* Clustering
* Dimension reduction

# kmeans

# Import KMeans

from sklearn.cluster import KMeans

# Create a KMeans instance with 3 clusters: model

model = KMeans(n\_clusters = 3)

# Fit model to points

model.fit(points)

# Determine the cluster labels of new\_points: labels

labels = model.predict(new\_points)

# Print cluster labels of new\_points

print(labels)

## plot clusters and centroids

# Import pyplot

import matplotlib.pyplot as plt

# Assign the columns of new\_points: xs and ys

xs = new\_points[:,0]

ys = new\_points[:,1]

# Make a scatter plot of xs and ys, using labels to define the colors

plt.scatter(xs, ys, c=labels, alpha=0.5)

# Assign the cluster centers: centroids

centroids = model.cluster\_centers\_

# Assign the columns of centroids: centroids\_x, centroids\_y

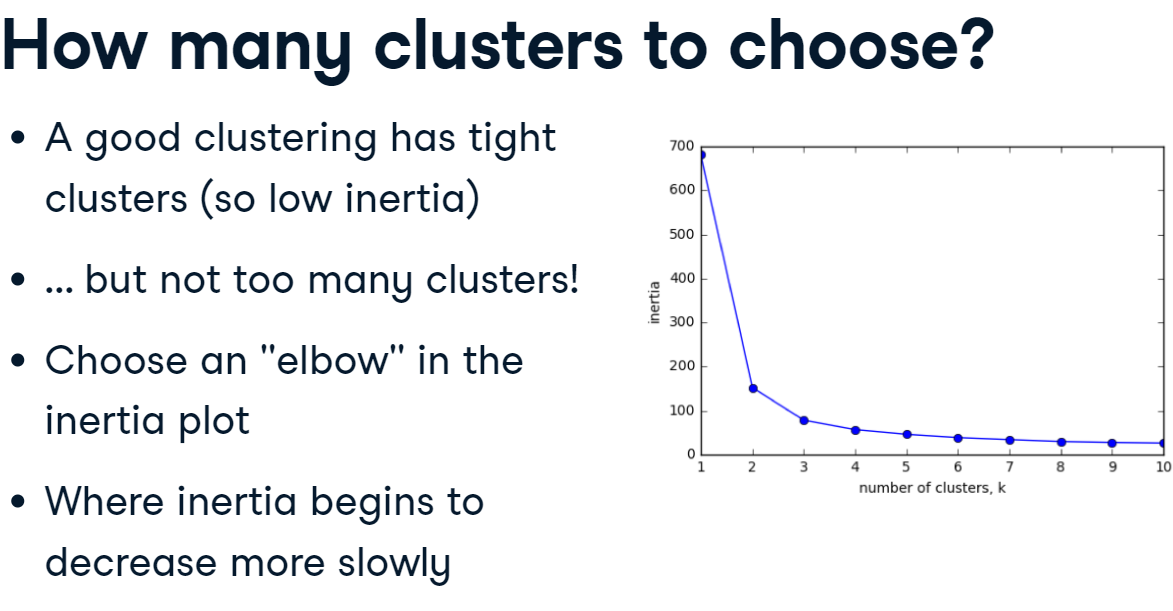
centroids\_x = centroids[:,0]

centroids\_y = centroids[:,1]

# Make a scatter plot of centroids\_x and centroids\_y

plt.scatter(centroids\_x, centroids\_y, marker='D', s=50)

plt.show()



Plot:

ks = range(1, 6)

inertias = []

for k in ks:

    # Create a KMeans instance with k clusters: model

    model = KMeans(n\_clusters = k)

    # Fit model to samples

    model.fit(samples)

    # Append the inertia to the list of inertias

    inertias.append(model.inertia\_)

# Plot ks vs inertias

plt.plot(ks, inertias, '-o')

plt.xlabel('number of clusters, k')

plt.ylabel('inertia')

plt.xticks(ks)

plt.show()

If data are labeld …..

# Create a KMeans model with 3 clusters: model

model = KMeans(n\_clusters = 3)

# Use fit\_predict to fit model and obtain cluster labels: labels

labels = model.fit\_predict(samples)

# Create a DataFrame with labels and varieties as columns: df

df = pd.DataFrame({'labels': labels, 'varieties': varieties})

# Create crosstab: ct

ct = pd.crosstab(df['labels'], df['varieties'])

# Display ct

print(ct)

!!!!!!!!!!!!!!!!!!!

!!! The varicance (spread) of the features (variables) has big impact on kmeans algortim performance. The data need to be transformed to have the same variance. Stadarizetion make variance eaual 1 and zero mean.

!!!!!!!