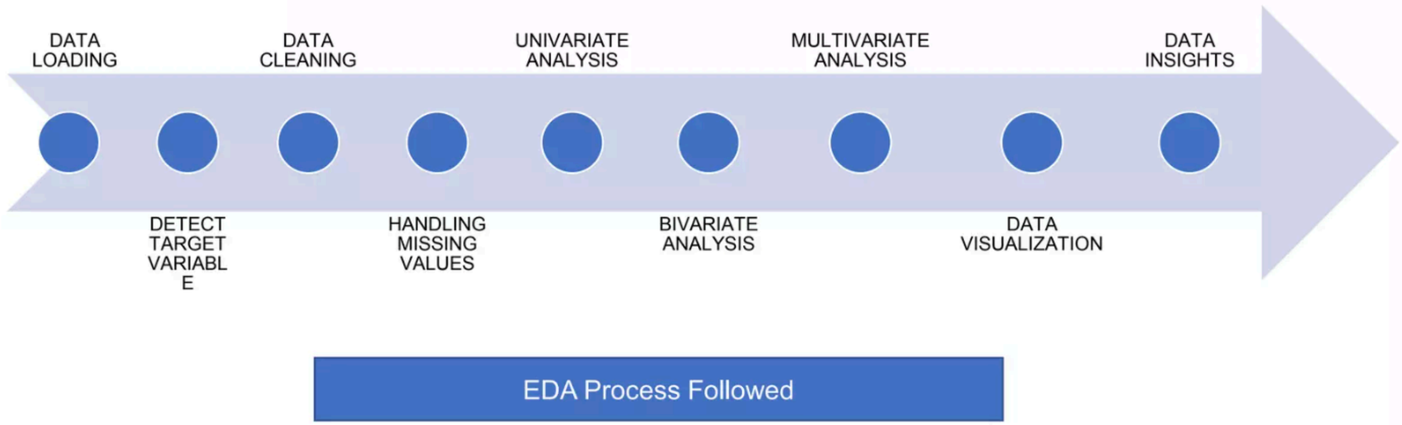
Credit EDA Case Study

By Mr Uttam Kumar

Purpose

Credit Risk Analysis:
Mitigating Losses and Ensuring Financial Stability and Making Informed Decisions for Loan Approval

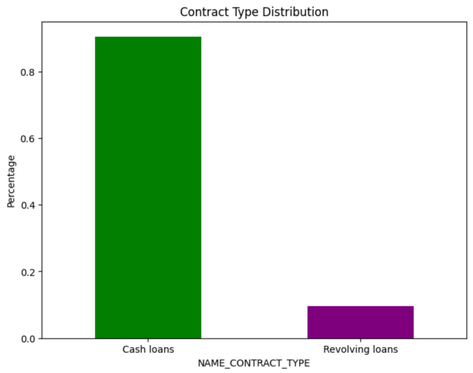
Exploratory Data Analysis Flow chart

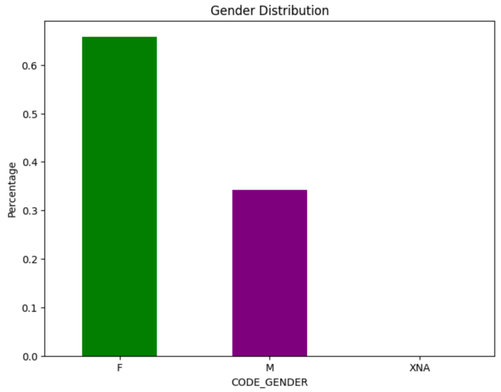


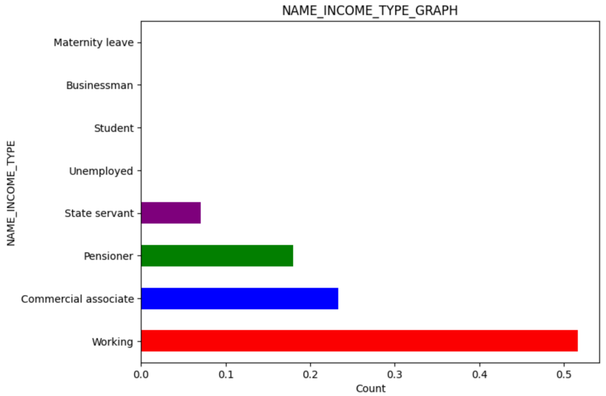
Exploratory Data Analysis of Application Data

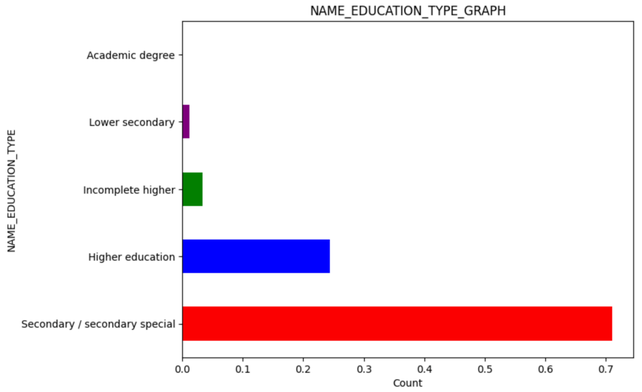
Univariate Analysis

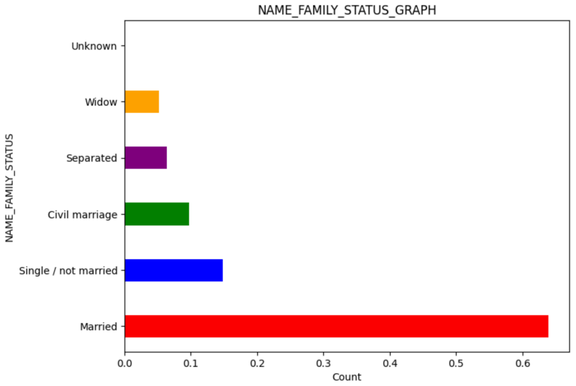
Categorical Nominal











key observations :

Most of the money offered is in the form of cash loans, while revolving loans make up a small portion.

The percentage of female loan applicants was 65%, which is higher than that of male applicants, by a margin of 31%. This finding suggests that there may be underlying reasons for the gender disparity.

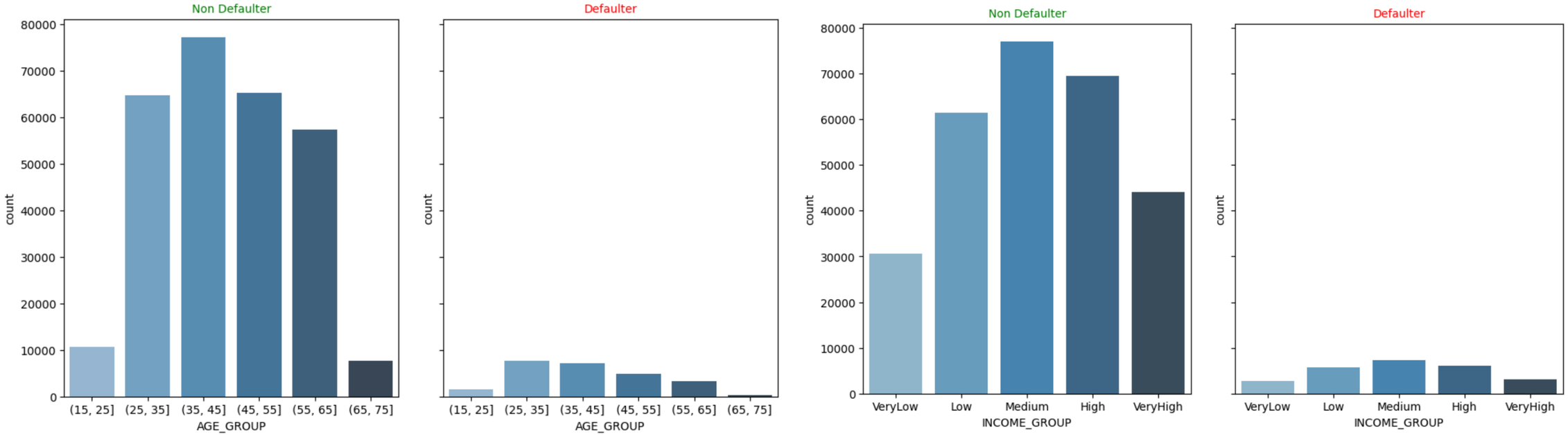
Although the majority of applicants are working class, it's interesting that 18% are pensioners. This group should be considered due to their varying risk profiles and financial needs.

