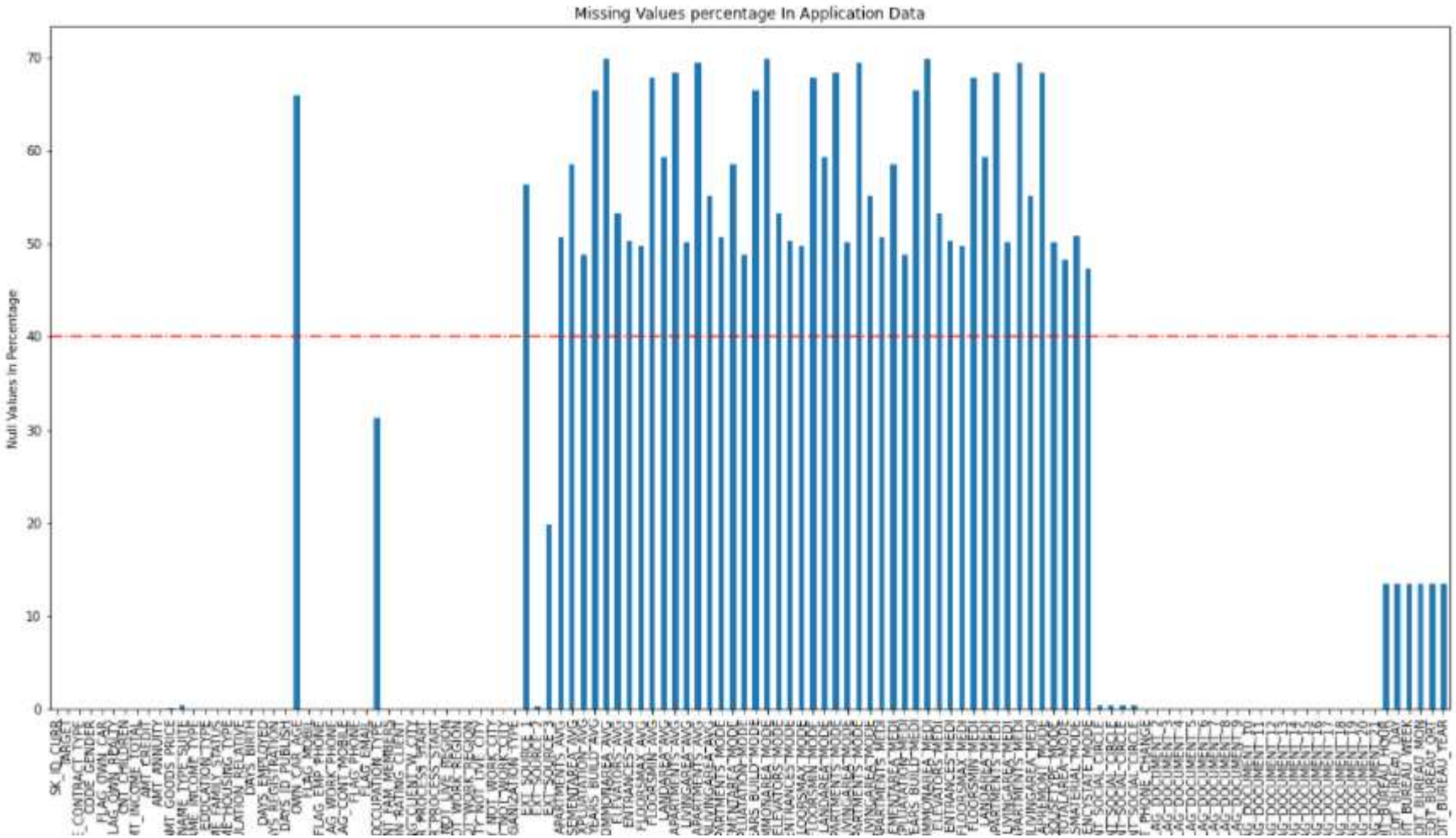
Credit EDA Case Study

by

Harsh Chindarkar and Ketan Kandalkar

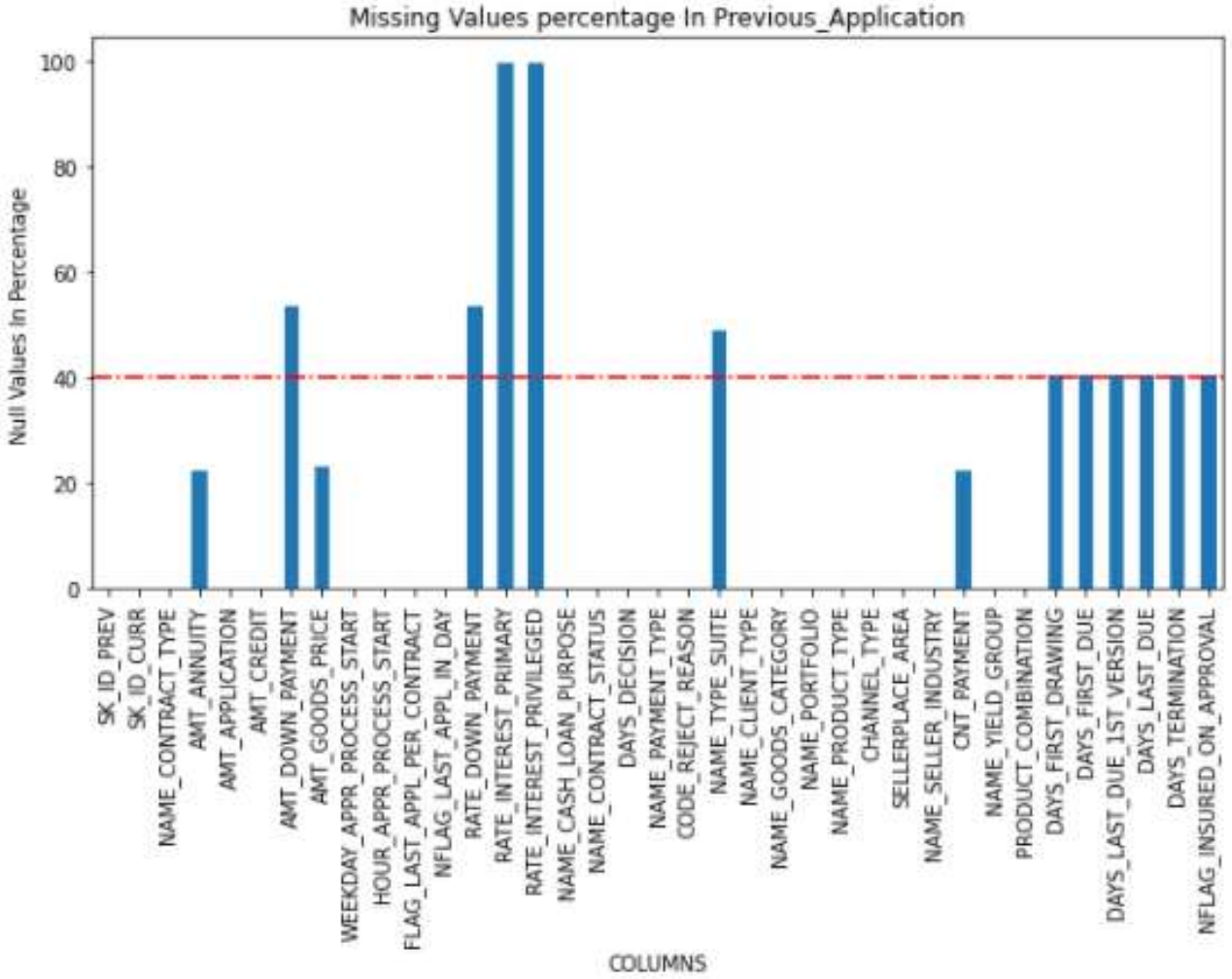
Missing values percentage of application data



INFERENCE

We can see from the plot that the columns above redline mark are the columns with missing values more 40% and the columns with less than 40% missing values are less than redline mark

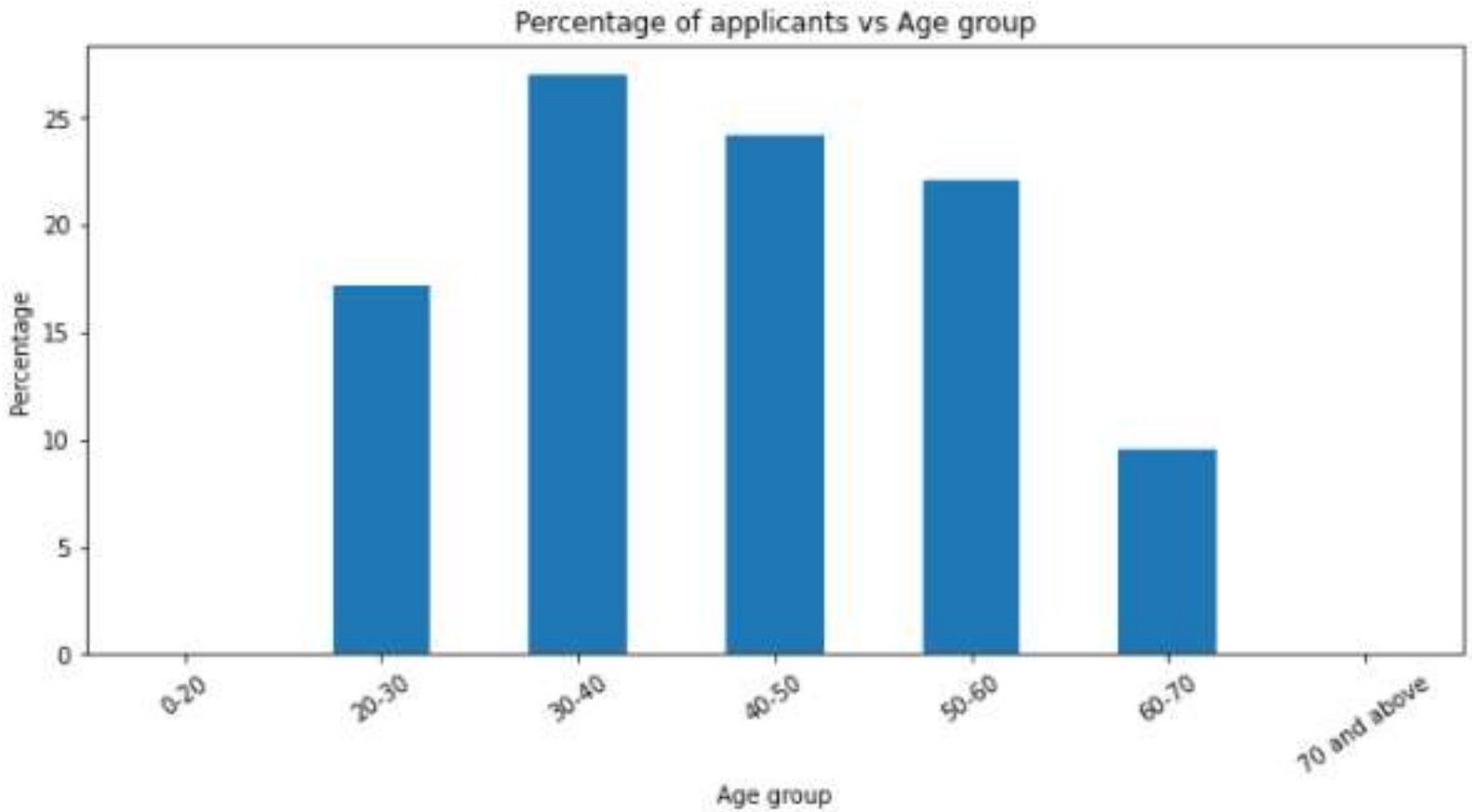
Missing values percentage of previous application data



INFERENCE

Here also as we can see their are many columns in previous application data whose null values are more than 40%

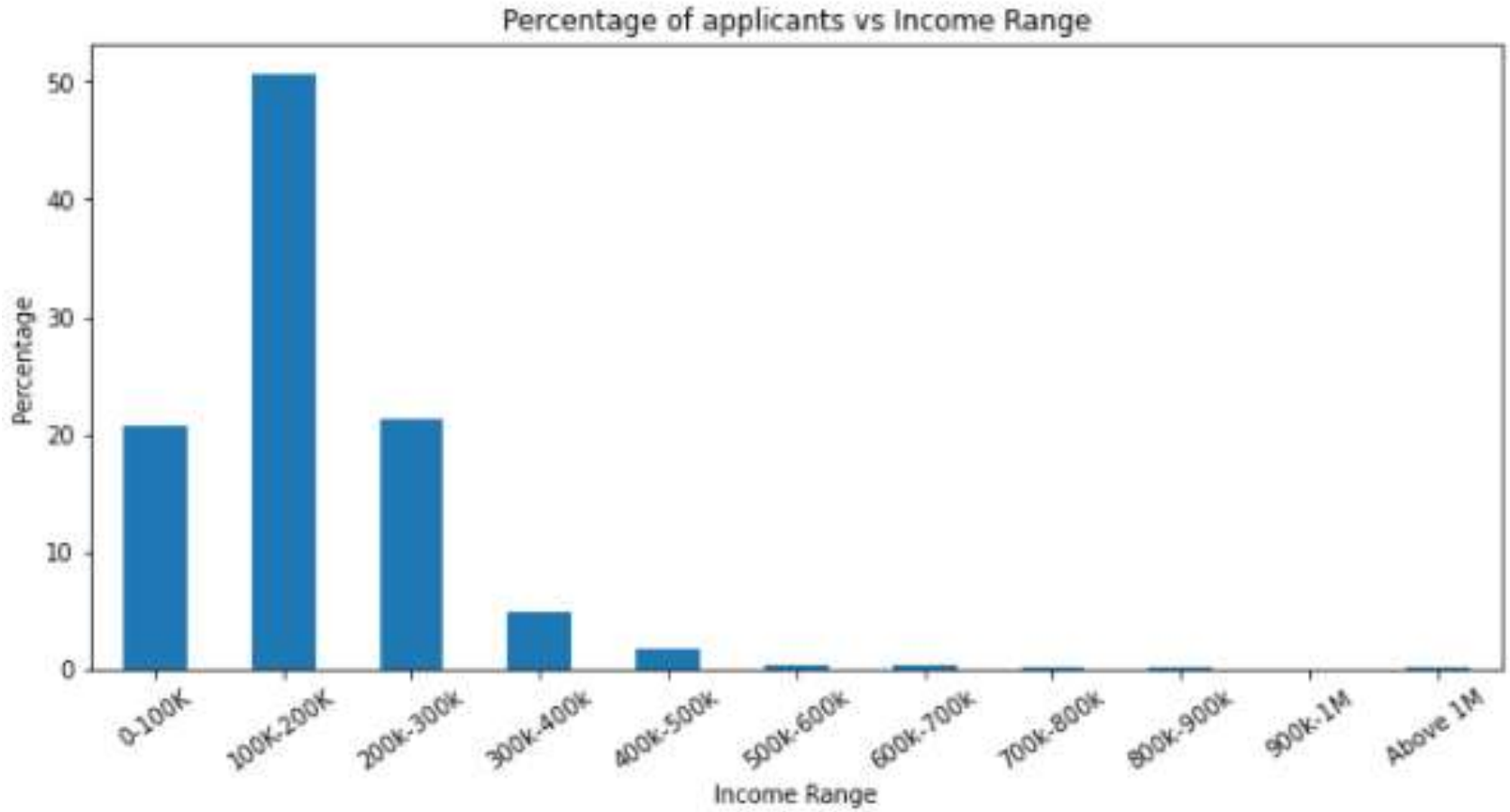
Countplot for Age Group



INFERENCE

More than 50% of the applicants belong to age group of 30 to 50 years

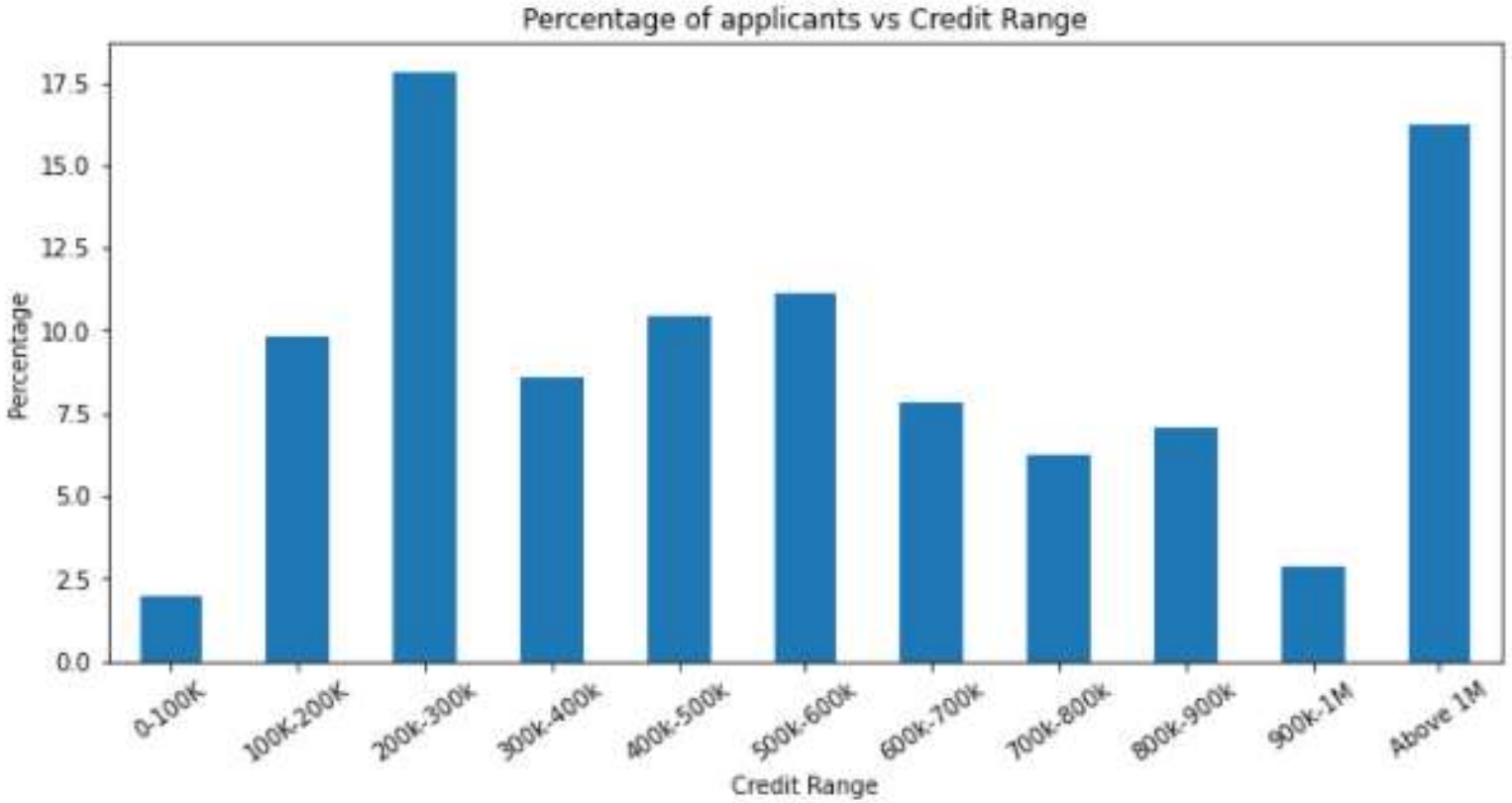
Countplot for Income Range



INFERENCES

1. 50% applicants have income in the range of 100k to 200k
2. Almost 91% applicants have income in the range of 0 to 300K
3. Less than 10& applicants have income more than 300K

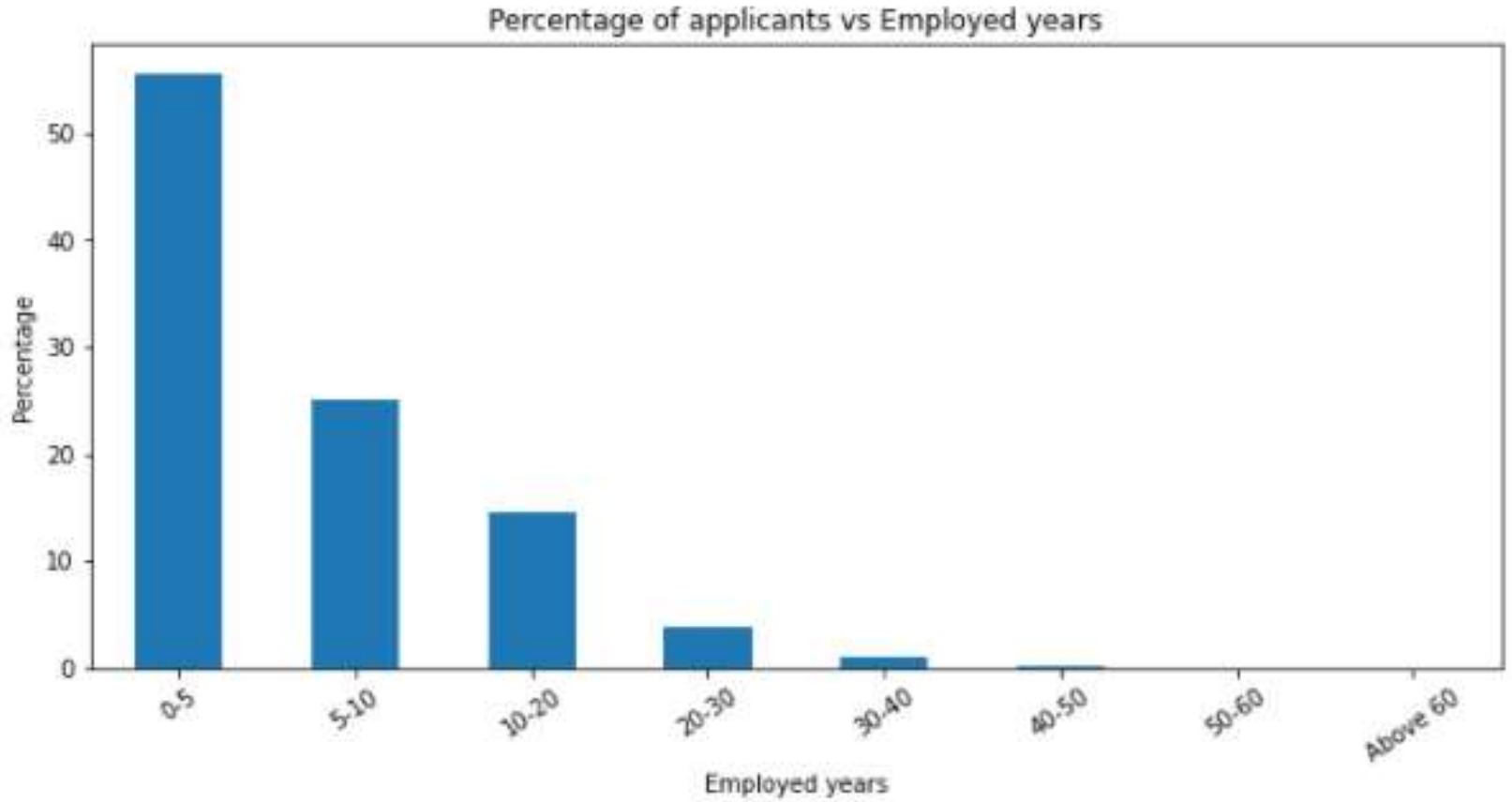
Countplot for Credit Range



INFERENCES

1. Around 16% applicants took loan of above 1M
2. Almost 18% applicants took loan in the range of 200k to 300k

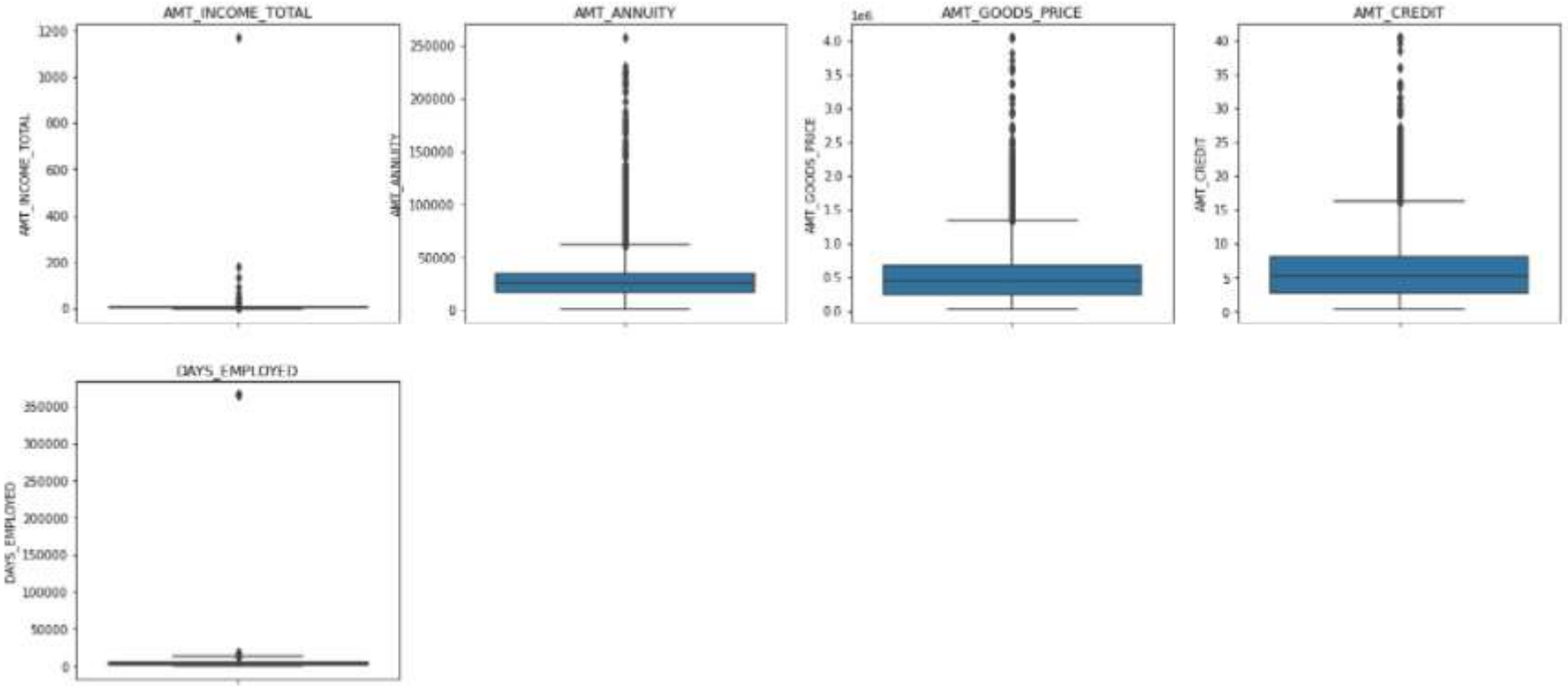
Countplot for Employed years



INFERENCE

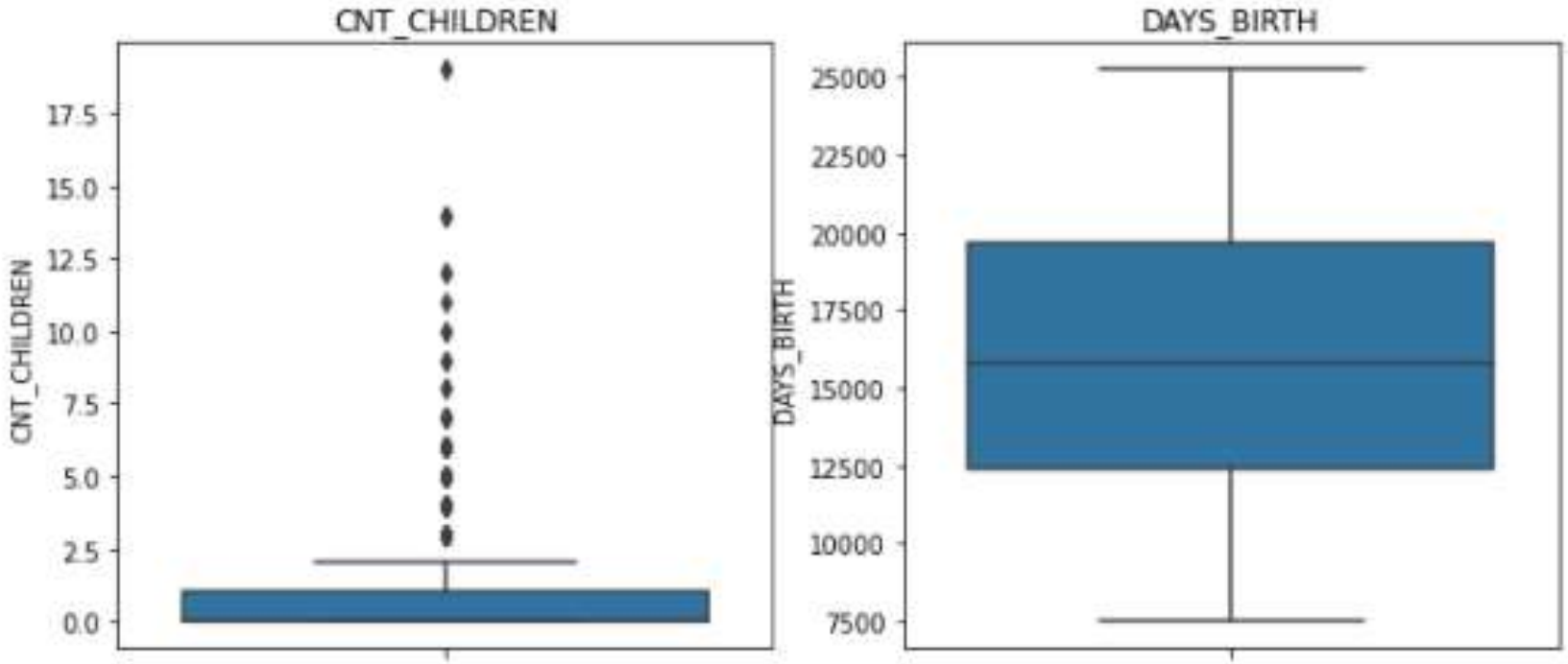
More than 50% applicants are employed for last 5 years whereas almost 80% applicants have 10 years work experience

IDENTIFYING OUTLIERS



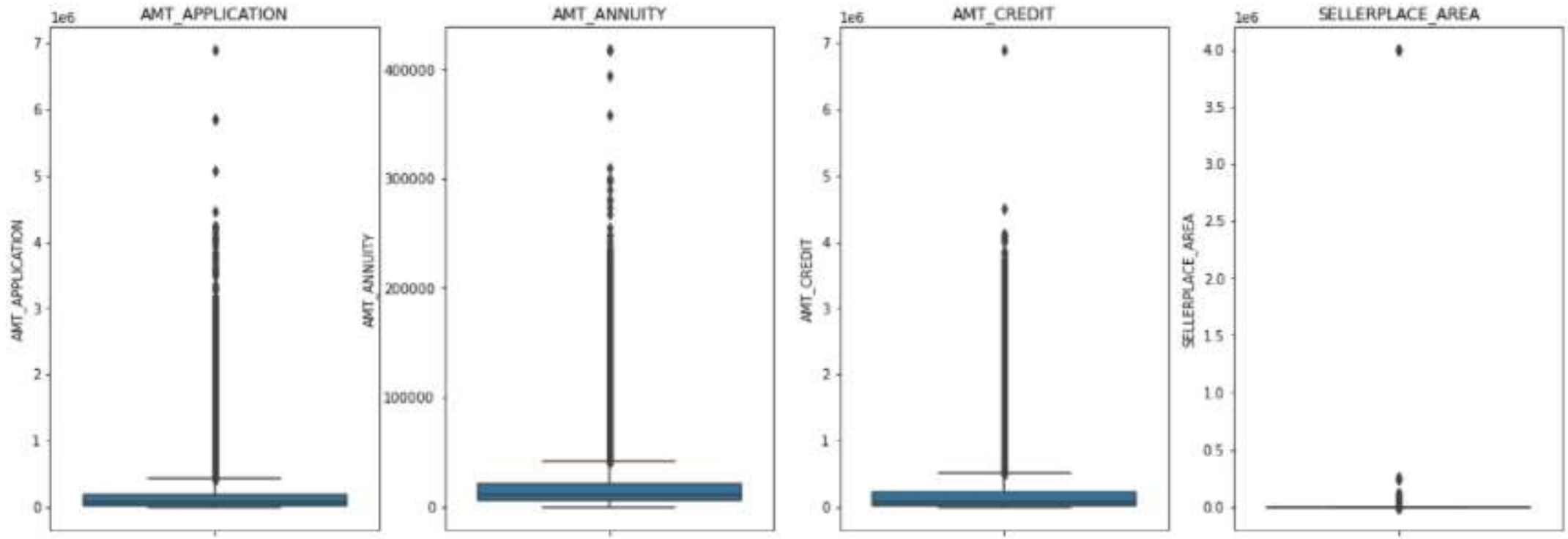
INFERENCES

1. AMT_INCOME_TOTAL has some outliers which shows that there are some applicants with very high income
2. AMT_ANNUITY, AMT_GOODS_PRICE, AMT_CREDIT have very high number of outliers DAYS_EMPLOYED column has less number of outliers with value around 350000 days which is impossible hence it is incorrect entry



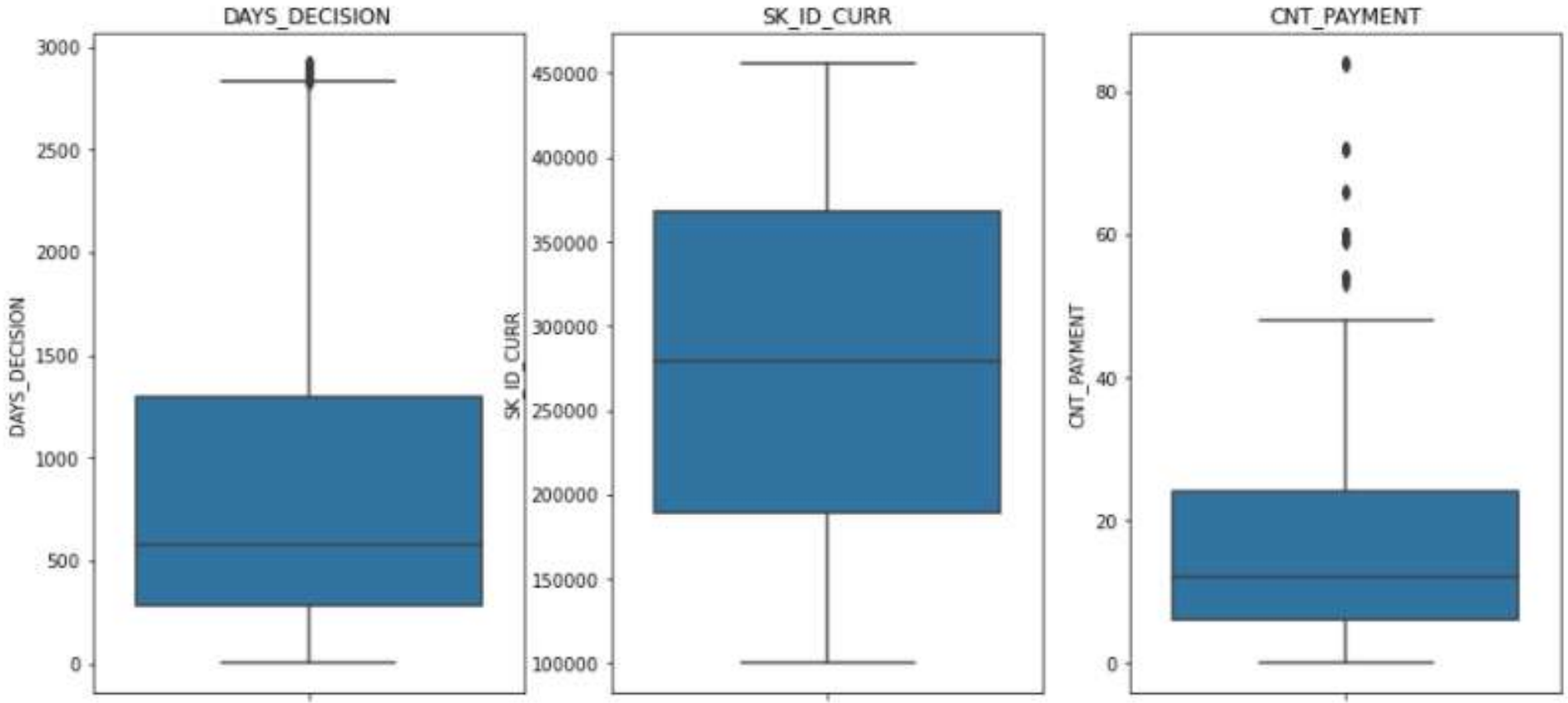
INFERENCES

1. CNT_CHILDREN have few outliers with value almost 20
2. DAY_BIRTH have no outliers which shows that data is uniformly spread in this column



INFERENCE

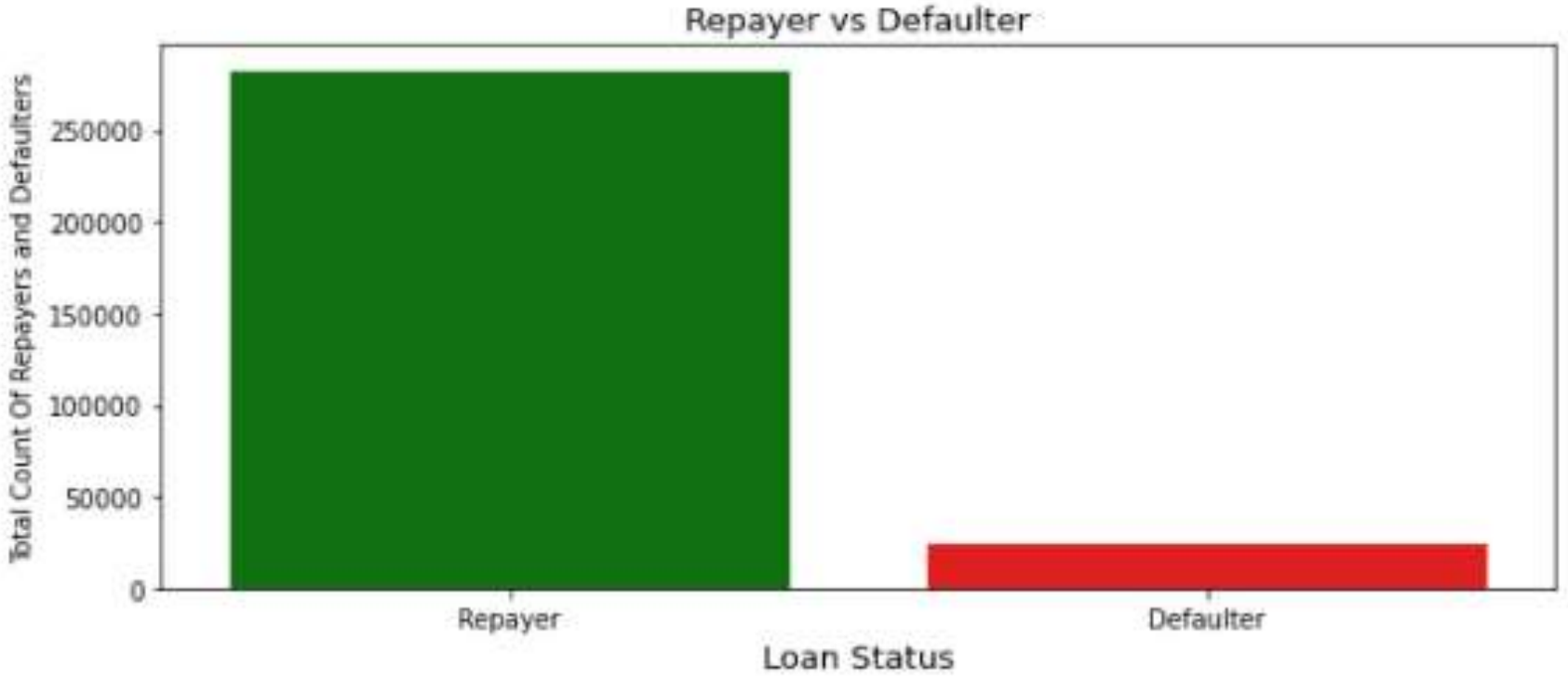
AMT_APPLICATION, AMT_ANNUITY, AMT_CREDIT, SELLERPLACE_AREA and AMT_GOODS_PRICE, have huge number of outliers



INFERENCES

DAYS_DECISION and CNT_PAYMENT have very few outliers
SK_ID_CURR has no outliers since it is an ID column so it is understood

DATA IMBALANCE ANALYSIS

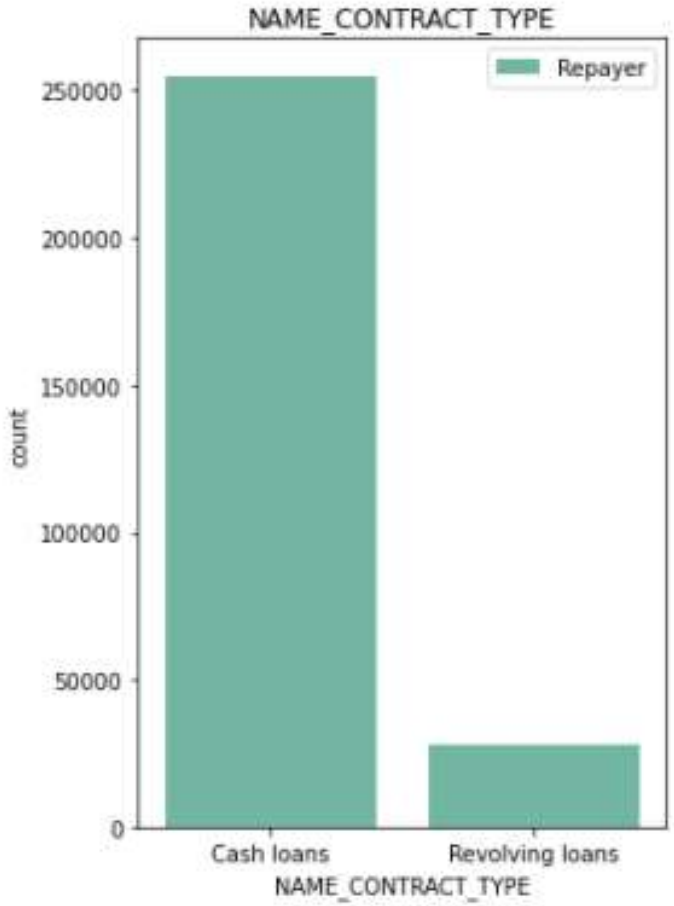


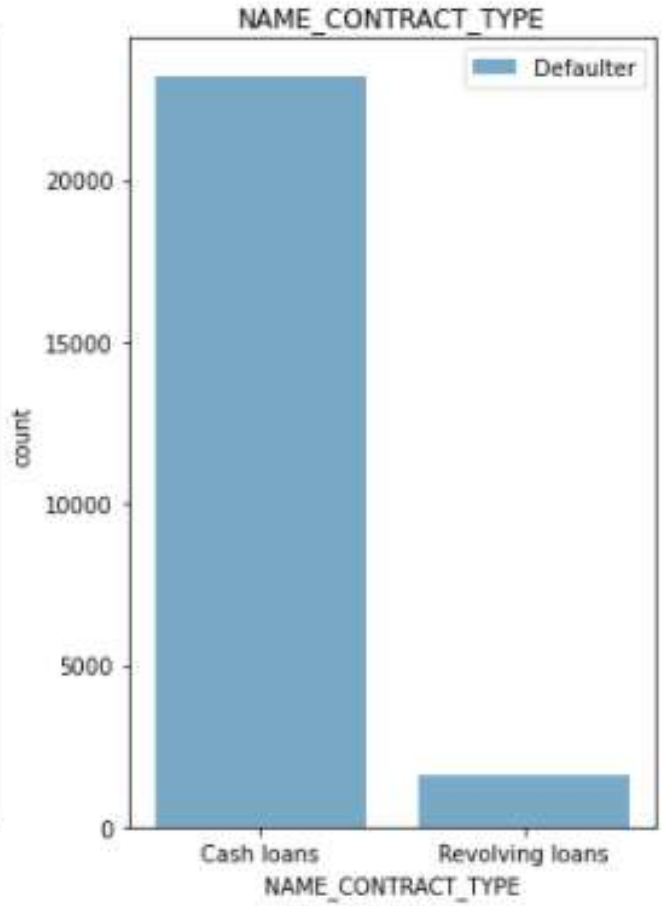
INFERENCE

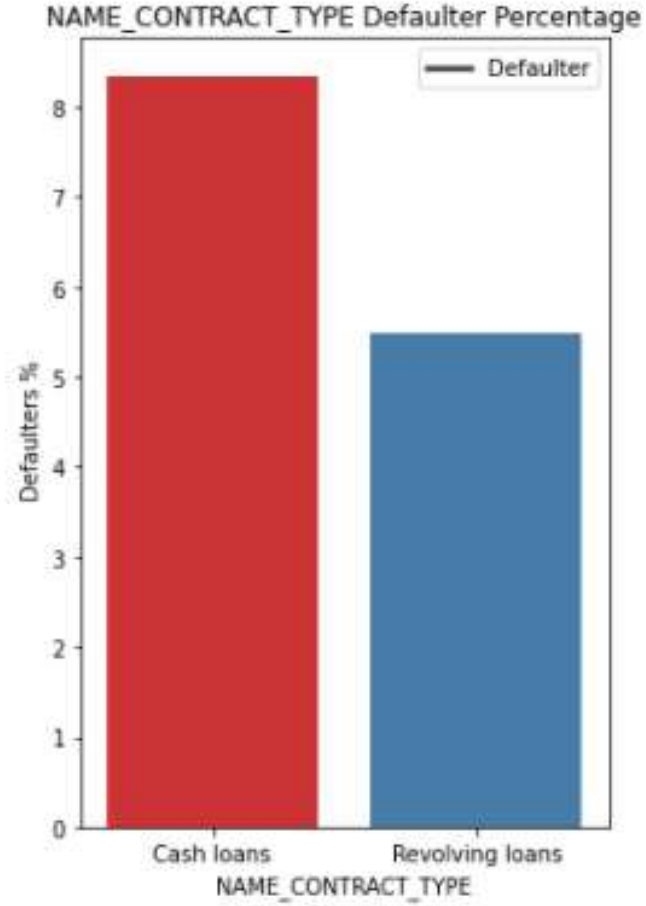
So from Above barplot now we know that the imbalance percentage of Repayer and Defaulter is 92% and 8% respectively.

UNIVARIATE ANALYSIS

NAME_CONTRACT_TYPE



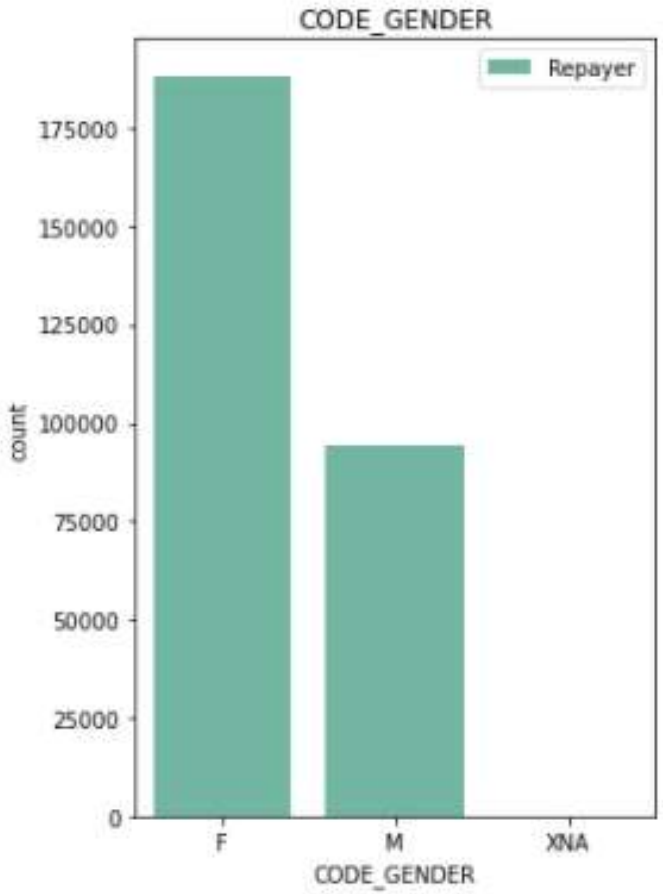


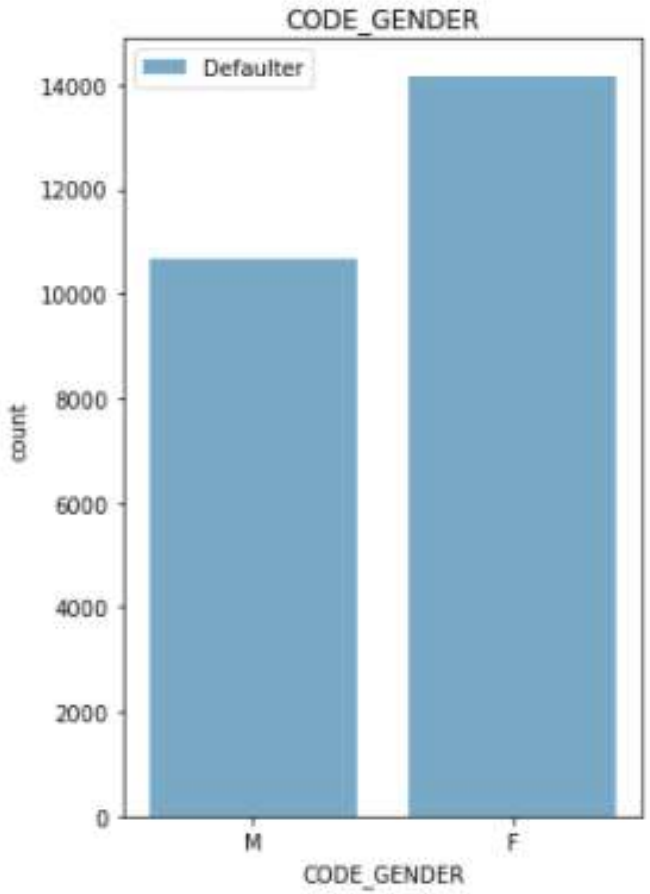


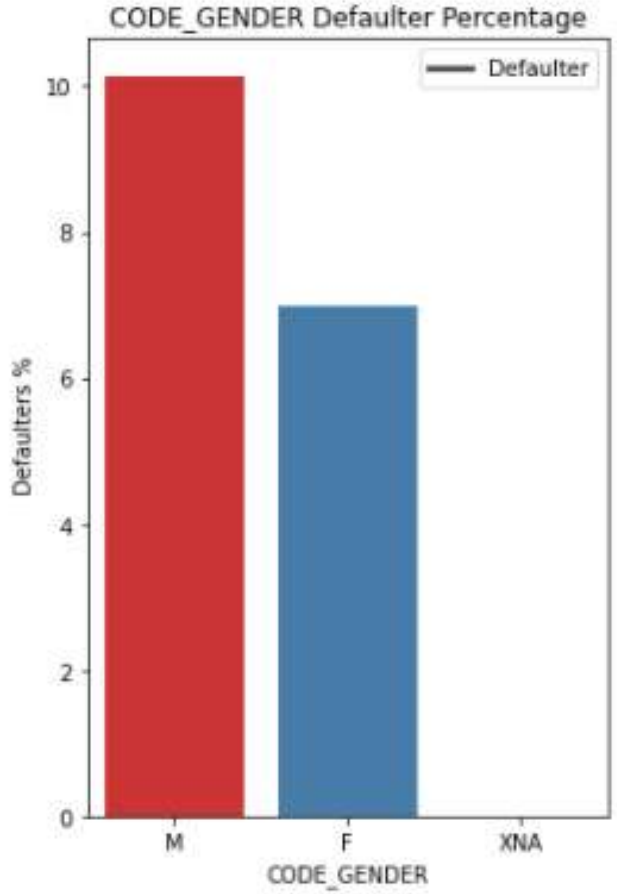
INFERENCES:

1.As we can see their are less amount of Revolving loans but still majority of the loans are not repaid.
2.Large amount of cash loans are seen and also defaulter percenatge is higher for cash loans.

CODE_GENDER



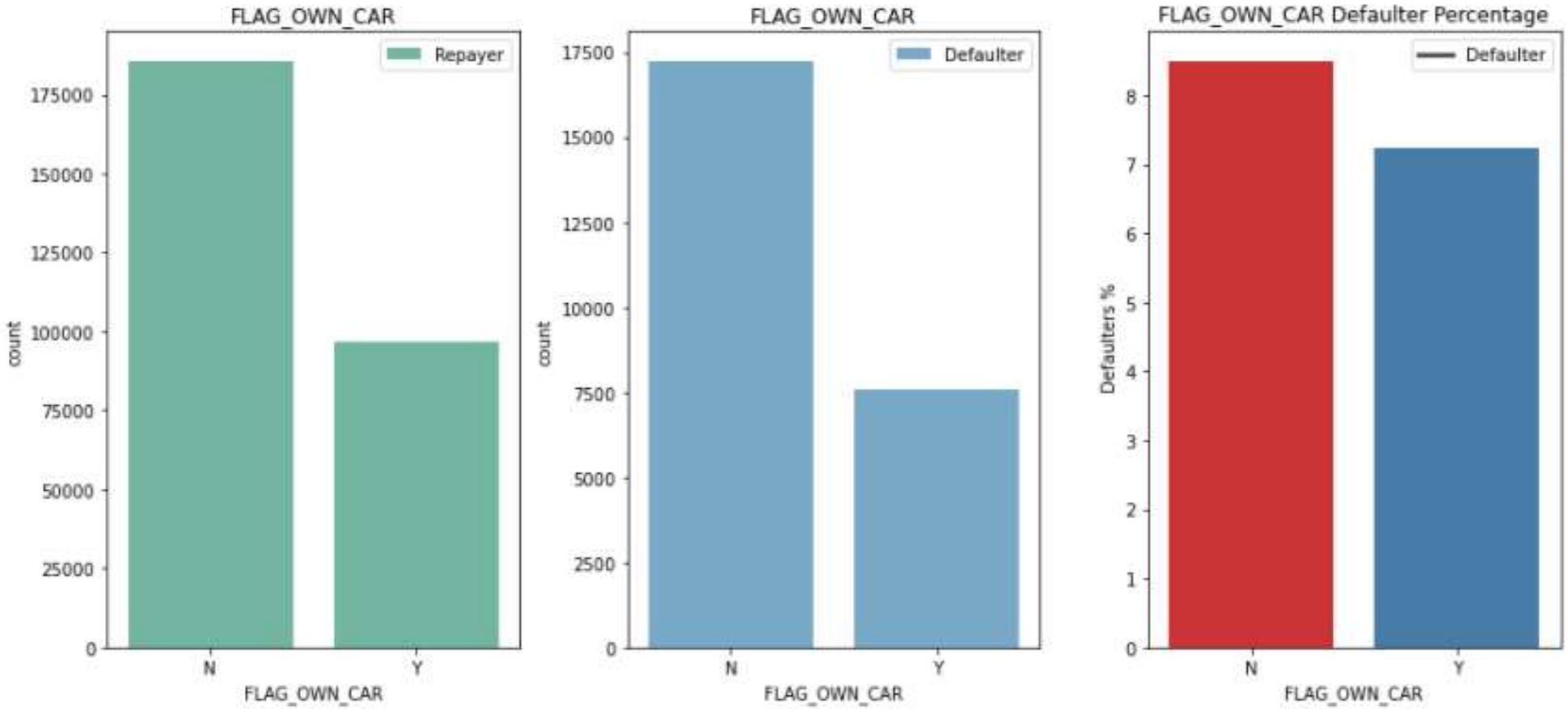




INFERENCES:

1. We can see number of female clients is more than male clients.
2. But in Defaulter % plot we can see that defaulter percentage of male clients (approx-10%) is more than female clients. This means males have less chance of returning their loans.

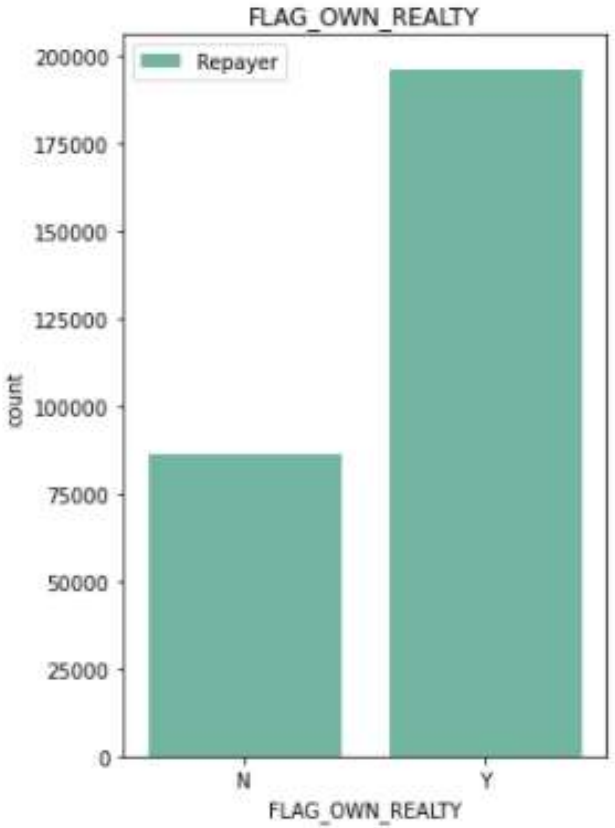
FLAG_OWN_CAR

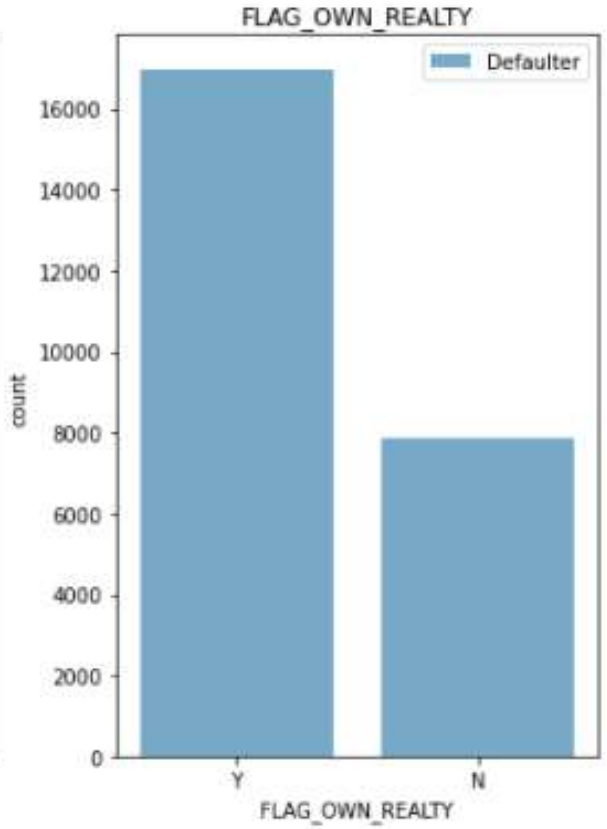


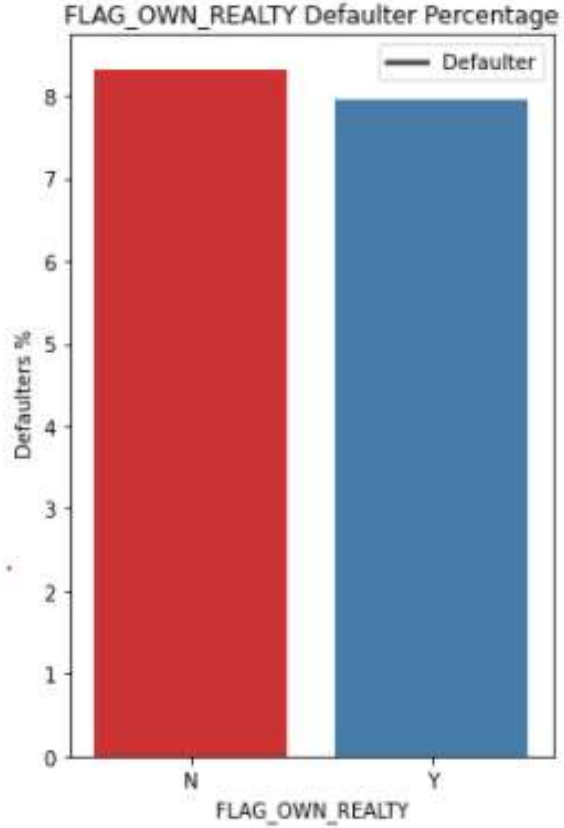
INFERENCES:

1.We can observe that Large Number of clients does not own cars.
2.In Defaulter plot client who does not own car have slightly high chances of not repaying the loan than client who owns the car but still we cant say theirs coorelation bewteen them as they both have almost same percentage of defaulters.

FLAG_OWN_REALTY



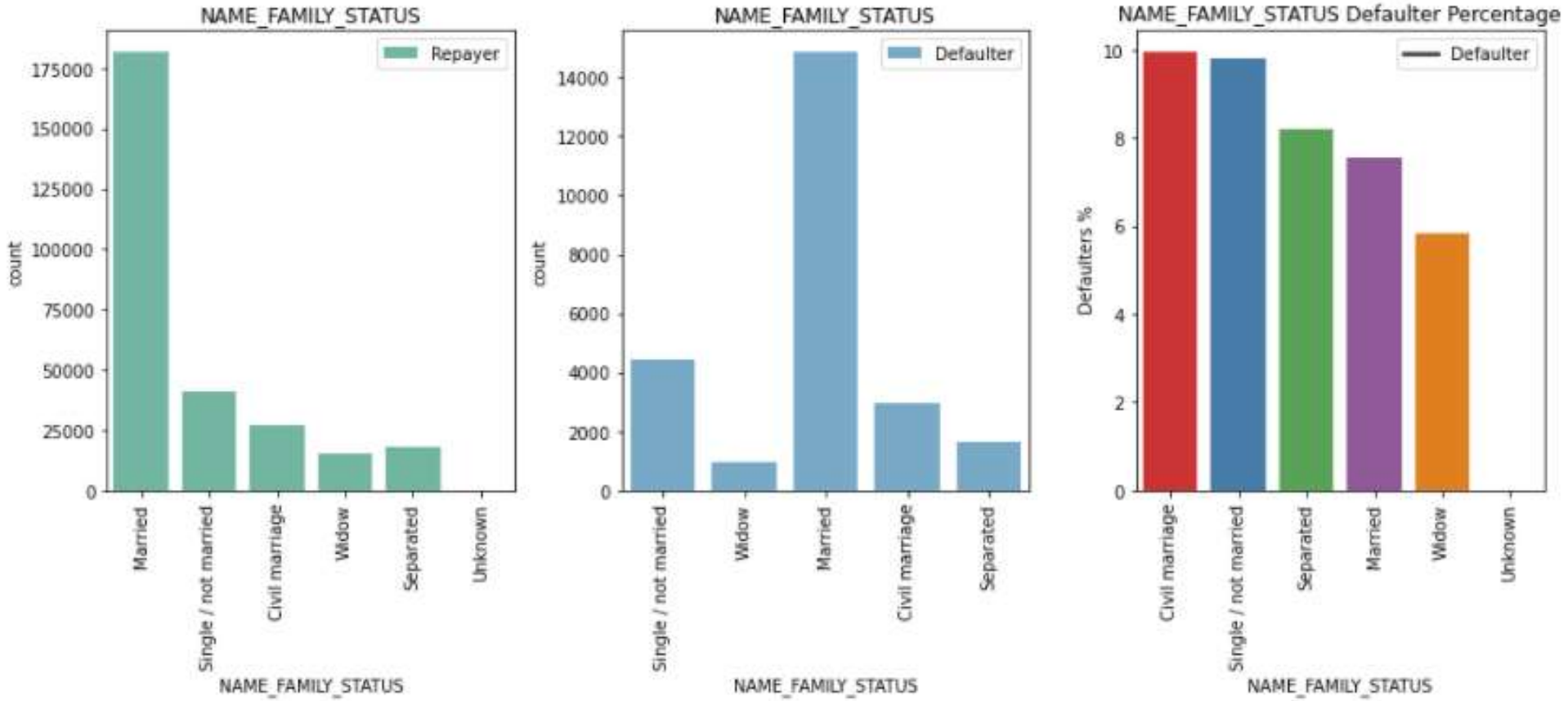




INFERENCES:

1. Here we can see clients who own real estate is more than half of the clients who don't own real estate. This means that majority of the clients own real estate.
2. Their is no correlation seen between the client owning real estate and defaulter of loan as their percentage of defaulters is almost same.

NAME_FAMILY_STATUS



INFERENCES:

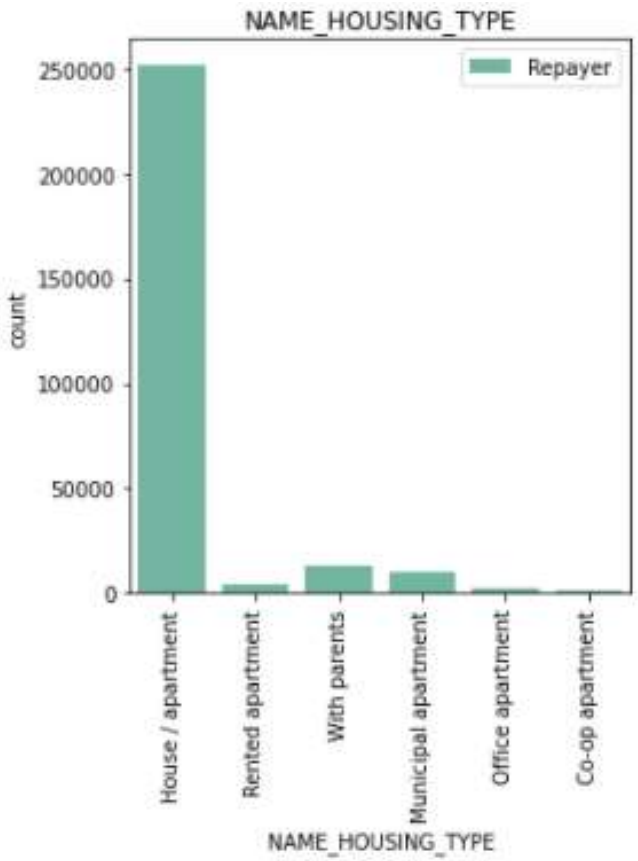
1.Majority of the loans are take by married clients as seen in the plot.

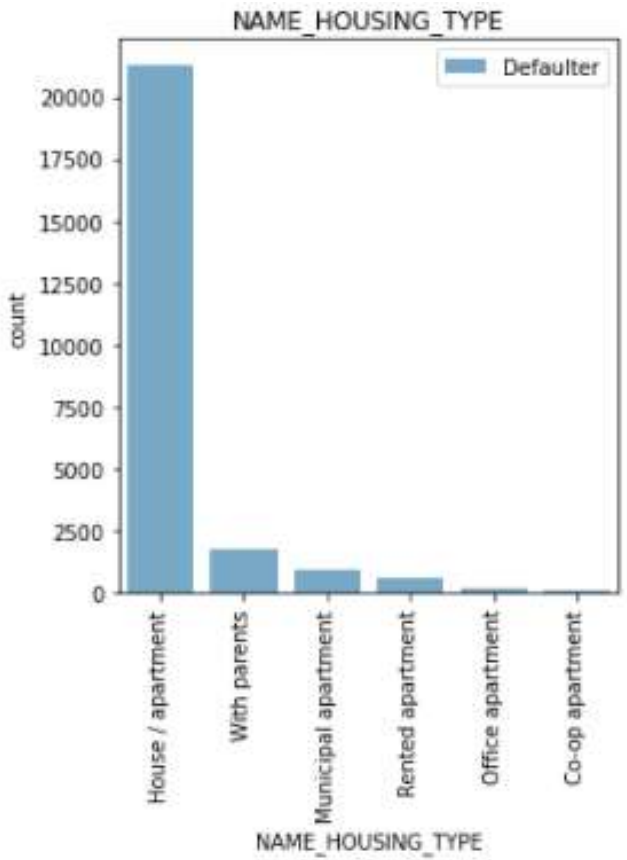
2.Second highest for taking the loans are Single/not Married clients and then Civil Marriages, Seperated and Widow.

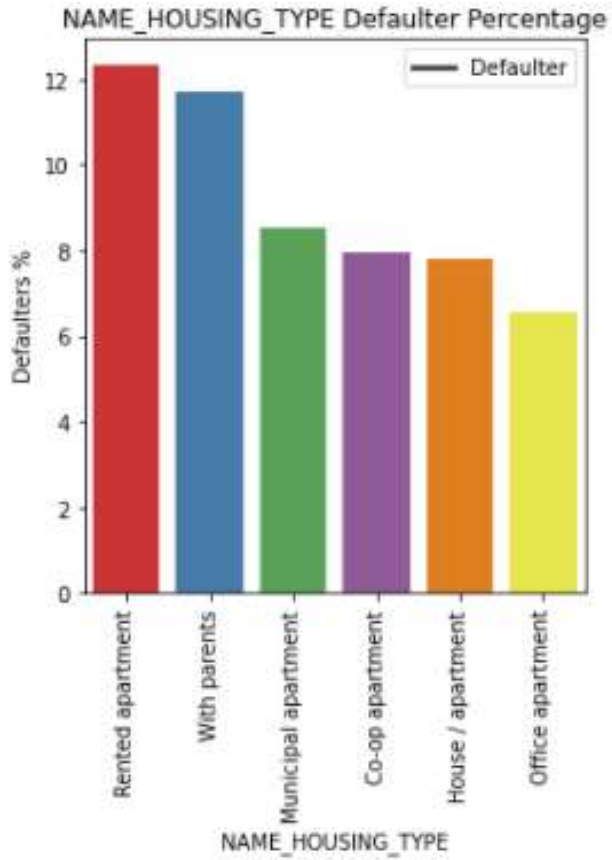
3.As seen in the Defaulter % plot Civil marriage (10%) has the highest percenatage of defaulters and then comes Single/not married.

4.Widow has the lowest percentage of defaulter.

NAME_HOUSING_TYPE







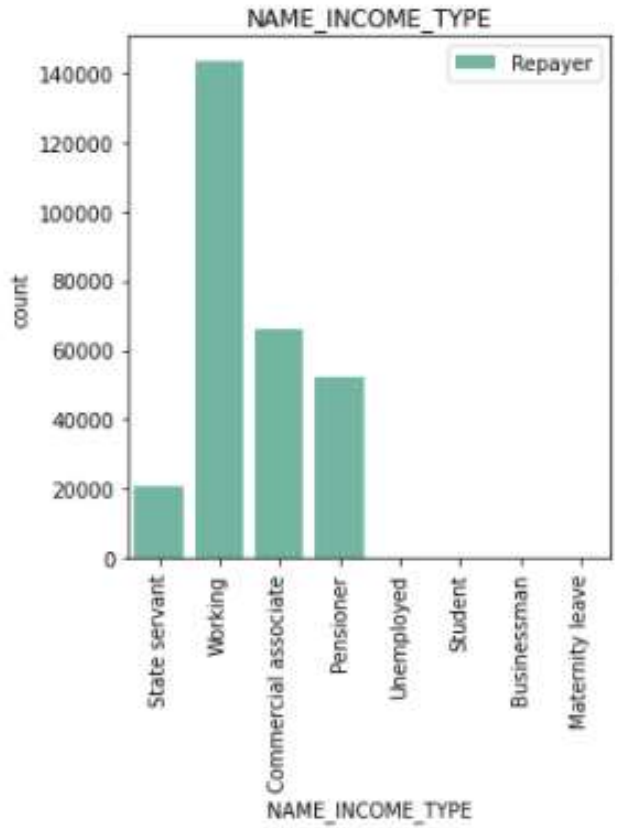
INFERENCES:

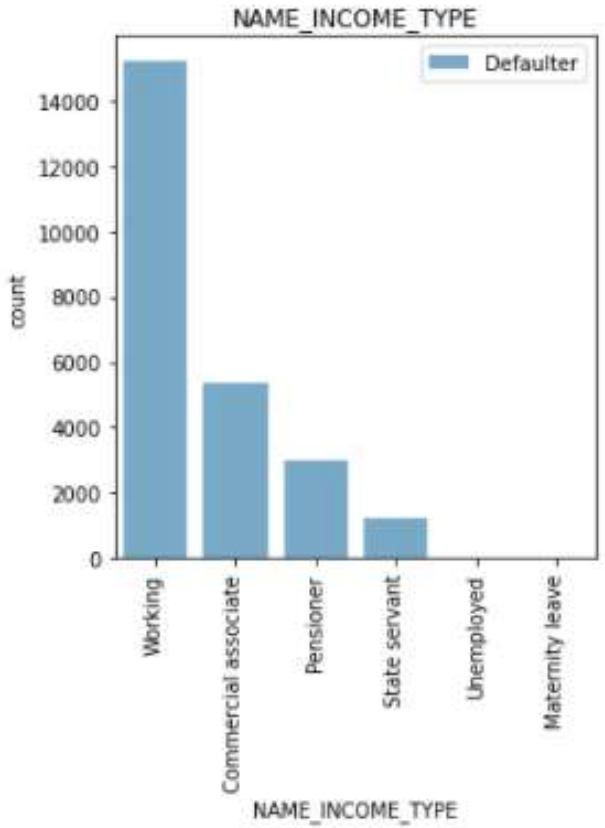
1.From this plot we get to know that majority of the clients stays in their own House/Apartment.

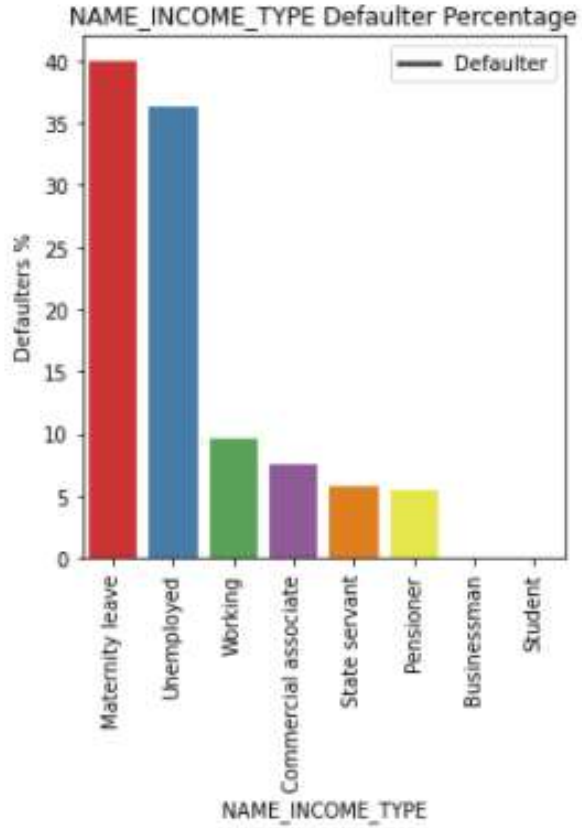
2.Then comes clients who lives with their parents and then comes clients who live in Municipal apartment.

3.Clients living in Rented apartment(12%) and clients living with parents(approx:11.8) have Highest default rates.

NAME_INCOME_TYPE



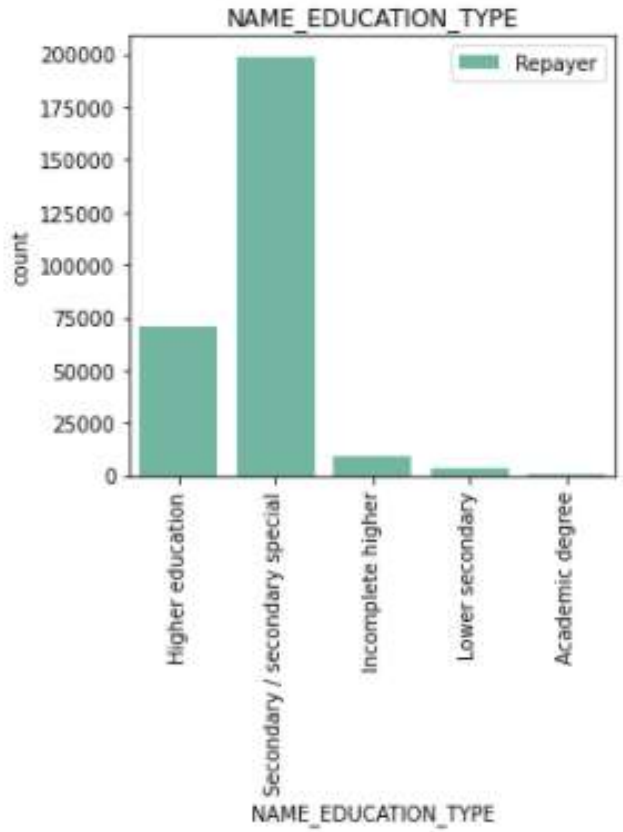


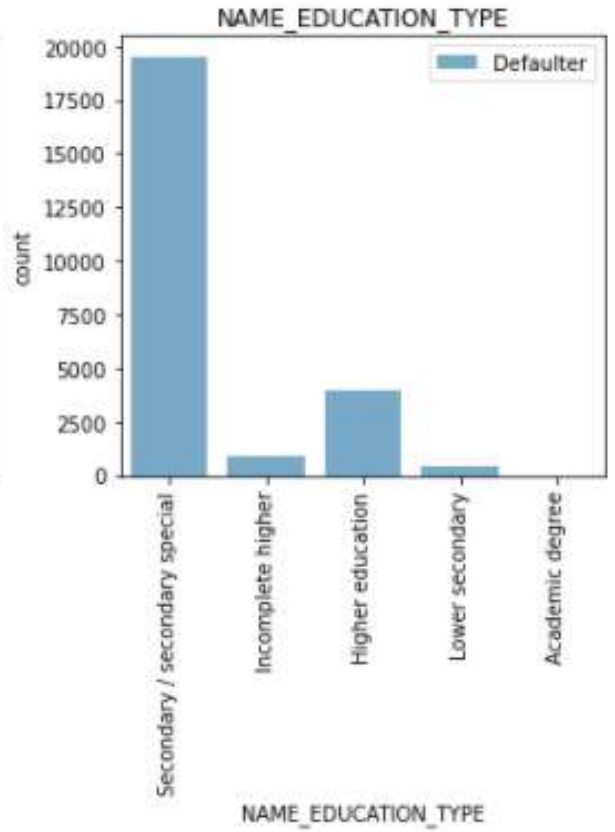


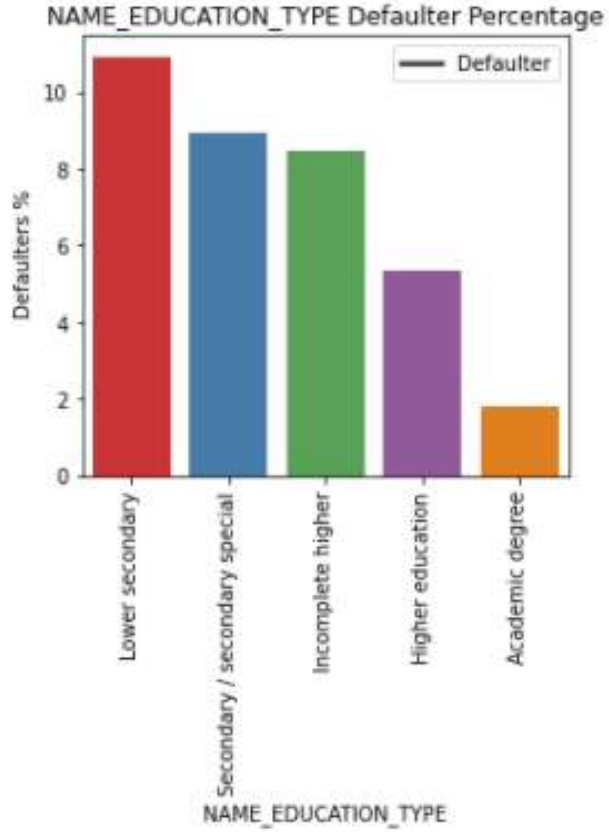
INFERENCES:

1.Large number of clients who are taking loans are working clients followed by Commercial associate and pensioner.
2.Clients with the type of Maternity leave (40%) are the clients not repaying the loan.
3.Loan can be provided to students and businessman as theirs almost no record of defaulters for them. So it could be the most reliable and safest option.

NAME_EDUCATION_TYPE



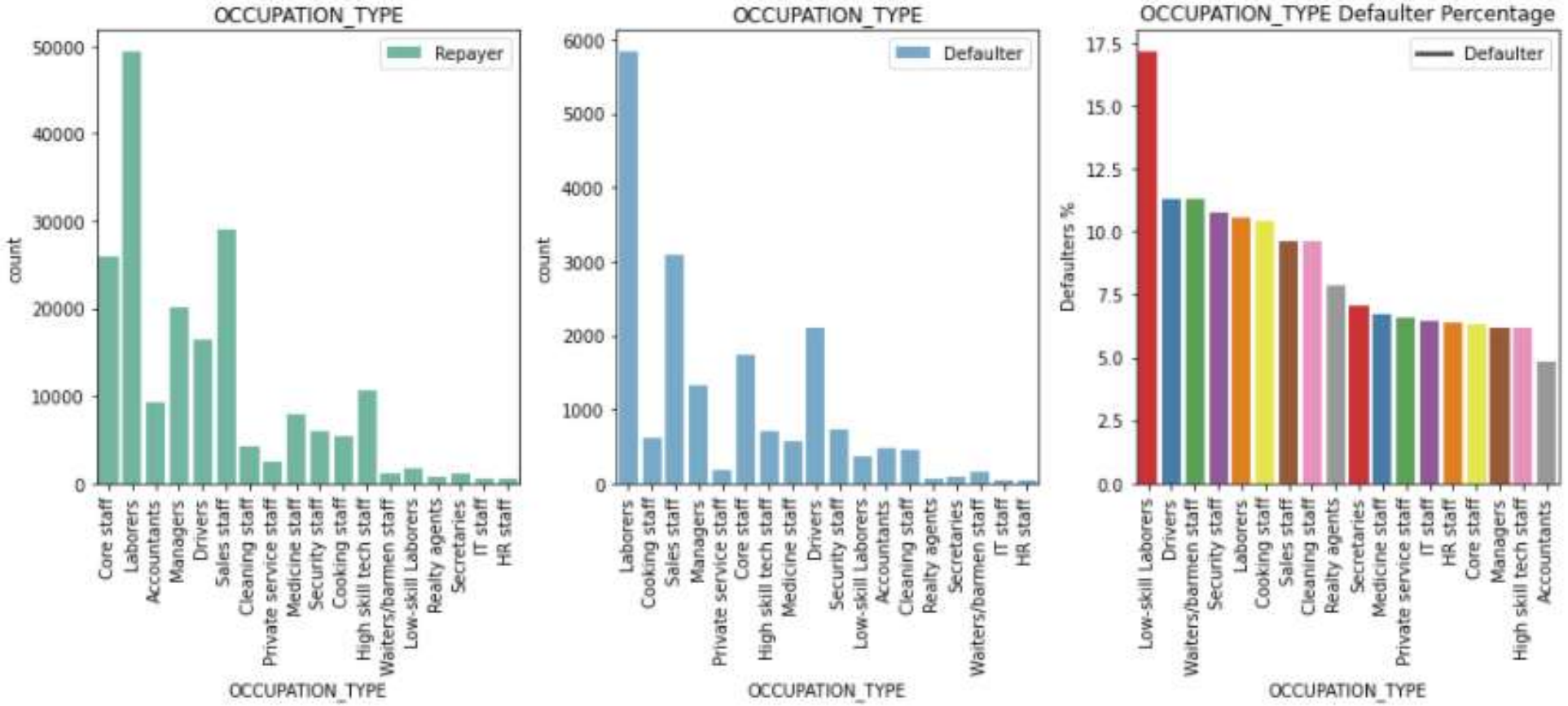




INFERENCES:

1.Majority of the clients have Secondary/Secondary special education then comes Higher education.
2.Defaulter % (approx:11%) of Lower secondary education type is the highest. This means that clients with lower education have higher chances of not repaying the loans.
3.Clients with Academic degree have the lowest defaulter %.

OCCUPATION_TYPE



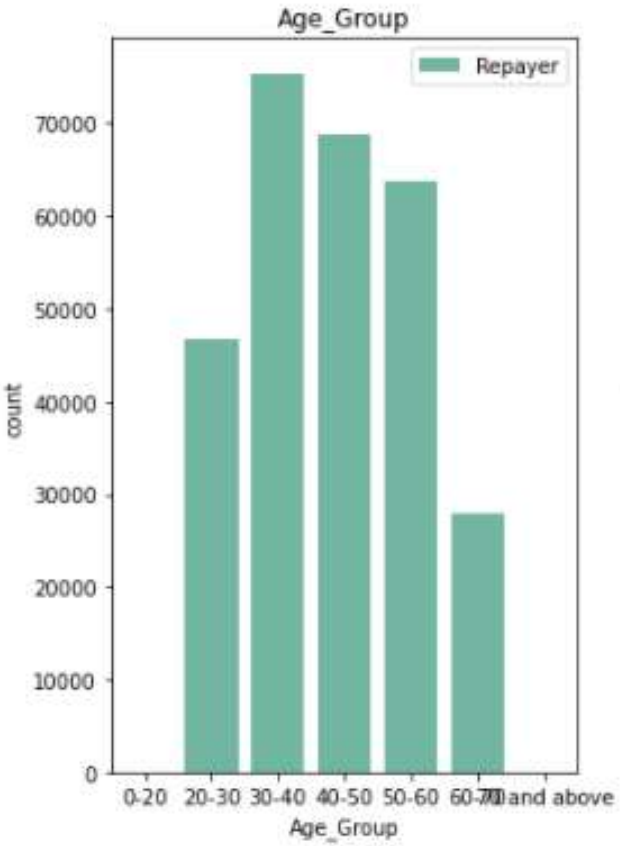
INFERENCES:

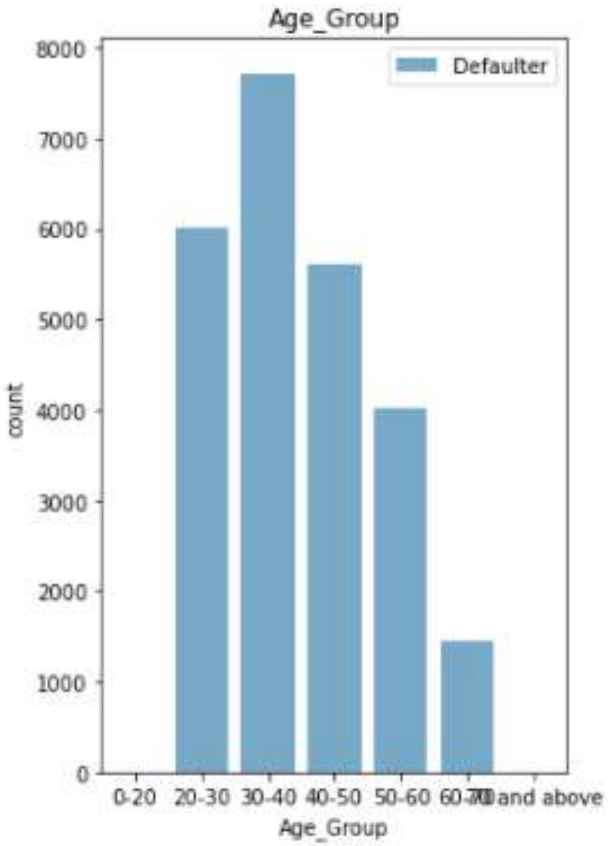
1.Majority of the loans are taken by Laborers followed by Sales Staff, Core Staff, Drivers, Accountants.

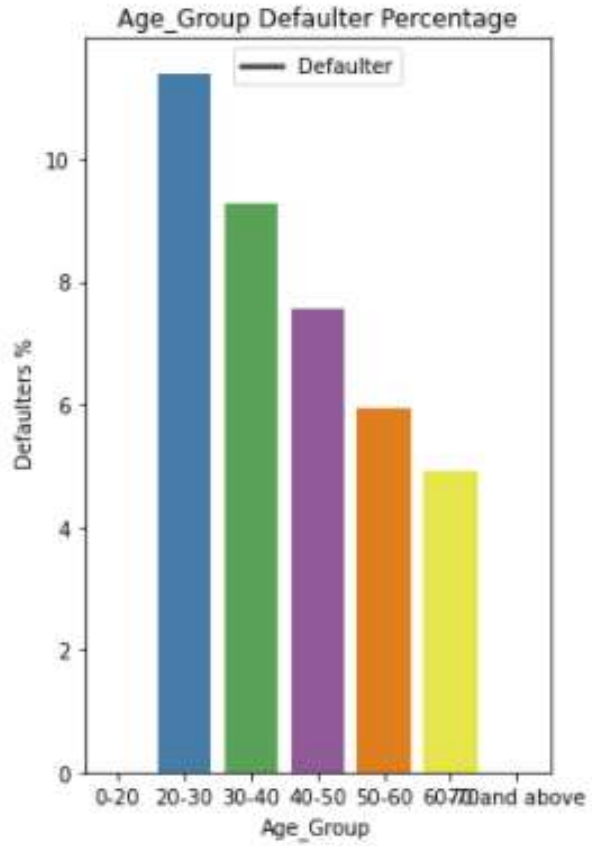
2.Though the percentage of loan take by low-skilled laborers are less, but it has the highest percentage of not repaying the loan.

3.Then followed by Drivers, Waiters/barmen staff, security staff, labourers, cooking staff and so on.

AGE_GROUP



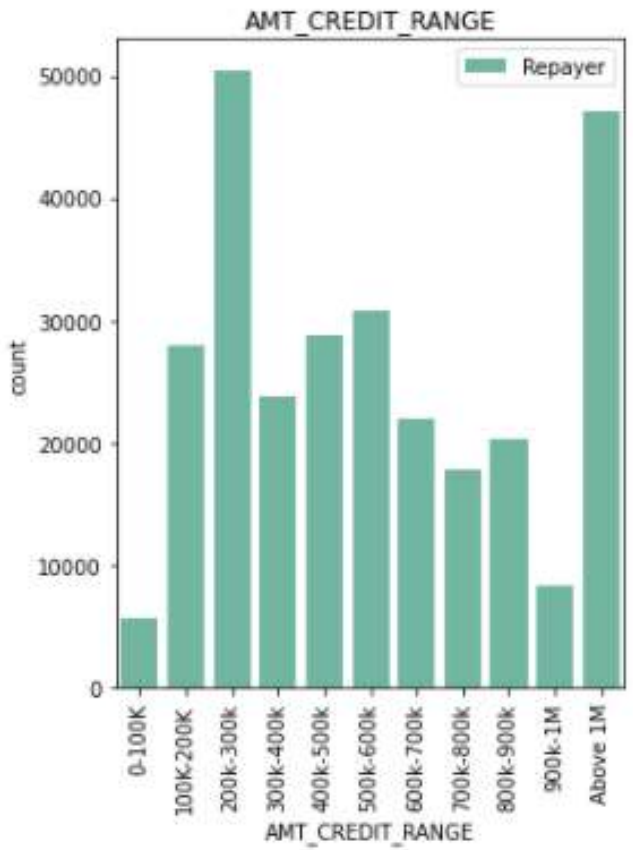


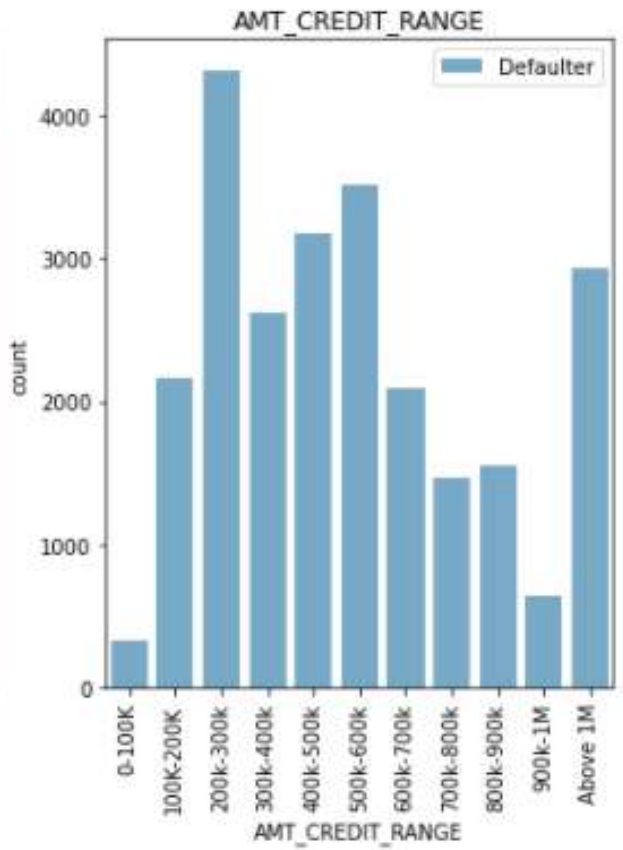


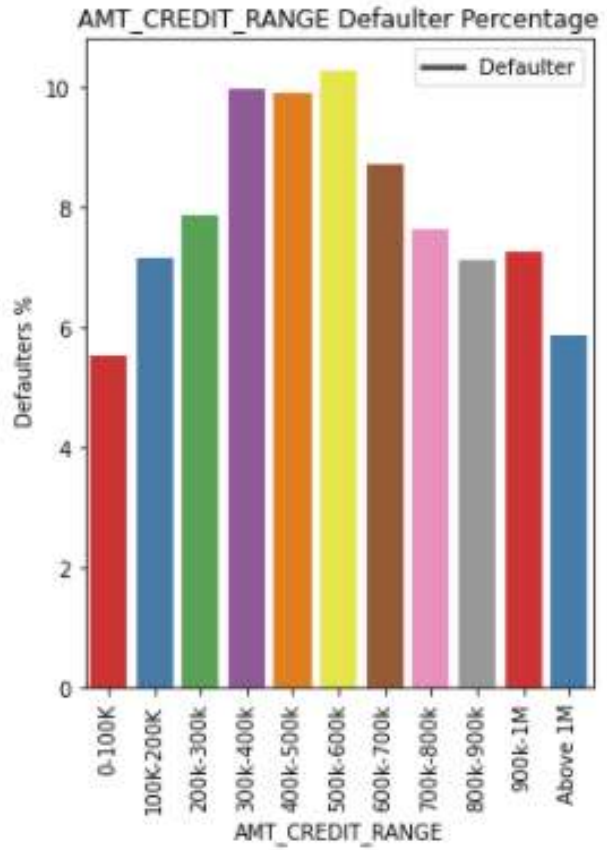
INFERENCES:

1.Clients within age group of 30 to 60 have the highest applicants.
2.Clients with age group of 20-30 have higher chances of not repaying the loan or we can say higher chances of being defaulters.

AMT_CREDIT_RANGE







INFERENCES:

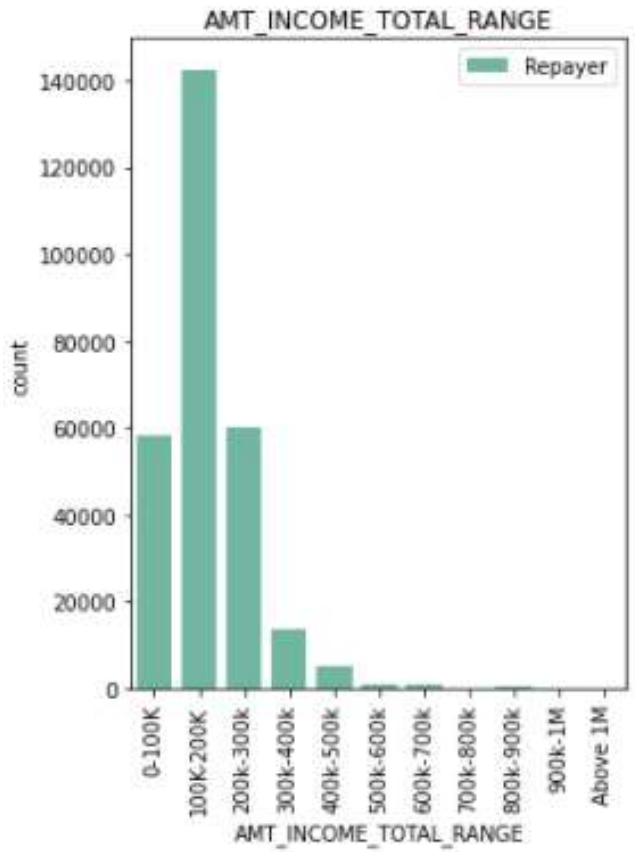
1.Majority of the clients have been granted loan for amount between 200k-300k.

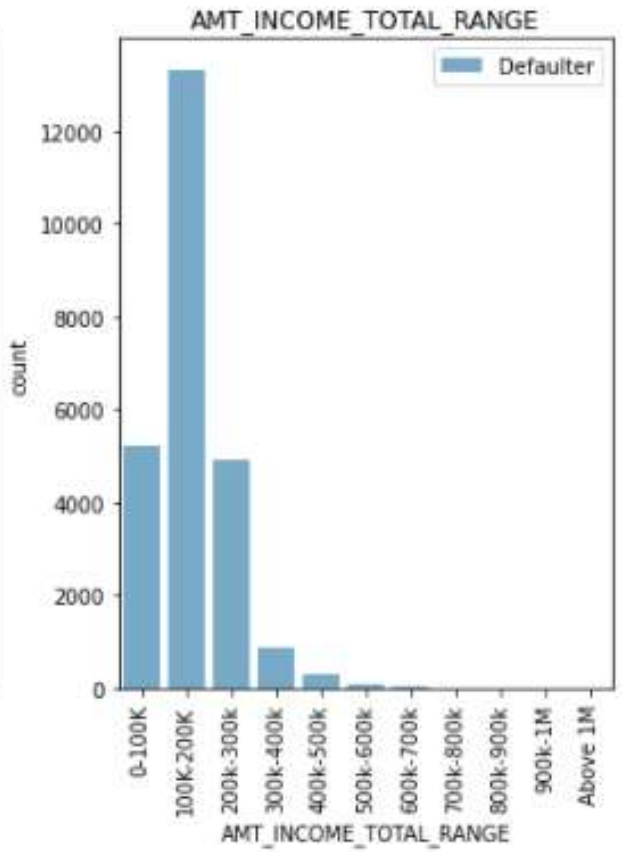
2.very less amounts of clients have beenn credited with the loan amount more than 1M.

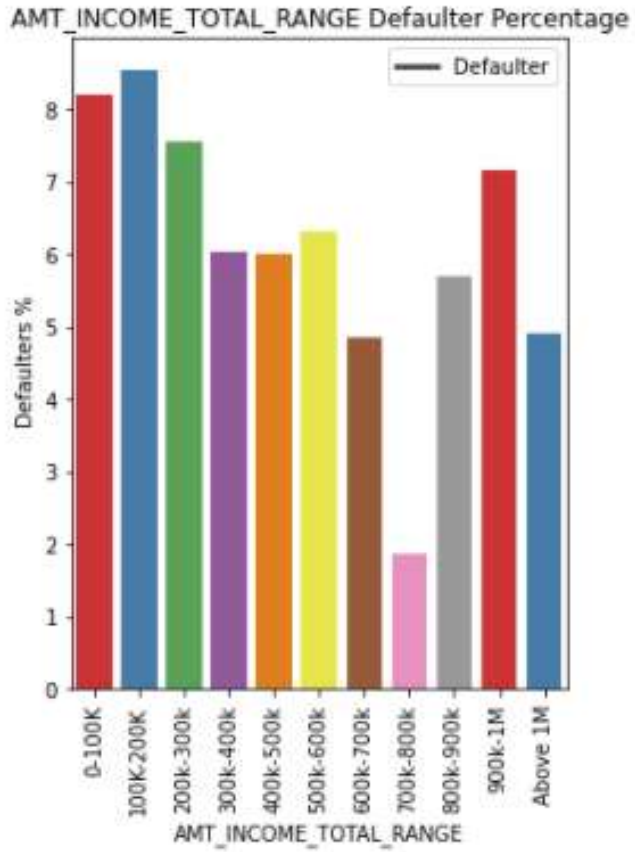
3.Mostly loan credited to the clients is less than 900k.

4.Cleints who have been credited with the loan between 300k to 600k has more percentage of defaulter.

AMT_INCOME_RANGE



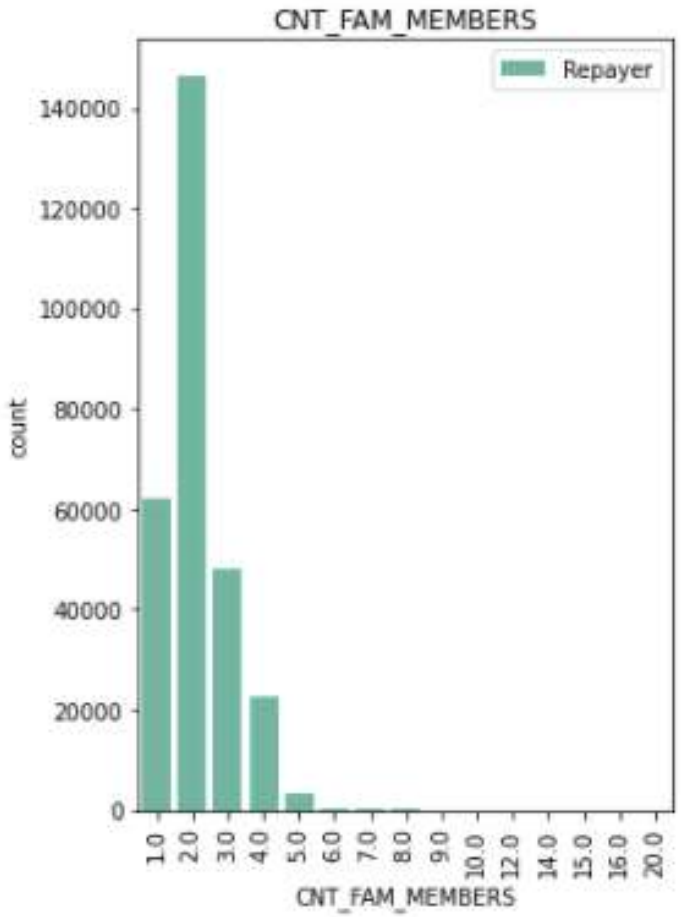


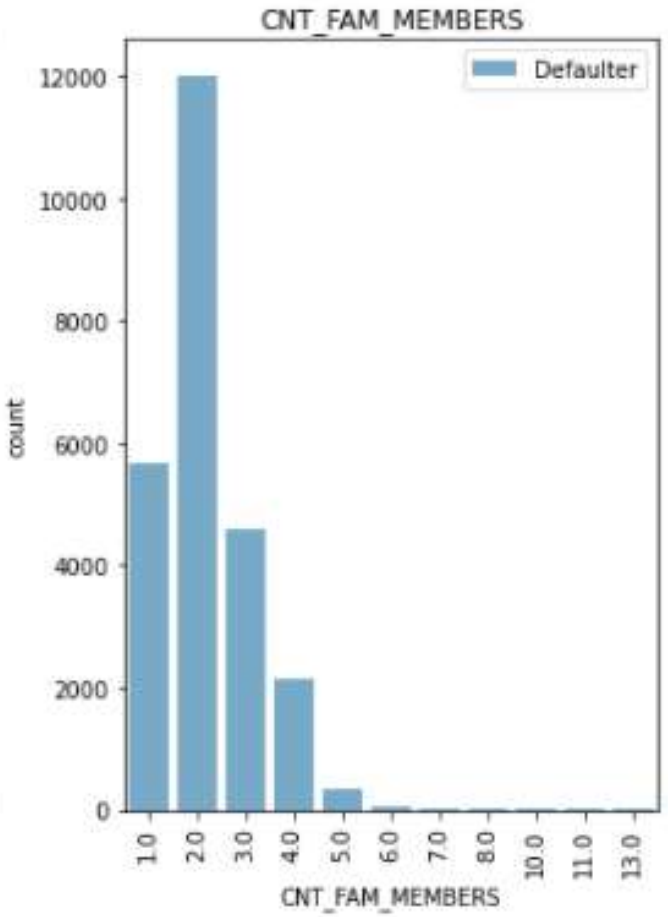


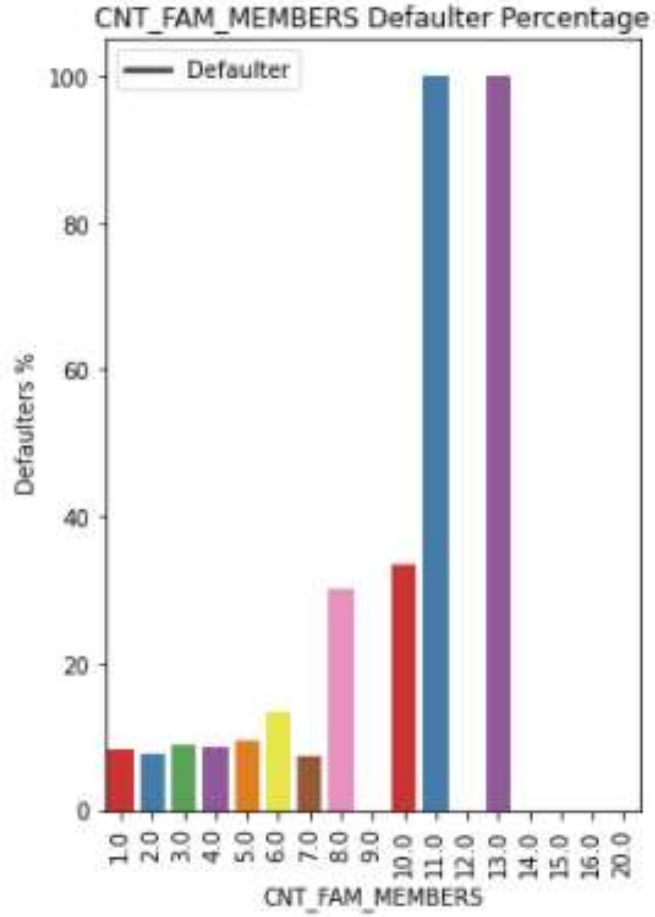
INFERENCES

1.Majority of the clients have the income range bewteen 100k-200k.
2.Highest default rate is also for the clients with income range of 100k-200k followed by the clients with income range of 0-100k

