Advanced Analytical Theory and Methods: Clustering

Overview of Clustering

- Clustering is the use of unsupervised techniques for grouping similar objects
  - Supervised methods use labeled objects
  - Unsupervised methods use unlabeled objects
- Clustering looks for hidden structure in the data, similarities based on attributes
  - Often used for exploratory analysis
  - No predictions are made

K-means Algorithm

- Given a collection of objects each with n measurable attributes and a chosen value k of the number of clusters, the algorithm identifies the k clusters of objects based on the objects proximity to the centers of the k groups.
- The algorithm is iterative with the centers adjusted to the mean of each cluster's n-dimensional vector of attributes

Use Cases

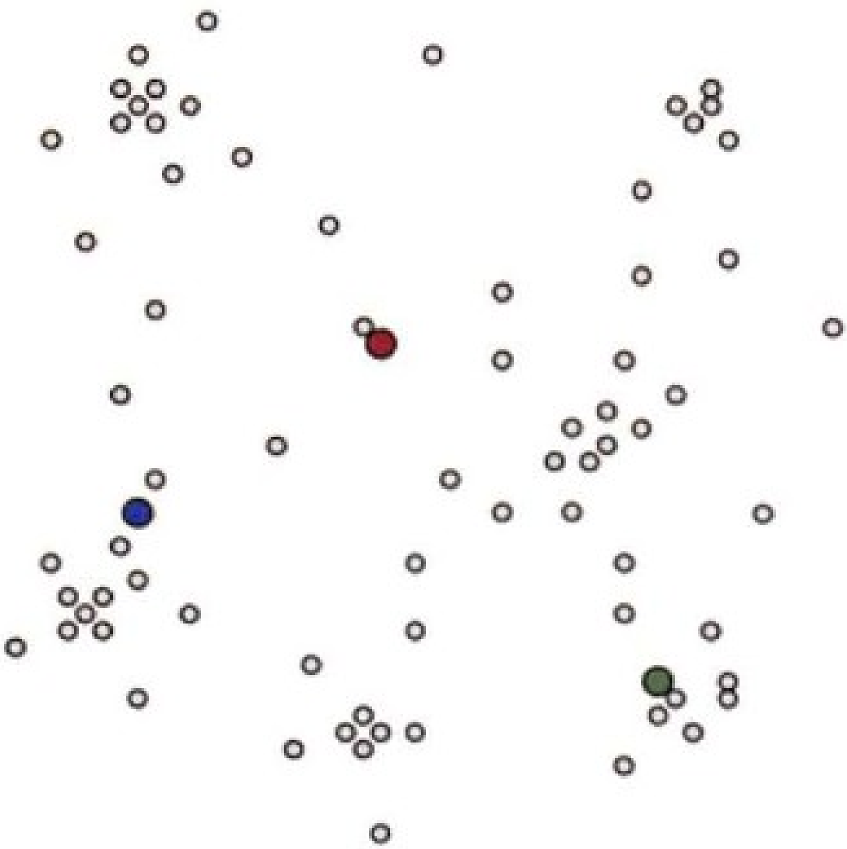
- Clustering is often used as a lead-in to classification, where labels are applied to the identified clusters
- Some applications
  - Image processing
    - With security images, successive frames are examined for change
  - Medical
    - Patients can be grouped to identify naturally occurring clusters
  - Customer segmentation
    - Marketing and sales groups identify customers having similar behaviors and spending patterns

Overview of the Method
Four Steps

1. Choose the value of k and the initial guesses for the centroids
2. Compute the distance from each data point to each centroid, and assign each point to the closest centroid
3. Compute the centroid of each newly defined cluster from step 2
4. Repeat steps 2 and 3 until the algorithm converges (no changes occur)

