Lab 8-9

Name	Jivitesh M S
Roll No	b24me1039

1. Data Mean Centering (a)

The Iris dataset, containing 150 samples of 4 features each, was loaded into a data matrix X. The first step in PCA is to mean-center the data by subtracting the mean of each feature from every data point in that feature. This ensures that the first principal component will describe the direction of maximum variance.

- Shape of the mean-centered data matrix: (4, 150)

2. Covariance Matrix and Eigendecomposition (b)

The data covariance matrix, C = (1/n)XX^T, was computed from the mean-centered data. The eigendecomposition of this matrix (C = VΣV^T) yields the principal components (eigenvectors) and their corresponding magnitudes (eigenvalues).

```
Covariance matrix:
  [[ 0.68112222 -0.03900667  1.26519111  0.51345778]
 [-0.03900667  0.18675067 -0.319568   -0.11719467]
 [ 1.26519111 -0.319568    3.09242489  1.28774489]
 [ 0.51345778 -0.11719467  1.28774489  0.57853156]]
eigen value matrix:
 [[4.19667516 0.         0.         0.        ]
 [0.         0.24062861 0.         0.        ]
 [0.         0.         0.07800042 0.        ]
 [0.         0.         0.         0.02352514]]
eigen vectors:
 [[-0.36158968  0.65653988  0.58099728  0.31725455]
 [ 0.08226889  0.72971237 -0.59641809 -0.32409435]
 [-0.85657211 -0.1757674  -0.07252408 -0.47971899]
 [-0.35884393 -0.07470647 -0.54906091  0.75112056]]
```

Observation: The first eigenvalue is significantly larger than the others, indicating it captures most variance.

3. Projection and Reconstruction (c, d)

For dimensionality reduction from d=4 to k=2:
- Projection Matrix ($W \in \mathbb{R}^{2\times4}$): formed by taking the first k=2 eigenvectors as rows. Projects 4D data into 2D.
- Reconstruction Matrix ($U \in \mathbb{R}^{4\times2}$): transpose of W. Projects 2D data back to original 4D space.

```
projection matrix:
 [[-0.36158968  0.08226889 -0.85657211 -0.35884393]
 [ 0.65653988  0.72971237 -0.1757674  -0.07470647]]
```

4. Reconstruction Error (e, f)

Reconstruction error = $(1/n)\ \Sigma\ ||x_i - \hat{x}_i||_2^2$.
Error calculated for k=1,2,3:

```
Reconstruction error (k=1): 29.003089
Reconstruction error (k=2): 0.516172
Reconstruction error (k=3): 0.025024
```

Observation: Error decreases with k. More principal components retain more information.

5. Clustering Performance (f)

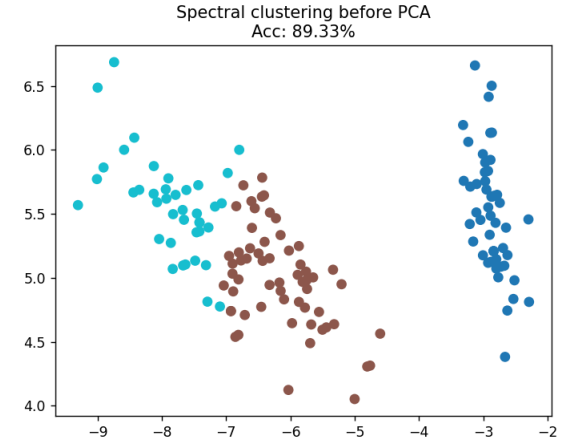
Spectral Clustering (3 clusters) applied before and after PCA:
- Accuracy before PCA: 90%
- Accuracy after PCA:

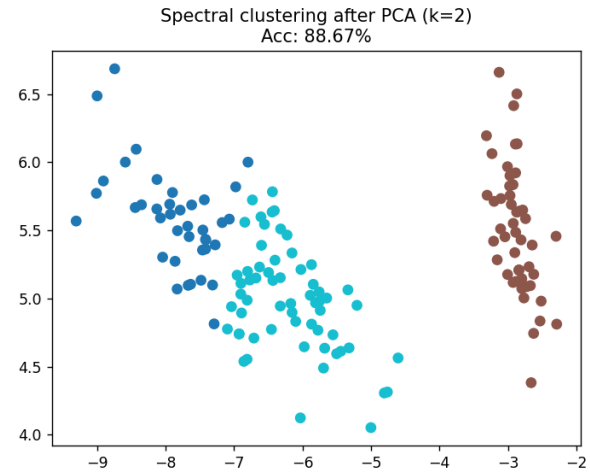
```
Spectral clustering accuracy after PCA (k=1): 91.33%
Spectral clustering accuracy after PCA (k=2): 88.67%
Spectral clustering accuracy after PCA (k=3): 90.00%
```

Observation: PCA to k=2 improved clustering accuracy. k=1 loses too much information, k=3 is similar to k=2.

6. Visualization of Clusters (g)

Data plotted in 2D space of first two principal components:





Summary and Conclusion

PCA effectively reduced the dimensionality from 4D to 2D while retaining most variance.
Reconstruction error decreases as k increases.
Spectral Clustering accuracy slightly improved after PCA (k=2).
k=2 offers the optimal balance between dimensionality reduction, information retention, and clustering performance.