Exploratory Data Analysis – Assignment 1 (Group-based except for part 2c)
Group Members: Andrew T Mavizha, Ernestina Hooper, Rellikson Kisyula
Dataset: Casedata.csv

Background and Business Importance

Millions of people with thin or no credit history still need safe and affordable loans. In this exploratory data analysis assignment, you will work with historical application and related credit data to uncover patterns that can help expand access responsibly. Your goal is to understand the data landscape including quality, coverage, and early signals of repayment behavior and to surface people focused insights that a lender could act on next without building a predictive model. Your EDA should highlight where safe access can be widened while protecting both borrowers and the lender from avoidable risk.

Deliverable: Submit one document per group, written in clear, plain, nontechnical language for the client. In Section 2c, clearly identify each member's individual contributions.

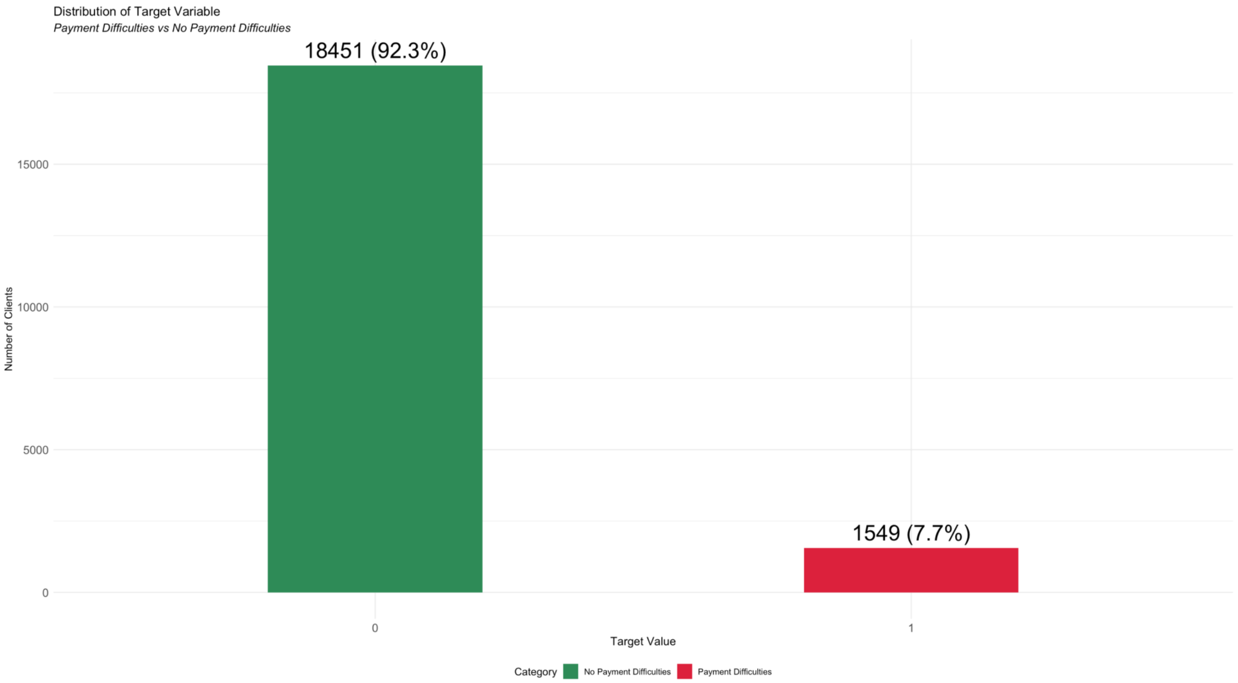
1. Case Context and Problem Framing

- What is the core business problem to solve? Explain the relevance of this problem.
  - *The core business problem is predicting loan default risk to help expand safe and affordable credit access to underserved populations with thin or no credit history. This is highly relevant because:*
    - *Millions of people are excluded from traditional lending due to lack of credit history*
    - *Lenders need to balance expanding access with protecting against defaults*
    - *Loan defaults directly affect profitability and sustainability of lending operations*
    - *Safe credit access can help people build financial stability and economic mobility*
  - *The goal is to identify patterns in historical data that can help lenders make better decisions about who to approve for loans while minimizing risk to both borrowers and lenders.*
- Identify whether this is a supervised or unsupervised learning problem. If supervised, specify whether it is a classification or regression problem and justify your reasoning.
  - *This is a supervised learning problem, specifically a binary classification problem.*
  - *Justification:*
    - *Supervised: We have a labeled target variable (TARGET) indicating whether clients had payment difficulties*
    - *Classification: The target variable is binary (0 = no payment difficulties, 1 = payment difficulties)*
    - *Binary: There are only two possible outcomes we're trying to predict*
  - *The model would learn from historical loan application data and outcomes to classify new applicants into "likely to repay" vs "likely to default" categories.*

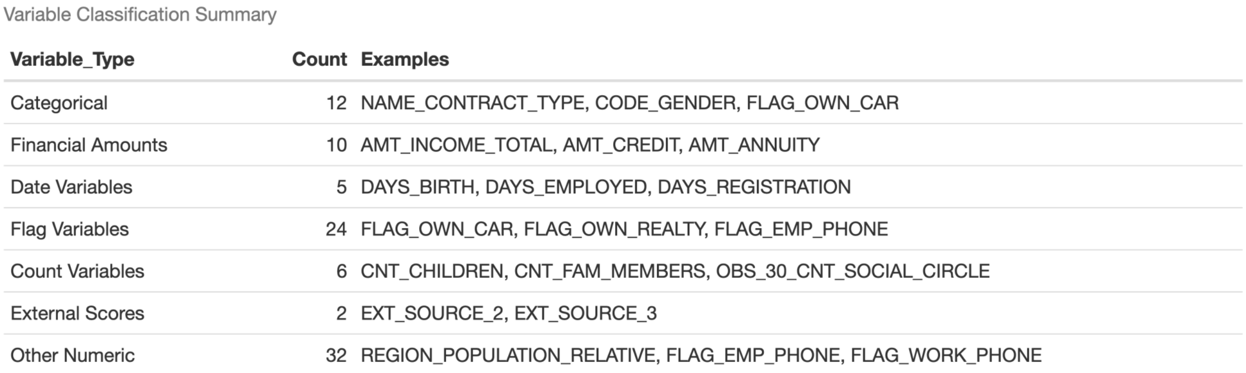
2. Data Exploration (EDA)

a. General Overview of the Dataset

- What is the response variable? Provide one appropriate visualization or table to show the distribution of the response variable.
  - *The response variable is TARGET, which indicates whether a client had payment difficulties:*
    - *0: Client had no payment difficulties (good repayer)*
    - *1: Client had payment difficulties (late payment more than X days)*



- State the number of observations and predictor variables in the dataset.
  - *Dataset Dimensions:*
    - *Total observations (rows): 20,000*
    - *Total variables (columns): 68*
    - *Predictor variables: 67*
    - *Response variable: 1 (TARGET)*
- Classify variables as categorical, continuous, date variables, or other type.