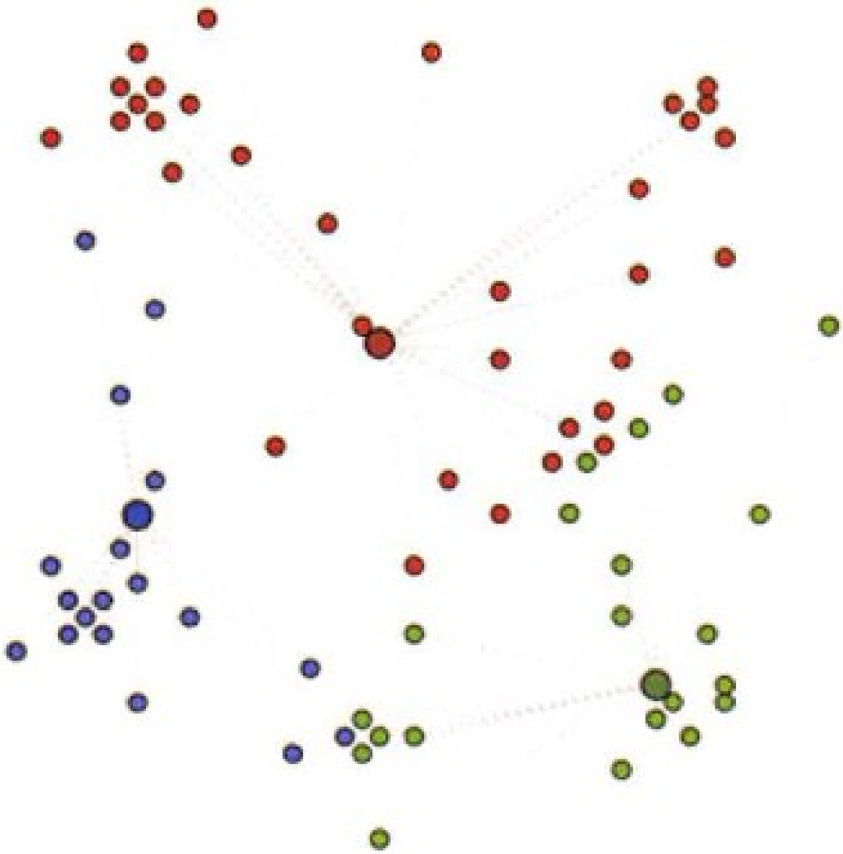
Example – Step 1

Set k = 3 and initial clusters centers



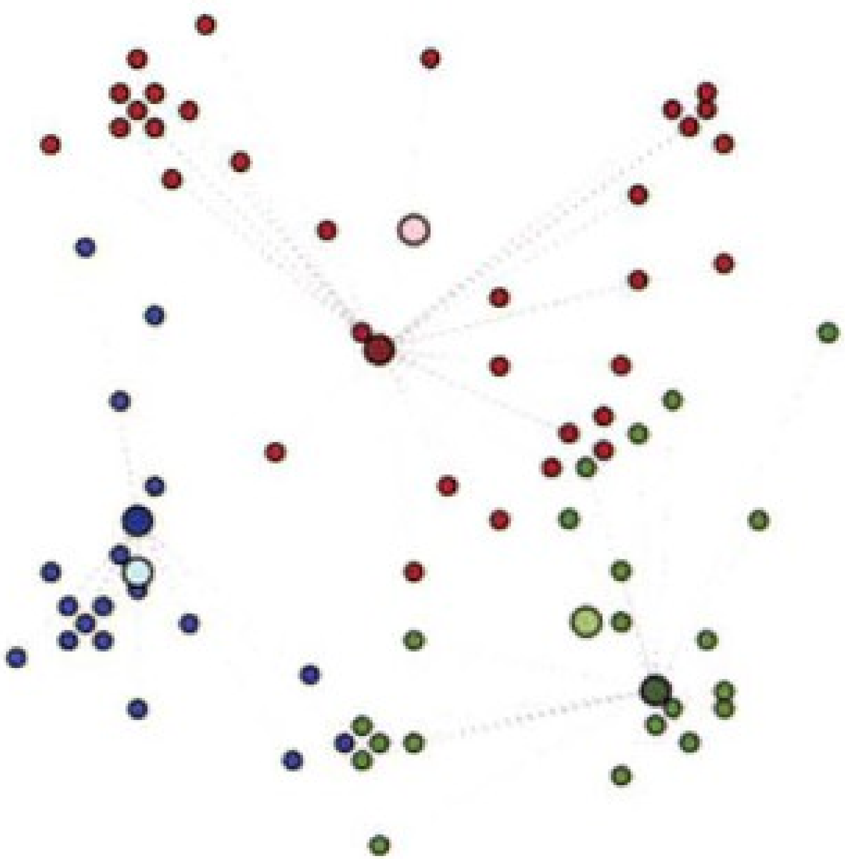
Overview of the Method
Example – Step 2

Points are assigned to the closest centroid



Overview of the Method
Example – Step 3

Compute centroids of the new clusters



Example – Step 4

- Repeat steps 2 and 3 until convergence
- Convergence occurs when the centroids do not change or when the centroids oscillate back and forth
  - This can occur when one or more points have equal distances from the centroid centers
- Videos

http://www.youtube.com/watch?v=aiJ8II94qck
  - https://class.coursera.org/ml-003/lecture/78

Common Distance measures:

- *Distance measure* will determine how the *similarity* of two elements is calculated and it will influence the shape of the clusters.

  They include:

1. The Euclidean distance (also called 2-norm distance) is given by:

2. The Manhattan distance (also called taxicab norm or 1-norm) is given by:

$$d(x,y) = \sum_{i=1}^{p} | x_i - y_i |$$

Common Distance measures:

if p = (p1, p2,..., pn) and q = (q1, q2,..., qn) are two points in Euclidean n-space, then the distance (d) from p to q, or from q to p is given by :

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

A Simple example showing the implementation of k-means algorithm (using K=2)

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Step 1:

Initialization: Randomly we choose following two centroids (k=2) for two clusters.

In this case the 2 centroid are: m1=(1.0,1.0) and m2=(5.0,7.0).

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

	Individual	Mean Vector
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

Step 2:

Thus, we obtain two clusters containing: {1,2,3} and {4,5,6,7}.

Their new centroids are: c1= (1.8,2.3), c2= (4.1,5.4)

	Cluster 1		Cluster 2	
Step	Individual	Mean Vector (centroid)	Individual	Mean Vector (centroid)
1	1	(1.0, 1.0)	4	(5.0, 7.0)
2	1, 2	(1.2, 1.5)	4	(5.0, 7.0)
3	1, 2, 3	(1.8, 2.3)	4	(5.0, 7.0)
4	1, 2, 3	(1.8, 2.3)	4, 5	(4.2, 6.0)
5	1, 2, 3	(1.8, 2.3)	4, 5, 6	(4.3, 5.7)
6	1, 2, 3	(1.8, 2.3)	4, 5, 6, 7	(4.1, 5.4)

Step 2:

- Thus, we obtain two clusters containing:
  {1,2,3} and {4,5,6,7}.
- Their new centroids are:

$$m_1 = \left(\frac{1}{3}(1.0+1.5+3.0), \frac{1}{3}(1.0+2.0+4.0)\right) = (1.83, 2.33)$$

$$m_2 = \left(\frac{1}{4}(5.0+3.5+4.5+3.5), \frac{1}{4}(7.0+5.0+5.0+4.5)\right)$$

$$= (4.12, 5.38)$$

Individual	Centroid 1	Centroid 2
1	0	7.21
2 (1.5, 2.0)	1.12	6.10
3	3.61	3.61
4	7.21	0
5	4.72	2.5
6	5.31	2.06
7	4.30	2.92

$$d(m_1, 2) = \sqrt{|1.0-1.5|^2 + |1.0-2.0|^2} = 1.12$$

$$d(m_2, 2) = \sqrt{|5.0-1.5|^2 + |7.0-2.0|^2} = 6.10$$

Step 3:

- Now using these centroids we compute the Euclidean distance of each object, as shown in table.

