Experiment No. 9

Title: Unsupervised Learning: Implement K-Means Clustering and Hierarchical clustering on proper data set of your choice. Compare their Convergence

Objectives:

1. Implement K-Means Clustering and Hierarchical Clustering on a chosen dataset.
2. Compare the convergence behavior of both clustering algorithms.
3. Evaluate clustering performance using appropriate metrics.
4. Visualize clusters and analyze computational efficiency.

Problem Statement:

This project aims to implement, compare, and analyze both algorithms(K means clustering and Hierarchical Clustering) to understand how they converge and under what conditions one outperforms the other.

Outcomes:

Implementation of K-Means and Hierarchical Clustering on a real-world dataset.
Visual comparisons of clusters formed by both methods.
Evaluation of convergence speed, SSE (Sum of Squared Errors), Silhouette Score, and other metrics.

Tools Required: 4GB RAM, Anaconda, Notebook

Theory:

K-Means Clustering

- Algorithm:
  1. Choose the number of clusters, K.
  2. Randomly initialize K centroids.
  3. Assign each data point to the nearest centroid.
  4. Recalculate centroids by averaging the assigned points.
  5. Repeat until convergence (centroids don't change or max iterations reached).
- Convergence: Converges when centroids stabilize. Affected by initial centroid selection.
- Computational Complexity: O(*nkt*) (n = points, k = clusters, t = iterations).

Hierarchical Clustering

- Types:
  1. Agglomerative (Bottom-Up):
     - Start with individual points, merge closest pairs iteratively.
  2. Divisive (Top-Down):
     - Start with all points in one cluster, split recursively.
- Convergence: Defined by a dendrogram; doesn't require pre-defined K.

- Computational Complexity: O(n² log n) (Higher than K-Means).

Algorithm:

Step 1: Load Dataset

- Choose a dataset and preprocess it.

Step 2: Implement K-Means Clustering

- Use K-Means from sklearn.cluster with different K values.
- Track inertia (SSE) to check convergence.

Step 3: Implement Hierarchical Clustering

- Apply Agglomerative Clustering using scipy.cluster.hierarchy.
- Visualize dendrogram to determine the best number of clusters.

Step 4: Compare Convergence

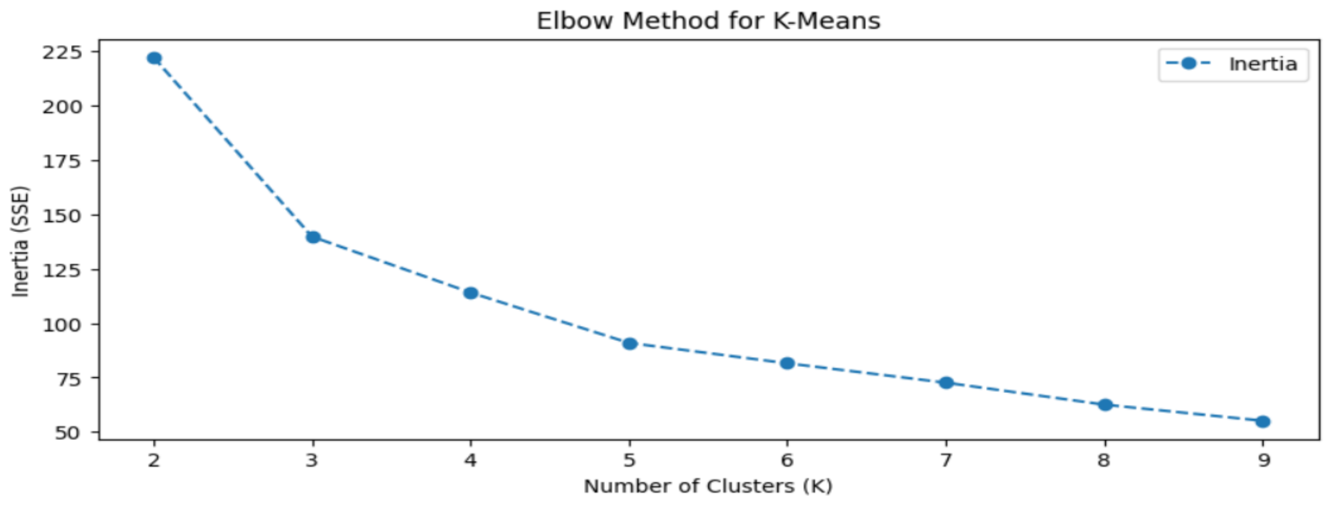
- Compare SSE, silhouette scores, and runtime.
- Plot elbow method, cluster assignment, and dendrograms.

