Author: Madhurima Rawat

Data Cleaning and Preprocessing with Pandas

Use the Pandas library in Python to clean and preprocess large datasets, addressing issues such as missing values, outliers, and inconsistent data.

Explanation of the Code

1. **Dataset**: The code uses the Titanic dataset, which contains information about passengers on the Titanic, such as age, sex, fare, survival status, and more.
2. **Missing Values Handling**:
   - We check for missing values in the dataset using `isnull().sum()` to identify columns with missing data.
   - For the `age` column, we fill missing values with the **median** (since age is numerical and the median is less sensitive to outliers).
   - For the `embarked` column, we fill missing values with the **mode** (the most frequent value).
   - The `cabin` column is dropped entirely due to a high percentage of missing values, which might not be useful for analysis.
3. **Outliers Handling**:
   - We use a **boxplot** to visualize the distribution of the `age` column and detect potential outliers.
   - Outliers are detected using the **Interquartile Range (IQR)** method: values outside 1.5 times the IQR are considered outliers and removed.
   - Another boxplot is displayed after removing outliers to show the cleaned data.
4. **Inconsistent Data Handling**:
   - The `sex` column may have inconsistent data (e.g., "Male", "male", "Female", "female"). We normalize this column by converting all values to lowercase to ensure consistency.
5. **Feature Engineering**:
   - A new feature called `family_size` is created by combining the `sibsp` (siblings/spouses aboard) and `parch` (parents/children aboard) columns. This feature gives insight into the size of the passenger's family.
6. **Statistical Measures**:

- Basic statistical measures such as mean, standard deviation, and percentiles are calculated using `describe()` on the cleaned dataset to summarize the distribution of numerical features.

7. **Data Distribution Visualization**:
   - **Age Distribution**: A histogram is plotted for the `age` column to visualize its distribution after cleaning.
   - **Family Size Distribution**: A histogram is plotted for the newly created `family_size` column to visualize the distribution of family sizes.

3. Output

Missing Values Before and After Handling

Before handling the missing values, we observe the following:

- **Age**: 177 missing values.
- **Embarked**: 2 missing values.
- **Cabin**: 687 missing values.

After handling missing values:

- **Age**: Missing values are filled with the median.
- **Embarked**: Missing values are filled with the mode.
- **Cabin**: The column is dropped due to excessive missing values.

Boxplot for Age (Before and After Removing Outliers)

- **Before**: A boxplot shows that the `age` column contains some outliers, particularly in the higher age range.
- **After**: The boxplot after outlier removal shows a cleaner distribution of age data without extreme values.

Summary Statistics for the Cleaned Dataset

The `describe()` function provides the following insights into the cleaned dataset:

- **Age**: The average age is around 29.5 years with a standard deviation of approximately 14.5 years.
- **Family Size**: The average family size is around 1.5, with a standard deviation of about 2.0.

Data Distribution Visualizations

1. **Age Distribution**: The histogram shows the distribution of ages after removing outliers. The distribution appears to be relatively normal, with more passengers in the 20-40 age range.

2. **Family Size Distribution**: The histogram of family sizes reveals that most passengers traveled alone or with a small family.

Explanation of the Concepts

1. Introduction

Data cleaning and preprocessing are essential steps in any data analysis or machine learning project. These steps help to ensure that the data is accurate, consistent, and ready for modeling. Common tasks include handling missing values, removing outliers, and addressing inconsistent data.

2. Handling Missing Values

- **Identifying Missing Data**: The `isnull()` function helps detect missing data in the dataset.
- **Imputation**:
  - For numerical data like `age`, we use the **median** to fill missing values, as it is less affected by outliers.
  - For categorical data like `embarked`, we use the **mode** (most frequent value) to fill missing values.

3. Handling Outliers

Outliers can significantly distort statistical analyses and machine learning models. We used the **Interquartile Range (IQR)** method to detect and remove outliers from the `age` column. The data outside 1.5 times the IQR is considered an outlier and removed.

4. Handling Inconsistent Data

Inconsistent data, such as different capitalization for categorical variables, can lead to incorrect analysis. The `sex` column was normalized to lowercase to ensure consistency.

5. Feature Engineering

Feature engineering involves creating new features that might provide more insights for analysis. We created a new `family_size` feature by combining the `sibsp` and `parch` columns.

6. Statistical Measures and Visualizations

After cleaning the data, we calculated basic statistical measures (mean, standard deviation, etc.) and visualized the distributions of `age` and `family_size` to understand their characteristics.

Conclusion

Data cleaning and preprocessing are fundamental steps in preparing data for analysis or machine learning. By addressing missing values, outliers, and inconsistent data, we can improve the quality of the dataset, leading to more accurate and reliable models.