



Credit Case Study

BY – G Ventakesh
Ashutosh Singh

Problem Statements

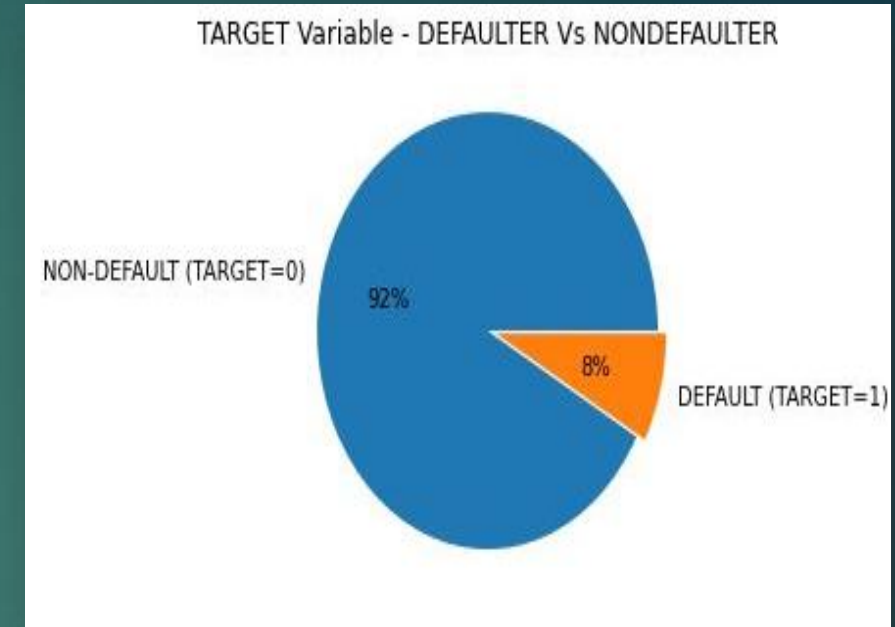
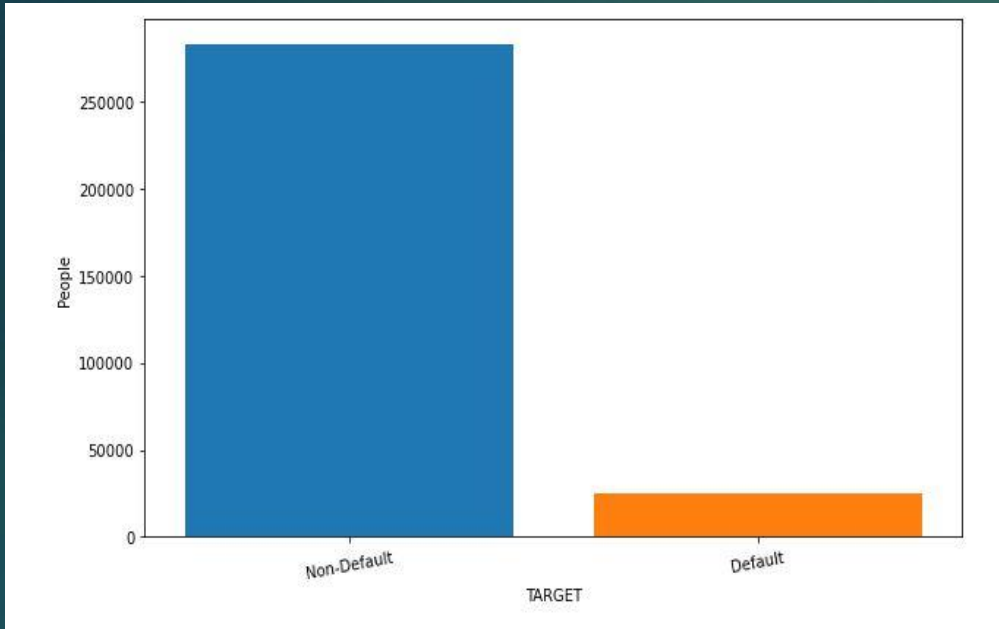
The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample.
- **All other cases:** All other cases when the payment is paid on time.
- 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.

Steps

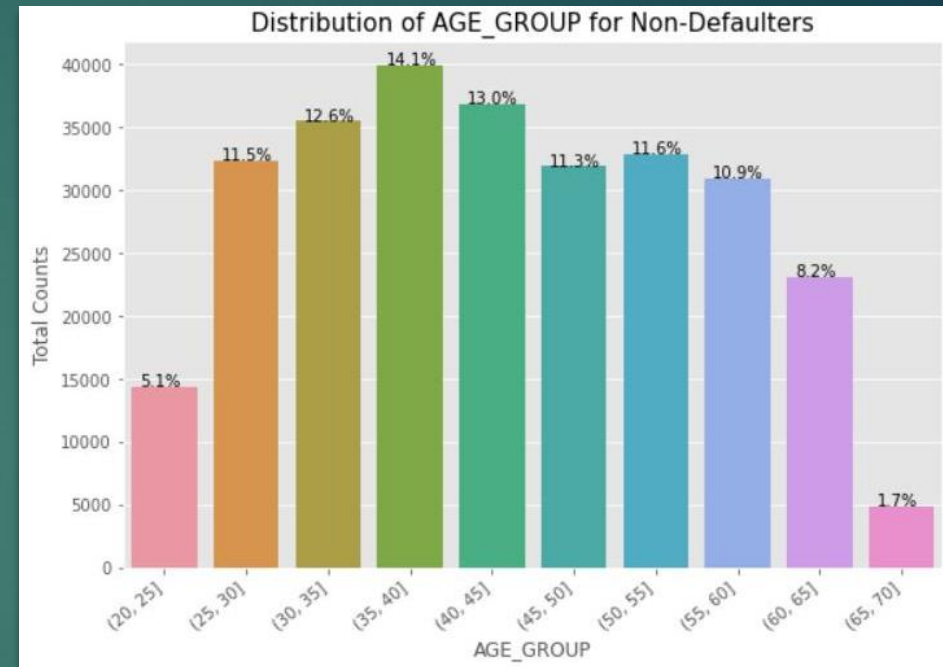
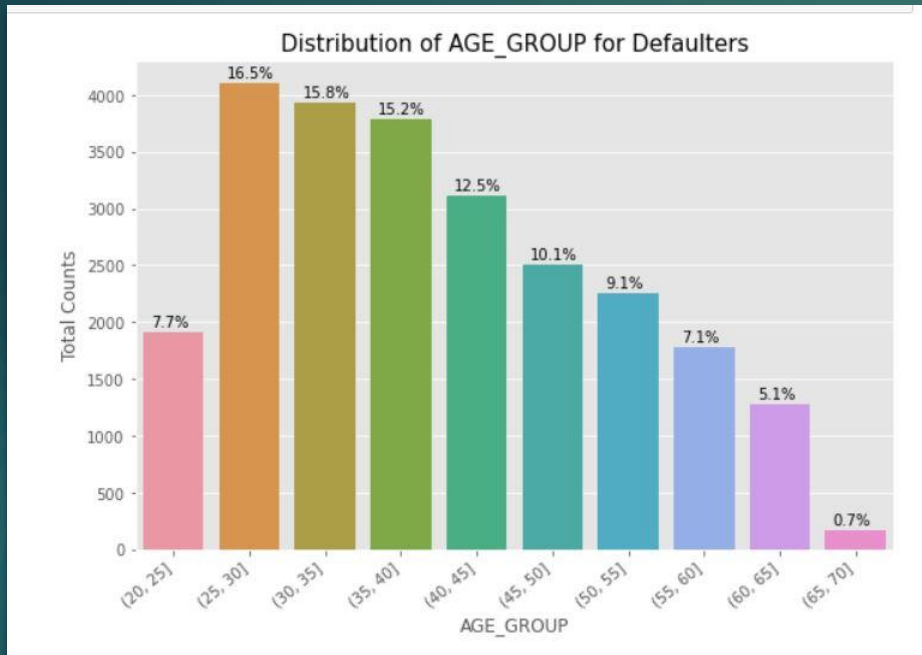
- 1.Dataunderstanding and Sourcing.
- 2.Check the data quality issues and binning.
- 3.Check the Data imbalance and univariate , segmented univarite and bivariate analysis ,correlation.
- 4.Merging application data with previous application data.
- 5.Data analysis by univariate , segmented univarite and bivariate analysis ,correlation.
- 6.Recommedation and Risks.

