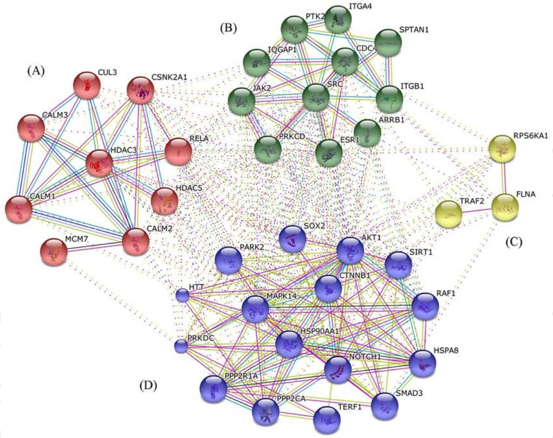
Module 0

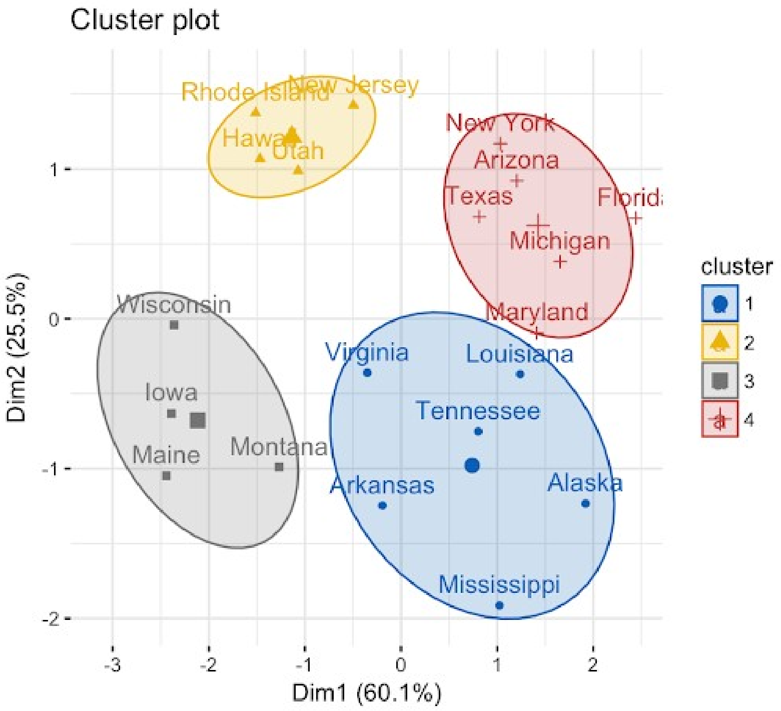
Matrix / Tensor Algebra

Cluster Analysis



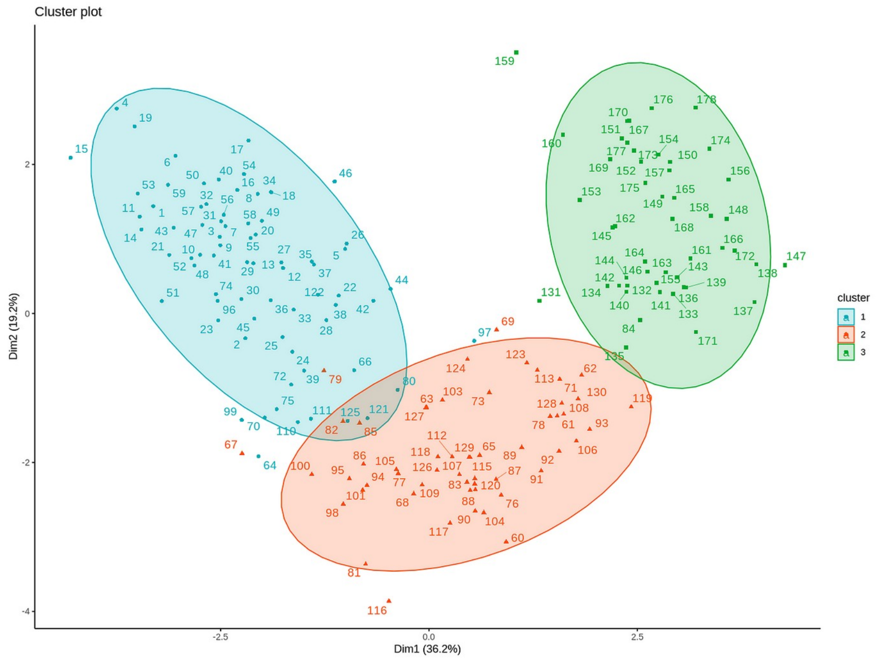
Cluster Analysis

- **Goal:** to segment the data into a set of homogenous groups (clusters) of observations



Measures of Distance

- ❑ We use distances to find out if observations are “alike”.
- ❑ We need to determine if observations with small distances to each other belong in the same group.



Euclidian Distance

▣ This is the most popular distance measure.

