Bank Loan Application Case Study

A Comprehensive Review of Loan Application Data

By – Aqib Jallal

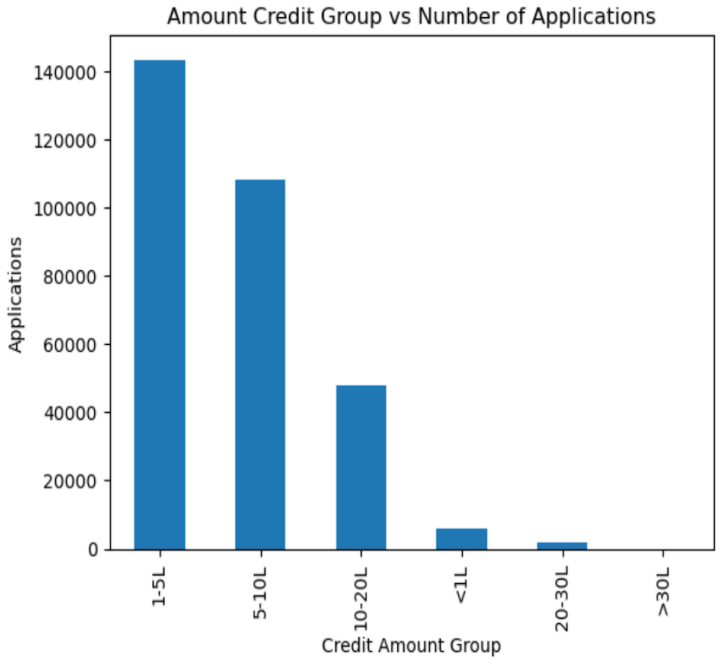
Introduction to data set

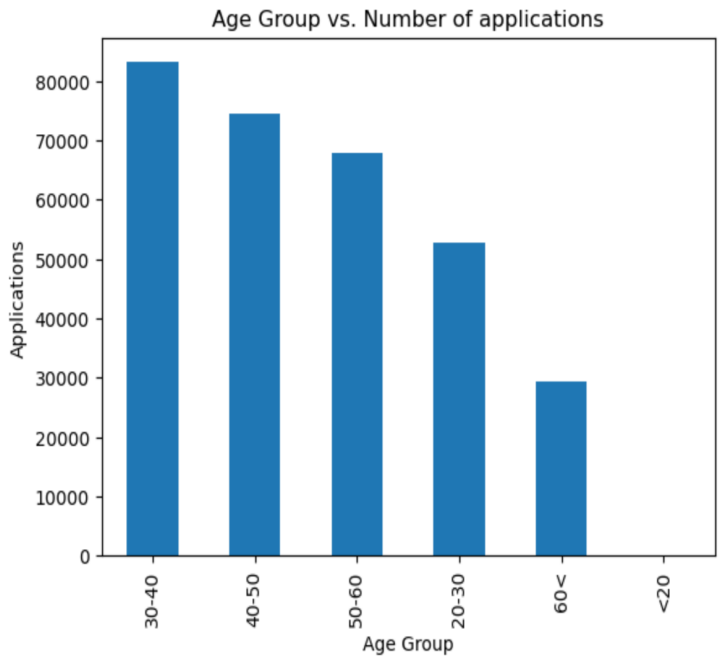
1. application_data.csv
   - Number of Rows = 307510
   - Number of Columns = 112
   - Contains details of the client at the time of application.
2. previous_application.csv’
   - Number of Rows = 1670213
   - Number of Columns = 37
   - contains information about the client’s previous loan data. It contains the data on whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

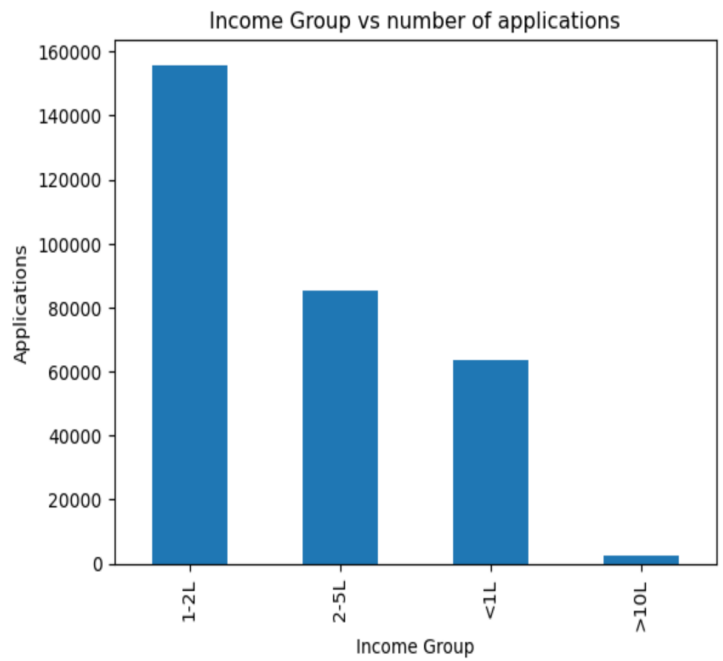
Data Cleaning & Preparation of data set application_data.csv

- Handling missing values:
  - In the data set, there were 33 columns with more than 40% missing values in. As these were not relevant to our analysis so I have dropped them from the data set. Also, the columns 'AMT_REQ_CREDIT_BUREAU_DAY', 'AMT_REQ_CREDIT_BUREAU_WEEK', 'AMT_REQ_CREDIT_BUREAU_MON', ' AMT_REQ_CREDIT_BUREAU_QRT', 'AMT_REQ_CREDIT_BUREAU_YEAR' were having 13.5% missing data so I have replaced the missing values with the median of the columns respectively.
- Updating the data type of the column from float to int:
  - The columns 'AMT_REQ_CREDIT_BUREAU_DAY', 'AMT_REQ_CREDIT_BUREAU_WEEK', 'AMT_REQ_CREDIT_BUREAU_MON', 'AMT_REQ_CREDIT_BUREAU_QRT', 'AMT_REQ_CREDIT_BUREAU_YEAR', 'DAYS_REGISTRATION' were having the data type as float which will create difficulty on analysis as there fields are not supposed to be in float so have updated the data type as int.
- Updating the negative values of the columns to positive:
  - The columns 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH', 'DAYS_LAST_PHONE_CHANGE', were having some negative values which I have changed to positive as there columns are not supposed to have negative values.
- Dropping columns not required for analysis:
  - Dropped 20 columns related to document as the are not necessary for the analysis.

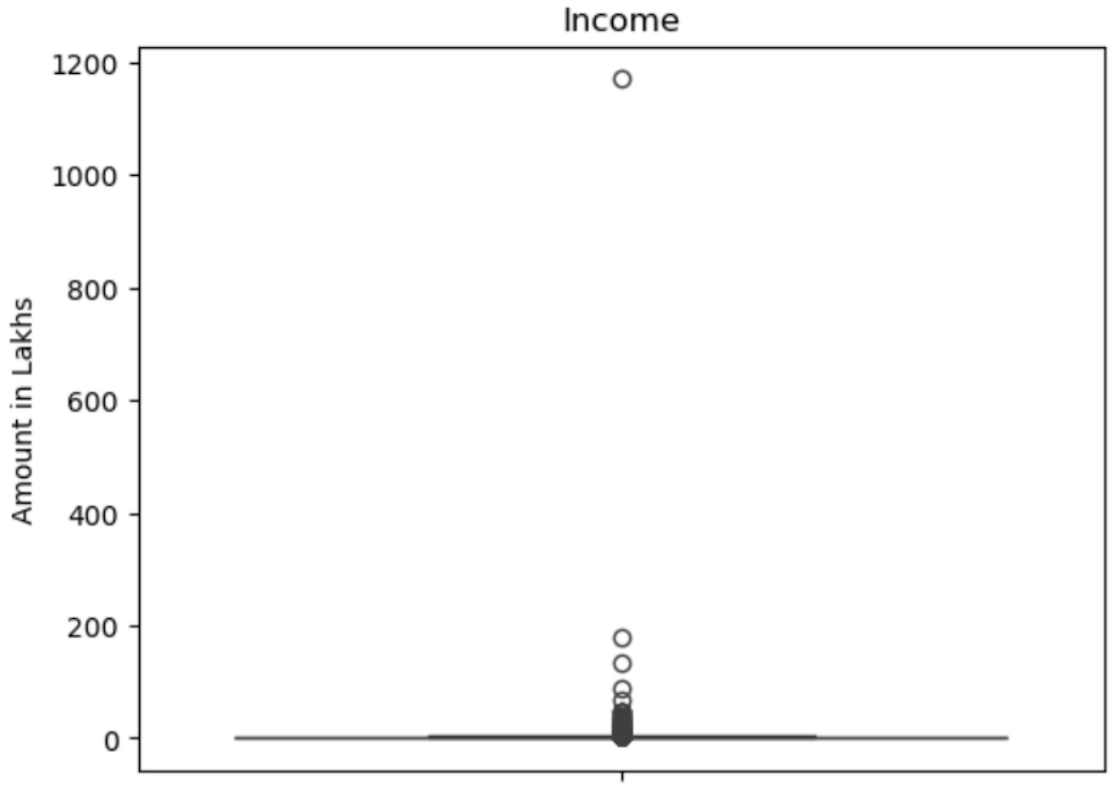
- Binning and plotting the data:
  - Created bins of the columns and plotted a graph showing the number of loan applications.

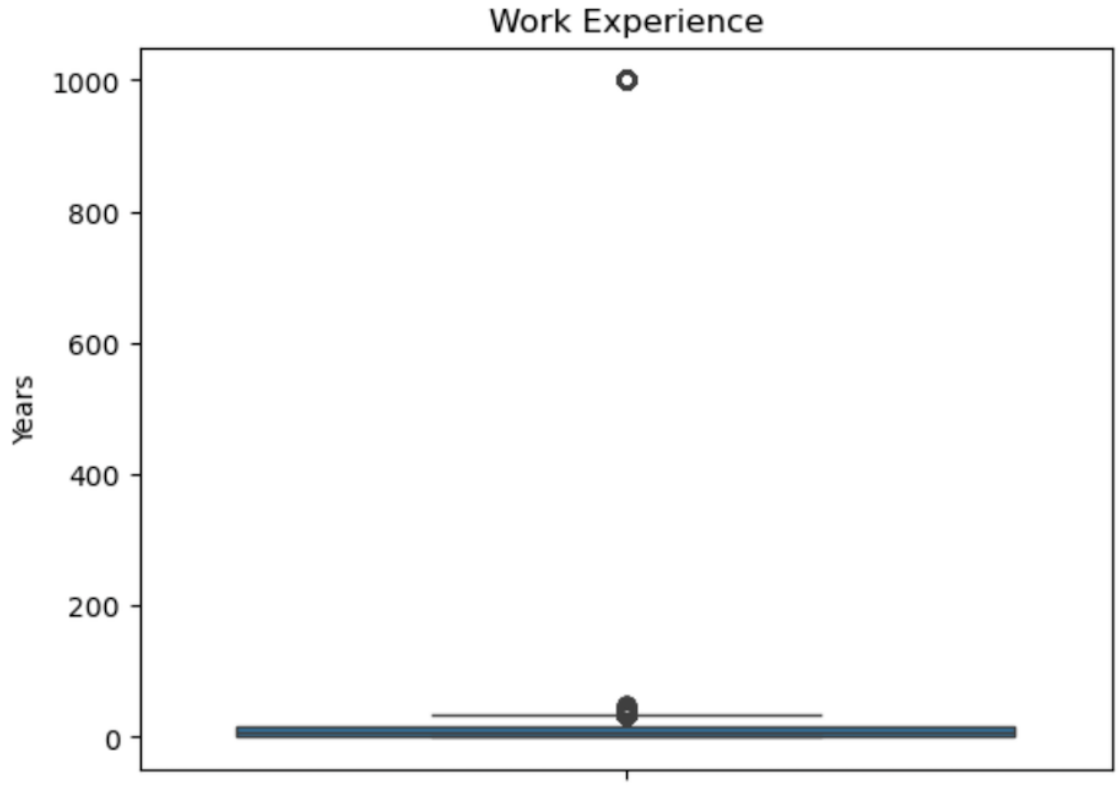




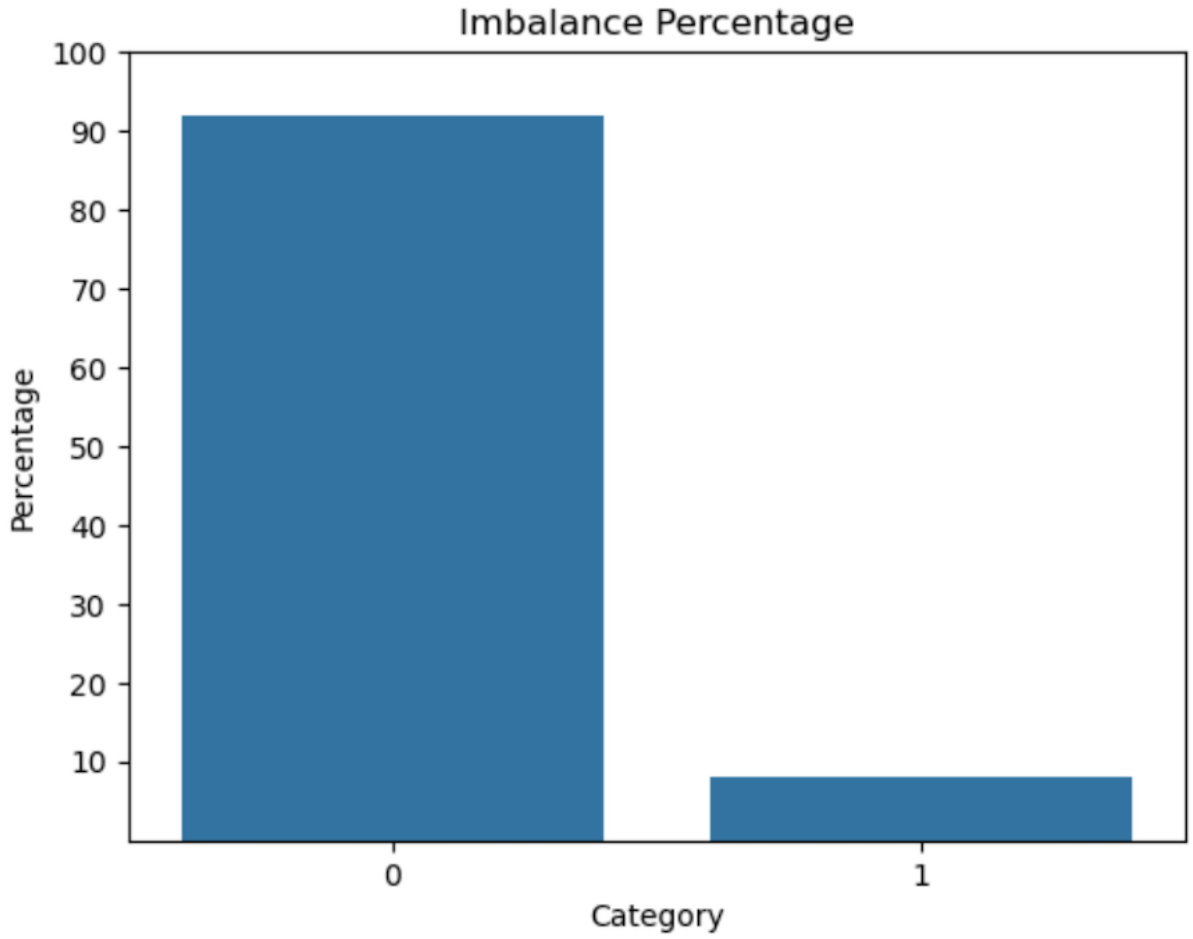


- Finding outliers:
  - AMT_INCOME_TOTAL has very high valued outliers.
  - DAYS_EMPLOYED has huge outliers. Few of the data points are showing close to 1000 years in service which is impossible.