The educational attainment of the applicants is noteworthy, with over 70% of them having completed their secondary education. This indicates that there is a well-educated pool of potential borrowers.

Most of the people who applied for a loan were married.

Segmented Univariate Analysis

Age and Income Category Segmented two Variables

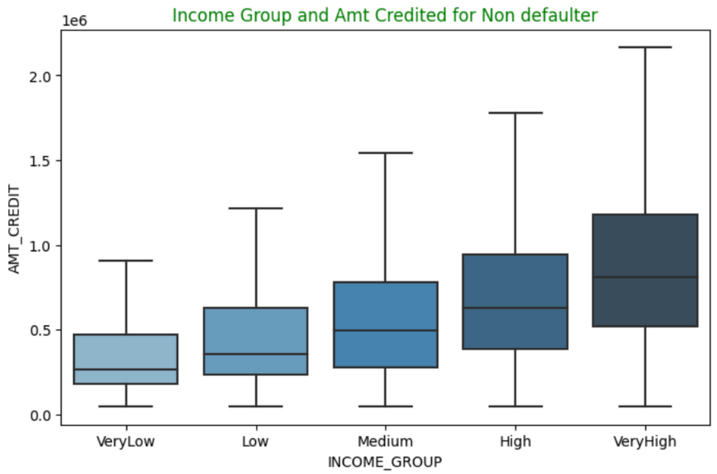


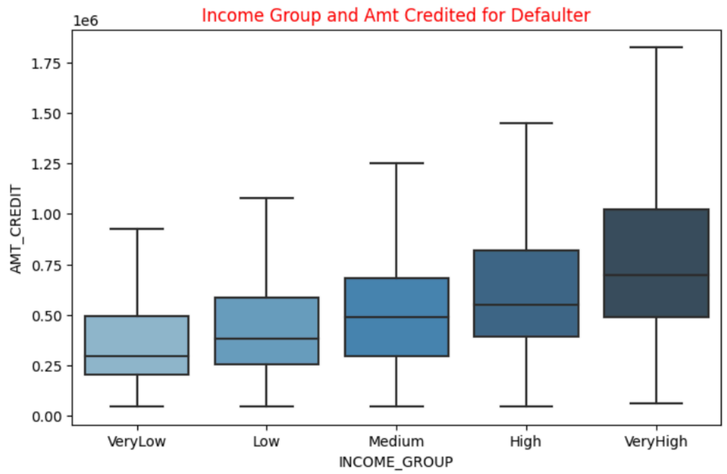
Key Observations:

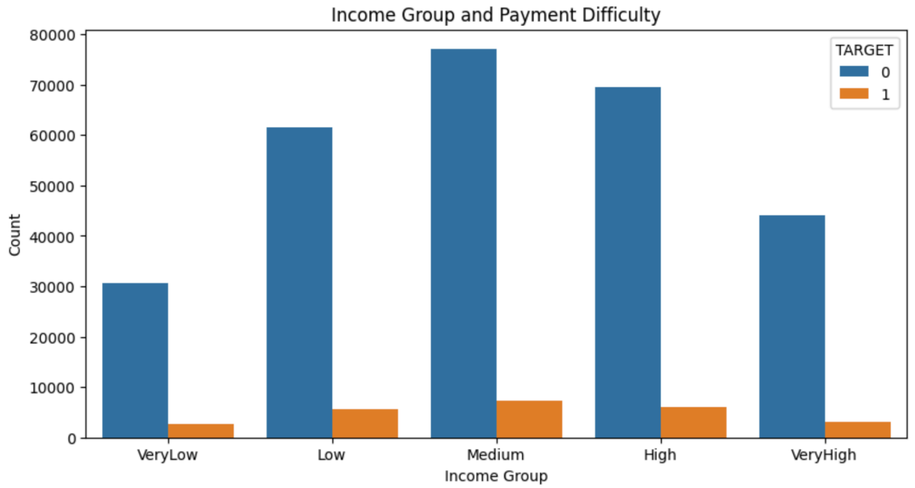
1. Age does seem like influencing default.
2. Medium income groups are most likely to be Non Defaulter.

Bivariate Analysis on Categorical and Continuous Variable

Income and Credit Category



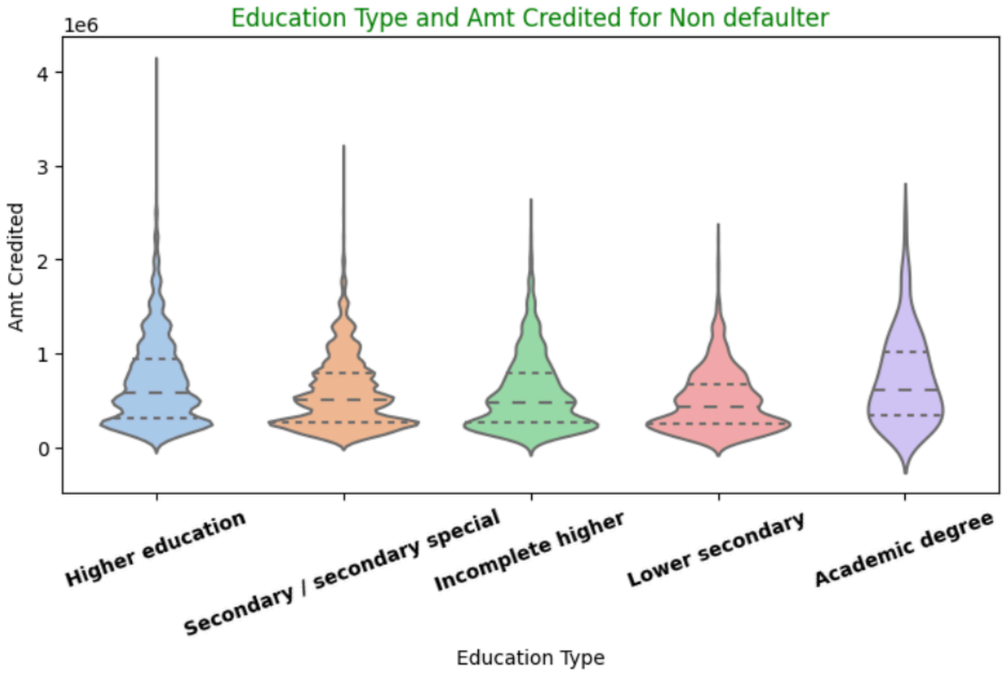


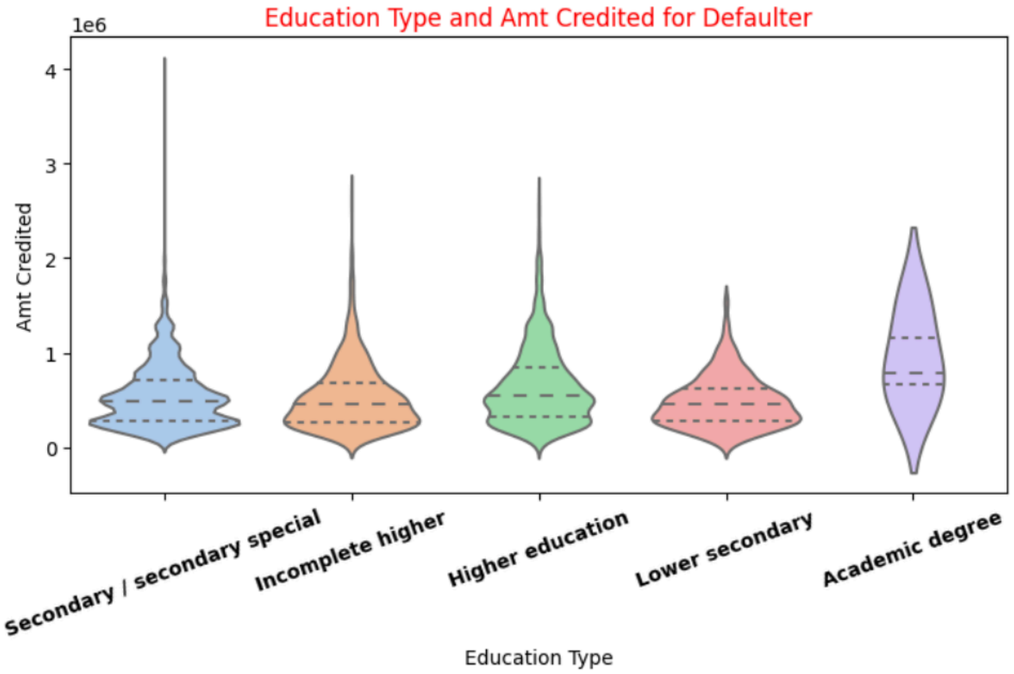


Key observations:

Based on the data, It's widely believed that a financial institution gives the most number of loans to the Medium-income group. The High-income group, on the other hand, has the highest default rate. This can have an impact on the loan book of the institution as a significant portion of the money may not be paid back.

