Data Analytics Assignment: Titanic Dataset Analysis

Objective

The purpose of this assignment is to practice data manipulation, analysis, and visualization using the Titanic dataset. You will explore passenger information to gain insights into survival rates based on different factors such as gender, passenger class, and age. You will also clean the data and perform exploratory data analysis (EDA).

Task 1: Data Exploration

1. **Load the data**: Import the dataset using Pandas.
    - Inspect the first few rows of the dataset.
    - Display the shape of the dataset (number of rows and columns).
    - Get a summary of the dataset using `.info()` and `.describe()` methods.
2. **Handling Missing Values**:
    - Identify which columns have missing values and how many missing values each column has.
    - Handle missing values for the `Age` column by filling them with the median age.
    - Drop the `Cabin` column as it contains too many missing values.
3. **Data Cleaning**:
    - Drop any remaining rows with missing data.
    - Ensure that columns like `Survived`, `Pclass`, and `Embarked` are of the correct data type.

Task 2: Exploratory Data Analysis (EDA)

1. **Analyse Survival Rates**:
    - Calculate the overall survival rate (the percentage of passengers who survived).
    - Compare the survival rates between different passenger classes (`Pclass`).
    - Compare the survival rates between male and female passengers.
    - Compare survival rates across different age groups (create bins for age ranges: 0-10, 11-20, etc.).
2. **Visualizations**:
    - Create a bar chart to show the number of passengers who survived vs. who did not survive.
    - Create a pie chart to show the proportion of survivors based on gender.
    - Create a bar chart to show the survival rate for each passenger class.

- Create a histogram to show the age distribution of passengers who survived and those who did not.

Task 3: Advanced Analysis

1. **Correlation Analysis**:
   - Use the `.corr()` method to calculate the correlation between different numerical features such as `Pclass`, `Age`, `Fare`, and `Survived`.
   - Plot a heatmap using Seaborn to visualize the correlation matrix.

2. **Survival Prediction Insights**:
   - Based on your analysis, what factors seem to have a strong correlation with survival? Provide an interpretation of the results based on gender, age, class, or other features.

Deliverables:

- Submit your Python code for the assignment in a `.py` or `.ipynb` file.
- Provide a brief report (PDF) summarizing your findings, with visualizations embedded.
- Include all visualizations in your report and describe your insights for each.