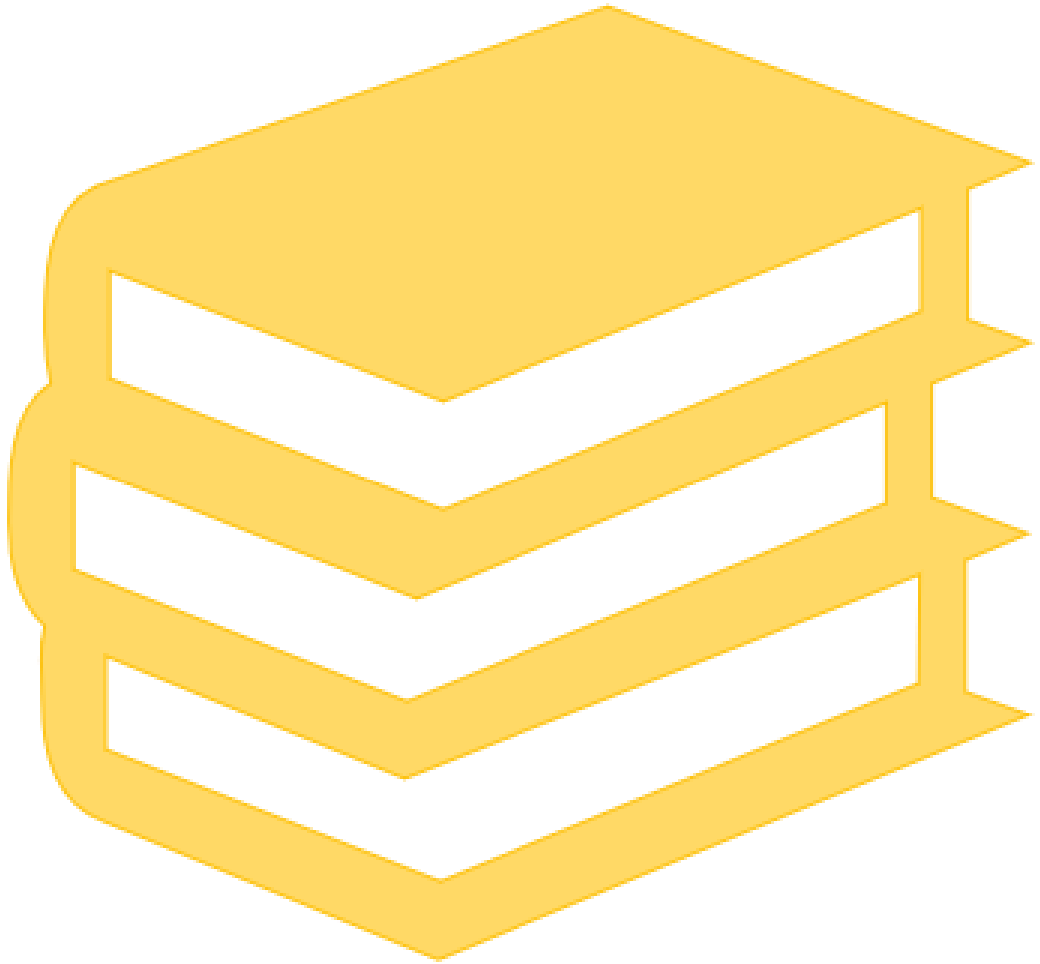


EDA CASE STUDY

Presented by Jitesh Garg

Problem Statement

Lending companies find it hard to give loans to people due to insufficient or non-existent credit histories. This decision of extend credit is associated with two distinct types of risks:

1. **Risk of Missed Opportunity:** If the applicant demonstrates a strong likelihood of loan repayment but is denied the loan. In such cases, declining the loan application may result in missed revenue and growth for the company.

2. **Risk of Default:** Conversely, if the applicant exhibits a higher probability of loan default, meaning they are likely to fail to repay the loan, approving the loan can expose the financial institution to potential financial losses.

Business Objective

- This case study aims 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.

- In other words, 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 utilise this knowledge for its portfolio and risk assessment.

Steps & Methodologies followed (1/2)

1. Understanding Dataset

- Imported relevant libraries.
- Analyzed headers, indexes, columns, top 5 & bottom 5 rows.
- Checked data size & statistical summary .
- Looked for data quality issues (if any).

2. Data Cleaning

- Identifying datatypes.
- Dropping irrelevant columns.
- Impute/remove missing values.
- Standardising the values.
- Fixing invalid values (if any).

Steps & Methodologies followed (2/2)

3. Outlier Handling & Data Imbalance Check

- Checked data for identifying outliers that may lead to biased results.
- Checked data for ratio and percentage imbalance.

4. Data Analysis

- Conducted Univariate Analysis, Bivariate Analysis, plotted various charts for inferences and useful insights.

5. Conclusion

- Derived insights on both aspects – (defaulters and non-defaulters) for business to take informed decisions.

Assumptions

XNA

- In column ‘CODE_GENDER’, we have assumed ‘XNA’ with 4 frequency as null value and replaced it with most frequent value i.e., ‘F’.

XNA & XAP

- In column ‘NAME_CASH_LOAN_PURPOSE’, ‘XNA’ & ‘XAP’ with high frequency. However, these values doesn’t make any sense. Therefore, we have dropped these values.

