EDA CASE STUDY

CREDIT EDA ASSIGNMENT

ABHIMANYU GANGANI

DSC 40

PROBLEM STATEMENT

- In this problem we are having two datasets as follow:

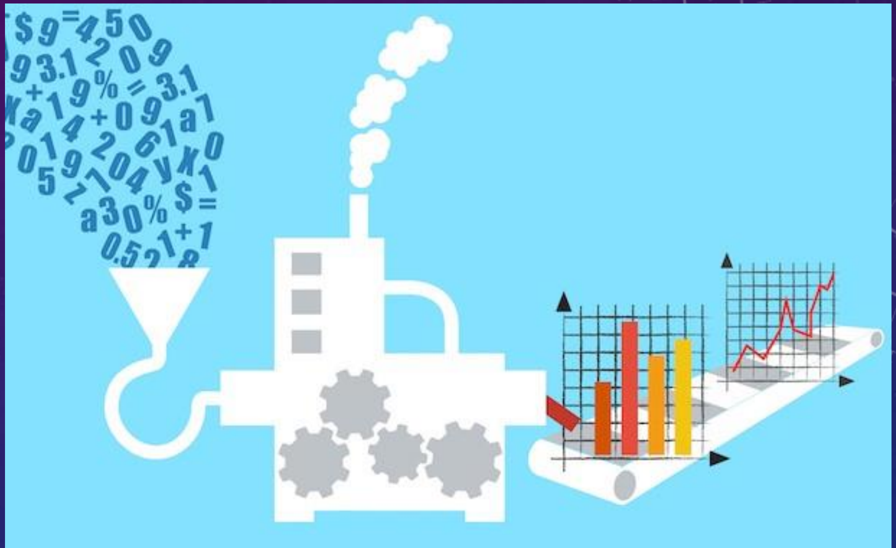
1. Application Data
2. Previous Application Data

- Applicants are bifurcated in two type one who clears the loan on time (TARGET : 0) and one’s who don’t clears the loan (TARGET : 1) or defaulters
- Using these data we have to perform EDA on what are the type of people to which bank may provide loan.
- Again there are two business scenarios:

1. If the loan is provided to defaulter, its business loss
2. If the loan is not provided to non-defaulter, then also its business loss

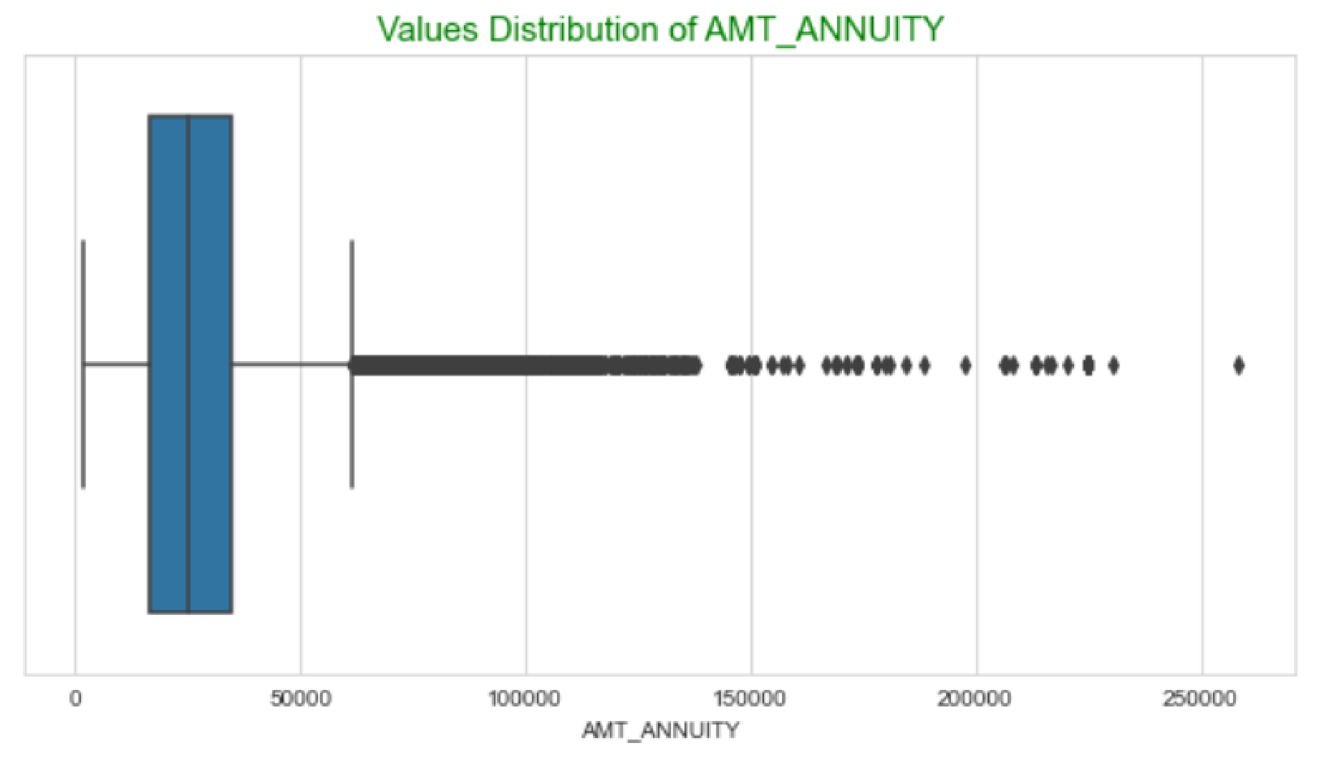
- We will be doing UNIVARIATE and BIVARIATE analysis to find out behaviour of different categories of the people

STEPS TO BE TAKEN

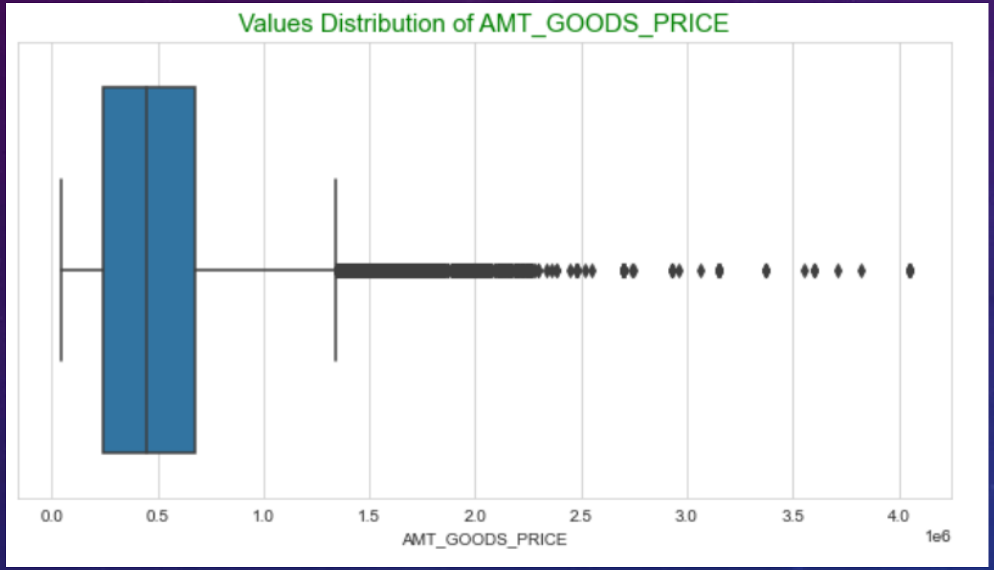


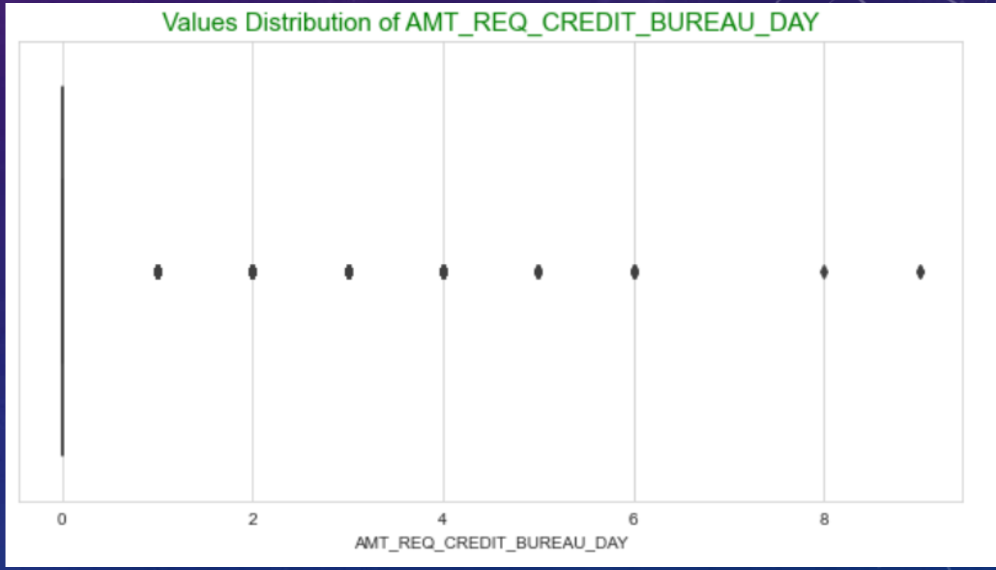
- Import the application dataset
- Check the structure of the application data dataset
- Drop all the columns which are not required
- Identify the outliers and fix them
- Impute the null values whenever required
- Correct the data and datatype wherever required
- Bifurcate the data into Defaulter and non-defaulter
- Perform Univariate analysis
- Perform different types of bivariate analysis
- Import the previous applications dataset and perform necessary action like dropping and imputing the nulls along with correcting the data
- Merge both the datasets on some common unique values
- Perform Univariate and Bivariate analysis on the merged dataset.
- Point the actions that should be taken by bank while providing credits or loans

