

12 **Unsupervised Learning: Clustering and Dimensionality Reduction**

Learning Without Labels

Where Chapters 8-11 modelled a known outcome (a disease status, a time-to-event, a risk category), unsupervised learning discovers structure in data with no designated outcome variable — grouping similar observations together (clustering) or reducing many correlated variables to a smaller set of underlying dimensions (dimensionality reduction). In health data science, these methods are used to identify patient or facility subgroups with distinct profiles, to detect natural strata in a population that were not part of the original study design, and to compress high-dimensional data (many biomarkers, many survey items) for visualisation and downstream modelling.

K-Means and Hierarchical Clustering

K-means clustering partitions observations into k groups by iteratively assigning each observation to the nearest cluster centroid and recomputing centroids, requiring the analyst to pre-specify k (guided by the elbow method or the silhouette coefficient) and to standardise variables first, since the algorithm is sensitive to scale. Hierarchical clustering instead builds a nested tree (dendrogram) of clusters by successively merging (agglomerative) or splitting (divisive) groups based on a chosen

distance metric and linkage method, and it has the practical advantage of not requiring k to be fixed in advance — the analyst can cut the dendrogram at whatever level of granularity is scientifically meaningful. Both methods have been applied, for example, to cluster health facilities by their diagnostic testing profile (proportion RDT, microscopy, and PCR-confirmed cases) to identify facility archetypes for targeted quality-improvement support, and to cluster dermatoglyphic (fingerprint pattern) profiles in a study comparing children affected by a congenital condition against unaffected controls.

R example

```
# R: k-means clustering of facilities by diagnostic testing profile
library(dplyr)

facility_profile <- facility_data %>%
  select(pct_rdt, pct_microscopy, pct_pcr, pct_positivity) %>%
  scale() # standardise before clustering

set.seed(42)
km <- kmeans(facility_profile, centers = 4, nstart = 25)
facility_data$cluster <- km$cluster
table(facility_data$cluster)
```

Model-Based Clustering and Mixture Models

K-means assumes roughly spherical, equally sized clusters, an assumption that often does not hold in health data. Gaussian mixture models instead assume the data arise from a mixture of several probability distributions (typically multivariate normal), estimated via the Expectation-Maximization (EM) algorithm, and yield a probabilistic (soft) cluster assignment for each observation rather than a hard label — a more honest representation of uncertainty when group boundaries are not sharp, as is common in clinical subtyping.

Principal Component Analysis and Dimensionality Reduction

Principal component analysis (PCA) transforms a set of correlated variables into a smaller set of uncorrelated components, each a linear combination of the original variables, ordered by the proportion of total variance they explain. In health survey data with many correlated items (for example, a multi-item household asset inventory used to construct a wealth index, or a multi-item adolescent depression symptom checklist), PCA both reduces dimensionality for downstream modelling and provides a principled way to construct a single composite index (the widely used DHS wealth index is itself constructed via PCA on household asset ownership) from many individually noisy indicators.

Validating Cluster Solutions

Because clustering has no ground truth outcome to validate against, cluster solutions must be evaluated on internal criteria (the silhouette coefficient, measuring how well-separated clusters are) and, more importantly, on external, substantive grounds: do the resulting clusters correspond to a meaningful and actionable distinction (for example, do facility clusters correspond to detectably different service-delivery needs), and are they stable under reasonable perturbations of the data (a bootstrap or subsampling stability check)? A clustering result that cannot be explained to, and recognised by, domain experts familiar with the facilities or patients in question should be treated with suspicion regardless of how clean it looks statistically.