❑ We determine the Euclidian distance of two vectors by:

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2}$$

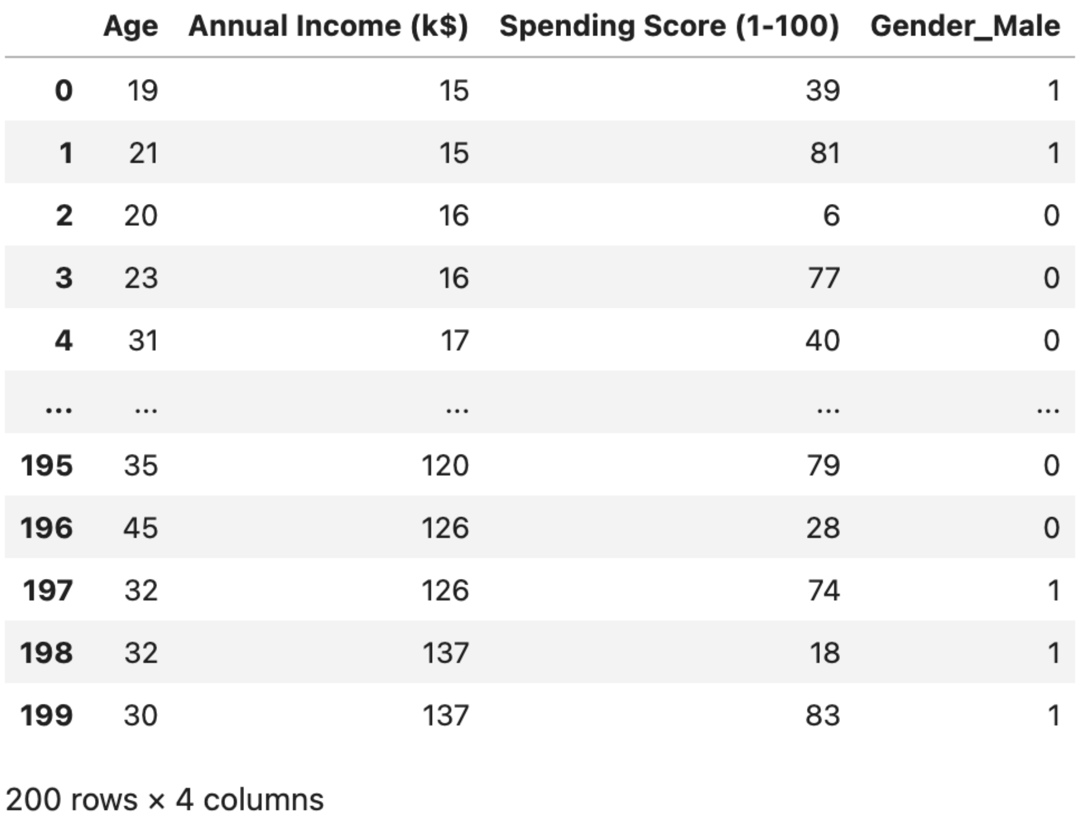
❑ We can also use vector / matrix algebra notation for vectors $\boldsymbol{x}$ and $\boldsymbol{z}$:

$$d = ||x - z||_2$$

Python

Euclidian Distance Example

- Clustering Customers

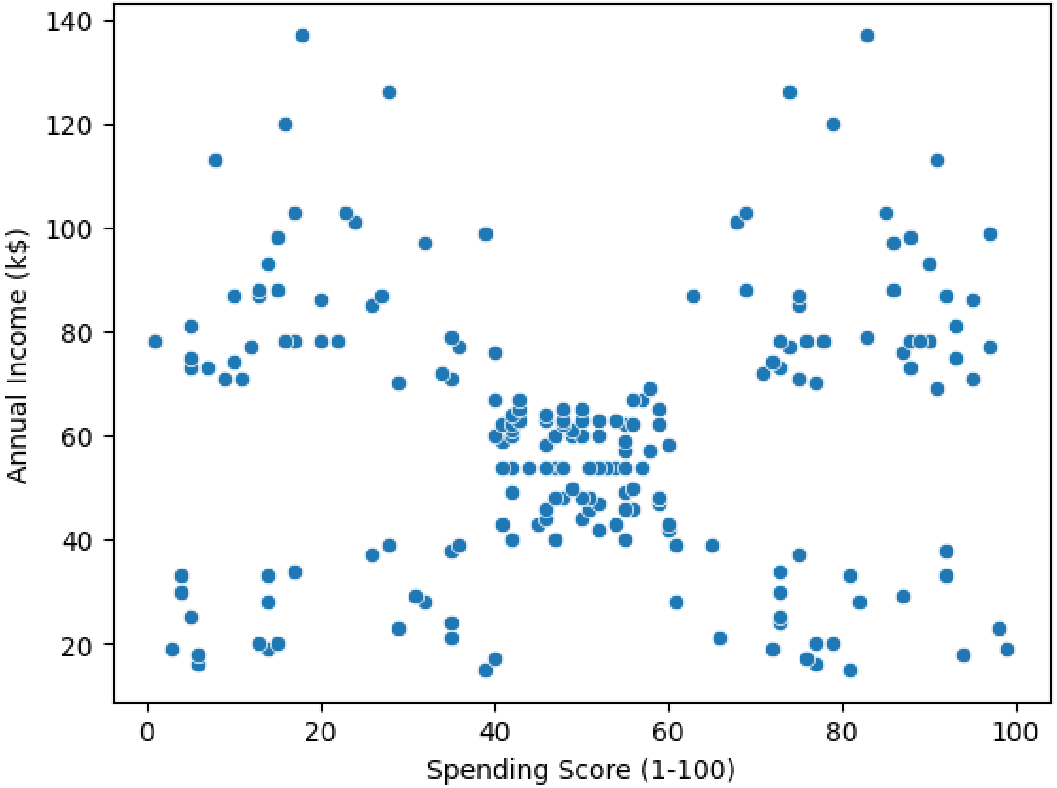


	Age	Annual Income (k$)	Spending Score (1-100)	Gender_Male
0	19	15	39	1
1	21	15	81	1
2	20	16	6	0
3	23	16	77	0
4	31	17	40	0
...	...	...	...	...
195	35	120	79	0
196	45	126	28	0
197	32	126	74	1
198	32	137	18	1
199	30	137	83	1

200 rows × 4 columns

Euclidian Distance Example

- Clustering Customers



Drawbacks of Euclidian Distance

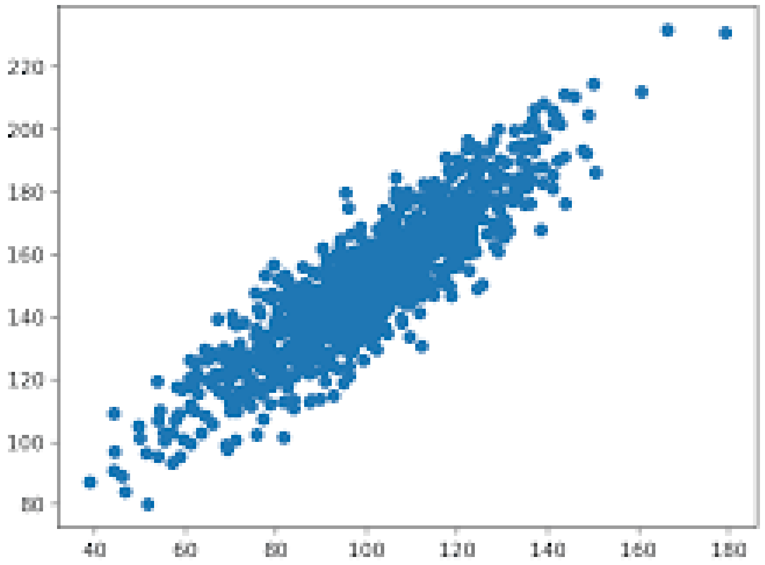
⊡ You must scale your data

❑ It completely ignores relationships between variables

❑ If you have outliers you might consider the Manhattan distance.

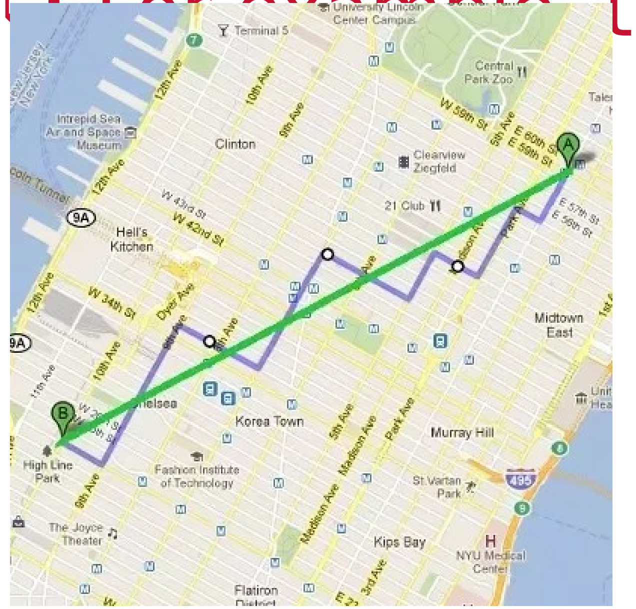
❑ For example, the statistical distance considers the covariance matrix Σ:

$$D = (x - \mu)^{\top}\Sigma^{-1}(x - \mu)$$



Drawbacks of Euclidian Distance

- You must scale your data
- It completely ignores relationships between variables
- If you have outliers you might consider the Manhattan distance.
- For example, the Manhattan distance uses absolute



$$D = \left\| x - z \right\|_M = |x_1 - z_1| + |x_2 - z_2| + \ldots + |x_p - z_p|$$

How do we Distance Between Clusters?

