Title: Loan Default Prediction Project

Project Description: In this in-class exercise, you will be working with a real-world dataset from a bank that contains information about loan borrowers. The objective of this project is to develop a machine learning model to predict whether a new borrower is likely to default on a loan. To achieve this goal, you will need to perform data preprocessing, which includes handling missing values, identifying and treating outliers, and exploring feature engineering possibilities.

Dataset Description: The dataset contains the following columns:

1. **ID**: A unique identifier for each entry or borrower.
2. **year**: The year in which the loan application was submitted (e.g., 2019).
3. **loan\_limit**: Indicates the loan limit or type of loan (e.g., "cf" - possibly a loan category).
4. **Gender**: The gender of the borrower (e.g., "Male," "Female," "Sex Not Available").
5. **approv\_in\_adv**: Approval status, possibly indicating if the loan was pre-approved (e.g., "nopre" - no pre-approval, "pre" - pre-approved).
6. **loan\_type**: Type of loan (e.g., "type1," "type2," "type3").
7. **loan\_purpose**: The purpose for which the loan is taken (e.g., "p1," "p4").
8. **Credit\_Worthiness**: An indication of the borrower's creditworthiness (e.g., "l1").
9. **open\_credit**: Information about open credit lines (e.g., "nob/c" - possibly "no business/commercial").
10. **business\_or\_commercial**: Indicates if the loan is for business or commercial purposes (e.g., "nob/c").
11. **loan\_amount**: The amount of the loan.
12. **rate\_of\_interest**: The interest rate on the loan.
13. **Interest\_rate\_spread**: The interest rate spread.
14. **Upfront\_charges**: Upfront charges associated with the loan.
15. **term**: The term of the loan in months (e.g., 360 months).
16. **Neg\_ammortization**: Indicates if there is negative amortization on the loan (e.g., "not\_neg" - not negative amortization).
17. **interest\_only**: Indicates if the loan has an interest-only payment option (e.g., "not\_int" - not interest-only).
18. **lump\_sum\_payment**: Indicates if there is a lump-sum payment option (e.g., "not\_lpsm" - not lump-sum payment).
19. **property\_value**: The value of the property associated with the loan.
20. **construction\_type**: The type of construction related to the property (e.g., "home").
21. **occupancy\_type**: The type of occupancy of the property (e.g., "1U" - possibly "1-unit").
22. **Secured\_by**: Information about the loan security (e.g., "sb").
23. **total\_units**: Total units related to the property.
24. **income**: The borrower's income.
25. **credit\_type**: Type of credit (e.g., "EXP," "CRIF," "CIB").
26. **Credit\_Score**: The borrower's credit score.
27. **co-applicant\_credit\_type**: Type of credit for the co-applicant (if applicable).
28. **age**: The age of the borrower.
29. **submission\_of\_application**: Information about the submission of the loan application.
30. **LTV**: Loan-to-value ratio.
31. **Region**: The region where the loan application was made (e.g., "south," "North," "central").
32. **Security\_Type**: The type of security associated with the loan.
33. **Status**: Indicates the status of the loan application (e.g., 0 for non-default, 1 for default).
34. **dtir1**: Possibly a debt-to-income ratio or related financial metric

Tasks:

1. **Data Loading and Exploration**:
   * Load the dataset into a suitable data structure (e.g., Pandas DataFrame).
   * Explore the dataset to understand its structure and summary statistics.
   * Identify any missing values and outliers in the dataset.
2. **Missing Values Treatment**:
   * Implement a strategy to handle missing values in the dataset. You can choose from methods like:
     + Imputation (e.g., mean, median, mode)
     + Removal of rows or columns with too many missing values
     + Interpolation
     + Advanced techniques like predictive modeling to impute missing values
3. **Outlier Detection and Treatment**:
   * Identify outliers in the dataset using appropriate visualization or statistical techniques.
   * Decide on a method to treat outliers, which may include:
     + Truncation (capping values)
     + Transformation (e.g., log transformation)
     + Removal of extreme outliers
4. **Feature Engineering**:
   * Explore feature engineering possibilities. This may include:
     + Creating new features based on existing ones.
     + Encoding categorical variables (e.g., one-hot encoding).
     + Scaling or standardizing numerical features.
     + Binning or discretization of continuous variables.
     + Interaction terms or polynomial features.