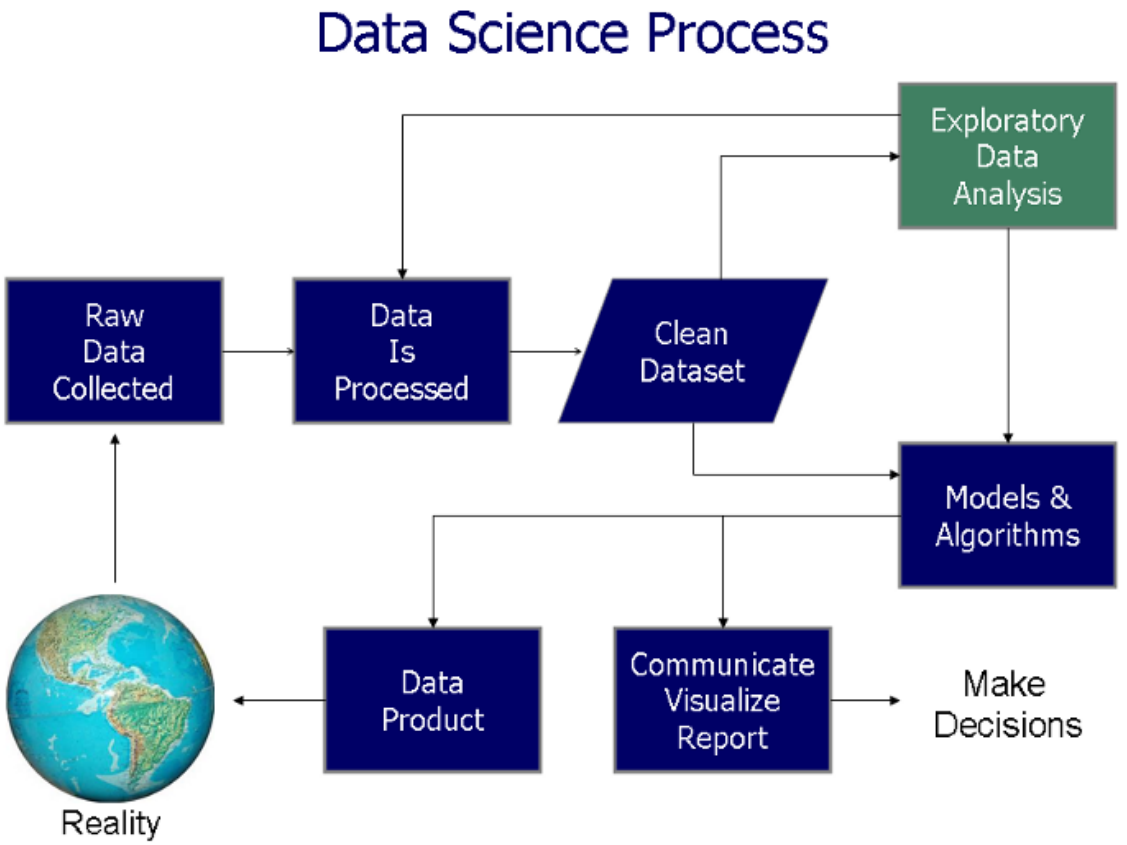
Credit EDA Case Study

Devesh Singh

Suchith K C

Flow Chart


Data Science Process
Exploratory Data Analysis
Raw Data Collected
Data Is Processed
Clean Dataset
Models & Algorithms
Reality
Data Product
Communicate Visualize Report
Make Decisions

Flow Chart

Raw Data was collected using Sampling Techniques from Real World Data

Application Data contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.

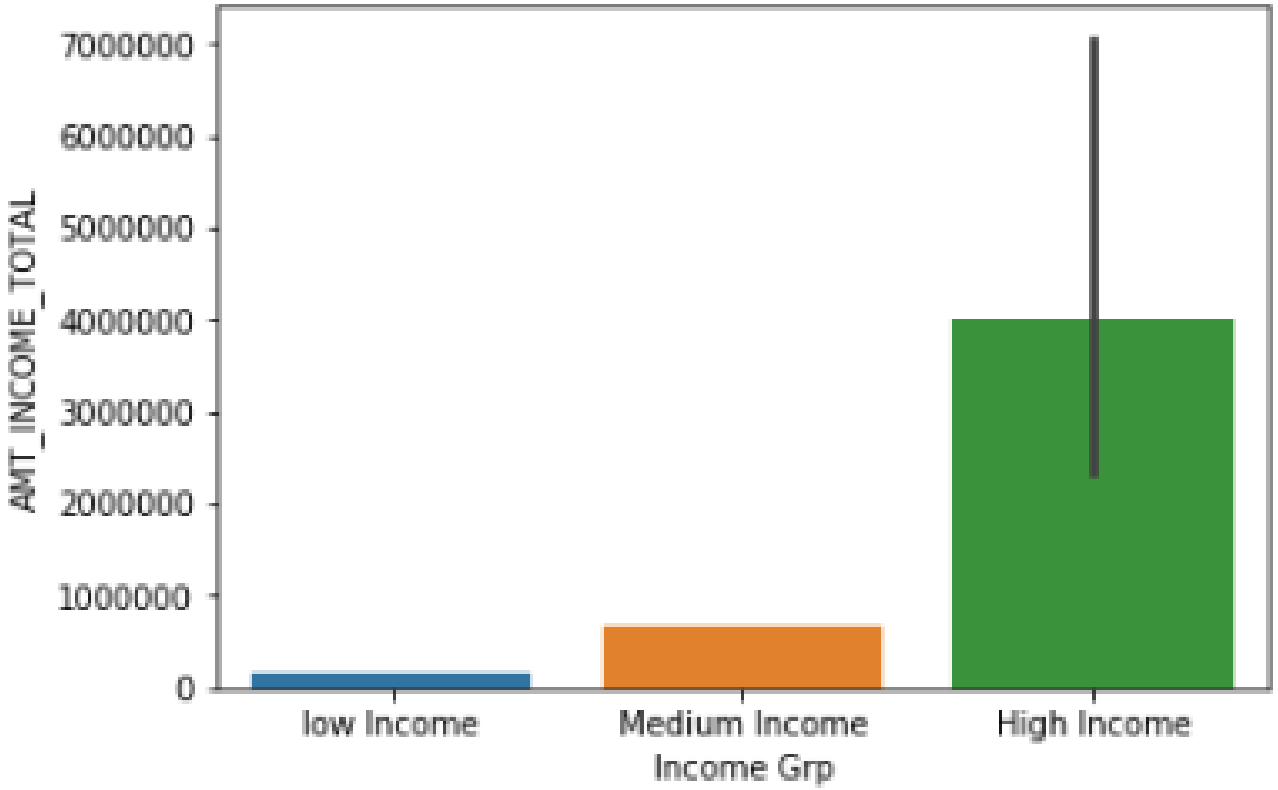
Previous Application contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

Columns Description is data dictionary which describes the meaning of the variables.

