Credit EDA Assignment

Submitted by Anand Umrani

Submitted on 1st Nov 2022

Data Science Program – August 2022

DS C47 Aug EPGDS

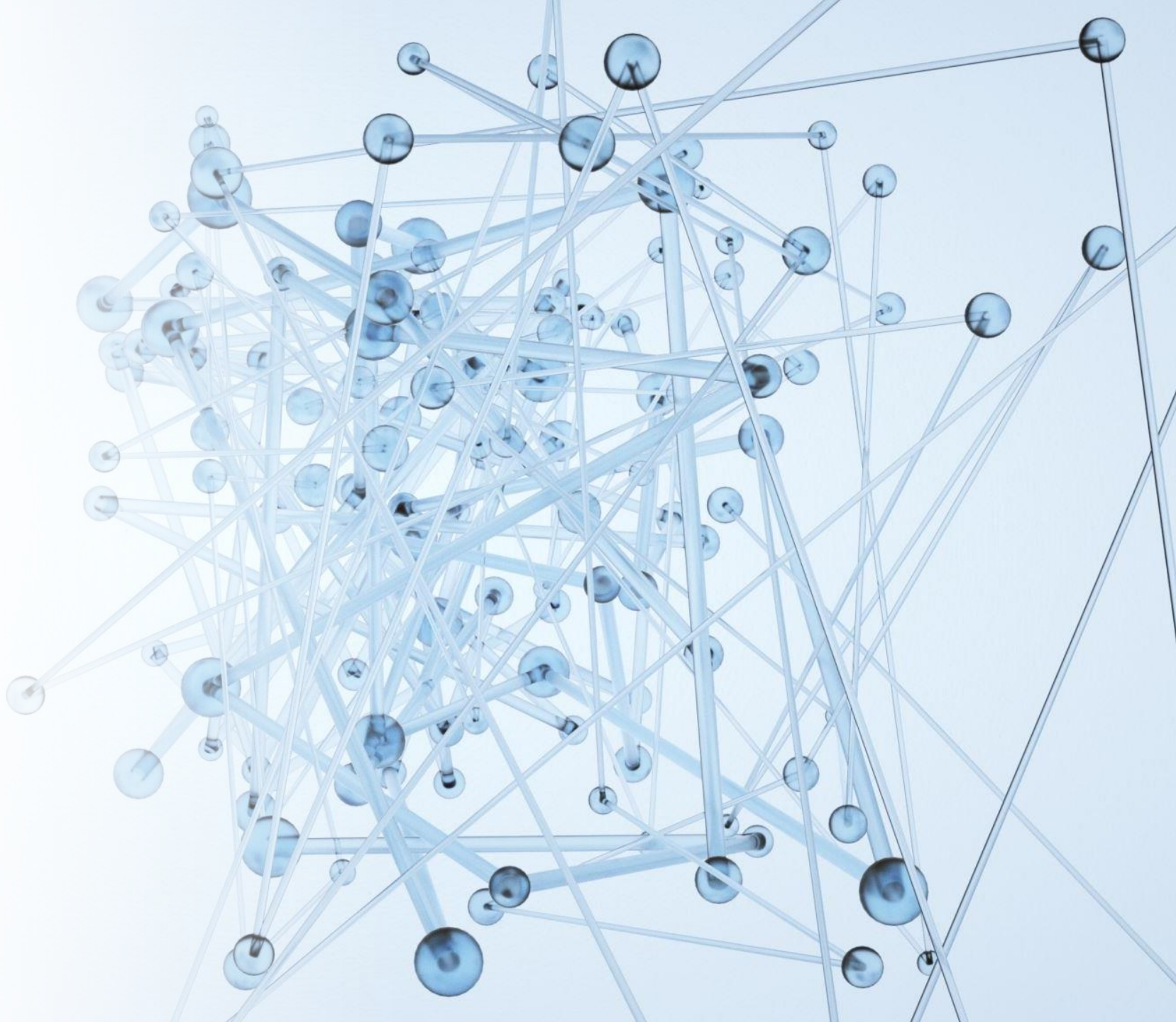


Table of Contents

Business Understanding

Problem Statement

- Consumer finance company which specialize in lending various types of loans to urban customers find it hard to give loans to the people due to:
  - their insufficient or
  - non-existent credit history
- When the company receives a loan application, the company must decide for loan approval based on the applicant's profile.
- Two types of risks are associated with the bank's decision:
  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
  - 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.

Business Objectives:

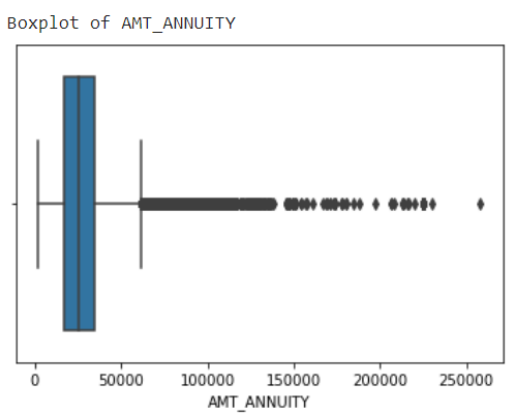
- To identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.
- This will ensure that the consumers capable of repaying the loan are not rejected.
- 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.