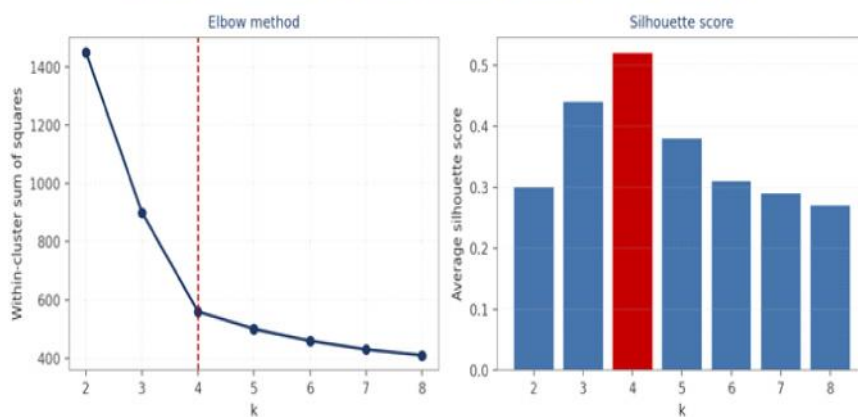
Worked Example

Choosing k with the Elbow Method and Silhouette Score

Applying k -means to the facility diagnostic-profile data from Section 12.2 across $k = 2$ to $k = 8$, the within-cluster sum of squares (WCSS) decreases sharply from $k=2$ to $k=4$, then flattens

noticeably from $k=4$ onward — the classic ‘elbow’ pattern suggesting $k=4$ is a reasonable choice. Computing the average silhouette score for each k confirms this: the silhouette score peaks at $k=4$ (0.52, indicating reasonably well-separated clusters) and declines for $k=5$ and beyond (0.38 and 0.31), giving converging evidence from two independent diagnostics that four facility archetypes are appropriate for this dataset, rather than a finer or coarser partition.

Figure 12.1 — Choosing k for Facility Clustering (Worked Example, Sec. 12.6)



PYTHON example

```
# Python: elbow method and silhouette score to choose k
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import silhouette_score

X_scaled = StandardScaler().fit_transform(facility_df[["pct_rdt", "pct_microscopy",
                                                         "pct_pcr", "pct_positivity"]])

wcss, silhouette = [], []
for k in range(2, 9):
    km = KMeans(n_clusters=k, n_init=25, random_state=42).fit(X_scaled)
    wcss.append(km.inertia_)
    silhouette.append(silhouette_score(X_scaled, km.labels_))

for k, w, s in zip(range(2, 9), wcss, silhouette):
    print(f"k={k}: WCSS={w:.1f}, silhouette={s:.3f}")
```

Worked Example

PCA-Based Wealth Index Construction

Consider a household asset inventory with 12 binary items (electricity, radio, television, refrigerator, bicycle, motorcycle, car, and six housing-quality indicators). Running PCA on these 12 standardised items, the first principal component explains 34% of total variance and loads positively on nearly all items (electricity: 0.31, television: 0.35, refrigerator: 0.33, car: 0.22, improved roofing: 0.29, and so on) — consistent with its interpretation as a general household wealth dimension, exactly the logic underlying the construction of the DHS wealth index. Households are then ranked by their score on this first component and divided into wealth quintiles, providing a single, defensible ordinal measure of relative socioeconomic status from 12 individually noisy binary indicators.

R example

```
# R: constructing a PCA-based wealth index from household asset items
library(dplyr)

asset_items <- household_data %>%
  select(electricity, radio, television, refrigerator, bicycle, motorcycle,
         car, improved_roof, improved_floor, improved_wall, piped_water, flush_toilet)

pca_result <- prcomp(asset_items, scale. = TRUE)
summary(pca_result) # variance explained by each component
household_data$wealth_score <- pca_result$x[, 1]
household_data$wealth_quintile <- ntile(household_data$wealth_score, 5)
```

DBSCAN and Density-Based Clustering

K-means and hierarchical clustering both struggle with clusters of irregular shape and cannot automatically identify outlier points that do not belong to any cluster. Density-based spatial clustering (DBSCAN) instead groups points that are closely packed together, marking points in low-density regions as noise rather than forcing them into the nearest cluster — a useful property when clustering,

for example, the geographic coordinates of confirmed case locations during an outbreak investigation, where a handful of geographically isolated cases genuinely should not be forced into the same cluster as a large, dense case cluster near a point source.

