Credit EDA Assignment

By Vaibhav Bhargava

DS C-42 Batch

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

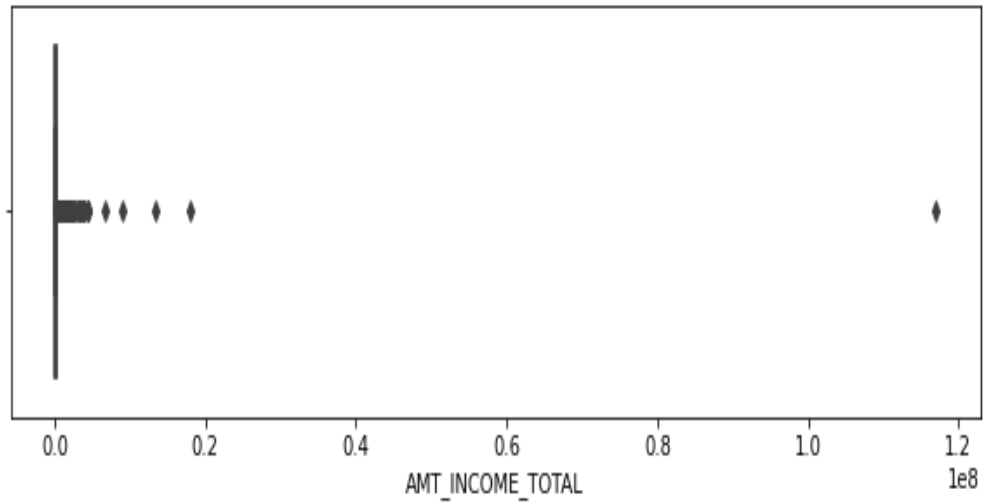
Business Objective

- This case study aims to identify patterns which indicate if a client has difficulty paying their installments 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.
- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

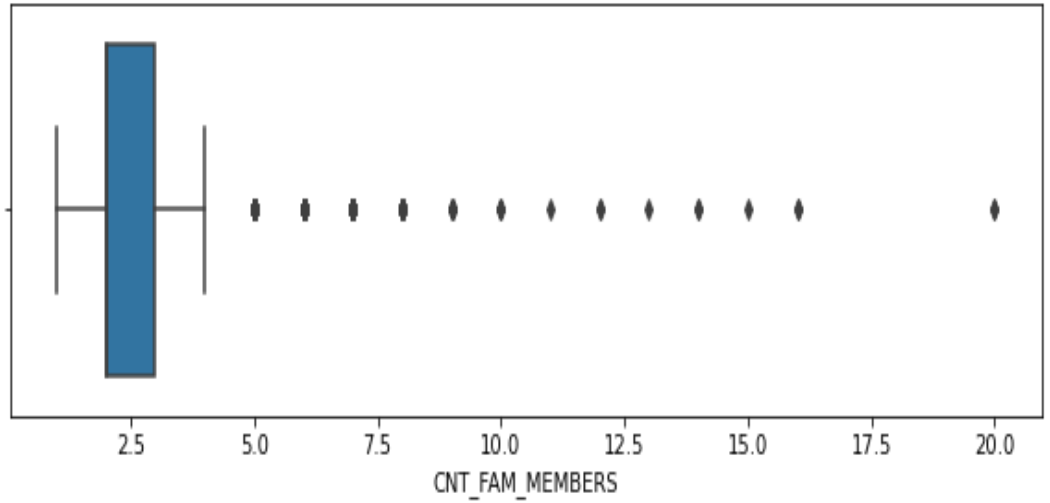
Various Steps Involved

- Importing the Required Libraries
- Importing the given Datasets.
- Do basic Data inspection like shape, types, count of null values etc.
- Data Pre-processing and Cleaning which includes following
  - Dropping Columns with High Missing Values Percentage approx. 50%
  - Imputation of Missing Values with Mean, Mode, Median or Zero with Missing Values is less if required.
  - Carry out Outlier Analysis
  - Binning & Bucketing
- Data Imbalance Check with respect to Target Variable
- Segmenting the Given Datasets into two Data Sets based on Target Variable Value
- Carrying out Univariate & Bivariate analysis on Categorical Variables
- Carrying out Correlation Check for Numerical Variables
- Joining the two datasets using inner join method
- Again Carrying out Univariate & Bivariate analysis
- Conclude the Insights

Outlier analysis-application dataset

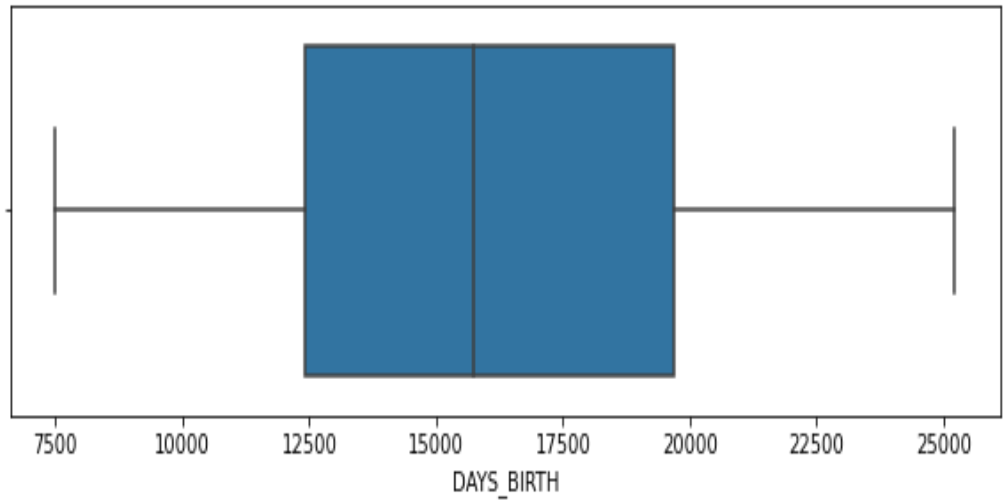


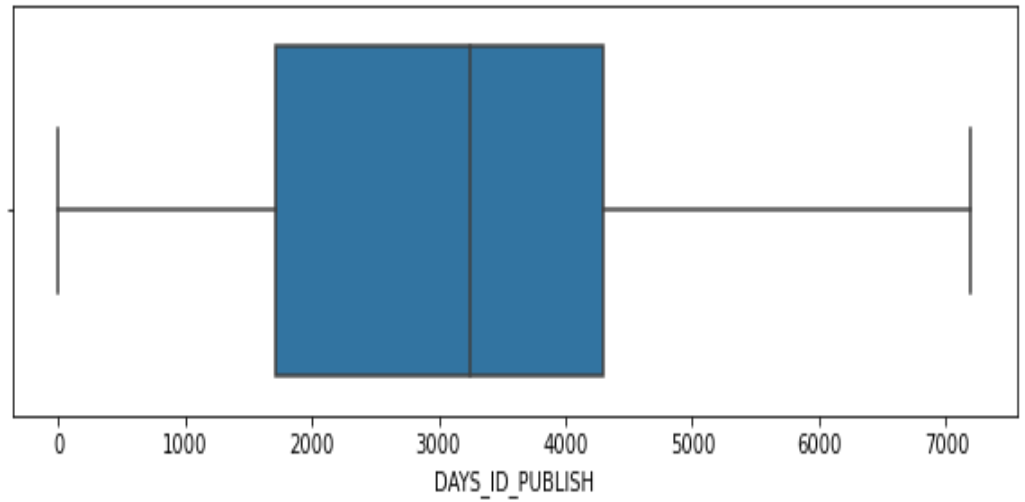
Variable AMT_INCOME_TOTAL is having very skewed box plot which may represent that the income has too much variation it from lowest to maximum values



Variable CNT_FAM_MEMBERS have outliers present in it

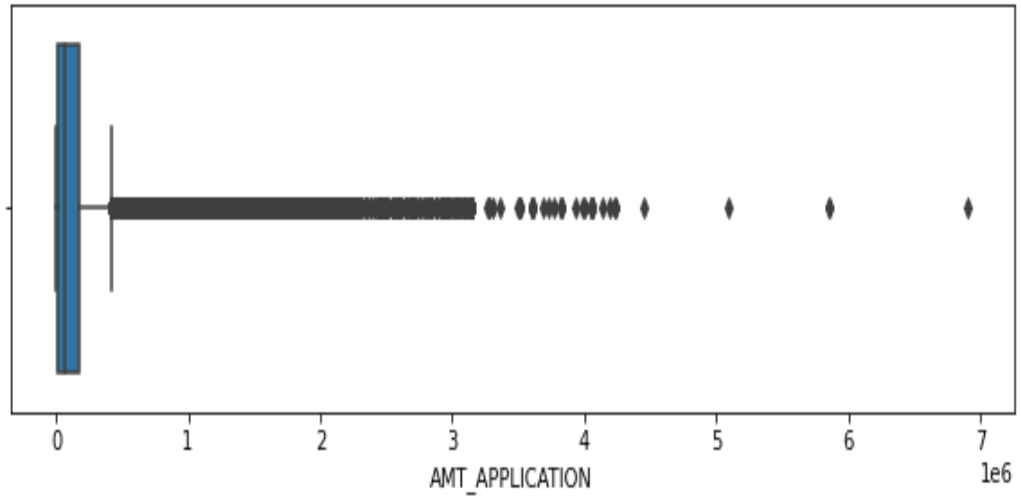
Outlier analysis-application dataset

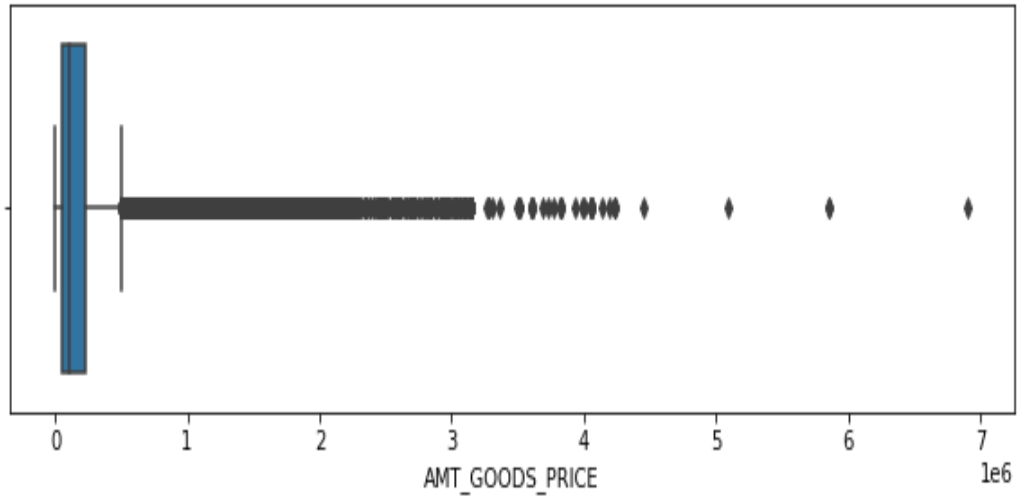




Variables DAYS_BIRTH and DAYS_ID_PUBLISH don't have any outlier in it. So we can use them as for further analysis purposes

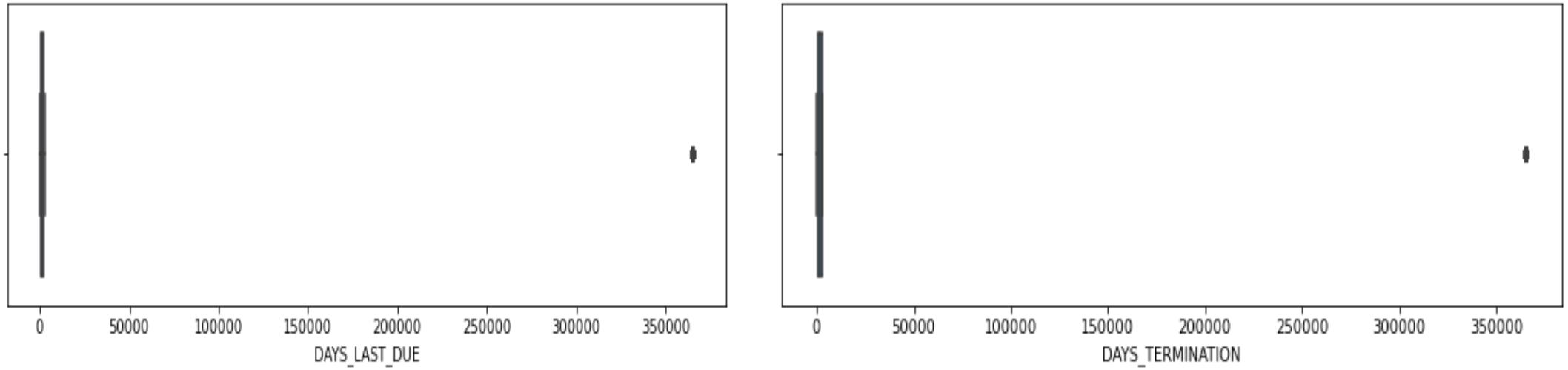
Outlier analysis-application dataset





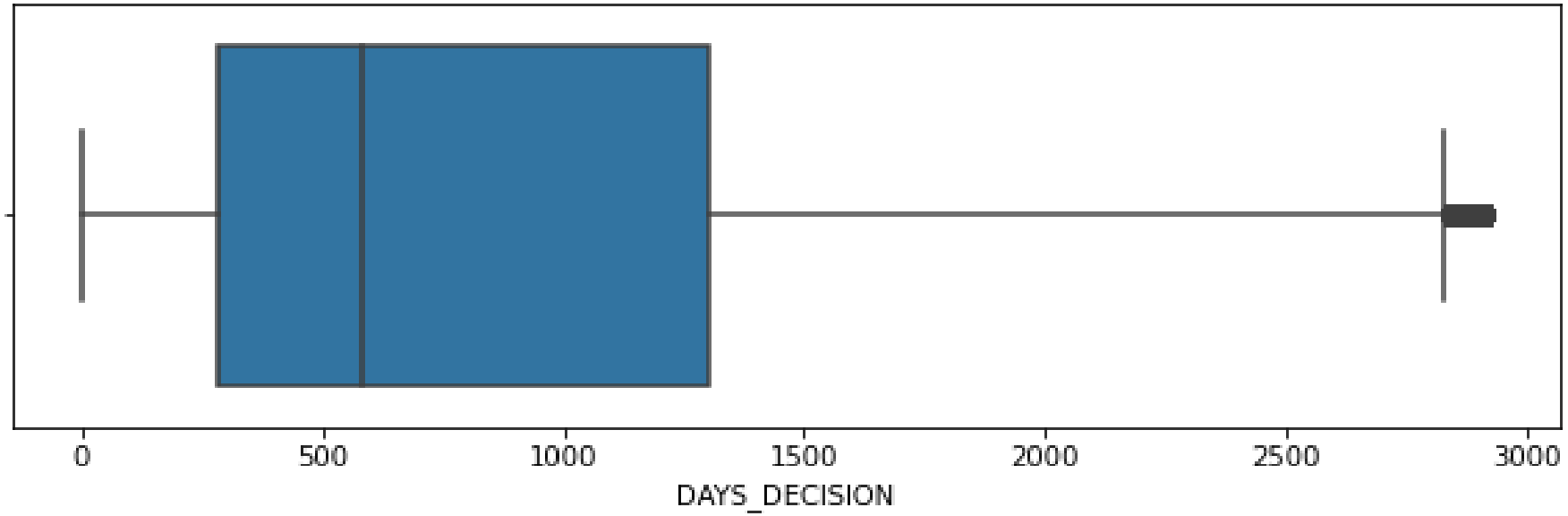
Variables AMT_APPLICATION and AMT_GOODS_PRICE have outliers in it.

Outlier analysis-application dataset



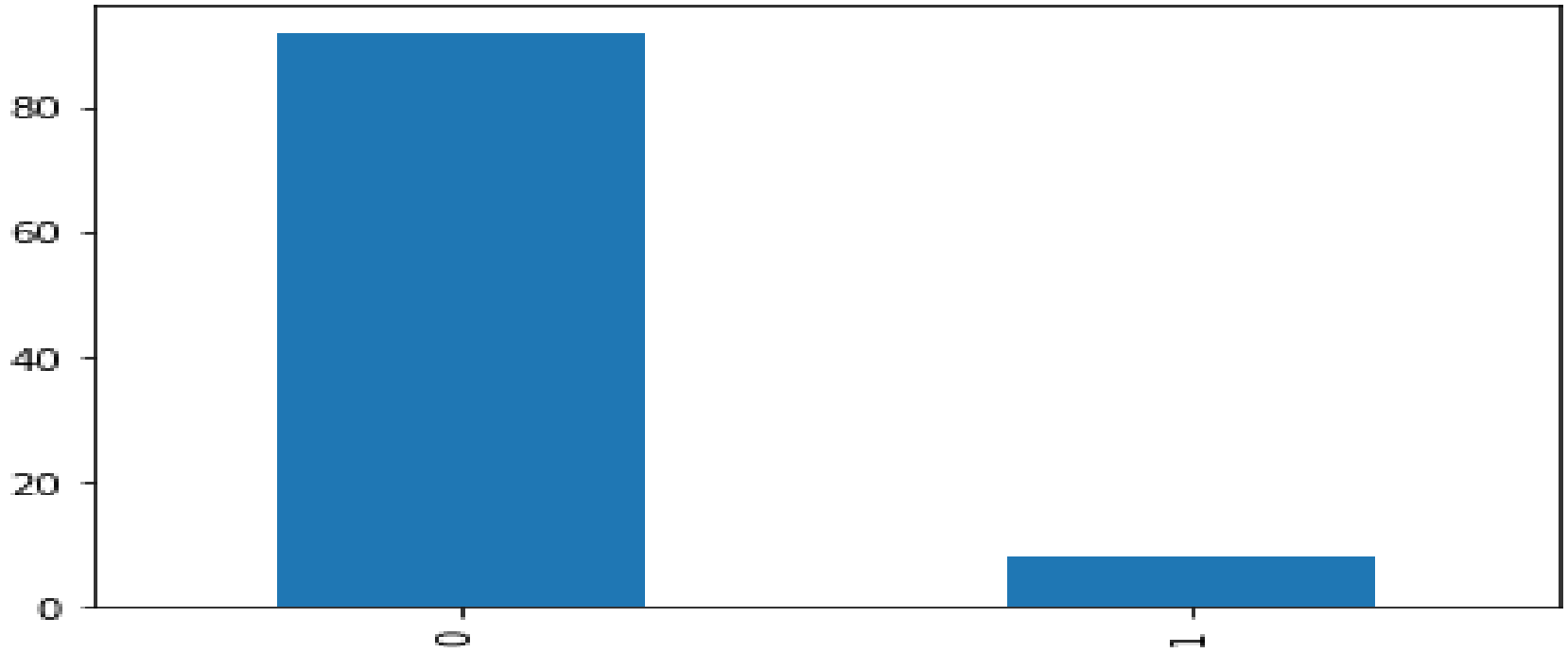
Variables DAYS_LAST_DUE and DAYS_TERMINATION have outliers in it.

