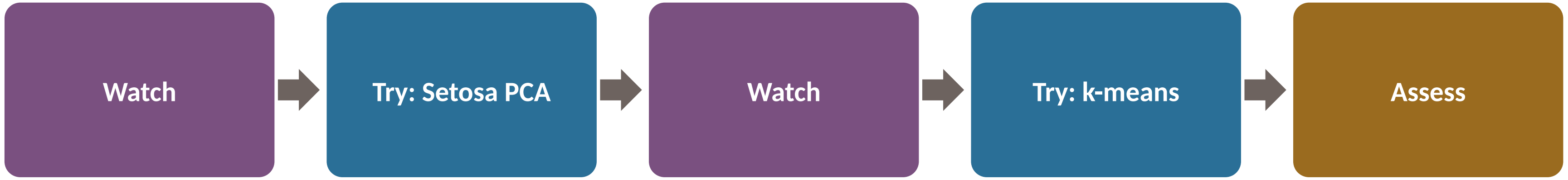
MODULE 6

Unsupervised learning: clustering and PCA

With no label there is no accuracy. Validity comes from stability, cohesion, and interpretability instead.

 ·

Watch a little, then do a little



Each recorded segment is short. You pause, practice the concept hands-on, then return for the next piece.

MODULE 6 · WHERE THIS SITS

Where this module sits in the lifecycle

Unsupervised methods live in the analyze layer, finding structure rather than predicting a target. Because there is no held-out accuracy, you argue for a result with converging evidence, never a single plot.

PART 1 OF 2

PCA and reading it

About 12 minutes on camera

PCA is a variance-preserving rotation

Standardize first so units don’t decide → Rotate to the axes of greatest variance → Read loadings, scree, cumulative variance → Keep the top few components

Principal components are the eigenvectors of the covariance matrix, ordered so the first axis captures the most variance. Loadings tell you what each component means.

Why standardize before you rotate

- Without standardization, a single large-variance column silently dominates the distance metric and decides the answer.
- The scree plot and cumulative variance tell you how many components to keep.
- Scores place each observation in the reduced space, which is where clustering becomes tractable.

NOW TRY IT

Setosa PCA: feel the rotation

- Drag the points and watch the principal components rotate to follow the direction of greatest variance.
- See how correlation shows up as a dominant first component.

About 12 minutes, then come back for Part 2.

PART 2 OF 2

k-means, choosing k, and t-SNE traps

About 13 minutes on camera

k-means is Lloyd's algorithm

- Alternate two steps: assign each point to the nearest centroid, then recompute each centroid as the mean of its members.
- It minimizes within-cluster sum of squares but only converges to a local optimum, so multiple restarts and k-means++ initialization matter.
- Standardize first, or one large-variance feature dominates the distance.

Argue for k, do not eyeball it

Stability

Bootstrap or subsample and measure how often the same points cluster together.

Silhouette

Read the silhouette width across candidate k values.

Gap statistic

Compare within-cluster dispersion against a null reference.

Agreement

Trust a k only where these independent signals converge.

The traps of t-SNE

Perplexity. changes the entire picture, so it is a tuning knob, not a truth

Distances. between clusters are not meaningful and must not be read as similarity

Cluster sizes. are distorted, so a big blob is not a big group

A lens, not evidence. run it twice and the layout changes; use it to generate hypotheses, never to count clusters

NOW TRY IT

Visualizing K-Means and t-SNE

- In Visualizing K-Means, step through Lloyd’s algorithm by hand: try a bad initialization and watch it settle into a poor local optimum.
- In How to Use t-SNE Effectively, move the perplexity sliders until you personally reproduce a misleading picture.

About 18 minutes, then build the formative.

Formative deliverable (completion points)

- On a small standardized dataset, run PCA and report the variance explained by the first two components.
- Then run k-means and choose k using one explicit stability or silhouette argument.
- Write two sentences defending your k against the evidence you gathered.

How it counts. Submit the work with a short reflection or evidence you ran it. Credited on completion, not scored for correctness. This is the practice that makes the graded simulation go well.

GRADED ASSESSMENT

How Many Segments, Really?

- You are handed a customer dataset with genuine structure, correlated noise columns, and non-spherical, unequally sized groups that punish naive k-means.
- Marketing asks: how many distinct groups actually exist, and how would you describe each so we can target them?
- Your work is judged the way segmentation is defended in industry: converging evidence for the count, not a single elbow plot.