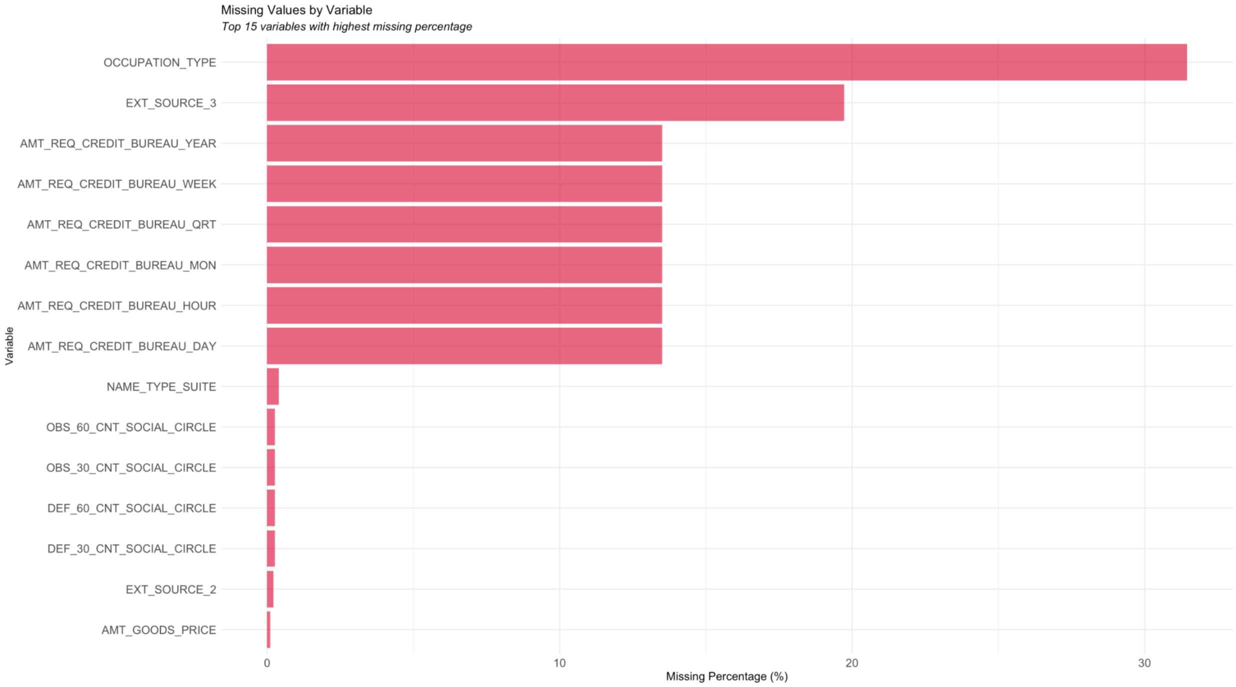
Variable Classification Summary

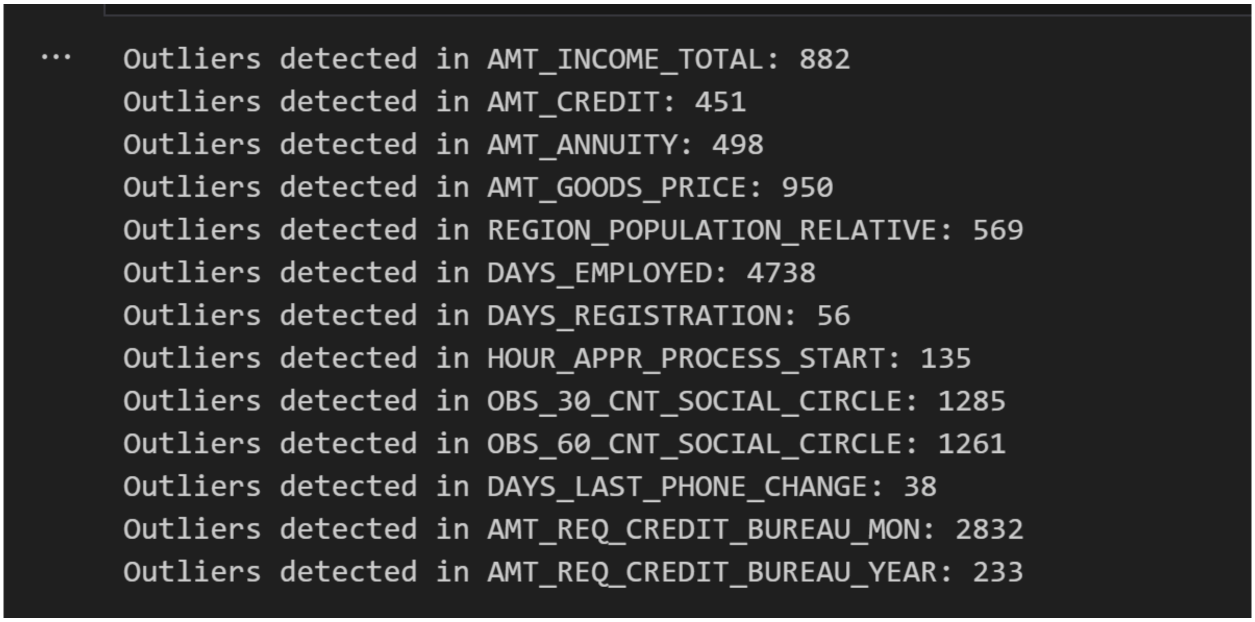
Variable_Type	Count	Examples
Categorical	12	NAME_CONTRACT_TYPE, CODE_GENDER, FLAG_OWN_CAR
Financial Amounts	10	AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY
Date Variables	5	DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION
Flag Variables	24	FLAG_OWN_CAR, FLAG_OWN_REALTY, FLAG_EMP_PHONE
Count Variables	6	CNT_CHILDREN, CNT_FAM_MEMBERS, OBS_30_CNT_SOCIAL_CIRCLE
External Scores	2	EXT_SOURCE_2, EXT_SOURCE_3
Other Numeric	32	REGION_POPULATION_RELATIVE, FLAG_EMP_PHONE, FLAG_WORK_PHONE

