Problem Statement: A financial institution wants to automate its loan approval process. We are tasked with building a machine learning model to predict whether a loan application should be approved or rejected based on the applicant's profile.

Objectives:

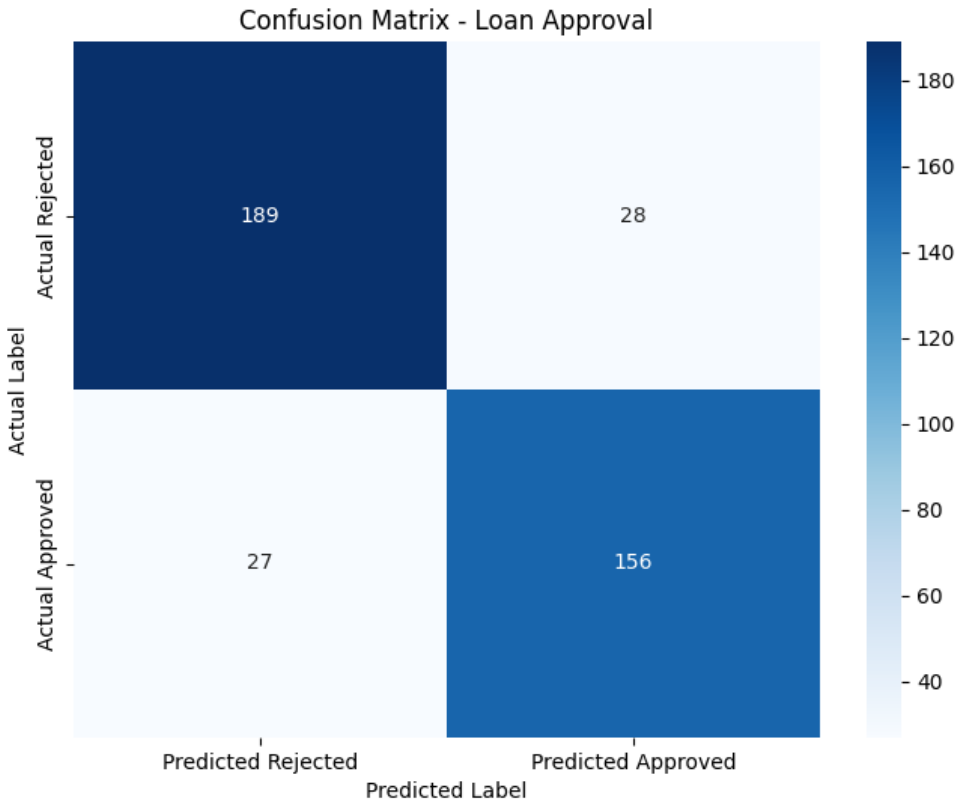
1. To perform data preprocessing to clean and prepare applicant data for modeling.
2. To build a classification model to predict loan approval status.
3. To evaluate the model's performance using key metrics like accuracy.
4. To visualize and identify the key factors that influence the loan approval decision.
5. To provide actionable recommendations to the bank based on the analysis.

Methodology: The dataset was loaded using pandas. We performed two key preprocessing steps:

1. **Feature Selection:** The name and city columns were dropped as they are unique identifiers and not useful for predicting a loan outcome.
2. **Data Conversion:** The target variable, loan_approved, was in a boolean format (True/False). This was converted into a binary integer format where **1** represents 'Approved' and **0** represents 'Rejected' for the model.
3. **Model:** We selected **Logistic Regression** for this task. It is a highly efficient and interpretable model, making it ideal for a binary classification problem (Approved/Rejected).
4. **Training:** The data was split into two parts: an **80% Training Set** (to teach the model) and a **20% Testing Set** (to evaluate its performance on unseen data).

The model performed very well, achieving an accuracy of **86.25%**.

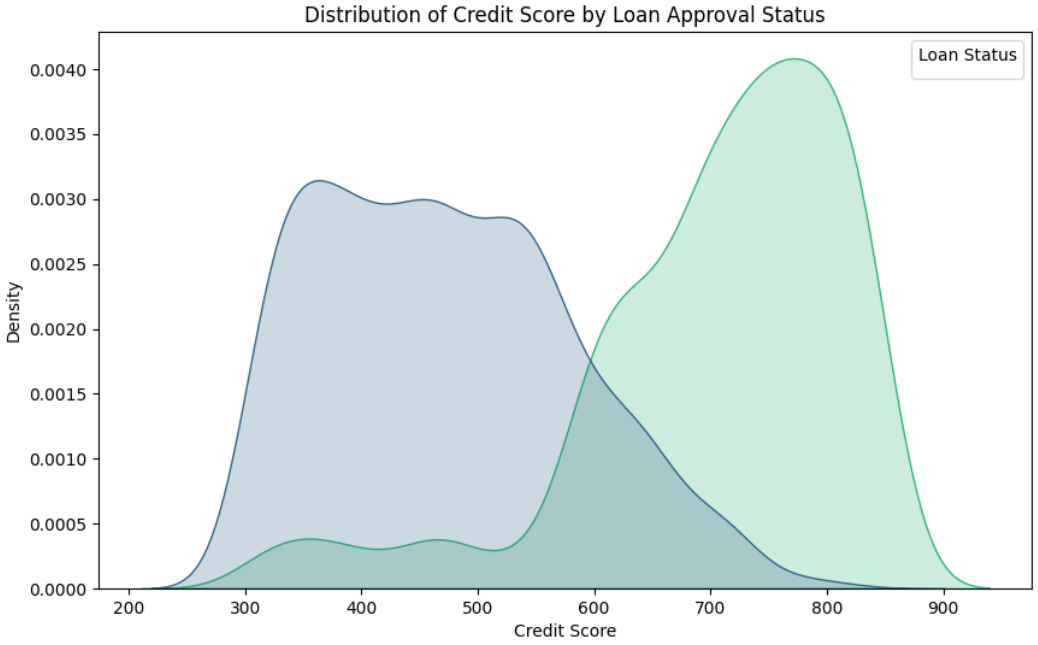
Insight 1: Model Performance (Confusion Matrix)



"The confusion matrix shows the model's performance in detail.

- **True Negatives:** 189 'Rejections' were correctly predicted.
- **True Positives:** 156 'Approvals' were correctly predicted.
- **Errors:** The model made only 28 false positives (incorrectly approved) and 27 false negatives (incorrectly rejected), demonstrating high reliability."

Insight 2: Key Factor for Approval (Credit Score)



"This visualization is the key insight. The two distributions show a clear separation:

- **Rejected (0) applicants** (the blue curve) are clustered around the lower credit scores.
- **Approved (1) applicants** (the green curve) are clustered around the higher credit scores.

This confirms that **credit score is the single strongest predictor** of loan approval."

Based on the analysis, we recommend the following:

1. **Automate Screening:** With an accuracy of **86.25%**, the model is a reliable tool for automating the initial screening of loan applications, which will save significant time.
2. **Prioritize Credit Score:** The 'credit_score' feature is the most important factor. The bank should use this model to automatically fast-track applications with high credit scores and flag those with low scores for immediate manual review or rejection.
3. **Deploy the Model:** We recommend deploying this Logistic Regression model as a first-pass filter for all incoming loan applications.