CREDIT EDA ASSIGNMENT

VARUNPRAKASH SHANMUGAM

Introduction

Objective

When evaluating a loan application, the company faces two types of risks that influence its decision:

1.Risk of lost business: If the applicant is likely to repay the loan but the loan is not approved, the company misses out on potential business and revenue.

2.Risk of financial loss: If the applicant is not likely to repay the loan and is at risk of default, approving the loan may result in a financial loss for the company.

Datasets

We have used total **2 datasets** they are,

1. Current Applications
2. Previous Applications

Index

1. Importing Data
2. Inspecting Data
3. Data Analysis
4. Checking Data
5. Analysis for categories in column
6. Analysis of numerical in column
7. Correlations
8. Loading previous applications
9. Conclusion
10. Recommendation

Description

Loan providers often face difficulties when granting loans to individuals who have insufficient or non-existent credit history. This situation creates an opportunity for some consumers to exploit it by intentionally defaulting on their loans.

As an employee of a consumer finance company that focuses on providing loans to urban customers, your task involves conducting exploratory data analysis (EDA) to examine patterns within the data. This analysis aims to ensure that applicants who can repay the loans are not unfairly rejected, thereby mitigating the risks associated with loan defaults.

The objective of this case study is to detect patterns that can indicate whether a client will encounter challenges in making loan instalments. These patterns can be utilized to make informed decisions, such as denying the loan, adjusting the loan amount, or offering loans to higher-risk applicants at a higher interest rate. By employing exploratory data analysis (EDA), the aim is to identify such applicants and ensure that those who are capable of repaying the loan are not unjustly rejected. The primary focus of this case study is to leverage EDA for the identification of these applicants.

Importing Data 1

To begin, our analysis involves utilizing the **'Application Data** Set' which encompasses all the relevant client information available at the time of application. This dataset primarily focuses on discerning whether a client encounters any challenges with payments.

Data cleaning

To prepare the data for analysis, we engage in the process of data cleaning. The following steps are executed during this stage:

Initially, we identify the essential columns required for our analysis, and subsequently, we endeavor to identify any Empty values present within each column.

Importing Data 2

In the present stage, we have acquired the second dataset, namely the **'Previous Application** Data Set,' for our analysis. This dataset encompasses relevant information regarding the client's previous loan data. It specifically includes data pertaining to the status of the previous application, whether it was approved, cancelled, refused, or an unused offer.

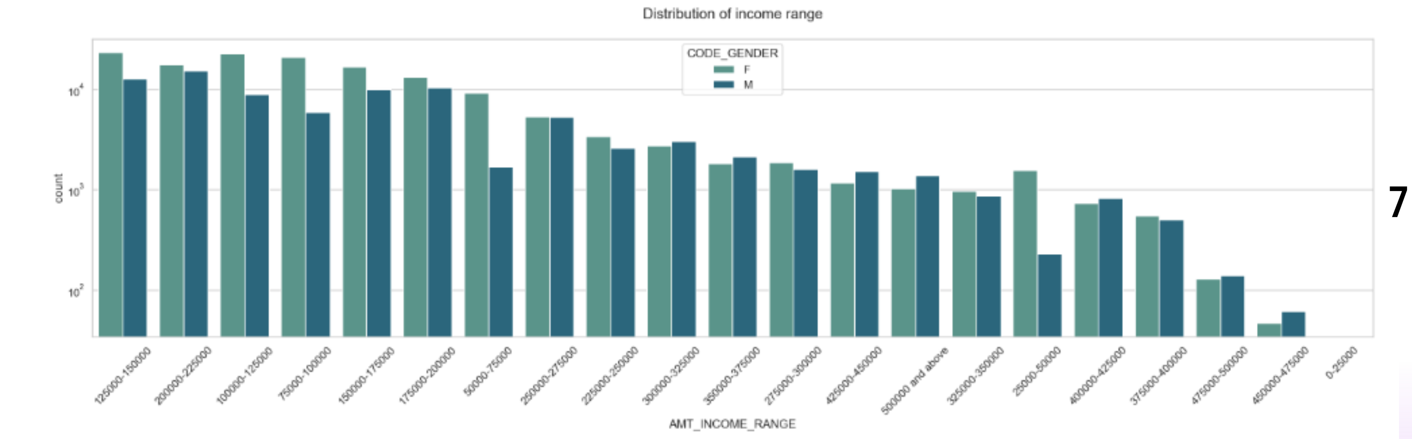
The 'Previous Application Data Set' comprises 1670214 rows and 37 columns.

Likewise, we have undertaken the process of **data cleaning** in this stage:

