Hierarchical Clustering Implementation

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn import datasets
```

In [3]:
```
# Import IRIS Dataset
iris = datasets.load_iris()
```

In [4]:
```
iris_data = pd.DataFrame(iris.data)
```

In [7]:
```
iris_data.columns = iris.feature_names
```

In [8]:
```
iris_data
```

Out[8]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
...	...	...	...	...
145	6.7	3.0	5.2	2.3
146	6.3	2.5	5.0	1.9
147	6.5	3.0	5.2	2.0
148	6.2	3.4	5.4	2.3
149	5.9	3.0	5.1	1.8

150 rows × 4 columns

In [9]:
```
# Standard Scaler
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
```

In [10]:
```
X_scaled = scaler.fit_transform(iris_data)
```

In [11]:
```
X_scaled
```

Out[11]:
```
array([[-9.00681170e-01,  1.01900435e+00, -1.34022653e+00,
        -1.31544430e+00],
       [-1.14301691e+00, -1.31979479e-01, -1.34022653e+00,
        -1.31544430e+00],
       [-1.38535265e+00,  3.28414053e-01, -1.39706395e+00,
        -1.31544430e+00],
       [-1.50652052e+00,  9.82172869e-02, -1.28338910e+00,
        -1.31544430e+00],
       [-1.02184904e+00,  1.24920112e+00, -1.34022653e+00,
        -1.31544430e+00],
       [-5.37177559e-01,  1.93979142e+00, -1.16971425e+00,
        -1.05217993e+00],
       [-1.50652052e+00,  7.88807586e-01, -1.34022653e+00,
        -1.18381211e+00],
       [-1.02184904e+00,  7.88807586e-01, -1.28338910e+00,
        -1.31544430e+00],
       [-1.74885626e+00, -3.62176246e-01, -1.34022653e+00,
        -1.31544430e+00],
       [-1.14301691e+00,  9.82172869e-02, -1.28338910e+00,
        -1.44707648e+00],
       [-5.37177559e-01,  1.47939788e+00, -1.28338910e+00,
        -1.31544430e+00],
       [-1.26418478e+00,  7.88807586e-01, -1.22655167e+00,
        -1.31544430e+00],
       [-1.26418478e+00, -1.31979479e-01, -1.34022653e+00,
        -1.44707648e+00],
       [-1.87002413e+00, -1.31979479e-01, -1.51073881e+00,
        -1.44707648e+00],
       [-5.25060772e-02,  2.16998818e+00, -1.45390138e+00,
        -1.31544430e+00],
       [-1.73673948e-01,  3.09077525e+00, -1.28338910e+00,