b. Data Quality Assessment

- Examine the dataset for missing values, NULL values, outliers, or inconsistencies. Summarize any issues found and discuss their implications for analysis.

```
## Top 10 variables with most missing values:
##
##
## |Variable                   | Missing_Count| Missing_Percentage|
## |:--------------------------|-------------:|------------------:|
## |OCCUPATION_TYPE            |          6290|              31.45|
## |EXT_SOURCE_3               |          3945|              19.73|
## |AMT_REQ_CREDIT_BUREAU_HOUR |          2701|              13.51|
## |AMT_REQ_CREDIT_BUREAU_DAY  |          2701|              13.51|
## |AMT_REQ_CREDIT_BUREAU_WEEK |          2701|              13.51|
## |AMT_REQ_CREDIT_BUREAU_MON  |          2701|              13.51|
## |AMT_REQ_CREDIT_BUREAU_QRT  |          2701|              13.51|
## |AMT_REQ_CREDIT_BUREAU_YEAR |          2701|              13.51|
## |NAME_TYPE_SUITE            |            81|               0.40|
## |OBS_30_CNT_SOCIAL_CIRCLE   |            54|               0.27|
```





```
...  Outliers detected in AMT_INCOME_TOTAL: 882
     Outliers detected in AMT_CREDIT: 451
     Outliers detected in AMT_ANNUITY: 498
     Outliers detected in AMT_GOODS_PRICE: 950
     Outliers detected in REGION_POPULATION_RELATIVE: 569
     Outliers detected in DAYS_EMPLOYED: 4738
     Outliers detected in DAYS_REGISTRATION: 56
     Outliers detected in HOUR_APPR_PROCESS_START: 135
     Outliers detected in OBS_30_CNT_SOCIAL_CIRCLE: 1285
     Outliers detected in OBS_60_CNT_SOCIAL_CIRCLE: 1261
     Outliers detected in DAYS_LAST_PHONE_CHANGE: 38
     Outliers detected in AMT_REQ_CREDIT_BUREAU_MON: 2832
     Outliers detected in AMT_REQ_CREDIT_BUREAU_YEAR: 233
```

Key Findings and Summary:

1. ***Missing Values****: The dataset shows some levels of missing data across the variables, with some variables having substantial missing rates that may require dropping the variable or inputting the missing values.*
2. ***Outliers****: Financial variables show expected outliers in income and credit amounts, which are typical in financial data and may represent legitimate high-value cases.*
3. ***Data Consistency****: Most variables show expected ranges and formats, with date variables properly encoded as days before application.*

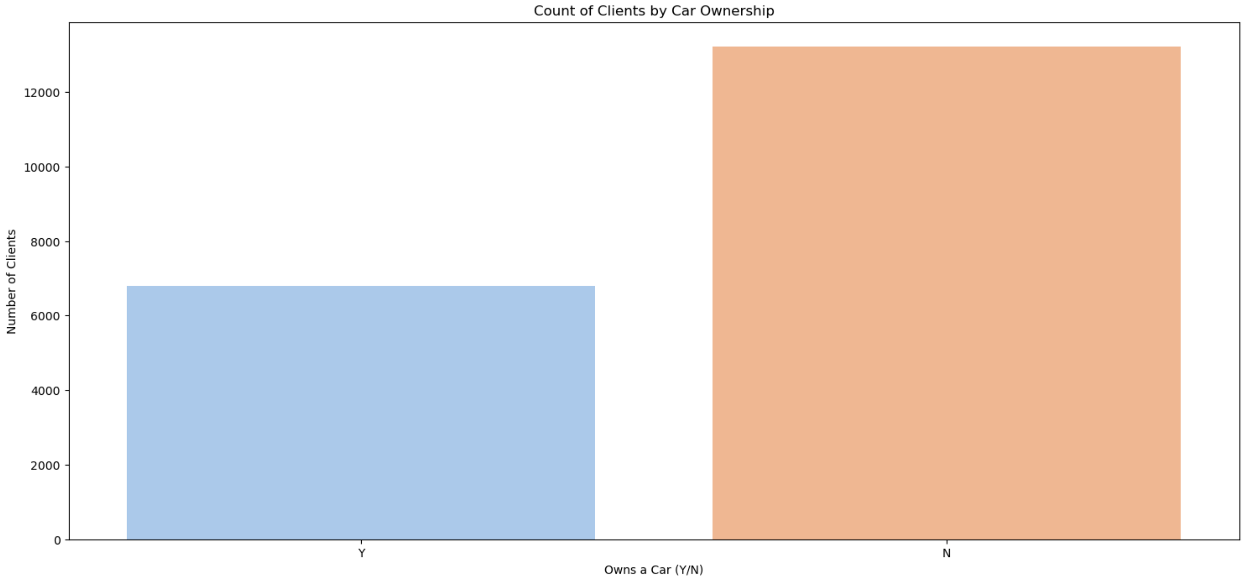
Implications for Analysis:

- *The data seems to be clean and consistent, with no major issues found. The two variables with the most missing data are OCCUPATION_TYPE and EXT_SOURCE_3*

c. Key Visualizations and Insights (Each group member should provide their own answers to this portion and please label who did what.)

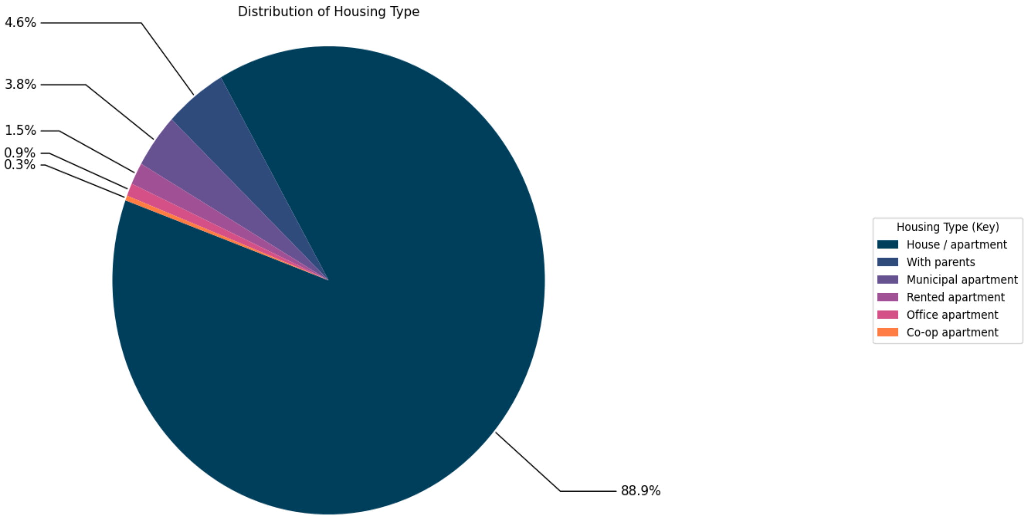
- Explore the data to discover meaningful patterns. Include:
    - Two plots examining trends, distributions of a single variable on its own.
    - Two plots showing the relationship between the response variable and one or more predictor variables, highlighting significant trends or patterns.
  - For each plot, include two brief notes:
      - Purpose: Why this chart is appropriate and what the pattern shows.
      - Business takeaway: What the result means for the client’s decisions.