- Therefore, the new clusters are:
  {1,2} and {**3**,4,5,6,7}

- Next centroids are: m1=(1.25,1.5) and m2 = (3.9,5.1)

Individual	Distance to mean (centroid) of Cluster 1	Distance to mean (centroid) of Cluster 2
1	1.5	5.4
2	0.4	4.3
3	2.1	1.8
4	5.7	1.8
5	3.2	0.7
6	3.8	0.6
7	2.8	1.1

- Step 4 :

  The clusters obtained are:
  {1,2} and {3,4,5,6,7}

- Therefore, there is no change in the cluster.
- Thus, the algorithm comes to a halt here and final result consist of 2 clusters {1,2} and {3,4,5,6,7}.

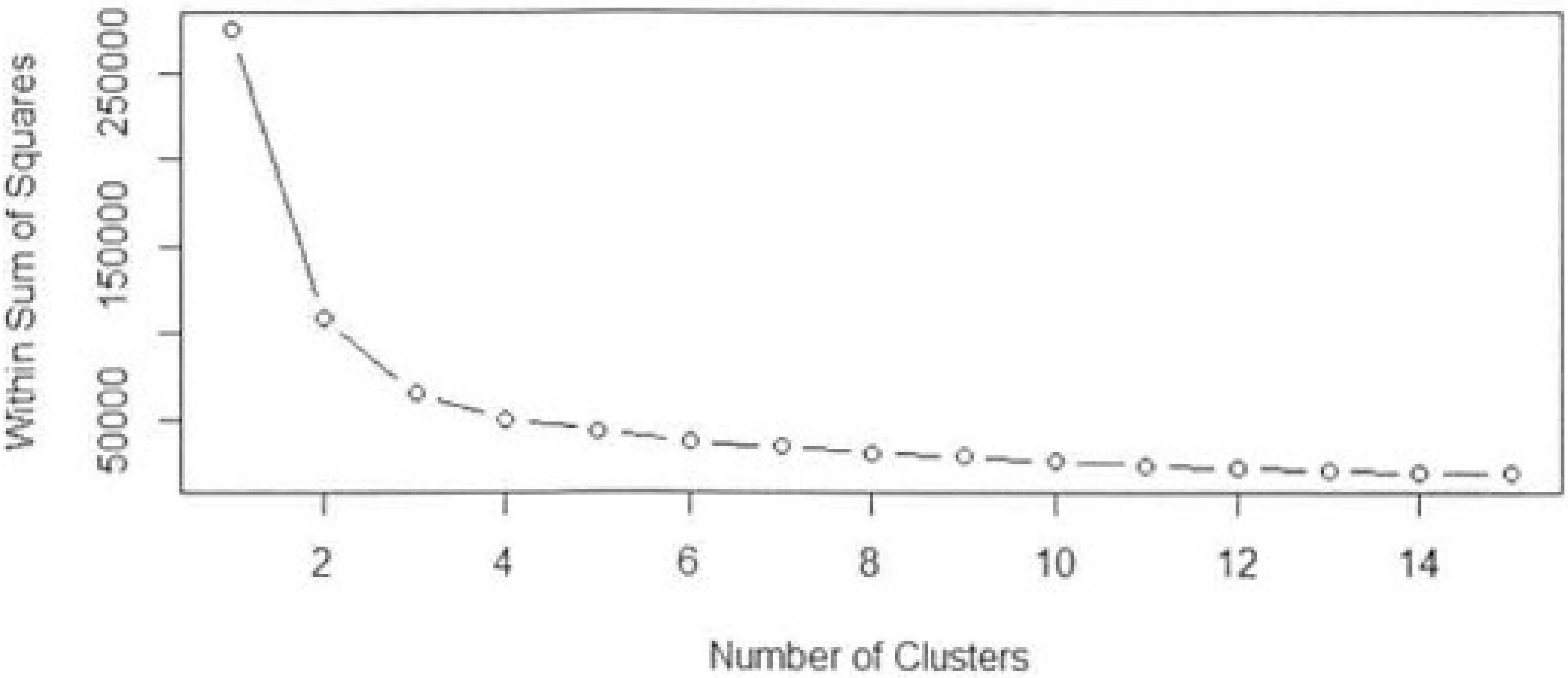
Determining Number of Clusters

- Reasonable guess
- Predefined requirement
- Use heuristic – e.g., Within Sum of Squares (WSS)
  - WSS metric is the sum of the squares of the distances between each data point and the closest centroid
  - The process of identifying the appropriate value of k is referred to as finding the “elbow” of the WSS curve

$$\sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{p} \left(x_{ij} - \bar{x}_{kj}\right)^2$$

