Home Loan Dataset Report

By

HASSAN TAIWO, MAJARO

Table of Contents

1. Executive Summary

This EDA on homeload dataset examined demographic, financial, and credit attributes of loan applicants to understand the patterns influencing loan approval.
The dataset contains applicant profiles (Gender, Marital Status, Dependents, Education, Employment), financial metrics (ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term), their credit behavior (Credit_History), and the outcome variable which is (Loan_Status).

Key Observations:

- **Applicant and Coapplicant Income Distributions:** These are highly skewed right, most earnings are under 10,000NGN while a few high-income earners stretch the scale.
- **LoanAmount:** This also shows right skewness and clear outliers, confirming the wide variation in requested loan sizes.
- **Loan_Amount_Term:** This is concentrated at 360months, implying that most applicants chose standard 30-year terms.
- **Credit_History:** This shows overwhelmingly positive (i.e most have credit history), showing that many applicants are not first-time borrowers.
- **Categorical analysis:** This shows most applicants are male, married, graduates, and live in semi-urban or urban areas.
- **Loan_Status:** This indicates around 70-75% approvals, suggesting the institution has a relatively generous approval policy.

Missing values were imputed (using mode for categorical, median for numeric) and outliers in *LoanAmount* were capped using IQR method.
This means that the dataset is now consistent, interpretable, and it's suitable for predictive modelling.

2. Data Assessment

2.1 Structure

- Train dataset: 13 columns
- Test dataset: 12 columns
- Both have a mix of categorical and numeric features

2.2 Missing Values

Missing values are mainly in:

- Gender, Dependents, Self_Employed, and Credit_History (categorical)
- LoanAmount and Loan_Amount_Term (numeric)

These missing values were handled with logical imputations

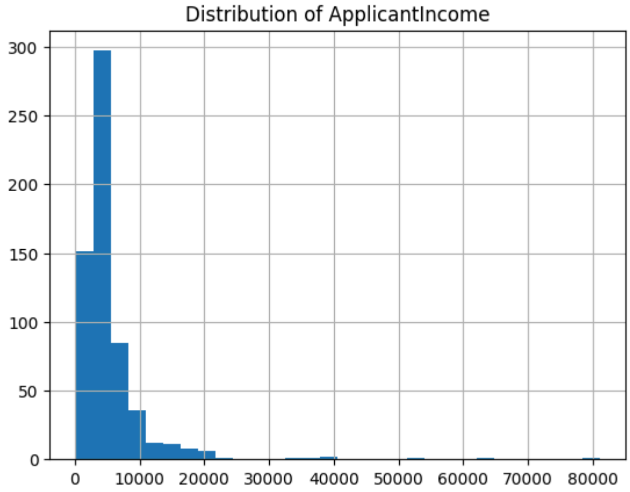
- Mode (most frequent) for categorical
- Median for numeric, since the distributions were skewed.

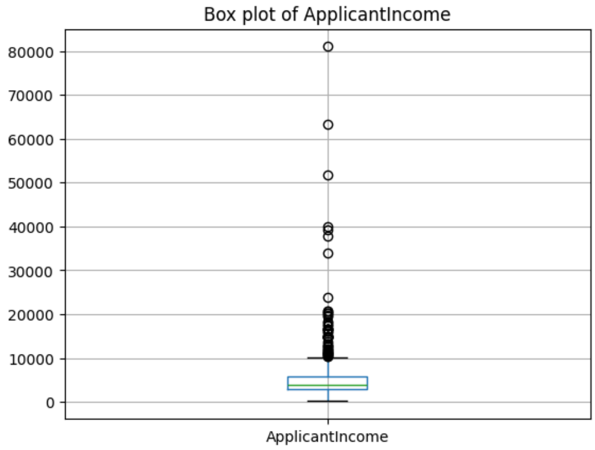
There are no missing values in key columns after cleaning.

3. Descriptive Analysis

3.1 ApplicantIncome

- Histogram and Boxplot show strong right skew i.e most applicants earn below 10,000NGN while a few exceed 50,000NGN - 80,000NGN
- Many outliers are visible above 20,000NGN, but those are not errors. They just represent a few high-income earners.
- This confirms that median income (not mean) represents the central tendency better.