        -1.05217993e+00],
       [-5.37177559e-01,  1.93979142e+00, -1.39706395e+00,
        -1.05217993e+00],
       [-9.00681170e-01,  1.01900435e+00, -1.34022653e+00,
        -1.18381211e+00],
       [-1.73673948e-01,  1.70959465e+00, -1.16971425e+00,
        -1.18381211e+00],
       [-9.00681170e-01,  1.70959465e+00, -1.28338910e+00,
        -1.18381211e+00],
       [-5.37177559e-01,  7.88807586e-01, -1.16971425e+00,
        -1.31544430e+00],
       [-9.00681170e-01,  1.47939788e+00, -1.28338910e+00,
        -1.05217993e+00],
       [-1.50652052e+00,  1.24920112e+00, -1.56757623e+00,
        -1.31544430e+00],
       [-9.00681170e-01,  5.58610819e-01, -1.16971425e+00,
        -9.20547742e-01],
       [-1.26418478e+00,  7.88807586e-01, -1.05603939e+00,
        -1.31544430e+00],
       [-1.02184904e+00, -1.31979479e-01, -1.22655167e+00,
        -1.31544430e+00],
       [-1.02184904e+00,  7.88807586e-01, -1.22655167e+00,
        -1.05217993e+00],
       [-7.79513300e-01,  1.01900435e+00, -1.28338910e+00,
        -1.31544430e+00],
       [-7.79513300e-01,  7.88807586e-01, -1.34022653e+00,
        -1.31544430e+00],
       [-1.38535265e+00,  3.28414053e-01, -1.22655167e+00,
        -1.31544430e+00],
```

```
[-1.26418478e+00,  9.82172869e-02, -1.22655167e+00,
 -1.31544430e+00],
[-5.37177559e-01,  7.88807586e-01, -1.28338910e+00,
 -1.05217993e+00],
[-7.79513300e-01,  2.40018495e+00, -1.28338910e+00,
 -1.44707648e+00],
[-4.16009689e-01,  2.63038172e+00, -1.34022653e+00,
 -1.31544430e+00],
[-1.14301691e+00,  9.82172869e-02, -1.28338910e+00,
 -1.31544430e+00],
[-1.02184904e+00,  3.28414053e-01, -1.45390138e+00,
 -1.31544430e+00],
[-4.16009689e-01,  1.01900435e+00, -1.39706395e+00,
 -1.31544430e+00],
[-1.14301691e+00,  1.24920112e+00, -1.34022653e+00,
 -1.44707648e+00],
[-1.74885626e+00, -1.31979479e-01, -1.39706395e+00,
 -1.31544430e+00],
[-9.00681170e-01,  7.88807586e-01, -1.28338910e+00,
 -1.31544430e+00],
[-1.02184904e+00,  1.01900435e+00, -1.39706395e+00,
 -1.18381211e+00],
