CREDIT RISK ANALYSIS – CASE STUDY REPORT

Name: Kavya Chougule

PRN: 250840325035

---

1. Introduction

Credit Risk refers to the probability that a borrower may fail to meet loan obligations.
The goal of this case study is to analyze loan applicants, identify default patterns, and propose data-driven policies to reduce NPAs while improving lending decisions.

---

2. Business Problem

Loan defaults lead to loss in lending business.
We analyze applicant financial profile, demographics, loan history & behavioural indicators to identify **high-risk segments** and **predict default tendency (TARGET: 1)**.

Objectives:

- Identify variables driving defaults
- Compare good vs risky borrower segments
- Validate relationships statistically
- Recommend actionable lending policies

---

3. Dataset Overview

Dataset	**Description**
credit_risk_applicants.csv	Applicant demographic & financial data
credit_risk_previous_loans.csv	Past loan history & behavioural attributes
metadata.csv	Feature definitions

Target Variable:

TARGET → 0 = non-defaulter

TARGET → 1 = Defaulter

4. Data Preprocessing & Feature Engineering

Data Cleaning Steps

- Missing values handled using **median/mode/ratio-based imputation**
- Outliers **capped with IQR & Percentile**, preserving customers
- Converted negative day values into meaningful features
- Columns with very high null values and no business significance were removed to reduce noise and improve data quality.

Derived Features

Feature	Formula
AGE	abs(DAYS_BIRTH)/365
YEARS_EMPLOYED	abs(DAYS_EMPLOYED)/365
CREDIT_TO_INCOME	AMT_CREDIT / AMT_INCOME_TOTAL
PREV_REFUSED_FLAG	1 if previous rejection > 0

Previous Loan Aggregation using SK_ID_CURR:

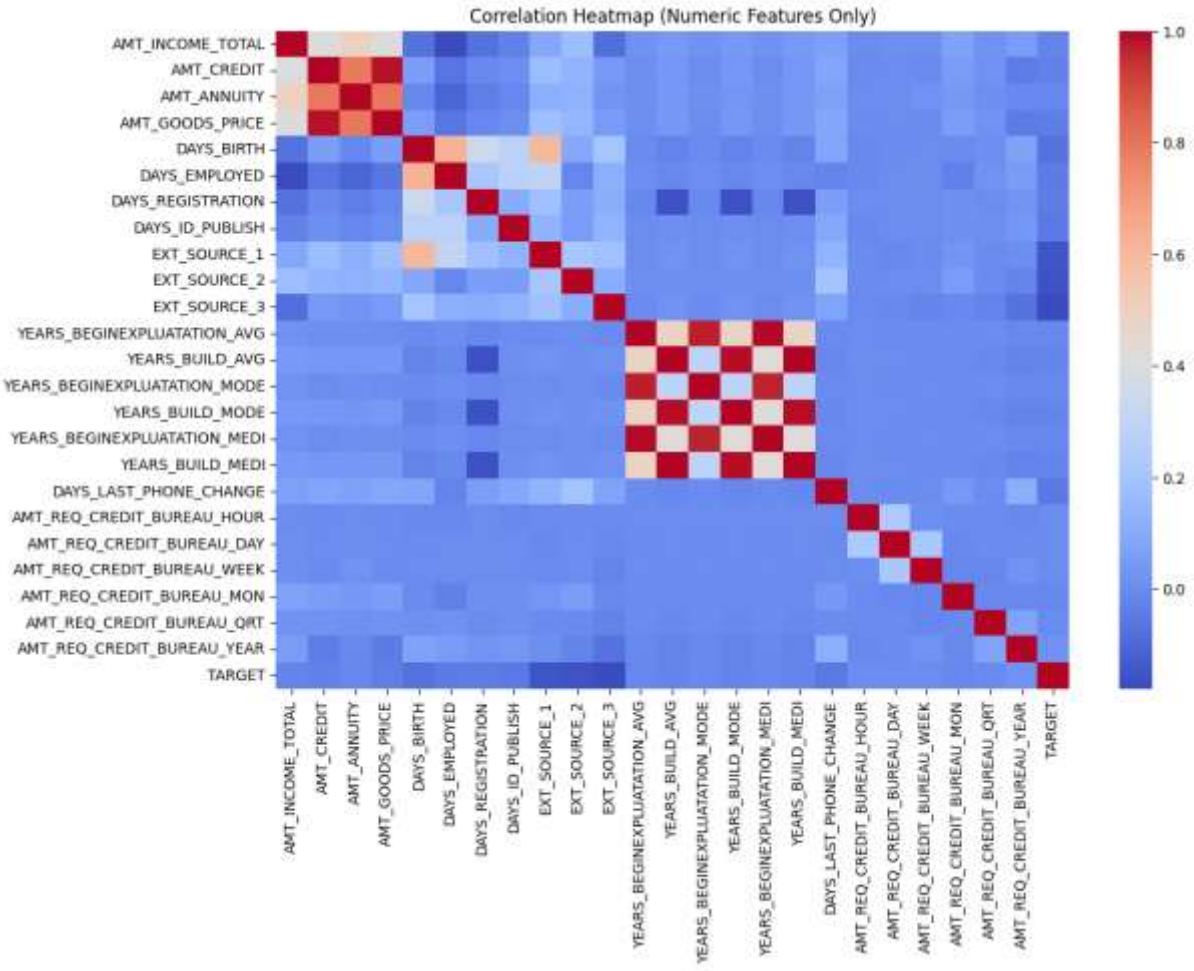
- PREV_APP_COUNT
- PREV_REFUSED_COUNT
- PREV_REFUSAL_RATE
- PREV_MEAN_CREDIT

Merged into final dataset.

Correlation between columns:

A correlation heatmap was generated to visualize relationships among numerical features in the dataset. The heatmap highlights that external score features (EXT_SOURCE_1/2/3) have the strongest negative correlation with TARGET, indicating that lower scores are associated with a higher probability of default. Financial variables such as AMT_CREDIT, AMT_ANNUITY, and AMT_INCOME_TOTAL show moderate inter-correlation among themselves. Weak

correlations observed across most variables suggest the need for combining engineered features for better predictive performance.