OUTLIERS IN THE DATASET

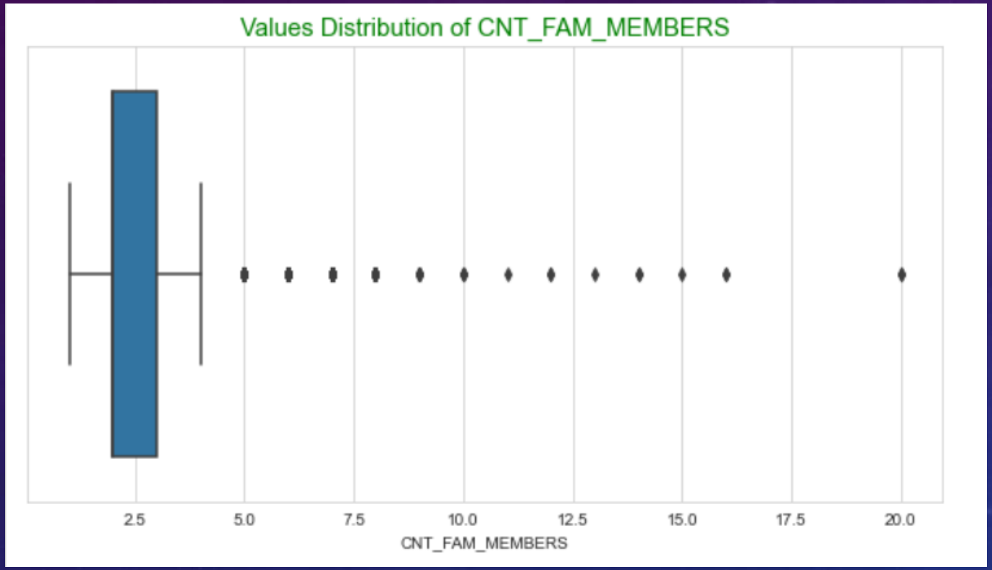


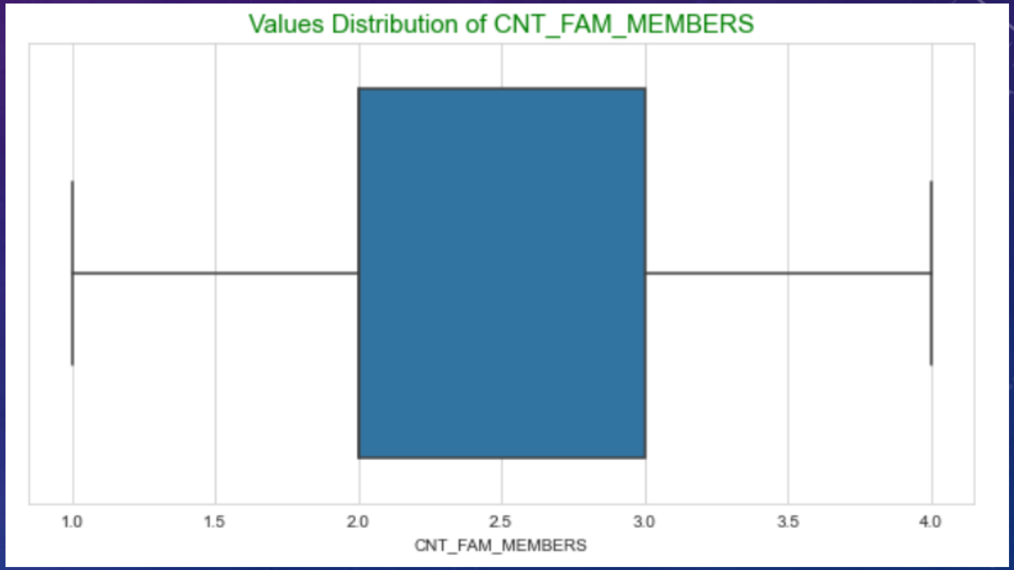
OUTLIERS IN THE DATASET





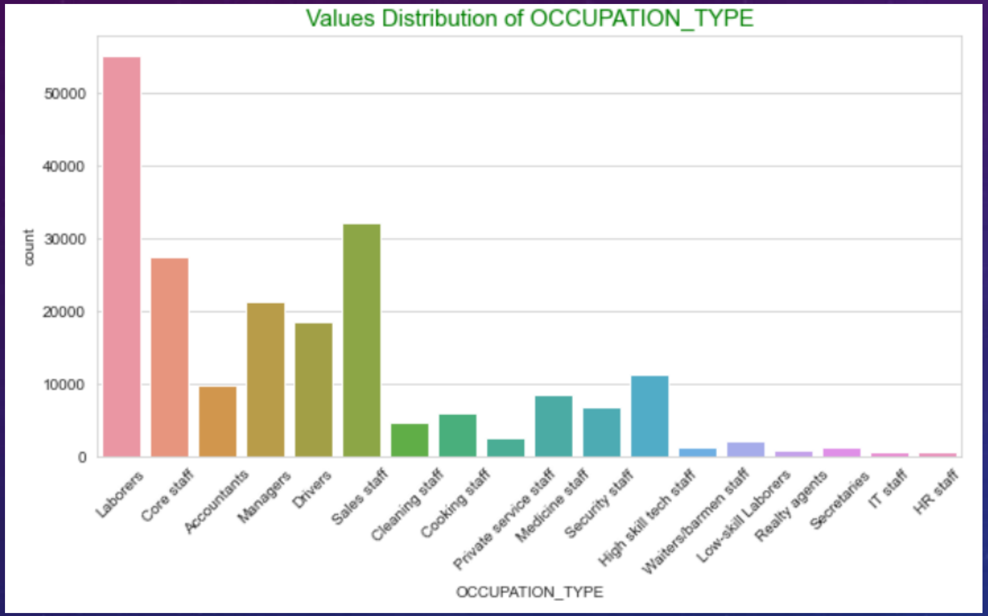
HANDLING OUTLIERS IN THE DATASET

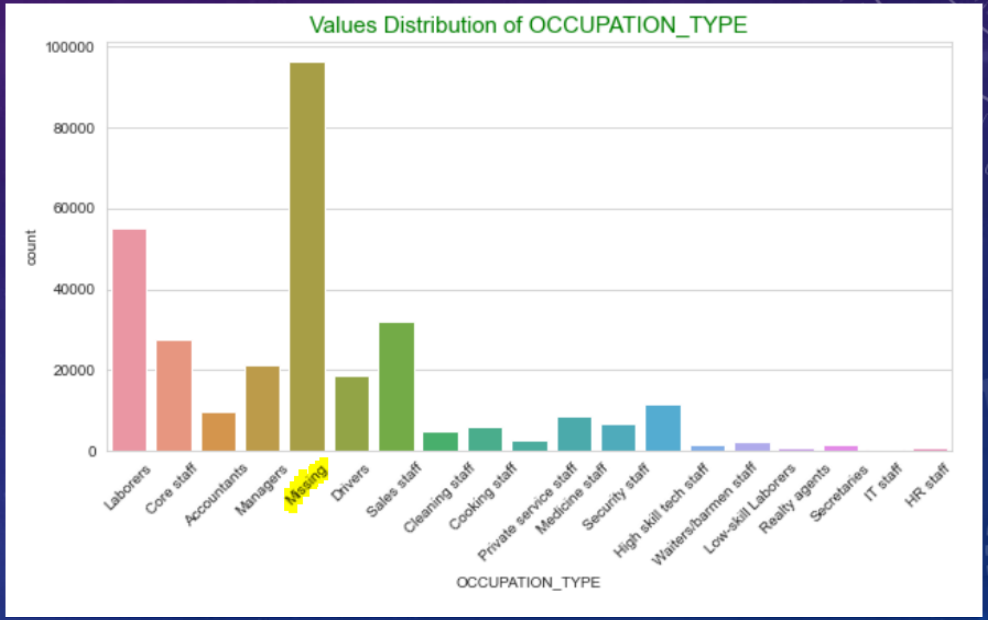




Handling outliers by capping them

IMPUTING THE NULL VALUES

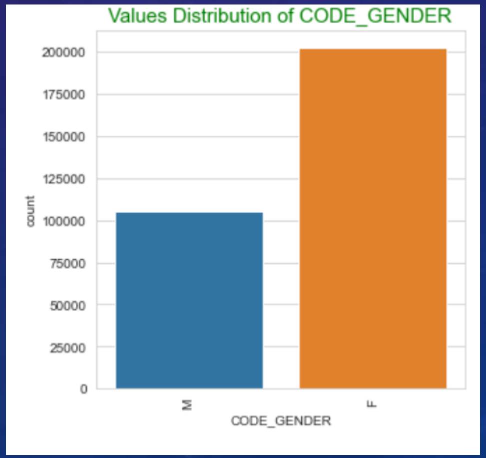




IMPUTING THE NULL VALUES with Keyword 'MISSING' as the count is high

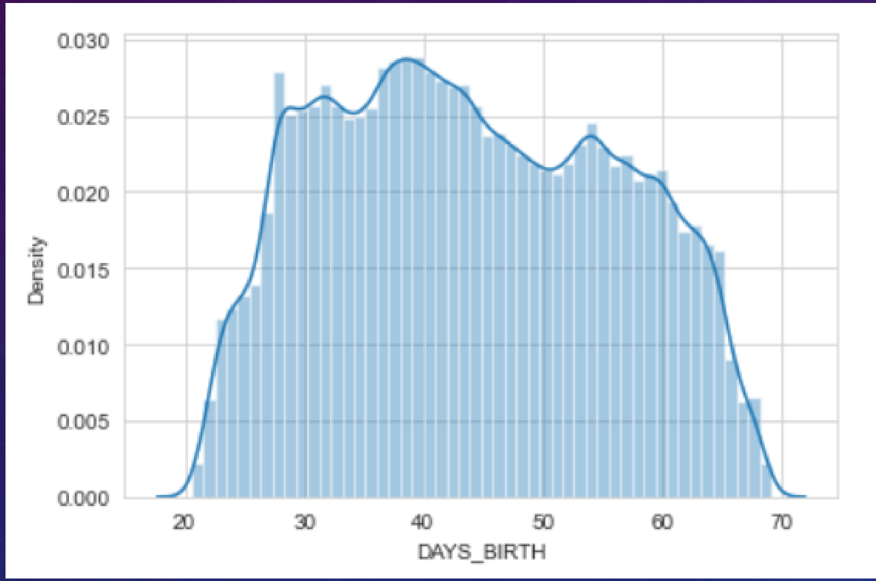
IMPUTING THE NULL VALUES

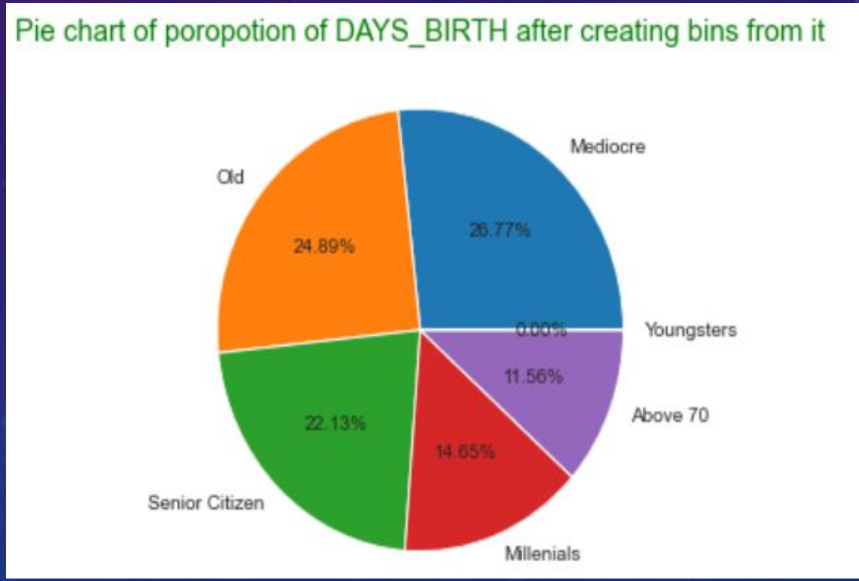
- IF the column is numerical we can either impute the null values with median or mean, depending on the count of outliers
- If there are many outliers we will impute with median and if there are less or no outliers we will replace it with mean
- If the count of null values in categorical column is considerable we will impute it with the mode value
- If the count of null values in categorical column is more than count of any other category we will assign it as new category like ‘Missing’



BINNING

- If we feel like that continuous data should be converted to categorical data we can perform binning like on age column

