TASK-2

# Import required libraries

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

# Load Titanic dataset (from seaborn for convenience, or download from Kaggle)

# If using Kaggle dataset: df = pd.read\_csv("titanic.csv")

df = sns.load\_dataset("titanic")

# Preview data

print(df.head())

# -------------------------

# Data Cleaning

# -------------------------

# Check missing values

print("\nMissing values:\n", df.isnull().sum())

# Fill missing Age with median

df['age'] = df['age'].fillna(df['age'].median())

# Fill missing Embarked with mode (most frequent value)

df['embarked'] = df['embarked'].fillna(df['embarked'].mode()[0])

# Drop rows with too many missing values if necessary

df = df.dropna(subset=['sex', 'class', 'survived'])

# -------------------------

# Data Analysis

# -------------------------

# 1. Who survived more: males or females?

survival\_by\_gender = df.groupby("sex")["survived"].mean()

print("\nSurvival rate by gender:\n", survival\_by\_gender)

# 2. Did passenger class affect survival chances?

survival\_by\_class = df.groupby("class")["survived"].mean()

print("\nSurvival rate by class:\n", survival\_by\_class)

# 3. What was the survival rate by age group?

df['age\_group'] = pd.cut(df['age'], bins=[0, 12, 18, 35, 60, 100],

labels=['Child', 'Teen', 'Young Adult', 'Adult', 'Senior'])

survival\_by\_age\_group = df.groupby("age\_group")["survived"].mean()

print("\nSurvival rate by age group:\n", survival\_by\_age\_group)

# -------------------------

# Visualizations

# -------------------------

# Bar chart: Survival by gender

plt.figure(figsize=(6,4))

sns.barplot(x="sex", y="survived", data=df, palette="pastel")

plt.title("Survival Rate by Gender")

plt.ylabel("Survival Rate")

plt.show()

# Bar chart: Survival by class

plt.figure(figsize=(6,4))

sns.barplot(x="class", y="survived", data=df, palette="muted")

plt.title("Survival Rate by Passenger Class")

plt.ylabel("Survival Rate")

plt.show()

# Histogram of passenger ages

plt.figure(figsize=(6,4))

sns.histplot(df['age'], bins=30, kde=True, color="skyblue")

plt.title("Distribution of Passenger Ages")

plt.xlabel("Age")

plt.ylabel("Count")

plt.show()

# Bar chart: Survival by Age Group

plt.figure(figsize=(6,4))

sns.barplot(x="age\_group", y="survived", data=df, palette="Set2")

plt.title("Survival Rate by Age Group")

plt.ylabel("Survival Rate")

plt.show()

OUTPUT:





