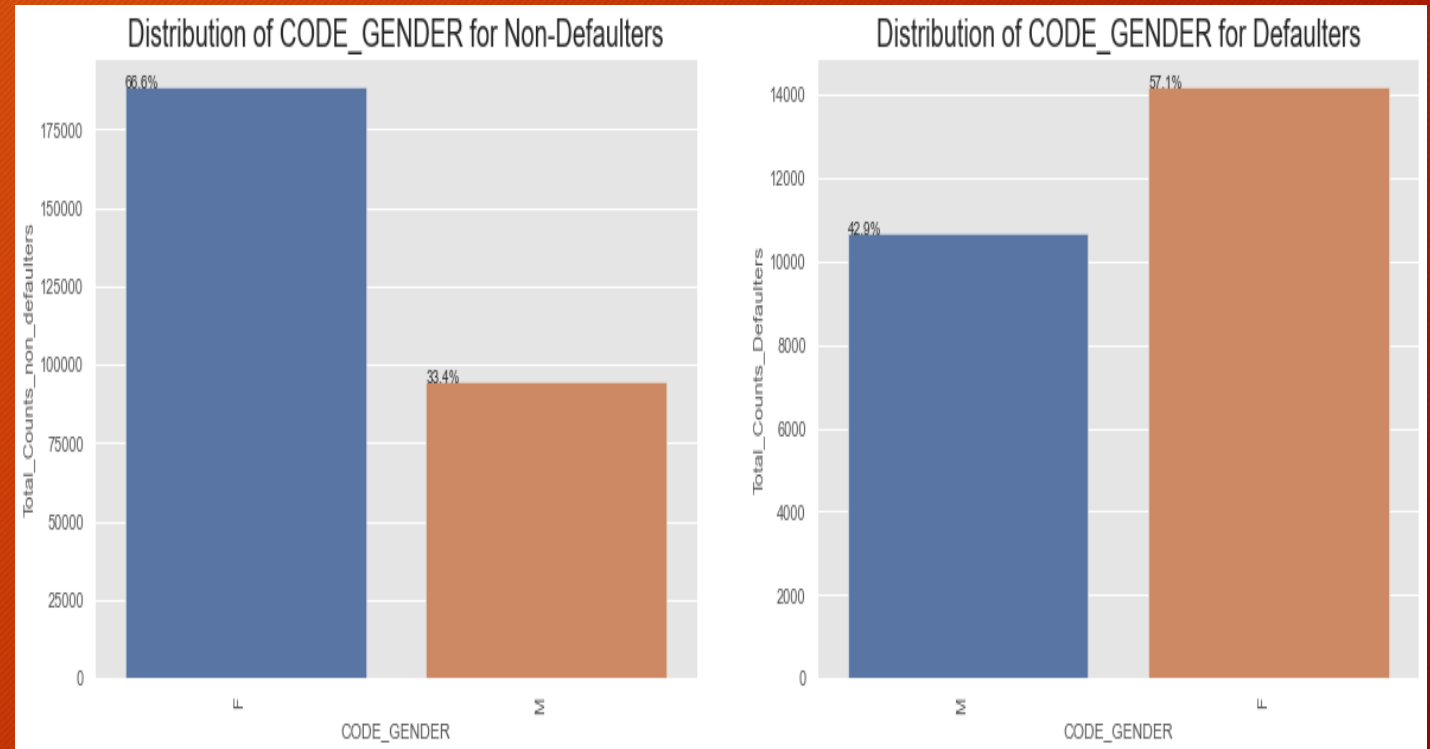
- We can observe from above plotted graph that the people whose highest education is secondary/secondary special have applied for the loan the most.
- People whose highest education is higher education are less likely to default.



To Check and Compare defaulters and Non-Defaulters based on their Gender

Inference :

- Females tend to default the least
- Females also apply for more number of loans than males



To Check and Compare defaulters and Non-Defaulters based on their Region Rating

Inference :

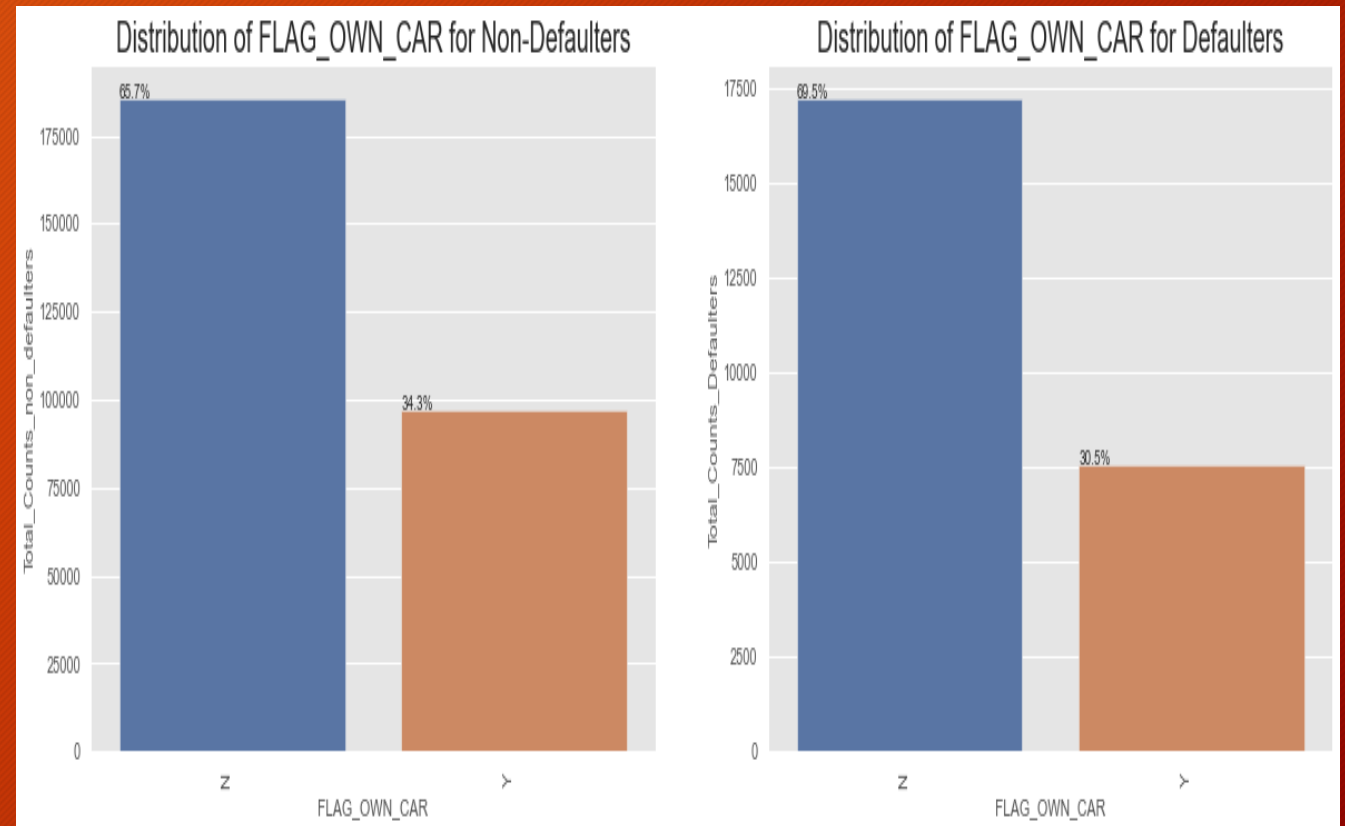
- People from tier 3 city are mostly likely to default more due to either low income or unemployment.
- We can observe that the majority of the loan applicants are from Tier 2 City.



To Check and Compare defaulters and Non-Defaulters based on their ownership of car

Inference :

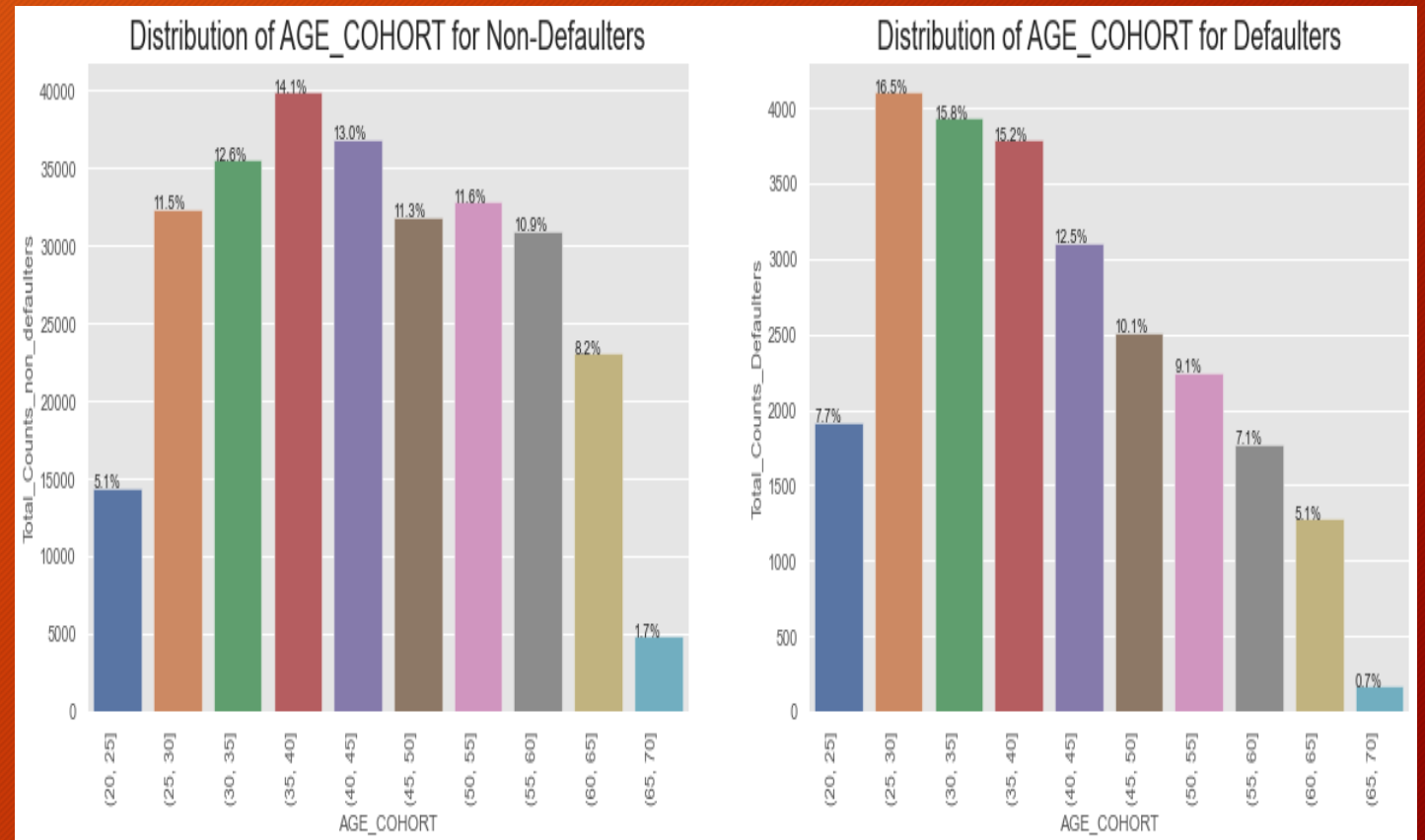
- We can observe from the plotted graphs that the non-defaulters owning a car are above 175000 and the people owning a car being a defaulter are less than 17500 and hence we can conclude that the more number of non-defaulters own a car.