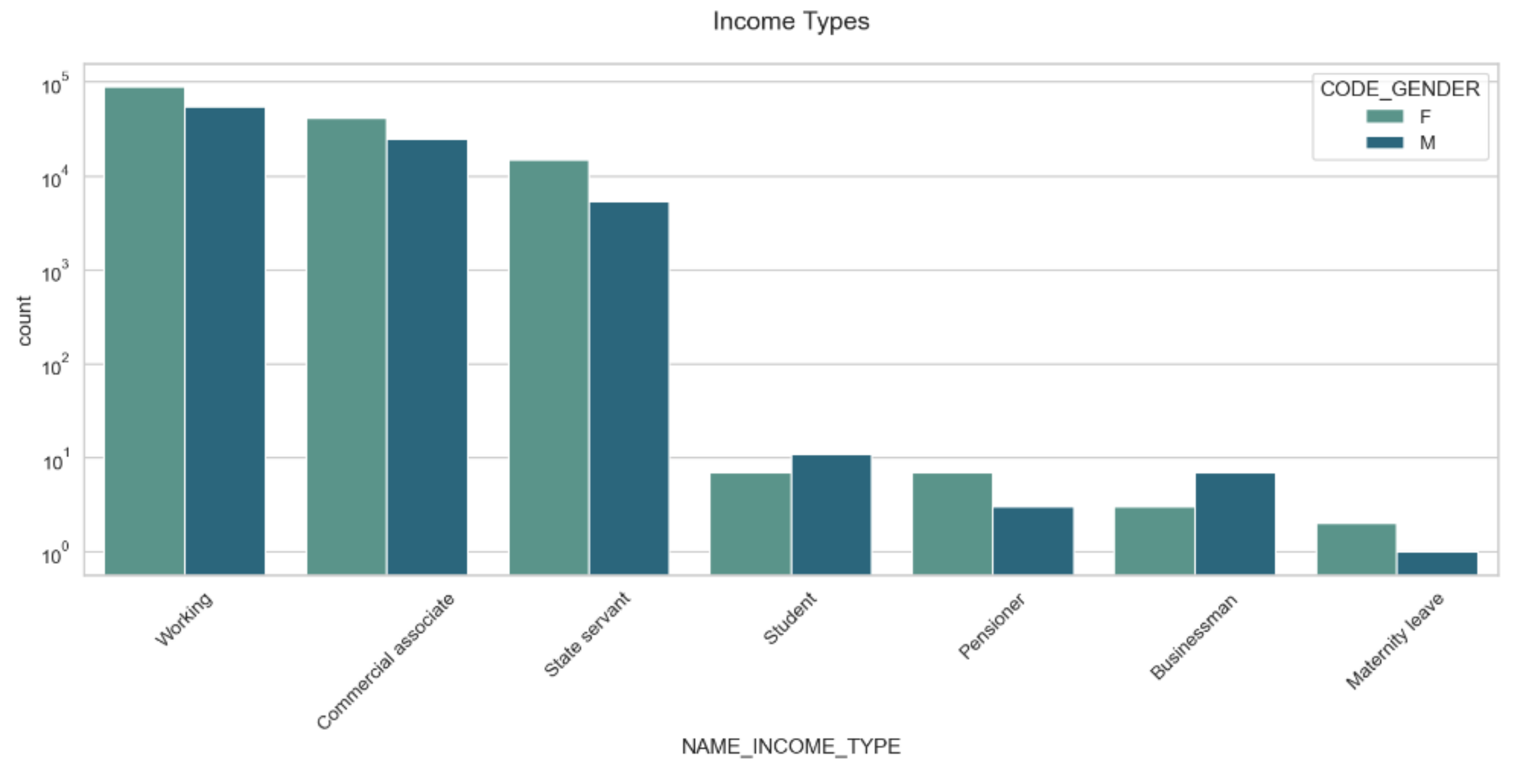
- We examined the percentage of null values present in each column to assess the necessary steps for data cleaning.
- The percentage of null entries in the provided DataFrame was calculated.
- We iterated over the columns in the DataFrame and deleted those where more than 20% of the values were null.
- The percentage of null values in each column was determined.

An outlier refers to an observation that deviates significantly from the other values in a random sample taken from a population. We have identified several values as outliers through the utilization of plots.

The plot focuses on the distribution of income range (AMT_INCOME_RANGE) and includes different colours (hue) based on the gender of individuals (CODE_GENDER). The title of the plot is set as "Distribution of income range." This plot helps visualize how the income range is distributed among different genders within the dataset.

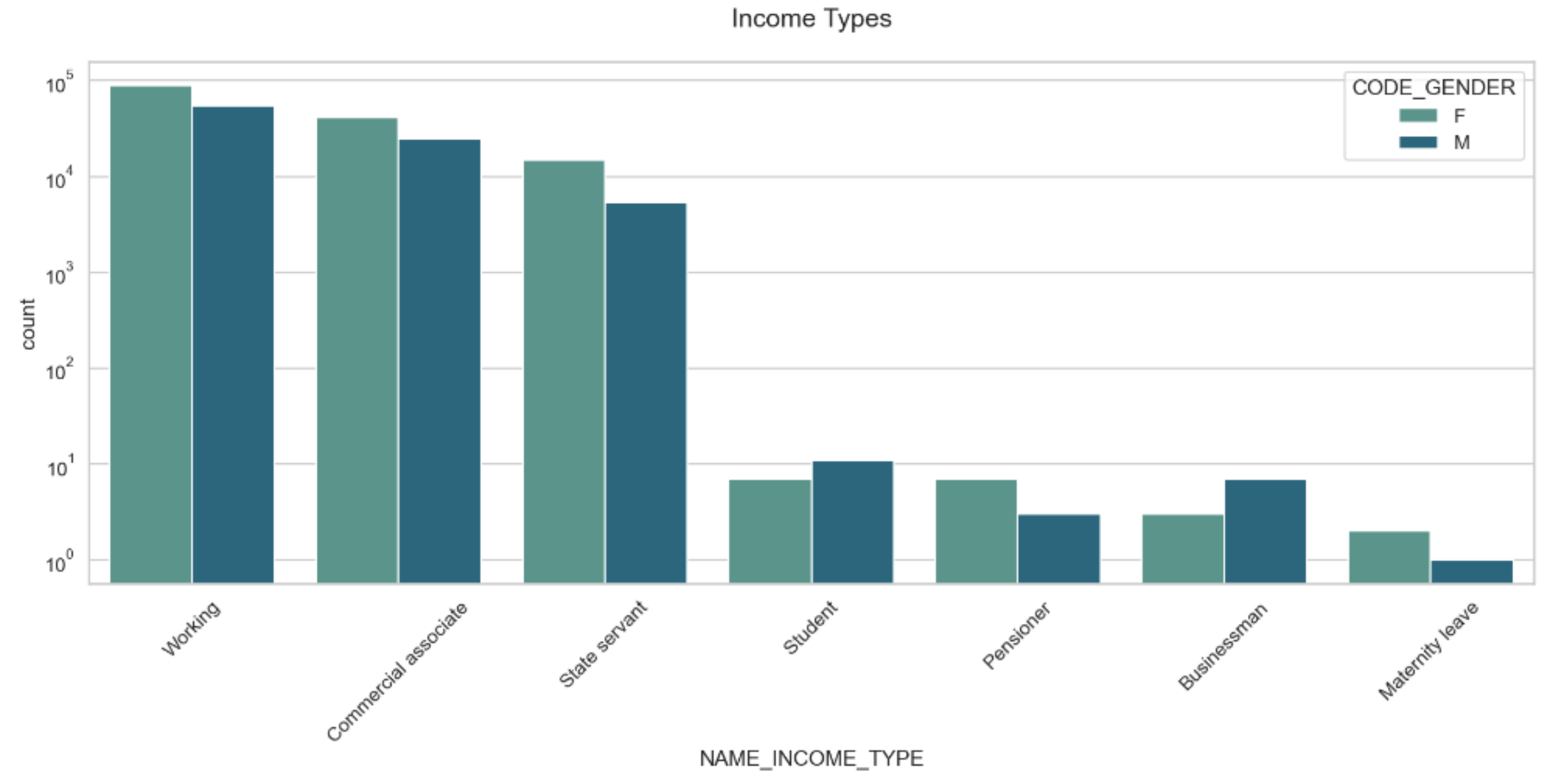


We now produces a univariate plot (uniplot) using the dataframe target0_df. The plot illustrates the distribution of different income types (NAME_INCOME_TYPE) and distinguishes them by colors based on the gender of individuals (CODE_GENDER). The title of the plot is set as "Income Types." This visualization provides insights into how various income types are distributed among different genders within the dataset.



