Homework Assignment 2 - Imputation Evaluation

Use the provided dataset `hw-2-training-data.csv` to build a binary classifier (e.g., logistic regression, SGDClassifier) that predicts the `has_disease` column. Make sure to:

- Carefully handle the missing data by choosing suitable imputation methods for both categorical and numeric columns.
- Check for outliers.
- Standardize numeric features.
- Encode categorical features appropriately before training.
- Make sure to build imputation models/scaling on the training set only.

Your aim is to maximize your F1 score! To do this, you might try:

- Removing columns that you don't think are necessary.
- Attempting different imputation methods.
- Creating new features (note - we haven't done this yet, but you can inspect the data to create "meta-features" that combine existing features).
- Try a different classifier (anything in Scikit-Learn is fair game).

Be careful not to overfit, as this might lead to a classifier that performs poorly on the test data! When you are done, train your model on the full training data, and then obtain predictions for the `hw-2-test-data.csv`. Save your predictions to a file called `answers.csv`, which should contain a single column of predictions. For the test data, impute any missing data using the transformers that have been trained on your training data before testing.

Submit both your notebook and your answers.csv file

Rubric

Points will given for each of the following sections present in your notebook. Each section must be named with a heading. Code must be documented where necessary, and markdown should be present to explain summary findings at the end of each section. Points will be awarded based on the thoroughness and quality of code in each section.

1. **Exploratory Data Analysis** - Did you:
    - Explore your data
    - Examine distributions
    - Look for correlations
    - Identify nulls
2. **Preprocessing** - As necessary, did you correctly:
    - Handle outliers
    - Encode features
    - Standardize features
    - Handle nulls (whether by imputation or some other method)
3. **Modeling & Evaluation** - Did you:
    - Try more than one classifier
    - Use a cross-validation procedure
    - Avoid data leakage
    - Correctly evaluate F1-Score
4. **Test data** - Did you:
    - Train you model appropriately
    - Impute data correctly
    - Achieve at least 65% F1 on the test data
    - Save your answers to an `answers.csv` file