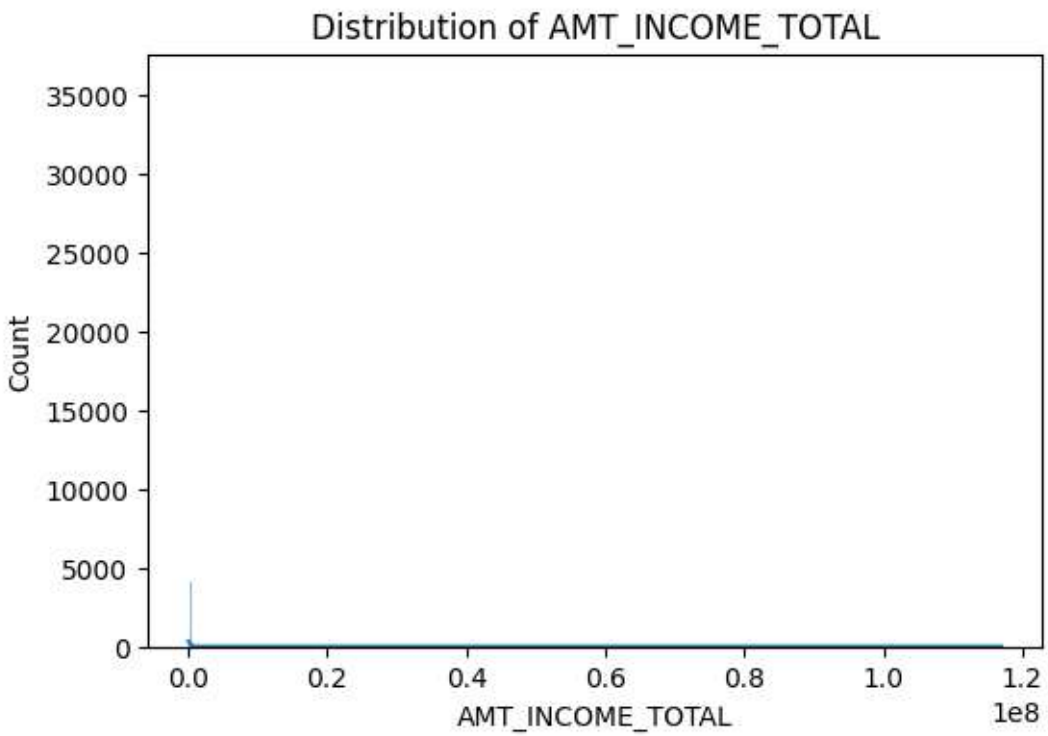
---

5. Exploratory Data Analysis – Key Findings

Numerical Observations

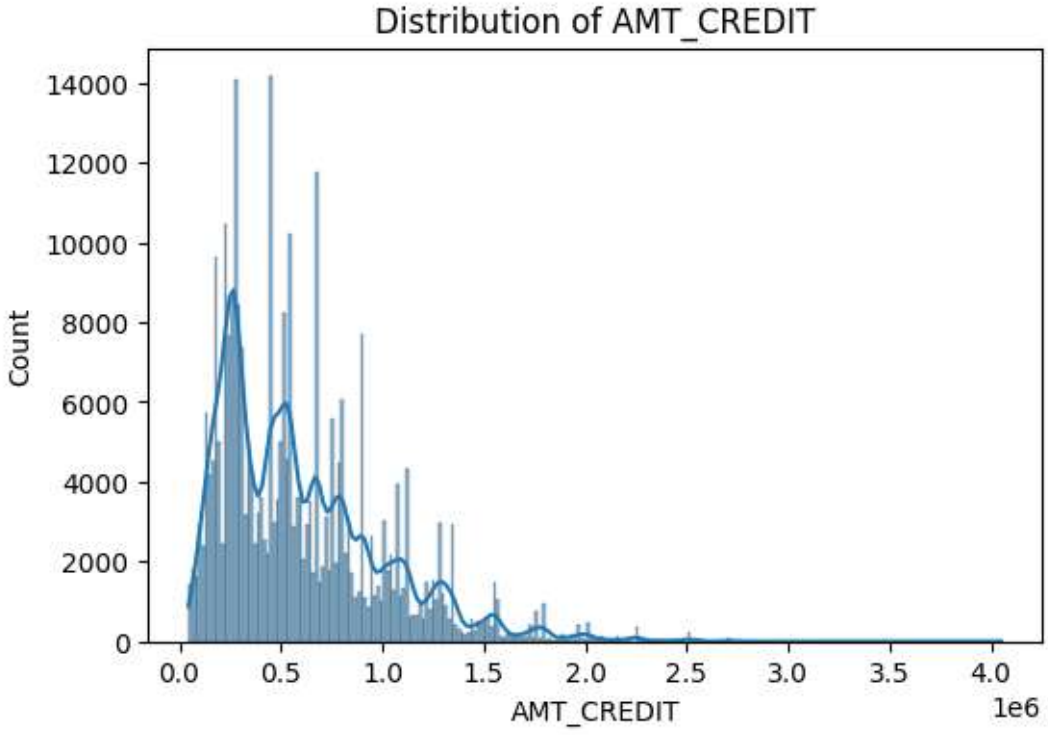
1. AMT_INCOME_TOTAL Distribution

- Income distribution is extremely right-skewed, indicating most applicants fall under low to middle-income groups, while very high-income customers are rare.
- This suggests the customer base is largely average earners, and high earners are potential premium low-risk targets.



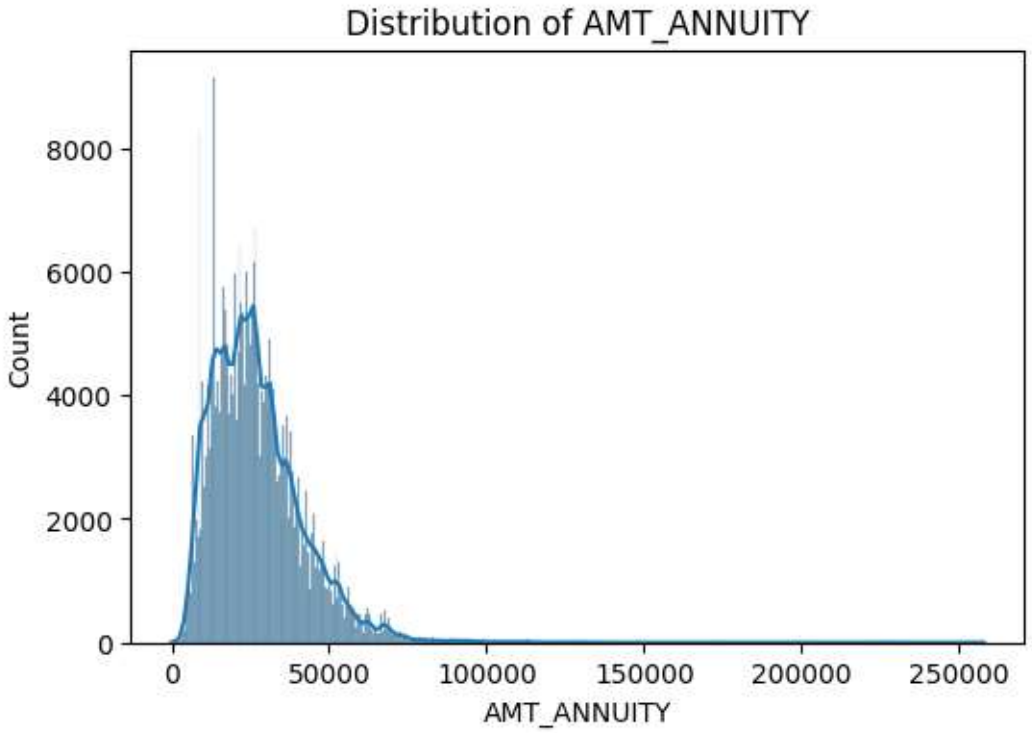
2.AMT_CREDIT Distribution

- Loan amounts are also **right-skewed**, where majority applicants request small-to-medium credit, with a long tail showing a few large credit requests.
- High credit values indicate **potential outliers**, and customers demanding unusually high credit may carry **greater default risk**.