To Check and Compare defaulters and Non-Defaulters based on their Age Group

Inference :

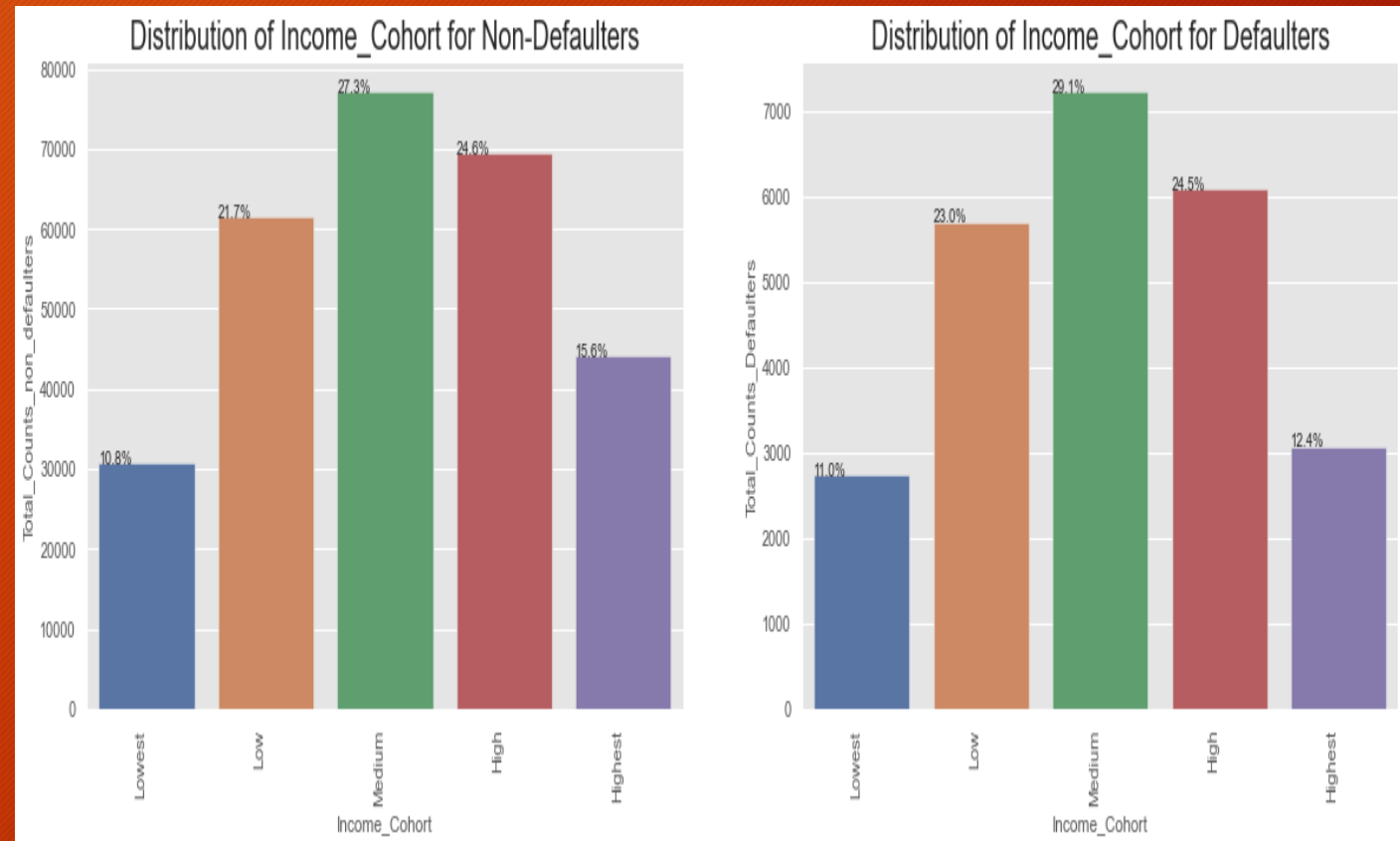
- We can observe from the defaulters graph that as the age group increases the number of people who default decreases.
- People of age group 25 to 30 are most likely to default cause they might be unemployed .



To Check and Compare defaulters and Non-Defaulters based on their Income Group

Inference :

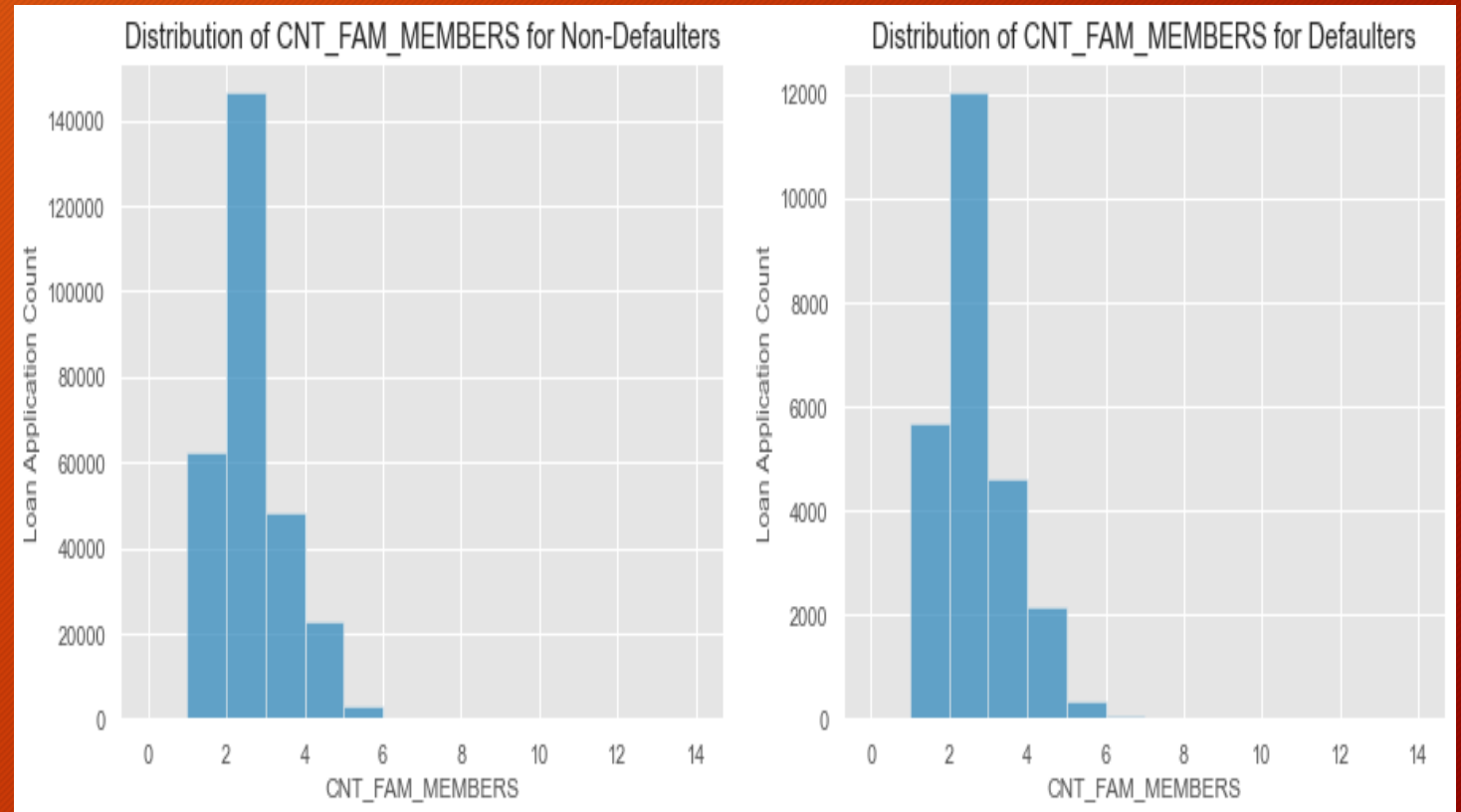
- People with lowest income tend to default the least.
- People having medium income tend to apply for loan the most.