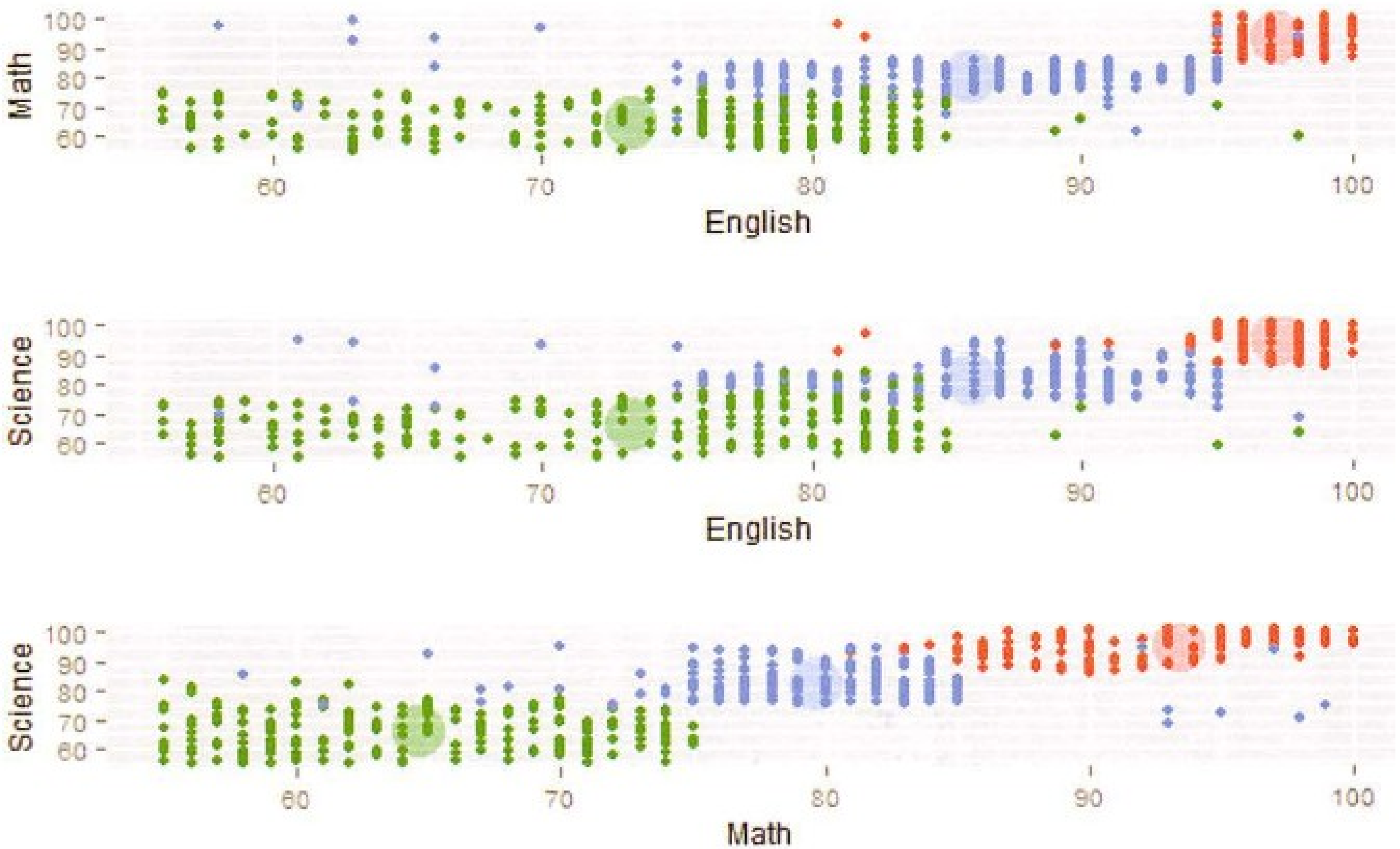
where S_k is the set of observations in the kth cluster and $\bar{x}_{kj}$ is the jth variable of the cluster center for the kth cluster.

Determining Number of Clusters
Example of WSS vs #Clusters curve



The elbow of the curve appears to occur at k = 3.