Extended Case Study

Clusters That Told a Different Story Than Expected

A research team applies k-means clustering to a set of health facilities using indicators of infrastructure and staffing (number of skilled birth attendants, availability of emergency obstetric care equipment, distance to referral hospital, average monthly patient volume), expecting the resulting clusters to align neatly with the facilities' official tier classification (primary, secondary, tertiary). Instead, the four-cluster solution that best satisfies the elbow and silhouette diagnostics of Section 12.6 cuts across the official tier system: one cluster contains a mix of officially 'primary' and 'secondary' facilities that, empirically, share a very similar infrastructure and staffing profile, while a second cluster contains some nominally 'secondary' facilities that, on the actual measured indicators, more closely resemble the primary-tier profile — suggesting that the official administrative tier classification, likely set some years earlier and not since revised, no longer accurately reflects several facilities' actual current operational capacity.

This is precisely the kind of substantively meaningful, actionable clustering result argued for in Section 12.5 — rather than a disappointing failure to replicate the expected administrative categories, the mismatch itself is the finding, and it is a directly useful one: it identifies specific facilities whose official tier and actual operational capacity have diverged, a mismatch that has concrete implications for how referral pathways, staffing allocations, and quality-improvement resources should realistically be directed, independent of what each facility's administrative paperwork currently states. Presenting this finding

required the research team to resist an initial instinct to treat the mismatch with the expected tier system as evidence the clustering had somehow ‘failed,’ and instead to trust the empirical, data-driven grouping and investigate why it diverged from the administrative expectation — precisely the kind of substantive interpretation, rather than mechanical pattern-matching to a pre-existing expectation, that responsible unsupervised learning requires.

Silhouette Scores for Individual Observations, Not Just the Overall Average

Section 12.6 introduced the average silhouette score as a tool for choosing k ; it is equally useful, and under-used, to examine the silhouette score for each individual observation rather than only the dataset-wide average. An observation with a low or negative individual silhouette score sits poorly within its assigned cluster (closer, in fact, to a neighbouring cluster than to its own), and examining which specific facilities or patients receive low individual silhouette scores often reveals genuinely ambiguous or borderline cases worth individual attention — for example, a facility whose testing profile places it right at the boundary between two archetype clusters might warrant a closer, more qualitative review rather than a mechanical assignment to whichever cluster centroid it happens to sit marginally closer to.

Cluster Analysis as an Input to, Not a Replacement for, Domain Expertise

A recurring theme across this chapter’s worked examples (Sections 12.6, 12.9, and 12.10) deserves a final, explicit statement: clustering algorithms are best understood as tools that organise and surface patterns in data for a domain expert’s subsequent judgment, not as tools that autonomously produce a final, actionable classification requiring no further human interpretation. Even the most methodologically careful clustering analysis — correctly standardised, validated with converging elbow

and silhouette diagnostics, checked for stability — produces groupings whose real-world meaning and appropriate use must still be established through discussion with people who understand the substantive domain (facility managers, clinicians, programme officers), and a data scientist who presents a clustering result as a final, self-evidently meaningful answer, without this collaborative interpretation step, risks either overstating the algorithm's autonomous insight or missing an opportunity to catch a spurious or misleading grouping before it influences a real decision.

Cluster Stability via Bootstrap Resampling

Beyond the elbow method and silhouette score (Section 12.6), a further, complementary check of a clustering solution's trustworthiness is bootstrap stability analysis: repeatedly resampling the original dataset with replacement, re-running the clustering algorithm on each resampled dataset, and measuring how consistently each observation is grouped with the same peers across resamples (commonly summarised via the Jaccard similarity between the original and resampled cluster assignments). A cluster that remains stable — consistently reappearing across the great majority of bootstrap resamples — is far more trustworthy as a genuine, reproducible feature of the underlying population than one that appears in the original analysis but dissolves or merges unpredictably under minor resampling perturbation, which would instead suggest the original cluster was closer to a statistical artefact of that one particular sample.