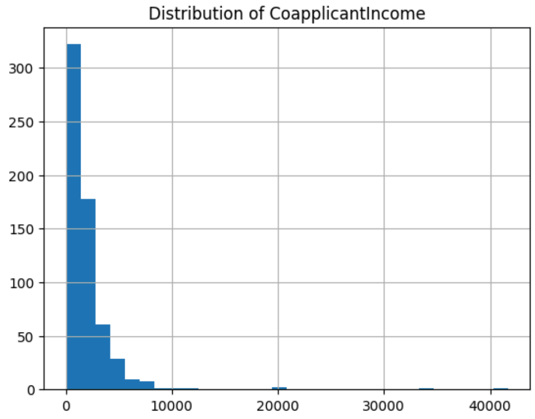


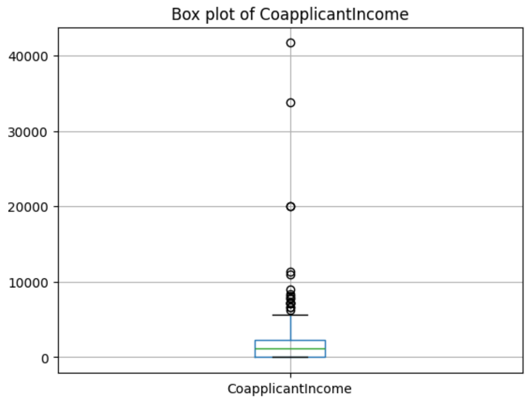


3.2 Coapplicant Income

- Similar pattern to ApplicantIncome
- Most coapplicants earn below 5,000NGN, and several records show 0NGN, indicating single applicants or non-earning coappliants.
- Outliers above 10,000NGN exist but are rare.

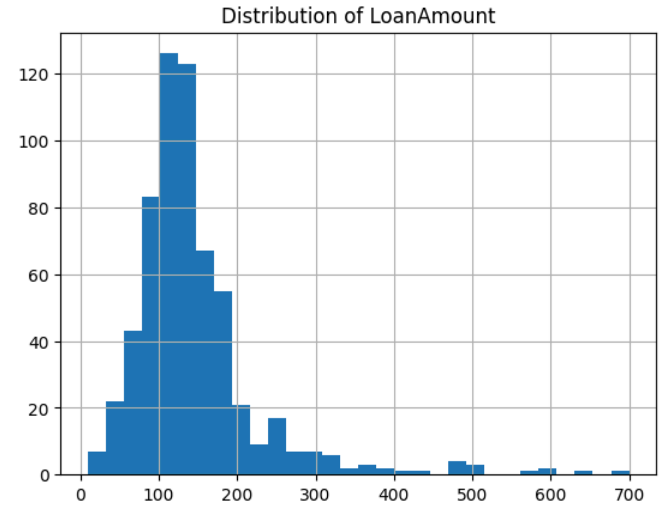
This means that the majority of the households depend highly on the primary coapplicant’s income.

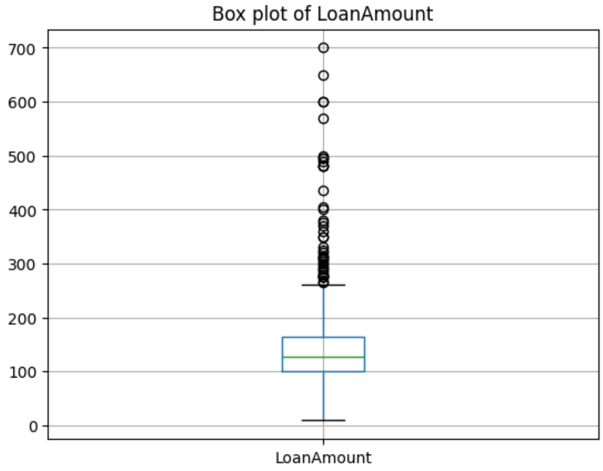




3.3 Loan_Amount

- Distribution is right-skewed, with many applicants requesting 100NGN - 150NGN range, and a few extreme values up to 600NGN - 700NGN
- Boxplot confirms clear outliers, which were corrected using IQR clipping (which results in LoanAmount_Clipped).
- After clipping, the distribution became more asymmetrical, indicating a more balanced loan size range for analysis.

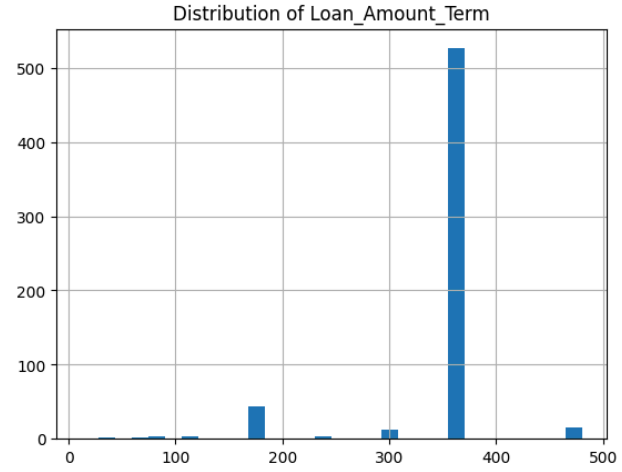


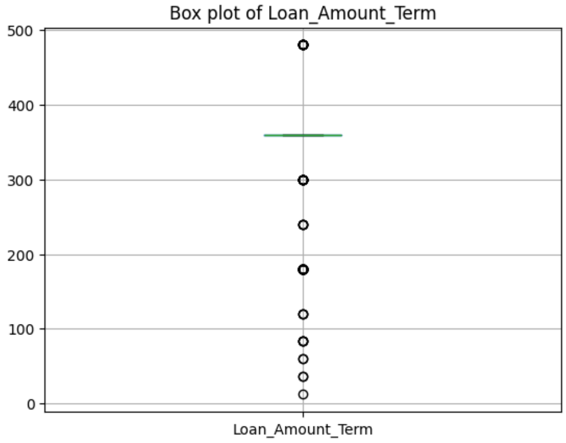


3.4 Loan_Amount_Term

- The histogram shows most loans at 360months (30 years)
- Few cases exist around 120, 180, or 240 months, suggesting short-term or mid-term loans are rare.

