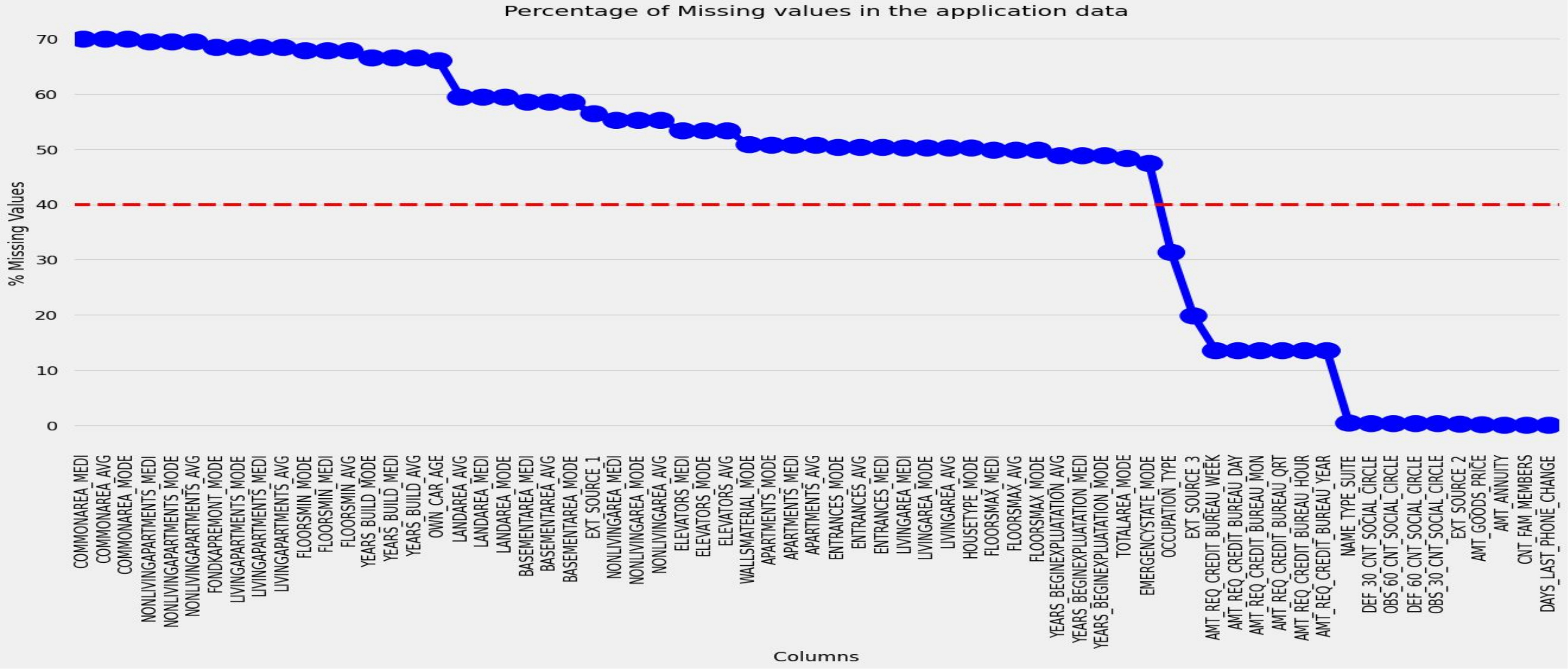
Exploratory Data Analysis

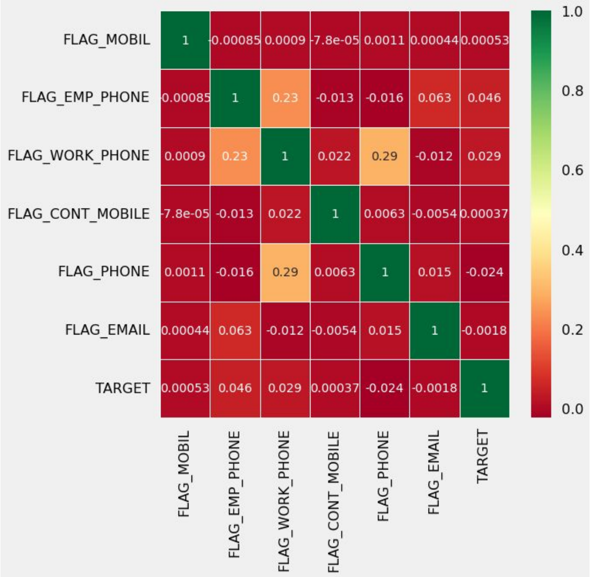
Bank Loan Default Risk Analysis

Tom Mathews

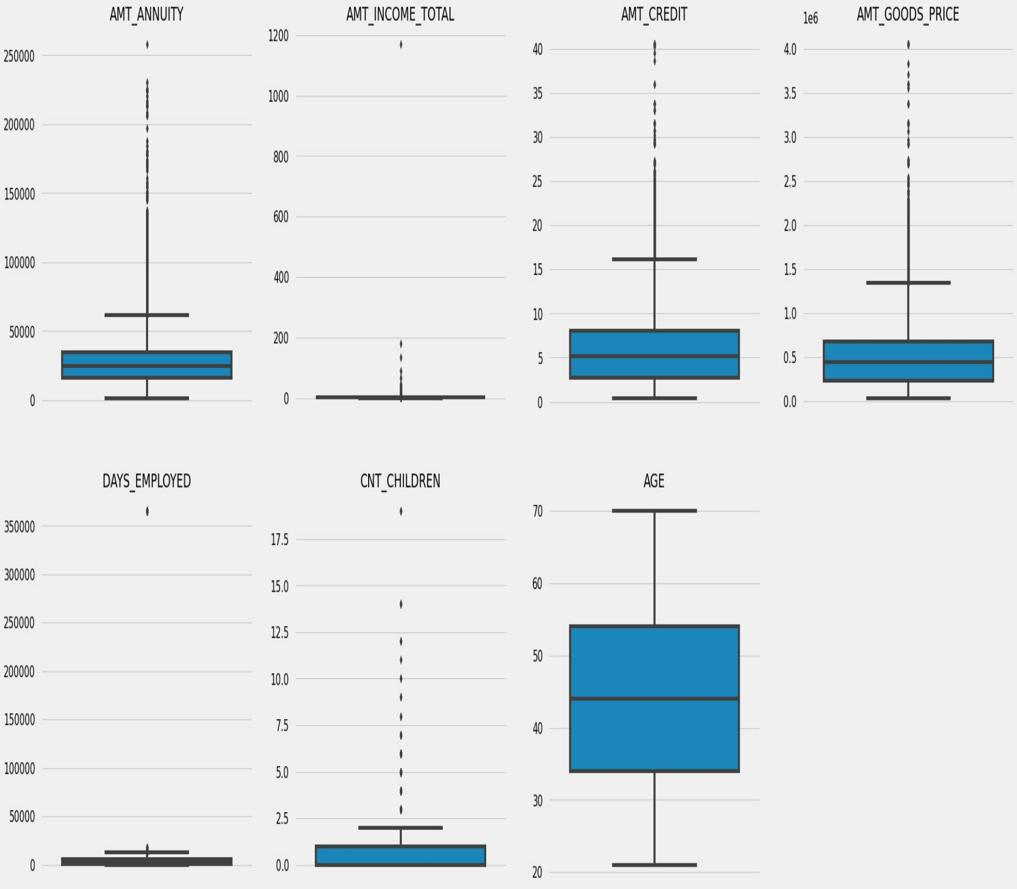
Missing Values in Application Data



There seems to be very low to no correlation between flags of mobile phone, email etc with loan repayment; thus these columns can be deleted

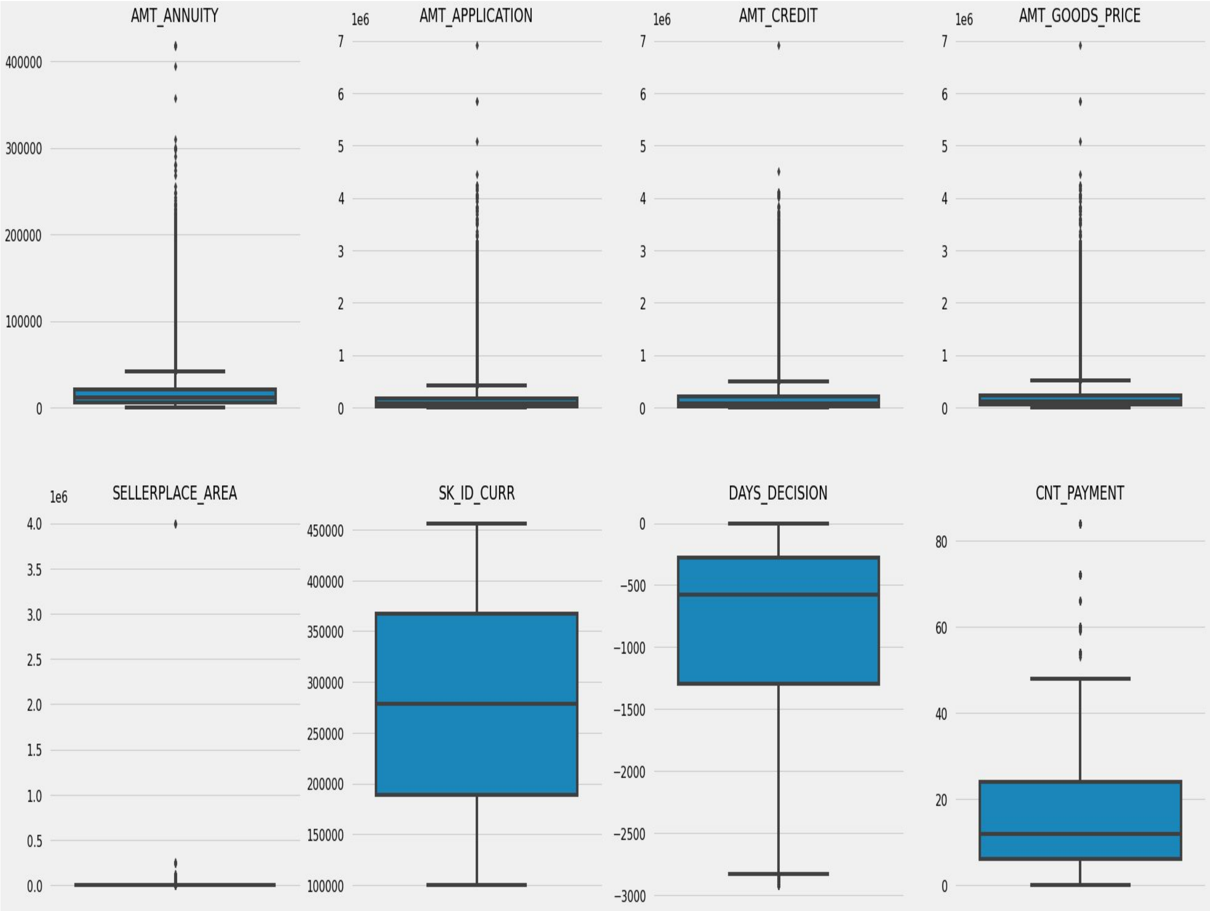


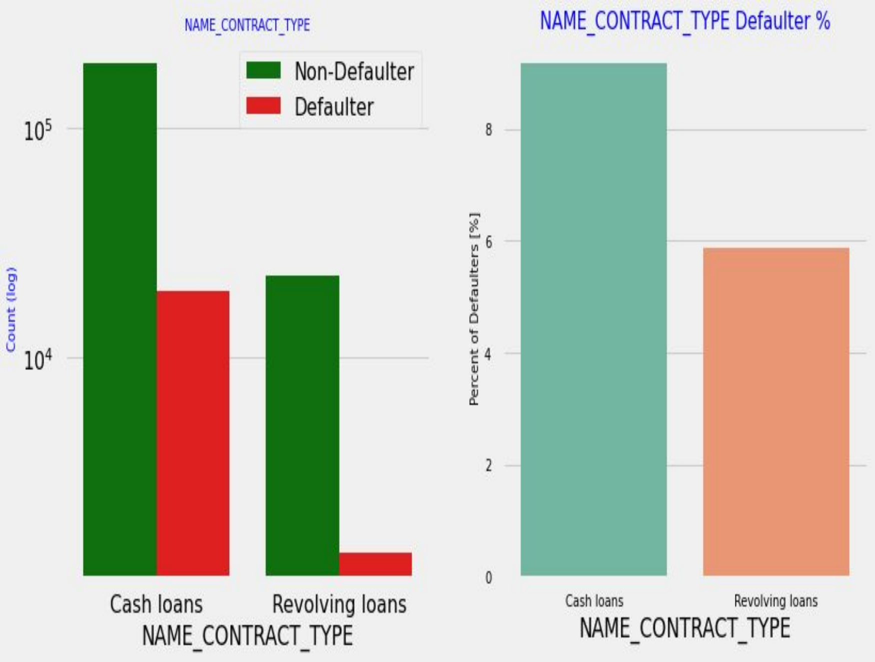
- It can be seen that in current application data
- AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE, CNT_CHILDREN have some number of outliers.
- AMT_INCOME_TOTAL has a huge number of outliers which indicate that few of the loan applicants have high income when compared to the others.
- AGE has no outliers, which means the data available is reliable.
- DAYS_EMPLOYED has outlier values around, 350000(days) which is around 958 years which is impossible and hence this has to be incorrect entry.