- ❑ Minimum Distance
- ❑ Maximum Distance
- ❑ Average Distance
- ❑ Centroid Distance

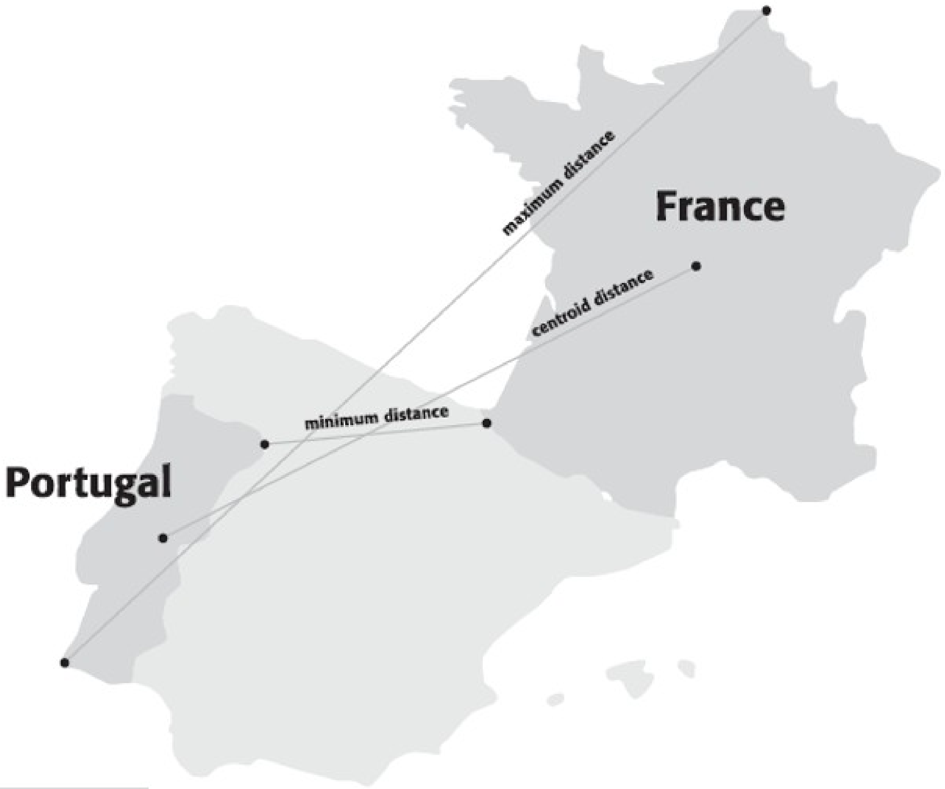
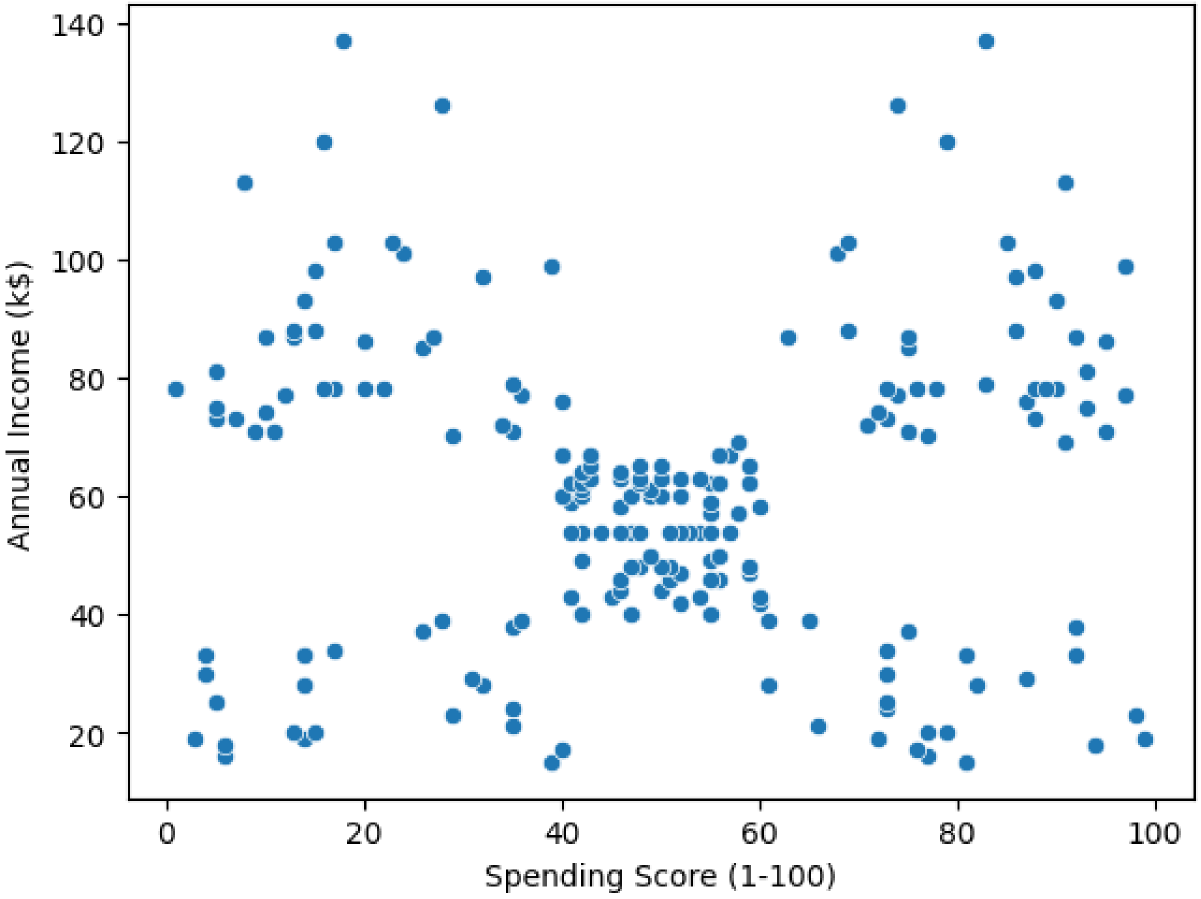