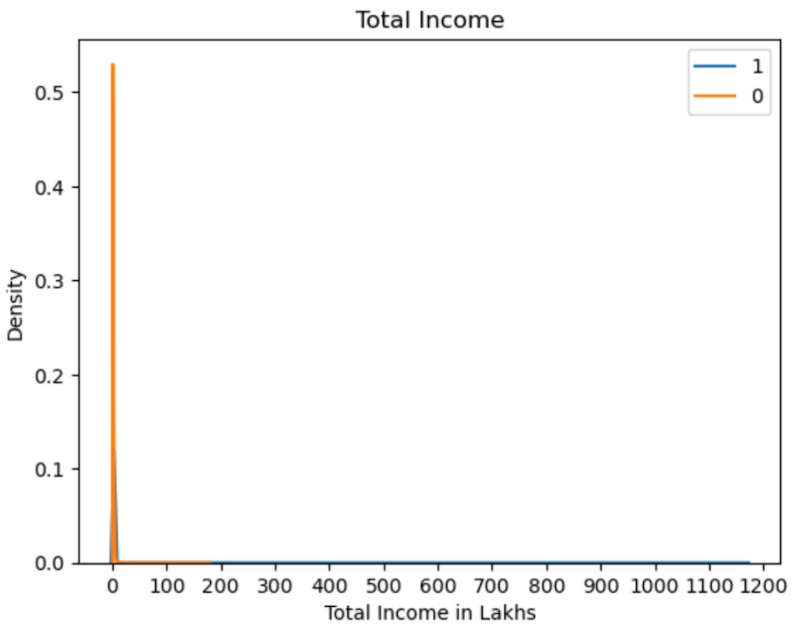


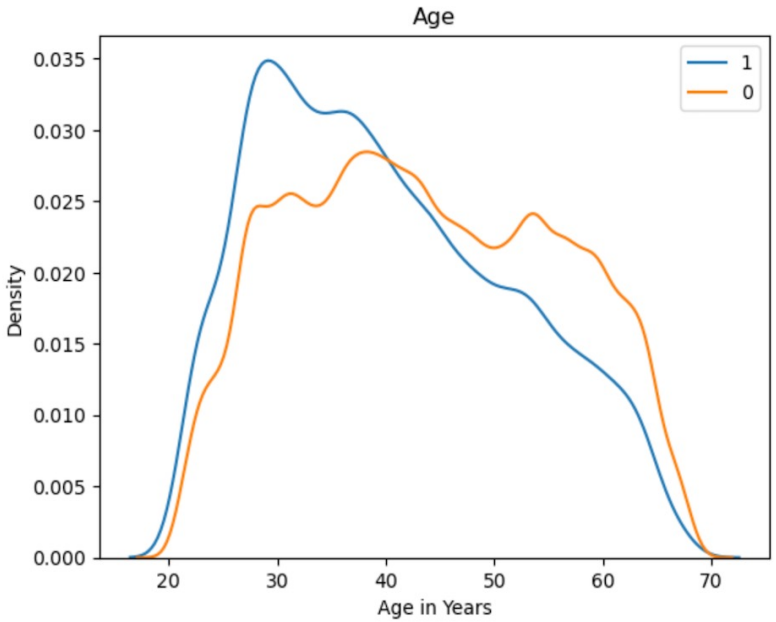


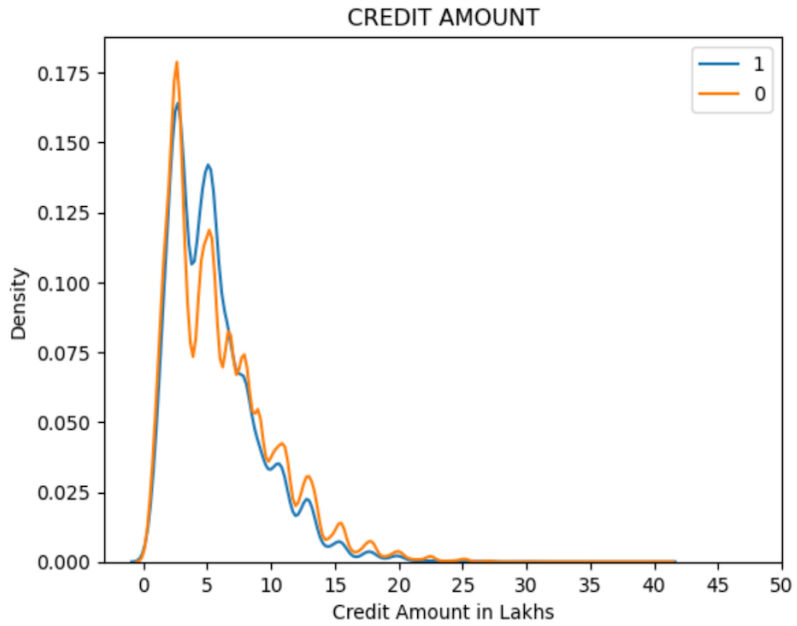
- **Performing Imbalance check:**
  - **The value of Target column in data set, the value '0' in the application_data data set is '92%' where as the value of '1' is '8%'. Which means 8% of the people were facing difficulties in repayment and became defaulter.**



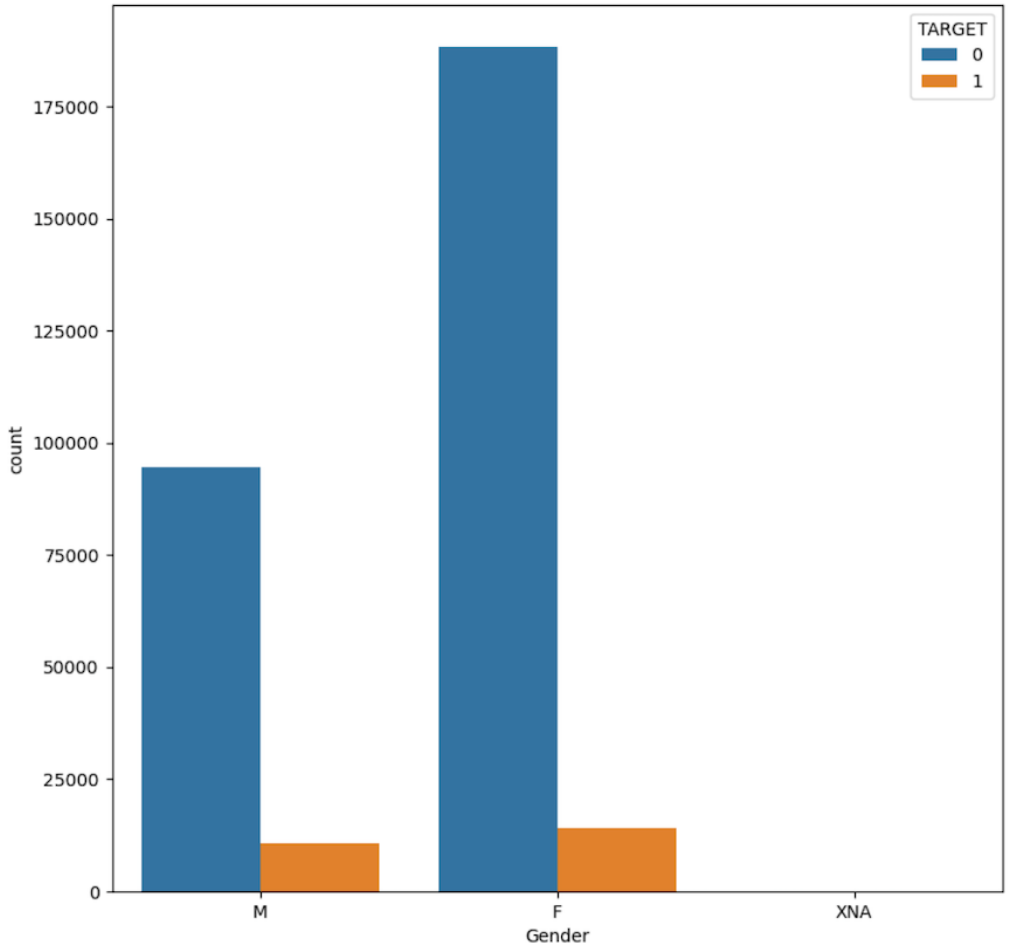
- # Univariate Analysis on Numerical Variables:
  - **Maximum applications are in low range of income along with the clients who were not able to repay.**
  - **Most of the clients who were not able to repay the loans are in the age group of 25-40.**
  - **Most of the loan amounts are under 15 lakhs**

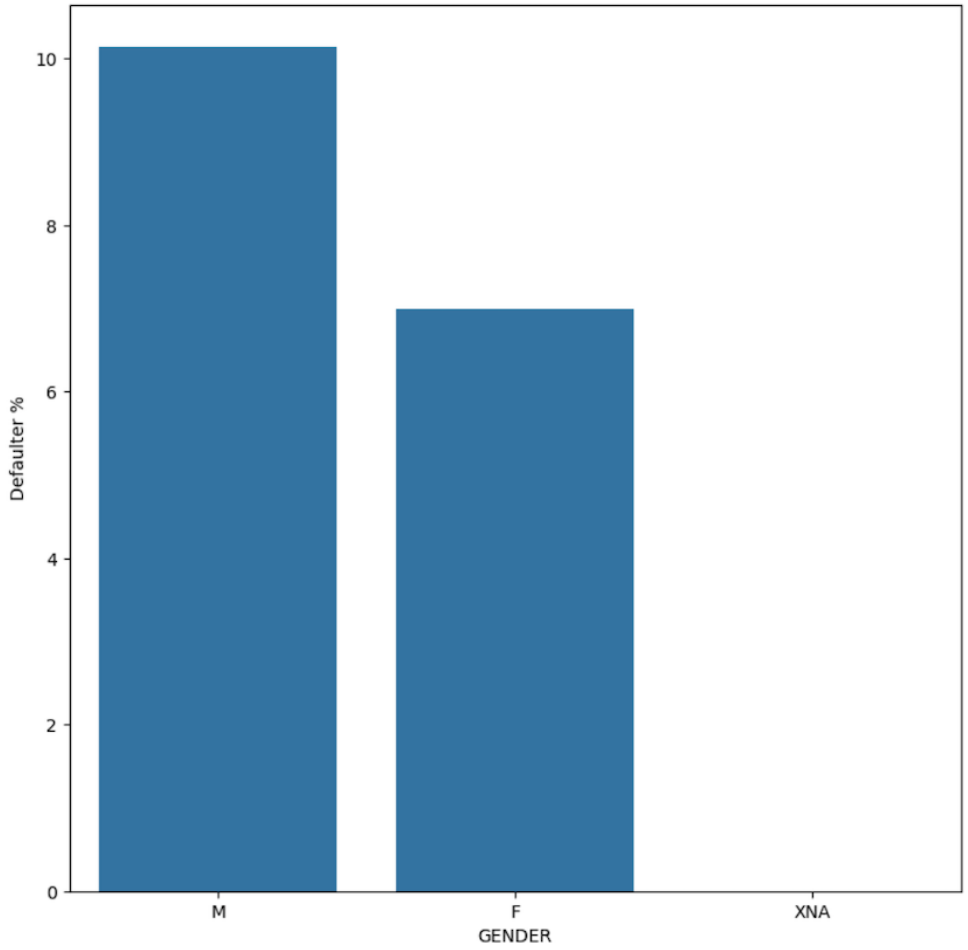




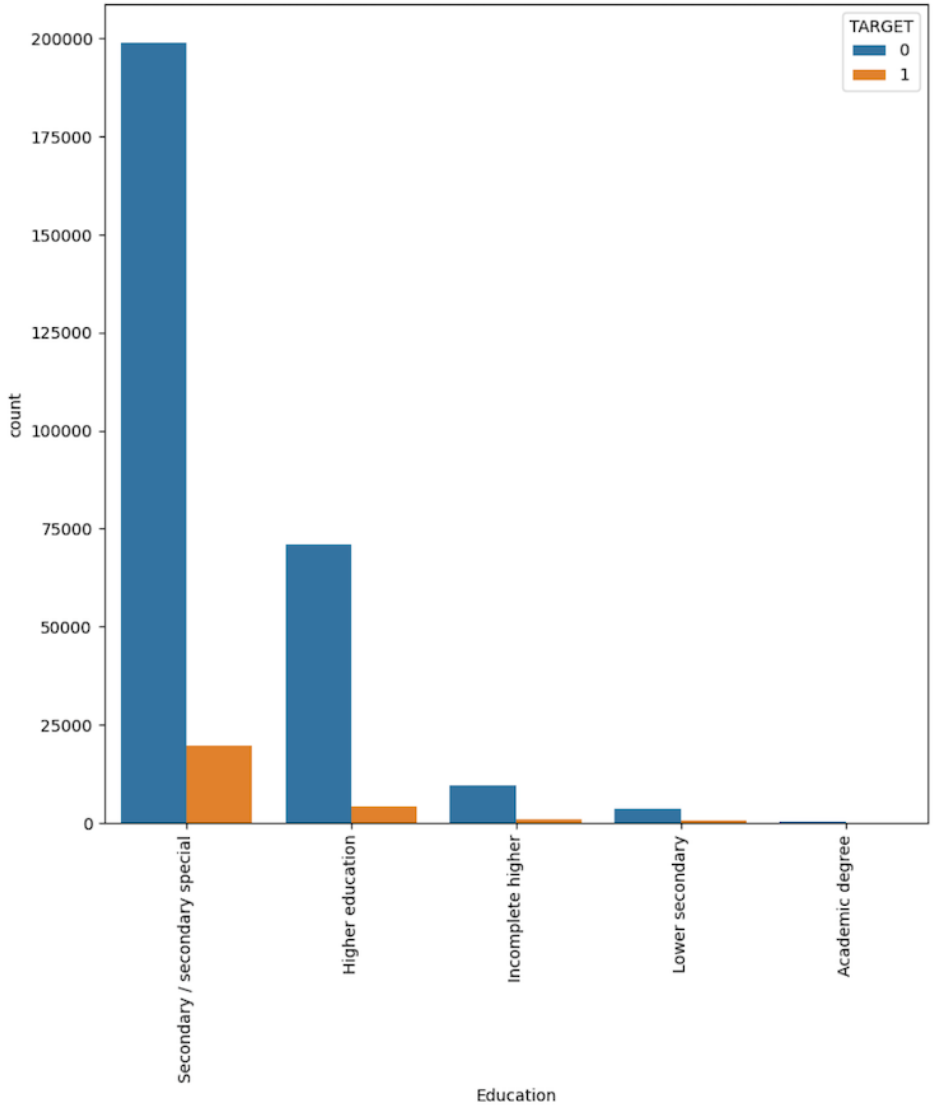


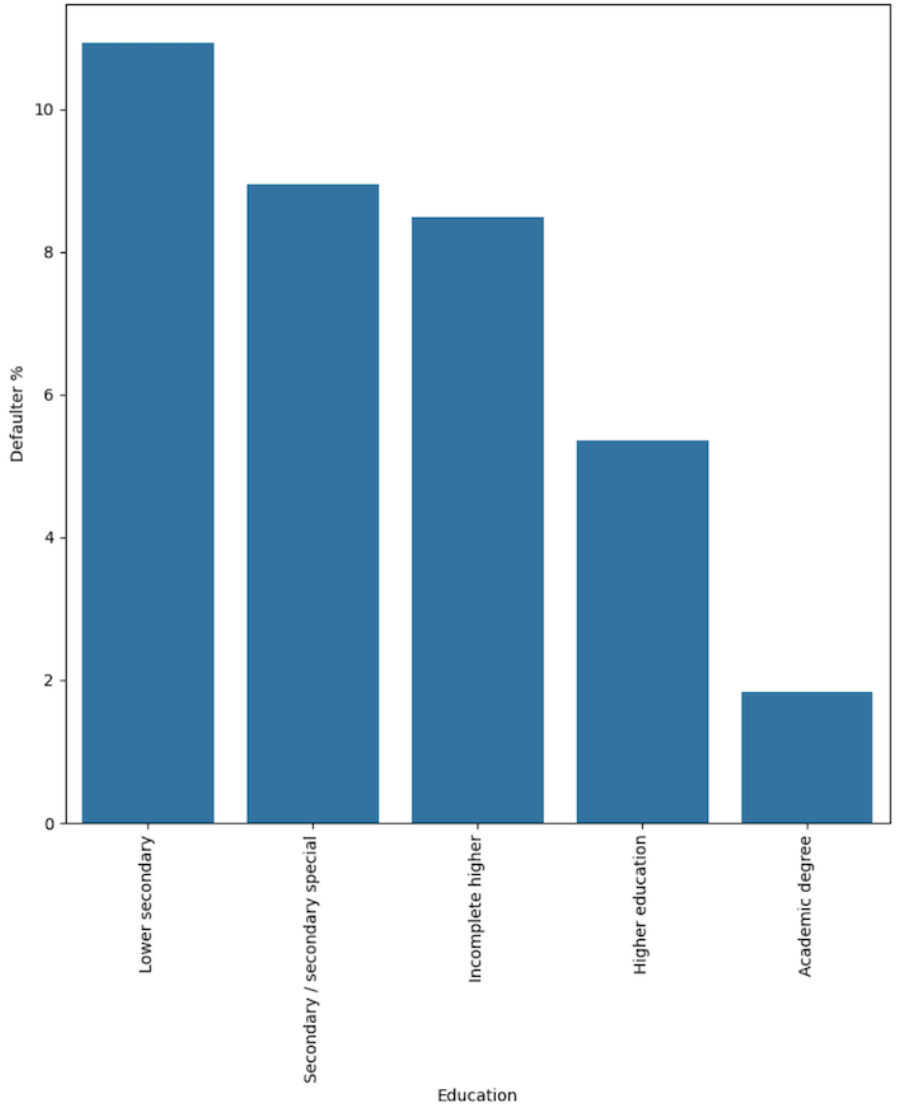
- **Univariate Analysis on Categorical Variable**
  - **Females clients are more in number than males for both the categories.**
  - **Male clients are more likely to be defaulter than Female.**