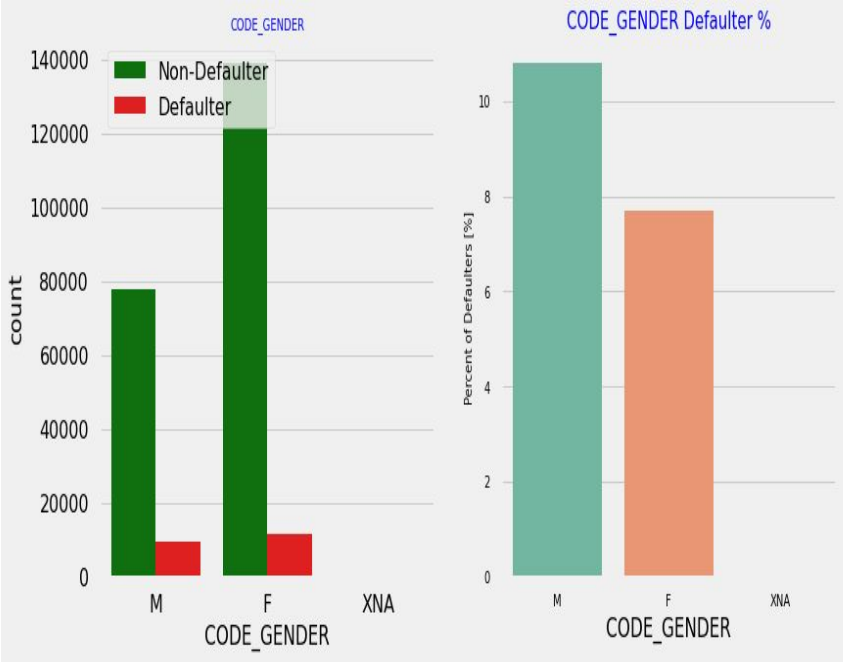


It can be seen that in previous application data

- AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.
- CNT_PAYMENT has few outlier values.
- SK_ID_CURR is an ID column and hence no outliers.
- DAYS_DECISION has little number of outliers indicating that these previous applications decisions were taken long back.

