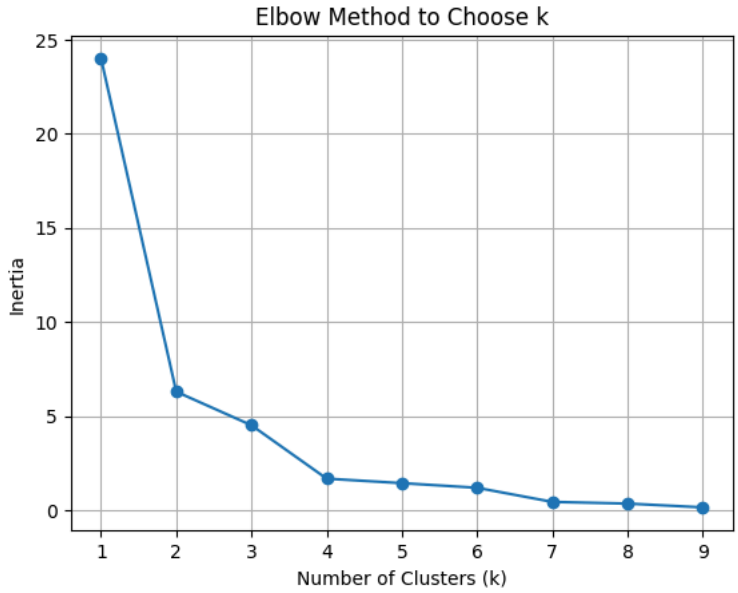
Activity 3: Patterns with K-means

Emiliano Torres Sandoval A01666136 Alejandro Salazar Loza A01665123

```
# Activity 3: Patterns with K-means
# Emiliano Torres Sandoval A01666136, Alejandro Salazar Loza A01665123

Repo https://github.com/ET771/arte_analitica

import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from scipy.spatial.distance import cdist

ronaldo = pd.read_csv('ronaldo_competition_goals.csv')
messi = pd.read_csv('messi_competition_goals.csv')

ronaldo.rename(columns={'Caps': 'Appearances'}, inplace=True)
messi.rename(columns={'Apps': 'Appearances'}, inplace=True)

ronaldo['Player'] = 'Ronaldo'
messi['Player'] = 'Messi'

df = pd.concat([ronaldo, messi], ignore_index=True)

features = df[['Appearances', 'Goals']]

scaler = StandardScaler()
scaled = scaler.fit_transform(features)

inertia = []
for k in range(1, 10):
    km = KMeans(n_clusters=k, random_state=42)
    km.fit(scaled)
    inertia.append(km.inertia_)

plt.plot(range(1, 10), inertia, marker='o')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Inertia')
plt.title('Elbow Method to Choose k')
plt.grid(True)
plt.show()

k = 3
kmeans = KMeans(n_clusters=k, random_state=42)
df['Cluster'] = kmeans.fit_predict(scaled)

centroids_scaled = kmeans.cluster_centers_
centroids_original = scaler.inverse_transform(centroids_scaled)
centroids_df = pd.DataFrame(centroids_original, columns=['Appearances', 'Goals'])

print("Centroids:\n", centroids_df)

distances = cdist(centroids_original, centroids_original)
print("\nDistances Between Cluster Centers:\n", distances)
```




```
Centroids:
    Appearances      Goals
0    46.666667  40.333333
1    17.571429   8.142857
2    56.000000  24.000000

Distances Between Cluster Centers:
 [[ 0.         43.39077825 18.81193475]
 [43.39077825  0.         41.57167403]
```

1. Do you think the centers are representative of the data? Why? Yes, I think the centers make sense because they group competitions where Messi and Ronaldo had similar performances. For example, some tournaments had a lot of games and goals, and others had fewer. The clusters help show those patterns clearly.
2. How did you get the k value? I used the elbow method. I made a graph of inertia and looked where the line starts to flatten. That happened at k = 3, so I chose that because it gives a good balance.
3. Would the centers be better with a higher or lower k? If I use a higher k, I might get more detail but it can be too complicated. A lower k is too simple and may not show important differences. So, I think k = 3 is a good number to use.
4. How far apart are the centers? Are some close? Two centers were kind of close, and one was farther away. That makes sense because some tournaments are more similar than others. One cluster had really different values, so it's more separated.
5. What happens if there are outliers (extreme values)? Outliers can move the center and make it less accurate. Like if a player scored a lot in one small tournament, it could pull the center away. So, it's a good idea to check for outliers first.
6. What can you say about the data from the centers? The clusters show us three main types of competitions:

Some with lots of games and goals

Some with low participation

And others that are in the middle It helps us see how their performance changes depending on the competition.