To Check and Compare defaulters and Non-Defaulters based on their Income Group

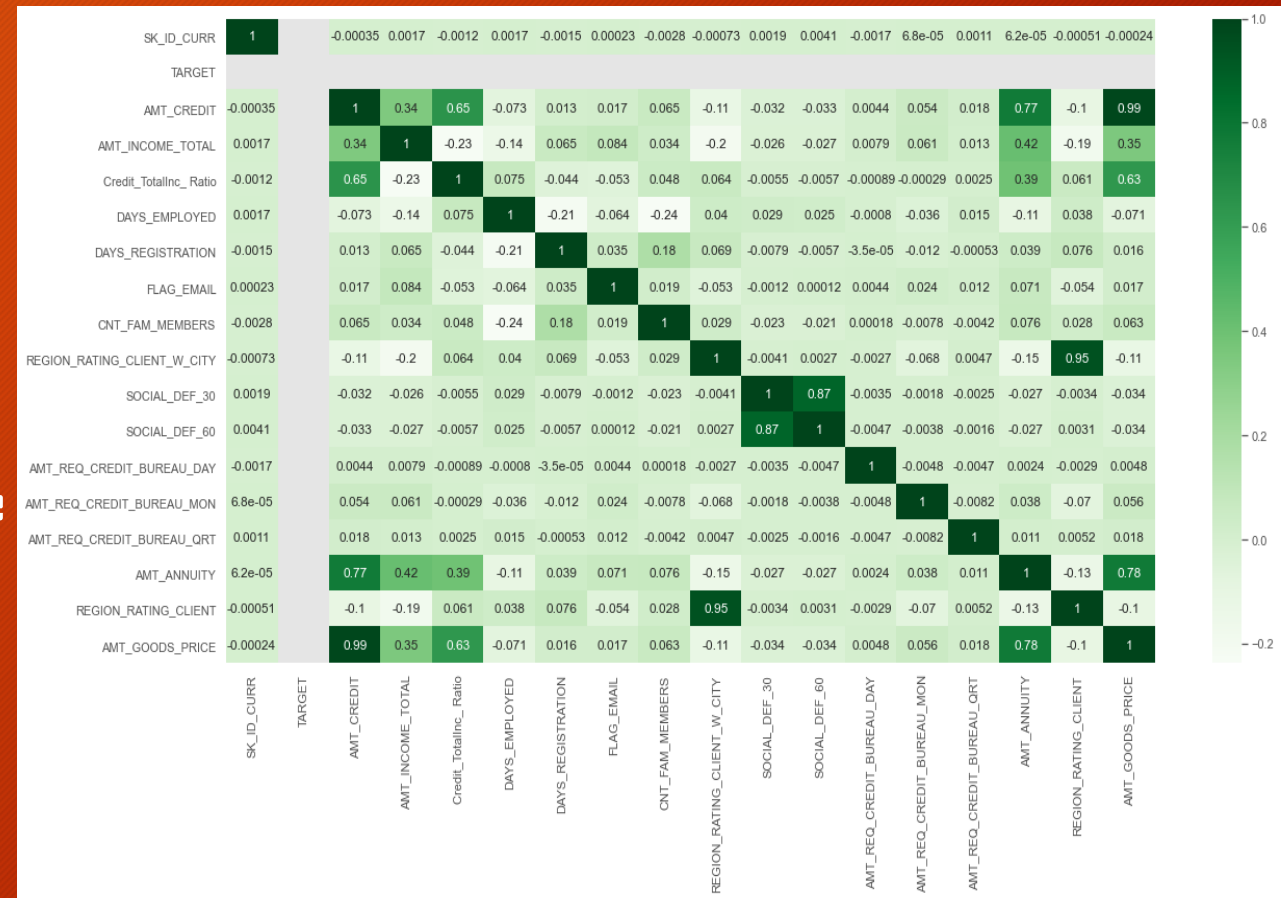
Inference :

- So from the graph we can say that the family of 3 members are most likely to apply for loan compared to other number of family members.



Checking Co-relation in Non-Default Data frame

Co-relations in Non-Default Data frame



Inference :

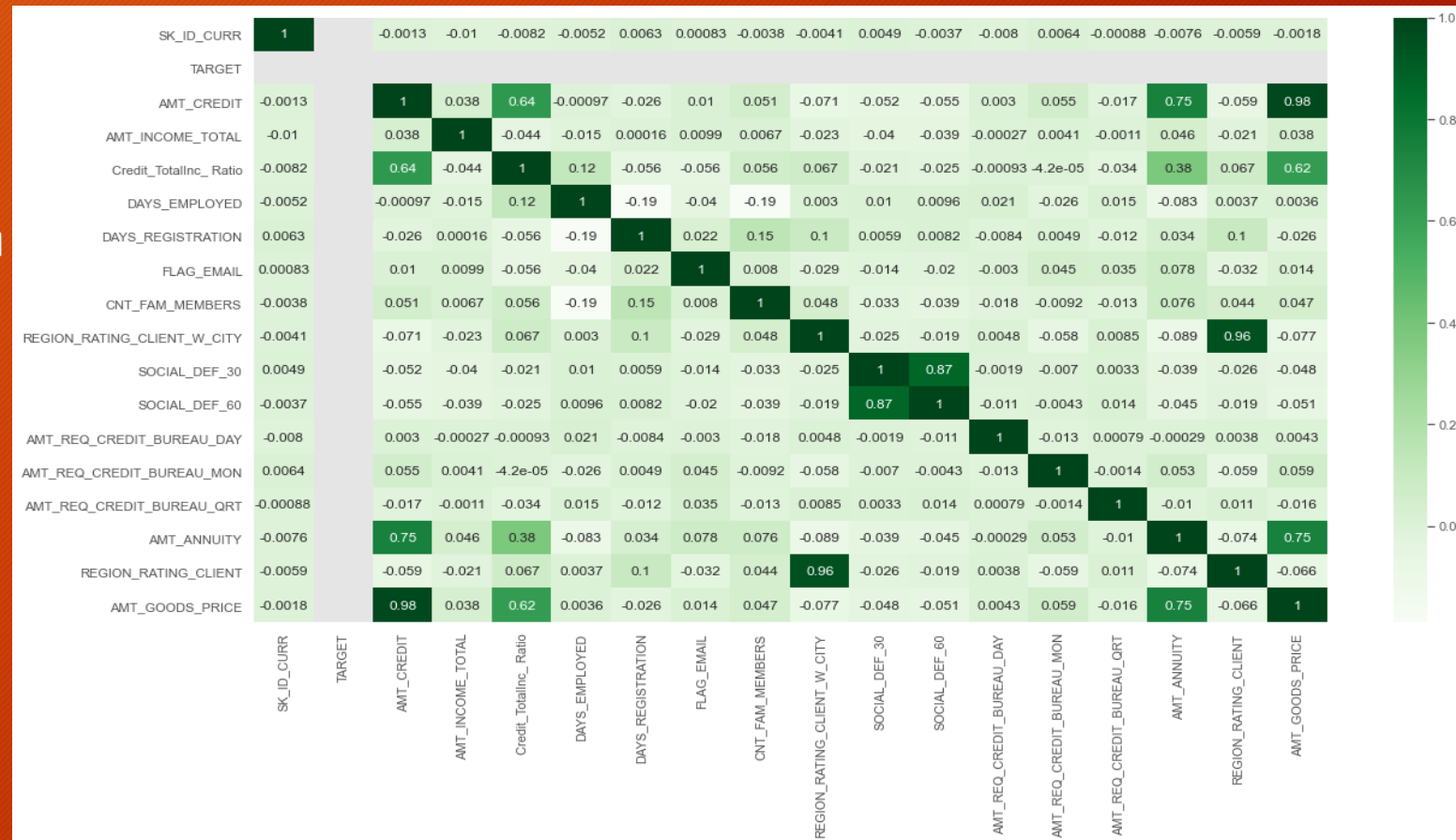
- We can observe from the above table that there is high co-relation between AMT_GOODS_PRICE and AMT_CREDIT , as the goods price increases the loan amount credited for those goods also increases.
- There is very low co-relation between AMT_GOODS_PRICE and AMT_INCOME_TOTAL , as income total does not depend on goods price or vice-versa.

Checking Co-relation in Default Data frame

Co-relations in Default Data frame

Inference :

- We can observe from the above table that there is high co-relation between AMT_GOODS_PRICE and AMT_CREDIT, as the goods price increases the loan amount credited for those goods also increases.
- There is very low co-relation between CNT_FAM_MEMBERS and DAYS_EMPLOYED, as there is no relation between number of days employed and number of family members the client has.