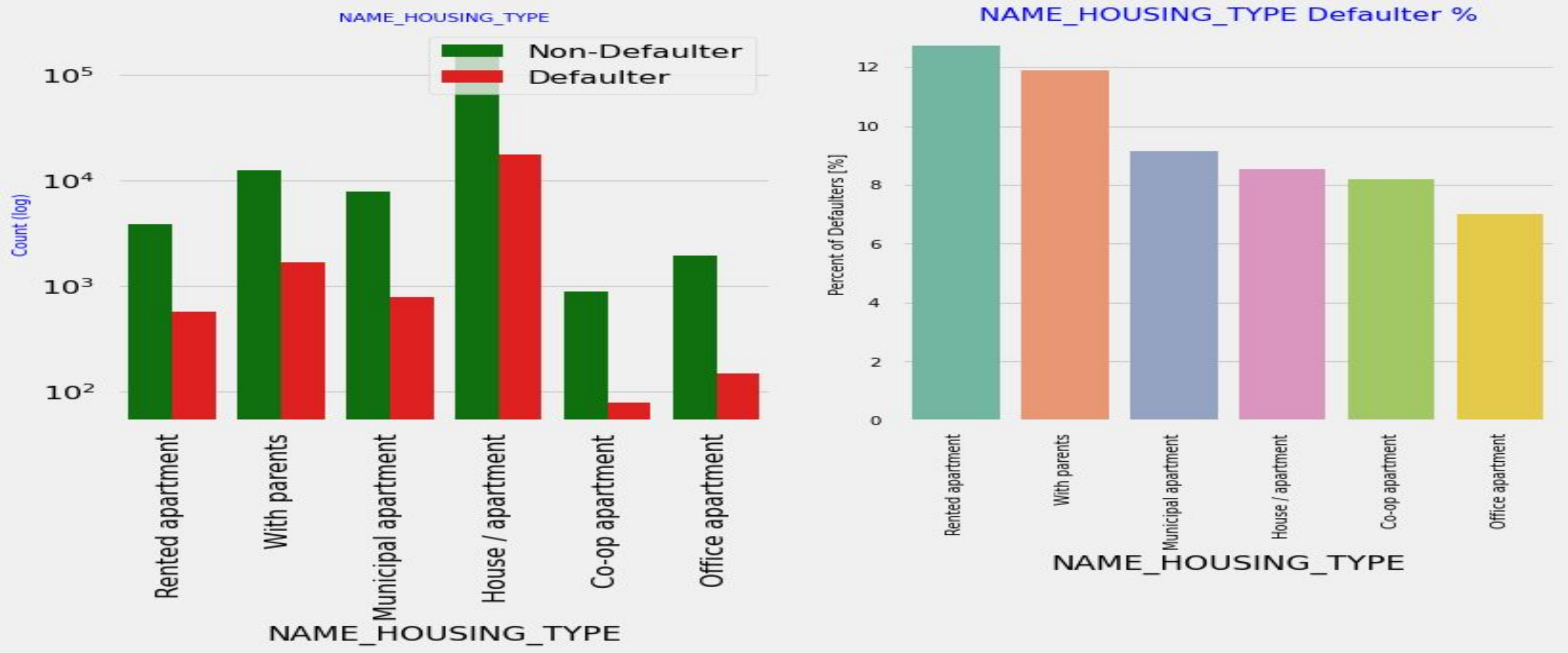




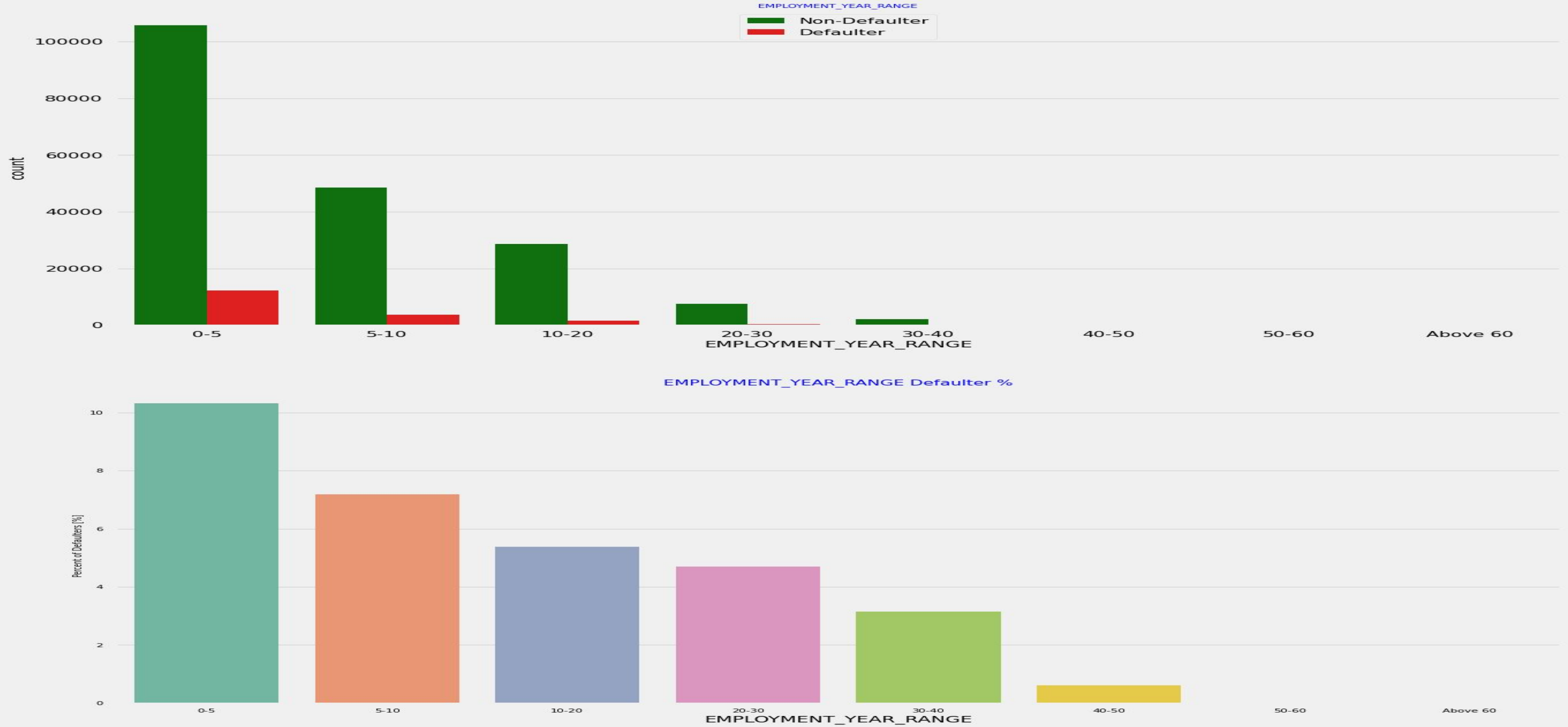


- Contract type: Revolving loans are just a small fraction (10%) from the total number of loans.
- In the same time, a larger amount of Revolving loans, comparing with their frequency, are not repaid.

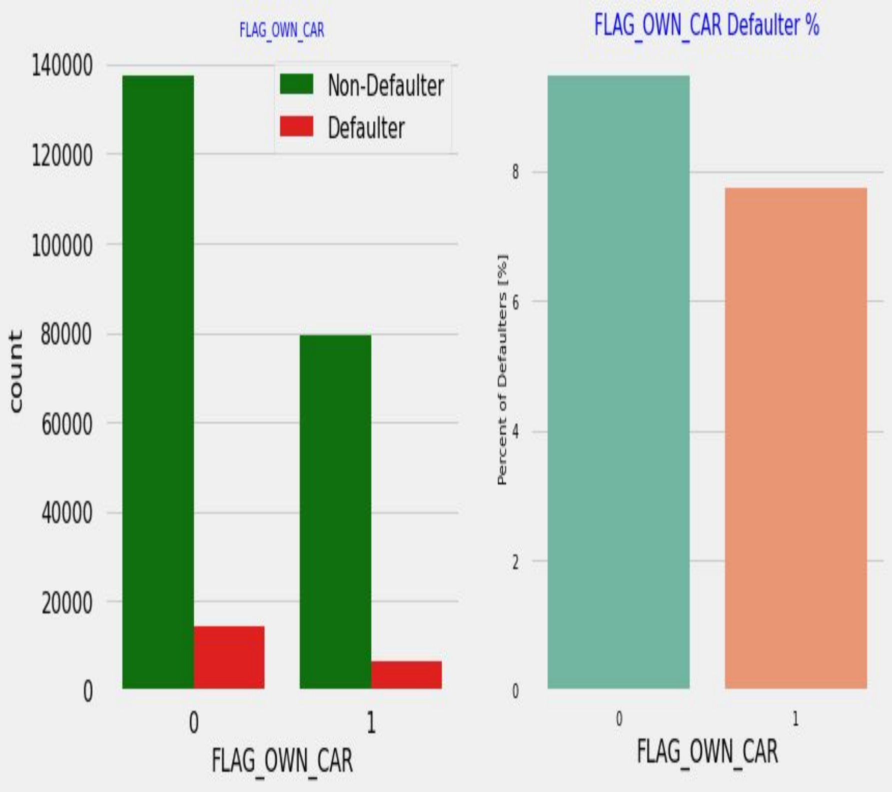
- The number of female clients is almost double the number of male clients.
- Based on the percentage of defaulted credits, males have a higher chance of not returning their loans (approx 10%), comparing with women (approx 7%)