[-1.62768839e+00, -1.74335684e+00, -1.39706395e+00,
 -1.18381211e+00],
[-1.74885626e+00,  3.28414053e-01, -1.39706395e+00,
 -1.31544430e+00],
[-1.02184904e+00,  1.01900435e+00, -1.22655167e+00,
 -7.88915558e-01],
[-9.00681170e-01,  1.70959465e+00, -1.05603939e+00,
 -1.05217993e+00],
[-1.26418478e+00, -1.31979479e-01, -1.34022653e+00,
 -1.18381211e+00],
[-9.00681170e-01,  1.70959465e+00, -1.22655167e+00,
 -1.31544430e+00],
[-1.50652052e+00,  3.28414053e-01, -1.34022653e+00,
 -1.31544430e+00],
[-6.58345429e-01,  1.47939788e+00, -1.28338910e+00,
 -1.31544430e+00],
[-1.02184904e+00,  5.58610819e-01, -1.34022653e+00,
 -1.31544430e+00],
[ 1.40150837e+00,  3.28414053e-01,  5.35408562e-01,
  2.64141916e-01],
[ 6.74501145e-01,  3.28414053e-01,  4.21733708e-01,
  3.95774101e-01],
[ 1.28034050e+00,  9.82172869e-02,  6.49083415e-01,
  3.95774101e-01],
[-4.16009689e-01, -1.74335684e+00,  1.37546573e-01,
  1.32509732e-01],
[ 7.95669016e-01, -5.92373012e-01,  4.78571135e-01,
  3.95774101e-01],
[-1.73673948e-01, -5.92373012e-01,  4.21733708e-01,
  1.32509732e-01],
[ 5.53333275e-01,  5.58610819e-01,  5.35408562e-01,
  5.27406285e-01],
[-1.14301691e+00, -1.51316008e+00, -2.60315415e-01,
 -2.62386821e-01],
[ 9.16836886e-01, -3.62176246e-01,  4.78571135e-01,
  1.32509732e-01],
[-7.79513300e-01, -8.22569778e-01,  8.07091462e-02,
  2.64141916e-01],
```

```
[-1.02184904e+00, -2.43394714e+00, -1.46640561e-01,
 -2.62386821e-01],
[ 6.86617933e-02, -1.31979479e-01,  2.51221427e-01,
  3.95774101e-01],
[ 1.89829664e-01, -1.97355361e+00,  1.37546573e-01,
 -2.62386821e-01],
[ 3.10997534e-01, -3.62176246e-01,  5.35408562e-01,
  2.64141916e-01],
[-2.94841818e-01, -3.62176246e-01, -8.98031345e-02,
  1.32509732e-01],
[ 1.03800476e+00,  9.82172869e-02,  3.64896281e-01,
  2.64141916e-01],
[-2.94841818e-01, -1.31979479e-01,  4.21733708e-01,
  3.95774101e-01],
[-5.25060772e-02, -8.22569778e-01,  1.94384000e-01,
 -2.62386821e-01],
[ 4.32165405e-01, -1.97355361e+00,  4.21733708e-01,
  3.95774101e-01],
[-2.94841818e-01, -1.28296331e+00,  8.07091462e-02,