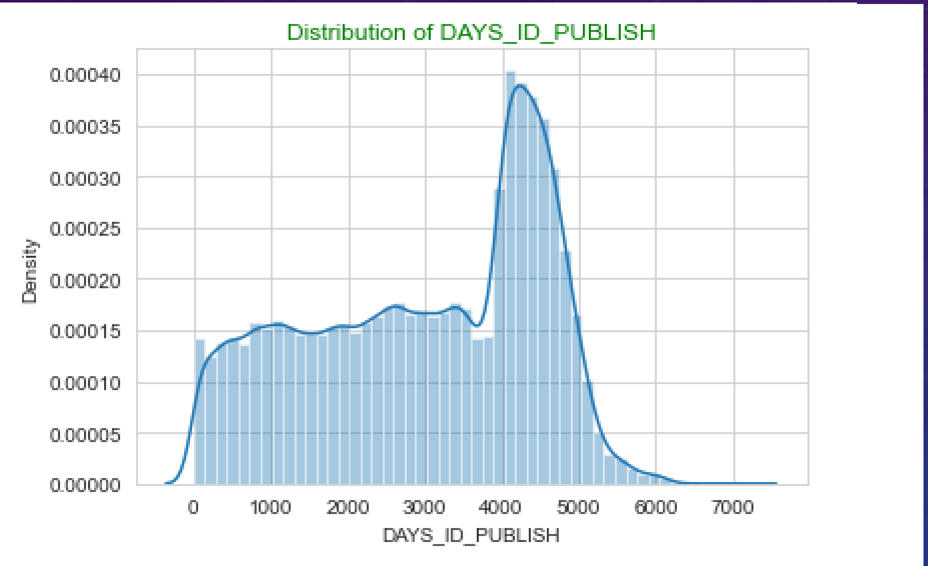


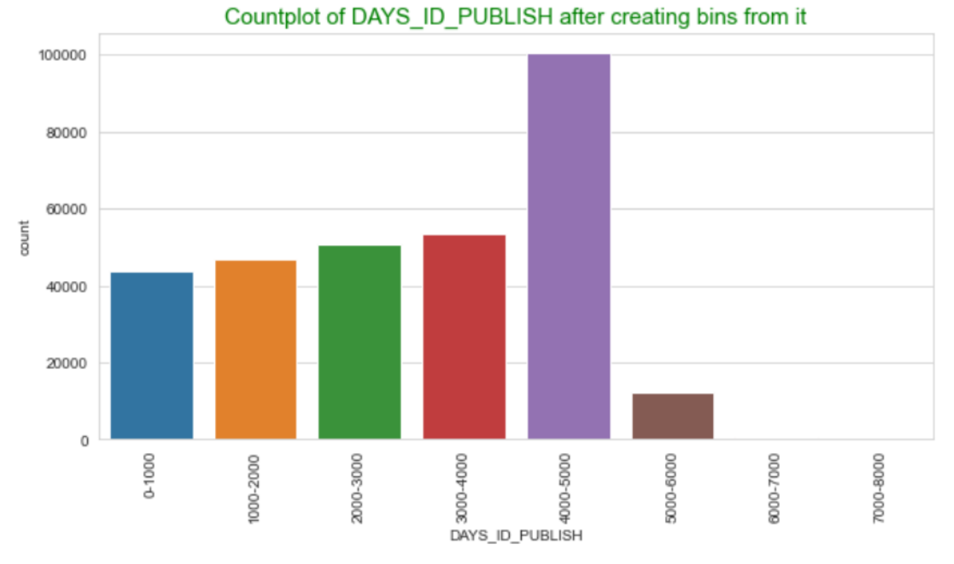


BEFORE BINNING

AFTER BINNING

BINNING



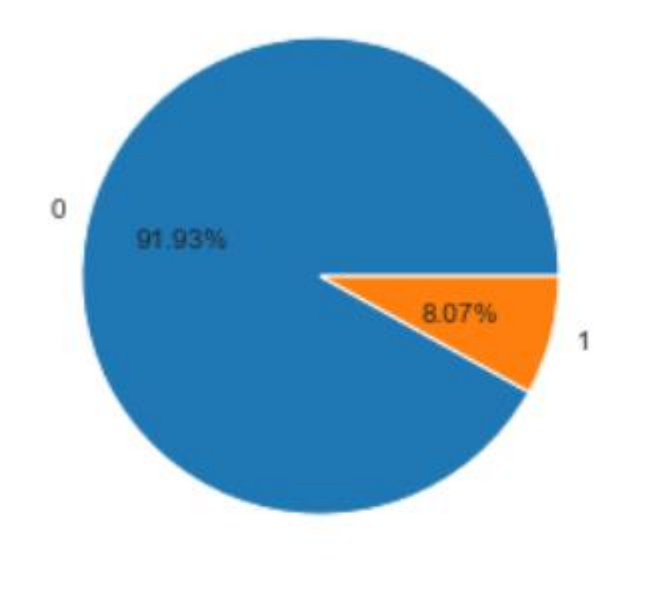


BEFORE BINNING

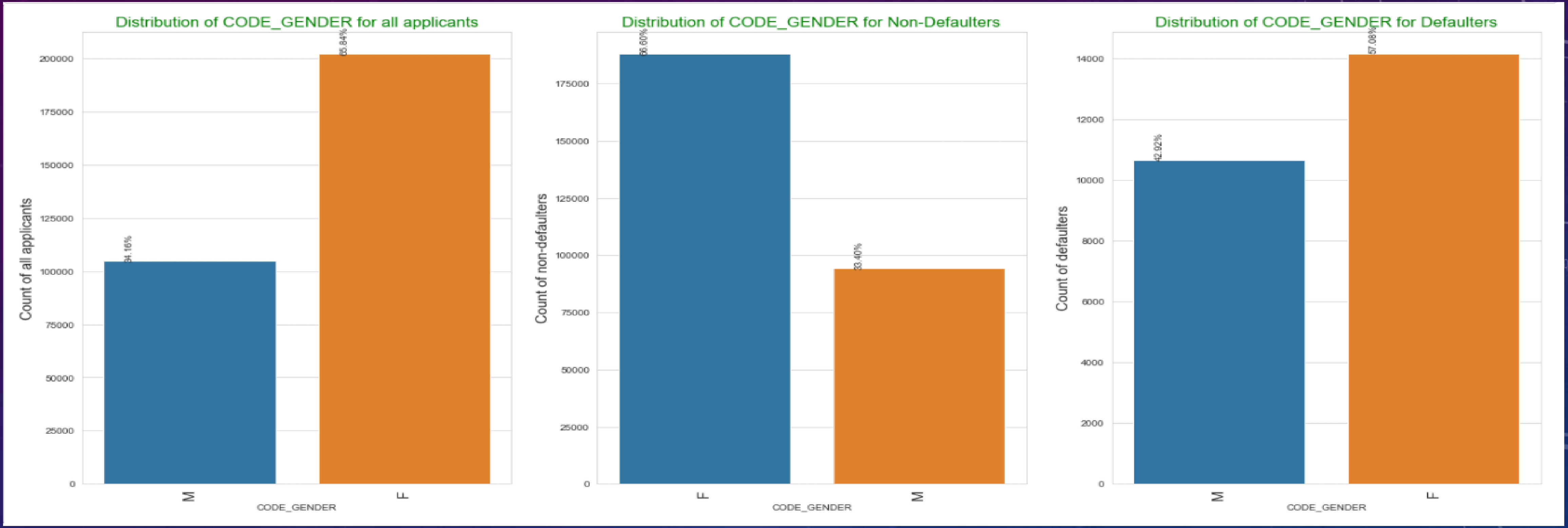
AFTER BINNING

EXPLORATORY DATA ANALYSIS

TARGET COLUMN PROPOTIONS/IMBALANCE PERCENTAGE

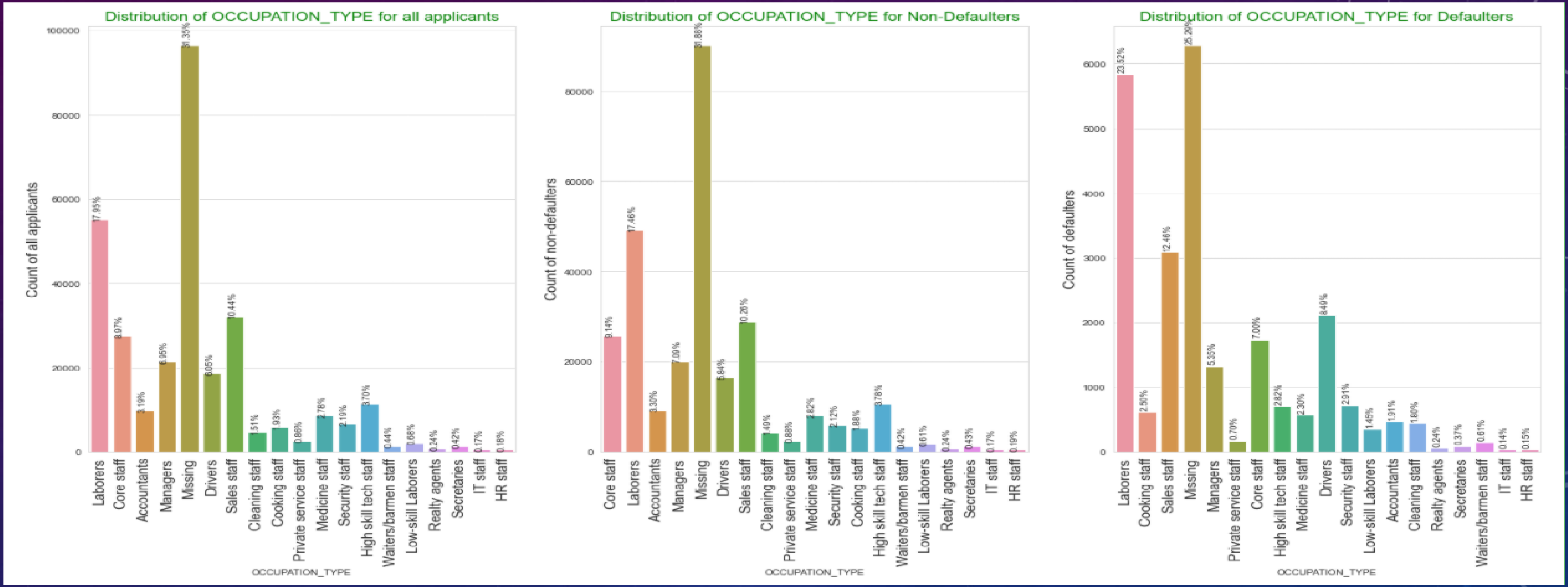


UNIVARIATE ANALYSIS ON GENDER



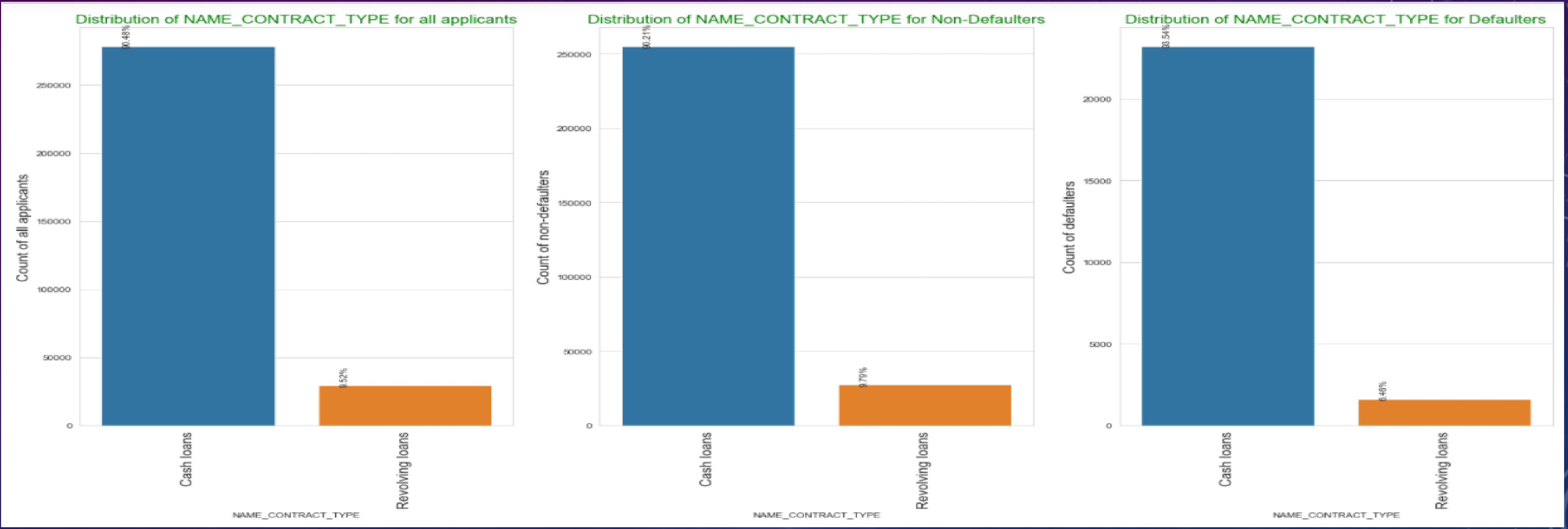
- In both the defaulters and non-defaulters analysis we can find that number of females are high.
- So it can be concluded on the basis of gender that females is more likely to become defaulter.
- Females are applying for more loans as compared to males.

UNIVARIATE ANALYSIS ON OCCUPATION_TYPE



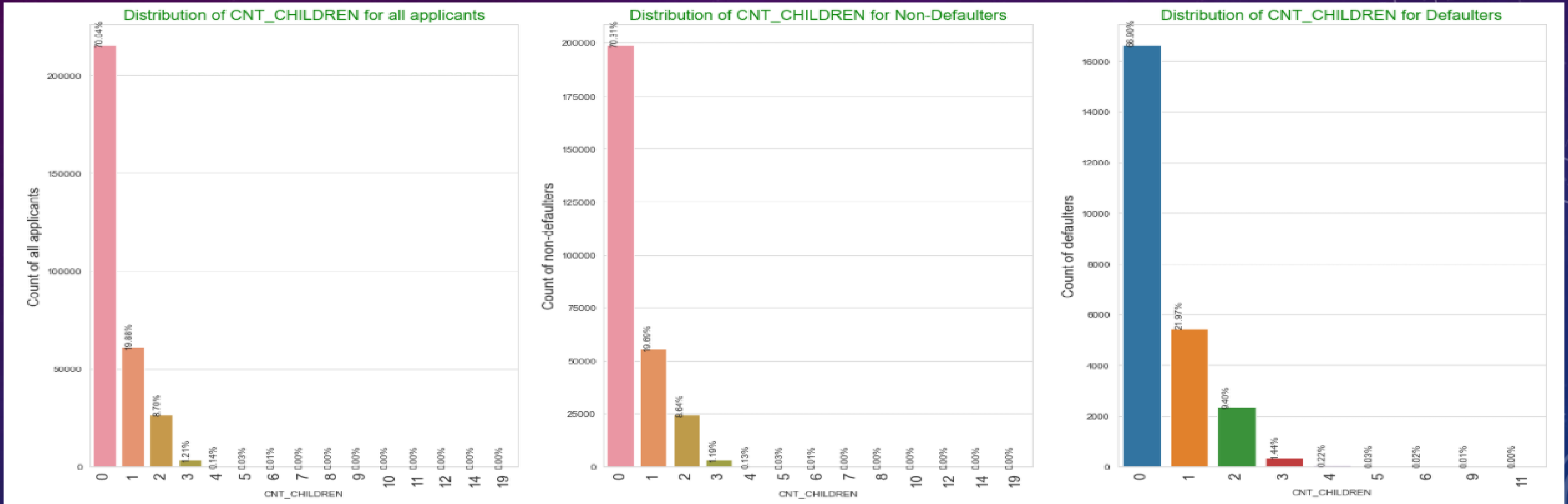
- Most of the people are not comfortable in telling their occupation.
- Laborers, Sales staff, Core staff, Driver are the occupation that mostly apply for loans and becaome defaulters
- HR staff, IT staff and Secretaries are less likely to apply for loan and if does pays loan on time

UNIVARIATE ANALYSIS ON NAME_CONTRACT_TYPE



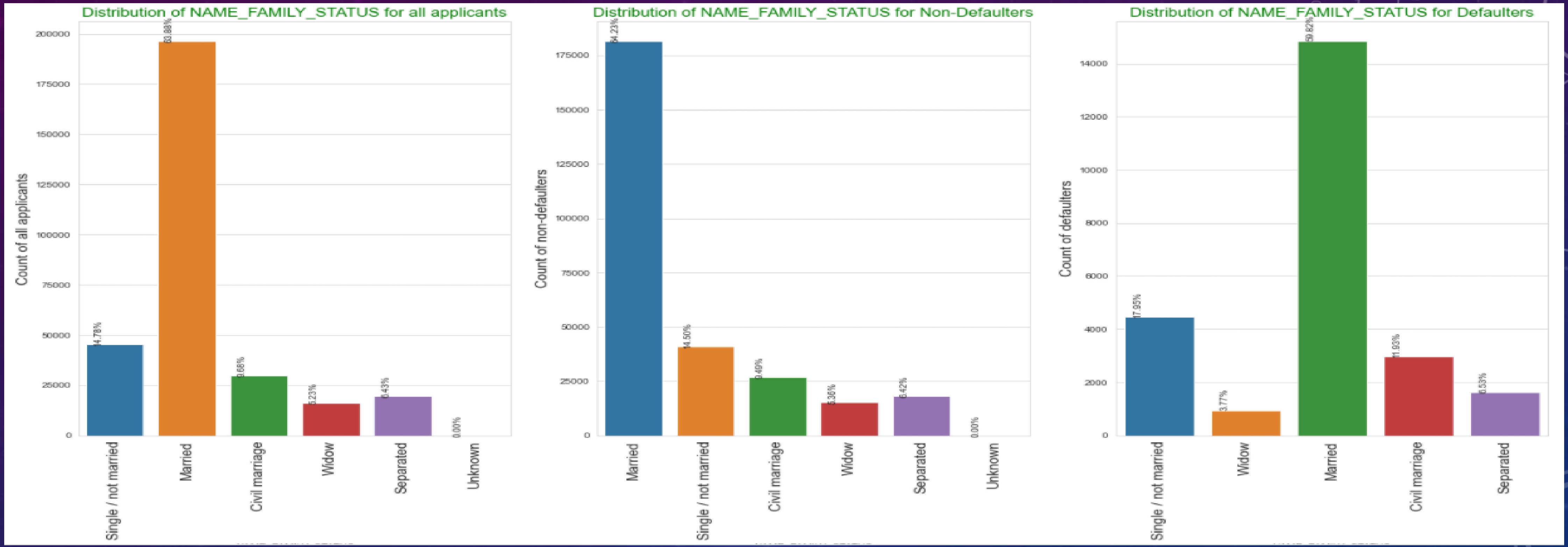
- Nearly 93% of the defaulters are prefering to apply for cash loans 6.46% of the defaulters are having revolving loans

