Credit EDA Assignment

Aditi Deshpande

Problem Statement

A bank receives various types of loan applications. Before taking any decision on application company has to verify applicant's profile.

Risk associated with bank's decision:

- If applicant is likely to repay the loan, then rejecting his application is loss of business

- If the applicant is not able to repay loan then person is likely to be default, then approving loan for such person is loss of business.
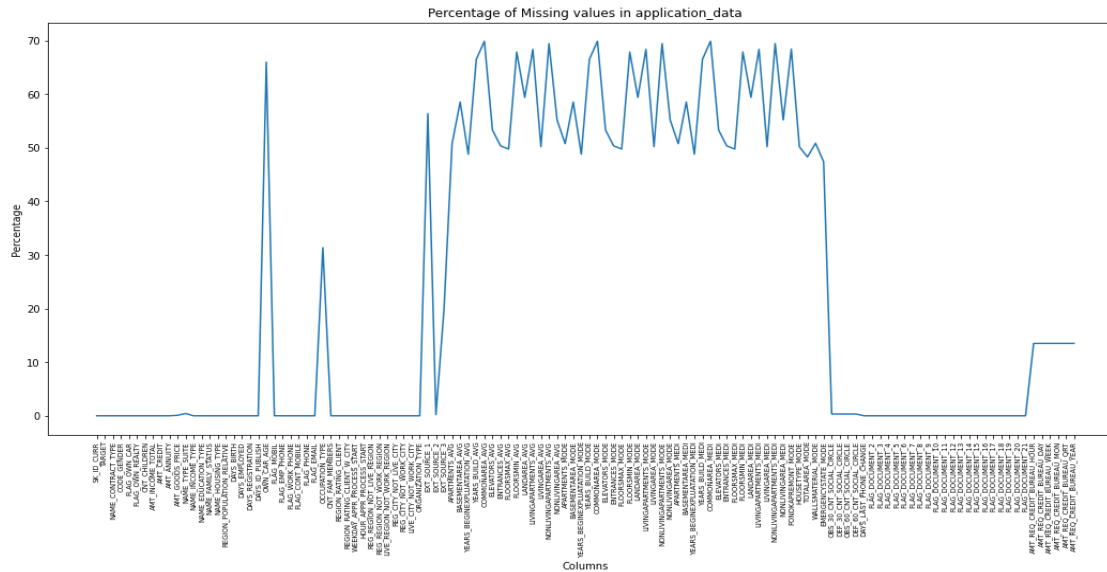
The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.
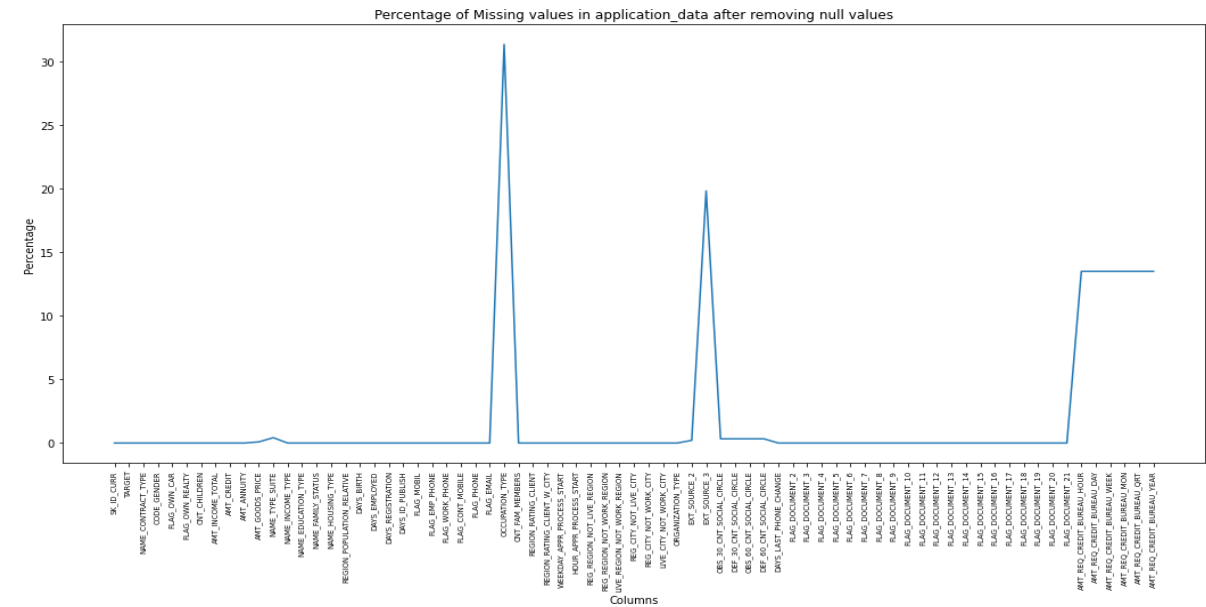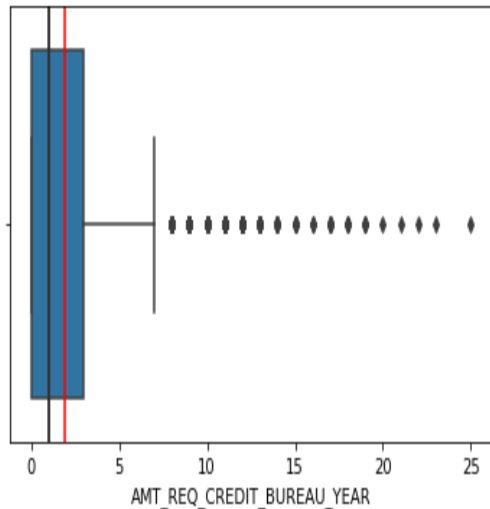
Analysis Approach
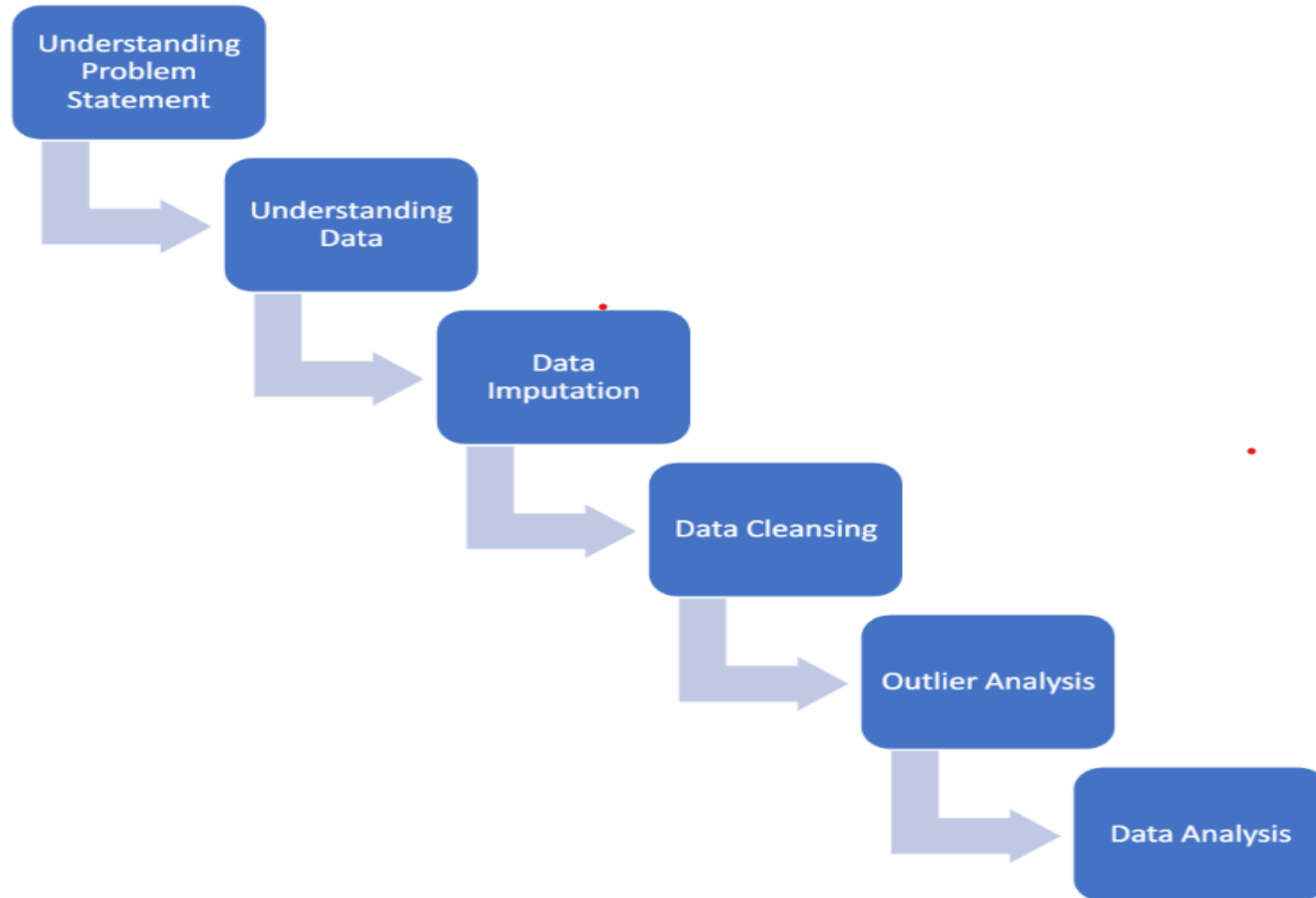
Missing Value Analysis