When plotting the count of the TARGET variable, we observe a significant imbalance within the variable.

The number of credits is higher for income types such as 'working,' 'commercial associate,' and 'State Servant' compared to others. Income types such as 'student,' 'pensioner,' 'businessman,' and 'maternity leave' have a lower number of credits compared to others.

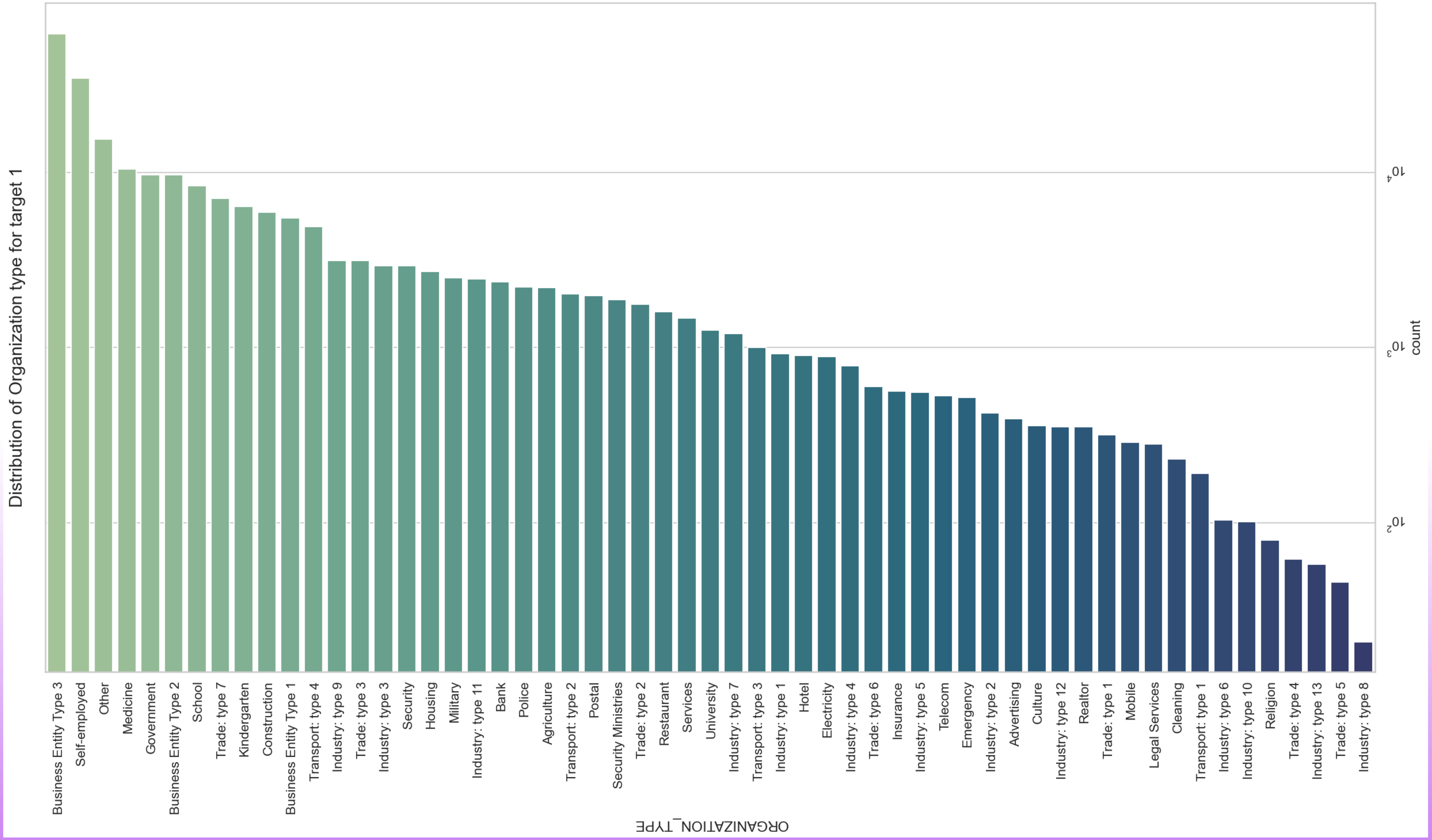


The given analysis Sets the style of the plot to a white grid (whitegrid) and adjusts the context to a talk level (set_context). It then creates a figure with a size of 15x30 inches (figsize). The label and title sizes are increased (axes.labelsize and axes.titlesize) and extra padding is added to the title (axes.titlepad).

