Titanic-like Data Processing Script Documentation

This script demonstrates how to process a small custom Titanic-like dataset using pandas and numpy in Python.

1. Dataset Initialization:

- A custom dataset is created using a dictionary with keys representing column names similar to the Titanic dataset.
- The dataset includes columns such as PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked.

2. Missing Values Check:

- The script prints the count of missing values in each column before any cleaning.

3. Missing Values Handling:

- Missing 'Embarked' values are filled using the mode (most frequent value).
- Missing 'Age' values are filled using the median of the Age column.
- Missing 'Cabin' values are replaced with "Not Assigned".

4. Age Binning:

- A function `categorize_age` is defined to classify passengers into age groups: Child, Teen, Adult, Middle-Aged, and Senior.
- This function is applied to the 'Age' column to create a new column called 'AgeCategory'.

5. Extracting Surnames:

- A new column 'Surname' is created by extracting the last name from the 'Name' column (i.e., the part before the comma).

6. Final Output:

- The script prints the missing values after processing and a sample of the cleaned and transformed data with selected columns.

Note:

- This code is meant for educational purposes and mimics a real-world Titanic dataset scenario.
- The values have been modified to ensure uniqueness and avoid direct duplication.