Program 9: Grouping objects by similarity using k-means

Problem Statement:

The goal of this analysis is to use K-Means clustering on the Wine dataset and evaluate how well it identifies different wine types. The process includes standardizing the features of the data and applying K-Means clustering with k = 3, which matches the number of wine types. The quality of the clustering is then measured using three metrics: Completeness Score (to see how well data points of the same wine type are grouped together), Silhouette Coefficient (to assess how distinct and well-separated the clusters are), and Calinski-Harabasz Index (to evaluate the separation and compactness of the clusters). These metrics help determine if k = 3 is an effective choice for clustering this dataset.

```
import pandas as pd

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

from sklearn.datasets import load_wine

from sklearn.metrics import completeness_score, silhouette_score, calinski_harabasz_score


# Load the Wine dataset

wine = load_wine()


# Extract features and target

X = pd.DataFrame(wine.data, columns=wine.feature_names)

y = wine.target


# Standardize the features

scaler = StandardScaler()

X_scaled = scaler.fit_transform(X)


# Set the number of clusters

k = 3


# Initialize KMeans with n_init explicitly set

kmeans = KMeans(n_clusters=k, n_init=10, random_state=42)
```

```
# Fit the model
kmeans.fit(X_scaled)

# Get cluster centers and labels
centroids = kmeans.cluster_centers_
labels = kmeans.labels_

# Calculate evaluation metrics
completeness = completeness_score(y, labels)  # Completeness Score
silhouette_avg = silhouette_score(X_scaled, labels)  # Silhouette Coefficient
calinski_harabasz = calinski_harabasz_score(X_scaled, labels)  # Calinski-Harabasz Index

# Print specific evaluation metrics
print(f'Silhouette Coefficient: {silhouette_avg:.2f}')
print(f'Calinski-Harabasz Index: {calinski_harabasz:.2f}')
print(f'Completeness: {completeness:.2f}')
```