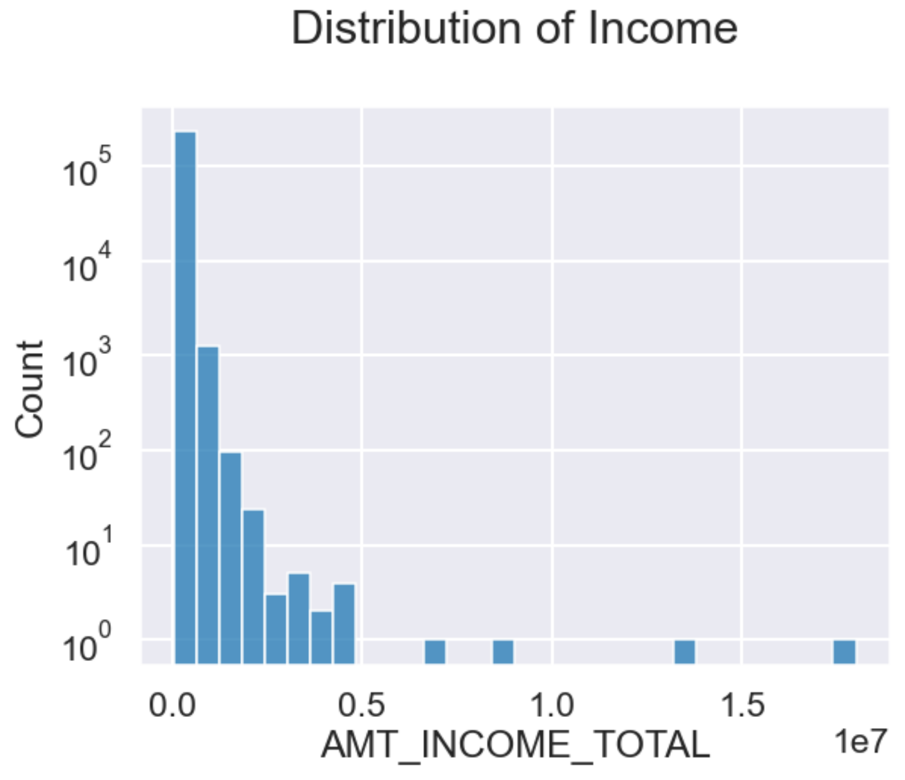
The plot focuses on the distribution of organization types for target 1 (target0_df). The y-axis represents the organization types (ORGANIZATION_TYPE), while the x-axis displays the count of occurrences. The organization types are sorted in descending order based on their counts (order=target0_df['ORGANIZATION_TYPE'].value_counts().index). The color palette used is 'crest’.

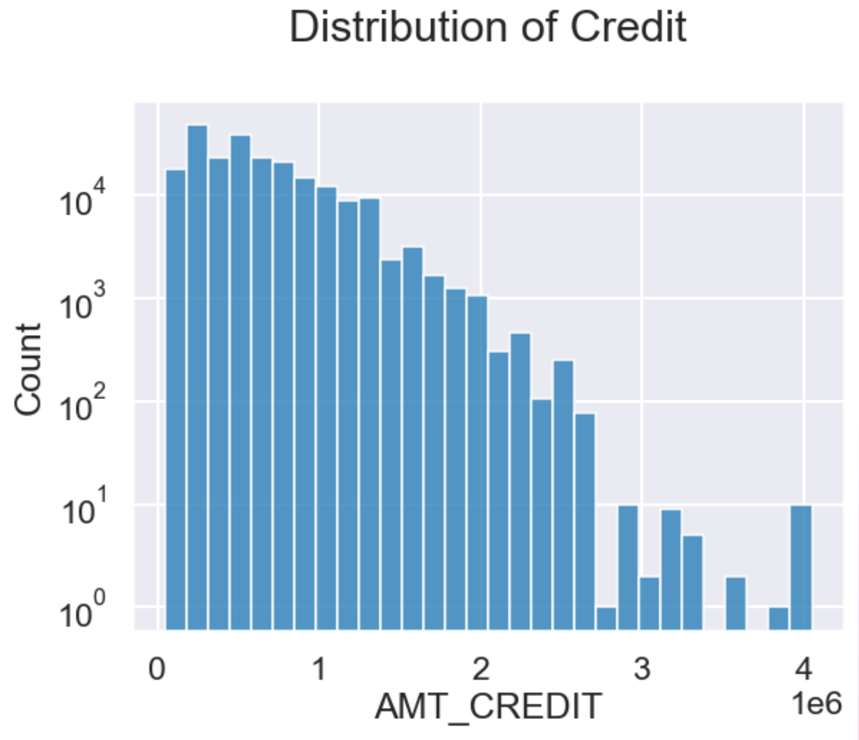
The plot is displayed with the title "Distribution of Organization type for target 1" and the x-axis labels are rotated by 90 degrees for better readability. The x-axis scale is set to logarithmic (xscale('log')).

