CREDIT EDA CASE STUDY

MANGESH SATISH KENDRE

Categorical Univariate analysis for Target = 0

Distribution of Income range

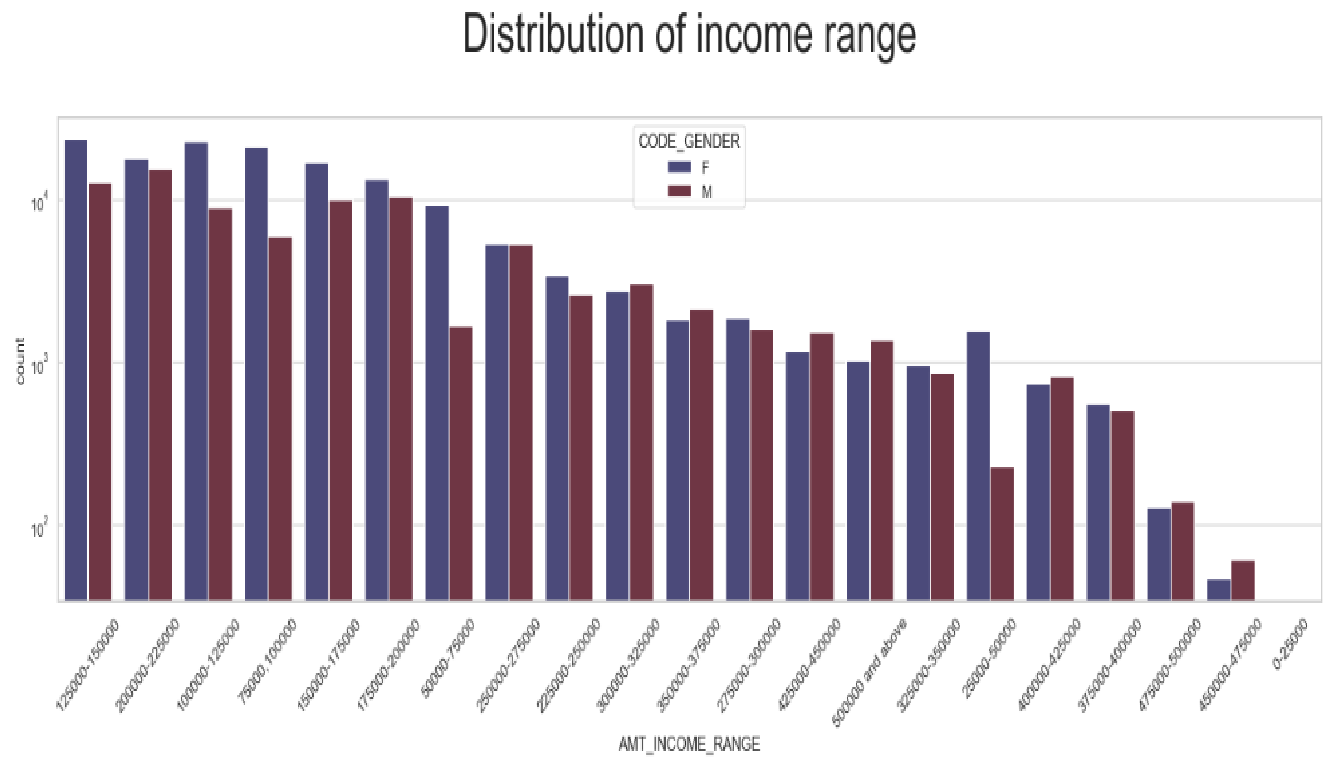
Key observations drawn from the provided graph include:

Gender Distribution: The graph illustrates a higher count of females compared to males.

Credit Distribution by Income: The range of income spanning from 100,000 to 200,000 exhibits a notable concentration of credit occurrences.

Gender and Credit: This depiction unveils that within the aforementioned income range (100,000 to 200,000), the number of females surpasses that of males in obtaining credits.

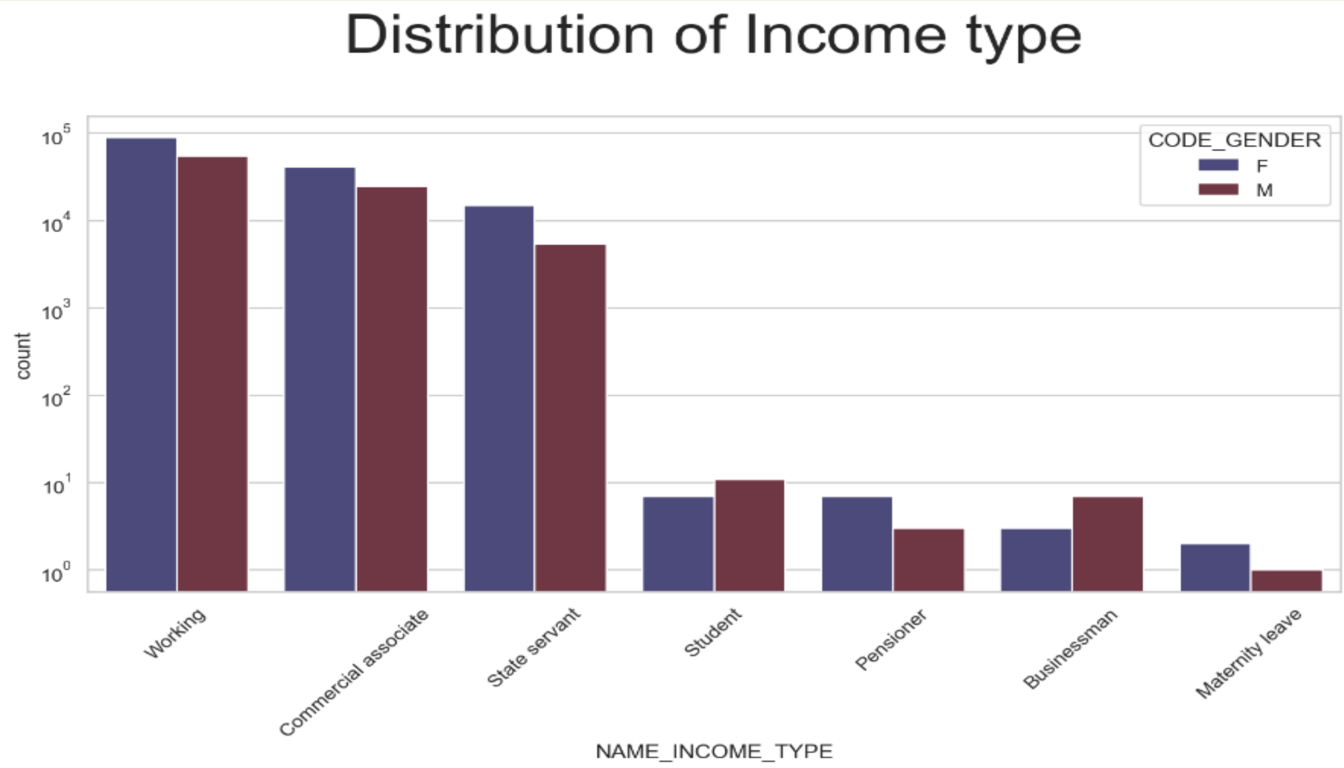
Sparse Credit Uptake: Notably, there is a significant scarcity in credit uptake for individuals with income levels of 400,000 and beyond.



Distribution of income type

Key takeaways from the presented graph include:

Income Type and CreditDistribution: The graph indicatesthat the categories 'working,' 'commercial associate,' and 'State Servant' exhibit a substantial prevalence of credit acquisition compared to the other income types.
Gender and Credit Distribution: Within the aforementioned income categories, females tend to secure a higher count of credits in comparison to males.
Limited Credit Uptake: Conversely, there is a noticeable dearth of credit instances among individuals with income types such as 'student,' 'pensioner,' 'Businessman,' and those on 'Maternity leave.'

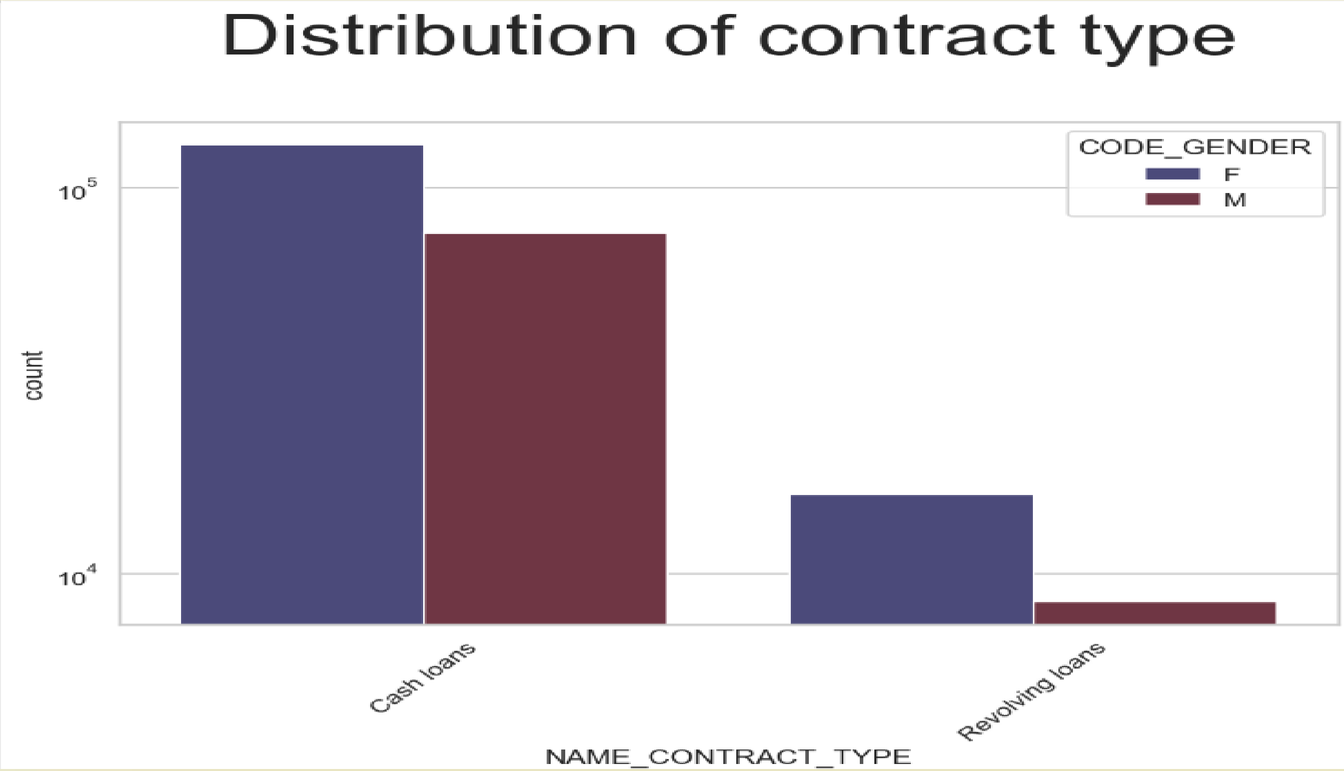


Distribution for contract type

Key observations drawn from the provided graph are as follows:

Credit Type Comparison: The graph reveals that the category 'cash loans' exhibits a notably higher count of credit occurrences in comparison to 'Revolving loans.'

Gender and Credit Type: Within the 'cash loans' category, females take the lead in applying for credits.

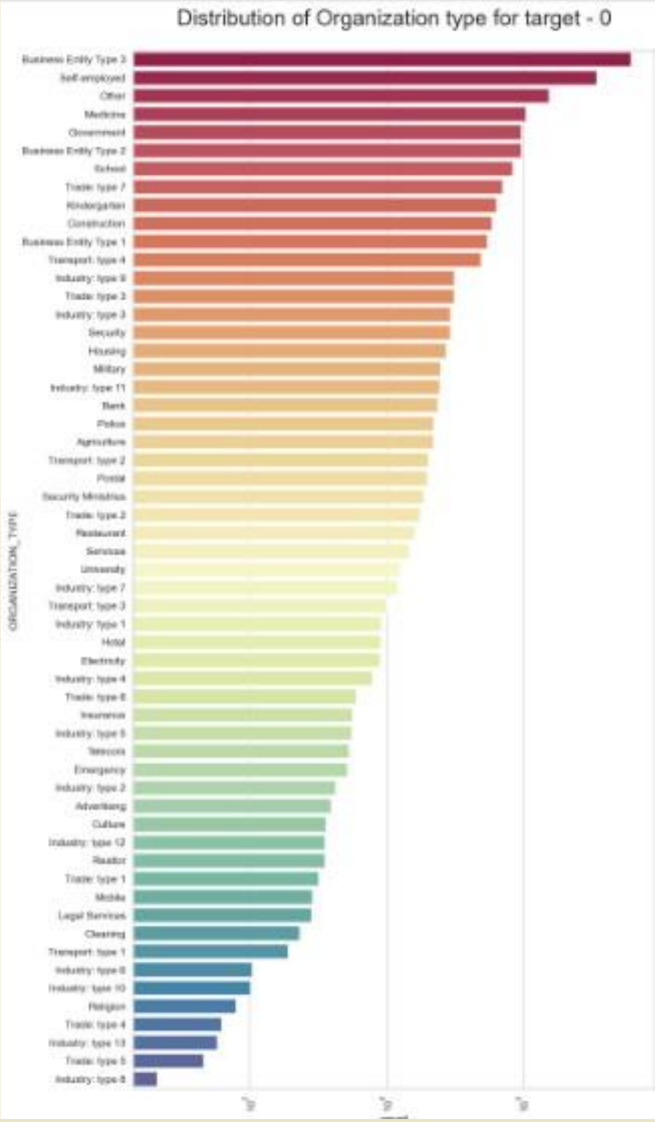


Distribution of organization type

Key observations derived from the presented graph include:

Prominent Organization Types: The graph illustrates that clients who have applied for credits predominantly belong to various organization types such as 'Business entity Type 3,' 'Self employed,' 'Other,' 'Medicine,' and 'Government.'

Limited Representation: Conversely, there is a lower representation of clients from organization types like 'Industry type 8,' 'type 6,' 'type 10,' 'religion and trade type 5,' and 'type 4.'



Categorical Univariate analysis for Target = 1

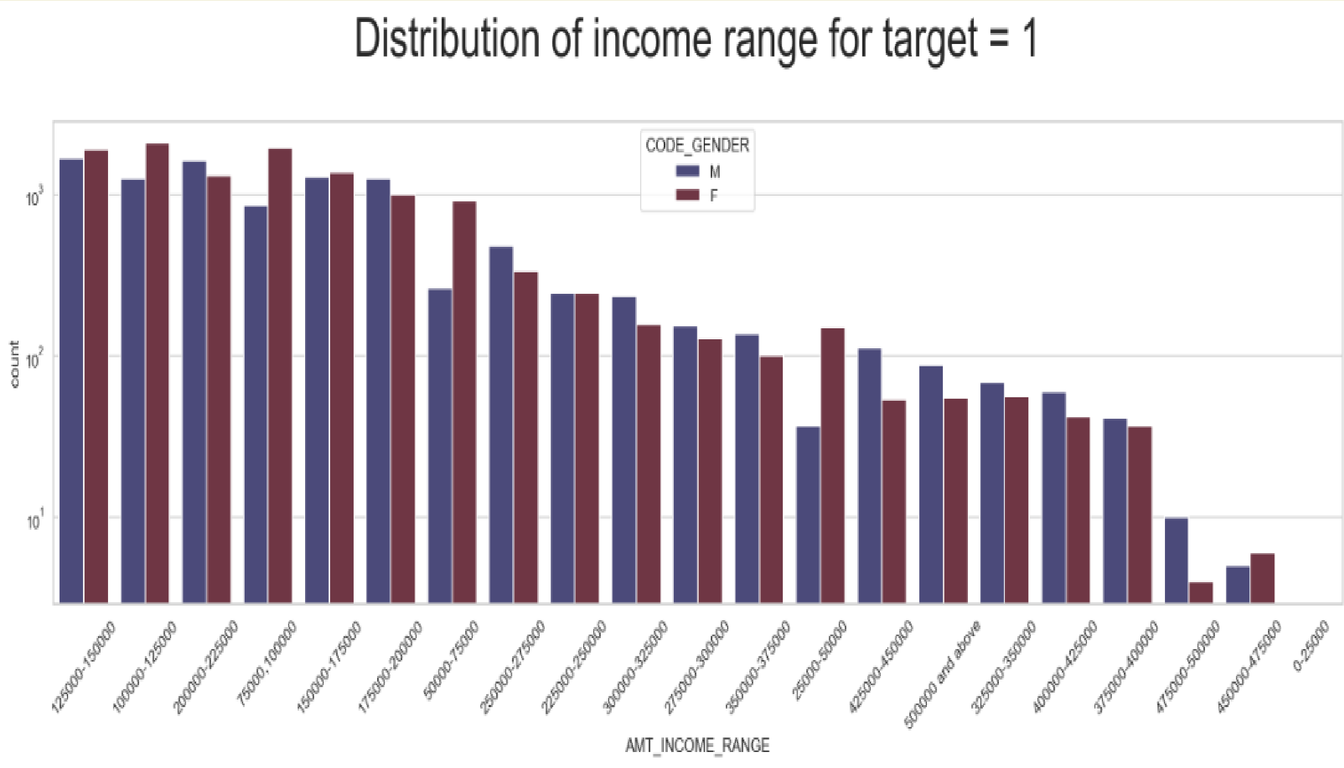
Distribution of Income range

Key takeaways from the presented graph are as follows:

Gender Distribution: The graph indicates that, overall, the count of males is greater than that of females.

Credit Distribution by Income: Within the dataset, there is a notable concentration of credit occurrences in the income range between 100,000 and 200,000.

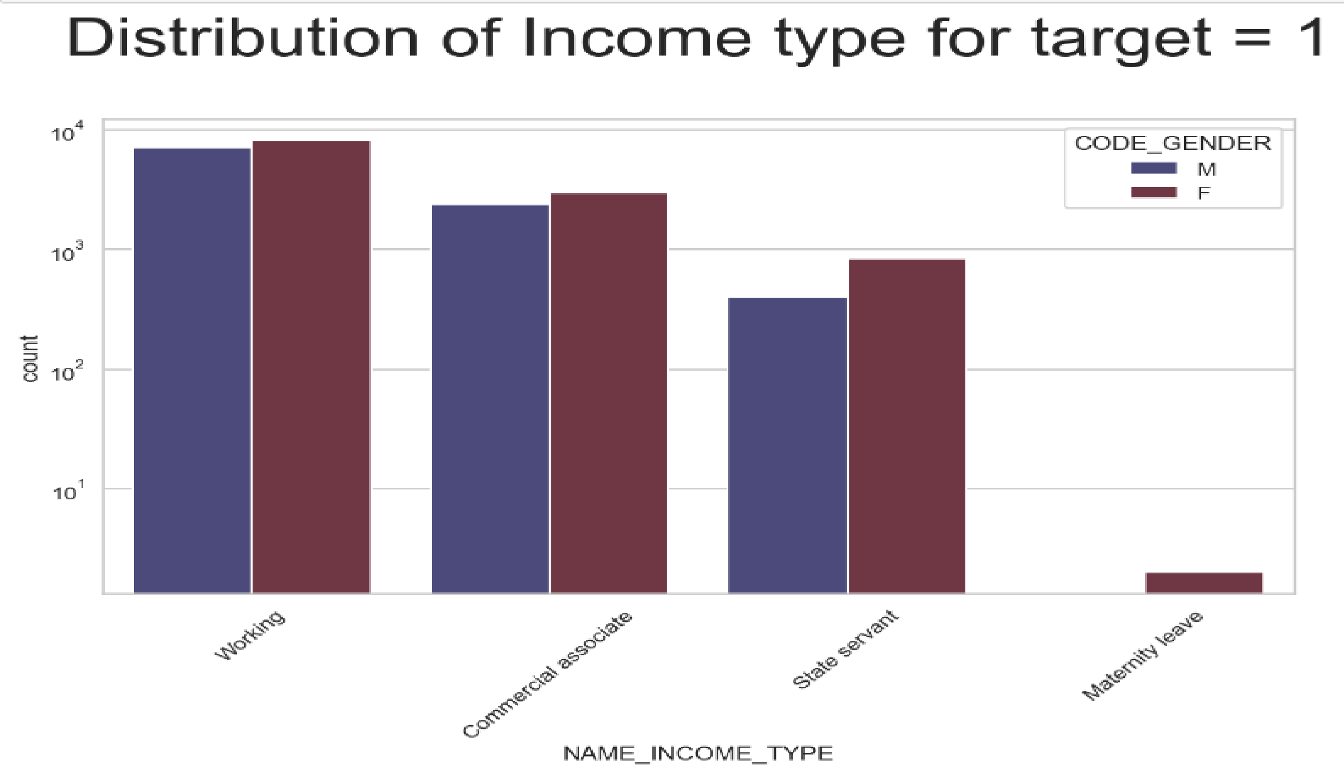
Limited High-Income Credits: Conversely, there is a considerably lower number of credit instances for individuals with income levels of 400,000 and beyond.



Distribution of income type

Key observations drawn from the provided graph include:

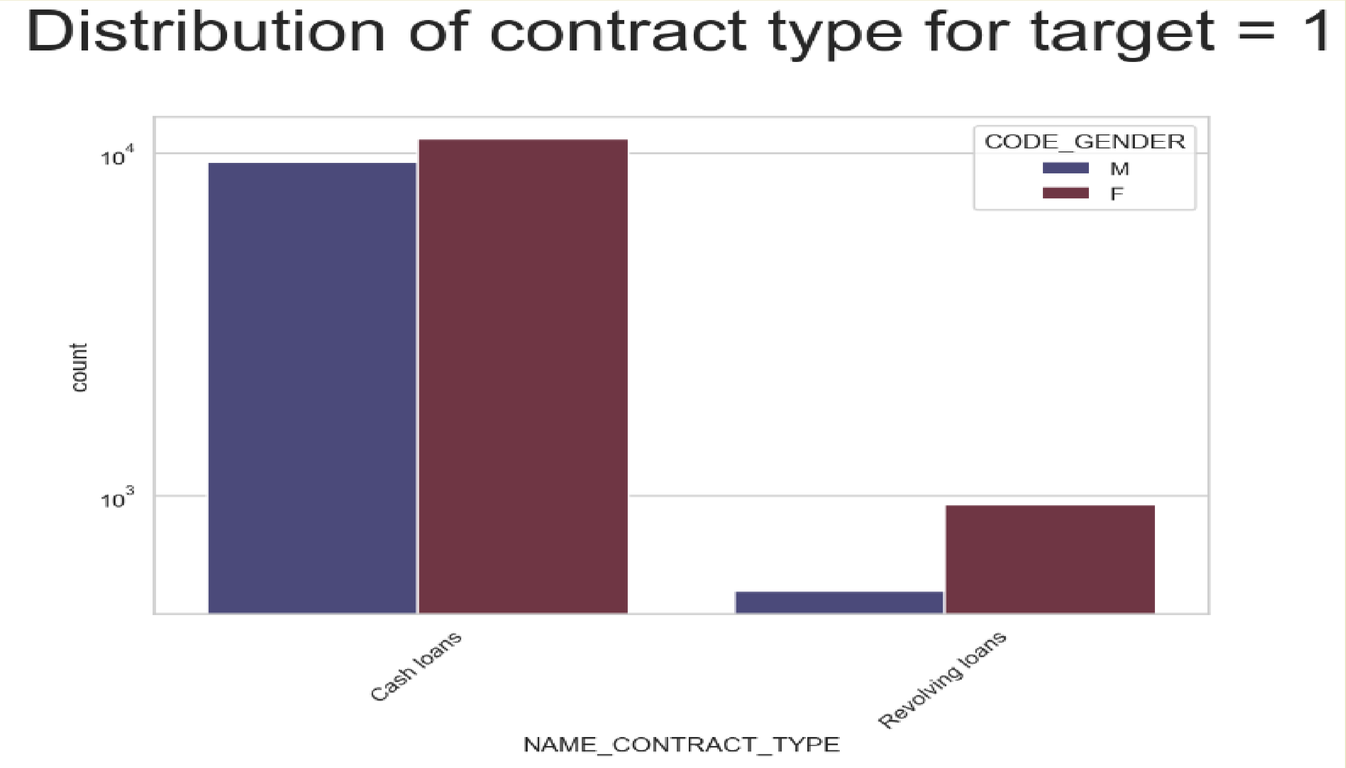
Income Type and Credit Distribution: The graph indicates that among various income types, such as 'working,' 'commercial associate,' and 'State Servant,' the number of credits is notably higher compared to the 'Maternity leave' income type.
Gender and Credit Distribution: Within the aforementioned income categories, females tend to secure a higher number of credits in comparison to males.
Limited Credit Uptake and Target Analysis: The 'Maternity leave' income type demonstrates a lower count of credits, potentially due to the specific circumstances associated with it. Furthermore, for the 'target = 1' group, there is an absence of income types like 'student,' 'pensioner,' and 'Businessman,' which implies a lack of late payments within these categories.



Distribution for contract type

Key observations drawn from the provided graph include:

Income Type and Credit Distribution: The graph indicates that among various income types, such as 'working,' 'commercial associate,' and 'State Servant,' the number of credits is notably higher compared to the 'Maternity leave' income type.
Gender and Credit Distribution: Within the aforementioned income categories, females tend to secure a higher number of credits in comparison to males.
Limited Credit Uptake and Target Analysis: The 'Maternity leave' income type demonstrates a lower count of credits, potentially due to the specific circumstances associated with it. Furthermore, for the 'target = 1' group, there is an absence of income types like 'student,' 'pensioner,' and 'Businessman,' which implies a lack of late payments within these categories.



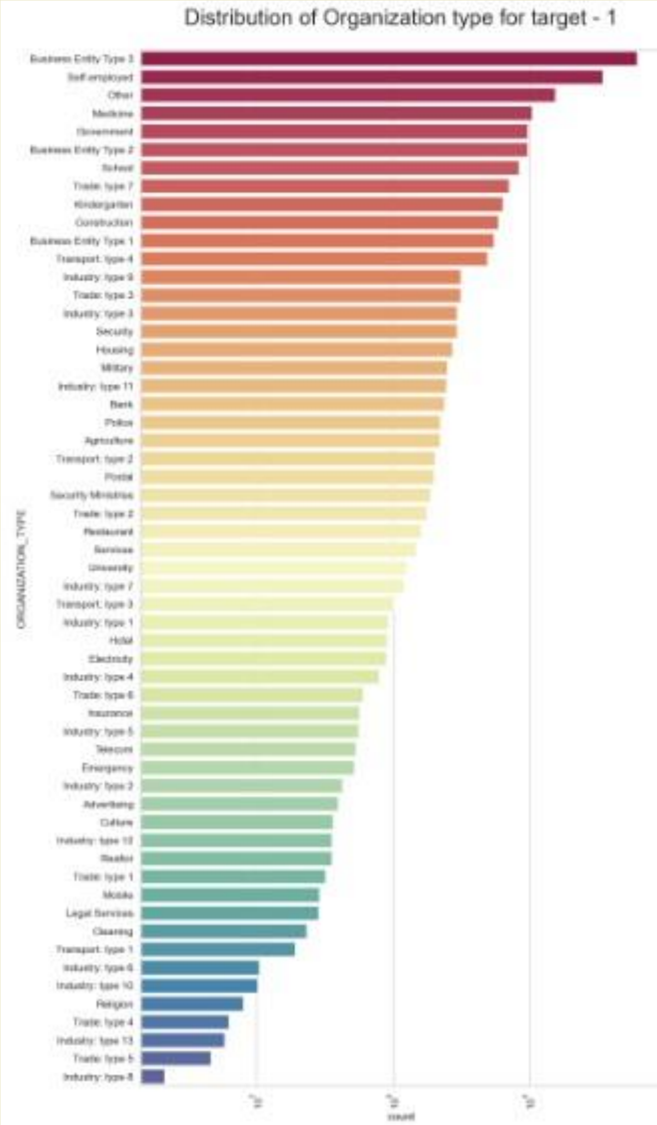
Distribution of organization type

Key takeaways from the provided graph are as follows:

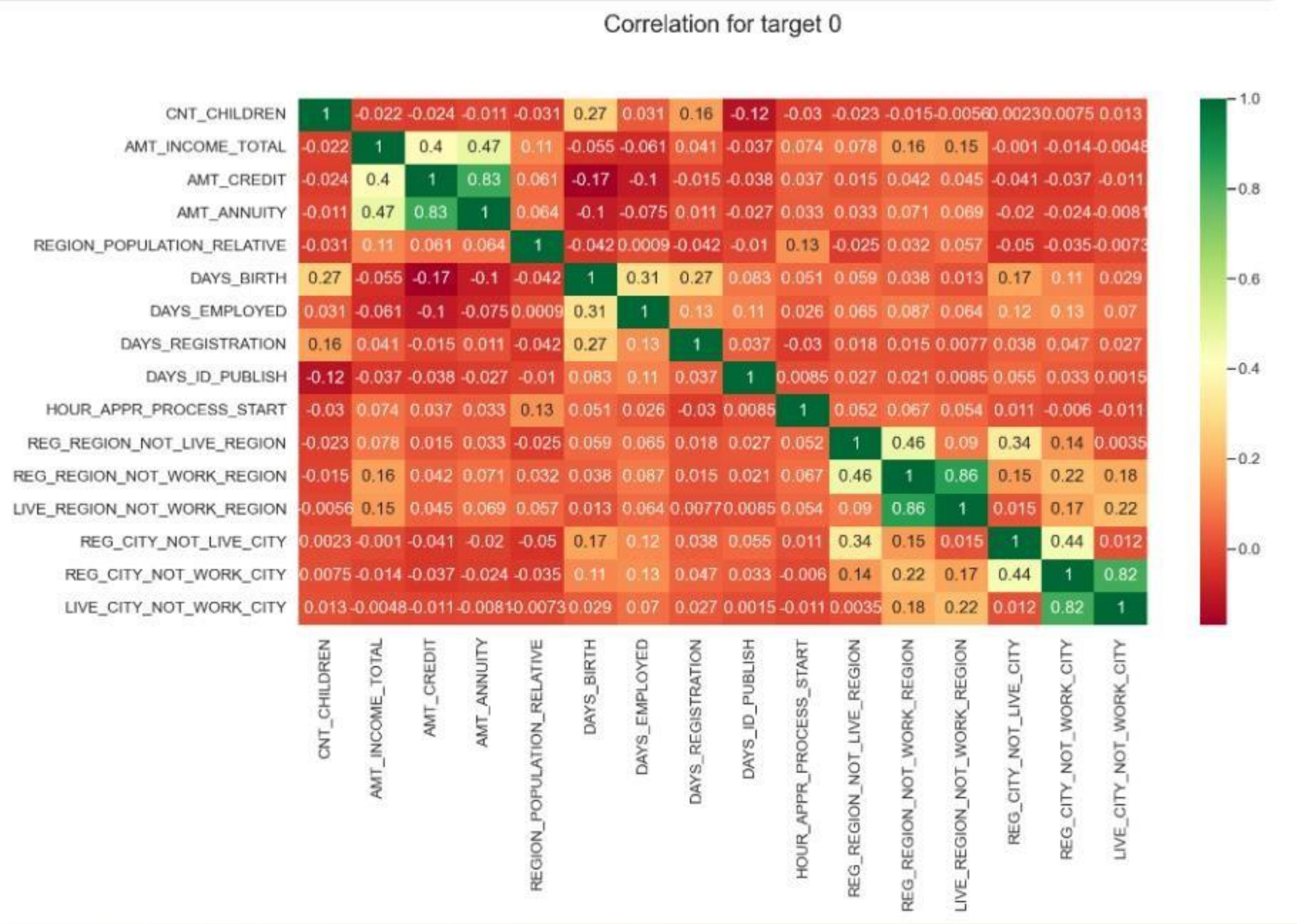
Prominent Organization Types: The graph indicates that clients who have applied for credits predominantly belong to organization types such as 'Business entity Type 3,' 'Self employed,' 'Other,' 'Medicine,' and 'Government.'

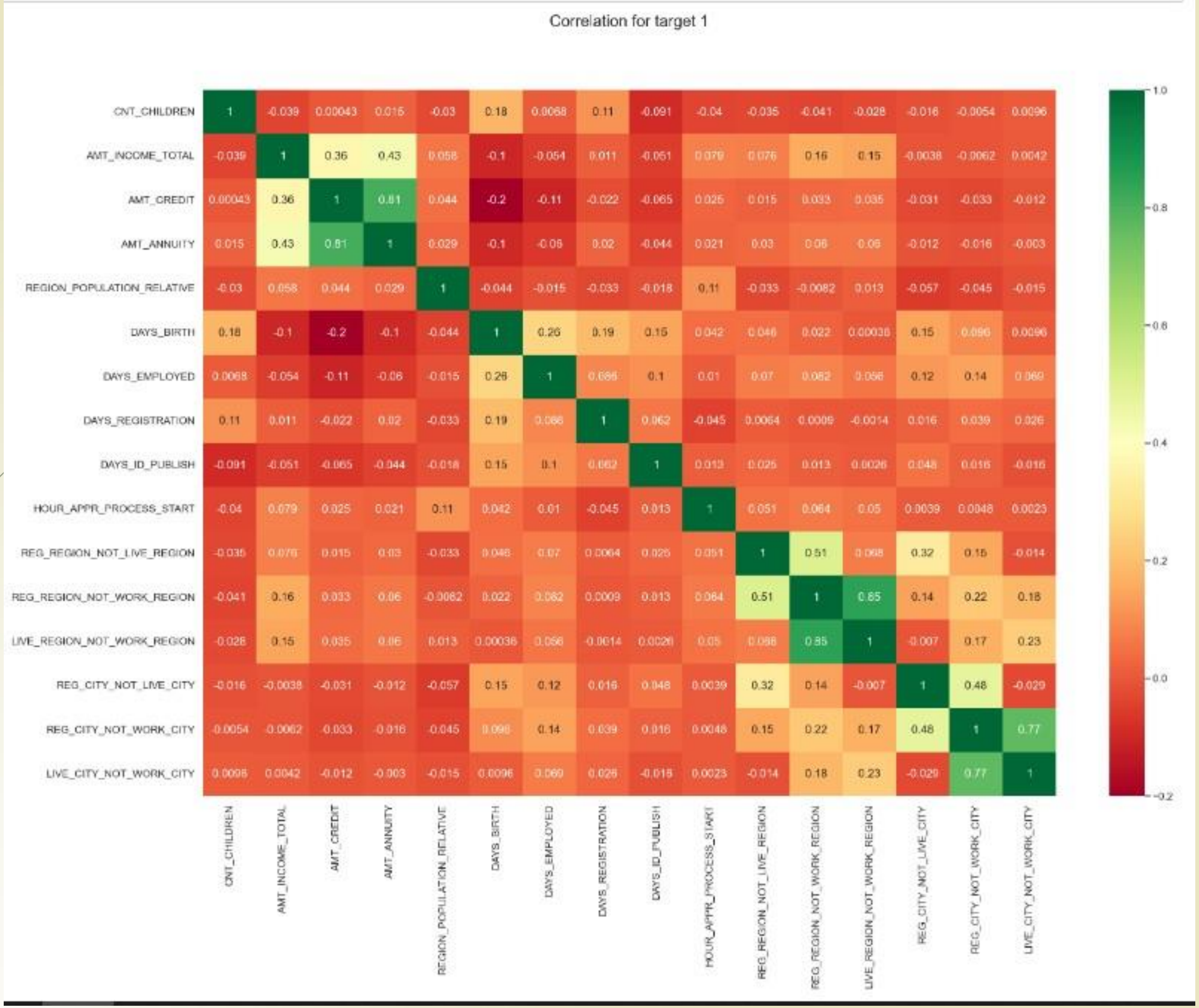
Limited Representation: Conversely, there is a lower representation of clients from organization types like 'Industry type 8,' 'type 6,' 'type 10,' 'religion and trade type 5,' and 'type 4.'