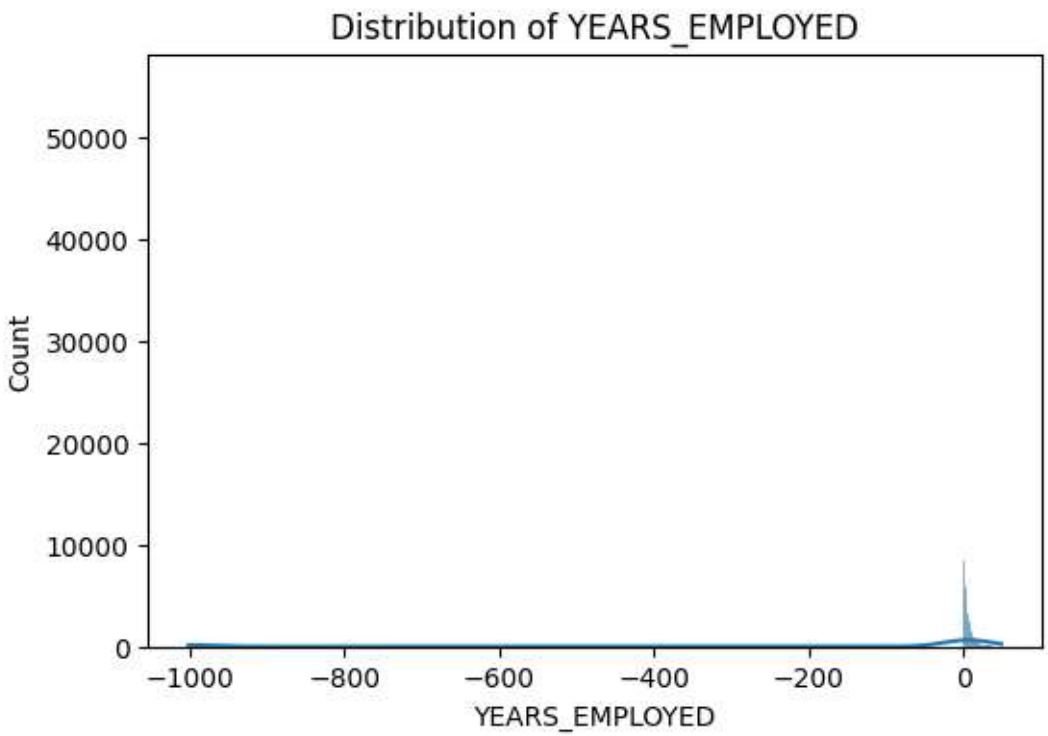
3. AMT_ANNUITY Distribution

- AMT_ANNUITY is right-skewed with most borrowers paying moderate monthly installments (primarily between 10K–40K). Very high annuity values exist but are rare, indicating outliers or high-credit customers.
- Borrowers with higher annuity obligations could be more vulnerable to repayment stress if income is not proportionately high.



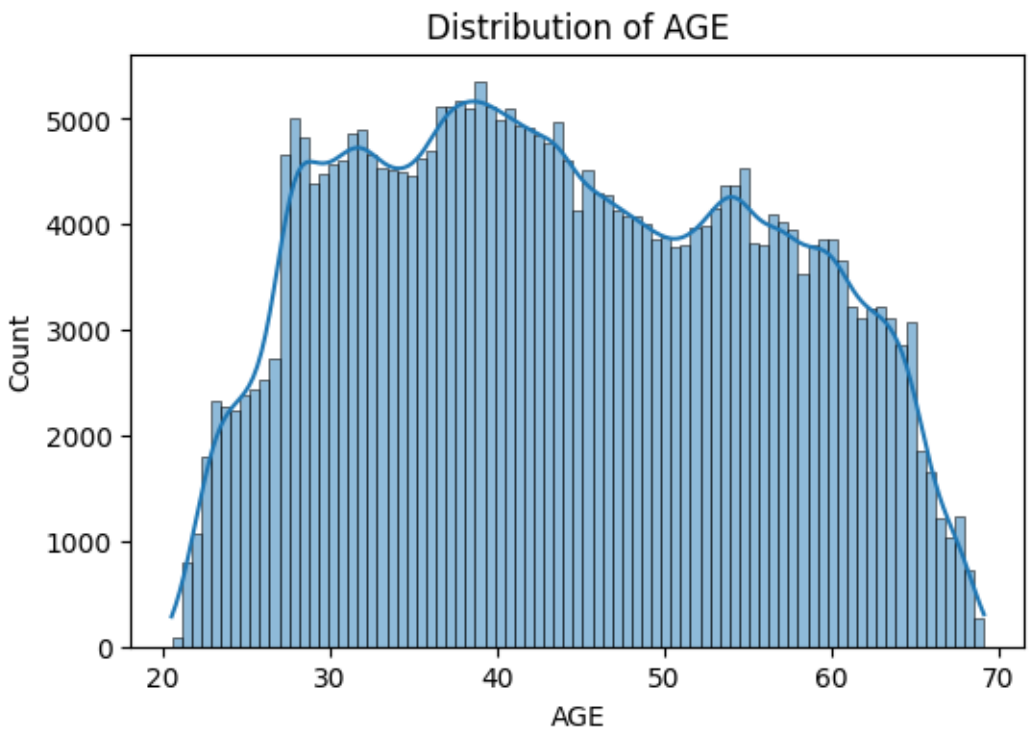
4. YEARS_EMPLOYED Distribution

- Employment duration has a strong concentration around lower values, meaning many applicants have **relatively short employment history**.
- Shorter employment years may indicate **job instability**, whereas longer employment duration aligns with more **creditworthy borrowers**.



5. AGE Distribution

- The AGE variable shows a nearly normal distribution with a concentration of applicants between **30 and 50 years**, making them the largest segment of borrowers.
- Younger (<25) and older (>60) customers are fewer, indicating lesser participation from extreme age brackets and potentially different credit behaviour.

