CREDIT EDA CASE STUDY

NITINKUMAR SHARMA
DS C58
IIIT Banglore

Project Introduction

Problem Statement and Analysis Approach:

- Clearly state the problem statement, which is to identify patterns that indicate if a client has difficulty paying their installments.
- Use exploratory data analysis (EDA) techniques to achieve this goal.

Data Preprocessing:

- Identify missing data in the dataset and determine the appropriate method for dealing with it. This could involve removing columns with excessive missing data or imputing missing values with appropriate methods. Clearly, mention the approach taken.

Outlier Detection:

- Identify outliers in the dataset and provide a rationale for why each data point is considered an outlier. Note that you are not required to remove outliers; this step is for detection and understanding.

Data Imbalance:

- Check for data imbalance in the target variable ("Client with payment difficulties" vs. "All other cases").
- Calculate the ratio of data imbalance.
- Use a mix of univariate and bivariate analysis to explore the data imbalance, which might include plotting in terms of percentage or absolute value.

Univariate and Segmented Univariate Analysis:

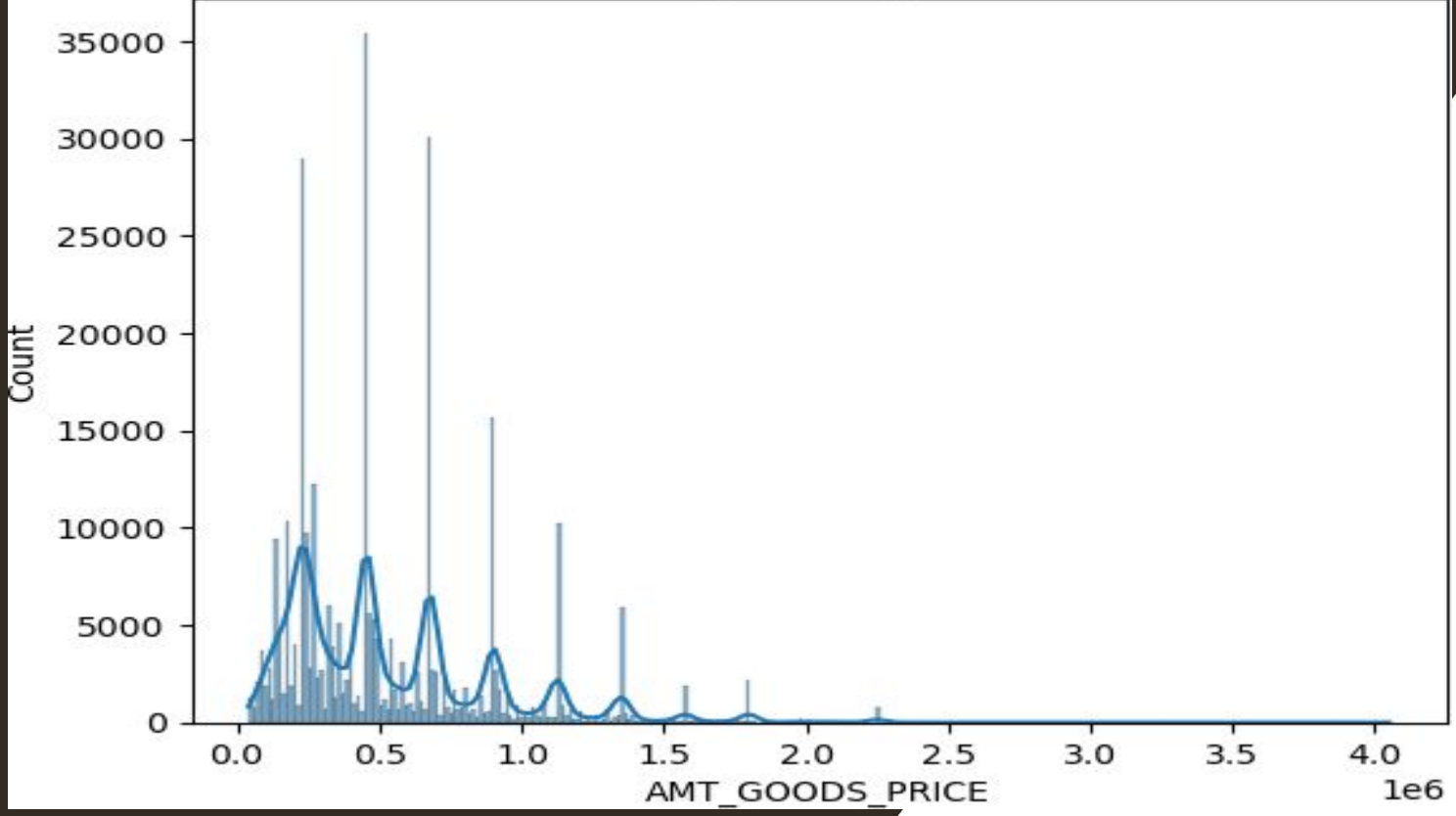
- Perform univariate analysis on relevant columns to understand their distributions.
- Use segmented univariate analysis to compare distributions between clients with payment difficulties and all other cases.
- Explain the results of these analyses in business terms.

Bivariate Analysis:

- Conduct bivariate analysis to explore relationships between variables.
- Explain the insights gained from bivariate analysis in the context of loan default risk.

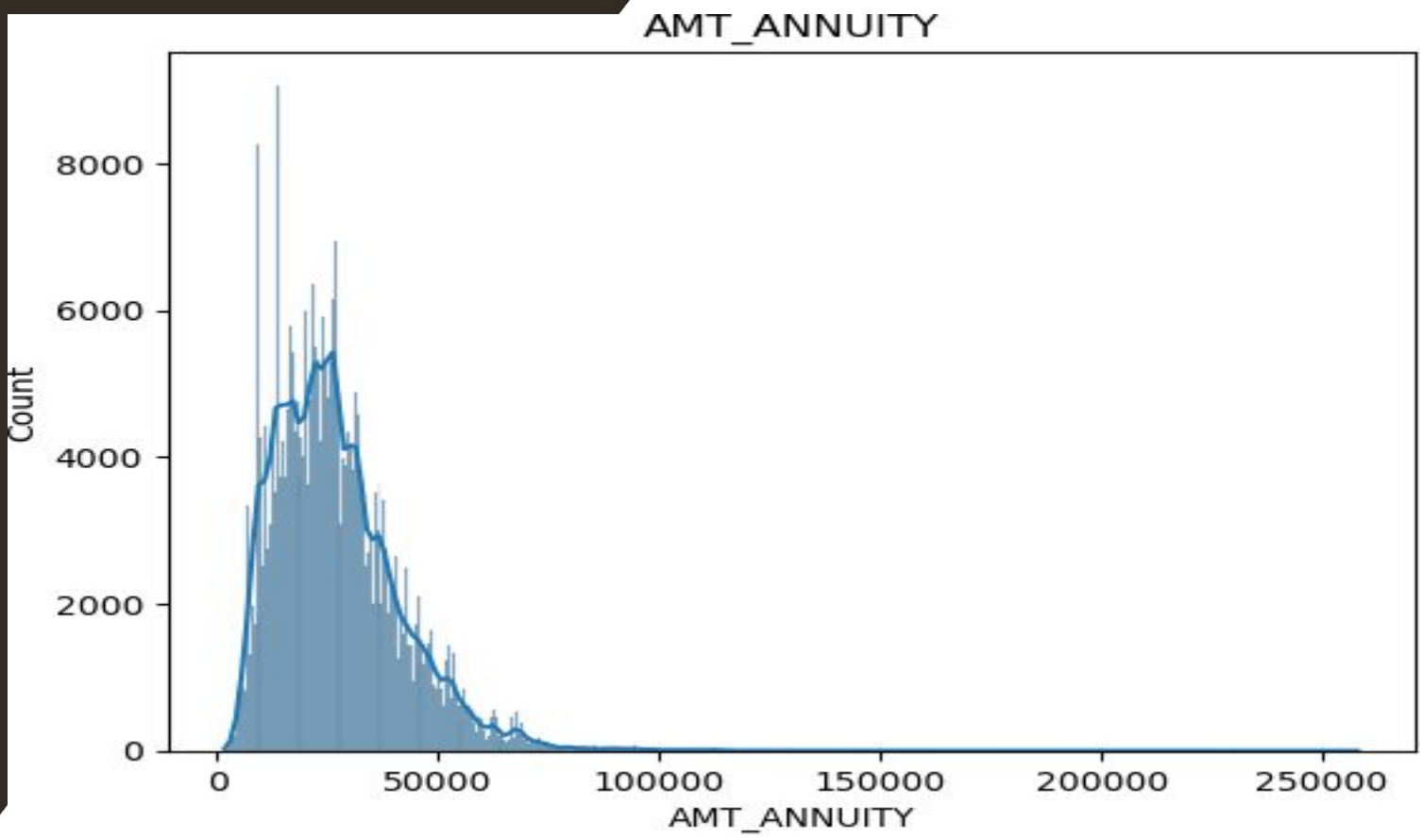
Correlation Analysis:

- Calculate and present the top correlations for both "Client with payment difficulties" and "All other cases" segments.
- Explain the significance of these correlations and how they relate to the target variable.



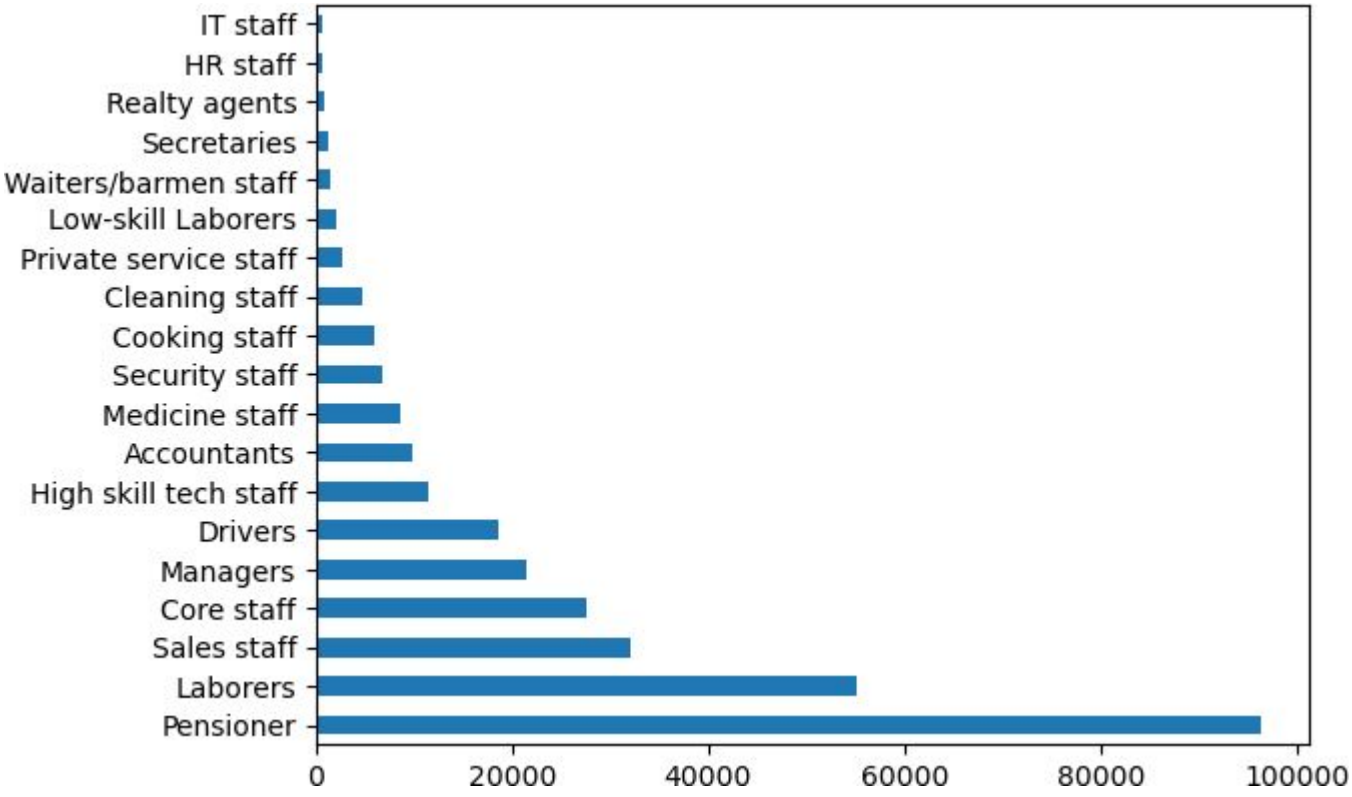
1: Majority of data had Skewness either it was positive or negatively skewed. Hence replaced those missing values with Mean, Median or Mode