After Data Processing, the data was cleaned for Exploratory Data Analysis

Models are developed to generate Data Insights

Communicated the Data Product through Visualizations and Reports



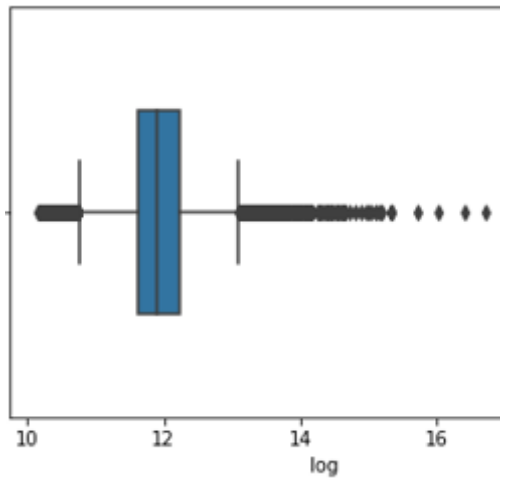
Loan Distribution Among Various Income Groups

- A study between Income and Loan disbursed shows that the High-Income Groups were able to secure more Loans than the Medium-Income and Low-Income Groups

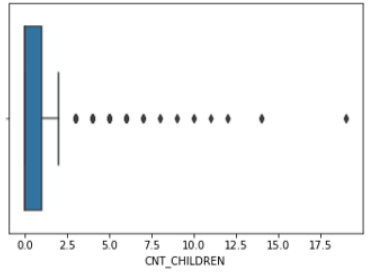
- Here we converted AMT_INCOME_TOTAL & AMT_CREDIT in to Low, medium and high by binning the loan amount to identify the distribution pattern

Outliers for the Numerical Variable

1. AMT_CREDIT
2. AMT_INCOME_TOTAL
3. CNT_CHILDREN

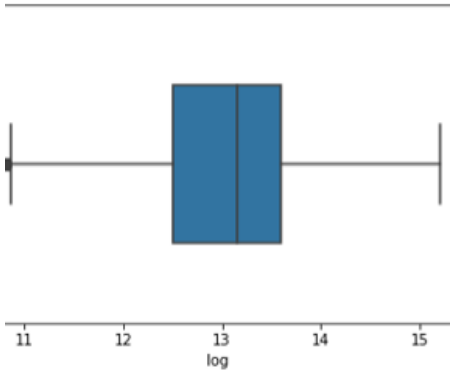


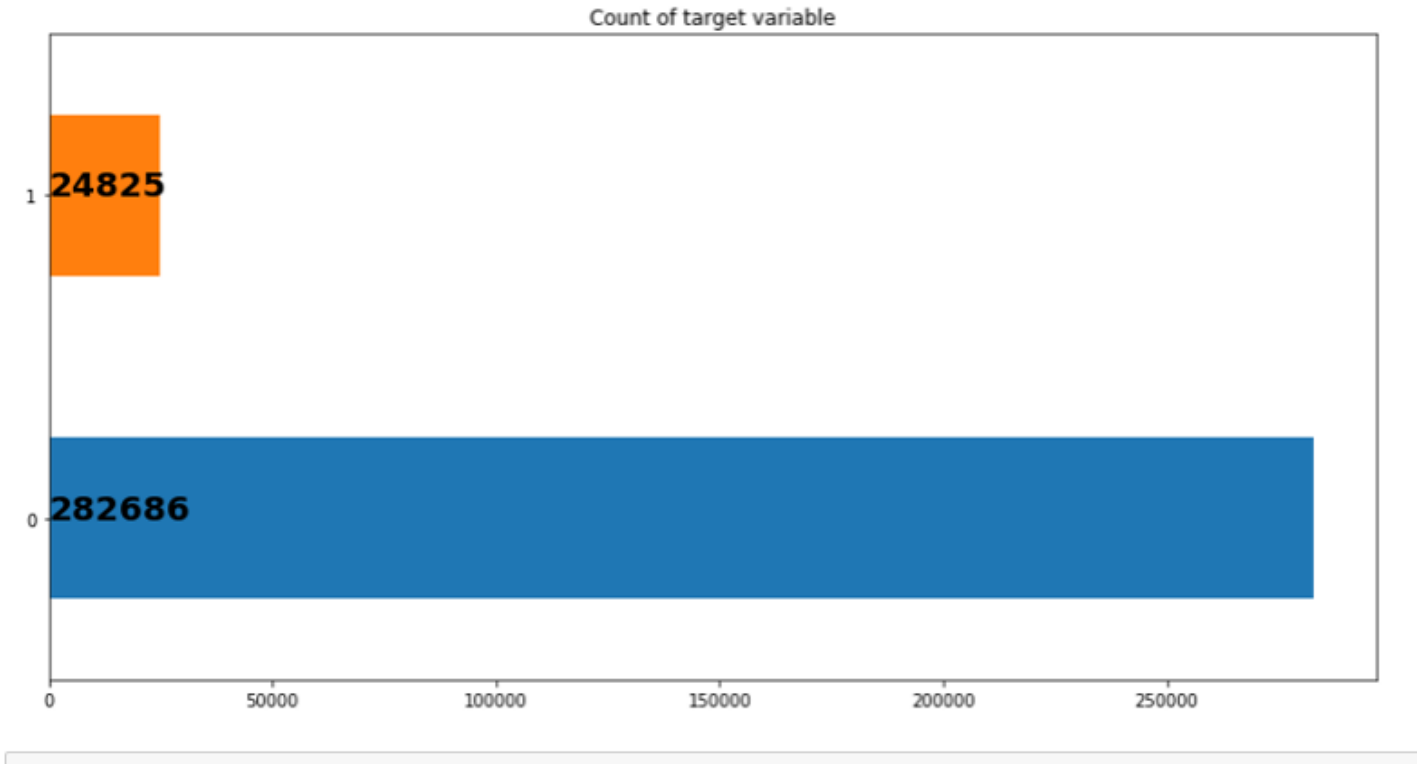
2



3

1



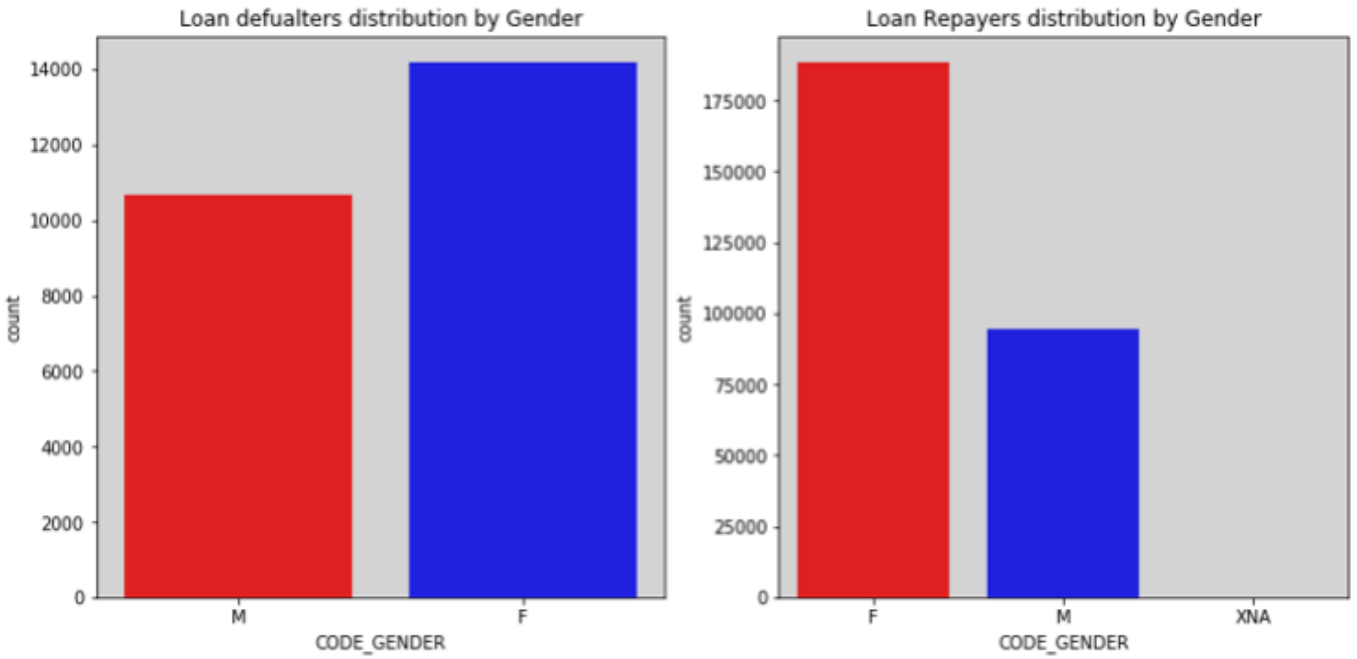


Target Variables

- The number of clients with Payment Difficulties are lesser than the ones with no payment issues
- 1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases
- Data is highly imbalanced at nearly a ration of 11.4

