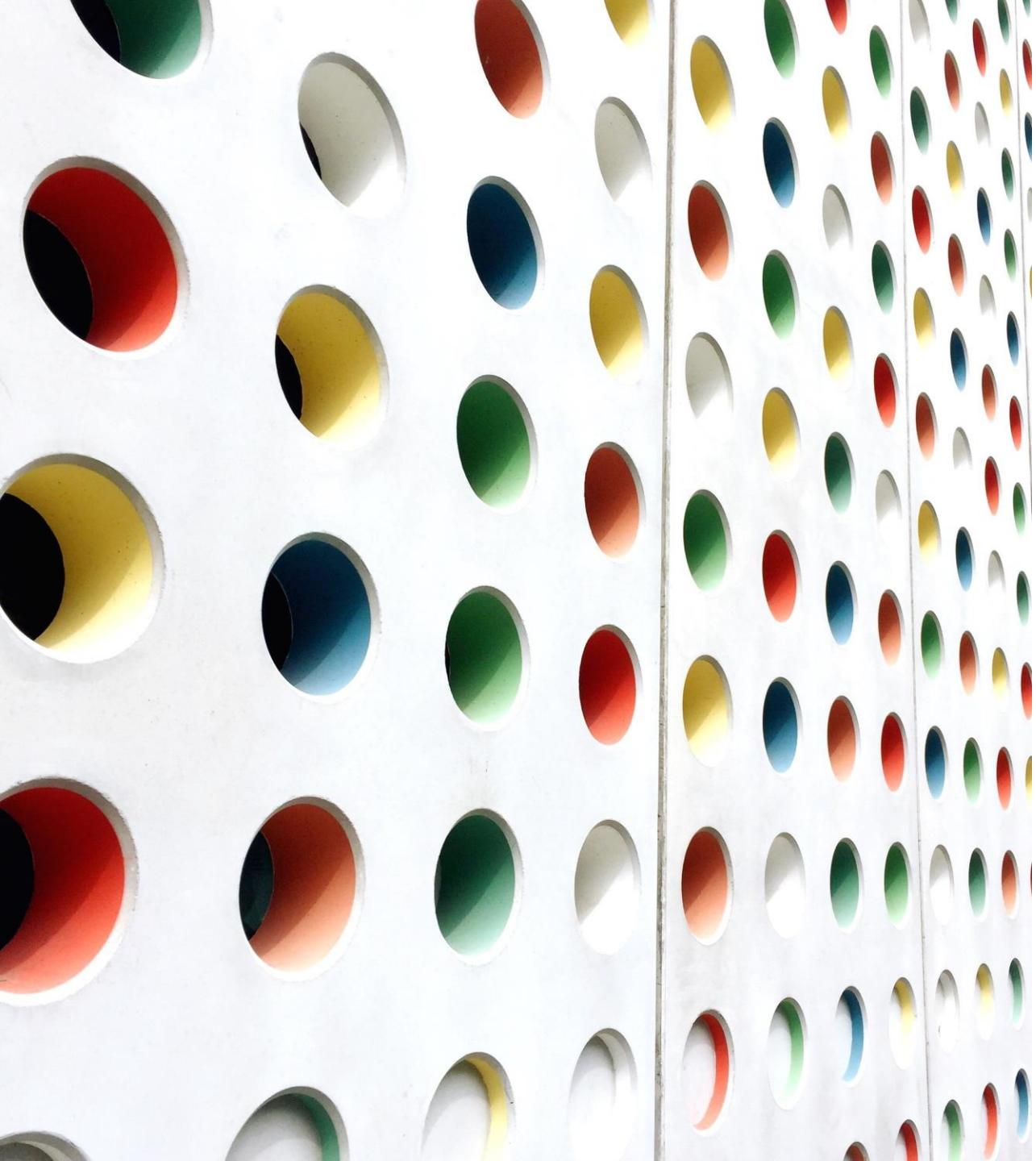


CREDIT RISK ANALYSER

BY-

Y. Swetha Ashervadam

Data science

Problem Statement

Challenges in Loan Approval for Applicants with Limited Credit History:

Lending institutions often encounter difficulties when assessing loan applications from individuals with inadequate or non-existent credit histories. Unfortunately, some applicants exploit this situation by deliberately defaulting on their loans. The company aims to use **Exploratory Data Analysis (EDA)** to scrutinize the data for patterns, ensuring that deserving applicants, with the capacity to repay their loans, are not unjustly declined.

Balancing Loan Approval Decisions: When the company receives a loan application, it must make a pivotal decision regarding loan approval. This decision is associated with two distinct risks:

1. **Risk of Non-Approval**: If the applicant is likely to repay the loan, declining their application results in a loss of potential business for the company.

2. **Risk of Default**: On the other hand, if the applicant is not likely to repay the loan and is at risk of defaulting, approving the loan may lead to a financial loss for the company.

Data Scenarios

The provided data is categorized into two scenarios:

1. **Clients with Payment Difficulties**: This category includes individuals who have a history of late payments of more than X days on at least one of the first Y installments of the loan in the sample.
2. **All Other Cases**: This category encompasses all other cases where loan payments are made on time.

- The ***goal of the analysis*** is to distinguish patterns and characteristics that differentiate these two scenarios, allowing the company to make more informed and balanced loan approval decisions, reducing the risk of financial loss while serving deserving applicants.

Approach & Methodology

Data understanding

- There are two datasets:
  - The dataset ***'application_data.csv'*** contains all the information of the client at the time of application. The data is about whether a client has payment difficulties. There are 307511 rows and 122 columns.
  - The dataset ***'previous_application.csv'*** contains information about the client's previous loan data. It contains the data on whether the previous application had been **Approved, Cancelled, Refused or Unused offer.** There are 1670214 rows and 37 columns.

Application_data.csv

- In this dataset, there are 307511 rows and 122 columns.
- Some important columns used for this analysis are -
  - **TARGET** - 0 or 1
    - A **Target** value of **`1`** is assigned to individuals who are unable to repay the loan and a **Target** value of **`0`** to those capable of repaying the loan.
  - **"AMT_INCOME_TOTAL"** - Income of the client
  - **"AMT_CREDIT"** - Credit amount of the loan
  - **"AMT_ANNUITY"** - Loan annuity
  - **"AMT_GOODS_PRICE"** - For consumer loans it is the price of the goods for which the loan is given
  - **"DAYS_BIRTH"** - Client's age in days at the time of application
  - **"DAYS_EMPLOYED"** - How many days before the application the person started current employment

Previous_application_data.csv

- In this dataset, there are 1670214 rows and 37 columns.
- Some important columns used for this analysis are –
  - **"NAME_CONTRACT_TYPE"** - Contract product type (Cash loan, consumer loan [POS] ,...) of the previous application
  - **"AMT_ANNUITY“** - Annuity of previous application

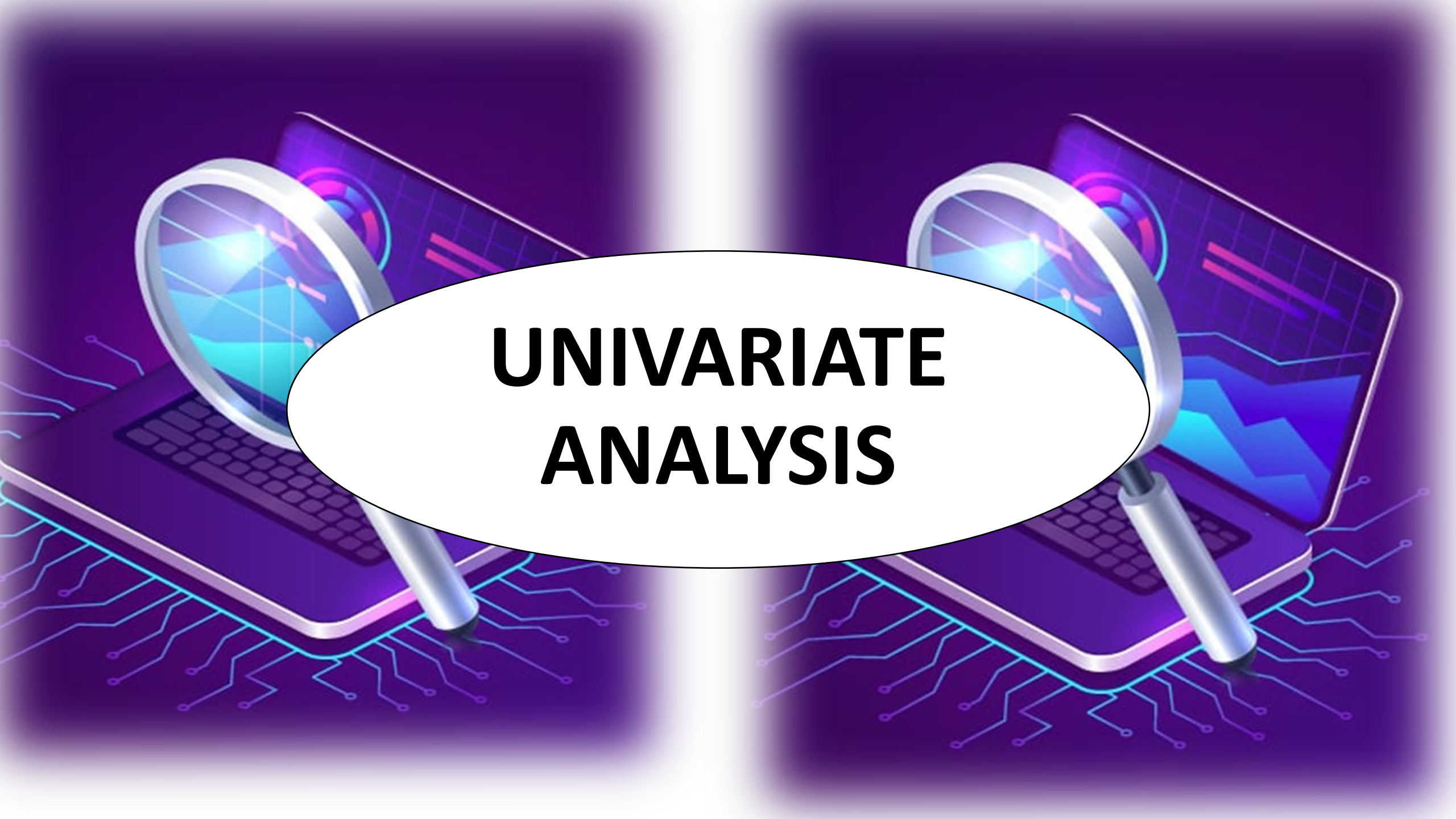
IMPUTE/REMOVE NULL VALUES

- In the **Application_data,** there were 49 columns which were having more than 40% of null values. So they were dropped.
- After dropping there I notice that 18 columns has some **NaN** values. Here, we inspect these columns which are required.
- It is seen that OCCUPATION_TYPE has 31.35% of null values so it is safe to replace all the missing values with `not_known` in the column.
- EXT_SOURCE_3 has 19.83% of null values and there are no outliers, so it is advised to impute the null values with the mean of the variable.
- I have compiled a list named `credit_counts` which have nearly the same percentage of null values – and imputed with median and mean according to the presence of the outliers.

- The next **four** columns give information about the **number of instances or observations where a client's social connections** or network show a noticeable or observable pattern of defaulting on their payments. The column with an **outlier** is replaced with a **median**, while the remaining with **mean**.
- The remaining top **five** columns with null values **can be ignored since the null values are minimal**, and they won't significantly impact the analysis.
- After dropping the null value columns, now the dataset has **307511** rows and **73** columns.

- In the **Previous_application** dataset, there were 49 columns which were having more than 40% of null values. So they were dropped.
- The **five** remaining columns with null values can be ignored since the **null values are minimal**, and they won't significantly impact the analysis.
- After dropping the null value columns, now the dataset has **1670214** rows and **26** columns.

