**project 4: Olympics Data Analysis**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# Load Excel file

df = pd.read\_excel('Summer-Olympic-medals-1976-to-2008.xlsx')

# Clean column names (remove extra spaces if any)

df.columns = [col.strip() for col in df.columns]

# Drop duplicates and rows where all fields are empty, fix Year dtype

df = df.dropna(how='all')

df['Year'] = df['Year'].astype(int)

# -- Step 1: Data Preparation --

print(df.head())

print(df.info())

print(df.describe())

# -- Step 2: Data Cleaning --

print(df.isnull().sum())

df\_cleaned = df.dropna()

print(df\_cleaned.info())

# -- Step 3: Exploratory Data Analysis (EDA) --

# Q1: Which city hosted max no. of Olympics (for your period)?

unique\_olympics = df\_cleaned[['City', 'Year']].drop\_duplicates('Year')

print(unique\_olympics)

# Q2: City hosting most events

events\_by\_city = df\_cleaned['City'].value\_counts()

plt.figure(figsize=(12,4))

events\_by\_city.plot(kind='bar', color='orange')

plt.title('Number of Events Hosted by Each City')

plt.xlabel('City')

plt.ylabel('Event Count')

plt.show()

# Q3: Type and number of unique events

unique\_events = df\_cleaned[['Sport', 'Discipline', 'Event']].drop\_duplicates()

print("Unique events held:", len(unique\_events))

events\_per\_sport = unique\_events['Sport'].value\_counts()

plt.figure(figsize=(14,5))

events\_per\_sport.plot(kind='bar', color='purple')

plt.title('Number of Unique Events Per Sport')

plt.ylabel('Event Count')

plt.show()

# Q4: Athletes with most medals

athlete\_medals = df\_cleaned.groupby('Athlete')['Medal'].count().sort\_values(ascending=False).head(10)

plt.figure(figsize=(12,6))

athlete\_medals.plot(kind='bar', color='silver')

plt.title('Top 10 Athletes by Medal Count')

plt.ylabel('Medals Won')

plt.show()

# Q5: Gender ratio among winners

gender\_counts = df\_cleaned['Gender'].value\_counts()

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

gender\_counts.plot(kind='pie', autopct='%1.1f%%', explode=[0.05, 0])

plt.title('Gender Distribution among Medal Winners')

plt.ylabel('')

plt.show()

# Q6: Country medals per year (by athlete medals, not just event medals)

country\_year = df\_cleaned.groupby(['Year', 'Country', 'Medal'])['Country'].count().reset\_index(name='Count')

country\_year['Medal'] = pd.Categorical(country\_year['Medal'], categories=['Gold','Silver','Bronze'], ordered=True)

pivot\_table = country\_year.pivot\_table(index=['Year', 'Country'], columns='Medal', values='Count', fill\_value=0)

pivot\_table['Total'] = pivot\_table.sum(axis=1)

country\_medal\_sum = pivot\_table.groupby('Country')['Total'].sum().sort\_values(ascending=False)

print(country\_medal\_sum.head(10))

plt.figure(figsize=(12,5))

country\_medal\_sum.head(10).plot(kind='bar', color='gold')

plt.title('Top 10 Countries by Total Medals (1976–2008)')

plt.ylabel('Total Medals')

plt.show()

# Q7: Country dominance by sport

dominance = df\_cleaned.groupby(['Sport', 'Country'])['Medal'].count().reset\_index(name='Count')

for sport in dominance['Sport'].unique():

print(f"{sport}:")

print(dominance[dominance['Sport']==sport].sort\_values('Count', ascending=False).head(3))

# Q8: Athletes winning medals in more than one sport

athlete\_sports = df\_cleaned[['Athlete','Sport']].drop\_duplicates()

multi\_sport\_athletes = athlete\_sports.groupby('Athlete').Sport.nunique()

print(multi\_sport\_athletes[multi\_sport\_athletes > 1])

# Q9: Top 5 countries per Olympic year (by athlete medal count)

top\_by\_year = pivot\_table.groupby('Year')['Total'].nlargest(5)

print(top\_by\_year)

# -- Additional: Medals Trend Over Time --

medals\_by\_year = df\_cleaned.groupby('Year')['Medal'].count()

plt.figure(figsize=(10,6))

plt.plot(medals\_by\_year.index, medals\_by\_year.values, marker='o')

plt.title('Total Medals Won Over the Years')

plt.xlabel('Year')

plt.ylabel('Total Medals')

plt.grid(True)

plt.show()