Prepare the Data

Notes:

- Work on copies of the data (keep the original dataset intact).
- Write functions for all data transformations you apply, for five reasons:

  So you can easily prepare the data the next time you get a fresh dataset

  So you can apply these transformations in future projects

  To clean and prepare the test set

  To clean and prepare new data instances once your solution is live

  To make it easy to treat your preparation choices as hyperparameters

  1. Data cleaning:
     - Fix or remove outliers (optional).
     - Fill in missing values (e.g., with zero, mean, median…) or drop their rows (or columns).
  2. Feature selection (optional):
     - Drop the attributes that provide no useful information for the task.
  3. Feature engineering, where appropriate:
     - Discretize continuous features.
     - Decompose features (e.g., categorical, date/time, etc.).
     - Add promising transformations of features (e.g., log(x), sqrt(x), x^2, etc.).
     - Aggregate features into promising new features.
  4. Feature scaling: standardize or normalize features.