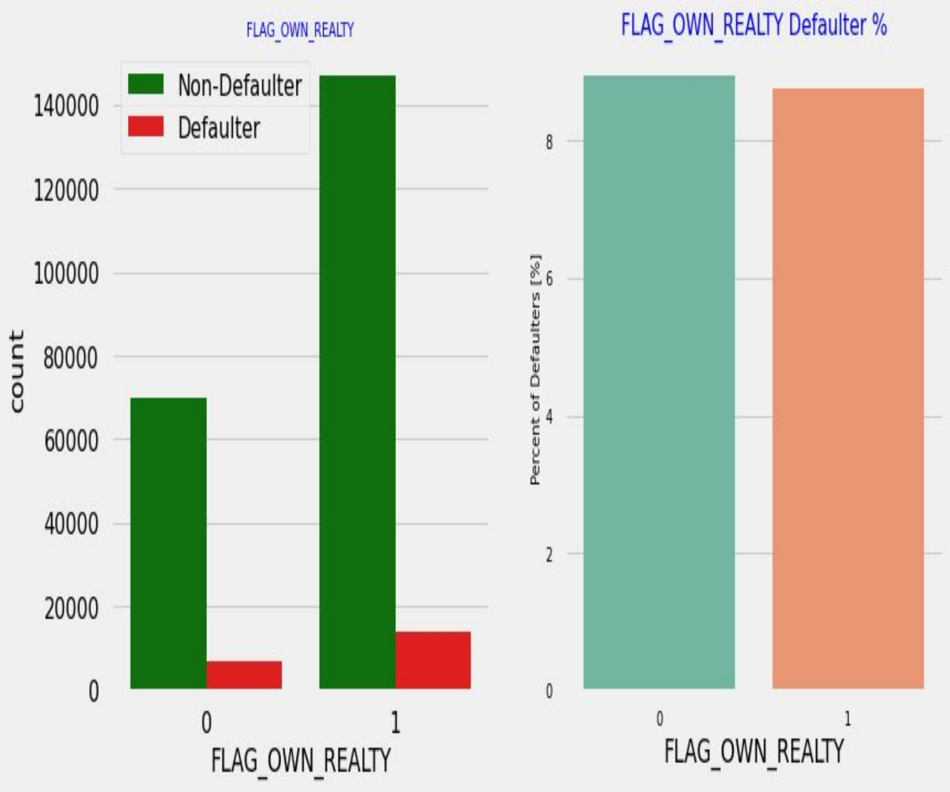


- Majority of people live in House/apartment
- People living in office apartments have lowest default rate
- People living with parents (approx 11.5%) and living in rented apartments (>12%) have higher probability of defaulting



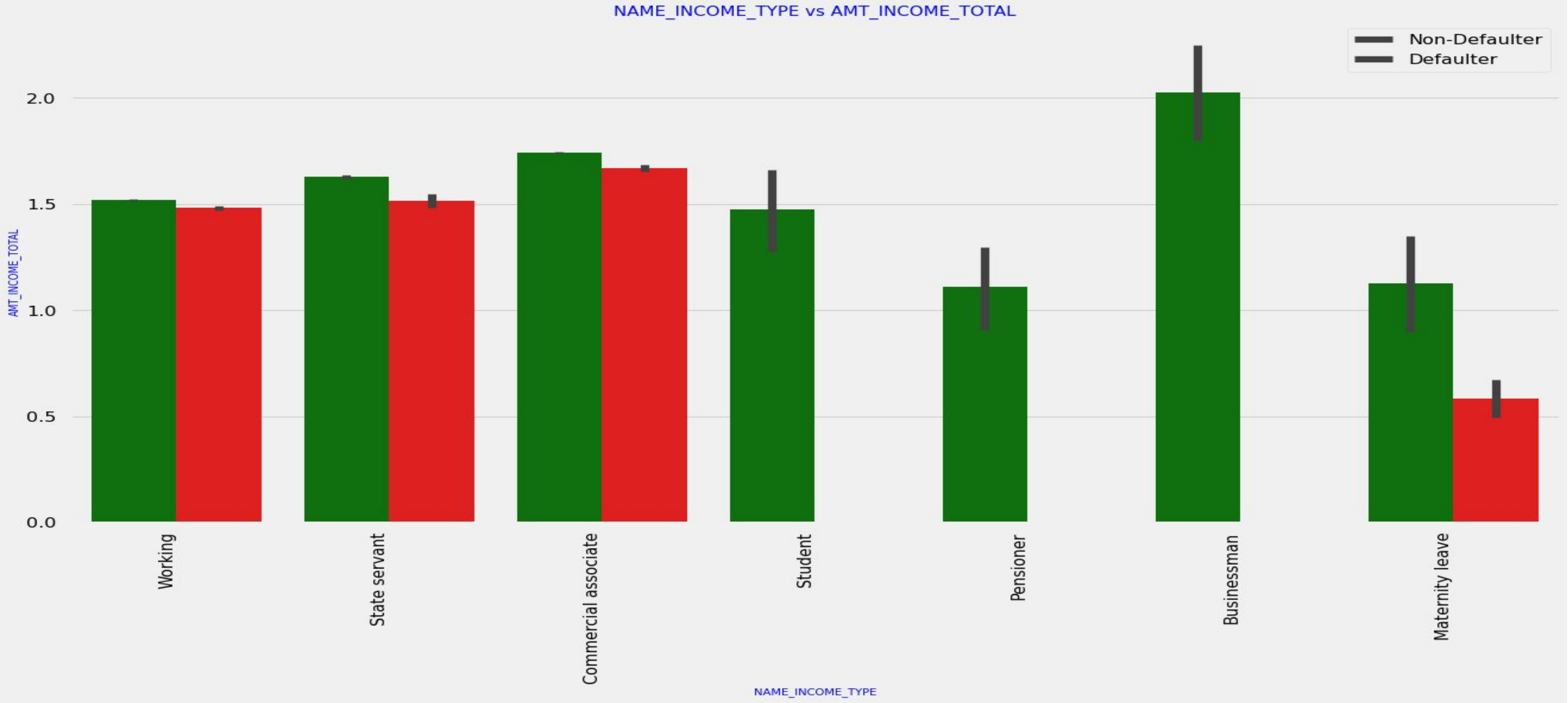
- Majority of the applicants have been employed in between 0-5 years. The defaulting rating of this group is also the highest which is 10%
- With increase of employment year, defaulting rate is gradually decreasing with people having 40+ year experience having less than 1% default rate





- Clients who own a car are half in number of the clients who don't own a car.
- But based on the percentage of defaulters, there is no correlation between owning a car and loan repayment as in both cases the default percentage is almost the same.

- The clients who own real estate are more than double of the ones that don't own.
- But the defaulting rate of both categories are around the same (approx 8%).
- Thus, there is no correlation between owning a reality and defaulting the loan.