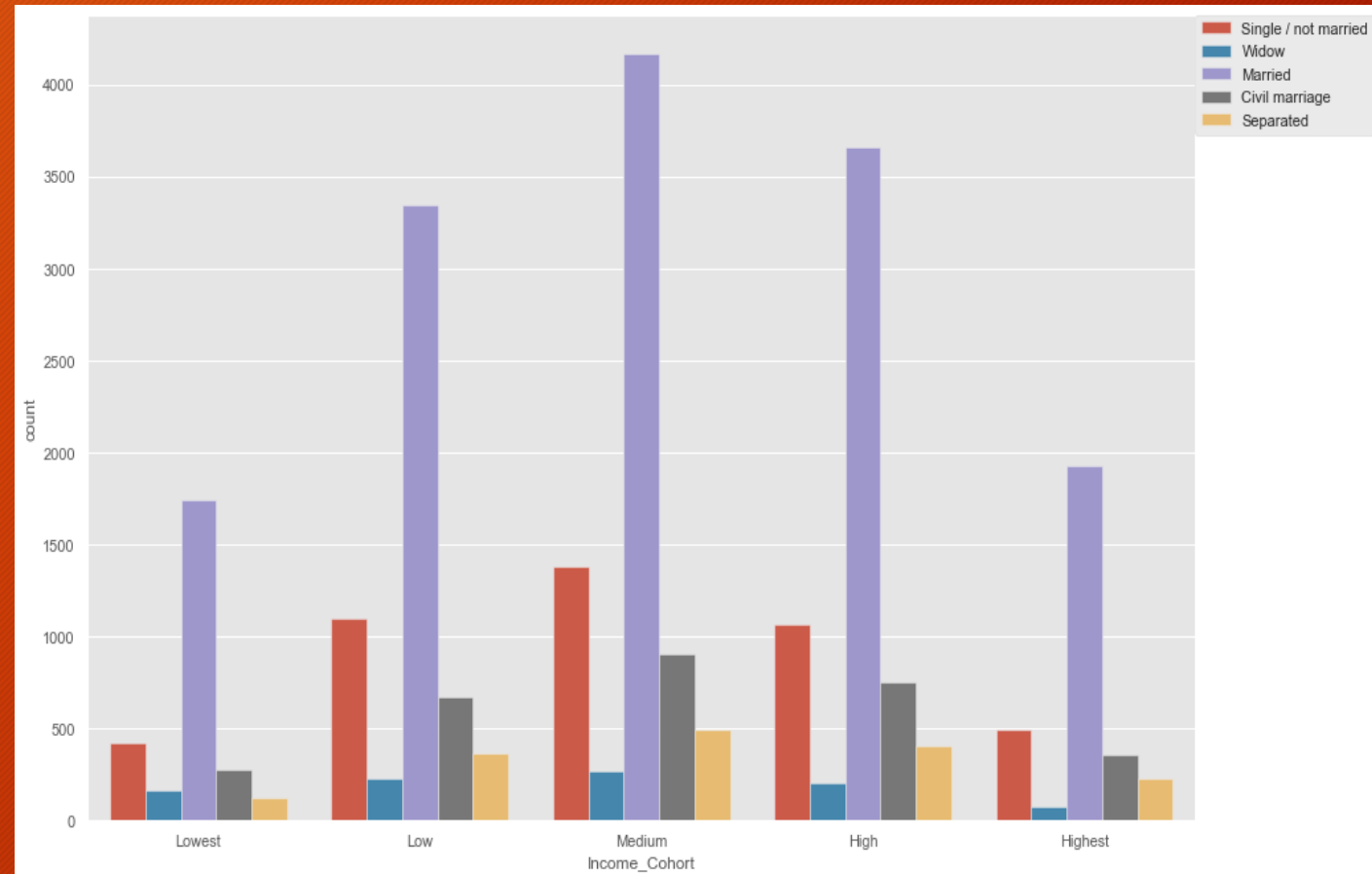


Bivariate Analysis of Variables

To Check the count of Income cohort based on Family Status

Inference :

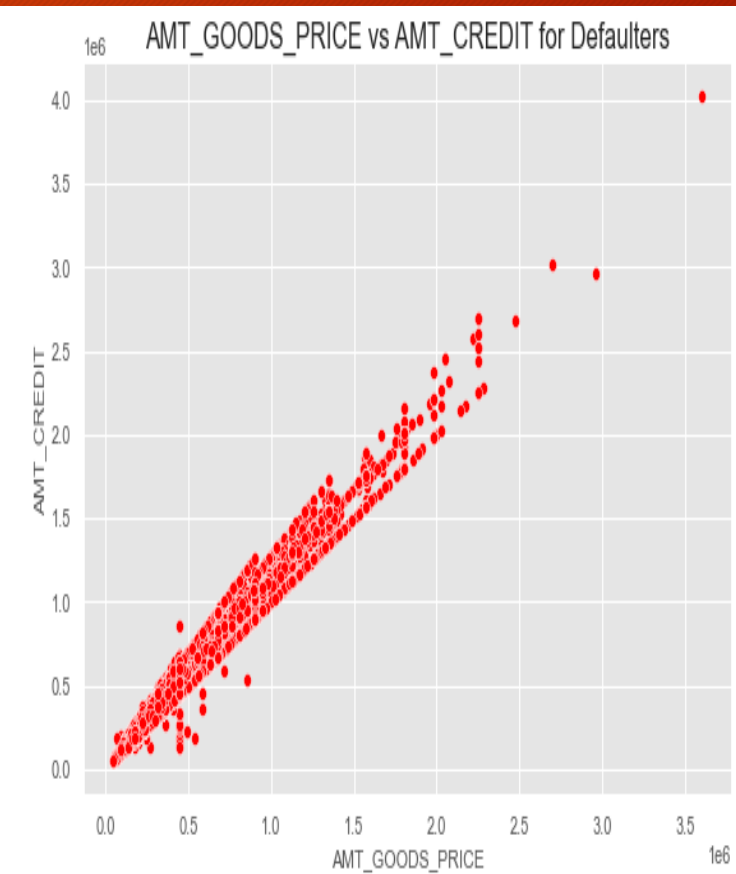
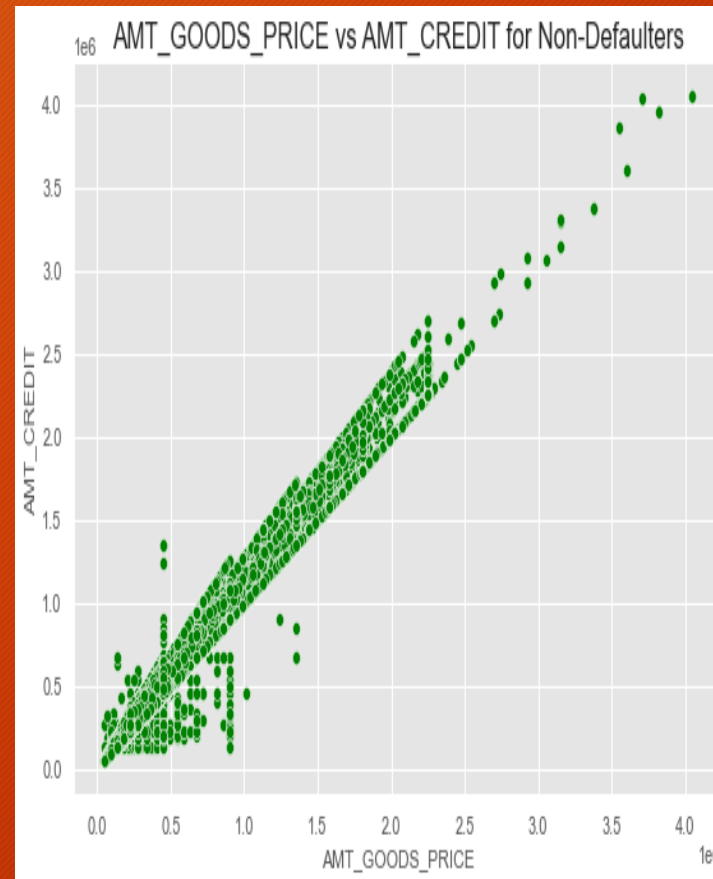
- The people who are married and have medium income are most likely to default and on all the bins of income the married people are most likely to default.
- Widows are least likely to default on all bins of income.



Comparison between AMT_GOODS_PRICE and AMT_CREDIT

Inference :

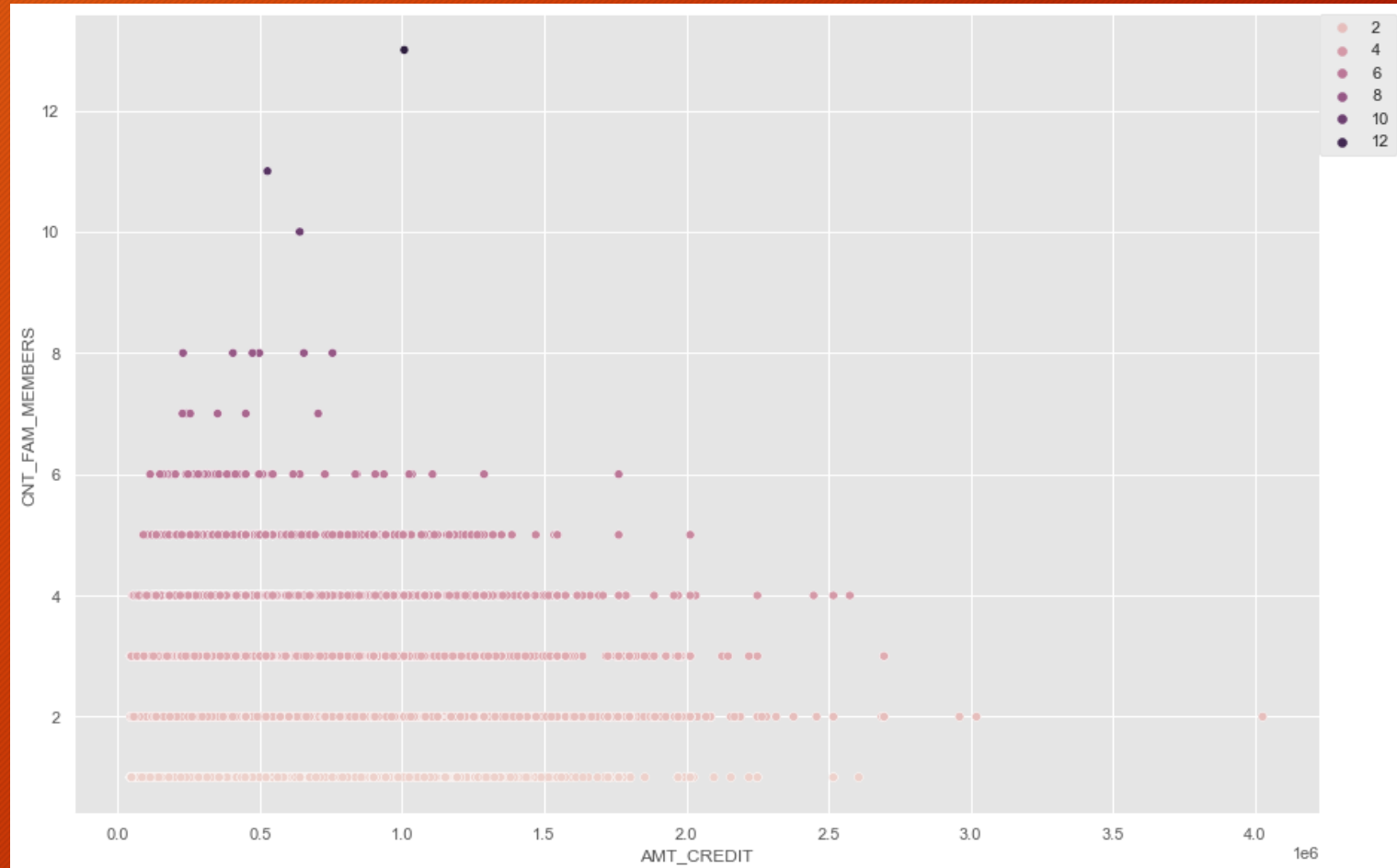
- The Number of Defaulters are reduced as the goods prices increases above 25 lakhs.