CNT_FAM_MEMBERS



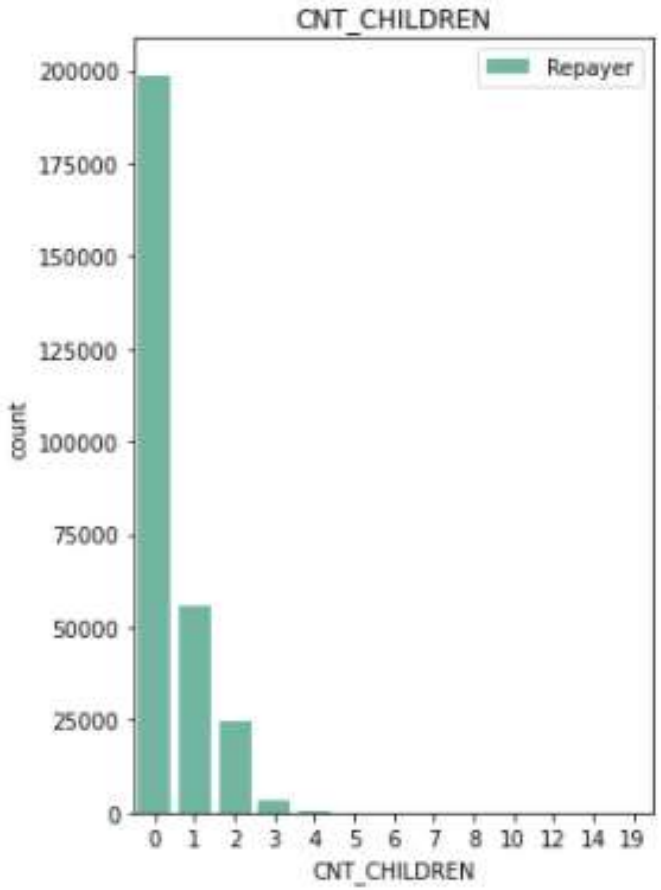


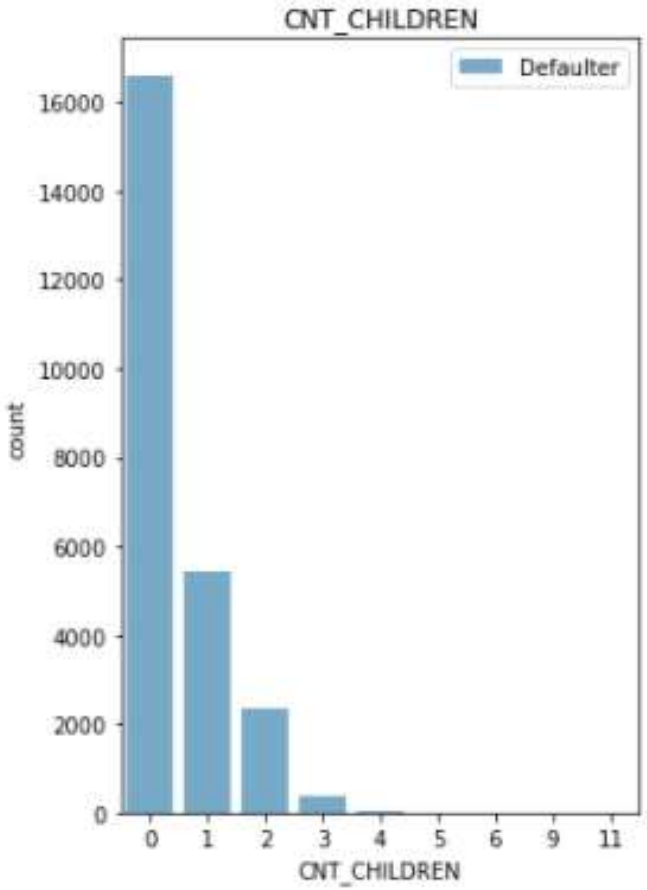


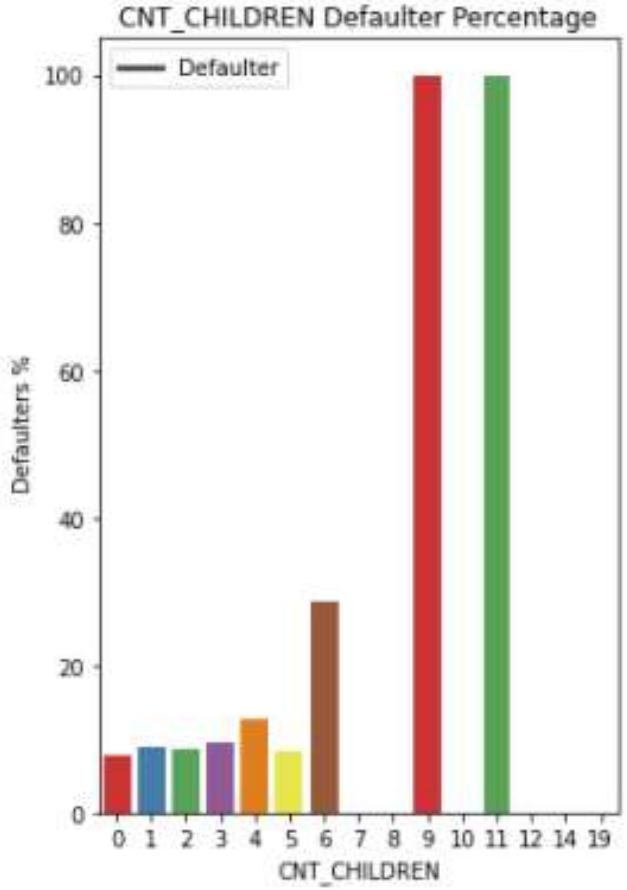
INFERENCES:

1.Majority of the clients are having 2 members in the family followed by 1,3 and 4.
2.We can see clients with members 11 and 13 have 100% of defaulter rate.

CNT_CHILDREN



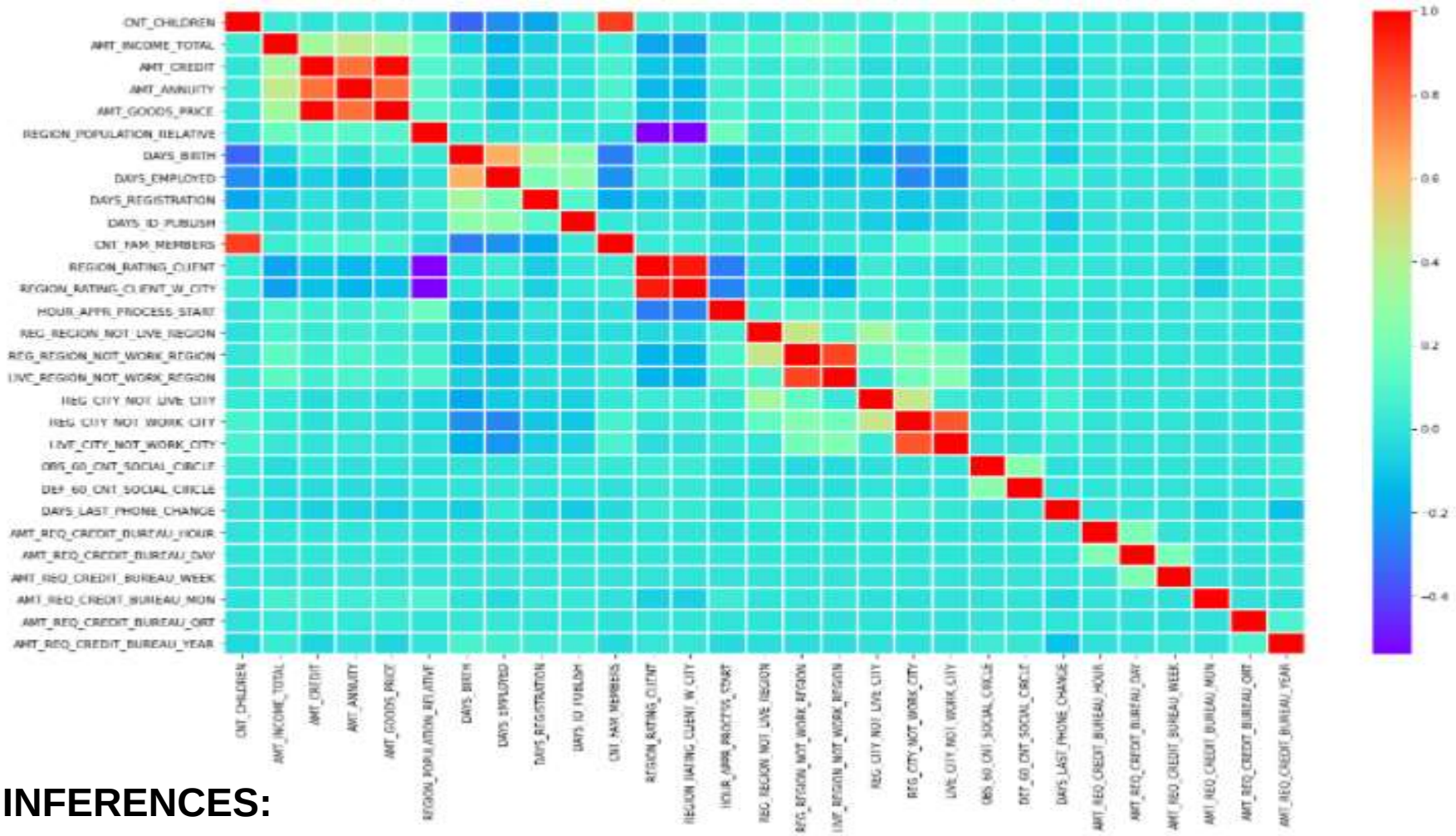




INFERENCES:

1.Majority of the clients have no children.
2.Very less clients can be seen who has 3 children.
3.Clients with the child count 9 and 11 are showing 100% defaulter rate..

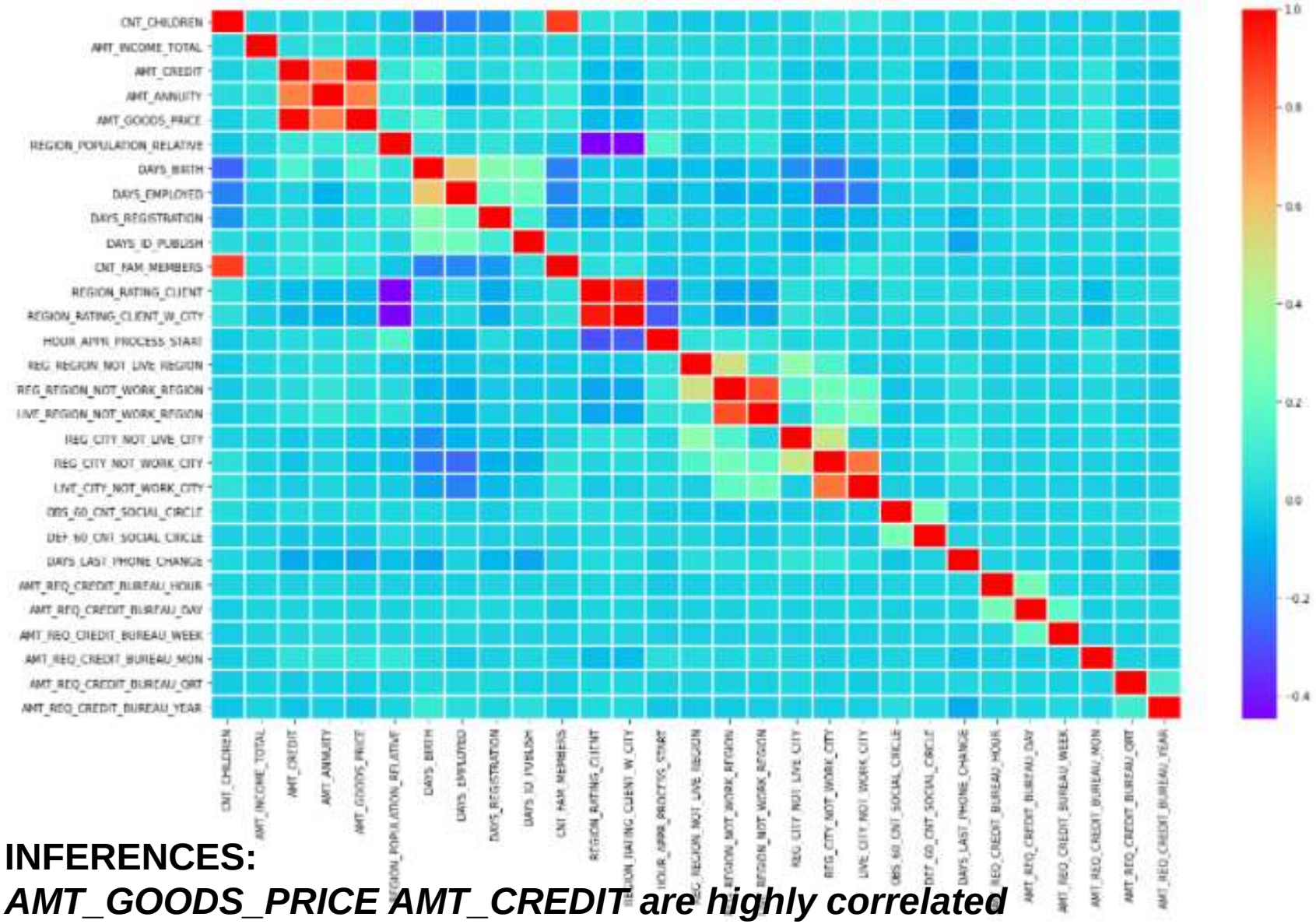
CORRELATION FOR NUMERICAL COLUMNS OF REPAYERS FROM APPLICATION DATA



INFERENCES:

1. AMT_GOODS_PRICE and AMT_CREDIT credit are highly correlated
2. CNT_FAM_MEMBERS and CNT_CHILDREN are also highly correlated
3. AMT_GOODS_PRICE and AMT_ANNUITY are Moderately correalted
4. AMT_ANNUITY and AMT_CREDIT are also moderately correlated
5. DAYS_EMPLOYED and DAYS_BIRTH are also correlated with each other

CORRELATION FOR NUMERICAL COLUMNS OF DEAFULTERS FROM APPLICATION DATA



INFERENCES:

AMT_GOODS_PRICE AMT_CREDIT are highly correlated

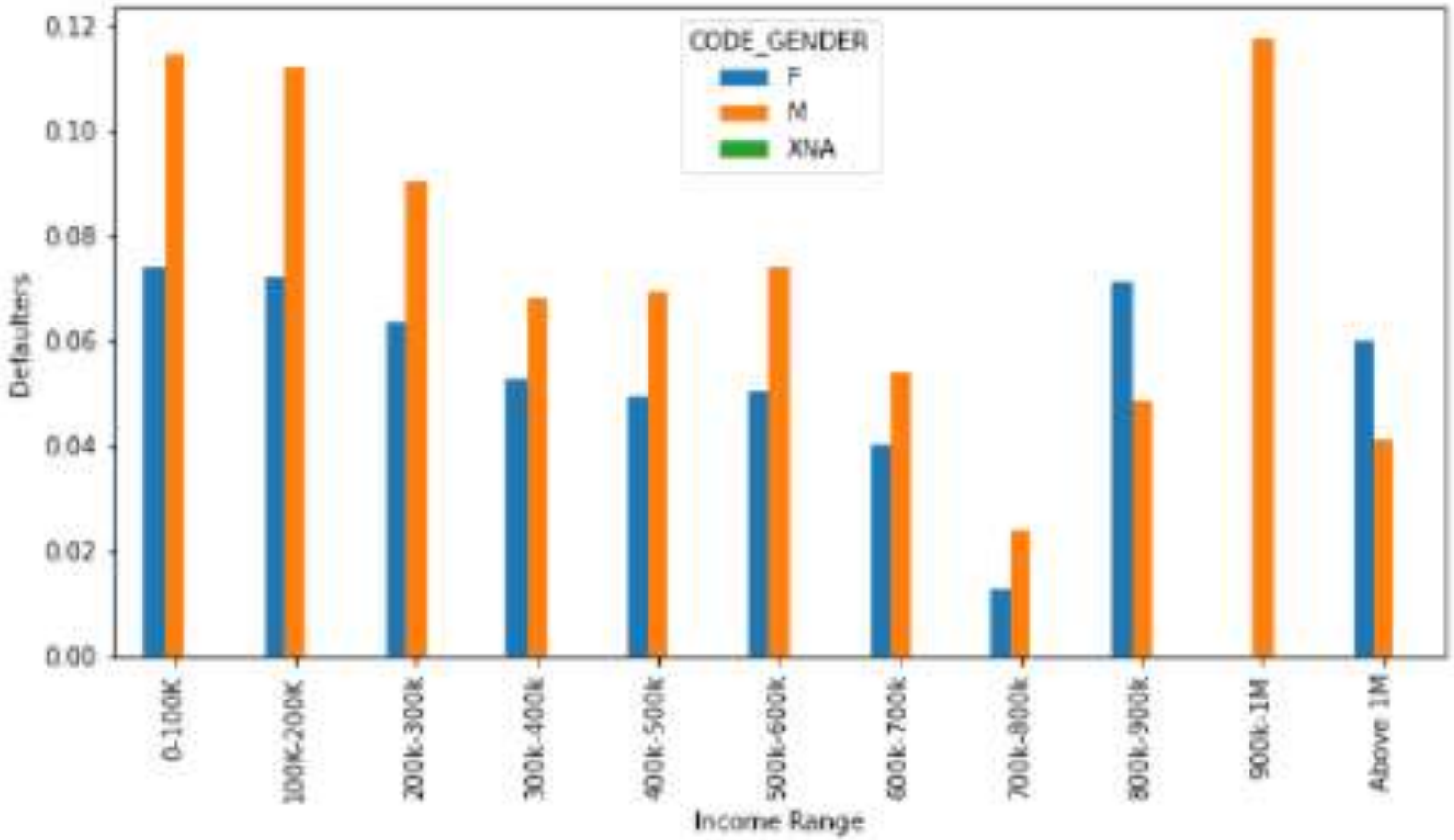
CNT_FAM_MEMBERS and CNT_CHILDREN are moderately correlated