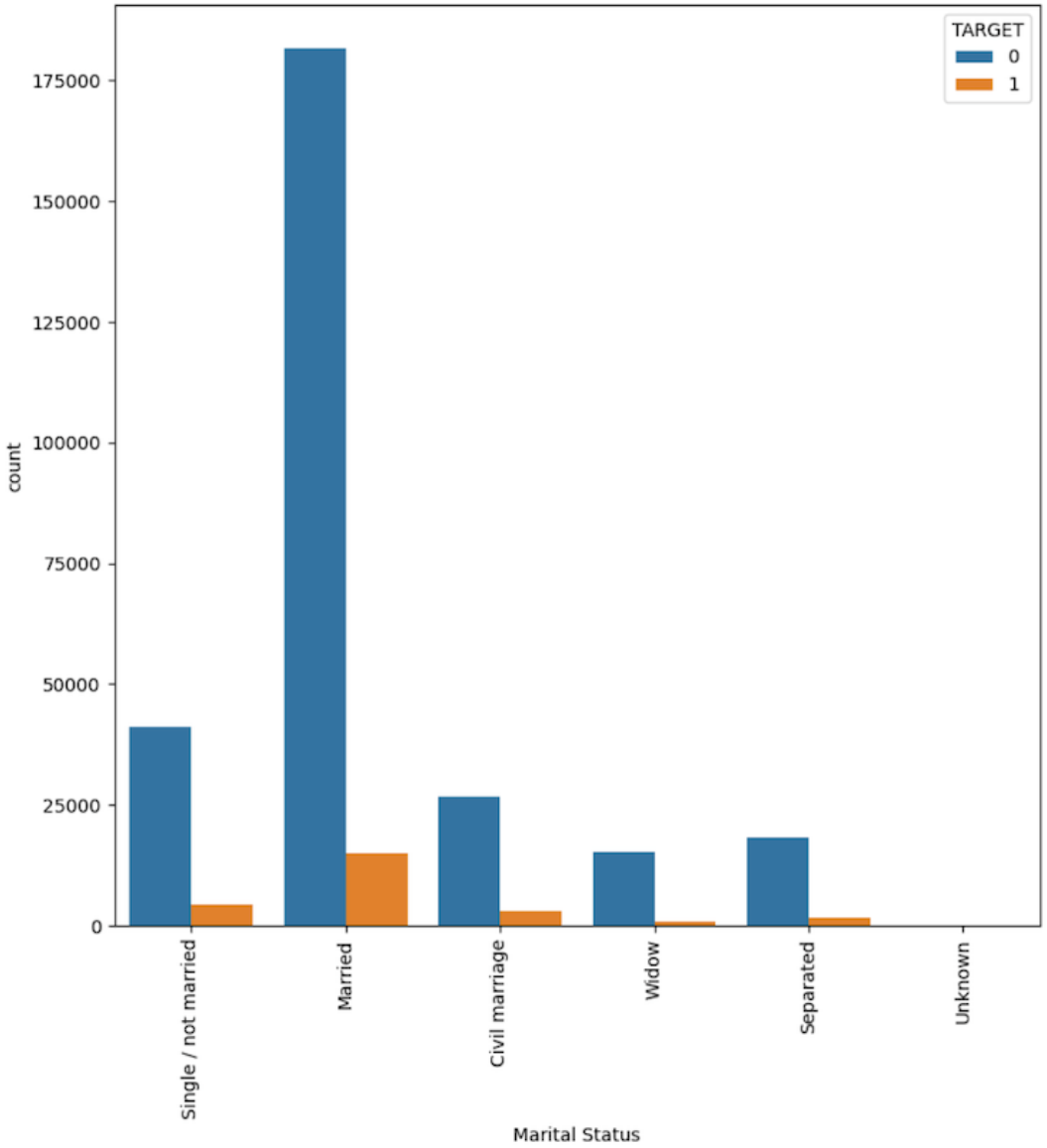


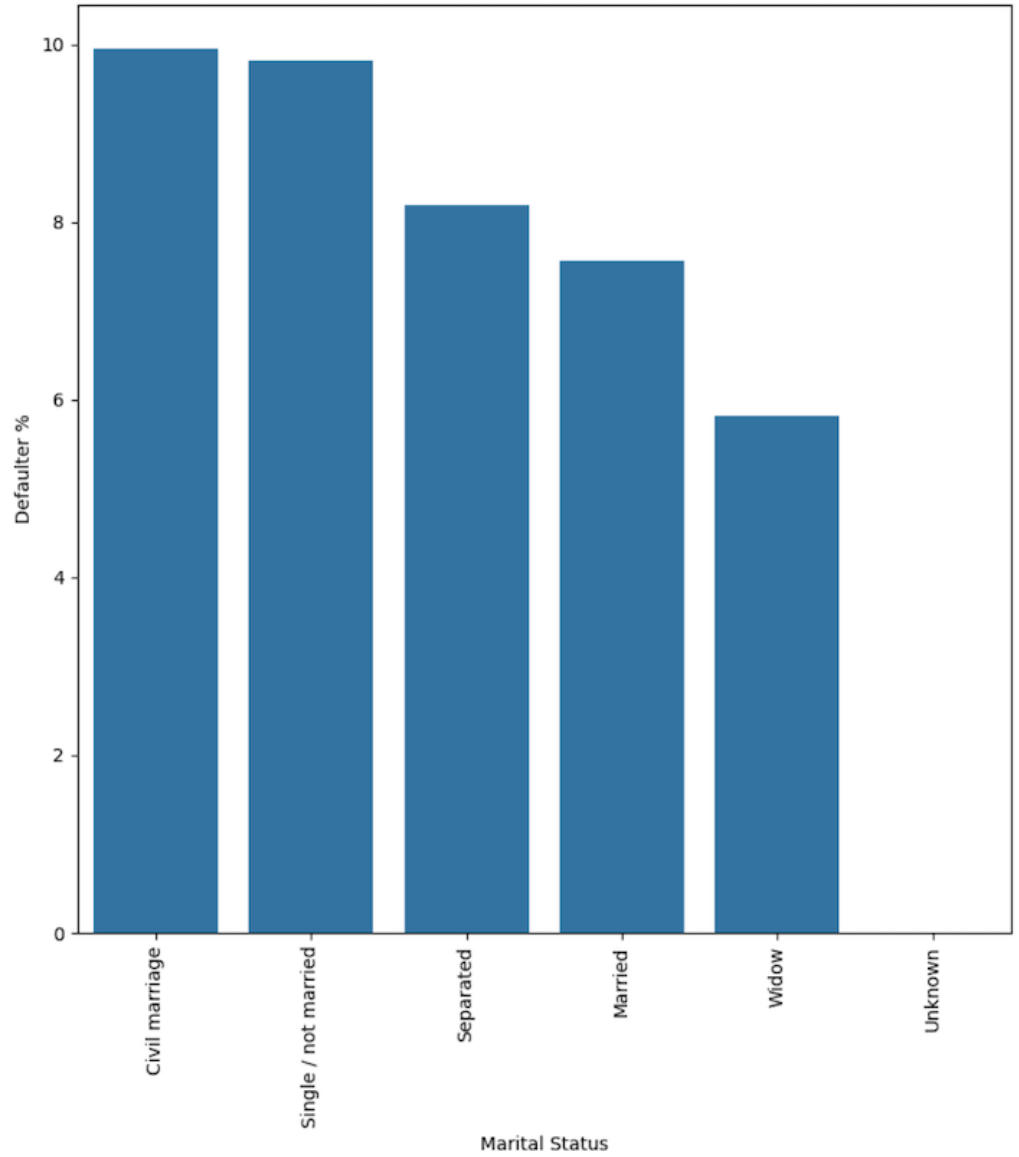
- **Working Category people are applying more for the loan.**
- **Maternity Leave and Unemployed clients are having high chances to become defaulter.**



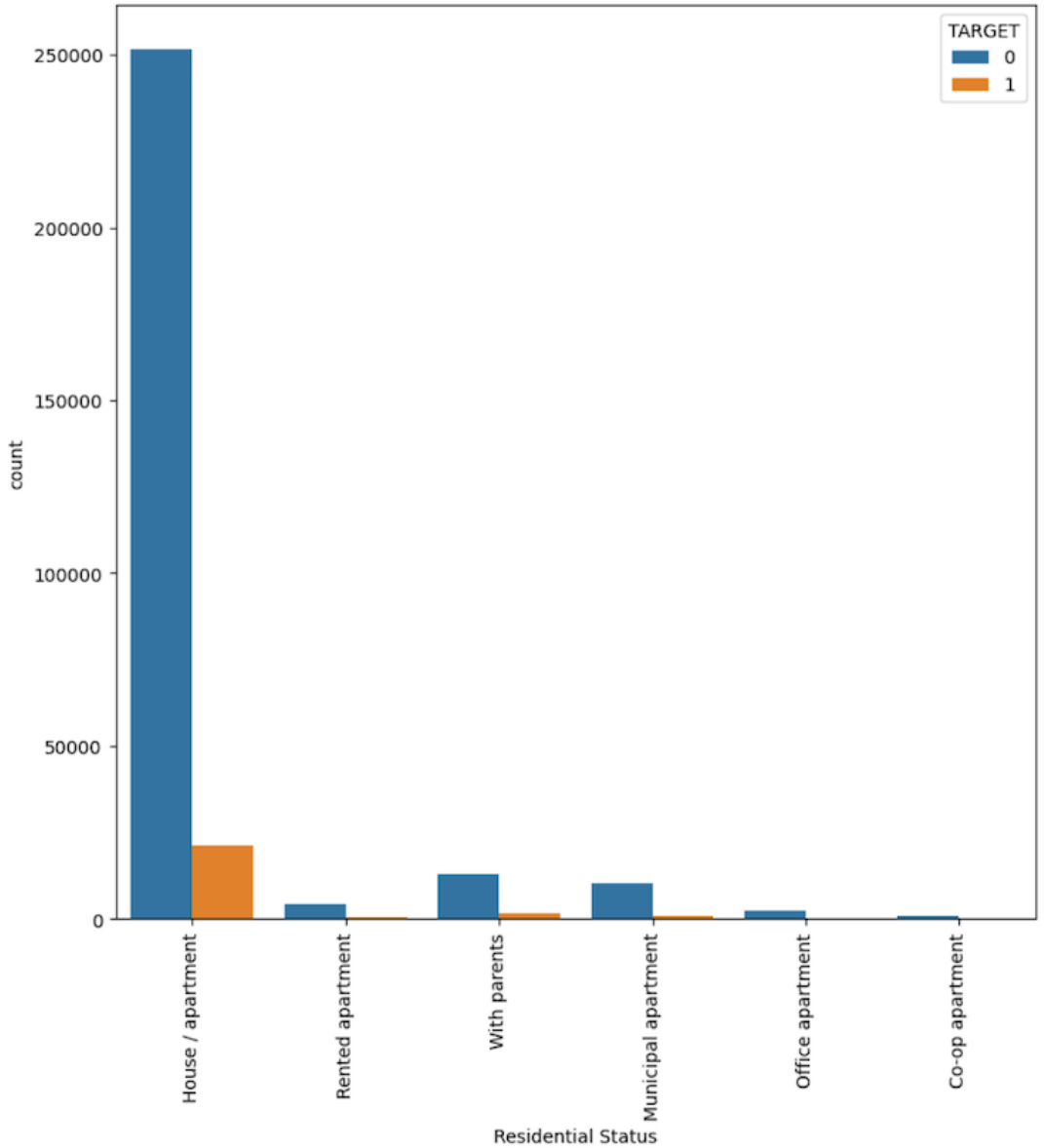


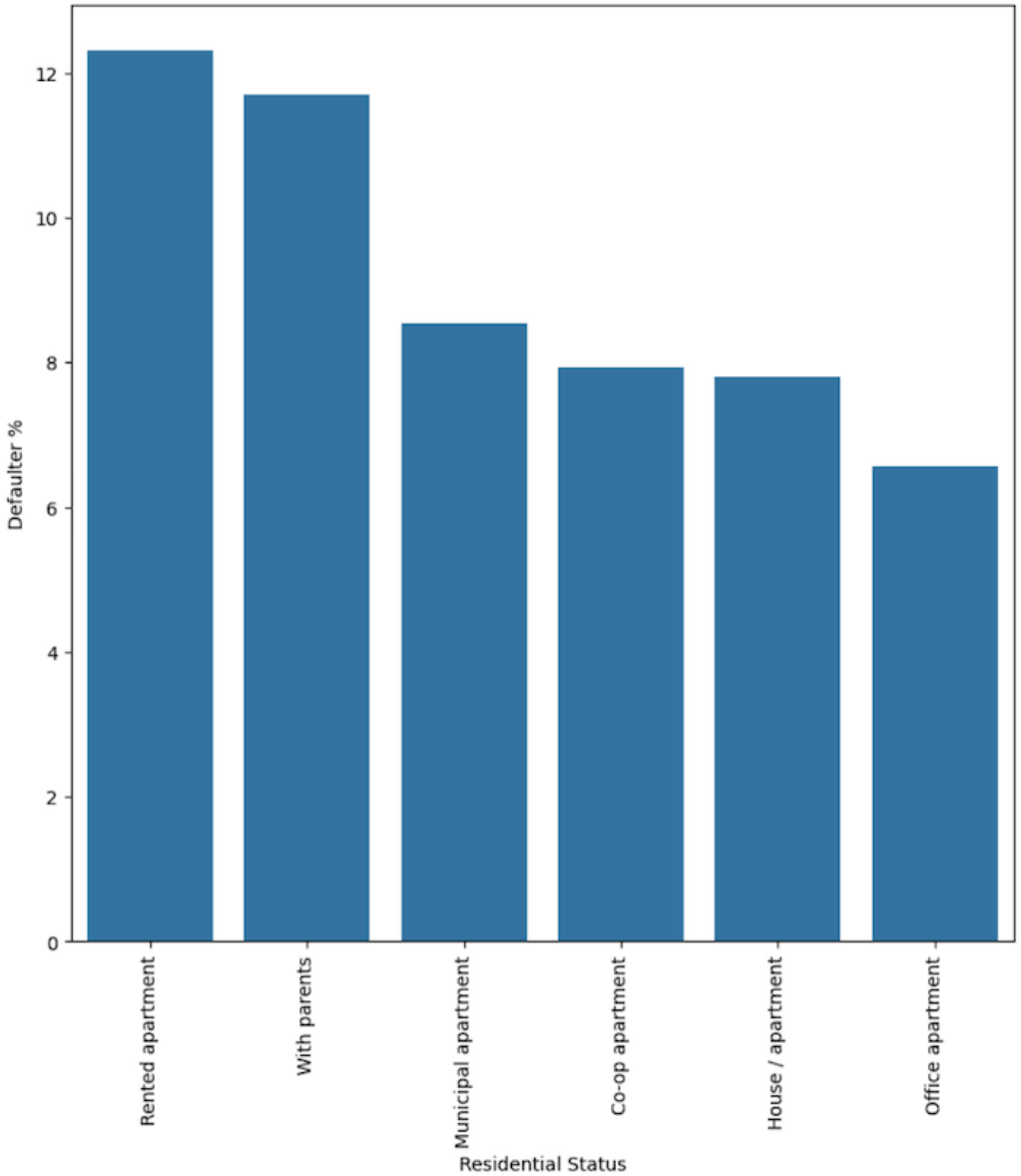
- **Married clients have more number of loan applications than others also Civil marriage and Single/not married clients are more likely to become defaulter.**