Graphs and Insights

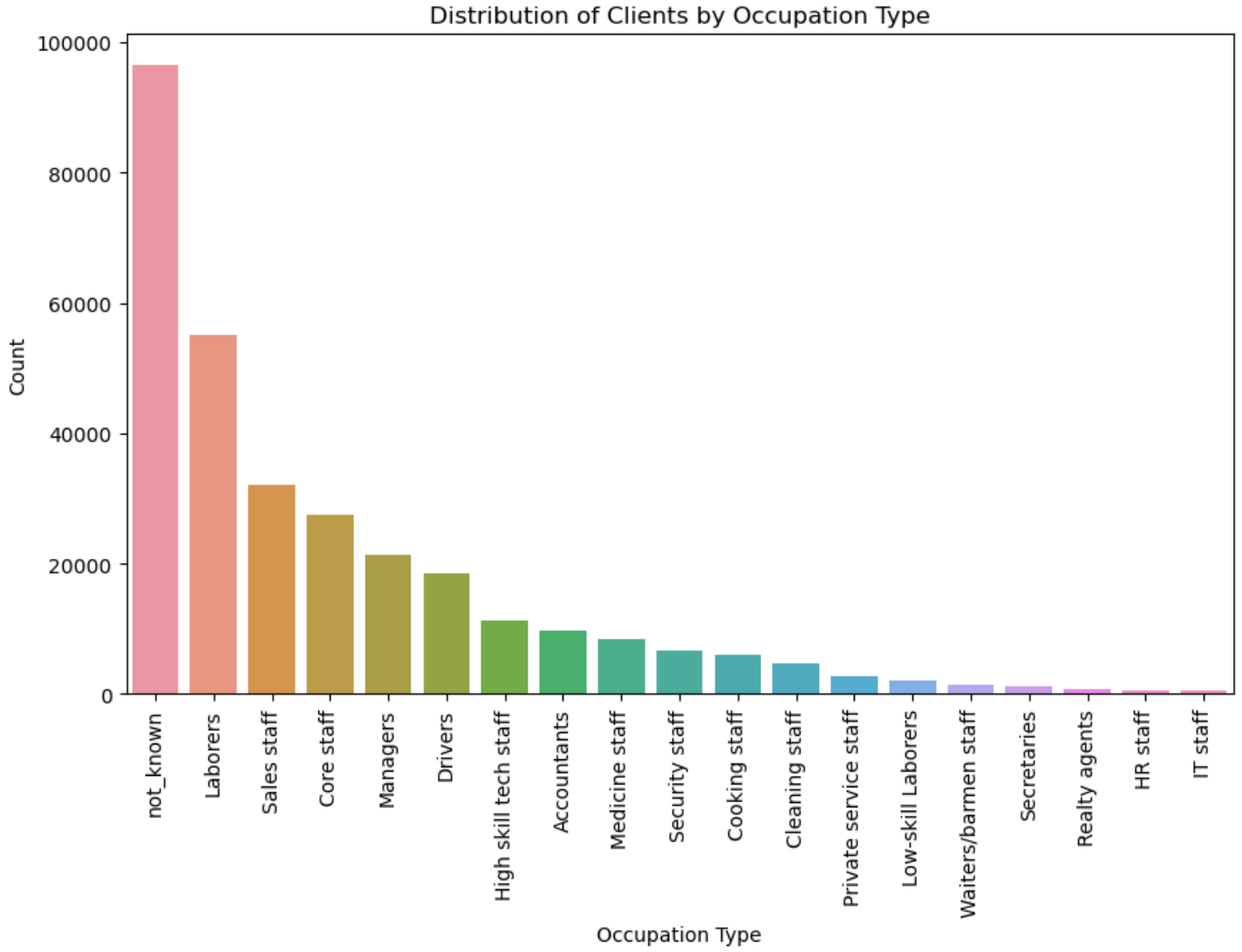

UNIVARIATE
ANALYSIS

Categorical Univariate Analysis

1. OCCUPATION_TYPE (Categorical):

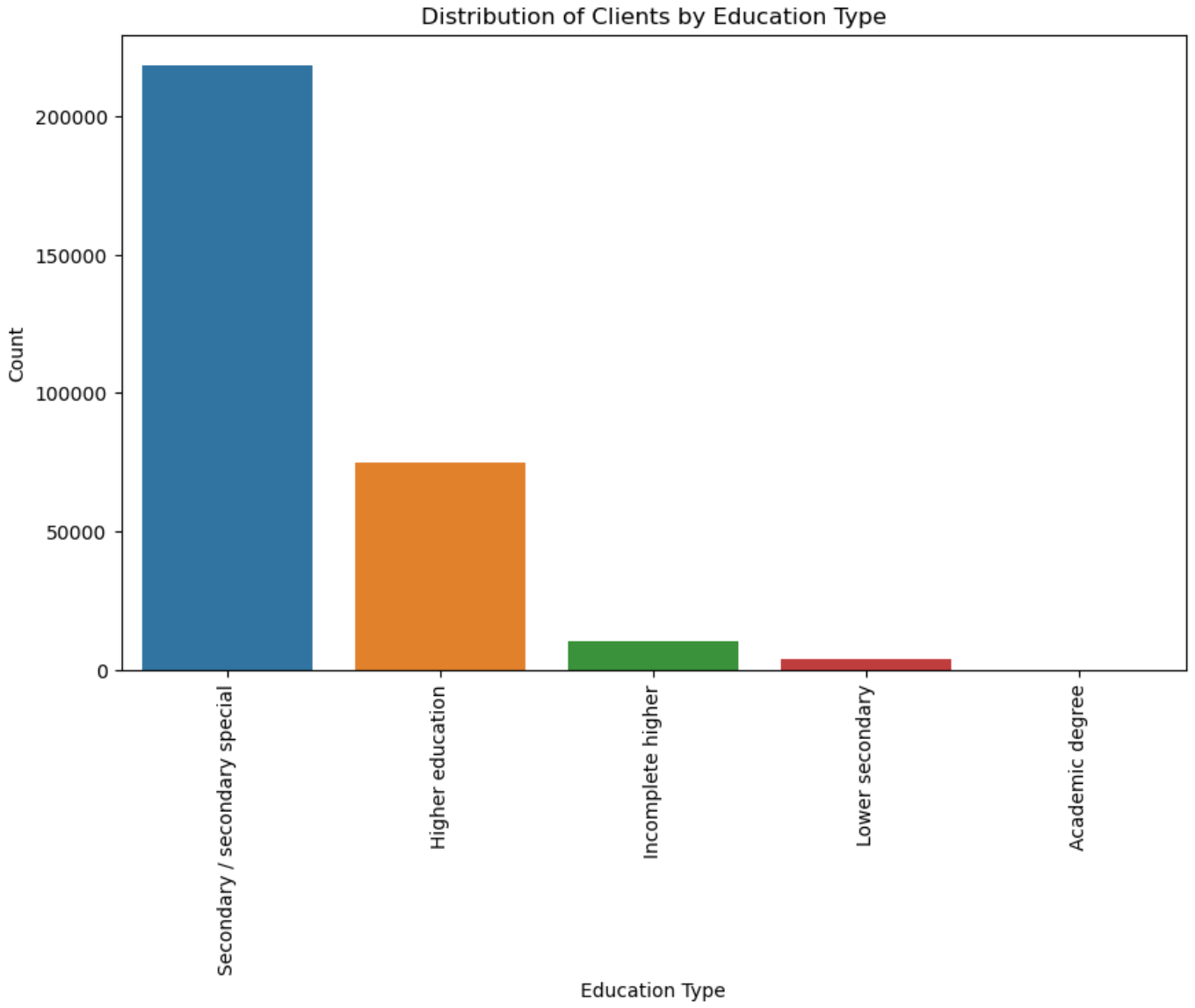
- **Graph Name:** Count Plot
- **Purpose:** Visualize the distribution of clients with different occupations.
- **Insight:** Check if certain occupations are more prone to payment difficulties.

➢ **Analysis:** The analysis reveals that individuals with an occupation labeled as 'unknown' are more likely to experience payment difficulties.



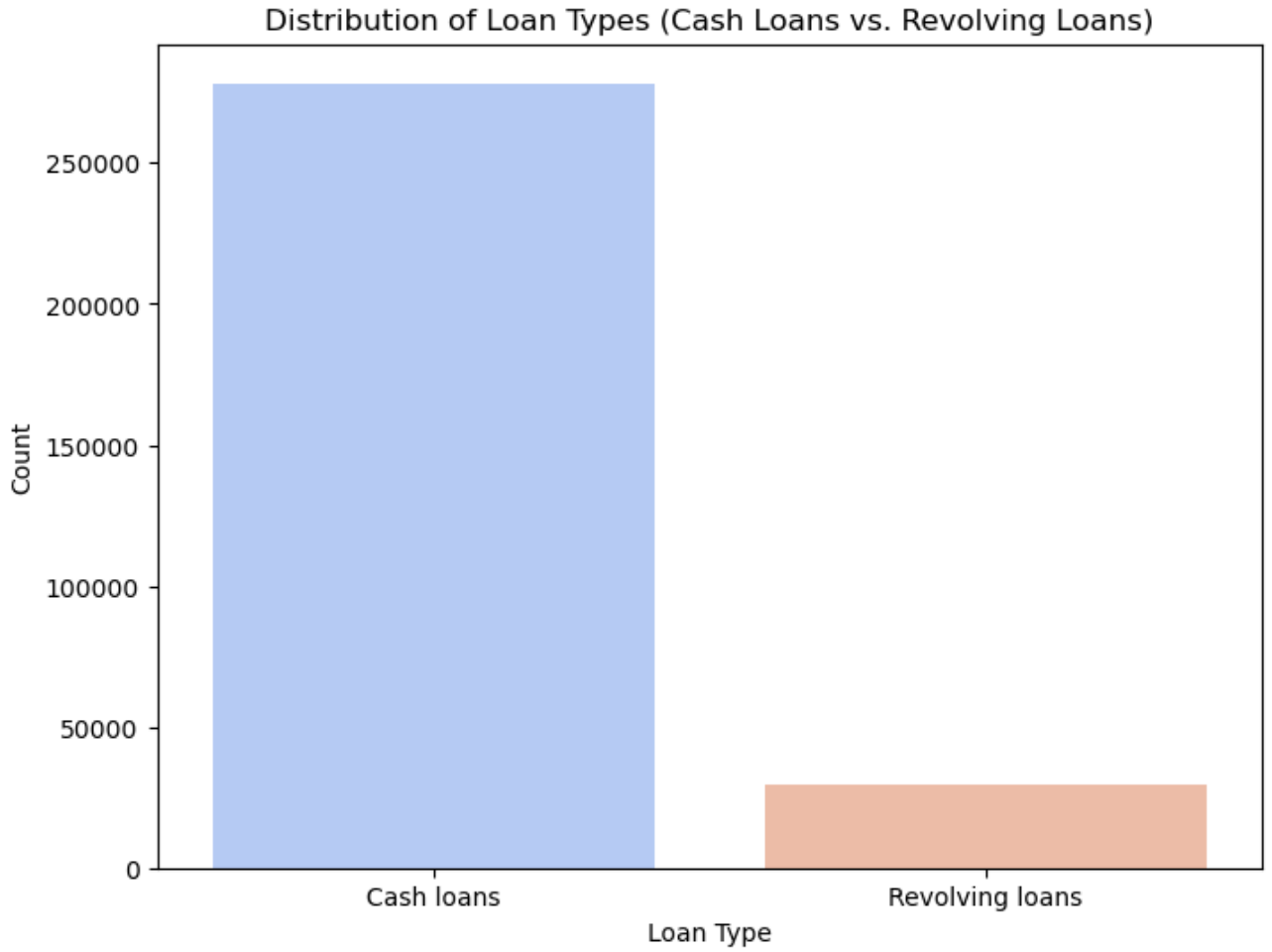
2. NAME_EDUCATION_TYPE (Categorical):

- **Graph Name:** Count Plot
- **Purpose:** Visualize the distribution of education types.
- **Insight:** Determine if education level affects payment difficulties.
- **Analysis:** From this dataset, it's evident that individuals applying for loans tend to have lower educational qualifications.



3. NAME_CONTRACT_TYPE (Categorical):

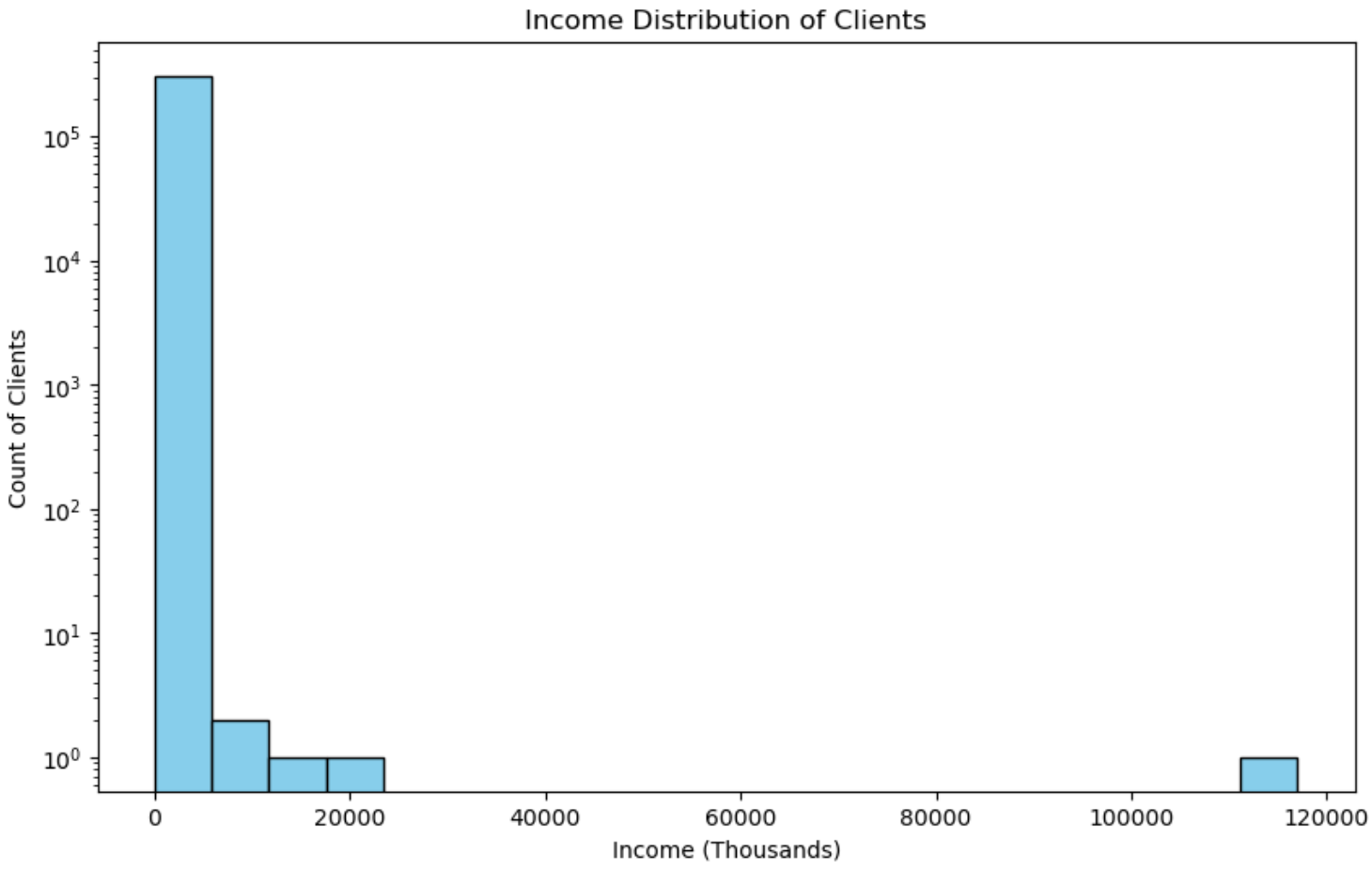
- **Graph Name:** Count Plot
- **Column:** "NAME_CONTRACT_TYPE"
- **Purpose:** To visualize the distribution of different loan types among clients.
- **Insight:** This analysis provides insights into the prevalence of different loan types in the client population It helps understand the proportion of clients who have Cash loans versus Revolving loans.
- **Analysis:**
- The analysis of this distribution suggests that the majority of clients have "Cash loans", while a smaller proportion have "Revolving loans".
- "Revolving loans," being a smaller segment, may warrant closer scrutiny due to their potential higher-risk nature.
- Based on the analysis, lenders can develop tailored risk mitigation strategies for each loan type. This may include adjusting lending criteria, setting appropriate interest rates, and implementing risk management measures specific to the characteristics of each loan category.



