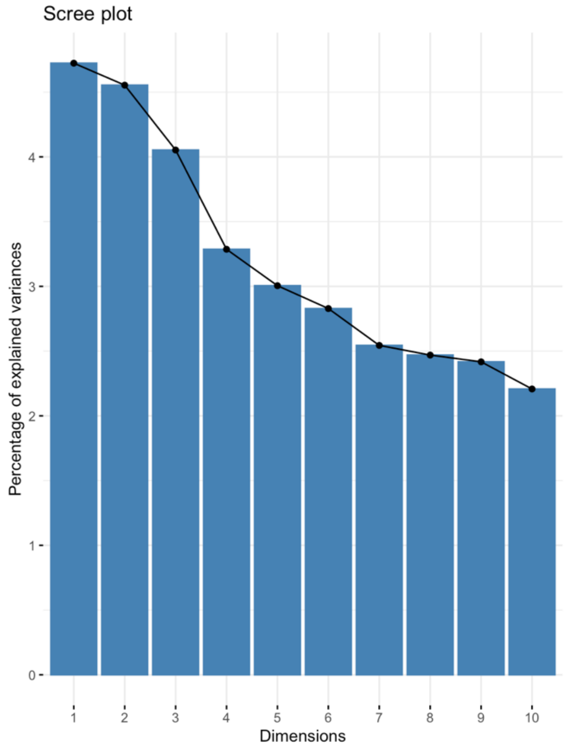
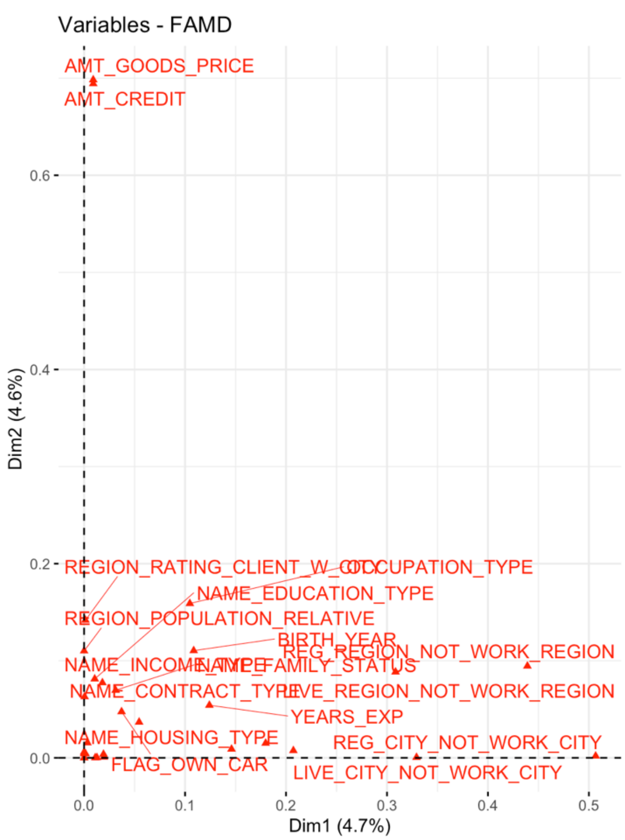
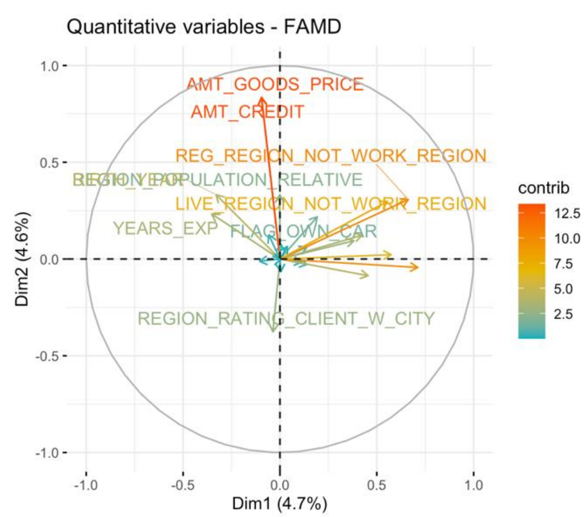
Appendix – Team 3 Section A – Data Science Team Project
Principal component analysis:



The percentage variance of the data set explained by the potential principal components is quite low.