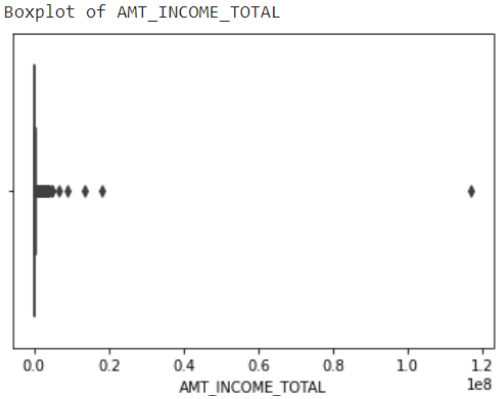
EDA Approach Taken

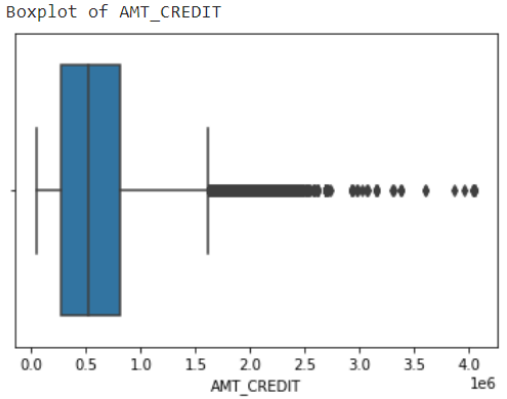
- Develop the understanding of domain - Business Problem and Objectives
- Understanding the Data provided – variables
- Importing the Data Sets
- Check the structure of the dataset
- Data Cleaning and Manipulation
  - missing value imputation analysis and removing data redundancies
  - Standardization of data – removing negative values, creating buckets, etc.
  - Converting to appropriate data types
- Identify outliers in the dataset*
  - Insights from outliers
- Data Analysis:
  - Data Imbalance
  - Perform univariate analysis
  - Perform bi/multivariate analysis
- Provide recommendations

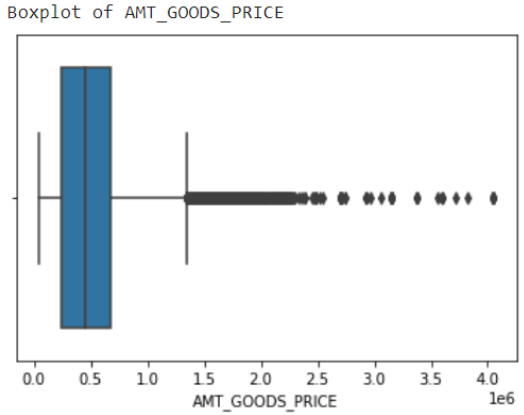
* *As per case study objective, removal of outliers is not expected, hence they are kept as is*

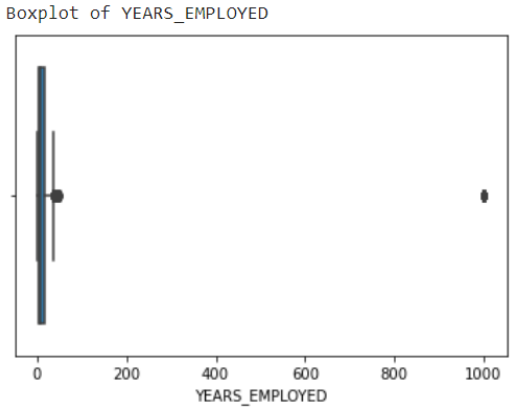
Current Application Data – Outliers Analysis

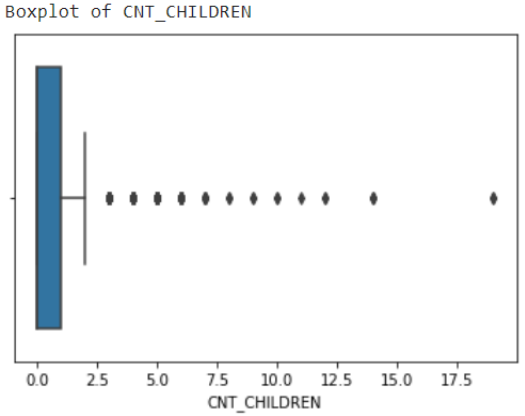


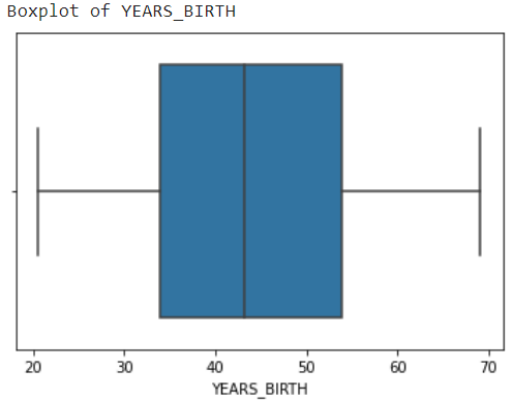






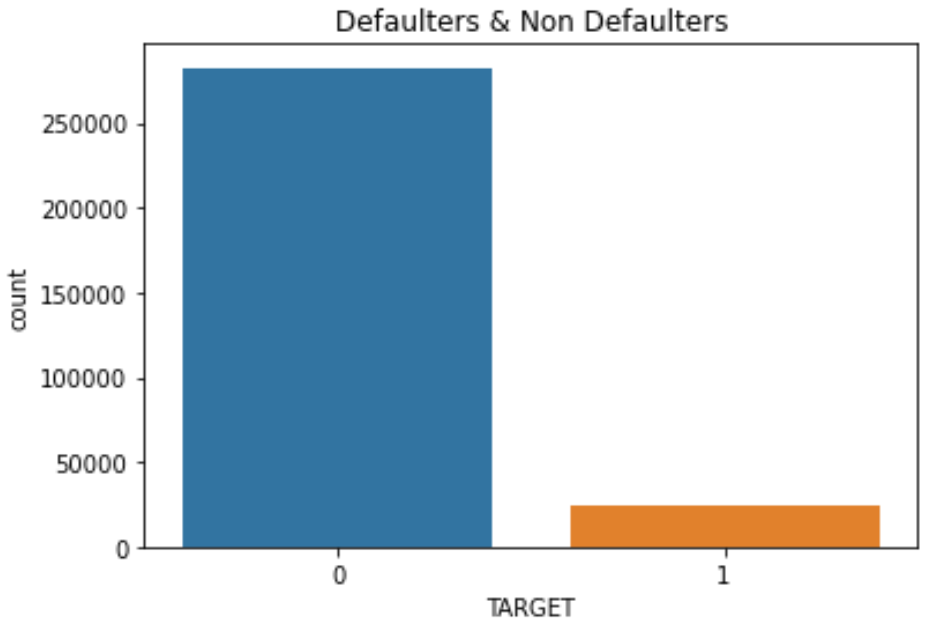






- Following have outliers: AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN
- There are huge outliers in AMT_INCOME_TOTAL indicating that there are people with extremely high income as compared to the others.
- YEARS_EMPLOYED has outlier values such as 1000 YEARS which is not natural. This may be because of incorrect entry or some other problems while capturing the data.
- CNT_CHILDREN indicates values beyond 15 which again seems to be not natural. It means having more than 15 children.
- Of all the plots DAYS_BIRTH has no outliers which means the data is good.

Current Application Data Analysis



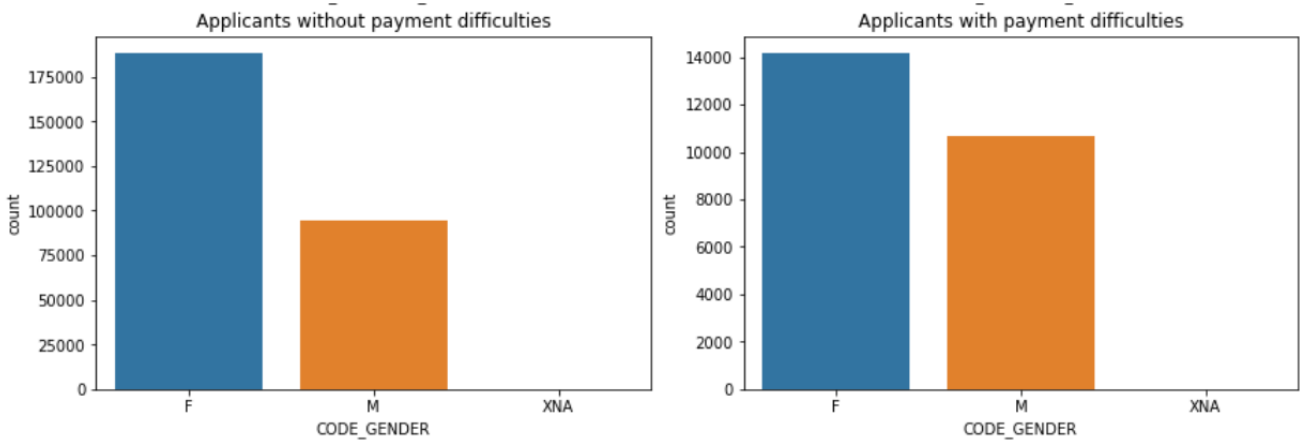

```
appdata1['TARGET'].value_counts(normalize=True)*100
```

```
0    91.927118
1     8.072882
Name: TARGET, dtype: float64
```

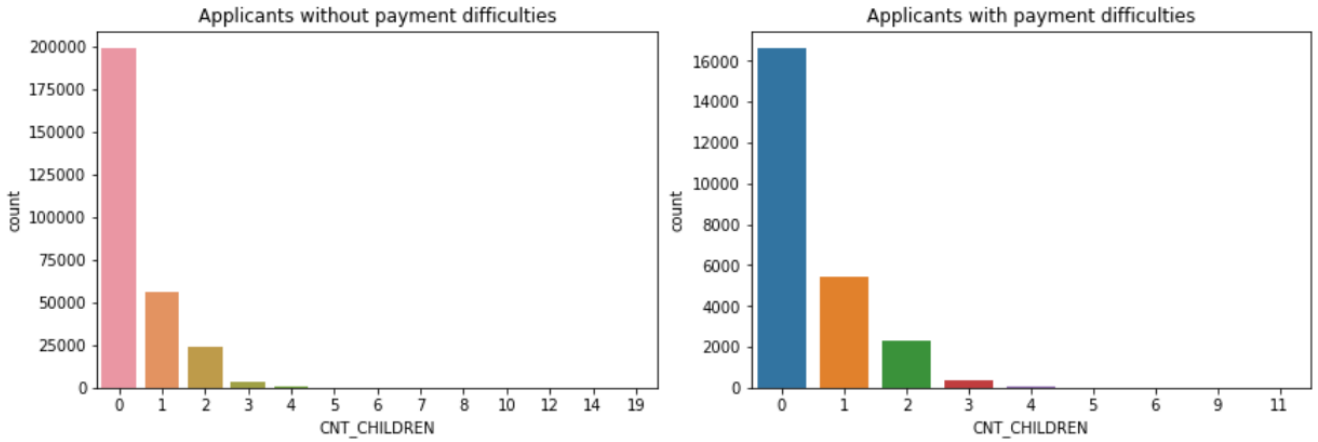
- 92% of the applicants have done payment on time and rest 8% have defaulted.
- This seems to be a practical scenario and there is problem in the data set at high level.
- This is also a good situation for bank.

Current Application Data Analysis

- Females are taking more loans
- Amongst defaulter's men are having higher chances of defaulting than females

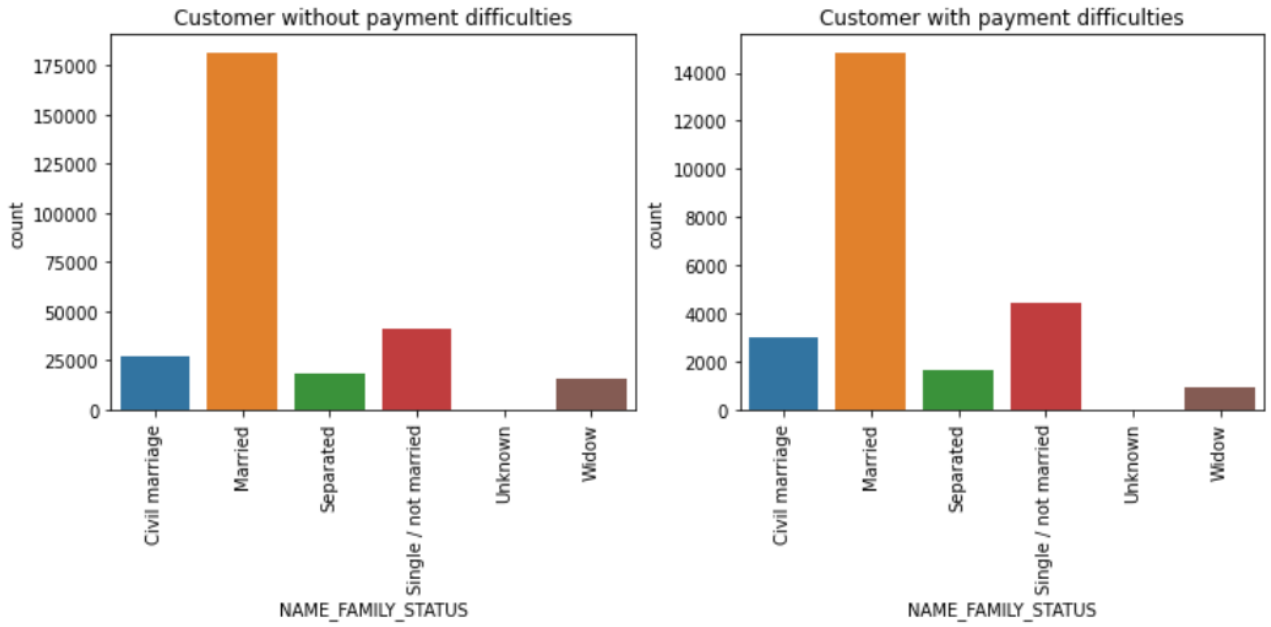


- Majority of the applicants do not have children.
- Applicants with zero to two children tend to repay the loans. Rest all may be risky