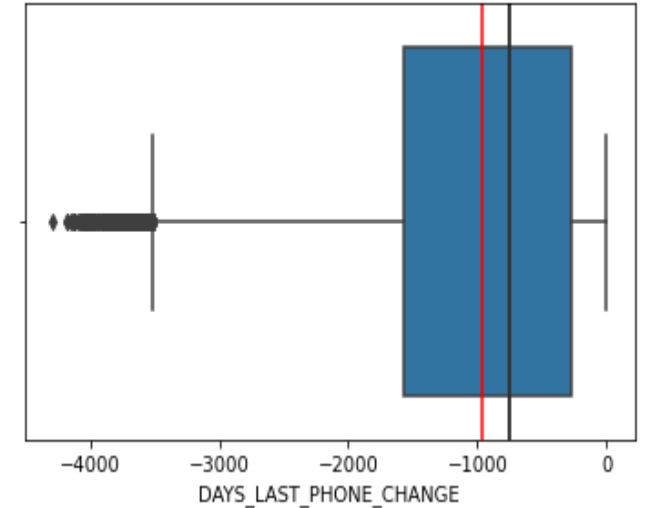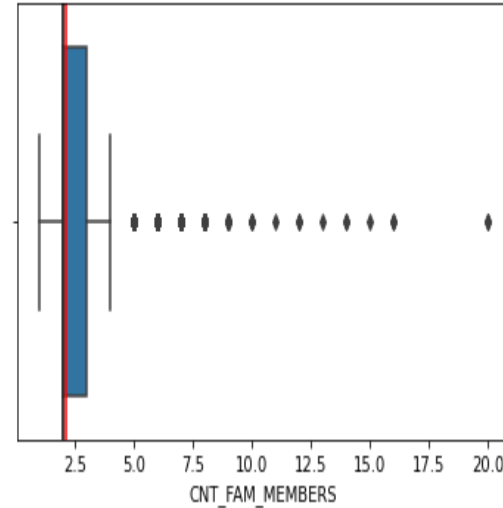Percentage of Missing values in application_data

Around 49 Columns contains null values greater than 40% so dropped these columns.