Determining Number of Clusters
High School Student Cluster Analysis

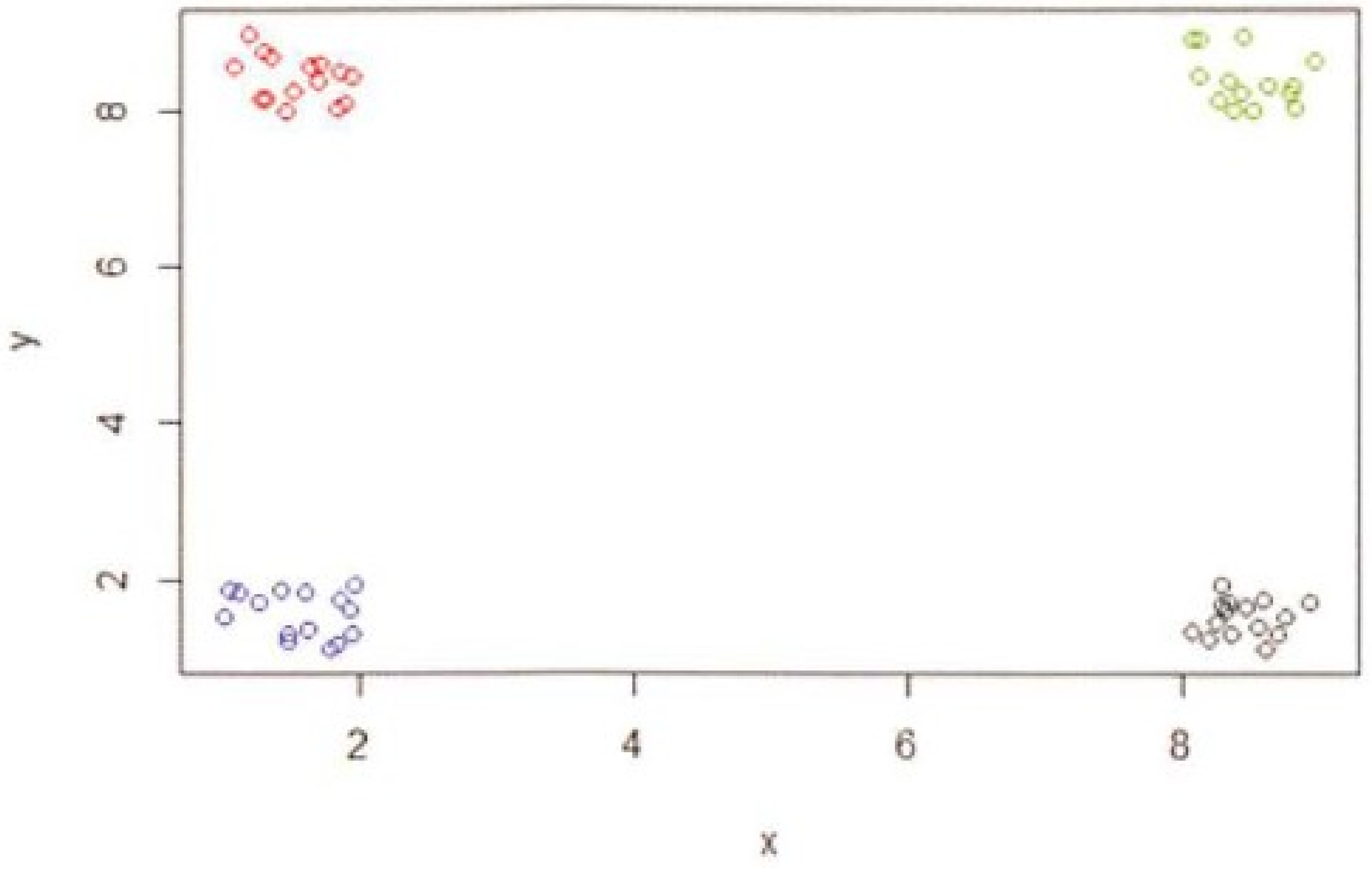


Diagnostics

- When the number of clusters is small, plotting the data helps refine the choice of k
- The following questions should be considered
  - Are the clusters well separated from each other?
  - Do any of the clusters have only a few points
  - Do any of the centroids appear to be too close to each other?