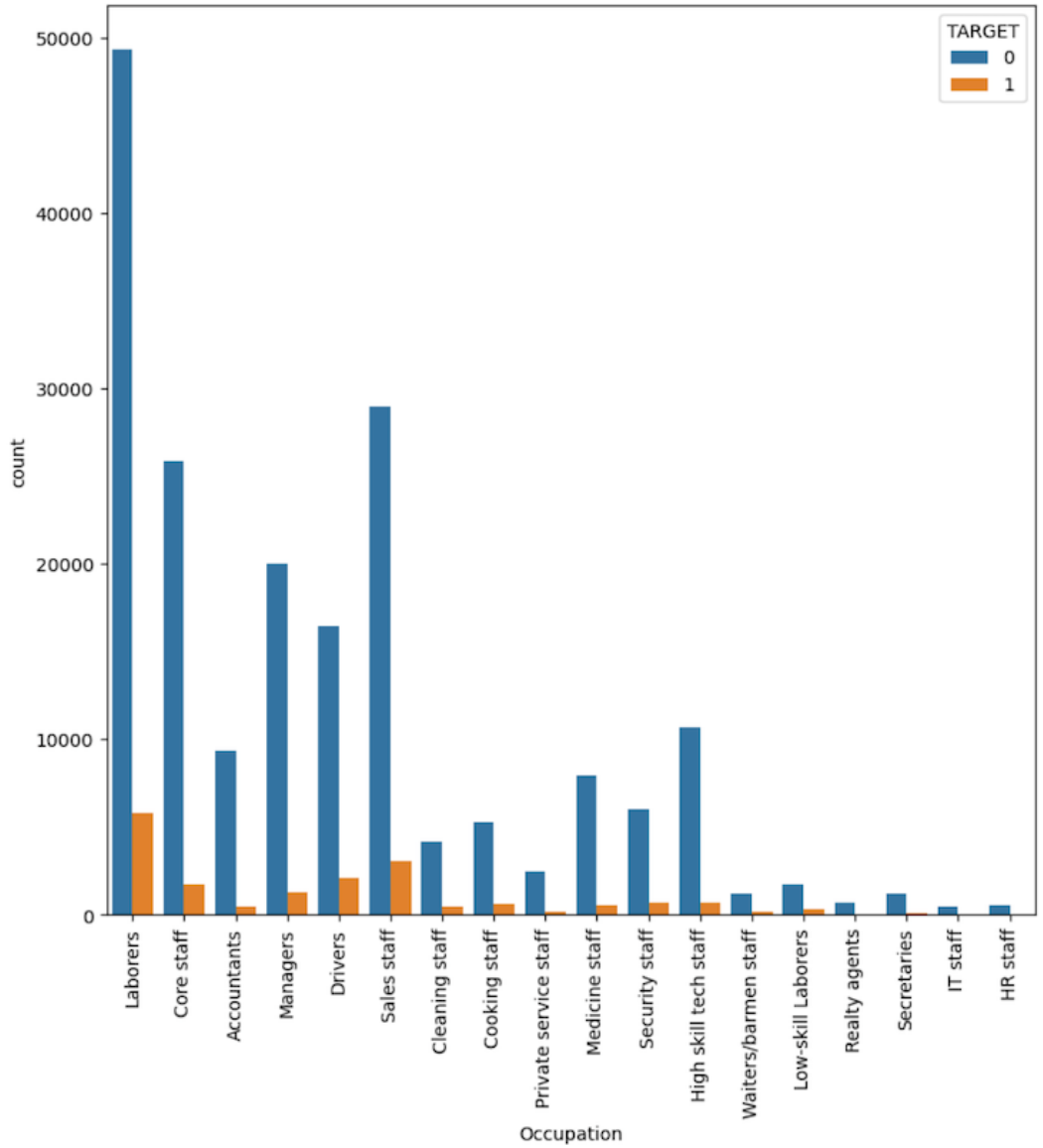


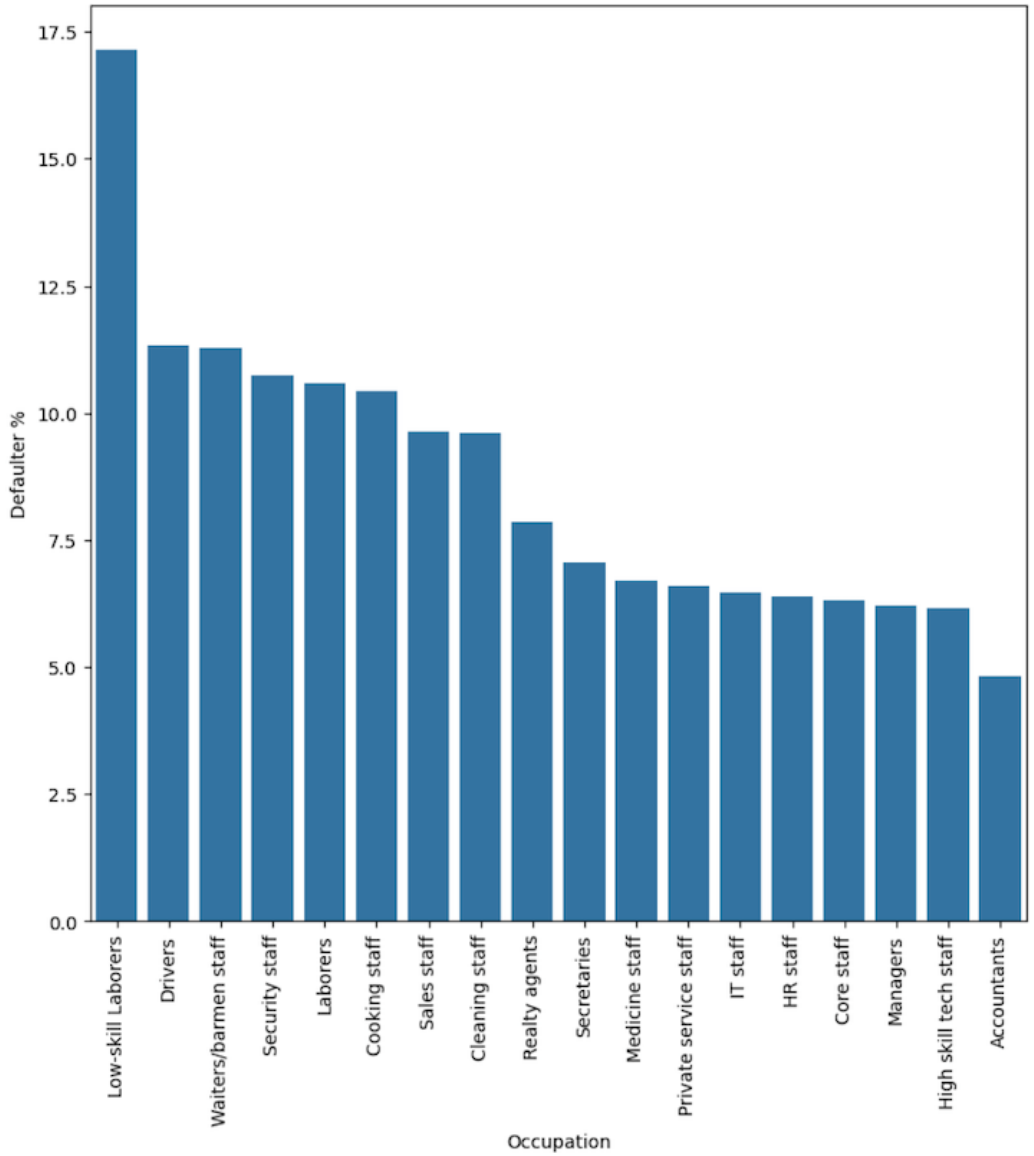
- **Clients with own House/Apartments have more number of loan applications where as clients with Rented Apartment or living with parents are more like to default the loan.**



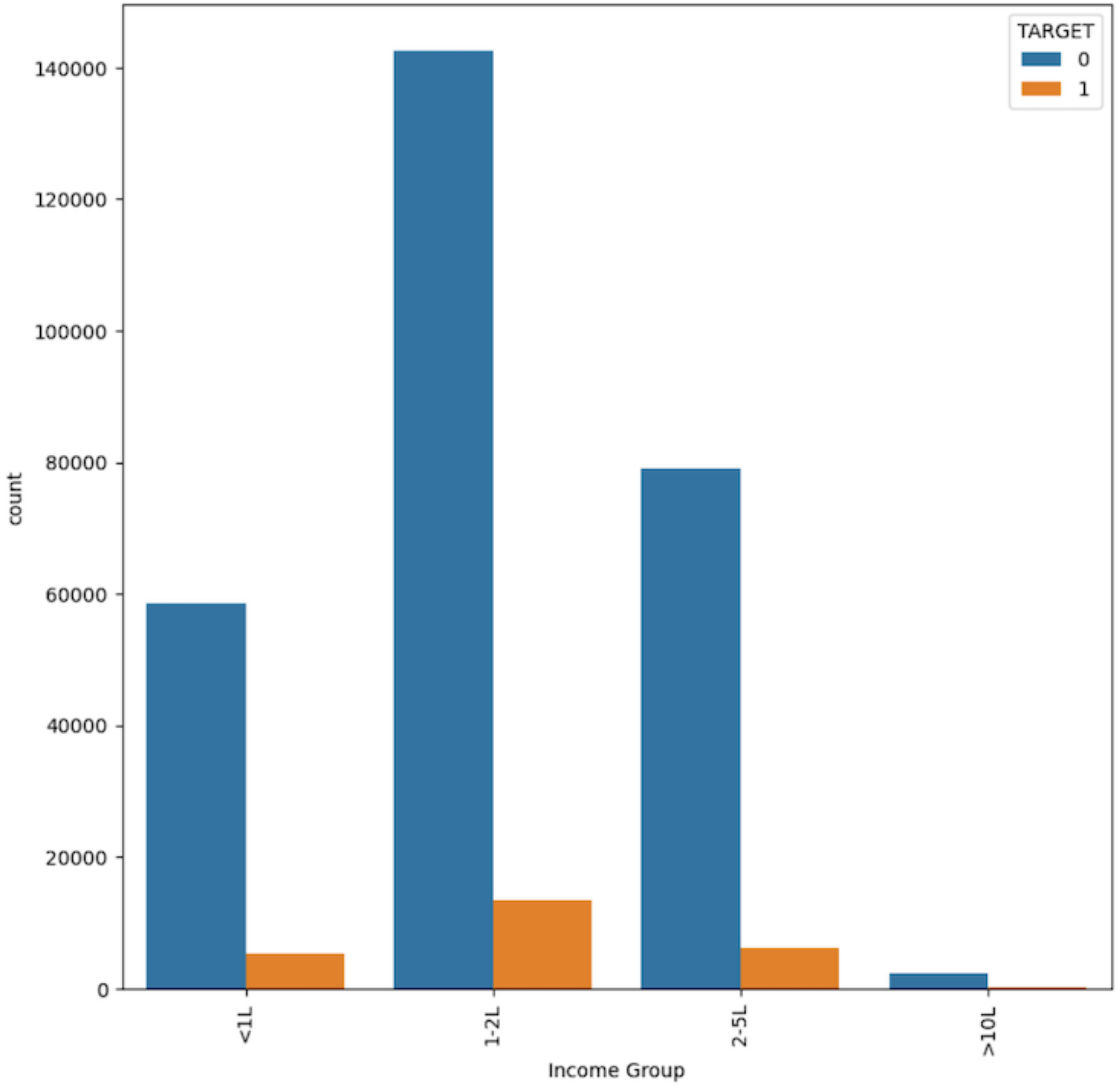


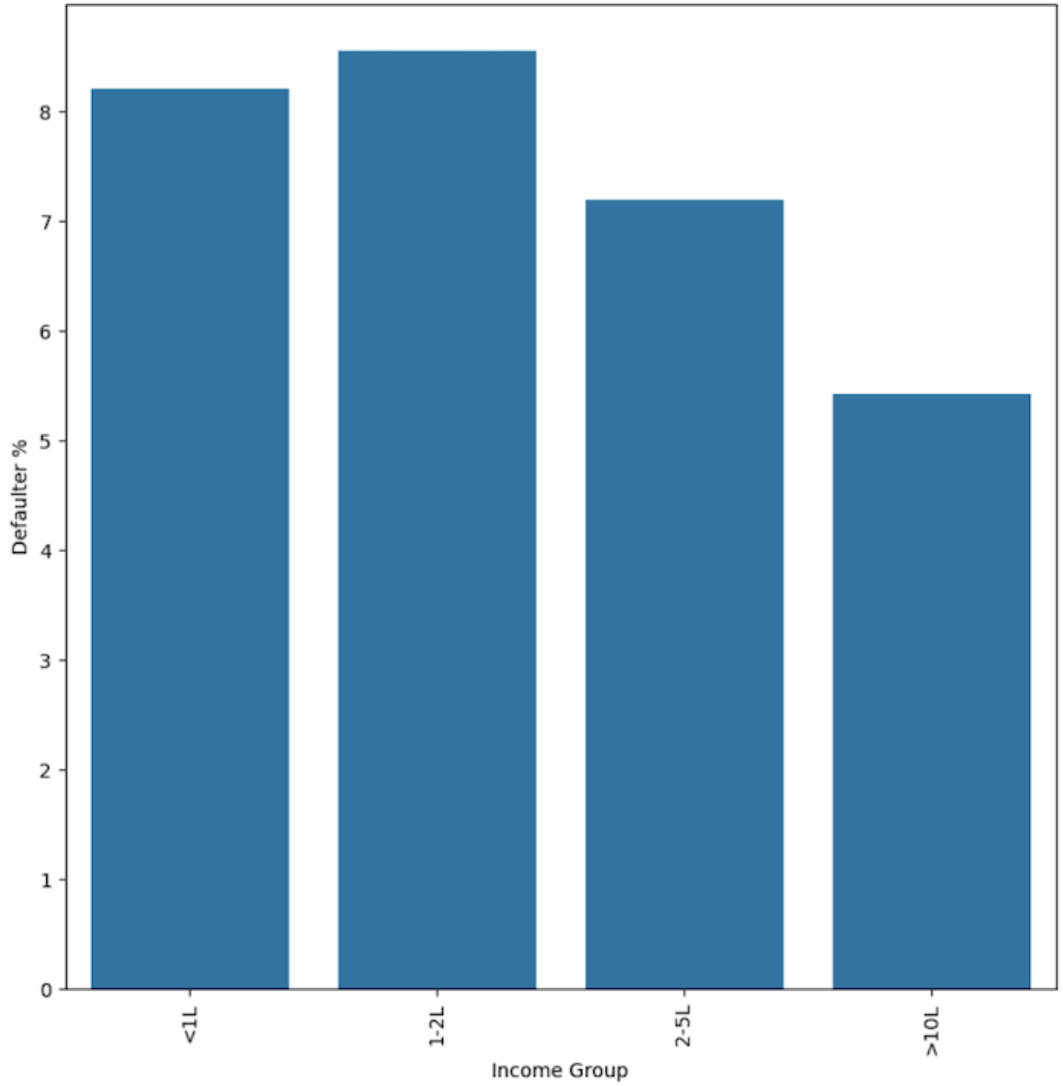
- **Clients who work as Laborers have more number of the loan applications and the are most likely to be defaulters.**