Andrew



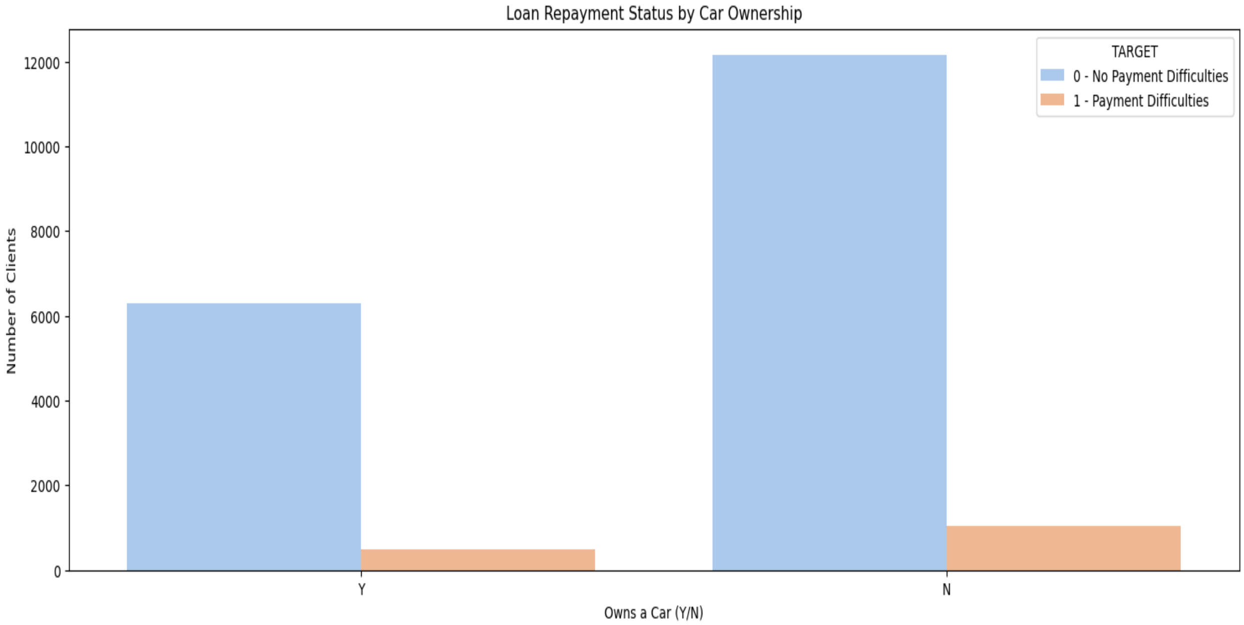
Purpose: This chart shows us whether our loan applicants own cars or not. It helps us understand the lifestyle and asset profile of our clients.

Business Takeaway: Most of our clients do not own cars, which suggests they may rely on public transportation or have lower asset ownership. This could mean they are more sensitive to financial pressures and need smaller, more affordable loan options. On the other hand, the group that does own cars may have higher asset values and could qualify for larger loans.



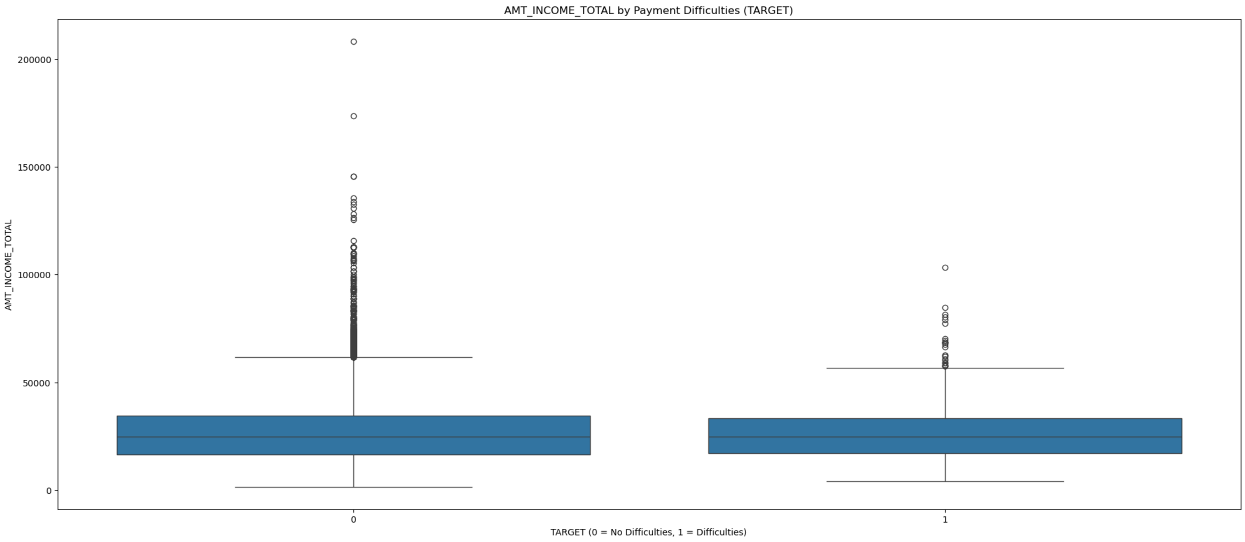
Purpose: This chart shows the distribution of housing types among our loan applicants. It helps us see whether clients are mostly homeowners, renters, or living with parents, which can signal financial stability and repayment capacity.

Business Takeaway: The majority of clients nearly 89% live in their own house or apartment, suggesting most have stable housing and possibly long-term roots in their communities. A smaller group lives with parents or in rented/municipal housing, which may indicate younger applicants or those with fewer assets. For these clients, loans may need to be structured with more flexible terms.



Purpose: This chart shows how loan repayment status varies by car ownership. It helps us see whether owning a car is linked to higher or lower likelihood of repayment difficulties.