 -1.30754636e-01],
[ 6.86617933e-02,  3.28414053e-01,  5.92245988e-01,
  7.90670654e-01],
[ 3.10997534e-01, -5.92373012e-01,  1.37546573e-01,
  1.32509732e-01],
[ 5.53333275e-01, -1.28296331e+00,  6.49083415e-01,
  3.95774101e-01],
[ 3.10997534e-01, -5.92373012e-01,  5.35408562e-01,
  8.77547895e-04],
[ 6.74501145e-01, -3.62176246e-01,  3.08058854e-01,
  1.32509732e-01],
[ 9.16836886e-01, -1.31979479e-01,  3.64896281e-01,
  2.64141916e-01],
[ 1.15917263e+00, -5.92373012e-01,  5.92245988e-01,
  2.64141916e-01],
[ 1.03800476e+00, -1.31979479e-01,  7.05920842e-01,
  6.59038469e-01],
[ 1.89829664e-01, -3.62176246e-01,  4.21733708e-01,
  3.95774101e-01],
[-1.73673948e-01, -1.05276654e+00, -1.46640561e-01,
 -2.62386821e-01],
[-4.16009689e-01, -1.51316008e+00,  2.38717193e-02,
 -1.30754636e-01],
[-4.16009689e-01, -1.51316008e+00, -3.29657076e-02,
 -2.62386821e-01],
[-5.25060772e-02, -8.22569778e-01,  8.07091462e-02,
  8.77547895e-04],
[ 1.89829664e-01, -8.22569778e-01,  7.62758269e-01,
  5.27406285e-01],
[-5.37177559e-01, -1.31979479e-01,  4.21733708e-01,
  3.95774101e-01],
[ 1.89829664e-01,  7.88807586e-01,  4.21733708e-01,
  5.27406285e-01],
[ 1.03800476e+00,  9.82172869e-02,  5.35408562e-01,
  3.95774101e-01],
[ 5.53333275e-01, -1.74335684e+00,  3.64896281e-01,
  1.32509732e-01],
[-2.94841818e-01, -1.31979479e-01,  1.94384000e-01,
  1.32509732e-01],
[-4.16009689e-01, -1.28296331e+00,  1.37546573e-01,
  1.32509732e-01],
```

```
[-4.16009689e-01, -1.05276654e+00,  3.64896281e-01,
  8.77547895e-04],
[ 3.10997534e-01, -1.31979479e-01,  4.78571135e-01,
  2.64141916e-01],
[-5.25060772e-02, -1.05276654e+00,  1.37546573e-01,
  8.77547895e-04],
[-1.02184904e+00, -1.74335684e+00, -2.60315415e-01,
 -2.62386821e-01],
[-2.94841818e-01, -8.22569778e-01,  2.51221427e-01,
  1.32509732e-01],
[-1.73673948e-01, -1.31979479e-01,  2.51221427e-01,
  8.77547895e-04],
[-1.73673948e-01, -3.62176246e-01,  2.51221427e-01,
  1.32509732e-01],
[ 4.32165405e-01, -3.62176246e-01,  3.08058854e-01,
  1.32509732e-01],