Categorical Numerical Analysis

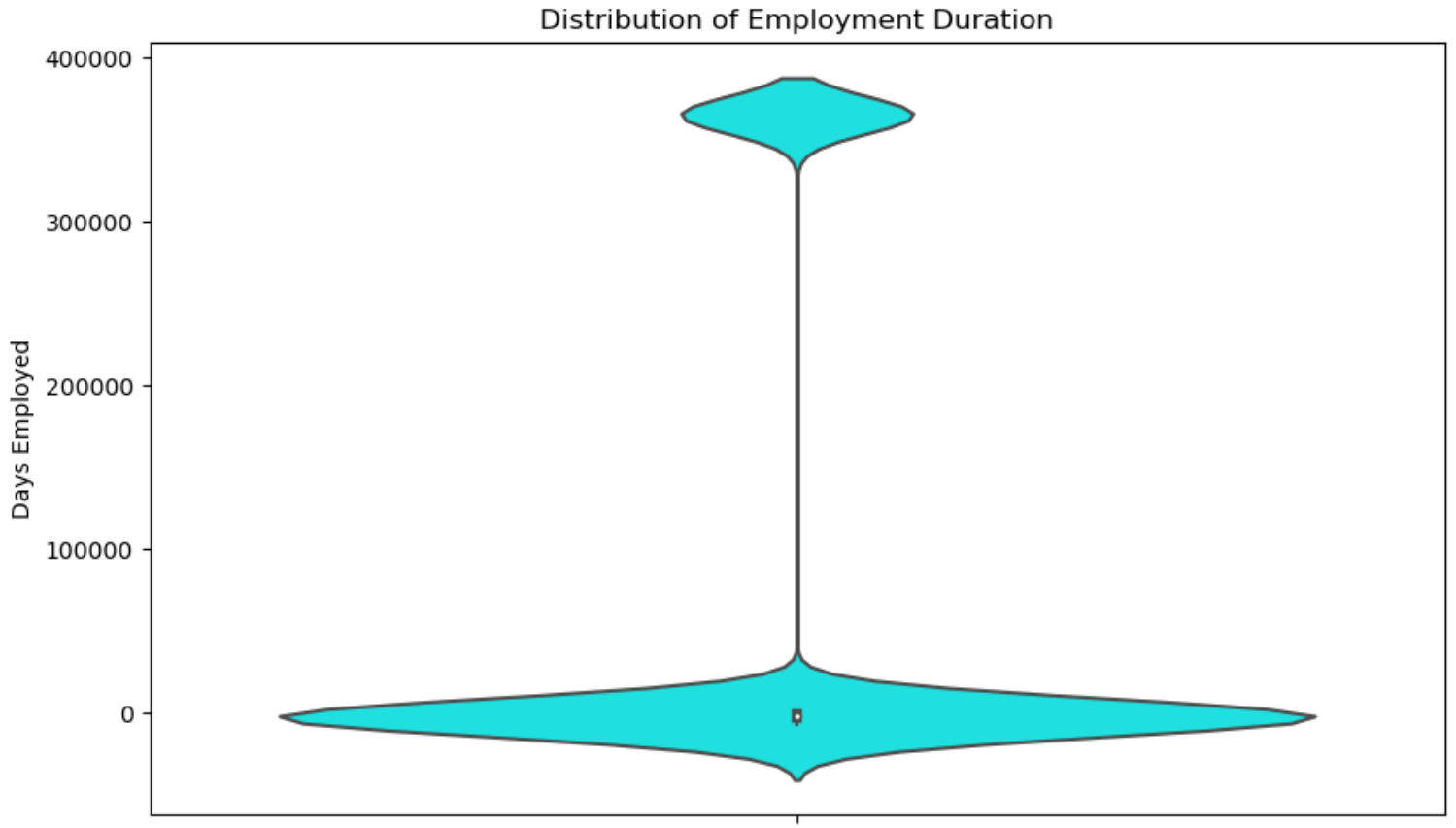
1. **AMT_INCOME_TOTAL (Numerical):**

- **Graph Name:** Histogram
- **Purpose:** Analyze the income distribution.
- **Insight:** Check if higher income clients are less likely to have payment difficulties.
- **Analysis:** The analysis of this data indicates that clients with higher incomes do not face challenges when it comes to repaying their loans.



2. DAYS_EMPLOYED (Numerical):

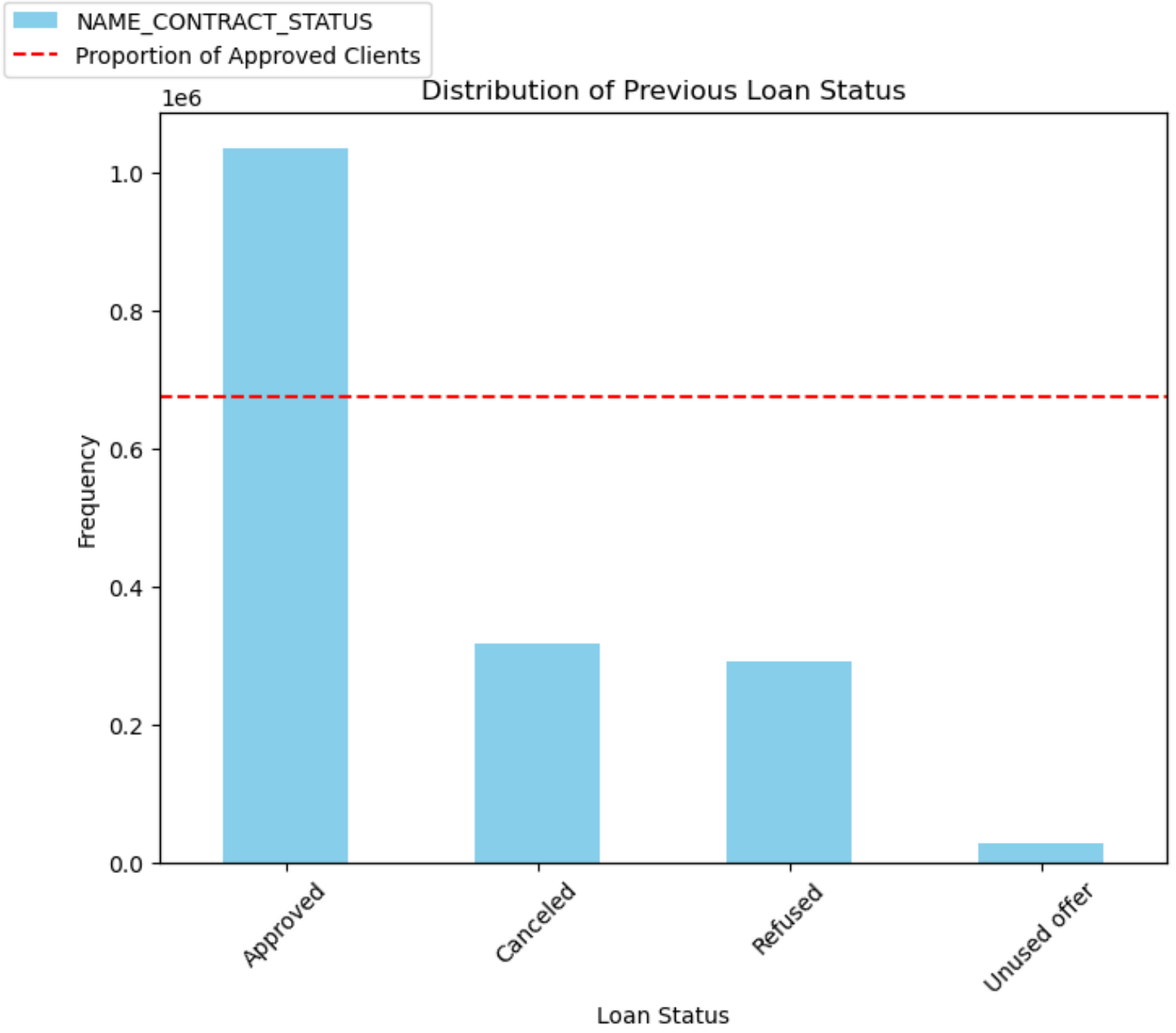
- **Graph Name:** Violin Plot
- **Purpose:** Analyze the employment duration distribution.
- **Insight:** Check if longer employment is correlated with lower payment difficulties.
- **Analysis:** From this dataset, analysis reveals some patterns where longer employment durations are associated with lower payment difficulties, which can be a valuable insight in credit risk assessment.



Univariate Analysis for 'previous_application.csv':

1. NAME_CONTRACT_STATUS (Categorical):

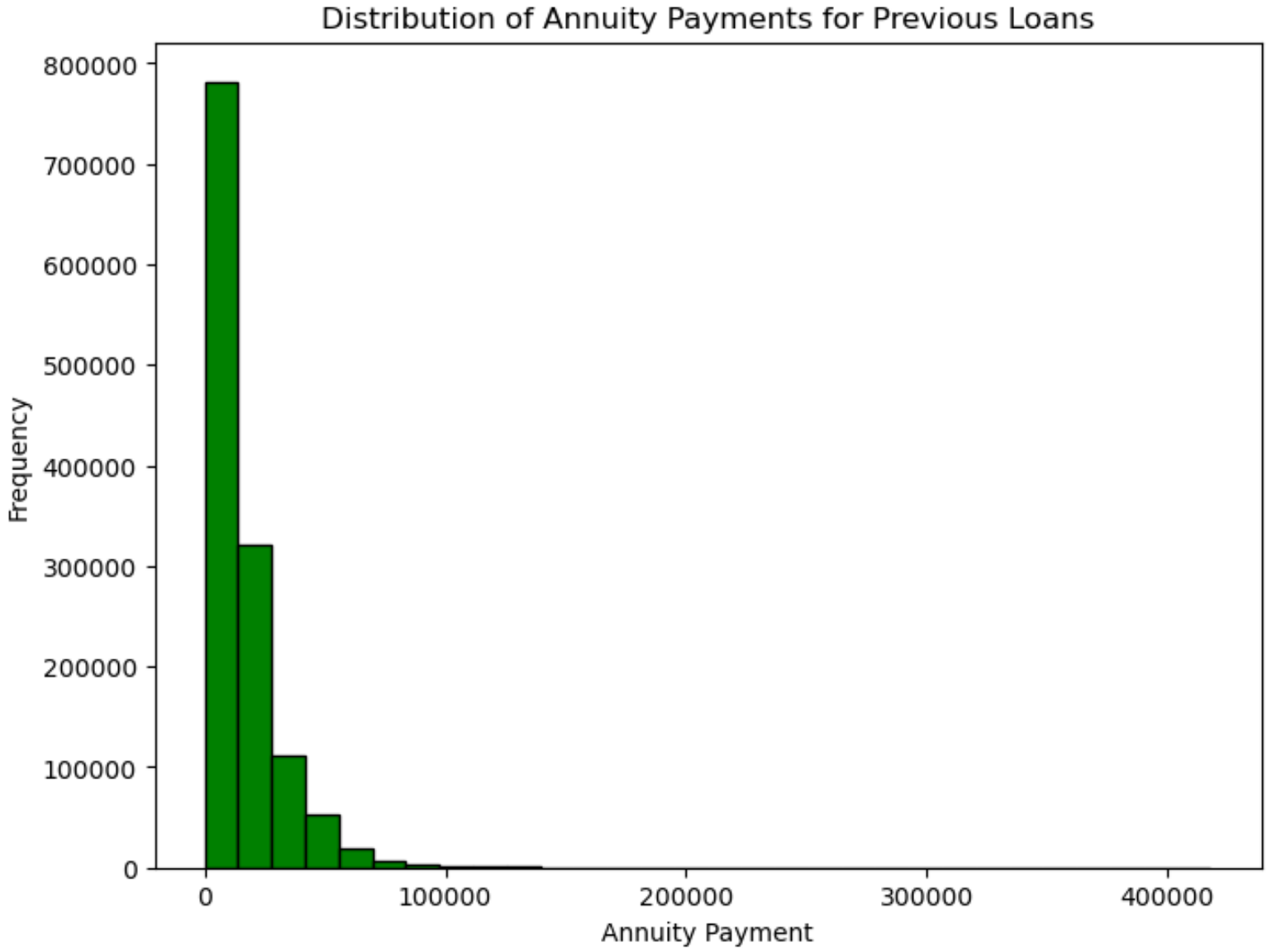
- **Graph Name:** Bar Plot
- **Purpose:** To visualize the distribution of previous loan statuses. Assess whether clients with previous approved loans are less likely to face payment difficulties..
- **Insight:** By comparing the proportion of clients with "Approved" status to the proportion of clients with other statuses (e.g., Canceled or Refused), you can gain insights into whether those with approved loans have a lower likelihood of facing payment difficulties.
- **Analysis:** The insight derived from the analysis is that a substantial proportion of clients have "Approved" loans, which suggests that many clients in the dataset have successfully secured loans, possibly indicating a higher likelihood of them not facing payment difficulties.