Determining the Right Number of Principal Components to Retain

Section 12.4 introduced PCA's first component as a wealth index; in general dimensionality-reduction applications, an analyst must decide how many of the (potentially many) principal components to retain for further analysis. A scree plot (the proportion of variance explained by each successive component, plotted in order)

showing a clear ‘elbow’ — a small number of components explaining most of the variance, followed by many components each explaining very little — provides a visual guide analogous to the elbow method for choosing k in Section 12.6; a more formal rule, the Kaiser criterion, retains only components with an eigenvalue greater than 1 (explaining more variance than a single original standardised variable would). Both approaches are heuristics rather than definitive statistical tests, and the ultimate decision should also be informed by whether the retained components remain substantively interpretable, per this chapter’s repeated emphasis on domain-expert validation.

An Automated k Selection Function

Section 12.6 chose k by visually inspecting an elbow plot and a silhouette score table side by side — a reasonable manual process for a one-off analysis, but one worth automating when clustering must be repeated across many similar datasets (for example, re-clustering facility profiles every reporting quarter). The function below selects the k with the highest silhouette score automatically, while still returning the full table so the choice remains fully auditable rather than a hidden, unexplained decision.

PYTHON example

```
# Python: automatic k selection, with the full comparison table returned for auditability
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score

def select_best_k(X_scaled, k_range=range(2, 9), random_state=42):
    results = []
    for k in k_range:
        km = KMeans(n_clusters=k, n_init=25, random_state=random_state).fit(X_scaled)
        results.append({"k": k, "wcss": km.inertia_, "silhouette": silhouette_score(X_scaled,
km.labels_)})
    best = max(results, key=lambda r: r["silhouette"])
    return best["k"], results

best_k, table = select_best_k(X_scaled)
print(f"Selected k={best_k}"); [print(r) for r in table]
```

Chapter Summary

- Unsupervised learning finds structure (clusters, reduced dimensions) without a labelled outcome variable.
- K-means requires pre-specifying k and standardised variables; hierarchical clustering builds a dendrogram without fixing k in advance.
- Gaussian mixture models provide probabilistic cluster assignments and relax k-means' spherical-cluster assumption.
- PCA reduces correlated variables to uncorrelated components and underlies standard tools such as the DHS wealth index.
- Cluster solutions require both internal validity checks (elbow method, silhouette score) and substantive, domain-expert interpretability before being acted upon.
- Choosing k should draw on converging evidence from multiple diagnostics (elbow, silhouette) rather than either one alone.

Mini-Project

Mini-Project 12: Clustering Facilities or Constructing a Composite Index

Choose one of two applied tasks: (a) cluster a set of facilities or patients into meaningful subgroups, or (b) construct a PCA-based composite index from a multi-item questionnaire.

Step 1. For task (a): select at least four numeric variables describing your units of analysis, standardise them, and run k-means across a range of k values.

Step 2. Use both the elbow method and silhouette score (Section 12.6) to justify your choice of k, and profile each resulting cluster in plain language.

Step 3. For task (b): select at least eight binary or ordinal items measuring a single underlying construct (wealth, symptom severity, service quality), and run PCA.

Step 4. Report the proportion of variance explained by the first component and interpret its loadings, following Section 12.7.

Step 5. Whichever task you chose, write a short paragraph on how you would validate your result's stability (e.g., via bootstrap resampling) before presenting it to a non-technical audience.

Deliverable: For task (a): a cluster profile table and elbow/silhouette plots. For task (b): a PCA loadings table and a distribution plot of the resulting index.