After Dropping columns having null values greater than 40%

Percentage of Missing values in application_data after removing null values

Outlier Analysis
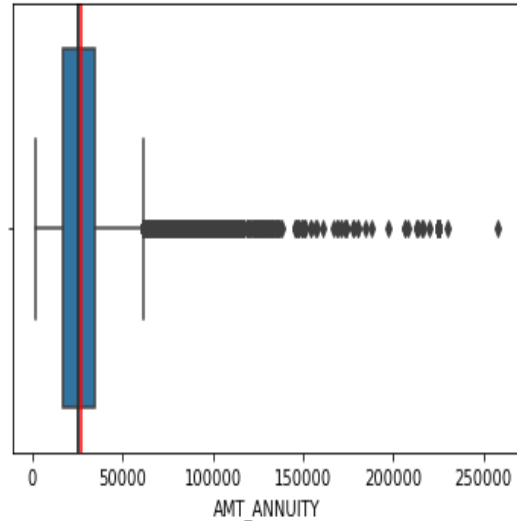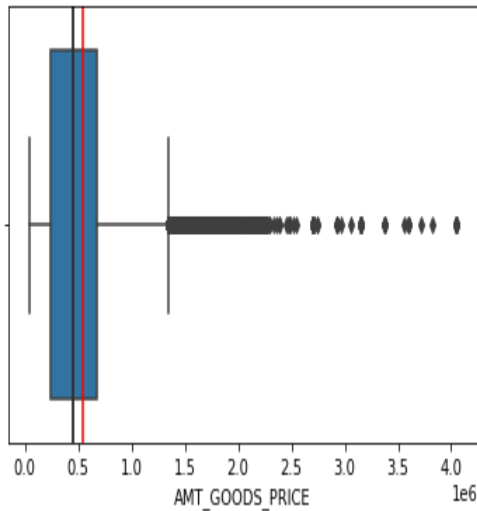
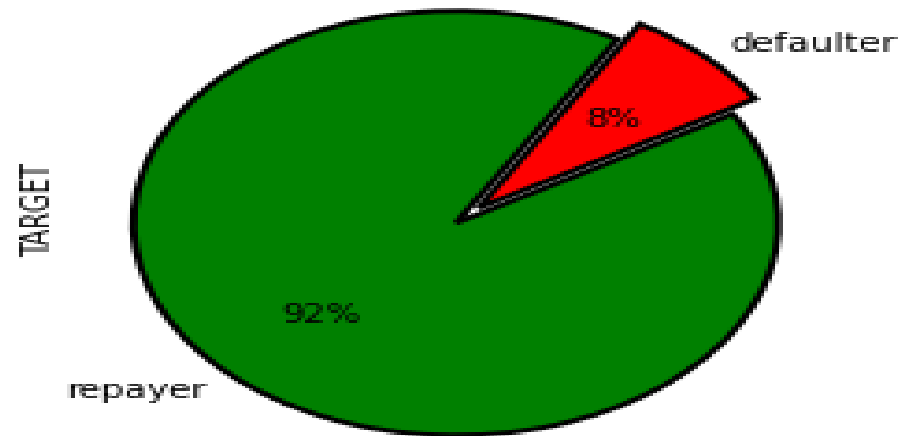**Observations:**
- AMT_GOODS_PRICE, AMT_ANNUITY, CNT_FAM_MEMBERS, AMT_REQ_CREDIT_BUREAU_YEAR and DAYS_LAST_PHONE_CHANGE have some number of outliers
- As these columns contain outliers replaced it with median().

Data Imbalance

Target Variable Data Imbalance
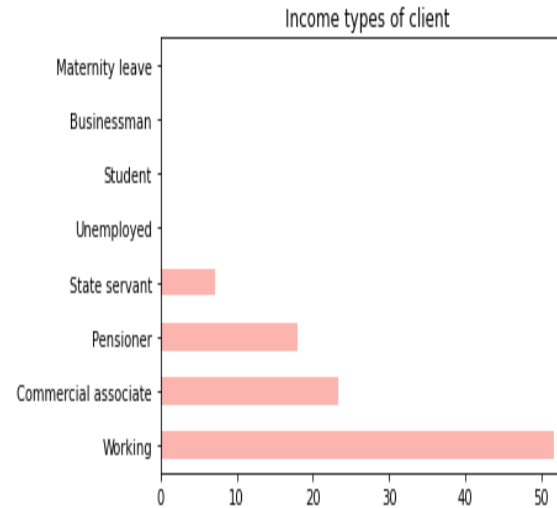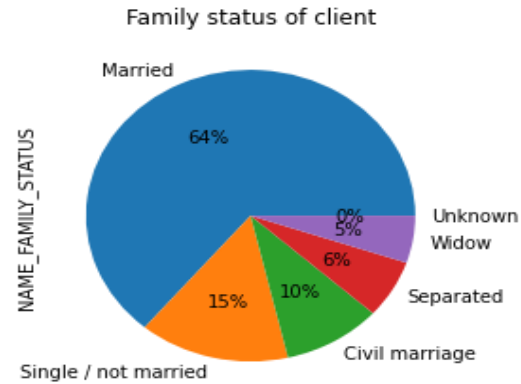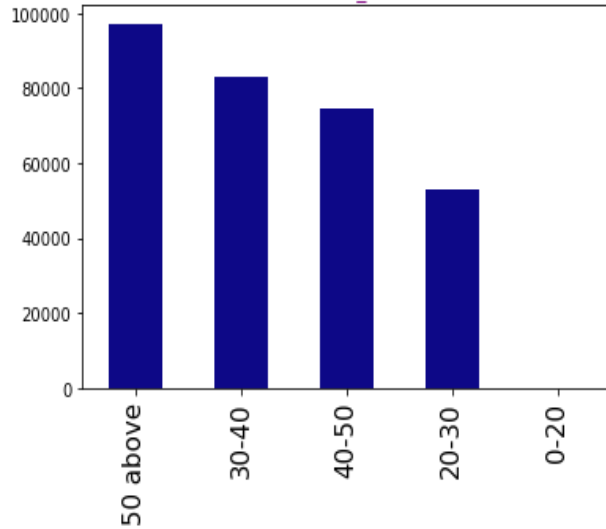