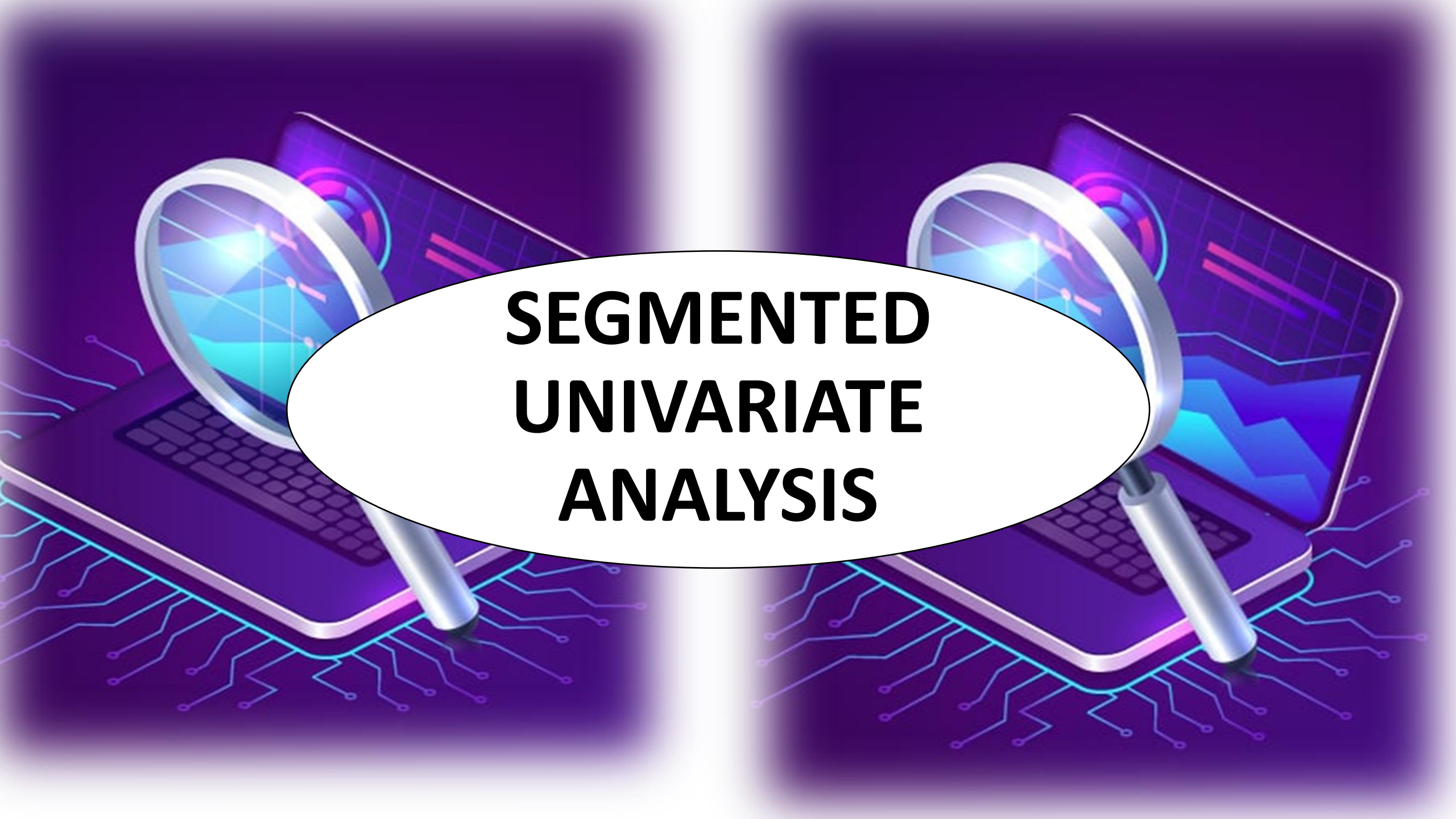



```
Proportion of clients with 'Approved' loans: 62.07%
```

2. AMT_ANNUITY (Numerical):

- **Graph Name:** Histogram
- **Purpose:** To analyze the annuity payment distribution for previous loans.
- **Insight:** By analyzing the shape of the distribution and the presence of any peaks or clusters, you can gain insights into the annuity payment patterns.
- **Analysis:** The annuity payment distribution reveals distinct peaks, with the highest frequency observed at 780,000, followed by a secondary peak at around 320,000, another peak near 120,000, and a final peak at approximately 50,000. These peaks and clusters in the distribution provide valuable insights into annuity payment patterns, suggesting that certain payment amounts are more common or have higher frequencies compared to others.

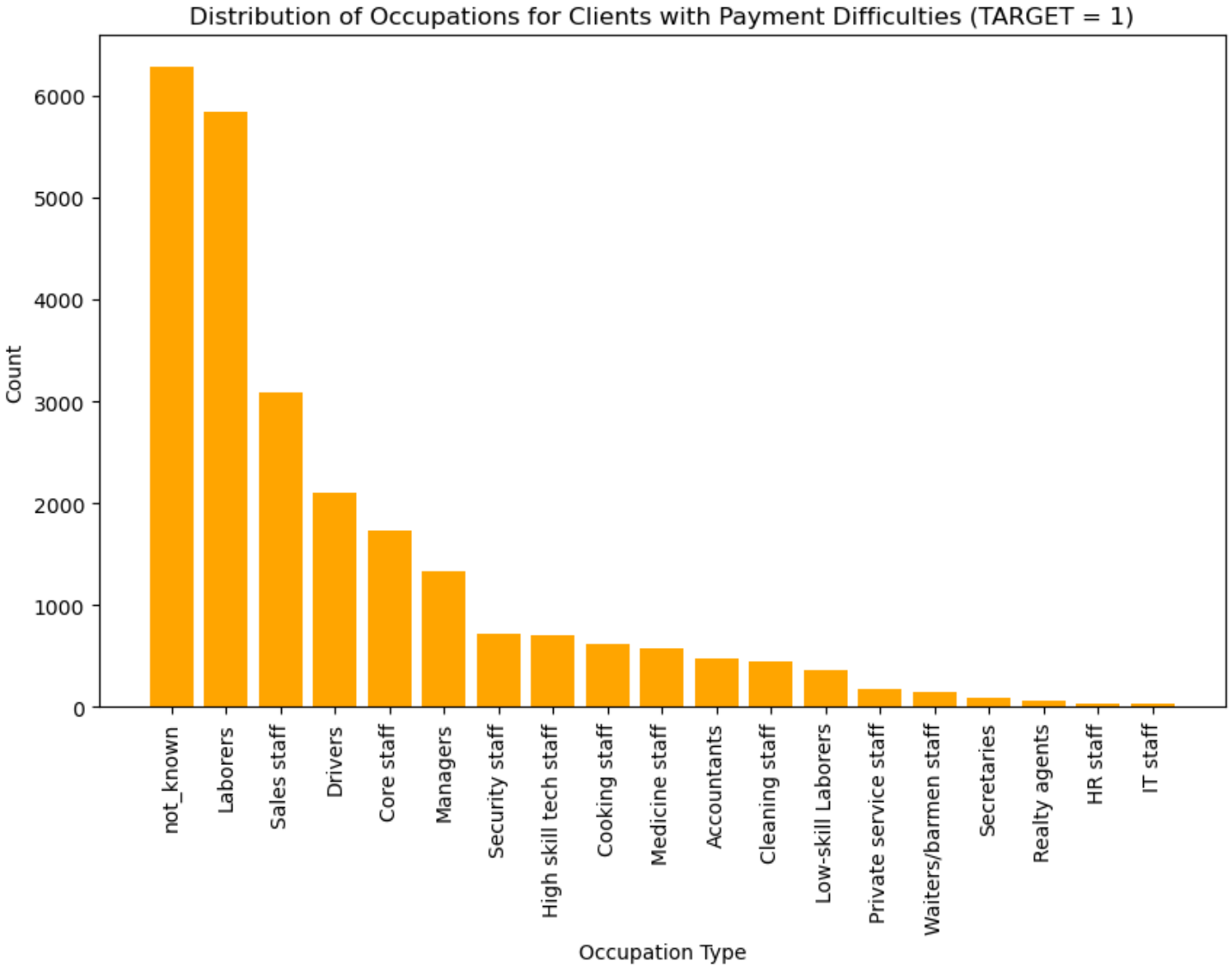



SEGMENTED
UNIVARIATE
ANALYSIS

Analysis of Clients with payment difficulties -- *Defaulters*

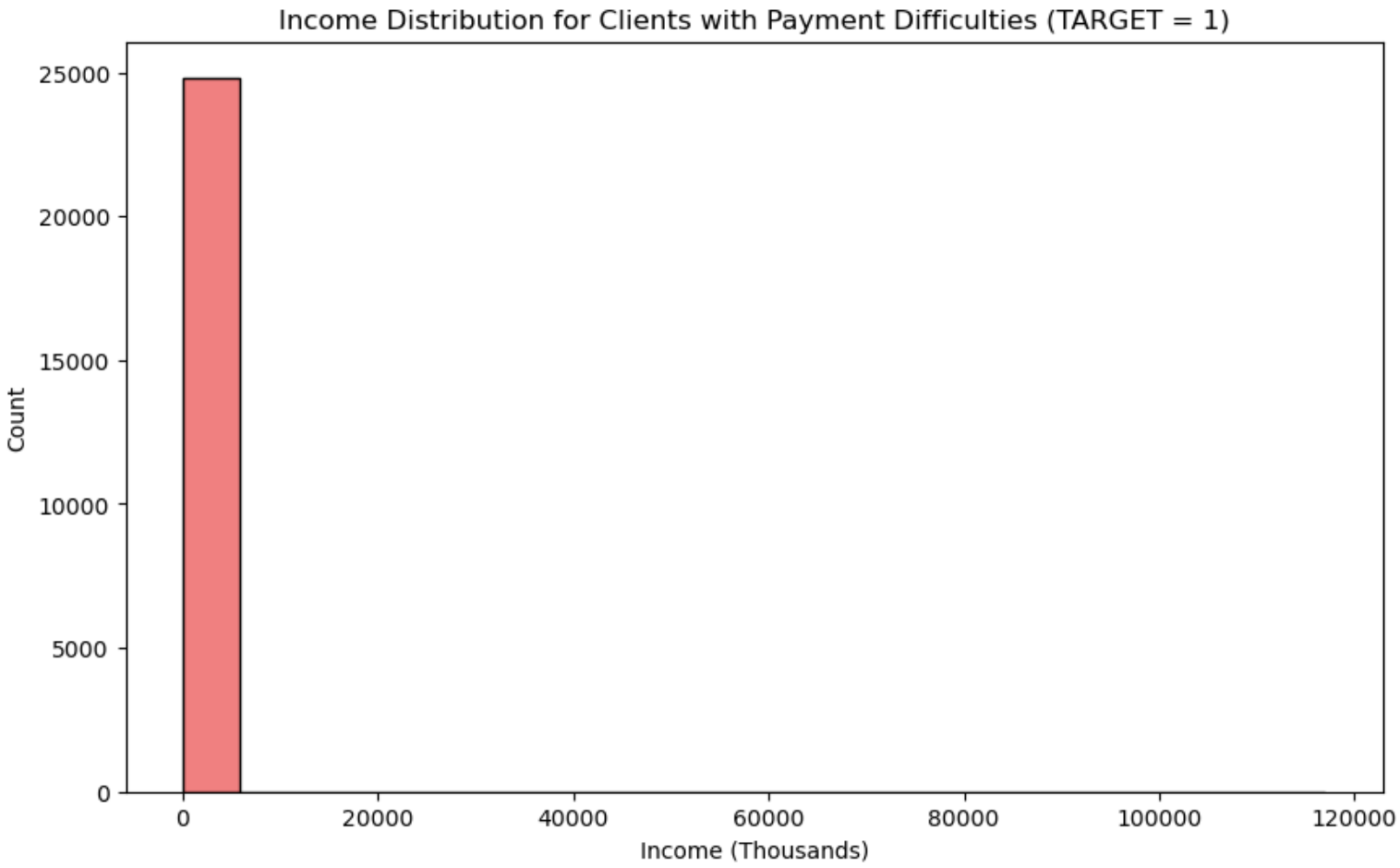
1. **Clients with Payment Difficulties ('TARGET' = 1) with their OCCUPATION_TYPE.**

- **Graph Name:** Bar Chart
- **Purpose:** Visualize the distribution of clients with different occupations who have payment difficulties.
- **Insight:** Identify if certain occupations are more prone to payment difficulties among defaulters.
- **Analysis:** The analysis indicates that individuals with an unknown occupation exhibit the highest defaulters when it comes to repaying loans. Following that, laborers and sales staff are the next groups with notable challenges in loan repayment.