FIGURE 14.2 TWO-DIMENSIONAL REPRESENTATION OF SEVERAL DIFFERENT DISTANCE MEASURES BETWEEN PORTUGAL AND FRANCE

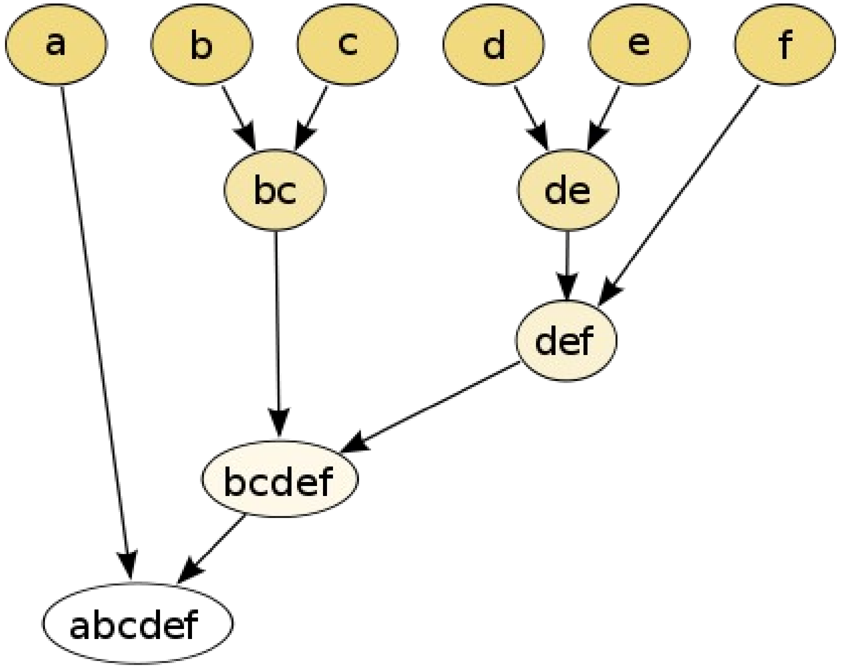
How do we Distance Between Clusters?

- ❑ Minimum Distance
- ❑ Maximum Distanc
- ❑ Average Distance
- ❑ Centroid Distance



