Credit Risk Exploratory Data Analysis Report

Data Bootcamp

Group 12

Skye Xi   Jonathan Cain   Sab Rajesh Krishnan

February 20, 2026

Contents

1 Business Goal

Consumer finance lenders must balance two competing risks:

- **Rejecting a reliable applicant** leads to opportunity loss.
- **Approving a risky applicant** increases expected credit losses.

The objective of this EDA is to identify variables and segments associated with repayment difficulties and decision outcomes.

2 Data Description

We use:

- Application data with repayment label (`TARGET`).
- Previous-application outcomes (`NAME_CONTRACT_STATUS`).
- A data dictionary (`columns_description.xlsx`).

2.1 Key Variables Used

- `TARGET`: 1 = payment difficulties, 0 = on-time payments.
- `DAYS_EMPLOYED`, `AMT_INCOME_TOTAL`.
- Age derived from `DAYS_BIRTH` and binned to `YEARS_BIRTH_CATEGORY`.
- `NAME_FAMILY_STATUS`, `NAME_EDUCATION_TYPE`, `NAME_INCOME_TYPE`, `CODE_GENDER`.
- `AMT_CREDIT`, `AMT_GOODS_PRICE`.
- `NAME_CLIENT_TYPE`, `NAME_CONTRACT_STATUS`.

2.2 Data Cleaning and Outlier Treatment

Cleaning steps included:

- Replacing known anomalous codes (e.g., `DAYS_EMPLOYED = 365243`) with missing values.
- Converting negative-coded day variables into positive, interpretable measures for plotting.
- Applying high-percentile caps for visualization to avoid axis compression.

3 Overall Distributions

Key numeric variables (income, credit, goods price) exhibit right-skewness and heavy tails. We used histograms/boxplots and applied (i) log scaling for visualization where appropriate and (ii) percentile-based capping for extreme outliers.

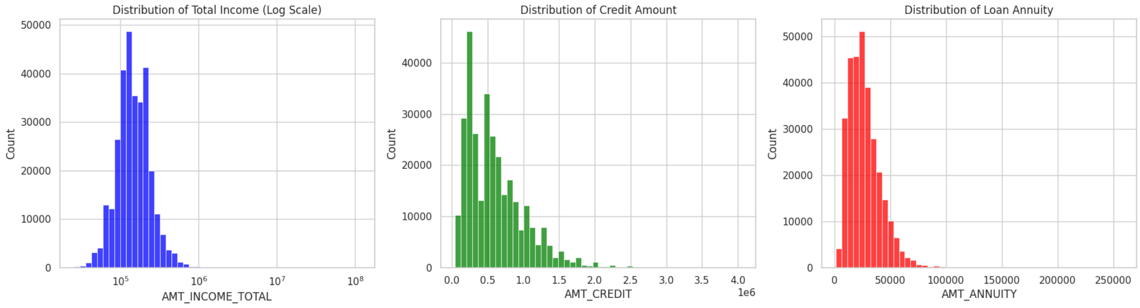
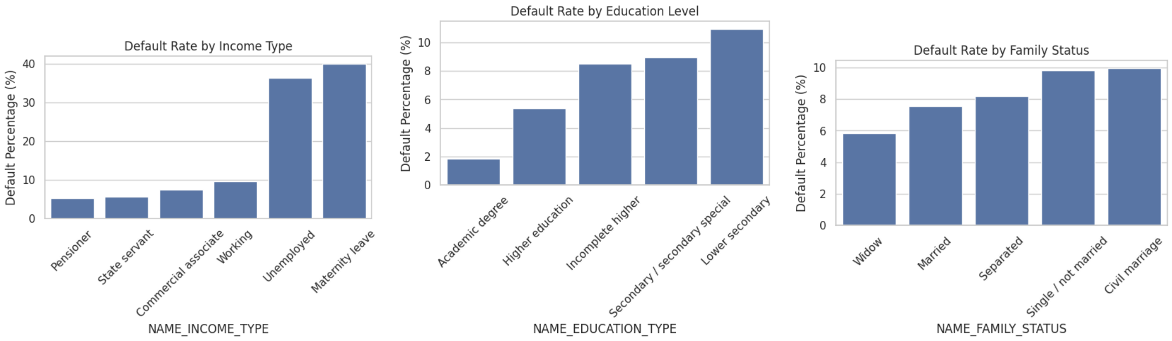


Figure 1: Distribution Analysis

4 Univariate Analysis

4.1 Categorical Variables



(a) (b) (c)