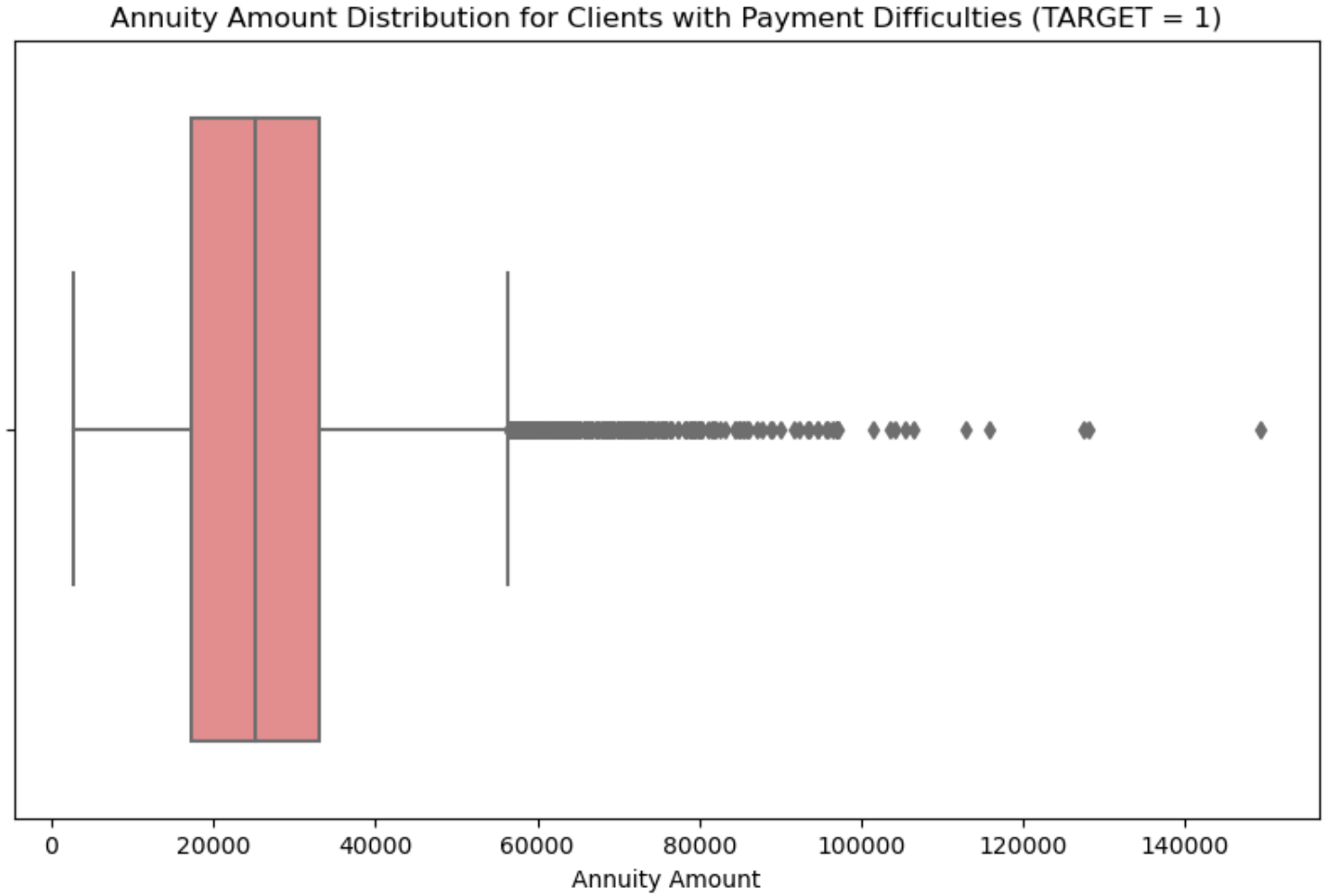
2. Clients with Payment Difficulties ('TARGET' = 1) with their Total Income -- 'AMT_INCOME_TOTAL'.

- **Graph Name:** Histogram
- **Purpose:** Analyze the income distribution for clients with payment difficulties.
- **Insight:** Check if higher income clients are less likely to have payment difficulties among clients with defaulters
- **Analysis:** The analysis reveals a significant challenge in loan repayment for individuals with incomes lower than Rs. 2,000. Therefore, individuals with such low incomes should not be considered eligible for loans.



3. Clients with Payment Difficulties with Annuity Amount -- 'AMT_ANNUITY'.

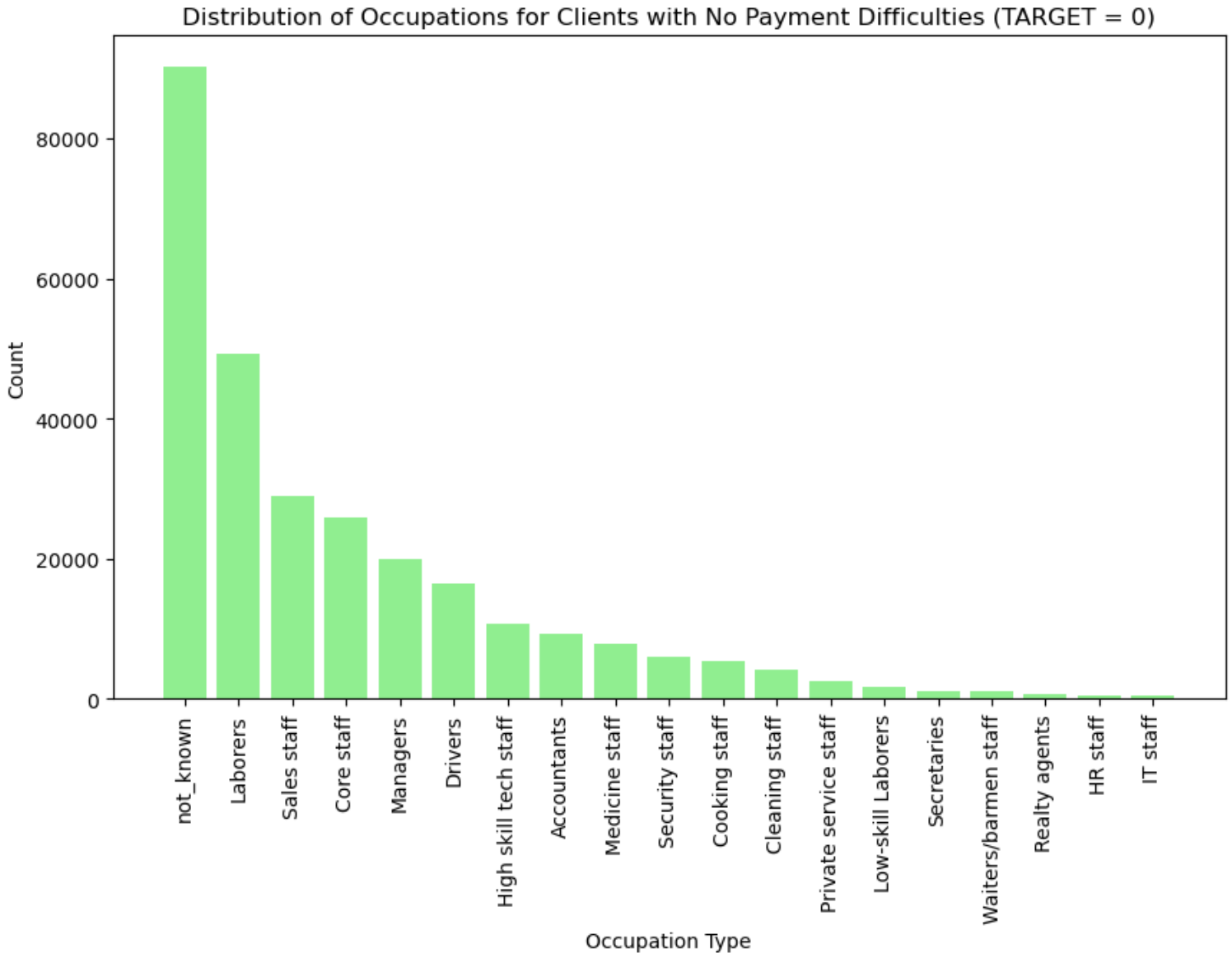
- **Graph Name:** Boxplot
- **Purpose:** It aims to identify any potential outliers or patterns in annuity amounts within this specific group of clients.
- **Insight:** This analysis can help identify whether the defaulters tend to have different annuity amounts compared to those without payment difficulties and can inform risk assessment and lending strategies.
- **Analysis:** The data suggests that clients with payment difficulties have annuity amounts that are **relatively consistent,** but there are still some outliers with higher annuity amounts. When assessing credit risk within this group, other factors such as the nature and **severity of their payment difficulties, credit history, and financial stability** should also be taken into account to make a comprehensive evaluation of their creditworthiness.



Analysis of Clients with No Payment Difficulties

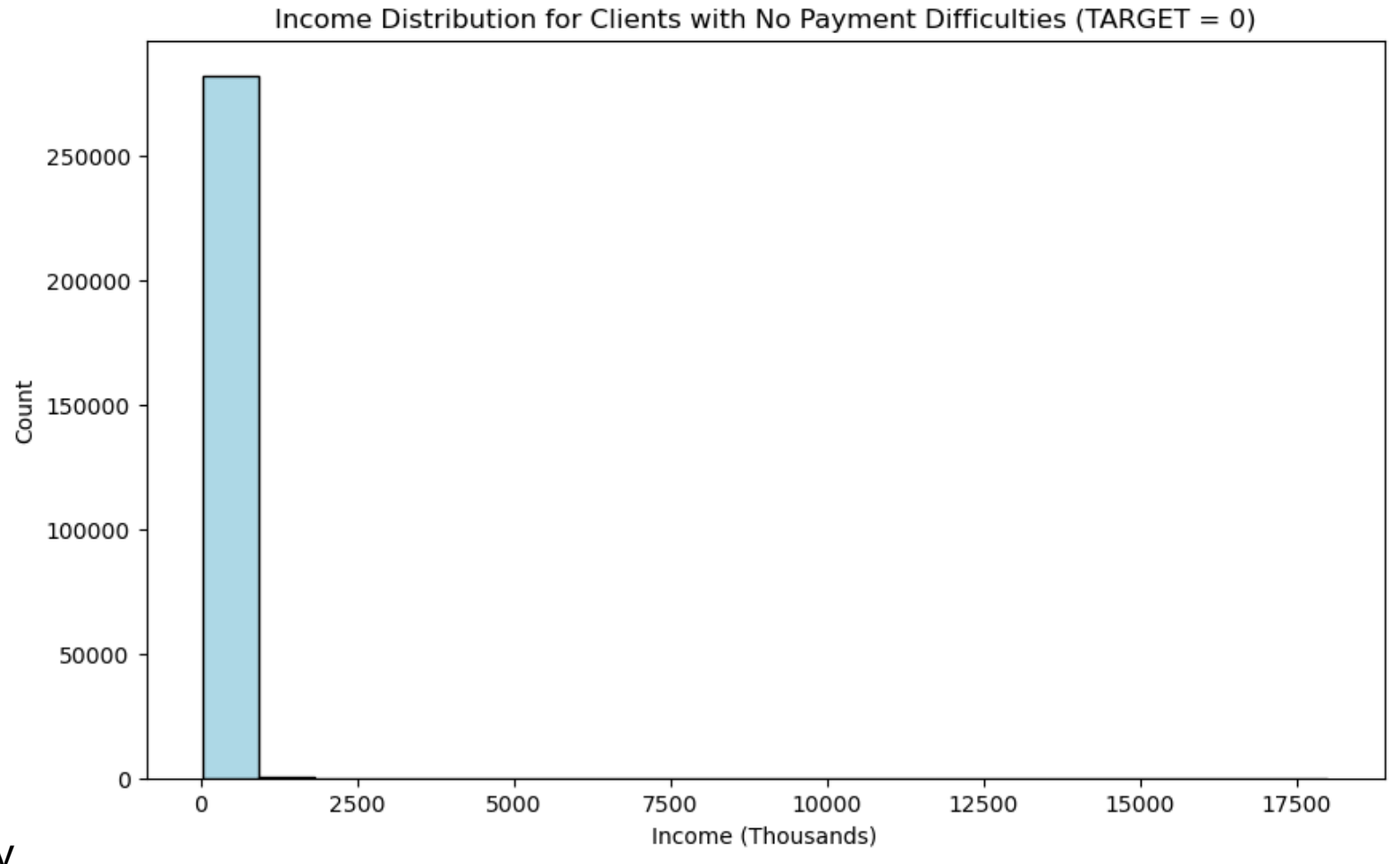
1. **Clients with No Payment Difficulties ('TARGET' = 0) with their OCCUPATION_TYPE.**

- **Graph Name:** Bar Chart
- **Purpose:** Visualize the distribution of clients with different occupations who do not have payment difficulties**.**
- **Insight:** Determine if there are different occupational distributions for clients with 'TARGET' = 0.
- **Analysis:** The graph shows that a significant number of individuals with an **'unknown'** occupation face no difficulty in repaying their loans. Following them are **laborers** and **sales** staff, who also exhibit a low incidence of payment difficulties.



