EXPLORATORY DATA ANALYSIS (EDA)

ON

A HOME LOAN DATASET

Prepared by

SAHEED OLAYEMI OLAYINKA

PROJECT OVERVIEW

In this project, I conducted an in-depth Exploratory Data Analysis (EDA) on a Home Loan dataset. The objective is to understand the underlying structure, trends, and relationships in the data through data cleaning, visualization, and statistical analysis. This initial investigation is essential for uncovering patterns that may influence loan approvals and risk assessment.

Project Objective

The primary goal of this project is to perform a thorough exploratory analysis of the Home Loan dataset. Specific objectives include:

- Data Cleaning and Preparation: Identify and handle missing values, inconsistencies, and outliers in the dataset.
- Descriptive Analysis: Understand the distribution of key features such as applicant income, loan amounts, and property characteristics.
- Correlation Analysis: Explore relationships between variables (e.g., the impact of credit history on loan approval) using correlation matrices and statistical measures.
- Visualization: Generate meaningful charts and plots (histograms, scatter plots, box plots, etc.) to visually represent data distributions and relationships.
- Insight Generation: Summarize and interpret findings to support subsequent predictive modeling and strategic decision-making in home loan processing.

Comprehensive EDA Report

(All recommendations are justified by the EDA results.)

1. Executive summary

The EDA reveals that credit history is the strongest determinant of loan approval. Applicants with a credit history of “1” are far more likely to have their loans approved. While income and loan amount have weak correlations with approval, they interact meaningfully when combined (e.g., Debt-to-Income ratio).

Other important observations:

- Most applicants are male, married, graduates, and salaried.
- Numeric variables such as income and loan amount are right-skewed and contain outliers.
- Property area and marital status also show mild associations with approval.

These findings suggest that credit history, applicant type, and property area are strong predictive features. Future modeling should emphasize engineered features (e.g., total income, EMI ratio) and tree-based algorithms that handle non-linearity effectively..

2. Univariate findings

Finding (evidence) fom Numerical variables

Feature	Shape	Observation	Outliers
ApplicantIncome	Right-skewed	High variance due to a few very high earners	Present

CoapplicantIncome	Right-skewed	Lower magnitude; some zeros (no co-applicant)	Present
LoanAmount	Right-skewed	Most loans between 100–150	Present
Loan_Amount_Term	Nearly constant	Mostly 360 months	Minimal
Credit_History	Binary	Dominantly 1 (good history)	None

Recommendations & Rationale

1. **Log-transform skewed variables** (e.g., *Application Income Coapplication Income* and *Loan amount*).
   - Rationale: Skewness and unstable variance use to have influence on distance/linear models. EDA showed these variables are right-skewed; log(1+x) reduces influence of extremes and makes distributions more Gaussian-like.
   - Action:df['ApplicationIncome'] = np.log1p(df['ApplicationIncome']) etc.
2. **Outlier handling (Winsorize or IQR-capping) for extreme features**
   - Rationale: boxplots show extreme tails. Outliers can distort many algorithms (especially KNN, SVM, linear models).
   - Action: Cap values at Q1 - 1.5 x IQR and Q3 + 1.5 x IQR, or Winsorize the top/bottom 1–2%.
3. **Scaling**
   - Rationale: Features have different ranges (e.g Credit_History 0~1, dependant 1~3 etc). For distance-based or regularized models, scale features.
   - Action:Use StandardScaler for models like Logistic/SVM or RobustScaler if outliers remain.

4. **Target encoding:**

- Rationale: Loan Status is a categorical data with Y (yes) and N (no).
- Action: For binary classification, we use the integer labels classifier. we can label it as binary 1 and 0. Where 1 will map to Y and 0 mapto N.

Finding (evidence) from the categorical features

Variable	**Observation**	**Count Summary**
Gender	Majority male	472 males, 106 females
Married	Majority married	377 married, 201 single
Education	Majority graduate	457 graduates
Self_Employed	Mostly not self-employed	501 not self-employed
Property_Area	Semiurban most common	225 semiurban, 184 urban, 169 rural
Loan_Status	69.7% approved	403 approved, 175 rejected

Recommendations & Rationale

- Rationale: The dataset contains a higher proportion of male applicants, Educated and salaried individuals compared to female,uneducated and self-employed applicants. This imbalance may introduce bias into the model, as the algorithm might overfit to the majority groups (males, educated and salaried applicants). Consequently, the model's predictions could be less accurate or fair for underrepresented groups such as females, non-educated and self-employed individuals.
- Action:To mitigate this imbalance, techniques such as resampling (oversampling the minority groups or undersampling the majority), using class weights, or collecting more

balanced data can be applied. Additionally, fairness metrics should be monitored during model evaluation to ensure that predictions remain equitable across all demographic groups.