Figure 2: Categorical Analysis

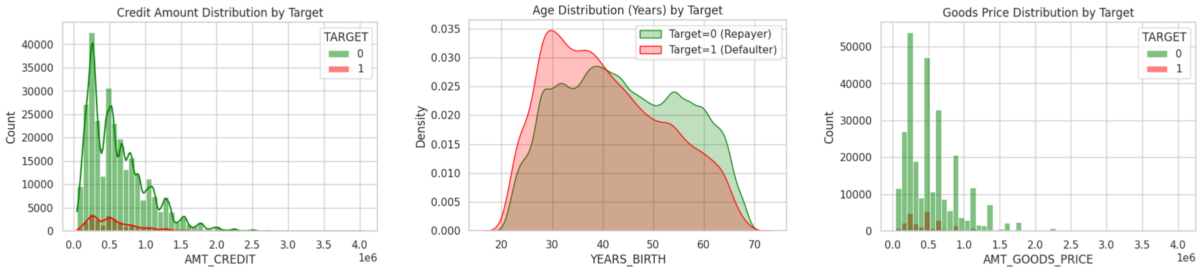
The bar chart distributions for Income Type, Education Level, and Family Status collectively indicate that **financial and structural stability** are the strongest predictors of loan repayment.

- **Income Type (Default Rate):** The default-rate chart shows substantial variation across income types. “Pensioner” and “State servant” exhibit the lowest observed default rates (approximately 5–6%), while “Commercial associate” and “Working” display moderate levels (roughly 7–10%), representing the largest and most statistically reliable segments. “Unemployed” and “Maternity leave” appear to have extremely high default rates (around 36–40%); however, these categories have very small sample sizes (n=22 and n=5), so their percentages are likely unstable and should be interpreted cautiously. Similarly, the 0% default rates for “Student” and “Businessman” may reflect limited observations rather than truly zero risk.

- **Education Level:** The chart indicates a strong monotonic pattern: default rates increase as education level decreases. “Academic degree” exhibits the lowest default percentage (around 2%), followed by “Higher education” (around 5%). Default rates rise substantially for “Incomplete higher” and “Secondary / secondary special” (roughly 8–9%), and peak at “Lower secondary” (around 11%). This gradient suggests education level is a meaningful risk proxy in the dataset.

- **Family Status:** Default rates vary moderately by family status. “Widow” shows the lowest observed default rate (about 6%), while “Married” is higher (about 7–8%), and “Separated” increases further (about 8%). The highest default rates appear in “Single / not married” and “Civil marriage” (both close to 10%). Overall, differences across family-status groups are smaller than the education gradient, but the ordering suggests less stable household structures are associated with slightly higher default risk.

4.2 Numerical Variables



(a) Credit Amount Distribution

(b) Age Distribution

(c) Goods Price Distribution

Figure 3: Numerical Analysis

- **Age as a Risk-Associated Factor (YEARS_BIRTH):**

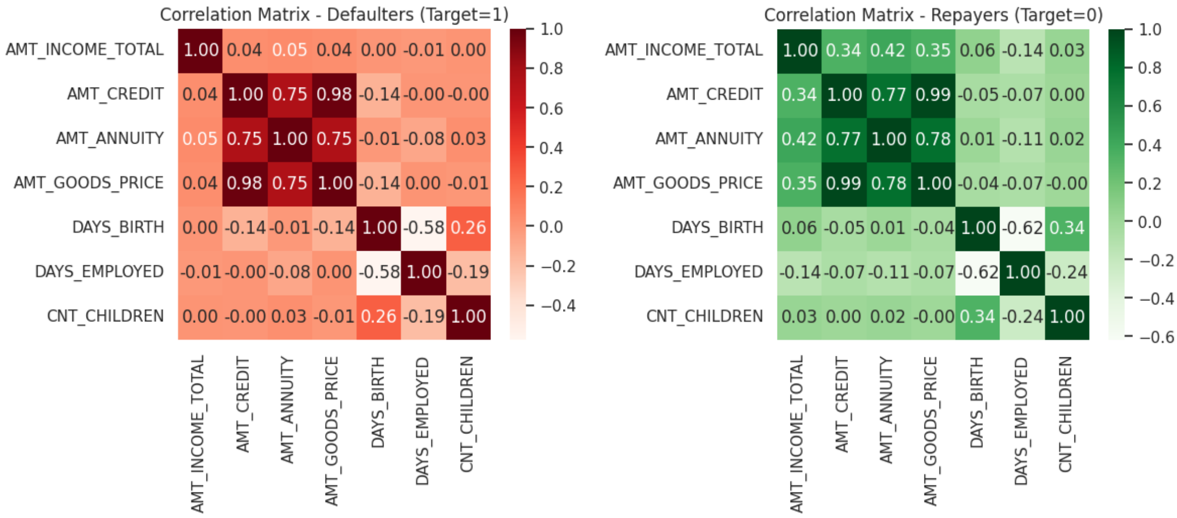
  The age density distributions differ noticeably between repayment groups. Defaulters (TARGET=1) are concentrated in younger age ranges, with a pronounced peak between roughly 25 and 35 years old, whereas repayers (TARGET=0) display a broader distribution extending more heavily into middle-age groups. This pattern indicates that younger applicants are overrepresented among defaulters, suggesting a strong association between age and repayment risk.