[-9.00681170e-01, -1.28296331e+00, -4.30827696e-01,
 -1.30754636e-01],
[-1.73673948e-01, -5.92373012e-01,  1.94384000e-01,
  1.32509732e-01],
[ 5.53333275e-01,  5.58610819e-01,  1.27429511e+00,
  1.71209594e+00],
[-5.25060772e-02, -8.22569778e-01,  7.62758269e-01,
  9.22302838e-01],
[ 1.52267624e+00, -1.31979479e-01,  1.21745768e+00,
  1.18556721e+00],
[ 5.53333275e-01, -3.62176246e-01,  1.04694540e+00,
  7.90670654e-01],
[ 7.95669016e-01, -1.31979479e-01,  1.16062026e+00,
  1.31719939e+00],
[ 2.12851559e+00, -1.31979479e-01,  1.61531967e+00,
  1.18556721e+00],
[-1.14301691e+00, -1.28296331e+00,  4.21733708e-01,
  6.59038469e-01],
[ 1.76501198e+00, -3.62176246e-01,  1.44480739e+00,
  7.90670654e-01],
[ 1.03800476e+00, -1.28296331e+00,  1.16062026e+00,
  7.90670654e-01],
[ 1.64384411e+00,  1.24920112e+00,  1.33113254e+00,
  1.71209594e+00],
[ 7.95669016e-01,  3.28414053e-01,  7.62758269e-01,
  1.05393502e+00],
[ 6.74501145e-01, -8.22569778e-01,  8.76433123e-01,
  9.22302838e-01],
[ 1.15917263e+00, -1.31979479e-01,  9.90107977e-01,
  1.18556721e+00],
[-1.73673948e-01, -1.28296331e+00,  7.05920842e-01,
  1.05393502e+00],
[-5.25060772e-02, -5.92373012e-01,  7.62758269e-01,
  1.58046376e+00],
[ 6.74501145e-01,  3.28414053e-01,  8.76433123e-01,
  1.44883158e+00],
[ 7.95669016e-01, -1.31979479e-01,  9.90107977e-01,
  7.90670654e-01],
[ 2.24968346e+00,  1.70959465e+00,  1.67215710e+00,
  1.31719939e+00],
[ 2.24968346e+00, -1.05276654e+00,  1.78583195e+00,
  1.44883158e+00],
[ 1.89829664e-01, -1.97355361e+00,  7.05920842e-01,
  3.95774101e-01],
```

```
[ 1.28034050e+00,  3.28414053e-01,  1.10378283e+00,
  1.44883158e+00],
[-2.94841818e-01, -5.92373012e-01,  6.49083415e-01,
  1.05393502e+00],
[ 2.24968346e+00, -5.92373012e-01,  1.67215710e+00,
  1.05393502e+00],
[ 5.53333275e-01, -8.22569778e-01,  6.49083415e-01,
  7.90670654e-01],
[ 1.03800476e+00,  5.58610819e-01,  1.10378283e+00,
  1.18556721e+00],
[ 1.64384411e+00,  3.28414053e-01,  1.27429511e+00,
  7.90670654e-01],
[ 4.32165405e-01, -5.92373012e-01,  5.92245988e-01,
  7.90670654e-01],