Education category for both default and non-default

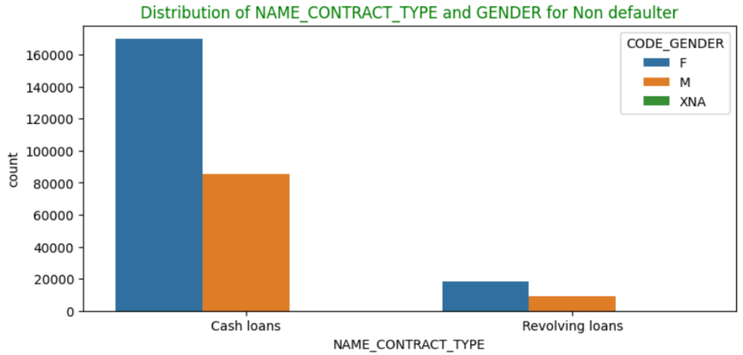


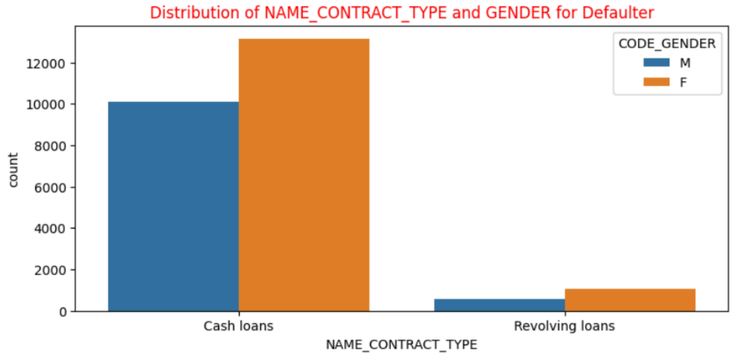


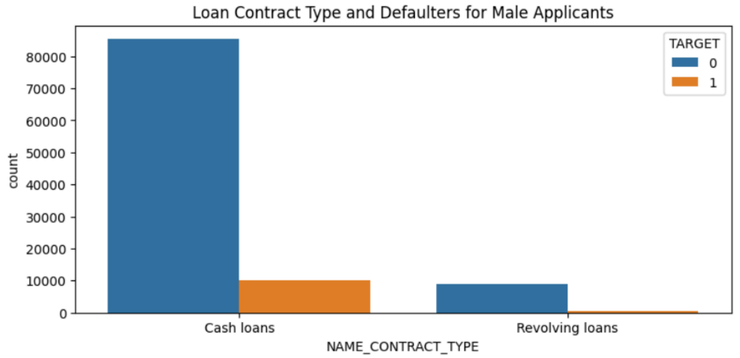
Key observations:
Higher loan values were observed for borrowers with an academic degree, who are also more prone to default. This finding is not conclusive, as the number of individuals with this degree is small, unlike other groups.