- **Loan and Goods Amount Distributions (AMT_CREDIT & AMT_GOODS_PRICE):** Both loan amounts and financed goods prices exhibit heavily right-skewed distributions, with most observations concentrated at lower values. Importantly, the overall shapes of these distributions appear similar across repayment groups, implying that absolute loan size alone does not visually differentiate defaulters from non-defaulters. This suggests that relative or interaction-based measures (e.g., loan-to-income relationships) may be more informative for risk assessment than raw loan amounts.

5 Bivariate Analysis

5.1 Pairwise Correlation Structure

Correlation matrices were computed separately for repayers (TARGET=0) and defaulters(TARGET=1) to compare how relationships among numeric variables differ across repayment outcomes.

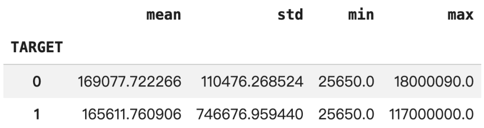


(a) Defaulters (TARGET=1)

(b) Repayers (TARGET=0)

Figure 4: Correlation Analysis

- The correlation structure differs across repayment groups. Among on-time payers (TARGET=0), AMT_INCOME_TOTAL is moderately correlated with both AMT_CREDIT (0.34) and AMT_ANNUITY (0.42), indicating income-aligned loan sizing. For defaulters (TARGET=1), these correlations are near zero (0.04 and 0.05), suggesting loan amounts were weakly associated with income within this subgroup.
- This contrast is unlikely to be driven by restricted income ranges, as both groups exhibit similar income means and bounds, but the defaulter group shows substantially higher variance. The pattern therefore points to greater heterogeneity or noise in the income distribution of defaulting applicants rather than a simple scaling effect.



	mean	std	min	max
TARGET				
0	169077.722266	110476.268524	25650.0	18000090.0
1	165611.760906	746676.959440	25650.0	117000000.0

Figure 5: Standard Deviation for AMT_INCOME_TOTAL

5.2 Numeric Variable Relationships

5.2.1 Employment Duration vs. Income

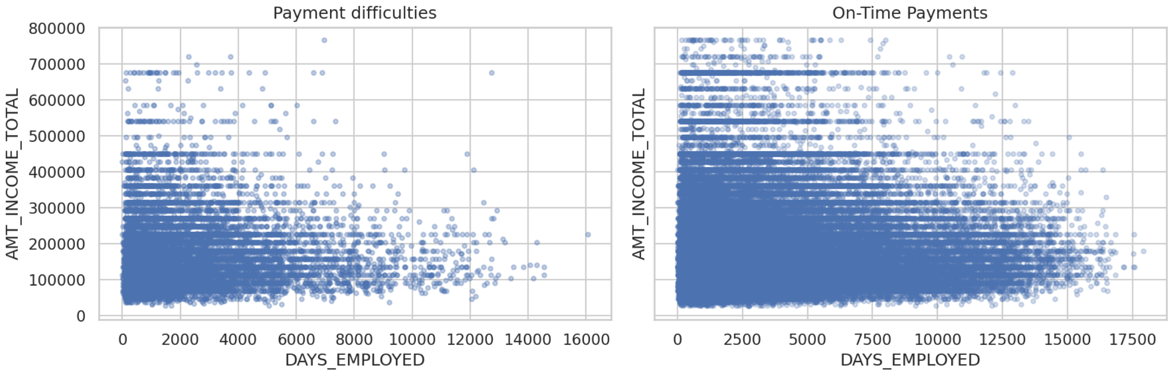


Figure 6: Employment duration vs. total income, split by repayment label.

