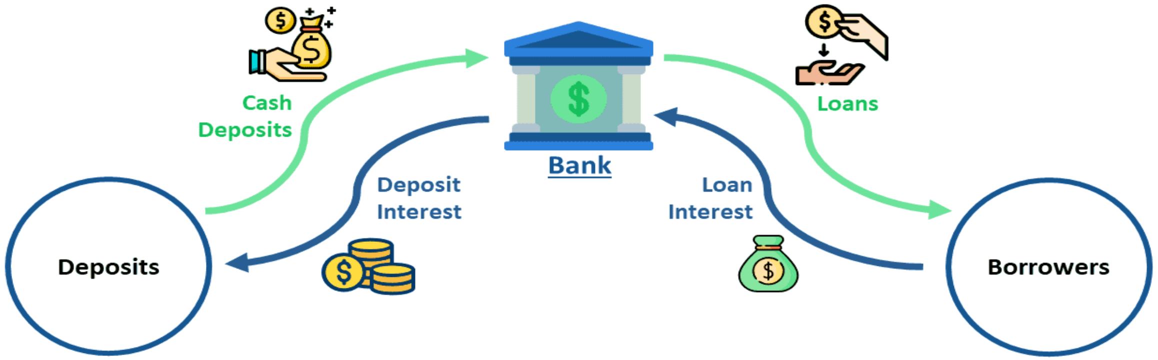
Credit EDA Assignment

Problem Statement

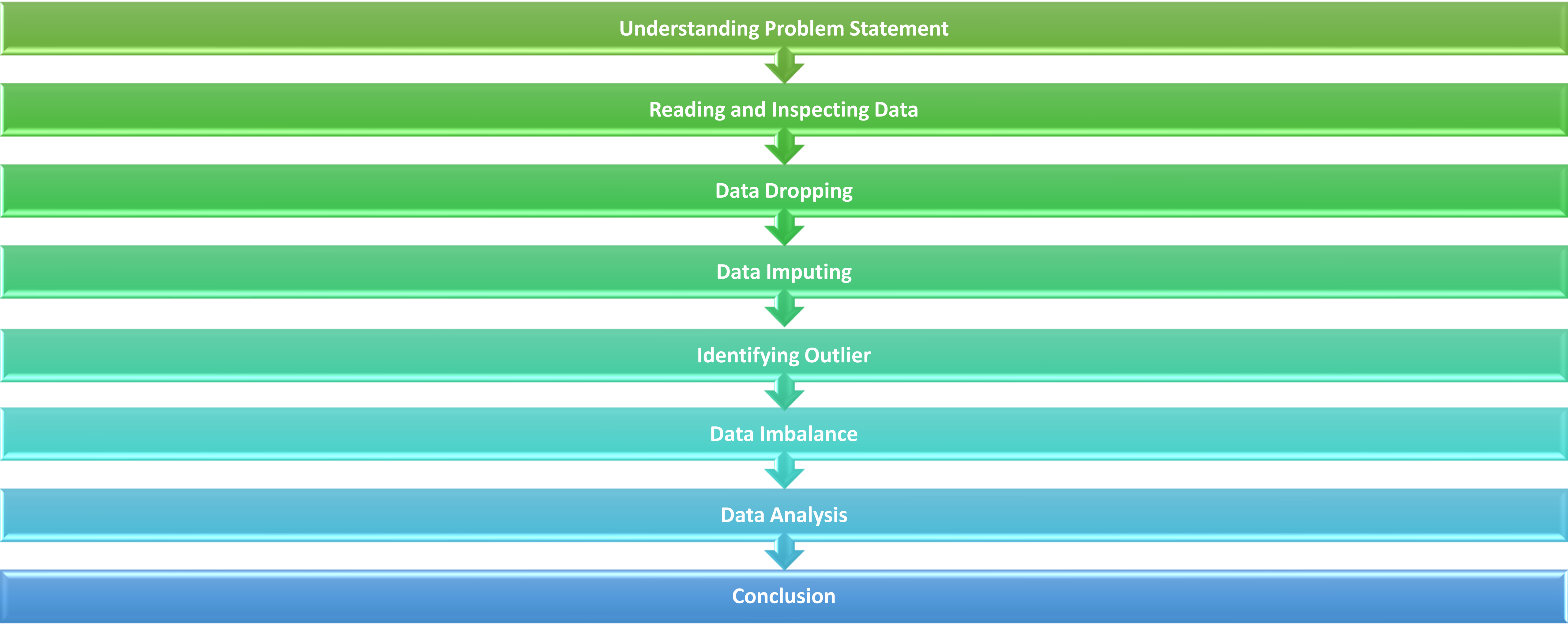


- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter

- 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:
  1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
  2. 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.

- 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.