Distribution of Target Column

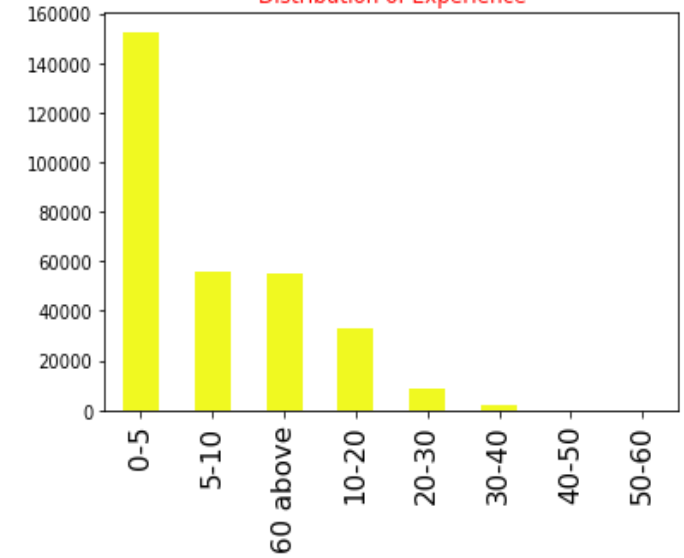**Ratio of imbalance for Re-payer and Defaulter in percentage: 92 & 8**

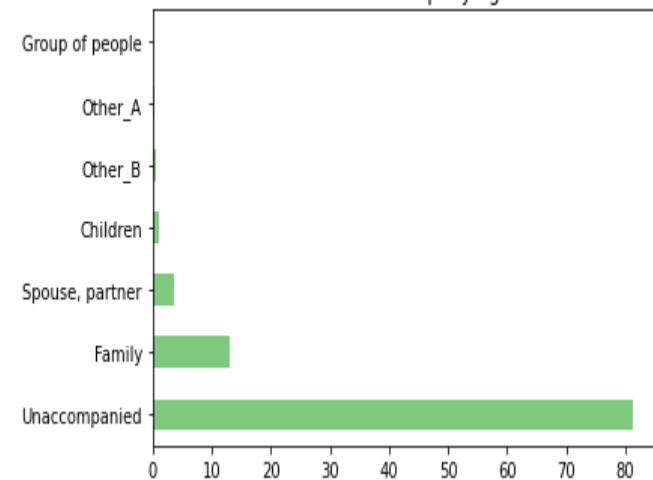**Ratio of imbalance of Re-payer Vs Defaulter: 11.5**