ANALYSIS OF THE ORGANIZATION TYPE DISTRIBUTION



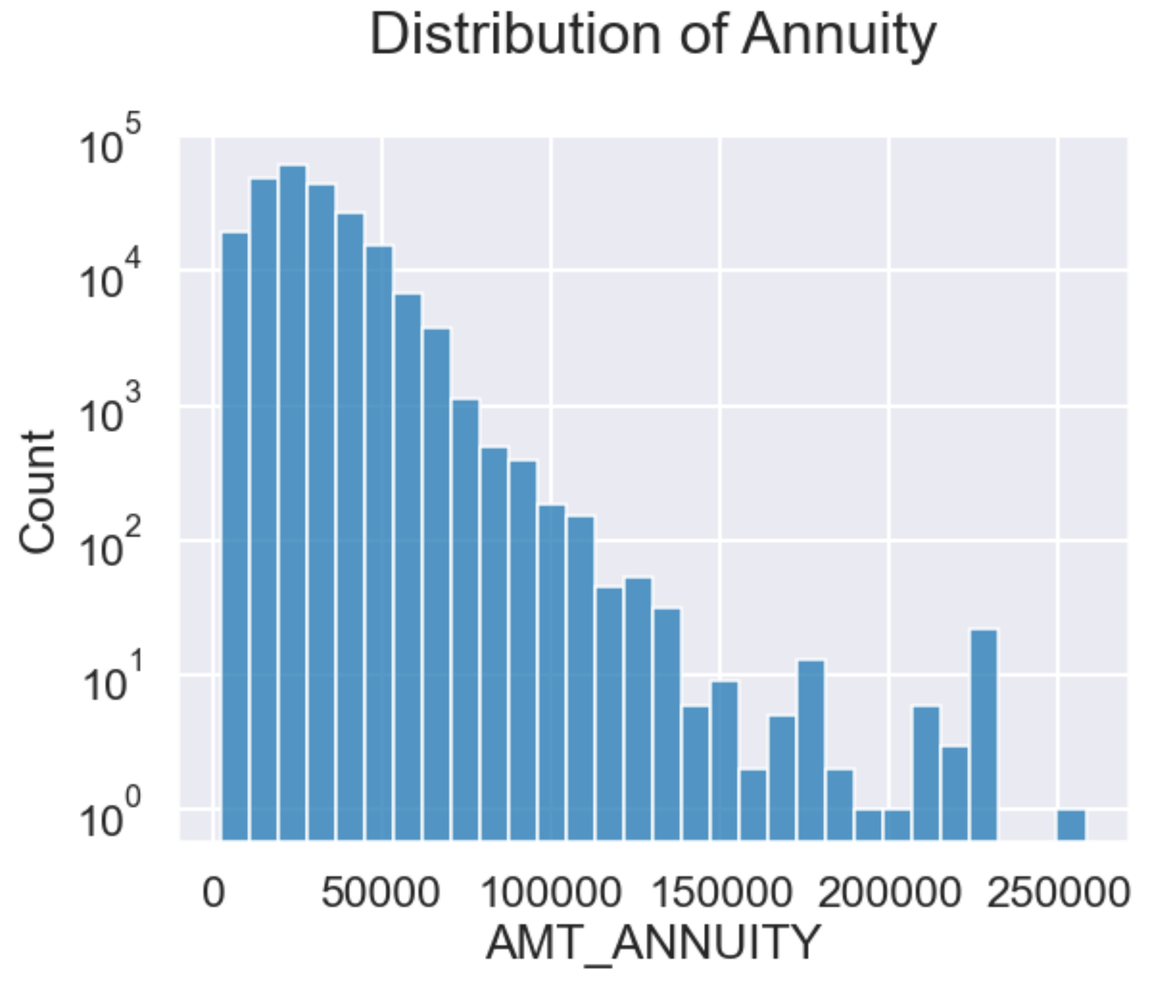
Distributions





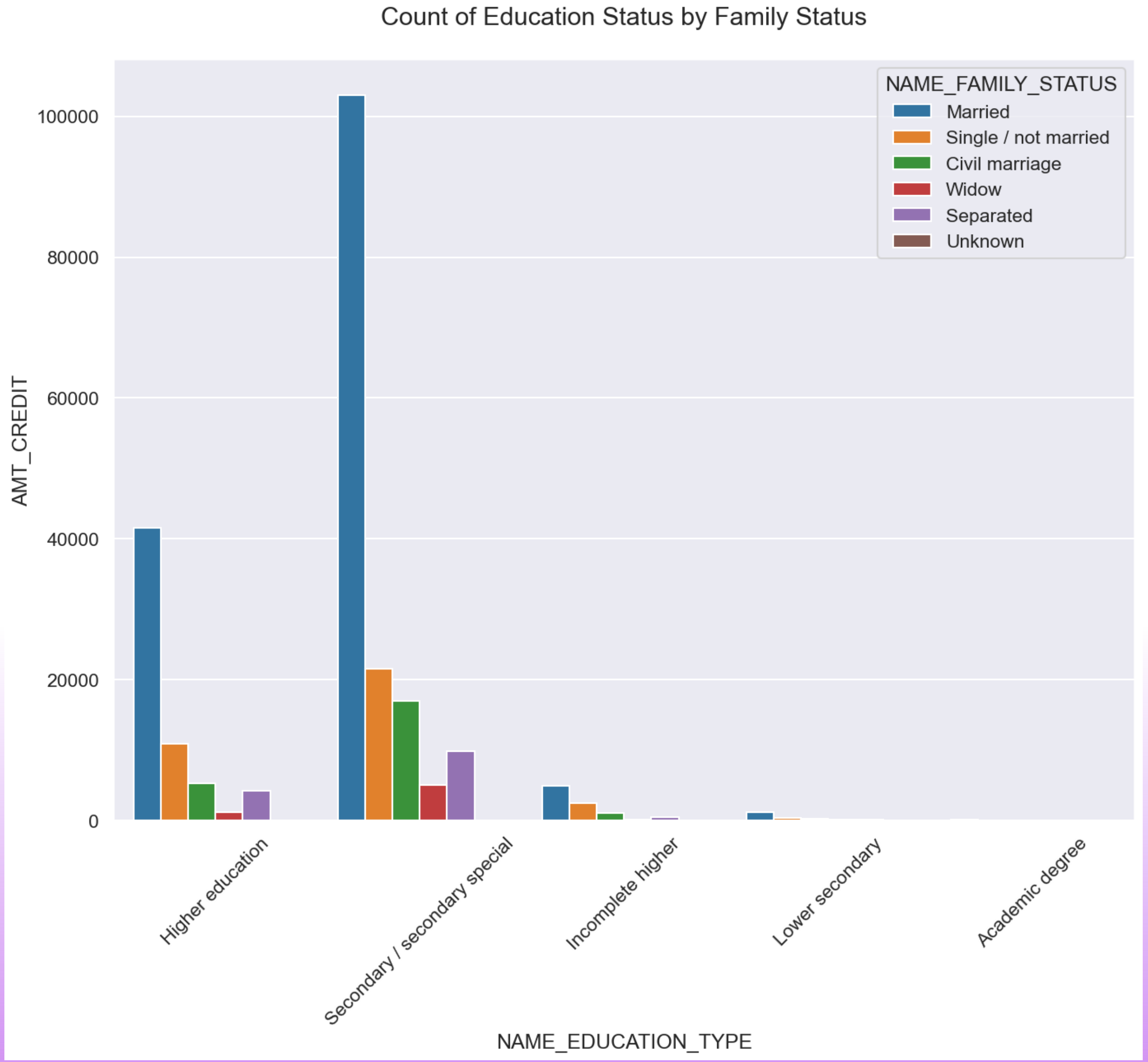
Distributions

The function is then called three times to create three separate distribution plots for the AMT_INCOME_TOTAL, AMT_CREDIT, and AMT_ANNUITY variables from the target0_df dataset. Each plot has a different title indicating the variable being visualized.