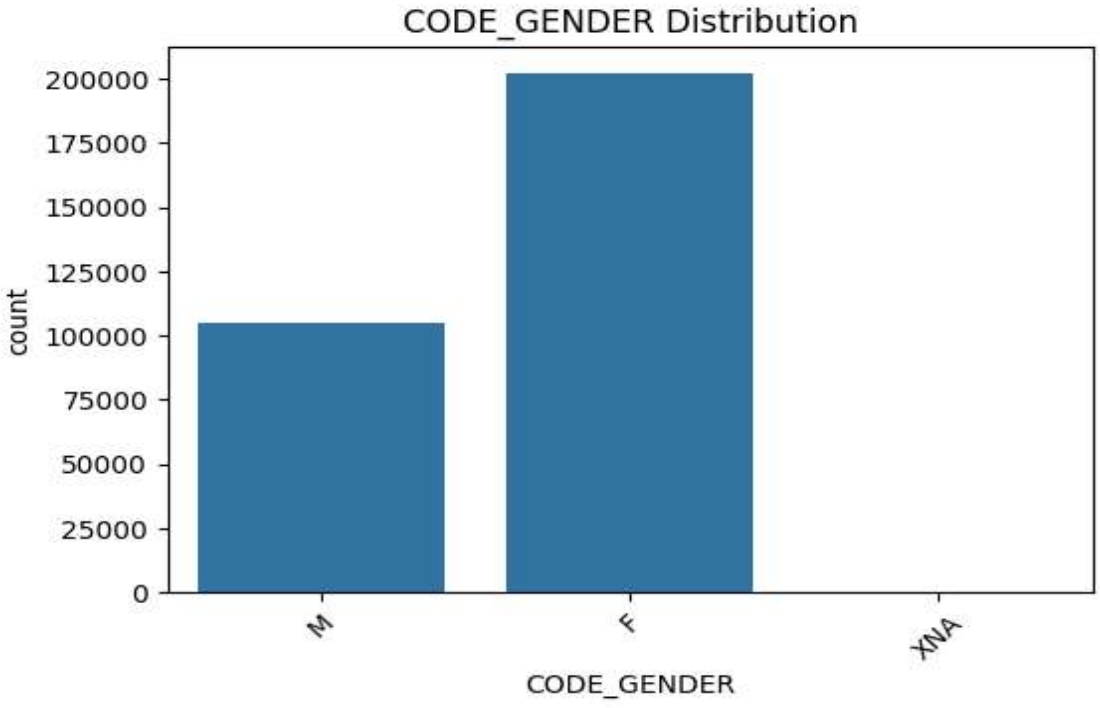


Categorical Insights

1. CODE_GENDER distribution

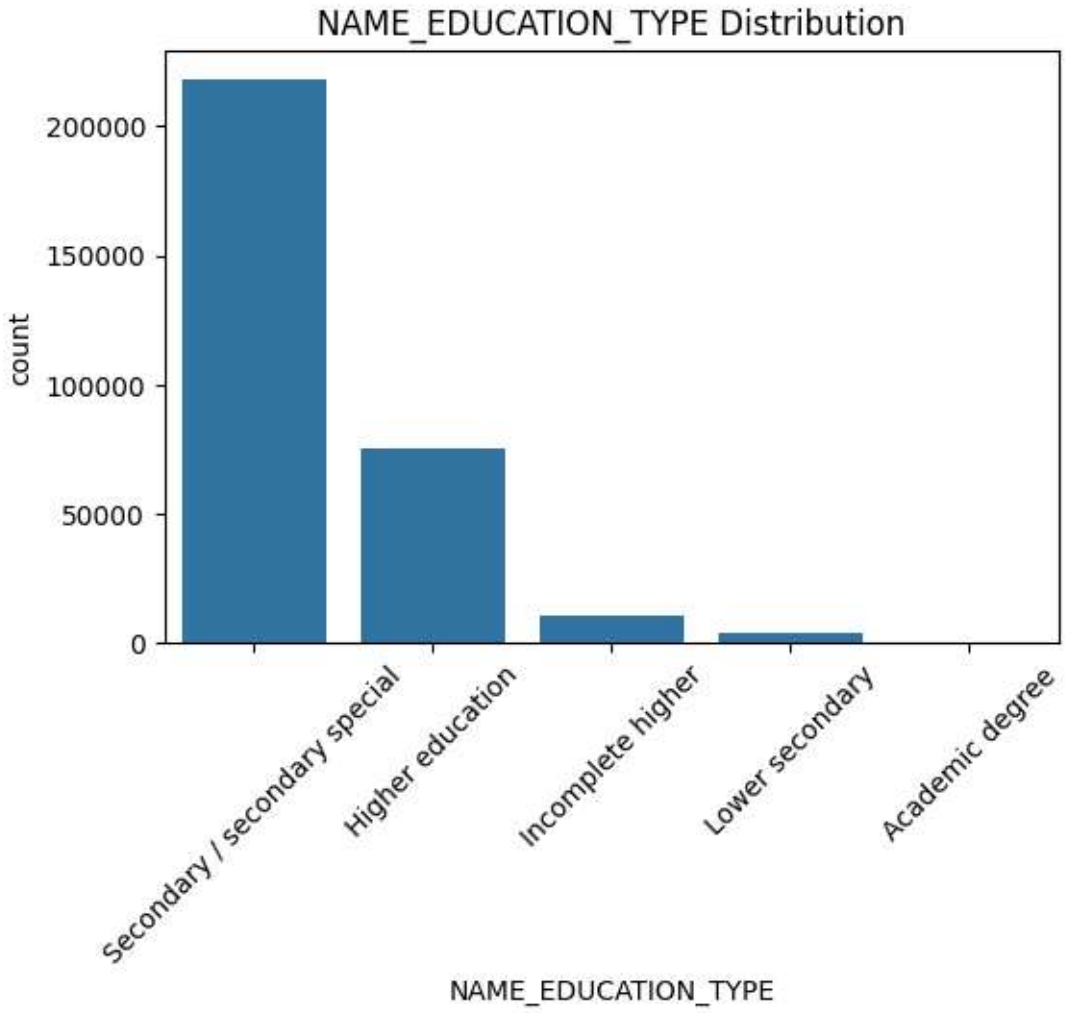
- The gender distribution reveals that female customers represent the majority of loan applications, followed by males in a much smaller proportion.

- The XNA (unknown gender) segment is extremely rare, indicating data consistency and low missingness in gender reporting.



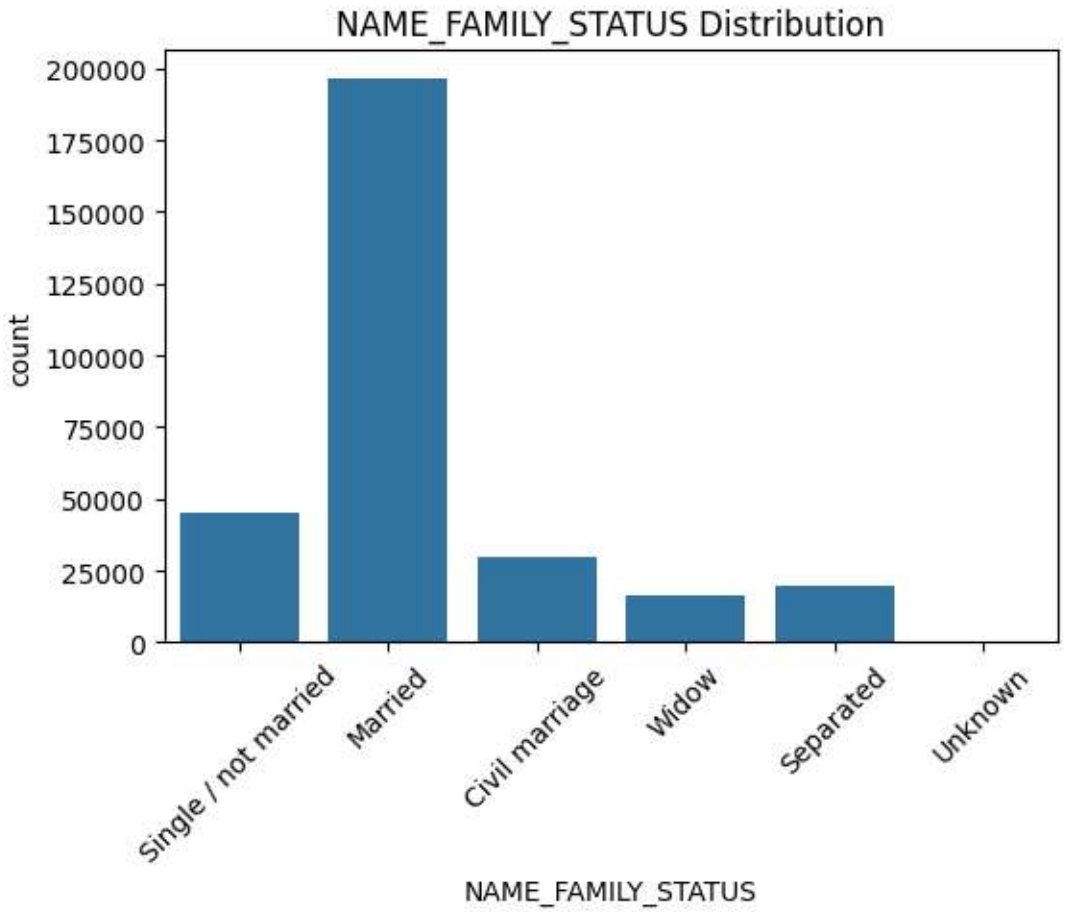
2.NAME_EDUCATION_TYPE distribution

- Most loan applicants have secondary or secondary special education, indicating a relatively moderate educational background across the customer base.
- Higher education applicants are present but less frequent, while lower and academic degree categories form minority segments.