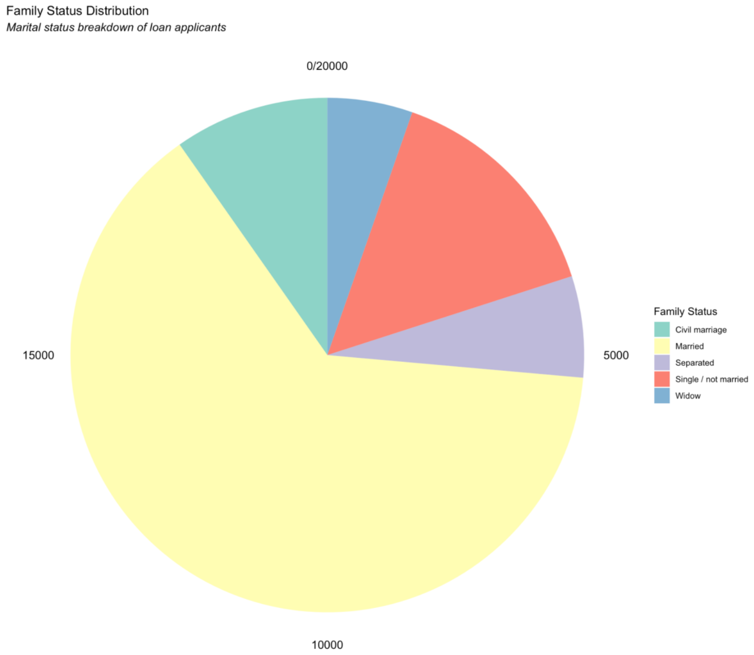
Business Takeaway: Most clients, whether they own cars or not, do not experience repayment difficulties. However, those without cars make up a larger share of borrowers overall and also account for more repayment difficulties. This suggests that clients without cars may face greater financial strain, making them slightly riskier borrowers. By contrast, clients who own cars not only are fewer in number but also show relatively fewer repayment problems, hinting at higher financial stability.



Purpose: This chart shows the distribution of applicants' total income, separated by whether they experienced payment difficulties (1) or not (0). It helps us understand whether income levels are linked to repayment challenges.

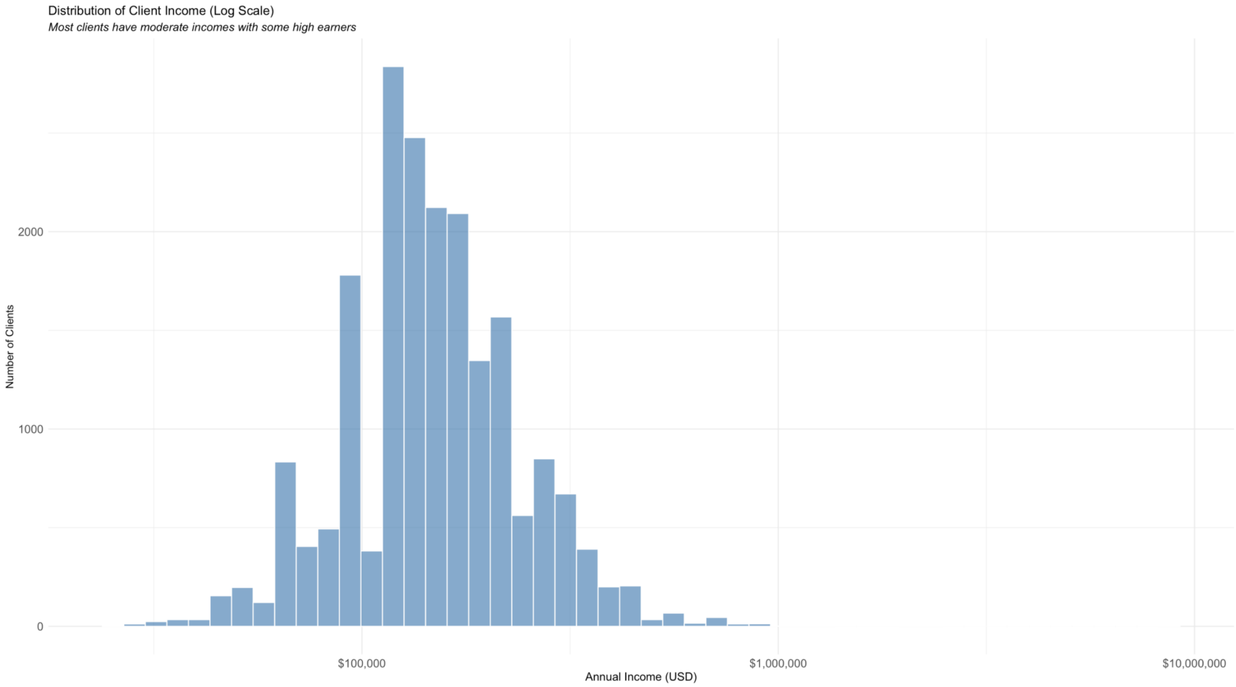
Business Takeaway: Most of our clients, both with and without payment difficulties, fall into a moderate-income range. However, applicants with no repayment issues (Target = 0) include more high-income outliers, while those with repayment difficulties (Target = 1) cluster around lower incomes. This suggests that income stability plays a role in repayment success: higher earners are less likely to default, while lower earners may struggle more. For the business, this means we should continue serving middle-income clients as our core base but design products that keep repayment manageable for lower-income borrowers while carefully monitoring loan sizes for risk management.