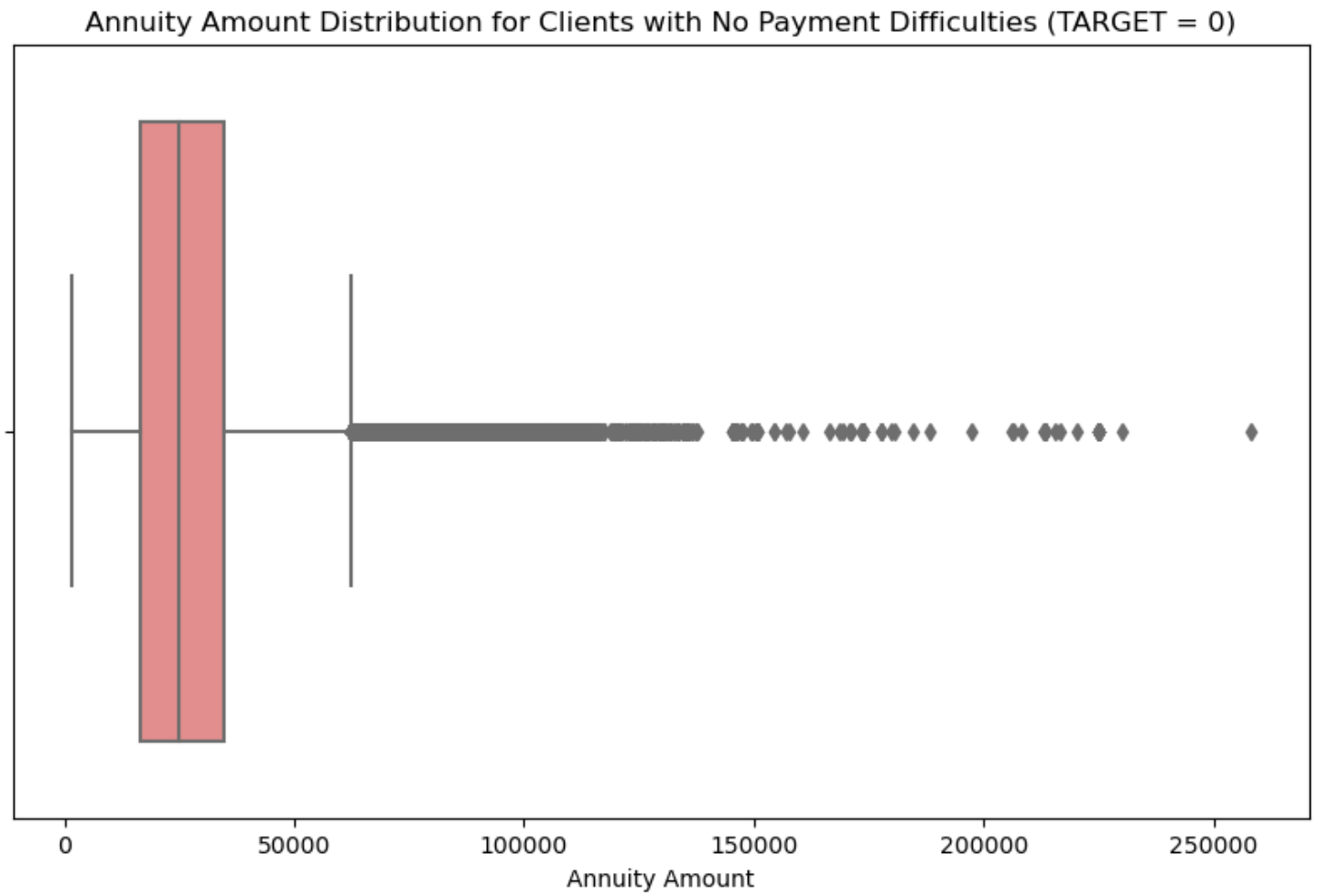
2. Clients with No Payment Difficulties ('TARGET’ = 0) with their Total Income -- 'AMT_INCOME_TOTAL’.

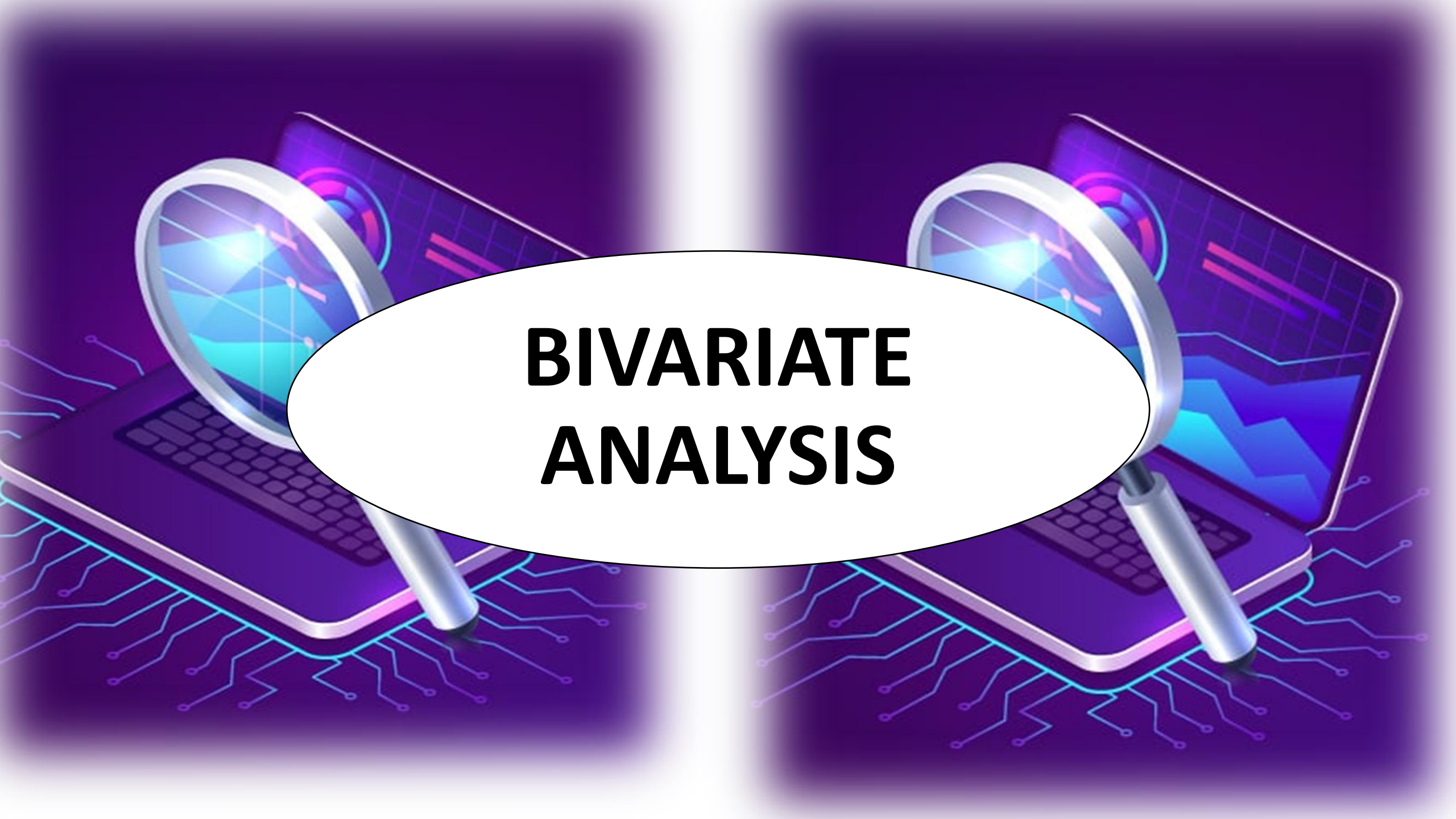
- **Graph Name:** Histogram
- **Purpose:** Analyze the income distribution for clients with no payment difficulties.
- **Insight:** Investigate if there are differences in income distribution for clients with 'TARGET' = 0.
- **Analysis:** The graph indicates that there is an equal income distribution among clients who are required to repay their loans.



3. Clients with No Payment Difficulties with Annuity Amount -- 'AMT_ANNUITY'.

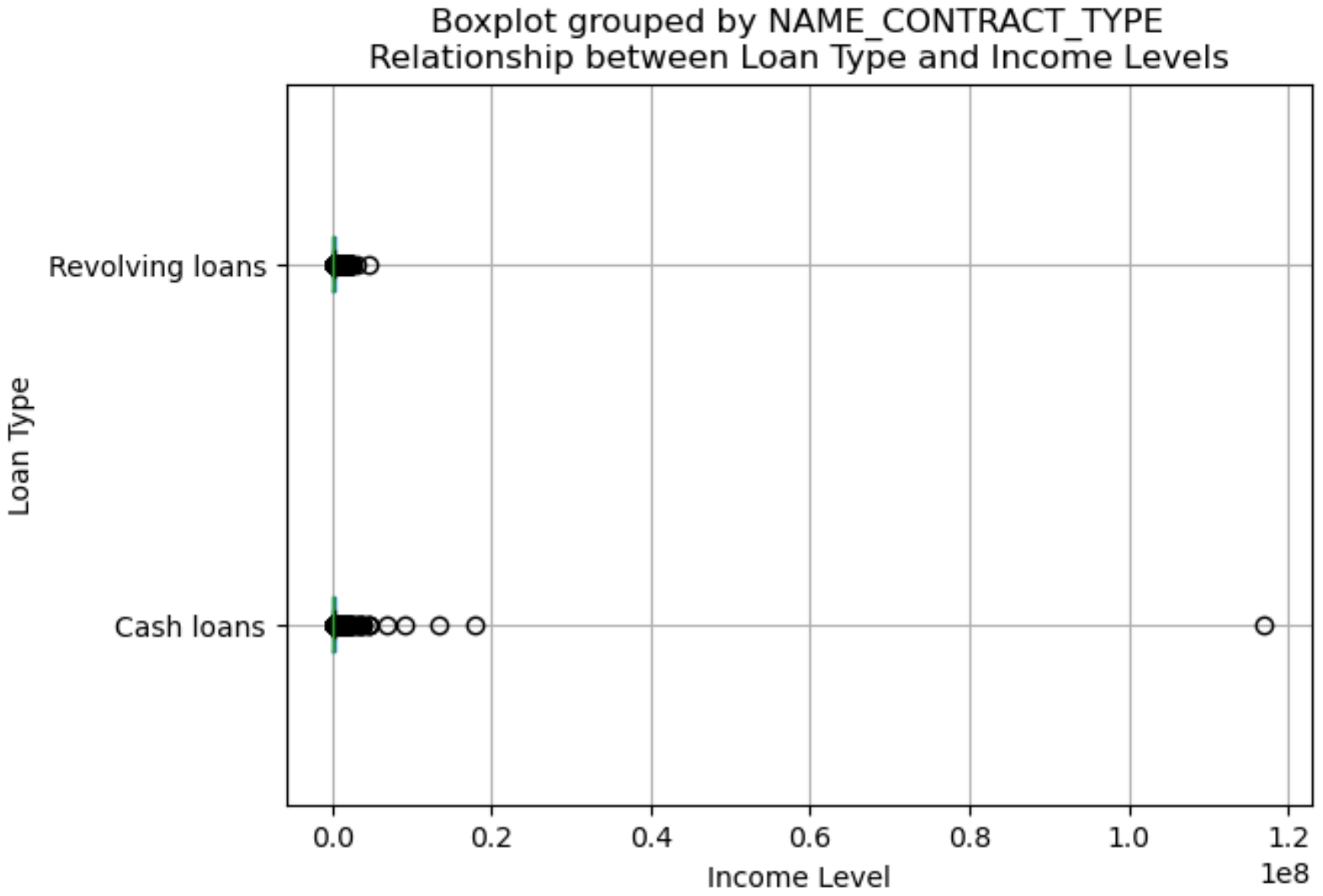
- **Graph Name:** Boxplot
- **Purpose:** To visually analyze the distribution of annuity amounts for clients who have no payment difficulties.
- **Insight:** This analysis helps us understand the range and spread of annuity amounts for clients who have successfully managed their payments, aiding in further insights into their financial profiles and payment behavior.
- **Analysis:** While the data suggests that clients with no payment difficulties tend to have annuity amounts **distributed** around the mean, the presence of *outliers and variability within the data* emphasizes the need for a more comprehensive credit risk assessment that considers other factors such as credit history, employment stability, and overall financial health.