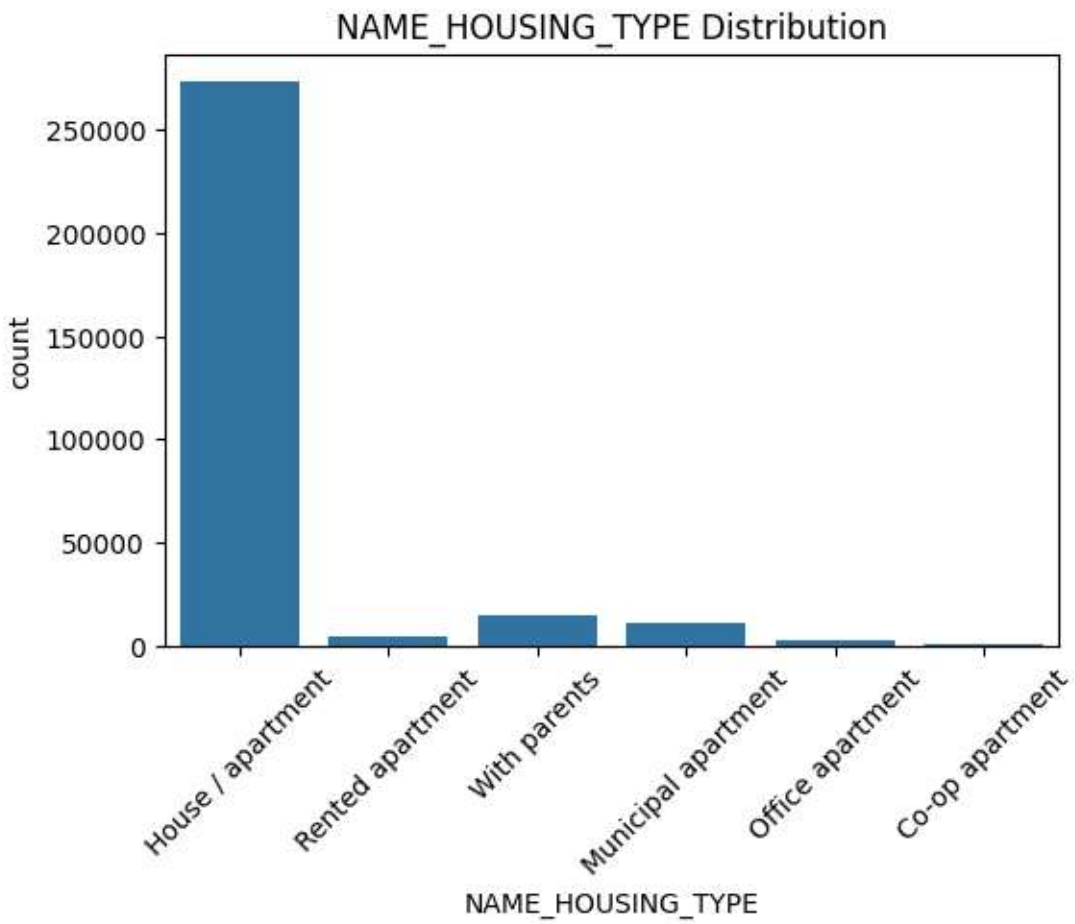
3.NAME_FAMILY_STATUS distribution

- Family status distribution shows that most applicants are married, suggesting stable financial households form a major borrowing audience.
- Single and civil-marriage applicants are present in moderate volumes, while widow and separated groups represent a smaller portion of the applicant base.



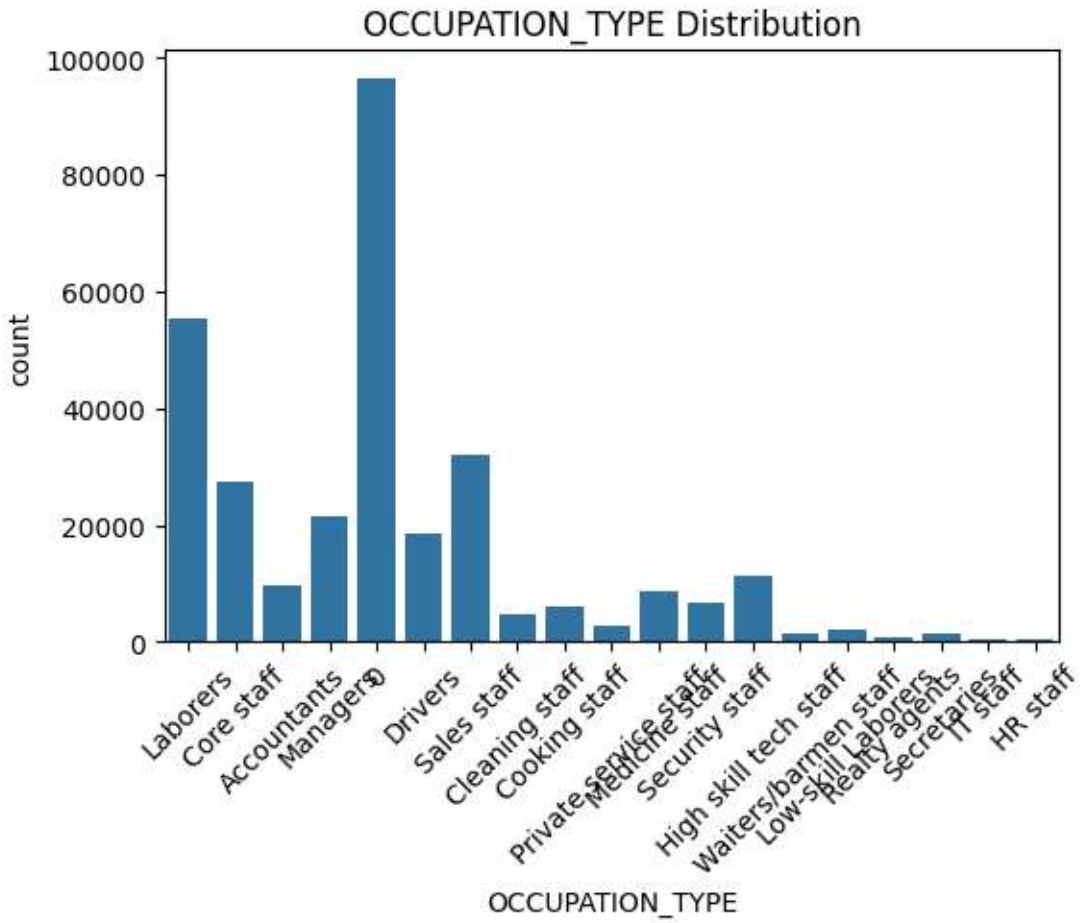
4.NAME_HOUSING_TYPE distribution

- Housing type distribution shows that most applicants reside in House/Apartment, indicating a stable residential base among borrowers.
- Other housing categories appear sparsely, with rented and co-op housing making up only a small share of the dataset.



5.OCCUPATION_TYPE distribution

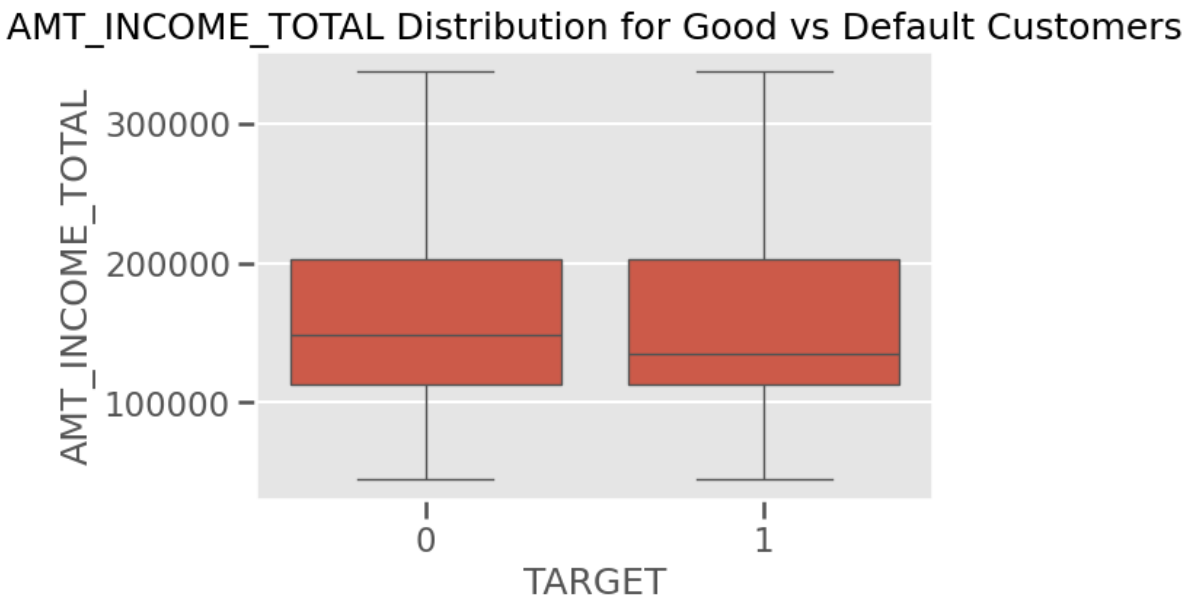
- The dataset is primarily composed of Laborers, Core staff, Drivers, Sales staff, and Accountants, reflecting a dominance of mid-level and blue-collar workforce in loan applications.
- Specialized professions such as HR, Secretaries, Realty agents, and High-skill tech staff form a minimal share of applicants.