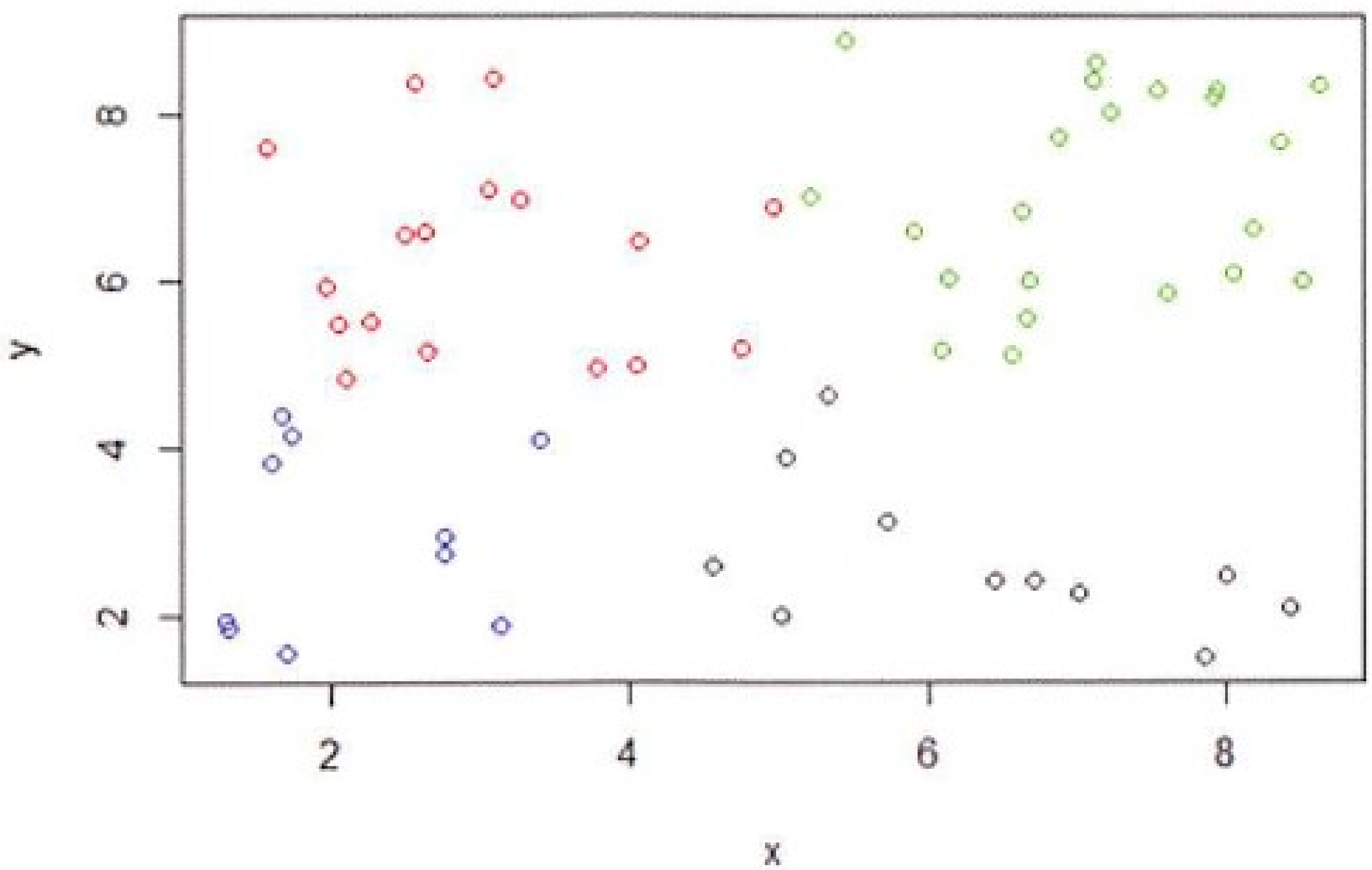
Diagnostics

Example of distinct clusters



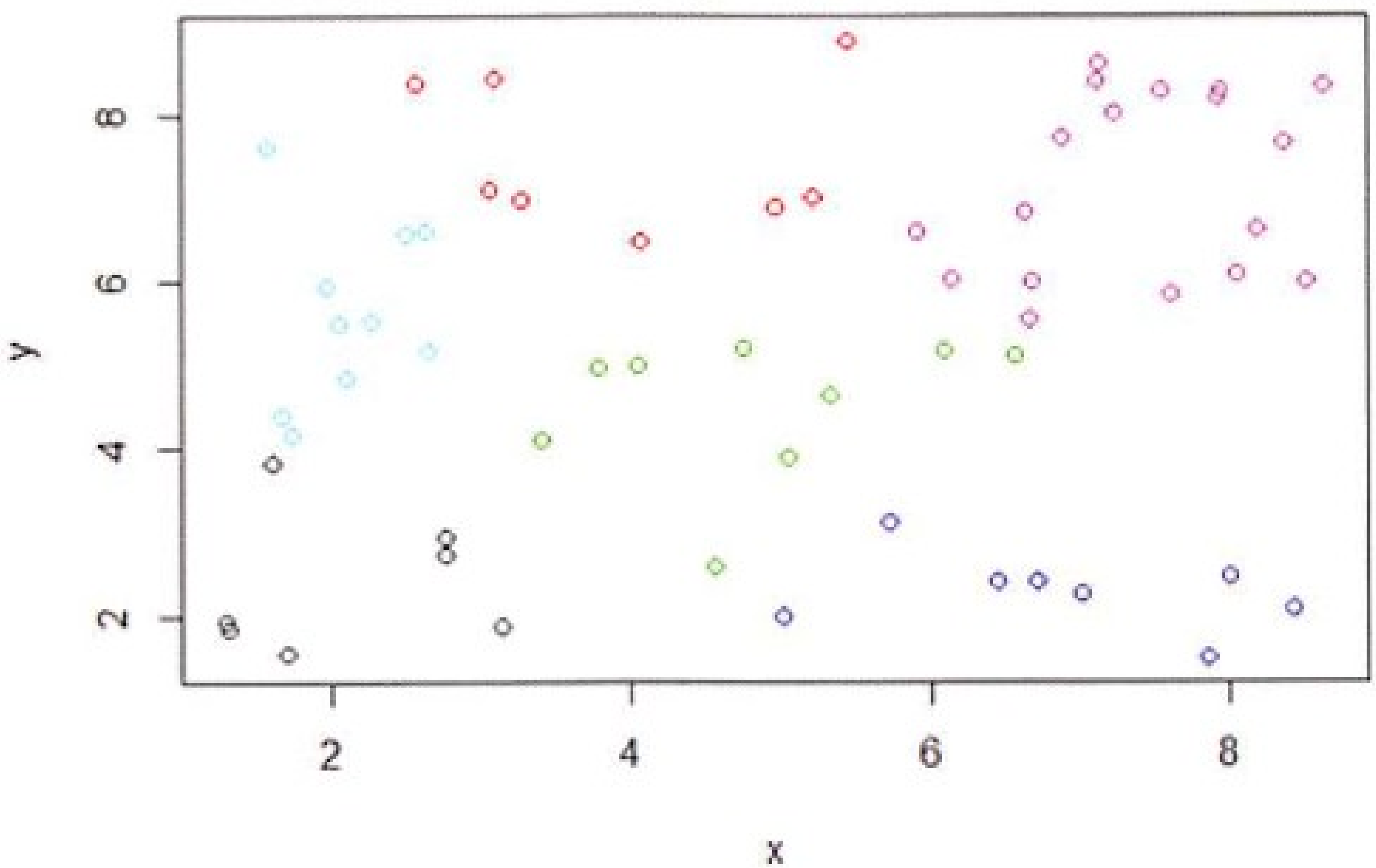
Diagnostics

Example of less obvious clusters



Diagnostics

Six clusters from points of previous figure



Reasons to Choose and Cautions