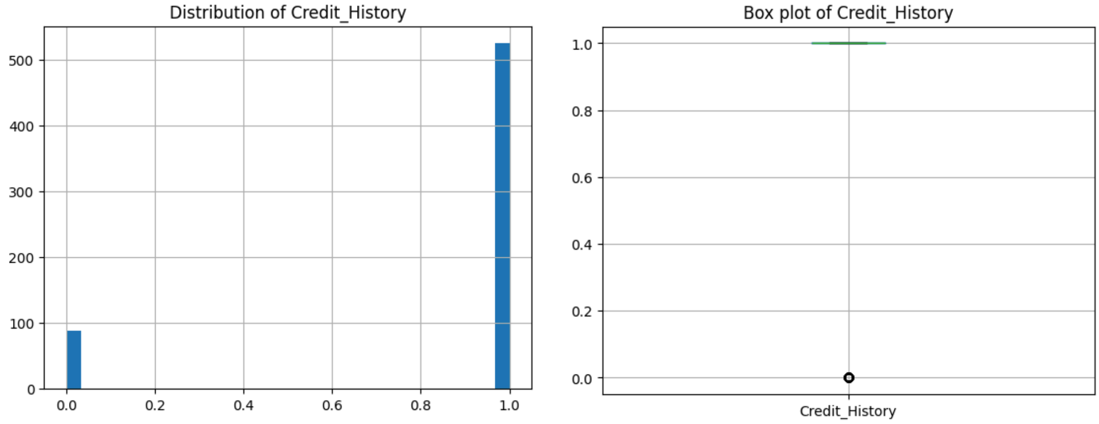
This means that the institution mainly issues long-term mortgages which is typical in housing finance.





3.5 Credit_History

- The chart shows major dominance of Credit_History, with very few applicants having 0.0
- This means most applicants have prior repayment records and likely access to formal credit systems.
- This implies that Credit_History will be a strong but possible biased predictor of Loan_Status.



3.6 Categorial Variables

Feature	**Observation**	**Insight**
Gender	Males = 500, Females = 100	Male applicants dominate.
Married	Majority are married	Suggests higher dual-income potential.
Dependents	Most have 0 dependents	Fewer dependents may indicate lower financial stress.
Education	Graduates > Non-graduates	Education positively correlated with access to formal loans.
Property_Area	Semi-urban leads, then Urban, then rural	Semi-urban applicants have highest loan representation
Loan_Status	Y (Approved) = 400+, N (Rejected) = 150-200	Roughly 70% approval rate

4. Loan Approval Behavior (Loan_Status)

General Trend

- 70-75% of applicants are approved (Loan_Status = Y).
- 25-30% are denied (Loan_Status = N).
  This suggests relatively approval criteria but also potential bias if approvals concentrate in certain applicant groups.

Influential Drivers

- **Credit_History:** is strongly linked with approvals, most successful applicants have Credit_History = 1.
- **Loan_Amount:** higher requested loans tend to have slightly higher rejection rates.
- **Income levels:** higher ApplicantIncome doesn’t automatically mean approval, credit behavior carries more weight.
- **Property_Area:** semi-urban and urban show more approvals than rural.
- **Education & Employment:** graduates and salaried/self-employed with stable income are more successful.

5. Conclusion & Recommendations

5.1 Conclusions

This EDA revealed clear patterns in applicant characteristics and loan approval behavior. Credit history emerged as the strongest determinant of approval, while income and loan amount showed wide variability with noticeable outliers. Most applicants are male, married, and graduates applying for standard 30year loans.

After addressing missing values and outliers, the dataset is now clean, consistent, and ready for modelling. Overall, the analysis provides a solid foundation for building predictive models and guiding fair, data-driven lending decisions.

5.2 Recommendations

Data & Quality

- Require Credit_History during application to avoid uncertainty.
- Add data validation on numeric fields (no zero or negative amounts).
- Consider features like *Credit_HIstory_Missing* (1 if missing originally) for future modelling.

Modelling Readiness

- Dataset is now clean, imputed, and ready for supervised learning
- Handle class imbalance (if Y:N ratio > 2:1) using weighted models or resampling.

Business Strategy

- Approval is heavily tied to Credit history. This ensures fairness review (avoid over-reliance on bureau score).
- Applicants from rural areas or without coapplicants may need special policy consideration or adjusted scoring.
- Income skew implies the lender serves mainly middle-to-low-income segments. Policies should reflect affordability.