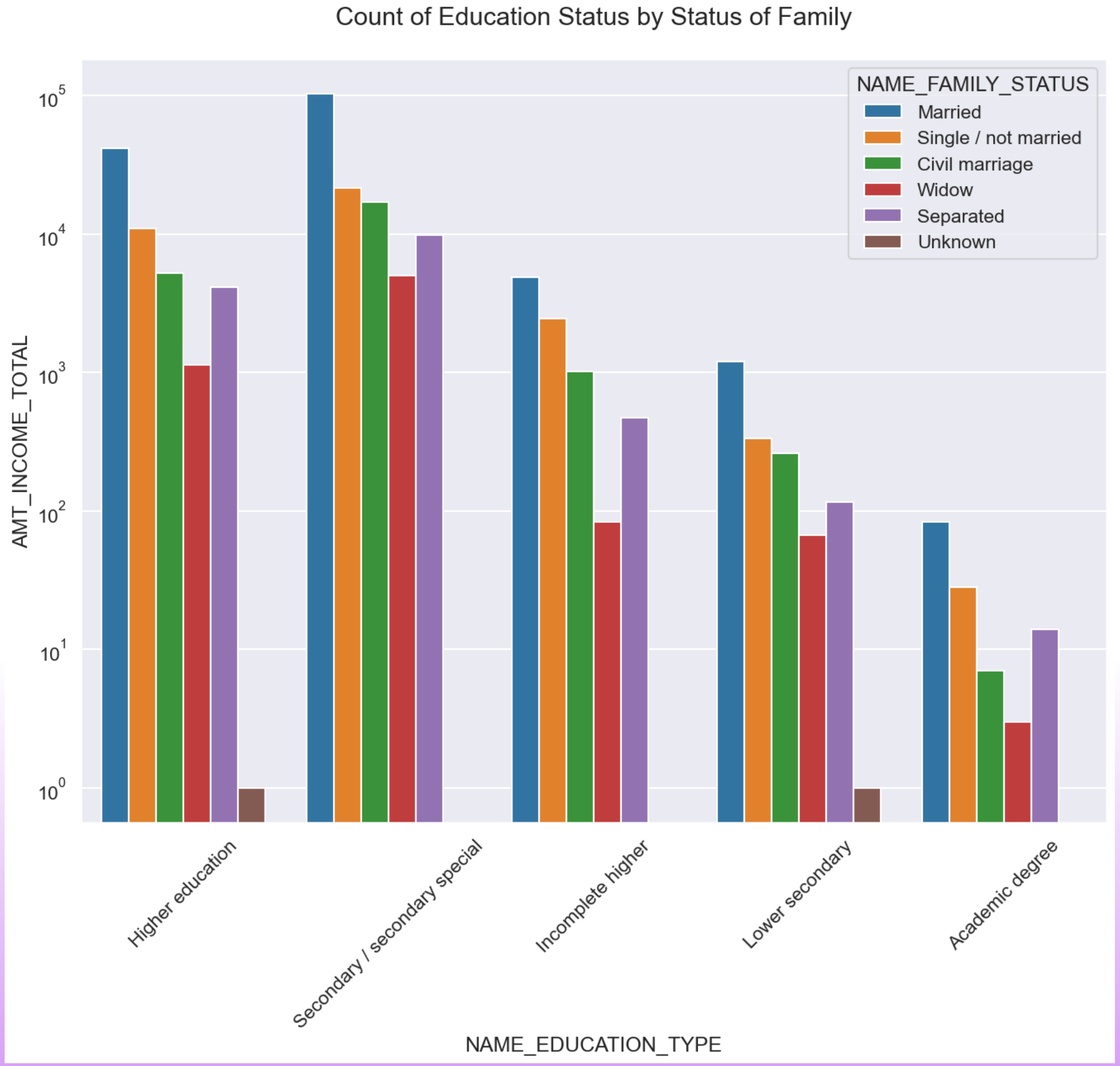
Status

Here’s bar plot to visualize the count of education status based on different family statuses, with the x-axis representing the education types and the bars grouped and coloured by family status.



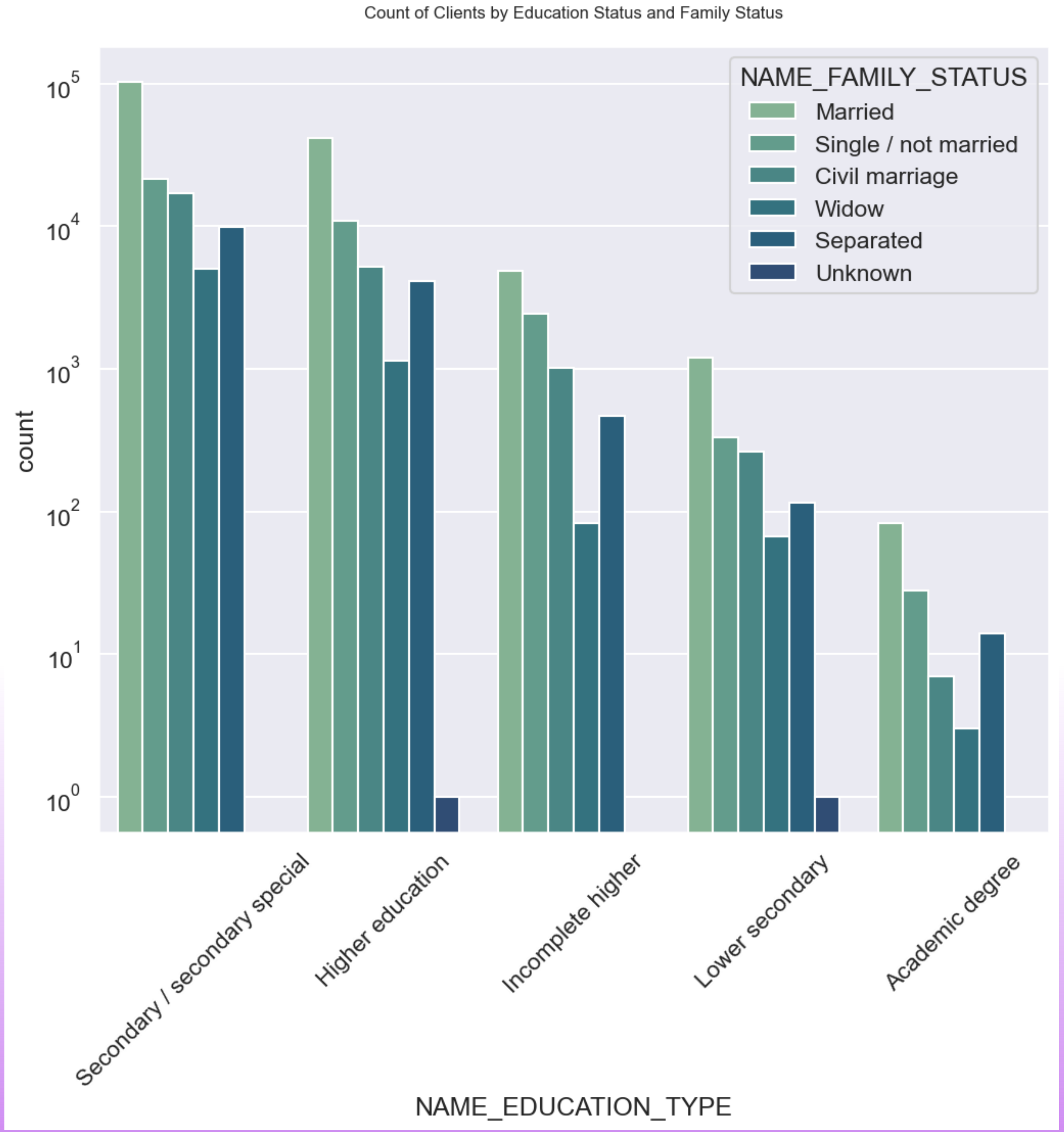
Status

Now generated a bar plot that shows the count of education status based on different family statuses, using a larger figure size, rotated x-axis labels, and a logarithmic scale for the y-axis.