Comparison between AMT_CREDIT and CNT_FAM_MEMBERS

Inference :

- So from the graph we can say that as the family members increases the amount credited decreases.
- People having more number of family members and higher credit are least likely to default.

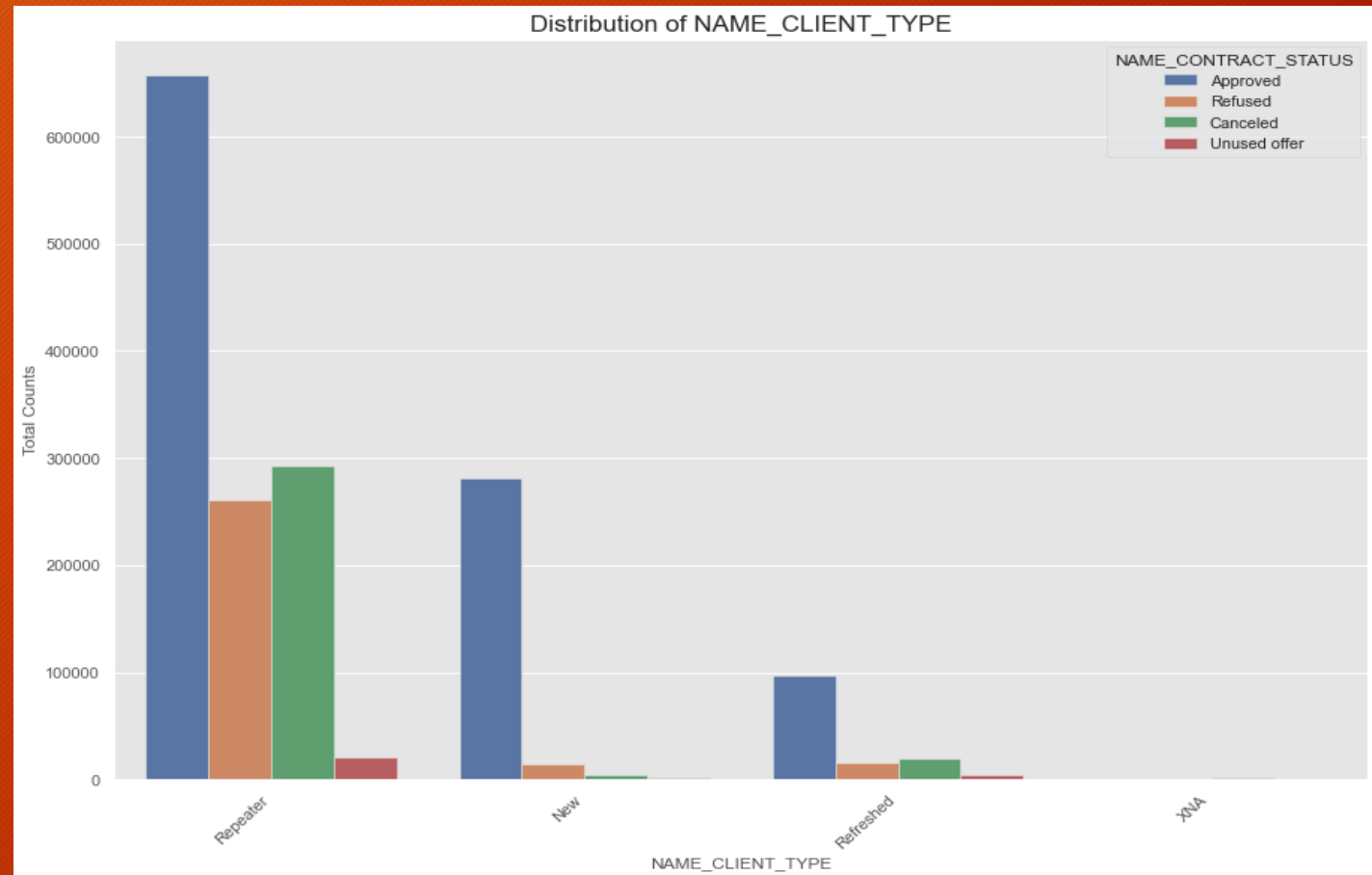


Univariate Analysis from Previous Application Data Frame

Analysis of NAME_CLIENT_TYPE using count plot

Inference :

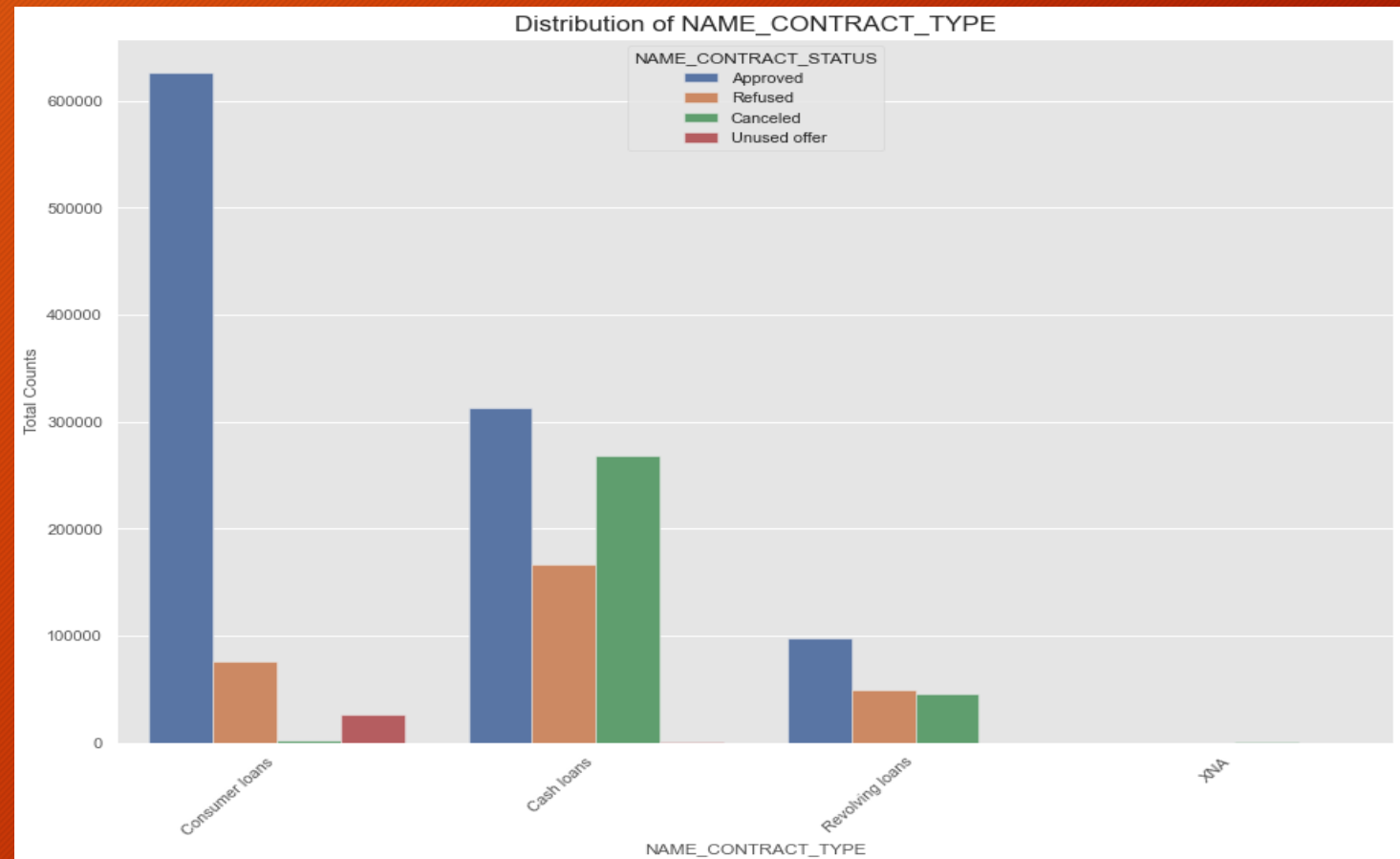
- Repeaters of the loan are approved for the new application the most.
- New Applicants use the offers the most as the unused offer count is least for them.



Analysis of NAME_CONTRACT_TYPE using count plot

Inference :