Conclusion: Long employment histories are more prevalent among on-time payers, while income ranges overlap strongly between repayment groups; tenure appears to be a more informative stability proxy than income alone.

5.3 Categorical Variables vs Repayment Outcome

5.3.1 Age vs. Previous Application Outcomes

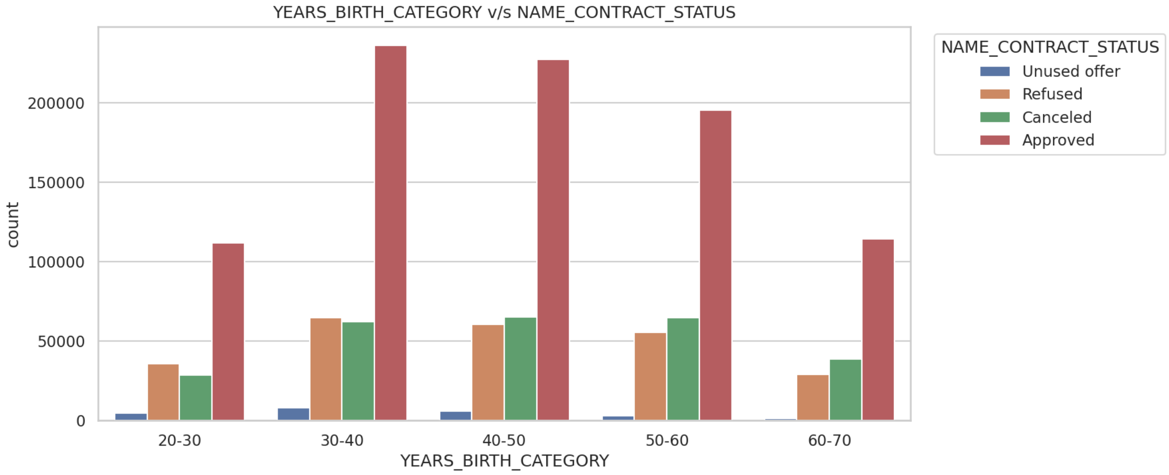


Figure 7: Previous application outcomes by age bucket.

Conclusion: Approval counts are highest for ages 30–50 and decline after age 50; approvals dominate outcomes in all age buckets.

5.3.2 Family Status vs. Previous Application Outcomes

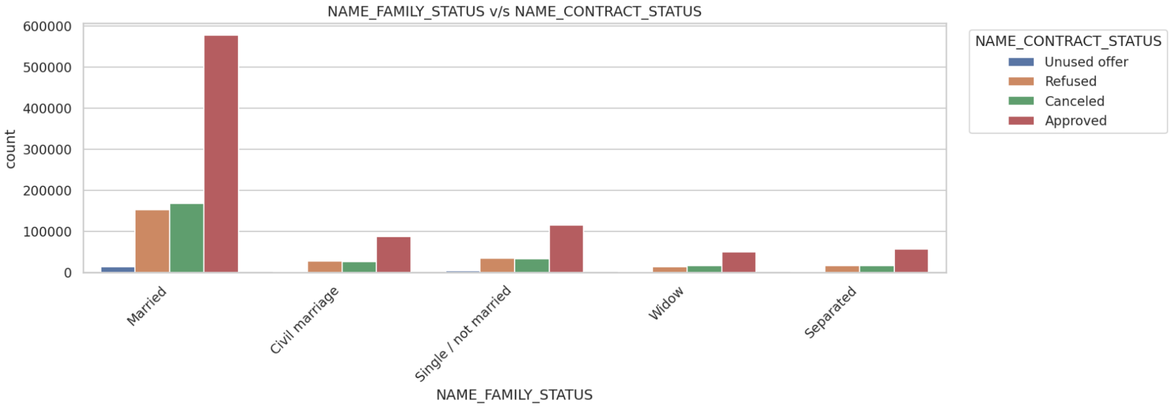


Figure 8: Previous application outcomes by family status.

Conclusion: Married clients account for the largest volume of approvals (count-based), but approvals remain the most frequent outcome across all family-status categories.