Status

Now generated a count plot that visualizes the count of clients based on their education status and family status. It uses a larger figure size, rotated x-axis labels, a dark grid background, and a larger font size for improved readability. The count values are displayed on a logarithmic scale, and the plot title indicates the purpose of the visualization.

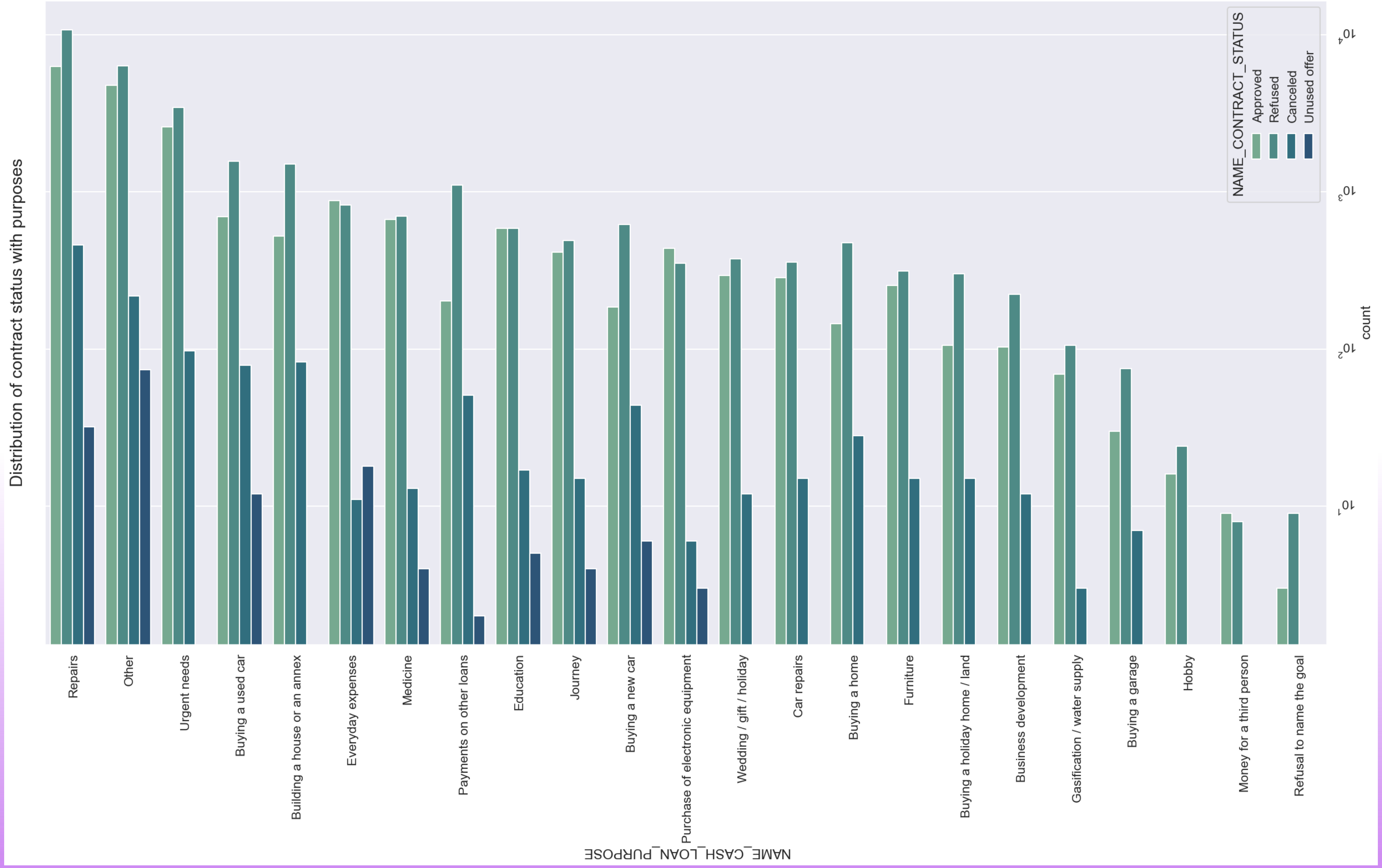


Removing empty column

Importing and reading from a CSV file (Data Set 1), identifies columns with more than 30% empty values, drops those columns from the DataFrame, and outputs the new shape of the modified DataFrame.

From this rows by using code removing df1 that contain specific values ('XNA' and 'XAP') in the 'NAME_CASH_LOAN_PURPOSE' column. The DataFrame df1 is modified accordingly, and its resulting shape can be examined using the df1.shape line.

