# Titanic Dataset - Data Cleaning & Exploratory Data Analysis (EDA)

## 1. Importing Libraries and Loading Data

We start by importing the necessary libraries and loading the Titanic dataset.

import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt  
import seaborn as sns  
  
# Load the Titanic dataset  
train\_path = "train.csv"  
test\_path = "test.csv"  
  
train\_df = pd.read\_csv(train\_path)  
test\_df = pd.read\_csv(test\_path)  
  
# Display basic information  
train\_df.info()  
train\_df.head()

## 2. Handling Missing Values

We check for missing values and handle them by filling or dropping columns accordingly.

# Check for missing values  
print(train\_df.isnull().sum())  
  
# Fill missing Age values with median  
train\_df["Age"].fillna(train\_df["Age"].median(), inplace=True)  
  
# Fill missing Embarked values with mode  
train\_df["Embarked"].fillna(train\_df["Embarked"].mode()[0], inplace=True)  
  
# Drop Cabin column due to too many missing values  
train\_df.drop(columns=["Cabin"], inplace=True)  
  
# Verify missing values are handled  
print(train\_df.isnull().sum())

## 3. Exploratory Data Analysis (EDA)

We explore the relationships between different variables using visualizations.

# Summary statistics  
print(train\_df.describe())  
  
# Survival rate by class  
sns.barplot(x="Pclass", y="Survived", data=train\_df)  
plt.title("Survival Rate by Passenger Class")  
plt.show()  
  
# Survival rate by gender  
sns.barplot(x="Sex", y="Survived", data=train\_df)  
plt.title("Survival Rate by Gender")  
plt.show()  
  
# Age distribution  
sns.histplot(train\_df["Age"], bins=30, kde=True)  
plt.title("Age Distribution")  
plt.show()

## 4. Feature Correlation Heatmap

We visualize the correlation between different numerical features.

# Correlation heatmap  
plt.figure(figsize=(10, 6))  
sns.heatmap(train\_df.corr(), annot=True, cmap="coolwarm", fmt=".2f")  
plt.title("Feature Correlation Heatmap")  
plt.show()