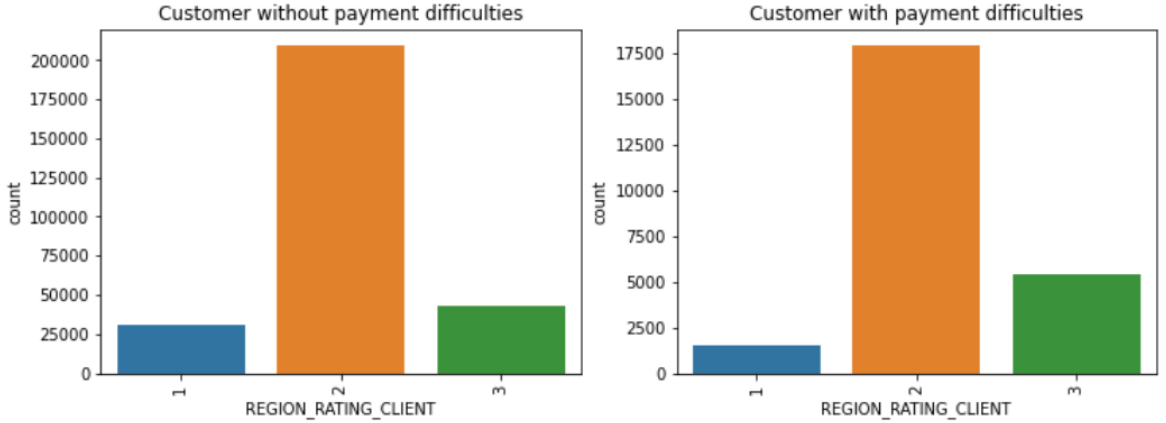


Current Application Data Analysis

- Married people apply for loans more than anyone else
- People who have civil marriage or who are single default a lot.

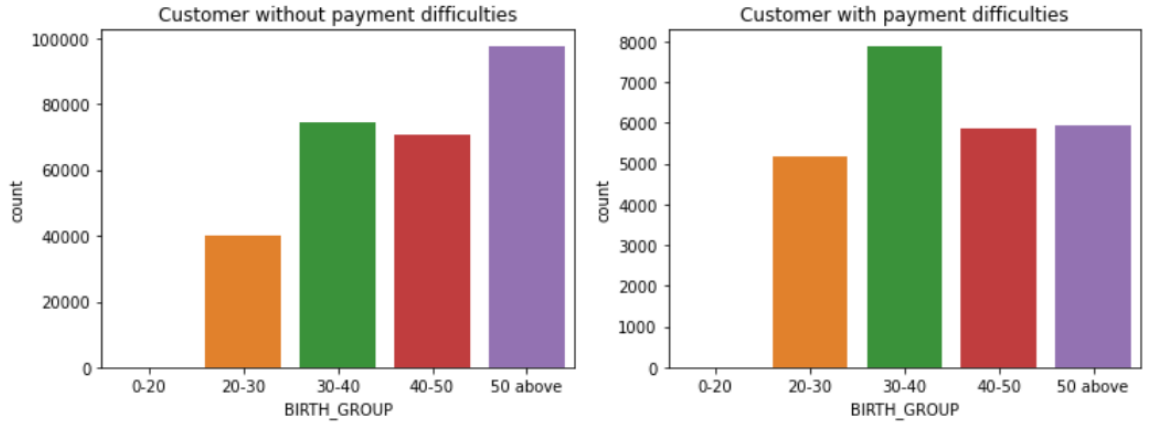


- Majority of the applicants are from region rating 2
- Applicants from region rating 3 are having max defaults
- Region rating 1 applicants seem to be less defaulters