BIVARIATE
ANALYSIS

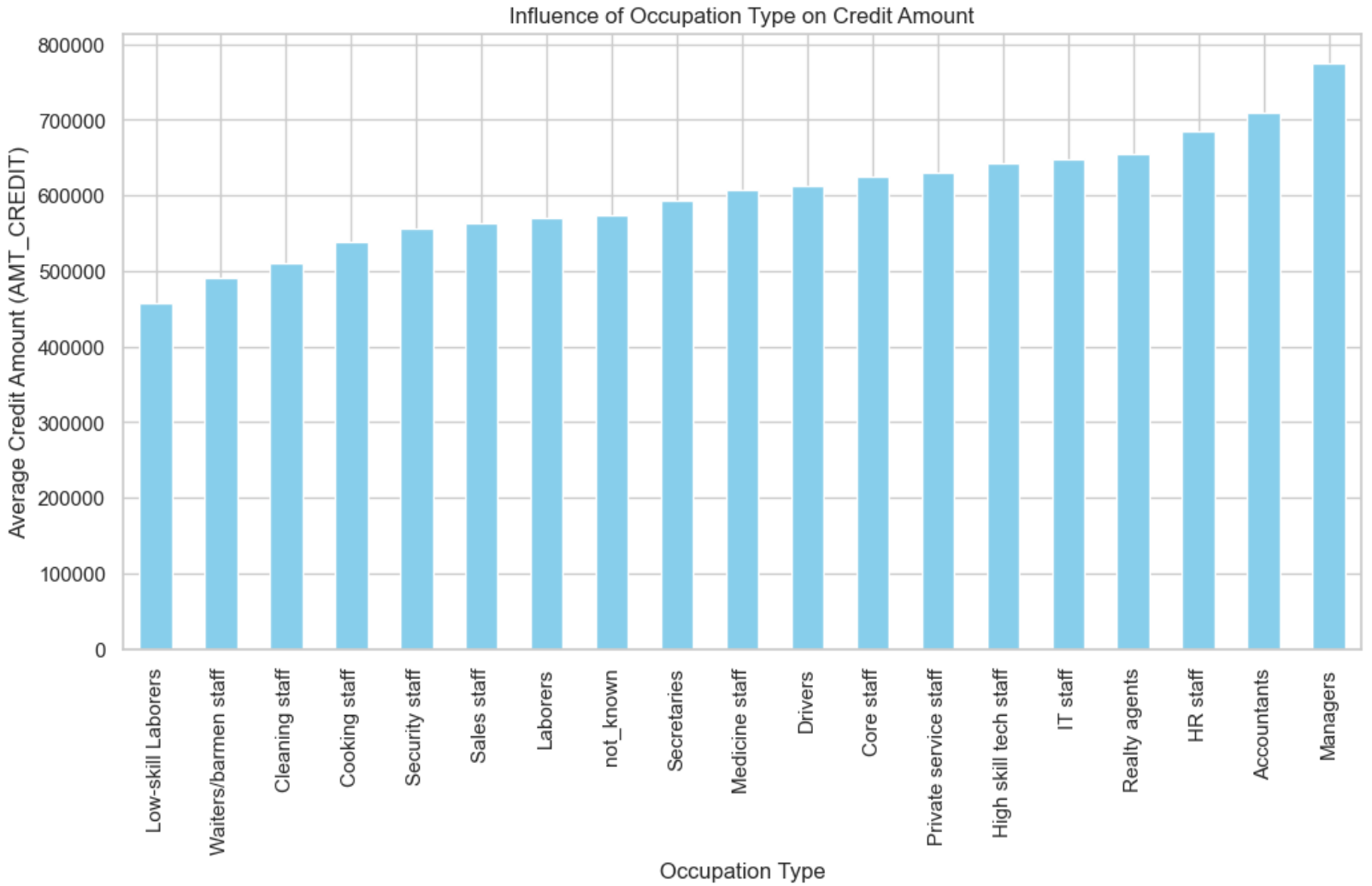
1. NAME_CONTRACT_TYPE vs. AMT_INCOME_TOTAL:

- **Graph Name:** Boxplot
- **Purpose:** To visually explore the relationship between the type of loan contract and the income levels of clients.
- **Insight:** The boxplot provides insights into the distribution of income levels among clients with different contract types.
- **Analysis:** Considering the overall shape of the boxplots and the distribution of income levels in the graph, it appears that the **"Cash loans"** category exhibits a **higher incidence of clients** with exceptionally high or low incomes compared to **"Revolving loans."**



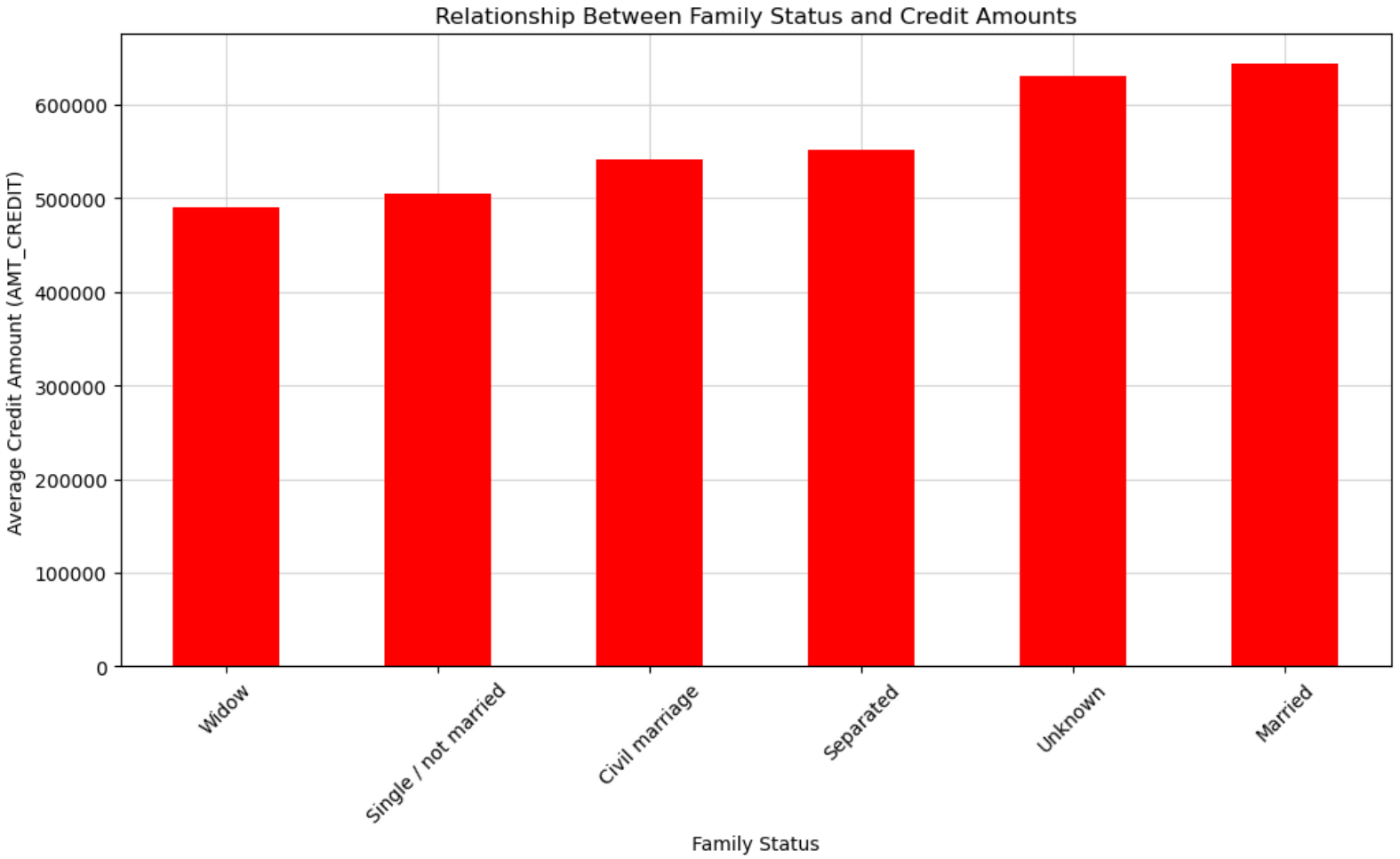
2. OCCUPATION_TYPE vs. AMT_CREDIT:

- **Graph Name:** Bar plot
- **Purpose:** To see how occupation influences the credit amount granted to clients of different professions.
- **Insight:** The bar plot visualizes the relationship between occupation type and the corresponding credit amounts, allowing for a comparison of credit amounts across different professions.
- **Analysis:** As seen in the graph, high-income occupations are associated with larger credit amounts, while others have more modest loan approvals. In summary, the analysis will help to understand the lending practices of financial institutions and how they can assess risk and determine loan amounts for different occupational groups.



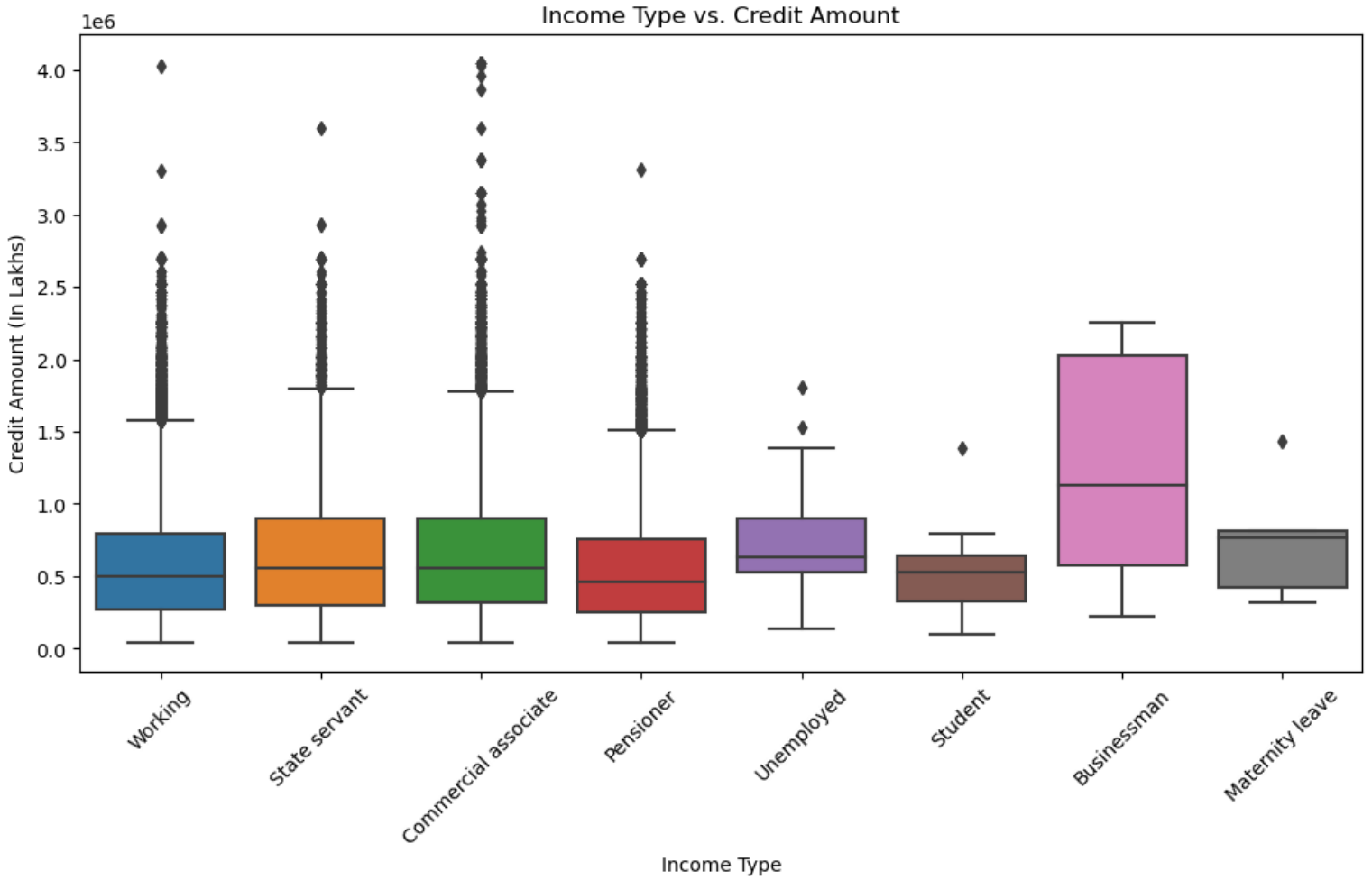
3. NAME_FAMILY_STATUS vs. AMT_CREDIT:

- **Graph Name:** Bar plot
- **Purpose:** To examine the relationship between family status and credit amounts granted to individuals and its influence of the credit amounts individuals receive.
- **Insight:** Family status provides insights into an applicant's financial stability and potential ability to meet repayment obligations.
- **Analysis:** When examining credit amounts across various family statuses, a pattern emerges where Married individuals tend to receive the highest credit amounts, followed by those with an unknown family status. Separated and Civil Marriage categories exhibit similar credit amounts, while Single individuals receive lower credit amounts. Widowed individuals tend to receive the lowest credit amounts among the family status groups.