PC1 Vs PC2





Exploratory data analysis continued:

Figure 1: Histogram of Type of Occupation of Applicant

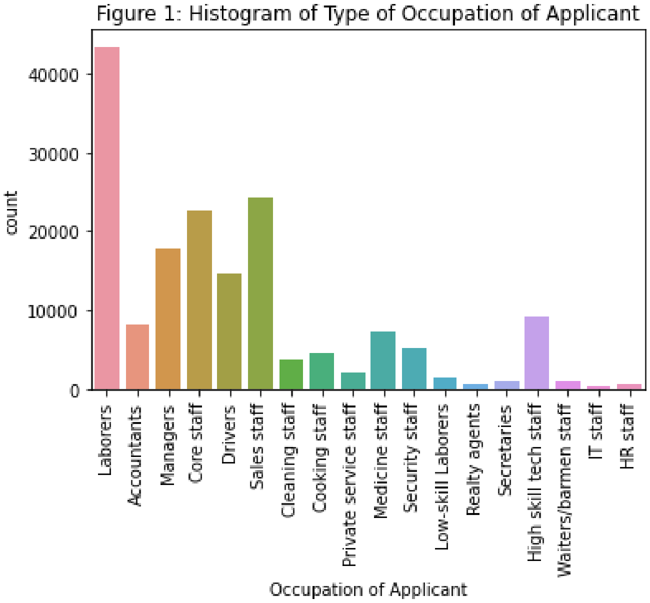


Figure 2: Histogram of The Individuals Previous Loan Status

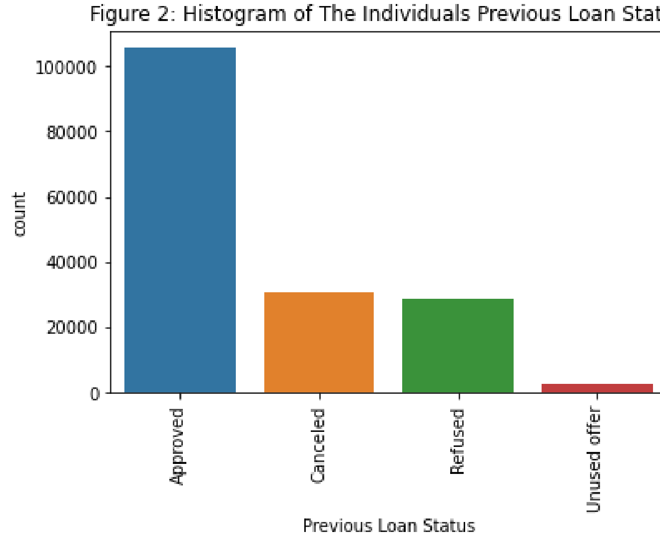


Figure 3: Histogram of The Type of Loan

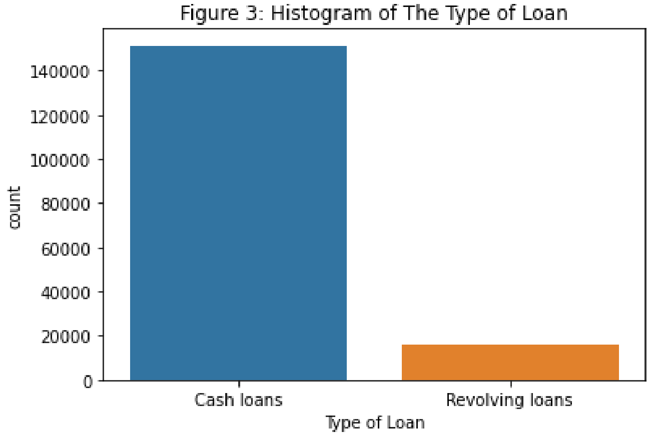


Figure 4: Histogram of The Education Level of Applicants

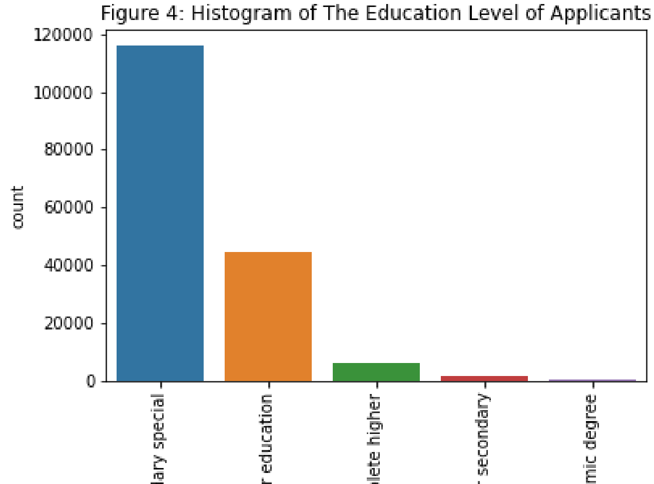


Figure 5: Plot of the Occupation Type for the Default Status

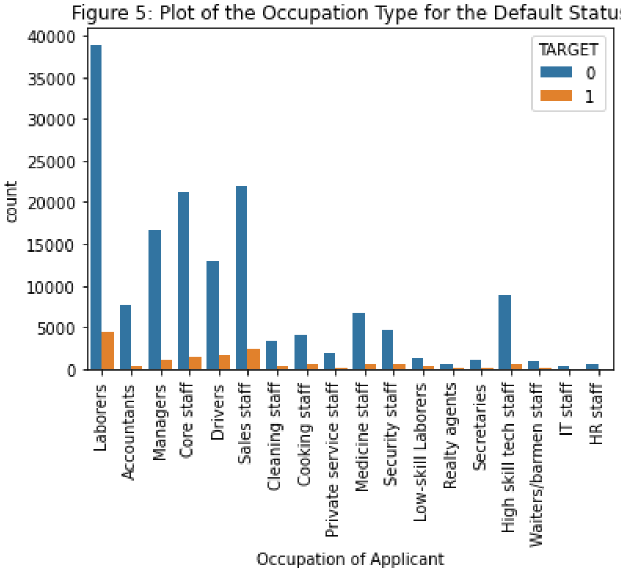
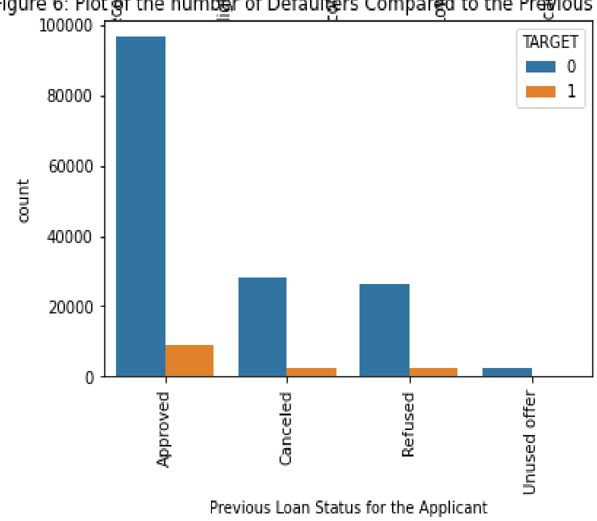


Figure 6: Plot of the number of Defaulters Compared to the Previous Loan Status



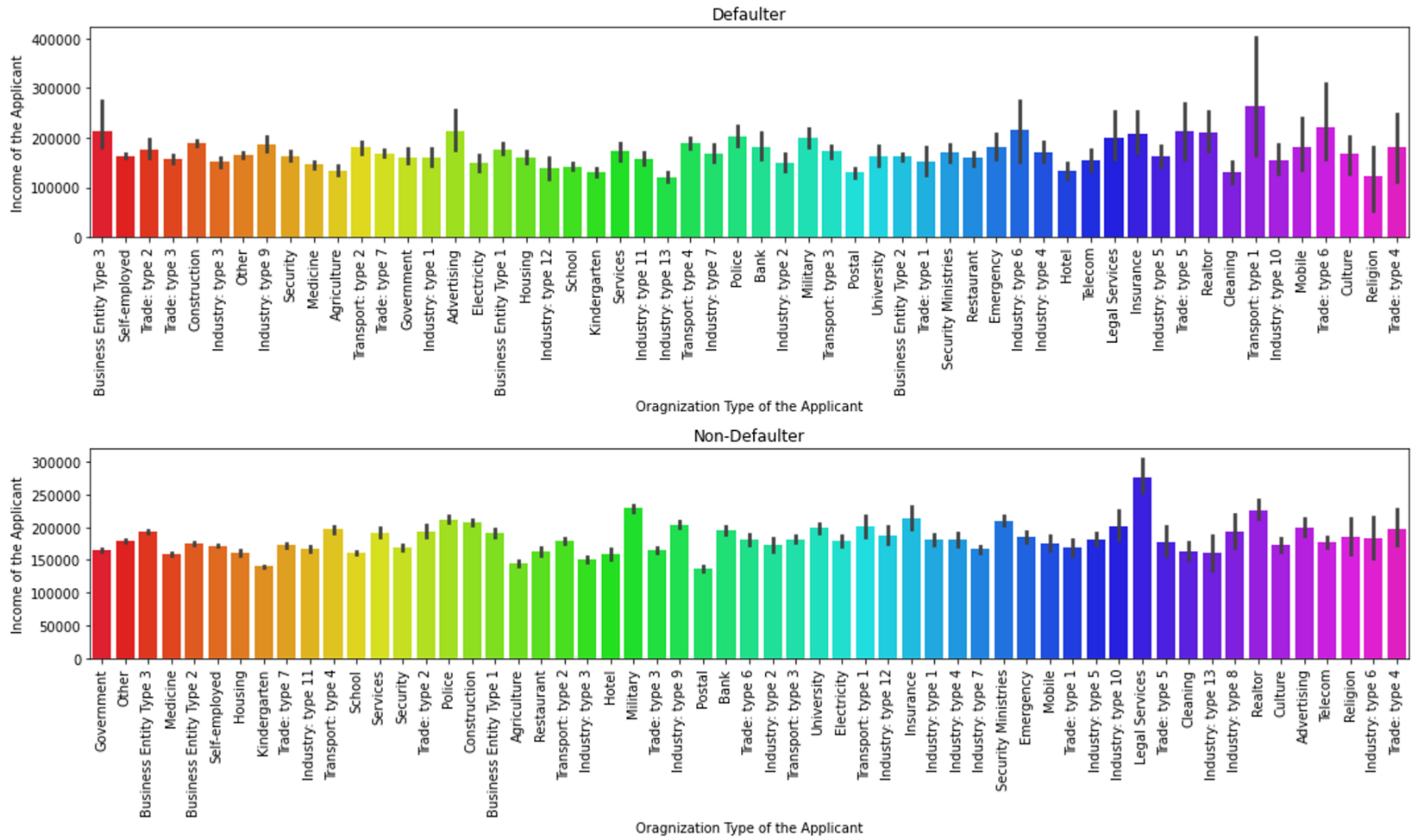


Figure 7: Comparing Organization Type and Income for Defaulters and Non-Defaulters