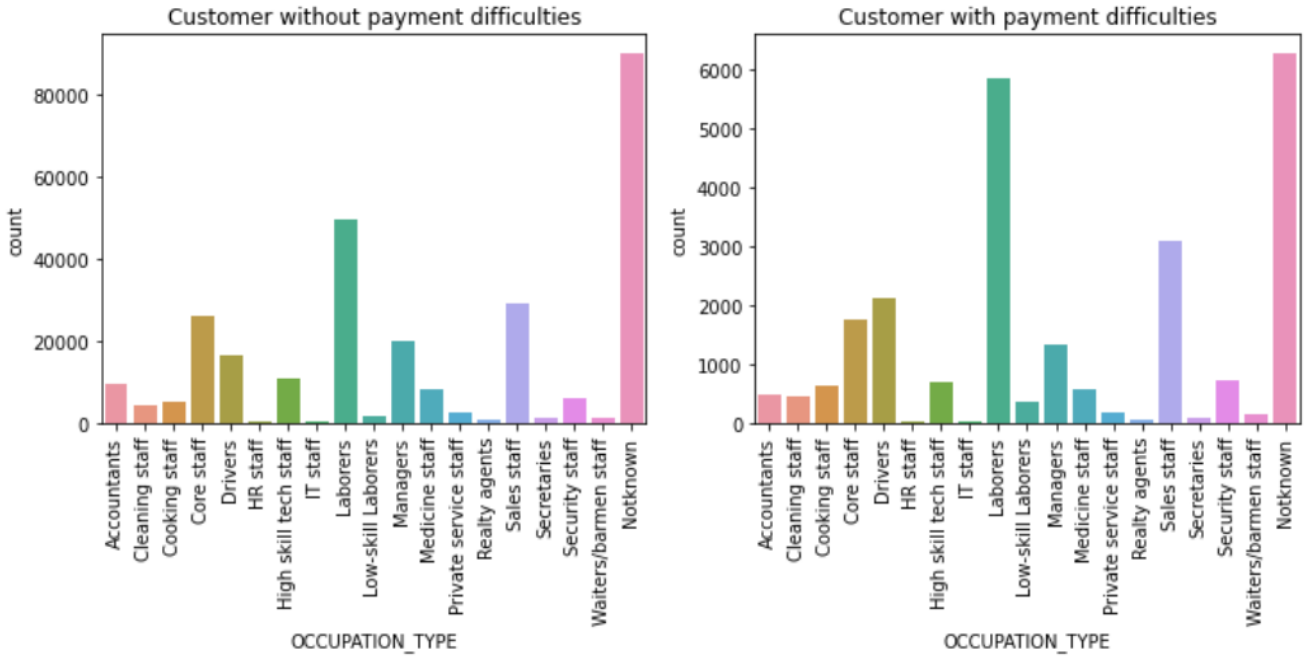


Current Application Data Analysis

- Majority of applicants is residing in middle age (30 to 50 years)
- Age group of 20-40 are having higher likelihood of defaulting
- Whereas above age of 50 have less likelihood of defaulting



- Following are high risk applicants Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff



Current Application Data Analysis

Numeric Variable Analysis - Top 10 Correlations Analysis

Non_defaulters

	pairs	corr
40	(AMT_CREDIT, AMT_GOODS_PRICE)	0.987022
23	(CNT_CHILDREN, CNT_FAM_MEMBERS)	0.878571
50	(AMT_ANNUITY, AMT_GOODS_PRICE)	0.776421
39	(AMT_CREDIT, AMT_ANNUITY)	0.771297
77	(YEARS_BIRTH, YEARS_EMPLOYED)	0.626114
28	(AMT_INCOME_TOTAL, AMT_ANNUITY)	0.418948
29	(AMT_INCOME_TOTAL, AMT_GOODS_PRICE)	0.349426
27	(AMT_INCOME_TOTAL, AMT_CREDIT)	0.342799
78	(YEARS_BIRTH, YEARS_REGISTRATION)	0.333151
85	(YEARS_EMPLOYED, YEARS_ID_PUBLISH)	0.276663