Univariate Analysis

Distribution of AGE_GROUP Variable
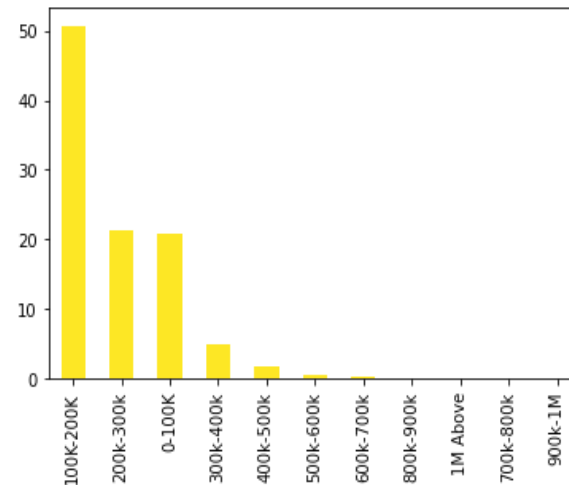

Family status of client

Income types of client

Distribution of Experience

Client accompanying

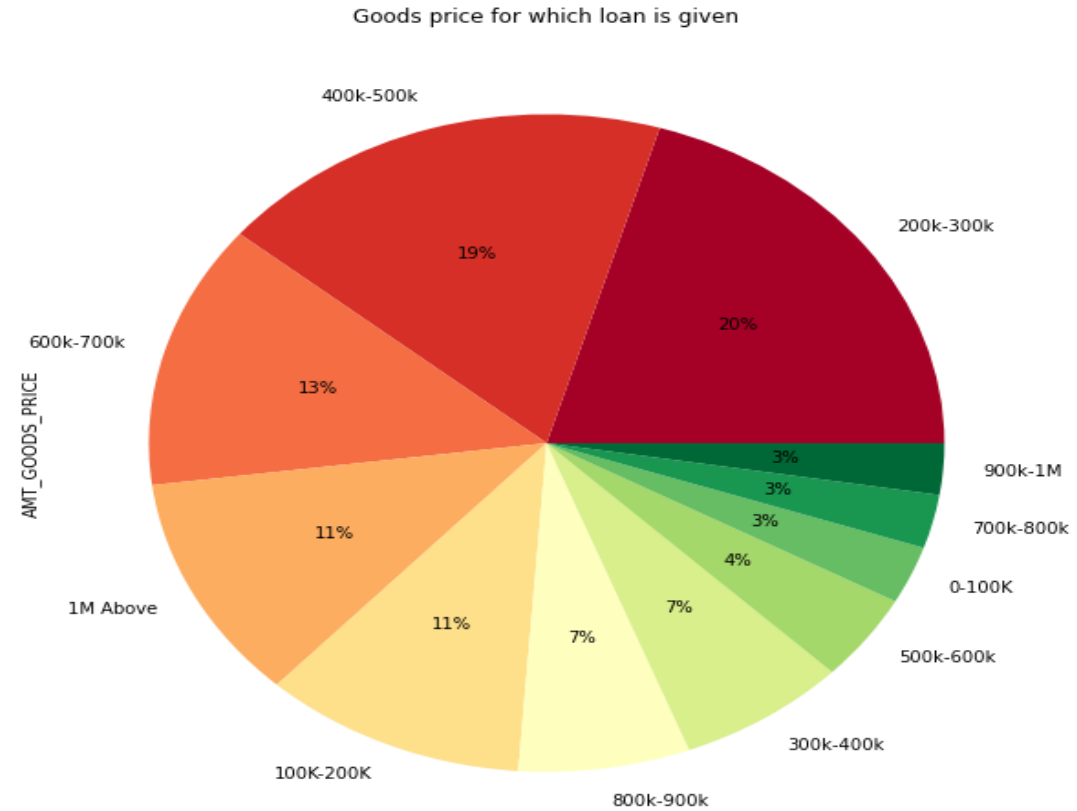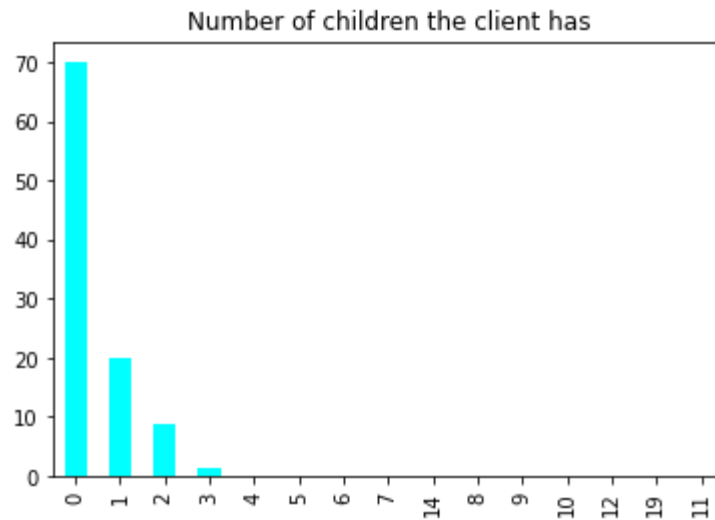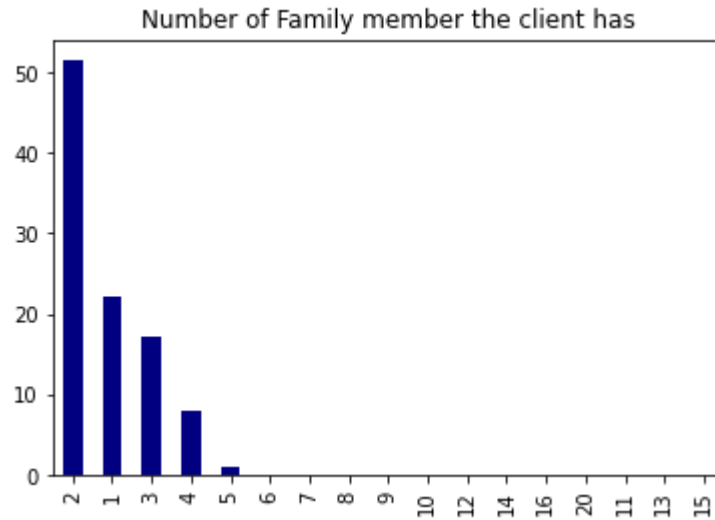Income of client

1) In Age distribution: 50 and Above people took more loans
2) Married people took more loan as compare to others
3) Working people has took most of the loans.
4) 0-5 years of experience people took loan as compare to others
5) Most of client are unaccompanied in their loan
6) 50% of client has income between 100k-200k