2: At some places it was replaced by correlation factor



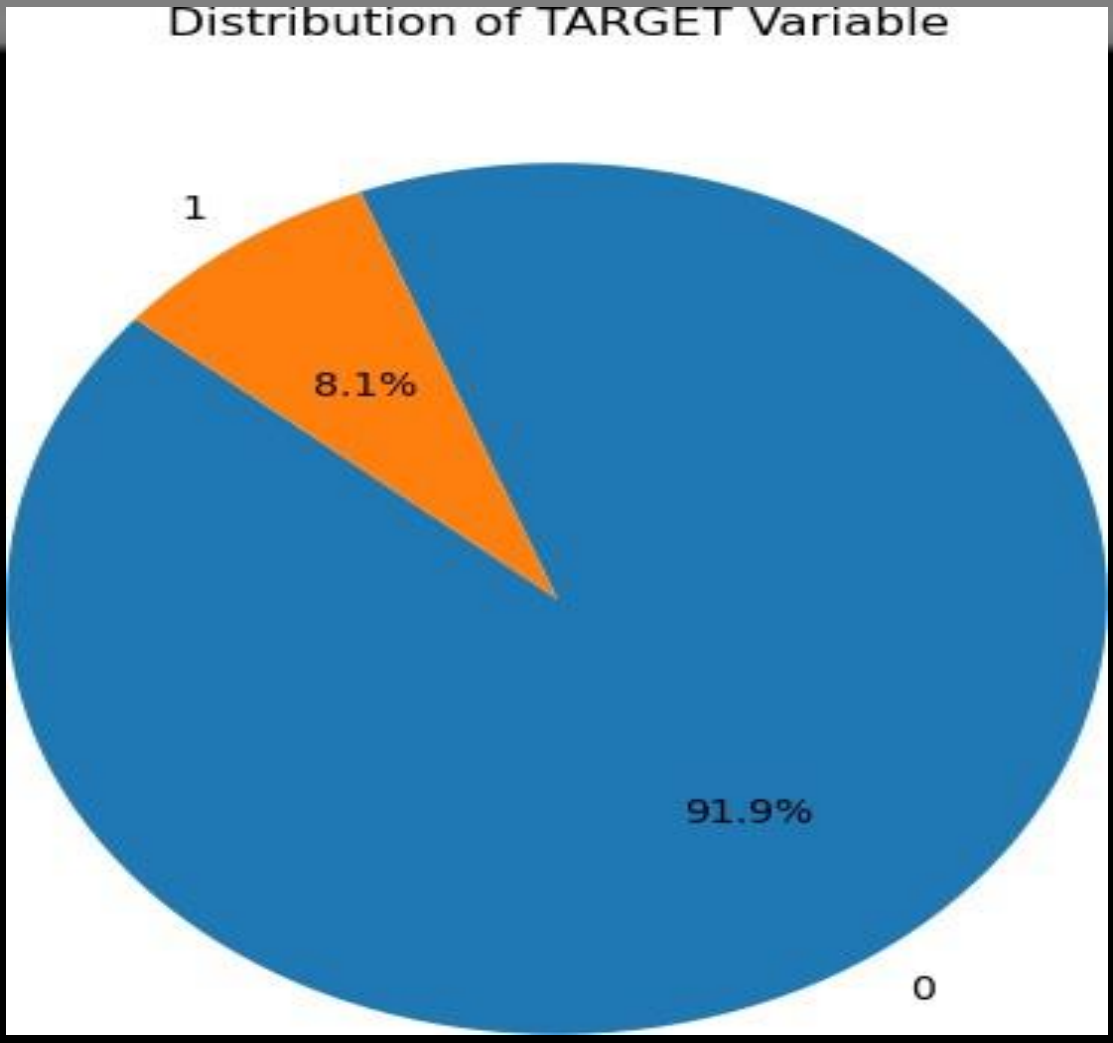
Assumption:

- In the "ORGANIZATION_TYPE" column, null values are closely associated with the "Pensioner" category, indicating that these null values are likely "Missing At Random."
- When analyzing the null values in the "OCCUPATION_TYPE" column, it is evident that the majority of these missing values are related to individuals classified as "Pensioners" under the "Income Type" variable.
- Given this strong correlation, imputing the null values in the "OCCUPATION_TYPE" column with the "Pensioner" category is a reasonable data handling strategy based on observed data patterns.




```
In [1985]: df[days_column]=abs(df[days_column])

In [1986]: df[days_column]=round(df[days_column]/ 365,0)

In [1987]: df[days_column]= df[days_column].astype(int)

In [1988]: df.DAYS_EMPLOYED.value_counts().head()

Out[1988]: 1001    55374
           1       34796
           2       30530
           3       27721
           4       22933
           Name: DAYS_EMPLOYED, dtype: int64
```

Above value represents pensioners. ie 1001. Considering there 40 years of work from age 20 to 60. Replacing them by 45 MAX.



In risk analytics, specifically in the context of credit risk assessment:

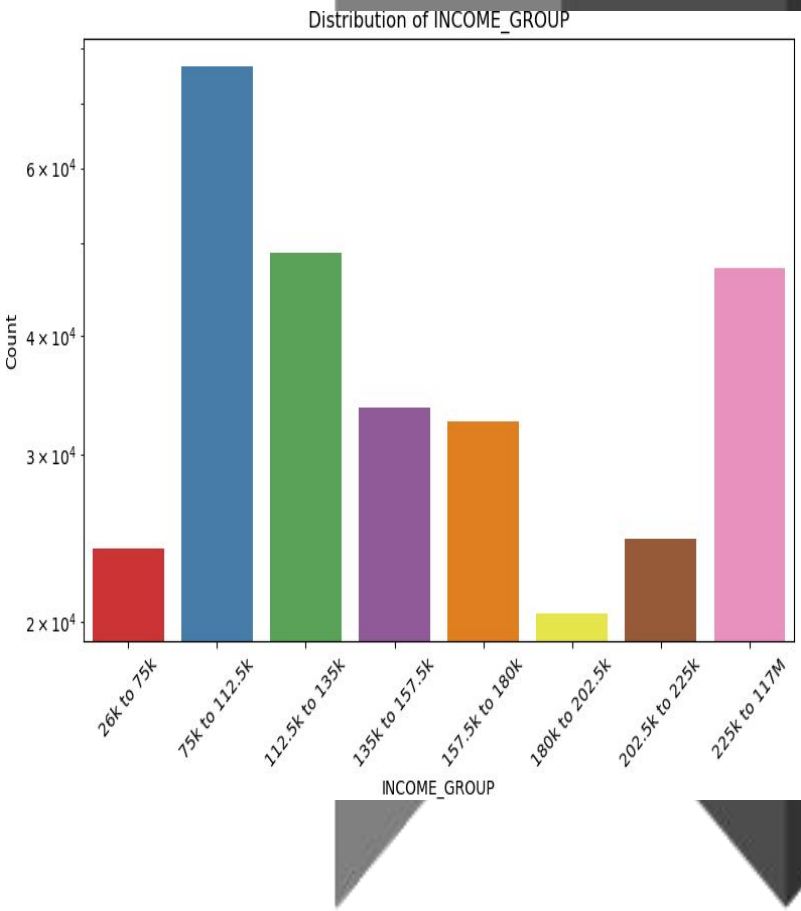
1 typically represents customers who are classified as "high risk" or those who at higher risk of getting loan default.

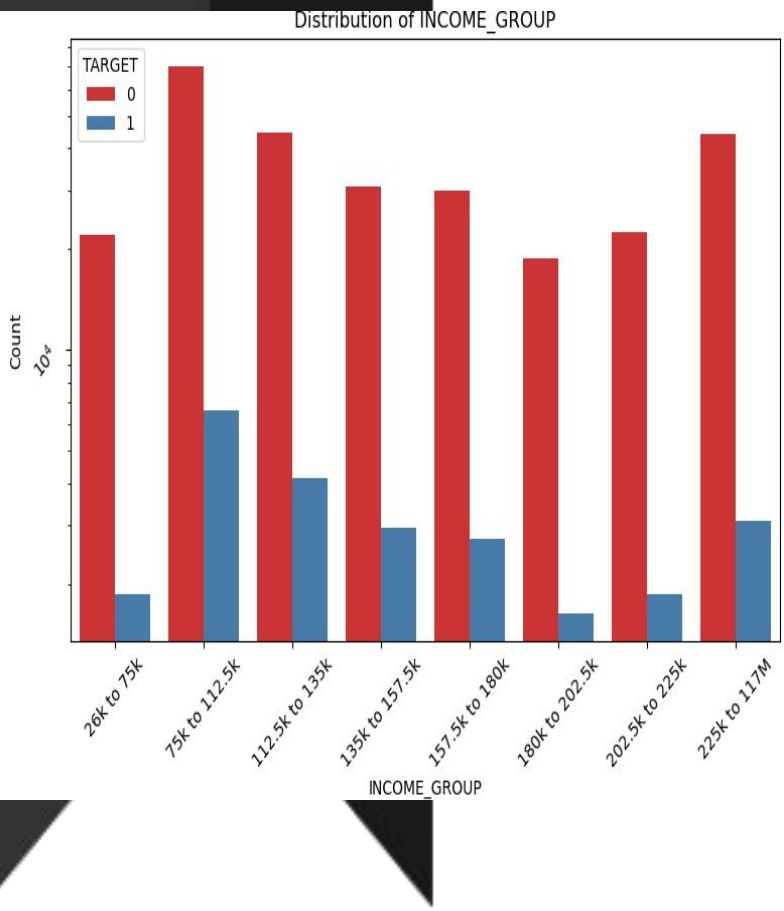
0 typically represents customers who are classified as "low risk" or those who have a lower risk of defaulting

These values are commonly used to create a binary classification of customers based on their creditworthiness or the likelihood of defaulting on a loan.

Imbalance Ratio with respect to Repayer(0) and Defaulter(1) is given: 11.345679

Univariate Analysis for 'AMT_INCOME_TOTAL':





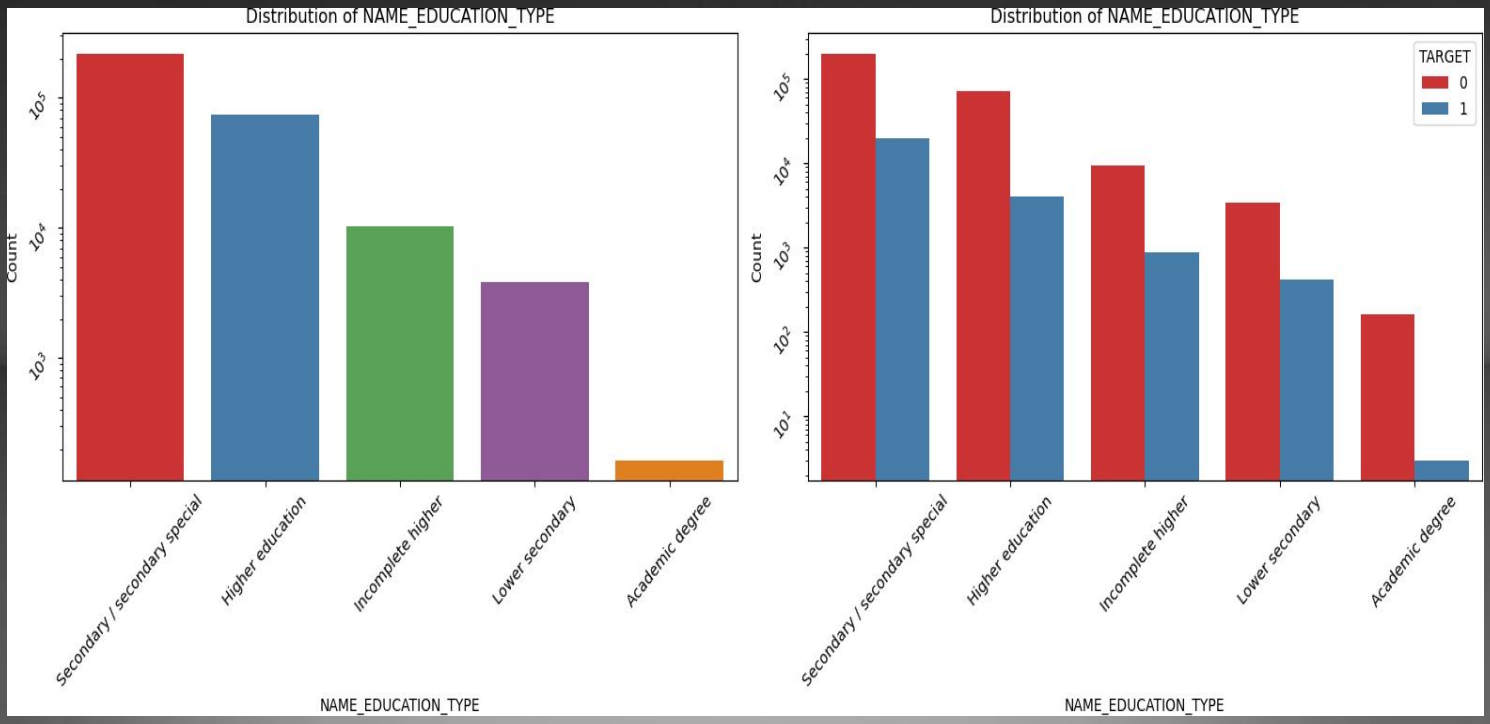
1: Majority of people have income between 75k to 112.5k and highest defaulters also lies in this region

2: People belonging to this category has highest defaulters.

3: People Belonging to category of 180k to 225k have less defaulter.

4: Bank should come up with various schemes to lend money to this group of people.

Univariate Analysis for 'NAME_EDUCATION_TYPE':



Analysis of Education Types and Default Rates

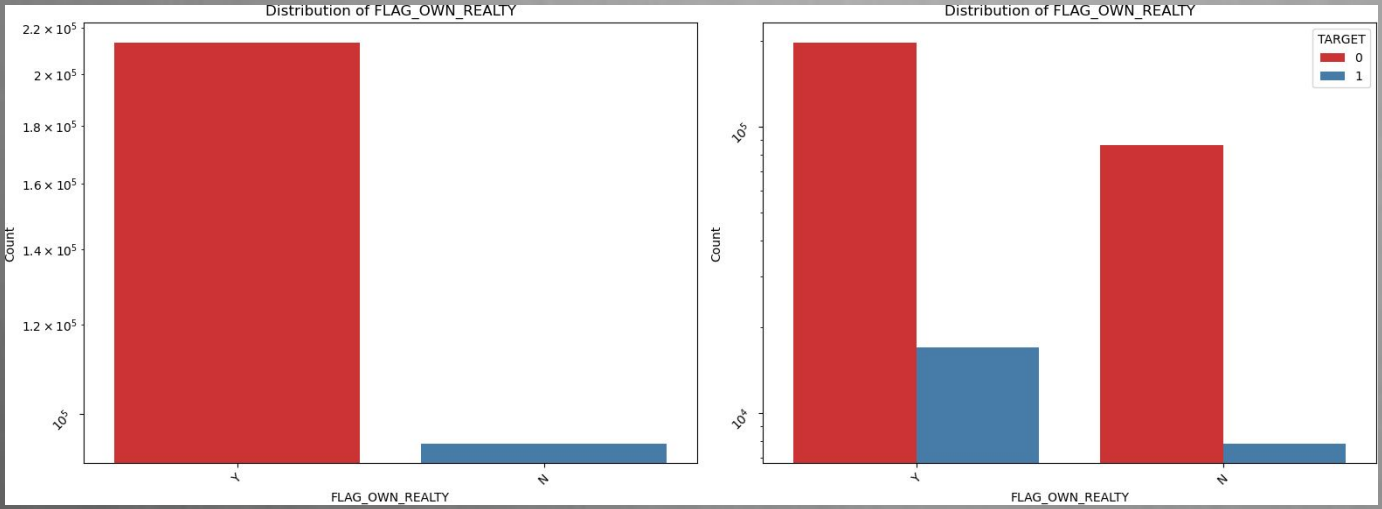
The majority of clients in the dataset have completed Secondary/secondary special education. Following closely, the second most common education level among clients is Higher education. Academic degree holders make up a relatively small portion of the dataset.

Default Rates by Education Type:

Among these education groups, clients with Lower secondary education have the highest default rate. In contrast, clients with Academic degrees have the lowest default rates.

It's worth noting that individuals with Academic degrees are the least likely to face issues with loan defaulting.

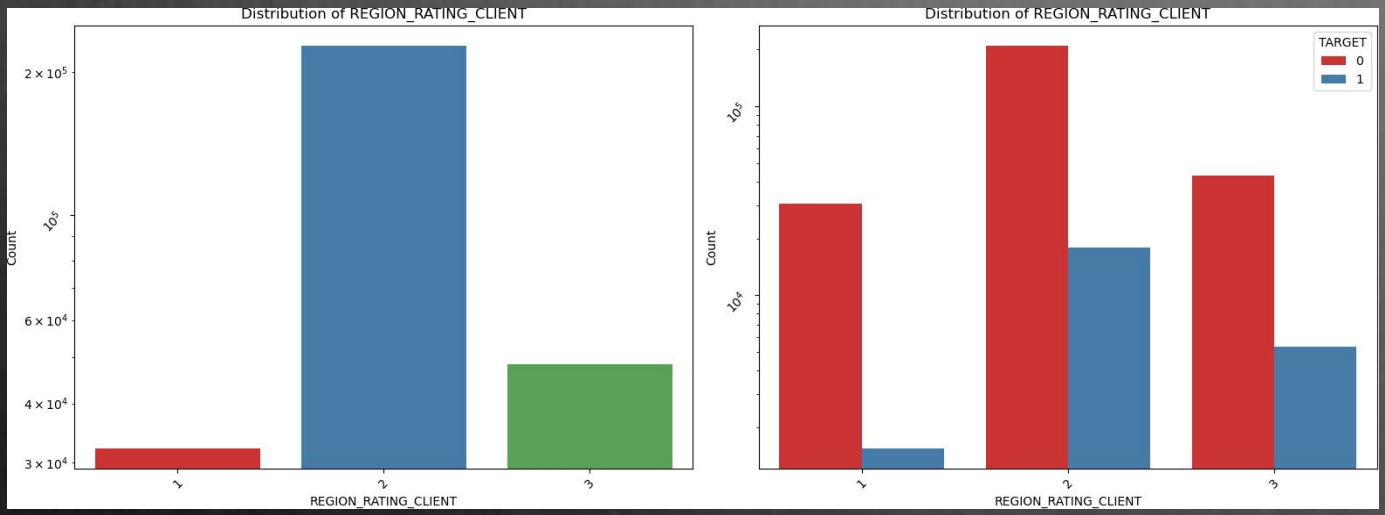
Univariate Analysis for "FLAG_OWN_REALTY":



Observations on Ownership of Real Estate and Loan Default Rates

- The majority of clients in the dataset own real estate.
- Clients who own real estate are more than double in number compared to those who don't own.
- People who won Real Estate are higher defaulters

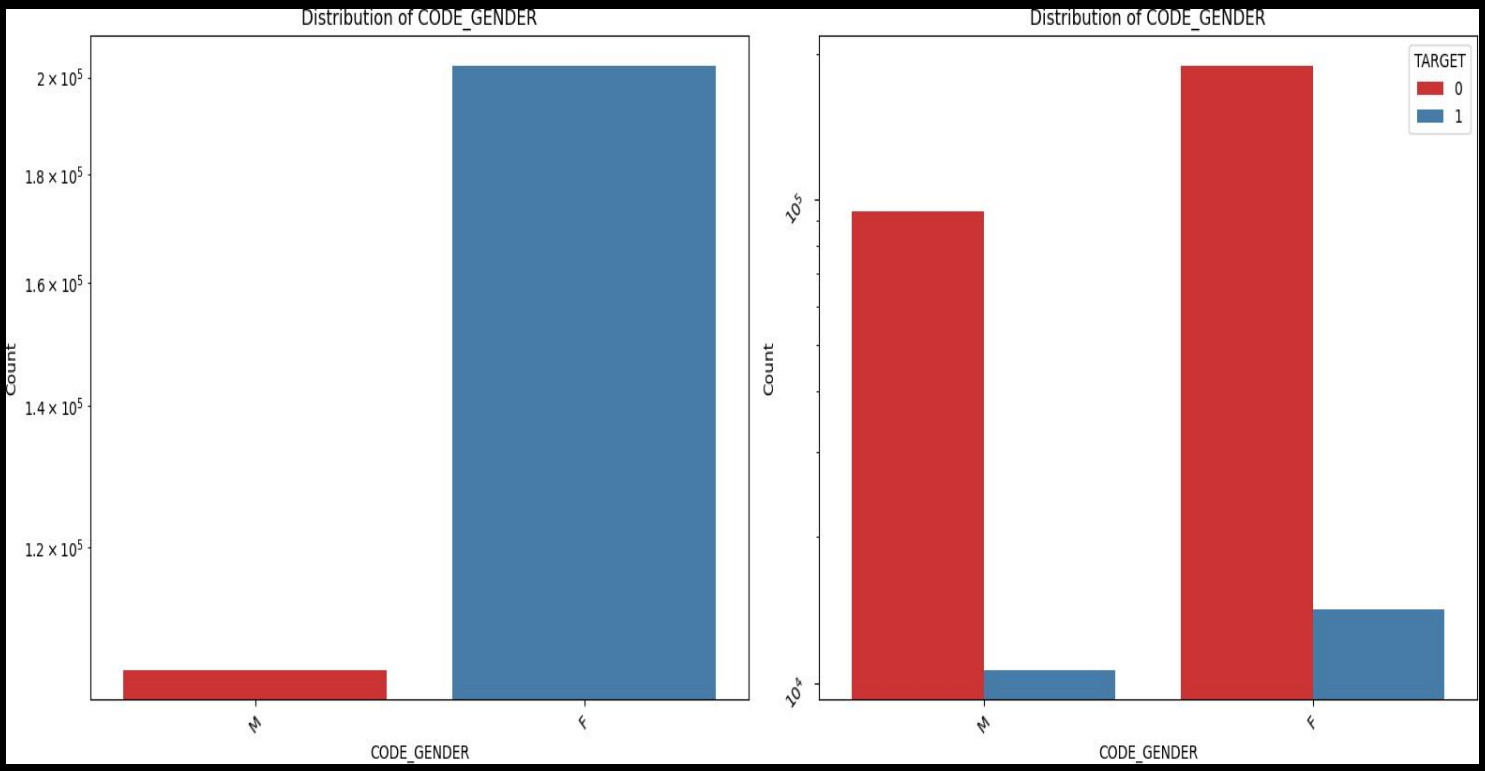
Univariate Analysis for "REGION_RATING_CLIENT":



Client Region Rating Insights

- Most applicants are concentrated in Region Rating 2.
- Region Rating 1 is associated with the lowest default rate, making it a safer choice for approving loans.

Univariate Analysis for "CODE_GENDER":



When we look at the data, interesting insights emerge regarding the gender of clients applying for loans:

1. **Gender Distribution:**

   The dataset comprises almost double the number of female clients compared to male clients. It suggests that there are more female applicants seeking loans.

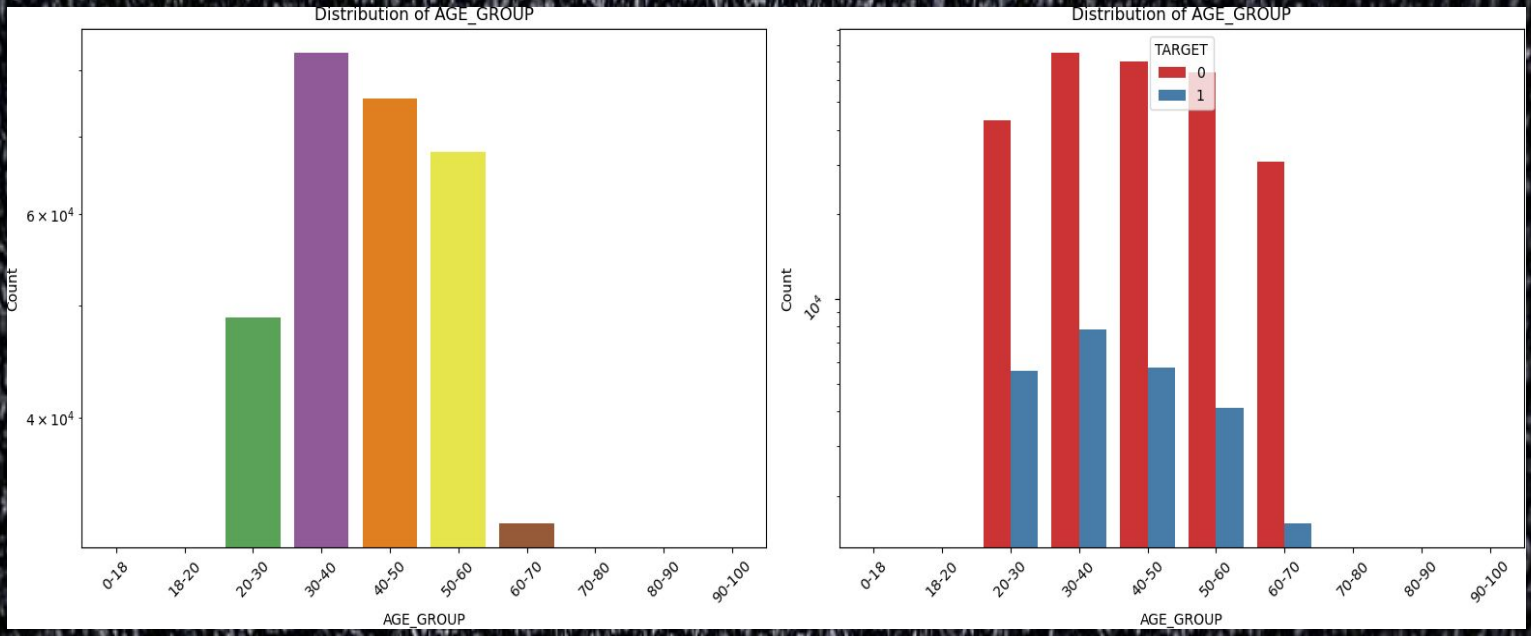
2. **Loan Default Rates:**

   Digging deeper into loan default rates, we notice a distinction based on gender:

   Male clients have a higher likelihood of not repaying their loans, with a default rate. In contrast, female clients exhibit a comparatively lower default rate.

These findings imply that gender might play a role in loan default tendencies, with males showing a slightly higher risk of not returning their loans.

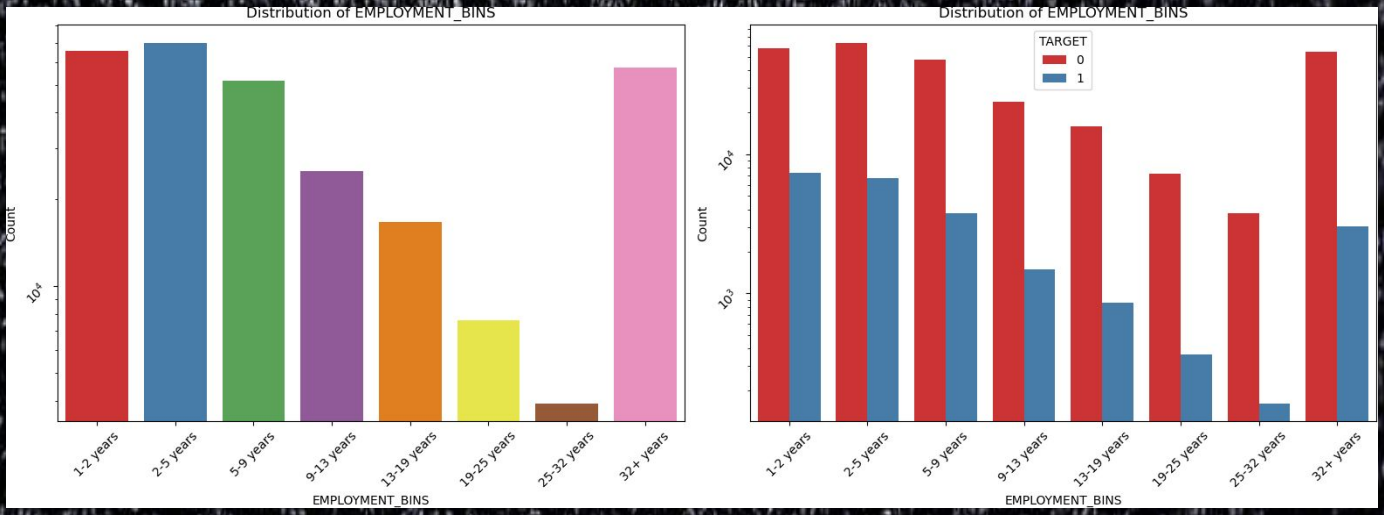
Univariate Analysis for "AGE_GROUP":



Age Rating Insights

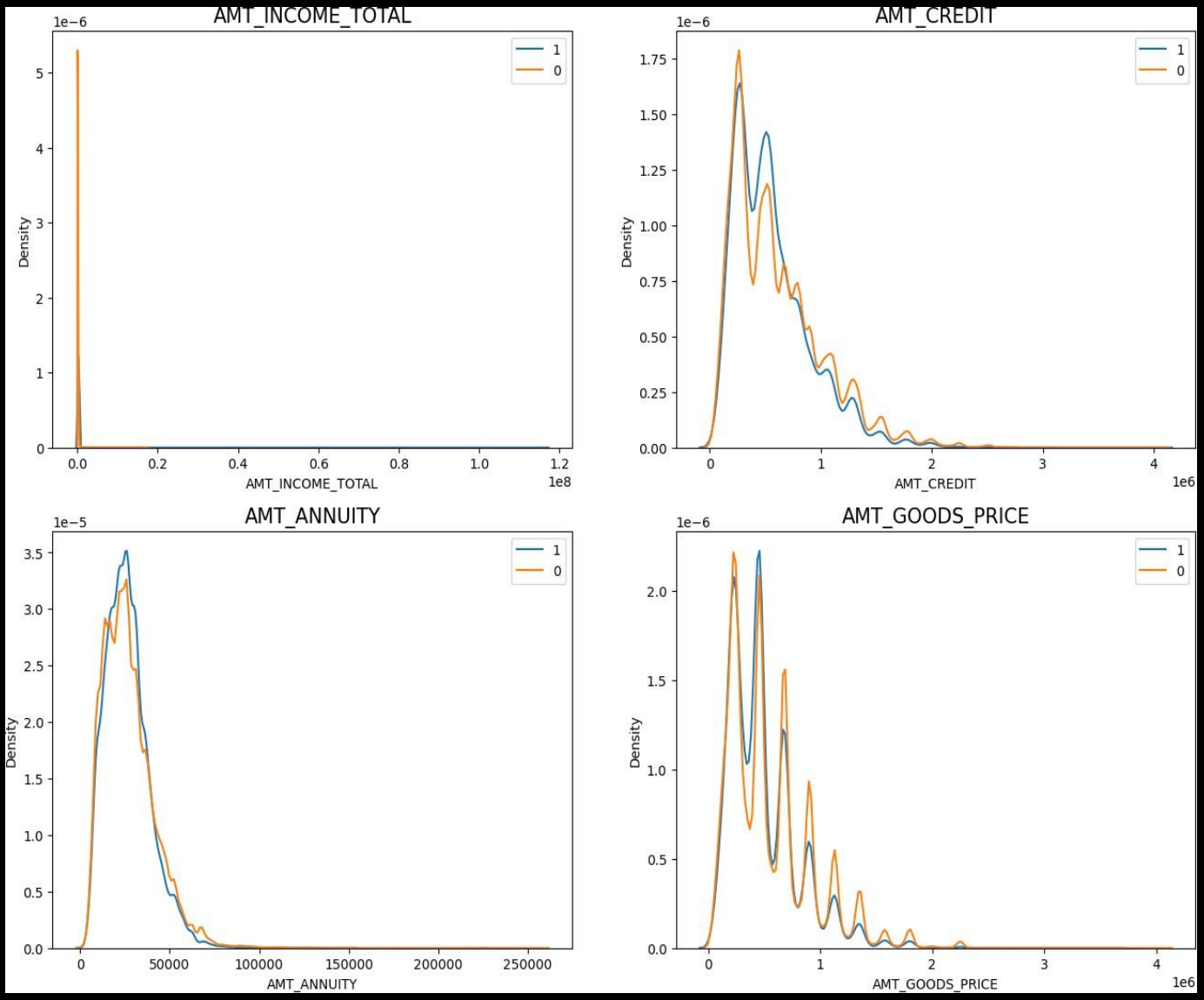
- Most applicants are from age between 30-40 People with higher age are less defaulting
- People between 60-70 years of Age have less default rates.

Univariate Analysis for "DAYS_EMPLOYED":



Days Employment

- Most applicants are having work experience between 1 to 5 years.
- People with work experience between 1-5 years have highest default rate.
- People with least default rate are people having experience more than 25+ Years



Univariate Analysis of Numerical variable

- In this bivariate analysis, we explore the relationships between various loan features and the loan repayment status (defaulters and repayers).

Loan Amount vs. Goods Price

- The majority of loans are granted for goods prices below 1 Million.

Annuity vs. Credit Amount

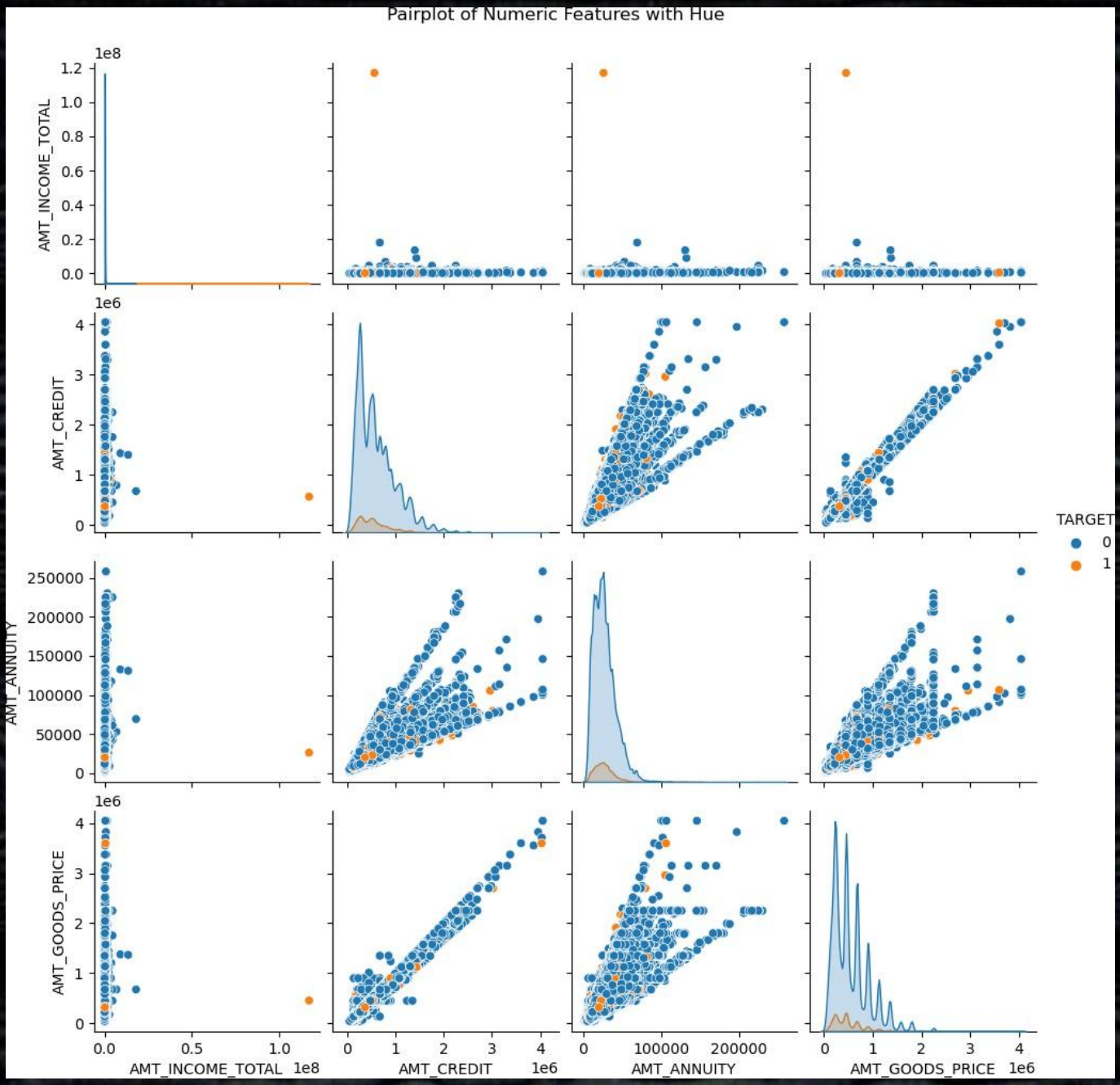
- Most borrowers pay an annuity below 50K for their credit loans.

Credit Amount Distribution

- The credit amount of the loan is mostly less than 1 Million .

Conclusion

- The distributions of repayers and defaulters overlap in all the plots.
- It is challenging to make a decision based solely on these individual variables.
- Further analysis and possibly additional variables are needed to assess loan repayment risk effectively.



1. **Annuity and Goods Price Influence:**

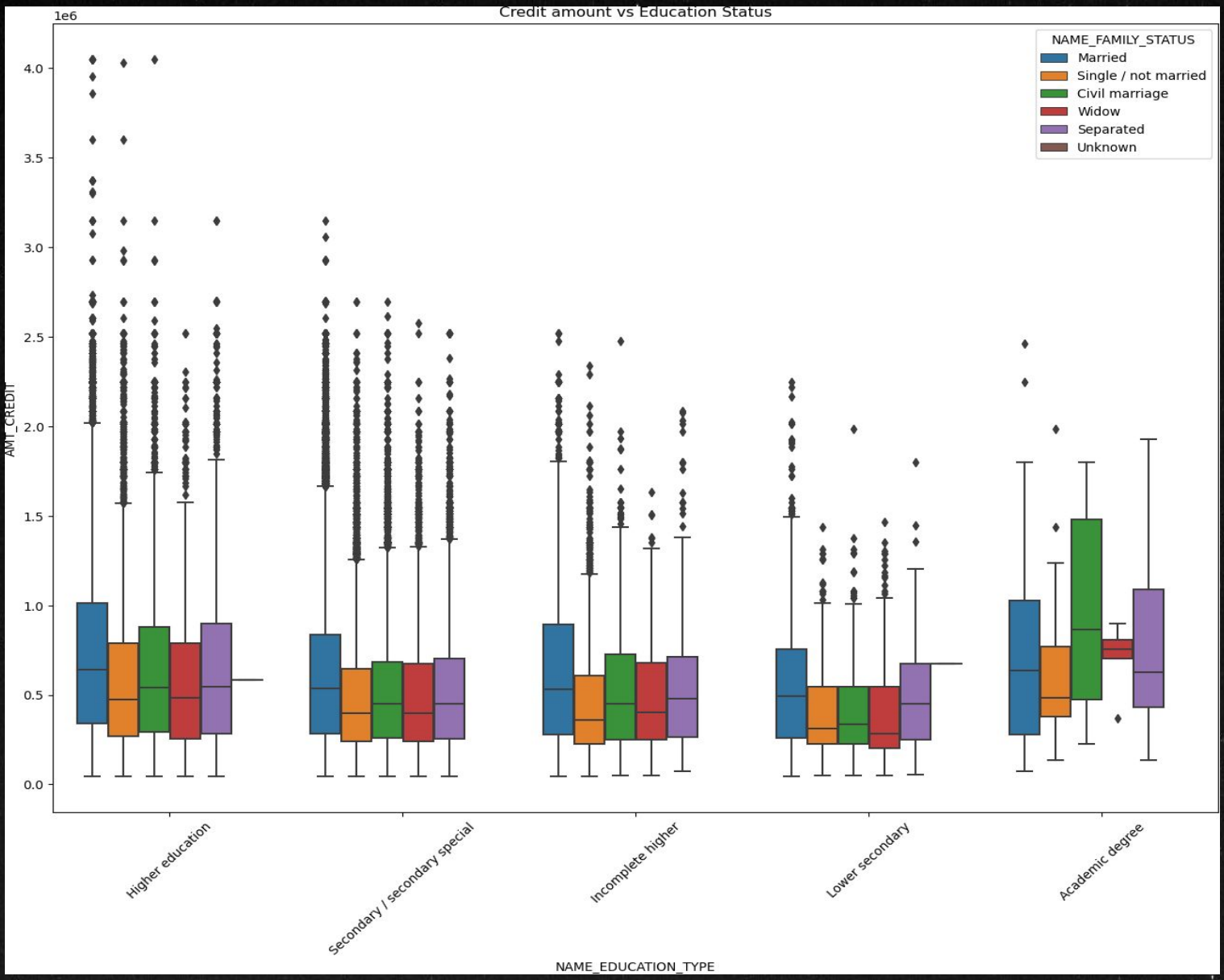
   When the Annuity Amount is greater than 15k and the Goods Price Amount is greater than 20 lakhs, there appears to be a lower chance of loan defaulters. This suggests that clients with higher annuity and goods price values may be more reliable in terms of loan repayment.

2. **Correlation between Loan Amount and Goods Price:**

   Loan Amount (AMT_CREDIT) and Goods Price (AMT_GOODS_PRICE) are highly correlated. This is evident from the scatterplot, where most of the data points are consolidated in the form of a line. This high correlation indicates that clients tend to borrow an amount close to the price of the goods they intend to purchase.

3. **Defaulters for High Loan Amounts:**

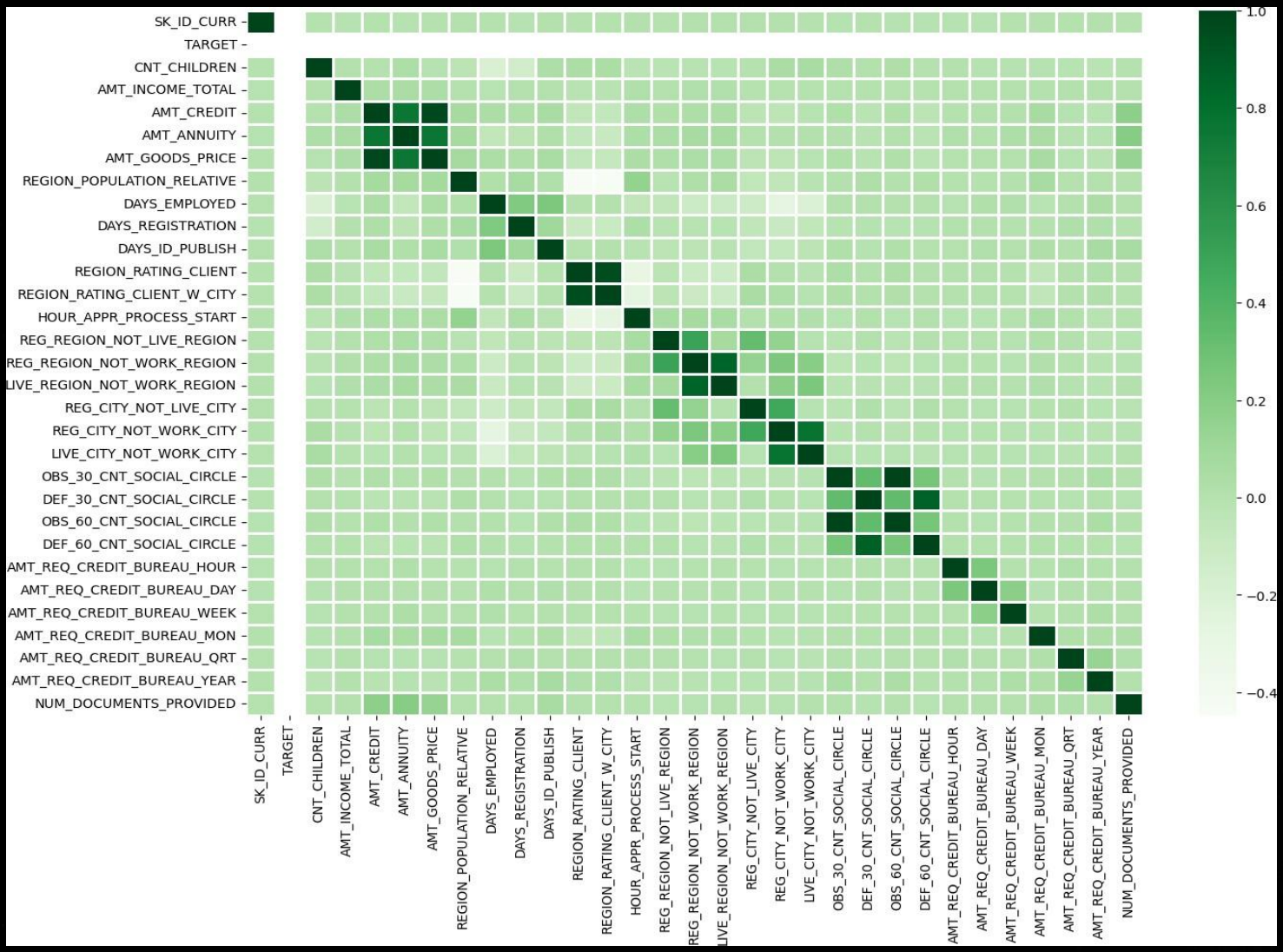
   There are very few defaulters for loan amounts exceeding 2 millions. This suggests that clients who borrow larger sums are less likely to default on their loans. It's important to note that this observation is based on the available data and may not imply causation.



Family Status and Education Analysis

- Clients with family statuses like 'civil marriage,' 'marriage,' and 'separated' who have achieved an 'Academic degree' education tend to have a higher number of credit applications.
- Among clients with 'Higher education,' family statuses such as 'marriage,' 'single,' and 'civil marriage' show more outliers in their credit application amounts, indicating greater variation.
- Notably, clients with 'Civil marriage' status and an 'Academic degree' education primarily fall within the third quartile range, suggesting a preference for higher credit amounts in this group.

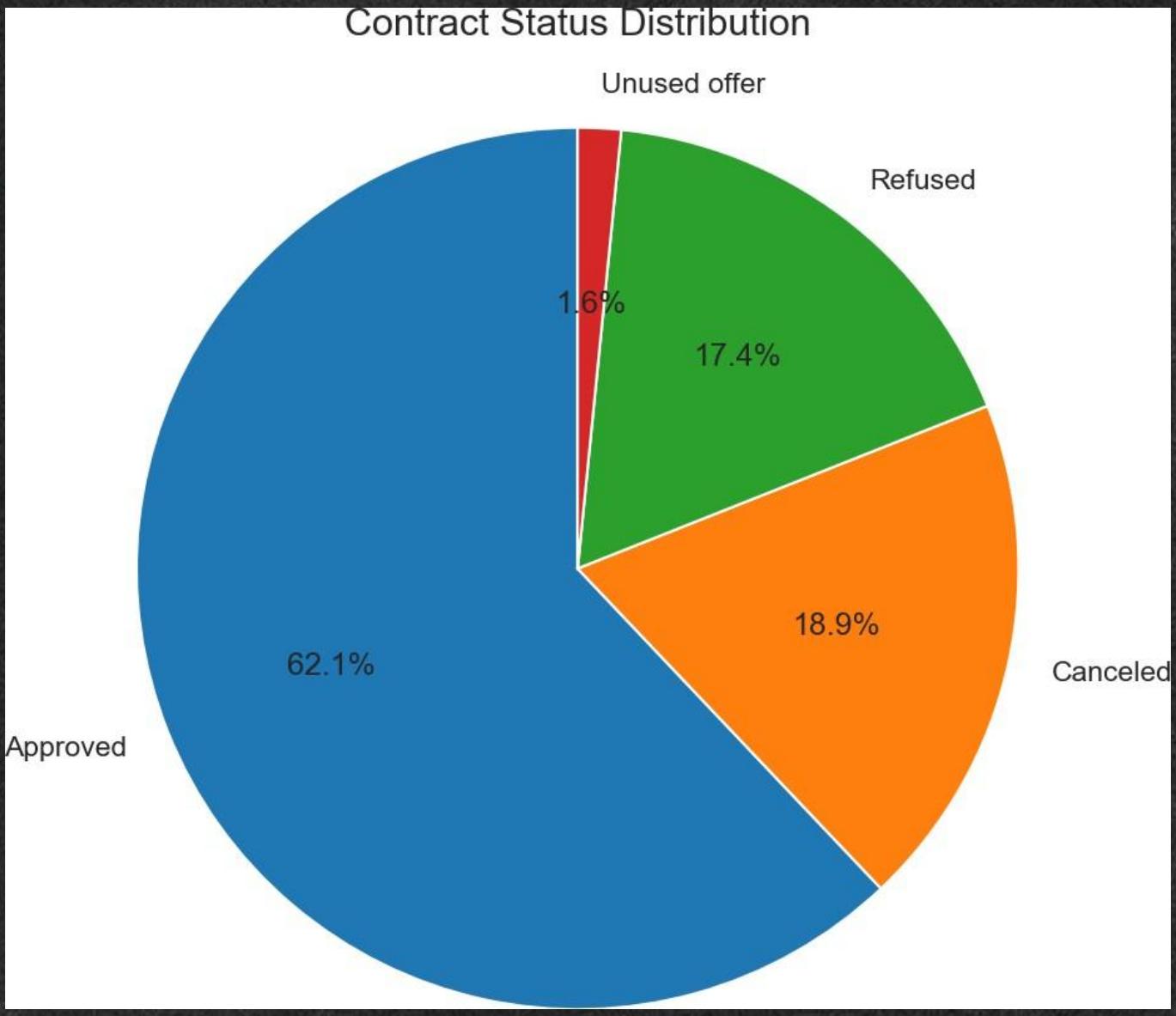
MULTIVARIATE ANALYSIS:



Inferences: Correlating Factors

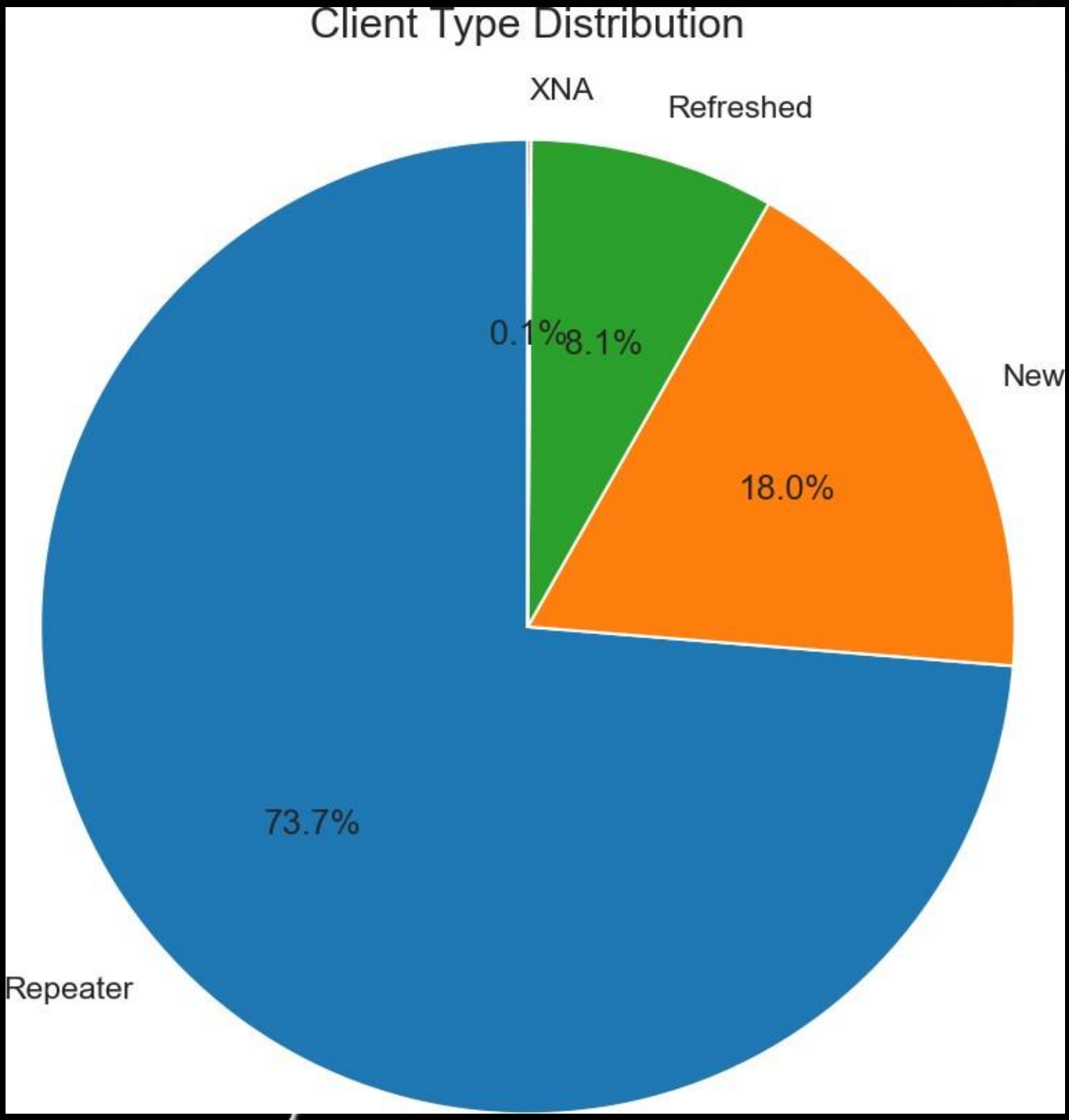
1. **Credit Amount and Goods Price Correlation:** The credit amount is highly correlated with the goods price amount, and this correlation remains consistent among repayers.
2. **Loan Annuity and Credit Amount Correlation:** The correlation between loan annuity and credit amount shows a slight reduction in defaulters when compared to repayers.
3. **Days Employed Correlation:** Repayers exhibit a higher correlation with the number of days employed when compared to defaulters.
4. **Total Income and Credit Amount Correlation:** Among defaulters, there is a significant drop in the correlation between the total income of the client and the credit amount, whereas among repayers, this correlation remains relatively strong.
5. **Days Birth and Number of Children Correlation:** The correlation between days_birth and the number of children is slightly lower in defaulters compared to repayers.
6. **Social Circle Correlation:** The correlation between the social circle and defaulting shows a slight increase among defaulters compared to repayers.

NAME_CONTRACT_STATUS



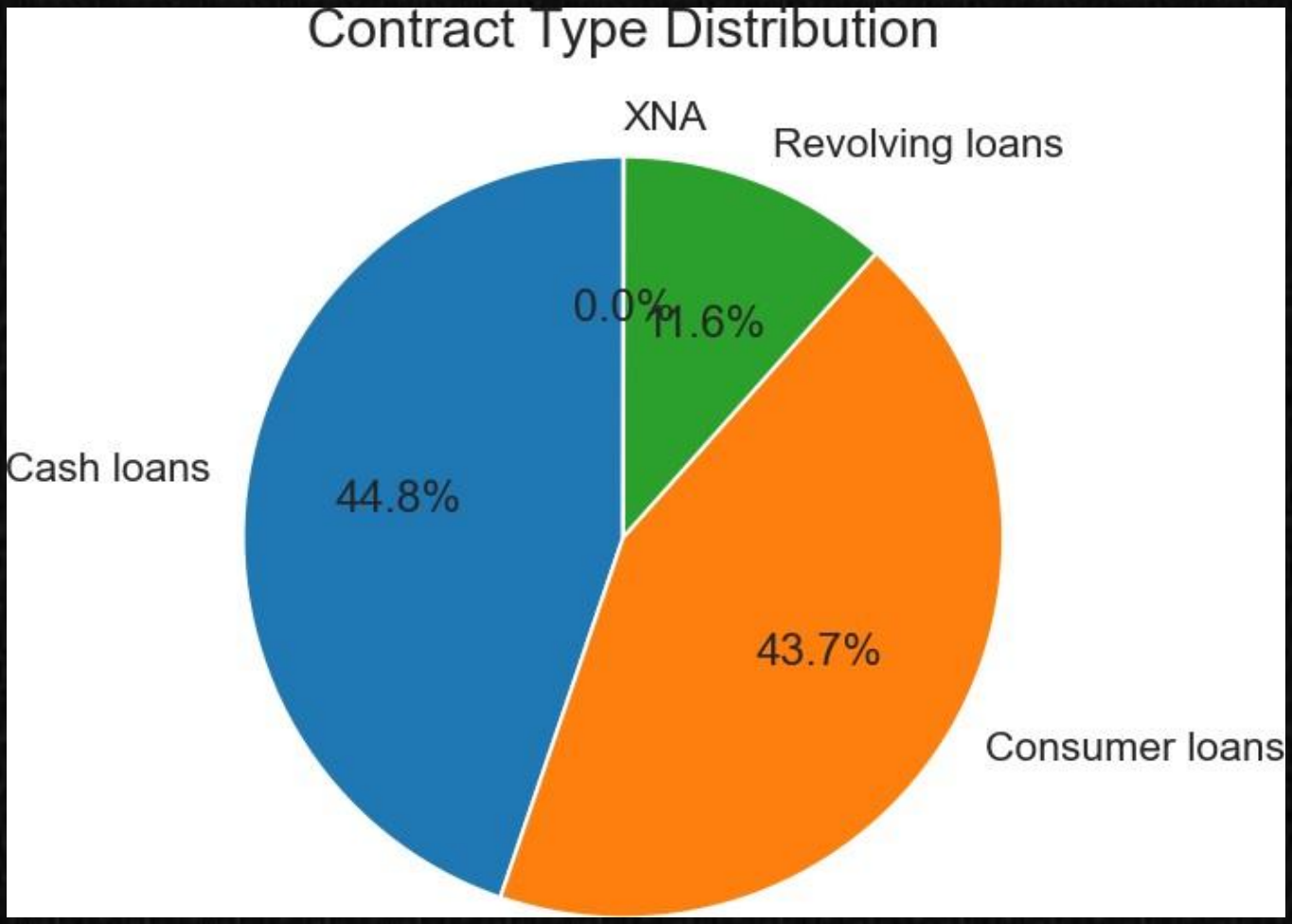
1. **Approved (62.07%)**: The majority of loan applications were successfully approved, indicating a healthy lending rate.
2. **Canceled (18.94%)**: Approximately 19% of applicants canceled their loan requests, suggesting the need to explore reasons for such cancellations.
3. **Refused (17.40%)**: About 17% of applications were denied, highlighting the importance of assessing rejection criteria and applicant profiles.
4. **Unused Offer (1.58%)**: A small percentage received offers but didn't proceed, warranting investigation into factors influencing this decision.

This analysis provides insights for optimizing the lending process and mitigating default risks.



- **Repeater:** 73.7% of clients fall into this category.
- **New:** Approximately 18.0% of clients are categorized as "New."
- **Refreshed:** About 8.1% of clients are classified as "Refreshed."
- **XNA:** A very small percentage, only 0.1%, is labeled as "XNA."

NAME_CONTRACT_TYPE



Contract Type Distribution

Let's analyze the distribution of contract types in the dataset. Contract types indicate the various types of loans offered by the company and their respective percentages:

- **Cash loans**: The most common contract type, accounting for 44.76% of all contracts.
- **Consumer loans**: A significant portion of contracts, making up 43.66% of the total.
- **Revolving loans**: Less common, representing 11.57% of contracts. Typically associated with credit cards or lines of credit.
- **XNA**: An extremely rare category, making up only 0.02% of the contracts. The exact meaning of "XNA" is unclear.

Insights from Previous Application Data

1. **Outliers in Financial Variables:**

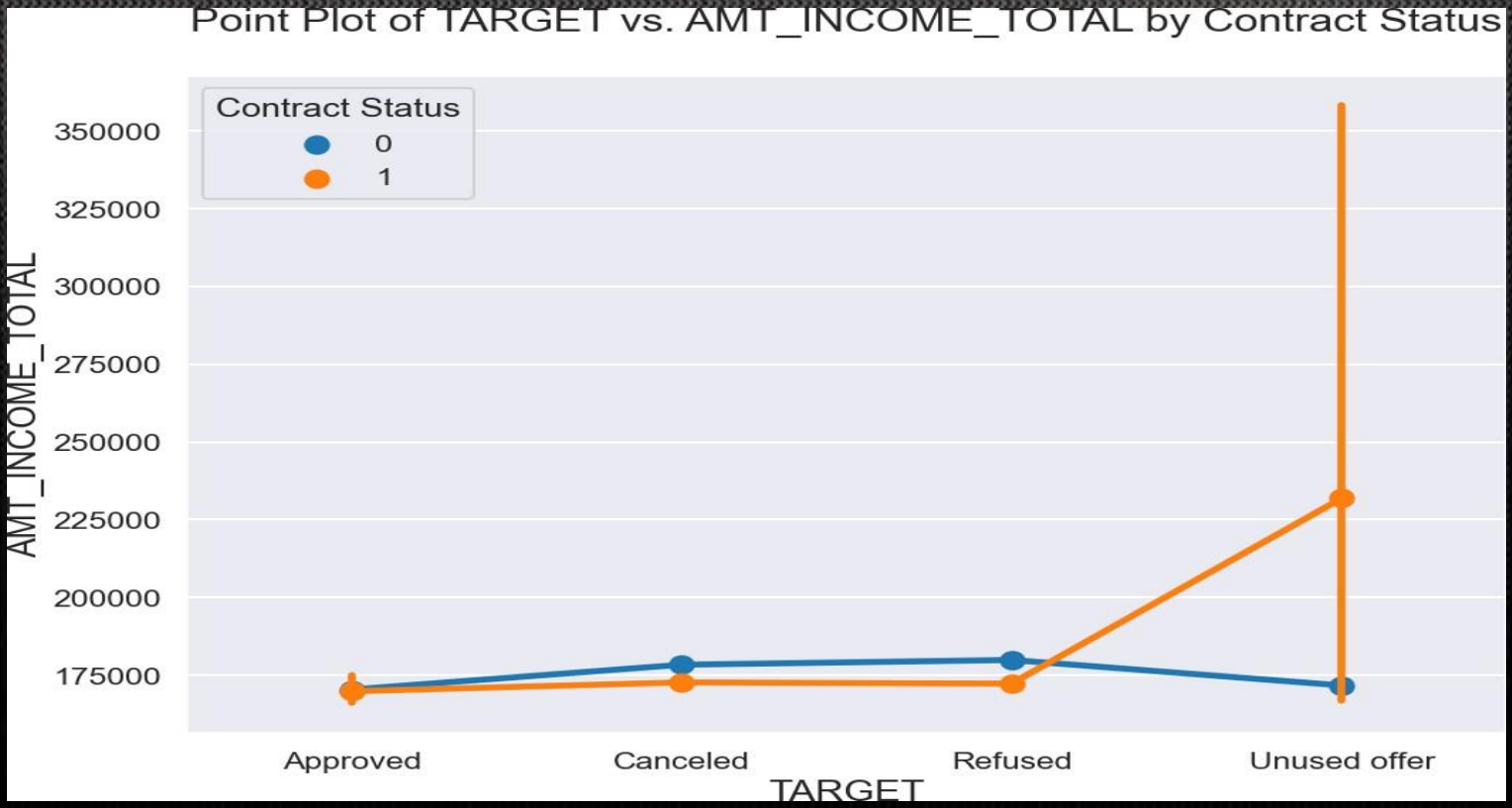
   It is evident that several financial variables, namely AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, and SELLERPLACE_AREA, contain a substantial number of outliers. These outliers indicate significant variations in loan and purchase amounts, which should be carefully considered in further analysis.

2. **Outliers in Payment Count (CNT_PAYMENT):**

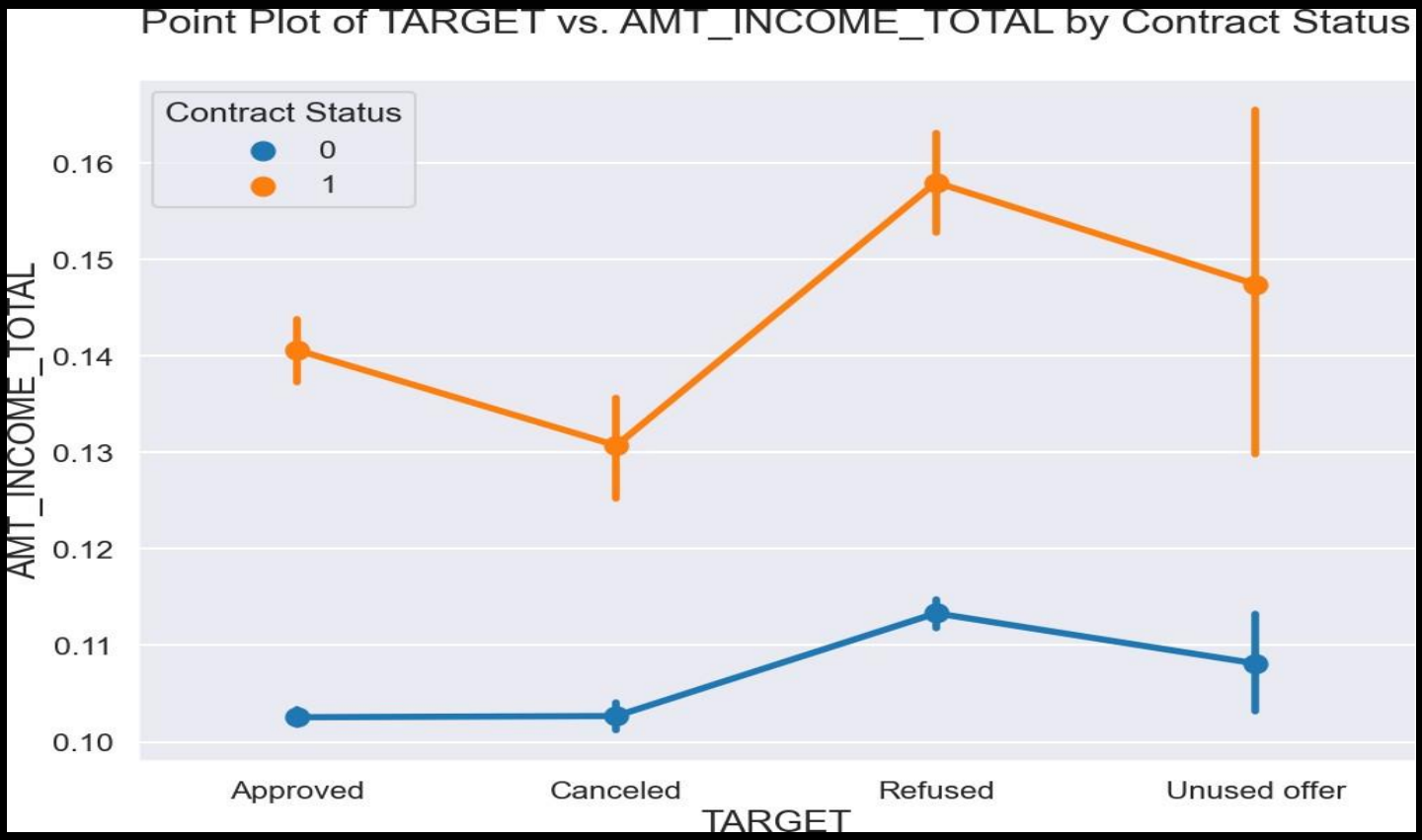
   The variable CNT_PAYMENT displays a few outlier values. These outliers suggest deviations from the typical payment counts associated with previous loan applications. Exploring the reasons behind these deviations may provide valuable insights.

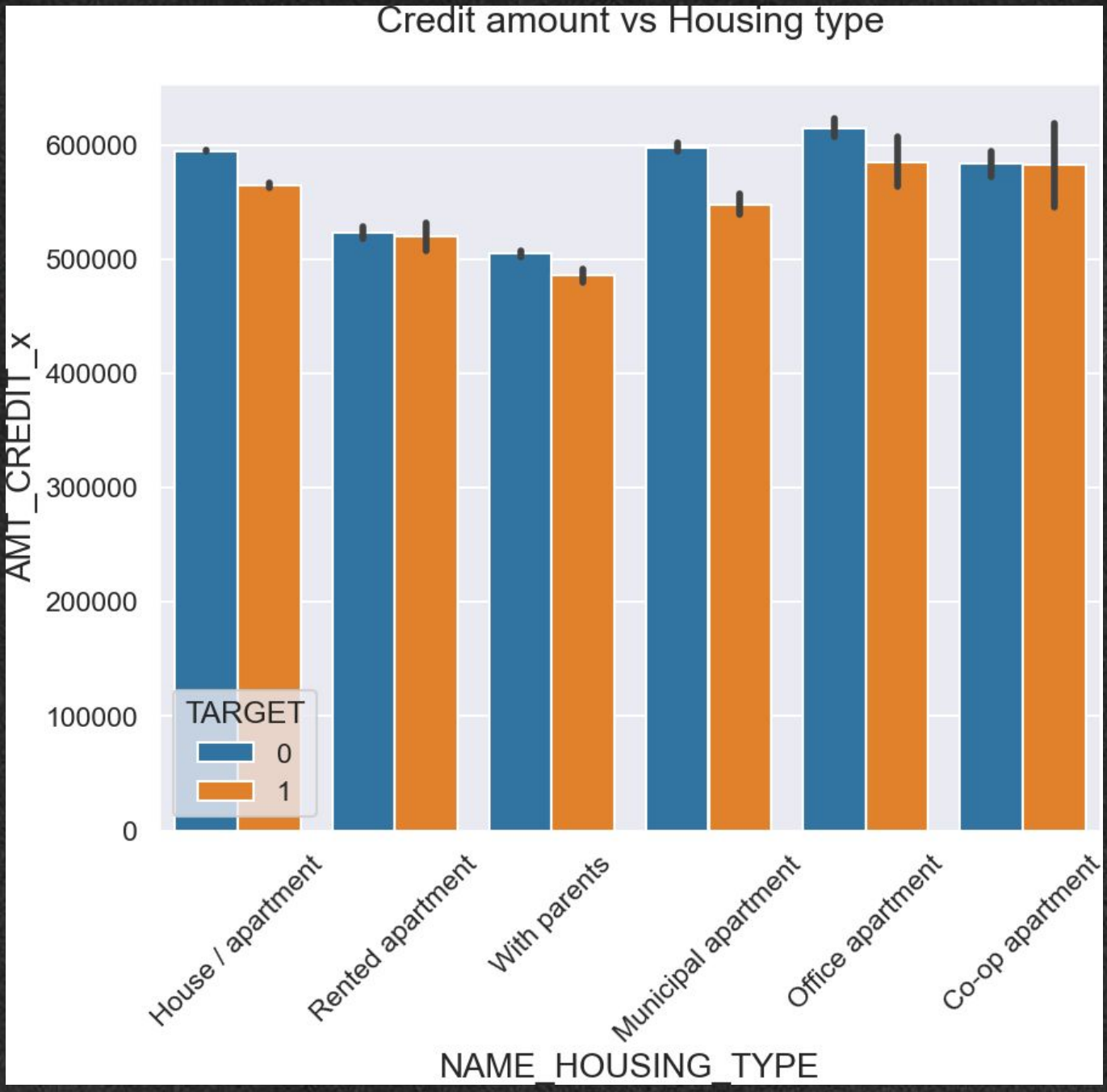
3. **DAYS_DECISION and Outliers:**

   When examining the variable DAYS_DECISION, it becomes evident that it has a relatively small number of outliers. This observation implies that decisions related to previous loan applications were made a long time ago. This aspect may be indicative of historical or infrequent decision- making processes.



The point plot reveals an interesting trend. It indicates that individuals who have not used an offer earlier ("Unused offer" category in NAME_CONTRACT_STATUS) are more likely to default on their loans, even when their average income is higher compared to other contract status categories. This suggests that the history of not using previous loan offers may be a significant factor in predicting loan defaults, potentially even more important than income alone.





Housing Type Analysis:

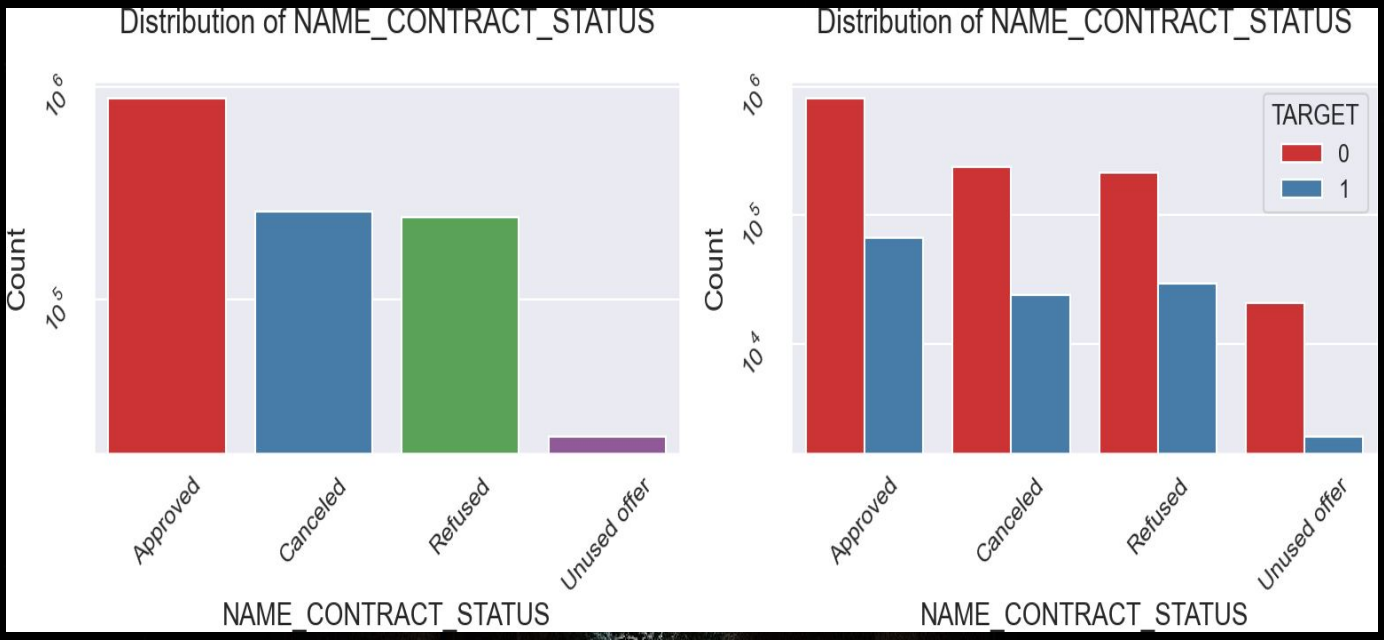
1. Among various housing types, "Office apartment" has a higher percentage of clients with target 0 (repaid loans) compared to other housing types, indicating a better repayment rate for this category.

2. Conversely, "Co-op apartment" exhibits a higher percentage of clients with target 1 (defaulted loans), suggesting a higher credit risk associated with this housing type.

3. Based on this analysis, it is advisable for the bank to exercise caution when granting loans to clients with "Co-op apartment" housing type due to their higher likelihood of payment difficulties.

4. To mitigate risk and improve loan repayment rates, the bank can focus its lending efforts on housing types such as "With parents," "House/apartment," or "Municipal apartment," which appear to have a better track record of successful payments.

In [224...



Observations on Previous Client Behavior:

1. Clients who had previously canceled their loan applications demonstrated a positive repayment behavior in the current case. Adjusting interest rates for these clients could potentially lead to increased business opportunities.

2. Similarly, clients who were previously refused a loan have shown a favorable repayment trend in the current scenario. It is advisable to record the reasons for their initial loan refusal, as these clients may become potential reliable borrowers in the future.

Key Factors for Successful Repayments

Education Matters: Applicants holding academic degrees exhibit a significantly lower default rate.

Income Diversity: Individuals in various income-generating roles, such as students and businessmen, demonstrate a strong repayment record.

Regional Influence: The geographical rating, specifically the top-rated region, is linked with safer loan repayment behavior.

Occupational Insights: Applicants affiliated with certain professional sectors like Trade Types 4 and 5, or Industry Type 8, tend to default less frequently.

Age Advantage: Loan applicants above the age of 50 exhibit a lower tendency to default.

Seasoned Employment: Clients with over four decades of work experience have a default rate of under 1%.

Financial Strength: Applicants with incomes exceeding 10 Millions experience a lower likelihood of default.

Loan Purpose Precision: Loans designated for hobbies or garage purchases demonstrate higher repayment reliability.

Family Influence: Clients with fewer dependents, ranging from zero to two children, consistently honor their loan commitments.

Strategic Loan Approval Considerations

- **Housing Insight:** Applicants residing in rented apartments or with their parents represent a significant portion of loan applicants. Offering loans to this group with slightly higher interest rates may help mitigate potential losses due to defaults.
- **Credit Range Caution:** Loans within the 300K-600K range have shown a consistent default trend. Implementing elevated interest rates for this specific credit bracket could be advisable.
- **Income-Tiered Approach:** Given that around 90% of applicants have a total income of less than 300K , applying higher interest rates to loans for this demographic might help offset potential default risks.
- **Family Size Impact:** Clients with larger families (4-8 children) have demonstrated a high default rate, warranting the imposition of higher interest rates on their loans.
- **Loan Purpose Prudence:** Loans intended for repairs have exhibited the highest default rate. This aligns with the bank's cautious approach in the past, either rejecting such applications or offering loans at prohibitive interest rates. Maintaining this approach may be advisable in the future.

Strategic Recommendations

- **Record Cancellation Reasons:** Investigate the reasons behind loan cancellations, as a significant proportion of previously canceled clients have repaid their loans. Understanding these reasons could open opportunities for renegotiating terms with these clients.
- **Reconsider Rejected Applicants:** Nearly 88% of clients previously rejected by the bank have become successful repayers. Documenting the reasons for rejection and revisiting these applicants could help mitigate business losses and expand lending opportunities.