



Unsupervised Learning: k-means

Mirco Schönfeld
mirco.schoenfeld@uni-bayreuth.de



Strategies of Clustering

Hierarchical Agglomerative Clustering

Each point is in its own cluster

Clusters are combined based on their “closeness”

Combination stops when undesirable clusters occur

Point assignment

Initial clusters are estimated

Points are considered in some order

Points are assigned to clusters into which they best fit



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..

Examples: k-Means Clustering



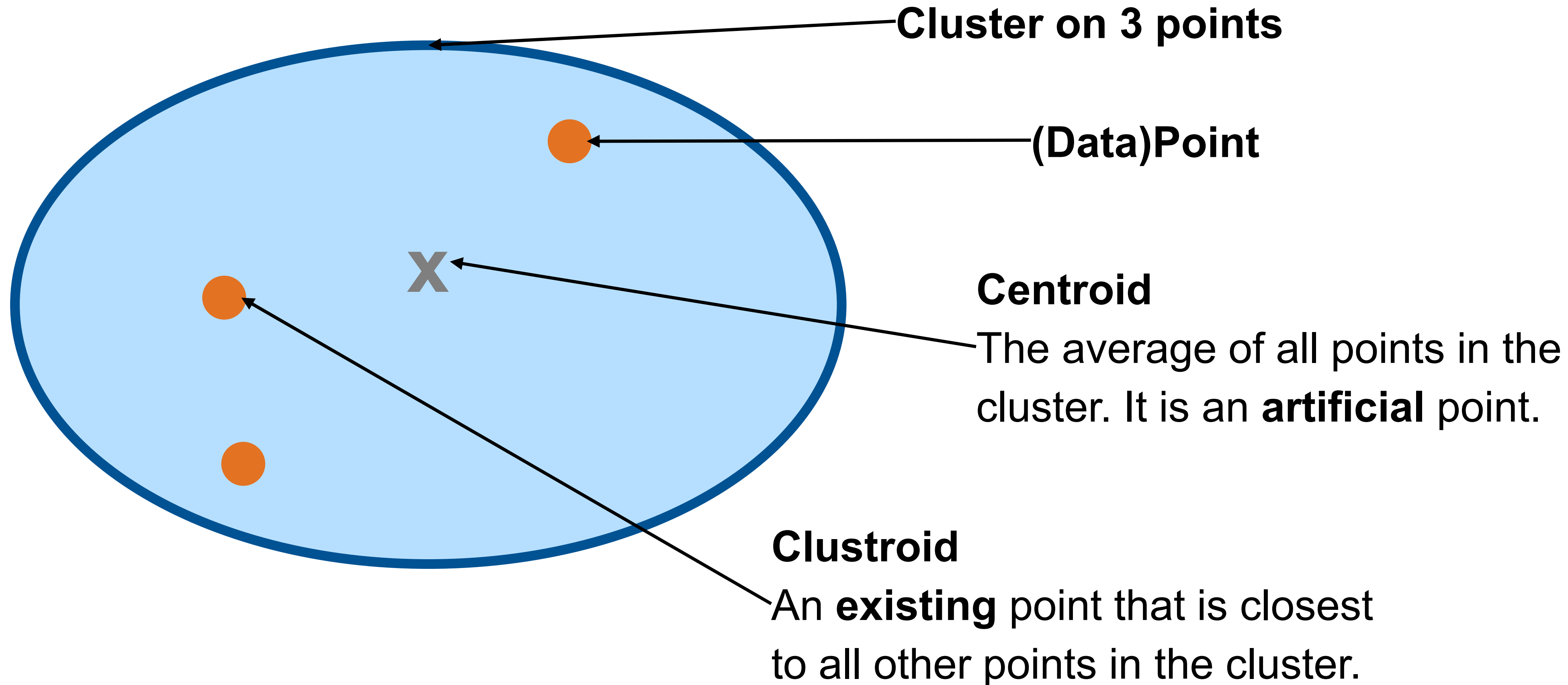
Clusters represented by their arithmetic mean

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



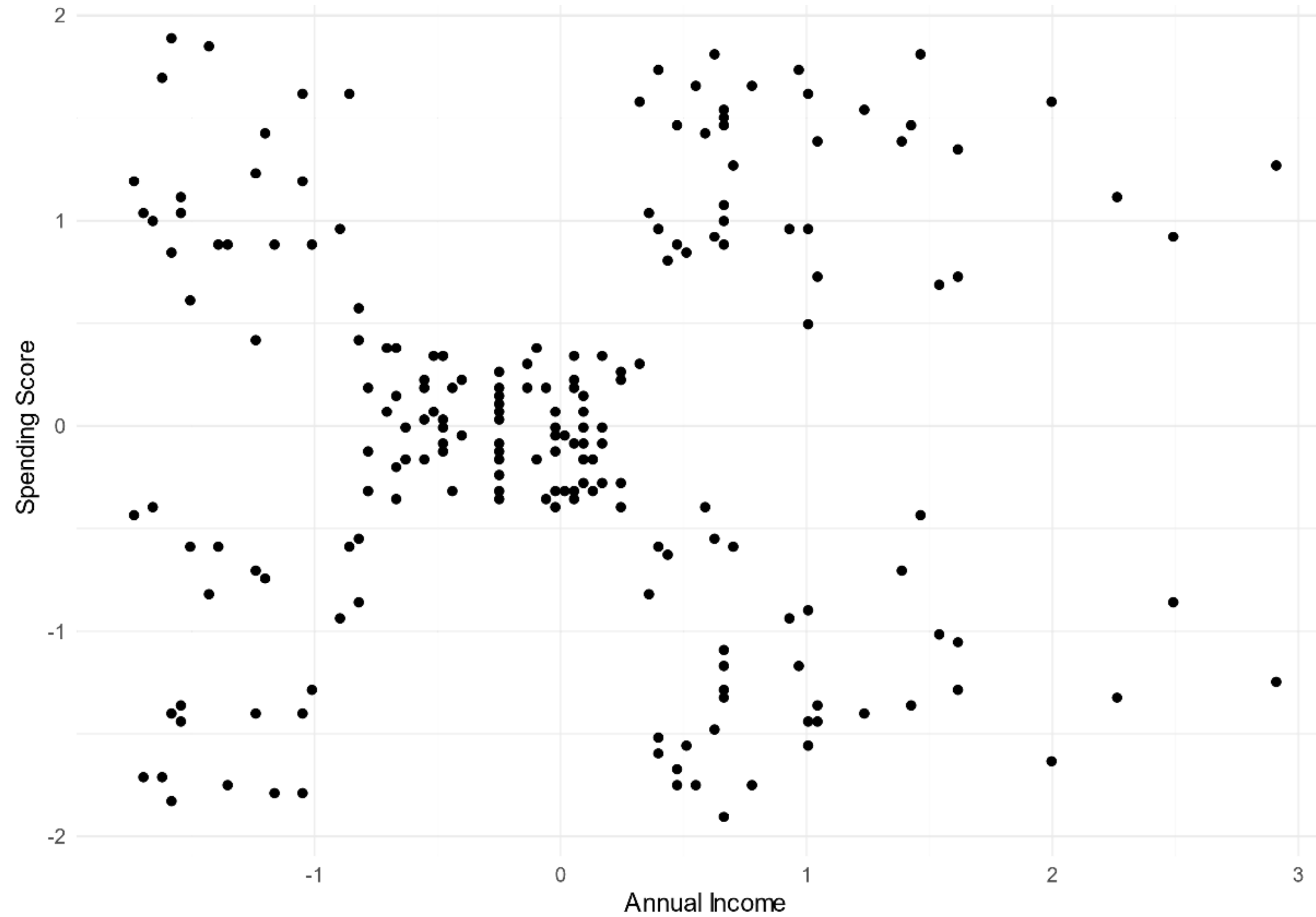
Centroids and Clustroids



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

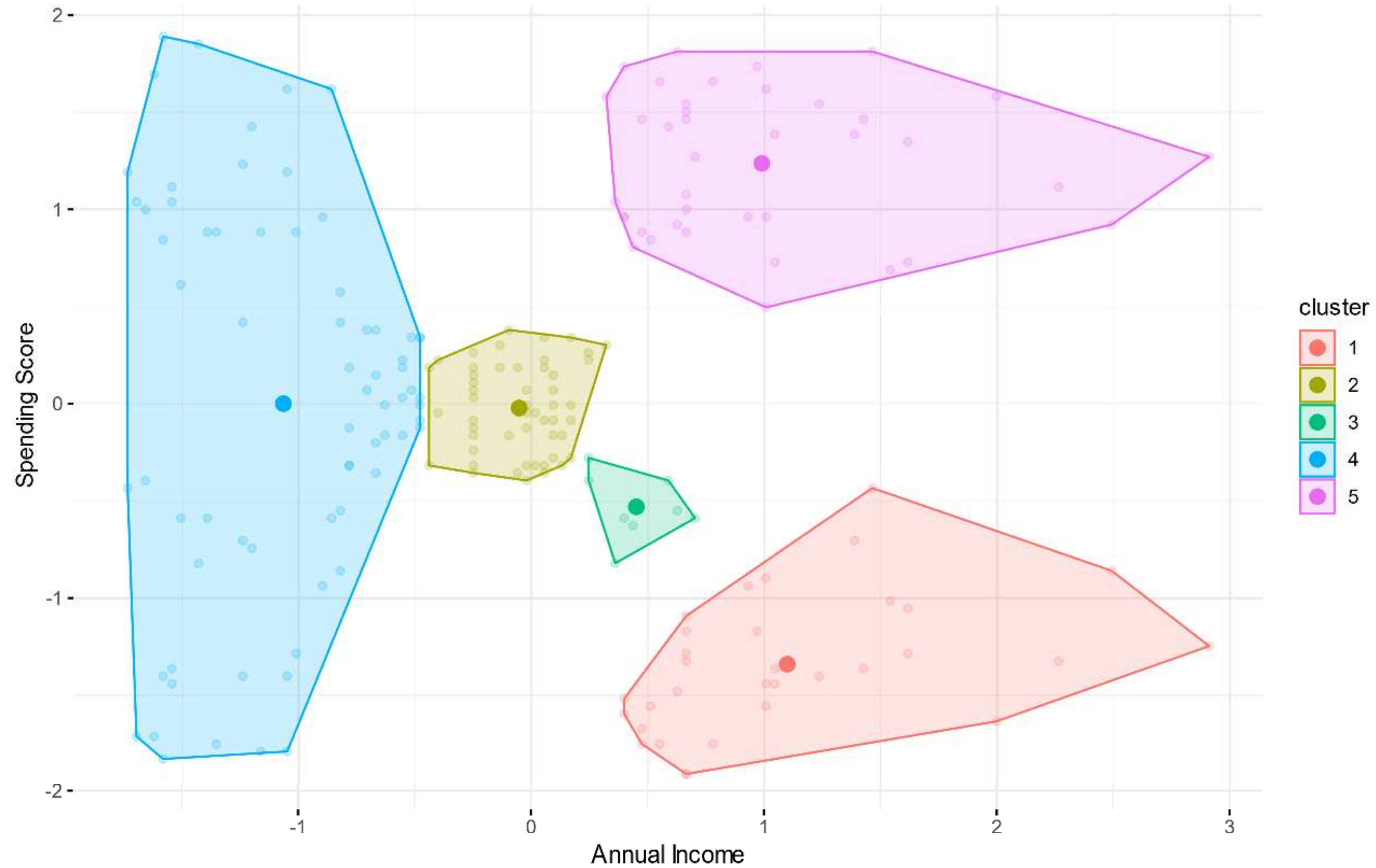
- Point with smallest maximum distance to other points
- Point with smallest average distance to other points
- More complicated notions

Examples: k-Means Clustering



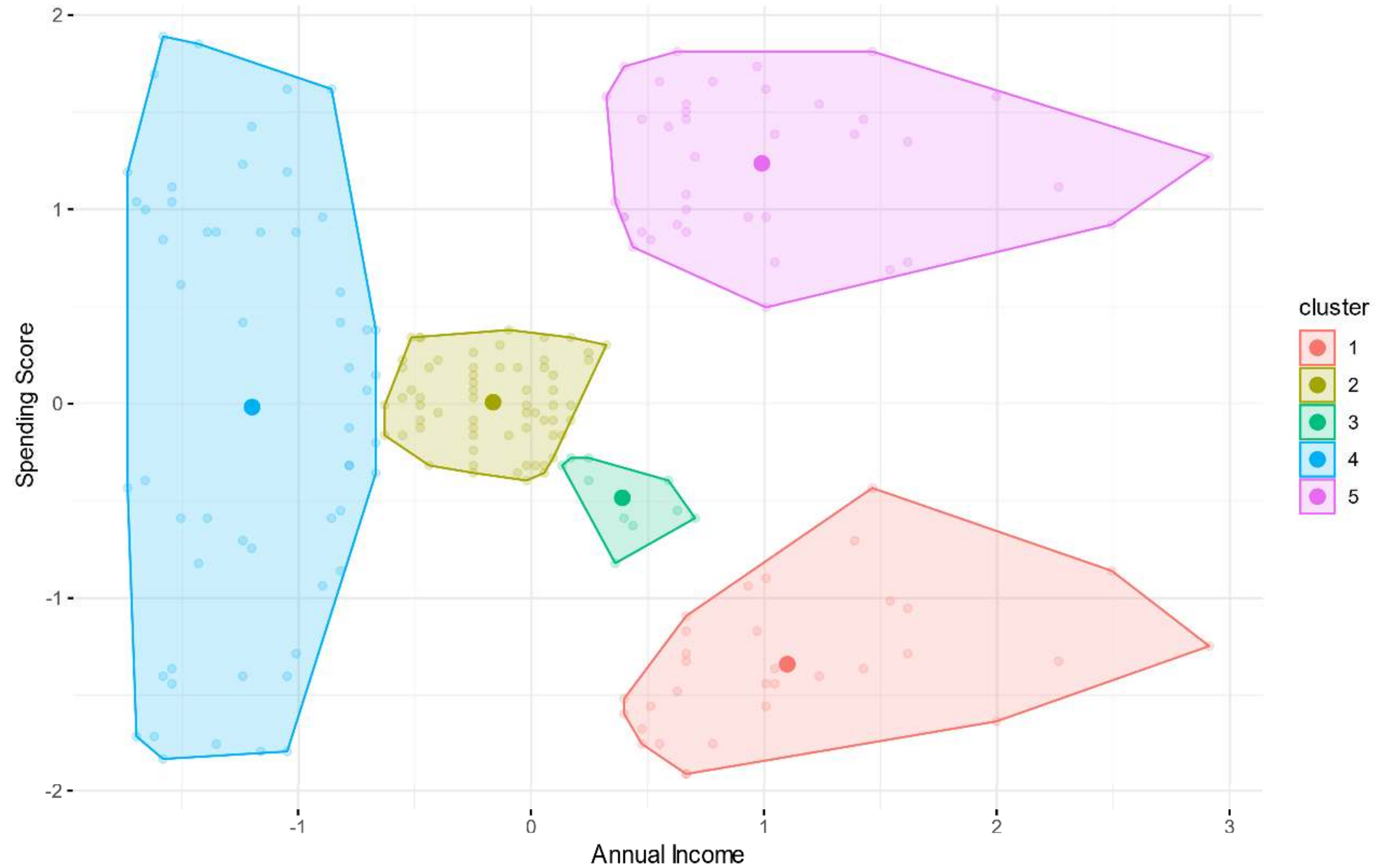


Examples: k-Means Clustering

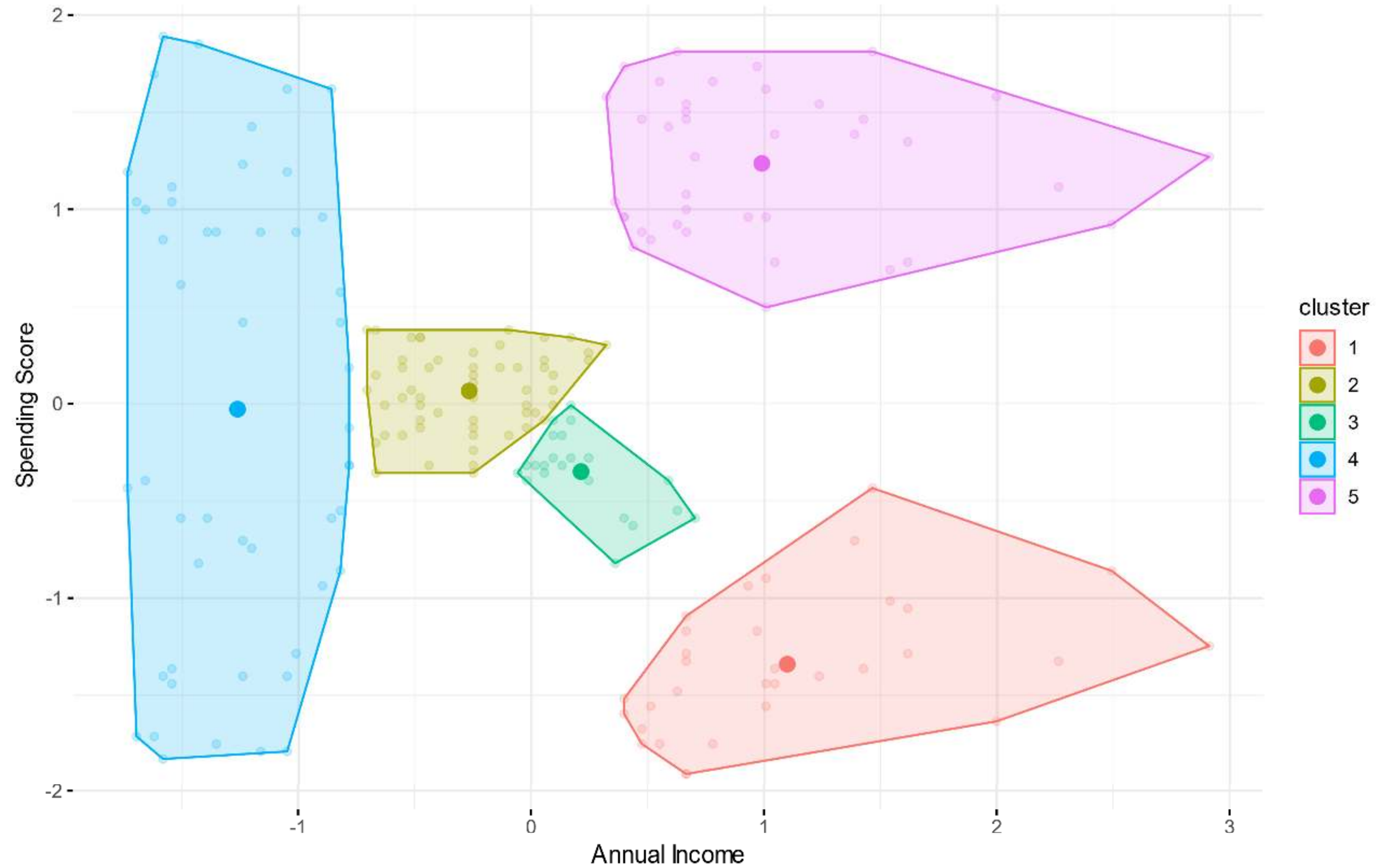




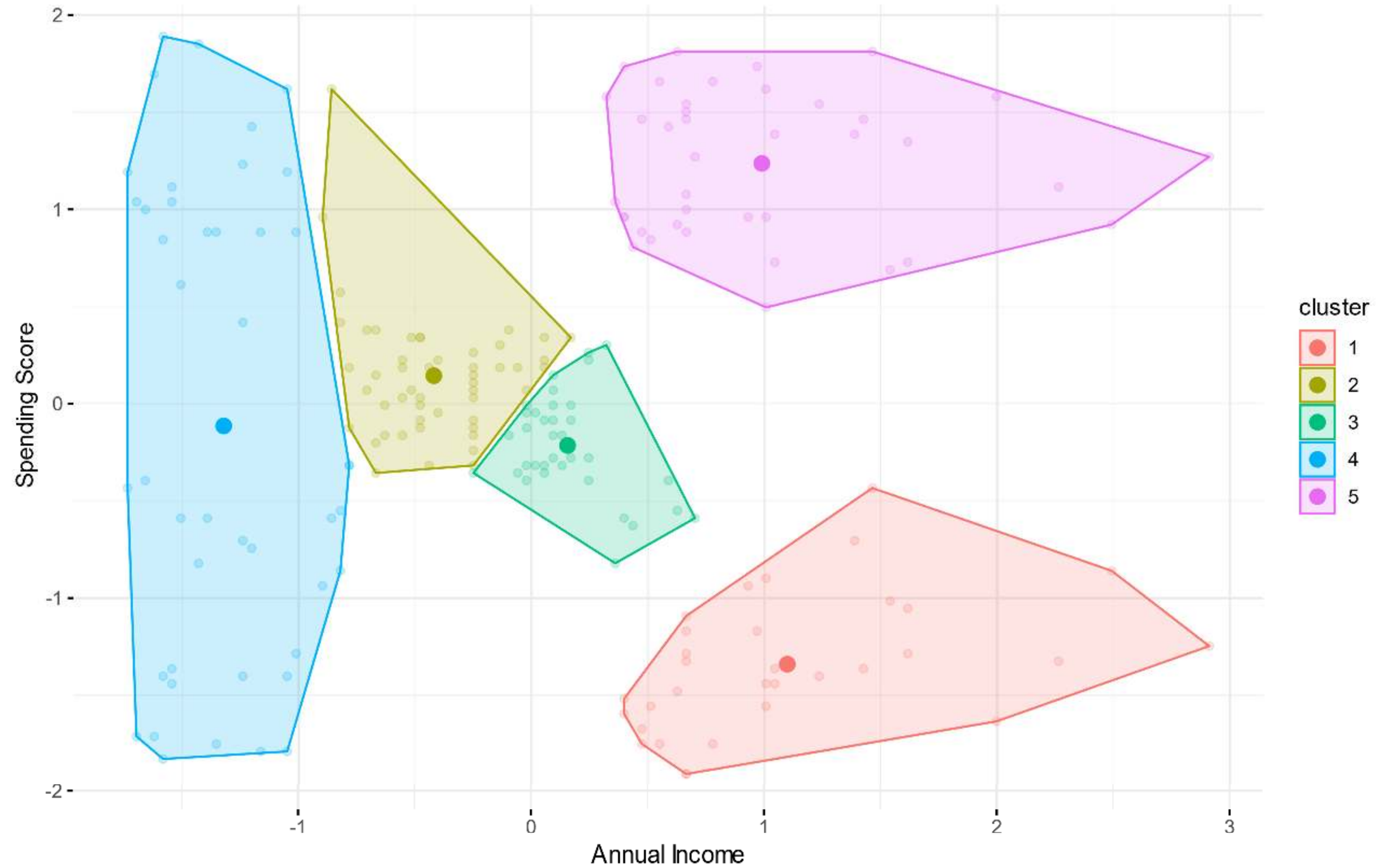
Examples: k-Means Clustering



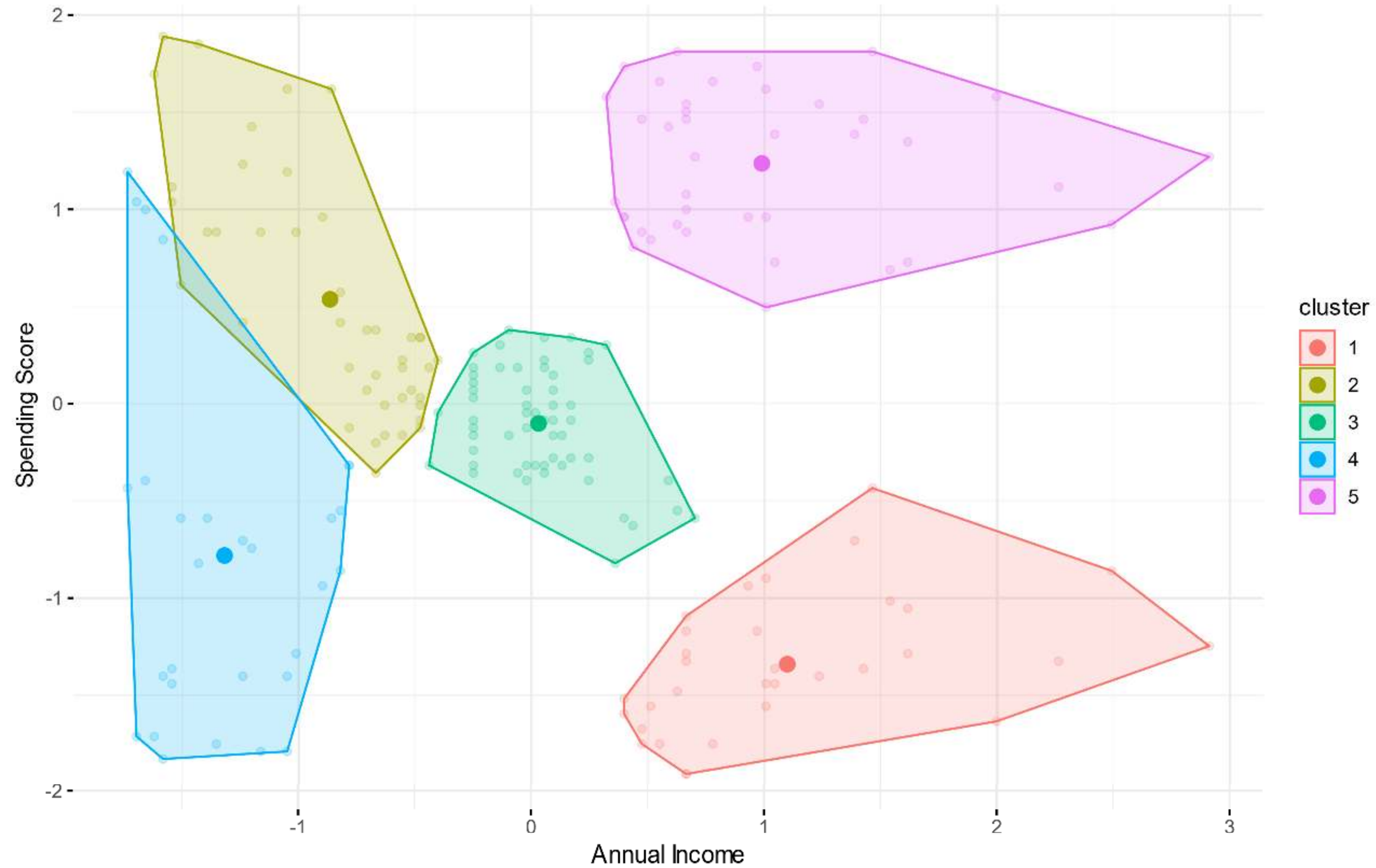
Examples: k-Means Clustering



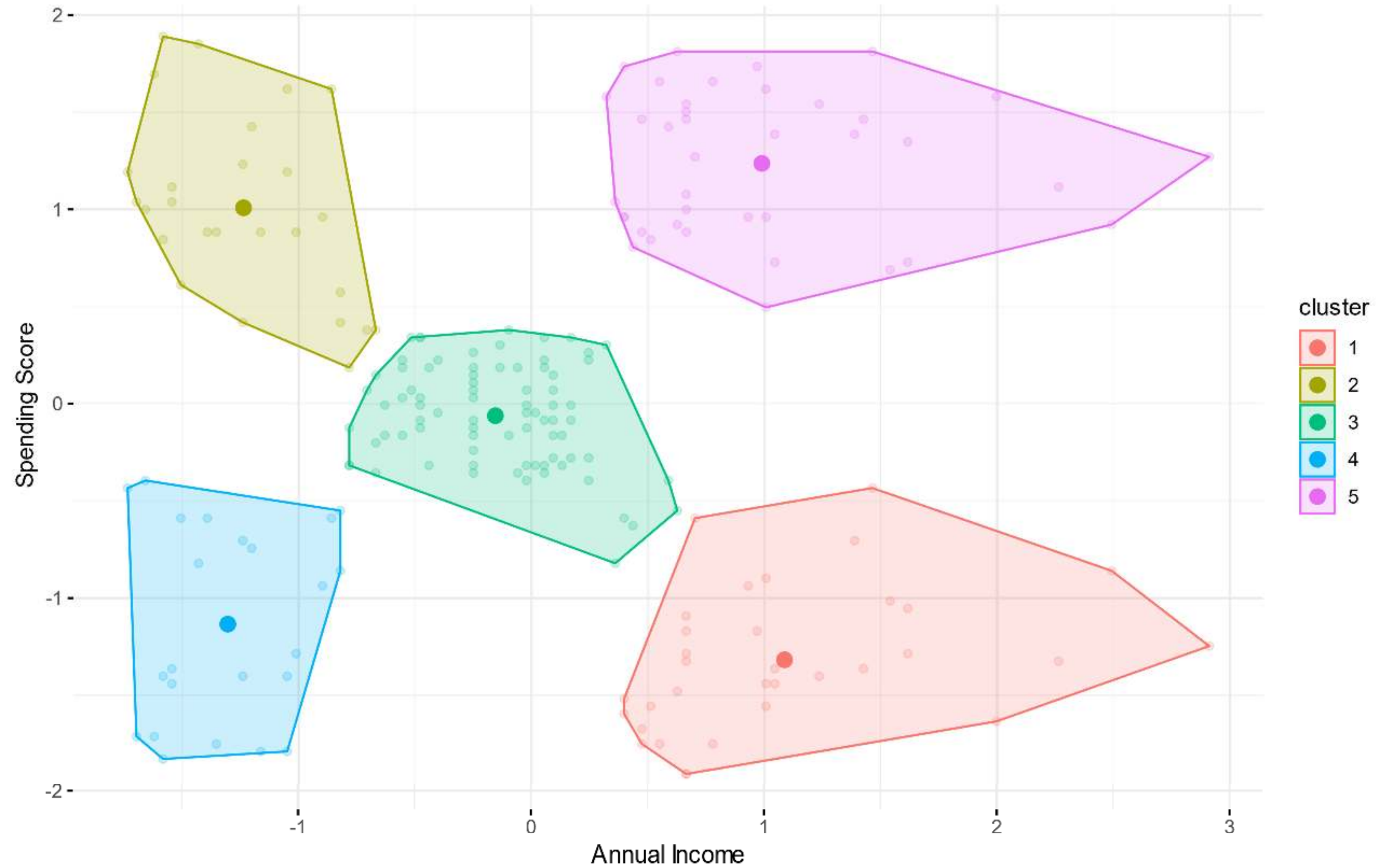
Examples: k-Means Clustering



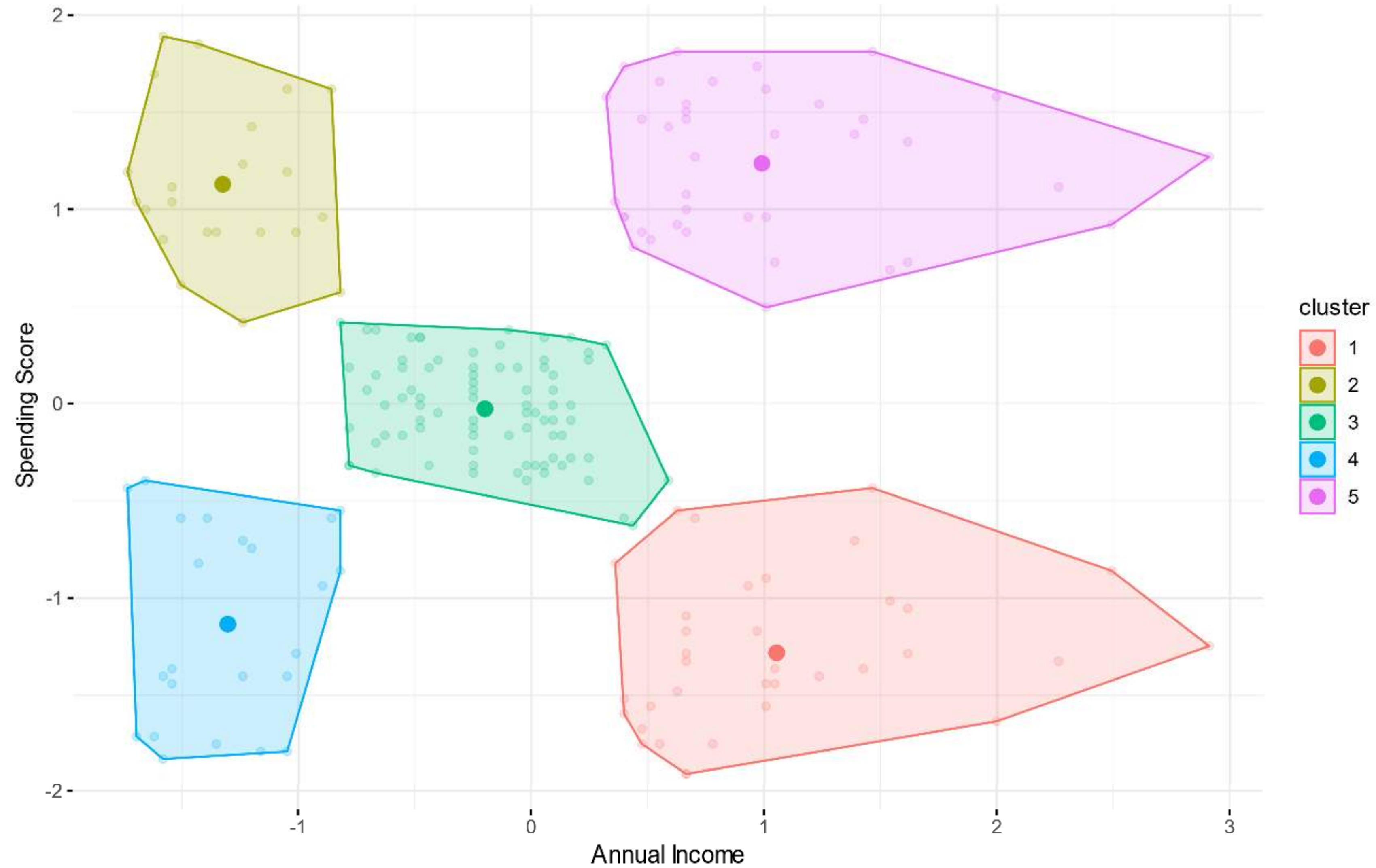
Examples: k-Means Clustering



Examples: k-Means Clustering



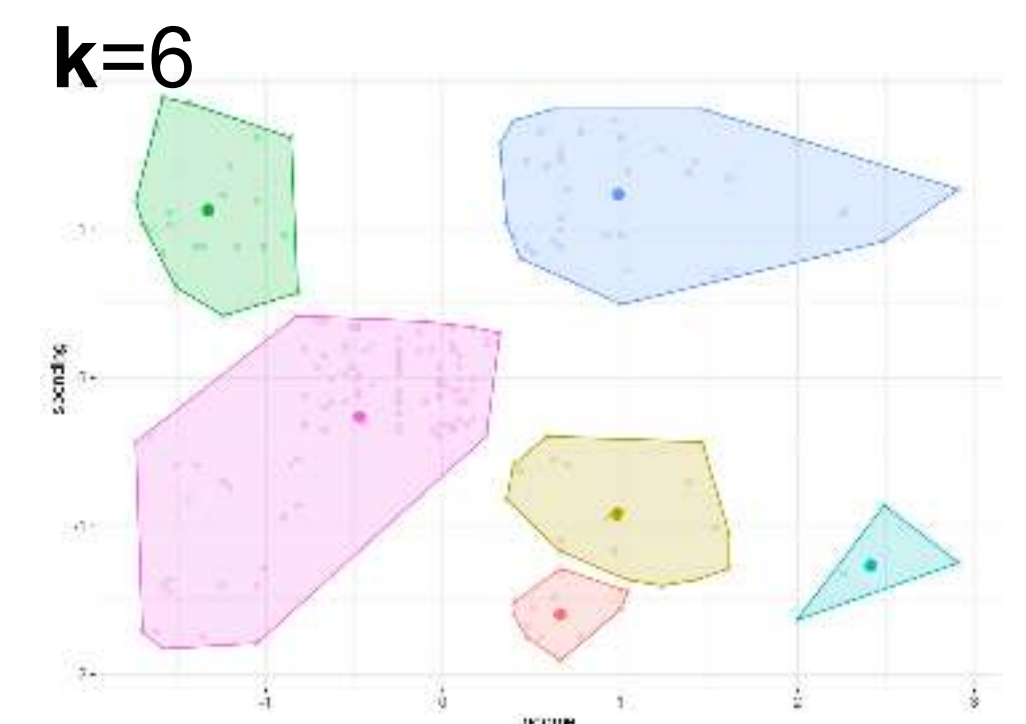
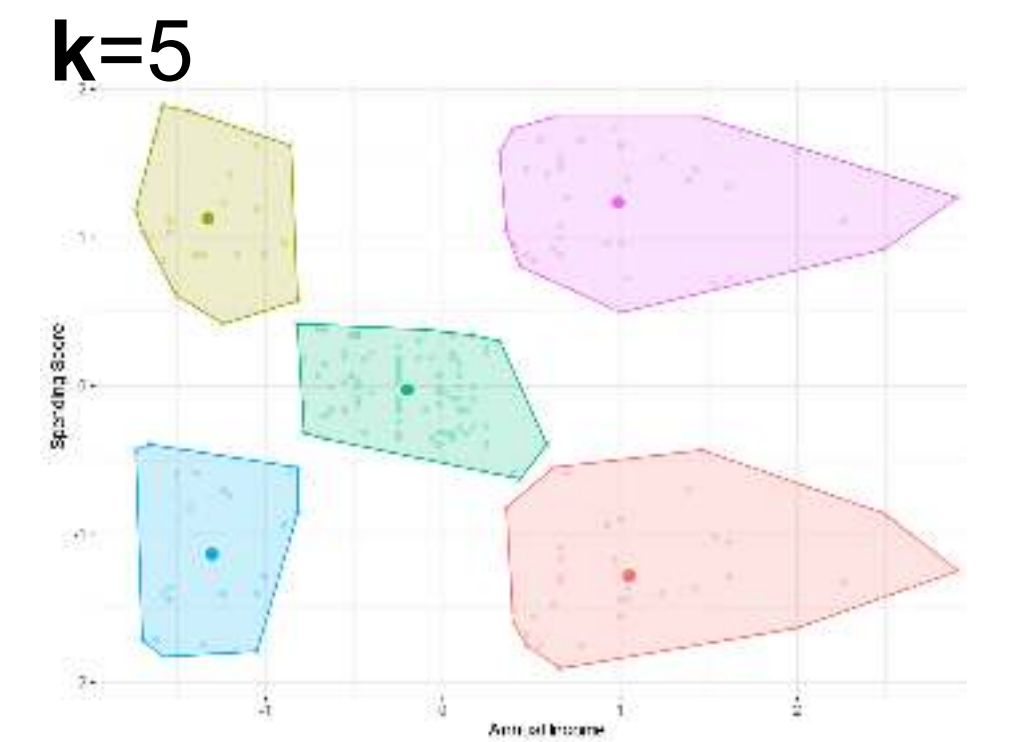
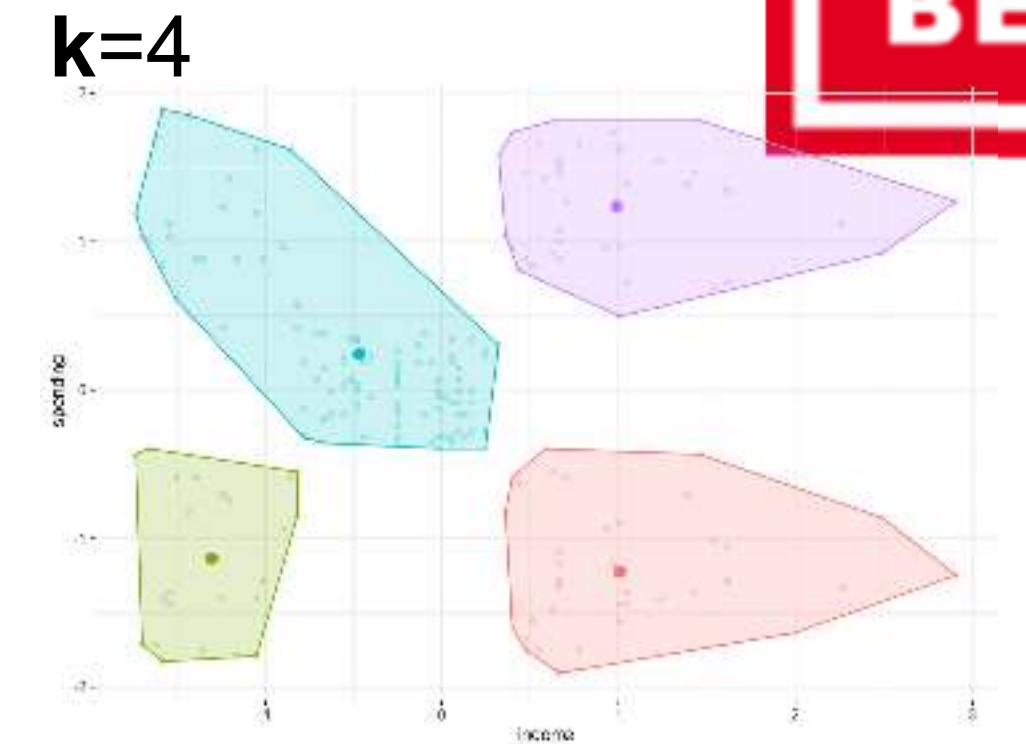
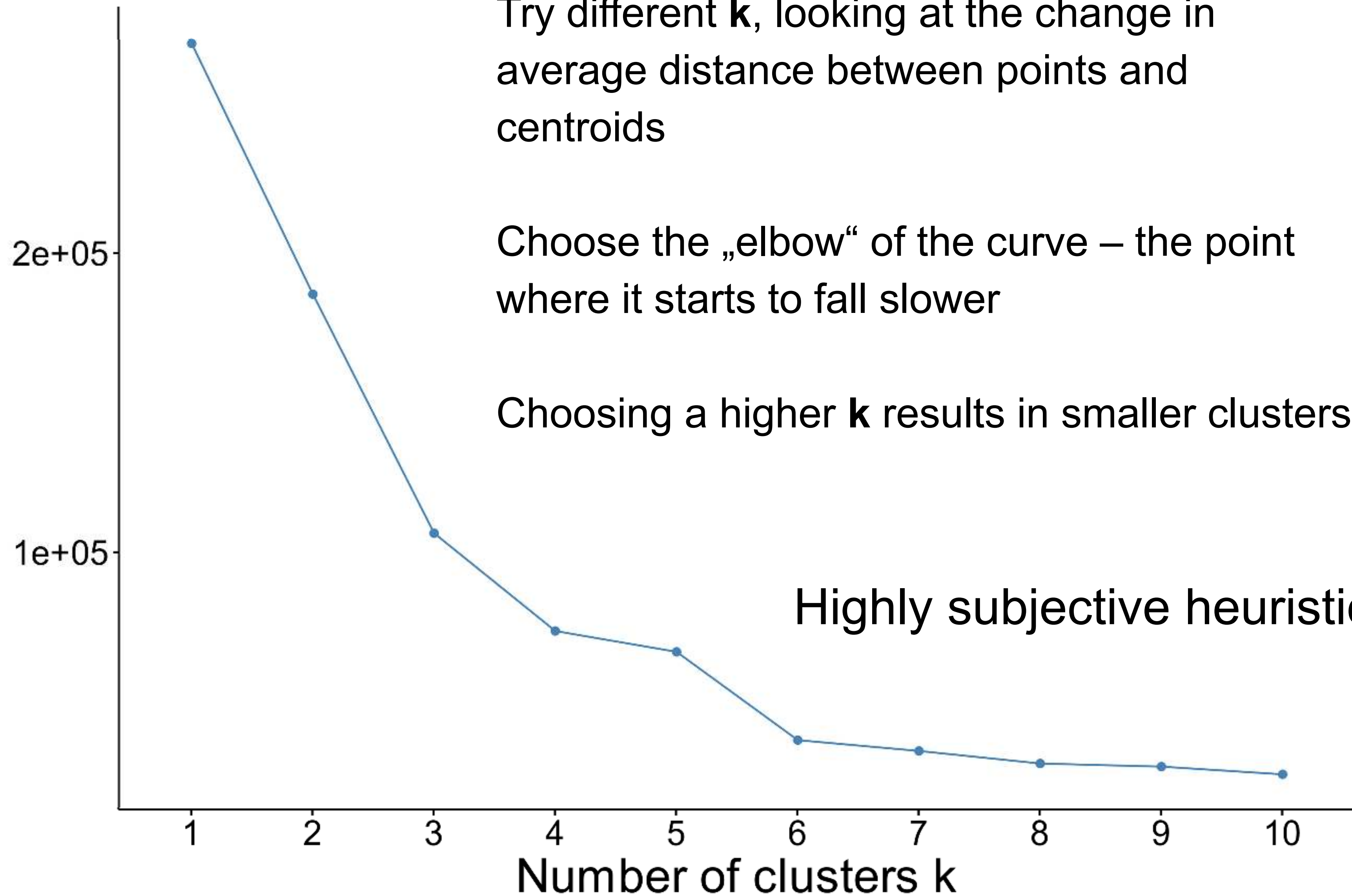
Examples: k-Means Clustering





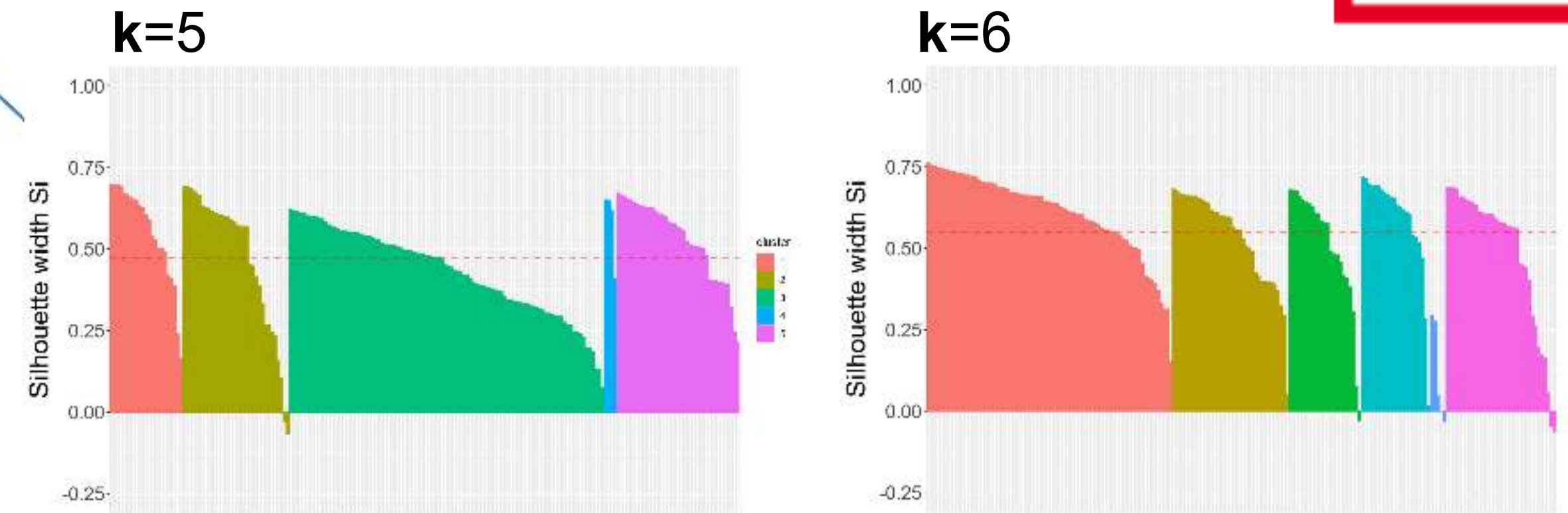
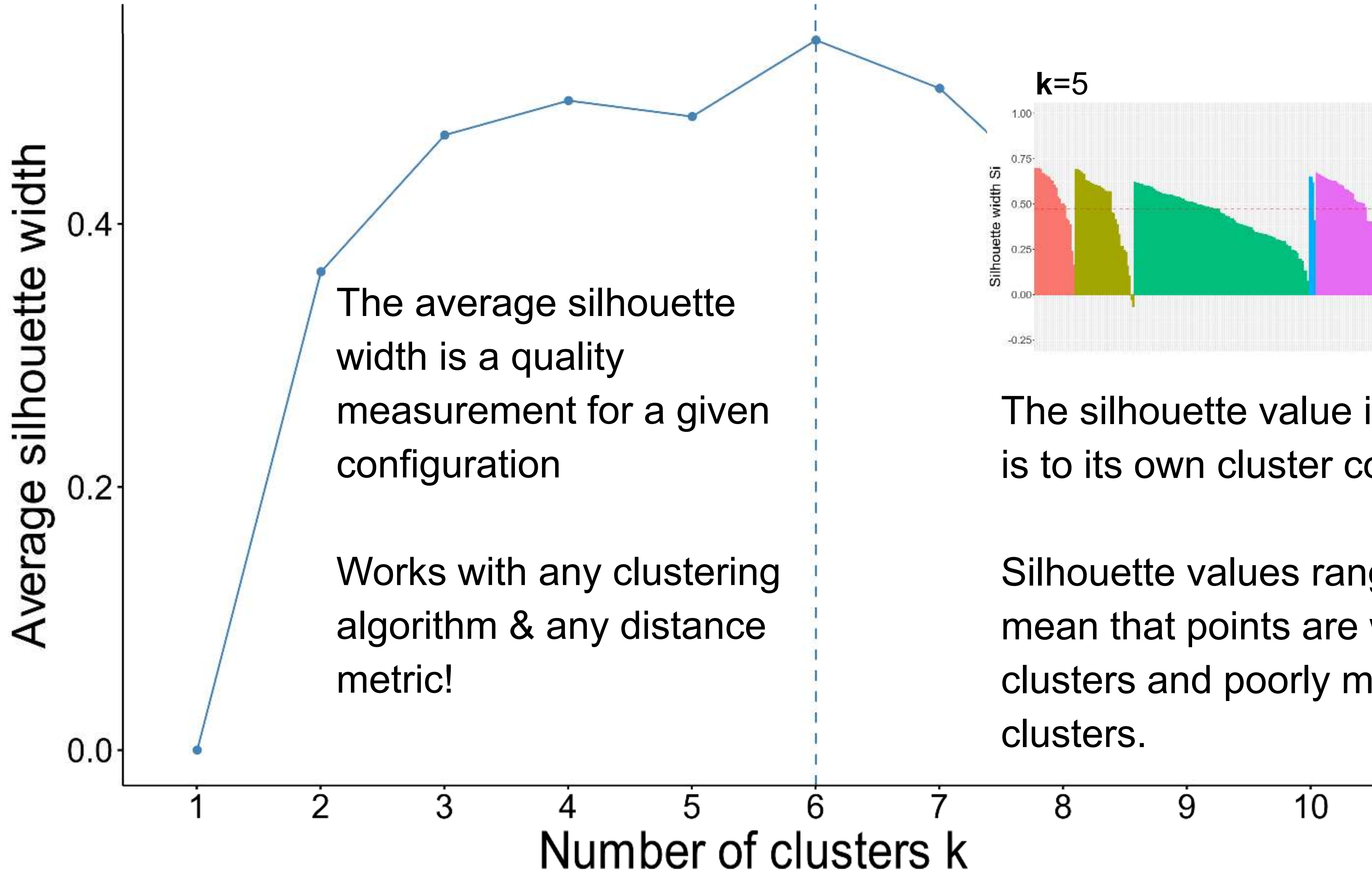
How to choose k : Elbow Criterion

Total Within Sum of Square





How to choose k : Silhouettes



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.