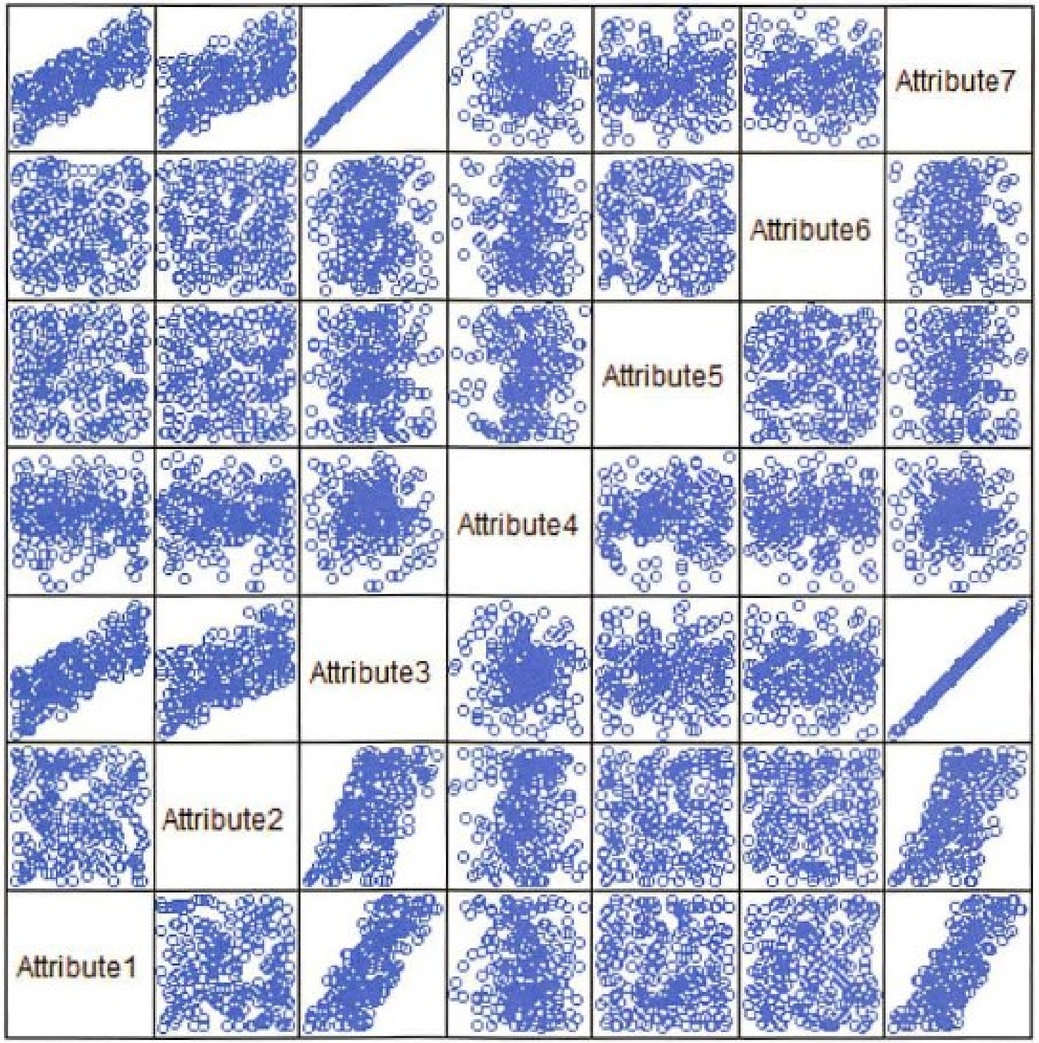
- Decisions the practitioner must make
  - What object attributes should be included in the analysis?
  - What unit of measure should be used for each attribute?
  - Do the attributes need to be rescaled?
  - What other considerations might apply?