Consistency Across Targets: The distribution of clients based on organization types remains consistent between 'target = 0' and 'target = 1,' indicating that organization type is not a significant discriminator when considering the target variable.



Correlation for Target = 0


Correlation for target 0
CNT_CHILDREN 1 -0.022 -0.024 -0.011 -0.031 0.27 0.031 0.16 -0.12 -0.03 -0.023 -0.015 -0.0056 0.0023 0.0075 0.013
AMT_INCOME_TOTAL -0.022 1 0.4 0.47 0.11 -0.055 -0.061 0.041 -0.037 0.074 0.078 0.16 0.15 -0.001 -0.014 -0.0048
AMT_CREDIT -0.024 0.4 1 0.83 0.061 -0.17 -0.1 -0.015 -0.038 0.037 0.015 0.042 0.045 -0.041 -0.037 -0.011
AMT_ANNUITY -0.011 0.47 0.83 1 0.064 -0.1 -0.075 0.011 -0.027 0.033 0.033 0.071 0.069 -0.02 -0.024 -0.0081
REGION_POPULATION_RELATIVE -0.031 0.11 0.061 0.064 1 -0.042 0.0009 -0.042 -0.01 0.13 -0.025 0.032 0.057 -0.05 -0.035 -0.0073
DAYS_BIRTH 0.27 -0.055 -0.17 -0.1 -0.042 1 0.31 0.27 0.083 0.051 0.059 0.038 0.013 0.17 0.11 0.029
DAYS_EMPLOYED 0.031 -0.061 -0.1 -0.075 0.0009 0.31 1 0.13 0.11 0.026 0.065 0.087 0.064 0.12 0.13 0.07
DAYS_REGISTRATION 0.16 0.041 -0.015 0.011 -0.042 0.27 0.13 1 0.037 -0.03 0.018 0.015 0.0077 0.038 0.047 0.027
DAYS_ID_PUBLISH -0.12 -0.037 -0.038 -0.027 -0.01 0.083 0.11 0.037 1 0.0085 0.027 0.021 0.0085 0.055 0.033 0.0015
HOUR_APPR_PROCESS_START -0.03 0.074 0.037 0.033 0.13 0.051 0.026 -0.03 0.0085 1 0.052 0.067 0.054 0.011 -0.006 -0.011
REG_REGION_NOT_LIVE_REGION -0.023 0.078 0.015 0.033 -0.025 0.059 0.065 0.018 0.027 0.052 1 0.46 0.09 0.34 0.14 0.0035
REG_REGION_NOT_WORK_REGION -0.015 0.16 0.042 0.071 0.032 0.038 0.087 0.015 0.021 0.067 0.46 1 0.86 0.15 0.22 0.18
LIVE_REGION_NOT_WORK_REGION -0.0056 0.15 0.045 0.069 0.057 0.013 0.064 0.0077 0.0085 0.054 0.09 0.86 1 0.015 0.17 0.22
REG_CITY_NOT_LIVE_CITY 0.0023 -0.001 -0.041 -0.02 -0.05 0.17 0.12 0.038 0.055 0.011 0.34 0.15 0.015 1 0.44 0.012
REG_CITY_NOT_WORK_CITY 0.0075 -0.014 -0.037 -0.024 -0.035 0.11 0.13 0.047 0.033 -0.006 0.14 0.22 0.17 0.44 1 0.82
LIVE_CITY_NOT_WORK_CITY 0.013 -0.0048 -0.011 -0.0081 -0.0073 0.029 0.07 0.027 0.0015 -0.011 0.0035 0.18 0.22 0.012 0.82 1
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
HOUR_APPR_PROCESS_START
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
1.0
0.8
0.6
0.4
0.2
0.0


Correlation for target 1
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
HOUR_APPR_PROCESS_START
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
1.0
0.8
0.6
0.4
0.2
0.0
-0.2

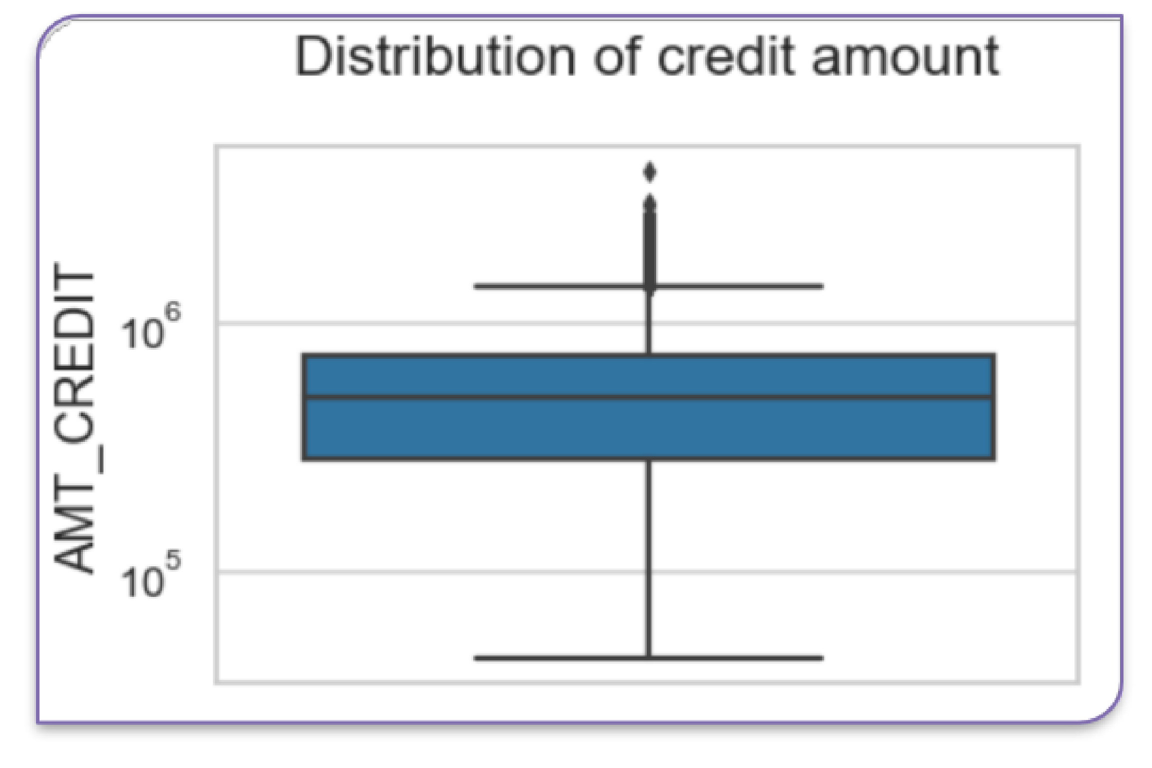
Categorical Univariate analysis of variables for Target = 0

Boxplot for credit amount

Based on the analysis, it can be concluded that there are some noticeable outliers in the credit amount data. These outliers might represent instances where clients have received significantly higher or lower credit amounts compared to the majority of the clients.

Additionally, the observation that the first quartile (25th percentile) of the credit amount distribution is larger than the third quartile (75th percentile) implies that a significant portion of the credits granted to clients falls within the lower range. In other words, most clients tend to receive smaller credit amounts, while fewer clients receive larger credit amounts.

This conclusion could have implications for decision-making within the context of credit assessment or risk management, as it suggests that the majority of clients may be eligible for smaller credit amounts based on historical data. Further analysis and investigation might be needed to understand the reasons behind these patterns and whether they hold consistent over time.

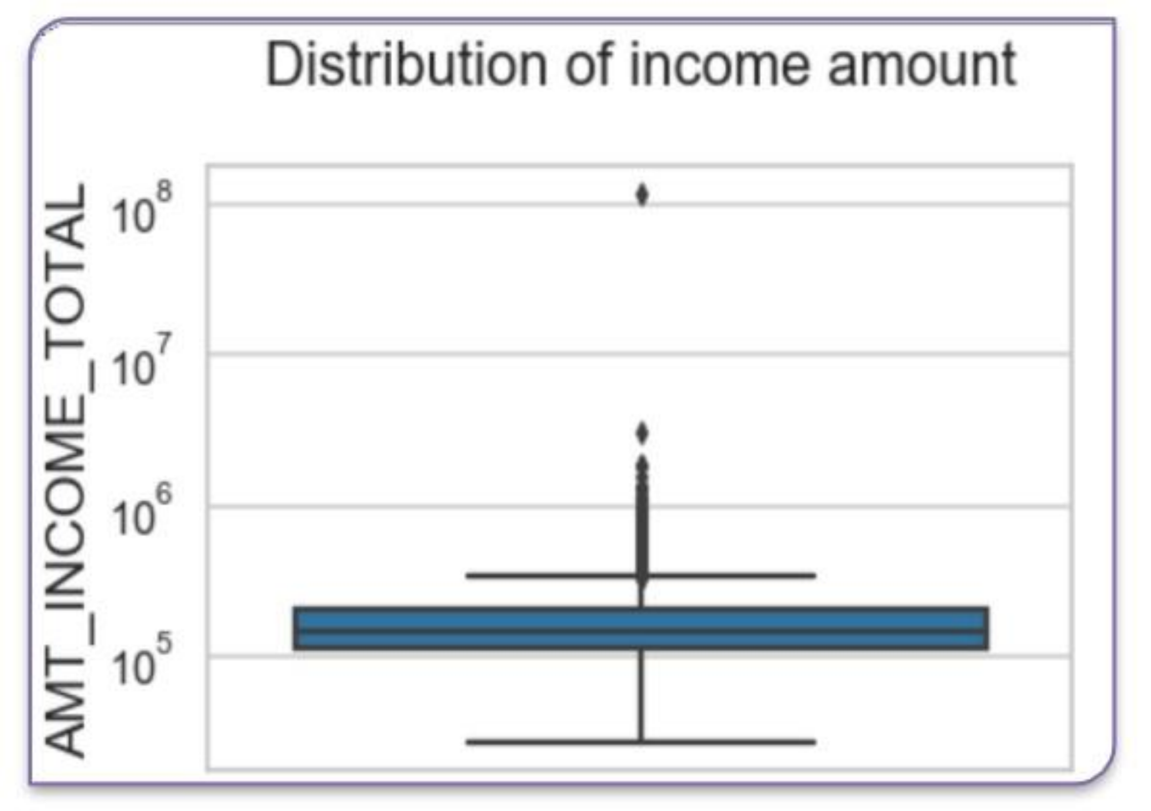


Boxplot for income amount

Based on the analysis, it can be concluded that there are outliers present in the income amount data. These outliers might represent cases where individuals have significantly higher or lower income amounts compared to the majority of the population.