Rellikson



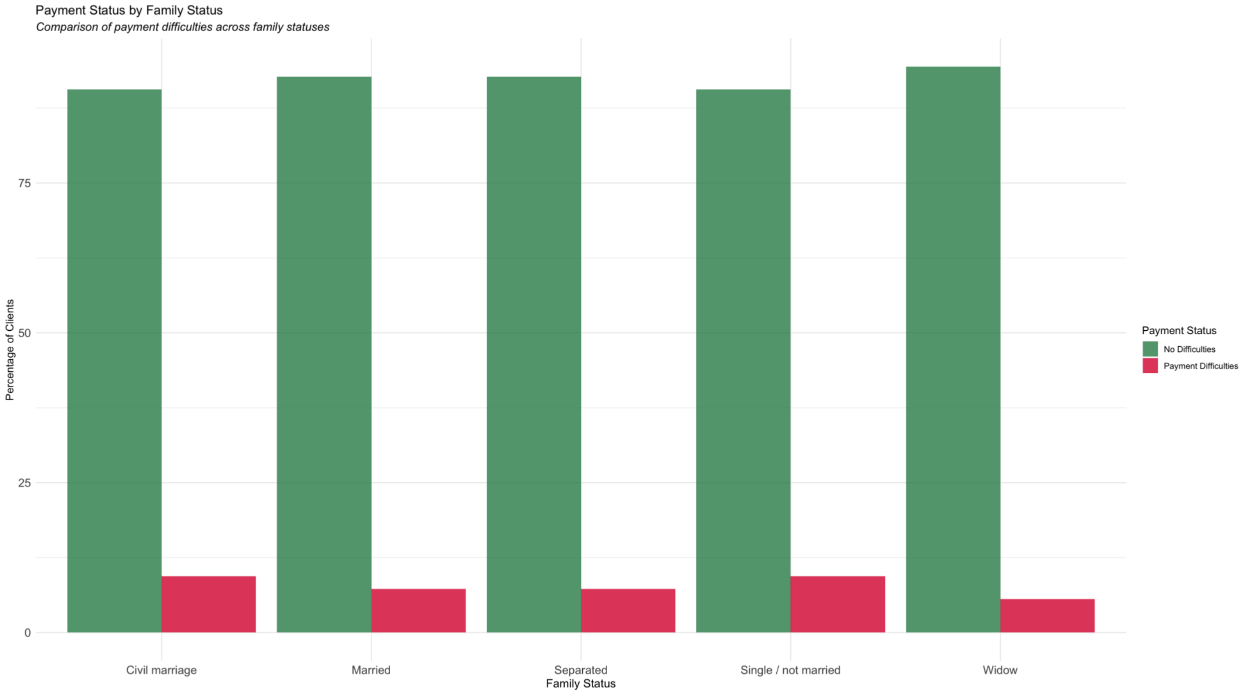
Purpose: This pie chart shows the family status composition to understand household structure patterns among applicants.

Business Takeaway: Married clients represent a significant portion, suggesting stable household income potential. Single clients may need different risk assessment criteria, while family status could influence loan terms and repayment capacity.



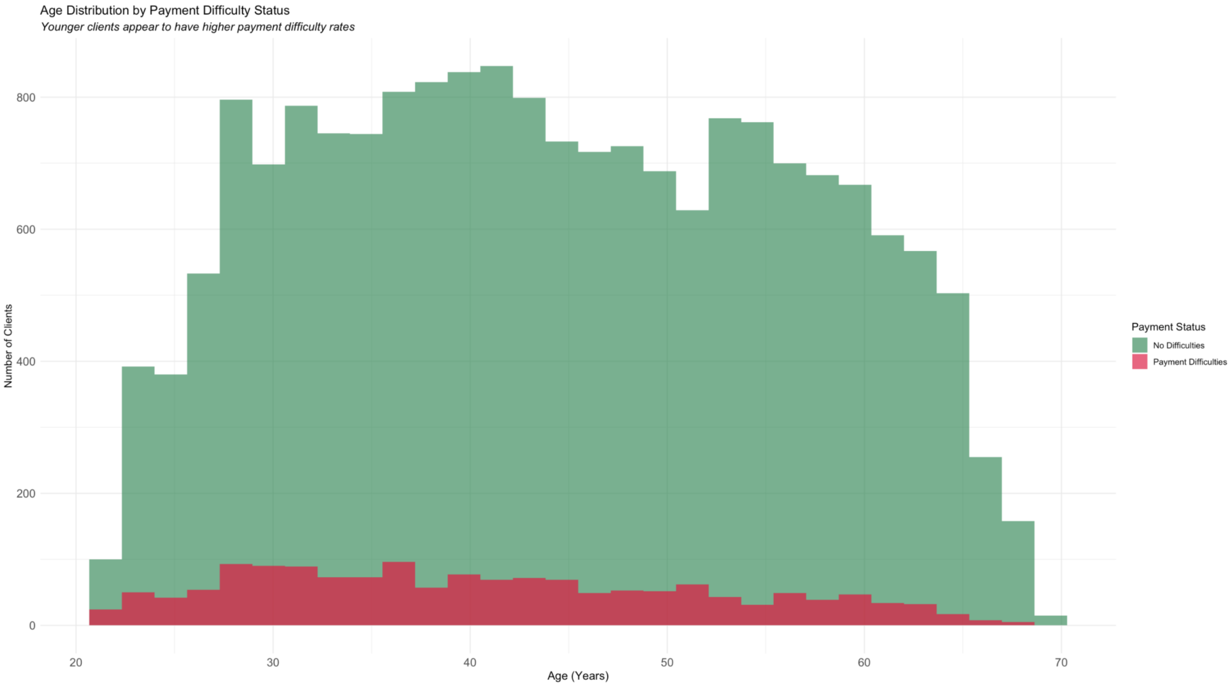
Purpose: This chart shows us what kinds of incomes our loan applicants have - basically, are we mostly serving wealthy people, poor people, or people in the middle?

Business Takeaway: Most of our clients earn moderate amounts of money, which means we're serving everyday working people rather than just the wealthy. We do have some high earners who might want bigger loans, and we also have lower-income people who need us to keep loans affordable and accessible.



Purpose: This grouped bar chart compares payment difficulty rates across different family statuses to understand relationship stability as a risk factor.