UNIVARIATE ANALYSIS ON CNT_CHILDREN



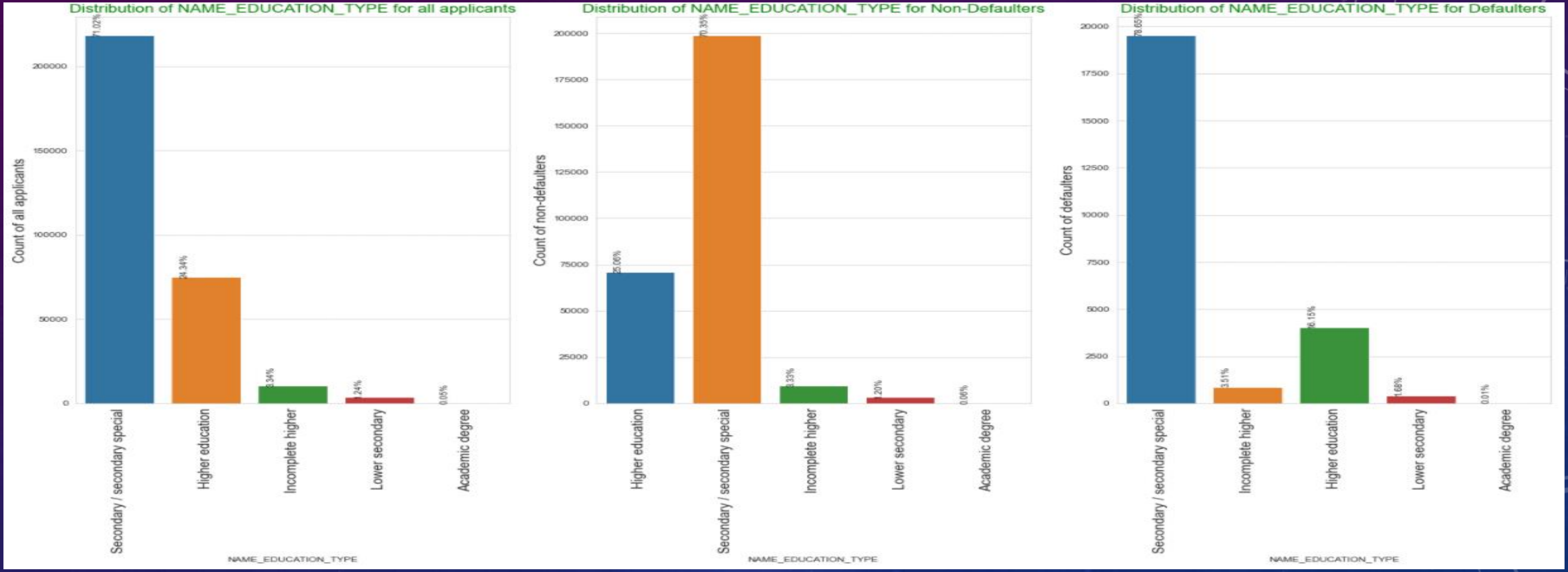
- 70% of the people applying for loan having no children and more likely to become defaulter(66.9%) as the count is high.
- As the count of children increases the chance of applying for loan and become defaulter reduces

UNIVARIATE ANALYSIS ON NAME_FAMILY_STATUS



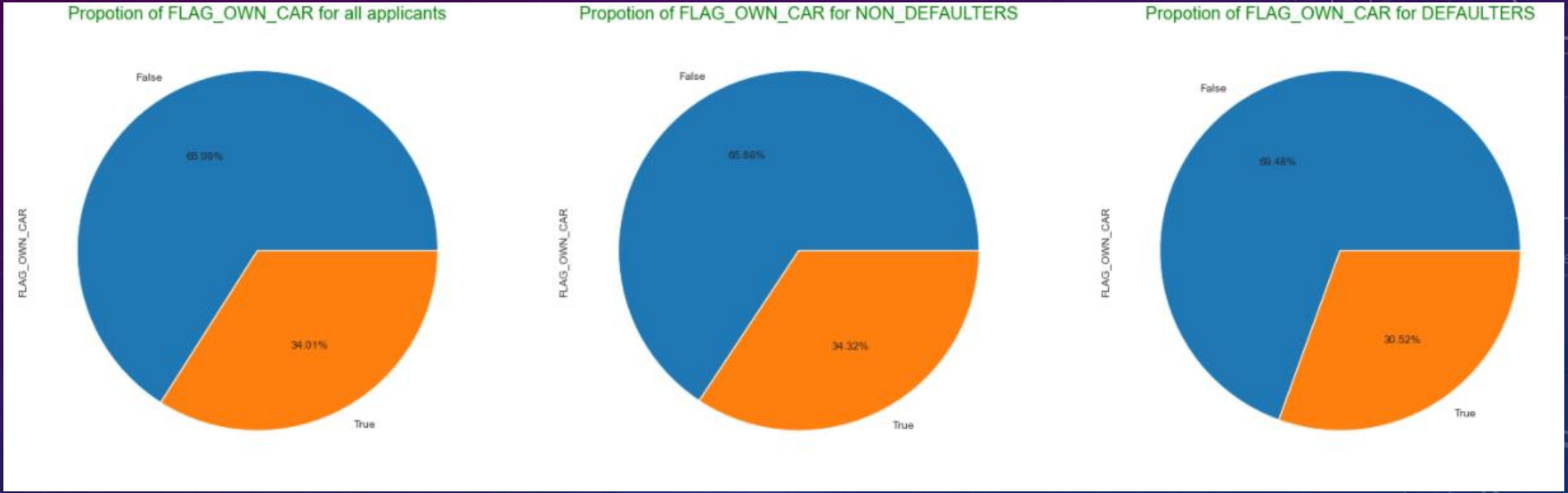
- Married people are the one's which are mostly applying for loan(63.88) and widows are very less likely to apply for loan(5.23%) and be defaulter
- 60% of the defaulters belongs to the married category and 3.77% of defaulters are widows which is compartivly less

UNIVARIATE ANALYSIS ON NAME_EDUCATION_TYPE



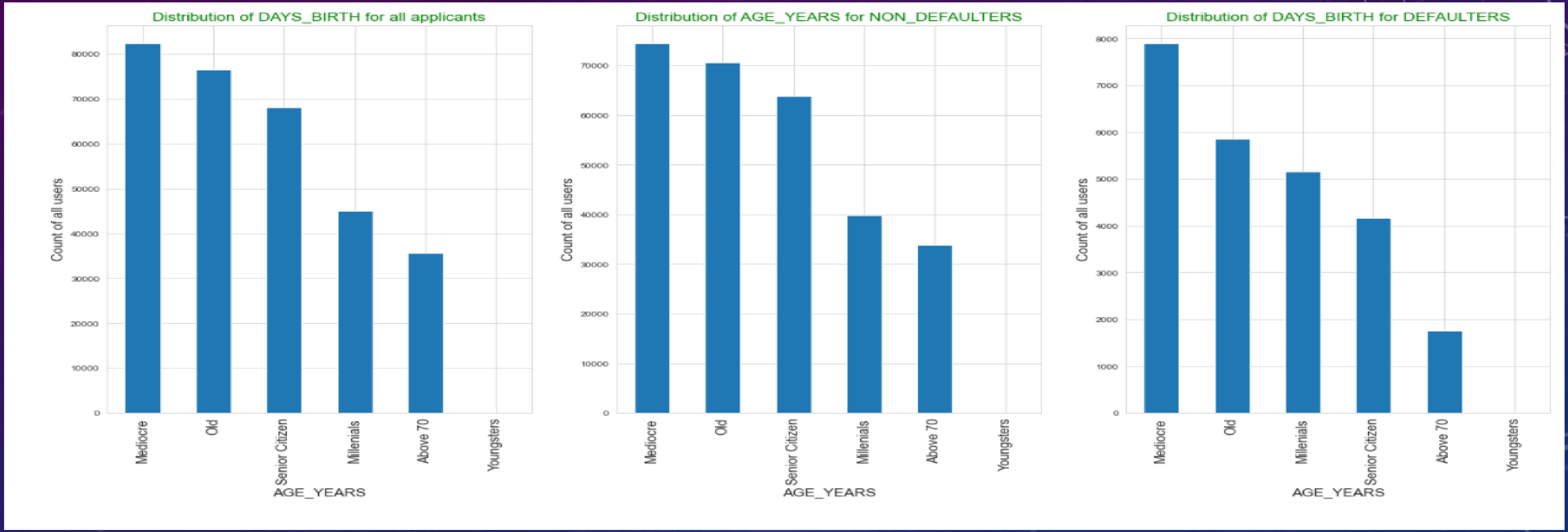
- Education has low casuation on becoming defaulter. Higher education means less chance of becoming defaulter
- People with secondary/secondary special education are most likely to apply for loan(71.08%) and become defaulter (78.65%)
- People with Academic degree are very less likely to apply for loan(0.05%) and become defaulter (0.01%)

UNIVARIATE ANALYSIS ON FLAG_OWN_CAR



- 67% of the people applying for loan doesn't owns car
- 70% of the people who are defaulter doesn't own's car

UNIVARIATE ANALYSIS ON DAYS_BIRTH



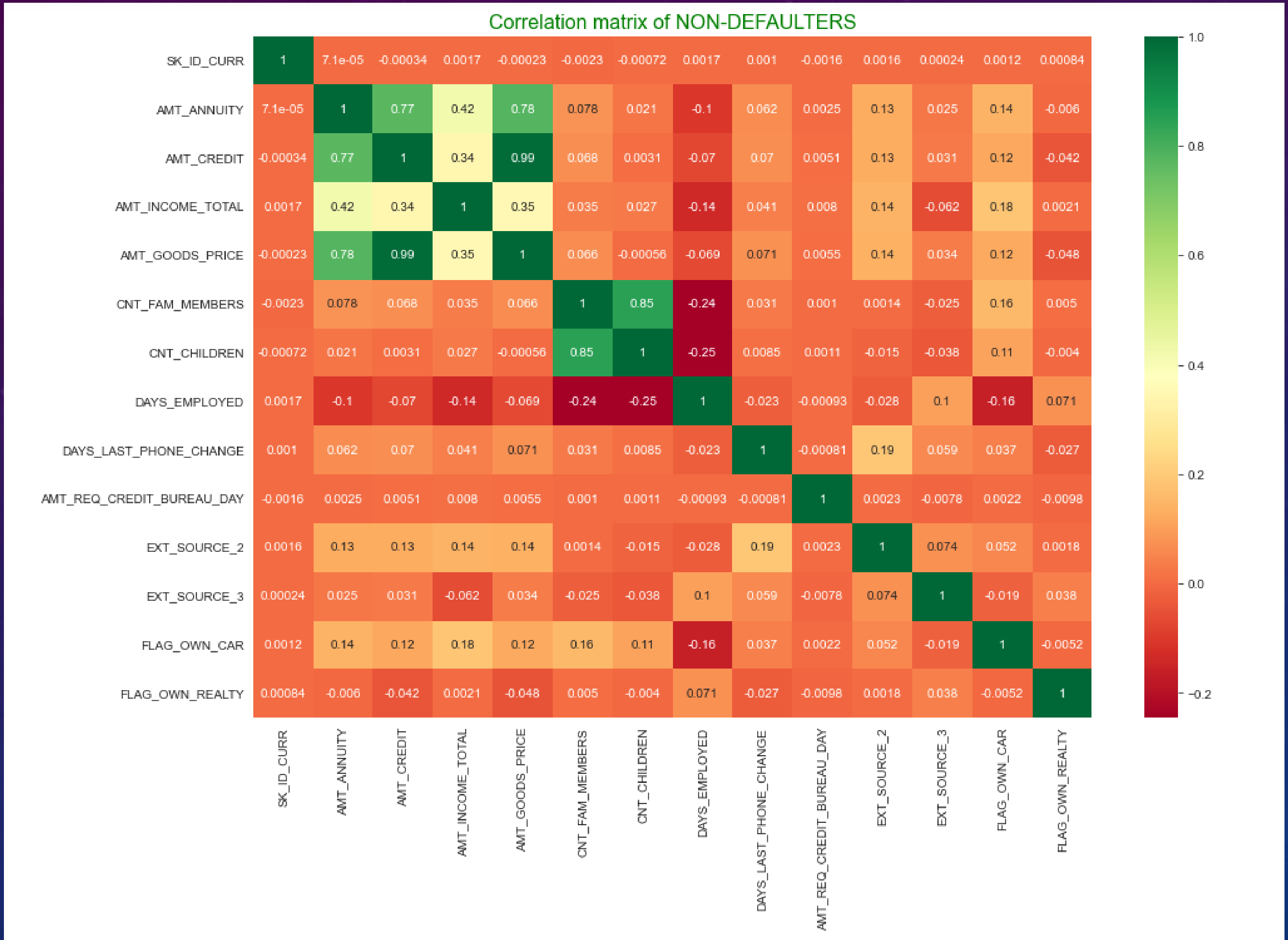
- Most of the people applying for loans are Mediocre i.e. (30-40 years of age)
- People above 70 years are having least chances of becoming defaulter
- There are less chance of people between (40-70) being defaulters as they pay loan on time

BIVARIATE ANALYSIS

There are 3 types of BIVARIATE analysis:

1. Numerical-Numerical
2. Numerical-Categorical
3. Cateogorical-Categorical

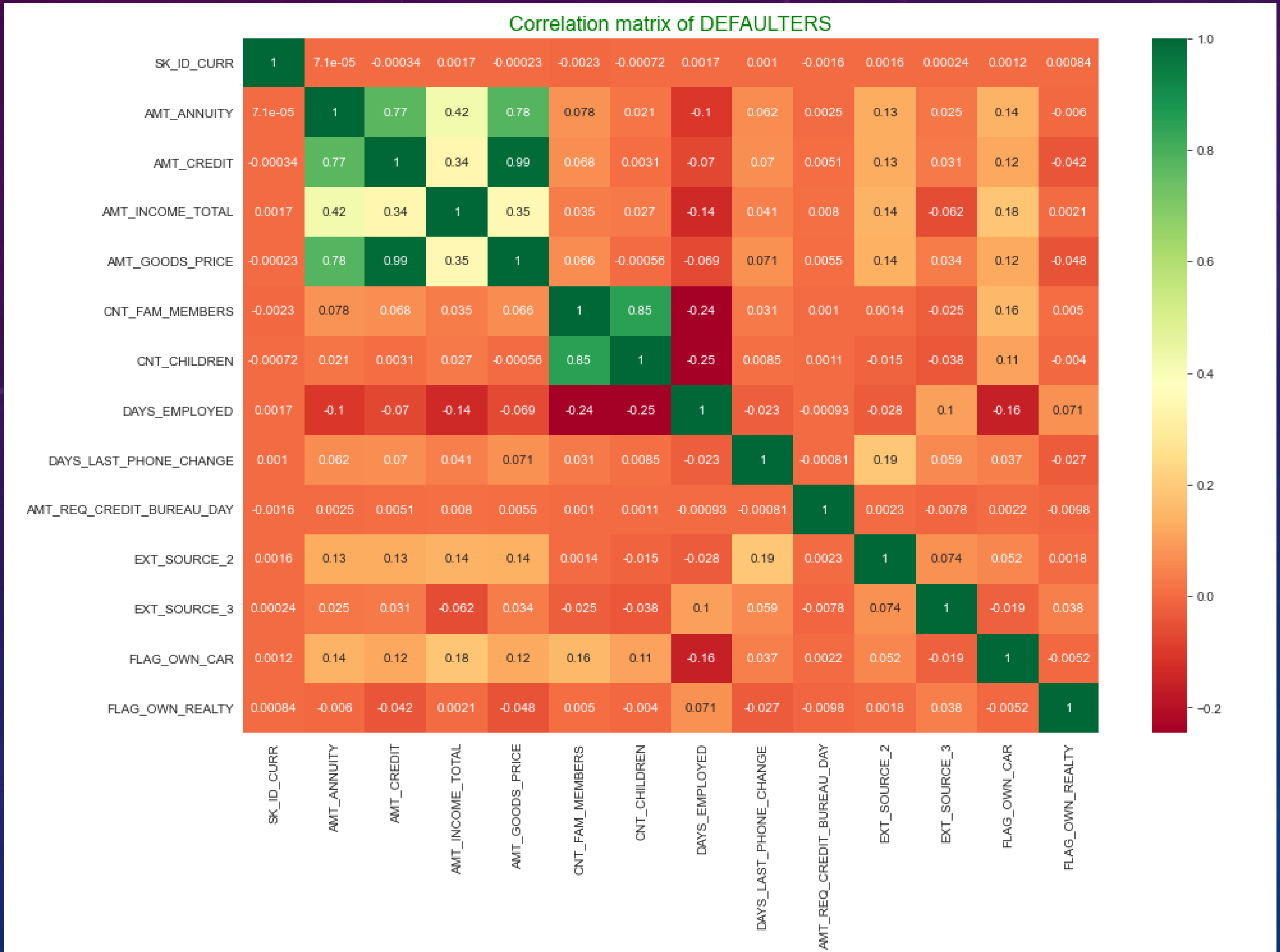
NUMERICAL-NUMERICAL BIVARIATE ANALYSIS(NON-DEFAULTERS)



UNDERSTANDING FROM CORRELATION MATRIX FOR NON DEFAULTERS

1. AMT_ANNUITY, AMT_GOODS_PRICE,AMT_INCOME_TOTAL,AMT_CREDIT are highly correlated
2. CNT_CHILDREN AND CNT_FAM_MEMBERS are highly correlated
3. DAYS_EMPLOYED is negatively correlated with CNT_FAMILY MEMBER AND CNT_CHILDREN

NUMERICAL-NUMERICAL BIVARIATE ANALYSIS(DEFAULTERS)



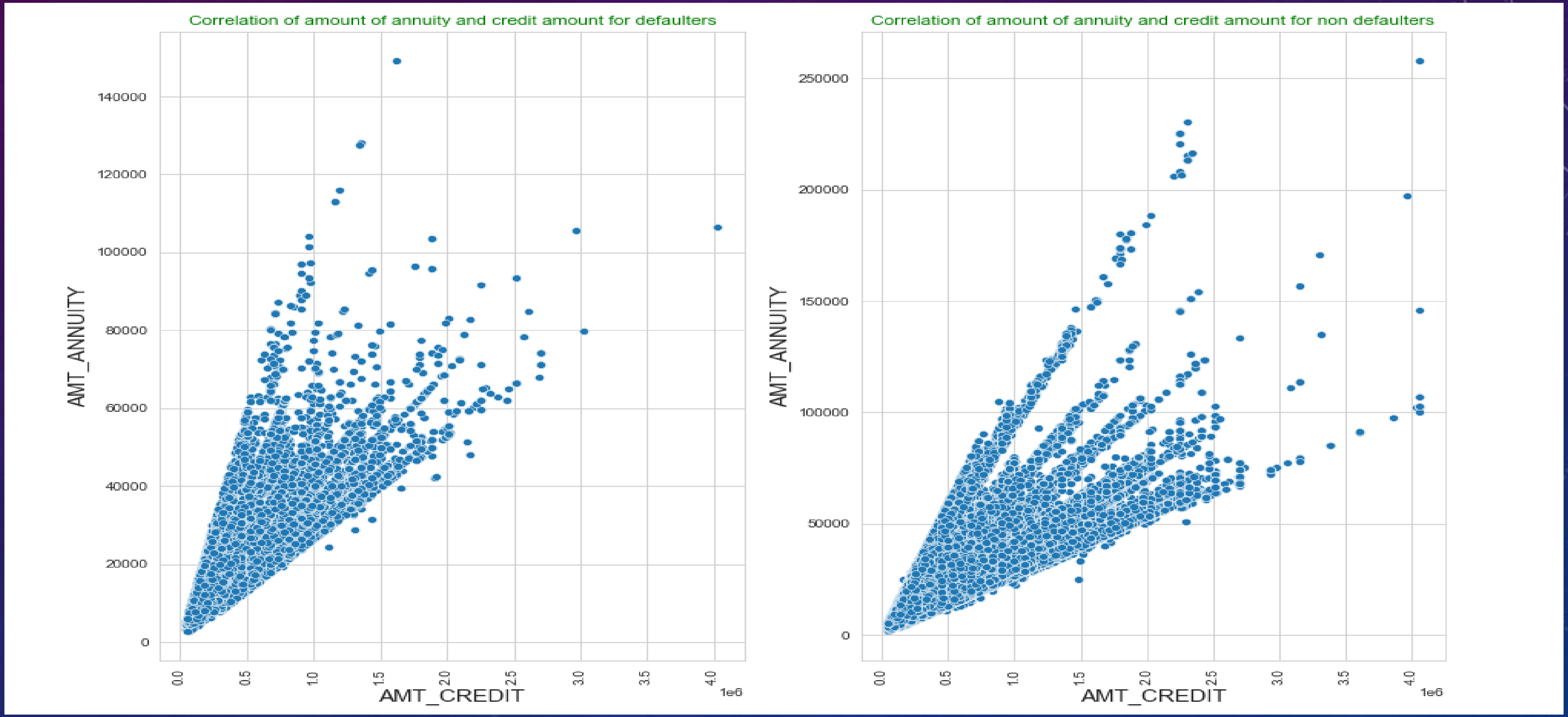
UNDERSTANDING FROM CORRELATION MATRIX FOR DEFAULTERS

1. AMT_ANNUITY, AMT_GOODS_PRICE,AMT_INCOME_TOTAL,AMT_CREDIT are highly correlated

2. CNT_CHILDREN AND CNT_FAM_MEMBERS are highly correlated

3. DAYS_EMPLOYED is negatively correlated with CNT_FAMILY MEMBER AND CNT_CHILDREN

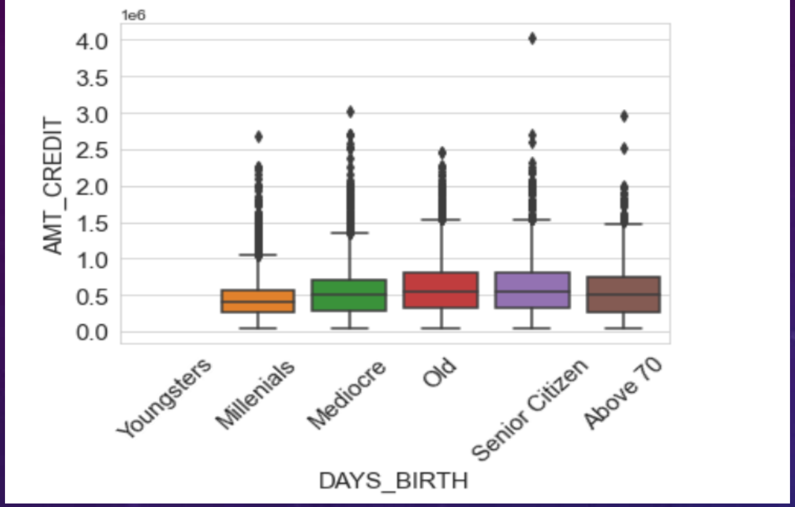
BIVARIATE ANALYSIS FOR AMT_CREDIT/AMT_ANNUITY

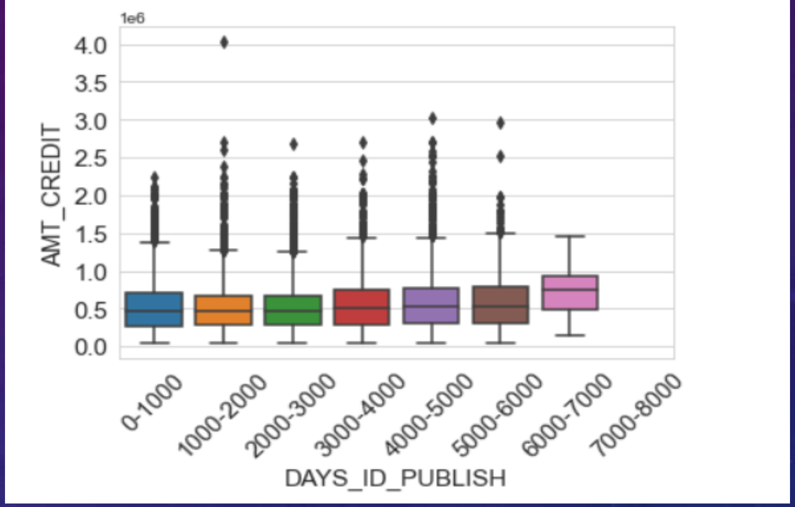


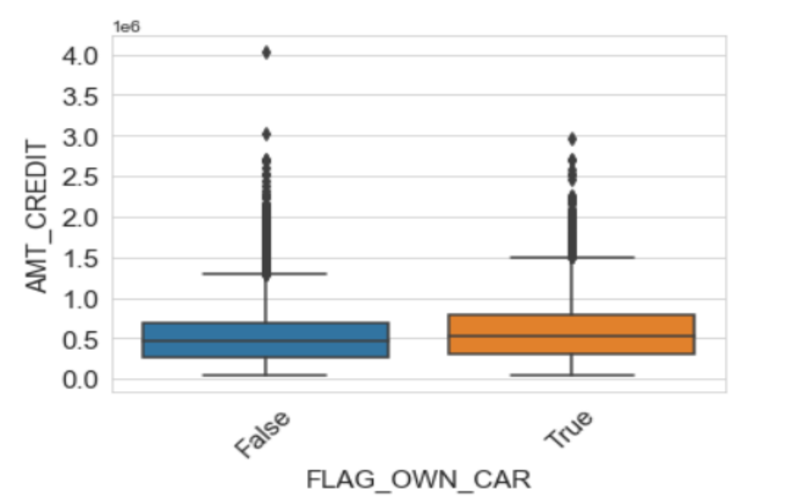
IT CAN BE CONCLUDED FROM ABOVE GRAPHS THAT AMT_CREDIT and AMT_ANNUITTY are highly correlated

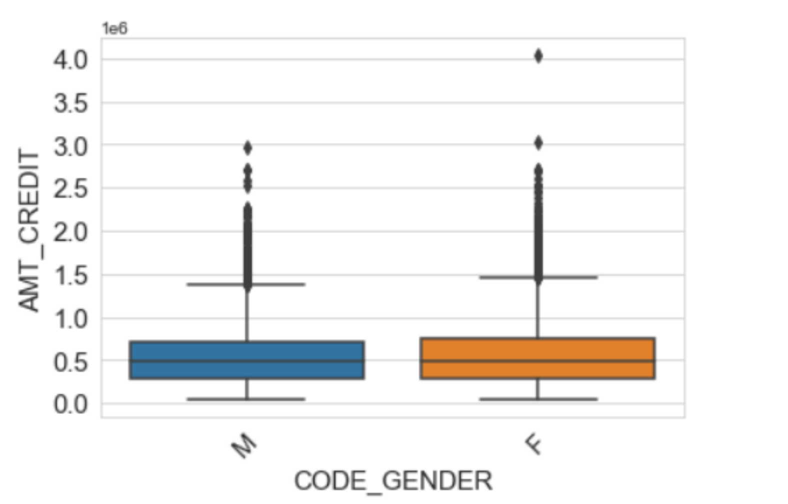
NUMERICAL - CATEGORICAL ANALYSIS(DEFAULTERS)

AMT_CREDIT v/s CATAGORICAL COLUMNS



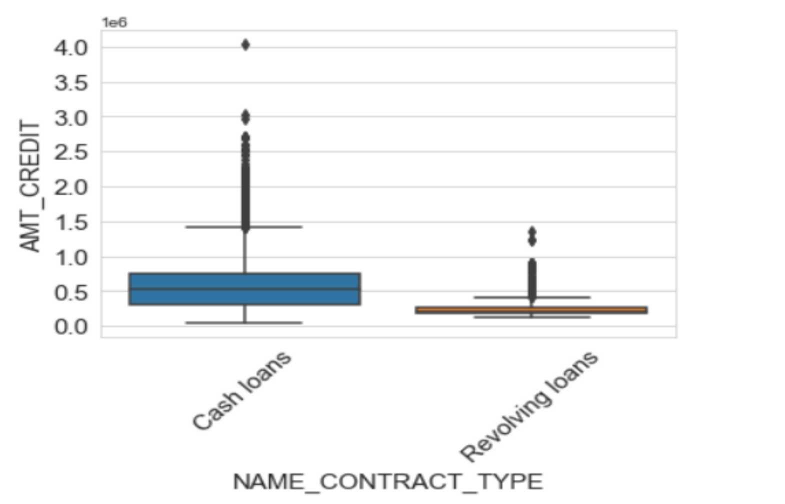


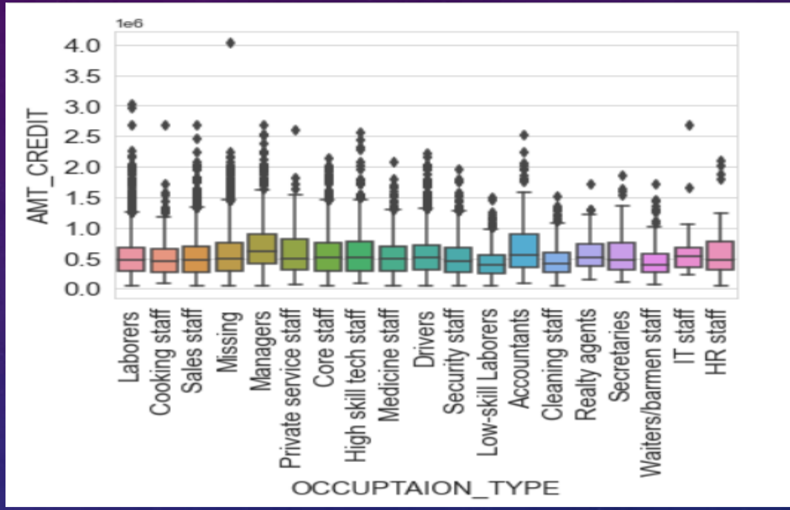


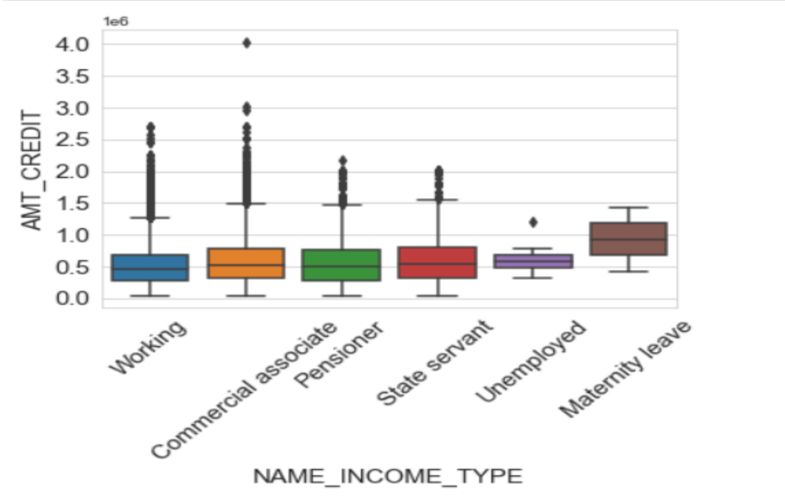


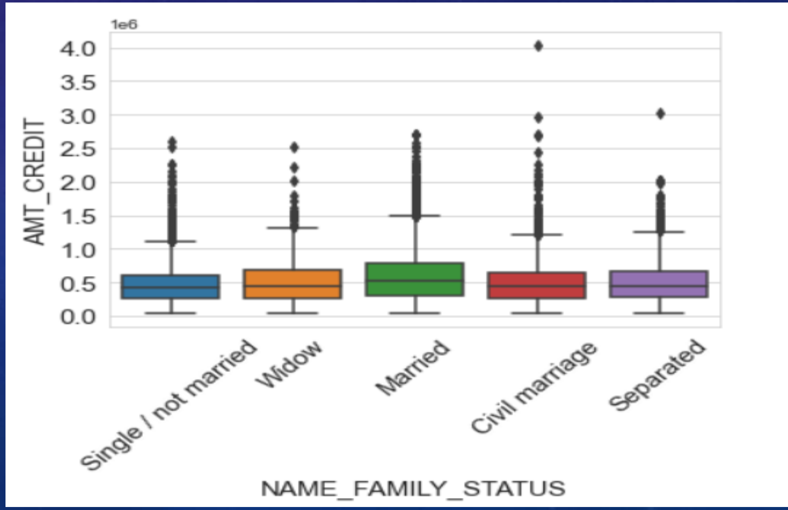
NUMERICAL - CATEGORICAL ANALYSIS(DEFAULTERS)

AMT_CREDIT v/s CATAGORICAL COLUMNS









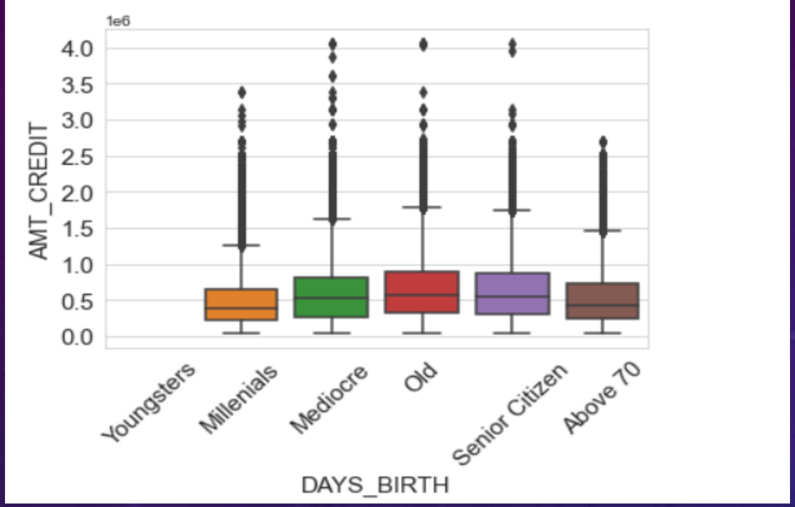
NUMERICAL – CATEGORICAL ANALYSIS(DEFAULTERS)

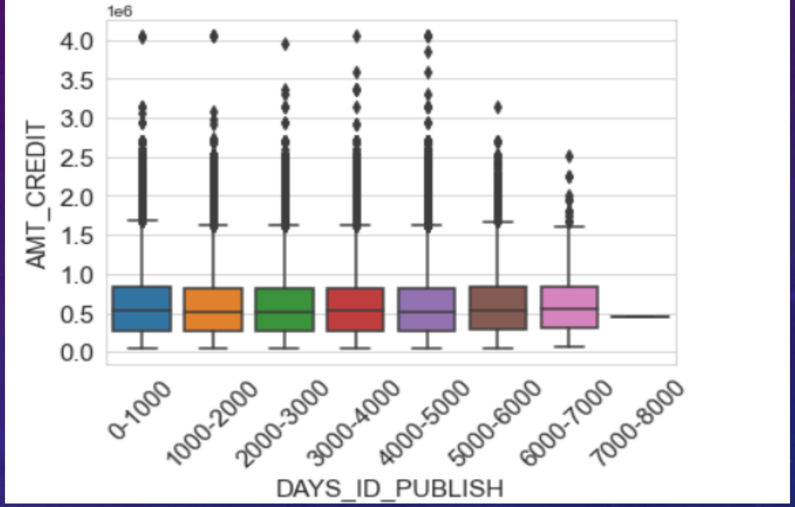
AMT_CREDIT v/s CATAGORICAL COLUMNS

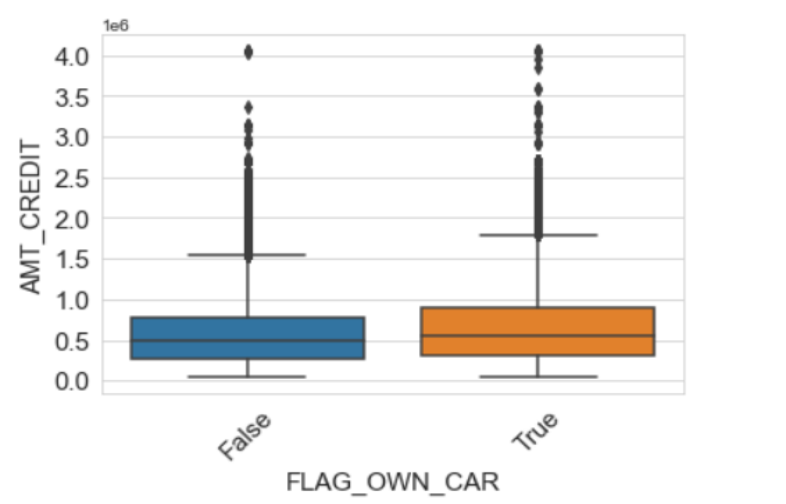
- People between 40-60 years of age are getting most amount credited.
- Amount of loan credited is consistent over the period of time of ID creation days.
- Females are getting slightly more amount credited than males
- Managers and accountants are getting most amount credited to their accounts
- Applicants who are on maternity leave are getting more amount loans
- Married people are getting more amount compared to others.

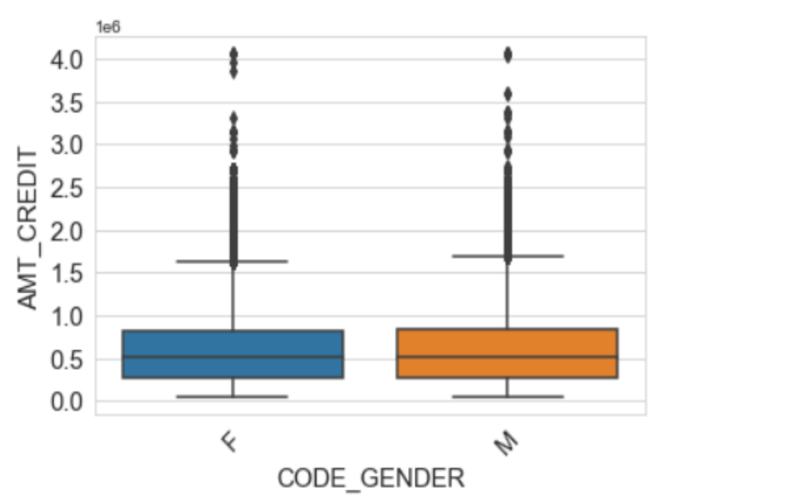
NUMERICAL - CATEGORICAL ANALYSIS(NON-DEFAULTERS)

AMT_CREDIT V/S CATAGORICAL COLUMNS



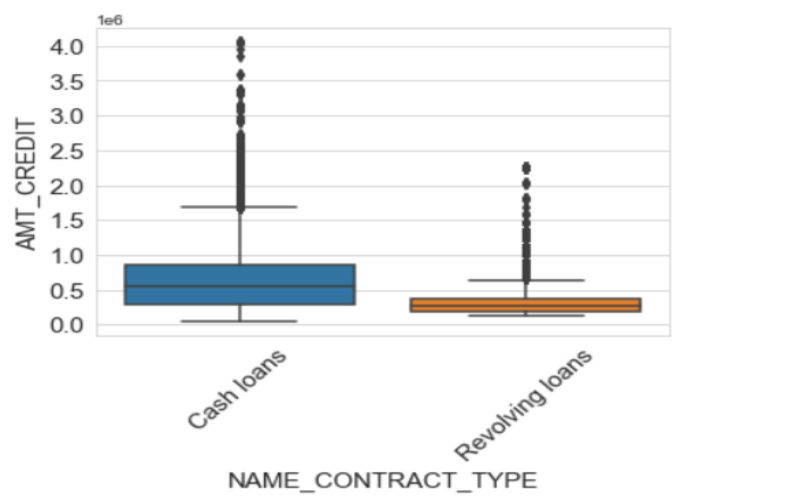


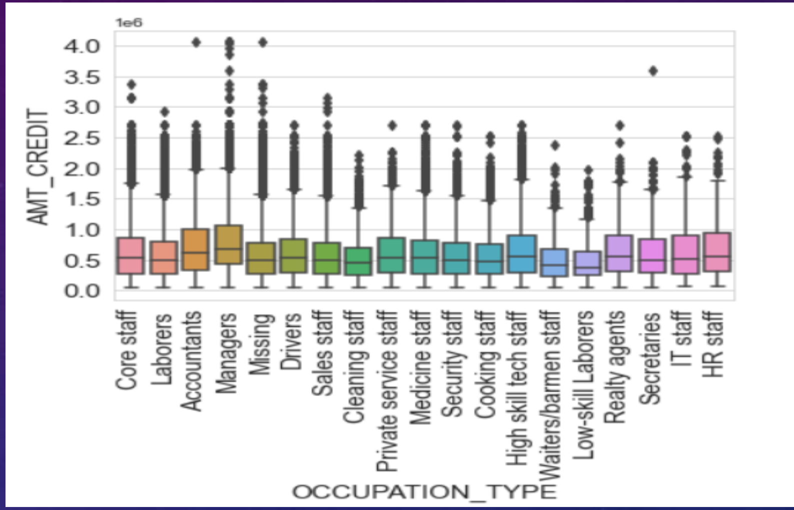


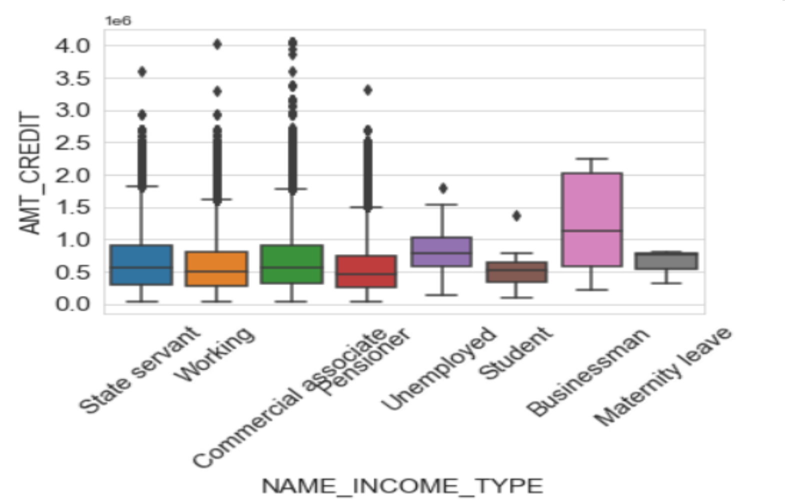


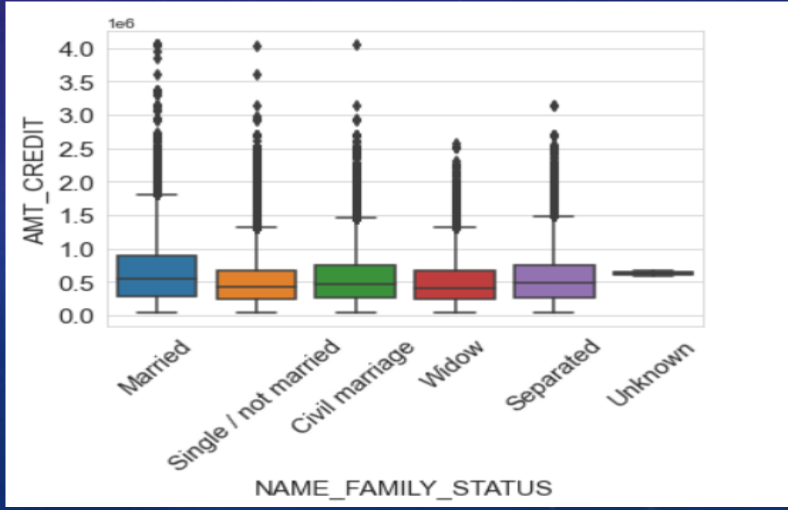
NUMERICAL - CATEGORICAL ANALYSIS(NON-DEFAULTERS)