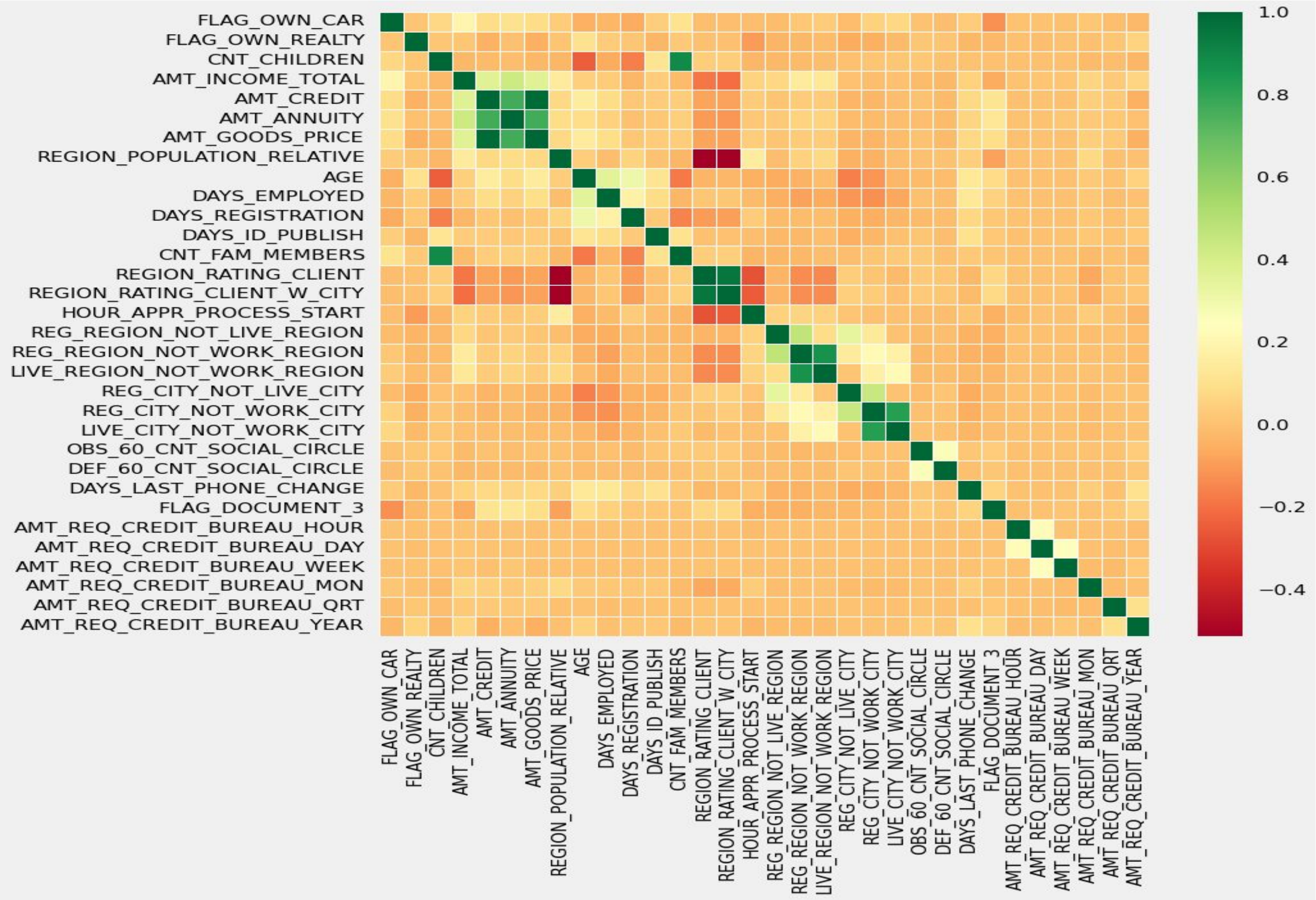
It can be seen that businessman's income is the highest and the estimated range with default 95% confidence level seem to indicate that the income of a businessman could be in the range of slightly close to 400K and slightly above 1M

Correlating factors amongst Non-Defaulters.

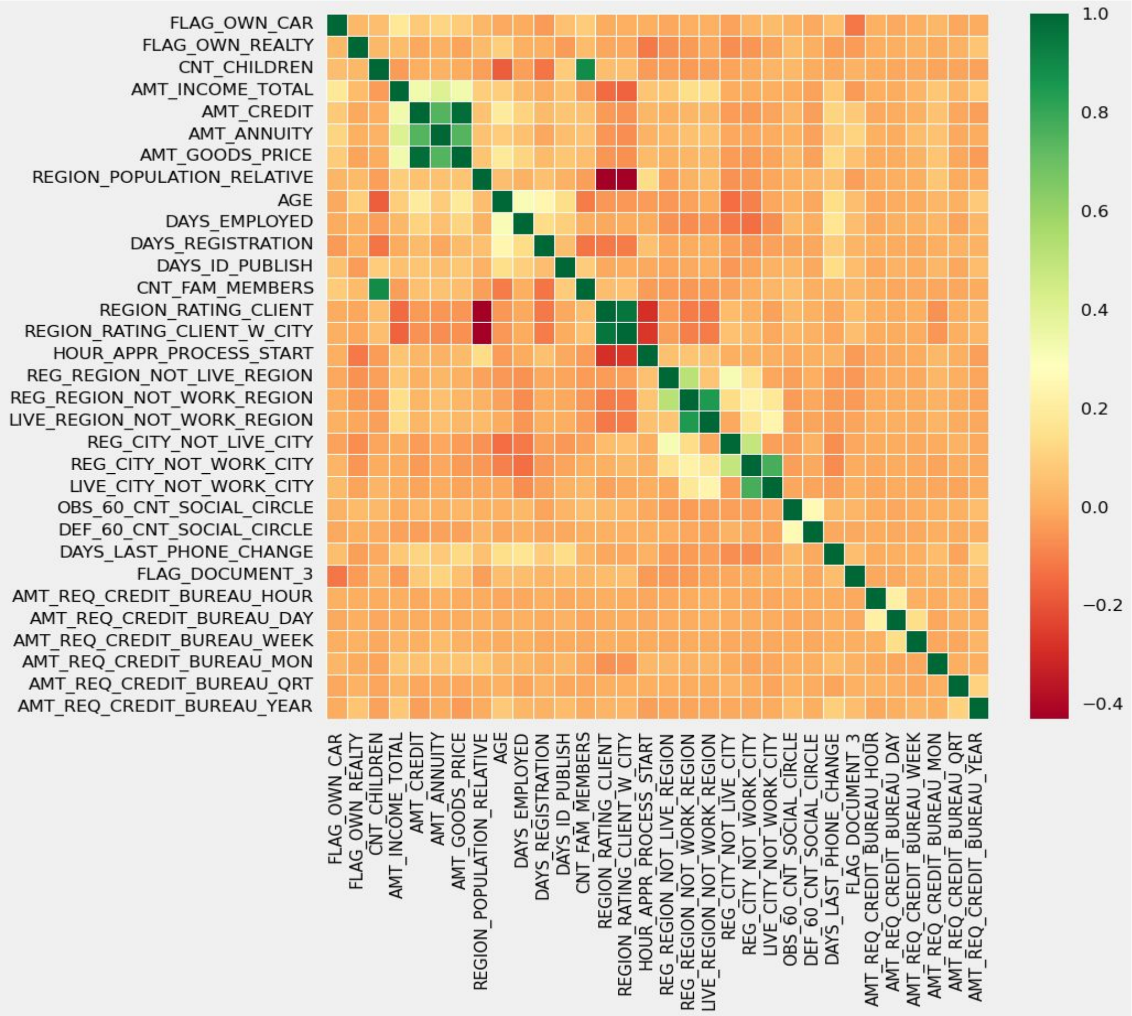
Credit amount is highly correlated with:

- amount of goods price
- loan annuity
- total income

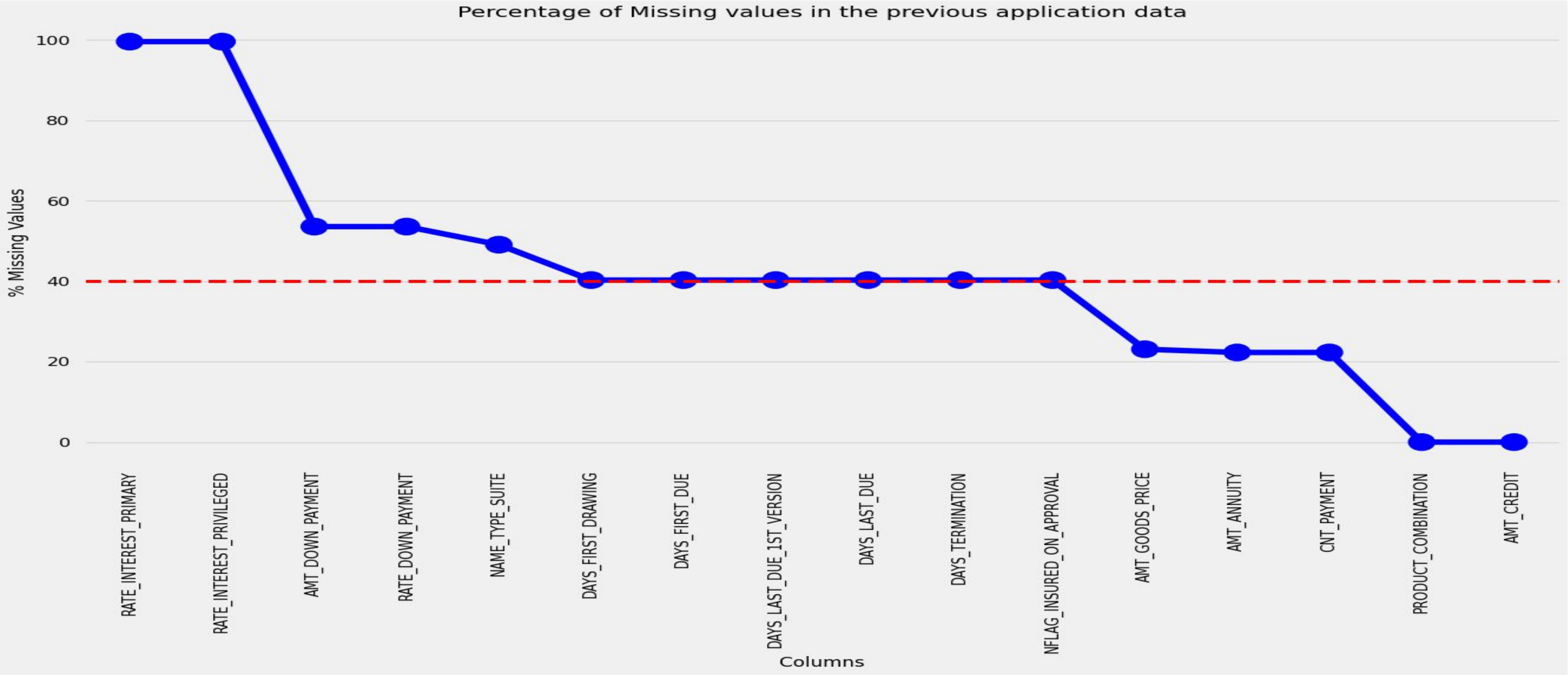
We can also see that Non-Defaulters have high correlation in number of days employed.