Univariate Analysis - Parameters

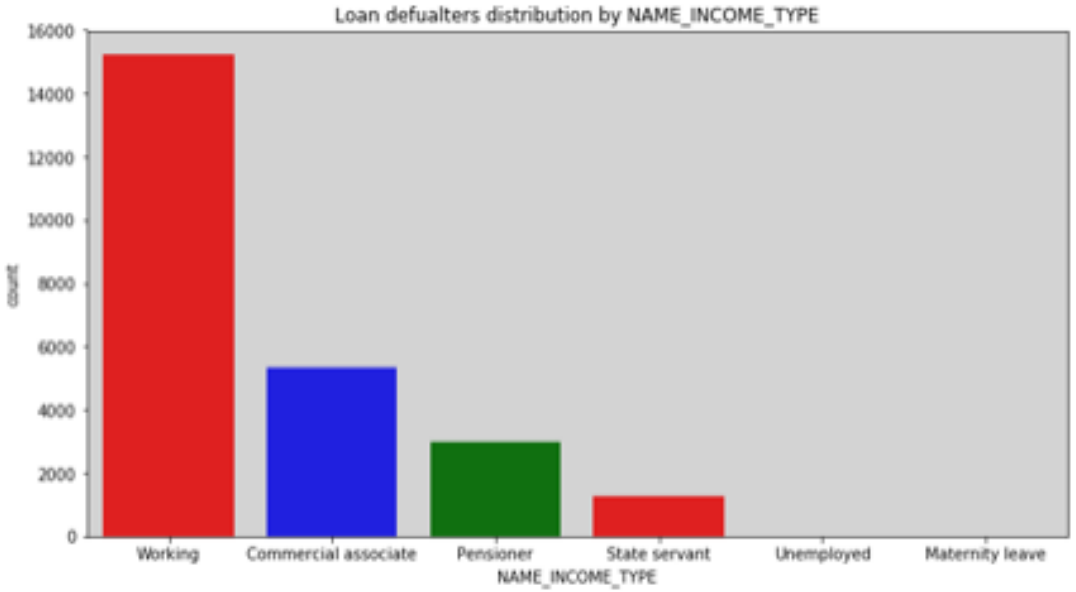
1. CODE_GENDER
2. FLAG_OWN_REALTY
3. NAME_CONTRACT_TYPE
4. NAME_EDUCATION_TYPE
5. NAME_INCOME_TYPE
6. NAME_TYPE_SUITE
7. ORGANIZATION_TYPE
8. AMT_ANNUITY
9. AMT_CREDIT
10. AMT_INCOME_TOTAL
11. CNT_CHILDREN
12. NAME_CONTRACT_TYPE
13. NAME_PAYMENT_TYPE
14. NAME_CLIENT_TYPE
15. NAME_PRODUCT_TYPE

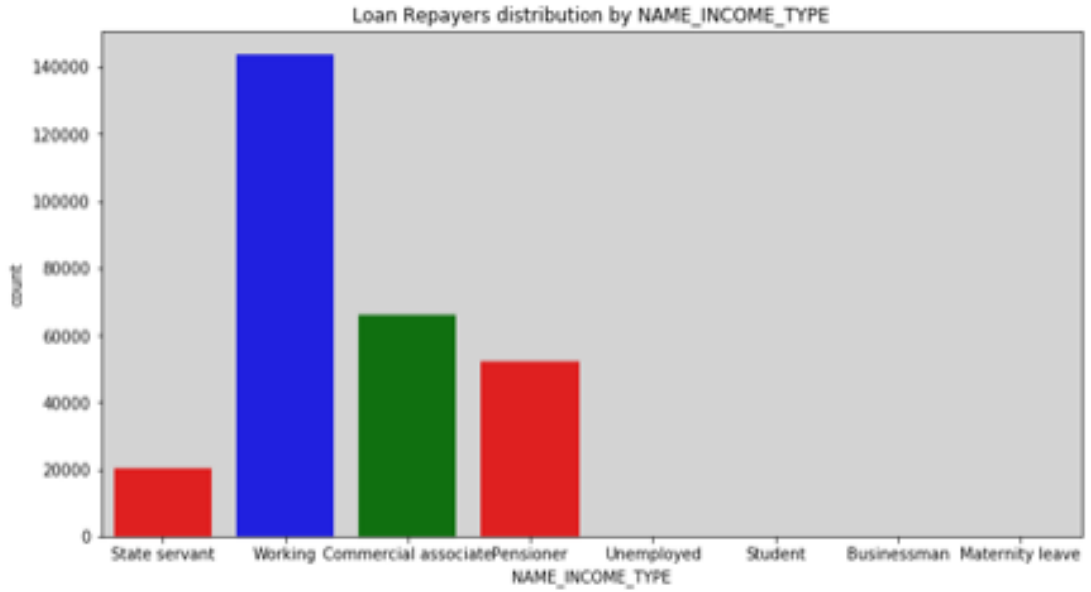


Univariate Analysis : Gender Distribution

- Among the Loan Defaulters, Females outnumber Males
- Even in case of Loan Repayment there are more females than males.
- So loan is disbursed more to females than males