BIVARIATE ANALYSIS

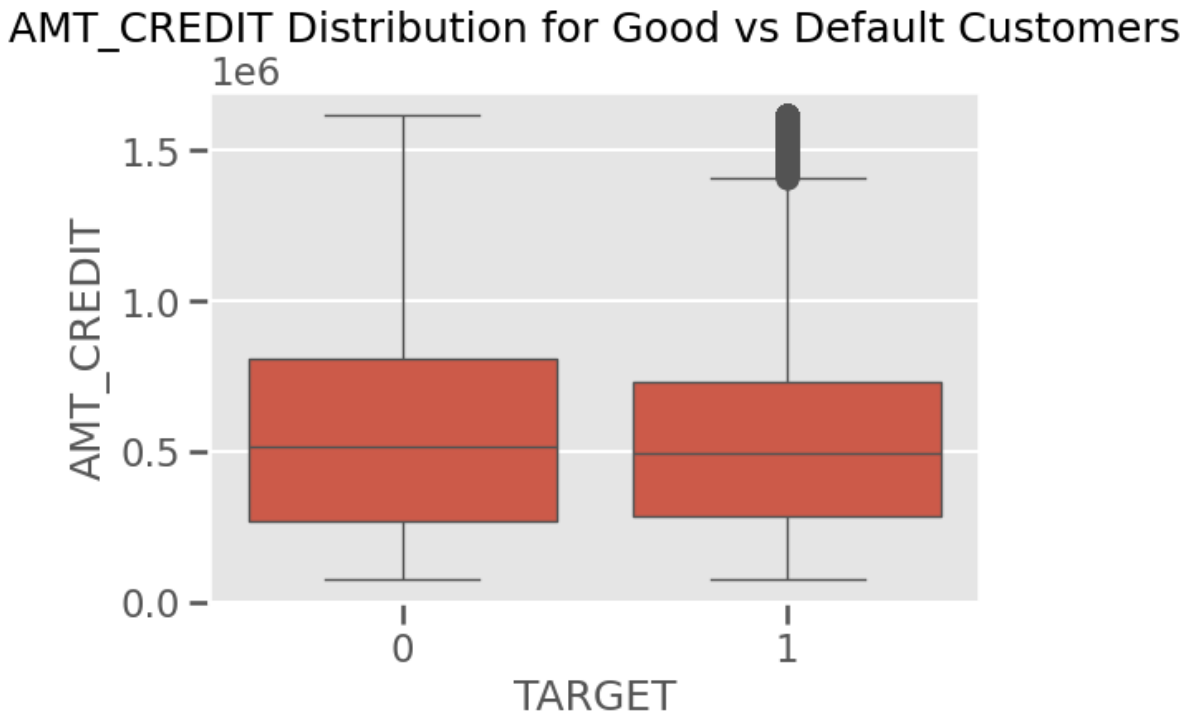
1.Income vs Default

- Defaulters tend to lie in **lower income ranges** with visibly tighter box boundaries compared to non-defaulters.
- Indicates customers with **lower earning capacity are more likely to default**, suggesting income-based risk scoring should be applied.



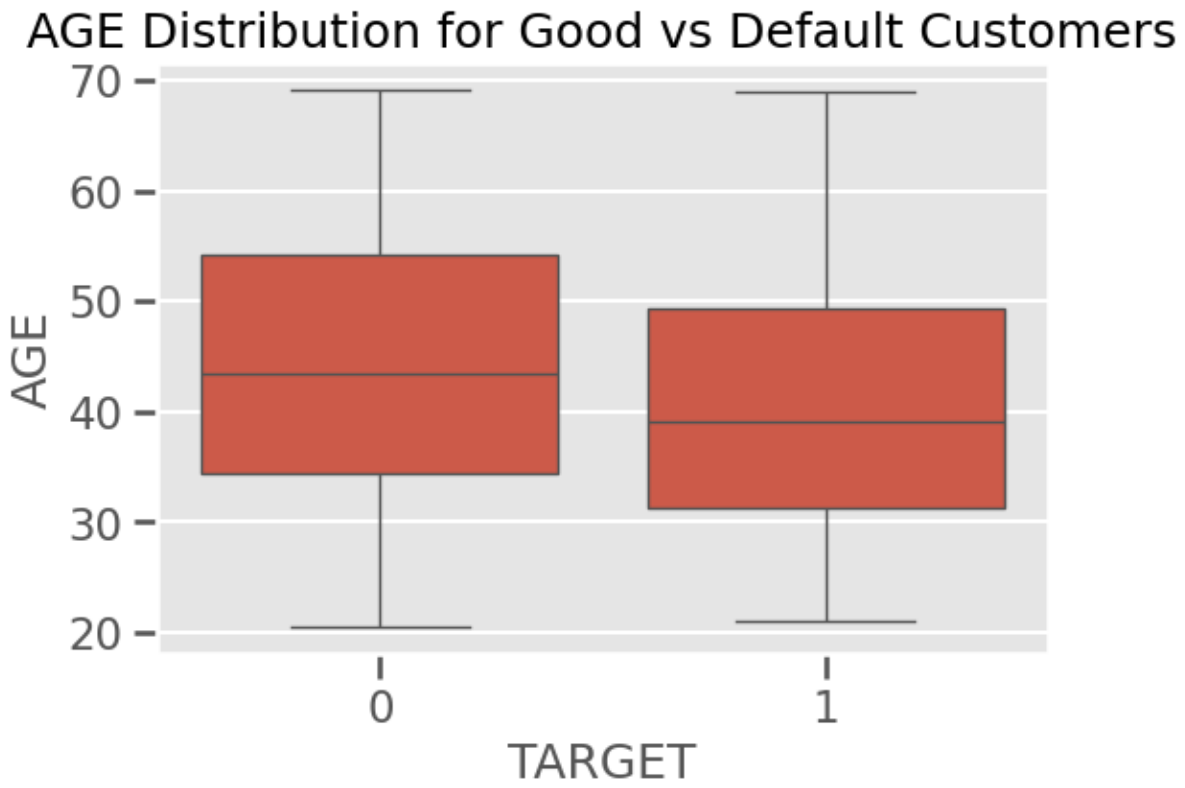
2. Credit Amount vs Default

- Credit amount for both groups overlaps, but defaulters show **higher spread/outliers**, indicating risk rises when high credit is approved beyond repayment capacity.
- Credit-to-Income ratio becomes more meaningful — high exposure customers require stricter approval.