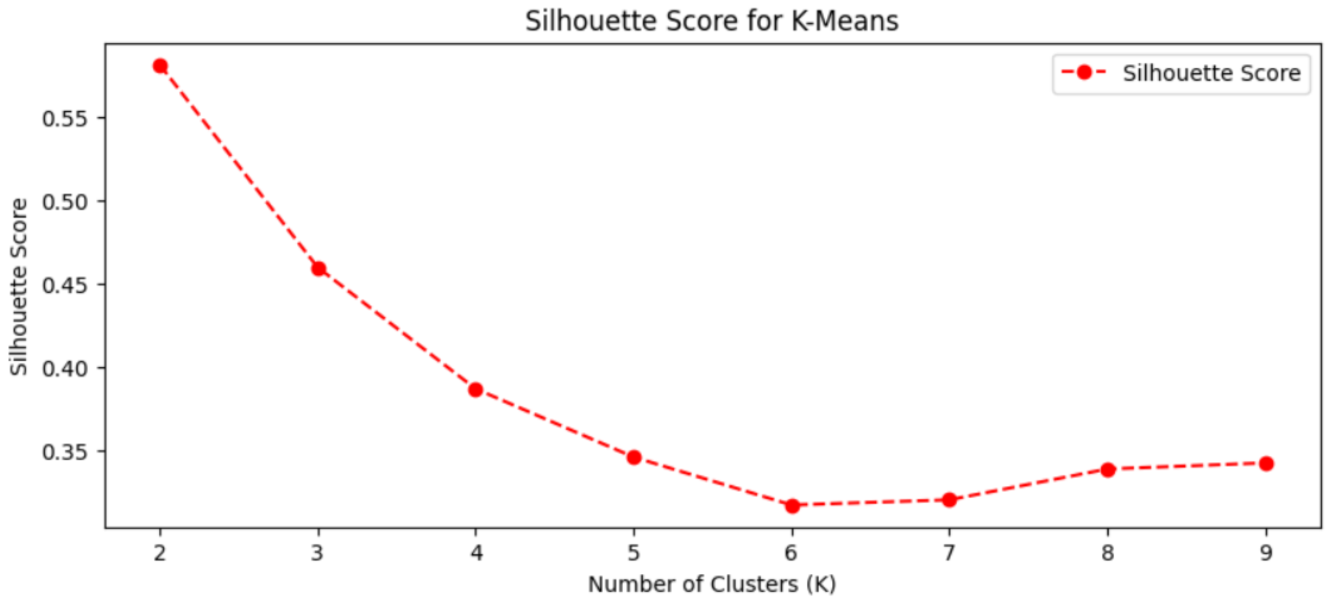
Source Code:

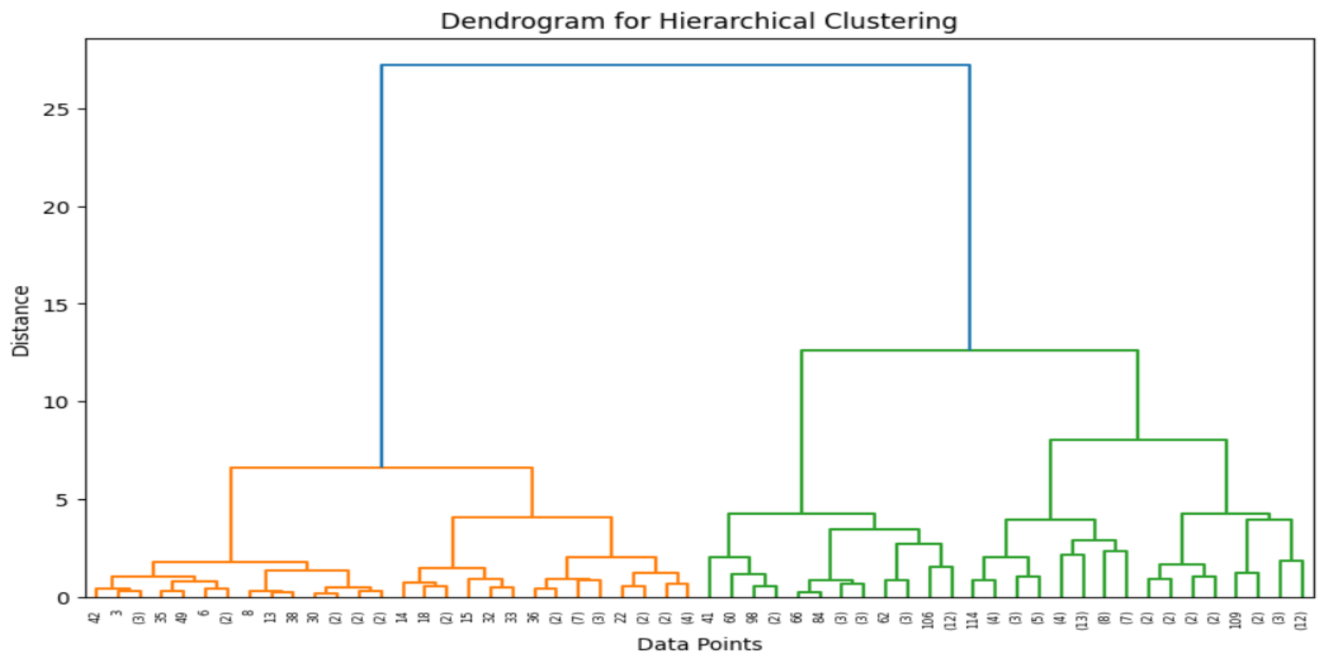
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans, AgglomerativeClustering
from scipy.cluster.hierarchy import dendrogram, linkage
from sklearn.metrics import silhouette_score
from sklearn.datasets import load_iris
from sklearn.preprocessing import StandardScaler
iris = load_iris()
X = iris.data
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
inertia = []
silhouette_scores = []
K_range = range(2, 10)
for k in K_range:
    kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
    kmeans.fit(X_scaled)
    inertia.append(kmeans.inertia_)
    silhouette_scores.append(silhouette_score(X_scaled, kmeans.labels_))
plt.figure(figsize=(10, 4))
plt.plot(K_range, inertia, marker='o', linestyle='--', label='Inertia')
plt.xlabel('Number of Clusters (K)')
plt.ylabel('Inertia (SSE)')
plt.title('Elbow Method for K-Means')
plt.legend()
plt.show()
plt.figure(figsize=(10, 4))
plt.plot(K_range, silhouette_scores, marker='o', linestyle='--', color='red', label='Silhouette
  Score')
```

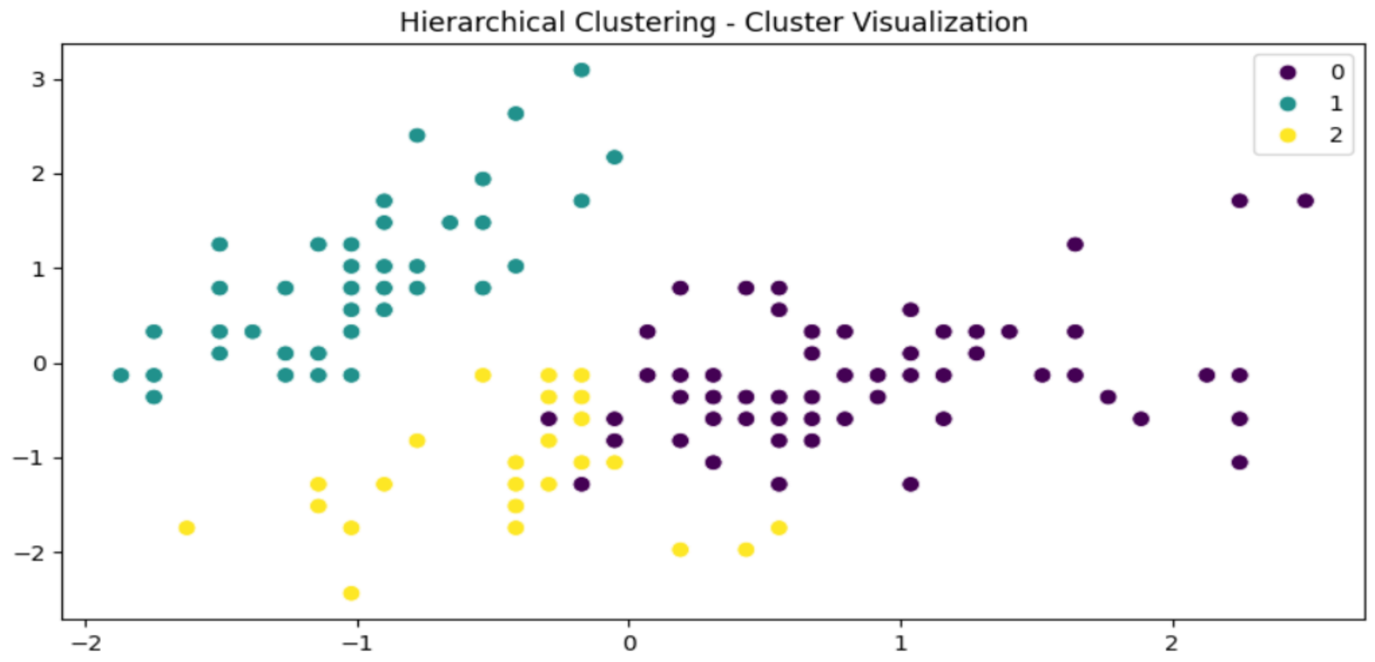
```
plt.xlabel('Number of Clusters (K)')
plt.ylabel('Silhouette Score')
plt.title('Silhouette Score for K-Means')
plt.legend()
plt.show()
plt.figure(figsize=(10, 6))
linked = linkage(X_scaled, method='ward')
dendrogram(linked, truncate_mode='level', p=5)
plt.title('Dendrogram for Hierarchical Clustering')
plt.xlabel('Data Points')
plt.ylabel('Distance')
plt.show()
hc = AgglomerativeClustering(n_clusters=3, linkage='ward')
hc_labels = hc.fit_predict(X_scaled)
plt.figure(figsize=(10, 5))
sns.scatterplot(x=X_scaled[:, 0], y=X_scaled[:, 1], hue=hc_labels, palette='viridis', s=60)
plt.title("Hierarchical Clustering - Cluster Visualization")
plt.show()
```

Output:






Dendrogram for Hierarchical Clustering
Distance
Data Points


Hierarchical Clustering - Cluster Visualization
0
1
2

Conclusion:

____

____

____

____