Distribution Status

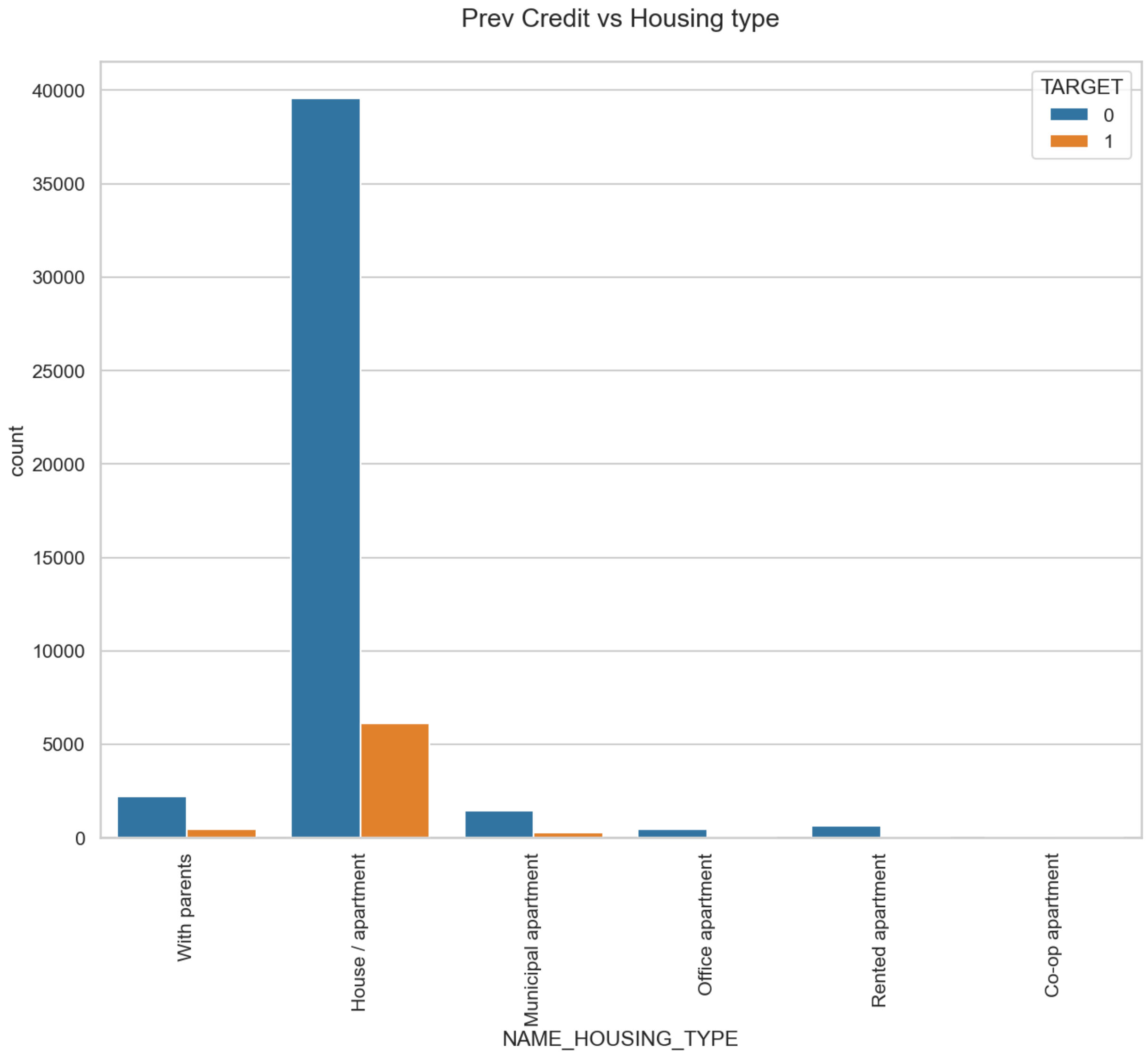


Distribution Status

It visualizes the distribution of contract status with respect to different purposes for cash loans. The data is taken from the new_df1 DataFrame. The y-axis represents the 'NAME_CASH_LOAN_PURPOSE' column, the order of bars is determined by the value counts of 'NAME_CASH_LOAN_PURPOSE', and the bars are colored based on the 'NAME_CONTRACT_STATUS' column. The 'crest' palette is used for coloring. The resulting plot object is assigned to ax.

Housing Types

In summary, the code generates a countplot that visualizes the distribution of previous credit status across different housing types, and the bars are colored based on a target variable. The plot provides insights into the relationship between housing type and credit status.



Observations

•Customers with a high rate of rejection and a default in the current application are considered high-risk customers. They may be denied a loan if their external credit score is low, their income is low, and the amount of credit they applied for is high.

•More than 25% of people who have not defaulted on their current loan have a lower rate of rejection (higher approval rate, >.75) across all age groups. These are low-risk customers who could be given higher-credit loans in future applications. However, it is important to consider the number of approved loans a particular customer has held to date.

•Customers with a high rate of rejection but no default in the current application are considered moderate-risk customers. They may be granted a loan if their external credit score is high and reliable, their income is high, and the amount of credit they applied for is low.

- They could be granted a loan with a lower credit amount.
- They could be granted a loan with a higher interest rate, provided their income is higher and their credit score is reliably higher.

•Customers who had a higher approval rate but defaulted on the current application could be granted a loan with a lower credit amount. A higher credit amount would attract higher interest charges, which could further stress the customer financially.

Recommendation

Based on the customer profile and credit default drivers, we identify and recommend the following:

•High-risk profiles: Customers with a low external credit score, low income, and a high credit amount applied for are considered high-risk. Their loan applications could be rejected.

•Low-risk profiles: Customers with a high external credit score, high income, and a low credit amount applied for are considered low-risk. They should be extended higher credit loans in future applications. However, it is important to inquire about the number of previously approved loans they have held to date.

•Moderately low-risk profiles: Customers with a medium external credit score, medium income, and a medium credit amount applied for are considered moderately low-risk. They could be granted loans with a lower credit amount. They could also be granted credit at a higher interest rate, provided their income is higher and their credit score is reliably higher.

•Medium-risk profiles: Customers with a medium external credit score, medium income, and a high credit amount applied for are considered medium-risk. They could be granted loans with a lower credit amount. A higher credit amount would attract higher annuity and that would further stress an already defaulting customer.

Conclusion

In conclusion we provide insights into the factors influencing payment difficulties and loan defaults, enabling banks to make more informed decisions in assessing loan applications, managing risks, and targeting client segments for better loan performance.

1. The proportion of Payment Difficulties among pensioners has decreased, while the proportion of Payment Difficulties among working individuals has increased, compared to both Payment Difficulties and non-Payment Difficulties.

2. There is a lower percentage of married and widowed individuals with Loan Payment Difficulties and a higher percentage of single and civil married individuals with Loan Payment Difficulties, compared to both Loan Payment Difficulties and Loan Non-Payment Difficulties.

3. The percentage of Loan Payment Difficulties with secondary/secondary special educational qualifications has increased, while the percentage of Loan Payment Difficulties with higher education completion has decreased, compared to both Loan Payment Difficulties and Loan Non-Payment Difficulties.

4. The count of 'Low skilled Laborers' in the 'OCCUPATION_TYPE' category is relatively low, but they have the highest percentage of payment difficulties, around 17%. Therefore, clients with the occupation type 'Low skilled Laborers' are significant contributors to Loan Defaulters.