AMT_CREDIT V/S CATAGORICAL COLUMNS









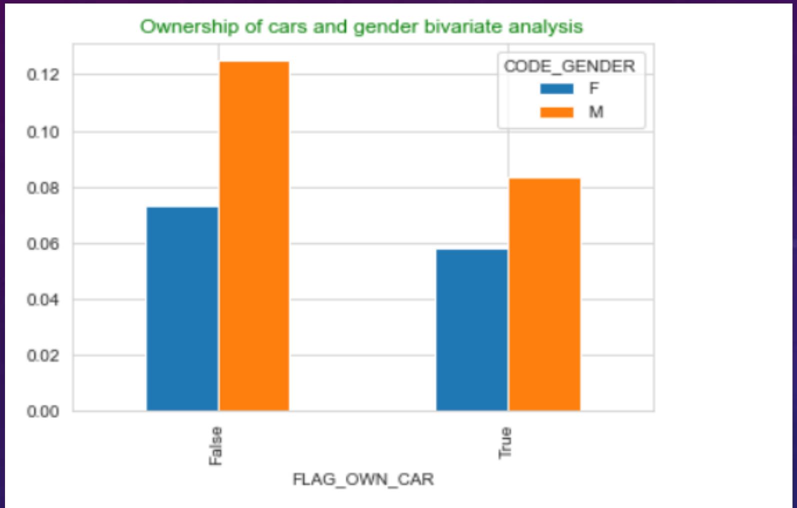
NUMERICAL – CATEGORICAL ANALYSIS(NON-DEFAULTERS)

AMT_CREDIT v/s CATAGORICAL COLUMNS

- People between 40-60 years of age are getting most amount credited.
- Amount of loan credited is consistent over the period of time of ID creation days.
- Females are getting slightly more amount credited than males
- Managers and accountants are getting most amount credited to their accounts
- Applicants who are on maternity leave are getting more amount loans
- Married people are getting more amount compared to others.

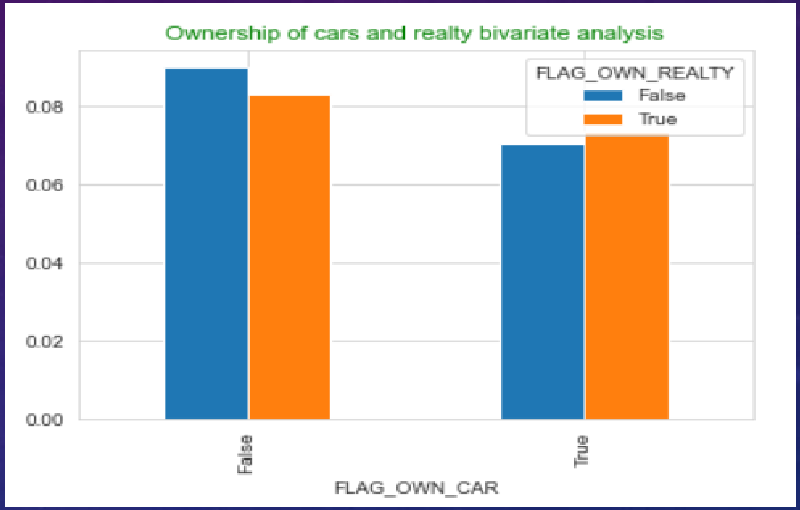
BIVARIATE CATEGORICAL-CATEGORICAL COLUMNS

Gender v/s car ownership bivariate analysis



We can see that number of males are higher in both owning car and not owning car

Realty ownership v/s car ownership bivariate analysis



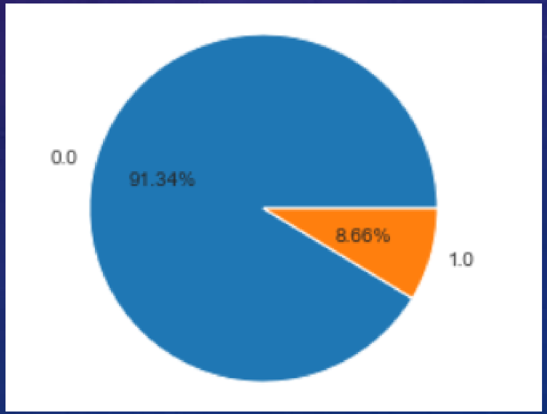
Most number of applicants don't have cars and realty both
Applicants with cars also have more chances of having realty or homes

IMPORTING PREVIOUS APPLICATION DATASET.

- Check the properties of structure of the previous application dataset.
- In total dataset is having 37 columns
- Remove columns having greater than 45% of values as null
- Replace the ‘XNA’ and ‘XMA’ in dataset with NaN so our analysis is not affected.
- Check for the outliers in columns and treat them by capping technique.
- Define a new subset of application_data which is needed to be merged to previous applications dataset.
- Left merge both the dataset on common unique value ‘SK_ID_CURR’ and name dataset as collective_df

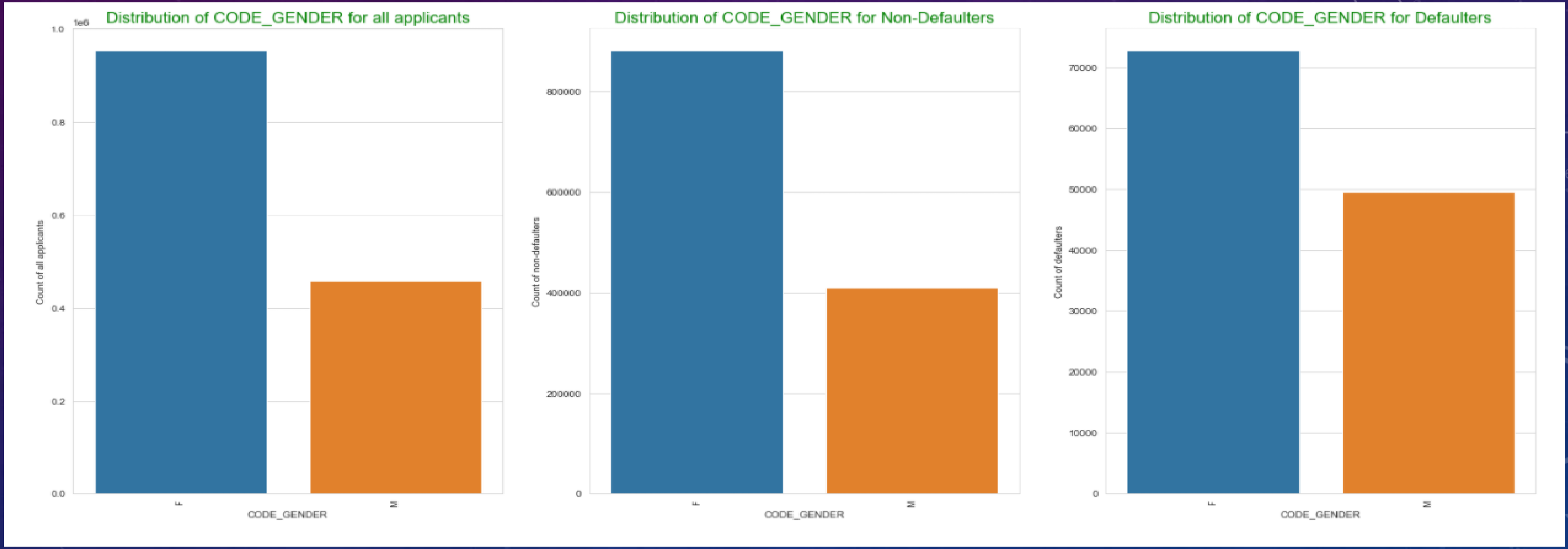
TREATING THE COLLECTIVE DATASET

- Remove the rows from the collective dataframe where values of Target column is null
- Check the imbalance percentage of the new dataset
- Divide the collective dataframe in two dataframe each for target : 0 and target : 1



UNIVARIATE ANALYSIS IN FINAL COLLECTIVE DATAFRAME

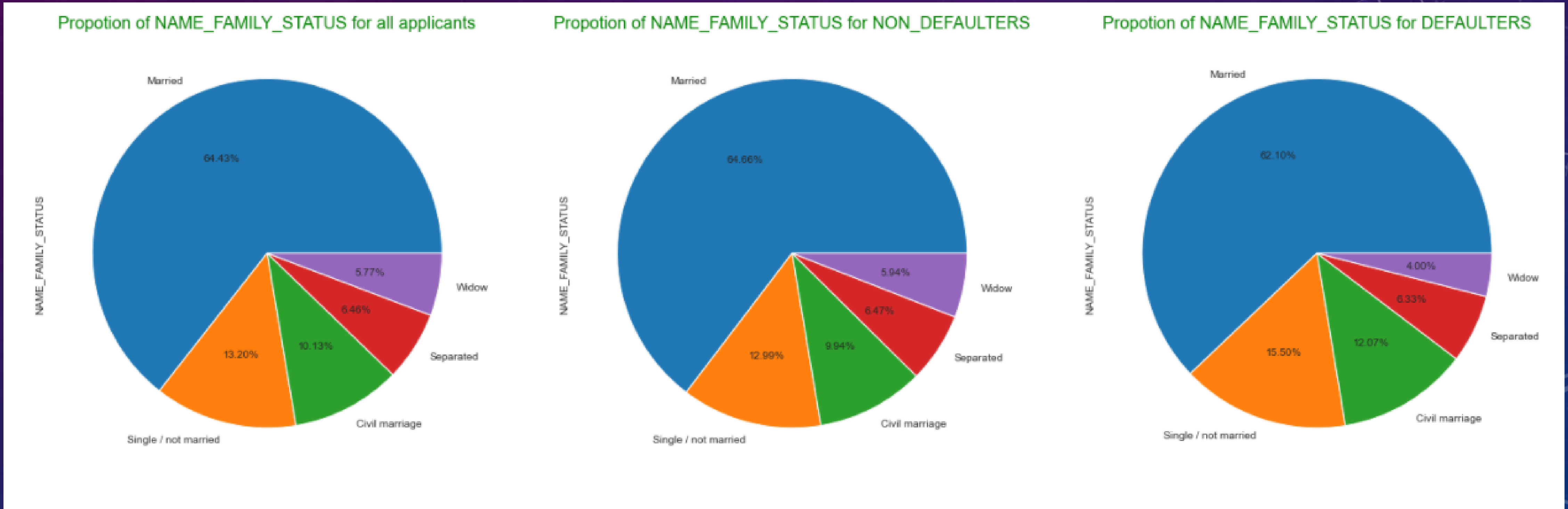
UNIVARIATE ANALYSIS OF THE GENDER COLUMN IN THE MERGED DATASET



We can see that number of females are still higher in case of applying loan and become defaulter

UNIVARIATE ANALYSIS IN FINAL COLLECTIVE DATAFRAME

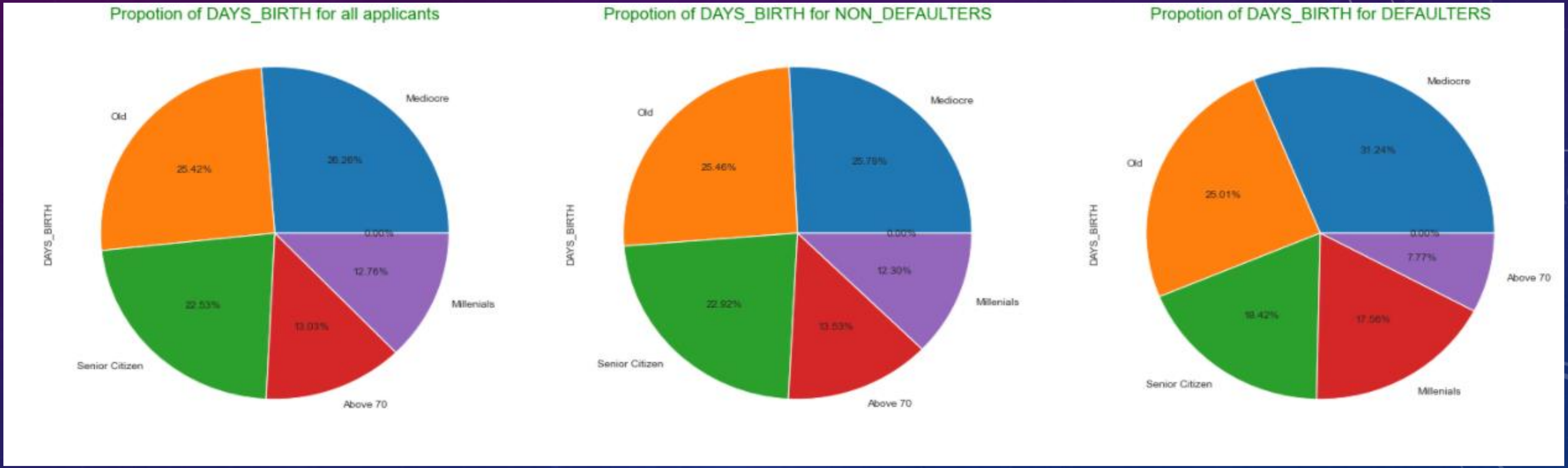
UNIVARIATE ANALYSIS OF THE NAME_FAMILY_STATUS COLUMN IN THE MERGED DATASET



1. We can see that most of the people applying for loans and become defaulters are married
2. Widows are less likely to become defaulter
3. No. of people single or non-married are more leaning towards becoming defaulters

