TASK DS_02

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

# Load the Titanic dataset from Kaggle
train_df = pd.read_csv('train.csv')
test_df = pd.read_csv('test.csv')

# Data Cleaning
print(train_df.info())
print(train_df.describe())

# Fill missing Age values with the median Age
train_df['Age'].fillna(train_df['Age'].median(), inplace=True)
test_df['Age'].fillna(test_df['Age'].median(), inplace=True)

# Drop the Cabin column since it has too many missing values
train_df.drop('Cabin', axis=1, inplace=True)
```

```
test_df.drop('Cabin', axis=1, inplace=True)

# Fill missing Embarked values with 'S'
train_df['Embarked'].fillna('S', inplace=True)
test_df['Embarked'].fillna('S', inplace=True)

# Exploratory Data Analysis (EDA)

# 1. Survival Rate
sns.countplot(x='Survived', data=train_df)
plt.title('Survival Rate')
plt.show()

# 2. Age Distribution
sns.distplot(train_df['Age'], kde=False)
plt.title('Age Distribution')
plt.show()

# 3. Class Distribution
sns.countplot(x='Pclass', data=train_df)
plt.title('Class Distribution')
plt.show()
```

```
# 4. Sex Distribution
sns.countplot(x='Sex', data=train_df)
plt.title('Sex Distribution')
plt.show()

# 5. Embarked Distribution
sns.countplot(x='Embarked', data=train_df)
plt.title('Embarked Distribution')
plt.show()

# 6. Survival Rate by Class
sns.barplot(x='Pclass', y='Survived', data=train_df)
plt.title('Survival Rate by Class')
plt.show()

# 7. Survival Rate by Sex
sns.barplot(x='Sex', y='Survived', data=train_df)
plt.title('Survival Rate by Sex')
plt.show()

# 8. Survival Rate by Embarked
```

```
sns.barplot(x='Embarked', y='Survived', data=train_df)
plt.title('Survival Rate by Embarked')
plt.show()
```