Unknown

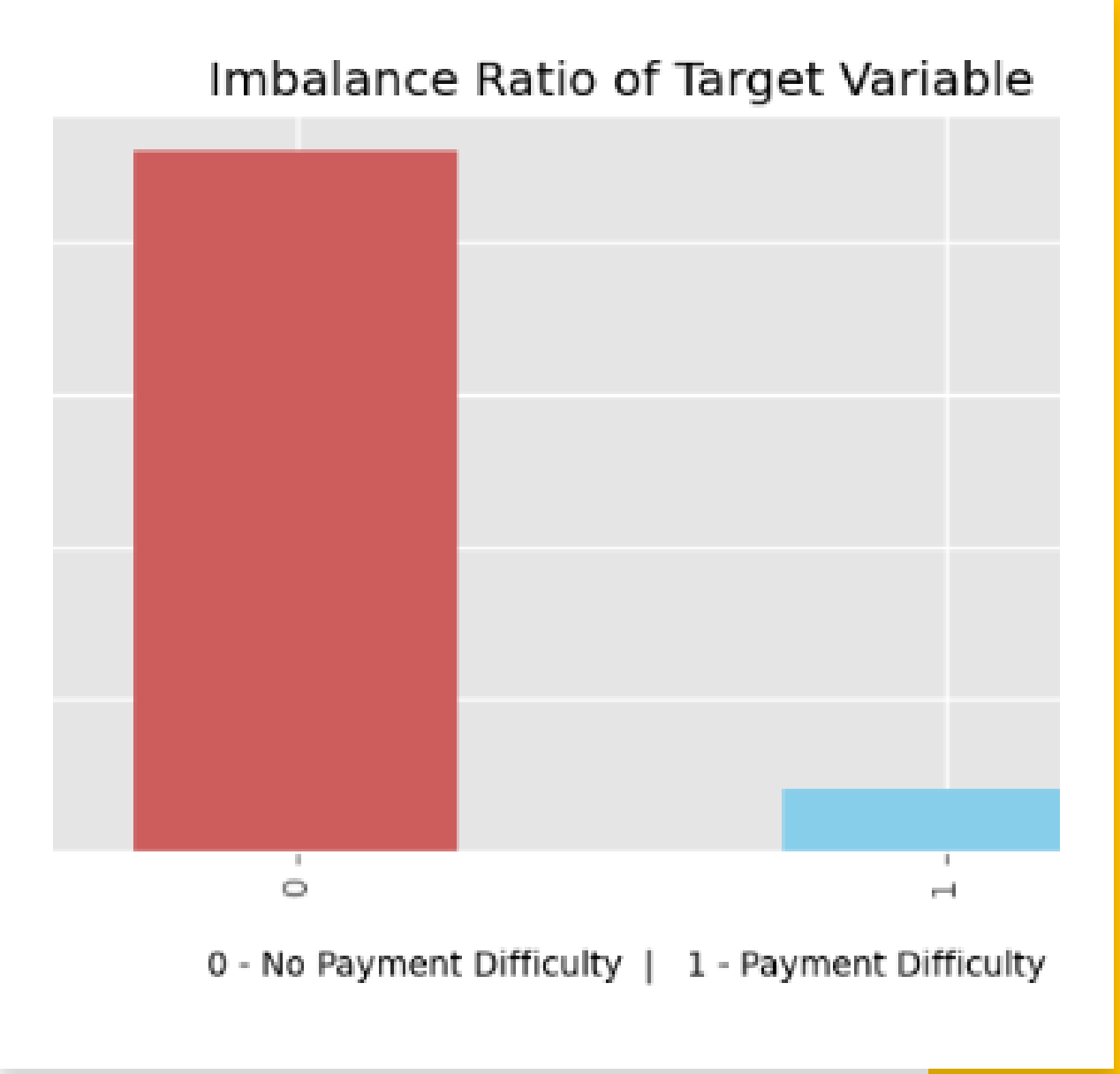
- In column ‘NAME_FAMILY_STATUS’, we have assumed ‘Unknown’ with 2 frequency as null value and replaced it with mode (i.e., ‘Married’).

Data Imbalance

Inference:

- 91.93% applicants pays on time while 8.07% applicants have payment difficulties.

- The ratio for people who have payment difficulties is 11.387150050352467 which means ----> 1 in every 11 applicant faces payment difficulty.



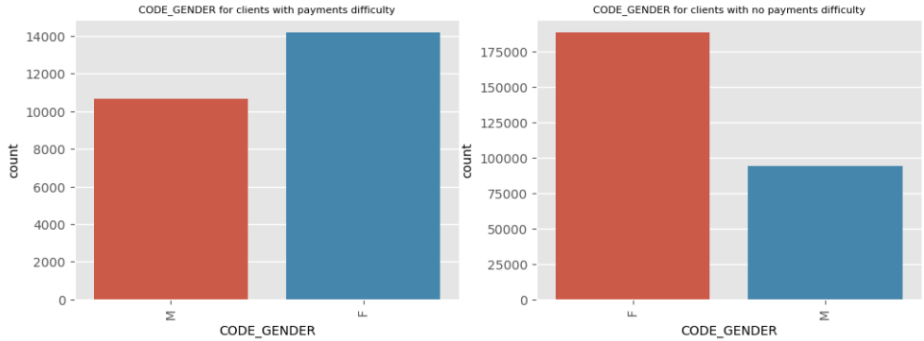
Segment Analysis - Male vs Female

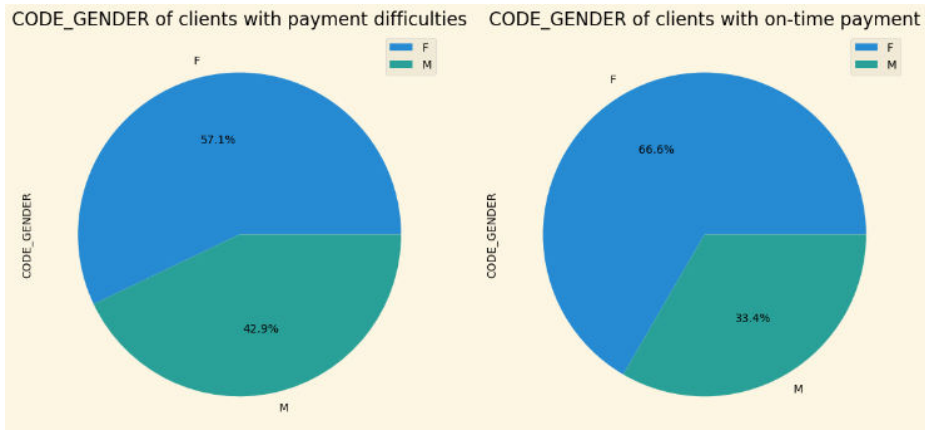
Inference:

- Looking at the first bar chart, there are around 14000 female vs 11,000 male faces payment difficulty. Whereas on increased scale in second bar chart around 1,90,000 female vs 90,000 male do not have payment difficulty. Hence, it is clear that female has lesser difficulty in repaying the loan.

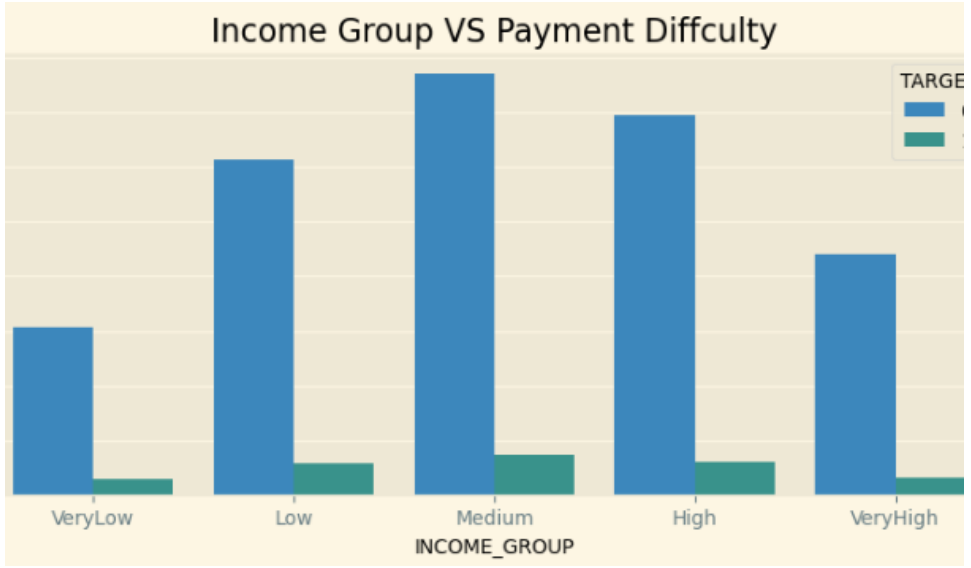
- Pie chart also clearly states that 66.6% female makes payment on time while only 33.4% male makes timely repayment.