Outlier analysis-application dataset



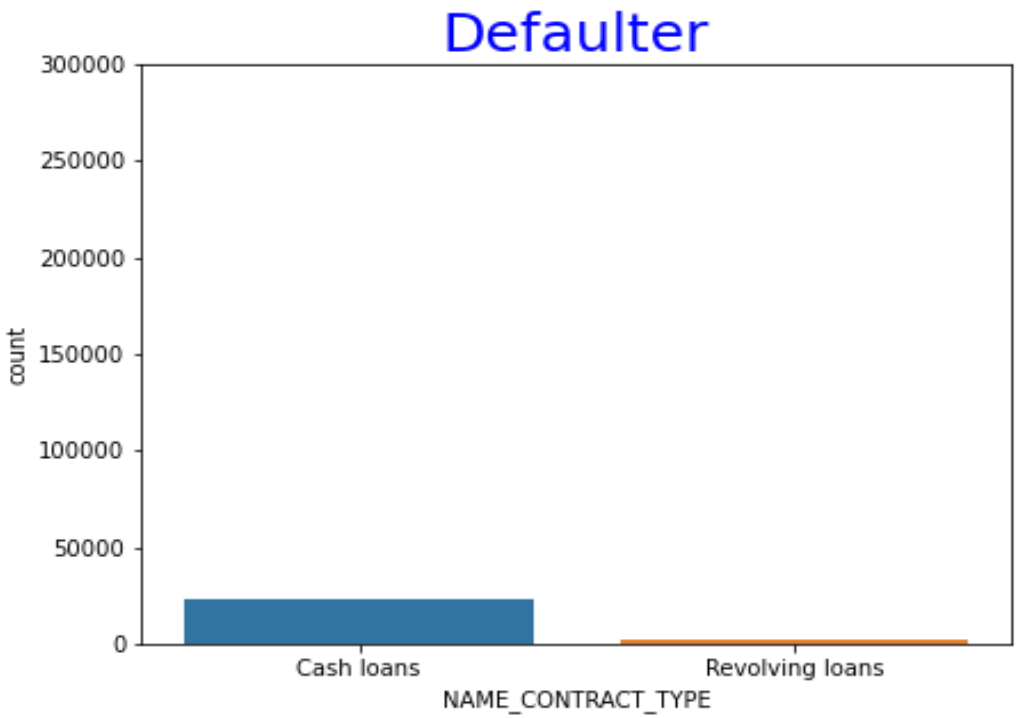
Variables DAYS_DECISION don't have any outlier in it.

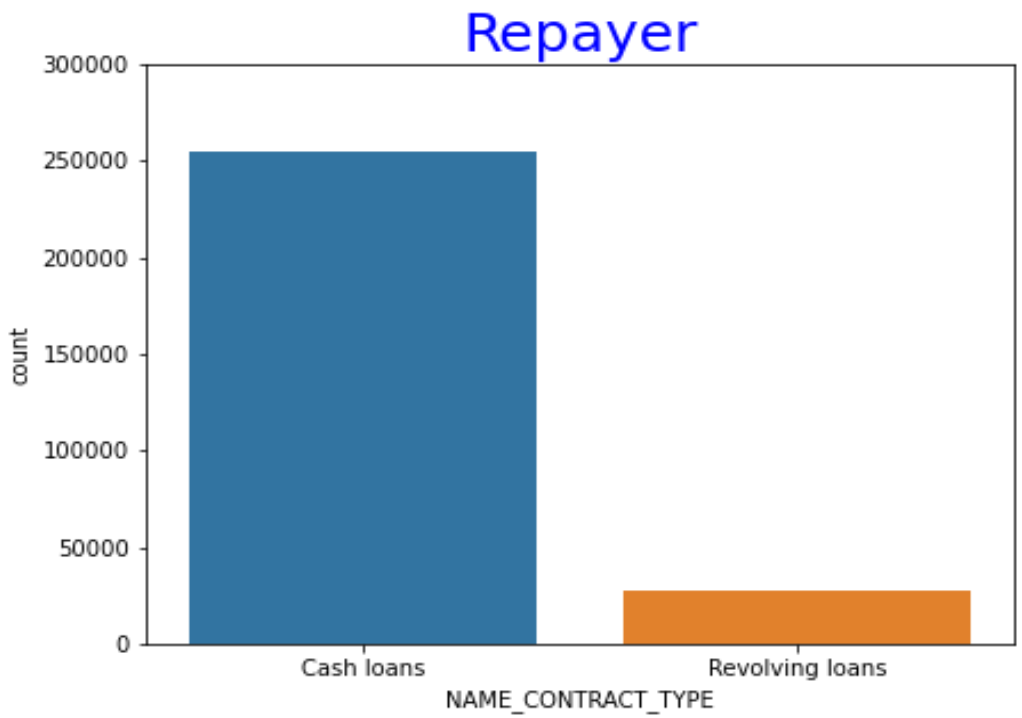
Data imbalance-TARGET VARIABLE



Target Variable has severe Data Balance. Approx. 91% data is with Target Value 0 and 9% is with Target Value 1. This Sort of Data Imbalance may impact our analysis. Hence we have subset our original data frame based on target variable value

UNIVARIATE ANALYSIS NAME_CONTRACT_TYPE

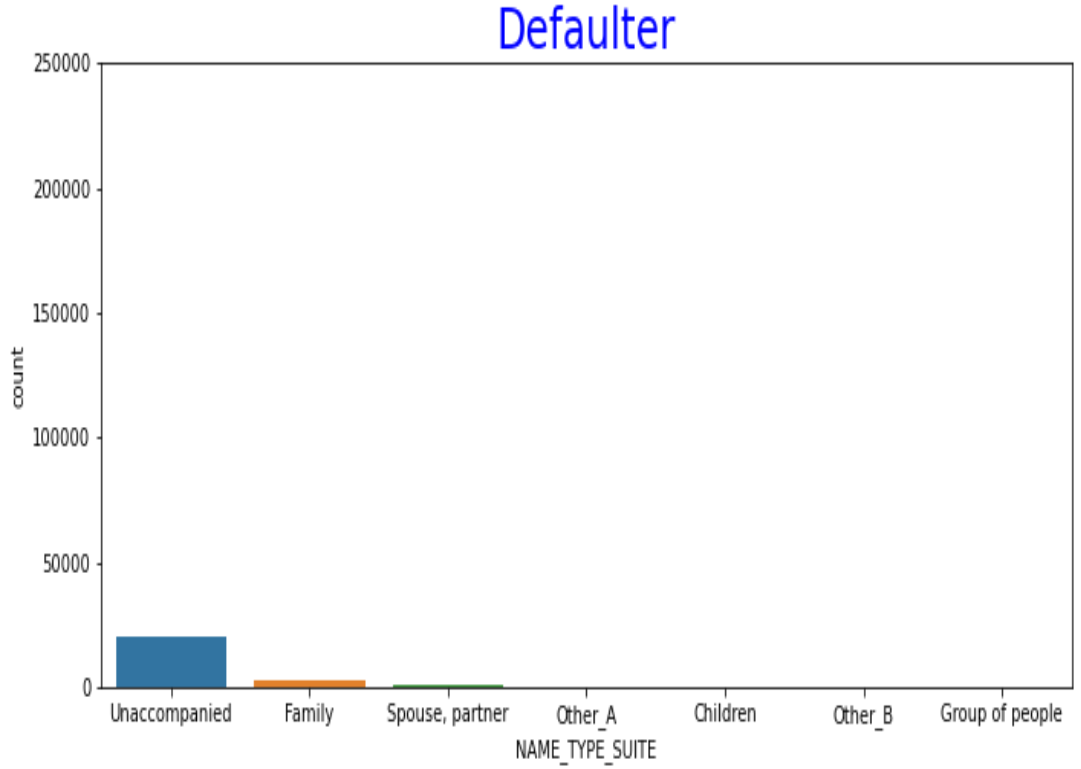


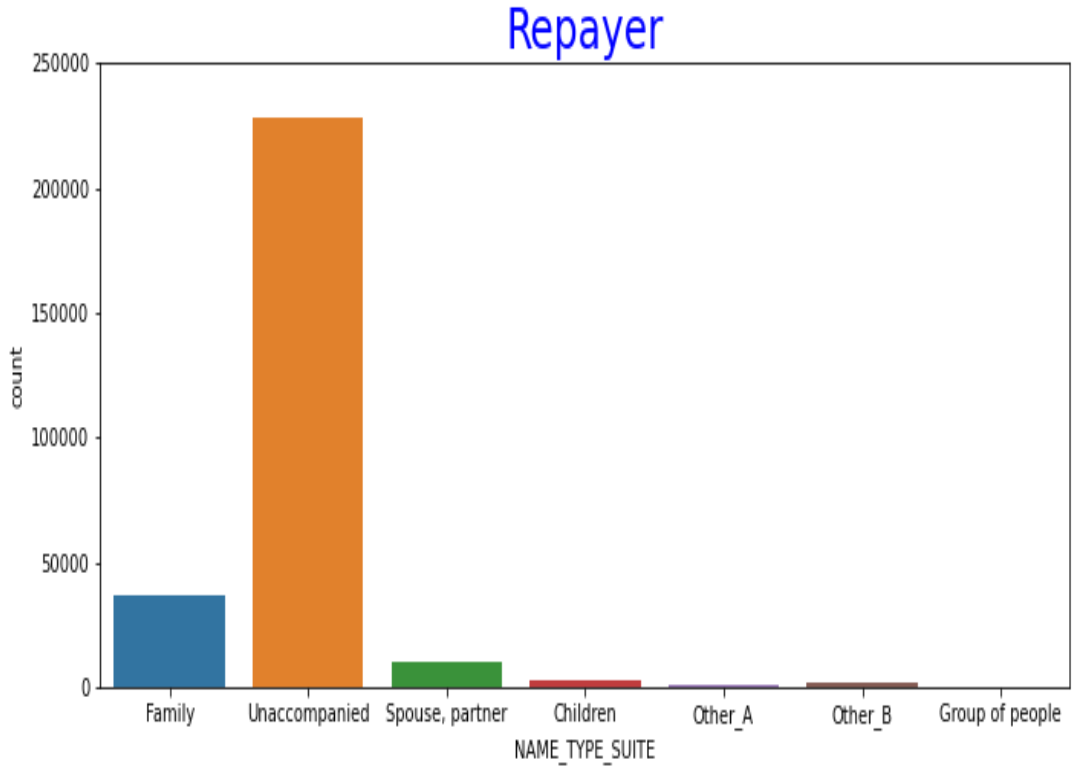


Insights:-

1. Cash Loans are more preferred compared to revolving Loans
2. The proportion of revolving loans is very less for the persons who has defaulted the loans.

UNIVARIATE ANALYSIS-NAME_CONTRACT_TYPE

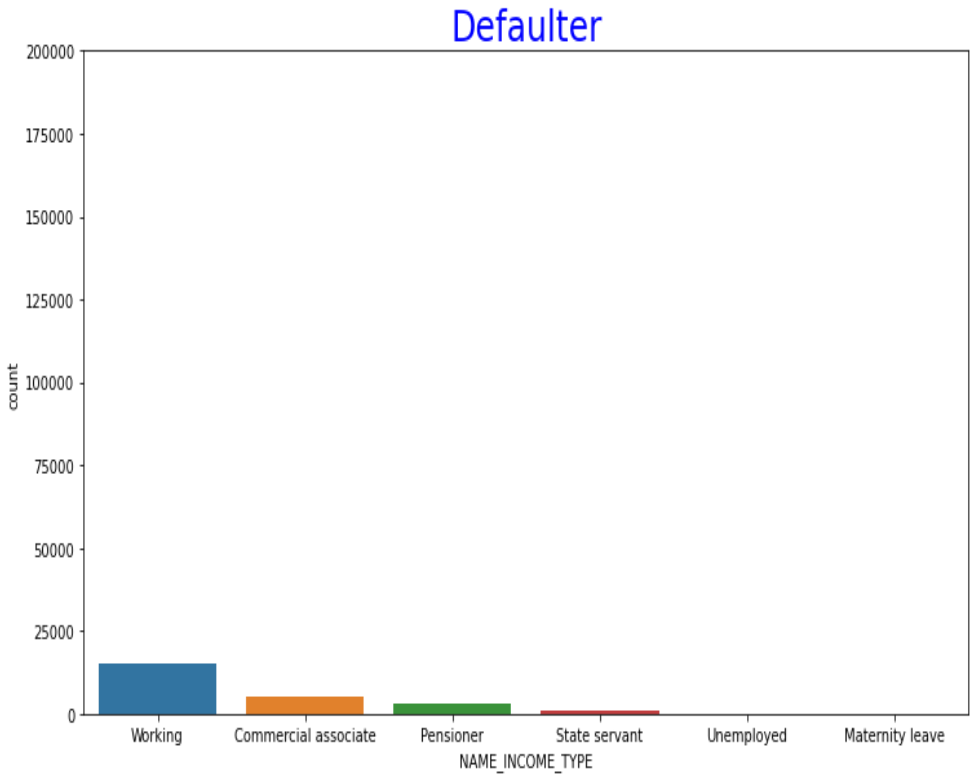


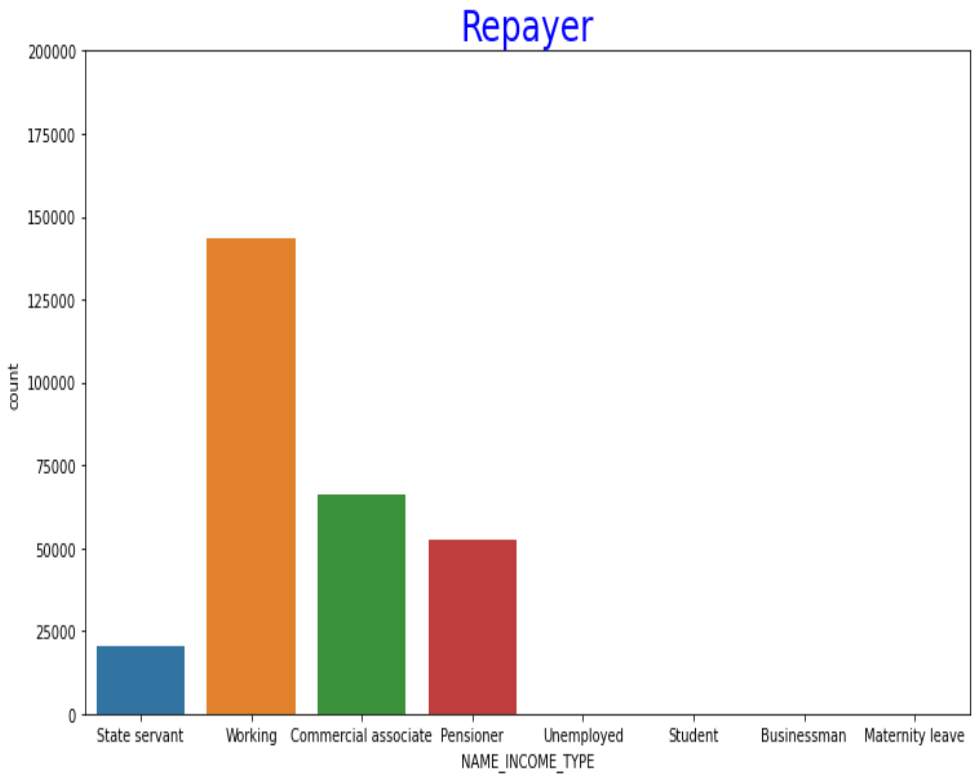


Insights:-

1. Most of the People applying loans are not carrying anybody as the companion
2. Some are carrying their Family for applying the loans
3. People accompanied by somebody have very high chances of repaying the loans

