Unsupervised Learning: k-means

Mirco Schönfeld
mirco.schoenfeld@uni-bayreuth.de

Strategies of Clustering

Hierarchical Agglomerative Clustering	Point assignment
Each point is in its own cluster	Initial clusters are estimated
Clusters are combined based on their “closeness”	Points are considered in some order
Combination stops when undesirable clusters occur	Points are assigned to clusters into which they best fit

Leskovec, J., Rajaraman, A. and Ullman, J.D., 2020. *Mining of massive data sets*. Cambridge university press.

Examples: k-Means Clustering


```
Place each point in the cluster whose current centroid is the nearest
WHILE points are moving between clusters and centroids not stabilized
DO
      Update locations of centroids of k clusters
      Reassign all points to their closest centroid
END
```

Disclaimer:

This is the standard k-means algorithm proposed by Lloyd (1982)

It is, however, not the most efficient variant..

Lloyd, Stuart. "Least squares quantization in PCM." IEEE transactions on information theory 28.2 (1982): 129-137.

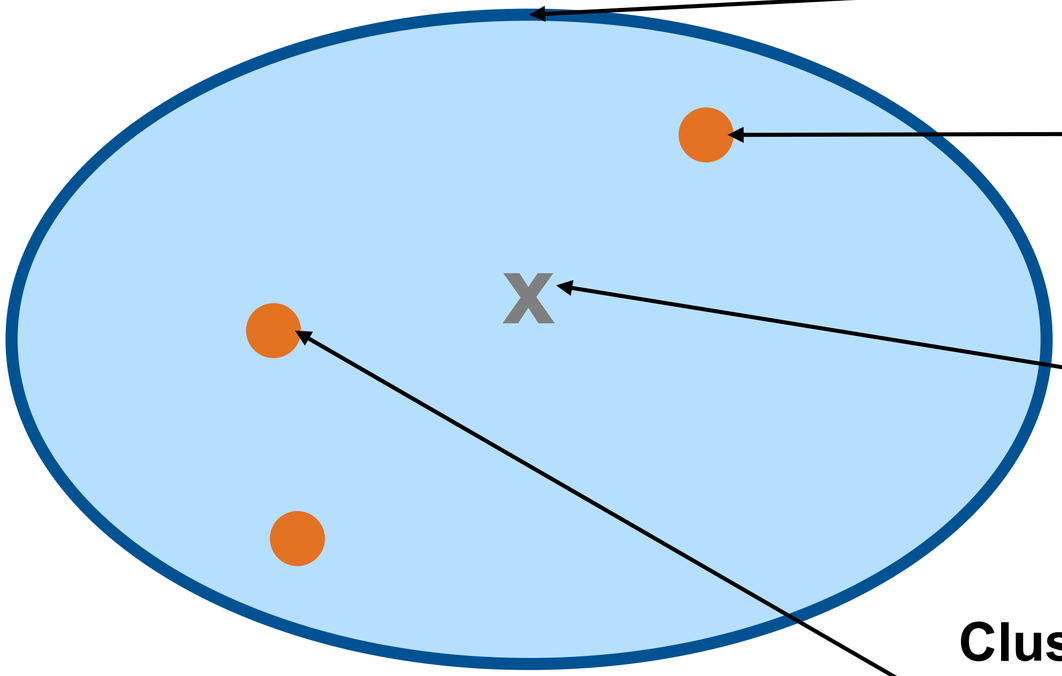
Examples: k-Means Clustering

Clusters represented by their arithmetic mean

Optimizes „least squared errors“, i.e. minimizes distance of points from centroids