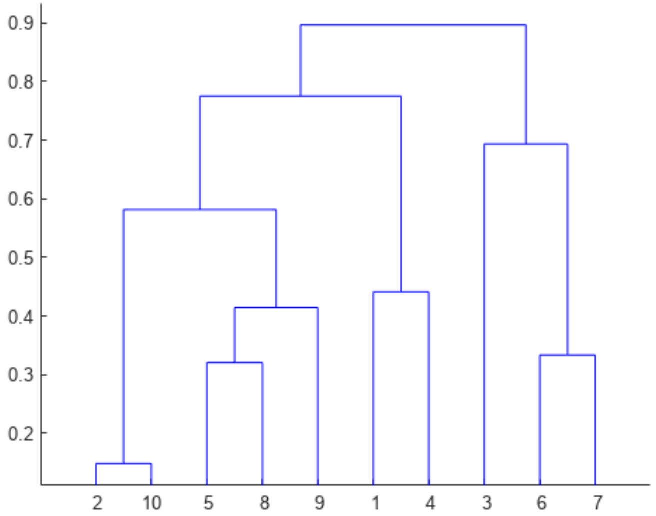
Types of Clustering Techniques

- ❑ Hierarchical Methods
- ❑ Non-Hierarchical Methods



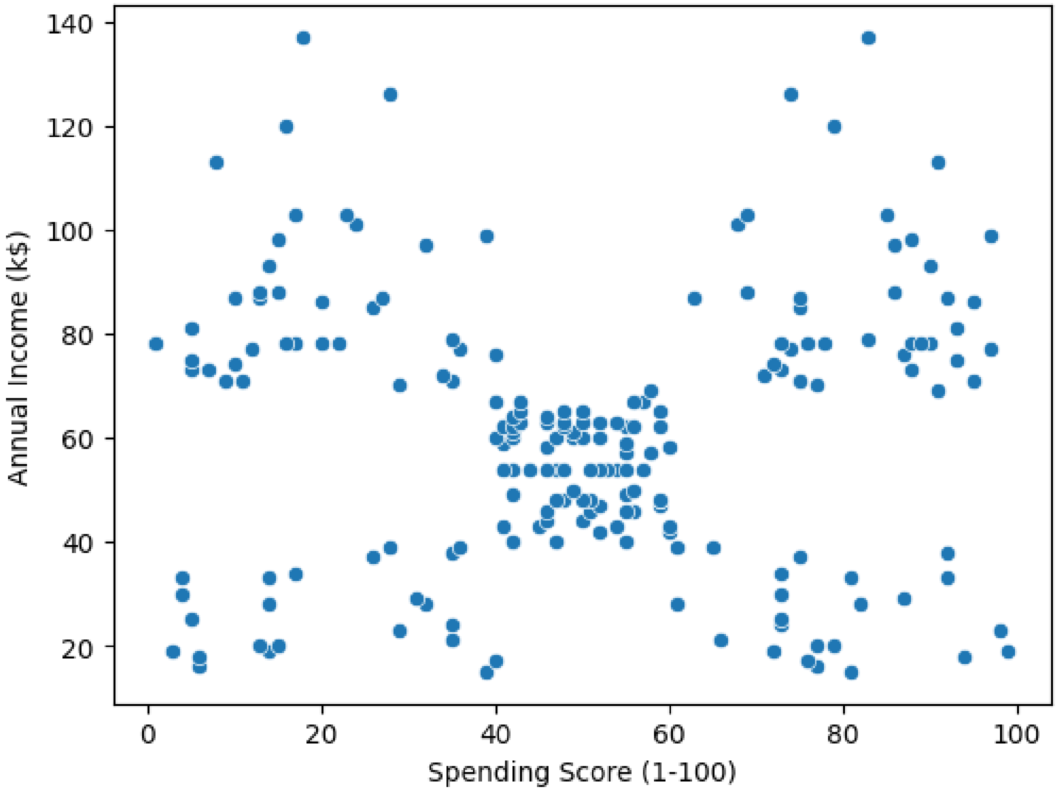
Hierarchical Methods

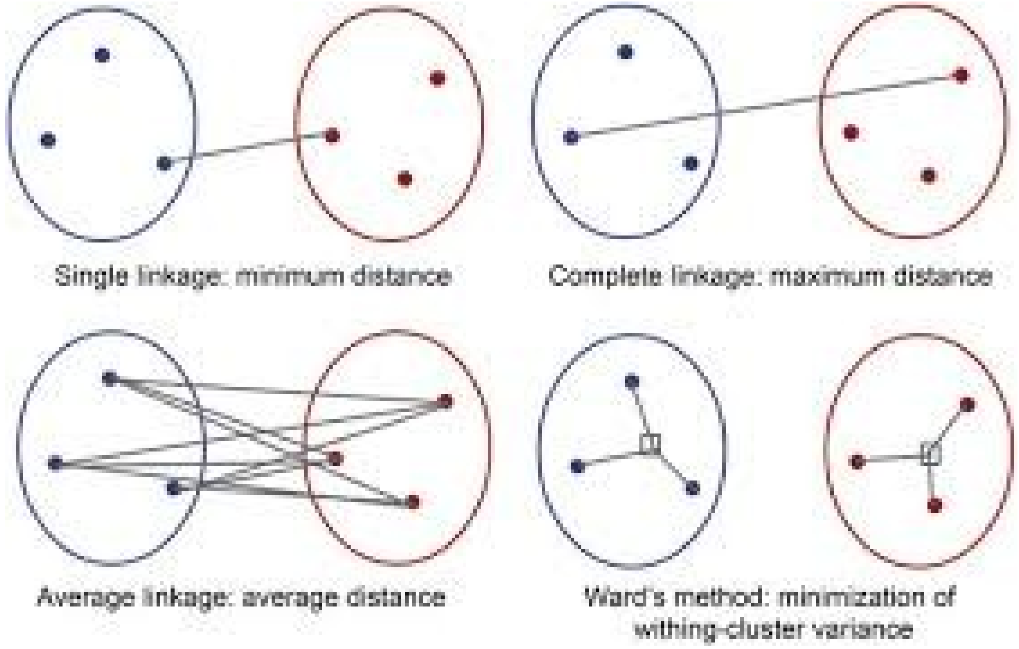
- Starts with each cluster comprising a small or one number of observation.
- Progressively combine the two nearest clusters until there is just one cluster left at the end which consists of all observations.



Hierarchical Methods

❑ Ward’s Linkage Algorithm





Single linkage: minimum distance

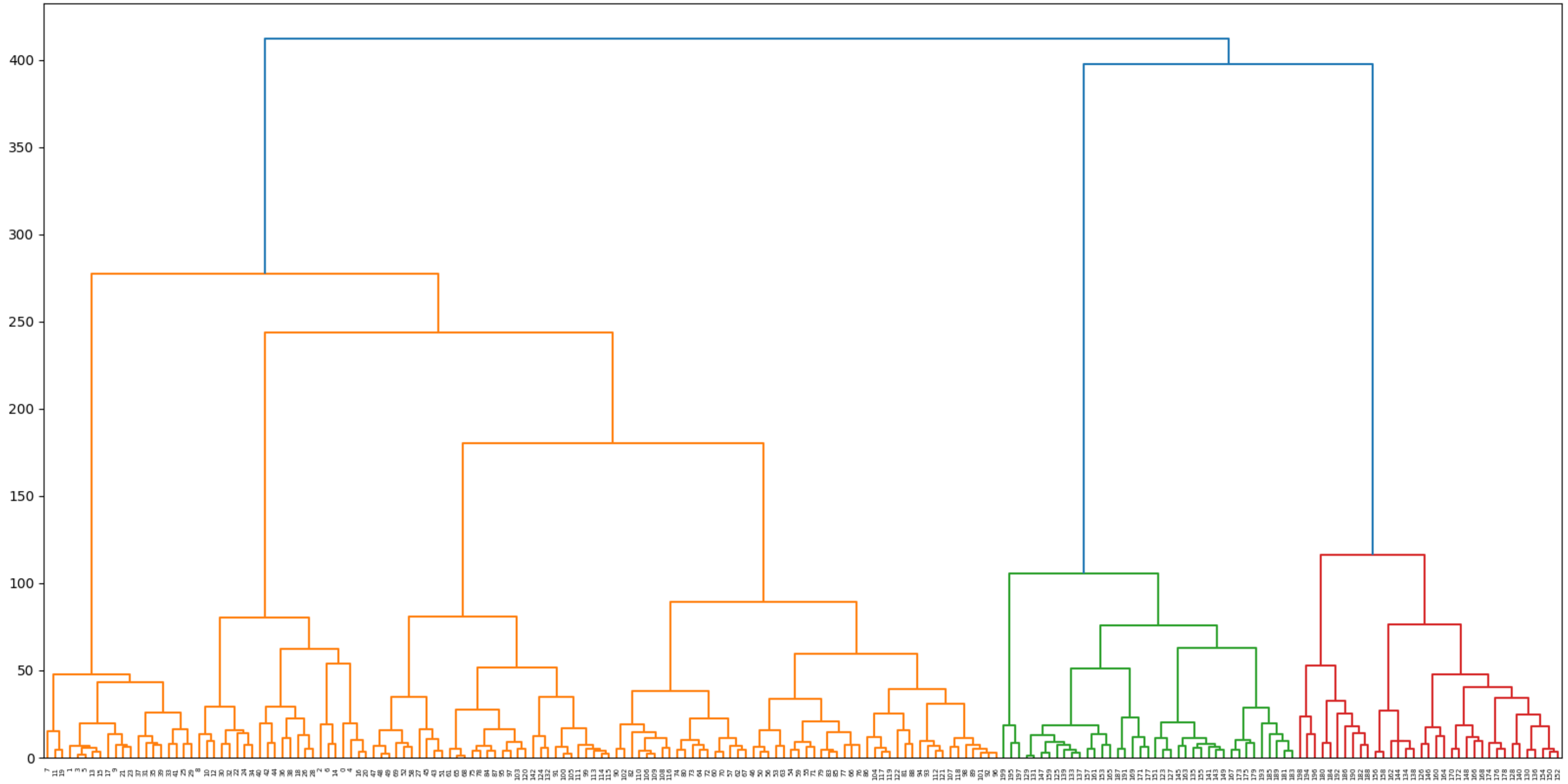
Complete linkage: maximum distance

Average linkage: average distance

Ward's method: minimization of withing-cluster variance

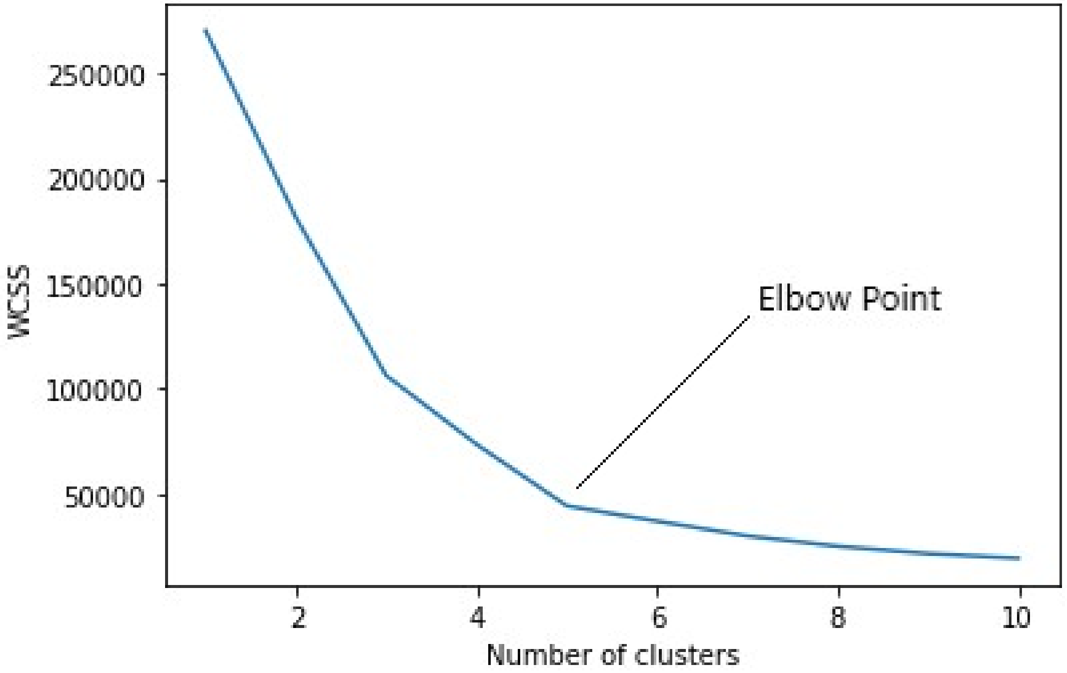
How to Find Number of Clusters

- Dendogram:



How to Find Number of Clusters

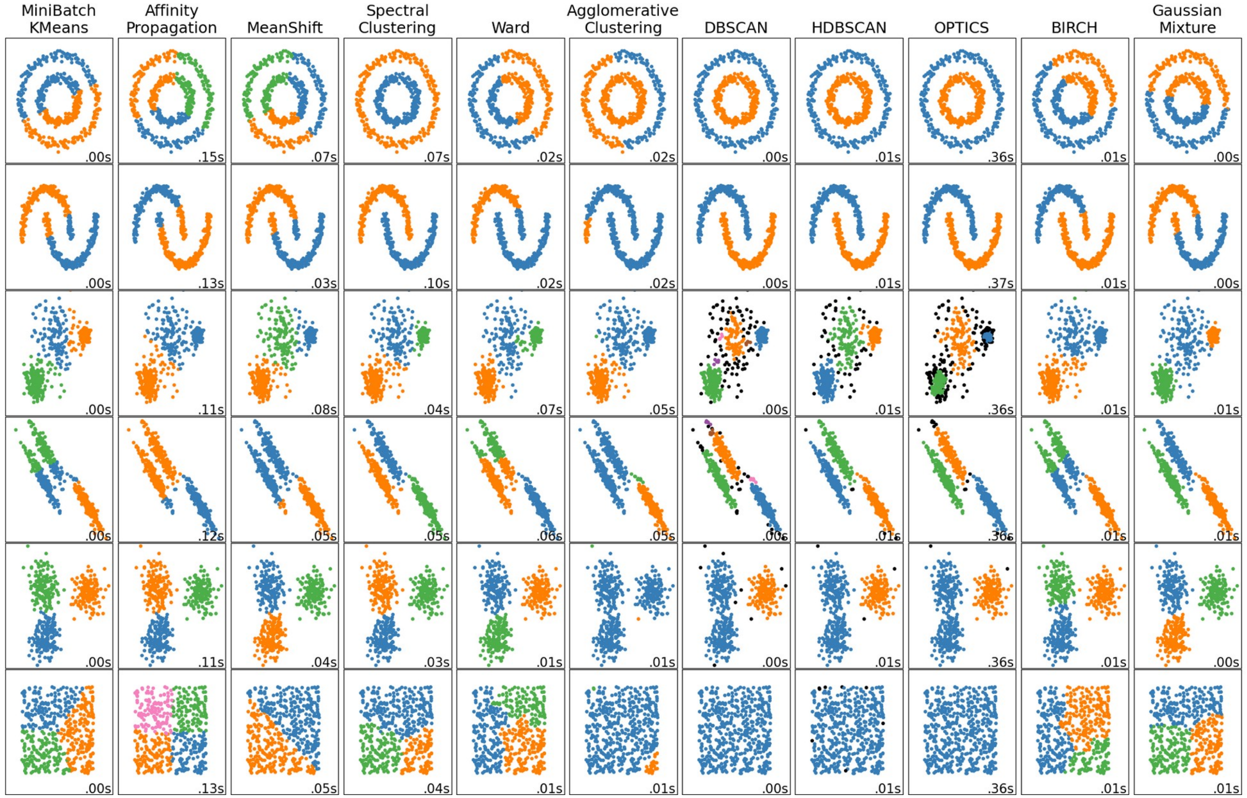
- ❑ Elbow Method: we fit the clustering algorithm for various values of �, say 1 to 10 and determine where there is a leveling of the the within-cluster sum of squares (WCSS).



Python

Non-Hierarchical Methods

- Iteratively assign objects to different groups while searching for some optimal value of the criterion
- K-means
- DBSCAN



Python