UNIVARIATE ANALYSIS-NAME_INCOME_TYPE

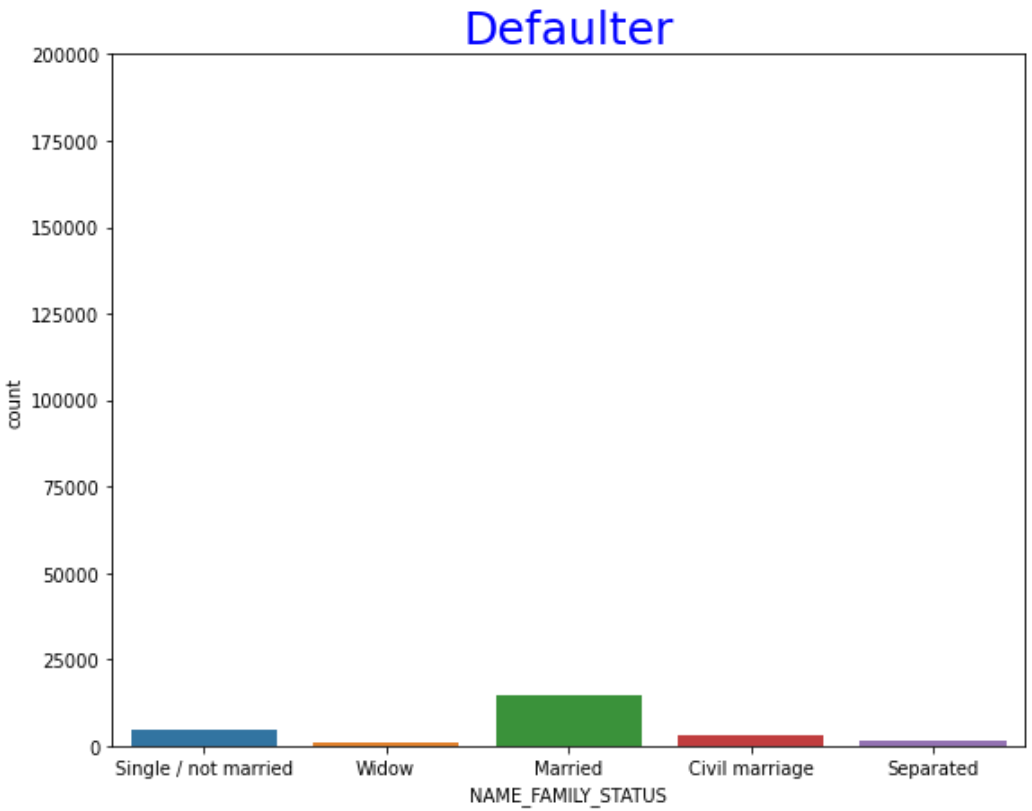


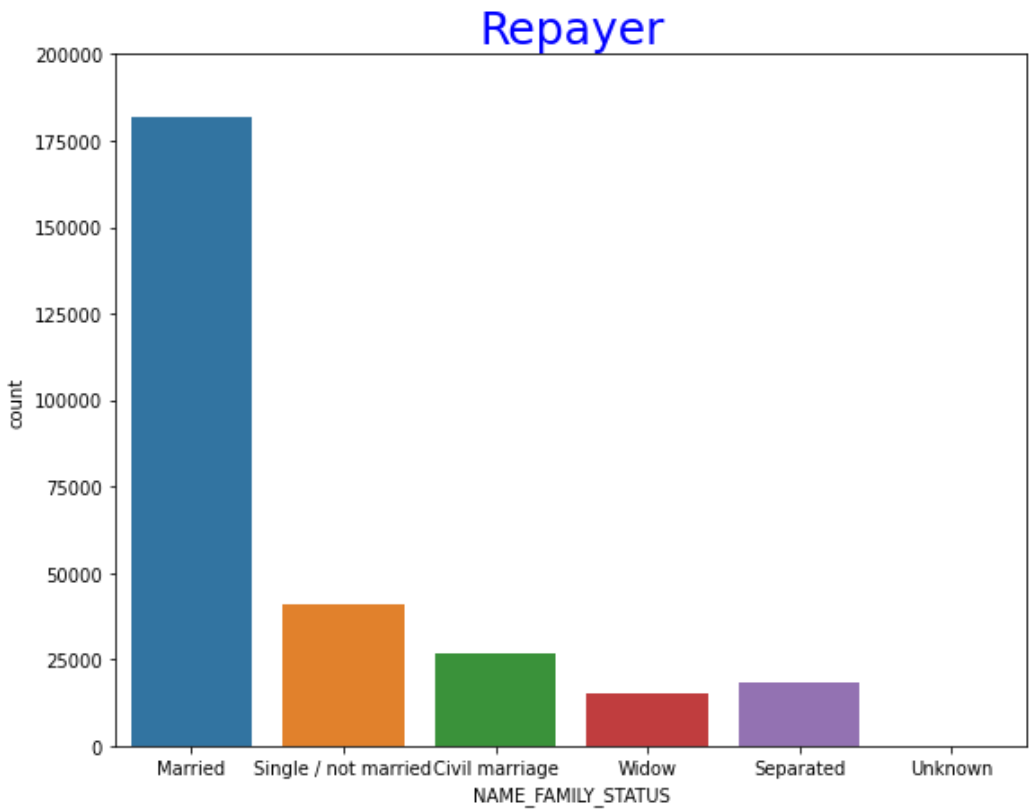


Insights:-

1. Businessman are never defaulting any loans
2. Working People are more likely to default the loans

UNIVARIATE ANALYSIS-NAME_FAMILY_STATUS

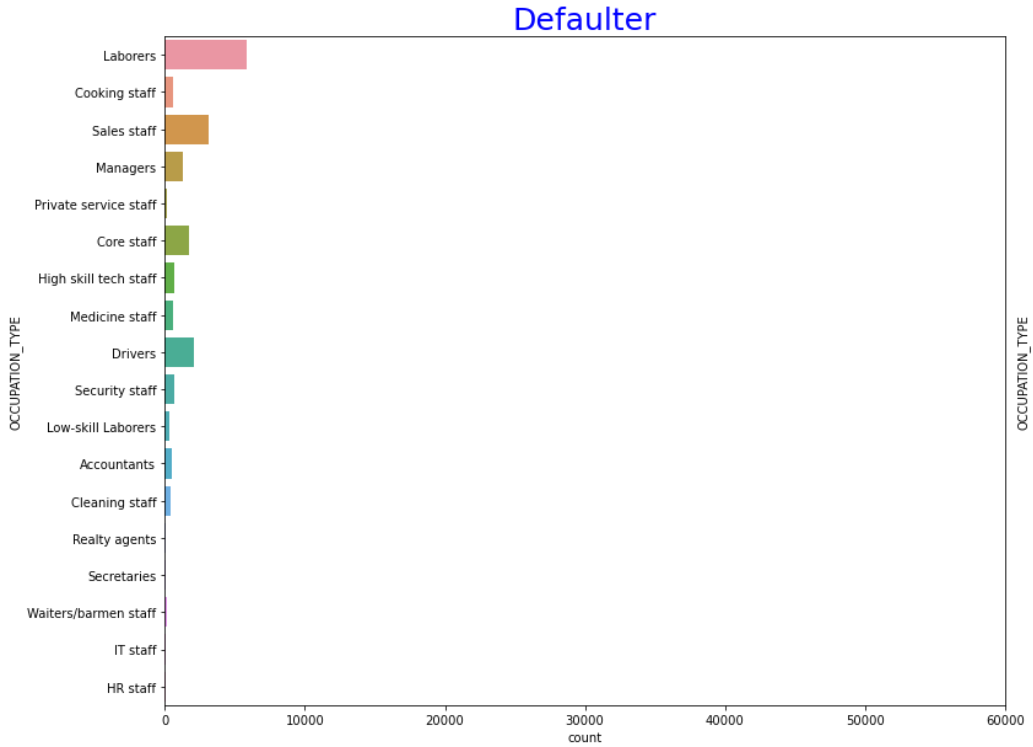


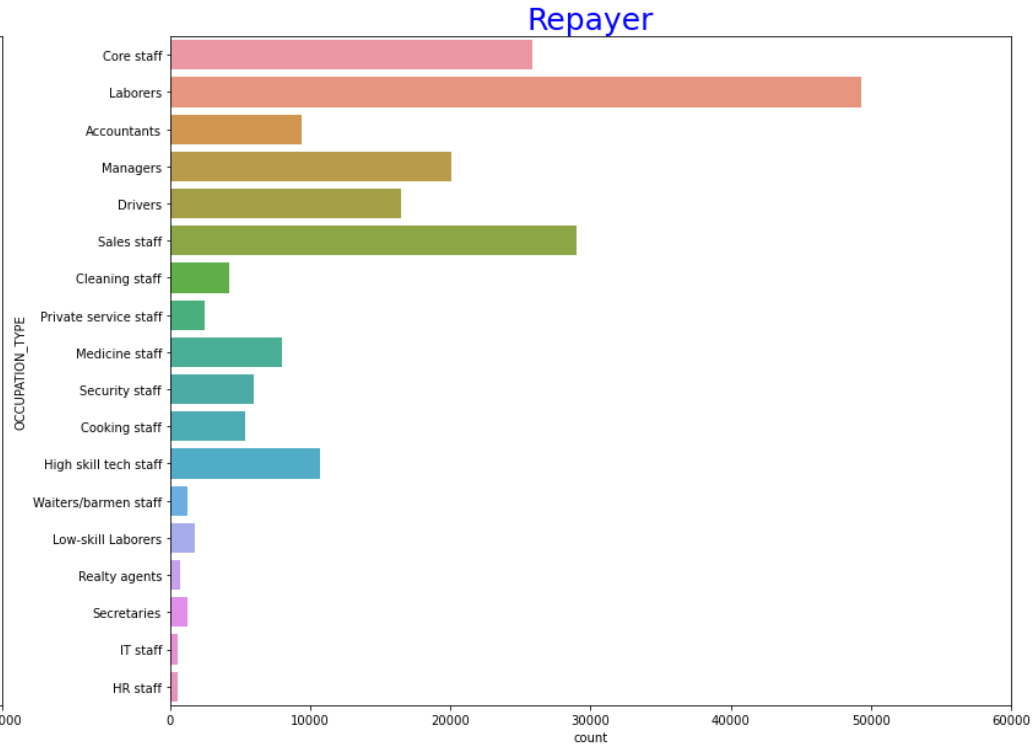


Insights:-

1. Widow, Separated are very least likely to default on loans
2. Married People are more likely to default and repay both
3. Single/Not married category is least likely to default however more likely to repay

UNIVARIATE ANALYSIS-OCCUPATION_TYPE

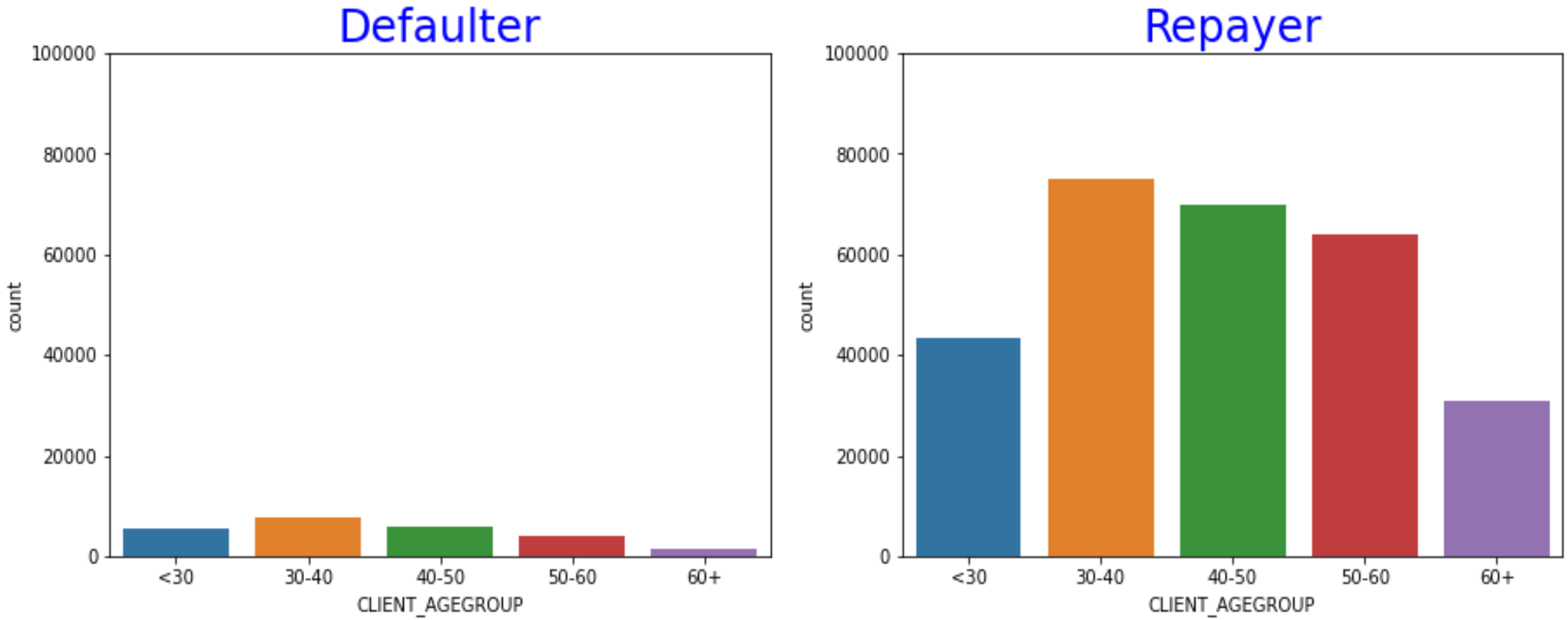




Insights:-

1. Laborers and Sales Staff are equally likely to default and repay
2. Realty agents, Secretaries, Walters/barmen staff, IT staff, HR Staff have very less chance of defaulting the loan

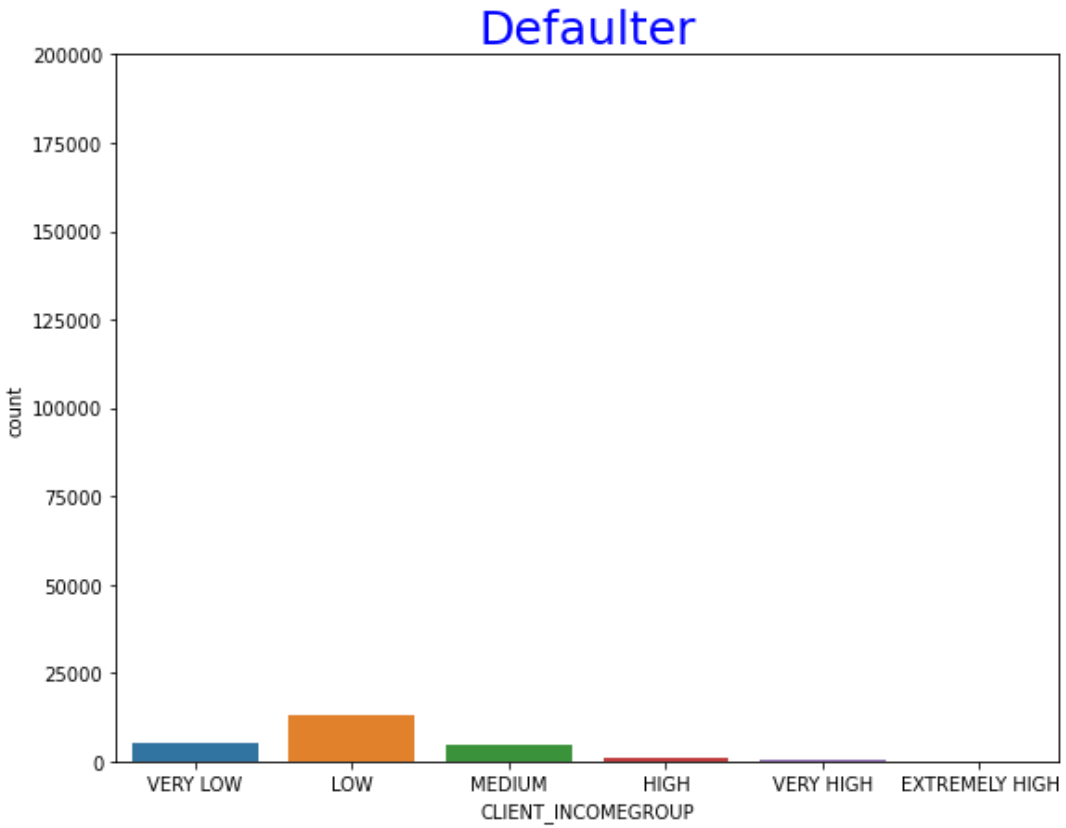
UNIVARIATE ANALYSIS-CLIENT_AGEGROUP

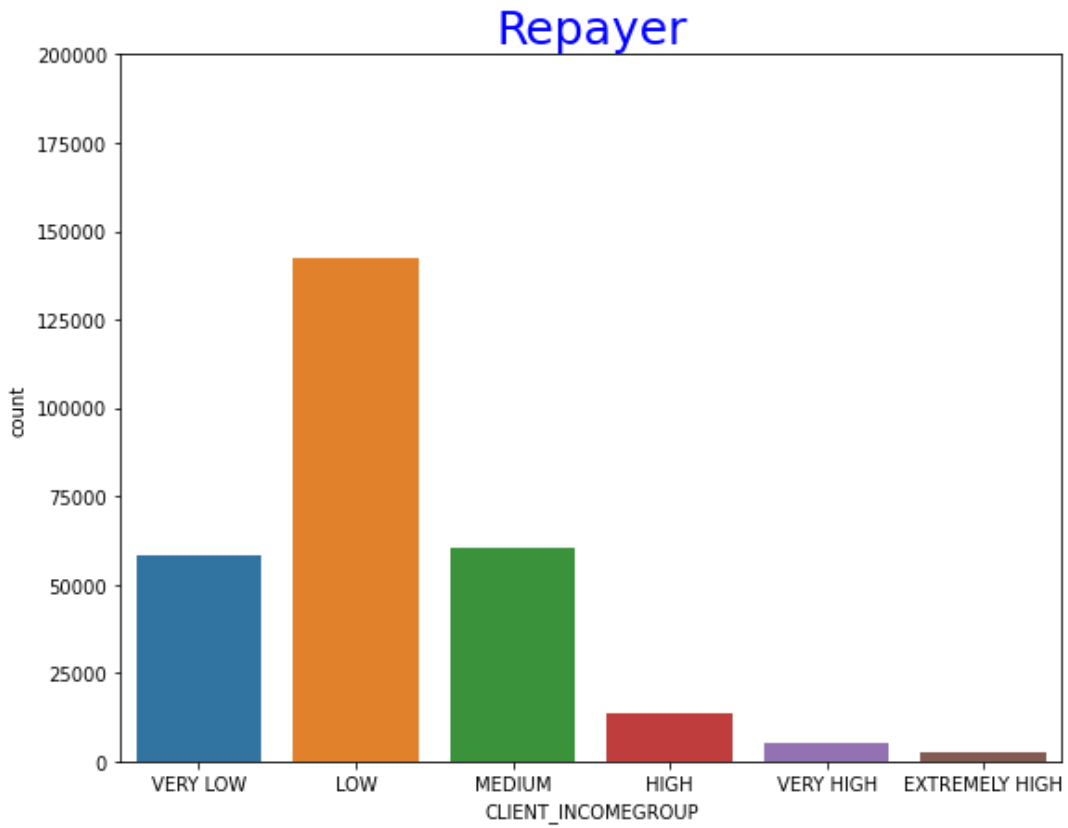


Insights:-

1. Loans are applied mostly be people in age range of 30-40,40-50
2. People with age 60+ are least likely to take loans. Moreover they have least chances of defaulting the loans
3. People with older age 60+ primarily senior citizens are less likely to take loans compared to other age groups
4. People in age group 30-40 and most likely to repay the loan