Target Variable

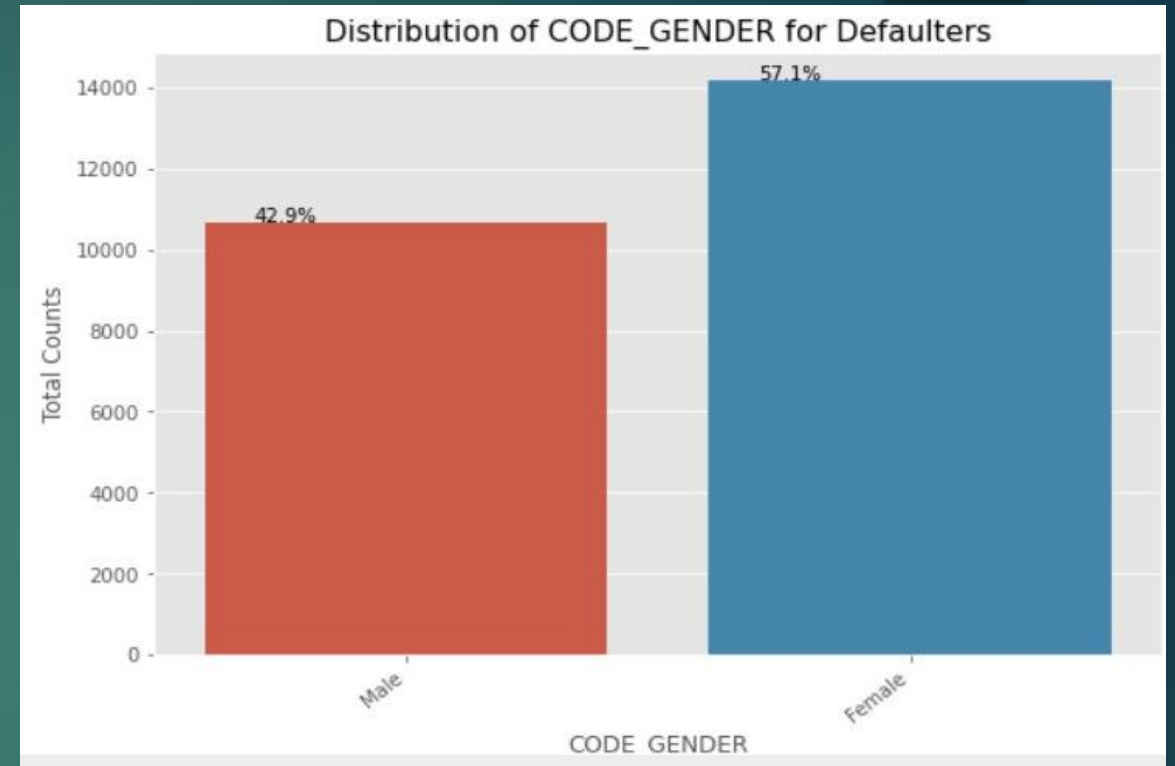
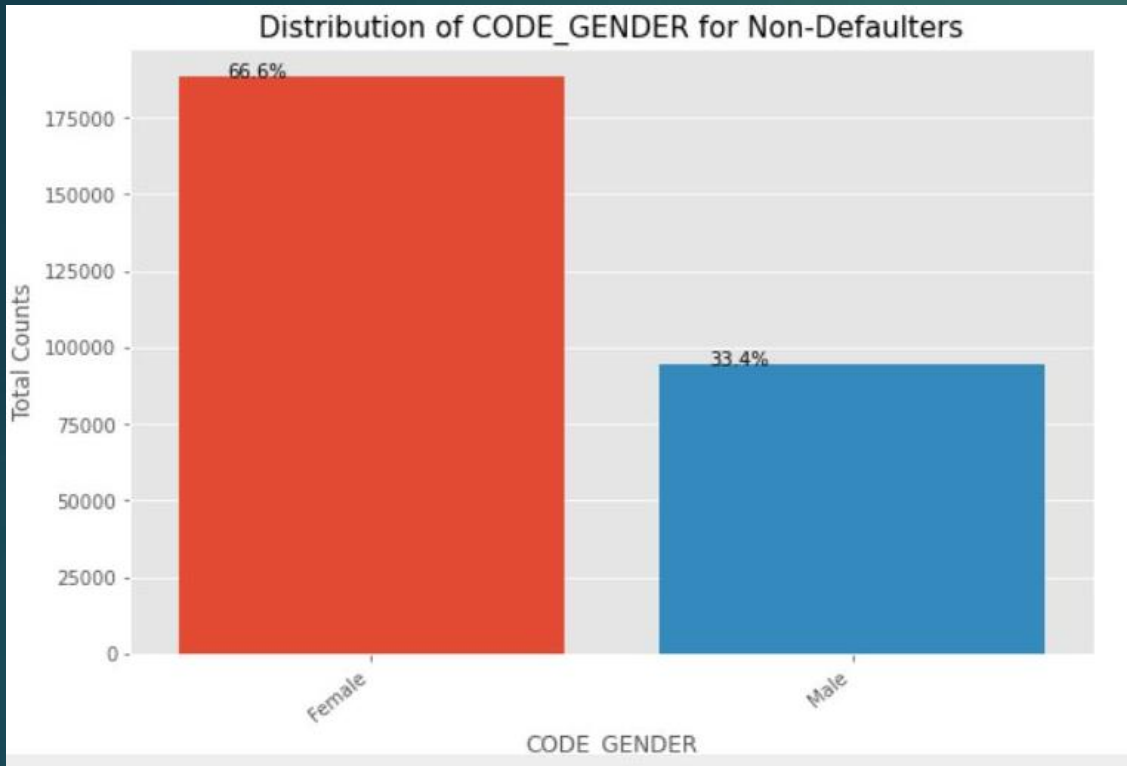


The target variable suggests the clients pay loan at time, Non-Default i.e. 0 where as clients who have difficulty to pay loan on time, they default i.e. 1. Here the Non-Default class and Default class has difference in their observations, so the data is very highly imbalanced.