Reasons to Choose and Cautions
Object Attributes

- Important to understand what attributes will be known at the time a new object is assigned to a cluster
  - E.g., customer satisfaction may be available for modeling but not available for potential customers
- Best to reduce number of attributes when possible
  - Too many attributes minimize the impact of key variables
  - Identify highly correlated attributes for reduction
  - Combine several attributes into one: e.g., debt/asset ratio

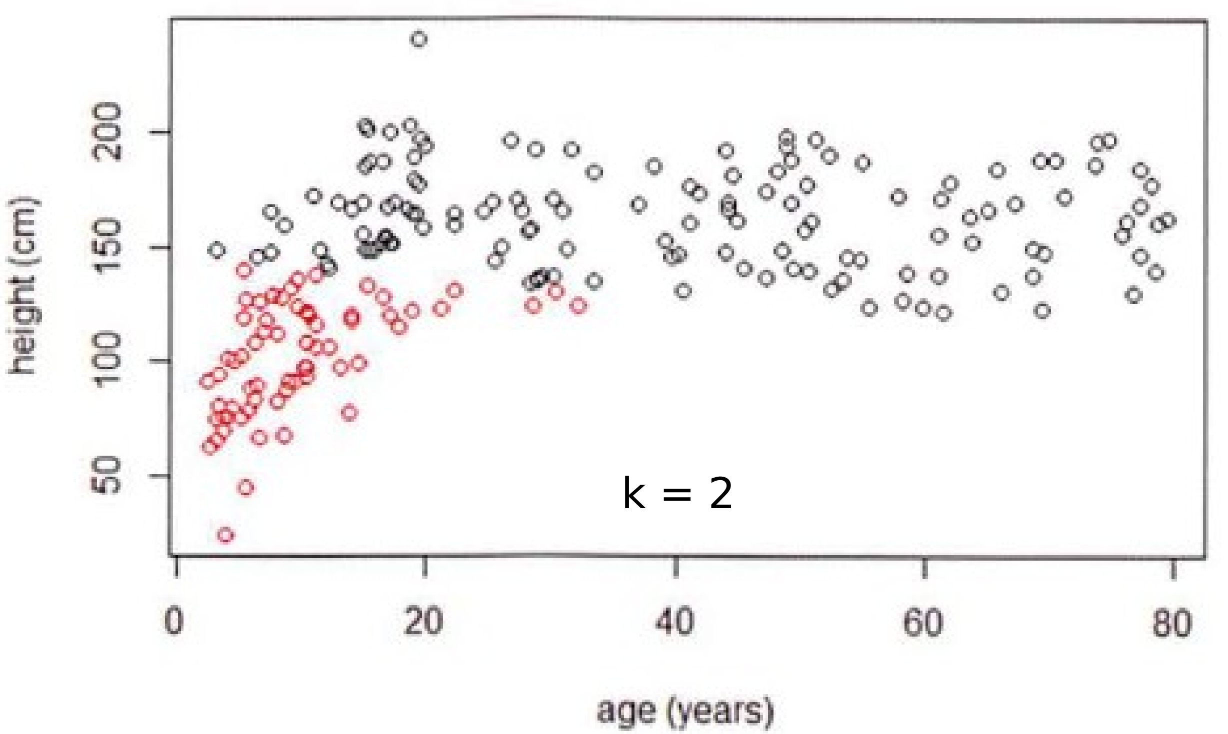
Reasons to Choose and Cautions

Object attributes: scatterplot matrix for seven attributes



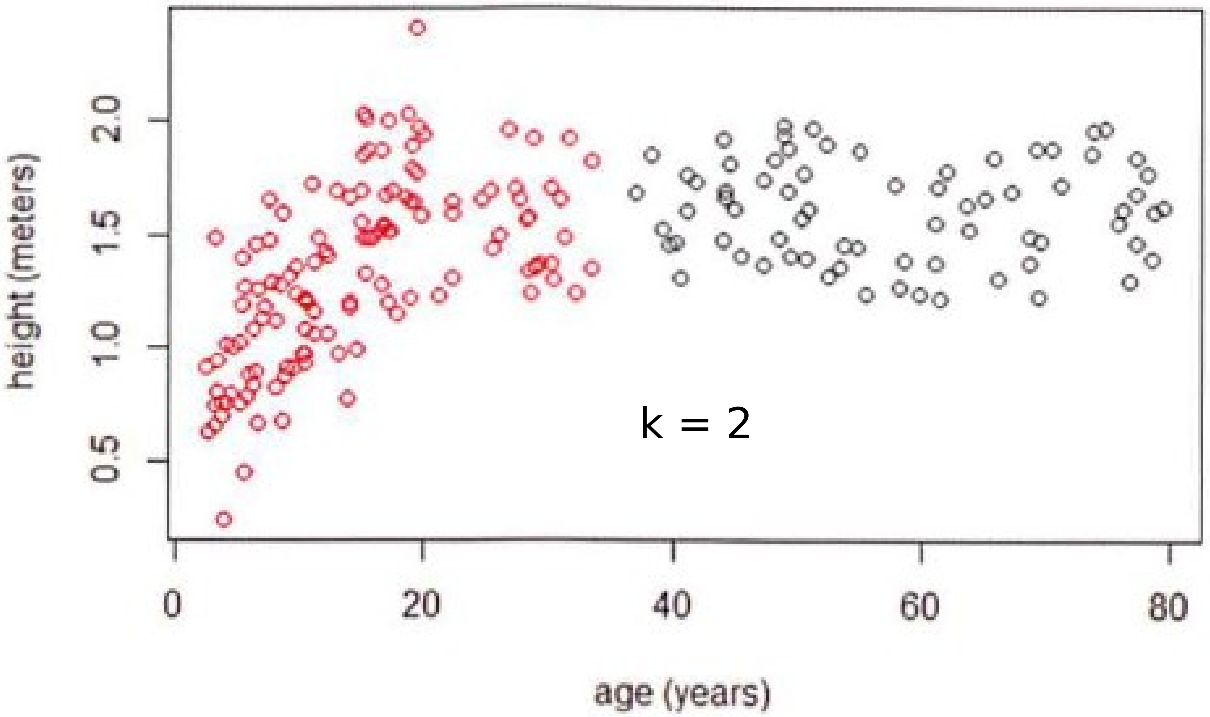
Reasons to Choose and Cautions
Units of Measure

- K-means algorithm will identify different clusters depending on the units of measure



Reasons to Choose and Cautions

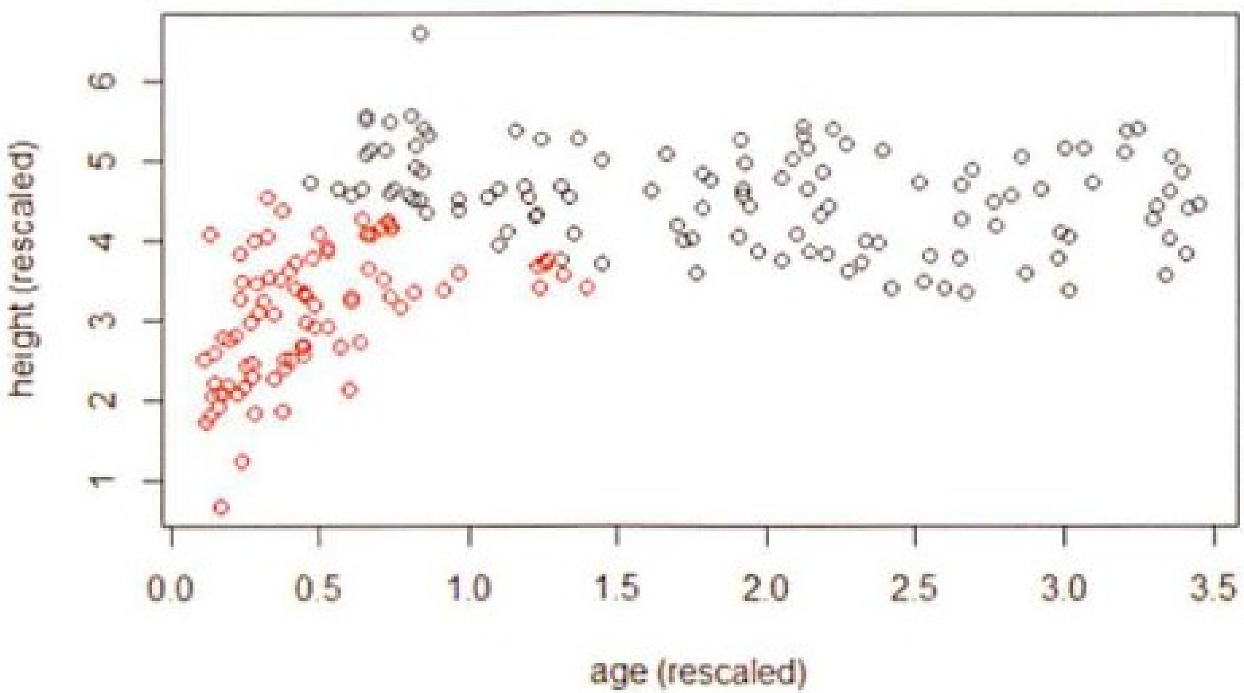
Units of Measure



Age dominates

Reasons to Choose and Cautions
Rescaling

- Rescaling can reduce domination effect
  - E.g., divide each variable by the appropriate standard deviation



Rescaled attributes

Reasons to Choose and Cautions
Additional Considerations

- K-means sensitive to starting seeds
  - Important to rerun with several seeds – R has the `nstart` option
- Could explore distance metrics other than Euclidean
  - E.g., Manhattan, Mahalanobis, etc.
- K-means is easily applied to numeric data and does not work well with nominal attributes
  - E.g., color

Additional Algorithms

- K-modes clustering
  - `kmod()`
- Partitioning around Medoids (PAM)
  - `pam()`
- Hierarchical agglomerative clustering
  - `hclust()`

Summary

- Clustering analysis groups similar objects based on the objects’ attributes
- To use k-means properly, it is important to
  - Properly scale the attribute values to avoid domination
  - Assure the concept of distance between the assigned values of an attribute is meaningful
  - Carefully choose the number of clusters, k
- Once the clusters are identified, it is often useful to label them in a descriptive way