TASK-2 (Data Cleaning and EDA on a Dataset)

PROGRAM:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the Titanic dataset
url =
'https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv'
titanic_df = pd.read_csv(url)

# Display the first few rows of the dataset
print(titanic_df.head())

# Display basic information about the dataset
print(titanic_df.info())

# Display summary statistics
print(titanic_df.describe())

# Data Cleaning
# Handling missing values
# Fill missing 'Age' values with the median age
titanic_df['Age'].fillna(titanic_df['Age'].median(), inplace=True)

# Fill missing 'Embarked' values with the most common embarkation port
titanic_df['Embarked'].fillna(titanic_df['Embarked'].mode()[0],
inplace=True)

# Drop the 'Cabin' column due to too many missing values
titanic_df.drop(columns='Cabin', inplace=True)

# Drop rows with missing 'Fare' values
titanic_df.dropna(subset=['Fare'], inplace=True)

# Converting categorical variables to numeric
# Convert 'Sex' to numeric (Male: 0, Female: 1)
titanic_df['Sex'] = titanic_df['Sex'].map({'male': 0, 'female': 1})
```

```
# Convert 'Embarked' to numeric
titanic_df['Embarked'] = titanic_df['Embarked'].map({'C': 0, 'Q': 1,
'S': 2})

# Exploratory Data Analysis (EDA)
# Correlation heatmap
plt.figure(figsize=(12, 6))
numerical_df=titanic_df.select_dtypes(include=['number'])
sns.heatmap(numerical_df.corr(), annot=True, cmap='coolwarm',
linewidths=0.5)
plt.title('Correlation Heatmap')
plt.show()

# Distribution of survivors vs. non-survivors
plt.figure(figsize=(10, 5))
sns.countplot(x='Survived', data=titanic_df, palette='Set2')
plt.title('Distribution of Survivors vs. Non-Survivors')
plt.show()

# Survival rate by sex
plt.figure(figsize=(10, 5))
sns.barplot(x='Sex', y='Survived', data=titanic_df, palette='Set2')
plt.title('Survival Rate by Sex')
plt.show()

# Survival rate by passenger class
plt.figure(figsize=(10, 5))
sns.barplot(x='Pclass', y='Survived', data=titanic_df, palette='Set2')
plt.title('Survival Rate by Passenger Class')
plt.show()

# Age distribution by survival status
plt.figure(figsize=(10, 5))
sns.histplot(data=titanic_df, x='Age', hue='Survived',
multiple='stack', palette='Set2')
plt.title('Age Distribution by Survival Status')
plt.show()
```