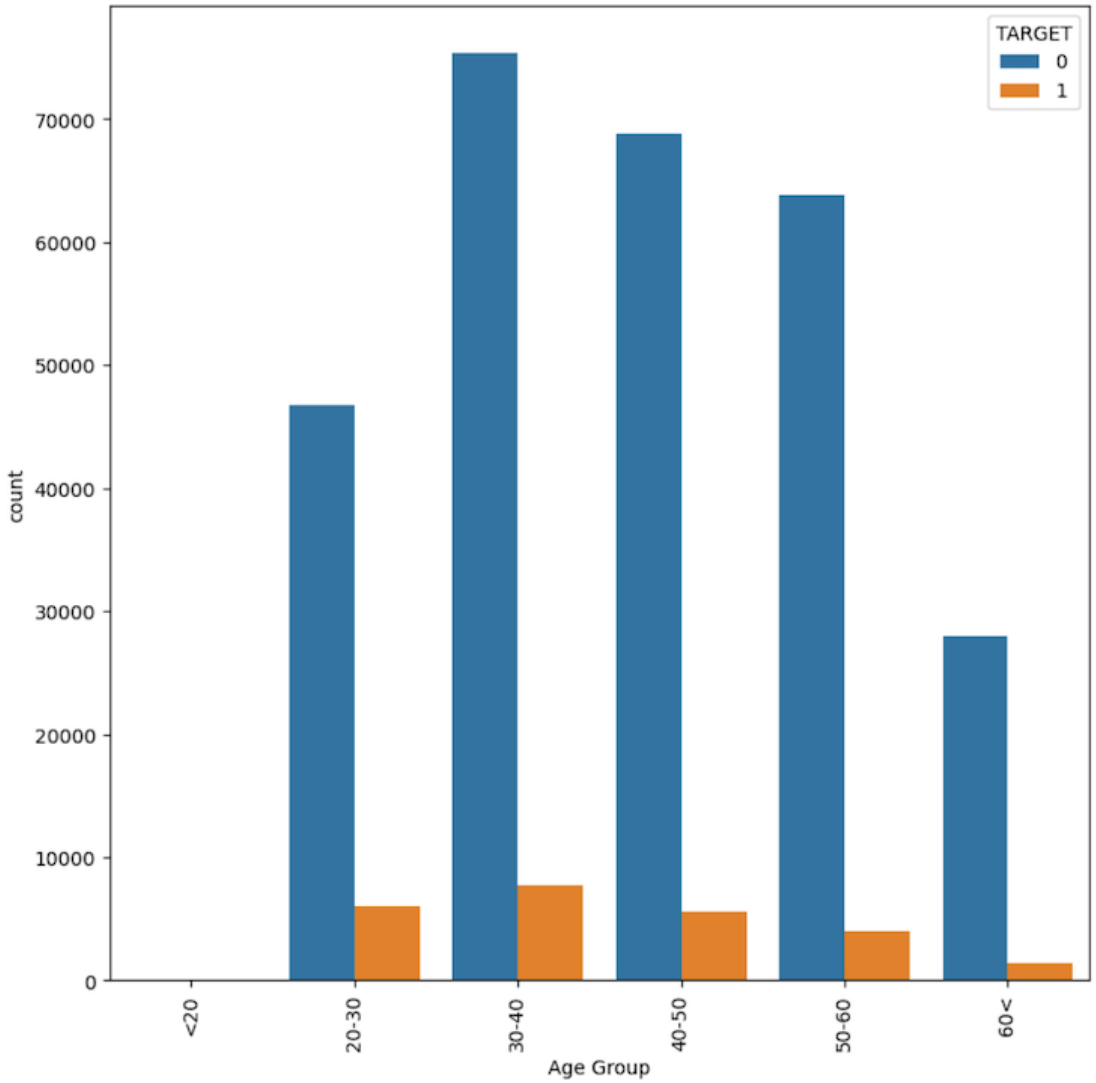


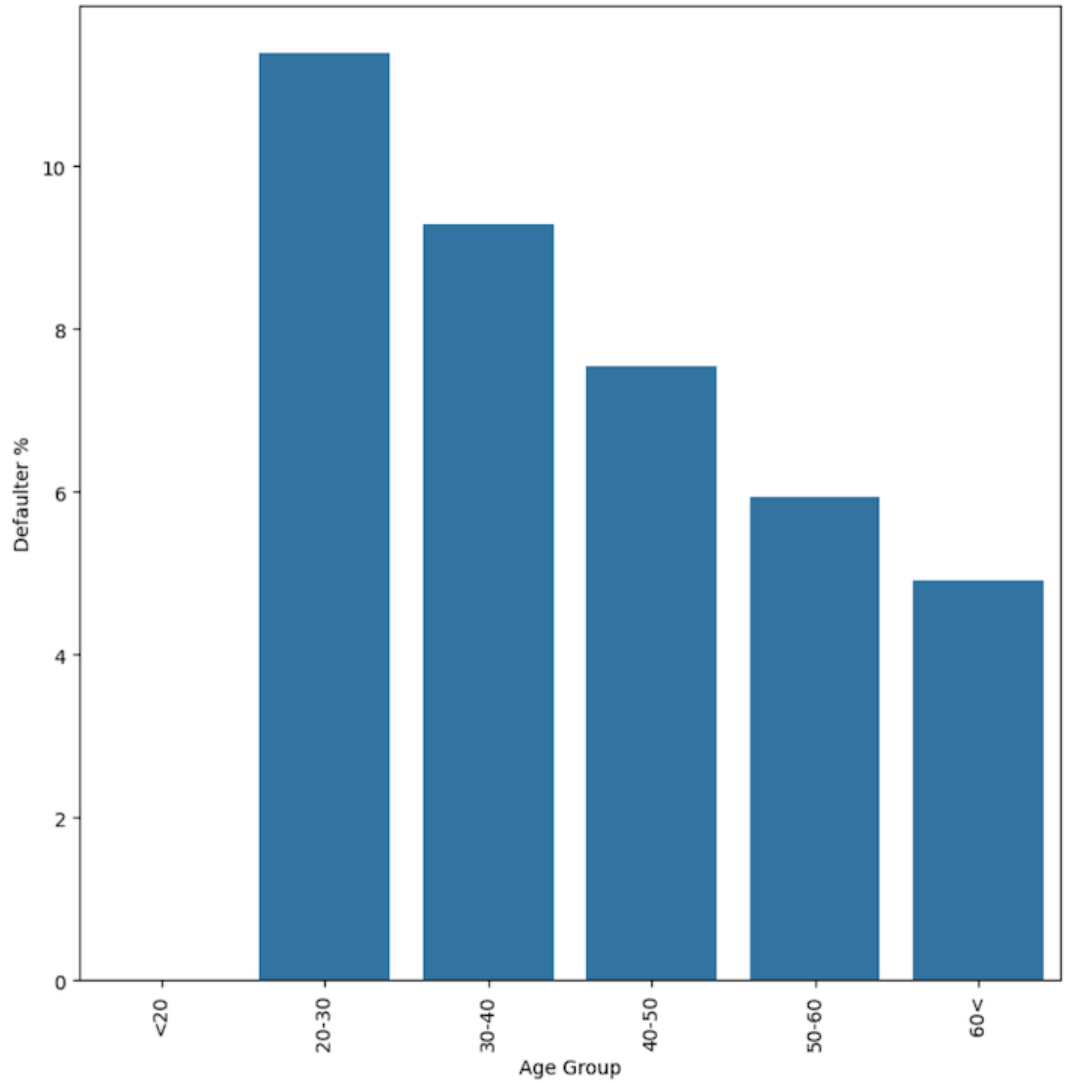
- **People having income between 1-2 Lakhs have most number of application in both the Target values where as people under income group 1-2 Lakhs & less than 1 Lakhs are most likely to default.**



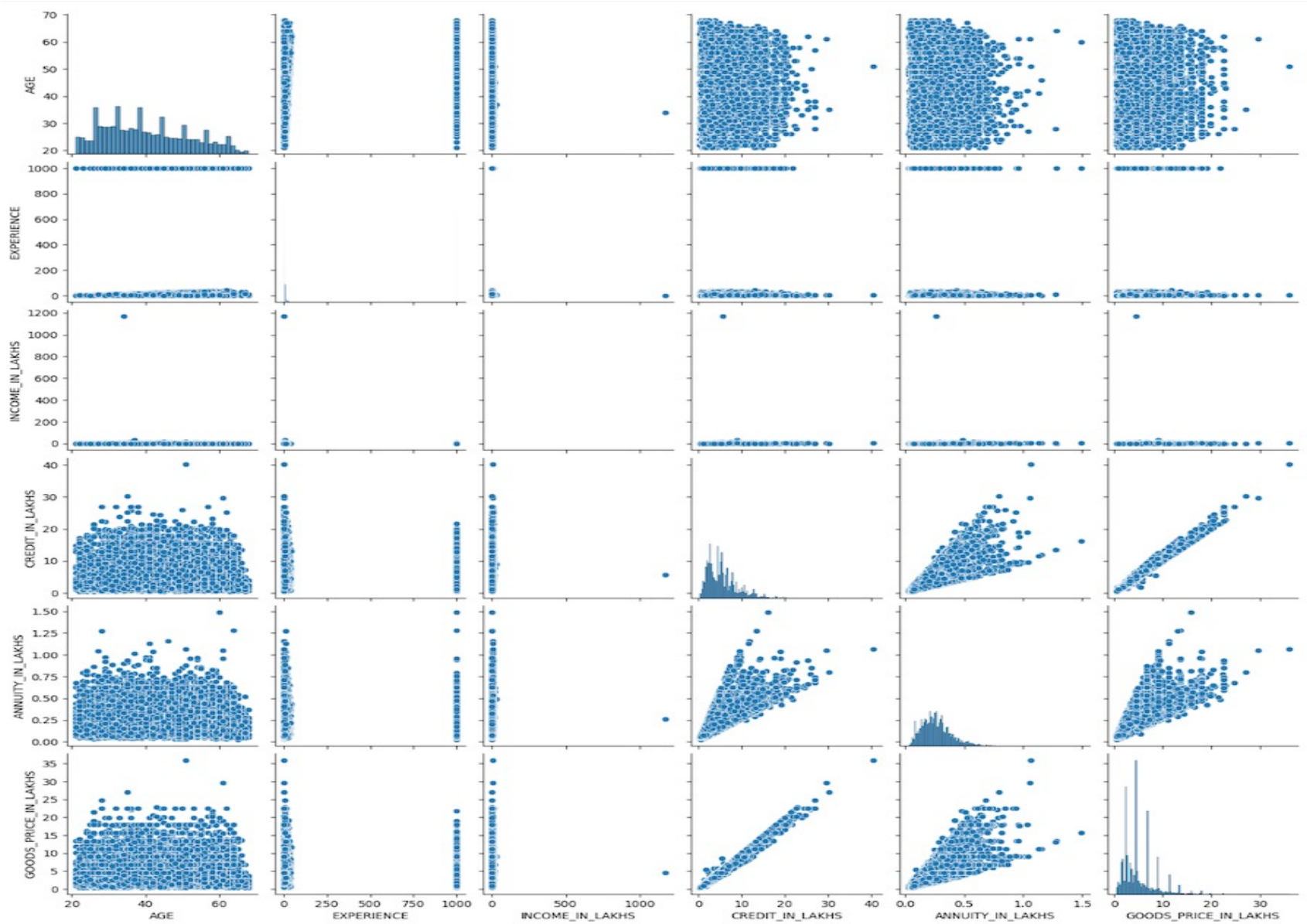


- **People withing age group 30-50 have most number of applications and people with in age group of 20-30 are most likely to default.**

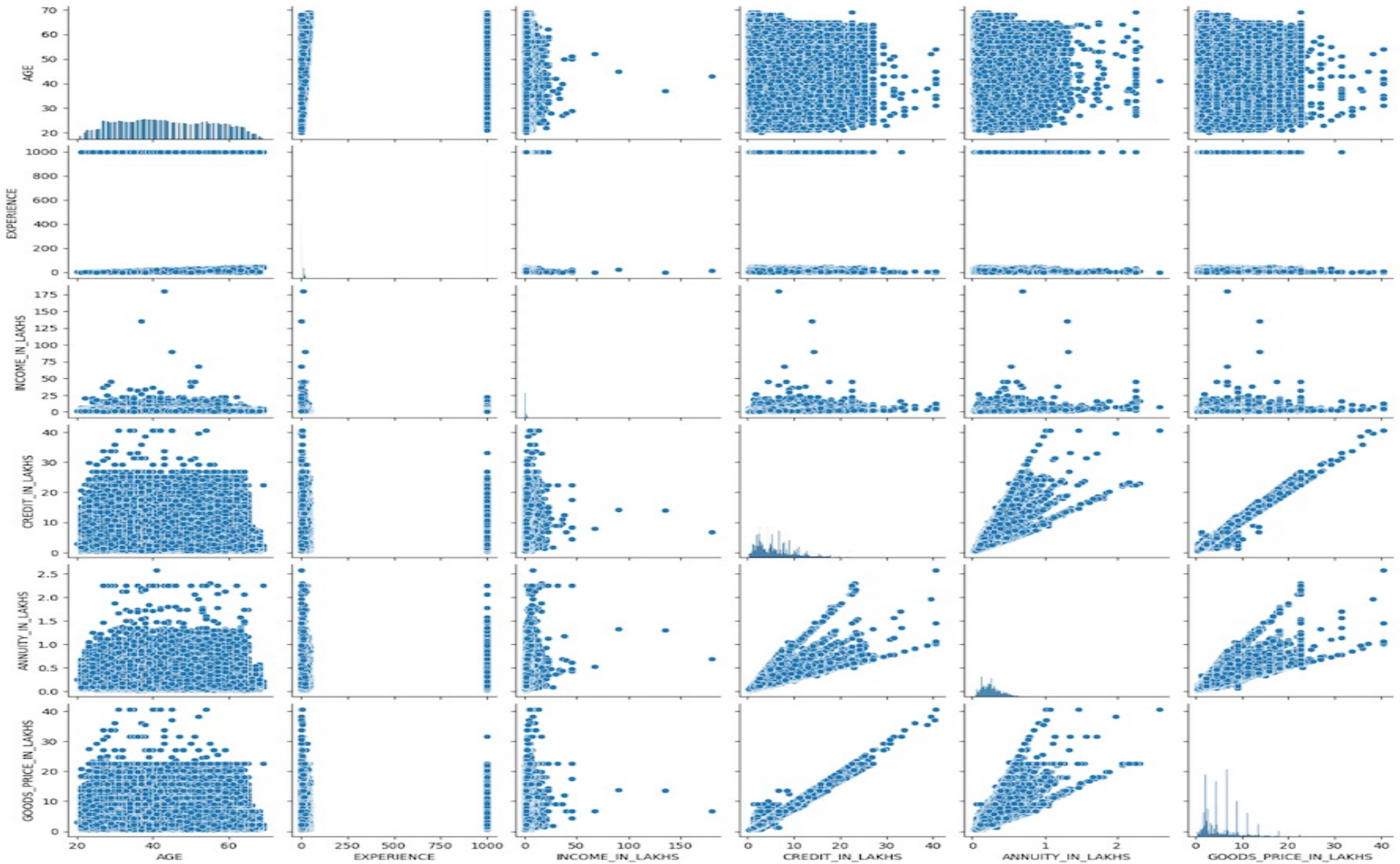




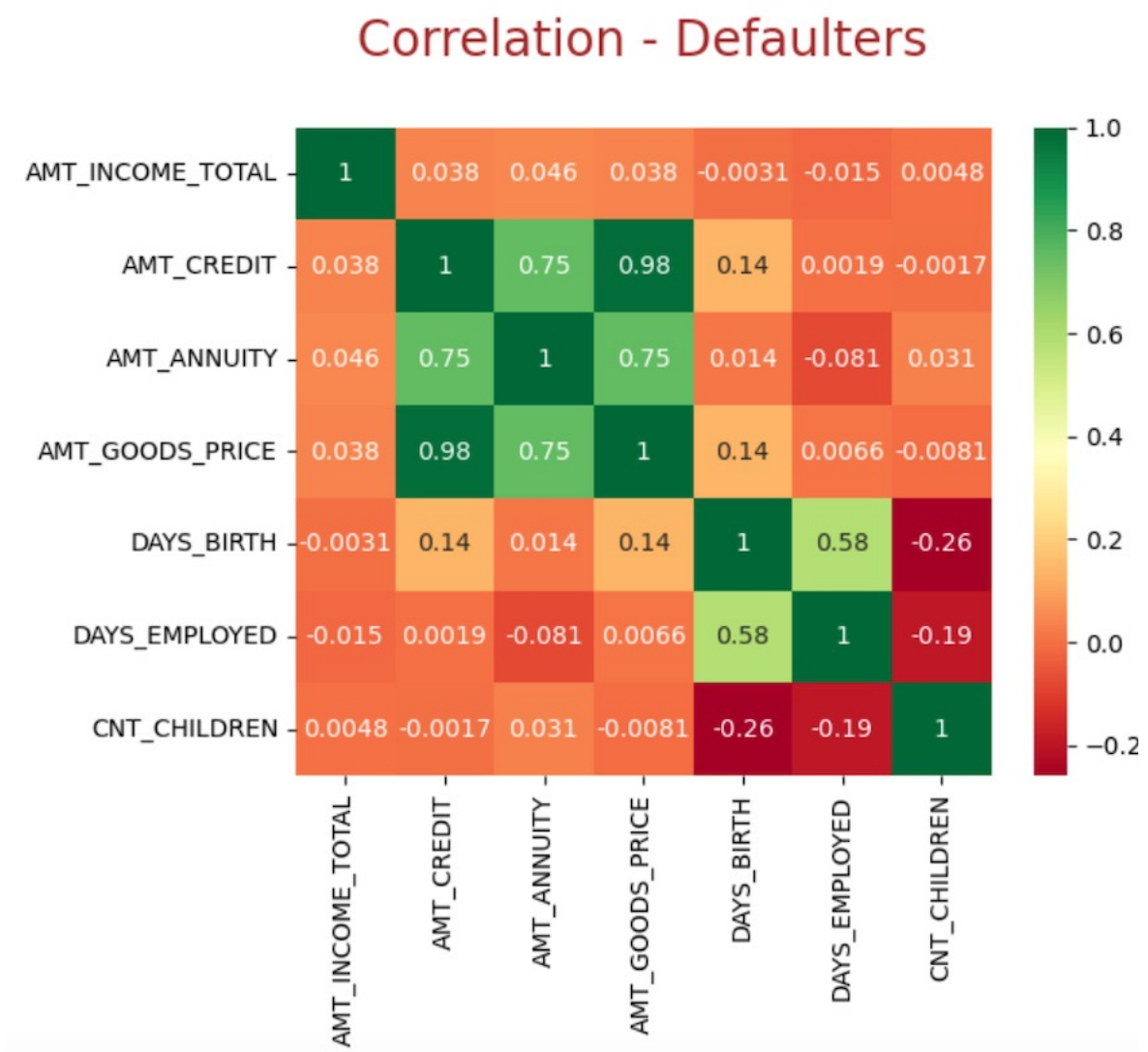
- Bivariate Analysis with Target = 1

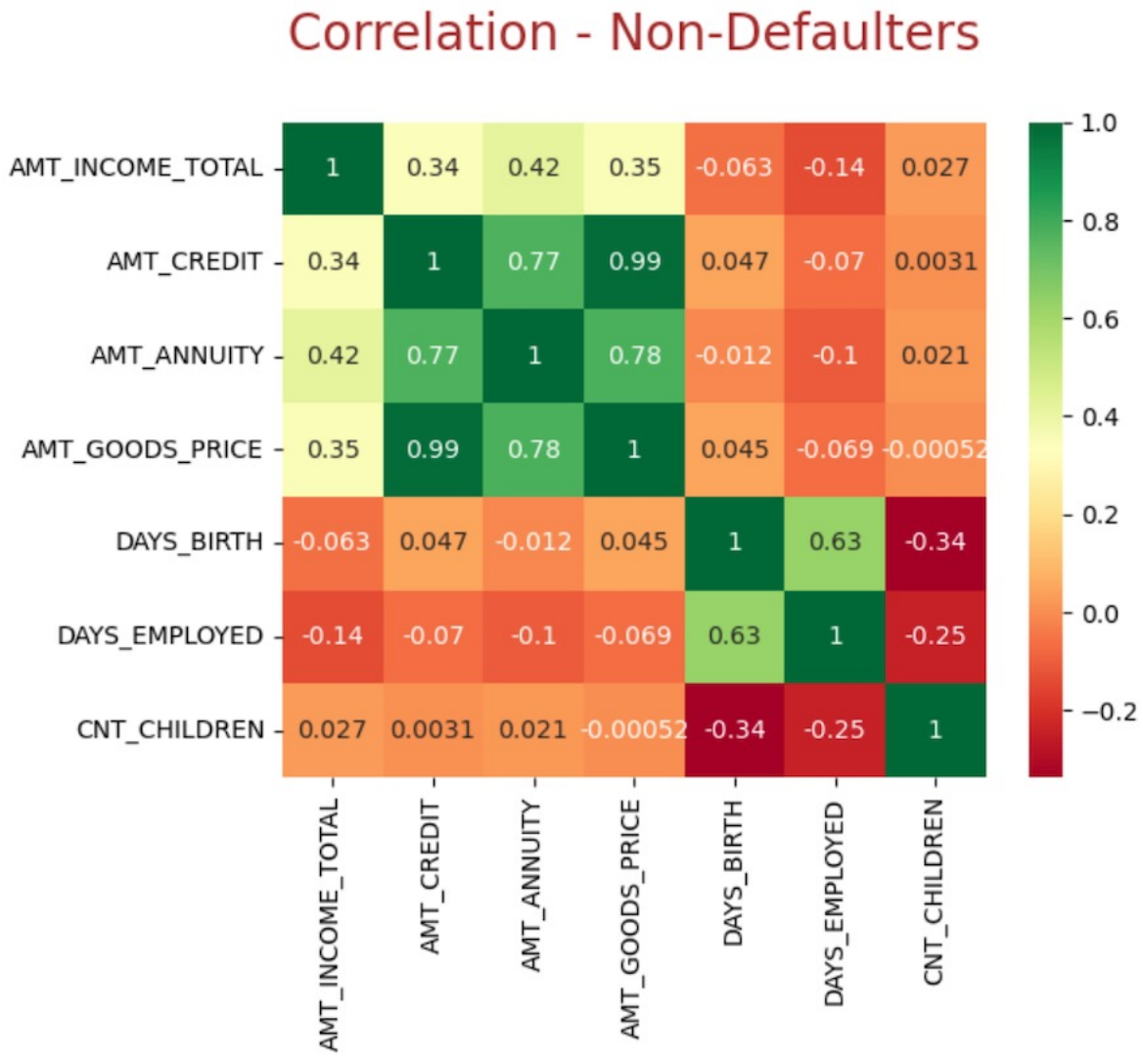


- Bivariate Analysis with Target = 0



- Correlation:



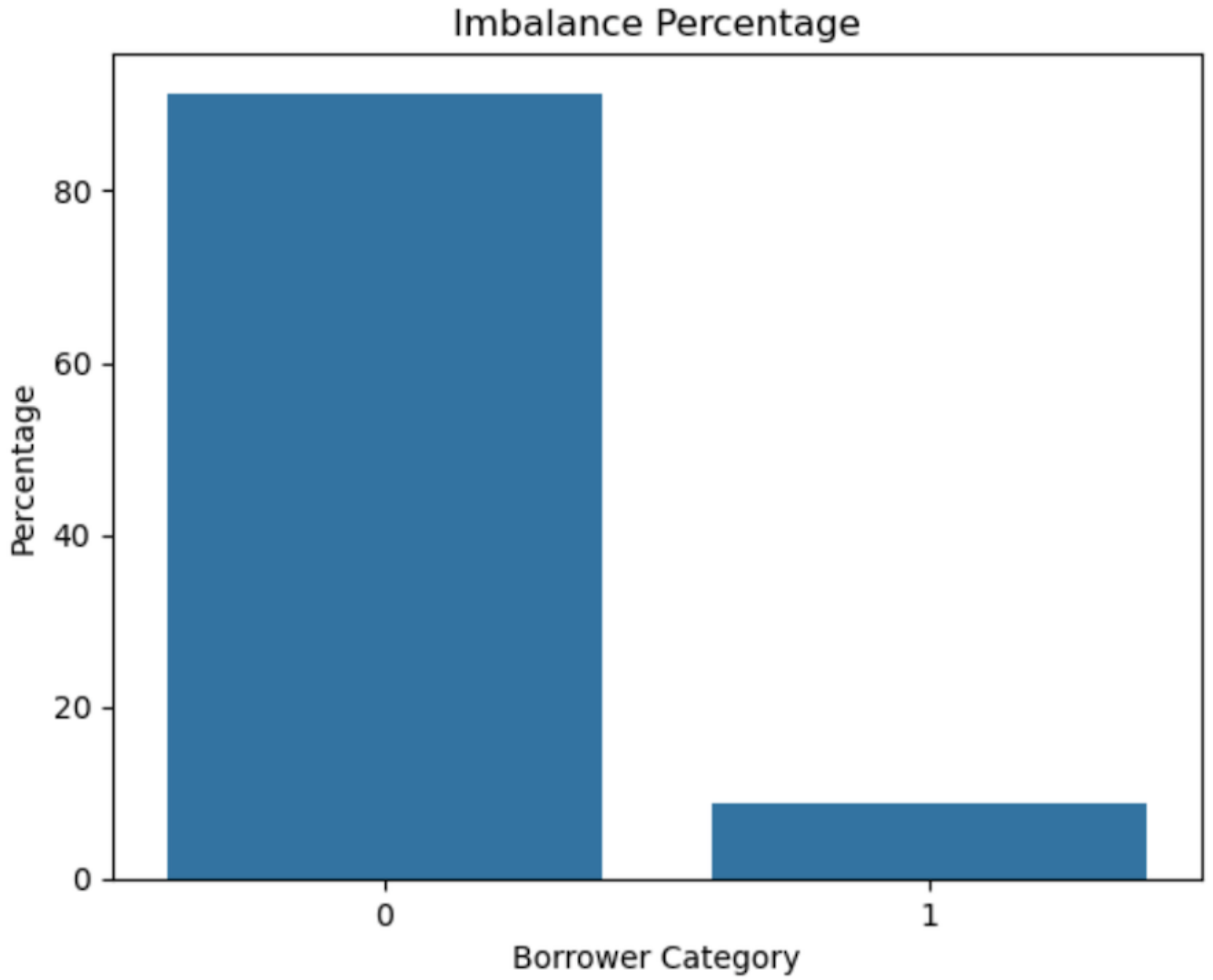


Analysis on Previous Application Data

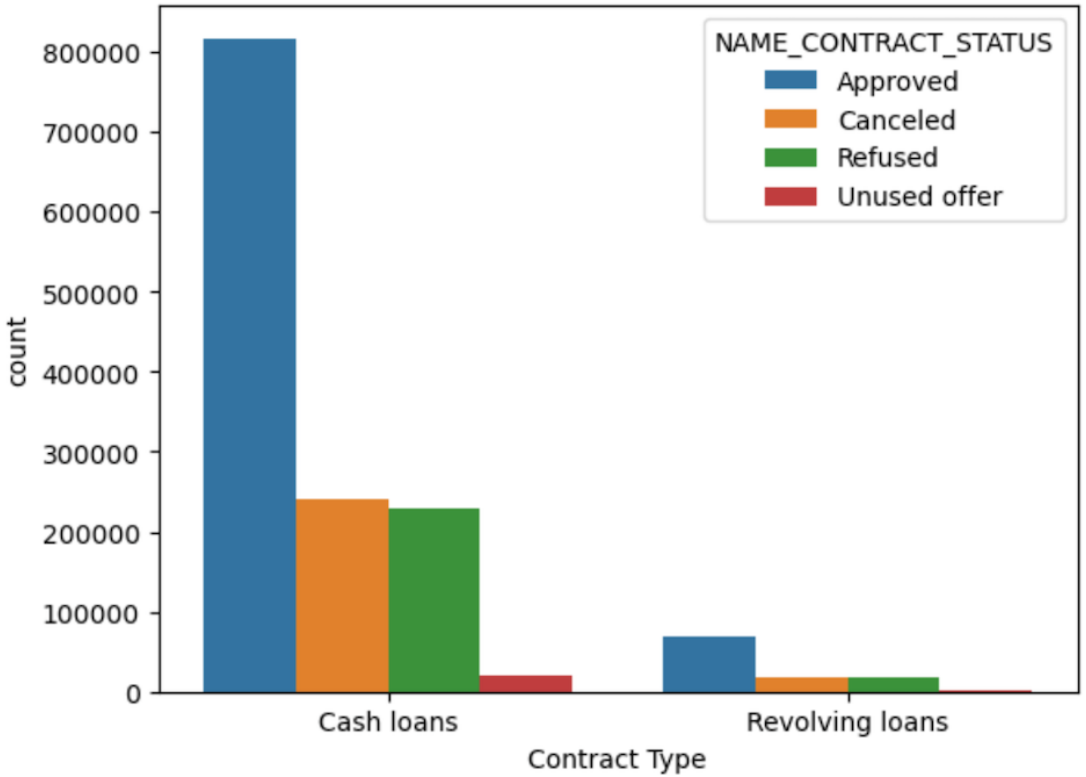
- Handling missing values:
  - In the data set the columns, ‘RATE_INTEREST_PRIMARY', 'RATE_INTEREST_PRIVILEGED', 'RATE_DOWN_PAYMENT', 'AMT_DOWN_PAYMENT’ were having more than 99% missing data. So have dropped them from the data set.
- Updating the negative values of the columns to positive:
  - The columns 'DAYS_FIRST_DRAWING’, 'DAYS_FIRST_DUE’, 'DAYS_LAST_DUE_1ST_VERSION’, 'DAYS_LAST_DUE’, 'DAYS_TERMINATION', were having some negative values which I have changed to positive as there columns are not supposed to have negative values.

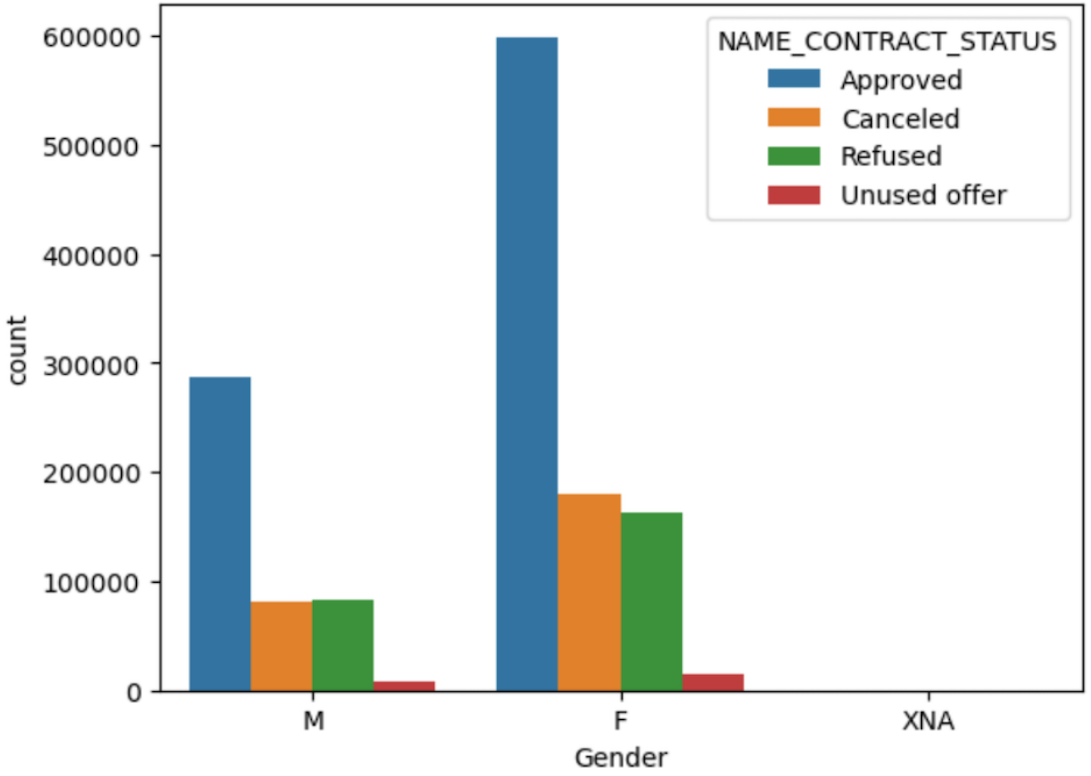
Merging the application_data and previous_application_data data set.

