# Task 3 - Mini EDA on Titanic Dataset

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

# Import libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# Load Titanic dataset (from seaborn for demo, in practice use Kaggle CSV)

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

# Preview data

print("First 5 rows of dataset:")

print(titanic.head())

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

# 1. Data Cleaning

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

# Fill missing Age with mean

titanic['age'].fillna(titanic['age'].mean(), inplace=True)

# Drop irrelevant columns

titanic.drop(['deck'], axis=1, inplace=True)   # deck ~ cabin

titanic.drop(['alive', 'class', 'embark\_town', 'adult\_male', 'who'], axis=1, inplace=True)

# Fill Embarked missing values with mode

titanic['embarked'].fillna(titanic['embarked'].mode()[0], inplace=True)

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

# 2. Feature Engineering

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

# Create Age Groups

bins = [0, 12, 18, 35, 50, 80]

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

titanic['age\_group'] = pd.cut(titanic['age'], bins=bins, labels=labels)

# Family Size (sibsp + parch)

titanic['family\_size'] = titanic['sibsp'] + titanic['parch']

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

# 3. Analysis

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

# Survival rate by Age Group

age\_survival = titanic.groupby('age\_group')['survived'].mean()

print("\nSurvival Rate by Age Group:\n", age\_survival)

# Survival rate by Embarkation Port

embark\_survival = titanic.groupby('embarked')['survived'].mean()

print("\nSurvival Rate by Embarked Port:\n", embark\_survival)

# Survival rate by Family size

family\_survival = titanic.groupby('family\_size')['survived'].mean()

print("\nSurvival Rate by Family Size:\n", family\_survival)

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

# 4. Visualizations

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

# Age distribution

plt.figure(figsize=(8,5))

sns.histplot(titanic['age'], bins=30, kde=True)

plt.title("Age Distribution of Titanic Passengers")

plt.show()

# Heatmap (numeric columns only)

plt.figure(figsize=(8,6))

numeric\_data = titanic.select\_dtypes(include=[np.number])

sns.heatmap(numeric\_data.corr(), annot=True, cmap="coolwarm")

plt.title("Correlation Heatmap")

plt.show()

# Survival by Family Size

plt.figure(figsize=(8,5))

sns.barplot(x='family\_size', y='survived', data=titanic, ci=None, palette="Set2")

plt.title("Survival Rate by Family Size")

plt.show()

# Survival by Age Group

plt.figure(figsize=(8,5))

sns.barplot(x='age\_group', y='survived', data=titanic, ci=None, palette="muted")

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

plt.show()

# Survival by Embarkation Port

plt.figure(figsize=(8,5))

sns.barplot(x='embarked', y='survived', data=titanic, ci=None, palette="pastel")

plt.title("Survival Rate by Embarkation Port")

plt.show()