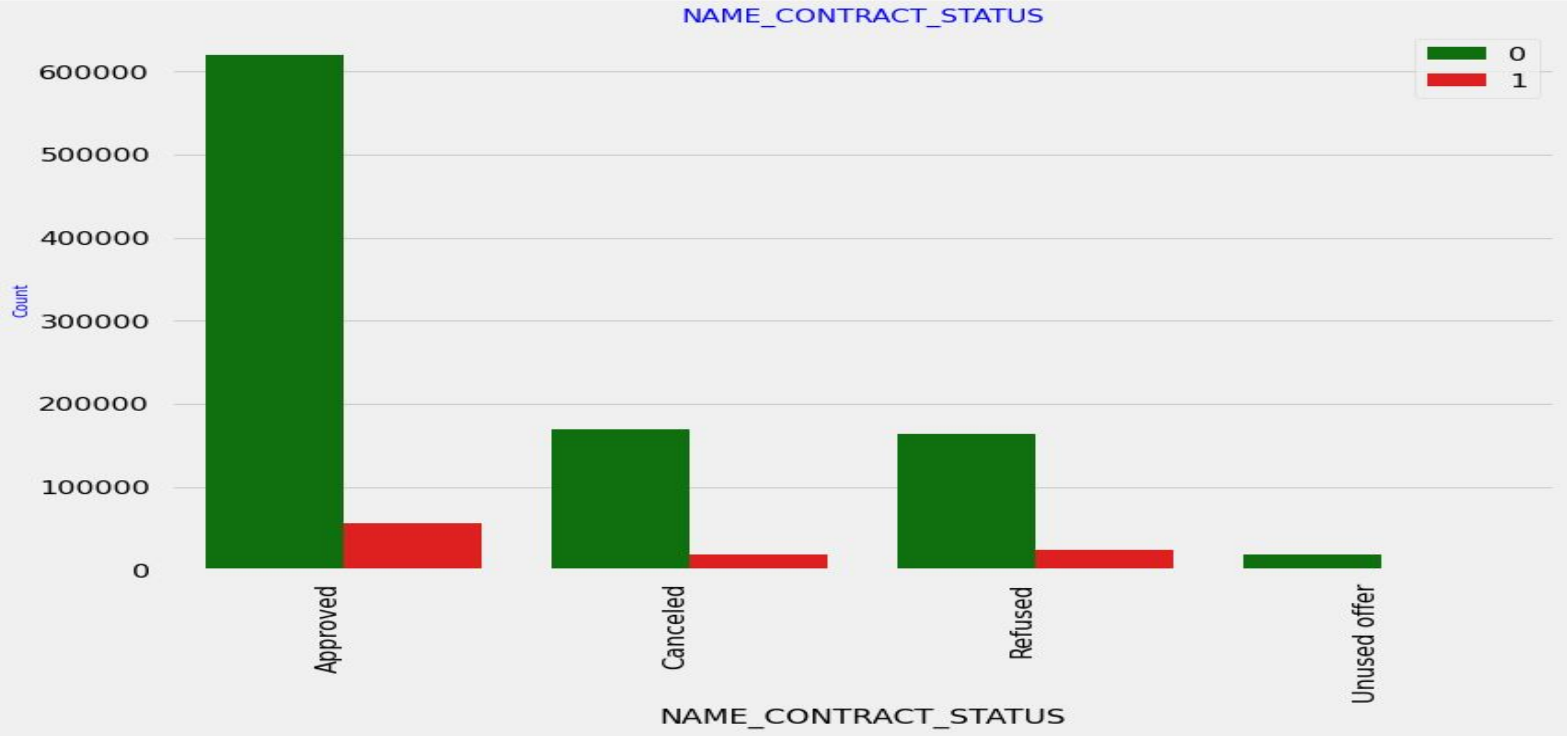


- Credit amount is highly correlated with amount of goods price which is same as non-defaulters.
- But the loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to non-defaulters(0.77)
- We can also see that repayers have high correlation in number of days employed(0.62) when compared to defaulters(0.58).
- There is a severe drop in the correlation between total income of the client and the credit amount(0.038) amongst defaulters whereas it is 0.342 among non-defaulters.
- Days_birth and number of children correlation has reduced to 0.259 in defaulters when compared to 0.337 in non-defaulters.
- There is a slight increase in defaulted to observed count in social circle among defaulters(0.264) when compared to non-defaulters(0.254)



Missing Data in Previous Application





- 90% of the previously cancelled client have actually repaid the loan. Revisiting the interest rates would increase business opportunity for these clients
- 88% of the clients who have been previously refused a loan has paid back the loan in current case.
- Refusal reason should be recorded for further analysis as these clients would turn into potential repaying customer

Conclusion - Non-Defaulters

- NAME_EDUCATION_TYPE: Academic degree has less defaults.
- NAME_INCOME_TYPE: Student and Businessmen have no defaults.
- REGION_RATING_CLIENT: RATING 1 is safer.
- ORGANIZATION_TYPE: Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%.
- DAYS_BIRTH: People above age of 50 have low probability of defaulting.
- DAYS_EMPLOYED: Clients with 40+ year experience having less than 1% default rate.
- AMT_INCOME_TOTAL:Applicant with Income more than 700,000 are less likely to default.
- NAME_CASH_LOAN_PURPOSE: Loans bought for Hobby, Buying garage are being non-defaulters mostly.
- CNT_CHILDREN: People with zero to two children tend to repay the loans.

Conclusion - Defaulters

- CODE_GENDER: Men are at relatively higher default rate.
- NAME_FAMILY_STATUS : People who have civil marriage or who are single default a lot.
- NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education.
- NAME_INCOME_TYPE: Clients who are either at Maternity leave OR Unemployed default a lot.
- REGION_RATING_CLIENT: People who live in Rating 3 has the highest defaults.
- OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as the default rate is huge.
- ORGANIZATION_TYPE: Organizations with the highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self-employed people have relative high defaulting rate, and thus should be avoided to be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting.
- AGE: Avoid young people who are in age group of 20-40 as they have higher probability of defaulting.
- DAYS_EMPLOYED: People who have less than 5 years of employment have high default rate.
- CNT_CHILDREN & CNT_FAM_MEMBERS: Client who have children equal to or more than 9 default 100% and hence their applications are to be rejected.
- AMT_GOODS_PRICE: When the credit amount goes beyond 3M, there is an increase in defaulters.

The following attributes indicate that people from these categories tend to default but then due to the number of people and the amount of loan, the bank could provide loan with higher interest to mitigate any default risk thus preventing business loss:

- NAME_HOUSING_TYPE: High number of loan applications are from the category of people who live in Rented apartments & living with parents and hence offering the loan would mitigate the loss if any of those default.
- AMT_CREDIT: People who get loan for 300-600k tend to default more than others and hence having higher interest specifically for this credit range would be ideal.
- AMT_INCOME: Since 90% of the applications have Income total less than 300,000 and they have high probability of defaulting, they could be offered loan with higher interest compared to other income category.
- CNT_CHILDREN & CNT_FAM_MEMBERS: Clients who have 4 to 8 children has a very high default rate and hence higher interest should be imposed on their loans.
- NAME_CASH_LOAN_PURPOSE: Loan taken for the purpose of Repairs seems to have highest default rate. A very high number applications have been rejected by bank or refused by client in previous applications as well which has purpose as repair or other. This shows that purpose repair is taken as high risk by bank and either they are rejected, or bank offers very high loan interest rate which is not feasible by the clients, thus they refuse the loan. The same approach could be followed in future as well.

Other Suggestions

- 90% of the previously cancelled client have actually repaid the loan. Record the reason for cancellation, which might help the bank to determine and negotiate terms with these repaying customers in future to increase business opportunity.
- 88% of the clients who were refused by bank for loan earlier have now turned into a repaying client. Hence, documenting the reason for rejection could mitigate the business loss and these clients could be contacted for further loans.