UNIVARIATE ANALYSIS- CLIENT_INCOMEGROUP

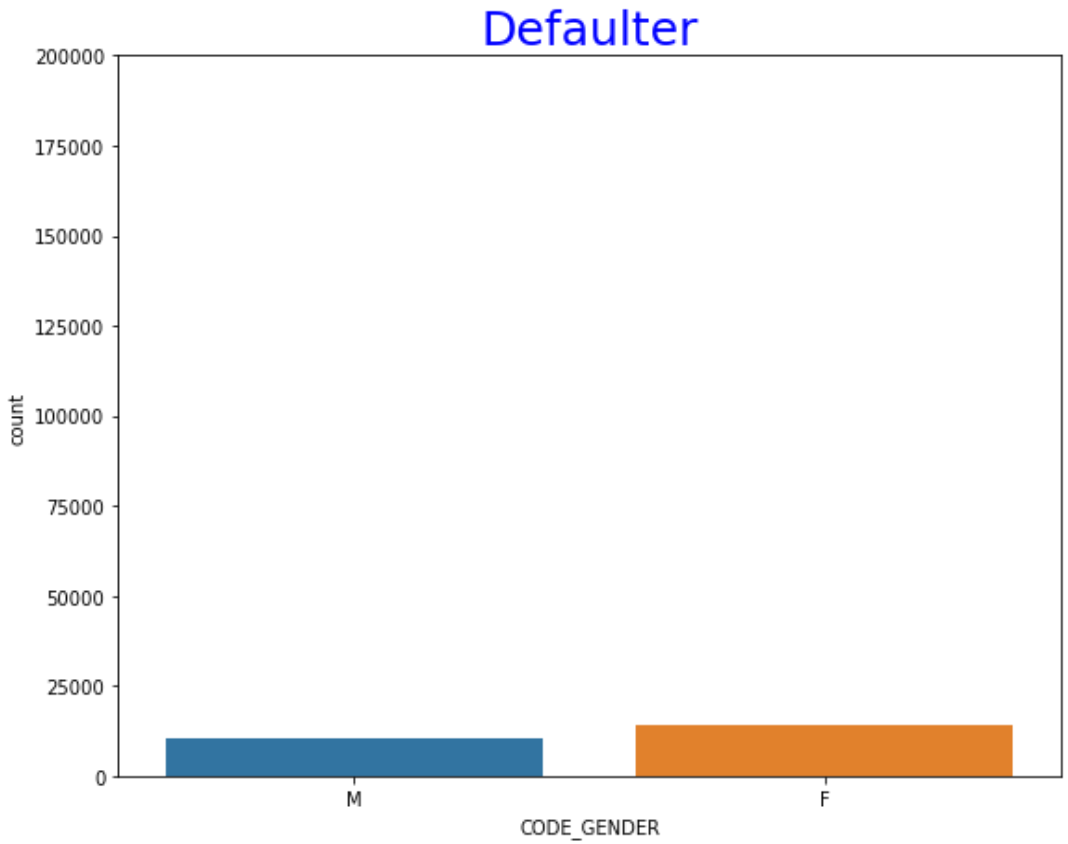


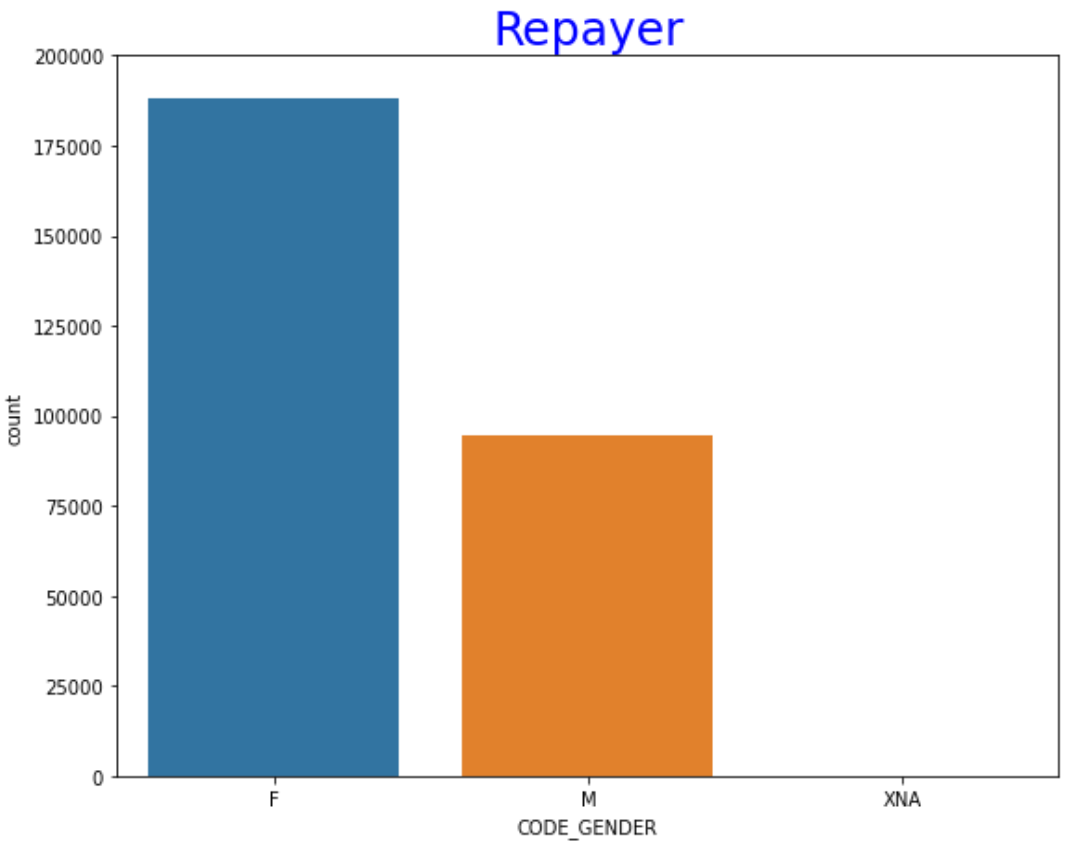


Insights:-

1. Loans are applied mostly be people in low income range(1 Lac to 2 Lac)
2. People with Low Income (1 Lac to 2 Lac) have more chances to defaults on loans
3. People with Very High (4 Lac to 5 Lac) and Extremely High (5 Lac+) are very less chance to become defaulter

UNIVARIATE ANALYSIS- CODE_GENDER

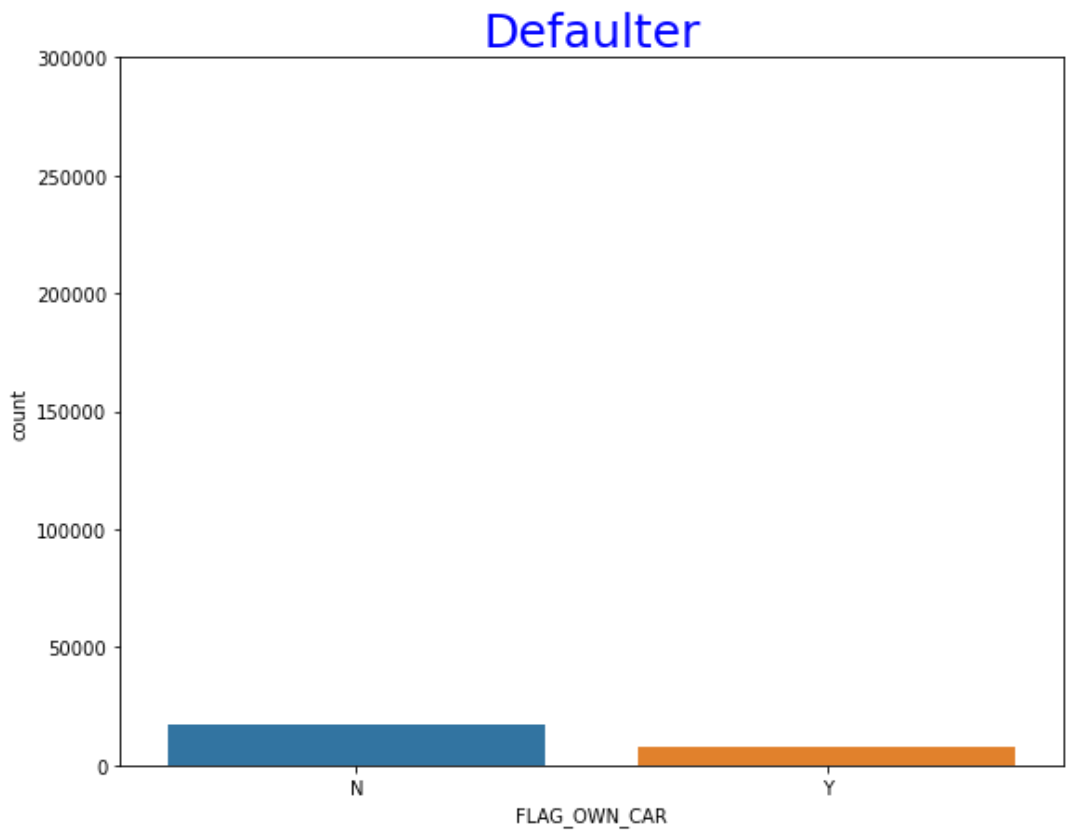


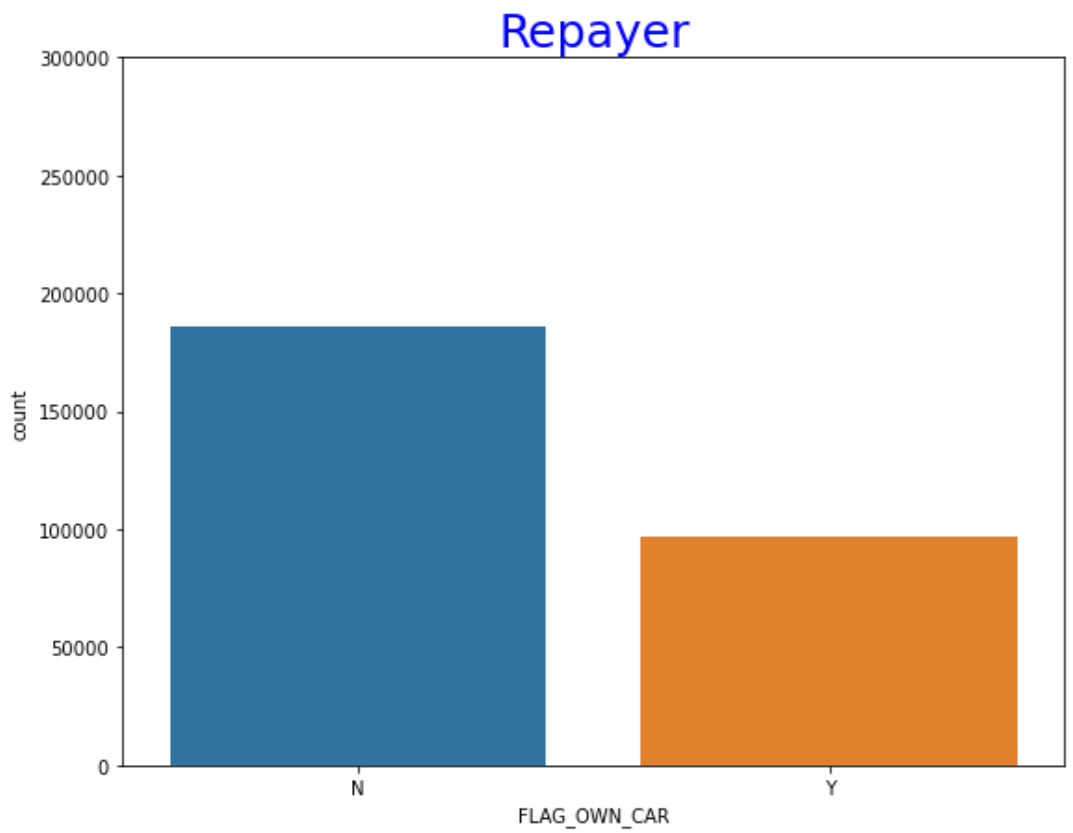


Insights:-

1. Females are most likely to repay the loan as well.
2. Females Re payers are approx. double of male Re payers

UNIVARIATE ANALYSIS- FLAG_OWN_CAR

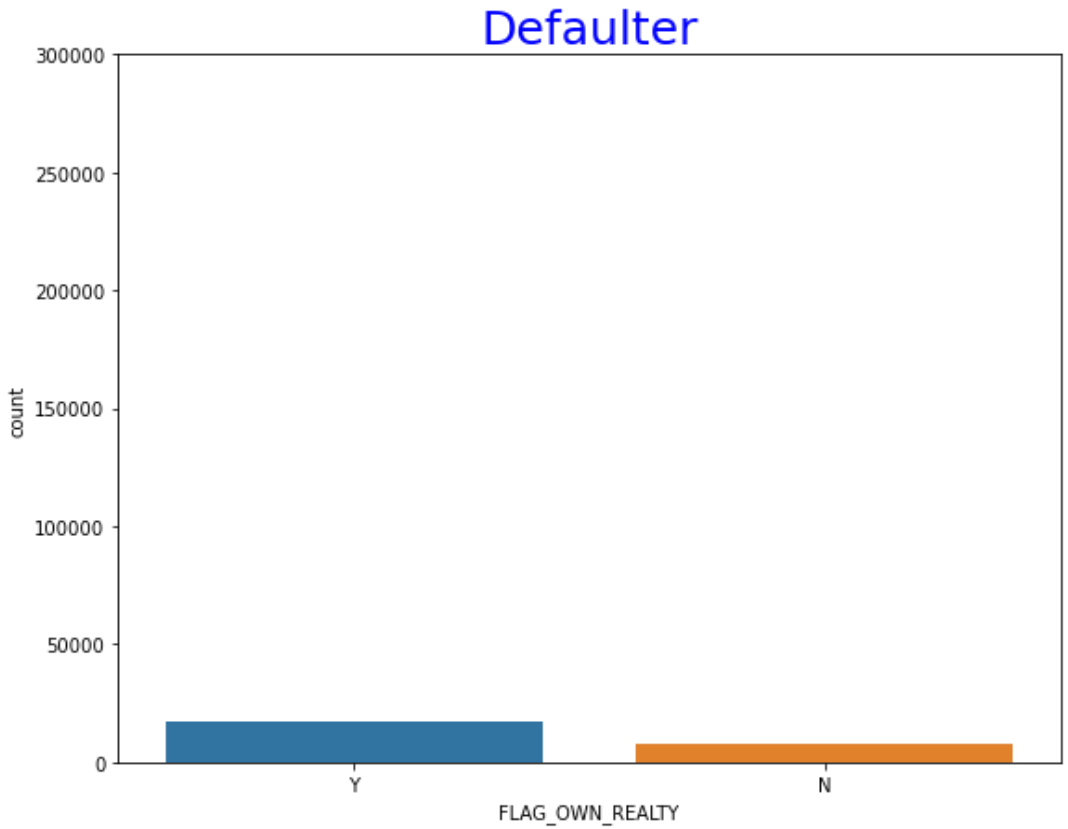


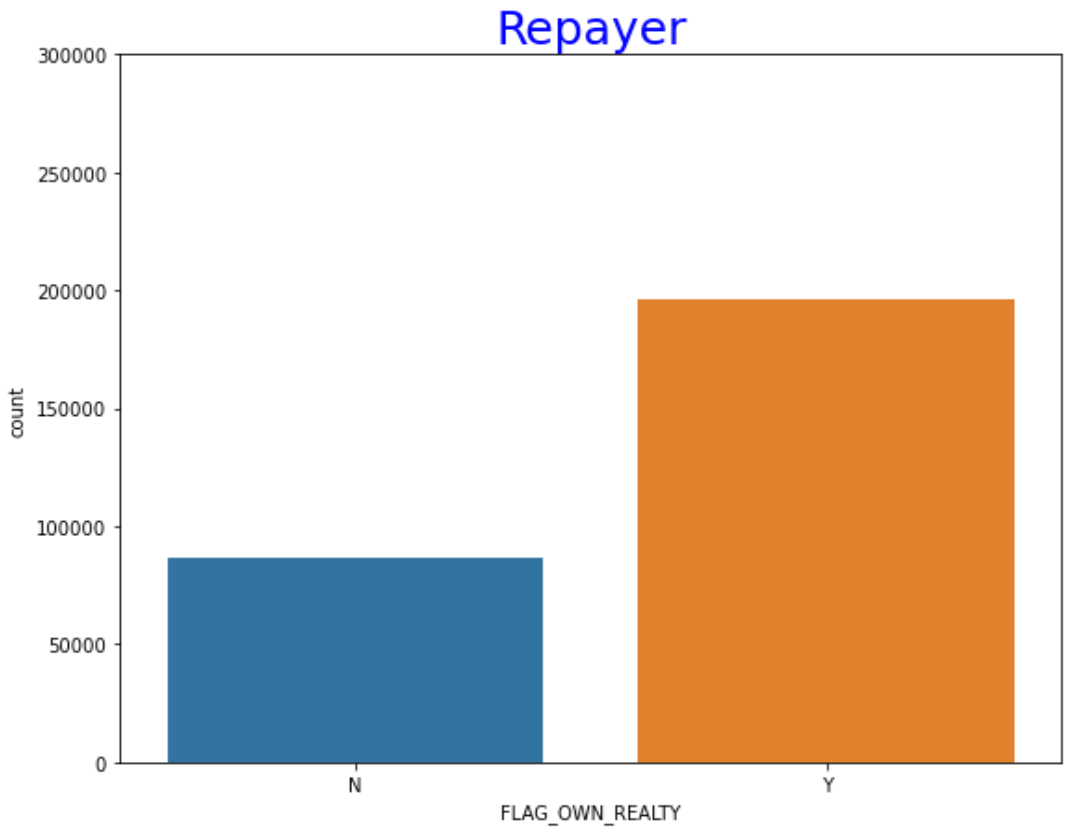


Insights:-

1. People without car have more chances to default the loans

UNIVARIATE ANALYSIS- FLAG_OWN_REALTY

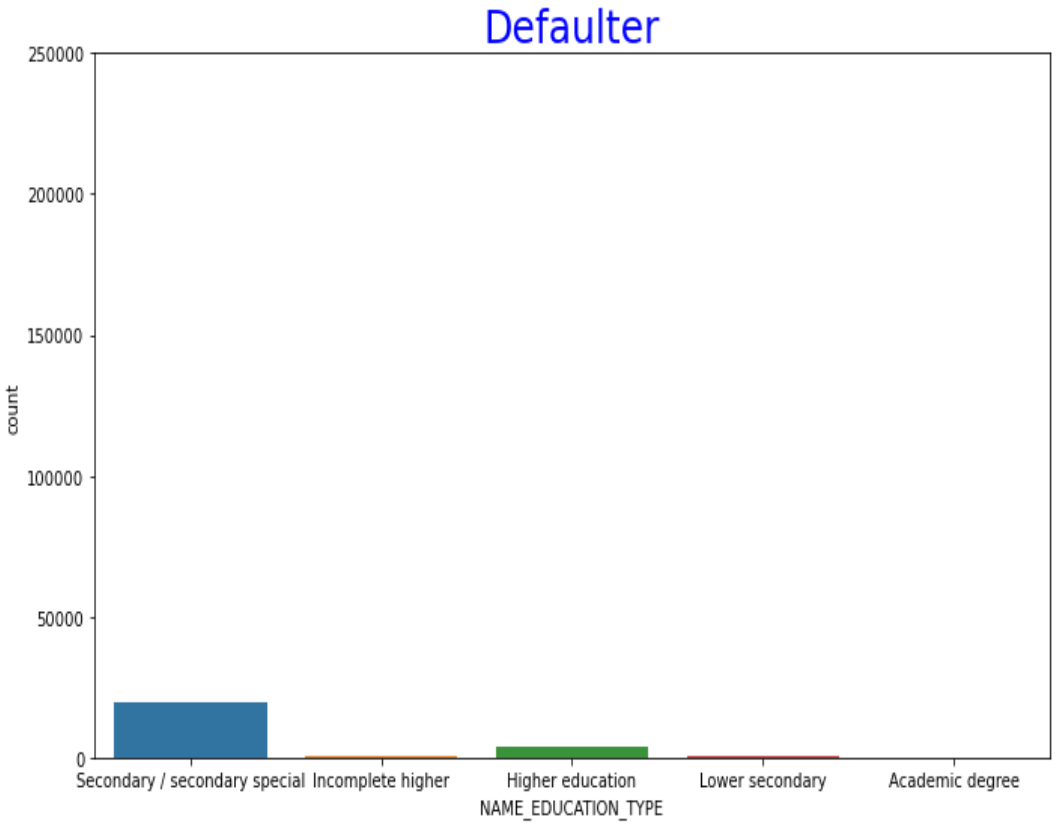


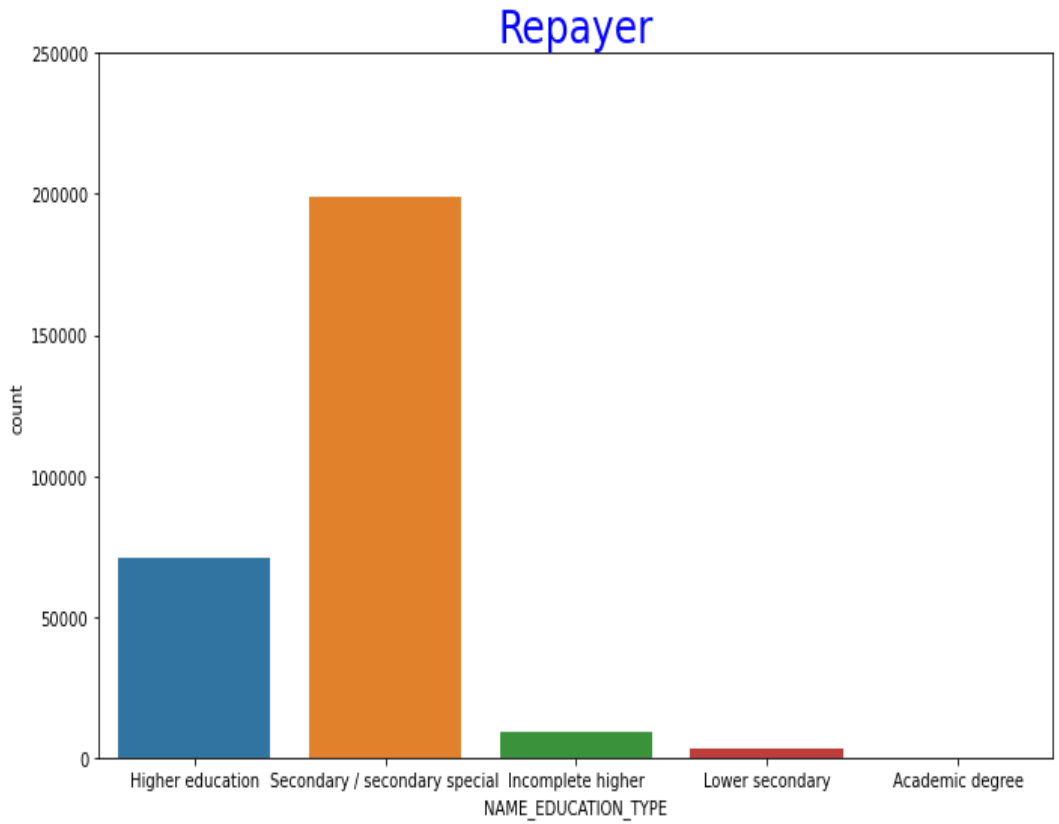


Insights:-

1. People who owns Flats have double the chances of repaying the loan

UNIVARIATE ANALYSIS- NAME_EDUCATION_TYPE

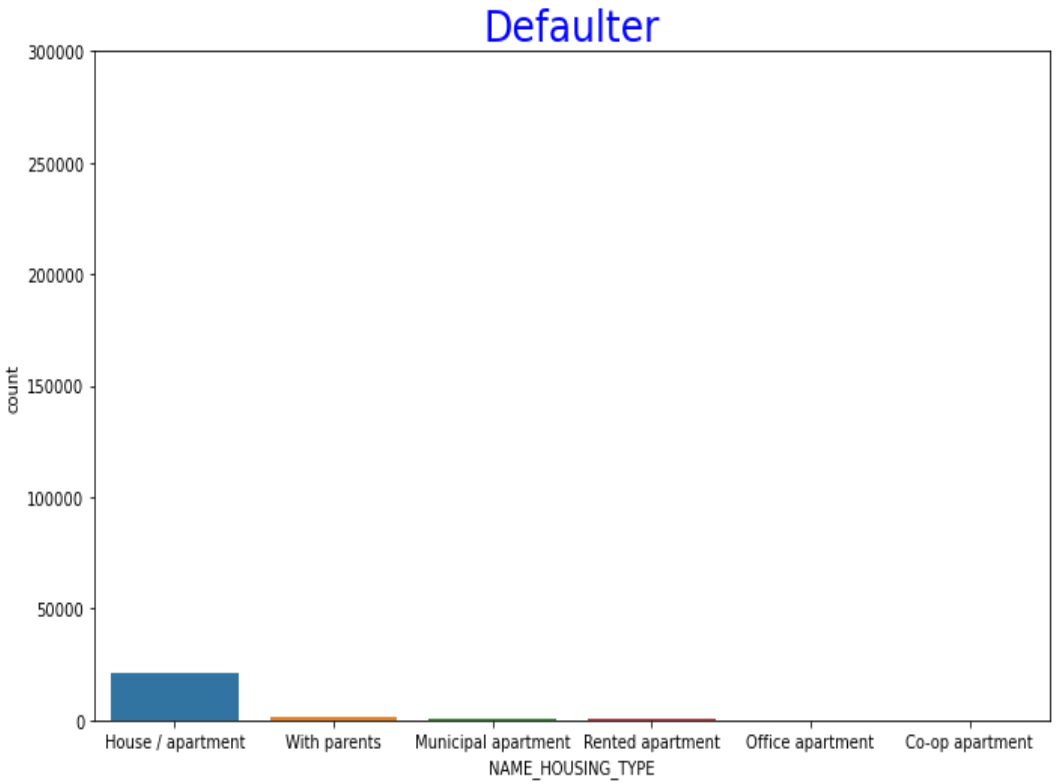


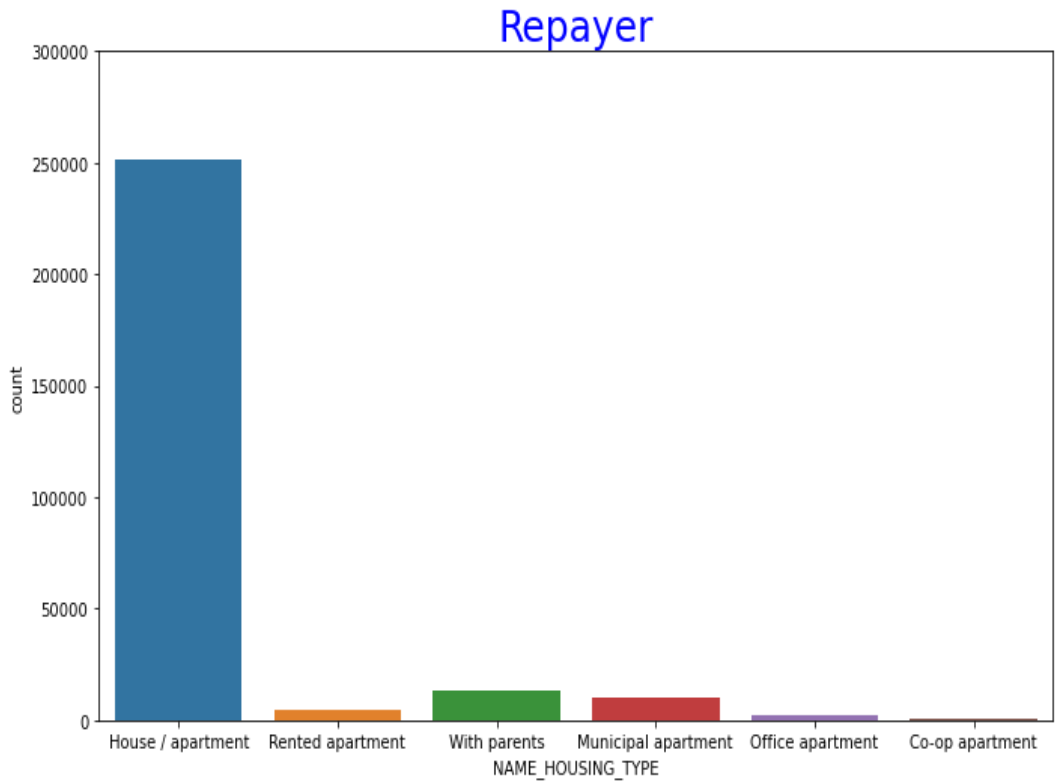


Insights:-

1. People with Secondary Education is more likely to repay the loan

UNIVARIATE ANALYSIS- NAME_HOUSING_TYPE

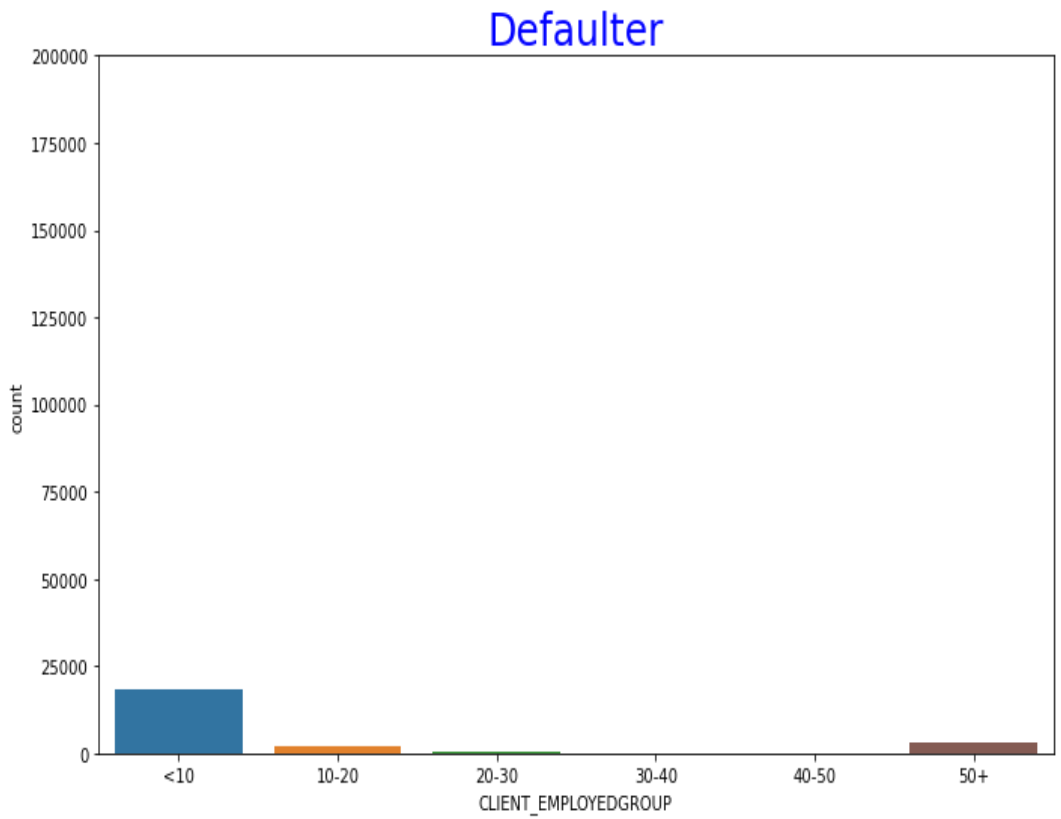


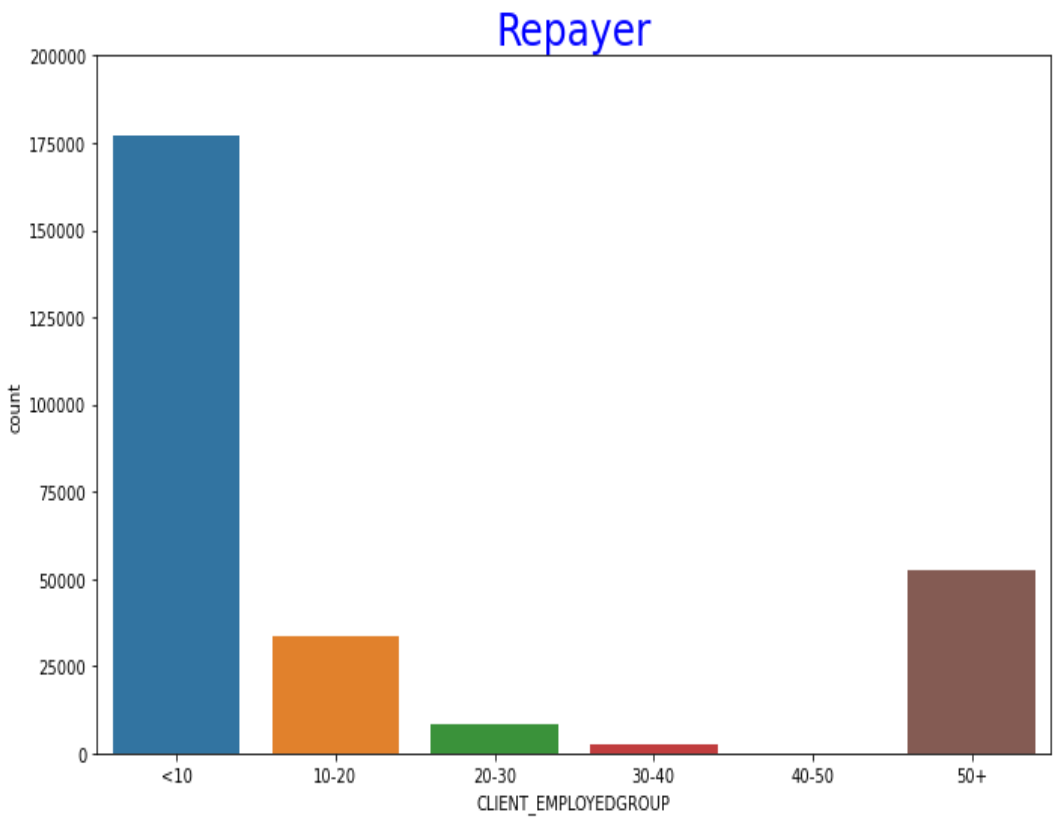


Insights:-

1. People with House/Apartment is more likely to repay the loan

UNIVARIATE ANALYSIS- CLIENT_EMPLOYEDGROUP

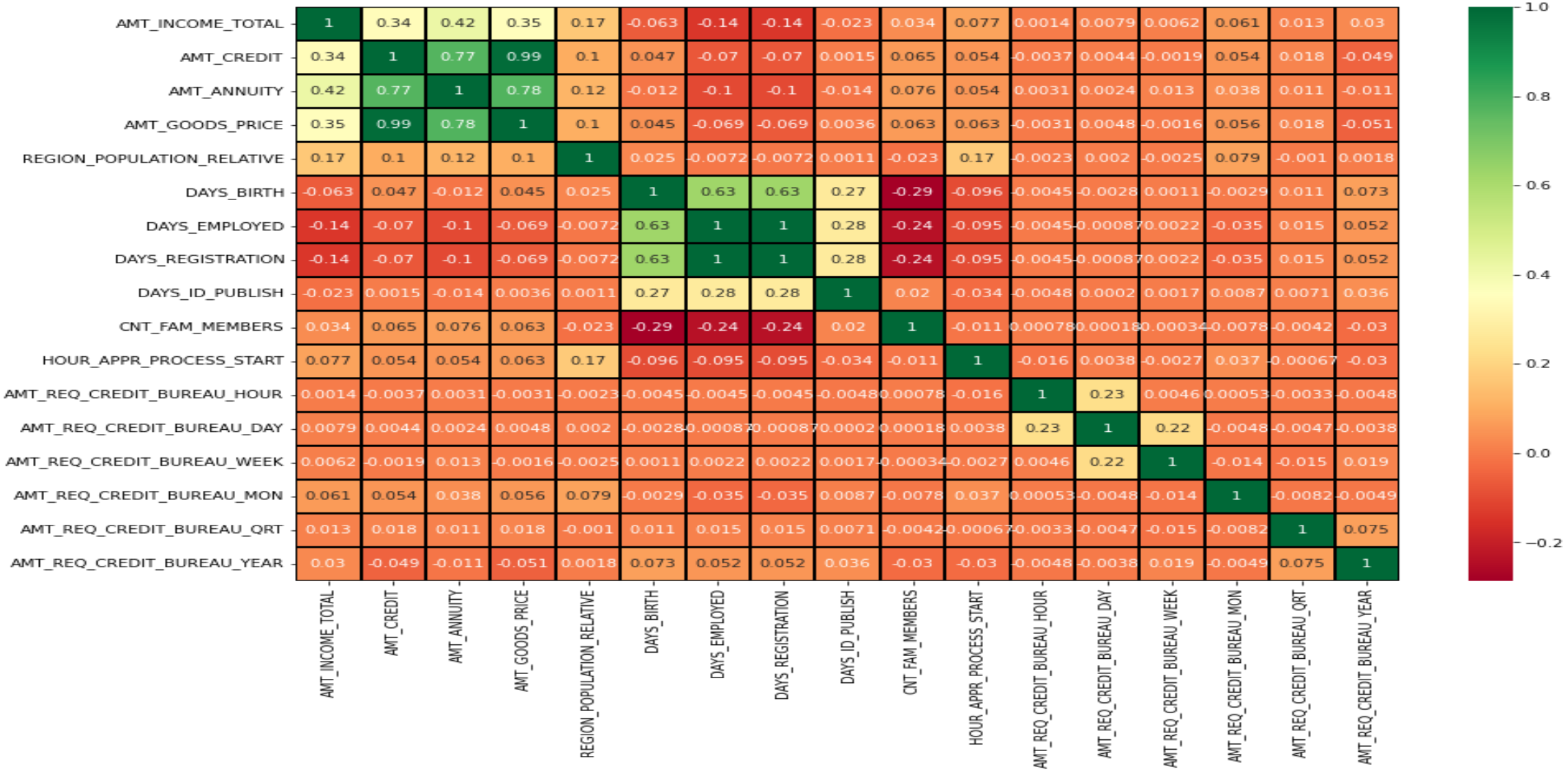




Insights:-

1. People who are employed with more than 50+Years are likely to default less however most likely to repay.
2. People who are employed with less than 10 years are most likely to default compared to other time periods

CORRELATION ANALYSIS- REPAYER

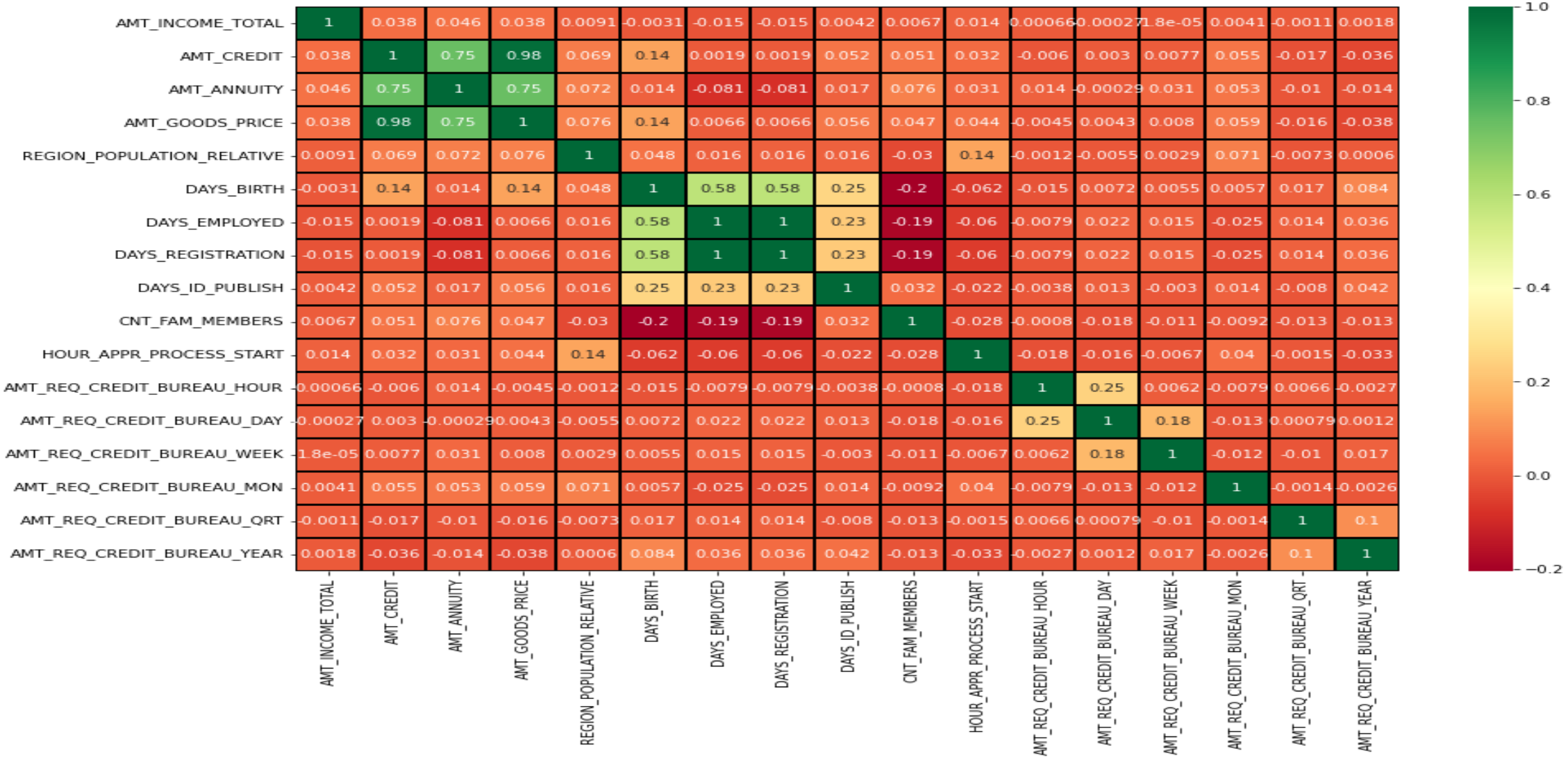


CORRELATION ANALYSIS- REPAYER

- Top Correlations are
  - AMT_CREDIT VS AMT_GOODS_PRICE
  - AMT_ANNUITY VS AMT_GOODS_PRICE
  - AMT_CREDIT VS AMT_ANNUITY

- Other Strong Correlations are as follows:
  - DAYS_BIRTH VS DAYS_EMPLOYED
  - DAYS_BIRTH VS DAYS_REGISTRATION
  - AMT_ANNUITY VS AMT_INCOME_TOTAL
  - AMT_INCOME_TOTAL VS AMT_GOODS_PRICE
  - AMT_CREDIT VS AMT_INCOME_TOTAL
  - DAYS_ID_PUBLISH VS DAYS_EMPLOYED
  - DAYS_ID_PUBLISH VS DAYS_REGISTRATION
  - DAYS_ID_PUBLISH VS DAYS_BIRTH
  - AMT_REQ_CREDIT_BUREAU_HOUR VS AMT_REQ_CREDIT_BUREAU_DAY
  - AMT_REQ_CREDIT_BUREAU_DAY VS AMT_REQ_CREDIT_BUREAU_WEEK

CORRELATION ANALYSIS- DEFAULTER



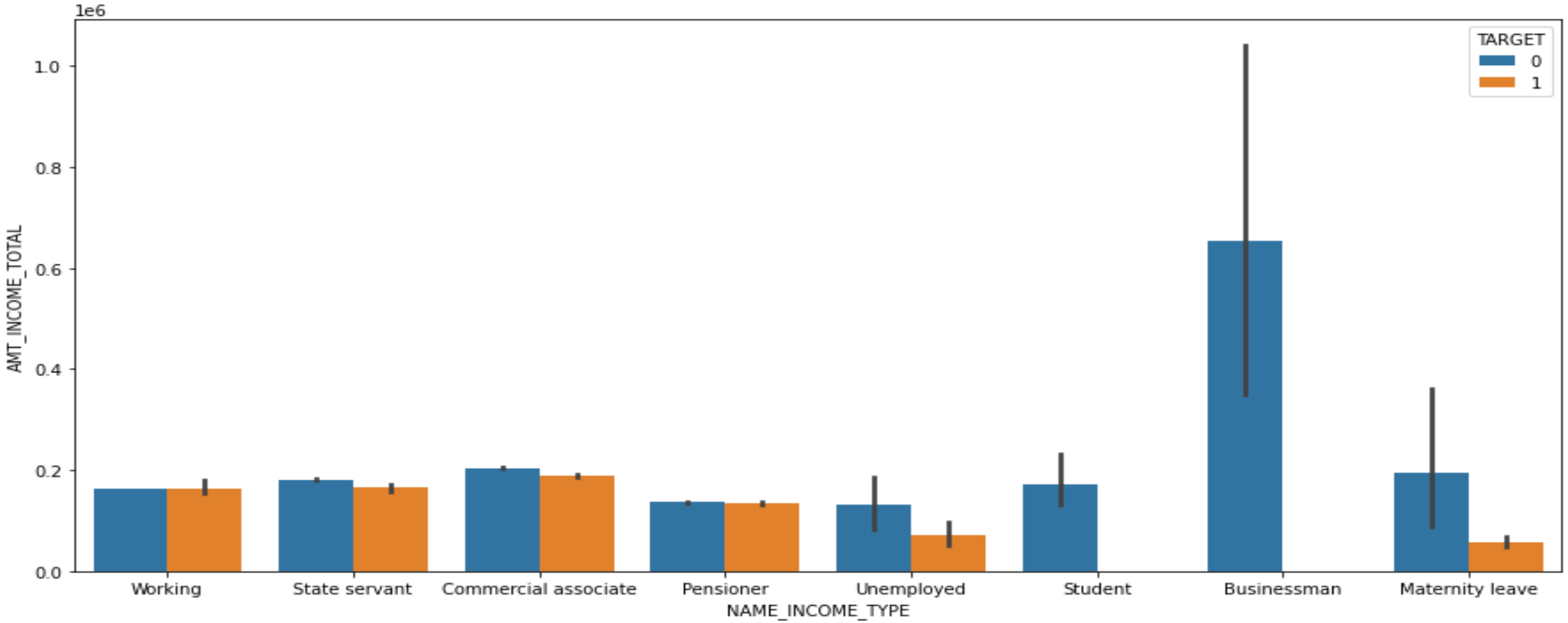
CORRELATION ANALYSIS- DEFAULTER

- Top Correlations are
  - AMT_CREDIT VS AMT_GOODS_PRICE
  - AMT_CREDIT VS AMT_ANNUITY
  - AMT_ANNUITY VS AMT_GOODS_PRICE

- Other Strong Correlations are as follows:
  - DAYS_BIRTH VS DAYS_EMPLOYED
  - DAYS_BIRTH VS DAYS_REGISTRATION
  - DAYS_BIRTH VS DAYS_ID_PUBLISH
  - AMT_REQ_CREDIT_BUREAU_HOUR VS AMT_REQ_CREDIT_BUREAU_DAY
  - DAYS_ID_PUBLISH VS DAYS_EMPLOYED
  - DAYS_ID_PUBLISH VS DAYS_REGISTRATION
  - AMT_REQ_CREDIT_BUREAU_DAY VS AMT_REQ_CREDIT_BUREAU_WEEK

BIVARIATE & MULTIVARIATE ANALYSIS

NAME_INCOME_TYPE VS AMT_INCOME_TOTAL

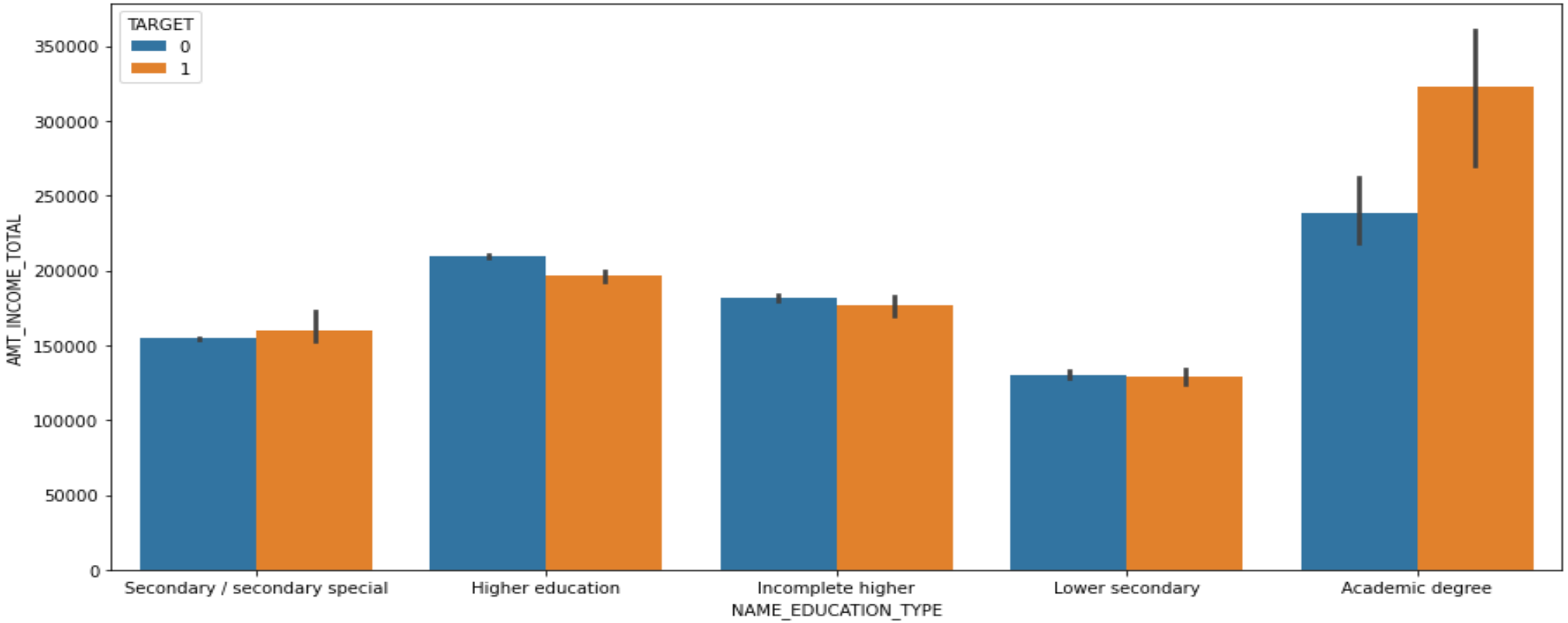


Insights:-

1. Businessman and students are never defaulting any loans
2. Working People and Pensioner have equal chances of defaulting and repaying the loans.
3. Women on Maternity leave have high chances of repaying the loan rather defaulting the loans

BIVARIATE & MULTIVARIATE ANALYSIS

NAME_EDUCATION_TYPE VS AMT_INCOME_TOTAL

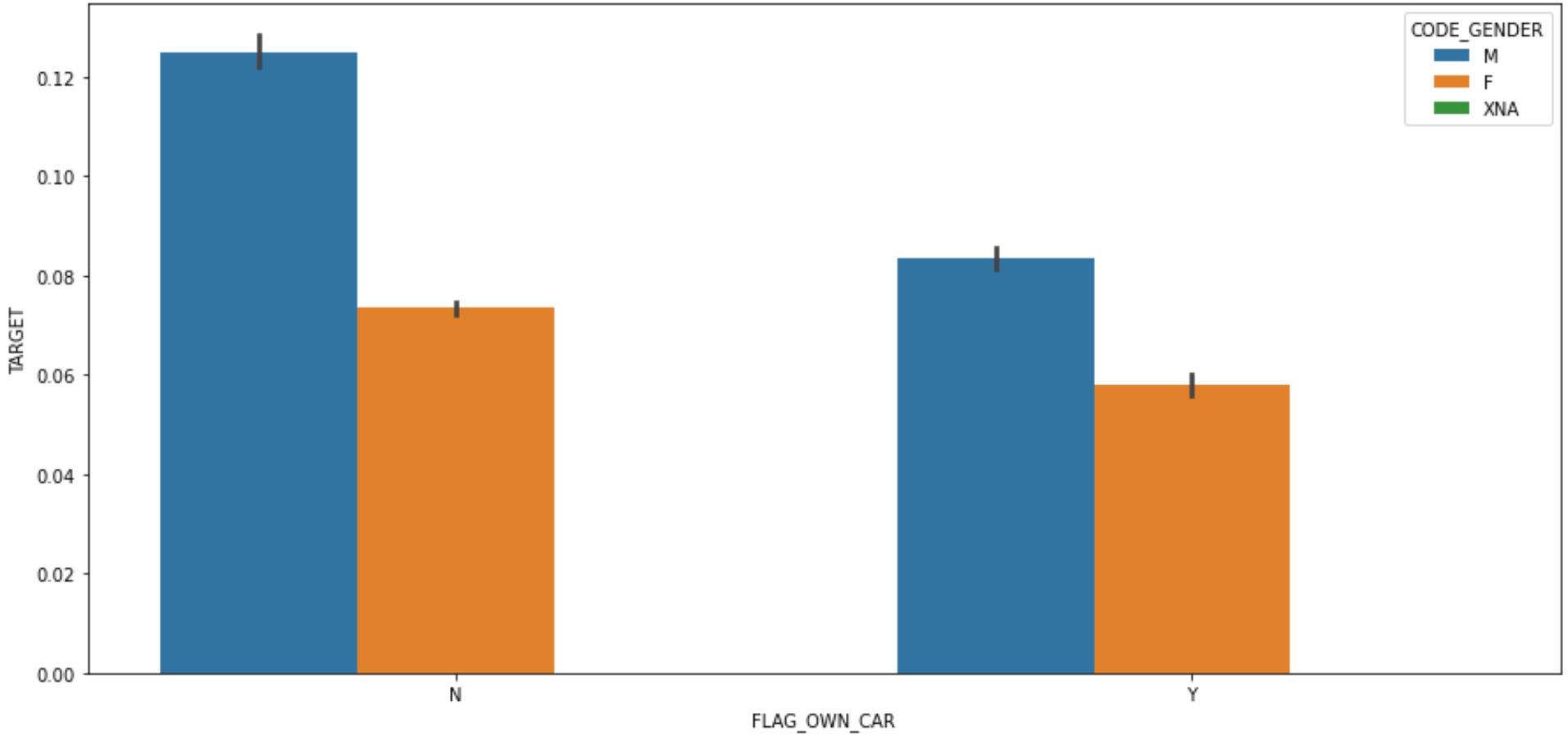


Insights:-

1. People with Academic Degree have more chances to default the loans

BIVARIATE & MULTIVARIATE ANALYSIS

FLAG_OWN_CAR VS CODE_GENDER

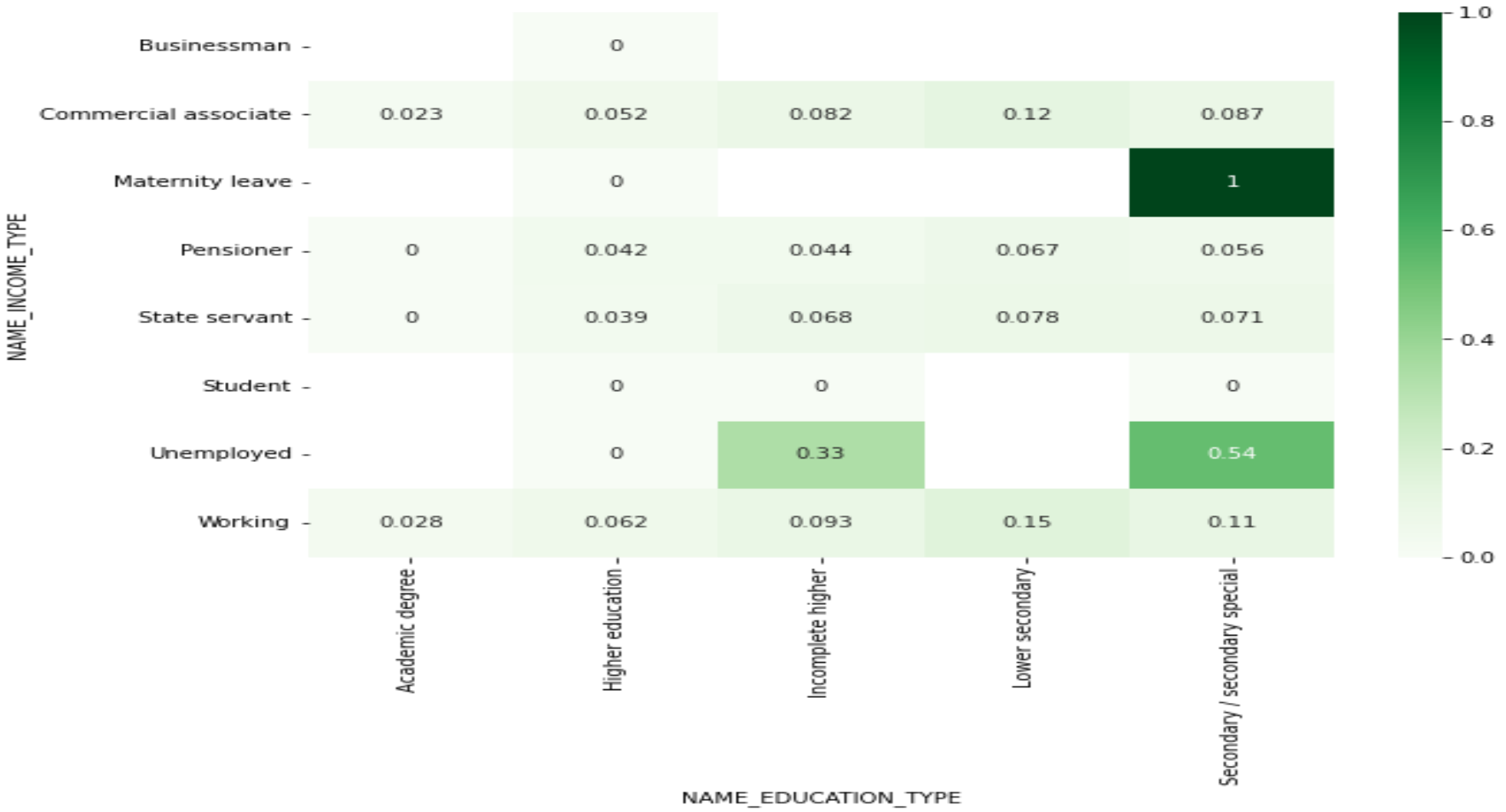


Insights:-

1. Males who do not own the car are most likely to default the loans compared to who owns the car.

BIVARIATE & MULTIVARIATE ANALYSIS

NAME_INCOME_TYPE VS NAME_EDUCATION_TYPE

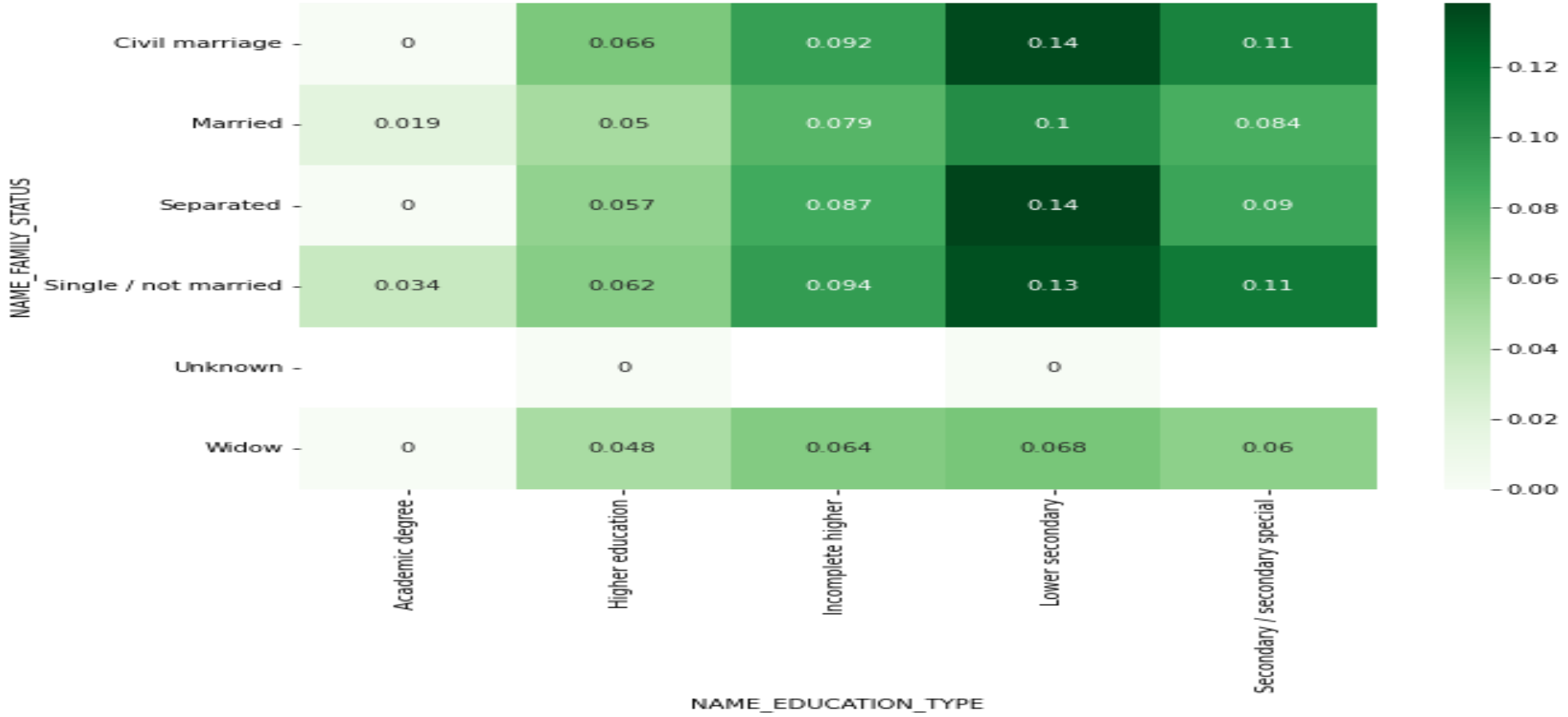


Insights:-

1. Unemployed people with Secondary Education are more likely the defaulters

BIVARIATE & MULTIVARIATE ANALYSIS

NAME_FAMILY_STATUS VS NAME_EDUCATION_TYPE

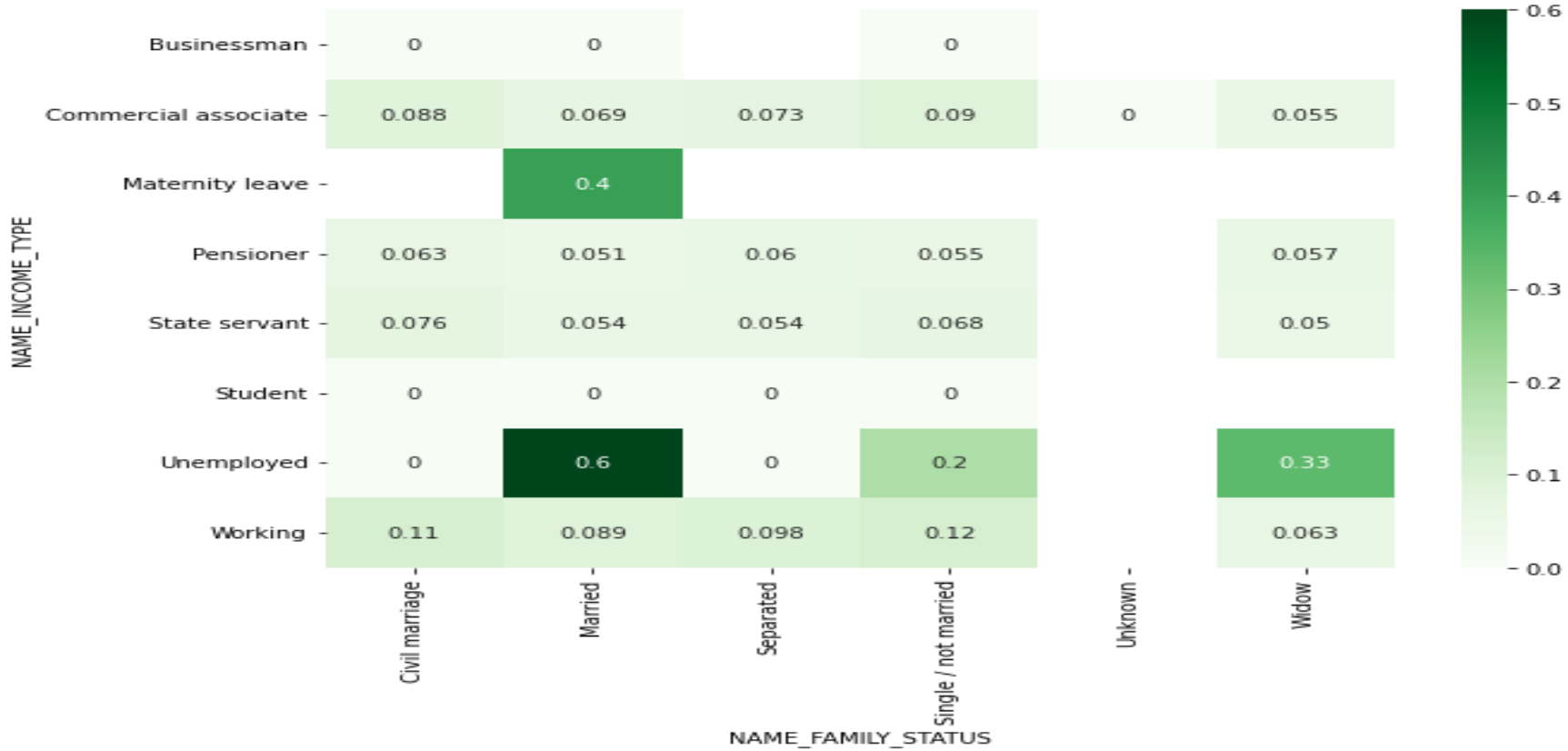


Insights:-

1. People with Lower Secondary education are more likely to defaulters
2. People who are single/not married and have Secondary Education are likely to be defaulters
3. People who have done civil marriage and have Secondary Education are likely to be defaulters

BIVARIATE & MULTIVARIATE ANALYSIS

NAME_INCOME_TYPE VS NAME_FAMILY_STATUS

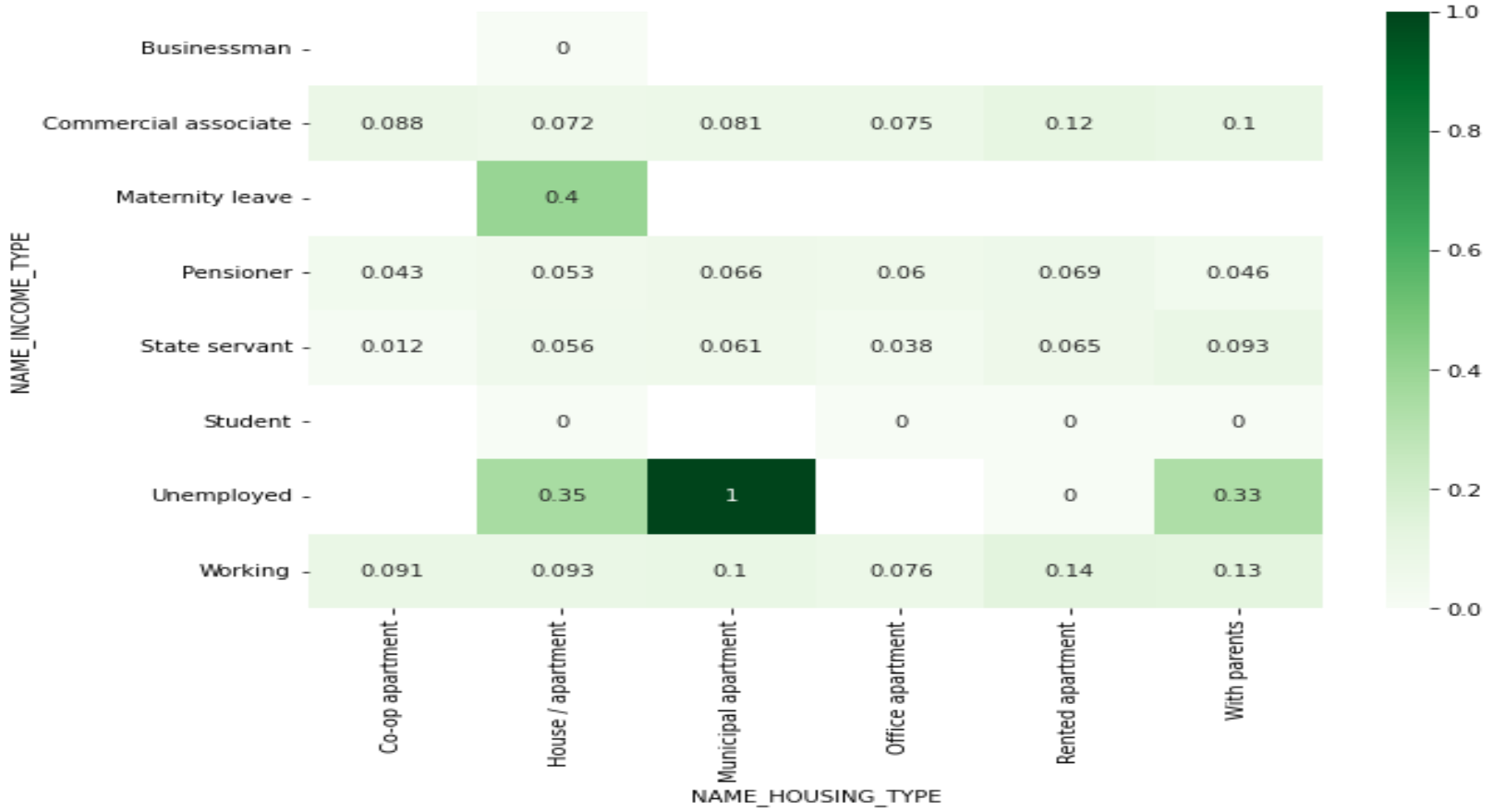


Insights:-

1. Married Unemployed people are more likely the defaulters
2. Unemployed Widow are likely to be defaulters

BIVARIATE & MULTIVARIATE ANALYSIS

NAME_INCOME_TYPE VS NAME_HOUSING_TYPE

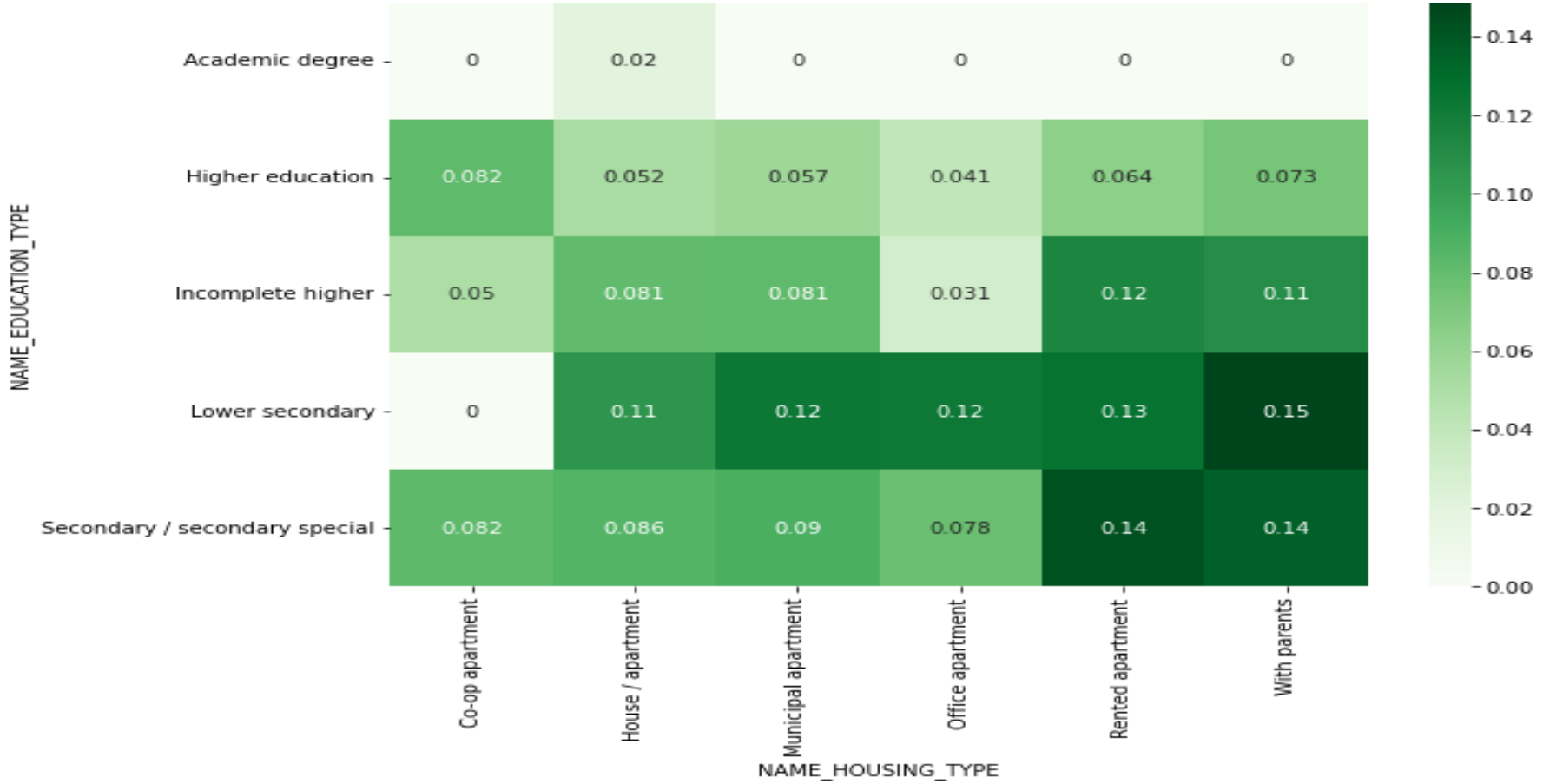


Insights:-

1. Unemployed people living in Municipal apartment will be defaulters always

BIVARIATE & MULTIVARIATE ANALYSIS

NAME_EDUCATION_TYPE VS NAME_HOUSING_TYPE

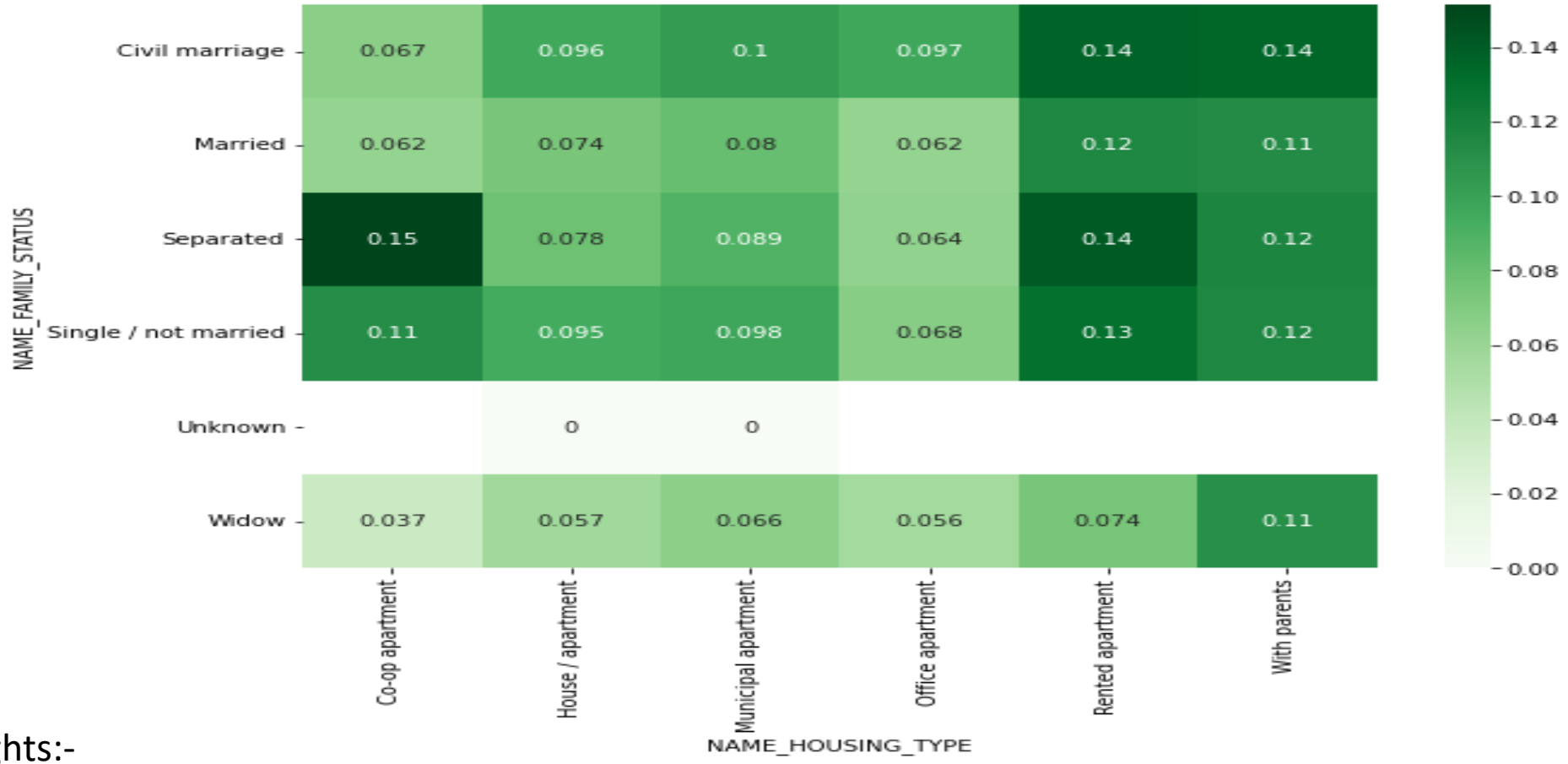


Insights:-

1. People with Lower Secondary education and living with parents are more likely to be defaulters
2. People with Secondary education and living in rented apartments or with parents are more likely to be defaulters

BIVARIATE & MULTIVARIATE ANALYSIS

NAME_FAMILY_STATUS VS NAME_HOUSING_TYPE

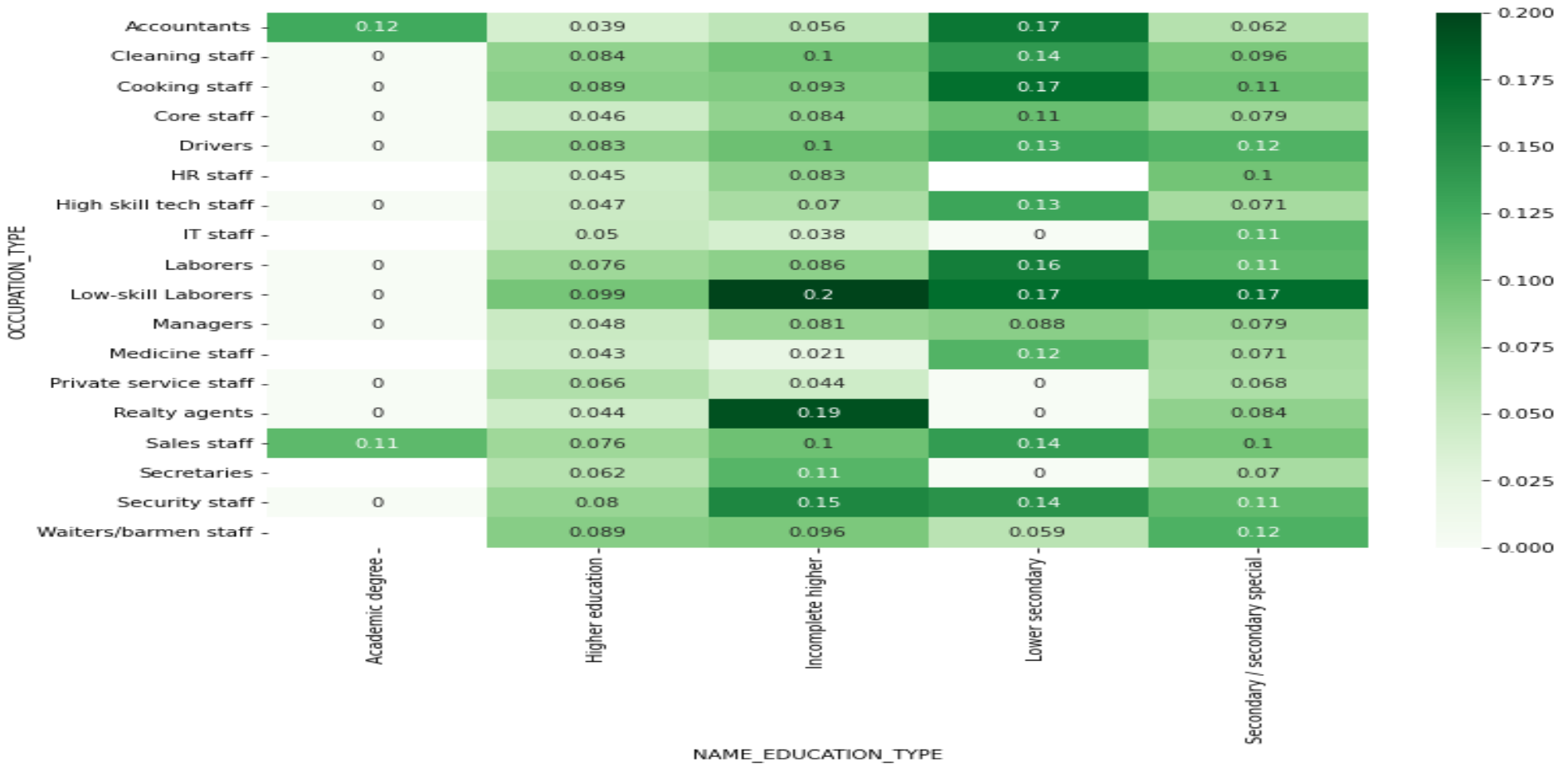


Insights:-

1. People who are separated and living in Co-op Apartments are more likely to be defaulters
2. People who are living in Rented Apartments are more likely to be defaulters except widows
3. People living with parents are more likely to be defaulters
4. People who are single/not married living in Co-op Apartment are more likely to be defaulters

BIVARIATE & MULTIVARIATE ANALYSIS

OCCUPATION_TYPE VS NAME_EDUCATION_TYPE

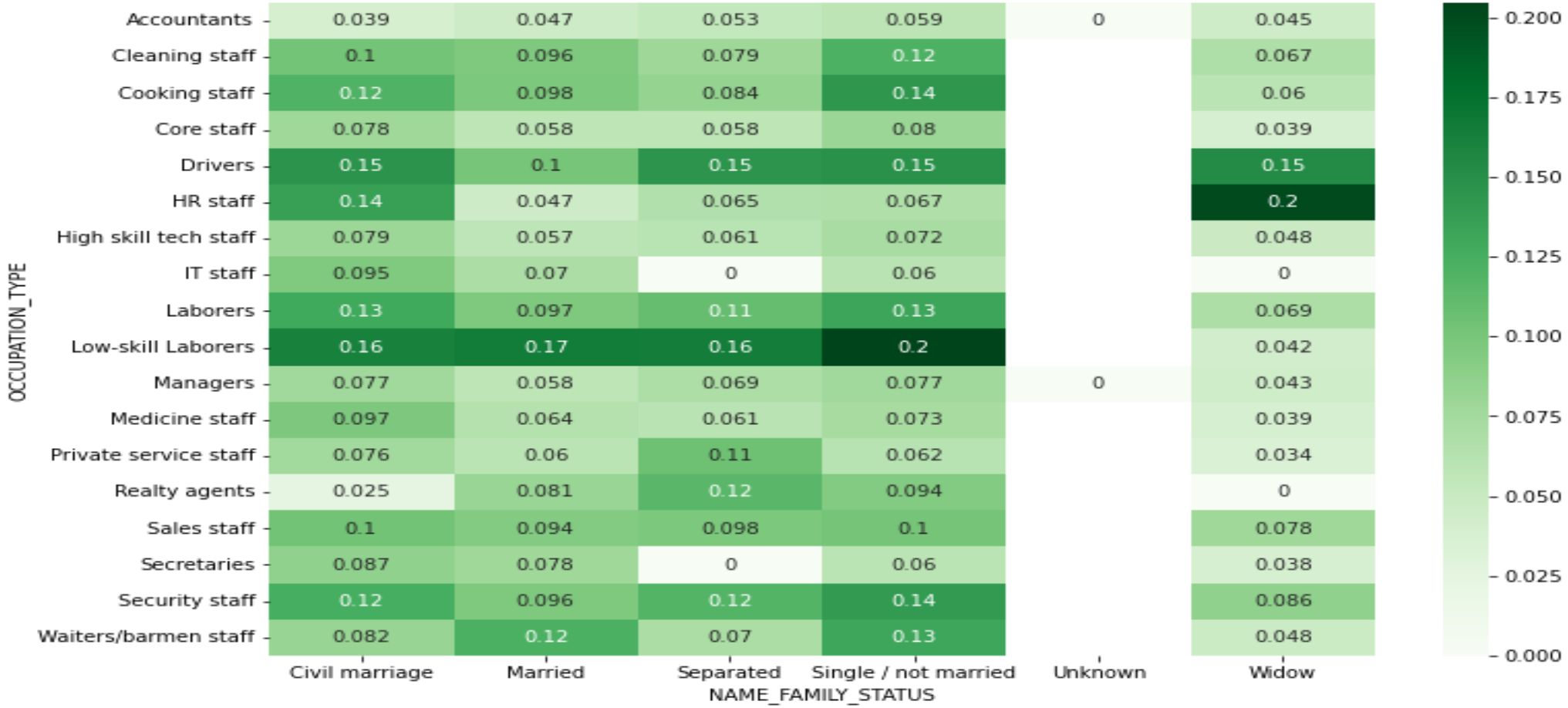


Insights:-

1. People who are Low Skill Laborer are more likely to be defaulters except those having higher education
2. People who are Realty Agents and have incomplete higher education are more likely to be defaulters

BIVARIATE & MULTIVARIATE ANALYSIS

OCCUPATION_TYPE VS NAME_FAMILY_STATUS

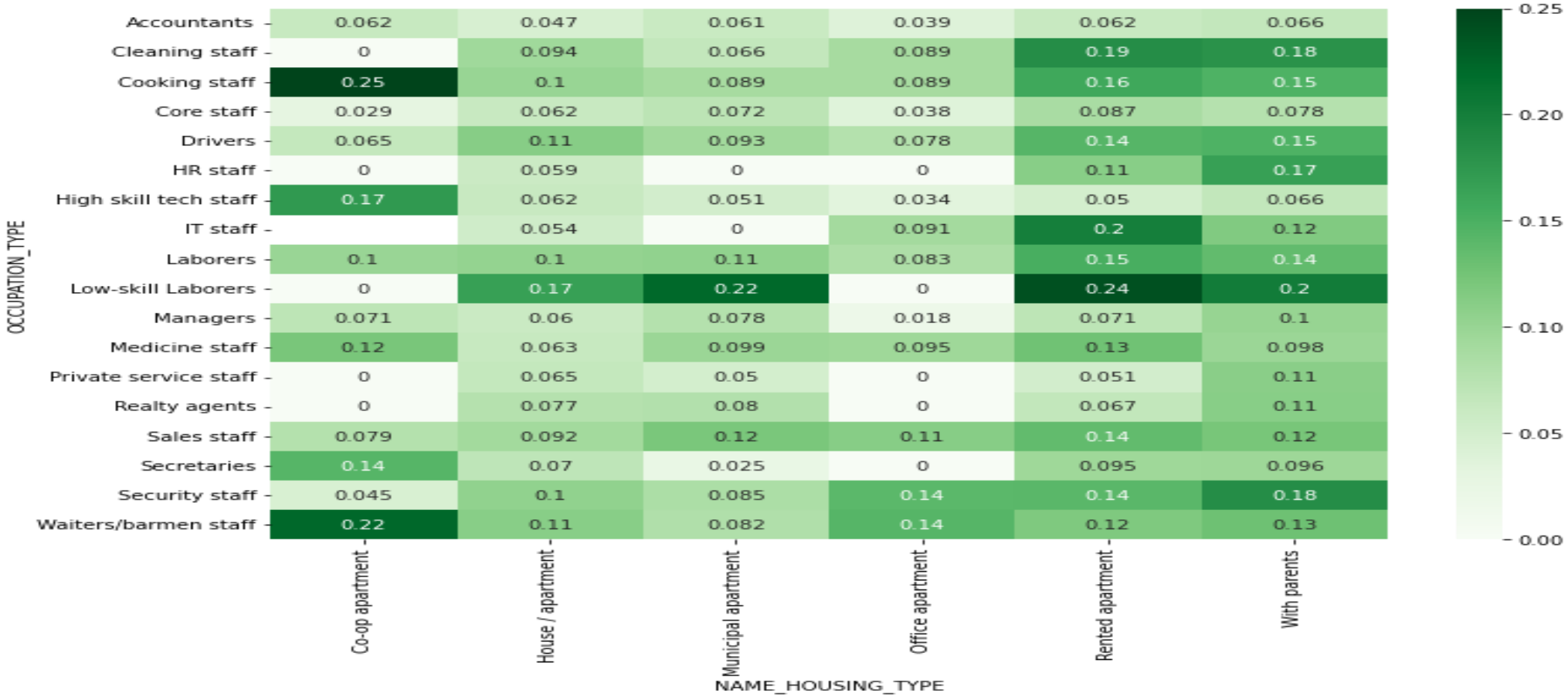


Insights:-

1. People who are Low Skill Laborer are more likely to be defaulters except Widows.
2. Widows who are working as Drivers and HR Staff are more likely to be defaulters
3. People who are working as drivers are more likely to be defaulters

BIVARIATE & MULTIVARIATE ANALYSIS

OCCUPATION_TYPE VS NAME_HOUSING_TYPE

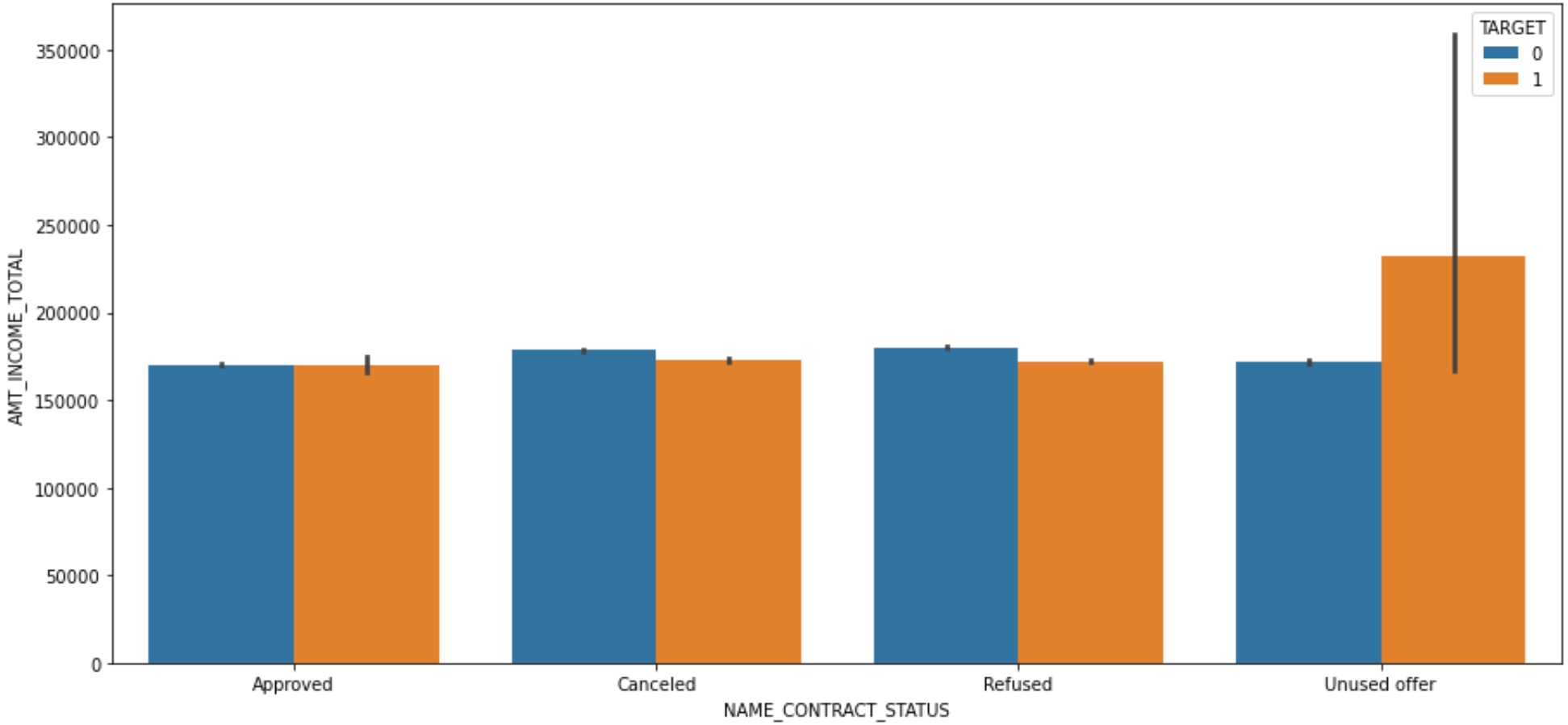


Insights:-

1. People who are living in Co-Op Apartment and working as Cooking Staff or Waiters/Barmen Staff are more likely to be defaulters
2. People who are living in Co-Op Apartment and working as High Skill Tech Staff are more likely to be defaulters
3. People who are living with Parents and working as Security Staff are more likely to be defaulters
4. People who are living with Parents or in Rented Apartments and working as Cleaning Staff are more likely to be defaulters

BIVARIATE & MULTIVARIATE ANALYSIS

NAME_CONTRACT_STATUS VS NAME_CONTRACT_STATUS



Insights:-

1. People who have Unused Offer are more likely to be defaulters

Conclusion- General Insights

- Cash loans widely preferred loan type
- Mostly People are unaccompanied by somebody while applying for loans
- Loans are applied mostly be people in age range of 30-40,40-50
- People with 60+ primarily senior citizens are very less likely to apply for loans
- Loans are applied mostly be people in low income range(1 Lac to 2 Lac)
- Females Re payers are approx. double of male Re payers

- Working People and Pensioner should be offered loans on higher rates as they have equal chances of repaying and defaulting the loans.

Conclusion- Defaulters

- Married People have more chances to default
- People with age 60+ are least likely to take loans. Moreover they have least chances of defaulting the loans
- People with Low Income (1 Lac to 2 Lac) have more chances to defaults on loans
- People without car
- People who are employed with less than 10 years
- People with Academic Degree
- Males who do not own the car
- Unemployed people with Secondary Education
- People with Lower Secondary education
- People who are single/not married and have Secondary Education
- People who have done civil marriage and have Secondary Education
- Married Unemployed people
- Unemployed Widow
- Unemployed people living in Municipal apartment
- People with Lower Secondary education and living with parents
- People with Secondary education and living in rented apartments or with parents
- People who are separated and living in Co-op Apartments
- People living with parents

Conclusion- Defaulters contd…

- People who are single/not married living in Co-op Apartments
- People who are Low Skill Laborer except those having higher education
- People who are Realty Agents and have incomplete higher education
- Widows who are working as Drivers and HR Staff
- People who are working as drivers
- People who are living in Co-Op Apartment and working as Cooking Staff or Waiters/Barmen Staff
- People who are living in Co-Op Apartment and working as High Skill Tech Staff
- People who are living with Parents and working as Security Staff
- People who are living with Parents or in Rented Apartments and working as Cleaning Staff
- People who have Unused Offer are more likely to be defaulters

Conclusion- Repayer

- People accompanied by somebody have higher chances
- Businessman and students are good choice to give loans
- Single Not Married people are good choice to give loans
- Realty agents, Secretaries, Walters/barmen staff, IT staff, HR Staff are good choice to give loans
- People in age group 30-40 and most likely to repay the loan
- People with Very High (4 Lac to 5 Lac) and Extremely High (5 Lac+)
- Females are most likely to repay the loan as well
- People who owns Flats
- People with Secondary Education
- People with House/Apartment
- People who are employed with more than 50+Years
- Women on Maternity leave