Presentation
On
Credit EDA Case Study

Umesh Pandey

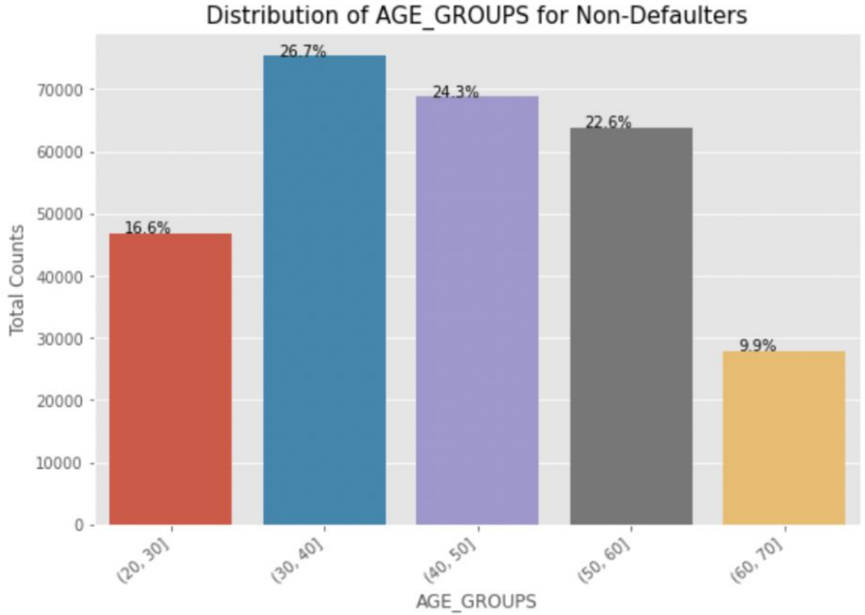
Purpose

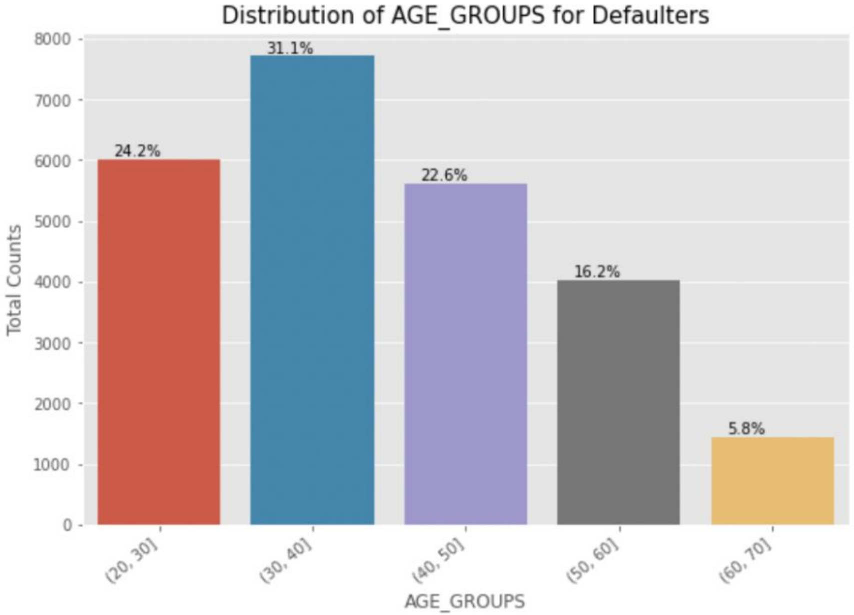
We have to understand how the bank approves or refuses any loan. Find out the different patterns and represent the outcomes to help the bank reduce the credit risk and interest risk.

Steps Involved:

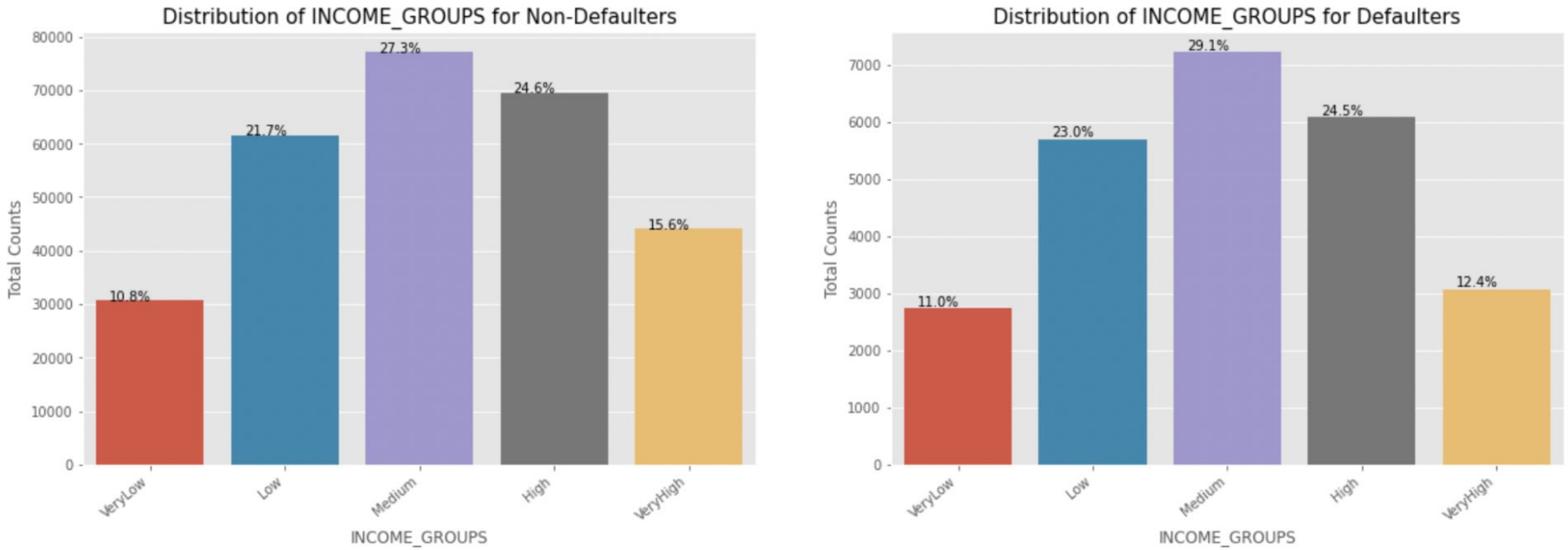
1. Data Understanding and sourcing.
2. Check for data quality issues and binning.
3. Check for data imbalance.
4. Check correlation, perform univariate, bivariate and multivariate analysis.
5. Merge both the datasets - application.csv and previous_application.csv
6. Perform univariate, bivariate and multivariate analysis.
7. Identify risks involved and provide recommendation

Univariate Analysis

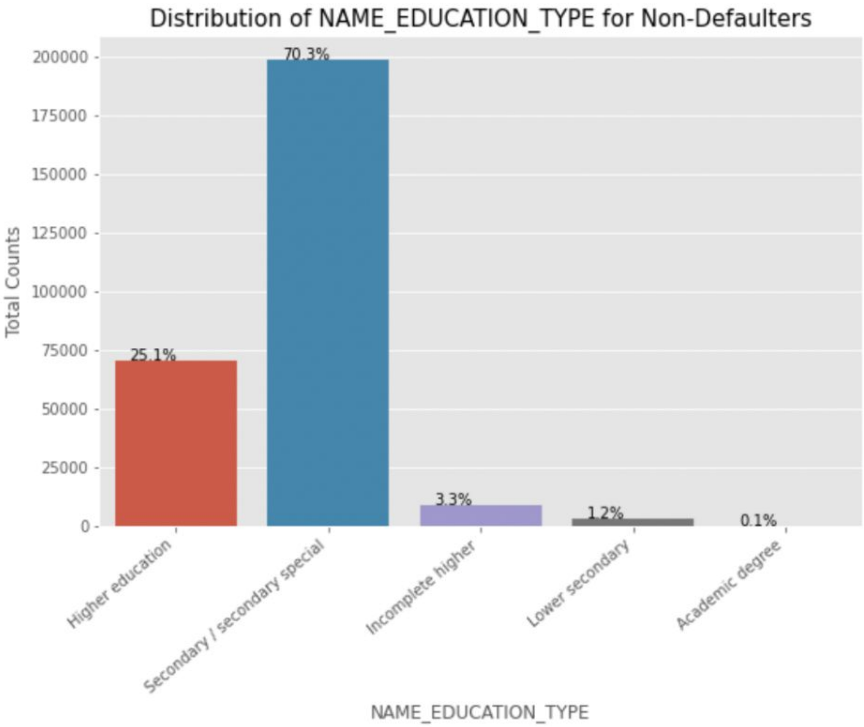


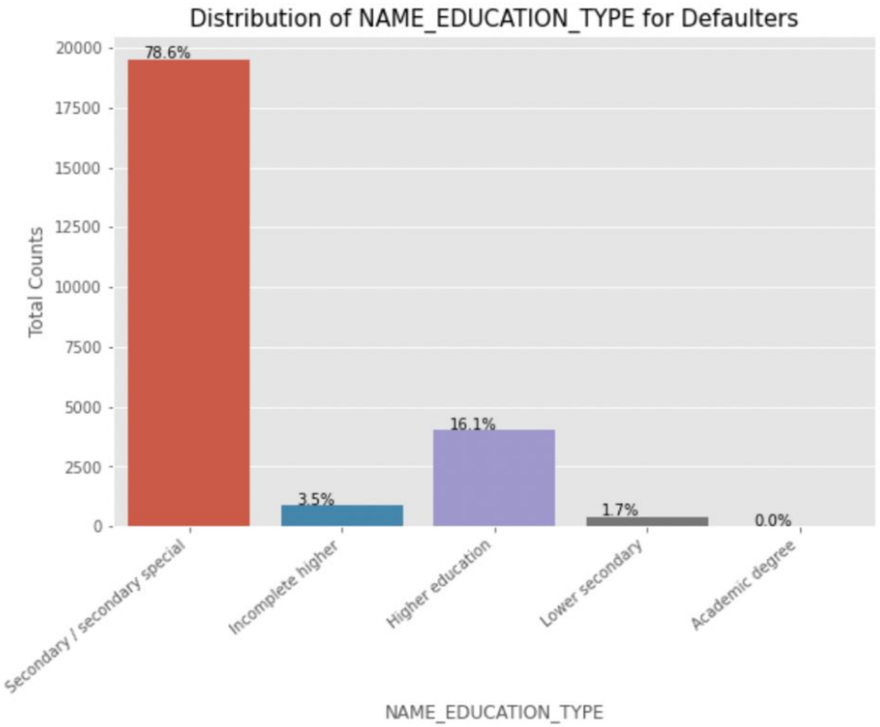


We see that (30,40) age group tend to default more often. So they are the riskiest people to loan to. With increasing age group, people tend to default less starting from the age 40. One of the reasons could be they get employed and settled around that age and with increasing age, their salary also increases.

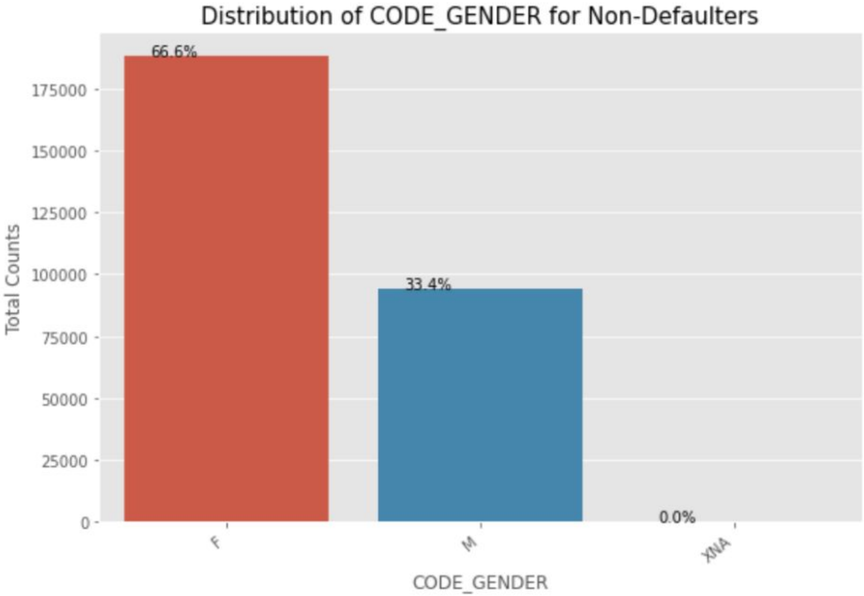


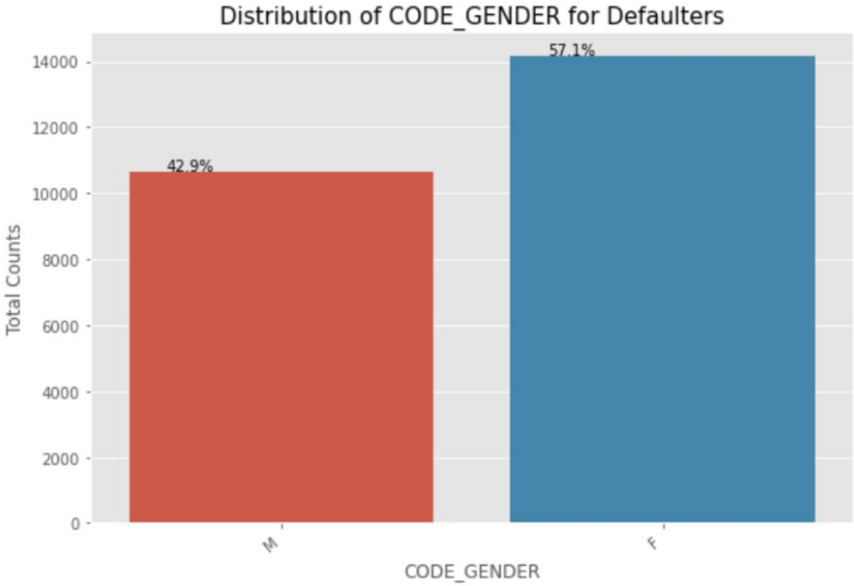
The Very High income group tend to default less often. They contribute 12.4% to the total number of defaulters, while they contribute 15.6% to the Non-Defaulters.





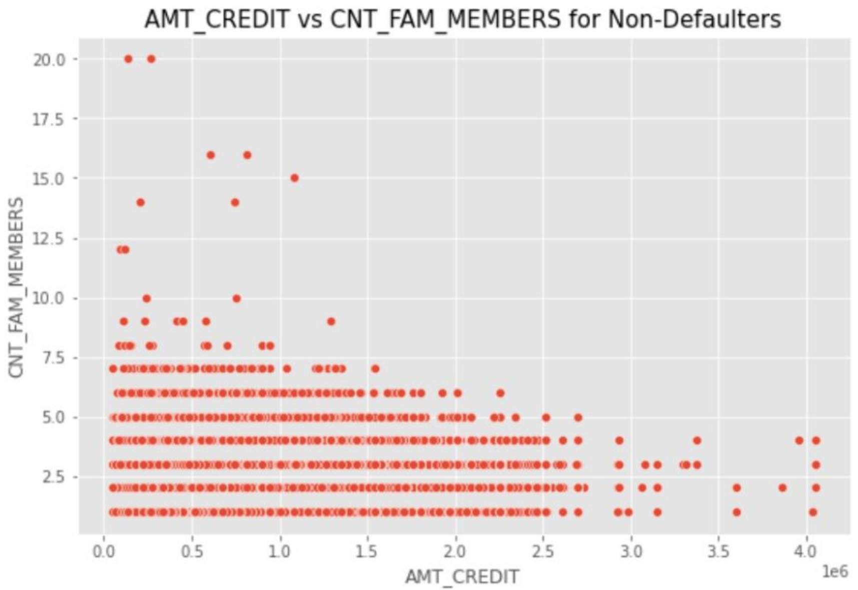
Almost all of the Education categories are equally likely to default except for the higher educated ones who are less likely to default and secondary educated people are more likely to default.

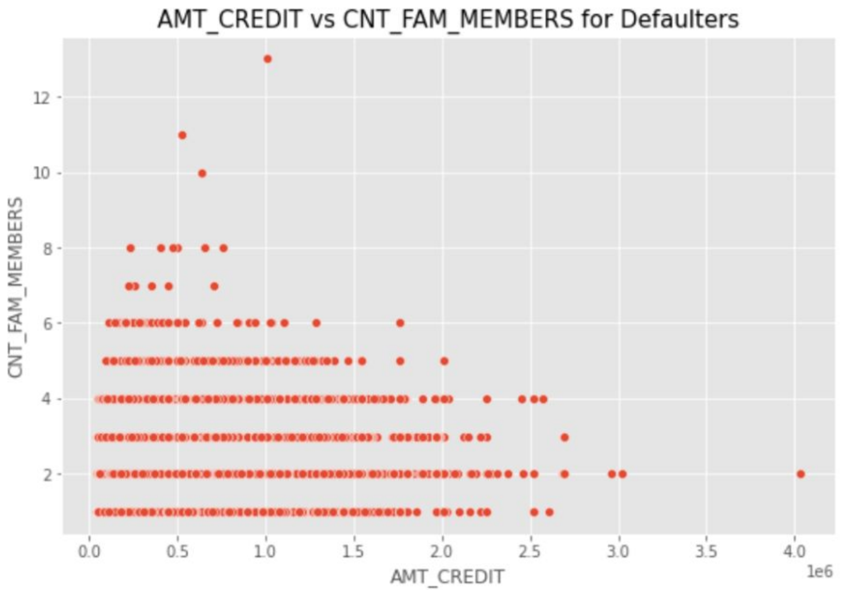




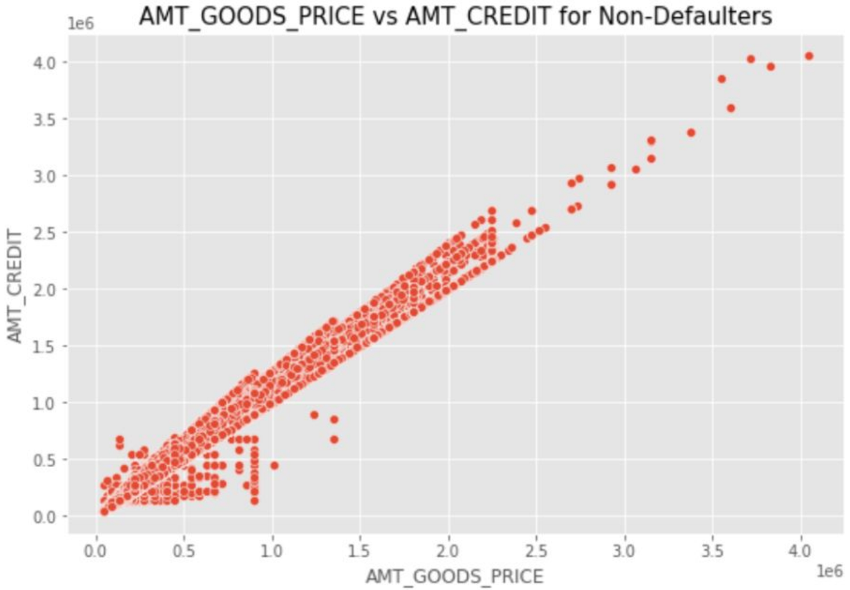
We can see that Females contribute 67% to the non-defaulters while 57% to the defaulters. We can conclude that we see more female applying for loans than males and hence the more number of female defaulters as well. But the rate of default of FEMALE is much lower compared to their MALE counterparts. Additionally we can drop the data of XNA as it's very minimal and won't affect our analysis.