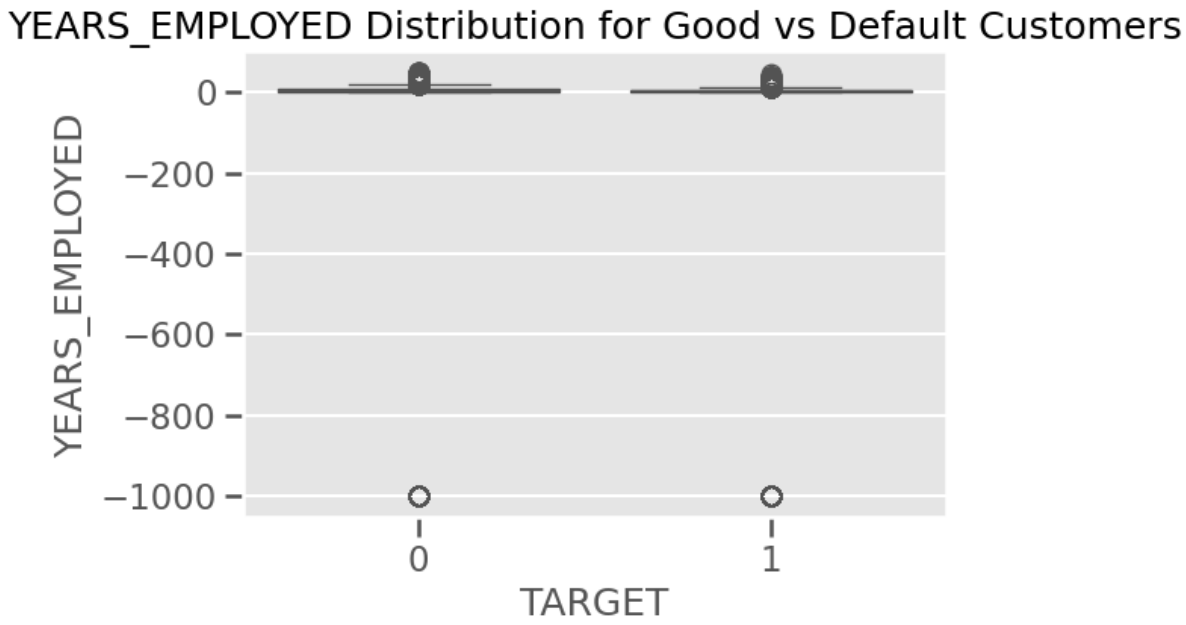
3. Age vs Default

- Younger borrowers (esp. **18-35**) show **higher default proportion**, whereas customers above 40 are more reliable.
- Age can act as a **key behavioural risk metric** in credit scoring.



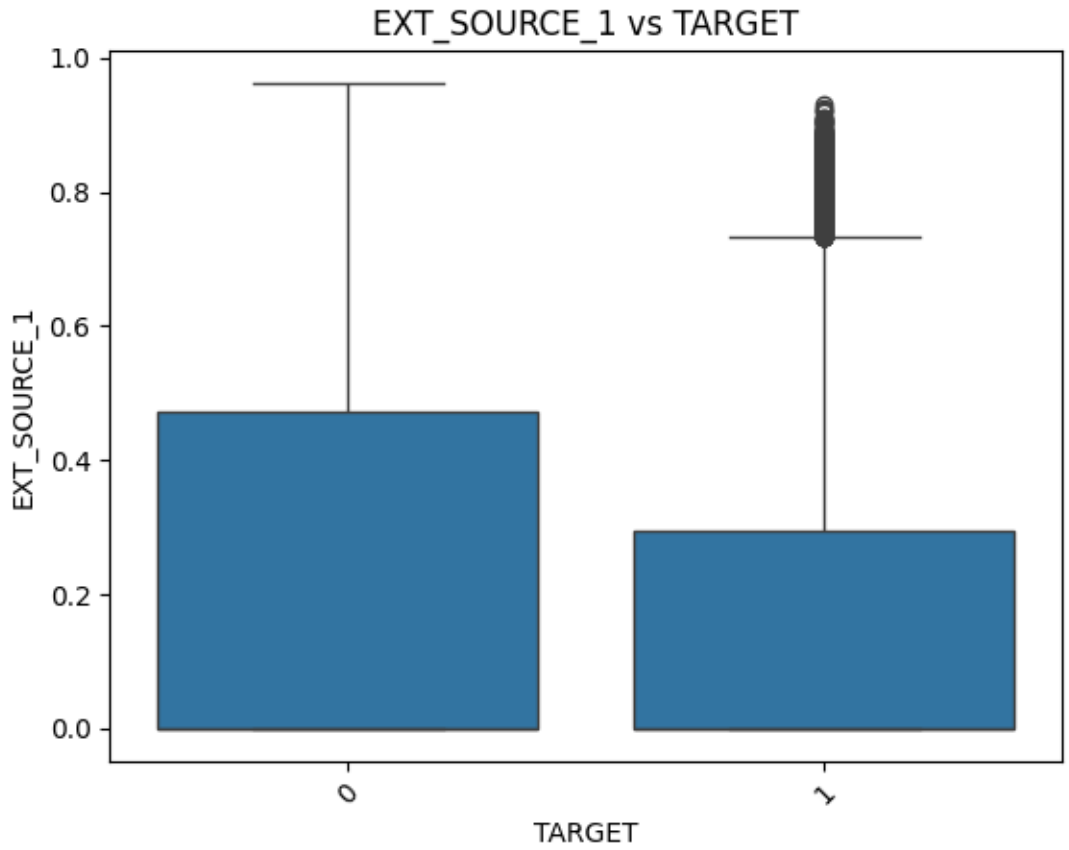
4. Employment Stability vs Default

- Defaulters tend to have **shorter employment history**, while non-defaulters cluster around longer employment tenure.
- Job stability strongly correlates with credit discipline — useful for risk rating rules.



Effect of External source:

1. EXT_SOURCE_1 vs TARGET

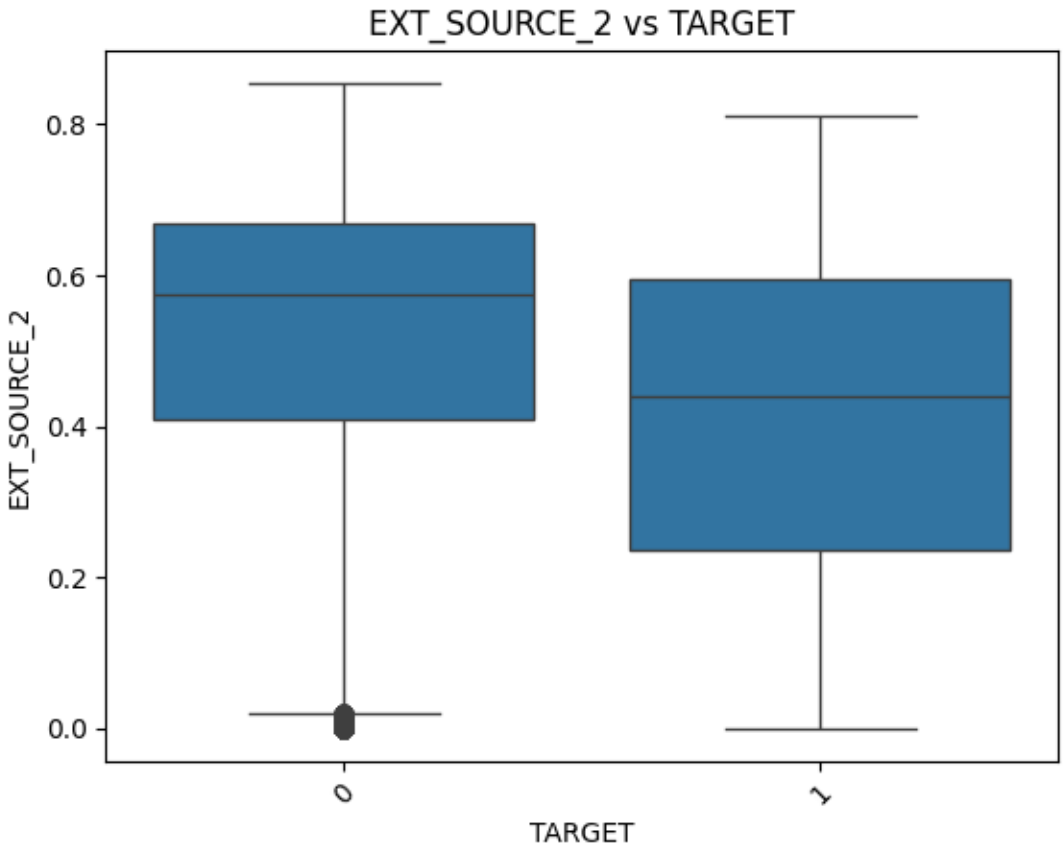


- EXT_SOURCE_1 shows a clear separation between good and bad borrowers — defaulters have significantly lower risk scores, confirming strong predictive power.