3. Bivariate findings

Correlation with Target (Loan_Status)

Feature	Correlation	Interpretation
Credit_History	**+0.41 (Strong)**	Key predictor — applicants with credit history=1 have high approval chances.
ApplicantIncome	-0.01 (Weak)	Minimal linear effect.
CoapplicantIncome	-0.07 (Weak)	Negligible relationship.
LoanAmount	-0.04 (Weak)	Slightly higher loan amounts more likely rejected, but weak trend.
Loan_Amount_Term	-0.03 (Weak)	Little impact.
Dependents	+0.03 (Weak)	Very slight positive association.

Interpretation:

Credit_History dominates predictive power; numeric income-related features show little correlation but may interact non-linearly with other factors.

Recommendations & rationale

1. Keep high-signal features: Credit_History and load amount.

2. Weak features can be transformed or combined rather than dropped immediately.

   Further investigations are required to create new features from features that have weak correlation with the target variable.

4. Trivariate & grouped multivariate findings

Core evidence: When analyzing grouped variables:

- **Married + Credit_History:** Married applicants with credit history=1 show the highest approval rates (~80%).
- **Education + Property_Area:** Graduates in semiurban areas tend to have higher approval chances.
- **Income + LoanAmount + Loan_Status:** Even at similar loan amounts, applicants with good credit history receive more approvals.

Feature engineering recommendations:

- Total Income (`TotalIncome = ApplicantIncome + CoapplicantIncome`)

  Combine both applicant and coapplicant incomes to provide a more comprehensive measure of the household's earning capacity.
- Log of Total Income (`Log_TotalIncome = np.log1p(TotalIncome)`)

  Apply a logarithmic transformation to normalize the highly skewed income distribution and reduce the influence of extreme values.
- Debt-to-Income Ratio (`DTI = LoanAmount / TotalIncome`)

  Represents the proportion of income allocated to loan repayment, serving as an indicator of the applicant's affordability and credit risk.

- Equated Monthly Instalment Feature (`EMI = LoanAmount / Loan_Amount_Term`)

  Estimates the monthly repayment burden for each applicant, helping to assess the financial strain relative to income.

  Implementation notes: These engineered features should be added before modeling. Apply one-hot encoding for categorical variables and log-transform numeric ones to reduce skew.

5. Model selection & evaluation

Evidence behind the recommendation:

EDA findings indicate:

- Nonlinear relationships (e.g., between income and approval).
- Dominance of categorical and binary predictors (e.g., Credit_History).
- Presence of outliers and skewed numeric variables.

Hence, **tree-based classifiers** (RandomForest, XGBoost, LightGBM) are suitable because they:

- Handle mixed data types.
- Are robust to skewness and outliers.
- Naturally perform feature selection and capture interactions.

Recommended Pipeline

1. **Prepare Dataset**: Clean data, handle missing values, encode categories.
2. **Train-Test Split**: 80/20 stratified by Loan_Status.
3. **Baseline Model**: Logistic Regression for interpretability.
4. **Primary Models**: RandomForest, XGBoost, or CatBoost for performance.

5. **Handle Class Imbalance**: Apply class weights or SMOTE.
6. **Hyperparameter Tuning**: Use GridSearchCV or RandomizedSearchCV.
7. **Evaluation Metrics**: Accuracy, Precision, Recall, F1-score, AUC.
8. **Explainability**: Use SHAP values or permutation importance for interpretation.

6. EDA evidence

Observation	**Evidence**	**Implication**
Credit history strongly predicts loan approval	Correlation = +0.41	Must be retained as a core feature
Income variables are right-skewed	Histograms & Boxplots	Apply log transformations
Outliers exist in income & loan amount	Boxplots	Apply capping or robust scaling
Categorical imbalance (Gender, Married, Education)	Frequency plots	Use class weighting or resampling
Weak numeric correlations	Correlation heatmap	Engineer derived features
Property_Area affects approval	Bar charts	Include as categorical variable

7. Production-oriented suggestions

- Feature Pipeline Automation: Implement preprocessing steps (encoding, imputation, scaling) using `sklearn.Pipeline`.
- Model Monitoring: Track approval prediction accuracy and fairness across gender/education.

- Explainable AI (XAI): Use SHAP/LIME to justify model decisions in production.
- Data Refresh Policy: Regularly retrain models with new data to capture evolving applicant behavior.
- Integration: Deploy trained model as an API for bank officers to query loan approvals in real time.
- Dashboard Visualization: Create Power BI or Streamlit dashboard showing approval rates by gender, area, and credit history.

Conclusion

The EDA successfully identifies key trends and relationships in the Home Loan dataset. Credit history remains the strongest determinant of approval, while demographic and financial attributes play secondary roles. Careful preprocessing (log transformation, scaling, and feature engineering) will improve downstream predictive modeling. The insights provide a solid foundation for developing accurate, fair, and production-ready credit risk models.