5.3.3 Client Type vs. Previous Application Outcomes

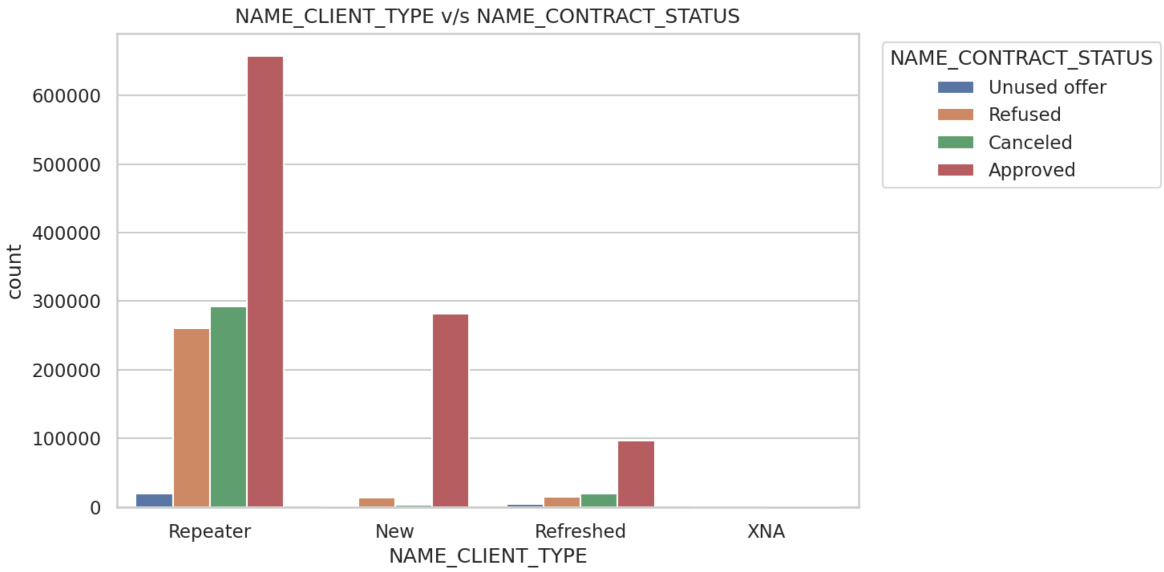


Figure 9: Previous application outcomes by client type.

Conclusion: Repeat clients dominate approval counts; however, this is count-based and should be complemented with approval-rate comparisons for inference.

5.4 Socioeconomic Factors vs Financial Magnitudes

5.4.1 Education vs. Credit Amount

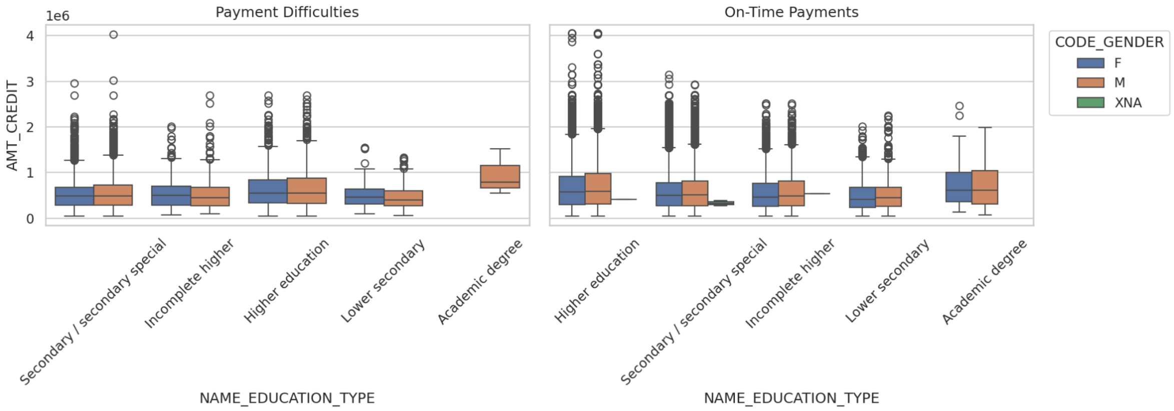
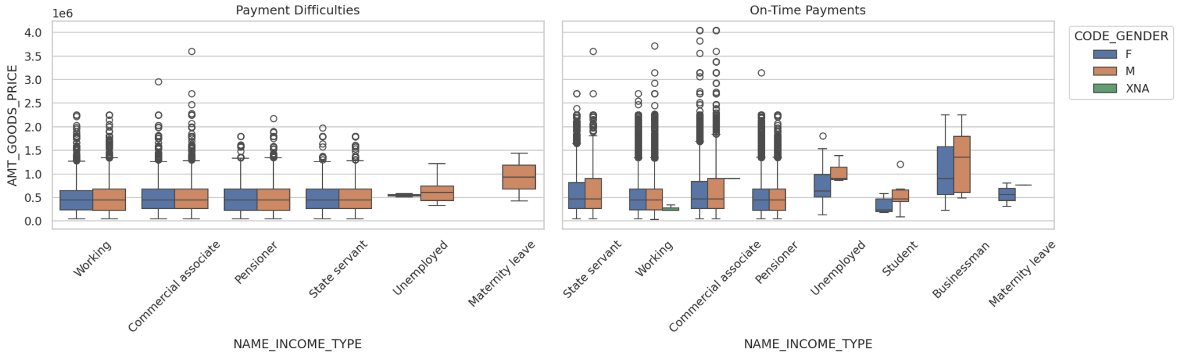