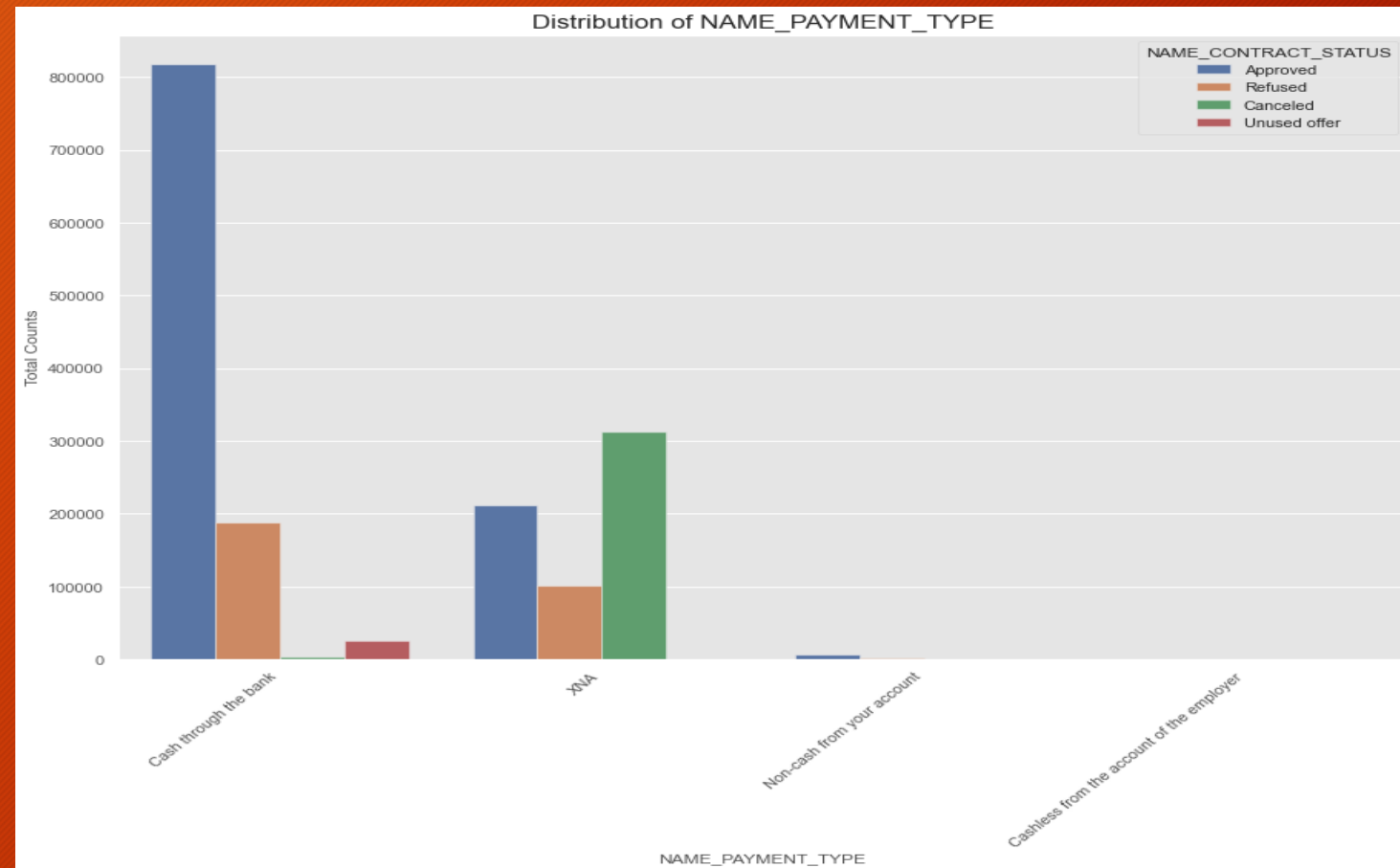
- Consumer Loans Applications and Cash Loan applications are approved the more than revolving loans.
- Cash Loans Applications are refused the most.
- Consumer Loans have most unused offer.



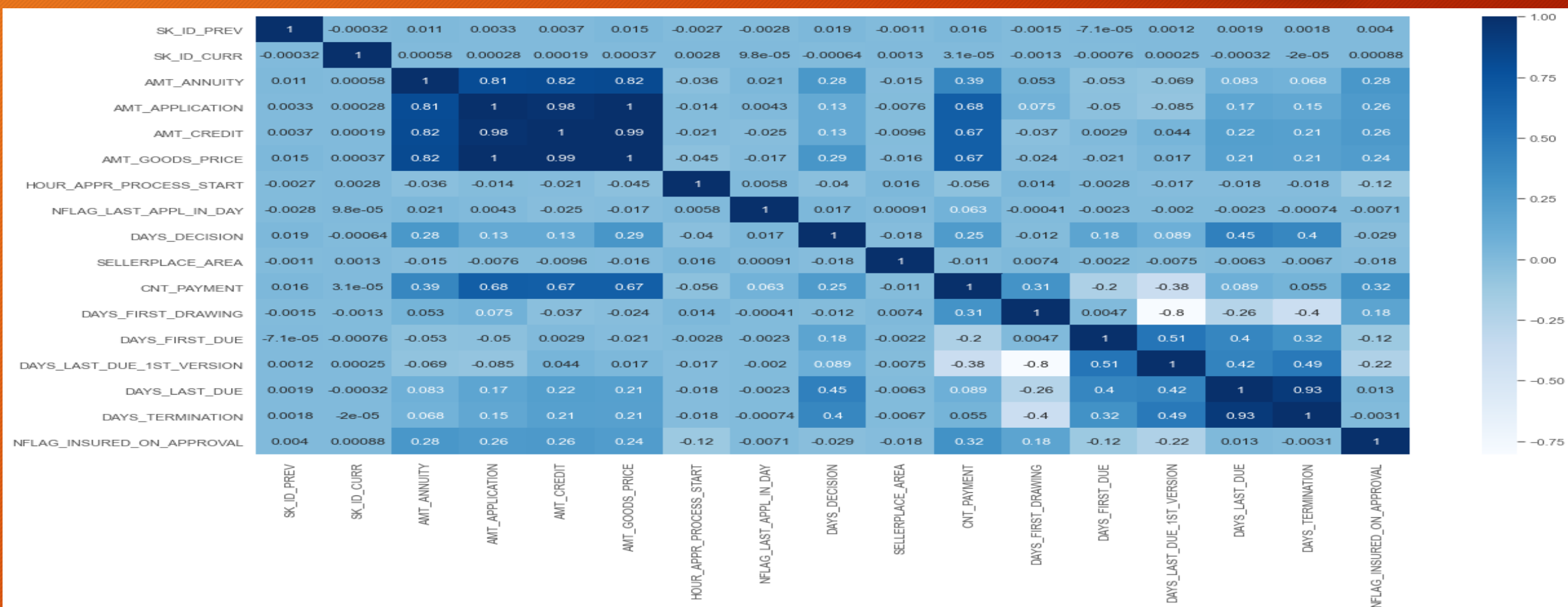
Analysis of NAME_PAYMENT_TYPE using count plot

Inference :

- Majority of clients paid through cash through the bank option.
- As we can observe from the plotted graph "Non-cash from your account" and "Cashless from the account of employer", these options are not famous amongst the client.



Checking Co-relation in Previous Application Data frame

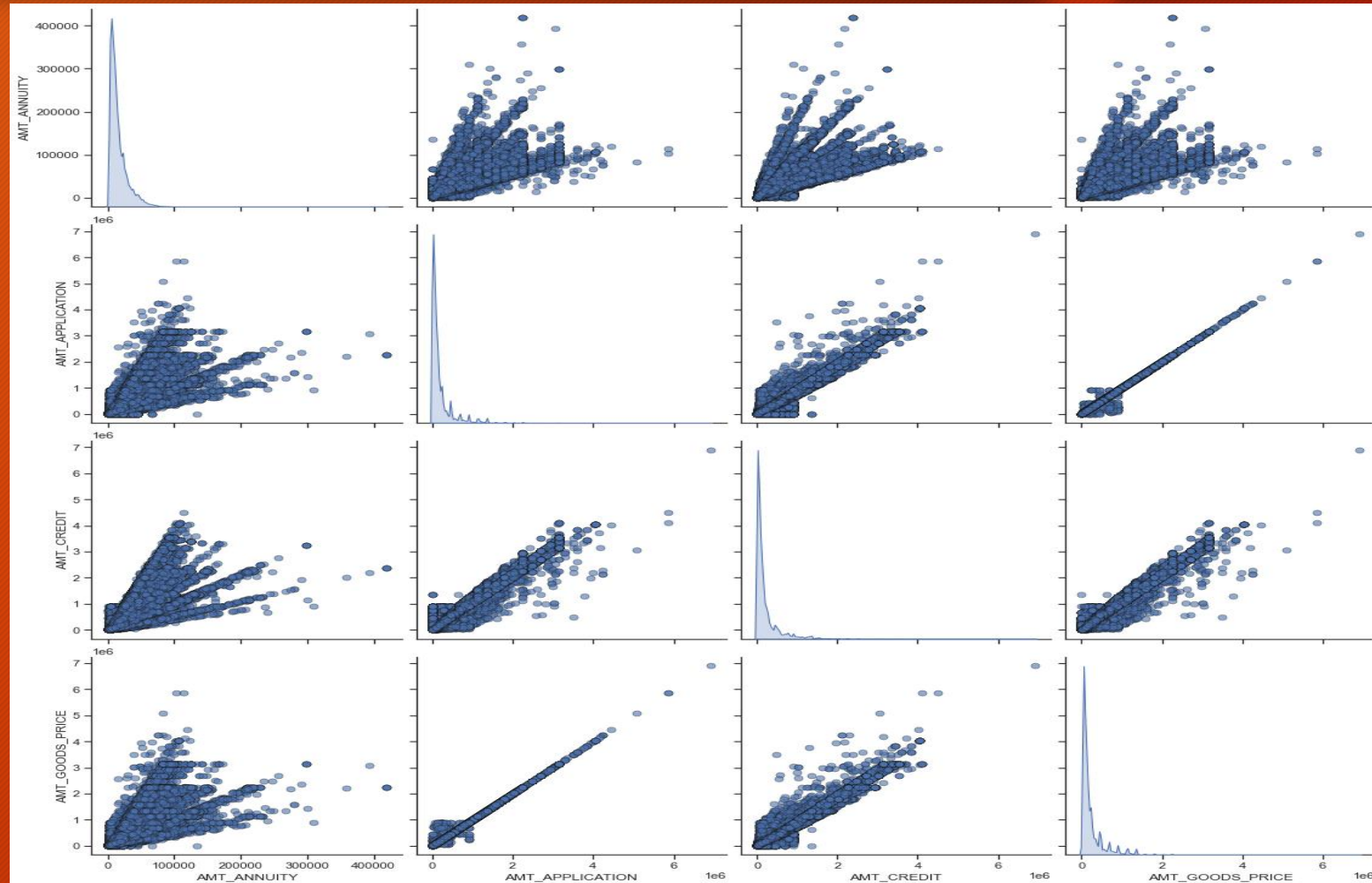


Bivariate Analysis of Variables in Previous Application Data Frame

Bivariate Analysis on numerical columns using pairplot

Inference :

