Code:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

from sklearn.decomposition import PCA

file\_path = "/content/Analytics Challenge Data.xlsx"

df = pd.read\_excel(file\_path)

print("Data Loaded Successfully!\n")

print(df.head())

print("\n Dataset Info:")

print(df.info())

print("\n Summary Statistics:")

print(df.describe())

print("\n Missing Values:")

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

df = df.dropna()

scaler = StandardScaler()

scaled\_data = scaler.fit\_transform(df.select\_dtypes(include=np.number))

print("\n Data Scaled Successfully")

inertia = []

K = range(1, 11)

for k in K:

    kmeans = KMeans(n\_clusters=k, random\_state=42)

    kmeans.fit(scaled\_data)

    inertia.append(kmeans.inertia\_)

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

plt.plot(K, inertia, 'bo-')

plt.xlabel('Number of Clusters (k)')

plt.ylabel('Inertia')

plt.title('Elbow Method for Optimal k')

plt.grid(True)

plt.show()

optimal\_k = 4

kmeans = KMeans(n\_clusters=optimal\_k, random\_state=42)

df['Cluster'] = kmeans.fit\_predict(scaled\_data)

print("\n K-Means Clustering Applied Successfully!")

print(df['Cluster'].value\_counts())

pca = PCA(n\_components=2)

pca\_result = pca.fit\_transform(scaled\_data)

df['PCA1'] = pca\_result[:, 0]

df['PCA2'] = pca\_result[:, 1]

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

sns.scatterplot(data=df, x='PCA1', y='PCA2', hue='Cluster', palette='viridis', s=80)

plt.title('Customer Segments (PCA Projection)')

plt.show()

numeric\_cols = df.select\_dtypes(include=np.number).columns

cluster\_summary = df.groupby('Cluster')[numeric\_cols].mean()

print("\n Cluster Summary (Average Values per Cluster):")

print(cluster\_summary)

print(" - Low engagement users. Retarget with ads or recommendation emails to increase activity.")

print(" - Columns differ in this dataset; review numeric features for marketing insights.")

output:

/usr/local/lib/python3.12/dist-packages/openpyxl/worksheet/\_reader.py:329: UserWarning: Unknown extension is not supported and will be removed

warn(msg)

✅ Data Loaded Successfully!

day site new\_customer platform visits distinct\_sessions \

0 2013-01-01 Acme 1.0 Android 24 16

1 2013-01-01 Acme 1.0 BlackBerry 0 0

2 2013-01-01 Sortly 1.0 iPad 0 0

3 2013-01-01 Acme 1.0 Windows 922 520

4 2013-01-01 Botly 1.0 Android 11 10

orders gross\_sales bounces add\_to\_cart product\_page\_views \

0 14 1287.0 4 16 104

1 0 13.0 0 0 1

2 0 98.0 0 0 0

3 527 60753.0 149 610 3914

4 11 1090.0 0 11 4

search\_page\_views

0 192

1 0

2 0

3 7367

4 19

Dataset Info:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 21061 entries, 0 to 21060

Data columns (total 12 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 day 21061 non-null datetime64[ns]

1 site 21061 non-null object

2 new\_customer 12802 non-null float64

3 platform 20651 non-null object

4 visits 21061 non-null int64

5 distinct\_sessions 21061 non-null int64

6 orders 21061 non-null int64

7 gross\_sales 11485 non-null float64

8 bounces 21061 non-null int64

9 add\_to\_cart 21061 non-null int64

10 product\_page\_views 21061 non-null int64

11 search\_page\_views 21061 non-null int64

dtypes: datetime64[ns](1), float64(2), int64(7), object(2)

memory usage: 1.9+ MB

None

Summary Statistics:

day new\_customer visits \

count 21061 12802.000000 21061.000000

mean 2013-07-30 13:23:22.839371264 0.448055 1934.708039

min 2013-01-01 00:00:00 0.000000 0.000000

25% 2013-06-10 00:00:00 0.000000 3.000000

50% 2013-08-21 00:00:00 0.000000 24.000000

75% 2013-10-27 00:00:00 1.000000 360.000000

max 2013-12-31 00:00:00 1.000000 136057.000000

std NaN 0.497314 7448.607191

distinct\_sessions orders gross\_sales bounces \

count 21061.000000 21061.000000 11485.000000 21061.000000

mean 1515.205024 62.378994 16473.395821 743.282085

min 0.000000 0.000000 1.000000 0.000000

25% 2.000000 0.000000 79.000000 0.000000

50% 19.000000 0.000000 851.000000 5.000000

75% 274.000000 7.000000 3145.000000 97.000000

max 107104.000000 4916.000000 707642.000000 54512.000000

std 5925.833287 260.279286 51111.354605 3154.697787

add\_to\_cart product\_page\_views search\_page\_views

count 21061.000000 21061.000000 21061.000000

mean 166.250890 4358.198234 8584.187788

min 0.000000 0.000000 0.000000

25% 0.000000 3.000000 4.000000

50% 4.000000 53.000000 82.000000

75% 43.000000 708.000000 1229.000000

max 7924.000000 187601.000000 506629.000000

std 505.186834 14327.287354 31120.321365

Missing Values:

day 0

site 0

new\_customer 8259

platform 410

visits 0

distinct\_sessions 0

orders 0

gross\_sales 9576

bounces 0

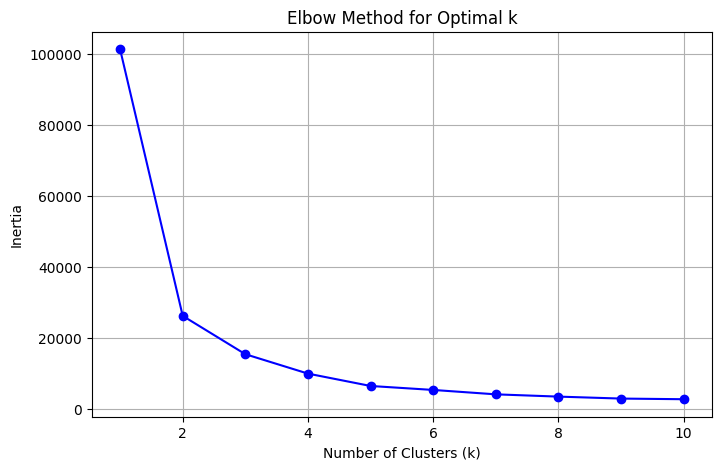
add\_to\_cart 0

product\_page\_views 0

search\_page\_views 0

dtype: int64

Data Scaled Successfully



K-Means Clustering Applied Successfully!

Cluster

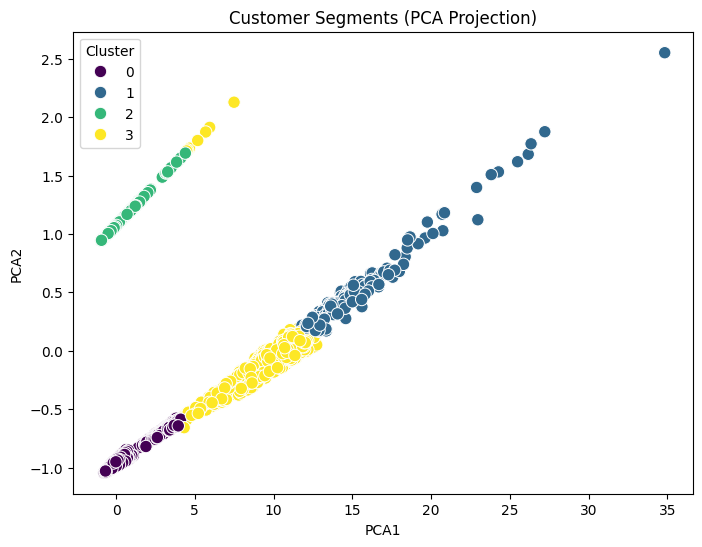
2 5432

0 5229

3 397

1 194

Name: count, dtype: int64



Cluster Summary (Average Values per Cluster):

new\_customer visits distinct\_sessions orders \

Cluster

0 0.00000 220.201759 167.498566 44.716963

1 0.00000 11896.546392 8187.876289 2035.886598

2 1.00000 76.961156 47.486561 46.419735

3 0.02267 7124.463476 4764.136020 1087.765743

gross\_sales bounces add\_to\_cart product\_page\_views \

Cluster

0 6727.618474 46.926946 71.631287 873.425320

1 299434.427835 2090.180412 3372.314433 40468.087629

2 5213.668630 12.102541 52.111561 292.613218

3 170222.984887 1917.596977 1800.350126 23598.322418

search\_page\_views Cluster PCA1 PCA2

Cluster

0 1234.582138 0.0 -0.434289 -1.012323

1 83221.201031 1.0 15.246718 0.516448

2 500.243925 2.0 -0.772183 0.966299

3 47388.052897 3.0 8.835088 -0.140282

- Low engagement users. Retarget with ads or recommendation emails to increase activity.

- Columns differ in this dataset; review numeric features for marketing insights.