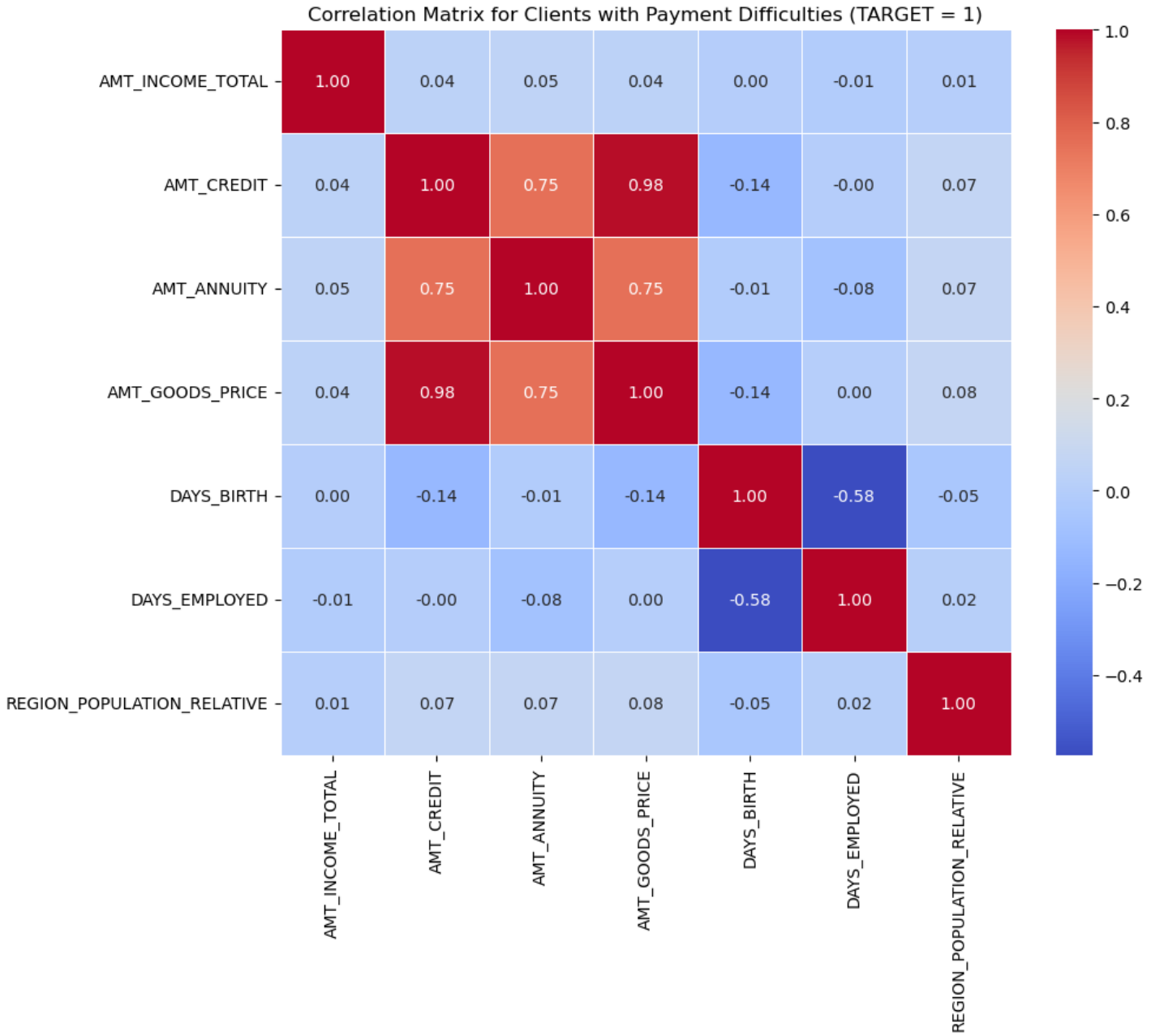
4. NAME_INCOME_TYPE vs. AMT_CREDIT:

- **Graph Name:** Boxplot
- **Purpose:** To visualize the relationship between **income sources** and the **credit amounts** granted to individuals. The goal is to understand how various income types are related to credit amounts.
- **Insight:** The box plot helps identify the distribution and central tendencies of credit amounts for different income sources, allowing for comparisons and insights into credit allocation based on income type.
- **Analysis:**
- According to the analysis of the Box plot, it is evident that **Businessmen** tend to secure the highest credit amounts, while **Students** tend to obtain the lowest credit amounts. There are numerous outliers in the income categories of **Commercial associate**, followed by **Working**, **Pensioner**, and **State servant**. These outliers may represent exceptional cases where individuals within these income sources receive either exceptionally high or low credit amounts.




MULTIVARIATE
ANALYSIS

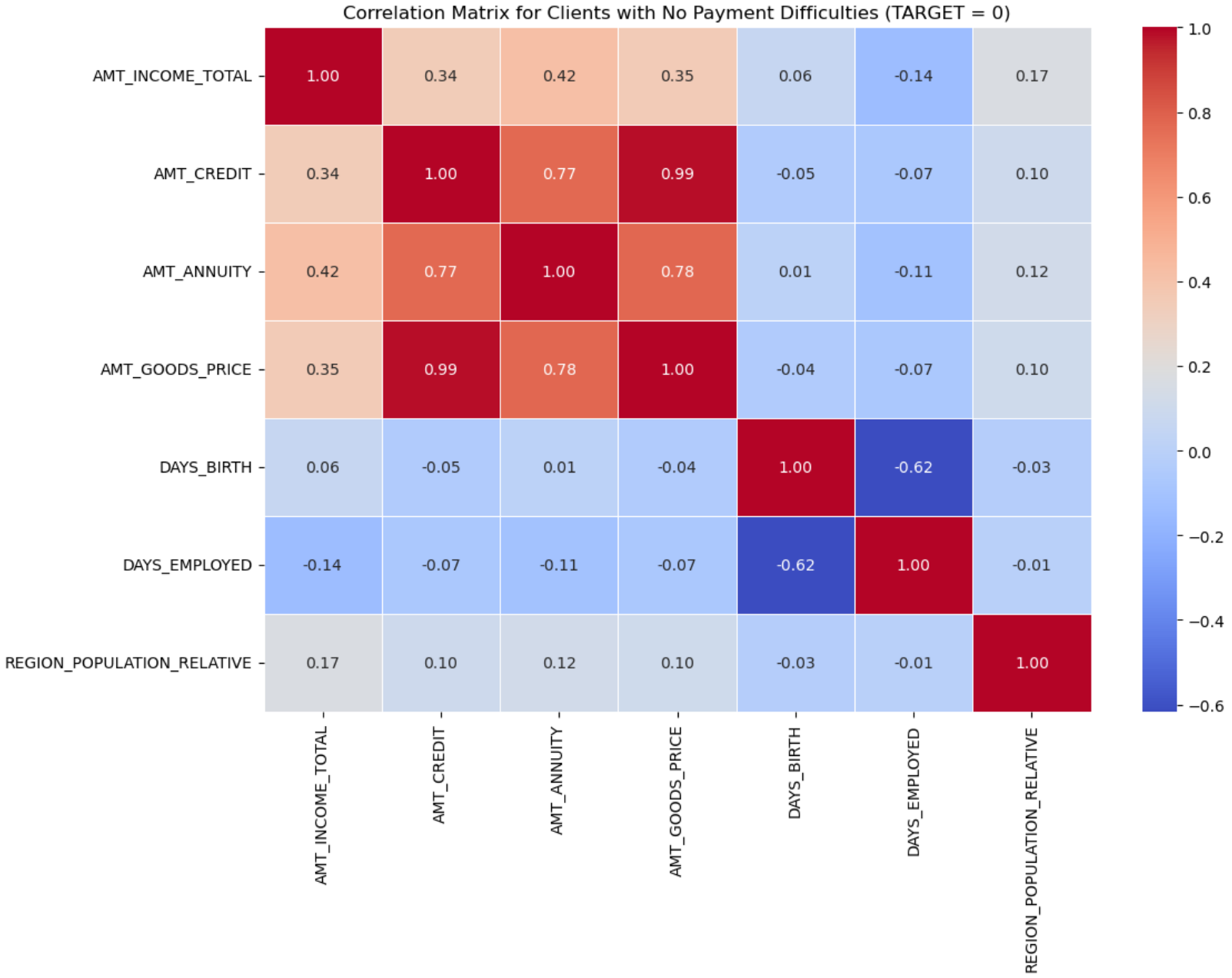
Correlation Analysis for Clients with Payment Difficulties ('TARGET' = 1) -- *Defaulters*



Analysis

- **AMT_INCOME_TOTAL:**
  - With AMT_ANNUITY and AMT_GOODS_PRICE, these correlations are relatively weak, indicating that income is slightly positively related to annuity and the price of goods for the loan.
  - With REGION_POPULATION_RELATIVE , indicating a slight positive association between income and the relative population density of the region.
- **AMT_CREDIT**:
  - With AMT_ANNUITY and AMT_GOODS_PRICE - This suggests a strong positive relationship between the credit amount and both the annuity and the price of goods.
  - With DAYS_BIRTH - The negative correlation implies that younger individuals tend to secure higher credit amounts.
- **AMT_ANNUITY**:
  - With AMT_CREDIT and AMT_GOODS_PRICE - This indicates that annuity amounts are strongly associated with credit amounts and borrower income levels.
  - With DAYS_EMPLOYED - This suggests that longer employment gaps or unstable employment histories may lead to lower annuity amounts.
- **AMT_GOODS_PRICE:**
  - With AMT_CREDIT - This suggests a strong positive relationship between the price of goods and the credit amount.
  - With REGION_POPULATION_RELATIVE - This indicates a positive association between the price of goods and the relative population density of the region.
- **DAYS_BIRTH:**
  - With AMT_CREDIT and AMT_GOODS_PRICE - These negative correlations suggest that younger individuals tend to secure higher credit amounts and request loans for less expensive goods.
- **DAYS_EMPLOYED:**
  - With AMT_ANNUITY - This suggests that longer employment gaps or unstable employment histories may lead to lower annuity amounts.
  - With REGION_POPULATION_RELATIVE - This indicates a slight positive association between employment history and the relative population density of the region.
- **REGION_POPULATION_RELATIVE:** - With AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, and AMT_GOODS_PRICE - These correlations suggest that as the population density of a region increases, there is a slight tendency for higher income levels, higher credit amounts, higher annuity payments, and more expensive goods for which loans are requested. This may indicate that regions with higher population density are associated with slightly improved financial metrics, but the associations are relatively weak.

Correlation Analysis for Clients with No Payment Difficulties ('TARGET' = 0)



Analysis

- **AMT_INCOME_TOTAL** :With AMT_ANNUITY, AMT_GOODS_PRICE, and AMT_CREDIT - This indicates that individuals with higher incomes tend to secure higher credit amounts, pay higher annuities, and request loans for more expensive goods.
- **AMT_CREDIT** :
  - With AMT_ANNUITY and AMT_GOODS_PRICE - This indicates that credit amounts are strongly correlated with the annuity to be paid and the price of goods.
  - With DAYS_BIRTH - The negative correlation implies that younger individuals tend to secure higher credit amounts.
- **AMT_ANNUITY** :
  - With AMT_CREDIT and AMT_GOODS_PRICE - This indicates a strong positive relationship between the annuity and both the credit amount and the price of goods.
  - With DAYS_EMPLOYED - This suggests that longer employment gaps or unstable employment histories may lead to lower annuity amounts.
- **AMT_GOODS_PRICE** :With AMT_CREDIT - This suggests a strong positive relationship between the price of goods and the credit amount.
- **DAYS_BIRTH** :
  - With AMT_INCOME_TOTAL - This suggests that older individuals tend to have slightly higher incomes.
  - With AMT_CREDIT and AMT_GOODS_PRICE - The negative correlations indicate that younger individuals tend to secure higher credit amounts and request loans for less expensive goods.
- **DAYS_EMPLOYED** : With AMT_INCOME_TOTAL , AMT_ANNUITY , and REGION_POPULATION_RELATIVE - This indicates that individuals with longer employment gaps or unstable employment histories tend to have lower incomes, lower annuity amounts, and may reside in regions with lower population density.
- **REGION_POPULATION_RELATIVE** : With AMT_INCOME_TOTAL, AMT_ANNUITY, and AMT_CREDIT - This suggests that individuals living in regions with higher population density tend to have higher incomes, higher annuity amounts, and are granted higher credit amounts.

Recommendation or Conclusion

- All the analyses highlight the importance of considering various factors in credit risk assessment, including income, age, employment history, and regional demographics.
- These insights can help in making informed lending decisions and managing credit risk.