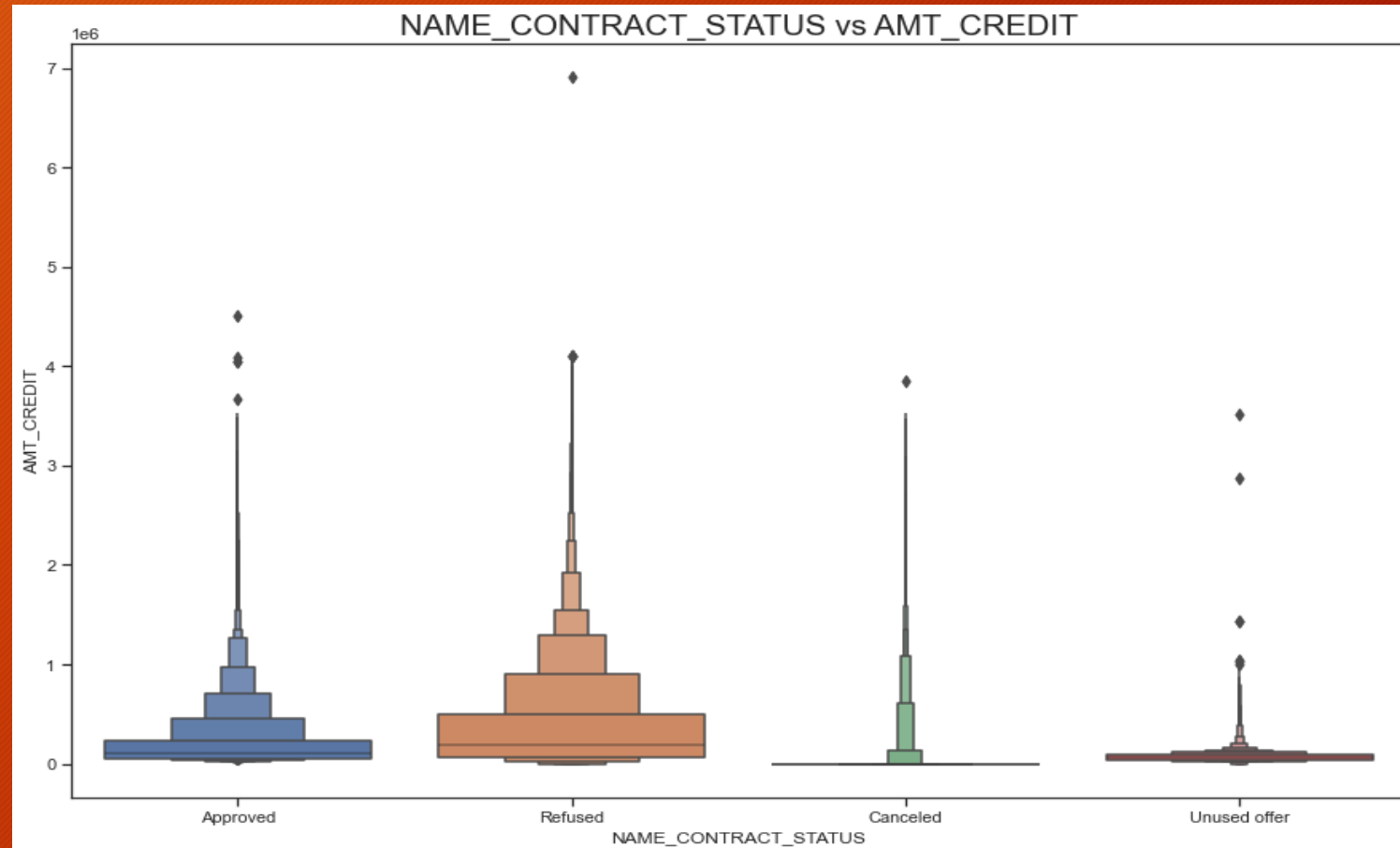
- There is a strong linear relation between goods price and application amount.
- The Amount of credit that the client asks for is highly dependent on goods price.
- We can see there is positive relation between application amount and credited amount.



Bivariate Analysis of categorical vs Numerical columns

Inference :

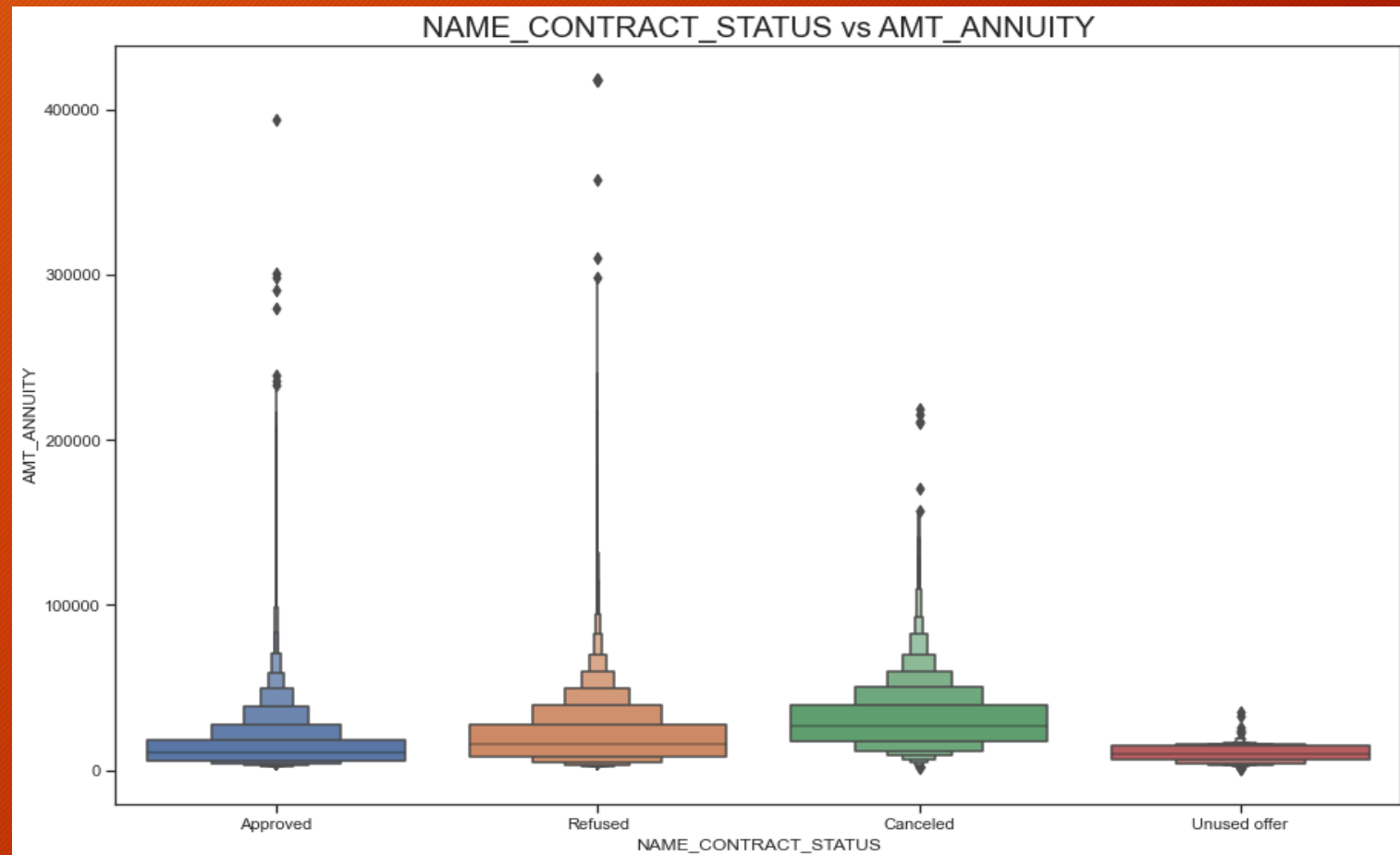
- As the credit amount goes low the offers are mostly unused and the loans are cancelled.



Bivariate Analysis of categorical vs Numerical columns

Inference :

- As the annuity amount increases there are high chances of loan application refusal.



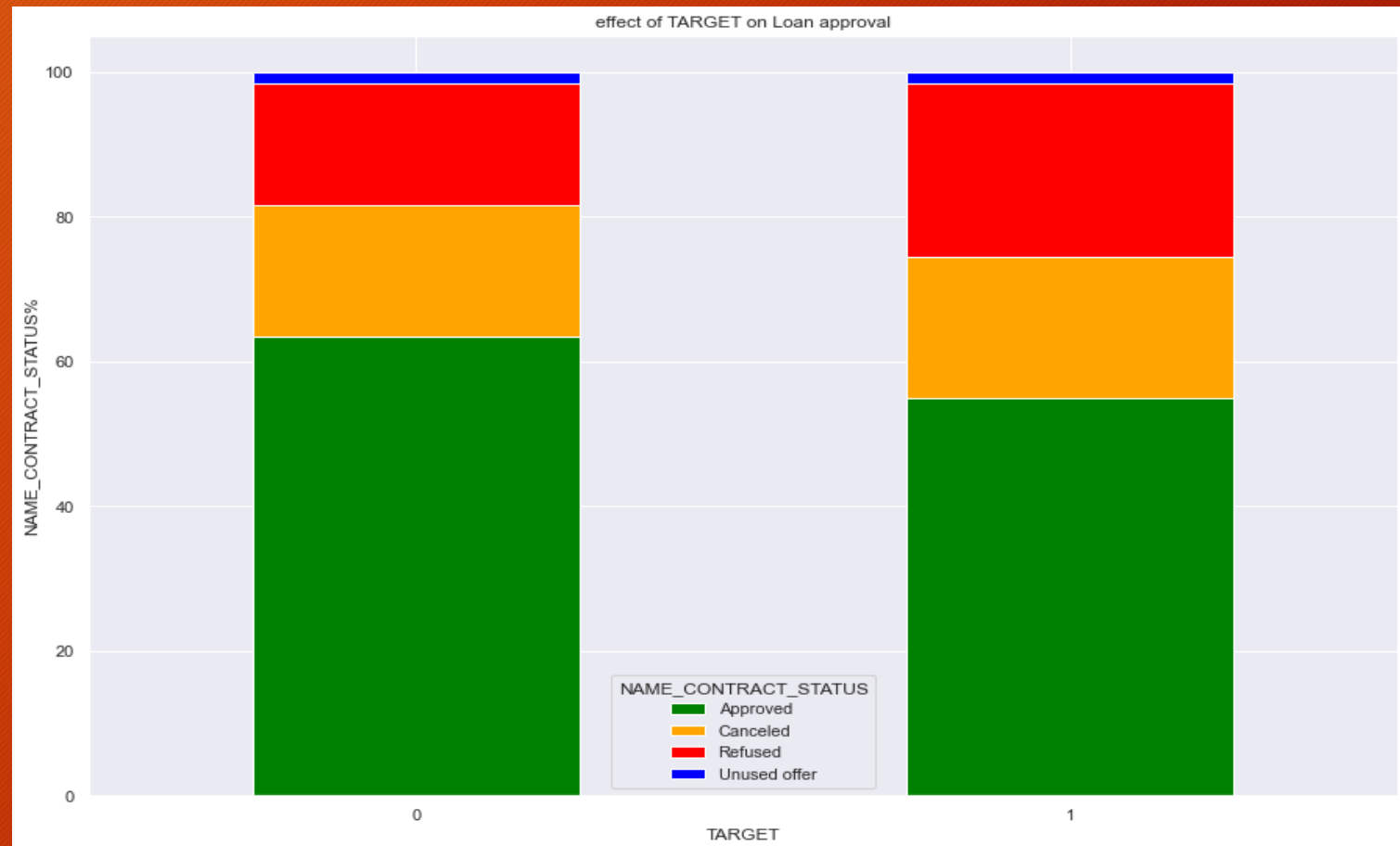


Analysis On Combined Data Frames of Application Data and Previous Application

Analysis of Target variable against NAME_CONTRACT_STATUS

Inference :

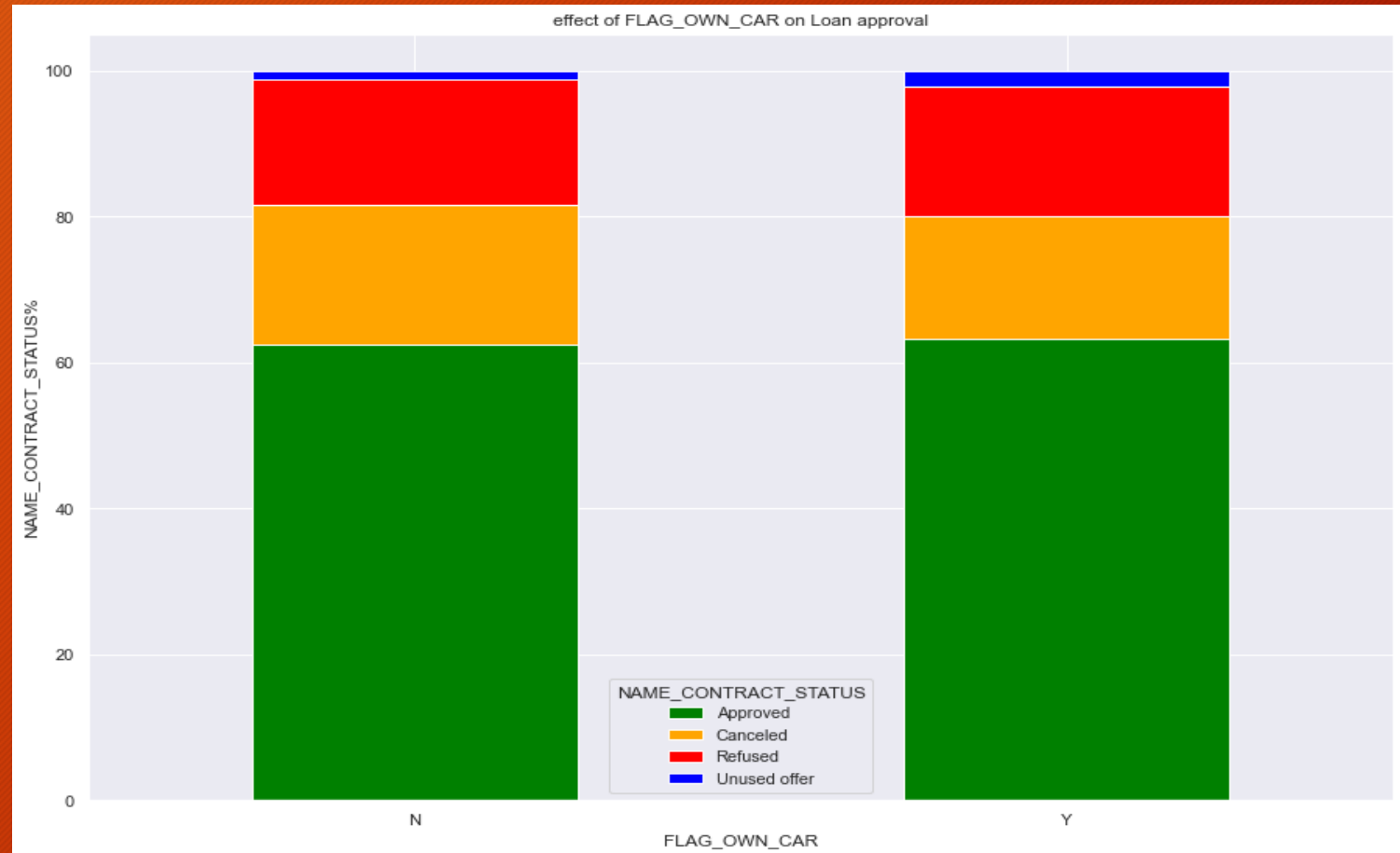
- So we can observe from the plotted graph that the people whose loans were approved in previous applications those people tend to default less after the approval of current application.



Analysis of FLAG_OWN_CAR against NAME_CONTRACT_STATUS

Inference :

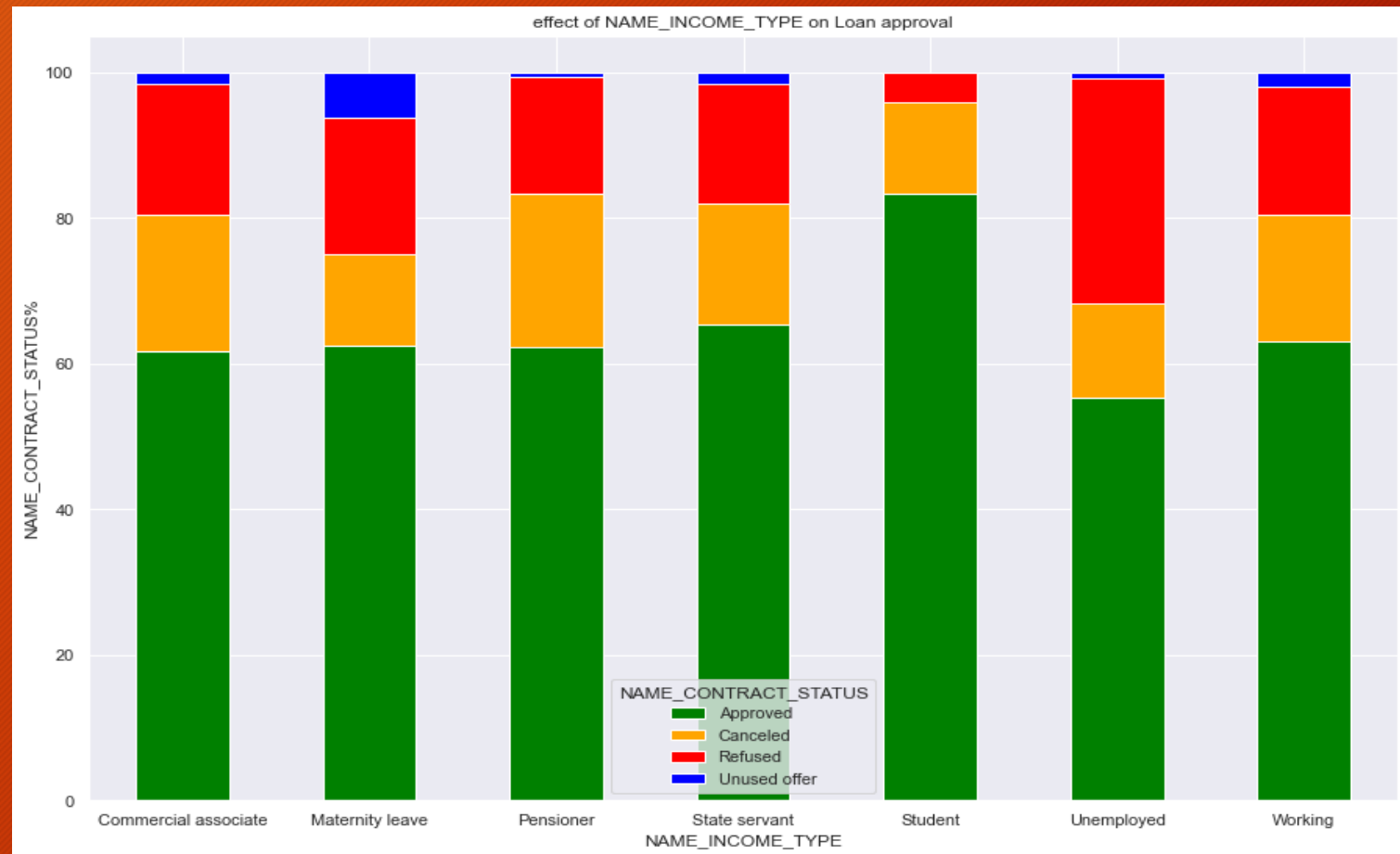
- The People who own a car are not likely to use the offers.



Analysis of NAME_INCOME_TYPE against NAME_CONTRACT_STATUS

Inference :

- The Students loans have higher approval rate compared to other income type.
- People who are on maternity leave have the most number of unused offers.
- The rejection rate for the unemployed people is the most as observed from the graph.



Conclusion

- The Percentage of defaulters are 8% and percentage of non-defaulters are 92%
- Highest number of females apply for loans
- The Bank should focus on promoting consumer loans and cash loans as they are more famous amongst the consumers
- Elder People are least likely to default , hence the bank should focus on promoting the loans to elder people
- The people whose loans were approved in previous applications those people tend to default less after the approval of current application.
- People with lower annuity amount get cancelled more often
- Married people default the most
- Majority of clients paid through cash through the bank option.

THANK
YOU