- Banks should focus more on female segment for successful payments.



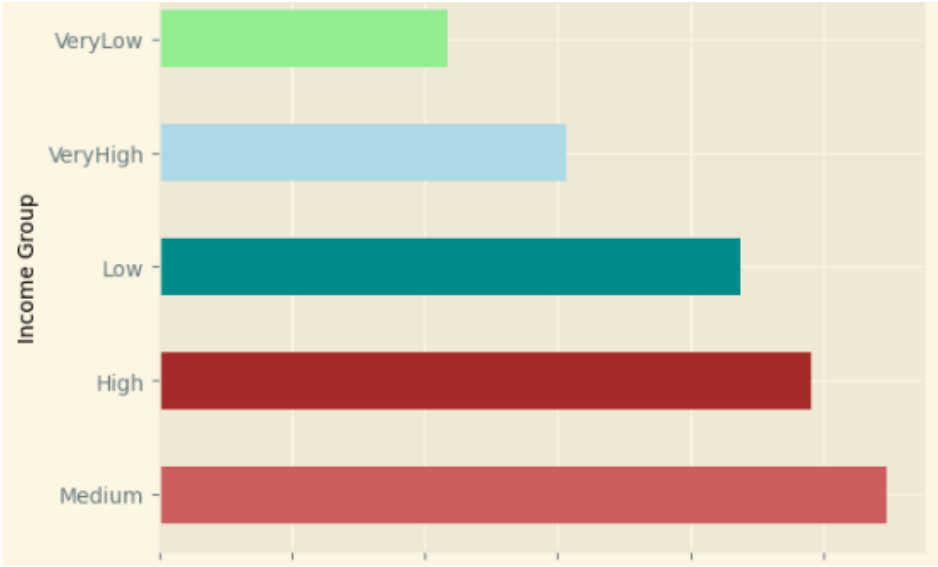


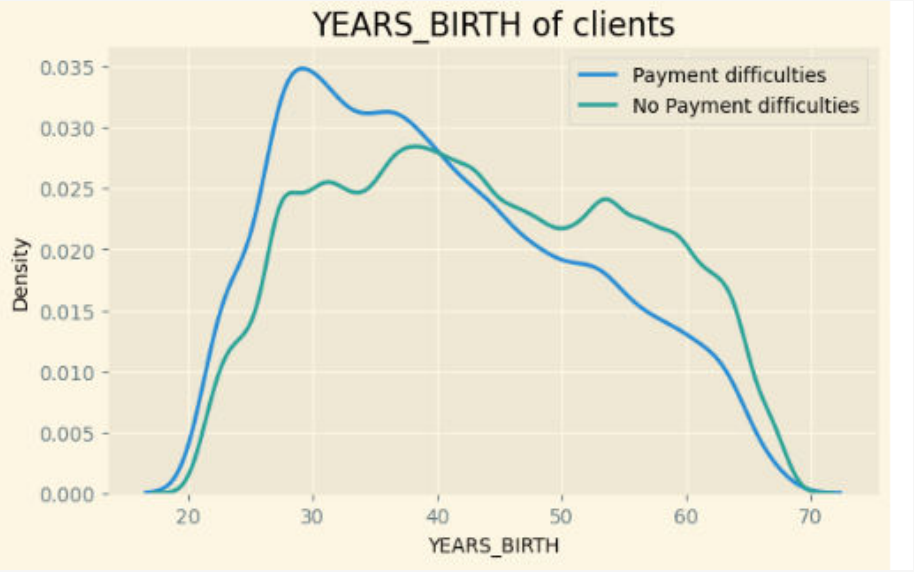
Income Group

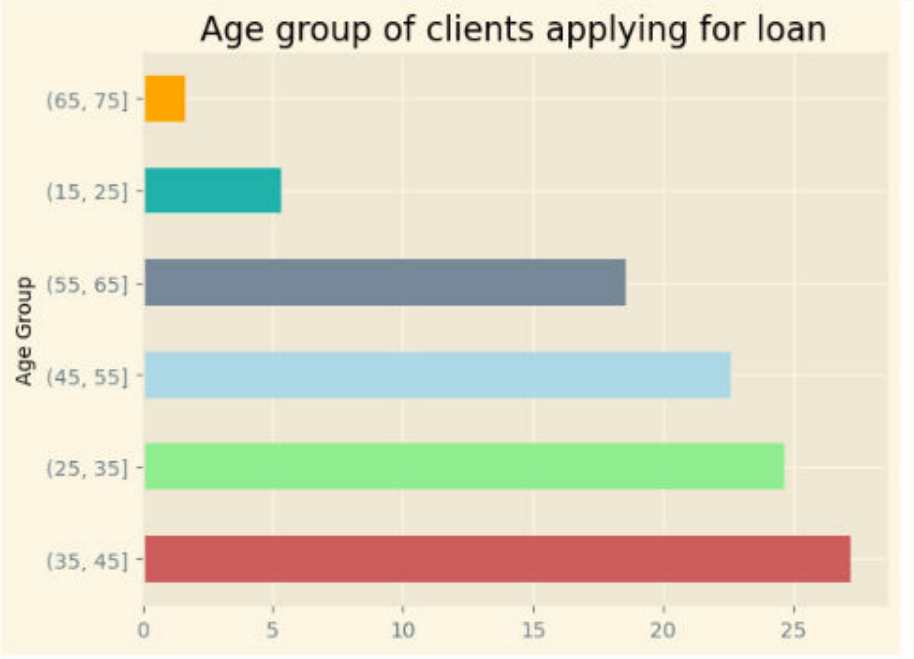


Inference:

- Medium Income group are the largest applicants for loan, followed by high income group. However, medium income group has higher difficulty in repaying the loan, followed by high income group.