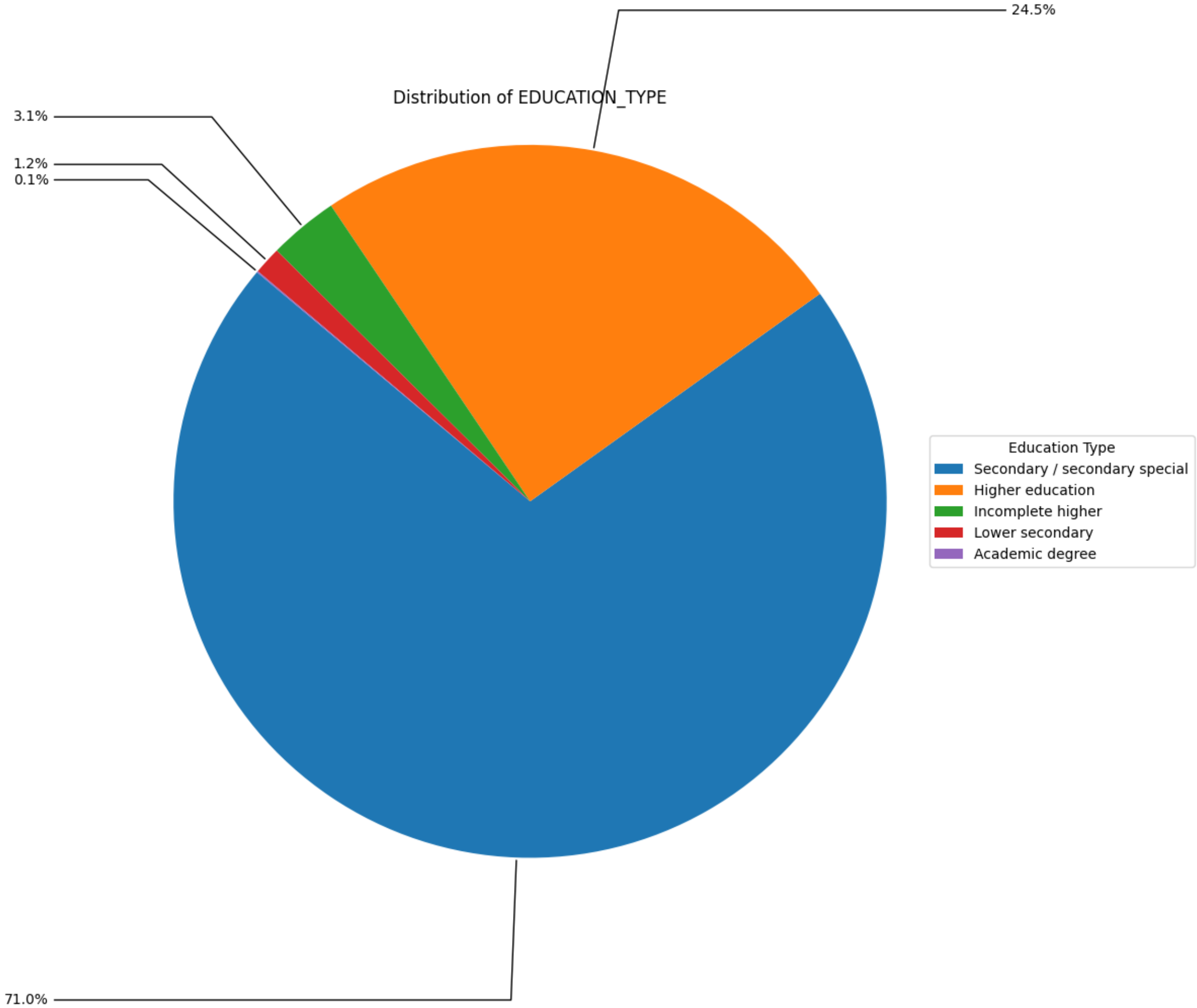
Business Takeaway: Married and Widow clients show lower default rates compared to single/divorced clients, suggesting relationship stability influences financial stability. This supports using family status as a risk assessment factor and potentially offering family-oriented financial products.



Purpose: This looks at whether age makes a difference in loan repayment - are younger or older borrowers more reliable?

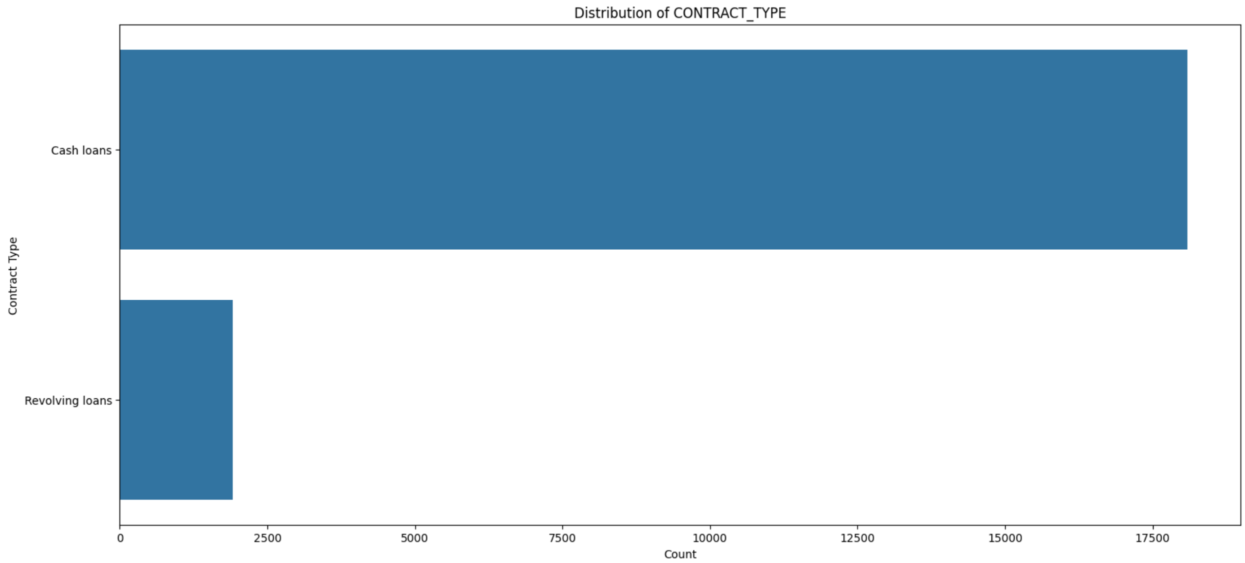
Business Takeaway: Experience really does matter! Younger people (especially those under 30) struggle more with loan payments than older, more experienced borrowers. This suggests we should provide extra support and guidance for younger borrowers, maybe even requiring financial education or mentorship programs.