AMT_GOODS_PRICE and AMT_ANNUITY are moderately correlated

AMT_ANNUITY and AMT_CREDIT are also correlated with each other

BIVARIATE ANALYSIS

COMPARING INCOME RANGE AND GENDER

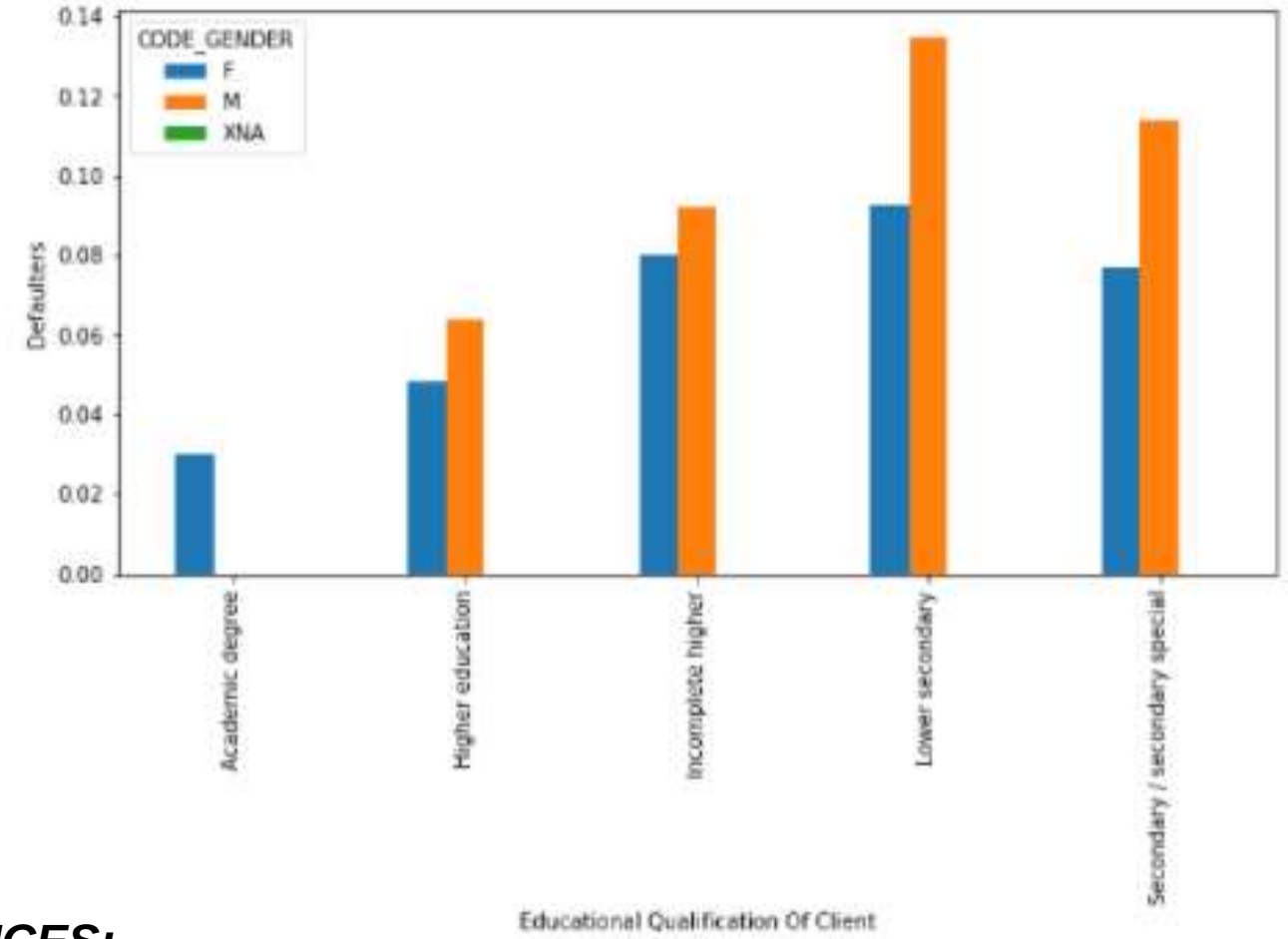


INFERENCES

1. Male and female in the income range 0 to 100k and 100k to 200k have equal default rate around 10%

2. Females earning above 1M are more likely to default than males

COMPARING EDUCATION TYPE AND GENDER



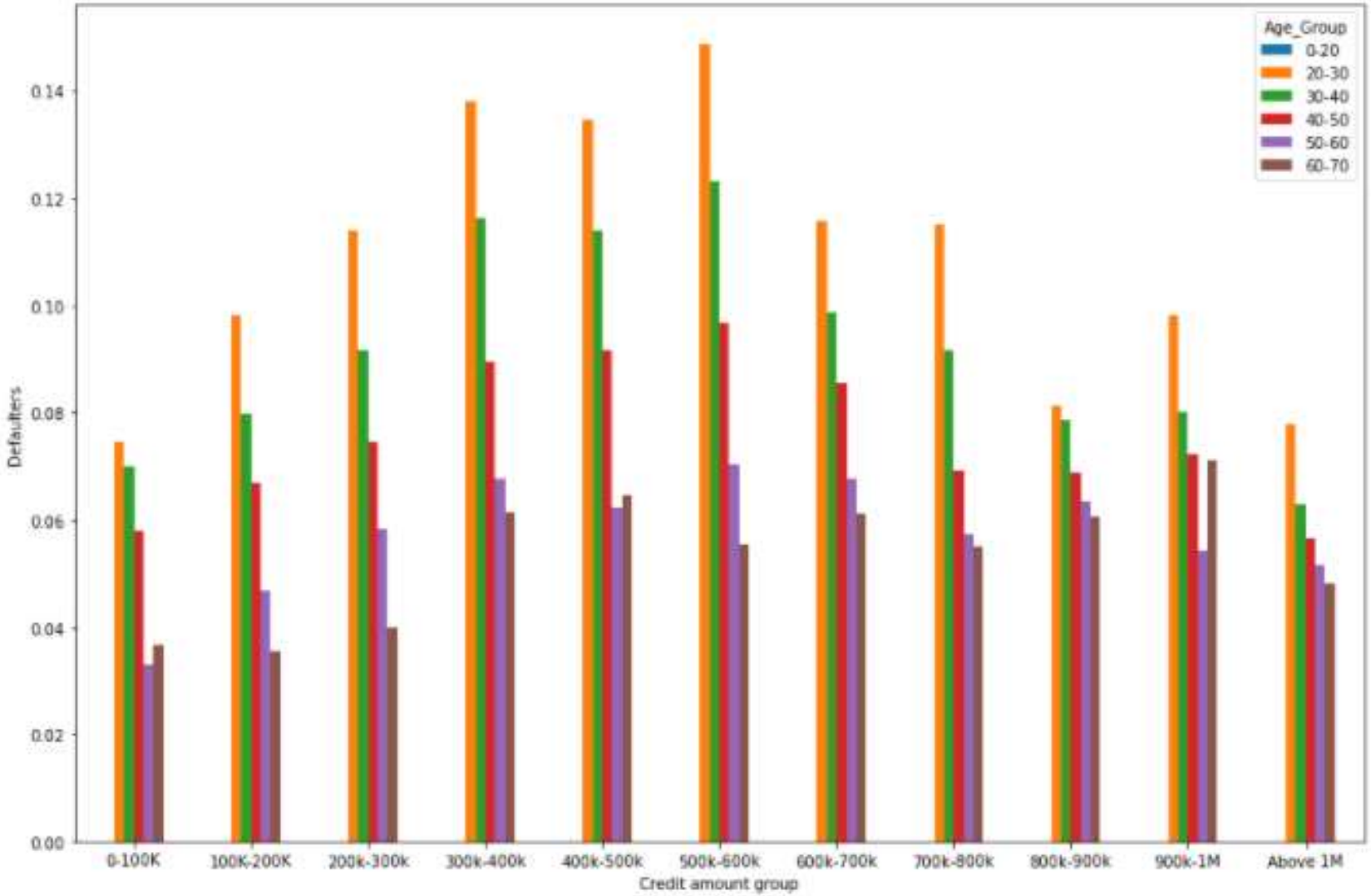
INFERENCES:

1.Males and female clients with Lower secondary Educational Qualification have higher defaulter rates.

2.Then comes males clients with Secondary/secondary special who have high defaulter rate.

3.People with higher education have low defaulter rate.

COMPARING AMOUNT CREDIT RANGE AND AGE GROUP

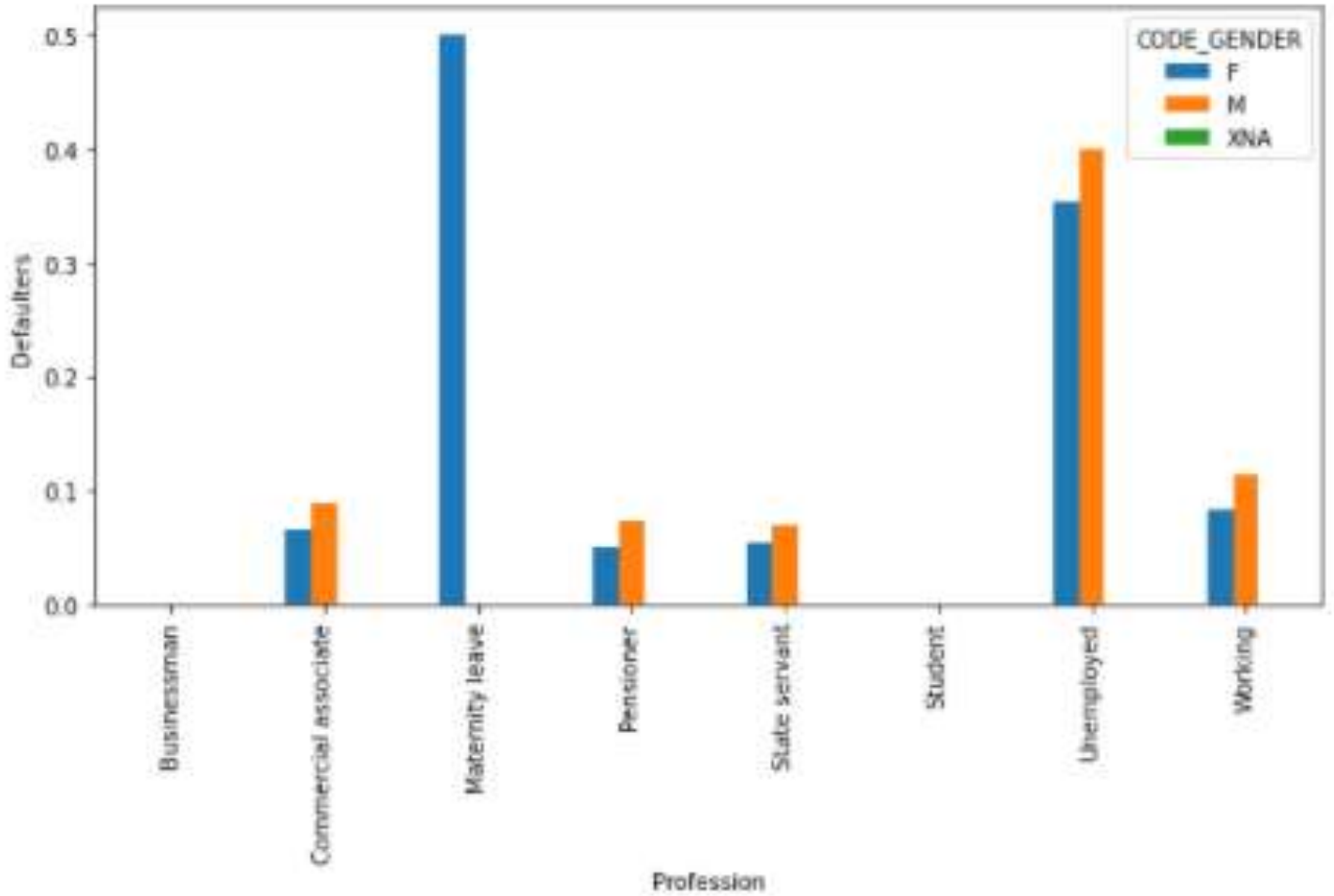


INFERENCES:

1.Form this plot we can see that young clients with the average income have the most default rates.

2.All the senior citizens are less likely to be defaulters.

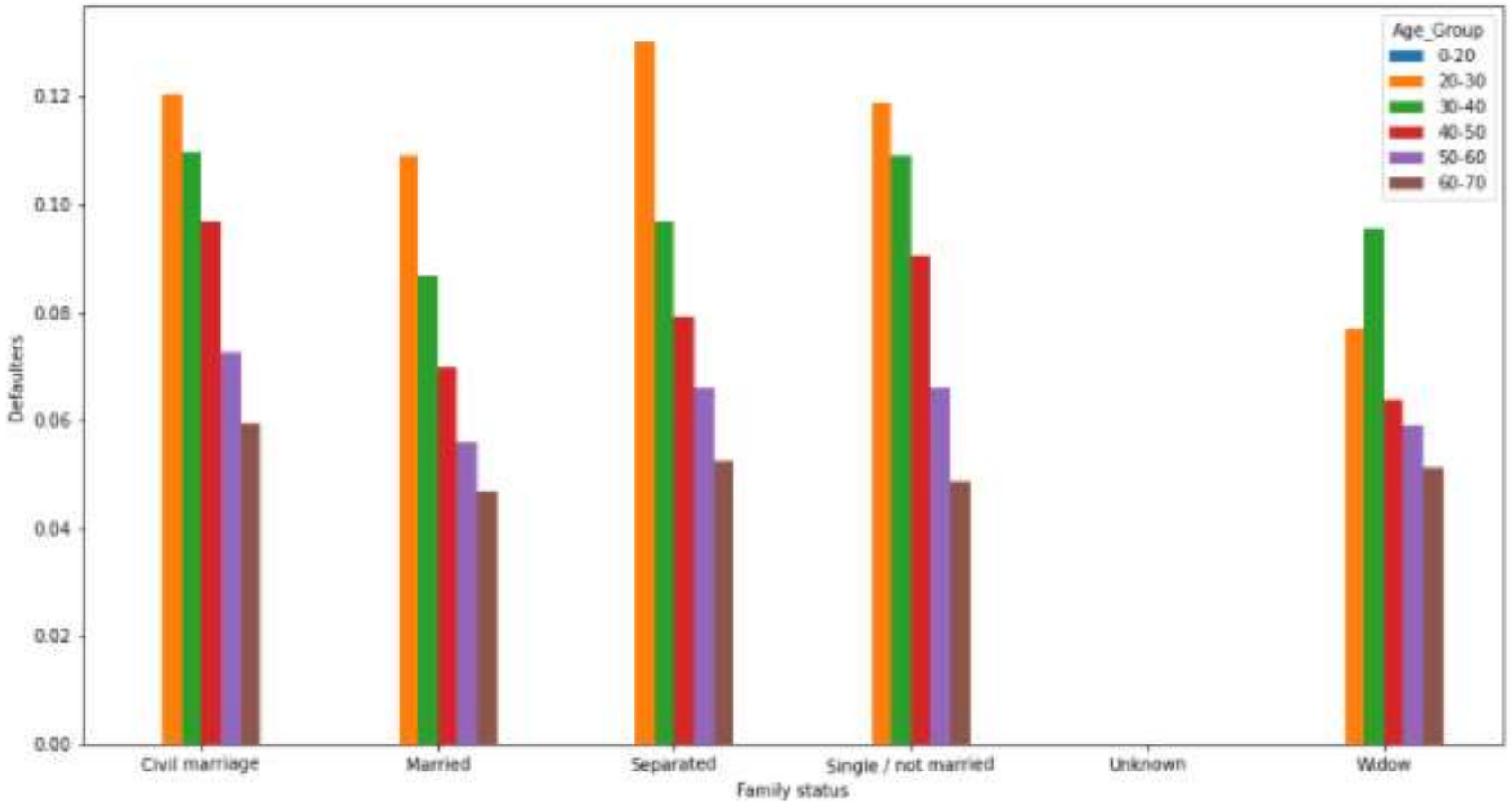
COMPARING INCOME TYPE AND GENDER



INFERENCES:

1.Unemployed clients have more percentage of defaulter rate.
2.Also clients who are at maternity leave have the highest chances of defaulter rate.
3.Overall males have more defaulter rate in all the Income types.

COMPARING FAMILY STATUS AND AGE GROUP

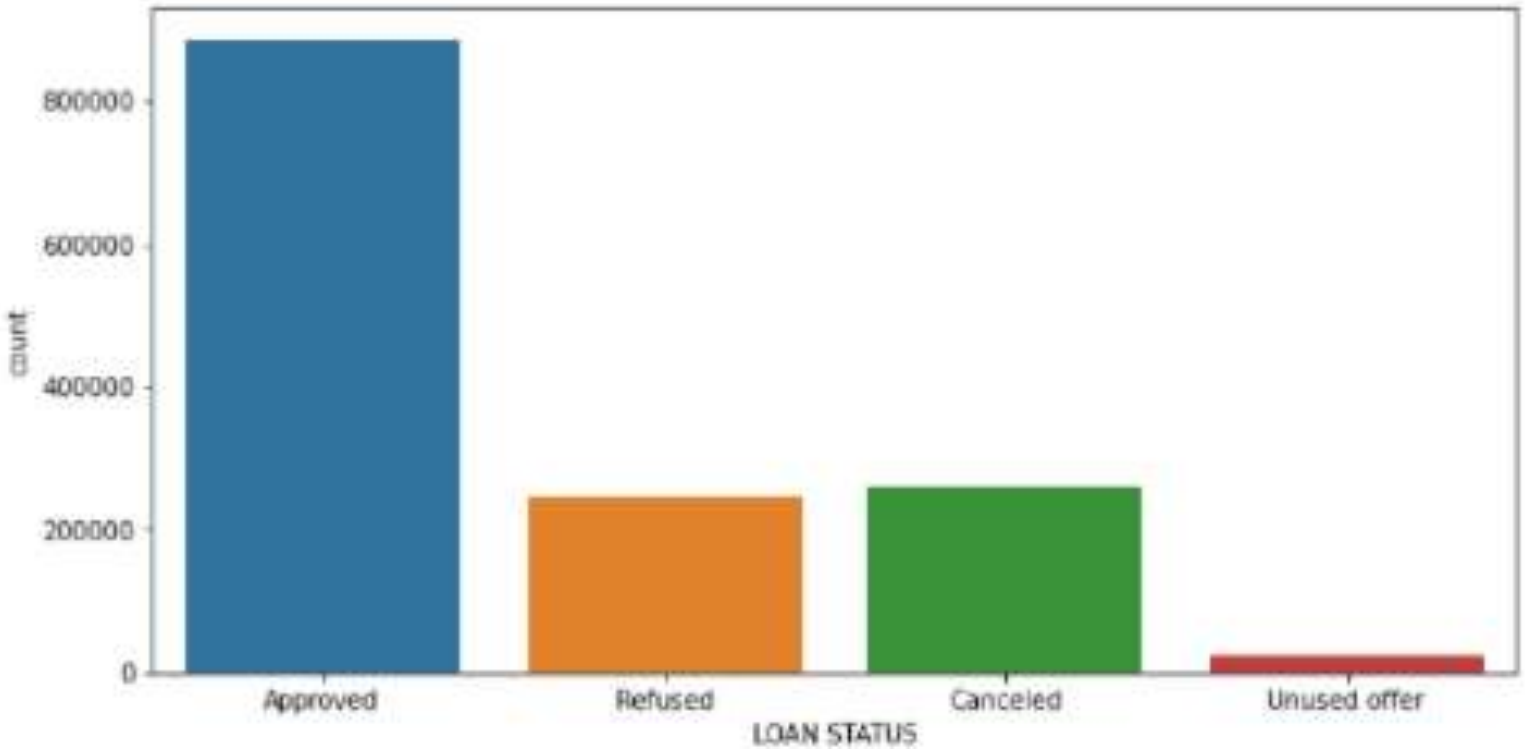


NFERENCES:

1. In this plot we can see that younger generations have higher percentage of defaulters in all the status except widow.
2. Lowest defaulter rate are the clients who are elderly age.

UNIVARTIE ANALYSIS (MERGED DATASET)

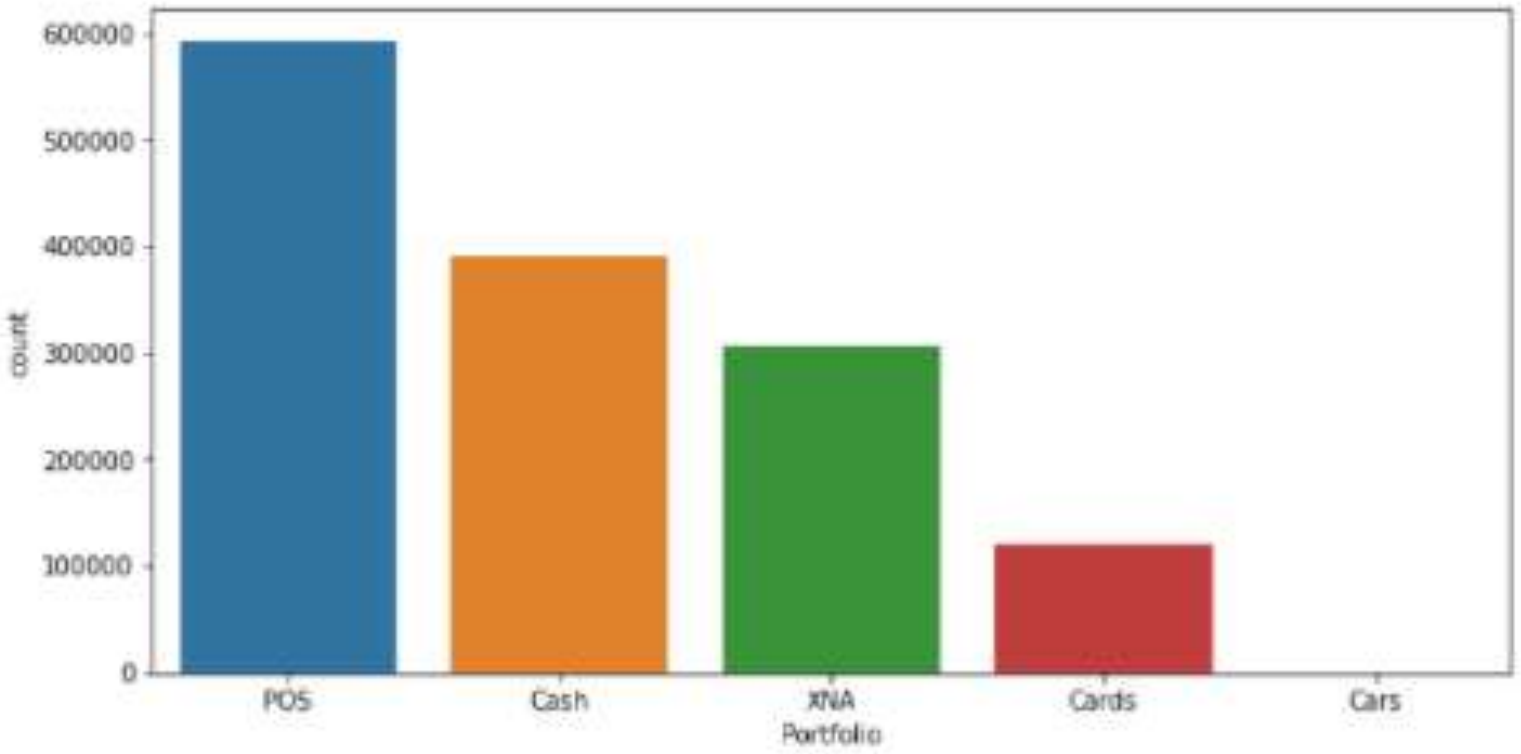
NAME_CONTRACT_STATUS



INFERENCES:

1. Mostly Loans are being approved
2. Almost same number loans are being refused or either got cancelled

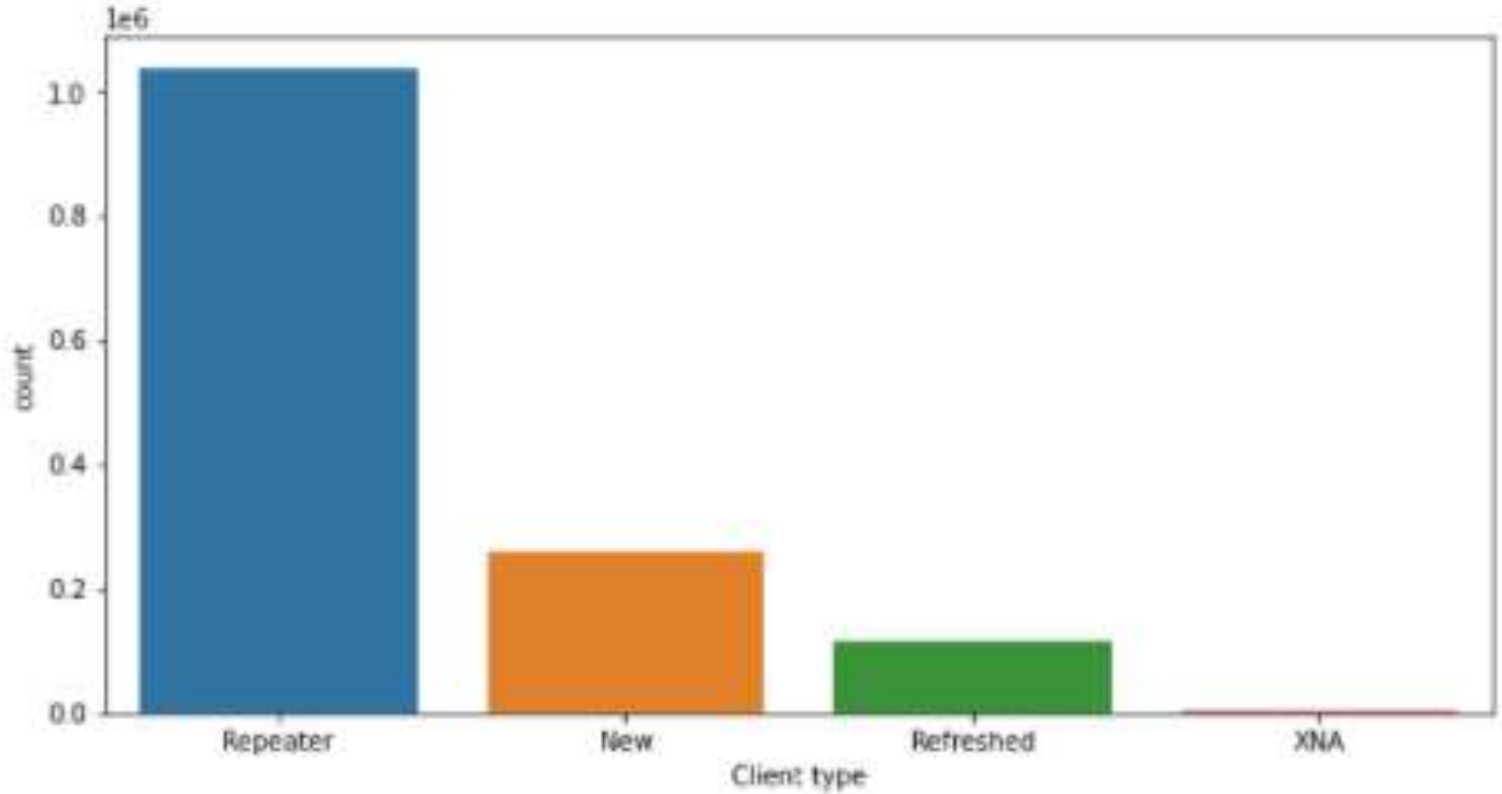
NAME_PORTFOLIO



INFERENCE

Most of the previous applications were for POS followed by cash, cards and cars

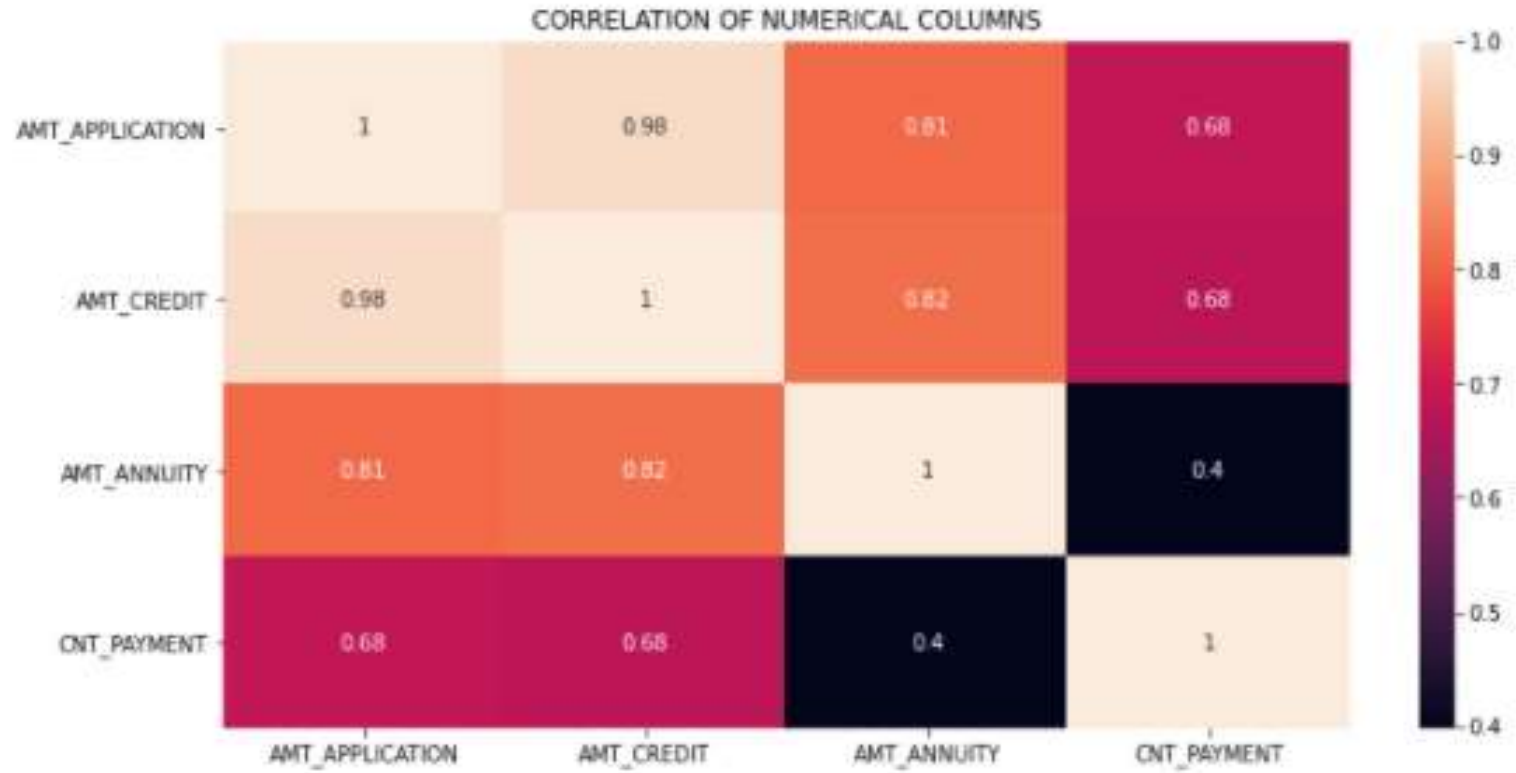
NAME_CLIENT_TYPE



INFERENCE

Most of the applications were repeater followed by new and refreshed.

CORRELATION OF NUMERICAL COLUMNS OF PREVIOUS APPLICATION DATA

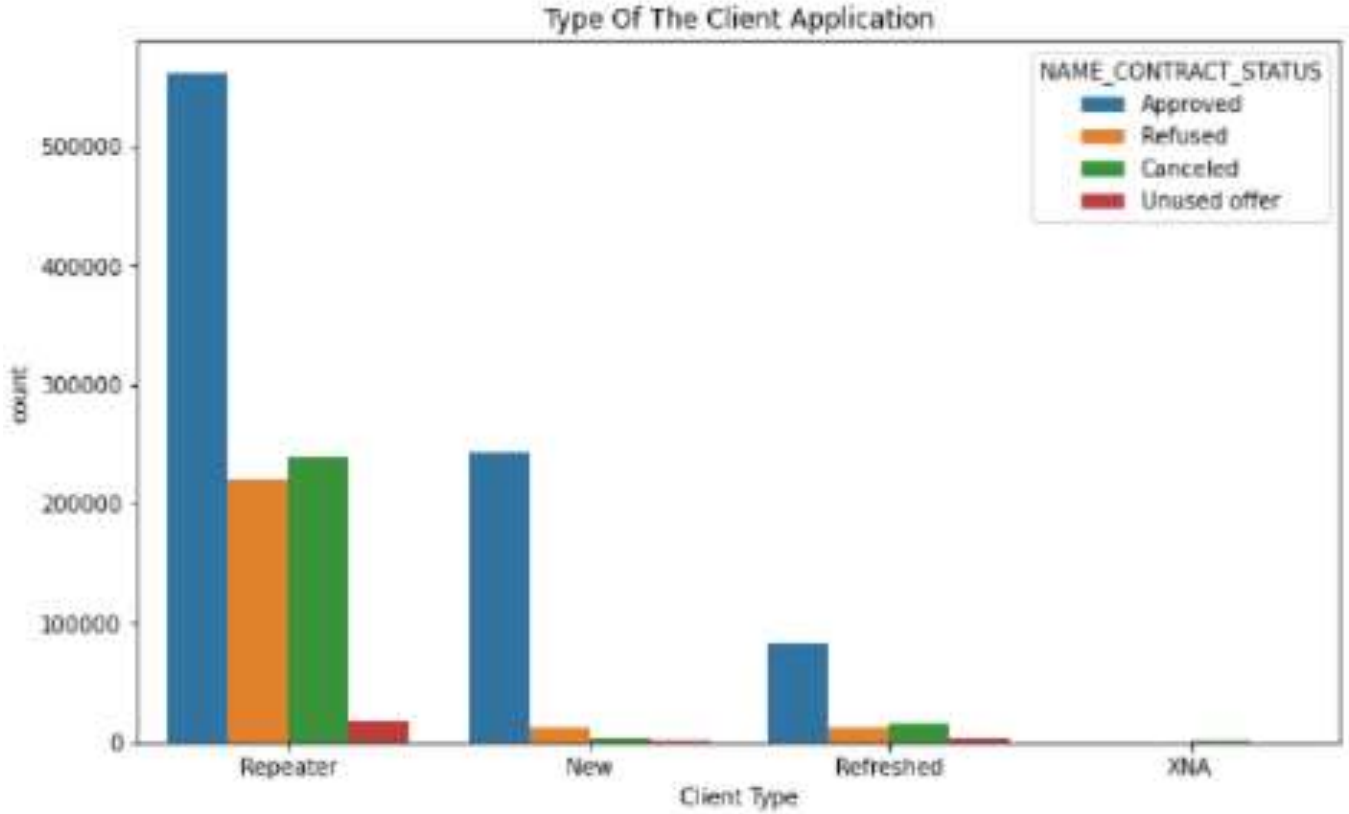


INFERENCE

According to the heat map highly correlated columns are AMT_APPLICATION and AMT_ANNUITY, AMT_CREDIT and AMT_ANNUITY, AMT_APPLICATION AND AMT_CREDIT.

BIVARIATE ANALYSIS (MERGED DATASET)

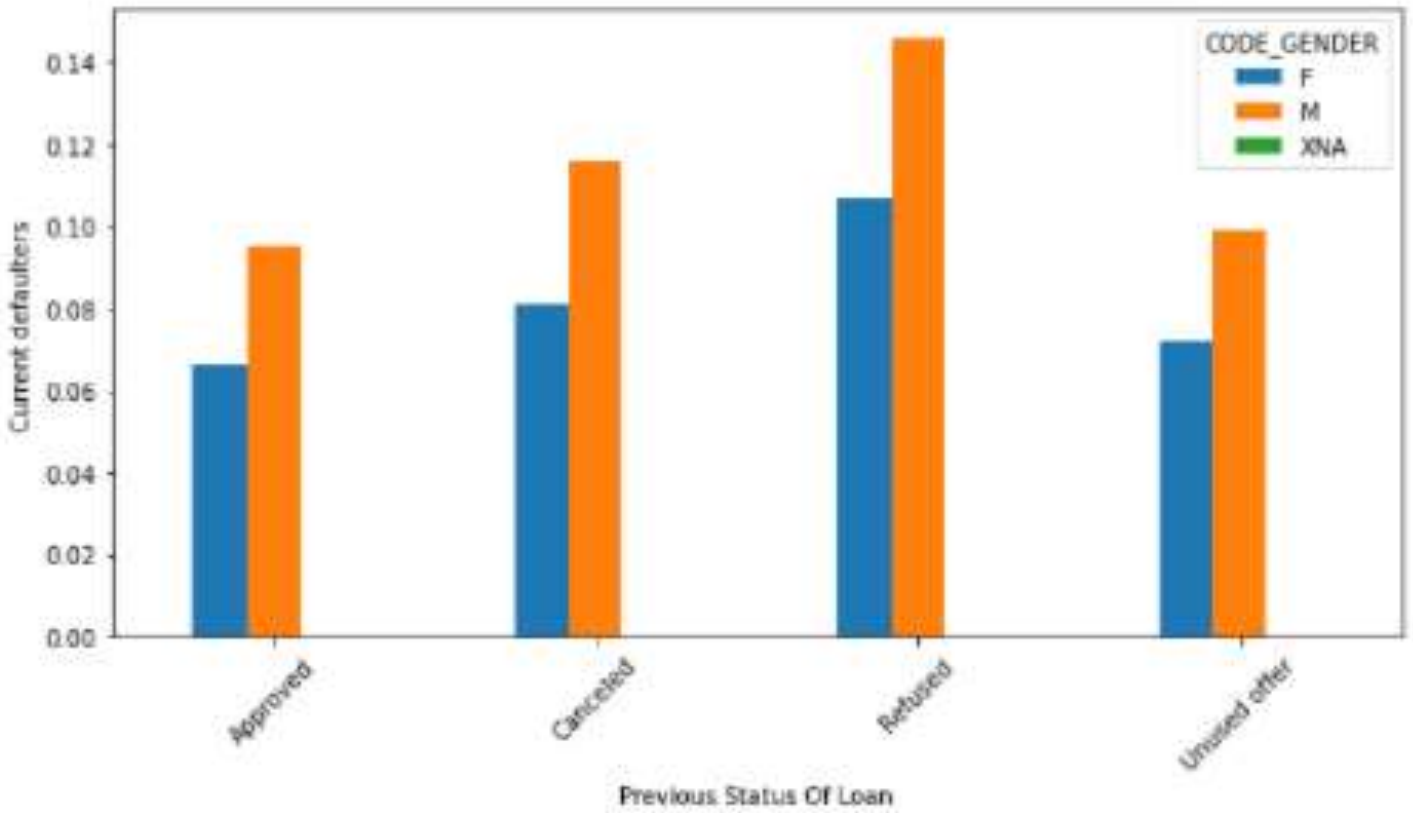
COMPARING CLIENT TYPE AND STATUS



INFERENCE

Most of the approved loans are off the repeater clients.

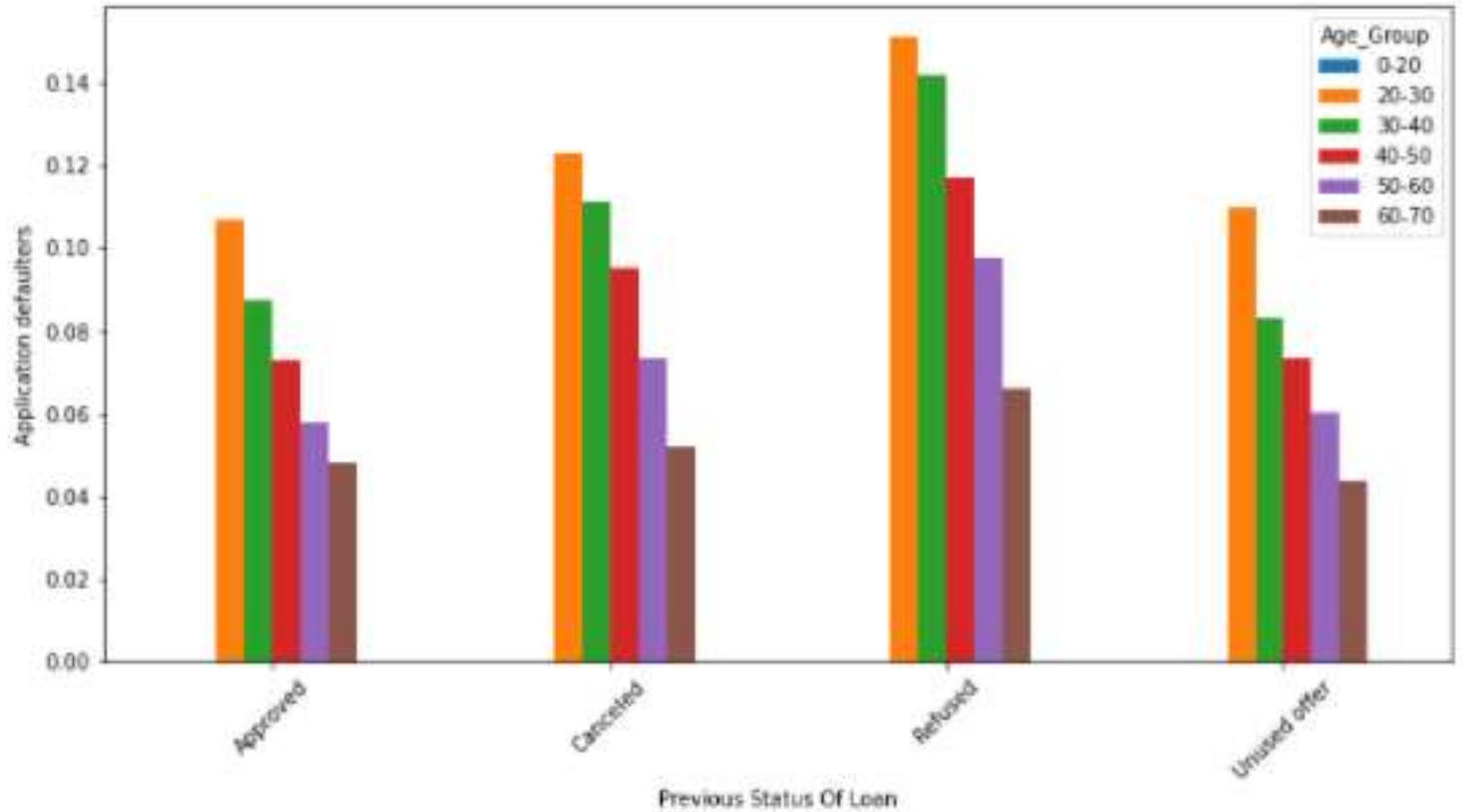
COMAPRING CONTRACT STATUS AND GENDER



INFERENCE

1.In all the cases here males can be seen more defaulted than females.

COMPARING CONTRACT STATUS AND AGE GROUP

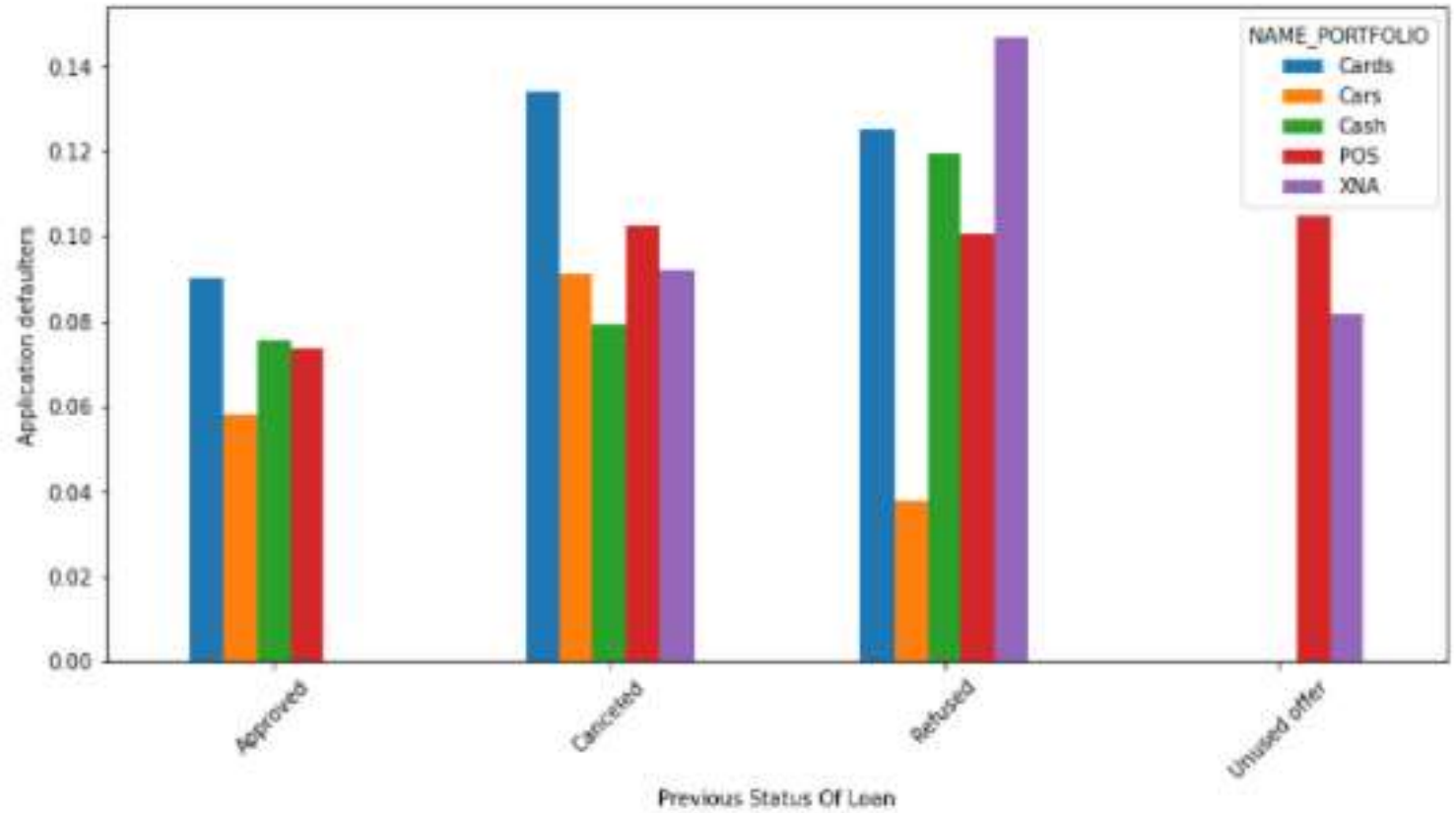


INFERENCES:

1.As we can see in all the cases higher percentage of young clients can be seen.

2.Refused clients are more likely to default than the previously approved clients.

Comparing NAME_CONTRACT_STATUS and NAME_PORTFOLIO



INFERENCE

1.Clients who previously applied for cards are most likely to default.
2.Also clients who have applied for XNA most of them are defaulters

FINAL RECOMMENDATIONS

Overall Recommendations

1. It is more safe to grant the loan to mid age clients and senior citizen clients with higher income.
2. Loan can be granted to highly educated clients because there is very less chance of them being a defaulter.
3. Overall females have less chance of being a defaulter than males so loan can be granted to them.
4. Married clients should also be given loans because it has less defaulter rate as compared to other family status.

Overall Risks

1. Clients with low income groups should be avoided because of higher chances of being defaulter.
2. Unemployed clients can also be a big risk factor for providing with loan.
3. External credit score should be also considered before approving the clients application as it consist of clients credit score.
4. Lower secondary and secondary educated clients should be avoided for loan as they have high defaulter rates.