Overall Approach



Understanding Problem Statement

Reading and Inspecting Data

Data Dropping

Data Imputing

Identifying Outlier

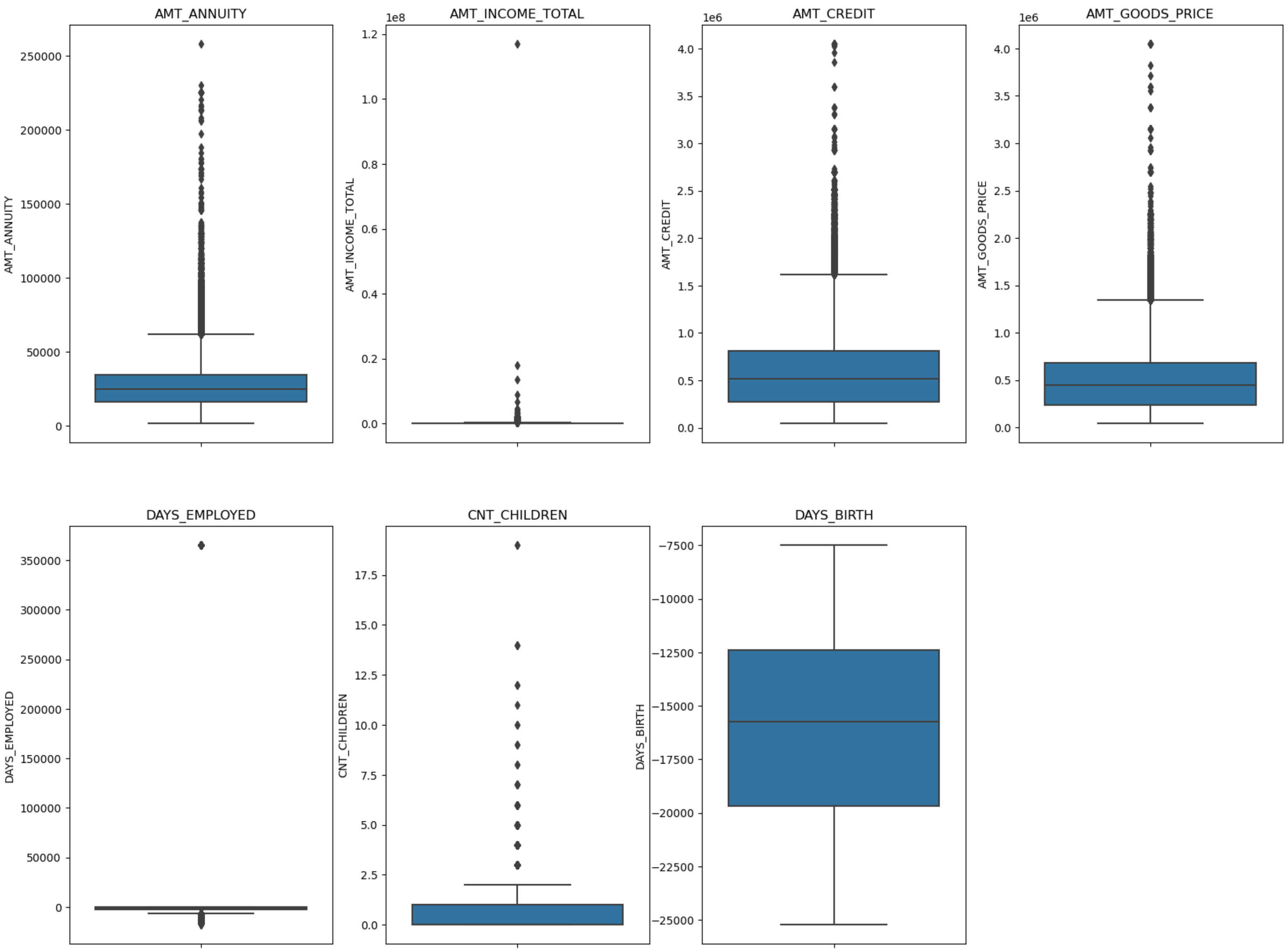
Data Imbalance

Data Analysis

Conclusion

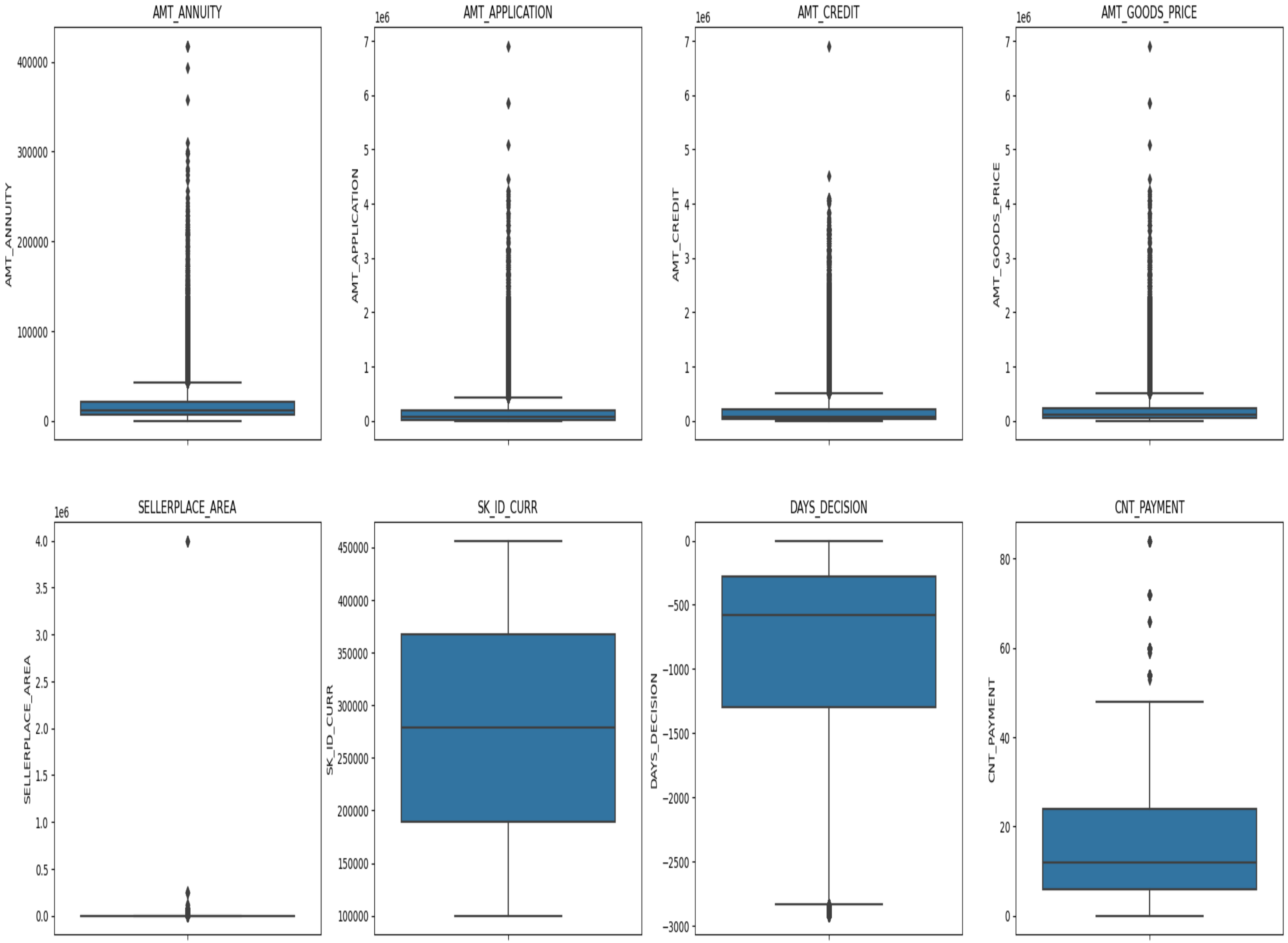
Outlier Analysis in application_data

- AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have few outliers in them Whereas AMT_INCOME_TOTAL has huge number of outliers.
- In AMT_ANNUITY column we can see outliers near 250000 and there isn't much difference between the mean and median so we can impute the outliers with Median here
- In DAYS_EMPLOYED column we can clearly see that the value of outlier is more than 35,0000 which is practically not possible and the data is not valid.
- The DAYS_BIRTH looks completely fine without any outliers

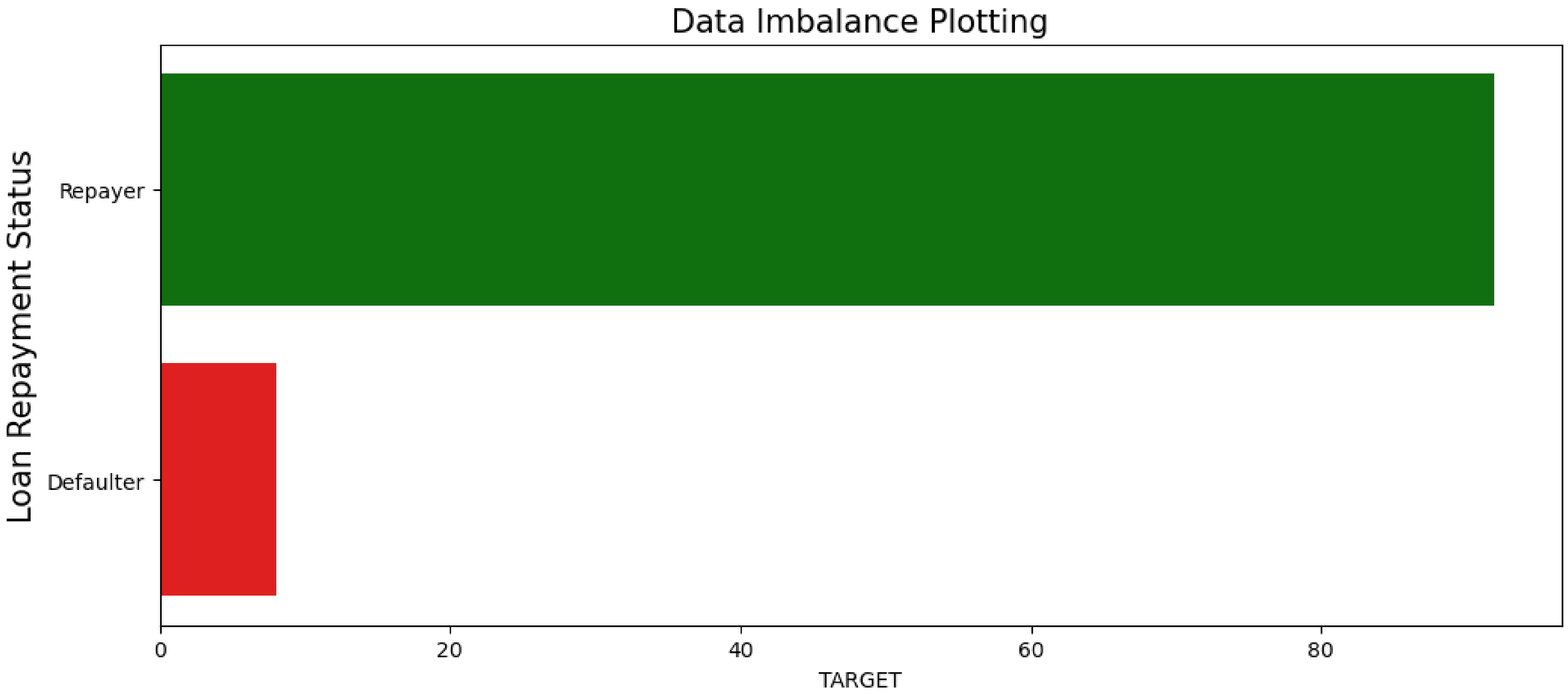


Outlier Analysis in previous_application

- From the above box plot we can see that there are huge number of outliers in AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA
- There aren't any outliers present in SK_ID_CURR column
- DAYS_DECISION and CNT_PAYMENT columns has very few outliers