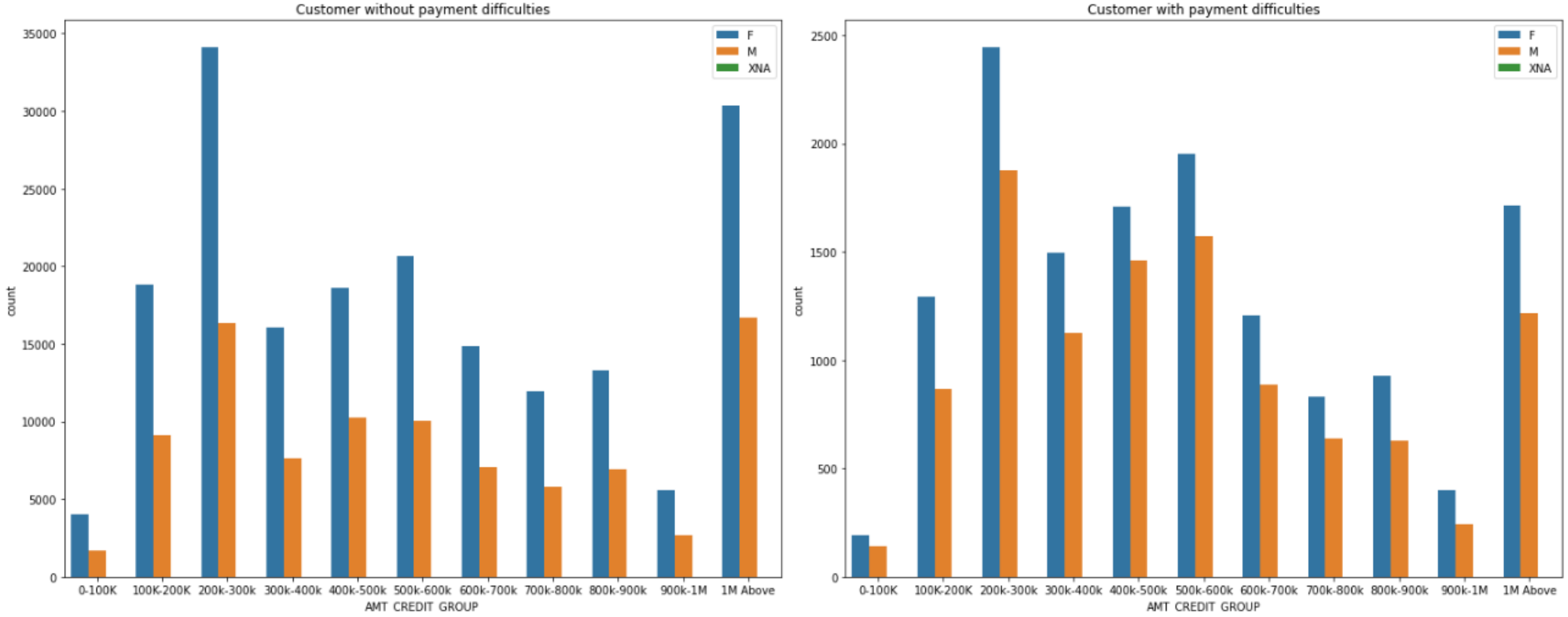
Defaulters

	pairs	corr
40	(AMT_CREDIT, AMT_GOODS_PRICE)	0.982783
23	(CNT_CHILDREN, CNT_FAM_MEMBERS)	0.885484
50	(AMT_ANNUITY, AMT_GOODS_PRICE)	0.752295
39	(AMT_CREDIT, AMT_ANNUITY)	0.752195
77	(YEARS_BIRTH, YEARS_EMPLOYED)	0.582185
78	(YEARS_BIRTH, YEARS_REGISTRATION)	0.289114
79	(YEARS_BIRTH, YEARS_ID_PUBLISH)	0.252863
85	(YEARS_EMPLOYED, YEARS_ID_PUBLISH)	0.229090
103	(EXT_SOURCE_2, YEARS_LAST_PHONE_CHANGE)	0.207071
84	(YEARS_EMPLOYED, YEARS_REGISTRATION)	0.192455

- 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 repayers(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 repayers.
- Days_birth and number of children correlation has reduced to 0.259 in defaulters when compared to 0.337 in repayers.
- There is a slight increase in defaulted to observed count in social circle among defaulters(0.264) when compared to repayers(0.254)