Figure 10: AMT_CREDIT by education type and gender, split by repayment segment.

Conclusion: Higher education levels are associated with larger typical credit amounts; gender differences are minor relative to education differences.

5.4.2 Income Type vs. Goods Price


Payment Difficulties

On-Time Payments

Figure 11: AMT_GOODS_PRICE by income type and gender, split by repayment segment.

Conclusion: Income source differentiates financed goods price more than gender; business/commercial categories show higher typical goods price, while student/unemployed categories are lower.

- **Gender Differences in Goods Price:** Across several employment categories, male applicants tend to finance slightly higher-priced goods than female applicants; however, these differences are modest and not uniform across all groups. In categories such as “Working” and “Pensioner,” gender differences are minimal, suggesting that gender alone is unlikely to be a strong determinant of financed goods value.
- **Variation Across Income Types:** Average financed goods prices differ more substantially by income type than by gender. Categories such as “Businessman” exhibit the highest values, while others remain clustered in lower ranges. Some economically vulnerable groups (e.g., “Unemployed”) also display relatively elevated averages; however, these observations should be interpreted cautiously without considering sample size and additional risk variables.
- **Data Quality Note:** A small number of observations fall under the “XNA” (unknown) gender category within the “Working” and “Commercial associate” groups. Because this category appears infrequently, it represents a minor data-quality issue.

Overall, income type appears to explain more variation in financed goods price than gender differences.

6 Business Implications

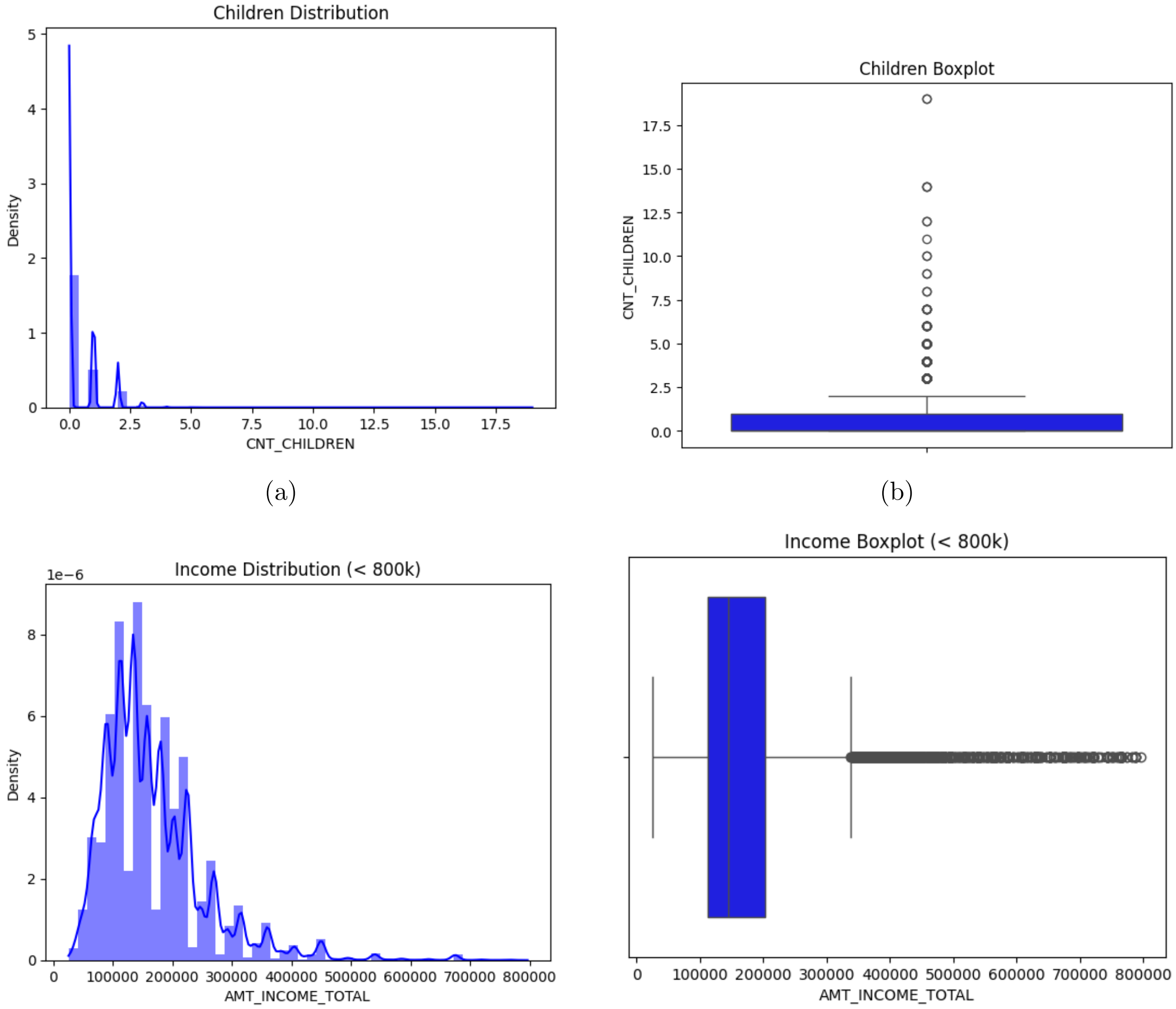
The exploratory analysis directly supports the lending objective of minimizing approval risk while avoiding unnecessary rejections.