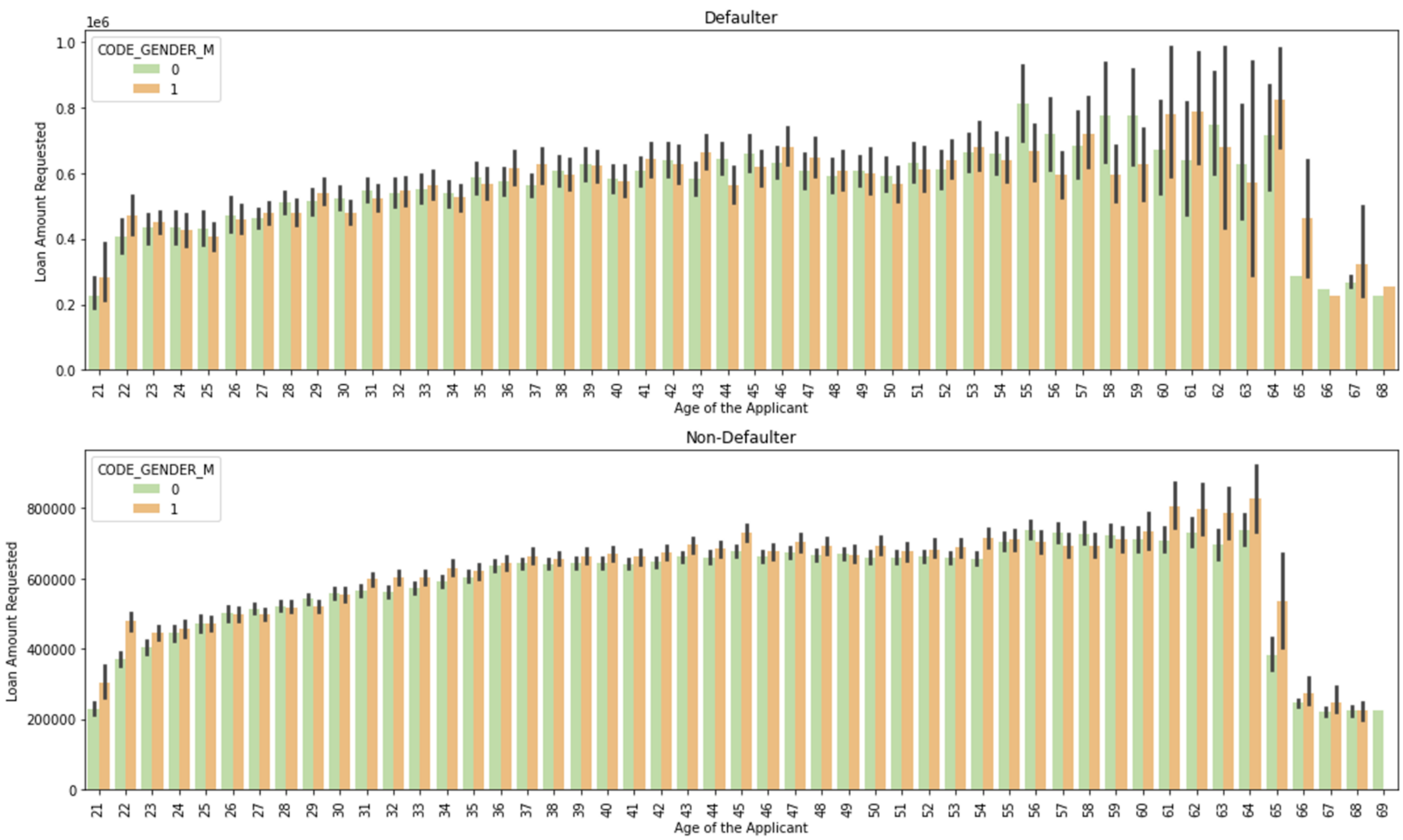


Figure 8: Comparing age group and amount of credit (amount loan applied) for defaulters and non-defaulters categorized by male and female.

Figure 9 : Histogram of Income Type

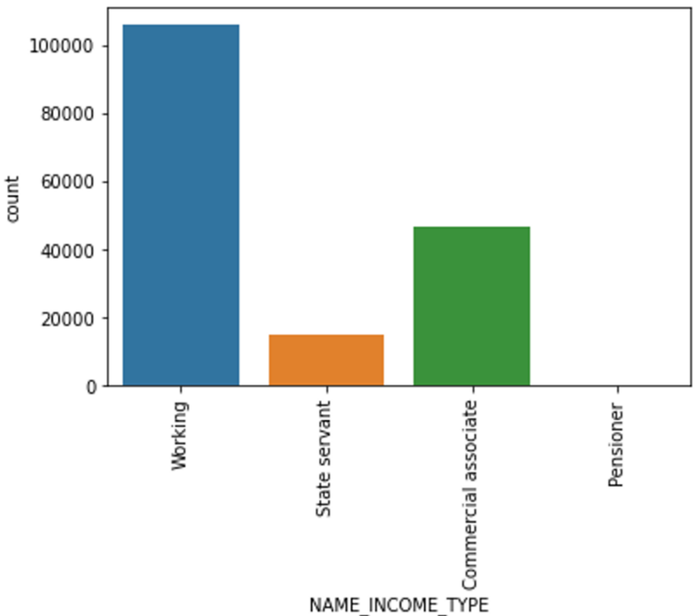


Figure 10: Histogram of the type of housing of the applicant

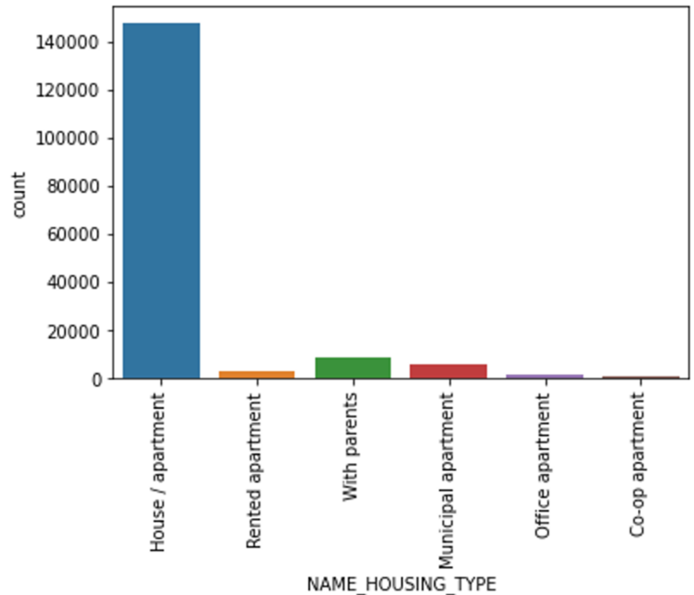


Figure 11: Histogram of the Type of accompaniment

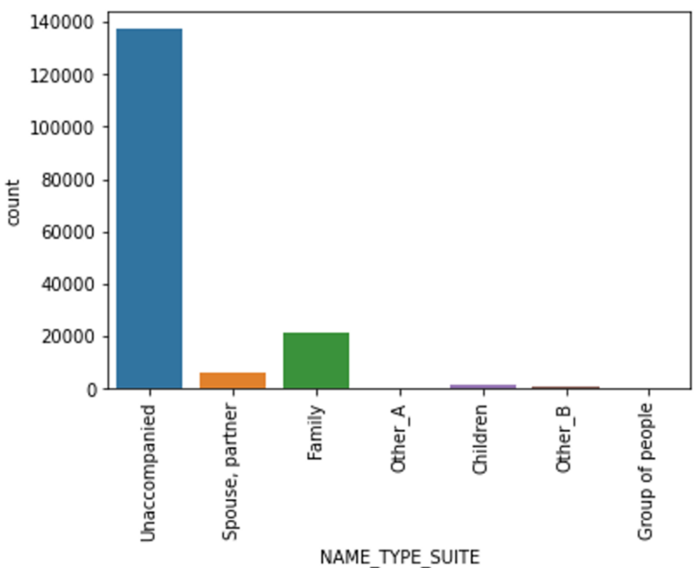


Figure 12: Family status of applicant

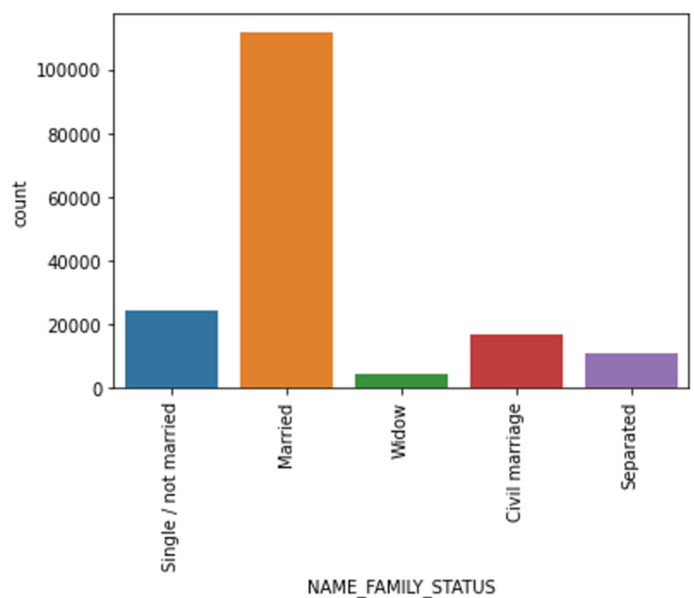
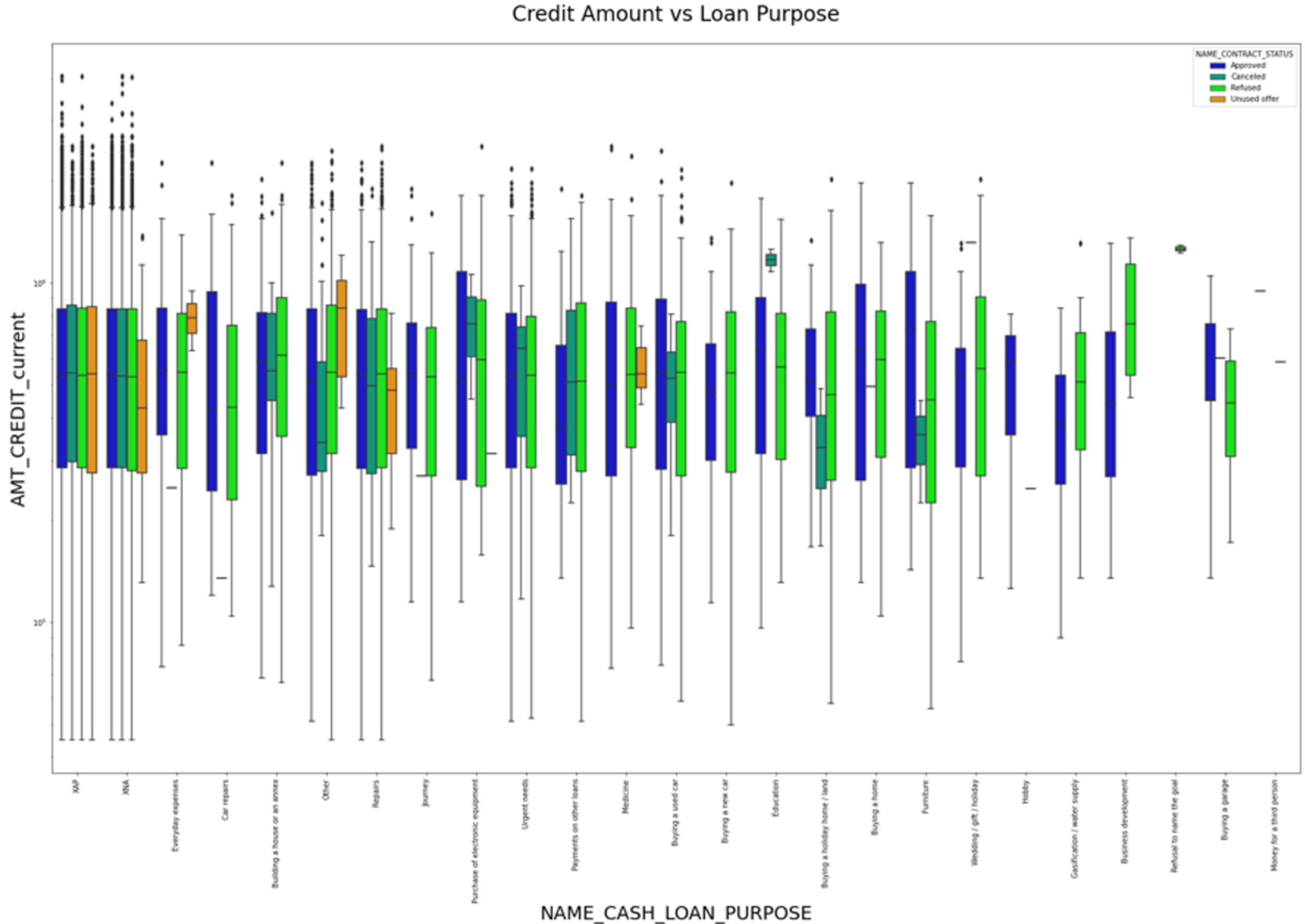
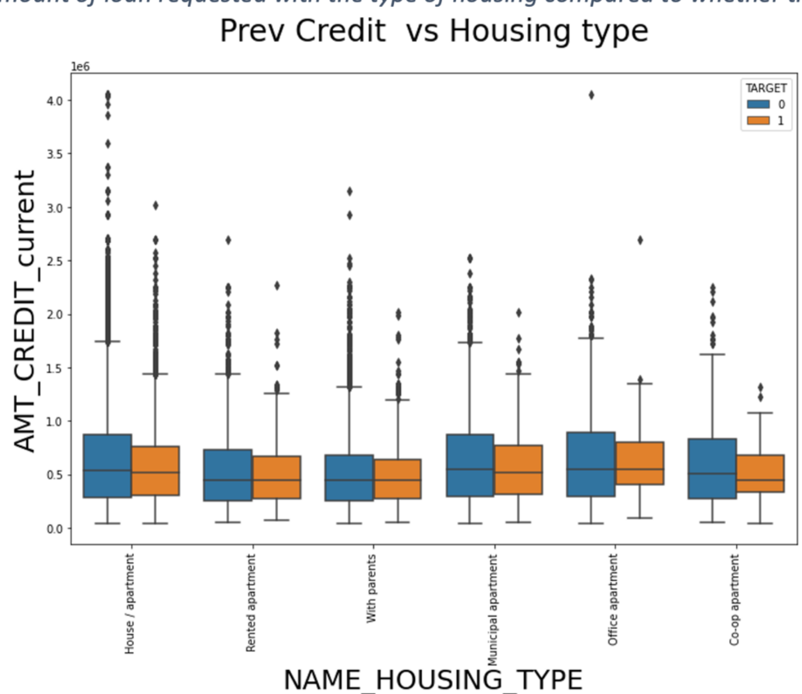


Figure 13: Comparing the loan amount with the reason for acquiring the loan with the status historically.



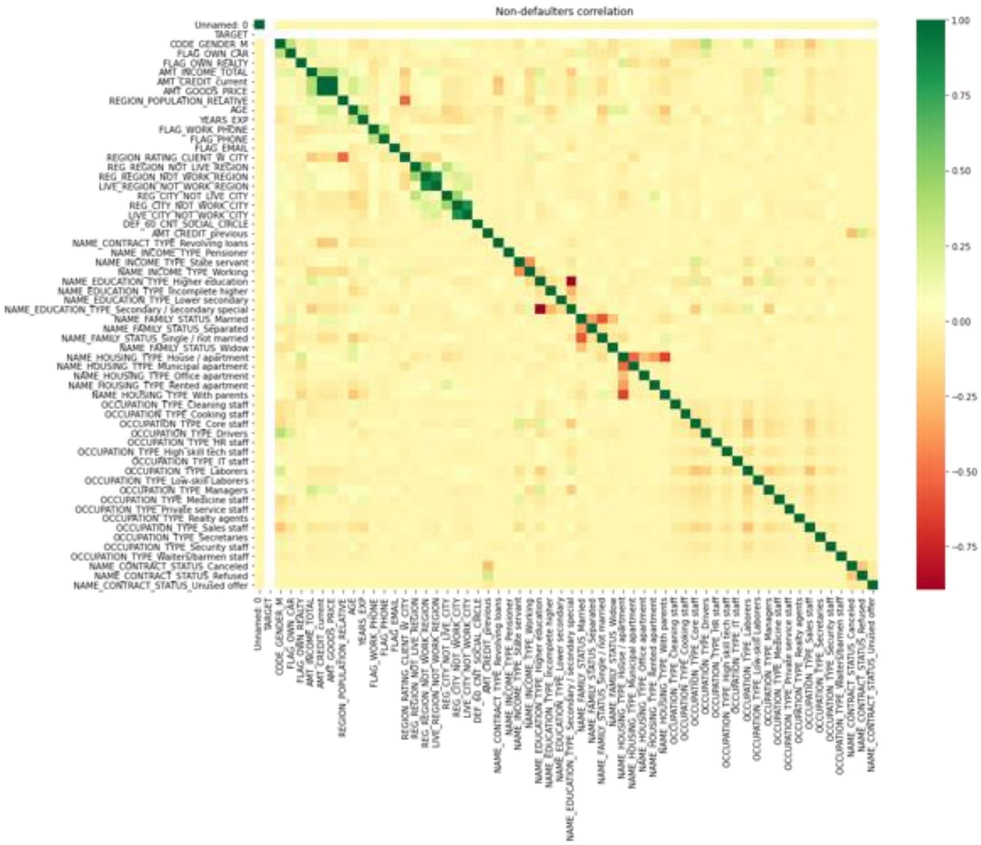
The credit amount of Loan purposes like Buying a home, Buying a land, Buying a new car and Building a house is higher in which refused and cancelled are more prominent.

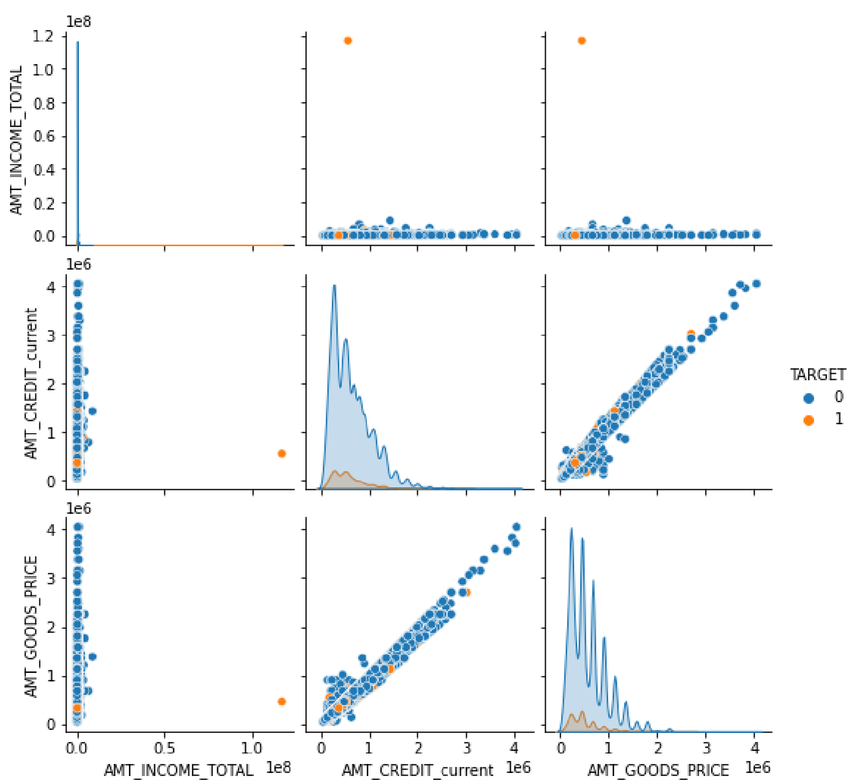
Figure 14: Amount of loan requested with the type of housing compared to whether they default or not.



People taking higher credit amount living in office apartment tend to be non- defaulters as must be having less expenses on rent and maintenance charges whereas co-op apartment people are prominently defaulters

Correlation plots:





Interest rate calculation :

We choose the interest rate of loan from Fair Isaac Corporation (FICO) website. FICO credit score is the most common credit scores used by many lenders to determine if a person qualifies for a credit card, mortgage, or other loan in the United States. Originally, FICO credit score divides people into 6 categories:

FICO Score	APR
760-850	6.481 %
700-759	6.703 %
680-699	6.88 %
660-679	7.094 %
640-659	7.524 %
620-639	8.07 %

In our project, combing the above interest rate in reality and available variables in our dataset, we use age and income to measure the credit level of clients. We assume that old people have lower possibility to default the loan than young people. The more income the clients have, the less likely they will default the loan.

Income credit level (112500 is 1st quartile and 225000 is 3rd quartile)
low: income<112500   medium:112500<income < 225000   high: income>225000

- Age credit level (40 is the median)
  low: age<40 high: age>40

Based on these two factors, the credit level will be

- Low credit level: low income + low age, low income + high age
- Medium credit level: medium income+ low age, high income +low age
- High credit level: medium income+ high age, high income +high age

According to the interest rate in reality,
We set 6% as the annual interest rate for client with high credit level,
7 % as the annual interest rate for client with medium credit level,
8% as the annual interest rate for client with low credit level.

We assume the number year each client apply for the loan is 10 years, so the total loan terms are 120.