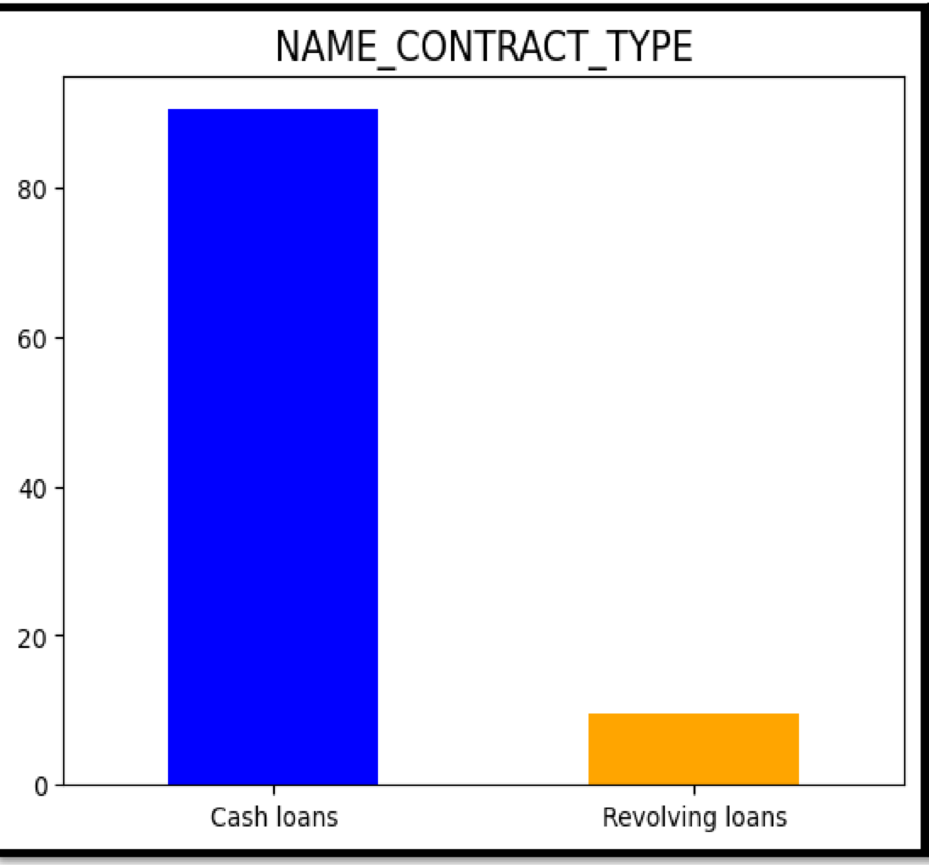


Data Imbalance Analysis

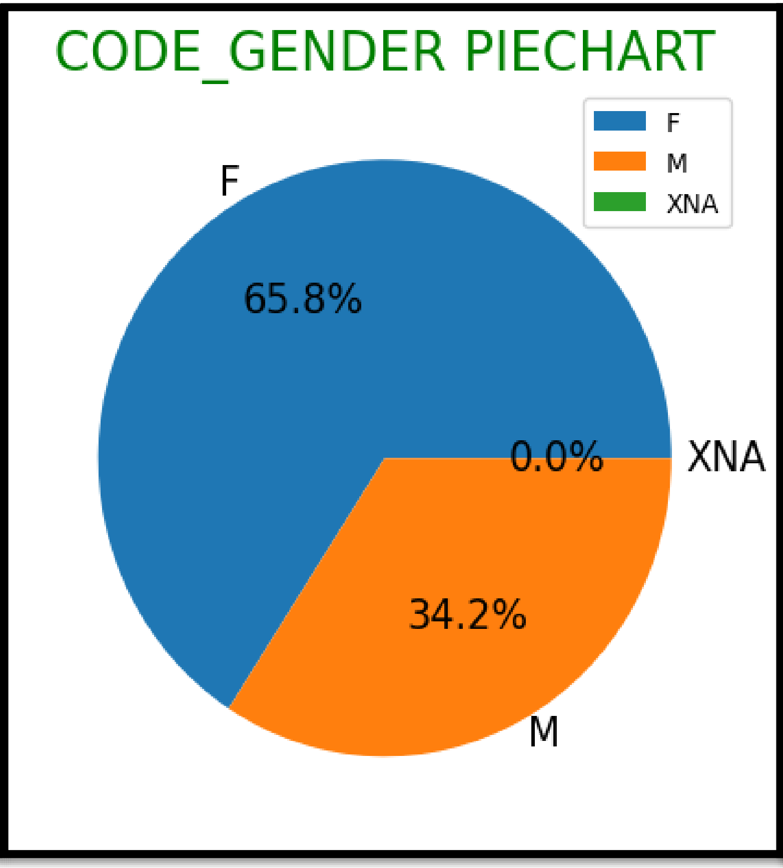


- From the above figure we can conclude that this is a real time and genuine data as 90% of customers are repaying and only 8% are defaulting

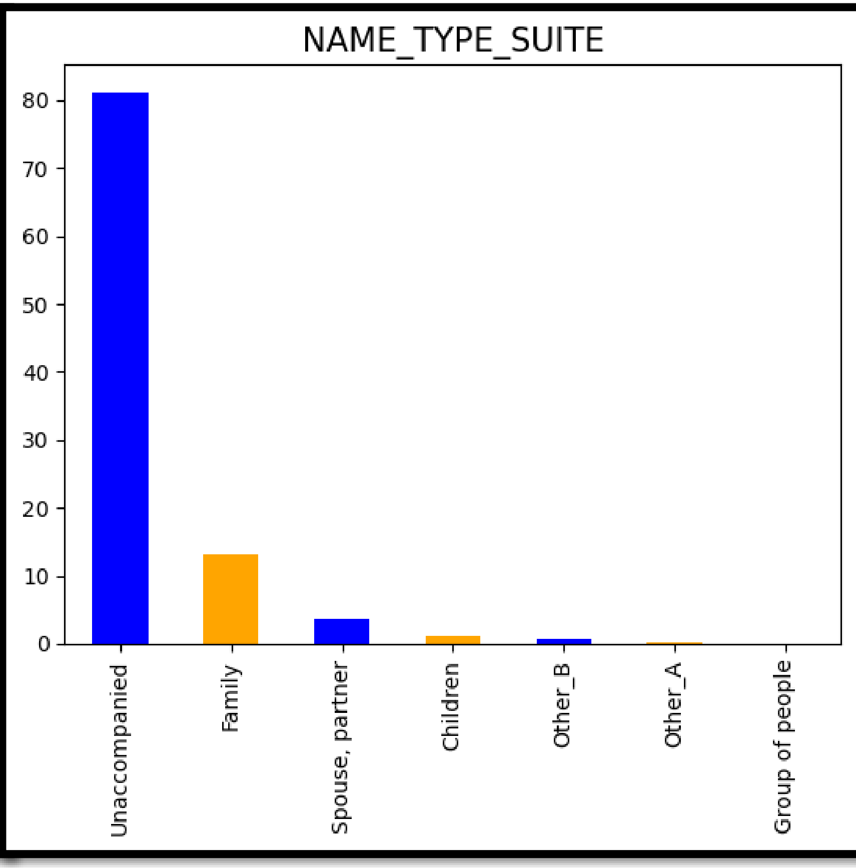
Categorical Univariate Analysis



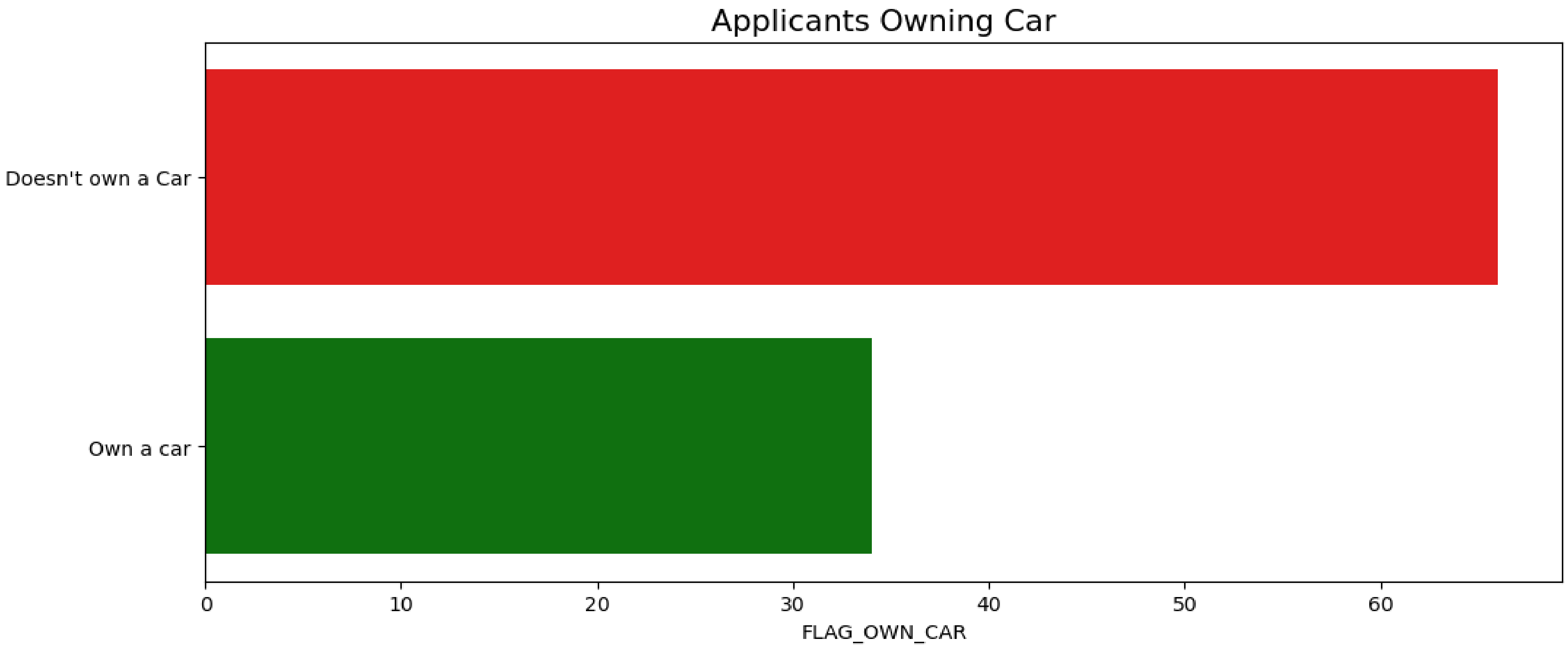
Only 10% of total loans are revolving loans rest are cash loans