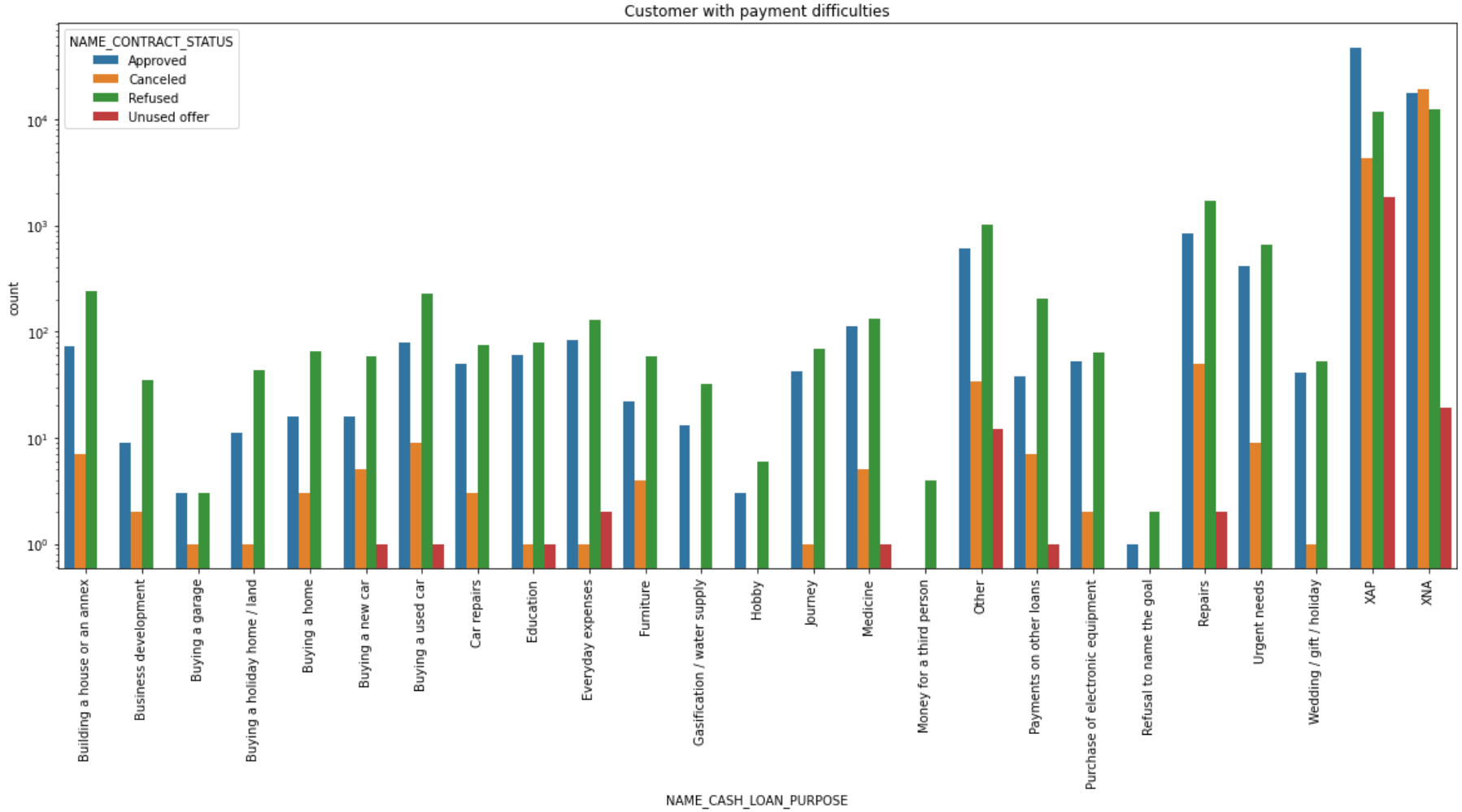
Current Application Data Analysis

Bivariate analysis on CODE_GENDER and AMT_CREDIT_GROUP



- Male applicants who get credit between 300K to 600K are the majority of the defaulters

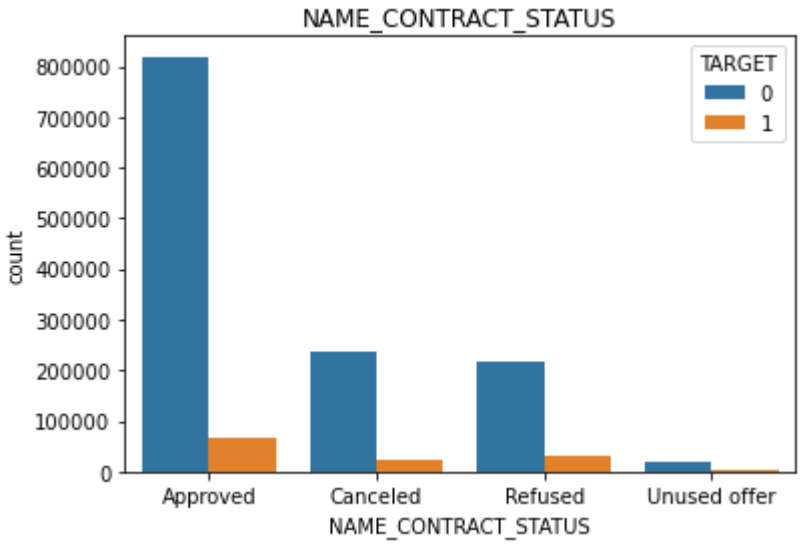
Merged Data Analysis



- CASH_LOAN_PURPOSE has high number of values - XAP and XNA
- Loan taken for repairs, urgent need and others are having highest defaulters
- Whenever loan has been taken for repair and other work, the bank has rejected or the applicant has refused.

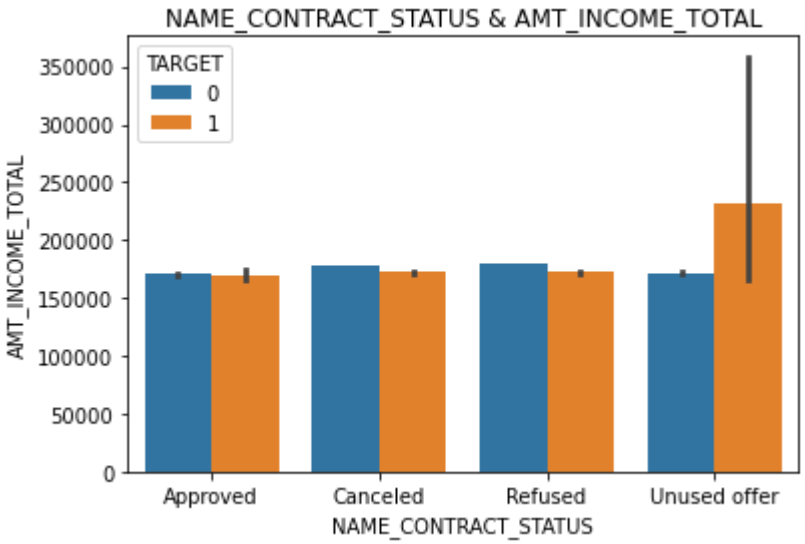
Merged Data Analysis

CONTRACT_STATUS & Loan Repayment



- Majority of the prev cancelled applicants have repayed the loan.
- Lot of applicants who have been previously refused a loan have done a repayment in current case

Total Income & Contract Status



- Applicants with higher income than others have defaulted with contract status - unused offer
- Lot of people are cancelling or leaving their offers unused. This needs further investigation

Recommendation

- **Gender**: Females are less likely to default and generally men are high risk, especially who get credit between 300K to 600K
- **Family**: Applicants with few children tend to pay back loans
- **Age**: Applicants above age of 50 are safest while age group of 20-40 is high risk
- **Credit Amount**: People (both male and females) 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.
- **Loan Purpose**: Loan taken for repairs, urgent need and others are having highest defaulters
- **Contract Status**:
  - Majority of the prev cancelled applicants have repayed the loan.
  - Lot of people are cancelling or leaving their offers unused. This needs further investigation