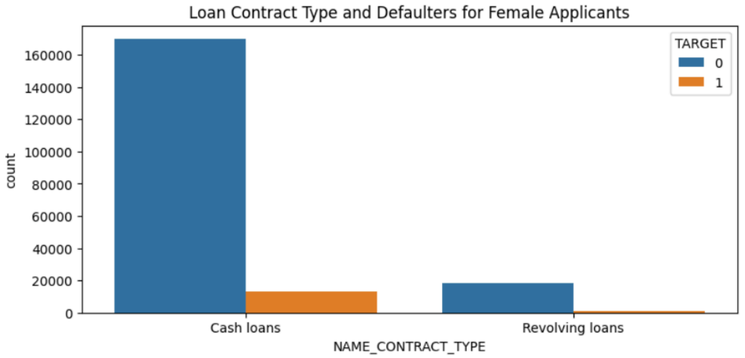
Bivariate categorical and categorical

Gender category for both default and non-default







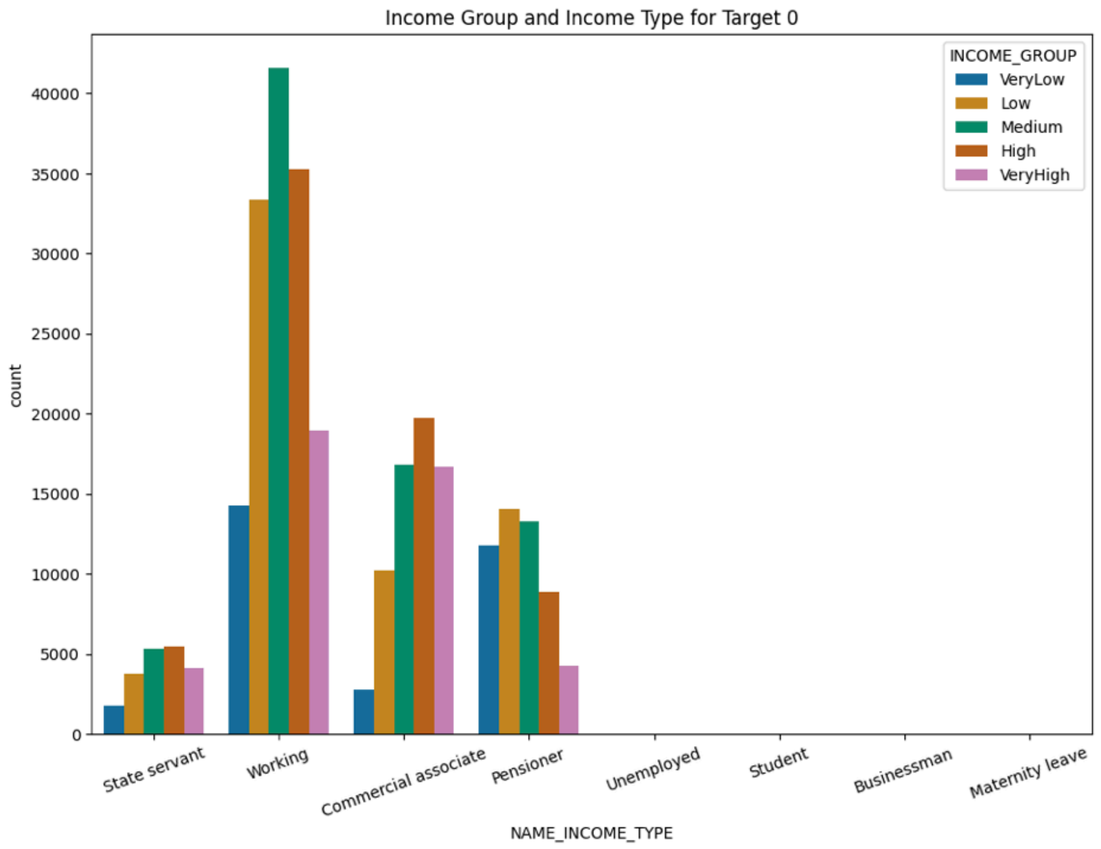


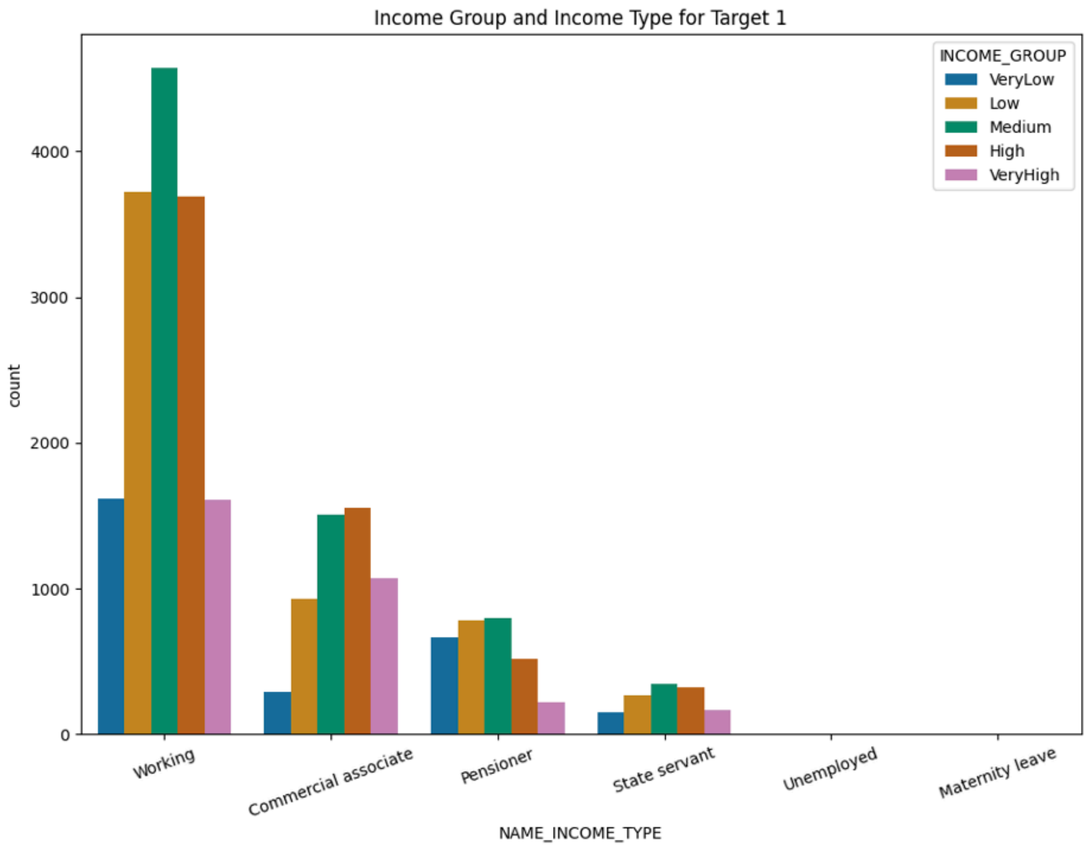
Key observations:
As noted above data has more females as loan applicant. Default rates are higher among male applicants compared to female applicants.

This dataset holds a higher number of females as evidenced in the previous study.

Although there were fewer male applicants, the default rate among men was higher than that of women.

Income Group and Type category for both default and non-default

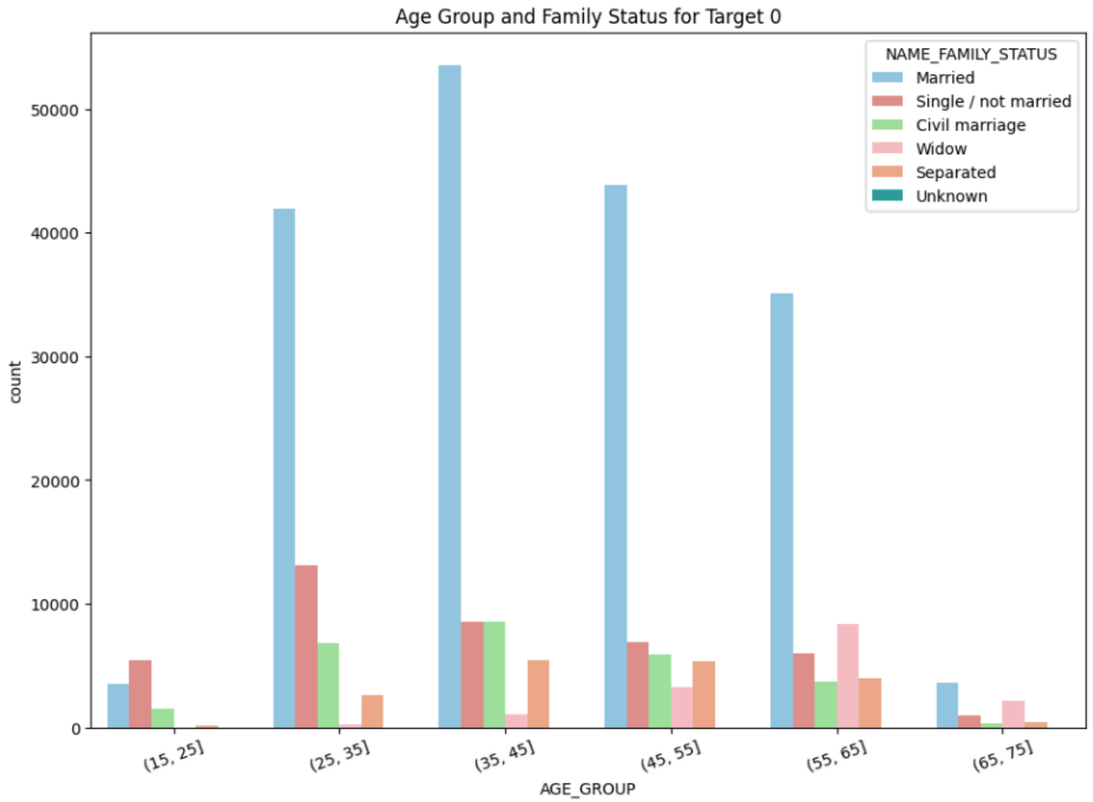


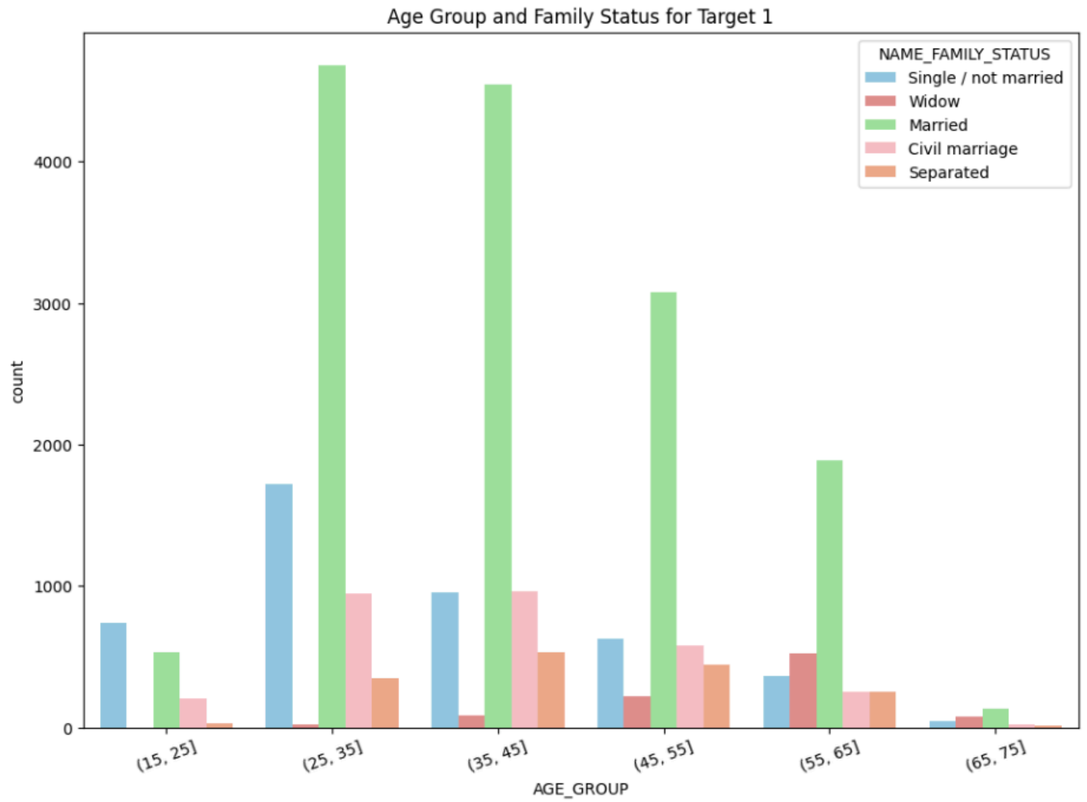


Key Observations:

The medium income group with a specific income type has a default rate of almost 1 in 12, which is higher than the average default rate of 1 in 11

Age Group and Family Status category for both default and non-default





Key Observations:

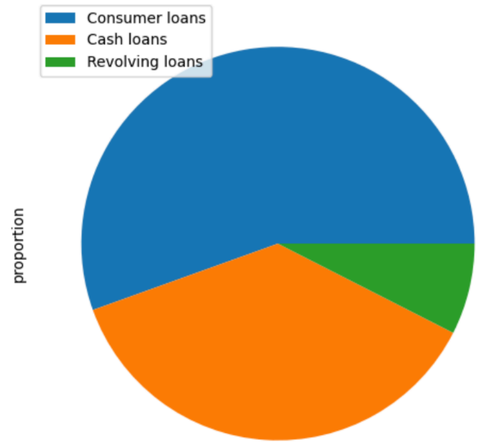
- The largest group of applicants facing payment difficulties falls under the age groups of 25-35 and 35-45.
- Within these age groups, married applicants are more likely to face payment difficulties.

Exploratory Data Analysis of Previous Application Data

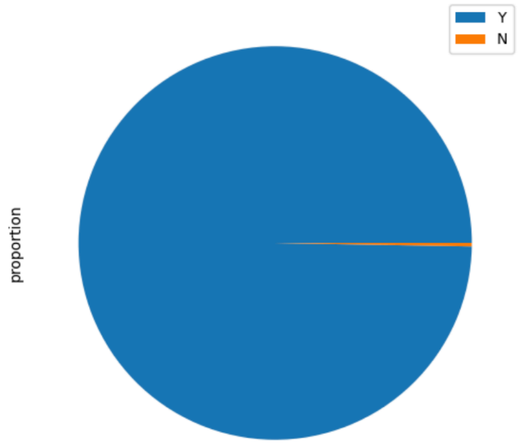
Univariate Analysis

Categorical Nominal

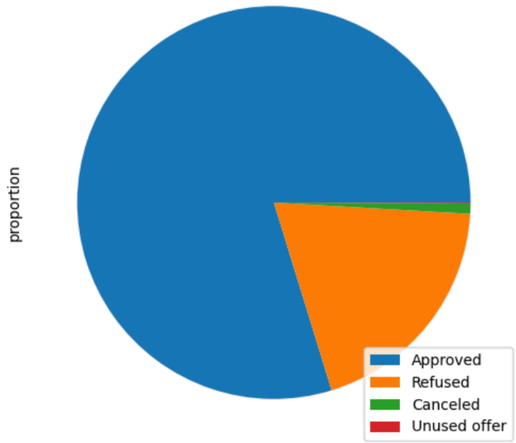
```
NAME_CONTRACT_TYPE
Consumer loans     0.554779
Cash loans         0.370341
Revolving loans    0.074880
Name: proportion, dtype: float64
```



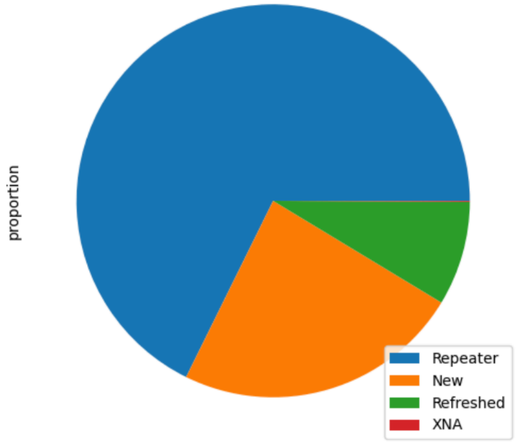

```
FLAG_LAST_APPL_PER_CONTRACT
Y    0.997222
N    0.002778
Name: proportion, dtype: float64
<Figure size 640x480 with 0 Axes>
```




```
NAME_CONTRACT_STATUS
Approved        0.797498
Refused         0.193344
Canceled        0.008427
Unused offer    0.000730
Name: proportion, dtype: float64
<Figure size 640x480 with 0 Axes>
```




```
NAME_CLIENT_TYPE
Repeater     0.677066
New          0.236334
Refreshed    0.085877
XNA          0.000723
Name: proportion, dtype: float64
<Figure size 640x480 with 0 Axes>
```



Key Points:

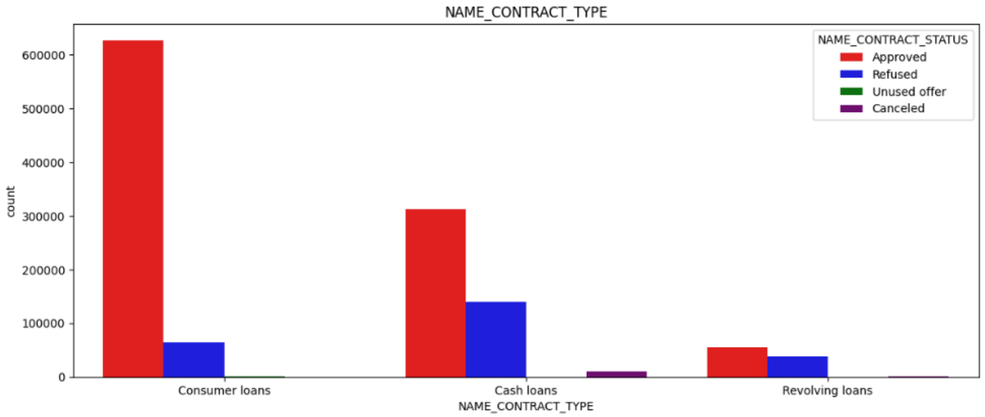
The data frame currently contains a consumer loan, which was not included in the application data frame. This type of loan accounts for 43 percent of all loans, while cash loans make up 44%. Revolving loans make up the remaining portion.

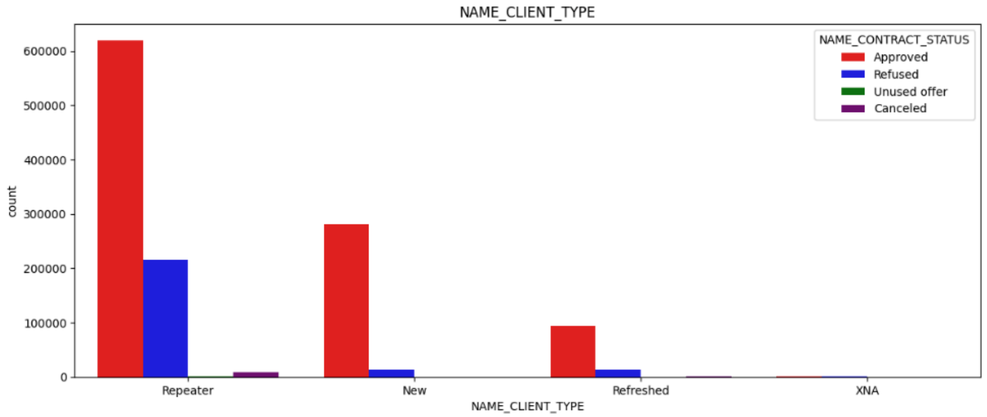
Around 79% of the loans were approved, while the remaining ones were either refused, canceled, or left unused, which indicates a discrepancy in the data.