Univariate Analysis

Number of Family member the client has

Number of children the client has

Goods price for which loan is given
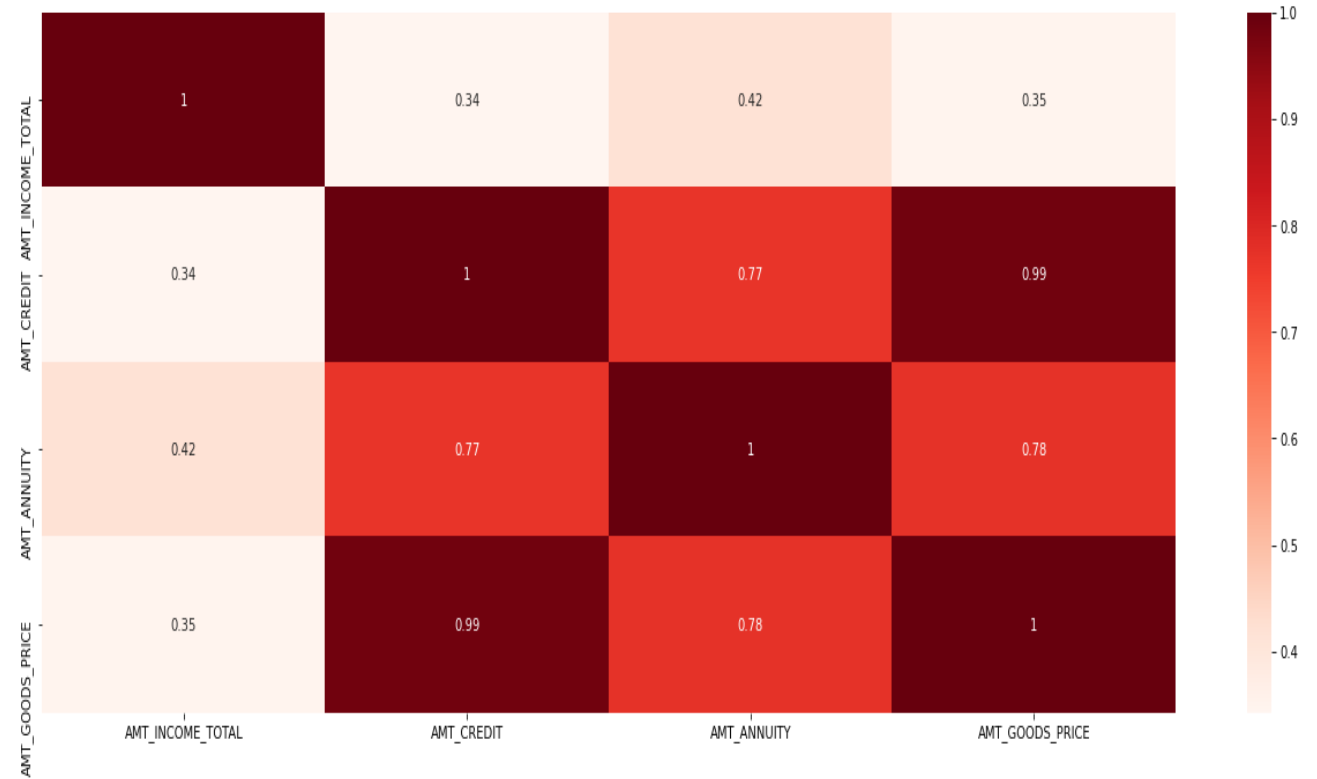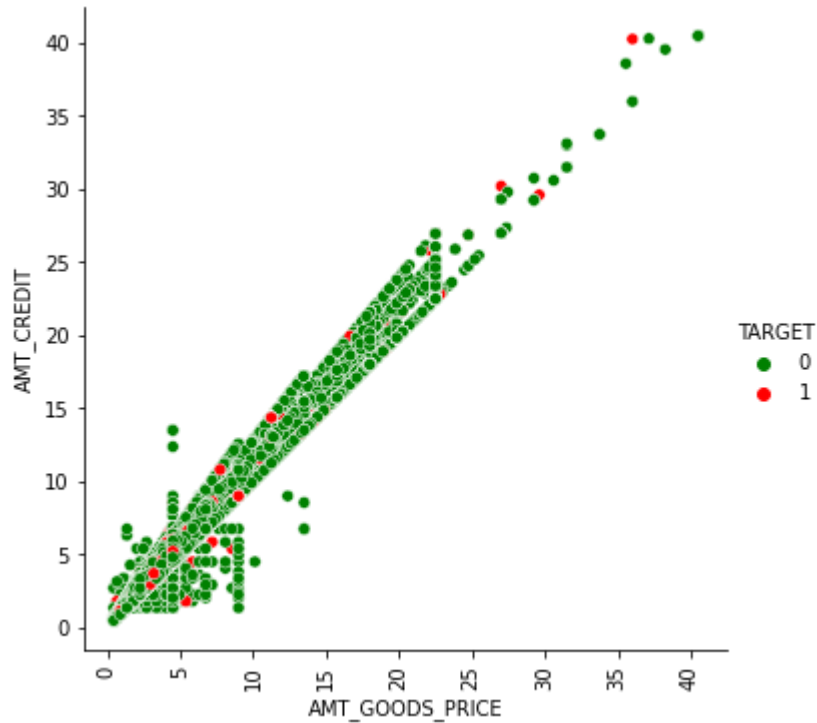

1) 50% client has 2 family members.
2) Around 70% client don't have a child
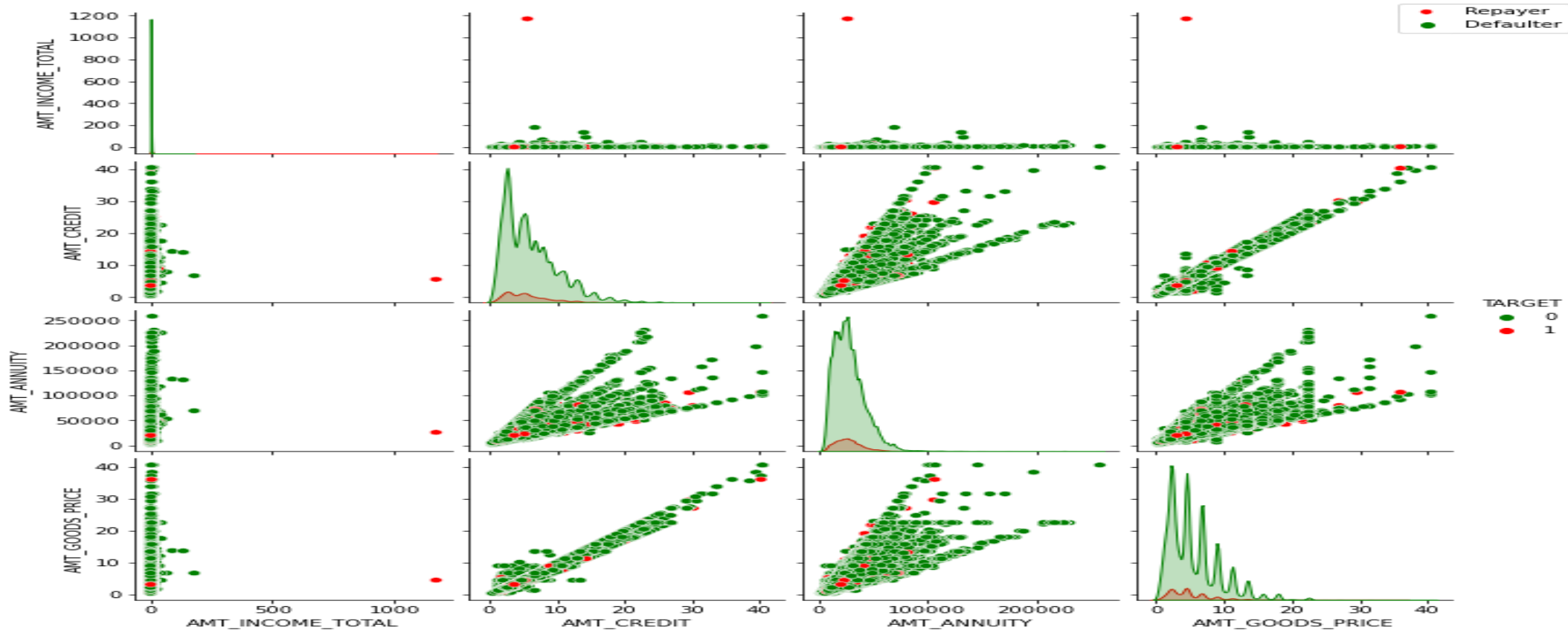3) Most of the clients took loan to buy goods in range of 200k-300k