End-of-Chapter Exercises

1. Explain why variables should be standardised before k-means clustering, using an example where one variable is measured in the hundreds (e.g., CD4 count) and another is a proportion between 0 and 1.
2. You cluster health facilities into 4 groups by testing profile. Describe two ways you would check whether this clustering is meaningful, beyond the silhouette coefficient.
3. Explain, conceptually, how PCA could be used to construct a single household wealth index from a 15-item asset ownership questionnaire.
4. Using the worked example in Section 12.6, explain why converging evidence from both the elbow method and the silhouette score is more persuasive than either method alone for choosing $k=4$.
5. In Section 12.7, explain what it means for the first principal component to ‘load positively on nearly all items,’ and why this supports interpreting it as a general wealth dimension rather than, say, a rural/urban distinction.
6. Using the R code in Section 12.7 as a guide, describe what would change in your interpretation if the second principal component (rather than the first) showed strong positive loadings on rural-associated items and negative loadings on urban-associated items.
7. A colleague clusters patients into 6 groups because ‘more clusters gives more detail.’ Using the logic of Sections 12.5-12.6, explain why more clusters is not automatically better.
8. Using Section 12.8, explain why DBSCAN might be more appropriate than k-means for clustering the geographic locations of outbreak cases.

9. What does it mean for DBSCAN to label some points as ‘noise’ rather than assigning them to a cluster, and why might this be a desirable property for outbreak case mapping?
10. Describe one limitation of DBSCAN compared to k-means (for example, in terms of parameter selection or its behaviour with clusters of very different densities).
11. In the Section 12.9 case study, explain why the mismatch between the clustering result and the official facility tier system was treated as a finding rather than a failure.
12. What practical actions might a state health administration take in response to discovering that some ‘secondary’ facilities empirically resemble ‘primary’ facilities?
13. Why would it have been a mistake for the research team to force the clustering to match the expected tier categories rather than reporting the empirical result?
14. Propose one additional facility-level indicator that could be added to strengthen this clustering analysis, and explain what it might reveal.
15. How would you communicate this finding sensitively to facility staff, given that it could be seen as a critique of their facility’s official classification?
16. Using Section 12.10, explain what a low or negative individual silhouette score indicates about a specific observation’s cluster assignment.
17. Why might a facility with a low individual silhouette score be worth individual, qualitative review rather than automatic reassignment to another cluster?
18. Propose how you would use individual silhouette scores to prioritise which facilities a quality-improvement team should visit first.

19. Explain the difference between the average silhouette score (Section 12.6) and the individual silhouette score (Section 12.10), and what each is used for.
20. If many observations in a cluster solution have low individual silhouette scores, what might this suggest about the overall choice of k ?
21. Using Section 12.11, explain why a clustering result should be treated as an input to domain expert judgment rather than a final, self-evidently meaningful answer.
22. What risk does a data scientist take by presenting a clustering result as a final answer without a collaborative interpretation step?
23. Propose how you would structure a meeting with facility managers to interpret a new clustering result together, drawing on Section 12.9's case study.
24. Why might even a methodologically flawless clustering analysis (correctly validated with elbow and silhouette diagnostics) still need substantive human interpretation?
25. Explain, using your own words, the difference between an algorithm 'surfacing a pattern' and an algorithm 'producing a final answer.'
26. Summarise, in your own words, why this chapter insists that clustering results must be validated against domain expertise, not statistical criteria alone.
27. Explain the common thread linking the individual silhouette scores discussion (Section 12.10) and the extended case study of clusters telling an unexpected story (Section 12.9).
28. Why does this chapter describe cluster analysis as an input to, rather than a replacement for, domain expertise (Section 12.11)?

29. Propose a clustering task from your own field where you would want to examine individual silhouette scores, not just the average, before trusting the result.
30. If a colleague presented a clustering result and asked whether it was ‘correct,’ how would you respond, drawing on this chapter’s arguments?
31. Using Section 12.13, explain the difference between a scree plot’s ‘elbow’ and the Kaiser criterion for choosing how many principal components to retain.
32. Why are both of these methods described as heuristics rather than definitive statistical tests?
33. Explain why substantive interpretability should also inform how many components are retained, not just the numeric criteria alone.
34. Describe a scenario where the scree plot and the Kaiser criterion might suggest a different number of components to retain, and how you would decide between them.
35. How does the logic of choosing the number of principal components relate to the logic of choosing k in k -means clustering (Section 12.6)?