UNIVARIATE ANALYSIS IN FINAL COLLECTIVE DATAFRAME

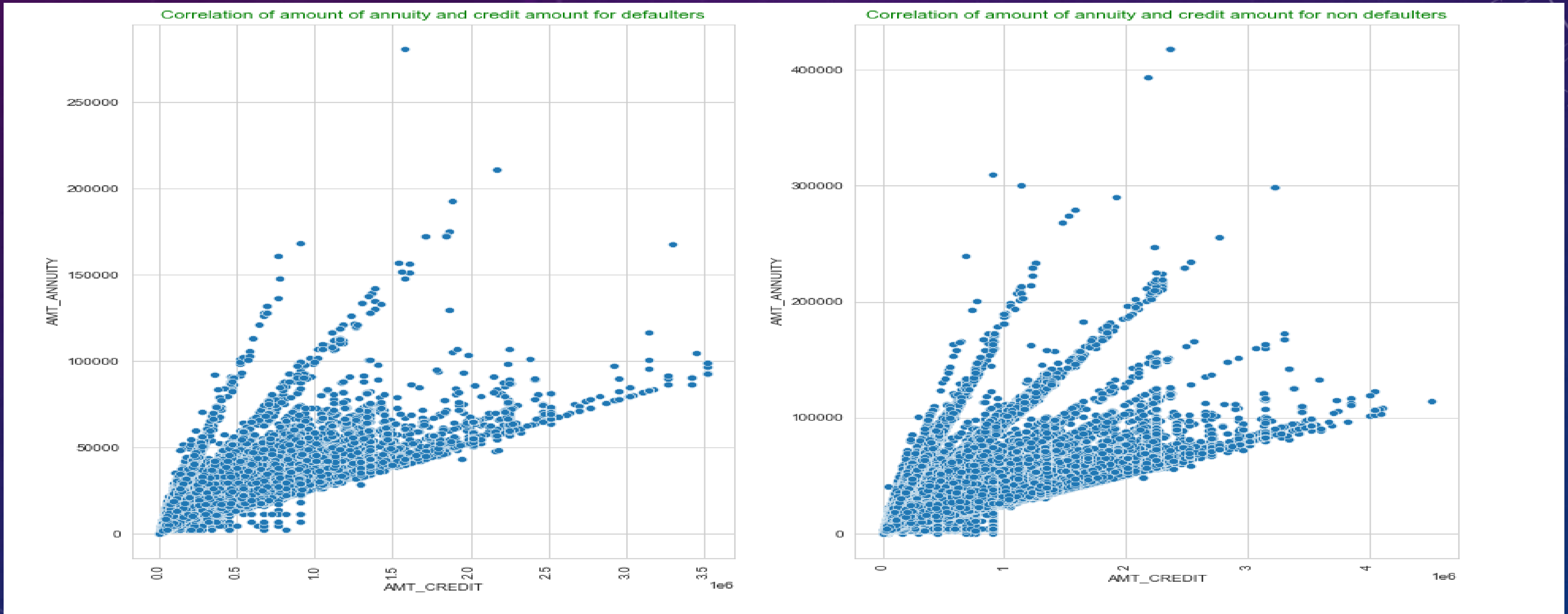
UNIVARIATE ANALYSIS OF THE DAYS_BIRTH COLUMN IN THE MERGED DATASET



1.Applicants above 70 years of age are most like to be non-defaulters
2.Applicatns between age 40-70 years of age i.e. Old, Senior Citizens, and above 70 have comparatively less chances of becoming defaulter
3.Youngsters below 20 years of age are not applying for loans

NUMERIC-NUMERIC BIVARIATE ANALYSIS

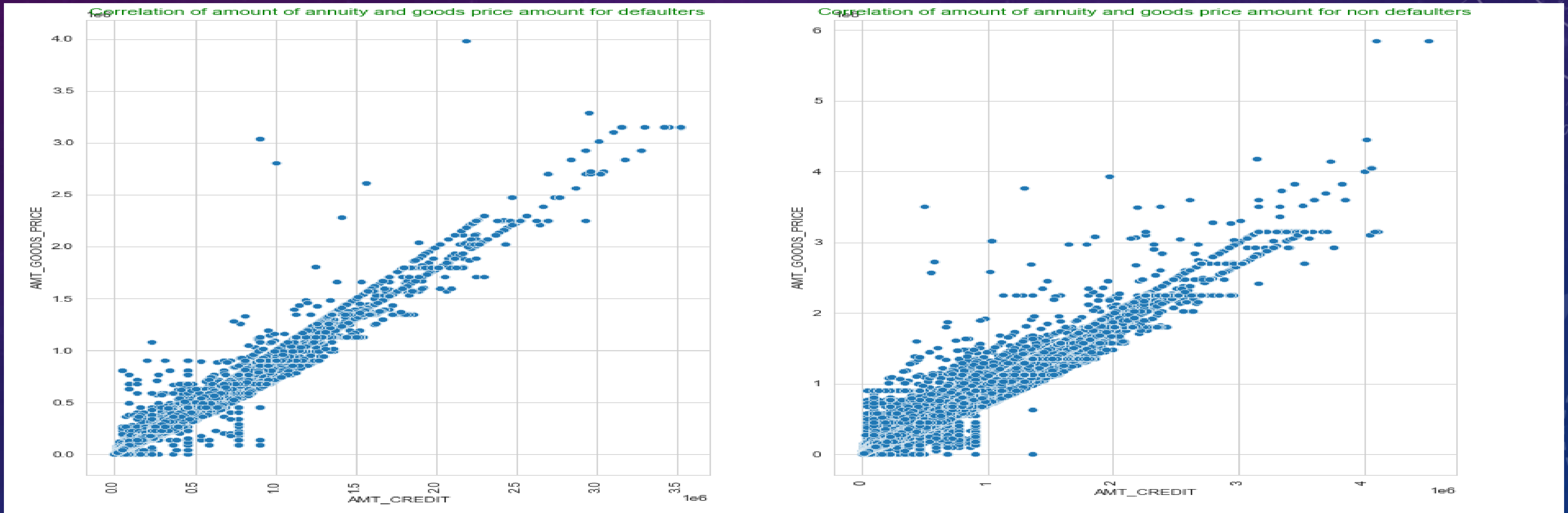
BIVARIATE ANALYSIS OF AMT_CREDIT/AMT_ANNUITY



- It cannot be said that AMT_CREDIT and AMT_ANNUITY are very well lineraly correlated because as the value increase graph gets more scattered.

NUMERIC-NUMERIC BIVARIATE ANALYSIS

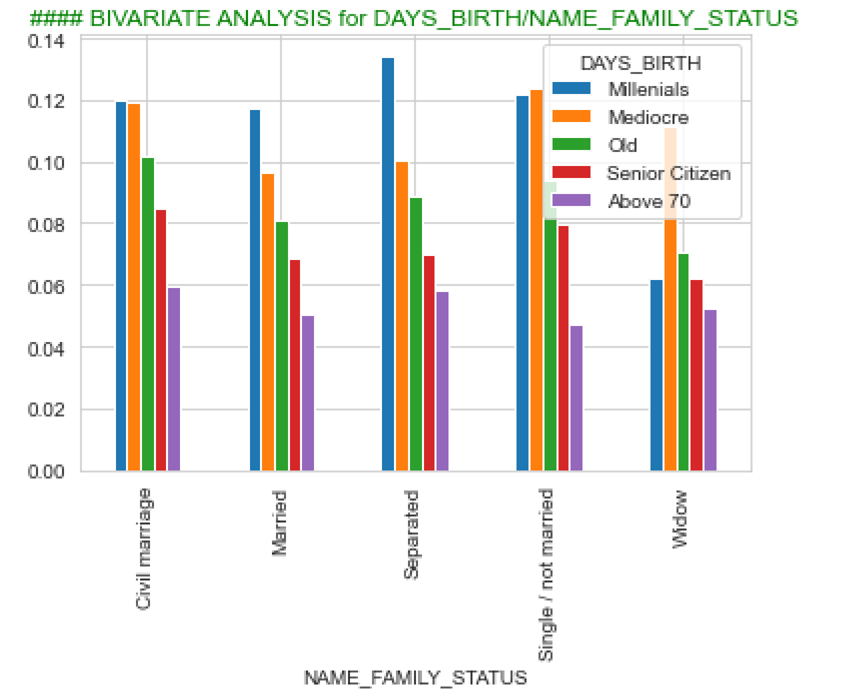
BIVARIATE ANALYSIS OF AMT_CREDIT/AMT_GOODS_PRICE



- It can be concluded that AMT_CREDIT and AMT_GOOD_PRICE are very highly correlated to each other as AMT_CREDIT increase AMT_GOODS_PRICE increases

CATEGORICAL-CATEGORICAL BIVARIATE ANALYSIS

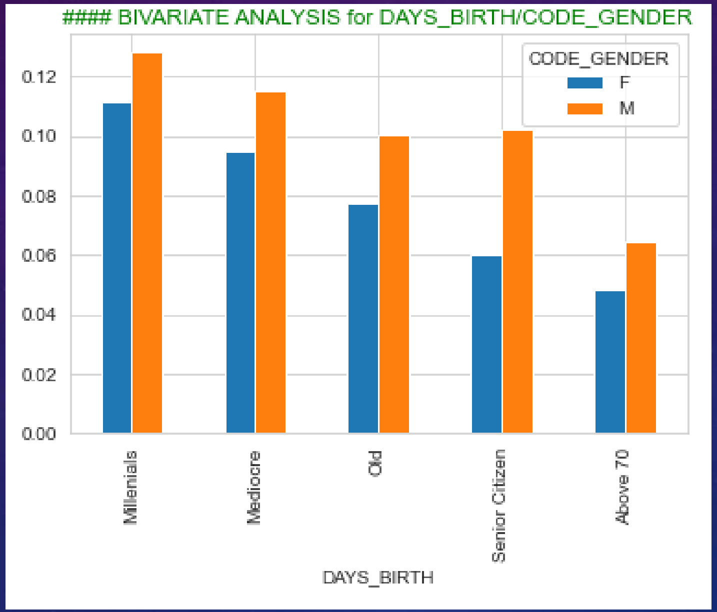
BIVARIATE ANALYSIS FOR DAYS_BIRTH/NAME_FAMILY_STATUS



- It can be concluded that separated millennials are having most chances of becoming defaulters.
- Widow mediocre are also having more chances of becoming defaulters.
- People above 70 years of age have least chances on becoming defaulters.

CATEGORICAL-CATEGORICAL BIVARIATE ANALYSIS

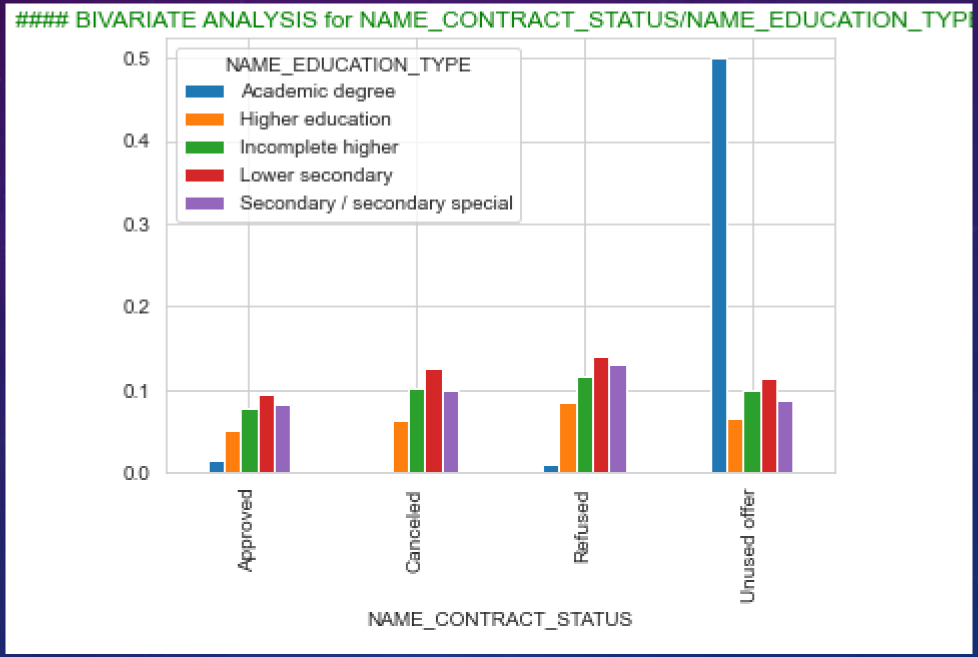
BIVARIATE ANALYSIS FOR DAYS_BIRTH/CODE GENDER



1. From overall dataset it can be concluded that males are more likely to become defaulters which is contradict of application_data dataset
2. Millennial males are having most chances of becoming defaulters
3. Females above 70 years of age having least chances of becoming defaulters

CATEGORICAL-CATEGORICAL BIVARIATE ANALYSIS

BIVARIATE ANALYSIS FOR NAME_CONTRACT_STATUS/NAME_EDUCATION_TYPE



1.People having Academic degrees are mostly using the offers beside having least approved loans.
2.Lower secondary educations are having most approved, canceled and refused loans

FINAL MAIN CONSIDERABLE CONCLUSIONS

1. Married people are taking most loans, and having medium chances of becoming defaulter
2. As the age of applicants increases chances of becoming defaulters reduces
3. Widowers are among the least defaulters
4. Older females have less chances of becoming defaulters
5. People between age 30-40 are most likely to become dafaulters
6. From the overall merged dataset it can be concluded that males are having more chances on becoming defaulters which is opposite of application datasets
7. All amount columns AMT_ANNUITY,AMT_CREDIT,AMT_GOODS_PRICE,AMT_INCOME_TOTAL are highly correlated columns
8. The proption of defaulters in final dataset is 8.66
9. More cash loans are provided by banks and becomes defaulters. Revolving loans should be considered more by banks.
10. Single people have most chances of becoming defaulters, avoid giving loans to them
11. People with higher education, i.e. having more chances of unusing the loans


THANK
YOU