EXPERIMENT 1

import pandas as pd

import matplotlib.pyplot as plt

from scipy import stats

file\_path = 'C:\\Users\\Lenovo\\Downloads\\1.csv'

data = pd.read\_csv(file\_path)

print("Initial Data Preview:")

print(data.head())

missing\_values = data.isnull().sum()

print("\nMissing values:")

print(missing\_values)

data['price'].fillna(data['price'].median(), inplace=True)

data['month'] = pd.to\_datetime(data['month'], format='%d-%b')

data['month'] = data['month'].apply(lambda x: x.replace(year=2022))

data.sort\_values('month', inplace=True)

z\_scores = stats.zscore(data['price'])

outliers = (z\_scores > 3) | (z\_scores < -3)

data = data[~outliers]

data.set\_index('month', inplace=True)

print("\nSummary Statistics of Cleaned Data:")

print(data.describe())

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

plt.plot(data.index, data['price'], marker='o', linestyle='-', color='b', label='Shampoo Sales')

plt.title('Shampoo Sales Over 3 Years')

plt.xlabel('Month')

plt.ylabel('Price')

plt.legend()

plt.grid(True)

plt.show()

print("\nCleaned Data Preview:")

print(data.head())