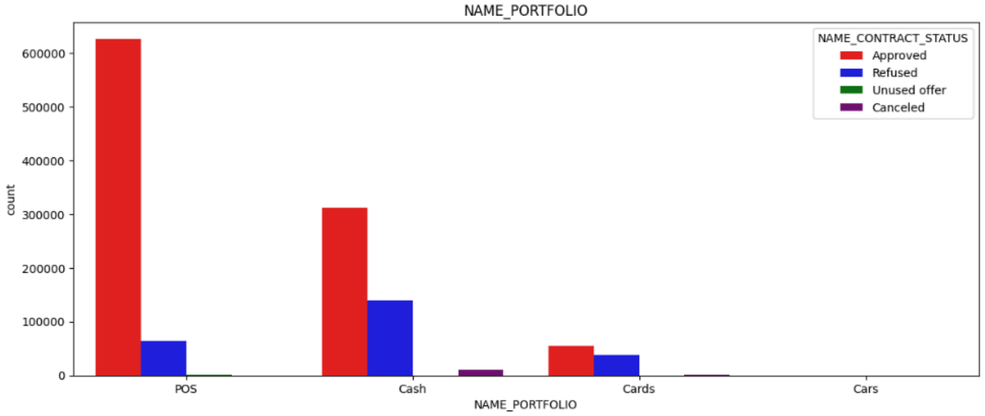
About 67% of the loan applicants are repeaters. The name_consumer_type column has null values.

Bivariate Analysis

Categorical Nominal



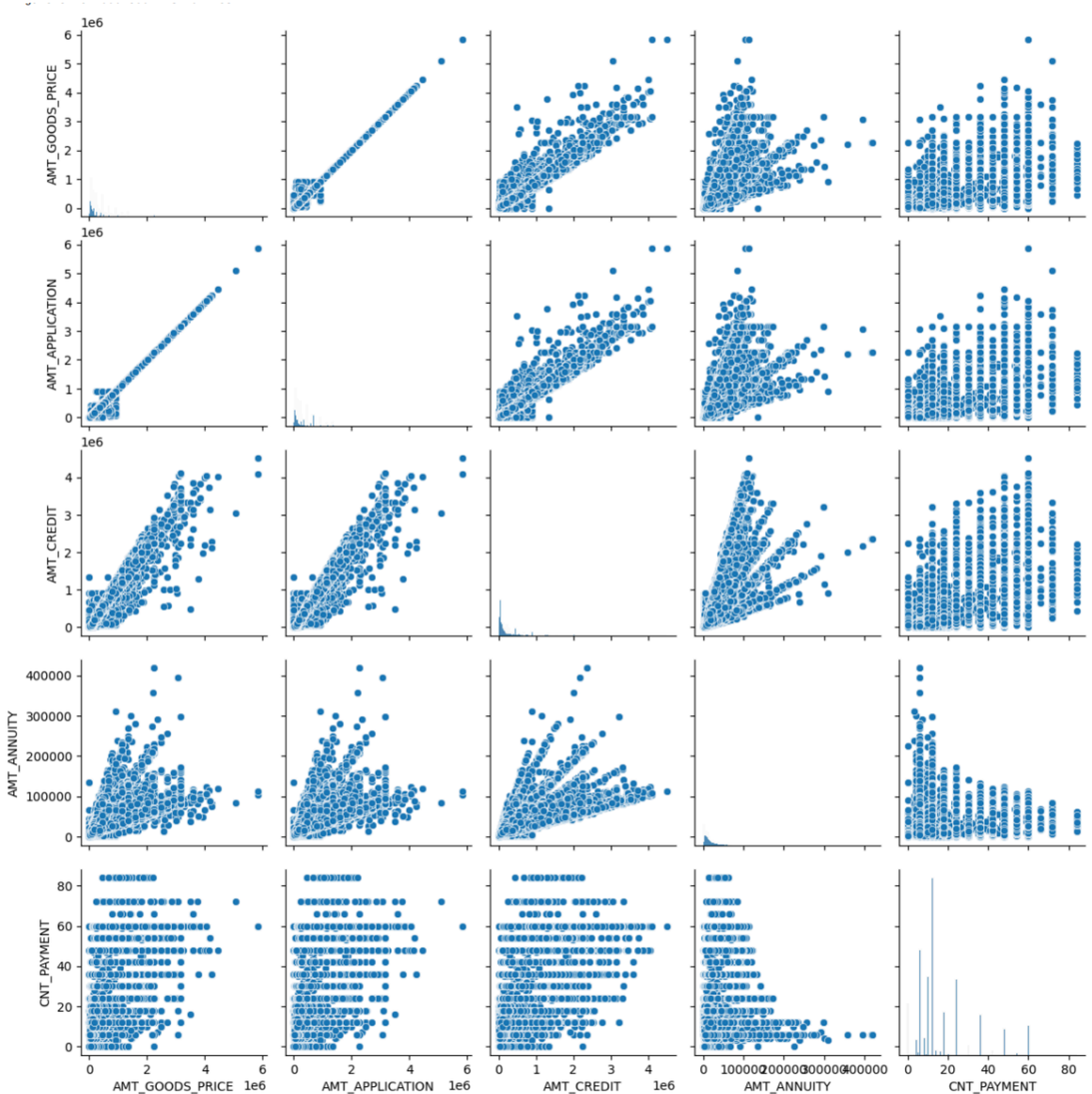




Key observation:
The number of people applying for consumer loans is higher than for cash loans. In contrast, cash loans do not have canceled loans as frequently as consumer loans. Also, the bank has a higher number of repeaters in each category, such as refused, canceled, and approved. POS transactions are mainly consumer loans, and cash advances have been rejected more often than those made using a POS device.

MultiVariate Analysis

Numerical Variable



Key Obervations based on the correlation analysis:
The various loan categories, such as AMT_ANNUITY, AMT_GOODS, and AMT_APPLICATION, exhibit a positive correlation. This is because higher prices of goods require larger loan amounts, and monthly payments. It's logical to expect that these variables would also positively correlate.

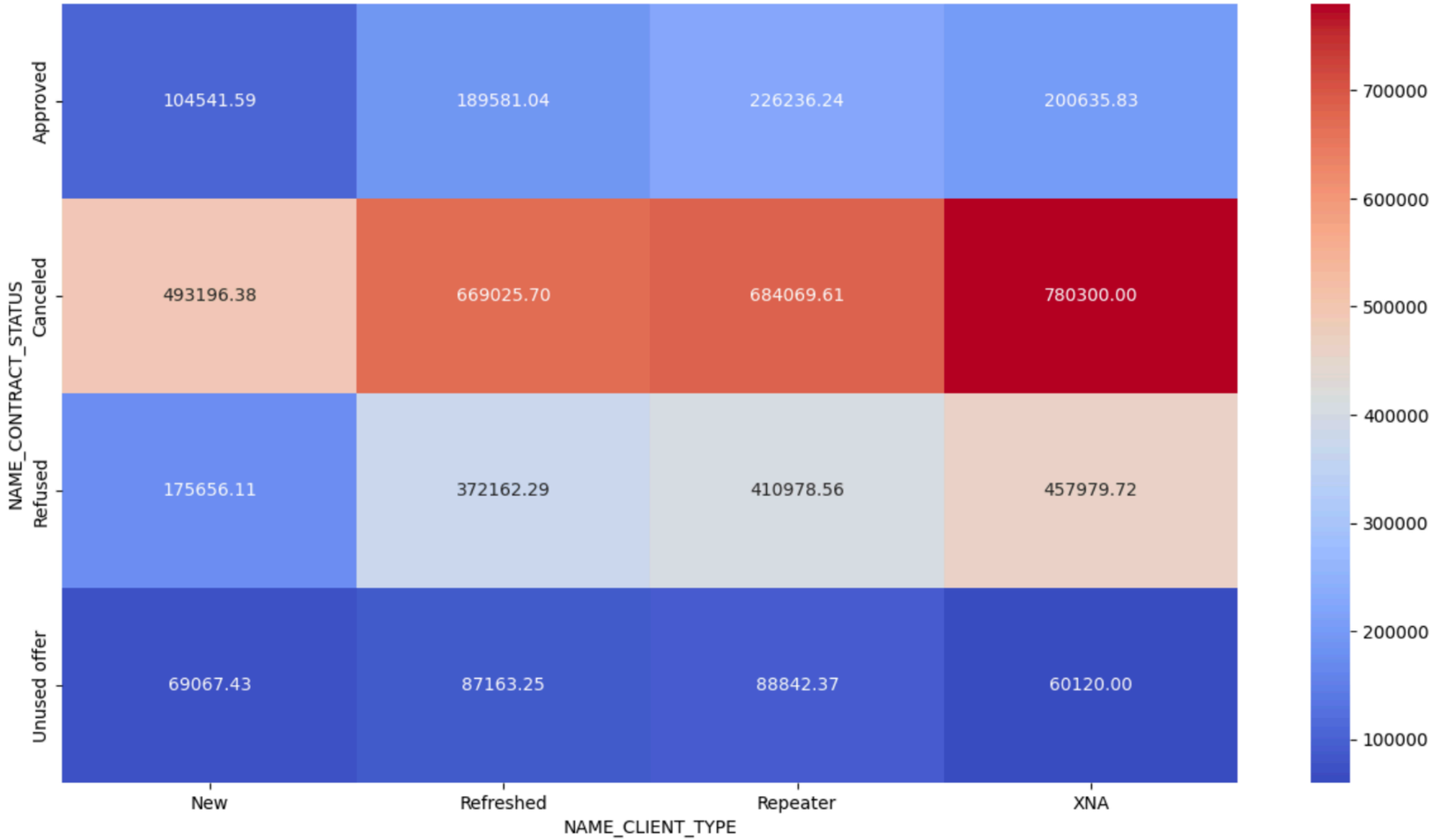
There's a strong correlation between the AMT_PRICE and AMT_CREDIT values. This suggests that the value of the products an individual is planning on purchasing directly correlates with the loan amount.

It is surprising, then, that there is a non-significant relationship between AMT_CREDIT and CNT_PAYMENTS. It was initially believed that as the loan term lengthens, the number of monthly payments would increase. However, this finding suggests that other factors may be affecting the payment terms.

The results of this analysis provide us with valuable information on the link between these variables and how they affect loan applications.

MultiVariate Analysis

Contract status vs name client type aggregating over application amount



Key points:

The low number of offer applications is a sign that many consumers are taking advantage of the loans that were offered.

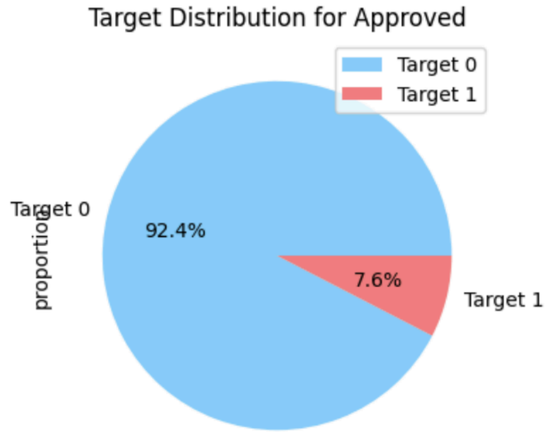
The number of loan applications that have been canceled is high. This could be a result of the bank refusing the loans due to the borrower's high debt-to-liability ratio.

The number of repeat customers who applied for a loan was higher than those who took out a new one. This suggests that the bank has a better policy or incentive program for these consumers, which encourages them to take out larger loans.

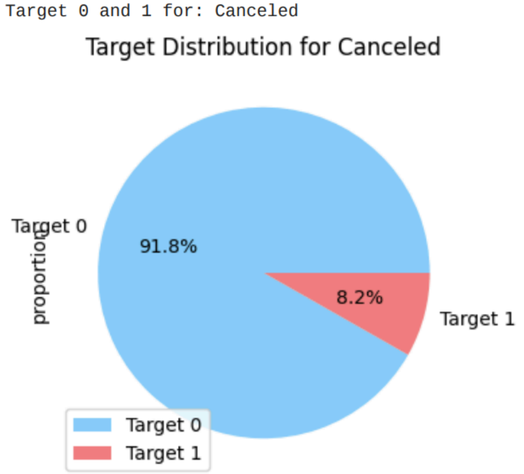
Merged Data frames Analysis

Univariate Analysis

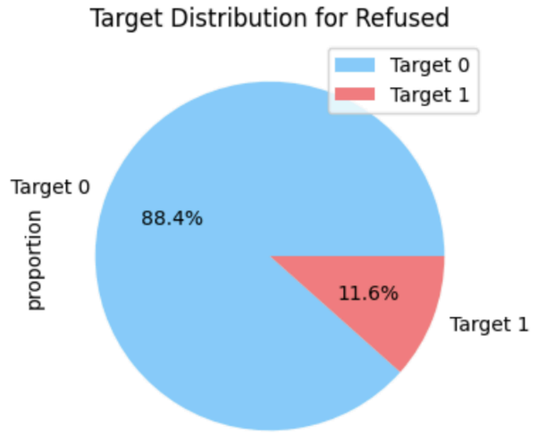
Target 0 and 1 for: Approved



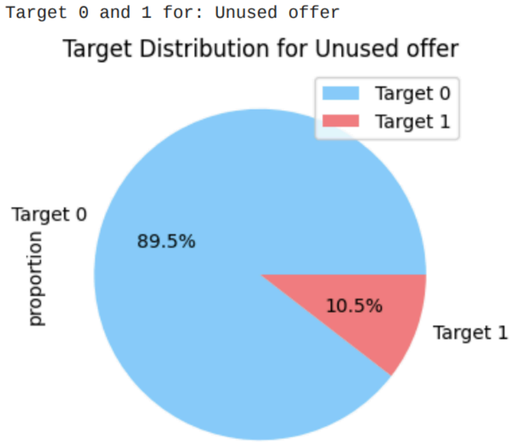
Target 0 and 1 for: Canceled



Target 0 and 1 for: Refused



Target 0 and 1 for: Unused offer

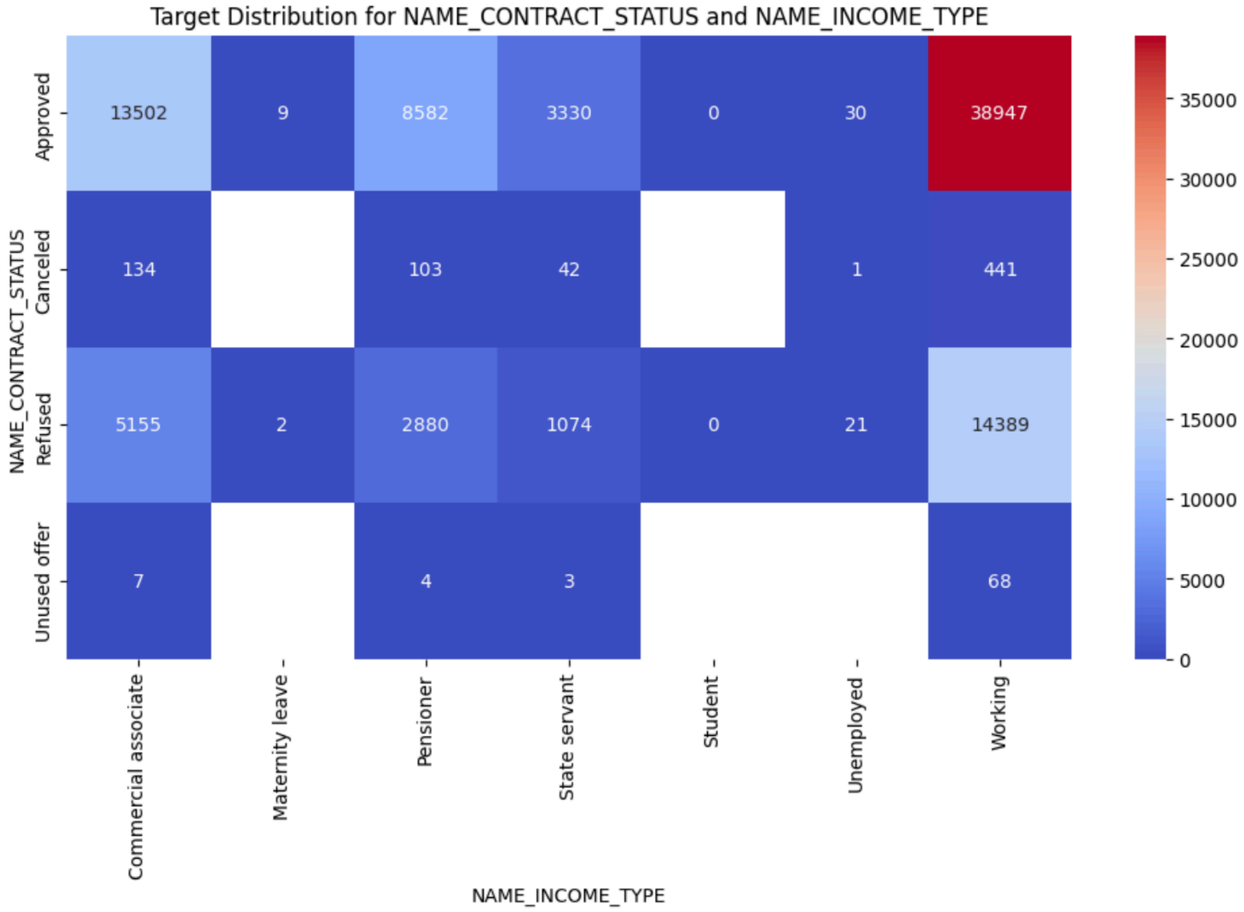


Key Observations:
Out of the loans that were approved, around 7.6% have resulted in default. This is a concerning figure.

The presence of past-due loans in the background of new applications is alarming.
It suggests that the bank's approval decisions may be leading to defaults.

MultiVariate Analysis

NAME_CONTRACT_STATUS, NAME_INCOME_TYPE,aggregating on Target



Key point:

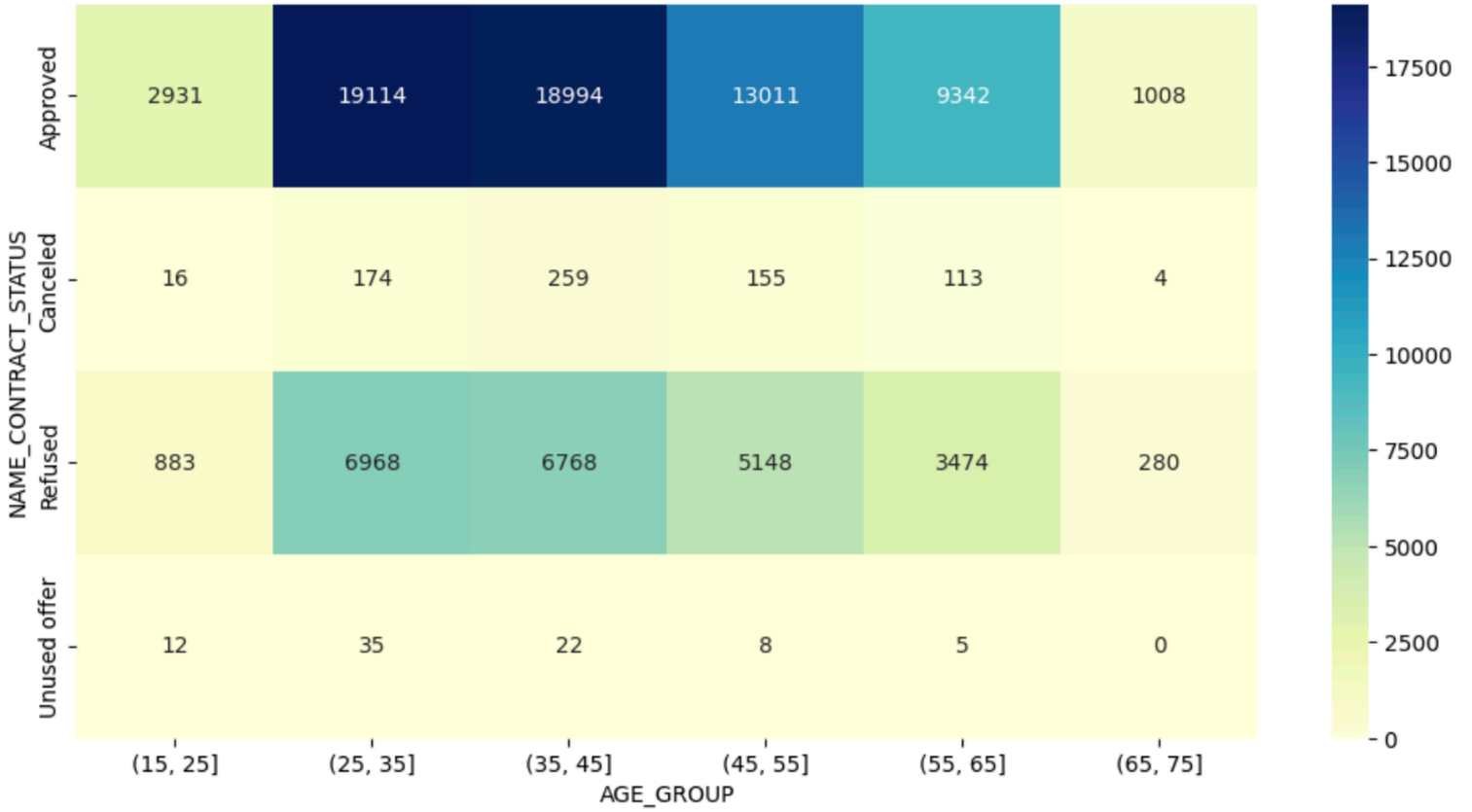
The values in the matrix above suggest a strong relationship between the number of defaults and the working income type of the applicant. For instance, people with an "Approved" contract status experienced higher default rates.

It's concerning to see that people who previously applied for a loan with the notations "Refused," "Cancelled," or "Unused" are experiencing default. This indicates that even though the financial institution rejected or canceled the previous application, it still approved the current one.

Among the borrowers whose previous applications were rejected, over 14,000 were working-income individuals who subsequently defaulted. This suggests that they have a credit risk.

By reviewing the heatmap, we may be able to identify possible risk factors related to the default behavior of borrowers.

"NAME_CONTRACT_STATUS", "AGE_GROUP", aggregating on Target



Key points:
Values in the matrix above indicate a strong correlation with the number of cases that have resulted in default, which is Target 1.

The default rates among individuals aged 25 to 35 and those aged 35 to 45 with approved loans were higher. This suggests a concerning trend.

Individuals who had previously applied with the notations “Refused,” “Cancelled,” or “Unused” have defaulted on their current loans, which suggests that the financial institution might have a risk with these types of applicants.

The heatmap offers valuable information on the correlation between the target variable "TARGET" and the various attributes of applicants, allowing us to identify distinct patterns of default.

Case Summary

Defaulters' Demography and Other Important Factors

The data collected from the heatmap revealed that there were distinct patterns of default in the approved applications.

Several factors were identified as possible risk factors that could lead to a borrower's default. These variables were then cross-referenced with the default cases and approved applications to confirm the findings. It was found that the high rate of defaults among the approved applications was due to the various factors that affected the applicant's credit history.

A. Individuals with a medium income level were more prone to experiencing default.
B. Individuals who are working are more prone to experiencing default.
C. Business Type 3 applicants are more prone to experiencing delinquency.
D. About 70% of applicants who don't own their homes are more prone to default.
E. The results of this study suggest that the various factors that affect an individual's credit risk can be used to predict a borrower's likelihood of experiencing default even after they have been approved.

Other important factors to be considered when evaluating loan applications include:

A. The lower the number of days that elapsed following the last phone call, the more concerns it raises about applicant stability.
B. The number of bureau hits in the last month or week is favorable, which indicates that there has been a decrease in credit inquiries.
C. Income-Goods Price Disparity can affect repayment capacity.
D. People who have previously applied for a loan with a notation labeled "Refused," "Unused," or "Cancelled" are more likely to default.
E. This suggests that financial institutions may have to take a closer look at these types of cases since they are still subject to default risks.

Credible Applications Refused

A. Unused applications are rejected as they may not meet the sanctioned amount. It is not clear why this happens, though further investigation is necessary.
B. Women applicants may be given a higher weightage in the assessment procedure as their default rate is lower. Although 60% of borrowers who have defaulted are working, this doesn't mean that every applicant should be rejected. It's important to scrutinize other factors such as income and credit history before making a final decision.
C. Some cases where the previous application was rejected or marked as "refused," "unused," or "cancelled" show that the current one has been paid on time. This raises questions about the decision-making process in the past, requiring further investigation.

THANK YOU