Bivariate Analysis : Income Type Distribution Vs Loan Default & Repayment



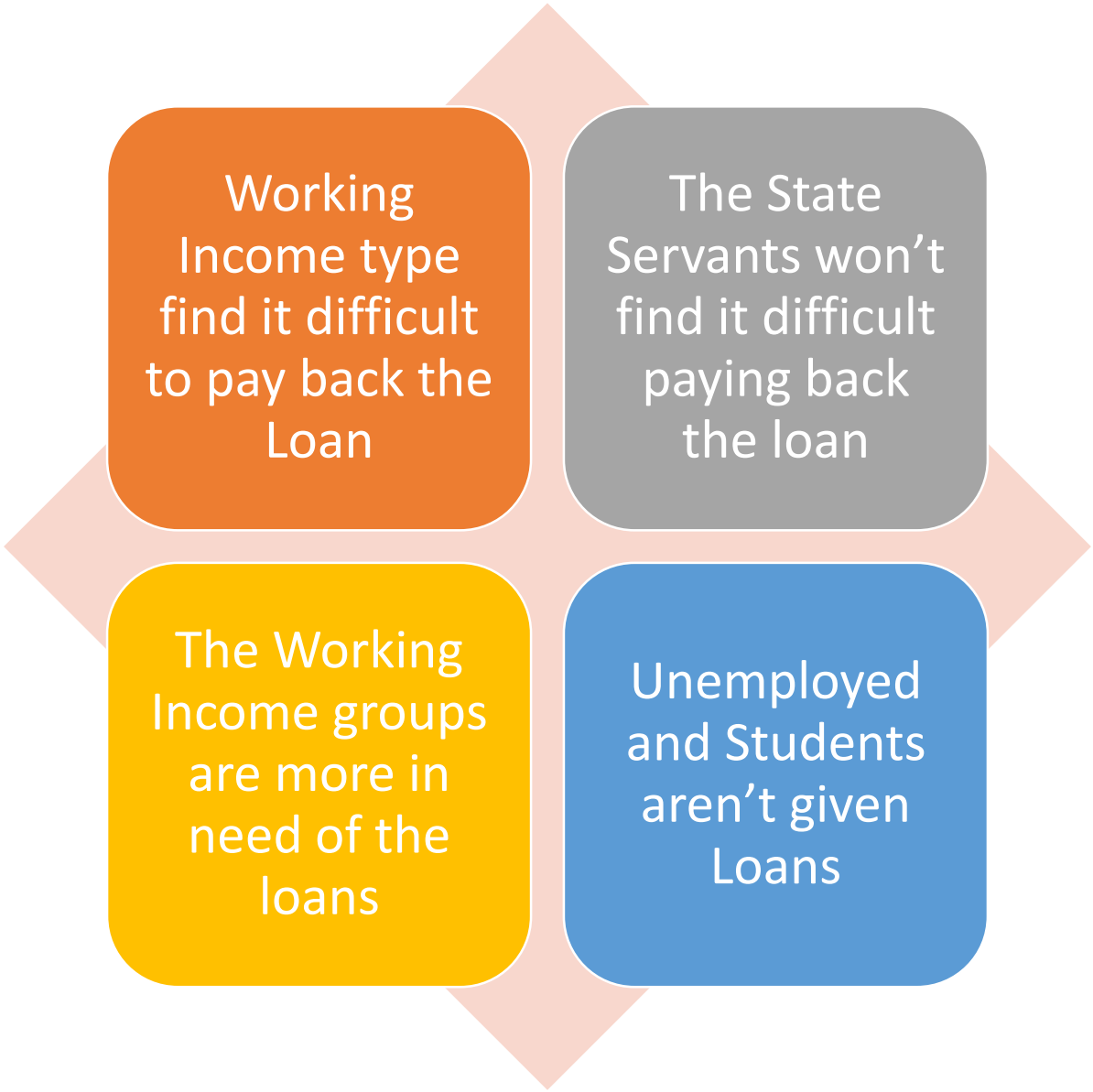


Bivariate Analysis : Income Type Distribution Vs Loan

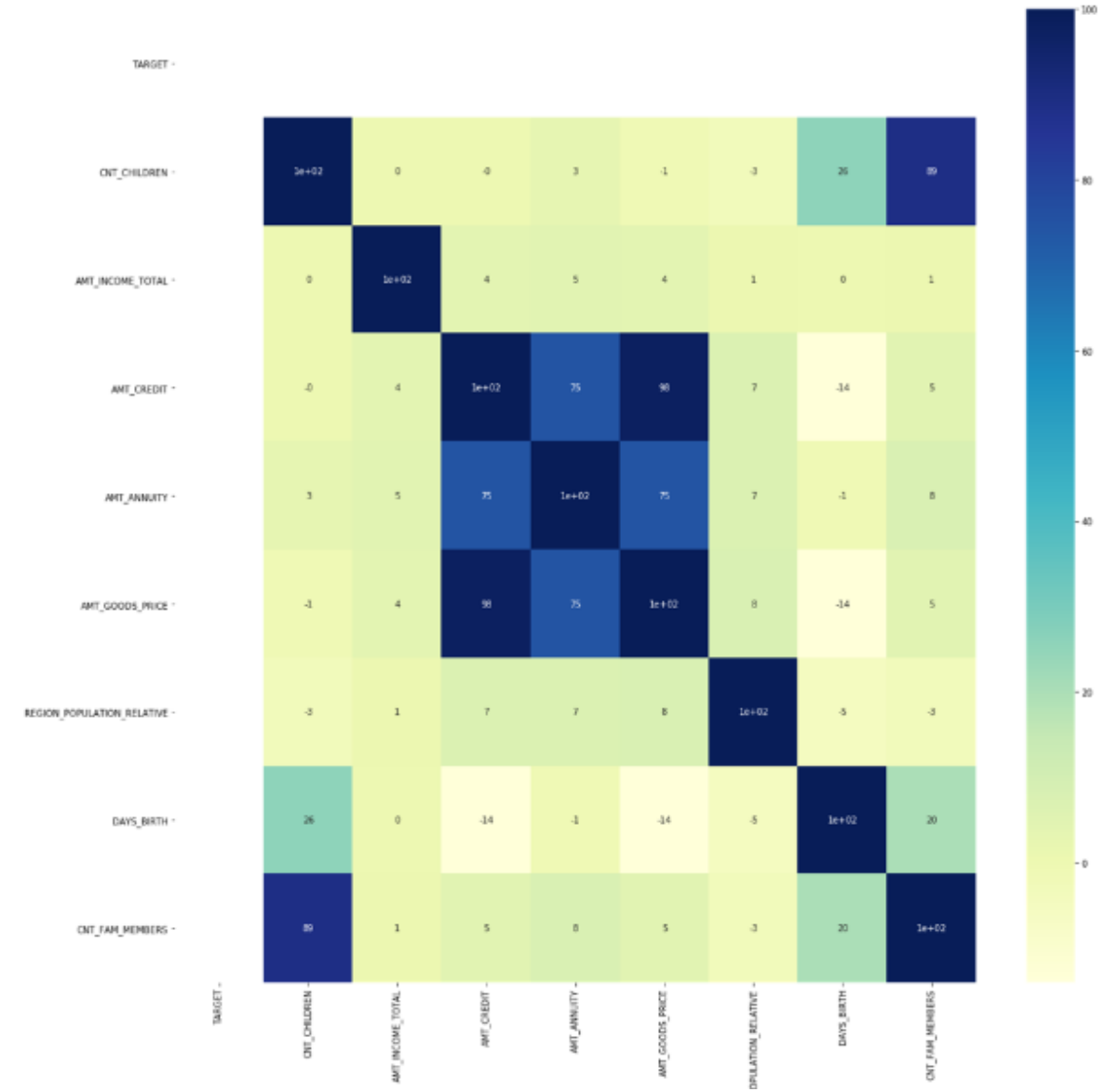
Bivariate Analysis : Income Type Distribution Vs Loan

Default & Repayment

Default & Repayment



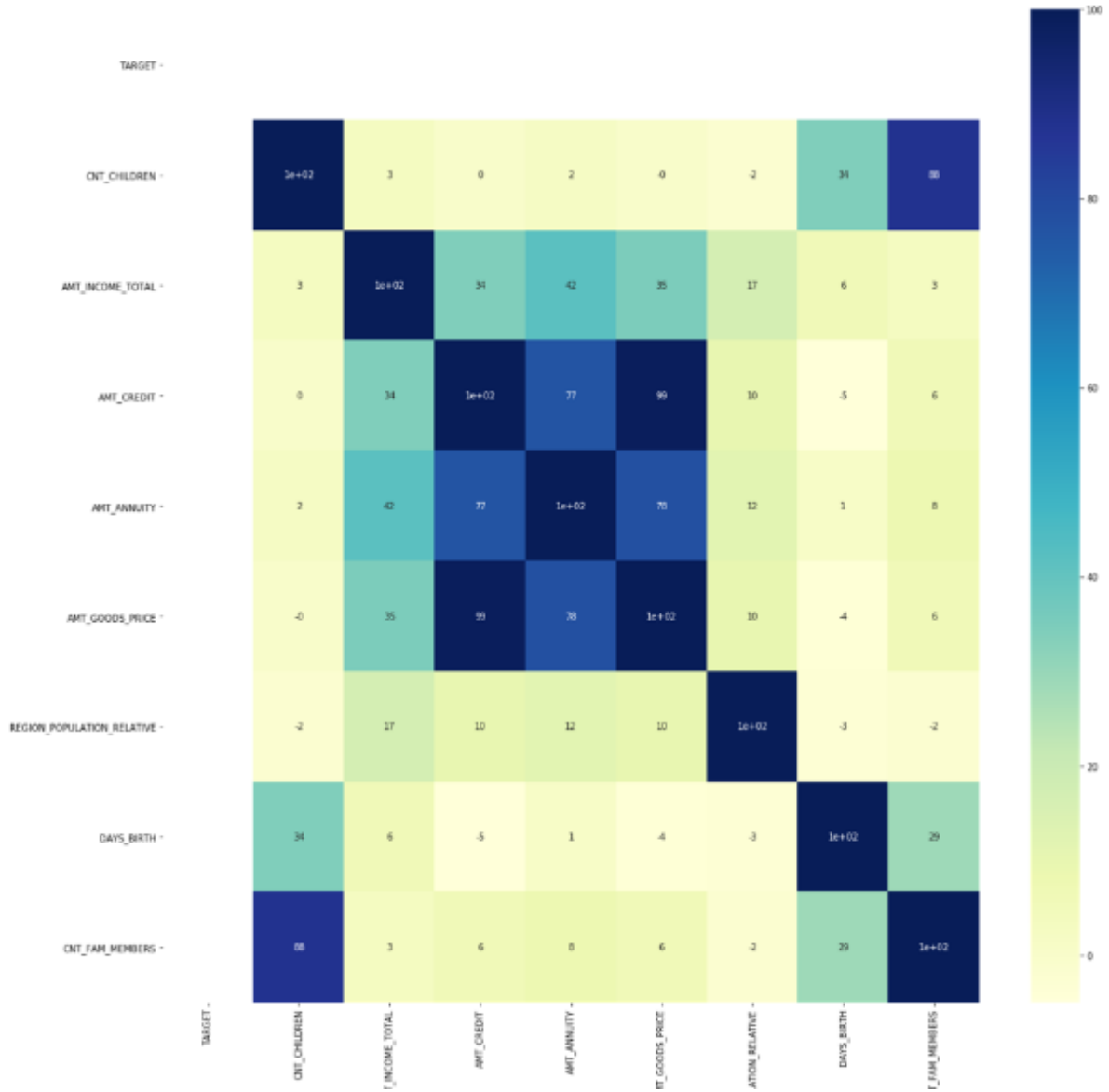
Correlation of Numerical Parameters for Target Variable = 1



