

DATA 202 — PCA · Week 5 Wednesday

Draw right on the plots. They show the same data as today's slides: 1,797 handwritten digits (8×8 pixels = **64 numbers** each). Your professor will tell you when to do each part.

1 · What's in each cluster?

	average	8 of its members								digit?
cluster 0										
cluster 1										
cluster 2										
cluster 3										
cluster 4										
cluster 5										
cluster 6										
cluster 7										
cluster 8										
cluster 9										

1. k-means split the digits into 10 clusters.
Each row: the cluster's **average image**, then 8 of its members.
2. In the empty box, write the **digit** each row mostly holds.
3. **Circle** every image that doesn't belong in its row.
4. Which clusters are the most mixed?

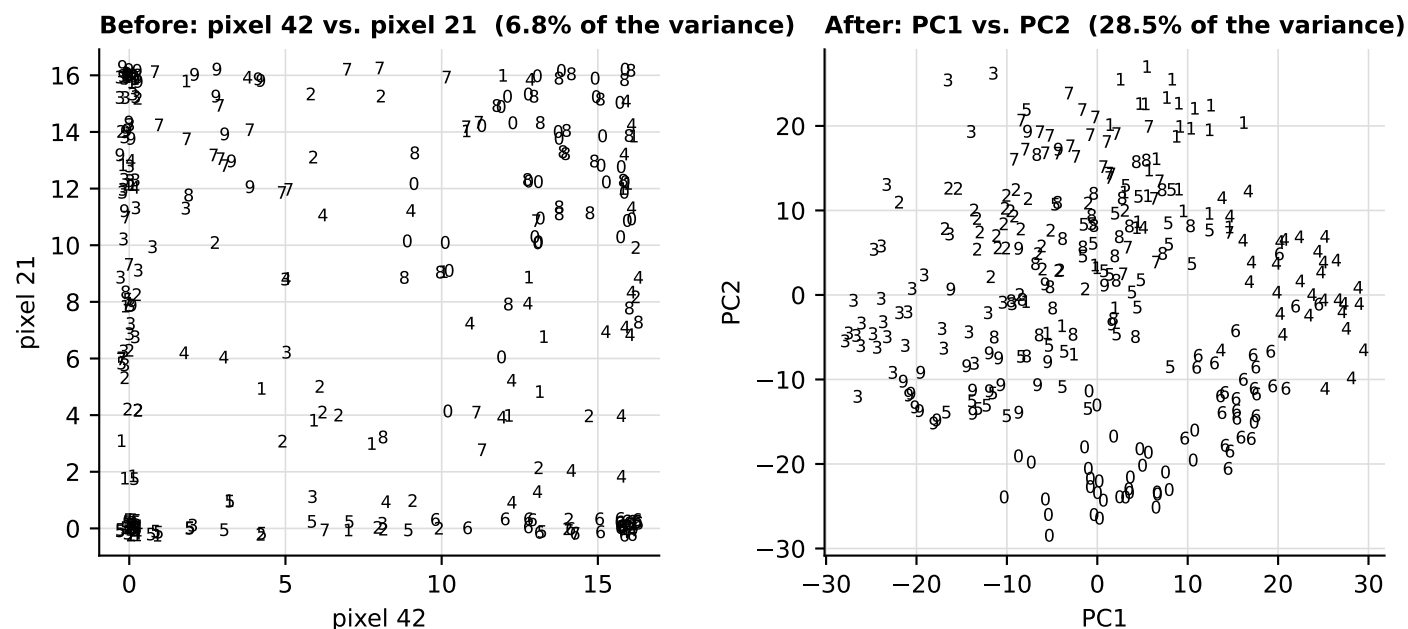
2 · One ruler for every column

Pixel 42 of the first image, measured two ways: the usual darkness scale (0–16), and a scanner that stores that pixel $\times 100$ (0–1,600). Compute its z-score in both units: $\mathbf{z} = (\mathbf{value} - \mathbf{column\ mean}) \div \mathbf{column\ standard\ deviation}$.

	0–16 scale	0–1,600 scale
column mean	6.9	690
column standard deviation	6.5	650
pixel 42, first image	11	1,100
z-score	_____	_____

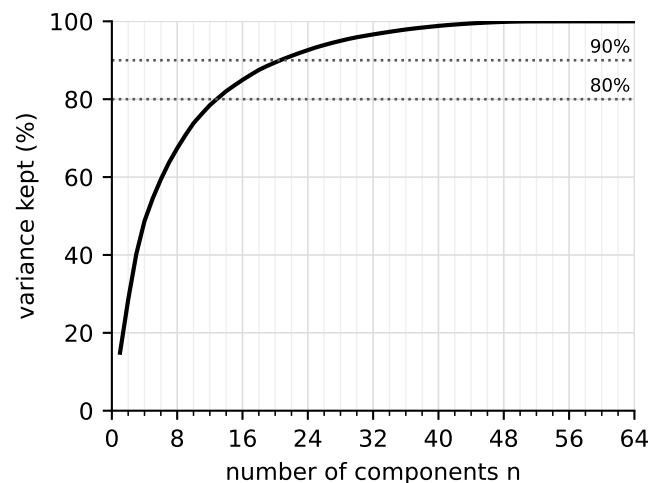
What do you notice?

3 · Two views of the same digits



1. Each character is one image (300 of the 1,797), drawn as its **true digit**.
2. **Circle the 0s** in both views. In which view do they form one group?
3. In the PCA view, name two digits that still **overlap**: _____ and _____

4 · How many components?



1. The share of the variance kept by the first n of the 64 components.
2. Mark where the curve reaches **80%**: _____ components.
3. Mark where it reaches **90%**: _____ components.