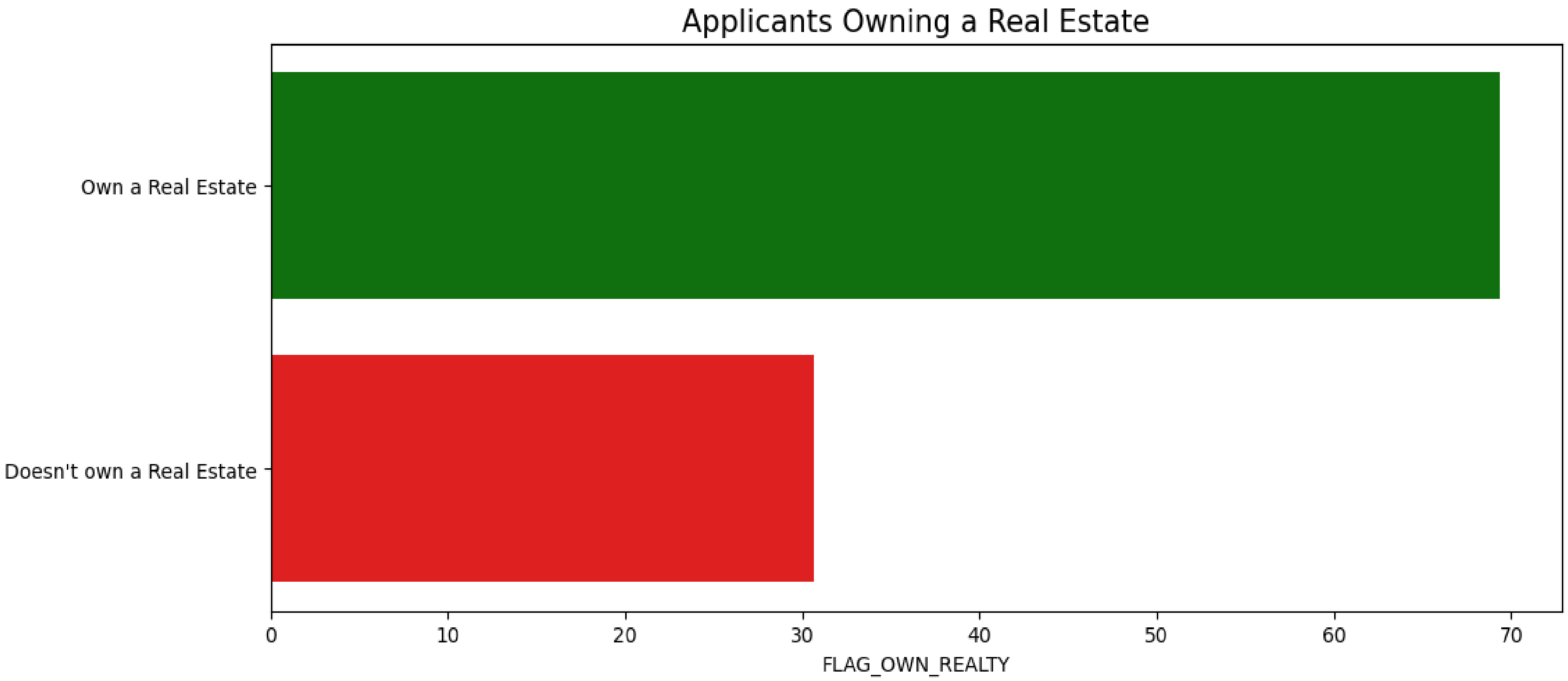
From the above pie chart we can clearly see that the number of female applicants are twice the size of male employees



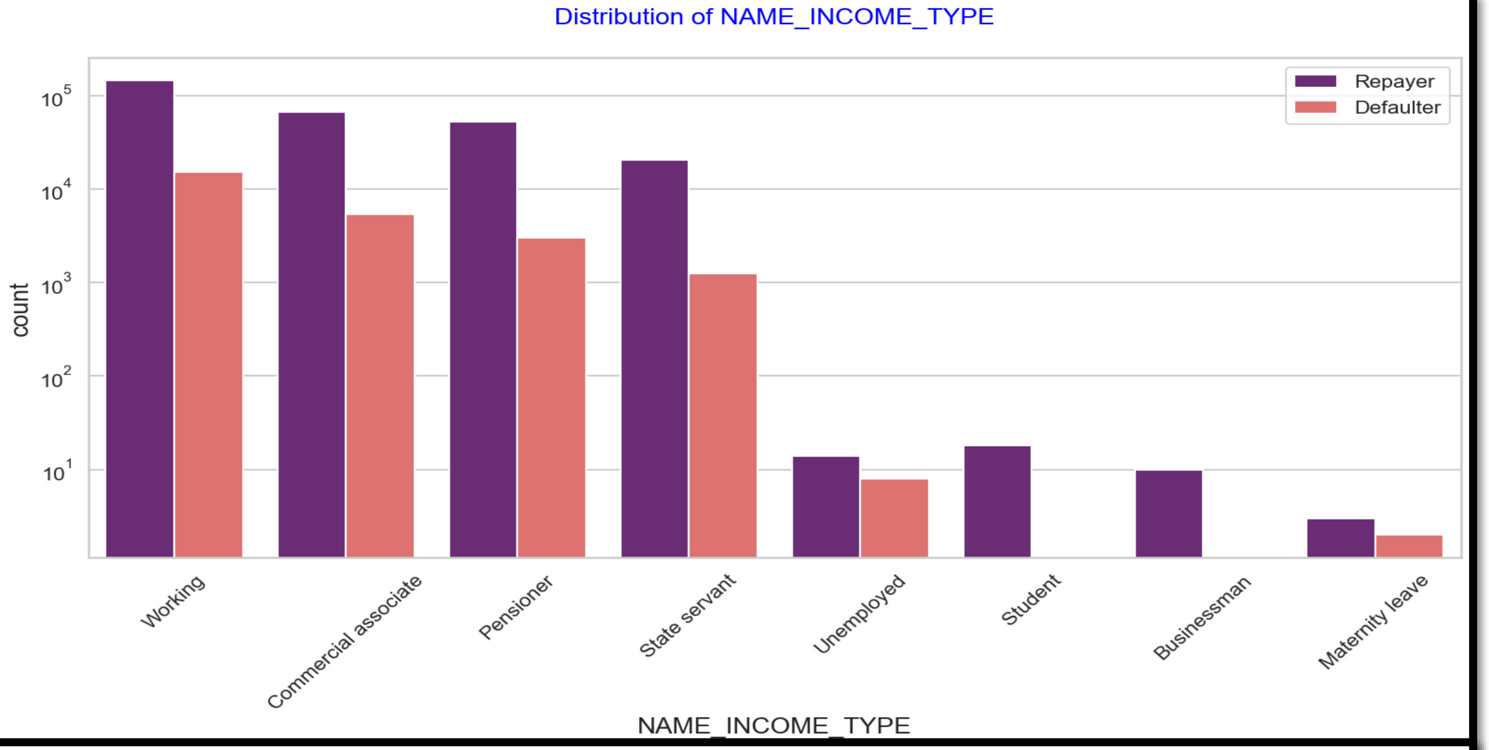
Majority of applicants are Unaccompanied followed by those with family



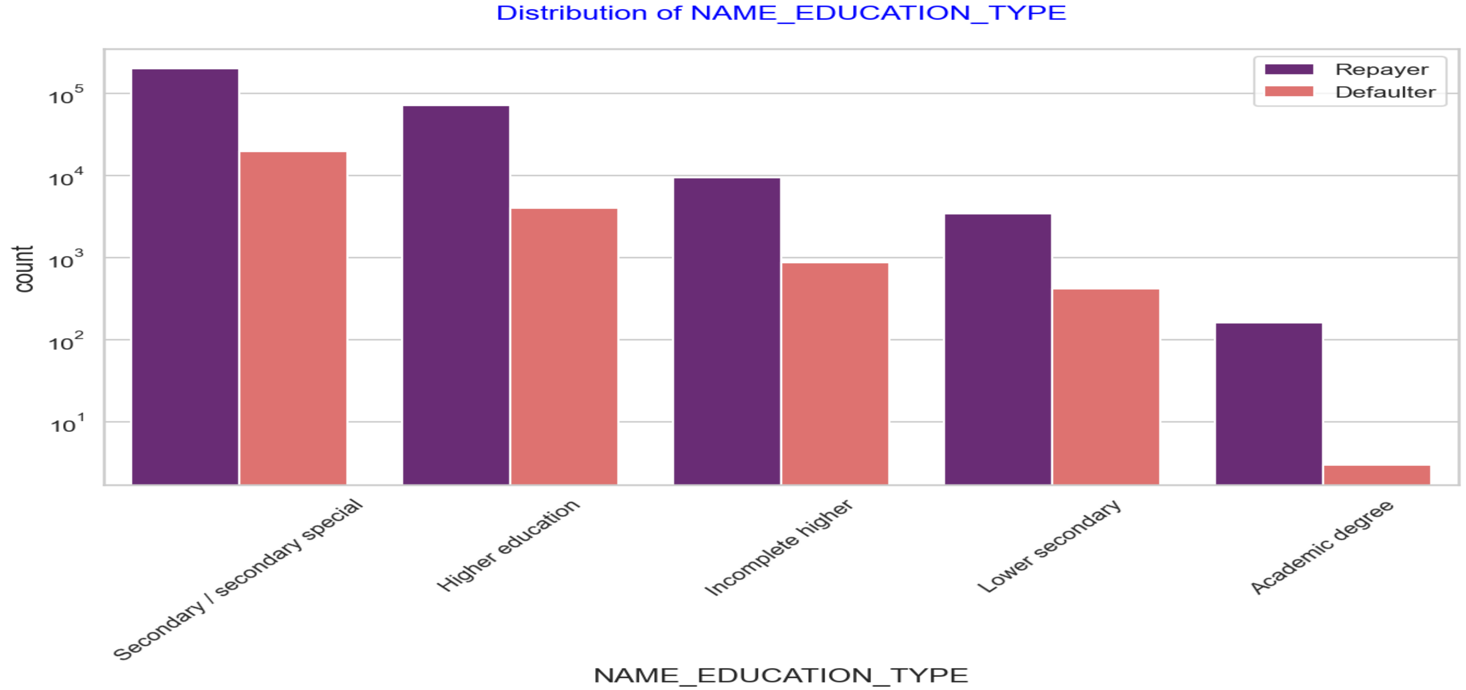
Only half of the total applicants own a car rest doesn't own a car