[ 3.10997534e-01, -1.31979479e-01,  6.49083415e-01,
  7.90670654e-01],
[ 6.74501145e-01, -5.92373012e-01,  1.04694540e+00,
  1.18556721e+00],
[ 1.64384411e+00, -1.31979479e-01,  1.16062026e+00,
  5.27406285e-01],
[ 1.88617985e+00, -5.92373012e-01,  1.33113254e+00,
  9.22302838e-01],
[ 2.49201920e+00,  1.70959465e+00,  1.50164482e+00,
  1.05393502e+00],
[ 6.74501145e-01, -5.92373012e-01,  1.04694540e+00,
  1.31719939e+00],
[ 5.53333275e-01, -5.92373012e-01,  7.62758269e-01,
  3.95774101e-01],
[ 3.10997534e-01, -1.05276654e+00,  1.04694540e+00,
  2.64141916e-01],
[ 2.24968346e+00, -1.31979479e-01,  1.33113254e+00,
  1.44883158e+00],
[ 5.53333275e-01,  7.88807586e-01,  1.04694540e+00,
  1.58046376e+00],
[ 6.74501145e-01,  9.82172869e-02,  9.90107977e-01,
  7.90670654e-01],
[ 1.89829664e-01, -1.31979479e-01,  5.92245988e-01,
  7.90670654e-01],
[ 1.28034050e+00,  9.82172869e-02,  9.33270550e-01,
  1.18556721e+00],
[ 1.03800476e+00,  9.82172869e-02,  1.04694540e+00,
  1.58046376e+00],
[ 1.28034050e+00,  9.82172869e-02,  7.62758269e-01,
  1.44883158e+00],
[-5.25060772e-02, -8.22569778e-01,  7.62758269e-01,
  9.22302838e-01],
[ 1.15917263e+00,  3.28414053e-01,  1.21745768e+00,
  1.44883158e+00],
[ 1.03800476e+00,  5.58610819e-01,  1.10378283e+00,
  1.71209594e+00],
[ 1.03800476e+00, -1.31979479e-01,  8.19595696e-01,
  1.44883158e+00],
[ 5.53333275e-01, -1.28296331e+00,  7.05920842e-01,
  9.22302838e-01],
[ 7.95669016e-01, -1.31979479e-01,  8.19595696e-01,
  1.05393502e+00],
[ 4.32165405e-01,  7.88807586e-01,  9.33270550e-01,
  1.44883158e+00],
[ 6.86617933e-02, -1.31979479e-01,  7.62758269e-01,
  7.90670654e-01]])
```

In [12]:
```
X_scaled.shape
```

Out[12]: (150, 4)

In [14]:
```
# apply the pca
from sklearn.decomposition import PCA
```

In [15]:
```
pca = PCA(n_components=2)
```

In [16]:
```
pca
```

Out[16]:
```
▾ PCA
▶ Parameters
```

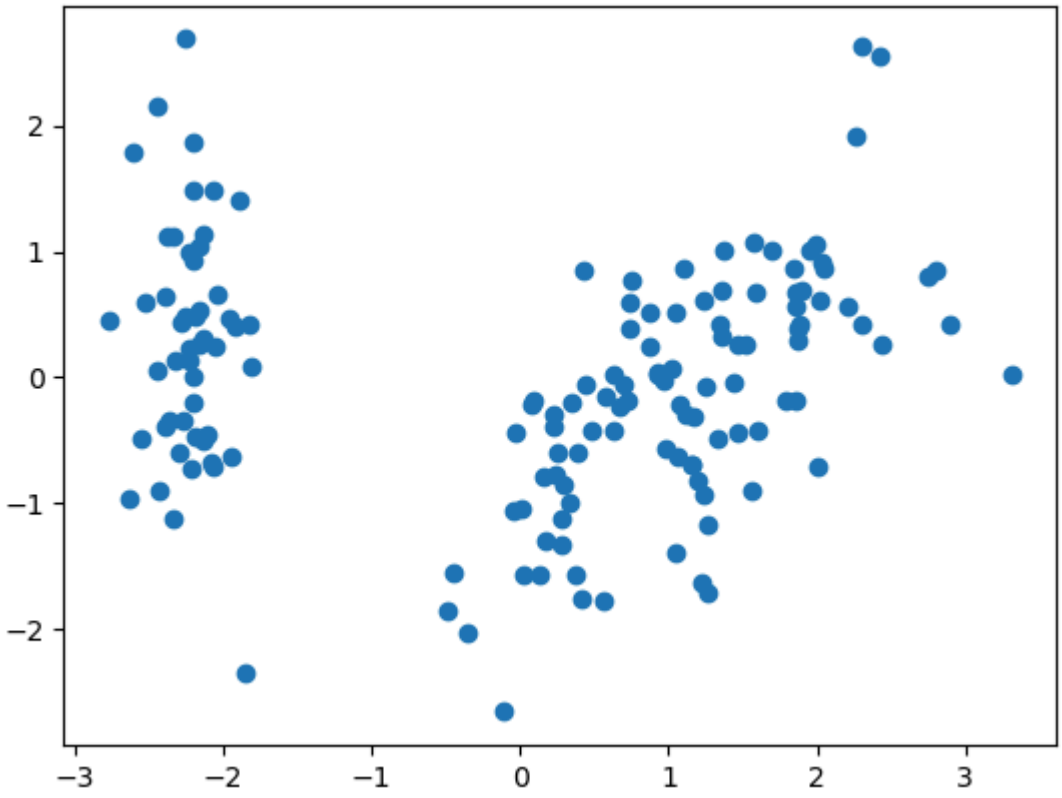
In [17]:
```
pca_scaled =pca.fit_transform(X_scaled)
```

In [19]:
```
pca_scaled.shape
```

Out[19]: (150, 2)

In [20]:
```
plt.scatter(pca_scaled[:,0],pca_scaled[:,1])
```

Out[20]: <matplotlib.collections.PathCollection at 0x237e818fc90>

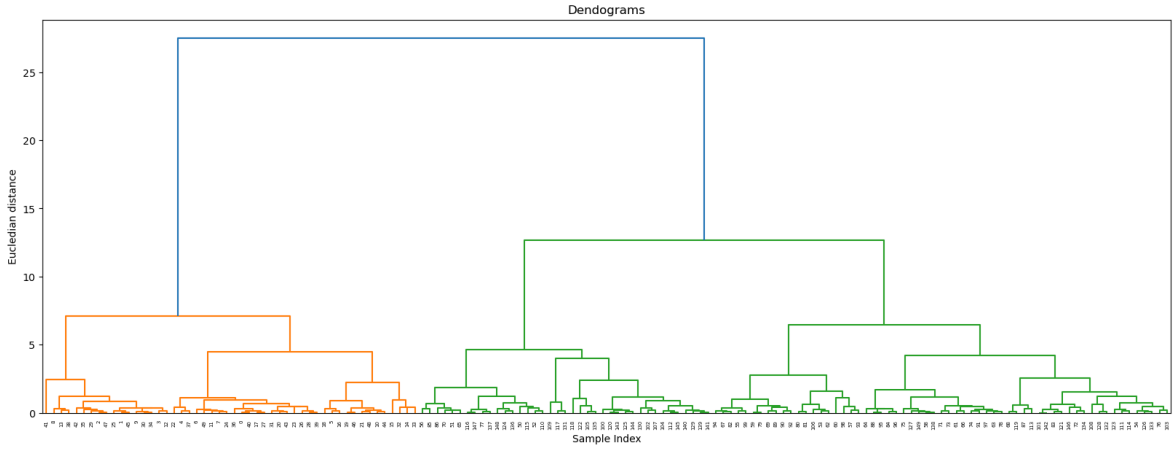


In [23]:
```
## Agglomerative Clustering
import scipy.cluster.hierarchy as sc

plt.figure(figsize=(20,7))
plt.title("Dendograms")
```

```
# craete dendograms
sc.dendrogram(sc.linkage(pca_scaled,method='ward'))
plt.title('Dendograms')
plt.xlabel('Sample Index')
plt.ylabel('Eucledian distance')
```

Out[23]:
```
Text(0, 0.5, 'Eucledian distance')
```



In [26]:
```
from sklearn.cluster import AgglomerativeClustering
cluster = AgglomerativeClustering(n_clusters=2,metric='euclidean',linkage='ward'
cluster.fit(pca_scaled)
```

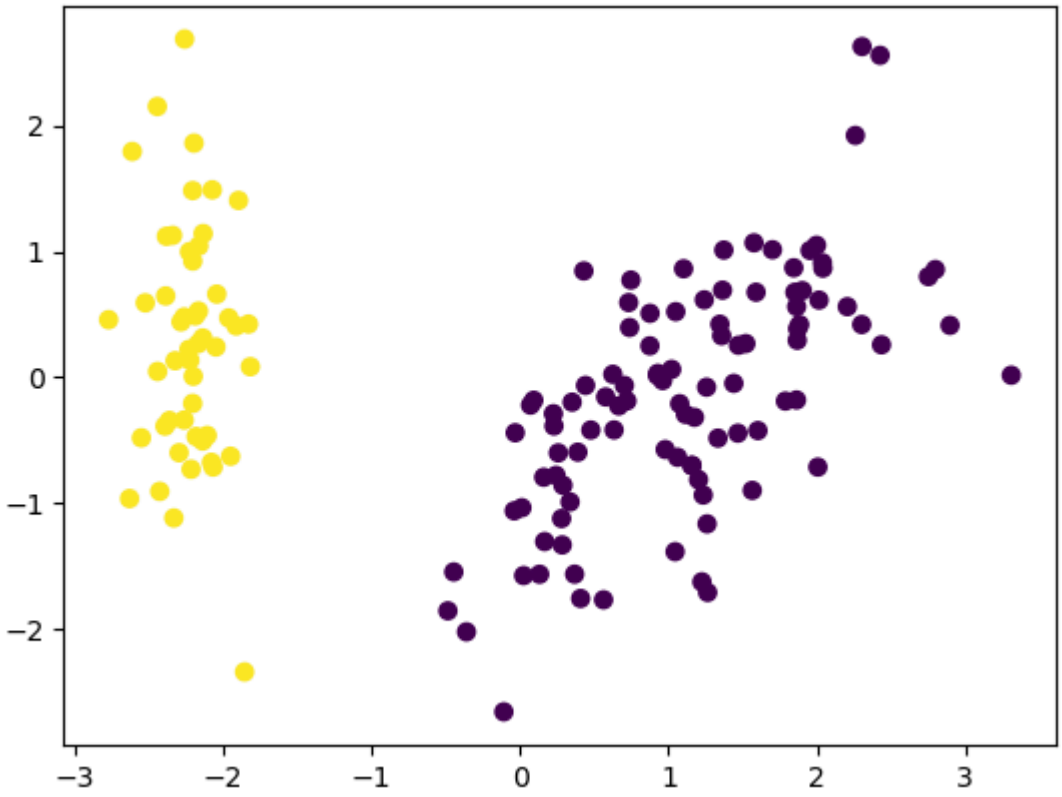
Out[26]:
```
AgglomerativeClustering
▶ Parameters
```

In [27]:
```
cluster.labels_
```

Out[27]:
```
array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], dtype=int64)
```

In [28]:
```
plt.scatter(pca_scaled[:,0],pca_scaled[:,1],c=cluster.labels_)
```

Out[28]:
```
<matplotlib.collections.PathCollection at 0x237ea7616d0>
```



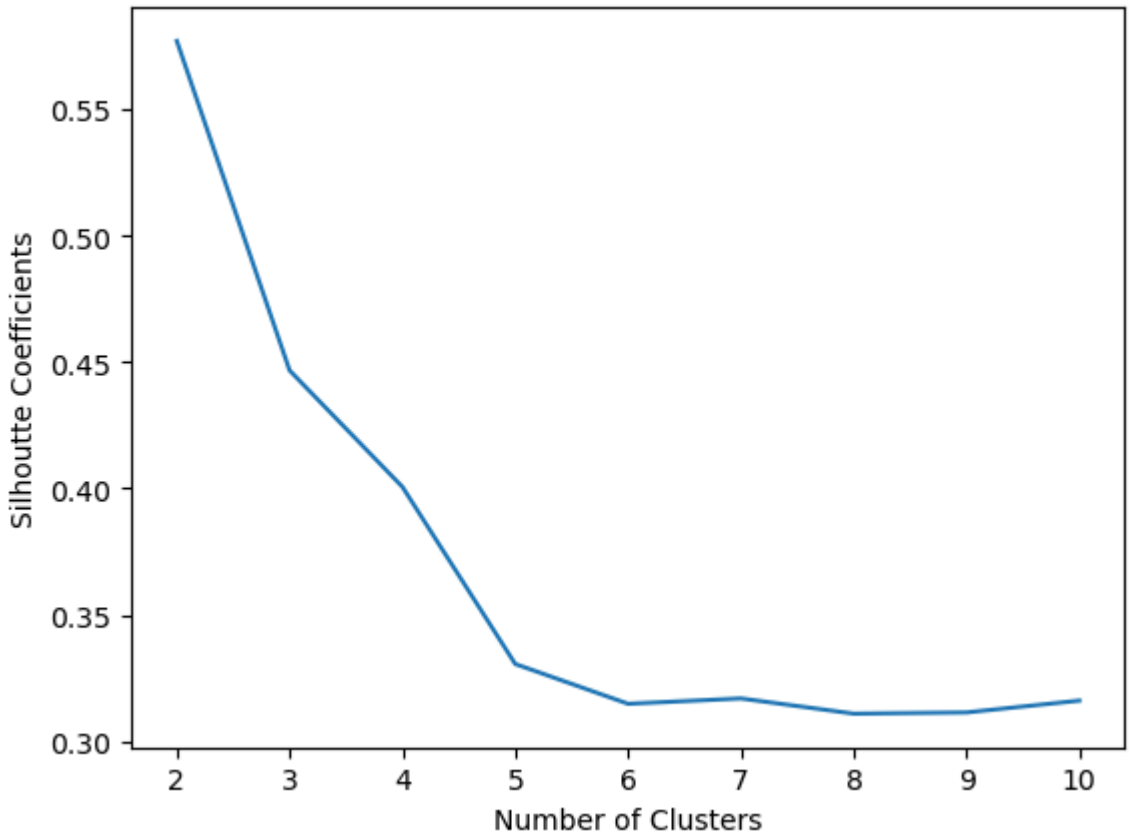
In [29]:
```
# performance metrics
# silhoutte score
from sklearn.metrics import silhouette_score
```

In [32]:
```
silhouette_coefficients = []
for k in range(2,11):
    agglo = AgglomerativeClustering(n_clusters=k,metric='euclidean',linkage='war
    agglo.fit(X_scaled)
    score = silhouette_score(X_scaled, agglo.labels_)
    silhouette_coefficients.append(score)
```

In [33]:
```
silhouette_coefficients
```

Out[33]:
```
[0.5770346019475988,
 0.4466890410285909,
 0.4006363159855973,
 0.33058726295230545,
 0.31485480100512825,
 0.316969830299128,
 0.310946529007258,
 0.31143422475471655,
 0.3161120375980681]
```

In [34]:
```
# plot the silhouette score
plt.plot(range(2,11),silhouette_coefficients)
plt.xticks(range(2,11))
plt.xlabel("Number of Clusters")
plt.ylabel("Silhoutte Coefficients")
plt.show()
```


Silhoutte Coefficients
0.55
0.50
0.45
0.40
0.35
0.30
2
3
4
5
6
7
8
9
10
Number of Clusters

In [ ]: