Credit EDA Case Study

By : Karishma Sahay

Problem Statement

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history.
- 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.
- This case study aims to identify patterns which indicate if a client has difficulty paying their instalments 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.

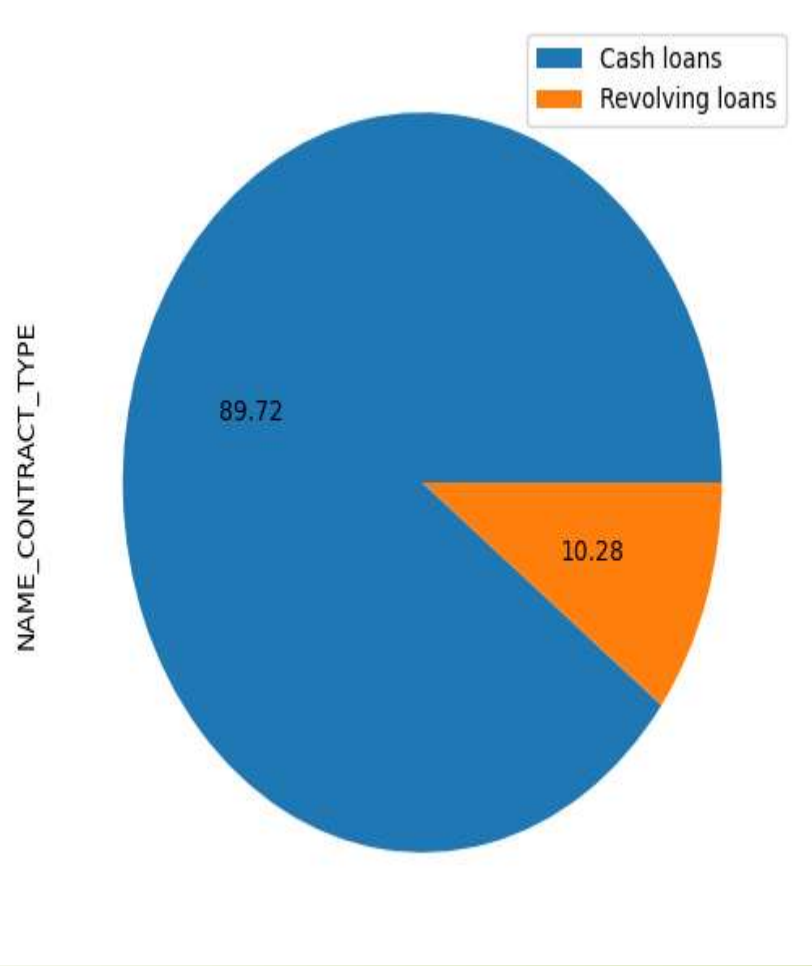
Steps

- Data Understanding and Data Sourcing
- Data Loading
- Data Inspection
- Data Cleaning
- Graphs and Insights
- Univariate Analysis
- Outlier Treatment
- Bivariate Analysis
- Multivariate Analysis
- Recommendation and Conclusion

Univariate analysis Graphs and Insights

- NAME_CONTRACT_TYPE

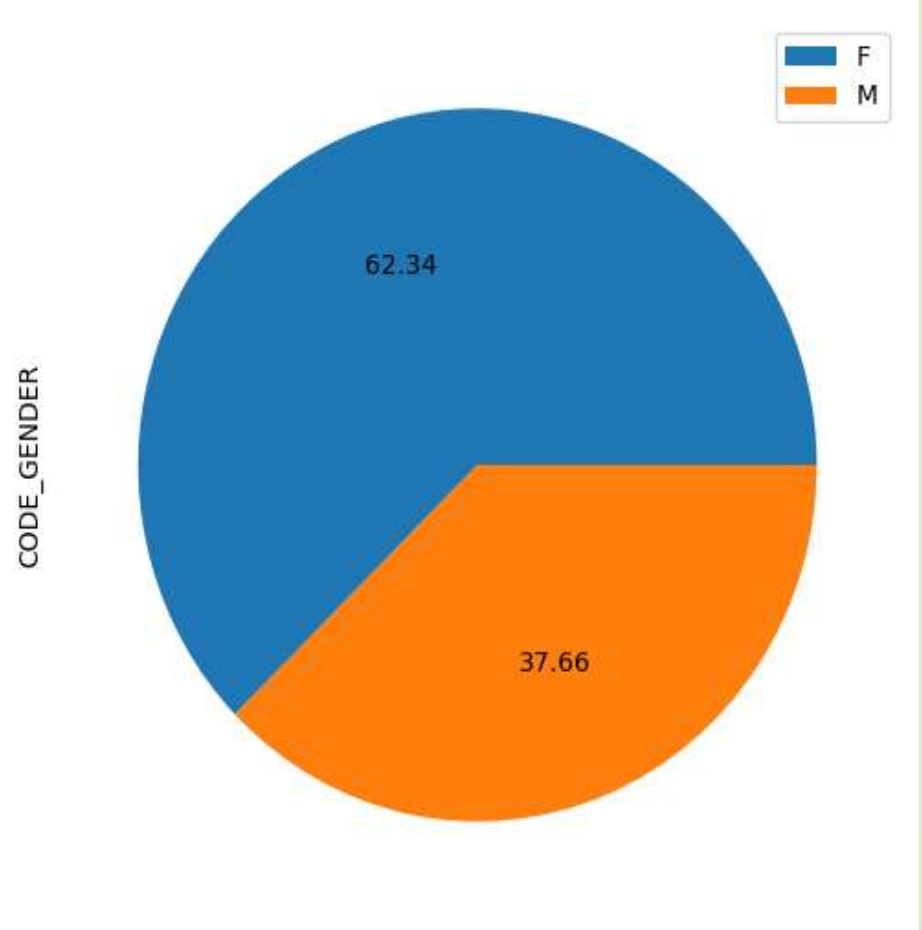
1. Cash loans – 89.72%
2. Revolving loans – 10.28%
3. Hence cash loans are greater than revolving loans



Univariate analysis Graphs and Insights

- CODE_GENDER

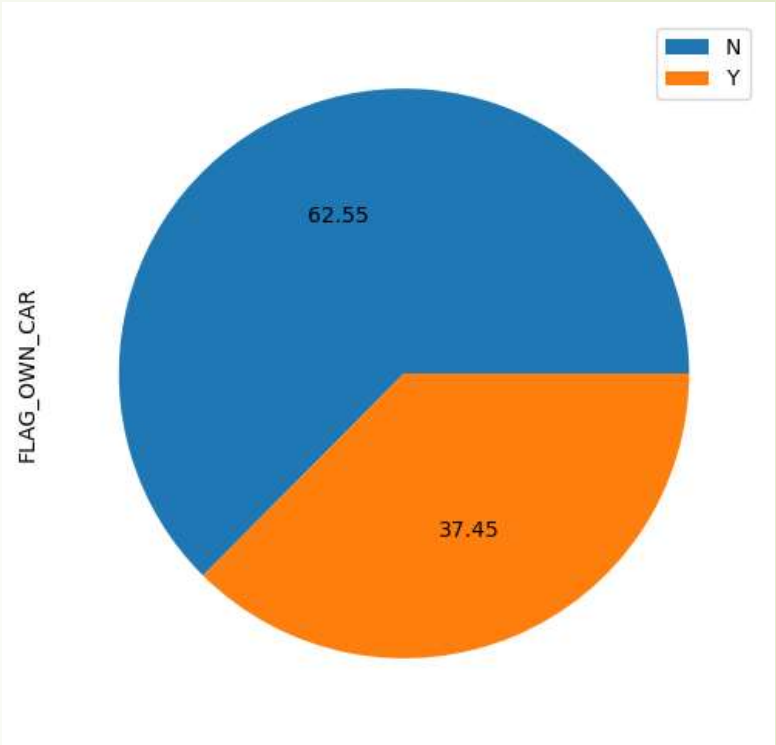
1. Females – 62.34
2. Male – 37.66
3. Females have higher chances of getting loan.



Univariate analysis Graphs and Insights

- **FLAG_OWN_CAR**

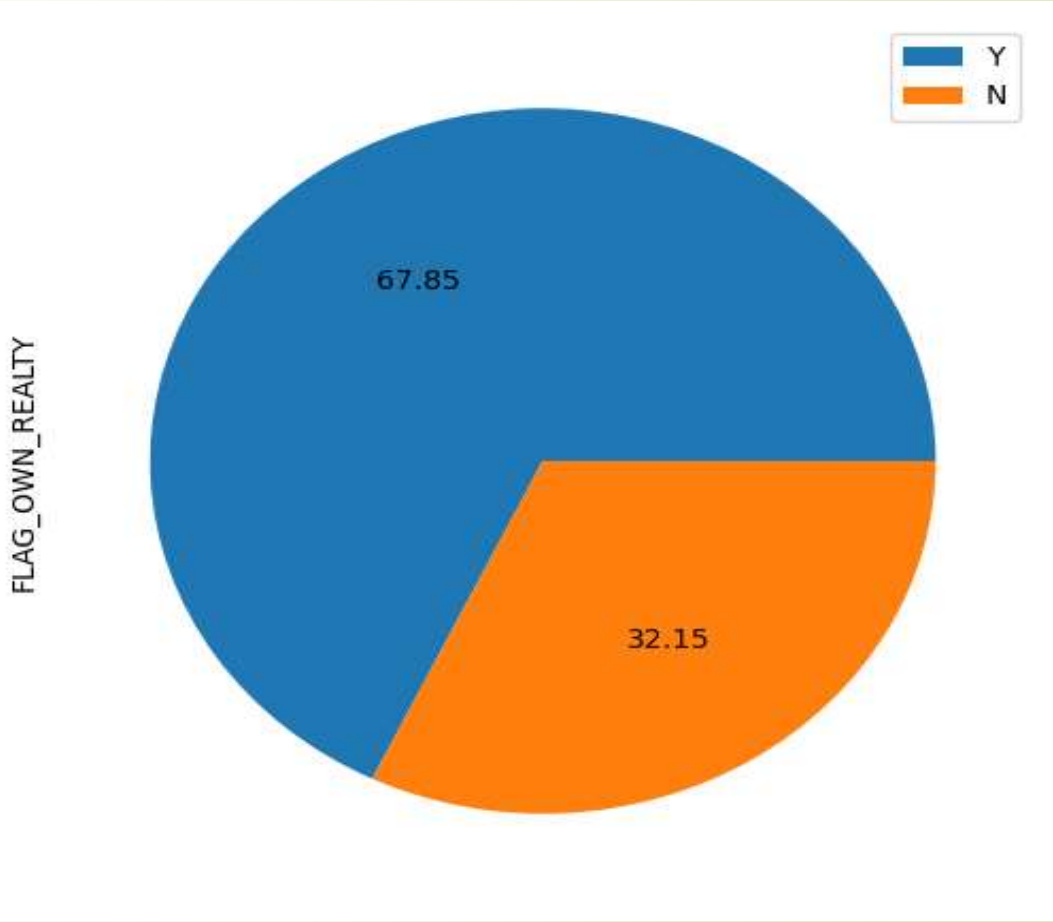
1. People having car have more chances to get loan



Univariate analysis Graphs and Insights

- **FLAG_OWN_REALITY**

1. People those who have there own house ,have higher chances to get loan.

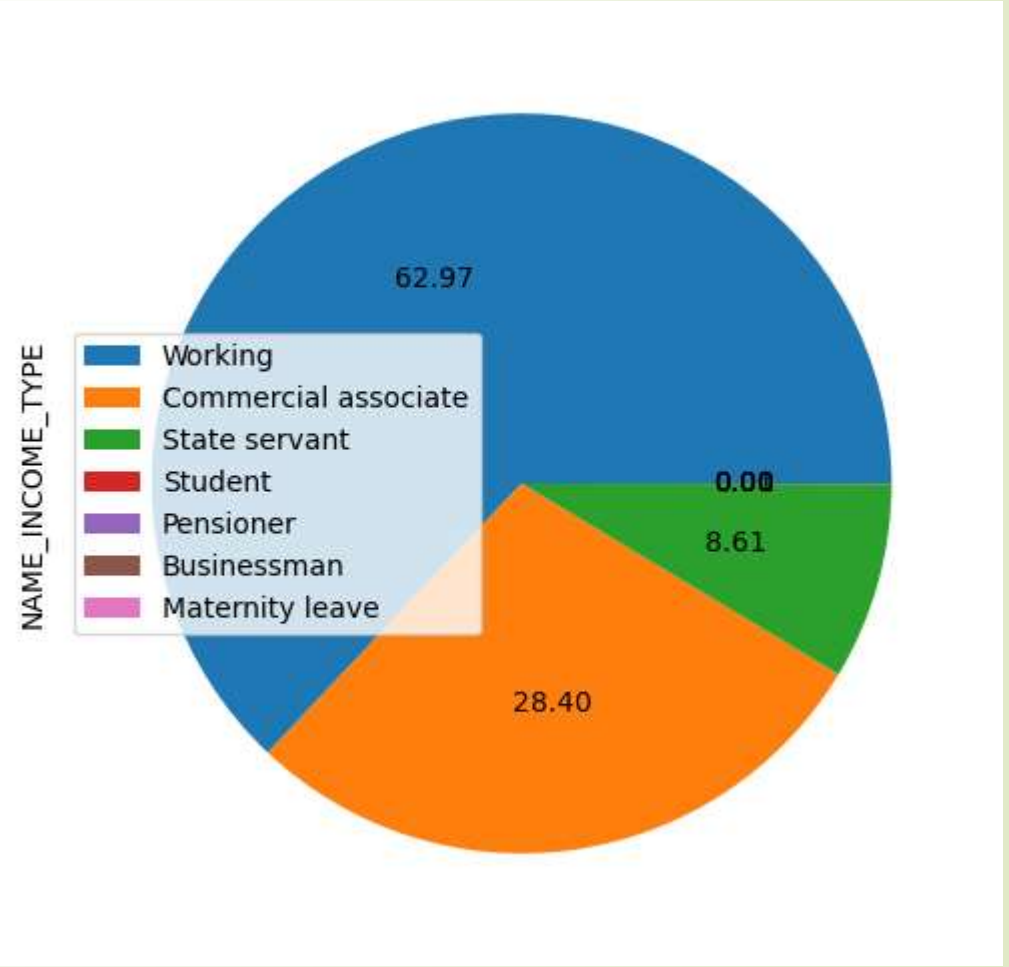


Univariate analysis Graphs and Insights

- **FLAG_INCOME_TYPE**

1. People those who are working and commerc

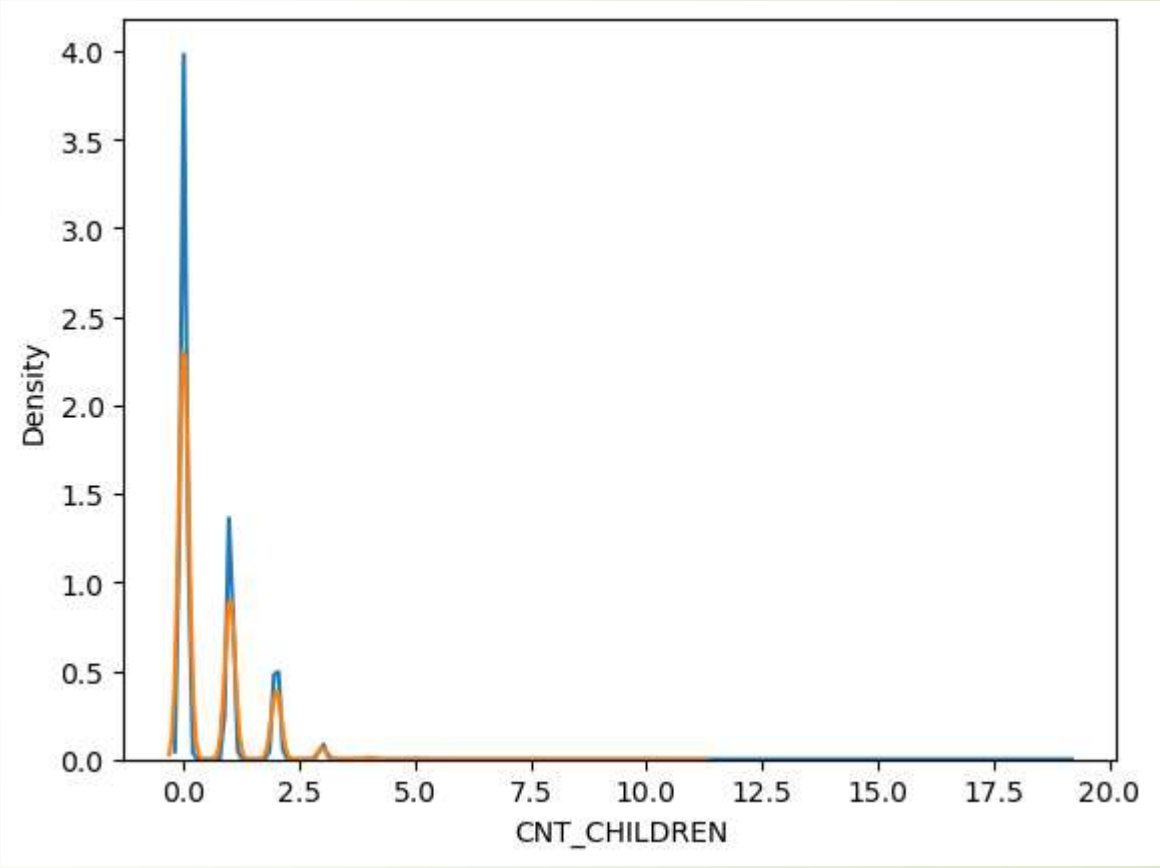
,have more chances to get loan.



Univariate analysis Graphs and Insights

- **CNT_CHILDREN**

1. People having no children have higher ,chances of getting loan.

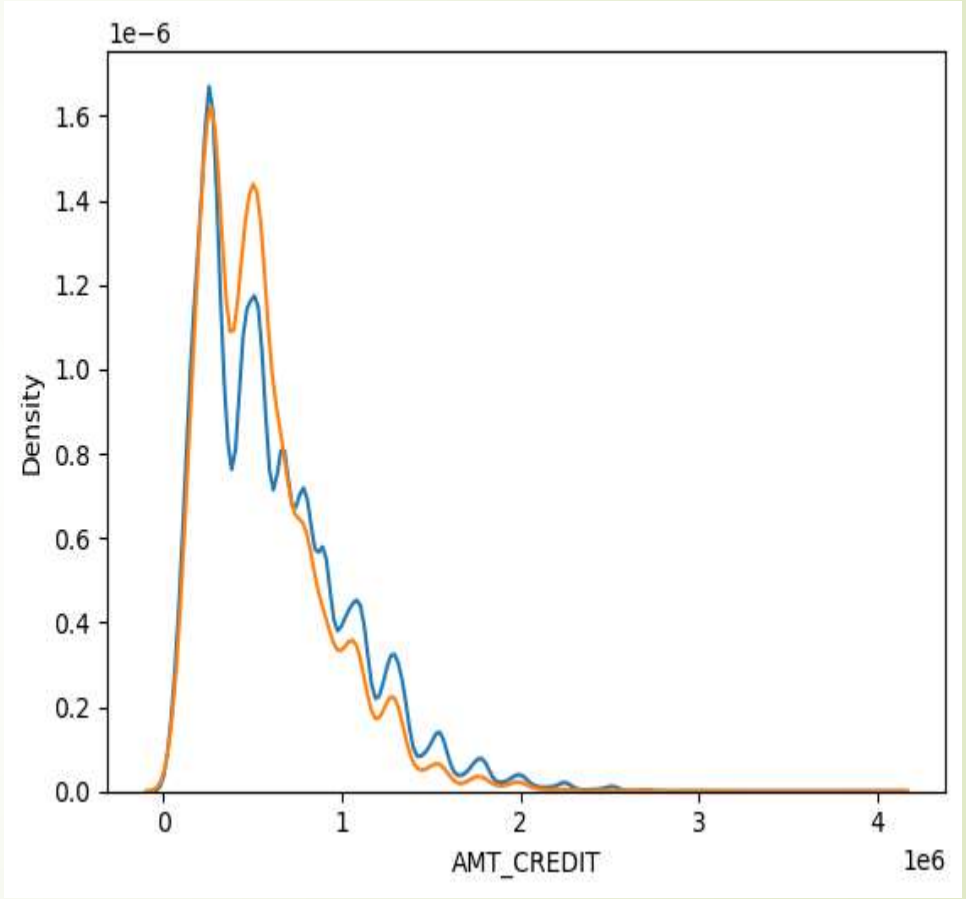


Univariate analysis Graphs and Insights

- **AMT_CREDIT**

1. Amt_CREDIT lower for Target 1,

which is a good as lesser default loss to the company.



Univariate analysis Graphs and Insights

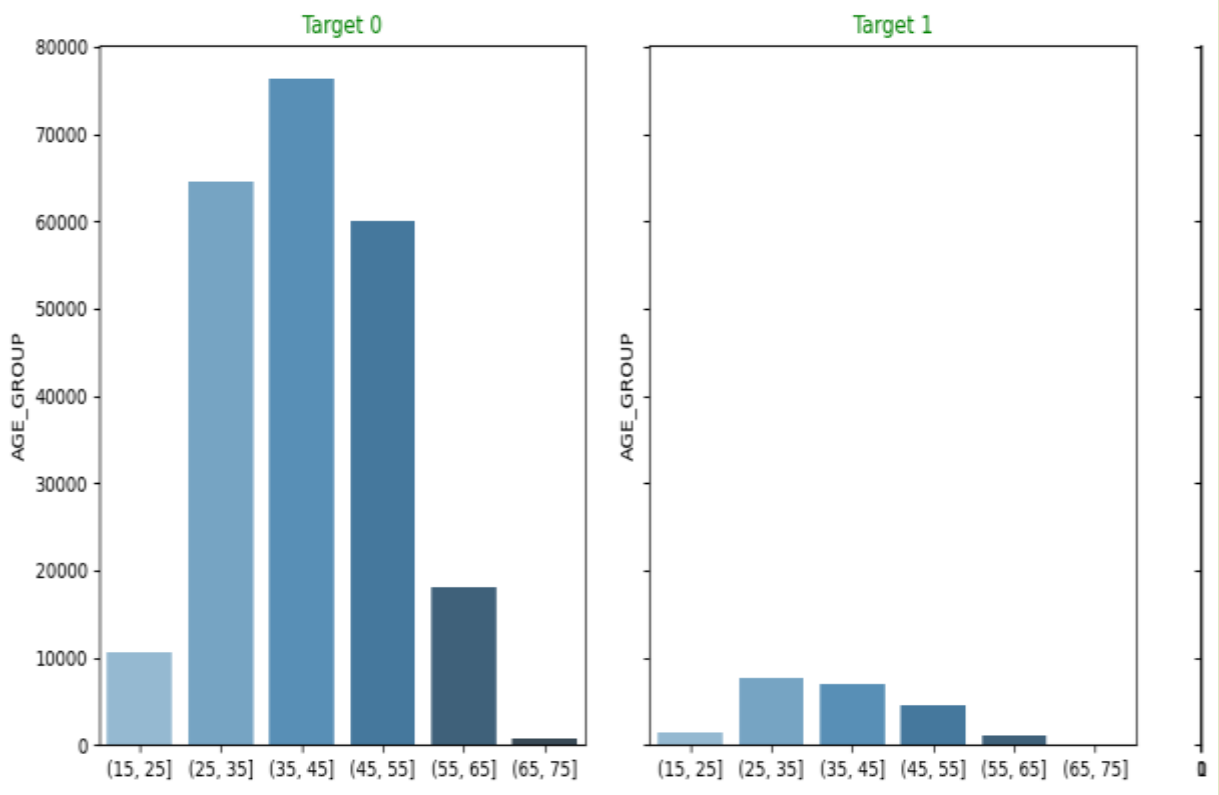
1. AGE_GROUP - 35-45 are more in TARGET 0

In Target 1- 25-35 have higher share.

Age does seem like Influencing default.

1. INCOME_GROUP - Medium income group

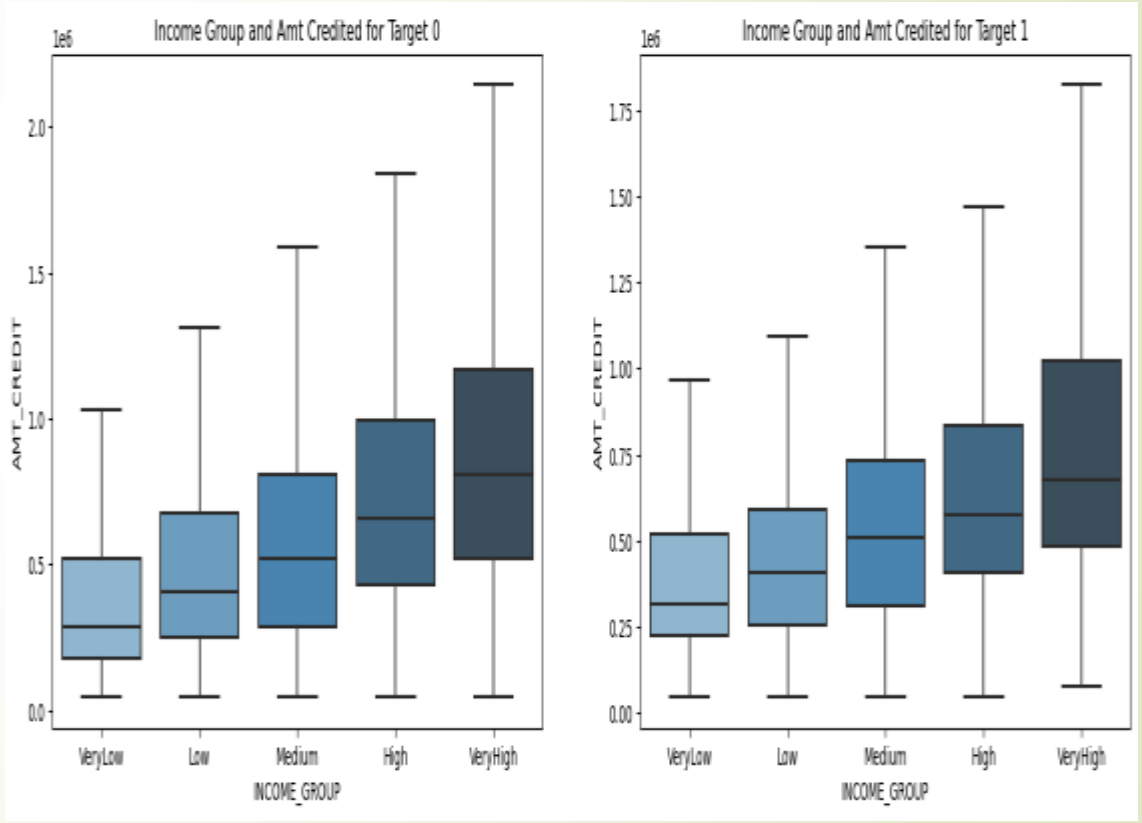
have more count in Target 0 and Target 1



Bivariate Analysis

- **Income Group and Amt Credited for Target 0 and Target 1**

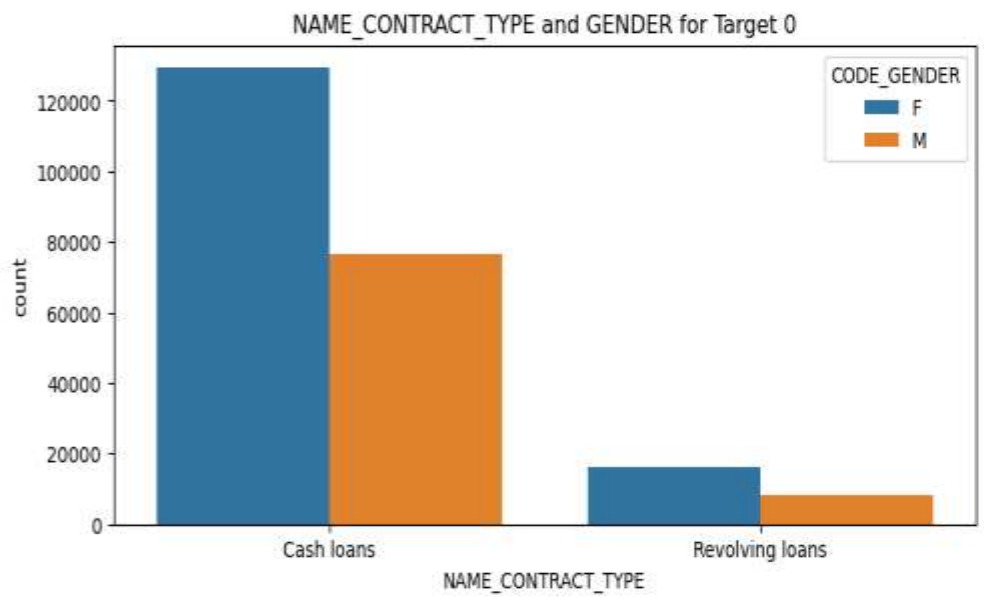
1. We can infer that though the maximum no of loans is given to Medium income group.
2. Default value per loan is highest in, High income group as the AMT_CREDIT is higher too.

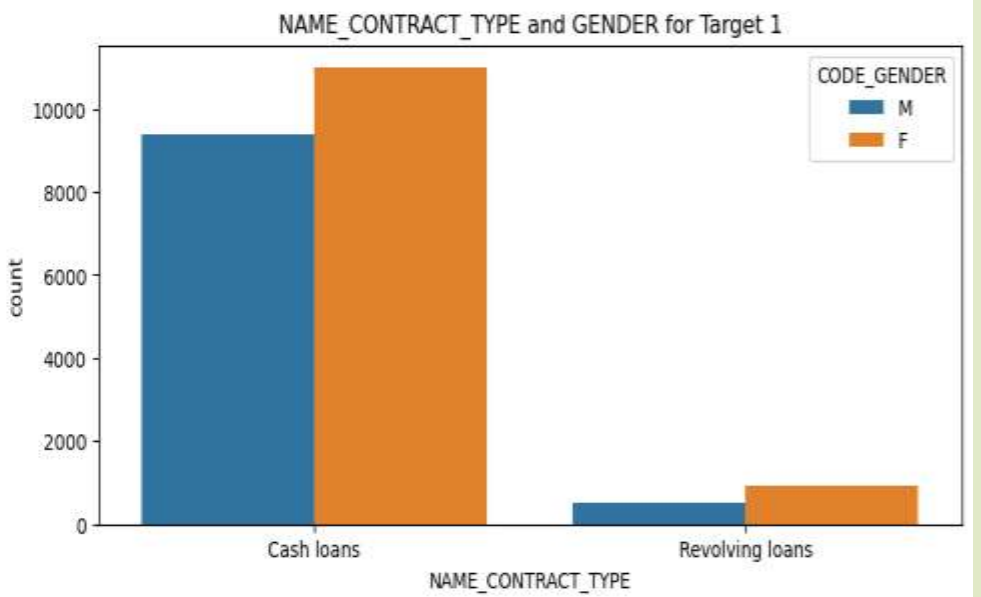


Bivariate Analysis

- NAME_CONTRACT_TYPE and GENDER for Target 0 and 1

1. As noted above data has more females as loan applicant.
2. As seen in plot above, though male applicants are lower, ratio of male applicants deafaulting is higher. Let us check this by another analysis



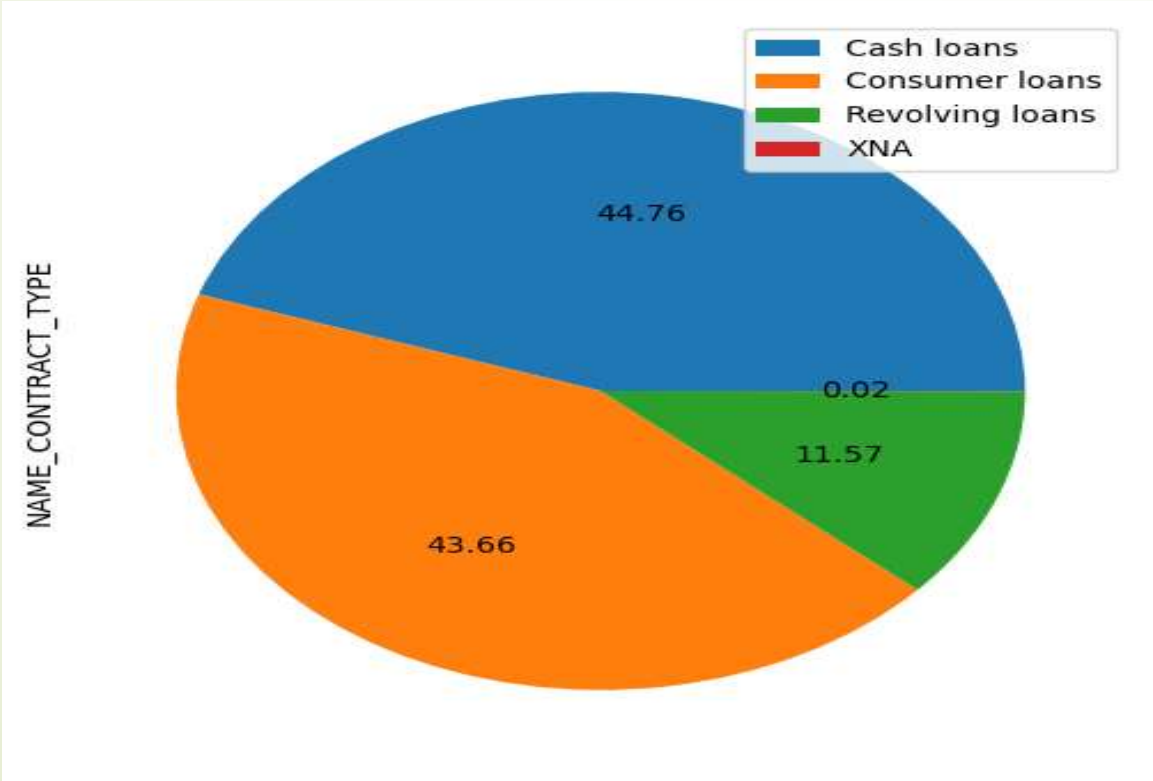


Graphs and Insights of Previous Application Dataset

Univariate Analysis

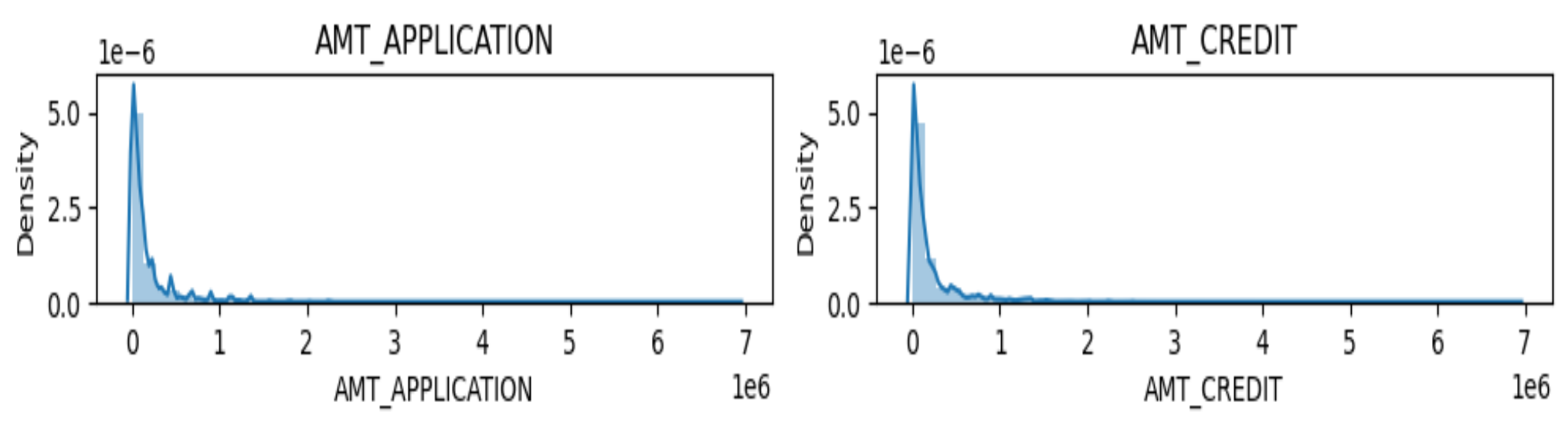
- Conclusion from graph.

1. This dataframe has a different type of loan called Consumer Loan, which was not there in Application data frame. 55% of loans are consumer loans. 37% cash loans and rest revolving.



Univariate Analysis

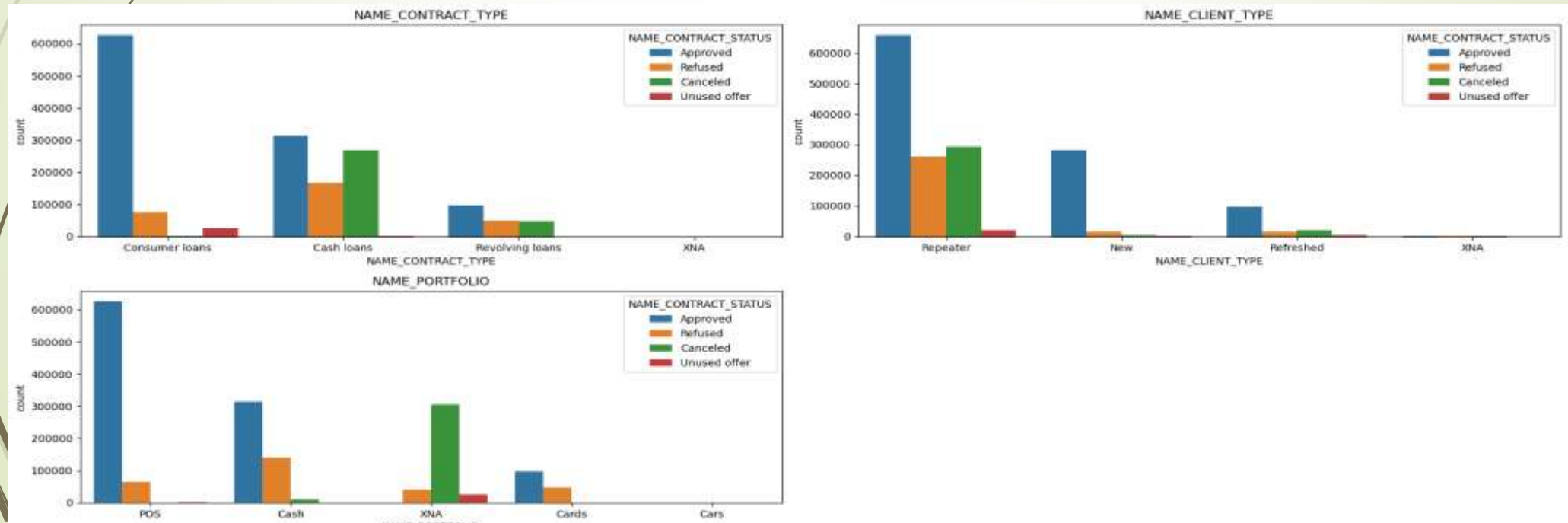
- Continuous Variables seem to have high percentage of outliers. Box plot and distribution both signify the same.



Bivariate Analysis

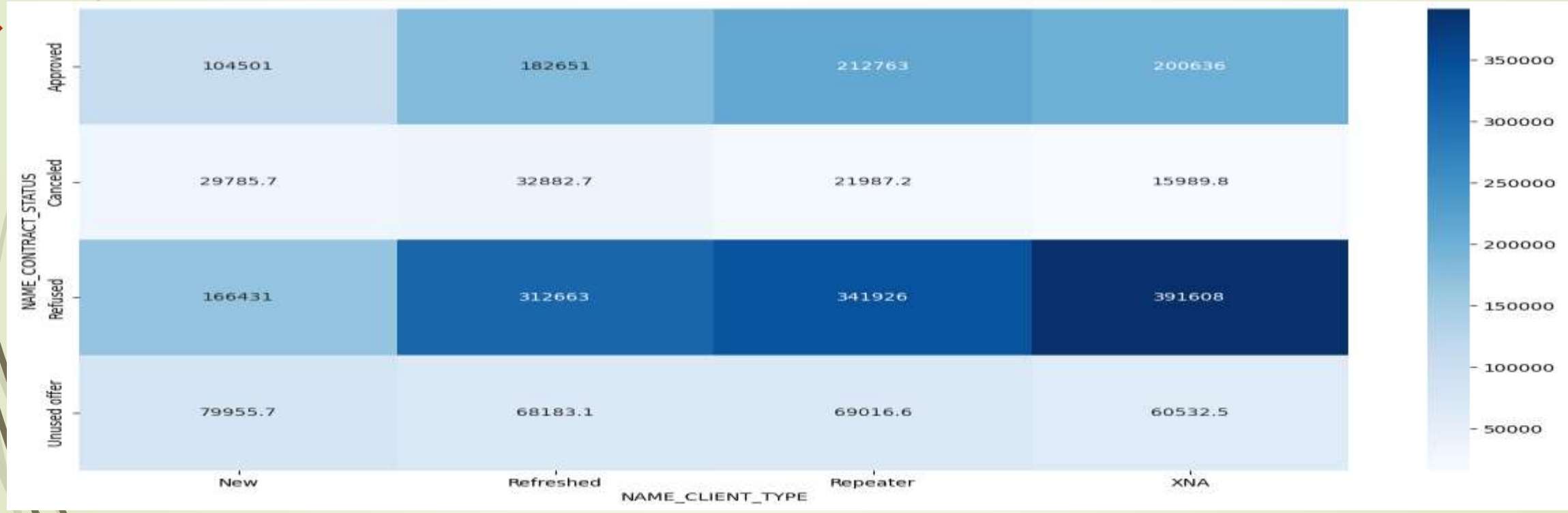
- Important points

1. In approved category, consumer loan has largest no of applicants.
2. There seem to be no cancelled loans in cash loan category than consumer loan.
3. More cash loans have been refused than consumer loans.
4. The bank has more repeaters in all approved, refused, unused, cancelled categories
5. POS transactions seem to be consumer loans and similar to point 2 - more cash loans have been refused than POS.



Multivariate Analysis

1. Unused offer application amount is low
2. Cancelled application amount is high. The bank may be refusing these possibly as the Debt liability ratio of consumer must be going high due to the high amount and thus credit default risk.
3. Repeater's application amount is higher than the New customers. This may indicate that the bank has more conducive policies/rate of interest etc. for repeat applicants.



Conclusion

- Defaulters

1. *All the variables were established in analysis of Application dataframe as leading to default. Checked these against the Approved loans which have defaults, and it proves to be correct.*

- *Medium Income*
- *25-35 years old , followed by 35-45 years age group*
- *Male*
- *Unemployed*
- *Laboure's, Salesman, Drivers*
- *Business type 3*
- *Own House-No*

Conclusion

- Important Factor
- *Last days phone number changed - Lower figure points at concern*
- *No of Bureau Hits in last week. Month etc. – zero hits is good.*
- *The Amount income is not corresponding equivalent to goods that bought – Income low and good value high is a concern.*
- *Applications that has been changed with refused, Cancelled, Unused loans also have default which is a matter of concern. It indicates that the financial company had Refused/Cancelled previous application but has approved the current and is facing default on these.*
- Credible Applications refused
- The Unused applications has lower loan amount. The Female applicants should be given extra weightage as defaults are lesser.60% of default people are Working applicants. This did not mean working applicants must be refused. Proper inspection of other parameters needed.