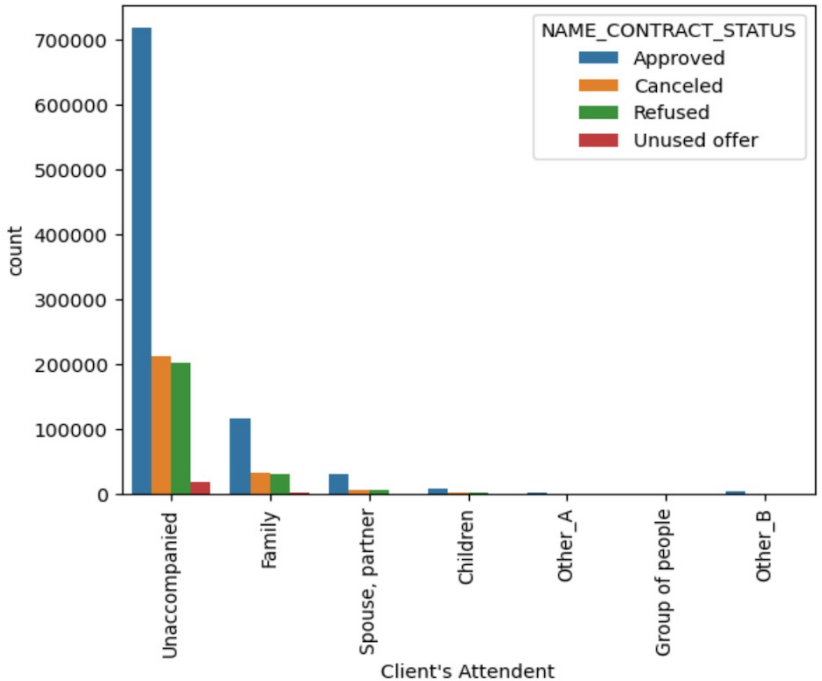
- Imbalance Percentage:

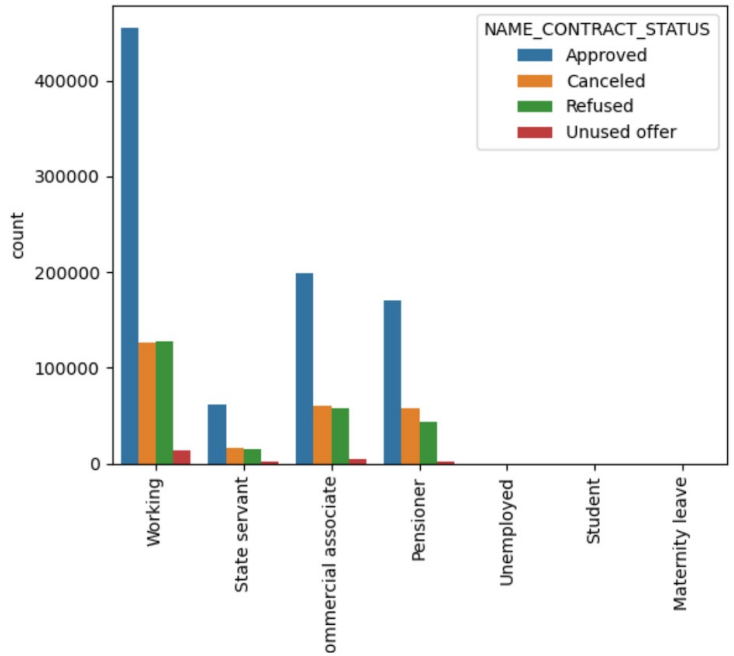


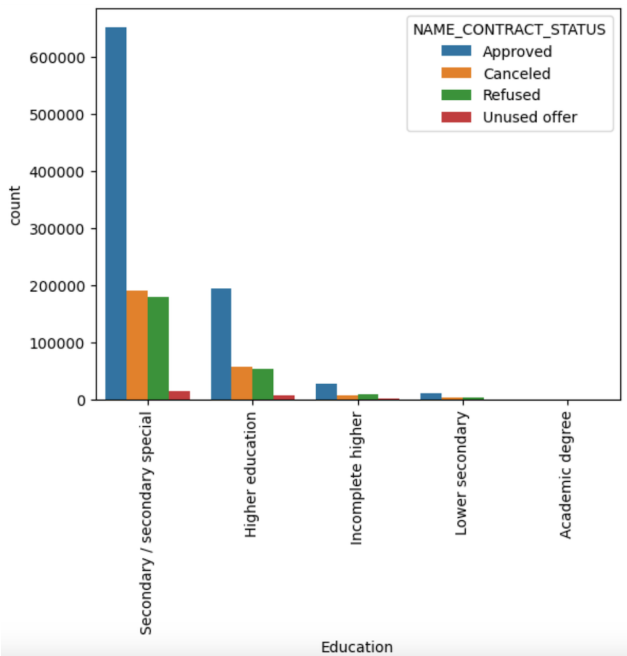
- **Performing Univariate Analysis on Categorical Variables:**
  - **Number of approved loans is higher in both the type. Also Cancelled and Refused loans have same proportion in both the types.**
  - **Number of approved loans in Female is higher than Male.**
  - **Clients with no attendant have higher numbers of loan approval.**
  - **Working class employees have more approved application.**
  - **Clients with Secondary/secondary special educations have more number of approved loan application.**
  - **Married clients have more number of approved loans**

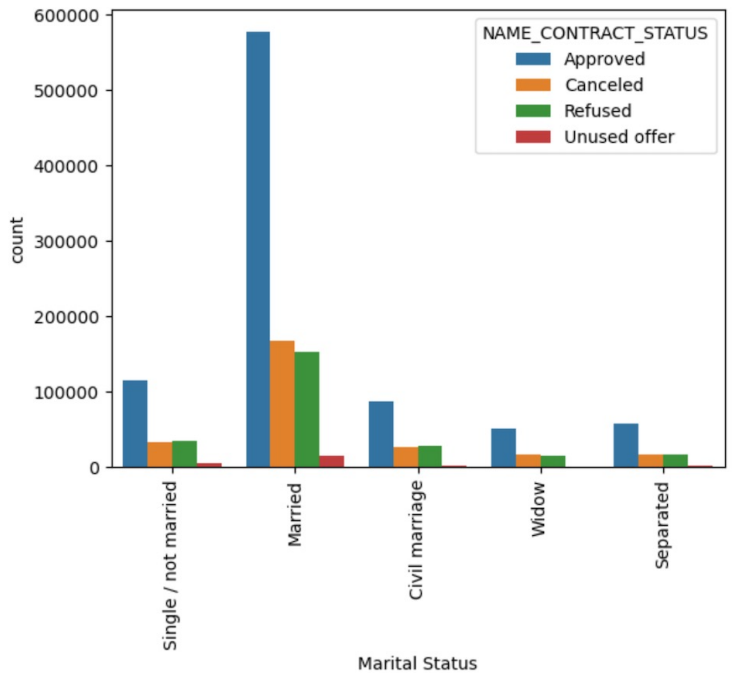




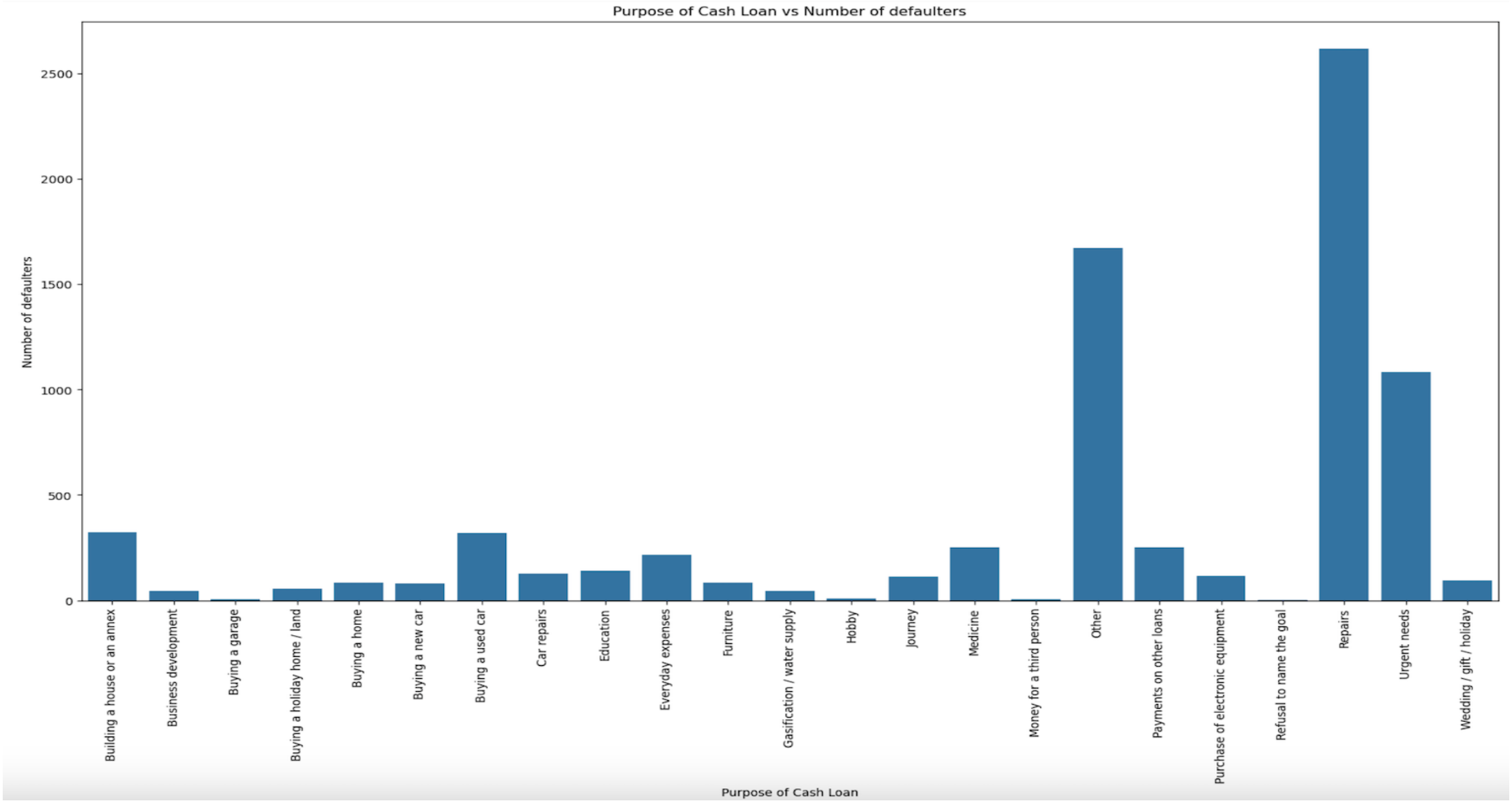

NAME_CONTRACT_STATUS
Approved
Canceled
Refused
Unused offer
count
700000
600000
500000
400000
300000
200000
100000
0
Unaccompanied
Family
Spouse, partner
Children
Other_A
Group of people
Other_B
Client's Attendent


NAME_CONTRACT_STATUS
Approved
Canceled
Refused
Unused offer
count
400000
300000
200000
100000
0
Working
State servant
ommercial associate
Pensioner
Unemployed
Student
Maternity leave


NAME_CONTRACT_STATUS
Approved
Canceled
Refused
Unused offer
count
600000
500000
400000
300000
200000
100000
0
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
Education

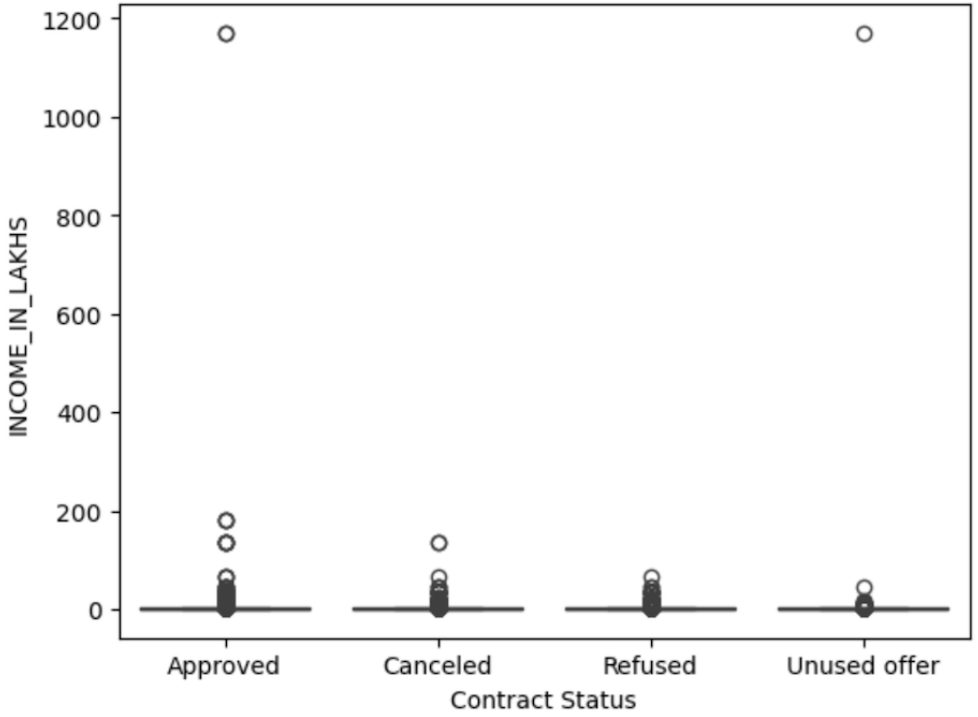

NAME_CONTRACT_STATUS
Approved
Canceled
Refused
Unused offer
count
600000
500000
400000
300000
200000
100000
0
Single / not married
Married
Civil marriage
Widow
Separated
Marital Status

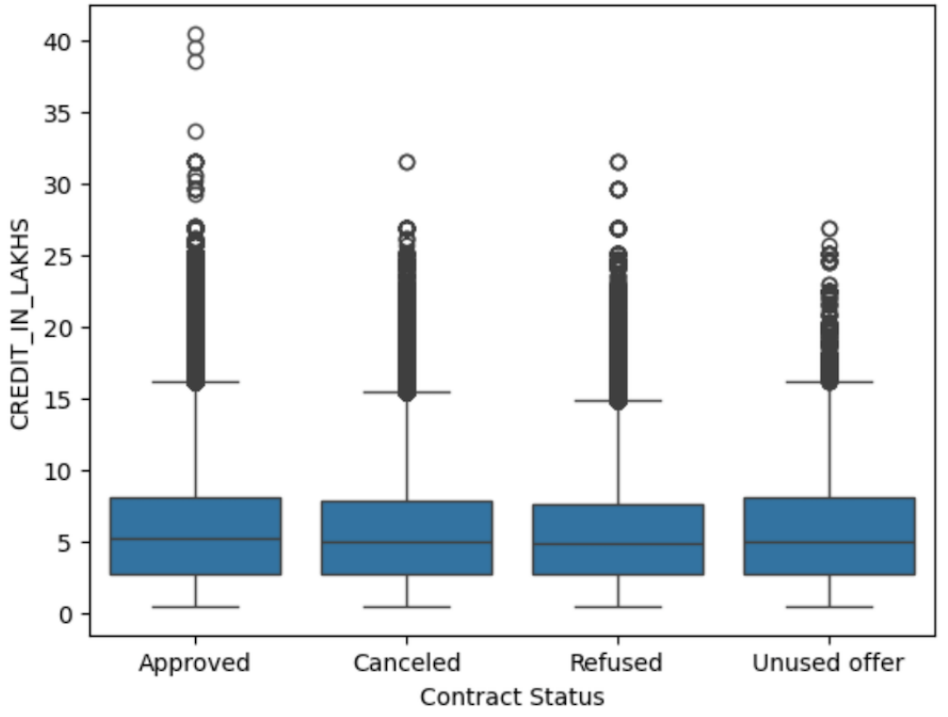
- # Plotting graph of Defaulters vs Purpose of cash loan:
  - **Clients having purpose of buying a garage, home and hobby etc, have no difficulty in repayment.**
  - **Clients with loan applications for Repairs, Urgent needs and Others etc. are more likely to default.**

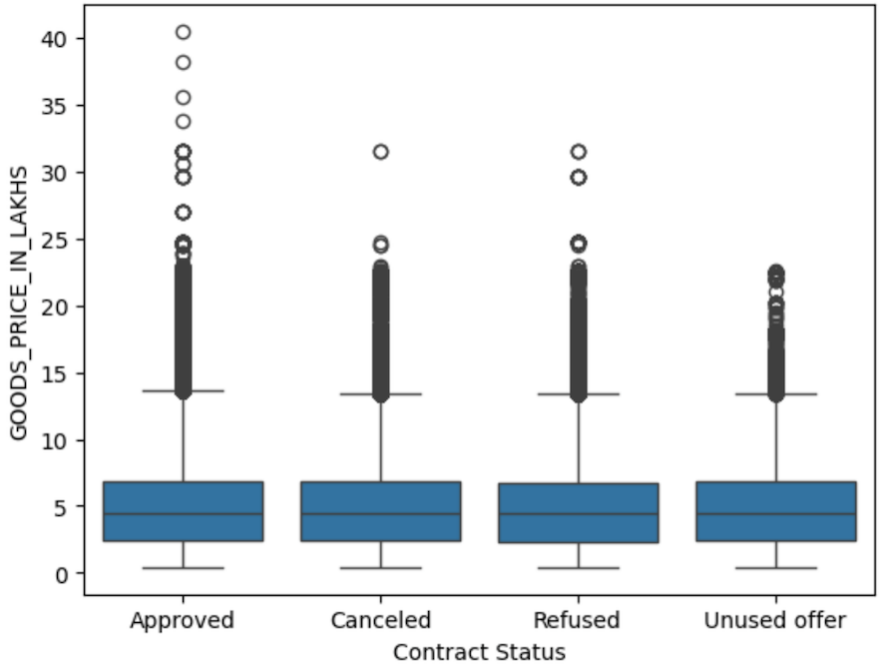


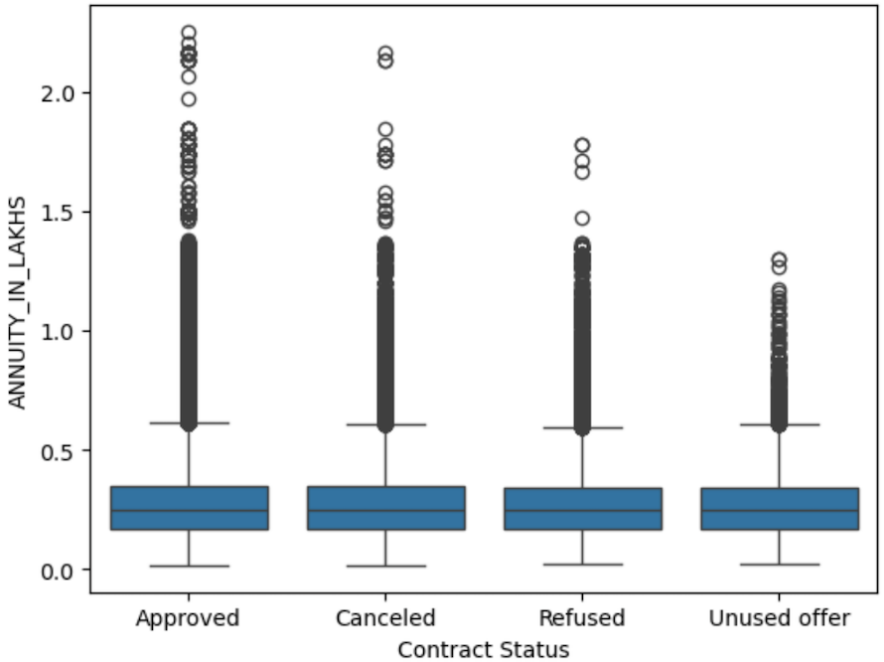
Univariate analysis on numerical variables

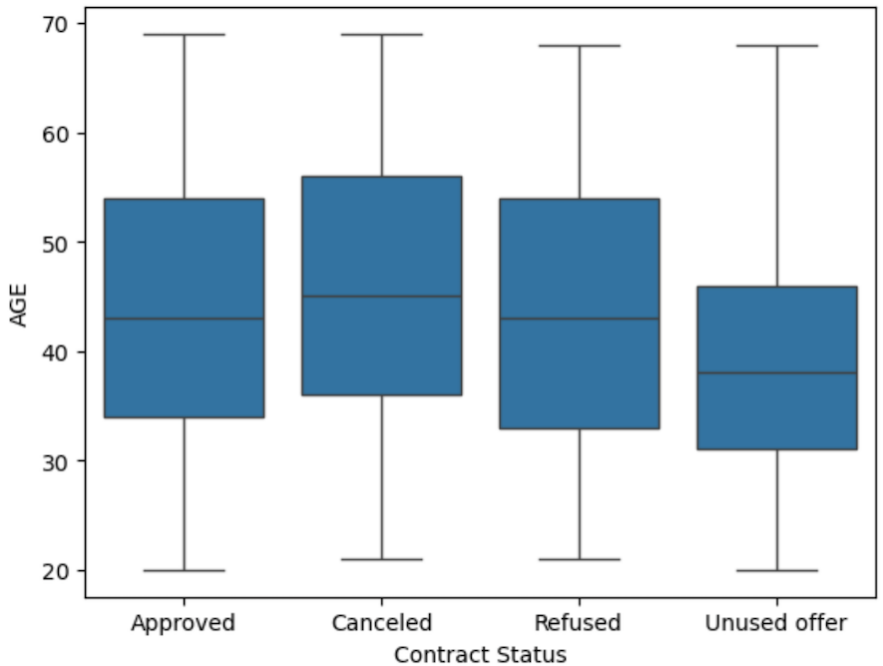
- **All variables seem to follow a similar distribution for all cases except for CLIENT_AGE where all four cases have a similar distribution but in different age ranges**

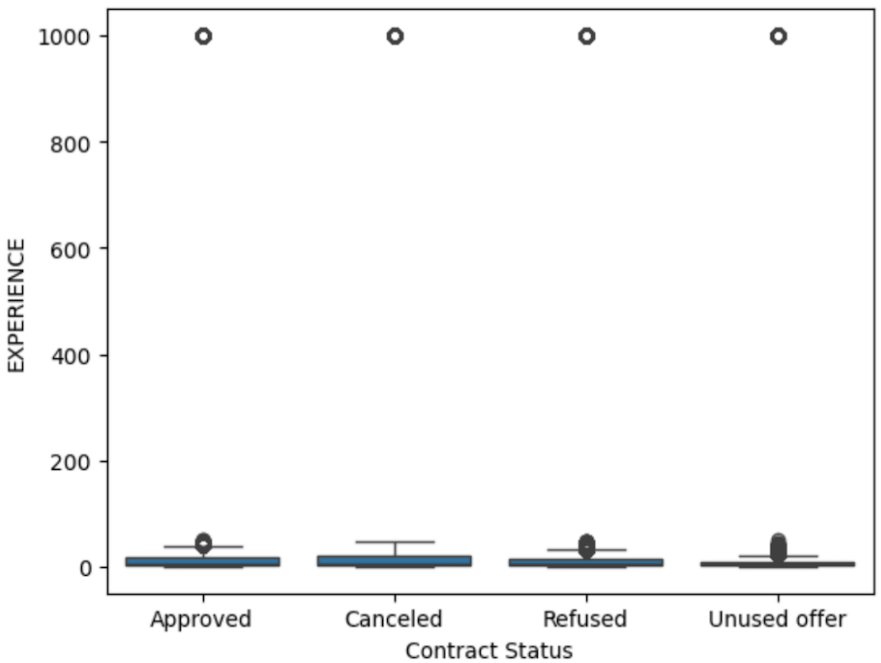




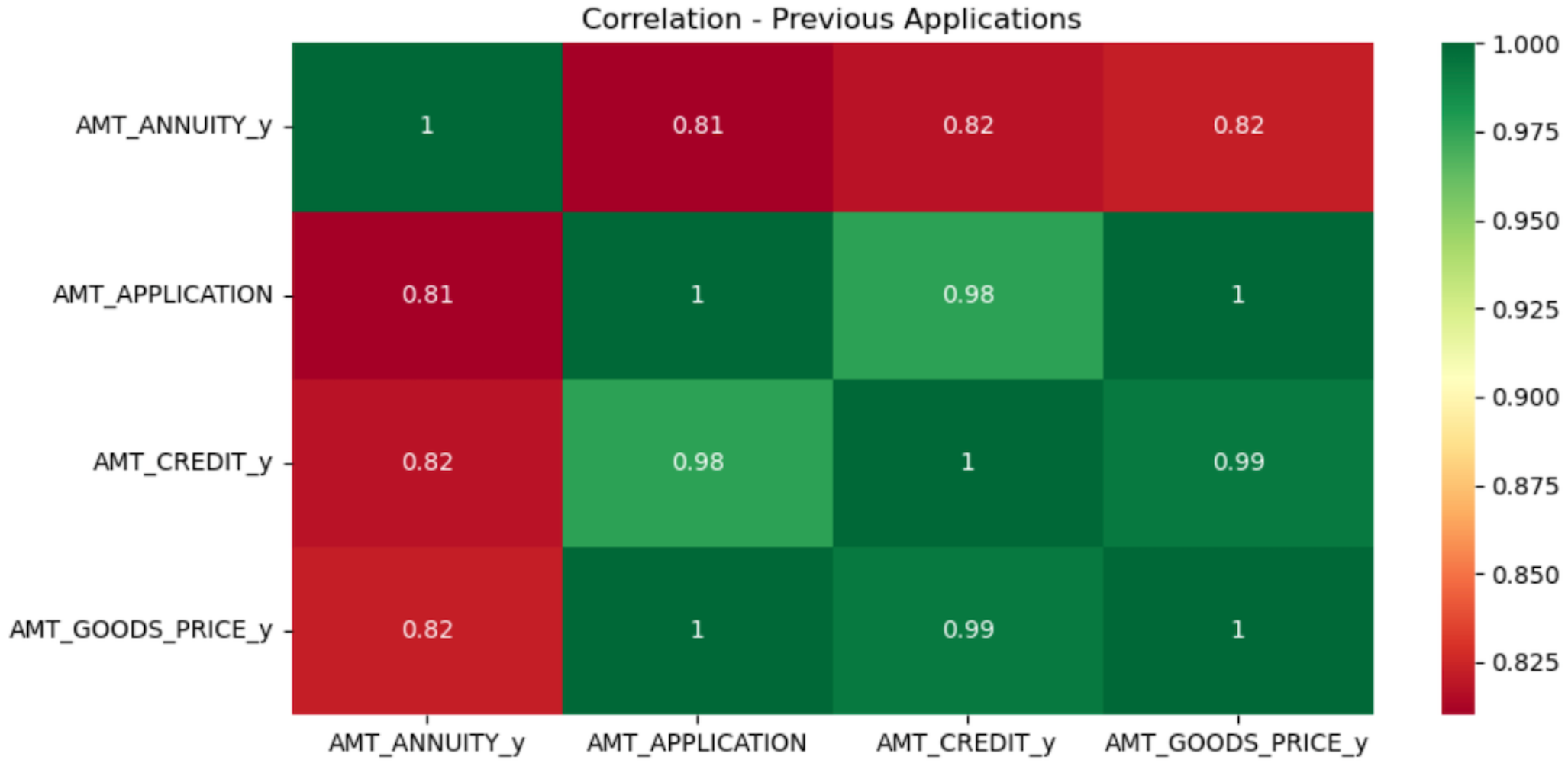

GOODS_PRICE_IN_LAKHS
0
5
10
15
20
25
30
35
40
Approved
Canceled
Refused
Unused offer
Contract Status


ANNUITY_IN_LAKHS
0.0
0.5
1.0
1.5
2.0
Approved
Canceled
Refused
Unused offer
Contract Status


AGE
20
30
40
50
60
70
Approved
Canceled
Refused
Unused offer
Contract Status


EXPERIENCE
0
200
400
600
800
1000
Approved
Canceled
Refused
Unused offer
Contract Status

Correlation heatmap for numerical variables



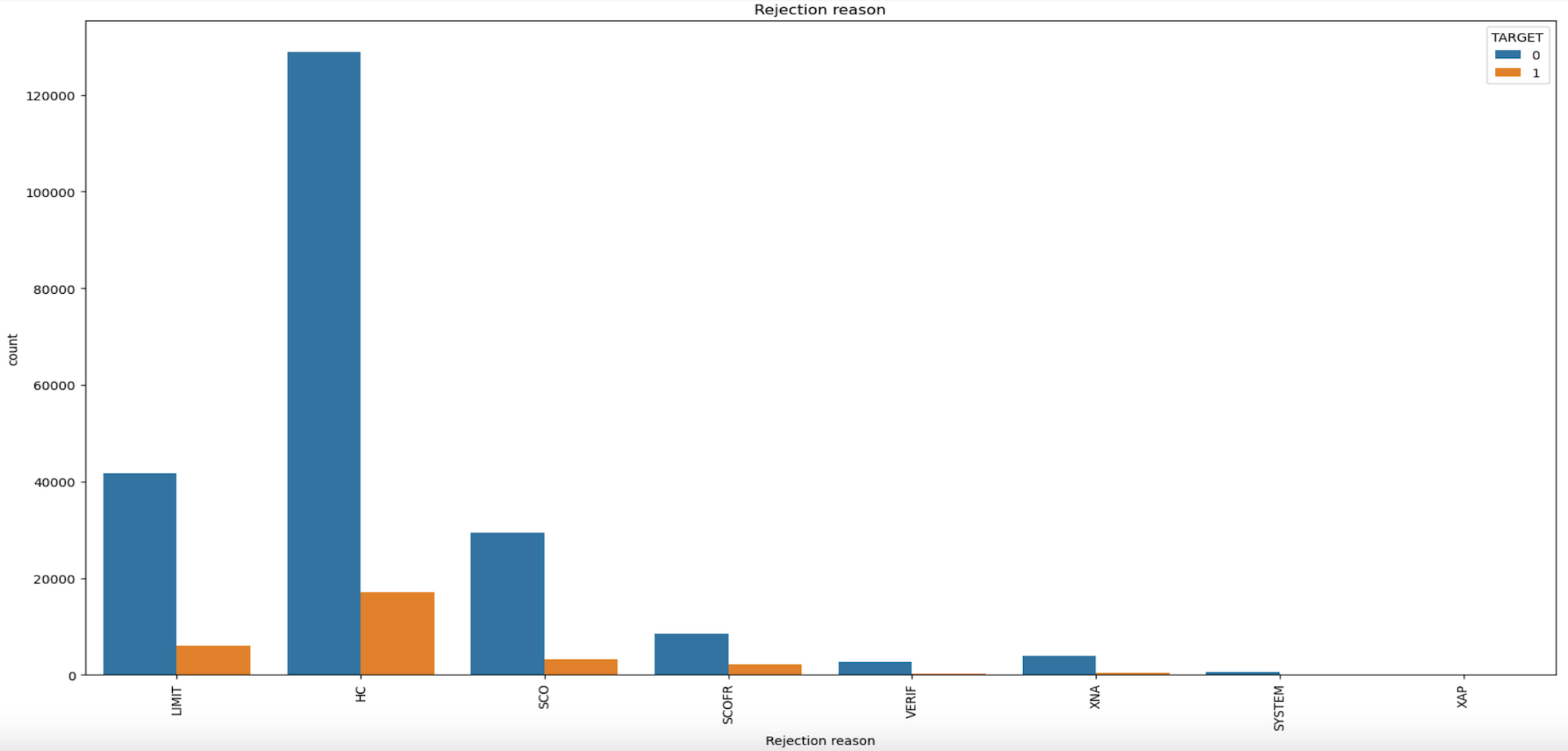
AMT_CREDIT_PREV is highly correlated to AMT_APPLICATION and AMT_GOODS_PRICE_PREV

Bivariate Analysis

- **High chances of defaulting for lower credit amount, applied amount and goods price.**
- **The chance of default decreases with an increase in credit amount, applied amount and goods price.**
- **Credit Amount is highly correlated to Goods Price and the Amount applied by the client on previous loan applications.**

Rejected Reasons:

- **Rejection by system is very less.**
- **Rejection code - HChas higher number of rejections. It also has the higher number of defaulters.**



Conclusion:

- **Based on the above Analysis, we can conclude the below:**
  - **Bank can approve loans taken for purpose of buying home or garage as there are less chances of defaulting.**
  - **Bank can refuse the applicants who are taking cash loans for the purpose of urgent need or repairs.**
  - **Bank can offer loan on higher rate to the people who have unused offers and comparatively high total income are more likely to default.**
  - **Bank can reduce the loan amount for female applicants who are on maternity leave.**
  - **Bank should reject the application of young males with lower secondary education and of lower income group and staying with parents or in a rented house, applying for low-range cash contract**
  - **Banks can target businessmen, students and working class people with academic degree/ higher education to provide loan.**