That‘s why k-means is bound to Euclidean distance in Euclidean spaces

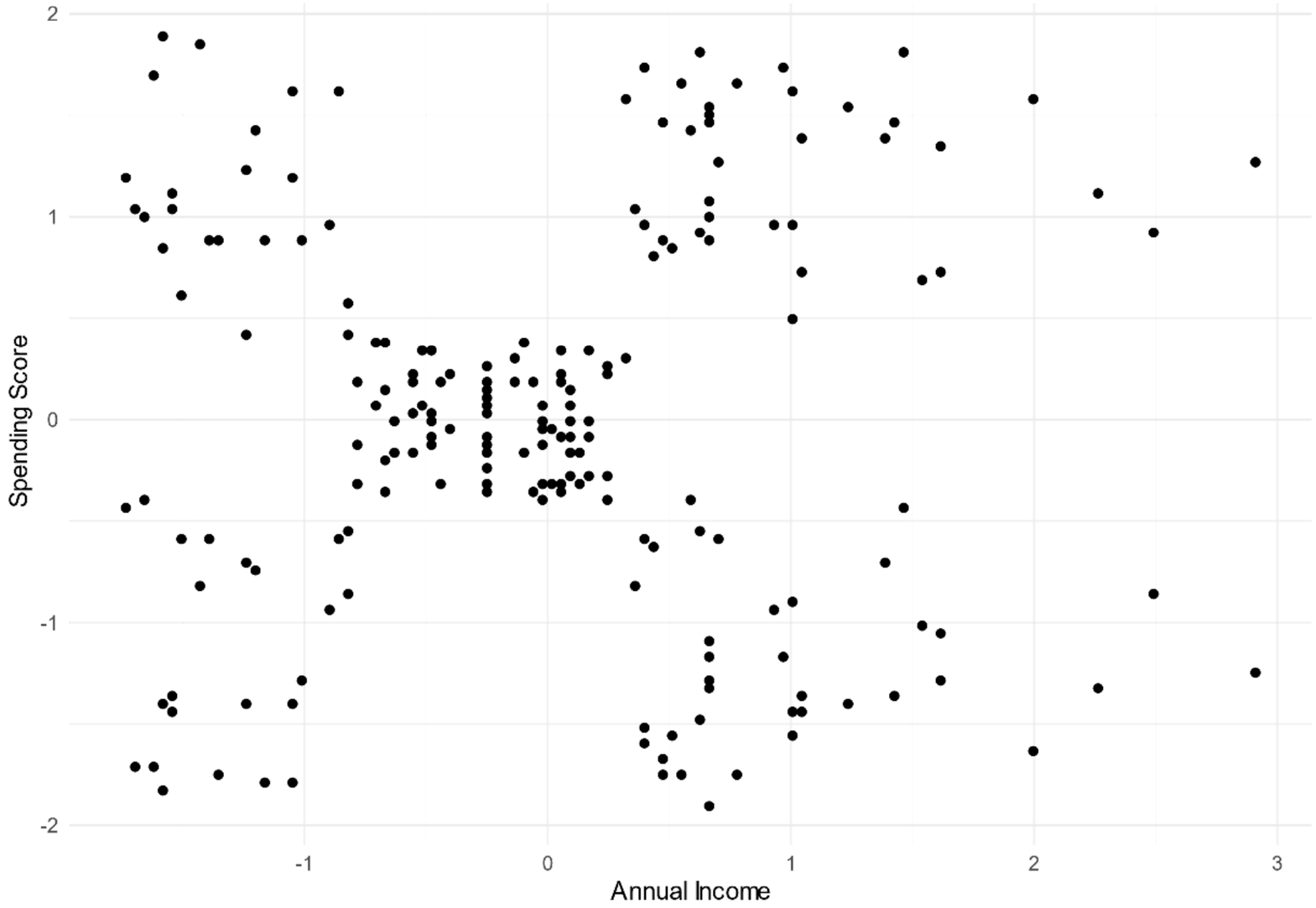
Centroids and Clustroids



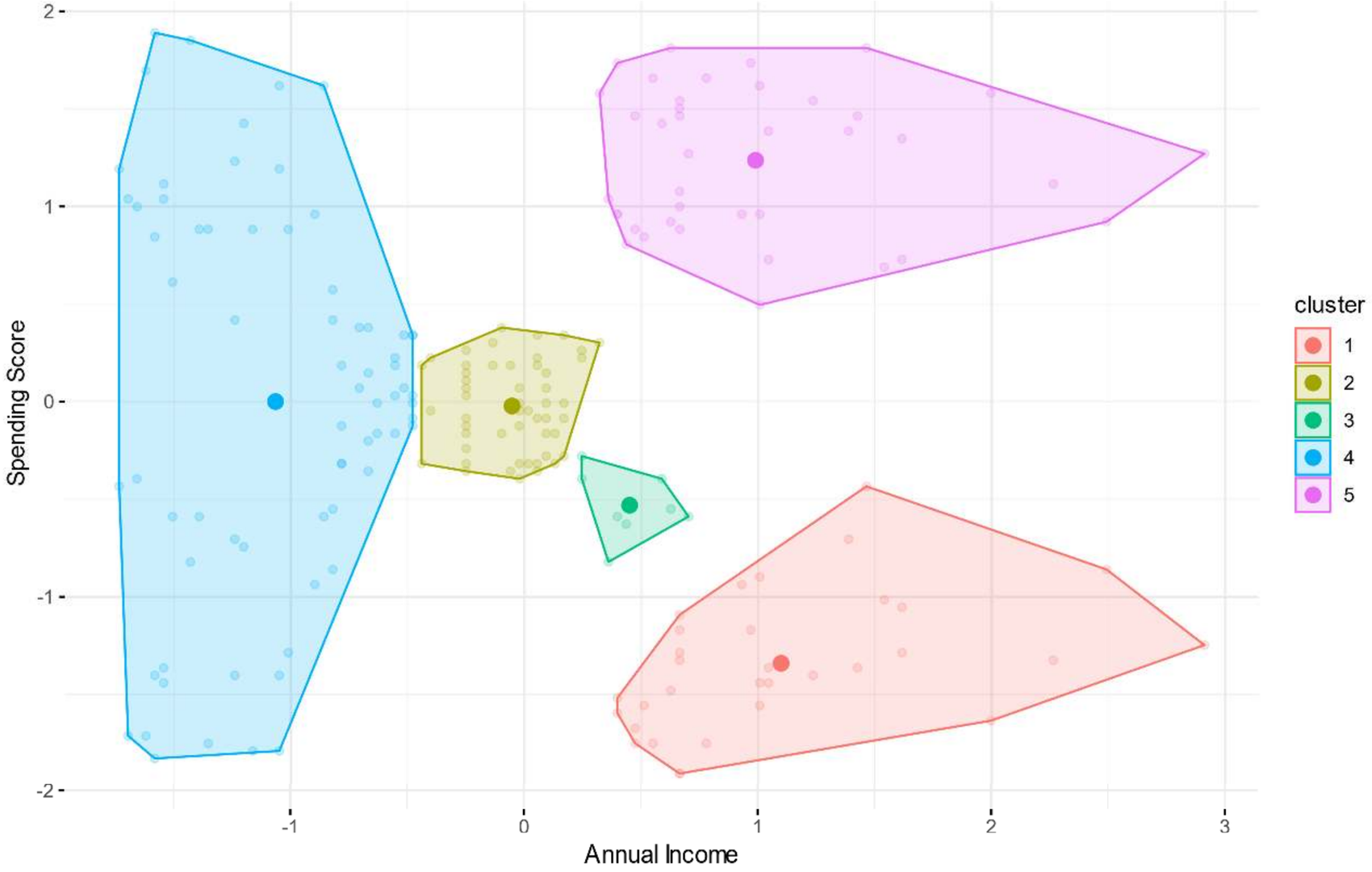
Determining the clustroid, i.e. the point being closest to all other points:

- Point with smallest maximum distance to other points
- Point with smallest average distance to other points
- More complicated notions

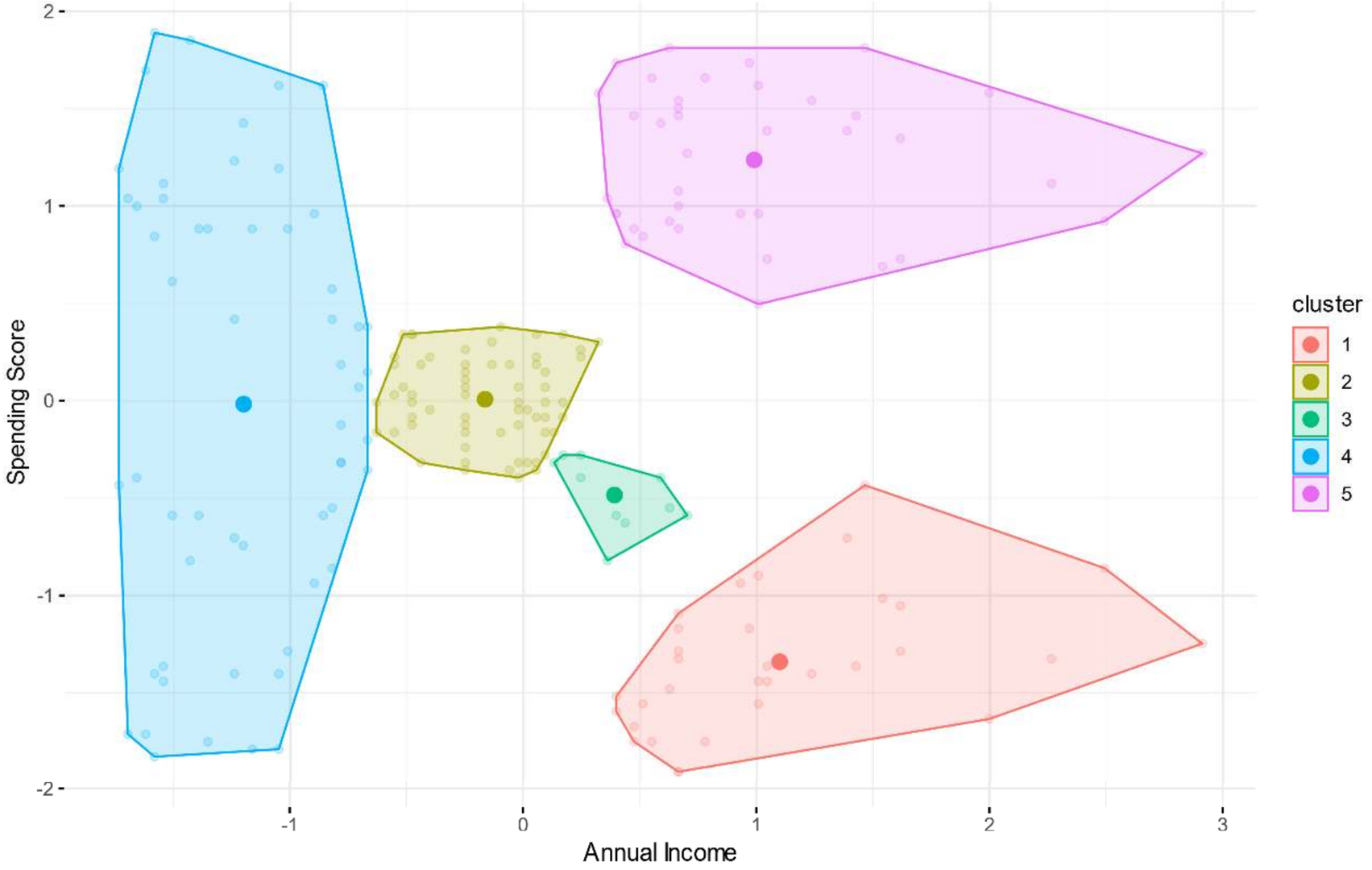
Examples: k-Means Clustering



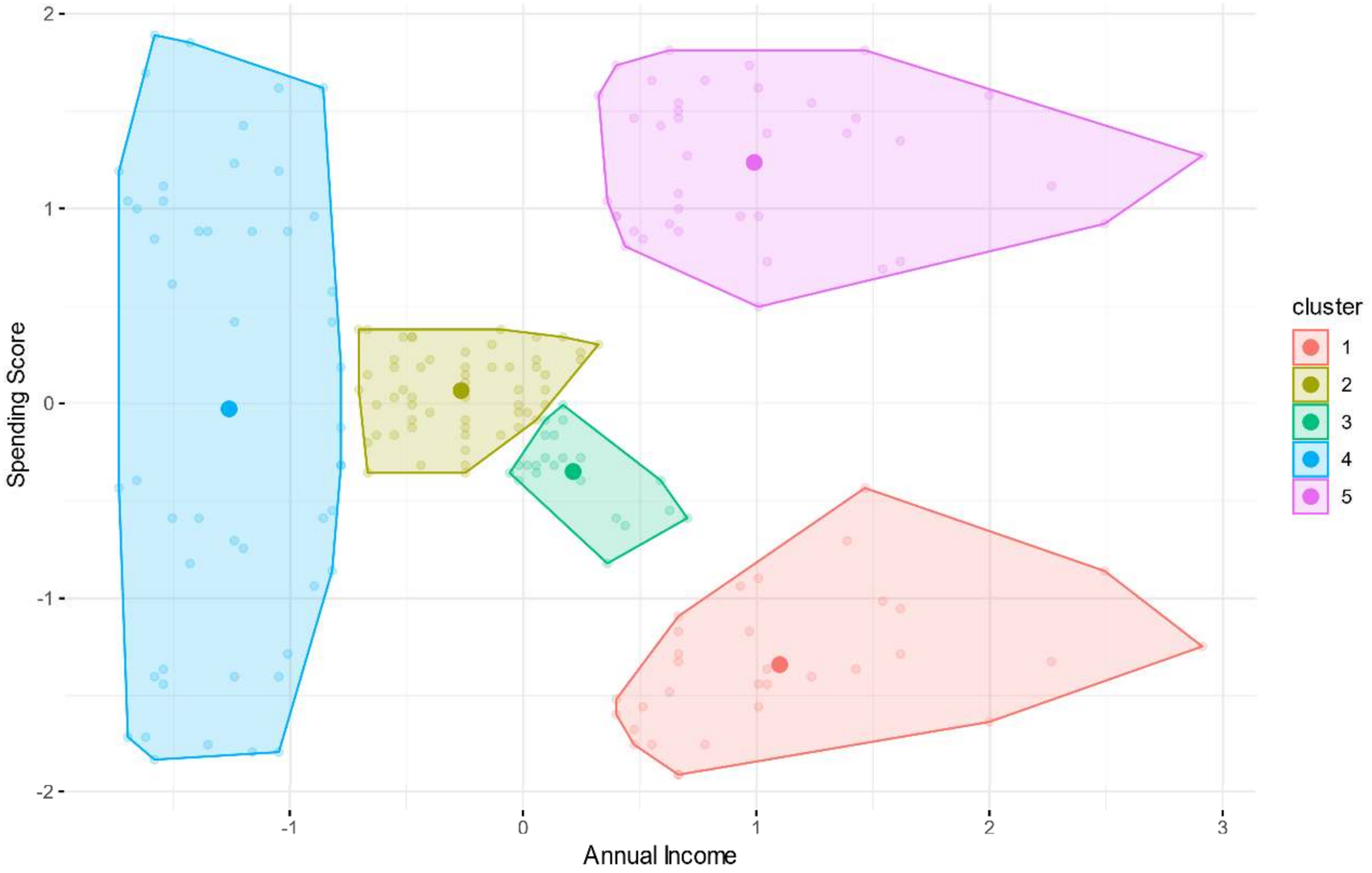
Examples: k-Means Clustering



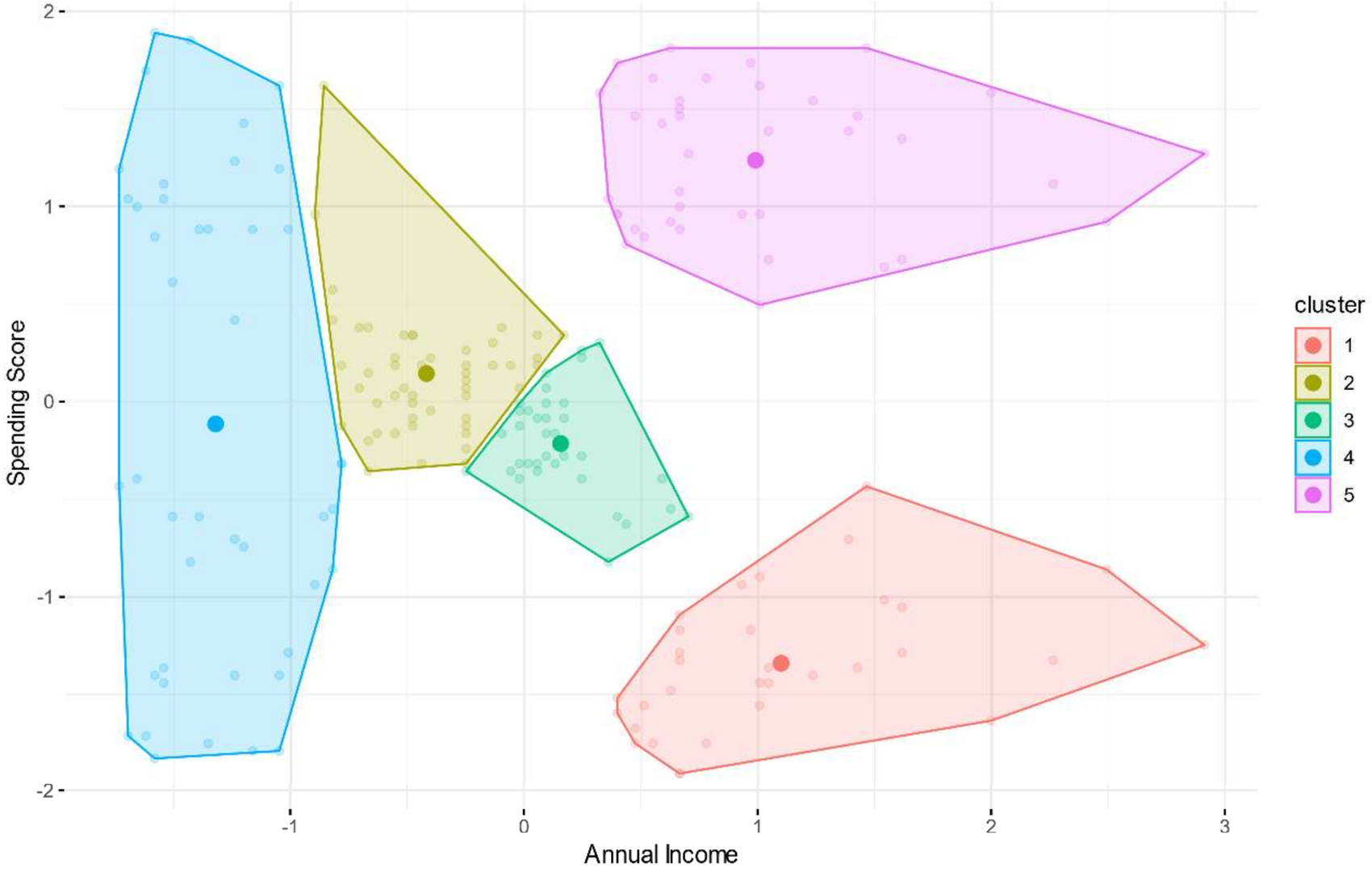
Examples: k-Means Clustering



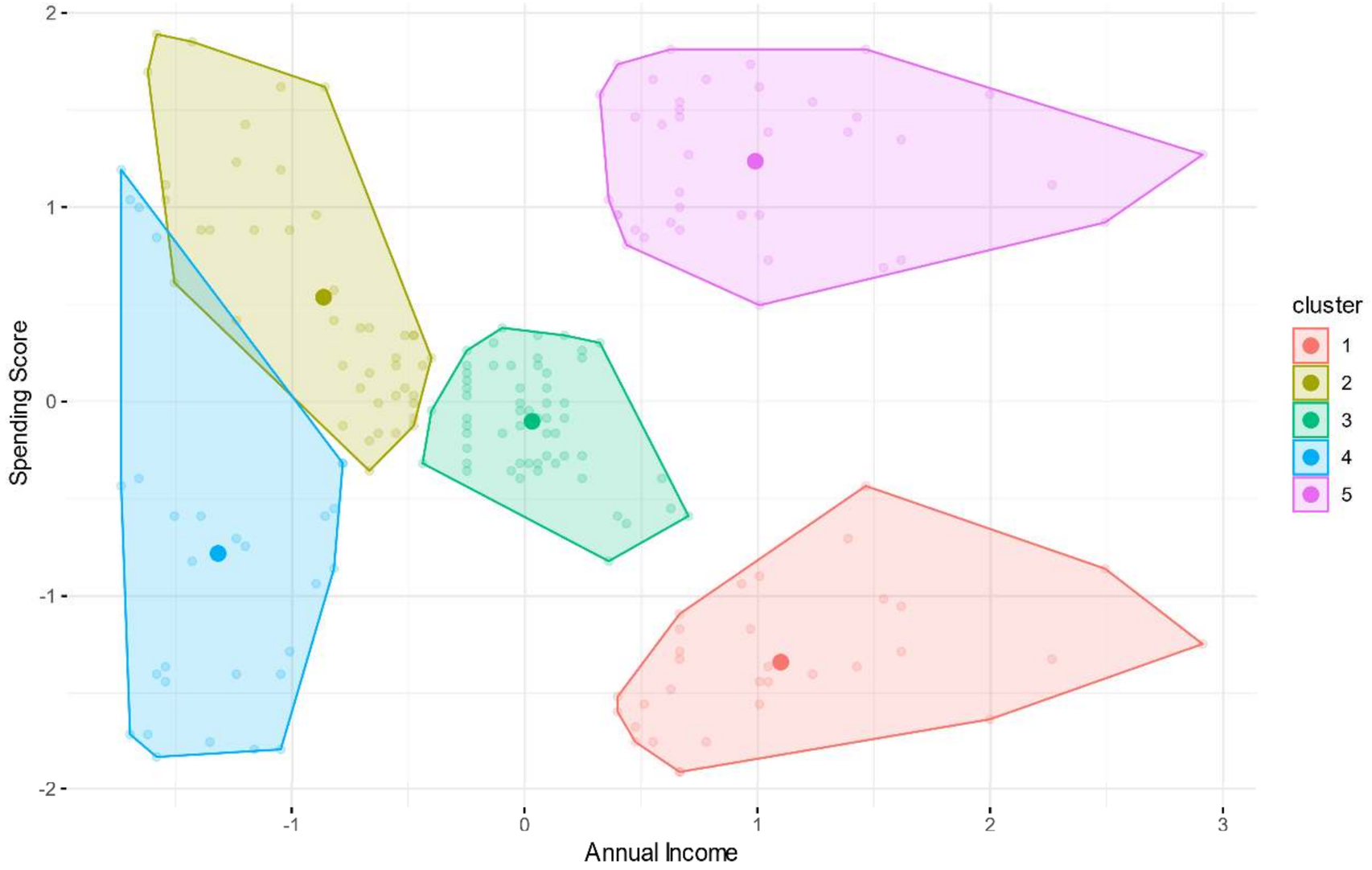
Examples: k-Means Clustering



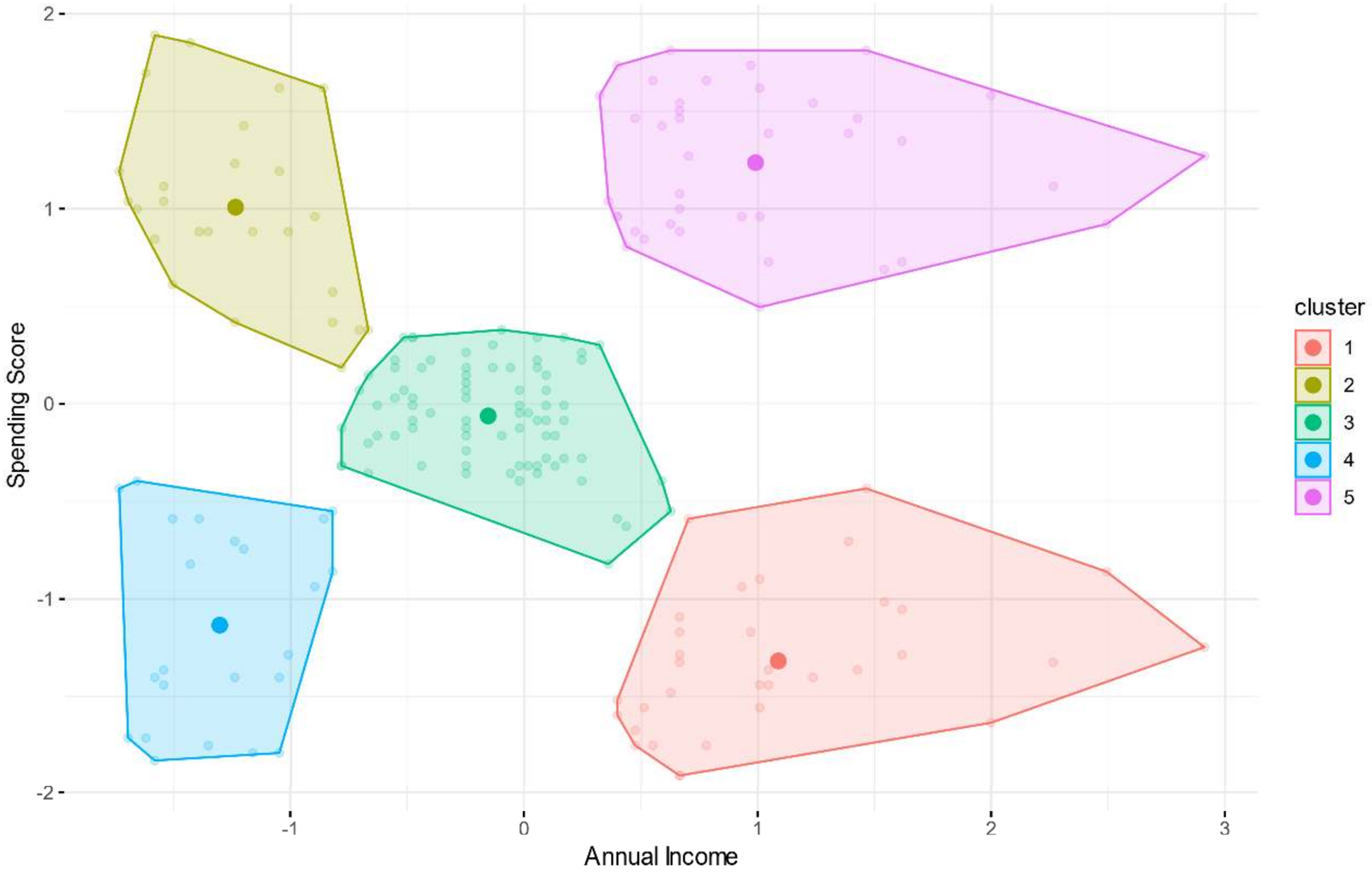
Examples: k-Means Clustering



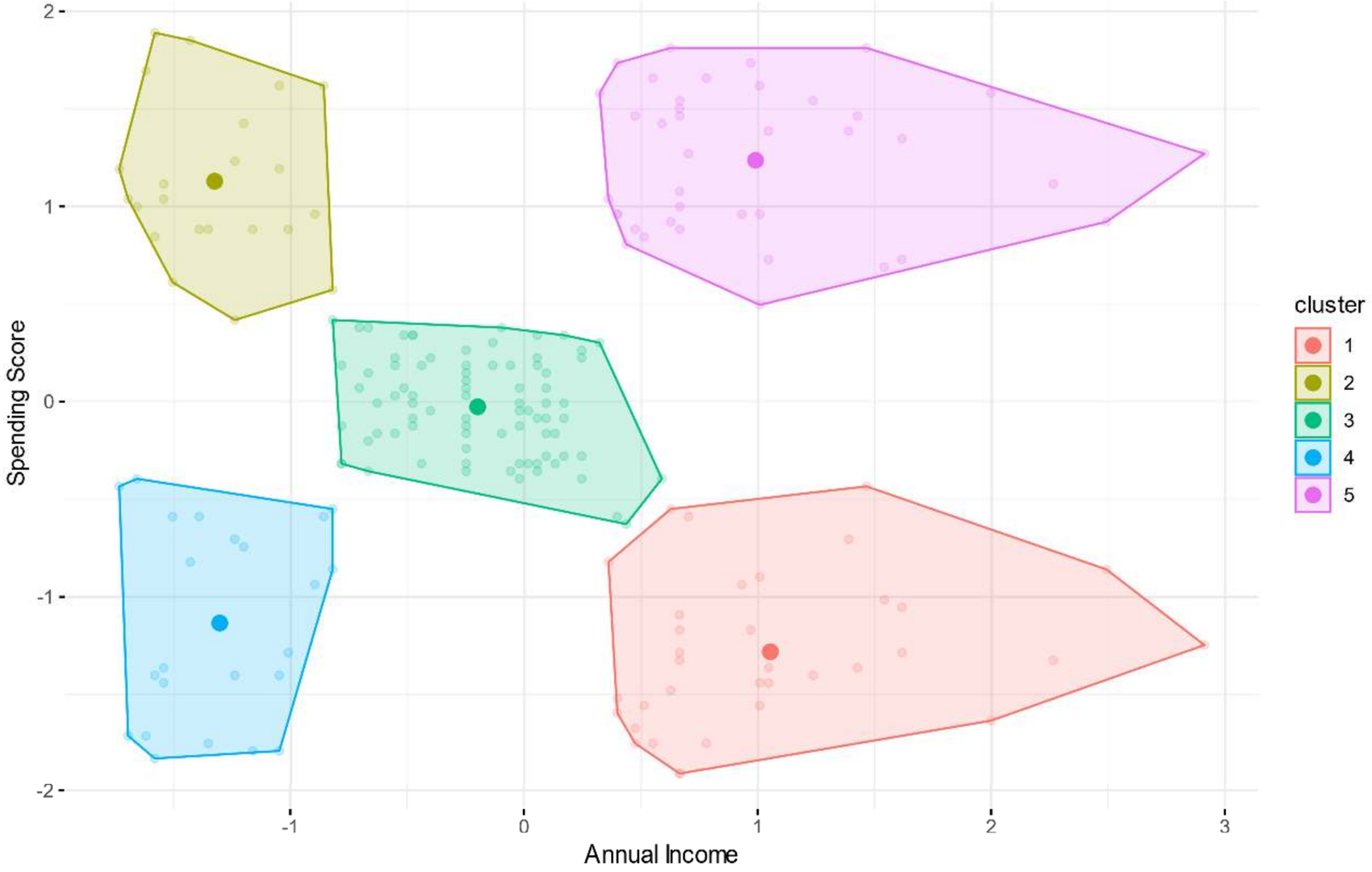
Examples: k-Means Clustering



Examples: k-Means Clustering



Examples: k-Means Clustering



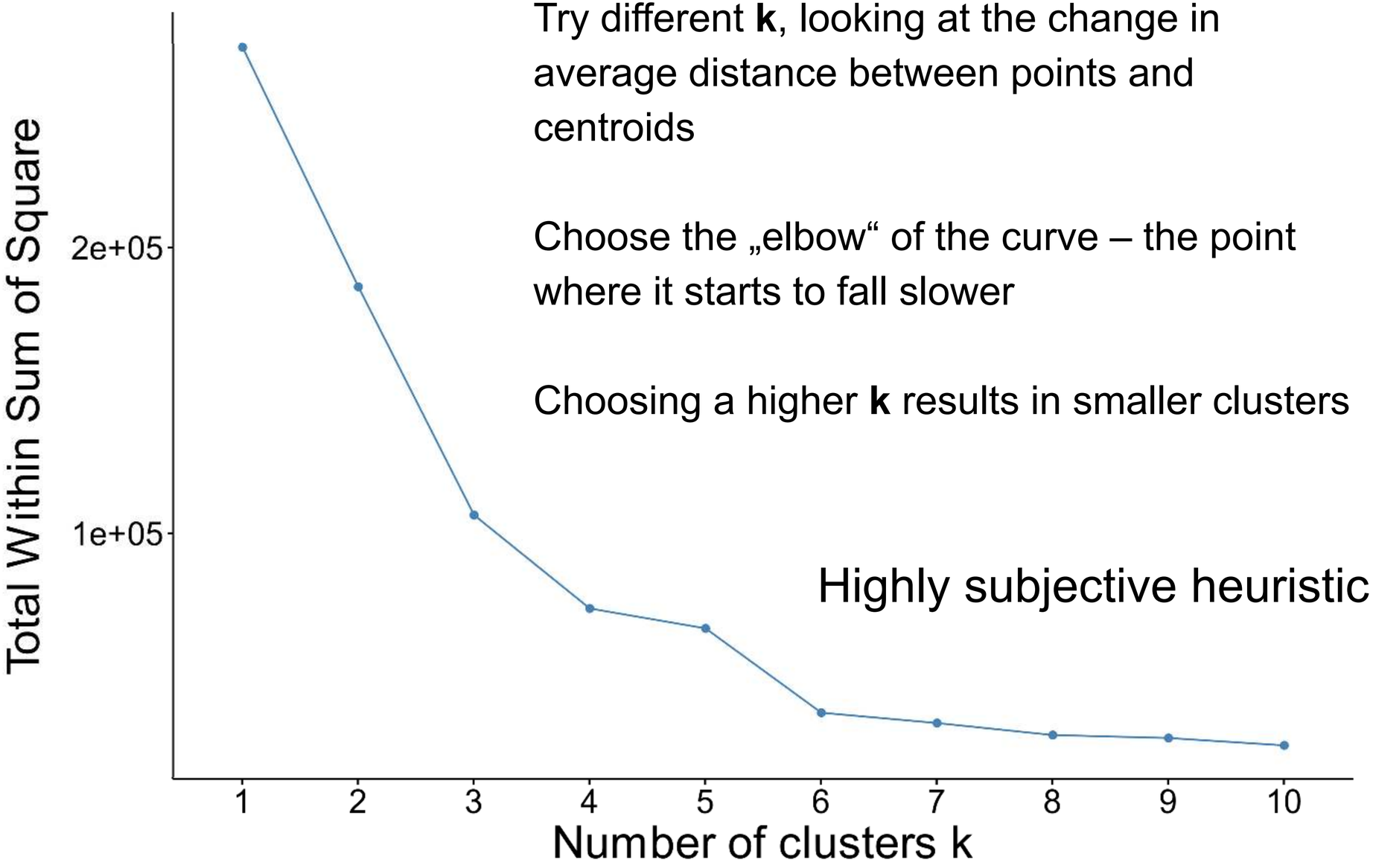
How to choose *k*: Elbow Criterion

Try different **k**, looking at the change in average distance between points and centroids

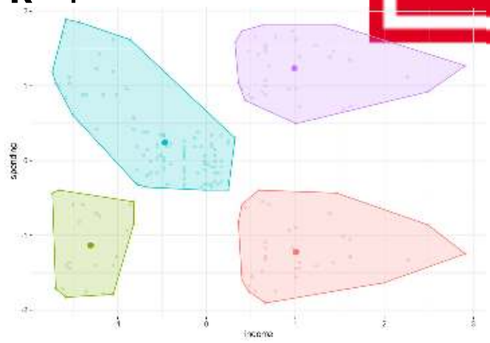
Choose the „elbow“ of the curve – the point where it starts to fall slower

Choosing a higher **k** results in smaller clusters

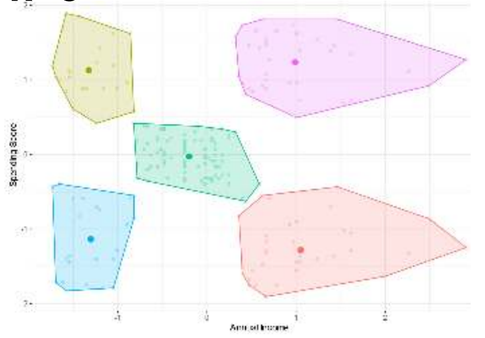
Highly subjective heuristic



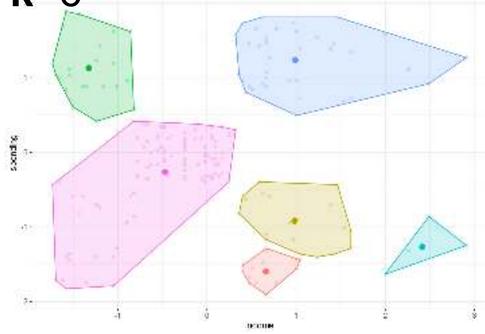
k=4



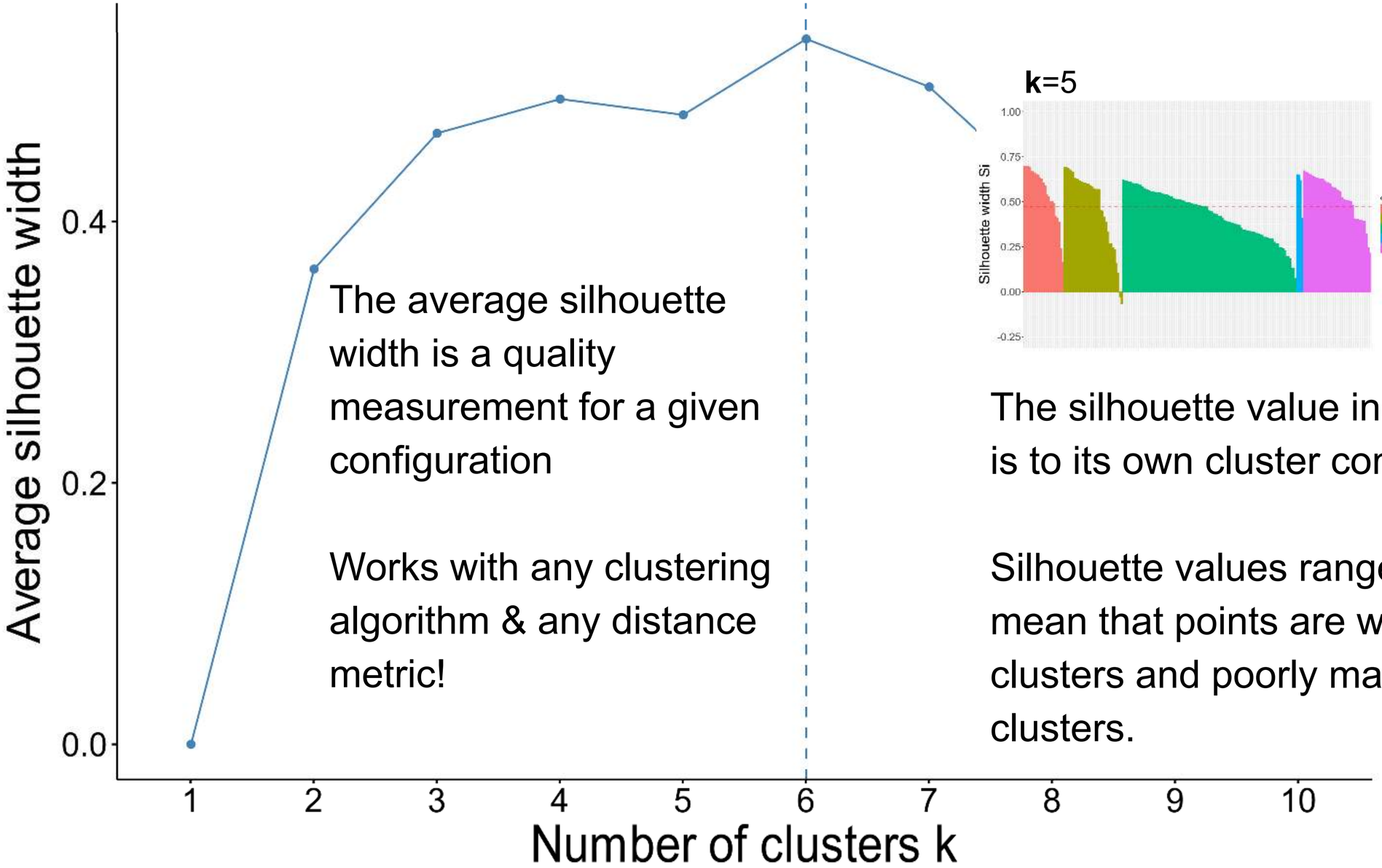
k=5



k=6



How to choose *k*: Silhouettes

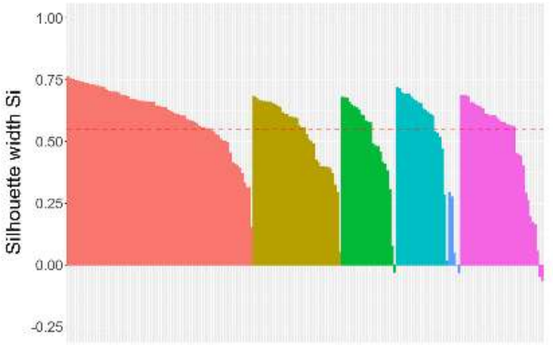


The average silhouette width is a quality measurement for a given configuration

Works with any clustering algorithm & any distance metric!

k=5

k=6



The silhouette value indicates how similar a point is to its own cluster compared to other clusters.

Silhouette values range from -1 to +1. High values mean that points are well matched to their own clusters and poorly matched to neighboring clusters.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.