Additionally, the observation that the third quartile (75th percentile) of the income amount distribution is very slim suggests that a majority of individuals in the population have relatively consistent and lower income amounts. In other words, most individuals seem to earn similar income levels up to the 75th percentile, beyond which there are relatively fewer individuals with significantly higher income amounts.

This conclusion could have implications for understanding the income distribution within the analyzed population. It might indicate a relatively uniform distribution of income for the majority of individuals, with only a smaller portion of the population having notably higher income levels. Further exploration and analysis could help uncover the underlying factors contributing to these income distribution patterns.

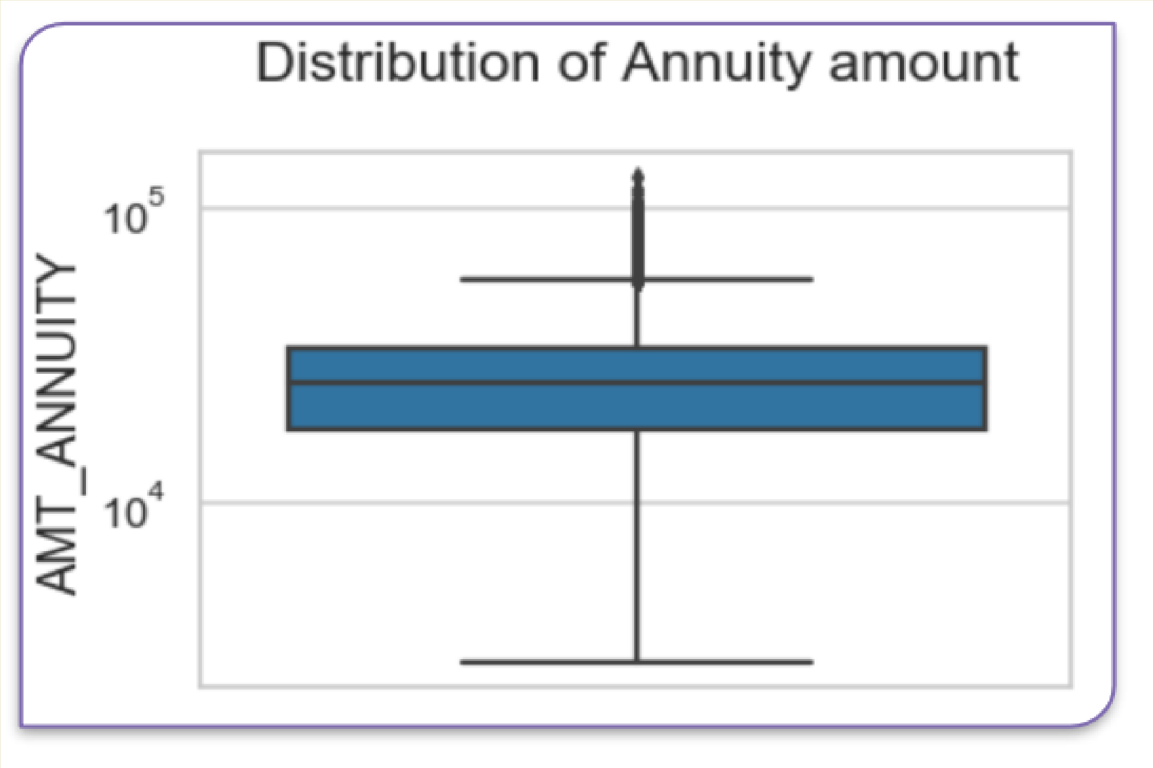


Boxplot for annuity amount

Based on the analysis, it can be concluded that there are outliers present in the annuity amount data. These outliers could represent instances where clients have received significantly higher or lower annuity amounts compared to the majority of the clients.

Furthermore, the observation that the first quartile (25th percentile) of the annuity amount distribution is larger than the third quartile (75th percentile) indicates that a majority of the annuity clients fall within the lower annuity payment range. In other words, most clients seem to have lower annuity amounts, while fewer clients have higher annuity amounts.

This conclusion could be significant for understanding the distribution of annuity amounts among clients. It suggests that the majority of clients are associated with lower annuity payments, with only a smaller portion of clients receiving higher annuity payments. Further analysis and investigation could shed light on the reasons behind these distribution patterns and whether they have remained consistent over time.



Categorical Univariate analysis of variables for Target = 1

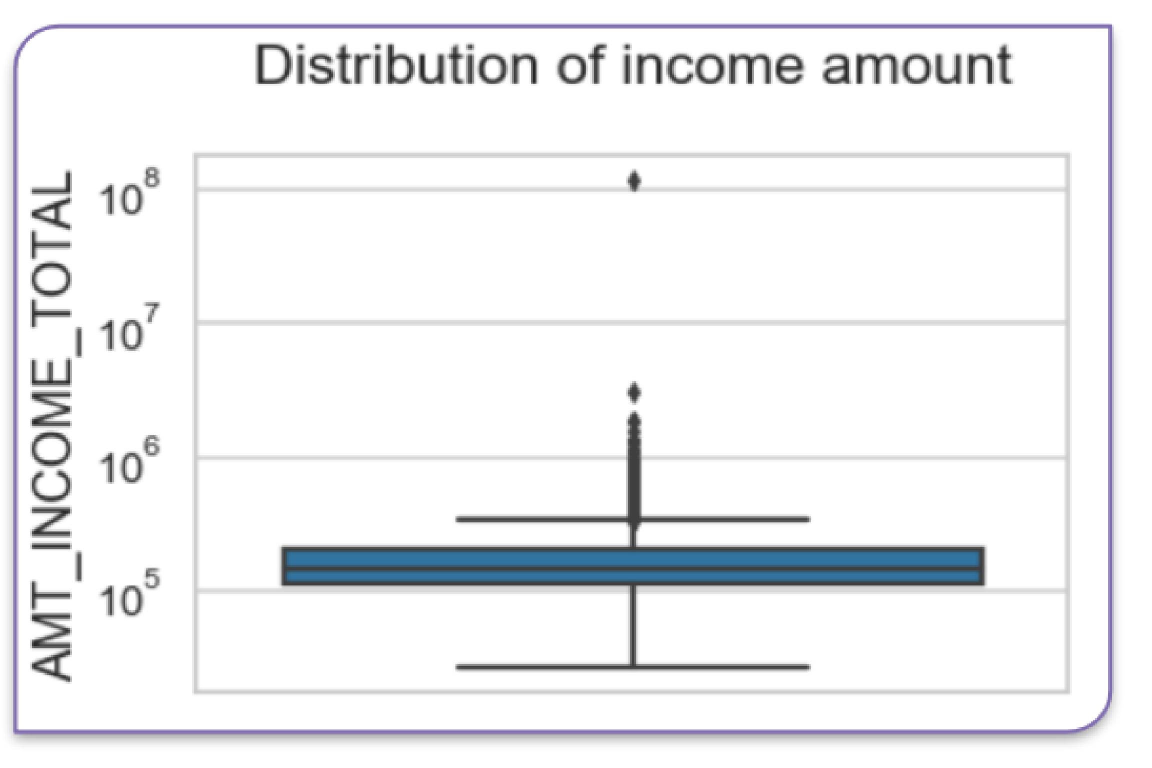
Boxplot for income amount

Based on the analysis, several key conclusions can be drawn:

Outliers in Income Amount: The presence of outliers in the income amount data suggests that there are individuals with significantly higher or lower income amounts compared to the majority of the population. These outliers might represent unique or exceptional cases.

Slim Third Quartile for Income Amount: The observation that the third quartile (75th percentile) of the income amount distribution is slim indicates that the majority of individuals in the population earn similar income levels up to this point. This suggests that a substantial portion of the population has relatively consistent income amounts.

Concentration in First Quartile: The fact that most of the clients' income amounts are present in the first quartile (25th percentile) implies that a significant majority of clients have lower income levels. This could suggest an unequal distribution of income, with a large portion of clients falling into the lower-income category.



Boxplot for credit amount

Based on the analysis, the following conclusions can be drawn:

Outliers in Credit Amount: The presence of outliers in the credit amount data indicates that there are instances where clients have received credit amounts that deviate significantly from the majority of the clients. These outliers could represent unique or exceptional cases with unusually high or low credit amounts.

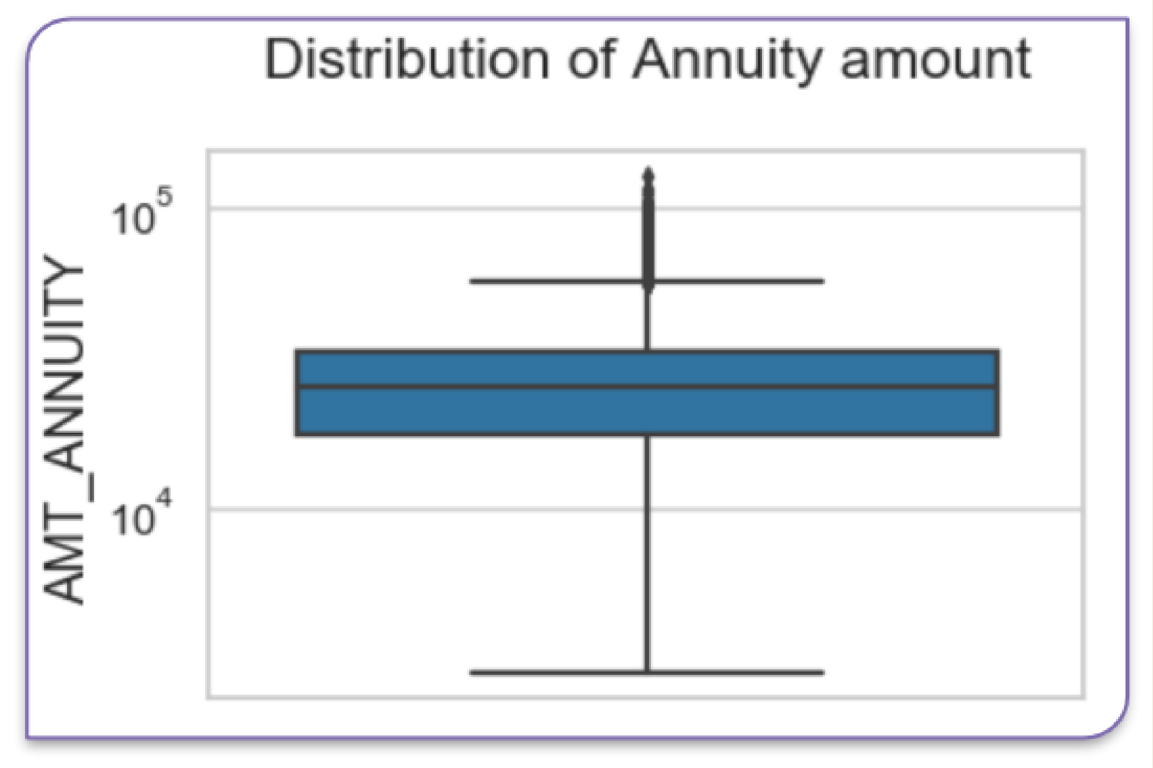
First Quartile vs. Third Quartile for Credit Amount: The observation that the first quartile (25th percentile) of the credit amount distribution is larger than the third quartile (75th percentile) suggests that the majority of clients have received credit amounts that fall within the lower range. In other words, most clients seem to have obtained smaller credit amounts, while fewer clients have received larger credit amounts.

Boxplot for annuity amount

Based on the analysis, the following conclusions can be made:

Outliers in Annuity Amount: The presence of outliers in the annuity amount data suggests that there are instances where clients have been associated with annuity amounts that deviate significantly from the majority of the clients. These outliers might represent exceptional cases with unusually high or low annuity amounts.

First Quartile vs. Third Quartile for Annuity Amount: The observation that the first quartile (25th percentile) of the annuity amount distribution is larger than the third quartile (75th percentile) indicates that most of the annuity clients fall within the lower range of annuity amounts. In other words, a significant majority of clients seem to have lower annuity payments, while fewer clients have higher annuity payments.



Bivariate analysis for Target = 0

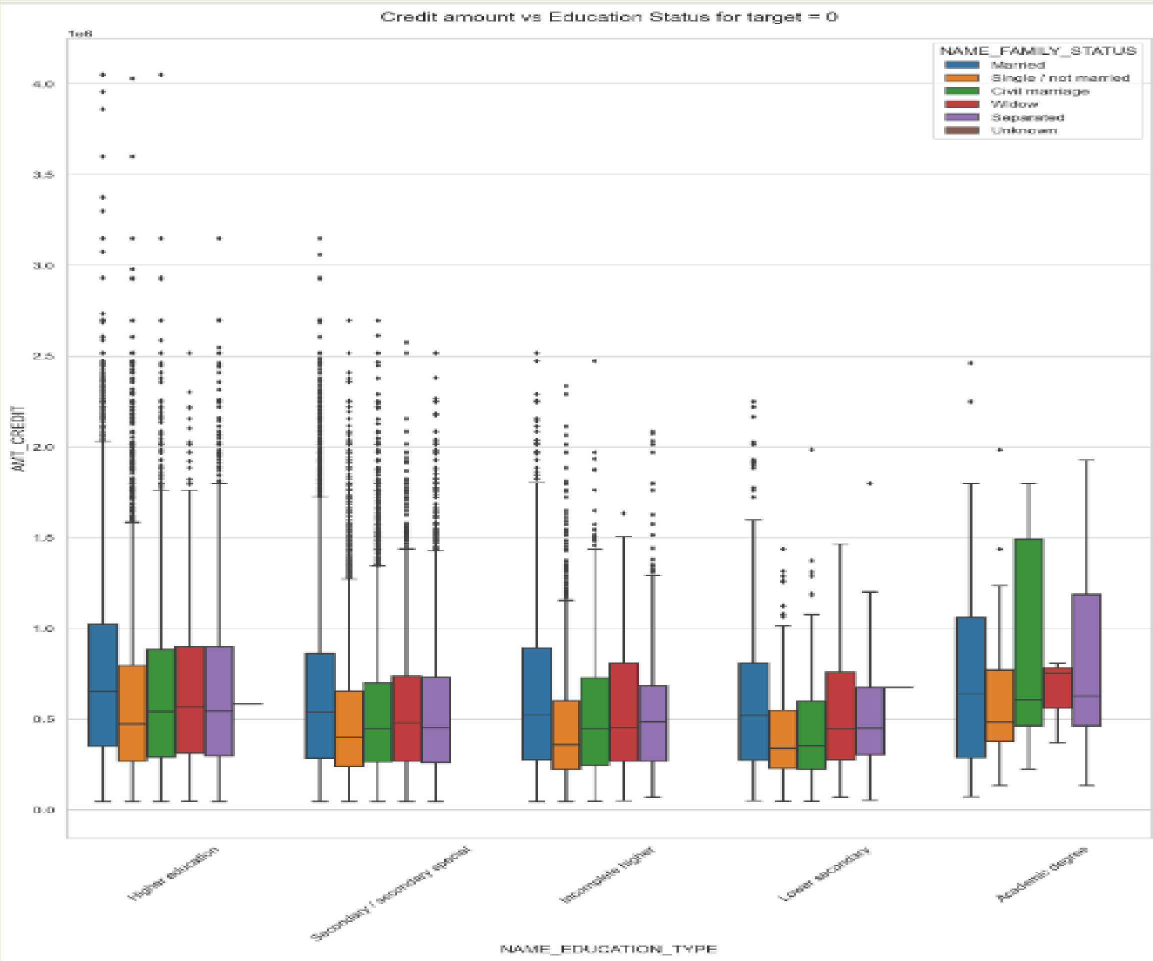
Credit amount vs Education Status

Key insights derived from the presented box plot analysis are as follows:

Family Status and Education: Individuals with family statuses labeled as 'civil marriage,' 'marriage,' and 'separated,' who possess an 'Academic degree' education level, demonstrate a higher count of credits compared to other categories.

Education and Outliers: Notably, 'Higher education' among individuals with family statuses of 'marriage,' 'single,' and 'civil marriage' exhibits a higher frequency of outliers in credit distribution. This suggests potential variations or anomalies within this subgroup.

Civil Marriage and Academic Degree: Specifically, the combination of 'civil marriage' within the 'Academic degree' category stands out with a significant concentration of credits in the third quartile.



Income amount vs Education Status

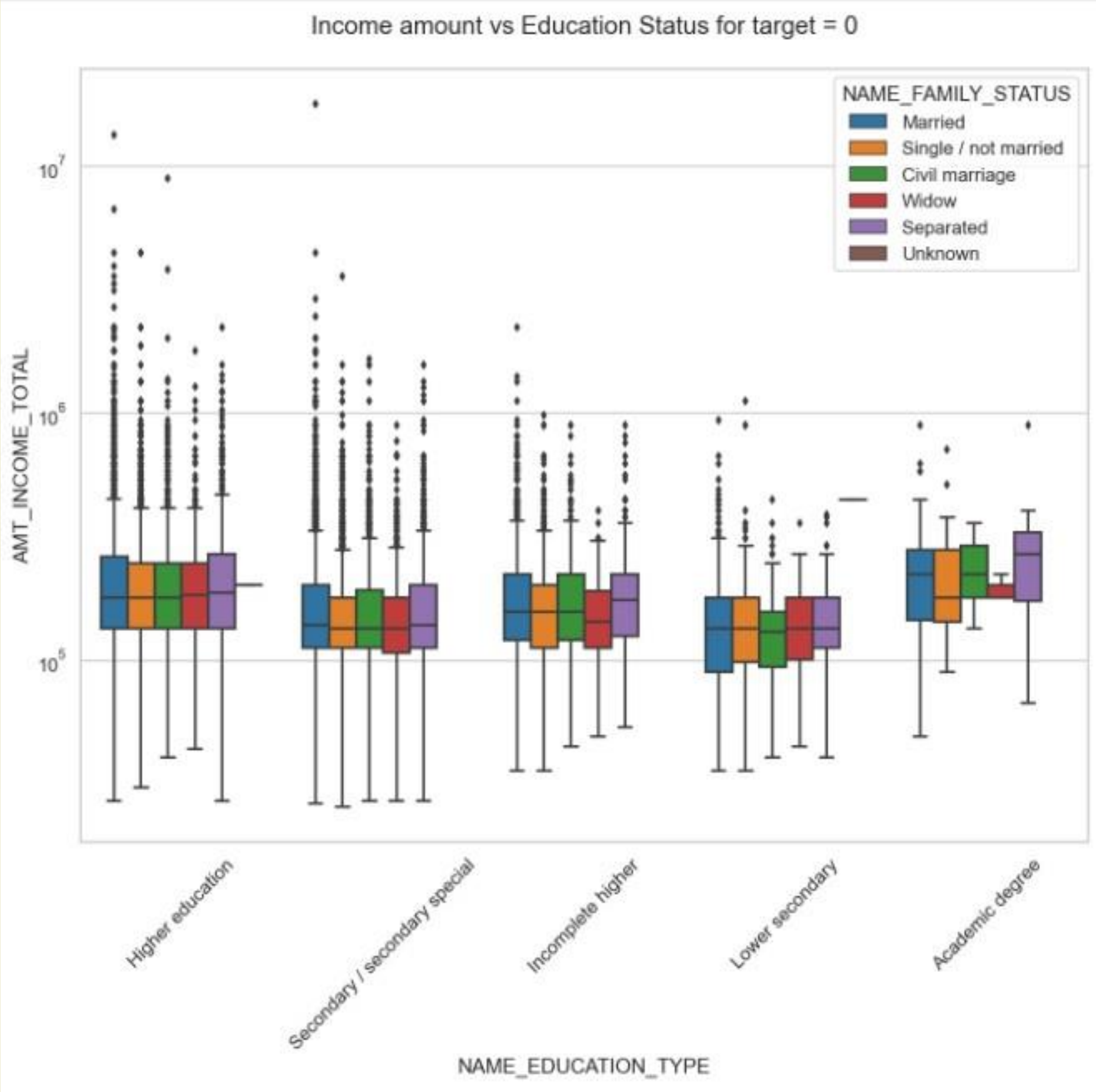
Key insights extracted from the provided box plot analysis for the 'Higher education' category are as follows:

Income and Family Status: For individuals with 'Higher education,' the income amounts are notably consistent across different family statuses. This suggests that education level is a stronger determinant of income than family status within this category.

Outliers: The presence of numerous outliers in the 'Higher education' group indicates potential variations or anomalies in income within this education level. These outliers might reflect exceptional cases with significantly higher or lower incomes.

Comparison with Academic Degree: Individuals with 'Academic degree' education show fewer outliers compared to 'Higher education.' However, their income amounts tend to be slightly higher.

Lower Secondary Education and Civil Marriage: Individuals with 'Lower secondary education' and a family status of 'civil marriage' demonstrate relatively lower income amounts compared to other combinations of education and family status.

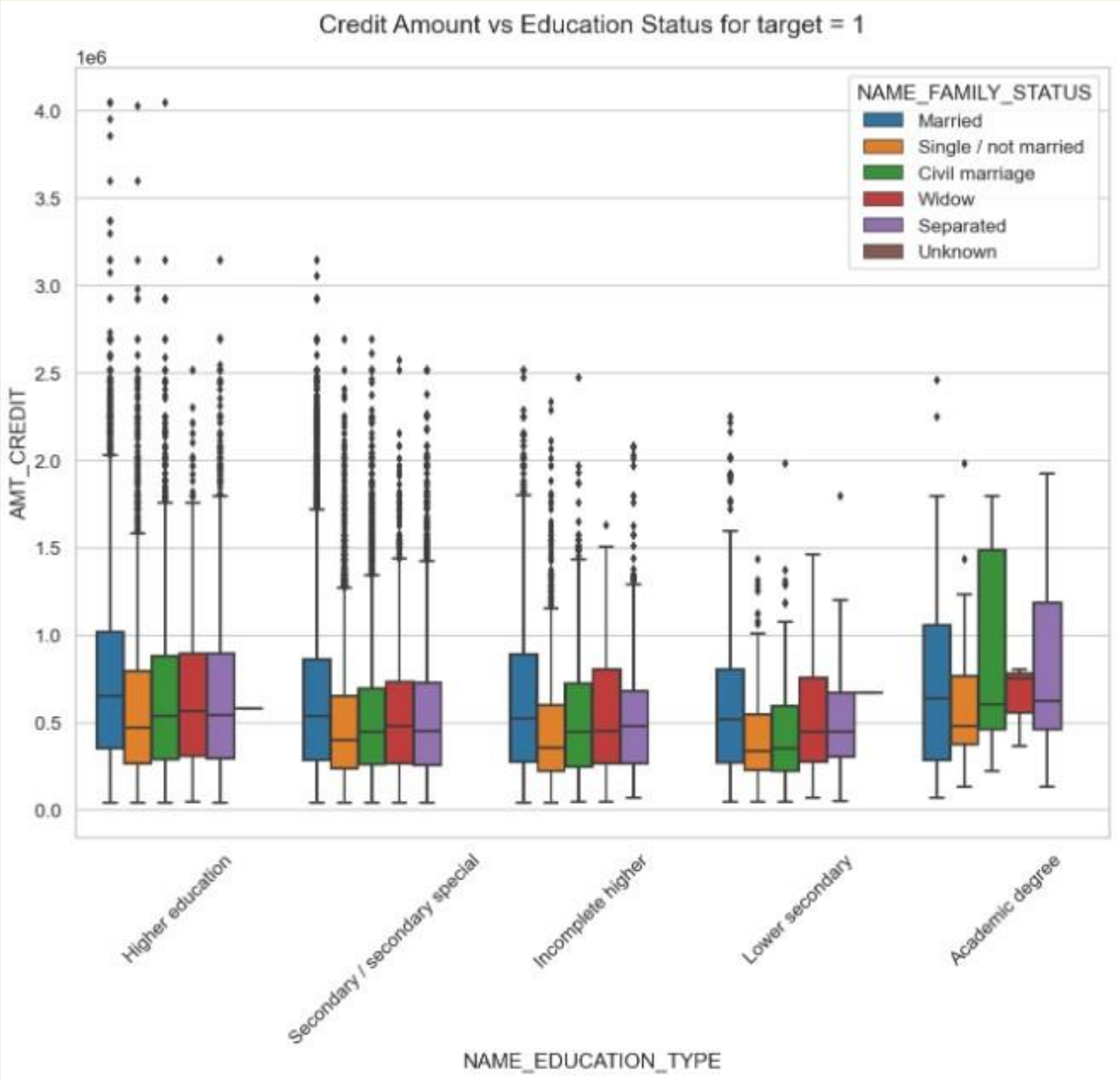


Bivariate analysis for Target = 1

Credit amount vs Education Status

The box plot analysis for 'Target = 1' appears to exhibit similarities to 'Target = 0.'

Family Status and Education: Similar to 'Target = 0,' individuals with family statuses 'civil marriage,' 'marriage,' and 'separated,' coupled with an 'Academic degree' education, tend to have a higher count of credits compared to other categories.
Outliers and Education: 'Higher education' and 'Secondary' education levels contain a significant number of outliers, indicating potential data anomalies or deviations from the norm in these categories.
Academic Degree and Civil Marriage: As observed in 'Target = 0,' the combination of 'civil marriage' within the 'Academic degree' category demonstrates a notable concentration of credits in the third quartile.



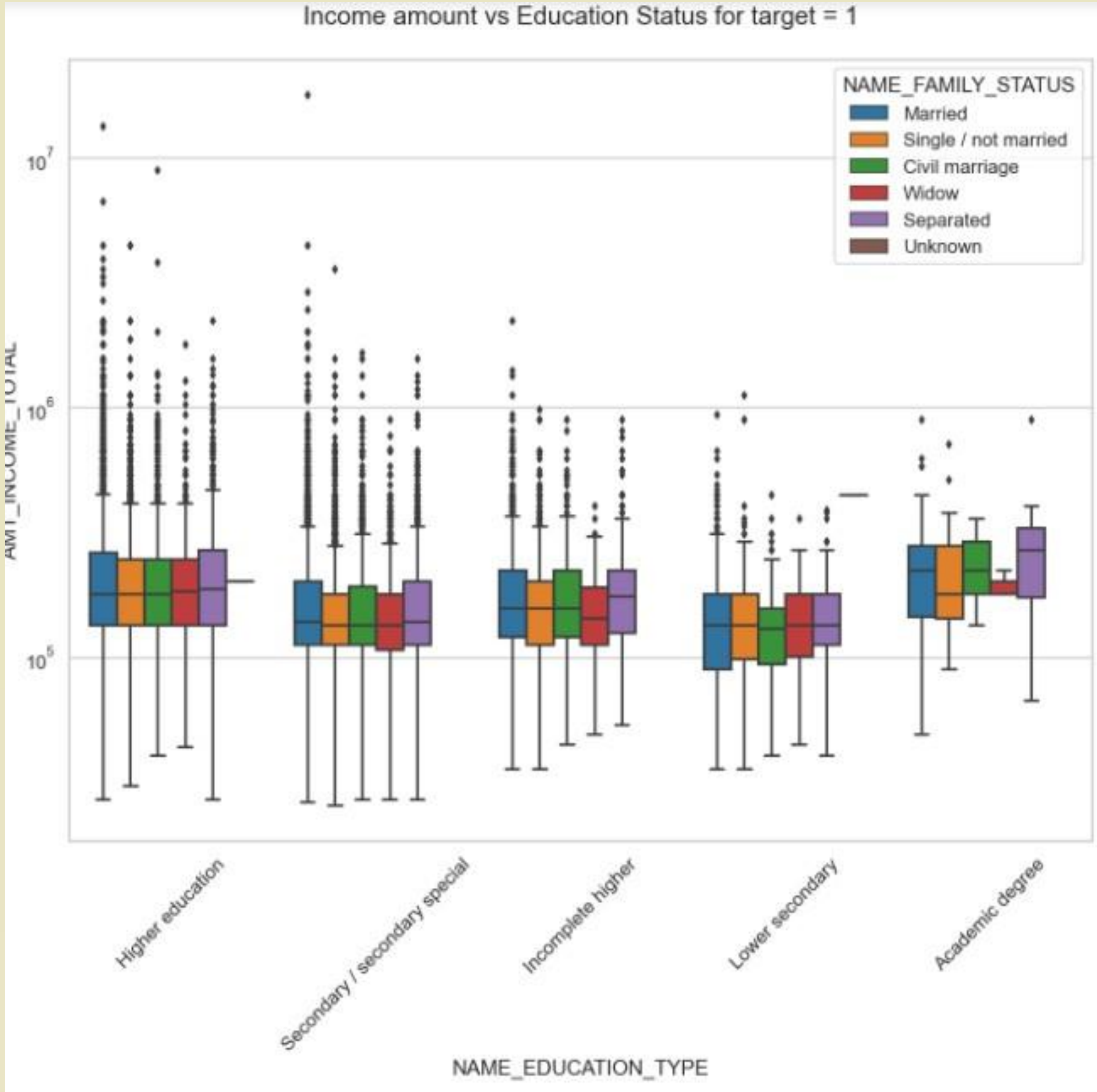
Income amount vs Education Status

Similarities with 'Target = 0' are evident in the box plot analysis.

Education and Family Status: As in 'Target = 0,' within the 'Higher education' category, income amounts are relatively consistent across various family statuses. This underscores the notion that education level is a stronger determinant of income within this subgroup.

Outliers: 'Higher education' experiences fewer outliers in comparison to 'Academic degree.' However, those with 'Academic degree' education showcase slightly higher income levels.

Lower Secondary Education: Individuals with 'Lower secondary education' exhibit lower income amounts compared to other education categories.



Univariate analysis (after merging data from previous application)

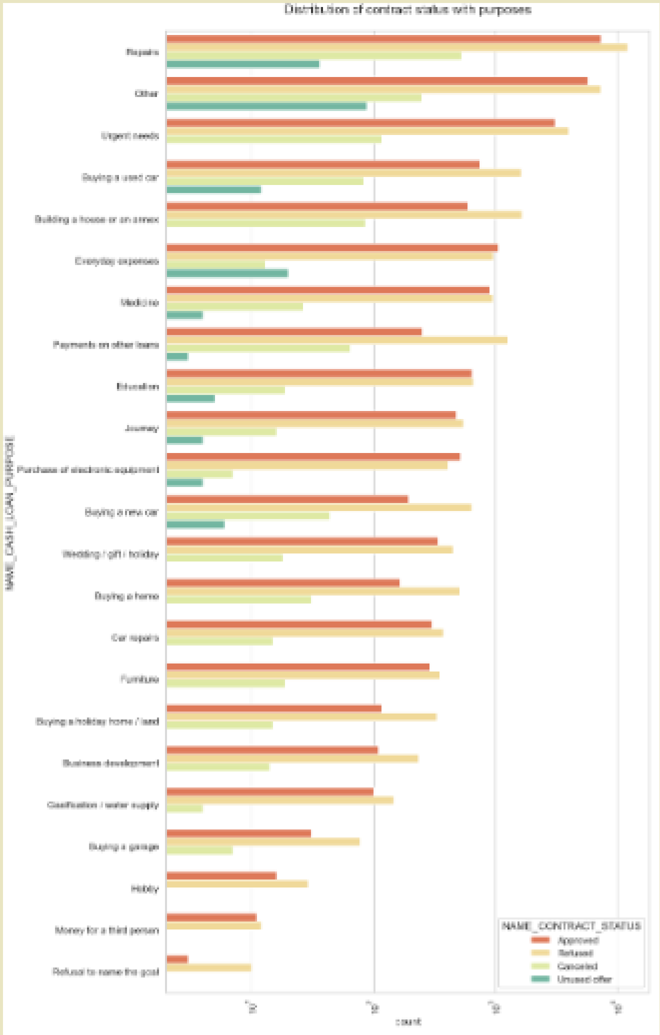
Distribution of contract status with purposes

Key observations inferred from the provided plot include:

Loan Rejection: The highest number of loan rejections stem from the purpose labeled 'repairs.'

Education Loan: The count of approvals and rejections for education-related loans appears to be balanced, suggesting that the decision on these loans is not significantly biased toward either outcome.

Loan Purposes with Significant Rejections: Loan purposes related to 'paying other loans' and 'buying a new car' exhibit notably higher counts of loan rejections compared to approvals.

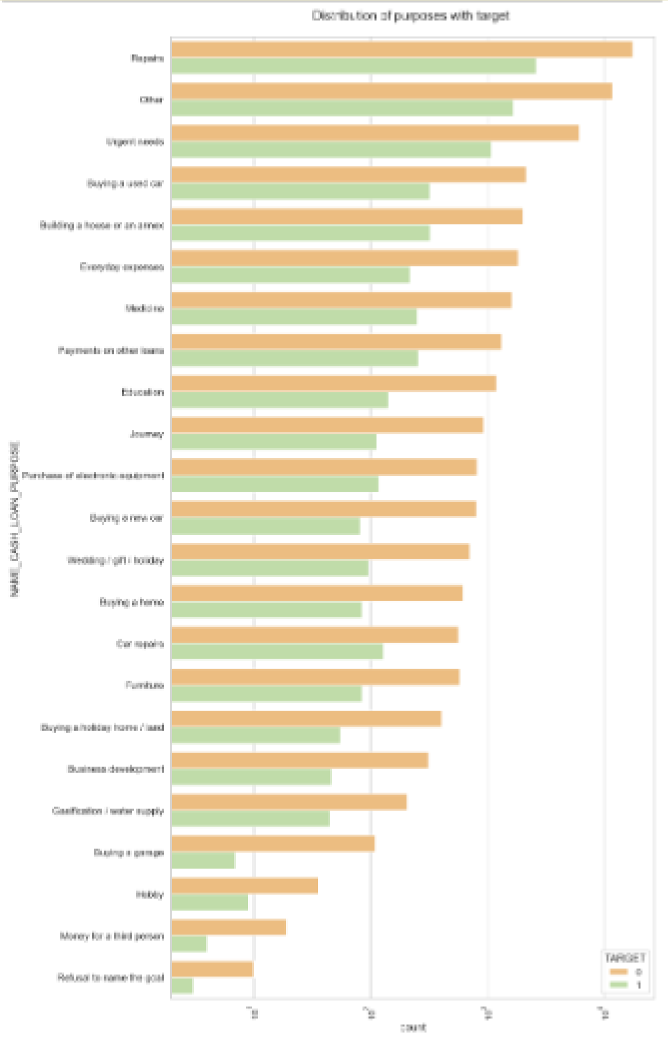


Distribution of purposes with target

Additional insights that can be drawn from the provided plot are as follows:

Loan Difficulties and Purpose: Loans intended for 'Repairs' encounter the greatest challenges in on-time payment, highlighting the payment difficulties associated with this particular loan purpose.

Successful Payment Purposes: Conversely, several loan purposes exhibit notably higher instances of successful on-time payments. These purposes include 'Buying a garage,' 'Business development,' 'Buying land,' 'Buying a new car,' and 'Education.'



Bivariate analysis

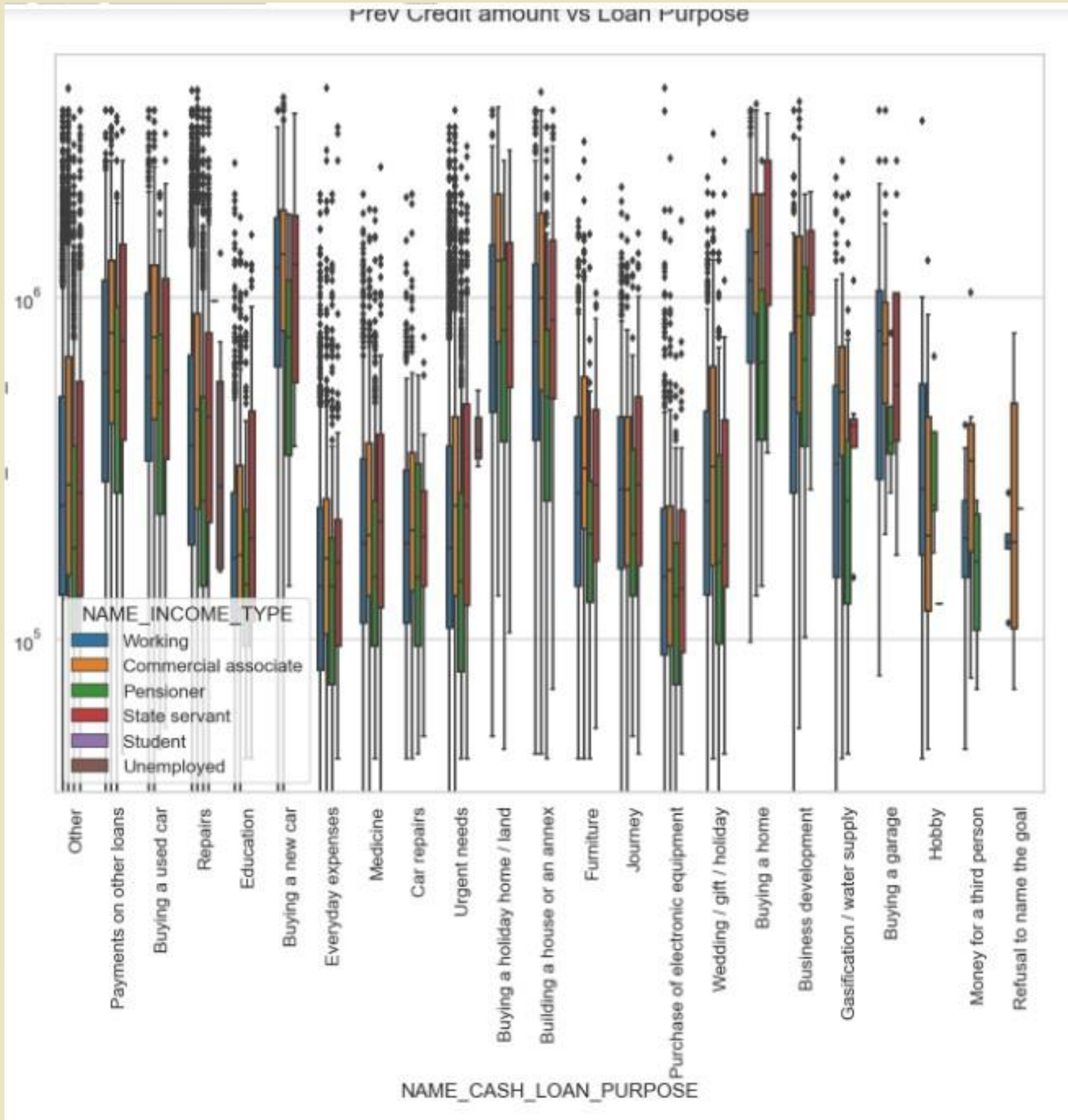
Income amount vs Education Status

Key conclusions that can be drawn from the provided information are:

Loan Purpose and Credit Amount: Loan purposes such as 'Buying a home,' 'Buying land,' 'Buying a new car,' and 'Building a house' are associated with higher credit amounts. These purposes likely involve substantial financial commitments, leading to larger credit applications.

Credit Application by State Servants: Individuals categorized as state servants exhibit a notable presence in terms of applying for significant credit amounts. This suggests that state servants are more likely to seek higher credit sums.

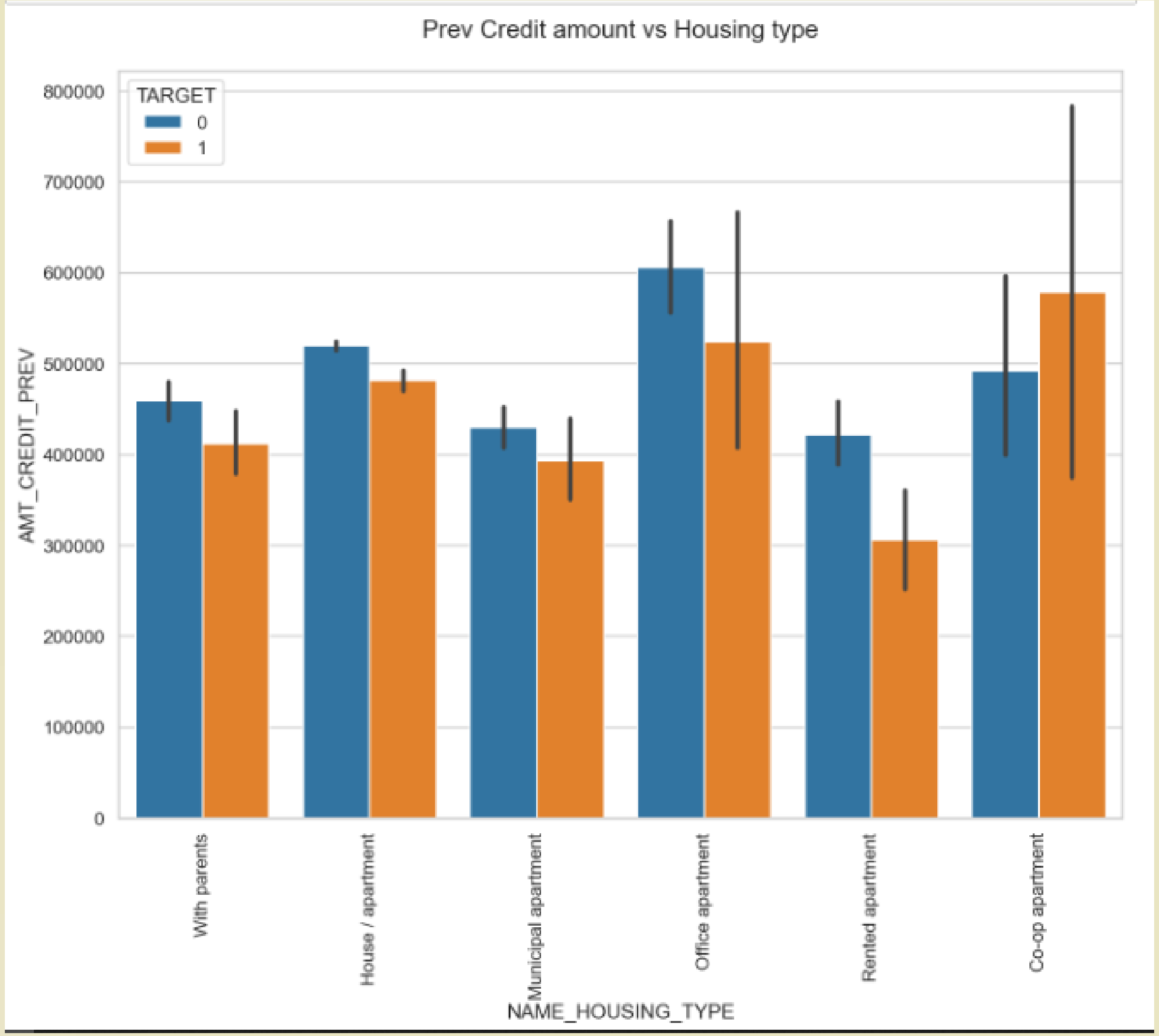
Limited Credit Applications for Third Person or Hobby: Purposes related to borrowing money for a third person or engaging in hobbies reflect lower credit application counts. These activities might involve smaller financial transactions compared to other loan intents.



Income amount vs Education Status

The analysis of housing types and their corresponding credit behaviors reveals valuable insights:

Target = 0 - Credit Behavior for Housing Types: In cases where the target variable is '0,' housing types like 'office apartment' exhibit higher credit activity. This suggests a propensity for successful payment among applicants with 'office apartments' in this target group.
Target = 1 - Credit Behavior for Housing Types: Conversely, when the target variable is '1,' the 'co-op apartment' housing type demonstrates elevated credit utilization. This indicates a potential difficulty in payment among applicants with 'co-op apartments' in this target group.



Final Remarks

Based on the provided information, the following strategies can be considered for focusing on specific groups of clients to ensure successful payments:

1. **Targeting Specific Income Types:** Focus on clients with income types 'Student', 'Pensioner', and 'Businessman' who have housing types other than 'Co-op apartment'. These income types might represent individuals with more stable financial situations, and by selecting housing types other than 'Co-op apartment', you could be narrowing down to those who are more likely to have stable living conditions.
2. **Priority on Housing Type 'With Parents':** Prioritize clients with the housing type 'With parents'. Since these clients have shown the least number of unsuccessful payments, they might be a reliable and low-risk group to target. It's possible that the financial stability provided by living with parents contributes to their ability to make successful payments.

Based on the provided insights, here are the recommended actions for banks:

Reduce Focus on Income Type 'Working': Considering that clients with the income type 'Working' have the highest number of unsuccessful payments, it might be beneficial for banks to reduce their focus on this particular income group. This could involve implementing stricter assessment criteria for clients with this income type or providing them with additional financial education and support to improve their payment performance.

Loan Purpose 'Repair': Since the loan purpose 'Repair' is associated with a higher number of unsuccessful payments, banks could consider conducting a more thorough risk assessment for clients seeking loans for repair purposes. This could involve evaluating the financial stability of clients with this loan purpose more rigorously and potentially offering alternative loan terms or conditions to mitigate the risk of default.