Correlation is highest for:

- AMT_CREDIT
- AMT_GOODS_PRICE

Correlation of Numerical Parameters for Target Variable = 0

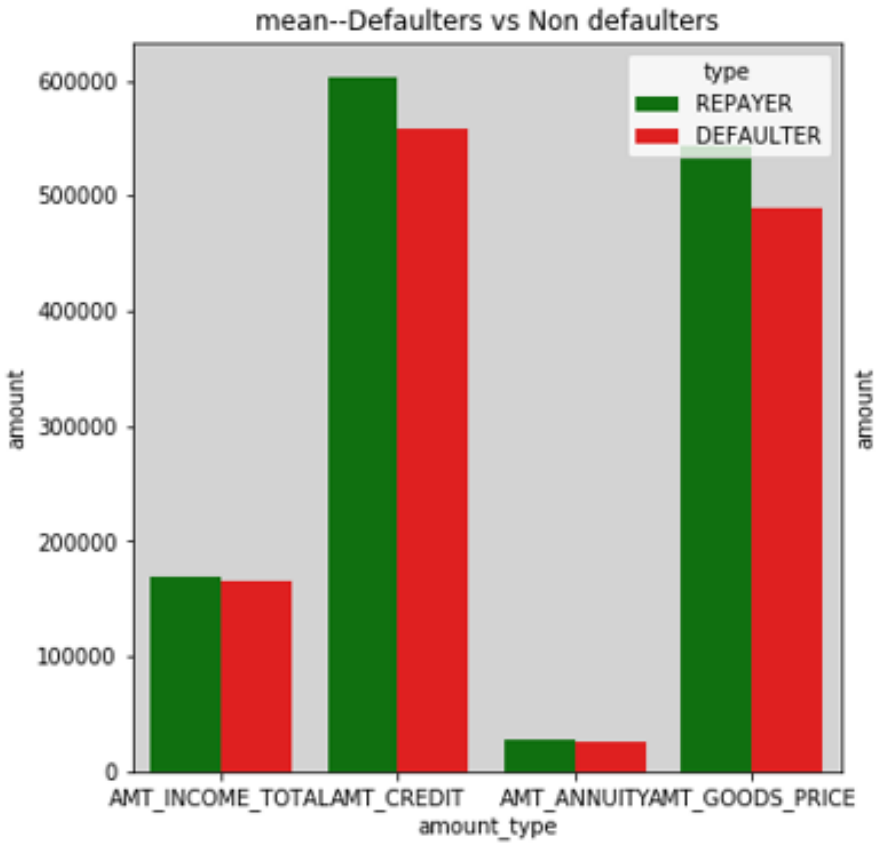


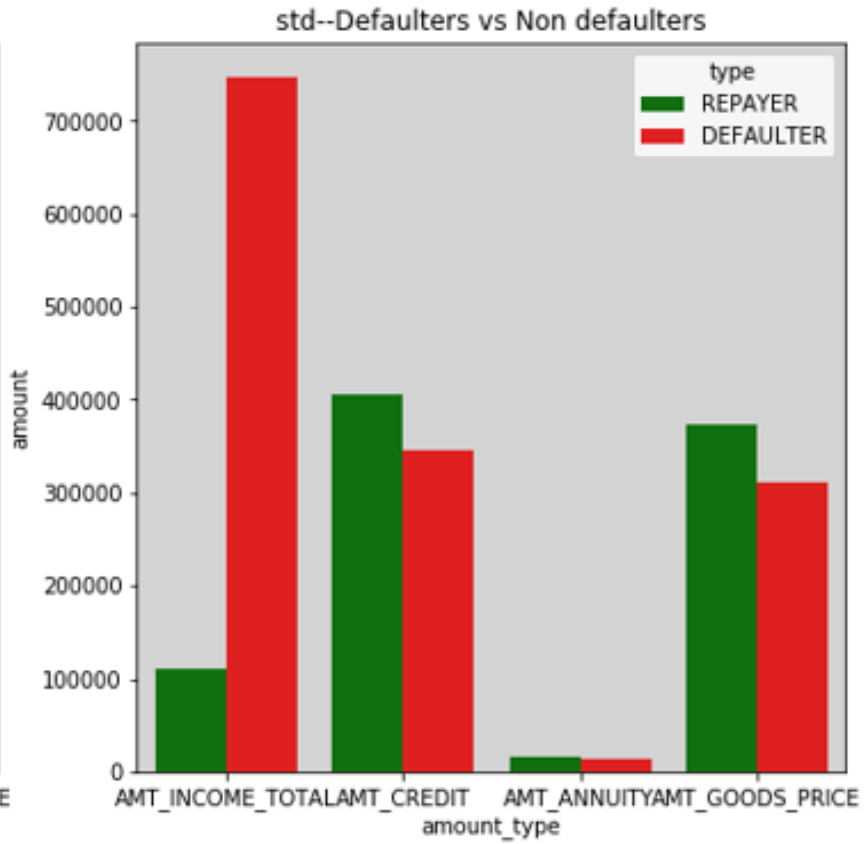
Correlation is highest for:

- AMT_GOODS_PRICE

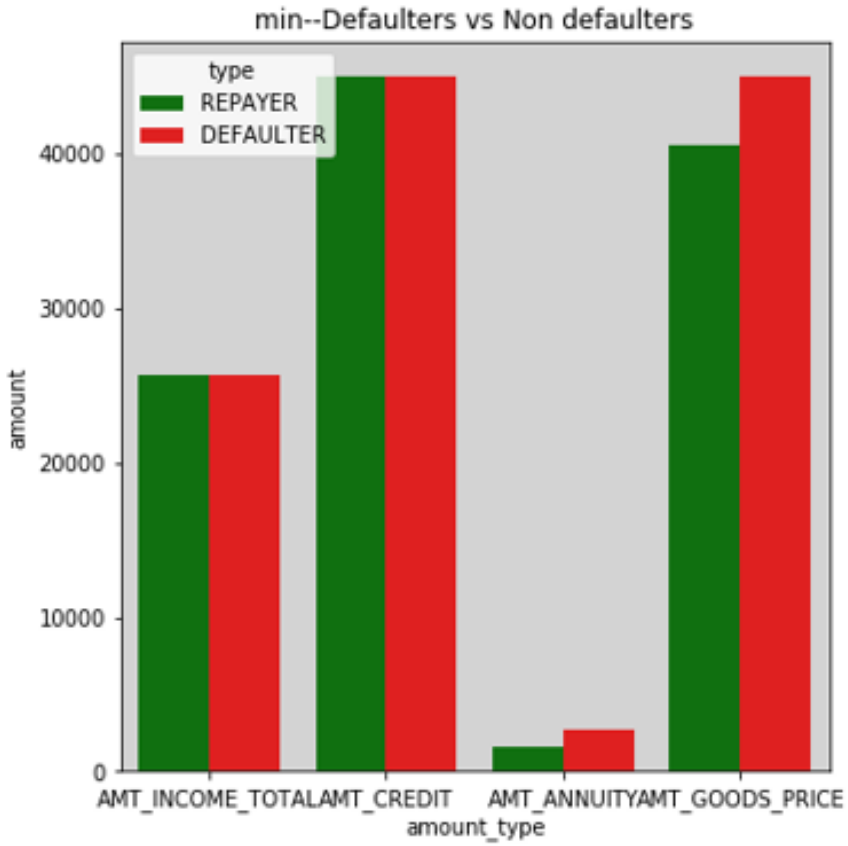
- Older people in new jobs are more likely to default

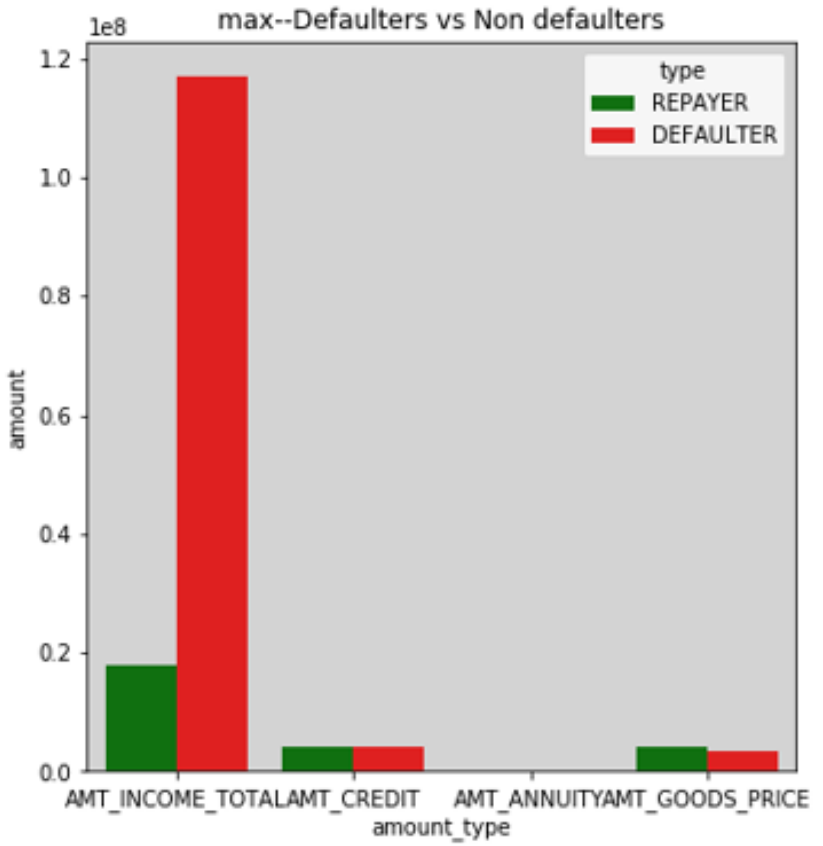
Univariate Segmented Analysis : Defaulters Vs Non-Defaulters





Univariate Segmented Analysis : Defaulters Vs Non-Defaulters





Bivariate Analysis

AMT_INCOME_TOTAL

AMT_CREDIT

'AMT_ANNUITY

AMT_GOODS_PRICE

DAYS_DECISION

HOUR_APPR_PROCESS_START

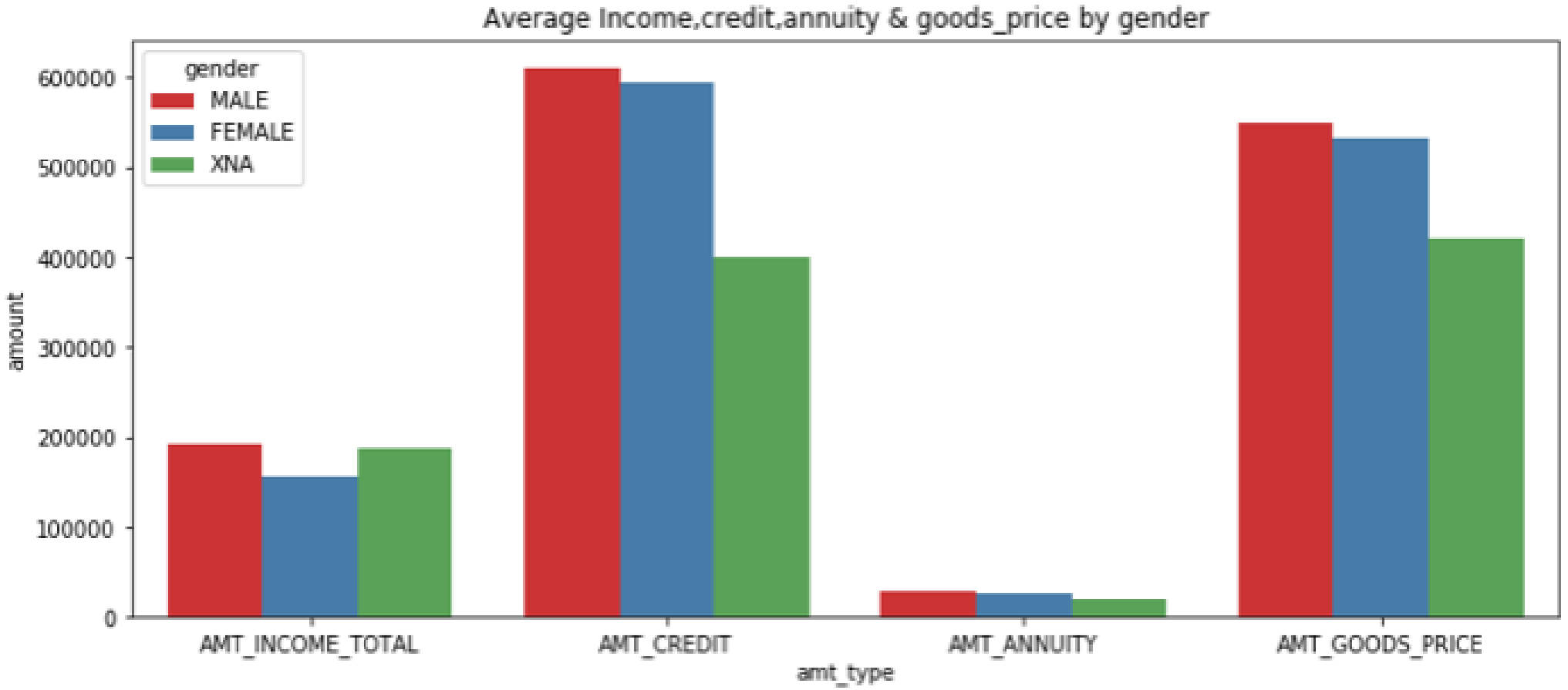
NAME_CONTRACT_TYPE

CODE_GENDER

FLAG_OWN_CAR

FLAG_OWN_REALTY

Bivariate Analysis : Gender Vs Average Income, Credit, Annuity and Goods Price



Bivariate Analysis : Gender Vs Average Income, Credit, Annuity and Goods Price

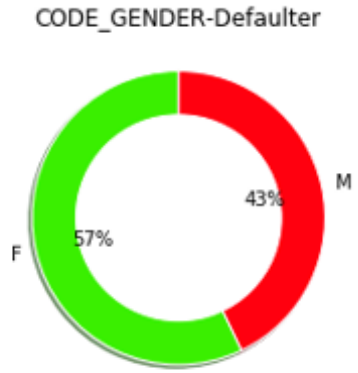
The Average Income of Males is larger than that of Females

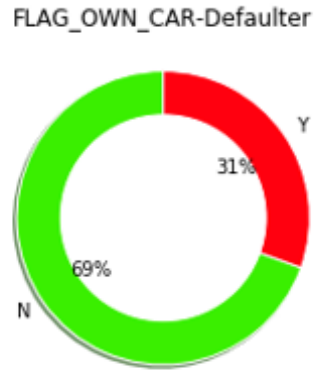
The Credit Amount received by Males is larger than Females

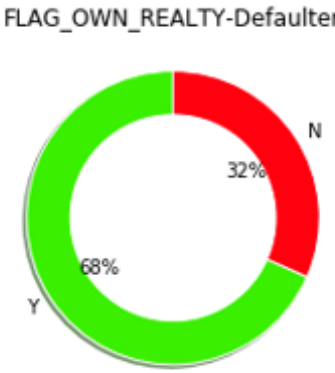
When it comes to Annual Annuity there isn't significant difference between both genders

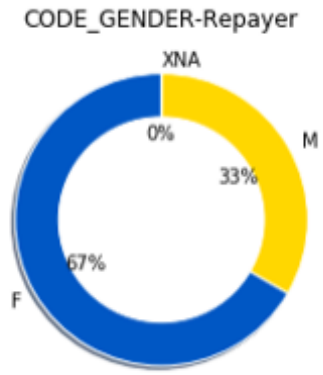
The Price of Goods for which the loan is availed is higher for Males than that of females

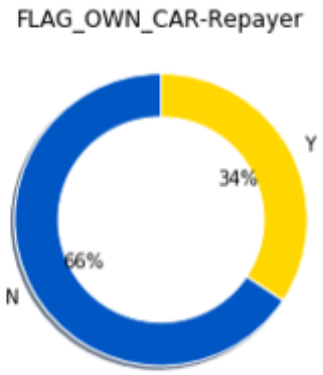
Bivariate Analysis : Distribution of contract type ,gender ,own car ,own house with respect to Repayment status

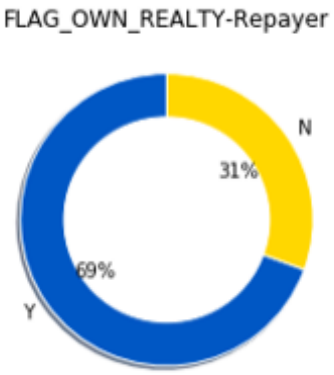






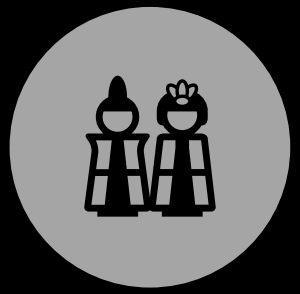






Bivariate Analysis : Distribution of contract type ,gender ,own car ,own house with respect to Repayment status

Cash Loans are more in Subscription than the Revolving Loans



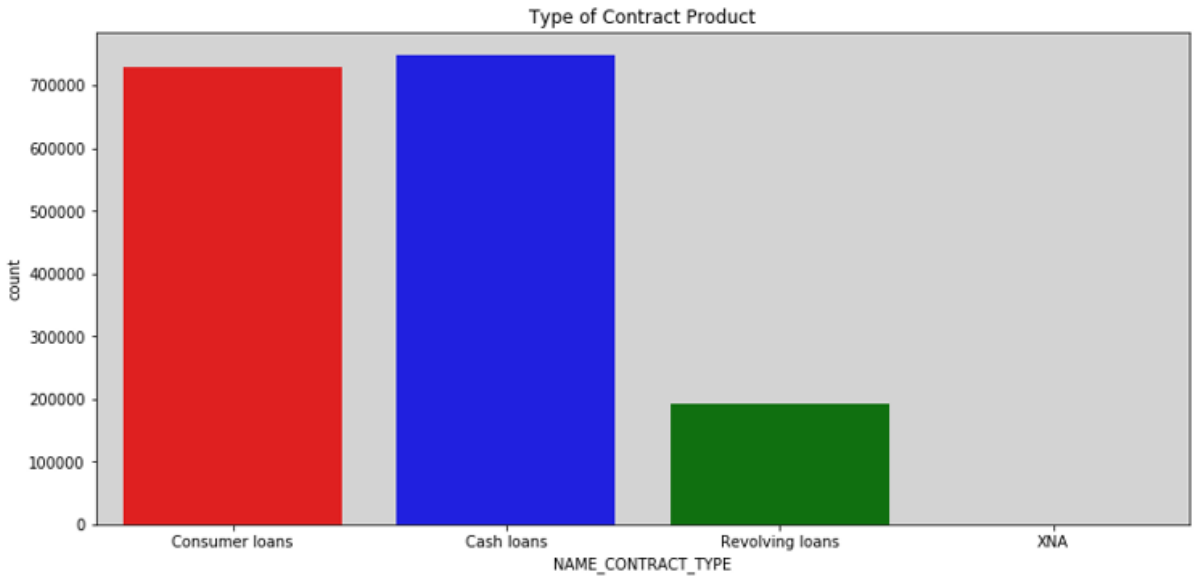
Females are the recipients of most loans

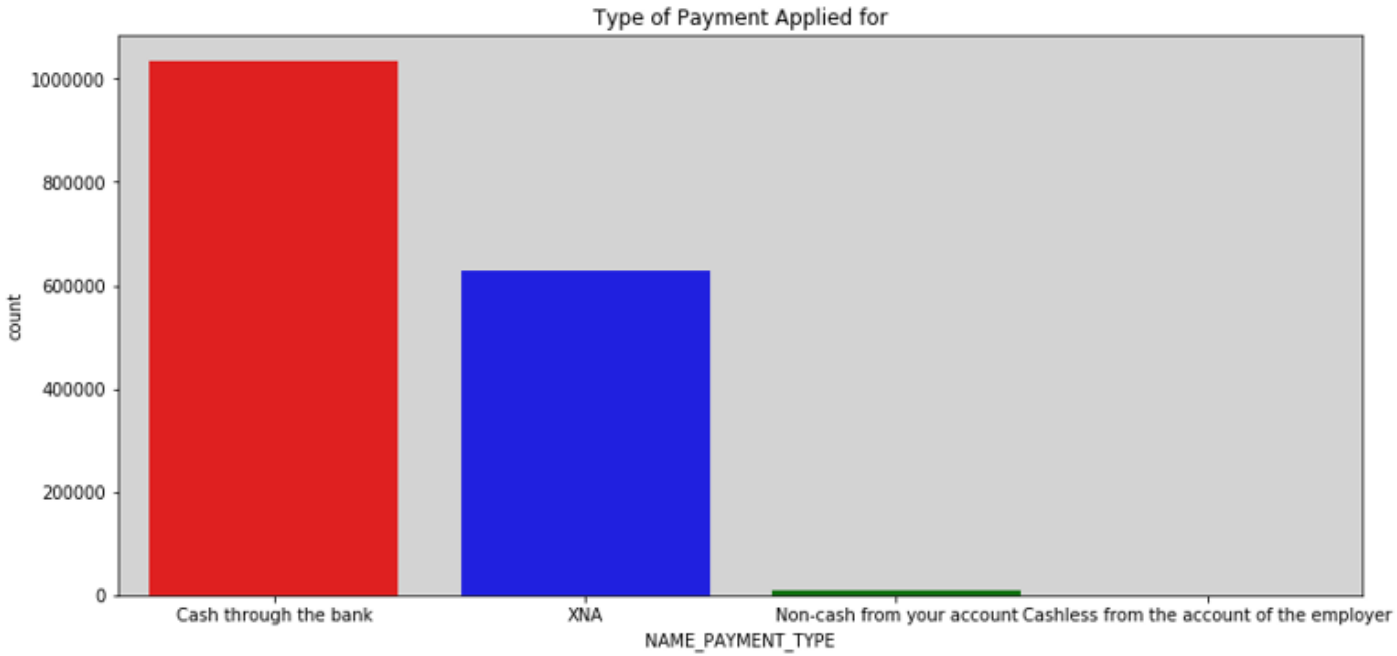
Car Owners don’t default on loan lesser than Non car owners

Owning a flat/ Home doesn’t affect the status of defaulting or repayment of the loan

Univariate Analysis : Contract Product Distribution based on Previous Applications

- The Loan Product Type in highest subscription is the Cash Loan followed by the Consumer Loans



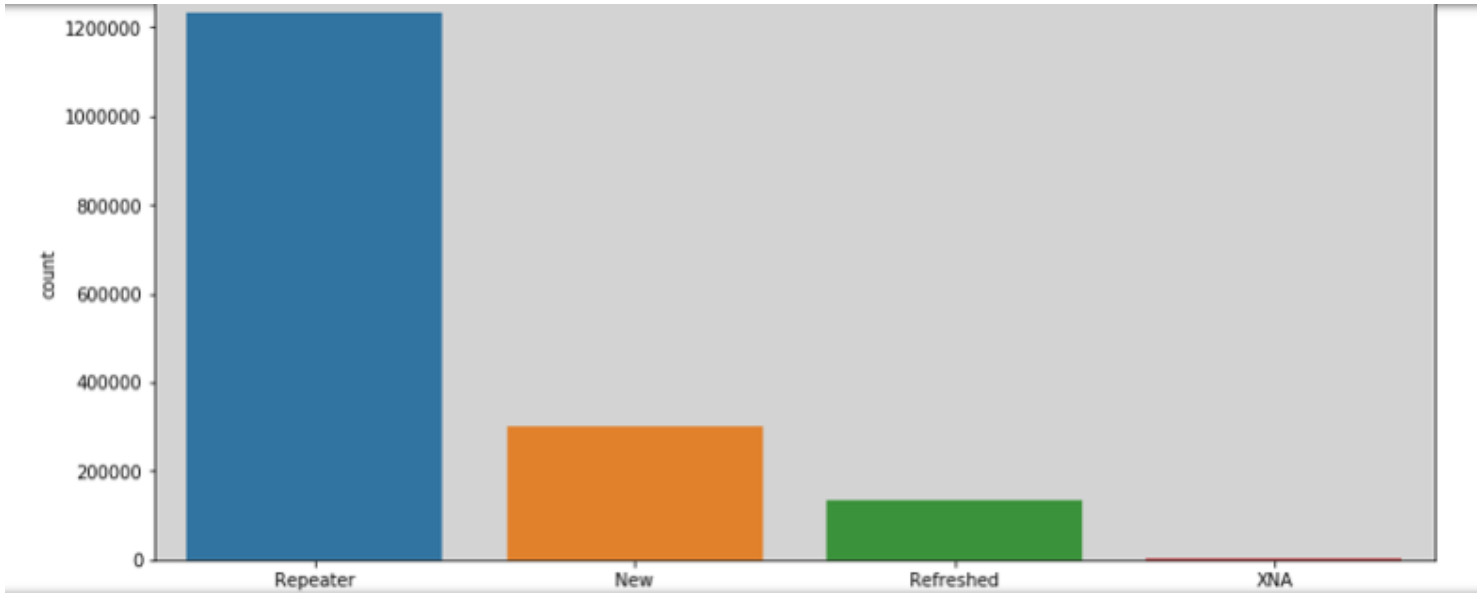


Univariate Analysis : Types of Payment applied for based on Previous Applications

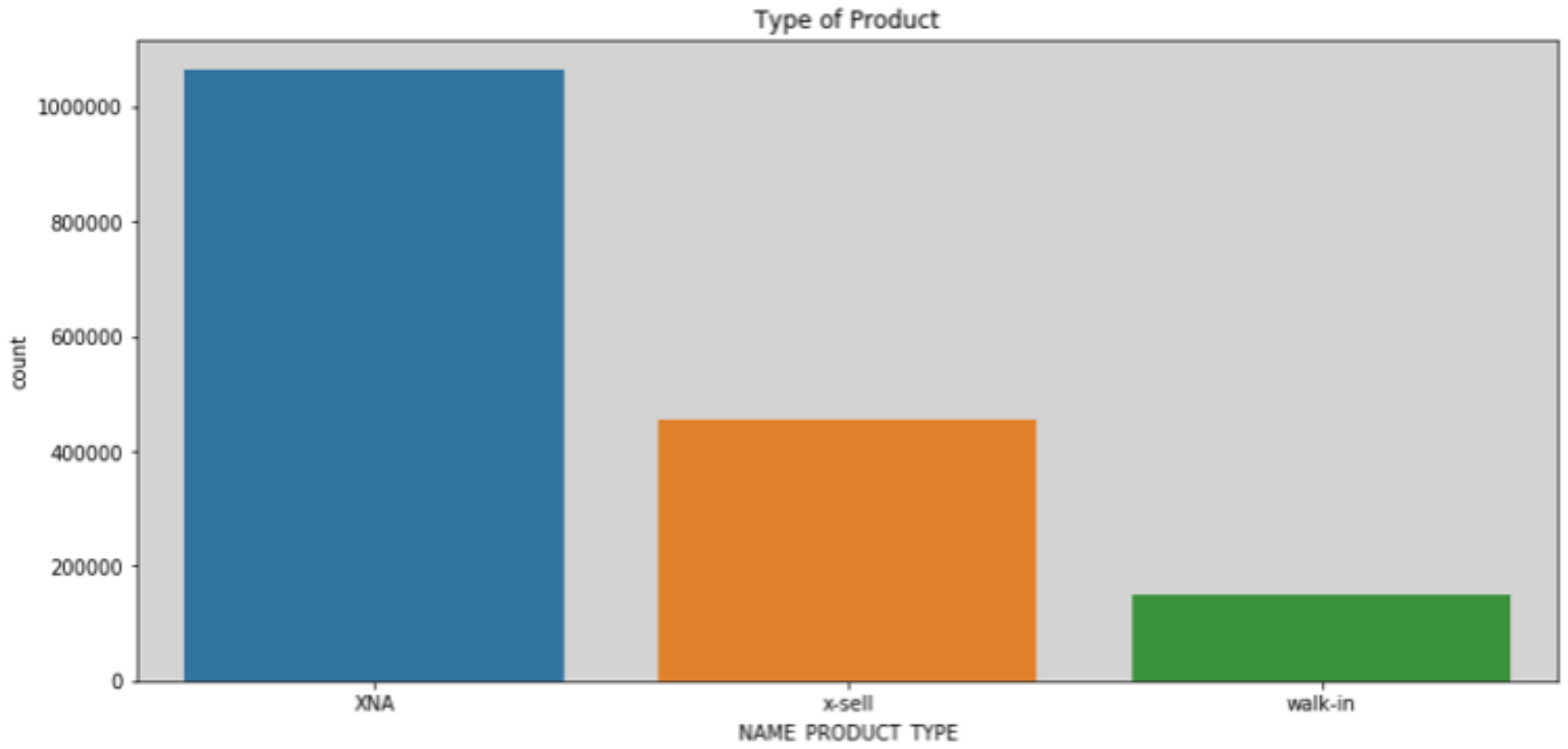
- Most of the Loans subscribed are cash loans.
- Non-Cash/ Cashless Payments have less penetration

Univariate Analysis : Client Type based on Previous Applications

- Its previous customers who keep on applying for loans again



Univariate Analysis : Product Type based on Previous Applications



Bivariate Analysis : Process Start Hours Vs Days to Decision in Previous Applications

- Bivariate Analysis between the Point of Loan applications and the number of days taken for the decision

Final Suggestions

- There is a huge untapped market for Cashless Loans
- Even though the females are the highest recipients of the loans, the expenditure and average income is higher for males
- The higher income groups secure most of the loans
- While analyzing the Previous applications we can see that it’s the repeated customers that mostly apply for loans
- Older people in new jobs are more likely to default