Data Analysis and Manipulation - Assignment

Problem Statement

This assignment focuses on utilizing the Titanic dataset to perform various data manipulation and analysis tasks. By exploring the dataset and implementing operations such as array manipulation, missing value handling, and data manipulation tasks, students will gain hands-on experience in working with real-world datasets.

Assignment Questions

1. Implementing Basic Array Operations:

a. Perform addition operation on the 'Fare' column of the Titanic dataset. What is the total fare paid by all passengers?

b. Subtract the 'Age' column from the 'Fare' column. What insights can you derive from this operation?

c. Multiply the 'SibSp' column (number of siblings/spouses aboard) with the 'Parch' column (number of parents/children aboard). What does the resulting column represent?

d. Divide the 'Fare' column by the 'Pclass' column (passenger class). Interpret the meaning of the values obtained.

2. Creating and Manipulating Multidimensional Arrays:

a. Create a 2D array using the 'Age' and 'Fare' columns of the Titanic dataset. What are the dimensions of this array?

b. Perform element-wise multiplication on this array with a scalar value of 10. What effect does this operation have on the array?

c. Transpose the array created in part (a). What are the dimensions of the transposed array?

3. Create a New DataFrame with 5 Records:

a. Create a new DataFrame named 'CustomRecords' with 5 records randomly selected from the Titanic dataset. Include relevant columns such as 'Name', 'Age', etc.

4. Read the Dataset and Display the Records:

a. Read the Titanic dataset from the provided source and display the first 10 records.

b. Identify the number of columns and rows in the dataset.

5. Find Missing Values and Replace it:

a. Identify the columns with missing values in the Titanic dataset.

b. For each column with missing values, replace them with the appropriate valuea.

c. Verify that there are no missing values remaining in the dataset.

6. Data Manipulation Tasks:

a. Filter the Titanic dataset to include only passengers with ages between 20 and 30 years.

b. Sort the dataset based on the 'Fare' column in descending order.

c. Group the dataset by 'Pclass' (passenger class) and calculate the mean fare for each class.

Guidelines

1. Foundational Knowledge
   - Understand basic array operations and multi-dimensional arrays.
   - Familiarize yourself with data manipulation tasks in Python, such as filtering, sorting, grouping, and merging datasets.

2. Data Exploration
   - Analyze the Titanic dataset's structure and characteristics.
   - Gain insights into the distribution of passenger classes, ages, genders, etc.

3. Preprocessing and Parameter Selection
   - Handle missing values in the dataset.
   - Perform tasks like filtering, sorting, grouping, and merging datasets as required.

4. Constructing a New DataFrame
   - Create a new DataFrame with 5 records containing relevant features such as passenger name, age, ticket class, etc.

5. Reading Dataset and Displaying Records
   - Read the Titanic dataset and display a sample of records.

6. Handling Missing Values
   - Find missing values in the dataset and replace them with appropriate values, such as the mean, median, etc...

7. Data Manipulation Tasks
   - Perform filtering, sorting, grouping, and merging operations on the Titanic dataset.

Step-by-Step Approach to Titanic Dataset Analysis

1. Setup and Data Preparation
   - Import necessary libraries: Pandas, NumPy, etc.
   - Load the Titanic dataset for analysis.

2. Data Exploration
   - Explore dataset characteristics using Pandas functions (e.g., `head()`, `info()`, `describe()`).
   - Visualize dataset features (e.g., age distribution, survival rate by class) using Matplotlib or Seaborn.

3. Preprocessing and Data Manipulation
   - Handle missing values using Pandas functions (e.g., `fillna()`, `dropna()`).
   - Perform data manipulation tasks like filtering, sorting, grouping, and merging datasets.

4. Creating a New DataFrame
   - Create a new DataFrame with 5 records containing relevant features.

5. Reading Dataset and Displaying Records
   - Read the Titanic dataset using Pandas' `read_csv()` function.
   - Display a sample of records using Pandas functions.

6. Handling Missing Values
   - Identify missing values in the dataset using Pandas functions.
   - Replace missing values with appropriate values (mean, median, etc..) using Pandas functions.

7. Data Manipulation Tasks
   - Perform filtering, sorting, grouping, and merging operations on the Titanic dataset using Pandas functions.

References

Dataset Link : https://www.kaggle.com/competitions/titanic/data
NumPy Documentation : https://numpy.org/doc/
Pandas documentation : https://pandas.pydata.org/docs/