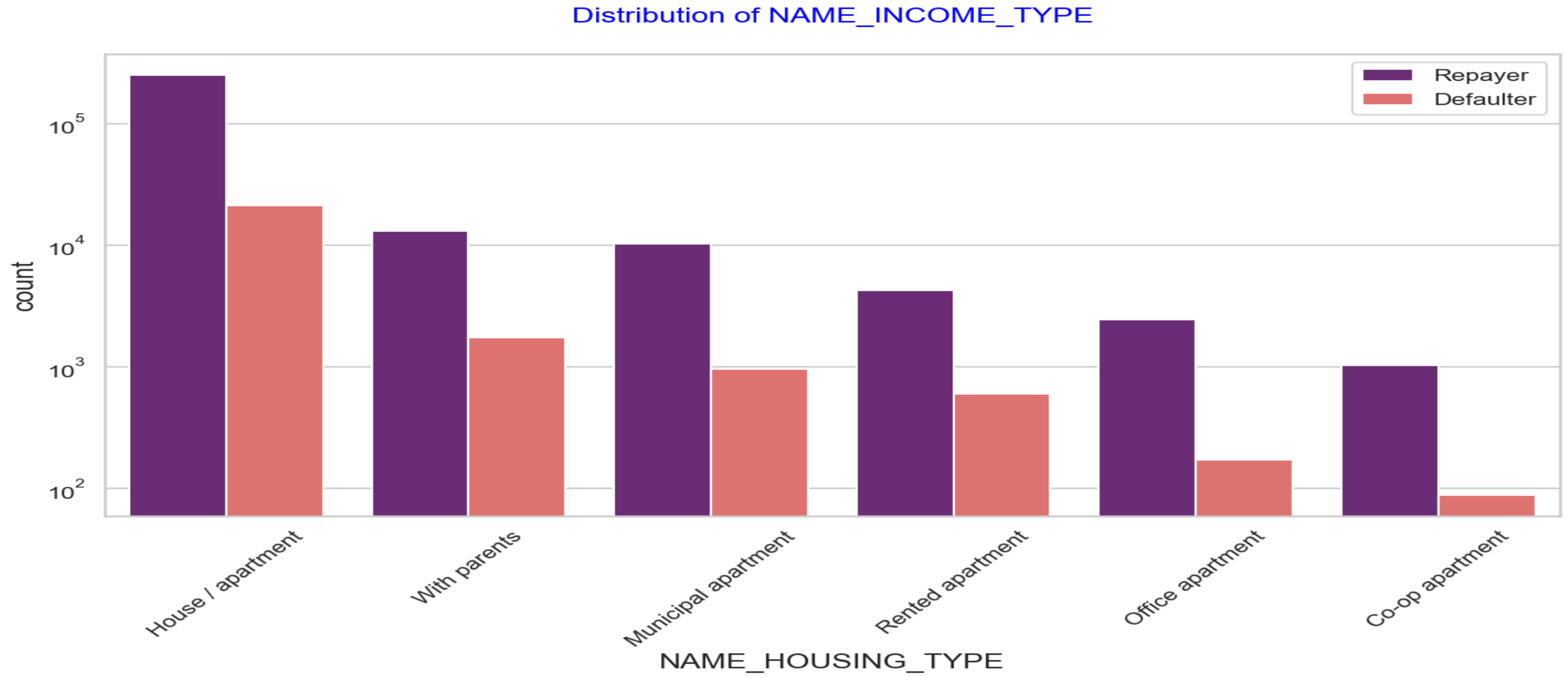
The applicants that own real estate is twice the size of applicant that doesn't own a real estate



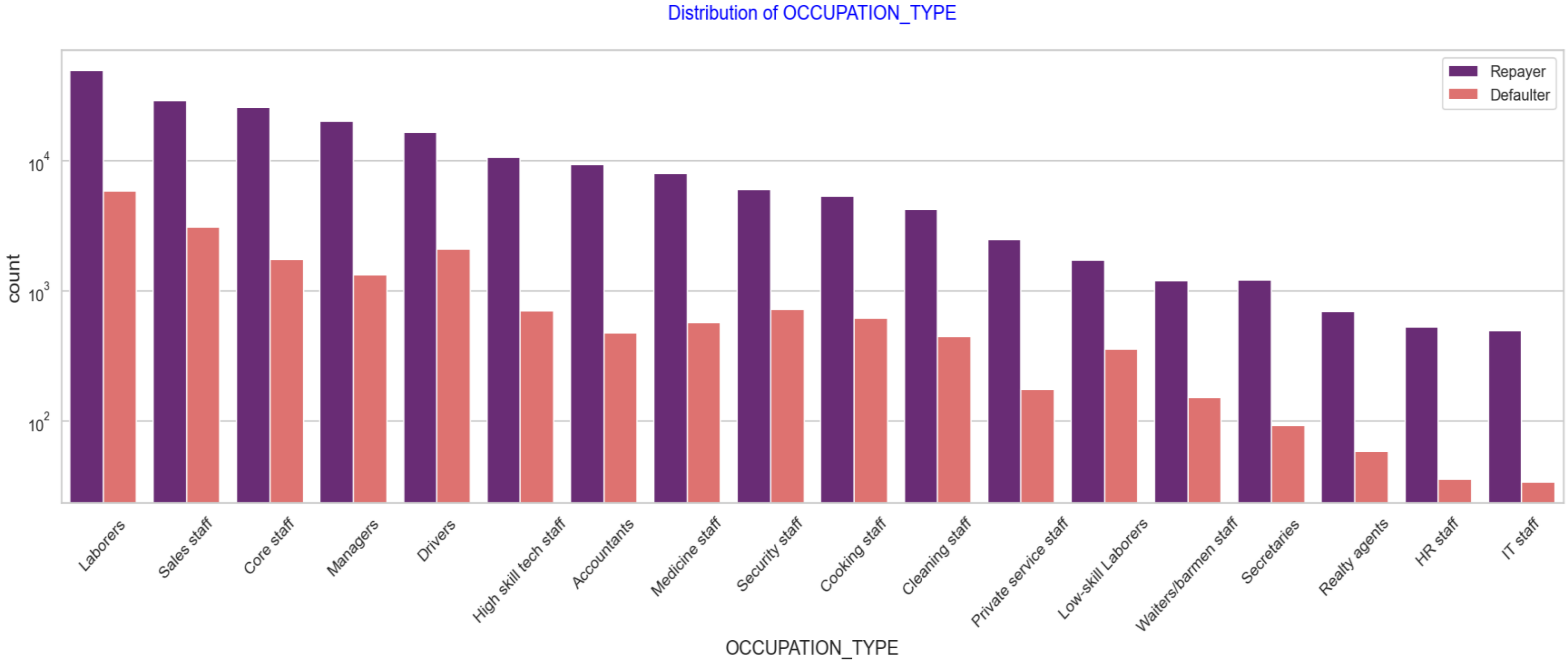
- Most number of applicants are from Working category followed by Commercial associate, Pensioner and State servants.
- Students and Businessman are the safest category to provide loans as they doesn't have any defaulters.
- The Unemployed and Maternity leave category are the riskiest as they have the highest defaulters.



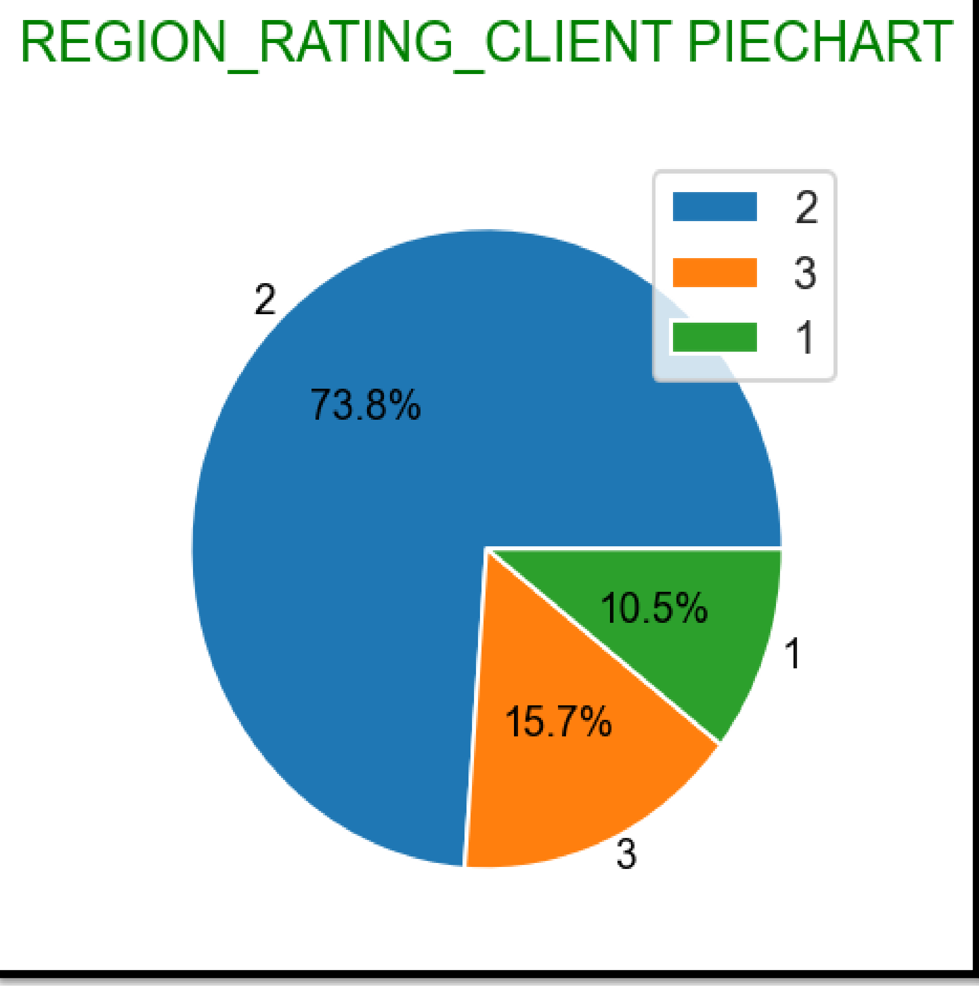
- Most applicants have completed their Secondary/secondary special followed by the applicants with Higher education.
- Very few number of people have lower secondary and Academic degree.
- According to the charts people with Academic degree have lowest defaulter.

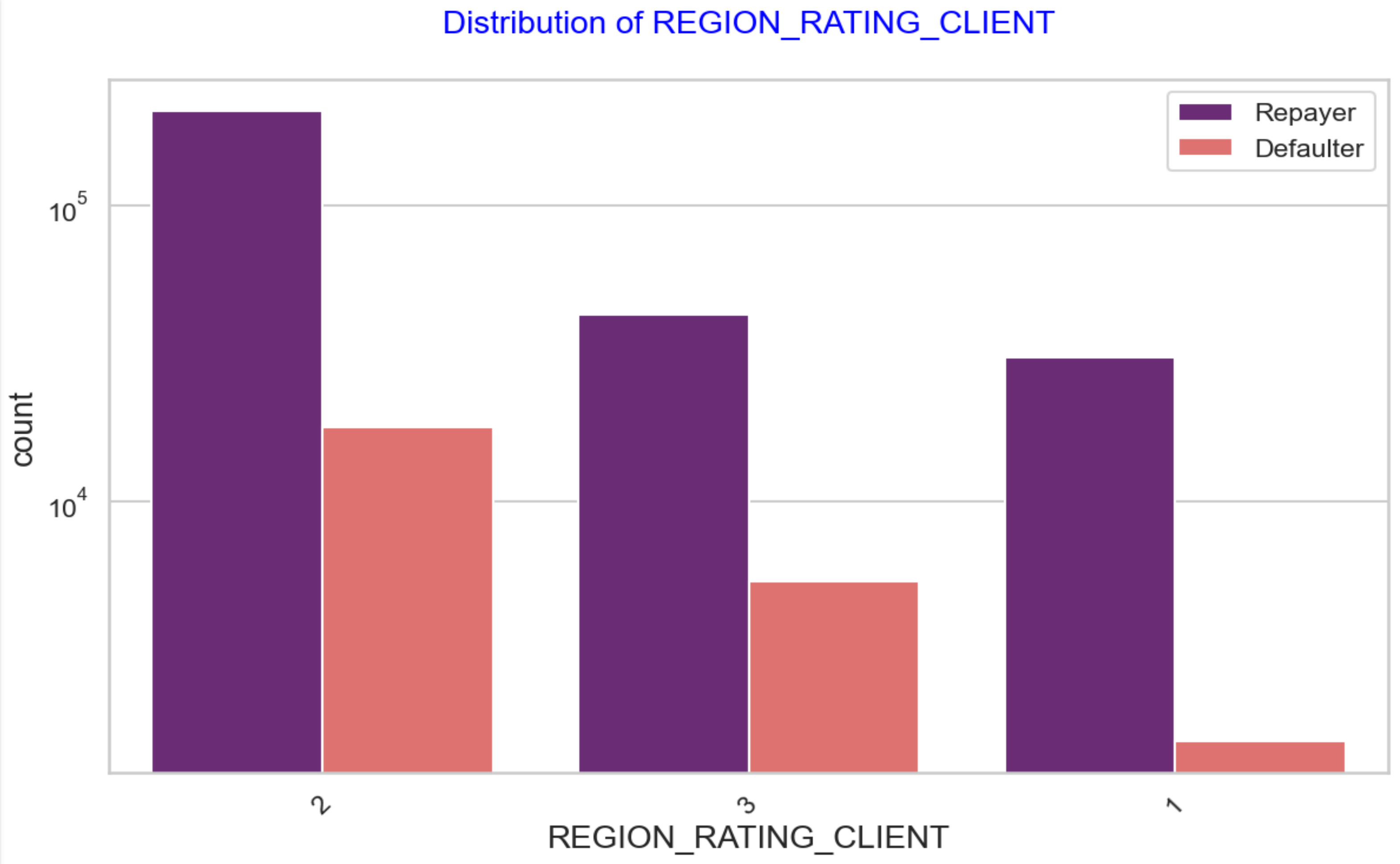


- From the above figure we found out that majority of applicant lice in House/apartment
- Only few of the applicants live in Co-op apartments and Office apartments these applicants have the lowest default rate

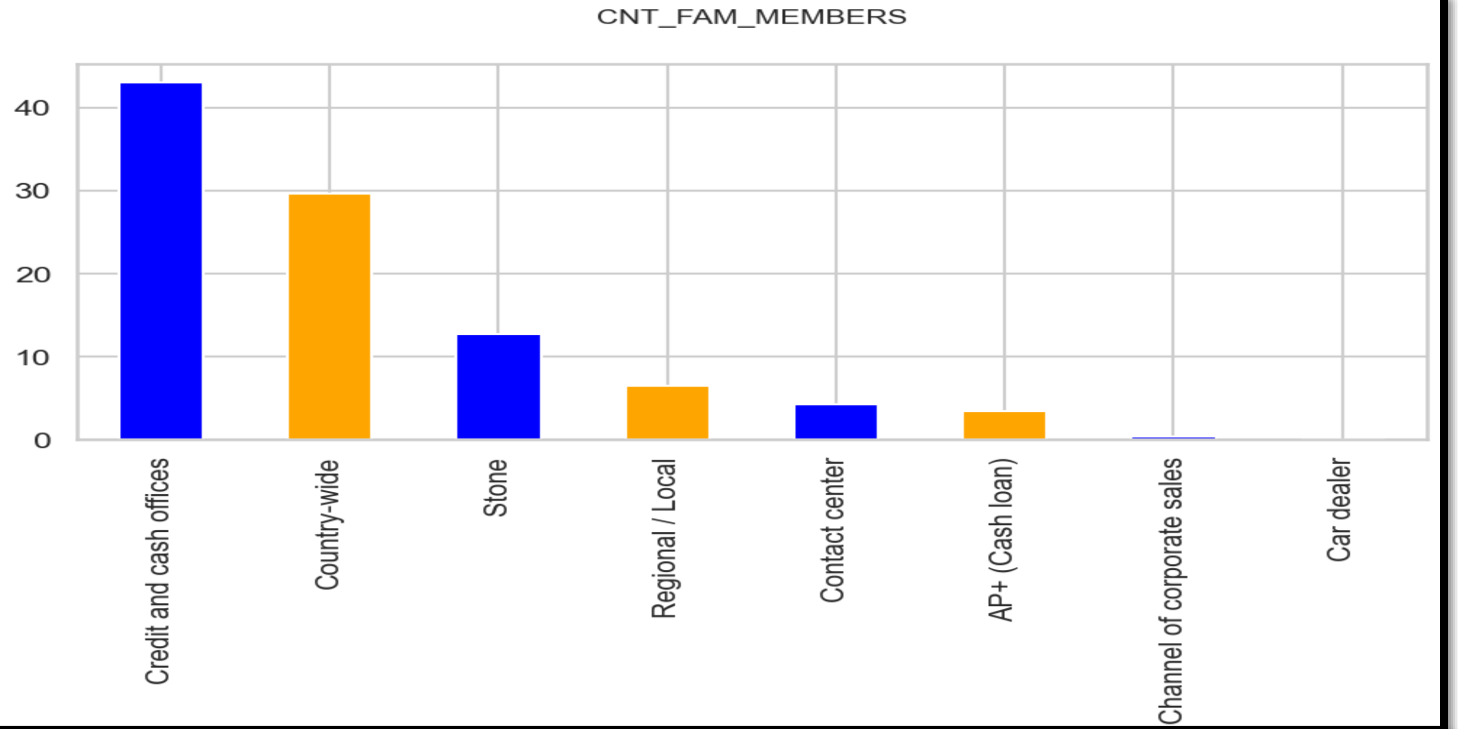


- Most of loan belongs to applicant who work as a Labourer, Sales staff and Core staff
- Applicants who work in IT and HR staff don't apply for loans much often

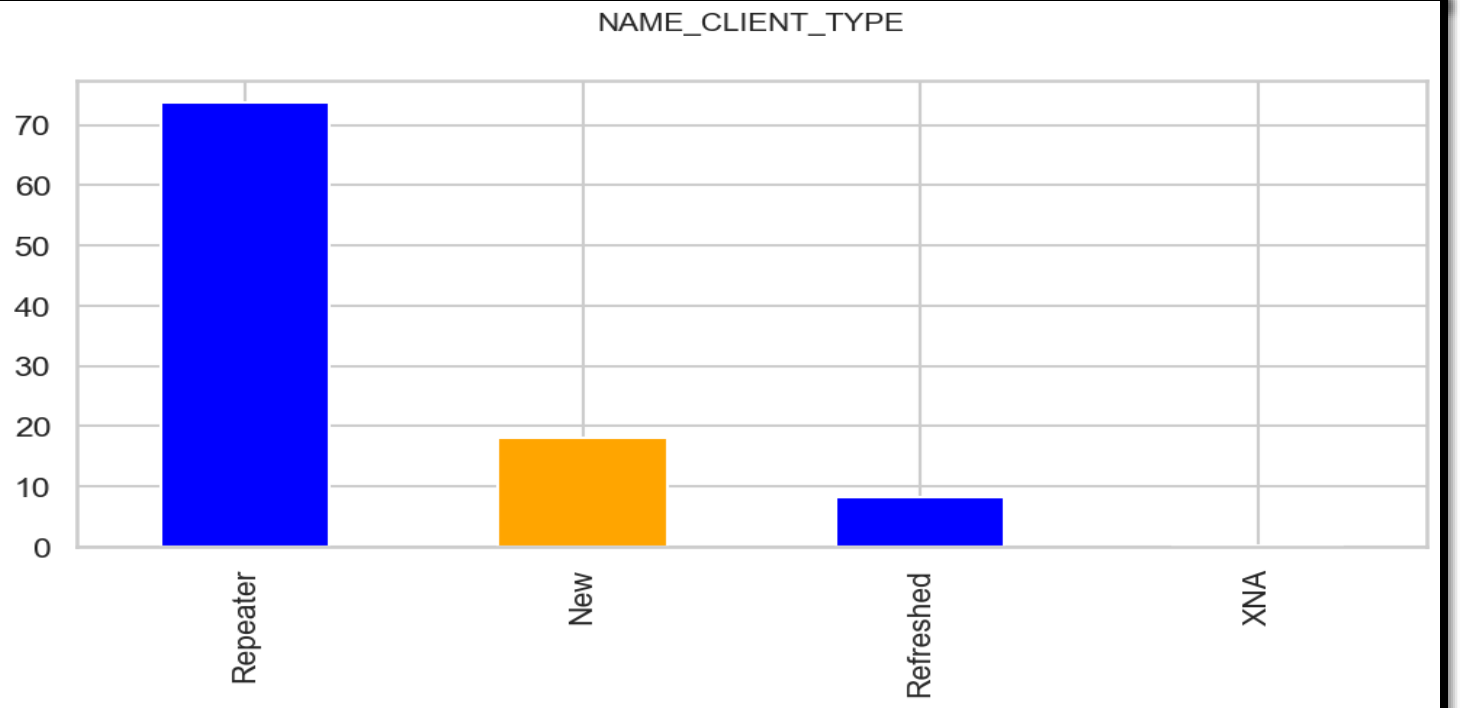




- Majority of applicant are from Region Rating 2
- It is safer to approve loan for the applicants who live in Region Rating 1 as they have lowest defaulter rate



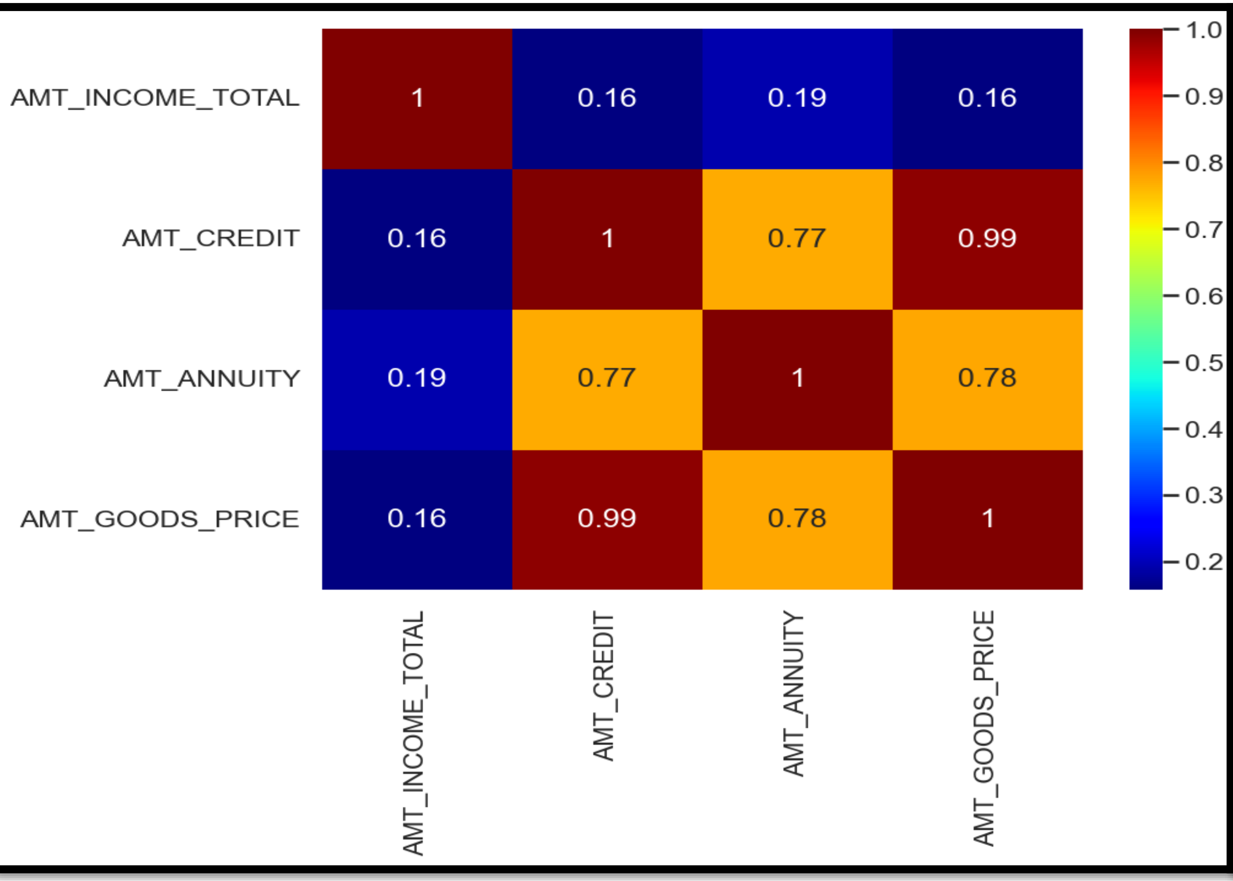
Majority of the loans are transacted through Credit and cash offices

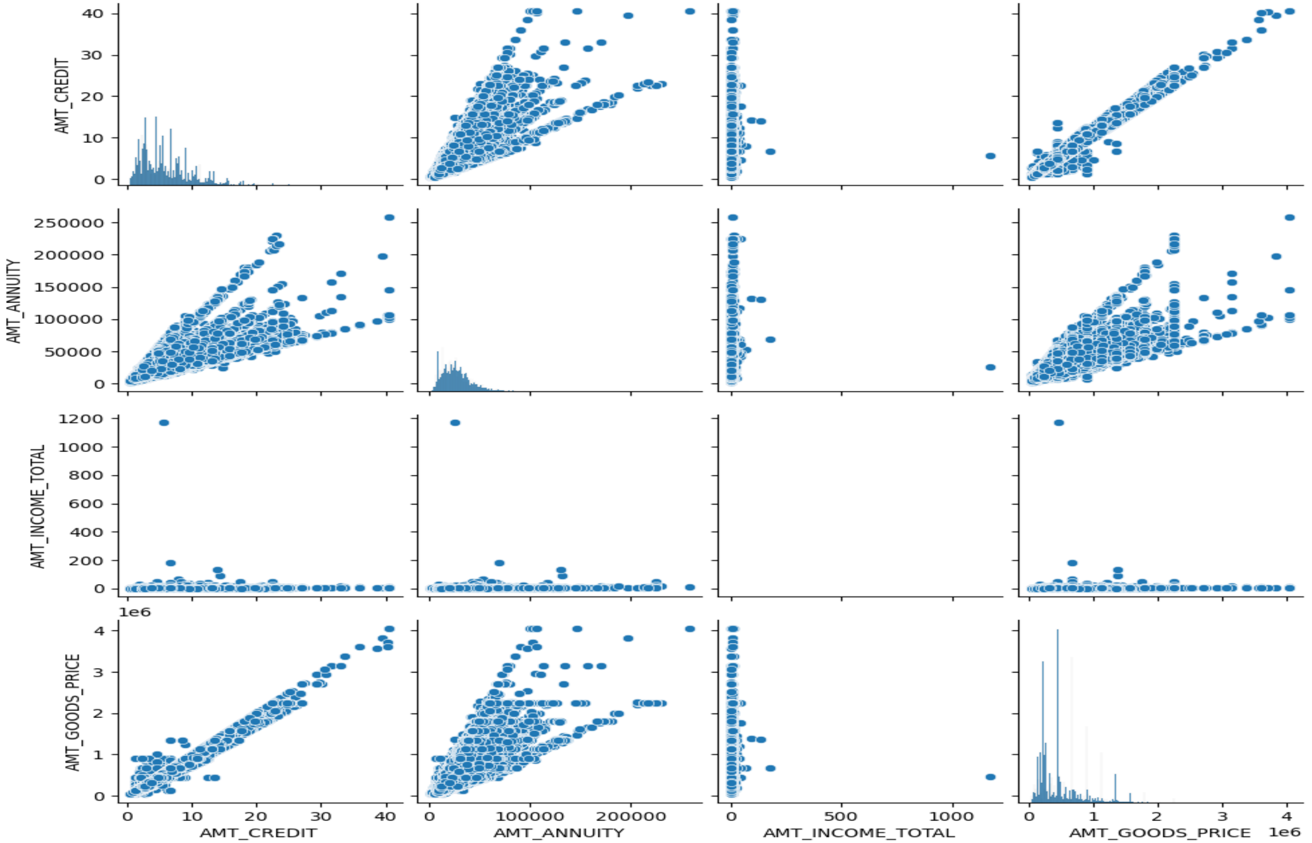


About 70% of applicant are repeaters which is a good thing and around 20% are new ones