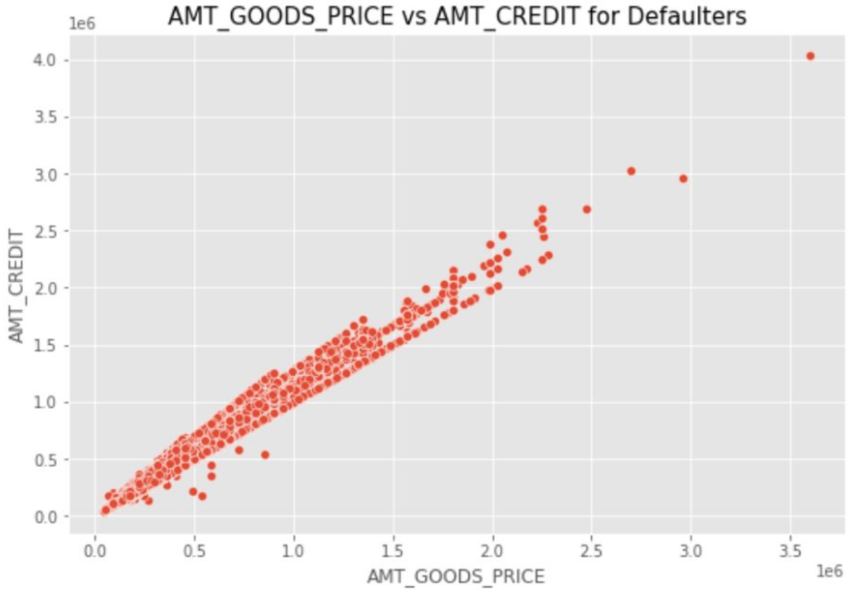
Bivariate Analysis





We can see that the density in the lower left corner is similar in both the case, so the people are equally likely to default if the family is small and the AMT_CREDIT is low. We can observe that larger families and people with larger AMT_CREDIT default less often.





AMT_CREDIT and AMT_GOODS_PRICE have linear relation. For lower range of AMT_CREDIT and AMT_GOODS_PRICE, amount of defaulters is less than that of non-defaulters

Analysis of Merged Dataset

We see that car ownership doesn't have any effect on application approval or rejection. But we saw earlier that the people who has a car has lesser chances of default. The bank can add more weightage to car ownership while approving a loan amount.

Target variable (0 - Non Defaulter 1 - Defaulter)

We can see that the people who were approved for a loan earlier, defaulted less often, where as people who were refused a loan earlier have higher chances of defaulting.

Risks:

1. Previous refused loan groups.
2. Lower Secondary educated groups.
3. People living with parents tend to default more.
4. Working class clients as students and businessmen default less.

Suggestions:

Loans can be provided to :

1. Old people above 40.
2. Client with higher education.
3. Client’s whose previous loans were approved.
4. Client with high income category.
5. Client having car.

Thank You