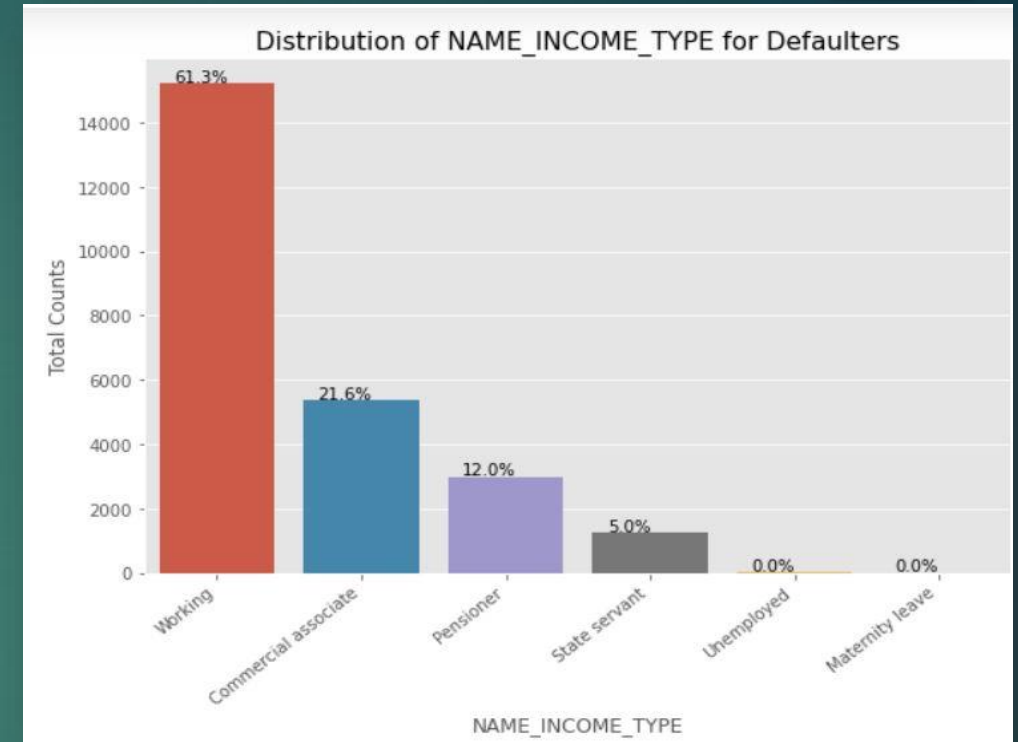
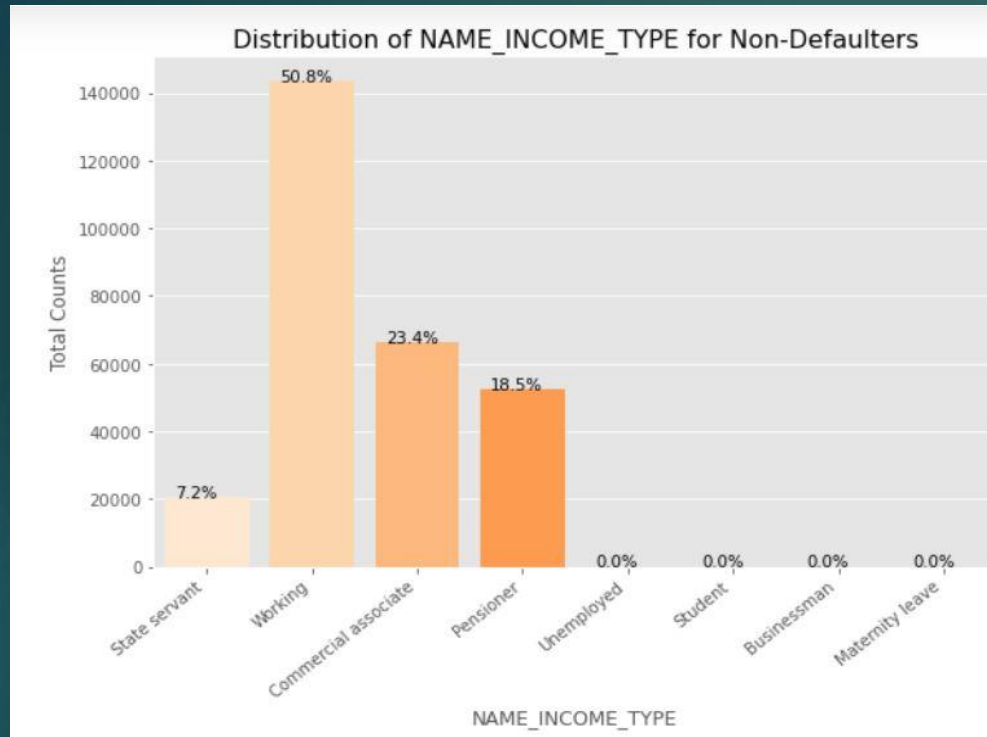
Univariate Analysis On Application Data



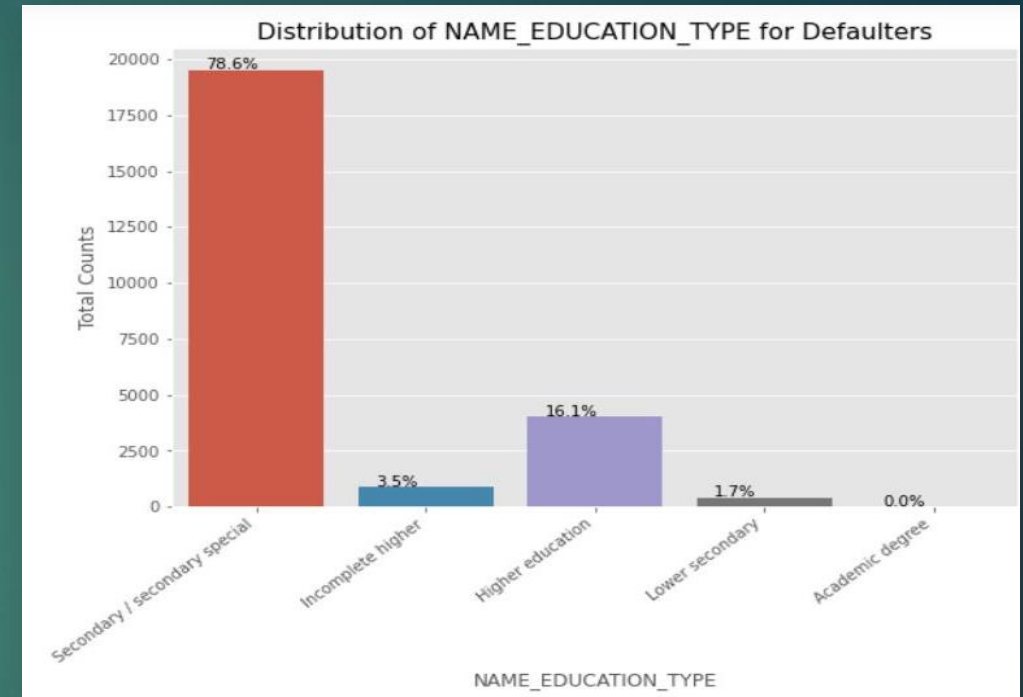
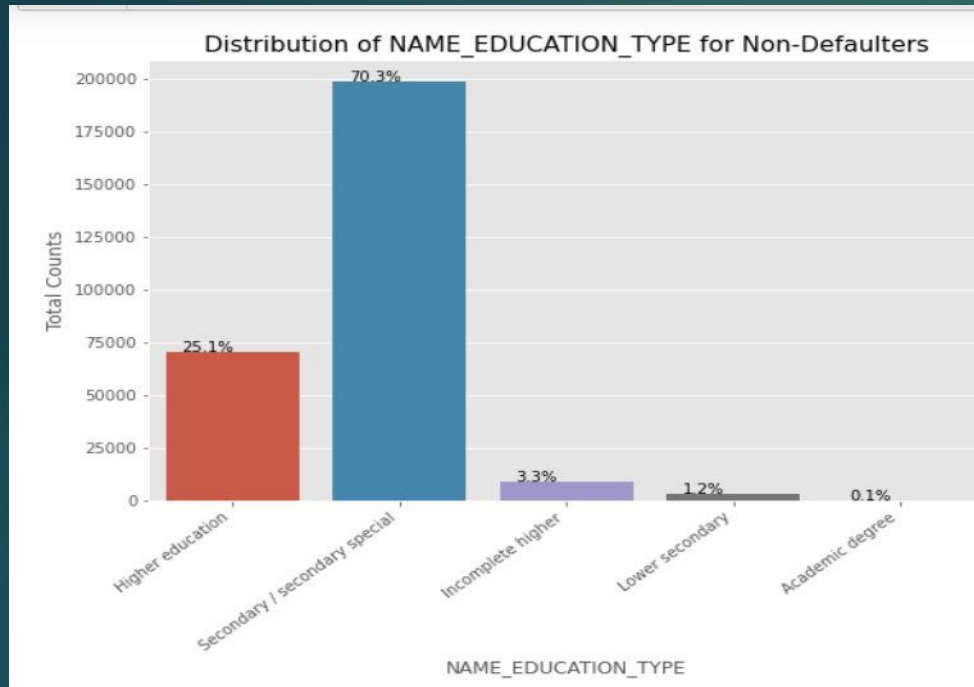
Age between 25 and 40 group the people are tends to default more. So this age group bank should not give the loan to. and as the age increasing the people tends to default less from starting the age of 25. Reason would be that the people are get employed around it and with increasing age there salary also increases



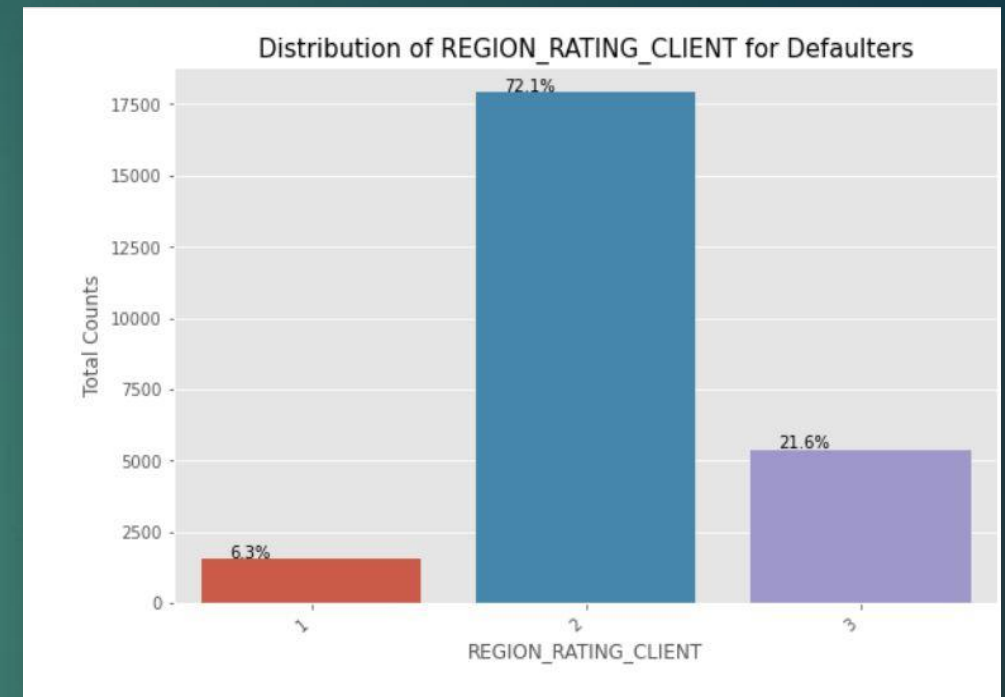
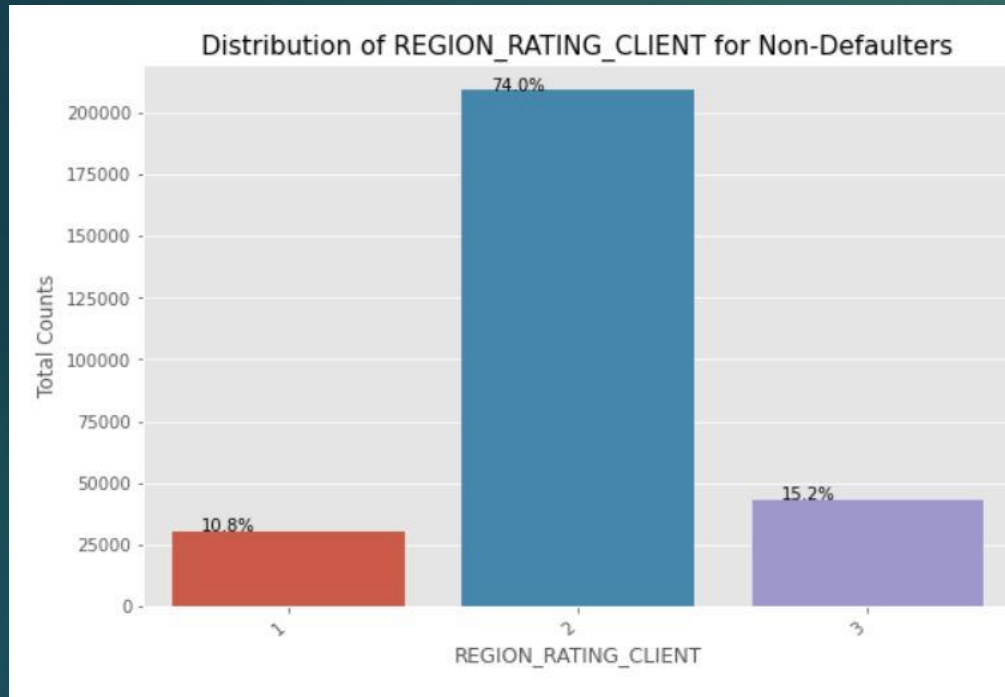
In Non-Defaulter the female has taking more loans while 57% for defaulter. we can say that that more number of female are applying for loans so more number of female are defaulting and graph show that the number of female defaulting is much lower than the male



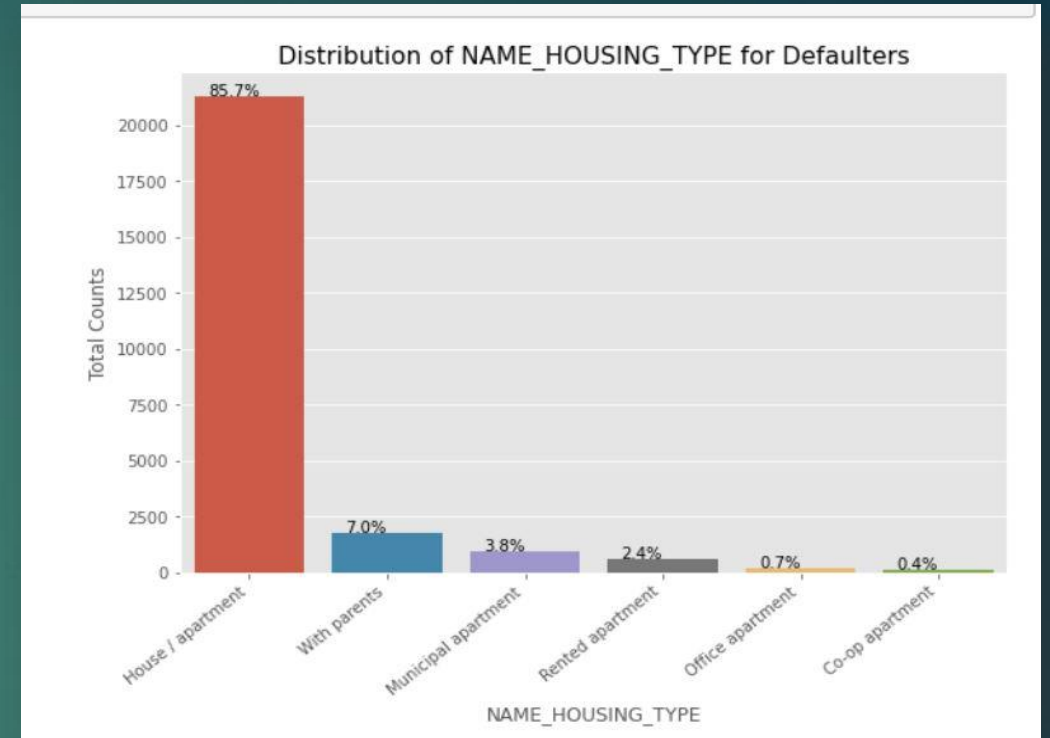
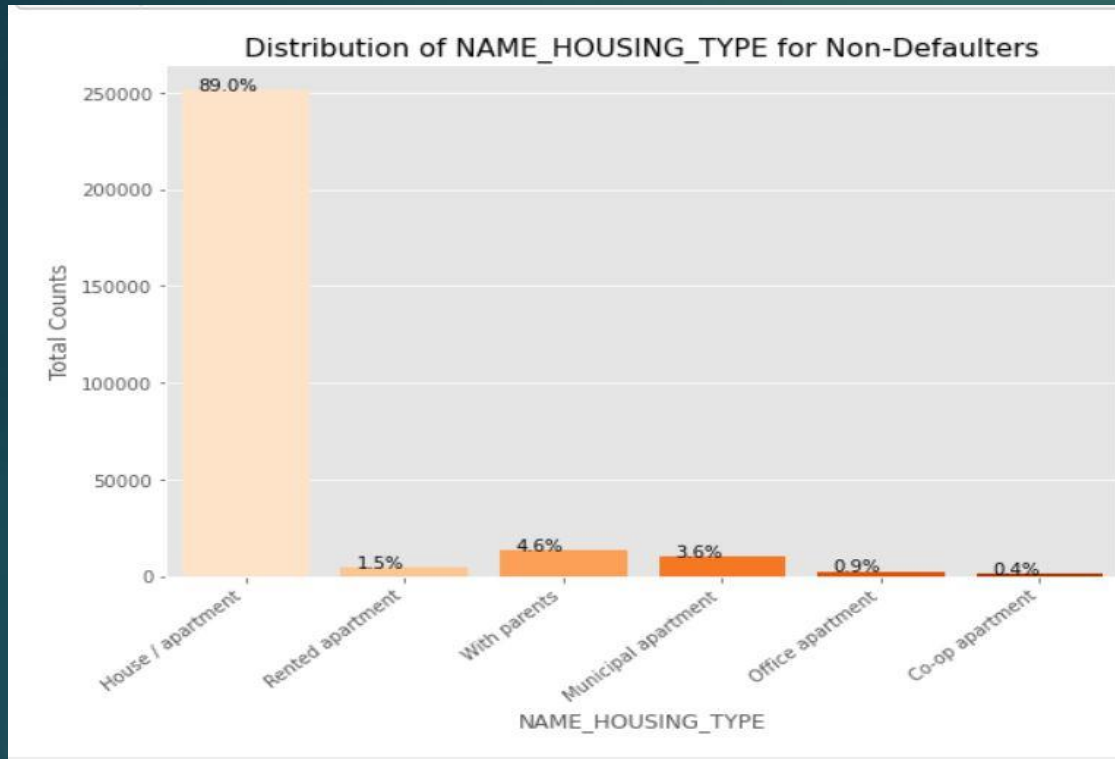
In Defaulter graph there is no column for student and businessmen reason is because that student didn't have to pay during the studying and businessmen never default. Many are from working class and 50.8% are to non-defaulter while in 61.3% are defaulter.



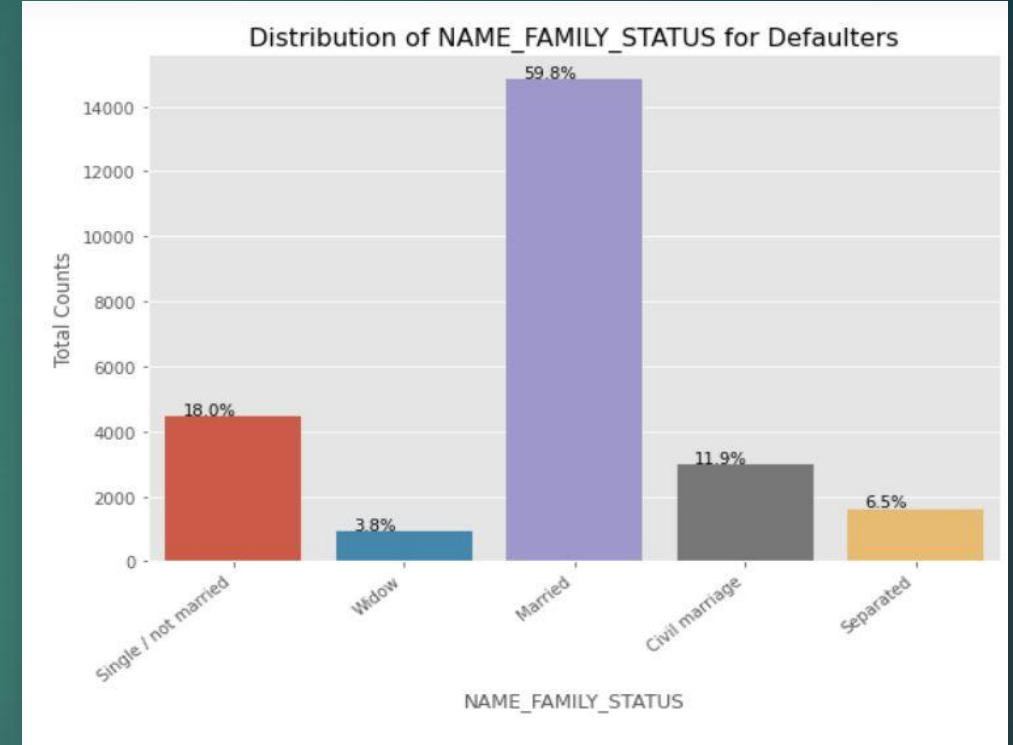
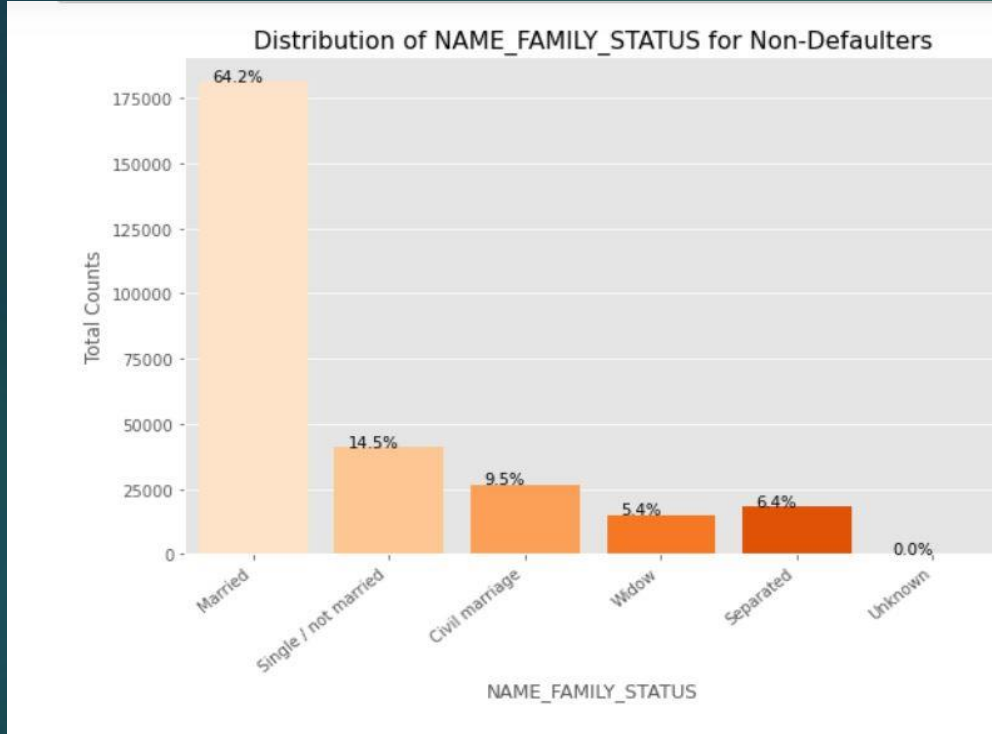
Higher education clients of making default are more as shown in the graph and other than the higher education all the education categories has equal chances to default.



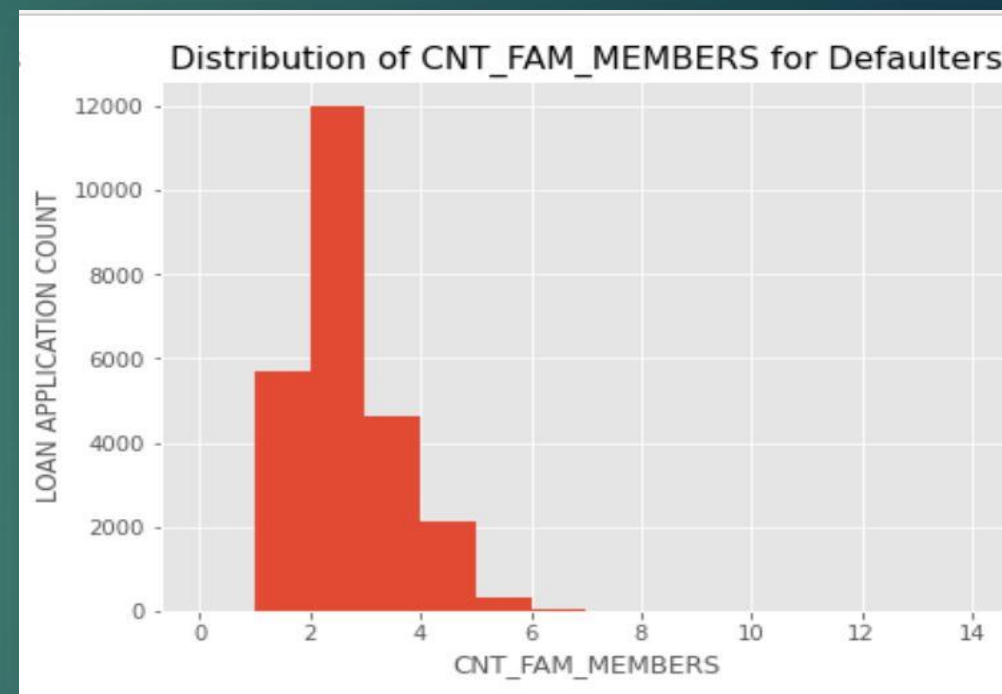
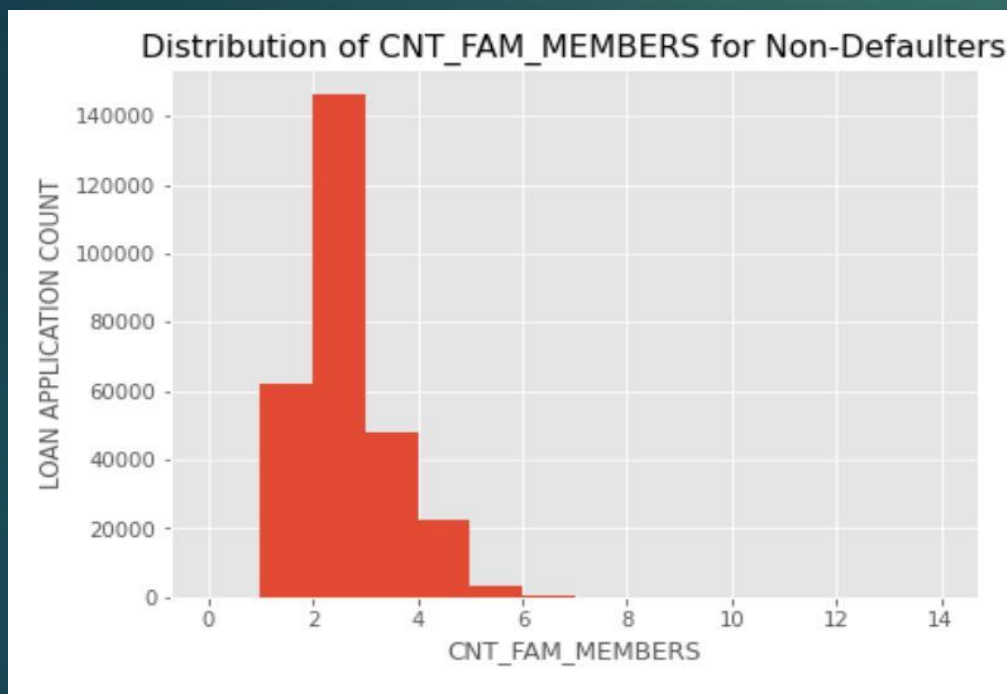
Graph show that more number of clients are applying for the loan are from 2 tier region and people living in 3 tier are tends to default more by their weightage.



The graph show that the people who have /Apartment are more likely to apply for more loans.people living with parents are more likely to default when it is compared to other and the reason could be their living expenses are more due to their parents living with them.



Married people are interested to take the loan more in comparative to other and graph also show that the single/not married people are 14.5% from non defaulter and the defaulter have 18% respectively so there is more risk with them.



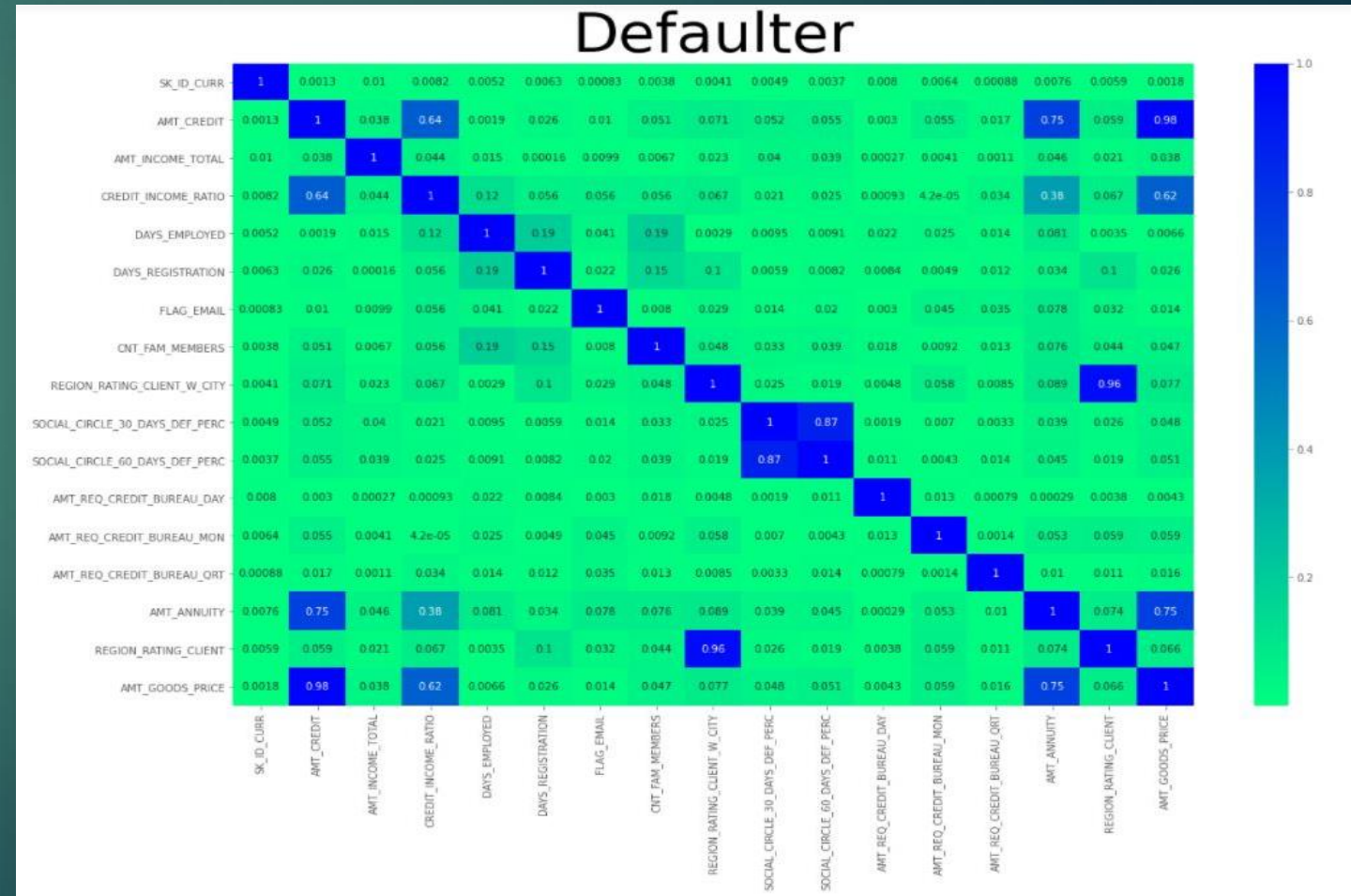
The clients which have 3 children are likely to apply for loans than other families. With 3 children they aslo default more.

Correlation of Continuous variable of Defaulter DataFrame

1) THERE IS HIGH CORRELATION BETWEEN THE AMT_GOOD_PRICE AND AMT_CREDITY i.e 0.98

2) THERE IS HIGH CORRELATION BETWEEN THE REGION_RATING_CLIENT AND REGION_RATING_CLIENT_W_CITY i.e 0.96

3) THERE IS HIGH CORRELATION BETWEEN THE AMT_CREDIT AND AMT_ANNUNITY i.e 0.75

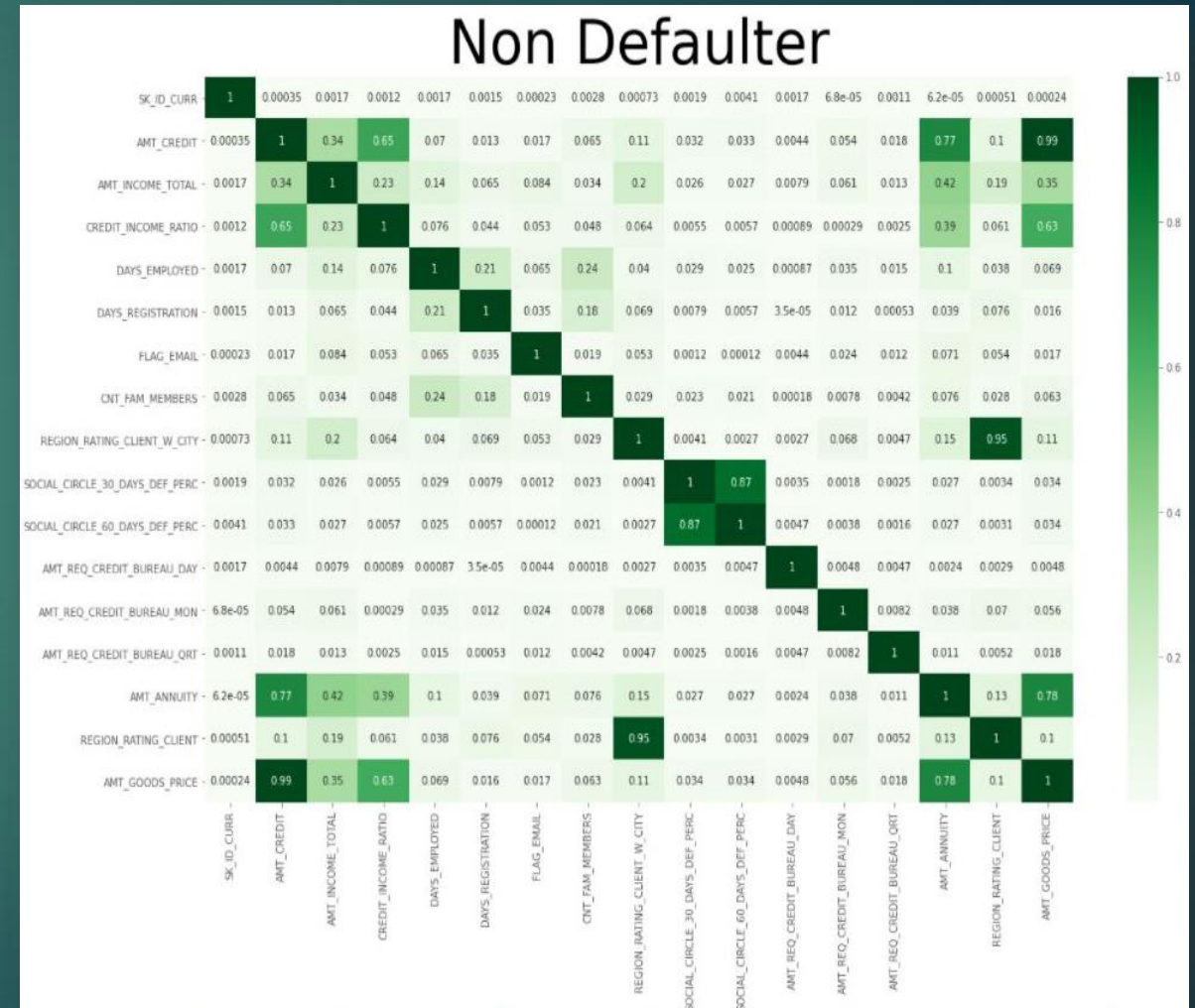


Correlation of Continuous variable of Non-Defaulter DataFrame

1) THERE IS HIGH CORRELATION BETWEEN THE AMT_GOOD_PRICE AND AMT_CREDITY i.e 0.99

2) THERE IS HIGH CORRELATION BETWEEN THE REGION_RATING_CLIENT AND REGION_RATING_CLIENT_W_CITY i.e 0.95

3) THERE IS HIGH CORRELATION BETWEEN THE AMT_CREDIT AND AMT_ANNUNITY i.e 0.77



Top 10 Correlation For Defaulter and Non-Defaulter

| | column1 | column2 | correlation | abs_correlation |
|-----|--------------------------------|--------------------------------|-------------|-----------------|
| 308 | AMT_GOODS_PRICE | AMT_CREDIT | 0.982783 | 0.982783 |
| 297 | REGION_RATING_CLIENT | REGION_RATING_CLIENT_W_CITY | 0.956637 | 0.956637 |
| 208 | SOCIAL_CIRCLE_60_DAYS_DEF_PERC | SOCIAL_CIRCLE_30_DAYS_DEF_PERC | 0.874562 | 0.874562 |
| 321 | AMT_GOODS_PRICE | AMT_ANNUITY | 0.752295 | 0.752295 |
| 272 | AMT_ANNUITY | AMT_CREDIT | 0.752195 | 0.752195 |
| 74 | CREDIT_INCOME_RATIO | AMT_CREDIT | 0.639744 | 0.639744 |
| 310 | AMT_GOODS_PRICE | CREDIT_INCOME_RATIO | 0.623100 | 0.623100 |
| 274 | AMT_ANNUITY | CREDIT_INCOME_RATIO | 0.381298 | 0.381298 |
| 113 | DAYS_REGISTRATION | DAYS_EMPLOYED | 0.192455 | 0.192455 |
| 149 | CNT_FAM_MEMBERS | DAYS_EMPLOYED | -0.186515 | 0.186515 |

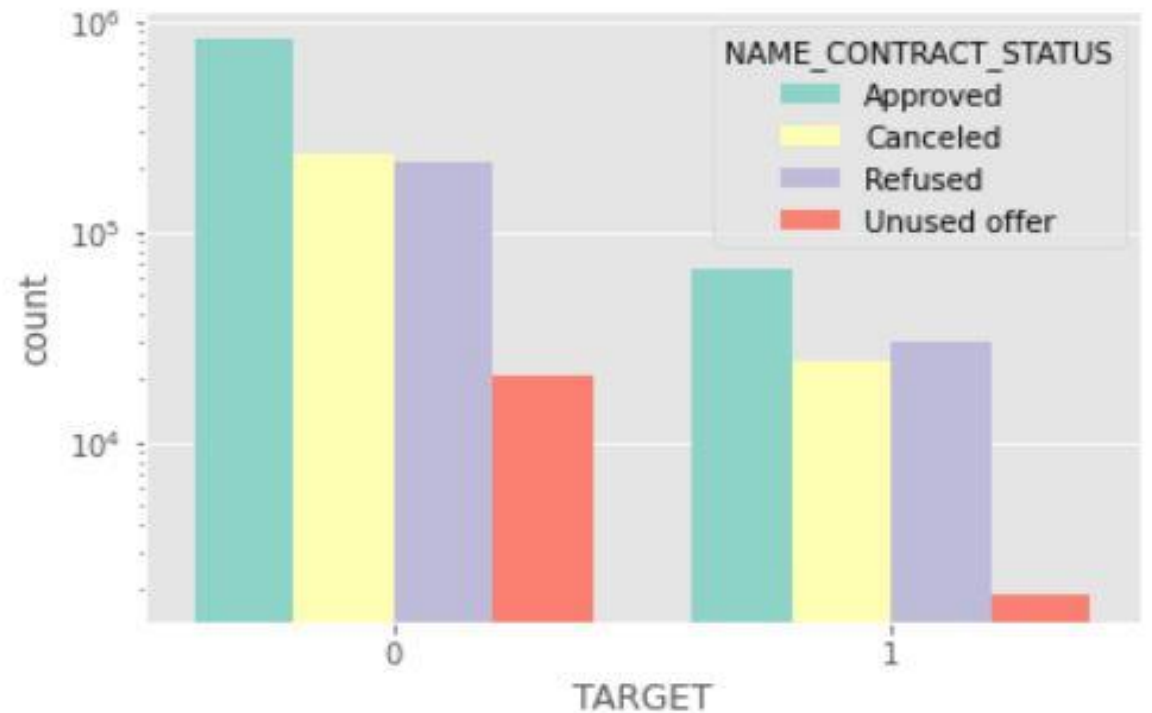
The above dataframe shows top 10 correlation for clients with payment difficulties

| | column1 | column2 | correlation |
|-----|--------------------------------|--------------------------------|-------------|
| 308 | AMT_GOODS_PRICE | AMT_CREDIT | 0.987024 |
| 297 | REGION_RATING_CLIENT | REGION_RATING_CLIENT_W_CITY | 0.950148 |
| 208 | SOCIAL_CIRCLE_60_DAYS_DEF_PERC | SOCIAL_CIRCLE_30_DAYS_DEF_PERC | 0.873003 |
| 321 | AMT_GOODS_PRICE | AMT_ANNUITY | 0.776421 |
| 272 | AMT_ANNUITY | AMT_CREDIT | 0.771296 |
| 74 | CREDIT_INCOME_RATIO | AMT_CREDIT | 0.648589 |
| 310 | AMT_GOODS_PRICE | CREDIT_INCOME_RATIO | 0.628732 |
| 273 | AMT_ANNUITY | AMT_INCOME_TOTAL | 0.418949 |
| 274 | AMT_ANNUITY | CREDIT_INCOME_RATIO | 0.391498 |
| 309 | AMT_GOODS_PRICE | AMT_INCOME_TOTAL | 0.349425 |