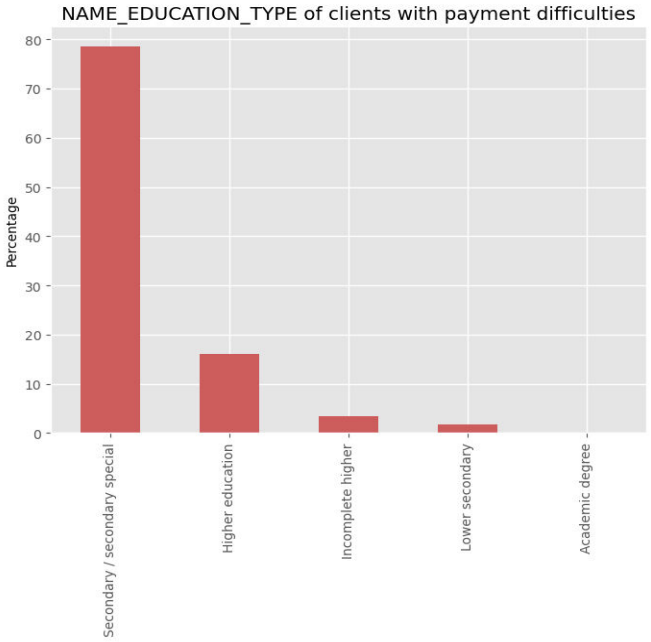


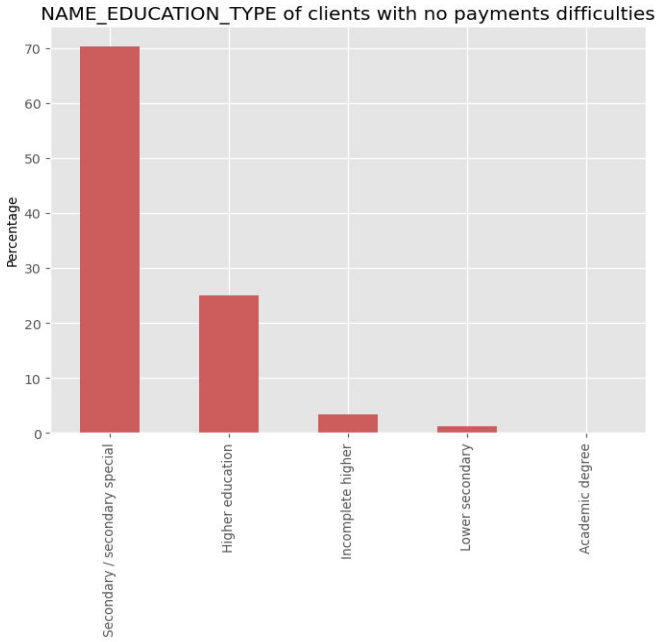
Age factor

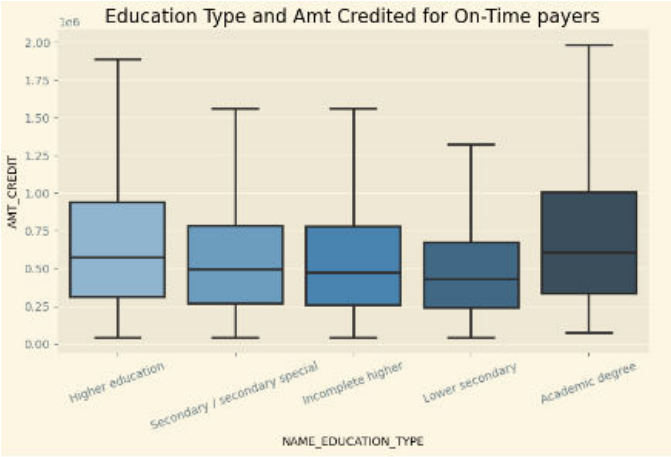
Inference:

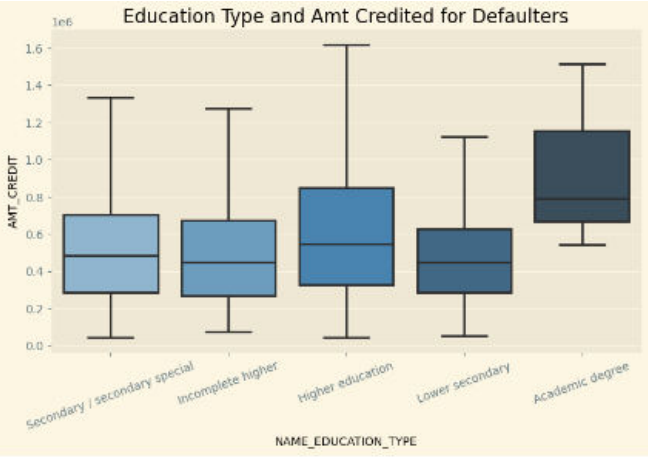
- 35-45 age group is the largest Group which apply for loans.
- There are more clients with payment difficulty for Years_Birth between 20 & 40. While Years_Birth > 40, there are more clients with on-time payments.
- Clients have payment difficulty in the age group of 25-35 years, followed by 35-45 years of age group.

Analysis on Education Type





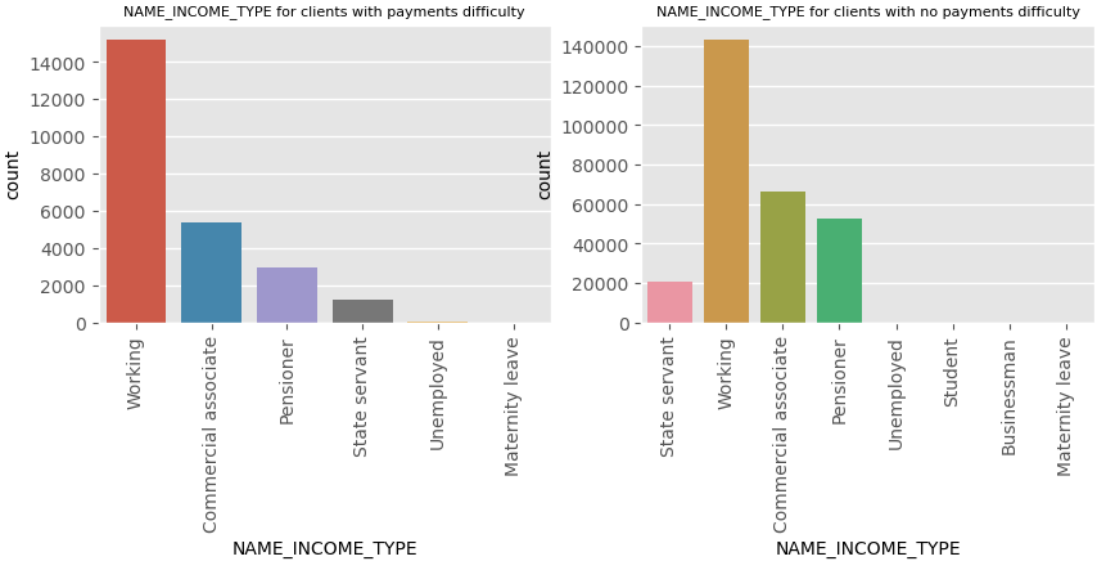


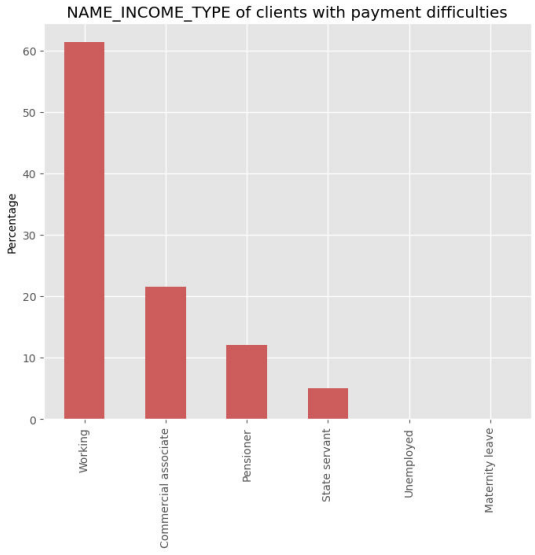


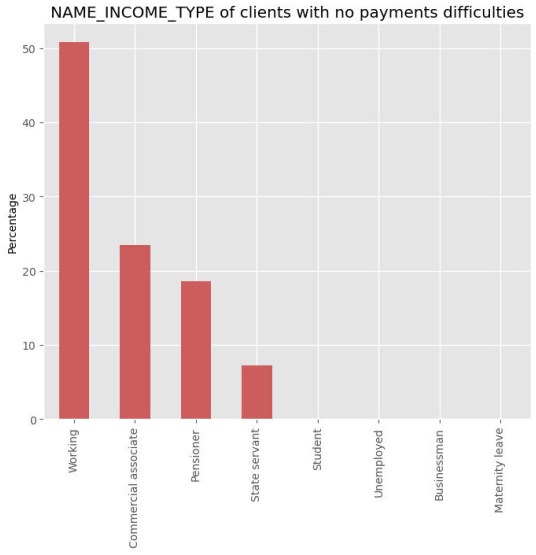
Inference:

- Clients with 'Higher education' have better on-time payments than others and have less payment difficulties.
- Median of Loan values defaulting for Applicants with Academic degree is higher.
- Clients with higher education should get higher preference.

Analysis on Job Type



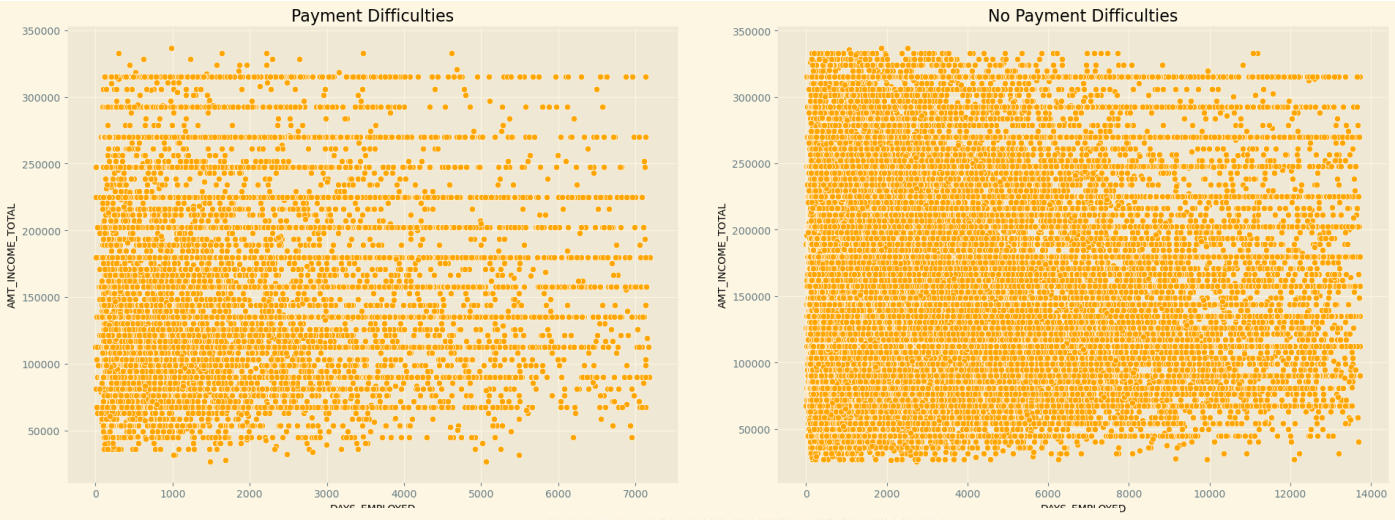


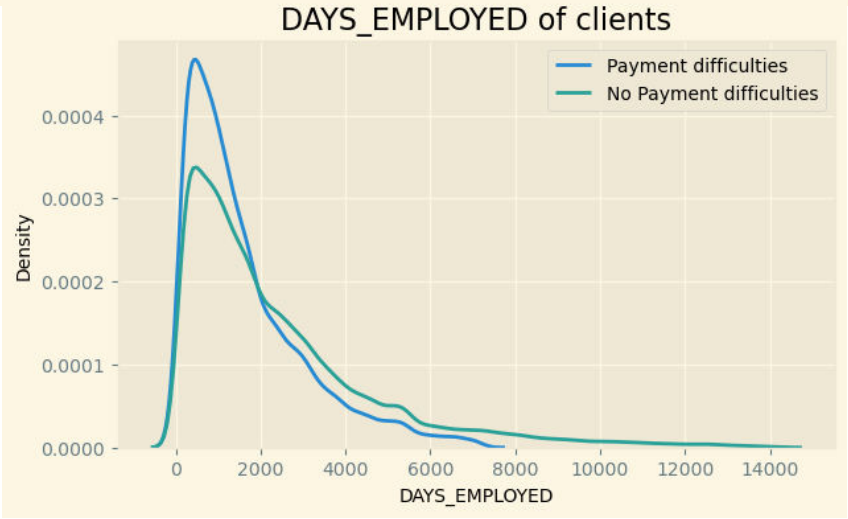


Inference:

- 'Working' class have processed more loans in comparison to other categories.
- Pensioners have better on-time payments.
- Students don't have Payment difficulties.
- Businessmen don't have Payment difficulties.

Analysis on Employment Years



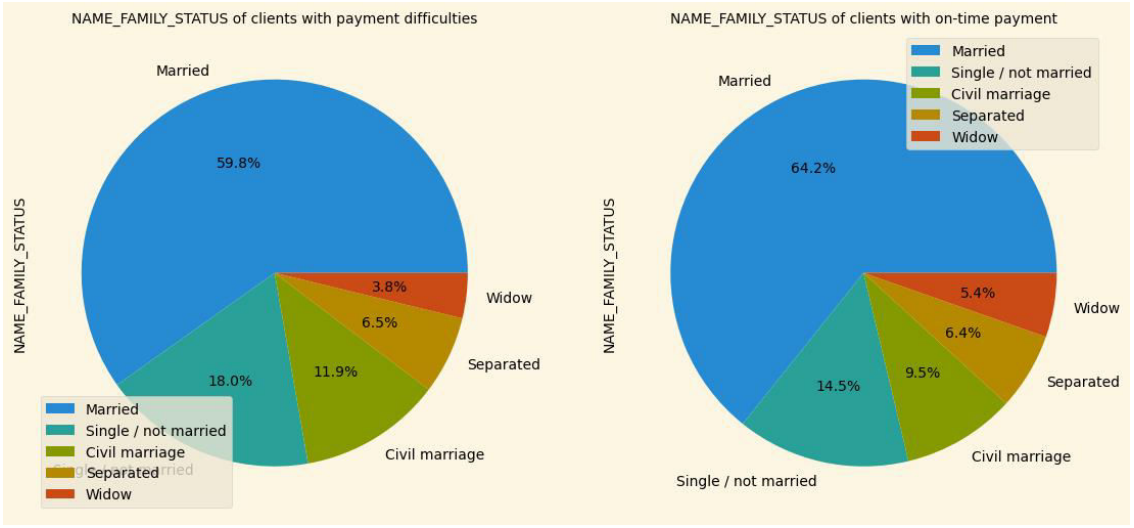


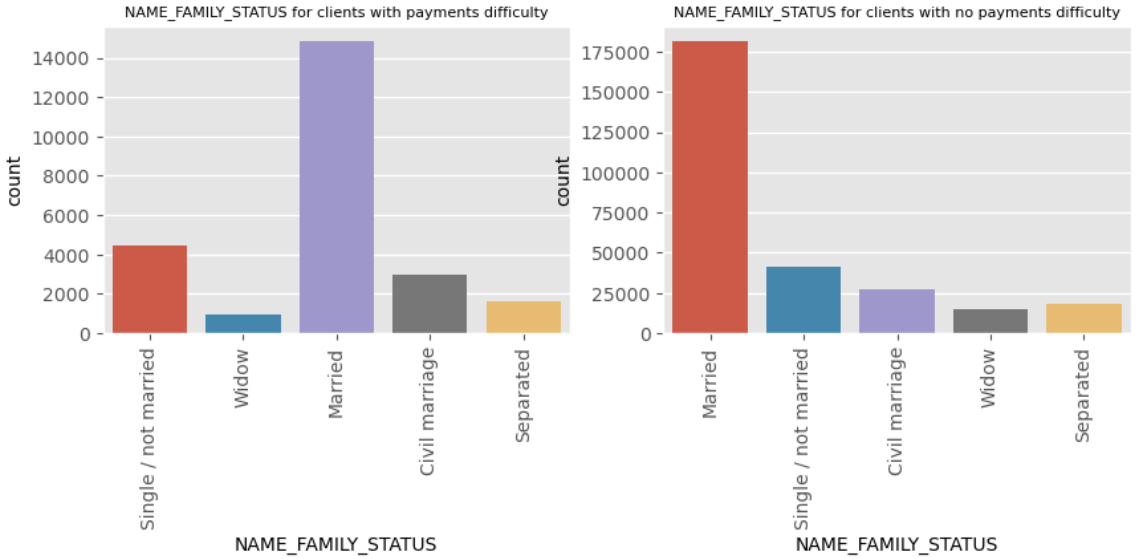
Inference:

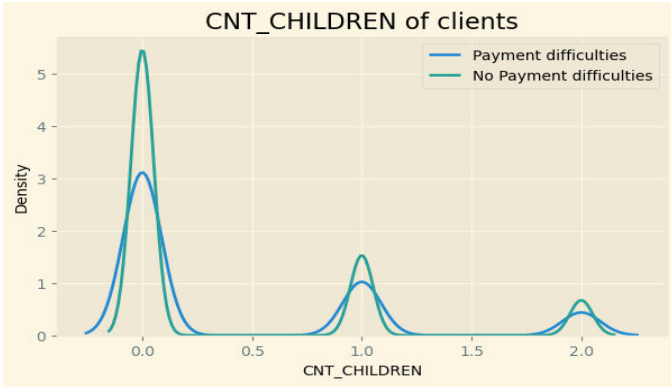
- Clients who are employed for a long time days are making their payments on-time but these category of clients do not exist in payments difficulties Group.

- For DAYS_EMPLOYED > 2000 , there are more clients with no payment difficulties, implying that those who are employed longer have better chances of repaying the loan.

Family Status



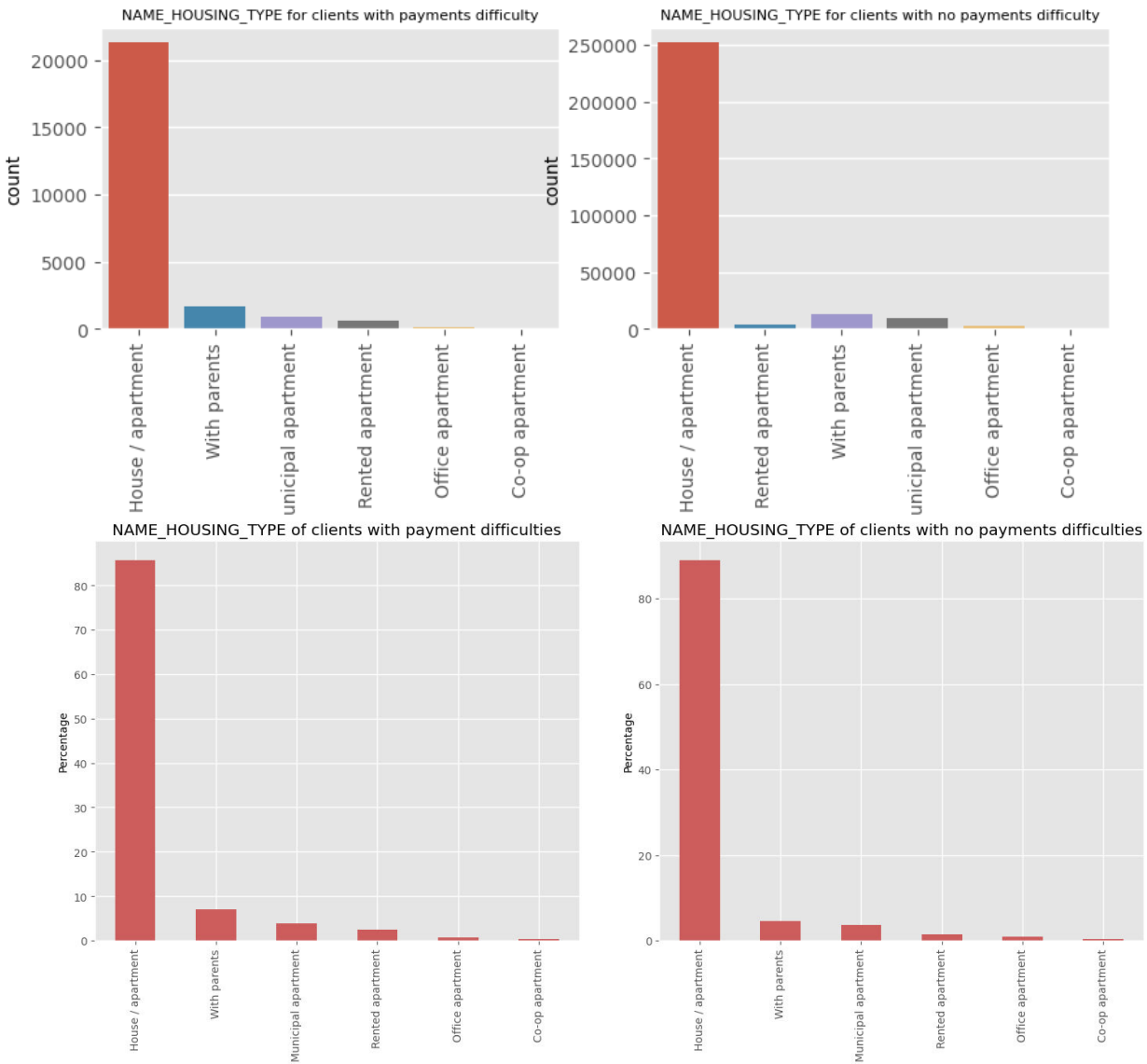




Inference:

- People with no children tends to make on-time payments.
- Clients who are 'Married' do on-time payments better comparatively.
- Clients who are 'Single/not married' have more difficulties with on-time payments comparatively.
- Married people are the ones mostly applying for loan followed by Single or Not married.
- Widows are the ones who have less number of application for a loan.

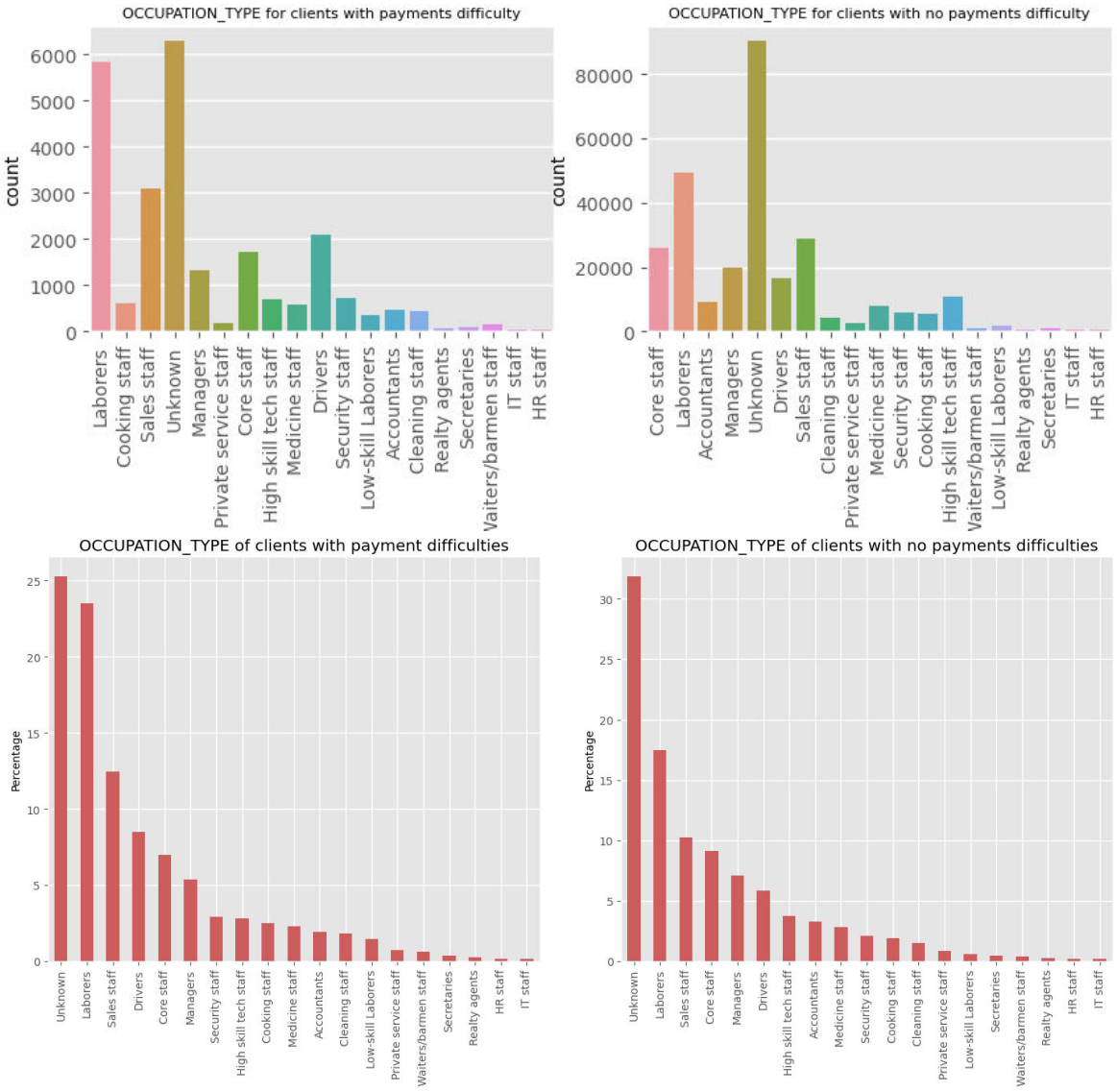
Analysis on Housing Type



Inference:

- Applicants who own a house have processed more loans in comparison to those who don't.
- People who own house have better on-time repayment record.
- People who live in rented flat or with parents tends to default more on loan payments than people living in House/Flat or office flats.

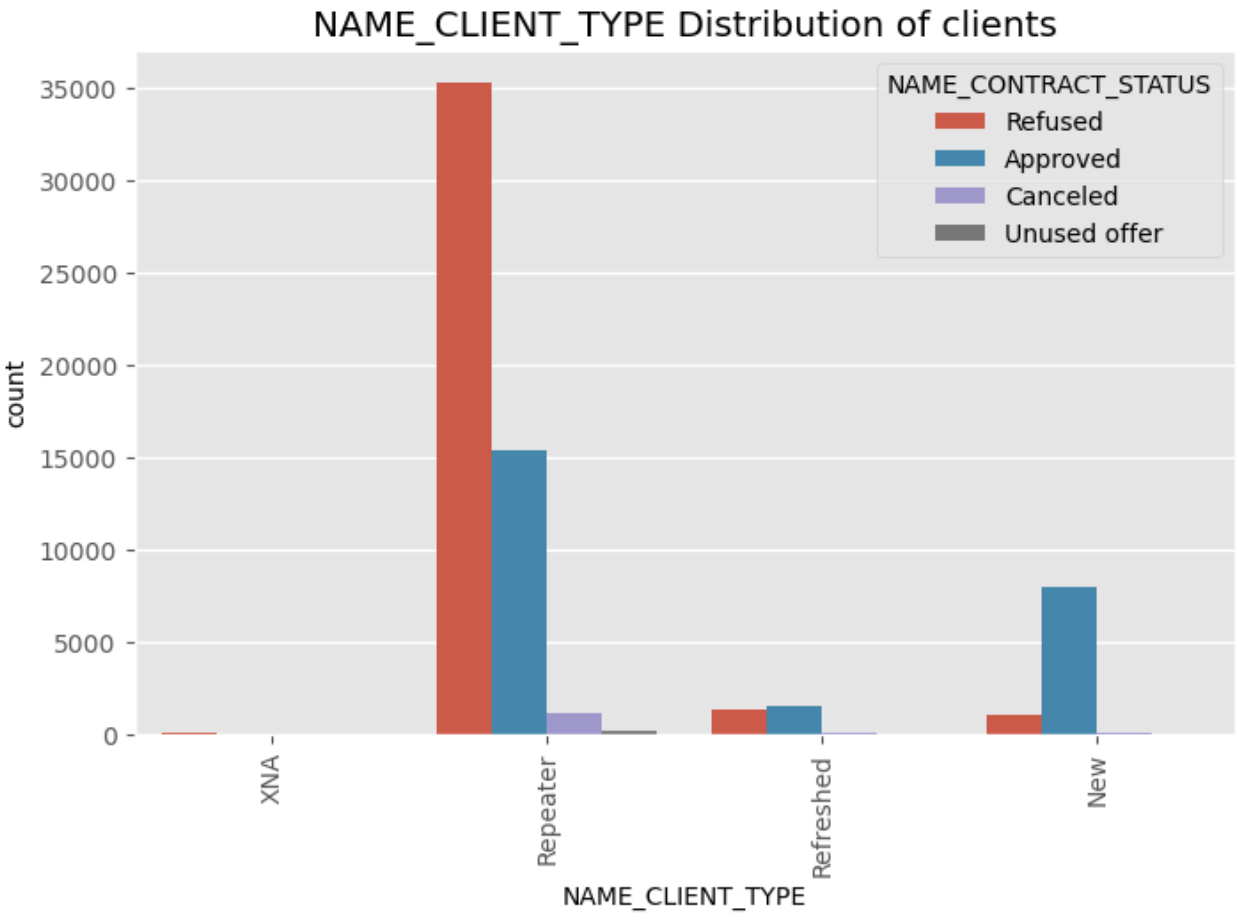
Analysis on Client Occupation



Inference:

- Labourers are on top to apply for loan followed by sales staff and core staff.
- Maximum percent of people defaulting on payments are from the occupations :Low skilled labourers, Drivers, waiters/barmen staff.
- Accountants, High skill tech staff and manager are very less likely to default.

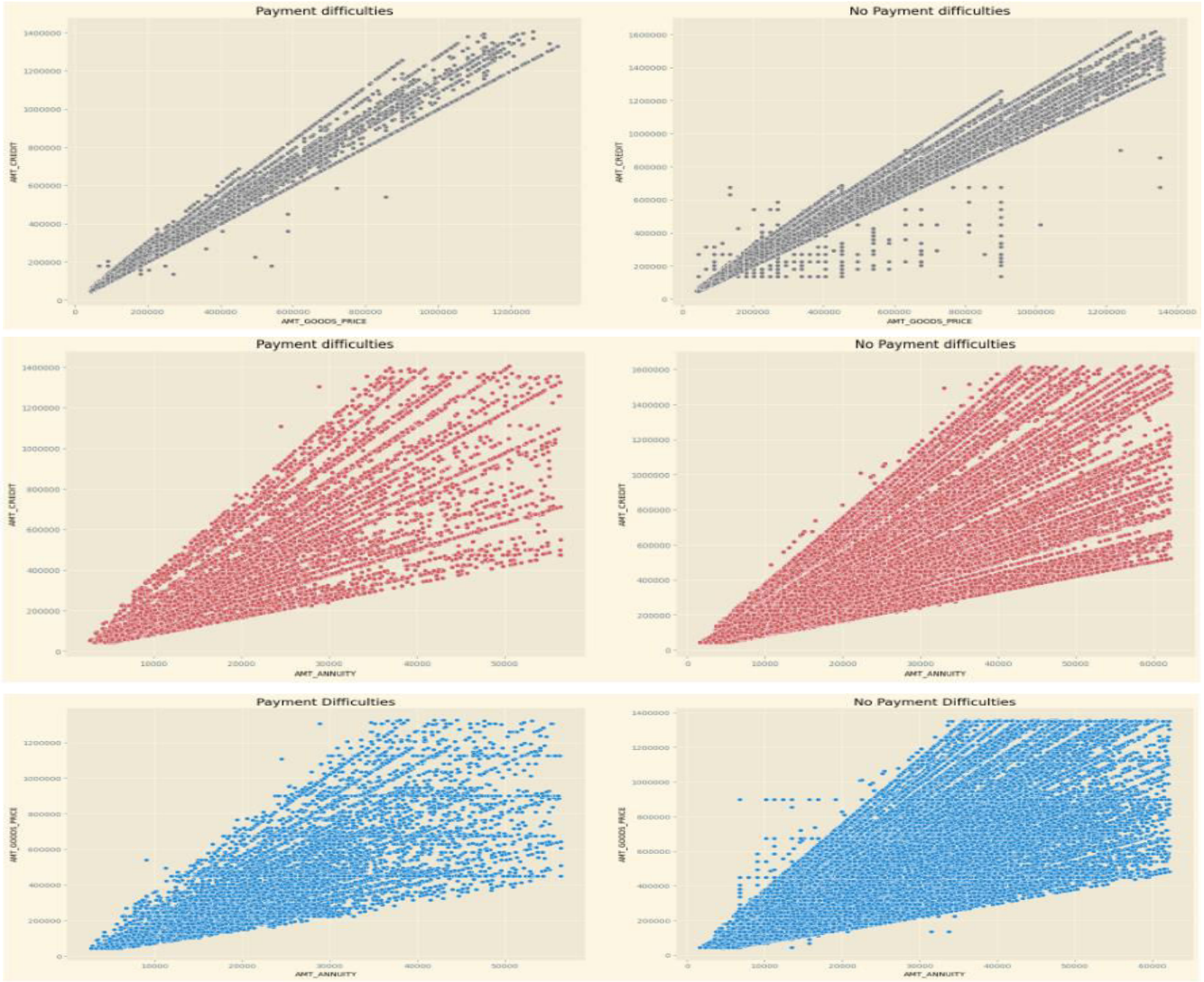
Distribution of clients basis contract status



Inference:

- The company has more repeaters in all approved, refused, unused, cancelled categories.
- PoS transactions seem to be more for repeaters, and more loans have been refused for repeaters than other group.
- Client cancelling the loan application is high in the case of repeaters

Positive Correlation between different variables



Inference:

- AMT_GOODS_PRICE and AMT_CREDIT showed a strong positive correlation. This implies that as the goods price increases, the credit amount also tends to increase.
- AMT_ANNUITY and AMT_CREDIT demonstrate a strong positive correlation. This suggests that as the annuity amount increases, the credit amount also tends to increase.
- AMT_ANNUITY and AMT_GOODS_PRICE exhibit a strong positive correlation. This indicates that as the annuity amount increases, the goods price also tends to increase.

Conclusion

Target Clients:

- Female applicants should be given extra weightage as defaults cases are lesser
- Pentioners tend to pay on time. Students & Businessmen also have no problem in repayment of the loan
- Clients who are Married
- Clients with higher education
- Clients who are employed for more years
- Clients who own a house
- Previously cancelled, refused, unused loan cases
- Repeater clients

Defaulter Clients:

- Male applicants
- Medium income
- 25-35 years - age group
- Single/not married
- Unemployed
- Labourers, Drivers, waiters/barmen staff
- Don't own house, stay with parents, rented flats


Thank you!