# Bivariate and Multivariate Analysis

1) When the credit amount goes above 2M, there is an increase in defaulters.
2) Very high correlation between AMT_CREDIT and AMT_GOODS_PRICE - Applicants owning goods of high value can take loans of higher amounts

Bivariate and Multivariate Analysis

1) AMT_CREDIT and AMT_GOODS_PRICE are highly correlated as shown in scatterplot most of the data are consolidated in form of line
2) There are very less defaulters for AMT_CREDIT < 2M
3) AS AMT_ANNUITY is increasing number of defaulter is also increasing
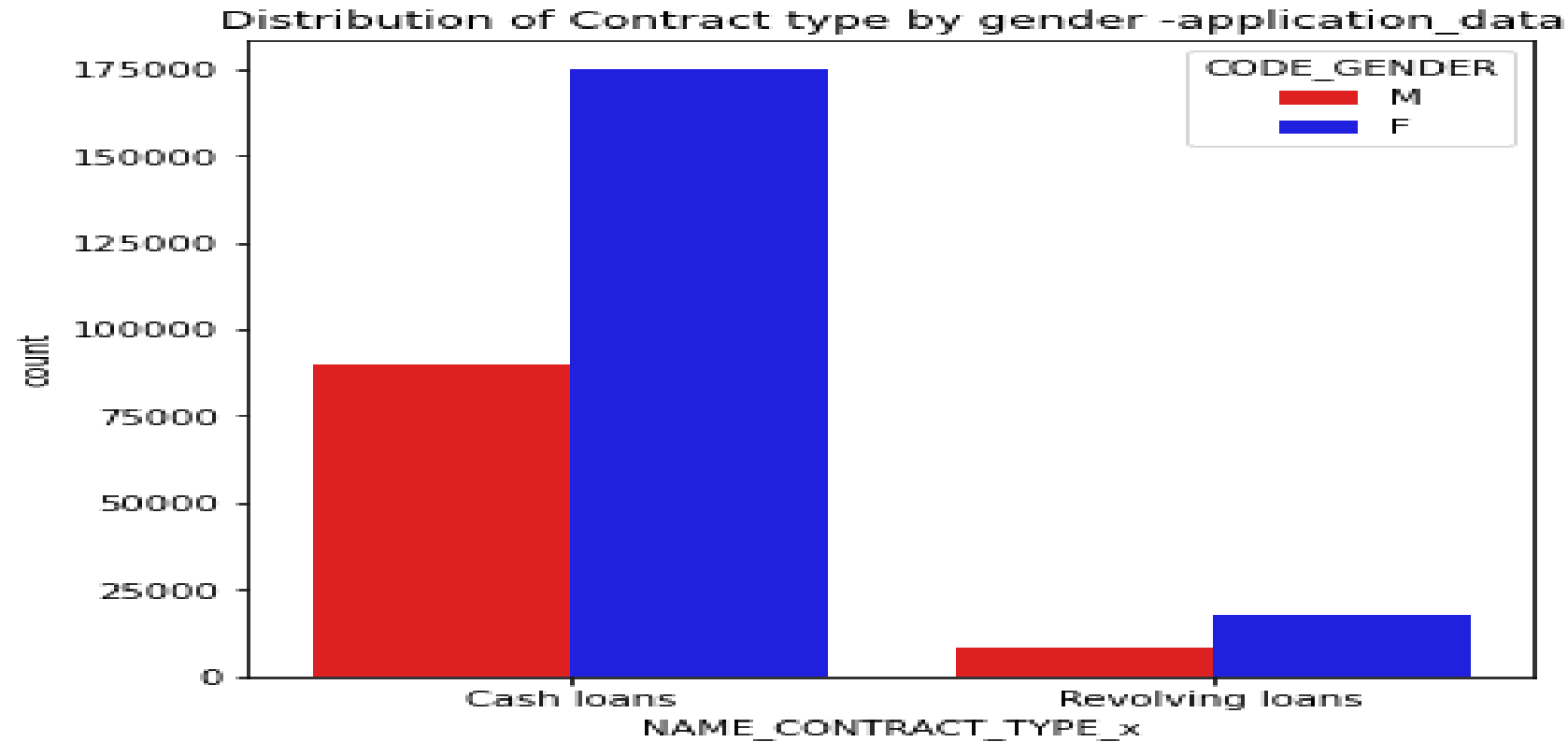
Multivariate Analysis(Re-payer)
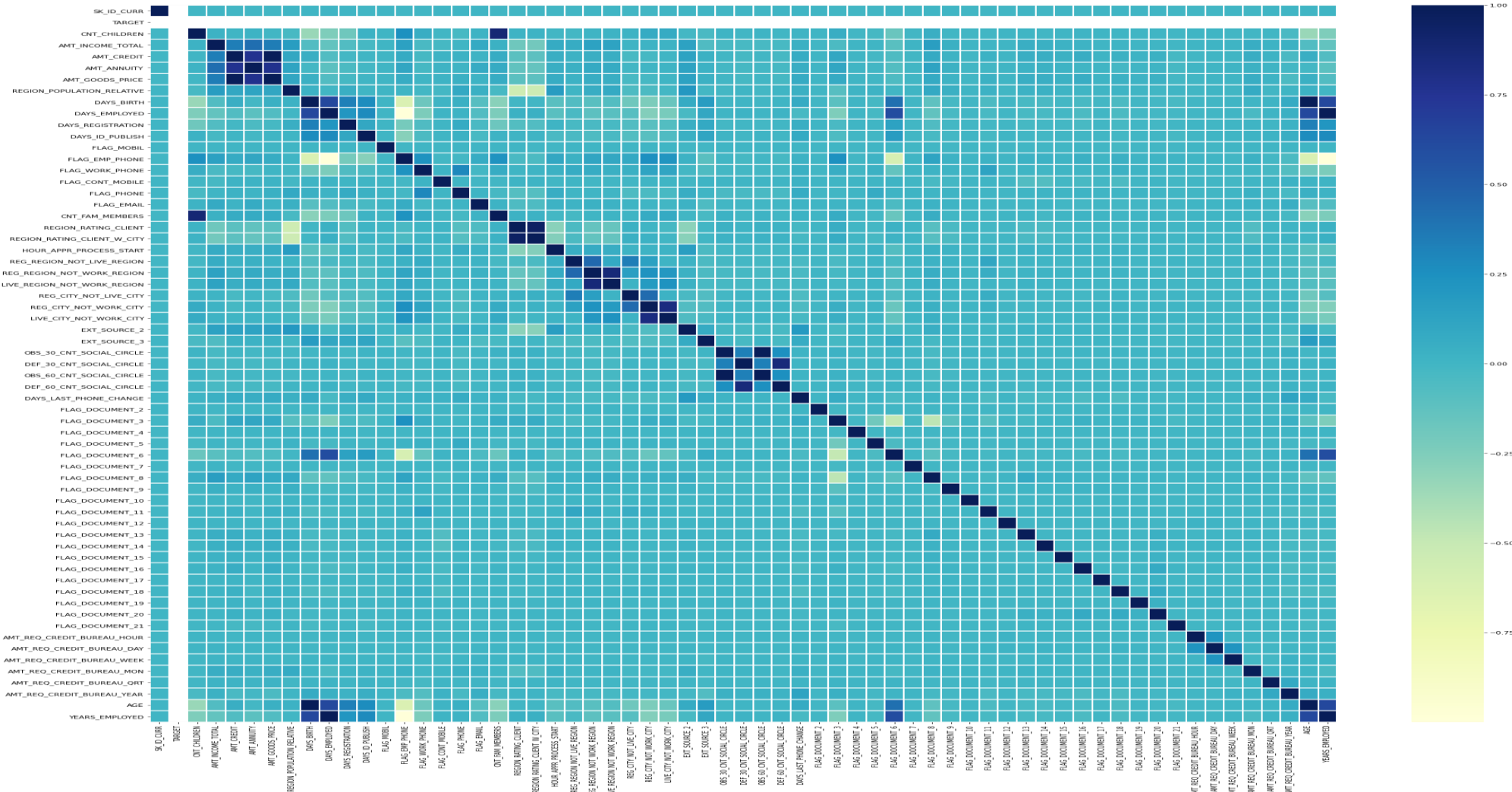
For loan re-payers credit amount is highly correlated with:
1) amount of goods price
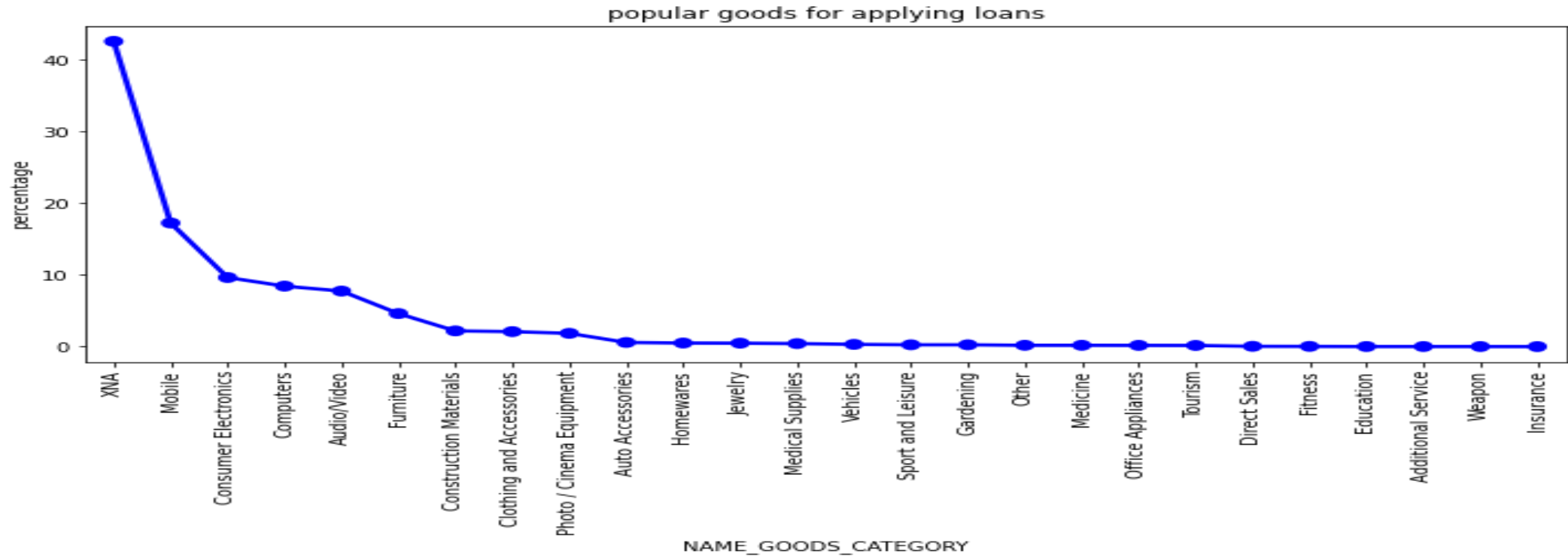2) total income
3) employment days

Gender wise loan type

Distribution of Contract type by gender -application_data

1) Female are taking more loans as compare to Male
2) Cash loans is always preferred over Revolving loans by both genders

Popular goods for loan

popular goods for applying loans

XNA ,Mobile,Computers and consumer electronics are popular goods for applying loans

SUMMARY

- *Females are taking more loans as compare to Males*

- *Income range in between 100k-200k have high rate of repaying loans*

- *Applicants age above 50 having low probability of defaulting.*

- *When the credit amount goes beyond 2M, there is an increase in defaulters.*

- *Applicants with 40+ year experience having around 1% default rate*

- *Applicant having Academic degree has rate of repaying loan.*