Bivariate Analysis

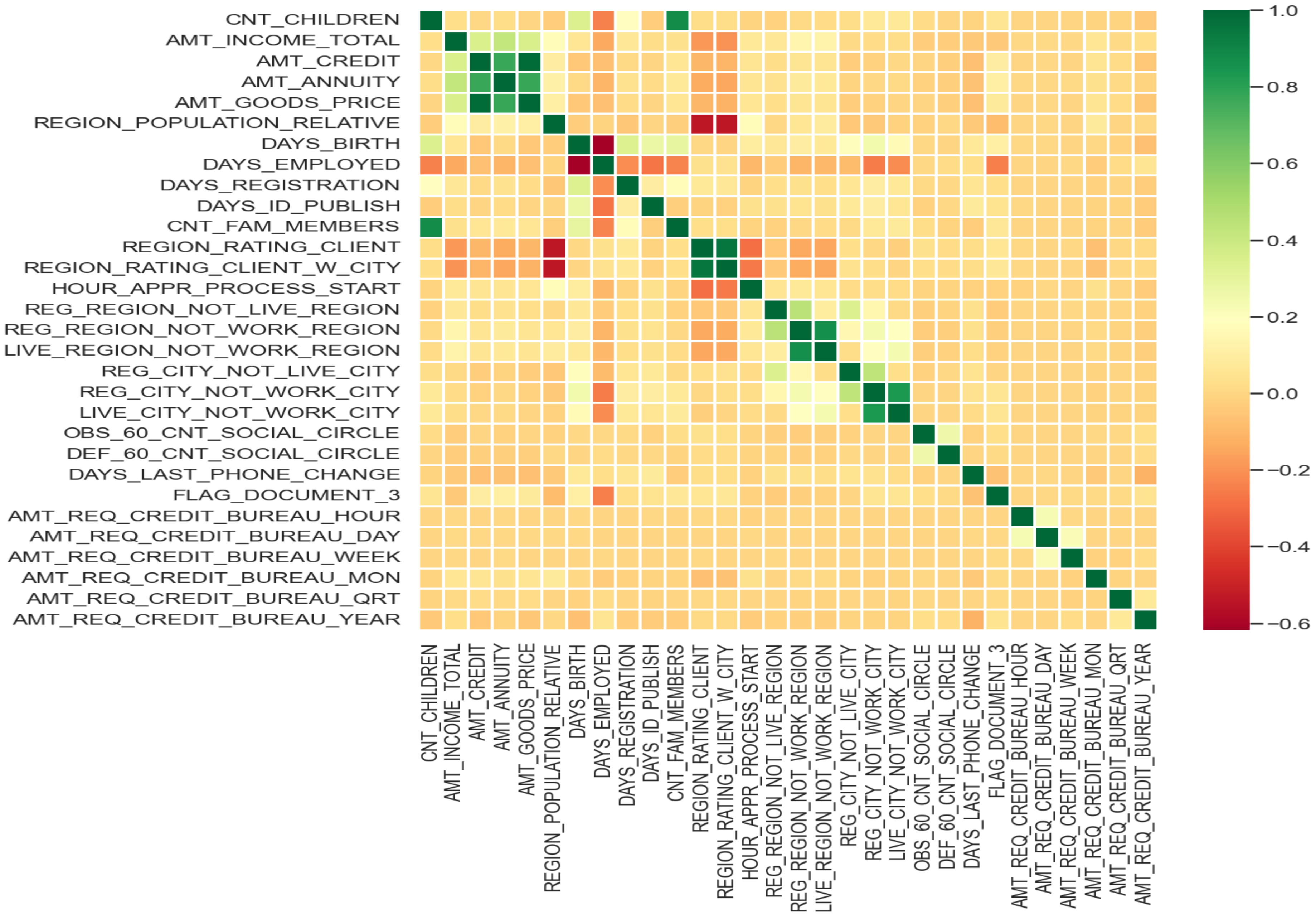
Comparing all Amount columns with each other





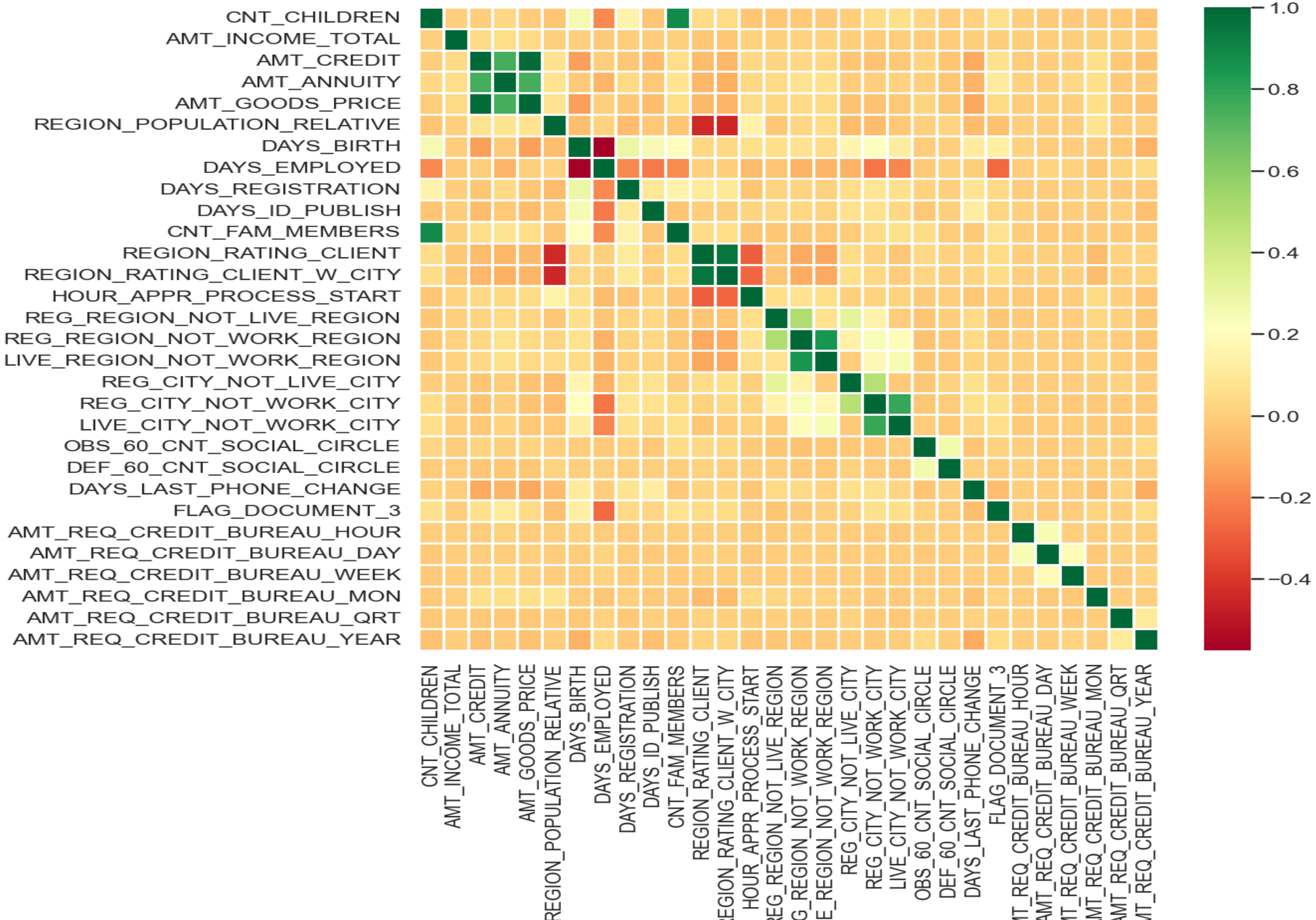
Looking at the above heat map we can say the there is a good corelation between AMT_CREDIT and AMT_GOODS_PRICE columns

Non Defaulters Data Heatmap



In the given HEATMAP we can see that AMT_CREDIT column is highly correlated with AMT_ANNUTY, AMT_GOODS_PRICE, AMT_IMCOME_TOTAL

Defaulter Data Heatmap



- There is a very good correlation between AMT_CREDIT and AMT_GOODS_PRICE

- AMT_ANNUITY also has decent correlation with AMT_CREDIT and AMT_GOODS_PRICE

CONCLUSION

After analysing the datasets we came to a conclusion that bank should keep in mind before approving or rejecting a loan so that they don't face any loss.

- It is safe to provide loans to Students and Businessmen as they don't default much often.
- There is a very good correlation between Amount of Goods and Credit Amount.
- Applicant with more Family members or children tend to default loan so higher interest should be charged on their loans.
- Applicants who live in Office apartments and those with Academic degree tend to default less.
- The Unemployed and Maternity leave categories are the riskiest as they have the highest defaulters.

SUGGESTION

- 70% of applicants are repeaters and new applicants are only 20% company should try to increase their new customer rate by conducting more marketing campaigns and introducing new offers to attract new customers

THANK
YOU