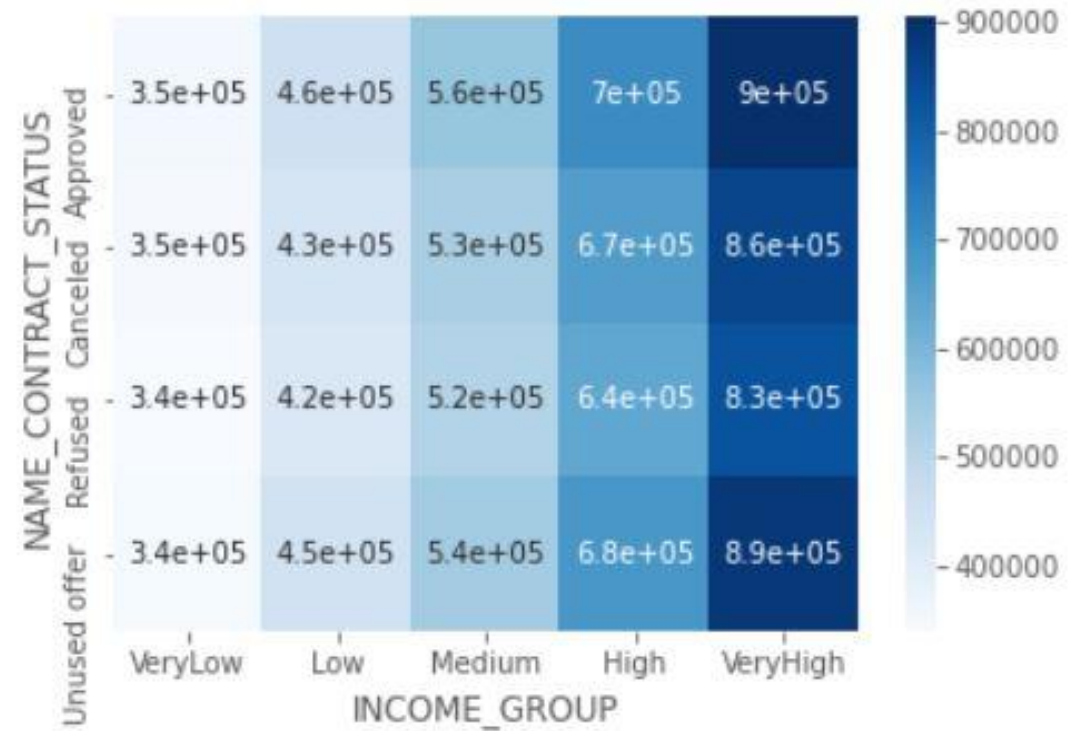
The above dataframe shows top 10 correlation for clients with No payment difficulties

Analysis with joining the previous application data

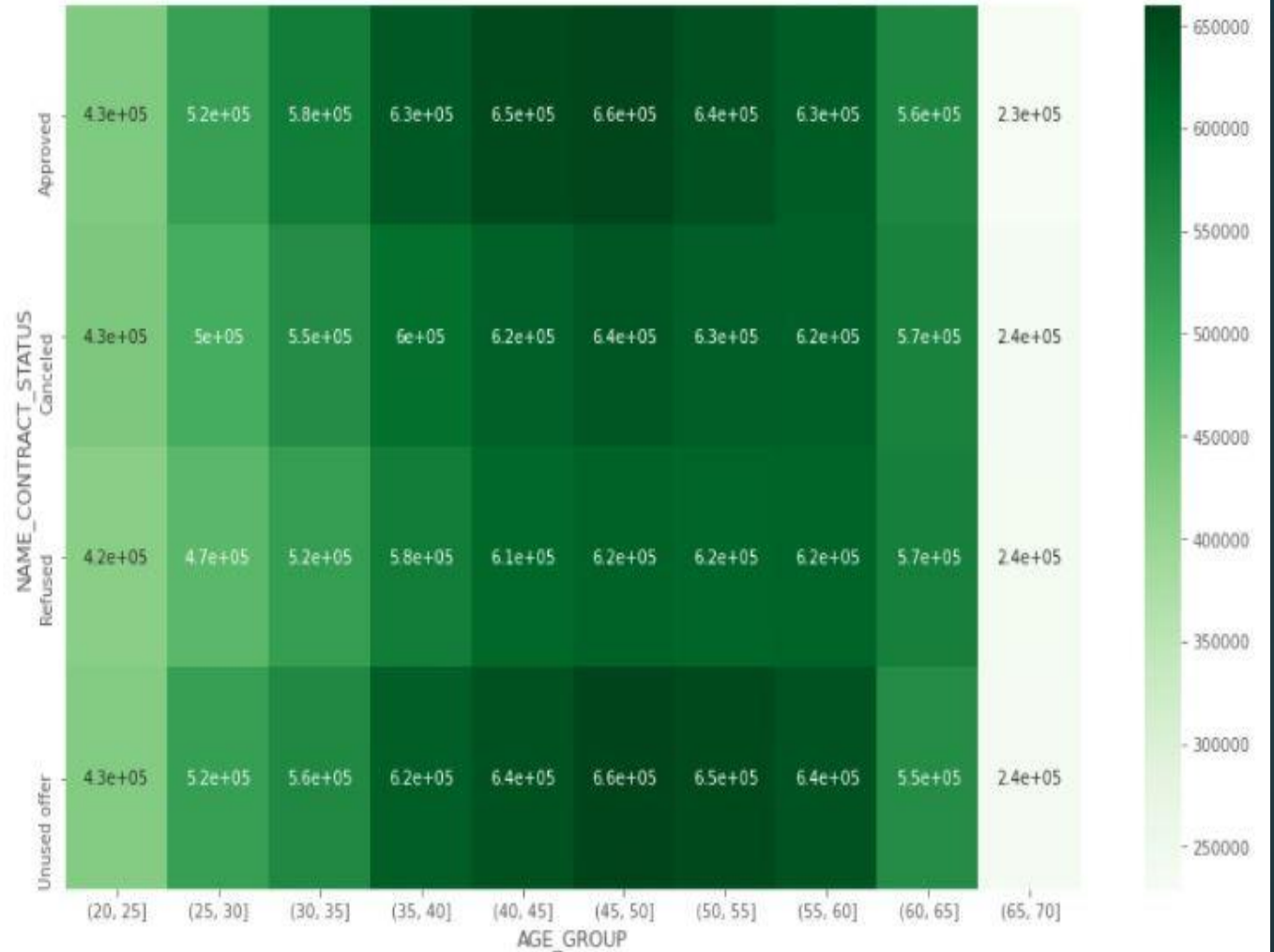
Payment with difficulties(1) the client are not much approved by bank in comparsion to payment at time (0)



Those in High Income and Very High Income groups are most preferred for credit in previous years- because of higher repayment capacity



Those between Age Groups 40-60 are most preferred for credit in previous years- because of higher repayment capacity.



Recommended group where loan can be credited(Less chance to be defaulter)

- Bank or insitution give the loans to those clients who are working as state servant.
- Client with high income category.
- Old female clients.
- Client with higher education(female).
- Old age client of any income group.
- Any client who`s previous loan was approved.
- Widow who has unused previous loan status.

Risky Group

- Single /not married are more likely to default.
- Male clients with civil marriage.
- Previously refused loan status group.
- Age group between the 25 to 45 more people to default.
- Family which has 3 children are more likely to default.

THANK YOU