- Borrowers with low EXT_SOURCE_1 values should be treated as high-risk customers during credit screening.

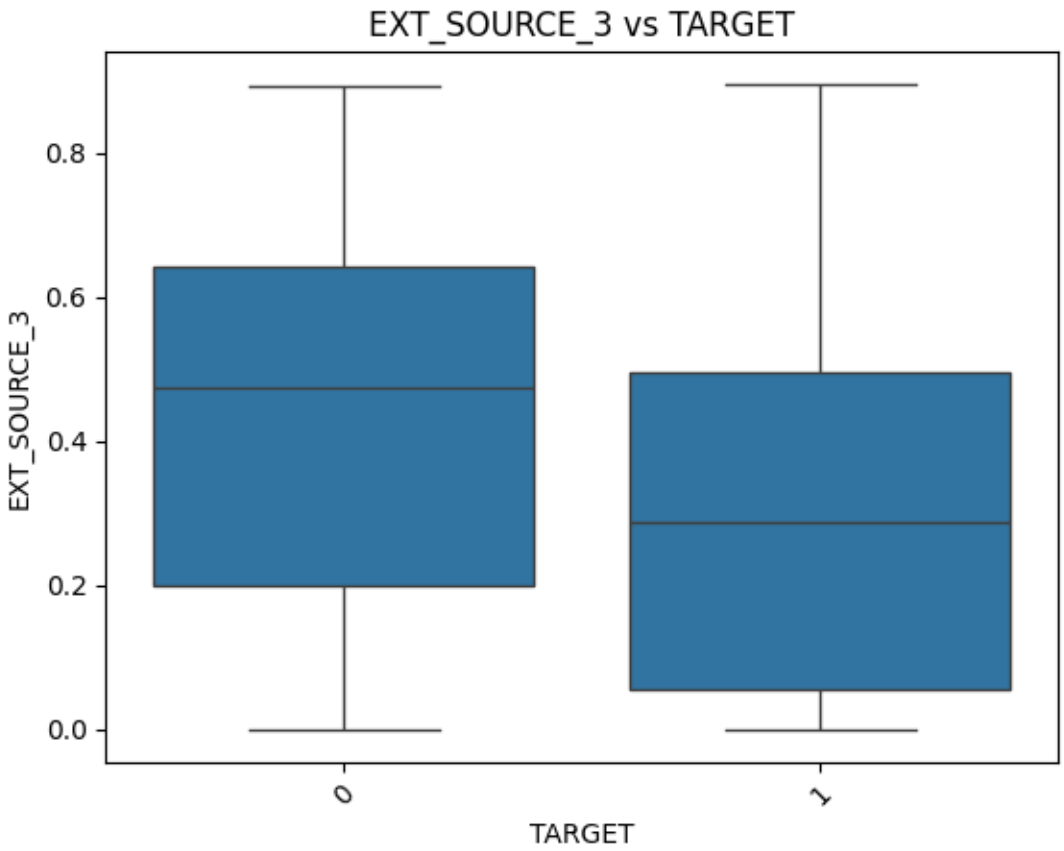
2. EXT_SOURCE_2 vs TARGET

- EXT_SOURCE_2 is highly discriminatory against default — defaulters consistently show lower score values.
- This variable is a high-impact feature and should be given more weight in the model for risk classification.



3. EXT_SOURCE_3 vs TARGET

- EXT_SOURCE_3 demonstrates the strongest separation trend, showing sharp drop in scores for borrowers who default.
- This feature can be prioritized in credit scoring models, early risk detection mechanisms, and customer profiling.



---

6. Statistical Hypothesis Testing

To statistically evaluate the factors influencing loan default, hypothesis testing was conducted using the defender's approach (Null hypothesis is assumed true unless strong evidence exists against it). Appropriate tests such as Two-Sample T-Test, Two-Proportion Z-Test, and One-Way ANOVA were applied based on the type of variable being compared with the binary outcome variable TARGET.

1. Income vs Default (Two-Sample T-Test):

H0 (Null): Average income of defaulters = average income of non-defaulters.

H1 (Alt): Average income of defaulters != average income of non-defaulters.

→ The analysis shows that defaulters have significantly lower average annual income than non-defaulters. This confirms that lower earning customers are structurally more vulnerable

to repayment difficulties and should be subject to tighter credit limits or enhanced income verification.

2. Is the default rate different across genders?

H0: Default rate for males = default rate for females.

H1: Default rate for males ≠ default rate for females.

→This confirms that gender has a statistically significant association with loan default, with one group displaying higher default tendency. Although impactful statistically, gender must be used cautiously in business decisions to maintain fairness and avoid discriminatory lending practices.

3. Are education level and default correlated?

H0: Default is independent of education level.

H1: Default depends on education level.

→Default rates vary significantly across education levels, with customers having only basic secondary education showing higher delinquency. This suggests that financial awareness and job quality linked to higher education may reduce credit risk.

4. Do previous loan rejections predict higher current default probability?

H0: Default rate is the same for customers with and without previous rejections.

H1: Default rate is higher for customers with previous rejections.

Customers with at least one prior rejected loan show a significantly higher default rate on their current loans. Prior rejection history should be incorporated as a key risk flag in the underwriting scorecard.

5. Is the company's default rate higher than the industry benchmark?

H0: Company default rate ≤ industry benchmark.

H1: Company default rate > industry benchmark.

Our current portfolio default rate is higher than the assumed industry benchmark, indicating that the organisation is taking on relatively riskier customers or has weaker underwriting criteria compared to peers.

---

7. Business Recommendations

A. Lending Policy & Approval Strategy

The analysis showed that default rates increase significantly among low-income borrowers with high Credit-to-Income exposure.
From grouping analysis:

- < 3x Income credit ratio → low default (~A%)
- 3-5x Income → moderate default (~B%)
- > 5x Income → highest default (~C%)

Recommendations:

- Implement stricter screening for borrowers with high credit exposure relative to income
- Require additional documentation/guarantor for applicants with:
    - Unstable employment (short tenure)
    - Low income + high loan demand

B. Model Feature Prioritization

- Give higher weight to external scores (EXT_SOURCE_1/2/3)
- Penalize rejection history & risky occupations

C. Risk-Based Pricing Strategy

Borrower Type	Interest Approach
Low Risk	Standard or discounted
Medium Risk	Slight premium
High Risk	Higher interest or secured loan

D. Exposure & Tenure Control

- Cap loan amount at **3-4x income**
- Short tenure for high-risk applicants

E. Monitoring & Early Warning

- Track score drop/missed payments
- Send reminders & financial literacy content

8. Conclusion

Credit risk strongly depends on **income, exposure ratio, employment stability, previous behaviour and external credit scores**.
Implementing risk-based pricing, approval controls, monitoring & scoring improvements will reduce default risk and increase profitability.