- **Risk segmentation signals:** Stability-related attributes such as employment tenure, education level, and age provide stronger separation between repayment groups than raw financial magnitudes. These variables can improve early screening rules.
- **Underwriting alignment insight:** Loan amounts appear income-aligned among reliable borrowers but not among defaulters, suggesting that enforcing income-consistency checks may reduce default exposure.
- **Feature prioritization:** Relative metrics (e.g., loan-to-income, tenure–income interactions) are likely more predictive than absolute values and should be prioritized in scoring models.

- **Interpretation caution:** Segments with extreme percentages but very small sample sizes should not drive policy decisions, as they may reflect statistical noise rather than true risk patterns.
- **Operational implication:** Count-dominant groups (e.g., married or repeat clients) should be evaluated using approval or default rates instead of raw volume when defining lending strategies.

Overall, the results indicate that incorporating structural stability indicators and interaction-based measures can help lenders more effectively distinguish reliable applicants from high-risk ones, directly supporting the business objective of optimizing approval decisions.

A Additional Plots for Outlier Analysis



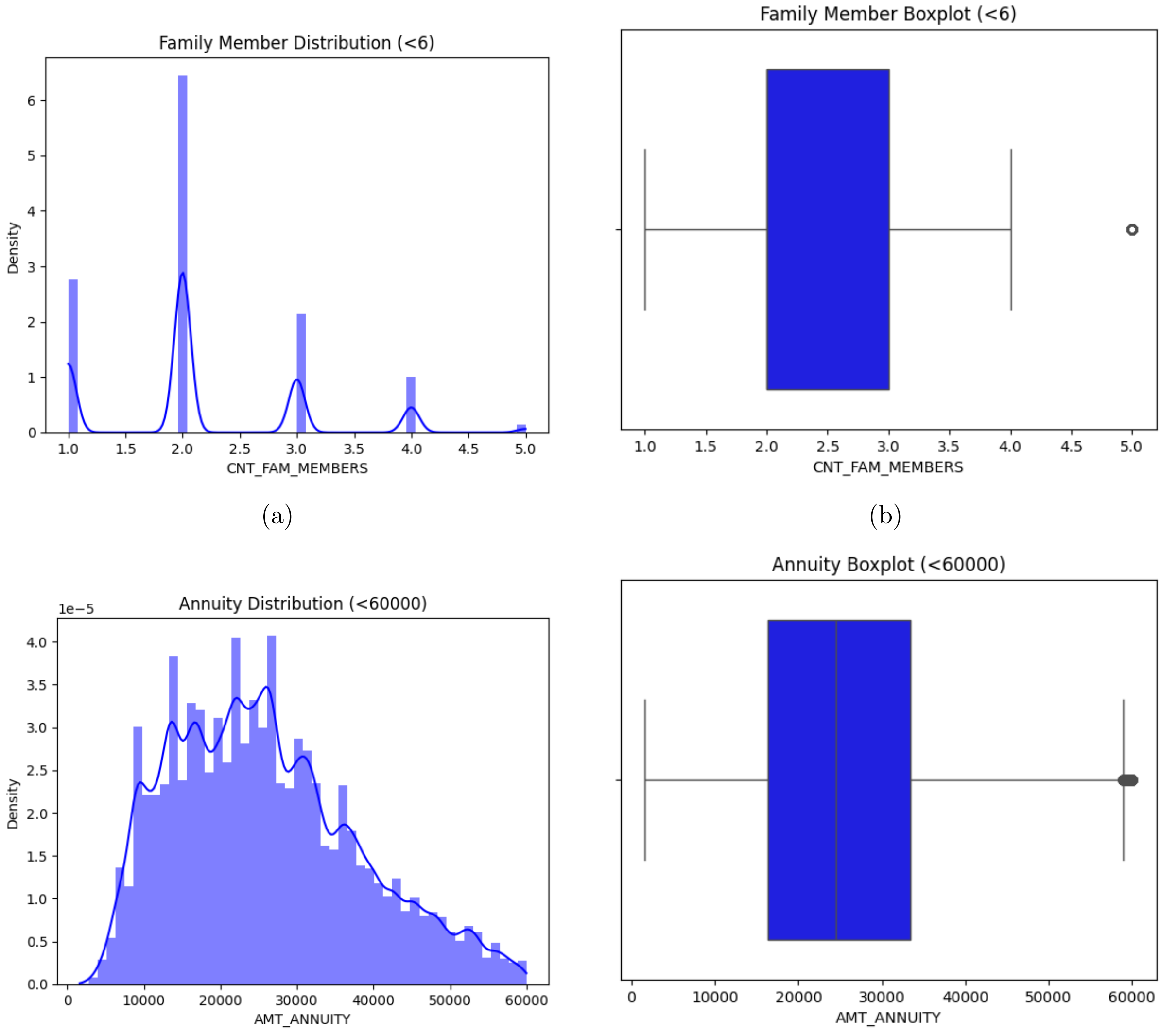
(a)

(b)

(c)

(d)

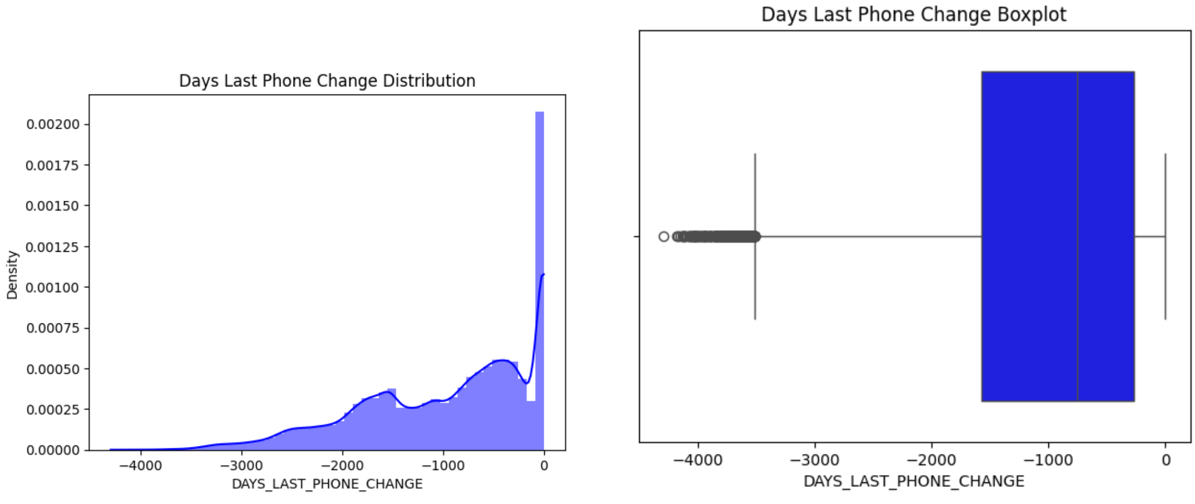
Figure 12: Outlier diagnostics (Part 1)



(a) (b)

(c) (d)

Figure 13: Outlier diagnostics (Part 2)



(a) (b)

Figure 14: Outlier diagnostics (Part 3)