import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# Step 1: Load the dataset

# For demonstration, let's create a synthetic dataset

data = {

'Region': np.random.choice(['North', 'South', 'East', 'West'], 100),

'Marketing\_Spend': np.random.randint(1000, 5000, 100),

'Sales': np.random.randint(5000, 20000, 100),

'Effectiveness': np.random.uniform(0.5, 1.0, 100),

'Age\_Group': np.random.choice(['18-25', '26-35', '36-45', '46-55', '56+'], 100),

'Trial\_Period': np.random.choice(['Q1', 'Q2', 'Q3', 'Q4'], 100)

}

df = pd.DataFrame(data)

# Step 2: Data Cleaning

# Check for missing values

print("Missing values in each column:")

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

# Check for duplicates

duplicates = df.duplicated().sum()

print(f"\nNumber of duplicate rows: {duplicates}")

# Remove duplicates if any

df = df.drop\_duplicates()

# Step 3: Create Visualizations

# A. Bar plot showing total sales per region

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

sns.barplot(x='Region', y='Sales', data=df, estimator=sum, ci=None)

plt.title('Total Sales per Region')

plt.ylabel('Total Sales')

plt.xlabel('Region')

plt.show()

# B. Scatter plot to visualize the relationship between Marketing\_Spend and Sales

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

sns.scatterplot(x='Marketing\_Spend', y='Sales', data=df)

plt.title('Relationship between Marketing Spend and Sales')

plt.xlabel('Marketing Spend')

plt.ylabel('Sales')

plt.show()

# C. Boxplot comparing drug effectiveness across different age groups

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

sns.boxplot(x='Age\_Group', y='Effectiveness', data=df)

plt.title('Drug Effectiveness across Age Groups')

plt.xlabel('Age Group')

plt.ylabel('Effectiveness')

plt.show()

# D. Line plot showing the sales trend for each product over different trial periods

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

sns.lineplot(x='Trial\_Period', y='Sales', data=df, estimator='mean', ci=None, hue='Region')

plt.title('Sales Trend for Each Region over Trial Periods')

plt.xlabel('Trial Period')

plt.ylabel('Average Sales')

plt.show()

# E. Heatmap of the correlation between Sales, Marketing\_Spend, and Effectiveness

correlation\_matrix = df[['Sales', 'Marketing\_Spend', 'Effectiveness']].corr()

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

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', fmt=".2f")

plt.title('Correlation Heatmap')

plt.show()