Ernestina

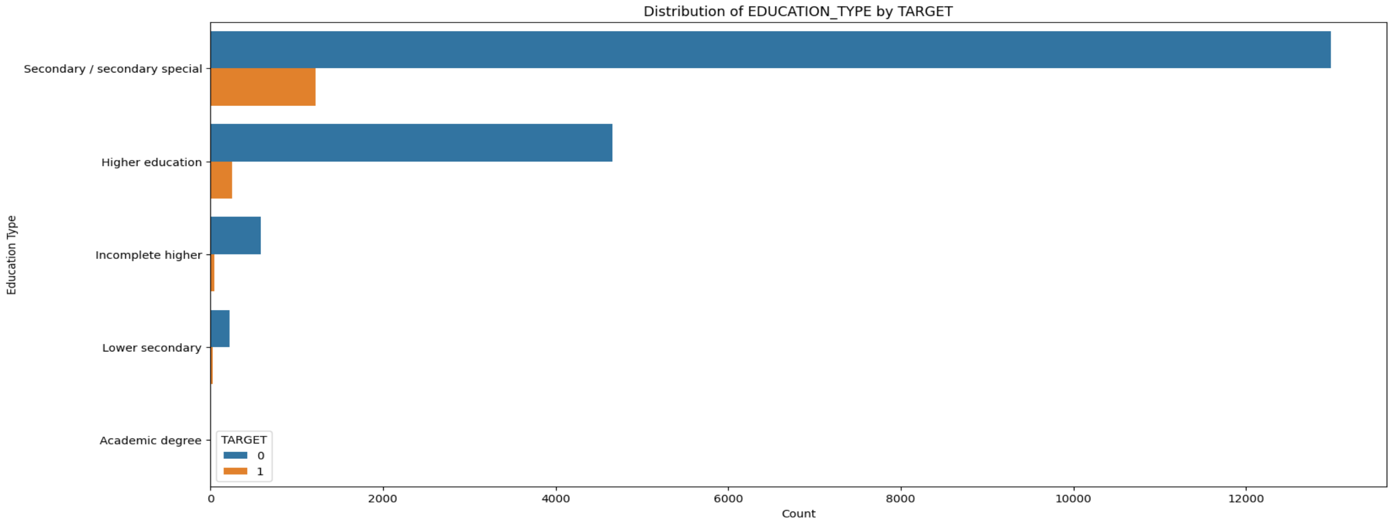


Purpose: This pie chart gives us a picture of the educational backgrounds of the people applying for loans. It shows that most applicants have a secondary education or some kind of specialized secondary training.

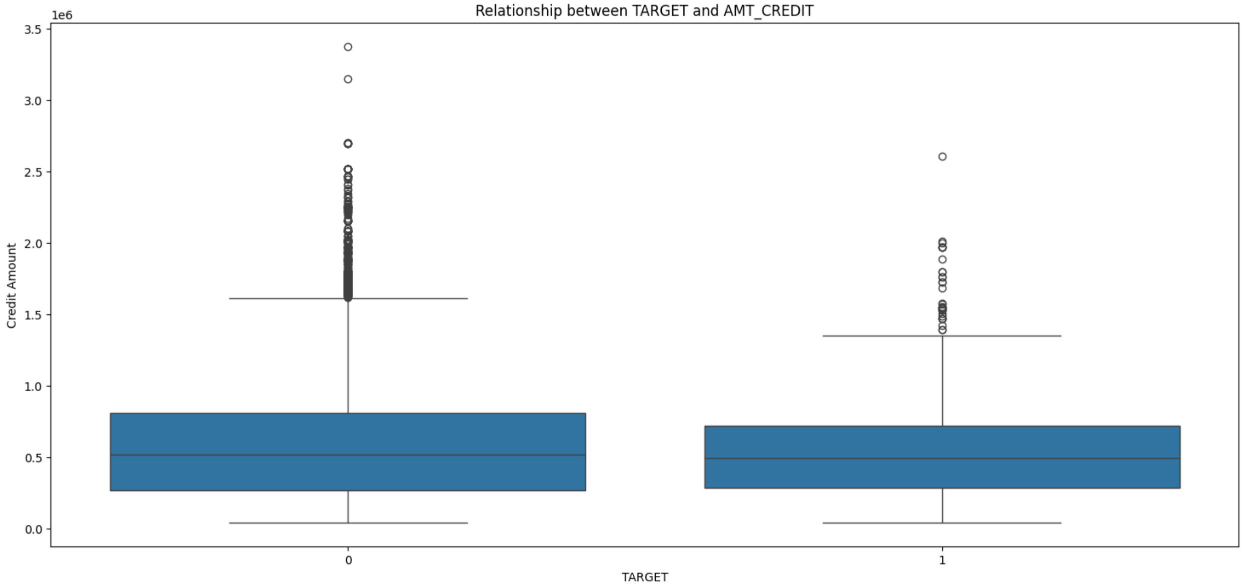
Business Takeaway: Most of the applicants fall into the secondary and higher education categories. Therefore, we should focus our marketing, and loan offers on these groups.



Purpose: The bar chart shows that applicants are much more likely to take out cash loans than revolving loans.
Business Takeaway: Most applicants prefer cash loans, which shows that this is the client's main area of business. This insight can help in deciding how to allocate resources and plan future product development.



Purpose: This bar chart explores the relationship between education level (NAME_EDUCATION_TYPE) and loan repayment behavior (TARGET). It shows the number of applicants in each education category: whether they repaid successfully (TARGET=0) or had repayment issues (TARGET=1).
Business Takeaway: For business decisions, the focus is on the default rate (the percentage of applicants who default within each education group). If certain education levels show higher default rates, that would signal higher-risk groups. This insight could be used in credit scoring models and adjusting loan conditions.



Target: This box plot compares loan amounts (AMT_CREDIT) between applicants who repaid their loans (TARGET=0) and those who didn't (TARGET=1).
Business Takeaway: This box plot helps us know if larger loan amounts are linked to a higher risk of default? If the group with repayment issues (TARGET=1) has higher loan amounts, that suggests bigger loans may carry greater risk. This guides business decisions about setting loan limits and improving risk strategies.

3. Recommendations for Actionable Insights

Based on your group's analysis, provide the most interesting findings to the client:

- What do your findings reveal about the characteristics of those repaying loans or not repaying loans?
    - After carefully studying 20,000 loan applications, we found clear patterns that separate successful borrowers from those who struggle with payments. Think of it like this: some people are naturally better positioned to handle a loan, while others face challenges that make repayment harder.
        - **People with steady, higher incomes**: When someone earns more money, they simply have more left over each month to make loan payments.
        - **Older, more experienced borrowers**: People over 35 tend to be more settled in their careers and have learned how to manage money through life experience.
        - **Education makes a difference**: Those with college or higher education seem to understand financial commitments better and plan accordingly.
    - People Who Face Payment Difficulties
        - **Lower-income borrowers:** When you're already stretching every dollar, unexpected expenses can quickly derail loan payments.
        - **Younger borrowers (18-30)**: Often still figuring out career paths and money management, making them more vulnerable to financial surprises.
        - **Limited education**: Without exposure to financial concepts, it's harder to fully understand long-term loan commitments.
- Our Recommendations
    - **Design Loans for Different Life Situations**:
        - Small amounts for young people or first-time borrowers to build their track record
        - Better terms for married couples since they're statistically safer
        - Automatic limits that ensure no one borrows more than they can realistically repay
    - **Build Safety Nets**

    - Allow trusted family members to back up higher-risk loans
    - Lower rates for safer borrowers, slightly higher for those who need extra support
    - Use data to spot payment problems before they become serious
  - **Use data**
    - Track how different groups are performing and adjust our approach
    - Test new approaches with small groups before expanding